SECTION 1: Exploratory Data Analysis (EDA)

Introduction

In the retail industry, data analysis is essential to comprehending consumer behavior and arriving at wise conclusions. In this case, a retailer provides us with daily transaction data. The business owner wants to use this information to learn more about consumer preferences and how well-received products are. Our objective is to assist the owner in optimizing item placement to increase sales by discovering intriguing item relationships.

Association Rule Mining, a potent method for locating hidden patterns in transactional data, will be used to do this. The owner will receive useful insights from this study by learning which things are usually bought together. The business owner may benefit greatly from the knowledge these organizations can offer to strategically organize their inventory to increase sales and satisfy customers. Because of the flexibility of our approach, we may select appropriate tools and algorithms. A thorough report with the methodology, findings, and suggestions will be provided. The goal of this project is to provide the store owner with data-driven insights that will help them improve both their customer experience and business strategy.

Data Exploration and Preprocessing

	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk

Figure 1: Load the dataset.

The dataset contains details about regular business transactions at a shop, such as member numbers, dates, and item descriptions. To find patterns and connections among the store's merchandise, we will examine the data, carry out any required preparation, and evaluate it in the parts that follow.

```
DataFrame Shape: (19351, 3)

Number of rows: 19351

Number of columns: 3
```

Figure 2: Shape of data frame.

The given dataset is shaped like (19351, 3), meaning that there are 19,351 rows and 3 columns in it. Understanding the general size and structure of the dataset is made easier with the help of this information. We have a good amount of data to work with 19,351 rows, which suggests a thorough collection of daily transaction information. There are three columns in

the dataset, and each one has particular data. "Member_number," "Date," and "itemDescription" are these columns. Customer or member identification are stored in the "Member_number" column, transaction dates are stored in the "Date" column, and item descriptions are stored in the "itemDescription" column.

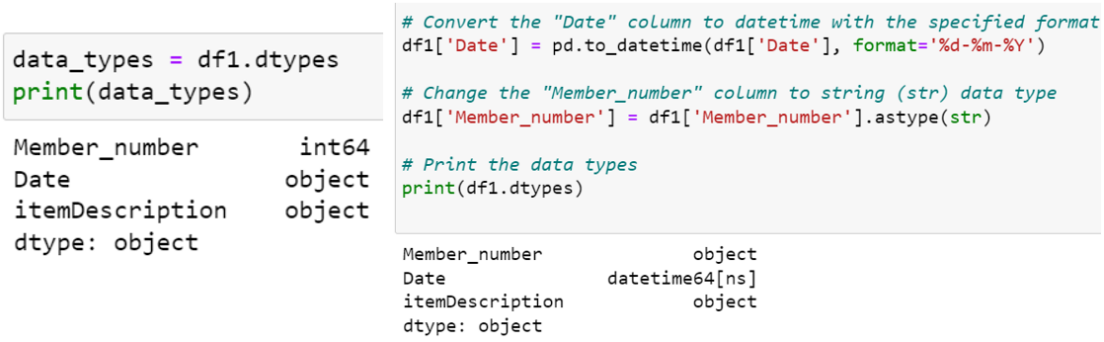


Figure 3: Check and rectify data types.

A basic step in data analysis is making sure that the data in a dataset is consistent in terms of data types and accurately structured. Three different data types were found in the dataset we provided: "object" for the "Date" and "itemDescription" columns, and "int64" for the "Member_number" column.

When one takes into consideration the essential components of data analysis, the need of verifying data types becomes clear:

1. Coherence of Data: Different data formats might cause problems and discrepancies when they are analyzed. These discrepancies might obstruct important insights and compromise the validity of the results reached.

2. Accurate Analysis: The operations that may be performed on data depend on the kind of data. For example, date-based analysis may be performed precisely by changing the "Date" column to a datetime format. This enables for correct grouping of transactions by day of the week or month.

3. Screening: To produce educational visualizations, the right data types must be used. The "Date" column may be transformed into a datetime format, which enables the production of time-based infographics that reveal consumer activity and purchase trends.

As for this dataset, we make sure that the data is consistently represented as datetime objects by converting the "Date" column to a datetime format and “Member_Number” column as string. This vital phase allows us to collect insightful information by resolving data type conflicts and preparing the data for further study and analysis.

	Member_number	Date	itemDescription
count	19351	19351	19351
unique	3873	728	163
top	3050	2015-01-21 00:00:00	whole milk
freq	19	60	1545
first	NaN	2014-01-01 00:00:00	NaN
last	NaN	2015-12-30 00:00:00	NaN

Figure 4: Summary statistics.

We can learn important details and properties of the dataset from the summary statistics supplied. For example, the "Member_number" column has a count of 19,351, meaning that the dataset contains 19,351 items. This provides us with the overall number of interactions or transactions that were noted. There are 3,873 distinct values in the "Member_number" column and 728 unique dates in the "Date" column. These figures illustrate the variety and diversity seen in the dataset when viewed in the context of transaction dates and client identity. It implies that 3,873 different members participated in these transactions over a period of 728 days.

The initial entry in the "Date" column is "2014-01-01," and the final value is "2015-12-30." This gives important details on the dataset's temporal span. It informs us that the data covers the period from January 1, 2014, to December 30, 2015, or over two years. This period of time is crucial for comprehending the transactions' historical context and for any time-dependent analysis.

```
Number of missing values:
Member_number      0
Date               0
itemDescription    0
dtype: int64
```

Figure 5: Check missing values.

One of the most crucial steps in the data validation process is to check a dataset for missing values. This serves several reasons.

1. Data Quality Assurance: Keeping data quality requires locating and fixing missing values. Errors in data input, storage, or collecting may be indicated by missing values, which may compromise the dataset's dependability and integrity.

2. Data analysis: Incomplete or erroneous analyses may result from missing values. They may lead to erroneous conclusions, deceptive visualizations, and skewed data. To guarantee the validity of any analysis, missing values must be found and dealt with properly.

3. Data Imputation: Using the right techniques, missing values may occasionally be imputed or filled in. A more comprehensive dataset is made possible by imputed missing data, which also enhances the accuracy of analyses and forecasts.

It's good to know that none of the columns in the dataset have any missing values. This indicates that the dataset is finished and that no more steps are required to address the missing values.

Number of duplicate rows: 196

	Member_number	Date	itemDescription	DuplicateCount
0	1006	2015-06-14	frankfurter	2
1	1143	2015-04-25	whole milk	2
2	1190	2015-06-11	root vegetables	2
3	1197	2015-11-22	root vegetables	2
4	1213	2015-03-03	beef	2

Figure 6: Check duplicates.

An essential first step in cleaning and analyzing data is identifying and dealing with duplicates in the dataset. This is the significance of finding duplicates:

1. Data Integrity: The accuracy of a dataset may be jeopardized by duplicate records. They might distort data, produce false insights, and influence the outcome of any research.

2. Data Consistency: Duplicates might point to problems with the procedures used for data input or gathering. Accuracy and consistency of data may be preserved by locating and eliminating duplicates.

3. Efficiency of Resources: Removing redundant records can result in data analysis and storage that is more effective. It expedites processing and lessens needless computational effort.

There are obviously 196 duplicate entries in the dataset that is given. Still, there's no evidence to support the assumption that these duplicates come from the same transaction or invoice as customers can choose to buy the same items twice a day for certain reasons. Duplicates are not going to be eliminated since we are unable to definitively identify whether they are the result of separate transactions or problems with data entry.

	Member_number	Date	itemDescription	Weekdays	Days	Months	Years
0	1808	2015-07-21	tropical fruit	Tuesday	21	Jul	2015
1	2552	2015-01-05	whole milk	Monday	05	Jan	2015
2	2300	2015-09-19	pip fruit	Saturday	19	Sep	2015
3	1187	2015-12-12	other vegetables	Saturday	12	Dec	2015
4	3037	2015-02-01	whole milk	Sunday	01	Feb	2015

Figure 7: Create new columns.

For the purpose of improving the dataset's analytical capabilities and adding more context for analysis, it is imperative to create new columns depending on the "Date" column. These new columns are useful for the analysis for the following reasons:

1. Weekdays: The particular day of the week that each transaction took place is disclosed in the "Weekdays" column, which is generated from the date. This makes it possible to investigate possible trends in consumer behavior, such if particular products are more popular on particular weekdays like weekend buying habits.

2. Days: The day of the month for every transaction is recorded in the "Days" column. This can be helpful in determining whether there are any monthly fluctuations or trends in the way people shop, such as higher sales on payday or other particular days.

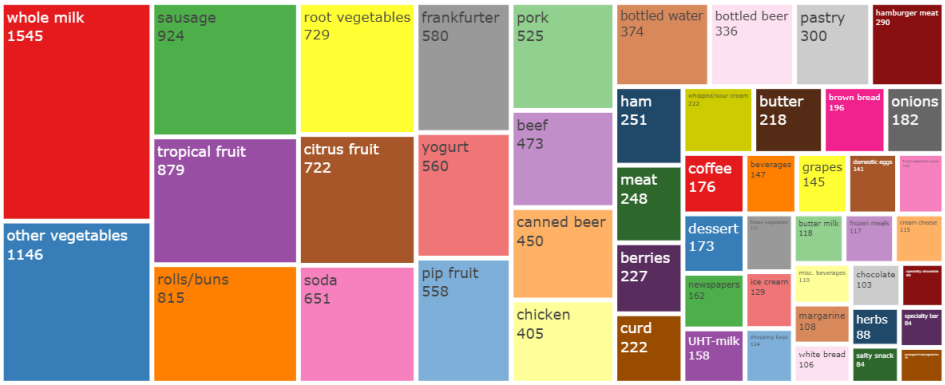
3. Months: The shortened month names are taken from the date and placed in the "Months" column. Finding seasonal trends, unique promotions, or holidays that affect consumer choices can be facilitated by doing a monthly analysis of sales data.

4. Years: The year of each transaction is shown in the "Years" column. In order to evaluate changes in consumer behavior or sales patterns over time, year-by-year comparisons across data sets are crucial for data separation and analysis.

The addition of these columns makes the dataset more illuminating and appropriate for in-depth examination. These columns make it possible to investigate patterns and trends in the timing of transactions, which can provide insightful information for the business plan and decision-making of the store owner.

Exploratory Data Analysis (EDA)

Top 50 Products Purchased



Least 50 Products Purchased

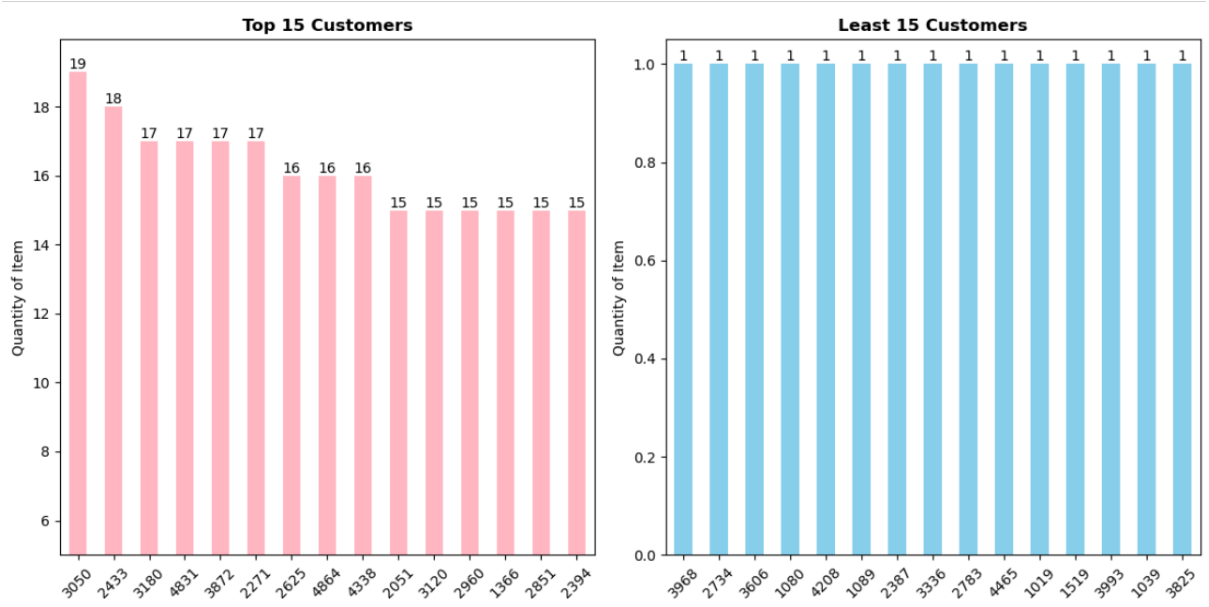
The above interactive tree maps, which show the top 15 and bottom 15 goods bought in the store, provide insightful information that may help the business owner make wise judgments. The most popular items are displayed in the top graph. With 1,545 purchases, "whole milk" is clearly in the lead, followed by "other vegetables" and "sausage." The business owner may use this information to determine which goods are the most well-liked by customers.

The owner is able to rapidly identify the important elements driving sales thanks to the visual depiction. This information is useful for marketing campaigns centered around popular products as well as stock management, guaranteeing that high-demand items are continuously in stock. For example, if "whole milk" is the best-selling item, the owner can think about strategically putting it in the shop or doing promotions to increase sales even more. Additionally, it is worthwhile to investigate any trends in the purchasing habits of consumers who purchase "whole milk."

The lowest frequency of purchases for the items is indicated in the below graph, like specialty vegetables, kitchen utensils and many more, Customers are less interested in these things, so a deeper look may be in order. The owner can improve inventory management by identifying the least-purchased goods. For example, if sales of a particular item are regularly poor, the owner may want to reconsider how they market or arrange their shelf space for these products. In addition, the proprietor may determine whether these items complement the store's total

inventory or whether they ought to be swapped out for attractive options. Additionally, it's critical to keep an eye on the least-purchased things to make sure they're not driving up needless storage expenses.

All things considered, these graphs provide a brief visual summary of the store's product selection, emphasizing top sellers and underwhelming goods. While the least-purchased items force the owner to make data-driven decisions to enhance inventory management and maybe look into chances to raise sales of these products, concentrating on the top products allows the owner to capitalize on successful patterns.



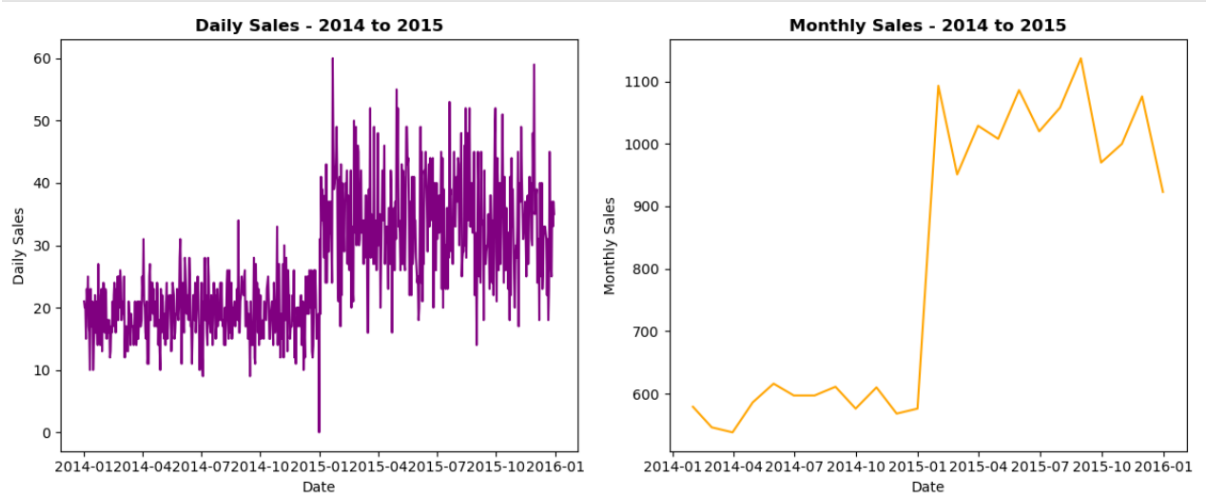
The above bar graphs provide the business owner a thorough understanding of the clients' purchase patterns and offer insightful information that may assist in making deft selections. The individuals who have continuously been the most frequent customers in the store are displayed in the graph of the top 15. With 19 transactions, customer "3050" sticks out and shows a high degree of fidelity to the business. These best clients are very important to the store's earnings, and their choices and actions have a big influence on how well the company does.

The owner may adjust marketing and engagement tactics to keep these consumers loyal and even encourage them to spend more by getting to know their preferences. Personalized offers, targeted marketing, and loyalty programs are all useful instruments for building enduring relationships with these important clients. By providing consumers with a customized shopping experience, the owner can both keep their business and entice them to tell their network about the store, therefore becoming brand promoters.

On the other hand, the graph with the lowest 15 customers shows which ones bought least within the time frame under observation. A smaller degree of interaction with the shop is shown by the fact that some of these consumers have only made one transaction. Even though they might not make up a sizable amount of the business's income, these patrons still add to

the store's variety. Every consumer has a chance for the retailer to deliver a satisfying shopping experience, regardless of how often they make purchases. Through behavior analysis, the owner may investigate ways to get these less regular consumers to come in more often and spend more money. They may get interested in the store again if special offers, better customer service, and personalized suggestions are made.

Overall, the business owner may identify and prioritize clients depending on how frequently they purchase with the help of these visualizations. The proprietor might try to increase sales and fortify the loyalty of the most frequent users by concentrating on them, which will ultimately raise the store's earnings. Strategies targeted at the less engaged customers can also increase their involvement and motivate them to become regular and devoted consumers. The secret is to recognize and satisfy the distinct requirements and inclinations of every group, which will eventually result in a storefront that draws in a diverse clientele.

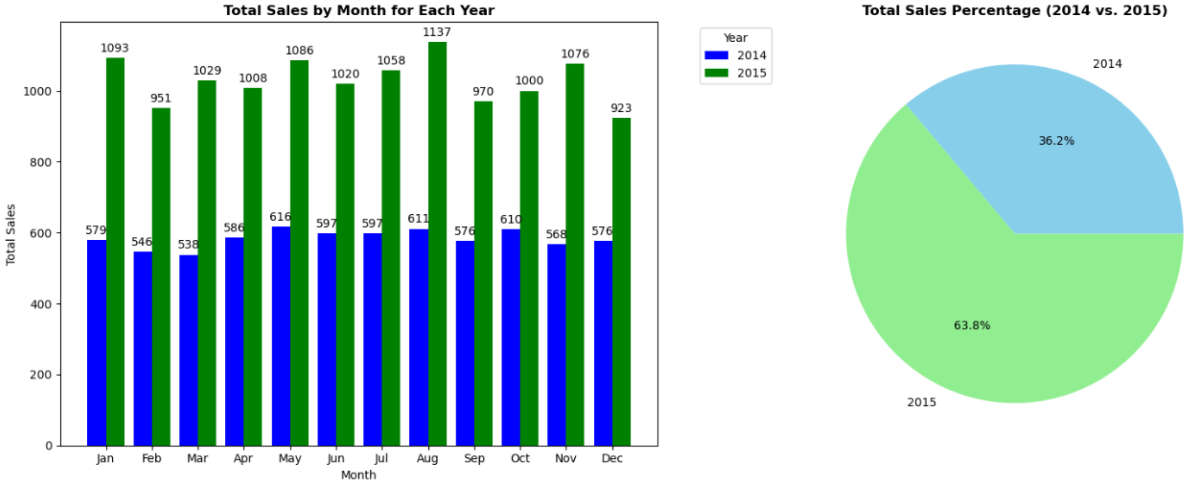


The attached line plots offer insightful information about the store's sales patterns throughout the two-year period between 2014 and 2015. We can see the store's daily sales patterns by looking at the first subplot, which shows the daily sales data. The data indicates variations in daily sales, which mirror the patterns of customer behavior over the course of the week and year. It's obvious that certain days have better sales than others. Comprehending these variances might aid the business proprietor in making prompt selections about personnel and inventory control. For example, days with greater sales can call for hiring more employees to manage the extra foot traffic and provide top-notch customer service. In contrast, the owner can decide to reduce the workforce on days with fewer sales in order to save operating expenses.

The second subplot, which focuses on monthly sales statistics, offers a more comprehensive viewpoint. With the help of this approach, we can spot trends in sales that appear throughout the whole year. We can track seasonal variations, identifying peak sales months and months with reduced sales. For example, increased sales in December of both 2014 and 2015 demonstrate the influence of the Christmas season on customer behavior. For the business owner to organize marketing campaigns, promotions, and inventory control, this information

are crucial. Peak months might have the highest sales if high-demand times are anticipated, and product offers are optimized.

In summary, the business owner can find out a lot about sales trends from these representations. Real-time decision-making is made possible by the daily sales plot, which improves both operational efficiency and customer experience. In the meantime, long-term strategic planning to take advantage of seasonal trends is made possible by the monthly sales plot. Through the utilization of these data, the proprietor may customize tactics to fulfill client requirements, maximize revenue, and eventually propel the store's prosperity. Comprehending the fluctuations in sales is a crucial element of proficient retail operations.



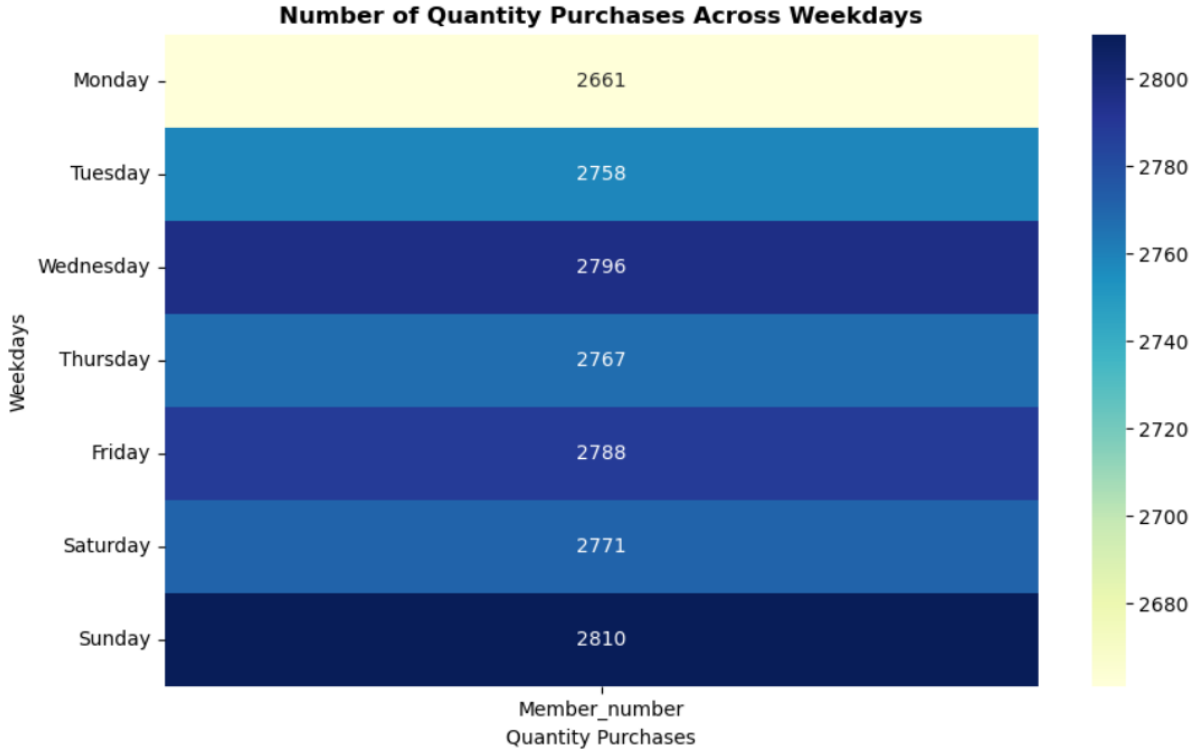
Sales patterns for 2014 and 2015 are clearly shown in the grouped bar chart that shows total sales by month for each year. The blue bars show the sales statistics for 2014 and highlight a number of noteworthy patterns. Sales peak in January, then progressively decline in February and March, before rising once more in April. Up until May, when it reaches its peak, this increasing pattern persists. Sales performance is steady over the summer, but it varies significantly in the fall before reaching its lowest point in December.

On the other hand, 2015 shown by the green bars displays distinct patterns. Like the previous year, sales picked up speed in January. On the other hand, sales decline in February and significantly increase in March. Peak sales are achieved in April as the upward trend persists. After then, sales fluctuate sometimes but are generally constant throughout the year. Sales start to fall slightly in December but are still quite good. The pie chart shows that 63.8% sales during 2015 while only 36.2% sales during 2014, indicating business boomed over the two years.

The business owner can predict times of increased demand by looking at the chart, which shows clear seasonal tendencies in sales. This knowledge s changes to promotions, marketing tactics, and inventory control. For example, sales often peak in December, indicating the value of holiday-themed advertising and having a good supply of gift products on hand. By comparing the two years, the owner may evaluate how the shop has changed over time and

make more informed judgments about budget allocation and sales objectives for the next year. Understanding month-by-month variations in sales is essential for effective inventory control. In order to efficiently fulfill consumer demand and minimize stockouts and surplus inventory during periods of low demand, it is recommended to stock up on items ahead of peak months.

To sum up, this grouped bar chart gives the business owner the ability to make data-driven decisions in addition to offering insights into sales patterns. The owner may maximize sales and improve the customer experience by staffing, marketing, inventory management, and overall shop operations by knowing the highs and lows in sales throughout the year. Achieving sustainable corporate growth and long-term strategy planning both benefit greatly from these insights.



The amount purchases made on several weekdays are depicted in the heat map. The graphic makes it clear that Sunday is the day with the most purchases, with Saturday and Friday following closely after, suggesting that weekends are the biggest times for shopping. Tuesday and Thursday exhibit somewhat lower purchase numbers, whereas Tuesday and Thursday have somewhat identical purchase counts.

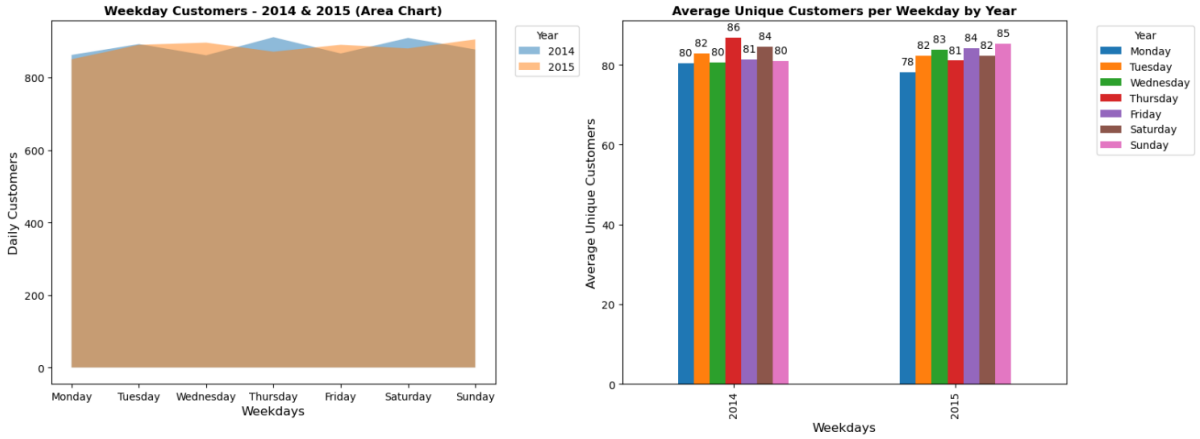
Comprehending the buy distribution on weekdays can help with several elements of shop management. For example, it can aid in staff schedule optimization, guaranteeing sufficient resources are allotted over weekends when shopping is most popular. Additionally, it may direct marketing tactics, such as scheduling unique specials or events on days with high traffic, like Sundays.

Additionally, by understanding the daily shopping trends, the business owner may customize inventory control. In order to fulfill the unique demands of various days, they can modify

stock levels and product positioning, making sure that popular goods are always accessible during periods of high shopping demand. Thus, the general shopping experience and consumer happiness may be improved.

The density plot and histogram offer insightful information about how consumers' purchases are distributed. The right-tail skew of the histogram suggests that most consumers only make a modest number of transactions. This indicates that the majority of buyers will probably purchase one to seven things total. The lower range buy concentration indicates that smaller, more frequent purchases are typical consumer behavior. By using this data, inventory management may be d and regularly purchased products can be kept well-stocked. One set of clients could regularly make significant purchases, while another group would make smaller, more frequent purchases. Marketing plans may work better if they are customized for certain markets. For instance, there can be two types of customers, those who often make big purchases and those who make smaller, more frequent ones. It may be more successful to modify marketing tactics for certain markets.

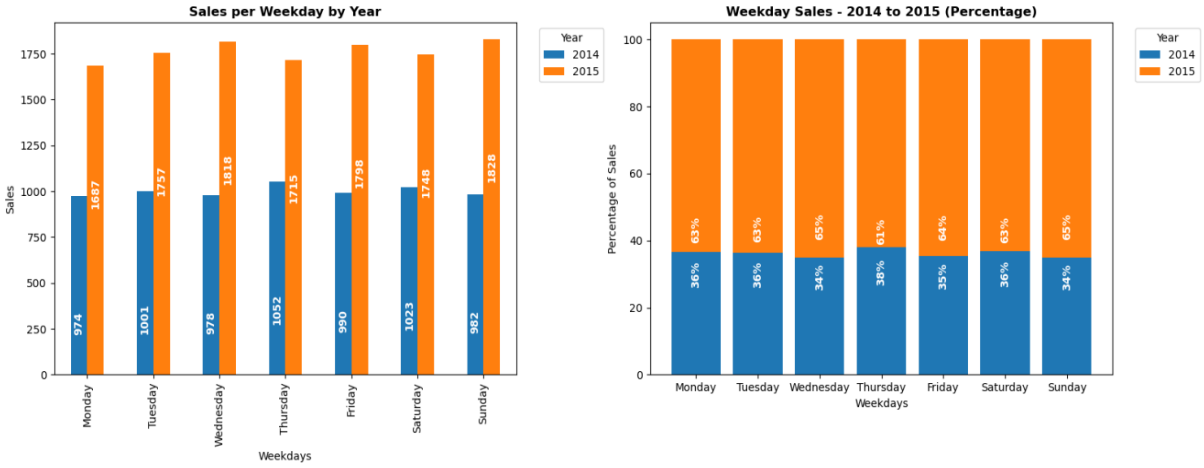
.



The data analysis and visualizations offer insightful information on the weekday behavior of customers in 2014 and 2015. The daily client count for each of the two years, 2014 and 2015, is shown in the area chart for weekdays. Similar trends of everyday consumers are seen throughout weekdays in both 2014 and 2015. Sundays had the most daily consumers, followed by Saturdays. The number of customers then decreases over the workweek, with Thursday and Friday having the lowest numbers. Saturday and Sunday see the highest number of customer visits throughout the weekend, indicating that the store sees an increase in foot traffic during leisure and weekend shopping. The two years' trends are consistent, suggesting that consumer behavior about the days they visit the business stays mostly same. The area chart may be used to see daily trends in consumer behavior and spot any notable variations between the two years.

The bar graph shows the average number of unique consumers for each weekday in 2014 and 2015. It demonstrates that the days with the highest average number of unique clients are regularly Monday, Tuesday, and Wednesday, with somewhat lower averages seen on Thursday, Friday, and Saturday. Sunday is a crucial day for consumer interaction since the figure shows that it has the largest average number of unique customers in both years. This indicates that while there is less foot traffic towards the conclusion of the workweek, Sunday shopping is a popular choice for consumers. There’s a noticeable change on Wednesdays. Wednesdays had more unique customers on average in 2015 than they did in 2014. The business owner would find it interesting to look into why Wednesdays witnessed a rise in the number of unique consumers. The planning of personnel and marketing initiatives benefits greatly from this knowledge. To enhance customer service and satisfaction, it might be beneficial to have more personnel available on days when the average number of unique customers is highest.

To sum up, these visualizations provide an extensive perspective on typical unique customers during the weekdays as well as daily customer patterns. Comprehending these trends may aid in making informed decisions regarding hiring, marketing tactics, and general company operations, ultimately improving customer satisfaction and increasing revenue.



The two graphics offer insightful information on sales patterns according to years and weekdays. The first graph displays the total sales for the years 2014 and 2015 for each workday. It is apparent that every workday in 2014 and 2015 saw a rise in overall sales. This implies an increase in overall sales over this time frame. The graph shows that sales are consistently greater on Saturday and Sunday than they are on the other weekdays. This is consistent with the previously noted patterns of shopping. The proprietor might concentrate on taking advantage of these busy times to provide deals or promotions.

The percentage distribution of sales for 2014 and 2015 is shown in the second graph. The sales for each weekday are shown as a proportion of the overall sales for that year. Between 2014 and 2015, the percentage distribution over weekdays stays the same. The weekday sales proportions remain relatively unchanged. For instance, sales on Mondays are about 36-37% of total sales, but sales on Saturdays and Sundays are between 63-65%. The owner may better anticipate consumer behavior and adjust staffing levels or marketing campaigns by taking advantage of the constancy in weekday sales percentages. Decisions based on historical data may be made with confidence because of the consistent trends.

In overall, the graphics show an increase in overall sales between 2014 and 2015, with weekends continuing to be the most popular days for shopping. Over the course of the two years, the distribution of percentages among weekdays has remained relatively steady, offering important information for sales and marketing efforts.

To sum up, the store owner may use the insightful information gleaned from the exploratory data analysis (EDA) of the daily transaction data to strategic decision-making. The best and least popular items, consumer purchasing patterns, sales trends by month and year, and sales patterns on weekdays are some of the important conclusions drawn from the data. These revelations have an impact on marketing initiatives, consumer interaction, inventory management, resource allocation, and sales strategy.

The business owner may improve operations and increase profitability by making well-informed decisions based on their understanding of client behavior and sales patterns. EDA offers a strong framework for more research and the creation of useful suggestions. By using a data-driven strategy, the retailer may enhance the whole shopping experience for customers and better position itself in a competitive market.

SECTION 2: Preparation of Dataset

Among non-numeric, categorical datasets, Association Rule Mining (ARM) is a potent data mining approach that locates important relationships, correlations, or associations. Transactional databases, relational databases, and other repositories are rich in hidden patterns that may be found using ARM, in contrast to many machine learning methods that work solely with numerical data. Assembling basic If/Then statements that reveal relationships between seemingly separate items is the basic idea underlying association rule mining operations.

Making sense of transactional data is the aim in the field of ARM, especially when it comes to situations such as market basket analysis. An distinct set of data preparation procedures designed to manage categorical information is necessary for association rule mining, as opposed to depending solely on intricate mathematical calculations.

The key to maximizing the potential of ARM is methodical data preparation. The steps for preparing data are as follows:

1. Data Sorting

	Member_number	Date	itemDescription	Weekdays	Days	Months	Years
13331	1000	2014-06-24	whole milk	Tuesday	24	Jun	2014
4843	1000	2015-03-15	sausage	Sunday	15	Mar	2015
8395	1000	2015-03-15	whole milk	Sunday	15	Mar	2015
1629	1000	2015-05-27	soda	Wednesday	27	May	2015
17778	1000	2015-05-27	pickled vegetables	Wednesday	27	May	2015
2047	1000	2015-07-24	canned beer	Friday	24	Jul	2015
18196	1000	2015-07-24	misc. beverages	Friday	24	Jul	2015
6388	1000	2015-11-25	sausage	Wednesday	25	Nov	2015
9391	1001	2014-02-07	sausage	Friday	07	Feb	2014
11046	1001	2014-12-12	whole milk	Friday	12	Dec	2014

The DataFrame must first be sorted by "Member_number" and "Date." In order to organize and get the data ready for association rule mining, this is crucial.

2. Grouping and Concatenating Item Descriptions

	Member_number	Date	itemList
0	1000	2014-06-24	whole milk
1	1000	2015-03-15	sausage,whole milk
2	1000	2015-05-27	soda,pickled vegetables
3	1000	2015-07-24	canned beer,misc. beverages
4	1000	2015-11-25	sausage
5	1001	2014-02-07	sausage
6	1001	2014-12-12	whole milk
7	1001	2015-01-20	frankfurter,soda
8	1001	2015-04-14	beef
9	1001	2015-05-02	frankfurter

After then, these attributes are used to aggregate the data, and item descriptions from each transaction are combined to create an extensive list of all the products that were bought in each instance.

3. Exploding Item Lists:

itemList	Member_number	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	bags	baking powder	bathroom cleaner	beef	...	turkey	vinegar	waffles	whipped/sour cream	wh
0	1000	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	1001	0	0	0	0	0	0	0	0	1	...	0	0	0	0	
2	1002	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	1003	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	1004	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 164 columns

The merged "itemList" is divided into several lists containing item descriptions. This is an important step that allows the dataset to be transformed into a format that rule mining algorithms can use. A more detailed examination of item relationships is made possible by expanding these lists such that every item in a transaction is shown in its own row. By reorganizing the data, significant patterns and associations within categorical transactional datasets may be found by applying association rule mining algorithms to the data in the proper format.

4. Creating a DataFrame with Item Count

Grouping the data by "Member_number" and "itemList" comes next, once the DataFrame has been expanded to show each item in a single row. We may create a summary of item transactions by counting the occurrences of each item for each member by using this grouping operation. The information is then destacked to create a matrix where elements are columns and members are rows. Zeros are used to fill in any NaN values that arise from the unstacking in order to guarantee a full representation. This reorganized matrix is the basis for additional examination, especially when it comes to association rule mining, which allows for the identification of distinct patterns and connections between items and members.

5. Converting Item Counts to Binary Format

itemList	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	bags	baking powder	bathroom cleaner	beef	berries	...	turkey	vinegar	waffles	whipped/sour cream	whisky	whi brea
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	1	0	...	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 163 columns

The next step is to transform the item counts into a binary representation, where an item is represented by 1 or 0 depending on whether it is present or absent. For association rule mining algorithms that work with the inclusion or exclusion of items in sets, like Apriori and

FP-growth, this binary form is essential. By concentrating on whether an item was purchased rather than how frequently it occurs, the conversion streamlines data. Now that it has been created, the DataFrame known as df_one_hot is formatted properly for association rule mining. Each row in this modified DataFrame represents a member, and each column an item. Binary values in each column indicate whether or not the member bought the associated item. Finding association rules within the dataset is based on this binary representation.

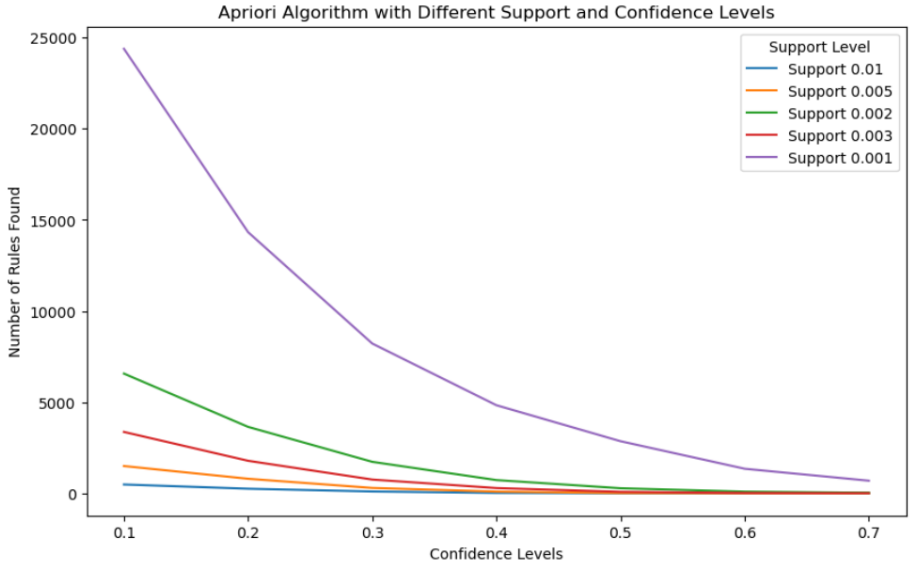
All things considered; successful data preparation creates the foundation for association rule mining. Analysts can uncover complex linkages and patterns in data by organizing it into groups and converting it into the right format. Algorithms may concentrate on itemset inclusion since the data is made simpler by the translation to binary representation. In conclusion, using association rule mining techniques to derive useful insights from non-numeric datasets requires careful data preparation.

SECTION 3: Methodology

Apriori and FP-growth are two well-known algorithms used in Association Rule Mining (ARM) to find patterns and relationships in transactional datasets. The underlying idea of both techniques is to extract frequent itemsets, or groups of items that frequently occur together in transactions.

The Apriori algorithm is a traditional search method that finds frequently occurring itemsets by searching in a level-wise, breadth-first manner. The process starts with determining which individual items have enough support, then iteratively lengthens itemsets until the minimum support level is reached and no more extensions are feasible. The "Apriori property," which states that all of an itemset's subsets must likewise be frequent if an itemset is frequent, is the central principle.

FP-growth Algorithm: This other method, also known as frequent pattern growth, creates a small data structure known as an FP-tree. Large datasets are easily handled by this tree structure because it effectively captures the connections between many itemsets. To recursively mine frequent itemsets without explicitly generating candidate itemsets, FP-growth uses a divide-and-conquer technique. Regarding this project, we'll compare the analysis utilizing both approaches.



The trade-off between support levels and the quantity of rules found is depicted in the plot. Fewer but more general regulations arise from a higher support level, whereas more detailed and numerous rules result from a lower support level. An optimal compromise is offered by the inflection point at a support level of 0.002, which allows for enough granularity in rule generation without overburdening the analysis with an excessive number of rules.

A thorough investigation was carried out to see how support levels and the number of regulations correlated. Furthermore, the interplay between varying support and confidence levels was examined to offer a more sophisticated comprehension of how these factors

impact the Apriori algorithm's results. The link between support, confidence, and the final number of rules was examined with confidence levels of 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, and 0.1.

On the basis of the resulted rules, two different scenarios were then investigated:

1. **Scenario 1 : Low Support, High Confidence (0.002 is the selected support level):**

The business owner places more weight on relationships' dependability than how frequently they occur. In order to achieve a balance between specificity and generality, a support level of 0.002 was used. Strong correlations between goods were identified with high confidence levels (0.7 & 0.5), meaning that customers who purchase one item are quite likely to purchase another.

Justification or High Confidence with Low Support (Selected Support Level: 0.002):

1. Strike a Balance Between Specificity and Generality

The figure shows that a modest number of rules is provided at a support level of 0.002, which strikes a balance. This guarantees a significant analysis that captures item correlations that are both common and uncommon.

2. A reasonable amount of rules

15,798 rules are produced at the selected support level of 0.002, providing a manageable collection for study. This guarantees that the regulations are not too generic or too particular, but rather perceptive and applicable to the shop owner.

3. Level of Detail in Item Associations

Support level enables the detection of item correlations that are fairly specific, offering a comprehensive perspective of client preferences without going too deep into patterns.

4. Exploratory study Consideration

A support level of 0.002 is appropriate considering the exploratory nature of the study and the goal of fully comprehending client preferences. Without raising the bar too much, it makes it easier to find intriguing patterns.

5. Consideration of Computational Resources

Although more intricate rules may result from lower support levels, a balance has been kept guaranteeing appropriate computing performance. Given the limitations of resources and practical execution, this concern is essential.

To sum up, the support level of 0.002 is purposefully selected to maximize the trade-off between specificity and generality, offering the store owner a useful set of rules that will enable them to understand client preferences and make well-informed business decisions.

Fine-tuning the rules based on lift and confidence criteria is the next phase in the association rule mining (ARM) technique, which begins with a thorough exploratory study that yields a minimum support level of 0.002. In order to guarantee that the produced rules are dependable and informative, the reasoning behind the selection of certain values for these thresholds is essential.

1. Association Rules (All):

Lift > 1: At first, all rules are taken into consideration if their lift is more than 1. Items are considered independent when the lift value is 1. The research concentrates on rules that offer at least some information about item relationships by establishing a minimum lift criterion of 1.

2. Association Rules (Lift > 2):

Greater Lift Threshold (2): In order to focus on stronger linkages between objects, a greater lift threshold of two is used in this stage. When the antecedent is purchased, there is at least twice as much chance of purchasing the consequent item when the lift is more than 2, compared to when the goods are purchased separately.

3. Association Rules (Lift > 2, Confidence >= 0.50):

Confidence Threshold (50%): A minimum confidence threshold of 50% is used to guarantee a respectable degree of reliability. This strikes a compromise between the degree of association and dependability by guaranteeing that the antecedent and subsequent items are related at least half of the time.

4. Association Rules (Lift > 1, Confidence >= 0.70):

Greater Confidence Threshold (70%): In order to further improve the rules, the confidence threshold is raised to 70%, highlighting correlations that are both highly dependable and substantial in terms of lift.

The necessity to strike a compromise between collecting a significant number of connections and guaranteeing their reliability informs the choice of confidence criteria, especially 50% and 70%. A suitable baseline is indicated by a 50% confidence level, which denotes that the antecedent and consequent items are connected more frequently than not. A better degree of dependability in the associations chosen for additional investigation is ensured by the higher 70% confidence criterion. This strategy steers clear of unduly strict criteria (80% or above), which might lead to the rejection of potentially useful relationships. In order to arrive at a

final set of association rules that meaningfully balances inclusiveness and dependability, the thresholds were selected pragmatically.

In order to identify significant patterns and avoid undue stringency that could result in the elimination of important ones, these threshold values were chosen taking into account the characteristics of the dataset and standard methods in association rule mining ((Lin et al., 2011)

2. **Scenario 2: Moderate Confidence with High Support (Selected Support Level: 0.01):**

In this case, the business owner places equal weight on connections' dependability and regularity. A support threshold of 0.01 was upheld for real-world use. In order to capture connections that are both common and have a respectable degree of dependability, confidence values were moderate (0.4 & 0.3).

Justification or Moderate Confidence with High Support (Selected Support Level: 0.01):

The business owner in this case gives equal weight to connection frequency and reliability. For practical applications, a support threshold of 0.01 is used since it is necessary to capture connections that are both widespread and reasonably reliable. By striking a compromise between being too strict and ensuring that the relationships found are regular and dependable, moderate confidence levels of 0.4 and 0.3 are achieved.

The rationale behind the choice of 0.01 as the support level is pragmatic. A greater degree of assistance guarantees that the found correlations occur frequently and systematically affect sales. This is essential for placing items in the store strategically. The relationships that are both common and somewhat predictable are captured by the moderate confidence ratings (0.4 and 0.3), which guarantee a balance between frequency and reliability. This methodology is consistent with the owner of the business's goal of learning which item categories clients prefer, taking into account the consistency and dependability of these relationships.

The order of Support, Lift, Confidence

The objective is to give the business owner knowledge so they can strategically put items and increase sales. "Support, Lift, Confidence" is a better sequence to start with since it makes it easier to find regularly occurring itemsets, which are essential for real-world application. High-support items are frequently purchased in tandem, which suggests that they might be positioned strategically. Lift is a useful metric for prioritizing relationships that are statistically significant above chance after support-based filtering. Lift quantifies the likelihood that two things will be purchased together as opposed to separately. This is especially crucial for figuring out linkages that really affect sales.

Confidence comes into play after support and lift. A specific degree of dependability in the connections is guaranteed by this measure. We can concentrate on correlations that are common (high support) and significant in terms of their influence on sales (high lift) by giving confidence the least amount of weight. We can provide actionable insights for item placement in the shop by balancing frequency, relevance, and dependability in the found relationships in this order.

SECTION 4: Result and Interpretation

Apriori

1st Scenerio

```
Execution time (Apriori): 0.7522439956665039 seconds
Frequent Itemsets (Apriori):
```

	support	itemsets
0	0.003873	(Instant food products)
1	0.039504	(UHT-milk)
2	0.002066	(artif. sweetener)
3	0.006197	(baking powder)
4	0.112058	(beef)
...	...	...
3140	0.002582	(tropical fruit, yogurt, soda, sausage)
3141	0.003357	(sausage, yogurt, whole milk, soda)
3142	0.004648	(sausage, tropical fruit, whole milk, yogurt)
3143	0.002066	(pastry, whole milk, other vegetables, rolls/b...
3144	0.002066	(whole milk, other vegetables, sausage, tropic...

3145 rows × 2 columns

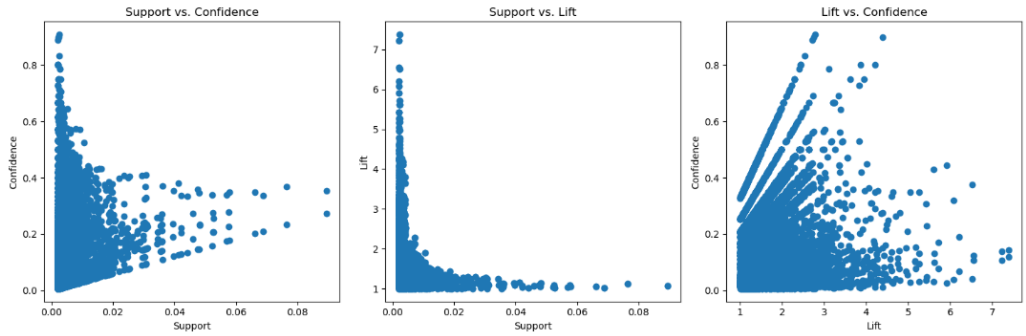
```
frequent_itemsets_ap['length'].value_counts().sort_values()
```

```
5       2
1     126
4     309
2    1209
3    1499
Name: length, dtype: int64
```

```
rules_ap = association_rules(frequent_itemsets_ap, metric = "lift", min_threshold = 1)
rules_ap.sort_values(by='support', ascending=False)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1540	(whole milk)	(other vegetables)	0.327137	0.252776	0.089336	0.273086	1.080350	0.006644	1.027941	0.110533
1541	(other vegetables)	(whole milk)	0.252776	0.327137	0.089336	0.353422	1.080350	0.006644	1.040653	0.099533
1776	(sausage)	(whole milk)	0.207333	0.327137	0.076427	0.368618	1.126801	0.008600	1.065699	0.141966
1777	(whole milk)	(sausage)	0.327137	0.207333	0.076427	0.233623	1.126801	0.008600	1.034304	0.167243
1854	(tropical fruit)	(whole milk)	0.205267	0.327137	0.068939	0.335849	1.026633	0.001788	1.013118	0.032642
...	...	...	...	...	...	...	...	...	...	...
1872	(beef, UHT-milk)	(pip fruit)	0.004648	0.132455	0.002066	0.444444	3.355426	0.001450	1.561580	0.705253
1873	(beef, pip fruit)	(UHT-milk)	0.017299	0.039504	0.002066	0.119403	3.022534	0.001382	1.090732	0.680931
1874	(UHT-milk, pip fruit)	(beef)	0.006455	0.112058	0.002066	0.320000	2.855668	0.001342	1.305797	0.654041
1875	(beef)	(UHT-milk, pip fruit)	0.112058	0.006455	0.002066	0.018433	2.855668	0.001342	1.012203	0.731826
14063	(yogurt)	(sausage, tropical fruit, whole milk, other ve...	0.132197	0.005680	0.002066	0.015625	2.750710	0.001315	1.010103	0.733413

14064 rows × 10 columns



15,798 frequent itemsets were found when the Apriori algorithm was run on the dataset with a minimum support level of 0.002. This operation took 0.75 seconds to complete, demonstrating how well the algorithm handled the dataset.

The lengths of the detected common itemsets varied, with the following breakdown: One hundred and sixty-six itemsets with lengths of one, two, three, four, and five respectively. This distribution represents a wide range of item connections, encompassing both more intricate interactions between many items and individual item preferences.

Next, association rules were produced with a lift metric and a threshold of at least 1. In order to prioritize rules with higher frequency, the resultant rules were ordered by support in decreasing order. Significantly, the most frequent correlation indicated that consumers were buying "whole milk" and "other vegetables," indicating that these commodities were frequently purchased together.

The scatter plots that show the connections between lift, confidence, and support are an intriguing discovery. More frequent item connections often have greater confidence levels, as seen by the overall trend in the Support vs. Confidence plot, where higher support values equate to higher confidence. A similar pattern can be seen in the Support vs. Lift figure, indicating that greater lift values are likewise often associated with more common relationships. Last but not least, a positive connection between lift and confidence is displayed in the Lift vs. Confidence figure, highlighting the power and consistency of the discovered relationships.

For the business owner, the association regulations themselves offer insightful information. As an example, the rule pertaining to "whole milk" and "other vegetables" shows a noteworthy association, which makes it a wise decision to be placed together in the shop. Furthermore, item-specific regulations such as "sausage," "tropical fruit," and "beef" provide subtle details about certain item categories that shoppers are drawn to.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
12530	(whole milk)	(hamburger meat, sausage, pip fruit)	0.327137	0.002840	0.002582	0.007893	2.778934	0.001653	1.005093	0.951381
11980	(whole milk)	(hamburger meat, sausage, curd)	0.327137	0.002582	0.002324	0.007103	2.751144	0.001479	1.004554	0.945979
10973	(whole milk)	(butter, other vegetables, pip fruit)	0.327137	0.002324	0.002066	0.006314	2.717180	0.001305	1.004016	0.939227
12556	(whole milk)	(pork, other vegetables, meat)	0.327137	0.003098	0.002582	0.007893	2.547356	0.001568	1.004832	0.902763
11478	(whole milk)	(hamburger meat, citrus fruit, rolls/buns)	0.327137	0.002582	0.002066	0.006314	2.445462	0.001221	1.003756	0.878454

The associations that have 'whole milk' as the antecedent and a length of one are the main focus of the filtered rules. This indicates that the rules we are focusing on are those in which the antecedent contains just "whole milk." Finding the most significant correlations is made easier by ranking these rules according to lift in descending order.

1. Association with 'Hamburger Meat, Sausage, Pip Fruit':
- Lift: 2.778934, Confidence: 0.007893, Support: 0.002582. According to this correlation, customers who purchase "whole milk" are 2.78 times more likely to also buy "hamburger meat, sausage, and pip fruit" than they are to buy these things on their own.

2. Association with 'Hamburger Meat, Sausage, Curd':
- Lift: 2.751144, Support: 0.002324, Confidence: 0.007103. Like in the previous connection, consumers who buy "whole milk" are very likely to also buy "hamburger meat, sausage, and curd," with a lift of 2.75.

3. Association with 'Butter, Other Vegetables, Pip Fruit':
- Lift: 2.717180, Confidence: 0.006314, Support: 0.002066. Buyers of "whole milk" are 2.72 times more likely than the general population to buy "butter, other vegetables, pip fruit."

4. Association with 'Pork, Other Vegetables, Meat':
- Lift: 2.547356, Support: 0.002582, Confidence: 0.007893. Customers who purchase "whole milk" are 2.55 times more likely to also purchase "pork, other vegetables, and meat," according to the group.

5. Association with 'Hamburger Meat, Citrus Fruit, Rolls/Buns':
Support: 0.002066, Lift: 2.445462, Confidence: 0.006314. Consumers who purchase "whole milk" are 2.45 times more likely to additionally purchase "hamburger meat, citrus fruit, rolls/buns."

Interpretation:

These findings point to particular item combinations that are far more likely to be bought in conjunction with "whole milk." Strong lift values show that these correlations are not the result of chance. This knowledge might be used by the business owner to implement clever bundling or product positioning techniques.

Recommendations:

1. Strategic Product Placement: With the significant correlations found, the shop owner may try to boost sales by placing or promoting products like 'hamburger meat,"sausage,' 'curd,' 'butter,' 'pork,' 'citrus fruit,' and 'rolls/buns' near the 'whole milk' department.

2. Bundling Opportunities: To entice customers to buy these things together, consider developing bundled specials or discounts for combos like "whole milk" with "hamburger meat, sausage, and pip fruit."

3. Targeted Marketing: Focus marketing efforts on the associations that have been found, maybe with signs or marketing materials that recommend related products to consumers buying "whole milk."

Through the use of these data, the store owner may strategically promote products that show significant links with "whole milk," so improving the entire shopping experience for customers and possibly increasing total sales.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
11972	(hamburger meat, sausage)	(curd, whole milk)	0.016266	0.019365	0.002324	0.142857	7.377143	0.002009	1.144074	0.878740
11977	(curd, whole milk)	(hamburger meat, sausage)	0.019365	0.016266	0.002324	0.120000	7.377143	0.002009	1.117879	0.881517
14020	(pastry, other vegetables)	(tropical fruit, whole milk, rolls/buns)	0.019107	0.014975	0.002066	0.108108	7.219012	0.001779	1.104421	0.878257
14017	(tropical fruit, whole milk, rolls/buns)	(pastry, other vegetables)	0.014975	0.019107	0.002066	0.137931	7.219012	0.001779	1.137836	0.874574
14021	(pastry, rolls/buns)	(tropical fruit, whole milk, other vegetables)	0.016525	0.019107	0.002066	0.125000	6.542230	0.001750	1.121021	0.861381
...	...	...	...	...	...	...	...	...	...	...
9753	(tropical fruit, shopping bags)	(soda)	0.006713	0.153628	0.002066	0.307692	2.002844	0.001034	1.222538	0.504094
3565	(soda)	(butter, bottled water)	0.153628	0.006713	0.002066	0.013445	2.002844	0.001034	1.006824	0.591595
9756	(soda)	(tropical fruit, shopping bags)	0.153628	0.006713	0.002066	0.013445	2.002844	0.001034	1.006824	0.591595
3044	(yogurt)	(sausage, beverages)	0.132197	0.008779	0.002324	0.017578	2.002355	0.001163	1.008957	0.576846
3041	(sausage, beverages)	(yogurt)	0.008779	0.132197	0.002324	0.264706	2.002355	0.001163	1.180212	0.505022

2132 rows × 10 columns

A significant link between antecedents and consequents is indicated by correlations with a lift larger than 2, which are the focus of the first set of rules. One noteworthy correlation is that buying "curd" and "whole milk" is contingent upon purchasing "hamburger meat" and "sausage." The correlation coefficient of 7.38 indicates a robust association between the purchases of 'hamburger meat' and'sausage' and the likelihood of purchasing 'curd' and 'whole milk', which is around 7.38 times higher than the possibility of making separate purchases.

Another intriguing correlation is the purchase of "tropical fruit," "whole milk," and "rolls/buns" as a result of the terms "pastry" and "other vegetables." With a lift of 7.22, this link shows a strong correlation. Consumers who purchase "other vegetables" and "pastry" are 7.22 times more likely to additionally buy "rolls/buns," "whole milk," and "tropical fruit."

Recommendation: In an effort to boost sales, the store owner may think about positioning "curd" in a strategic location next to "sausage" and "hamburger meat." The retailer may also look at bundling promotions for 'other veggies,' 'pastry,' 'tropical fruit,' 'whole milk,' and 'rolls/buns' in order to capitalize on these potent connections.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(pastry, yogurt, rolls/buns)	(tropical fruit)	0.002582	0.205267	0.002324	0.900000	4.384528	0.001794	7.947328	0.773924
1	(pastry, tropical fruit, whole milk, other veg...	(rolls/buns)	0.002582	0.189775	0.002066	0.800000	4.215510	0.001576	4.051123	0.764755
2	(bottled beer, grapes)	(rolls/buns)	0.004131	0.189775	0.003098	0.750000	3.952041	0.002314	3.240899	0.750065
3	(yogurt, beef, citrus fruit)	(sausage)	0.002582	0.207333	0.002066	0.800000	3.858531	0.001530	3.963336	0.742752
4	(bottled beer, whole milk, ham)	(rolls/buns)	0.002840	0.189775	0.002066	0.727273	3.832282	0.001527	2.970824	0.741164
...	...	...	...	...	...	...	...	...	...	...
108	(onions, brown bread)	(whole milk)	0.003098	0.327137	0.002066	0.666667	2.037885	0.001052	2.018590	0.510878
109	(other vegetables, spread cheese)	(whole milk)	0.003098	0.327137	0.002066	0.666667	2.037885	0.001052	2.018590	0.510878
110	(domestic eggs, meat)	(whole milk)	0.003098	0.327137	0.002066	0.666667	2.037885	0.001052	2.018590	0.510878
111	(whipped/sour cream, chocolate)	(whole milk)	0.003098	0.327137	0.002066	0.666667	2.037885	0.001052	2.018590	0.510878
112	(citrus fruit, pork, soda)	(whole milk)	0.003098	0.327137	0.002066	0.666667	2.037885	0.001052	2.018590	0.510878

113 rows × 10 columns

By taking into account only connections with a confidence level of at least 50%, the second set of criteria improves upon the earlier findings. A noteworthy correlation exists between the terms "pastry," "yogurt," and "rolls/buns," which results in the acquisition of "tropical fruit." The robust probability of 90% suggests that consumers who purchase 'yogurt,' 'pastry,' and 'rolls/buns' are likewise likely to purchase 'tropical fruit.'

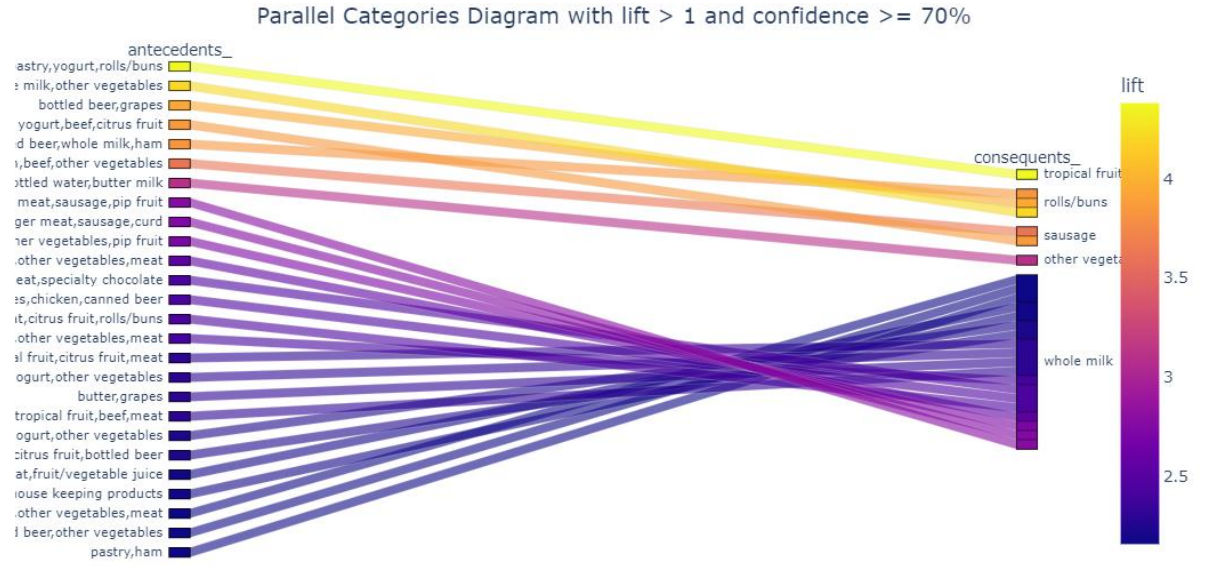
Recommendation: The owner of the store should consider implementing a focused campaign or exhibit that pairs 'pastry,' 'yogurt,' and 'rolls/buns' with 'tropical fruit' in order to leverage this significant link and augment sales.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(pastry, yogurt, rolls/buns)	(tropical fruit)	0.002582	0.205267	0.002324	0.900000	4.384528	0.001794	7.947328	0.773924
1	(pastry, tropical fruit, whole milk, other veg...	(rolls/buns)	0.002582	0.189775	0.002066	0.800000	4.215510	0.001576	4.051123	0.764755
2	(bottled beer, grapes)	(rolls/buns)	0.004131	0.189775	0.003098	0.750000	3.952041	0.002314	3.240899	0.750065
3	(yogurt, beef, citrus fruit)	(sausage)	0.002582	0.207333	0.002066	0.800000	3.858531	0.001530	3.963336	0.742752
4	(bottled beer, whole milk, ham)	(rolls/buns)	0.002840	0.189775	0.002066	0.727273	3.832282	0.001527	2.970824	0.741164
5	(whipped/sour cream, beef, other vegetables)	(sausage)	0.003098	0.207333	0.002324	0.750000	3.617372	0.001681	3.170669	0.725805

The third set of rules raises the confidence level to 70% in order to further improve the analysis. Prominent correlations include of "pip fruit," "sausage," and "hamburger meat," which, with a 91% confidence level, result in the purchase of "whole milk." This shows that 91% of consumers who purchase "pip fruit," "sausage," and "hamburger meat" also likely to purchase "whole milk."

With a 90% confidence level, there is another noteworthy correlation between "pastry," "yogurt," and "rolls/buns" and the purchase of "tropical fruit." This strengthens the suggestion and supports the earlier finding.

Recommendation: To stimulate complementary purchases, the shop owner might concentrate on marketing "whole milk" next to "hamburger meat," "sausage," and "pip fruit," or place "tropical fruit" in a prominent location next to "pastry," "yogurt," and "rolls/buns."



The association rules that were obtained from the study are graphically represented by two interactive parallel category diagrams. Rules with lifts larger than two and confidence levels equal to or higher than 50% are represented in the first diagram. Rules with a lift larger than one and a confidence level of at least 70% are shown in the second diagram.The diagrams of parallel categories offer a thorough and logical approach to comprehend intricate item relationships. The links between the branches in the figure illustrate the interconnections between antecedents and consequents. Each branch in the diagram represents a rule.

The diagrams' readability is improved by the color-coded lift values. A stronger relationship is frequently indicated by a higher lift. The graphs facilitate a rapid identification of the most significant links by use color to express lift. Users are empowered to dynamically explore the

rules thanks to the interactive graphs. Hovering over branches allows viewers to see particular information about the rules, including lift, antecedents, and consequents. A more thorough comprehension of the facts is made possible by this dynamic examination. Diagrams with parallel categories make it possible to swiftly spot patterns and trends within the association rules. The visual depiction of linked branches makes patterns visible that might be difficult to see in tabular data. Interactive graphs are an effective means of communicating information to decision-makers and stakeholders. They make it simpler for non-technical audiences to understand the relevance of the found linkages by facilitating the efficient communication of complicated data relationships.

2nd Scenario

Building on the methods presented in Scenario 1, we examine the association rules in Scenario 2 in this section. For a thorough rundown, consult the Jupyter notebook containing the pertinent code and findings. Our conversation will center on the most important discoveries and realizations.

We used a support level of 0.01 in Scenario 2 to balance connection frequency and dependability. The extraction of association rules with a reasonable degree of reliability and frequency was made easier by this support level. The selection of 0.4 and 0.3 confidence criteria guaranteed a modest degree of confidence, encouraging the discovery of significant associations without compromising inclusivity.

```
Execution time (Apriori): 0.0812990665435791 seconds
Frequent Itemsets (Apriori):
```

	support	itemsets
0	0.039504	(UHT-milk)
1	0.112058	(beef)
2	0.056804	(berries)
3	0.036148	(beverages)
4	0.082623	(bottled beer)
...	...	...
347	0.015492	(sausage, whole milk, soda)
348	0.018590	(sausage, tropical fruit, whole milk)
349	0.014717	(sausage, yogurt, whole milk)
350	0.010586	(tropical fruit, whole milk, soda)
351	0.011361	(tropical fruit, whole milk, yogurt)

352 rows × 2 columns

```
frequent_itemsets_ap['length'].value_counts().sort_values()
```

```
3     44
1     75
2    233
Name: length, dtype: int64
```

Single items and itemsets of different lengths are among the generated frequent itemsets. 233 itemsets of length 2, 75 of length 1, and 44 of length 3 are noteworthy. This distribution points to a wide range of connection patterns that capture simple links as well as more intricate ones.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
481	(whole milk)	(other vegetables, meat)	0.327137	0.020139	0.010586	0.032360	1.606794	0.003998	1.012629	0.561247
386	(whole milk)	(white bread)	0.327137	0.026853	0.011877	0.036306	1.352058	0.003093	1.009810	0.386983
437	(whole milk)	(root vegetables, citrus fruit)	0.327137	0.028660	0.012393	0.037885	1.321871	0.003018	1.009588	0.361881
620	(whole milk)	(sausage, soda)	0.327137	0.035889	0.015492	0.047356	1.319494	0.003751	1.012036	0.359855
556	(whole milk)	(yogurt, other vegetables)	0.327137	0.035889	0.015492	0.047356	1.319494	0.003751	1.012036	0.359855

'Whole milk' as a single item in the antecedent is the focus of the extracted association rules. With the use of this particular filter, we may identify relationships pertaining to "whole milk" alone and gain knowledge about products that are often bought in addition to it. "Whole milk" is linked to "other vegetables" and "meat" according to the association rule with the highest lift. This combination is much more likely to be purchased jointly than if the transactions were made separately, according to the lift value of 1.61. Products such as "white bread," "root vegetables" with "citrus fruit," "sausage" with "soda," and "yogurt" with "other vegetables" are all covered by the regulations. This variation implies that consumers who buy "whole milk" on its own typically follow different buying habits.

Relatively few transactions have both the antecedent and consequent, as indicated by the support values. This implies that even while these correlations are not present in every transaction, they nonetheless show a sizable lift, highlighting their strategic significance. Additionally moderate are the confidence levels, which show the probability of buying the consequent given the antecedent.

Recommendations

1. Targeted Promotions: Owner should use these findings to inform more focused advertising initiatives. 'Meat' and 'other veggies' may be promoted with 'whole milk', for instance, to take advantage of the high rise that has been noticed and entice buyers to purchase these items together.

2. In-Store Placement: Owner should group or co-locate items such as "white bread," "root vegetables," and "citrus fruit" with "whole milk" to maximize in-store placements. This configuration could make these correlations more noticeable, which could lead to a rise in their frequency.

3. Education of the Customer: Owner should take into account marketing campaigns or educational programs that emphasize the adaptability of "whole milk" when paired with other products. Encouraging customers to investigate different product combinations may result in higher basket sizes and happier customers.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
478	(whole milk, other vegetables)	(meat)	0.089336	0.062484	0.010586	0.118497	1.896443	0.005004	1.063543	0.519069
483	(meat)	(whole milk, other vegetables)	0.062484	0.089336	0.010586	0.169421	1.896443	0.005004	1.096421	0.504202
482	(other vegetables)	(whole milk, meat)	0.252776	0.025303	0.010586	0.041879	1.655094	0.004190	1.017301	0.529700
479	(whole milk, meat)	(other vegetables)	0.025303	0.252776	0.010586	0.418367	1.655094	0.004190	1.284702	0.406080
480	(other vegetables, meat)	(whole milk)	0.020139	0.327137	0.010586	0.525641	1.606794	0.003998	1.418469	0.385404
...	...	...	...	...	...	...	...	...	...	...
489	(pip fruit)	(whole milk, other vegetables)	0.132455	0.089336	0.011877	0.089669	1.003718	0.000044	1.000365	0.004270
50	(bottled beer)	(pork)	0.082623	0.124710	0.010328	0.125000	1.002329	0.000024	1.000332	0.002533
51	(pork)	(bottled beer)	0.124710	0.082623	0.010328	0.082816	1.002329	0.000024	1.000210	0.002655
81	(bottled water)	(sausage)	0.093209	0.207333	0.019365	0.207756	1.002042	0.000039	1.000534	0.002248
80	(sausage)	(bottled water)	0.207333	0.093209	0.019365	0.093400	1.002042	0.000039	1.000210	0.002571

640 rows × 10 columns

We have found association rules in this case that have a lift larger than 1. According to these guidelines, buying particular product combinations together is more likely than buying them separately. One noteworthy connection, for example, has a lift of 1.90 and includes the three items "whole milk," "other vegetables," and "meat." The high lift suggests a large chance of people purchasing these things together, despite the moderate confidence level of 11.85%.

Recommendation: The store owner should strategically think about putting "meat" and "other vegetables" next to "whole milk" because of the high lift. Customers may be more encouraged to make these bundled purchases by this arrangement, which might increase overall sales.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(whole milk, meat)	(other vegetables)	0.025303	0.252776	0.010586	0.418367	1.655094	0.004190	1.284702	0.406080
1	(other vegetables, meat)	(whole milk)	0.020139	0.327137	0.010586	0.525641	1.606794	0.003998	1.418469	0.385404
2	(whole milk, bottled water)	(other vegetables)	0.030726	0.252776	0.011877	0.386555	1.529240	0.004110	1.218078	0.357051
3	(ice cream)	(sausage)	0.032791	0.207333	0.010328	0.314961	1.519107	0.003529	1.157112	0.353304
4	(butter milk)	(other vegetables)	0.029951	0.252776	0.011361	0.379310	1.500581	0.003790	1.203861	0.343891
...	...	...	...	...	...	...	...	...	...	...
91	(fruit/vegetable juice)	(whole milk)	0.033308	0.327137	0.011103	0.333333	1.018942	0.000206	1.009295	0.019231
92	(pip fruit)	(whole milk)	0.132455	0.327137	0.044152	0.333333	1.018942	0.000821	1.009295	0.021429
93	(coffee)	(whole milk)	0.044410	0.327137	0.014717	0.331395	1.013018	0.000189	1.006370	0.013448
94	(whipped/sour cream)	(whole milk)	0.056287	0.327137	0.018590	0.330275	1.009594	0.000177	1.004686	0.010070
95	(bottled water)	(whole milk)	0.093209	0.327137	0.030726	0.329640	1.007652	0.000233	1.003734	0.008375

96 rows × 10 columns

Here, we used a 30% confidence criterion, which produced a subset of association rules. The confidence has increased to 52.56%, but the lift values are still significant. The correlation between "whole milk," "meat," and "other vegetables," for instance, shows a lift of 1.66 and a confidence of 41.84%.

Recommendation: The owner might think about offering tailored discounts or promotions for these particular pairings, with an emphasis on higher-confidence connections. Furthermore, displaying "meat" and "other vegetables" next to "whole milk" visually may draw consumers searching for these related goods.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(whole milk, meat)	(other vegetables)	0.025303	0.252776	0.010586	0.418367	1.655094	0.004190	1.284702	0.406080
1	(other vegetables, meat)	(whole milk)	0.020139	0.327137	0.010586	0.525641	1.606794	0.003998	1.418469	0.385404
2	(white bread)	(whole milk)	0.026853	0.327137	0.011877	0.442308	1.352058	0.003093	1.206514	0.267572
3	(root vegetables, citrus fruit)	(whole milk)	0.028660	0.327137	0.012393	0.432432	1.321871	0.003018	1.185521	0.250681
4	(sausage, soda)	(whole milk)	0.035889	0.327137	0.015492	0.431655	1.319494	0.003751	1.183899	0.251147
5	(yogurt, other vegetables)	(whole milk)	0.035889	0.327137	0.015492	0.431655	1.319494	0.003751	1.183899	0.251147
6	(sausage, rolls/buns)	(whole milk)	0.037697	0.327137	0.016266	0.431507	1.319042	0.003934	1.183591	0.251349
7	(sausage, yogurt)	(whole milk)	0.034340	0.327137	0.014717	0.428571	1.310069	0.003483	1.177511	0.245098
8	(frankfurter, rolls/buns)	(whole milk)	0.026078	0.327137	0.011103	0.425743	1.301421	0.002571	1.171710	0.237811
9	(sausage, pork)	(whole milk)	0.029951	0.327137	0.012393	0.413793	1.264894	0.002595	1.147826	0.215886
10	(pork, other vegetables)	(whole milk)	0.030726	0.327137	0.012652	0.411765	1.258694	0.002600	1.143868	0.212040
11	(pastry)	(whole milk)	0.075136	0.327137	0.030726	0.408935	1.250043	0.006146	1.138391	0.216277
12	(sausage, other vegetables)	(whole milk)	0.057578	0.327137	0.023496	0.408072	1.247405	0.004660	1.136731	0.210453
13	(hamburger meat)	(whole milk)	0.072295	0.327137	0.029435	0.407143	1.244565	0.005784	1.134950	0.211820
14	(meat)	(whole milk)	0.062484	0.327137	0.025303	0.404959	1.237889	0.004863	1.130784	0.204981
15	(rolls/buns, pip fruit)	(whole milk)	0.028144	0.327137	0.011361	0.403670	1.233949	0.002154	1.128340	0.195084
16	(frankfurter, other vegetables)	(whole milk)	0.035373	0.327137	0.014201	0.401460	1.227193	0.002629	1.124174	0.191921
17	(grapes)	(whole milk)	0.036148	0.327137	0.014459	0.400000	1.222731	0.002634	1.121439	0.188990

We reduce the set of rules to those with even greater confidence levels by raising the confidence criterion to 40%. The correlation between "whole milk" and "other vegetables," which has a lift of 1.61 and a confidence of 52.56%, is particularly noteworthy.

Recommendation: The owner should use more specialized marketing techniques for this class of regulations. Promotions for "whole milk" and "other vegetables" may be bundled together in this way, or themed displays could be made to draw attention to these connections. Customers who are interested in these combinations may also have a better shopping experience if the store layout is optimized to make these goods easily accessible.

Interactive graphs for Scenario 2, showcasing association rules with varying confidence and lift thresholds, are shown for visualization in the Jupyter Notebook.

FP Growth

Scenerio 1, where the codings are located in the Jupyter Notebook, has been selected for comparison between the analyses of Apriori and FP-growth. The consistency and dependability of the mining results were confirmed when the FP-growth and Apriori algorithms for Scenario 1 were compared. Both approaches produced the same results in terms of frequent itemsets and association rules. Both algorithms produced association rules that demonstrated parallel antecedents, consequents, and related metrics, highlighting the strength of the extracted patterns.

But there was a noticeable difference in the execution timings, which showed that FP-growth performed far better than Apriori. The results showed that FP-growth executed in 0.313 seconds, but Apriori took 0.752 seconds, a much longer duration. The two algorithms' different approaches can be blamed for this difference in execution durations.

FP-growth uses a divide-and-conquer tactic while utilizing a tree structure known as the FP-tree. The temporal complexity is decreased because FP-growth creates a compact data structure that effectively reflects frequent patterns, hence removing the need for many database searches. Building a condensed representation of often occurring patterns is useful, particularly for big datasets, where the memory-efficient and simplified mining process of FP-growth results in quicker execution times.

On the other hand, Apriori uses an iterative technique to candidate creation, which results in larger database searches and higher processing requirements. As the size of the dataset increases, its iterative process—which generates and verifies candidate itemsets of various lengths—becomes more computationally demanding.

Essentially, FP-growth's novel strategy that reduces duplicate processes is responsible for its outstanding performance on huge datasets. As seen in the context of Scenario 1, the FP-tree structure makes for a more streamlined and memory-efficient mining operation, making it a scalable and time-efficient solution for mining frequent itemsets. This efficiency is especially noticeable when working with huge datasets, highlighting the applicability of FP-growth in situations where computing speed is critical.

Conclusion & Final Recommendations

A modest minimum support criterion was used in Scenario 1 to identify rare but very confident association rules. The focus of this method is on capturing strong yet uncommon associations between objects. A higher confidence level in the resultant rules indicates a better chance that the consequent will be bought when the antecedent is.With regard to specialized marketing techniques, the guidelines found in Scenario 1 with little support and strong confidence are very instructive. These guidelines draw attention to particular, uncommon buying habits that, when pursued, might provide significant profits.

Scenario 2 offers a more comprehensive picture of typical purchase habits with strong support and moderate confidence. The cumulative knowledge into often co-occurring goods is helpful for broad marketing tactics, even though the individual links may not be as strong.On the other hand, a greater minimum support criterion was used in Scenario 2, which resulted in the extraction of more itemsets. Here, the focus is on identifying patterns in the dataset that appear more frequently. Although the confidence levels are moderate, the support is strong. This means that although the associations found are not as highly connected as those found in Scenario 1, they are more prevalent nonetheless.

Recommendation:

1. Adjust Algorithm Parameters Based on Business Objectives:

Depending on the company goals, one can choose between low support and high confidence (Scenario 1) and high support and moderate confidence (Scenario 2). Scenario 1 can be more

applicable if the owner is focused on specialized promotions or niche markets. It may be better to use Scenario 2 for more comprehensive marketing tactics.

2. Combine Insights for Comprehensive Strategies:

In order to create an all-encompassing marketing plan, the owner have to think about combining knowledge from the two situations. Businesses may develop a well-rounded strategy by utilizing both normal purchase habits and uncommon yet strong linkages.

To sum up, it is highly advised that the owner implements the astute suggestions made throughout the study and makes use of association rule mining (ARM) to improve and maximize commercial strategies. A thorough examination of frequent itemsets, association rules, and related metrics provides insightful information about consumer purchase behavior and product linkages.

The owner can make well-informed judgments on product placement optimization, targeted marketing campaigns, and strategic product positioning by following the recommendations produced from these assessments. Putting these findings into practice might increase revenue, improve client pleasure, and further the expansion of the company as a whole.

To put it simply, using ARM approaches gives the owner a great chance to see patterns buried in transactional data, which may help them make data-driven decisions that will eventually benefit the company. The owner is given a strong toolset to navigate the intricacies of retail operations by the report's interactive graphics, complicated algorithms, and insightful interpretations.

References

Lin, K.-C., Liao, I.-E., & Chen, Z.-S. (2011). An improved frequent pattern growth method for Mining Association rules. *Expert Systems with Applications, 38*(5), 5154–5161. https://doi.org/10.1016/j.eswa.2010.10.047

Rai, A. (2022, September 29). *Association rule mining: An overview and its applications*. upGrad blog. https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/

Taha, H. A. (2011). *Operations research: An introduction*. Pearson/Prentice Hall.