

CHAPTER 1

Section 1.1: Understanding The Industry

This section offers a detailed summary of an expert interview conducted over three hours across two days, regarding a supermarket within the UK. The information has been paraphrased to highlight the main insights and findings, yet still includes enough detail to lean into further investigation and reasonable justification for chapters to come. This summary is intended for both the current project and as a resource for future research and analysis. The original interview was extensive, so this summary helps to capture and organise the most important points, yet still give a substantial amount of detail. My goal is to make these insights accessible for ongoing and deeper exploration.

1.1 Day-to-Day Operations

1.1.1 Daily Operations

1.1.1.1 Typical Working Day

A typical day for a retail assistant within the expert's supermarket begins with a partial restock. The supermarket opens to the public at 8am, yet the "morning shift" employees will start work from 6am, with their main priority being restocking of items. Primarily, this concerns fresh produce such as fruit and vegetables, in addition to everyday items such as bread, milk, newspapers, etc. Their duties also include the rotating of stock by bringing the products with the most recent date to the front, alongside ensuring the items on the shelves are 'forward-facing' for visual appeal and easier to find for customers, also alluding to fuller shelves. The retail assistants who work the morning shift will usually finish their shift around 2pm, with a changeover of employees who will then work from 2pm until 10pm, i.e. store closure.

At this point in the interview, the expert was asked to elaborate on the restocking process, specifically regarding the time it takes, and the process involved. They explained that restocking products in the morning usually takes around one hour, all retail workers on shift are involved, and since they have worked within that supermarket there has been no issue with timing concerning restock since it is a rather small delivery and very few updates are made to the inventory records, but the extra hour in the morning allows retail workers to assist with returns by inspecting the returned items and placing them back within the supermarket or assisting with product recall. The expert did not expand on any more activities although prompted, and stated they are sure there may be more activities they are able to do, yet those are the ones they are aware of and have completed themselves. The

expert also noted that product recall was not a daily operation and just an example of a task. The expert was asked about the potential for implementing automated processes during the morning restocking or having staff from the later shift prepare non-perishable products in advance. While the expert acknowledged that automation could be beneficial, they expressed uncertainty about how it could be effectively integrated into the current operations. They mentioned that the existing workflow among staff is already well-coordinated, and there are concerns that the costs of implementing automation might outweigh the potential benefits. Additionally, the expert noted that there seems to be a general reluctance to embrace automation at the store level, even though it has been discussed. This reluctance is largely due to these decisions being made at the store level rather than by higher management.

When asked about stock shortage, the expert explained that “section leaders” are responsible for addressing stock shortages by monitoring inventory, ordering more stock as needed, and ensuring prompt restocking. However, they often fail to act, resulting in empty shelves. The expert noted they are under significant pressure and might overlook minor stock shortages, such as cereal or biscuits, compared to essential items like bread, fruit, vegetables, or newspapers. The expert was then asked if this ‘shortage’ was due to section leader’s workload or possible communication gaps within the team, particularly with higher management. The expert responded that, as far as they know, the issue seems to stem more from forgetfulness rather than any serious underlying problems. They did not provide much further elaboration.

The expert continued to go on and say that a retail assistant is trained in all departments within the umbrella term of a “retail worker”. Their initial training allows them to be versatile in various duties within the supermarket. Upon opening the supermarket to the public (8am), retail assistants can be allocated any duty such as checkout, customer service desk, self-scan or “shelve stacker” which entails continuous stock of the shelves within a specific section, for example fresh produce, fridges, freezers, or general merchandise, usually areas of high purchase rates. The expert was then asked to elaborate on each duty a retail assistant may have on a daily basis.

Firstly, operating the checkout is a core component within the role of a retail worker. Handling various payment methods securely, processing transactions, resolving payment issues and being proficient in point-of-sale systems are the main tasks associated with the duty of being a cashier. In addition, retail workers are also allocated to ‘self-scan’ which refers to the self-service system where customers can scan and pay for their items themselves. However, not all transactions are successful, and retail workers are there to help with technical issues that may arise, to guide customers who are not tech-savvy or check identification for age-restricted products, although any product scanned through the self-scan will be flagged and need to be cleared via the scanning of the retail assistants’ employee card, which verifies their identity, and thus can track their actions of approval or refusal of sale. The expert was then asked about the types of issues that customers may face. They stated that the self-service systems have continuous issues surrounding item weight within the bagging area, where the customer may not place the item in the bagging area or the item is seen as not the correct weight from the systems data, hence causing the system to flag the purchase and thus needing approval from a retail assistant (and their

employee card) to approve the remainder of the sale. Although prompted, no further examples were given by the expert, however they did note that the self-scan area often lacks sufficient staffing, leading to long wait times for customers needing assistance.

The expert then went on to talk about the "shelve stacking" element of the role. Retail assistants will bring large cages from the warehouse and restock the shelves, ensuring all items are forward facing for visual appeal and easier to find products for customers due to label recognition. The expert did mention here that the duty of a "shelve stacker" will essentially be the same as what the morning shift retail assistants do, however it will continue for the remainder of their shift and allocated to a specific section such as fresh produce, fridges, freezers or general merchandise. The duty of restocking the shelves also includes ensuring products are correctly labelled and priced. If not, retail assistants will print out a new label and replace the old one. They also maintain digital inventory records by checking delivered products against the purchase orders and then update the inventory system to reflect the new stock levels.

The expert was asked, "Do you find that customer interactions are important and a daily occurrence for a retail assistant? How much time do you think you spend on these interactions, and what are the typical activities involved?" They responded that time engaging with customers is deemed very important by the supermarket, and thus employees must do everything to assist with the customer's needs and create a positive shopping experience by being friendly and helpful, especially when concerning feedback and complaints. When asked about frequent customer complaints and how they address them, the expert mentioned that complaints are usually centred around long checkout times and out-of-stock items. They try to assist customers as quickly as possible and report stock shortages to section leaders. Each interaction can range from a few minutes to over 15 minutes, especially if a customer needs help finding specific products that the retail assistant is not aware of, providing in-depth information about a product, or assisting with a complicated return or exchange. When asked about returns, the expert stated that most returns and exchanges are not simple or quick, even if the item is basic, as they usually need to be approved by management, which takes a bit of time to get, especially if management are already busy. Moreover, the customer service desk is the only place where returns can be processed, manned by a maximum of two people. The returns are usually centred around either clothing or household appliances.

Additionally, all staff have the responsibility to reduce theft. While a significant part of their day is not physically reducing theft, they remain alert for signs of theft. This vigilance involves monitoring for suspicious behaviours such as loitering in aisles, frequenting high-theft areas, and carrying large bags or wearing oversized clothing. They may need to run to and from the back wall, where high-value goods are usually kept, several times during their shift to ensure these items are secure.

The expert was asked "Can you elaborate on the 'signs of theft' you have spoken about? What does the anti-theft protocols entail?". In the past 12 months, anti-theft hanging arms have been introduced. These devices are designed to deter theft by making it more difficult for shoplifters to discreetly remove items from displays. When an item is removed incorrectly, a click sound is heard, alerting staff to potential theft.

The anti-theft protocol includes several key components. One significant measure is the installation of one-way barriers near exits to prevent quick escapes. These barriers require customers to walk around the supermarket to the checkouts before exiting, reducing the likelihood of theft as customers cannot easily run out the same door with stolen goods. Security staff now monitor entry and exit points more diligently, enhancing overall supermarket security.

Additionally, high-value goods, originally located near the front of the supermarket in the general merchandise section, have been strategically moved to the back of the supermarket. This relocation makes it harder for potential thieves to run from the back to the front unnoticed. Regular inventory audits and random checks are conducted to identify any discrepancies that might indicate theft, helping to maintain accurate stock levels and identify problem areas.

The expert used the term 'now' to imply recent changes, thus it was important to elaborate on the timeframe. While the number of cameras has not been increased, security guards now work until 10pm rather than 6pm, with the number of guards increasing from one to two. This staffing change occurred mid-2023. Additionally, all staff members have been trained to deal with security measures, not just anti-theft hooks. For the past two years, staff have been trained and retrained every six months to recognise suspicious behaviours such as loitering in aisles, frequenting high-theft areas, and carrying large bags or wearing oversized clothing.

When the expert was asked about recent theft incidents, they reported that it had been over a year since any actual theft was recorded. Most issues were related to errors with scan-and-go or self-checkout systems, typically involving inexpensive items. It was also noted that security usually handles disruptive customers rather than actual theft cases.

Given the high emphasis on security at the supermarket, the expert was asked whether they felt customers might feel overly monitored due to these new measures, perceiving them as more about store safety than customer safety, and whether there have been any complaints. The expert stated that while no formal complaints have been recorded, there is a concern that some customers might feel uncomfortable with the increased security presence, particularly with the visible surveillance systems and anti-theft measures. When asked who was 'concerned' about the increased security measures, the expert stated that it was mainly the retail staff who interact directly with customers. However, they noted that these concerns were minimal and largely subsided after the first week of implementation. When further questioned about what actions should be taken in light of these concerns, the expert suggested that while it might be beneficial for customer comfort to reduce the visible level of security in some areas, it is highly important to maintain strong security measures for the protection of the store and its financial assets.

Furthermore, it is also within the retail assistant's role to setup promotional displays which highlight special offers or seasonal products, in collaboration with the marketing team who send the briefs to all supermarket locations which is then shown to the retail assistants. The displays can either be standalone at the entrance to the supermarket or on the end of an

aisle which contains related products, also known as “end-bay” promotions. This process is time consuming, often taking between 2 to 8 hours to complete, depending on the size of the display, with changes typically occurring every 3 weeks, and is present within every section of the supermarket. The expert stated that 2 hours for a small project and 8 hours for a large project are the supermarkets expected timeframe yet can double or triple in time depending on the circumstance. When probing the cause of this variation of timeframes it became apparent that it becomes particularly time consuming when stock does not arrive on time or in full, requiring multiple adjustments and causing delays. When stock does arrive on time and in full the retail assistants are able to complete the tasks complying with company standards. Additionally, once the display is set up, it needs to gain management approval. If management is unsatisfied, sales assistants must dismantle and redo the display. This duty is completed alongside their regular duties, such as assisting customers or restocking shelves. When asked how they manage their time during peak hours, the expert did not suggest anything that could be done or why it was occurring, but they did say that just time management was a factor that was hard to control. During seasonal events, displays can change daily based on customer interest and trends. It is currently early summer when conducting this interview, so with the summer range coming out this is considered a daily operation for the expert at this moment in time. The expert did mention that they have personally noticed over the past few weeks of the summer event that the frequency of detailed customer interactions has increased, and these interactions are fewer and quicker on regular days in comparison, implying that large events are a primary factor regarding prolonged customer interactions.

1.1.1.2 Attraction & Retention of Customers

Customer service is deemed as a fundamental aspect of customer retention within the expert’s supermarket. The retail assistants are enrolled every 6 months in a mandatory training session based solely upon customer service skills and approaches. This training ranges from simple conversation between employees and customers, to showing compassion and empathy when assisting the elderly or disabled. It is not only for retail assistants, but other employees such as section leaders and domestics must also take part. Additionally, employees receive monthly newsletters highlighting the importance of customer service, alongside photographs of those within the company who have excelled in customer service that month. The initiative here is that employees who receive exceptional feedback, regarding their customer service from the public or another employee has witnessed and reported it, will get rewarded a voucher containing money to be used within the supermarket, thus receiving a reward for their dedication.

Another strategy to attract and retain customers is the new ‘reward’ program, which offers exclusive discounts and personalised offers. When asked how ‘new’ the reward system was, the expert stated it was only implemented in all stores within the past 12 months. Customers use an app with a scannable barcode to collect points towards spending in-store. Each time a customer shops and scans the barcode at checkout, 5% of their total expenditure is added to their balance. Additionally, novelty products and special weekly offers contribute to their balance, for example, purchasing five items of a particular coffee brand within a week will add £1.50 to their balance. The reward system encourages customers to choose this supermarket over competitors to gain more rewards and earn

cash back, in addition to encouraging them to return to use their rewards they have accumulated. These offers change weekly, and according to the expert, they typically focus on products that are not performing well rather than seasonal displays, though this observation is based on personal experience rather than evidence.

Promotions displayed throughout the supermarket are also shown on the customer's app, and notifications are sent via the app when new promotions or discounts become available. This approach ensures that customers are consistently informed about offers, encouraging repeat visits. The expert was asked about how the balance was displayed to the customers. They stated that they were displayed as 'pounds and pennies', rather than a point system that emulated money. When asked the reasoning for this, they said the supermarket perceives displaying a cash balance is more motivating than points, it shows customers how much money they are gaining and how much money they can spend, in a much more direct manner.

When asked to elaborate more on the effectiveness of promotional strategies the supermarket uses, such as large displays or end-bays, the expert gave an example of 'buy one, get one free' offers and bulk purchase discounts which attract significant attention from the public.

When asked if they observed any differences in spending patterns among different types of customers, the expert mentioned that regular customers generally spend more per visit than non-regular customers. Regular customers often buy a combination of everyday essentials and higher-margin items, frequently using loyalty program discounts. On the other hand, occasional shoppers typically purchase fewer items, focusing mainly on promotional or discounted products. They were confident that converting new customers into regular shoppers requires exceptional customer service, emphasising that they try to go 'above and beyond' for those customers who are seemingly new, or rarely visit the supermarket. When asked about how they determine if a customer may be 'new', the expert replied that when a customer does not know where the bread or milk are, or essential items, it is usually a good indication they are new to shopping with the supermarket. When asked if they could recall any new customers who didn't come back, the expert found it difficult to remember specific instances. When further questioned about why local or convenience-driven customers might not return, they did not have an answer. When asked about marketing strategies, the expert explained that the supermarket tailors its marketing efforts to different customer segments. For families, this includes promotions on bulk items, family-sized packages, and essential household products, often advertised on social media such as Facebook or television. Marketing campaigns for students or young professionals focus on convenience, such as ready-meals, usually advertised again on social media such as Facebook. For elderly customers, they are also targeted with promotions such as convenience and simplicity, emphasising traditional foods such as potatoes or coffee. The expert used the term 'old-school media' to describe the conventional way in which elderly were targeted by the supermarket in marketing strategies, relating to newspapers or magazine.

1.1.1.3 KPIs & Metrics

The expert was then asked to identify and explain the Key Performance Indicators within their supermarket. They began by stating that they monitor the amount of damaged goods within a supermarket, either through returns or when inspecting deliveries. This also encompasses undesirable or passed expiration dates. When asked how they track this they explained that their digital inventory management system allows items to be recorded as damaged or unfit for purpose. This also includes customer returns if they were damaged upon purchase, however the expert was unsure how products are deemed to be damaged before or after purchase, unless from a product recall, and stated that 'management' usually complete the inventory updates for those instances. The expert then explained that associated metric was the cost of damaged goods as a percentage of sales, indicating the total sales revenue that has been lost.

The expert then went on to talk about customer satisfaction being a measurement used throughout the supermarket. They explained that it shows their overall experience and the quality of staff interactions, ultimately aiming for returned visits and increase expenditure, boosting profits. They measure this via feedback forms and online reviews via the app. There are links to 'satisfaction' questionnaires to fill out at the bottom of each receipt, with questions relating to helpfulness of staff and success of visit to supermarket based on if they found all the products they were looking for. The expert also mentioned that when a customer fills in a feedback form of any variation, they are entered into a 'prize draw' for that week, implying that it is marketed as a way to encourage customers participation in providing feedback via a chance to win prizes.

Analysing unit sales is another way which the expert stated the supermarket tracks KPIs. It informs the supermarket about product performance and how inventory turnover affects profitability. It helps to prevent overstocking and markdowns, which is the reduction of selling price of products below their desired retail price. Additionally, with regards to overstocking, the expert stated that the supermarket buys in bulk to reduce unit cost, boosting profit margins per item sold. Unit sales and inventory management are tracked through the point-of-sale (POS) systems, which captures data in real-time, tracking inventory levels and unit sales throughout the supermarket. The only measurement the expert has observed regarding unit sales was the total sales revenue being subtracted from the cost of products sold, within categories such as fresh produce, thus providing gross profit margins.

Finally, the expert states that theft is heavily monitored within the supermarket. It is mainly conducted through inventory audits, whereas surveillance cameras and security systems are more for deterrence than actual measurement. The expert did not provide any further information, other than it is monitored on a weekly basis.

1.1.1.4 Customer Interaction Process

Handling a typical customer interaction in the supermarket begins with a warm greeting if customers are near to or walk past a member of staff. If the customer wants to ask the retail assistant a question, they will listen to try to better understand the customer's needs. The retail assistant will then provide clear information and guidance relevant to the customer's needs. Another retail assistant will be at the checkout area, ready to warmly greet the

customer again and process the transaction. This is also the case for customer service desk or self-checkout. After the transaction is completed, they offer additional assistance if needed, and thanks them.

1.1.2 Challenges and Decision Making

1.1.2.1 Routine Decisions

When asked about the daily routine decisions made in the supermarket, the expert began by addressing the replenishment of stock. This involves monitoring product levels throughout the day, and deciding if more products need to be moved from the warehouse into the main supermarket, or if new orders with suppliers need to be made, due to a substantial amount of stock being sold. The expert recalls that last week, due to the warmer weather, and 'emergency' delivery was made to order both ice and BBQs from their supplier. Non-regular (emergency) orders cost more for delivery and the supplier charges a fee to the supermarket for the delivery. The expert is unaware of how much this service-charge is yet said that the managers do not question the need for stock if things are running out and put 'emergency' orders around twice a month. Therefore, the expert was asked about any strategies the supermarket had implemented to address these issues, such as stock forecasting. They explained that they are not involved in that aspect, as it is handled by a separate team. The expert believes that managers only receive the results rather than gaining detailed insights into the findings. However, they did mention that this has been a recurring issue for several years and has become more pronounced as the supermarket's popularity has grown, indicating that either no successful strategies have been implemented, or that any strategies that were initially effective have been outpaced by the increasing demand.

Moreover, markdowns are a key routine decision within the supermarket, dealing with perishable or seasonal inventory. These markdowns involve reducing the prices of items to stimulate sales and clear out stock that is approaching its sell-by date, going out of season, or not selling as expected, thus attempting to maintain cash flow. The expert also mentioned that a small number of customers will prefer to shop in the reduced section, and the majority of those customers are regular shoppers who primarily buy from the reduced section.

Another component of routine decisions is the monitoring of checkout processes in relation to queue lengths. The expert confessed they once seen this as trivial, but from an experienced point of view they can see that adding a retail assistant to checkouts or additional help at self-service means taking away retail assistants from other supermarket activities or their current duties. The supermarket always has two checkouts manned at all times, in addition to one member of staff operating the customer service desk. The expert noted that during the Christmas period, especially during the few days before the supermarket closes over the holidays, almost all checkouts are manned and operating throughout the day, which is why temporary staff during this season is especially useful.

1.1.2.2 Common Challenges

When asked about common challenges faced in daily operations, the expert referred back to some examples they had given earlier in the interview, namely 'novelty' stock shortage, issues at self-service and frustrations regarding the setup of promotional displays. They did not elaborate on these previously mentioned challenges, yet when prompted and given some time to think, they did explain some other challenges which would occur frequently in the daily running of the supermarket.

Firstly, the expert spoke about the challenges with communication between different sections or departments of the supermarket. The expert gave an example that the supermarket is currently holding a promotion on meal-deals. However, retail assistants that did not work at the customer service desk where these meal-deals are located were unaware of this promotion. So, when customers came to the checkout they were asking if certain items were included in the meal-deal, to which the staff were unable to help. The promotion began on the Friday, yet the other departments were unaware until the following Tuesday, only because management had received a complaint about the staff's lack of knowledge regarding the promotion and thus an email was sent to all retail assistants regarding all promotions for the upcoming week. The expert emphasised that such emails should now become a regular practice to ensure all departments are consistently informed and can provide accurate information to customers. When prompted about the effect this had on the supermarket, the expert said that this communication issue could have led to customer dissatisfaction and potentially lost sales, as customers might have been discouraged from purchasing items included in the meal deal due to the staff's uncertainty, yet no substantial impact has come from it.

It is worth noting that during this segment of the interview, the expert was asked about the perceived cause of the communication issues. The expert suggested that, in hindsight, communication itself is not the primary issue, nor is it the fault of any one person. Instead, the problem lies in the absence of a standardised system within the supermarket that lets all employees know, regardless of department or contract, what promotions and changes they need to be aware of, as well as all other changes within the supermarket, and a way to go back and look at the changes or promotions if staff forget.

Moreover, high employee turnover rate is a significant challenge for the challenge. Although turnover itself is not a daily occurrence, the repercussions are. The frequent addition of new staff disrupts the established routines and collaboration among existing employees. The long-term retail workers, in particular, may have an effective and efficient routine or system within their workday, which allows them to work to the best of their ability and manage their time effectively, but the regular influx of new staff can upset this flow. certain department which works for them, yet with new staff entering the workplace regularly it can upset this flow. When asked for an example of this 'flow' the expert was unable to think of one. The expert empathises with new hires, acknowledging the difficulty of adjusting, yet also understands the frustration experienced by long-term employees. Additionally, the expert highlighted that the continual need for training new hires consumes time and resources, leading to increased operational costs and diverting these resources away from other critical areas of the business. To address this challenge, the supermarket has started to pair new retail workers with experienced ones to provide additional support and guidance, helping the new hires integrate more smoothly and understand their role in

relation to current systems. The expert noted that during the initial period of this new “scheme”, new hires were being paired with both part-time and full-time staff. This approach, while well-intentioned, caused several perceived issues. The expert explained that part-time staff found it overwhelming as they are used to having fewer responsibilities, in addition to not knowing the supermarket or daily operations as well as full-time staff. This was quickly corrected, and even if a new hire was contracted to part-time, they were still paired with fulltime staff for their induction period.

Furthermore, the supermarket has a wide range of products, which retail assistants come in contact with every day. Each of these products have a specific requirement with regards to handling and storage, especially frozen food and fresh produce which require strict temperature controls to maintain freshness. Failure to adhere to these regulations may lead to waste and health hazards, detrimental to both the consumer and the supermarkets reputation. In recent years, the supermarket has been integrating “smart refrigeration and temperature control systems” which regulate the storage conditions for these items. The expert states that retail assistants in the frozen food department are required to ensure that all freezer doors are securely closed every 15 minutes, due to customers leaving them open or ajar, indicating the cause of this issue. However, with the advanced refrigeration systems in place, there is some flexibility as these systems are designed to compensate for rising temperatures by automatically adjusting the temperature if any doors are left open.

When asked about their views on the industry’s shift towards technology, particularly with the integration of the ‘app’ within their supermarket, the expert expressed confidence in their alignment with market leaders. They noted that technological upgrades are not a concern and in fact they are welcomed, as long as they prove beneficial to operations. The expert emphasised that both they and the ‘higher ups’ believe in thoroughly trailing and testing new technologies due to their associated costs. They are optimistic about technology that simplifies operations, but they emphasise that any new system must be both efficient and clearly demonstrate its effectiveness before being fully implemented, especially when it involves significant financial investment with the expectation of increased profits in return.

1.1.2.3 Customer Involvement

Previously in the interview, the expert had already discussed the importance and involvement of customers in daily operations, namely how crucial it is for the supermarket to maintain exceptional customer service, for the employees to present themselves in a way that is helpful and welcoming. The expert states how highly important customer satisfaction is for the supermarket, but alludes to their shopping experience more than anything, namely that the customer feels like the retail assistants are doing everything they can to help the customer, even if the outcome is not desirable.

When asked a question regarding the influence customers have on daily activities, the expert discussed the use of feedback when adding new products to the supermarkets inventory. Essentially, if unit sales are increasing, the supermarket may decide to increase not only the delivery size of those products, but also find some alternatives or similar products to gain customer interest. Additionally, if retail assistants are being questioned regularly about a particular product which the supermarket doesn’t stock, then

management encourages the retail assistants to submit a “product suggestion form”, which will be sent to the supermarket’s headquarters, and will be reviewed to see if this is location specific or if all supermarkets could benefit from this product. The expert states that even if it is location specific, it is usually granted, depending on regularity of questions and demand. An example the expert gave was the recent popularity in the “Asevi cleaning range”, which was approved and delivered to their supermarket 3 days after a product suggestion form was sent. When asked what information the form contained such as frequency of questions about the cleaning products, the expert did not know the contents, only that one was sent and followed up promptly.

The expert also noted that customers often provide subtle feedback to employees regarding signage, for example category labels within the aisles or price tags on promotional displays. For example, this month, many customers have been asking where the ice was located. This indicated that, despite the presence of freezers, the ice had been moved to a portable freezer near the BBQ section without clear signage. As a result, the ice was returned to its usual place in the freezers with improved signage to prevent further confusion.

1.1.2.4 Notable Problems & Successful Strategies

When asked about the successful strategies and notable problems the expert had observed recently which they had not mentioned before, they were unable to think about some straight away. However, when prompted by recapping the previous challenges they faced and addressed, they were able to add some additional ones in those areas.

They began by explaining the issue of implementing new technologies into the supermarket, with regards to both employees and customers. Introducing a new system or updating an existing one often proves challenging for employees considering the variety of systems and devices the employees interact with daily, based on the expectation that employees must be proficient across all departments within their role. When asked how the supermarket has provided a solution to this issue, the expert stated that the training provided is sufficient for understanding the technologies. Additionally, management is always willing to assist those who need further support. When asked to give an example of this happening recently, the expert explained that around 6 weeks ago the supermarket updated its inventory management system to allow employees to directly add new products, rather than management having to add them when deliveries containing new products arrived. Many struggled to adapt due to the diverse age range and varying technological proficiency among staff. With this new update, the supermarket ran training sessions for all staff during their shift, with hands-on demonstrations of how to use the updated system. By monitoring the implementation via employee feedback, the supermarket maintained operational efficiency. The expert was then asked how this impacted the supermarket’s performance. They noted that while there were initial challenges, the overall impact has been positive. However, during the early stages, there was a significant loss of time and a decrease in operational efficiency related to inventory management. This was due to some employees continuing to use the old manual methods while others adopted the new system, which resulted in inconsistencies in inventory levels and even led to a stockout of some popular canned foods, such as soups. An emergency delivery was required, which reduced profits.

However, the expert is confident that, in the long run, the new system will outweigh this one-off issue by significantly improving operational efficiency.

Moreover, in the past two and a half years, there has been a recurring issue with managing waste in the colleague inventory system. When, for example, a food jar breaks or is otherwise rendered unusable, retail assistants are supposed to clean it up, scan the barcode or label, and record the quantity of the broken item. This information should then be entered into the inventory management software under a 'waste entry.'

The process includes specifying the reason for the waste, such as out of temperature range, damaged, expired or spilt multipack. After entering the details, the staff member submits the entry. Properly accounting for waste is crucial because it affects stock levels. If waste is not recorded correctly, it can result in significant discrepancies in stock counts. Moreover, the supermarket can claim money back for damaged items if they are recorded and salvaged correctly. For example, if a TV is broken during transit and immediately marked as waste, the supermarket can claim a refund from the supplier. Additionally, the inventory system is not perpetual, meaning it does not automatically update stock levels in real-time. If wasted items are not accurately recorded, the system may incorrectly show that there is plenty of stock when, in reality, shelves are empty. This leads to frequent complaints about empty shelves, despite the system indicating sufficient inventory. The expert said that staff often fail to recognise the full extent of inventory issues or do not fully understand the impact on the supermarket, especially when considering the long-term effects of profitability and revenue loss being effected so heavily, which could be avoided if they took the correct approach. When asked about a solution, the expert mentioned that while the frequency and content of training on the systems have increased, there has been no noticeable improvement. Currently, the expert cannot think of a better way to address this issue.

Despite further prompts, regarding topics such as sustainability and managerial duties, the expert was unable to think of any additional problems they had faced that had not already been mentioned. However, they did share a successful strategy they had recently observed. The expert explained that, over the past eight months, their supermarket has begun to engage with the local community. They began sponsoring local football teams, placed a trolley at the end of the checkout lanes where customers can donate non-perishable items to local food banks, and last Christmas they ran a toy drive where, for every toy a customer donated to the "toy bin", the supermarket contributed two additional toys. These efforts made it to the local newspaper. The expert stated that having a good reputation is a priority for those in the "headquarters". When asked why they believed this, they explained that the supermarket now has a 'community' wall where cut-outs of media and awards are places, located at the entrance of the supermarket. Additionally, they feel that engaging with the community is a strategic move, particularly because other large supermarkets are also adopting similar practices.

1.1.3 Seasonal Variations and Special Events

With regards to seasonal variations and special events, the expert had previously spoken about the summer event and the increase of foot traffic and customer interactions. This was

mentioned within the section of “Daily Operations”, above. However, at this stage of the interview, I asked the expert to focus specifically on the events, rather than the associated promotions. The expert began by implying the overall store layout remains consistent throughout the year by saying “aisle 12 will always be pasta, aisle 3 will be raw meat”, however the variation of products within these sections will change. An example the expert provided was that during the summer, the supermarket decreases its stock of bread-related items like pancakes and increases its stock of baps and buns. This seasonal adjustment also applies to other parts of the supermarket. For instance, general merchandise from the previous season, such as Easter products, is replaced with BBQs and garden sets. The supermarket also capitalises on the back-to-school season by stocking stationery items only from the end of July to the beginning of October. During the Christmas season there is a noticeable increase in the stock of specialty foods, gift sets, decorations, and seasonal drinks. Moreover, alcohol sales significantly increase during the Christmas period. Special promotions featuring premium brands are displayed to attract customers. The expert states that Christmas is the only time the supermarket prominently promotes luxury over budget options. Additionally, during the New Year, there is a significant shift towards promoting health and wellness products, taking advantage of New Year’s resolutions by stocking vitamins, home gym equipment, and fresh fruits and vegetables. Due to the supermarket’s diverse range of products, the expert notes that there is considerable flexibility to experiment with new non-food items. They receive small quantities of new products each season to test their popularity, and if these items sell well, they continue to restock them. This approach makes maintaining accurate inventory levels crucial during these times.

The expert was then prompted to discuss some changes in the actual running of the supermarket during this time. They suggested several key adjustments, particularly in staffing and operations. Firstly, during the Christmas season, the supermarket increases its staffing levels to handle the higher number of customers. This predominantly consists of hiring temporary employees, yet they still have to go through all the training required to be a permanent retail assistant, so the hiring process commences late September every year. There is a heightened focus on inventory management to ensure that popular items are always in stock during this period, in addition to an increase in the frequency of deliveries. The expert originally stated that the number of deliveries at Christmas ranged from 2-5 per day, yet when questioned about a specific time this happened, they said that 5 was only on one occasion, and has never exceeded 3 except on the one occasion of 5 where it was 5 small deliveries due to a shortage of large trailers stemming from bad weather.

Finally, the expert explained that the supermarket extends its operating hours during the Christmas season. This allows more time for shopping and reduces congestion during peak hours, in addition to more flexibility for restocking shelves. They also operate a one-way system around the supermarket on the days leading up to Christmas to prevent congestion at the checkouts and for marketing purposes to ensure customers are exposed to all products within the supermarket.

Section 1.2: Historic Changes

1.2.1 Expert Knowledge

To begin, the expert was asked, "What changes have occurred within your supermarket over the years?". The encouragement of both positive and negative changes was made apparent to the expert.

Firstly, the expert talked about the main campaign within the supermarket, which offers lower prices to selected items. They explain that the campaign was used to give a 'name' to the sales which the supermarket was promoting, introducing a new way of engaging with customers regarding both promotions and lower prices. The products for the campaign are selected based on consumer demand and competitive analysis, ensuring that popular items are the main focus. The items which the discount is applied to change every 3 weeks. The aim of the campaign is to increase sales on these items, as well as gain new customers. This strategy reflects a broader industry trend where retailers use targeted promotions to compete against discount supermarkets like Aldi and Lidl, which have reshaped consumer expectations by consistently offering low prices, pushing traditional supermarkets to innovate in how they communicate value to customers as they managed to reshape consumer behaviour in terms of cost-consciousness.

The expert was then asked, "Can you identify any subtle trends or rare events that have affected your supermarket over the years?". They began by explaining the repercussion Brexit has had on the supermarket, particularly since 2021. They explained that their supply chain for fresh fruit and vegetables had been significantly disrupted. The border processes had gotten more difficult, leading to delivery delays and one of their main suppliers no longer delivering to Non-EU countries. The delayed deliveries also means that perishable items are being unfit for sale once arriving to the supermarket. When asked about ongoing issues, the expert noted that, just last week, a large shipment of strawberries were delayed by 4 days due to new documentation requirements stemming from Brexit. The majority of the strawberries had to be wasted, and they were not in stock within the supermarket for a further 2 days until another delivery arrived.

In the wider industry, following Brexit, supermarkets in the UK faced considerable disruptions. An article by Kantar (2022) highlighted that almost 80% of the UK's food imports originate from the EU, resulting in increased costs and delays for stores such as supermarkets. Additionally, the increased costs for deliveries will eventually imply increased costs for consumers, with many customers becoming more aware of the implications of Brexit, particularly for those who are cost-conscious, especially with the cost-of-living crisis. These changes reflect broader industry trends where retailers across the UK are struggling with the post-Brexit reality, leading to a re-evaluation of supply chain strategies and a push for more localised sourcing options.

Furthermore, the expert states that product returns have increased steadily but noticeably over the past few years. The supermarket's return policy has stayed the same yet returns continue to increase. They observed that returns have not been specific to one specific section of the supermarket and is unable to explain why this may be occurring.

This rise in returns could be attributed to several factors, such as changes in consumer behaviour post-pandemic, where customers are more likely to buy in bulk and return excess items, or heightened quality expectations driven by increased competition from online

retailers. Additionally, the ease of return policies might inadvertently encourage more frequent returns. According to the National Retail Federation (2023), returns in the grocery sector have increased by 15% over the past five years. They state that this trend is attributed to the increase in online shopping, where discrepancies between what consumers expect and the actual products often result in higher return rates. This is something to keep in mind when considering the experts limited knowledge surrounding this change.

Finally, the expert was asked to discuss the ‘competitive’ aspect between their supermarket and other industry leaders. They state that low-pricing has remained the main competitive factor between the ‘big 5’, yet states their supermarket is more focused on customer experience and the advancement of both customer-facing and employee-facing technologies. Yet, the expert did state that their supermarket still attempts to keep prices low.

1.2.2 Further Research

With the expert's insights now established, independent research through literature reviews was carried out to gain a deeper understanding of the historic changes within the supermarket industry. It is worth noting that I asked AI (ChatGPT) to verify the causes mentioned, hence the screen shots following the points.

With regards to the expert stating that their supermarket chooses customer experience over cost-orientated strategies, research was conducted to see the opposite of this approach. Rice (2021) explained that the “discount supermarket chain” Aldi has had a giant impact on the way supermarkets are perceived. Aldi has adopted low-prices as their main objective. They maintain a streamlined selection of products to reduce their inventory costs and maintain higher quality control. Their goal is to offer high-quality products at lower prices than current market leaders. Since discount retailers like Aldi have adopted this model, shopping habits have changed within Britain with customers now focused on finding the best discounts and lowest prices to save money. There is now an ongoing price war between discount retailers and other supermarkets, whereby traditional retailers are constantly adjusting their pricing and values to compete with the discount stores, and to match the current consumers shopping habits.

When conducting literature reviews, articles surrounding sustainability were prevalent, which was a factor the expert had not touched on, although prompted. To understand this better, it's useful to examine the historical context of sustainability and evaluate how supermarkets are adapting to sustainable practices, as well as the benefits of these adaptations. The Retail Bulletin (2024) discusses the growing emphasis on sustainability in grocery retail, driven by environmental concerns and consumer expectations. While the industry faces challenges, the shift toward sustainable practices offers significant benefits. These include enhancing brand image, increasing consumer loyalty, and achieving long-term cost savings through efficiency improvements. Additionally, embracing sustainability can align them with regulatory trends and provide access to new markets and product offerings.

Further investigation into sustainability practices sheds light on Marks & Spencer's approach [1], particularly through their "Plan A" initiative. This program shows the company's commitment to becoming a leader in sustainable retailing. Marks & Spencer aims to significantly reduce their carbon footprint by implementing strategies to lower greenhouse gas emissions throughout their supply chain. Their efforts extend to ethical sourcing practices, ensuring that the materials they use are obtained in a way that respects both environmental and human rights standards. Additionally, the company is focused on minimising waste through various waste reduction strategies and promoting recycling and circular economy principles. Marks & Spencer presents itself as an industry leader in sustainability, a position that enhances their brand reputation and strengthens their public image.

Regarding the use of loyalty cards, The Good Growth (2023) article explores the rising use of supermarket loyalty cards, focusing on how they help retailers gather detailed customer data. This data collection allows supermarkets to enhance customer retention by offering personalised marketing and promotions. Loyalty programs enable stores to gain insights into individual shopping habits, tailoring their offerings to meet specific customer needs and preferences. This targeted approach not only boosts customer satisfaction but also improves inventory management and promotional strategies. However, there is a risk of "data blindness." This occurs when retailers become overly dependent on quantitative data, potentially missing out on broader market trends and qualitative insights that haven't been captured, leading to strategies that might overlook important aspects of customer behaviour.

The article by Pinochet (2022) explores how grocery stores and supermarket retailers have used technology to enhance various aspects of their business. For instance, the use of automated systems has improved operational efficiency and customer experience, similar to the benefits observed within 1.1.1.1, regarding self-service machines. Notably, the study highlights the increased adoption of technology in response to challenges presented by the COVID-19 pandemic. One example is the implementation of 'traffic light systems' in stores, which helped maintain social distancing and ensured customer safety, thereby streamlining the shopping experience and increasing efficiency. The authors observed that the implementation of smart technologies in supermarkets was positively linked to an enhanced quality of life for people during the pandemic, and suggested this dramatic increase of technology in supermarkets would continue. In conjunction with the thoughts and assumptions from this article, some further research around increased self-service technology was conducted. Findings showed that some leading supermarkets, such as Tesco [2], have implemented 'Scan & Go', which is a system used by customers to directly scan an item as they place them in their carts, eliminating the need for traditional checkout lines, with ASDA [3] now incorporating this within their mobile app, thus removing the need for a physical bar-code scanner.

This strategy aligns with broader industry shifts where retailers are investing in technology not only to reduce costs but also to differentiate themselves in a highly competitive market, where factors like convenience and service quality are becoming as important as price.

Research surrounding the downfall of supermarkets within the retail industry uncovered a significant issue when considering adaptation to technology. Kivlehan (2019) showed the impact that the transition to online shopping had on Woolworths. The store was known for its brick-and-mortar appeal. However, in the mid-2000's this became less appealing to customers who sought convenience. The market conditions were changing rapidly, with the main emphasis being on the integration of online shopping. Woolworths failed to implement this new consumer preference, and eventually went into administration in 2008. Additionally, Woolworths also failed to provide higher discounts, with competitors such as Poundland gaining increased market share during this period. This combination of competitive prices and a changing retail environment was the reason Woolworths is no longer a retail store. This situation highlights a general change in consumer behaviour during economic downturns, where shoppers become more focused on finding the best prices. Retailers that can offer good deals during these times are more likely to succeed.

1.2.3 Artificial Intelligence

It's worth noting that I used some AI prompts with ChatGPT to explore questions in this section before diving into further research. This approach helped me identify potential 'dead-ends' and also recognise areas where additional investigation was needed due to possible AI biases. Nevertheless, it guided me in the correct direction. Please note that the 'AI Prompt' refers to the prompt used to generate the results, and the image that follows the prompt displays the outcomes produced by ChatGPT.

AI Prompt: "Can you provide an analysis of recent trends in the supermarket industry, focusing on changes in customer behaviour, technological advancements, and competition? How do these trends compare to historical data?"



Certainly! Here's a summary of recent trends in the supermarket industry:

Customer Behavior:

- **Online Shopping Growth:** Increased preference for ordering groceries online, with delivery and curbside pickup options gaining popularity.
- **Health Focus:** Higher demand for organic, plant-based, and health-conscious products.
- **Sustainability:** Greater consumer interest in eco-friendly packaging and sustainable practices.
- **Personalization:** Expectation for personalized shopping experiences, including tailored promotions and recommendations.

Technological Advancements:

- **AI and Automation:** Adoption of AI for inventory management, personalized marketing, and automated checkout systems.
- **Omnichannel Integration:** Seamless shopping experience across online and physical stores, with synchronized inventories.
- **Blockchain:** Use of blockchain for improved supply chain transparency and product traceability.

Competition:

- **E-Commerce Giants:** Increased competition from online retailers like Amazon with advanced logistics and tech efficiencies.
- **Discount Retailers:** Growth of discount chains like Aldi and Lidl driving price competition.
- **Niche and Local Stores:** Rising competition from specialty and local stores offering unique products and experiences.

Historical Comparison:

- **Customer Behavior:** Shift from traditional in-store shopping to a mix of online and personalized experiences.
- **Technology:** Move from manual processes to advanced technologies like AI and blockchain.
- **Competition:** Transition from regional competitors to a global and digital competitive landscape.

AI Prompt: "How are supermarkets adapting to sustainability trends? What are some successful strategies implemented by industry leaders, and what challenges are they facing?"



Certainly! Here's a summary of how supermarkets are adapting to sustainability trends:

Adaptations:

- **Sustainable Sourcing:** Focus on ethical, local, and organic products.
- **Eco-Friendly Packaging:** Reduction of single-use plastics and introduction of reusable packaging options.
- **Energy Efficiency:** Adoption of green building standards and renewable energy sources.
- **Waste Reduction:** Programs for food waste management, recycling, and composting.
- **Consumer Engagement:** Educational campaigns and incentives for sustainable behaviors.

Successful Strategies:

- **Walmart:** Sustainable sourcing goals, waste reduction initiatives.
- **Tesco:** Plastic reduction, commitment to carbon neutrality.
- **Whole Foods Market:** Emphasis on local and organic products, comprehensive waste reduction programs.

AI Prompt: "What factors are contributing to the rise in customer returns in the supermarket industry? Are there any recent studies or data that highlight the reasons behind this trend?"



The rise in customer returns in the supermarket industry can be attributed to several key factors:

1. **Online Shopping Growth:** Issues with online orders, like incorrect items or quality concerns, lead to more returns.
2. **Higher Consumer Expectations:** Customers expect better product quality and service, prompting returns when expectations are not met.
3. **Product Quality Problems:** Issues with freshness, expiration, or packaging defects increase returns.
4. **Flexible Return Policies:** More lenient return policies encourage higher return rates.
5. **Pricing and Promotions:** Fluctuations in pricing and promotions can lead to returns if customers find better deals elsewhere.
6. **Ease of Returns:** Simplified return processes make it easier for customers to return items.
7. **Consumer Behavior Trends:** Trends like "try before you buy" influence return behaviors.

Recent studies and reports from sources like McKinsey & Company, the National Retail Federation (NRF), and Deloitte highlight these factors and their impact on return rates in the supermarket industry.



AI Prompt: "How has the adoption of new technologies influenced the supermarket industry? Specifically, how have advancements like automation and digital payment systems changed the shopping experience and store operations?"

Advancements in technologies like automation and digital payment systems have transformed the supermarket industry in several key ways:

1. **Store Operations:**
 - **Automation:** Enhances inventory management, streamlines checkout with self-service kiosks, and improves supply chain efficiency through robotics and automated systems.
 - **Data Analytics:** Provides insights into customer behavior and preferences, aiding in personalized marketing, inventory management, and demand forecasting.
2. **Shopping Experience:**
 - **Digital Payments:** Speeds up transactions with mobile wallets and contactless payments, and enhances security with advanced encryption.
 - **Personalization:** Uses data to offer tailored promotions and conveniences like smart carts and digital shelves.
 - **Online Integration:** Facilitates click-and-collect services and improved delivery options, offering greater flexibility and convenience for shoppers.

Overall, these technologies have made store operations more efficient and improved the overall shopping experience.

1.2.4 Further Artificial Intelligence

Now, I will show my use of Ai to verify the causes of historical changes mentioned above.

When asking about the cause of supermarket supply disruptions, ChatGPT did state that Brexit has had a significant effect on supply chains and issues due to the movement restrictions, which has been discussed in section 1.2.2, as shown in the conversation below:

The increase in supply disruptions for supermarkets in recent years can be attributed to a combination of global events, structural changes in supply chains, and evolving market dynamics. These factors have made supermarket supply chains more vulnerable to disruptions, impacting the availability of products on shelves. Here's a breakdown of the key reasons:

1. **Global Pandemic (COVID-19)**
 - **Impact on Global Supply Chains:** The COVID-19 pandemic severely disrupted global supply chains. Lockdowns, factory closures, and restrictions on movement affected the production, distribution, and transportation of goods worldwide. Supermarkets faced difficulties in sourcing products, leading to shortages and delays.
 - **Panic Buying and Stockpiling:** The pandemic also led to periods of panic buying and stockpiling, which created sudden spikes in demand that supermarkets were unprepared for. This exacerbated supply chain pressures and led to empty shelves for certain essential items.

Next, I asked ChatGPT about the cause of rising discount retailers, to which its reply corresponds to the understanding and research conducted earlier in this section, as shown below:

The rise of discount retailers in the supermarket industry over the past 20 years can be attributed to several factors, including changing consumer behavior, economic conditions, and evolving business models. Here's a closer look at the key reasons:

- 1. Economic Pressures**
 - **Global Financial Crisis (2007-2008):** The economic downturn that began in 2007 had a profound impact on consumer spending. With reduced disposable incomes, many consumers became more price-sensitive, leading to a shift towards discount retailers, which offered lower prices for basic goods. This trend continued even as economies recovered, as consumers became more accustomed to the value offered by discount stores.
 - **Austerity Measures and Wage Stagnation:** In many countries, particularly in Europe, austerity measures and wage stagnation have continued to squeeze household budgets. Discount retailers have benefited from this environment by providing cost-effective alternatives to traditional supermarkets.
- 2. Changing Consumer Behavior**
 - **Increased Price Sensitivity:** Consumers have become more price-conscious and are actively seeking ways to stretch their budgets. Discount retailers have capitalized on this by offering low-cost, no-frills shopping experiences that appeal to budget-conscious shoppers.
 - **Focus on Value:** Shoppers are increasingly prioritizing value over brand loyalty. Discount retailers often offer high-quality private-label products at lower prices than branded goods found in traditional supermarkets, attracting customers looking for the best value for their money.
 - **Perception of Quality:** Over time, discount retailers have improved the quality of their products, particularly their private-label offerings, which has helped change consumer perceptions. The stigma once associated with shopping at discount stores has diminished, making these retailers more appealing to a broader audience.
- 3. Efficient Business Models**
 - **Lean Operations:** Discount retailers typically operate with leaner business models than traditional supermarkets. They often have smaller store footprints, limited product ranges, and fewer staff. These efficiencies allow them to pass savings on to consumers.
 - **Private Labels:** Many discount retailers focus heavily on private-label products, which are often cheaper to produce and offer higher profit margins than branded goods. This focus on private labels allows them to keep prices low while maintaining profitability.

Then, I asked to detail the causes for increased product returns in supermarkets. Again, ChatGPT's reply aligns with the insights I had previously uncovered, as shown below:

Product returns in supermarkets have increased in recent years due to a combination of factors, although the extent of this increase can vary depending on the region and specific supermarket. Here's a look at the reasons behind the rise in product returns:

- 1. Growth of Online Grocery Shopping**
 - **E-Commerce Expansion:** The rise of online grocery shopping, especially accelerated by the COVID-19 pandemic, has contributed to an increase in product returns. Customers ordering online may receive items that differ from what they expected, either in quality, size, or freshness, leading to higher return rates.
 - **Substitutions and Errors:** When certain products are unavailable, supermarkets often substitute them with similar items. However, these substitutions may not meet the customer's preferences, leading to returns. Additionally, errors in picking and packing orders can result in customers receiving incorrect items, prompting returns.
- 2. Increased Consumer Expectations**
 - **Higher Standards:** Modern consumers have increasingly high expectations for the quality and accuracy of the products they purchase. If a product fails to meet these expectations—whether it's due to freshness, taste, or appearance—customers are more likely to return it.
 - **Greater Awareness of Return Policies:** Many supermarkets have more lenient return policies today, encouraging customers to return products they are not satisfied with. Increased awareness of these policies has made consumers more comfortable with the idea of returning products.
- 3. Product Quality and Safety Concerns**
 - **Quality Issues:** In some cases, product quality issues, such as spoilage or damage during transportation, have led to an increase in returns. This can be particularly prevalent in fresh produce, meat, and dairy products.
 - **Food Safety:** Concerns over food safety, including contamination scares or issues with packaging, can lead to an uptick in product returns. If consumers perceive a product as potentially unsafe, they are likely to return it for a refund or exchange.

I then asked a follow up question concerning consumer behaviour in relation to bulk buying, in addition to less strict return policies, to delve deeper into the causes:



Yes, both bulk buying during the post-pandemic period and easy return policies can contribute to an increase in product returns in supermarkets. Let's explore how these factors play a role:

1. Bulk Buying and Post-Pandemic Behavior

- **Overbuying During the Pandemic:** During the height of the COVID-19 pandemic, many consumers engaged in bulk buying and stockpiling, particularly of essential items like groceries, cleaning supplies, and toiletries. This behavior was driven by uncertainty and fear of shortages. As a result, many households ended up with more products than they needed.
- **Excess Inventory Post-Pandemic:** After the initial panic subsided and supply chains began to stabilize, some consumers found themselves with excess inventory of certain products. This led to a situation where they might decide to return unopened or unused items to supermarkets, especially if those products had long shelf lives or if they were purchased in quantities far exceeding their actual needs.
- **Shift in Consumer Behavior:** As the pandemic's immediate impact waned, consumer buying patterns shifted. Some people who initially bulk bought items later found that they either no longer needed those items in such large quantities or preferred different products. This shift can result in an increase in returns as consumers seek to recoup money spent on unnecessary or unwanted goods.

2. Easy Return Policies and Their Impact

- **Encouragement of Returns:** Supermarkets with easy and lenient return policies are more likely to experience higher return rates. When customers know that returning products is hassle-free and without significant penalties, they may feel more comfortable purchasing items in bulk or trying new products, knowing they can easily return them if they change their minds or if the products do not meet their expectations.
- **Lack of Deterrence:** Lenient return policies can reduce the deterrent effect that typically prevents unnecessary purchases. For example, a consumer might buy a large quantity of an item because they perceive little risk—if they don't use it, they can just return it. This can lead to an increase in both intentional over-purchasing and subsequent returns.
- **Impact on Product Testing:** Easy return policies also encourage consumers to try new products or brands they might otherwise avoid. If the product doesn't meet their expectations, they are more likely to return it, contributing to overall return rates.

Next, I questioned the causes surrounding sustainability. The answer aligned with my understanding of customer demand and supermarket image:



The increase in sustainability practices in supermarkets over the past years can be attributed to a combination of consumer demand, regulatory pressures, competitive advantage, and corporate responsibility. Here's a closer look at the key drivers behind this trend:

1. Consumer Demand for Sustainable Products

- **Increased Environmental Awareness:** Consumers are becoming more aware of environmental issues, such as climate change, plastic pollution, and deforestation. This awareness has translated into demand for more sustainable products and practices. Shoppers increasingly prefer products that are ethically sourced, have minimal environmental impact, and come with reduced or recyclable packaging.
- **Health and Wellness Trends:** Alongside environmental concerns, many consumers are focused on personal health and wellness, which often aligns with sustainability. Organic, locally-sourced, and non-GMO products, which are perceived as healthier, are also viewed as more sustainable, driving supermarkets to increase their offerings in these areas.

2. Regulatory and Legislative Pressure

- **Government Regulations:** Governments around the world are introducing stricter regulations aimed at reducing carbon emissions, improving waste management, and promoting sustainable sourcing. For example, bans on single-use plastics, requirements for recycling, and emissions targets are pushing supermarkets to adopt more sustainable practices to remain compliant.
- **Sustainability Reporting Requirements:** In many countries, companies, including supermarkets, are required to report on their sustainability practices. This transparency has encouraged supermarkets to adopt and publicize their sustainable initiatives to meet regulatory requirements and improve their public image.

3. Corporate Social Responsibility (CSR)

- **Reputation and Brand Image:** Sustainability is now a key component of corporate social responsibility (CSR). Supermarkets are increasingly aware that their brand image and reputation are closely tied to their environmental and social impact. By adopting sustainable practices, they can enhance their reputation, attract eco-conscious consumers, and build long-term brand loyalty.
- **Ethical Leadership:** Many supermarket chains are driven by a genuine commitment to ethical practices and the desire to play a positive role in society. This includes reducing their carbon footprint, supporting fair trade, and ensuring that their supply chains are free from exploitation and environmental degradation.

Then, I had a longer-winded conversation surrounding loyalty cards and just rewards in general. The conclusion of this was directed back to loyalty cards, aligning with the reasons I had spoken about surrounding a want for customer details and retention:

Loyalty cards have increased in prevalence in supermarkets over the past years primarily due to their value in gathering detailed customer data and fostering customer retention. These programs offer numerous benefits for both supermarkets and consumers, which has driven their widespread adoption. Here's a closer look at the reasons behind the increase in loyalty cards and their role:

1. Data Collection and Customer Insights

- **Detailed Customer Data:** Loyalty cards are a powerful tool for collecting detailed data about customers' shopping habits, preferences, and behaviors. Every time a customer uses their loyalty card, supermarkets can track what they purchase, how often they shop, and how much they spend. This data is invaluable for understanding customer preferences and trends.
- **Personalized Marketing:** The data collected through loyalty programs allows supermarkets to create personalized marketing campaigns. By analyzing purchase history, supermarkets can tailor promotions, discounts, and product recommendations to individual customers, increasing the likelihood of repeat purchases and higher spending.

2. Customer Retention and Loyalty

- **Increased Customer Loyalty:** Loyalty cards encourage repeat business by offering rewards, discounts, or points that can be redeemed for future purchases. Customers who perceive value in the rewards are more likely to continue shopping at the same supermarket rather than switching to competitors.
- **Building Long-Term Relationships:** Loyalty programs help supermarkets build long-term relationships with customers. By offering consistent rewards and personalized experiences, supermarkets can foster a sense of loyalty and attachment, making customers feel valued and appreciated.

Finally, I asked about the causes that are shifting supermarkets towards technology integration. This again was a long-winded conversation and was difficult to portray the meaning of 'cause' within advancement of technology. Nevertheless, the answer does relate back to operational efficiency, general industry trends and consumer wants :

1. Rapid Technological Advancements

- **Moore's Law:** The continuous advancement in computing power, as predicted by Moore's Law, has driven the development of faster, smaller, and more affordable technologies. This has enabled widespread adoption across various sectors.
- **Innovation in Software and Hardware:** Advances in software development, cloud computing, artificial intelligence (AI), machine learning (ML), and the Internet of Things (IoT) have revolutionized how businesses operate and how people interact with technology. Hardware innovations, including more powerful processors, better sensors, and improved connectivity, have further accelerated the integration of technology into everyday life.

2. Digital Transformation of Industries

- **Business Efficiency and Automation:** Companies across industries have increasingly adopted technology to improve efficiency, reduce costs, and automate routine tasks. This digital transformation has led to the integration of advanced technologies like AI, robotics, and big data analytics into business operations.
- **E-Commerce and Online Services:** The rise of e-commerce and online services has dramatically changed consumer behavior, leading to the widespread adoption of technology in retail, banking, and other service sectors. The convenience of online shopping, digital banking, and remote services has made technology indispensable.

3. Global Connectivity and Communication

- **Internet and Mobile Penetration:** The expansion of the internet and mobile networks has made technology accessible to a vast global audience. The proliferation of smartphones and mobile devices has enabled constant connectivity, allowing people to access information, communicate, and conduct transactions from virtually anywhere.
- **Social Media and Digital Platforms:** Social media platforms and other digital communication tools have become central to personal and professional interactions. The ability to connect, share, and collaborate online has made technology an integral part of daily life.

4. Consumer Expectations and Behavior

- **Demand for Convenience and Personalization:** Modern consumers expect convenience, speed, and personalization in their interactions with businesses and services. Technology enables companies to meet these expectations through personalized marketing, instant access to services, and seamless user experiences.

Section 1.3: Challenging Unquestioned Assumptions

Before diving into this section, it's important to acknowledge that challenging unquestioned assumptions is difficult, especially in an industry as established and complex as supermarkets and retail. Market leaders often emerge by maximising something already in use, rather than something extremely radical. However, given the maturity and constant evolution of this industry, where businesses are continuously analysing and adapting, uncovering new assumptions and opportunities for innovation can be challenging.

In my exploration, I wanted to identify expert, general and anecdotal assumptions by reading between the lines of expert interviews and industry analysis. The goal was to uncover assumptions that might have been overlooked and think creatively about how to address them. To assist with this process, I asked ChatGPT to help identify general assumptions within the industry, as these could serve as a foundation for deeper analysis.

However, as my research progressed, I found that many of the general assumptions identified by ChatGPT had already been optimised for by industry players. This realisation pushed me to get more creative with my assumptions. In this section, I aim to present a blend of expert insights, general assumptions and anecdotal observations that could offer fresh perspectives and drive innovative thinking in the supermarket and retail sectors, rather than simply optimising already existing assumptions, like the wider industry is already doing.

To begin, the expert was asked "What are some things that everyone in the industry assumes is the case, in terms of what is conventional within your business or industry". These were the leading questions to uncovering some unquestioned assumptions. It is worth noting that the expert only provided insight into the assumptions, as when they were asked to elaborate on any radical ideas for change or different ways of working, they were unable to think of any.

Assumption by Expert: The need for security within their supermarkets, and supermarkets alike. They explained that security is needed, both in physical form for shop lifters and disturbances, as well as cameras for theft and disorderly conduct, in addition to the things they had previously spoken about, such as anti-theft hanging arms. This assumption demonstrates a conventional belief that visible security infrastructure is essential for loss prevention.

Challenge to Assumption: Changing the engagement of the business's employees could achieve better overall security outcomes. The desired result is to potentially reduce the reliance on costly technology and equipment, making them less necessary or prevalent within the supermarket. This approach aims to lower security-related expenses, contributing to overall cost savings for the supermarket.

Alternative: Supermarkets already have a plethora of staff, some of which have minimal tasks to attend to. There is an opportunity to engage these employees more actively in security roles, rather than relying heavily on surveillance cameras which only record incidents after they occur or with a delay. The benefit of having increased employee

engagement is that they can view and intercede in 'real-time'. While training staff in security awareness is beneficial, it is not always effective if employees are preoccupied with other duties.

Instead, supermarkets could implement more automated systems, such as fully digital and automated inventory management. This would not only help with store operations but also allow employees to focus more on monitoring for security issues. Additionally, improving store layouts to potentially eliminate blind spots from certain areas of the store where employees can be more engaged with their surroundings (such as shelf-stackers as opposed to cashiers), could prevent theft. For example, instead of locating anti-theft devices at the back of the store, positioning them where they can be easily monitored might deter theft more effectively. The expert previously mentioned that staff often "listen" for the sound of the anti-theft hanging arm clicking, indicating a potential theft. However, an alternative approach could involve positioning these anti-theft devices in more visible locations within the store, rather than at the back. This would allow for constant visual monitoring by staff, reducing the need for employees to rush towards the area when a device is triggered. This change would make it more efficient for staff to address potential thefts and enhancing the overall customer experience by avoiding the impression of constant surveillance or accusations.

Data: To evaluate the cost-effectiveness of a new approach to security, we could compare the expenses of high-tech security systems with those of customer service training and digital improvements. We can also look at the additional benefits associated with a technology-focused store, in terms of operation efficiency, which is a sub-factor that would hopefully improve the entire supermarket.

Considering 'TransLink', they use a blend of staff presence (within their trains) and technology to ensure safety and security while maintaining a welcoming environment. Their presence serves as a deterrent to misconduct and provides immediate assistance to passengers. TransLink also employ surveillance cameras on trains and in stations to monitor activity. In addition, automated ticketing systems help at platforms and reduce the need for on-site staff, allowing them to be more engaged with security and safety.

Just as TransLink combines staff presence with technology, supermarkets could implement a hybrid approach to security. This would involve increasing employee engagement on the shop floor, where staff actively monitor customer behaviour and respond to potential security issues in real-time. By doing so, supermarkets can reduce their reliance on surveillance cameras and anti-theft devices.

As for blind spots, libraries enhance security by using open floor plans and strategic layout designs. This setup improves visibility, allowing staff to monitor the entire area by avoiding hidden areas and maintaining clear sightlines. Within the supermarket, strategically positioning security devices can help eliminate blind spots and ensure continuous staff visibility, for example, having aisles more spread out or less of them, or in the case of some stores, having two floors where one overlooks the other, so there is a birds-eye view of the shop floor. However, this is more of a 'development' aspect to future supermarkets.

A combination of technology and layout adjustments can create a more secure store environment, aligning with the hybrid approach seen in other industries while enhancing the customer experience.

Furthermore, without giving away any information that could lead to identification, a recent review of the expert's supermarket website reveals that significant security measures (such as one-way systems) were rolled out in 80% of stores in 2023. However, the necessity for implementing these measures across all locations gives further concerns. Observing and analysing data related to the area in which the supermarkets are situated that study crime rates and demographic profiles of different areas may provide a better insight regarding the extremity and necessity of these security measures. For example, stores situated in neighbourhoods with higher crime rates or specific demographics, like a large elderly population, may require varying levels of security. Additionally, examining theft data from other businesses in the same area can help gauge actual security needs more accurately, thus being able to tailor security approaches to the specific risks and characteristics of each location, rather than applying a one-size-fits-all solution.

Assumption by Expert: The expert stated, in terms of their own supermarket, that excellent customer service is a main driving force for increased customer loyalty, subject to things such as a welcoming, personal, and approachable manner when engaging with customers. This value was also demonstrated throughout a lot of the expert's interview within 1.1.1.

Challenge to Assumption: Although excellent customer service is important, customer loyalty is more than likely influenced by a combination of factors, and customer service may not be the primary driver but rather just one of many contributing elements. In the broader context, customer service could even be relatively insignificant in comparison to other factors. The expert highlighted that their company takes great pride in providing superior service, making it a core focus. However, it's worth considering whether factors like lower prices might be more crucial, especially given the intense competition among retailers striving to offer the lowest prices. Market leaders and discount retailers emphasise price competitiveness, which might be more influential in driving customer loyalty. Despite this, the expert's business appears to place less emphasis on pricing strategies, as noted in the section 1.2.1 of the report, which suggests that while competitive pricing is acknowledged, it is not a primary focus for them.

Data: Firstly, we could analyse data from the loyalty programs to see if changes in service, pricing, or product offerings correlate with increases or decreases in customer retention and satisfaction. This can be achieved by focusing on specific feedback questions within the loyalty app, such as "Did you interact with any of our staff today?". A straightforward 'yes or no' answer to this question can provide clearer insights into customer interactions with staff, compared to a more complex array of questions that may not be easily comparable or could be rather long-winded. Additionally, incorporating a small incentive, such as adding 1p or 10p to a cash pot for each response, can encourage greater participation in the survey, with more customers actually engaging in the questionnaires.

Thinking of this in terms of competitors, Aldi's model (Sherman, 2024) demonstrates how efficiency and competitive pricing can effectively attract and retain customers, even with minimal service levels. By focusing on a limited product selection and minimal operations, Aldi keeps their costs low and brands themselves as a discount retailer. This approach appeals to budget-conscious shoppers who prioritise savings over extensive service, and this EDLP (everyday low prices) model could be adopted within the expert's supermarket.

Additionally, Amazon shows how combining great customer service with competitive pricing and convenience builds customer loyalty. Amazon uses data to offer personalised pricing and recommendations, and it focuses on convenience with fast shipping and easy returns. This strategy demonstrates that while good service is important, it's even more effective when paired with fair pricing and convenience. Amazon even has a "Tell us about a lower price" section on their website, so they know about competitors and try to get you the best price possible (Amazon, 2024). Amazon's success proves that a mix of service quality, pricing, and convenience can strongly drive customer loyalty and satisfaction, directly relating to the hybrid approach in which the expert's supermarket could opt for.

Moreover, Amazon Go stores utilise advanced technology, vision sensors, and movement sensors, to automatically track items taken by customers, eliminating the need for traditional security measures. Is there a gap in the UK market for this kind of technology? Or even advancing in this direction of technology implementation? Embracing similar technologies could transform the security approach in the expert's supermarket. Implementing AI and sensors to monitor store activity and prevent theft can reduce dependence on physical security measures. This shift not only aligns with modern technological trends but also creates a more seamless shopping experience for customers.

Anecdotal Assumption: Markdowns are merely a routine practice rather than a strategic opportunity.

Challenge to Assumption: During the interview, the expert highlighted that markdowns are a routine and essential decision-making process within the supermarket, especially when managing perishable or seasonal inventory (1.1.2.1). These markdowns involve reducing prices to stimulate sales and clear stock that is nearing its sell-by date, going out of season, or not selling as expected. The expert noted that while a small subset of customers specifically seek out markdown food, most of these shoppers are regulars who primarily buy from the 'reduced' section. This indicates a consistent demand for discounted products among a specific customer segment.

The primary challenge lies in the existing stigma around purchasing reduced-price items, often perceived as undesirable, especially concerning food. This stigma can deter customers from engaging with these items (Edwards, 2020). Current supermarket practices generally treat these items as afterthoughts, marked down and placed in less prominent areas. Having conducted further investigation, research suggests no major UK supermarkets have fully integrated a model that prioritises food reductions as a core marketing strategy, indicating a potential gap in the market.

Data: Too Good To Go (2024) is an app that lets customers buy leftover food from restaurants and stores at a lower price. It markets these meals as "rescued," highlighting how buying them helps reduce food waste. This approach not only saves customers money but also promotes a sense of environmental responsibility.

Moreover, Misfits Market (2024) sells near-expiry produce, marketing these items as budget-friendly and environmentally conscious. They emphasise that these "imperfect" foods are still high-quality, appealing to consumers interested in sustainability. By purchasing from these services, customers can help reduce food waste while saving money, making it an attractive option for those who value both cost savings and environmental impact.

By framing the purchase of near-expiry food items as both an environmental and community-friendly choice, supermarkets can encourage more sustainable shopping behaviours and contribute to a positive local impact. Highlight the positives, not the stigma around it. The marketing approach could change, as although there is a niche customer population buying these items regularly, by making it more desirable or removing aspects of the stigma via marketing and prompting it in ways that makes people feel 'comfortable' (their image, ethical standpoint and food quality) about buying the items, there could be a greater influx of customers, more profit, and a reduction of waste, in addition to improving brand image surrounding sustainability (which is another matter which the expert could not address, although there is a growing concern).

Artificial Intelligence: Given that two-thirds of the assumptions came from experts, I decided to explore what AI could contribute to this topic. The following images are insights from ChatGPT and Perplexity AI.

What are some assumptions in the supermarket industry which everyone believes to be the case?

In the supermarket industry, several assumptions are widely held, though they may not always reflect the full complexity of the market. Here are some common assumptions:

- Price Sensitivity:** Consumers are assumed to be highly sensitive to prices and will switch brands or stores for better deals or discounts.
- Convenience is Key:** There's a belief that convenience is a major driver for customer choice, so stores prioritize location, store layout, and efficient checkout processes.
- Brand Loyalty:** It's often assumed that customers will be loyal to specific brands, which is why supermarkets stock popular brands and use them for promotional strategies.
- Product Freshness is Paramount:** There is a strong belief that freshness, particularly in perishables like fruits, vegetables, and meats, is crucial for customer satisfaction and sales.
- Private Labels are Inferior:** A common assumption is that private label (store brand) products are seen as inferior compared to national brands, even though this isn't always true for every customer.
- Sales and Promotions Drive Traffic:** It's assumed that sales, discounts, and special promotions are the primary drivers for attracting customers to the store.
- Customer Loyalty Programs Work:** Many supermarkets believe that loyalty programs effectively increase customer retention and spending.
- Technology Will Solve Problems:** There's an assumption that adopting new technologies, such as self-checkout systems or sophisticated inventory management tools, will solve operational inefficiencies and enhance the shopping experience.

Sources

Data-Driven Pricing Strategies for Better... linkedin • 1 From Insights to Impact: How Data-Driven Pricin... linkedin • 2 How to define a data-driven pricing strategy ... reactev • 3 View 2 more

Data-Driven Pricing Strategies for Better Bottom Line - LinkedIn
How can data-driven pricing strategies improve your bottom line? Pricing is one of the most important decisions for any business, as it directly affects revenue...

It price is the only factor influencing consumer decisions, ride valuable insights by considering other factors such as per service. Here's how a data-driven pricing strategy can pricing strategies and understand the impact of various

Data Collection and Analysis

- Customer Behavior and Preferences:** Collect data on customer behavior, preferences, and purchase history. This includes analyzing factors such as customer demographics, purchase frequency, and feedback on product quality and service. Data sources can include sales transactions, customer surveys, web analytics, and social media interactions ① ② ③.
- Market and Competitor Analysis:** Monitor market conditions and competitor pricing strategies. This involves gathering real-time data on competitor prices, promotional activities, and market trends using tools like SEMrush, MOZ, and SpyFu. Understanding competitor moves helps in setting competitive prices and identifying gaps in the market ③ ④.
- Supply and Demand Dynamics:** Analyze supply and demand patterns to understand price sensitivity and optimal pricing points. This includes tracking inventory levels, production costs, and seasonal demand variations. Historical data can help detect patterns and predict future trends ① ③ ④.

Implementing Pricing Strategies

- Dynamic Pricing:** Utilize dynamic pricing algorithms to adjust prices based on real-time data. This approach allows businesses to respond quickly to changes in demand, competitor pricing, and market conditions. Dynamic pricing ensures that prices are always aligned with the current market scenario, maximizing profitability

Product Uniqueness

1. 44% of consumers say they're more likely to become repeat buyers after a personalized shopping experience (Segment)
2. 91% of consumers are more likely to shop with brands that provide relevant offers and recommendations (Accenture)

These statistics highlight that unique, tailored product offerings can be a significant draw for customers, often outweighing pure convenience.

Store Atmosphere

1. 78% of millennials would rather spend money on a desirable experience than on desirable things (Harris Group)
2. 74% of Americans prioritize experiences over products (Eventbrite)

This data suggests that creating an engaging in-store atmosphere can be a powerful tool for attracting and retaining customers, especially among younger demographics.

Community Involvement

1. 70% of consumers believe it's important for brands to take a public stand on social and political issues (Sprout Social)
2. 64% of consumers choose, switch, avoid or boycott a brand based on its stand on societal issues (Edelman)

These figures indicate that a store's community involvement and social stance can significantly influence customer loyalty and attraction.

Customer Feedback and Trend Analysis

1. Companies that excel at customer experience grow revenues 4-8% above their market (Bain & Company)
2. 89% of companies see customer experience as a key factor in driving customer loyalty and retention (Gartner)

This data underscores the importance of continuously gathering and acting on customer feedback to improve the overall shopping experience.

Multi-faceted Approach

1. 86% of buyers will pay more for a better customer experience (PWC)
2. Loyal customers are 5x more likely to repurchase, 5x more likely to forgive, 4x more likely to refer, and 7x more likely to try a new offering (Temkin Group)

These statistics highlight that a multi-faceted approach addressing various aspects of the customer experience can lead to significant benefits in terms of customer loyalty and business growth.

Sources

4 Sales Promotion Customer Retention... 	How to Increase Customer Loyalty: Best... 	How Retail Promotions Drive Sales and... 
 salespromotions	 loyaltylion	 newstore

[View 2 more](#)



4 Sales Promotion Customer Retention Tactics (Loyalty vs ...)

Table of Contents This post may contain affiliate links. As an Amazon Associate we earn from qualifying purchases.* Sales promotions' goal is to bring about a desire...

...ve in driving short-term sales, you're right to challenge the best way to create lasting customer loyalty. Data-driven alternative strategies focused on quality, service, and effectiveness in attracting and retaining loyal customers:

1. Quality and service trump discounts: Research shows that 52% of American consumers will join loyalty programs for brands they frequently purchase from, but more importantly, 79% say these programs make them more likely to continue doing business with the brand. This indicates that customers value ongoing relationships and benefits beyond just discounts.
2. Personalization drives loyalty: Gartner's 2022 Consumer Priorities Survey emphasizes that fostering loyalty among customers is crucial for encouraging repeat business and establishing a base of long-term buyers. This suggests that personalized approaches, rather than generic promotions, can be more effective.
3. Unique experiences create emotional connections: In-store events and personalized services can create stronger emotional bonds with customers. For example, 56% of consumers are more likely to switch brands due to loyalty or reward scheme promotions, compared to only 48% for pure price discounting. This indicates that value-added experiences are more compelling than simple price cuts.
4. Long-term brand perception: Continual discounting can negatively impact brand perception. If customers become accustomed to discounted prices from the outset, it can be challenging to maintain loyalty without ongoing discounts, potentially damaging the brand's perceived value.
5. Data collection opportunities: Alternative strategies like loyalty programs or personalized services provide valuable customer data. This information can be used to tailor future offerings and create more targeted, effective marketing strategies.

Here's a brief discussion on the three assumptions I selected from ChatGPT & Perplexity AI to deepen my understanding and examine some of the widely accepted beliefs in the industry.

Price Sensitivity

Assumption: Customers always pick the cheapest option and will switch stores for better deals.

Challenging the Assumption: This assumption might ignore other important factors like quality, convenience, and customer service. Some shoppers might prefer to pay more for a better experience or higher quality. Testing different pricing strategies, like adding value or offering loyalty rewards, can help determine if price is the only factor or if other benefits matter too.

Convenience is Key

Assumption: The main reason customers choose a store is its location and layout.

Challenging the Assumption: While convenience is important, it might not be the only factor. Customers might also care about unique products, store atmosphere, or community involvement. Gathering customer feedback and analysing trends can help identify additional ways to attract and keep shoppers. This is in correlation with the previous assumption of the correlation between customer service and customer satisfaction. Again, this has a multifaceted approach.

Sales and Promotions Drive Traffic

Assumption: Discounts and special offers are the biggest draw for customers.

Challenging the Assumption: Relying heavily on promotions might not create lasting customer loyalty or set a store apart. Customers might value consistent quality, great service, or unique products more. Trying other strategies, like in-store events or personalised services, can offer new ways to attract and retain shoppers.

Homing in on success factors, a prompt used in ChatGPT was "What are some unconventional success stories in the supermarket industry that have deviated from traditional business models or strategies? Please provide examples of companies that have achieved remarkable success through innovative approaches, unique business practices, or technological advancements."



Certainly! Here's a concise summary of four unconventional success stories in the supermarket industry:

1. Trader Joe's

Trader Joe's stands out with its:

- **Private Label Products:** Focuses on its own brands, which control quality and pricing.
- **Unique Store Atmosphere:** Features nautical-themed decor and friendly staff.
- **Curated Selection:** Offers a select range of unique and exotic items, creating a distinctive shopping experience.

2. Aldi

Aldi has transformed the discount supermarket model through:

- **Efficient Operations:** Uses a no-frills store design, minimal staffing, and a streamlined supply chain.
- **Private Labels:** Dominates with its own brands to keep costs low and quality high.
- **Cost Savings:** Passes on savings to customers with lower prices, achieved through bulk purchasing and a limited product range.

3. Ocado

Ocado has innovated with its:

- **Online-Only Model:** Operates exclusively online, bypassing physical stores.
- **Advanced Technology:** Utilizes robotics and automation in warehouses for efficient order fulfillment.
- **Global Partnerships:** Partners with other retailers worldwide to offer its technology and logistics solutions.

4. Thrive Market

Thrive Market differentiates itself with:

- **Membership Model:** Offers discounts on natural and organic products through a subscription service.
- **Social Mission:** Uses membership fees to provide free memberships to low-income families, focusing on accessibility and sustainability.



Looking at the retail industry as a whole, due to potential future issues surrounding existing data in supermarkets, ChatGPT was asked "Provide examples of companies in the retail and supermarket industries that have achieved significant success through unconventional or innovative business models. Describe their unique strategies, technological advancements, or business practices that set them apart from traditional approaches. Include a brief summary of how these strategies contributed to their success."

 In the retail industry, several companies have achieved remarkable success through unconventional approaches that deviate from traditional business models. Here are a few standout examples:

1. Warby Parker

Warby Parker revolutionized the eyewear market with:

- **Direct-to-Consumer Model:** Sells glasses online and in branded stores, bypassing traditional retail channels and reducing costs.
- **Home Try-On Program:** Offers customers the ability to try on frames at home before purchasing, enhancing convenience and reducing return rates.
- **Social Mission:** For every pair of glasses sold, Warby Parker donates a pair to those in need, aligning the business with a social cause.

2. Dollar Shave Club

Dollar Shave Club disrupted the shaving industry with:

- **Subscription Model:** Provides a subscription service for razors and grooming products, delivering them directly to customers' doors on a regular basis.
- **Digital Marketing:** Used viral marketing and humorous content to build brand recognition and customer loyalty.
- **Cost Efficiency:** Focuses on cost-effective products and streamlined distribution to offer lower prices compared to traditional retail.

3. Glossier

Glossier has redefined beauty retail through:

- **Direct-to-Consumer Approach:** Sells products primarily online, leveraging social media for marketing and customer engagement.
- **Community-Driven Development:** Builds its product line based on customer feedback and real-life beauty routines, creating a strong sense of community and brand loyalty.
- **Minimalist Branding:** Emphasizes a minimalist aesthetic and user-friendly product designs, differentiating itself from traditional beauty brands.

4. REI

REI has set itself apart in the outdoor retail sector with:

- **Co-op Model:** Operates as a consumer co-op where members pay a one-time fee for a share and receive annual dividends, fostering customer loyalty and a sense of ownership.
- **#OptOutside Campaign:** Closed stores on Black Friday to encourage outdoor activities and promote a lifestyle centered around nature and sustainability.
- **Focus on Experience:** Provides expert advice and experiences, such as outdoor classes and gear rentals, enhancing the customer experience beyond just selling products.

In summary, the AI insights in this section provides a solid foundation on the widespread, unquestioned assumptions within the supermarket industry. This understanding will aid in further investigation of market gaps and opportunities for analysis and capitalisation.

Section 1.4: Conclusions & Further Work

1. Conclusion

Firstly, the supermarkets current system of collecting customer feedback through forms, online reviews, and incentive driven prize draws highlights a commitment to understanding and improving customer experiences. However, analysing the effectiveness of this feedback system is crucial for ensuring it translates into actionable improvements. Throughout the interview, the expert highlights the supermarket's thoughts surrounding the correlation between customer satisfaction and returning customers, ultimately leading to increased profits. Examining how the prize draw incentives impact the quality of feedback would help us to understand if incentivised responses are more or less reliable compared to non-incentivised ones. Furthermore, there is an option to study the relationship between customer satisfaction scores and key performance indicators like return rates and unit sales, to see how strong this correlation is, which could include regression analysis to determine which aspects of customer satisfaction most strongly impact KPIs. A thorough investigation into the root causes of customer satisfaction can provide deeper insights into areas needing improvement, beyond the expert's emphasis on creating a 'welcoming atmosphere.' By analysing how other businesses within the retail and grocery sectors approach customer satisfaction and their performance (such as the practices discussed in section 1.2 regarding discount retailers) we can uncover more concrete evidence and factors that genuinely influence customer loyalty and repeat business at a supermarket.

Moreover, the supermarket's recent anti-theft initiatives, including new security devices and protocols, warrant thorough evaluation to gauge their effectiveness in reducing theft and improving overall security. It would be useful (as described in section 1.3) to conduct a cost-benefit analysis of the new security measures to determine their financial viability and overall impact on store security, such as assessing the reduction in theft since the implementation of the new security measures and compare the current theft rates to those before the new measures were introduced. Or we could quantify the financial savings from reduced theft. For instance, if theft was costing the supermarket £50,000 annually and the new measures have reduced this by 30%, calculate the savings (£15,000). Then, having these figures, compare them to other stores or supermarkets in the area or around the UK, for example. This will give us an insight into the benefits of these security measures, rather than simply perceiving them to be reducing theft and losing profit and may also reveal if the cost of the measures is actually detrimental to supermarket profits considering the usefulness of the security, whilst also considering the Payback period regarding how long it will take for the financial benefits to cover the initial costs of the new security measures. As a side note, consider any changes in the supermarket's brand image and reputation resulting from the enhanced security measures. Improved security could boost customer loyalty and attract new shoppers. Alternatively, heavy surveillance might negatively impact customer perceptions, making them feel uncomfortable or overly scrutinised. We could track customer reviews on platforms like Google Reviews, Yelp, and social media channels, as well as customer feedback forms. Look for mentions of security and any changes in tone or sentiment related to the new measures. In addition, track the average time customers spend in the supermarket. A significant change could indicate how security measures are

affecting the shopping experience. Also, we could compare the supermarket's brand image and reputation metrics with industry benchmarks to identify any shifts caused by the security measures.

Finally, effective management of inventory waste is crucial for accurate stock levels and financial accuracy. Current issues with waste management and inventory tracking highlight the need for improved processes. Optimising inventory waste management can lead to more accurate stock levels, reduce financial discrepancies, and improve customer satisfaction by ensuring better stock availability. There was a suggestion (in section 1.3) about food waste and the incentives surrounding the improvement of stigma, as well as items perceived as 'broken' could actually be sold for a profit, considering the supermarket does not gain any return of funds if products are broken within the supermarket, rather than during transit. We could start by calculating the financial impact of current waste practices, including the cost of lost inventory and any associated operational inefficiencies. In addition, we could categorise the types of waste and their reasons (e.g., broken packaging, expired items, etc.). Identify which types could potentially be salvaged or repurposed. Also, research potential markets for selling these 'damaged' or 'expiring' items, such as subscription services for damaged goods. This might include discount stores, specialised outlets, or donation programs, as mentioned in section 1.3. Finally, we could calculate the costs involved in repurposing or reselling wasted items, including any modifications needed to make them sell, and estimate the potential revenue from selling previously wasted items whilst comparing this with the current financial impact of waste. This would also incorporate the inclusion of profit margins, ensuring that the selling price covers costs and generates a profit, even for the majority of items. Considering if it is cheaper for the supermarket to sell the items or dispose of them correctly also comes into play here, which includes measuring the time and resources required for waste management.

2. Further Work

Firstly, it may be better to interview a retail assistant on each shift, i.e. the morning shift and evening shift, as shown in section 1.1.1.1. Asking targeted questions such as, "Are there any stock issues during the changeover?" or "What is the most common or recurring challenge or annoyance you face when starting your shift? Is there anything that colleagues from the previous shift consistently leave for you to handle?" the supermarket can gather valuable data. Interviewing retail assistants across different shifts can provide a comprehensive view of the challenges and opportunities unique to each time period. This approach not only highlights operational inefficiencies but can also uncover variations in customer demographics and shopping behaviours, which are crucial for tailored marketing strategies.

Next, In addition to interviewing retail assistants from each shift, interviewing a section leader would also be beneficial. While the expert mentioned certain job aspects, they did not delve deeply into the pressures faced by section leaders, who often manage significant stress. Insights from a section leader could reveal challenges in coordinating shifts, managing stock levels, and maintaining smooth operations. This interview could cover difficulties in keeping consistent standards across shifts, overseeing employee performance, and dealing with customer complaints. It would also be useful to understand what the

section leader finds most stressful, such as meeting sales targets or handling staff shortages. This perspective can highlight areas where support and training are needed, in addition to more understanding about the way the supermarket operates at a slightly more managerial level, and what improvements to operations could be made.

It could be valuable to delve deeper into the KPIs and metrics used by management, which may not be apparent to lower-level staff. While daily operational metrics like unit sales and customer satisfaction are important, management likely monitors broader indicators. Interviewing a manager of the supermarket, that also directly works with employees, could reveal how these KPIs influence key aspects of store operations such as inventory management, staffing, and floor space allocation. For example, understanding revenue per square meter, a measure of how much income is generated relative to the area occupied by products, may be crucial in optimising floor space. By analysing this metric, management can identify which product categories generate the most revenue relative to their space. This insight could lead to reallocating more floor space to high-turnover, high-margin items, thereby increasing overall profitability.

Additionally, it would be good to engage with customers who regularly shop in the reduced-price sections of the supermarket. This could involve gathering their thoughts on any stigma associated with purchasing discounted items and exploring their interest in potential offerings such as subscription services for these products. Understanding their experiences and preferences could reveal valuable opportunities for customer engagement and loyalty-building strategies. For instance, customers might appreciate a subscription service that provides discounted fresh produce nearing its expiration date, directly from the supermarket, which could not only cater to budget-conscious shoppers but also reduce food waste. Furthermore, inquiring whether these customers frequent other supermarkets for similar deals can provide insights into competitive shopping behaviours, and certain times when markdowns occur could be capitalised on. Besides conducting traditional interviews, using sentiment analysis on social media and customer reviews can uncover hidden concerns and preferences. This approach helps gain a deeper understanding of how customers feel about value and quality, especially regarding discounted items. By addressing the needs and perceptions of this idea, the supermarket can enhance its values surrounding sustainability and better compete or establish a niche in the marketplace. Furthermore, analysing the business models of companies like Too Good to Go (2024), which specialise in selling discounted food items nearing expiration or with cosmetic imperfections, could provide valuable insights. These companies have built successful businesses around the concept of reducing food waste by offering these products at reduced prices, and thus by studying their strategies we could uncover potential practices for adoption within the supermarket.

APPENDIX

- Amazon (2024) Amazon Customer. Available at:
<https://www.amazon.co.uk/gp/help/customer/display.html?nodeId=GW9VEM42882GUCHK>
- OpenAI (2024) ChatGPT [Large language model]. <https://chat.openai.com/chat>
- Edwards, F. (2020) 'Overcoming the social stigma of consuming food waste by dining at the Open Table,' *Agriculture and Human Values*, 38(2), pp. 397–409. Available at:
<https://doi.org/10.1007/s10460-020-10176-9>
- Fresen, N. (2024) 'Sustainability in grocery retail: Challenges and opportunities,' *Retail Bulletin | Daily UK Retail News*. Available at:
<https://www.theretailbulletin.com/grocery/sustainability-in-grocery-retail-challenges-and-opportunities-21-06-2024/>
- Ggwpadmin (2023) 'The Data-Driven rise of supermarket loyalty cards and the risk of data blindness,' *Good Growth*. Available at: <https://goodgrowth.co.uk/articles/the-data-driven-rise-of-supermarket-loyalty-cards-and-the-risk-of-data-blindness/>
- Kantar (2022) 'How is Brexit affecting UK supermarkets and shoppers?' Kantar. Available at:
<https://www.kantar.com/uki/inspiration/retail/the-impact-of-brexit-on-uk-grocery-industry-and-shoppers/>
- Kivlehan, N.P. (2019) '10 years after Woolworths' collapse, what can retailers learn?' *Retail Gazette*, 4 March. Available at: <https://www.retailgazette.co.uk/blog/2018/12/woolworths-collapse-10-years-ago-first-domino-physical-retail-armageddon/>
- Misfits Market (2024) 'Save up to 30% off your groceries,' Misfits Market. Available at:
<https://www.misfitsmarket.com/?redirect=true&promo=240610X2>
- National Retail Federation (NRF) and Appriss Retail (2023) '2023 Consumer returns in the retail industry,' National Retail Federation (NRF) Report. Available at:
<https://cdn.nrf.com/sites/default/files/2024-01/2023%20Consumer%20Returns%20in%20the%20Retail%20Industry.pdf>
- Perplexity (2024) Perplexity AI [Artificial Intelligence Search Engine].
<https://www.perplexity.ai/>
- Pinochet, L.H.C., et al. (2022) 'Smart technologies in supermarket retail and their influence on citizens' quality of life during the COVID-19 pandemic,' *REGE Revista De Gestão*, 31(1), pp. 80–100. Available at: <https://doi.org/10.1108/rege-09-2021-0178>
- Rice, X. (2021) 'The Aldi effect: how one discount supermarket transformed the way Britain shops,' *The Guardian*, 17 December. Available at:

<https://www.theguardian.com/business/2019/mar/05/long-read-aldi-discount-supermarket-changed-britain-shopping>

Sherman, M. (2024) 'Aldi's Generic Competitive Strategy & Growth Strategies,' Panmore Institute. Available at: <https://panmore.com/aldi-generic-competitive-strategy-intensive-growth-strategies-case-study#:~:text=While%20Aldi%20faces%20competitive%20challenges,on%20a%20larger%20market%20share>

Sustainability (2024) 'Marks & Spencer,' Marks & Spencer. Available at: <https://corporate.marksandspencer.com/sustainability>

Too Good to Go (2024) 'Save good food from going to waste,' Too Good to Go. Available at: <https://www.toogoodtogo.com/>

[1] Asda (2024) 'Asda.com - Online Food Shopping, george, & more,' Asda. Available at: <https://www.asda.com/about/instore/scan-and-go>

[2] Security (2024) 'Tesco.com,' Tesco. Available at: <https://secure.tesco.com/clubcard/scan-as-you-shop>

Chapter 3

Section 3.1: Finding a useful dataset

3.1.1 Preparation

In preparation for this analysis, I utilised Miro to brainstorm and identify the type of data that would be most valuable for my dataset (figure 1). This approach ensured that the data I was searching for aligned with the core issues and values discussed in chapter 2 and incorporated relevant metrics and potential causes. The underlying purpose of this was to have a ‘quick reference’ guide, to glance at when searching for and delving into different data sets.

The data points identified during the brainstorming session included both numeric and qualitative aspects. I believe that the qualitative elements were crucial for providing context and supporting a thorough investigation. This open minded approach was intended to prevent any preconceived biases from influencing the data search and analysis or attempting to search for a data set related to only one issue and its associated metric. The goal was to consider a wide range of data sources, which would enable a better understanding of the supermarket's operations and potential areas for optimisation. By starting with a broad perspective, I aimed to capture a complete view of the relevant data.

Figure 1

Customer satisfaction Survey results ART metrics Customer inquiries Product return rates Reasons for returns Trolley disengagement??	Community engagement Records of local community events Participation records Sales data for branded merch via supermarkets or companies who sponsor sports teams (could link if needed)	Supermarket image NPS to gauge loyalty of customers SOV data from media that relate to supermarket or wider industry of retail? Customer review and social media mention metrics	Financial performance Sales data Unit sales Revenue by product category or product singly Gross profit margins ROA Inventory management data (stock + waste data) Employee turnover rates
Technology integration MTTR Downtime logs Tech integration error rates Technology incident reports Employee training records Feedback new technology adoption	Employee turnover Employee turnover rate Exit interview feedback Data on training costs	Supply issues OTDR Supplier performance metrics Stockout rates Inventory depletion records Supply chain issue records Chain disruption cause records	

In the context of the supermarket industry, one significant opportunity for innovation and better financial performance regarding profitability and revenue lies in the effective management and promotion of markdowns on near expiry food items, or products that are deemed as lacking in quality compared to the supermarket's standards. Chapter 1 proposed

and explored the unquestioned assumption of maximising the potential of markdowns, not only as a means to minimise waste but also to reduce the profit lost on selling items at a reduced price, in addition to any of these items remaining unsold, by proposing a strategic marketing scheme to reduce the stigma associated with these products. By encouraging customers to purchase markdown items, the supermarket can recover revenue that would otherwise be lost and improve their overall financial performance.

The ideal dataset for this analysis would include detailed information on markdowns for near expiry items, or products that were 'reduced to clear', or reduced for a reason related to quality. For example, columns related to revenue generated from markdowns, timing of markdowns (time of day), profitability of markdown, gross profit margins or another metric that may detail the impact of markdowns on overall growth of the supermarket. Again, this is only an example, but gives insight into what I was looking for.

This ideal dataset is valuable because it directly addresses the potential for increasing profit through effective markdown strategies, potentially resulting in positive financial outcomes for the supermarket, as mentioned in chapter 1. By understanding the patterns related to markdown purchases, the expert's supermarket can optimise their markdown processes, reduce stigma surrounding reduced items and ultimately drive higher sales from what is currently a less profitable section.

Additionally, regarding other challenges, discussed in chapter 1 and 2, improving the markdown strategy can significantly reduce losses, waste and contribute to better sustainability, thus indirectly affecting other values of the supermarket such as supermarket image and customer satisfaction.

Using the dataset, I would test assumptions about customer behaviour towards markdowns. For instance, I would investigate whether the timing of markdowns affects purchase rates and if certain product categories are more likely to be purchased when marked down. These analyses would provide insights into optimising markdown strategies and understanding customer preferences.

In addition, within the values section of chapter 2, I was able to subliminally relate all other values back to financial performance, although the other values themselves are standalone values. Therefore, a dataset that addresses the components and indicators of financial performance, such as profitability and revenue, along with specific metrics like gross profit margins, would be highly advantageous.

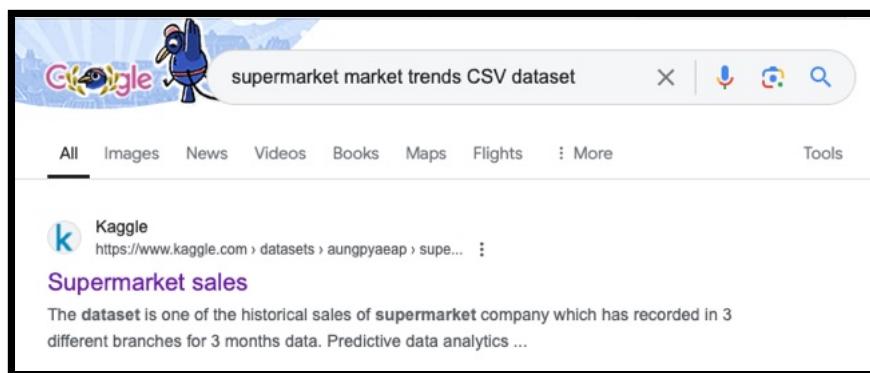
This data is valuable as it provides an overview of the supermarket's financial health, allowing for an assessment of how other supermarket aspects influence profitability, revenue, etc. By starting with financial data, I can gain a broader understanding of the industry landscape before diving into specific, innovative recommendations. Moreover, as I have previously addressed in the 'ideal' dataset section, the core reason for innovations surrounding markdowns is to attempt reduce financial losses, which is a direct element of financial performance.

Although a dataset which has a primary correspondence with financial performance, I believe it is a very good alternative to the ‘ideal’ dataset. In addition, due to financial performance data likely being publicly available and considering the time constraints of this project, a ‘financial performance’ focused dataset is, in foresight, more feasible to obtain compared to one focused on markdowns. This practical aspect supports my decision to prioritise financial performance data initially, ensuring a more efficient and productive research process.

As an additional note, it is worth mentioning that gathering and linking datasets is time consuming and challenging, especially for linking datasets, because data sources often lack consistent unique identifiers. Given the time constraints of this module, it is impractical to optimise my searching for this task. Therefore, I was mainly focused on identifying and utilising a single dataset, while keeping the possibility of linking data together as a secondary consideration.

3.1.2 Searching for datasets

To start my search, I initially used Google to explore potential datasets with the search condition of “supermarket market trends CSV dataset” to see where it would lead. This general search aimed to uncover publicly available datasets that might be useful, or websites which could be visited that contained datasets of similar aspects. Below is an example of the search, alongside the suggestion of using ‘Kaggle.com’.



Following this initial Google search, I specifically searched Kaggle, since the Google search recommended that website the most.

Within Kaggle, I simply used broad search terms, such as ‘supermarket industry dataset’, alongside the use of filters to initially exclusively look for datasets. To determine the contents of the dataset for consideration, I reviewed the ‘data card’ heading of the page, which allowed me to view a general overview of the contents, context and column headings. Once I had discovered a few possibilities, I examined the associated ‘notebooks’ to further gain insight into the datasets with corresponding visualisations and research, in addition to understanding what data analysts had used it for previously.

I was able to cover a wide range of datasets relating to my industry, however, they were rather ‘niche’ in the sense that they were predominantly datasets that had already been ‘cleaned’ or manipulated in some way to meet the needs of the analyst and their project scope, rather than having a less compromised dataset that could be analysed in multiple other ways. Additionally, the ‘original’ datasets were mostly behind paywalls, which is outside the means of my project.

It is worth noting that the searches below are comprised of switching between ‘Notebooks’ and ‘Datasets’ mainly, upon the results.

Kaggle Searches

Google Search “supermarket market trends CSV dataset” ->
<https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales> -> Home -> Kaggle
Search “Supermarket” -> <https://www.kaggle.com/discussions/accomplishments/516849> ->
Home -> Kaggle Search “Supermarket” ->
<https://www.kaggle.com/code/fareedalianwar/supermarket-sales> -> Home -> Kaggle Search
“supermarket industry dataset” -> <https://www.kaggle.com/code/ridamahmood005/indian-grocery-supermarket-big-basket-eda> -> Back ->
<https://www.kaggle.com/code/ragishehab/supermarket-sales-exploration> -> back ->
<https://www.kaggle.com/datasets/rupakroy/market-basket-optimization> -> Home -> Kaggle
Search “Retail Supermarket” -> <https://www.kaggle.com/datasets/roopacalistus/superstore>
-> back -> <https://www.kaggle.com/datasets/akashdeepkuila/big-mart-sales> -> Back ->
<https://www.kaggle.com/datasets/trainingdatapro/ocr-receipts-text-detection> -> Back ->
<https://www.kaggle.com/datasets/arunjangir245/retail-and-service-industries-dataset> ->
Home -> Kaggle Search “supermarket markdown” (no datasets to look into) -> Home ->
Kaggle Search “supermarket reduced” (no datasets to look into) -> Home -> Kaggle Search
“supermarket financial performance” -> <https://www.kaggle.com/code/teyang/drivers-of-hdb-resale-price-and-prediction> -> Back ->
<https://www.kaggle.com/datasets/prasad22/retail-transactions-dataset> -> Home -> Kaggle
Search “supermarket sales” -> <https://www.kaggle.com/datasets/alexhuitron/supermarket-sales> -> Back ->
<https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales> -> Back ->
<https://www.kaggle.com/datasets/armanmanteghi/supermarket-sales-insights-and-optimization> -> Back ->
<https://www.kaggle.com/datasets/muhammadehabmuhammad/supermarket-sales> -> Back ->
<https://www.kaggle.com/datasets/mohamedelaziz/supermarket-sales> -> Back ->
<https://www.kaggle.com/datasets/agasca/retail-sales>

I have added screenshots below to show these filters, search terms and navigations around Kaggle, and is only supposed to provide a snapshot of my efforts, in addition to demonstrating how other individuals could use this website.

← Supermarket industry dataset

Filter by

CREATOR

- You 0
- Others 48

DATASET SIZE

- small 45
- large 2
- medium 1

DATASET FILE TYPES

- xlsx 45
- jpg 2
- png 2

view more

DATASET LICENSE

- Other 46
- Commercial 2

48 Results

Relevance ▾

	Dataset	Downloads
	A Large Scale Fish Dataset Dataset - 3y ago · by Öğuzhan Ulucan from a supermarket in Izmir, Turkey for a university-industry collaboration project at Izmir University	983 23,979 downloads
	Meat Quality Assessment Dataset Dataset - 3y ago · by Öğuzhan Ulucan contains 2 classes, fresh and spoiled red meat samples collected from a supermarket in Izmir, Turkey	52 2,229 downloads
	Fish data Dataset - 4mo ago · by Sagnik Kayal-CSE-52 This dataset contains 9 different seafood types collected from a supermarket in Izmir, Turkey for a university-indust... 9 downloads	2
	Corporate Actions Market Data Kenya Techsalerator Dataset - 1y ago · by Techsalerator How do I pay for this dataset?	5 84 downloads
	Corporate Actions Data Bolivia Techsalerator Dataset - 1y ago · by Techsalerator How do I pay for this dataset?	1 20 downloads
	Corporate Actions Market Data Italy Techsalerator Dataset - 1y ago · by Techsalerator How do I pay for this dataset?	1 30 downloads

≡ kaggle

+ Create

- Home
- Competitions
- Datasets
- Models
- Code
- Discussions
- Learn
- More

Search

Retail Supermarket

Data Card Code (22) Discussion (1) Suggestions (0)

89 New Notebook Download

SampleSuperstore.csv (1.11 MB)

Detail Compact Column 10 of 13 columns ▾

About this file

The sales details of different stores of a supermarket chain that has multiple stores in different parts of the US.

# Ship Mode	# Segment	# P Country	# City	# State	# Postal
Shipping class	Customer type	USA	The city in USA where the store is located	Which State of the US	Postal code
Standard Class	Consumer	USA	New York City	California	20%
Second Class	Corporate	USA	Los Angeles	New York	11%
Other (2081)	Other (1783)	USA	Other (8332)	Other (6865)	69% 1040
Second Class	Consumer	United States	Henderson	Kentucky	42428
Second Class	Consumer	United States	Henderson	Kentucky	42428
Second Class	Corporate	United States	Los Angeles	California	90836
Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311
Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311
Standard Class	Consumer	United States	Los Angeles	California	90832
Standard Class	Consumer	United States	Los Angeles	California	90832
Standard Class	Consumer	United States	Los Angeles	California	90832

← supermarket markdown

[Notebooks 18](#) [Comments 2](#) [Topics 1](#)

← supermarket reduced

[Notebooks 57](#) [Topics 3](#)

← supermarket financial performance

[Datasets 106](#) [Notebooks 33](#)

Filter by

139 Results

Relevance ▾

DATE

Last 90 days

5



Topic 9. Part 2. Time series with Facebook Prophet
Notebook · 4y ago · by Yury Kashnitsky

194

6 comments

CREATOR

You

0



Supermarket Sales Exploration 📈💡
Notebook · 7mo ago · by Ragi Shehab

19

2 comments

OTHERS

Others

139



Drivers of HDB Resale Price and Prediction
Notebook · 4y ago · by TeYang Lau

32

0 comments

DATASET SIZE

small

105



TwoSigmaRenthop - Advanced Feature Engineering
Notebook · 3y ago · by ChrisCC

5

0 comments

DATASET FILE TYPES

xlsx

104



Predicting ENERGY STAR Score - Feb/2019
Notebook · 5y ago · by Daniela A.

6

2 comments

CSV

docx

1



Supermarket markdown
Notebook · 4y ago · by Aung Pyae

271,530 downloads

DOCX

supermarket sales

[Notebooks 1,102](#) [Comments 353](#) [Datasets 108](#) [Topics 62](#)

Filter by

108 Results

Relevance ▾

DATE

Last 90 days

8



Supermarket sales

Dataset · 5y ago · by Aung Pyae
The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches

2464

271,530 downloads

THIS WEEK

This week

1



Supermarket store branches sales analysis

Dataset · 2y ago · by SJ
StoreSales: Sales in (US \$) that stores made.

694

42,416 downloads

CREATOR

You

0



Supermarket Sales Data

Dataset · 1y ago · by YapWH
Sales Data of Vegetables in Supermarket

116

11,961 downloads

OTHERS

Others

108



Sales of a Supermarket

Dataset · 4mo ago · by Lovish Bansal
The dataset is one of the historical sales of supermarket company which has recorded in 3 different branches

94

7,606 downloads

DATASET SIZE

small

101



Supermarket Sales

Dataset · 10mo ago · by Alex Hutton
The original "Supermarket Sales" dataset exists out there and has been widely used to learn and practice

38

2,379 downloads

FILE TYPES

CSV

76



Supermarket Sales: Insights & Optimization

Dataset · 15d ago · by Arman Manteghi
In this project, we analyze supermarket sales data to derive actionable insights and develop a strategic

43

1,755 downloads

XLSX

xlsx

23



Supermarket Sales

Dataset · 8y ago · by Arturo Gasca
Sales tickets from Gasca Supermarket

36

3,673 downloads

PBIX

pbix

5



Supermarket Sales

Dataset · 8y ago · by Arturo Gasca
Sales tickets from Gasca Supermarket

36

3,673 downloads

Next, I explored Wikipedia. Although Wikipedia is not a primary source for datasets, it often links to valuable external resources. By using ‘Ctrl F’ I was able to search for ‘dataset’ on ‘supermarket’ yet proved unsuccessful. I have included an example below.

The other Wikipedia pages I searched for ‘Dataset’ were:

https://en.wikipedia.org/wiki/Grocery_store
<https://en.wikipedia.org/wiki/Asda>
<https://en.wikipedia.org/wiki/Tesco>
<https://en.wikipedia.org/wiki/Sainsbury>

Again, dataset was not found.



Additionally, I used Perplexity AI to search for datasets and information on how to attain certain datasets.

A screenshot of the Perplexity AI interface. The search query 'Supermarket dataset' is at the top. On the left, there is a sidebar with 'Sources' and a 'Sources' section containing links to Kaggle, Best Groceries Datasets for Market Research, and Data.world. Below this is an 'Answer' section with text about supermarket datasets being valuable for analyzing retail trends, consumer behavior, and sales patterns. It lists three main sources: 1. Kaggle offers a "Supermarket sales" dataset, 2. Another dataset on Kaggle titled "Supermarket dataset for predictive marketing 2023," and 3. Data.world hosts 5 different supermarket datasets contributed by various users and organizations. On the right, there is another 'Answer' section titled 'Answer' which provides insights on how markdowns affect profitability in supermarkets, listing datasets from Kaggle and Data.world, and an article from Airmart E-Commerce Article on Markdown Strategies.

Based on the results from Perplexity AI, I visited Data.gov, which is known for providing industry specific data, but mostly comes with a price. Again, no substantial result.

Data.gov Searches

Data.gov -> Search -> “supermarket sales data CSV” (no significant results, mainly electrical energy and utility) -> Search -> “supermarket” -> <https://catalog.data.gov/dataset/price-spreads-from-farm-to-consumer> -> Back -> <https://catalog.data.gov/dataset/supermarket-tax-incentives> -> Home -> Search -> “grocery store” -> <https://catalog.data.gov/dataset/food-price-outlook> -> Back -> <https://catalog.data.gov/dataset/grocery-store-locations> -> Back -> <https://catalog.data.gov/dataset/grocery-stores-fb200> -> Back -> <https://catalog.data.gov/dataset/grocery-stores-ee743> -> Back -> <https://catalog.data.gov/dataset/grocery-store-status>

The screenshot shows the Data.gov catalog interface. At the top, there's a search bar containing "supermarket sales data CSV". Below it, a sidebar on the left includes filters for "Filter by location" (with a map of the United States), "Topics" (Local Government), "Topic Categories" (none listed), "Dataset Type" (prospectus), and "Tags" (open-data, supermarket, open-government, district-of-columbia, electric-rates, housing). The main content area displays a list of datasets under the heading "101 datasets found for "supermarket sales data CSV"". The first dataset is "U.S. Electric Utility Companies and Rates: Look-up by Zipcode (2020)" with 205 recent views. The second is "U.S. Electric Utility Companies and Rates: Look-up by Zipcode (2022)" with 101 recent views. The third is "Chemicals in Cosmetics" with 52 recent views. The fourth is "United States Stateplane Zones - NAD83" with 36 recent views. Each dataset entry includes a preview, download links (CSV, XLS, PDF, etc.), and a "View Details" button.

I then reverted back to Google searches to try and find a dataset:

Google Search -> “supermarket dataset” -> data.world -> <https://data.world/codefordc/supermarket-tax-credit-zone> -> Back -> <https://data.world/dcopendata/1c5d4b467eaa4301b976547c65cd7d06-24> (the remaining 3 were unrelated) -> Back -> [data.world search](https://data.world/dcopendata/1d7c9d0e3aac49c1aa88d377a3bae430-4) -> “grocery store” -> <https://data.world/dcopendata/1d7c9d0e3aac49c1aa88d377a3bae430-4> -> Back -> <https://data.world/usda/grocery-stores> -> <https://data.world/d3detroit/detroit-grocery-stores> -> back -> Google Search -> “Grocery store dataset” -> <https://paperswithcode.com/dataset/grocery-store>

The screenshot shows the data.world search results page. At the top, there's a header with the data.world logo, a search bar containing "Search data.world", and buttons for "Sign in" and "Join". The main content area features a large message: "There are 5 supermarket datasets available on data.world." Below this, a sub-message reads: "Find open data about supermarket contributed by thousands of users and organizations across the world." The overall layout is clean and modern.

Many datasets I found contained very few columns or data points, making them insufficient for in depth analysis. Attempts to visualise the data using Python often revealed a lack of comprehensive information. Furthermore, many datasets were behind paywalls, rendering them unavailable for this project. While further research might yield better results, including direct requests to companies for data, these barriers limited my immediate options.

Even among the freely available datasets, they lacked key metrics of value, such as gross profit margins and detailed financial performance indicators. The absence of these metrics meant that the datasets could not support the detailed analysis needed. Additionally, some datasets were overly comprehensive, with extensive product information aggregated into single columns, making it challenging to extract novel insights.

Realising these limitations, I turned to Companies House, anticipating that it might offer more structured and detailed data. I examined various datasets available there, focusing on those that seemed most relevant.

3.1.3 Companies House Dataset

By searching the company name, i.e. the market leaders for supermarkets in UK, I found information on Companies House relating to datasets I could potentially use. I used Miro to display all of the overview of the market leaders and proceeded to highlight the SIC codes (figure 2) that corresponded to a specific dataset, which had previously been scrapped.

Figure 2

Company Name	Address	Status	Incorporation Date	Nature of business (SIC)
J SAINSBUERY PLC	33 Holborn, London, EC1N 2HT	Active	10 November 1922	47190 - Other retail trade with food, beverages or tobacco products
ASDA STORES LIMITED	Asda House, South Bank, Great Wilson Street, Leeds, LS1/9AD	Active	18 February 1949	47190 - Other retail trade with food, beverages or tobacco products
MARKS AND SPENCER P.L.C.	Waterloo House, 35 North Wharf Road, London, W2 1NW	Active	17 June 1926	47190 - Other retail trade with food, beverages or tobacco products
LIDL GREAT BRITAIN LIMITED	Lidl House, 14 Kingston Road, Turbздton, England, KT9 5NU	Active	10 May 1993	47190 - Other retail trade with food, beverages or tobacco products
TESCO PLC	Tesco House, Shire Park, Kestrel Way, Weybridge, Surrey, KT12 2SD	Active	27 November 1947	47190 - Other retail trade with food, beverages or tobacco products
ALDI STORES LIMITED	Holy Lane, Atherton, Warrington, Cheshire, WA2 8QZ	Active	25 November 1988	70100 - Activities of head offices

From this, I was able to determine that two SIC codes were used for these companies: SIC Code 47110 Retail sale in non-specialised stores with food (figure 3), & beverages or tobacco predominating and SIC Code 47190: Other retail sale in non-specialised stores (figure 4).

Figure 3

ACTIVITIES FROM UK SIC CODE 47110

SIC Code 47110: Retail sale in non-specialised stores with food, beverages or tobacco predominating

List of activities classified inside the UK SIC Code 47110

- Cinema kiosk (retail)
- Village general store (selling mainly foodstuffs) with alcohol licence (retail)
- Superstore (selling mainly foodstuffs) without alcohol licence (retail)
- Superstore (selling mainly foodstuffs) with alcohol licence (retail)
- Supermarket (selling mainly foodstuffs) without alcohol licence (retail)
- Supermarket (selling mainly foodstuffs) with alcohol licence (retail)
- NAAFI shop without alcohol licence (retail)
- NAAFI shop with alcohol licence (retail)

miro

Figure 4

ACTIVITIES FROM UK SIC CODE 47190

SIC Code 47190: Other retail sale in non-specialised stores

List of activities classified inside the UK SIC Code 47190

- Department stores (retail)
- General stores in which the sale of food beverages or tobacco products is not predominant (retail)
- Household stores (retail)
- Mixed business retailing both food and non food goods but non-food predominating (retail)

miro

From the 'list of activities' detailed in figure 3 and figure 4, in conjunction with the actual datasets that had been scrapped as fallback solutions, I was able to determine that the dataset relating to SIC code 47110, pertaining to the retail sale in non-specialised stores with food, beverages, or tobacco predominating, was the most appropriate classification for this project. This dataset specifically encompasses various types of food-focused retail establishments, such as supermarkets and superstores, which are central to the analysis of aspects such as financial performance, in addition to general analysis of supermarkets. The

data scrapped within the CSV is relevant to understanding key metrics and operational aspects of the supermarket. This includes detailed insights into assets, equity, employee numbers and thus providing a snapshot of the financial health and operational scale of these retail companies, encompassing supermarkets.

Section 3.2: How data was gathered

Understanding how the dataset was created is key in regards to evaluating its reliability and relevance. As previously mentioned, the dataset I will be evaluating is obtained from Companies House, which relates to the retail sale in non-specialised stores with food, beverages, or tobacco predominating. This dataset encompasses detailed financial and operational data for businesses, such as supermarkets and superstores, that come under the 'activities' list, shown in 3.1.3.

Companies House is an executive agency of the Department for Business and Trade, who are responsible and the overseer of examining and publishing company information, within the UK (Companies House, 2023), who control and run what a Limited company can do. They register all UK Ltd businesses' information, and make this data freely available for the public to inspect (Companies House, 2024 [1]). Due to being an executive agency for the DBT, Companies House collects data to ensure transparency, regulatory compliance, economic analysis, fraud prevention and business support (Companies House, 2024 [2])(Hourihane, 2016).

The Companies House blog post "Big insights: collecting, analysing and presenting data" (Hourihane, 2016) explains how they effectively manage and use the large amounts of data it collects. Instead of just reporting data, Companies House conducts detailed analyses to find valuable insights. Additionally, they also combine data from different sources and work with experts to identify patterns, trends, and anomalies, which lead to suggestions for improving performance.

Companies House asks companies to provide information about their activities, annual accounts, and control structure by providing personal details of their directors (Companies House, 2024 [2]).

They collect data through mandatory 'filings' submitted by registered Ltd companies in the UK. These filings include annual financial statements and balance sheets, alongside profit and loss accounts (Government Digital Service, 2017).

The data collection method is the use of online or paper 'filings' that only directors can compose, which includes choosing an SIC code that corresponds with their businesses' activities, and then filling in the related documents, as discussed above, regarding the annual financial statements, etc (Companies House, 2022). The results of these filings are added to a database for the corresponding SIC code, allowing a dataset to be created.

Companies are legally required to submit accurate and up-to-date information upon the time of submission, ensuring a level of reliability in the data collected. Companies House reviews this information for entries such as the correct financial year, and if the information

or document are not accurate or up to date, they will withdraw the information until the company resends accurate information (Companies House, 2021).

The documents are collected and published/updated annually by Companies House, providing an overview of the financial performance of these businesses, under their associated SIC code. The dataset I am using covers recent years, showing both the latest year (the year the data corresponds to and was collected) and the previous year (the year prior to the current year, from which comparisons can be made). The most recent year data was collected and published on was 2022, and the oldest year was 2016. However, further investigation into the columns showed that companies that have records dating back as far as 2016 also have records for 2017, 2018, etc (Section 3.3). These years were easily found by formulas preformed on the columns in the dataset, shown below.



The image contains two side-by-side screenshots of Microsoft Excel cells. The left cell displays the formula '=MAX(BF2:BF1316)' in blue text. The right cell displays the formula '=MIN(BF2:BF1316)' in blue text. Both formulas are enclosed in black-bordered boxes.

Since this is a scrapped dataset, the latest update within the drive is April 2022 (figure 5), this can have some effect on the 'up-to-date' aspects of the current data available, however it is still fairly recent and should not cause much concern, in addition to indirectly showing that data collected in 2022 was also published in 2022.

Figure 5

Name	Owner	Last modified
NI_47110.csv	Couldn't load user	Apr 21, 2022

However, I noticed on the Companies House website that they only perform 'basic checks' on the documents submitted to them (figure 6). Therefore, if companies submit financial information that is incorrect or misleading, whether intentionally or unintentionally, this will lead to skewed results in analysis, which in turn will have effect on the reliability of the results. For example, a company who want to present themselves in a better financial state may skew the information slightly, such as increasing asset value, thus creating a biased and untruthful account for their performance.

Additionally, Companies House only includes companies that meet filing requirements. This creates a sampling bias, as it may mostly represent larger more established companies that are more likely to follow these regulations. Smaller or newer companies, which might have difficulties with compliance, could miss from the dataset. This bias can skew the analysis towards performance of larger firms.

Also, the dataset I chose only covers the years from 2016 to 2022, but not all companies update their reports annually. This time-related bias can impact the analysis, particularly if significant changes occurred in the industry during the gaps in data reporting. For example, in 2016, a new law was established that required all companies to submit information on people who have significant control over the business, thus any records collected before 26

June 2016 will not have this information within the dataset (Lloyd, 2016). Even in 2024, there is a new law to state that all businesses must also provide Companies House with a registered email address, thus previous datasets, even the one I have picked, does not include this information as it was not mandatory (Changes to UK company law, n.d.).

Figure 6

We carry out basic checks on documents received to make sure that they have been fully completed and signed, but we do not have the statutory power or capability to verify the accuracy of the information that companies send to us. The fact that the information has been placed on the public record should not be taken to indicate that Companies House has verified or validated it in any way.

When Companies House examines accounts it only makes basic checks on them, for example to ensure that the appropriate documents are there, that they are for the correct financial year and that they have been signed. If, subsequently, we receive a complaint that the accounts do not comply with the Companies Acts in some way we have a duty to draw this to the company's attention and, if appropriate, to ask for a set of revised accounts which do comply with the requirements.

Section 3.3: Understanding the columns

First of all, on Miro, I made a list of all the columns in the dataset (figure 7), to ensure I had an overview of the columns, and was able to 'tick' them off as I went through attempting to understand them. This list also served as a reminder of the columns and data I wanted to look at when inspecting the original reports and filings from companies.

Figure 7

Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2
Nature of business (SIC) 3	No. Officers	Officer Name 1	Role 1	Status 1	DOB 1	Appointed On 1	Nationality 1	Country of Residence 1
Officer Name 2	Role 2	Status 2	DOB 2	Appointed On 2	Nationality 2	Country of Residence 2	Officer Name 3	Role 3
Status 3	DOB 3	Appointed On 3	Nationality 3	Country of Residence 3	Officer Name 4	Role 4	Status 4	DOB 4
Appointed On 4	Nationality 4	Country of Residence 4	Officer Name 5	Role 5	Status 5	DOB 5	Appointed On 5	Nationality 5
Country of Residence 5	Officer Name 6	Role 6	Status 6	DOB 6	Appointed On 6	Nationality 6	Country of Residence 6	XBRL Available?
Link to Report	Report Number	Report Type	Current Year	Currency	Fixed Assets(CUR)	Net Fixed Assets(CUR)	Current Assets(CUR)	Creditors(CUR)
Equity(CUR)	Employee(CUR)	Previous Year	Fixed Assets(PREV)	Net Fixed Assets(PREV)	Current Assets(PREV)	Current Liabilities(PREV)	Equity(PREV)	Employee(PREV)

I visually inspected the data for around 120 minutes, so these observations are not statistical but rather initial reactions that provide a fair insight into the dataset. To verify the accuracy of my interpretation of the values, I cross checked a sample row with a second source, namely the actual documents used to produce this dataset, accessed directly from Companies House. This helped confirm that the values were being interpreted correctly.

Below is the method in which I was able to attain the filings and reports from these companies, and gain access to the PDF's for certain years. I gathered the 'Company Number' from the dataset, used Google to find the Companies House related resource, then navigated to 'Filing history' and accessed the document corresponding to the 'current year' column of the row I was wanting to investigate. It is worth noting at this point that this was also the method I used to inspect outliers and 'not available' data entries within the dataset, which will be discussed in further detail later.

The screenshot illustrates the workflow for verifying company data. At the top, a Google search bar shows the query "NI620056". Below it, the search results page for GOV.UK displays the company overview for "727 YOUR STORE LIMITED". The "Filing history" tab is selected, showing a list of documents:

Date	Description	View / Download
13 May 2024	Confirmation statement made on 7 April 2024 with no updates	View PDF (3 pages)
30 Sep 2023	Micro company accounts made up to 31 December 2022	View PDF (3 pages) Download XBRL
07 Apr 2023	Confirmation statement made on 7 April 2023 with no updates	View PDF (3 pages)
30 Sep 2022	Micro company accounts made up to 31 December 2021	View PDF (3 pages) Download XBRL
03 May 2022	Confirmation statement made on 9 April 2022 with no updates	View PDF (3 pages)
01 Feb 2022	Termination of appointment of Deborah Victoria Patterson as a director on 19 May 2021	View PDF (1 page)

The following is the breakdown and my understanding of what the columns mean. The italicised title refers to the column name, the variable refers to attribute or characteristic of the entities being studied, the data type refers to the kind of data stored in each column and Missing/invalid refers to any entries which are absent or void, thus detailing how the dataset represents them in each column. I would also like to mention here that this part of the report is not 'beautified', yet simply a more clean version of how I was writing and interpreting the information on Miro. I believe this will help others who may want to do further investigation see how rough and 'imperfect' analysing columns can be, yet still show the extensive work that went into understanding them.

Company Name

- Variable: Name of company
- Data type: text
- Missing/invalid: No data missing or seemingly invalid

Company Number

- Variable: Unique identifier for company
- Data type: text
- Missing/invalid: No data missing or seemingly invalid

Registered Office Address

- Variable: Official address of company (not the actual place of the store)
- Data type: text
- Missing/invalid: No data missing or seemingly invalid

Company Status

- Variable: Operational status of the company (e.g., Active, Dissolved)
- Data type: categorical text + nominal
- Note: It seems to be deemed under categorial data type, as the responses are either 'Active', 'Liquidation', Voluntary arrangement', 'Dissolved' or 'Active proposal to strike off', thus detailing the companies activity status, regarding if it is operating or not, during that time
- Missing/invalid: No data missing or seemingly invalid

Company Type

- Variable: Type of company
- Data type: categorical text + nominal
- Note: Categorical as either 'private limited company' or private unlimited company', which indicates their specific type of corporate structure, also note there is a handful of 'Private Limited Company by guarantee without share capital use of 'Limited' exemption' and one instance of 'Private company limited by guarantee without share capital'
- Missing/invalid: No data missing or seemingly invalid

Incorporated On

- Variable: Date the company was incorporated i.e. date the company was legally formed and registered with authorities.
- Data type: date, hyphenated i.e. 26/01/2006 in cell but displayed in CSV as 26-Jan-06

- Missing/invalid: No data missing or seemingly invalid

Nature of Business (SIC 0,1,2,3)

- Variable: "Standard Industrial Classification" i.e. a four digit numerical code assigned to businesses to classify their primary industry and the type of economic activities they are involved in
- Data type: Integer
- Note: SIC 0, 1, 2, 3 relates to the 4 columns where these are input i.e. SIC 1 is the primary Sic, SIC 2 is the next column along that has the secondary SIC code
- Missing/invalid: There is no data missing but there is a few that have different SIC codes to 47110, which is the dataset we are currently analysing. The other SIC codes are 46390, 43991, 46900, 46711, 10890, 46341, 10710, 46170, 45111, 46342, 46380, 10520, 46900. I bier the reasoning for this as within columns of SIC 1 2 and 3 we can see that if 47110 is not mentioned in the primary column SIC 1) then it is usually mentioned in the secondary or tertiary etc columns related to the nature of the business. Missing data is then attributed to N/A.

No. Officers

- Variable: Number of officers in the company and number of resignations of those officer
- Data type: text + integer e.g. '4 officers / 2 resignations'
- Missing/invalid: No data missing or seemingly invalid

Please note: The brackets with numbers (e.g., 1-6) indicate the number of associated columns under a particular column name, such as 'Officer Name 1' having six related columns. The dataset organises these columns in a section-like format, such as 'Officer Name 1, Role 1... Officer Name 2, Role 2...'. If the missing/invalid section indicates no issues, it means all columns within that section are error free. However, in later sections, some companies may not have the specified number of officers, resulting in 'not applicable' entries in those cells.

Officer Name (1,2,3,4,5,6)

- Variable: name of an officer
- Data type: text, comma separated i.e. 'CLARKE, Jacqueline'
- Missing/invalid: No data missing yet some are 'Not applicable', which is seemingly due to privacy concerns, yet other areas of importance such as financial records also have 'Not applicable', so should not be a cause for concern.

Role (1,2,3,4,5,6)

- Variable: Role of the officer
- Data type: categorical text as seemingly only two options, director or secretary
- Missing/invalid: No data missing or seemingly invalid

Status (1,2,3,4,5,6)

- Variable: Activity status of officer
- Data type: categorical text as seemingly only two options, active and resigned

- Missing/invalid: No data seemingly invalid, however there are 2 officers which have 'not available' as their status, which seems to be an incomplete filing when looking at records.

DOB (1,2,3,4,5,6)

- Variable: officer's date of birth
- Data type: date, hyphenated i.e. 01/12/1987 in cell but displayed in CSV as Dec-87
- Missing/invalid: No data missing in the sense of all cells are complete, yet there are a significant amount that say 'Not available', due to data of birth not being a requirement to disclose due to privacy concerns.

Appointed on (1,2,3,4,5,6)

- Variable: Date officer was appointed i.e. date when they were officially assigned to their role as an officer of the company
- Data type: date, hyphenated i.e. 27/08/2013 in cell but displayed in CSV as 27-Aug-13
- Missing/invalid: No data missing or seemingly invalid

Nationality (1,2,3,4,5,6)

- Variable: nationality of officer
- Data type: text
- Missing/invalid: No data missing in the sense of all cells are complete, yet there are a significant amount that say 'Not available'. Further research showed that certain types of companies have different reporting obligations, which could result in the nationality of officers not being recorded in all cases, such as 'Micro-companies' in addition to nationality seemingly not being a requirement for officers to disclose.

Country of Residence (1,2,3,4,5,6)

- Variable: country of residence of officer
- Data type: text
- Missing/invalid: No data missing in the sense of all cells are complete, yet there are a significant amount that say 'Not available'. Further inspection into these reports show that it was never mentioned as most were submitted as "Micro-company" reports, which do not usually detail that information. Also, some cells say 'Northern Ireland' and others say 'N Ireland' and 'N.Ireland', which is something to be cautious of when analysing the data is based on country of residence.

iXBRL Available?

- Variable: Availability of Inline XBRL (eXtensible Business Reporting Language) report – i.e. the filing that they complete
- Data type: categorical (yes/No-PDF only) (i.e. yes there is a iXBRL file available, or no there is only a PDF (seemingly indicating that if there is only a PDF, the company has not provided or updated their financial information)).
- Missing/invalid: No data missing or seemingly invalid

Link to report

- Variable: URL link to report

- Data type: text (url) – some are URL which show a download query parameter, others are simply a link to filing history, based on if the iXBRL is available. The response to the iXBRL available seemingly corresponds to the URL presented in row.
- Missing/invalid: No data missing or seemingly invalid

Report Number

- Variable: Unique identifier for the report?
- Data type: integer (0-5) – unable to distinguish how this is used, no research on this has come back with a result either, but seemingly related to iXBRL link and report, thus potentially no cause for concern
- Missing/invalid: No data missing or seemingly invalid

Report Type

- Variable: Type of report in terms of if it is registers, unaudited, with some numbers as well if registered, giving a description of the report type.
- Data type: text
- Missing/invalid: No data missing, some say unapplicable or not available. Also, some are 'unaudited' and others are 'UNAUDITED FINANCIAL STATEMENTS' so again be aware of this is report is significant.

Current Year

- Variable: the year for which the data in the row is reported
- Data: date (year) e.g. '2020'
- Missing/invalid: Some cells within this column have 'Not applicable', however the remainder of the financial information is also 'Not applicable', and the report type is also not applicable and report number is 0 potentially indicating not applicable, as the iXBRL is not available, only PDF download. Therefore this has a knock-on effect, but can be avoided as most columns are then rendered not applicable.
- Note: CUR is an abbreviation, meaning columns with CUR in title is relating to the year in 'Current Year' of that row.

Please note that in Companies House records, each row does not represent a unique company but rather a company in a specific year. As a result, you might see multiple rows for the same company. If there appears to be repetition in the rows and entries, it's because the company's name, officers, and other details may remain the same across years, while the financial aspects corresponding to different years will vary. Thus, eradicating these seemingly 'duplicated' rows would result in the loss of company change over time.

Currency

- Variable: type of currency used for finance reports to follow
- Data type: text
- Missing/invalid: The majority if this column is 'Not available' or 'Non applicable'. The non applicable entries are related to those who do not have financial reports, as stated with 'Current year' section above, however the not available ones suggest it was never mentioned within the report. There are also a few £ and some other symbols that are unidentifiable such as 'ØΩ'.

Fixed Assets (CUR)

- Variable: value of fixed assets in current year – measured as currency
- Data type: integer , comma separated currency based i.e. 11,000
- Missing/invalid: Around half of the entries are ‘not available’, and some ‘not applicable’, due to no financial information being available for those companies. However, upon further investigation into some companies that have ‘not available’ as their ‘Fixed Assets (CUR)’, it seemed that on the financial report the Fixed Assets were reported, yet not input into the data set. Figure 8 shows the data records in the dataset regarding A & A Harding Co, and figure 9 shows the filed document. The Net Fixed assets and Current Assets alight with the dataset, yet it seems there is a more underlying reason as to why the fixed assets were not included, since this data is absent from all data entries for this company, thus it alludes to an internal matter between the company and companies house.

Figure 8

2017	Not Available	Not Available	1,153	23,228
------	---------------	---------------	-------	--------

Figure 9

A & A Harding Co Ltd		
Company Number: NI057200		
STATEMENT OF FINANCIAL POSITION		
as at 30 November 2017		
	2017 £	2016 £
Fixed Assets	62,442	65,190
Current assets	23,228	38,494
Prepayments and accrued income	541	532
Creditors: amounts falling due within one year	(83,378)	(88,245)
Net Current Liabilities	(59,609)	(49,219)
Total Assets less Current Liabilities	2,833	15,971
Accruals and deferred income	(1,680)	(2,012)
Net Assets	1,153	13,959
Capital and Reserves	1,153	13,959
Notes to the financial statements		

Net Fixed Assets (CUR)

- Variable: Net value of fixed assets in the current year – measured as currency
- Data type: integer , comma separated currency based i.e. 5,000
- Missing/invalid: A significant amount of the entries are ‘not available’ and some ‘not applicable’, due to no financial information being available for those companies. Also, some companies entries that have been ‘not applicable’ in previous finance related

columns sometimes have '1' in the column, thus I need to be aware of this, potentially causing outliers.

Current Assets (CUR)

- Variable: Value of current assets in current year – measured as currency
- Data type: integer , comma separated currency based i.e. 6,000
- Missing/invalid: A few of entries are 'not available' and some 'not applicable', due to no financial information being available for those companies, and if the fixed assets were not included, then the current assets seemingly are not included.

Creditors (CUR)

- Variable: Value of liabilities/creditors in the current year – measured as currency
- Data type: integer , comma separated currency based i.e. 6,000
- Missing/invalid: A few of entries are 'not available' and some 'not applicable', due to no financial information being available for those companies. I do believe this is down to the amount of disclosure a company wants to undertake, and can be relative to the size of the company. Figure 10 shows the report submitted for 77 retail. This 'Creditors' information corresponds to what is found in the dataset. They undertook a 'Total exemption full accounts' filing, whereas companies who do not have this 'creditor' entry can be seen taking 'Micro company account' filings, which leads me to believe the scale and reach of a company directly correlates to the necessity of creditors being included in a filing, leading to 'not available' entries, although the likelihood of smaller companies having creditors is quite high due to cash flow management and limited access to capital.

Figure 10

77 RETAIL 2 LIMITED (REGISTERED NUMBER: NI636843)					
BALANCE SHEET 30 NOVEMBER 2018					
	Notes	30.11.18		30.11.17	
FIXED ASSETS		£	£	£	£
Intangible assets	4		266,893		302,893
Tangible assets	5		205,383		245,531
			472,276		548,424
CURRENT ASSETS					
Stocks		80,235		50,427	
Debtors	6	58,441		38,979	
Cash at bank and in hand		65,800		79,360	
		204,476		168,766	
CREDITORS					
Amounts falling due within one year	7	647,918		714,408	
NET CURRENT LIABILITIES			(443,442)		(545,642)
TOTAL ASSETS LESS CURRENT LIABILITIES			28,834		2,782
PROVISIONS FOR LIABILITIES			5,189		6,134
NET ASSETS/(LIABILITIES)			23,645		(3,352)
CAPITAL AND RESERVES					
Called up share capital	10		100		100
Retained earnings			23,545		(3,452)
SHAREHOLDERS' FUNDS			23,645		(3,352)

Equity (CUR)

- Variable: Value of liabilities/creditors in the current year – measured as currency
- Data type: integer , comma separated currency based i.e. 6,000

- Missing/invalid: A few of entries are ‘not available’ and some ‘not applicable’, due to no financial information being available for those companies. Although, smaller companies may have smaller equity in comparison to larger companies, yet there is still a drastic change in numbers. Further investigation shared that those companies who had their equity shared as ‘1’ (figure 11), ‘2’, or ‘100’(figure 12) etc have this information put into the field of their filing being ‘called up share capital’. Whereas companies who have their Total net assets and Capital reserves input currently seemingly have their equity calculated off of that (figure 13). It is fair so say that there are discrepancies in the way equity is calculated, and if companies are putting in the amount of shares, a random filler number, or the actual value, which is clearly the value duplicated from the ‘Net Fixed Assets’ column entry. Here are some company filings I found that show the numbers

Figure 11

ADC ENTERTAINMENTS (N.I.) LIMITED			
Balance sheet			
As at 31 March 2019			
	Notes	2019	2018
		£	£
Fixed assets			
Intangible assets:	2	30,000	30,000
Tangible assets:	3	433,372	259,707
Total fixed assets:		<u>463,372</u>	<u>289,707</u>
Current assets			
Stocks:		5,063	4,635
Debtors:		15,771	15,428
Cash at bank and in hand:		49,868	32,006
Total current assets:		<u>70,702</u>	<u>52,069</u>
Creditors: amounts falling due within one year:		(63,235)	(70,555)
Net current assets (liabilities):		<u>7,467</u>	<u>(18,586)</u>
Total assets less current liabilities:		470,839	271,121
Creditors: amounts falling due after more than one year:		(211,975)	(149,071)
Total net assets (liabilities):		<u>258,864</u>	<u>122,050</u>
Capital and reserves			
Called up share capital:	1	1	1
Profit and loss account:		258,863	122,049
Shareholders funds:		<u>258,864</u>	<u>122,050</u>

Figure 12

77 RETAIL 2 LIMITED (REGISTERED NUMBER: NI636843)					
BALANCE SHEET 30 NOVEMBER 2018					
	Notes	30.11.18		30.11.17	
FIXED ASSETS					
Intangible assets	4		266,893		302,893
Tangible assets	5		<u>205,383</u>		<u>245,531</u>
			472,276		548,424
CURRENT ASSETS					
Stocks		80,235		50,427	
Debtors	6	58,441		38,979	
Cash at bank and in hand		<u>65,800</u>		<u>79,360</u>	
		204,476		168,766	
CREDITORS					
Amounts falling due within one year	7	<u>647,918</u>		<u>714,408</u>	
NET CURRENT LIABILITIES		(443,442)		(545,642)	
TOTAL ASSETS LESS CURRENT LIABILITIES		28,834		2,782	
PROVISIONS FOR LIABILITIES		<u>5,189</u>		<u>6,134</u>	
NET ASSETS/(LIABILITIES)		<u>23,645</u>		<u>(3,352)</u>	
CAPITAL AND RESERVES					
Called up share capital	10	<u>100</u>		<u>100</u>	
Retained earnings		23,545		(3,452)	
SHAREHOLDERS' FUNDS		<u>23,645</u>		<u>(3,352)</u>	

Figure 13

727 YOUR STORE LIMITED		Registered Number NI620056	
Micro-entity Balance Sheet as at 31 December 2019			
	Notes	2019	2018
Fixed Assets		13,830	17,288
Current Assets		17,578	3,558
Creditors: amounts falling due within one year		(16,630)	(18,330)
Net current assets (liabilities)		<u>948</u>	<u>(14,772)</u>
Total assets less current liabilities		<u>14,778</u>	<u>2,516</u>
Creditors: amounts falling due after more than one year		(11,130)	(2,007)
Total net assets (liabilities)		<u>3,648</u>	<u>509</u>
Capital and reserves		<u>3,648</u>	<u>509</u>

Employees (CUR)

- Variable: Average number of employees in the current year (Figure 11) – measured as a total number
- Data type: integer
- Missing/Invalid: ‘Not Available’, ‘Not applicable’, or zero (0), causing concern if company is still active.

- Upon investigation, the 'Not available' entries are correctly associated with the absence of employee numbers in the filings. The 'Not applicable' entries result from the company not providing information due to liquidation, dissolution, or simply not submitting any financial information to Companies House, reflected by the 'Not applicable' entries within subsequent finance related columns. Therefore, it is appropriate that these figures do not apply to those companies. However, I conducted further investigation into the zeros present in the dataset. Figure 15 illustrates an example of a company that reported '0' average employees, which aligns with the actual data in the company's filing. In all cases I've examined, companies with a '0' in the dataset have a corresponding '0' in their financial filings. This may be because smaller companies, in particular, have directors and secretaries who are seemingly not counted as employees under an employment contract, and therefore are reported as having '0' employees.

Figure 14

3. EMPLOYEES AND DIRECTORS
The average number of employees during the year was 18 (2018 - 24) .

Figure 15

AZAC CASTLEROCK LIMITED NOTES TO THE ACCOUNTS FOR THE PERIOD FROM 3 FEBRUARY 2020 TO 28 FEBRUARY 2021
--

8 Average number of employees

During the period the average number of employees was 0.

There are seven additional columns related to the previous year's financial performance. These columns show results that are very similar to the current year's data in terms of outliers, 'not available' entries, and reporting inconsistencies, which I have already discussed in detail above. For example, in the 'Previous Year' column, most dates are reported as '20XX,' but some are formatted differently, such as '31/08/2016'. These 'previous year' finance related columns have the same names and types of entries as the current year (CUR) financial data, but they are labelled as (PREV) beside the column name, to distinguish them as relating to the date within the 'previous year' column. The only unique aspect is that these columns use the same 'currency' type as the other columns, so they do not have a preceding currency column. To save space, I will not provide an in depth analysis of these columns, as they are practically identical to the current year's financial statistics, differing only in figures.

With regards to the relative importance of the dataset, it shows a detailed view of various aspects of companies in the supermarket sector, but the usefulness and reliability of each column does vary. Key columns like 'Current Assets (CUR)', 'Creditors (CUR)', 'Equity (CUR)', and 'Net Fixed Assets (CUR)' are important for evaluating a company's financial health, giving insights into aspects such as long-term asset investment or the financial risks they

may be facing. However, issues such as missing data and inconsistencies, especially in the equity column, require careful interpretation when analysing as this can cause a major issue with outliers and reading the visualisation wrongly. Moreover, additional columns, such as those detailing company officers and operational status, provide useful context for understanding governance structures and company types, which can influence strategic decisions.

Furthermore, columns related to financial reports, like 'iXBRL Available' and 'Link to Report,' are more administrative rather than associated with any of the values. Temporal data from the "Current Year" and "Previous Year" columns is essential for conducting trend analysis, ensuring that comparisons are made within the correct time frame. The "Currency" column is crucial for comparing financial data across companies that report in different currencies, but inconsistencies here also limit its reliability.

With regards to the columns correspondence to metrics of value, the columns related to financial data, like 'Fixed Assets (CUR)', 'Net Fixed Assets (CUR)', 'Current Assets (CUR)', 'Creditors (CUR)', 'Equity (CUR)' and 'Employees (CUR)', are closely tied to the financial performance of a supermarket. These figures provide insight into the supermarket's financial health and operational scale, which are essential for decision makers when evaluating profitability and making strategic choices. Although this document emphasises financial metrics, these numbers also indirectly affect customer satisfaction and the supermarket's image. For instance, effective asset management and strong financial health can lead to better customer service by allowing resources to be more effectively directed towards enhancing customer experience, ultimately improving customer satisfaction and the supermarket's image. Furthermore, the data in these columns not only represent financial outcomes but also hint at potential factors driving these outcomes. For example, the average number of employees within the company ('Employees (CUR)') can impact operational efficiency and customer satisfaction, both of which are deemed key by the decision makers to ensure financial success. Similarly, 'Creditors (CUR)' and 'Equity (CUR)' reflect financial obligations and ownership, respectively, which influence the supermarket's financial stability and its ability to invest in initiatives that focus on customer satisfaction, thus causing a 'chain reaction' like approach.

Section 3.4: Missing and Null Values

1. Identification and Analysis of Missing/Null Data

I began my search for missing or null values by identifying special codes used in the columns of the CSV file. I had already completed this to a degree in section 3.3, but I revisited this task, focusing on identifying any codes I might have missed and understanding their extent as a precursor to the visualisations. The codes I identified were "Not Available" (e.g. Figure 16), "Not Applicable" (e.g. Figure 17), and "N/A" (e.g. Figure 18). Notably, "N/A" was found exclusively in the "Nature of business" columns, and only within the last three columns. It is also important to mention that when the CSV was opened in 'MacOS Numbers', a 'blank field' was observed (Figure 19), which appeared as 'ØΩ' when opened in MS Excel (Figure 20). This entry was located in column 3, specifically within the 'Currency' column. While I will delve further into the implications of this for analysis later in this section, it is worth

noting here that this seems to be the only column with a genuinely missing value, i.e. a null entry.

Figure 16

Currency	Fixed Assets(CUR)	Net Fixed Assets(CUR)	Current Assets(CUR)	Creditors(CUR)
₪	11,064	6,425	64,782	18,364
₪	13,830	3,648	17,578	16,630
₪	17,288	509	3,558	18,330
₪	17,859	17,323	106,263	54,157
Not Available	337,671	167,892	418,863	586,318
Not Available	399,066	36,028	213,046	571,869
Not Available	472,276	23,645	204,476	647,918
Not Available	548,424	3,352	168,766	714,408
Not Available	625,907	60,543	210,231	773,282
Not Available	61,858	169	75,566	Not Available
Not Available	Not Available	6,198	19,687	Not Available
Not Available	Not Available	1,145	24,089	Not Available
Not Available	Not Available	1,153	23,228	Not Available
Not Available	Not Available	Not Available	35,737	Not Available
Not Available	Not Available	Not Available	37,387	Not Available
Not Available	Not Available	Not Available	38,701	Not Available

Figure 17

Nationality	3	Country of Re	Officer Name	Role	4	Status	4	DOB	4
British		Northern Ire	Not Applicab	Not Applicab	Not Available	Not Applicab			
British		Northern Ire	Not Applicab	Not Applicab	Not Available	Not Applicab			
British		Northern Ire	Not Applicab	Not Applicab	Not Available	Not Applicab			
British		Northern Ire	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	Not Applicab	Not Applicab	Not Available	Not Applicab			
Not Applicab		Not Applicab	HARDING, A	Director	Resigned			Feb-71	
Not Applicab		Not Applicab	HARDING, A	Director	Resigned			Feb-71	
Not Applicab		Not Applicab	HARDING, A	Director	Resigned			Feb-71	
Not Applicab		Not Applicab	HARDING, A	Director	Resigned			Feb-71	

Figure 18

Figure 19

Currency	
£	
£	
£	

Figure 20

	BG
1	Currency
2	£
3	ØøΩ
4	£
5	£

Figure 21 displays a heatmap indicating which variables have missing values. If a section of the heatmap is yellow, it signifies the presence of a missing or null value in that variable's column for a particular row, based on criteria such as blank cells, "Not Available," "Not Applicable," or "N/A." The bar chart in Figure 22, illustrates the percentage of missing data for each variable within the CSV file, shown by the use of the orange bars. Using my prior knowledge of Python, ChatGPT for errors with python being previously installed and confliction with Anaconda Python installation, in conjunction with the visualisation resource from GeeksforGeeks (2023), I was able to create both the heatmap and the bar chart using Jupyter notebooks.

The visualisations clearly show that the missing values mostly come from the lack of information on additional SIC codes and company officer details and demographics, which is expected. I'll discuss this in more detail later. However, the financial variables are more problematic. It seems that businesses that didn't report one financial entry often missed others as well. The visualisations do not differentiate between "Not Applicable" and "Not Available," so this will need to be clarified and investigated further. This is concerning because the main purpose of this dataset is to provide insights into financial aspects. Homing in fully on the bar chart, it seems that categorical data is the only aspect of the dataset that contains almost all data, with no missing or null aspects.

Figure 21

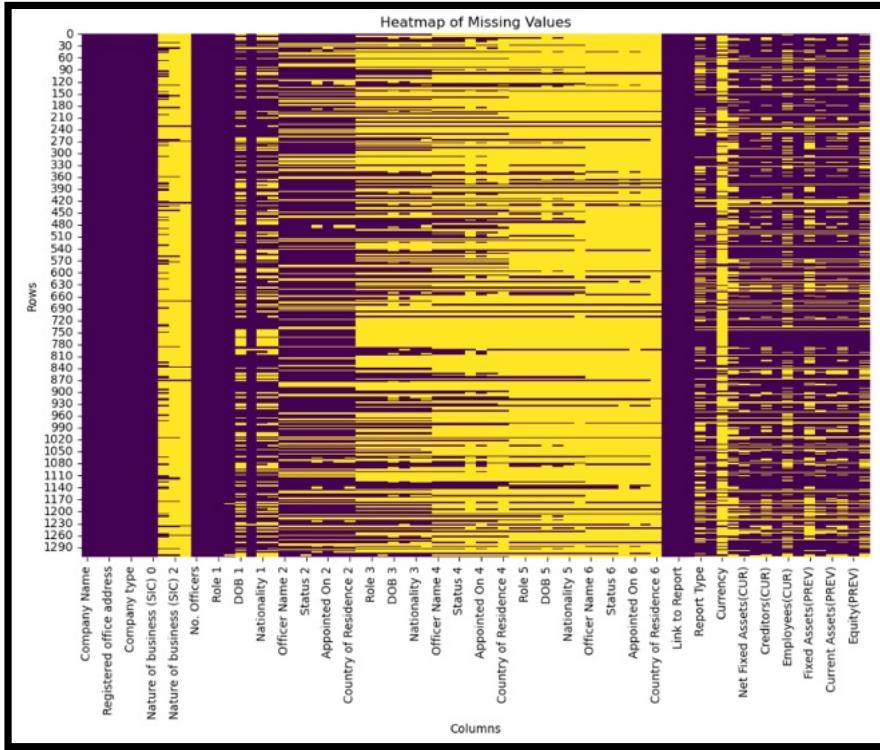
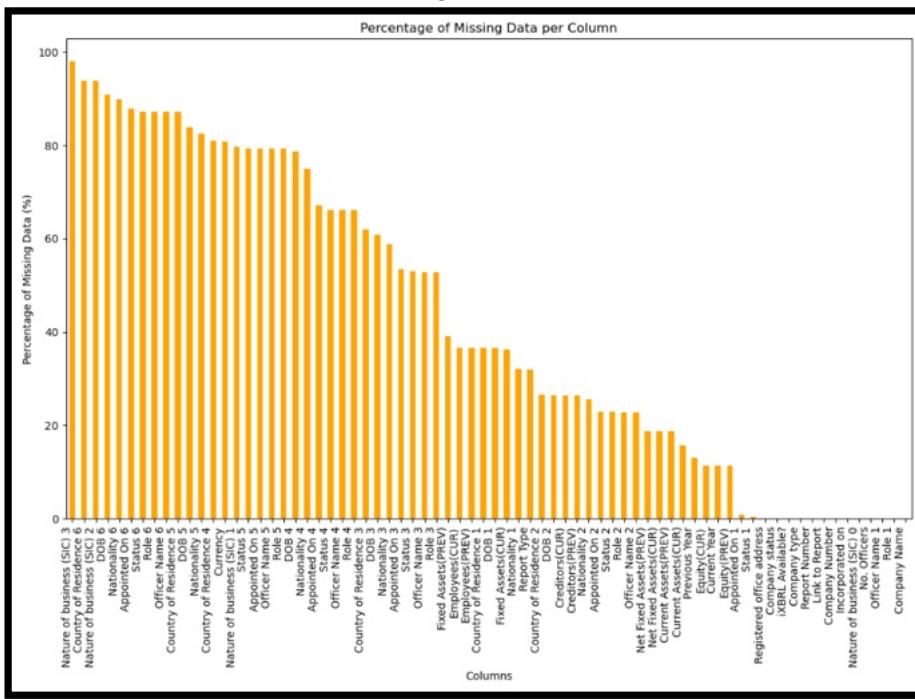


Figure 22



2. Impact Assessment of Missing Data

Missing data in financial variables like ‘Fixed Assets (CUR)’, ‘Net Fixed Assets (CUR)’, ‘Current Assets (CUR)’, ‘Creditors (CUR)’, and ‘Equity (CUR)’ can lead to biased conclusions about the financial health and performance of companies in the supermarket industry. For example, with a large portion of the dataset lacks information on equity or creditors, any analysis of financial stability or risk could be skewed. Similarly, missing values in average employee counts can lead to incorrect assessments of a company's operational scale.

This missing data, particularly in financial and operational columns, can result in the dataset not being fully representative of the industry. For instance, if smaller companies are more likely to have missing financial data (perhaps due to less intense reporting requirements), the analysis may overrepresent larger, more established companies. This could distort our understanding of industry trends, making conclusions more reflective of big companies rather than the entire industry. Additionally, if data from some smaller companies is included but incomplete, these could be misinterpreted as outliers.

Some missing data, such as the absence of additional officer details or SIC codes, appears to be systematic and tied to specific reporting practices. For instance, if a company doesn't have additional SIC codes, it will be marked as ‘N/A’ which is understandable, as they simply do not associate with any other subset of an industry. However, for companies that operate under secondary or tertiary SIC codes that are 47110, failing to include those fields could lead to misrepresentation, as the absence of these codes could be misleading, as face-value interpretation would persuade analysts that these companies are not in the same SIC code as the others in the dataset, having not included additional fields for second and tertiary etc SIC codes. Similarly, some companies might only have one director, so the first officer column is filled with the director's name, etc, while subsequent officer columns remain empty. In contrast, companies with multiple directors will have these columns filled accordingly. This variation reflects the company's structure rather than an error in the data. Another example is that ‘Not Applicable’ or ‘Not Available’ entries in financial data might occur because companies choose different types of filings, like “Micro company account” filings, which require less detailed financial disclosures. These are not necessarily errors but reflect differences in reporting standards and company size. This is also true for financial data, when a company marks one financial statistic as ‘Not Applicable,’ the rest of the related data is also be marked the same. Likewise with the chain reaction of no ‘Current Year’ or ‘Previous year’ entries, if they are ‘Not applicable’ the remainder of the financial data is also no existent. However, this is not always the case, as some companies simply do not report certain financial aspects.

The presence of genuinely missing data, such as a null entry in the “Currency” column or missing financial details despite their availability in company filings, suggests potential issues during the data collection process. These might be mistakes or inconsistencies in how the data was extracted. However, this should not cause much concern since the currency might not be directly relevant to the analysis. For instance, even if the currency type is not recorded, I can still compare financial trends across companies without needing to match values side by side and can look at the significance in different between the values in rows and compare this difference to another row. Moreover, since all the companies in the

dataset are based in the UK, it is likely that they report their finances in GBP, though this assumption is not fully supported as some companies may be based in UK but operate overseas or are recorded in a different currency due to stock-market aspects, so there are some risks associated. I believe I will continue to keep this column in my dataset to ensure that I can compare some companies based on the same currency on a factual account, being that they are both '£', and the column will also serve as a reminder, that not all finance related data is measured in the same currency, for myself or others who do further work on the refined dataset.

3. Strategies for Handling Missing Data

Next, I considered various ways to handle the missing data, knowing that leaving it as-is could lead to biased results. I also want to mention some methods that I found unsuitable for my dataset, to help others avoid the same dead ends I encountered. However, I recognise that these methods might be appropriate for different datasets under the right conditions, and thus can be helpful for further work.

Firstly, I considered the imputation of data with statistical measures, which involved replacing the missing values with a calculated statistic, such as the mean, median or mode of the available data. While imputation is a common method for handling missing data, I do not believe it is a good strategy for my dataset, especially consider the financial variables. Smaller companies do not report certain data because it could 1) reveal too much sensitive information or 2) they are not required to due to the specific filings they fill out, as mentioned previously. Imputing missing values with the median or mean of other companies would significantly skew results, as these values will not represent the missing data's true nature. For example, imputing small, unreported financial figures with the average values of larger companies could distort the analysis, making smaller companies appear more financially stable than they are. Since my missing data is intentionally not recorded due to privacy or lack of legal obligations, estimating it through imputation is possibly not the best course of action.

Moreover, after a browse online for some ideas, I can across listwise deletion (Grace-Martin, 2023). Essentially, this involved excluding rows of data with missing data from the analysis. It is recommended when the amount of missing data is minimal and does not significantly impact the overall dataset size. However, in my dataset, this exclusion will disproportionately affect smaller companies (which already has less data due to less intense reporting requirements), and also further skew the analysis regarding larger companies. This method is also not to be used if the missing data is not randomly missing, which also raises issues in terms of my dataset.

I also considered the indicator method (Wikipedia, 2024) which involves creating an indicator variable that flags the presence of missing data. This method allows for the retention of data while accounting for missing values in the analysis. By creating a separate indicator variable, I could explore whether the absence of certain data correlates with other variables, such as company size or financial performance. This method is particularly useful when missing data is systematic, as it is in my dataset, where certain financial figures might

be missing due to different reporting requirements. This is more likely to apply in relation to my dataset.

Finally, I looked into analysing subsets of data where missingness is not an issue, such as focusing only on large companies with complete data. This method allows for the analysis of data where it is most complete, thus avoiding the issues of imputation or exclusion. In my dataset, this could involve conducting separate analyses for smaller and larger companies. While this approach might limit the generalisations the findings, it ensures that the analysis remains accurate and relevant to each subset, thus more applicable for my dataset.

Now, I would like to touch on the things that I would find appropriate for my dataset. With regards to ‘Not applicable’ data, it is clear that, as mentioned before, a significant number of businesses do not record financial information within Companies House, thus is not reflected in this revenue orientated, finance related dataset. I believe that removing these rows would simplify the dataset and allow for some of these ‘missing values’ to be addressed. However, rows where only a few columns are marked as ‘Not Available’ will remain, as they still provide valuable financial information, such as revenue, in other relevant columns, and will simply not be picked up in the analysis within the jupyter notebook, as they can be converted to (N/A) and will simply not be used for visualisations.

Below are some examples of the rows that could be removed, or targeted:

2020	-	Not Available	2 Not Available	Not Available	2 Not Available	2019	Not Available	2 Not Available	Not Available	2 Not Available	2 Not Available
2019	-	Not Available	2 Not Available	Not Available	2 Not Available	2018	Not Available	2 Not Available	Not Available	2 Not Available	2 Not Available
2018	-	Not Available	2 Not Available	Not Available	2 Not Available	2017	Not Available	2 Not Available	Not Available	2 Not Available	2 Not Available
Not Applicable											

-	-	0 -	0	0	0	2019	-	0 -	0	0	0
-	-	0 -	0	0	0	2018	-	0 -	0	0	0

Also, I considered the removal of the ‘Nature of Business’ columns, or at least the non-primary columns, which contain the SIC codes for companies, due to a significant number of ‘Not Applicable’ entries in the later columns. However, upon further reflection, I realised that these columns could be valuable for future data linking, particularly with other datasets or companies associated with the secondary or tertiary SIC codes of the businesses in my dataset. Additionally, removing rows corresponding to datasets where SIC code 47110 is listed as a secondary or tertiary nature of business might be a mistake, as it could prevent the identification of patterns or causes that are not evident in businesses solely identified by SIC code 47110. Therefore, I concluded that these columns should be retained for now, but later chapters will explore this.

Section 3.5: Converting data into a more analysable form

My chosen dataset is in CSV format, which is fully compatible with both Excel and Python. This makes the CSV format acceptable and practical for analysis, as it can be easily imported, read, and manipulated in Python, ensuring smooth integration. In Section 3.4, I was able to create bar charts and heat maps without any issues, confirming that the file type and format are appropriate for my analysis.

Furthermore, that the dataset primarily consists of financial information from Companies House, there is no need for pivot tables. Pivot tables are typically useful for summarising data related to items, products, or transactions over time, which are situations that do not apply to this dataset.

Moreover, the only location data available in the dataset is the registered office address of each company, not the actual store locations. Converting these addresses to GPS coordinates would be impractical and unnecessary, particularly since there are over 1,300 rows, and most store locations are not directly linked to these registered addresses. Furthermore, manually searching for and verifying each store's location is not feasible due to the lack of disclosure and the potential mismatch between the registered and actual locations. Dropping this column could be useful for data cleaning, due to the fact it is not significant enough for the investigation.

The dataset contains limited unstructured text, primarily limited to company names and addresses, with no significant references to brands or products, as the focus is on Companies House filings. As such, there is no need for further analysis of unstructured text in this context. However, at this point, it could be good to remove the columns related to officer details, as many entries are marked as 'Not Applicable,' and it is unlikely that any significant correlations or insights could be derived from the details of multiple officers and their roles. These columns do not provide valuable information, such as officer values. The only aspect this could indicate is a subset of time-series related data, however, since I have the actual 'Incorporated on' and '... Year' entries, I do not believe this is significant and is not key attributes. While this data might be useful for other forms of analysis, it is not relevant to the objectives of my project.

While the dataset does not currently include detailed external data, there is potential for data linking in future work. For example, linking revenue data from Companies House with manually researched information, such as a company in a given year that implements new security, seasonal promotions, or aspects of manual investigation, could provide insights into causal relationships. However, this would require significant manual effort to gather and integrate this additional data. For the current analysis, we will acknowledge these possibilities and consider them, particularly for anecdotal analysis where correlations between trends could be explored after the initial findings. I do believe that comparing my anecdotal investigation appoints from Chapters 1 and 2 will provide a significant amount of anecdotal data linking, which is not ideal, yet certainly beneficial in terms of finding causal factors of decision maker values.

A column under consideration for removal is the 'Company Number.' While this is a unique identifier used by Companies House, it's not particularly useful for analysis. The same information can be accessed through the company name, which is more intuitive and meaningful for the analysis process. Keeping the 'Company Number' column adds redundancy, so by removing it, I can focus on data that directly contributes to generating meaningful insights. If more detailed information is needed about a specific company, the company name can serve as a sufficient identifier.

While some datasets require extensive conversion to be useful, this dataset does not need significant changes for the current analysis. However, as mentioned in Section 3.3, dates in my CSV are seemingly presented in different formats across different columns and thus can cause issues when working with the data in Python. Python, specifically libraries like pandas, requires consistency in date formats for effective processing and analysis. I, again, visually inspected these date columns. The ‘Appointed on’ column, which records the date when an officer was appointed, is standardised in the format DD/MM/YYYY. The ‘Incorporated on’ column, which indicates the date when the company was legally formed and registered with authorities, is also standardised in DD/MM/YYYY format. This consistent formatting is essential for analysing the longevity and growth of companies over time, as well as for performing any chronological comparisons. The ‘Current Year’ and ‘Previous Year’ columns are formatted as YYYY, reflecting the year in which financial data was recorded. This format is consistent and straightforward, helping the analysis of financial trends across different time periods. However, if analysis through python does flag any obscured or abnormal date layouts, they will be corrected to the ‘majority’ of the columns layout.

On the topic of years, an observation I have made is that the dataset currently includes both ‘Current Year’ and ‘Previous Year’ data for each company, spread across multiple rows. This duplication, where the ‘Previous Year’ data essentially mirrors information available in the ‘Current Year’ rows, seems to add unnecessary complexity to the dataset. Given that python allows for easy referencing of any year’s data dynamically by simply accessing the relevant rows, these ‘Previous Year’ columns appear redundant. Including them not only clutters the dataset but also risks complicating the analysis process. By removing these columns, the dataset becomes more straightforward and easier to navigate, which should help prevent potential errors and confusion during analysis. This approach also aligns better with the capabilities of python, where historical data can be efficiently accessed without the need for these extra columns.

Additionally, removing the columns related to the original report from which the data was scraped, specifically ‘iXBRL Available?’, ‘Link to Report’, ‘Report Number’, and ‘Report Type’, could be wise. The information in these columns is not useful for data linking, and even if it were, the correlation would likely be weak given the nature of the content. Furthermore, these columns do not provide any meaningful insights for the analysis of the financial data and serve more as background information, as utilised in section 3.3. Therefore, I believe they should be removed from the dataset.

Furthermore, I’ve noted that the dataset includes a ‘Registered Office Address’ column. Upon reviewing multiple entries, it is clear that this address often does not represent the actual location where the business operates. In many cases, particularly for smaller businesses, this address is a requirement for filing purposes and might be the owner’s residence or an accountant’s office rather than a business location. Since this information does not contribute to the analysis and could introduce confusion, it seems likely I will drop this column to keep the dataset focused on relevant, actionable data.

Section 3.6: Conclusions & Further Work

1. Conclusions

The visual analysis conducted in this chapter has focused on identifying and understanding a suitable dataset relevant to the supermarket industry, specifically concerning financial performance and operational efficiencies. The data collection and preparation process provided a clear emphasis on aligning the dataset with the core issues and values identified in previous chapters. The decision to prioritise a dataset centred on financial performance, despite the challenges in obtaining the ideal dataset, demonstrates a realistic approach that ensures the relevance of the analysis.

The dataset sourced from Companies House, although not without its limitations, offers a good foundation for analysing the financial health and operational scale of supermarkets in the Northern Ireland. One of the most valuable areas identified is the financial health of supermarkets, specifically the analysis of key financial metrics such as fixed assets, current assets, creditors, equity, and employee numbers. These metrics provide critical insights into the supermarkets' operational efficiency and financial stability, which are essential for understanding their overall performance. This area is particularly valuable for further analysis because it directly correlates with the supermarkets' ability to sustain and grow their market presence, making it a crucial factor for decision makers in the industry, and especially as one of my expert's decision maker's core values is financial performance. .

The time series nature of the dataset also presents a valuable area for further analysis. Analysing trends over time, particularly in financial performance, could provide insights into how the supermarket industry's financial health has evolved. This analysis could be particularly useful for forecasting future trends and identifying long term patterns that may not be immediately apparent through simple visual analysis. The fact that each row does not represent a unique company but rather a company in a specific year, rather than having columns associated with a specific year is good as you can easily track changes over time without needing to reshape the data, in addition to more straightforward filtering and grouping data over time, which is essential for accurate time series analysis. An insight that can be gained, for example, is comparing employee numbers and current assets which can reveal how workforce changes impact overall operational efficiency, potentially identifying optimal staffing levels for profitability yet maintain customer satisfaction, or understanding what employee numbers really equate to financial metrics.

Another aspect of the visual analysis has been the identification of potential causal relationships through anecdotal information linked to the dataset. This chapter has highlighted the importance of aligning anecdotal evidence with quantitative data to explore underlying causes, that impact financial performance. This approach will be brought forward and further developed in later chapters, where a more systematic linkage of anecdotal information to causal factors will be explored, providing a better understanding of the dynamics within the supermarket industry.

However, several challenges were encountered during the data preparation phase, including missing or null values and inconsistencies in the reporting of financial data. These issues

were thoughtfully addressed through consideration of various strategies, including data imputation, deletion, and indicator methods. The current decision to retain certain columns despite missing values, based on their potential for future data linking, balances current analytical needs with future research opportunities.

Another valuable area is the impact of missing and null values in the dataset. The presence of missing data, particularly in financial columns, could significantly skew the results and lead to biased conclusions. This issue is vital for further analysis because addressing these gaps could enhance the accuracy and reliability of the findings. Future chapters could focus on developing more sophisticated methods for handling missing data, such as statistical approaches that account for data gaps.

The visual analysis also highlighted the importance of understanding the data collection process, particularly the potential biases introduced by Companies House's reliance on company filings. The dataset's coverage of larger, more established companies, coupled with the exclusion of smaller firms due to reporting requirements, may skew the analysis towards reflecting the performance of more dominant companies in the industry. This bias must be acknowledged when interpreting the results of this analysis.

Furthermore, the exploration of markdown strategies and their impact on profitability represents another critical area for further analysis. While the current dataset primarily focuses on financial performance, future chapters could delve deeper into how effective markdown strategies can optimise sales and reduce waste, thereby improving overall financial outcomes, potentially being adapted by the wider industry.

2. Further Work

One promising area for further work is the integration of unstructured data, such as customer reviews, social media sentiment and feedback from customer satisfaction questionnaires. These data sources can provide valuable insights into customer perceptions, preferences, and overall satisfaction, which are factors influencing a supermarket's brand image and customer loyalty, which in turn are core values for my expert's company's decision maker. For instance, linking financial performance data with customer reviews from platforms like Google Reviews or Trustpilot could reveal correlations between customer satisfaction and financial outcomes. Text analysis techniques could be applied to these reviews to extract themes related to product quality, customer service or shopping experience. This unstructured data could then be quantified and compared with financial metrics, allowing for a deeper understanding of how customer perceptions impact profitability. Additionally, feedback from customer satisfaction surveys could be linked to specific periods or campaigns, helping to identify how marketing efforts or changes in service delivery affect customer sentiment. This analysis could be particularly valuable for tailoring strategies to enhance customer satisfaction and brand image.

Customer satisfaction and supermarket image are values of business success, and further work could focus on linking these qualitative aspects with quantitative financial data. For example, datasets containing Net Promoter Scores, customer loyalty metrics, or brand reputation surveys could be integrated with financial performance data to explore how

customer perceptions influence sales, profitability, and market share. Moreover, the exploration of online engagement metrics, such as social media activity or website traffic data, could provide additional context for understanding customer satisfaction. Analysing the correlation between online engagement and financial performance could help identify which digital strategies are most effective in building a positive supermarket image and driving customer loyalty.

While the current dataset focuses on financial performance, further work could involve a more detailed investigation into markdown strategies and their impact on profitability. This could be achieved by sourcing or creating a dataset specifically related to markdowns, potentially through primary data collection or by requesting specific data from supermarkets. Analysing the relationship between markdown timing, product categories, and sales performance could yield actionable insights for optimising markdown practices.

Further work should focus on investigating and verifying the anecdotal information which may be linked to causal factors, through manual data collection, rather than simply relying on the expert's information to fuel causes. This could involve gathering primary data, such as conducting interviews, surveys, or direct observations, to provide a better basis for linking anecdotal evidence to specific causal factors. This approach will help transform anecdotal observations into reliable data points that can be integrated with the existing dataset, providing the evidence based factors driving supermarket success.

Given the identified biases in the dataset, particularly the underrepresentation of smaller companies, future research could focus on addressing these gaps. This might involve sourcing additional data on smaller firms or adjusting the analysis to account for the biases, such as through weighting techniques or the use of statistical methods that show the impact of missing data.

APPENDIX

Kaggle (2024) *Kaggle: your machine learning and data science community*. Available at: <https://www.kaggle.com/>

Wikipedia (2024) *Wikipedia, the free encyclopaedia*. Available at: <https://www.wikipedia.org/>

Perplexity (2024) Perplexity AI [Artificial Intelligence Search Engine]. Available at: <https://www.perplexity.ai/>

Statista (2024). *Statista - the statistics portal*. Available at: <https://www.statista.com/>

Data.gov (2024). Available at: <https://www.data.gov/>

Google (2024) Available at: <https://www.google.com/>

Companies House (2023) *About Companies House*. Available at: <https://www.gov.uk/government/organisations/companies-house/about>

[1] Companies House (2024) *Home: Companies House*. Available at: <https://www.gov.uk/government/organisations/companies-house>

[2] Companies House (2024) *Your Personal information on the Companies House Register*. Available at: <https://www.gov.uk/guidance/your-personal-information-on-the-public-record-at-companies-house>

Hourihane, H. (2016) *Big insights: collecting. Analysing and presenting data*. Available at: <https://companieshouse.blog.gov.uk/2016/05/31/big-insights-collecting-analysing-and-presenting-data/>

Government Digital Service (2017) *Prepare annual accounts for a private limited company*. Available at: <https://www.gov.uk/annual-accounts>

Companies House (2022) *Frequently Asked Questions*. Available at: <https://ewf.companieshouse.gov.uk/help/en/stdwf/webIncHelp.html>

Companies House (2021) *Service Information*. Available at: <https://resources.companieshouse.gov.uk/serviceInformation.shtml>

Lloyd, G. (2016) *The new 'people with significant control' register*. Available at: <https://companieshouse.blog.gov.uk/2016/04/13/the-new-people-with-significant-control-register/>

Changes to UK company law. (n.d.). *Changes to UK company law*. Available at: <https://changestoukcompanylaw.campaign.gov.uk>.

GeeksforGeeks (2023) *Data Visualization in jupyter notebook*. Available at:
<https://www.geeksforgeeks.org/data-visualization-in-jupyter-notebook/>.

Grace-Martin, K. (2023) *When Listwise Deletion works for Missing Data*. Available at:
<https://www.theanalysisfactor.com/when-listwise-deletion-works/>.

Wikipedia (2024) *Dummy variable (statistics)*. Available at:
[https://en.wikipedia.org/wiki/Dummy_variable_\(statistics\)](https://en.wikipedia.org/wiki/Dummy_variable_(statistics)).

CHAPTER 2

Section 2.1: Understanding Values, Their Metrics, and Any Causal Factors

1. Identification of Business Values

After thoroughly reviewing the interview I conducted with my expert in the supermarket industry, I identified specific points that clearly illustrate the value of decision makers within this sector (individuals who control and have the final say in critical choices, such as financial resources, operations, and personnel). By examining the expert's supermarket's website, along with those of other supermarkets in the UK, I discovered that decision makers often belong to the Board of Directors, commonly referred to as "(company's name) House." Building on this, I will now delve into the aspects highlighted in the interview that seemingly depict the values associated with these individuals.

Please note, I did not include all the points from the interview, as it was 'anecdotal' and based on first-hand accounts. Instead, I used this information to guide further research into industry values. By verifying that these values are somewhat common within the industry, I ensured that the expert's anecdotal insights had external validity and were supported by real industry evidence, making the values more credible and plausible.

Here is a list of values that were identified from the expert's interview but not elaborated on due to both time constraints of the project and the reasons listed below, yet should not be ruled out of further work:

- Security and loss prevention : Although the expert discussed security and loss prevention measures in detail, the absence of comprehensive industry wide data on this specific value limited my ability to present compelling evidence within the project's time constraints. The expert mentioned that no thefts had been recorded over the past year, indicating a robust system in place, but without broader industry insights or comparative data, the full impact and effectiveness of these measures could not be thoroughly assessed, and there were more credible sources for other values.
- Effective inventory management : While not explicitly identified as a standalone value, effective inventory management is closely tied to broader values such as financial performance. From the expert's interview, it is evident that this value underpins various aspects of the supermarket's operations. For instance, the expert discussed the supermarket's approach to adjusting inventory based on seasonal demand and managing waste for damaged goods. This shows that effective inventory management is a fundamental element supporting and influencing other values discussed in this report, such as financial performance.

- Operational efficiency : From the expert's insights, it becomes apparent that operational efficiency plays a significant role in areas such as customer satisfaction and financial performance, thus I decided to show operational efficiency as an underlying factor to other key values, rather than a standalone value. For example, the expert states that during the Christmas season, increased staffing and extended operating hours were implemented to handle higher customer volumes and improve service quality. This demonstrates that operational efficiency is not a standalone value but an underlying factor that supports and enhances other critical aspects of supermarket operations, including customer satisfaction and financial outcomes.

That being said, below are the values I decided to focus on, with evidence:

1) Customer satisfaction

Firstly, in the interview with the expert, there was a strong indication that the supermarket prioritises delivering high quality customer service to enhance the shopping experience and foster customer loyalty, highlighted by the emphasis on regular training sessions every six months focused on customer service skills, in addition to feedback mechanisms, such as surveys and feedback forms, which are used to assess customer satisfaction and reward staff who excel in service, showcasing its value to the business. Moreover, the expert highlighted that customer satisfaction is an instance of a 'metric' for the supermarket, as well as the expert discussing how interaction with customers is deemed highly important, and this a companywide view. This focus on customer satisfaction is not only aimed at encouraging repeat visits but also at increasing overall customer expenditure.

To support the claims from the interview, I examined market leaders and their Key Performance Indicators (KPIs), as well as relevant articles. Currently, the only UK supermarkets with publicly available, up-to-date KPIs are Tesco, Sainsbury's, and Morrisons. Therefore, I also relied on additional articles to gather more insights throughout section 2.1. To begin, Figure 1 shows Tesco's "Big 6 KPIs", taken from Tesco's PLC Annual Report (2023). The one I have highlighted corresponds to the importance of customer satisfaction and how this then relates to customer loyalty. The evidence from Tesco's KPIs shows that supermarkets with high customer satisfaction tend to experience better sales growth and market share. Therefore, Tesco's focus on customer service can be seen as a strategic move to boost both customer loyalty and financial performance.

With Perplexity AI (Figure 2) I decided to see if I could find any more resources for customer satisfaction being a core value for supermarkets, within the UK.

Perplexity AI directed me to a LinkedIn article by Causon (2017) highlighting the importance of customer service in supermarkets. Quotes from this article include, "Supermarkets with a strong focus on customer experience (CX) saw their market share increase more than their rivals - and conversely the worst performing brands for CX saw their market share decrease," and "Supermarkets with a strong focus on CX saw their sales figures increase faster than their rivals." Additionally, an article by Retail Gazette explains that Aldi's rise in market share is due to its strong focus on customer experience.

Figure 1

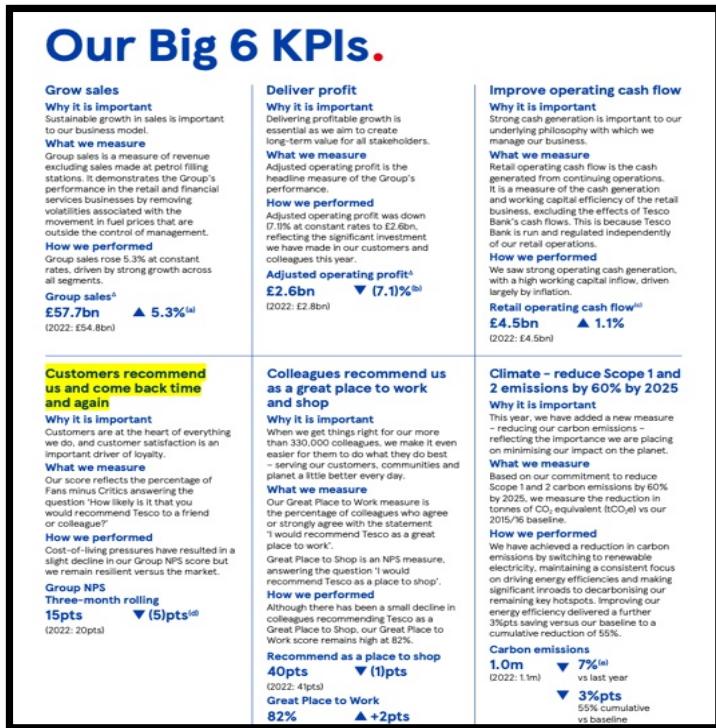
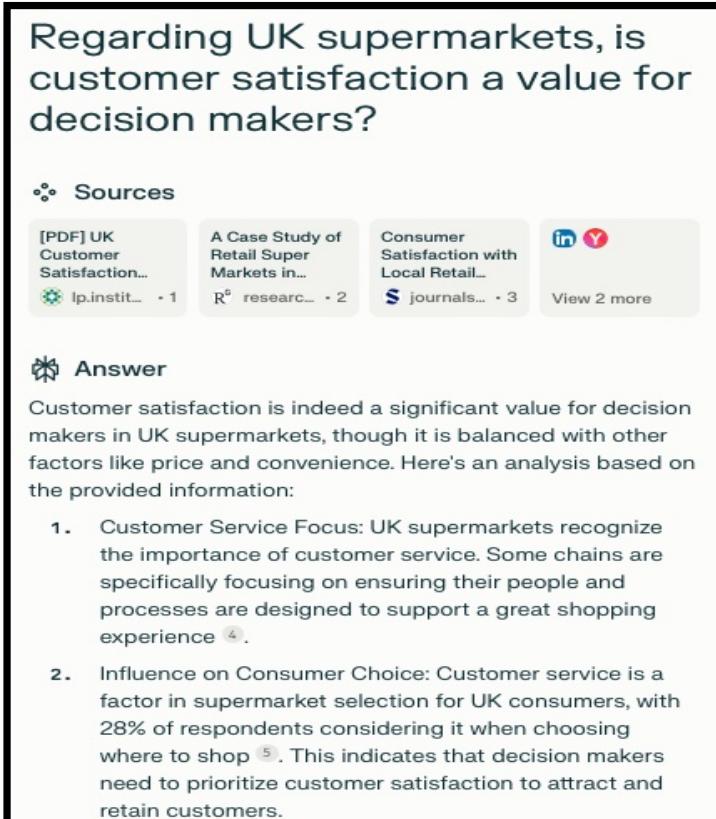


Figure 2



Causon (2017) states that supermarkets with low customer service levels, such as Tesco, Asda, and Co-Op Food, saw drops in market share, while those with satisfied customers enjoyed a sales growth of 10.7%, compared to 1.8% for others. High customer satisfaction consistently correlates with better sales and market share performance. Ignoring customer service can lead to significant financial losses, as shown by the clear link between customer satisfaction and business success.

These insights show that customer satisfaction, which directly relates to customer experience, could well be a core value for decision makers in the supermarket industry, as a strong focus on “customer experience” directly correlates with increased market share and sales, which is something that supermarkets also value and will be discussed in more detail later.

Finally, I decided to ask ChatGPT about The UK Customer Satisfaction Index (UKCSI) as I was unable to gain access to it myself, yet I was still curious of the content (Figure 3). I found that the trends detailed within this result, such as the reference to employee engagement, closely aligns with the expert’s statement regarding customer interactions are an important factor to the supermarket.

Figure 3

The UK Customer Satisfaction Index (UKCSI) typically uncovers several key trends in customer service, which can vary from year to year. While specific trends may change based on the latest data, here are some common trends that the UKCSI often reveals:

1. **Rising Expectations:** Customers' expectations for service quality and responsiveness tend to increase over time. Businesses must continually adapt to meet these higher expectations to maintain satisfaction.
2. **Impact of Digital Transformation:** The growing importance of digital channels and technologies often appears in the data. Customers expect seamless and efficient digital interactions, and companies that excel in this area generally score higher in satisfaction.
3. **Personalization:** There is a strong trend toward the need for personalized service. Customers appreciate when businesses tailor their interactions and offerings to their specific needs and preferences.
4. **Customer Service Consistency:** Consistency across different touchpoints and interactions is increasingly important. Customers expect a uniform level of service regardless of how or when they interact with a company.
5. **Service Recovery:** The ability of a company to effectively resolve issues and complaints often impacts overall satisfaction. Businesses that handle problems well tend to have higher customer satisfaction scores.
6. **Employee Engagement:** The satisfaction of employees can influence customer satisfaction. Engaged and motivated employees are more likely to provide better service, which positively affects customer perceptions.
7. **Sector-Specific Changes:** Different sectors may experience unique trends. For example, in the retail sector, trends might include the growth of omnichannel experiences, while in the financial sector, there could be a focus on trust and security.

For the most current and detailed trends, you would need to refer to the latest UKCSI report available on the Institute of Customer Service's website or similar publications. These reports will provide the most up-to-date insights into customer service trends based on recent data.

High customer satisfaction is crucial for supermarkets to stand out in a competitive market. Great service keeps customers coming back and leads to positive word-of-mouth referrals. This loyalty provides steady revenue and helps supermarkets shine compared to those with poorer service. Investing in customer satisfaction also builds a strong brand reputation, making customers more likely to continue shopping at the supermarket even if prices go up or during tough economic times. In short, focusing on customer satisfaction fosters lasting relationships and supports ongoing growth.

Reflecting on all the information above, it is reasonable to conclude that customer satisfaction is a core value for decision-makers within the expert's supermarket, as it is supported by both anecdotal evidence and industry-wide standards.

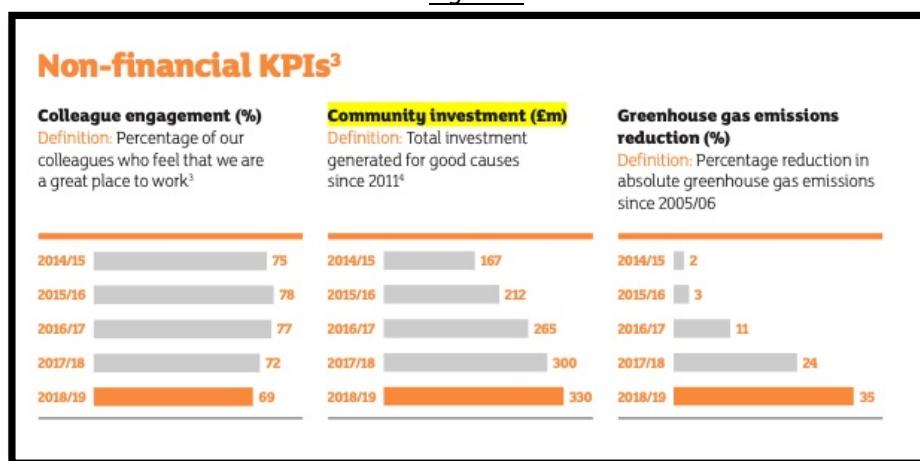
2) Community engagement

As a preface, while researching the company's website, I found that they have implemented additional community engagement initiatives that the expert had not observed. Although I prefer not to discuss these specifics to keep the expert's supermarket anonymous, I believe it is worth mentioning at this point. Additionally, it is important to note that the expert only touched on this value to a limited extent. Consequently, I believe this value is not specifically tracked or prioritised within the expert's supermarket. However, it is a value that decision makers should consider based the chain reaction mentioned later and the current market leader's attention to this value.

Within the interview, the expert highlighted that, over the past eight months, their supermarket has increasingly engaged with the local community. They have sponsored local football teams, set up a donation trolley for food banks, and organised a toy drive with a matching donation scheme. They also believe that community engagement is a strategic move, given that other large supermarkets are adopting similar practices.

Figure 4 illustrates Sainsbury's Non-Financial KPIs (2019), highlighting community engagement as measured by financial investment in local initiatives, dating back to 2011. This demonstrates the significance of community support for major supermarkets and highlights how long this engagement has been in place.

Figure 4



Fast-forward to 2023, Sainsburys still considers community engagement to be a factor worth sharing within their annual report, highlighting the importance of it to the supermarket (Figure 5).

Figure 5

Community and partnerships

Sainsbury's has a presence in communities across the country and with the help of our colleagues and customers we support our communities through fundraising, volunteering and raising awareness. This year alone we raised a total of £34.5 million for good causes.

Our primary focus is to tackle food poverty. In response to the rising cost of living we launched Nourish the Nation with our longstanding charity partner Comic Relief. This long-term community initiative provides funding to tackle food insecurity and ensure communities have access to balanced, nutritional and sustainable food sources now and in the future. Through our Nourish the Nation community programme we have raised £7.2 million to support Comic Relief, our food redistribution partners, such as FareShare and The Felix Project, and local community groups to help mitigate the impact of the rising cost of living. In addition, through our customer food donation scheme, 97 per cent of our stores are supporting 524 number of local community groups with essential grocery items.

£34.5m
raised for good causes this year

4. Discrimination. We will improve the position of the most vulnerable workers in our priority value chains

5. Grievance mechanisms. We will ensure access to effective and trusted grievance mechanisms across all Tier 1 suppliers

To live up to our vision of being the most trusted retailer it is critical that there is complete transparency in our value chain. Having previously published our Tier 1 clothing and food sites, this year we also published our Tier 1 general merchandise sites with information such as addresses, number of workers, gender split and union membership. We will be publishing a list of our Sainsbury's branded Goods Not Resale sites in the year ahead, as well as working to achieve greater visibility of the challenges faced further down the value chain.

We updated and published our whistleblowing and ethical trade policies. These define our position and expectations of suppliers on ethical sourcing practices and access to safe and independent means of raising concerns.

For more information on our approach to social sustainability and to read our Human Rights Policy, Ethical Sourcing Policy and Modern Slavery Statement, please visit www.about.sainsburys.co.uk.

Moreover, when considering Morrison's annual report (2022), they also strongly highlight their value of community support and engagement (Figure 6).

Figure 6

Responding to local needs

At the heart of Morrisons community work is our team of Community Champions who play a vital role responding to local needs and supporting good causes. We allocate hours for them to work directly with local organisations and key stakeholders to ensure we are playing our full part in the communities we serve.

As part of our strategy to become locally integrated, our Community Champions continued their extensive listening programme to better understand local priorities and needs. The information was used to ensure their Community Plans remained locally relevant.

Tackling food poverty

We know lots of people in our communities experience food poverty and this is being exacerbated by the cost of living crisis. As a food maker and shopkeeper we see this as an area on which we should focus our support.

The school holidays can be a particular pinch point for families in hardship and we want to help community partners relieve some of the pressure. Over the summer, we donated thousands of food products and essential items to local organisations and clubs delivering support for low income families over the school holidays. Community Champions partnered with local schools and charities, as well as local authority activities organised as part of the UK Government's Holiday Activities and Food ('HAF') Programme.

Our Pick Up Pack scheme continues to resonate with customers, allowing them to add a pre-packed donation bag to their shopping. Our Community Champions work closely with local food banks and community groups to understand the items that they need the most. They then make dedicated packs of these items, and place them at the front of the store for customers to pick up on their way in. The scheme generated over £4 million worth of donations in the financial period.

Morrisons Foundation

In the financial period ended 30 October 2022, the Morrisons Foundation donated nearly £3 million in grants (over £3 million including match funding) to registered charities across England, Scotland and Wales. Since its launch, it has donated over £35 million in grants (over £38 million including match funding) and supported over 3,100 charities which are making a positive difference in local communities.

During the financial period, the Morrisons Foundation placed particular focus on grant funding to organisations and projects that are helping to tackle food insecurity such as food banks and community kitchens.

Supporting charities

Our colleagues, customers and suppliers raised over £1 million in the financial period for our previous national charity partner Young Lives vs Cancer. Our partnership ended in February 2022 and we raised a total of over £18 million since it began in February 2017. Our new charity partner for the next three years is Together for Short Lives. Our partnership will raise vital funds for children's hospices across the country. Over £2 million has been raised in the financial period since the partnership launched in February 2022.

As a strategic partner of the Poppy Appeal we were pleased to welcome volunteers back into our stores in November 2021. Alongside implementing systems to collect donations at our checkouts and online, many of our store colleagues also volunteered to fundraise and together we raised nearly £2 million. We also raised over £670,000 for Marie Curie's Great Daffodil Appeal in March 2022 and were the headline partner for the NHS Big Tea in July 2022, raising over £130,000 for NHS Charities Together.

Figures 4, 5, and 6 illustrate that major supermarkets like Sainsbury's and Morrisons invest significantly in community engagement. This financial commitment is not merely for public relations but is strategically aimed at building a positive image and fostering local loyalty. The integration of community investment with business results highlights a strategic effort to improve brand perception and boost market performance.

Engaging with the community boosts a supermarket's image as caring and responsible, which can increase customer trust and loyalty. By participating in local events, supporting causes, and addressing community needs, the supermarket can stand out from competitors and build a positive brand reputation. Additionally, being active in the community offers valuable insights into local preferences, helping decision-makers adjust their offerings and strategies to better serve their customers.

With these reasons in mind, it is fair to say that community engagement is a significant value for supermarkets, and thus corresponds with what the expert states within their interview when discussing the implementation of successful strategies.

3) Supermarket image

Building on the topic of community engagement, it is worth discussing its impact on supermarket image. The expert mentioned that maintaining a positive reputation is a priority for the supermarket's headquarters, who are the key decision makers, evidenced by the new community wall in the store that showcases their achievements.

While community engagement and business image are distinct concepts, they are interrelated. Effective community engagement can improve the company's image by presenting the supermarket as a responsible, community-focused organisation.

Research by Shamsher (2021) shows that a store's image affects customer loyalty through customer satisfaction and trust. A good store image improves customer satisfaction, which helps build trust. This trust then leads to stronger customer loyalty. In short, a positive store image makes customers happier, more trusting, and more loyal.

Using Perplexity AI, I was able to find some suggestions of articles as to where to look for this connection between these two concepts (Figure 7).

A study examining the impact of community involvement on brand perception found that there was a positive correlation between the two concepts (Chen et al. 2017). Specifically, consumers who view a brand as actively engaged in their community are more loyal, which in turn enhances marketing performance. Strong community engagement also boosts brand identification and valuation, leading to improved marketing outcomes. Essentially, the benefits of community involvement are mediated through increased brand loyalty and perceived value, which contribute to better market performance.

Essentially, a positive supermarket image helps attract new customers, keep existing ones loyal, and set the supermarket apart from competitors. It affects how customers view the supermarket and can influence their buying choices, including their willingness to pay more

or travel further. It also builds a strong 'brand' reputation, boosting the supermarket's market position and supporting long-term growth.

Figure 7

How does community engagement influence brand image

❖ Sources

[PDF] The Effects of Community Involvement, Brand Identity, and Bran...
ijrbsm • 1

The Influence of Online Community Experience on Brand Image and ...
atlantis-press • 2

The roles of brand community and community engagemen...
sciedirect • 3

R^E E
View 2 more

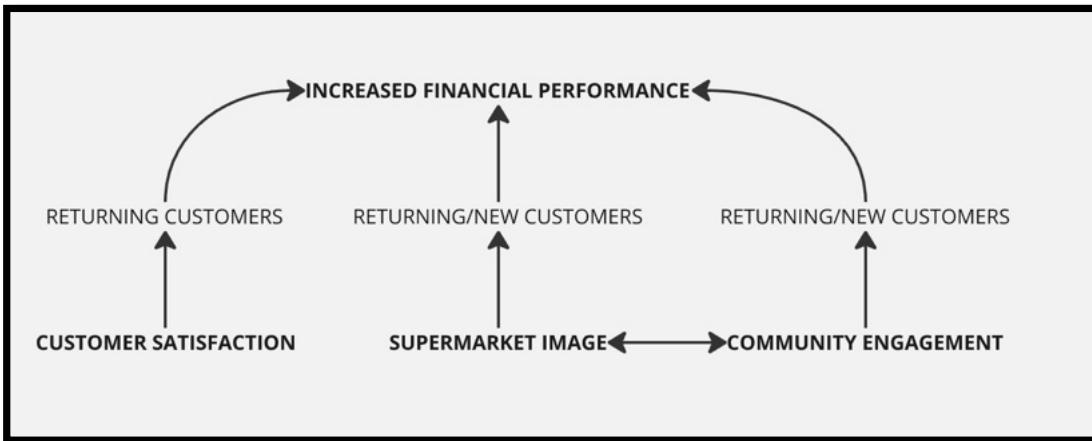
Following the research above, there is now a common belief that community engagement and image are both essential values for a business, but they are separate concepts. Community engagement focuses on supporting and participating in the local community, while image deals with the overall perception and reputation of the company. Although they both influence the company's identity and success, each plays a different role in shaping how the company is viewed and interacts with its stakeholders, hence supermarket image being deemed as a core value for the decision makers.

4) Financial performance

During the interview, the expert consistently linked the three core values, discussed above, to financial performance. They explained that customer satisfaction leads to returning customers and increased spending, which boosts revenue and profit. When discussing community engagement, they noted that it enhances the supermarket's image, attracting more customers and fostering loyalty. They also mentioned that adopting these principles is important because market leaders in the sector emphasise community and image, and the reason these other supermarkets are considered market leaders is due to their strong customer base and consistent financial success. In figure 8, I have drawn a diagram to illustrate this flow and 'chain reaction'.

The common denominator in figure 8 is the impact of customers on sales, which ties all the core values to one main value: financial performance. It is worth noting that having 'returning customers' or 'new customers' is not considered a core value for the supermarket. Instead, these outcomes are byproducts of other core values and the link between them and financial performance and would not be core values by themselves as they depend on the foundation provided by the other core values.

Figure 8



Furthermore, each KPI discussed by the expert, relating to the KPI's used within their supermarket, is closely tied to financial performance. Monitoring the amount of damaged goods, including returns and expired items, relates to financial performance by indicating the total sales revenue lost due to unsellable products. This KPI aligns with the core value of operational efficiency, as reducing damaged goods minimises losses and maximises profitability. Customer satisfaction, measured through feedback forms and online reviews, is directly linked to financial performance by driving returning visits and increased expenditure, which boost profits. Analysing unit sales informs the supermarket about product performance and inventory turnover, hoping to prevent overstocking and unneeded markdowns, thus maintaining optimal profit margins. Lastly, monitoring theft through inventory audits relates to financial performance by reducing losses due to stolen goods. With these KPIs in mind, it is clear that financial performance is something decision makers highly value.

In Figure 9, Sainsbury's KPIs for group measures (2019) are shown. At a glance, it is easy to see that these KPIs are all focused on finance and profits, such as examining margins and other financial metrics. This emphasis clearly highlights the importance of financial performance for Sainsburys.

Again, figure 10 demonstrates Sainsbury's KPIs for 2023. The one thing all these KPIs have in common is their relation to financial performance. Retail sales growth, at 5.2% including fuel, indicates increased total revenue, which boosts financial performance. The statutory profit before tax of £327 million highlights profitability before tax, though a decrease from previous years suggests potential declines in profitability. The underlying basic earnings per share (EPS) of 23.0p measures profit per share, showing improved earnings from 2019/20 but a recent decline. The underlying profit before tax of £690 million, excluding non-recurring items, provides a clearer view of ongoing profitability, despite a recent decrease. Retail operating profit of £926 million represents profit from core operations, with recent declines indicating reduced profitability. The return on capital employed (ROCE) of 7.6% measures efficiency in generating profit from capital, showing a decrease from the previous year but improvement from 2019/20.

Figure 9

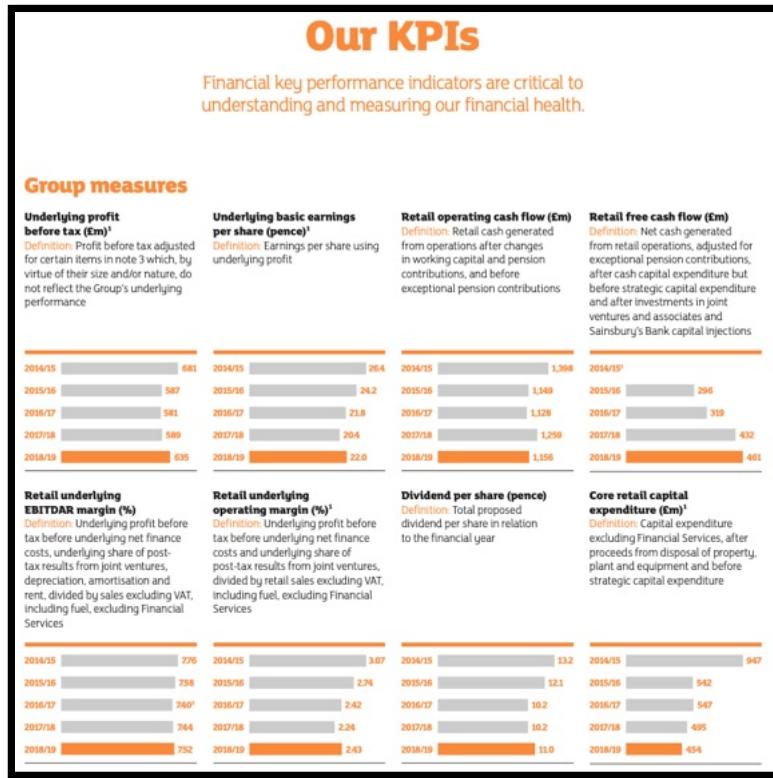


Figure 10



Figure 11 shows the “Big 6” KPI’s for Tesco (2023). Again, three out of six of these KPIs are directly related to financial performance of the supermarket.

Figure 11



In summary, supermarkets that successfully connect customer satisfaction, community engagement, and brand image to financial results can achieve greater profitability and growth. By continuously optimising these areas, they can maintain strong financial health and stay ahead of competitors. Monitoring financial performance through various KPIs ensures alignment with strategic goals, enabling effective management. Insights from customer satisfaction and operational efficiencies help supermarkets invest in future growth initiatives, such as expanding store locations, enhancing technology, or exploring new market segments.

With all these KPIs aligning across different supermarkets, focusing on profitability, margins, revenue, and other financial metrics, in addition to the expert’s interview points and the supermarket’s KPIs, it is clear that financial performance is a primary value. Financial performance is essential for a supermarket because it demonstrates stability to investors, helping secure future funding. For decision makers, tracking financial performance helps evaluate strategies, cut costs, and make decisions that drive long-term success.

Please note that although it may seem that financial performance is the primary core value for decision makers, this does not imply that other core values are any less important. Each value has its own successes and should be evaluated on an equal footing. Ignoring this could lead to missed insights and overlooked opportunities. Recognising the interplay between values can reveal chain reactions and benefits that drive further improvements in the supermarket.

2. Development of Metrics

For the development of metrics, I decided to focus on outcome metrics which measure the results and impacts of our processes. I avoided process metrics because, despite being easier to measure, they do not fully reflect the success. While outcome metrics are more

challenging to assess, they provide a clearer understanding of overall performance and goal achievement, thus prioritising the effectiveness of actions over efficiency. In other words, measuring the outcomes shows if the process used to produce the outcome is beneficial or detrimental to a business.

Value : Customer satisfaction

Average response time to customer inquiries

Firstly, we could track the average time taken to respond to customer inquiries or complaints. This could be more difficult to do in store, but could gain data from feedbacks forms, asking the customer how long the response took and depict if this matches with the time recorded by the employee.

I desire to focus on outcome metrics, so recording response times during the interaction, or shortly after, may not be accurate, as the customer could deem that the issue has not actually been 'resolved'. For email or telephone inquiries, these should also be tracked using a dedicated system that logs the time taken to respond and marks the completion of each 'case'. This way, we can accurately measure the response time and ensure that issues are fully resolved.

A potentially useful formula I partially created is: Average response time = Total time taken to response to enquiries (hours or minutes – decision maker preference) / Total number of enquiries (the division gives the 'average')

Quick response times are a key element for maintaining customer satisfaction, especially when addressing issues or complaints. Faster response times typically lead to higher customer satisfaction, as customers feel their concerns are addressed promptly and efficiently. By evaluating this metric, the supermarket can assess their responsiveness and identify any inefficiencies in their customer service. This metric also helps to benchmark the supermarket's performance against wider industry standards if the metric were to be compared against other supermarket's response times, and pinpoint areas needing improvement, ensuring that customer inquiries are managed effectively. To do this aspect of normalisation, you could divide the average response time of the supermarket by the industry average to get a percentage difference, thus allowing to signify areas for improvement. If the supermarket did not want to optimise for customer satisfaction, but just simply be closer or alongside market leaders, this metric would give a good indication on their current position.

Data regarding customer response wait times can be found in mystery shopper reports. For instance, a mystery shopping report by FM Outsource (2022) highlighted that the average telephone response time for UK supermarket brands is notably slower than industry standards, averaging 5 minutes. These mystery shopper reports also include some data relating to in store inquiries, such as the length time it takes for a store associate to respond to a mystery shopper's question, the relativity and effectiveness of the answer, or the length of time the mystery shopper waits in line at the checkout. Regardless, all of these insights help to draw conclusions around response time, and can help to categorise the

sections of the supermarket with their associated response time. Likewise, we could actually gather this data ourselves if needs be, due to their straight-forward, real-life nature.

When asking Perplexity AI, the following image captures the response to the prompt, "Average response time to customer inquiries - how could I attain a dataset/data relating to this? Or which parts of a dataset would allude to this?"

 Answer

To attain a dataset or data relating to the average response time to customer inquiries, you can focus on several sources and key metrics. Here's how you can proceed:

Potential Data Sources

1. **Customer Service Platforms:** Platforms like Zendesk, Metabase, and Pexly provide tools and dashboards to track response times. These platforms often offer built-in analytics that can give you insights into average response times across different communication channels [1](#) [4](#) [5](#).
2. **Market Research Reports:** Research firms and platforms like Statista often publish reports on customer service metrics, including response times. These reports can provide benchmark data and industry standards [3](#).
3. **Internal Company Data:** If you have access to a company's customer service data, you can extract response times from customer service logs, ticketing systems, or CRM software.

Relevant Data Points in a Dataset

To analyze average response times, you should look for the following parts of a dataset:

- **Timestamp of Inquiry:** The exact time when a customer inquiry was received.
- **Timestamp of First Response:** The exact time when the first response was sent to the customer.
- **Response Time Calculation:** The difference between the inquiry timestamp and the first response timestamp.
- **Communication Channel:** The medium through which the inquiry was made (e.g., email, phone, live chat, social media).
- **Customer Service Agent ID:** Identifier for the agent who responded to the inquiry.

Product return rate

To monitor the product return rate, we should track the percentage of products returned by customers compared to the total number of products sold.

Tracking this metric helps identify potential issues with product quality or customer expectations. For accuracy, we should collect data on returns from various sources, including in-store returns and online purchases, and categorise these returns by product type or category, for example, fresh produce, bulk-buys, products on promotion. This will help pinpoint specific products or categories with higher return rates.

It is worth mentioning at this point that tracking this metric with regards to a time frame could be useful for the supermarket. By reviewing the return rate each 'month', you can identify trends and patterns, such as higher returns during specific times of the year, following promotions, or in response to new product launches.

High return rates can point to issues like faulty products, inaccurate descriptions, or, more importantly, unmet customer expectations. Analysing why products are returned indicates why customers were unsatisfied, thus helping identify common problems and make improvements to reduce future returns and improve customer satisfaction. By addressing high return rates and understanding their causes, the supermarket can enhance product quality, align offerings with customer expectations, and improve overall customer experience.

Regarding the collection of data, POS systems within most major supermarkets have transaction records and return receipts held within them. Due to the role of technology now in supermarkets, this becomes easier to attain, and also allows for ‘filtering’ such as suing an export via CSV, etc.

Data on return rates and industry-wide statistics within supermarkets can be found through the Office for National Statistics (2024). ONS has an entire report published or updated each year regarding the ‘Retail Sales Index’ which encompasses comprehensive data on retail sales, including statistics that infer product return rates.

When asking Perplexity AI, the following image captures the response to the prompt, “Product return rate metrics for supermarkets, how could I attain a dataset/data relating to this? Or which parts of a dataset would allude to this?”

Sources of Data

1. **Point of Sale (POS) Systems:**
 - Most POS systems track sales and returns, providing data on the number of items returned, reasons for returns, and specific products frequently returned.
2. **Returns Management Systems:**
 - Specialized software designed to handle returns can offer detailed data on return rates, reasons for returns, and related metrics.
3. **Inventory Management Systems:**
 - These systems track product movement, including returns, and can provide insights into which products are returned most often.
4. **Customer Relationship Management (CRM) Systems:**
 - CRM systems can give insights into customer behavior related to returns, including frequency and reasons for returns.
5. **Financial Reports:**
 - Look for sections in financial reports that discuss returns or adjustments in sales figures due to returns.
6. **Customer Satisfaction Surveys:**
 - Surveys can provide indirect data on returns by highlighting customer satisfaction and potential issues with products.

Key Data Points to Focus On

1. **Total Number of Products Sold:**
 - This helps in understanding the scale of operations and provides context for return data.
2. **Number of Products Returned:**
 - This is the primary data point for understanding return rates.
3. **Reasons for Returns:**
 - Categorizing reasons for returns (e.g., damaged goods, expired products, customer dissatisfaction) can help identify problem areas.
4. **Product Categories with Highest Return Rates:**
 - Identifying which categories (e.g., perishable vs. non-perishable) have higher return rates can help in inventory and quality control.
5. **Seasonal Variations in Return Rates:**
 - Understanding how return rates fluctuate with seasons can help in planning and inventory management.
6. **Customer Demographics Associated with Returns:**
 - Analyzing which customer segments are more likely to return products can provide insights for targeted interventions.

Additional Tips

- **Industry Reports:** Consider accessing industry reports and studies from market research firms or retail associations. These often contain aggregated data and insights on return rates.
- **Benchmarking:** Compare your data against industry benchmarks to understand how your return rates stack up against competitors.

By focusing on these sources and data points, you can gather a comprehensive understanding of product return rates in supermarkets without needing to perform calculations.

Trolley disengagement rate

To assess the trolley disengagement rate, we need to monitor the percentage of trolleys that are initiated, yet not purchased, and abandoned within the supermarket. This metric is good for understanding customer satisfaction, as high disengagement rates can indicate issues in the checkout process or concerns about the purchasing experience, thus relating to customer satisfaction. By tracking this rate as a percentage, we can easily compare abandonment rates across different periods and identify trends or areas for improvement, hence the multiplication of 100.

A potentially useful formula I partially created is: Trolley disengagement rate = $\frac{\text{Number of abandoned trolleys (with items - decision makers preference)}}{\text{Number of trolleys taken from trolley bay (with items - decision makers preference)}} \times 100$

You can also perform this metric with online shopping, regarding online trolleys that are not checked-out within a 30-day period. This may get more difficult due to discrepancies with what a suitable time period would be considered and what decision makers determine as 'disengaged' when online shopping. However, it is not impossible, but would just need more insight from the decision maker themselves, regarding their preferences for boundaries.

High trolley disengagement rates in store often indicate issues with customer satisfaction, such as long checkout lines, confusing store layouts, or poor customer service. These rates can also reflect frustrations with out-of-stock items or inconvenient store policies. Identifying and addressing these issues is key for improving the overall shopping experience, alongside gauging and improving customer satisfaction.

The trolley disengagement rate is not commonly tracked in UK supermarkets, so implementing this metric would require direct communication with decision makers rather than relying on publicly available records. To effectively monitor this rate, you would need to integrate specific data collection processes within the store and collaborate with store managers and other stakeholders to ensure accurate tracking. This could be implemented by having staff complete a form when they encounter an abandoned trolley, whilst also noting the location (aisle) this occurred, and time.

When asking Perplexity AI the following image captures the response to the prompt, "How could I attain a dataset/data relating to trolley disengagement within supermarkets? Or which parts of a dataset would allude to this?"

 **Answer**

To attain a dataset or data relating to supermarket trolley disengagement metrics, you would need to focus on several key areas that can provide insights into trolley usage and disengagement patterns. Here are some suggestions on where to find such data and what parts of a dataset would be relevant:

Potential Data Sources

- Supermarket Operations Data:** Supermarkets often collect data on trolley usage, including the number of trolleys in use, trolleys left in parking lots, and the time spent collecting trolleys. This data can be obtained directly from supermarket chains or through industry reports.
- Technology Solutions Providers:** Companies like McCue, which offer solutions like CartMagnet to reduce stray trolleys and associated costs, might have relevant data and case studies on trolley disengagement metrics ³.
- Market Research Reports:** Reports on the supermarket trolley market or related segments might include data on trolley usage and disengagement. These can be found through market research firms or industry publications ⁴.

Relevant Data Points

To analyze trolley disengagement, you should look for the following parts of a dataset:

- **Trolley Usage Statistics:** Number of trolleys used per day, peak usage times, and average usage duration.
- **Trolley Collection Times:** Time spent by employees collecting trolleys from parking lots and other areas.
- **Trolley Return Rates:** Percentage of trolleys returned to designated areas versus those left in parking lots.
- **Customer Behavior Data:** Insights into customer habits regarding trolley usage and return, possibly gathered through surveys or observational studies.

It is important to note that there are some limitations regarding optimising for this proposed metric. This metric focuses solely on the abandonment of trolleys, potentially missing other factors that contribute to customer dissatisfaction, such as product quality or staff behaviour. Furthermore, trolley disengagement data can be influenced by temporary issues or outliers, leading to potential misunderstanding of broader customer satisfaction trends. Thus, it should be used in conjunction with direct feedback mechanisms or other customer satisfaction metrics to obtain a more accurate understanding of customer satisfaction.

Value : Community engagement

Local Media Coverage

In this metric, I aim to measure the extent of local media coverage (including newspapers, social media, and TV) that the supermarket receives for its community engagement efforts.

I have come up with no real ‘formula’ for this, yet it is a numeric metric. Simply count the total number of media mentions across various medias. This total count provides a numeric measure of how extensively the supermarket’s community initiatives are covered in the media. To further this, you could also categorise the coverage, using headings such as ‘positive’, ‘negative’, and ‘neutral’, to get a better grasp of content within the coverage. This normalisation via categorisation helps in understanding the nature of media coverage.

Increased visibility through media coverage reflects the supermarket's successful community involvement, as it suggests that these activities are attracting attention and generating interest. This metric helps assess the effectiveness of community engagement in enhancing the supermarket's reputation and public image. By tracking media coverage, the supermarket can evaluate the reach and impact of its community efforts, gaining insights into how these activities contribute to business growth and customer engagement.

As for data, the University of Cambridge’s (2022) report on retail and community engagement offers a framework for evaluating community investments, though it doesn’t provide direct data. Local studies on supermarket engagement can provide detailed insights, similar to ones like Weaver (2013) which studied leadership and community engagement in supermarkets. For collecting data, tools like Google News can track mentions of community initiatives, while APIs from platforms such as Facebook can deliver data on public engagement as a numeric result. These sources help measure the impact of community engagement strategies on public perception and business outcomes.

When asking Perplexity AI the following image captures the response to the prompt, “How could I attain a dataset/data relating to trolley disengagement within supermarkets? Or which parts of a dataset would allude to this?”

Potential Data Sources

1. **Academic and Research Papers:** For example, the dataset presented in "A Dataset of U.S. Local News Articles for the Study of County-level..." includes over 1.4 million online news articles from 313 local U.S. news outlets ¹.
2. **Public Interest News Foundation:** Their Local News Map project provides detailed information on local media coverage for each Local Authority District in the UK, including reports on the best and worst covered districts ².
3. **Local Media Works Database:** This database offers detailed information on UK local news brands, including circulation breakdowns, web traffic data, and other relevant metrics for local newspapers and their companion websites ³.
4. **Journalist's Resource:** This platform provides a running list of useful datasets, tools, and resources for data journalism, which can help fuel stories on local media coverage ⁴.

Relevant Data Points in a Dataset

To analyze local media coverage, you should look for the following parts of a dataset:

- **Publication Details:** Information about the local news outlets, including names, addresses, contact information, and publication frequency.
- **Article Metadata:** Data on the number of articles published, their topics, and publication dates.
- **Circulation and Readership:** Audited circulation figures, readership statistics, both total and by area, and web traffic data.
- **Geographical Coverage:** Information on the geographical areas covered by the local news outlets.
- **Content Analysis:** Details on the types of content published, such as political reporting, local events, sports, and community news.
- **Engagement Metrics:** Data on reader engagement, including comments, shares, and likes on digital platforms.

Branded merchandise sales

Within the interview, the expert mentioned that the supermarket has sponsored local football teams. So, I decided that a potential way to measure community engagement would be to develop a metric that measures the sales of football merchandise branded with the supermarket's logo.

By comparing current merchandise sales to past figures, you can see if the sponsorship is increasing interest in the supermarket and building brand loyalty.

If sales of branded merchandise go up or stay neutral, it shows the sponsorship is likely having a desired effect. This boost or neutrality in sales not only helps increase profits (based on if the supermarket will generate commission from the sales or drive more customer attention) but also indicates that the supermarket is successfully connecting with the local community. In other words, strong community engagement, such as supporting local teams, helps raise the supermarket's profile. Additionally, dividing total merchandise sales by the number of stores selling these items helps in understanding the impact of in-store availability on sales. This ensures that sales data reflects actual engagement and not just availability, and thus normalising this metric slightly, depending on the decision maker's preference.

The expert did not mention this within the interview, but through further online research, the supermarket also sells the merchandise in store, leading to increased visibility and direct customer interaction. This in store availability makes it easier for customers to purchase the

branded items while shopping, potentially boosting sales even further. It also provides a connection between the supermarket's community involvement and its image, enhancing customer loyalty and reinforcing the supermarket's presence within the local community.

After consulting with ChatGPT for ideas of datasets and data collection, the main points that stood out were to seek existing sales data from the supermarket's marketing or sales departments and analyse historical sales figures.

Additionally, we could utilise internal reports such as inventory and sales reports from POS systems, and e-commerce sales data, in addition to reviewing industry reports and benchmarks on the impact of sports sponsorships on retail sales.

However, Focusing too heavily on merchandise sales could lead to prioritising sales tactics over more meaningful community initiatives. Additionally, external factors such as trends in merchandise popularity or seasonal changes can influence sales figures, potentially giving a skewed view of the supermarket's community engagement. Therefore, this metric should be used alongside other indicators of community engagement to provide a better understanding of the supermarket's level and effectiveness of community engagement.

To gather a dataset or obtain data relating to the sales of branded merchandise and assess the impact of community engagement efforts like sponsoring local football teams, you can follow these steps:

- 1. Contact the Supermarket Directly**
 - **Sales Data Request:** Reach out to the supermarket's marketing or sales department and request data on branded merchandise sales. Explain your purpose and how the data will be used to measure community engagement impact.
 - **Historical Sales Figures:** Ask for historical sales figures for similar merchandise before and after the sponsorship began. This helps you establish a baseline for comparison.
- 2. Utilize Internal Reports**
 - **Company Reports:** If you have access to internal reports or business analytics tools used by the supermarket, check for any available data on merchandise sales and sponsorship performance.
 - **Sales Tracking Systems:** If the supermarket uses a POS (Point of Sale) system or CRM (Customer Relationship Management) system, these may provide detailed sales data.
- 3. Analyze In-Store Data**
 - **Inventory and Sales Reports:** Check inventory and sales reports for branded merchandise sold in-store. This data might be accessible through inventory management systems.
 - **Foot Traffic and Sales Correlation:** Analyze foot traffic data (if available) and correlate it with branded merchandise sales to understand the impact of in-store availability.
- 4. Conduct Surveys and Customer Feedback**
 - **Customer Surveys:** Implement surveys to gauge customer awareness of the sponsorship and its impact on their purchasing decisions.
 - **Focus Groups:** Conduct focus groups with customers to gather qualitative data on their perception of the sponsorship and its influence on their buying behavior.

Value : Supermarket image

Net Promoter Score (NPS)

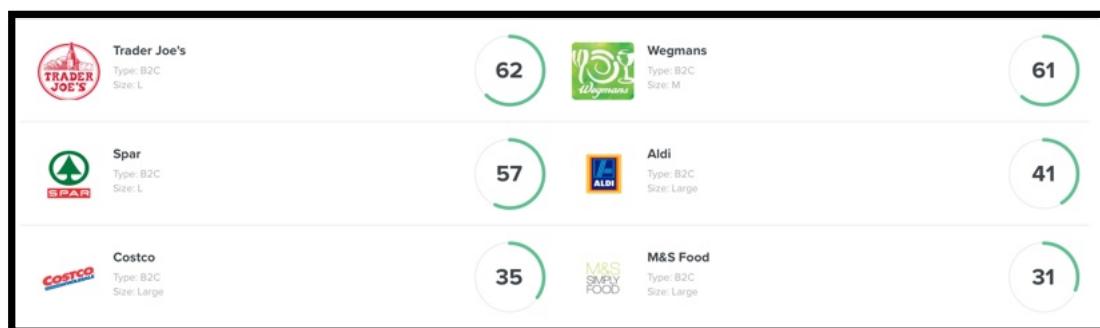
Net Promoter Score (NPS) assesses customer loyalty and supermarket image. It measures how likely customers are to recommend the supermarket to others, giving a clear picture of its reputation and customer satisfaction. A high NPS means customers are more likely to promote the supermarket, reflecting loyalty and a positive image.

An article by SurveyMonkey (2024) details how a business may calculate their NPS, and online tools like these help businesses do it automatically based on information they have collected. NPS measures how likely customers are to recommend a company or product. Customers rate their likelihood from 0 to 10. Those who rate 9 or 10 are called Promoters, 7 or 8 are Passives, and 0 to 6 are Detractors. To get the NPS, subtract the percentage of Detractors from the percentage of Promoters. The result shows how happy customers are with the company, and in our case, the supermarket. In addition, it helps compare the supermarket's performance with industry standards and competitors, giving a clear view of where the supermarket stands in the market. For example, divide the supermarket's gross profit margin by the industry average to determine how it performs relative to competitors.

Below is an image taken from Tesco's Annual Report (2023), demonstrating the importance of the NPS metric.

Key engagement metrics	Customers recommend us and come back time and again: our customer net promoter score (NPS) , which is measured based on customers recommending us as a place to shop.	Colleagues recommend us as a great place to work and shop; our Great Place to Work score, which is measured through our Every Voice Matters colleague engagement survey.	Our Supplier Viewpoint survey results continue to reflect our progress on building trusted relationships with our suppliers. Meeting our Scope 3 net zero commitments by 2050 and supporting suppliers in improving diversity within their businesses.	Drive top-line growth and grow absolute profits while maintaining sector leading margins. This is expected to generate between £1.4bn and £1.8bn of Retail free cash flow.
-------------------------------	--	--	--	--

As for the analysis of data, CustomerGauge (2022) collects and publishes publicly available NPS scores from supermarkets. According to their benchmarks, the average NPS in supermarkets is 41, thus showing a benchmark for supermarkets who value having a competitive edge. Below is an image taken from their website, showing the NPS scores of supermarkets.



Share of voice

Share of voice (SOV) measures how often a supermarket is mentioned in industry-specific media compared to its competitors. This metric helps gauge the supermarket's visibility and influence in the market. Essentially, SOV reflects how prominent and relevant the supermarket is within its industry.

SOV indirectly reflects a supermarket's image because it measures how often the supermarket is talked about in industry media. Although SOV doesn't directly assess the supermarket's image, frequent mentions usually indicate a strong market presence and relevance. If a supermarket has a high SOV, it often means it's receiving significant media attention, which can positively impact public perception. This visibility helps build a better supermarket image, as being frequently discussed suggests that the supermarket is influential in the industry. So, while SOV doesn't measure image directly, it serves as a good indicator of how well a supermarket is seen and talked about.

However, there is a drawback to this metric due to time constraints. After researching datasets and places to find data relating to supermarket SOV, as well as using Perplexity AI, I found that there are no available public records relating to supermarket SOV that I can find. However, I still want to include this metric, in the case that future investigation allows for this to become a data-backed metric. Below is some suggestions from perplexity AI on how to make this a data-backed metric.

To measure a supermarket's Share of Voice, you would need to create a custom dataset by collecting data from various sources, as there isn't a readily available public dataset specifically for supermarket SOV. Here's how you could approach this:

1. Media Monitoring:
 - Use media monitoring tools to track mentions of the supermarket and its competitors across various channels (TV, radio, print, online news, social media).
 - Collect data on the frequency and reach of these mentions.
2. Social Media Data:
 - Use social media analytics tools to gather data on mentions, hashtags, and engagement related to the supermarket and its competitors.
 - Platforms like Twitter, Facebook, and Instagram can provide valuable data.
3. Search Engine Data:
 - Use SEO tools to track search volume and rankings for keywords related to the supermarket and its competitors.
 - Google Trends can provide comparative search interest data.
4. Advertising Spend:
 - If available, collect data on advertising spend for the supermarket and its competitors across different channels.
5. Customer Surveys:
 - Conduct surveys to measure brand awareness and recall among consumers.

It is worth mentioning that if the supermarket optimises for SOV it can be very misleading, due to it being an indirect metric. SOV focuses on how often the supermarket is mentioned in the media rather than the quality or context of those mentions, thus allowing negative mentions to be accounted for and potentially deemed as a positive. Also, not all of the data collected around the SOV of the supermarket will be relevant to shaping supermarket image regarding customer and stakeholder perceptions.

Value : Financial performance

Gross Profit Margin

Gross profit margin shows how well a supermarket is doing financially by measuring the percentage of revenue left after covering the cost of goods sold. A higher margin means the supermarket is keeping more money from each sale, which suggests it's good at controlling production costs and setting prices. This metric helps us see how profitable the core operations are before considering other expenses. By tracking this margin over time, we can spot trends and see if cost management and pricing strategies are improving or need attention, directly relating to financial performance of the supermarket.

An article by Bloomenthal (2024) demonstrates the use and calculation of a gross profit margin. For a supermarket gross profit margin is calculated by subtracting the cost of goods sold from the total revenue, and then dividing that number by the total revenue. The result is then multiplied by 100 to get a percentage. For example, to keep it simple, if a supermarket earns £1,000 in revenue and its cost of goods sold is £700, the gross profit is £300. Dividing £300 by £1000 gives 0.3, or 30%. This means the supermarket's gross profit margin is 30%, showing what portion of revenue remains after covering the cost of goods sold.

To actually collect or access the data needed to comprise the metric, sales records from the supermarket's point-of-sale systems can provide detailed information on total revenue generated from transactions. Moreover, financial statements, such as income statements or profit and loss reports show revenue. Finally, inventory management systems track the product purchases which can be combined with the price of products to calculate the cost of goods sold.

In regards to datasets, there is a dataset by GTS (2024) which includes the Gross margin Percentage of supermarket sales across the industry. This can be used to calculate average margins across product categories and time periods. An image of the information contained is below.

This dataset provides a historical record of three months of sales data from three different supermarket branches, capturing competitive market trends in populous cities. It's structured to facilitate predictive analytics in retail.

Dataset Attributes:

- **Invoice ID:** Unique identifier for each sales transaction.
- **Branch:** Identifier for each of the three supermarket branches (A, B, and C).
- **City:** Supermarket locations.
- **Customer Type:** Distinguishes between 'Member' and 'Normal' customers.
- **Gender:** Customer gender.
- **Product Line:** Categories include Electronic Accessories, Fashion Accessories, Food and Beverages, Health and Beauty, Home and Lifestyle, Sports and Travel.
- **Unit Price:** Cost per item in dollars.
- **Quantity:** Items purchased per transaction.
- **Tax:** Applied sales tax (5%).
- **Total:** Total transaction cost, including tax.
- **Date:** Date of transaction, from January 2019 to March 2019.
- **Time:** Time of transaction, from 10 AM to 9 PM.
- **Payment Method:** Includes Cash, Credit Card, and Ewallet.
- **COGS:** Cost of goods sold.
- **Gross Margin Percentage:** Percentage of gross margin.
- **Gross Income:** Revenue generated from sales.

This structured data supports analysis of buying patterns, customer behavior, and financial performance across different locations and times.

Return on assets

Return on assets (ROA) is a metric for supermarkets to evaluate their operational efficiency and profitability. By consistently monitoring ROA, supermarket managers and stakeholders can compare performance across different stores or with competitors, assess the impact of strategic decisions on overall profitability, identify areas for operational improvements, and guide financial planning and resource allocation. A high ROA shows that the supermarket is using its assets effectively and has strong operations. A low ROA, on the other hand, might suggest that there's room to improve how assets are managed or that investment decisions need to be rethought.

To calculate ROA, you can gather data from several sources, including financial statements, which provide net income from the income statement and total assets from the balance sheet. Internal accounting systems offer detailed breakdowns of both income and asset values, while annual reports contain financial data for ROA calculation.

Return on Assets (ROA) is a useful measure for understanding a supermarket's financial performance because it shows how well the supermarket uses its assets to generate profit. By comparing the net income (profit) to the total assets (like inventory and equipment), ROA indicates how efficiently the supermarket is turning its resources into earnings. This metric makes it easy to see if the supermarket is using its assets effectively, compare performance with other supermarkets, and evaluate the impact of strategic decisions. Tracking ROA over time also helps identify whether improvements are needed in asset management to boost profitability.

I asked Perplexity AI to find some areas where I could find data relating to the calculation of ROA for certain supermarkets, based on the fact that this information is usually not contained within datasets for supermarkets. Below is the response.

Answer

Based on the search results, there isn't a specific dataset that directly provides Return on Assets (ROA) data for supermarkets. However, we can identify some relevant sources and approaches to compile this information:

1. Financial Databases:

While not mentioned in the search results, financial databases like Bloomberg, S&P Capital IQ, or Refinitiv often contain detailed financial information for public companies, including ROA for major supermarket chains.

2. Annual Reports:

For publicly traded supermarket companies, their annual reports typically contain the necessary information to calculate ROA. You could compile this data manually from multiple years and companies to create your own dataset.

3. Industry Statistics:

The CSIMarket website (result ²) provides some industry-level statistics for grocery stores, including ROA. For Q2 2024, they report:

- Grocery Stores Industry average ROA: 4.59%
- ROA range: High of 15.23% (Q2 2016) to Low of 2.09% (Q3 2015)

3. Analysis of Causal Factors

Value : Customer satisfaction

Metric : Average response time to customer inquiries

Causal factor : Employee workload

Cashiers, primarily stationed in one area, have limited capacity to assist with inquiries beyond their immediate vicinity. In contrast, floor staff, who are mobile, are better positioned to handle inquiries but are often have significant stocking responsibilities. This workload can delay their response to customer inquiries.

Analysing and optimising staff schedules to align with times of high inquiry volume could improve response times, and thus higher customer satisfaction, due to higher response times contributing to a better shopping experience. However, this requires a thorough analysis of inquiry patterns, which could lead to increased labour costs.

For instance, as the expert stated in their interview, inquiries can range from a few minutes to over 15 minutes, especially if the customer needs detailed information. By identifying peak inquiry times and ensuring adequate staffing during these periods, we can enhance response times and, consequently, customer satisfaction.

For example, if we identify that inquiries peak during certain hours or days, we can allocate more staff during these times, unlike what the expert stated where all morning staff are in the supermarket two hours before it opens, to help with restocking, yet it only takes around an hour to do. This could lead to faster response times and higher customer satisfaction.

However, this must be balanced against potential increases in labour costs. Also, increasing staff during peak times could lead to higher labour costs. Furthermore, overstaffing during non-peak times may lead to underutilisation of employees, as shown by how the expert describes the time it takes to complete the morning routine, in comparison to the time dedicated to it, leaving many staff with a significant amount of free time (around 1 hour). Finally, staff who's schedule mainly align with peak inquiring times may suffer from burnout very quickly if their tasks are not changed in relation to those peak times.

Causal factor : Knowledge of staff

The knowledge level of staff significantly impact the average response time to customer inquiries. Employees who possess good knowledge about the store's products and services can provide quicker and more accurate responses to customer inquiries. Thus, employees who are less knowledgeable may take longer to find the necessary information or may need to seek assistance from more experienced colleagues, therefore increasing response times.

Providing regular and thorough training sessions can ensure that staff are well-equipped to handle a wide range of customer inquiries efficiently. For instance, introducing a bi-weekly

training session focused on different product categories can keep staff informed about new products and promotions, since the expert did mention in their interview about the lack of knowledge of products staff have concerning different areas of the store, making customer interactions go for as long as 15 minutes, in addition to different departments not communicating about their products and promotions effectively.

The primary trade-off is between the time and resources spent on training and the potential improvement in response times. However, the long-term benefits of having a knowledgeable staff may outweigh these costs, as it can lead to higher customer satisfaction and potentially increased sales.

As an additional note, confounding variables such as the complexity of customer inquiries or complaints can impact the response time, regardless of how much knowledge staff may have.

Metric : Product return rate

Causal factor : Product quality and defects

The quality of products and the occurrence of defects significantly impact the product return rate. High quality products with minimal defects lead to fewer returns and higher customer satisfaction, whereas poor quality and defective products result in higher return rates and dissatisfaction.

Ensuring good quality control inspection before products are displayed on shelves can reduce the incidence of defective products. This, in turn, minimises returns and enhances customer satisfaction. Implementing more quality control measures at the supermarket level can help reduce the return rate.

For instance, conducting random product inspections can ensure that only products meeting high standards are sold. While these measures can have additional costs, the reduction in returns and increased customer satisfaction can outweigh these expenses by building customer trust and loyalty.

Furthermore, focusing on product defects and quality brings the risk of overlooking supplier or delivery complications. While supermarkets are responsible for any damage or quality issues that occur on their premises, they are not accountable for issues caused by suppliers or delivery companies. However, these external factors can significantly affect product quality, such as items being damaged during transit.

However, by investigating the root causes of product quality and defects in their store, supermarkets can identify the need to change suppliers or improve delivery processes. This proactive approach can lead to a positive change in product return rates, ultimately enhancing customer satisfaction and reducing losses.

As a first-hand account regarding product returns is that the expert mentions in their interview that product returns at the supermarket have been rising consistently over the

past few years, despite the return policy remaining unchanged, and also states that the reason for this trend is unclear. Finding a causal relationship for this implication is potentially a major win.

Causal factor : Competitor pricing

In the first chapter, I explored the supermarket's unquestioned assumption that increased customer satisfaction primarily stems from welcoming and friendly engagement. However, this satisfaction can be attributed to various factors, such as pricing, as seen in discount supermarkets.

To maintain customer satisfaction, the supermarket should also focus on competitive pricing. Customers who find similar or identical products at lower prices in other supermarkets are more likely to return products and shop elsewhere. Thus, the supermarket should consider the significant impact of pricing on customer satisfaction and product return rates.

To improve customer satisfaction and reduce product return rates, the supermarket could adopt more competitive pricing strategies. This may involve price matching, offering frequent sales or discounts, and implementing more personalised loyalty programs that provide value that other supermarkets may not. However, while aggressive pricing strategies can lower the return rate directly, it can also close profit margins.

Causal factor : Product descriptions and labels

A mismatch between product descriptions and the actual products purchased by customers can lead to higher return rates and decreased customer satisfaction. When customers purchase items that do not match their expectations based on product descriptions, they are more likely to return these products.

Ensuring accurate and detailed product descriptions, whether that be on the product itself or on the label under the product, can help set customer expectations and reduce return rates.

Implementing enhanced product description protocols can significantly reduce mismatches. Ensuring that product descriptions remain consistent and accurate across all platforms can be challenging but is necessary to maintain customer trust. There is also a risk of technical issues with technology orientated label printers etc, which could frustrate customers if not properly managed, such as having a label for an item with the incorrect measurements.

It is also worth noting that confounding factors such as packaging or labelling errors could be solely due to supplier issues, affecting product descriptions and return rates.

Moreover, there is often a lack of detail on products. In the section on unquestioned assumptions in chapter one, I discussed how positively promoting reduced or markdown items that are nearing expiration or are not up to supermarket's quality standards could potentially increase sales and reduce the stigma surrounding them. When items are placed in the reduced section of the supermarket, customers often wonder, "What is wrong with

this product? Is it expired, broken, or not up to quality standards?". Providing detailed information about these products can encourage customers to buy them, thus increasing sales and reducing waste.

Regarding product return rate and customer satisfaction, according to the expert's supermarket policy, non-food items that are reduced or marked down can be returned if they do not meet customer expectations. These increased descriptions and more informative labels can decrease the product return rate, as customers will have more information about the product before purchasing it at a discount, leading to greater satisfaction and fewer returns.

Metric : Trolley disengagement rate

Causal factor : Store layout and navigation with time constraints on customers

The layout of a store and ease of navigation play a crucial role in the trolley disengagement rate, particularly for customers shopping with time constraints. An organised store with good signage and good product placement can make shopping more enjoyable and efficient, reducing the likelihood of customers abandoning their trolleys. On the other hand, a confusing or poorly organised store layout can frustrate customers, especially those in a hurry, leading to higher disengagement rates and lower customer satisfaction.

Optimising store layout involves trade offs, such as frequent changes to store layout upsetting regular customers or cost of new signage. However, the benefits of reduced trolley disengagement and improved customer satisfaction can outweigh these costs. Clear and efficient store layouts can enhance the shopping experience by making it easier for customers to find products quickly, which is particularly beneficial for those with limited time. Also, the store would ensure to communicate any upcoming changes clearly to customers, some time in advance, to minimise the frustration.

A confounding variable with this cause is the level of staff assistance available. Even with a well organised layout, a lack of helpful staff to guide customers can still lead to high trolley disengagement rates.

Value : Community engagement

Metric : Local media coverage

Causal factor: Extent of community involvement and events

The level of community involvement and the frequency of local events hosted or sponsored by the supermarket significantly impact the amount of local media coverage it receives. This increased media coverage not only raises the supermarket's profile but also fosters stronger community ties, enhancing overall community engagement.

Increasing community involvement through sponsorship of events or participation in local initiatives can boost local media coverage. For example, as the expert states within their

interview, hosting toy drives at Christmas for less fortunate children and donating food to local food banks generate positive media.

This engagement not only improves media coverage but also strengthens community relationships. There are risks associated with increased community involvement, including the cost and time required to organise and sponsor events.

Additionally, there is a risk of media coverage being less favourable if events are poorly executed. Ensuring that events are well organised and align with the supermarket's value of community engagement, in contrast to doing it solely for profitability or financial gain can help maximise benefits.

A notable example from the expert was sponsoring a significant local football team, which not only boosts media coverage due to the team's prominence but also enhances the supermarket's reputation as a key community supporter.

It is also important to note that, according to the expert, the toy drive and local food bank support were primarily funded by customer donations. To enhance media coverage and community impact, the supermarket could consider a more direct provision approach. For instance, implementing a policy where the supermarket matches customer donations, such as donating five tins of food for every tin donated by a customer, could generate even greater media attention and demonstrate a stronger commitment to community support.

This more 'involved' approach could further influence the metric of local media coverage and the value of community engagement.

Metric : Branded merchandise sales

Causal factor : Sponsored football team performance

The performance of the local football team which the supermarket sponsors significantly impacts the sales of branded merchandise, reflecting community engagement. When the team achieves notable successes, such as winning a championship, fans tend to buy more team-related products to show their support. This surge in sales indicates increased community interest and boosts the supermarket's standing in the community.

Steady sales of branded merchandise, regardless of the team's performance, suggest a strong and lasting connection between the supermarket's branding and community support. By analysing these sales patterns, we can gauge how well the supermarket's sponsorship enhances its presence and engagement in the community.

However, the team's varying performance can lead to unpredictable sales changes. Lower sales during a losing streak might indicate reduced community support or less effective sponsorship. However, these fluctuations could also be due to other factors like competing promotions, merchandise prices, people choosing a different team, already owning all of that season's merchandise, or not liking the available products.

Value : Supermarket image

Metric : Net promoter score

Causal factor: Customer service quality

The quality of customer service influences the net promoter score and, consequently, the supermarket's image. This is because when customers have positive interactions with staff, it enhances the overall shopping experience, leading to higher customer satisfaction and a more favourable supermarket image.

High quality customer service can be considered as knowledgeable, welcoming, friendly staff, as described by the expert in the interview, as well as getting desired resolutions to issues or concerns. Improving customer service through comprehensive staff training programs, regular performance reviews, and implementing feedback mechanisms can enhance NPS.

Investing in customer service improvements involves trade offs such as the cost of training and the time required for implementation. There is also a risk that changes may not immediately reflect in NPS scores if customer expectations are not fully met.

Ensuring that staff are supported and have a focus on consistent, high quality customer service can lessen these risks. Balancing these investments with the potential long term benefits of improved customer loyalty and a stronger supermarket image is key.

Causal factor: Store cleanliness

Store cleanliness is a fundamental factor that influences net promoter score and thus impacts the supermarket's image. Customers are more likely to have a positive shopping experience and recommend the supermarket to others if they find the store clean, organised, and easy to navigate, contributing to a good impression and higher customer satisfaction, which can enhance the supermarket's overall image, as well as the NPS.

For example, if improvements in cleanliness lead to higher NPS scores, it indicates that these factors are positively affecting the supermarket's image. Maintaining high standards of cleanliness involves ongoing costs for staffing and supplies, and it requires consistent attention to detail.

I assume that there is also a risk that fluctuations in store conditions due to staff performance may affect NPS scores and the supermarket's image.

Additionally, as a note to be aware of, a confounding factor such as high foot traffic during seasonal times could temporarily affect cleanliness and influence NPS.

Metric: Share of voice

Causal factor : Customer loyalty programs

When the supermarket introduces successful loyalty programs, or personalises existing ones, it can lead to customers feeling appreciated which often results in them sharing their positive experiences both on social media and in person.

Additionally, a well received loyalty program can make the supermarket seem more attentive to customer needs and preferences, improving supermarket image around customer care. A rise in positive SOV can show that the loyalty program is enhancing the supermarket's public perception.

However, if the program fails to meet customer expectations, fails to evolve or overly personalised, its impact on customer's appreciation and the supermarket's image may decrease.

It is essential to keep the program engaging and consistently aligned with customer preferences to preserve its positive influence. Offering unique and meaningful rewards can further enhance the supermarket's reputation by increasing customer satisfaction and loyalty.

As a side note, certain times of the year such as winter holidays usually see increased social media mentions, for example sharing bargains and 'in-stock' items in certain supermarkets. It is good to be aware of these confounding variables as these seasonal spikes will increase SOV, regardless of the loyalty scheme. Thus, it is difficult to see if the SOV is being impacted by the loyalty discounts at this time, or just increased social media in general.

Causal factor : Sales promotions

Promotions that resonate well with the community, such as charity events or themed sales such as back to school range, often generate customer attention and sharing on social media, for example, mothers sharing their children's back to school range and prices.

Promotions that engage customers can lead to increased positive mentions on social media, improving the supermarket's online reputation and visibility. Promotions can reinforce the image that the supermarket offers value beyond just products, strengthening its image as community and customer focused, which is a positive image to have. For example, if the supermarket had promotions, like the toy drive, where charities were involved, or by buying certain products, customers are donating to a good cause.

However, there is a risk that some promotions may not achieve the desired level of engagement or may not align with customer preferences, potentially affecting their impact on the supermarket's image, as well as the potential loss of profit with overstocking or other aspects. Also, overemphasis on short-term promotions might detract from long-term brand-building efforts, affecting the overall perception of the supermarket.

Value : Financial performance

Metric : Gross profit margin

Causal factor : Dynamic pricing

Within almost all supermarkets in the UK, dynamic pricing models are used to adjust prices in real-time based on factors such as competitor pricing, and thus maximising revenue and improving GPM. This approach allows for more flexible and strategic pricing decisions that can enhance financial performance.

For example, discounts or promotions can be offered for products nearing expiration to boost sales and reduce waste. As shown when I was challenging assumptions, this ‘markdown’ aspect can be used as a strategy to boost more sales and profits when considering promoting and changing the stigma around these markdowns, which in turn will increase GPM, assisted by dynamic pricing, and allowing for a significant amount of time for customers to view the price reductions, creating increased exposure, rather than reducing product prices them 1 hour before store closure.

However, dynamic pricing introduces trade offs, such as the need for better technology as systems evolve, and potential customer dissatisfaction with prices varying across the supermarket, as well as being frustrated that, if they had bought that item a week prior or two days later, they could have gotten it cheaper. There is also the question of “Does the supermarket relate to local models of dynamic pricing with local competitors, or does the pricing reflect all stores in UK or country?”. Whatever the decision makers of the supermarket choose will ultimately have a desired or undesired effect on the gross profit margin for their company, thus effecting financial performance.

Additionally, as a side note, any increases or decreases in supplier costs can also impact GPM, making this a confounding variable that the supermarket must be aware of.

Causal factor : Store format and space

Optimising store layout to enhance the product visibility and use of retail space can lead to higher sales and reduced operational costs. For example, placing high margin items in prominent locations or dividing the actual layout of the store into departments that it should reflect, or having the essential items at the back of the store to ensure customers walk past promotions and other products before purchasing their items, may improve profitability and enhance GPM, and overall supermarket financial performance.

As previously mentioned in chapter one further work, it could be ideal to analyse the supermarket’s revenue per square meter, especially considering the causal impact of space utilisation on financial performance.

Tracking sales data and customer flow patterns before and after layout changes can provide valuable insights. However, there is a risk that redesigns might not produce the desired improvements in GPM if they do not align with customer preferences or operational needs. Also, during redesigns there may be increased customer dissatisfaction due to the disruption of their regular shopping routine. Finally, customers who frequently visit a store for essential items might not come as often if those items are inconvenient to access. For

instance, if the products they need are located far the entrance to the shop, they might be discouraged from shopping there regularly. In contrast, convenience stores like fuel station shops are designed to be more accessible and convenient, which can encourage more frequent visits, thus increasing profits, GPM and financial performance.

Metric : Return on assets

Causal factor : Operational cost management

When considering operational costs, the factor of managing them efficiently and effectively plays a role with return on assets. By focusing on efficient management of components such as labour costs, utilities, and technology expenses, a supermarket can significantly enhance its net income, thereby improving its ROA.

For example, managing labour costs through scheduling, reducing overtime and getting the correct staffing levels can lead to cost savings. Using workforce management technologies can ensure that staffing aligns with demand, preventing both understaffing and overstaffing.

If effective cost management strategies lead to reduced expenses and higher profitability, it indicates a positive effect on ROA and financial performance of the supermarket.

Moreover, when considering trade offs, the implementation and upgrading of technologies to align with operations have substantial upfront costs which can strain short term financial impact, but in the long term can balance out the costs and increase profits, thus improving the value of financial performance. An example of a risk would be that when managing labour costs by reducing overtime rates can frustrate employees would like additional hours, thus excessive cut backs are not recommended from my point of view, however it is the decision makers decision when it comes to prioritising employee satisfaction or financial performance.

Section 2.2: Understanding Issues, Their Metrics, and Any Causal Factors

1. Identification of Business Issues

After further review of the interview with the expert, I focused on identifying the core issues within the business, specifically, the negative outcomes that decision makers are most concerned about and want to mitigate or avoid. I did not include every issue mentioned in the interview. Some were not backed by broader market evidence and seemed more like individual opinions rather than widely accepted concerns. I only included issues that were supported by industry-wide research and evidence to ensure they were truly important, not just personal viewpoints.

As a note, here is a list of values that were identified from the expert's interview but not elaborated on due to both time constraints of the project and the reasons listed below, yet should not be ruled out of further work:

- Communication gaps : The mention of communication issues, such as the delayed instance of promotional information, indicates their potential significance. However, without further detailed analysis and little to no industry wide evidence available, the full impact and broader implications of these gaps remain underdeveloped in this report.
- Operational adjustments during peak seasons : Operational adjustments during peak seasons were noted in the interview but not extensively covered. The expert highlighted specific changes, such as increased staffing and extended operating hours during busy periods. From conducting wider industry research, it seems like a rather niche or personal issue, and thus remains anecdotal for now.

Thus, here are the issues I identified, with elaboration and evidence:

1) Technology integration

Within the interview, the expert stated on multiple occasions that their supermarket faces challenges with integrating new technologies, affecting both employees and customers. Implementing new systems, such as an updated inventory management system, has been problematic due to the diverse range of employee ages and varying technological skills. While training sessions are conducted to help employees adapt, recent updates—like enabling employees to directly add new products—required substantial support and hands-on demonstrations. Additionally, the self-scan system for customers encounters issues, particularly with item weight discrepancies in the bagging area, which often result in system flags that need retail assistants to resolve.

Firstly, regarding the credibility of these claims, Fic (2023) highlights that self-service technologies in supermarkets often need a significant amount of staff intervention. These systems frequently run into problems like verifying age-restricted purchases, dealing with glitches, and assisting customers who are not comfortable with technology. Self-service tills, in particular, often need employees to intervene and fix issues or manually override the system. This shows that while self-service can be useful, it still relies heavily on staff to keep everything running smoothly, leading to operation inefficiency in the long-term. These issues often lead to longer wait times, decreased customer satisfaction, and ultimately, a potential loss of customer loyalty, which is obviously issues that decision makers want to avoid.

Moreover, in March 2024, both Tesco and Sainsbury's faced major disruptions due to technical problems with their systems (Stavrou, 2024). Sainsbury's experienced a significant failure in their contactless payment system following a problematic software update, which led to widespread issues for customers trying to make payments. Meanwhile, Tesco had to cancel numerous online orders because of severe system malfunctions that disrupted their order processing and fulfilment. These incidents highlight the challenges that even large retailers face with technology integration and the impact such disruptions can have on customer experience and operational efficiency.

Furthermore, according to Progressive Grocer (2024), 40% of grocery retailers have trouble adopting new technologies because of the high costs and complexity. This issue is part of a bigger challenge in the grocery industry, where retailers are trying to update their systems

while dealing with tight budgets and complicated new technologies. The main problem is struggling to introduce new tech and train employees to use it effectively.

To further this research, I decided to use Perplexity AI to delve into some of the main issues supermarkets face when integrating new technology (Figure 12).

From this research, I found a study by Hunter (2024), which demonstrates the issue with technology integration within supermarkets. Some supermarkets are trying to modernise by combining different systems like inventory, point-of-sale, and supply chain into one system, which is complicated and requires careful planning to keep data accurate and secure, and often causes a lot of issues concerning data corruption upon transfer. Additionally, employees also need to adjust to new systems and workflows which is time consuming and often does not have the desired outcome within a small timeframe. Handling these challenges well is essential for staying competitive and keeping up with future technology.

Figure 12

What are the main challenges supermarkets face when integrating new technology

Sources

- Trends and Challenges Shaping the Supermarkets of the... vaimo • 1
- 6 ways UK supermarket brands are using technology to drive reta... Isretail • 2
- Grocery Digital Transformation: Navigating the... rizing • 3

Answer

Supermarkets face several significant challenges when integrating new technology. These challenges span various operational and strategic areas:

- 1. Legacy Systems and Integration**
 - Compatibility Issues:** Many supermarkets operate on outdated, monolithic backend systems that are not easily compatible with modern digital technologies. This makes integration complex and costly, often requiring extensive overhauls of existing IT infrastructure to support new solutions ⁵.
- 2. Data Management and Migration**
 - Data Discrepancies:** Successful technology integration often necessitates a well-defined data migration plan. Retailers must ensure that all data, including products, vendors, customers, and transactional information, is accurate and reliable. Discrepancies in data can hinder the functionality of new systems and lead to operational inefficiencies ³.
- 3. Change Management**
 - Adoption and Training:** Introducing new technology requires significant changes in processes and workflows. Ensuring that all employees are adequately trained and that there is buy-in from top management is crucial for smooth adoption. Resistance to change and inadequate training can lead to underutilization of new technologies ³.

As I could not find any KPIs or performance reviews online that relate to the aspect of technology integration being an issue for supermarkets, I decided to use ChatGPT as a source of confirmation, whilst remaining aware of bias. The research I had done prior helped

to prepare me for any unrelated or divergent aspects that may have been biased. The main parts of this conversation can be found in figures 13 and 14.

The results presented through figure 13, relating to technology integration issues within supermarkets, heavily indicate that both the expert's view and research I have found are in agreement that this is a core issue for the supermarket and decision makers. Aspects such as employee training, customer satisfaction (such as the expert's indication to the issues with self-service machines), inventory management, and security are all byproduct issues of technology integration.

Additionally, Figure 14 illustrates that prominent supermarkets have encountered difficulties with technology integration. Notably, all the supermarkets listed are industry leaders, indicating that this is not merely a localised problem but a widespread issue recognised across the sector. These results show the importance of addressing technology integration challenges as a core concern for decision makers in the wider supermarket industry, as well as the expert's supermarket.

Figure 13

A screenshot of a digital communication platform, likely a messaging app or a specialized survey tool. The message is framed by a black border. At the top right, there is a small circular icon with a question mark and the text "Is a main issue for a supermarkets \"technology integration?\"". The message itself starts with a small profile picture of a person with short hair and a light blue background. The text reads: "Yes, technology integration is a significant issue for supermarkets. Here are a few key aspects where it can be challenging:" followed by a numbered list of six points. After the list, there is a concluding sentence and a row of small, faint icons at the bottom.

Is a main issue for a supermarkets "technology integration?"

Yes, technology integration is a significant issue for supermarkets. Here are a few key aspects where it can be challenging:

1. **Point-of-Sale Systems:** Upgrading or integrating new POS systems can be complex, especially if it needs to work seamlessly with existing inventory and customer data systems.
2. **Inventory Management:** Implementing real-time inventory tracking and automated ordering systems can improve efficiency but requires robust technology and integration with supplier systems.
3. **Customer Experience:** Integrating technology like self-checkout kiosks, mobile payment systems, and personalized promotions requires careful planning to enhance the shopping experience without causing disruptions.
4. **Data Security:** Protecting customer data and transaction information is crucial, requiring advanced cybersecurity measures and compliance with regulations.
5. **Supply Chain Coordination:** Technology integration helps streamline the supply chain, but coordinating between multiple vendors and systems can be complex.
6. **Employee Training:** New technologies often require training for employees to ensure they are used effectively, which can be a resource-intensive process.

Overall, successful technology integration can lead to improved efficiency, better customer service, and a competitive edge, but it requires careful planning and execution.

⋮ ⏴ ⏵ ⏷ ⏸ ⏹ ⏺

Figure 14

The figure shows a screenshot of a digital interview platform. At the top, there is a question bubble containing the text: "What supermarkets have faced these issues? Can you please give concrete examples?". Below this, there is a response area with a user icon and the text: "Certainly! Here are some concrete examples of supermarkets that have faced issues related to technology integration:". A numbered list follows, detailing issues faced by five supermarket chains:

1. **Walmart:**
 - **Issue:** Walmart's implementation of RFID (Radio Frequency Identification) technology aimed at improving inventory management encountered difficulties with system integration and accuracy. Early challenges included discrepancies in inventory data and disruptions in the supply chain. **The company is still working on finding a solution.**
2. **Kroger:**
 - **Issue:** Kroger has faced issues integrating new POS systems and self-checkout kiosks across its stores. The initial rollout of self-checkout systems experienced reliability issues and customer acceptance problems. **Kroger continues to seek effective solutions.**
3. **Tesco:**
 - **Issue:** Tesco struggled with integrating its online shopping platform with in-store inventory systems, leading to inventory discrepancies and challenges with home delivery logistics. **The company is still addressing these issues.**
4. **Target:**
 - **Issue:** Target faced significant problems with inventory management and POS systems during its expansion into Canada. The technology integration issues resulted in stock shortages and operational inefficiencies. **Target is actively working to resolve these challenges.**
5. **Ahold Delhaize (including brands like Stop & Shop and Giant):**
 - **Issue:** Ahold Delhaize has encountered difficulties with integrating various technologies across its brands and locations, which has led to inconsistencies in customer experience and inventory management. **The company is still in the process of finding a resolution.**

At the bottom of the response area, there are small icons for navigating through the interview transcript.

In summary, technology integration is a core issue for supermarkets, impacting their efficiency, customer experience, and overall business operations. The research I have found on the wider industry paired with the expert's interview points clearly indicate technology integration being a core issue for the supermarket. Furthermore, poor technology integration is a large concern for decision makers because it causes inefficiencies, higher costs, and dissatisfied customers. When technology systems are not integrated as desired, it disrupts operations and makes it harder to make educated decisions, which can diminish the supermarket's success and competitive edge. Effective technology integration is crucial for staying competitive in a rapidly evolving market. Poor integration can lead to a significant competitive disadvantage, particularly as consumer expectations for seamless technology interactions rise.

2) Employee turnover

Within the interview, the expert notes that high employee turnover presents a major challenge for the supermarket, disrupting established routines and collaboration among long-term staff who have developed efficient work systems. This constant turnover of new employees interrupts these workflows, leading to inefficiencies and increased operational costs due to the ongoing need for training and adjustment. Although the supermarket has introduced a system to pair new hires with experienced staff to aid their integration, turnover remains high, and the underlying issues continue. This instability affects both the effectiveness of the team and the supermarket's overall performance.

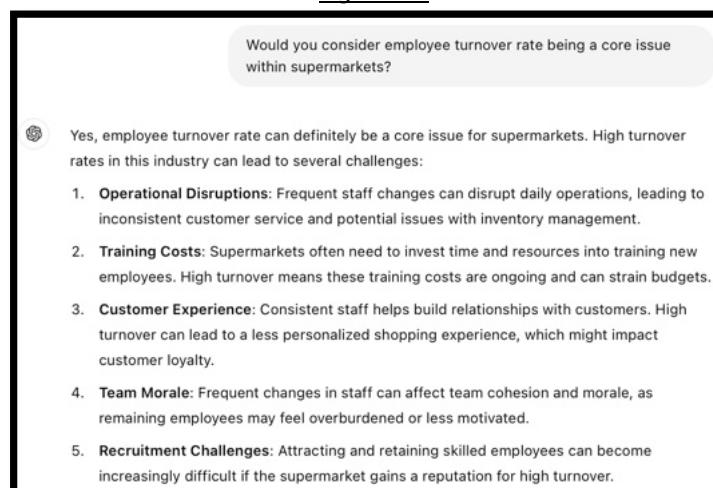
Regarding industry-wide evidence, an article by Katrina (2022) shows that supermarkets have an average employee turnover rate of 100%. This is somewhat due to the temporary and seasonal staff employed, as well as better opportunities arising elsewhere, and also low pay and better benefits elsewhere in different industries. The 100% turnover rate in the supermarket industry indicates a systemic issue affecting overall productivity and operational efficiency.

Moreover, a case study on Tesco by Cram (2020) highlighted that the industry giant has a significant challenge when considering employee turnover rates, and the affect it has on productivity and operational efficiently. A key point within this study is that Tesco sees a clear issue in the balance between company performance and employee turnover. Additionally, knowing that market leaders also struggle with this issue adds credit to the expert's interview surrounding this issue.

As a side note, Whitfield (2024) states that it takes an average of 13 weeks for a new retail employee to undergo the necessary training and have a desired rate of performance. This training period results in increased costs for training resources and time, while lower productivity levels within the 13 weeks reduces overall efficiency and disrupts established workflows. Additionally, during training, new hires may not provide the level of customer service expected, potentially affecting customer satisfaction and loyalty.

I decided to consult ChatGPT about this issue in supermarkets, as shown in figure 15. The results aligned well with the points raised in the expert interviews, particularly concerning team morale, and supported existing research on broader supermarket industry challenges, such as operational inefficiency, training costs, and customer experience. Based on this alignment and the significant impact of these factors, I concluded that this is indeed a legitimate problem for supermarkets and should be considered a core issue. High employee turnover not only disrupts daily operations but also affects long-term strategic goals by preventing the development of a skilled and experienced workforce, ultimately reducing the supermarket's ability to compete effectively, thus something that decision makers and supermarkets in general would prefer to avoid.

Figure 15



3) Supply issues

During the interview with the expert, I noticed they had alluded to something much bigger than what they saw on the surface. They stated that, since Brexit, their supermarket has faced significant supply chain disruptions, particularly for fresh fruit and vegetables. Increased border complexities have led to delivery delays and the withdrawal of a key suppliers. These delays and supply issues continue to result in perishable items often being unsuitable for sale upon arrival. For instance, a recent delay in a large strawberry shipment, caused by new documentation requirements, led to the waste of most of the strawberries and a prolonged shortage in the supermarket. So, this aspect of Brexit mentioned alludes to the aspect of an industry wide supply chain issue which I wanted to investigate more to see if this could potentially be a core issue for the supermarket. If Brexit can affect supply issues significantly, surely other world-affairs can too.

As approximately 80% of food in the UK is imported from the EU, Brexit has significantly impacted Tesco's supply chain, resulting in increased costs and delays (Course Hero, 2024). The added complexities and new border processes have disrupted the timely delivery of goods, affecting the availability and freshness of products in-store.

When viewing and analysing Tesco's financial report (2023), I observed the risk management concerns, shown in figure 16. Based on the information in this section of the report, it is clear that supply chain problems are a major issue for supermarkets, when looking at Tesco, as it is one of the market leaders. The complexity of their supply chains means they often face disruptions, which can lead to shortages of products. Factors like political events, economic changes, and natural disasters can make these problems worse. To address these challenges, Tesco officially identified "security of supply" as a key risk. This means ensuring a steady supply of products is now a top priority for their risk management efforts.

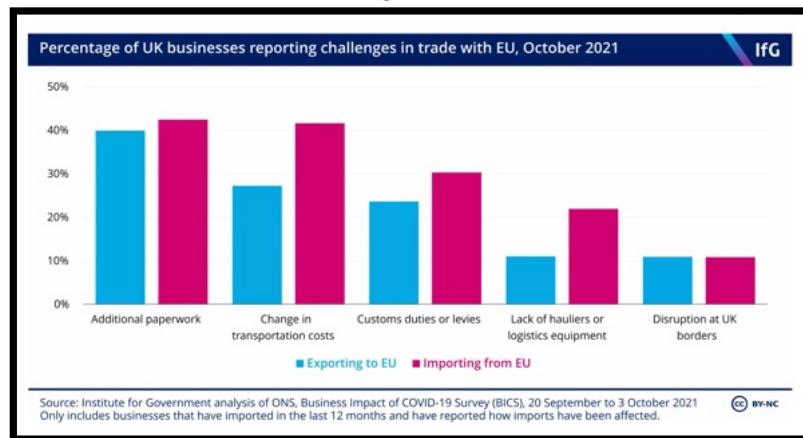
Figure 16

Risk management

The Committee reviewed the Group's principal and emerging risks and mitigation strategies, with particular discussion of prioritised risks and risk movements. We also discussed the complexity of the Group's supply chains and the threat of frequent disruption to product availability due to increased volatility in external conditions. It was considered appropriate to adopt security of supply as a new principal risk, which was considered and approved by the Board. The Committee received updates on various emerging risk themes in areas such as technology, economics, political impacts, talent, climate and sustainability, including periodic deep dive sessions, see page 39. These are considered by management in connection with the risk assessment process. An assessment of the Group's principal risks and detailed scenario analysis work to stress test liquidity was performed as part of the viability scenario modelling. For further information on the Group's risk management framework, see page 38.

Furthermore, an article by the Institute for Government (2023) demonstrates that supermarket supply chains have faced significant challenges due to several factors. Brexit and the COVID-19 pandemic have led to a shortage of workers, particularly truck drivers, which has disrupted timely deliveries. Additionally, the war in Ukraine has increased energy prices and contributed to inflation. These issues have resulted in product shortages, higher costs, and less availability on supermarket shelves. To address these problems, supermarkets are adapting by diversifying their suppliers and increasing their inventory levels to ensure they can continue to provide essential products. However, it still highlights that supply issues are a major issue for supermarkets. Figure 17 (Institute for Government, 2023) further illustrates these points by showing the percentage of UK businesses that reported difficulties in trading with the EU in October 2021, highlighting that importing from the EU was notably more problematic.

Figure 17



Finally, figure 18 shows my results from Perplexity AI when considering the supply issues in UK supermarkets last year. Regarding this information, UK supermarkets faced major supply issues and food shortages, especially for fresh produce like tomatoes, peppers, cucumbers, lettuce, and berries. This led retailers such as Asda, Morrisons, Tesco, and Aldi to implement purchase limits. Key factors behind these shortages included poor weather in Europe affecting crop yields, high energy costs reducing domestic production, and disruptions in global supply chains and transportation. While Brexit has added to costs, it was not the main driver of these issues. Experts warned that shortages could persist or worsen into 2024 due to climate change, new border checks, and ongoing shipping disruptions.

Figure 18

The screenshot shows a search results page with the title "Supply Issues Supermarkets Uk 2023". Below the title, there is a section titled "Sources" with three news items:

- Food shortages 'alarmingly likely' in the UK next... - telegraph... · 1
- Why are UK supermarkets facing fresh... - theguardian... · 2
- The Worst Meat and Dairy Greenwashing... - feedback... · 3

Below the sources, there is a section titled "Answer" which provides a summary of supply issues in UK supermarkets in 2023. The summary includes:

- Shortages of fresh produce: Supermarkets experienced shortages of various fruits and vegetables, including tomatoes, peppers, cucumbers, lettuce, and berries. This led some major retailers like Asda, Morrisons, Tesco, and Aldi to introduce purchase limits on certain items.
- Causes of shortages:
 - Poor weather conditions in Spain and other European countries affected crop yields.
 - Reduced UK domestic production due to high energy costs for greenhouse cultivation.
 - Transportation issues and global conflicts disrupted supply chains.
 - The UK's heavy reliance on imports (46% of food consumed in 2020 was imported).
- Impact of Brexit: While not considered the main cause, Brexit has increased costs and bureaucracy for importers, potentially putting the UK at a disadvantage when supplies are limited across Europe.
- Ongoing concerns: Food security experts warn that shortages may persist or worsen in 2024 due to climate change, transportation issues, and global conflicts.

Considering the cost implications and ongoing supply issues faced by UK supermarkets, it is fair to say that these challenges are a central concern, thus making supply disruption a core issue for the supermarket. Addressing supply chain disruptions is crucial for maintaining customer trust and ensuring the supermarket can meet consumer demand consistently. Failure to do so could lead to significant revenue losses and damage to the brand's reputation, which again is something decision makers for supermarkets want to avoid.

2. Development of Metrics

Issue : Technology integration

Mean time to repair (MTTR)

MTTR is a metric to reflect the average time it takes to repair a failed system, in this case a technological integrated one. The time starts from when the issue is raised to when the system is operating as desired.

ChatGPT states that the mean time to repair is calculated by the total repair time divided by the number of repairs, as shown below.

1. Define the Scope
<ul style="list-style-type: none"> • Technology Types: Identify the specific types of technology you are focusing on (e.g., POS systems, self-checkout machines, refrigeration units). • Repair Events: Define what constitutes a repair event or incident.
2. Collect Data
<ul style="list-style-type: none"> • Record Keeping: Track each repair event, noting the start and end times for each repair. This can be done through maintenance logs, incident reports, or repair tickets. • Data Points: Gather data points for the total time taken for each repair. Ensure you have consistent records for the time each issue was reported and when it was resolved.
3. Calculate Individual Repair Times
<ul style="list-style-type: none"> • For each repair event, calculate the time taken by subtracting the start time from the end time. • Formula: $\text{Repair Time} = \text{End Time} - \text{Start Time}$
4. Sum of Repair Times
<ul style="list-style-type: none"> • Add up all the individual repair times to get the total repair time.
5. Count of Repairs
<ul style="list-style-type: none"> • Count the total number of repair events that occurred during the period you are analyzing.
6. Calculate the Mean Time to Repair (MTTR)
<ul style="list-style-type: none"> • Divide the total repair time by the number of repair events. • Formula: $\text{MTTR} = \frac{\text{Total Repair Time}}{\text{Number of Repairs}}$

Supermarkets use technological systems like POS and inventory management. When these systems fail, it disrupts operations. MTTR is a good measure of technology integration issues because it shows how effectively and quickly the supermarket can address and resolve technology related problems, reflecting the overall efficiency of the integration process. Monitoring MTTR will help pinpoint specific issues related to the time taken to resolve technology problems. By tracking this metric, supermarkets can identify delays in the repair process and areas where improvements are needed. This approach enables targeted improvements in support processes and integration strategies, leading to quicker resolutions and minimise disruptions.

I asked ChatGPT for guidance on locating data related to MTTR in a supermarket dataset. I wanted to better understand what to look for when implementing this metric and to gain insights into areas within the supermarket where such data might already be recorded, which would aid in calculating this metric. The response from ChatGPT was, in summary, that to calculate Mean Time to Repair (MTTR) in a supermarket, you need data on system failures, downtime, and repair activities. Key sources include incident or trouble tickets, system downtime logs, maintenance records, repair and support tickets, and operational performance reports. The process involves gathering this data, identifying relevant records, calculating the downtime for each issue, and then computing MTTR by dividing the total downtime by the number of repairs. This approach helps in assessing the efficiency of the repair process and the impact of technology issues. The results of this conversation are as follows.



To find data related to Mean Time to Repair (MTTR) in a supermarket dataset, you'll want to look for information related to system failures, maintenance activities, and repair logs. Here are the key data points and sources you should consider:

1. Incident or Trouble Tickets

- **Data to Look For:** Records of technology-related incidents or trouble tickets, including timestamps for when the issue was reported and when it was resolved.
- **Where to Find It:** Check the IT support or helpdesk system used by the supermarket. This system often logs detailed information about system outages, hardware failures, and software issues.

2. System Downtime Logs

- **Data to Look For:** Logs that record periods when technology systems (POS, inventory management, etc.) were not operational. These logs should include start and end times for each downtime event.
- **Where to Find It:** System monitoring tools or software management systems often track downtime events. This data might be available in system administration dashboards or logs.

3. Maintenance Records

- **Data to Look For:** Documentation of maintenance activities, including scheduled maintenance and unscheduled repairs. Records should indicate the duration of each maintenance activity.
- **Where to Find It:** Facilities management systems or maintenance management software used by the supermarket.

4. Repair and Support Tickets

- **Data to Look For:** Information on repairs and support tickets, including the date and time of the request, the resolution time, and any notes on the type of issue and repair process.
- **Where to Find It:** IT service management systems or customer support platforms where repair tickets are logged and managed.



Technical integration error rate

The metric of technical integration error rate directly measures the reliability and functionality of the technology systems within the supermarket. It tracks the proportion of systems successfully implemented within a given time frame or the speed at which the integration milestones are completed.

Integration errors happen when systems don't communicate properly, causing data inconsistencies, process interruptions, and manual fixes. By counting these errors, we get a clear picture of how severe the integration issues are. High error rates mean the integration process has significant problems, which can reduce efficiency and increase operational costs.

Monitoring the technical integration error rate helps reflect issues in technology integration by identifying problem areas in the process. It reveals stages where delays or difficulties occur, offering insights into where improvements are needed. A high integration error rate suggests that new technologies are being adopted smoothly, with minimal disruption. In contrast, a low rate might indicate compatibility issues or inefficiencies in integration strategies.

Additionally, tracking this metric can highlight if there are needs for additional resources or adjustments, such as more training or enhanced support systems, to accelerate integration. By analysing trends and patterns in the integration error rate, alongside other measures such as employee adaptation, supermarkets can better plan and manage future technology integrations. This focused approach to monitoring this metric enables targeted improvements and more efficient technology adoption.

To further understand the data required for the calculation and the formula itself, I asked ChatGPT to outline these areas. The response from ChatGPT was, in summary, that calculating the technical integration error rate involves determining the total number of integration projects and the number of those successfully completed. The formula is the number of successfully integrated systems divided by the total number of integration projects, multiplied by 100. To calculate this, supermarkets can use data from project management records, IT integration logs, maintenance tickets, system documentation, operational reports, and integration checklists. The response of this conversation is detailed below.

Formula for Technical Integration Rate

1. **Determine the Total Number of Integration Projects or Systems:** This includes all the technology integration efforts planned or underway within a specific period.
2. **Count the Number of Successfully Integrated Systems:** This includes the number of new systems that have been fully integrated and are functioning as intended.
3. **Calculate the Integration Rate:**

$$\text{Technical Integration Rate} = \left(\frac{\text{Number of Successfully Integrated Systems}}{\text{Total Number of Integration Projects}} \right) \times 100$$

Example Calculation

If a supermarket had 10 technology integration projects planned in a given period and 7 of them were successfully completed and fully integrated:

$$\text{Technical Integration Rate} = \left(\frac{7}{10} \right) \times 100 = 70\%$$

Data Readily Available in Supermarkets for Calculation

To calculate the Technical Integration Rate, you can use the following data sources:

1. **Project Management Records:** These records track the progress of technology integration projects. They should include information on project goals, timelines, milestones, and completion status.
2. **IT and System Integration Logs:** These logs provide details on the integration process of new technology systems, including start and end dates, issues encountered, and successful integration indicators.
3. **Maintenance and Support Tickets:** Data from support tickets can indicate whether integration projects were completed successfully and whether any issues were reported.
4. **System Documentation:** Documentation related to new technology implementations often includes information about integration status and project outcomes.
5. **Operational Reports:** Reports on technology performance and integration progress can offer insights into which systems have been integrated successfully and which are still in progress.

Issue : Employee turnover

Employee turnover rate

ETR is a metric used to numerically represent the proportion of employees who leave a company within a given year. Resources, such as the one from Pavlou (2024), help companies and businesses calculate their employee turnover rate with ease. Essentially, “It's calculated by dividing the number of employees who left by the average number of employees, then multiplying by 100” (Pavlou 2024). Calculating the employee turnover rate helps companies understand how stable their workforce is and spot issues related to employee satisfaction, management concerns, and so forth.

A high turnover rate often signals problems like low employee engagement or poor management, suggesting the need to investigate why employees are leaving. By tracking this metric over time, organisations can identify trends, compare their performance to industry standards, and take steps to address the causes of high turnover.

I decided to ask ChatGPT to help me understand the additional implications of this metric. I asked “My supermarkets has, seemingly, an issue with employee turnover. What does this infer for my business, and what would calculating employee turnover metric help with this issue?”. ChatGPT responded that high employee turnover in a supermarket can lead to increased costs, decreased productivity, lower morale, inconsistent customer service, and loss of knowledge. Calculating the employee turnover metric helps by identifying patterns, measuring the impact of retention initiatives, benchmarking against industry standards, performing root cause analysis, and aiding in budgeting and forecasting. These insights can guide targeted actions to improve employee retention and overall business performance. The results of this conversation are shown below.

High employee turnover in a supermarket can infer several potential issues for your business, such as:

- Increased Costs:** Recruiting, hiring, and training new employees can be expensive. High turnover means these costs are incurred more frequently.
- Decreased Productivity:** New employees typically take time to reach the productivity levels of more experienced workers.
- Lower Morale:** High turnover can negatively impact employee morale, leading to a less cohesive and motivated workforce.
- Inconsistent Customer Service:** Frequent changes in staff can lead to inconsistencies in customer service, potentially harming customer satisfaction and loyalty.
- Loss of Knowledge:** When experienced employees leave, they take valuable knowledge and skills with them, which can affect operations and efficiency.

Importance of Calculating Employee Turnover Metric

Calculating the employee turnover metric helps in several ways:

- Identify Patterns:** Understanding the rate at which employees leave can help identify patterns or trends. For instance, if turnover spikes at certain times of the year or in specific departments, it might indicate underlying issues.
- Measure Impact of Initiatives:** If you implement strategies to improve retention, measuring turnover can help assess the effectiveness of these initiatives.
- Benchmarking:** Comparing your turnover rate to industry standards can help determine if your turnover is unusually high or within a normal range.
- Root Cause Analysis:** By analyzing turnover data, you can identify common reasons for leaving (e.g., compensation, work environment, management practices) and address these issues.
- Budgeting and Forecasting:** Knowing your turnover rate helps in budgeting for recruitment, training, and other related costs.

Issue : Supply issues

On time delivery rate

Lorena (2024) defines OTDR as "...a key performance indicator ecommerce and other delivery businesses use to assess their ability to fulfil a customer order by the promised delivery date... divide the total number of orders delivered by the number of deliveries that arrived after the promised delivery date."

So, in terms of supermarkets, OTDR is a performance metric used to measure the percentage of deliveries made to supermarkets on or before the promised delivery date, providing insights into the reliability and performance of the delivery process. While this metric is often used by companies delivering products, supermarkets can also monitor it by tracking the time frame between when an order is marked as 'filled' when the order actually arrives. Most supermarkets receive notifications about these statuses from the delivery company, similar to how customers receive 'dispatched' and 'order confirmation' notifications from Amazon.

The OTDR is crucial for supermarkets because timely deliveries are key to keeping shelves stocked with fresh products. Delays or inaccuracies in deliveries can cause empty shelves, lost sales, and unhappy customers. By tracking the OTDR, supermarkets can evaluate how well their supply chain is working and spot any issues that might affect their ability to keep products in stock for shoppers.

When ChatGPT was asked about where to find data on the OTDR in existing supermarket data sets, the response was that you can look into Order Management Systems for tracking order dates and delivery statuses, Supply Chain Management Software for performance monitoring and OTDR reports, Logistics and Shipping Records for detailed shipment timestamps, and Vendor Management Systems for supplier performance metrics. The results of this conversation can be found below.

Where to Look for This in Supermarket Data Sets:

In supermarket data sets, you can find OTDR-related information in several areas:

1. **Order Management Systems:** These systems track the dates when orders are placed, scheduled for delivery, and actually delivered.
2. **Supply Chain Management Software:** This software often includes modules for monitoring delivery performance and generating OTDR reports.
3. **Logistics and Shipping Records:** These records provide detailed timestamps and statuses of shipments, helping to verify if deliveries were made on time.
4. **Vendor Management Systems:** These systems may include performance metrics from suppliers, which can be useful in assessing the overall supply chain efficiency.

Stockout rate

SR is used by supermarkets to demonstrate the frequency at which products become out of stock. Essentially, it reflects the proportion of instances when a product was not available on the shelf compared to the times it was expected to be available. This metric is particularly valuable for reflecting supply issues because it directly ties stock availability to supply chain performance. When a supermarket consistently experiences high stockout rates, it indicates that there may be underlying issues such as supplier delays, logistical challenges, or inadequate stock levels. Addressing these issues can help improve inventory turnover, reduce customer frustration, and enhance overall shopping satisfaction. Additionally, a supermarket could calculate the percentage of stockouts relative to the total number of sales opportunities and compare it to industry standards to assess inventory management effectiveness, giving even further insight into stockout rate, but this is entirely down to the decision maker's preference of normalisation.

An example of calculating stockout rate for a supermarket is shown below, regarding the conversation I had with ChatGPT regarding a potential formula. In summary, ChatGPT says to calculate the SR, first determine how many times a specific item was out of stock during a specific period, then identify the total number of times the item was supposed to be available for purchase. To find the Stockout Rate, divide the number of stockouts by the total number of sales opportunities and then multiply by 100 to get a percentage.

To calculate the Stockout Rate, follow these steps:

1. **Determine the Number of Stockouts:** Count the number of times a specific item was out of stock during the period of interest. This can be tracked through inventory management systems or point-of-sale (POS) data.
2. **Identify the Total Number of Sales Opportunities:** This is the total number of times the item was intended to be available for purchase. It can be calculated by summing the number of sales instances or inventory checks where the item was expected to be in stock.
3. **Calculate the Stockout Rate:** Use the following formula:
$$\text{Stockout Rate} = \frac{\text{Number of Stockouts}}{\text{Total Number of Sales Opportunities}} \times 100$$

Example Calculation:

Let's say you want to calculate the Stockout Rate for a particular product over a month.

1. **Number of Stockouts:** Assume the item was out of stock 12 times during the month.
2. **Total Number of Sales Opportunities:** Suppose the item was supposed to be on the shelves 150 times during the same period (i.e., the item was scheduled to be available 150 times based on sales projections and restocking plans).
3. **Apply the Formula:**
$$\text{Stockout Rate} = \frac{12}{150} \times 100 = 8\%$$

In this example, the Stockout Rate is 8%, meaning the item was unavailable for purchase 8% of the time it was supposed to be on the shelf.

Share, Print, Copy, Download, Close

To calculate the SR, the supermarket can use POS data which records each transaction and indicates when items are sold out. Also, inventory management systems track current stock levels and document previous stockouts, showing how often items are unavailable. Moreover, delivery records provide insight into delivery delays or issues that could impact

stock levels. To find this information within a supermarket related data set, we could look at sections related to inventory levels and sales transactions, for example.

3. Analysis of Causal Factors

Issue : Technology integration

Metric : Mean time to repair

Causal factor : Diagnostic tool effectiveness

Advanced diagnostic tools can significantly enhance the ability to diagnose issues, including potential future problems, and speed up their identification, thereby reducing MTTR, and conversely positively impacting the issue of technology integration.

However, investing in new diagnostic tools involves trade offs between the benefits of improved repair times and the costs associated with purchasing and maintaining these tools. The cost of these tools and the training required for their use can be substantial, potentially straining budgets and resources.

Despite these challenges, more effective diagnostic tools usually improve technological integration and facilitate quicker and easier repairs. This, in turn, positively impacts both MTTR and overall technological integration, especially if the diagnostic tools are effective in identifying problems during the integration process.

It is worth noting that a confounding variable for this cause is that any existing maintenance processes may affect the usefulness of the new or updated diagnostic tools. They must be updated together. If not, any maintenance processes which are currently implemented within the supermarket that are outdated or less efficient may overshadow the improvements made by new or upgraded diagnostic tools.

Metric : Technology integration error rate

Causal factor : Technology compatibility

When choosing new technology to implement into the supermarket, it is crucial that it is compatible with both supermarket software and current systems, if a full or partial upgrade is not being completed. Selecting technology that is highly compatible with existing systems can significantly reduce integration errors. Ensuring that the new technology can interface smoothly with current systems and accommodate future updates will help in achieving a lower technology integration error rate.

If the systems are designed to work well together, technology integration error rate will decline, and this will have a positive impact on the issue of technological integration.

However, if the supermarket mainly focuses on cost effectiveness or compatibility with all current supermarket systems, then options are limited and this could have a negative effect on the technology integration error rate.

Additionally, compatibility issues with future updates could also arise and again impact technology integration error rates.

Causal factor : User training

Well educated staff surrounding new technology being implemented into the supermarket has a positive impact on technological integration error rates. The employees are the individuals who will, more than likely, be using this new system the most, such as cashiers with new POS services.

Providing adequate training is essential but must be balanced with avoiding training overload. Training needs to be comprehensive enough to be effective but not overly extensive. Proper training improves user competence, decreases the likelihood of mistakes, and accelerates the adoption of new technology. This leads to a lower technology integration error rate. However, over-extensive training or pressure on staff regarding 'perfect' performance with the new system can cause fear and delays in the rollout, thus ensuring that there is a balance between desired and suitable training standards is essential to keep technology integration error rates at a minimum.

Causal factor : Complexity of security systems

As demonstrated within my unquestioned assumptions within chapter one, I challenged the increased amount of security, including technology orientated security. Complex security systems may lead to higher integration error rates. Simplifying security technology and implementing more 'employee' based security can help minimise this error rate.

While simplifying security systems can reduce integration errors, it might also lower the overall security level, potentially leaving the system vulnerable to attacks. Therefore, it is important to strike a balance between technology and in-person security, as well as the appropriate amount of security within the supermarket.

Implementing security systems that balance complexity with usability will result in a lower technological integration error rate. Moreover, avoiding the excessive deployment of complex technological security measures can further reduce this rate. This balanced approach ensures that security systems are effective against threats while being integrated and updated with fewer errors. Simplifying or reducing the complexity of these systems enables the supermarket to focus more efficiently on addressing any errors that do occur, and also simplifies the planning and execution of future integrations.

Issue : Employee turnover

Metric : Employee turnover rate

Causal factor : Employee job satisfaction

Job satisfaction has various parts, such as work environment, work relationships, job role fulfilment, and overall work-life balance, or even the aspect of working from home.

High job satisfaction typically leads to lower turnover rates as employees who are content with their jobs are more likely to stay with the company. On the other hand, low job satisfaction can lead to higher turnover as employees seek more fulfilling or supportive work environments elsewhere.

However, changes to improve job satisfaction can cost a lot and may require significant time to implement. Additionally, there is a risk of over promoting employees or making changes that do not address the root causes of dissatisfaction, for example, longer break times but the breaks were adequate.

Additionally, regarding trade offs, investing money into employee satisfaction with regards to their role in the company means using money from other aspects of the supermarket that could have been more beneficial, thus creating more or new issues elsewhere, especially if the parts that were improved were not the root cause of employee dissatisfaction. However, implementing this correctly also has a positive impact on supermarket image and reputation, regarding the fact that the majority of employees are satisfied, which is another benefit as opposed to simply reducing employee turnover.

Moreover, confounding factor here could be changes to wage levels in the wider industry, which could indirectly affect turnover rates independently.

Issue : Supply issues

Metric : On time delivery rate

Causal factor : Weather

Bad weather conditions, particularly during winter months, can severely impact the on time delivery rate. Heavy rain and icy roads can cause delays in transportation, making it challenging for delivery companies to maintain schedules.

When delivery companies face weather related issues, the likelihood of maintaining an on time delivery significantly decreases. The increased risk of accidents and slower travel speeds result in increased transit times, which directly impacts the ability to meet promised delivery windows. This, in turn, disrupts the supply chain and leads a significantly reduced number of on time deliveries, thus reducing on time delivery rate within supermarkets during peak seasonal demand periods like Christmas.

Diversifying supply sources and delivery companies can reduce reliance on any single company or route, thereby improving flexibility around deliveries, when bad weather is a factor. If another company is supplying half of the stock and another company the other half,

then the on time delivery rate may only be affected slightly or less important if only half or a quarter of an order arrived late.

However, unexpected weather or unseasonal conditions may still lead to delays despite preparation, effecting both on time delivery rate and adding to supply issues, thus highlighting the need for flexibility in the supply chain.

Causal factor : World affairs

The ongoing conflict in Ukraine has significantly disrupted global supply chains, resulting in delays in on-time deliveries. The war has led to the blockage or rerouting of transportation routes and damage to critical infrastructure, complicating the movement of goods.

Additionally, sanctions and trade restrictions imposed on Russia have added to these issues, leading to increased transit times and delays at customs. Therefore, supermarkets are finding it hard to meet delivery times and maintain product availability.

Moreover, the recent blockage of the Suez Canal, a global shipping route, has had an impact on the on time delivery rate, making the flow of good complicated. This incident has shown the direct effect that global disruptions can have regarding on time delivery rates and overall supply chain issues, contributing to issues in inventory management and stock availability for supermarkets.

Similarly to the causal factor of weather, increased supplier flexibility can improve these rates to make sure even partial deliveries are on time, thus somewhat addressing the underlying issue of product availability, when considering supply issues. But again, this needs planning and potentially increased costs of delivery rates, and so on.

Metric : Stockout rate

Causal factor : Seasonal demand

Around Christmas, customers tend to stock up on supplies and buy in large quantities. Supermarkets need to plan for these seasonal changes to avoid stockouts. This planning can be done by analysing historical sales data and monitoring inventory counts and depletion during these periods, allowing for increased orders at that time.

However, inaccurate forecasts can lead to overstocking, resulting in higher holding costs and increased waste. Despite accounting for these seasonal demands, supermarkets often still experience stockouts. This can be due to supply issues during Christmas, such as adverse weather conditions or delivery companies being overwhelmed with high volumes, causing delays. Better inventory management reduces the likelihood of stockouts and overstocking, improving customer satisfaction and also when supply issues arise it can be combated. However, there are some high costs around upgrading or implementing forecasting, in addition to managing increased stock can be challenging.

Therefore, seasonal variations, particularly in the winter months, contribute to stockouts and highlight supply issues even when higher purchases are anticipated. Accurate demand forecasting and flexible supply chain practices are essential to balance these factors effectively.

However, a confounding variable for this cause may be that suppliers or delivery companies may not have the products or limited transport vehicles when considering seasonal demand, thus effecting stockouts indirectly.

Causal factor : Regulations

In terms of supply issues, the expert noted in their interview that Brexit has introduced new regulations, causing suppliers to face difficulties when importing goods into the EU. Delays caused by customs and the potential for perishable goods to spoil upon arrival contribute to stockouts. Compliance with new regulations requires investment in training, additional staff, and increased shipping costs, which can further strain inventory management.

These challenges can lead to stockouts, as customs regulations cause delivery delays. Additionally, perishable foods may need to be discarded upon arrival at supermarkets. However, improved logistics and understanding of compliances with regulations lead to more reliable inventory levels and reduced chances of stockouts.

Quality control is also a concern, even if products haven't expired, it may not be in the supermarket's best interest to display items that don't reflect their high standards.

Adapting to these regulations demands significant effort and resources. Companies may need to invest in compliance training, hire additional staff for customs clearance, and deal with increased shipping costs. However, failing to adapt can result in frequent stockouts and lost sales.

Section 2.3: Conclusions & Further Work

1. Conclusions

The insights gained from analysing key metrics such as product return rates and technology integration error rates have significant implications for improving customer satisfaction and operational efficiency in the supermarket industry. Ensuring good quality control measures before products reach the shelves can significantly reduce returns, thereby enhancing customer satisfaction. Implementing regular random product inspections and fostering partnerships with reliable suppliers can mitigate the risk of defects, leading to fewer returns and higher customer loyalty. According to Progressive Grocer (2024), 40% of grocery retailers have trouble adopting new technologies because of the high costs and complexity, further highlighting the importance of effective technology integration.

Technology integration remains a critical issue impacting both operational efficiency and customer experience. Selecting technologies that are compatible with existing supermarket systems can significantly reduce integration errors and enhance overall operational

performance. Longitudinal studies tracking the impact of implemented changes on key metrics over time will provide insights into their long term effectiveness and areas needing adjustment. Addressing these technological challenges is crucial for maintaining a competitive edge in an increasingly digital market. Seamless technology integration can improve overall customer experience, reduce operational costs, and increase staff efficiency, thereby contributing to the supermarket's long term success.

Moreover, focusing on improving product quality and technology integration can have a broader impact on the supermarket's reputation and customer loyalty. Enhanced product quality not only reduces returns but also builds customer trust and loyalty over time.

Similarly, successful technology integration can streamline operations, making the shopping experience smoother and more efficient for customers. By learning from other supermarkets that have successfully implemented similar strategies, we can draw valuable lessons and best practices to apply in our own context.

2. Further Work

Future work should explore additional metrics, such as the Customer Effort Score (CES), to gain a more overall understanding of customer interactions. Implementing CES through digital feedback systems or in store forms can provide valuable insights into areas needing improvement. This metric measures the ease with which customers can interact with the supermarket and can highlight specific areas where the shopping experience can be enhanced to boost customer satisfaction. Unfortunately, there is no current measures to really go by thus far, and the ones I have come across are rather indirect, so further work into scores like CES would be beneficial.

Refining causal factors like supplier performance scorecards is also a strong suggestion of mine. Developing detailed scorecards for suppliers can help track and analyse the quality and reliability of deliveries, providing insights into supplier related product defects and stockouts. By including specific criteria such as delivery timeliness, product quality, and responsiveness, and conducting regular reviews and adjustments, supermarkets can ensure that these scorecards remain relevant and useful. Examples of supermarkets that have successfully used similar tools to improve supplier performance can provide valuable insights and best practices.

Conducting longitudinal studies to track the impact of implemented changes on key metrics over time will provide insights into their long term effectiveness and areas needing adjustment. Detailed data collection and analysis plans should be developed to monitor various metrics such as product return rates, customer satisfaction, and operational efficiency. Regular tracking and analysis will enable supermarkets to make informed decisions and continuously improve their operations.

My suggestion is to implement AI driven analytics to predict supplier performance and potential issues. This method can offer more precise and actionable insights compared to traditional approaches mentioned within the chapter. Additionally, conducting longitudinal studies to track the impact of changes on key metrics over time will provide valuable data on the longevity of these implemented strategies.

Recognising the interplay between different values and metrics can reveal chain reactions and benefits that drive further improvements in the supermarket industry, ensuring sustainable growth and success.

Finally, pursuing these identified areas for further investigation could lead to significant advancements in the supermarket industry. Improved supplier performance and customer interaction metrics could enhance operational efficiency, reduce costs, and increase customer satisfaction. These changes would not only benefit the supermarket in question but also set a standard for the industry, potentially influencing competitors' strategies and driving broader industry innovation. Enhanced customer satisfaction and operational efficiency ensure sustainable growth and success, positioning the supermarket as a leader in the industry.

APPENDIX

- Bloomenthal, A. (2024) Gross profit margin: formula and what it tells you. Available at: https://www.investopedia.com/terms/g/gross_profit_margin.asp
- Causon, J. (2017) 'COMMENT: Why has Aldi overtaken Waitrose and M&S?', Retail Gazette, 16 August. Available at: <https://www.retailgazette.co.uk/blog/2017/08/aldi-overtaken-waitrose-ms/>
- Centres, C. (2022) Customer Service ranked in Mystery shopping report, contact. Available at: <https://contact-centres.com/customer-service-ranked-in-new-mystery-shopping-report/>
- Chen, C.-Y. et al. (2017) 'The effects of community involvement, brand identity, and brand valuation on marketing performance', International Journal of Research in Business Studies and Management, 4(9), pp. 10–24. Available at: <https://www.ijrbsm.org/papers/v4-i9/2.pdf>
- Coca-Cola Retailing Research Councils (CCRRCC). (2017) The Importance of Employee Engagement. Available at: <https://www.ccrrc.org/2017/08/29/importance-employee-engagement/>
- Course Hero (2024) Tesco: Contemporary Issues - Brexit. Available at: <https://www.coursehero.com/file/174469924/TESCO-CONTEMPORARY-ISSUES-BREXIT-filedocx/>
- CustomerGauge (2022). Benchmarks. Available at: https://customergauge.com/benchmarks/industry/consumer_brands/supermarkets
- Fic, S.R. (2023) The paradox of automation: Self-Checkouts and staffing struggles in UK supermarkets. Available at: <https://www.linkedin.com/pulse/paradox-automation-self-checkouts-staffing-struggles-uk-roy-fic-xdgze>
- GLOBOSE TECHNOLOGY SOLUTIONS PRIVATE LIMITED (2024) Supermarket Sales Dataset: Insights from Three Branches. Available at: <https://gts.ai/dataset-download/supermarket-sales-dataset/>
- Hunter, D. (2024) Grocery Digital Transformation: Navigating the Challenges - Rizing. Available at: <https://rizing.com/consumer-industries-fashion-retail-grocery/grocery-digital-transformation-navigating-challenges/>
- Institute for government (2023) Supply chain problems. Available at: <https://www.instituteforgovernment.org.uk/explainer/supply-chain-problems>
- Katrina (2022) Employee turnover rates by industry comparison. Available at: <https://www.e-days.com/news/employee-turnover-rates-an-industry-comparison>
- Lee, C. and The University of Cambridge Institute for Sustainability Leadership (2022) Supermarkets and community well-being: Developing a framework to guide investment,

implementation, and impact of community actions. Available at: https://www.cisl.cam.ac.uk/files/supermarkets_and_community_wellbeing_framework_261022.pdf

Lorena (2024) Use On-Time delivery metrics to improve customer satisfaction. Available at: <https://optimoroute.com/on-time-delivery-metric/#what-is-it>

McTaggart, J. (2024) 'SPECIAL REPORT: 2024 Grocery Tech Trends,' Progressive Grocer, 2 July. Available at: <https://progressivegrocer.com/special-report-2024-grocery-tech-trends>

Morrisons (2022) Annual Report and Financial Statements for the 52 weeks ended 30 October 2022. Available at: <https://www.morrisons-corporate.com/globalassets/corporatesite/about-us/biographies/final-annual-report.pdf>

Office for National Statistics (2024). Available at: <https://www.ons.gov.uk/>

OpenAI (2024) ChatGPT [Large language model]. Available at: <https://chat.openai.com/chat>

Oxford English Dictionary (2016). Available at: <http://www.oed.com/>

Pavlou, C. (2024) How to calculate employee turnover rate. Available at: <https://resources.workable.com/tutorial/calculate-employee-turnover-rate#:~:text=Employee%20turnover%20rate%20is%20a,retention%20and%20overall%20management%20effectiveness>

Perplexity (2024) Perplexity AI [Artificial Intelligence Search Engine]. Available at: <https://www.perplexity.ai/>

Sainsbury, J., plc (2019) Strategic Report J Sainsbury plc Annual Report 2019. Available at: <https://www.about.sainsburys.co.uk/~/media/Files/S/Sainsburys/documents/reports-and-presentations/2019/ar-docs/sainsburys-ar2019-financialkpis.pdf>

Sainsbury, J.J. and Sainsbury, M.A. (2023) Annual Report and Financial Statements 2023. Available at: <https://www.about.sainsburys.co.uk/~/media/Files/S/Sainsburys/documents/reports-and-presentations/2023/annual-report-2023/annual-report-and-financial-statements-2023.pdf>

Shamsher, R. (2021) 'The effect of store image on store loyalty mediated by customer satisfaction and trust,' Business Perspective Review, 3(1), pp. 54–72. <https://doi.org/10.38157/businessperspectivereview.v3i1.330>

Stavrou, A. (2024) 'Tesco and Sainsbury's hit with technical issues disrupting deliveries,' The Independent, 16 March. Available at: <https://www.independent.co.uk/news/uk/home-news/tesco-sainsbury-shopping-it-glitch-b2513807.html>

SurveyMonkey (2024) NPS calculator—work out your Net Promoter Score. Available at: <https://uk.surveymonkey.com/mp/nps-calculator/>

Tesco (2023) Tesco Annual Report & Financial Statements 2023. Available at: <https://www.tescoplccom/media/u1wlq2qf/tesco-plc-annual-report-2023.pdf>

Tesco Employee Turnover - 1426 words | Cram (2020). Available at: <https://www.cram.com/essay/Relationship-Between-Turnover-And-Organizational-Productivity-Case/FCJ965YHWU>

Weaver, A. R. (2013) Exploring Community Engagement in Supermarket Recruitment Processes. Available at: <https://krex.k-state.edu/server/api/core/bitstreams/89066ef2-f19f-4884-93c1-33a1970a5b58/content>

Whitfield, D. (2024) 'How retailers can increase employee retention in 2024', HR Data Hub. Available at: <https://www.hrdatabhub.com/blog/increase-employee-retention-in-retail>

Chapter 4

4.1 Format data into a more analysable form

The work in this section directly reflects all suggestions in Chapter 3 , regarding formatting data in a more analysable form. However, I will continue to state my objects upon code execution, simply for clarity and linkage between Chapter 3 (section 3.5) and this notebook.

The code below reads a CSV file named NI_47110.csv, which is the csv I am using for my dataset, into a pandas data frame and shows the first few rows and a summary of the data. data.head() displays the first 5 rows, giving an overview of the columns and their values. data.info() provides details about the data frame, including the number of rows , the types of each column, and how many non-null values each column contains. For example, the Current Year column has no missing values and is of type int64, while columns like Fixed Assets(CUR) can handle missing values and are of type Int64. int64 is a standard integer type from NumPy and does not support missing values, so it's used for columns where all data is present and valid integers. On the other hand, Int64 is a pandas extension type that can handle missing values, represented as pd.NA. , which will be explained more later

```
In [754]: import pandas as pd

file = 'NI_47110.csv'
data = pd.read_csv(file)

print(data.head())
print(data.info())

    Company Name Company Number \
0  727 YOUR STORE LIMITED      NI620056
1  727 YOUR STORE LIMITED      NI620056
2  727 YOUR STORE LIMITED      NI620056
3  727 YOUR STORE LIMITED      NI620056
4    77 RETAIL 2 LIMITED      NI636843

                    Registered office address Company status \
0  Ruddell & Company, 50 Portmore Street, Portad...      Active
1  Ruddell & Company, 50 Portmore Street, Portad...      Active
2  Ruddell & Company, 50 Portmore Street, Portad...      Active
3  Ruddell & Company, 50 Portmore Street, Portad...      Active
4  Unit 1 Enler Centre Craigleith Drive, Dundonal...      Active

    Company type Incorporated on \
0  Private limited Company  27-Aug-13
1  Private limited Company  27-Aug-13
2  Private limited Company  27-Aug-13
3  Private limited Company  27-Aug-13
4  Private limited Company  03-Mar-16

    Nature of business (SIC) 0 \
0  47110 - Retail sale in non-specialised stores ...
1  47110 - Retail sale in non-specialised stores ...
2  47110 - Retail sale in non-specialised stores ...
3  47110 - Retail sale in non-specialised stores ...
4  47110 - Retail sale in non-specialised stores ...

    Nature of business (SIC) 1 Nature of business (SIC) 2 \
0                NaN          NaN
1                NaN          NaN
2                NaN          NaN
3                NaN          NaN
4                NaN          NaN

    Nature of business (SIC) 3 ... Creditors(CUR) Equity(CUR) Employees(CUR) \
0                NaN   ...     18,364      6,425          5
1                NaN   ...     16,630      3,648          6
2                NaN   ...     18,330      509  Not Available
3                NaN   ...     54,157     17,323  Not Available
```

4 NaN ... 586,318 100 22

	Previous Year	Fixed Assets(PREV)	Net Fixed Assets(PREV)	\
0	2019	13,830	3,648	
1	2018	17,288	509	
2	2017	17,859	17,323	
3	31/08/2016	22,323	28,415	
4	2019	399,066	36,028	

	Current Assets(PREV)	Creditors(PREV)	Equity(PREV)	Employees(PREV)
0	17,578	11,130	3,648	5
1	3,558	2,007	509	6
2	106,263	52,642	17,323	Not Available
3	59,329	30,816	28,415	Not Available
4	213,046	571,869	36,028	18

[5 rows x 72 columns]

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1315 entries, 0 to 1314

Data columns (total 72 columns):

#	Column	Non-Null Count	Dtype
0	Company Name	1315	non-null object
1	Company Number	1315	non-null object
2	Registered office address	1315	non-null object
3	Company status	1315	non-null object
4	Company type	1315	non-null object
5	Incorporated on	1315	non-null object
6	Nature of business (SIC) 0	1315	non-null object
7	Nature of business (SIC) 1	268	non-null object
8	Nature of business (SIC) 2	82	non-null object
9	Nature of business (SIC) 3	27	non-null object
10	No. Officers	1315	non-null object
11	Officer Name 1	1315	non-null object
12	Role 1	1315	non-null object
13	Status 1	1315	non-null object
14	DOB 1	1315	non-null object
15	Appointed On 1	1315	non-null object
16	Nationality 1	1315	non-null object
17	Country of Residence 1	1315	non-null object
18	Officer Name 2	1315	non-null object
19	Role 2	1315	non-null object
20	Status 2	1315	non-null object
21	DOB 2	1315	non-null object
22	Appointed On 2	1315	non-null object
23	Nationality 2	1315	non-null object
24	Country of Residence 2	1315	non-null object
25	Officer Name 3	1315	non-null object
26	Role 3	1315	non-null object
27	Status 3	1315	non-null object
28	DOB 3	1315	non-null object
29	Appointed On 3	1315	non-null object
30	Nationality 3	1315	non-null object
31	Country of Residence 3	1315	non-null object
32	Officer Name 4	1315	non-null object
33	Role 4	1315	non-null object
34	Status 4	1315	non-null object
35	DOB 4	1315	non-null object
36	Appointed On 4	1315	non-null object
37	Nationality 4	1315	non-null object
38	Country of Residence 4	1315	non-null object
39	Officer Name 5	1315	non-null object
40	Role 5	1315	non-null object
41	Status 5	1315	non-null object
42	DOB 5	1315	non-null object
43	Appointed On 5	1315	non-null object
44	Nationality 5	1315	non-null object
45	Country of Residence 5	1315	non-null object
46	Officer Name 6	1315	non-null object
47	Role 6	1315	non-null object
48	Status 6	1315	non-null object
49	DOB 6	1315	non-null object
50	Appointed On 6	1315	non-null object
51	Nationality 6	1315	non-null object
52	Country of Residence 6	1315	non-null object
53	iXBRL Available?	1315	non-null object
54	Link to Report	1315	non-null object
55	Report Number	1315	non-null int64
56	Report Type	1315	non-null object
57	Current Year	1315	non-null object
58	Currency	1315	non-null object
59	Fixed Assets(CUR)	1315	non-null object
60	Net Fixed Assets(CUR)	1315	non-null object

```

61 Current Assets(CUR)      1315 non-null   object
62 Creditors(CUR)          1315 non-null   object
63 Equity(CUR)             1315 non-null   object
64 Employees(CUR)          1315 non-null   object
65 Previous Year           1315 non-null   object
66 Fixed Assets(PREV)      1315 non-null   object
67 Net Fixed Assets(PREV)  1315 non-null   object
68 Current Assets(PREV)    1315 non-null   object
69 Creditors(PREV)          1315 non-null   object
70 Equity(PREV)            1315 non-null   object
71 Employees(PREV)          1315 non-null   object
dtypes: int64(1), object(71)
memory usage: 739.8+ KB
None

```

```

In [755]: import numpy as np
# I imported NumPy library for the number operations and for when I am working with arrays

columns_to_drop = [
    'Officer Name 1', 'Role 1', 'Status 1', 'DOB 1', 'Appointed On 1', 'Nationality 1', 'Country of Residence 1',
    'Officer Name 2', 'Role 2', 'Status 2', 'DOB 2', 'Appointed On 2', 'Nationality 2', 'Country of Residence 2',
    'Officer Name 3', 'Role 3', 'Status 3', 'DOB 3', 'Appointed On 3', 'Nationality 3', 'Country of Residence 3',
    'Officer Name 4', 'Role 4', 'Status 4', 'DOB 4', 'Appointed On 4', 'Nationality 4', 'Country of Residence 4',
    'Officer Name 5', 'Role 5', 'Status 5', 'DOB 5', 'Appointed On 5', 'Nationality 5', 'Country of Residence 5',
    'Officer Name 6', 'Role 6', 'Status 6', 'DOB 6', 'Appointed On 6', 'Nationality 6', 'Country of Residence 6'
]
# columns_to_drop is a list that was created to specify the columns i wanted to remove from the dataframe.
# This is solely for the officer-related columns

# I then removed the specified columns from the data frame and create a new dataframe with the remaining data,
data_cleaned = data.drop(columns=columns_to_drop)

data_cleaned.head()
# Below is the output of this function which displays the first 5 rows of the cleaned data frame, to verify the

```

	Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	... Creditors(CU)	
0	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	18,3
1	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	16,6
2	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	18,3
3	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	54,1
4	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleith Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	586,3

5 rows × 30 columns

I removed the officer data as my analysis is centered on understanding the causes behind decision maker values. Officer details (such as names, roles, and personal information) are not directly relevant to this type of analysis unless I was specifically studying the impact of individual officers on company outcomes. Officer details like names, dates of birth, nationality, and country of residence are often more relevant for headoffice related analyses rather than for understanding industry performance metrics. By removing officer details, I have simplified the dataset, making it easier to focus on the key variables that align with my project, thus avoiding distractions from irrelevant data.

```
In [757]: # I made a list of additional columns to drop, mainly linked to the report
additional_columns_to_drop = [
    'iXBRL Available?',
    'Link to Report',
    'Report Number',
    'Report Type'
]

# Again, datacleaned is the dataframe I am applying this to, using the drop() function I removed these columns
# data frame, yet applied the name of a data frame which is the same as the one I am working on to avoid confusion
data_cleaned = data_cleaned.drop(columns=additional_columns_to_drop)

# Here I am displaying the first 20 rows of the cleaned data to ensure that it has been updated properly, using
data_cleaned.head(20)
```

Out[757]:

	Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	... Creditors(C)
0	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
1	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
2	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
3	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
4	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
5	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
6	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
7	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
8	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
9	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
10	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...

11	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	Not Avail.
12	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	Not Avail.
13	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	Not Avail.
14	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	Not Avail.
15	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	Not Avail.
16	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	47,
17	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	112,
18	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	101,
19	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	118,

20 rows × 26 columns



Again, the report columns are very unlikely to provide useful insights, especially since they are referencing external data reports, which was helpful in the raw excel file, but unhelpful when it comes to analysis. Thus, removing these columns reduced the complexity of my dataset, without removing any valuable information, making it easier to focus on the data that matters.

```
In [759]: # Here, I am counting the total number of rows in the data frame before any rows are removed, to make a total
# of rows removed to ensure it matches with the COUNTIF in my excel sheet.
initial_row_count = data_cleaned.shape[0]

# I defined a list of financial columns that I want to check, since having not applicable within them can cause
financial_columns = [
    'Fixed Assets(CUR)', 'Net Fixed Assets(CUR)', 'Current Assets(CUR)',
    'Creditors(CUR)', 'Equity(CUR)', 'Employees(CUR)',
    'Fixed Assets(PREV)', 'Net Fixed Assets(PREV)', 'Current Assets(PREV)',
    'Creditors(PREV)', 'Equity(PREV)', 'Employees(PREV)'
]

# The condition below checks if all values across these columns in a row are 'Not Applicable'. The '~' negates
# meaning I keep rows that do not meet this condition, effectively removing those that do. The all() function checks
# values along a certain axis are True, axis=1 specifies the axis along which to apply the function i.e. all in
# than axis=0 meaning that it would be down the rows in each column
data_cleaned = data_cleaned[~(data_cleaned[financial_columns] == 'Not Applicable').all(axis=1)]

# The .shape function below counts the rows remaining in the data frame after the removal of rows, to subtract the
# as mentioned at the start of this code
final_row_count = data_cleaned.shape[0]

# The rows_removed subtraction calculates how many rows were removed by comparing the row count before the operation
# The result of this gives me the total number of rows that were removed because they contained 'Not Applicable'
# finance related columns. This number of 149 reflects the number in my COUNTIF within excel, showing successful
rows_removed = initial_row_count - final_row_count
rows_removed
```

Out[759]: 149

In [760]: # Display the first 30 rows of the cleaned DataFrame to check for the removal of columns
 data_cleaned.head(30)

Out[760]:

	Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	...
0	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
1	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
2	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
3	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	27-Aug-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
4	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
5	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
6	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
7	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
8	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	03-Mar-16	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
9	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
10	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
11	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
12	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	15-Nov-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
13	A & C MAGINN	NI057968	42 Mill Street, Gilford, Co	Active	Private limited	26-Jan-06	47110 - Retail sale in non-	NaN	NaN	NaN	...

	LIMITED		Down, BT66 6HQ	Company	specialised stores ...				
14	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
15	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
16	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active Private limited Company	26-Jan-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
17	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
18	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
19	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
20	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active Private limited Company	22-May-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
22	ACE SCAFFOLDING BALLYMONEY LIMITED	NI647113	63 Garryduff Road Garryduff Road, Ballymoney, ...	Active Private limited Company	26-Jul-17	43991 - Scaffold erection	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN ...
23	ACE SCAFFOLDING BALLYMONEY LIMITED	NI647113	63 Garryduff Road Garryduff Road, Ballymoney, ...	Active Private limited Company	26-Jul-17	43991 - Scaffold erection	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN ...
24	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	17-Sep-15	47110 - Retail sale in non-specialised stores ...	NaN	NaN ...
25	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	17-Sep-15	47110 - Retail sale in non-specialised stores ...	NaN	NaN ...
26	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	17-Sep-15	47110 - Retail sale in non-specialised stores ...	NaN	NaN ...
27	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	17-Sep-15	47110 - Retail sale in non-specialised stores ...	NaN	NaN ...
28	ADRIAN & MAURA MCNALLY LTD	NI056696	15 Main Street, Middletown, Armagh, BT60 4JF	Active Private limited Company	04-Oct-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...

29	AEC RETAIL LIMITED	NI647001	Maghera Street, Kilrea, Coleraine, North...	Active	Private limited Company	20-Jul-17	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
30	AJ TRADING LIMITED	NI656774	32 Chapel Road, Killeavy, Newry, Down, Northern...	Active	Private limited Company	31-Oct-18	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...

30 rows × 12 columns

The decision to remove the rows where 'Not Applicable' was found in all finance-related column entries was driven by the understanding that if a business in the dataset had not provided these financial values, there would be little to gain from including them in the analysis. These entries clutter the dataset and add unnecessary complexity without offering meaningful insights. Additionally, pandas can be unforgiving when handling non numeric data in numeric columns, which complicates analysis. In previous attempts to analyse the dataset, the presence of these 'Not Applicable' entries caused issues, leading to analysis failures. Therefore, the first step I took was to remove rows that lacked relevant financial information to ensure a cleaner and more manageable dataset for subsequent analysis.

```
In [762]: # Check the data types of each column in the DataFrame make sure no unnamed ones now just incase there was some
# when manipulating the columns, as had been made aware of via youtube tutorials
data_cleaned.dtypes
```

```
Out[762]: Company Name          object
Company Number        object
Registered office address    object
Company status         object
Company type           object
Incorporated on        object
Nature of business (SIC) 0    object
Nature of business (SIC) 1    object
Nature of business (SIC) 2    object
Nature of business (SIC) 3    object
No. Officers           object
Current Year            object
Currency                object
Fixed Assets(CUR)        object
Net Fixed Assets(CUR)      object
Current Assets(CUR)       object
Creditors(CUR)           object
Equity(CUR)              object
Employees(CUR)           object
Previous Year            object
Fixed Assets(PREV)        object
Net Fixed Assets(PREV)      object
Current Assets(PREV)       object
Creditors(PREV)           object
Equity(PREV)              object
Employees(PREV)           object
dtype: object
```

Next, I focused on further cleaning the financial columns because there were still string entries like "Not Available" present in the dataset. These entries would obviously cause issues, as they prevent any meaningful numeric operations. Since the data types of these columns are currently set as objects, as shown above, due to the presence of string data, performing numeric operations would be invalid. By replacing 'Not Available' and '-' (which was in the data set as a place holder for a null entry but not picked up in initial excel sheet) with NA (recognised by python as a null entry), I am able to ensure the financial columns of the dataset were better for analysis.

```
In [764]: # I defined the financial columns to be cleaned, via a list that contained all the financial columns i wanted
# to manipulate
financial_columns = [
    'Fixed Assets(CUR)', 'Net Fixed Assets(CUR)', 'Current Assets(CUR)',
    'Creditors(CUR)', 'Equity(CUR)', 'Employees(CUR)',
    'Fixed Assets(PREV)', 'Net Fixed Assets(PREV)', 'Current Assets(PREV)',
    'Creditors(PREV)', 'Equity(PREV)', 'Employees(PREV)'
]

# Replace 'Not Available' and '-' with NaN, remove commas, then convert to numeric
# As a side note, NaN is used to represent missing or undefined values in numeric data
# Remove commas from the data and convert the strings to floats (which will be updated later as i wanted to see
# Commas were originally used as thousand separators but they need to be removed for proper numeric conversion
# The .str.replace(',', '') method removes commas from the string representations of the numbers
for column in financial_columns:
    data_cleaned[column] = data_cleaned[column].replace(['Not Available', '-'], np.nan)
    data_cleaned[column] = data_cleaned[column].str.replace(',', '').astype(float)
# Explanation of the loop:
# The loop iterates over each column in the financial_columns list
```

```
# For each column, it first replaces any occurrences of 'Not Available' or '-' with NaN, which pandas recognises.  
# It then removes any commas from the data and converts the resulting strings to floats
```

```
In [765]: # Check data types to ensure they are now numeric  
data_cleaned[financial_columns].dtypes  
  
# Display the first 30 rows to verify  
data_cleaned[financial_columns].head(30)
```

	Fixed Assets(CUR)	Net Fixed Assets(CUR)	Current Assets(CUR)	Creditors(CUR)	Equity(CUR)	Employees(CUR)	Fixed Assets(PREV)	Net Fixed Assets(PREV)	C Assets(f
0	11064.0	6425.0	64782.0	18364.0	6425.0	5.0	13830.0	3648.0	11
1	13830.0	3648.0	17578.0	16630.0	3648.0	6.0	17288.0	509.0	3
2	17288.0	509.0	3558.0	18330.0	509.0	NaN	17859.0	17323.0	106
3	17859.0	17323.0	106263.0	54157.0	17323.0	NaN	22323.0	28415.0	59
4	337671.0	167892.0	418863.0	586318.0	100.0	22.0	399066.0	36028.0	213
5	399066.0	36028.0	213046.0	571869.0	100.0	18.0	472276.0	23645.0	204
6	472276.0	23645.0	204476.0	647918.0	100.0	24.0	548424.0	3352.0	168
7	548424.0	3352.0	168766.0	714408.0	100.0	23.0	625907.0	60543.0	210
8	625907.0	60543.0	210231.0	773282.0	100.0	23.0	625907.0	60543.0	210
9	61858.0	169.0	75566.0	NaN	169.0	11.0	79385.0	6199.0	18
10	NaN	6198.0	19687.0	NaN	6198.0	19.0	NaN	1144.0	24
11	NaN	1145.0	24089.0	NaN	1145.0	13.0	NaN	1153.0	25
12	NaN	1153.0	23228.0	NaN	1153.0	13.0	NaN	13959.0	38
13	NaN	NaN	35737.0	NaN	2.0	0.0	NaN	NaN	37
14	NaN	NaN	37387.0	NaN	2.0	0.0	NaN	NaN	38
15	NaN	NaN	38701.0	NaN	2.0	0.0	NaN	NaN	5
16	NaN	3915.0	51118.0	47203.0	3915.0	NaN	NaN	7150.0	12
17	29709.0	11600.0	149240.0	112369.0	2.0	14.0	39762.0	20494.0	88
18	39762.0	20494.0	88562.0	101605.0	2.0	14.0	51452.0	26078.0	106
19	51452.0	26078.0	100593.0	118186.0	2.0	14.0	61544.0	35836.0	92
20	61544.0	35836.0	92348.0	109023.0	2.0	14.0	76930.0	57775.0	88
22	352828.0	644741.0	432446.0	138538.0	644741.0	8.0	303674.0	515561.0	312
23	221883.0	271901.0	229093.0	271036.0	271901.0	NaN	221883.0	271901.0	229
24	561255.0	407049.0	78076.0	72133.0	1.0	23.0	463372.0	258864.0	70
25	463372.0	258864.0	70702.0	63235.0	1.0	NaN	289707.0	122050.0	52
26	289707.0	122050.0	52069.0	70655.0	1.0	NaN	82697.0	23911.0	20
27	82697.0	23911.0	20533.0	46183.0	1.0	NaN	37113.0	11956.0	58
28	NaN	239635.0	196165.0	NaN	239635.0	5.0	NaN	207184.0	145
29	38028.0	18232.0	150200.0	206460.0	100.0	0.0	49232.0	44150.0	70
30	17162.0	1952.0	63742.0	51910.0	1952.0	9.0	22883.0	3315.0	52

```
In [766]: # Convert the financial columns to integers  
for column in financial_columns:  
    # Convert to float first (to handle NaNs), then to int  
    data_cleaned[column] = data_cleaned[column].astype(float).astype('Int64')
```

```
In [767]: # Display the first 30 rows to verify  
data_cleaned[financial_columns].head(30)
```

Out[767..]

	Fixed Assets(CUR)	Net Fixed Assets(CUR)	Current Assets(CUR)	Creditors(CUR)	Equity(CUR)	Employees(CUR)	Fixed Assets(PREV)	Net Fixed Assets(PREV)	C Assets(f)
0	11064	6425	64782	18364	6425	5	13830	3648	
1	13830	3648	17578	16630	3648	6	17288	509	
2	17288	509	3558	18330	509	<NA>	17859	17323	1
3	17859	17323	106263	54157	17323	<NA>	22323	28415	
4	337671	167892	418863	586318	100	22	399066	36028	2
5	399066	36028	213046	571869	100	18	472276	23645	2
6	472276	23645	204476	647918	100	24	548424	3352	1
7	548424	3352	168766	714408	100	23	625907	60543	2
8	625907	60543	210231	773282	100	23	625907	60543	2
9	61858	169	75566	<NA>	169	11	79385	6199	
10	<NA>	6198	19687	<NA>	6198	19	<NA>	1144	
11	<NA>	1145	24089	<NA>	1145	13	<NA>	1153	
12	<NA>	1153	23228	<NA>	1153	13	<NA>	13959	
13	<NA>	<NA>	35737	<NA>	2	0	<NA>	<NA>	
14	<NA>	<NA>	37387	<NA>	2	0	<NA>	<NA>	
15	<NA>	<NA>	38701	<NA>	2	0	<NA>	<NA>	
16	<NA>	3915	51118	47203	3915	<NA>	<NA>	7150	1
17	29709	11600	149240	112369	2	14	39762	20494	
18	39762	20494	88562	101605	2	14	51452	26078	1
19	51452	26078	100593	118186	2	14	61544	35836	
20	61544	35836	92348	109023	2	14	76930	57775	
22	352828	644741	432446	138538	644741	8	303674	515561	3
23	221883	271901	229093	271036	271901	<NA>	221883	271901	2
24	561255	407049	78076	72133	1	23	463372	258864	
25	463372	258864	70702	63235	1	<NA>	289707	122050	
26	289707	122050	52069	70655	1	<NA>	82697	23911	
27	82697	23911	20533	46183	1	<NA>	37113	11956	
28	<NA>	239635	196165	<NA>	239635	5	<NA>	207184	1
29	38028	18232	150200	206460	100	0	49232	44150	
30	17162	1952	63742	51910	1952	9	22883	3315	

In [768..]: # This is just for visual testing to see if all years were in the same format, which clearly they were not
data_cleaned[['Current Year', 'Previous Year']].head(50)

Out[768..]

	Current Year	Previous Year
0	2020	2019
1	2019	2018
2	2018	2017
3	2017	31/08/2016
4	2020	2019
5	2019	2018
6	2018	2017
7	2017	2016
8	2016	2015
9	2020	2019
10	2019	2018
11	2018	2017
12	2017	2016
13	2021	2020
14	2020	2019

15	2019	2018
16	2018	2017
17	2020	2019
18	2019	2018
19	2018	2017
20	2017	2016
22	2020	2019
23	2018	2017
24	2020	2019
25	2019	2018
26	2018	2017
27	2017	2016
28	2021	2020
29	2020	2019
30	2020	2019
31	2019	2018
32	2021	2020
33	2020	2019
34	2021	2020
35	2020	2019
36	2019	2018
37	2018	2017
38	2021	2020
39	2020	2019
40	2019	2018
41	2020	2019
42	2019	2018
43	2018	Not Available
44	2020	2019
45	2019	2018
47	2021	2020
48	2020	Not Available
49	2021	2020
50	2020	2019
51	2019	2018

Clearly, as shown above, the ...Years columns are varied in the data they have. The fact that some are string values, others are YYYY and others are DD/MM/YYYY poses a big issue for analysis as they are as strings, which makes it difficult to perform operations like filtering, summarising, or comparing values. These inconsistencies also make it hard to do calculations, such as determining time intervals or spotting trends, because year and date values cannot be easily compared or used together. Additionally, entries that are not actual dates, like "Not Available," can cause errors or inaccurate results. To make the data easier to work with and ensure accurate analysis, I standardised these columns to a single format (YYYY) which simplifies processing and analysis.

```
In [770]: # checking unique values in "Current Year" and "Previous Year" columns
# These lines retrieve all the unique values present in the "Current Year" column of the data_cleaned data frame
# and thus by calling .unique(), I can see the different formats or values that exist in this column,
# such as years, full dates or strings etc
unique_current_year = data_cleaned['Current Year'].unique()
unique_previous_year = data_cleaned['Previous Year'].unique()

# line displays the unique values found in both the "Current Year" and "Previous Year" columns
unique_current_year, unique_previous_year
```

```
Out[770]: (array(['2020', '2019', '2018', '2017', '2016', '2021', '2022'],
      dtype=object),
 array(['2019', '2018', '2017', '31/08/2016', '2016', '2015', '2020',
       'Not Available', '31/10/2017', '2021', '31/01/2019', '31/08/2017'],
      dtype=object))
```

The output indicates that the data is still being treated as objects rather than the integers we need. I sought help from ChatGPT and CodeAcademy for this task, as the complexities of the data were beyond my current knowledge. However, with the detailed inline comments, I now understand how the process works and feel confident in applying these techniques in the future. It took several adjustments and numerous prompts to refine this code, making it more challenging than a straightforward date function, which is where the knowledge exceeded me.

```
In [772]: import re
# Import the regular expression module to assist with pattern matching in strings

# Define a function to clean the year data in the 'Current Year' and 'Previous Year' columns
def clean_year(value):
    # Step 1: Handle missing or non-applicable values
    if pd.isna(value) or value in ["Not Available", "Not Applicable"]:
        # If the value is NaN (missing) or explicitly labeled as "Not Available" or "Not Applicable",
        # we return pd.NA, which is pandas' representation of a missing value.
        return pd.NA

    # Step 2: Check if the value is a valid year number or string
    if isinstance(value, (int, float)) and 1000 <= value <= 9999:
        # Here, we check if the value is an integer or float, and whether it falls within the range of valid years.
        # If it does, the value is returned as an integer.
        return int(value)

    # Step 3: Handle cases where the value is a string
    if isinstance(value, str): # Check if the value is a string
        # Check if the string represents a 4-digit year
        if value.isdigit() and len(value) == 4:
            # If the string consists of digits only and is 4 characters long, it's assumed to be a valid year.
            # We convert it to an integer and return it.
            return int(value)

    # Step 4: Handle date strings in the DD/MM/YYYY format
    match = re.match(r'^\d{2}/\d{2}/(\d{4})$', value)
    if match:
        # This regular expression checks if the string matches the pattern of a date in DD/MM/YYYY format.
        # If it matches, the year part (YYYY) is extracted using match.group(1) and returned as an integer.
        return int(match.group(1))

    # Step 5: Default case for any other format
    return pd.NA # If the value doesn't match any of the above conditions, it is considered invalid, and we return pd.NA

# Apply the clean_year function to clean both 'Current Year' and 'Previous Year' columns
for col in ['Current Year', 'Previous Year']:
    if col in data_cleaned.columns: # Check if the column exists in the DataFrame
        # Apply the clean_year function to each entry in the column
        # This loop ensures that both 'Current Year' and 'Previous Year' columns are processed
        # The apply function applies clean_year to each value in the column.
        data_cleaned[col] = data_cleaned[col].apply(clean_year)

# Display the cleaned columns to verify the changes
# This command will output the first 50 rows of the cleaned 'Current Year' and 'Previous Year' columns,
# allowing us to verify that the cleaning function has worked correctly.
data_cleaned[['Current Year', 'Previous Year']].head(50)
```

	Current Year	Previous Year
0	2020	2019
1	2019	2018
2	2018	2017
3	2017	2016
4	2020	2019
5	2019	2018
6	2018	2017
7	2017	2016
8	2016	2015
9	2020	2019
10	2019	2018
11	2018	2017
12	2017	2016
13	2021	2020
14	2020	2019
15	2019	2018

16	2018	2017
17	2020	2019
18	2019	2018
19	2018	2017
20	2017	2016
22	2020	2019
23	2018	2017
24	2020	2019
25	2019	2018
26	2018	2017
27	2017	2016
28	2021	2020
29	2020	2019
30	2020	2019
31	2019	2018
32	2021	2020
33	2020	2019
34	2021	2020
35	2020	2019
36	2019	2018
37	2018	2017
38	2021	2020
39	2020	2019
40	2019	2018
41	2020	2019
42	2019	2018
43	2018	<NA>
44	2020	2019
45	2019	2018
47	2021	2020
48	2020	<NA>
49	2021	2020
50	2020	2019
51	2019	2018

As you can see, the code above successfully standardised the “Current Year” and “Previous Year” columns, ensuring that the data is consistent and ready for time based analysis. This standardisation allows for accurate and reliable analysis across different time periods.

With dates in mind, I focused on the “Incorporated On” column to ensure that the dates were standardised. Fortunately, this was much simpler to achieve using the `.to_datetime` function. I also decided to standardise the layout for all dates, originally in the DD/MM/YYYY format, to a consistent YYYY-MM-DD format. This change made it easier to visually verify that the conversion had been correctly applied.

```
In [774]: # Function to standardise the dates in the "Incorporated on" column
def standardize_date(value):

    if pd.isna(value) or value in ["Not Available", "Not Applicable"]:
        return pd.NA
    # if the value is missing (NaN) or labeled as "Not Available" or "Not Applicable", then return pd.NA to
    # Handling errors appropriately with try catch below
    try:
        # Converting the date to YYYY-MM-DD format via pandas object with 'coerce' error handler (which turns i
        # them I format the date as a string in 'YYYY-MM-DD' format using strftime.
        standardized_date = pd.to_datetime(value, errors='coerce').strftime('%Y-%m-%d')
    except Exception as e:
        return pd.NA # if convert fails then return NA

    # apply the function to standardize the "Incorporated on" column
    data_cleaned['Incorporated on'] = data_cleaned['Incorporated on'].apply(standardize_date)
```

```
# Display the first 50 rows to verify the changes
data_cleaned['Incorporated on'].head(50)
```

```
Out[774]: 0    2013-08-27
1    2013-08-27
2    2013-08-27
3    2013-08-27
4    2016-03-03
5    2016-03-03
6    2016-03-03
7    2016-03-03
8    2016-03-03
9    2005-11-15
10   2005-11-15
11   2005-11-15
12   2005-11-15
13   2006-01-26
14   2006-01-26
15   2006-01-26
16   2006-01-26
17   2009-05-22
18   2009-05-22
19   2009-05-22
20   2009-05-22
22   2017-07-26
23   2017-07-26
24   2015-09-17
25   2015-09-17
26   2015-09-17
27   2015-09-17
28   2005-10-04
29   2017-07-20
30   2018-10-31
31   2018-10-31
32   2019-02-11
33   2019-02-11
34   2019-09-06
35   2046-06-28
36   2046-06-28
37   2046-06-28
38   2018-01-15
39   2018-01-15
40   2018-01-15
41   2017-12-12
42   2017-12-12
43   2017-12-12
44   2006-03-07
45   2006-03-07
47   2019-04-04
48   2019-04-04
49   2013-07-18
50   2013-07-18
51   2013-07-18
Name: Incorporated on, dtype: object
```

Simply for visually checking that the actual data entries match this conversion

```
In [776]: data_cleaned.head(30)
```

	Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	...
0	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
1	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
2	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...
			Ruddell &				47110 -				

3	727 YOUR STORE LIMITED	NI620056	Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
4	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
5	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
6	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
7	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
8	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
9	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	2005-11-15	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
10	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	2005-11-15	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
11	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	2005-11-15	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
12	A & A HARDING & CO LTD	NI057200	9 Atlantic Avenue, Belfast, BT15 2HN	Active	Private limited Company	2005-11-15	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
13	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	2006-01-26	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
14	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	2006-01-26	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
15	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	2006-01-26	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
16	A & C MAGINN LIMITED	NI057968	42 Mill Street, Gilford, Co Down, BT66 6HQ	Active	Private limited Company	2006-01-26	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
17	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	2009-05-22	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
18	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	2009-05-22	47110 - Retail sale in non- specialised stores ...	NaN	NaN	NaN	...
			Unit 7,				47110 -				

19	A&G RASOOL LIMITED	NI072701	Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	2009-05-22	Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
20	A&G RASOOL LIMITED	NI072701	Unit 7, Woodborne Crescent, Belfast, BT11 9TG	Active	Private limited Company	2009-05-22	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
22	ACE SCAFFOLDING BALLYMONEY LIMITED	NI647113	63 Garryduff Road Garryduff Road, Ballymoney, ...	Active	Private limited Company	2017-07-26	43991 - Scaffold erection	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN ...
23	ACE SCAFFOLDING BALLYMONEY LIMITED	NI647113	63 Garryduff Road Garryduff Road, Ballymoney, ...	Active	Private limited Company	2017-07-26	43991 - Scaffold erection	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN ...
24	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	2015-09-17	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
25	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	2015-09-17	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
26	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	2015-09-17	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
27	ADC ENTERTAINMENTS (N.I.) LIMITED	NI633671	26 Ballymoney Road, Ballymena, Antrim, Norther...	Active — Active proposal to strike off	Private limited Company	2015-09-17	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
28	ADRIAN & MAURA MCNALLY LTD	NI056696	15 Main Street, Middletown, Armagh, BT60 4JF	Active	Private limited Company	2005-10-04	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
29	AEC RETAIL LIMITED	NI647001	39-43 Maghera Street, Kilrea, Coleraine, North...	Active	Private limited Company	2017-07-20	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...
30	AJ TRADING LIMITED	NI656774	32 Chapel Road, Killeavy, Newry, Down, Norther...	Active	Private limited Company	2018-10-31	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN ...

30 rows × 26 columns

```
In [777]: # Again, checking it matches the conversion and looking over entire dataset again to regain knowledge of wider issues to address when thinking about making the data into a more analysable format
print(data_cleaned.head())
print(data_cleaned.info())
```

	Company Name	Company Number	\
0	727 YOUR STORE LIMITED	NI620056	
1	727 YOUR STORE LIMITED	NI620056	
2	727 YOUR STORE LIMITED	NI620056	
3	727 YOUR STORE LIMITED	NI620056	
4	77 RETAIL 2 LIMITED	NI636843	

	Registered office address	Company status	\
0	Ruddell & Company, 50 Portmore Street, Portado...	Active	

```

1 Ruddell & Company, 50 Portmore Street, Portado... Active
2 Ruddell & Company, 50 Portmore Street, Portado... Active
3 Ruddell & Company, 50 Portmore Street, Portado... Active
4 Unit 1 Enler Centre Craigleith Drive, Dundonal... Active

    Company type Incorporated on \
0 Private limited Company 2013-08-27
1 Private limited Company 2013-08-27
2 Private limited Company 2013-08-27
3 Private limited Company 2013-08-27
4 Private limited Company 2016-03-03

    Nature of business (SIC) 0 \
0 47110 - Retail sale in non-specialised stores ...
1 47110 - Retail sale in non-specialised stores ...
2 47110 - Retail sale in non-specialised stores ...
3 47110 - Retail sale in non-specialised stores ...
4 47110 - Retail sale in non-specialised stores ...

    Nature of business (SIC) 1 Nature of business (SIC) 2 \
0 NaN NaN
1 NaN NaN
2 NaN NaN
3 NaN NaN
4 NaN NaN

    Nature of business (SIC) 3 ... Creditors(CUR) Equity(CUR) Employees(CUR) \
0 NaN ... 18364 6425 5
1 NaN ... 16630 3648 6
2 NaN ... 18330 509 <NA>
3 NaN ... 54157 17323 <NA>
4 NaN ... 586318 100 22

    Previous Year Fixed Assets(PREV) Net Fixed Assets(PREV) \
0 2019 13830 3648
1 2018 17288 509
2 2017 17859 17323
3 2016 22323 28415
4 2019 399066 36028

    Current Assets(PREV) Creditors(PREV) Equity(PREV) Employees(PREV)
0 17578 11130 3648 5
1 3558 2007 509 6
2 106263 52642 17323 <NA>
3 59329 30816 28415 <NA>
4 213046 571869 36028 18

[5 rows x 26 columns]
<class 'pandas.core.frame.DataFrame'>
Index: 1166 entries, 0 to 1311
Data columns (total 26 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Company Name     1166 non-null   object  
 1   Company Number   1166 non-null   object  
 2   Registered office address  1166 non-null   object  
 3   Company status    1166 non-null   object  
 4   Company type     1166 non-null   object  
 5   Incorporated on   1166 non-null   object  
 6   Nature of business (SIC) 0  1166 non-null   object  
 7   Nature of business (SIC) 1  239 non-null   object  
 8   Nature of business (SIC) 2  73 non-null   object  
 9   Nature of business (SIC) 3  25 non-null   object  
 10  No. Officers     1166 non-null   object  
 11  Current Year     1166 non-null   int64  
 12  Currency          1166 non-null   object  
 13  Fixed Assets(CUR) 828 non-null   Int64  
 14  Net Fixed Assets(CUR) 1067 non-null   Int64  
 15  Current Assets(CUR) 1104 non-null   Int64  
 16  Creditors(CUR)     968 non-null   Int64  
 17  Equity(CUR)       1164 non-null   Int64  
 18  Employees(CUR)    833 non-null   Int64  
 19  Previous Year     1144 non-null   object  
 20  Fixed Assets(PREV) 781 non-null   Int64  
 21  Net Fixed Assets(PREV) 1067 non-null   Int64  
 22  Current Assets(PREV) 1065 non-null   Int64  
 23  Creditors(PREV)    944 non-null   Int64  
 24  Equity(PREV)      1165 non-null   Int64  
 25  Employees(PREV)   828 non-null   Int64 

dtypes: Int64(12), int64(1), object(13)
memory usage: 259.6+ KB
None

```

At this stage, I noticed that 'Previous Year' was still an object, so I ensured to convert it to Int64 (due to the fact of if containing missing data)

via N/A). I then visually checked the conversion

```
In [779]: data_cleaned['Previous Year'] = data_cleaned['Previous Year'].astype('Int64')

In [780]: print(data_cleaned.head())
print(data_cleaned.info())

    Company Name Company Number \
0  727 YOUR STORE LIMITED      NI620056
1  727 YOUR STORE LIMITED      NI620056
2  727 YOUR STORE LIMITED      NI620056
3  727 YOUR STORE LIMITED      NI620056
4    77 RETAIL 2 LIMITED      NI636843

                    Registered office address Company status \
0  Ruddell & Company, 50 Portmore Street, Portado...      Active
1  Ruddell & Company, 50 Portmore Street, Portado...      Active
2  Ruddell & Company, 50 Portmore Street, Portado...      Active
3  Ruddell & Company, 50 Portmore Street, Portado...      Active
4  Unit 1 Enler Centre Craigleith Drive, Dundonal...      Active

    Company type Incorporated on \
0  Private limited Company  2013-08-27
1  Private limited Company  2013-08-27
2  Private limited Company  2013-08-27
3  Private limited Company  2013-08-27
4  Private limited Company  2016-03-03

    Nature of business (SIC) 0 \
0  47110 - Retail sale in non-specialised stores ...
1  47110 - Retail sale in non-specialised stores ...
2  47110 - Retail sale in non-specialised stores ...
3  47110 - Retail sale in non-specialised stores ...
4  47110 - Retail sale in non-specialised stores ...

    Nature of business (SIC) 1 Nature of business (SIC) 2 \
0            NaN          NaN
1            NaN          NaN
2            NaN          NaN
3            NaN          NaN
4            NaN          NaN

    Nature of business (SIC) 3 ... Creditors(CUR)  Equity(CUR) Employees(CUR) \
0            NaN   ...     18364      6425        5
1            NaN   ...    16630      3648        6
2            NaN   ...    18330       509      <NA>
3            NaN   ...    54157     17323      <NA>
4            NaN   ...    586318      100        22

    Previous Year Fixed Assets(PREV) Net Fixed Assets(PREV) \
0           2019        13830          3648
1           2018        17288          509
2           2017        17859        17323
3           2016        22323        28415
4           2019        399066        36028

    Current Assets(PREV) Creditors(PREV)  Equity(PREV) Employees(PREV)
0            17578        11130        3648        5
1            3558         2007        509        6
2           106263        52642        17323      <NA>
3            59329        30816        28415      <NA>
4           213046        571869        36028        18

[5 rows x 26 columns]
<class 'pandas.core.frame.DataFrame'>
Index: 1166 entries, 0 to 1311
Data columns (total 26 columns):
 #   Column           Non-Null Count Dtype
 --- 
 0   Company Name      1166 non-null  object
 1   Company Number    1166 non-null  object
 2   Registered office address  1166 non-null  object
 3   Company status    1166 non-null  object
 4   Company type      1166 non-null  object
 5   Incorporated on   1166 non-null  object
 6   Nature of business (SIC) 0  1166 non-null  object
 7   Nature of business (SIC) 1  239 non-null   object
 8   Nature of business (SIC) 2  73 non-null   object
 9   Nature of business (SIC) 3  25 non-null   object
 10  No. Officers      1166 non-null  object
 11  Current Year      1166 non-null  int64
 12  Currency          1166 non-null  object
 13  Fixed Assets(CUR) 828 non-null   Int64
```

```

14 Net Fixed Assets(CUR)      1067 non-null   Int64
15 Current Assets(CUR)       1104 non-null   Int64
16 Creditors(CUR)           968 non-null    Int64
17 Equity(CUR)              1164 non-null   Int64
18 Employees(CUR)            833 non-null    Int64
19 Previous Year             1144 non-null   Int64
20 Fixed Assets(PREV)        781 non-null    Int64
21 Net Fixed Assets(PREV)    1067 non-null   Int64
22 Current Assets(PREV)     1065 non-null   Int64
23 Creditors(PREV)           944 non-null    Int64
24 Equity(PREV)              1165 non-null   Int64
25 Employees(PREV)            828 non-null    Int64
dtypes: Int64(13), int64(1), object(12)
memory usage: 260.8+ KB
None

```

At this point, I noticed that No. Officers column was still present in the dataset. Since my analysis does not involve any information related to officers or the amount of them in a business, I determined that this data is not useful. Additionally, the officer numbers are stored as numerical strings, combining both integers and text, which makes them more complex and less practical for analysis. For a simpler analysis, I decided to drop this column.

```
In [782]: # Drop the "No. Officers" column from the data frame using drop method as before
if 'No. Officers' in data_cleaned.columns:
    data_cleaned = data_cleaned.drop(columns=['No. Officers'])

# Verify the change by displaying the first few rows of the cleaned data frame
data_cleaned.head()
```

Out[782]:

	Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	... Creditors(CU	
0	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	183
1	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	166
2	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	183
3	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	541
4	77 RETAIL 2 LIMITED	NI636843	Unit 1 Enler Centre Craigleith Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	...	5863

5 rows × 25 columns

```
In [783]: # Again, using the print command and .info function I could see the effect this change had on the data set and
print(data_cleaned.info())
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 1166 entries, 0 to 1311
Data columns (total 25 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   Company Name     1166 non-null   object  
 1   Company Number   1166 non-null   object  
 2   Registered office address 1166 non-null   object  
 3   Company status   1166 non-null   object  
 4   Company type    1166 non-null   object  
 5   Incorporated on 1166 non-null   object  
 6   Nature of business (SIC) 0 1166 non-null   object  
 7   Nature of business (SIC) 1 239 non-null   object  
 8   Nature of business (SIC) 2 73 non-null   object  
 9   Nature of business (SIC) 3 25 non-null   object  
 10  Current Year    1166 non-null   int64  
 11  Currency        1166 non-null   object  
 12  Fixed Assets(CUR) 828 non-null   Int64  
 13  Net Fixed Assets(CUR) 1067 non-null   Int64  
 14  Current Assets(CUR) 1104 non-null   Int64  
 15  Creditors(CUR)   968 non-null   Int64  
 16  Equity(CUR)     1164 non-null   Int64  
 17  Employees(CUR)  833 non-null   Int64  
 18  Previous Year   1144 non-null   Int64  
 19  Fixed Assets(PREV) 781 non-null   Int64  
 20  Net Fixed Assets(PREV) 1067 non-null   Int64  
 21  Current Assets(PREV) 1065 non-null   Int64  
 22  Creditors(PREV)  944 non-null   Int64  
 23  Equity(PREV)    1165 non-null   Int64  
 24  Employees(PREV)  828 non-null   Int64  
dtypes: Int64(13), int64(1), object(11)
memory usage: 251.6+ KB
None

```

```
In [784]: # saving data to csv to ensure I have a backup and also to look more closely at the data available before proceeding
data_cleaned.to_csv('cleaned_data_for_review.csv', index=False)
```

```
In [785]: print(data_cleaned.info())
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 1166 entries, 0 to 1311
Data columns (total 25 columns):
 #   Column           Non-Null Count Dtype  
--- 
 0   Company Name     1166 non-null   object  
 1   Company Number   1166 non-null   object  
 2   Registered office address 1166 non-null   object  
 3   Company status   1166 non-null   object  
 4   Company type    1166 non-null   object  
 5   Incorporated on 1166 non-null   object  
 6   Nature of business (SIC) 0 1166 non-null   object  
 7   Nature of business (SIC) 1 239 non-null   object  
 8   Nature of business (SIC) 2 73 non-null   object  
 9   Nature of business (SIC) 3 25 non-null   object  
 10  Current Year    1166 non-null   int64  
 11  Currency        1166 non-null   object  
 12  Fixed Assets(CUR) 828 non-null   Int64  
 13  Net Fixed Assets(CUR) 1067 non-null   Int64  
 14  Current Assets(CUR) 1104 non-null   Int64  
 15  Creditors(CUR)   968 non-null   Int64  
 16  Equity(CUR)     1164 non-null   Int64  
 17  Employees(CUR)  833 non-null   Int64  
 18  Previous Year   1144 non-null   Int64  
 19  Fixed Assets(PREV) 781 non-null   Int64  
 20  Net Fixed Assets(PREV) 1067 non-null   Int64  
 21  Current Assets(PREV) 1065 non-null   Int64  
 22  Creditors(PREV)  944 non-null   Int64  
 23  Equity(PREV)    1165 non-null   Int64  
 24  Employees(PREV)  828 non-null   Int64  
dtypes: Int64(13), int64(1), object(11)
memory usage: 251.6+ KB
None

```

While reviewing the cleaned dataset in Excel, I noticed that it still contained the current year and previous year columns, which caused almost duplicate data across multiple rows. This happens because the dataset includes 'Previous Year' data for each company, even though the same information is already available in 'Current Year' rows, just one row below each. In Python, we don't need these "Previous Year" columns because the data for previous years can be easily accessed by simply referencing the appropriate row. These extra columns just add clutter, making the dataset harder to manage and analyse. They also lead to unnecessary visualisations that don't provide new insights. Further research into the dataset leads me to believe that the layout is for the access within Excel, rather than a programming language. By removing the 'Previous Year' column and all associated columns, we can make the dataset more straightforward and easier to work with. This is especially useful in Python, where we can dynamically access any year's data without

needing these redundant columns. Given the large amount of company data, keeping these extra columns only increases the chances of confusion and errors. Removing them will make the dataset cleaner and more efficient for analysis, without losing any important information. In summary, dropping these 'Previous Year' columns simplifies the dataset, reduces the risk of mistakes, and makes it easier to analyse the data, while still allowing full access to historical data when needed. This conclusion was reached via the help of both Stack Overflow and ChatGPT, given the cs file and the nature of analysis.

```
In [787]: # we will begin by performing data manipulation on data_cleaned data frame
# this list is made of the columns i want to drop from the dataset i.e. previous year
columns_to_drop = [
    'Fixed Assets(PREV)',
    'Net Fixed Assets(PREV)',
    'Current Assets(PREV)',
    'Creditors(PREV)',
    'Equity(PREV)',
    'Employees(PREV)',
    'Previous Year'
]
# using the drop function as before, i remove the columns specified in the above list, assigning it back to the
data_cleaned = data_cleaned.drop(columns=columns_to_drop)
# after removing these columns, it is best practice to rename the remaining as there is no need
# for "CUR" anymore, doing so via the rename() function
# as a side note, I renamed the current year column to just year, for clarity
data_cleaned = data_cleaned.rename(columns={
    'Fixed Assets(CUR)': 'Fixed Assets',
    'Net Fixed Assets(CUR)': 'Net Fixed Assets',
    'Current Assets(CUR)': 'Current Assets',
    'Creditors(CUR)': 'Creditors',
    'Equity(CUR)': 'Equity',
    'Employees(CUR)': 'Employees',
    'Current Year': 'Year'
})
# printing out the modified dataframe to ensure changes were completed
print ("Columns after modification:", data_cleaned.columns)
```

```
Columns after modification: Index(['Company Name', 'Company Number', 'Registered office address',
       'Company status', 'Company type', 'Incorporated on',
       'Nature of business (SIC) 0', 'Nature of business (SIC) 1',
       'Nature of business (SIC) 2', 'Nature of business (SIC) 3', 'Year',
       'Currency', 'Fixed Assets', 'Net Fixed Assets', 'Current Assets',
       'Creditors', 'Equity', 'Employees'],
      dtype='object')
```

```
In [788]: print(data_cleaned.info())
data_cleaned.head()

<class 'pandas.core.frame.DataFrame'>
Index: 1166 entries, 0 to 1311
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype  
 ---  -- 
 0   Company Name     1166 non-null    object  
 1   Company Number   1166 non-null    object  
 2   Registered office address  1166 non-null    object  
 3   Company status   1166 non-null    object  
 4   Company type    1166 non-null    object  
 5   Incorporated on  1166 non-null    object  
 6   Nature of business (SIC) 0  1166 non-null    object  
 7   Nature of business (SIC) 1  239 non-null    object  
 8   Nature of business (SIC) 2  73 non-null    object  
 9   Nature of business (SIC) 3  25 non-null    object  
 10  Year             1166 non-null    int64  
 11  Currency         1166 non-null    object  
 12  Fixed Assets     828 non-null    Int64  
 13  Net Fixed Assets 1067 non-null    Int64  
 14  Current Assets   1104 non-null    Int64  
 15  Creditors        968 non-null    Int64  
 16  Equity            1164 non-null    Int64  
 17  Employees         833 non-null    Int64  
dtypes: Int64(6), int64(1), object(11)
memory usage: 179.9+ KB
None
```

Out[788..]

	Company Name	Company Number	Registered office address	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency
0	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	£
1	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	£
2	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	£
3	727 YOUR STORE LIMITED	NI620056	Ruddell & Company, 50 Portmore Street, Portado...	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	£
4	77 RETAIL 2 LIMITED	NI636843	Unit 1 Erler Centre Craigleath Drive, Dundonal...	Active	Private limited Company	2016-03-03	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available

I have decided to remove the 'Registered office address' column because it does not reflect where the business actually operates. In many cases, especially for small businesses, this address is just a requirement for the filing and is sometimes even the owner's home or an accountant's office, not the store or workplace, as I have observed through looking at 20 different businesses and their corresponding registered office address. Since this information is not useful for the analysis and can cause confusion, I have chosen to drop the column, hopefully leading to clearer and more accurate results.

In [790..]

```
# same code as previously, only now with 'Registered office address' being dropped

data_cleaned = data_cleaned.drop(columns=['Registered office address'])

print("Columns after modification:", data_cleaned.columns)
```

Columns after modification: Index(['Company Name', 'Company Number', 'Company status', 'Company type', 'Incorporated on', 'Nature of business (SIC) 0', 'Nature of business (SIC) 1', 'Nature of business (SIC) 2', 'Nature of business (SIC) 3', 'Year', 'Currency', 'Fixed Assets', 'Net Fixed Assets', 'Current Assets', 'Creditors', 'Equity', 'Employees'], dtype='object')

In [791..]

```
print(data_cleaned.head())
```

```

      Company Name Company Number Company status \
0 727 YOUR STORE LIMITED           NI620056      Active
1 727 YOUR STORE LIMITED           NI620056      Active
2 727 YOUR STORE LIMITED           NI620056      Active
3 727 YOUR STORE LIMITED           NI620056      Active
4    77 RETAIL 2 LIMITED           NI636843      Active

      Company type Incorporated on \
0 Private limited Company        2013-08-27
1 Private limited Company        2013-08-27
2 Private limited Company        2013-08-27
3 Private limited Company        2013-08-27
4 Private limited Company        2016-03-03

      Nature of business (SIC) 0 \
0 47110 - Retail sale in non-specialised stores ...
1 47110 - Retail sale in non-specialised stores ...
2 47110 - Retail sale in non-specialised stores ...
3 47110 - Retail sale in non-specialised stores ...
4 47110 - Retail sale in non-specialised stores ...

      Nature of business (SIC) 1 Nature of business (SIC) 2 \
0          NaN          NaN
1          NaN          NaN
2          NaN          NaN
3          NaN          NaN
4          NaN          NaN

      Nature of business (SIC) 3   Year       Currency Fixed Assets \
0          NaN  2020         £      11064
1          NaN  2019         €      13830
2          NaN  2018         £      17288
3          NaN  2017         £      17859
4          NaN  2020  Not Available      337671

  Net Fixed Assets Current Assets Creditors   Equity Employees
0      6425        64782     18364     6425      5
1      3648        17578     16630     3648      6
2      509         3558      18330     509      <NA>
3     17323        106263    54157     17323     <NA>
4     167892       418863    586318     100      22

```

The 'Company Number' column in the dataset is a unique identifier used by Companies House to track and manage company records. However, for the purposes of analysis, this identifier is not particularly useful. The same information can be accessed using the company name, which is more meaningful for analysis. Additionally, keeping the 'Company Number' column adds unnecessary redundancy. By removing this column, I aim to focus only on relevant information that directly contributes to meaningful insights. If detailed information on a specific company is needed, it can be easily retrieved using the company name instead.

```
In [793]: # again, same code as above but dropping company numbers
data_cleaned = data_cleaned.drop(columns=['Company Number'])

print("Columns after modification:", data_cleaned.columns)

Columns after modification: Index(['Company Name', 'Company status', 'Company type', 'Incorporated on',
       'Nature of business (SIC) 0', 'Nature of business (SIC) 1',
       'Nature of business (SIC) 2', 'Nature of business (SIC) 3', 'Year',
       'Currency', 'Fixed Assets', 'Net Fixed Assets', 'Current Assets',
       'Creditors', 'Equity', 'Employees'],
      dtype='object')
```

4.2 Analysis of Univariate Distributions

I would like to focus on analysing the univariate distributions within my dataset, particularly through visualisations such as histograms and density plots. This involves examining individual columns of data to gain insights into their distributions. To start, I want to concentrate on the "Nature of Business" columns, as analysing these will provide valuable insight into the most common SIC codes that companies in the dataset are classified under.

NB: I did not reload the 'cleaned_data_for_reviews2.csv', it was simply for checkpoint purposes

```
In [796]: # histogram creation for first nature of business column (0)

import matplotlib.pyplot as plt # import the Matplotlib library for creating the visualisations in Python
import seaborn as sns # seaborn is a statistical data visualisation library based on Matplotlib, making graphical analysis easier
# thus i needed to be imported

# this code sets the figure size of the plot, specifying in inches for width and height
plt.figure(figsize=(10, 6))

# now i will create the histogram via histoplt function, thus plotting distribution of the column in dataset
```

```
# 'data['Nature of business (SIC) 0']' selects the column to be plotted
# 'dropna()' removes any missing values (NaN) from the data before plotting, ensuring
# that only valid SIC codes are included in the histogram, but it is only temporary and will not drop from data
# only effects the data being passed to the plot, thus handling the missing values in this dataset
# 'kde=False' specifies not want to include a kernel density estimate line, which
# would represent the data's distribution with a continuous curve
sns.histplot(data_cleaned['Nature of business (SIC) 0'].dropna(), kde=False)

# setting title of plot ie .title
plt.title('Histogram for Nature of Business (SIC) 0')

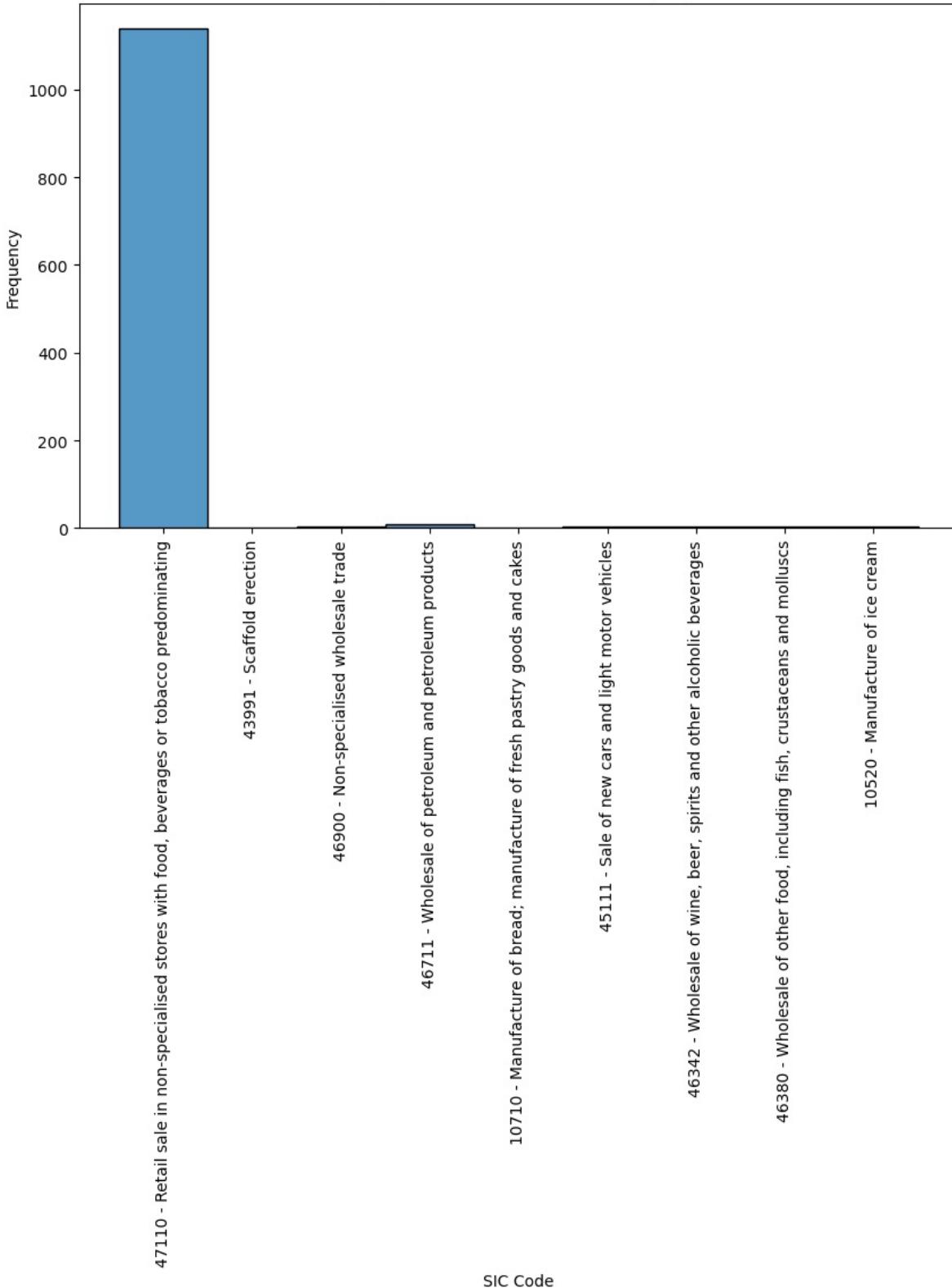
# setting x label axis as sic code
plt.xlabel('SIC Code')

# setting y label axis as the frequency
plt.ylabel('Frequency')

# x-axis labels rotated by 90 degrees
# making them vertical for easier to read labels since the sic codes can be longer
plt.xticks(rotation=90)

# displays the plot
plt.show()
```

Histogram for Nature of Business (SIC) 0



The histogram above illustrates the primary SIC code (industry classification) for each company, indicating their main area of business. Initially, I was concerned that a significant number of companies with different primary SIC codes might skew the dataset, as companies in other industries, like scaffolding or petroleum wholesale, would likely offer products vastly different from supermarkets. However, the histogram reveals that only a small number of companies identify with primary SIC codes other than 47110 (Retail sale in non-specialised stores with food, beverages, or tobacco predominating). This finding suggests that the dataset is highly reliable for analysing the supermarket and retail industry, with minimal influence from companies in unrelated sectors.

Likewise, I believe keeping the three additional SIC code columns (secondary, tertiary, and quaternary) is beneficial because it allows for a broader analysis of the companies. These columns help us see if businesses are involved in multiple industries, which can provide insights into how they diversify their operations. This also makes it easier to compare companies across different sectors and understand how various industries interact within the same market. By retaining these columns, we gain a more complete picture of each company's

activities beyond just their primary focus.

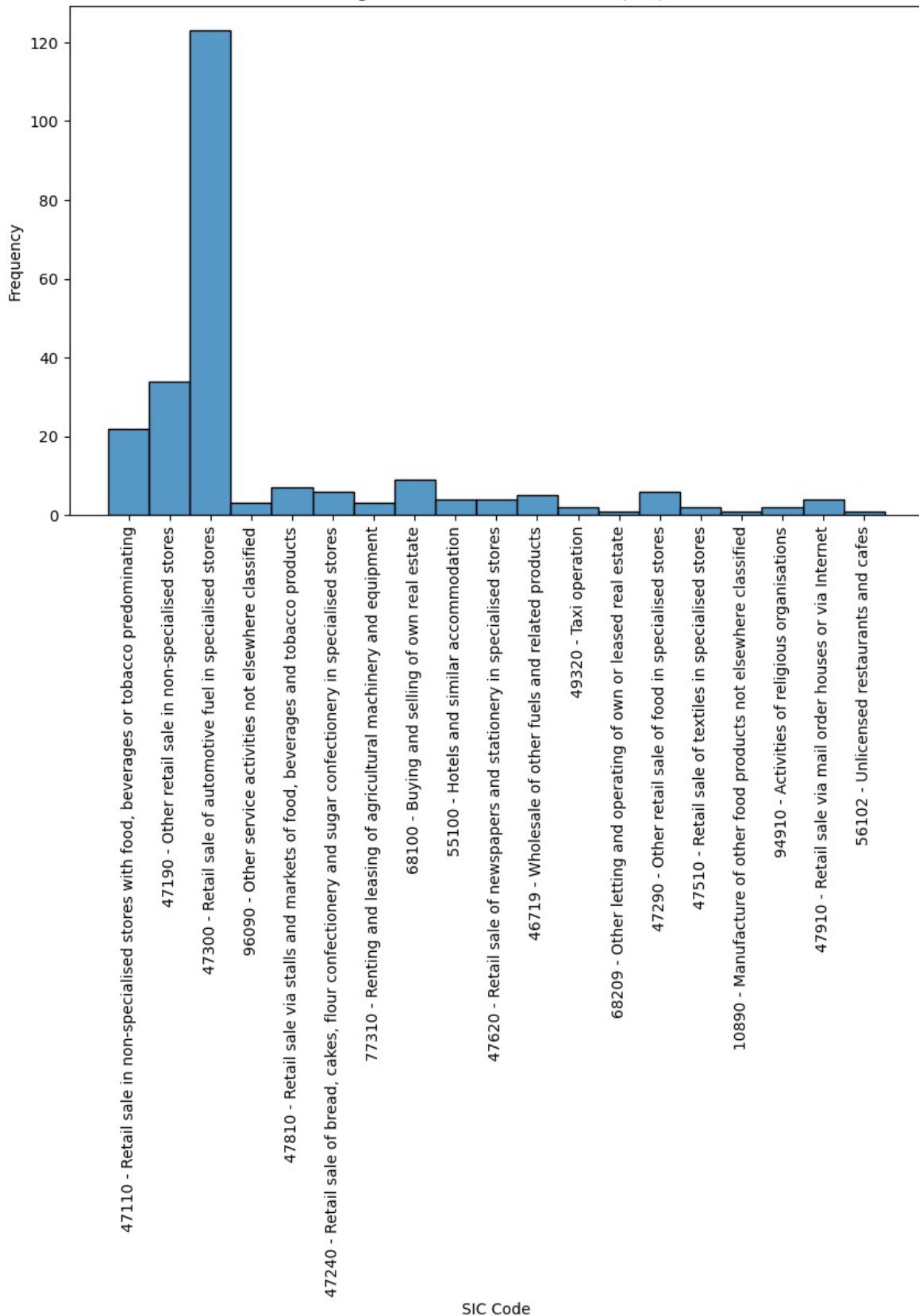
Therefore, I have created visualisations for these additional columns as well, using the same Python code as above, with only the column names adjusted. Overall, there is a diverse range of entries in each column, making direct comparisons challenging. However, these columns provide a general overview of the additional operations these companies engage in, along with insights into the products and services they offer. This isn't a major issue, as the number of companies identifying with secondary or tertiary etc SIC codes is relatively small compared to SIC 0.

```
In [798]: # second nature of buisness column (1)
plt.figure(figsize=(10, 6))
sns.histplot(data_cleaned['Nature of business (SIC) 1'].dropna(), kde=False)
plt.title('Histogram for Nature of Business (SIC) 1')
plt.xlabel('SIC Code')
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()

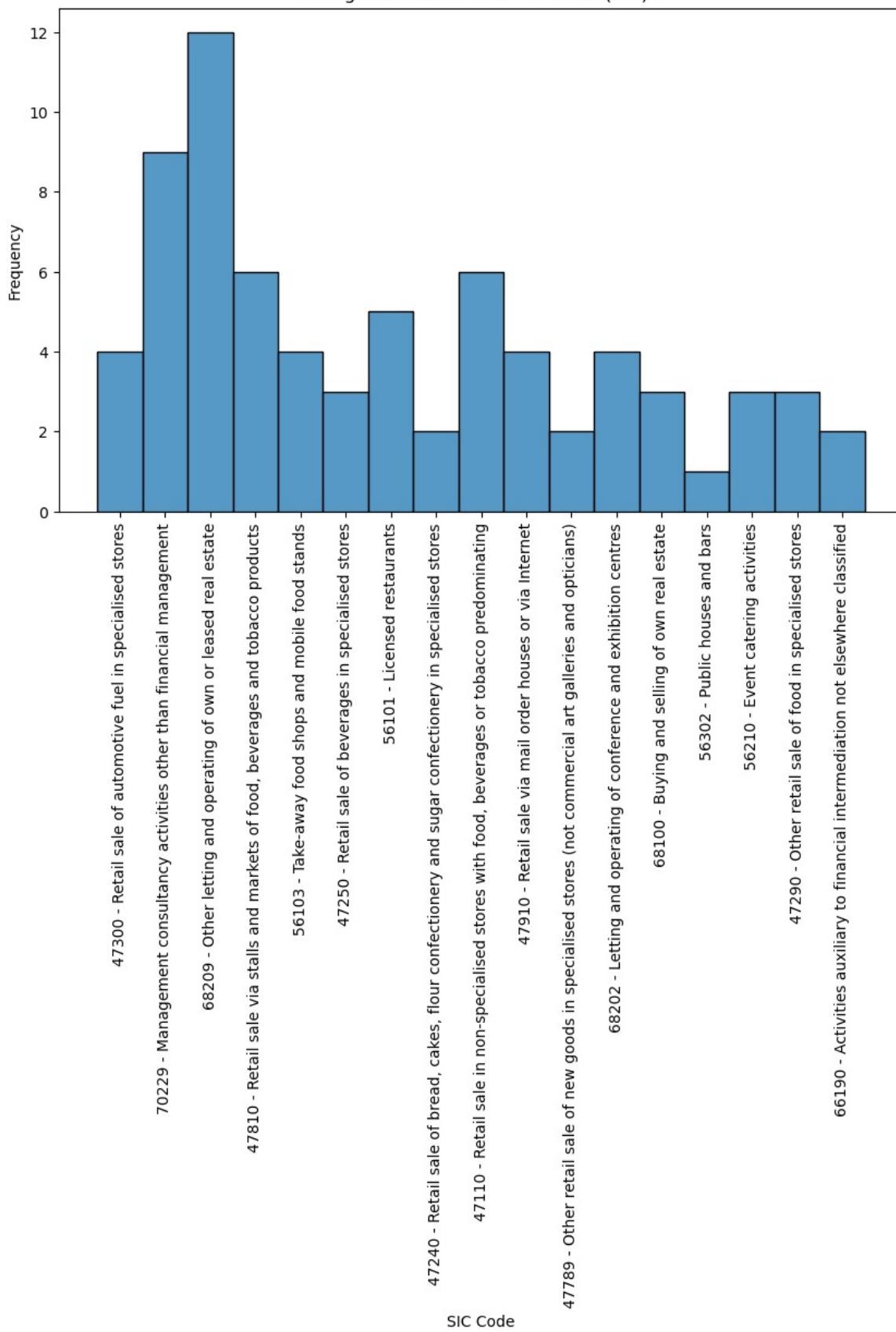
# third nature of buisness column (2)
plt.figure(figsize=(10, 6))
sns.histplot(data_cleaned['Nature of business (SIC) 2'].dropna(), kde=False)
plt.title('Histogram for Nature of Business (SIC) 2')
plt.xlabel('SIC Code')
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()

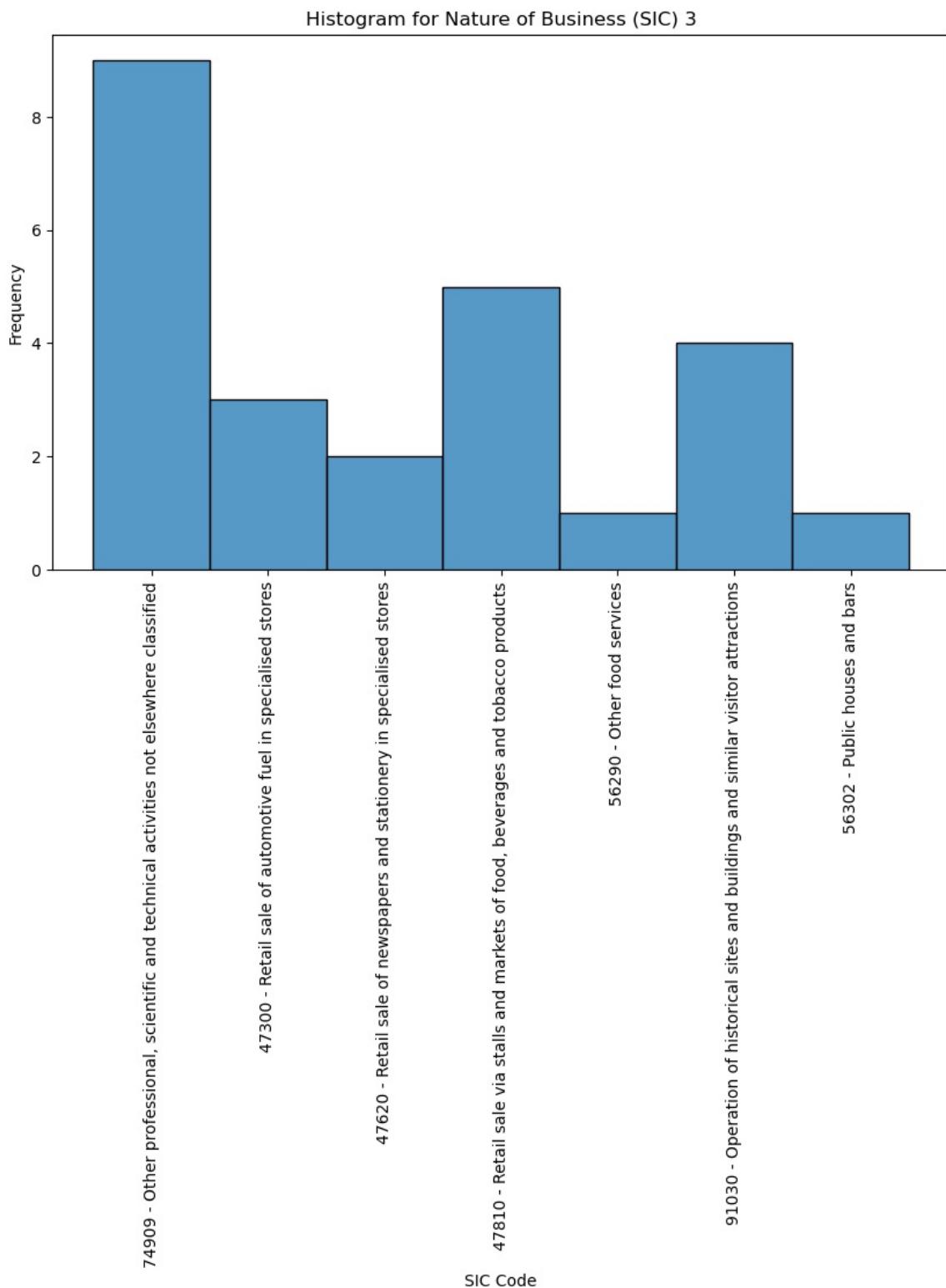
# fourth nature of buisness column (4)
plt.figure(figsize=(10, 6))
sns.histplot(data_cleaned['Nature of business (SIC) 3'].dropna(), kde=False)
plt.title('Histogram for Nature of Business (SIC) 3')
plt.xlabel('SIC Code')
plt.ylabel('Frequency')
plt.xticks(rotation=90)
plt.show()
```

Histogram for Nature of Business (SIC) 1



Histogram for Nature of Business (SIC) 2





Next, I decided to delve into the distribution of currencies. The visualisation below highlights that the majority of companies in my dataset have their currency marked as 'Not Available' indicating significant missing or unreported data, which has challenges for financial analysis. Among the reported currencies GBP £ is the most common, suggesting that many companies are British or operate in the UK, hence the dataset being Northern Irish companies, while other symbols appear infrequently and may represent data inconsistencies or errors. I will retain the currency column as a reminder that this assumption might skew results, especially for anyone who wishes to

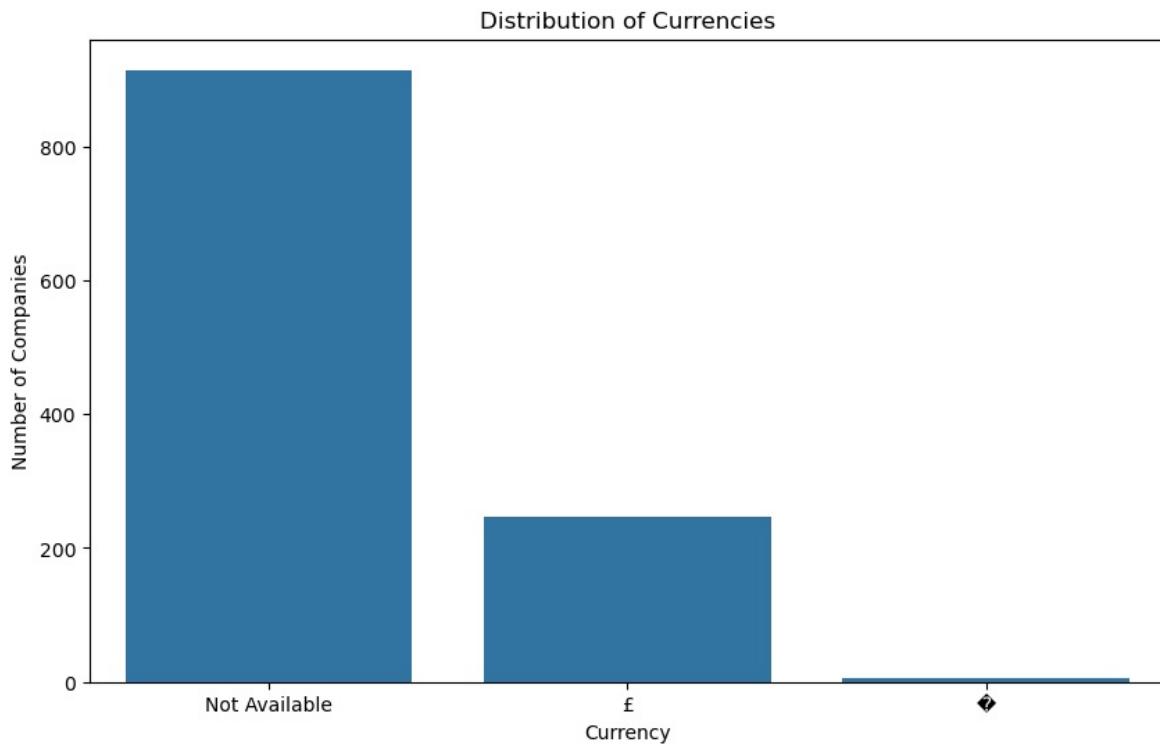
conduct further analysis either in this notebook or with the dataset, yet presume that they are all in GBP, as there is no indication for conversion to a single currency, removing the rows would remove most of the data and the dataset is from Northern Irish companies, thus local currency is GBP.

```
In [800]: # count occurrences of each currency in the 'Currency' column
# 'value_counts()' func counts the number of occurrences of each unique value in the 'Currency' column
# creates series where the index is the unique currency symbols and the values are the counts
currency_counts = data_cleaned['Currency'].value_counts()

# create bar chart to show the distribution of currencies
# 'plt.figure(figsize=(10, 6))' initialises a new figure with the size making sure it is readable
plt.figure(figsize=(10, 6))

# 'sns.barplot' creates a bar chart
# 'x=currency_counts.index' sets the x axis to display the symbols
# 'y=currency_counts.values' sets the y axis to display the counts of each currency symbol in dataset
# thus visualising the frequency of each currency
sns.barplot(x=currency_counts.index, y=currency_counts.values)

# same for the previous histograms
plt.title('Distribution of Currencies')
plt.xlabel('Currency')
plt.ylabel('Number of Companies')
plt.show()
```



Moreover, I wanted to investigate the distribution of company types.

The code below was partially generated with the help of ChatGPT due to issues with the hue parameter in Seaborn, which were beyond my understanding. While I grasped the concepts, I encountered difficulties in debugging and resolving the error. However, the code is now functioning correctly.

Here is my original code that encountered issues:

```
type_counts = data['Company type'].value_counts() plt.figure(figsize=(10, 6)) ax = sns.barplot(x=type_counts.index,
y=type_counts.values, palette='Blues') plt.title('Distribution of Company Types') plt.xlabel('Company Type') plt.ylabel('Number of
Companies') short_labels = ['Ltd.', 'LLP', 'Plc.', 'Other', 'Private Ltd. by Guarantee']
ax.set_xticklabels(short_labels, rotation=45, ha="right") plt.show()
```

After visually analysing the visualisation of the distribution of company types, it was clear that Private limited companies (Ltd.) were the vast majority of businesses in the dataset, possibly due to the advantages of limited liability and the relative ease of setting up such companies, which are particularly popular among small and medium sized businesses (<https://realbusiness.co.uk/disadvantages-of-private-limited-company>). As a result, any analysis or insights derived from this dataset will be most relevant to Private Limited Companies, given their overwhelming representation. It is also worth noting that Private limited companies can include supermarkets, especially smaller or independently owned ones. However, larger supermarket chains are often structured as Public limited companies because they need to raise capital from the public through stock markets.

However, after browsing the Companies House website, I discovered that major supermarket chains like ASDA, Aldi, and Lidl are classified as Private limited companies, while Tesco and Sainsbury's are Public limited companies (Plc). This finding reassures me that the dataset, which predominantly features Ltd companies, is still relevant to large supermarket chains, even market leaders.

ASDA: <https://find-and-update.company-information.service.gov.uk/company/00464777> Lidl: <https://find-and-update.company-information.service.gov.uk/company/02816429> Aldi: <https://find-and-update.company-information.service.gov.uk/company/02321869>
Tesco: <https://find-and-update.company-information.service.gov.uk/company/00445790> Sainsbury's: <https://find-and-update.company-information.service.gov.uk/company/00185647>

```
In [802]: # Count the occurrences of each unique company type
type_counts = data_cleaned['Company type'].value_counts().reset_index()

# Rename the columns for clarity
type_counts.columns = ['Company Type', 'Count']

# Create a bar chart to show the distribution of company types
plt.figure(figsize=(10, 6))

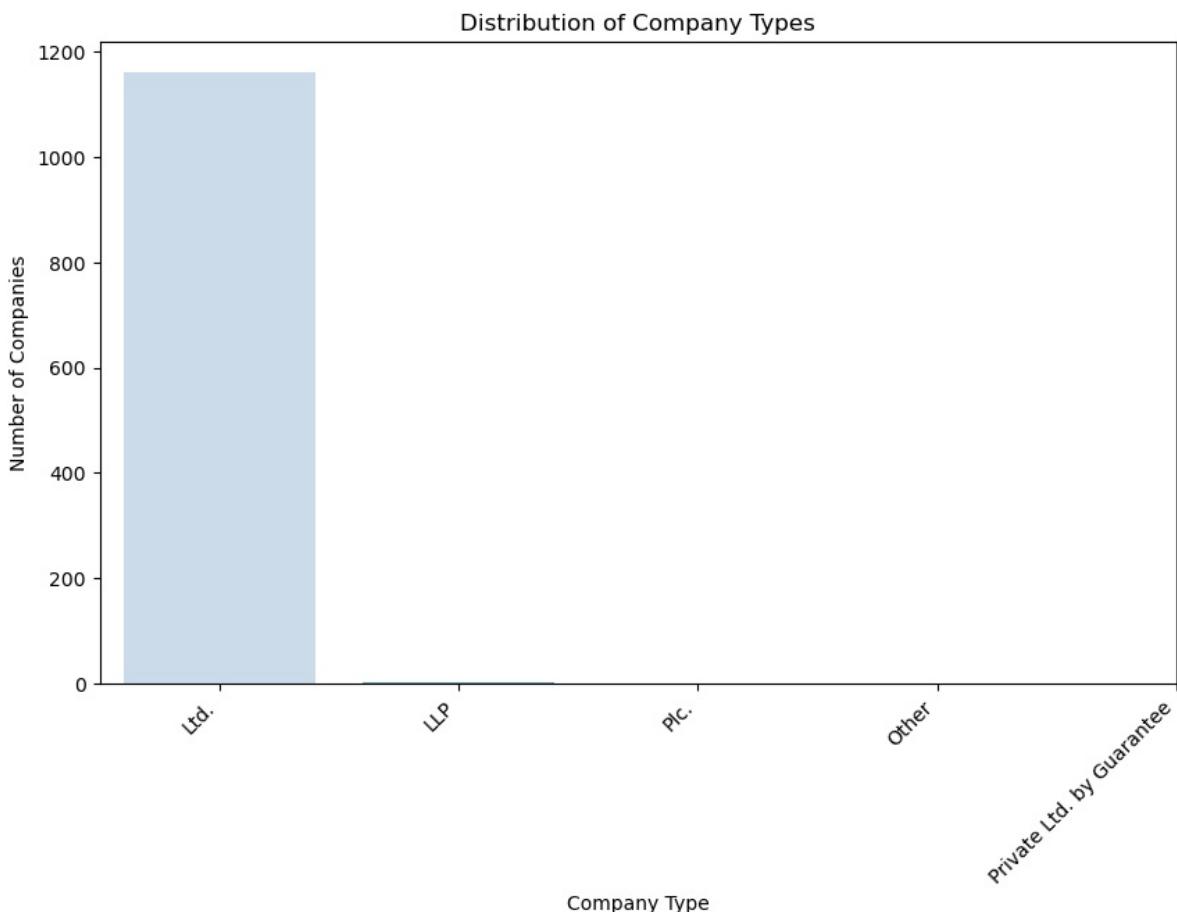
# Use 'hue' and set 'legend=False' to avoid the FutureWarning
ax = sns.barplot(x='Company Type', y='Count', data=type_counts, hue='Company Type', dodge=False, palette='Blues')

# Set the title and labels
plt.title('Distribution of Company Types')
plt.xlabel('Company Type')
plt.ylabel('Number of Companies')

# Shorten the labels if necessary (adjust these to fit your actual data)
short_labels = ['Ltd.', 'LLP', 'Plc.', 'Other', 'Private Ltd. by Guarantee']

# Apply the shortened labels to the x-axis ticks
ax.set_xticks(range(len(short_labels))) # Set the number of ticks to match the number of labels
ax.set_xticklabels(short_labels, rotation=45, ha="right")

# Display the plot
plt.show()
```



Furthermore, using the information provided by ChatGPT, as demonstrated in previous code, I was able to create a similar visualisation for company statuses. Although there were still challenges with the hue parameter in Seaborn, I successfully applied the knowledge to resolve these issues in the new code I used.

This visualisation reveals that the vast majority of companies in my dataset are classified as 'Active' meaning that most businesses are currently in operation, suggesting that any analysis or insights derived from my dataset will primarily reflect the characteristics and behaviors of ongoing businesses. While the dataset allows for some exploration into companies that are no longer operating or are in the process of closing, the primary conclusions should be applied to active businesses. This is beneficial since my expert's supermarket is classified as 'Active' and also aligning with the status of almost all supermarkets in the UK, and definitely major ones and market leaders.

```
In [804]: status_counts = data_cleaned['Company status'].value_counts().reset_index()

# renaming columns
status_counts.columns = ['Company Status', 'Count']

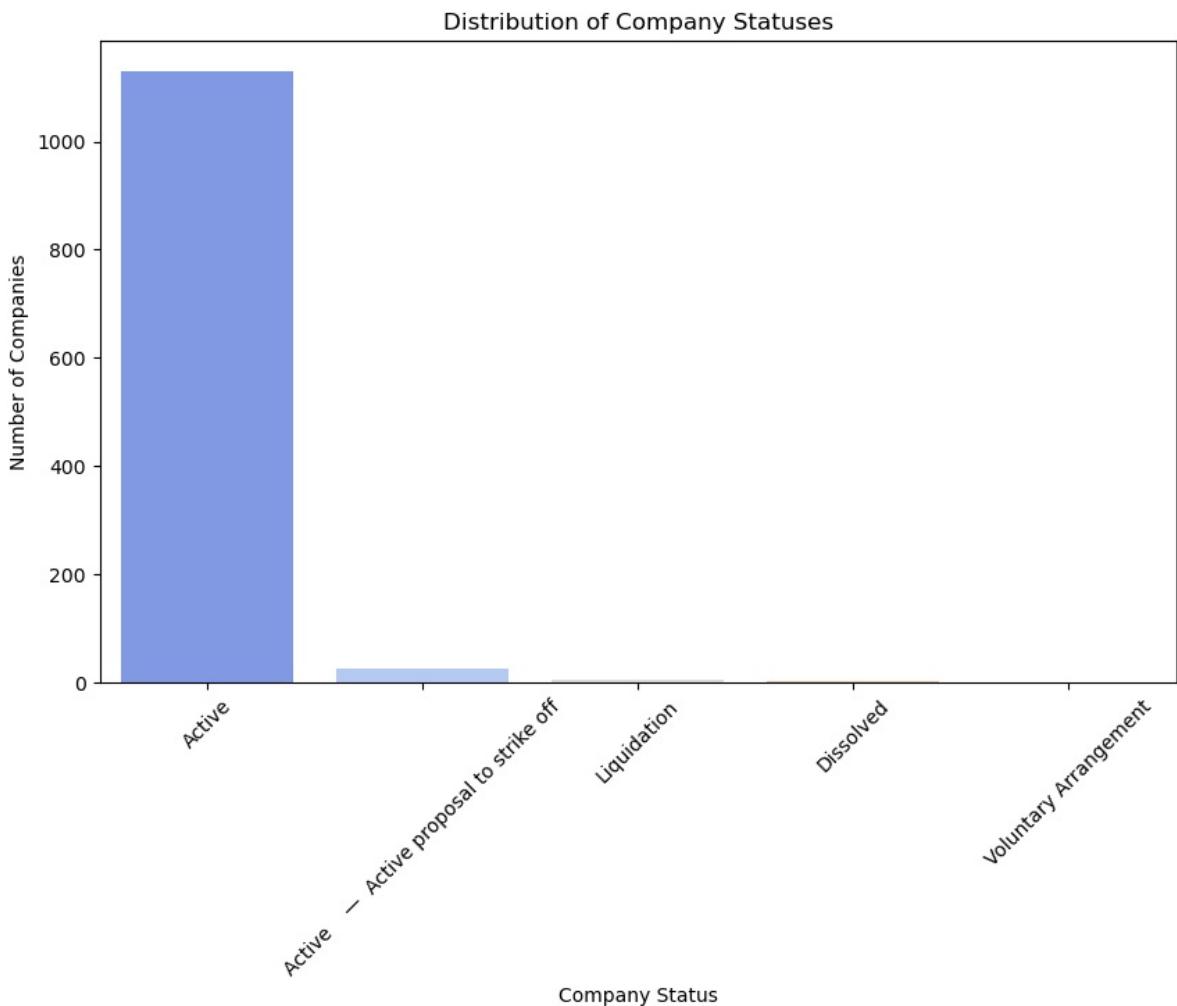
# plotting bar chart height and width
plt.figure(figsize=(10, 6))

# using 'hue' to assign colors based on 'Company Status' and set 'legend=False', as shown above, to avoid hue error
sns.barplot(x='Company Status', y='Count', data=status_counts, hue='Company Status', dodge=False, palette='cool')

# setting titles and labels
plt.title('Distribution of Company Statuses')
plt.xlabel('Company Status')
plt.ylabel('Number of Companies')

# rotating the x labels for readability
plt.xticks(rotation=45)

# display plot
plt.show()
```



I then began to visualise company name. The bar chart below provides insight into the distribution of companies based on the number of rows they occupy in the dataset, via the occurrence of their name, with companies like McAuley Botanic and JMK Enterprises appearing most frequently with 6 rows. The presence of multiple rows for these companies suggests that they have more historical records, hence the duplication per row. This could indicate that these 'top 10' companies have undergone several reporting periods, making them particularly relevant for in-depth analysis when considering time related evaluation. However, it also means that any conclusions drawn from the dataset should consider whether these frequently represented companies are reflective of the broader industry. If a company has only two records while another has five, the company with more records offers a better view of changes over time, allowing for more detailed analysis. This could show trends or growth more clearly. However, companies with fewer records might not provide as much

insight into time related factors, so it's important to consider these differences to ensure the analysis isn't biased towards companies with more data.

```
In [806]: # the value_counts() method is used to count how many times each company name appears
company_name_counts = data_cleaned['Company Name'].value_counts()

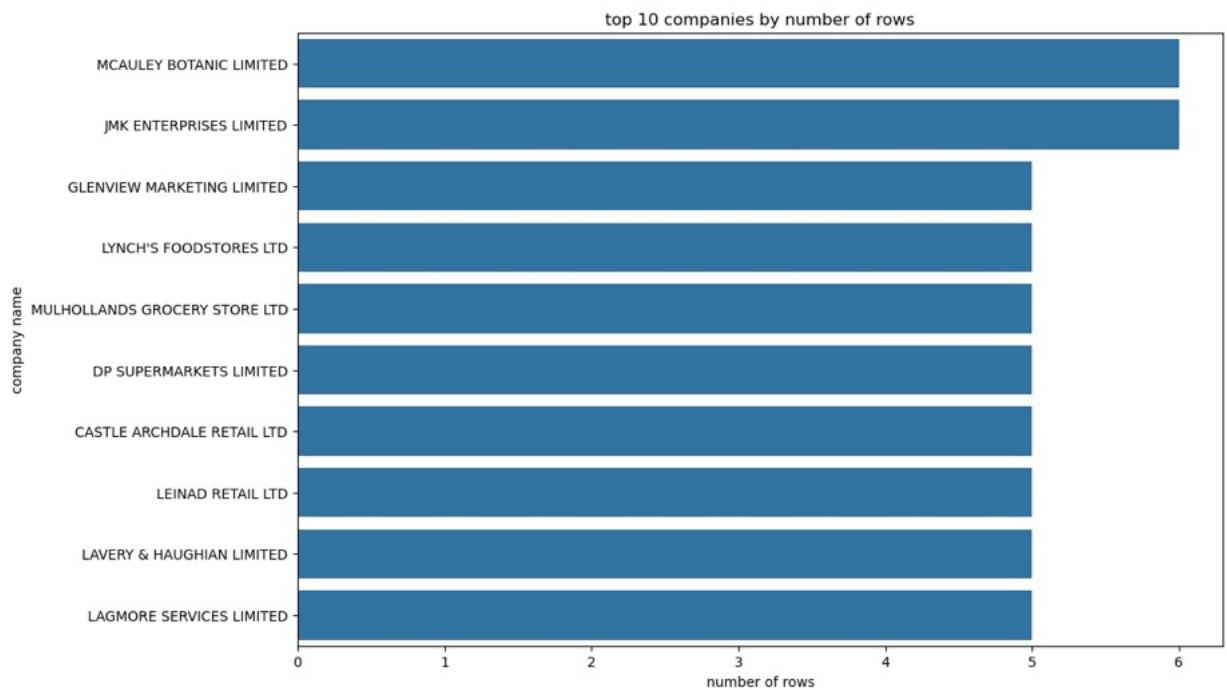
# get the top 10 companies with the most rows in the dataset
# head(10) selects the first 10 entries in the sorted series
top_company_names = company_name_counts.head(10)

# creating new figure for plot with width and height
plt.figure(figsize=(12, 8))

# creating a bar plot using seaborn
sns.barplot(x=top_company_names.values, y=top_company_names.index)

# label axis and set title as shown in previous code
plt.title('top 10 companies by number of rows')
plt.xlabel('number of rows')
plt.ylabel('company name')

# display the plot
plt.show()
```



Furthermore, the heatmap below shows the number of company incorporations by year and month, clearly showing when businesses in the dataset were most commonly established. The most significant concentration of incorporations occurred between 2010 and 2018, with especially high activity in 2010 and 2014. This indicates that these years were marked by substantial business growth or economic conditions that encouraged company formation. In contrast, earlier years, especially before 2000, show much lower levels of company incorporations. This focus on more recent companies could shape the analysis by highlighting trends, challenges and opportunities relevant to newer businesses.

```
In [808]: # convert the 'incorporated on' column to datetime format
data_cleaned['Incorporated on'] = pd.to_datetime(data_cleaned['Incorporated on'])

# extract the year and month, then store them in temporary columns
data_cleaned['year'] = data_cleaned['Incorporated on'].dt.year
data_cleaned['month'] = data_cleaned['Incorporated on'].dt.month

# create a pivot table to organise the data
# aggfunc='size' counts the number of companies incorporated in each year-month combination
# fill_value=0 fills any missing values with 0, ensuring the heatmap has no gaps
incorporation_pivot = data_cleaned.pivot_table(index='year', columns='month', aggfunc='size', fill_value=0)

# drop the temporary 'year' and 'month' columns from the dataframe to avoid keeping them permanently
data_cleaned = data_cleaned.drop(columns=['year', 'month'])

# remainder of code is the same as before to generate the heatmap
plt.figure(figsize=(12, 8))
```

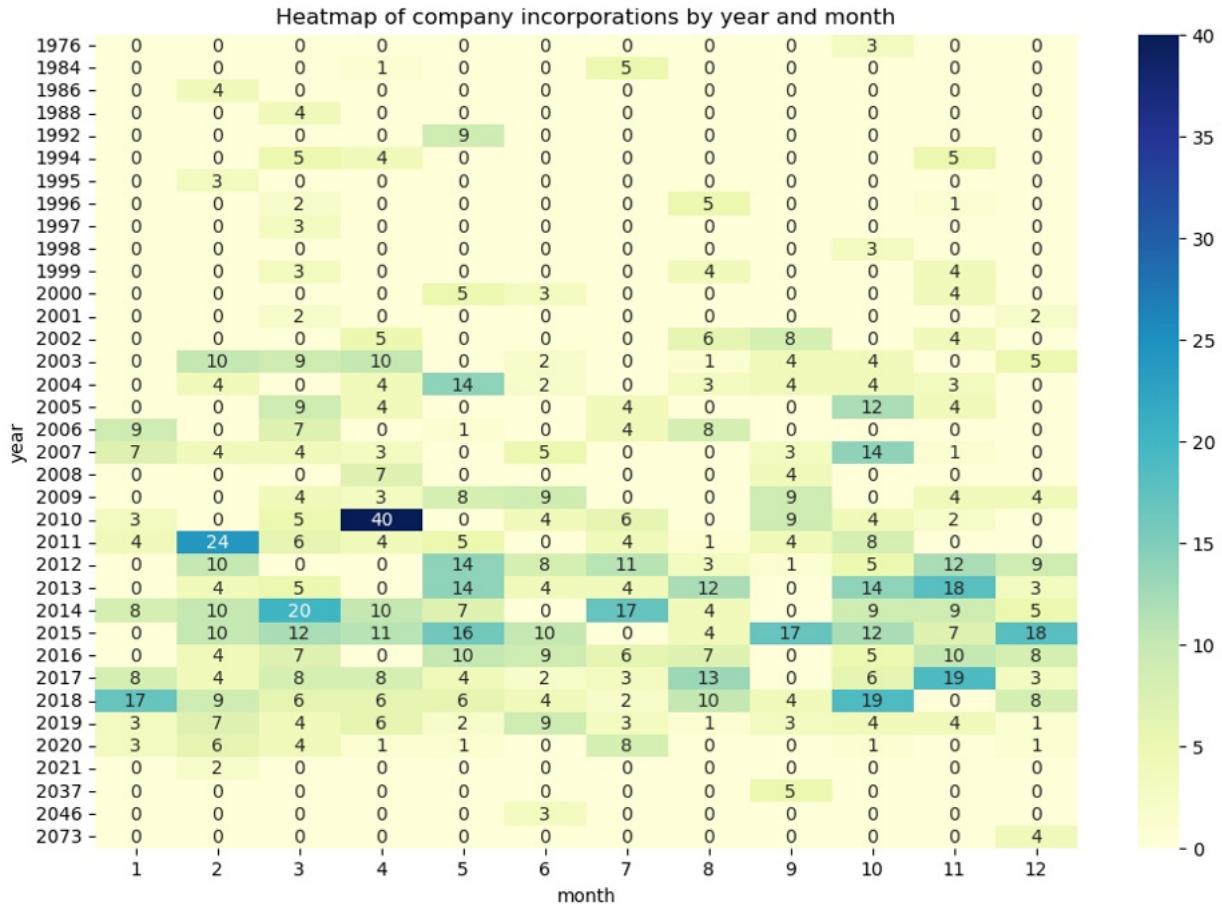
```

sns.heatmap(incorporation_pivot, cmap='YlGnBu', annot=True, fmt="d")

plt.title('Heatmap of company incorporations by year and month')
plt.xlabel('month')
plt.ylabel('year')

plt.show()

```



I will now begin analysing the financial columns using density plots, starting with fixed assets. Density plots are used here to show how financial values are distributed across the dataset, giving a clear view of how the data points are spread out. This helps to easily spot patterns, groupings and any outliers within the dataset. It's important to note that the code is fully commented in the initial examples (all rows and 2020), but for subsequent plots, the code is essentially duplicated with changes only to the colour and column names. Therefore, detailed comments are not repeated for each plot.

```

In [810]: # creating a density plot for fixed assets column, plt.figure() creates a new figure object and I have specified
plt.figure(figsize=(10, 6))

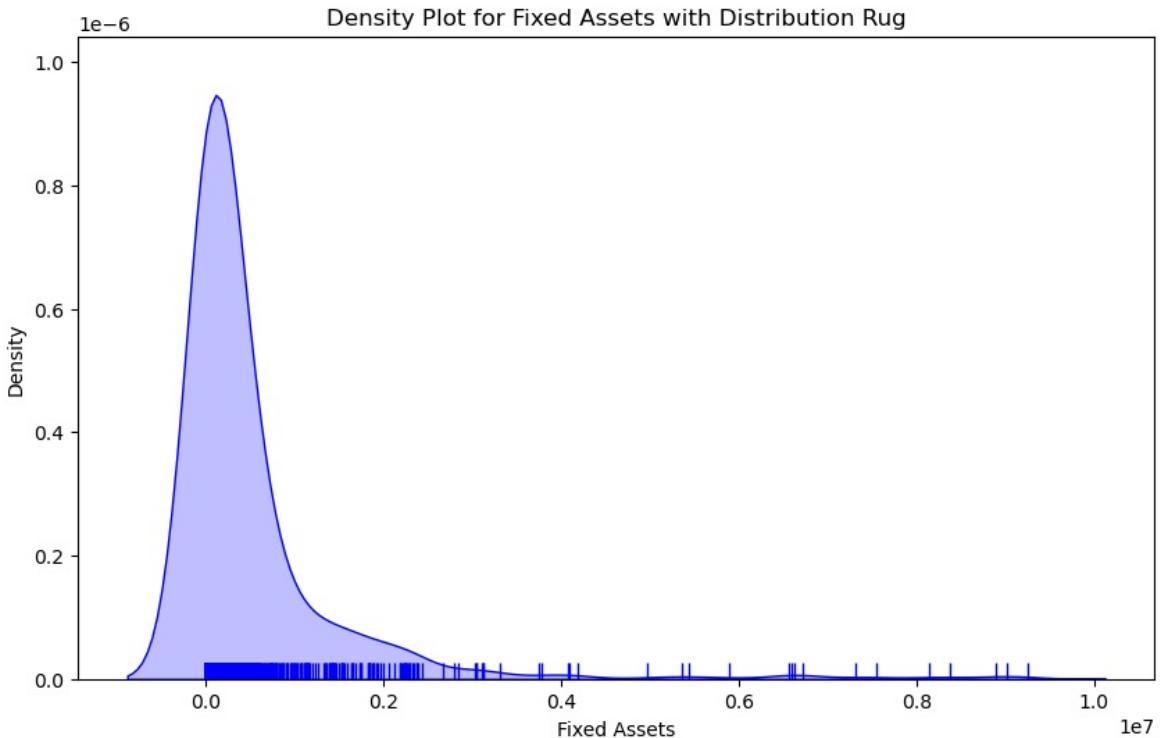
# creating a kernel density estimate plot showing distribution of data, selecting fixed assets column
# .dropna() is used to temporarily remove any missing values from fixed asset column, otherwise issue with plot
# not getting rid of them entirely as the rows can be used in other ways
# 'fill=True' fills the area under the kde to make more visually apparent
# 'color="blue"' set colour of kde to blue (this changes in later code due to showing the different in the densities
# for the financial columns
sns.kdeplot(data_cleaned['Fixed Assets'].dropna(), fill=True, color="blue")

# sns.rugplot() adds small vertical lines (a rug plot) at the bottom of the kde to show the distribution of individual
# I am selecting the fixed assets column again, dropping the missing values temporarily, and make the rug plot blue
sns.rugplot(data_cleaned['Fixed Assets'].dropna(), color="blue")

# setting titles and labels as before
plt.title('Density Plot for Fixed Assets with Distribution Rug')
plt.xlabel('Fixed Assets')
plt.ylabel('Density')

# displaying the plot
plt.show()

```



This density plot for "Fixed Assets" reveals that most companies in the dataset have relatively low fixed assets, with the distribution being heavily right skewed. The sharp peak on the left indicates that the majority of companies have their fixed assets clustered around lower values, typically between 0 and 1 million. However, there is a long tail extending to the right, showing that a few companies have significantly higher fixed assets, up to 10 million or more, suggesting the presence of outliers. The rug plot at the bottom, with its dense clustering of lines on the left, further confirms that most companies fall into the lower range of fixed assets, while the more sparsely spaced lines towards the right indicate the less common, higher fixed assets. This distribution highlights the wide disparity in fixed assets among the companies in the dataset, which could be important for understanding their financial health and scale.

I then created a Number of Company Filings by Year bar chart. The reason for creating the bar chart below is to account for the fact that companies in the dataset appear multiple times across different rows, each representing a different year with corresponding financial aspects. The chart highlights 2020 as the year with the most company filings, which provides a focus for analysis. By identifying 2020 as a year with the highest number of records, I gained a more meaningful snapshot of the dataset during a key time. By displaying the most active year each time I create an entire column visualisation, I can see if the trends and patterns observed in 2020 align with those seen across the entire column, offering a less biased insight into the dataset, and address those issues if there are any.

```
In [812]: # counting the numbers of filings per year
year_counts = data_cleaned['Year'].value_counts().sort_index()

# displaying counts
print(year_counts)
```

Year	count
2016	5
2017	127
2018	230
2019	276
2020	303
2021	223
2022	2

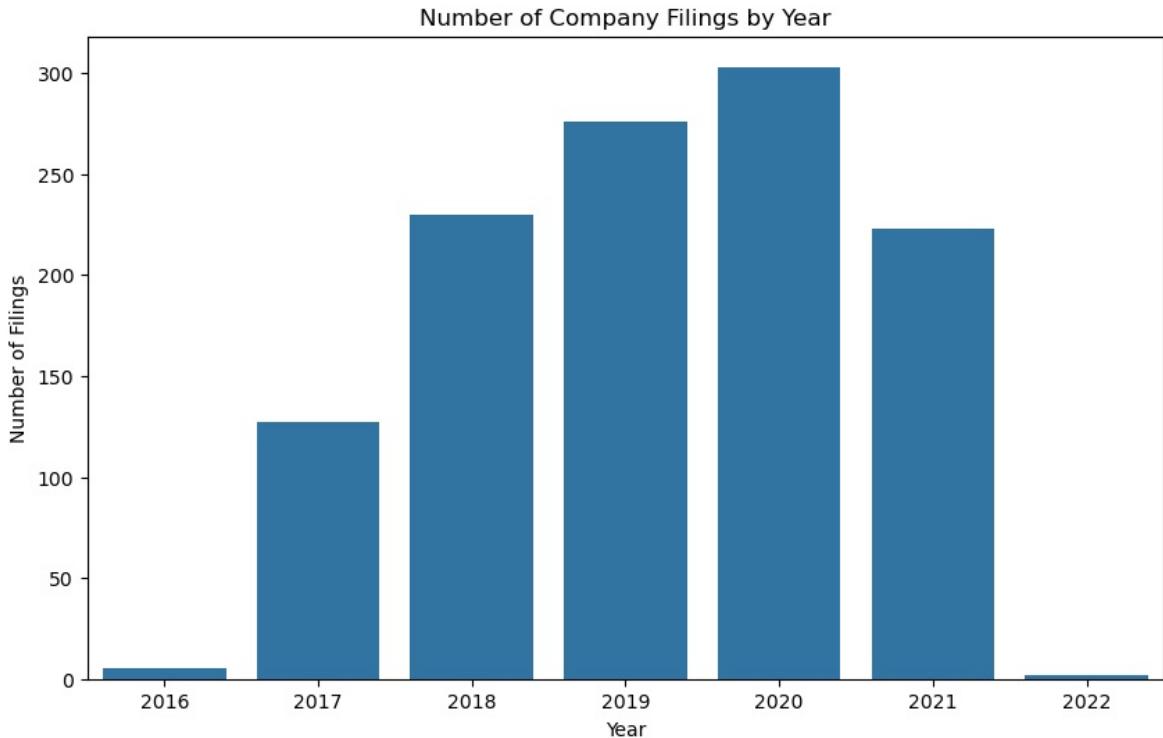
Name: count, dtype: int64

```
In [813]: plt.figure(figsize=(10, 6))
# creating a new figure object and specified sizing

# creating a bar plot showing the number of company filings by year
sns.barplot(x=year_counts.index, y=year_counts.values)
# sns.barplot() generates a bar chart
# 'x=year_counts.index' uses the years as the x-axis (categories)
# 'y=year_counts.values' uses the number of filings as the y-axis (values to be plotted)

# assigning labels and titles
plt.title('Number of Company Filings by Year')
plt.xlabel('Year')
plt.ylabel('Number of Filings')

# displaying the plot
plt.show()
```



```
In [814]: # filtering the dataset to include only rows where the 'Year' is 2020
data_2020 = data_cleaned[data_cleaned['Year'] == 2020]
# creates a new data frame 'data_2020' where the 'Year' column is equal to 2020 for the rows

# creating the density plot for the fixed assets column specifically for the year 2020 and specify sizing
plt.figure(figsize=(10, 6))

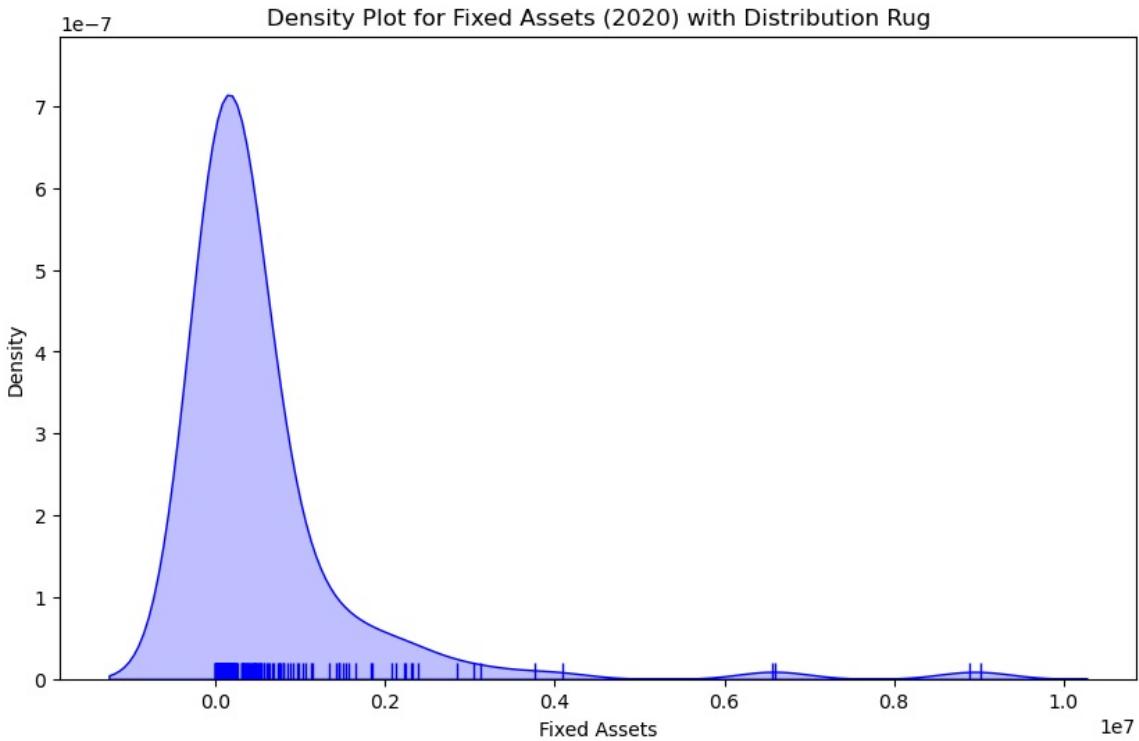
sns.kdeplot(data_2020['Fixed Assets'].dropna(), fill=True, color="blue")
# sns.kdeplot() creates a kde plot for the fixed assets column from the filtered 2020 data
# 'data_2020['Fixed Assets']' selects the fixed assets column from the 'data_2020' data frame
# .dropna() is used as seen before

sns.rugplot(data_2020['Fixed Assets'].dropna(), color="blue")
# again, rug plot added

# set title and labels
plt.title('Density Plot for Fixed Assets (2020) with Distribution Rug')

plt.xlabel('Fixed Assets')
plt.ylabel('Density')

# display the plot
plt.show()
```



This density plot for "Fixed Assets" in 2020 shows a distribution similar to the overall fixed assets distribution, with a heavy right-skew and a sharp peak near the lower end of the value range. Most companies in 2020 had relatively low fixed assets, clustered around values close to zero, which is consistent with the broader dataset. The rug plot at the bottom further confirms this concentration of companies with lower fixed assets. However, the plot also shows a long tail extending towards higher values, indicating that, as with the overall dataset, a few companies in 2020 held significantly larger fixed assets. This suggests that 2020, a year with the most company records, follows the same trend as the entire dataset, with the majority of companies maintaining modest fixed assets while a smaller number have much larger values.

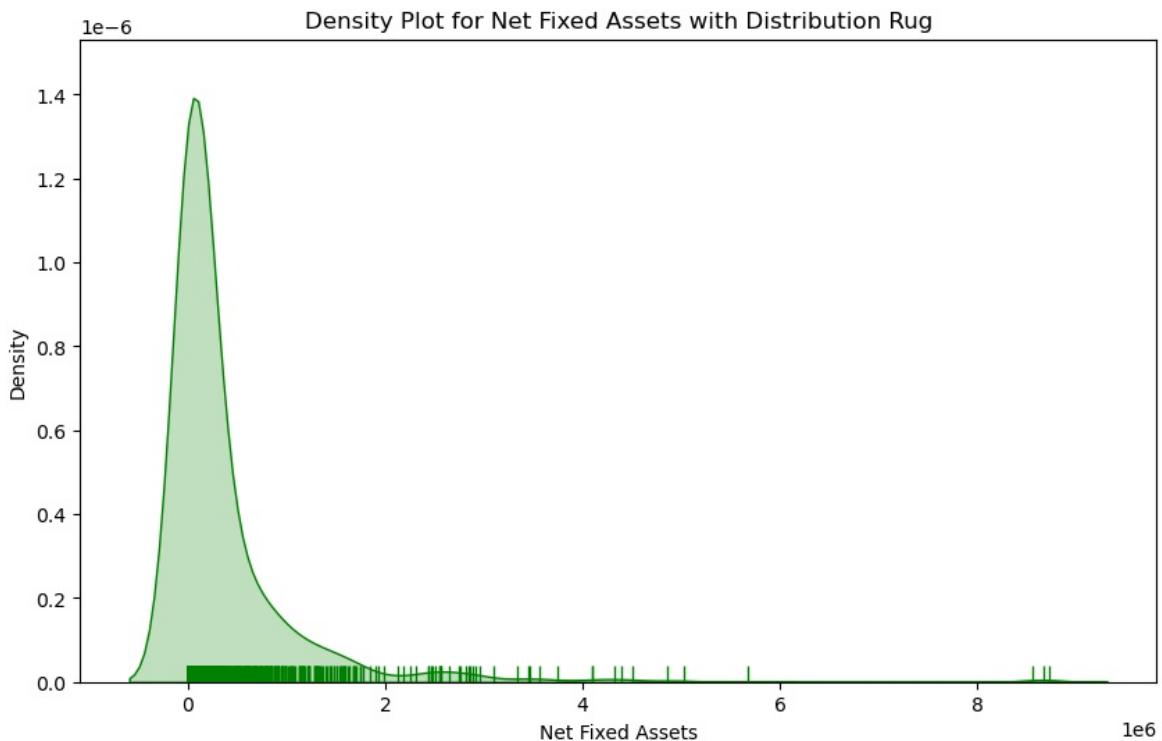
It is also worth mentioning here, regarding the y axis, that if most companies have fixed assets clustered tightly around a certain value, the density plot will have a high peak, with values like 1, 2, or higher, but if companies have fixed assets spread widely across a range, the density plot will have lower peaks, with values like 0.2, 0.4, etc. This is just for clarification.

Also, the scale on the x axis is automatically adjusted based on the range of the data being plotted. The key takeaway here is that 1e6 refers to millions, and 1e7 refers to tens of millions, so the plots are showing data in different ranges, leading to different axis labeling. So, for example, the 'Density Plot for Fixed Assets (2020) with Distribution Rug' has 0.X for 1e7, meaning 0.2 is 2 million. Similarly, the 'Density Plot for Net Fixed Assets with Distribution Rug' shows 1e6, meaning that on the x axis 1 is equal to one million, 2 to two million. If it was 0.1 for the x axis with a 1e6, it would be 100,000.

```
In [816]: # Code for density plot for net fixed assets (all rows)
plt.figure(figsize=(10, 6))
sns.kdeplot(data_cleaned['Net Fixed Assets'].dropna(), fill=True, color="green")
sns.rugplot(data_cleaned['Net Fixed Assets'].dropna(), color="green")

plt.title('Density Plot for Net Fixed Assets with Distribution Rug')
plt.xlabel('Net Fixed Assets')
plt.ylabel('Density')

plt.show()
```

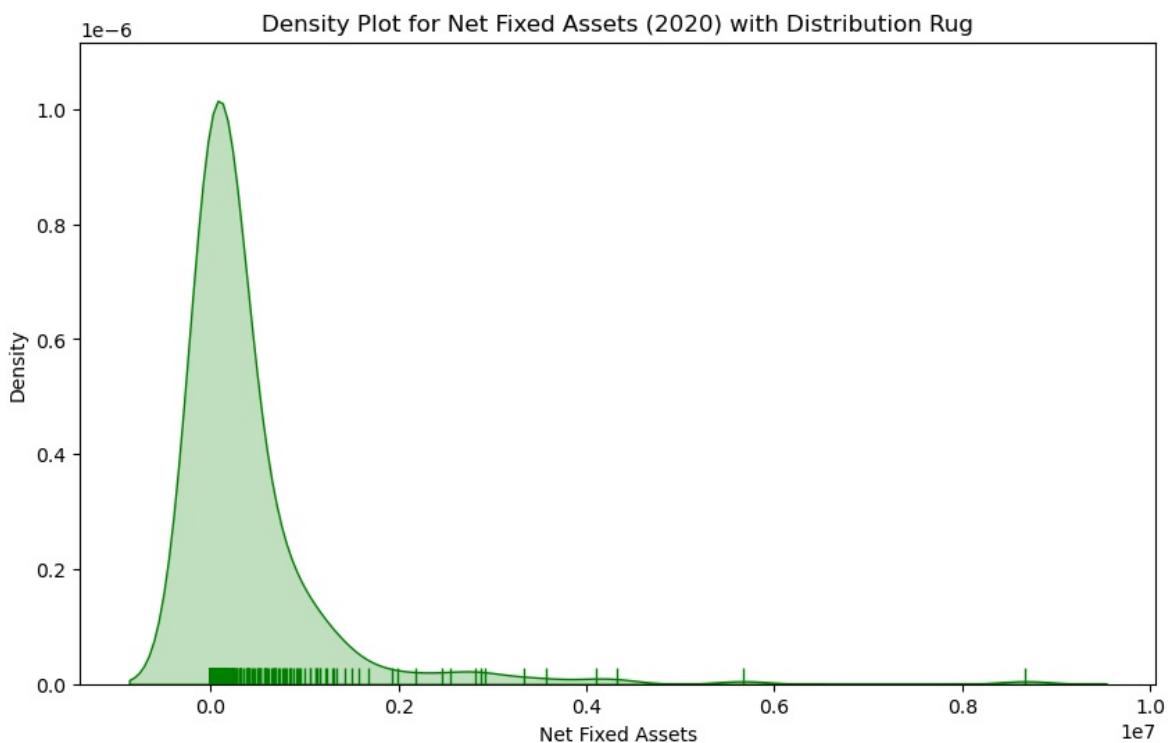


```
In [817]: # Code for density plot for net fixed assets (only rows that are dated 2020(Year))
data_2020 = data_cleaned[data_cleaned['Year'] == 2020]

plt.figure(figsize=(10, 6))
sns.kdeplot(data_2020['Net Fixed Assets'].dropna(), fill=True, color="green")
sns.rugplot(data_2020['Net Fixed Assets'].dropna(), color="green")

plt.title('Density Plot for Net Fixed Assets (2020) with Distribution Rug')
plt.xlabel('Net Fixed Assets')
plt.ylabel('Density')

plt.show()
```



```
In [818]: # Code for density plot for current assets (all rows)
plt.figure(figsize=(10, 6))
sns.kdeplot(data_cleaned['Current Assets'].dropna(), fill=True, color="purple")
```

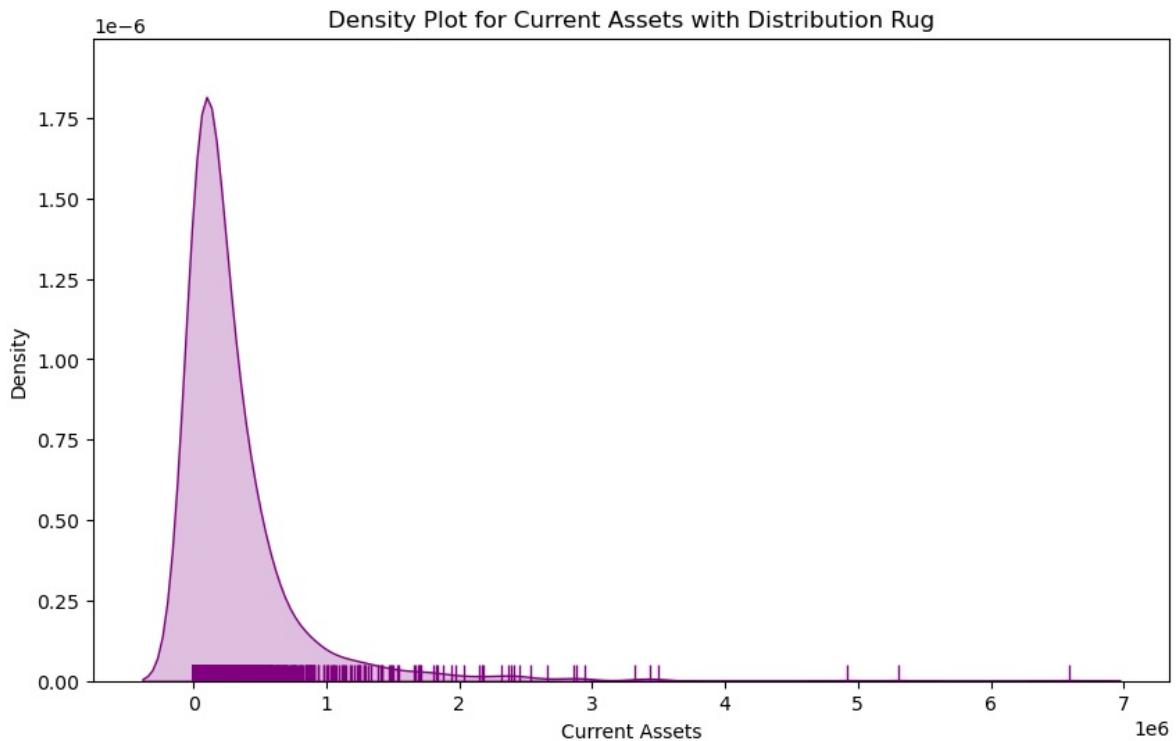
```

sns.rugplot(data_cleaned['Current Assets'].dropna(), color="purple")

plt.title('Density Plot for Current Assets with Distribution Rug')
plt.xlabel('Current Assets')
plt.ylabel('Density')

plt.show()

```



```

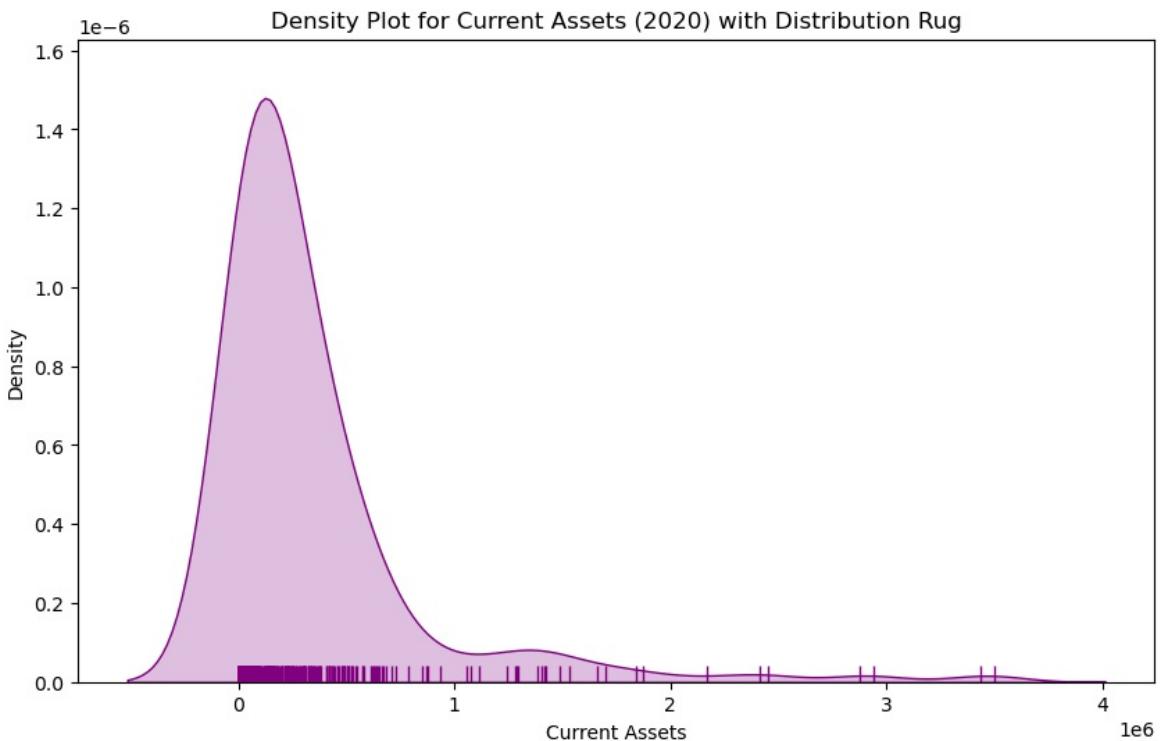
In [819]: # Code for density plot for current assets (only rows that are dated 2020(Year))
data_2020 = data_cleaned[data_cleaned['Year'] == 2020]

plt.figure(figsize=(10, 6))
sns.kdeplot(data_2020['Current Assets'].dropna(), fill=True, color="purple")
sns.rugplot(data_2020['Current Assets'].dropna(), color="purple")

plt.title('Density Plot for Current Assets (2020) with Distribution Rug')
plt.xlabel('Current Assets')
plt.ylabel('Density')

plt.show()

```



In the retail sector, different types of assets are crucial for ensuring a business's financial health and smooth operations. These assets can be grouped into three main categories, being fixed assets, net fixed assets and current assets, as shown by the density plots above, based on the columns from the dataset.

Fixed assets refer to long term investments like buildings and equipment. These are essential for running the business but aren't usually sold off. Net fixed assets represent the value of these fixed assets after accounting for the reduction in value overtime, giving a more accurate picture of what they're worth now. Current assets are short term resources like inventory that a business expects to turn into cash within a year.

When I look at the distribution of these assets in dataset, a pattern shows that most businesses hold low levels of these assets, while a few have much higher amounts. This suggests that the data set includes a wide range of businesses, from small businesses with minimal assets to large businesses with large investments. We also can account for outliers here due to different primary industries being present.

Fixed assets, such as buildings and equipment, are key for long term business operations. Companies with high fixed assets often own their retail locations rather than leasing them. For large retail chains like Aldi and Tesco, owning their stores reduces reliance on rent and thus cutting long term costs and increasing financial stability.

But, businesses with lower fixed assets might lease their premises or use temporary setups, minimising their investment in physical infrastructure. For example, Derrybeg Shop, which reported fixed assets of £6,946 in 2020, likely operates with minimal physical investments, possibly leasing space or sharing equipment. This low level of fixed assets helps keep overhead costs down, offering financial flexibility, though it may limit the business's ability to scale up or grow in the long term.

In retail, net fixed assets help assess the remaining 'life' of assets and how well the business maintains or reinvests. For example, McAnallen's of Benburb Limited, operating under the SPAR brand, reported net fixed assets of £1.2 million in 2021. This suggests a moderate investment in physical infrastructure, like store premises and necessary equipment.

As a SPAR franchisee, McAnallen's likely owns or has heavily invested in its store location and equipment to maintain operations. Their net fixed assets reflect the company's efforts to keep their physical assets in good shape. This significant investment indicates a commitment to maintaining a strong operational base, which is good for delivering consistent and reliable service to customers, potentially a cause for increased levels of customer satisfaction.

Current assets are essential for the day today operations of retail businesses, providing the liquidity needed to manage inventory, meet short term needs and respond to market demand. In the retail sector, where quickly selling and restocking inventory is crucial, having a high level of current assets can indicate a business's ability to manage its stock effectively and keep a good flow of sales opportunities.

For example, Centra Magheralin Ltd reported current assets of £48,308, which is close to the average for businesses in this sector. This suggests that Centra Magheralin operates with a small but effective inventory management system. Many small to medium sized retailers in this sector have similarly low levels of current assets, focusing on maintaining just enough inventory to meet customer demand without gaining too much capital.

Operating with lower levels of current assets, such as inventory, can be a strategic choice for businesses like Centra Magheralin, allowing them to manage resources efficiently and minimise holding costs. However, this approach also carries certain risks, including potential

stockouts, limited flexibility to respond to market fluctuations and increased dependency on a reliable supply chain. While it can optimise cash flow and reduce unnecessary expenditures, businesses must carefully balance these benefits against the potential downsides to ensure they can meet customer demand and sustain long term growth, thus affecting both financial performance and customer satisfaction.

The management of fixed assets, net fixed assets and current assets in the retail sector is deeply influenced by a company's priorities, which in turn affect key metrics. For example, a supermarket chain that aims to cultivate a good image may decide to invest heavily in owning and maintaining retail spaces in prime locations. SPAR does this by targeting local communities or more suburban parts of cities in the UK. This decision boosts fixed and net fixed assets.

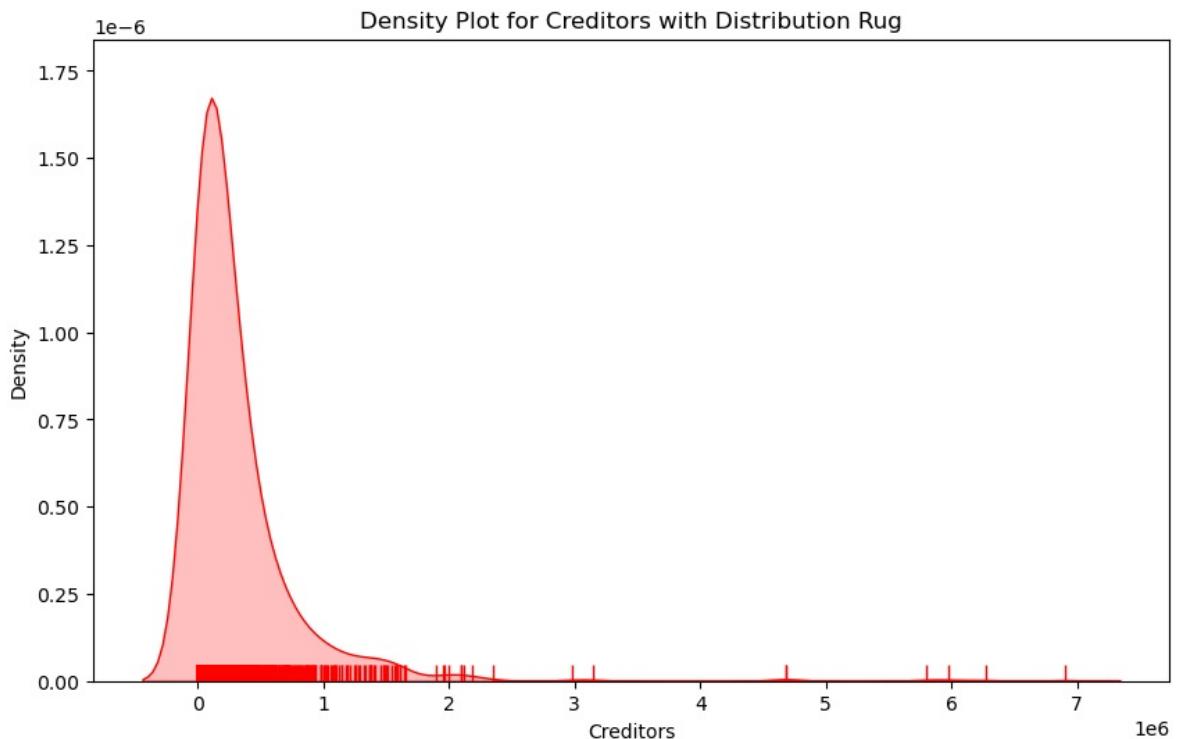
Although, a company focused on financial performance might choose to rent properties or limit investments in physical assets, keeping fixed asset levels low. This approach can improve financial performance by reducing expenditure. However, it might also result in a less distinctive supermarket image and lower customer satisfaction, as leased or minimally invested stores may not offer the same level of quality or consistency.

```
In [821]: # Code for density plot for creditors (all rows)
plt.figure(figsize=(10, 6))
sns.kdeplot(data_cleaned['Creditors'].dropna(), fill=True, color="red")

sns.rugplot(data_cleaned['Creditors'].dropna(), color="red")

plt.title('Density Plot for Creditors with Distribution Rug')
plt.xlabel('Creditors')
plt.ylabel('Density')

plt.show()
```



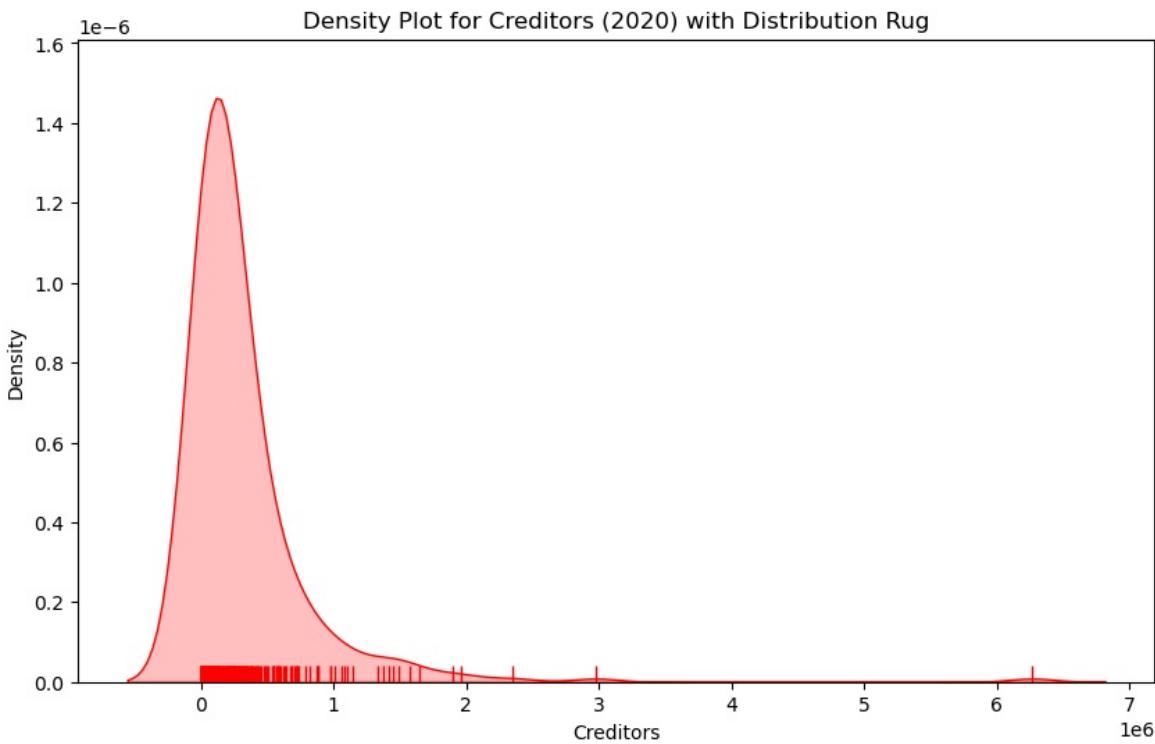
```
In [822]: # Code for density plot for creditors (only rows that are dated 2020(Year))
data_2020 = data_cleaned[data_cleaned['Year'] == 2020]

plt.figure(figsize=(10, 6))
sns.kdeplot(data_2020['Creditors'].dropna(), fill=True, color="red")

sns.rugplot(data_2020['Creditors'].dropna(), color="red")

plt.title('Density Plot for Creditors (2020) with Distribution Rug')
plt.xlabel('Creditors')
plt.ylabel('Density')

plt.show()
```



The term creditors refers to any individuals or entities to who the business owes money for goods or services received on credit. The efficient management of creditor balances is crucial for retail businesses, especially those operating on tight margins, as it allows them to optimise cash flow, maintain relationships with suppliers and ensure that they can meet their financial obligations without straining their resources.

The density plot for creditors in both 'all rows' and 2020 shows a clear right skewed distribution. This pattern indicates that most businesses within the retail sector have relatively low levels of creditors, while a smaller number of businesses owe significantly larger amounts.

Mulhollands Grocery Store Ltd, which has creditors amounting to approximately £32,000, is an example of a business on the lower end of the creditor spectrum within the retail sector. This relatively modest level of creditors suggests that Mulhollands operates with a cautious approach to managing its responsibilities, possibly prioritising prompt payment to suppliers and minimising its reliance on credit. By maintaining lower creditor balances, Mulhollands likely ensures stronger control over its cash flow, reducing the risks associated with higher debt levels, such as potential liquidity issues or strained relationships with suppliers.

Operating with a lower level of creditors can be particularly advantageous for smaller retailers like Mulhollands, as it allows them to maintain financial flexibility and avoid the financial strain that can come from excessive debt. However, this approach may also indicate a more conservative growth strategy, where the business wants to limit its credit use to manage costs and risks carefully. While this strategy supports long term financial stability, it might also constrain the business's ability to invest in expansion or take advantage of growth opportunities, which may be demonstrated by the fact that Mulhollands is an independent store.

On the other hand, the long tail of the distribution, where a few businesses hold relatively high creditor balances, likely represents larger retail chains or businesses that operate on a more extensive scale or outliers from businesses that have a different primary SIC code or simply miscellaneous results.

The level of creditors on a business's balance sheet can provide insights into its financial strategy and operational efficiency. Lower creditor balances might indicate that a business is paying its suppliers promptly, which can be a sign of strong cash flow management. It also suggests that the business may not be heavily relying on credit to finance its operations, thus reducing its exposure to interest costs.

Derrybeg Shop presents an interesting case when examining the changes in their financial metrics between 2020 and 2021. In 2020, the business reported creditors of £7,567 and fixed assets of £6,946. By 2021, these figures had changed significantly, with creditors increasing to £59,579 and fixed assets rising to £33,977. The substantial increase in both creditors and fixed assets suggests a period of significant investment and growth for Derrybeg Shop Ltd. The increase in fixed assets from £6,946 to £33,977 indicates that the company likely invested in long term assets such as property, equipment or renovations. However, this investment seems to have been financed largely through credit, as evidenced by the jump in creditors from £7,567 in 2020 to £59,579 in 2021. The fact that the creditors increased so significantly suggests that Derrybeg Shop may have utilised credit lines or other forms of borrowing to fund their fixed asset purchases.

The significant rise in creditors shows that Derrybeg Shop is using debt strategically to finance its growth. While this increases financial risk, it also provides opportunities for better financial performance. By taking on more debt, the shop likely aims to boost its operations and

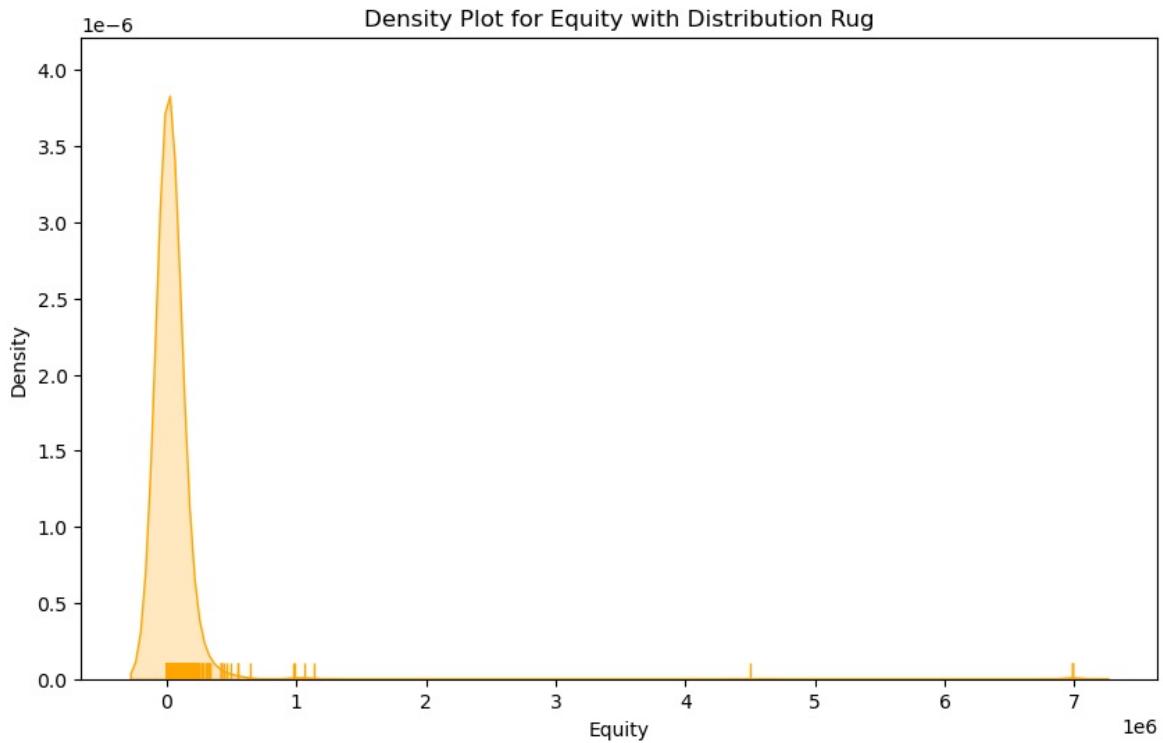
improve quality, inventory and customer experience, which could lead to higher sales and profits. The larger creditor balance indicates a planned decision to take on more debt now, expecting that the benefits like a better store or improved services will outweigh the costs of the debt.

This strategy also has indirect benefits on customer satisfaction and the store's image. The improvements funded by the increased creditors, such as enhanced facilities or expanded offerings, can lead to a better shopping experience for customers. As a result, customer satisfaction may increase, which in turn can enhance the store's reputation and potentially attract more business, further contributing to the financial performance and growth of the company.

```
In [824]: # Code for density plot for equity (all rows)
plt.figure(figsize=(10, 6))
sns.kdeplot(data_cleaned['Equity'].dropna(), fill=True, color="orange")
sns.rugplot(data_cleaned['Equity'].dropna(), color="orange")

plt.title('Density Plot for Equity with Distribution Rug')
plt.xlabel('Equity')
plt.ylabel('Density')

plt.show()
```

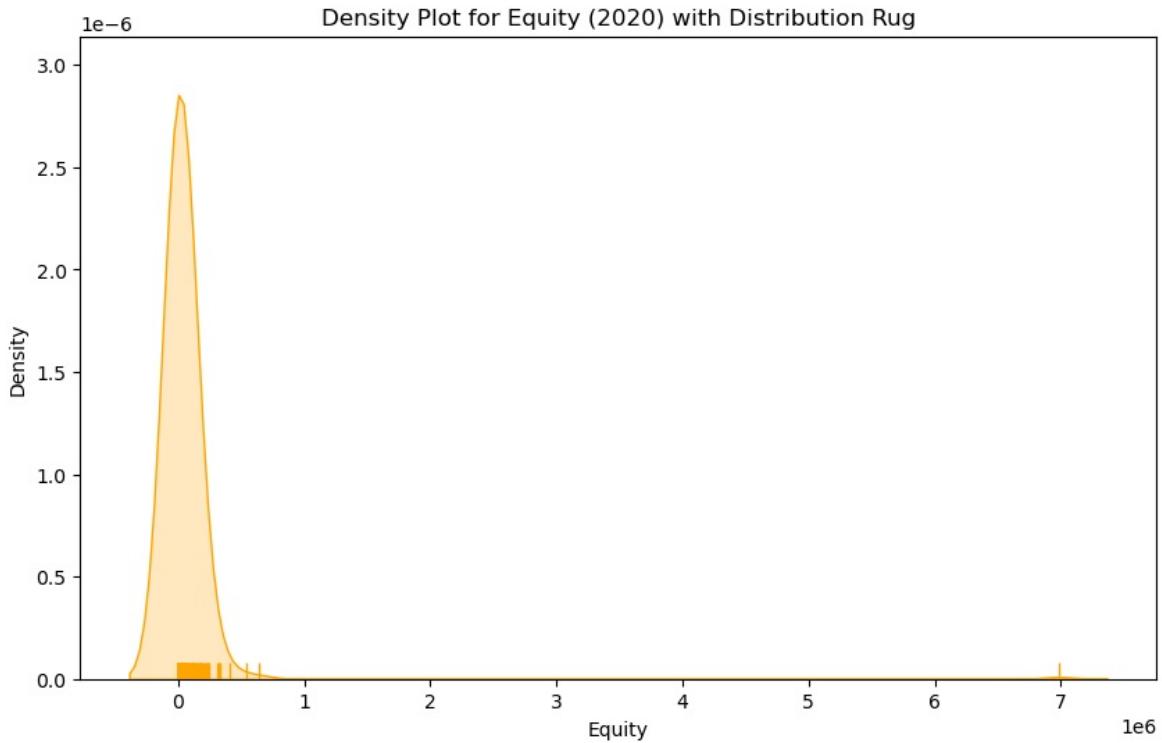


```
In [825]: # Code for density plot for equity (only rows that are dated 2020(Year))
data_2020 = data_cleaned[data_cleaned['Year'] == 2020]

plt.figure(figsize=(10, 6))
sns.kdeplot(data_2020['Equity'].dropna(), fill=True, color="orange")
sns.rugplot(data_2020['Equity'].dropna(), color="orange")

plt.title('Density Plot for Equity (2020) with Distribution Rug')
plt.xlabel('Equity')
plt.ylabel('Density')

plt.show()
```



Equity is a key measure of a company's financial strength, showing the ownership value held by shareholders. Both density plots for equity reveal a right skewed distribution, with the majority of businesses clustered around lower equity levels and a few outliers with significantly higher equity.

It's important to note that a significant number of entries in the dataset appear as an obscure value, such as '100' or '1', likely due to companies being required to fill in this field but either unwilling or unable to disclose their actual equity values. The consistent repetition of this number across multiple years suggests it was intentionally input. While other financial metrics fluctuate over time for many companies, these equity figures remain fixed at '1', for example.

Businesses with low equity typically operate with limited financial resources, which can strain their ability to invest in improvements or expansions. This often results in a more cautious approach to operations, focusing on maintaining daily procedures without needing to spend a lot of money. But, businesses with higher equity enjoy greater financial stability and flexibility. This allows them to invest in areas that directly impact customer satisfaction or their image such as upgrading facilities, expanding product offerings or enhancing customer service, regardless of the industry. Higher equity also provides a fallback during economic downturns, enabling these businesses to maintain operations without compromising on quality or financial performance.

An example from the dataset is the Parish Community Shop, with equity of £8,490 in 2021 and £10,494 in 2020, it has limited financial resources, which means it needs to be careful with how it spends money. This lower equity can actually help its financial performance because it encourages the shop to focus on running smoothly and not taking big risks. By managing its money wisely, the shop can consistently meet customer needs, which can lead to higher customer satisfaction. For small businesses, having less equity can encourage smart spending and help keep the business stable over time. However, this lower equity can also be a drawback, as it limits the shop's ability to invest in growth, new products or improvements, which could eventually affect its ability to compete and satisfy customers in the long run.

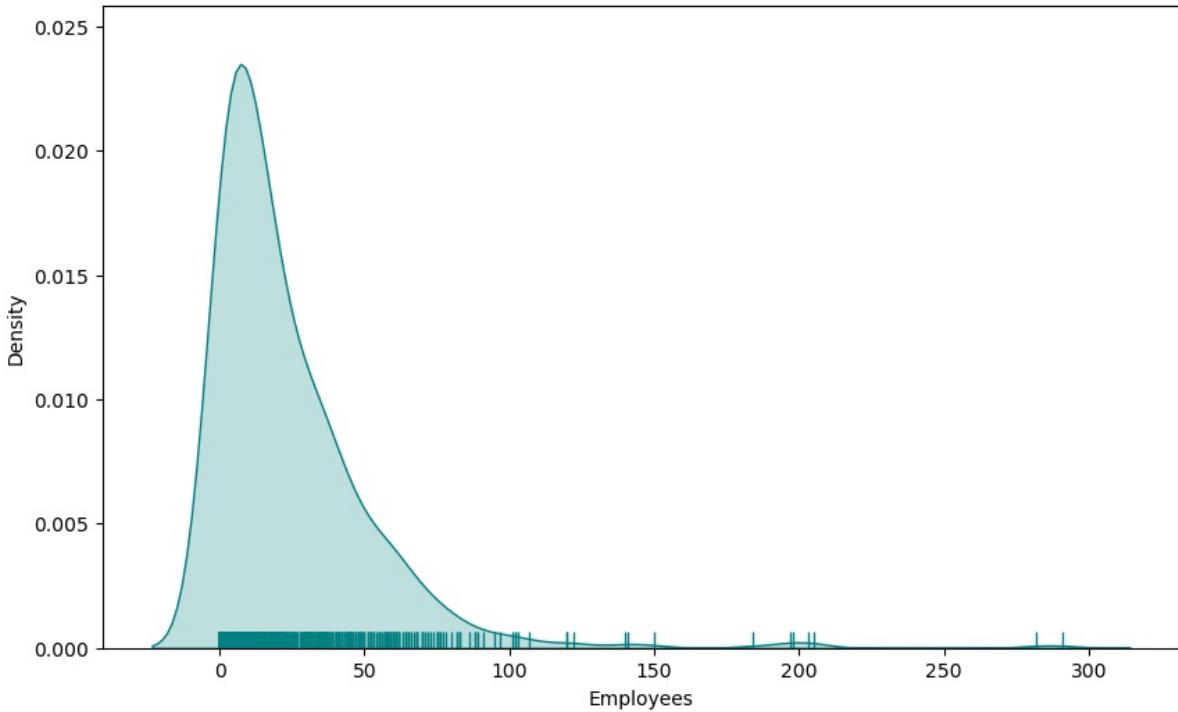
```
In [827]: # Code for density plot for employees (all rows)
plt.figure(figsize=(10, 6))
sns.kdeplot(data_cleaned['Employees'].dropna(), fill=True, color="teal")

sns.rugplot(data_cleaned['Employees'].dropna(), color="teal")

plt.title('Density Plot for Employees with Distribution Rug')
plt.xlabel('Employees')
plt.ylabel('Density')

plt.show()
```

Density Plot for Employees with Distribution Rug



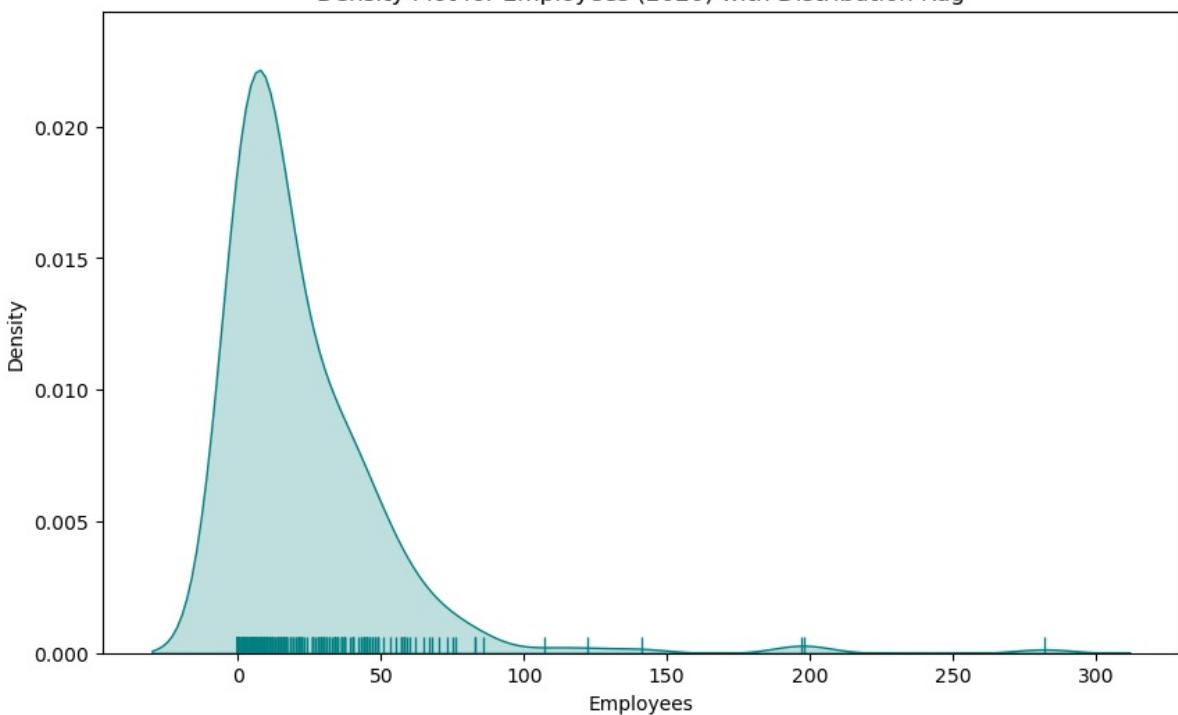
```
In [828]: # Code for density plot for employees (only rows that are dated 2020(Year))
data_2020 = data_cleaned[data_cleaned['Year'] == 2020]

plt.figure(figsize=(10, 6))
sns.kdeplot(data_2020['Employees'].dropna(), fill=True, color="teal")
sns.rugplot(data_2020['Employees'].dropna(), color="teal")

plt.title('Density Plot for Employees (2020) with Distribution Rug')
plt.xlabel('Employees')
plt.ylabel('Density')

plt.show()
```

Density Plot for Employees (2020) with Distribution Rug



The average number of employees a business employs is closely linked to its ability to deliver high levels of customer satisfaction, maintain a strong financial performance and cultivate a positive store image. Employees are the backbone of any retail operation, directly influencing how customers perceive the business, how efficiently operations are run, and ultimately, the financial health of the company.

The employee distribution in the retail sector, as shown by the density plots, reveals a strong right skewed pattern, with most businesses employing fewer than 50 employees. This distribution indicates that many retailers operate with lean staffing, which can have significant implications for their overall business outcomes, or may simply be small or medium sized businesses, as mentioned in previous density plot discussions.

Customer satisfaction is heavily influenced by the number of employees a retail store employs. Businesses with smaller workforces may struggle to provide the level of service that larger, better staffed stores can offer. In these smaller operations, each employee often has to do multiple roles, which can lead to longer wait times, less personalised service and a generally lower customer experience. However, in some cases, smaller stores can use their size to create a more intimate, community focused shopping experience, which can be highly valued by customers seeking personalised attention or simply convenience.

The number of employees also plays a crucial role in a business's financial performance. Maintaining a smaller workforce helps reduce labor costs, which is a significant expense for any retail operation. This approach can improve profit margins, particularly for small businesses operating on tight budgets. However, while this strategy can boost financial performance in the short term by keeping costs low, it must be balanced carefully. If taken too far, under staffing can negatively impact customer service, potentially leading to lower sales and harming long term financial performance and service quality.

The number of employees also directly impacts the public image of a retail store. Stores with a sufficient number of staff on the floor are generally perceived as well managed and customer focused, as discussed in previous chapters. This positive image can enhance the brand reputation, attract more customers, and increase customer retention rates. However, stores that appear understaffed can give off an impression of neglect or poor management, potentially damaging the store's reputation and deterring customers.

The comparison between Bell's Supermarket (Garvagh) and McBride Retail illustrates the varying workforce strategies employed by different types of retail businesses and how these strategies correlate with their operational scale. Bell's Supermarket, a more localised grocery store being a superstore, which seemingly targets less population dense areas, maintained an average of 22 employees from 2017 to 2020, with a slight dip to 21 in 2019. This consistent staffing level reflects a stable, community focused operation where personalised service and a close team are central. However, McBride Retail, a store that is part of a larger SPAR franchise, experienced a steady reduction in its average number of employees from 76 in 2018 to 68 in 2019, 48 in 2020, and further down to 42 in 2021. The decline in workforce starting in 2019 suggests that McBride Retail was already making adjustments in response to market conditions or internal strategies well before the significant disruptions of 2020 with the pandemic. This ongoing reduction indicates a continued focus on cost efficiency, possibly due to sustained changes in consumer behavior, such as increased online shopping or a strategic decision to cut down operations.

4.3: Normalising data

When assessing the financial performance of businesses, it is good to analyse specific financial stats that provide insights into how a company utilises its resources and manages its financial responsibilities. I did recap back to the chapter 2, but no metrics identified truly aligned with the information I had available, so I had to find new metrics of value to create. In my case, the focus is on two key metrics, them being fixed assets to employees and debt to equity. These metrics offer a detailed view of how well a business uses its physical assets relative to its workforce and how it balances debt with equity to manage financial risk. Through a conversation with ChatGPT, I decided these were the best metrics to use, especially with the data I had and could get hold of.

Retail businesses can range in size and scale. The value of fixed assets and the number of employees can differ significantly between these businesses. Without normalisation, comparing a small store with a large chain would be challenging, as the differences in scale would overshadow any meaningful insights into efficiency or financial performance.

Fixed assets to employees ratio provides insight into how efficiently a company is utilising its physical assets in relation to its workforce. In the retail sector, where physical assets like inventory are the main factor to operations, this ratio can indicate whether a business is over or under utilising its assets. For example, a higher ratio might suggest that the business is relying more on its assets, such as machinery and inventory, than its workforce, while a lower ratio could indicate that the business is more workforce focused.

Debt to equity ratio is a measure of financial stability and risk. It shows the extent of company's operations are financed through debt compared to equity. A higher debt to equity ratio suggests that a company is relying more on loans, which can increase financial risk. By normalising the debt to equity ratio, we can compare the financial risk across companies regardless of their size or the levels of debt and equity. This comparison is essential for understanding how different businesses manage their financial aspects and risks.

Furthering my conversation with ChatGPT, I was able to combine my knowledge with AI to produce code that would address these equations. I wanted to make two new columns with these ratios in them, but also found difficulty while practicing in another notebook regarding floating data and converting, as well as conditional logic that needed applied for rows that did not contain the data needed for the equation. Below is the code that was generated to create these new columns.

NB: In non-code terms, the equation for Fixed assets to employee ratio is Fixed assets divided by the average number of employees, within that row (year). The equation for Debt to equity ratio is creditors divided by equity.

```
In [832]: # calculate the fixed assets to employees ratio  
# np.where() applies the conditions:  
# the ratio is calculated only when both 'fixed assets' and 'employees' are not missing  
# and when 'employees' is not zero (to avoid division by zero)
```

```

# if the conditions are met, it calculates the ratio 'fixed assets' / 'employees'
# otherwise, it assigns np.nan (missing value)
data_cleaned['Fixed assets to employees'] = np.where(
    (data_cleaned['Fixed Assets'].notna()) & (data_cleaned['Employees'].notna()) & (data_cleaned['Employees'] != 0),
    data_cleaned['Fixed Assets'] / data_cleaned['Employees'],
    np.nan
)

# calculate the debt-to-equity ratio (creditors to equity ratio)
# np.where() applies the conditions:
# the ratio is calculated only when both 'creditors' and 'equity' are not missing
# and when 'equity' is not zero (to avoid division by zero)
# if the conditions are met, it calculates the ratio 'creditors' / 'equity'
# otherwise, it assigns np.nan (missing value)
data_cleaned['Debt to equity'] = np.where(
    (data_cleaned['Creditors'].notna()) & (data_cleaned['Equity'].notna()) & (data_cleaned['Equity'] != 0),
    data_cleaned['Creditors'] / data_cleaned['Equity'],
    np.nan
)

# round the new 'fixed assets to employees' column to one decimal place
data_cleaned['Fixed assets to employees'] = data_cleaned['Fixed assets to employees'].round(1)

# round the new 'debt to equity' column to one decimal place
data_cleaned['Debt to equity'] = data_cleaned['Debt to equity'].round(1)

# convert specified columns back to Int64 (pandas nullable integer type)
# this ensures that any previously calculated ratios or data manipulations don't affect the integer type consistency
data_cleaned['Fixed Assets'] = data_cleaned['Fixed Assets'].astype('Int64')
data_cleaned['Net Fixed Assets'] = data_cleaned['Net Fixed Assets'].astype('Int64')
data_cleaned['Current Assets'] = data_cleaned['Current Assets'].astype('Int64')
data_cleaned['Creditors'] = data_cleaned['Creditors'].astype('Int64')
data_cleaned['Equity'] = data_cleaned['Equity'].astype('Int64')
data_cleaned['Employees'] = data_cleaned['Employees'].astype('Int64')

# display the first few rows of the dataframe to verify the changes
data_cleaned.head()

```

Out[832]:

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	Net Fixed Assets	Current Assets
0	727 YOUR STORE LIMITED	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	£	11064	6425	647
1	727 YOUR STORE LIMITED	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	◆	13830	3648	175
2	727 YOUR STORE LIMITED	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	£	17288	509	35
3	727 YOUR STORE LIMITED	Active	Private limited Company	2013-08-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	£	17859	17323	1062
4	77 RETAIL 2 LIMITED	Active	Private limited Company	2016-03-03	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	337671	167892	4188

With the new columns now created, I shall produce some visualisations.

In [834]:

```

# set the figure plot size as usual
plt.figure(figsize=(10, 6))

# creating the density plot as before
sns.kdeplot(data_cleaned['Fixed assets to employees'].dropna(), fill=True, color="blue")

# adding a rug as before
sns.rugplot(data_cleaned['Fixed assets to employees'].dropna(), color="blue")

# setting titles and labels
plt.title('Density Plot for Fixed Assets to Employees with Distribution Rug')

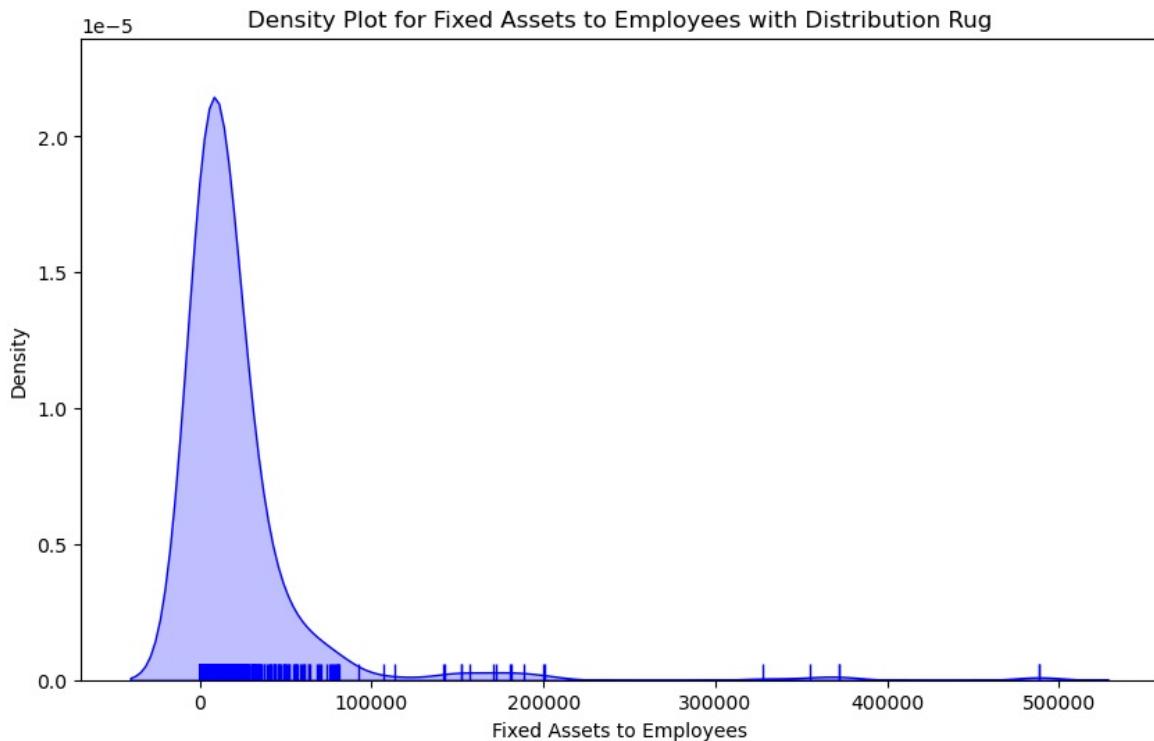
```

```

plt.xlabel('Fixed Assets to Employees')
plt.ylabel('Density')

# displaying the plot
plt.show()

```



The density plot above shows that, with regards to fixed assets to employee ratios, it is heavily right skewed indicating that the majority of businesses within my dataset have a relatively low ratio of fixed assets to employees, while a smaller subset of companies exhibit much higher ratios. This suggest that most of the businesses are either operating with minimal reliance on physical assets or the value of their assets, such as inventory or equipment, is low compared to the size of their workforce. This is considered normal in sectors where workforce is more valuable than physical assets, such as service-oriented retail operations, playing a key role in customer service, sales and day to day operations.

These companies may prefer to invest in their workforce rather than in physical assets, which allows them to remain adaptable in a competitive market, with new technologies and equipment emerging regularly, but having a solid workforce may be deemed more beneficial to the companies decision makers.

However, the outliers could skew stats like the mean, which might not accurately show the typical company in the dataset. Moreover, these high ratio businesses may follow a different operational strategy compared to those with lower ratios, as well as primarily identifying with another industry. These companies may be exposed to greater financial risks, especially if market conditions change and the value of their assets decreases.

It is worth while to perform deeper analysis on understanding the underlying values of the companies in the outliers and thus being able to comprehend what drives these ratios and how they impact other financial performance metrics such as profit.

```

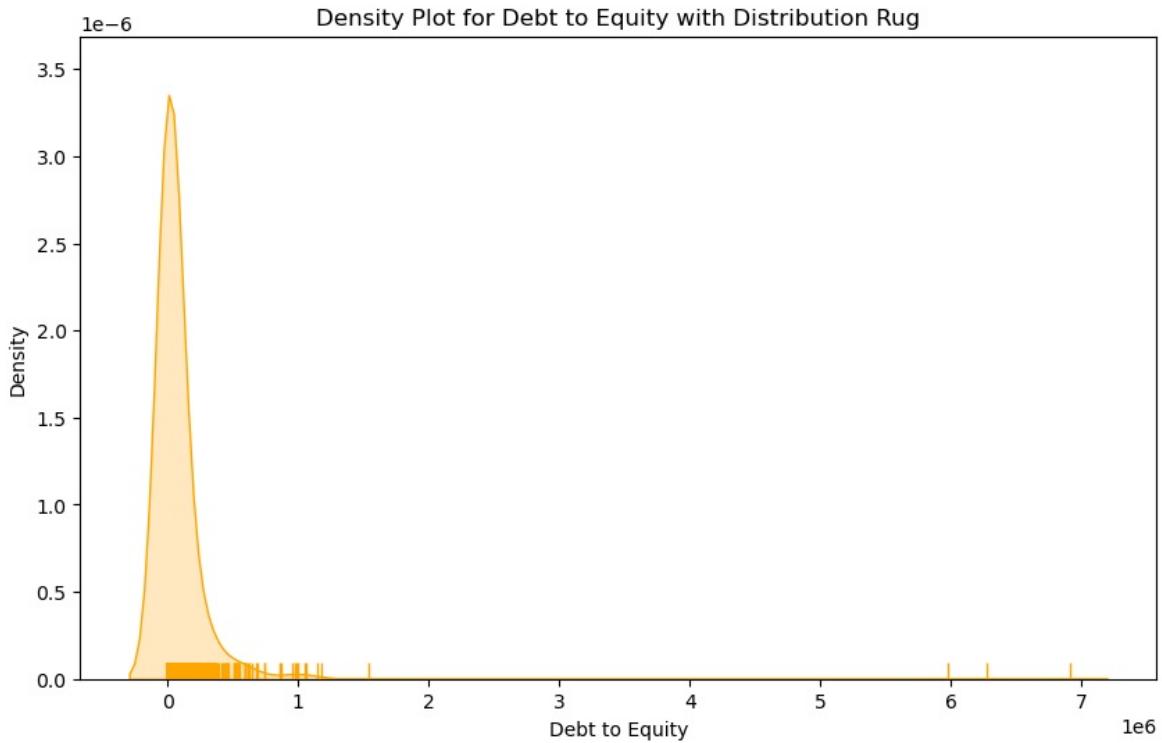
In [836]: # same as before, but for Debt to equity column
plt.figure(figsize=(10, 6))

sns.kdeplot(data_cleaned['Debt to equity'].dropna(), fill=True, color="orange")
sns.rugplot(data_cleaned['Debt to equity'].dropna(), color="orange")

plt.title('Density Plot for Debt to Equity with Distribution Rug')
plt.xlabel('Debt to Equity')
plt.ylabel('Density')

plt.show()

```



The density plot above relating to the debt to equity ratio reveals a highly right skewed distribution. This indicates that the vast majority of companies in the dataset have low levels of debt relative to their equity.

Companies with lower debt to equity ratios might rely more on their own funds because they have fewer opportunities or resources to take on debt, or they might operate in industries where growth is slower and more steady, reducing the need for aggressive borrowing. For example, in the supermarket or grocery industry, smaller shops typically don't expand quickly, if at all. Growth for these businesses is often slow and incremental, as the market is dominated by large supermarket chains, making it difficult for small shops to scale significantly. This lack of rapid growth opportunities means they are less likely to take on debt and more likely to focus on sustaining their operations with their own resources, which also limits their ability to compete on a larger scale.

Many businesses in the dataset seemingly prefer to use their own funds or equity to run their operations, keeping their debt levels low. This approach helps them stay financially stable and reduces the risk of running into financial problems, especially when the economy is unpredictable. In contrast, a smaller group of companies have much higher debt compared to their equity. These businesses borrow more money to finance their activities, which can be a smart move if they operate in a stable industry, potentially increasing their profits, making it a good strategy for those looking to grow quickly. However, this also comes with higher financial risk, such as changes in interest and fluctuation in income. The analysis of these companies could provide insights into the effectiveness of such 'aggressive' strategies and their impact on financial performance, particularly in comparison to the more cautious companies.

4.4: Subgroups & Outliers

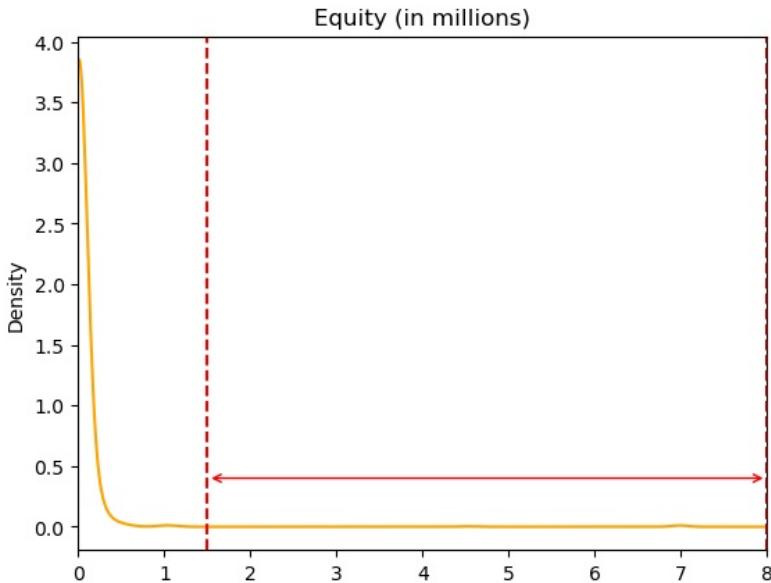
I will now focus on examining subgroups and outliers. While working through Sections 4.2 and 4.4, I took note of the density plots that displayed prominent outliers based on visual analysis. Due to time constraints and the need to highlight key findings, I have not included all the univariate columns as before. Instead, I concentrated on the ones that appeared to be the most significant, particularly those with noticeable outliers. Also, as I delved deeper in the findings, I found a recurring theme with these 'subgroups' that could be applied to the rest of the density plots, yet is not truly accurate.

Firstly, as observed in Section 4.2, the equity density plot shows a prominent peak near zero that rapidly slopes downward, leveling off at around 0.5 Million. However, there is a long tail with a small bump, suggesting the presence of outliers or potential subgroups that need further investigation. To explore this further, I will plot the density again, this time using vertical lines (axvline) to highlight the areas where these outliers likely exist, focusing on the regions farther from the peak or the steep slope.

```
In [840]: ax = (data_cleaned['Equity'] / 1e6).plot(kind='density', xlim=(0, 8), title='Equity (in millions)', color='orange')
ax.axvline(x=1.5, color='red', linestyle='--')
ax.axvline(x=8, color='red', linestyle='--')

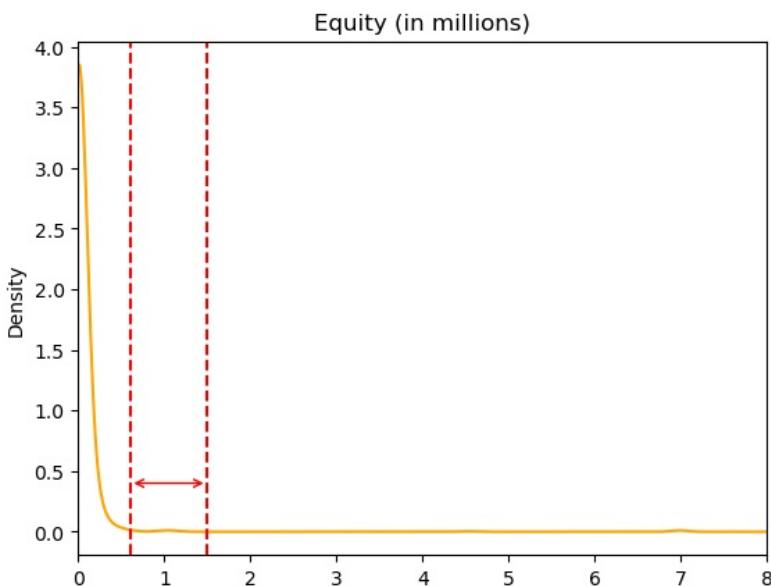
ax.annotate(' ', xy=(1.5, .4), xytext=(8, .4), arrowprops=dict(arrowstyle='<->', color='red'))

plt.show()
```



```
In [841]: ax = (data_cleaned['Equity'] / 1e6).plot(kind='density', xlim=(0, 8), title='Equity (in millions)', color='orange')
ax.axvline(x=0.6, color='r', linestyle='--')
ax.axvline(x=1.5, color='r', linestyle='--')

ax.annotate(' ', xy=(0.6, .4), xytext=(1.5, .4), arrowprops=dict(arrowstyle='<->', color='r'))
plt.show()
```



Having marked these sections on the plot, the next step is to examine the specific entries that fall within these identified outlier regions. As indicated by the rug plot in Section 4.2, there appear to be relatively few data points in these areas. By displaying these entries, I aim to gain a clearer understanding of the characteristics and potential significance of these outliers. This deeper analysis will help determine whether these points represent a distinct subgroup or are simply anomalies in the dataset.

```
In [843]: from IPython.display import display

equity_filtered_1 = data_cleaned[(data_cleaned['Equity'] >= 1.5e6) & (data_cleaned['Equity'] <= 8e6)]
print(f"Number of data points between 1.5M and 8M: {len(equity_filtered_1)}")
display(equity_filtered_1.head())

equity_filtered_2 = data_cleaned[(data_cleaned['Equity'] >= 0.6e6) & (data_cleaned['Equity'] <= 1.5e6)]
print(f"Number of data points between 0.6M and 1.5M: {len(equity_filtered_2)}")
display(equity_filtered_2.head())

equity_filtered_3 = data_cleaned[(data_cleaned['Equity'] >= 0) & (data_cleaned['Equity'] <= 0.6e6)]
print(f"Number of data points between 0M and 0.6M: {len(equity_filtered_3)}")
```

Number of data points between 1.5M and 8M: 4

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	F As
231	CHARIOTEER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2020	Not Available	6591322	4096
232	CHARIOTEER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2019	Not Available	6727431	310'
233	CHARIOTEER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2018	Not Available	<NA>	2760
870	MONTALTO PROPERTIES LIMITED	Active	Private limited Company	1999-08-23	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	68202 - Letting and operating of conference an...	91030 - Operation of historical sites and buil...	2018	£	7553364	4500

Number of data points between 0.6M and 1.5M: 5

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	F As
22	ACE SCAFFOLDING BALLYMONEY LIMITED	Active	Private limited Company	2017-07-26	43991 - Scaffold erection	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN	2020	£	352828	644
186	C & O MCGINLEY LIMITED	Active	Private limited Company	2009-09-28	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	56103 - Take-away food shops and mobile food s...	NaN	2017	£	1271517	988
362	DRUMHARVEY LIMITED	Active	Private limited Company	2011-01-12	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	1843182	1531
363	DRUMHARVEY LIMITED	Active	Private limited Company	2011-01-12	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	Not Available	1887071	1473
364	DRUMHARVEY LIMITED	Active	Private limited Company	2011-01-12	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	1931409	1453

Number of data points between 0M and 0.6M: 1155

This analysis considers outliers on both the left and right sides of zero, including those within specific categories or the bumps observed in the distribution. According to the output, there are 4 entries between '£1.5M and '£8M, 5 entries between '£0.6M and '£1.5M, and 1155 entries between '£0M and '£0.6M, excluding those marked as (N/A) within the dataset.

To delve deeper, I will examine some of the companies within these ranges. Starting with the '£1.5M to '£8M category, two businesses stand out, 'CHARIOTEER LIMITED' and 'MONTALTO PROPERTIES LIMITED'. Both companies primarily identify with the 47100 SIC code but also have 47190 as their secondary SIC code, indicating "Other retail sale in non specialised stores."

Looking further into Charioteer Limited, it becomes clear that this company is not just a typical 'store'. Until the 2000s, it operated under the name 'CEDARMOUNT CONSULTANTS LIMITED' (<https://find-and-update.company-information.service.gov.uk/company/NI038821>). According to their strategic report, their main activity revolves around the retail sale of automotive fuel, and they operate a supermarket. Additionally, Charioteer Limited is the parent company of several Spar stores, particularly those attached to petrol stations, indicating that their operations extend beyond just running a supermarket. For example, their investment in a new Spar store in Dundonald highlights their role in retail, particularly in convenience stores linked to fuel stations (<https://www.newsletter.co.uk/business/new-spar-store-has->

[opened-its-doors-in-dundonald-thanks-to-ps22-million-investment-and-creating-22-jobs-4227297](#).

Additionally, the company's current assets (432446) are much greater than its current liabilities (creditors of 138538), which indicates that the company has sufficient resources available to cover its short term responsibilities, ensuring it can meet its financial requirements without issues. This excess of assets over liabilities not only secures the company in the event of liquidation but also directly boosts its overall equity value. Also, the relatively low debt to equity ratio of 0.2 highlights that the company has financed its operations more through equity than debt. This financial strategy reduces financial risk and increases equity as the company retains more earnings within the business rather than distributing them as interest payments.

Moving away from statistics, their dual focus on automotive retail and convenience store management places the company on a larger and more diversified scale compared to many others in the same industry, likewise in the dataset. As a result, the financial figures, including the high equity values, are not due to any mistakes but instead accurately reflect the company's extensive operations and important role in the market.

Now, for the £'0.6M to £'1.5M range, where I will investigate 'Ace Scaffolding Ballymoney Limited.' Their primary SIC code is 43991, which indicates their involvement in scaffold erection, while their secondary SIC code is 47110. It is clear that this company operates primarily as a scaffolding business. Scaffolding companies like Ace Scaffolding Ballymoney Limited specialise in providing temporary structures used to support workers and materials during the construction, maintenance, or repair of buildings and other structures (<https://www.acescaffoldingballymoney.co.uk>).

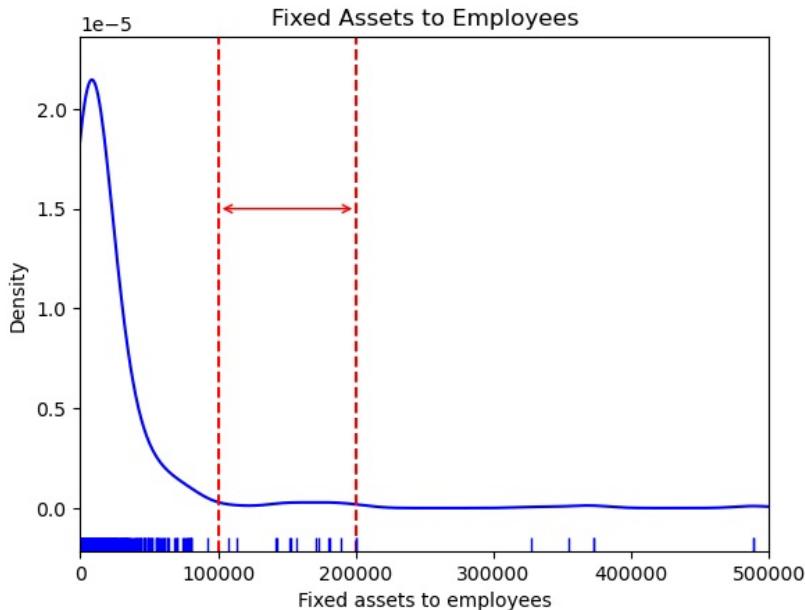
As a scaffolding company, Ace Scaffolding Ballymoney Limited needs to invest heavily in fixed assets, such as scaffolding equipment. These assets are key for their operations and must be durable and reliable to meet the demands of their projects. The company's fixed assets are valued at £'352828, indicating a significant amount of money in essential equipment. This is typical for businesses in construction.

Furthermore, the net fixed assets are reported at £'644741, which shows that even after accounting for loss of value over time on the assets, the value of these long term assets remains high, contributing to their overall equity.

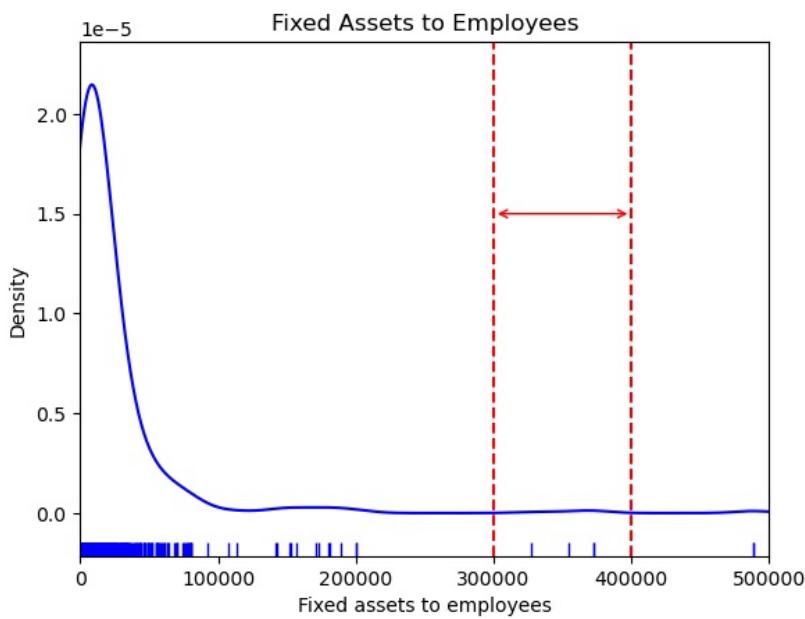
While these figures are likely reflective of Ace Scaffolding Ballymoney Limited's actual financial situation, it is worth noting that this company is an outlier compared to others in the dataset. Unlike most other companies, which are primarily retail stores, this company operates in a subset of the construction industry, as indicated by its primary SIC code. This distinction explains why the company's financial structure and equity levels differ significantly from those of typical retail sector.

Now, I will examine the Fixed Assets to Employees ratio, considering that Section 4.3 reveals potential outliers and subgroups based on a visual inspection. I will begin by highlighting the specific ranges we will focus on, as I have done before, only this time I will also introduce the rug due to subgroups being harder to distinguish. Any outliers caused by skewed results in either fixed assets or employee numbers should, in theory, be caught here.

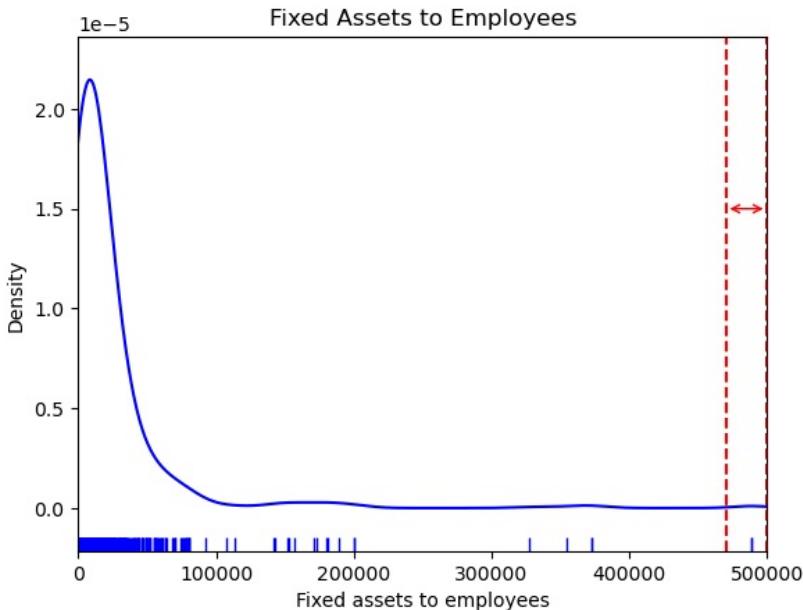
```
In [846]: ax = (data_cleaned['Fixed assets to employees']).plot(kind='density', title='Fixed Assets to Employees', color="#4CAF50")
ax.set_xlim(0, 500000)
ax.axvline(x=100200, color='r', linestyle='--')
ax.axvline(x=200300, color='r', linestyle='--')
sns.rugplot(data_cleaned['Fixed assets to employees'].dropna(), color="blue")
ax.annotate(' ', xy=(100200, 0.000015), xytext=(200300, 0.000015), arrowprops=dict(arrowstyle='<->', color='r'))
plt.show()
```



```
In [847]: ax = (data_cleaned['Fixed assets to employees']).plot(kind='density', title='Fixed Assets to Employees', color='blue')
ax.set_xlim(0, 500000)
ax.axvline(x=300000, color='r', linestyle='--')
ax.axvline(x=400000, color='r', linestyle='--')
sns.rugplot(data_cleaned['Fixed assets to employees'].dropna(), color="blue")
ax.annotate(' ', xy=(300000, 0.000015), xytext=(400000, 0.000015), arrowprops=dict(arrowstyle='<->', color='r'))
plt.show()
```



```
In [848]: ax = (data_cleaned['Fixed assets to employees']).plot(kind='density', title='Fixed Assets to Employees', color='blue')
ax.set_xlim(0, 500000)
ax.axvline(x=470000, color='r', linestyle='--')
ax.axvline(x=500000, color='r', linestyle='--')
sns.rugplot(data_cleaned['Fixed assets to employees'].dropna(), color="blue")
ax.annotate(' ', xy=(470000, 0.000015), xytext=(500000, 0.000015), arrowprops=dict(arrowstyle='<->', color='r'))
plt.show()
```



Given the presence of a very long and distant outlier, I decided to focus specifically on the results within the '£470,000 to '£500,000 range for closer manual investigation. Below, you can see the entries within this range.

```
In [850]: filtered_rows = data_cleaned[(data_cleaned['Fixed assets to employees'] >= 470000) &
                               (data_cleaned['Fixed assets to employees'] <= 500000)]

print(f"Displaying rows where 'Fixed assets to employees' is between 470,000 and 500,000:")
display(filtered_rows)
```

Displaying rows where 'Fixed assets to employees' is between 470,000 and 500,000:

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	Net Fixed Assets
1182	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	£	488447	412772
1183	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	£	488447	412774

The statistics for SuperValu Aughnacloy suggest that the data may not accurately reflect the true nature of the business. The record shows only one employee, which is highly unlikely for a retail operation of this scale, particularly one associated with a well known brand like SuperValu. Additionally, the "2" listed under current assets is abnormal and unrealistic, as it does not align with the financial structure typically observed in similar businesses. This suggests the possibility of human error in data entry, where values might have been filled in simply because the field required an input, rather than reflecting the actual financial situation.

Given these inconsistencies, it is clear that this entry is an outlier, but not in the sense of a store with a different operational status or a unique business model. Instead, it appears to be an unrealistic outlier, likely caused by inaccuracies during data entry. These factors suggest that the figures for SuperValu Aughnacloy should be interpreted with caution, as they do not provide a reliable representation of the company's actual operations or financial health.

Additionally, further investigation (see below) reveals that there are other years with similar types of data inconsistencies, though they lack the necessary information to perform accurate calculations, which is why they are not displayed. This pattern suggests that incomplete or inaccurate data entries are a recurring issue for SuperValu Aughnacloy, indicating that omitting or inaccurately reporting data is not uncommon for this business.

```
In [852]: filtered_rows = data_cleaned[data_cleaned['Company Name'] == 'SUPERVALU AUGHNACLOY LTD']

print(f"Displaying all rows for 'SuperValu Aughnacloy LTD':")
display(filtered_rows)
```

Displaying all rows for 'SuperValu Aughnacloy LTD':

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	Net Fixed Assets
						(SIC) 1	(SIC) 2	(SIC) 3				
1182	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	£	488447	412772
1183	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	£	488447	412774
1184	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	£	488447	412774
1185	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	£	488447	412774
1186	SUPERVALU AUGHNACLOY LTD	Active	Private limited Company	2011-02-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	<NA>	<NA>

Now, I will look at the other two ranges by breaking them down into smaller subgroups. This helps to understand the data better and spot any unusual patterns or outliers that might not be obvious when looking at the entire dataset. By focusing on specific ranges of the 'Fixed assets to employees' ratio, I can create clear visualisations, like density plots, and closely examine the individual data points within each subgroup, using temporary tables to group these values into one descriptiv plot.

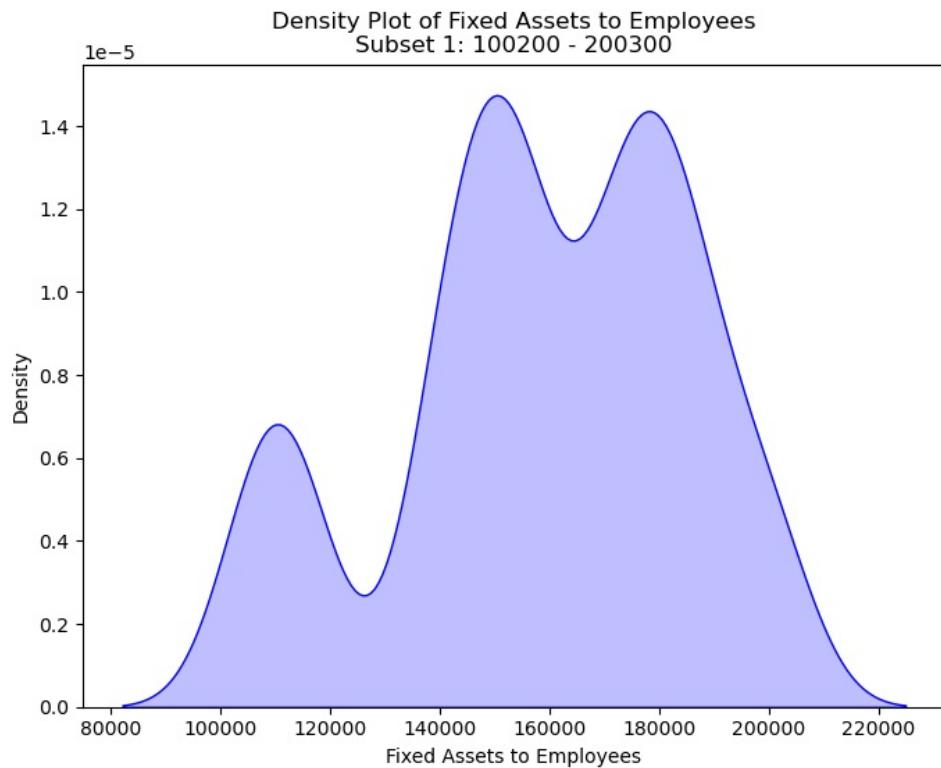
```
In [854]: # defining subgroups and their ranges
subgroups = [
    {'start': 100200, 'end': 200300, 'label': '100200 - 200300'},
    {'start': 300000, 'end': 400000, 'label': '300000 - 400000'}
]

# for loop to loop over the subgroups
for i, group in enumerate(subgroups, start=1):
    # filter data in subgroup
    subgroup_filtered = data_cleaned[(data_cleaned['Fixed assets to employees'] >= group['start']) &
                                      (data_cleaned['Fixed assets to employees'] <= group['end'])]

    # density plot
    plt.figure(figsize=(8, 6))
    sns.kdeplot(subgroup_filtered['Fixed assets to employees'], fill=True, color='blue', bw_adjust=0.5)
    plt.title(f'Density Plot of Fixed Assets to Employees\nSubset {i}: {group["label"]}')
    plt.xlabel('Fixed Assets to Employees')
    plt.ylabel('Density')

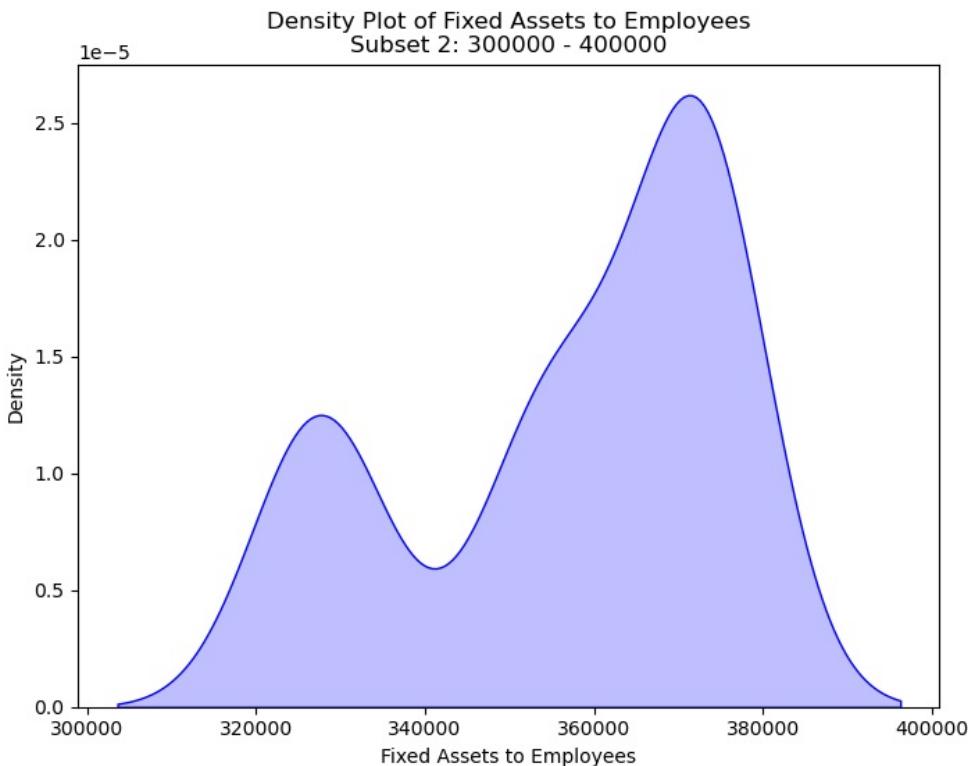
    plt.show()

print(f"\nAll rows of the subset {i} with 'Fixed assets to employees' between {group['start']} and {group['end']}")
display(subgroup_filtered)
```



All rows of the subset 1 with 'Fixed assets to employees' between 100200 and 200300:

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	A
194	CAMERON STEWART (NORTHERN IRELAND) LIMITED	Active	Private limited Company	1999-03-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	2799336	168
195	CAMERON STEWART (NORTHERN IRELAND) LIMITED	Active	Private limited Company	1999-03-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	Not Available	1842425	139
196	CAMERON STEWART (NORTHERN IRELAND) LIMITED	Active	Private limited Company	1999-03-05	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	1880306	118
231	CHARIOOTER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2020	Not Available	6591322	409
232	CHARIOOTER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2019	Not Available	6727431	310
673	LEINAD RETAIL LTD	Active	Private limited Company	2014-03-24	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	428077	8
710	M1 SERVICE AREA LIMITED	Active	Private limited Company	2003-09-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	997292	78
861	MILLVALE SERVICES LIMITED	Active	Private limited Company	1992-05-29	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN	NaN	2021	Not Available	1828070	162
862	MILLVALE SERVICES LIMITED	Active	Private limited Company	1992-05-29	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN	NaN	2020	Not Available	1825565	158
928	NTH RETAIL LTD	Active	Private limited Company	2009-03-23	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	180310	8
1134	SESSIA LTD	Active	Private limited Company	2008-04-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	8374474	858
1135	SESSIA LTD	Active	Private limited Company	2008-04-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	8893972	861
1136	SESSIA LTD	Active	Private limited Company	2008-04-28	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	9249432	874



All rows of the subset 2 with 'Fixed assets to employees' between 300000 and 400000:

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	% Fix Assets
674	LEINAD RETAIL LTD	Active	Private limited Company	2014-03-24	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	372299	461
675	LEINAD RETAIL LTD	Active	Private limited Company	2014-03-24	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	372299	267
709	M1 SERVICE AREA LIMITED	Active	Private limited Company	2003-09-06	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	982751	7975
871	MONTALTO PROPERTIES LIMITED	Active	Private limited Company	1999-08-23	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	68202 - Letting and operating of conference an...	91030 - Operation of historical sites and buil...	2017	Not Available	4966162	34524

Firstly, in the £'300000 to £'400000 range, I observe that Leinad Retail oversees several Spar locations, as confirmed by matching the office address (<https://www.spar-ni.co.uk/store-locator-ni/hen93525-spar-ballynure-road/>). The reported average of just one employee is likely referring to the director or possibly serving as a placeholder. This skews the fixed assets to employees ratio, especially given the company's substantial size and fixed assets exceeding £'300,000. Since the division is by just one, this results in an inflated ratio of 372299 for fixed assets to employees. This highlights an outlier caused by human error or a misrepresentation of employee numbers, reflecting the distinction between the stores and the parent company, Leinad Retail. The presence of information to calculate this ratio means the row was filled in, making it an outlier due to data inaccuracies rather than a true subgroup based on different business operations or financial strategies.

This pattern is similar within the M1 Service Area company, and thus the same issue applies. With help from Perplexity AI I found that M1 Service Area operates motorway service areas along the M1 motorway, managed by Applegreen (<https://motorwayservices.uk/Lisburn>). Thus the '14' employees compared to the nearly half a million in Fixed assets does not add up.

This pattern is also evident with the M1 Service Area company, and the same issue arises. With assistance from Perplexity AI, I discovered that M1 Service Area operates motorway service areas along the M1 motorway, managed by Applegreen (<https://motorwayservices.uk/Lisburn>). Given this context, the reported 14 employees compared to almost 1 Million in fixed assets seems disproportionate. This disproportion suggests that the data may not accurately reflect the true operational scale of the business,

suggesting possible data entry errors or misrepresentations similar to those seen in the 300000 to 4000000 outlier case.

Moving on, looking at the 100200 to 200300 range, I observed that there were 7 individual companies present. So I decided to home in on some of these. Firstly, on the upper end, Cameron Stewart was observed to have a fixed asset to employee ratio of 199952.2. I know from previous knowledge of this sector that they overlook all sites in certain areas, thus their employees are not entirely 14, but only accounting for maybe those in the large head office, not the actual stores. So again, this is a case of misrepresentation.

Moving on to the 100200 to 200300 range, I observed that there were 7 individual companies within this bracket. I decided to take a closer look at some of them. At the upper end, Cameron Stewart was found to have a fixed assets to employee ratio of 199952.2. Based on my prior knowledge of this sector, I know that they oversee Nisa stores in certain areas. Therefore, the reported 14 employees likely only reflect those working in the large head office, not the employees in the actual stores. This suggests another instance of misrepresentation. As I continue my analysis, I found that for companies like Sessia Ltd and Charioteer Limited, the average employee numbers provided more than likely reflect the actual staff within those companies. However, since these companies oversee multiple businesses, the data can be misleading. Again, we only see the overall financial information for all the stores or locations they manage, not the individual details of each store. This means that while the employee counts are likely accurate, they only account for the staff directly employed by the parent company, not those working in the businesses they oversee. This distinction can skew the results, as it doesn't fully capture the broader workforce involved in their operations, and thus leading these outliers.

Now, I will investigate the debt to equity ratio, as the previous density plot revealed slight bumps in the tail. To begin, I will outline the specific sections I will focus on, using the rug plot to identify some very small outliers for further analysis, as I had previously observed them in the first density plot within 4.3. Any outliers caused by skewed results in either creditors or equity columns should be seen here.

```
In [857]: ax = (data_cleaned['Debt to equity']).plot(kind='density', title='Debt to Equity', color='red')

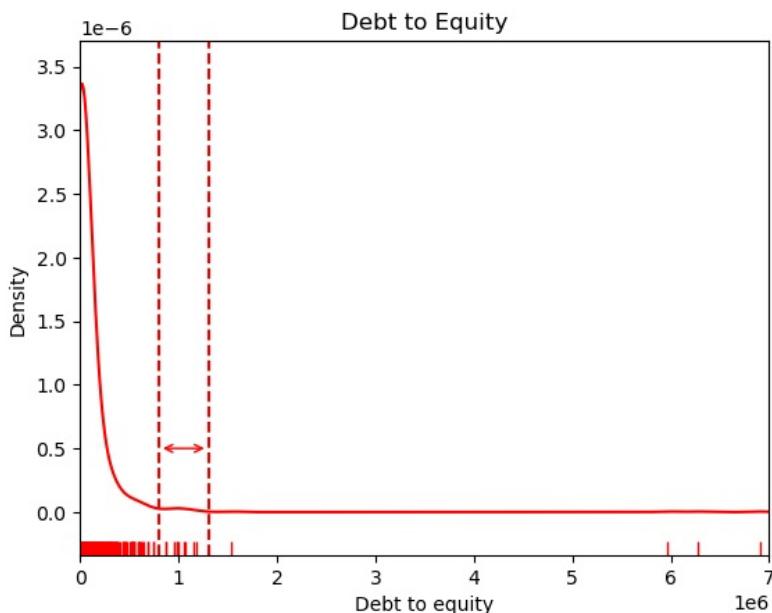
ax.set_xlim(0, 7e6)

ax.axvline(x=0.8e6, color='r', linestyle='--')
ax.axvline(x=1.3e6, color='r', linestyle='--')

sns.rugplot(data_cleaned['Debt to equity'].dropna(), color="red")

ax.annotate(' ', xy=(0.8e6, 0.0000005), xytext=(1.3e6, 0.0000005), arrowprops=dict(arrowstyle='<->', color='r'))

plt.show()
```



```
In [858]: ax = (data_cleaned['Debt to equity']).plot(kind='density', title='Debt to Equity', color='red')

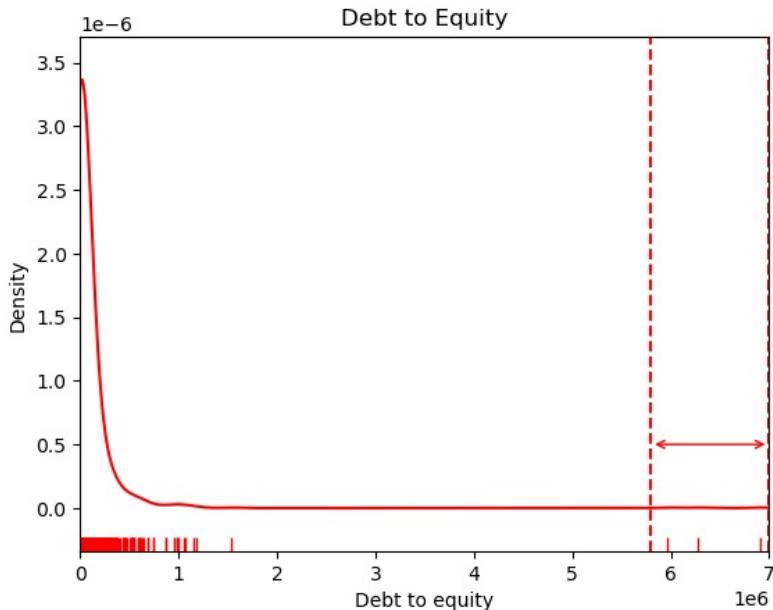
ax.set_xlim(0, 7e6)

ax.axvline(x=5.8e6, color='r', linestyle='--')
ax.axvline(x=7e6, color='r', linestyle='--')

sns.rugplot(data_cleaned['Debt to equity'].dropna(), color="red")

ax.annotate(' ', xy=(5.8e6, 0.0000005), xytext=(7e6, 0.0000005), arrowprops=dict(arrowstyle='<->', color='r'))

plt.show()
```



Before moving forward, I want to focus on the small rug between 1 and 2 on the plot. This represents McDade Retail, which, according to the rows presented below, has an equity of just 1. This figure is clearly unmeasurable given the company's substantial other financial metrics, and the fact they they overlook Centra's (<https://neighbourhoodretailer.com/mcdades-centra-clogher-hungry-for-more/>). As a result, dividing the creditors by this equity value produces a debt to equity ratio of 1539340.0. This is likely due to the company either not calculating or not reporting their actual equity, rather than it being a true reflection of their financial position.

```
In [860]: filtered_rows = data_cleaned[(data_cleaned['Debt to equity'] >= 1.3e6) &
                               (data_cleaned['Debt to equity'] <= 2e6)]
print(f"Displaying rows where 'Debt to Equity' is between 1.3e6 and 2e6:")
display(filtered_rows)
```

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	Net Fixed Assets	C
818	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	3788825	1058326	11

Additionally, further investigation (as shown below) reveals that all entries for McDade Retail Ltd over the years consistently report an equity of '1'. It's worth noting that three of these entries will not appear in the report because they lack the creditor information needed to calculate the debt to equity ratio.

```
In [862]: filtered_rows = data_cleaned[data_cleaned['Company Name'] == 'MCDADE RETAIL LTD']
print(f"Displaying all rows for 'McDade Retail LTD':")
display(filtered_rows)
```

Displaying all rows for 'McDade Retail LTD':

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	Net Fixed Assets	CA
818	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	3788825	1058326	11
819	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	3754432	991778	8
820	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	9136	808675	8
821	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	Not Available	14618	477877	13
822	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	20099	266597	8

Now, I will examine the other two subgroups I have identified within the debt to equity density plot, via the use of temporay tables to home in o the distribution within these sections.

```
In [864]: # same code as shown before
subgroups_e = [
    {'start': 0.8e6, 'end': 1.3e6, 'label': '0.8e6 - 1.3e6'},
    {'start': 5.8e6, 'end': 7e6, 'label': '5.8e6 - 7e6'}

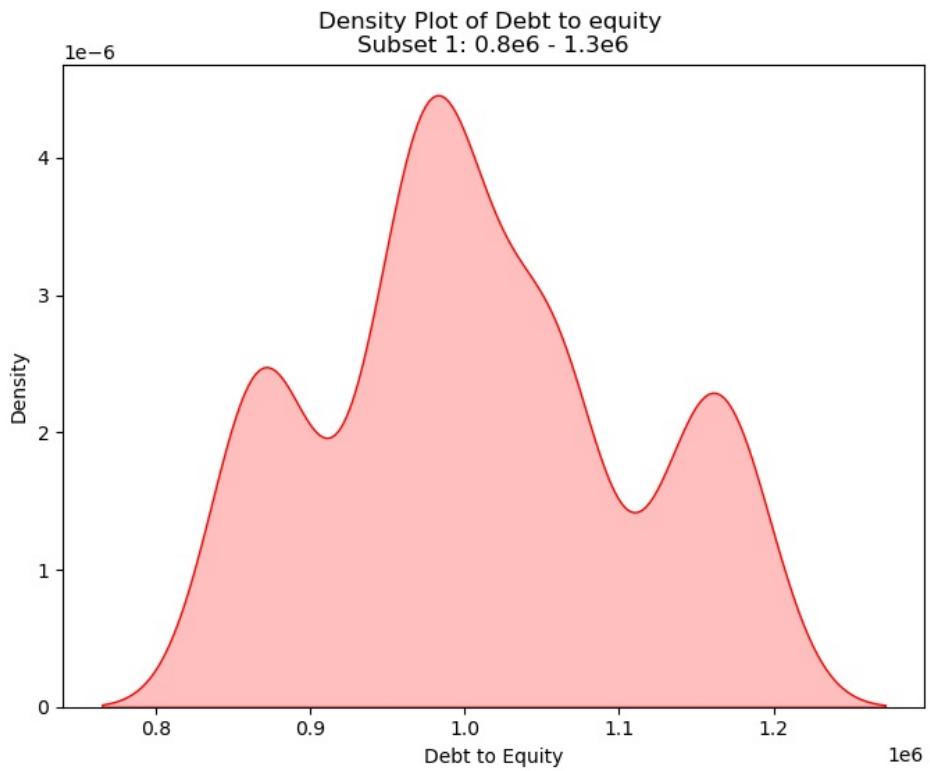
]

for i, group in enumerate(subgroups_e, start=1):
    subgroup_filtered_eq = data_cleaned[(data_cleaned['Debt to equity'] >= group['start']) &
                                         (data_cleaned['Debt to equity'] <= group['end'])]

    plt.figure(figsize=(8, 6))
    sns.kdeplot(subgroup_filtered_eq['Debt to equity'], fill=True, color='red', bw_adjust=0.5)
    plt.title(f'Density Plot of Debt to equity \nSubset {i}: {group["label"]}')
    plt.xlabel('Debt to Equity')
    plt.ylabel('Density')

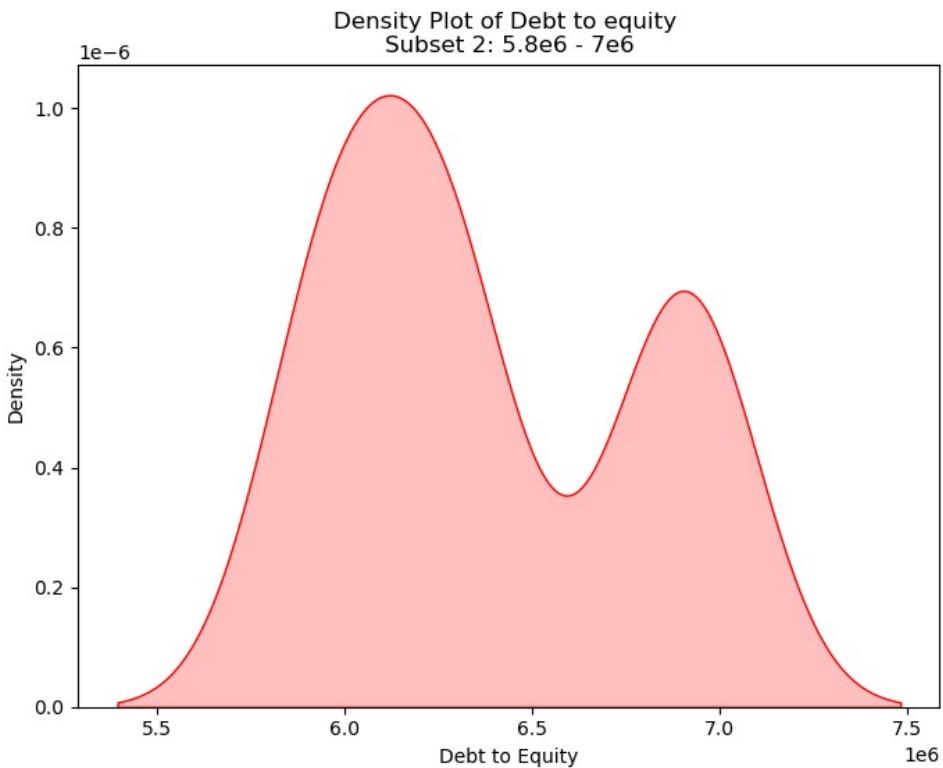
    plt.show()

print(f"\nAll rows of the subset {i} with 'Debt to equity' between {group['start']} and {group['end']}:")
display(subgroup_filtered_eq)
```



All rows of the subset 1 with 'Debt to equity' between 800000.0 and 1300000.0:

		Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	F As
442	G. P. MARKETING LIMITED	Active	Private limited Company	1994-11-01	46711 - Wholesale of petroleum and petroleum p...	46719 - Wholesale of other fuels and related p...	47110 - Retail sale in non-specialised stores ...	NaN	2021	Not Available	<NA>	3738	
443	G. P. MARKETING LIMITED	Active	Private limited Company	1994-11-01	46711 - Wholesale of petroleum and petroleum p...	46719 - Wholesale of other fuels and related p...	47110 - Retail sale in non-specialised stores ...	NaN	2020	Not Available	<NA>	3564	
444	G. P. MARKETING LIMITED	Active	Private limited Company	1994-11-01	46711 - Wholesale of petroleum and petroleum p...	46719 - Wholesale of other fuels and related p...	47110 - Retail sale in non-specialised stores ...	NaN	2019	Not Available	<NA>	2421	
514	GREENS OF LISBURN LIMITED	Active	Private limited Company	2017-03-30	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	329704	12%	
515	GREENS OF LISBURN LIMITED	Active	Private limited Company	2017-03-30	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	191601	30%	
516	GREENS OF LISBURN LIMITED	Active	Private limited Company	2017-03-30	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	34487	14%	
762	MC BRIDE RETAIL IRVINESTOWN LIMITED	Active	Private limited Company	2018-10-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	2237868	47%	
763	MC BRIDE RETAIL IRVINESTOWN LIMITED	Active	Private limited Company	2018-10-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	2300000	18%	
819	MCDADE RETAIL LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	3754432	99%	
856	MILESTONE (RATHFRILAND) LIMITED	Active	Private limited Company	2003-08-21	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	5894209	284%	



All rows of the subset 2 with 'Debt to equity' between 5800000.0 and 7000000.0:

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets	N Fixe Asse
915	NEWELL STORES (COALISLAND) LIMITED	Active	Private limited Company	2009-09-07	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN	NaN	2021	Not Available	7310123	12986!
916	NEWELL STORES (COALISLAND) LIMITED	Active	Private limited Company	2009-09-07	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN	NaN	2020	Not Available	6570164	13328!
917	NEWELL STORES (COALISLAND) LIMITED	Active	Private limited Company	2009-09-07	47110 - Retail sale in non-specialised stores ...	47300 - Retail sale of automotive fuel in spec...	NaN	NaN	2019	Not Available	6631109	12207!

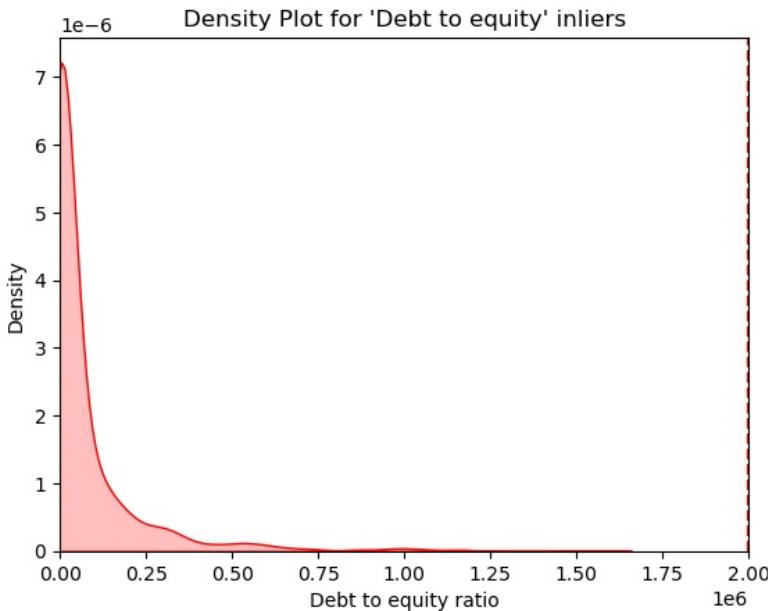
Looking at the subset between 5.8M and 7M, it is one company, Newell Stores (Coalisland). They are a large supermarket in Dungannon (https://www.facebook.com/NewellStores/?locale=en_GB). The substantial amounts listed under fixed assets, which range from approximately 6M+ to 7.3M+, indicate significant investments that typically require substantial equity support. Additionally, the current assets, which range from around 599782 to 1800000, suggest that the company is in a strong position for covering their liabilities, further implying that an equity of "1" is unrealistic. The large creditor amounts, ranging from approximately 5.9M to 6.9M, also point to the need for a much larger equity base to support these responsibilities. The low equity value has led to an increased debt to equity ratio, essentially duplicating the creditors value as it is simply being divided by 1, as we have seen in previous companies. In reality, if the equity were accurately reported, the debt to equity ratio would likely be much more reasonable and reflective of Newell Stores' financial performance, as of the rest of the dataset. Hence, the result is down to misrepresentation of the equity value, seemingly input due to a required field.

In the 800,000 to 1,300,000 subgroup, a similar pattern emerges, with equity values being listed as either '1' or '2', which appear to be placeholders rather than accurate figures. Given that these are large corporations with substantial fixed assets and creditor amounts, their actual equity would likely be much higher. This doesn't suggest better business decisions or different operational strategies but rather a misrepresentation of the financial data.

I will now examine the inliers within this group more closely to identify any patterns that may have been overlooked in the dataset. This analysis will help me uncover additional insights and inform potential directions for further work.

```
In [867]: inlier_subset = data_cleaned[(data_cleaned['Debt to equity'] >= 0) &
                                    (data_cleaned['Debt to equity'] <= 2e6)]

# Plot the density of the inliers
ax = sns.kdeplot(inlier_subset['Debt to equity'], fill=True, color="red")
ax.set_xlim(0, 2e6)
ax.axvline(x=0, color='r', linestyle='--')
ax.axvline(x=2e6, color='r', linestyle='--')
plt.title("Density Plot for 'Debt to equity' inliers")
plt.xlabel('Debt to equity ratio')
plt.ylabel('Density')
plt.show()
```



The plot does not reveal any distinct subgroups of inliers. Instead, it shows a single, highly concentrated group of data points near the low end of the 'Debt to equity' ratio spectrum. The sharp peak near zero and the rapid decline suggest that having a low 'Debt to equity' ratio is typical or normal within this dataset, while higher values are less common and taper off quickly. It is safe to conclude that the data points are "bundled" near zero, making this the norm for the dataset. There doesn't appear to be a need to identify or separate any further inlier subgroups based on the current plot.

4.5: Conclusions & Further Work

1. Conclusions

Outlier detection is a critical insight derived from the density plots. Companies that significantly deviate from the norm in terms of financial metrics, such as debt levels or asset values, should be closely monitored. These outliers may represent either potential risks or unique opportunities, however in this current dataset it was seemingly misrepresentation of data. Implementing early warning systems to monitor these outliers can help identify them early, preventing efforts to replicate or optimise strategies based on inaccurate data, which could lead to negative outcomes.

In terms of potential causal relationships, the data suggested the an important relationship to explore is the correlation between the number of employees a company has and its overall financial performance. The analysis could investigate whether larger companies, in terms of employee numbers, tend to be more financially stable or profitable, or if there is an optimal company size for specific industries. Understanding this relationship could help companies optimise their workforce for better financial outcomes, rather than just simply viewing solely finance metrics.

The creditors density plot shows that most companies in the dataset have low levels of debt. For a large supermarket like my expert's, this highlights the importance of careful debt management. While borrowing can be useful for growth and investment, high levels of debt increase financial risk, especially if market conditions change or revenues decline. To maintain financial stability, it's crucial to balance debt with the ability to generate cash flow. By keeping debt at manageable levels, the supermarket can invest in growth while minimising the risk of financial strain. This approach ensures long term resilience in a competitive market, especially given the ongoing need for costly technology upgrades that supermarkets are adopting. These investments are often necessary but expensive, and may require taking on loans to finance them. Therefore, it's essential to make calculated decisions about borrowing, ensuring that debt remains manageable while supporting growth and innovation.

The density plot for the debt to equity ratio shows that most companies maintain a relatively low level of debt compared to their equity. For a large supermarket like my expert's, this indicates the importance of keeping debt in check relative to equity. Maintaining a low debt

to equity ratio can enhance financial stability, making it easier to manage requirements and also survive in a fluctuating market. Supermarkets must be cautious not to overextend themselves, ensuring that their equity base is strong enough to sustain any debt taken on.

For a large supermarket like my expert's, the level of equity plays a crucial role in its ability to expand, whether that means opening new store locations, investing in cutting edge technologies or entering new markets. With higher equity levels, a supermarket is better equipped to undertake such expansion initiatives without heavily depending on debt. Therefore, another potential causal relationship to explore is how these equity levels directly impact the success of these expansion efforts, helping the supermarket identify the optimal financial strategies that best support sustainable growth.

1. Further Work

One key area for further work is to explore the relationships between different financial metrics in more detail. For example, investigating how debt levels impact overall financial stability or growth could be enhanced by using more advanced statistical techniques, such as time series analysis. This does not necessarily apply to univariate analysis, but it would be useful and data linking from another dataset or reformatting this dataset to include columns relating to years, rather than simply a year row, would be good. This would help better understand how borrowing affects long term success.

Another valuable next step would be to develop predictive models that can forecast financial outcomes based on the current data. By using machine learning or other forecasting methods, it's possible to predict how changes in factors like debt or equity levels might impact future performance. These models could help decision makers anticipate financial risks and opportunities, allowing for more proactive management.

Additionally, looking beyond just financial data could provide a more complete picture. Including factors such as customer satisfaction, employee performance, or supply chain efficiency in the analysis via data linking might uncover new areas for improvement. Understanding how these non financial elements interact with financial outcomes could lead to better and more comprehensive decision making.

Expanding the dataset to include more companies and a longer period of time, especially larger companies in the industry that directly reflect the market leaders, even from different countries, could also enhance the analysis. Including external data, such as economic trends or consumer behavior, might offer new insights that are not immediately obvious from the current dataset. This broader perspective could help the expert's supermarket's decision makers better understand their position in the market and identify new opportunities.

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Chapter 5

5.1 Causal Relationships

5.1.1 Causal Relationship Overview

In this section I will be delving into the use of pair plots to visualise all possible combinations of columns within the dataset using scatterplots or histograms, depending on datatypes. Although these visualisations may not uncover all causal relationships, they are highly effective for detecting clear and significant associations, and thus forming a good basis for analysis.

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns

In [54]: # reading in pdf from chapter 4
df = pd.read_csv('cleaned_data_chapter4.csv')

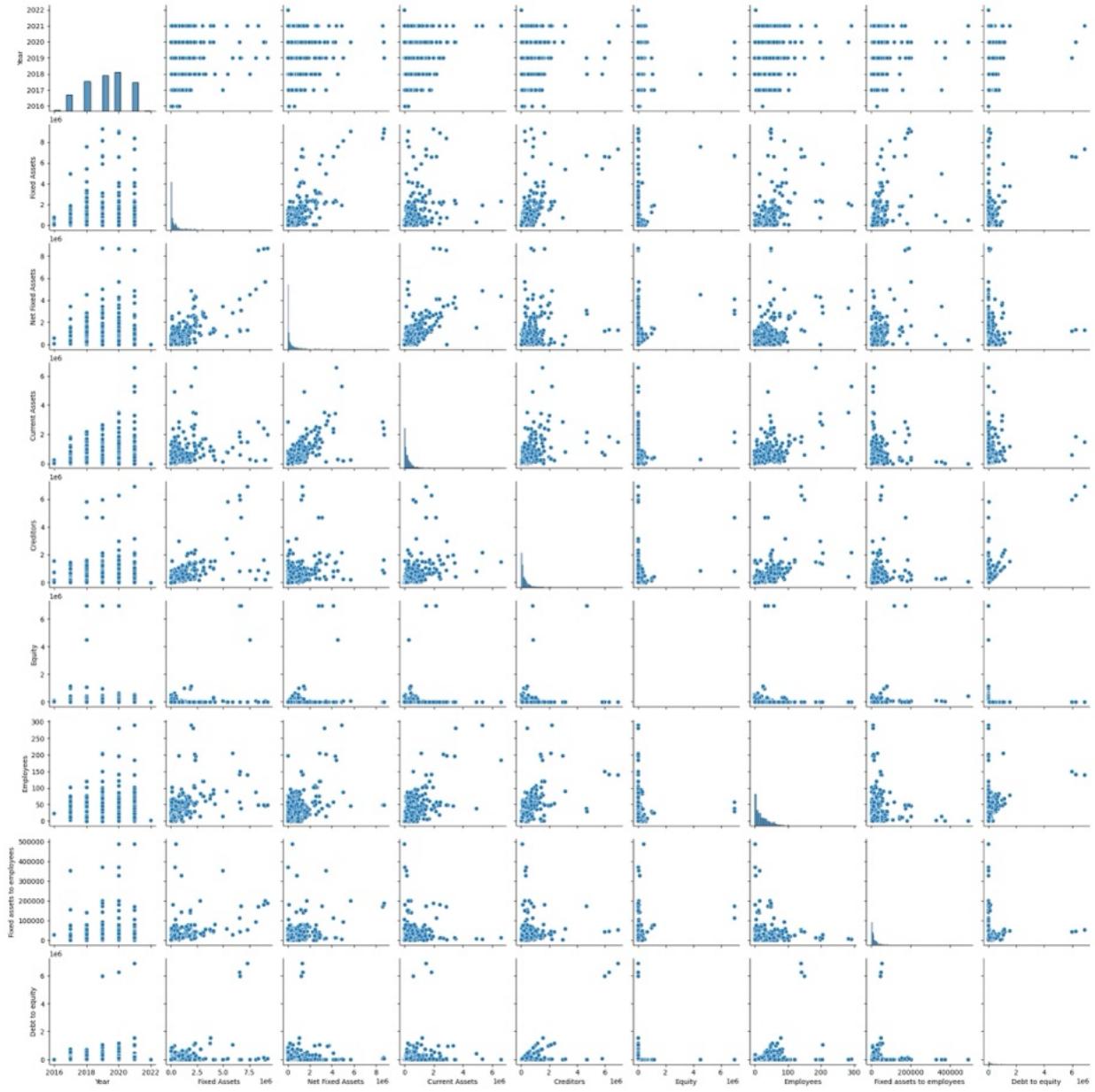
In [10]: df.columns

Out[10]: Index(['Company Name', 'Company status', 'Company type', 'Incorporated on',
       'Nature of business (SIC) 0', 'Nature of business (SIC) 1',
       'Nature of business (SIC) 2', 'Nature of business (SIC) 3', 'Year',
       'Currency', 'Fixed Assets', 'Net Fixed Assets', 'Current Assets',
       'Creditors', 'Equity', 'Employees', 'Fixed assets to employees',
       'Debt to equity'],
      dtype='object')

In [60]: import seaborn as sns
import matplotlib.pyplot as plt

# using seaborn pair plot func
sns.pairplot(df)

plt.show()
```



Given the absence of distinct upward or downward sloping lines in the pair plot, it is evident that there are no simple and strong causal relationships apparent in the data. This suggests that any potential causal links are likely more complex and not immediately visible through basic visualization. Consequently, further investigation is required to uncover any underlying factors that may be influencing these variables at a more detailed and nuanced level.

However, from the generated pair plot, I can identify visualisations that need a little more explanation.

First and foremost, in the pair plots involving the 'Year' variable, distinct horizontal bands are evident across several scatterplots. These bands represent clusters of data points where the values of other financial variables, such as 'Fixed Assets', 'Net Fixed Assets' and 'Equity' tend to fall within specific ranges for each given year. This pattern suggests that, during particular years, certain financial metrics stabilise or are more common within those ranges, leading to these visible lines. However, despite the presence of these bands, there is no clear directional trend, either upward or downward, across the scatterplots. This lack of a consistent slope indicates that the variables may not be causally related. Instead, these bands likely reflect the typical financial characteristics of companies in specific years without directly influencing other variables in a predictable manner. The financial metrics for each year may exhibit particular characteristics, but these do not appear to cause or correspond to changes in other financial variables. Also, it is also possible that these consistent bands across years might point to something else. For instance, the way the data was collected could mean that certain years have more entries than others, which might create these visible clusters. If the dataset has more companies reporting financial data in specific years, it could make those years look more stable or influential than they actually are.

When examining the scatterplot between 'Employees' and 'Fixed Assets' it is evident that there is no strong or consistent pattern linking the two variables. Some might expect a company with a higher number of employees to have more fixed assets, such as equipment, to support their operations. However, the data does not align with this assumption. The scatterplot shows a wide dispersion of points, indicating that companies with similar numbers of employees can have vastly different levels of fixed assets. For example, some companies with relatively few employees possess substantial fixed assets.

In the visualisations involving 'Creditors' and the 'Debt to equity' it is important to note that there is naturally some structure in the data. This is because the 'Debt to equity' ratio is calculated by dividing creditors by equity, leading to the observed patterns. Similarly, for visualisations involving 'Fixed Assets' and the 'Fixed assets to employees' ratio, the relationship is driven by the fact that the ratio is calculated by dividing fixed assets by the average number of employees column, creating a direct link between the variables, which explains why these patterns appear in the pair plot. Because these relationships are a result of the formula used to calculate the ratios, they do not require as much detailed analysis beyond acknowledging their inherent connection.

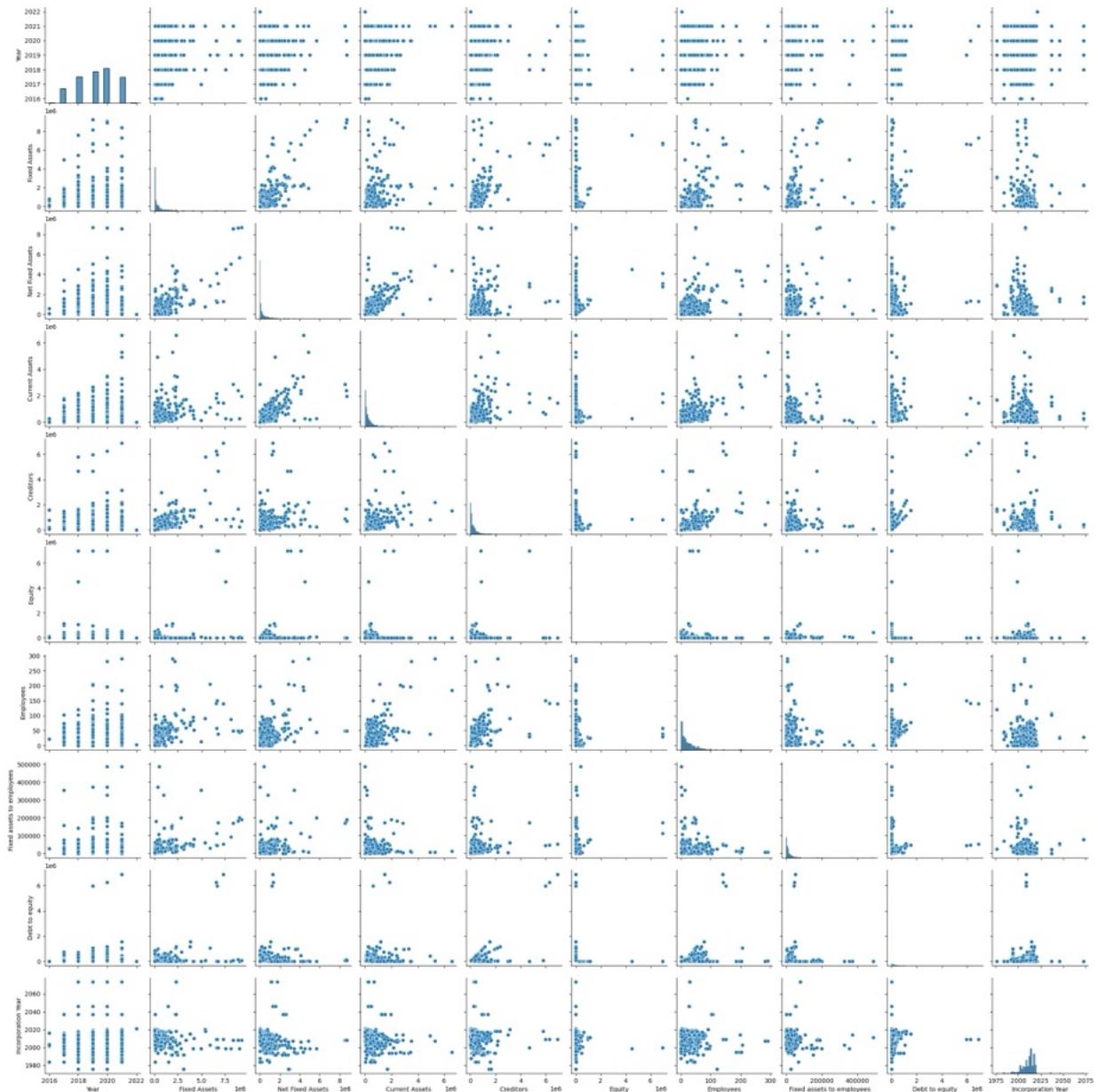
It is also important to note that pair plots can only display numerical data, whether it be categorical etc. As a result, variables such as company names and SIC codes are not included in the visualisation. However, I noticed that the 'Incorporated on' column was absent from the analysis. To address this, I have decided to write some code with the help of ChatGPT to include this column and reanalyse the data accordingly.

```
In [97]: # Convert 'Incorporated on' to datetime format
df['Incorporated on'] = pd.to_datetime(df['Incorporated on'], errors='coerce')

# Extract the year from 'Incorporated on' and add it to the dataset of further analysis
df['Incorporation Year'] = df['Incorporated on'].dt.year

# Select only the numerical columns, including the new 'Incorporation Year' variable
numerical_data = df.select_dtypes(include=['float64', 'int64'])
numerical_data['Incorporation Year'] = df['Incorporation Year']

# Create the pair plot with the new variable included
pair_plot_with_year = sns.pairplot(numerical_data)
plt.show()
```



As displayed, the addition of the 'Incorporation Year' (derivative from the Incorporated on column) did not reveal any strong causal

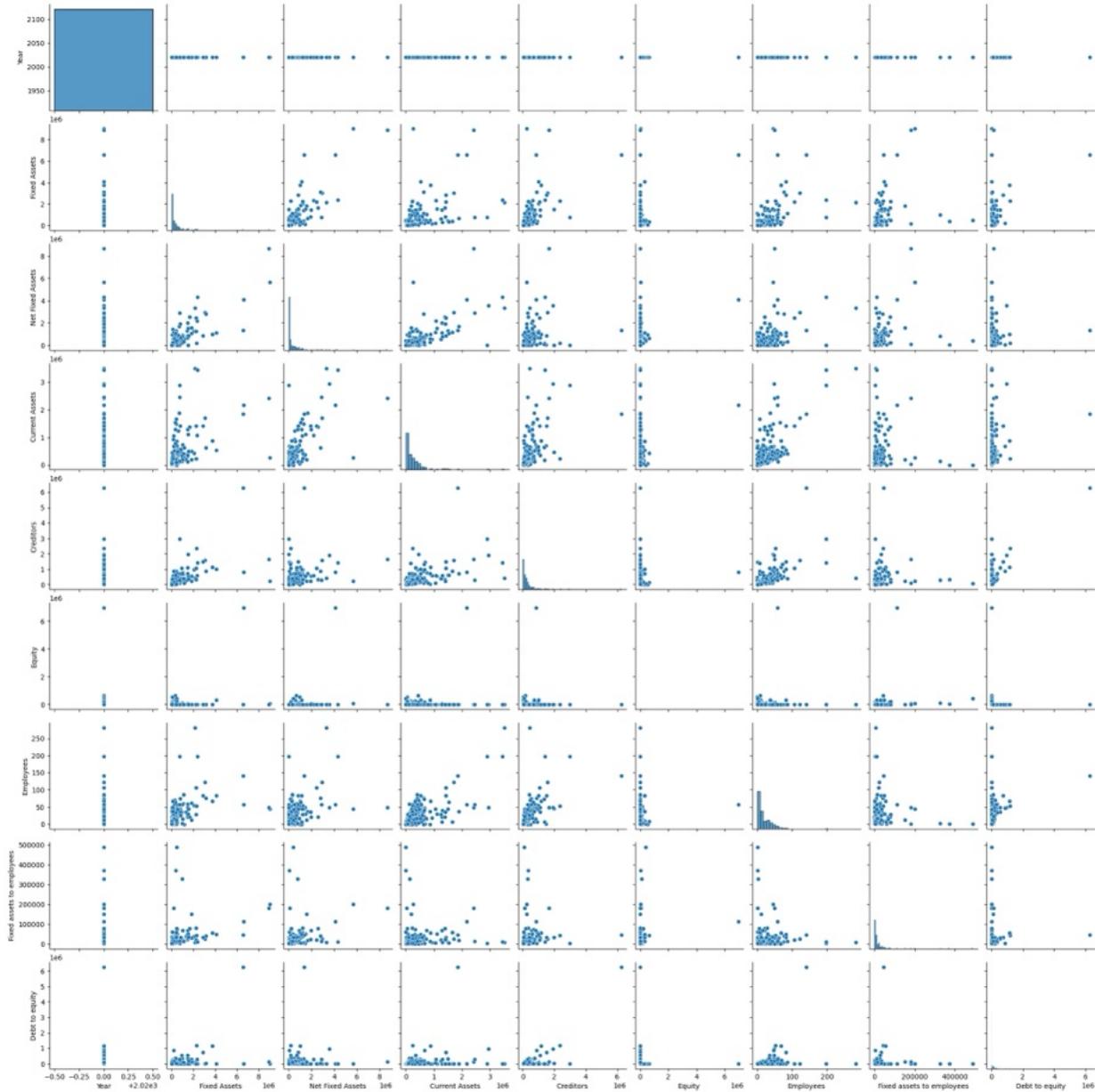
relationships, but it did provide more context about potential areas to explore, considering the bundling of data points.

To explore these relationships further, I will focus on pair plotting data from the year 2020 only. As identified in Chapter 4, 2020 is the year with the most company filings in the dataset. By isolating this year, I can determine if the variation across different years has an impact on these observed relationships, or if the patterns hold true.

```
In [62]: # filter data to only contain rows that have 2020 in the year column
filtered_df = df[df['Year'] == 2020]

# pair plotting this filtered dataframe
sns.pairplot(filtered_df)

plt.show()
```



The pair plot above indicates that focusing on the year 2020 does not significantly change the overall patterns observed in the data. The plots remain consistent with what was seen across all years, but with fewer data points and fewer bars in the year related charts since only 2020 is being considered. Additionally, some outliers are reduced or less pronounced due to the smaller dataset being visualised, which naturally cuts down on the extremes that might have been more visible when all years were included. This suggests that the relationships and patterns observed are relatively stable, regardless of the year, but focusing on a single year like 2020 provides a cleaner view with fewer outliers.

Further investigation into more 'underlying' causes will be discussed further next.

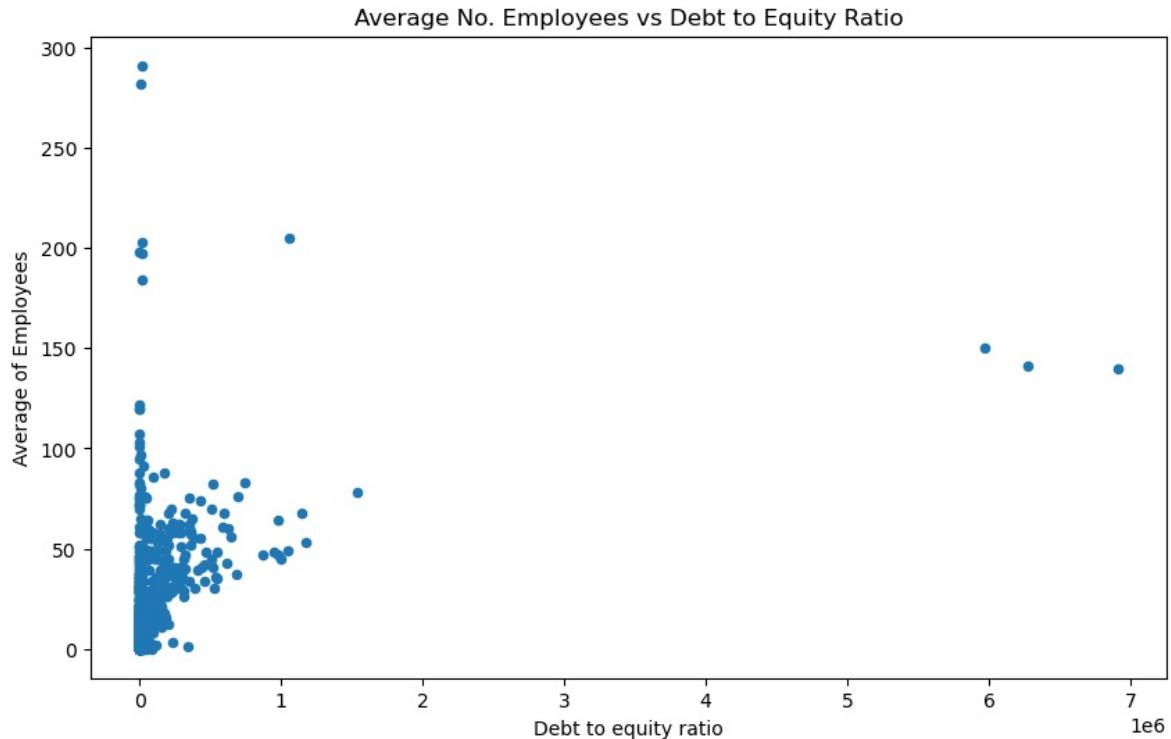
5.1.2 Probability Based Analysis

Before diving into the analysis of causal relationships, it is crucial to consider the primary objective of this report, i.e. to create visualisations

that highlight potential causes or outcomes that the decision maker at my expert's supermarket would find useful in relation to their care values or issues. In earlier stages of the project, I identified financial performance as one of the core values held by the decision maker. Fortunately, the dataset we are working with includes relevant financial performance metrics. With this in mind, I can now explore potential causal relationships based on the pair plot, focusing on those that could directly impact financial performance.

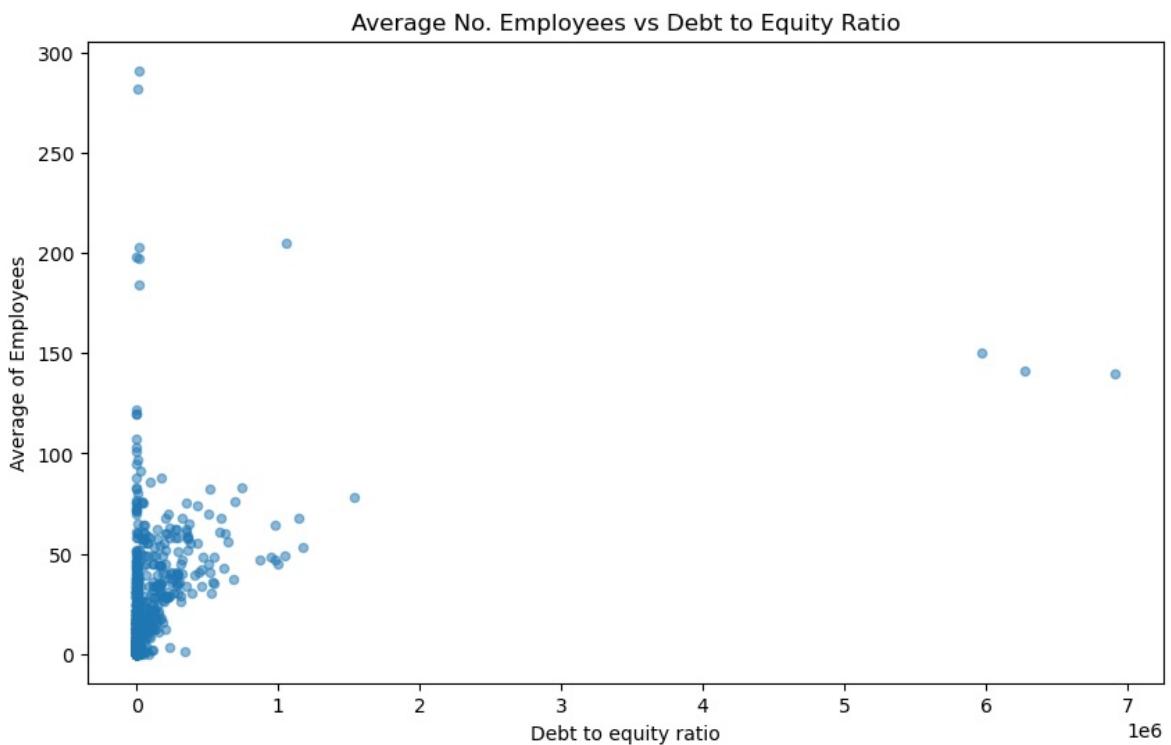
Firstly, I will investigate Debt to equity ratio and Employees (i.e. average employee's in the company). Investigating the relationship between the Debt to equity ratio and the average number of employees is important for understanding how the supermarket's financial decisions impact its operations. The Debt to equity ratio shows how much the company relies on borrowed money compared to its own funds. A higher ratio means the company is taking on more debt, which can increase financial risk. By looking at how this ratio relates to the number of employees, the decision maker can see if hiring more staff is being supported by taking on more debt. If there's a strong connection, it might mean the company is borrowing money to grow its workforce, which could affect its financial stability. Understanding this link helps ensure that employee growth is financially sustainable. I will begin by making the scatter plot on a larger scale here.

```
In [125]: df.plot(x='Debt to equity', y='Employees', kind='scatter', figsize=(10, 6), title='Average No. Employees vs Debt to Equity Ratio')
Out[125]: <Axes: title={'center': 'Average No. Employees vs Debt to Equity Ratio', xlabel='Debt to equity ratio', ylabel='Average of Employees'>
```



Now, let's add an alpha value of 0.5 to see more clearly the overlapping points. Using 0.5 makes the points transparent, making areas with more data points seem darker.

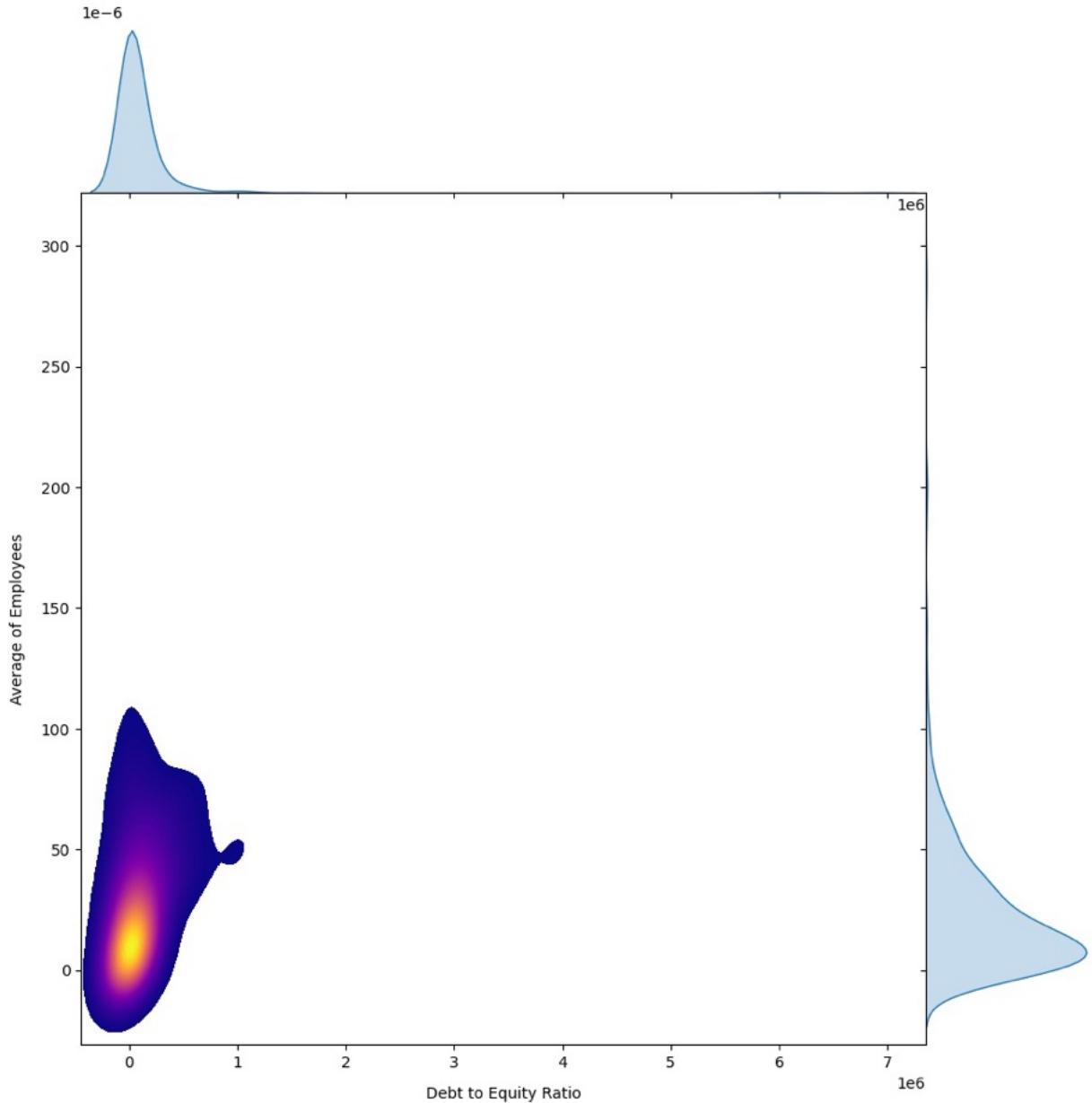
```
In [127]: df.plot(x='Debt to equity', y='Employees', kind='scatter', figsize=(10, 6), title='Average No. Employees vs Debt to Equity Ratio', alpha=0.5)
Out[127]: <Axes: title={'center': 'Average No. Employees vs Debt to Equity Ratio', xlabel='Debt to equity ratio', ylabel='Average of Employees'>
```



To better visualise the concentration of points, I will add a 2D density plot. This will make it easier to see the distribution and density of the data points, as the alpha value does not overly help with the issue of data point dsitribution and congregation.

```
In [222]: # using jointplot from seaborn to plot the 2d desnity plot
sns.jointplot(data=df, x='Debt to equity', y='Employees', kind='kde', fill=True, space=0, cmap='plasma', height=10)
plt.suptitle('2D Density Plot of Debt to Equity Ratio vs Average No. Employees', y=1.05, fontsize=16)
plt.ylabel('Average of Employees', labelpad=10)
plt.xlabel('Debt to Equity Ratio', labelpad=10)
plt.show()
```

2D Density Plot of Debt to Equity Ratio vs Average No. Employees



The 2D density plot shows that most companies have a low Debt to equity ratio, especially close to zero, and these companies also tend to have a smaller number of employees. This suggests that companies with fewer employees often choose to operate with minimal debt. As the Debt to Equity ratio increases, there are significantly fewer companies, indicating that not many companies maintain high levels of debt while having a large workforce. This pattern suggests that companies might prefer to manage their debt conservatively, especially as their employee numbers increase. However, lets home in on this figure, ignoring the outliers shown by the size of the plot.

```
In [252]: max_employees_threshold = 130
max_debt_to_equity_threshold = 0.7e6

filtered_df = df[(df['Employees'] <= max_employees_threshold) & (df['Debt to equity'] <= max_debt_to_equity_threshold)]

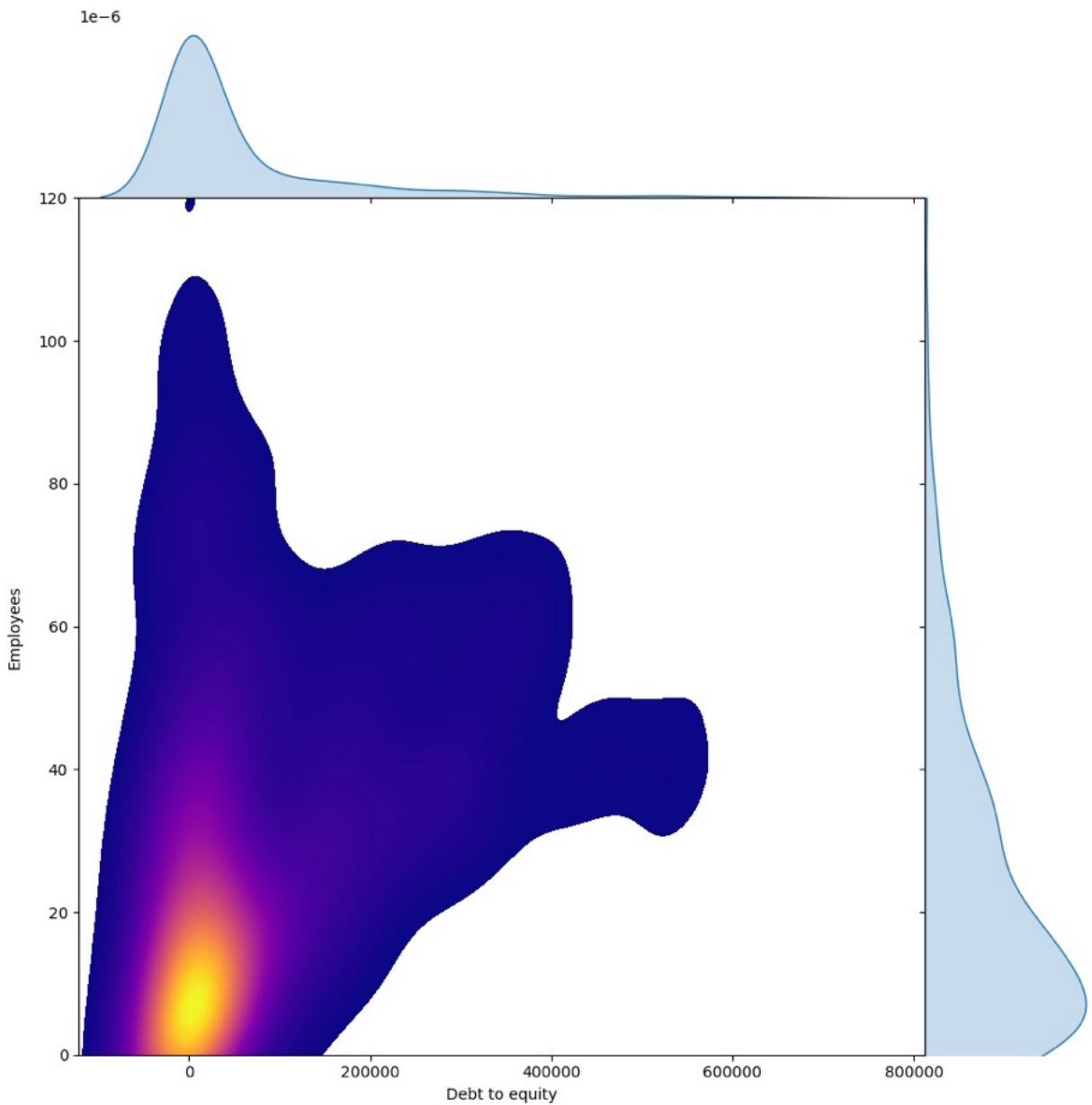
# Now plot the filtered data
sns.jointplot(
    data=filtered_df,
    x='Debt to equity',
    y='Employees',
    kind='kde',
    fill=True,
    space=0,
    cmap='plasma',
    height=10,
    ratio=5,
    levels=100,
    ylim=(0, 120)
)
```

```

plt.suptitle('2D Density Plot of Debt to Equity Ratio vs Average No. Employees (Outliers Removed)', y=1.05, fontweight='bold')
plt.ylabel('Employees', labelpad=10)
plt.show()

```

2D Density Plot of Debt to Equity Ratio vs Average No. Employees (Outliers Removed)



These plots highlight the concentration of data points and provide a clearer picture of how these variables interact. The density plots revealed that the majority of companies operate with a low Debt to equity ratio and a smaller number of employees. There is a noticeable drop off in the number of companies as the Debt to equity increases, accounting for outliers as seen in chapter 4. This trend suggests that most companies prefer to maintain lower debt levels, especially when managing a sizable employee base. The 2D density plots reinforce the hypothesis that companies with higher employee numbers may take on more debt, but such instances are less common, indicating a cautious approach towards considering debt for workforce expansion. However, I will now take 'vertical slices' of this 2d density plot at specific points along the Debt to equity ratio axis. These vertical slices allow me to examine the distribution of the average number of employees for specific ranges of Debt to equity ratio. By normalising the density values within each slice, I can better understand how the number of employees varies within companies that have similar financial gain.

```
In [258]: df = df[df['Employees'] >= 0]
```

```
In [262]: ratios_to_plot = [0.1, 0.2, 0.3]
```

```

plt.figure(figsize=(10, 6))
for specific_ratio in ratios_to_plot:
    selected_companies = df[df['Debt to equity'] == specific_ratio]
    if selected_companies['Employees'].var() > 0:
        sns.kdeplot(data=selected_companies['Employees'], fill=True, label=f'Debt to Equity Ratio = {specific_ratio}')
    else:
        sns.kdeplot(data=selected_companies['Employees'], label=f'Debt to Equity Ratio = {specific_ratio}')

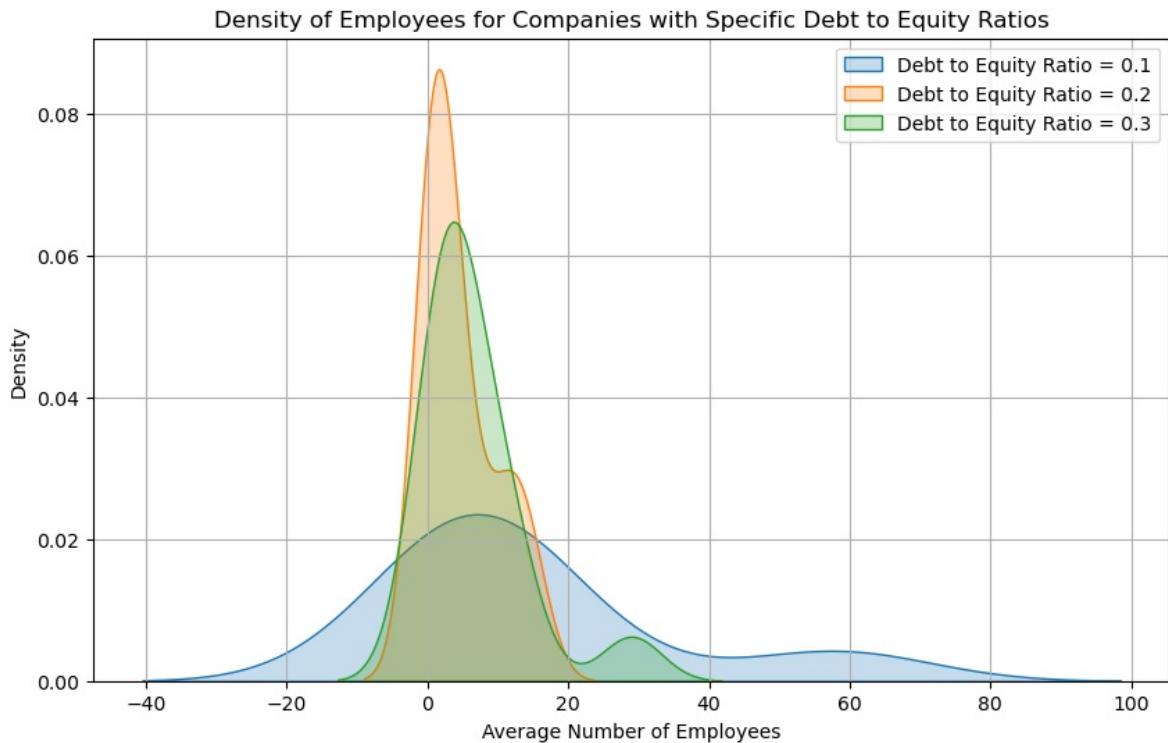
```

```

print(f'Skipping Debt to Equity Ratio = {specific_ratio} due to zero variance or insufficient data.')

plt.title('Density of Employees for Companies with Specific Debt to Equity Ratios')
plt.xlabel('Average Number of Employees')
plt.ylabel('Density')
plt.grid(True)
plt.legend()
plt.show()

```



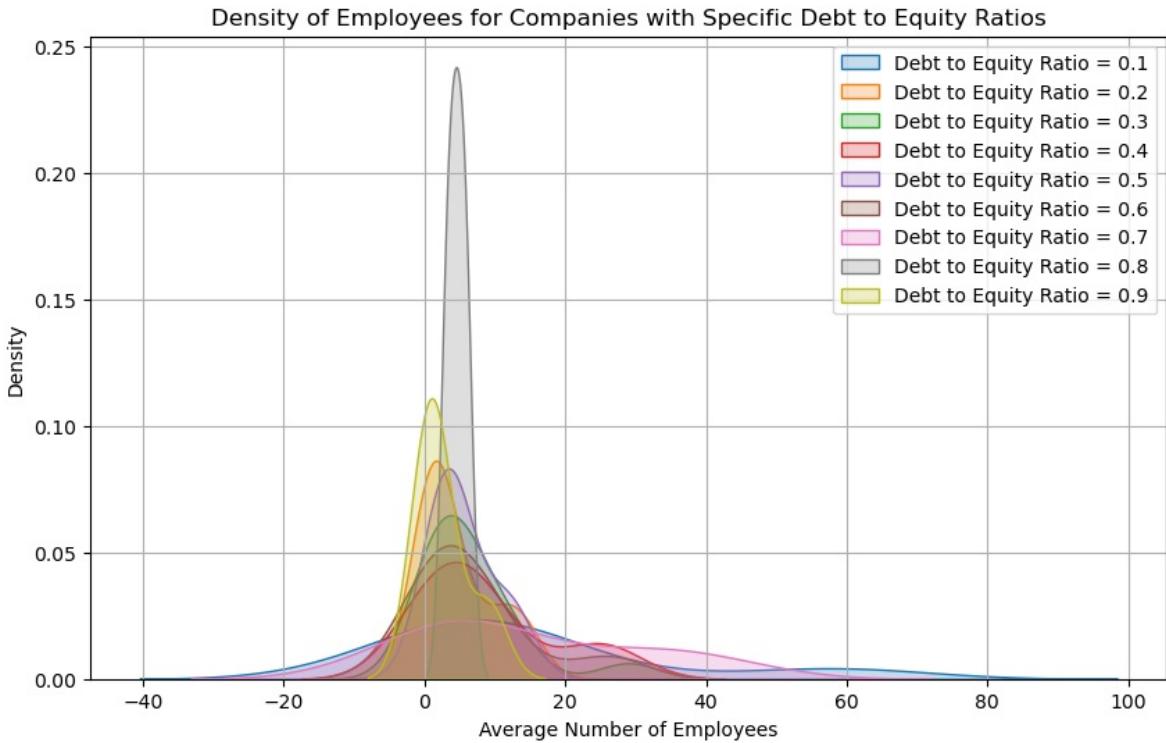
```

In [269]: ratios_to_plot = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]

plt.figure(figsize=(10, 6))
for specific_ratio in ratios_to_plot:
    selected_companies = df[df['Debt to equity'] == specific_ratio]
    if selected_companies['Employees'].var() > 0:
        sns.kdeplot(data=selected_companies['Employees'], fill=True, label=f'Debt to Equity Ratio = {specific_ratio}')
    else:
        print(f'Skipping Debt to Equity Ratio = {specific_ratio} due to zero variance or insufficient data.')

plt.title('Density of Employees for Companies with Specific Debt to Equity Ratios')
plt.xlabel('Average Number of Employees')
plt.ylabel('Density')
plt.grid(True)
plt.legend()
plt.show()

```



My density plot was specifically designed to provide a detailed examination of the relationship between the variables by slicing the data at distinct debt levels, at first for 0.1, 0.2, 0.3, to see the different relationships at these density level slices. This slicing approach allowed for a granular view of how employee distributions vary across these specific financial gain points. It took multiple code attempts to ensure I was in the correct range for the points, pin pointing the different points. However, the more inclusive plot is the most important.

By taking these 'slices' at key point, the density plot revealed how companies distribute their workforce within these financial frameworks. The broader spread of the density curves for higher ratios suggests that companies with more employees are more likely to have varying debt to equity ratios. This could imply that as companies scale up, their financial strategies become more diverse, with some relying heavily on debt while others maintain lower leverage. The tight clustering of companies with ratios between 0.1 and 0.5 near low employee counts suggests a potential causal relationship that companies that are more conservative in managing debt may focus on maintaining smaller, more stable workforces. This could be due to a preference for internal financing over expansion through debt. As potential reasoning for this investigation, for companies aiming to grow quickly, examining how retail stores with higher debt ratios have successfully scaled their workforce could offer strategies for balancing growth with financial risk. The density plot suggests that as workforce size increases, debt levels become more variable, indicating the need for flexible financial management when scaling.

However, this plot very difficult to comprehend any more information and would benefit from a different kind of plot.

```
In [275]: df_filtered = df[(df['Debt to equity'] > 0) & (df['Debt to equity'] < 5)]
ratios_to_plot = np.linspace(0.1, 4.0, 20)

mean_employees = []
median_employees = []
for specific_ratio in ratios_to_plot:
    selected_companies = df_filtered[df_filtered['Debt to equity'].round(1) == round(specific_ratio, 1)]
    if not selected_companies.empty:
        mean_employees.append(selected_companies['Employees'].mean())
        median_employees.append(selected_companies['Employees'].median())
    else:
        mean_employees.append(np.nan)
        median_employees.append(np.nan)

ratios_to_plot = np.array(ratios_to_plot)
mean_employees = np.array(mean_employees)
median_employees = np.array(median_employees)

mask = ~np.isnan(mean_employees) & ~np.isnan(median_employees)

fig, (ax1, ax2) = plt.subplots(2, 1, figsize=(10, 12), sharex=True, gridspec_kw={'height_ratios': [3, 1]})

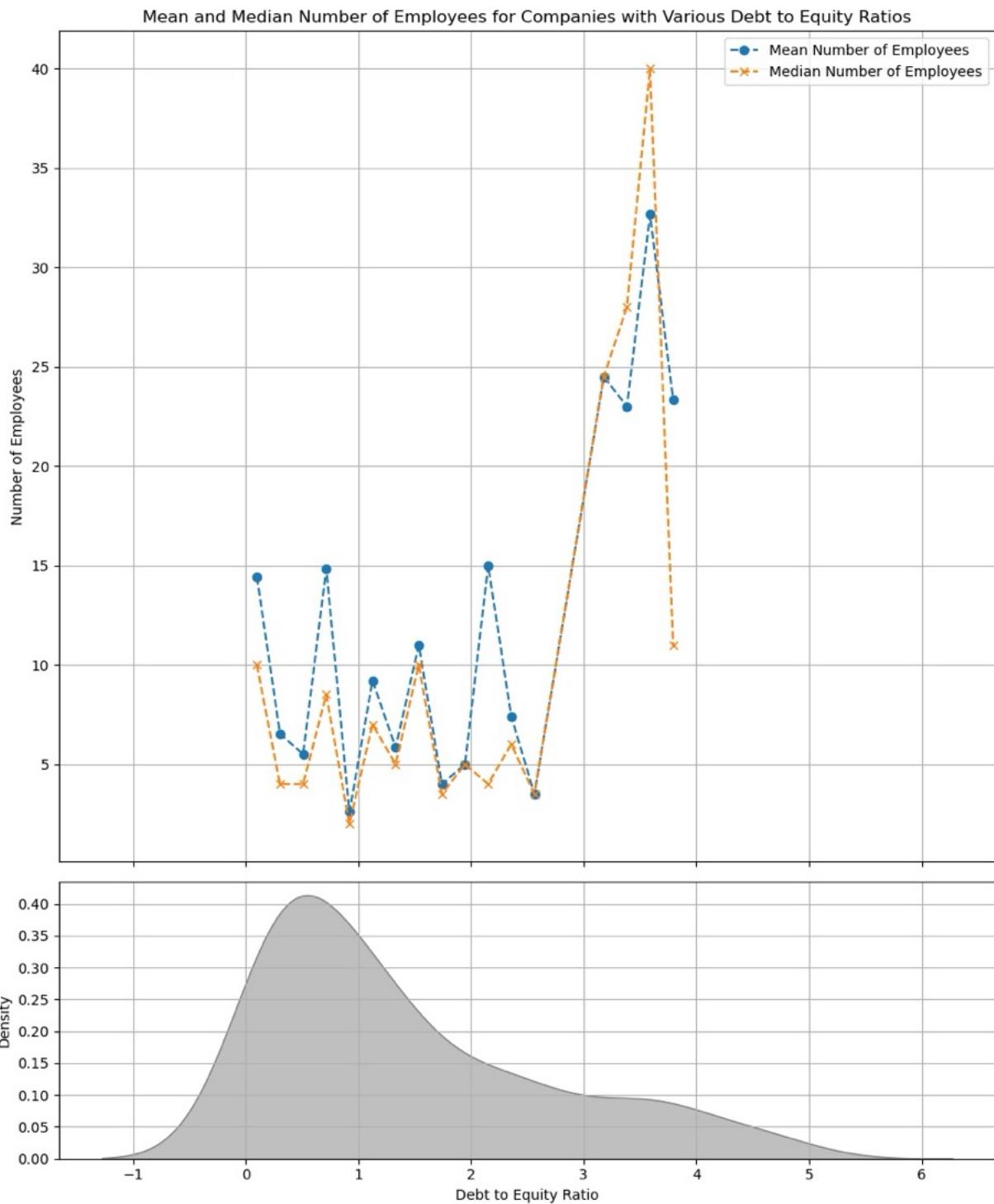
ax1.plot(ratios_to_plot[mask], mean_employees[mask], label='Mean Number of Employees', marker='o', linestyle='-.')
ax1.plot(ratios_to_plot[mask], median_employees[mask], label='Median Number of Employees', marker='x', linestyle='--')
ax1.set_ylabel('Number of Employees')
ax1.grid(True)
ax1.legend()
ax1.set_title('Mean and Median Number of Employees for Companies with Various Debt to Equity Ratios')
```

```

sns.kdeplot(data=df_filtered, x='Debt to equity', ax=ax2, fill=True, alpha=0.5, color='gray')
ax2.set_xlabel('Debt to Equity Ratio')
ax2.set_ylabel('Density')
ax2.grid(True)

plt.tight_layout()
plt.show()

```



The top plot, which displays the mean and median number of employees across various Debt to equity ratios, provides a valuable comparison between the median and mean workforce size within companies at different debt levels. At lower debt levels, the mean and median are closely aligned, indicating that companies with conservative debt practices tend to maintain a consistent and similar number of employees. However, as the Debt to Equity Ratio increases, especially beyond 2.0, a noticeable change between the mean and median appears. This change suggests that while some companies significantly expand their workforce with higher debt, others do not, leading to greater variability and skewness in the distribution of employee numbers.

Beneath this, the density plot of the Debt to equity shows that most companies cluster around a ratio of 1.0, with the frequency decreasing as the ratio increases. This distribution indicates that moderate debt levels are the norm, with fewer companies willing to take

on higher leverage. When viewed together with the top plot, it becomes clear that companies maintaining moderate debt levels generally have stable workforce sizes. However, those that opt for higher debt levels tend to experience a wider range of workforce sizes, reflecting differing strategies in increasing debt for potential growth.

Now, I will create a normalized 2D density plot, which displays the relationships independent of the x-axis distribution. The plot will zoom in on the densely populated regions and include contour lines for better clarity.

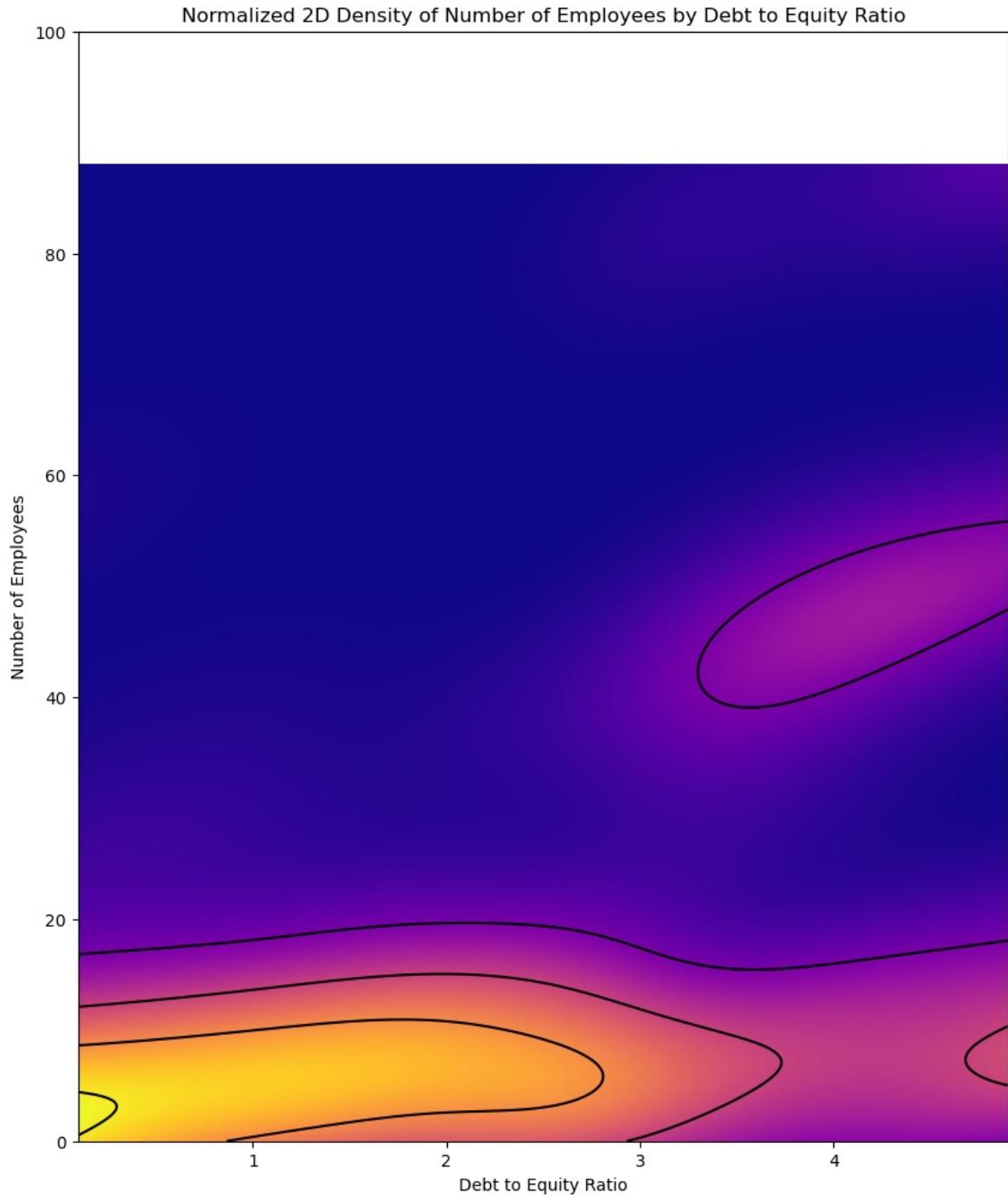
```
In [208]: data = df_filtered[['Debt to equity', 'Employees']].dropna().values

kde = gaussian_kde(data.T, bw_method='scott')

x = np.linspace(data[:, 0].min(), data[:, 0].max(), 1000)
y = np.linspace(data[:, 1].min(), data[:, 1].max(), 1000)
X, Y = np.meshgrid(x, y)
positions = np.vstack([X.ravel(), Y.ravel()])
Z = np.reshape(kde(positions).T, X.shape)

for i, x_val in enumerate(x):
    slice_density = Z[:, i]
    slice_density /= slice_density.sum()
    Z[:, i] = slice_density

# Plotting
plt.figure(figsize=(10, 12))
plt.pcolormesh(X, Y, Z, shading='auto', cmap='plasma')
plt.ylim(0, 100)
plt.xlabel('Debt to Equity Ratio')
plt.ylabel('Number of Employees')
plt.title('Normalized 2D Density of Number of Employees by Debt to Equity Ratio')
plt.contour(X, Y, Z, levels=5, colors='k')
plt.show()
```



Now I will create more plots, and overlay current findings.

```
In [361]: data = df_filtered[['Debt to equity', 'Employees']].dropna().values

kde = gaussian_kde(data.T, bw_method='scott')

x = np.linspace(data[:, 0].min(), data[:, 0].max(), 1000)
y = np.linspace(data[:, 1].min(), data[:, 1].max(), 1000)
X, Y = np.meshgrid(x, y)
positions = np.vstack([X.ravel(), Y.ravel()])
Z = np.reshape(kde(positions).T, X.shape)

for i, x_val in enumerate(x):
    slice_density = Z[:, i]
    slice_density /= slice_density.sum()
    Z[:, i] = slice_density

fig, axs = plt.subplots(2, 1, figsize=(12, 15), gridspec_kw={'height_ratios': [3, 1]})
```

```

c = axs[0].pcolormesh(X, Y, Z, shading='auto', cmap='plasma')
axs[0].contour(X, Y, Z, colors='k')
axs[0].set_xlabel('Debt to Equity Ratio', fontsize=0)
axs[0].set_ylabel('Number of Employees', fontsize=25)
axs[0].set_title('Normalised 2D Density of Number of Employees by Debt to Equity Ratio', fontsize=30)
axs[0].set_ylim(0, 100)

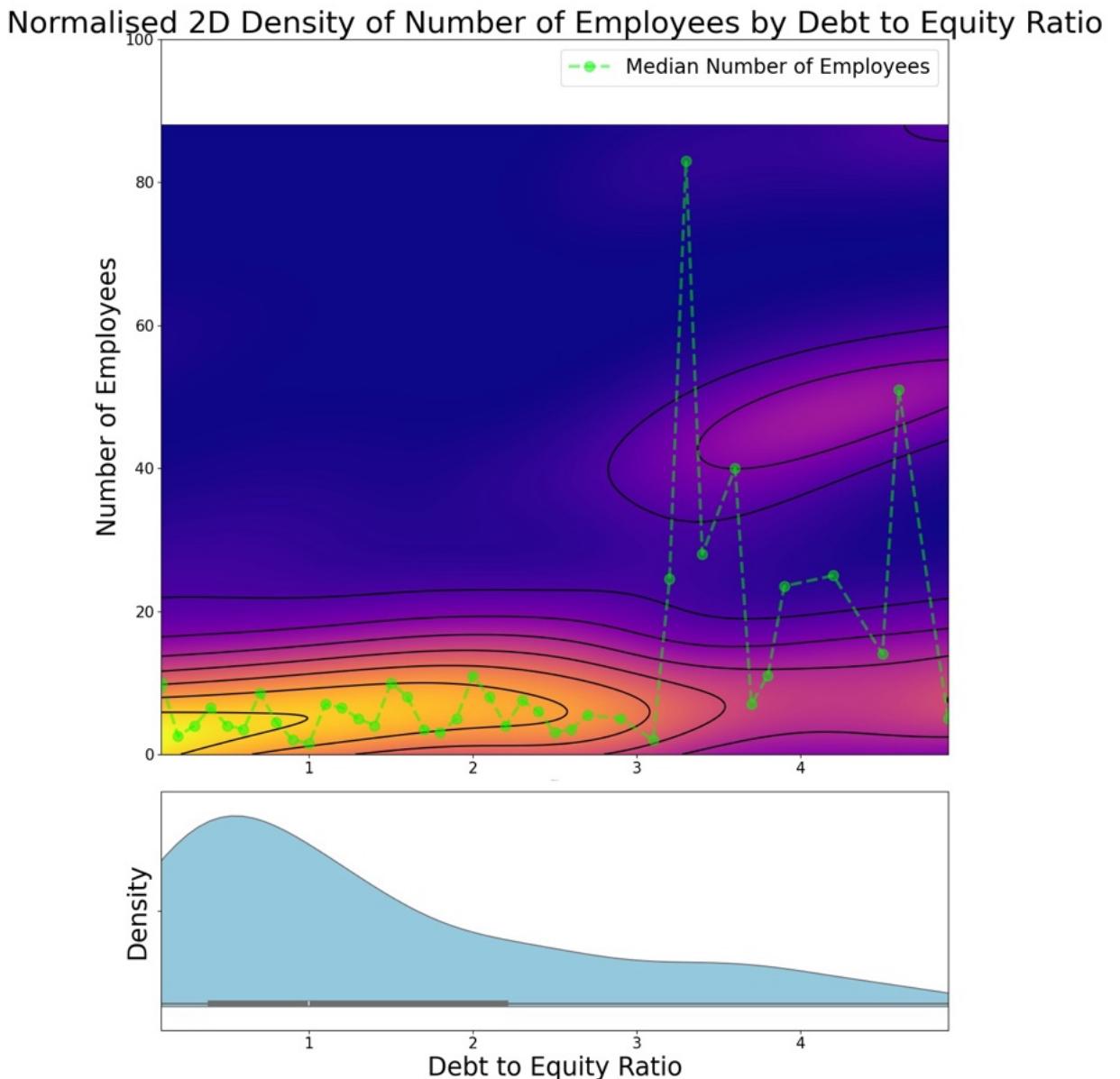
median_employees_per_ratio = df_filtered.groupby('Debt to equity')['Employees'].median()
median_employees_per_ratio.plot(kind='line', marker='o', color='lime', linestyle='--', linewidth=3, markersize=10)
axs[0].set_xlim(axs[0].get_xlim())
axs[0].legend(['Median Number of Employees'], fontsize=20)

sns.violinplot(x=df_filtered['Debt to equity'], color='skyblue', ax=axs[1], split=True)
axs[1].set_xlabel('Debt to Equity Ratio', fontsize=25)
axs[1].set_ylabel('Density', fontsize=25)
axs[1].set_xlim(axs[0].get_xlim())

for ax in axs:
    ax.tick_params(axis='both', which='major', labelsize=15)

plt.tight_layout()
plt.show()

```



The top plot is a normalized 2D density plot that shows the relationship between the debt to equity ratio (x-axis) and the number of employees (y-axis). The color intensity indicates the density of data points, with warmer colors (yellow, orange) representing areas of higher concentration. The contour lines further highlight regions of similar density. The plot reveals that companies with lower debt-to-equity ratios (around 1) are mostly clustered within lower employee counts (below 20), suggesting that smaller workforces are common among companies with more conservative financial leverage. The erratic green line represents the median number of employees at each debt to equity ratio, indicating some volatility in workforce sizes as debt levels increase, but the majority of the data remains concentrated at lower employee counts.

The bottom plot is a density plot (violin plot style) that displays the distribution of debt to equity ratios. The shape of the plot shows that the majority of companies have debt to equity ratios between 1 and 2, with the density tapering off as the ratio increases. This reinforces the finding from the top plot that most companies operate with moderate levels of debt relative to equity.

All in all, the visualisations suggest a potential causal relationship where the financial leverage (debt to equity ratio) of a company might influence its workforce size. Specifically, the density plot reveals that companies with low debt to equity ratios (around 1) are typically associated with smaller workforces (below 20 employees). This pattern implies that companies with lower financial leverage tend to maintain leaner teams. The likely causal explanation is that companies that are more conservative in taking on debt prioritise financial stability and therefore avoid rapid workforce expansion. These businesses may rely more on internal funding, leading to a streamlined operational model focused on sustainability rather than aggressive growth. As the debt to equity ratio increases beyond 2, the plot indicates more variability in employee count, as reflected by the fluctuations in the green median line and the spread of the density contours. This suggests that companies taking on higher debt might be doing so to support expansion, leading to fluctuating workforce sizes. The causal relationship here could be that companies willing to take on more financial risk do so to fuel rapid growth, which involves hiring more employees. However, the observed variability also indicates that some companies struggle to sustain large teams when leveraging higher debt, possibly due to financial strain or the inability to generate sufficient returns to cover the additional liabilities.

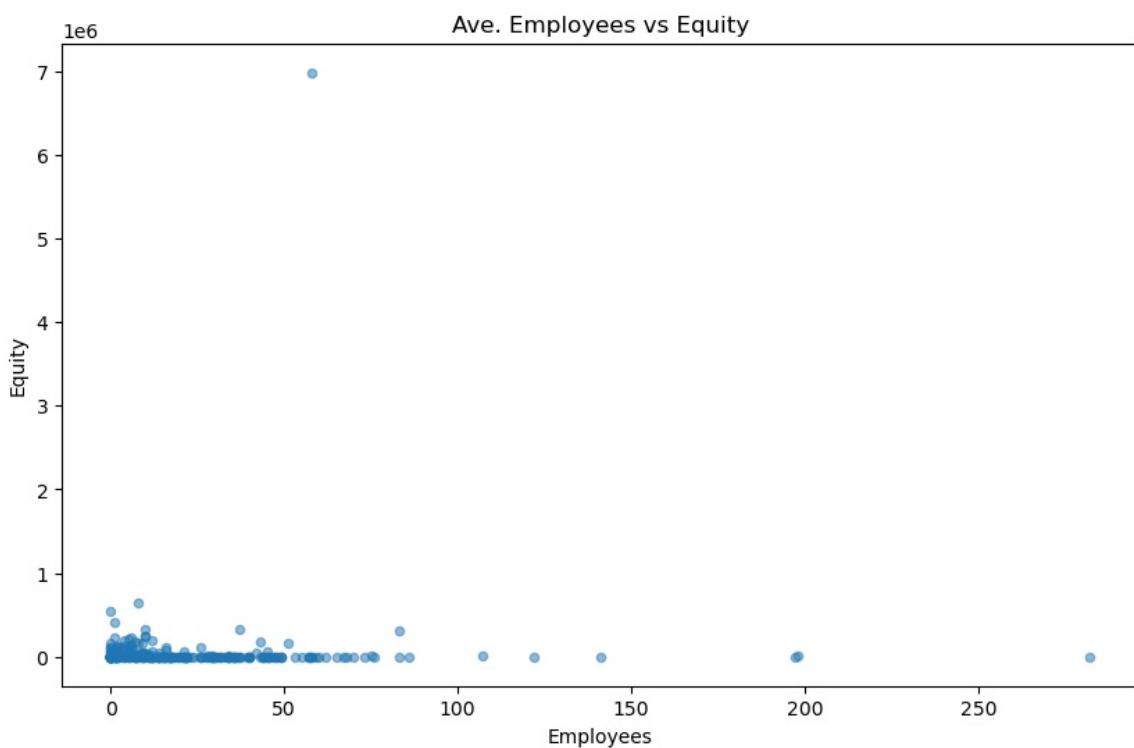
For a decision maker in a large supermarket such as my expert's, understanding this relationship is good. Supermarkets often operate with significant employee counts and face decisions regarding expansion and financial strategy. Knowing that increasing debt could lead to variability in workforce stability can help guide decisions about financing growth. If debt is used strategically, it can help fund workforce expansion, but excessive use could introduce risks that destabilise operations. By recognising these patterns, the decision maker can better balance financial performance with sustainable workforce management, ensuring the company grows without compromising stability.

There were some insights in the above visualisations concerning debt to equity and employee numbers, but I would still like to explore further causal relationships, as the insights were fairly weak and lacked relevance in ways. I will now commence by producing similar visualisations of that of the debt to equity vs employee ones above, but concerning Equity vs Employees. To ensure more accurate results, I will specifically focus on the year 2020. As discussed in Chapter 4, 2020 is the year with the most complete filings, and I believe analysing only this year will provide a sufficient number of data points while reducing the risk of skewed results that could occur if I include data from different years. By focusing solely on 2020, I can handle outliers more effectively and draw clearer conclusions.

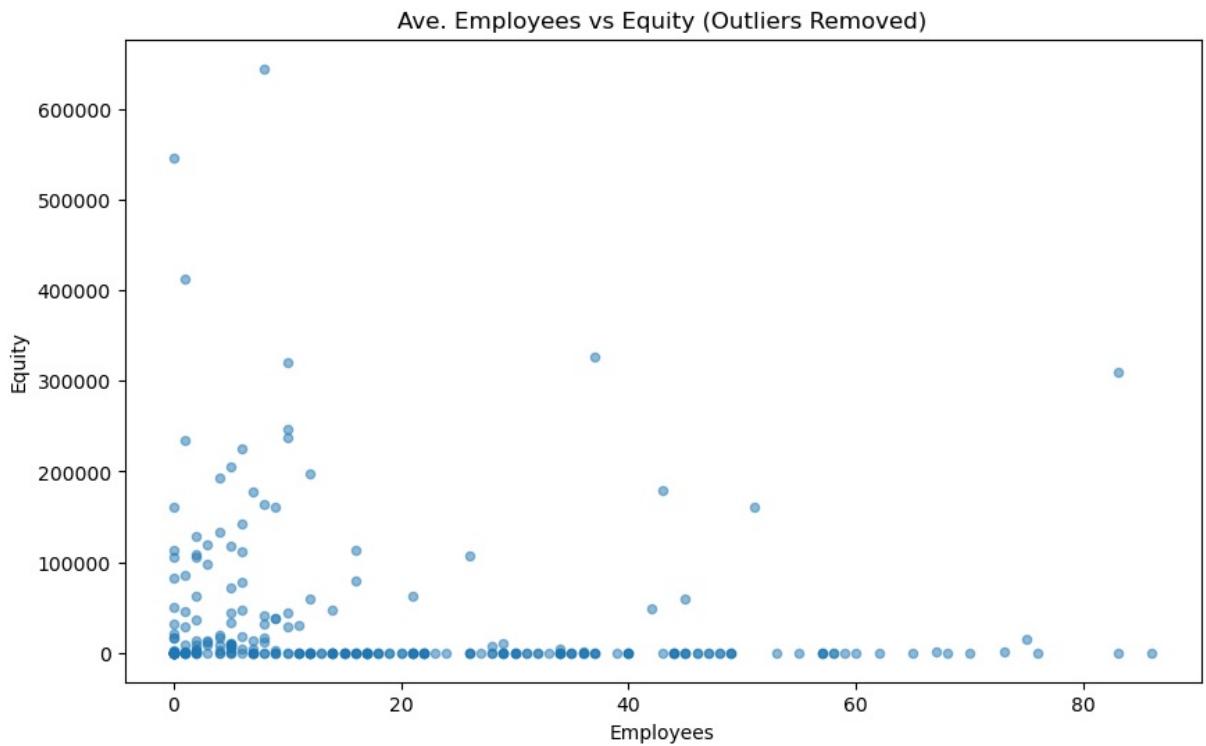
```
In [280]: df_2020 = df[df['Year'] == 2020]

df_2020.plot(
    x='Employees',
    y='Equity',
    kind='scatter',
    figsize=(10, 6),
    title='Ave. Employees vs Equity',
    xlabel='Employees',
    ylabel='Equity',
    alpha=0.5
)

Out[280]: <Axes: title={'center': 'Ave. Employees vs Equity'}, xlabel='Employees', ylabel='Equity'>
```

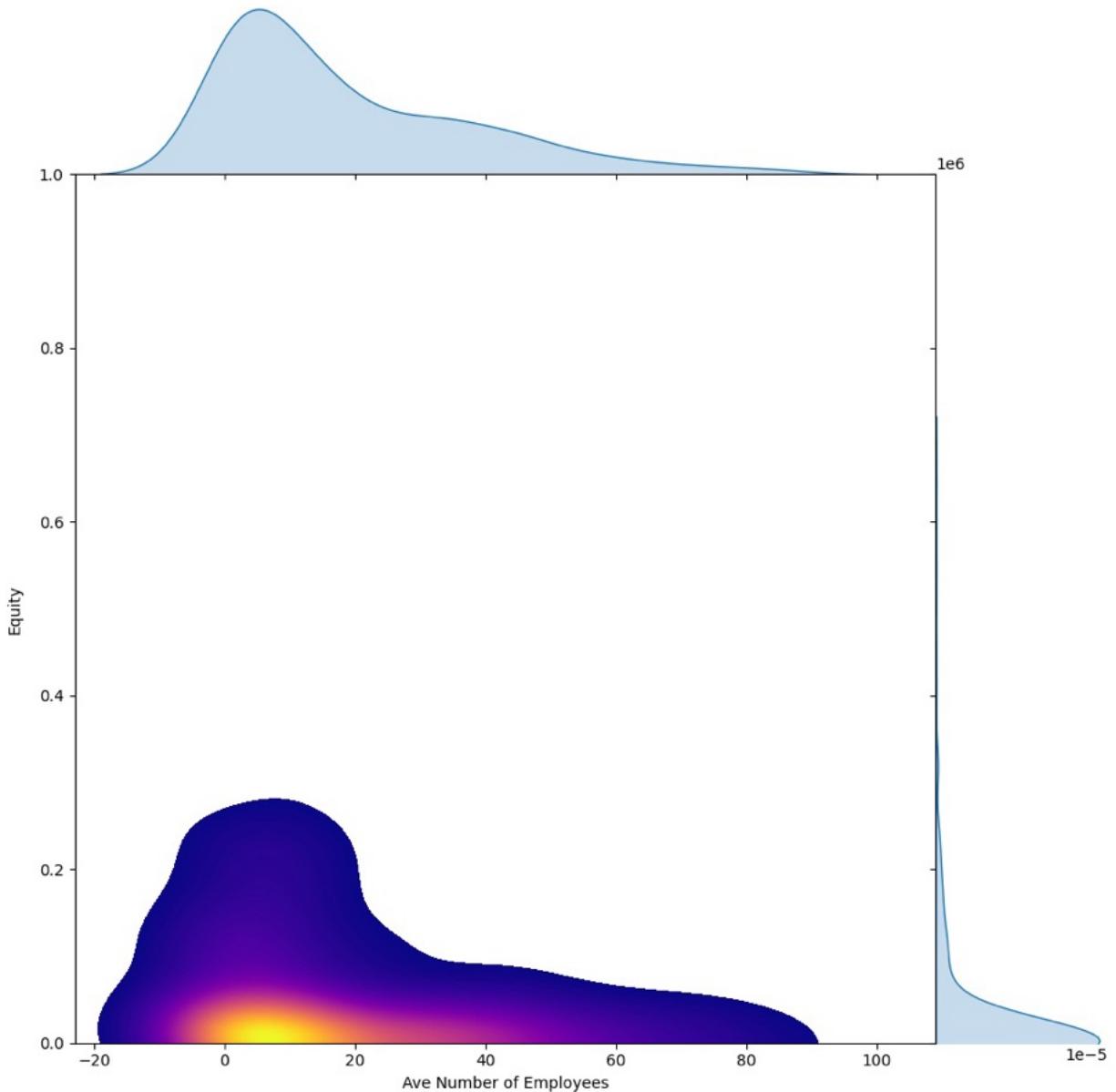


```
In [312]: filtered_df_2020 = df_2020[(df_2020['Equity'] <= 1e6) & (df_2020['Employees'] <= 100)]  
filtered_df_2020.plot(  
    x='Employees',  
    y='Equity',  
    kind='scatter',  
    figsize=(10, 6),  
    title='Ave. Employees vs Equity (Outliers Removed)',  
    xlabel='Employees',  
    ylabel='Equity',  
    alpha=0.5  
)  
plt.show()
```



```
In [310]: sns.jointplot(  
    data=filtered_df_2020,  
    x='Employees',  
    y='Equity',  
    kind='kde',  
    fill=True,  
    space=0,  
    cmap='plasma',  
    height=10,  
    ratio=5,  
    levels=100,  
    ylim=(0, 1e6)  
)  
  
plt.suptitle('2D Density Plot of Employees vs Equity(Outliers Removed)', y=1.05, fontsize=16)  
plt.ylabel('Equity', labelpad=10)  
plt.xlabel('Ave Number of Employees')  
  
plt.show()
```

2D Density Plot of Employees vs Equity(Outliers Removed)



The density plot shows that most companies have low equity and small workforces, which suggests that many businesses operate with limited financial resources and prefer to keep their teams small. This is common for smaller companies that focus on staying financially stable and avoid growing too quickly because of limited resources.

As equity increases, the spread of employee numbers widens slightly, but most companies still have relatively small workforces. This indicates that even when companies have more financial capital, they often prioritise being efficient rather than hiring large teams. For decision makers in a large supermarket, this suggests that it's more important to focus on efficient operations rather than just increasing staff as financial resources grow. For example, adding more security staff or checkout workers might improve safety and customer satisfaction, but it's essential to consider the balance. Expanding staff could provide some benefits, but decision makers need to weigh these against the potential downsides, such as increased costs or management challenges. Ultimately, it is about finding the right balance between operational efficiency and strategic growth without losing sight of what really drives financial success.

The lack of companies with both high equity and large workforces suggests that increasing financial resources does not automatically lead to hiring more employees. This highlights that growth needs to be strategic and careful, not just about expanding for the sake of it. For decision makers, the key takeaway is that workforce expansion should be balanced with smart financial management to ensure long term stability.

Now, I want to explore whether there are instances where companies have disproportionately high equity compared to others with different average employee numbers, which could indicate that having a certain number of employees may impact equity. I will investigate this by performing a "vertical slice" using equity bins, as pinpointing these variations was difficult.

```
In [322]: import matplotlib.pyplot as plt
```

```

import seaborn as sns

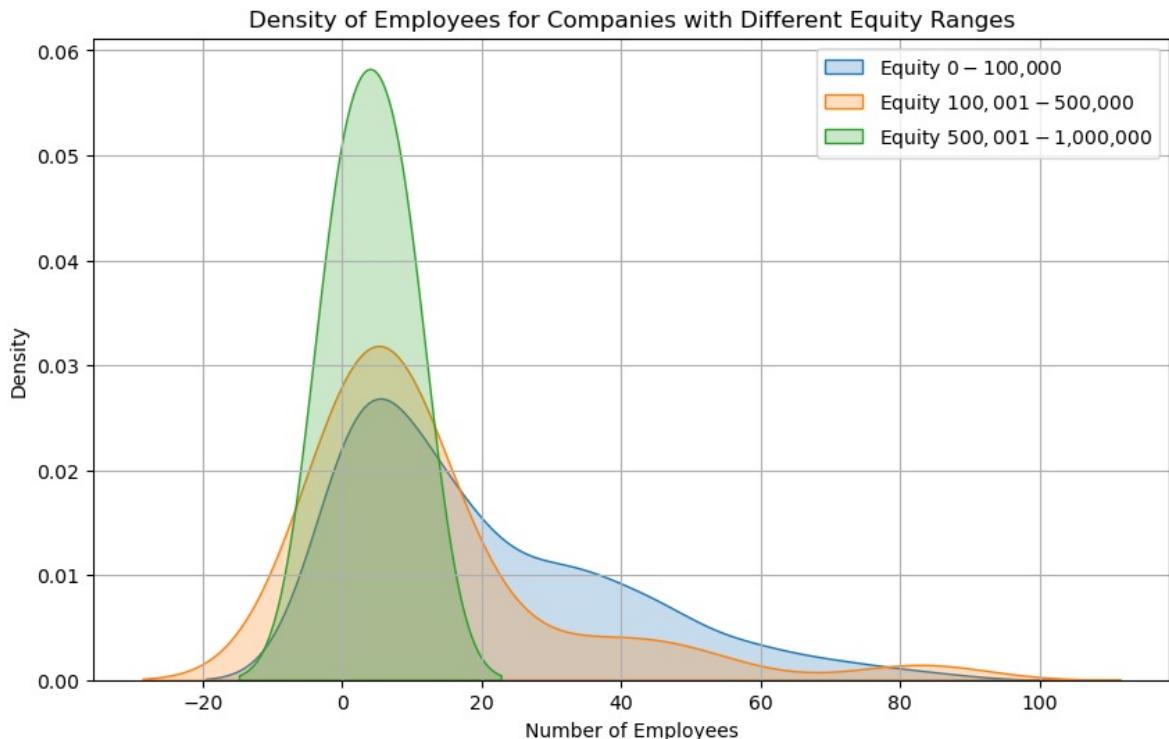
equity_bins = [(0, 100000), (100001, 500000), (500001, 1000000), (1000001, 2000000)]

plt.figure(figsize=(10, 6))
for lower, upper in equity_bins:
    selected_companies = filtered_df_2020[(filtered_df_2020['Equity'] >= lower) & (filtered_df_2020['Equity'] <= upper)]
    if selected_companies['Employees'].var() > 0:
        sns.kdeplot(data=selected_companies['Employees'], fill=True, label=f'Equity ${lower:,} - ${upper:,}')
    else:
        print(f'Skipping Equity ${lower:,} - ${upper:,} due to zero variance or insufficient data.')

plt.title('Density of Employees for Companies with Different Equity Ranges')
plt.xlabel('Number of Employees')
plt.ylabel('Density')
plt.grid(True)
plt.legend()
plt.show()

```

Skipping Equity \$1,000,001 - \$2,000,000 due to zero variance or insufficient data.



The sharp peak in the green curve implies that companies within this equity range have likely optimised their operations, maintaining around 10-15 employees while generating substantial equity. This might indicate that they are focusing on high value services or efficient processes. The broader distributions in the blue (£0 - £100,000) and orange (£100,001 - £500,000) curves suggest that companies with more employees do not necessarily achieve higher equity. These firms might rely more on labor intensive operations where equity growth is slower. The data suggests that there is a sweet spot (around 10-15 employees) where mid sized companies can generate significant equity (£500,001 - £1,000,000). Beyond this range, the returns on increasing workforce size seem to diminish, indicating that focusing on efficiency rather than expansion might be a better strategy for equity growth. I attempted to create a graph using the mean values, but encountered several issues with the code and logic. Despite multiple attempts, neither I nor AI were able to resolve these challenges. As a result, I decided to revert to a 2D density plot with normalisation instead.

```

In [335]: # Prepare the data with 'Equity' and 'Employees'
data = filtered_df_2020[['Equity', 'Employees']].values

# Apply Gaussian KDE to the data
kde = gaussian_kde(data.T, bw_method='scott') # Transpose because gaussian_kde expects shape (n_features, n_samples)

# Prepare the grid for plotting
x = np.linspace(data[:, 0].min(), data[:, 0].max(), 1000) # Grid for Equity
y = np.linspace(data[:, 1].min(), data[:, 1].max(), 1000) # Grid for Employees
X, Y = np.meshgrid(x, y)
positions = np.vstack([X.ravel(), Y.ravel()])
Z = np.reshape(kde(positions).T, X.shape)

# Normalize each x slice so that the integral over y is 1
for i, x_val in enumerate(x):
    slice_density = Z[:, i]

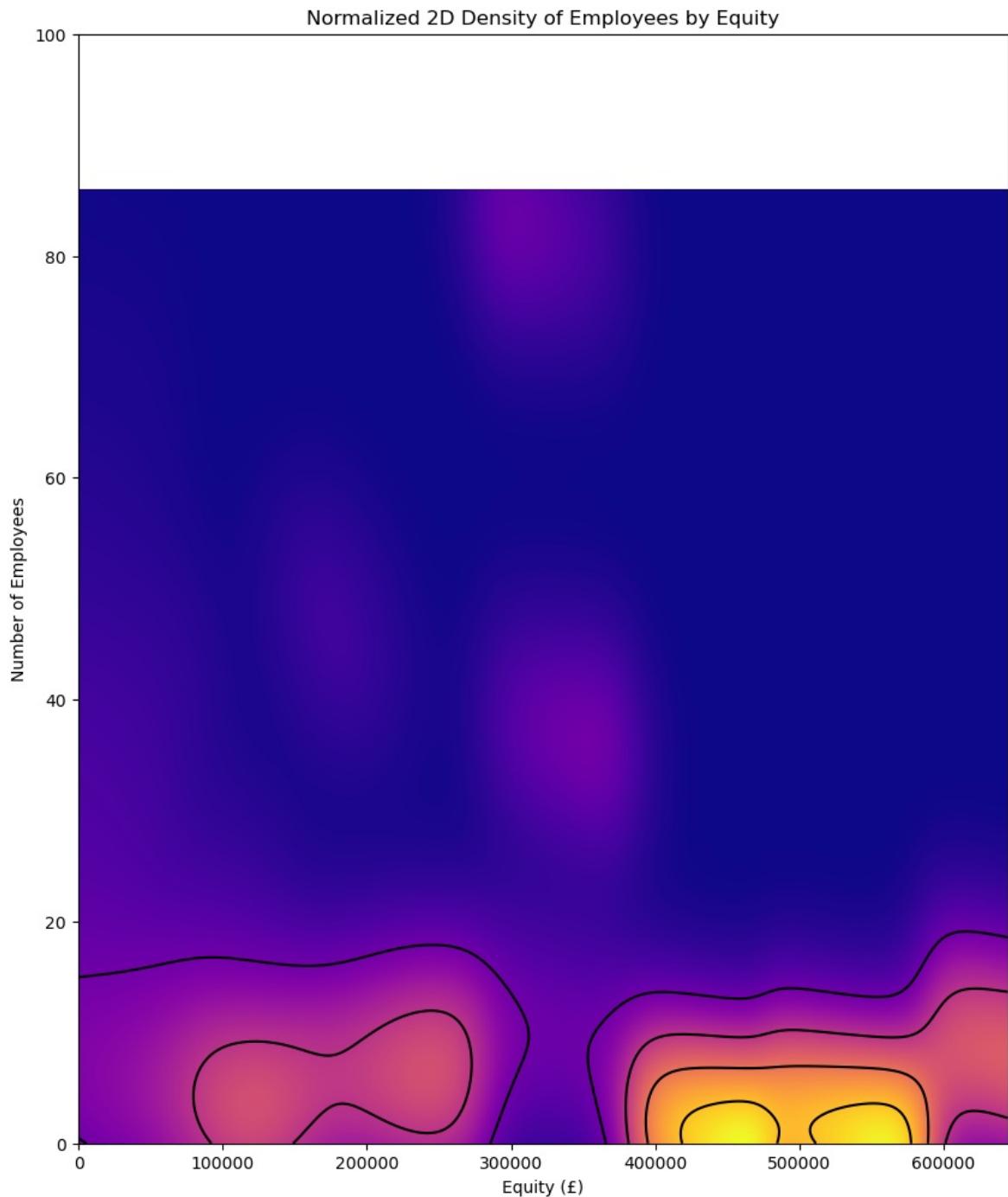
```

```

slice_density /= slice_density.sum() # Normalize the slice
Z[:, i] = slice_density

# Plotting
plt.figure(figsize=(10, 12))
plt.pcolormesh(X, Y, Z, shading='auto', cmap='plasma')
plt.ylim(0, 100) # Adjust this y-limit to zoom in on employee numbers if needed
plt.xlabel('Equity (£)')
plt.ylabel('Number of Employees')
plt.title('Normalized 2D Density of Employees by Equity')
plt.contour(X, Y, Z, levels=5, colors='k') # Add contour lines
plt.show()

```



The highest density areas (highlighted in yellow and orange) are clustered where equity is below £200,000 and the number of employees is below 20. This suggests that many companies in the dataset are small both in terms of workforce and capital. Even as equity increases to £300,000-£600,000, the concentration remains in the lower employee ranges (0-20 employees). This could indicate that companies with higher equity are often able to generate or maintain that equity without needing to significantly expand their workforce. The density rapidly decreases as the number of employees increases beyond 40, regardless of equity. This suggests that companies with larger workforces are less common in this dataset, even if they have substantial equity. The fact that equity increases without a proportional increase in the number of employees indicates that companies might be relying more on capital efficiency or increased labor rather than simply expanding their workforce. This could imply that these businesses focus on maximizing output per employee rather than scaling

through hiring. Again, the concentration in the lower equity and employee ranges suggests that the dataset is skewed toward smaller companies.

```
In [385]: # Preparing the data with 'Employees' and 'Equity'
data = filtered_df_2020[['Employees', 'Equity']].values

# Apply log transformation to Equity to better visualize it
data[:, 1] = np.log1p(data[:, 1]) # log1p is used to handle zeros by applying log(1 + x)

# Applying Gaussian KDE to the data
kde = gaussian_kde(data.T, bw_method='scott')

# Preparing the grid for plotting
x = np.linspace(data[:, 0].min(), data[:, 0].max(), 1000) # Grid for Employees
y = np.linspace(data[:, 1].min(), data[:, 1].max(), 1000) # Grid for log-transformed Equity
X, Y = np.meshgrid(x, y)
positions = np.vstack([X.ravel(), Y.ravel()])
Z = np.reshape(kde(positions).T, X.shape)

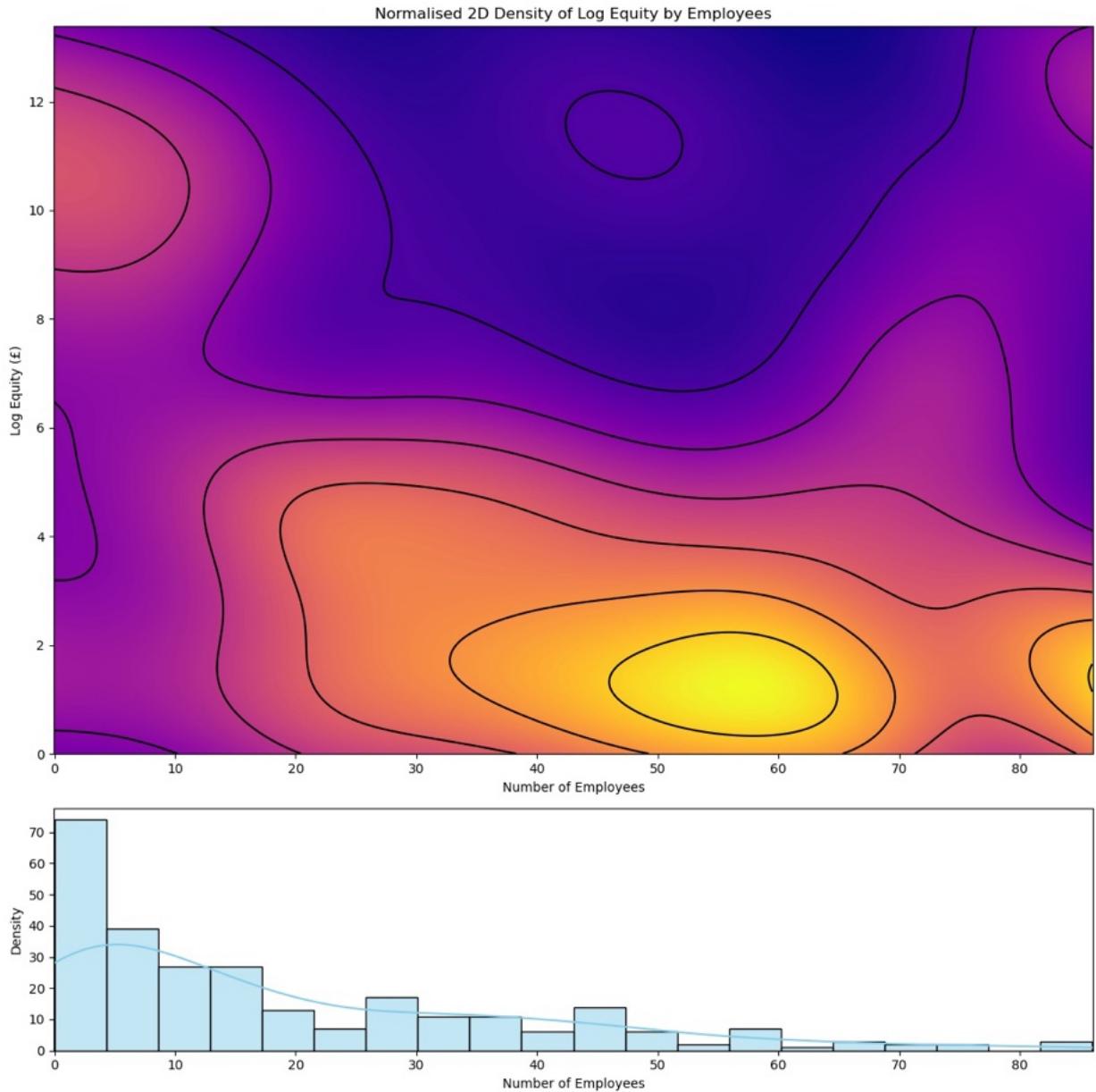
# Normalize each x slice so that the integral over y is 1
for i, x_val in enumerate(x):
    slice_density = Z[:, i]
    slice_density /= slice_density.sum() # Normalize the slice
    Z[:, i] = slice_density

# Plotting
fig, axs = plt.subplots(2, 1, figsize=(12, 12), gridspec_kw={'height_ratios': [3, 1]})

# 1. Plotting the 2D density plot of Employees vs Log Equity
c = axs[0].pcolormesh(X, Y, Z, shading='auto', cmap='plasma')
axs[0].contour(X, Y, Z, colors='k') # Add contour lines to the plot
axs[0].set_xlabel('Number of Employees')
axs[0].set_ylabel('Log Equity (£)')
axs[0].set_title('Normalised 2D Density of Log Equity by Employees')

# 2. Plotting the histogram of Employees below the 2D density plot
sns.histplot(filtered_df_2020['Employees'], bins=20, kde=True, color='skyblue', ax=axs[1])
axs[1].set_xlabel('Number of Employees')
axs[1].set_ylabel('Density')
axs[1].set_xlim(axs[0].get_xlim()) # Line up the x-axis with the plot above

plt.tight_layout()
plt.show()
```



The decision to plot this normalised 2D density chart came after challenges with other visualisation methods, particularly when trying to represent mean values across different equity ranges. The 2D density plot was chosen because it offers a clearer view of the relationship between employee numbers and equity while handling outliers and varying data densities more effectively. By normalising the data, the plot focuses on the core structure of the data, minimising the impact of extreme values.

The plot shows that companies with lower employee numbers (0 to 20) tend to have modest equity levels (log values between 2 and 4), indicating that smaller teams are common among businesses with limited resources. As employee numbers increase beyond 20, equity values become more variable, suggesting that larger companies either attract more capital or need more financial resources to sustain operations. Interestingly, some companies maintain high equity with only moderate employee counts (30 to 60), reflecting a focus on capital efficiency rather than workforce expansion. Identifying these patterns is valuable for understanding causal relationships between equity and workforce size, helping decision makers recognise that financial resources alone don't drive growth. But again, as a whole this was also fairly unsuccessful in understanding the causal relationships within the data set, however due to time constraints I will try to work with what I have.

5.1.3 Causal Relationship Anecdotal

Firstly, within this part of the the manual anecdotal analysis section, my goal is to take a closer look at specific instances within the dataset that stand out due to their relationship between equity and employee numbers, focusing specifically on the year 2020. By diving deeper into these data points, I can gain additional insights beyond what purely statistical analysis might reveal. The purpose here is to examine individual examples and use external research or anecdotal evidence to refine my understanding of the causal relationships. The analysis hints at a "sweet spot" where companies with 10-15 employees achieve higher equity levels, suggesting that mid sized companies can optimise their workforce to generate substantial equity. Beyond this range, increasing the number of employees does not necessarily lead to proportional equity growth, emphasising the importance of efficiency over expansion. However, there was a major issue. Alsonside not getting valuable insights before, I came across the fact that alot of these higher equity and relationships were due

to either companies who overlook lots of smaller companies, thus the employees were disproportionate due to misrepresentation of data, companies who are primarily in different industries but remained in the dataset due to a secondary etc SIC code, or companies who were 'outliers' in a way because of incorrect data input into the filing, and was not checked by companies house prior to submission, with random numbers to fill columns that are required. I knew this information before I commenced this chapter, however when looking at other relationships (not just the ones in this notebook) I found that there were also a major amount of issues. However, removing columns or inputation is extremely risky due to the nature of the data within this dataset and the way it has been structured via the years in different rows. Coming into 2020 gave me insight, but it did not solve these other issues.

Firstly, let's home in on this sweet spot:

```
In [414]: companies_bw_10_15 = filtered_df[(filtered_df['Employees'] >= 10) & (filtered_df['Employees'] <= 15)]
companies_bw_10_15.sort_values('Equity', ascending=True)
```

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fixed Assets ,
1158	WILKINSON OPERATIONS LIMITED	Active	Private limited Company	2009-06-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	321483.0 37
1157	WILKINSON OPERATIONS LIMITED	Active	Private limited Company	2009-06-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	Not Available	382623.0 38
512	J M N RETAILERS LTD	Active	Private limited Company	2014-07-21	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	NaN 6
155	BRADBURY STORES LIMITED	Active	Private limited Company	2013-02-26	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	NaN
154	BRADBURY STORES LIMITED	Active	Private limited Company	2013-02-26	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	Not Available	NaN
...
908	QUINN'S (DUNGANNON) LTD	Active	Private limited Company	2014-12-11	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2017	Not Available	195689.0 22
572	L. S. STORES LIMITED	Active	Private limited Company	2007-02-09	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	£	388868.0 27
762	MILLTOWN TRADING LIMITED	Active	Private limited Company	2011-02-07	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	£	321959.0 32
761	MILLTOWN TRADING LIMITED	Active	Private limited Company	2011-02-07	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	£	305861.0 32
938	RICE'S SUPERMARKET LIMITED	Active	Private limited Company	2002-09-12	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2019	Not Available	489042.0 57

106 rows × 19 columns

After analysing the data, I identified these companies as being within the "sweet spot" range. However, after further investigation, I found limited publicly available information that could clearly illustrate the relationships causes, in addition to not really having a good base to uncover for these relationships. Additionally, when researching companies like Rice's Supermarket, I discovered that it is actually a large supermarket (<https://www.yellowtom.co.uk/144236>). This calls into question the dataset's accuracy, particularly the reported figure of just 15 employees, which seems highly unlikely for a business of that size.

These issues highlights the limitations of the dataset, prompting me to consult ChatGPT for further analysis on the potential causes and factors influencing these relationships.

ChatGPT

*The “sweet spot” of companies with 10-15 employees showing balanced or higher levels of equity could be influenced by several external and internal factors. Here are some potential outbound causes:

1. Operational Efficiency and Lean Management

Companies with around 10-15 employees often find themselves in a position where they can operate efficiently without the complexity of managing larger teams. At this size, businesses can avoid excessive overhead costs, maintain close-knit communication, and make quick decisions. This lean structure allows these companies to maximize equity by minimizing unnecessary expenses, thus creating a balanced financial environment.

2. Focus on High-Value, Specialized Services

Businesses with a smaller, specialized workforce tend to focus on niche markets or high-value offerings. For example, a tech consultancy with 10-15 highly skilled employees can command higher fees while keeping operational costs manageable. This allows them to achieve a good equity position relative to their size without needing to scale rapidly.

3. Owner-Operated and Family-Owned Businesses

Companies in this size range are often owner-operated or family-run businesses. In these cases, the owners are highly invested in both financial stability and operational control. The motivation to balance growth with financial prudence leads to a stable equity situation while avoiding excessive expansion that might require more complex management or external financing.

4. Access to Favorable Financing or Grants

Companies with 10-15 employees might be well-positioned to access government grants, favorable loans, or local business incentives. Smaller companies are sometimes more agile and better suited to qualifying for specific financial programs aimed at encouraging small business growth. These additional financial inputs can boost equity without the need to scale the workforce dramatically.

5. Retention of Key Talent

A business with a small, cohesive team may experience lower turnover rates. If the employees are experienced and committed, the company can maintain high productivity levels without the frequent costs associated with recruiting, training, and onboarding. This stability helps build and sustain equity by optimizing operational costs.

6. Controlled Growth and Scalability

Companies in this range often prioritize controlled, organic growth. They may expand through strategic partnerships, smart investments, or focusing on market segments with less competition. By growing strategically rather than pursuing aggressive expansion, they can build equity steadily while keeping a balanced workforce.

7. Flexible and Adaptive Business Models

Small companies can be more adaptive to market changes, allowing them to pivot their offerings, restructure operations, or realign resources quickly. This agility allows them to capture profitable opportunities or mitigate risks, keeping their equity strong even with a limited number of employees.

8. Owner Equity Investment and Low Leverage

In some cases, the equity may be high because it is largely made up of owner or founder investments rather than external debt. This conservative financial strategy avoids the risks associated with high leverage, leading to healthier equity metrics relative to the company's size and employee base.*

So what are we looking at here? Well, there is seemingly a sweet spot where companies that balance their workforce at around 10-15 employees tend to maintain high levels of equity. These businesses seemingly achieve this by focusing on operational efficiency, keeping overhead low, and making quick decisions without the complexity of managing larger teams.

I then considered the aspect of how small to medium sized businesses impact equity to employee relationships. Specifically, I questioned whether the dominance of these smaller companies in my dataset skews the observed trends. Since the dataset does contain a significant amount of misrepresentations where larger companies are either inaccurately portrayed or overshadow smaller ones I wanted to ensure I understood how this affects the relationship between equity and company size. This led me to investigate whether smaller companies generally tend to have lower equity compared to larger firms, which could be distorting the insights drawn from the data.

ChatGPT

1. The prevalence of smaller businesses (in terms of employees) could cause the overall trends to lean toward lower equity levels.

Larger companies might either be underrepresented or inaccurately portrayed, which could create an imbalance in the observed relationship between equity and employee size. 2. Smaller Companies and Equity:

Generally, smaller companies do tend to have lower equity compared to larger firms. This is because smaller businesses often have limited access to capital, lower revenue, and less accumulated assets. They may also prioritize staying lean and focusing on efficiency rather than building significant equity reserves. Consequently, this could naturally result in the observed trend where smaller companies dominate the lower equity ranges in your dataset.

3. Impact on Equity Ratios: The skew caused by an abundance of smaller businesses could mean that the equity ratios are disproportionately influenced by businesses that are more conservative in taking on debt or raising capital. Smaller firms often rely more on internal funding or small-scale investments, leading to lower equity levels compared to larger enterprises with greater access to funding sources like venture capital or large loans.

4. Comparing Smaller and Larger Companies: If your dataset had a more balanced representation of both small and large companies, you would likely see a clearer distinction in equity levels. Larger companies tend to accumulate more equity due to their ability to scale, expand operations, and generate higher profits. The disparity between small and large firms in terms of equity would become more apparent in a dataset that reflects this range more accurately.*

Furthermore, having a view of this article '<https://usprogram.gatesfoundation.org/news-and-insights/articles/the-powerful-role-small-and-medium-sized-businesses>' I was surprised to see that " 60% of workers earning low wages are employed by small and medium-sized businesses – those with fewer than 100 employees, including 35% of workers at micro-businesses with fewer than 10 employees." Thus, with regards to my dataset, this will directly impact their profitability, and lead to equity growth, bearing in mind the outgoings are maybe not what they 'should' be.

Next I decided to consider investigating low equity with high employee levels (30+). As shown below, a lot of these businesses are overlaid or franchises, as demonstrated by the contents of chapter 4. This insight suggests that analyzing these types of businesses may not provide valuable insights into the financial performance of the expert's supermarket. For instance, even when researching companies like Kashmire Ltd, there is limited information available—mainly just a basic image that appears to be a hotel and some simple financial details. There is virtually no information related to the actual business operations or concepts. This makes it difficult to draw meaningful conclusions that could inform the expert's supermarket strategy.

```
In [432]: companies_30 = filtered_df[(filtered_df['Employees'] >= 30)]  
companies_30.sort_values('Equity', ascending=False)
```

Out [432...]

	Company Name	Company status	Company type	Incorporated on	Nature of business (SIC) 0	Nature of business (SIC) 1	Nature of business (SIC) 2	Nature of business (SIC) 3	Year	Currency	Fix Assets
209	CHARIOTEER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2018	Not Available	Ni
208	CHARIOTEER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2019	Not Available	672743
207	CHARIOTEER LIMITED	Active	Private limited Company	2000-06-19	47110 - Retail sale in non-specialised stores ...	47190 - Other retail sale in non-specialised s...	47250 - Retail sale of beverages in specialise...	47300 - Retail sale of automotive fuel in spec...	2020	Not Available	659132
937	RICE'S SUPERMARKET LIMITED	Active	Private limited Company	2002-09-12	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	46302
1060	SWIFTS SUPERMARKETS LTD	Active	Private limited Company	1999-11-24	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	408535
...
557	KASHMIRE LIMITED	Active	Private limited Company	2011-10-04	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2018	Not Available	Ni
978	SALLEYS FILLING STATION LIMITED	Active	Private limited Company	2017-11-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	38899
720	MCDADE DERRYLIN LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2020	Not Available	Ni
719	MCDADE DERRYLIN LTD	Active	Private limited Company	2015-05-13	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	49202
977	SALLEYS FILLING STATION LIMITED	Active	Private limited Company	2017-11-27	47110 - Retail sale in non-specialised stores ...	NaN	NaN	NaN	2021	Not Available	40226

216 rows × 19 columns

However, Swifts supermarkets has an equity of £310,000 and an average employee count of 83. This is alot more reasonable, since the compnay seeminly only seems to be tied to one main store, that being a euro spar https://www.facebook.com/swiftseurosparlismaskeal/?locale=en_GB.

For further investigation to see if these statistic were aligned, I briefly asked ChatGPT about the statistics in comparison to Swifts, which suggested the figures were infact reasonable to assume true.

*ChatGPT In Northern Ireland, the average number of employees for small supermarkets typically ranges between 10 to 20 workers, reflecting their role as community-focused businesses. When considering financial metrics, such as equity and assets, the values can vary significantly depending on the location, scale, and operational strategy of the supermarket. On average:

1. **Assets:** Small to medium-sized grocery stores often maintain modest assets, typically in the range of £100,000 to £500,000. This includes inventory, property, equipment, and cash reserves.
2. **Equity:** Equity values for small supermarkets can also range widely, generally between £50,000 to £300,000, depending on whether the business is self-funded or benefits from external investments.*

For the expert's supermarket, the Swifts model provides a useful reference point. It suggests that in a similar operational context (community-focused, focusing on a main single-store setup), a well performing supermarket could maintain equity around £300,000 with a higher employee count, potentially due to offering additional services or focusing on customer experience, but it does not necessarily show the relationship or an underlying cause for the relationship to be this way. <https://neighbourhoodretailer.com/swifts-lisnaskea-setting-standards-in-excellence/>

This is where my insights drew closed. Due to the nature of the dataset, and the unclear relationships to begin with, it became practically impossible to either see the underlying causes such as locations or other niche qualities due to the business names being not the real businesses, or the fact that the businesses were misrepresentative of either their numbers or their actual status i.e. not a store but an overlooker. However, I do want to make known that, by looking at supermarkets and stores in NI in general, rather than simply the businesses in my dataset, there are some seemingly related insights that I came up with. Firstly, location may play a role in determining the success and financial structure of smaller shops. In larger cities, small stores tend to perform better financially (<https://www.centreforcities.org/reader/small-business-outlook-2015/cities-small-businesses-and-new-work/>). The potential reason for this is that they can operate with fewer employees due to limited physical space, yet generate higher profits due to increased footfall and demand. This may lead to a higher equity to employee ratio, as these stores probably maximise efficiency and profitability despite their smaller workforce. Another significant factor that I came up with is the focus on niche markets, particularly stores specialising in organic products or premium goods (<https://www.untaylored.com/post/the-long-tail-business-model-explained-unlocking-success-in-niche-markets>). Businesses in this sector more than likely have higher equity despite having a relatively small number of employees. This is largely because these niche stores can command premium prices, leading to better margins and higher equity. The specialised nature of their offerings allows them to remain competitive with fewer employees, driving an efficient equity to employee ratio. Now, with this being said, I did manually investigate my dataset, search many business names to see if I could find more information on these insights, or potentially another dataset I could link to investigate these potential causal factors further. Yet I could not find either of these aspects.

5.2 Confounding Factors

I will now begin by accounting a confounding factor that can be found in my dataset, rather than factors that I am unable to source or link to.

- Incorporated On : Older companies (incorporated earlier) might have more equity due to accumulated assets and stable growth. These companies might also have more employees because of their maturity. Therefore, the Incorporation Year could act as a confounding variable influencing both equity and the number of employees.

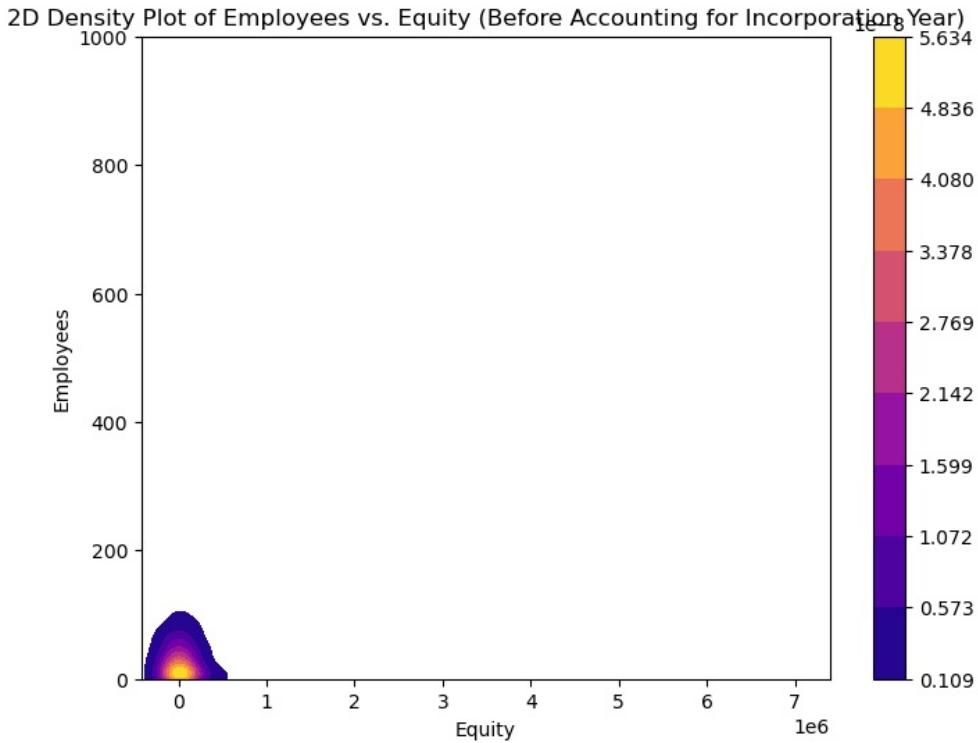
The first part of the code below handles the conversion and preparation of the data. It converts the 'Incorporated on' column, which contains the incorporation date of each company, into a datetime format, just in the case any of these had been reverted or corrupted during previous data manipulations. From this, it extracts the year of incorporation and creates a new column called 'Incorporation Year', allowing us to analyse companies based on when they were established, which serves as a derivative for company age. Next, the code filters out only the relevant columns that are necessary for the analysis, and drops any rows which don't have these 3 variables or are N/A via .dropna().

NB: ChatGPT and StackOverflow has a huge prevalence within this code. Using the samples for previous plots, as well as logic and normalisation points from StackOverflow, in conjunction with my limited python knowledge in comparison, and help from ChatGPT, the code was written for this section, hence the in code comments are very obscure, yet they will show the errors etc one may encounter when doing plots such as these.

```
In [469]: # Convert 'Incorporated on' to datetime and extract the year
df['Incorporated on'] = pd.to_datetime(df['Incorporated on'], errors='coerce')
df['Incorporation Year'] = df['Incorporated on'].dt.year

# Filter out relevant columns and remove missing values
df_filtered = df[['Incorporation Year', 'Equity', 'Employees']].dropna()

# 2D Density Plot (Before Accounting for Incorporation Year)
plt.figure(figsize=(8, 6))
sns.kdeplot(data=df_filtered, x='Equity', y='Employees', fill=True, cmap='plasma', cbar=True) # Added cbar=True
plt.title('2D Density Plot of Employees vs. Equity (Before Accounting for Incorporation Year)')
plt.xlabel('Equity')
plt.ylabel('Employees')
plt.ylim(0, 1000) # Focus on the range with most data
plt.show()
```



The plot above shows a spread of companies with different equity levels and varying numbers of employees. There is a concentration of companies with low equity and few employees, but the relationship is not entirely clear. The lack of clear trends or patterns indicates that there may be underlying factors (like incorporation year) influencing this relationship. It is difficult to draw a direct correlation between employees and equity in this unnormalised form, and thus needs this normalisation element incorporated.

```
In [479]: # Convert 'Incorporated on' to datetime and extract the year
df['Incorporated on'] = pd.to_datetime(df['Incorporated on'], errors='coerce')
df['Incorporation Year'] = df['Incorporated on'].dt.year

# Filter out relevant columns and remove missing values
df_filtered = df[['Incorporation Year', 'Equity', 'Employees']].dropna()

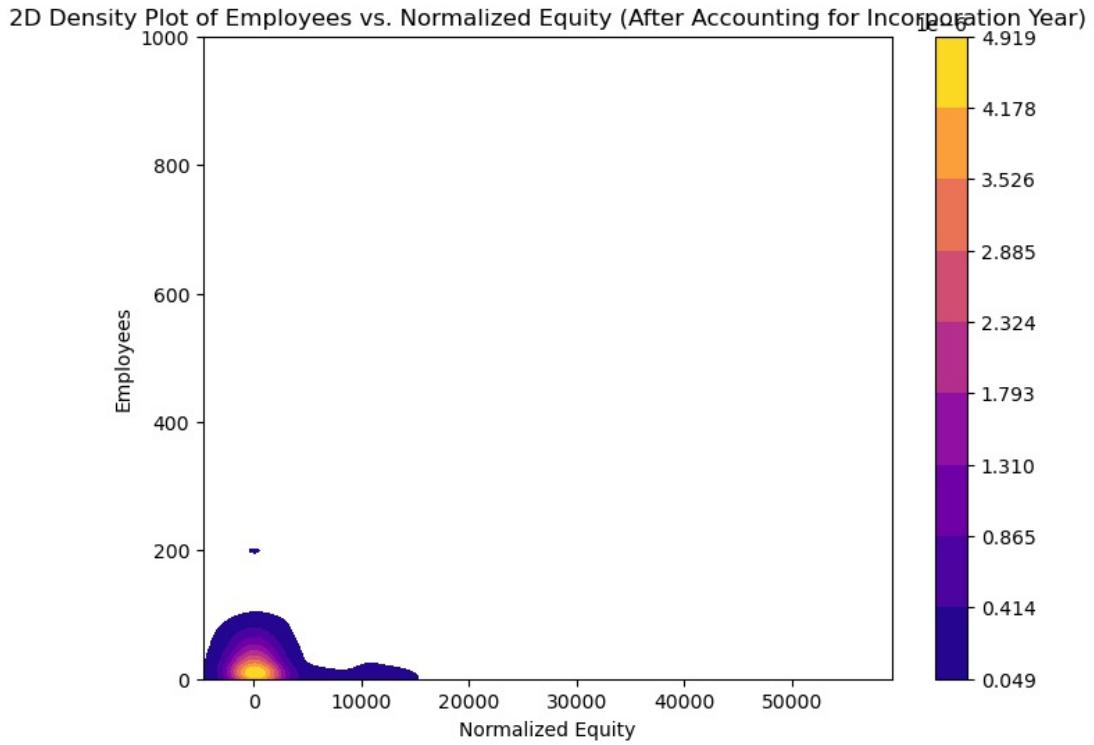
# Stratify by Incorporation Year
year_bins = [1900, 1950, 2000, 2010, 2024]
df_filtered['Year Group'] = pd.cut(df_filtered['Incorporation Year'], bins=year_bins, labels=["<1950", "1950-2000", "2000-2010", "2010-2024", "2024+"])

# Remove rows with NaN values in 'Year Group'
df_filtered = df_filtered.dropna(subset=['Year Group'])

# Calculate median equity for each year group, explicitly setting observed=True
median_equity_by_year_group = df_filtered.groupby('Year Group', observed=True)[['Equity']].median()

# Normalize equity within each year group
df_filtered['Normalized Equity'] = df_filtered.apply(lambda row: row['Equity'] / median_equity_by_year_group[row['Year Group']], axis=1)

# 2D Density Plot (After Accounting for Incorporation Year)
plt.figure(figsize=(8, 6))
sns.kdeplot(data=df_filtered, x='Normalized Equity', y='Employees', fill=True, cmap='plasma', cbar=True)
plt.title('2D Density Plot of Employees vs. Normalized Equity (After Accounting for Incorporation Year)')
plt.xlabel('Normalized Equity')
plt.ylabel('Employees')
plt.ylim(0, 1000) # Focus on the range with most data
plt.show()
```



In the plot above i.e. post-normalisation, the equity values are adjusted within each year group to remove any bias introduced by differences in financial scales across time. The plot shows that, even after accounting for incorporation year, the majority of companies are still clustered in the lower range of equity and employees. However, there is now a more consistent distribution across the x axis, which relates to the normalised equity. This may highlight or atleast provide some indication that the year the company was established does not entirely explain the relationship, at this point. However, due to the slight difference in density and contour structure, there may be some effect that I can look into. Below I have plotted the side-by-side comparison to show the differences. Next, I will zoom in to these 2d density plots.

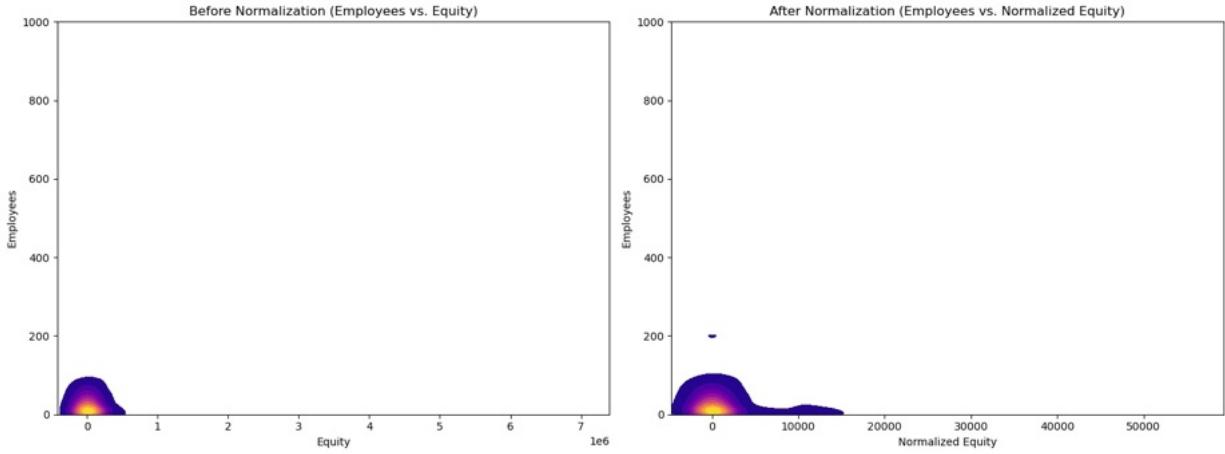
```
In [481]: # Plotting side-by-side comparisons

plt.figure(figsize=(16, 6))

# Before Normalization
plt.subplot(1, 2, 1)
sns.kdeplot(data=df_filtered, x='Equity', y='Employees', fill=True, cmap='plasma')
plt.title('Before Normalization (Employees vs. Equity)')
plt.xlabel('Equity')
plt.ylabel('Employees')
plt.ylim(0, 1000)

# After Normalization
plt.subplot(1, 2, 2)
sns.kdeplot(data=df_filtered, x='Normalized Equity', y='Employees', fill=True, cmap='plasma')
plt.title('After Normalization (Employees vs. Normalized Equity)')
plt.xlabel('Normalized Equity')
plt.ylabel('Employees')
plt.ylim(0, 1000)

plt.tight_layout()
plt.show()
```



```
In [495]: # Convert 'Incorporated on' to datetime and extract the year
df['Incorporated on'] = pd.to_datetime(df['Incorporated on'], errors='coerce')
df['Incorporation Year'] = df['Incorporated on'].dt.year

# Filter out relevant columns and remove missing values
df_filtered = df[['Incorporation Year', 'Equity', 'Employees']].dropna()

# Stratify by Incorporation Year
year_bins = [1900, 1950, 2000, 2010, 2024]
df_filtered['Year Group'] = pd.cut(df_filtered['Incorporation Year'], bins=year_bins, labels=["<1950", "1950-2000", "2000-2010", "2010-2024", "2024+"])

# Remove rows with NaN values in 'Year Group'
df_filtered = df_filtered.dropna(subset=['Year Group'])

# Calculate median equity for each year group, explicitly setting observed=True
median_equity_by_year_group = df_filtered.groupby('Year Group', observed=True)[['Equity']].median()

# Normalize equity within each year group
df_filtered['Normalized Equity'] = df_filtered.apply(lambda row: row['Equity'] / median_equity_by_year_group[row['Year Group']], axis=1)

# Prepare the data for KDE
data_before = df_filtered[['Equity', 'Employees']].dropna().values
data_after = df_filtered[['Normalized Equity', 'Employees']].dropna().values

# Kernel Density Estimation for both before and after normalization
kde_before = gaussian_kde(data_before.T, bw_method='scott')
kde_after = gaussian_kde(data_after.T, bw_method='scott')

x_before = np.linspace(data_before[:, 0].min(), data_before[:, 0].max(), 1000) # Equity
y_before = np.linspace(0, 200, 1000) # Employees - Zooming in on employee range
X_before, Y_before = np.meshgrid(x_before, y_before)
positions_before = np.vstack([X_before.ravel(), Y_before.ravel()])
Z_before = np.reshape(kde_before(positions_before).T, X_before.shape)

x_after = np.linspace(data_after[:, 0].min(), data_after[:, 0].max(), 1000) # Normalized Equity
y_after = np.linspace(0, 200, 1000) # Employees - Zooming in on employee range
X_after, Y_after = np.meshgrid(x_after, y_after)
positions_after = np.vstack([X_after.ravel(), Y_after.ravel()])
Z_after = np.reshape(kde_after(positions_after).T, X_after.shape)

# Normalize each slice for better visualization
for i, x_val in enumerate(x_before):
    slice_density = Z_before[:, i]
    slice_density /= slice_density.sum()
    Z_before[:, i] = slice_density

for i, x_val in enumerate(x_after):
    slice_density = Z_after[:, i]
    slice_density /= slice_density.sum()
    Z_after[:, i] = slice_density

# Plotting before and after normalization side by side
plt.figure(figsize=(16, 6))

# Before Normalization
plt.subplot(1, 2, 1)
plt.pcolormesh(X_before, Y_before, Z_before, shading='auto', cmap='plasma')
plt.xlim(0, 1e6) # Zoom in to the densest area of equity
plt.ylim(0, 200) # Zoom in to focus on employee range from 0 to 200
plt.xlabel('Equity')
plt.ylabel('Employees')
plt.title('Before Normalization (Employees vs. Equity)')
plt.contour(X_before, Y_before, Z_before, levels=10, colors='k') # Add contour lines

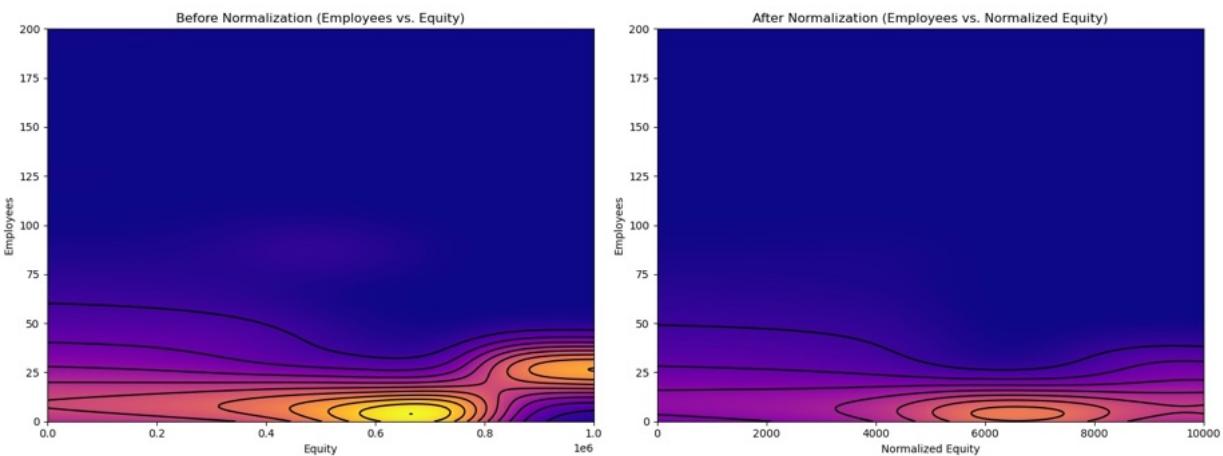
# After Normalization
plt.subplot(1, 2, 2)
plt.pcolormesh(X_after, Y_after, Z_after, shading='auto', cmap='plasma')
plt.xlim(0, 50000) # Zoom out to cover the full range of normalized equity
plt.ylim(0, 1000) # Zoom out to cover the full range of employees
plt.xlabel('Normalized Equity')
plt.ylabel('Employees')
plt.title('After Normalization (Employees vs. Normalized Equity)')
```

```

# After Normalization
plt.subplot(1, 2, 2)
plt.pcolormesh(X_after, Y_after, Z_after, shading='auto', cmap='plasma')
plt.xlim(0, 10000) # Zoom in to the densest area of normalized equity
plt.ylim(0, 200) # Zoom in to focus on employee range from 0 to 200
plt.xlabel('Normalized Equity')
plt.ylabel('Employees')
plt.title('After Normalization (Employees vs. Normalized Equity)')
plt.contour(X_after, Y_after, Z_after, levels=10, colors='k') # Add contour lines

plt.tight_layout()
plt.show()

```



To understand this more, I will start with a general overview of the plots in relation to the entire confounding factor element of this section. The primary objective of this analysis was to determine whether the year the company was incorporated on, which is a stand in point for company age, acts as a confounding factor in the relationship between Equity and Employees. The zoomed in plots above, created from previous analysis in the section, focus on the densest regions of the data, where the majority of companies have relatively low equity and a small number of employees. In the first zoomed in plot, which visualises the relationship before any normalisation, we see a strong clustering effect. Most companies are densely concentrated in this region, and there is little differentiation based on when they were incorporated. Without accounting for incorporation year, the plot suggests that the observed equity to employee relationship might simply reflect the presence of a large number of small, potentially younger companies. However, this observation remains superficial because it fails to control for differences in financial scales and growth opportunities that are typically influenced by when a company was established.

The second zoomed in plot addresses these concerns by normalising equity based on incorporation year groups. This normalisation controls for the inflationary effects and financial scaling differences that occur over time, allowing us to see whether the relationship between equity and employees persists across different periods. After normalisation, I still observe a similar clustering effect in the densest region of the plot, but there are notable differences that indicate the impact of incorporation year. The normalised equity distribution is more consistent across different incorporation periods, suggesting that the variability in the initial plot was partly driven by differences in when companies were established. Older companies naturally have higher equity due to accumulated growth, and this skewed the equity to employee relationship in the unnormalised state.

By adjusting for these differences, the relationship between equity and employees becomes more stable across time periods, highlighting that incorporation year was indeed introducing some bias. However, after normalisation, the remaining distribution indicates that the relationship between equity and employee count is largely consistent. In essence, incorporation year does have a minor confounding effect, but once adjusted for, the equity to employee relationship is revealed to be more stable and largely independent of when the company was founded.

Overall, while incorporation year does contribute to the variability in the initial plots, it is not a dominant factor in determining the relationship between equity and employee count once adjusted. This insight emphasises the importance of accounting for such factors when analysing financial and operational data, ensuring that the core relationships are not masked by time related biases.

5.3 Conclusions & Further Work

1. Conclusions

The analysis in this chapter did not lead to groundbreaking discoveries, but it did offer some insights that can guide further investigation and potential recommendations. The findings highlight trends and patterns that, while not definite, can be useful stepping stones toward better conclusions in the future.

The data analysis suggests that companies with lower debt to equity ratios tend to maintain smaller workforces, which points to a conservative financial approach. These businesses appear to prioritise stability over aggressive growth, focusing on lean operations that

minimise risk. While this observation is not surprising, it reinforces the importance of balancing financial health with operational efficiency, especially in industries where instability is largely a concern.

One interesting, though somewhat limited, finding was the identification of a potential "sweet spot" among mid sized companies with 10-15 employees. These businesses showed higher levels of equity, indicating that this size range might represent an optimal balance between efficient management and financial performance. While this insight could be valuable, it needs further validation across more diverse datasets and industries to confirm whether this pattern holds consistently. Decision makers in industries like retail could consider this as a pre indicator when evaluating workforce and growth strategies, although more particular analysis would be necessary to develop solid recommendations.

Moreover, the analysis of the "Incorporated On" variable revealed patterns that suggest financial metrics tend to stabilise within specific year ranges, hinting at possible periods of economic stability or reporting consistency. This finding could help decision makers better understand temporal influences on financial performance, but more work needs to be carried out for validity.

A significant limitation encountered in the analysis was the presence of outliers and data inconsistencies. Many companies in the dataset were either misrepresented or had inaccurate data entries, making it challenging to draw reliable conclusions. This issue highlights the importance of data quality in any analysis. Before making any concrete recommendations based on the findings, further steps should be taken to clean and validate the dataset. Improved data quality would lead to more trustworthy insights and could significantly enhance the recommendations drawn from future analyses.

Additionally, the dataset is heavily skewed toward smaller companies, which could distort the trends observed. The dominance of small and medium sized companies led to an emphasis on lower equity levels and more conservative financial strategies. While these insights are valuable, they may not be broadly applicable across all business sizes. For decision makers looking to apply these findings, it would be essential to ensure that the recommendations are tailored to the specific size and scale of their operations.

Given these limitations, the current analysis suggests that decision makers should lean toward cautious, efficiency focused strategies. While increasing the workforce is often seen as a growth strategy, the data indicates that for mid sized businesses, prioritising operational efficiency may be more effective. Running leaner operations with well managed resources and balanced financial strategies could lead to more sustainable outcomes. However, this recommendation remains provisional due to the dataset's limitations and should be revisited with more accurate data and broader analysis.

In practice, this could involve decision makers focusing on investments in staff training, automation, and process improvements within their current teams to boost productivity without significantly raising costs. For example, enhancing customer service or upgrading logistics with better technology could offer more value than simply hiring more employees. Utilising a well coordinated, moderately sized team could lead to more stable growth and financial health than expanding too rapidly. This approach could be particularly relevant to certain departments within the expert decision maker's supermarket.

1. Further Work

There are several directions for extending and improving the analysis presented in this chapter. One area involves expanding the dataset to include more diverse company sizes, particularly larger firms. The current analysis was limited by a dataset predominantly composed of smaller businesses, which likely influenced the trends that I observed with large fluctuations, seemingly outliers, that would maybe benefit from having median values in the dataset or other aspects linked to make it more of a well-rounded dataset. Incorporating data from larger enterprises could provide a more balanced perspective and enable a clearer understanding of how equity and workforce dynamics change as companies grow. Such an expansion would also allow for more comparisons between small, medium, and large companies.

Also obtaining a different dataset that includes not just financial metrics but also qualitative data, such as the geographical location of stores, the specific products sold, or other operational characteristics, would have a major benefit to this section. This kind of dataset could reduce some of the challenges encountered in this chapter, such as the presence of companies that oversee multiple stores, which skewed the results. A dataset that focuses specifically on individual stores rather than having companies who are owners of multiple stores, and their data stored based on an accumulation of finances but only office-related employee numbers, would provide more targeted insights and allow for more accurate analysis of causal relationships at the store level. The inclusion of qualitative data would also enable better contextual analysis, helping to explain why certain financial patterns emerge in specific locations or product segments. This could be done by either analysing another data set completely, or linking it to this one, potentially through company names of single stores, which would reduce the rows significantly, but also increase the likelihood of finding causal factors and confounding variables. While on the topic of data linking, combining the current dataset with another that includes more straightforward financial metrics such as profit margins, annual revenue, or cost structures could enhance the analysis. These metrics are often easier to normalise and analyse, offering clearer insights into business performance.

Data quality also needs addressing. The presence of outliers, misreported entries, and inconsistent categorisation significantly impacted the reliability of the analysis. Future work should prioritise data cleaning and validation techniques, such as advanced outlier detection, imputation strategies, or manual verification. By enhancing the quality of the data, subsequent analyses would contain more credible and actionable insights.

Another approach would involve manually cleaning the existing dataset by identifying and removing rows that represent companies unrelated to the scope of the analysis. For instance, companies that primarily operate outside of retail but are included due to secondary industry classifications could be removed. By refining the dataset in this way, the analysis could focus more precisely on the types of businesses relevant to the decision maker's context, leading to more relevant findings.

In addition to refining the dataset, directly engaging with companies included in the dataset could provide valuable qualitative insights. If public data is limited or inaccurate, approaching companies themselves could offer deeper understanding through interviews, surveys, or internal reports. This kind of primary data collection would complement the existing analysis and allow for more accurate conclusions about the causes behind financial and operational trends.

Finally, an alternative approach could involve experimenting with different visualisation techniques and statistical models. Although this chapter primarily used pair plots and density plots, advanced models like decision trees or causal inference models could reveal more complex relationships that were not immediately apparent with simpler methods. By expanding the range of analytical tools used, the analysis could uncover more subtle patterns and provide stronger recommendations within the causes and variables currently in this chapter, and also further ones that could be introduced following the other recommendations I have mentioned above.

In []:

Loading [MathJax]/extensions/Safe.js