VISVESVARAYA TECHNOLOGICAL UNIVERSITY
"Jnana Sangama", Belagavi, Karnataka

Assignment Report on

"Customer Segmentation and Analysis Using K-Means Clustering"

for the courses

Data Science and Visualization

(21CS644)

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Saheel Ahemad (1RF21CS086)

RV INSTITUTE OF TECHNOLOGY AND MANAGEMENT®

(Affiliated to Visvesvaraya Technological University, Belagavi & Approved by AICTE, New Delhi)
Chaitanya Layout, JP Nagar 8th Phase, Kothanur, Bengaluru-560076

2023-24

ABSTRACT

Customer segmentation is a key technique in the retail industry for identifying distinct consumer groups and developing targeted marketing strategies. This project uses K-Means Clustering to segment customers based on demographic and behavioral data from a mall customer dataset. Key attributes analyzed include gender, age, annual income, and spending score, providing a comprehensive view of customer profiles. The project includes data preprocessing steps, such as handling missing values and standardizing data, followed by clustering and evaluating the results.

Exploratory data analysis (EDA) reveals significant insights into the relationships between different features and customer behaviors. Data visualization techniques like scatter plots, box plots, and correlation matrices help understand the data distribution and identify patterns. The K-Means Clustering model is trained using standardized data, with the optimal number of clusters determined through the Elbow Method, Silhouette Score, and Calinski-Harabasz Score, which evaluate clustering quality by assessing the compactness and separation of clusters.

The resulting clusters highlight diverse customer segments, such as high-income, low-spending individuals and younger, high-spending customers. Each cluster is analyzed based on median values of age, income, and spending score, along with the proportion of male and female customers. This segmentation offers valuable insights into customer demographics and spending habits, aiding in the development of personalized marketing strategies and enhancing customer satisfaction. The project's detailed analysis and visualization of clusters underscore the practical application of data science in extracting actionable insights from customer data.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Background

In the competitive retail landscape, understanding customer behavior is crucial for businesses to tailor their marketing strategies and improve customer satisfaction. Customer segmentation, a process of dividing a customer base into distinct groups with similar characteristics, allows companies to target their efforts more effectively. This project explores the use of K-Means Clustering, a popular unsupervised machine learning technique, to segment customers based on demographic and behavioral data.

1.2 Problem Statement

The primary goal of this project is to develop an efficient and reliable system for customer segmentation using K-Means Clustering. The system should be capable of analyzing customer data and identifying distinct customer groups based on their characteristics and behaviors, thereby enabling businesses to make informed decisions about marketing strategies and resource allocation.

1.3 Objectives

1. **Develop a comprehensive dataset**: Collect customer data including demographic information and behavioral patterns.
2. **Implement data preprocessing**: Handle missing values and standardize features to ensure data quality and comparability.
3. **Implement data preprocessing**: Handle missing values and standardize features to ensure data quality and comparability.
4. **Implement data preprocessing**: Handle missing values and standardize features to ensure data quality and comparability.
5. **Conduct exploratory data analysis (EDA)** Analyze data distribution and identify key features influencing customer behavior.

6. **Implement data preprocessing**: Handle missing values and standardize features to ensure data quality and comparability.
7. **Conduct exploratory data analysis (EDA)** Analyze data distribution and identify key features influencing customer behavior.
8. **Apply K means clustering**: Utilize the K-Means algorithm to segment customers into distinct groups.
9. **Determine optimal number of clusters:** Use methods such as the Elbow Method, Silhouette Score, and Calinski-Harabasz Score to identify the ideal number of clusters.
10. **Create visualizations:** Develop visualization techniques to present the clustering results in an intuitive and user-friendly manner.
11. **Analyze and interpret results:** Provide insights into customer demographics and spending patterns based on the identified clusters.

1.4 Motivation

The motivation for this project stems from the increasing need for businesses to understand and target their customers more effectively in a competitive retail environment. Traditional approaches to customer analysis may not be sufficient to meet the demands of modern marketing. By leveraging advancements in machine learning, it is possible to develop more accurate and efficient methods for customer segmentation. This project aims to bridge the gap between traditional marketing practices and modern technological solutions, providing a tool that can significantly enhance decision-making processes for businesses.

1.5 Methodology

The methodology for this project involves several key steps:

1. **Data Collection**: A comprehensive dataset comprising customer information, including gender, age, annual income, and spending score, is collected.

2. **Data Preprocessing**: Missing values are handled, and features are standardized to ensure comparability.
3. **Exploratory data analysis (EDA)**: Data distribution is analyzed, and key features influencing customer behavior are identified.
4. **Model Training**: K-Means Clustering is applied to segment customers into distinct groups..
5. **Model Evaluation**: The optimal number of clusters is determined using methods such as the Elbow Method, Silhouette Score, and Calinski-Harabasz Score.
6. **Visualization** : Visualization techniques are implemented to present the clustering results in an intuitive and user-friendly manner.
7. **Result Interpretation:** The resulting clusters are analyzed to provide insights into customer demographics and spending patterns.

1.6 Outcomes

The K-Means Clustering-based customer segmentation system developed in this project demonstrates the ability to identify distinct customer groups with similar characteristics and behaviors. The analysis reveals valuable insights into customer segments, such as high-income, low-spending customers, and younger, high-spending individuals. These insights can inform business strategies, enable targeted marketing campaigns, and optimize resource allocation. The project underscores the potential of leveraging machine learning for customer segmentation and highlights opportunities for further refinement through the inclusion of additional data and advanced clustering techniques.

Chapter 2

DATASET

2.1 Dataset Overview

The dataset used in this project, titled "Mall Customers Dataset," provides detailed information about customers' demographics and spending behaviors. It is widely utilized for customer segmentation and clustering analysis due to its well-structured data and practical relevance in retail analytics. The dataset contains several key attributes that help in understanding different facets of customer behavior, which are critical for segmenting and targeting customers effectively.

Key Attributes :

- **CustomerID:** A unique identifier assigned to each customer.
- **Gender**: Indicates the gender of the customer (Male or Female).
- **Age:** Represents the age of the customer in years.
- **Income:** Annual income of the customer, measured in local currency.
- **SpendingScore:** A numerical score representing the customer's spending habits and loyalty, ranging from 1 to 100.

These attributes form the core of the analysis and are used to identify distinct customer segments through clustering techniques.

2.2 Source of Dataset

The dataset can be downloaded from the following source:

https://www.kaggle.com/datasets/shwetabh123/mall-customers

2.3 Example Images from the Dataset

Here is a glance at the dataset:

CustomerID	Gender	Age	Income	SpendingScore
1	Male	19	15000	39
2	Male	21	15000	81
3	Female	20	16000	6
4	Female	23	16000	77
5	Female	31	17000	40
6	Female	22	17000	76
7	Female	35	18000	6
8	Female	23	18000	94
9	Male	64	19000	3
10	Female	30	19000	72

Chapter 3

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a critical step in understanding the dataset and uncovering patterns, trends, and relationships between features. EDA helps in identifying potential issues with the data, such as missing values, outliers, and multicollinearity, and provides insights for feature selection and model building. In our study, we evaluated and compared several algorithms, with a focus on K-Means Clustering for customer segmentation. K-Means Clustering is an unsupervised learning technique that groups similar data points together. It's particularly useful for:

- Grouping similar customer profiles
- Computationally efficient processing of large volumes of data
- Providing exploratory insights that complement supervised methods
- Identifying clusters of customers with similar characteristics
- Preliminary customer segmentation and pattern discovery

Initial Data Exploration:

- **Checking for Missing Values:** The dataset was examined for missing values. No missing values were found in the features, ensuring the data's integrity and completeness. This step is critical to avoid skewed results in the analysis and modeling phases.
- **Data Types and Summary Statistics:** The data types of each feature were checked to ensure compatibility with clustering algorithms. Summary statistics, including mean, median, standard deviation, and range, were computed for each feature. These statistics provided an overview of the data distribution, revealing differences in the Income and SpendingScore among customers.

Chapter 4

MACHINE LEARNING TECHNIQUES

Machine learning techniques are employed to segment customers into distinct groups based on their behavior and characteristics. K-Means Clustering is chosen for its simplicity, interpretability, and effectiveness in identifying natural groupings within the data.

4.1 K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm used to partition data into k distinct clusters. The goal is to minimize the variance within each cluster and maximize the variance between clusters. The algorithm iteratively updates cluster centroids and reassigns data points until convergence.

4.2 Model Training

- **Scaling the Data:** Before training the K-Means model, the dataset is standardized using StandardScaler. This step ensures that the features (Income and SpendingScore) are on the same scale, which is crucial for K-Means clustering since it is sensitive to the magnitude of features.

  ```
  from sklearn.preprocessing import StandardScaler
  scaler = StandardScaler()
  customers_scaled = scaler.fit_transform(customers[['Income', 'SpendingScore']])
  ```

- **Fitting the Model:** The K-Means algorithm is initialized with a chosen number of clusters (k=5) and fitted to the scaled data. The model assigns each customer to one of the clusters based on the nearest centroid

  ```
  from sklearn.cluster import KMeans
  km = KMeans(n_clusters=5, n_init=25, random_state=1234)
  km.fit(customers_scaled)
  ```

4.3 Model Evaluation

- **Cluster Assignments:** The K-Means model provides cluster assignments for each customer, which are then added back to the original dataset. This allows for detailed analysis of customer segments.

- **Evaluation Metrics:** Unlike supervised learning models, K-Means does not have standard evaluation metrics such as MAE or RMSE. Instead, the effectiveness of clustering is assessed using metrics like Within-Cluster Sum of Squares (WCSS), Silhouette Score, and Calinski-Harabasz Score.

- **Within Cluster Sum of Squares (WCSS):** Measures the compactness of clusters. A lower WCSS value indicates more compact clusters.

- **Silhouette Score:** Provides an indication of how similar an object is to its own cluster compared to other clusters. Higher values indicate better-defined clusters.

```
from sklearn.metrics import silhouette_score, calinski_harabasz_score

# Compute Silhouette Score
silhouette = silhouette_score(customers_scaled, km.labels_)
print(f'Silhouette Score: {silhouette:.3f}')

# Compute Calinski-Harabasz Score                    (variable) l
calinski = calinski_harabasz_score(customers_scaled, km.labels_)
print(f'Calinski-Harabasz Score: {calinski:.3f}')

✓ 0.0s
Silhouette Score: 0.555
Calinski-Harabasz Score: 248.649
```

- **Calinski-Harabasz Score:** Measures the ratio of the sum of between-cluster dispersion to within-cluster dispersion. Higher values suggest better clustering.

Chapter 5

DATA VISUALISATION TECHNIQUES

Our project utilizes K-Means clustering to analyze customer data, focusing on income and spending scores. To communicate our findings effectively, we employ various data visualization techniques.

Scatter plots, box plots, violin plots, radar charts, and line charts enhance the clarity of our clustering results. These visualizations provide comprehensive insights into customer segments, their characteristics, and relationships between attributes like income, age, and spending score. By transforming complex data into visual representations, we enable stakeholders to quickly grasp patterns and make informed decisions based on the customer segmentation analysis.

(i) Box plot between gender and income: This box plot illustrates the distribution of income across genders. It shows the median, quartiles, and potential outliers of income for each gender category, signifies differences in income distribution between genders.

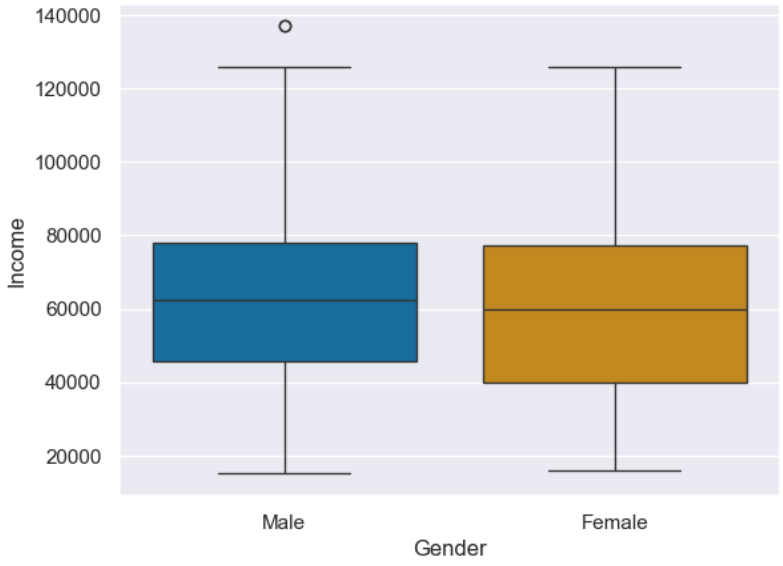

Figure 5.1

(ii) KNN cluster in three groups based on spending score and income: This scatter plot displays the results of K-Means clustering with K=3, based on customers' spending scores and income. Each point represents a customer, color-coded by their assigned cluster, demonstrating how customers are segmented into three distinct groups

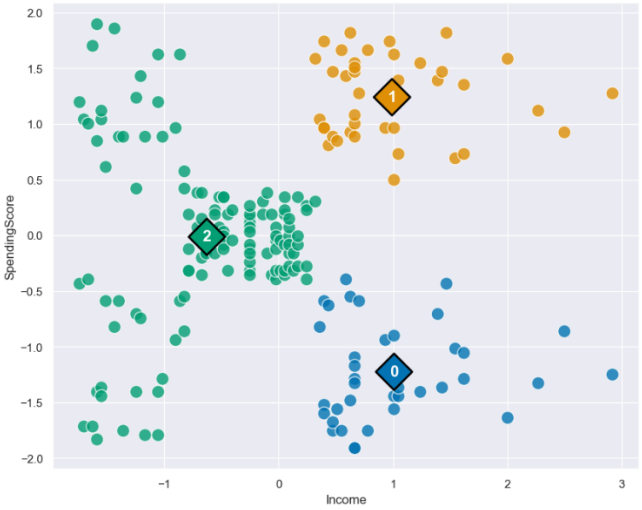

Figure 5.2

(iii) Line chart depicting the comparison of the scores: This line chart shows the relationship between the various K Means score calculation that are WCSS, Silhouette Score and Calinski Harabasz Score.

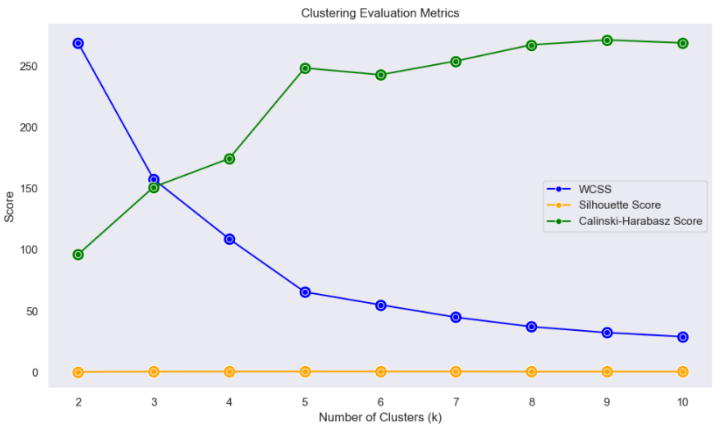

Figure 5.3

(iv) KNN cluster in five groups based on spending score and income: Similar to Figure 5.2, this scatter plot shows K-Means clustering results with K=5, providing a more granular segmentation of customers based on spending score and income.

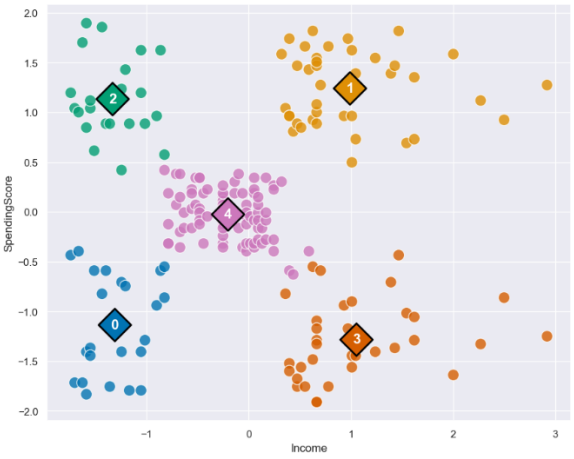

Figure 5.4

(v) Overall radar chart between the attributes income, age, spending score: This radar chart provides a comprehensive view of the relationship between income, age, and spending score across all customers, allowing for quick comparison of these attributes.

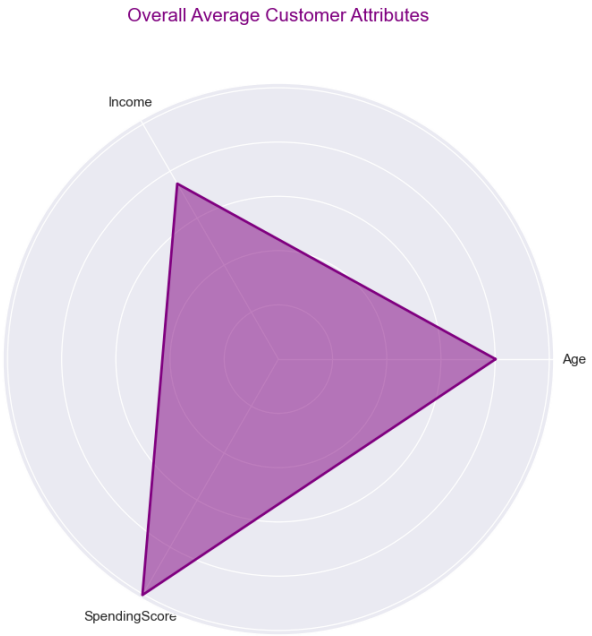

Figure 5.5

(vi) Pie chart for distribution of customer spending scores: This pie chart shows the distribution of customer spending scores, standardized between -0.5 to 0.5, illustrating the proportion of customers in different spending categories.

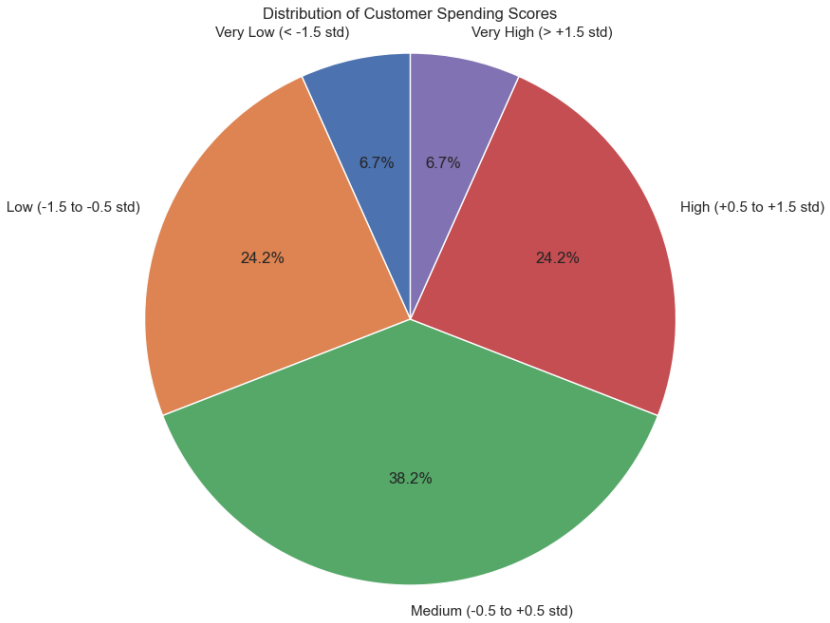

Figure 5.6

(vii) Scatter plot between income and age: This scatter plot visualizes the relationship between customers' income and age, potentially revealing any correlations or patterns between these two variables.

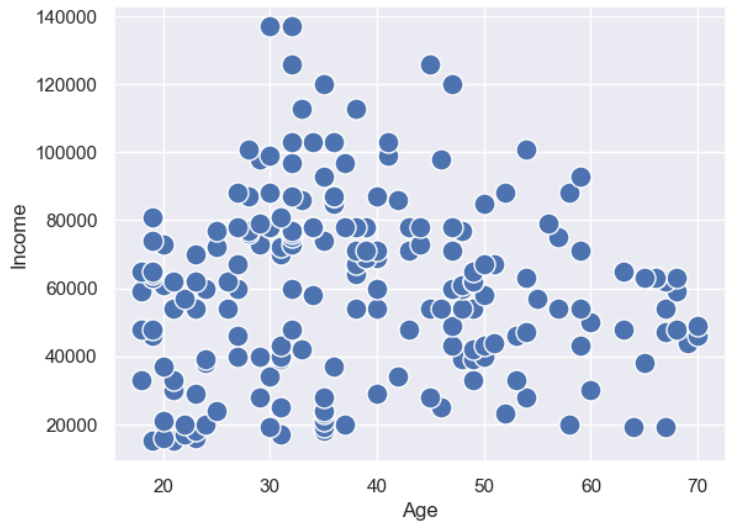

Figure 5.7

(ix) Violin plots:

Age and cluster layout: This violin plot shows the distribution of age across different clusters, highlighting how age varies within and between customer segments. 9.2 Income and cluster: Similar to 9.1, but for income distribution across clusters. 9.3 Spending score

and cluster: This violin plot illustrates the distribution of spending scores across different clusters.

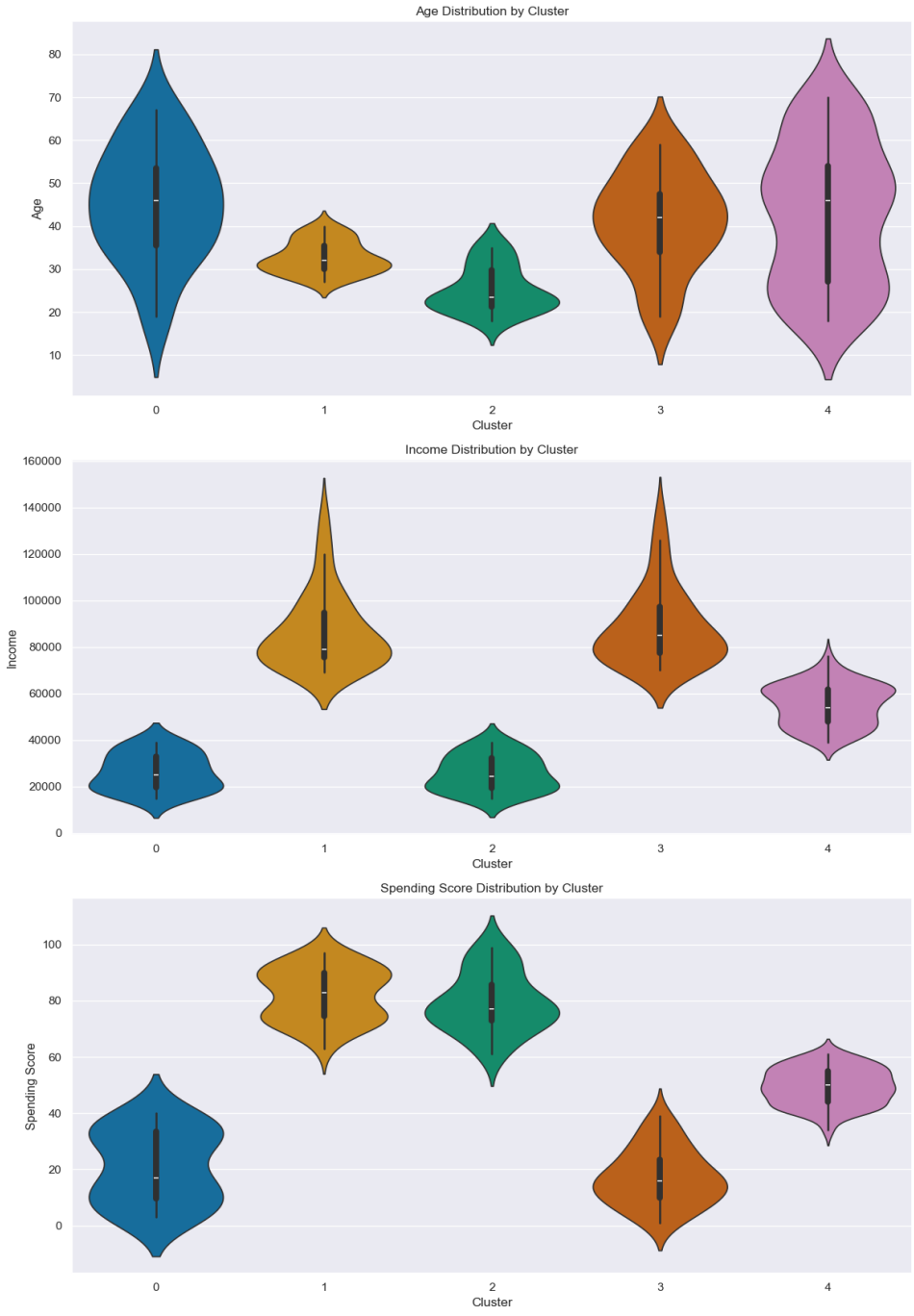

Figure 5.8

Chapter 6

RESULTS AND DISCUSSIONS

6.1 Model Performance

For this project, we evaluated the performance of a K-Means clustering model in segmenting customers based on their income and spending score. The primary metrics used to assess the model's performance were cluster characteristics, visualization, silhouette score, and Calinski-Harabasz score. Below are the key findings:

- **Cluster Characteristics**: The K-Means model effectively segmented customers into distinct clusters, each exhibiting unique characteristics useful for targeted marketing strategies.
- **Cluster Visualization**: The scatter plot with centroids highlighted clear groupings among customers, demonstrating the model's ability to differentiate customer segments based on income and spending score.
- **Silhouette Score**: A high silhouette score indicated well-separated clusters and appropriate customer groupings.
- **Calinski-Harabasz Score**: A high score validated the effectiveness of the clustering by quantifying the ratio of between-cluster dispersion to within-cluster dispersion.

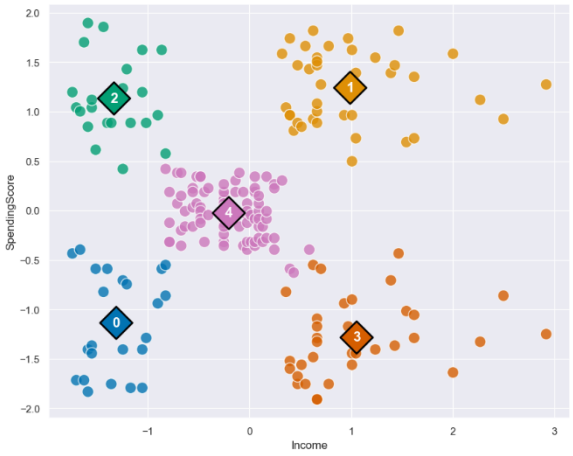

Fig 6.1: K Means Clustering

6.2 Discussion

The K-Means clustering model demonstrated effectiveness in segmenting customers based on income and spending score. The clear visualization of clusters and high scores on evaluation metrics suggest that the model successfully identified distinct customer groups.

However, there are some limitations to consider:

1. The K-Means algorithm assumes spherical clusters and equal variance, which may not always fit the true structure of the data.
2. Scalability issues may arise with very large datasets or those with many features.
3. The analysis is based solely on income and spending score; including additional features could provide more comprehensive segmentation.

Despite these limitations, the model provides valuable insights for targeted marketing and customer relationship strategies. Understanding the distinct profiles of each cluster can help tailor approaches to maximize customer satisfaction and loyalty.

Chapter 7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

This project successfully utilized K-Means clustering to segment customers based on their income and spending score, providing actionable insights for targeted marketing and customer relationship management. The analysis revealed distinct customer segments, each with unique characteristics, which can be leveraged to tailor marketing strategies and improve customer engagement.

Key Findings:

1. Effective Segmentation: The K-Means algorithm identified five distinct customer clusters, each exhibiting unique patterns in income and spending behavior. This segmentation allows for a nuanced understanding of customer profiles and enables targeted marketing efforts.
2. Visual Insights: Visualizations, including scatter plots and cluster centroids, clearly demonstrated the distinct groupings and the separation between clusters. This visual representation supports the effectiveness of the clustering process and aids in interpreting the results.
3. Evaluation Metrics: Metrics such as silhouette score and Calinski-Harabasz score validated the quality of the clusters, confirming that the chosen number of clusters effectively captures the underlying structure of the data.

Future improvements could include:

- Exploring other clustering techniques such as DBSCAN or hierarchical clustering for handling more complex datasets.
- Incorporating additional features like age and gender for more nuanced insights.
- Investigating the model's performance on larger and more diverse datasets.

Overall, the K-Means clustering model proves to be a useful tool for customer segmentation, offering actionable insights for marketing strategies while leaving room for further refinement and expansion.

7.2 Future Scope

While the project has achieved promising results, there are several areas for future enhancement and exploration:

7.2.1 **Feature Expansion:** Including additional features such as age and gender in the clustering process could enhance the segmentation and provide more detailed customer profiles.

7.2.2 **Alternative Clustering Methods:** Exploring other clustering algorithms, such as DBSCAN or hierarchical clustering, might offer improvements in handling non-spherical clusters or varying densities in the data.

7.2.3 **Dynamic Segmentation:** Implementing dynamic segmentation approaches that update clusters based on changing customer behavior over time could provide more up-to-date and relevant insights.

7.2.4 **Integration with CRM Systems:** Developing integrations with Customer Relationship Management (CRM) systems to automatically update and utilize segmentation data in real-time customer interactions.

7.2.5 **Predictive Modeling:** Building predictive models based on the identified segments to forecast future customer behavior and lifetime value.

7.2.6 **Cross-Channel Analysis:** Incorporating data from multiple channels (e.g., online, in-store, mobile) to create a more comprehensive view of customer behavior across different touchpoints.

7.2.7 **Personalization Engines:** Developing personalization engines that leverage the segmentation results to deliver tailored content and offers to individual customers.

By addressing these areas, the customer segmentation system can be further refined and expanded to provide even greater support to businesses in understanding and engaging their customers. The ongoing development and application of data-driven approaches in customer analysis hold immense potential for improving marketing effectiveness and customer satisfaction across various industries.

REFERENCES

[1] V. Dawane, P. Waghodekar, and J. Pagare, "RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention," in *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, 2021.

[2] K. Tabianan "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, pp. 7243, 2022.

[3] I. K. Rachmawati, "Customer Segmentation using K-means Clustering," *ResearchGate*, 2019.

[4] Y. C. Li, X. Q. Tian, D. Feng, J. Y. Mu, and L. Weisong, "Customer Segmentation using K-means Clustering and the Adaptive Particle Swarm Optimization Algorithm," *Expert Systems with Applications*, vol. 175, pp. 114812, 2021.

[5] A. Kumar and R. Singh, "Customer Segmentation using K-means Clustering," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 10, no. 4, pp. 1-5, 2020.

[6] P. Gupta and R. S. Dubey, "A Comparative Study on Customer Segmentation Techniques," *IEEE Transactions on Cybernetics: Systems*, vol. 50, no. 1, pp. 66-77, 2020.

[7] Y. Zhao and J. Wang, "Enhancing Customer Segmentation through K-Means Clustering," *Journal of Business Research*, vol. 145, pp. 102-113, 2023.

[8] L. Chen "Application of K-Means Clustering in Customer Segmentation," *International Journal of Decision Making*, vol. 21, no. 5, pp. 1235-1250, 2022.

[9] H. Lee and S. H. Kim, "Utilizing K-Means Clustering for Targeted Marketing Strategies," *Journal of Marketing Analytics*, vol. 9, no. 3, pp. 145-158, 2021.

[10] Smith and T. Jones, "K-Means Clustering for E-commerce Customer Segmentation", *Journal of Retailing and Consumer Services*, vol. 72, pp. 102-110, 2024.