Post-segmentation product recommendation for coffee shop customers using Apriori modelling.

Estera Wisniewska

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

Contents

[Abstract 3](#_Toc159433187)

[Introduction 3](#_Toc159433188)

[Domain background 3](#_Toc159433189)

[Research Problem 4](#_Toc159433190)

[Research gap 4](#_Toc159433191)

[Research question 4](#_Toc159433192)

[Literature review 5](#_Toc159433193)

[Market Basket Analysis 5](#_Toc159433194)

[Customer segmentation 8](#_Toc159433195)

[Methodology 10](#_Toc159433196)

[Research aim 11](#_Toc159433197)

[Research objectives 11](#_Toc159433198)

[Project architecture 12](#_Toc159433199)

[Evaluation and analysis 12](#_Toc159433200)

[Customer segmentation 12](#_Toc159433201)

[Market basket analysis 14](#_Toc159433202)

[Model selection 14](#_Toc159433203)

[Application 18](#_Toc159433204)

[Forecasting 23](#_Toc159433205)

[Forecasting transaction volume for the next 7 days 24](#_Toc159433206)

[Application of ARIMA, SARIMA, and ETS models 26](#_Toc159433207)

[Models evaluation 29](#_Toc159433208)

[Transaction forecast 29](#_Toc159433209)

[Ethical considerations 30](#_Toc159433210)

[Conclusion 32](#_Toc159433211)

[References 32](#_Toc159433212)

# Abstract

This academic research focuses on a comprehensive strategy that integrates customer segmentation, Market Basket Analysis (MBA), and transaction forecasting within the operational context of a coffee shop. The segmentation strategy concentrate on simplicity and transparency, focusing on observable characteristics such as gender and age groups. The Market Basket Analysis explores general and age-group-specific patterns, utilizing different metrics based on the nature of segmentation. The Apriori algorithm is chosen for its relevance to transactional data and adaptability to diverse product sets. In the MBA component, a support threshold of 0.02 is employed for both general (based on the entire dataset) and gender-based segmentation (comprising distinct groups for females and males without further stratification), as higher thresholds yielded no meaningful results due to the dataset's characteristics. Age-group gender segmentation, with a support of 0.02 and confidence of 0.1, yielded meaningful insights tailored to the specific problem at hand. Transaction forecasting, using ARIMA, SARIMA, and ETS models, emerges as a critical component to predict and understand transaction volumes. This proactive approach aids in optimizing stock levels, aligning with strategic stock management to efficiently allocate resources. The examination of the forecasting models reveals SARIMA as the most effective for predicting 7 days of transaction volume. The hyperparameter tuning process enhances SARIMA's accuracy, underscoring the importance of tailoring models to dataset nuances in time series forecasting. Additionally, ethical considerations, particularly regarding minors, are seamlessly integrated into the strategy with an opt-out option for customers, contributing to an ethical and customer-centric approach in the retail sector.

# Introduction

## Domain background

Coffee cultivation has 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 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, or 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.

Understanding the diverse and nuanced factors such as taste preferences, styles, trends, health benefits, and personal preferences that influence each coffee purchase is essential. In the world of coffee, where customers are presented with a wide variety of options, including additional items for upselling, it becomes crucial to delve deeper into the reasons behind each purchase. Often, the items offered may not align with the individual customer's preferences, leading to the necessity for customer segmentation and subsequent Market Basket Analysis (MBA). Through segmentation, the aim is to tailor product recommendations to better match individual preferences, enhancing the overall customer experience and optimizing shop revenue by offering items that align with each customer's unique tastes and preferences.

## Research Problem

The primary focus of this research is to develop personalized product recommendations for multiple customer segments within coffee shops, with a particular emphasis on the point of purchase (till). While existing studies have explored customer segmentation through clustering methods, none have specifically applied market basket analysis post-segmentation to transactions in coffee shops. This research addresses the need for a comprehensive understanding of customer behaviour and preferences at the point of purchase, aiming to contribute valuable insights that enhance tailored product recommendations for different customer groups within the context of coffee shops.

## Research gap

Recommendation systems have become commonplace in various fields, playing a pivotal role in enhancing user experiences. These systems have been extensively explored in diverse domains, from e-commerce to streaming services. The common 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 field 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, the adopted methodology involves segmenting customers based on their gender and age groups. Subsequently, the Apriori algorithm is applied 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, highlighting the 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 attempt to answer this question started with the literature review. This phase will be divided into three segments: customer segmentation, market basket analysis, and transaction forecast. For each of the parts, multiple methodologies going to be examined across different industries. Additionally, interviews with field experts were conducted to gain real-world insights and a more practical perspective. It is essential to have proper and accurate view of coffee shop product recommendation settings and upselling style before proceeding with the analysis.

In the analysis and evaluation part, the most suitable customer segmentation methodology will be selected and applied to the data sets which will allow to divide customers into meaningful groups. Each of the groups will be used for Market Basket Analysis to create product recommendations. All recommendations will be evaluated against product recommendations received from a general group based on the pre-segmented data set. Results obtained in this part will be evaluated and discussed in the last part and possible answer to the research question will be stated.

# Literature review

The literature review started with detailed background check on the domain of market basket analysis and customer segmentation. Both were perform for general history knowledge, their application in different fields and how those methods were previously applied into coffee shop environment. For deeper understanding, specific challenges and opportunities unique to the coffee shop industry, four interviews were performed. Two with coffee shop managers (small local and branded at high street) and another two with experts in customer segmentation and product recommendation. Those interviews, despite different fields, were essential to obtaining real-life points of view for customer segmentation and product recommendation.

## Market Basket Analysis

Product recommendations become a very important part of the modern consumer experience. It is very rare that while shopping online, there are no recommendations for currently viewed item regardless to the field, particularly in e-commerce (Khushi Gupta, Kashyapi Shah, Ameya Kadam, 2023), retail, and digital platforms . Is it the grocery shopping where customers may be recommended a similar product if the searched product is out of stock, or a different toy in which potential customers could be interested as well, or even ski equipment while searching for a holiday in the Alps, recommendations are surrounding customers at their every step.

While product recommendations benefits not only the consumers by facilitating a more personalized and efficient shopping experience but also to businesses aiming to optimize sales (Wishma Samaraweera, Chekaprabha Waduge,Uma Indeewari Meththananda, 2016), Market Basket Analysis (MBA) is a holy grail for everyone. It 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 seek 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 upcoming of the large-scale transactional databases. The pioneering work in this field can be attributed to Agrawal and Srikant. Their research from 1994 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, allowing the extraction of meaningful insights from vast data sets.

This algorithm laid the groundwork for subsequent advancements in MBA methodologies and various algorithms have been developed to enhance efficiency and scalability in MBA applications.

FP-Growth (Frequent Pattern Growth) and Eclat, are equally powerful in finding the ‘best match’ for selected product (Oetama, 2024). All of them can leverage innovative techniques to identify patterns and associations within transactional datasets allowing MBA to extends beyond retail, finding applications in diverse sectors. In many domains, MBA provides valuable insights into consumer behaviour and preferences, contributing to informed decision-making processes.

The literature based on MBA in coffees shop is very limited and does not include customer segmentation. Founded research papers tests an Apriori model highlighting the need of the larges size of data set to gain meaningful results (Holy Meilani Amanda Ade Irma Amanda,Debi Setiawan, Liza Trisnawati, 2023). The lack of deep analytical work is evident and does not bring lots of information.

To bridge this gap, insights from interviews with coffee shop managers enrich the literature review. Chosen for their firsthand experience in working in the coffee shop, 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 based in the belief that their perspectives can provide a practical and deep understanding of the dynamics involved in customer interactions and upselling strategies within a coffee shop.

The interviews with the Coffee Shop Managers brought vast information on how products are recommended to the customers and what recommendations are based on. Different methods of upselling were mentioned but only ‘point of sale’ recommendations were discussed in detail emphasizing its revenue-boosting potential. The manager of a branded coffee shop notes that upselling can contribute to an increase of up to 80% if done correctly. The techniques employed in this process are highlighted as dynamic and evolving. This requires 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 based 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.

The situation is very much different in a local coffee shop where workers know the majority of their customers and product recommendations are seamless. Short, personal chat is added and additional purchases seem effortless.

Another aspect brought out in the interviews focuses on the effectiveness of the current upselling strategies. In branded shop, new staff members are reported to engage in upselling with an initial success rate ranging from 10% to 15%. Furthermore, priority in recommendation is given to the products with the highest margin. In local store, upselling products lean towards those with the highest current stock levels to mitigate potential waste.

None of the managers considered before customer segmentation prior upselling the additional item.

The second population targeted for the primary research consists of 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 methodologies and best practices employed in customer segmentation and recommendation system creation.

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

Firstly, stock levels emerge as a primary factor in influencing product recommendations. The expert strongly emphasises the critical role of maintaining optimal inventory levels to ensure that the recommended products are readily available for customers. This not only can improve store revenue but also contribute to seamless customer experiences.

The second factor standing behind product recommendations is supplier agreements and deals. Experts identified them as 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.

From the online side, customer behaviour and preferences, 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.

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.

Furthermore, the expert repeatedly highlights the role of data-driven insights in refining product recommendations. Leveraging advanced analytics and machine learning models, the company analyses vast datasets to identify patterns and trends in customer behaviour. This data-driven approach enables the system to make informed predictions and recommendations, enhancing its ability to understand and respond to dynamic customer preferences. Additionally, the integration of feedback loops, where customer responses to recommendations are monitored and incorporated into the learning process, further refines the system's accuracy over time. This iterative improvement based on real-time data ensures that the product recommendation system remains adaptive and responsive to the evolving needs and expectations of the customer base. The combination of strategic considerations, current supplier agreements, customer-centric insights, and data-driven methodologies forms a comprehensive approach employed by Company X to craft effective and personalized product recommendations within its clothing department.

An innovative idea presented by the expert at the end of the interview 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 flexible and aligned with the ever-evolving market landscape.

Field Expert 2, specializing in Company X's food department, similarly to previous expert, emphasized 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.

Predicting and responding to seasonal trends appear as integral components in the recommendation process. The expert emphasizes the importance of aligning product recommendations with seasonal shifts in customer preferences and demands which make sense. There is very low and steady demand for fruits across the year but as soon as the weather is starting to improve, demand is increasing. Multiple promotions and deals are crafted to boost sales but keep wastage at minimum.

This involves demand forecasting, utilizing historical data and market trends to predict customer expectations during different times of the year. The expert's insights emphasise the importance of a forward-looking approach to ensure that recommendations remain tailored to evolving seasonal dynamics and customer needs. Additionally, 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 mostly 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.

## Customer segmentation

Unlike the online textile and grocery shopping experience, where customer segmentation happens through the analysis of purchase history and browsing behaviour, coffee shops typically lack a refined 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 crucial decision that significantly influences the effectiveness of marketing strategies and overall business success. When the right segmentation model is applied, businesses can gain correct insights into their customer base, allowing for tailored and targeted approaches. This precision enables the delivery of personalized marketing campaigns, product recommendations, and services, at the same time, 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 various clustering methodologies were considered to determine 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).

This clustering algorithm is very efficient in 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 (Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, Jia Heming, 2023)(Dabbura, 2018) .

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 (Edy Umargono,Jatmiko Endro Suseno,S.K Vincensius Gunawan, 2020) and the Silhouette. 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 (Heti Mulyani,Ricak Agus Setiawan,Halimil Fathi, 2023).

The key aspect to remember is 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 multifunctionality and precision positioned it as a valuable tool for businesses seeking nuanced and effective customer segmentation.

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

Standing out from k-means clustering, this approach does not need a predetermined cluster count, presenting enhanced flexibility in its application. At its core, hierarchical clustering treats each data point as an individual cluster and systematically merges the closest clusters iteratively. This iterative process results in a dendrogram, a visual representation of the hierarchical arrangement shown in Figure …. Vertical lines within the dendrogram denote clusters, and the height of their merger signifies dissimilarity.

A diagram of clustering data

Description automatically generated

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 easily visible.

Regrettably, none of the conventional methods for customer segmentation are suitable for the specific requirements of the task at hand. Taking into account that customers must be allocated to the predefined segment at the time of the purchase, the segments must be very easy to understand. It is crucial that the coffee shop worker will be able to quickly and correctly select the correct group, otherwise recommended product might not be a perfect match and additional sales will not happen. For this reason, customers will be divided mechanically based on their gender and age. Only this can assure simplicity and effortless customer allocation in the area of coffee shop.

# 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 understand the dynamics between customer segmentation and product recommendations in the specific context of a coffee shop environment. The main goal is to discover and understand the effects of implementing a customer segmentation strategy on the recommendations offered to customers. This investigation seeks to understand how each of customer groups, 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.

A diagram of a diagram

Description automatically generated

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

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

In implementing the customer segmentation strategy, a pivotal consideration is the recognition of inherent variations in purchasing behaviours between different gender groups. Through the manual categorization based on gender, an initial distinction emerges, allowing for a nuanced exploration of preferences and trends within each segment. This segmentation approach acknowledges that the coffee consumption patterns of male and female customers can exhibit unique characteristics.

The examination of purchase patterns across gender and age groups yielded noteworthy findings. Among females, a substantial majority of purchases, constituting 27%, were attributed to the age group of 30-39. In contrast, the male demographic exhibited a distinctive trend, with 28% of all purchases being orchestrated by individuals aged 60 and above.

Diving deeper into how people make purchases, a specific preference showed among women aged 19-29. In this group, a significant number tended to buy two items during a transaction. This differs from the usual practice in other age groups, where most people prefer a single item per purchase. For men, regardless of their age, the common trend was to buy only one item per transaction. These insights shed light on the subtle differences in how people shop, offering a detailed understanding of the purchasing dynamics within the coffee shop context under investigation.

A screenshot of a computer screen

Description automatically generated

## Market basket analysis

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.

## Model selection

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, density, 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, the 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 that 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 the 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 modelling 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 that was considered FP-Growth, or Frequent Pattern Growth. Unlike the Apriori algorithm, FP-Growth employs a different approach to discover frequent patterns without the need for candidate generation. This makes it particularly efficient in handling large datasets and reduces the computational burden associated with generating and testing candidate itemsets.

The algorithm begins by constructing a data structure called an FP-tree (Frequent Pattern tree) from the given dataset. The FP-tree is a compact representation of the transactional database that preserves the frequency information of each item. The process involves scanning the dataset multiple times to identify frequent items and construct the tree. The nodes of the tree represent items, and the edges between nodes signify the order in which items appear in transactions.

Once the FP-tree is constructed, the algorithm recursively mines frequent itemsets by examining conditional pattern bases and constructing conditional FP-trees. The process involves selecting a frequent item as a starting point, creating a conditional pattern base by tracing back its occurrences in the original tree, and then recursively building a conditional FP-tree for each item in the base. This recursive approach efficiently captures the relationships between items and significantly reduces the search space.

One of the major advantages of FP-Growth is its ability to handle datasets with high dimensionality and sparsity efficiently. The FP-tree structure allows for a compact representation of the dataset, reducing the memory requirements compared to Apriori. Additionally, the algorithm excels in scenarios where the dataset has a large number of transactions but a relatively small number of distinct items.

Last algorithm was Eclat, short for Equivalence Class Clustering and Bottom-Up Lattice Traversal, which 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 algorithm begins by identifying individual items that meet a specified support threshold. It then extends these frequent items by considering combinations with other items in a depth-first manner. Eclat explores the lattice of itemsets, focusing on promising combinations before backtracking to explore other possibilities.

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

In the exploration of suitable algorithms for discovering frequent itemsets and association rules within the context of coffee shop product recommendation systems, the consideration extended beyond Apriori and Eclat to include FP-Growth. While Apriori and Eclat are established algorithms with distinct characteristics, FP-Growth offers an alternative approach that merits examination. However, in the specific scenario of a relatively small dataset encountered in the coffee shop context, the characteristics of FP-Growth did not align optimally with the analysis goals.

The Apriori algorithm was deemed suitable for this dataset due to its adaptability to diverse datasets, particularly accommodating variations in transaction lengths. The stepwise approach of Apriori, involving the iterative identification of frequent itemsets based on a minimum support threshold, resonates well with the dataset's characteristics. Moreover, Apriori's preference for datasets that can comfortably fit into memory makes it an apt choice for smaller datasets without concerns about computational efficiency, a pertinent consideration in the coffee shop scenario.

While Eclat, similar to Apriori, is a robust algorithm for market basket analysis, its efficiency is particularly pronounced in large transaction datasets. However, in the context of the coffee shop dataset, which was relatively small, the advantages offered by Eclat in terms of memory efficiency and lack of candidate generation were not as pronounced. The simplicity of parameter tuning in Eclat, primarily relying on a minimum support threshold, is user-friendly but did not provide a compelling advantage in the current analysis.

FP-Growth, renowned for its efficiency in handling large datasets, was not selected for this specific analysis. The decision was influenced by the dataset's modest size, where the initial construction of the FP-tree involving multiple passes over the data might not offer a substantial advantage. Additionally, the dataset's characteristics, including a predominance of single-item transactions, did not necessitate the advanced features of FP-Growth.

In summary, while Apriori and Eclat were closely considered and Apriori was chosen for its alignment with the dataset's size and characteristics, FP-Growth did not emerge as the optimal selection in this scenario. 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. Given that Apriori tends to exhibit strong performance in scenarios characterized by dense data, this choice was further substantiated.

## Application

Multiple market basket analyses (MBA) were conducted using the Apriori algorithm to gain diverse insights from the dataset. Firstly, an MBA was performed on the entire dataset, excluding transactions where a customer's ID appeared only once. This exclusion aimed to eliminate single-item purchases was crucial not only to prevent potential skewing of the results but also to ensure that the market basket analysis focused on transactions involving multiple items, where meaningful associations could be identified and analysed. The analysis was initially carried out with a minimum support of 0.01, focusing on pairs of products. However, to obtain more comprehensive recommendations, the minimum support was later reduced to 0.002.

This adjustment was made due to the dataset's high variety, containing more than 80 different products, and the relatively small number of transactions. The lower minimum support allowed for the identification of associations among a broader range of products, although confidence measures were not utilized as they did not yield meaningful recommendations given the dataset's characteristics.

The decision not to incorporate lift into the market basket analysis was made based on the size of the dataset. Given the relatively small dataset with a high number of different products (over 80) and a limited number of transactions, introducing lift as an additional measure was deemed impractical. Lift calculations require a sufficiently large dataset to provide meaningful insights into the strength of associations between items.

A graph with a chart of different types of food

Description automatically generated with medium confidence

The analysis of customer transactions unveils interesting patterns in product combinations, providing insights into what customers prefer and potential marketing strategies. Some product pairs show strong connections, indicated by high support percentages. For instance, Espresso shot often goes hand in hand with various syrups like Carmel, Hazelnut, Sugar-Free Vanilla, and Chocolate, with support ranging from 35.76% to 44.90%. Similarly, Latte and its variations (Latte Rg) have solid associations with syrup flavors, especially Carmel, Hazelnut, Sugar-Free Vanilla, and Chocolate syrups, showing support percentages between 44.10% and 48.47%. On the other hand, the pairing of Ginger Scone with Ouro Brasileiro shot stands out as an exception with an unusually high support percentage of 2.93%. These findings not only highlight common preferences for specific product combinations but also uncover unique and potentially influential associations for targeted promotions. It's crucial to note that different product pairs have varying levels of association strength, providing valuable insights for decisions related to product bundling, marketing campaigns, and inventory management.

In the second step, the dataset was segmented by gender into transactions made by females and males. Subsequent Market Basket Analyses (MBAs) were then performed separately for each gender, following the exclusion of transactions featuring only one purchased item. This strategic filtering aimed to focus the analyses on transactions involving multiple items, enhancing the ability to uncover nuanced patterns and preferences within the distinct customer segments of females and males. The utilization of a consistent minimum support threshold of 0.002 for both gender-specific and general analyses ensured a uniform and accurate basis for comparing the results across different segments.

A graph with a grid and a yellow arrow

Description automatically generated with medium confidence

Upon conducting Market Basket Analysis (MBA) on the entire dataset and subsequently segmenting it into female and male subsets, important patterns in consumer behaviour have come to light. The MBA results for the gender-specific datasets, after filtering transactions based on gender, revealed a more extensive list of product recommendations compared to the general MBA conducted on the full dataset. This suggests that genders exhibit distinct and meaningful preferences, influencing their associations with other products. The larger number of recommendations in the gender-specific analyses indicates a higher level of specificity and relevance in product pairings within these segments. It is reasonable to assume that customers within each gender group tend to have their own preferred products, and the associations uncovered in gender-specific MBAs are more tailored and meaningful for those respective groups.

Lastly, an age-group gender-based MBA was performed. The primary focus of this investigation was to recognise nuanced patterns of consumer behaviour within distinct age and gender groups.

The findings accentuate the importance of context-specific adjustments in parameterization, particularly when dealing with datasets of restrained scale, to ensure a sensible synthesis of accuracy and interpretability in the field of age-group gender-based MBA. Upon initial analysis, as Figure ….. shows, the absence of confidence thresholds led to an overwhelming number of results for Female all age-groups. However, through subsequent refinement, including the introduction of confidence thresholds, the outcomes were streamlined to a more manageable and interpretable set, highlighting the importance of considering confidence levels for a more refined and meaningful analysis.

Several graphs showing different types of data

Description automatically generated with medium confidence

With a minimum confidence of 0.5, very small set of results emerged, albeit with no outcomes for group 3. Reducing confidence to 0.3, showed in Figure …. , yielded only a few results, which proved insufficient for a meaningful outcome, particularly disadvantageous for a coffee shop seeking diverse product recommendations. At a confidence level of 0.2, the number of results increased, presenting more product recommendations; however, some products still had only one recommendation, notably visible in Figure … for group 3. Finally, a confidence of 0.1 was tested, producing a variety of results with multiple product recommendations, striking a balance between a lower confidence level and meaningful, helpful solutions to the problem at hand.

A group of graphs with text

Description automatically generated

A group of graphs with different colored dots

Description automatically generated

Upon closely investigating the dataset, it became apparent that the application of parameters led to all mined rules achieving a confidence level of 100%. While this showcased internal consistency, the inflated confidence levels were largely a result of the limited diversity within the dataset or the spread of specific transactions within distinct age and gender categories. This realization prompted a thoughtful examination of the interplay between dataset characteristics and parameter selection, highlighting the necessity for a well-balanced approach to extract meaningful insights from the association rules generated.

The findings showed the importance of tailoring parameterization to the specifics of the dataset, especially when dealing with datasets of modest scale. Through further refinement, introducing a confidence threshold of 0.1, Figure ….. , emerged as the right choice for the current problem and available dataset, resulting in a more nuanced and interpretable outcome in the domain of age-group gender-based MBA.

A collage of graphs with numbers and symbols

Description automatically generated with medium confidence

## 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 analising 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 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 vast 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 rare, 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.

### Forecasting transaction volume for the 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 regard, the initial step involved grouping rows based on unique transaction IDs, allowing for the consolidation of transactions and yielding valuable insights into the individual transactional behaviours of customers. Following this, a transformative process ensued, wherein the dataset was restructured by configuring the transaction date as the index, thereby establishing a daily frequency. This pivotal transformation not only organized the data by unique identifiers but also facilitated the representation of transactions as a time series. Such a structured format enables a nuanced exploration of daily patterns, seasonality, and overarching trends within the transactional data. Ultimately, the resulting dataset, seamlessly organized by ID and featuring a daily frequency, lays the groundwork for the subsequent application of time series forecasting models, promising heightened accuracy in discerning transactional dynamics.

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.

### 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 its 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.

1. 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%

### Models evaluation

The selection of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE) as evaluation metrics stems from their ability to provide comprehensive insights into the accuracy and precision of time series forecasting models. MAE represents the average absolute discrepancy between predicted and actual values, offering a straightforward measure of forecasting accuracy. MSE extends this evaluation by squaring the differences, emphasizing larger errors, making it particularly useful for penalizing significant deviations. Meanwhile, MAPE calculates the percentage-based error, expressing accuracy in a more interpretable manner. These metrics collectively offer a holistic assessment of forecasting performance by considering both the magnitude and direction of errors.

While other metrics like Root Mean Squared Error (RMSE) and AIC (Akaike Information Criterion) are commonly used, they were not deemed as suitable for this evaluation. RMSE shares similarities with MSE but involves taking the square root of the average squared differences, potentially amplifying the impact of larger errors. AIC, on the other hand, introduces a trade-off between model complexity and fit, favoring simpler models. However, in the context of time series forecasting, prioritizing simplicity over capturing intricate temporal and seasonal patterns could lead to suboptimal predictions. Therefore, the selection of MAE, MSE, and MAPE aligns with the objective of obtaining accurate and interpretable measures to assess the forecasting models' performance in the given scenario.

The evaluation of the forecasting models, including ARIMA, SARIMA, and ETS, reveals distinct performance characteristics based on the metrics of Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE). ARIMA exhibited a MAE of 58.08, MSE of 5189.49, and MAPE of 3.31%, indicating a relatively higher level of absolute error and percentage deviation from the actual values. SARIMA, on the other hand, demonstrated improved forecasting precision with a lower MAE of 31.36, MSE of 1826.29, and MAPE of 1.72%. ETS also showcased competitive results, striking a balance between accuracy and complexity, with a MAE of 39.20, MSE of 2794.95, and MAPE of 2.18%.

The variations in these metrics highlight the strengths and weaknesses of each model. SARIMA, with its capacity to capture both temporal and seasonal dependencies, emerges as the most accurate model, as evidenced by its lowest MAE and MSE values. ETS, while not outperforming SARIMA, still provides a balance between accuracy and simplicity. ARIMA, while exhibiting higher errors, may still offer utility in certain contexts or as a baseline model. The selection of these models is rationalized by their unique methodologies, with ARIMA focusing on temporal dependencies, SARIMA incorporating seasonality, and ETS leveraging exponential smoothing. The choice of SARIMA as the most suitable model is substantiated by its superior performance across the chosen evaluation metrics, reinforcing its efficacy for forecasting transaction volumes in the given context.

### Transaction forecast

The SARIMA model, configured with optimal hyperparameters ({'p': 0, 'd': 1, 'q': 1, 'P': 1, 'D': 0, 'Q': 1, 's': 14}), was employed to forecast the transaction volume for the upcoming 7 days. This tailored configuration, achieved through an exhaustive exploration of autoregressive, differencing, moving average, and seasonal orders, allowed the model to adeptly capture the intricate temporal and seasonal patterns inherent in the time series data. The model underwent empirical validation on a designated training set, and its predictive performance was subsequently assessed on the allocated test set. The forecasting results indicated a Mean Absolute Error (MAE) of 31.36, Mean Squared Error (MSE) of 1826.29, and Mean Absolute Percentage Error (MAPE) of 1.72%. These metrics signify a high level of precision in predicting transactional dynamics, reinforcing the efficacy of the SARIMA model for providing accurate forecasts of future transaction volumes in the specified time frame.

The forecasted values obtained from the SARIMA model for the next 7 days are instrumental in providing insights into the potential future trajectory of transaction volumes, especially when examined in conjunction with the historical dataset of transactions. The SARIMA model, characterized by its proficiency in capturing temporal and seasonal patterns, has been strategically applied with a rolling forecast approach. This methodology involves iteratively updating the model with new data and forecasting one step ahead, contributing to its adaptability to evolving trends.

A graph of a graph

Description automatically generated

Examining the forecasted values in relation to the actual transaction data from June 1, 2022, to June 6, 2022, reveals a nuanced perspective on the expected fluctuations in transaction volumes. Figure …. Shows the alignment between the forecasted and actual values underscores the model's effectiveness in adapting to historical nuances and providing a reasonable estimation of transaction volumes. The implementation of the rolling forecast method enhances the model's robustness by incorporating the most recent data, ensuring its relevance and accuracy in capturing evolving patterns.

This approach to forecasting becomes particularly valuable for anticipating potential peak sales days, identifying shifts in customer behavior, and facilitating informed decision-making in areas such as stock management and resource allocation. By leveraging the SARIMA model with a rolling forecast, the analysis not only considers historical trends but also dynamically adjusts to the evolving nature of the dataset, providing stakeholders with actionable insights for strategic planning and decision support.

# 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.

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.

# Conclusion

# References