UNIVERSITY OF HERTFORDSHIRE

School of Computer Science
MSc Data Science and Analytics

Final Project Report

7COM1039 - Computer Science Masters Project

Title

Customer Segmentation in the Banking Industry using machine learning

Student Name: Rajashakhar Nampelli

Student ID: 19059824

Project supervisor: Imran Khan

MSc Final Project Declaration

This report is submitted in partial fulfilment of the requirement for the degree of Master of Science in Data Science and Analytics at the University of Hertfordshire (UH).

It is my own work except where indicated in the report.

I did not use human participants in my MSc Project.

I hereby give permission for the report to be made available on the university website provided the source is acknowledged.

Abstract

Banks globally are faced with intense competition from numerous financial products such as digital banking, cryptocurrencies, and financial technologies. Therefore, banks need to become agile and innovative to stay afloat and have a competitive edge. In light of this, customer segmentation is a critical strategy to allow banks to avail better-individualized experiences to their customers. This research paper encompasses the use of machine learning for the segmentation of bank customers. The study reviews literature on customer segmentation using machine learning, the key factors that influence customer segmentation in banks, and the benefits banks enjoy from improved customer clustering. Finally, the paper will explain the steps taken in the development of the machine learning model using unsupervised learning and two algorithms such as PCA and K-nearest neighbours. K-means was able to cluster clients based on common financial characteristics. The algorithm was used to segment customers because it is a simple and efficient method for identifying groups or clusters within a dataset. It helped segment customers based on their financial behaviours and transactions, as it allows for the creation of homogenous groups of customers with similar characteristics.

Table of Contents

List of tables

List of Figures

Chapter 1: Introduction

"Data is the new oil," a famous phrase used to elaborate on the significance of data and how it can increase its value once carefully processed and analysed, just like oil. The current global business environment faces numerous challenges, with the banking sector not spared. Customers globally are different, and with each comes different needs and wants. According to Nobar and Rostamzadeh (2018), customer power has risen significantly due to increased access to information and access to more alternatives. In addition to increased customer power, businesses globally face strong global competition that demands new innovative ways to have a competitive edge and increased profit margins (Crowley and Jordan, 2017).

Effective communication is essential to standing out from competitors in a competitive business environment. Dividing a business's target market into well-defined approachable clusters can help better understand customer needs and effectively fulfil communication. Customers leave behind a tonne of data that can segment them into distinct groups for marketing campaigns and propositions (Brito et al., 2015). Different artificial intelligence techniques, such as machine learning, can analyse vast volumes of data for better decision-making and insights synthesis (Sun et al., 2015). Analysing big data for customer behaviour and preferences using machine learning algorithms can help unearth crucial patterns that help better a customer's customer experience. This research aims to find out the role of big data analytics in creating a personalized banking experience for clients in banking institutions.

This research is about customer segmentation in the banking sector. Customer segmentation is the practice of dividing a customer base into smaller groups based on shared characteristics, such as demographics, preferences, behaviours, and needs (Raiter, 2021). This allows businesses to tailor their marketing, sales, and service efforts to better serve and target specific segments of their audience. By segmenting their customers, businesses can better understand their customers' needs and preferences, and create more personalized and effective marketing campaigns. Customer segmentation in banking is the practice of dividing a bank's customer base into smaller groups based on common characteristics. This allows banks to tailor their products and services to the specific needs of each group, which can help improve customer satisfaction and loyalty. For example, a bank might segment its customers based on factors such as age, income, geographic location, and account activity. By understanding the needs and preferences of each customer segment, banks can create targeted marketing campaigns and offers that are more likely to resonate with their customers (Raiter, 2021). This can also help banks to identify opportunities for cross-selling and upselling additional products and services.

1.1 Research Background

The banking sector is among the industries largely hit by strong global competition and multiple alternative products. Financial technology companies dubbed FinTechs, which are a huge threat to banking institutions, have risen significantly. According to Cortina Lorente and Schmukler (2018), FinTechs owe their success to intuitive and simplified customer experience. Customer experience is the relationship between a client and business in each interaction, no matter how brief. Customer experience significantly affects a customer's perception and feelings about a business. How seamless and simple these interactions remarkably affect customers' loyalty and can help boost revenue (Keiningham et al., 2020). According to Min et al. (2016), retaining an existing customer is five times cheaper than acquiring a new one. Therefore, the banking sector must enhance its customer experience to help them retain customers and foster loyalty.

Keeping clients happy calls for proper segmentation. Market segmentation is a process of dividing customers into different categories and discerning how customers sharing the same attributes engage or use a product. Segmentation helps discern how different customers think and their preferences. This information can then be used to implement bespoke customer experiences (Ziafat and Shakeri, 2014). Vast amounts of data characterize the banking sector. Customers interact with banks on a digital level due to access to the computer infrastructure. In addition to structured transactional data, banks collect data from a variety of sources such as social media, image data, emails, texts, voice messages, stream data, and customer relationship management systems (Bholat, 2015). The foundation of personalization requires the ability to use data and analytics to drive all customer engagement. This process calls for making personalization a major priority, creating and enforcing strategies and operational structures that allow for personalization at a scale. The availability of this big data has hugely impacted the banking sector due to ever-changing customer demands (Mihova and Pavlov, 2018).

Currently, banking sectors use a generalised approach to finding customer needs and wants instead of a more targeted and specific approach. Putting big data into context by processing and analysing data such as spending habits, demographic information, geographic locations, preferences, customer attitudes, and product and service usage is key in availing personalized experiences for bank clients (Cuadros and Domínguez, 2014). Leveraging machine learning algorithms is instrumental in big data analytics of data about the banking sector to uncover customer information that would be relevant in attaining desired banking goals and objectives.

Machine learning entails developing machines or programs that are not explicitly programmed to perform defined tasks but rather learn and improve from previous experiences (Jordan and Mitchell, 2015). Machine learning has a subset that is based on artificial neural networks referred to as deep learning. Machine and deep learning can be used to train a model using big data that can be used to segment bank clients appropriately to enable individualized experience accurately (Sabbeh, 2018). Big data analytics enables the analysis of large volumes of data, data from multiple sources, and different data formats and helps model the uncertainty or ambiguity of data. It is essential for banks globally to ensure that the right message is pitched to the right customer at the perfect moment.

1.2 Problem Statement

Banking institutions face intense global competition that requires high adaptability to survive. There is a need for banking institutions to analyse the data they collect from their clients and remodel their customer segmentation into a model that works best for the interest of the organization and clients. The typical or basic customer segmentation puts the needs and wants of the clients in a general manner without focusing on key areas that pain the clients. Communicating financial advertisements, financial plans, or new product recommendations or promotions to clients without properly targeting them can cause dissatisfaction and lead them to seek alternatives.

Most customer segmentations currently applied focus on large filters rather than specific ones. Common current segmentations are age, location, gender, marital status, and living place. This calls for a more focused, detailed, and precise way of segmenting bank clients to help banks provide a better-individualized customer experience to their clients. When the right messages are available or pushed out to the right customers, it creates context, which helps ensure a better response. Finance technology companies are becoming a major threat to bank institutions due to their focus on personalization. According to Mbama and Ezepue (2018), while most banks acknowledge and communicate they realize the need for personalization, less than twenty percent acknowledge doing a good job at it.

1.3 Project Rationale and Significance

The project aims to study how artificial intelligence can be applied in the banking industry and the role of big data analytics in improving the customer experience. The study will help determine how customer segmentation can be carried out in banking institutions using machine learning algorithms and big data clustering. With fast and powerful computers at our disposal, computer infrastructure has grown immensely over the years. Therefore, it is possible to process data tones fast to gain insights. Data relevant to properly segmenting bank clients are readily available from multiple sources. Numerous algorithms, such as K-nearest neighbours, are available for customer clustering (Qaddoura et al., 2020). Customer segmentation in banking is important and necessary since it will help banks better give product recommendations, give out loans, execute wealth management, design better chatbots, and customer relationship management systems (CRMs), and ultimately improve customer satisfaction (Raju et al., 2014). Machine learning and deep learning can be used with statistical models for forecasting and making predictions that can be used to evaluate how different customer segments can be administered tailored experiences.

1.4 Research Questions

The following are the research questions that will be used to study the research topic and narrow down the research scope.

i. What are the key factors that influence customer segmentation in machine learning?

ii. What are the benefits associated with proper customer segmentation in the banking industry?

iii. Can machine learning algorithms accurately predict customer segments based on their behaviour and demographics?

iv. How does the use of machine learning in customer segmentation improve marketing efforts and customer targeting?

v. How do different machine learning algorithms compare in terms of their effectiveness in customer segmentation?

vi. How can customer segmentation using machine learning be integrated into a company's overall marketing strategy?

vii. How does the inclusion of unstructured data, such as customer reviews and social media interactions, impact the accuracy of customer segmentation using machine learning?

1.5 Research Aim and Objectives

This research project aims to find out and discern the role of big data analytics and machine learning techniques in the banking sector and how they can be leveraged to segment customers to allow for individualized customer experience for clients.

1.6 Objectives

Below are the research objectives

- To investigate whether machine learning algorithms can accurately predict customer segments based on their behavior and demographics?
- To highlight the key factors that influence customer segmentation in machine learning?
- To understand the benefits associated with proper customer segmentation in the banking industry?
- To study whether machine learning in customer segmentation can improve marketing efforts and customer targeting?
- To evaluate how different machine learning algorithms compare in terms of their effectiveness in customer segmentation?
- To investigate how customer segmentation using machine learning can be integrated into a bank's overall marketing strategy?
- To understand whether the inclusion of unstructured data, such as customer reviews and social media interactions, can impact the accuracy of customer segmentation using machine learning?

Chapter 2: Literature review

2.1 Big Data and its relevance in the banking industry

Big data refers to large quantities of data that are usually difficult to manage (Zakir et al., 2015). The data can either be structured or unstructured from multiple sources. Structured data is characterized by how data is clearly defined with its data types and patterns that make it easily searchable (Kalyanpur et al., 2012). On the other hand, unstructured data is essentially everything else. Unstructured data do not have a predefined format or schema. They include emails, text files, website data, mobile data, media such as video and audio, and sensor data (Eberendu, 2016). Normally, structured data is stored in relational databases such as MySQL, while unstructured data may be stored within NoSQL databases.

Additionally, semi-structured data entail data that would ordinarily be characterized as unstructured but also contains metadata that identifies its characteristics. The metadata contains crucial information that can be used to index and catalog the data for easy searching and analysis (Kettouch et al., 2015). Semi-structured data is a go-between structured and unstructured data.

Big data is usually characterized by four V's: volume, variety, velocity, and veracity. According to Hurwitz et al. (2013), the volume describes the size of data generated and stored in big data systems. Big data is usually in Petabytes or Exabytes. This vast amount of data requires more powerful tools to process than laptops or desktop computers. For instance, people spend loads of time on social applications such as Instagram or Twitter, uploading, commenting, and liking posts. The data generated from these interactions is massive and exponential. The velocity of data describes the rate at which data is generated. The data accumulated rate influences whether the data is categorized as big data or regular data (Sun et al., 2018). Systems should be able to handle the pace and amount of data created, and data processing needs to be considered with respect to velocity. The variety of data describes how big data comes in different formats and sources.

Furthermore, the veracity of big data refers to the trustworthiness or unambiguity of the data. Veracity describes how accurate the data is and is used to conclude the confidence level in data. Reliability, accuracy, and data quality are critical as they influence the results obtained after analysis (Hurwitz et al., 2013). The last characteristic of big data is its value. Value describes the amount of data that organizations and companies keep that is necessary or important. This data is usually highly valued and must be saved and processed to gain insights (Song and Zhu, 2016). A lot of data can be obtained regularly, but it is necessary to identify valuable data relevant to the organization's goals and objectives.

Companies globally use big data to streamline their operations and create personalized campaigns and customer experiences to improve sales and generate more revenue. Businesses that use big data analytics will hold a competitive advantage over others since they can adapt and make crucial decisions quicker than the rest. In big data analytics, both historical and real-time data can be used to generate insights. This is made possible by the use of online analytical processing OLAP and online transaction processing OLTP (Cuzzocrea, 2013). OLAP are systems that deal with historical data and capture snapshots of data. OLAP systems perform multi-dimensional analyses of high speeds and volumes of data. Usually, this data is stored in a data warehouse, which can be subdivided into data marts. The core of most online analytical processing systems is the OLAP cube which enables querying, reporting, and analysing multi-dimensional data (Song et al., 2015). Data stored in OLAP systems (data warehouses) is vast, usually in petabytes or terabytes, and is subject-oriented (Krishnan, 2013).

Online transaction systems are usually transaction-oriented and allow for the real-time processing of database transactions. OLTP systems are behind most everyday transaction systems, such as automatic teller machines. The key difference between OLTP and OLAP systems is that OLAP is analytical and subject-oriented while OLTP is transaction-oriented (Erl et al., 2016). OLAP systems are optimized for complex data analysis and data mining, whereas OLTP systems are optimized for daily transactions and must have high availability. Additionally, data scientists and knowledge workers are designed to use OLAP systems. In contrast, OLTP systems are majorly used by frontline or operational-level workers such as cashiers and bank tellers (Sadeghi et al., 2016).

As indicated, most data analytics and data mining processes are conducted using online analytical systems. This is mostly done by the use of a data warehouse. A data warehouse is a database management system used to collect, store, and manage data from various sources to facilitate meaningful analytics and gain insights (Edastama et al., 2021). Data warehouses a context-based and usually hold historical data which are vast in nature. Furthermore, data warehouses are time-variant, suggesting that they allow users to view changes over time. Once data is stored in a data warehouse, it cannot be changed, unlike in online transaction systems where data is volatile. According to Mukherjee and Kar (2017), before data can be stored correctly in a data warehouse, it has to go through the extraction, transformation, and loading process.

The extraction, transformation, and loading process dubbed ETL is a process of correctly extracting data from different sources, transporting it to the staging area, and finally integrating or loading it to the data warehouse. The first step entails identifying and extracting needed data from the different available sources (Vaisman and Zimányi, 2014). These sources can be relational databases, applications, NoSQL databases, XML, web data, and flat files. The required data is first extracted to a staging area before to a data warehouse. According to Hofer et al. (2016), it is important first to extract the data to the staging area because the data have different formats, can be corrupt, or has errors. Therefore, pushing corrupt data to the warehouse might be damaging, whereas rollbacks are expensive and inefficient.

After data is extracted to the staging area, the transformation process begins where rules and functions are made on the data to convert it to a unified single standard format. Some of the functions include cleaning, splitting, joining, sorting, and filtering (Astriani and Trisminingsih, 2016). Filtering entails loading only selected attributes to the warehouse, while cleaning involves filling up some default values and ensuring that similar attributes with different names are standardized. The final step is now the loading process, where the transformed data is loaded into the data warehouse (Sharma and Jain, 2014). Data warehousing and the ETL process are at the core of big data analytics. With the large amount of customer data banks generate daily, it is necessary to store it in a data warehouse and combine it with other data sources such as demographics and social media data to add context and gain better insights.

2.2 The key factors that influence customer segmentation in banks using machine learning

According to He and Li (2016), customer segmentation is a process whereby customers are grouped based on similar characteristics, allowing businesses to conduct marketing in each group efficiently. The banking industry and other financial companies like insurance, credit card, and credit bureaus were among the first organizations to acknowledge the need to target distinct consumer groups with appropriate goods as a substitute for product differentiation. Over the years, bank customer segmentation has been done using various variables such as demographics, spatial differences, and psychological or benefit basis. According to Hassan and Craft (2012), market segmentation entails perceiving a heterogeneous market as a collection of small homogeneous markets. Customer segmentation plays a major role in delivering a competitive edge for banks if done well.

Below are the key factors that influence customer segmentation in the banking industry.

- ***Demographics***

Demographics such as age, gender, income level, education level, and location can influence customer segmentation in banks using machine learning (Abidar et al., 2020). For example, a younger demographic may be more likely to use mobile banking services and be interested in investing in high-risk, high-reward financial products, while an older demographic may be more interested in traditional banking services and low-risk investment options. Income level may also impact a customer's ability to access certain financial products and services, such as mortgages or investment portfolios. Additionally, education level may influence a customer's understanding and usage of financial products, while location can determine a customer's access to physical branches and ATM networks (Raiter, 2021). Machine learning algorithms can analyze these demographic factors and use them to accurately segment customers into relevant groups for targeted marketing and product offerings.

- ***Behavioural data***

Customer behavior can be a valuable source of information for segmentation (Kovács et al., 2021). For example, data on purchase history, website interactions, and engagement with marketing campaigns can be used to identify patterns and trends among different customer groups. Behavioural data can influence customer segmentation in banks using machine learning by providing insight into the actions and behaviours of customers. This information can be used to identify patterns and trends, which can then be used to group customers into segments based on their behavior. For example, if a bank notices that a group of customers consistently use mobile banking, they can create a segment for mobile banking users and tailor their marketing and product offerings to that group. Additionally, behavioural data can help banks better understand the needs and preferences of their customers, allowing them to create more personalized and targeted marketing campaigns.

- ***Psychographics***

Psychological characteristics, such as attitudes, values, and lifestyles, can also be used to segment customers. This can help identify customers with similar motivations and preferences, allowing marketers to create more targeted and effective campaigns. Machine learning algorithms can analyze customer data, such as spending habits, credit history, and demographic information, and use this information to create more accurate and personalized customer segments (Abidar et al., 2020). For example, a bank may use machine learning to segment customers based on their psychographic characteristics, such as their risk aversion, spending habits, and financial goals. This can help the bank identify customers who are more likely to be interested in certain products or services, such as high-yield savings accounts or investment portfolios.

Additionally, by understanding the psychographics of their customers, banks can create more targeted marketing campaigns and personalized offers that are more likely to resonate with specific customer segments. This can help increase customer engagement and loyalty, leading to increased revenue and customer satisfaction. Overall, incorporating psychographics into customer segmentation using machine learning can provide banks with valuable insights into their customers, enabling them to tailor their products, services, and marketing efforts to better meet the needs and preferences of their customers.

- ***Location***

Geographical location can be a key factor in customer segmentation, as customers in different regions may have different needs and preferences. Location plays a significant role in customer segmentation in banks using machine learning because it helps to identify and understand the unique needs and preferences of customers in different regions (Kovács et al., 2021). For example, a bank operating in a rural area may have a different customer base than one operating in an urban area. Machine learning algorithms can be used to analyze customer data and identify patterns and trends based on location. This information can be used to create customized marketing strategies and offers that cater to the specific needs of customers in different regions.

Additionally, location-based data can also be used to identify potential new customers and target them with personalized marketing campaigns (Kovács et al., 2021). For example, a bank can analyze data on economic indicators and population density in different regions to identify areas with high potential for growth and target these areas with tailored marketing efforts. Overall, using machine learning to analyze location-based data can help banks to effectively segment their customer base and provide personalized services that meet the unique needs of customers in different regions.

- ***Customer preferences***

Understanding customer preferences and needs are crucial for creating effective segments. This can be achieved through surveys, focus groups, and other methods that provide insights into customer attitudes and behaviours. Customer preferences play a significant role in the customer segmentation process in banks using machine learning. By understanding the preferences of customers, banks can segment them into different groups based on their common characteristics and needs (Raiter, 2021). For example, a bank may use machine learning algorithms to analyze customer data and identify patterns in their spending habits, financial goals, and preferences for certain products or services. By segmenting customers based on their preferences, banks can tailor their marketing and sales efforts to specific customer groups. This allows them to offer targeted products and services that are more likely to be relevant and appealing to each group, increasing customer satisfaction and loyalty. Furthermore, by understanding customer preferences, banks can also improve their risk management strategies. For example, if a bank identifies a group of customers who are at high-risk for defaulting on loans or credit cards, they can implement targeted interventions to help these customers manage their finances better and avoid financial difficulties. Overall, the use of machine learning in customer segmentation can help banks better understand and serve their customers, leading to improved customer satisfaction and retention.

- ***Brand loyalty***

Customers who are loyal to a particular brand are likely to have different characteristics and needs compared to those who are less loyal. This can be an important factor in segmenting customers based on their level of brand loyalty. Brand loyalty can influence customer segmentation in banks using machine learning in several ways (Raiter, 2021). First, loyal customers are more likely to stick with a particular bank and engage in a variety of financial products and services offered by the bank. This allows the bank to gather more data on loyal customers' preferences and behavior, which can be used to segment them into different groups based on their needs and preferences.

Second, loyal customers are more likely to provide positive feedback and recommend the bank to others, leading to increased brand awareness and customer acquisition. This can help the bank to target and attract new customers who are similar to the loyal ones, leading to more effective customer segmentation (Raiter, 2021). Third, machine learning algorithms can be used to analyze the data gathered from loyal customers and identify common patterns and trends in their behavior.

This can help the bank to segment the loyal customers into different groups based on factors such as their transaction history, account usage, and product preferences. Overall, brand loyalty can influence customer segmentation in banks using machine learning by providing more data on loyal customers, attracting new customers who are similar to the loyal ones, and identifying patterns and trends in the behavior of loyal customers (Raiter, 2021). This can help the bank to better understand and serve the needs of its customers, leading to increased customer satisfaction and loyalty.

2.3 The benefits associated with proper customer segmentation in the banking industry

Bank customers are generally different, and all require different tailored services from banks. Simply classifying customers based on their age, gender, geographical location, and other vague attributes is not adequate. There is a need to use more personalized information to profile bank clients (Wang et al., 2017). Artificial intelligence techniques such as machine learning, deep learning, and other procedures have proven impactful in segmentation (Paruchuri, 2019). Due to the growth of the internet and the use of computers, customers globally leave behind a massive trail of data that can be used to gain insights. This data can be analysed using big data analytics concepts to remove the guesswork involved in client interactions. For instance, by analysing customers' data and categorizing them appropriately within their profiles or persona, banks may be able to successfully conduct email or social media campaigns and offer top-notch customer service, which ultimately improves the customer experience (Gichuru and Limiri, 2017). This effectiveness is attributed to artificial intelligence, enabling banks to have better contact points with customers.

Additionally, using big data analytics to foster artificial intelligence marketing in banks can help improve the return on investment of their marketing campaigns. According to Osei et al. (2021), there are several key ways in which customer segmentation helps banks. First, it allows banks to discover critical data points for customer segmentation. These key data points are powerful since they give insights into given customer segments. The data points can be simple, from age to more complex lifestyle patterns, daily expenditure, and behaviour and interest. Second, customer segmentation allows banks to classify new customers into user-specific segments that are already classified based on various data connections (Osei et al., 2021).

Additionally, segmentation in banks allows for real-time monitoring, evaluation, and reallocation of segments. A bank's customer behaviour needs to be continuously monitored to detect any changes in their preferences or expenditure (Jayasree and Balan, 2013). Machine learning models can make switching or alternating customers in the set segments possible with continuous monitoring. Finally, customer segmentation enables banks to continuously learn about their customers, which enables them to know the relationship between the different services they offer, be it savings or loans (Taherparvar et al., 2014). Furthermore, proper segmentation makes it possible to predict the next best action for each customer, which goes a long way in delivering personalized experiences.

2.4 Machine Learning Algorithms and Techniques for Segmentation

One of the most powerful methods of segmentation is using the RFM model. The model is used to identify customer groups for special treatment. It uses three metrics, recency, frequency, and monetary. According to Dogan et al. (2018), the RFM model is commonly used in banks due to three key reasons. First, the model uses objective numerical scales that yield a concise high-level customer representation. Secondly, it is simpler to use without the need to complex software or data science. Finally, the RFM model is intuitive since the results of the segmentation process are easy to read, interpret, and understand. The RFM model bases its arguments on the fact that banks can gain an understanding using the three aforementioned quantifiers, recency, frequency, and monetary (Nikumanesh and Albadvi, 2014). Recency entails how much time has elapsed since a customer's last activity with the bank. These activities include deposits, savings, loans, or payments. Frequency entails how often a customer interacts with the bank's services during a certain period of time. Finally, monetary reflects how much a customer transacts with the bank over a certain time.

Binary classification is another method used to carry out categorization or segmentation. This method entails predicting categorical variables where the output is restricted to two (Smeureanu et al., 2013). Some of the popular algorithms that can be used for binary classification include decision trees and K-nearest neighbours. However, binary classification has some limitations; for instance, the process becomes more cumbersome as the generated data needs to be classified as large (Santoso et al., 2020). Machine learning algorithms have thus become more popular in solving binary classification problems, such as they are non-parametric.

Artificial neural networks are another way of classification that can be instrumental for bank segmentation. Artificial neural networks encompass the use of computational methods that are inspired by or mimic the human brain (Da Silva et al., 2017). Neural networks model statistical data in a non-linear manner and allow interconnected elements to process information simultaneously. Moreover, they have a learning agent that has the ability to learn and adapt from past experiences. A multilayer perceptron, a fully connected multilayer neural network, contains three layers: an input, one or more hidden layers, and an output layer. The input layer is concerned with initial data, while the output layer relays the answer of the network. According to Lolli et al. (2017), one hidden layer is capable of solving complex classification problems well, but in extreme cases, up to three hidden layers can be used. There is no clear way to identify the number of required hidden layers; most times, the number is selected through trials and error comparisons (Cilimkovic, 2015).

In addition to artificial neural networks, the best-worst method (BWM) is an efficient technique used to derive criterion weights during decision-making (Pamučar et al., 2020). This method entails identifying a set of decision criteria to be used, followed by the identification of the best and worst criteria. The preference of the best criteria, overall criteria and all criteria over the worst criteria is determined by a number between one and nine experts. Finally, optimal weights are established by solving the non-linear model (Omrani et al., 2022). Furthermore, clustering is a common method of classification of data and segmentation. It involves placing data that have the most resemblance in some features into groups. Each cluster holds data that is similar to other data in that cluster but different from data in other clusters (Kashwan and Velu, 2013). Distance is used to measure similarity in clusters. There are two key types of clustering, namely, hard and soft clustering. Hard clustering ensures that each data point belongs to a cluster completely or not. On the other hand, soft clustering entails using a probability or a likelihood of a data point belonging to a certain cluster instead of explicitly putting it in a cluster (Ferraro and Giordani, 2020).

Chapter 3: Research Methodology

The research methodology contains essential information on the various ways and steps followed in the analysis. The methodology draws its foundations from the problem statements stated above, the study's objectives, the research questions, and the overall aim. The major aim of this study was to find out how machine learning can be applied to banking dataset to cluster or segment customers based on similar characteristics. If successful, the insights generated can be leveraged to deliver a personalised banking experience to the bank clients.

3.1 Methodology

The research will use qualitative and quantitative approaches to collect and analyse banking data, for instance, to find out how sources of information and the extent to which artificial intelligence, in conjunction with big data analytics, has been and can be used in the banking sector. Data collection will involve looking for text data to gain both quantitative and quantitative data. The research study uses a descriptive design to try and explain the observed phenomena of how banks apply big data analytics in their customer segmentation and why the state of customer segmentation is not at its current best.

The analysis uses publicly available dataset at Kaggle to further study how customer segmentation can be done deeply and more thoroughly to enable a personalized customer experience for bank clients. The dataset contains over one million customer transactions from India. The transactions will help me understand the best way to segment them appropriately using machine learning algorithms and libraries. The dataset contains attributes such as gender, age, location, account balance, transaction times, transaction details, and amount spent. The report aims to find the best way or model to cluster or segment the customers based on their common banking experiences using these attributes. This dataset will be part of the secondary data that will be used in the research. Secondary data is data that has already been collected by someone else other than the primary user. This, therefore, means that necessary data is available, and appropriate tools can be used to analyse and uncover hidden insights.

Python Jupyter notebook has been used as the platform to analyse the data and some machine learning libraries to enable statistical analysis of the above dataset. The study will also use Matplotlib and Seaborn libraries to visualize, deduce insights, and observe trends within the dataset. Classifications algorithms will be used to cluster customers into appropriate clusters and create a model that can be used to predict where best a customer should be clustered based on the banking information and history of transactions. The study will use precision, recall, and accuracy metrics to gauge the model's performance before settling for the best one.

3.3 Flow chart of the implementation process

Below are the steps that will be followed to illustrate the process of segmenting customers in a banking institution using machine learning.

1. **Collecting data:** The dataset used is downloaded from Kaggle repository. It contains customer data from various sources such as transaction records, customer financial behaviour, and demographic information, among others.
2. **Pre-processing the data:** The dataset will be cleaned and preprocessed to remove any inconsistencies and outliers. This will ensure it is in a format suitable for machine learning. The activities that characterise this stage are removing or imputing missing values, handling outliers, normalizing or standardising the data, encoding categorical variables, splitting the dataset into training and test sets, dimensionality reduction through feature selection or feature extraction, and balancing the dataset to ensure equal representation of different classes.
3. **Applying machine learning algorithms:** Many different algorithms can be used for customer segmentation, such as clustering algorithms, decision trees, and neural networks. The choice of algorithm will depend on the specific goals and characteristics of the data.
4. **Training the model:** Once an algorithm has been chosen, the model must be trained on the data using a process called "fitting." This involves providing the model with a large number of examples and allowing it to learn the patterns and relationships in the data.
5. **Evaluating the model:** After the model has been trained, it must be evaluated to determine its performance. This can be done by comparing the model's predictions

to known outcomes and calculating metrics such as accuracy, precision, recall, and f1 score.

6. **Making predictions using the chosen model:** Once the model has been trained and evaluated, it can be used to make predictions on new data. This can be used to identify customer segments and make personalized recommendations to individual customers.

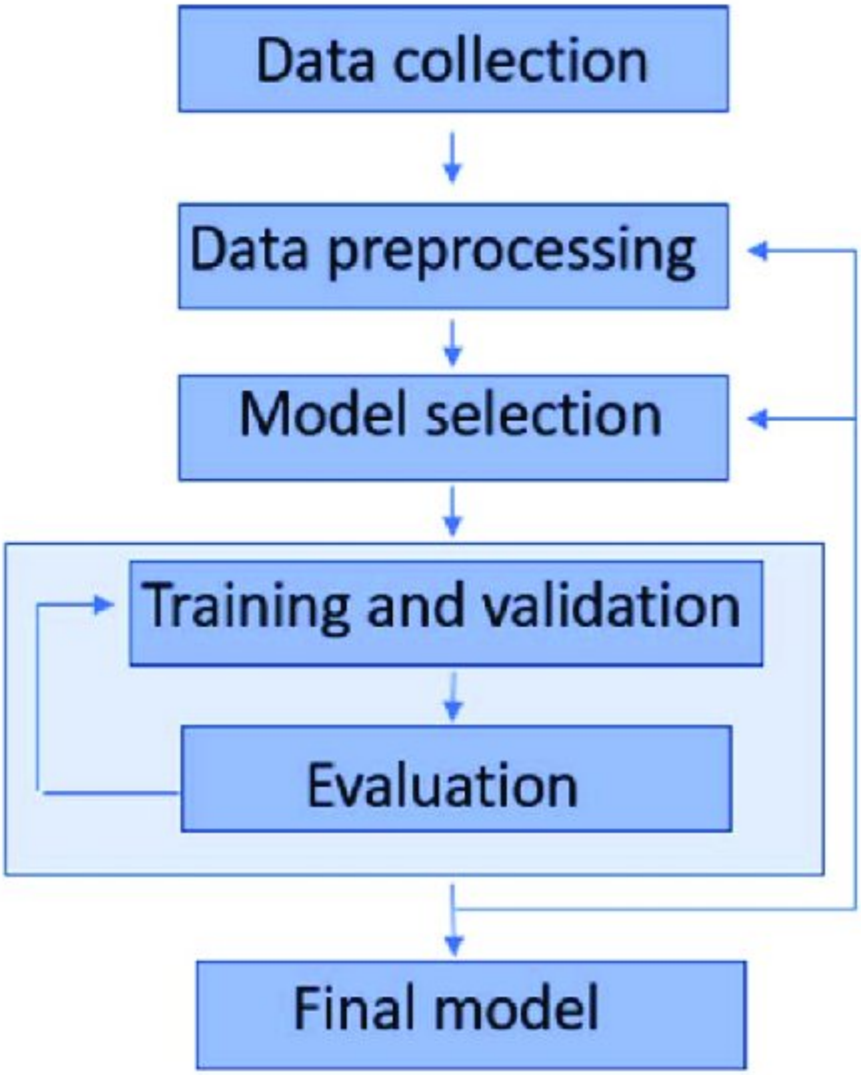


Figure 1: Methodology in a flow chart

3.4 Ethical Considerations

Ethics

While conducting this research, some of the crucial ethical concerns are privacy and consent (Wieringa et al., 2021).

Privacy and confidentiality: It is important to ensure that the customer data used for segmentation is collected and processed in accordance with privacy laws and regulations. This includes obtaining appropriate consent from customers and properly securing their personal information. This research uses publicly available dataset from Kaggle, where data about individuals is anonymised, guaranteeing that the privacy of the data subjects is upheld.

Fairness and non-discrimination: The machine learning algorithms used will be carefully designed and tested to avoid bias and ensure that customer segmentation is based on objective criteria and not on sensitive characteristics such as race, gender, or age.

Data quality and accuracy: The accuracy and reliability of the data used for customer segmentation ensured that the segments are meaningful and relevant. The data from Kaggle is collected from a banking institution that has robust data quality and governance processes in place, guaranteeing the integrity of the data.

Legal

Data privacy and protection laws, such as the General Data Protection Regulation (GDPR) must be followed to ensure that customer data is collected, used, and shared in a legal and transparent manner. The machine learning algorithms used in this study are not biased or discriminatory, and comply with anti-discrimination laws, such as the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA). Since the data used is anonymised, no need to have customers explicitly give consent for their personal data for machine learning purposes.

Social

Privacy concerns: collecting and analyzing customer data for segmentation purposes may raise concerns about personal privacy and data protection. Data used is unanimous and does not contain personally identifiable information.

Inclusivity and diversity: the segmentation process has considered and accounted for the diverse backgrounds and needs of the customers, to avoid potential biases or discrimination.

Impact on customer relationships: the use of machine learning in customer segmentation may affect the way banks interact with and serve their customers, potentially leading to changes in customer loyalty and satisfaction.

Professional

Data privacy and security: customer data has been collected, processed, and stored in a secure manner in a publicly available repository, Kaggle.

Accuracy and reliability: This study will validate and test the accuracy and reliability of the machine learning algorithms used for customer segmentation.

Data quality: The quality of data used in the customer segmentation process is of high quality.

Customer feedback and engagement: The results of this analysis can be used by banks to seek customer feedback and engage with customers to understand their needs and preferences, and to ensure that the customer segmentation process aligns with their expectations and priorities.

Chapter 4: Data Analysis and Code Implementation

4.1 Overview

The main purpose of this chapter is to describe the process and steps that will be taken to build a customer segmentation algorithm system using a banking dataset with an appropriate dataset from an online repository. The application will use provided data to group customers into several optimized segments, each with its own characteristics based on similarities. Hence, the research will focus on how Machine Learning (Unsupervised Learning) is used to improve a way of grouping banking customers based on their record similarities so that banking can reduce the workload involved in the process of segmenting customers based on common transaction behaviors. Using Feature engineering and machine learning, the system will use the customers' record features and group them into *K* optimum segments.

4.2 About the dataset

The dataset used to train the clustering algorithm for this project was downloaded from Kaggle (an online repository for public and private datasets for machine learning) using the following URL link. The dataset is collected from a banking institution based in India. It consists of 1 Million+ transactions with over 800K customers for a bank in India collected in 2016 at the 4th quarter of the year. The data contains information such as - customer age (DOB), location, gender, account balance at the time of the transaction, transaction details, transaction amount, etc.

4.2.1 Dataset description

The dataset contained a total of 9 attributes, with each of varied data types representing a unique feature about the transaction that happened. In general, it had 1048567 total records for different customers. It was recorded in the year 2016 in August and September. The attributes were composed of some numerical, categorical, or string data attributes.

4.2.2 Attributes Details

The selected dataset had the following attributes with their descriptions;

Attribute	Description

TransactionID	This is an attribute that is uniquely used to identify each transaction that was taking place. It is a string that starts with the letter T, and the following characters are number incremented for each transaction.
CustomerID	It is a string composed with an alphabet C as the starting character and followed by numeric values. It is used as an identifier for customers who made transactions. It is unique for each customer hence the values are repeated if the customer did more than one transaction during the data collection period.
CustomerDOB	It is an attribute that indicates when a customer was born. It stores the date of birth for the customer reference but CustomerID in the format date/month/year.
CustGender	It is a string containing a single character, either F and M, which indicates the birth gender of a referenced customer. F stands for Female while M is male.
CustLocation	A string that shows the location where the customer made the transaction. A single customer can perform a transaction in more than one location.
CustAccountBalance	It represents the current monetary value present in the customer's account. This value should not be below zero.
TransactionDate	It is a date attribute that indicates the date of the year when a certain transaction reference occurred. It stores data in the format data/month/year.
TransactionTime	This numerical value indicates the time of the day when a transaction happened. It indicates the time in seconds.
TransactionAmount	This indicates the amount a customer transacted within a single transaction. The value is recorded as Indian Rupees (INR).

Table 1: The dataset attributes

4.3 Resources and Libraries used

To complete the research and achieve the objectives, python language was used as the core development tool. The following libraries were to perform varied tasks as follows;

- Seaborn and matplotlib were used for creating plots for analysis.
- Pandas and NumPy used for loading and manipulating the dataset.
- Sklearn was used mainly as a resource for machine learning tools like PCA for dimension reduction, K-means for clustering, and other processing functions.

4.4 Data cleaning and analysis process

To create the segmentation system for analysis, the dataset was first loaded into memory for processing at first. Various processes were first checked to ensure that the dataset was prepared well for machine learning. Feature engineering was also done to ensure all possible feature attributes are created. At first, the dimension of the dataset was checked and then ensured that each attribute has its required data type. Attribute cleaning was done to ensure that they conform to the required form i.e. removing spaces between attribute names and converting them to lowercase to allow easier referencing.

To confirm if all values in the attributes were present, null values were checked, and their presence percentage analysed. The rule of thumb was that if the total records with nulls are less than 1%, these records were assumed to have no much effects on the data and, hence, were removed instead of imputing them. After dealing with nulls, columns with duplicates were also expulged. Since there needs to be no transaction with less than zero, the "Transaction Amount" and "Account Balance" attributes were checked, and if a value is less than zero, it was removed from the dataset. This is because the number involved must be greater than zero for a transaction to happen. Also, there are rare chances that a bank account will have a negative balance.

Since not all attributes presented can be used in machine learning, some attributes that are not important for clustering and analysis were removed. Attributes that identify a transaction or a customer have no use in modelling. In this case, "TransactionID" and "CustomerID" had no importance, so they were removed to remain with the important features only for clustering.

The next process was dealing with date features. "TransactionDate" and "CustomerDOB" had to be converted into DateTime format in order to allow for some feature engineering. The first feature to be generated from the customerDOB is his current age. This was done by subtracting the customerDOB date from the current date. After generating customer ages, all ages that were abnormal i.e. records with negative dates were removed since there cannot be an age in the future. Also, it was assumed that an individual can own a bank account once he attains an adult age. Based on India's requirements, adulthood starts at 18 years. So, any record with age less than 18 was removed to retain accounts created only by adult individuals. From the "TransactionDate" and "TransactionTime" features, some attributes were generate and they include;

Attribute	Description
Transaction Month	Represents the month in which the transaction took place
TransactionDay	The date of month of the transaction
TransactionWeekDay	The of the week when transaction happened
TransYear	The year in which the transaction happened
TransYearDay	Day of the year when a transaction happened
TransHour	Time in hour of the day when the transaction happened

Table 2: New attributes after feature engineering

After feature engineering, exploratory data analysis is done before selecting the columns to be used for clustering. In this phase, each feature was analysed by plotting simple graphs like pie charts, bar plots, KDE plots, line plots, etc. This made the final processing and analysis of the dataset, which paved the way for clustering data into segments after understanding the dataset.

4.5 Building the algorithm

This report intends to build a data mining tool that can understand the characteristics and similarities of banking transaction records intending to group these records into optimum selected clusters, where the characteristics of each cluster can be studied and used for further recommendation by the bank and on the way forward. Two machine learning algorithms are unsupervised (They do not require labels since they use data points characteristics i.e, similarities to create clusters). The first algorithm was PCA (Principal Component Analysis) which helps to reduce the features to appropriate sizes that can be visualized well. The final algorithm was K-means which clusters the data into appropriate clusters based on the number selected.

The implementation utilized each customer's transaction detail and deduced a comparative distance-based value that can be used to cluster the data into various segments based on a selected cluster centroid. PCA was selected among various dimension reduction techniques because it is easier to understand and build. In contrast, K-means was selected for clustering because it is one of the most classical algorithms in data mining that uses a distance-based technique to group various data points into unique clusters.

4.6 PCA Algorithm

4.6.1 How the PCA Algorithm works

PCA is an unsupervised machine learning model that is used for dimensionality reduction. It is based on some mathematical concepts such as;

- Variance and covariance between attributes
- Eigen values and eigenvectors

It is a statistical algorithm that is used to turn data points of correlated features into a set of new linearly uncorrelated features with the help of orthogonal transformation methods. Newly formed uncorrelated features are called Principal components. One needs to select the number of Principal components that require the algorithm to return. It helps gather some strong relationships from data by reducing the variances and dimensions of the data. The following steps are used for this PCA implementation.

- Standardize the data. Provided data X with columns' features, normalizing it into a specific range.
- Calculate the Covariance of Matrix X, which is standardized.
- Calculate the Eigen Values and Eigen Vectors of the Covariance Matrix of X. Eigenvectors or the covariance matrix are the directions of the axes with high information. And the coefficients of these eigenvectors are defined as the eigenvalues.
- Sort the Eigen Vectors in decreasing order i.e., largest first and smallest as last as Xe
- Calculate Principal Components/ New features by multiplying Xe by X. Each resultant observation will be a linear combination of original features and is independent of each other.

4.7 K-means Algorithm

4.7.1 How K-means Algorithm works

Kmeans is an iterative algorithm that divides given data into *k* clusters or segments where each data point belongs to a single cluster. It gives each data point into a cluster so that the sum of the squared distance between the cluster's centroid and data point is minimal. Less variation within clusters means more homogenous or similarities between data points. Since this algorithm uses distance-based methods to find similarities, it is always a good idea to standardize the data to have a mean of zero and a standard deviation of 1. The following are steps for the working of the K-means Algorithm.

- Select the value of *K*. In this case, the value of *k* will be selected using the elbow curve. *K* values from 2 to 20 will be iterated, and their sum squared errors will be compared by plotting them.
- Initialize *k* centroids from random selections from the dataset.
- Iterate through the data until there is a minimal chance in the centroids by assigning each data point to clusters and compute the squared distance between the data points and all centroids in each iteration.
- Assign each data point to its closest centroid based on computed distances.
- Compute new centroid points by finding the mean of all data points that belong to each cluster.

4.8 Algorithm implementation as described by the steps above

The dimension of the data representing at least 70% of the dataset using PCA is first selected. It was done by 15 components first and then explained the variance ratio used to select the number of components to be used. Below is the implementation.

```
mms = MinMaxScaler()
mms.fit(df)
df1 = mms.transform(df)
n_components = df.shape[1]
pca = PCA(n_components=n_components, random_state=SEED)
pca.fit(df)
#transform the df using the PCA fit above
pca_df = pca.transform(df)
# get variance/data representability of each of the component
for i in range(n_components):
    first_n = pca.explained_variance_ratio_[0:i+1].sum()*100
print(Percent variance explained by first {i+1} components:
{round(first_n , 4)}%')
```

After identifying the number of components to be used, the PCA algorithm was implemented using the selected component as follows in order to reduce the dimension of the dataset by first scaling using the minimax algorithm.

```
# reducing the data to the 2 dimensions
mms = MinMaxScaler()
mms.fit(df)
df2 = mms.transform(df)
pca = PCA(n_components=2, random_state=SEED)
pca.fit(df2)
#transform wines_norm using the PCA fit above
traindf = pca.transform(df2)
```

The best value of *K* groups was then selected by training the model with multiple *k* values from 1 to 20 and then plotting their SSE values. *K* values were determined using the elbow curve method as follows.

```
# determine the best k value using kmeans
# model params
kmeans_kwargs = {"init": "random","n_init": 10,"max_iter":
300,"random_state": 42}
 # A list holds the SSE values for each k
sse = []
for k in range(1, 20):
    kmeans = KMeans(n_clusters=k, **kmeans_kwargs)
    kmeans.fit(traindf)
    sse.append(kmeans.inertia_)
# plot the elbow curve
plt.plot(range(1, 20), sse)
plt.xticks(range(1, 20))
plt.xlabel("Number of Clusters")
plt.ylabel("SSE")
plt.show()
```

The model was then trained using the best selected kmean values and the clusters plotted on a scatter plot to observe their distinguishability characteristics. This marked the end of the training, mapping each record to its predicted clusters and observing their characteristics for further recommendation.

```
kmeans = KMeans(n_clusters=4,init="k-means++")
kmean4.fit(traindf)
labels_4 = kmean4.labels_
# plot
plt.figure(figsize=(15 , 10))
sns.scatterplot(traindf[:, 0], traindf[:, 1] , c = labels_4 ,
label="n_cluster-"+str(len(set(labels_4))))
plt.title("Kmeans Cluster=4   algorithm" , fontsize =25)
plt.show()
```

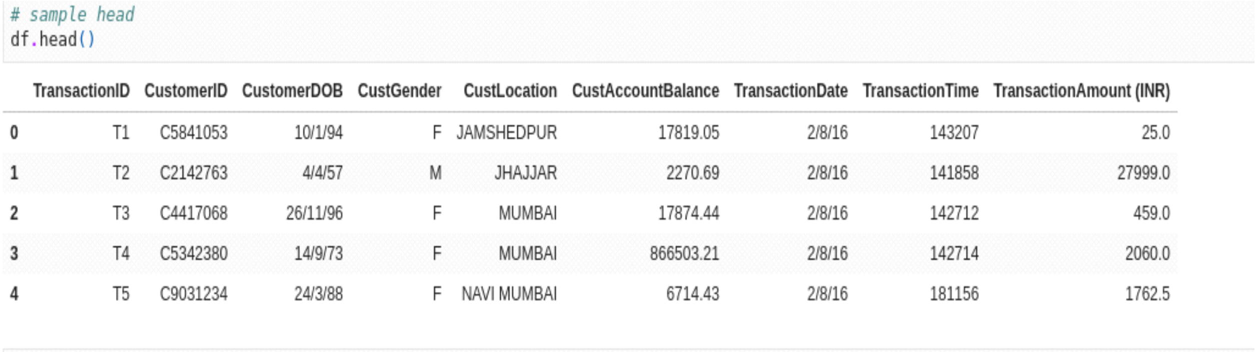
4.9 Chapter Summary

This chapter has described how processing, analysis, and modelling were done up to the final steps. The data was cleaned through various steps such as identifying and handling missing or incorrect data, standardising and normalising data values, removing or correcting outliers and anomalies, encoding categorical variables, removing irrelevant or redundant data such as “TransactionID” and “CustomerID”, and creating new features or transforming existing ones for better model performance, such as date. Principal Component Analysis (PCA) as a dimensionality reduction algorithm that seeks to find the underlying structure in a dataset by projecting the data onto a lower-dimensional space has been used. This is achieved by identifying the directions in the data that have the most variance and constructing new, uncorrelated variables called principal components. These components can be used to represent the original data with a reduced number of dimensions, while retaining as much of the original information as possible. PCA improves the performance of an algorithm by reducing the complexity of the data. K-Means is a popular unsupervised learning algorithm in machine learning that is used for clustering. It divides a given dataset into a specified number of clusters (K) based on the similarity of the data points. K-Means starts by randomly selecting K initial cluster centers, and then assigns each data point to the closest cluster center. The resulting clusters represent groupings of similar data points, which can be useful for clustering and classification.

Chapter 5: Results and discussions

5.1 Introduction

This chapter will focus on reporting immediate results obtained from chapter 4 above. The results are shown in the output snippet. From the dataset observation, some sample top 5 records obtained by *df.head()* are shown below, showing all columns with some 5 sample values in each.



```
# sample head
df.head()
```

	TransactionID	CustomerID	CustomerDOB	CustGender	CustLocation	CustAccountBalance	TransactionDate	TransactionTime	TransactionAmount (INR)
0	T1	C5841053	10/1/94	F	JAMSHEDPUR	17819.05	2/8/16	143207	25.0
1	T2	C2142763	4/4/57	M	JHAJJAR	2270.69	2/8/16	141858	27999.0
2	T3	C4417068	26/11/96	F	MUMBAI	17874.44	2/8/16	142712	459.0
3	T4	C5342380	14/9/73	F	MUMBAI	866503.21	2/8/16	142714	2060.0
4	T5	C9031234	24/3/88	F	NAVI MUMBAI	6714.43	2/8/16	181156	1762.5

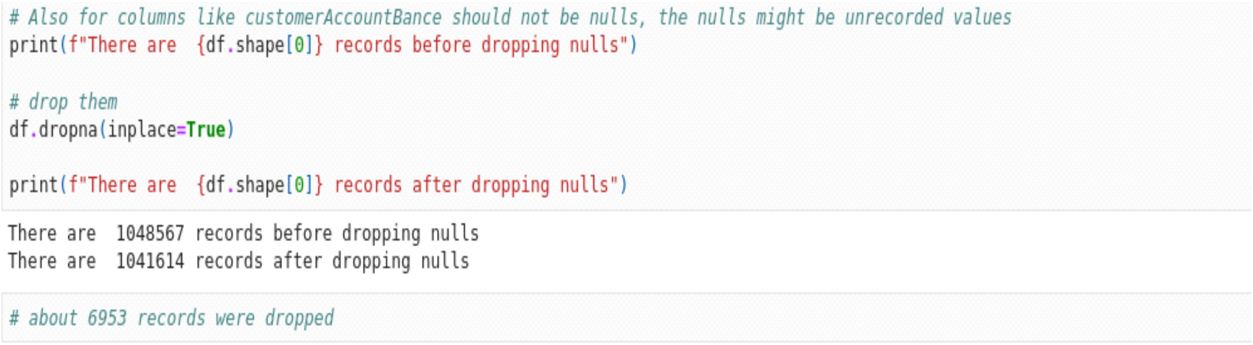
Figure: Sample dataset records

5.2 Data cleaning and processing

Upon looking at the number of records present, it was observed that there were 1,048,567 records, each with 9 attributes of different data types as described in the data description part. Some columns were not in their correct data type i.e., data columns were in string format instead of DateTime, while others had nulls. From the null analysis, it is observed that null values were present in the following.

Attribute	**Count**
CustomerDOB	3397
CustGender	1100
CustLocation	151
CustAccountBalance	2369

As the data keeps records of transactions, a record with "CustAccountBalance" as null is not relevant so it was dropped. Since the null records did not even accumulate over 1% of the whole records, removing them was an easier way to go. To make work easier, all nulls were removed. After removing all null records, 1,041,614 records remained.



```
# Also for columns like customerAccountBance should not be nulls, the nulls might be unrecorded values
print(f"There are  {df.shape[0]} records before dropping nulls")

# drop them
df.dropna(inplace=True)

print(f"There are  {df.shape[0]} records after dropping nulls")
```

```
There are  1048567 records before dropping nulls
There are  1041614 records after dropping nulls
```

```
# about 6953 records were dropped
```

Figure: Number of records before and after dropping nulls

There were also no values with negative values for Customer Balance and Transaction amount since if there are negatives, they might imply an overdraft. The data did not also have duplicate values. Columns that store and are kept to identify a record are not that important in these tasks since the analysis will be generalizing the data. "TransactionID" and "customerID" columns were removed since they are unique Identifiers. On observing the numerical columns, it was clearly seen that TransactionAmount, accountBalance, and Transaction time are highly skewed. This is when the value observed between the mean and maximum is very big (also they have a large std value).

5.3 Exploratory Data Analysis

The records with gender value other than F or M were removed since only Females and Males are documented in the dataset. Below is an analysis of the number of females and males present.

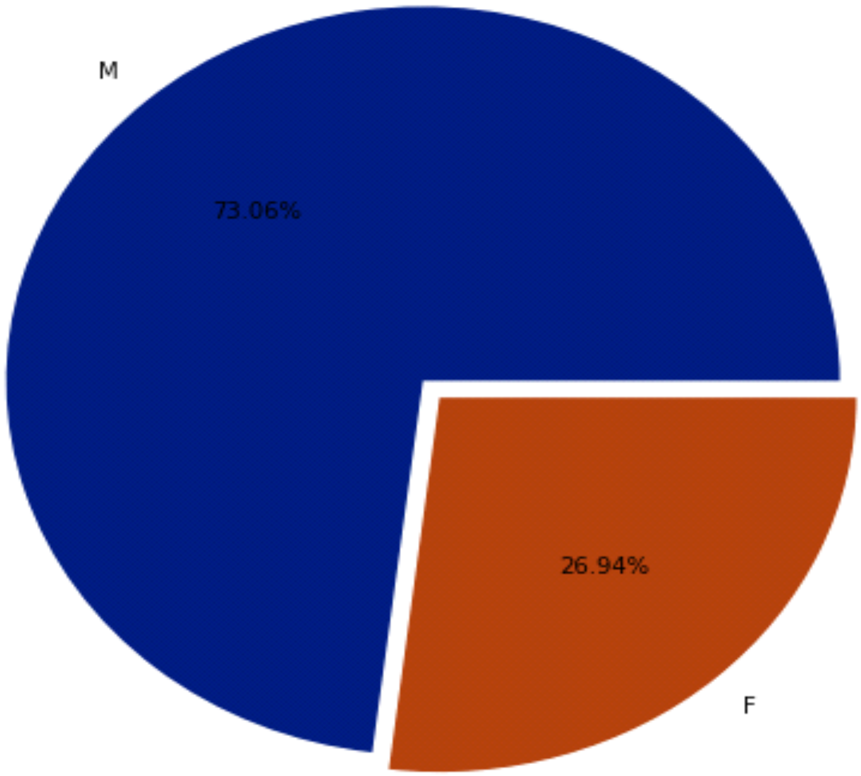
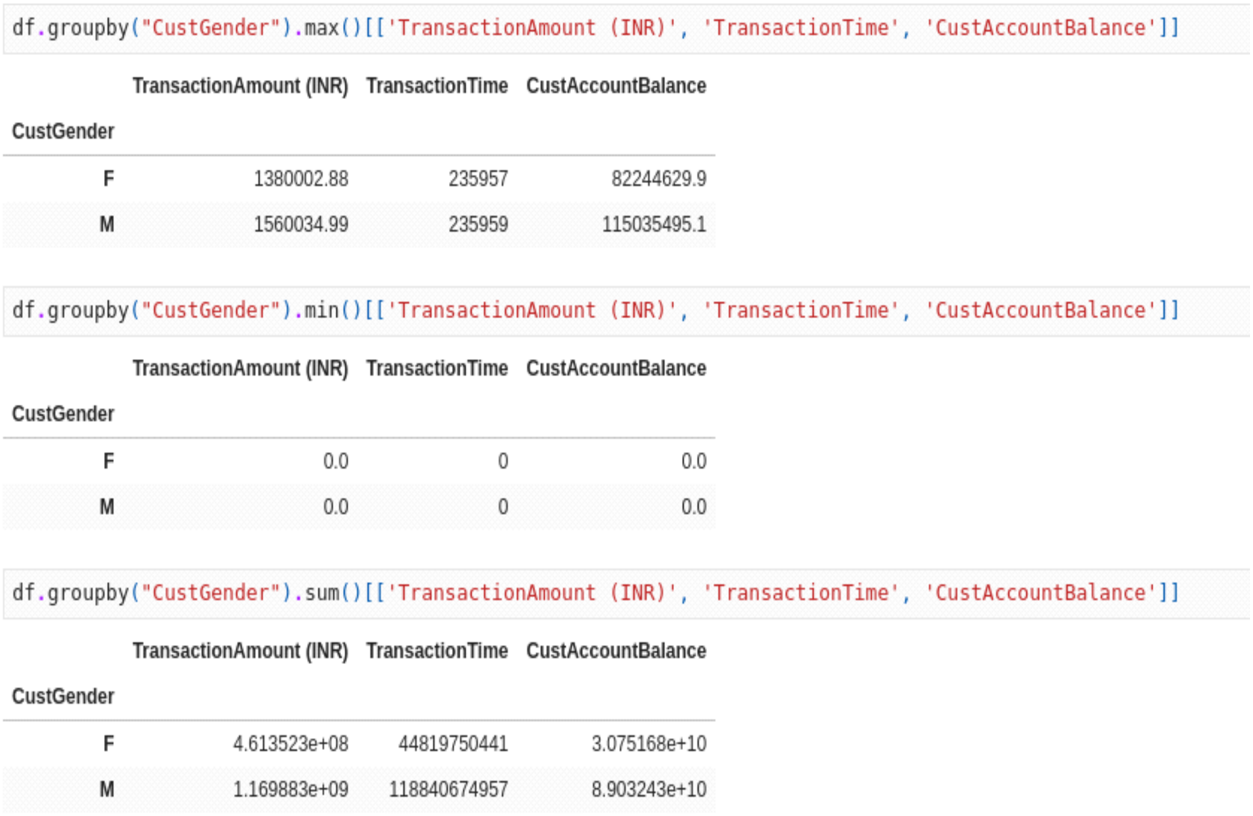


Figure: Pie chart showing Males Vs. Females distributions

About 73% of the dataset is male and 27% female. It shows that males transact more than Females. Some observations of the contribution of both genders are shown below.



```
df.groupby("CustGender").max()[['TransactionAmount (INR)', 'TransactionTime', 'CustAccountBalance']]
```

CustGender	TransactionAmount (INR)	TransactionTime	CustAccountBalance
F	1380002.88	235957	82244629.9
M	1560034.99	235959	115035495.1

```
df.groupby("CustGender").min()[['TransactionAmount (INR)', 'TransactionTime', 'CustAccountBalance']]
```

CustGender	TransactionAmount (INR)	TransactionTime	CustAccountBalance
F	0.0	0	0.0
M	0.0	0	0.0

```
df.groupby("CustGender").sum()[['TransactionAmount (INR)', 'TransactionTime', 'CustAccountBalance']]
```

CustGender	TransactionAmount (INR)	TransactionTime	CustAccountBalance
F	4.613523e+08	44819750441	3.075168e+10
M	1.169883e+09	118840674957	8.903243e+10

Figure: Results on Average, Sum, and Min of all attributes against gender

From the analysis above, Males always have larger value transactions than Females. For example, Males Transacted a sum of 1.16e9 while females had 4.6e8 records of transaction amount. In terms of average amounts, Females had the largest average amount transacted of about 1643 and males 1537, while their male counterparts had the largest amount of account balances average i.e M = 116997 while F=109578. This shows that males keep more money while Females transact more. When checking on the 3 main parameters, TransactionAmount was the last. Time taken for transactions in seconds was way higher than the parameters. The customer Account balance is higher than the transaction amount, hence indicating that the amount being transacted is way lower than the amount in their balance. The amount of distribution is low when compared to stored amounts. In summary from the 3 attributes:-

- Females took more time for transactions than males.
- Males have higher account balances than females.
- Females have a large transaction amount than males.

On analysis of the "CustomerDOB" attribute, it had records with 1/1/1800 as date of birth which was perceived as unrealistic since people born at that time might not be alive now (it is the largest and might have about 222yrs). From the conclusion, the date might be a placeholder. Age column was then created by counting the number of years from date of birth attribute. Based on the age columns, some of the values were negative, it seemed unrealistic since there cannot be a date in future that has already occured. All values with negative were removed. The records also that had an age value less than 18 were removed i.e It was assumed that in order to create an account, one must be attained the adulthood age since he needs to use his details for account registration so any value less than it are removed e.g 75020 records were removed. Some features like month, day and month were engineered to be used to do some analysis. It will help in gaining insights on when the transactions were made.

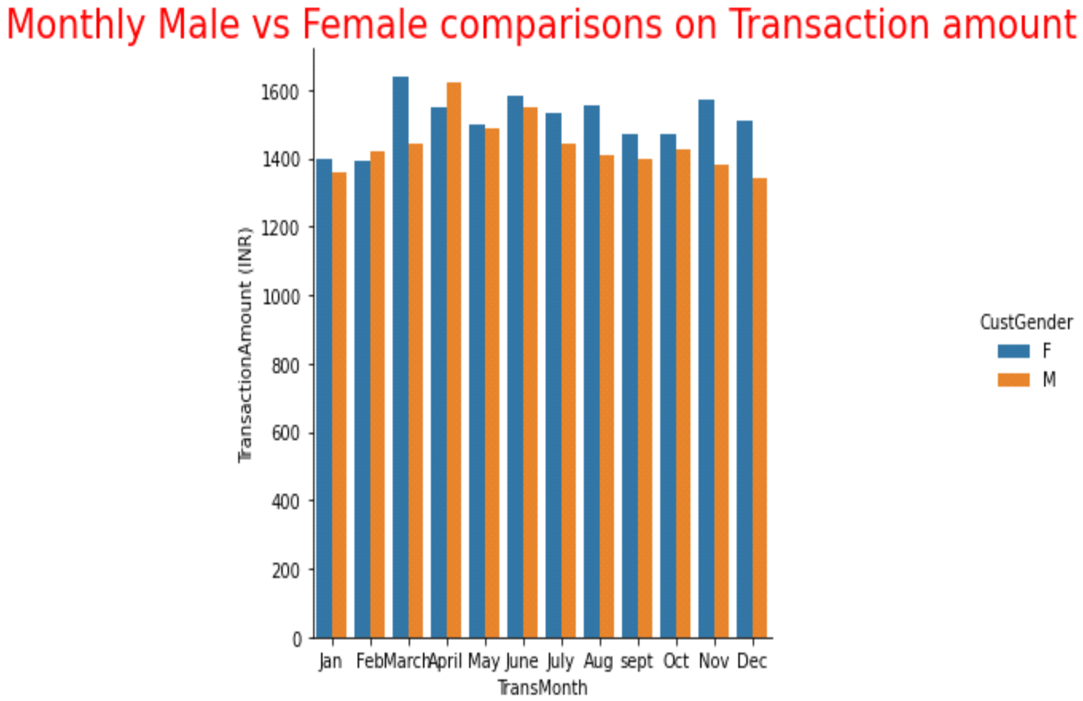


Figure: Males Vs Female Monthly Transaction amount.

Based on individual month transaction amount on total basis, unlike for averages, male customers have higher transaction rate than females. Banks should then consider providing more special incentives to males as they are likely to be active in terms of the number of customers. August and September had the most amount of transactions.

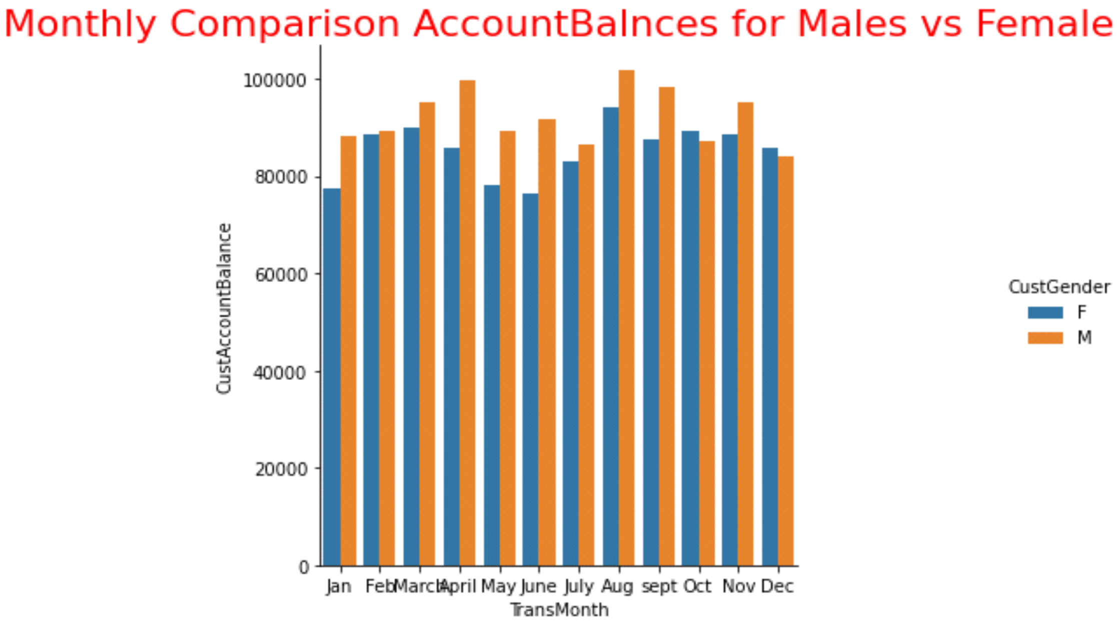


Figure: Male vs Female Monthly Account Balances

Based on the above visual, male customers predominantly have greater account balances as compared to their female counterparts. This can be attributed to the fact that their average transaction rates are low as compared to females. What makes them more in terms of TransactionAmount is that they are many unlike females. In the month of December, Females outweighed males in amount since maybe in December males tend to give out more money.

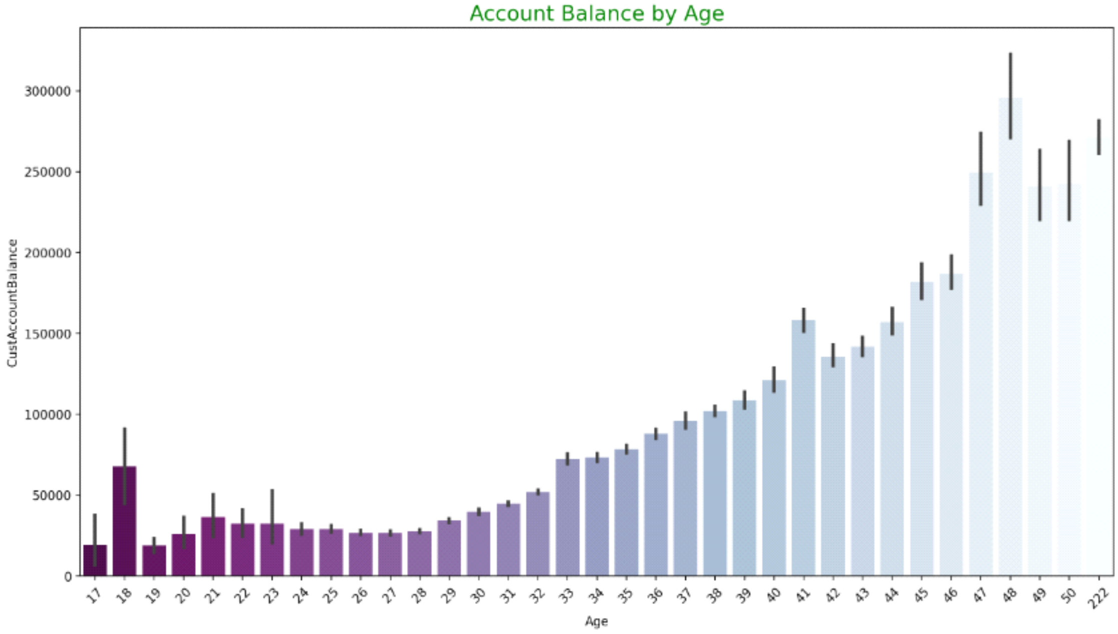


Figure: Account Balances Against Age

Based on the above, assuming that there is no age 222, which was assumed as a placeholder. The trend implies that account balances of customers increases with age, with a peak at 50 with few exemptions.

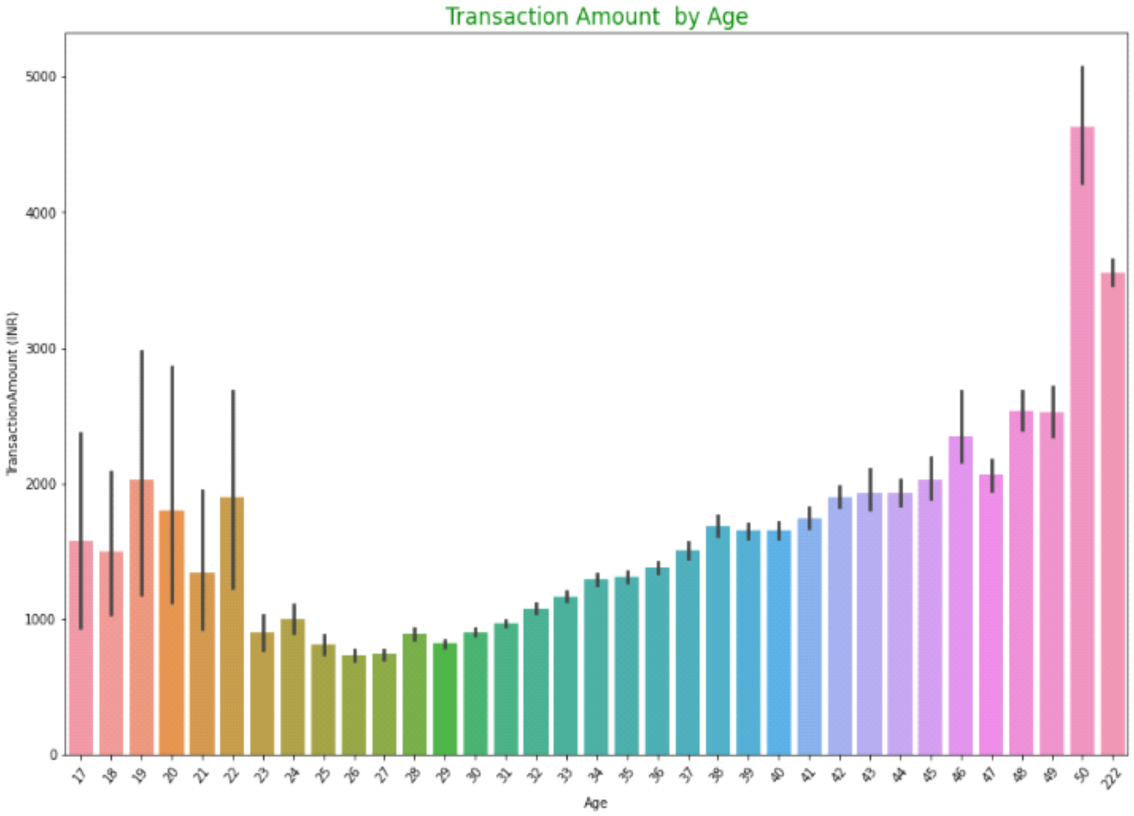


Figure: Transaction Amount Against Age

From the above graph, there is a similar trend in the case of the account balance. However, younger adult customers aged 19 to 24 years performed exorbitant transactions compared to their surrounding age groups. This might be a reason since younger adults between the ages of 18 to 24 years have their own career aspirations to fulfill as a consequence of which they generally have lavish and extravagant demands for fulfilling their passions and interests and for facilitating their development by all means. Also, from 29, the transaction amount increases, which might be due to the individual having a family which he is looking after and managing his livelihood.

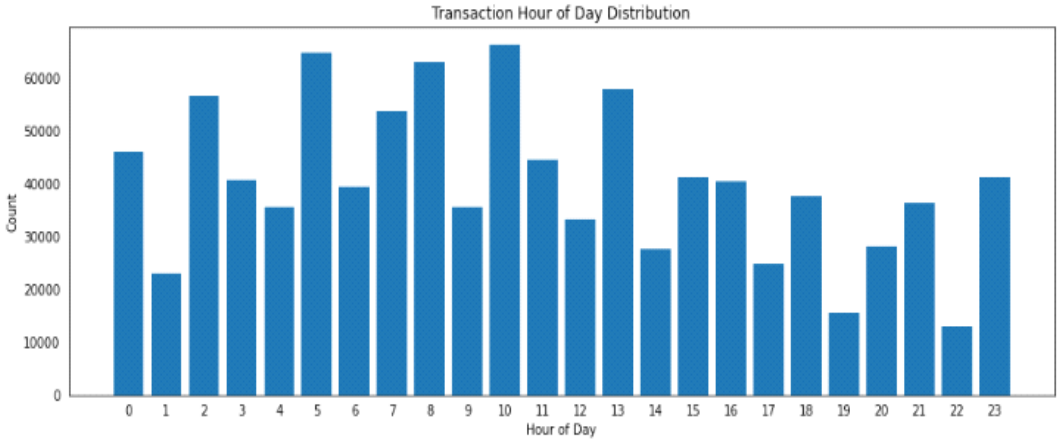
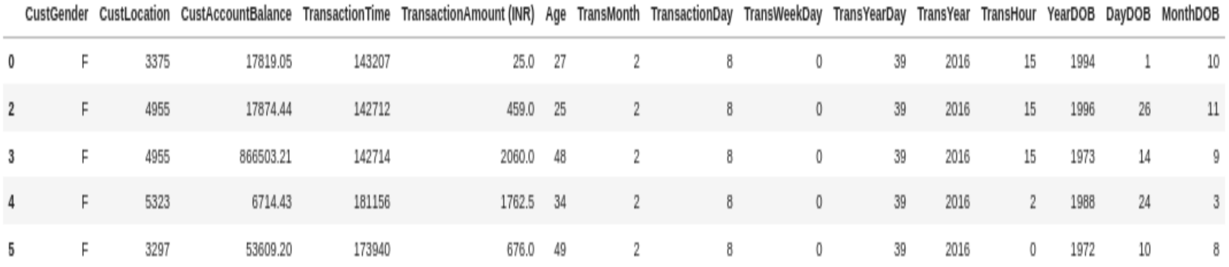


Figure: Transactions recorded in each Hour of the day

Using the above, more transactions are done at 0500hrs, 1000hrs, and 13hrs. This is probably the time in which they are waking up, going for morning tea and lunch. This made the end of EDA and the segmentation process start. At first, all columns that were not required were removed, and those that are only needed for clustering. The following are sample data outputs that resulted after final processing.

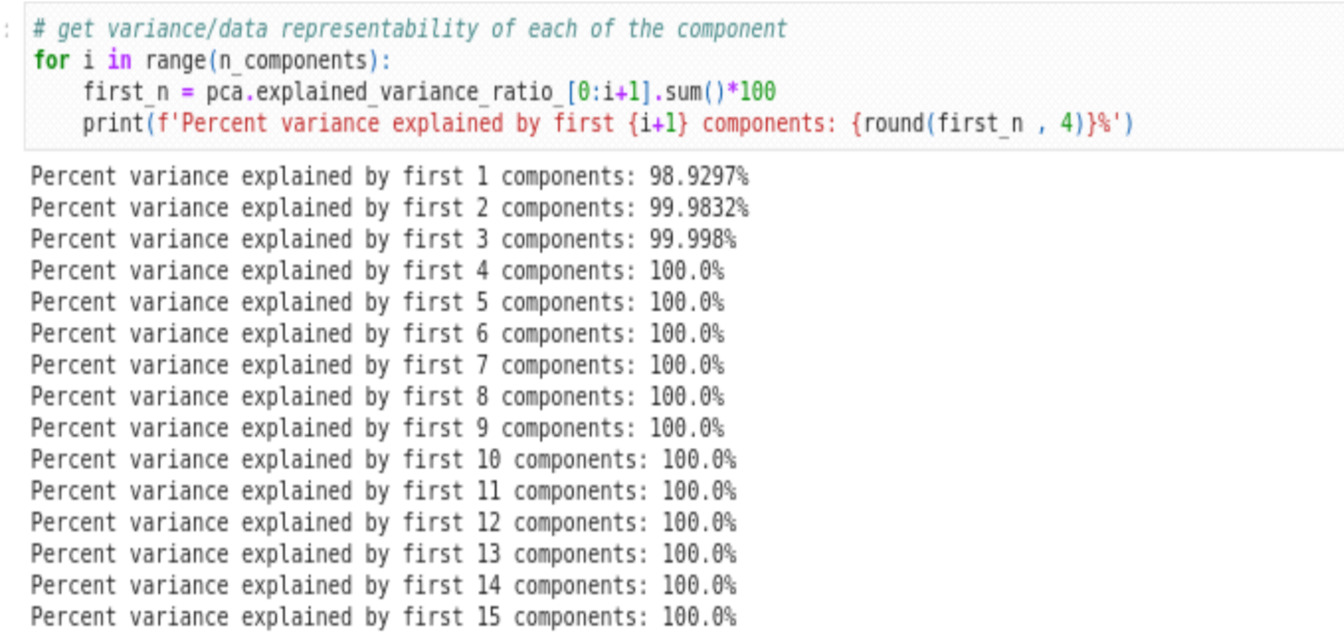


	CustGender	CustLocation	CustAccountBalance	TransactionTime	TransactionAmount (INR)	Age	TransMonth	TransactionDay	TransWeekDay	TransYearDay	TransYear	TransHour	YearDOB	DayDOB	MonthDOB
0	F	3375	17819.05	143207	25.0	27	2	8	0	39	2016	15	1994	1	10
2	F	4955	17874.44	142712	459.0	25	2	8	0	39	2016	15	1996	26	11
3	F	4955	866503.21	142714	2060.0	48	2	8	0	39	2016	15	1973	14	9
4	F	5323	6714.43	181156	1762.5	34	2	8	0	39	2016	2	1988	24	3
5	F	3297	53609.20	173940	676.0	49	2	8	0	39	2016	0	1972	10	8

Figure: Sample Final Data after Processing

5.3 Components Reduction and Clustering

The first process was to determine the number of new features that could represent most of the data in terms of explained variance. PCA was used for this purpose, and the following results were obtained.



```
# get variance/data representability of each of the component
for i in range(n_components):
    first_n = pca.explained_variance_ratio_[0:i+1].sum()*100
    print(f'Percent variance explained by first {i+1} components: {round(first_n , 4)}%')
```

```
Percent variance explained by first 1 components: 98.9297%
Percent variance explained by first 2 components: 99.9832%
Percent variance explained by first 3 components: 99.998%
Percent variance explained by first 4 components: 100.0%
Percent variance explained by first 5 components: 100.0%
Percent variance explained by first 6 components: 100.0%
Percent variance explained by first 7 components: 100.0%
Percent variance explained by first 8 components: 100.0%
Percent variance explained by first 9 components: 100.0%
Percent variance explained by first 10 components: 100.0%
Percent variance explained by first 11 components: 100.0%
Percent variance explained by first 12 components: 100.0%
Percent variance explained by first 13 components: 100.0%
Percent variance explained by first 14 components: 100.0%
Percent variance explained by first 15 components: 100.0%
```

Figure: Cumulative Explained variance for each Principal component

Majority of the variance in our data (>99%) can be encoded in 2 of the 15 dimensions. That's over 99% of the variance present can be encoded in 2 dimensions. This suggests that, there can be some underlying structure in a 2D visualization although some information of about 1% will be hidden. All the data can be encoded with only 3 dimensions (3D) by 100% score. Also at 2D 99.98% of the information can be encoded, by being able to gather much information using the first 2 components. So in this case, only 2 Dimensions will be created by reducing it using the PCA algorithm with 2 PCA components. Then compile the labels to this compressed dataset after assigning labels to all of our records to visualize the results of our clustering efforts.

After using PCA with 2 principal components features, the scatter plot below shows the results of the resultant 2 features.

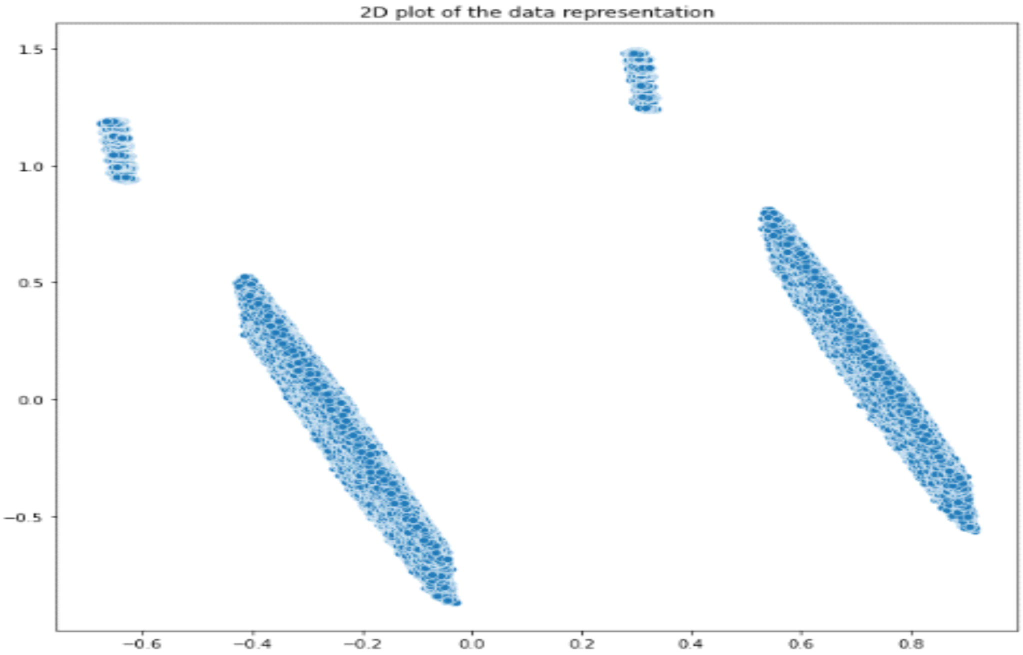


Figure: Final 2 Component Features scatter plot

The data is divided into some groups when visualized in 2D. From the above plot there seems to be a distinguishable pattern for groups. Some algorithms for clustering (In this case KMEANS will be used) will be used to perform checks for distinguishable patterns for groups/segments. An analytic method will be used to determine the best number of groups. To come up with the best number of clusters was done by fitting a clustering algorithm (Kmeans) using several different values of k, where k is the number of clusters with k range from 2 to 20. For each value of k, Evaluate the clustering results using the average Sum of the squared distance score. Then plot the results against each k value and identify which number of clusters leads to the best results in an elbow curve. Clusters will then be assigned to the required data point. The graph below shows the elbow curve method.

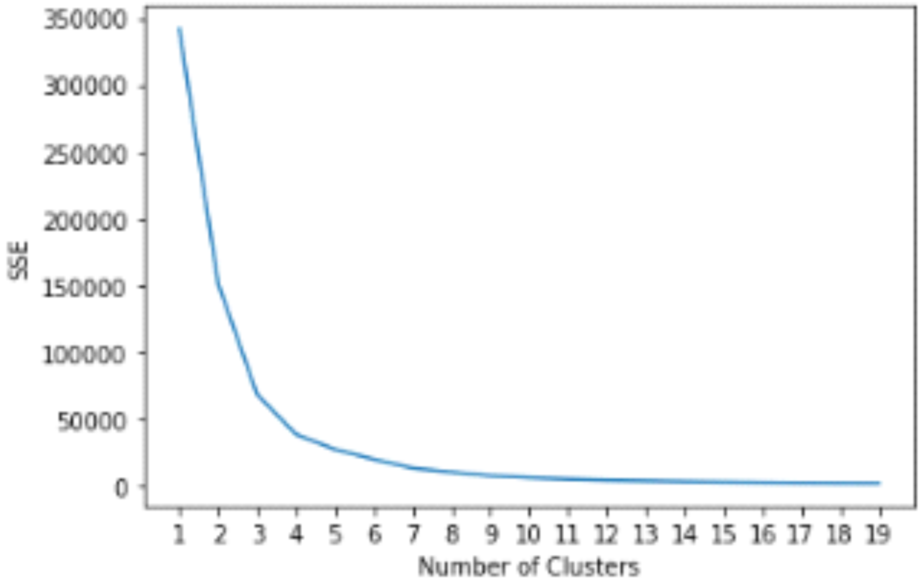


Figure: Elbow Curve of K=2 upto k=20

From the curve above, the elbow curve forms in about the 3rd or 4th cluster. These 2 might be the best clusters for segmentation. For easier analysis, both 2 clusters were used in the analysis. On training with 3 and 4 clusters, the graph below shows the results obtained of the data as a scatter plot for each cluster.

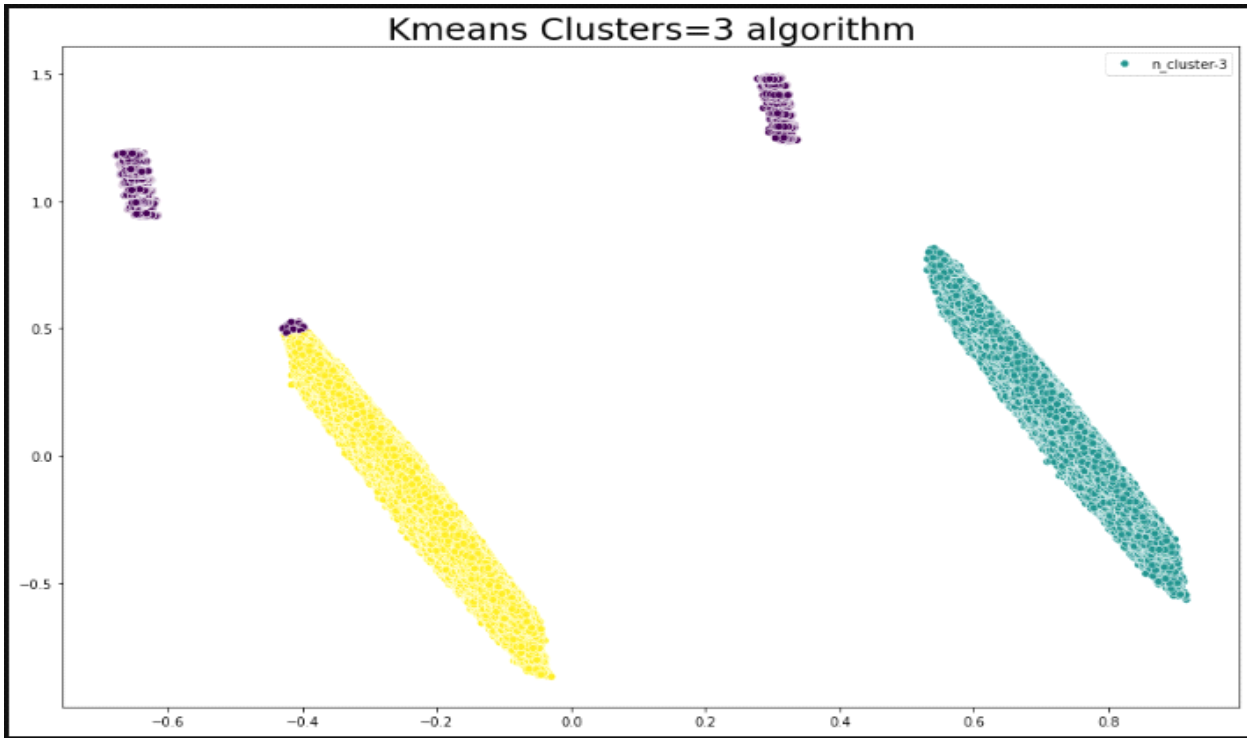


Figure: Scatter Plot with 3 clusters

There are well distinguishable features in this case although the centroid of one of the clusters seems to not be well placed when cluster 3 is observed.

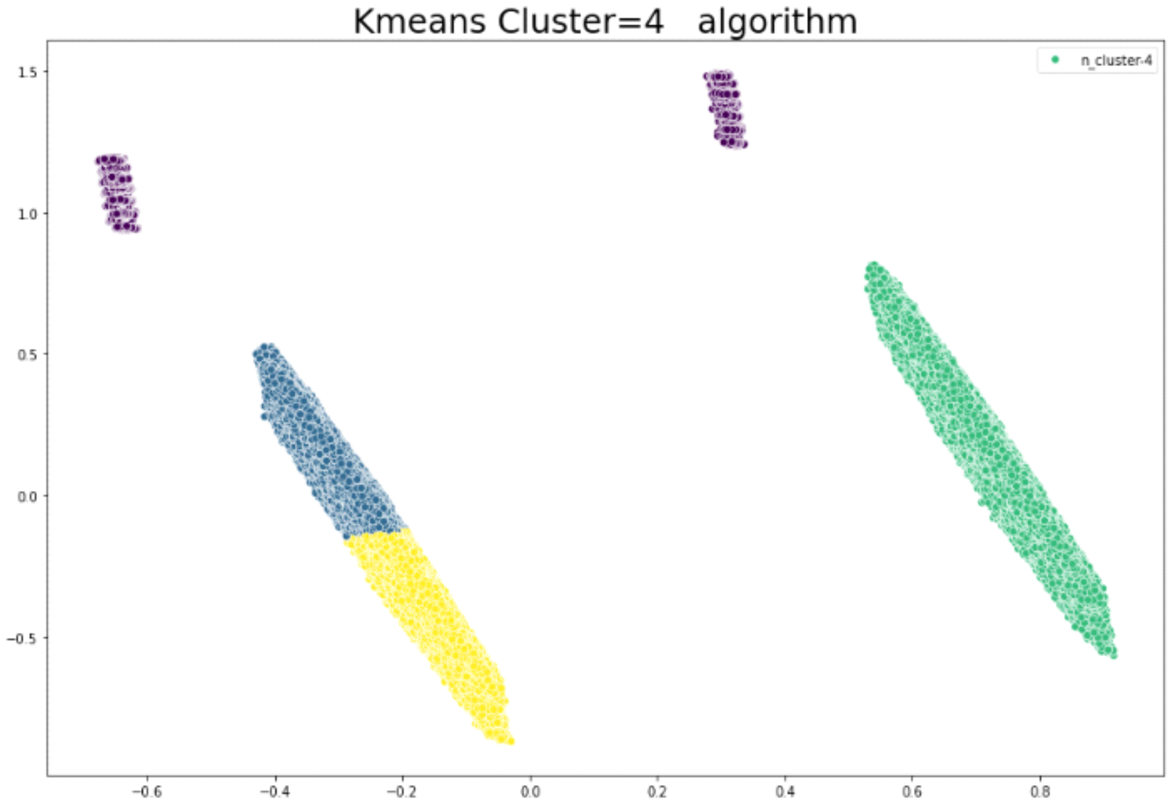


Figure: Scatter Plot with 4 clusters

Cluster 4 seemed to be the best. Segmentation will focus on 4 clusters but 3 will also be included. The clusters 3 and can split the data into distinct clusters as seen above by the graphs. Below graphs show distribution of each cluster for the two models trained.

5.3.1 Distribution of 4 cluster model

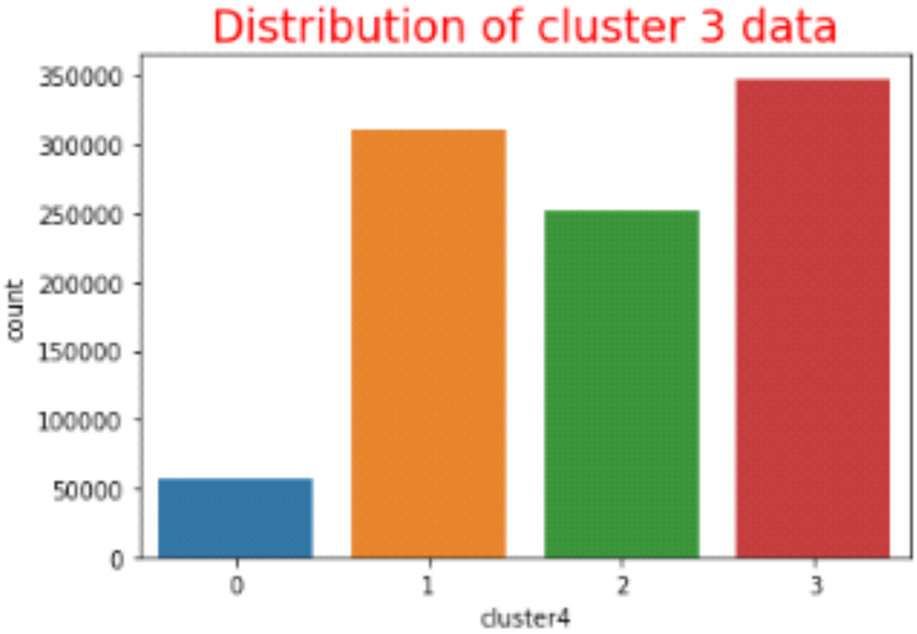


Figure: Cluster 3 labels distribution

The 4 clusters are not normally distributed. Each cluster has the following percentage distribution

0: 35%, 3 :32%, 1 :26%, 2: 5%

Clusters labeled 2 had the lowest data points of about 5%.

5.3.2 Distribution for the 3-cluster model

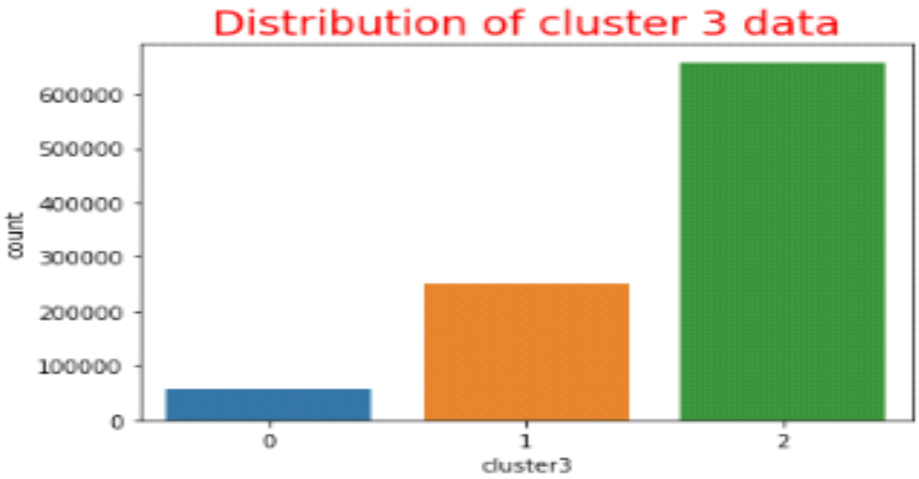


Figure: 3 clusters labels distribution

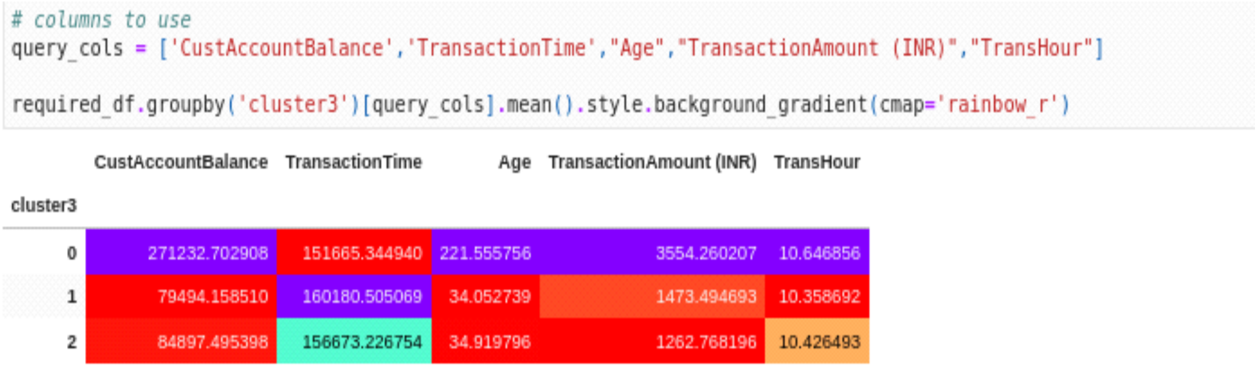
The 3 clusters are not normally distributed as some labels have large amounts of records. Each cluster has the following percentage distribution;

0 :26%, 2: 68%, 1: 5%

Much of the data is clustered at label 2, covering about 68% of the whole dataset.

5.4 Segmentation Results

5.4.1 Below are results that were obtained for 3 clusters



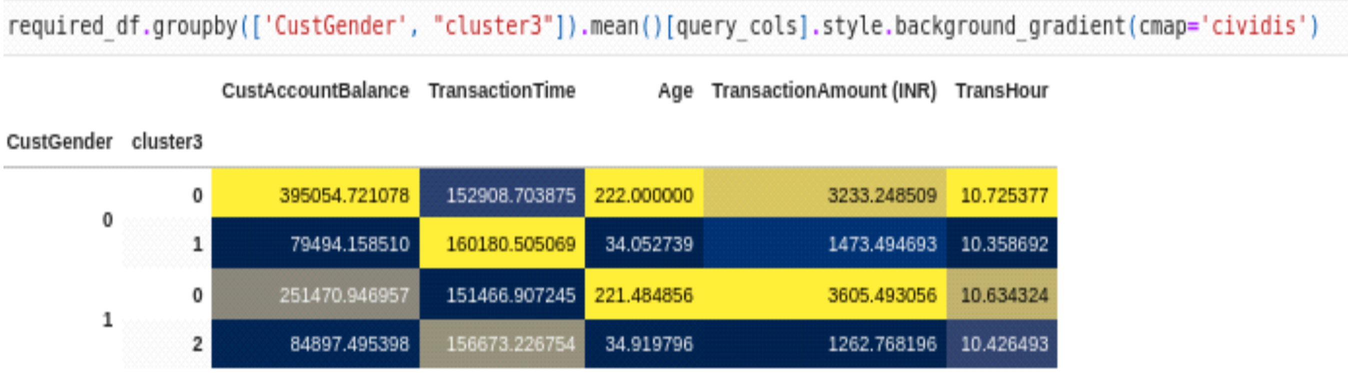
```
# columns to use
query_cols = ['CustAccountBalance','TransactionTime',"Age","TransactionAmount (INR)","TransHour"]

required_df.groupby('cluster3')[query_cols].mean().style.background_gradient(cmap='rainbow_r')
```

	CustAccountBalance	TransactionTime	Age	TransactionAmount (INR)	TransHour
cluster3					
0	271232.702908	151665.344940	221.555756	3554.260207	10.646856
1	79494.158510	160180.505069	34.052739	1473.494693	10.358692
2	84897.495398	156673.226754	34.919796	1262.768196	10.426493

Figure: 3 Clusters average analysis

There is a clear distinction between clusters labeled 1 and the rest. In terms of transaction Hour, both almost take a similar time on average. The values for Account Balance, Transaction Time and Transaction Amount is way higher than those labeled 0 and 2. This is the group that had their age put as placeholder i.e their age is 222 years. There is no big difference between clusters labeled 0 and 2.

When grouped by gender, the results were as follows.



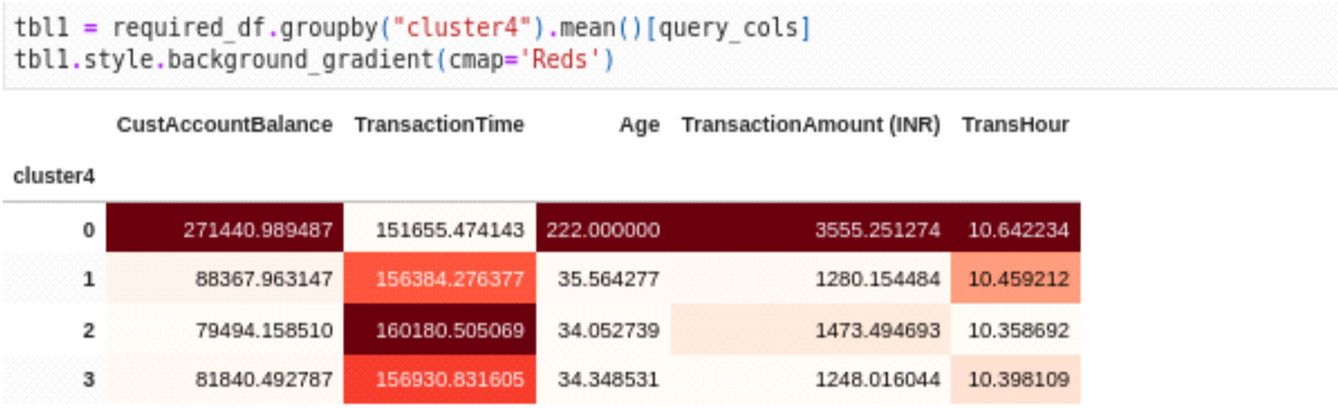
```
required_df.groupby(['CustGender', "cluster3"]).mean()[query_cols].style.background_gradient(cmap='cividis')
```

		CustAccountBalance	TransactionTime	Age	TransactionAmount (INR)	TransHour
CustGender	cluster3					
0	0	395054.721078	152908.703875	222.000000	3233.248509	10.725377
	1	79494.158510	160180.505069	34.052739	1473.494693	10.358692
1	0	251470.946957	151466.907245	221.484856	3605.493056	10.634324
	2	84897.495398	156673.226754	34.919796	1262.768196	10.426493

Figure: Figure 19. Three clusters Average analysis against gender

There was no Cluster 0 in Males. Majority of males up cluster 1 and 3. Clusters labeled 0 have only a few females. Clusters 0 and 2 had lowest amount of Transaction and Account Balance while cluster 1 had the highest. This suggests that both females and males in cluster 1 were heavy investors and transactors.

5.4.2 Below are results obtained for the 4 segments

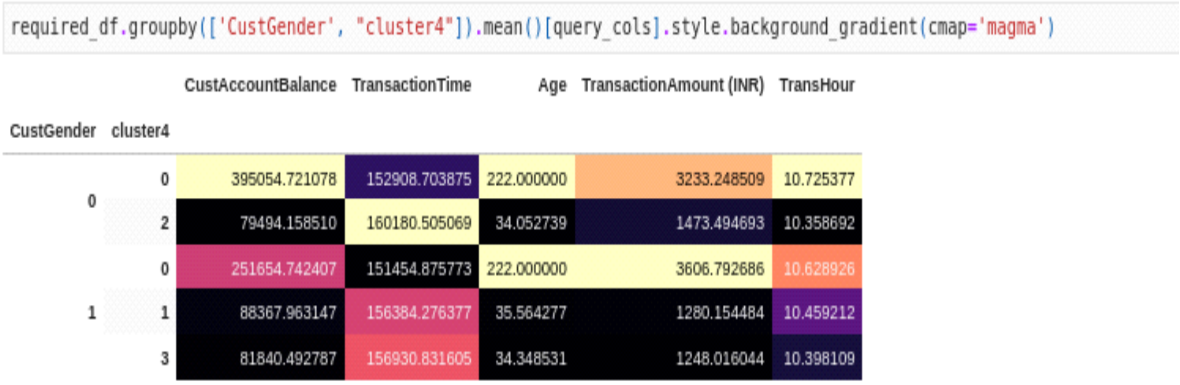


```
tbl1 = required_df.groupby("cluster4").mean()[query_cols]
tbl1.style.background_gradient(cmap='Reds')
```

cluster4	CustAccountBalance	TransactionTime	Age	TransactionAmount (INR)	TransHour
0	271440.989487	151655.474143	222.000000	3555.251274	10.642234
1	88367.963147	156384.276377	35.564277	1280.154484	10.459212
2	79494.158510	160180.505069	34.052739	1473.494693	10.358692
3	81840.492787	156930.831605	34.348531	1248.016044	10.398109

Figure: Four Clusters Average Analysis

A similar trend is observed. It can be seen that the clusters cannot be segmented by their age on the hour in which they did their transaction. Clusters labeled 0, 3, and 4 are closely related while cluster labeled 2 has the most distinct among them. When observed against the gender, the following results were obtained.



```
required_df.groupby(['CustGender', "cluster4"]).mean()[query_cols].style.background_gradient(cmap='magma')
```

CustGender	cluster4	CustAccountBalance	TransactionTime	Age	TransactionAmount (INR)	TransHour
0	0	395054.721078	152908.703875	222.000000	3233.248509	10.725377
	2	79494.158510	160180.505069	34.052739	1473.494693	10.358692
1	0	251654.742407	151454.875773	222.000000	3606.792686	10.628926
	1	88367.963147	156384.276377	35.564277	1280.154484	10.459212
	3	81840.492787	156930.831605	34.348531	1248.016044	10.398109

Figure: Four Clusters Average Analysis Against Gender

A similar trend is observed as above where clusteds 0, 1 and 3 have small values of account balances while clusters 2 have large sums of account balances and transaction amounts. Cluster 1 only has Females while cluster 0 only has Males.

5.5 Discussions

This research was performed to study the possibility of applying machine learning in segmenting customers based on similar financial characteristics in the banking sector. Several factors influence customer segmentation in the banking industry. This study has used a dataset with characteristics like transaction and customer IDs, gender, location, account balance, transaction time, and transaction amount. Many of these fall under demographics. Cluster analysis has helped in revealing specific characteristics and insights that will help the bank to understand their customers' needs and requirements so that the bank can create customized offers and custom plans to attract potential and profitable customers to sell their products and services to those holding few products that will lead to higher product penetration, higher customer retention, and reduce customer turnover. This will empower banks to nurture by targeting specific segments with suitable products and services, thus providing a more personalized approach that might lead the bank with appropriate marketing propositions, growth, and profitability.

In terms of gender, males transact more frequently than females. Males also have more account balances than females. The bank can then offer incentives that can lure more females to bank with them. For instance, the bank can offer flexible and convenient banking options, such as online and mobile banking, to make transactions more accessible for busy females. Educational resources and financial literacy workshops can also be offered by the bank to female account owners to empower them to make informed financial decisions and feel confident in their banking transactions. The bank can also create a welcoming and inclusive environment at bank branches and in advertising materials, featuring diverse female representation and highlighting the bank's commitment to gender equality and financial empowerment. Partnering with organizations and community groups that focus on women's empowerment and financial inclusion, to reach and engage with more female customers can be helpful. However, females transacted more amounts than the males. To encourage more savings for the females, the bank can offer special promotions and incentives, such as higher interest rates on savings accounts or lower fees on certain transactions, to encourage females to engage with the bank.

In terms of age, youth aged between 19 and 24 years old perform exorbitant transactions as compared to their surrounding age groups. Youths are often less experienced with managing money and may not have developed good financial habits. They tend to be more impulsive and may be more likely to make purchases without considering the long-term consequences. Also, they have less disposable income than older individuals, so they may feel the need to spend more to keep up with their peers or maintain a certain lifestyle. Youths are also more likely to be influenced by advertising and peer pressure, which can encourage them to spend more on fashion, technology, and other consumer goods. People beyond 33 years of age transact more. Which implies they also spend more. This can be explained by their financial stability. People in their mid-to-late thirties and beyond may have more established careers and higher income levels, allowing them to make larger purchases and investments. Also, as people get older, they may have different financial needs and priorities. For example, they may be saving for retirement, buying a home, or supporting a family. These life events may require more financial transactions. Banks can target this population for credit. Older individuals may have a longer credit history, making them more likely to be approved for loans and credit cards with higher limits. This can enable them to make larger purchases. People in this category have higher risk tolerance. As people age, they may become more risk-averse, leading them to make more conservative financial decisions and transact more frequently to protect their assets.

The behavioural aspect of the dataset was recorded as transaction time, or the length of time that clients spend when transacting with the bank. Females took more time to make a transaction than males. Studies have shown that men and women tend to approach decision making differently, with men often being more decisive and taking less time to make a decision. This could lead to men being more efficient and faster at completing transactions in a bank. Women may feel more pressure to be polite and patient in social interactions, which could lead to them taking more time to ask questions, clarify instructions, or engage in small talk with bank staff. Also, females are often expected to take on a greater share of household and caregiving responsibilities, which can leave them with less time and energy for tasks like banking. As a result, they may prioritize their transactions and take more time to ensure that they are completed accurately and efficiently. Also, access to financial knowledge and resources differs. Women may have less access to financial education, support, and resources, which could lead to them feeling less confident and competent in financial transactions. This could result in them taking more time to ask questions and seek guidance from bank staff. Banks can leverage these aspects to empower women transact quickly.

Geographical location can also be a key factor in customer segmentation. Customers in different regions may have different needs and preferences. Location plays a significant role in customer segmentation because it helps to identify and understand the unique needs and preferences of customers in different regions. Location-based data can also be used to identify potential new customers and target them with personalized marketing campaigns. For example, a bank can analyze data on economic indicators and population density in different regions to identify areas with high potential for growth and target these areas with tailored marketing efforts.

Several machine learning algorthms can be used for the clustering tasks. K-means clustering algorithm is commonly used because it is simple, easy to implement, and computationally efficient. It is also able to handle large datasets and can produce reliable results. Additionally, it is flexible and can be used for a variety of clustering tasks. K-means clustering algorithm was used to segment customers in banks because it is a simple and efficient method for identifying groups or clusters within a dataset. It is particularly useful for segmenting customers based on their financial behaviors and transactions, as it allows for the creation of homogenous groups of customers with similar characteristics. This can help banks better understand and target their customer base, and provide more personalized products and services. Additionally, the algorithm is fast and easy to implement, making it a popular choice for customer segmentation in the banking industry.

Chapter 6. Conclusion and Recommendations

6.1 Conclusion

Customer segmentation is critical in helping institutions such as banks better offer deals and services to their customers. As discovered, banks have in the past used trivial methods in segmenting their customers with few features such as age and demographics used to divide their customers into distinct groups for personalized experiences. Banks have failed to adapt to new approaches of personalization and have instead focused on how to provide the next best product and offers. A personalized experience is essential in abstracting the complex relationship between customers and business entities as banks. Bank customers seek consistent and bespoke experiences across all their touchpoints while buying financial products and managing their financial assets. Therefore, it is important for banks to critically evaluate and provide the most appropriate next best experience for their customers.

It is important to understand that personalization in banking is not merely about selling. Personalization in banking involves providing the best services, products, and advice that is within the right context relevant to the customer. This calls for recursive learning of customers to observe changing customer needs and wants, consequently offering the best solutions to their problems and pain points. Artificial intelligence techniques used in this research study have depicted how machine learning techniques and feature engineering on bank customers can be employed to understand different customer classifications better. Classifying customers using the most appropriate attributes or features is crucial to obtaining better classification models with greater accuracy and efficiency.

Applying machine learning to big data such as the one used in the study with over one million records can prove valuable in better understanding trends and patterns and handling multi-dimensional and multi-variety data. Viewing bank data in alternative dimensions such as weeks and times of transactions gives beneficial and novel insights, as observed in the data analysis and exploration. However, using primitive ways of viewing and segmenting bank customer data such as gender is still encouraged, and machine learning techniques are advocated to augment the current processes of customer segmentation in banking. As observed in the analysis, beneficial insights can be acquired from banking customer data by viewing different features with respect to gender. For example, it was observed from the analysis that males transact and keep more money in their accounts. This information is relevant in allowing banks to offer tailored products to male customers in India.

Artificial intelligence techniques like artificial neural networks, deep learning, and machine learning are yet to be fully leveraged in the banking sector to execute its customer segmentation better. Banks need to change their current way of selling products to customers and focus their product selling on more personalized and custom customer experiences. This will ensure better responses as their selling will be more targeted, making less uncalled-for sales for customers. The future of better banking lies in how well banks will utilize their ever-growing data to streamline their processes. Artificial intelligence methods are integral to this.

6.2 Recommendations

Following the research study and analysis of the dataset aforementioned, the following are some recommendations for the betterment of customer segmentation in banking institutions to improve their customers' experiences:

- Banks should endeavour to collect more complete data. Mechanisms should be put in to ensure there are no missing values in the data. These mechanisms may include ensuring key fields are marked as required and ensuring proper technologies are used to collect, store, and process data. Various integrity checks such as domain, referential, and data value integrity should be checked for any violations before storage and processing.
- Banks should assemble data quality assurance teams responsible for ensuring bank data is of high quality. Key metrics such as relevance, timeliness, completeness, accuracy, and reliability of data should be upheld.
- Banks should ensure that information that is critical is determined beforehand; this will ensure that the data collected is rich and useful insights can be derived from it. Early capturing of information needs will prevent changes from occurring later and allows for the creation of better systems that are data-driven.
- Banks should strive to employ artificial intelligence techniques such as machine learning, deep learning, feature engineering, and artificial neural networks to analyse their data and perform critical operations such as correlation analysis, clustering, anomaly detection, sensitivity analysis, and dependency modeling.

- More emphasis should be put on better-personalized experiences by banks instead of selling products to their customers. Customer experiences should spearhead banks' endeavours in selling financial products.
- Banking institutions ought to ensure that they use different dimensions of data to visualize their customers. As opposed to the primitive ways of viewing customers, where attributes such as location, gender, and age were the only ones used to segment customers, banks should utilize other dimensions such as time of transactions, days of the year, and occupation to view and segment customers.
- Measures should be put during the analysis of banks' data to ensure that security, confidentiality, and privacy are not compromised. Procedures in the extraction, transformation and loading stages should be done securely with appropriate tools to guarantee data safety.
- Banks should improve their relationships with their customers by becoming more in touch with them. Banks should communicate better with their customers and have better data-driven approaches to understand what makes their customers happy. Banks ought to provide useful information to their customers that are relevant to their problems and needs. Therefore, banks need to understand their customers better to give them relevant information within context.
- Banks should integrate better customer support and employee feedback systems as important information can be collected through such channels. Banks need to use intelligent customer relationship management systems and chatbots that learn and adapt to customer queries and behaviour.

7. References

Abidar, L., Zaidouni, D., and Ennouaary, A. 2020. Customer segmentation with machine learning: New strategy for targeted actions. Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications,

Aliyu, A. A., Bello, M. U., Kasim, R., and Martin, D. 2014. Positivist and non-positivist paradigm in social science research: Conflicting paradigms or perfect partners. *J. Mgmt. & Sustainability*, *4*, 79.

Astriani, W., and Trisminingsih, R. 2016. Extraction, transformation, and loading (ETL) module for hotspot spatial data warehouse using Geokettle. *Procedia Environmental Sciences*, *33*, 626-634.

Bholat, D. 2015. Big data and central banks. *Big Data & Society*, *2*(1), 2053951715579469.

Brito, P. Q., Soares, C., Almeida, S., Monte, A., and Byvoet, M. 2015. Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing*, *36*, 93-100.

Bunge, M. 2012. *Epistemology & Methodology I:: Exploring the World* (Vol. 5). Springer Science & Business Media.

Cilimkovic, M. 2015. Neural networks and back propagation algorithm. *Institute of Technology Blanchardstown, Blanchardstown Road North Dublin*, *15*(1).

Cortina Lorente, J. J., and Schmukler, S. L. 2018. The fintech revolution: a threat to global banking? *World Bank Research and Policy Briefs* (125038).

Crowley, F., and Jordan, D. 2017. Does more competition increase business-level innovation? Evidence from domestically focused firms in emerging economies. *Economics of Innovation and New Technology*, *26*(5), 477-488.

Cuadros, A. J., and Domínguez, V. E. 2014. Customer segmentation model based on value generation for marketing strategies formulation. *Estudios Gerenciales*, *30*(130), 25-30.

Cuzzocrea, A. 2013. Analytics over big data: Exploring the convergence of datawarehousing, OLAP and data-intensive cloud infrastructures. 2013 IEEE 37th Annual Computer Software and Applications Conference,

Da Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., and dos Reis Alves, S. F. 2017. Artificial neural networks. *Cham: Springer International Publishing*, *39*.

Dogan, O., Ayçin, E., and Bulut, Z. 2018. Customer segmentation by using RFM model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*, *8*.

Eberendu, A. C. 2016. Unstructured Data: an overview of the data of Big Data. *International Journal of Computer Trends and Technology*, *38*(1), 46-50.

Edastama, P., Dudhat, A., and Maulani, G. 2021. Use of Data Warehouse and Data Mining for Academic Data: A Case Study at a National University. *International Journal of Cyber and IT Service Management*, *1*(2), 206-215.

Erl, T., Khattak, W., and Buhler, P. 2016. *Big data fundamentals: concepts, drivers & techniques*. Prentice Hall Press.

Ferraro, M. B., and Giordani, P. 2020. Soft clustering. *Wiley Interdisciplinary Reviews: Computational Statistics*, *12*(1), e1480.

Gammie, E., Hamilton, S., and Gilchrist, V. 2017. Focus group discussions. *The Routledge companion to qualitative accounting research methods*, 372-386.

Gichuru, M. J., and Limiri, E. K. 2017. Market Segmentation As A Strategy For Customer Satisfaction And Retention. *International Journal of Economics, Commerce and Management. United Kingdom Vol. V*, *12*, 544-553.

Hassan, S. S., and Craft, S. 2012. Examining world market segmentation and brand positioning strategies. *Journal of Consumer marketing*.

He, X., and Li, C. 2016. The research and application of customer segmentation on e-commerce websites. 2016 6th International Conference on Digital Home (ICDH),

Hofer, I. S., Gabel, E., Pfeffer, M., Mahbouba, M., and Mahajan, A. 2016. A systematic approach to creation of a perioperative data warehouse. *Anesthesia & Analgesia*, *122*(6), 1880-1884.

Hurwitz, J., Nugent, A., Halper, F., and Kaufman, M. 2013. Big Data. *New York*.

Jayasree, V., and Balan, R. V. S. 2013. A review on data mining in banking sector. *American Journal of Applied Sciences*, *10*(10), 1160.

Jordan, M. I., and Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Kalyanpur, A., Boguraev, B. K., Patwardhan, S., Murdock, J. W., Lally, A., Welty, C., Prager, J. M., Coppola, B., Fokoue-Nkoutche, A., and Zhang, L. 2012. Structured data and inference in DeepQA. *IBM Journal of Research and Development*, *56*(3.4), 10: 11-10: 14.

Kashwan, K. R., and Velu, C. 2013. Customer segmentation using clustering and data mining techniques. *International Journal of Computer Theory and Engineering*, *5*(6), 856.

Keiningham, T., Aksoy, L., Bruce, H. L., Cadet, F., Clennell, N., Hodgkinson, I. R., and Kearney, T. 2020. Customer experience driven business model innovation. *Journal of Business Research*, *116*, 431-440.

Kettouch, M. S., Luca, C., Hobbs, M., and Fatima, A. 2015. Data integration approach for semi-structured and structured data (Linked Data). 2015 IEEE 13th international conference on industrial informatics (INDIN),

Kovács, T., Ko, A., and Asemi, A. 2021. Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis. *Journal of Big Data*, *8*(1), 1-25.

Krishnan, K. 2013. *Data warehousing in the age of big data*. Newnes.

Lolli, F., Gamberini, R., Regattieri, A., Balugani, E., Gatos, T., and Gucci, S. 2017. Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics*, *183*, 116-128.

Mbama, C. I., and Ezepue, P. O. 2018. Digital banking, customer experience and bank financial performance: UK customers' perceptions. *International Journal of Bank Marketing*.

Mihova, V., and Pavlov, V. 2018. A customer segmentation approach in commercial banks. AIP conference proceedings,

Mills, J., and Birks, M. 2014. *Qualitative methodology: A practical guide*. Sage.

Min, S., Zhang, X., Kim, N., and Srivastava, R. K. 2016. Customer acquisition and retention spending: An analytical model and empirical investigation in wireless telecommunications markets. *Journal of Marketing Research*, *53*(5), 728-744.

Mukherjee, R., and Kar, P. 2017. A comparative review of data warehousing ETL tools with new trends and industry insight. 2017 IEEE 7th International Advance Computing Conference (IACC),

Nikumanesh, E., and Albadvi, A. 2014. Customer's life–time value using the RFM model in the banking industry: a case study. *International Journal of Electronic Customer Relationship Management*, *8*(1-3), 15-30.

Nobar, H. B. K., and Rostamzadeh, R. 2018. The impact of customer satisfaction, customer experience and customer loyalty on brand power: empirical evidence from hotel industry. *Journal of Business Economics and Management*, *19*(2), 417-430.

O'Grady, P. 2014. *Relativism*. Routledge.

Omrani, H., Alizadeh, A., Emrouznejad, A., and Oveysi, Z. 2022. A novel best-worst-method two-stage data envelopment analysis model considering decision makers' preferences: An application in bank branches evaluation. *International Journal of Finance & Economics*.

Osei, F., Ampomah, G., Kankam-Kwarteng, C., Bediako, D. O., and Mensah, R. 2021. Customer Satisfaction Analysis of Banks: The Role of Market Segmentation. *Science Journal of Business and Management*, *9*(2), 126.

Östman, L., and Wickman, P.-O. 2014. A pragmatic approach on epistemology, teaching, and learning. *Science Education*.

Pamučar, D., Ecer, F., Cirovic, G., and Arlasheedi, M. A. 2020. Application of improved best worst method (BWM) in real-world problems. *Mathematics*, *8*(8), 1342.

Paruchuri, H. 2019. Market segmentation, targeting, and positioning using machine learning. *Asian Journal of Applied Science and Engineering*, *8*, 7-14.

Qaddoura, R., Faris, H., and Aljarah, I. 2020. An efficient clustering algorithm based on the k-nearest neighbors with an indexing ratio. *International Journal of Machine Learning and Cybernetics*, *11*(3), 675-714.

Raiter, O. 2021. Segmentation of Bank Consumers for Artificial Intelligence Marketing. *International Journal of Contemporary Financial Issues*, *1*(1), 39-54.

Raju, P. S., Bai, D. V. R., and Chaitanya, G. K. 2014. Data mining: Techniques for enhancing customer relationship management in banking and retail industries. *International Journal of Innovative Research in Computer and Communication Engineering*, *2*(1), 2650-2657.

Sabbeh, S. F. 2018. Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications*, *9*(2).

Sadoghi, M., Bhattacherjee, S., Bhattacharjee, B., and Canim, M. 2016. L-store: A real-time OLTP and OLAP system. *arXiv preprint arXiv:1601.04084*.

Santoso, L., Singh, B., Rajest, S., Regin, R., and Kadhim, K. 2020. A genetic programming approach to binary classification problem. *EAI Endorsed Transactions on Energy Web*, *8*(31), e11.

Sharma, S., and Jain, R. 2014. Modeling ETL Process for data warehouse: an exploratory study. 2014 Fourth International Conference on Advanced Computing & Communication Technologies, Smeureanu, I., Ruxanda, G., and Badea, L. M. 2013. Customer segmentation in private banking sector using machine learning techniques. *Journal of Business Economics and Management*, *14*(5), 923-939.

Smith, B. 2012. Ontology. In *The furniture of the world* (pp. 47-68). Brill.

Song, I. Y., and Zhu, Y. 2016. Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364-373.

Song, J., Guo, C., Wang, Z., Zhang, Y., Yu, G., and Pierson, J.-M. 2015. HaoLap: A Hadoop based OLAP system for big data. *Journal of Systems and Software*, *102*, 167-181.

Sternberg, R. J., Guyote, M. J., and Turner, M. E. 2021. Deductive reasoning. *Aptitude, learning, and instruction*, 219-246.

Sun, Z., Strang, K., and Li, R. 2018. Big data with ten big characteristics. Proceedings of the 2nd International Conference on Big Data Research,

Sun, Z., Zou, H., and Strang, K. 2015. Big data analytics as a service for business intelligence. Conference on e-Business, e-Services and e-Society,

Taherparvar, N., Esmaeilpour, R., and Dostar, M. 2014. Customer knowledge management, innovation capability and business performance: a case study of the banking industry. *Journal of knowledge management*.

Vaisman, A., and Zimányi, E. 2014. Data warehouse systems. *Data-Centric Systems and Applications*.

Walter, M., and Andersen, C. 2016. *Indigenous statistics: A quantitative research methodology*. Routledge.

Wang, M., Cho, S., and Denton, T. 2017. The impact of personalization and compatibility with past experience on e-banking usage. *International Journal of Bank Marketing*.

Wieringa, J., Kannan, P., Ma, X., Reutterer, T., Risselada, H., and Skiera, B. 2021. Data analytics in a privacy-concerned world. *Journal of Business Research*, *122*, 915-925.

Xiao, Y., and Watson, M. 2019. Guidance on conducting a systematic literature review. *Journal of Planning Education and Research*, *39*(1), 93-112.

Zakir, J., Seymour, T., and Berg, K. 2015. Big Data Analytics. *Issues in Information Systems*, *16*(2).

Ziafat, H., and Shakeri, M. 2014. Using data mining techniques in customer segmentation. *Journal of Engineering Research and Applications*, *4*(9), 70-79.