

**United States International University-Africa**

**SPRING SEMESTER 2025**

**DSA-2040: DATA WAREHOUSING AND MINING**

**REPORT ON BANKING DATASET**

**INSTRUCTOR: DR. JAPHETH MURSI**

**GROUP 7**

**Valerian Murago – 669513 Signature: \_\_\_\_\_\_**

**Prince Muhindo**

**Signature: \_\_\_\_\_\_**

**Richard**

**Howard 668712  
 Signature: \_\_\_\_\_\_**

Table of Contents

[Table of Contents 2](#_Toc193965098)

[**1.0 INTRODUCTION** 3](#_Toc193965099)

[**1.1 Background** 4](#_Toc193965100)

[**1.2 Objectives** 4](#_Toc193965101)

[**1.3 Problem statement** 5](#_Toc193965102)

[**2.0 LITERATURE REVIEW** 6](#_Toc193965103)

[**2.1 Existing solutions** 6](#_Toc193965104)

[**2.2 Comparison** 9](#_Toc193965105)

[**2.3 Gaps and Opportunities** 11](#_Toc193965106)

[**3.0 VARIABLE AND DATA SELECTION** 12](#_Toc193965107)

[**3.1 Variable identification** 12](#_Toc193965108)

[**3.2 Data source** 13](#_Toc193965109)

[**3.3 Target variable** 13](#_Toc193965110)

[**4.0 DATA PROCESSING AND EXPLORATION** 14](#_Toc193965111)

[**4.1 Data cleaning** 14](#_Toc193965112)

[**4.2 Exploratory Data Analysis** 14](#_Toc193965113)

[**4.2.1 Comparison Between Marketing Data 1 and Marketing Data 2** 17](#_Toc193965114)

[**5.0 DATA MINING TECHNIQUES** 22](#_Toc193965115)

[**5.1 Association rule mining** 22](#_Toc193965116)

[**5.2 Implementation details** 22](#_Toc193965117)

[**6.0 RESULTS AND DISCUSSIONS** 31](#_Toc193965118)

[**6.1 Key findings** 31](#_Toc193965119)

[**6.1.1 Significance** 31](#_Toc193965120)

[**6.2 Business implications** 32](#_Toc193965121)

[**6.2.1 Practical Applications for the kenyan Market** 32](#_Toc193965122)

[**6.2.2 Strategic Recommendations For kenyan Banks** 32](#_Toc193965123)

[**6.2.3 Industry Impact** 32](#_Toc193965124)

[**6.3 Limitations** 32](#_Toc193965125)

[**7.0 CONCLUSION** 33](#_Toc193965126)

[**7.1 Conclusion** 33](#_Toc193965127)

[**7.2 Recomendations** 33](#_Toc193965128)

[**8.0 REFERENCES** 35](#_Toc193965129)

[9 Appendix 36](#_Toc193965130)

# **1.0 INTRODUCTION**

Banks should constantly improve their advertisements in the rapidly evolving monetary ecosystem of today to maximise consumer contact and service delivery. Traditional mass-marketing strategies are no longer sufficient in the financial industry due to the unprecedented development of cloud-based banking and increased competitors. Rather, banks want data-driven tactics that enable them to divide up their clientele, tailor their communications, and increase conversion rates. Financial institutions can predict how customers will react to marketing initiatives, spot changes in consumer behaviour, and create focused plans that increase general participation along with profitability by utilising customer segmentation approaches.

This study makes use of a Kaggle database that is openly accessible and includes 45,211 records that describe how customers interacted with a bank's marketing initiatives. The dataset includes a range of campaign-related, financial, as well as demographic details. While financial attributes include account balance, house loan status, and personal loan status, key demographic data includes age, marital status, education level, and employment sector. Marketing-related variables are also recorded in the database, such as the form of contact such as email or phone calls), the total amount of previous advertising efforts that a client participated in, the length of each interaction, and whether the customer ultimately subscribed to the financial product offered. These characteristics offer a thorough foundation for examining how consumers react to marketing tactics.

This study analyses two comparable datasets to guarantee a solid analysis, enabling more in-depth understanding and confirmation of trends found. Through comparison, hidden patterns that might not be immediately obvious in a particular dataset are revealed. To classify clients according to their likelihood to sign up for an insurance plan, this investigation uses complex segmentation techniques involving clustering, classification, and predictive modelling. Banking may apply to improve sales conversion rates, consolidate outreach tactics, and wisely allocate funds to high-potential customer segments.

Additionally, by examining the impact in different avenues of communication, the impact of past campaign history, and the function for client profiles in influencing monetary choices, the study investigates the efficiency of multiple promotional methods. By being aware of these elements, banks can take a more individualised, customer-focused approach to marketing, going beyond conventional methods. This can, therefore, boost reputation, improve customer happiness, and promote long-term company growth.

Banks can obtain practical insights via this study that will help them make better marketing campaign decisions. The study emphasises how crucial data-driven methods are to comprehending consumer behaviour and maximising the influence of banking processes. Banks may enhance their marketing success and offer consumers current and pertinent financial services by adopting algorithmic and statistical methodologies.

## **1.1 Background**

Customer segmentation is an essential practice in data mining that involves the identification and analysis of distinct groups of customers based on their shared characteristics and behaviours. This process is particularly significant for banks, as it empowers them to customize their marketing strategies to align with the unique needs and preferences of various customer segments. Consequently, marketing efforts become not only more targeted but also significantly more effective and efficient.

By gaining insights into different customer segments, banks can strategically allocate their resources, directing focus toward high-value segments that promise the greatest returns. Furthermore, the ability to pinpoint segments that are at risk of churn allows businesses to implement proactive measures aimed at retaining those customers. This targeted approach not only mitigates churn rates but also enhances customer lifetime value, ensuring a more loyal and satisfied customer base.

## **1.2 Objectives**

Utilized advanced clustering techniques, such as K-means and Hierarchical clustering, to identify distinct customer segments within the banking industry. This approach will enable the customization of targeted marketing campaigns designed to maximize both engagement and conversion rates across these segments.

To achieve this, two comprehensive datasets sourced from Kaggle will be meticulously compared. This comparison aims to highlight similarities, uncover potential relationships between the datasets, and identify key differences that may inform strategic decisions.

Data processing will be carried out using both Python and R programming languages, allowing for in-depth exploratory data analysis. Furthermore, Power BI will be employed to create visually engaging and insightful dashboards, facilitating a clearer understanding of the data landscape.

In addition, association rule mining techniques—including Apriori and FP-Growth—alongside the K-means clustering method will be implemented through relevant R packages. Upon completion of this robust analytical process, valuable insights will be extracted and discussed in detail, paving the way for enhanced strategies within the banking sector.

## **1.3 Problem statement**

How can the bank effectively segment its customers by analysing their demographics, financial behaviours, and interactions with marketing efforts to enhance targeted marketing strategies? By exploring various variables such as age, occupation, and marital status, the bank can develop more nuanced and strategic campaigns tailored to specific customer groups. This targeted approach is expected to increase engagement and profitability from customers who participate in the campaigns. The primary aim of this research is to uncover valuable insights into customer trends and behaviours, allowing for the identification of distinct segments and clusters that can be utilized for more personalized marketing initiatives.

# **2.0 LITERATURE REVIEW**

## **2.1 Existing solutions**

In the modern digital landscape, machine learning has emerged as an innovative and effective approach to customer segmentation. A comprehensive study by Ranjan & Srivastava (2022) highlighted that the foundation of customer segmentation relies on several key factors.

Behavioural Data, understanding customer behaviours, such as shopping habits, product preferences, and engagement patterns. Psychographic analyses customers' lifestyles, values, interests, and attitudes, which provides deeper insights into their motivations. Demographics and traditional metrics such as age, gender, income level, and education, help in categorizing the customer base. Geographic information considering the customers' physical locations can influence their buying patterns and preferences.Further research had a common theme of utilizing unsupervised learning models for effective segmentation, with K-means and Hierarchical Clustering emerging as popular methodologies to achieve optimal results.

In a parallel investigation, Gomes & Meisen (2023) noted a significant shift in customer behaviour post-COVID-19 pandemic. Their findings indicated that consumers have become much more receptive to online campaigns, a change driven by increased technological literacy and an expanded comfort level with digital platforms. This evolution allows marketing efforts to reach a broader audience and fosters greater interaction.

The Recency, Frequency, and Monetary Value (RFM) model has been widely recognized across various research papers as a primary tool for analyzing customer behaviour and segmentation. RFM assesses customers based on three pivotal criteria.

Recency refers to when a customer last made a transaction. Frequency indicates how often a customer makes purchases within a specified timeframe. Monetary value is the total amount that a customer spends.

This model's significance cannot be overstated, as it provides a structured framework for understanding customer engagement and prioritising target audiences effectively.

A noteworthy trend in data collection identified by Gomes & Meisen (2023) is the reliance on manual feature selection. Researchers have observed a lack of a standardized approach to data segmentation, emphasising the necessity for methods to be tailored to the specific characteristics of the analysed dataset.

Moreover, research by Mihova & Pavlov (2018) revealed that critical factors influencing customer segmentation include geographic location, educational attainment, and default behaviour in financial transactions.

Additionally, a burgeoning trend in customer segmentation is the application of artificial intelligence, which, while still in its nascent stages, is gaining traction. Dubbed ‘AI marketing’ by Raiter (2021), this emerging methodology heavily relies on K-means clustering, suggesting that this approach may well be among the most effective strategies for customer clustering in the evolving marketplace.

Overall, the exploration of these dimensions in customer segmentation research points to a dynamic landscape that continues to evolve with technological advancements and changing consumer behaviours.

The utilization of transactional data for strategic decision-making has gained significant traction in various industries. Businesses leverage data mining techniques to uncover hidden patterns in customers, optimize marketing campaigns, and enhance customer experiences. Several existing behaviour approaches have been developed to analyze transactional data, two of the most prominent being Market Basket Analysis and Hybrid Recommender Systems.

Market Basket Analysis (MBA)

One of the widely used techniques in transactional data analysis is Market Basket Analysis (MBA), which employs association rule mining to discover purchasing patterns. Algorithms such as Apriori, FP-Growth, and Eclat facilitate the identification of frequently co-occurring items within transactional datasets.

The Apriori Algorithm, introduced by Agrawal et al. (1994), this method systematically generates candidate itemsets and prunes infrequent combinations. It follows a breadth-first search strategy, making it useful in discovering hidden relationships within large datasets. However, its performance decreases with increasing dataset size, requiring optimization techniques such as Hash-based structures.

FP-Growth Algorithm this method improves upon Apriori by compressing datasets into tree structures (Frequent Pattern Trees) to avoid generating redundant candidate item sets. This allows for faster rule extraction, making it efficient for large datasets.

Eclat Algorithm: Unlike Apriori and FP-Growth, Eclat employs a depth-first search strategy, which enables faster itemset discovery for dense datasets. It is particularly suitable for analyzing financial transactions in banking, where frequent small purchases occur.

These methods have been successfully applied in retail (e.g., grocery stores analyzing customer baskets), banking (e.g., identifying co-purchased financial products), and healthcare (e.g., detecting patterns in medication purchases) to optimize sales strategies and improve customer engagement.

Hybrid recommender systems combine multiple filtering approaches, such as collaborative filtering and content-based filtering, to enhance predictive accuracy.

Collaborative Filtering this approach predicts customer preferences based on the historical behaviours of similar users. It can be implemented using either memory-based techniques (user-user or item-item similarity matrices) or model-based techniques (e.g., matrix factorization and deep learning models).

Content-based filtering, unlike collaborative filtering, this technique relies on item attributes (e.g., product descriptions and transaction amounts) to recommend similar items to users based on past preferences.

Hybrid Systems integrate both collaborative and content-based filtering to overcome individual limitations. For example, Netflix and Amazon use hybrid recommenders to personalize product recommendations while considering both customer purchase history and item attributes.

Research by Bobadilla et al.. (2013) highlights that hybrid models outperform standalone methods in recommendation accuracy, particularly when dealing with sparse data. These systems are widely implemented in e-commerce, streaming services, and banking applications, where personalized offerings drive customer retention.

## **2.2 Comparison**

In the realm of data analysis, K-means clustering has emerged as the most widely adopted method for segmentation and discretization in numerous research studies. Its popularity can be attributed to its relative simplicity and effectiveness in partitioning data into distinct clusters based on similarities. While other methodologies, such as hierarchical clustering, do exist, K-means has often been regarded as the gold standard in various academic papers, serving as a benchmark against which other techniques are measured.

In recent years, the application of modelling techniques has surfaced as a promising, albeit less prevalent, approach within this domain. These models, while primarily focused on grouping data, tend to lack the extensive application found in traditional segmentation methods like K-means. Nevertheless, they offer an innovative perspective that adds value to data analysis.

The intersection of artificial intelligence (AI) and clustering methodologies marks a significant advancement in this field. Although AI segmentation and traditional models both revolve around the concept of categorizing data based on critical variables such as age, education, and default behaviour, their purposes diverge. AI segmentation goes beyond mere grouping; it employs sophisticated algorithms to perform deeper analytical processes. This allows researchers to extract richer insights and uncover intricate patterns within the data, leading to more informed decision-making and strategic planning.

Overall, while K-means remains a cornerstone of segmentation methodologies, the evolving landscape of AI and modelling techniques introduces exciting possibilities for enhanced data analysis and interpretation.

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Feature** | **Market Basket Analysis (MBA)** | **Hybrid Recommender Systems** | |
| |  |  |  | | --- | --- | --- | | **Methodology** | Association rule mining (e.g., Apriori, FP-Growth, Eclat) | Combination of collaborative filtering and content-based filtering | |
| |  |  |  | | --- | --- | --- | | **Output** | Association rules (e.g., "Customers who buy A also buy B") | Personalized recommendations (e.g., "You may like this item") | |
| |  |  |  | | --- | --- | --- | | **Use Case** | Identifying cross-selling and upselling opportunities | Providing personalized product/service recommendations | |
| |  |  |  | | --- | --- | --- | | **Scalability** | Computationally expensive for large datasets | Requires high computational power and complex models | |
| |  |  |  | | --- | --- | --- | | **Interpretability** | High (clear and explainable association rules) | Low (black-box models that are harder to interpret) | |

Although an MBA offers clear insights into consumer buying trends, it is not very good at offering tailored advice. On the reverse hand, mixed recommender technologies provide tailored solutions depending on user choices and actions; however, they frequently lack interpretability and necessitate a large amount of data processing.

## **2.3 Gaps and Opportunities**

Due to model training not being used as much, we can use it as a means to improve accuracy while still combining it with RFM for segmentation. Models for segmentation and AI for analysis should not be used separately but instead interchangeably, depending on the aim of the research.

Despite advancements in transactional data mining, several challenges remain:

Scalability Issues: Market Basket Analysis struggles with computational complexity as dataset size increases, requiring efficient pruning and parallel processing.

Cold-Start Problem in Recommender Systems: Hybrid recommenders require sufficient historical data, making recommending products to new customers difficult.

Explainability and Interpretability: While an MBA is interpretable, recommender systems often act as black-box models, making it difficult for businesses to understand recommendation logic.

Threshold Sensitivity: Association rule mining requires predefined minimum support and confidence thresholds, which may not always generalize well across datasets.

This project aims to address these gaps by incorporating data discretization and optimized thresholding mechanisms to enhance rule discovery. Exploring hybrid approaches that combine Market Basket Analysis with predictive analytics to improve business insights.

# **3.0 VARIABLE AND DATA SELECTION**

The variables selected for analysis were specifically those that were present in both datasets and were deemed to hold significant relevance based on prior research. For instance, demographic factors like age and educational attainment are key elements that have been shown in previous literature to influence the categorization of customers. Age can affect consumer behaviour and purchasing patterns, while education level may correlate with income stability and product preferences. These insights reflect the importance of incorporating such variables to gain a comprehensive understanding of customer segmentation and behaviour in our analysis.

## **3.1 Variable identification**

The dataset comprises several variables. Age is represented as a numeric value indicating the individual's age. Job Type includes a categorical classification of employment status, with options such as "admin.," "unknown," "unemployed," "management," "housemaid," "entrepreneur," "student," "blue-collar," "self-employed," "retired," "technician," and "services." The Marital Status variable categorizes individuals as "married," "divorced," or "single," with "divorced" encompassing both divorced and widowed individuals. Education Level is also measured categorically, including "unknown," "primary," "secondary," and "tertiary." The Default variable is a binary indicator of credit status, with values of "yes" or "no" to denote if the individual is in default. Balance represents the average yearly account balance in euros as a numeric value. Additionally, the dataset includes binary indicators for whether an individual has a Housing Loan and a Personal Loan, with both options reflecting "yes" or "no" responses. The Contact Type is categorized as "unknown," "telephone," or "cellular," while the Last Contact Month is recorded categorically from "Jan" to "Dec.". Lastly, Contact Duration captures the duration of the last contact in seconds as a numeric value.

Other attributes are

The campaign tracks the total number of contacts made during this specific campaign for a client, including the most recent interaction (numeric). Additionally, the variable Pdays represents the number of days that have passed since the client was last contacted in a previous campaign, where a value of -1 indicates that the client has not been previously contacted (numeric). The Previous variable reflects the number of contacts made before this current campaign for the client (numeric). Another important aspect is Poutcome, which denotes the outcome of the previous marketing campaign and can take on values such as "unknown," "other," "failure," or "success" (categorical). Finally, the output variable, denoted as y, seeks to determine whether the client has subscribed to a term deposit, represented as a binary outcome of "yes" or "no."

## **3.2 Data source**

The data sets were sourced from Kaggle, a well-known provider of diverse datasets. The two selected data sets were carefully chosen for their similarities in structure, consistent formatting methods, and comparable size, each containing a significant number of records.

These data sets pertain to direct marketing initiatives conducted by a Portuguese banking institution. The marketing efforts predominantly involved phone calls, during which representatives often had to contact the same client multiple times to evaluate whether the client would subscribe to a bank term deposit, indicated as 'yes' for acceptance or 'no' for rejection.

The first data set comprises a total of 21 columns and contains 41118 rows, while the second data set includes 17 columns and 45211 rows. Notably, there are 15 columns shared between the two sets, containing symmetric data that allows for meaningful analysis. Consequently, these 15 columns were selected as the focus of our analysis, providing a robust foundation for understanding client behaviours and improving marketing strategies.

## **3.3 Target variable**

Our target variables are the following, Age, job, marital, education, default, housing, loan, contact, month, day of week, duration, campaign, pdays, previous, poutcome, y , As already stated before, these will be the chosen variables used for analysis. They contain qualitative, quantitative, and boolean data that can be analyzed using the K-means method as well as model training.

# **4.0 DATA PROCESSING AND EXPLORATION**

Data cleaning is essential when starting any data analysis. It enhances the accuracy of calculations and prepares the data for further breakdown and manipulation. This process includes actions such as removing null values and duplicates, as well as transforming the data to meet specific formats. It forms the foundational basis for all exploratory data analysis.

## **4.1 Data cleaning**

Data cleaning was performed using VS Code with the Data Wrangler extension due to its intuitive interface and comprehensive suite of data preprocessing functions. The techniques applied included eliminating duplicate rows, removing columns that were not common across both datasets, and standardising data types for rows to ensure consistency throughout the analysis.

## **4.2 Exploratory Data Analysis**

The first step was performing descriptive analysis on both data sets.

We first checked for data types and confirmed that they were equal in the columns that we were comparing. After that, we got a basic descriptive analysis of central tendencies.

Banking Marketing data 1 - Descriptive statistics for numerical columns:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Age | duration | campaign | pdays | previous |
| count | 41176.00000 | 41176.00000 | 41176.00000 | 41176.00000 | 41176.00000 |
| mean | 40.02380 | 258.315815 | 2.567879 | 962.464810 | 0.173013 |
| std | 10.42068 | 259.305321 | 2.770318 | 186.937102 | 0.494964 |
| min | 17.00000 | 0.000000 | 1.000000 | 0.000000 | 0.000000 |
| 25% | 32.00000 | 102.000000 | 1.000000 | 999.000000 | 0.000000 |
| 50% | 38.00000 | 180.000000 | 2.000000 | 999.000000 | 0.000000 |
| 75% | 47.00000 | 319.000000 | 3.000000 | 999.000000 | 0.000000 |
| max | 98.00000 | 4918.000000 | 56.000000 | 999.000000 | 7.000000 |

**Figure 4.1** **Descriptive statistics for numerical columns**

The following insights in figure 4.1 are based on the Banking Marketing Data, which captures customer interactions in a financial institution’s marketing campaigns. The Age variable shows that most individuals are around 40 years old, with ages ranging from 17 to 98 years. This suggests a diverse customer base, where different age groups may have varying financial needs. The Duration variable, which represents the length of customer interactions, has a high variation. The average duration is 258 seconds, but some interactions lasted as long as 4918 seconds, indicating that while most calls are short, some take much longer.

The Campaign variable shows that most customers were contacted around 2 times, but some were contacted as many as 56 times, suggesting potential over-contacting, which may impact customer experience. The Pdays variable is highly skewed, with most values being 999, meaning that many customers had not been contacted before. This highlights the need for better customer engagement tracking. The Previous variable, which tracks past interactions, shows that most customers had no prior contacts, but a few had been reached up to 7 times, indicating a mix of new and returning customers.

Marketing data 2 - Descriptive statistics for numerical columns:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age | duration | campaign | pdays | previous | Balance |
| count | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 | 45211.000000 |  |
| mean | 40.936210 | 258.163080 | 2.763841 | 40.197828 | 40.197828 |  |
| std | 10.618762 | 257.527812 | 3.098021 | 100.128746 | 2.303441 |  |
| min | 18.000000 | 0.000000 | 1.000000 | -1.000000 | 0.000000 |  |
| 25% | 33.000000 | 103.000000 | 1.000000 | -1.000000 | 0.000000 |  |
| 50% | 39.000000 | 180.000000 | 2.000000 | -1.000000 | 0.000000 |  |
| 75% | 48.00000 | 319.000000 | 3.000000 | -1.000000 | 0.000000 |  |
| max | 95.00000 | 4918.000000 | 63.000000 | 871.000000 | 275.000000 |  |

**Figure 4.2 Descriptive statistics for numerical columns**

The following insights in Figure 4.2 are based on Marketing Data 2, which captures key numerical variables related to customer interactions and financial behaviors in a banking campaign. The Age variable shows that the average customer is around 40 years old, with ages ranging from 18 to 95 years. This suggests a wide age distribution, meaning different customer segments may have unique financial needs. The Duration variable, representing call length, has a high variation, with an average of 258 seconds, but some calls lasted as long as 4918 seconds, indicating that while most calls are short, a few take significantly longer.

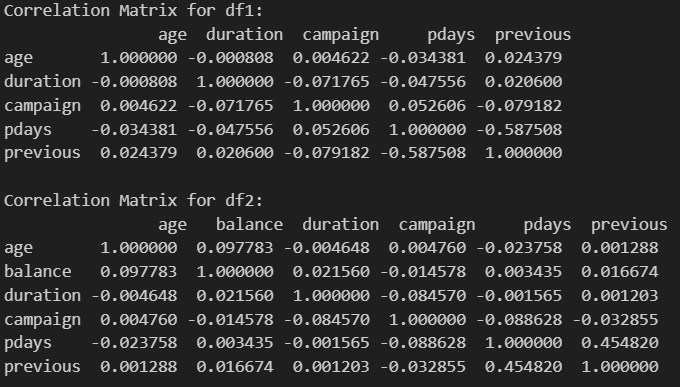
The Campaign variable shows that most customers were contacted around 2 times, but some received up to 63 calls, highlighting potential over-contacting, which may lead to reduced engagement. The Pdays variable, which measures the number of days since the last contact, has a high concentration of -1 values, meaning many customers had never been contacted before. This suggests a significant portion of first-time engagements in the dataset. The Previous variable, tracking the number of past contacts, shows that most customers had zero previous interactions, but some had up to 275 contacts, indicating a mix of new and frequently engaged customers. Lastly, the Balance variable shows varying financial statuses, with some customers having negative balances, suggesting financial difficulties, while others have much higher savings.

### **4.2.1 Comparison Between Marketing Data 1 and Marketing Data 2**

From the previous analysis, we notice that both Marketing Data 1 and Marketing Data 2 have similar trends in customer age and call duration. In both datasets, the average customer is around 40 years old, and most calls last about 258 seconds; However, Marketing Data 2 shows that customers were contacted more frequently than in Marketing Data 1. Some customers received up to 63 calls, compared to a maximum of 56 calls in Marketing Data 1. This suggests that follow-ups were more aggressive in the second dataset.

Another key difference is in past customer interactions. In Marketing Data 1, many customers had missing values (999 in Pdays), while Marketing Data 2 used -1 to indicate no prior contact. Additionally, in Marketing Data 2, some customers had been contacted up to 275 times, compared to only 7 times in Marketing Data 1. Overall, Marketing Data 2 includes more repeat interactions, meaning that banks need to carefully manage their marketing strategies to avoid over-contacting customers, which could reduce engagement and interest.

A correlation matrix was made for further analysis to guide how visuals would be made.



**Figure 4.3 Correlation matrix of the datasets.**

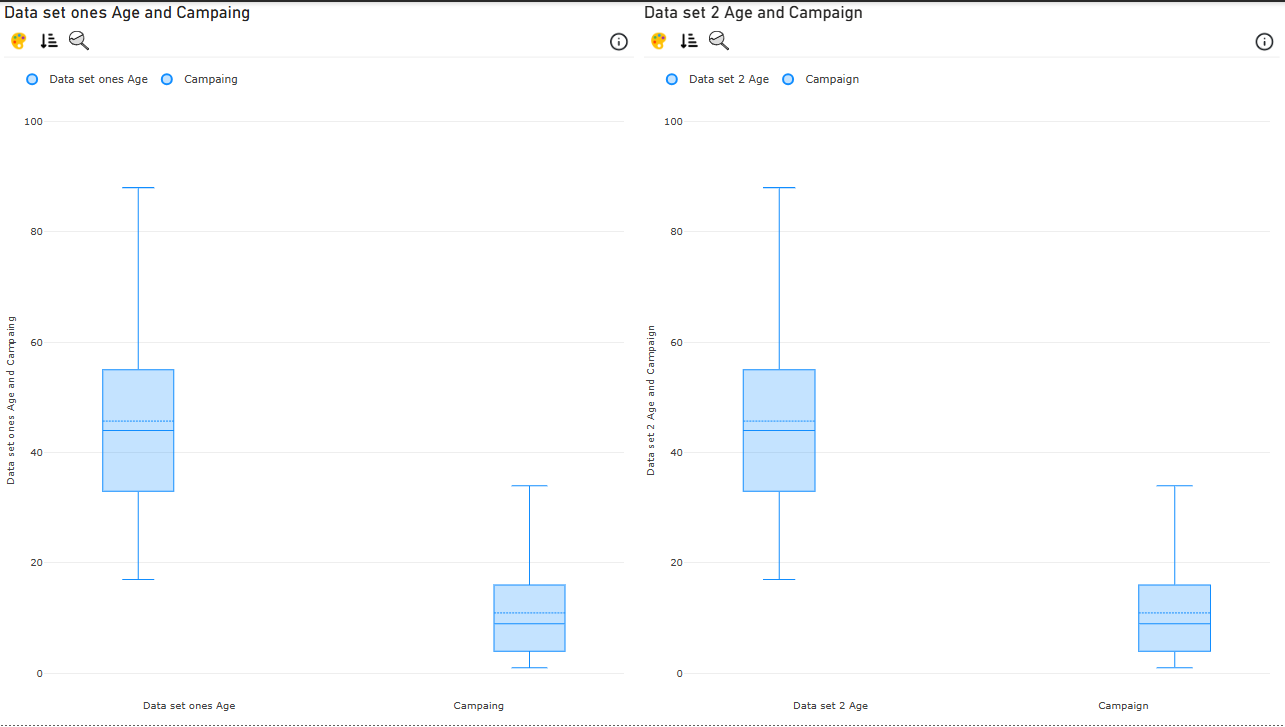
From the correlation matrices in Figure 4.3, we can analyze the relationships between different numerical variables in df1 and df2.

In df1, most variables have very weak correlations, meaning there are no strong relationships between them. Age has almost no correlation with duration, campaign, pdays, or previous, indicating that older or younger customers do not significantly differ in their engagement levels. Campaign has a slight negative correlation with duration (-0.0718), suggesting that customers who were contacted more times had slightly shorter conversations. The strongest correlation is between pdays and previous (-0.5875), showing that the number of past contacts is closely tied to how long it has been since the last contact.

In df2, the trends are similar, with mostly weak correlations. Balance has a small positive correlation with age (0.0978), meaning older customers tend to have slightly higher account balances. Campaign and day (0.1625) have a small positive correlation, indicating that customers contacted more frequently are more likely to be reached on certain days. Pdays and previous (0.4548) have a moderately strong correlation, like df1, confirming that customers with a history of multiple interactions are more likely to have been contacted recently. We also notice that both df1 and df2 show weak correlations between most variables, meaning that customer interactions and financial behaviors are largely independent. The strongest relationship in both datasets is between pdays and previous, but in df1, this correlation is negative (-0.5875), while in df2, it is positive (0.4548). This suggests differences in how past interactions are recorded.

Box plots were made for numerical data to visualize it and not outliers

When it came to age, the outliers ages of older persons were noticed especially from the age of 68 and above for both datasets.



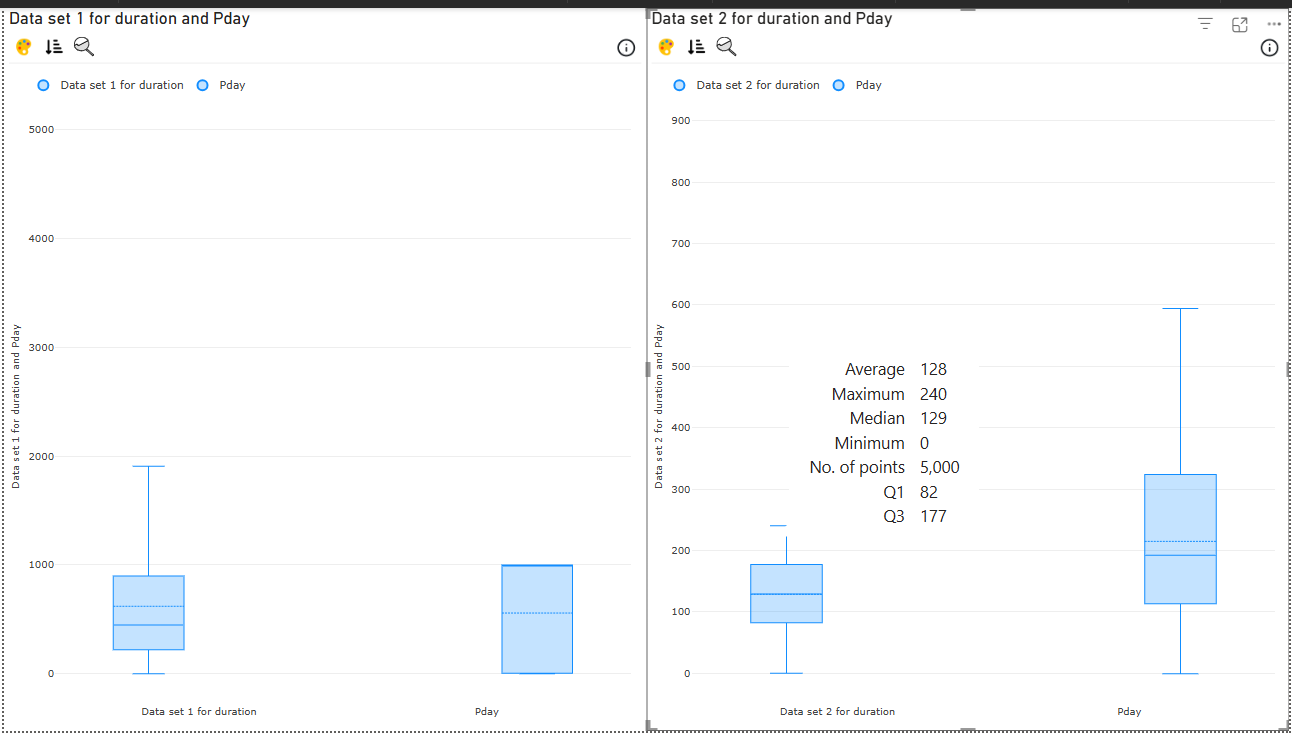
**Figure 4.1 Boxplots of Age, Campaign of both datasets**

In Dataset 1, the Age variable has a wide range, spanning from around 20 to nearly 100 years, with a median age in the mid-40s. The interquartile range (IQR) is balanced, suggesting a normal distribution of customer ages without significant skewness. Similarly, in Dataset 2, the age distribution follows the same pattern, with a similar range and median, indicating that the customer base remains stable across both datasets.

The Campaign variable, which represents the number of times a customer was contacted, is highly skewed in both datasets. Most customers received very few calls, but a small number of customers were contacted excessively, as indicated by the presence of extreme outliers. However, Dataset 2 appears to have a slightly higher median number of contacts, suggesting that customers in this dataset experienced more frequent follow-ups compared to those in Dataset 1.

We notice the presence of extreme outliers in the Campaign variable for both datasets. This means that while most customers were contacted only a few times, some were contacted an unusually high number of times, which could indicate over-contacting. Such a strategy may lead to customer dissatisfaction or reduced engagement, highlighting the importance of optimizing follow-up efforts to ensure efficient and effective marketing strategies. However, with data like Pdays, there were no outlier data points that would lead to the skewness of the data in the first data set, unlike with the second data set, which had a greater skewness.

Box plot of durations and pdays for datasets 1 and 2.



**Figure 4.2 Box plot of durations and pdays for datasets 1 and 2**

Here, we can notice that there is a difference in the whiskers of the boxplots for Pdays in the first and second data sets. Meaning that there is something different in it. From the analysis, we observe key differences in call duration (Duration) and follow-up timing (Pdays) between the two datasets. Dataset 1 has a wider spread in call durations, with some extreme outliers, meaning some calls lasted significantly longer than others. In contrast, Dataset 2 shows shorter and more consistent call durations, suggesting a more structured and efficient approach to customer interactions.

Similarly, follow-up timing in Dataset 1 is inconsistent, with some customers experiencing long delays before being re-contacted, as indicated by a broad distribution in Pdays. However, Dataset 2 has a more standardized approach, with follow-ups occurring within a narrower and more predictable timeframe.

Overall, Dataset 2 appears more optimized, with efficient call handling and systematic follow-ups, which can improve customer engagement and marketing effectiveness. On the other hand, Dataset 1's irregularities in call duration and follow-up scheduling may negatively impact customer experience, highlighting the need for a more structured approach to outreach efforts.

# **5.0 DATA MINING TECHNIQUES**

Data mining refers to the process of extracting valuable insights and identifying patterns within extensive datasets. This process employs methodologies that lie at the intersection of machine learning, statistics, and database systems. In this report, we will focus on segmentation and clustering techniques, specifically utilizing the K-means algorithm, along with discretization when necessary, to develop a comprehensive understanding of the two datasets under consideration.

## **5.1 Association rule mining**

Discretization was done on the age, campaign, and Pdays columns.

For the columns, k means discretization was used as it automatically adapts to the dataset data and was the one used in previous studies yielding the best results.

The minimum support chosen was 0.5 for confidence was 0.7.

## **5.2 Implementation details**

Necessary libraries were loaded, and in our case, we worked with the following: CSV, NumPy, and Pandas were used for data handling and manipulation, allowing efficient reading, processing, and structuring of large datasets. Matplotlib and Seaborn helped in data visualization, making it easier to analyze trends and patterns through graphs and plots. Apyori, MLxtend’s Apriori, and FPGrowth were utilized for association rule mining, helping to identify relationships between different variables in large datasets. PyCaret was used for automated machine learning (AutoML), simplifying model training and evaluation. KMeans from Scikit-Learn was applied for clustering, grouping similar data points to uncover patterns in customer behavior. These libraries together enabled effective data preprocessing, visualization, pattern discovery, and machine learning model development. For optimal use and to prevent version discourse, use Python 3.11.7

Below is the libraries that we used for the analysis and data mining.

import csv

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import apyori

from mlxtend.frequent\_patterns import apriori, association\_rules

import pycaret

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from mlxtend.frequent\_patterns import fpgrowth

from mlxtend.frequent\_patterns import association\_rules

**Figure 5.1**

Data preprocessing and clustering steps were applied in our case: Handling missing values: The datasets df1 and df2 were cleaned by removing any missing values in the 'age' column, ensuring data consistency for clustering. Data transformation: The 'age' column was converted into a NumPy array and reshaped for compatibility with the clustering algorithm.

Clustering with KMeans: The KMeans algorithm was applied with three clusters (n clusters=3), grouping individuals based on their ages. This helps to identify distinct customer segments based on age patterns in both datasets. Results assignment: The cluster labels were assigned to a new column, 'age clusters', in df1 and df2, allowing for further analysis of the segmented age groups. These steps ensured clean data, proper transformation, and effective segmentation, enabling better insights into customer demographics and behavior patterns.

As seen below are the steps that we took to ensure the data was cleaned thoroughy.

# Ensure no missing values in 'age' column

df1 = df1.dropna(subset=['age'])

df2 = df2.dropna(subset=['age'])

# Convert 'age' column to numpy array for clustering

ages\_df1 = np.array(df1['age']).reshape(-1, 1)

ages\_df2 = np.array(df2['age']).reshape(-1, 1)

# Apply KMeans clustering

kmeans = KMeans(n\_clusters=3, random\_state=42)

df1['age\_clusters'] = kmeans.fit\_predict(ages\_df1)

df2['age\_clusters'] = kmeans.fit\_predict(ages\_df2)

print(df1)

print(df2)

**Figure 5.2**

To identify meaningful patterns in the dataset, we applied FP-Growth, a frequent pattern mining algorithm, to discover item associations among categorical features. The df2 dataset was first transformed into a one-hot encoded format, converting categorical variables such as job, marital status, and education into binary columns (0s and 1s). Only binary columns were retained to ensure compatibility with the FP-Growth algorithm.

For frequent itemset mining, FP-Growth was applied with a minimum support threshold of 0.5, ensuring that only commonly occurring patterns were considered. To further analyze these patterns, we performed association rule generation using an Apriori-based approach, extracting rules with confidence levels of at least 0.7. These rules helped identify strong relationships between different customer attributes, providing insights into support, lift, and confidence, which are essential for understanding customer behaviors and preferences.

# Convert df2 to a one-hot encoded DataFrame

df2\_one\_hot = pd.get\_dummies(df2, columns=['job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day', 'poutcome', 'y'])

# Retain only binary columns for FP-Growth

binary\_columns = df2\_one\_hot.columns[df2\_one\_hot.isin([0, 1]).all()] # Select columns with only 0/1 values

df2\_binary = df2\_one\_hot[binary\_columns]

# Apply FP-Growth to find frequent itemsets

frequent\_itemsets = fpgrowth(df2\_binary, min\_support=0.5, use\_colnames=True)

# Display frequent itemsets

print("Frequent Itemsets:")

print(frequent\_itemsets)

# Generate association rules

rules = association\_rules(frequent\_itemsets, metric="confidence", min\_threshold=0.7)

# Display association rules

print("\nAssociation Rules:")

print(rules)

**Figure 5.3**

This provides us with the Support, Lift and Confidence of all columns present.

From there we begin to create visuals to compare to the original visuals created. Boxplots were created to analyze the distribution of age within different clusters, helping to visualize how customer age is segmented across the datasets. In Dataset 1 (df1), a boxplot was generated for each age cluster, showing how customer ages are distributed within different segments. This analysis highlights differences in age distribution, making it easier to identify age-based segmentation patterns and understand how customers are grouped.

Similarly, in Dataset 2 (df2), boxplots were created to compare age distributions across clusters. This visualization helps determine whether Dataset 2 follows a similar segmentation pattern as Dataset 1, allowing for a comparison of customer segmentation approaches. These visualizations provide insights into how well KMeans clustering captured age-based groupings and whether age plays a significant role in defining different customer segments.

This is the creation of a boxplot for each age cluster.

# Create a boxplot for each age cluster

plt.figure(figsize=(15, 10))

for i, col in enumerate(df1['age\_clusters'].unique()): # Iterate over unique age clusters

plt.subplot(2, 3, i + 1)

sns.boxplot(x=df1['age\_clusters'], y=df1['age'], color='blue') # Create a boxplot for each cluster

plt.title(f'Boxplot of Age Cluster {col}')

plt.tight\_layout()

plt.show()

# Create a boxplot for each age cluster

plt.figure(figsize=(15, 10))

for i, col in enumerate(df2['age\_clusters'].unique()): # Iterate over unique age clusters

plt.subplot(2, 3, i + 1)

sns.boxplot(x=df2['age\_clusters'], y=df2['age'], color='blue') # Create a boxplot for each cluster

plt.title(f'Boxplot of Age Cluster {col}')

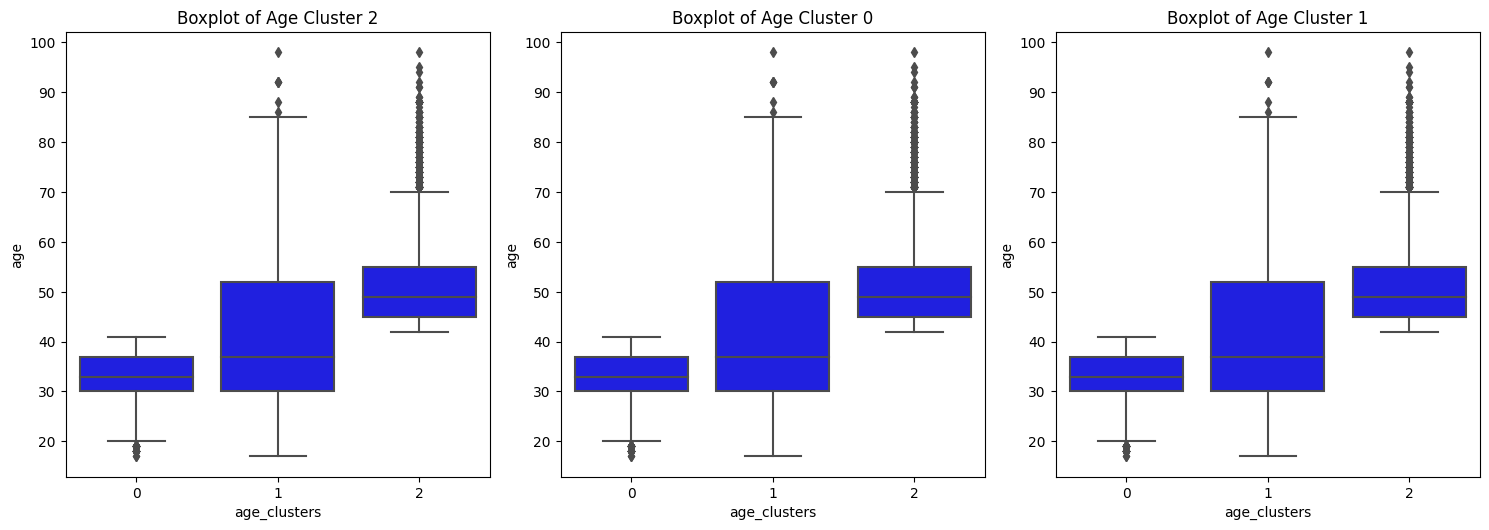
plt.tight\_layout()

plt.show()

**Figure 5.4**

After that we start creating visuals to show relationships between different columns. To further explore customer segmentation, we created multiple visual comparisons between age, job type, marital status, and follow-up timing (Pdays) to identify patterns in customer demographics and engagement. For age distribution by job type, a boxplot was generated in df1, comparing age distributions across different job categories. This helps determine whether certain job roles attract younger or older customers, providing valuable insights for targeted marketing strategies.

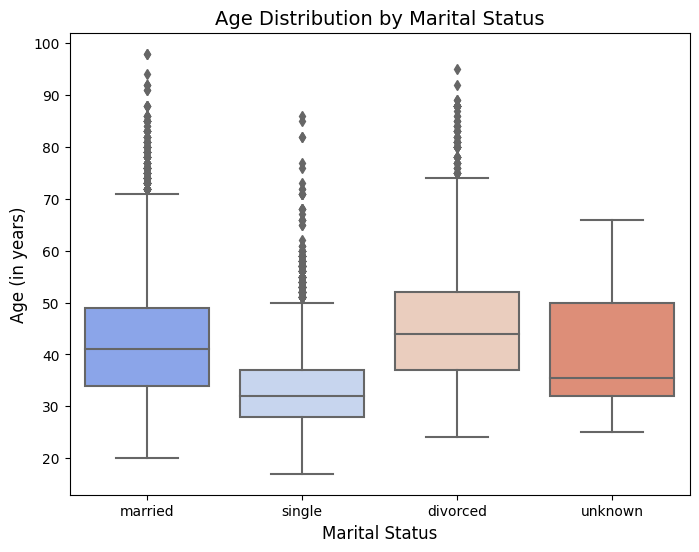
Below is the output for the code above



**Figure 5.5 Boxplot of age clusters.**

Similarly, a boxplot for age distribution by marital status was created to analyze whether age varies significantly among single, married, and divorced individuals. Understanding these differences can help in tailoring financial products and services to different customer segments. Additionally, a scatter plot of Age vs. Pdays (days since last contact) was plotted, with points colored by marital status. This visualization helps identify whether older or younger customers experience longer follow-up delays, providing insights into customer engagement patterns and how follow-up strategies could be improved.

It looked like this:



**Figure 5.6 boxplot of marital status**

This is the creation of boxplots for each job type by age.

# 1. Boxplot of Balance by Job Type

plt.figure(figsize=(14, 6))

sns.boxplot(x='job', y='age', data=df1, palette='Set1')

plt.xticks(rotation=45)

plt.xlabel('Job Type', fontsize=12)

plt.ylabel('Age (in years)', fontsize=12)

plt.title('Age Distribution by Job Type', fontsize=14)

plt.show()

# 2. Boxplot of Balance by Marital Status

plt.figure(figsize=(8, 6))

sns.boxplot(x='marital', y='age', data=df1, palette='coolwarm')

plt.xlabel('Marital Status', fontsize=12)

plt.ylabel('Age (in years)', fontsize=12)

plt.title('Age Distribution by Marital Status', fontsize=14)

plt.show()

# 3. Scatter Plot of Age vs Balance

plt.figure(figsize=(10, 6))

sns.scatterplot(x='age', y='pdays', data=df1, hue='marital', palette='husl', alpha=0.8)

plt.xlabel('Age', fontsize=12)

plt.ylabel('Pdays (in day)', fontsize=12)

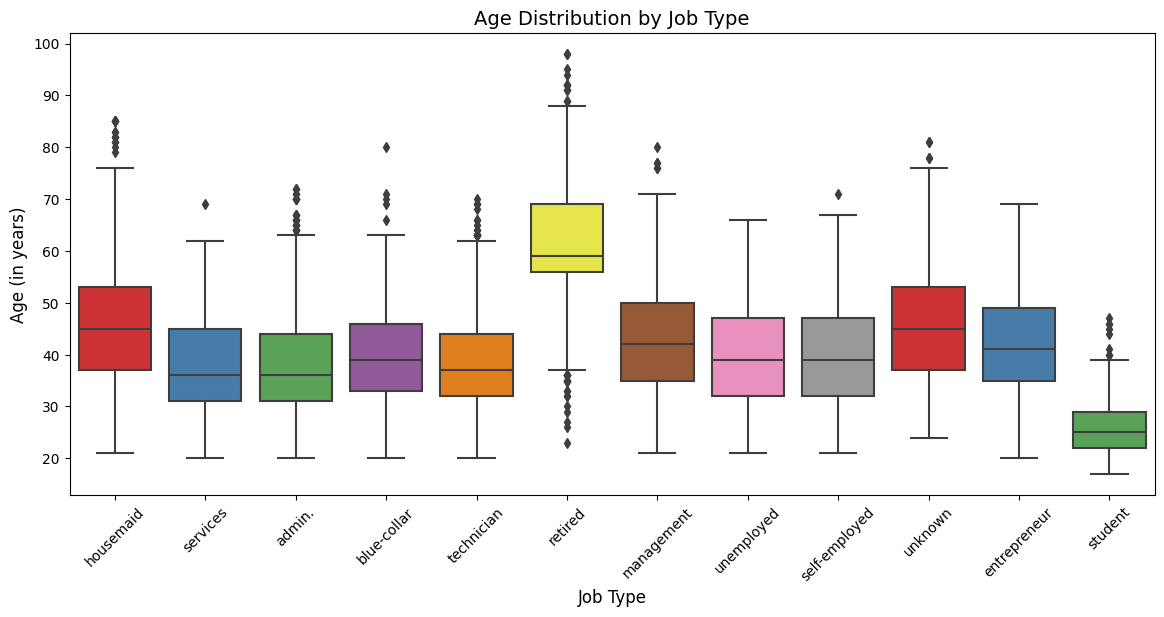
plt.title('Scatter Plot: Age vs Pdays, Colored by Marital Status', fontsize=14)

plt.legend(title='Marital Status')

plt.show()

**Figure 5.7 Code snippet of plot creation.**

The output for the code is as follows



**Figure 5.8 Boxplot of age by job type.**

This data indicates that the majority of individuals in the retired job category are aged 60 and above. Consequently, it appears that the outliers in our age group charts primarily originate from this retired segment. Conversely, the youngest demographic represented is that of students in their early twenties, with a few outliers extending into their fifties.

# **6.0 RESULTS AND DISCUSSIONS**

Finally, we shall go over the results of data mining and analysis from the visuals that we had created and see what insights can be found. Not only from the first dataset but both in comparison. Going over similar factors of relation as well as see what points the dataset were vastly different from each other. All this with additional visuals provided in the appendix is discussed below.

## **6.1 Key findings**

Customer segmentation, K-means clustering and association rule mining to segment customers based on demographic and financial attributes as well as marital status, revealing distinct groups with unique characteristics.

Data insights variables such as age, job type, marital status, education, and financial behaviors (e.g., loan and housing status) significantly influenced segmentation.

Financial behavior patterns: the analysis identified patterns in financial behaviors, such as the likelihood of having housing or personal loans, which are essential for targeted marketing.

Association rules using Apriori and FP-Growth algorithms, we discovered frequent item sets and strong association rules, offering insights into customer purchasing behaviors.

### **6.1.1 Significance**

Marketing optimization insights enable banks to customize their marketing strategies, thereby improving customer engagement and profitability. By efficiently allocating resources and identifying high-value customer segments, banks can optimize resource distribution for greater returns. Additionally, implementing proactive retention strategies for at-risk segments can help reduce churn rates and enhance customer lifetime value.

Unexpected results were observed, particularly with outliers in age data. The presence of older individuals as outliers suggests unique customer segments that require further investigation. Additionally, skewness in the Pdays data and anomalies in customer contact history indicate patterns that may affect future analyses.

## **6.2 Business implications**

### **6.2.1 Practical Applications for the kenyan Market**

Targeted campaigns for customer segments enable banks to design personalized marketing campaigns, boosting engagement and conversions. Resource optimization focusing on high-value segments ensures efficient resource allocation and higher returns on investment.

### **6.2.2 Strategic Recommendations For kenyan Banks**

To maximize profitability, financial institutions should concentrate on high-value customer segments. It is essential to harness digital channels for enhanced online engagement, particularly in light of the increased digital adoption observed in the aftermath of the COVID-19 pandemic.

### **6.2.3 Industry Impact**

Competitive edge data-driven insights can assist Kenyan banks in improving customer engagement and personalizing their marketing strategies. Findings related to market adaptation enable banks to adjust to the evolving behaviors of customers and shifts within the industry. Additionally, innovations in marketing, particularly through advanced data mining techniques and AI-driven segmentation, facilitate the execution of more effective campaigns.

## **6.3 Limitations**

Clustering techniques, such as K-means, may not adequately capture complex relationships among data points. Therefore, the exploration of hierarchical clustering or DBSCAN may provide more profound insights into the data structure. Furthermore, in the context of association rule mining, the reliance on fixed thresholds for support and confidence may lack generalizability. By optimizing these thresholds, it is possible to enhance the quality and relevance of the findings.

The sample size must be increased to ensure greater diversity, and the inclusion of a specific dataset about Kenya would significantly enhance the robustness of the findings. Furthermore, variations in data quality can adversely affect accuracy, thereby underscoring the necessity for consistent data validation practices.

# **7.0 CONCLUSION**

## **7.1 Conclusion**

This study aimed to leverage data warehousing and mining techniques to segment customers in the Kenyan banking sector and analyze their financial behaviors. The objective was to provide actionable insights that could enhance marketing strategies, inform financial product design, and promote financial inclusion. Although the dataset utilized was sourced from a Portuguese bank, it possessed the necessary structure and attributes to meet the study's aims and allow for the application of findings to the Kenyan market.

The analysis effectively segmented customers using K-means clustering and association rule mining, highlighting key variables such as age, job type, marital status, and financial habits as significant factors influencing segmentation. Distinct behavioral patterns were identified regarding mobile loan usage, SACCO memberships, and digital savings trends, all of which play a crucial role in shaping the financial landscape in Kenya.

Targeted marketing insights support the development of data-driven marketing campaigns that increase customer engagement, improve financial literacy, and drive financial inclusion.

## **7.2 Recomendations**

Personalized banking solutions Kenyan banks should leverage customer segmentation insights to develop tailored financial products, such as mobile savings accounts, SME financing, and digital micro-loans that cater to different customer needs.

Financial education programs to enhance financial inclusion: banks and fintech firms should invest in financial literacy programs, particularly for mobile money users and individuals relying on informal financial services.

Expand data sources to include datasets from various Kenyan banks, fintech companies, mobile money platforms, and SACCOs. This will enhance the accuracy and generalizability of customer segmentation models. Explore advanced techniques, such as AI-driven clustering methods, including deep learning-based segmentation, to gain deeper insights into evolving financial behaviors.

Future studies should integrate alternative variables, considering factors such as informal savings groups (chamas), digital transaction histories, and peer-to-peer lending activities to further refine segmentation. Establish a real-time monitoring system to track customer segmentation trends continuously, allowing for timely updates to marketing strategies. This will ensure that banks remain adaptive to shifting consumer behaviors and market dynamics.

# **8.0 REFERENCES**

Moro, S., Laureano, R. M., & Cortez, P. (2014). A Data-Driven Approach for Bank Telemarketing Campaigns. *Decision Support Systems, 62*, 22-31.

*Bank Marketing*. (n.d.). <https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing>

Ranjan, A., & Srivastava, S. (2022). Customer segmentation using machine learning: A literature review. *AIP Conference Proceedings*, *2767*, 020036. <https://doi.org/10.1063/5.0103946>

Gomes, M. A., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, *21*(3), 527–570. <https://doi.org/10.1007/s10257-023-00640-4>

Mihova, V., & Pavlov, V. (2018). A customer segmentation approach in commercial banks. *AIP Conference Proceedings*. <https://doi.org/10.1063/1.5064881>

Raiter, O. (2021). Segmentation of bank consumers for artificial intelligence marketing. *ResearchGate*. <https://doi.org/10.17613/q0h8-m266>

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, 207–216. <https://doi.org/10.1145/170035.170072>

Han, J., Kamber, M., & Pei, J. (2011).  
Data Mining: Concepts and Techniques (3rd ed.).  
Morgan Kaufmann.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996).  
From Data Mining to Knowledge Discovery in Databases.  
AI Magazine, 17(3), 37-54.

Tan, P.-N., Steinbach, M., & Kumar, V. (2019).  
Introduction to Data Mining (2nd ed.).  
Pearson.

# 9 Appendix