Developing customised product recommendations for coffee shop customers post-segmentation using Apriori modelling.

Estera Wisnieweska

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

February 2024

Supervisor: Taufique Ahmed

# Introduction

## Domain background

Coffee cultivation boasts a rich history, tracing its origins to the ancient coffee forests of Ethiopia around the 9th century. According to legend, Kaldi, a goat herder, stumbled upon the potential of coffee when his goats exhibited heightened energy after consuming berries from a specific tree (Lewis, 2021). This serendipitous discovery led Kaldi to share his findings with a local monastery abbot, giving rise to the creation of a beverage that proved invaluable for enhancing alertness during evening prayers. The dissemination of this knowledge marked the beginning of coffee's journey across the Arabian peninsula, eventually establishing itself as a globally cherished beverage.

By the 17th century, coffee had found its way into Europe, gaining widespread popularity across the continent. Over time, coffee evolved into a globally cherished and irreplaceable part of daily life. The recent surge in global coffee production during the 2020/21 period is a testament to its enduring popularity. The global production crossed a significant milestone, surpassing 175 million bags, each weighing 60 kilograms. South America, led by the coffee powerhouse Brazil, played a pivotal role in this achievement, contributing more than half of the global coffee output.

Notably, while Brazil led the world in coffee production, the United States emerged as the leader in coffee market revenue, generating a substantial 85 billion U.S. dollars compared to Brazil's 35 billion U.S. dollars (Ridder, 2023). This economic shift underscores the global significance of coffee beyond its agricultural roots.

The dynamics of the global coffee trade further revealed a robust demand, with major importers in 2021 including the United States, Germany, France, and Italy. Key exporting nations, such as Brazil, Switzerland, and Colombia, played crucial roles in meeting this demand. This surge in international coffee trade aligns with the growing emphasis on Fairtrade and sustainability within the industry. Notably, the increase in UTZ certified production exceeded 1.2 million metric tons in 2021, signalling a collective commitment to ethical and sustainable coffee practices.

Coffee comes in a rich array of styles like espresso, cappuccino, and americano, thanks to a diverse range of beans, with Arabica and Robusta taking the lead. Beyond being a caffeine kick, coffee is gaining acclaim for potential health perks, from reducing the risk of liver cancer to safeguarding against Parkinson’s disease (Hu, 2021) and promoting heart health, coffee's multifaceted advantages are gaining appreciation.

All these diverse factors play a pivotal role, in influencing customers' choices. Comprehending these complex dynamics is crucial for accurate customer segmentation, as it forms the foundation for tailored product recommendations.

## Research Problem

The primary focus of this research is to develop personalized product recommendations for diverse customer segments in the context of coffee shops. The research endeavours to employ customer segmentation techniques and market basket analysis, specifically utilizing the Apriori algorithm. Despite existing research on customer segmentation and market analysis for various industries, a noticeable research gap exists in the domain of coffee shop sales points, particularly at the point of purchase (till). While prior studies have explored customer grouping through clustering methods, none have applied market basket analysis post-segmentation to coffee shop transactions. This research aims to fill this gap by providing insights into customer behaviour and preferences at the till, ultimately contributing to enhanced product recommendations tailored for different customer groups within coffee shops.

## Research gap

Recommendation systems have become commonplace in various fields, playing a pivotal role in enhancing user experiences. In scholarly discourse, these systems have been extensively explored in diverse domains, from e-commerce to streaming services. The prevalent methodologies involve customer segmentation and subsequent market basket analysis. Clustering methods are often employed to group customers based on similarities, providing insights into their preferences. However, in the distinctive realm of coffee shop retailing, there is a noticeable gap in existing research.

Most existing studies primarily concentrate on broader customer grouping, neglecting the immediate and unique dynamics of transactions occurring at the point of sale in coffee shops. Unlike online retail environments where customers have more time for consideration, coffee shop sales involve quick, on-the-spot decisions. This immediacy in decision-making sets coffee shop transactions apart and requires a specialized approach for effective recommendation systems. The identified gap in the literature underscores the need for research specifically tailored to the nuances of coffee shop retailing.

To address this gap, I am employing a methodology that involves segmenting customers based on their gender and age groups. Subsequently, I am applying the Apriori algorithm to conduct a market basket analysis for each specific customer segment. This approach aims to uncover patterns and associations in customer transactions at coffee shop sales points, shedding light on preferences and potential cross-product relationships within distinct demographic categories.

## Research question

*Can post-segmentation market basket analysis, based on gender and age groups, enhance the accuracy and suitability of product recommendations for coffee shop customers?*

The exploration into the potential enhancement of product recommendation accuracy for coffee shop customers post-segmentation commenced with a thorough literature review. This phase involved an exhaustive examination of methodologies employed across diverse industries for customer segmentation, market basket analysis, and product recommendations. Following this insightful groundwork, the methodology section unfolds the intricacies of how customers were segmented. The Apriori algorithm is then applied to delve into post-segmentation market basket analysis, searching for potential patterns within the diverse groups.

The subsequent chapter, the discussion, will thoroughly explore and interpret the results. It will carefully connect the findings to the existing body of knowledge, providing insights derived from the research. The subsequent segment, the concluding part, will summarize the study's findings, highlighting potential contributions to the field and outlining avenues for future research. This structured approach aims to systematically understand the complexities of post-segmentation market basket analysis, contributing to a more nuanced comprehension of effective product recommendations in the coffee shop domain. The study seeks to enhance the body of knowledge in this field, offering valuable insights for researchers, practitioners, and stakeholders alike.

# Literature review

The literature review begins by providing a comprehensive background on the domain, tracing the historical development of market basket analysis and customer segmentation within the context of coffee shops. By drawing insights from interviews with domain experts, including coffee shop managers and professionals in recommendation system design for clothing and grocery sectors, this section aims to contextualize the specific challenges and opportunities unique to the coffee shop industry.

The literature review starts with a thorough exploration of the historical development of market basket analysis (MBA) and customer segmentation, particularly within the unique context of coffee shops.

Product recommendations have become an integral facet of modern consumer experiences, finding widespread application across various industries, particularly in e-commerce, retail, and digital platforms. In an era filled with choices, product recommendations play a pivotal role in enhancing user engagement, streamlining decision-making processes, and ultimately contributing to customer satisfaction. Leveraging sophisticated algorithms and data analytics, these recommendations tailor suggestions based on various factors.

The prevalence of product recommendations extends benefits not only to consumers by facilitating a more personalized and efficient shopping experience but also to businesses aiming to optimize sales, improve customer retention, and foster brand loyalty. As a result, the interplay between consumers seeking curated choices and businesses striving for enhanced profitability underscores the importance of product recommendations in the current marketplace.

Market Basket Analysis (MBA) is a data analysis technique that explores patterns and relationships among items that are frequently purchased together. It originated from the retail industry, where retailers sought to understand the purchasing behaviour of customers. The primary objective is to identify associations between products and discover which items tend to be bought together during a shopping transaction. This analysis helps businesses make informed decisions, such as optimizing product placement, designing effective marketing strategies, and enhancing the overall shopping experience.

Market Basket Analysis has its roots in the early 1990s, with the advent of large-scale transactional databases. The pioneering work in this field can be attributed to the seminal paper titled "Market Basket Analysis" by R. Agrawal and R. Srikant in 1994. Their research introduced the Apriori algorithm, a fundamental method for discovering frequent itemsets, which are sets of items that appear together in transactions. This algorithm laid the groundwork for efficient association rule mining, enabling the extraction of meaningful insights from vast datasets.

Since then, Market Basket Analysis (MBA) has evolved alongside advancements in data mining and machine learning. The introduction of the Apriori algorithm by Agrawal and Srikant in 1994 marked a significant milestone in MBA's development, offering an effective method for discovering frequent itemsets. This algorithm laid the groundwork for subsequent advancements in MBA methodologies.

In addition to Apriori, various algorithms, such as FP-Growth (Frequent Pattern Growth) and Eclat, have been developed to enhance efficiency and scalability in MBA applications. These algorithms leverage innovative techniques to identify patterns and associations within transactional datasets. The versatility of MBA extends beyond retail, finding applications in diverse sectors, including e-commerce, telecommunications, and healthcare. In these domains, MBA provides valuable insights into consumer behaviour and preferences, contributing to informed decision-making processes. This ongoing evolution of MBA methodologies underscores its adaptability and relevance in contemporary data-driven landscapes.

To complement this historical narrative, insights from interviews with coffee shop managers enrich the literature review. Chosen for their firsthand experience in managing coffee shop environments, these experts provide practical insights into the challenges and opportunities associated with implementing customer segmentation and recommendation systems. The decision to include this population in the study is grounded in the belief that their perspectives can provide a practical and contextual understanding of the dynamics involved in customer interactions, upselling strategies, and overall customer satisfaction within a coffee shop.

The interviews with the Coffee Shop Managers expose the existing landscape of upselling strategies within the coffee shop industry. Emphasizing its revenue-boosting potential, the manager notes that upselling can contribute to an increase of up to 80%. The techniques employed in this process are highlighted as dynamic and evolving, necessitating continuous training and coaching for staff members to enhance their upselling skills. The focus of this training is on recommending more suitable products to customers. The selection of offered products is contingent upon several factors, including current stock levels, seasonal items, promotions, and the time of day. Notably, the approach currently lacks a pronounced focus on individualized customer preferences.

Another aspect brought to light in the interviews focuses on the effectiveness of the current upselling strategies New staff members are reported to engage in upselling with an initial success rate ranging from 10% to 15%. Furthermore, the discretion afforded to shops in configuring upselling products leans towards those with the highest margin and highest current stock levels to mitigate potential waste.

The second population targeted for the primary research comprises individuals recognized as experts actively involved in the creation of customer segmentation and recommendation systems. These experts possess a wealth of knowledge and experience in developing strategies to categorize customers effectively and design recommendation systems tailored to specific industries. The selection of this population is driven by the necessity to gain in-depth insights into the sophisticated methodologies and best practices employed in customer segmentation and recommendation system creation.

Field Expert 1, specializing in Company X's clothing department, shed light on key factors shaping product recommendations within the company.

Firstly, stock levels emerge as a primary determinant in influencing product recommendations. The expert underscores the critical role of maintaining optimal inventory levels to ensure that the recommended products are readily available for customers, contributing to seamless customer experiences.

Supplier agreements and deals are identified as second, but equally important elements in shaping product recommendations. The expert highlights the strategic alignment with suppliers, allowing the company to offer exclusive products or negotiate favourable deals. This strategic partnership with suppliers significantly impacts the range of recommended products.

Customer behaviour and preferences, particularly observed through online interactions, constitute another critical aspect of the recommendation process. The expert emphasizes the importance of leveraging historical purchase data, browsing behaviour, and personalized preferences to tailor recommendations for individual customers, enhancing the relevance and personalization of the suggestions.

In addition, user engagement metrics, such as click-through rates and conversion rates, hold significant weight in the recommendation system. The company places high importance on monitoring these metrics to continuously refine and optimize recommendations, ensuring their ongoing effectiveness and alignment with customer needs.

An innovative idea presented by the expert emphasizes the need for dynamic adaptation to market trends. This entails regular updates based on market analysis, competitor movements, and emerging consumer preferences, ensuring that the recommendation system remains agile and aligned with the ever-evolving market landscape.

The expert also underlined the importance of effective communication and collaboration with marketing teams. Aligning product recommendations with ongoing marketing campaigns and strategies enhances the overall coherence of the customer experience. This collaboration ensures that recommendations seamlessly integrate with broader marketing initiatives, contributing to a cohesive and impactful customer journey.

Field Expert 2, specializing in Company X's food department, emphasised various promotions as a key factor in product recommendation. The ongoing promotions within the store or on the website, whether they involve special deals, discounts, or bundled offers, play a pivotal role in shaping the recommendations presented to customers. The expert underscores the dynamic nature of these promotional strategies in creating the direction of product recommendations. Notably, in the context of online food shopping, the connection to customer profiles adds another layer to this dynamic. Customer profiles store information about the last purchases, and recommendations are crafted based on these past preferences, in conjunction with real-time data on currently available stock. This integration ensures that online recommendations are not only influenced by ongoing promotions but also personalized to align with the customer's historical preferences and the present stock availability.

Anticipating and responding to seasonal trends emerge as integral components in the recommendation process. The expert emphasizes the significance of aligning product recommendations with seasonal shifts in customer preferences and demands. This involves demand forecasting, utilizing historical data and market trends to anticipate customer expectations during different times of the year. The expert's insights underscore the importance of a forward-looking approach to ensure that recommendations remain attuned to evolving seasonal dynamics and customer needs. Importantly, in the realm of online food shopping, this alignment extends to the customer profile, where past purchases serve as valuable data points influencing personalized recommendations, alongside considerations of currently available stock. This comprehensive approach ensures that online recommendations dynamically adapt to both seasonal trends and individual customer preferences.

The insights garnered from interviews show that product recommendations are predominantly driven by factors such as profit margins, current stock levels, and the need to minimize potential waste. Unlike in the textile or grocery sectors, where recommendations are influenced by ongoing promotions, deals, and supplier agreements, the coffee shop's approach tends to prioritize financial considerations and inventory management.

Unlike the online textile and grocery shopping experience, where customer segmentation plays a role through the analysis of purchase history and browsing behaviour, coffee shops typically lack a sophisticated customer segmentation strategy. In the absence of such segmentation, recommendations are less tailored to individual preferences and behaviours, focusing more on optimizing financial outcomes and managing inventory effectively. This distinction highlights the unique challenges and practices within the coffee shop domain compared to other retail sectors.

Selecting the correct model for customer segmentation is a pivotal decision that significantly influences the effectiveness of marketing strategies and overall business success. When the right segmentation model is employed, businesses can gain profound insights into their customer base, allowing for tailored and targeted approaches. This precision enables the delivery of personalized marketing campaigns, product recommendations, and services, thereby enhancing customer satisfaction and loyalty (Kari, 2022) (Anon., 2024). The correct model ensures that marketing efforts are aligned with the actual needs, preferences, and behaviours of specific customer segments, maximizing the impact of promotional activities and resource allocation.

The consideration of various clustering methodologies was integral to determining the most suitable approach for the current research, where coffee shop workers need to seamlessly allocate customers to the proper groups. The K-means clustering algorithm, an unsupervised machine learning technique, exhibited remarkable success in categorizing customers efficiently (E.Y.L. Nanadapala and K.P.N. Jayasena, 2020), (Vardhan, Gandhodi & Kala, Morthala & Reddy, Rushitha & Teja, Maddali & Mokshitha, Sadda & Pavani, Ponnada, 2022), (Prof. Nikhil Patankar, Soham Dixit, Akshay Bhamare, Ashutosh Darpel, Ritik Raina, 2021).

The K-means clustering algorithm plays a pivotal role in efficiently categorizing customers based on complex patterns within datasets. This method initiates with the careful selection of the desired number of clusters (k). Subsequently, the algorithm randomly initializes k centroids, representing the central points of these clusters. Each data point within the dataset is then carefully examined for its distance to each centroid, leading to the assignment of the point to the cluster whose centroid exhibits the minimum distance. This process establishes preliminary clusters.

Following the initial assignment, the algorithm proceeds to update the centroids, recalculating them as the mean of all data points within their respective clusters. This marks the completion of one iteration. The algorithm iteratively repeats the assignment and centroid update steps until a stopping criterion is met, often indicated by minimal changes in data point assignments or reaching a predefined number of iterations.

Choosing the appropriate number of clusters (k) is a critical consideration in the application of K-means. Two common methods for this determination are the Elbow Method and the Silhouette Score. The Elbow Method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and identifying the "elbow" point, signifying a suitable k. Alternatively, the Silhouette Score evaluates the cohesion and separation of data points within clusters, with a higher score indicating well-defined clustering.

It is important to note that the K-means algorithm is sensitive to the initial placement of centroids, leading to potentially different outcomes with varied initializations. To mitigate this sensitivity, the algorithm is often executed multiple times with diverse initializations, and the best result is selected.

By identifying patterns and segments within datasets, K-means facilitated a comprehensive understanding of consumer behaviour, preferences, and characteristics, enabling informed decision-making and refinement of marketing strategies. This algorithm's versatility and precision positioned it as a valuable tool for businesses seeking nuanced and effective customer segmentation.

Despite the instrumental success of the K-means algorithm, hierarchical clustering emerged as an alternative with distinctive advantages.

Hierarchical clustering stands out as a robust methodology employed for the systematic organization of similar data points into clusters, culminating in a hierarchical structure visually represented through a dendrogram. Distinguished from k-means clustering, this approach dispenses with the need for a predetermined cluster count, presenting enhanced flexibility in its application. A comprehensive exploration of hierarchical clustering necessitates an examination of its foundational principles.

At its core, hierarchical clustering treats each data point as an individual cluster, systematically merging the closest clusters iteratively. This iterative process results in a dendrogram, a visual representation of the hierarchical arrangement. Vertical lines within the dendrogram denote clusters, and the height of their merger signifies dissimilarity.

A pivotal facet of hierarchical clustering involves the selection of distance metrics, such as Euclidean or Manhattan distance, shaping the calculation of similarity between clusters or data points. This choice significantly influences the clustering outcome. Additionally, the method relies on linkage methods, including single, complete, and average linkage, to determine how the distance between clusters is computed during the merging process.

The interpretation of the dendrogram is crucial for extracting meaningful insights from hierarchical clustering, offering a snapshot of the relationships between data points or clusters. Tailoring the dendrogram cut at a specific height facilitates the determination of the desired number of clusters, a decision informed by the intended granularity.

Hierarchical clustering manifests in two forms: agglomerative and divisive. Agglomerative clustering commences with individual data points as clusters, progressively merging them, while divisive clustering initiates with all data points in a single cluster, recursively dividing them. The choice between these forms hinges on the specific analytical requirements.

The choice between these clustering approaches should be guided by the specific characteristics of the dataset, particularly in scenarios involving extensive data (Anifa, Mansurali & Prem, Mj & Hack-Polay, Dieu & Mahmoud, Ali & Grigoriou, Nicholas, 2022). For extensive datasets, K-means clustering, with its efficiency and scalability, may be preferable. K-means relies on centroid-based grouping, partitioning data into a predetermined number of clusters, making it computationally less intensive for large datasets. However, its reliance on a pre-defined number of clusters might lead to challenges in scenarios where the optimal number is uncertain or variable.

On the other hand, hierarchical clustering, while more computationally intensive, provides greater flexibility in revealing structures at varying granularity levels. It does not require a predefined number of clusters, allowing for a more adaptive approach. This flexibility is particularly advantageous when dealing with extensive datasets where the optimal number of clusters may not be readily apparent.

Regrettably, none of the conventional methods for customer segmentation proved suitable for the specific requirements of the task at hand. Given the necessity for a straightforward and easily applicable segmentation approach, the method chosen for coffee shop segmentation at the point of sale involved establishing clear and uncomplicated rules. This ensures that cashiers can accurately assign customers to predefined segments with ease.

# Methodology

## Research aim

To investigate the effects of customer segmentation on product recommendation in a coffee shop environment.

The primary aim of this research is to explore the intricate dynamics between customer segmentation and product recommendations in the specific context of a coffee shop environment. The overarching goal is to understand the effects and implications of implementing a customer segmentation strategy on the recommendations offered to customers. This investigation seeks to understand how distinct customer groups, categorized based on gender and age, influence the nature and effectiveness of product recommendations. By closely examining transactional data and applying advanced analytics, the research aims to reveal nuanced patterns in consumer behaviour within the coffee shop setting. The ultimate objective is to cultivate a more tailored and refined product recommendation system that not only aligns with the diverse preferences of individual customers but also contributes to an enhanced and gratifying overall experience. The aspiration is to go beyond conventional practices, strategically leveraging analytics to craft personalized recommendations that resonate with each identified customer segment, thereby promoting heightened satisfaction and enduring customer loyalty.

## Research objectives

To establish relevant customer segments that assist coffee shop staff in categorizing new customers during the purchase process.

To create relevant product recommendations tailored for the identified customer segment.

To forecast the number of transactions for the next 7 days for effective resource planning, staffing optimization, and inventory management, ensuring the coffee shop is well-prepared to meet anticipated demand.

The first objective focuses on establishing pertinent customer segments and providing valuable assistance to coffee shop staff in categorizing new customers during the purchase process. By implementing segmentation techniques based on visual aspects such as gender and age groups, distinct customer groups with similar purchasing behaviours are discerned. These identified customer segments form the foundation for subsequent analyses, facilitating the development of more targeted and personalized strategies.

Moving to the second objective, the focus shifts to creating relevant product recommendations tailored to the identified customer segments. This involves the strategic application of market basket analysis, utilizing the powerful Apriori algorithm. The goal is to unveil crucial patterns and associations within each discerned customer segment, providing insights into nuanced preferences and purchasing behaviours. Subsequently, personalized product recommendations are crafted to align seamlessly with the identified preferences and behaviours within each segmented customer group. The application of these personalized recommendations is anticipated to significantly enhance the overall customer experience, foster customer loyalty, and potentially drive increased sales.

The third and additional objective revolves around forecasting the number of transactions for the next 7 days. This forecasting task utilizes advanced time series forecasting models such as ARIMA, SARIMA, and ETS. The objective is to predict the number of transactions for the upcoming week, enabling effective resource planning, staffing optimization, and inventory management. This proactive approach ensures that the coffee shop is well-prepared to meet anticipated demand, aligning with the tailored product recommendations to enhance customer satisfaction and optimize operational efficiency.

## Project architecture

The research process included interviews with domain experts, such as coffee shop managers and product recommendation specialists from textile and grocery departments. A comprehensive dataset was systematically assembled, serving as the foundational resource for all research tasks. The customer segmentation phase involved categorizing customers based on gender and age groups and implementing clear and straightforward rules at the point of sale to facilitate cashier operations. The Apriori algorithm was then applied for Market Basket Analysis (MBA), both separately for each customer group and comprehensively for the entire dataset. Results of the MBA were systematically compared across genders, age groups, and the general dataset, with a focus on a specific product, such as a latte. Additionally, a detailed examination of product recommendations for the chosen product was conducted, contrasting findings between genders, age groups, and the overall dataset.

As an extension to the research, a transaction forecast for the next 7 days was incorporated to assess the anticipated unit's activity. This forecasting element aimed to provide insights into the potential business volume, offering a comprehensive perspective on the coffee shop's expected operational load. The final stages of the research involved an in-depth analysis and interpretation of patterns and variations in consumer behaviour, specifically exploring the efficacy of personalized upselling practices within the coffee shop setting.

# Evaluation and analysis

## Customer segmentation

## Market basket analysis

## Forecasting

The dataset reveals distinct temporal trends in the number of products sold, illustrating fluctuations across different dates. Notably, certain days, such as April 13, 2022, and April 27, 2022, stand out with significantly higher quantities sold, suggesting potential peak sales days. The analysis also indicates the presence of patterns, prompting further exploration of consistent trends or variations between different months.

A graph showing the price of sold products

Description automatically generated

Upon scrutinizing the data from May 1st onwards, a notable pattern emerges, showcasing a sustained and relatively lower level of sold products compared to the concluding days of April. Interestingly, this stable pattern persists without reverting to the elevated quantities observed in the last week of April. This observed shift implies a potential alteration in customer behaviour, staff upselling style, demand dynamics, or external factors influencing sales.

A graph showing the number of transaction

Description automatically generated

The transaction count remains consistent throughout the week, experiencing a slight decrease of less than 2,000 on Tuesday, Wednesday, and Thursday. Notably, Sunday emerges as the day with the highest transaction activity.

A graph of a transaction number

Description automatically generated

The analysis of transaction volume throughout the day reveals distinctive patterns. From 6 am to 10 am, there is a notable surge in transaction volume, reaching its pinnacle at 10 am. Subsequently, between 11 am and the early evening hours, the volume remains relatively steady, hovering around 7,000 transactions. However, a drastic drop happens at 8 pm. These fluctuations in transaction volume at different hours of the day provide insights into the temporal dynamics of customer engagement and highlight peak periods of transaction activity.

A graph with a green line

Description automatically generated with medium confidence

The chart above represents information of various coffee shop products and their corresponding transaction volumes, providing insights into customer preferences and purchasing behaviour. Among the 80 items in the dataset, the majority, exhibit a robust transaction volume, averaging around 2000 purchases. Additionally, there are 10 items characterized by a moderate volume, approximately 1500 transactions each. Another subset of 10 items falls within the range of 500 to 1000 transactions, representing a mid-range popularity. Lastly, 20 items in the dataset have a relatively lower transaction volume, each recording around 100 purchases. This diversified distribution highlights the varying degrees of popularity and demand for different coffee shop products.

A graph with blue squares

Description automatically generated

The examination of the frequency of items bought per transaction reveals distinctive patterns in purchasing behaviour. A noteworthy observation is that a significant majority, exceeding 60% of all transactions, consist of a single item purchase. The dominance of single-item transactions underscores the prevalent occurrence of customers opting for a singular product during their interactions with the coffee shop. On the contrary, transactions involving two items constitute approximately 26% of the dataset, indicating a notable but lesser frequency. The occurrences of transactions involving three or more items are notably scarce, collectively comprising less than 10% of the dataset.

The observed distribution indicates a notable preference for individualized purchases among customers. This trend may signify a customer base that consistently chooses straightforward and specific options rather than opting for more complex combinations or bundled products. Alternatively, it could be indicative of potential gaps in upselling techniques or skills among coffee shop employees, as customers seem to gravitate towards single-item transactions.

# Conclusion

Chapter 3: Methodology(Aim of the research, Objectives to achieve this aim, explain step by step your methodology)

Chapter 4: Evaluatio and analysis

Chapter 5: Conclusion( here you need to write contribution to the body of knowledge and future work)

3.1

3.2 Analysing sold products by gender

3.2 Analysing transactions

3.3 Forecasting transaction volume for next 7 days

1. Data preparation for transaction forecasting

Transaction forecasting relies heavily on the careful preparation and understanding of the underlying data. In this context, the data underwent a detailed process to unveil its inherent patterns and characteristics, ensuring that subsequent forecasting models could effectively capture and predict transactional trends.

-should I add what exactly was done?

1. Seasonal Decomposition:

The initial step involved the application of the seasonal decomposition method to disentangle the various components within the transactional data. Two decomposition approaches, multiplicative and additive, were employed to assess the data's behaviour under different assumptions.

The trend component represented the overall trajectory of transactional data, the seasonal component captured recurring patterns, and the residuals represented the unexplained variance. Examining these components provides insights into the data's inherent dynamics, aiding in the interpretation of forecasting results.

A group of graphs showing different types of data

Description automatically generated with medium confidence

The results indicated an additive structure as the residuals displayed a more widely and randomly spread pattern, contrasting the multiplicative model's residuals that formed a nearly straight line at value 1. This insight into the data's nature is crucial for selecting appropriate forecasting models.

1. Statistical Stationarity Assessment:

The assessment of statistical stationarity within a time series constitutes a foundational aspect critical to the robustness of forecasting models. The significance of stationarity lies in its pivotal role in enabling forecasting models to make accurate predictions, safeguarding against the influencing factor of changing statistical characteristics over time.

In the context of time series forecasting models such as ARIMA and SARIMA, the assumption of stationarity is inherent to their design and functionality. The constancy of statistical properties, including mean, variance, and autocorrelation, provides a stable foundation for these models to discern genuine patterns from the temporal data. Consequently, the stationarity assessment conducted via the Dickey-Fuller test not only affirms the compliance to a fundamental model assumption but also substantiates the reliability of subsequent predictions.

The p-value obtained from the Dickey-Fuller test, 0.000003 in this instance, undergoes examination against a predetermined significance level, conventionally set at 0.05. A p-value below this threshold signifies the rejection of the null hypothesis of non-stationarity, thus confirming the stationarity of the time series. This standardized criterion offers a lucid parameter for evaluating the stability of the time series, establishing a robust foundation for subsequent forecasting endeavours.

In scenarios where stationarity is compromised, it has far-reaching consequences for the accuracy of forecasting models. Non-stationarity introduces the risk of misleading correlations, adding complexity to distinguishing genuine patterns amid random fluctuations. The misinterpretation of dynamic statistical properties as meaningful trends becomes a threat, underscoring the imperative role of stationarity in fortifying the precision and reliability of forecasting models.

4.2 Application of ARIMA, SARIMA, and ETS models

1. ARIMA

The AutoRegressive Integrated Moving Average (ARIMA) model operates by integrating three key components: autoregression, differencing, and moving averages. Each of these elements contributes to the model's effectiveness in capturing and predicting temporal dependencies within sequential data. ARIMA's strength lies in its flexibility to accommodate diverse time series patterns. Whether the data exhibits linear trends, seasonality, or more complex temporal dependencies, ARIMA can adapt by adjusting the values of its order parameters (p, d, q). This adaptability makes it suitable for capturing the nuances of various datasets and forecasting accurately in the presence of changing trends over time.

Autoregression (AR) – the autoregressive component assesses the relationship between an observation and its previous values in a time series. It leverages the concept that the current value of a variable can be expressed as a linear combination of its past values. The ARIMA model considers the autoregressive order, denoted as "p," which signifies the number of lag observations included in the model. A higher "p" value implies a more extensive consideration of past observations.

Integrated (I) – the integrated component involves differencing the time series data to achieve stationarity. Differencing calculates the differences between consecutive observations, helping stabilize the mean and rendering the data more amenable to modelling. The order of differencing, denoted as "d," represents the number of times differencing is applied to attain stationarity. The integrated component ensures that the temporal patterns in the data are captured effectively.

Moving Averages (MA) – the moving average component considers the relationship between an observation and a residual error from a moving average model applied to lag observations. It smoothens out short-term fluctuations in the data and aids in identifying underlying trends. The order of the moving average, denoted as "q," indicates the number of lagged forecast errors considered in the model. A higher "q" value implies a greater emphasis on past forecast errors.

A comparison of a graph

Description automatically generated with medium confidence

The initial exploration involved examining the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots to discern potential temporal dependencies and guide the selection of hyperparameters for the ARIMA model. Through careful examination of these plots, the order parameters were determined to be (1, 1, 9), indicating a first-order differencing, an autoregressive component of order 1, and a moving average component of order 9. Additionally, seasonal order parameters (1, 1, 1, 12) were chosen to account for potential seasonality with a periodicity of 12 units.

A graph showing a graph

Description automatically generated with medium confidence

Subsequently, the Arima model, utilizing the prescribed hyperparameters, underwent empirical validation through testing on the designated training set. The assessment of predictive performance ensued on the allocated test set, with a meticulous examination of key metrics. The computed mean absolute error (MAE) yielded a value of 58.08, signifying the average absolute disparity between actual and predicted values. Concurrently, the mean squared error (MSE) was quantified at 5189.49, furnishing a comprehensive insight into the collective squared discrepancies between the forecasted and observed values. Additionally, the mean absolute percentage error (MAPE), computed at 3.31%, provided a percentage-based evaluation of the model's precision.

1. SARIMA

The Seasonal Auto Regressive Integrated Moving Average (SARIMA) model stands as a robust methodology for time series forecasting, particularly in scenarios where historical patterns significantly influence future trends. SARIMA augments autoregression, differencing, and moving averages, making it adept at capturing and predicting temporal dependencies within sequential data. Its strength lies in its flexibility to accommodate various time series patterns and the adaptability to changing trends over time.

SARIMA retains the autoregressive, integrated, and moving average components found in ARIMA. The autoregressive aspect captures the relationship between current and past observations, the integrated component ensures stationarity through differencing, and the moving averages smooth out short-term fluctuations. However, SARIMA further refines these components by introducing seasonal orders (P, D, Q), allowing it to account for periodic variations in the data

Seasonal Orders (P, D, Q):

* Seasonal Autoregressive (P): Represents the number of lag observations for seasonal autoregression.
* Seasonal Integrated (D): Denotes the number of seasonal differences applied to achieve stationarity.
* Seasonal Moving Averages (Q): Indicates the number of lagged forecast errors for seasonal moving averages.

SARIMA's strength lies in its ability to address both the temporal dynamics captured by ARIMA and the seasonal variations inherent in many time series datasets. By introducing seasonal orders, SARIMA accommodates recurrent patterns, such as those occurring yearly, monthly, or at other fixed intervals. This enhances the model's adaptability to datasets with complex, intertwined temporal and seasonal dependencies.

The exhaustive search for optimal hyperparameters was imperative to fine-tune the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and enhance its predictive accuracy on the test set. The comprehensive exploration spanned autoregressive (AR), differencing (D), moving average (MA), and seasonal orders (P, D, Q, s). The iterative process systematically tested various combinations to identify the configuration yielding the minimum Mean Absolute Error (MAE) during model evaluation on the test set.

The performance of each SARIMA model was assessed based on MAE, providing a quantitative measure of its predictive accuracy. The culmination of this hyperparameter search identified the optimal configuration as {'p': 0, 'd': 1, 'q': 1, 'P': 1, 'D': 0, 'Q': 1, 's': 14}. Applying these refined hyperparameters to the SARIMA model resulted in a notable improvement in forecasting accuracy. The model, with the optimized parameters, achieved a reduced MAE of 31.36 on the test set, demonstrating its enhanced ability to capture and predict the intricate temporal and seasonal patterns inherent in the time series data. This outcome underscores the significance of the hyperparameter tuning process in tailoring the SARIMA model to the specific nuances of the dataset, contributing to its robust performance in time series forecasting.

ETS

Exponential smoothing is a statistical method for analysing and forecasting time series data. The fundamental idea behind exponential smoothing is to give more weight to recent observations while gradually decreasing the influence of older observations.

ETS retains the simplicity and interpretability of exponential smoothing while addressing the limitations of ARIMA models. The error term captures random fluctuations, the trend term accounts for systematic variations, and the seasonality term accommodates periodic patterns. This adaptive approach allows ETS to effectively model a diverse range of time series data.

The model's strength is underscored by its ability to address both the temporal dynamics inherent in ARIMA models and the inherent seasonality characterizing various time series datasets. By introducing smoothing parameters (α, β, γ), ETS facilitates dynamic adjustments to evolving trends over time.

These parameters control the weights assigned to the most recent observations for error, trend, and seasonality, respectively.

* α (alpha): Controls the smoothing of the error term. A higher alpha places more weight on recent observations, making the model more responsive to short-term fluctuations.
* β (beta): Governs the smoothing of the trend component. Similar to alpha, a higher beta gives more weight to recent observations, allowing the model to adapt to changes in the trend.
* γ (gamma): Manages the smoothing of the seasonality component. A higher gamma emphasizes recent seasonal patterns, making the model more adaptable to evolving seasonal variations.

The exploration for optimal hyperparameters involved the consideration of smoothing parameters and the error type, with specific values such as alpha values = [0.2, 0.4, 0.6, 0.8], beta values = [0.2, 0.4, 0.6, 0.8], and gamma values = [0.2, 0.4, 0.6, 0.8]. The search yielded the optimal configuration of: (0.6, 0.2, 0.2).

Testing the ETS model with these parameters resulted with:

Mean Absolute Error (MAE): 39.20

Mean Squared Error (MSE): 2794.95

Mean Absolute Percentage Error (MAPE): 2.18%

1. Methodology

4.1 Customer segmentation (including lit review)

While various clustering methodologies, such as K-means and hierarchical clustering, have proven effective in diverse scenarios, it is crucial to acknowledge that the unique context of this research necessitates a distinct approach. The present study does not rely on automated algorithms like K-means or hierarchical clustering. Instead, the segmentation process is personally curated, aligning with the practical needs of coffee shop workers who must promptly allocate customers to specific groups at the till point. In this hands-on approach, information easily observable by the cashier, such as gender and approximate age group, becomes the basis for segmentation. This bespoke strategy recognizes that, despite the effectiveness of advanced clustering methods, the immediate and visual nature of customer attributes at the point of sale demands an intuitive and simplified segmentation process. Therefore, in this particular scenario, the emphasis shifts away from automated algorithms, underscoring the importance of practical and visually noticeable criteria for the smooth allocation of customers.

Following this approach, customers are first categorized based on their gender, and subsequently, they are assigned to predefined age groups. This sequential and visually discernible criteria-driven strategy ensures an organized and efficient customer allocation system. By prioritizing practicality and immediate visual recognition, businesses can tailor their engagement based on gender-specific and age-specific characteristics.

A diagram of a general structure

Description automatically generated

4.2 Product recommendation using Apriori

Market basket analysis is a crucial aspect of data analytics that involves examining customer purchase patterns to identify associations between different products (Loshin, 2013) (Cavique, 2007). It helps businesses understand which items are frequently bought together, providing valuable insights for effective marketing strategies (Hana Bernika Sabila, Feri Candra, 2023) and inventory management. In a broader context, this analysis aids in enhancing the overall customer shopping experience and increasing revenue.

When applied to a specific domain, such as a coffee shop, market basket analysis becomes particularly insightful. In the context of a coffee shop, understanding customer preferences for various coffee blends, add-ons, and accompanying snacks can notably enhance coffee shop sales through strategies like upselling and cross-selling (Shweta, Cassie Bottorff, 2022). Analysing the purchase behaviour of coffee shop customers can unveil patterns like the correlation between certain types of coffee and particular pastries or the popularity of specific combos during different times of the day. This information empowers the coffee shop to tailor its offerings, promotions, and customer interactions, ultimately fostering customer satisfaction and loyalty. Additionally, the implementation of market basket analysis in a coffee shop setting can contribute to more efficient inventory management, reducing waste and ensuring that popular items are consistently available.

In the exploration of customer purchase behaviour within a coffee shop dataset, two algorithms were considered for market basket analysis: Eclat and Apriori. These algorithms play a crucial role in discovering associations between products and identifying frequent itemsets, offering valuable insights into the relationships and preferences exhibited by customers. Eclat and Apriori present distinct approaches to extract meaningful patterns from transaction data, and the selection between these algorithms depends on the characteristics of the dataset and the specific goals of the analysis. In this context, the decision-making process involved an evaluation of the dataset's size, sparsity, desity, and other relevant features to determine the most suitable algorithm for uncovering meaningful associations within the coffee shop's transaction data.

Apriori, a classic algorithm in association rule mining, plays a key role in finding patterns and relationships within datasets (Jiawei Han, Jian Pei, Micheline Kamber, 2012), particularly in the context of the food and coffee industry. This algorithm is adept at identifying frequent item sets, showcasing the co-occurrence of items in transactions. This intricate procedure involves an exhaustive enumeration, encompassing the counting of occurrences for individual items and progressively extending to more extensive itemsets. Subsequently, items or itemsets that fail to meet the established minimum support criteria undergo a pruning process, excluding them from further consideration within the algorithmic framework.

The minimum support is a critical parameter that determines the threshold for considering an itemset as frequent. It represents the proportion of transactions in which a particular itemset must occur to be deemed significant. Additionally, confidence level is another parameter, indicating the likelihood that the presence of one item in a transaction implies the presence of another.

The confidence level is another parameter of this algorithm. It is a measure thet quantifies the strength of an association rule. Confidence is calculated based on the support of both the antecedent and the consequent of a rule.

Lift is a last metric in association rule mining that compares the likelihood of both items being purchased together against the likelihood of them being purchased independently. It is particularly valuable for identifying significant patterns and dependencies between items. A lift value greater than 1 suggests that the items are more likely to be bought together than would be expected by chance, while a lift less than 1 indicates a weaker association.

While lift can provide valuable insights in various contexts, its omission in this particular analysis is a strategic decision driven by the specific characteristics of the dataset. The dataset is characterized by its small size, a predominant occurrence of single-item transactions (primarily coffee), and a large number of available products. In such a scenario, the co-occurrence of items may be limited, making lift less informative. The focus on support and confidence metrics is chosen for its alignment with the goal of efficient analysis and effective communication of findings. Additionally, given the resource constraints, excluding lift contributes to computational efficiency, ensuring that the modeling process remains tailored to the unique aspects of the dataset and the desired outcomes of the analysis.

Following the identification of frequent itemsets, the algorithm proceeds to employ the Breadth-First Search (BFS) traversal methodology. BFS, a fundamental concept underpinning Apriori, is a graph traversal algorithm widely employed in computer science. In the context of Apriori, BFS entails systematically exploring the transactional dataset level by level, ensuring that all neighbours of a node are visited before moving on to their neighbours. The algorithm utilizes a queue data structure to manage the nodes to be visited, enqueuing the neighbours of the current node for future exploration. The exploration progresses in a level-order manner, horizontally traversing the levels before descending to the next level. Nodes are marked as visited to prevent revisiting, and the process continues until all reachable nodes have been explored.

A diagram of a search engine

Description automatically generated

BFS is not an inherent part of the Apriori algorithm itself, rather, it is a subsequent step used for efficient visualization and interpretation of the discovered frequent itemsets. After obtaining the frequent itemsets, BFS helps organize and present the relationships between items in a hierarchical manner, providing a clear representation of the association rules.

A diagram of a diagram

Description automatically generated

The Apriori algorithm, driven by the principles of breadth-first search, emerges as a robust and versatile tool for association rule mining in the domain of coffee shop customer product recommendations. Its systematic exploration of transactional datasets and efficient candidate generation process position it as a valuable asset in uncovering meaningful associations and patterns within diverse datasets.

Second algorithm which was considered is Eclat, short for Equivalence Class Clustering and Bottom-Up Lattice Traversal, is a frequent itemset mining algorithm that holds distinct strengths in certain scenarios. Eclat is particularly advantageous when dealing with large transaction datasets, as it efficiently discovers frequent itemsets without the need for candidate generation, making it more memory-efficient than some other algorithms.

The primary application of Eclat lies in market basket analysis, similar to Apriori. It is widely used in retail, e-commerce, and recommendation systems, where understanding the associations between items in transactions is crucial for optimizing product recommendations and enhancing customer experience.

Eclat operates by first identifying frequent items and their occurrences in the dataset. It then recursively extends these frequent items into larger itemsets, forming a lattice structure. Unlike Apriori, Eclat does not generate candidate itemsets explicitly, which contributes to its efficiency, especially in datasets with high dimensionality.

One of the notable features of Eclat is its simplicity in terms of parameter tuning. It mainly relies on the minimum support threshold, representing the minimum frequency required for an itemset to be considered frequent. This simplicity makes it user-friendly and easy to implement.

While Eclat does not employ a Breadth-First Search (BFS) approach like Apriori for visualization, it organizes itemsets in a depth-first manner.

The choice of the Apriori algorithm over Eclat for association rule mining was made based on several considerations, tailored to the characteristics of the dataset at hand.

Apriori and Eclat are both popular algorithms for discovering frequent itemsets and association rules, but their suitability may depend on the specific characteristics of the dataset. In this particular scenario, the dataset was relatively small, and there was no significant sparsity in the transaction data. Moreover, Apriori algorithm adapts well to diverse datasets, accommodating variations in transaction lengths, underscores its significance in the context of coffee shop product recommendation systems.

Its stepwise approach, which involves iteratively identifying frequent itemsets based on a minimum support threshold, aligns well with the dataset's characteristics. Additionally, Apriori is often preferred when the dataset can comfortably fit into memory, making it a first choice for smaller datasets without concerns about computational efficiency. The absence of sparsity in the dataset further supported the decision, as Apriori tends to perform well when data is dense. While Eclat is known for its efficiency in sparse datasets, the specific features of the dataset in question led to the selection of the Apriori algorithm for association rule mining in this context.

1. Evaluation

5.1 Performance evaluation of forecasting models

5.2 Performance evaluation of ARIMA, SARIMA, ETS, and LSTM models

5.3 Evaluation of product recommendation:

After saying which algorithm was selected:  
For this research, a minimum support of 0.03 was set, ensuring that only itemsets occurring in at least 3% of transactions were considered. The initial confidence level was set to 0.5 but was later reduced to 0.03 to yield more meaningful and diverse results.

- between age groups of the same gender

- between genders

- between general and genders

- between general and age-specific gender recommendation

6. Ethical considerations

Within the operational structure of the envisioned recommendation system for a coffee shop, cashiers engaged in processing customer orders may find themselves in a situation where the selection of a predefined customer segment is integral to generating personalized product suggestions. In this process, cashiers are inevitably tasked with making assumptions about customers, relying on specific characteristics such as age group, and gender. This scenario raises potential ethical concerns that align with the fundamental principles outlined in the ethical considerations. The act of assuming characteristics without explicit customer consent touches upon the delicate balance between personalization and privacy. It prompts an exploration of how businesses can navigate this ethical landscape, ensuring that the benefits of tailored recommendations do not infringe upon individual privacy expectations. Addressing this ethical concern requires a comprehensive strategy that combines transparency, customer consent mechanisms, and a robust communication framework to build and maintain trust between the coffee shop and its patrons.

In the scenario where a cashier selects predefined customer segments to offer personalized suggestions, a key challenge emerges from the assumption-making process without explicit customer consent. To address this issue, businesses can implement a transparent communication strategy at the point of sale, ensuring customers are well-informed about the use of predefined segments for tailored recommendations. Neglecting to tackle this concern carries the risk of violating privacy expectations, eroding customer trust, and potentially leading to legal ramifications. A robust approach involves providing customers with the option to opt-in or opt-out of such profiling, respecting their choices, and enhancing overall transparency in data processing practices. This proactive strategy not only aligns with privacy regulations but also emphasizes a customer-centric approach, reinforcing trust and ethical standards in the business-customer relationship.

The selection of predefined customer segments by a cashier falls under the umbrella of profiling, and potential issues arise if this process is not handled ethically. It is crucial to strike a balance between personalization and customer privacy. Handling this involves implementing clear policies on how customer information is used for profiling, ensuring that the criteria used are non-sensitive and non-discriminatory. Failing to address this issue may lead to customer discomfort, loss of trust, and reputational damage for the business. Moreover, there is a risk of perpetuating biases if the predefined segments are based on inappropriate or discriminatory factors. Addressing these concerns involves regular audits of the profiling process and continuous improvement to align with ethical guidelines and customer expectations.

Additionally, cashiers, in making assumptions about customers through predefined segments, encounter a specific ethical challenge when dealing with teenagers or minors. This introduces a critical concern regarding the responsible collection and processing of data related to this specific age group. To address this ethical dilemma, businesses should prioritize the implementation of stringent age verification measures at the point of sale. This step ensures that personalized recommendations based on predefined segments are exclusively offered to customers who have reached the legal age, aligning with privacy and data protection regulations. Moreover, the system should incorporate an option to bypass segmentation if the age of the customer or any other reason hinders certainty. Failing to proactively address this issue could lead to potential breaches of data protection regulations, thereby compromising the privacy rights of minors. Implementing robust age verification mechanisms becomes paramount in mitigating ethical risks associated with the processing of data concerning teenage or minor customers, safeguarding both legal compliance and the well-being of the individuals involved.

When considering the transfer of personal data to non-EU countries, a pertinent scenario arises when a coffee shop brand operates its main location outside the EU and needs to transfer customer data. In such instances, it becomes crucial for the brand to navigate the complexities of international data protection laws. The potential issue lies in ensuring a seamless transfer of customer information while complying with the legal frameworks of both the EU and the destination country. This involves the establishment of robust data protection agreements that clearly outline the terms and conditions of secure data transfer, assuring customers that their information is handled with diligence and care.

To mitigate risks, businesses should implement stringent security measures, including encryption and secure networks, to safeguard customer data during the transfer process. Obtaining explicit customer consent and transparently communicating the reasons for data transfers further builds trust and aligns with ethical data practices. By proactively addressing these considerations, coffee shop brands with operations outside the EU can uphold privacy standards, legal requirements, and customer expectations in the realm of international data transfers.

Another ethical issue relies on data security, the potential risks associated with collecting sensitive information are considerable and should not be overlooked. Mishandling or inadequate protection of sensitive data, such as personal identification details, financial information, or health records, can lead to severe consequences. Unauthorized access, data breaches, or misuse of such sensitive information could result in violations of privacy, identity theft, or financial fraud, leading to reputational damage and legal ramifications for the coffee shop. However, it's crucial to highlight that there is generally no valid reason for a coffee shop to collect sensitive information like social security numbers, medical records, or financial details. Common examples of non-sensitive data that might be collected include customer preferences, purchase history, and demographic information. By focusing solely on non-sensitive data, the coffee shop can maintain ethical practices, safeguard customer trust, and reduce the potential risks associated with data collection.

Additionally, a potential risk in the context of cashiers making assumptions about customers based on predefined segments lies in the potential reinforcement of stereotypes and biases. If the predefined segments are not carefully crafted and validated, there is a risk of reinforcing existing biases or introducing new ones into the decision-making process. For instance, assumptions based on gender and age, may inadvertently contribute to unfair treatment or perpetuate stereotypes. To address this risk, businesses should ensure that the predefined segments are created with a thoughtful and unbiased approach, involving diverse perspectives and avoiding generalizations. Regular audits and reviews of the predefined segments can help identify and rectify any unintended biases, contributing to a fair and equitable customer segmentation process. By addressing this risk, businesses not only uphold ethical standards but also foster a more inclusive and respectful customer experience.

7. Conclusion

8. References