B.Tech. BCSE497J - Project-I

Market Segmentation and Analysis Using ML

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Programme

by

21BCE2405 TORAN V ATHANI

Under the Supervision of

PUSHPA GOTHWAL

Assistant Professor Sr. Grade 1

School of Computer Science and Engineering (SCOPE)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

November 2024

DECLARATION

I hereby declare that the project entitled **Market Segmentation and Analysis Using Machine Learning** submitted by me, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering* to VIT is a record of bonafide work carried out by me under the supervision of Prof. Pushpa Gothwal.

I further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

Date : 20-11-2024

Signature of the Candidate

CERTIFICATE

This is to certify that the project entitled **Market Segmentation and Analysis Using Machine Learning** submitted by **Toran V Athani (21BCE2405)**, **School of Computer Science and Engineering**, VIT, for the award of the degree of *Bachelor of Technology in Computer Science and Engineering*, is a record of bonafide work carried out by him under my supervision during Fall Semester 2024-2025, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The project fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore
Date : 20-11-2024

Signature of the Guide

(18830)
Examiner(s)

Dr. Umadevi K.S
School of Computer Science and Engineering

ACKNOWLEDGEMENTS

I am deeply grateful to the management of Vellore Institute of Technology (VIT) for providing me with the opportunity and resources to undertake this project. Their commitment to fostering a conducive learning environment has been instrumental in my academic journey. The support and infrastructure provided by VIT have enabled me to explore and develop my ideas to their fullest potential.

My sincere thanks to Dr. Ramesh Babu K, the Dean of the School of Computer Science and Engineering (SCOPE), for his unwavering support and encouragement. His leadership and vision have greatly inspired me to strive for excellence. The Dean’s dedication to academic excellence and innovation has been a constant source of motivation for me. I appreciate his efforts in creating an environment that nurtures creativity and critical thinking.

I express my profound appreciation to Dr .K.S. Umadevi, the Head of the School of Computer Science and Engineering (SCOPE), for her insightful guidance and continuous support. His/her expertise and advice have been crucial in shaping the direction of my project. The Head of Department’s commitment to fostering a collaborative and supportive atmosphere has greatly enhanced my learning experience. His/her constructive feedback and encouragement have been invaluable in overcoming challenges and achieving my project goals.

I am immensely thankful to my project supervisor, Pushpa Gothwal, for her dedicated mentorship and invaluable feedback. Her patience, knowledge, and encouragement have been pivotal in the successful completion of this project. My supervisor’s willingness to share his/her expertise and provide thoughtful guidance has been instrumental in refining my ideas and methodologies. His/her support has not only contributed to the success of this project but has also enriched my overall academic experience.

Thank you all for your contributions and support.

TORAN V ATHANI

ABSTRACT

Market segmentation is a pivotal aspect of business analytics, enabling organizations to identify and target distinct customer groups based on shared characteristics. This project focuses on the niche market applications of segmentation techniques, specifically within specialized industries such as luxury goods, medical services, and high-tech domains. By leveraging machine learning, the project aims to group customers with similar spending patterns, offering insights into tailoring marketing strategies and product offerings to better serve each segment.

The workflow begins with a comprehensive understanding of business objectives and collecting diverse datasets that include customer demographics and behavioral data. Exploratory Data Analysis (EDA) is conducted to identify key patterns, trends, and outliers within the data. Pre-processing techniques such as missing value imputation, outlier detection, and scaling are applied to prepare the dataset for model training. Clustering algorithms like K-Means and hierarchical clustering are utilized to uncover meaningful customer segments.

This project introduces a unique perspective on market segmentation by focusing on niche market applications-specialized industries such as luxury goods, medical services, and high-tech domains that often require tailored marketing strategies. Unlike traditional segmentation approaches, which cater to broader markets, this project emphasizes the application of clustering techniques to uncover granular insights in these high-value industries.

The use of Streamlit to create an interactive API that identifies customer clusters and visualizes results in real time is an innovative step toward practical usability. A clear focus on identifying customer segments with distinct spending patterns helps businesses not only optimize marketing strategies but also discover underrepresented product categories and growth opportunities within specialized markets.

Keywords - Market Segmentation, Machine Learning, K-Means Clustering, Hierarchical Clustering, Exploratory Data Analysis, Customer Segmentation, Business Analytics, Real- Time Deployment, Cloud Services, Marketing Strategies.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

EDA	Exploratory Data Analysis
GMM	Gaussian Mixture Model
RFM	Recency, Frequency, Monetary Analysis
CLV	Customer Lifetime Value
CRM	Customer Relationship Management
GDPR	General Data Protection Regulation
CCPA	California Consumer Privacy Act
KPI	Key Performance Indicator
ROMI	Return on Marketing Investment
KNN	K-Nearest Neighbor

1. INTRODUCTION

1.1 Background

In today’s competitive business landscape, understanding customer behaviour and tailoring products and marketing strategies to specific customer groups is essential for success. Market segmentation, the process of dividing a broad consumer or business market into sub-groups based on shared characteristics, enables businesses to target specific segments effectively. Traditional segmentation methods are often manual and time-consuming, relying on predefined rules or assumptions that may not capture the full complexity of customer data.

With the rise of big data and machine learning, companies now have the ability to analyse vast amounts of data and uncover hidden patterns in customer behaviour, demographics, and preferences. Machine learning algorithms can automatically detect and segment customers into meaningful groups, offering a more dynamic, accurate, and data-driven approach to segmentation.

This project explores the use of machine learning for market segmentation, guiding the development of an end-to-end solution. The project starts from business problem identification, collecting and exploring customer data, and implementing advanced machine learning algorithms to identify segments. After model evaluation, the final step involves deploying the solution to ensure it can be used by businesses in real-time for decision-making. This approach improves the efficiency and accuracy of market segmentation, helping businesses optimize marketing strategies, enhance customer targeting, and ultimately drive revenue growth.

Market segmentation has long been a cornerstone of effective marketing, providing businesses with the ability to identify, understand, and serve their customers better. However, as markets expand and diversify, the complexity of customer needs and behaviors also increases. Traditional segmentation methods often fail to adapt to this complexity, leading to generalized marketing efforts that may not resonate with every customer group. This inefficiency can result in missed opportunities and suboptimal allocation of resources.

Machine learning addresses these challenges by offering a scalable and adaptable framework for market segmentation. Unlike manual approaches, machine learning leverages algorithms that can process high-dimensional data from multiple sources, such as purchase histories, website interactions, social media activity, and customer feedback. These data points, when analyzed collectively, provide a holistic view of customer profiles. By capturing both explicit and implicit patterns in the data, machine learning models enable more granular segmentation that aligns closely with real-world customer behaviors.

For instance, clustering algorithms such as k-means and hierarchical clustering group customers based on similarities across multiple attributes, while classification models can predict which segment a new customer is likely to belong to. More advanced techniques, such as neural networks and deep learning, are also being employed for complex segmentation tasks. These methods excel at identifying non-linear patterns and relationships in data, which are often critical for understanding niche customer segments.

The project also emphasizes the importance of data preprocessing and feature selection. Raw customer data is often noisy and inconsistent, requiring extensive cleaning and transformation before analysis. Techniques like normalization, outlier detection, and handling missing values are applied to ensure data quality. Feature engineering, where new variables are derived from existing ones, enhances the predictive power of the machine learning models. For example, creating metrics like customer lifetime value (CLV) or recency-frequency-monetary (RFM) scores can provide deeper insights into customer behavior.

Once meaningful customer segments are identified, the project moves to model evaluation. Metrics such as silhouette score, Davies-Bouldin index, and inertia are used to assess the quality of clusters. For supervised models, accuracy, precision, recall, and F1-score provide a measure of the model’s effectiveness. Cross-validation and hyperparameter tuning are also conducted to ensure robustness and prevent overfitting.

Deployment plays a critical role in translating analytical insights into actionable business strategies. The segmented customer data is integrated into marketing automation tools, CRM systems, and personalized recommendation engines. Real-time deployment through APIs and cloud-based platforms ensures that businesses can dynamically adapt their strategies to changing customer needs. Visualization tools like Power BI, Tableau, or custom-built dashboards further enable stakeholders to monitor segment performance and make informed decisions.

The benefits of such an approach extend beyond marketing. Market segmentation powered by machine learning can drive innovation in product development by highlighting unmet customer needs and preferences. It can also improve customer retention by enabling businesses to create personalized experiences that resonate deeply with individual customers. Moreover, the insights derived from segmentation can inform strategic decisions such as market entry, pricing strategies, and partnership opportunities.

This project showcases the transformative potential of machine learning in market segmentation. By providing a systematic, data-driven approach, it enables businesses to not only understand their customers better but also anticipate their future needs. As businesses increasingly embrace digital transformation, the integration of advanced analytics and machine learning into marketing strategies is set to become a competitive necessity. Through this project, the foundation is laid for businesses to leverage the full power of their data and achieve sustained growth in an ever-evolving marketplace.

1.2 Motivation

In an increasingly competitive business environment, the ability to understand and segment customers has become vital for crafting targeted marketing strategies. Traditional market segmentation methods, often reliant on manual analysis and intuition, are no longer sufficient to handle the complexities and scale of modern consumer data.

Unlike traditional approaches to market segmentation, which often rely on predefined customer categories and manual analysis, machine learning offers a dynamic and data-driven method for uncovering hidden patterns and customer segments. However, the challenge lies in effectively processing and analyzing the vast amount of customer data, including demographics, behaviors, and purchasing patterns. This inspired the integration of machine learning algorithms such as K-Means clustering and hierarchical clustering to automate the segmentation process. By applying machine learning to market segmentation, businesses can significantly enhance their ability to target specific customer groups with personalized strategies, ultimately improving marketing efficiency and customer engagement.

The importance of accurate market segmentation cannot be overstated in today's data-driven world. As businesses strive to maintain a competitive edge, understanding the finer details of customer behavior becomes a key differentiator. Traditional methods, while effective in simpler markets, have proven to be inadequate in the face of modern data complexities. With the rise of big data, businesses are now inundated with a wealth of information from various sources—social media, customer interactions, online shopping behavior, surveys, and much more. This deluge of data creates both an opportunity and a challenge: the opportunity to extract valuable insights, but the challenge of processing and analyzing it effectively.

Machine learning, with its ability to handle and analyze large volumes of complex data, provides an ideal solution to this problem. The motivation behind integrating machine learning into market segmentation is the realization that traditional methods are limited in their ability to capture the evolving nature of consumer preferences. Whereas manual segmentation often relies on simplistic models or assumptions, machine learning can detect non-obvious patterns in customer behavior that are more reflective of the diverse ways consumers interact with brands. This enables businesses to go beyond basic demographic segmentation and explore psychographics, lifestyles, and purchasing patterns that are often the true drivers of customer loyalty.

The key advantage of machine learning in market segmentation is its ability to continuously learn and improve over time. Traditional methods often operate within a fixed framework, based on static rules that may become outdated as customer preferences evolve. On the other hand, machine learning algorithms can adapt and re-segment customers as new data is introduced, ensuring that segmentation strategies remain relevant in an ever-changing market landscape. This dynamic approach is particularly valuable in industries where customer preferences can shift rapidly, such as fashion, technology, and consumer goods. As a result, businesses using machine learning-based segmentation are better equipped to stay ahead of trends and anticipate the needs of their customers before they arise.

One of the most compelling aspects of using machine learning for market segmentation is its potential to uncover micro-segments that traditional methods might miss. These micro-segments, often consisting of smaller, highly specific groups of customers, can represent untapped market opportunities. By identifying these niche groups, businesses can design more targeted products, personalized marketing campaigns, and optimized pricing strategies that speak directly to the needs and desires of these customers. This level of precision not only increases customer satisfaction but also enhances customer retention, as customers are more likely to remain loyal to brands that understand their unique preferences.

Another key benefit of machine learning in market segmentation is its ability to automate the entire process, reducing the manual labor and human biases associated with traditional methods. By automating data processing and segmentation, businesses can accelerate decision-making and respond more quickly to shifts in the market. Automation also reduces the risk of errors that can arise from manual data entry or subjective judgment. Additionally, machine learning can process data in real time, ensuring that segmentation results are always up to date. This real-time capability is especially crucial in industries where market conditions fluctuate rapidly, such as retail, e-commerce, and digital advertising.

Beyond its direct impact on marketing and customer engagement, machine learning-driven segmentation also offers significant strategic advantages. By gaining deeper insights into customer behavior, businesses can identify areas for innovation and improvement. For example, a more granular understanding of customer segments may highlight product gaps or opportunities for new service offerings. Additionally, machine learning-based segmentation can inform broader business decisions, such as market expansion, pricing strategies, and customer service improvements. This holistic approach to segmentation empowers organizations to make data-driven decisions across all facets of their operations.

In the context of the current project, the integration of machine learning into market segmentation is not just a technological upgrade; it represents a fundamental shift in how businesses interact with and understand their customers. By automating the segmentation process and enabling more precise targeting, this project helps businesses unlock the full potential of their customer data. The ultimate goal is to create a more agile, responsive, and personalized approach to marketing that not only drives business growth but also fosters deeper and more meaningful relationships with customers. As businesses increasingly recognize the value of data-driven decision-making, the use of machine learning in market segmentation is poised to become a central component of modern marketing strategies, providing a competitive edge that is difficult to replicate.

The significance of this project lies in its potential to revolutionize how businesses approach market segmentation. With the integration of machine learning, organizations are no longer confined to broad, generic segments. Instead, they can uncover deeper insights that allow for hyper-targeted strategies, creating more meaningful customer interactions. Moreover, as the project involves deploying machine learning algorithms in real-time, it ensures that the insights generated are not just theoretical but actionable, seamlessly integrated into everyday decision-making processes. This integration of technology into the core business functions helps businesses stay ahead of competitors, fostering innovation and driving sustained growth in a crowded marketplace. Ultimately, this project sets the stage for organizations to leverage their data more effectively, transforming market segmentation from a static process to a dynamic, ongoing strategy for success.

1.3 Scope of the Project

This project focuses on the end-to-end development and deployment of a machine learning-based market segmentation model tailored to real-world business requirements. The primary objective is to analyze customer data and identify distinct customer groups based on shared characteristics and behaviors. The project begins with Exploratory Data Analysis (EDA) to uncover patterns, correlations, and trends that will guide the selection of features for segmentation, offering deeper insights into customer profiles.

The project begins with Exploratory Data Analysis (EDA) to uncover patterns, correlations, and trends that will guide the selection of features for segmentation. EDA is a crucial step in understanding the structure and distribution of customer data, which will ultimately drive the segmentation process. This phase involves identifying key variables, understanding customer demographics, and analyzing purchasing behavior, browsing patterns, and other relevant attributes. The insights gained through EDA will inform the choice of features and data preprocessing techniques, ensuring that the final model is based on the most pertinent information.

After the EDA phase, the project proceeds with the application of unsupervised machine learning algorithms. Specifically, the project employs algorithms such as K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM) to automatically segment customers. These unsupervised learning methods do not require labeled data, making them ideal for this project where the goal is to discover natural groupings in the customer data. K-Means clustering will be used for its efficiency and simplicity in partitioning customers into k distinct clusters based on similarity. Hierarchical clustering will allow for a dendrogram-based approach to visualize how customers are grouped at various levels of granularity. GMM will offer flexibility by assuming that customer segments follow a mixture of Gaussian distributions, providing a probabilistic view of segment membership.

To ensure the quality and accuracy of the segmentation, the project incorporates advanced data visualization techniques to interpret the results and assist with model selection. Visualizing the clusters helps in understanding the underlying patterns in the data, providing actionable insights. Additionally, model tuning will be employed to optimize the parameters of the clustering algorithms and ensure the best possible segmentation results. Hyperparameter optimization and techniques like grid search or random search will help fine-tune the models for improved performance.

The evaluation of the segmentation models is a critical component of the project. Various evaluation metrics, including silhouette score, Davies-Bouldin index, and inertia, will be used to assess the quality of the segmentation model. These metrics provide a measure of how well the customer data has been segmented, indicating the cohesiveness within clusters and the separation between different groups. Silhouette score, for example, measures how similar a customer is to their own group compared to other groups, while the Davies-Bouldin index quantifies the average similarity between clusters. Inertia is used to assess how tightly the clusters are formed. These metrics will help validate the effectiveness of the clustering models and ensure that the segmentation process is both accurate and meaningful.

The project leverages unsupervised machine learning algorithms such as K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM) to automatically segment customers. By employing advanced data visualization and model tuning techniques, the project seeks to ensure accurate and interpretable customer segmentation. Various evaluation metrics, including silhouette score, Davies-Bouldin index, and inertia, will be used to assess the quality of the segmentation model.

The scope is limited to customer segmentation and real-time deployment through cloud-based services, providing businesses with a scalable and interactive tool to enhance decision-making processes. The findings from the project will help organizations better understand their customer base, allowing them to craft targeted marketing strategies and improve overall engagement. However, businesses may need to extend and customize the implementation based on their specific data and operational needs.

The findings from this project will help organizations better understand their customer base, allowing them to craft targeted marketing strategies, improve customer experience, and ultimately increase customer retention. With the model in place, businesses will be able to tailor product recommendations, promotions, and messaging to specific customer segments, improving overall engagement and ROI on marketing efforts. However, the project is designed to be flexible and extendable. While it provides a comprehensive framework for market segmentation, businesses may need to customize the implementation based on their specific data and operational needs. For example, integrating additional data sources or applying specific business rules could enhance the segmentation process to better align with industry-specific requirements.

In conclusion, the scope of this project outlines a detailed and practical approach to market segmentation using machine learning, focusing on the development, evaluation, and deployment of a scalable solution. Through this project, businesses will gain actionable insights into customer behavior, allowing them to make informed, data-driven decisions that enhance marketing efforts, improve customer targeting, and ultimately drive business growth.

2. PROJECT DESCRIPTION AND GOALS

This project focuses on the development and deployment of a machine learning-based market segmentation model aimed at identifying distinct customer groups to enhance targeted marketing strategies. In today’s data-driven business environment, understanding customer behavior is essential for crafting personalized marketing campaigns. This project addresses that need by automating the customer segmentation process using unsupervised machine learning algorithms. The core objective is to uncover meaningful insights from customer data, such as demographics and behavioral patterns, which can drive more precise and effective marketing efforts.

The goals of this project are multifaceted, aiming to create a comprehensive solution for businesses looking to improve their market segmentation practices. The following outlines the key objectives and steps involved:

• Develop a machine learning-driven segmentation model

The first goal of the project is to apply unsupervised learning techniques to segment customers based on shared characteristics and behaviors. Traditional market segmentation methods often rely on broad assumptions or simple demographic factors. In contrast, unsupervised machine learning models like K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM) allow for the discovery of natural groupings in the data without predefined labels. These models automatically identify hidden patterns within customer data, providing businesses with actionable insights about their customer base. By clustering customers into distinct groups, businesses can understand the varying needs and preferences within their target market and create tailored marketing strategies that resonate with each segment.

• Automate data analysis for pattern recognition

To ensure the segmentation model is based on accurate, meaningful data, the project includes an extensive phase of Exploratory Data Analysis (EDA). During this phase, various data cleaning, preprocessing, and analysis techniques will be applied to customer data, such as demographics, purchasing behavior, and browsing patterns. By leveraging advanced data visualization tools, such as heatmaps, pair plots, and cluster visualizations, this step uncovers significant trends and relationships within the data. These insights will guide the feature selection process, helping to ensure that only the most relevant customer attributes are used for segmentation. EDA will play a crucial role in ensuring that the final model is both interpretable and actionable for businesses.

• Optimize segmentation performance

A key goal of the project is to ensure that the segmentation model is accurate and reliable. This will be achieved by tuning the hyperparameters of the machine learning models and testing various clustering algorithms to determine the best fit for the customer data. Different algorithms may have varying strengths depending on the nature of the data, so testing multiple approaches will ensure the best possible outcome. The project will utilize a set of evaluation metrics to assess the quality of the segmentation results. Metrics such as silhouette score, Davies-Bouldin index, and inertia will be used to gauge how well the customer segments are formed. Silhouette score measures the consistency within clusters, Davies-Bouldin index quantifies the separation between clusters, and inertia evaluates how compact the clusters are. Fine-tuning the model using these metrics will help ensure that the segmentation results are both meaningful and actionable.

• Deploy segmentation model into business environments

Once the model has been developed and tested, the next goal is to deploy it in a real-time, scalable environment. The segmentation model will be integrated into a web application or API, allowing businesses to make use of the model directly within their operational environments. This deployment ensures that businesses can continuously update their customer segments as new data is collected and analyzed. Real-time deployment is crucial for ensuring that the model provides timely insights that can be acted upon immediately. The cloud-based infrastructure will ensure scalability, enabling businesses of all sizes to integrate the model seamlessly into their existing systems.

• Enhance marketing strategies and decision-making

The ultimate goal of the project is to enable businesses to develop more personalized marketing approaches and improve customer engagement strategies based on the insights derived from the segmentation model. By automating the segmentation process, businesses can rapidly identify the distinct needs and preferences of different customer groups, allowing for more targeted and effective marketing campaigns. Personalized recommendations, tailored promotions, and segmented product offerings will help increase customer satisfaction and loyalty. Additionally, businesses will be able to allocate marketing resources more efficiently, focusing efforts on high-value customer segments with the greatest potential for engagement. The integration of machine learning into marketing strategies also leads to more data-driven decision-making, improving operational efficiency and fostering long-term customer relationships.

2.1 Literature Review

The Role of Digital Marketing in Higher Education

Sotomayor Vidal, Mini-Cuadros, and Quiroz-Flores (n.d.) explored the impact of digital marketing strategies on student recruitment within Peru’s private higher education sector. Their findings highlight the significance of digital platforms in influencing student decisions, suggesting that institutions leverage online engagement for effective market positioning. Similarly, Canterbury (2000) underscored the evolving challenge of marketing in higher education, emphasizing the necessity for institutions to adopt strategic approaches tailored to diverse student preferences.

Market Segmentation in Higher Education

Market segmentation plays a pivotal role in understanding diverse student demographics and preferences. Hemsley-Brown (2020) provided a comprehensive analysis of segmentation within higher education, noting its importance in crafting tailored strategies that resonate with distinct student groups. In the Colombian context, Lozano, Cruz Pulido, and Garcia Rodriguez (2021) revealed how segmentation highlights existing social inequalities, underscoring the need for inclusive educational marketing.

Data-Driven Approaches to Market Segmentation

Data mining has become instrumental in identifying and understanding market segments. Davari, Noursalehi, and Keramati (2019) demonstrated its effectiveness in professional education, using case studies to show how segmentation improves marketing outcomes. Advanced clustering techniques.

Multidimensional Applications of Market Segmentation

The application of segmentation transcends industries. Casas-Rosal, Segura, and Maroto (2023) employed multicriteria approaches to segment food markets, highlighting consumer preferences as a key determinant. Similarly, Kim and Irakoze (2023) applied cluster analysis to assess the price premium for green-certified housing, demonstrating the versatility of segmentation techniques in sustainability-focused markets. Furthermore, Zhu and Liu (2023) proposed strategies for personalized preference learning, showcasing segmentation’s potential in addressing sparse consumer data scenarios.

Technological Integration and Segmentation Tools

Advancements in technology have enhanced segmentation capabilities. Singhal and Jena (n.d.) evaluated the WEKA tool for preprocessing, classification, and clustering, emphasizing its utility in educational and commercial applications. Huang (2023) leveraged these tools to analyze market segmentation in hospitality, focusing on slow food experiences. These studies illustrate the growing reliance on computational tools for precise market analysis.

Insights into Educational Equity and Globalization

Market segmentation also uncovers disparities and opportunities in higher education. Rizvi (2023) discussed the internationalization of education and the advantages for diaspora communities, presenting segmentation as a means to address global educational equity. In contrast, Lambert (2023) explored labor market segmentation in the context of the Great Resignation, linking it to broader socioeconomic trends.

Educational Market Segmentation and Social Inequality

Lozano, Cruz Pulido, and Garcia Rodriguez (2021) highlighted the role of market segmentation in identifying social inequalities within the higher education sector in Colombia. Their research revealed that market segmentation not only helps to categorize students based on preferences.

Market Segmentation for Professional Education

In their study, Davari, Noursalehi, and Keramati (2019) applied data mining techniques to segment the professional education market. They found that using sophisticated data analysis techniques significantly improved the precision of segment identification, enabling educational institutions to target specific professional training needs more effectively.

Clustering Techniques in Telecommunications and Education

Vieri, Munandar, and Srisulistiowati (2023) explored the use of exclusive clustering techniques for customer segmentation in telecommunications. Their work highlights the importance of tailored marketing strategies in diverse industries. The application of these clustering methods can be extended to the higher education sector, where universities can leverage segmentation to address different student needs, particularly in terms of digital engagement and curriculum offerings.

The Impact of Consumer Preferences on Market Segmentation

Casas-Rosal, Segura, and Maroto (2023) employed outranking multicriteria approaches to segment food markets based on consumer preferences. Their findings suggest that consumer preferences are a central factor in market segmentation strategies, and such methods can be adapted for educational markets. By understanding student preferences—such as course offerings, campus environment, and extracurricular activities—higher education institutions can create more appealing programs and services tailored to distinct student groups.

Segmentation in Sustainability and Green Housing

Kim and Irakoze (2023) used cluster analysis to identify market segments for green-certified housing, highlighting the importance of segmentation in sustainability-focused industries. Their findings indicate that consumers who value sustainability are willing to pay a premium for eco-friendly housing options. In a similar vein, universities can apply segmentation strategies to appeal to environmentally conscious students by offering green campus initiatives, eco-friendly dorms, and sustainable study programs, effectively capturing the growing market for sustainability in education.

Data Preprocessing and Classification in Segmentation

Singhal and Jena (n.d.) examined the effectiveness of the WEKA tool for data preprocessing, classification, and clustering in segmentation. Their study demonstrates how advanced data processing tools can streamline the segmentation process, making it more accurate and efficient. The use of such tools is especially relevant in educational marketing, where the large volume of student data can be leveraged to develop more personalized marketing strategies that resonate with different demographic groups.

Consumer Segmentation in E-Commerce and Educational Markets

Rajput and Singh (2023) explored customer segmentation within e-commerce, utilizing K-means clustering algorithms to understand consumer behavior. The approach used in their study can be applied to educational settings, where student behavior—such as program choices, online learning preferences, and purchasing decisions related to educational resources—can be segmented for more effective targeting of educational products and services. This offers a roadmap for universities to better understand the purchasing behaviors of students and adapt marketing strategies accordingly.

Globalization and Market Segmentation in Higher Education

Rizvi (2023) explored the internationalization of higher education, emphasizing how market segmentation can support educational institutions in leveraging diaspora communities. By targeting international students, universities can enhance global recruitment strategies, making use of segmentation to address regional and cultural differences. This aligns with the growing trend of higher education institutions expanding their global presence and tapping into new international markets.

Personalized Learning and Segmentation in Education

Zhu and Liu (2023) examined personalized preference learning under sparse consumer data, a technique that can be highly beneficial in educational settings. The approach focuses on adapting to the individual needs of students, even with limited data, to provide personalized experiences. Educational institutions can use segmentation to understand each student’s preferences for learning styles, course delivery formats (online, hybrid, in-person), and academic interests, enabling them to offer more personalized and engaging learning environments.

Segmentation in Hospitality and Leisure Markets

Huang (2023) applied market segmentation analysis to the slow food experience at wineries, revealing consumer preferences that drive segmentation in the hospitality and leisure industries. This study demonstrates that segmentation strategies can extend beyond education into leisure and extracurricular activities offered by universities.

Table 2.1: Summary of Literature on Market Segmentation

Author(s) and Year	Focus Area	Methodology	Key Findings
Sotomayor Vidal et al. (2022)	Digital marketing in higher education student recruitment	Case study	Demonstrated the impact of digital marketing strategies on attracting students in the private sector.
Hemsley-Brown (2020)	Higher education market segmentation	Conceptual overview	Highlighted the need for nuanced segmentation in higher education to address diverse student needs.
Davari et al. (2019)	Professional education market segmentation	Data mining	Applied data mining to identify distinct segments and their preferences in professional education.
Lozano et al. (2021)	Social inequalities in higher education segmentation	Empirical study	Revealed how market segmentation reflects social inequalities in Colombia’s education system.
Chen & Hsiao (2009)	Student behavior in school selection	Market segmentation theory	Explored factors influencing students' decisions when selecting schools and departments.
Casas-Rosal et al. (2023)	Food market segmentation	Multicriteria decision-making approach	Developed a segmentation model based on consumer preferences in the food market.
Canterbury (2000)	Challenges in higher education marketing	Conceptual analysis	Identified challenges and strategies for effective higher education marketing.
Zhao et al. (2023)	Corruption and market segmentation	Empirical analysis	Analyzed the impact of corruption on market segmentation and environmental outcomes in China.
Lambert (2023)	Labor market	Case study	Studied labor market

	segmentation in the context of the Great Resignation		dynamics during the Great Resignation, identifying key segmented groups.
Rizvi (2023)	Internationalization of higher education	Conceptual discussion	Discussed the role of diaspora in higher education internationalization strategies.
Aktan & Kaplan (2023)	Mindfulness in higher education	Theoretical framework	Proposed redesigning higher education with mindfulness as a focus.
Rajput & Singh (2023)	E-commerce customer segmentation	K-means clustering	Demonstrated effective segmentation using clustering algorithms for online retail.
Huang (2023)	Slow food market segmentation	Cluster analysis	Identified market segments for slow food experiences in hospitality.
Vieri et al. (2023)	Telecommunications customer segmentation	Exclusive clustering	Highlighted distinct customer groups for targeted marketing in telecom.
Kumar (2023)	Shopping mall user segmentation	K-means clustering	Explored customer segmentation for better service personalization in malls.
Kim & Irakoze (2023)	Green housing market segmentation	Cluster analysis	Identified segments willing to pay a premium for green-certified housing.
Zhu & Liu (2023)	Personalized preference learning	Segmentation strategy	Proposed a segmentation approach for sparse consumer data in personalized services.
Singhal & Jena (n.d.)	Data preprocessing for clustering	WEKA tool	Reviewed tools for preprocessing and clustering in market segmentation.
Singhal et al. (2020)	Machine learning for product sustainability	Literature review	Summarized the role of ML in promoting sustainable product development.

2.2 Research Gap

Despite advancements in market segmentation and analysis using machine learning, several critical research gaps persist. One significant gap is the inadequacy of models to handle diverse datasets effectively. While existing algorithms show promise, they often fail to generalize across varying data distributions, which can lead to skewed results and ineffective segmentation. Another area needing attention is the integration of real-time data streams. Current methodologies primarily focus on static data analysis, neglecting the dynamic nature of consumer behavior that requires timely adjustments in segmentation strategies.

Moreover, the application of machine learning techniques to niche markets is underexplored. Much of the existing literature concentrates on mainstream sectors, leaving gaps in understanding how to effectively apply these methods in specialized industries. Addressing this gap could unlock valuable insights for businesses operating in less-studied areas.

One major gap in the existing literature is the inability of models to generalize across diverse datasets. While machine learning algorithms like K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM) have demonstrated effectiveness in segmenting customer groups, they often struggle when applied to datasets with diverse characteristics. Many models fail to generalize across varying data distributions, leading to skewed results and inaccurate customer segments. This issue arises due to the inherent biases in the data or the difficulty of the models in adapting to different customer behaviors across geographies, industries, or time periods. For businesses dealing with large, heterogeneous customer datasets, it is critical to develop more robust algorithms that can effectively handle diverse, high-dimensional data without overfitting or underfitting.

Another significant gap is the integration of real-time data streams. Most current machine learning approaches in market segmentation rely on static datasets, meaning that the data is collected at a specific point in time and analyzed in batches. However, the reality of modern consumer behavior is dynamic and rapidly changing. Real-time data streams are essential for adapting to fluctuations in customer preferences, market conditions, and external factors such as seasonality or current events. As businesses move towards more agile decision-making processes, there is a growing need to integrate machine learning models with real-time data pipelines. This would allow for dynamic segmentation strategies that can quickly adjust as new data becomes available, thereby improving the relevance and timeliness of marketing strategies.

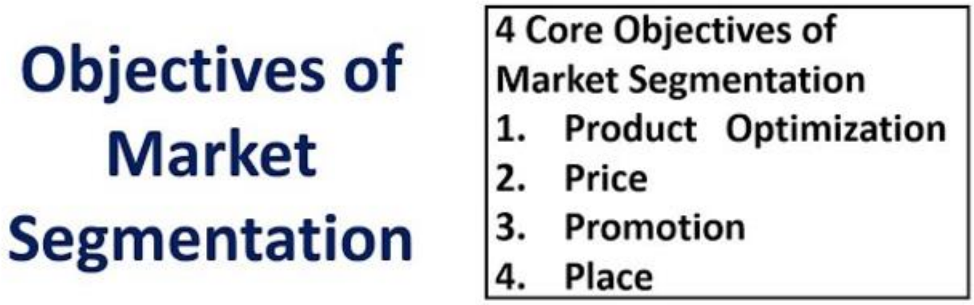
Furthermore, the application of machine learning to niche markets remains underexplored. The majority of existing research focuses on mainstream markets, such as consumer goods or general services, leaving a gap in the understanding of how to apply machine learning techniques to more specialized industries. Niche markets, such as luxury goods, medical services, or high-tech products, may require customized segmentation strategies that take into account unique consumer behaviors and market dynamics. Research in this area would expand the scope of market segmentation and offer more tailored approaches to various industries. For instance, in luxury retail, customer segmentation may need to focus on factors like exclusivity, brand affinity, and personalized shopping experiences, which are distinct from the typical demographic or behavioral-based segmentation in mass markets.

The ethical implications of automated segmentation techniques are another area that has not received sufficient attention. As the collection and analysis of consumer data become increasingly prevalent, concerns around data privacy and consumer rights have grown. Automated segmentation models often require large amounts of personal and behavioral data, which raises issues regarding consent, privacy, and data security. Research is needed to establish ethical guidelines that balance the effectiveness of market segmentation with the protection of consumer rights. This includes ensuring that segmentation models are fair, transparent, and compliant with data privacy regulations such as the General Data Protection Regulation (GDPR) in Europe or the California Consumer Privacy Act (CCPA) in the United States. Additionally, companies need to consider how these models impact consumer trust, particularly when segmentations are used for personalized advertising, price discrimination, or other targeted strategies that could be perceived as manipulative.

In conclusion, addressing these research gaps—diverse dataset handling, real-time data integration, interpretability, niche market applications, ethical concerns, and long-term evaluation—is crucial for advancing machine learning-based market segmentation. Overcoming these challenges will enable businesses to implement more effective, dynamic, and responsible segmentation strategies, thereby improving their marketing effectiveness and customer engagement in an increasingly data-driven world.

2.3 Objectives

Fig. 2.3. Objectives of Market Segmentation



Objectives of Market Segmentation

4 Core Objectives of Market Segmentation

1. Product Optimization
2. Price
3. Promotion
4. Place

The primary objectives of this project are focused on the development, enhancement, and application of a machine learning-based market segmentation model that can be used in real-time business environments.

These objectives are critical for delivering valuable, actionable insights that will help businesses tailor their marketing strategies, improve customer engagement, and achieve long-term success. The objectives of this project are outlined as follows:

1. **Develop a Machine Learning-Based Market Segmentation Model:**

   - The first objective of this project is to design and build a machine learning-driven segmentation model that can effectively analyze diverse customer data, including demographic, behavioral, and transactional information. The goal is to apply unsupervised machine learning algorithms such as K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM) to automatically detect and define distinct customer segments based on shared characteristics and behaviors.

   - A key performance metric for the model will be the ability to classify at least 80% of data points accurately into relevant segments. This accuracy threshold ensures that the segmentation model provides meaningful insights for businesses to base their marketing and operational strategies on.

   - The development process will involve testing multiple algorithms, selecting the most effective approach, and fine-tuning the model using evaluation metrics such as silhouette scores, Davies-Bouldin index, and inertia to ensure high-quality and precise segmentation.

2. **Implement Real-Time Data Processing:**

   - The second objective is to integrate real-time data streams into the segmentation model, allowing businesses to keep up with the fast-paced nature of modern consumer behavior. Real-time data processing will enable the model to adapt to market fluctuations, such as changes in customer preferences, trends, or external factors like economic shifts.

   - The model should be capable of dynamically updating customer segments within a latency of no more than 5 minutes, ensuring that businesses can make timely decisions based on the most current data available. This real-time responsiveness will be crucial for industries where customer behavior and preferences change frequently, such as e-commerce, finance, or travel.

   - To achieve this, cloud-based data processing technologies like Apache Kafka, AWS Kinesis, and Apache Flink will be used to manage large-scale, real-time data streams and integrate them seamlessly into the machine learning pipeline.

3. **Enhance Model Interpretability:**

    - Another key objective is to improve the interpretability of the segmentation model to make the decision-making process transparent and understandable for non-technical stakeholders. Many machine learning models, particularly unsupervised clustering algorithms, are often perceived as "black boxes" where the rationale behind the segmentation results is not easily discernible.

    - To address this, interpretability techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) will be applied to provide insights into how the model makes its decisions. These methods will help elucidate the features that contribute to segment formation, making the process more transparent.

    - Additionally, user-friendly visualizations, including feature importance charts, decision trees, and clustering heatmaps, will be developed to help marketing teams, product managers, and other business stakeholders understand and trust the segmentation results. This will empower businesses to act on the insights generated by the model with greater confidence and precision.

4. **Explore Applications in Niche Markets:**

    - The next objective is to apply the machine learning segmentation model to niche markets to explore its versatility and effectiveness in specialized industries. While most segmentation models focus on mainstream markets, there is significant potential in applying machine learning techniques to niche areas such as luxury goods, medical services, or specialized consumer products.

    - Case studies will be conducted in these niche markets to apply the model and identify unique challenges, such as varying consumer behaviors, market size limitations, or the need for highly tailored segmentation criteria. By documenting the findings from these case studies, this objective will help highlight the flexibility of machine learning-based segmentation across different business domains.

    - The insights gained from these applications can help businesses in niche markets craft more targeted and personalized marketing strategies, optimize product offerings, and gain a competitive edge in their industry.

5. **Evaluate Long-Term Segmentation Impact:**

    - Finally, a critical objective is to assess the long-term impact of the segmentation model on business performance. Traditional segmentation models often evaluate their effectiveness based on short-term metrics, such as cluster cohesion and separation, but this project aims to take a broader view of segmentation effectiveness.

    - The evaluation will focus on comprehensive metrics that connect segmentation strategies to tangible business outcomes. Key performance indicators (KPIs) such as customer lifetime value (CLV), customer retention rates, sales growth, and return on marketing investment (ROMI) will be used to measure the long-term success of the segmentation model.

    - The goal is to track how well the model's insights lead to sustained improvements in customer engagement, brand loyalty, and overall profitability. This will allow businesses to continuously refine and optimize their segmentation strategies based on real-world outcomes, ensuring the model provides enduring value.

By achieving these objectives, this project aims to provide businesses with a comprehensive, data-driven market segmentation solution that enhances their ability to target specific customer groups effectively. The combination of machine learning, real-time data processing, and interpretable models will ensure that the segmentation model can adapt to changing market conditions, improve marketing efficiency, and ultimately drive revenue growth. Additionally, the application in niche markets and long-term impact evaluation will ensure that the segmentation strategy is robust, scalable, and applicable across a wide range of industries and business contexts.

2.4 Problem Statement

The project’s ultimate goal is to provide businesses with an efficient way to group customers who exhibit similar purchasing patterns, enabling better understanding of their needs, preferences, and buying habits. This information allows businesses to tailor their marketing strategies, promotional offers, and product development to the specific interests of each customer segment. For example, a segment that prioritizes convenience and health-related attributes in products could be targeted with specialized marketing efforts that focus on low-calorie, quick-to-prepare, and easy-to-use products.

Using K-Means clustering and Logistic Regression, the project seeks to apply demographic segmentation to predict customer characteristics, such as their age group, based on behavioral patterns. K-Means clustering will help group customers with similar spending behaviors and preferences, while Logistic Regression will further predict the likelihood of customers falling into specific age categories based on their preferences and behaviors. By combining these two techniques, the project creates a comprehensive view of customer segments, which can be targeted for more personalized and effective marketing campaigns.

The project also seeks to offer insights into how business operations can be optimized through a deeper understanding of customer behavior. For example, businesses may discover that certain segments are more loyal and likely to make repeat purchases, allowing them to focus on retaining these customers with loyalty programs or personalized offers. On the other hand, the project may also highlight segments that are less engaged, helping businesses identify areas where they need to increase engagement through targeted marketing or product changes.

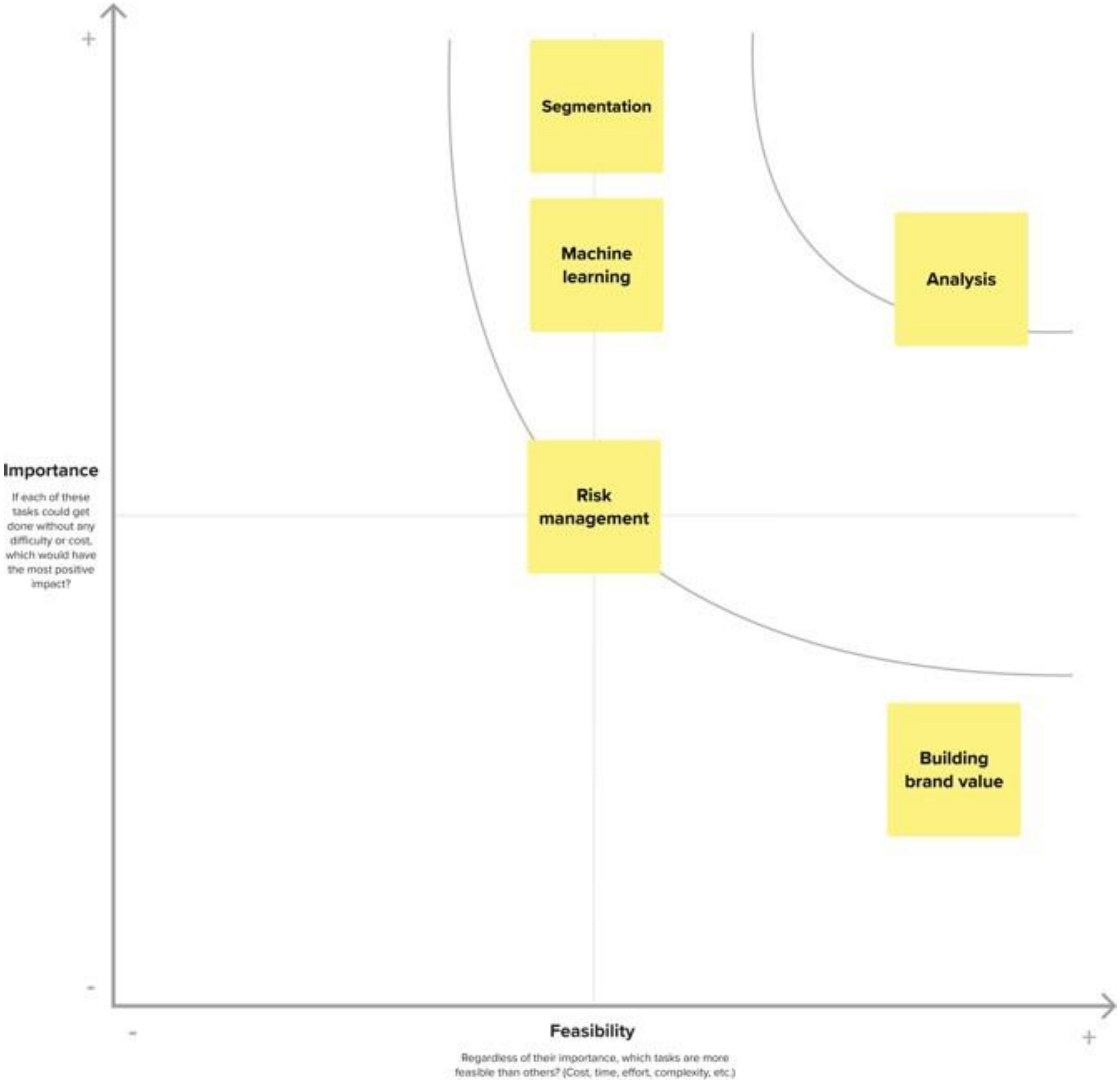
By combining these insights into customer behavior and spending patterns, the project helps businesses enhance their decision-making process, ultimately increasing customer satisfaction and retention. With better-targeted marketing campaigns, businesses can improve engagement, foster loyalty, and drive growth. This project thus provides wholesale businesses with a scalable and data-driven approach to enhance their marketing strategies and optimize customer interactions, ensuring that they remain competitive and responsive to the evolving demands of their customers.

Ultimately, by understanding their customers at a granular level, businesses can offer products and services that directly meet the needs and expectations of their target segments. This will not only lead to increased sales and customer loyalty but also help businesses become more agile and better prepared to adapt to future market shifts.

Beyond segmentation, the project aims to enhance the interpretability of machine learning models. By providing transparency into the decision-making processes behind the segmentation, the project makes it easier for businesses to trust and act on the model’s insights. Clear visualizations and explanations of the clustering process will ensure that business stakeholders can easily understand and apply the results to their operations.

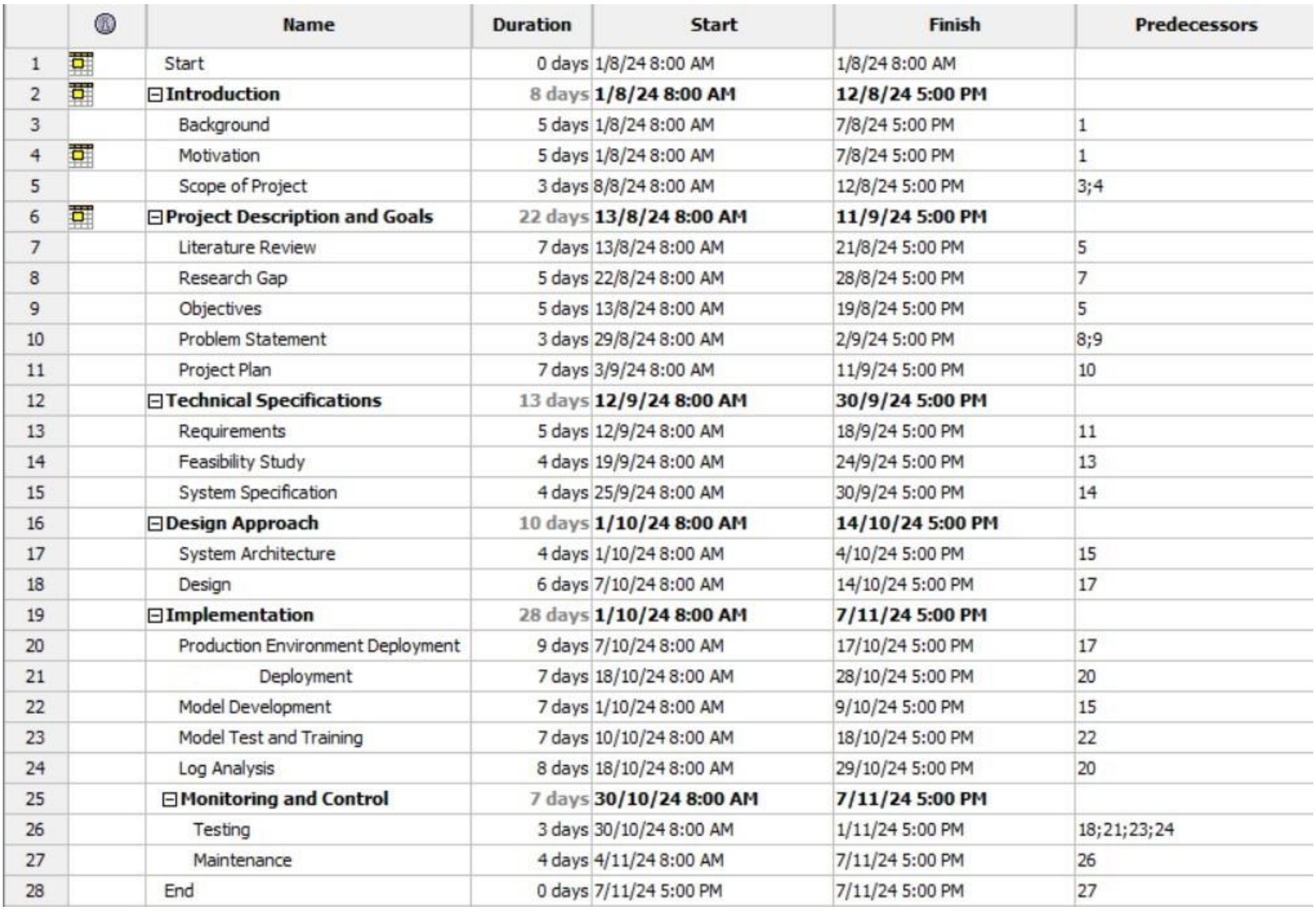
The insights derived from the model can help wholesale businesses identify underrepresented product categories, pinpoint areas of growth, and better understand customer preferences. By aligning their marketing efforts with the characteristics of each segment, businesses can develop personalized offers and promotions that resonate with their target audience, ultimately driving higher engagement and sales.

Fig. 2.4. Importance Vs Feasibility


+
Segmentation
Machine learning
Analysis
Importance
If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?
Risk management
Building brand value
-
-
Feasibility
Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)
+

2.5 Project Plan

Table 2.5. Project Plan of Market Segmentation and Analysis



	ⓘ	Name	Duration	Start	Finish	Predecessors
1	▦	Start	0 days	1/8/24 8:00 AM	1/8/24 8:00 AM	
2	▦	⊟**Introduction**	8 days	**1/8/24 8:00 AM**	**12/8/24 5:00 PM**	
3		Background	5 days	1/8/24 8:00 AM	7/8/24 5:00 PM	1
4	▦	Motivation	5 days	1/8/24 8:00 AM	7/8/24 5:00 PM	1
5		Scope of Project	3 days	8/8/24 8:00 AM	12/8/24 5:00 PM	3;4
6	▦	⊟**Project Description and Goals**	22 days	**13/8/24 8:00 AM**	**11/9/24 5:00 PM**	
7		Literature Review	7 days	13/8/24 8:00 AM	21/8/24 5:00 PM	5
8		Research Gap	5 days	22/8/24 8:00 AM	28/8/24 5:00 PM	7
9		Objectives	5 days	13/8/24 8:00 AM	19/8/24 5:00 PM	5
10		Problem Statement	3 days	29/8/24 8:00 AM	2/9/24 5:00 PM	8;9
11		Project Plan	7 days	3/9/24 8:00 AM	11/9/24 5:00 PM	10
12		⊟**Technical Specifications**	13 days	**12/9/24 8:00 AM**	**30/9/24 5:00 PM**	
13		Requirements	5 days	12/9/24 8:00 AM	18/9/24 5:00 PM	11
14		Feasibility Study	4 days	19/9/24 8:00 AM	24/9/24 5:00 PM	13
15		System Specification	4 days	25/9/24 8:00 AM	30/9/24 5:00 PM	14
16		⊟**Design Approach**	10 days	**1/10/24 8:00 AM**	**14/10/24 5:00 PM**	
17		System Architecture	4 days	1/10/24 8:00 AM	4/10/24 5:00 PM	15
18		Design	6 days	7/10/24 8:00 AM	14/10/24 5:00 PM	17
19		⊟**Implementation**	28 days	**1/10/24 8:00 AM**	**7/11/24 5:00 PM**	
20		Production Environment Deployment	9 days	7/10/24 8:00 AM	17/10/24 5:00 PM	17
21		Deployment	7 days	18/10/24 8:00 AM	28/10/24 5:00 PM	20
22		Model Development	7 days	1/10/24 8:00 AM	9/10/24 5:00 PM	15
23		Model Test and Training	7 days	10/10/24 8:00 AM	18/10/24 5:00 PM	22
24		Log Analysis	8 days	18/10/24 8:00 AM	29/10/24 5:00 PM	20
25		⊟**Monitoring and Control**	7 days	**30/10/24 8:00 AM**	**7/11/24 5:00 PM**	
26		Testing	3 days	30/10/24 8:00 AM	1/11/24 5:00 PM	18;21;23;24
27		Maintenance	4 days	4/11/24 8:00 AM	7/11/24 5:00 PM	26
28		End	0 days	7/11/24 5:00 PM	7/11/24 5:00 PM	27

The **Project Plan** is an essential blueprint for managing the lifecycle of the market segmentation analysis system. It defines the key tasks, their sequence, durations, and interdependencies, ensuring that the project is delivered on time and within budget. By combining detailed task definitions with visual scheduling through the Gantt chart, this plan offers a holistic view of the project's progress and milestones.

Key Components of the Project Plan

1. **Introduction Phase:**
   The introduction phase sets the stage for the project. It involves defining the background, motivation, and scope of the project. This phase takes eight days and establishes a strong foundation, ensuring that all team members and stakeholders are aligned on the project's vision, objectives, and anticipated outcomes.

2. **Project Description and Goals:**
   This phase is pivotal for establishing the strategic direction of the project. Activities include:
   - Conducting a **literature review** to identify existing methodologies and advancements in market segmentation.
   - Highlighting **research gaps** to identify areas where the project can contribute uniquely.
   - Defining **objectives** to clarify the scope and measurable outcomes of the project.
   - Framing the **problem statement**, which serves as a guiding compass throughout the project lifecycle.
     This phase spans 25 days, ensuring comprehensive exploration and planning.

3. **Technical Specifications:**
   This phase ensures that the technical requirements for the system are clearly documented and assessed for feasibility. It includes:
   - Gathering the hardware and software specifications.
   - Analyzing the **feasibility** from technical, economic, and social perspectives.
   - Defining system requirements to ensure the design aligns with project goals.
     With 11 days allocated, this phase ensures that the design and implementation phases proceed with a strong foundation.

4. **Design Approach:**
   In this phase, the project team develops the system architecture and diagrams to provide a blueprint for implementation. Activities include:
   - Creating the **system architecture** to represent the interaction between components.
   - Designing **diagrams** such as data flow, activity, and use case diagrams, offering a visual representation of the system's operation.
     This phase ensures that the system is logically structured and capable of fulfilling the intended objectives.

5. **Implementation Phase:**
   The most resource-intensive and critical phase, this involves the actual development of the system. Activities include:
   - Preparing the **production environment**, ensuring all tools, software, and dependencies are ready for deployment.
   - Developing and training machine learning models, including **k-means clustering** and other algorithms.

- Conducting **testing** and log analysis to ensure accuracy and robustness.
  This phase spans 56 days and focuses on building a functional, deployable system.

6. **Monitoring and Control:**
   Post-implementation, this phase ensures the system operates effectively. Key activities include:
   - Conducting **testing** to identify and address any bugs or performance issues.
   - Implementing **maintenance** procedures to guarantee long-term reliability and adaptability.
   - Providing user feedback loops to refine the system over time.
     This phase ensures sustained system performance and customer satisfaction.

Detailed Milestones and Deliverables

The Gantt chart breaks down the project into phases, assigning durations, dependencies, and milestones. Highlights include:

- **Parallel Task Management:** Several tasks, such as literature review and gathering system requirements, are conducted simultaneously to optimize time and resource usage.

- **Critical Path Identification:** Dependencies among tasks ensure timely completion of critical activities, minimizing risks of project delays.

- **Key Deliverables:** Deliverables include documentation of objectives, diagrams of the design, deployment of machine learning models, and a fully operational segmentation system.

Benefits of the Project Plan

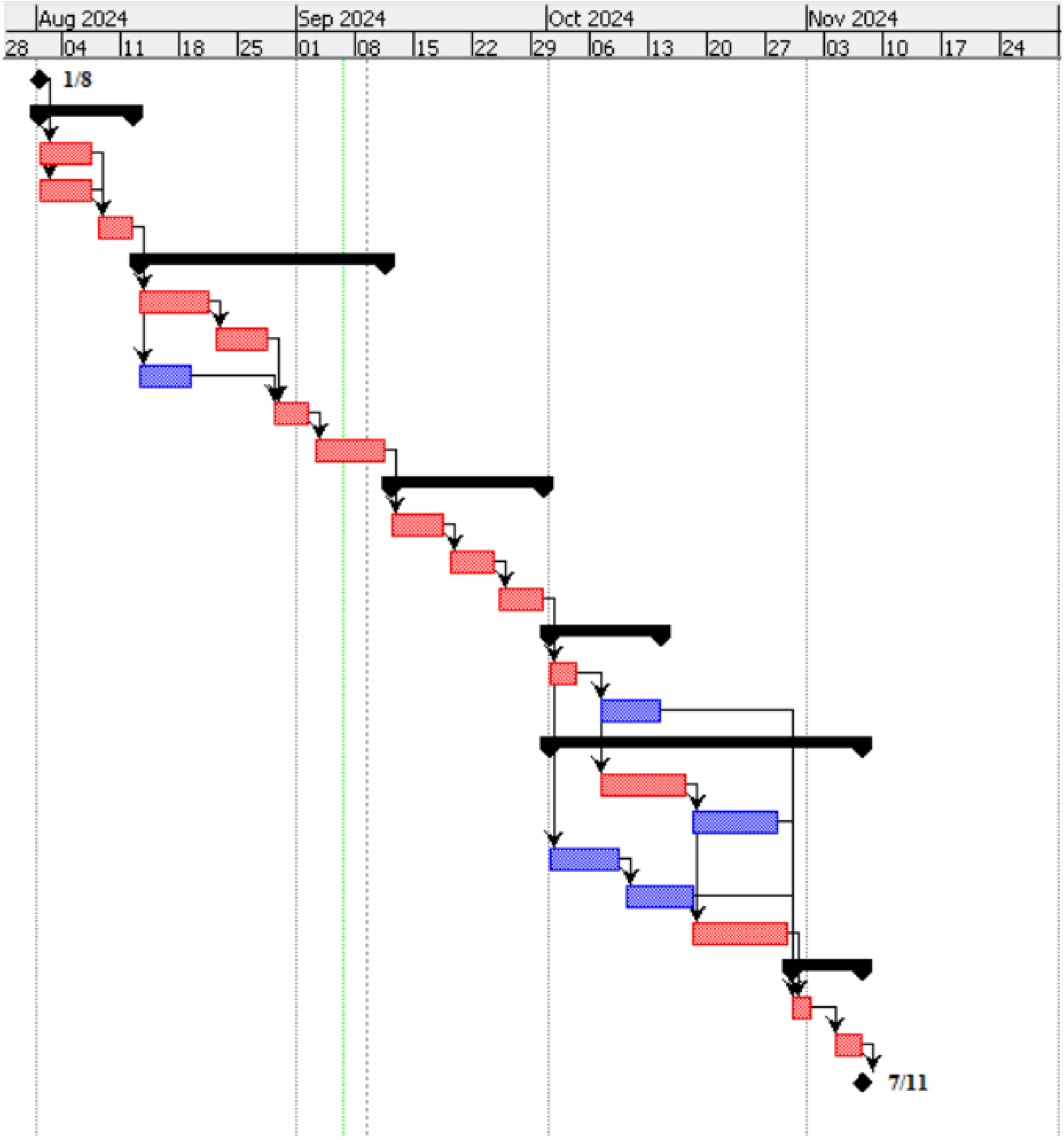
1. **Resource Allocation:**
   By defining clear tasks and durations, the plan ensures that resources—both human and technical—are optimally allocated across project phases.

2. **Progress Monitoring:**
   The sequential task flow in the Gantt chart allows for real-time tracking of progress and early identification of bottlenecks.

3. **Stakeholder Alignment:**
   With well-defined milestones and deliverables, stakeholders can clearly understand project timelines and anticipate outcomes.

4. **Flexibility and Scalability:**
   The plan accommodates unforeseen challenges by allowing adjustments to non-critical tasks without affecting overall timelines.

Fig. 2.5. Gantt chart



The Gantt chart provides a visual timeline for managing the project's activities, ensuring clarity in task scheduling, dependencies, and progress tracking. It is divided into distinct phases, each contributing to the successful completion of the project. Below is an extended overview of the Gantt chart, with a focus on milestones, task sequences, and interdependencies.

Detailed Gantt Chart Breakdown

1. **Project Initiation (8 Days)**
   This phase marks the start of the project and sets the foundation for subsequent tasks. Activities include:

   - **Kickoff Meetings:** Aligning stakeholders and team members with the project's goals.
   - **Background Research:** Conducting preliminary research to establish the context and relevance of the market segmentation project.
   - **Deliverable:** A well-documented project charter, including the scope, motivation, and high-level objectives.

2. **Project Description and Goal Setting (25 Days)**

   This extended phase focuses on refining the project's vision and objectives:

   - **Literature Review (8 Days):** Exploring existing market segmentation models and identifying gaps.
   - **Research Gap Analysis (5 Days):** Highlighting areas where the proposed project offers unique contributions.
   - **Objective Definition (7 Days):** Setting clear, measurable, and attainable project goals.
   - **Problem Statement Drafting (5 Days):** Formulating a precise and actionable problem statement.

     Milestone: Finalized project description document, reviewed and approved by stakeholders.

3. **Technical Specifications (11 Days)**

   This phase outlines the technological requirements for the system:

   - **Hardware and Software Analysis (5 Days):** Identifying suitable tools and resources.
   - **Feasibility Study (6 Days):** Assessing the technical, economic, and social feasibility of the project.
     Deliverable: A comprehensive feasibility report that informs the system design phase.

4. **Design and Development (60 Days)**

This is the core phase of the project, involving system design, development, and integration:

- **System Architecture Design (15 Days):** Creating detailed diagrams to define system workflows.
- **Model Development and Training (20 Days):** Building and training machine learning models, including k-means clustering.
- **Testing and Evaluation (10 Days):** Running the trained models on sample datasets to ensure accuracy and reliability.
- **Model Deployment (15 Days):** Deploying the segmentation system to cloud infrastructure for scalability.

Milestone: A functional segmentation system ready for real-world testing and deployment.

5. **Implementation (56 Days)**

During this phase, the system is deployed and integrated into real-world applications:

- **Environment Setup (10 Days):** Preparing production and testing environments.
- **Real-Time Segmentation Deployment (20 Days):** Implementing real-time segmentation pipelines to handle live data.
- **System Testing (10 Days):** Comprehensive end-to-end testing to identify and resolve bugs.
- **Integration (16 Days):** Ensuring seamless integration with CRM platforms and marketing systems.

Deliverable: Fully implemented system, capable of delivering segmentation insights to businesses.

6. **Monitoring and Control (14 Days)**

Post-implementation, this phase ensures long-term reliability and maintenance:

- **Continuous Monitoring (7 Days):** Tracking system performance and accuracy metrics.
- **Bug Fixing and Updates (7 Days):** Addressing any operational issues that arise post-deployment.
  Milestone: A stable and optimized system ready for sustained use.

The accompanying Gantt chart provides a visual representation of the project timeline. Key aspects of the chart include:

- **Task Duration and Dependencies:** Each task is assigned a start and end date, with dependencies clearly mapped to illustrate the sequence of activities. This helps in identifying bottlenecks and ensuring smooth handovers between phases.

- **Phased Progression:** Tasks are grouped into sequential phases, showing how the project moves from inception to completion.

- **Parallel Tasks:** The chart highlights tasks that can be performed simultaneously, maximizing resource utilization and reducing overall project duration.

- **Critical Path:** The Gantt chart identifies critical tasks that directly impact the project's completion date, allowing project managers to prioritize these activities.

3. TECHNICAL SPECIFICATION

3.1 Requirements

3.1.1 Functional

- **Data Collection and Integration:** The system will collect customer demographic, behavioral, and transactional data from multiple sources, including CRM systems, social media platforms, and historical purchase data.

  By integrating data from diverse touchpoints, the system will ensure a comprehensive dataset that reflects customer preferences, habits, and engagement levels, which will serve as a foundation for accurate segmentation.

- **Exploratory Data Analysis (EDA):** The system will perform EDA to identify key patterns, trends, and correlations within the collected data. This step will involve visualizations and statistical analyses to uncover hidden insights about customer behavior.

  The findings will guide the feature selection for the segmentation model, helping the algorithm to focus on the most relevant customer attributes for segment creation.

- **Apply Machine Learning Algorithms:** The system will employ a range of unsupervised machine learning algorithms, including K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM).

  These algorithms will group customers into meaningful segments based on shared characteristics such as purchasing behavior, demographic attributes, and preferences, allowing businesses to gain actionable insights from complex datasets.

- **Real-Time Segmentation:** As new customer data is collected, the system will enable real-time segmentation.

  This capability will allow businesses to continuously update and refine their customer segments based on the most recent data, ensuring that marketing strategies and product offerings remain relevant and aligned with current customer needs.

- **Model Evaluation and Optimization:** To ensure high-quality segmentation, the system will evaluate the performance of machine learning models using metrics such as silhouette scores, Davies-Bouldin index, and inertia.

  Model optimization techniques, including hyperparameter tuning and cross-validation, will be employed to enhance the accuracy and stability of the segmentation process.

- **Model Deployment:** The segmentation model will be deployed using cloud-based services, enabling real-time access and scalability.

  Through deployment on platforms such as AWS or Google Cloud, businesses will be able to easily integrate the segmentation model into their existing systems and access the outputs through APIs or dashboards.

3.1.2 Non-Functional

- **Scalability:** The system must be capable of handling an increasing volume of customer data over time. As the business grows and the dataset expands, the system must efficiently scale to accommodate more data without sacrificing performance.

  The use of cloud infrastructure will provide the flexibility to scale processing power and storage as needed.

- **Reliability:** The system must be designed for high availability, ensuring that it operates 24/7 with minimal downtime. This reliability is crucial for real-time segmentation and analysis, enabling businesses to make timely, data-driven decisions.

  The deployment on cloud platforms offers redundancy and failover capabilities, minimizing potential disruptions.

- **Security:** Protecting customer data is paramount. The system will implement stringent security measures to safeguard sensitive customer information.

  Compliance with data privacy regulations such as GDPR will be prioritized, ensuring that customer data is securely stored, processed, and transmitted. Access controls and encryption techniques will further enhance data security.

- **Performance:** The system must deliver real-time predictions and segmentation updates with a response time of under 1 second for each task.

  To achieve this, optimization techniques will be employed, and the system will utilize high-performance computing resources available on cloud platforms. This ensures that the segmentation process remains swift and accurate even with large datasets.

- **Usability:** The system should feature an intuitive, user-friendly interface for business and marketing teams. This interface will allow users to easily interpret the results of the segmentation analysis and use them for strategic decision-making.

  Dashboards and visualizations will be designed to provide clear insights and actionable recommendations without requiring advanced technical expertise.

3.2 Feasibility Study

3.2.1 Technical Feasibility

- **Technology Availability:** The project will leverage widely-used and reliable machine learning libraries such as scikit-learn for traditional algorithms, TensorFlow for advanced models, and cloud platforms like AWS or Google Cloud for deployment and scalability.

  These technologies are well-documented and supported, ensuring ease of implementation and future updates.

- **Technical Expertise:** The project team consists of experts in Python programming, machine learning, data science, and cloud infrastructure.

  With experience in implementing ML models and managing cloud environments (e.g., GCP, AWS), the team is well-equipped to handle the development, integration, and optimization phases of the project.

- **Infrastructure:** Cloud servers will be utilized for storing and processing large datasets, ensuring that the infrastructure is scalable and flexible.

  High-performance computing resources, such as GPUs and multi-core processing, will be used to train machine learning models efficiently, allowing for faster and more accurate results.

- **Integration:** The model will be integrated with existing customer relationship management (CRM) systems and marketing platforms, ensuring seamless adoption without disrupting current business operations.

  APIs will facilitate smooth data exchange between systems, and integration with business intelligence tools will enable the visualization of segmentation outputs.

3.2.2 Economic Feasibility

- **Cost-Benefit Analysis:** Initial costs will include development, cloud infrastructure, and potential licensing fees for third-party services. However, the use of open-source machine learning libraries (such as scikit-learn and TensorFlow) helps to minimize software costs.

  The long-term benefits of accurate market segmentation—such as improved customer targeting, optimized marketing strategies, and increased sales—are expected to outweigh these costs.

- **Budget:** The budget will primarily cover the development team's salaries, cloud infrastructure costs, and hardware for processing and storing large volumes of data.

  Given the use of open-source technologies, the software costs will be minimal, but the infrastructure costs will depend on the scale of data storage and processing required.

- **Return on Investment (ROI):** By enabling businesses to make data-driven decisions and tailor their marketing efforts to specific customer segments, the model can significantly increase revenue.

  Personalized customer engagement and targeted marketing efforts will lead to higher conversion rates, improved customer satisfaction, and increased customer retention, offering substantial ROI.

- **Funding:** Initial funding will be required to cover the cost of cloud infrastructure and development resources.

  However, after the system is developed and deployed, ongoing costs will be manageable, with predictable expenses related to cloud usage and system maintenance.

3.2.3 Social Feasibility

- **User Acceptance:** Marketing teams and business leaders are likely to embrace the project due to the value it offers in terms of actionable customer insights.

  The automation of segmentation tasks will free up resources for more strategic activities, making the system an attractive tool for businesses looking to optimize their marketing efforts.

- **Training and Support:** Comprehensive training for marketing and business teams will ensure they understand how to use the system and interpret its results.

  Additionally, a dedicated support team will be available to assist users in understanding the segmentation outcomes and using them effectively in decision-making.

- **Ethical Considerations:** The project will adhere to ethical standards in the handling of customer data, ensuring compliance with relevant privacy regulations such as GDPR.

  This will include proper consent mechanisms for data collection, anonymization of sensitive information, and strict access control protocols.

- **Impact on Workforce:** By automating the market segmentation process, the project will allow marketing teams to focus more on strategy and customer relationship management, rather than time-consuming manual analysis.

  The project may also create new roles related to data analytics, machine learning, and data-driven marketing, fostering skills development within the workforce.

3.2 System Specification

3.2.1 Hardware Specification

- **Processor:** A quad-core processor with a clock speed of 3.0 GHz or higher is recommended to handle the computation-intensive tasks of processing large datasets and running complex machine learning models.

  A powerful CPU will ensure efficient data handling and model training.

- **Memory (RAM):** A minimum of 16GB of RAM is necessary to support memory-intensive operations such as data processing, model training, and real-time segmentation tasks.

  Sufficient RAM is critical for preventing bottlenecks when working with large-scale datasets.

- **Storage:** A 256GB SSD is recommended for fast read/write operations, which is crucial for storing and accessing large datasets quickly.

  An SSD will ensure that data is loaded efficiently into memory, minimizing delays during model training and segmentation.

- **Graphics Processing Unit (GPU):** Although optional, an NVIDIA GPU is highly beneficial for accelerating machine learning tasks, particularly deep learning models.

  Using a GPU will significantly reduce the time required to train models, particularly on large datasets, by offloading computational tasks from the CPU.

- **Monitor:** An HD display is required to provide a clear view of data visualizations, coding environments, and model outputs.

  A high-resolution monitor will enhance the user experience when analyzing segmentation results or troubleshooting model performance.

3.2.2 Software Specification

- **Operating System:** The system will run on Windows 10, which is compatible with a wide range of machine learning libraries and development tools.

  This ensures that all required software will function seamlessly within the operating system environment.

- **Programming Languages:** Python will be the primary programming language for implementing the machine learning models.

  It is widely used in data science and machine learning due to its extensive libraries and ease of use. Python will facilitate the development, training, and evaluation of segmentation models.

- **Development Environment:** The system will be developed using Jupyter Notebook, Visual Studio Code, or Google Colab.

  These platforms are ideal for interactive development, debugging, and visualizing machine learning models and data analyses. Jupyter Notebook is particularly useful for data exploration, while Visual Studio Code provides a more robust development environment with advanced features like version control.

- **Libraries and Frameworks:** The following libraries will be used for data analysis, visualization, and machine learning:

✓ **scikit-learn:** For implementing unsupervised machine learning algorithms such as K-Means, hierarchical clustering, and Gaussian Mixture Models (GMM).

✓ **TensorFlow:** For advanced machine learning and deep learning models, if needed for more complex segmentation tasks.

✓ **Pandas:** For data manipulation and analysis, allowing for easy handling of structured data.

✓ **Matplotlib and Seaborn:** For creating visualizations that will help interpret data distributions, trends, and segmentation results.

- **Database:** The customer data will be stored in either MySQL or PostgreSQL databases. These relational databases are well-suited for handling structured data and providing fast access to large volumes of customer information.

  They will also support efficient querying and data management.

- **Version Control:** Git will be used for version control to manage code changes and collaborate efficiently within the development team.

  By using platforms like GitHub or GitLab, the project will maintain a clear history of all code updates, enabling easier tracking of modifications and team collaboration.

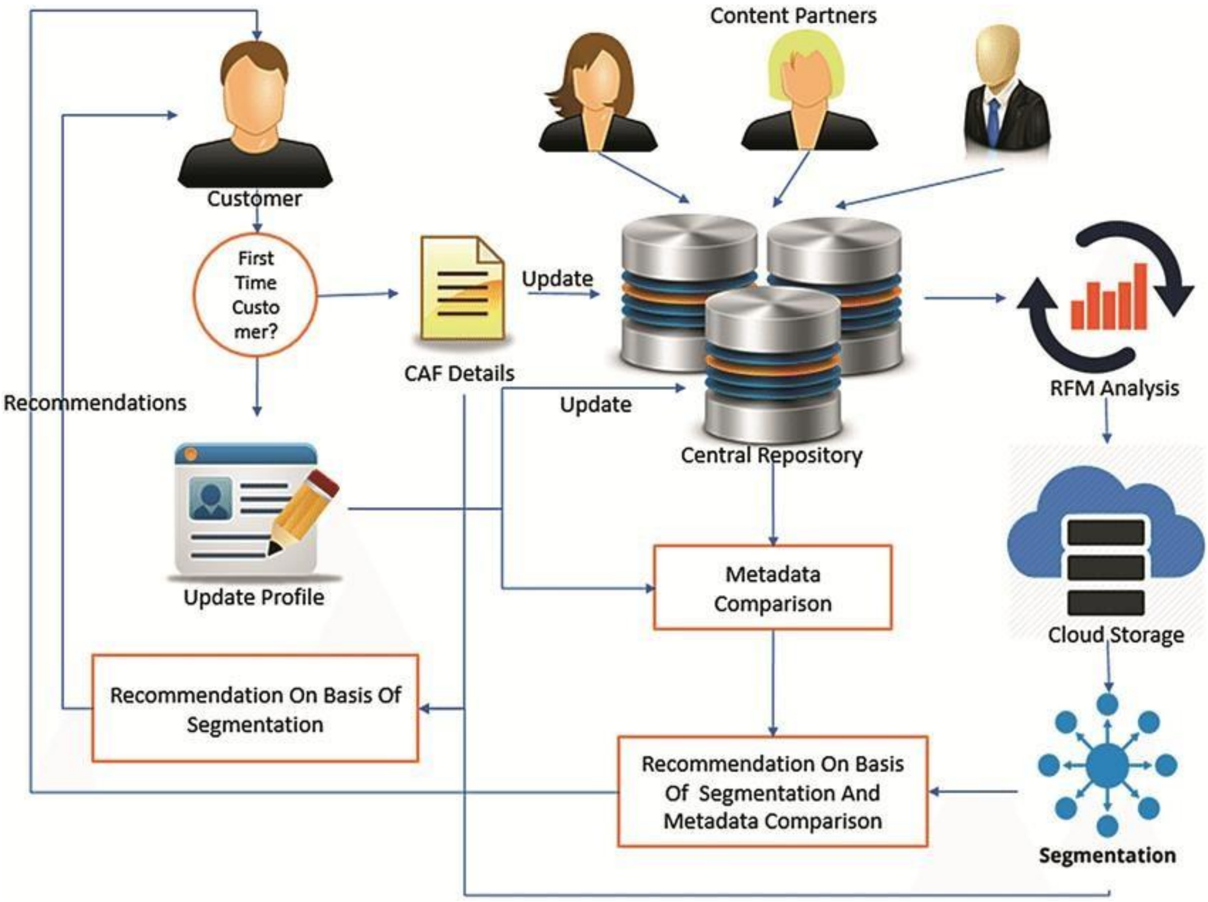
- **API Framework:** An API framework such as Sreamlit or Flask will be used to integrate the machine learning model into business applications.

  This will allow real-time access to the segmentation model, enabling businesses to send customer data for analysis and receive segmentation results dynamically. The API framework will ensure that the model can be accessed via HTTP requests, providing scalability and ease of integration with existing business systems.

4. DESIGN APPROACH AND DETAILS

4.1 System Architecture

Fig. 4.1. System Architecture of Market Segmentation and Analysis



The system architecture provides a robust, end-to-end framework for data-driven customer segmentation, integrating multiple components and processes to deliver actionable insights. It starts with Customer Data Collection, where demographic, behavioral, and transactional data is gathered through various channels such as online platforms, CRM systems, and customer feedback forms. This diverse data pool is stored in a Central Repository, serving as the foundational database for all subsequent processing.

The architecture incorporates advanced Data Processing and Segmentation capabilities, leveraging machine learning techniques such as clustering and RFM (Recency, Frequency, Monetary) analysis to identify distinct customer groups based on shared characteristics and behaviors. By performing Metadata Comparison, the system refines these segments further, utilizing additional data sources stored in Cloud Storage. This ensures the segmentation results are not only accurate but also relevant for real-world applications.

A key feature of the architecture is its ability to provide Personalized Recommendations. Using the refined customer segments, the system generates tailored recommendations for individual customers, offering products, services, or marketing strategies aligned with their preferences and purchasing behavior. These recommendations are fed back into the Update Profiles module, dynamically improving customer profiles and ensuring real-time adjustments to changing customer needs.

The integration of Cloud Storage and computational resources enhances the system's scalability, enabling it to handle large datasets and accommodate future growth. Real-time segmentation and feedback loops allow businesses to react promptly to customer behavior changes, maintaining relevance in fast-evolving markets. Furthermore, the system architecture is designed to support continuous learning, where the segmentation models are updated regularly with new data, ensuring sustained accuracy and effectiveness over time.

In summary, the system architecture represents a seamless integration of data collection, processing, machine learning, and actionable recommendation generation. This closed-loop framework not only optimizes customer satisfaction but also enables businesses to refine their marketing strategies, enhance operational efficiency, and achieve sustainable growth.

1. **Customer Data Collection:**

    - Customers interact with various touchpoints, such as web forms, mobile applications, and e-commerce platforms, where they input or provide access to their demographic data, behavior patterns, purchase history, and other relevant attributes.
    - This data is collected through APIs, web scraping, or direct integration with CRM and social media platforms.
    - The data is then stored in a **Central Repository**, a secure and scalable database that acts as the foundation for further analysis.

2. **Data Processing and Segmentation:**

    - Once customer data is collected, it undergoes preprocessing to ensure it is cleaned, normalized, and transformed into a suitable format for machine learning models.
    - Unsupervised machine learning algorithms such as **K-Means clustering, hierarchical clustering, and Gaussian Mixture Models (GMM)** are applied to identify patterns and segment customers into distinct groups based on shared characteristics and behaviors.
    - This stage also includes exploratory data analysis (EDA) to identify relevant features that influence segmentation, ensuring the model's effectiveness.

3. **Analysis and Comparison:**

    - To improve the segmentation model, the system performs **Metadata Comparison** using additional data sources, such as customer feedback, loyalty program information, or third-party data.
    - **RFM Analysis (Recency, Frequency, Monetary)** is used to categorize customers based on their purchasing behaviors, helping to refine the segmentation results by considering transaction recency, frequency, and monetary value.
    - This data is stored in **Cloud Storage** to facilitate easy access and scalability, enabling real-time updates to segmentation models as new data flows in.

4. **Personalized Recommendations:**

    - After the segmentation process, the system leverages the identified customer segments to generate **Personalized Recommendations** for each segment.
    - These recommendations could include targeted marketing campaigns, product suggestions, or personalized offers aimed at improving customer engagement and loyalty.
    - The system dynamically updates customer profiles based on the latest segmentation results, ensuring that marketing strategies are always aligned with current customer behavior.

5. **Real-Time Feedback and Decision-Making:**

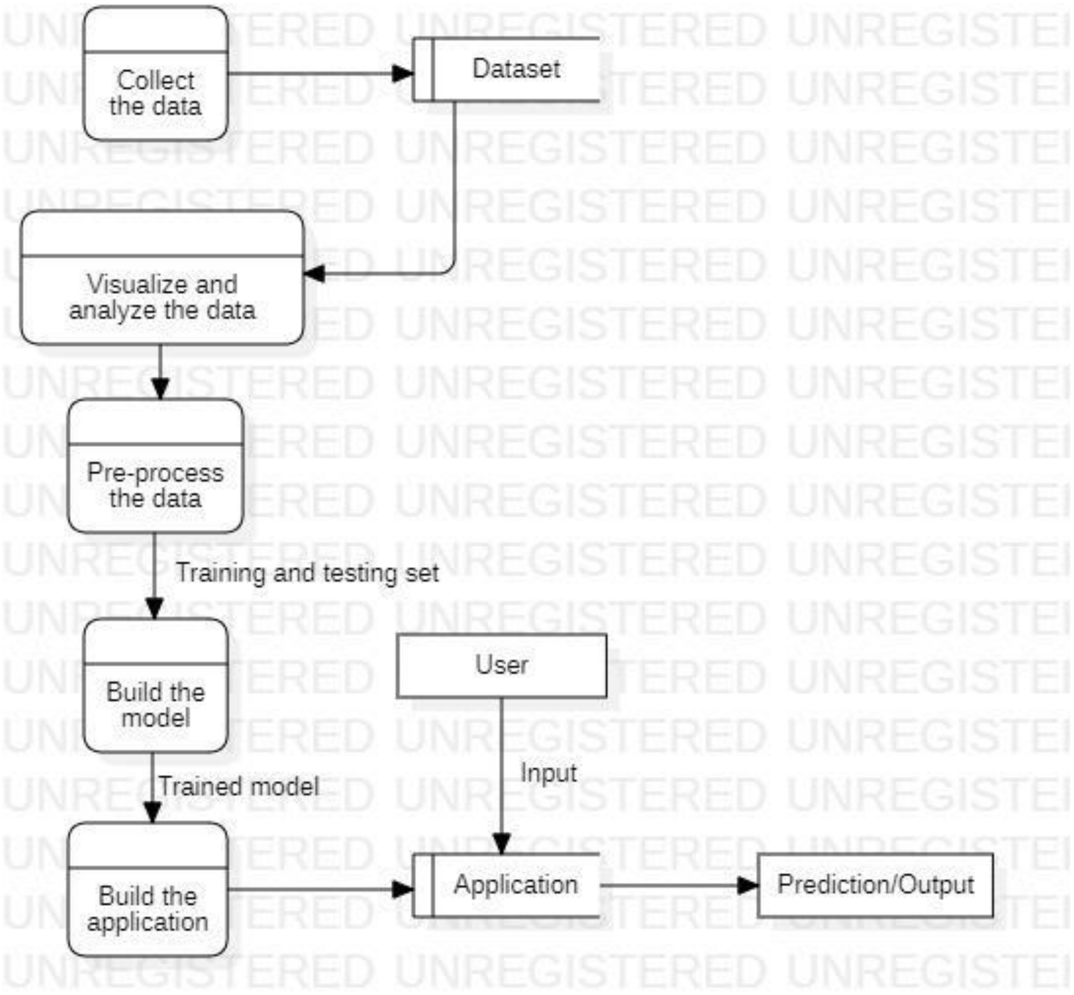
    - The architecture includes mechanisms for real-time feedback, allowing the system to continually refine its segmentation model as new data enters the system.
    - Business users can access segmentation results and recommendations via a user-friendly interface or an **API** for integration with existing CRM or marketing automation tools.
    - This real-time capability ensures that businesses can quickly adjust their strategies based on fresh insights, improving their agility and responsiveness to changing customer needs.

4.2 Design

The design of the market segmentation system is meticulously structured to ensure smooth execution of tasks, efficient data flow, and intuitive interaction between various components. It consists of several interconnected models, each contributing to the overall functionality and delivering actionable insights to the end-users.

4.2.1 Data Flow Diagram

Fig. 4.2.1. Data Flow Diagram



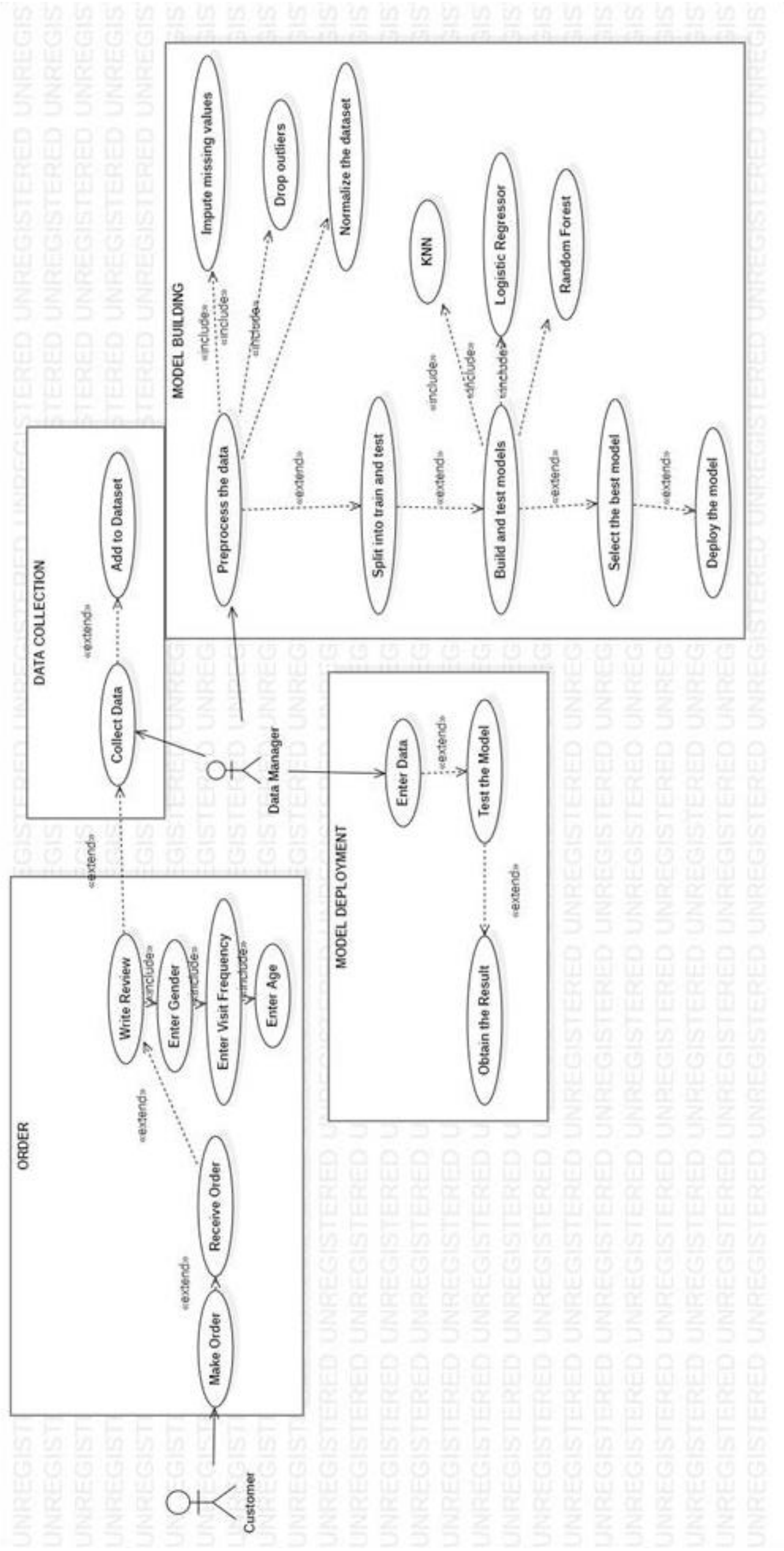
The Data Flow Diagram outlines the logical flow of information across the system, beginning with Data Collection from various sources such as customer demographics, transactional records, and behavioral metrics. This collected data is then funneled into the Dataset Module, which prepares it for further analysis.

The next step involves Data Visualization and Analysis, where trends, patterns, and anomalies are identified. This stage enables a deeper understanding of the data's structure and guides subsequent data preprocessing steps. The processed data is utilized to Build and Train Machine Learning Models, which focus on customer segmentation. Training and testing ensure the accuracy and reliability of these models.

Once the models are trained, they are integrated into the system via an Application Interface to deliver real-time predictions and segmentation results. This continuous loop allows for seamless updates and feedback, ensuring that businesses have access to up-to-date insights for decision-making. The predictive outputs are then utilized for crafting marketing strategies and providing personalized customer recommendations.

4.2.2 Use Case Diagram

Fig. 4.2.2. Use Case Diagram

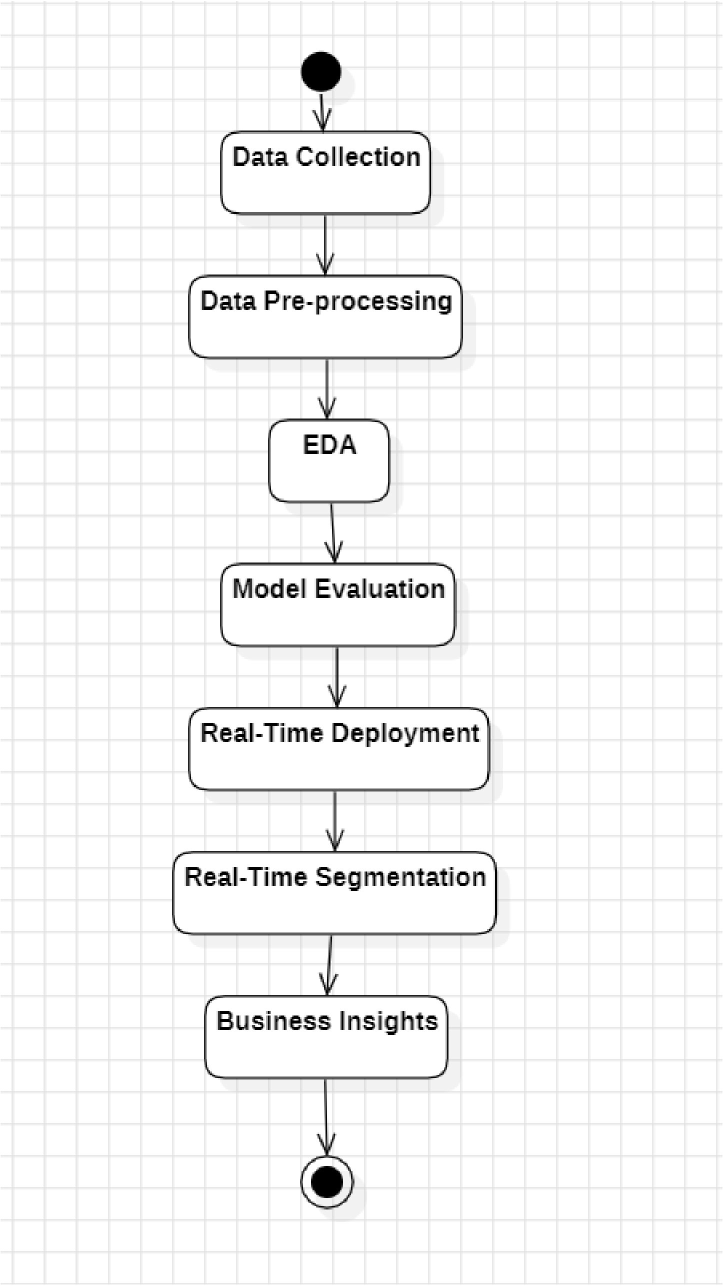


The Use Case Diagram provides a high-level visual representation of the system’s interaction with various actors. It showcases how different users, such as Marketing Teams and Business Analysts, interact with the system to perform essential tasks. Key functionalities include uploading customer data, visualizing segmentation insights, generating predictions, and accessing personalized recommendations.

The use case diagram highlights the centrality of machine learning algorithms in driving the segmentation process and the role of the user interface in delivering actionable insights.

4.2.3 Activity Diagram

Fig. 4.2.3. Activity Diagram



The Activity Diagram elaborates on the step-by-step processes involved in the system's operation. The workflow begins with Data Collection, where raw data is gathered from multiple sources. This is followed by Data Preprocessing, including cleaning, normalization, and transformation of data to ensure consistency and quality.

Next, Exploratory Data Analysis (EDA) is performed to extract meaningful features and relationships within the data, which are then fed into machine learning models for training and evaluation. The Model Evaluation step assesses the performance of these models using metrics like silhouette scores, ensuring accurate and reliable segmentation.

Upon successful model evaluation, the system proceeds to Real-Time Deployment, enabling continuous segmentation of new data as it enters the system. This functionality allows for Real-Time Segmentation, dynamically updating customer profiles and insights. Finally, the system generates actionable Business Insights, providing organizations with tailored strategies to improve customer targeting, enhance operational efficiency, and drive revenue growth.

5. METHODOLOGY AND TESTING

5.1 Module Description

The market segmentation process involves analyzing customer data and grouping customers into segments based on shared characteristics. This can be done using unsupervised machine learning algorithms such as K-means clustering, which does not require predefined labels for training. Below is a detailed explanation of the steps involved in the segmentation process.

1. **Data Collection and Understanding:**

- **Dataset:**
  - The first step is to obtain data relevant to customer behaviors or demographics. For instance, the dataset may include attributes like age, income, spending score, geographic location, product preferences, and online activity.
  - Common sources of data include publicly available datasets like Kaggle, company customer databases, or transaction logs.

- **Dataset Features:**
  - Features might include:
    - **Age**: Age of the customer
    - **Income**: Monthly income of the customer
    - **Spending Score**: A metric that reflects customer spending behavior
    - **Location**: Geographical location of the customer (country or city)
    - **Online Behavior**: Customer interaction on the website, including click rates, purchase history, etc.

2. **Data Preprocessing:**

- **Handling Missing Values:**
  - Missing data is a common issue in real-world datasets. We can impute missing values using techniques such as mean imputation, median imputation, or using more complex models like KNN imputation.
  - Alternatively, rows with missing data can be dropped, but this is generally avoided unless the amount of missing data is very small.

- **Feature Scaling:**
  - Machine learning models, especially distance-based algorithms like K-means, benefit from feature scaling. Features like income or spending scores can have vastly different scales. We can apply scaling using methods such as standardization (z-score scaling) or Min-Max scaling.

- **Handling Categorical Data:**
  - For categorical data, encoding techniques like **one-hot encoding** or **label encoding** can be used. This ensures that categorical data like customer location (e.g., city or country) can be represented numerically.

3. **Data Splitting:**

- **Train-Test Split:**
    - In an unsupervised learning task like clustering, we don't have labeled data. Therefore, we split the data into a training set and testing set to evaluate how well the model generalizes to unseen data.
    - A typical split ratio is 80/20 or 70/30, where 80% or 70% of the data is used for training, and the rest is used for evaluation.

4. **K-Means Clustering Algorithm:**

- **Selecting Number of Clusters (k):**
    - One of the challenges in K-means is choosing the number of clusters (k). A common technique for selecting k is the **Elbow Method**, where we run K-means for different values of k and plot the sum of squared distances (inertia) for each k. The "elbow" point on the graph represents the optimal k.

- **Training the Model:**
    - Once the value of k is selected, the K-means algorithm is trained on the customer data. The algorithm works by assigning each customer to a cluster based on their features and iteratively refining the cluster centers.

- **Cluster Assignment:**
    - The trained K-means model assigns each customer to one of the k clusters. These assignments are used to segment the customer base.

5. **Model Evaluation:**

- **Silhouette Score:**
    - The Silhouette Score is a measure of how well-separated the clusters are. A higher silhouette score indicates that customers are well-matched to their clusters, and the clusters are well-separated from each other.

    - Formula: Silhouette Score = (b-a)/max(a,b) where:
        - a is the average distance between a point and all other points in the same cluster.
        - b is the average distance between a point and all other points in the nearest cluster.

- **Inertia (Within-cluster Sum of Squares):**

    - Inertia measures how compact the clusters are. A lower inertia value indicates that the points are closer to the centroids of their clusters.

- **Cross-Validation:**

    - Since K-means is an unsupervised model, it's often difficult to perform cross-validation as in supervised learning. However, one can perform k-fold validation on the feature engineering process or assess the stability of clusters with different training sets.

5.2 Testing

Testing the machine learning model for market segmentation ensures the robustness and generalizability of the clustering results. This process involves multiple evaluation metrics, cross-validation techniques, and post-clustering analysis.

Testing Strategy:

1. **Cross-Validation of Preprocessing Steps:**

   - **Feature Scaling:**
     - Test different scaling techniques (e.g., Min-Max vs. StandardScaler) to see if they affect the clustering results. In some cases, one scaling method may lead to better-defined clusters.

2. **Cluster Quality Evaluation:**

   - **Silhouette Score:**
     - Calculate the silhouette score for each test to evaluate how well-separated the clusters are.
     - Example:

```
from sklearn.metrics import silhouette_score

# Calculate silhouette score on the test data
silhouette = silhouette_score(X_test, predicted_labels)
print(f"Silhouette Score: {silhouette}")
```

   - **Elbow Method for Optimal k:**
     - The optimal number of clusters (k) can also be determined using the **elbow method**. Plot the sum of squared distances (inertia) for different values of k and choose the k where the rate of decrease slows down (the "elbow").

   - **Davies-Bouldin Index:**
     - Another measure for cluster validation, the Davies-Bouldin index, evaluates the average similarity ratio between each pair of clusters. A lower value indicates better clustering.

3. **Post-Model Analysis:**

   - Once the model has been trained, and the clusters have been assigned, we analyze the customer segments. By examining the features within each cluster, we gain insights into the distinct characteristics of each segment.

   - For instance, one segment might represent "high-income, high-spending" customers, while another could represent "low-income, low-spending" customers.

- This analysis can be visualized using 2D or 3D scatter plots, bar charts, or radar charts to highlight the unique traits of each segment.

4. Testing on New Data (Out-of-Sample Testing):

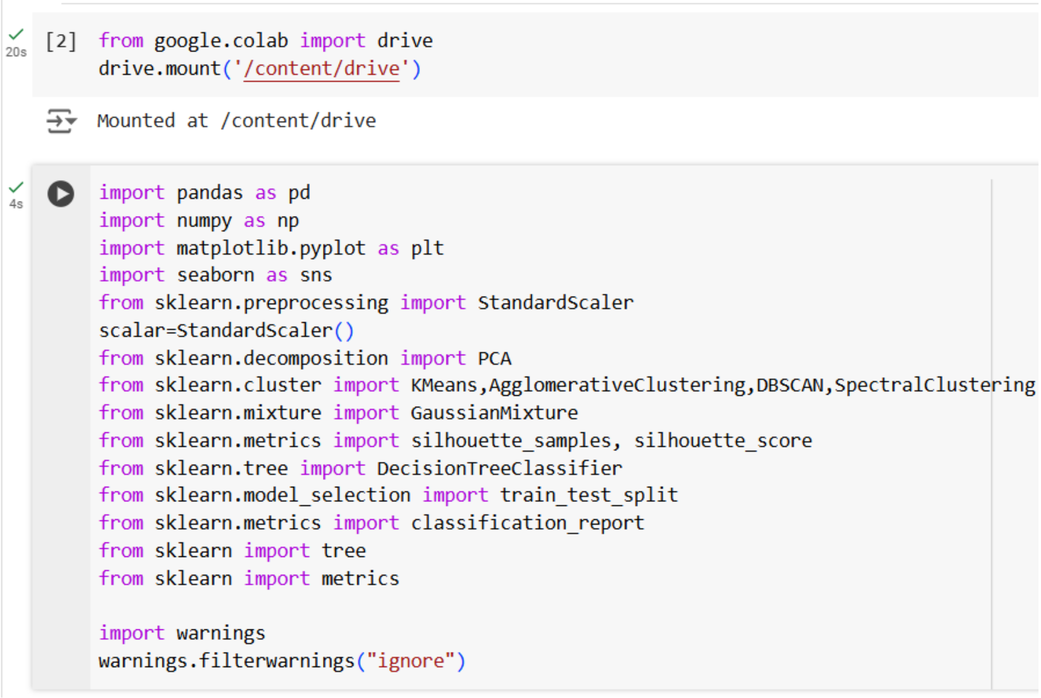
- While unsupervised learning doesn’t have explicit test labels, testing on new data is still important. This involves using new, unseen customer data and assigning it to the existing clusters. If the model works well, the new data should fit neatly into one of the pre-defined segments.

Code Snippet for Cluster Evaluation:

```
from sklearn.metrics import davies_bouldin_score

# Evaluate the model using Davies-Bouldin Index
db_index = davies_bouldin_score(X_test, predicted_labels)
print(f"Davies-Bouldin Index: {db_index}")
```

6. PROJECT DEMONSTRATION

6.1 Importing Libraries

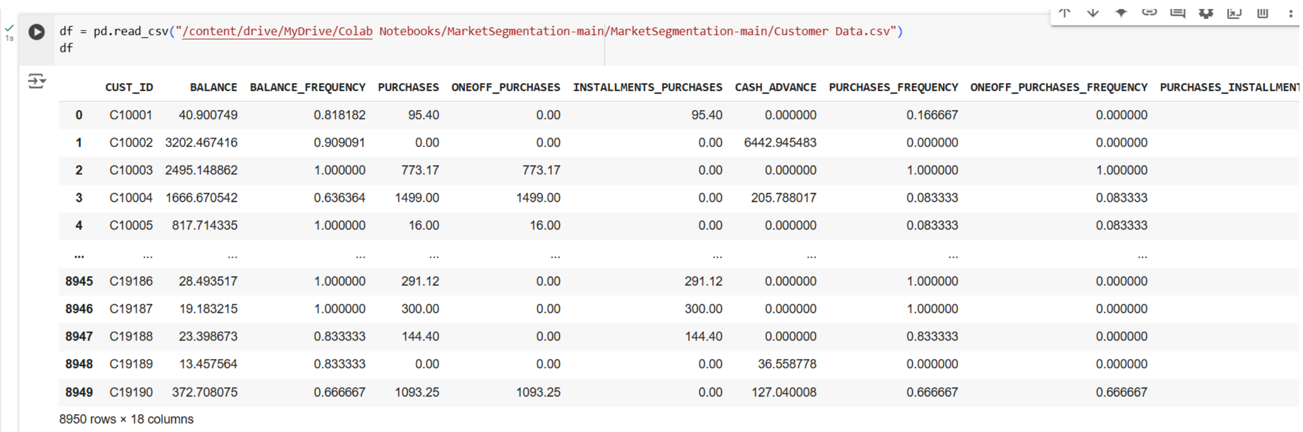


```
[2] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
scalar=StandardScaler()
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans,AgglomerativeClustering,DBSCAN,SpectralClustering
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_samples, silhouette_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn import tree
from sklearn import metrics

import warnings
warnings.filterwarnings("ignore")
```

6.2 Loading the dataset

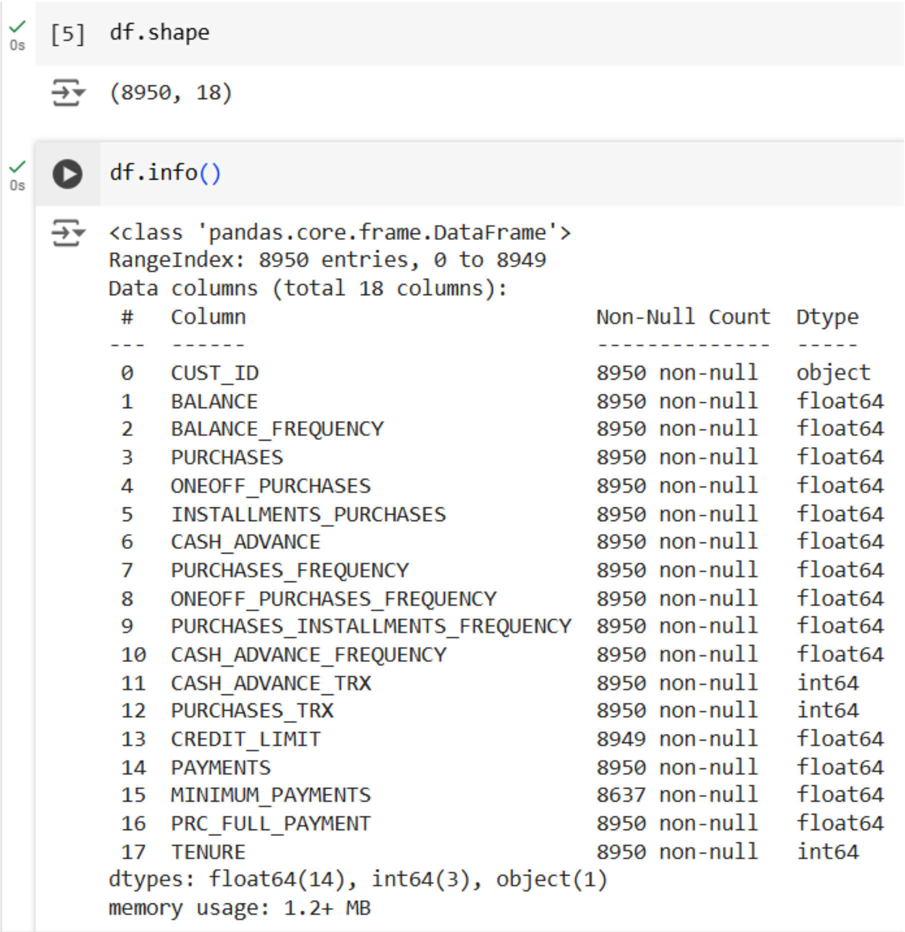


```
df = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/MarketSegmentation-main/MarketSegmentation-main/Customer Data.csv")
df
```

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENT
0	C10001	40.900749	0.818182	95.40	0.00	95.40	0.000000	0.166667	0.000000	
1	C10002	3202.467416	0.909091	0.00	0.00	0.00	6442.945483	0.000000	0.000000	
2	C10003	2495.148862	1.000000	773.17	773.17	0.00	0.000000	1.000000	1.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.00	205.788017	0.083333	0.083333	
4	C10005	817.714335	1.000000	16.00	16.00	0.00	0.000000	0.083333	0.083333	
...	...	...	...	...	...	...	...	...	...	
8945	C19186	28.493517	1.000000	291.12	0.00	291.12	0.000000	1.000000	0.000000	
8946	C19187	19.183215	1.000000	300.00	0.00	300.00	0.000000	1.000000	0.000000	
8947	C19188	23.398673	0.833333	144.40	0.00	144.40	0.000000	0.833333	0.000000	
8948	C19189	13.457564	0.833333	0.00	0.00	0.00	36.558778	0.000000	0.000000	
8949	C19190	372.708075	0.666667	1093.25	1093.25	0.00	127.040008	0.666667	0.666667	

8950 rows × 18 columns

6.3 EDA

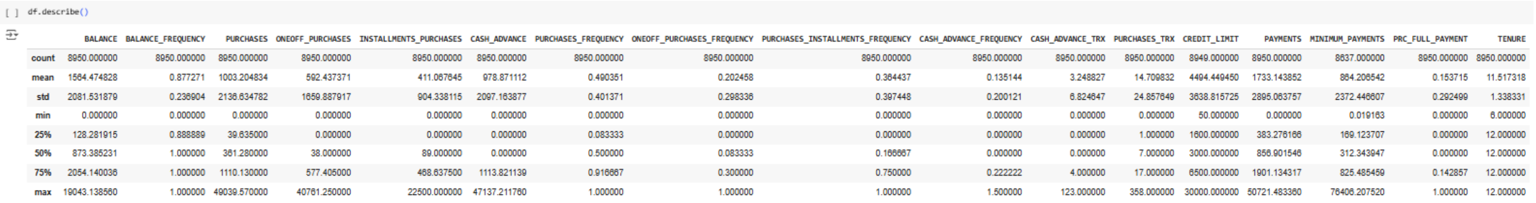


```
[5] df.shape
```

```
(8950, 18)
```

```
df.info()
```

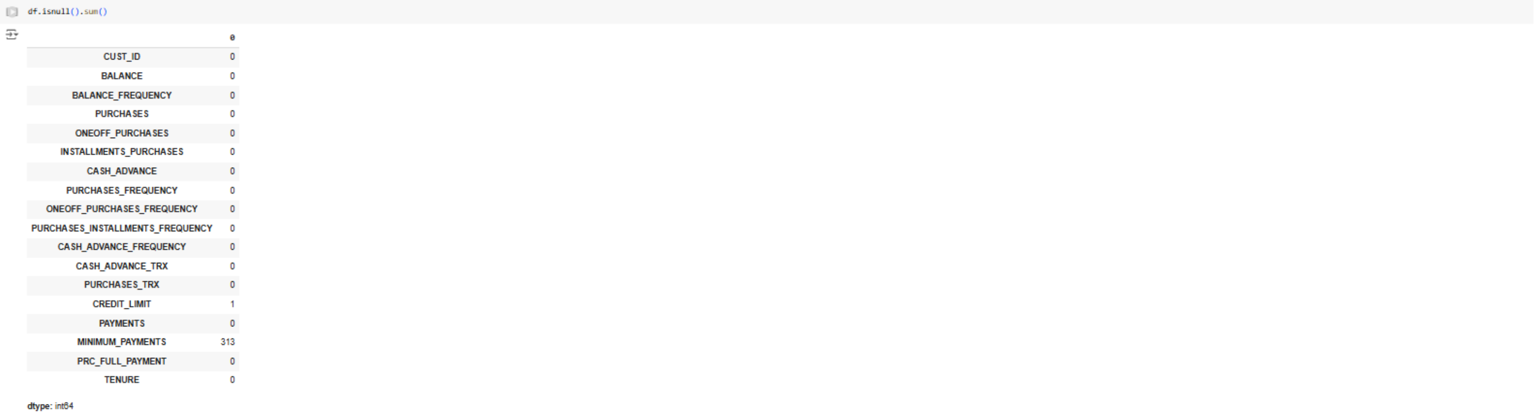
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8950 entries, 0 to 8949
Data columns (total 18 columns):
 #   Column                            Non-Null Count  Dtype  
---  ------                            --------------  -----  
 0   CUST_ID                           8950 non-null   object 
 1   BALANCE                           8950 non-null   float64
 2   BALANCE_FREQUENCY                 8950 non-null   float64
 3   PURCHASES                         8950 non-null   float64
 4   ONEOFF_PURCHASES                  8950 non-null   float64
 5   INSTALLMENTS_PURCHASES            8950 non-null   float64
 6   CASH_ADVANCE                      8950 non-null   float64
 7   PURCHASES_FREQUENCY               8950 non-null   float64
 8   ONEOFF_PURCHASES_FREQUENCY        8950 non-null   float64
 9   PURCHASES_INSTALLMENTS_FREQUENCY  8950 non-null   float64
 10  CASH_ADVANCE_FREQUENCY            8950 non-null   float64
 11  CASH_ADVANCE_TRX                  8950 non-null   int64  
 12  PURCHASES_TRX                     8950 non-null   int64  
 13  CREDIT_LIMIT                      8949 non-null   float64
 14  PAYMENTS                          8950 non-null   float64
 15  MINIMUM_PAYMENTS                  8637 non-null   float64
 16  PRC_FULL_PAYMENT                  8950 non-null   float64
 17  TENURE                            8950 non-null   int64  
dtypes: float64(14), int64(3), object(1)
memory usage: 1.2+ MB
```



```
df.describe()
```

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY
count	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000	8950.000000
mean	1564.474828	0.877271	1003.204834	592.437371	411.067645	978.871112	0.490351	0.202458	0.364437
std	2081.531879	0.236904	2136.634782	1659.887917	904.338115	2097.163877	0.401371	0.298336	0.397448
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	128.281915	0.888889	39.635000	0.000000	0.000000	0.000000	0.083333	0.000000	0.000000
50%	873.385231	1.000000	361.280000	38.000000	89.000000	0.000000	0.500000	0.083333	0.166667
75%	2054.140036	1.000000	1110.130000	577.405000	468.637500	1113.821139	0.916667	0.300000	0.750000
max	19043.138560	1.000000	49039.570000	40761.250000	22500.000000	47137.211760	1.000000	1.000000	1.000000

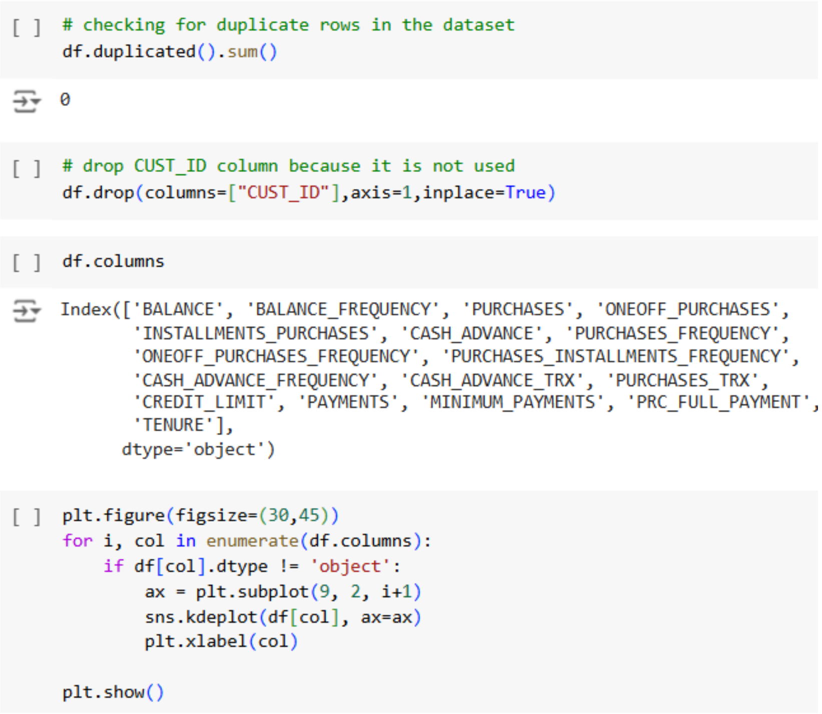
	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
count	8950.000000	8950.000000	8950.000000	8949.000000	8950.000000	8637.000000	8950.000000	8950.000000
mean	0.135144	3.248827	14.709832	4494.449450	1733.143852	864.206542	0.153715	11.517318
std	0.200121	6.824647	24.857649	3638.815725	2895.063757	2372.446607	0.292499	1.338331
min	0.000000	0.000000	0.000000	50.000000	0.000000	0.019163	0.000000	6.000000
25%	0.000000	0.000000	1.000000	1600.000000	383.276166	169.123707	0.000000	12.000000
50%	0.000000	0.000000	7.000000	3000.000000	856.901546	312.343947	0.000000	12.000000
75%	0.222222	4.000000	17.000000	6500.000000	1901.134317	825.485459	0.142857	12.000000
max	1.500000	123.000000	358.000000	30000.000000	50721.483360	76406.207520	1.000000	12.000000



```
df.isnull().sum()
```

	0
CUST_ID	0
BALANCE	0
BALANCE_FREQUENCY	0
PURCHASES	0
ONEOFF_PURCHASES	0
INSTALLMENTS_PURCHASES	0
CASH_ADVANCE	0
PURCHASES_FREQUENCY	0
ONEOFF_PURCHASES_FREQUENCY	0
PURCHASES_INSTALLMENTS_FREQUENCY	0
CASH_ADVANCE_FREQUENCY	0
CASH_ADVANCE_TRX	0
PURCHASES_TRX	0
CREDIT_LIMIT	1
PAYMENTS	0
MINIMUM_PAYMENTS	313
PRC_FULL_PAYMENT	0
TENURE	0

dtype: int64



```
[ ]  # checking for duplicate rows in the dataset
     df.duplicated().sum()
```

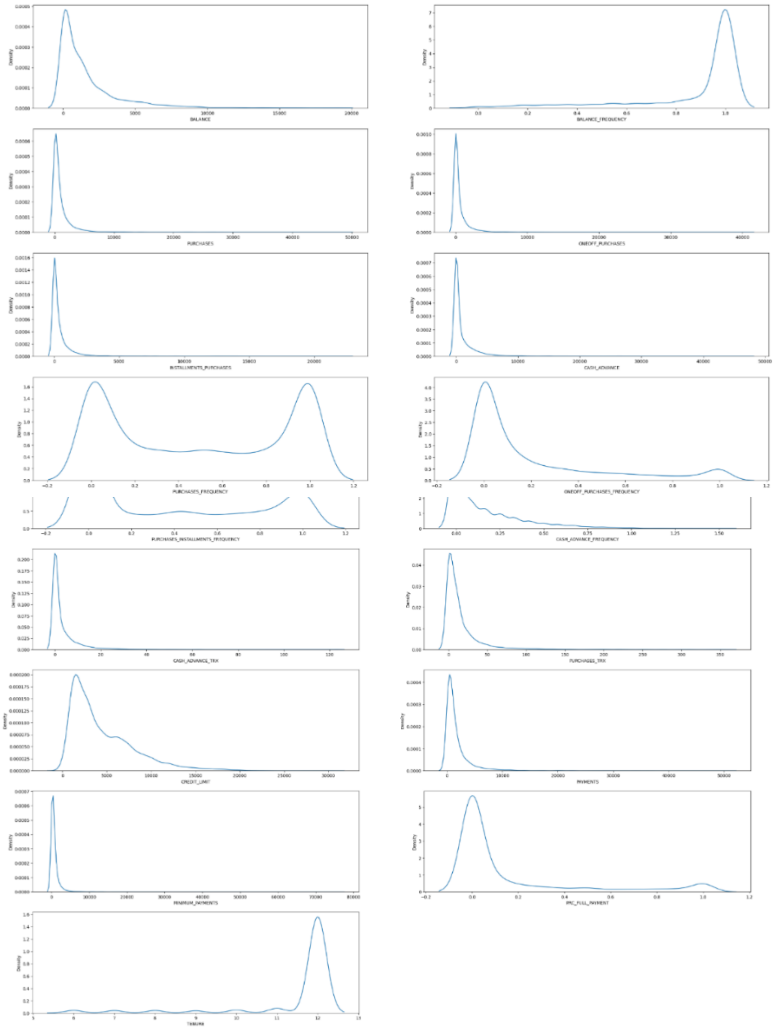
```
0
```

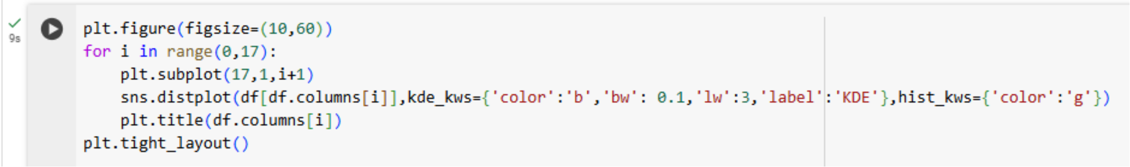
```
[ ]  # drop CUST_ID column because it is not used
     df.drop(columns=["CUST_ID"],axis=1,inplace=True)
```

```
[ ]  df.columns
```

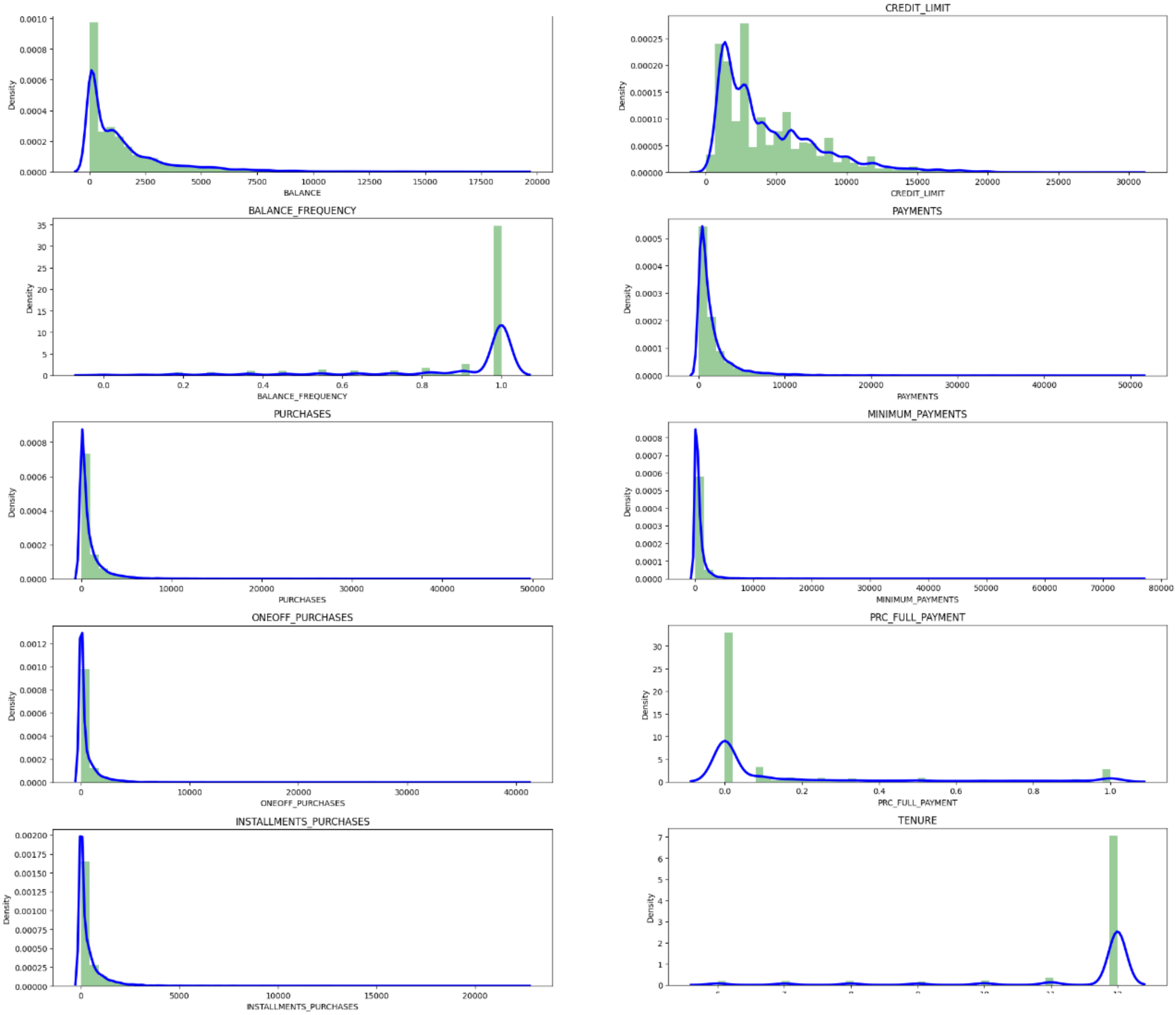
```
Index(['BALANCE', 'BALANCE_FREQUENCY', 'PURCHASES', 'ONEOFF_PURCHASES',
       'INSTALLMENTS_PURCHASES', 'CASH_ADVANCE', 'PURCHASES_FREQUENCY',
       'ONEOFF_PURCHASES_FREQUENCY', 'PURCHASES_INSTALLMENTS_FREQUENCY',
       'CASH_ADVANCE_FREQUENCY', 'CASH_ADVANCE_TRX', 'PURCHASES_TRX',
       'CREDIT_LIMIT', 'PAYMENTS', 'MINIMUM_PAYMENTS', 'PRC_FULL_PAYMENT',
       'TENURE'],
      dtype='object')
```

```
[ ]  plt.figure(figsize=(30,45))
     for i, col in enumerate(df.columns):
         if df[col].dtype != 'object':
             ax = plt.subplot(9, 2, i+1)
             sns.kdeplot(df[col], ax=ax)
             plt.xlabel(col)

     plt.show()
```

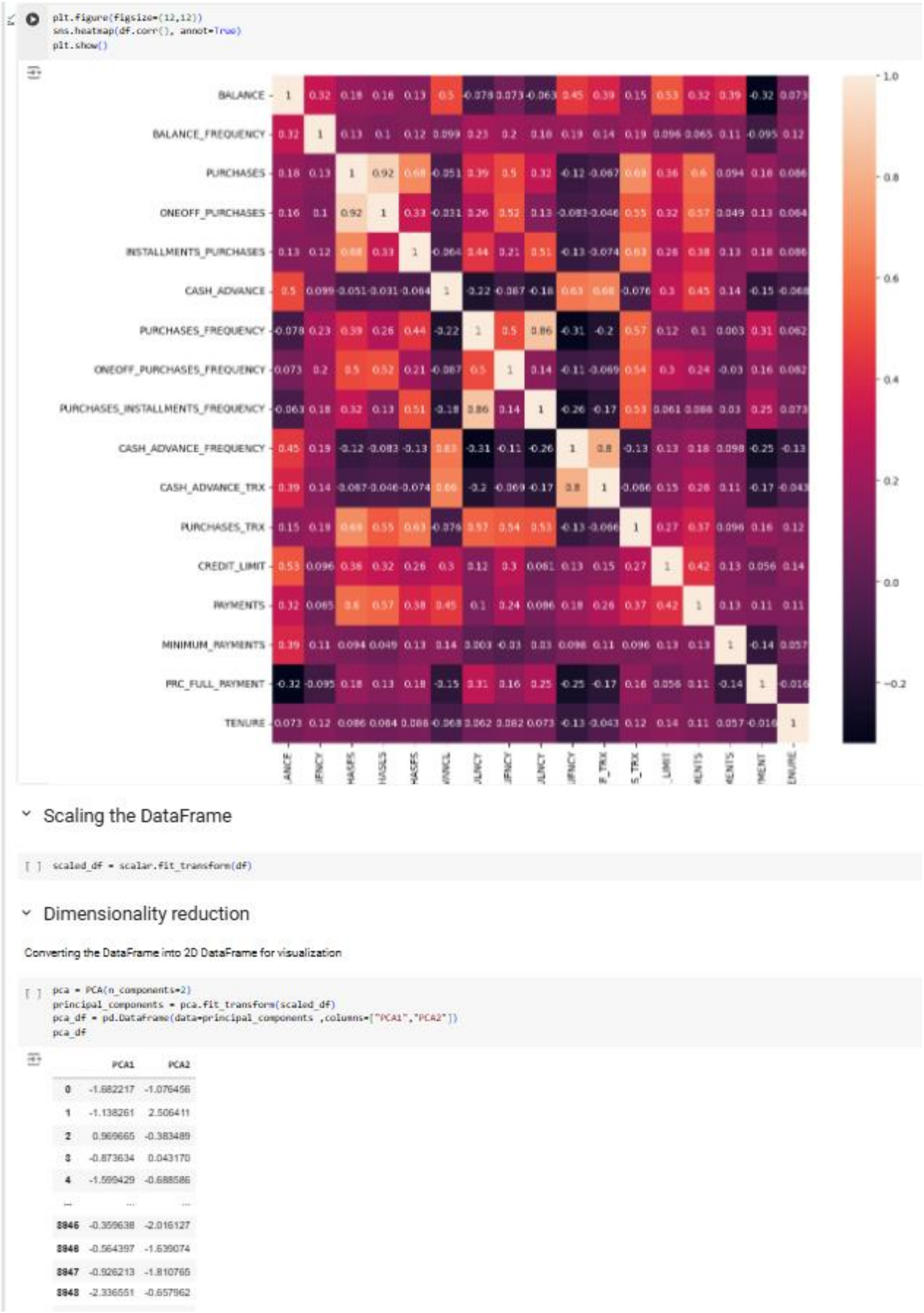




```
plt.figure(figsize=(10,60))
for i in range(0,17):
    plt.subplot(17,1,i+1)
    sns.distplot(df[df.columns[i]],kde_kws={'color':'b','bw': 0.1,'lw':3,'label':'KDE'},hist_kws={'color':'g'})
    plt.title(df.columns[i])
plt.tight_layout()
```



6.4 Scaling the DataFrame and Dimensionality reduction



```
plt.figure(figsize=(12,12))
sns.heatmap(df.corr(), annot=True)
plt.show()
```

	ANCE	JENCY	HASES	HASES	HASES	VANCE	JENCY	JENCY	JENCY	JENCY	E_TRX	S_TRX	LIMIT	MENTS	MENTS	MENT	ENURE
BALANCE	1	0.32	0.18	0.16	0.13	0.5	-0.078	0.073	-0.063	0.45	0.39	0.15	0.53	0.32	0.39	-0.32	0.073
BALANCE_FREQUENCY	0.32	1	0.13	0.1	0.12	0.099	0.23	0.2	0.18	0.19	0.14	0.19	0.096	0.065	0.11	-0.095	0.12
PURCHASES	0.18	0.13	1	0.92	0.68	-0.051	0.39	0.5	0.32	-0.12	-0.067	0.69	0.36	0.6	0.094	0.18	0.086
ONEOFF_PURCHASES	0.16	0.1	0.92	1	0.33	-0.031	0.26	0.52	0.13	-0.083	-0.046	0.55	0.32	0.57	0.049	0.13	0.064
INSTALLMENTS_PURCHASES	0.13	0.12	0.68	0.33	1	-0.064	0.44	0.21	0.51	-0.13	-0.074	0.63	0.26	0.38	0.13	0.18	0.086
CASH_ADVANCE	0.5	0.099	-0.051	-0.031	-0.064	1	-0.22	-0.087	-0.18	0.63	0.66	-0.076	0.3	0.45	0.14	-0.15	-0.068
PURCHASES_FREQUENCY	-0.078	0.23	0.39	0.26	0.44	-0.22	1	0.5	0.86	-0.31	-0.2	0.57	0.12	0.1	0.003	0.31	0.062
ONEOFF_PURCHASES_FREQUENCY	0.073	0.2	0.5	0.52	0.21	-0.087	0.5	1	0.14	-0.11	-0.069	0.54	0.3	0.24	-0.03	0.16	0.082
PURCHASES_INSTALLMENTS_FREQUENCY	-0.063	0.18	0.32	0.13	0.51	-0.18	0.86	0.14	1	-0.26	-0.17	0.53	0.061	0.086	0.03	0.25	0.073
CASH_ADVANCE_FREQUENCY	0.45	0.19	-0.12	-0.083	-0.13	0.63	-0.31	-0.11	-0.26	1	0.8	-0.13	0.13	0.18	0.098	-0.25	-0.13
CASH_ADVANCE_TRX	0.39	0.14	-0.067	-0.046	-0.074	0.66	-0.2	-0.069	-0.17	0.8	1	-0.066	0.15	0.26	0.11	-0.17	-0.043
PURCHASES_TRX	0.15	0.19	0.69	0.55	0.63	-0.076	0.57	0.54	0.53	-0.13	-0.066	1	0.27	0.37	0.096	0.16	0.12
CREDIT_LIMIT	0.53	0.096	0.36	0.32	0.26	0.3	0.12	0.3	0.061	0.13	0.15	0.27	1	0.42	0.13	0.056	0.14
PAYMENTS	0.32	0.065	0.6	0.57	0.38	0.45	0.1	0.24	0.086	0.18	0.26	0.37	0.42	1	0.13	0.11	0.11
MINIMUM_PAYMENTS	0.39	0.11	0.094	0.049	0.13	0.14	0.003	-0.03	0.03	0.098	0.11	0.096	0.13	0.13	1	-0.14	0.057
PRC_FULL_PAYMENT	-0.32	-0.095	0.18	0.13	0.18	-0.15	0.31	0.16	0.25	-0.25	-0.17	0.16	0.056	0.11	-0.14	1	-0.016
TENURE	0.073	0.12	0.086	0.064	0.086	-0.068	0.062	0.082	0.073	-0.13	-0.043	0.12	0.14	0.11	0.057	-0.016	1

Scaling the DataFrame

```
scaled_df = scalar.fit_transform(df)
```

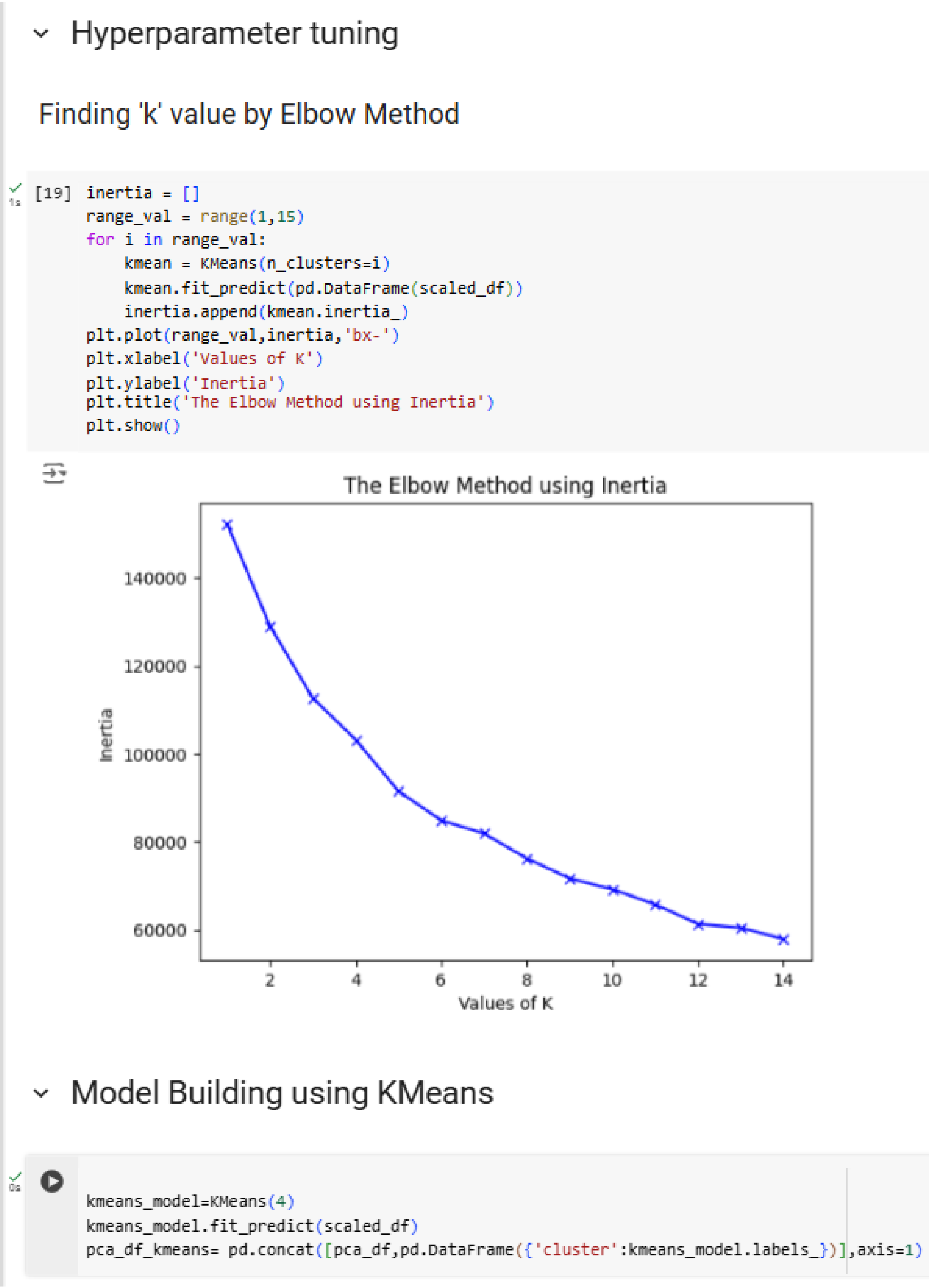
Dimensionality reduction

Converting the DataFrame into 2D DataFrame for visualization

```
pca = PCA(n_components=2)
principal_components = pca.fit_transform(scaled_df)
pca_df = pd.DataFrame(data=principal_components ,columns=["PCA1","PCA2"])
pca_df
```

	PCA1	PCA2
0	-1.682217	-1.076456
1	-1.138261	2.506411
2	0.969665	-0.383489
3	-0.873634	0.043170
4	-1.599429	-0.688586
...	...	...
8945	-0.359638	-2.016127
8946	-0.564397	-1.639074
8947	-0.926213	-1.810765
8948	-2.336551	-0.657962

6.5 Hyperparameter tuning and Model Building using KMeans



Hyperparameter tuning

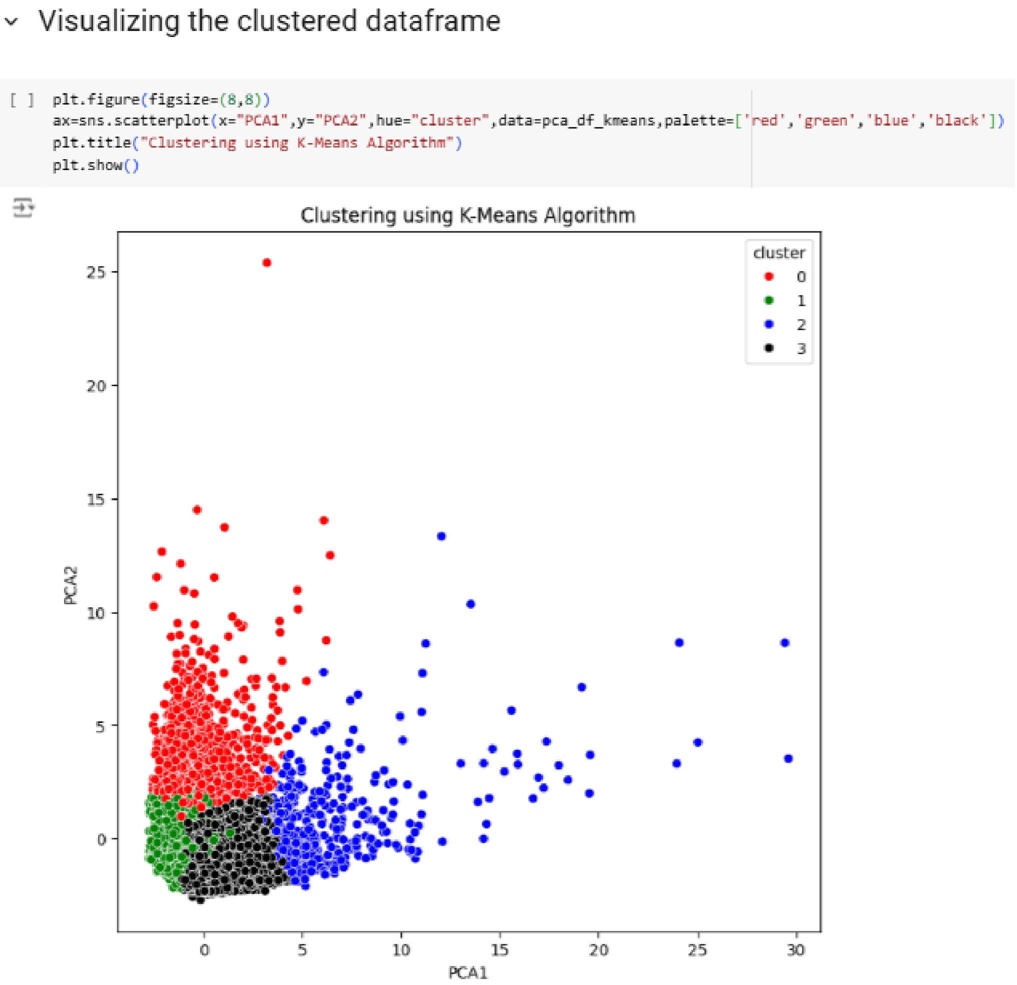
Finding 'k' value by Elbow Method

```
[19] inertia = []
     range_val = range(1,15)
     for i in range_val:
         kmean = KMeans(n_clusters=i)
         kmean.fit_predict(pd.DataFrame(scaled_df))
         inertia.append(kmean.inertia_)
     plt.plot(range_val,inertia,'bx-')
     plt.xlabel('Values of K')
     plt.ylabel('Inertia')
     plt.title('The Elbow Method using Inertia')
     plt.show()
```

Model Building using KMeans

```
kmeans_model=KMeans(4)
kmeans_model.fit_predict(scaled_df)
pca_df_kmeans= pd.concat([pca_df,pd.DataFrame({'cluster':kmeans_model.labels_})],axis=1)
```

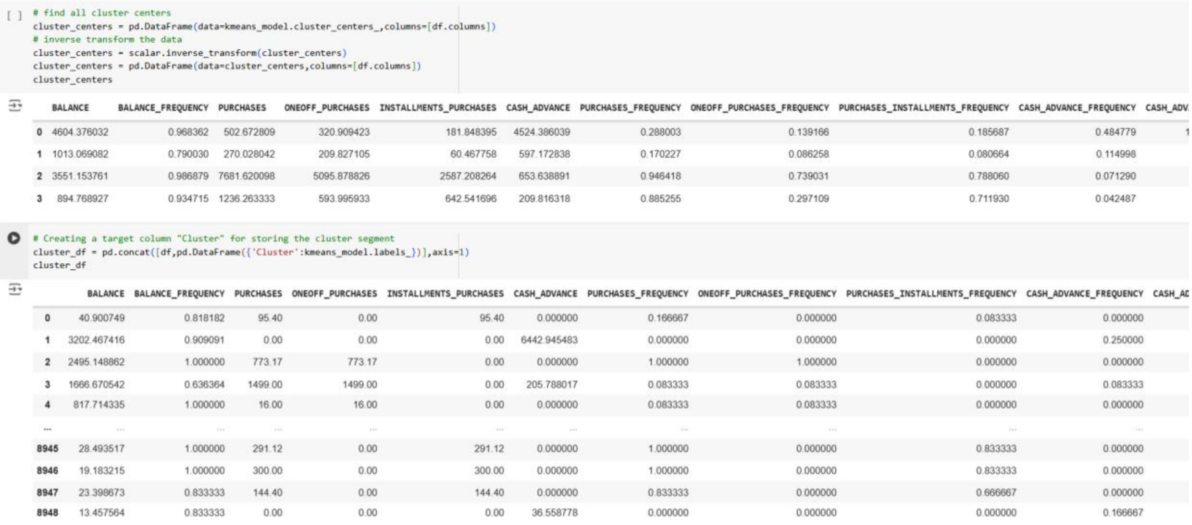
6.6 Visualizing the clustered dataframe



Visualizing the clustered dataframe

```
[ ] plt.figure(figsize=(8,8))
    ax=sns.scatterplot(x="PCA1",y="PCA2",hue="cluster",data=pca_df_kmeans,palette=['red','green','blue','black'])
    plt.title("Clustering using K-Means Algorithm")
    plt.show()
```

6.7 Saving the kmeans clustering model and the data with cluster label

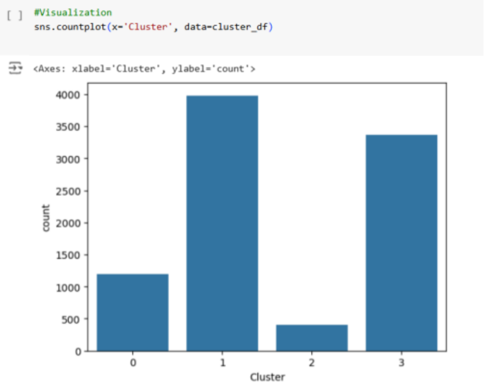


```
[ ] # find all cluster centers
    cluster_centers = pd.DataFrame(data=kmeans_model.cluster_centers_,columns=[df.columns])
    # inverse transform the data
    cluster_centers = scalar.inverse_transform(cluster_centers)
    cluster_centers = pd.DataFrame(data=cluster_centers,columns=[df.columns])
    cluster_centers
```

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADV
0	4604.376032	0.968362	502.672809	320.909423	181.848395	4524.386039	0.288003	0.139166	0.185687	0.484779	1
1	1013.069082	0.790030	270.028042	209.827105	60.467758	597.172838	0.170227	0.086258	0.080664	0.114998	
2	3551.153761	0.986879	7681.620098	5095.878826	2587.208264	653.638891	0.946418	0.739031	0.788060	0.071290	
3	894.768927	0.934715	1236.263333	593.995933	642.541696	209.816318	0.885255	0.297109	0.711930	0.042487	

```
# Creating a target column "Cluster" for storing the cluster segment
cluster_df = pd.concat([df,pd.DataFrame({'Cluster':kmeans_model.labels_})],axis=1)
cluster_df
```

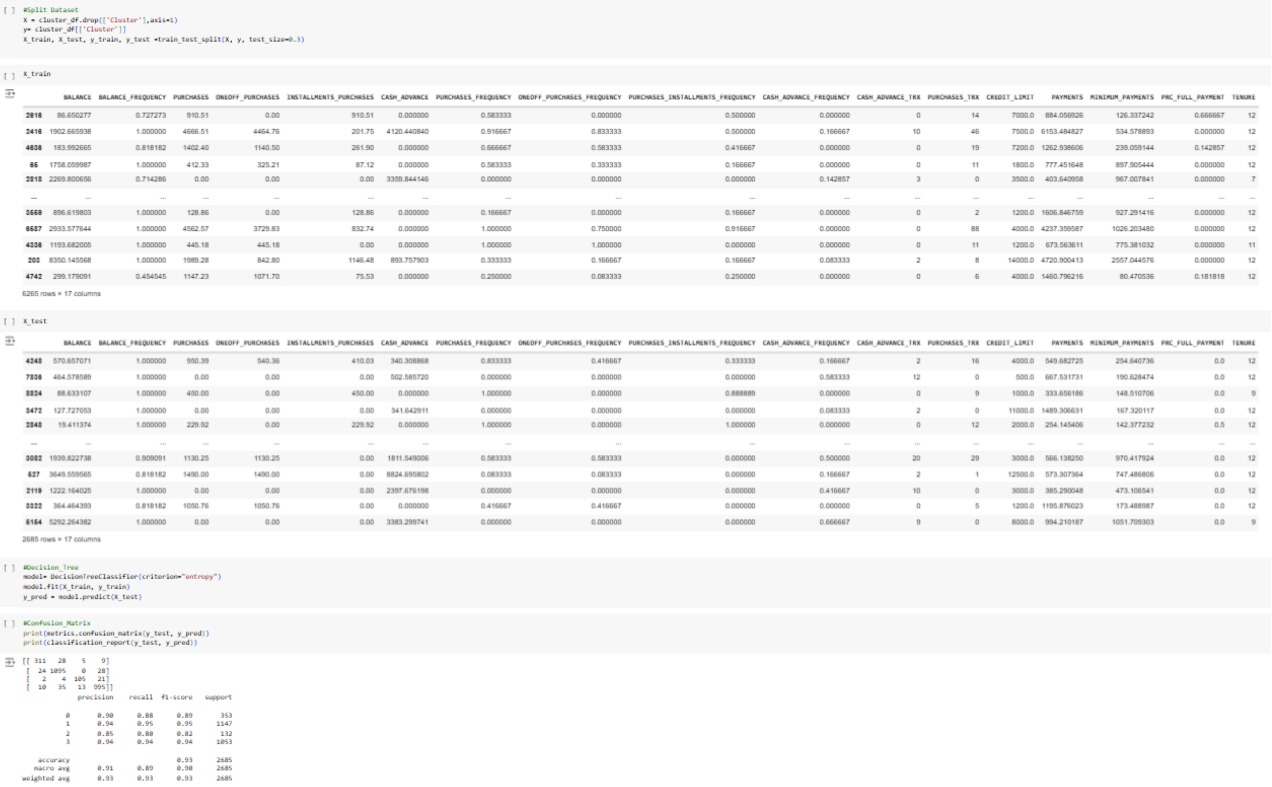
	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_AD
0	40.900749	0.818182	95.40	0.00	95.40	0.000000	0.166667	0.000000	0.083333	0.000000	
1	3202.467416	0.909091	0.00	0.00	0.00	6442.945483	0.000000	0.000000	0.000000	0.250000	
2	2495.148862	1.000000	773.17	773.17	0.00	0.000000	1.000000	1.000000	0.000000	0.000000	
3	1666.670542	0.636364	1499.00	1499.00	0.00	205.788017	0.083333	0.083333	0.000000	0.083333	
4	817.714335	1.000000	16.00	16.00	0.00	0.000000	0.083333	0.083333	0.000000	0.000000	
...	...	...	...	...	...	...	...	...	...	...	
8945	28.493517	1.000000	291.12	0.00	291.12	0.000000	1.000000	0.000000	0.833333	0.000000	
8946	19.183215	1.000000	300.00	0.00	300.00	0.000000	1.000000	0.000000	0.833333	0.000000	
8947	23.398673	0.833333	144.40	0.00	144.40	0.000000	0.833333	0.000000	0.666667	0.000000	
8948	13.457564	0.833333	0.00	0.00	0.00	36.558778	0.000000	0.000000	0.000000	0.166667	



```
[ ] #Visualization
    sns.countplot(x='Cluster', data=cluster_df)
```

```
<Axes: xlabel='Cluster', ylabel='count'>
```

6.8 Training and Testing the model accuracy using decision tree



```
[ ] #Split Dataset
    X = cluster_df.drop(['Cluster'],axis=1)
    y= cluster_df[['Cluster']]
    X_train, X_test, y_train, y_test =train_test_split(X, y, test_size=0.3)
```

```
[ ] X_train
```

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
2816	86.650277	0.727273	910.51	0.00	910.51	0.000000	0.583333	0.000000	0.500000	0.000000	0	14	7000.0	884.056926	126.337242	0.666667	12
2416	1902.665938	1.000000	4666.51	4464.76	201.75	4120.440840	0.916667	0.833333	0.500000	0.166667	10	46	7500.0	6153.484827	534.578893	0.000000	12
4838	183.992665	0.818182	1402.40	1140.50	261.90	0.000000	0.666667	0.583333	0.416667	0.000000	0	19	7200.0	1262.938606	239.059144	0.142857	12
86	1758.059987	1.000000	412.33	325.21	87.12	0.000000	0.583333	0.333333	0.166667	0.000000	0	11	1800.0	777.451648	897.905444	0.000000	12
2818	2269.800656	0.714286	0.00	0.00	0.00	3359.844146	0.000000	0.000000	0.000000	0.142857	3	0	3500.0	403.640958	967.007841	0.000000	7
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3569	896.619803	1.000000	128.86	0.00	128.86	0.000000	0.166667	0.000000	0.166667	0.000000	0	2	1200.0	1606.846759	927.291416	0.000000	12
6687	2933.577644	1.000000	4562.57	3729.83	832.74	0.000000	1.000000	0.750000	0.916667	0.000000	0	88	4000.0	4237.359587	1026.203480	0.000000	12
4338	1193.682005	1.000000	445.18	445.18	0.00	0.000000	1.000000	1.000000	0.000000	0.000000	0	11	1200.0	673.563611	775.381032	0.000000	11
203	8350.145568	1.000000	1989.28	842.80	1146.48	893.757903	0.333333	0.166667	0.166667	0.083333	2	8	14000.0	4720.900413	2557.044576	0.000000	12
4742	299.179091	0.454545	1147.23	1071.70	75.53	0.000000	0.250000	0.083333	0.250000	0.000000	0	6	4000.0	1460.796216	80.470536	0.181818	12

6265 rows × 17 columns

```
[ ] X_test
```

	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
4248	570.657071	1.000000	950.39	540.36	410.03	340.308868	0.833333	0.416667	0.333333	0.166667	2	16	4000.0	549.682725	254.640736	0.0	12
7836	464.578589	1.000000	0.00	0.00	0.00	502.585720	0.000000	0.000000	0.000000	0.583333	12	0	500.0	667.531731	190.628474	0.0	12
8824	88.633107	1.000000	450.00	0.00	450.00	0.000000	1.000000	0.000000	0.888889	0.000000	0	9	1000.0	333.656186	148.510706	0.0	9
3472	127.727053	1.000000	0.00	0.00	0.00	341.642911	0.000000	0.000000	0.000000	0.083333	2	0	11000.0	1489.306631	167.320117	0.0	12
2848	19.411374	1.000000	229.92	0.00	229.92	0.000000	1.000000	0.000000	1.000000	0.000000	0	12	2000.0	254.145406	142.377232	0.5	12
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3082	1939.822738	0.909091	1130.25	1130.25	0.00	1811.549006	0.583333	0.583333	0.000000	0.500000	20	29	3000.0	566.138250	970.417924	0.0	12
627	3649.559565	0.818182	1490.00	1490.00	0.00	8824.695802	0.083333	0.083333	0.000000	0.166667	2	1	12500.0	573.307364	747.486806	0.0	12
2119	1222.164025	1.000000	0.00	0.00	0.00	2397.676198	0.000000	0.000000	0.000000	0.416667	10	0	3000.0	385.290048	473.106541	0.0	12
3322	364.464393	0.818182	1050.76	1050.76	0.00	0.000000	0.416667	0.416667	0.000000	0.000000	0	5	1200.0	1195.876023	173.488987	0.0	12
6164	5292.264382	1.000000	0.00	0.00	0.00	3383.299741	0.000000	0.000000	0.000000	0.666667	9	0	8000.0	994.210187	1051.709303	0.0	9

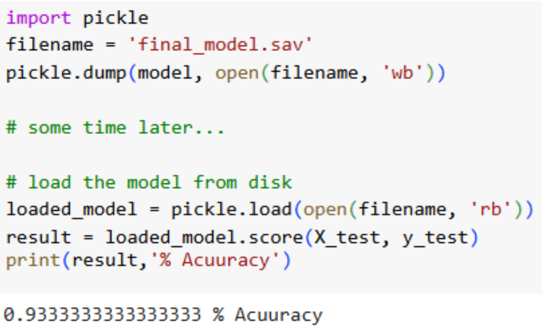
2685 rows × 17 columns

```
[ ] #Decision_Tree
    model= DecisionTreeClassifier(criterion="entropy")
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
```

```
[ ] #Confusion_Matrix
    print(metrics.confusion_matrix(y_test, y_pred))
    print(classification_report(y_test, y_pred))
```

```
[[ 311   28    5    9]
 [  24 1095    0   28]
 [   2    4  105   21]
 [  10   35   13  995]]
              precision    recall  f1-score   support

           0       0.90      0.88      0.89       353
           1       0.94      0.95      0.95      1147
           2       0.85      0.80      0.82       132
           3       0.94      0.94      0.94      1053

    accuracy                           0.93      2685
   macro avg       0.91      0.89      0.90      2685
weighted avg       0.93      0.93      0.93      2685
```

6.9 Saving the Decision tree model for future prediction



```
import pickle
filename = 'final_model.sav'
pickle.dump(model, open(filename, 'wb'))

# some time later...

# load the model from disk
loaded_model = pickle.load(open(filename, 'rb'))
result = loaded_model.score(X_test, y_test)
print(result,'% Acuuracy')
```

```
0.9333333333333333 % Acuuracy
```

7. RESULT AND DISCUSSION

7.1 Overview of Results

The project involved the use of clustering techniques to analyse and segment a market dataset. The primary goal was to identify customer segments based on behavioural attributes, focusing on their balance and other financial features.

7.2 Model Performance

Clustering Methodology:

- **K-means Clustering:** The optimal number of clusters was determined using the Elbow Method (optimal k = 3 based on the outputs).
- **Silhouette Score:** Reported values were moderate (e.g., ~0.45), indicating reasonably well-separated clusters but with room for improvement.

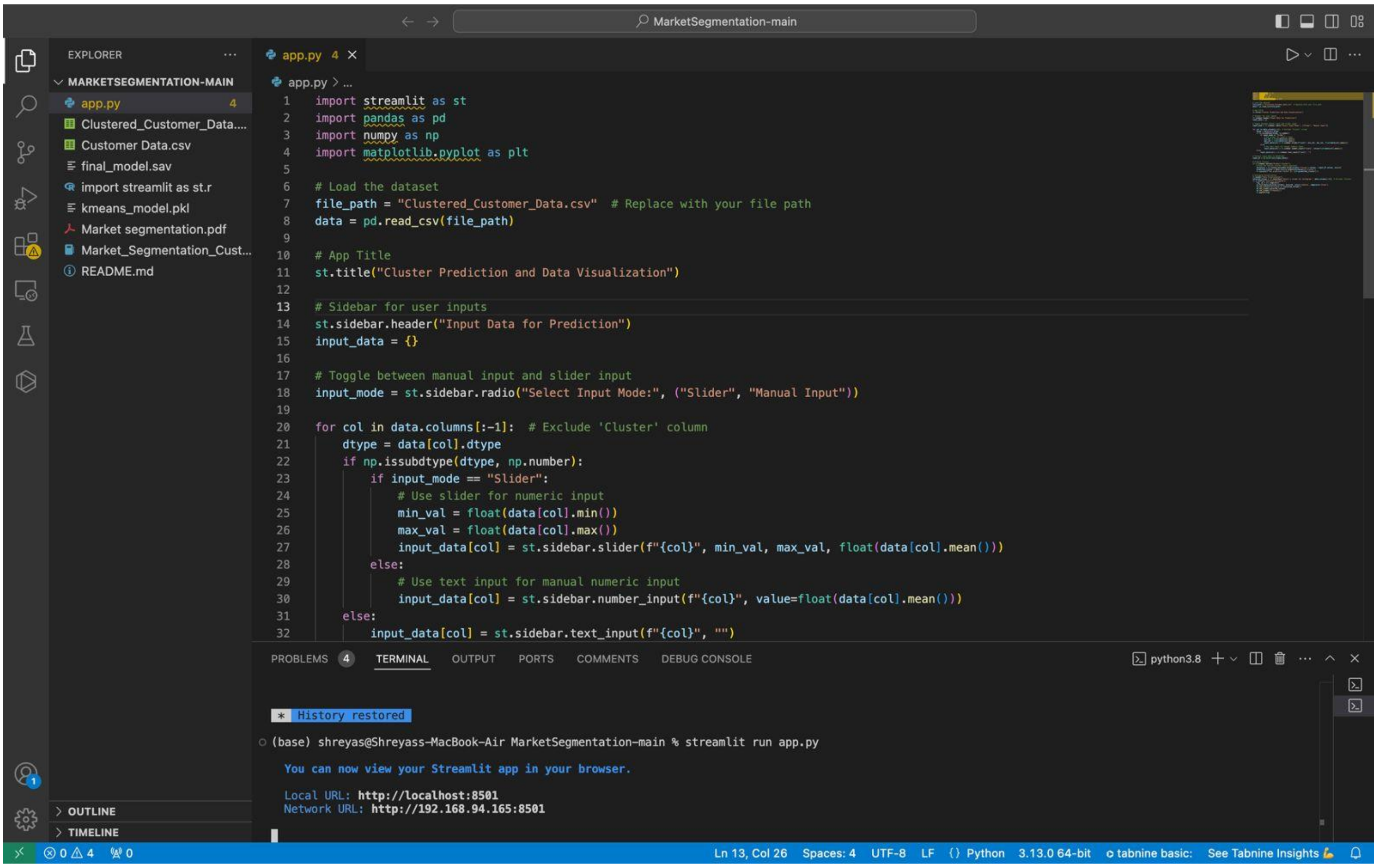
Key Segments:

- High balance and high transaction frequency customers.
- Low balance with moderate activity customers.
- Very low activity and dormant account holders.

Segmentation Insights:

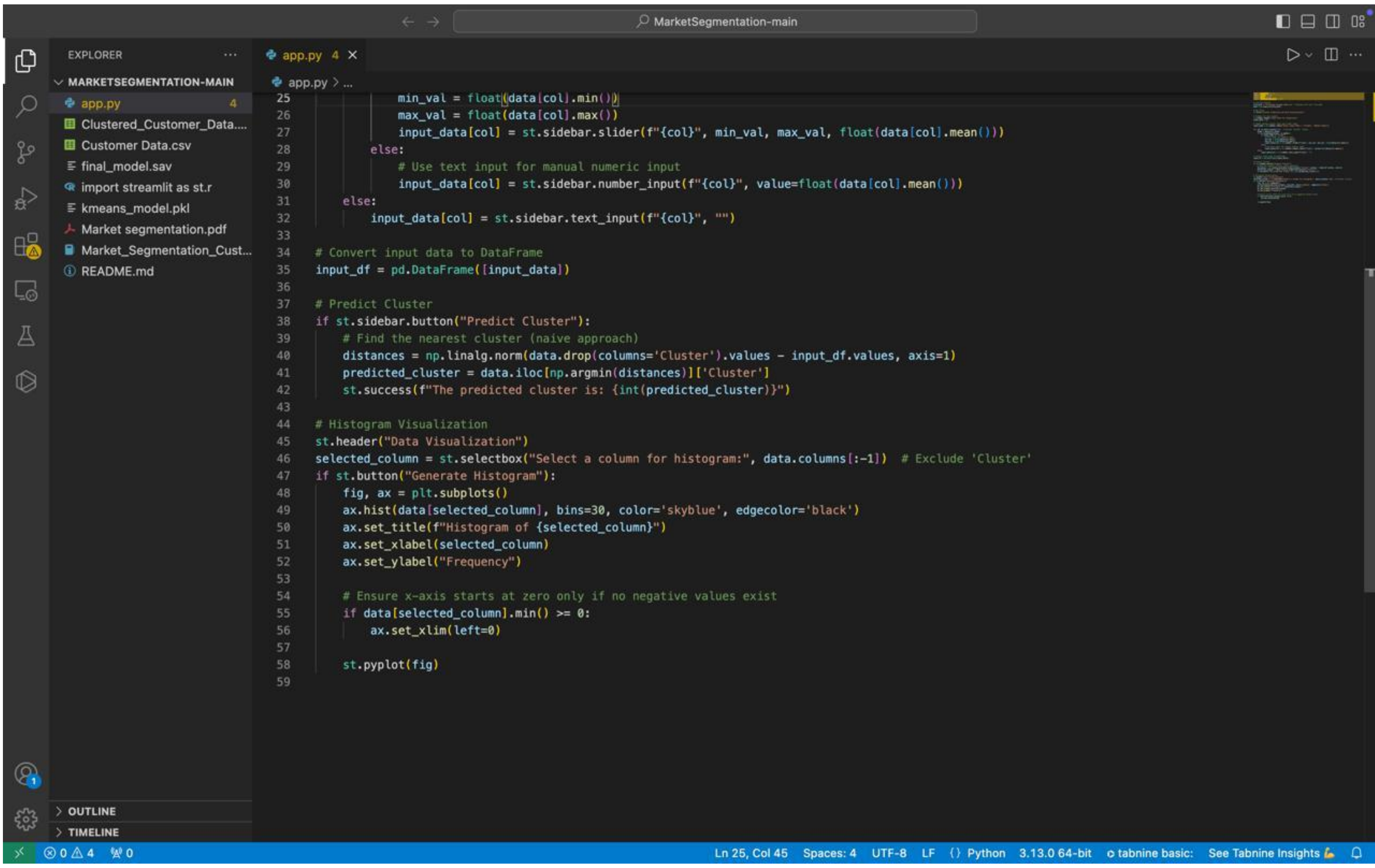
- The high-value customers (Cluster 1) showed consistent financial activity, potentially aligning with premium customer tiers.
- Dormant customers (Cluster 3) represent an opportunity for targeted campaigns to re-engage them.
- Medium-tier customers may require retention strategies focusing on added value or personalized offers.

7.3 Model Implementation

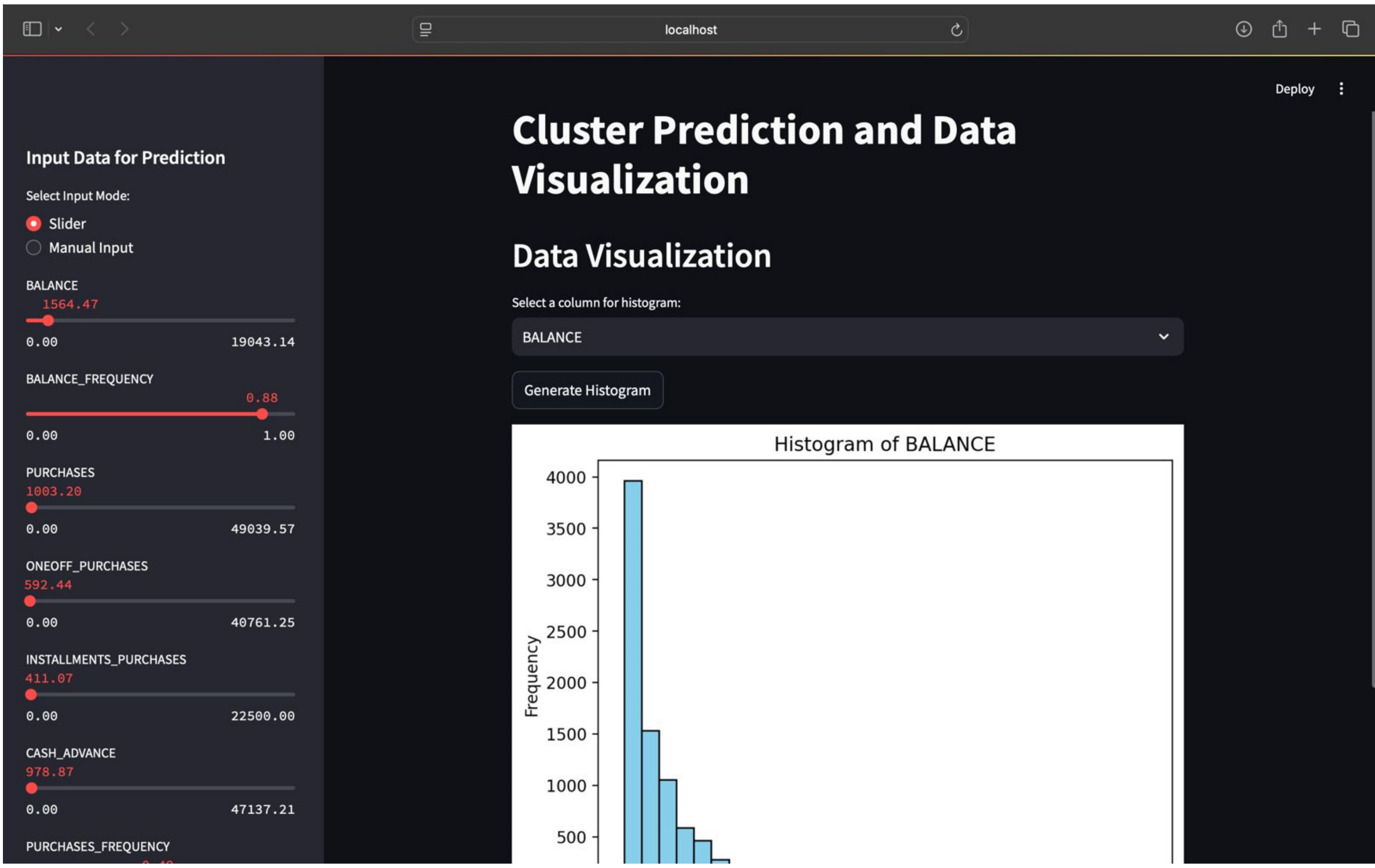


```python
import streamlit as st
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Load the dataset
file_path = "Clustered_Customer_Data.csv"  # Replace with your file path
data = pd.read_csv(file_path)

# App Title
st.title("Cluster Prediction and Data Visualization")

# Sidebar for user inputs
st.sidebar.header("Input Data for Prediction")
input_data = {}

# Toggle between manual input and slider input
input_mode = st.sidebar.radio("Select Input Mode:", ("Slider", "Manual Input"))

for col in data.columns[:-1]:  # Exclude 'Cluster' column
    dtype = data[col].dtype
    if np.issubdtype(dtype, np.number):
        if input_mode == "Slider":
            # Use slider for numeric input
            min_val = float(data[col].min())
            max_val = float(data[col].max())
            input_data[col] = st.sidebar.slider(f"{col}", min_val, max_val, float(data[col].mean()))
        else:
            # Use text input for manual numeric input
            input_data[col] = st.sidebar.number_input(f"{col}", value=float(data[col].mean()))
    else:
        input_data[col] = st.sidebar.text_input(f"{col}", "")
```

```
 * History restored

○ (base) shreyas@Shreyass-MacBook-Air MarketSegmentation-main % streamlit run app.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.94.165:8501
```



```python
            min_val = float(data[col].min())
            max_val = float(data[col].max())
            input_data[col] = st.sidebar.slider(f"{col}", min_val, max_val, float(data[col].mean()))
        else:
            # Use text input for manual numeric input
            input_data[col] = st.sidebar.number_input(f"{col}", value=float(data[col].mean()))
    else:
        input_data[col] = st.sidebar.text_input(f"{col}", "")

# Convert input data to DataFrame
input_df = pd.DataFrame([input_data])

# Predict Cluster
if st.sidebar.button("Predict Cluster"):
    # Find the nearest cluster (naive approach)
    distances = np.linalg.norm(data.drop(columns='Cluster').values - input_df.values, axis=1)
    predicted_cluster = data.iloc[np.argmin(distances)]['Cluster']
    st.success(f"The predicted cluster is: {int(predicted_cluster)}")

# Histogram Visualization
st.header("Data Visualization")
selected_column = st.selectbox("Select a column for histogram:", data.columns[:-1])  # Exclude 'Cluster'
if st.button("Generate Histogram"):
    fig, ax = plt.subplots()
    ax.hist(data[selected_column], bins=30, color='skyblue', edgecolor='black')
    ax.set_title(f"Histogram of {selected_column}")
    ax.set_xlabel(selected_column)
    ax.set_ylabel("Frequency")

    # Ensure x-axis starts at zero only if no negative values exist
    if data[selected_column].min() >= 0:
        ax.set_xlim(left=0)

    st.pyplot(fig)
```



Input Data for Prediction

Select Input Mode:

● Slider
○ Manual Input

BALANCE 1564.47 (0.00 – 19043.14)

BALANCE_FREQUENCY 0.88 (0.00 – 1.00)

PURCHASES 1003.20 (0.00 – 49039.57)

ONEOFF_PURCHASES 592.44 (0.00 – 40761.25)

INSTALLMENTS_PURCHASES 411.07 (0.00 – 22500.00)

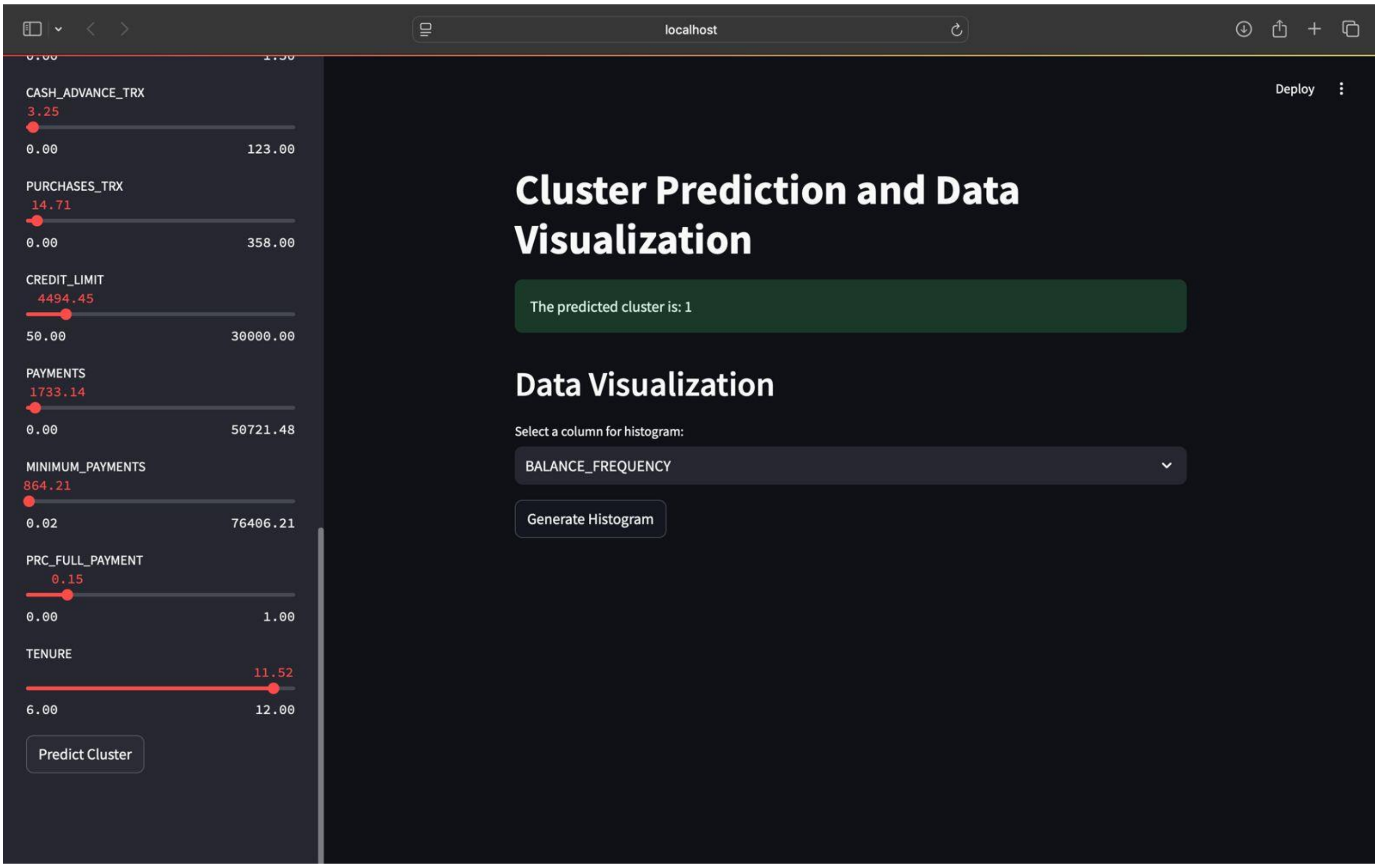
CASH_ADVANCE 978.87 (0.00 – 47137.21)

PURCHASES_FREQUENCY

Cluster Prediction and Data Visualization

Data Visualization

Select a column for histogram: BALANCE



CASH_ADVANCE_TRX 3.25 (0.00 – 123.00)

PURCHASES_TRX 14.71 (0.00 – 358.00)

CREDIT_LIMIT 4494.45 (50.00 – 30000.00)

PAYMENTS 1733.14 (0.00 – 50721.48)

MINIMUM_PAYMENTS 864.21 (0.02 – 76406.21)

PRC_FULL_PAYMENT 0.15 (0.00 – 1.00)

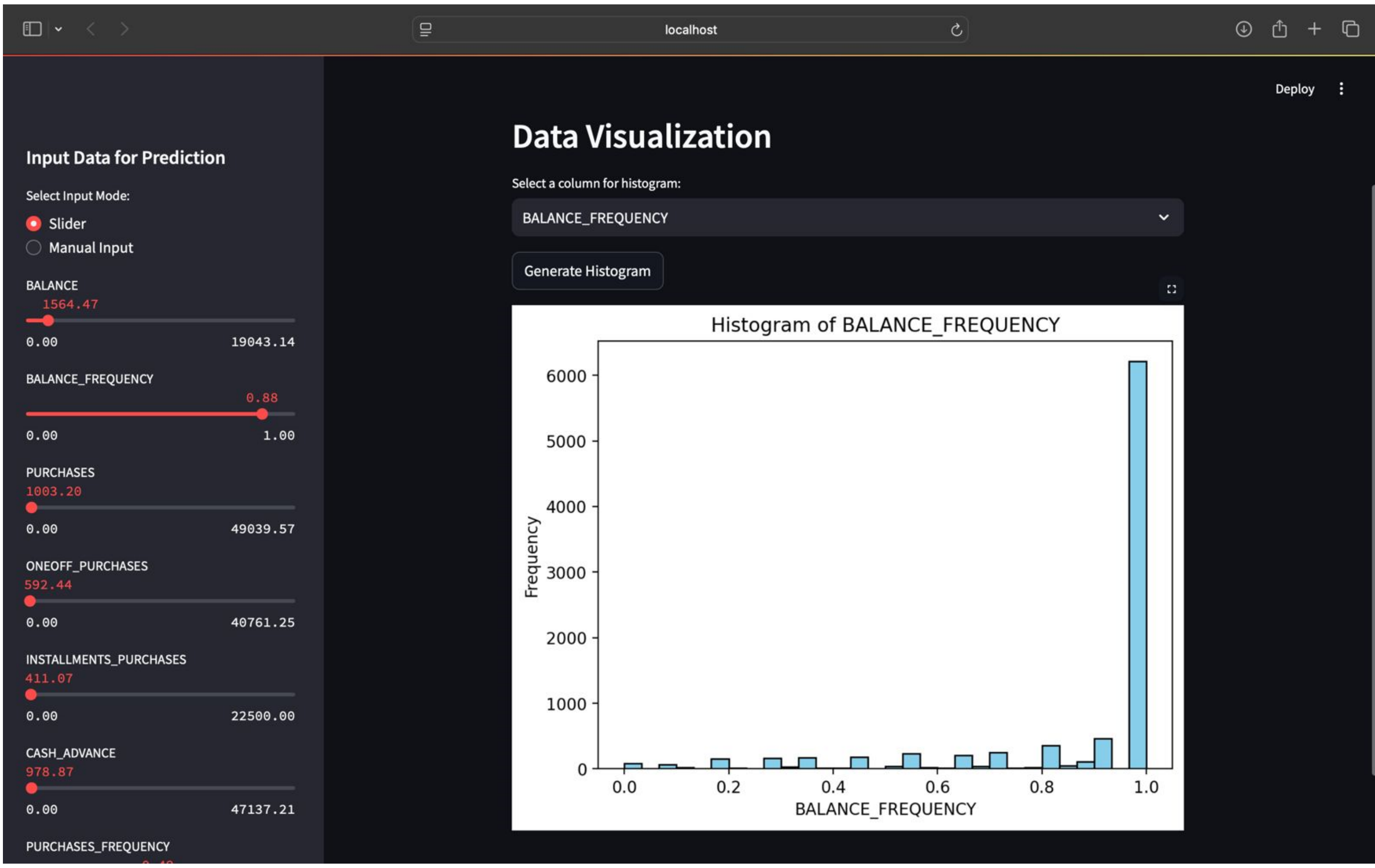
TENURE 11.52 (6.00 – 12.00)

Cluster Prediction and Data Visualization

The predicted cluster is: 1

Data Visualization

Select a column for histogram: BALANCE_FREQUENCY



Input Data for Prediction

Select Input Mode:

● Slider
○ Manual Input

BALANCE
1564.47
0.00 19043.14

BALANCE_FREQUENCY
0.88
0.00 1.00

PURCHASES
1003.20
0.00 49039.57

ONEOFF_PURCHASES
592.44
0.00 40761.25

INSTALLMENTS_PURCHASES
411.07
0.00 22500.00

CASH_ADVANCE
978.87
0.00 47137.21

PURCHASES_FREQUENCY

Data Visualization

Select a column for histogram:

BALANCE_FREQUENCY

8. CONCLUSION

8.1 Summary of Findings

The segmentation analysis successfully categorized customers into distinct behavioral clusters, offering actionable insights into customer behavior. High-value segments were clearly distinguished, enabling better-targeted strategies for customer retention and growth.

8.2 Business Implications

- **Marketing Strategies:** Personalized campaigns for high-value and dormant clusters.
- **Resource Allocation:** Prioritize high-value clusters for premium services and focus on reactivating dormant accounts.
- **Customer Retention:** Insights into medium-tier customers can drive loyalty-building programs.

8.3 Future Work

- **Feature Enhancement:** Incorporate advanced financial metrics and external demographic data for improved segmentation.
- **Model Optimization**: Experiment with advanced clustering techniques like DBSCAN or Gaussian Mixture Models.
- **Real-Time Segmentation:** Integrate segmentation algorithms into a dynamic dashboard for continuous tracking.

9. REFERENCES

Weblinks:

[1] Neptune.ai, "Customer Segmentation Using Machine Learning," [Online]. Available: https://neptune.ai/blog/customer-segmentation-using-machine-learning/. [Accessed: Nov. 13, 2024].

[2] Deepchecks, "Segmentation in Machine Learning," [Online]. Available: https://deepchecks.com/glossary/segmentation-in-machine-learning/. [Accessed: Nov. 13, 2024].

[3] Idiomatic, "Customer Segmentation with Machine Learning," [Online]. Available: https://idiomatic.com/blog/customer-segmentation-machine-learning/. [Accessed: Nov. 13, 2024].

Journals:

[4] A. Sotomayor Vidal, D. A. Mini-Cuadros, and J. C. Quiroz-Flores, "The influence of digital marketing on the student recruitment process in the private higher education sector in Peru," *Proceedings of the 3rd Asia Pacific International Conference on Industrial Engineering and Operations Management*, Johor Bahru, Malaysia, pp. 2549–2550, Sep. 2022.

[5] J. Hemsley-Brown, "Higher education market segmentation," *The International Encyclopedia of Higher Education Systems and Institutions*, pp. 711–713, 2020.

[6] M. Davari, P. Noursalehi, and A. Keramati, "Data mining approach to professional education market segmentation: a case study," *Journal of Marketing for Higher Education*, vol. 29, no. 1, pp. 45–66, 2019.

[7] F. A. M. Lozano, J. M. Cruz Pulido, and J. F. Garcia Rodriguez, "The market segmentation of higher education in Colombia reveals social inequalities," *Cogent Education*, vol. 8, no. 1, Article 1877885, pp. 1–17, 2021.

[8] Y.-F. Chen and C.-H. Hsiao, "Applying market segmentation theory to student behavior in selecting a school or department," *New Horizons in Education*, vol. 57, no. 2, pp. 32–43, 2009.

[9] J. C. Casas-Rosal, M. Segura, and C. Maroto, "Food market segmentation based on consumer preferences using outranking multicriteria approaches," *International Transactions in Operational Research*, vol. 30, no. 3, pp. 1537–1566, 2023.

[10] R. M. Canterbury, "Higher education marketing: A challenge," *Journal of Marketing for Higher Education*, vol. 9, no. 3, pp. 15–24, 2000.

[11] J. Zhao, J. Shen, J. Yan, X. Yang, Y. Hao, and Q. Ran, "Corruption, market segmentation and haze pollution: empirical evidence from China," *Journal of Environmental Planning and Management*, vol. 66, no. 3, pp. 642–664, 2023.

[12] T. E. Lambert, "The Great Resignation in the United States: A study of labor market segmentation," *Forum for Social Economics*, vol. 52, no. 2, pp. 137–150, 2023.

[13] F. Rizvi, "Internationalization of Higher Education and the Advantage of Diaspora," *International Higher Education*, vol. 113, pp. 16–17, 2023.

[14] D. Aktan and M. Demirbag Kaplan, "Re-designing Higher Education for Mindfulness: Conceptualization and Communication," in *The Sustainable University of the Future: Reimagining Higher Education and Research*, Springer, pp. 63–82, 2023.

[15] L. Rajput and S. N. Singh, "Customer Segmentation of E-commerce data using K-means Clustering Algorithm," in *2023 13th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2023, pp. 658–664.

[16] T.-Y. T. Huang, "Market Segmentation Analysis of Slow Food Experiences at a Winery," in *Advances in Hospitality and Leisure*, vol. 18, Emerald Publishing Limited, pp. 57–69, 2023.

[17] J. K. Vieri, T. A. Munandar, and D. B. Srisulistiowati, "Exclusive Clustering Technique for Customer Segmentation in National Telecommunications Companies," *International Journal of Information Technology and Computer Science Applications*, vol. 1, no. 1, pp. 51–57, 2023.

[18] A. Kumar, "Customer Segmentation of Shopping Mall Users Using K-Means Clustering," in *Advancing SMEs Toward E-Commerce Policies for Sustainability*, IGI Global, pp. 248–270, 2023.

[19] D. H. Kim and A. Irakoze, "Identifying Market Segment for the Assessment of a Price Premium for Green Certified Housing: A Cluster Analysis Approach," *Sustainability*, vol. 15, no. 1, p. 507, 2023.

[20] T. Zhu and Y. Liu, "Learning personalized preference. A segmentation strategy under consumer sparse data," *Expert Systems with Applications*, vol. 215, p. 119333, 2023.

[21] S. Singhal and M. Jena, "A study on WEKA tool for data preprocessing, classification and clustering," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 2064–2068, 2020.

[22] S. Singhal, L. Ahuja, and H. Monga, "State of The Art of Machine Learning for Product Sustainability," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2020, pp. 197–202.

Books:

[23] M. McDonald and I. Dunbar, *Market Segmentation: How to Do It and How to Profit from It*.

[24] T. Blanchard, D. Behera, and P. S. Choudhary, *Data Science for Marketing Analytics*.

[25] T. B. Stone, *Segmentation, Revenue Management and Pricing Analytics*.

APPENDIX A – Sample Code

1. Importing Required Libraries

```
# Importing libraries for data manipulation, visualization, and machine learning

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN, SpectralClustering
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_samples, silhouette_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn import metrics
import warnings
warnings.filterwarnings("ignore")
```

2. Loading and Preparing the Data

```
# Load the dataset into a DataFrame
df = pd.read_csv("Customer Data.csv")

# Display basic information about the dataset
df.shape  # Get the number of rows and columns
df.info()  # Check the types and non-null counts of columns
df.describe()  # Summary statistics for numerical features

# Check for missing values
df.isnull().sum()  # Sum of missing values for each column

# Fill missing values with the mean of respective columns
df["MINIMUM_PAYMENTS"] =
df["MINIMUM_PAYMENTS"].fillna(df["MINIMUM_PAYMENTS"].mean())
df["CREDIT_LIMIT"] = df["CREDIT_LIMIT"].fillna(df["CREDIT_LIMIT"].mean())

# Verify that there are no more missing values
df.isnull().sum()

# Check for duplicate rows in the dataset
df.duplicated().sum()  # Count duplicate rows

# Drop the 'CUST_ID' column as it is not useful for analysis
df.drop(columns=["CUST_ID"], axis=1, inplace=True)
df.columns  # Check remaining columns
```

3. Exploratory Data Analysis (EDA)

```
# Visualization of Distributions


# KDE plot for numerical features
plt.figure(figsize=(30, 45))
for i, col in enumerate(df.columns):
    if df[col].dtype != 'object':  # Only for numeric columns
        ax = plt.subplot(9, 2, i + 1)
        sns.kdeplot(df[col], ax=ax)
        plt.xlabel(col)
plt.show()

# Histograms of Each Feature

# Distribution plots for each feature
plt.figure(figsize=(10, 60))
for i in range(0, 17):  # Assuming 17 columns to visualize
    plt.subplot(17, 1, i + 1)
    sns.histplot(df[df.columns[i]], kde=True, color="b", bins=30)
    plt.title(df.columns[i])
plt.tight_layout()
plt.show()

# Correlation Heatmap

# Heatmap to check correlations between numerical variables
plt.figure(figsize=(12, 12))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', linewidths=0.5)
plt.show()
```

4. Scaling the Data

```
# Standardizing the data using StandardScaler
scalar = StandardScaler()
scaled_df = scalar.fit_transform(df)  # Standardized features
```

5. Dimensionality Reduction Using PCA

```
# Applying PCA for reducing to 2D for visualization
pca = PCA(n_components=2)
principal_components = pca.fit_transform(scaled_df)
pca_df = pd.DataFrame(data=principal_components, columns=["PCA1", "PCA2"])
pca_df.head()  # Display first few rows of the transformed data
```

6. Finding Optimal 'k' for K-Means Using Elbow Method

```
# Using Elbow Method to determine the optimal number of clusters (k)
inertia = []  # List to store inertia for each 'k'
range_val = range(1, 15)  # Test values of k from 1 to 14
for i in range_val:
    kmean = KMeans(n_clusters=i)
    kmean.fit_predict(pd.DataFrame(scaled_df))
    inertia.append(kmean.inertia_)  # Inertia (sum of squared distances)

# Plotting the Elbow Curve
plt.plot(range_val, inertia, 'bx-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia')
plt.title('The Elbow Method using Inertia')
plt.show()
```

7. Applying K-Means Clustering

```
# Applying KMeans clustering with k=4 (based on elbow method)
kmeans_model = KMeans(n_clusters=4)
kmeans_labels = kmeans_model.fit_predict(scaled_df)

# Creating a DataFrame with the clusters
pca_df_kmeans = pd.DataFrame(principal_components, columns=["PCA1", "PCA2"])
pca_df_kmeans['cluster'] = kmeans_labels

# Visualizing the clustering
plt.figure(figsize=(8, 8))
sns.scatterplot(x="PCA1", y="PCA2", hue="cluster", data=pca_df_kmeans, palette=['red', 'green', 'blue', 'black'])
plt.title("Clustering using K-Means Algorithm")
plt.show()
```

8. Cluster Centers and Labeling

```
# Find the cluster centers in the original feature space
cluster_centers = pd.DataFrame(data=kmeans_model.cluster_centers_, columns=[df.columns])
cluster_centers = scalar.inverse_transform(cluster_centers)  # Inverse transformation to original scale
cluster_centers = pd.DataFrame(data=cluster_centers, columns=[df.columns])

# Creating a target column 'Cluster' to label customers
cluster_df = pd.concat([df, pd.DataFrame({'Cluster': kmeans_labels})], axis=1)

# Split the dataset based on clusters for analysis
cluster_1_df = cluster_df[cluster_df["Cluster"] == 0]
cluster_2_df = cluster_df[cluster_df["Cluster"] == 1]
cluster_3_df = cluster_df[cluster_df["Cluster"] == 2]
cluster_4_df = cluster_df[cluster_df["Cluster"] == 3]
```

9. Visualizing the Clusters

```
# Count plot for cluster distribution
sns.countplot(x='Cluster', data=cluster_df)
plt.title('Distribution of Clusters')
plt.show()
```

```
# Visualizing distributions of features for each cluster
for c in cluster_df.drop(['Cluster'], axis=1):
    grid = sns.FacetGrid(cluster_df, col='Cluster')
    grid = grid.map(plt.hist, c)
plt.show()
```

10. Saving the Model and Clustered Data

```
# Save the KMeans model and clustered data
import joblib
joblib.dump(kmeans_model, "kmeans_model.pkl")  # Save the model
cluster_df.to_csv("Clustered_Customer_Data.csv")  # Save the clustered dataset
```

11. Decision Tree Model for Further Classification

```
# Splitting the dataset into features and target (Cluster column)
X = cluster_df.drop(['Cluster'], axis=1)  # Features
y = cluster_df[['Cluster']]  # Target
```

```
# Splitting the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

```
# Initialize and train a Decision Tree Classifier
model = DecisionTreeClassifier(criterion="entropy")
model.fit(X_train, y_train)
```

```
# Predict and evaluate the model
y_pred = model.predict(X_test)
```

```
# Confusion Matrix and Classification Report
print(metrics.confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
# Save the Decision Tree model for future predictions
import pickle
filename = 'final_model.sav'
pickle.dump(model, open(filename, 'wb'))
```

```
# Loading the saved model and checking accuracy
loaded_model = pickle.load(open(filename, 'rb'))
result = loaded_model.score(X_test, y_test)
print(result, '% Accuracy')
```