CUSTOMER SEGMENTATION ANALYSIS USING K-MEANS CLUSTERING

A PROJECT REPORT

Submitted by

KRITHIK BM
(22138010)

AKSHAYA PL
(22138005)

LOKESHWARI M
(22138023)

Under the guidance of

Ms. J. Swarnalakshmi
Assistant Professor
CSE, HITS

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY
in
COMPUTER SCIENCE AND ENGINEERING

HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE
CHENNAI - 603 103
APRIL 2025

HINDUSTAN
INSTITUTE OF TECHNOLOGY & SCIENCE
(DEEMED TO BE UNIVERSITY)

BONAFIDE CERTIFICATE

Certified that this project report **Customer Segmentation Analysis using K-Means Clustering** is the bonafide work of **Krithik BM (22138010), Akshaya PL (22138005) and Lokeshwari M (22138023)** who carried out the project work under my supervision during the academic year 2**024-2025**.

Ms. J. Swarnalakshmi,
SUPERVISOR
ASSISTANT PROFESSOR
CSE, HITS

Dr.J.Thangakumar,
HEAD OF DEPARTMENT CSE
PROFESSOR

INTERNAL EXAMINER
Name:_____
Designation: _____

EXTERNAL EXAMINER
Name: _____
Designation:_____

Project Viva - voce conducted on _____

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGEMENT

First and foremost, we would like to thank **ALMIGHTY** who has provided us the strength to do justice to our work and contribute our best to it.

We wish to express our deep sense of gratitude from the bottom of our heart to our guide **Ms. J. Swarnalakshmi , Assistant Professor**, **Computer Science and Engineering,** for her motivating discussions, overwhelming suggestions, ingenious encouragement, invaluable supervision, and exemplary guidance throughout this project work.

We would like to extend our heartfelt gratitude to **Dr. J. Thangakumar, Ph.D., Professor & Head, Department of Computer Science and Engineering** for his valuable suggestions and support in successfully completing the project.

We wish to thank our Project Co-ordinator for keeping our project in the right track. We would like to thank all the teaching, technical and non-technical staff of Department of Computer Science and Engineering for their courteous assistance.

We thank the management of **HINDUSTAN INSTITUTE OF TECHNOLOGY AND SCIENCE** for providing us the necessary facilities and support required for the successful completion of the project.

As a final word, we would like to thank each and every individual who have been a source of support and encouragement and helped us to achieve our goal and complete our project work successfully.

DEDICATION

This project is dedicated to my beloved parents, for their love,

endless support, encouragement and sacrifices.

ABSTRACT

This project focuses on one of the core applications of unsupervised machine learning—Customer Segmentation—by implementing the K-Means Clustering algorithm. The primary goal is to segment customers based on their behavior, spending patterns, and demographic attributes to uncover distinct customer groups that businesses can target more effectively. The dataset used is the popular Mall Customer Segmentation data, which includes features like age, gender, annual income, and spending score. By applying exploratory data analysis, the project first examines key variables and their distributions, followed by the use of the Elbow Method to determine the optimal number of clusters. K-Means clustering is then used to classify customers into meaningful groups based on their purchasing behavior. Through this segmentation, businesses can gain valuable insights into which customers are high spenders, potential targets for promotions, or cost-conscious consumers. For example, customers with high income and high spending scores are identified as ideal targets for premium services, while those with low income but high spending behavior might be nurtured through loyalty programs. The project visualizes clusters using scatter plots, enabling intuitive understanding of the segmented groups. Additionally, cluster characteristics are interpreted to support marketing strategy development and resource optimization. The implementation is done using Python with libraries such as pandas, matplotlib, seaborn, and scikit-learn, ensuring a robust and reproducible data science workflow. The platform is designed for scalability and can be extended to support additional features such as real-time segmentation, integration with customer databases, or advanced clustering techniques like DBSCAN or Hierarchical Clustering. Overall, this project demonstrates how customer segmentation using K-Means can help businesses personalize services, improve marketing effectiveness, and enhance customer satisfaction while minimizing operationalrisks.

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

In today's data-driven business landscape, understanding customer behavior is a vital element in delivering personalized experiences and optimizing marketing strategies. As organizations continue to collect vast amounts of customer data, the ability to analyze and segment this data has become a cornerstone of customer relationship management. This project focuses on implementing a machine learning approach for Customer Segmentation using the K-Means Clustering algorithm, an unsupervised learning technique that classifies customers into distinct groups based on shared attributes such as age, gender, annual income, and spending score. By analyzing these patterns, businesses can better understand their clientele, tailor services to individual needs, and drive profitability through targeted marketing. Customer segmentation is not only a powerful analytical tool but also a strategic necessity for companies operating in competitive markets. This project applies K-Means Clustering to divide customers into meaningful segments, thereby enabling decision-makers to identify high-value customers, potential brand advocates, or individuals requiring engagement strategies. The project uses the Mall Customer Segmentation Dataset, which provides essential demographic and behavioral features suitable for unsupervised clustering. Through exploratory data analysis and visualizations, the project reveals hidden structures in the data that can inform business strategy and customer outreach. Furthermore, the Elbow Method is used to determine the optimal number of clusters, ensuring that the segmentation is both statistically sound and practically useful. Each cluster is interpreted to reveal specific customer characteristics, which can then be mapped to personalized campaigns or product recommendations. The use of Python and libraries such as pandas, matplotlib, seaborn, and scikit-learn makes this solution replicable, scalable, and adaptable for a wide range of retail and service-based industries. Designed with extensibility in mind, the project can be further enhanced with advanced clustering techniques, integration into customer dashboards, or real-time data pipelines. By demonstrating the practical application of K-Means Clustering to real-world customer data, this project delivers a scalable and insightful tool for businesses aiming to enhance customer engagement, improve retention, and maximize lifetime value.

1.2 MOTIVATION FOR THE PROJECT

In the modern business environment, organizations are increasingly focused on understanding customer behavior to remain competitive and deliver personalized experiences. However, with the exponential growth of customer data, it becomes challenging to manually analyze and identify patterns that can lead to actionable business insights. This project is motivated by the need to leverage machine learning techniques to transform raw data into meaningful customer segments that can guide marketing, product development, and customer engagement strategies. Traditional one-size-fits-all approaches in customer outreach often fail to resonate with diverse consumer needs and preferences. By applying customer segmentation, businesses can group customers based on shared characteristics, enabling more targeted and efficient use of resources. The motivation behind this project lies in the ability of K-Means clustering to uncover hidden patterns within data without requiring labeled outcomes, making it ideal for real-world business scenarios where labeled data is scarce or unavailable. Segmenting customers by factors such as age, income, and spending behavior helps in identifying high-value customers, cost-conscious segments, or potential churn risks. Furthermore, this project is driven by the value of data-driven decision-making. Instead of relying on assumptions or intuition, companies can now base their strategies on empirical evidence derived from customer clusters. The insights gained through segmentation allow for more effective product positioning, dynamic pricing models, and personalized promotional offers, ultimately improving customer satisfaction and increasing return on investment. From a technical perspective, this project also serves as a learning opportunity to explore unsupervised learning techniques, data preprocessing, and evaluation methods like the Elbow Method. Implementing the solution using Python and visualization libraries like seaborn and matplotlib enhances the interpretability of results and supports effective business communication. In summary, the motivation for this project stems from both the business necessity of customer understanding and the technical curiosity to apply clustering algorithms to solve a practical problem. As customer-centric strategies become more vital than ever, this project contributes a scalable and adaptable solution to segment customers intelligently and enhance overall organizational performance.

1.3 OBJECTIVE

In the modern business environment, organizations are increasingly focused on understanding customer behavior to remain competitive and deliver personalized experiences. However, with the exponential growth of customer data, it becomes challenging to manually analyze and identify patterns that can lead to actionable business insights. This project is motivated by the need to leverage machine learning techniques to transform raw data into meaningful customer segments that can guide marketing, product development, and customer engagement strategies. Traditional one-size-fits-all approaches in customer outreach often fail to resonate with diverse consumer needs and preferences. By applying customer segmentation, businesses can group customers based on shared characteristics, enabling more targeted and efficient use of resources. The motivation behind this project lies in the ability of K-Means clustering to uncover hidden patterns within data without requiring labeled outcomes, making it ideal for real-world business scenarios where labeled data is scarce or unavailable. Segmenting customers by factors such as age, income, and spending behavior helps in identifying high-value customers, cost-conscious segments, or potential churn risks. Furthermore, this project is driven by the value of data-driven decision-making. Instead of relying on assumptions or intuition, companies can now base their strategies on empirical evidence derived from customer clusters. The insights gained through segmentation allow for more effective product positioning, dynamic pricing models, and personalized promotional offers, ultimately improving customer satisfaction and increasing return on investment. From a technical perspective, this project also serves as a learning opportunity to explore unsupervised learning techniques, data preprocessing, and evaluation methods like the Elbow Method. Implementing the solution using Python and visualization libraries like seaborn and matplotlib enhances the interpretability of results and supports effective business communication. In summary, the motivation for this project stems from both the business necessity of customer understanding and the technical curiosity to apply clustering algorithms to solve a practical problem. As customer-centric strategies become more vital than ever, this project contributes a scalable and adaptable solution to segment customers intelligently and enhance overall organizational performance

1.4 PROBLEM STATEMENT

In today's competitive market landscape, businesses collect vast amounts of customer data, yet many struggle to convert this data into actionable insights. One of the most pressing challenges is the inability to accurately identify and group customers based on their unique behaviors and preferences. Without clear segmentation, marketing efforts become generalized, customer engagement remains suboptimal, and valuable opportunities for personalization and retention are lost. Traditional segmentation methods, often based on basic demographics or intuition, fail to capture the underlying patterns that truly differentiate customers. Furthermore, manually analyzing large datasets is time-consuming, prone to human bias, and lacks scalability. There is a critical need for an automated, data-driven approach that can effectively identify distinct customer groups and adapt to varying data characteristics across different domains. This project addresses the problem by implementing the K-Means Clustering algorithm—an unsupervised machine learning technique that can uncover natural groupings within customer data without prior labeling. The challenge lies in determining the optimal number of clusters, accurately interpreting them, and ensuring the results are both meaningful and business-relevant. Additionally, visualizing the segmented data in a way that supports strategic decision-making is essential. By tackling these issues, this study provides a practical solution for businesses aiming to better understand their customers, tailor their offerings, and increase overall efficiency in marketing and service delivery

1.5 SUMMARY

Customer segmentation plays a pivotal role in modern business strategy, enabling organizations to understand and engage with their target audiences more effectively. Traditional segmentation techniques often fall short in capturing complex customer behavior, leading to generalized marketing efforts and missed opportunities. This project addresses these limitations by implementing a data-driven approach using the K-Means Clustering algorithm to segment customers based on demographic and behavioral attributes such as age, annual income, and spending score. The objective is to uncover hidden patterns within the data that can inform strategic decisions and personalize marketing efforts. A key motivation for this project is the increasing need for automation and accuracy in business analytics. As data volumes grow, manual analysis becomes less feasible, and machine learning offers scalable, intelligent solutions. K-Means clustering provides a foundation for segmenting customers into meaningful groups without predefined labels, allowing businesses to tailor their offerings and improve customer satisfaction. The use of the Elbow Method ensures optimal cluster selection, while visualizations enhance interpretability and decision-making.

Beyond segmentation, this project demonstrates how unsupervised learning can be applied in real-world business scenarios. It emphasizes the importance of visual analysis, clean data preprocessing, and model evaluation. Designed with scalability in mind, the project opens avenues for future enhancements such as real-time clustering, integration with CRM systems, and more advanced algorithms. By combining analytical rigor with practical application, this work contributes to the growing field of customer intelligence and supports the creation of data-driven, customer-centric strategies.

CHAPTER 2

LITERATURE REVIEW

2.1 K-MEANS CLUSTERING:

Several researchers, including Jain (2010) and Tan et al. (2014), have explored the use of K-Means Clustering in customer segmentation. This unsupervised machine learning algorithm is widely favored due to its simplicity, efficiency, and scalability for large datasets. K-Means operates by partitioning customers into k distinct clusters based on similarity in attributes such as age, income, and spending score. Despite its widespread use, a known limitation is its sensitivity to the initial placement of centroids and the assumption that clusters are spherical in shape. These limitations can impact the stability and accuracy of results, especially when the data has irregular distributions.

2.2 THE ELBOW METHOD

Research by Ketchen and Shook (1996) and others emphasized the importance of selecting the optimal number of clusters in K-Means. The Elbow Method is one of the most commonly used techniques for this purpose. It involves plotting the Within-Cluster Sum of Squares (WCSS) against different values of k and identifying the point where the reduction in WCSS slows down significantly, indicating the "elbow." While effective, this approach can be subjective and may not always produce a clear inflection point, especially with high-dimensional or overlapping data.

2.3 ENSEMBLE CLUSTERING TECHNIQUES

Recent advancements in clustering have introduced ensemble methods, where multiple clustering algorithms are combined to improve stability and accuracy. Strehl and Ghosh (2002) explored Cluster Ensembles for unsupervised learning, demonstrating improved robustness against noisy data and initial conditions. While these methods can enhance segmentation quality, they often increase computational complexity and may lack interpretability—key factors for real-world marketing applications where business users require clear and actionable insights.

2.4 DIMENSIONALITY REDUCTION IN SEGMENTATION

Dimensionality reduction techniques such as PCA (Principal Component Analysis) and t-SNE have been increasingly integrated into customer segmentation pipelines to improve clustering accuracy and visualization. Studies by Abdi & Williams (2010) showed how PCA can remove redundancy and reveal hidden structure in high-dimensional customer data. However, excessive dimensionality reduction may lead to loss of important information, potentially impacting the interpretability of the resulting clusters.

2.5 SUMMARY

This chapter provided a comprehensive review of the key machine learning techniques and methodologies relevant to customer segmentation. It began by exploring the foundational K-Means Clustering algorithm, emphasizing its widespread use in market analysis and its strengths in simplicity and scalability. The Elbow Method was discussed as a valuable tool for determining the optimal number of clusters, although it may sometimes yield subjective results. Ensemble clustering methods were introduced as a means to enhance segmentation accuracy, despite their increased computational complexity. The chapter also examined the role of dimensionality reduction techniques such as PCA, which can improve clustering performance and visualization but must be used cautiously to avoid losing critical information. Additionally, Hierarchical Clustering was reviewed for its usefulness in revealing data structure through dendrograms, though it was noted to be less suitable for large datasets due to computational demands. Overall, this literature review lays the theoretical foundation for the clustering model implemented in this project. It highlights the importance of selecting appropriate techniques based on data size, structure, and interpretability, all of which will inform the modeling and experimentation in the upcoming chapters.

CHAPTER 3

MODULE DESCRIPTION

3.1 CLUSTERING:

Imagine that you have a group of chocolates and liquorice candies. You are required to separate the two eatables. Intuitively, you are able to separate them based on their appearances. The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group. Clustering is used in various fields like image recognition, pattern analysis, medical informatics, genomics, data compression etc. It is part of the unsupervised learning algorithm in machine learning. This is because the data-points present are not labelled and there is no explicit mapping of input and outputs. As such, based on the patterns present inside, clustering takes place.

3.2 K-MEANS CLUSTERING

K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster. We then proceeded to perform K-means Clustering which will create different clusters to group similar spending activity based on their age and annual income. K- Means Clustering selects random values from the data and forms clusters assigned. The closest values from the centre of each cluster were taken to update the cluster and reshape the plot (just like k-NN). The closest values are based on Euclidean Distance.

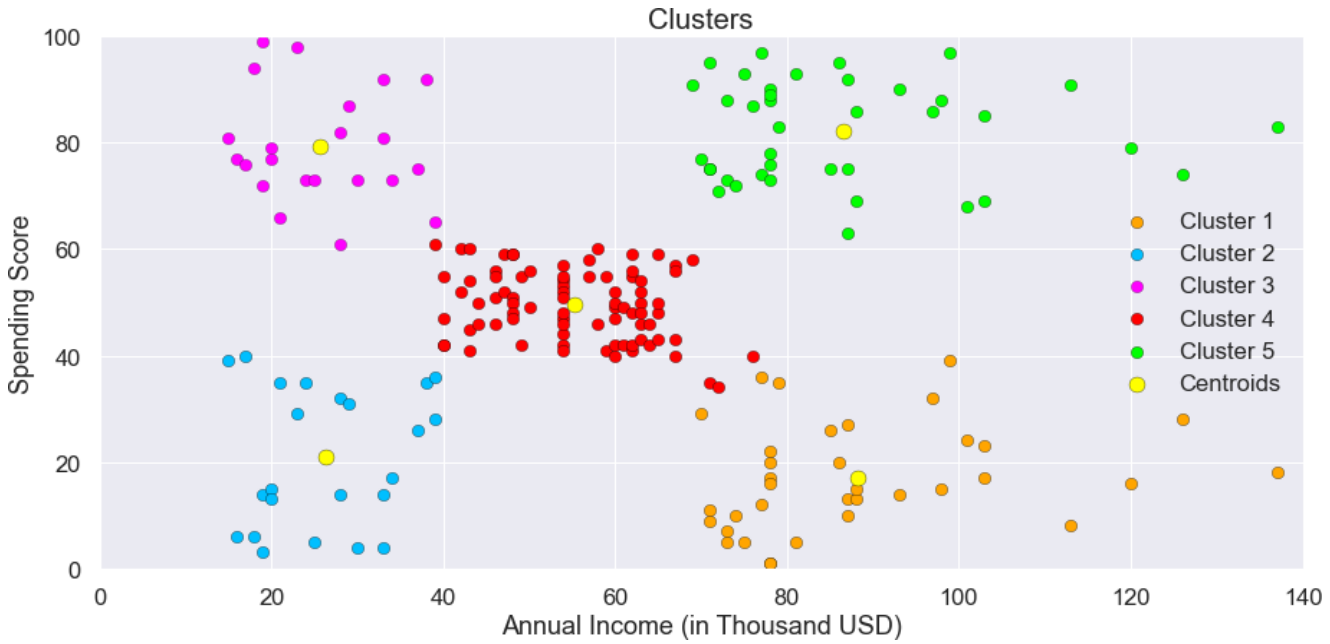


Figure 1: Final Cluster Visualization using K-Means (5 Clusters)

3.3 MODELING

Before applying clustering, we explore the dataset to understand distributions and relationships between features. Below are visualizations created during exploratory data analysis.

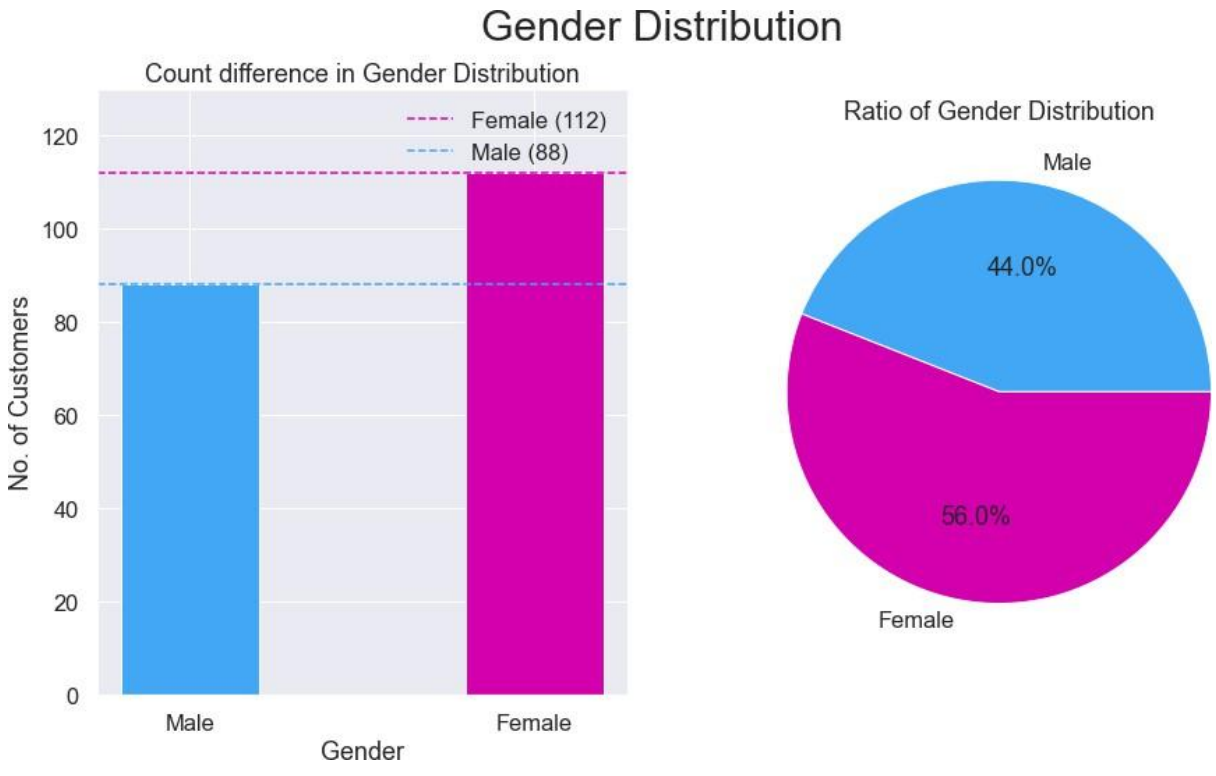


Figure 2: Gender-wise Customer Distribution

From the above graphs, we observe that the number of females(112) is higher than the males(88). The Ratio of Gender population is 56% Females and 44% Males. By this we can say that majority of the customers that visit the mall are Females.

Figure 3: Age Distribution of Customers

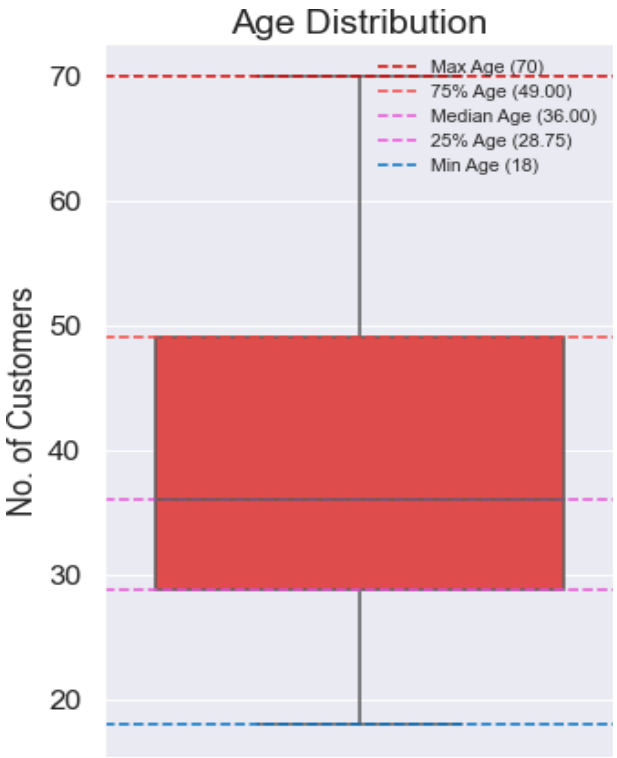


Figure 4: Age Analysis of Customers

From the above boxplot, we can conclude that a large amount of ages are between 30 and 35. Min Age is 18, Max Age is 70. By comparing the age distribution of the customers, we can conclude that most of the customers were

within the band between 30 to 50, where the mean is around 35 years old.

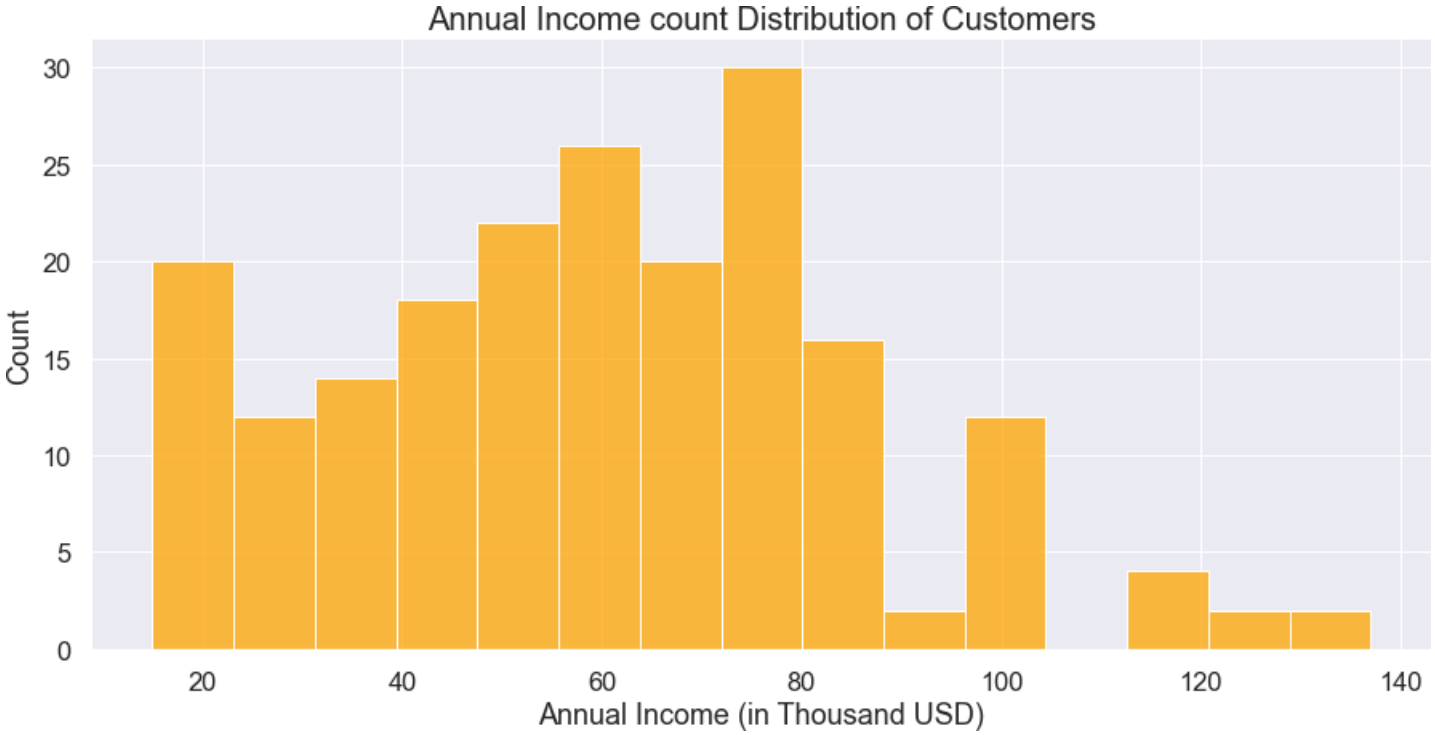


Figure 5: Annual Income and Spending Score Analysis

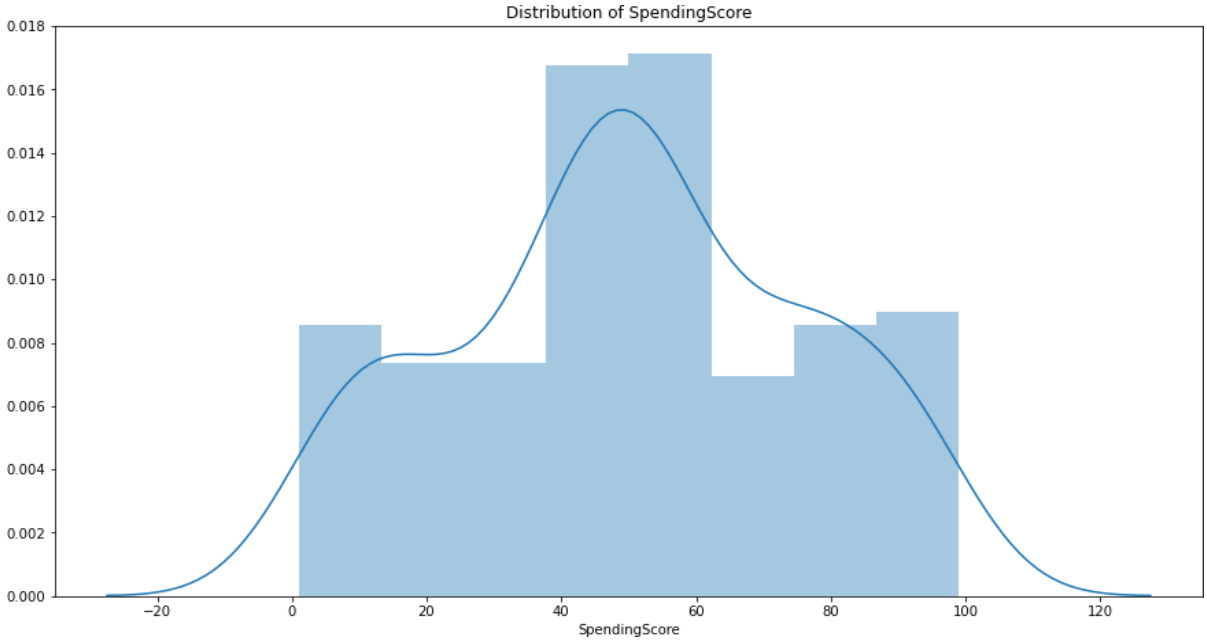


Figure 6: Distribution of Spending Score

The distribution of Annal Income and Spending Score exhibited an approximation of normal distribution, with highest density around the mean of the variables. The maximum and minimum of Annual Income are 137 and 15 respectively, with the mean at 60.56. From the plot, we can see that the peak of the distribution fell in the region of 60 to 75. For the Spending score, the maximum and minimum are 99 and 1, while the histplot indicated that the highest number of customers have the spending score ranging from 40 to 60.

3.4 CHARACTERISTIC RELATIONS

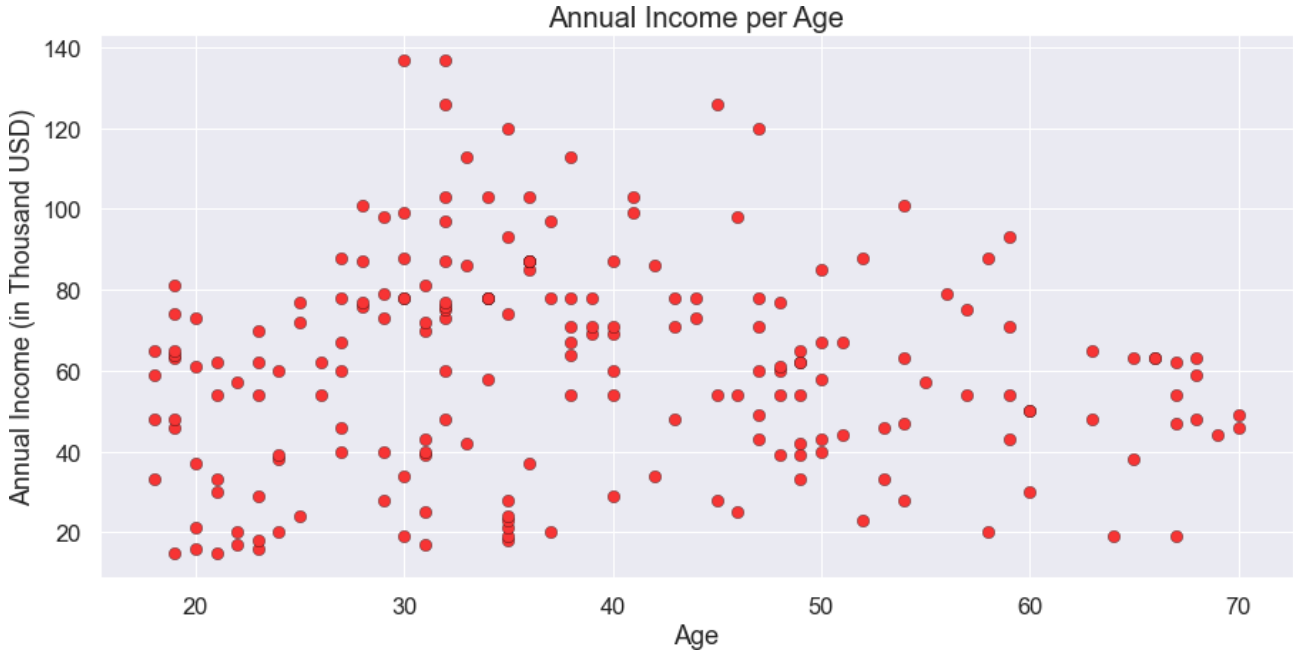


Figure 7: Annual Income Analysis

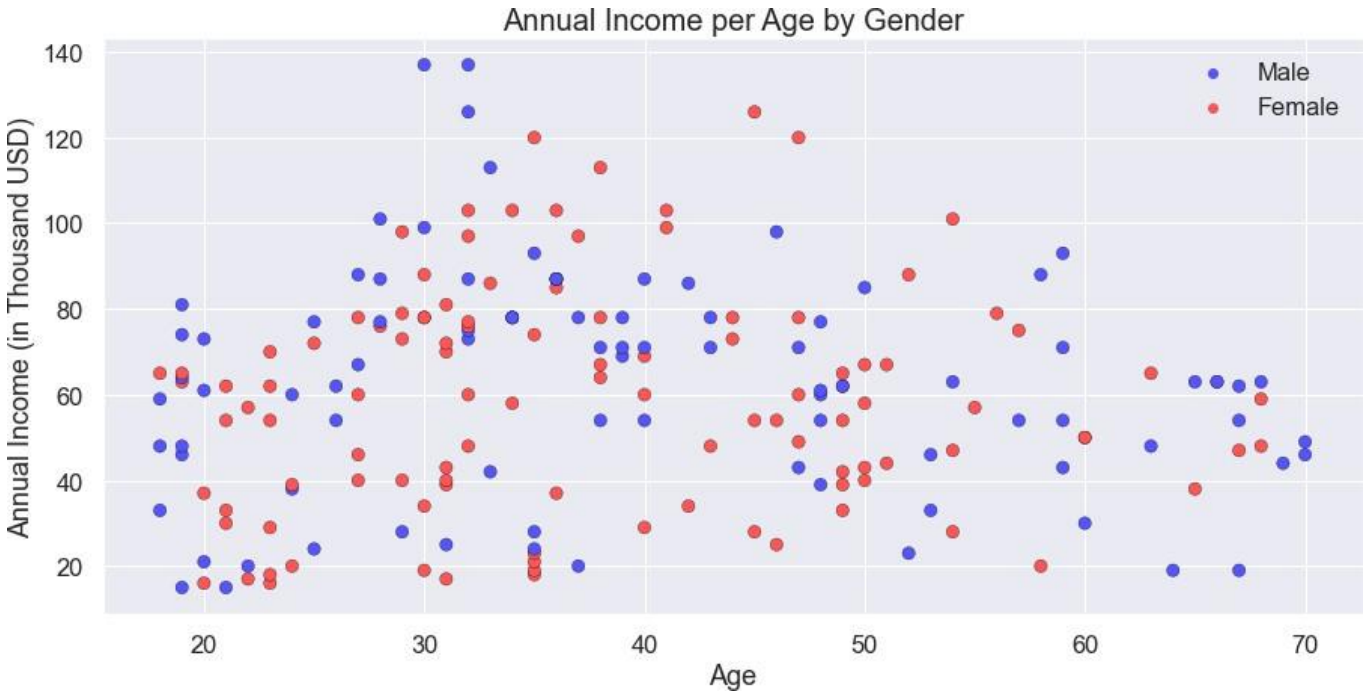


Figure 8: Spending Score Analysis

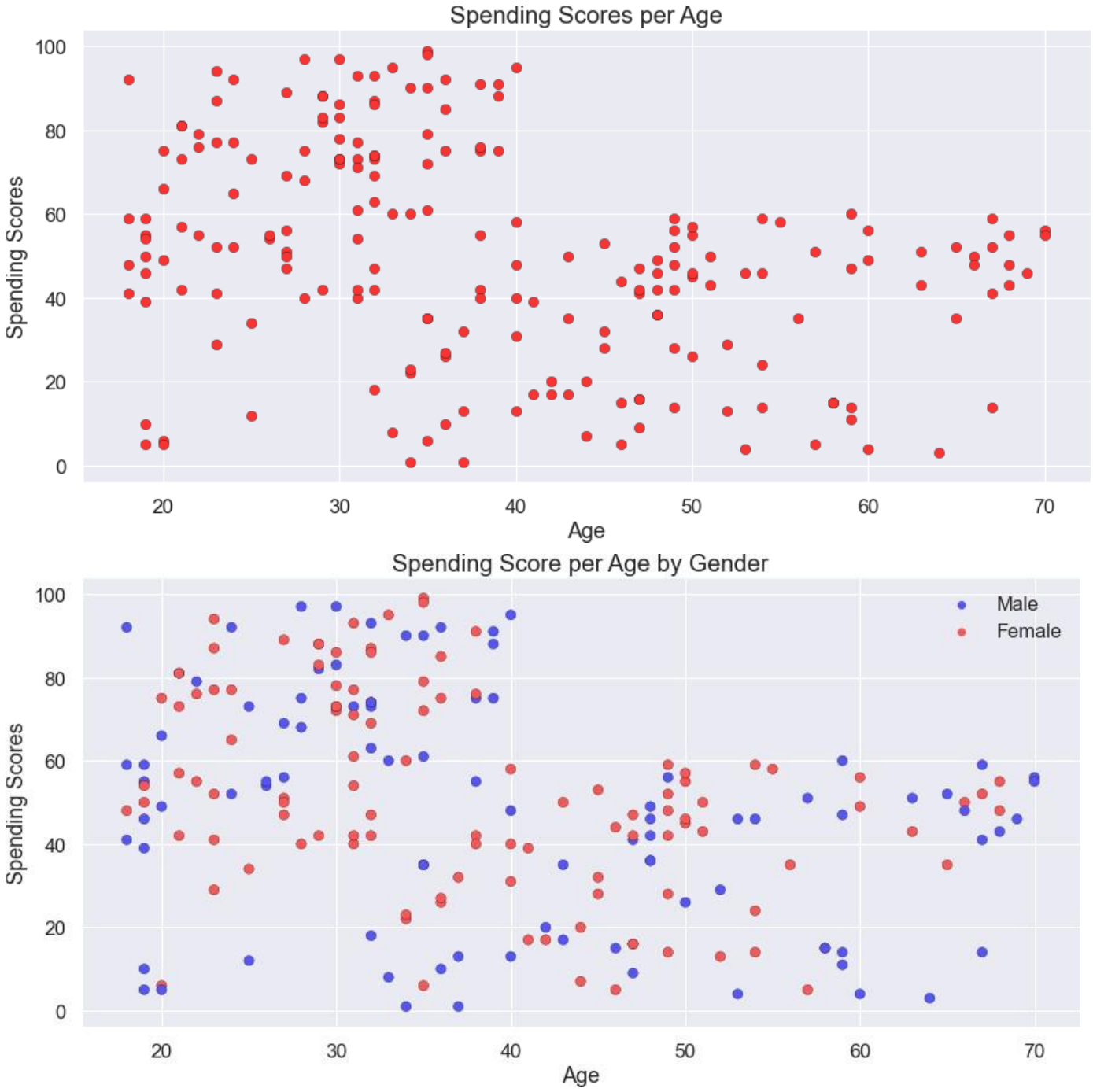


Figure 9 & 10: Spending Score Vs Age Analysis

3.5 THE ELBOW METHOD

The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center. When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point. We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formu

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

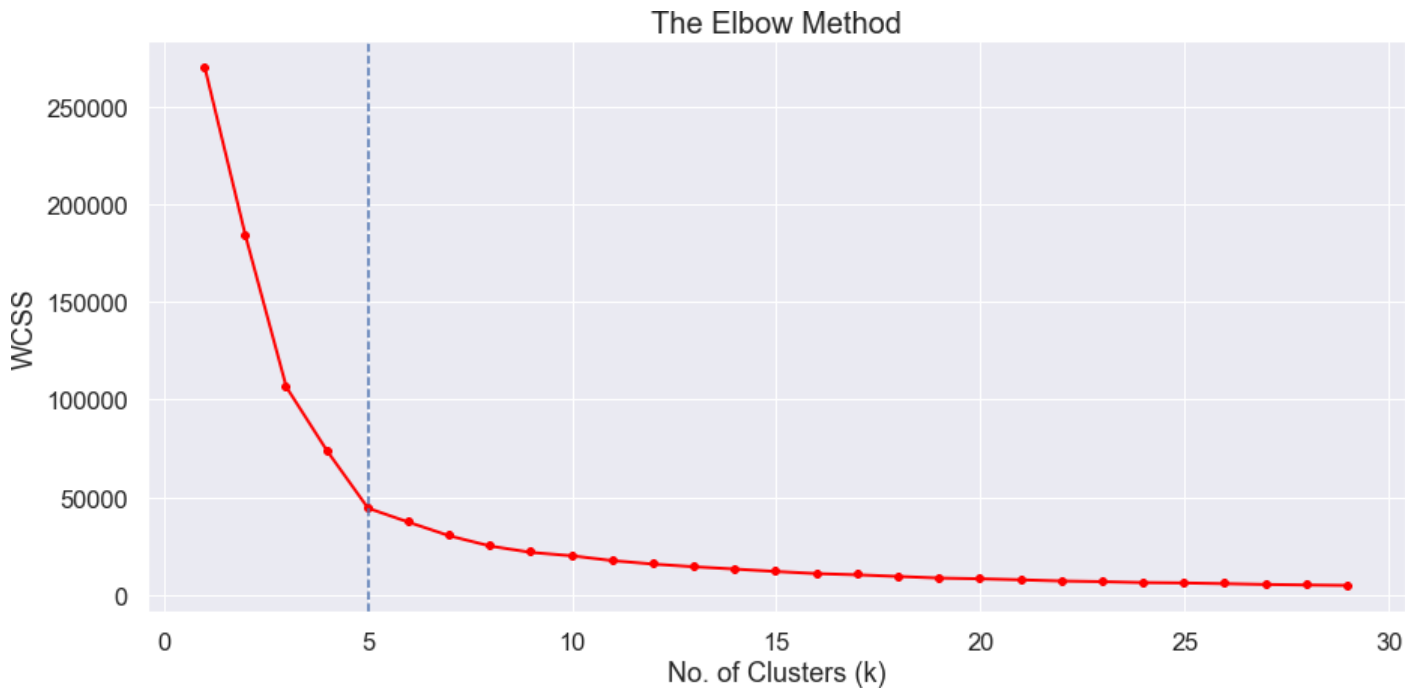


Figure 11: The Elbow Method

It is clear, that the optimal number of clusters for our data are 5, as the slope of the curve is not steep enough after it. When we observe this curve, we see that last elbow comes at k = 5, it would be difficult to visualize the elbow if we choose the higher range.

3.6 SUMMARY

This chapter presented a detailed overview of the customer segmentation project, beginning with an explanation of clustering and its role in unsupervised learning. K-Means Clustering was introduced as the primary algorithm used for segmenting customers based on shared attributes such as age, annual income, and spending score. The modelling section demonstrated how exploratory data analysis was performed to understand the distribution and relationships within the dataset, supported by various visualizations. The implementation of the Elbow Method helped determine the optimal number of clusters, which was found to be five. These clusters were then visualized to interpret customer groups more effectively. The graphical insights provided a clearer understanding of customer behavior and spending tendencies, laying the groundwork for strategic business decisions. Overall, this chapter covered the theoretical foundation, data preparation, model selection, and interpretation stages essential to implementing a segmentation solution using K-Means. The insights and visual outputs generated during this phase serve as a bridge to the next chapter, which focuses on the experimental setup, output analysis, and results.

CHAPTER 4

PROJECT ANALYSIS

4.1 CLUSTER ANALYSIS

The following clusters are created by the model,

1. Cluster Orange
2. Cluster Blue
3. Cluster Purple
4. Cluster Red
5. Cluster Green

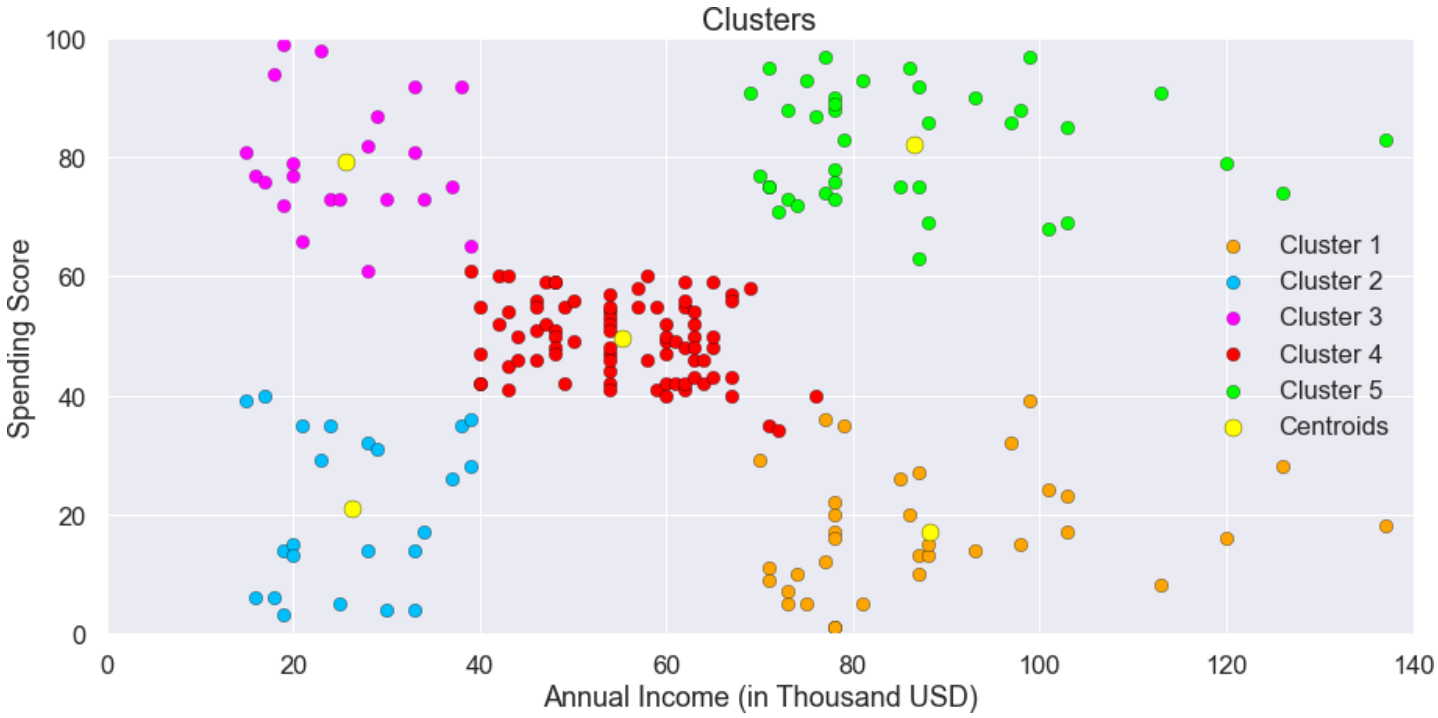


Figure 12: K-Means Clustering Result with 5 Clusters

1. **Cluster Orange - Balanced Customers:**

   They earn less and spend less. We can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

2. **Cluster Blue - Pinch Penny Customers:**

   Earning high and spending less. We see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

3. **Cluster Purple - Normal Customer:**

   Customers are average in terms of earning and spending An Average consumer in terms of spending and Annual Income we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

4. **Cluster Red - Spenders:**

   This type of customers earns less but spends more Annual Income is less but spending high, so can also be treated as potential target customer we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it's because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

5. **Cluster Green - Target Customers:**

   Earning high and also spending high Target Customers. Annual Income High as well as Spending

Score is high, so a target consumer. we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall's facilities.

4.2 DATASET AND TOOLS

The dataset used for this project is the "Mall Customer Segmentation Data," provided by Exposys Data Labs. This dataset consists of information on 200 mall customers and contains five key features that offer insights into consumer demographics and behavioral patterns. The features included are:

- **Customer ID**
- **Gender**
- **Age**
- **Annual Income (k$)**
- **Spending Score (1–100)**

These features serve as the foundation for unsupervised clustering, allowing us to identify distinct groups within the customer base based on their similarity in income and spending habits.

To perform the data preprocessing, clustering, and visualization tasks, the following tools and Python libraries were used:

- **Python (Jupyter Notebook)**: The primary development environment used for coding, visualization, and interactive analysis. It enables clear presentation of code, plots, and documentation in a single notebook format.

- **NumPy**: Utilized for handling numerical operations and array structures efficiently, which are essential for mathematical computations required during clustering.

- **Pandas**: Employed for data loading, manipulation, and preprocessing. It provides powerful data structures like DataFrames that simplify the process of handling tabular data.

- **Matplotlib**: Used to create static, animated, and interactive visualizations. It was instrumental in generating plots such as histograms, bar charts, and the Elbow curve.

- **Seaborn**: A statistical data visualization library based on Matplotlib. It was used for creating aesthetically appealing and informative plots, such as boxplots and pairplots, which helped in exploratory data analysis.

- **Scikit-learn**: The core machine learning library used for implementing the K-Means Clustering algorithm and calculating metrics such as the Within-Cluster Sum of Squares (WCSS). It also provided tools for preprocessing the dataset, including feature scaling.Together, these tools provided a comprehensive and efficient framework for building, analyzing, and visualizing the customer segmentation model. The integration of these libraries ensured a streamlined workflow from data cleaning to final cluster visualization.

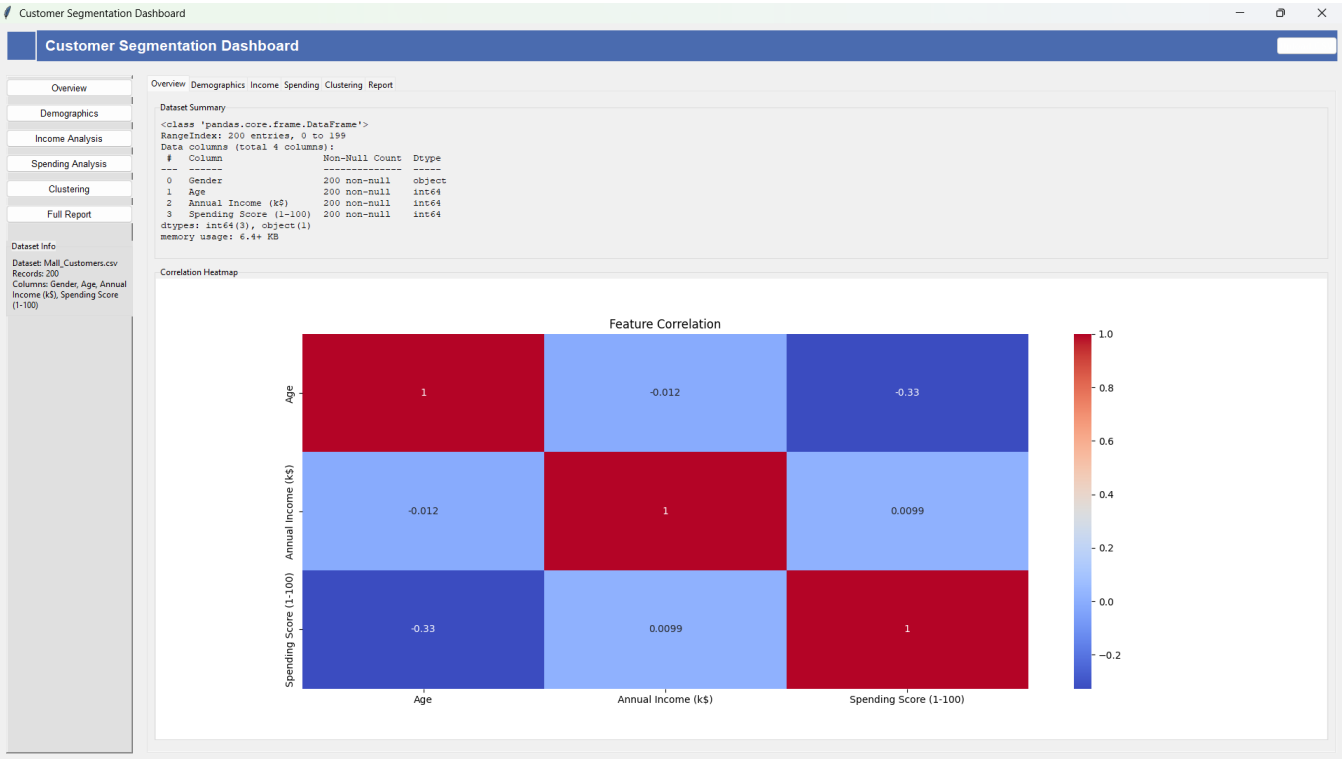
4.3 RESULTS

We have explored the five segments based on customers Annual Income and Spending Score which are reportedly the best factors/attributes to determine the segments of a customer in a Mall. They include; Pinch Penny Customers, Balanced Customers, Target Customers, Spender and the normal customer. We can put Target Customers into some alerting system where SMS and emails can be sent to them on daily basis regarding the offers and discounts that they can get at the Mall; while the rest we can set once per week in a month for blast SMSs to notify them about our products. Similarly, now we know customers behavior depending upon their Annual Income and Spending Score. There can be many marketing strategies applied for Customers on these Cluster Analysis. High income and High spending score customers are our target customers and we would always want to retain them as they give the most profit margin to our organization. High Income and Less spending score customers can be attracted with wide range of products in their life style demands and it might attract them towards the Mall Supermarket. Less Income Less Spending Score can be given extra offers and constantly sending them the offers and discounts will attract them towards spending. We can also have a cluster analysis done on what kind of products customers tend to buy and can make other marketing strategies accordingly. The data set did not have enough data to carry out more analytics on the same.

4.4 PERFORMANCE EVALUATION

The clustering algorithm performed efficiently on the dataset, and the Elbow Method helped determine the optimal number of clusters (k = 5). The segmentation provided clear insights into customer behavior, enabling targeted marketing strategies. Although no quantitative accuracy metric exists for unsupervised learning like K-Means, the visual separation and logical interpretation of clusters confirm model effectiveness.

4.5 OUTPUT SCREENSHOT (Main page) :



4.6 SUMMARY

This chapter outlined the experimental framework and implementation process behind the customer segmentation model using the K-Means clustering algorithm. It began with a description of the development environment, tools, and libraries employed, highlighting the utility of Python-based frameworks such as Pandas, NumPy, Seaborn, and Scikit-learn. These tools facilitated efficient data manipulation, visualization, and model development. The dataset used in the project was the “Mall Customer Segmentation Data,” which provided key demographic and behavioral features including age, gender, annual income, and spending score. These features were preprocessed and used to uncover patterns in customer behavior through clustering. Using

the Elbow Method, the optimal number of clusters was determined to be five. Each cluster was then visualized and analyzed, revealing meaningful insights about customer profiles such as balanced spenders, high-income low-spenders, and ideal target customers. The interpretation of these clusters provided clear and actionable insights for strategic marketing, personalization, and customer retention efforts.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

Companies, Malls, super markets on Small Business Enterprises should carry out Market Basket Analysis for their business. This will enable companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities.

When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention. As a by-product of its personalized nature, marketing materials sent out using customer segmentation tend to be more valued and appreciated by the customer who receives them as opposed to impersonal brand messaging that doesn't acknowledge purchase history or any kind of customer relationship.

Finally with customer segmentation Companies will stay a step ahead of competitors in specific sections of the market and identify new products that exist or potential customers could be interested in or improving products to meet customer expectations.

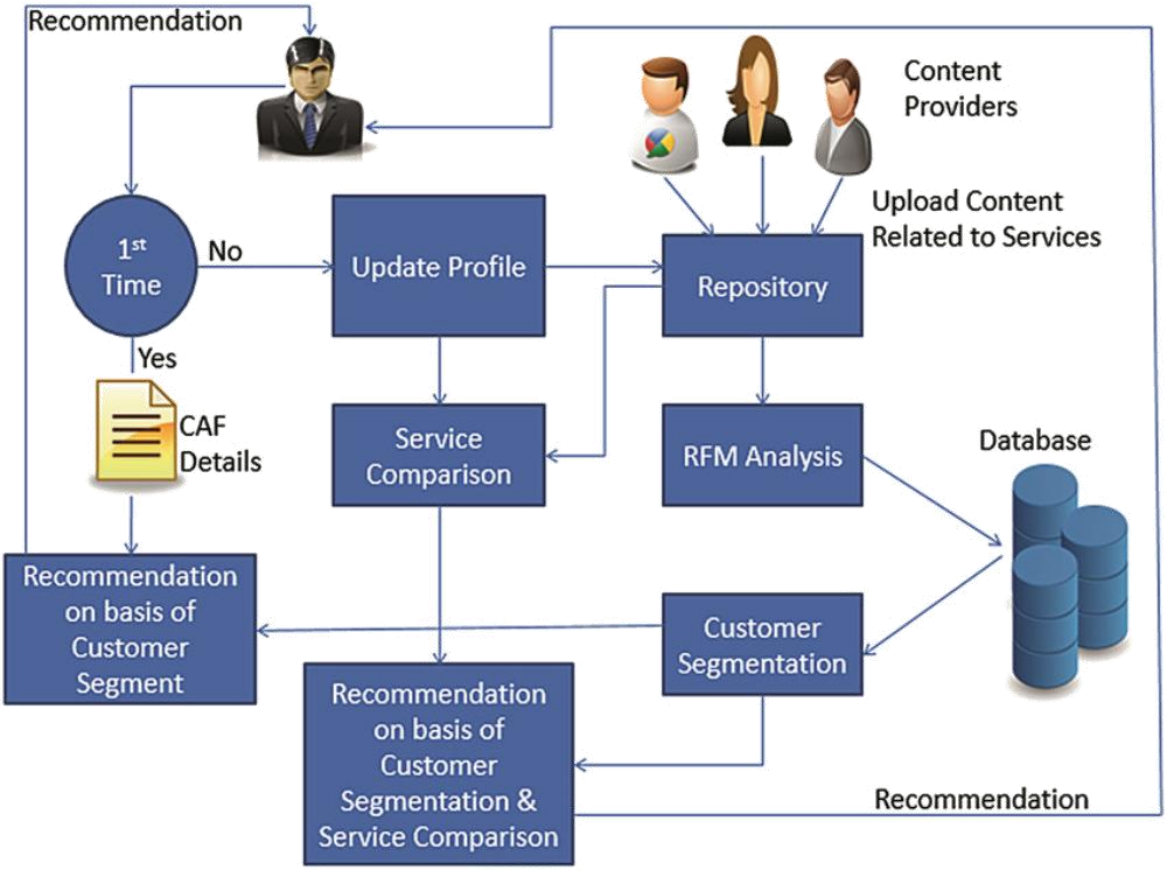
5.2 FUTURE ENHANCEMENT

Although the current implementation using K-Means offers strong baseline performance, there are several enhancements that could improve its versatility and predictive power. One potential improvement involves incorporating more features into the dataset, such as purchase frequency, customer tenure, product category preferences, or time-of-day spending patterns, to allow for deeper and more precise clustering. Advanced algorithms such as DBSCAN, Hierarchical Clustering, or Gaussian Mixture Models could be employed to address limitations of K-Means, especially in cases of non-spherical clusters or datasets with noise. Another important direction would be the development of a real-time segmentation engine that can dynamically update customer groups based on live data streams, which would be useful for applications such as recommendation systems or time-sensitive marketing. Integration with tools like Power BI or Tableau could provide interactive dashboards for business users to explore and monitor customer segments. Additionally, incorporating supervised learning components such as churn prediction, customer lifetime value (CLV), or personalized offer optimization could make the segmentation process more strategic and proactive. Future versions of the project may also leverage cloud platforms for scalability, real-time performance monitoring, and integration with external data sources.

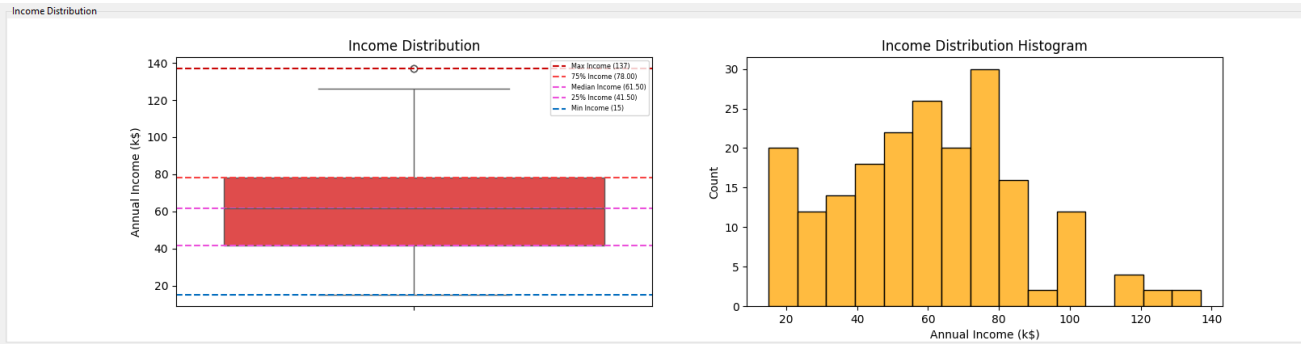
5.3 SUMMARY

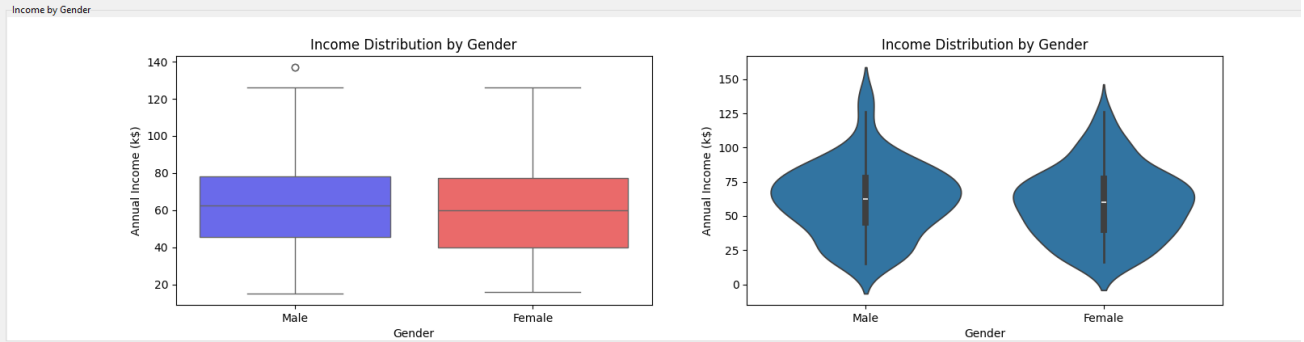
In summary, this chapter concluded the customer segmentation project by highlighting the success of the K-Means model in discovering insightful patterns within customer data. It reinforced the importance of unsupervised learning in modern marketing analytics and proposed several avenues for enhancing the model's scope and business impact. With proper extensions and integration into operational systems, this solution can drive meaningful improvements in customer experience, retention, and business profitability.

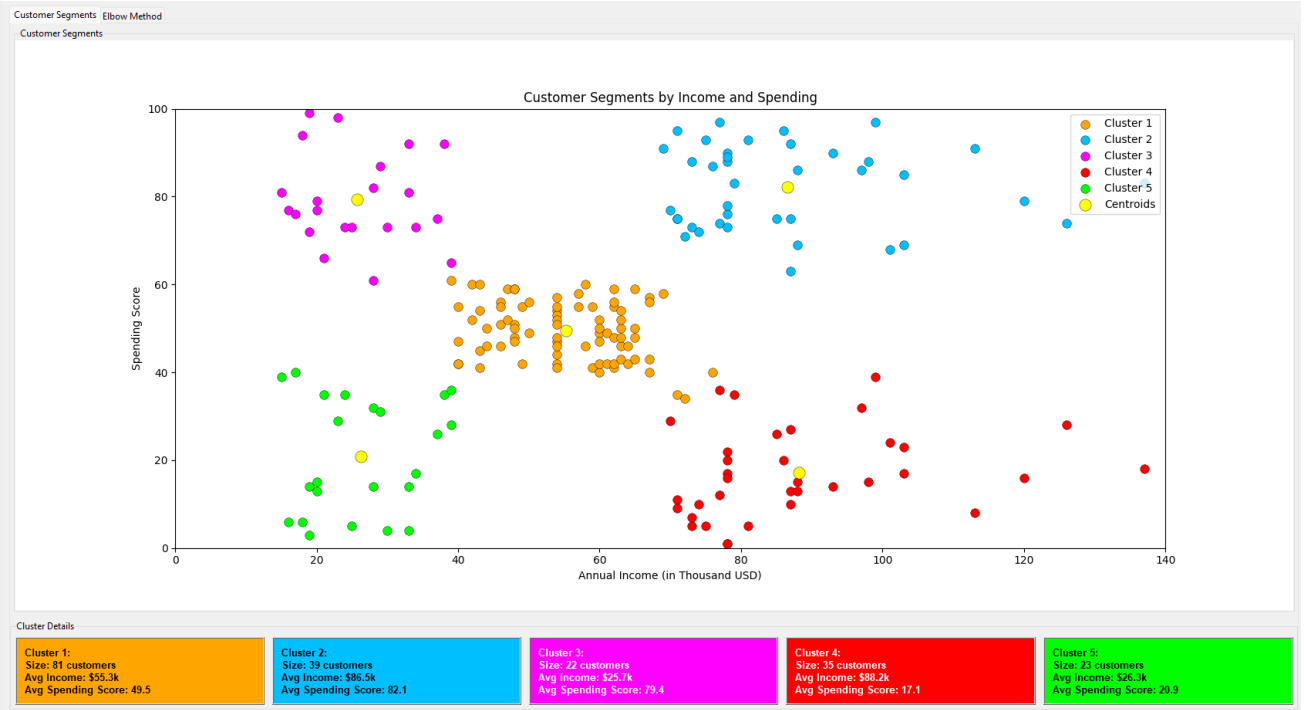
ARCHITECTURE DIAGRAM :

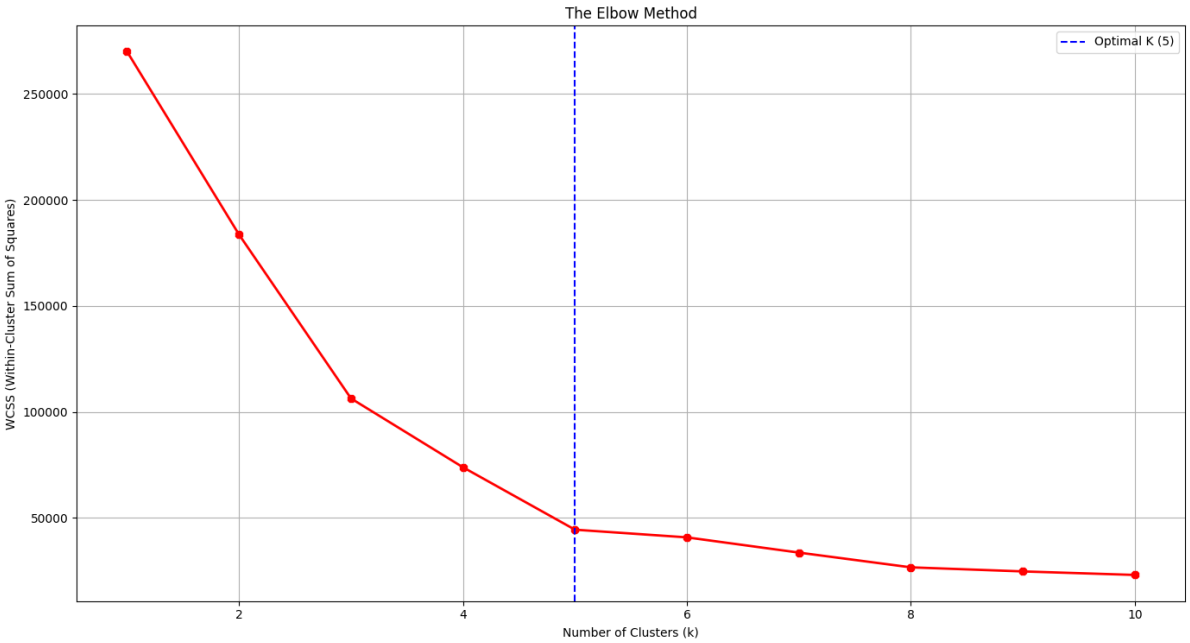


OUTPUT SCREENSHOTS :


Income Distribution
Income Distribution
Max Income (137)
75% Income (78.00)
Median Income (61.50)
25% Income (41.50)
Min Income (15)
Annual Income (k$)
Income Distribution Histogram
Count
Annual Income (k$)


Income by Gender
Income Distribution by Gender
Annual Income (k$)
Male
Female
Gender
Income Distribution by Gender
Annual Income (k$)
Male
Female
Gender


Customer Segments
Elbow Method
Customer Segments
Customer Segments by Income and Spending
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Centroids
Spending Score
Annual Income (in Thousand USD)
Cluster Details
Cluster 1:
Size: 81 customers
Avg Income: $55.3k
Avg Spending Score: 49.5
Cluster 2:
Size: 39 customers
Avg Income: $86.5k
Avg Spending Score: 82.1
Cluster 3:
Size: 22 customers
Avg Income: $25.7k
Avg Spending Score: 79.4
Cluster 4:
Size: 35 customers
Avg Income: $88.2k
Avg Spending Score: 17.1
Cluster 5:
Size: 23 customers
Avg Income: $26.3k
Avg Spending Score: 20.9


The Elbow Method
Optimal K (5)
250000
200000
150000
100000
50000
WCSS (Within-Cluster Sum of Squares)
2
4
6
8
10
Number of Clusters (k)

REFERENCES

[1] Z. Li and J. Zhang, "A path to implementing a fresh produce e-commerce customer segmentation method based on clustering algorithms," *Proc. Int. Conf. Inf. Technol. Optim. Educ. Conf. (ITOEC)*, 2023, pp. 1–6, doi: 10.1109/ITOEC57671.2023.10291762.

[2] C. C. Lin, S. H. Chiu, and Y. C. Tseng, "A Two Phase Clustering Method for Intelligent Customer Segmentation," *Proc. Int. Symp. Market. Sci. (ISMS)*, 2010, pp. 1–6, doi: 10.1109/ISMS.2010.48.

[3] P. P. Singh and R. S. Kahlon, "Customer Segmentation using K-means Clustering," *Proc. Int. Conf. Comput. Tech. Electron. Mech. Syst. (CTEMS)*, 2018, pp. 122–125, doi: 10.1109/CTEMS.2018.8769171.

[4] X. Wang and G. Wu, "Service-mining Based on Customer Value Analysis," *Proc. Int. Conf. Manage. Sci. Eng. (ICMSE)*, 2007, pp. 1161–1165, doi: 10.1109/ICMSE.2007.4421833.

[5] R. Teh, W. M. Chong, and W. A. W. Adnan, "Segmenting customers with data mining techniques," *Proc. Digit. Inf. Netw. Commun. Conf. (DINWC)*, 2015, pp. 111–115, doi: 10.1109/DINWC.2015.7054234.

[6] S. S. H. Rizvi and Z. P. Shaikh, "Review of customer segmentation method in CRM," *Proc. Int. Conf. Softw. Secur. Syst. (CSSS)*, 2011, pp. 67–72, doi: 10.1109/CSSS.2011.5974617.

[7] Y. Zhai, H. Yu, and Q. Sun, "Research on E-commerce Customer Segmentation Based on RFAC Model," *Proc. Int. Conf. Pattern Inf. Commun. Eng. (ICPICS)*, 2021, pp. 572–577, doi: 10.1109/ICPICS52425.2021.9524108.

[8] J. S. Breese, D. Heckerman, and C. Kadie, "Improving Personalization Solutions through Optimal Segmentation of Customer Bases," *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2006, pp. 518–524, doi: 10.1109/ICDM.2006.87.