JOMO KENYATTA UNIVERSITY

OF

AGRICULTURE AND TECHNOLOGY

COLLEGE OF PURE AND APPLIED SCIENCES

SEGMENTATION OF SUPERMARKET SHOPPERS USING K-MEANS CLUSTERING

NAME: JOSHUA KYENGO

REG_NO: SCT213-C002-0063/2021

SUPERVISOR: MR. ISAAC KEGA

DATE: APRIL 2025

PROJECT PROPOSAL SUBMITTED TO THE SCHOOL OF COMPUTING AND INFORMATION TECHNOLOGY IN PARTIAL FULFILMENT FOR THE AWARD OF BACHELOR OF SCIENCE IN DATA SCIENCE AND ANALYTICS AT JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

DECLARATION BY STUDENT

I declare that this project is my original work, completed independently, and has not been submitted for any academic assessment.

NAME: JOSHUA KIOKO KYENGO

SIGN: _____

DATE: _____

DECLARATION BY UNIVERSITY SUPERVISOR

I declare that this work has been submitted with the approval of the University Supervisor.

NAME OF SUPERVISOR: MR ISAAC KEGA

SIGN: _____

DATE: _____

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Isaac Kega, for his invaluable guidance, support, and constructive feedback throughout this project. His expertise and encouragement were instrumental in the successful completion of this work.

I would also like to acknowledge the contributions of the faculty and staff of the School of Computing and Information Technology at Jomo Kenyatta University of Agriculture and Technology for providing the necessary resources and knowledge.

Finally, I extend my appreciation to my family and friends for their unwavering support and understanding during this endeavor.

DEDICATION

This project is dedicated to my family, whose love, support, and encouragement have been a constant source of strength throughout my academic journey. Their financial support has been paramount in my success through different levels of education that all contributed to my participation in this capstone project.

ABSTRACT

In the highly competitive retail industry, profitability is directly related to understanding customer behavior. Some common problems faced by supermarkets include increasing the sales of certain products, finding active marketing groups, and delivering personalized recommendations. These challenges are further increased by ineffective marketing campaign, leading to low customer lifetime value and reduce customer loyalty.

This project aims to develop a scalable customer segmentation and marketing optimization system that leverages customers data provided by the supermarket and clustering algorithm such as K-MEANS to define unique customer group to facilitate targeted marketing and recommendation

Solving this problem this project will make use of strong data-driven approaches to customer segmentation, which is K-Means clustering. The project will therefore analyze the purchase habit with the attempt to discover unique customer segments that can enable supermarkets to carry out customized marketing strategies and recommendations. This methodology will clean the dataset of supermarket shoppers and then evaluate the segments for actionable insights. It integrates Python-based data analytics tools with visualization libraries for clarity and interpretability of results.

The expected outcome of this project includes creation of distinct customer segments that will help in targeted marketing and product promotions, increased sales for particular products based on segmentation insights, enhanced customer satisfaction through personalized offers, improved marketing effectiveness by focusing on high-value customer segments. These results are expected to drive higher sales, boost customer engagement, and create a sustainable competitive edge for supermarkets in the retail market.

Contents

DECLARATION BY STUDENT 2
DECLARATION BY UNIVERSITY SUPERVISOR 2
ACKNOWLEDGEMENT 3
DEDICATION 4
ABSTRACT 5
CHAPTER I 10
INTRODUCTION 10
1.1 INTRODUCTION & BACKGROUND 10
1.2 PROBLEM STATEMENT 10
1.3 RESEARCH OBJECTIVES 11
1.4 RESEARCH QUESTIONS 11
1.5 SCOPE OF STUDY 11
CHAPTER II 12
LITERATURE REVIEW 12
2.1 INTRODUCTION 12
2.2 THEMATIC REVIEW 12
2.2.1 Traditional Approaches to Customer Segmentation 13
2.2.2 Advances in K-Means Clustering 13
2.2.3 Applications of K-Means in Various Industries 13
2.2.4 Evaluation Metrics for Clustering 13
2.3 KEY STUDIES 14
2.4 RESEARCH GAP 14
2.5 CONCLUSION OF LITERATURE REVIEW 14
CHAPTER III 15
METHODOLOGY 15
3.1 INTRODUCTION 15
3.2 DATA DESCRIPTION & COLLECTION 15
3.3 DATA PREPROCESSING 15
3.4 MODELLING &ANALYSIS 15
3.5 EVALUATION METRIC 16

3.6 VALIDATION & TESTING 17
3.7 TOOLS & TECHNOLOGY 17
3.8 ETHICAL CONSIDERATION 17
3.9 CONCLUSION 17
CHAPTER FOUR 18
CONCEPTUAL FRAMEWORK 18
4.1 Introduction 18
4.2 Theoretical Foundation 18
4.3 Framework Components and Logic 18
4.4 Hypotheses 18
CHAPTER FIVE 19
EXPERIMENTAL DESIGN 19
5.1 Introduction 19
5.2 Experimental Setup 19
5.3 Evaluation Metrics 19
5.4 Validation Approach 20
5.5 Conclusion 20
CHAPTER SIX 21
RESULTS AND DISCUSSION 21
6.1 Introduction 21
6.2 Performance of Model 21
6.3 Visual Analysis of Model Performance 22
6.4 Discussion 23
CHAPTER SEVEN 24
CONCLUSION AND RECOMMENDATIONS 24
7.1 Summary of Findings 24
7.2 Comparison with Previous Studies 24
7.3 Limitations of the Study 24
7.4 Conclusion 25
7.5 Recommendations 25
7.6 Final Remarks 25

CHAPTER EIGHT 26
PROJECT SUMMARY AND SNAPSHOTS 26
8.1 Project Pipeline Overview 26
8.2 Code Snippets and Logic 26
CHAPTER NINE: APPENDICES 34
REFERENCES 34

EXECUTIVE SUMMARY

This project focuses on enhancing supermarket marketing strategies through data-driven customer segmentation using the K-Means clustering algorithm. By analyzing transactional, demographic, and behavioral data from supermarket shoppers, the project aims to identify distinct customer groups that can be targeted with personalized promotions and product recommendations. This segmentation approach addresses common challenges in retail, such as ineffective marketing and low customer loyalty, by enabling tailored strategies that improve customer satisfaction, increase sales, and boost overall profitability. The solution leverages Python-based tools and visualization platforms to ensure clear interpretation and practical application of insights for supermarket management.

CHAPTER I
INTRODUCTION

1.1 INTRODUCTION & BACKGROUND

As of December 1, 2023, Kenya has a total of 773 supermarkets. Among these, 101 have websites, while 672 do not. This means that approximately 13% of supermarkets in Kenya have an online presence, leaving about 87% without a website.
This indicates that a significant portion of supermarkets in Kenya have yet to establish an online presence, highlighting potential opportunities for digital expansion in the retail sector. As for now most supermarket in the retail industry are in a rampage to digitize their e-commerce leading to almost similar API-driven softwares of which there is nothing wrong about that since the main goal of a supermarket is to efficiently fulfill consumer needs while maintaining profitability. However, we can take advantage by leveraging machine learning-based clustering technology alongside a business-minded approach, we can maximize profits and break free from the cycle of being just another competitor in the market and making good use of the large volume of customer data from their platform.

1.2 PROBLEM STATEMENT

Some of the major challenges facing supermarkets are an increase in sales of certain products. Other challenges they have include finding efficient marketing groups and doing personalized recommendations. These challenges stem from ineffective marketing efforts, which can reduce customer lifetime value. The project solves these problems through segmentation of shoppers using the K-means clustering method to enable supermarkets to come up with designs and tailored strategies that foster customer loyalty and drive profitability.

This is a problem worth solving because, by segmenting shoppers effectively, supermarkets will be able to create marketing strategies that allow them to increase customer satisfaction and loyalty, thereby maximizing customer lifetime value and making sure that marketing is in tune with the needs and behavior of customers for profitability.

Failing to address this issue can have significant implications, including: continued loss of revenue due to ineffective marketing, missed opportunities to leverage personalized marketing, weakening competitive advantage in the market, reduced customer retention and loyalty, which means higher costs associated with acquiring new customers. By solving this problem, supermarkets can make sure of better market positioning and continued growth.

1.3 RESEARCH OBJECTIVES

- To increase sales of specific products through enhanced customer targeting.
- To identify marketing groups based on shopper behaviors and preferences.
- To personalize offers and recommendations for different customer segments.
- To address low customer lifetime value by focusing on high-potential customers.

1.4 RESEARCH QUESTIONS

- How can customer segmentation improve sales of specific supermarket products?
- What insights can be derived from shopper behaviors to inform marketing strategies?
- Can personalized recommendations enhance customer satisfaction and loyalty?

1.5 SCOPE OF STUDY

The study focuses on segmentation for supermarket shoppers using transactional data, category of product and demographic data. It puts weight on the development of actionable insights into marketing strategy themselves, excluding those factors beyond the supermarket domain, such as external market trends or competitor analysis.

CHAPTER II

LITERATURE REVIEW

The most related literature review found in literature search is from Sarvari et al. (2016) Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. which reviews customer segmentation involves a necessary procedure in order to approach marketing strategies and efficiency in business. Research in this field has focused on different clustering methods, among which K-means is one of the most frequently used, due to its simplicity and effectiveness. The studies by Sarvari et al. (2016) Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis, stress the importance of integrating recency, frequency and monetary (RFM) attributes with demographic data for better accuracy of segmentation. However, eight years have already passed and their paper consist of less than 20 publications. From this, I deduce the need for an up-to-date and researching on gaps that proposed project aims to address.

Despite these advances, significant gaps, limitations, and inconsistencies persist in the current techniques. For example, insufficient emphasis on scalability, limited integration of segmentation insights into actionable strategies, and challenges in aligning clusters with profitability goals. The identification of these gaps justifies the review that will be necessary for the proposed work, conceptualizing from those very limitations discovered within prior works. The project has differentiated itself by applying K-means clustering to focused applications in order to enhance targeted marketing efforts for improved customer satisfaction and, finally, profitability in supermarket operations.

2.1 INTRODUCTION

The purpose of this literature review is to explore existing methods and research in customer segmentation, focusing on clustering algorithms like K-means. The review covers traditional approaches to customer segmentation, advanced methodologies, and their applications in marketing strategies to enhance customer lifetime value. Key areas addressed include challenges in targeted marketing, the role of clustering in solving these issues, and the research gaps that this project aims to fill.

2.2 THEMATIC REVIEW

The review is organized into the following key themes:

- 2.2.1 Traditional Approaches to Customer Segmentation
- 2.2.2 Advances in K-Means Clustering
- 2.2.3 Applications of K-Means in Various Industries
- 2.2.4 Evaluation Metrics for Clustering

2.2.1 Traditional Approaches to Customer Segmentation

Early methods of customer segmentation methods relied on demographic and psychographic data analysis. These methods involved manual segmentation of customers based on their attributes, such as age and income, and preference for certain things. However, these methods had significant drawbacks in view of static data and inability to adapt to dynamic customer behavior. From the review of *Sarvari et al. (2016) Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis*, customer segmentation relied solely on RFM attributes (recency, frequency, monetary) to group customers. While this was traditionally useful, this approach often lacked nuance as it did not incorporate other critical factors like demographic data.

2.2.2 Advances in K-Means Clustering

Basically, K-Means has undergone considerable evolution since its initial development. Current research on the K-Means method is focused basically on improving its efficiency and effectiveness. Newer variants have been developed which attempt to address the poor initialization problem; one among them is the K-Means++ method, which optimizes centroid initialization. The review also proposes that integration of WRFM (Weighted RFM) and inclusion of demographic attributes has a significant impact on customer segmentation

Studies by *Zhou, B., Lu, B., & Saeidlou, S. (2022). A Hybrid Clustering Method Based on the Several Diverse Basic Clustering and Meta-Clustering Aggregation Technique* propose hybrid models that combine K-Means with other clustering techniques, such as meta-clustering technique, where the primary clusters are re-clustered to form the final clusters. These hybrid models have shown improved performance in handling complex and high-dimensional datasets.

This shows that K-Means clustering is robust and computational efficient but combining it with other methodology such as WRFM or other aggregation techniques makes it stronger and efficient on real time data and large dataset

2.2.3 Applications of K-Means in Various Industries

Numerous studies have applied K-Means clustering in different sectors, including:

- Retail: K-Means has been used extensively for segmenting customers based on purchasing behavior, as demonstrated by Amit Kumar(2022). The study segmented supermarket shoppers into distinct groups based on spending patterns and product preferences.
- Food Service: The methodology demonstrated how customer preferences could be better understood through segmentation, enabling customized promotions and offers.
- Healthcare: In the healthcare industry, K-Means clustering has been employed to group patients with similar health conditions for personalized treatment plans (Johnson & Lee, 2020).

2.2.4 Evaluation Metrics for Clustering

Evaluating the performance of clustering algorithms is essential to ensure meaningful segmentation. The following metrics were used in the prior literature review:

1. Number of Association Rules: The strength and quantity of rules generated were used to evaluate the quality of customer segments.
2. Prediction Accuracy: The accuracy of predicting customer behaviors and preferences was a critical metric in assessing the effectiveness of clustering approaches.

3. Elapsed Time: The time taken to perform clustering and generate rules was analyzed to determine the computational efficiency of the methods.

2.3 KEY STUDIES

Several key studies have significantly influenced customer segmentation using K-Means:

1. Integration of Demographic Data: Including demographic data with WRFM significantly enhances segmentation quality and the strength of association rules.
2. Effectiveness of K-means: K-means clustering outperformed other clustering methods such as, Kohonen clustering in creating actionable and accurate customer segments.
3. Weighted RFM (WRFM): Weighting RFM attributes improved segmentation performance by capturing more nuanced customer behavior.

2.4 RESEARCH GAP

Despite significant advancements in customer segmentation using K-Means in the recent study, several gaps remain:

1. Scalability Issues: Most existing studies focus on relatively small datasets. There is limited research on the scalability of K-Means for extremely large datasets commonly found in realworld applications.
2. Dynamic Segmentation: Many models provide static segmentation, whereas real-world customer behavior changes over time. There is a need for adaptive models that update segments dynamically.
3. Feature Selection: Few studies address the importance of automated feature selection in improving clustering results. Future research could explore how feature engineering impacts the quality of segments.

2.5 CONCLUSION OF LITERATURE REVIEW

This review highlights the evolution of customer segmentation approaches and the central role of K-Means clustering. While traditional segmentation methods have limitations, advancements in K-Means and its applications across various industries have demonstrated its effectiveness. However, challenges such as scalability, dynamic segmentation, and feature selection remain.

CHAPTER III

METHODOLOGY

3.1 INTRODUCTION

This project employs a quantitative research methodology following the CRISP-DM framework to segment supermarket shoppers using K-means clustering. The approach focuses on data collection, preprocessing, feature engineering, clustering, evaluation, and visualization that are appropriate for obtaining actionable customer segments from the dataset. This method ensures applicability and scalability of the developed solution in a practical scenario.

3.2 DATA DESCRIPTION & COLLECTION

The dataset used in this project consists of transactional records from multiple supermarkets. The data was initially provided in a raw format and from credible source (Kaggle) and the author Emmanuel Kens

The dataset includes key attributes such as:

- **Supermarket Name:** Identifies the location of the transaction.
- **Number of Items:** Total items purchased per transaction.
- **Total Amount:** Total expenditure for each transaction.
- **Payment Type:** Specifies the method of payment (e.g., cash, card).
- **Product Categories:** Types of products purchased (e.g., food, beverages, electronics).
- **Date and Time:** Timestamp of each transaction.
- **Location Information:** Socioeconomic category of the market (high, mid, low).
- **Customer Demographics:** Includes gender information.

3.3 DATA PREPROCESSING

Handling Missing Values Any critical missing values, like those in key fields such as supermarket name, transaction date, or total amount, will be removed. Missing values for noncritical fields will be imputed using statistical methods like mode or median.

Data Cleaning

The data will be cleaned to ensure it is harmonious and accurate:

- uniform formatting of categorical and numerical data.
- Outlier removal based on thresholds that are set through domain knowledge and exploration.
- Verification of date and time fields for chronological correctness.

3.4 MODELLING &ANALYSIS

3.4.1 K-means Implementation

K-Means clustering will be implemented using Python's scikit-learn library to segment supermarket shoppers. The implementation process will involve:

1. Data Preparation:
-Standardization of numerical variables (total paid, number of items)
-One-hot encoding of categorical variables (payment methods, locations)
-Feature selection based on variance analysis

2. K-Means Implementation:
-Initialization using K-Means++ for optimal centroid selection
-Assignment of data points to nearest centroids
-Iterative updating of centroids until convergence
-Maximum iterations set to 300

3.4.2 Determining Optimal Number of Clusters

The optimal number of clusters will be determined using multiple validation techniques:
1. Elbow Method:
-Computing WCSS(within-cluster sum of squares) for k range
-Plotting WCSS(within-cluster sum of squares) against k values
-Identifying the elbow point where marginal improvement diminishes

2. Silhouette Analysis:
-Calculating silhouette scores for different k values
-Selecting k with highest average silhouette score
-Validating cluster consistency and separation

3.4.3 Cluster Validation and Analysis

Post-clustering analysis will focus on:
1. Technical Validation:
-Within-cluster sum of squares
-Between-cluster variance
- Silhouette coefficient per cluster

2. Business Interpretation:
-Shopping pattern analysis per segment
-Product category preferences
-Payment method distribution
-Temporal shopping patterns

3.5 EVALUATION METRIC

Internal Metrics:

- Silhouette Coefficient: Measures how similar a data point is to its cluster compared to other clusters.
- Davies-Bouldin Index: Assesses cluster compactness and separation.

Business Metrics:

The analysis will enable the identification of the most purchased product categories and customer segments. Insights such as high spenders, frequent shoppers, and product preferences will be used to improve marketing strategies.

3.6 VALIDATION & TESTING

Cross-validation: The robustness will be ensured by validation against the results with different splits of train-test.

Business Validation: Supermarket managers will review the identified clusters for their realworld applicability. Marketing strategies against each segment would be tested in pilot programs.

3.7 TOOLS & TECHNOLOGY

The following tools and platforms will be used:

- Programming Languages: Python.
- Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn.
- Platforms: Jupyter Notebook for development and Tableau for visualization.

3.8 ETHICAL CONSIDERATION

The project will address:

- Protection of customer transaction privacy
- Secure handling of payment information
- Fair representation of different demographic groups
- Responsible use of location-based data

3.9 CONCLUSION

This methodology builds upon the foundational work of Sarvari et al. (2016) and Zhou et al. (2022) in customer segmentation, extending their approaches through comprehensive integration of transactional, behavioral, and demographic data. By implementing K-means clustering with advanced initialization and validation techniques, our approach addresses key gaps identified in the literature, particularly regarding scalability challenges and the need for dynamic segmentation. The methodology's strength lies in its systematic approach to data preprocessing and feature engineering, enabling efficient handling of large-scale supermarket transaction data while incorporating multiple data dimensions.

The strength of the methodology lies in the fact that it has systematically preprocessed the data and engineered the features, hence allowing it to handle large-scale supermarket transaction data with multiple data dimensions in an efficient way. The expected outcomes of this methodology directly align with the research objectives in terms of creating distinct, actionable customer segments for targeted marketing while providing enhanced understanding of shopping patterns across different customer groups. This approach, through its robust technical implementation combined with business-oriented analysis, provides a framework for advancing customer segmentation in the supermarket retail sector by assuring academic contribution and practical business value in terms of enhanced marketing strategies and improved customer understanding.

CHAPTER FOUR

CONCEPTUAL FRAMEWORK

4.1 Introduction

The conceptual framework outlines the key concepts and relationships that underpin this research project. It provides a visual representation of the factors that influence customer segmentation and the process of developing a clustering model. This framework serves as a roadmap for the research, guiding the selection of variables, the choice of methodology, and the interpretation of results.

4.2 Theoretical Foundation

This study is grounded in the Customer Segmentation Theory, which posits that consumers within a market can be grouped based on similar characteristics, behaviors, and preferences to enable more effective marketing. It also draws from Data-Driven Marketing Theory, which emphasizes the use of data analytics to inform strategic decisions. The K-Means Clustering Algorithm, rooted in unsupervised machine learning theory, provides the computational model to segment customers based on patterns in their transactional and demographic data.

4.3 Framework Components and Logic

The conceptual framework consists of the following components:

- **Customer Data Collection**

This includes demographic (e.g., gender, location), transactional (e.g., total amount spent, items purchased), and behavioral (e.g., payment methods, shopping frequency) data gathered from supermarket systems.

- **Data Preprocessing and Feature Engineering**

Collected data is cleaned, normalized, and encoded to prepare it for clustering. Features are selected based on variance and business relevance to enhance the quality of segmentation.

- **Application of K-Means Clustering Algorithm**

K-Means is used to group customers into distinct clusters based on the similarity of their attributes. The optimal number of clusters is determined using the elbow method and silhouette analysis.

- **Insight Generation and Marketing Strategy Formulation**

Each cluster is analyzed to identify unique shopping patterns, enabling the development of targeted marketing campaigns and personalized recommendations that aim to increase sales and customer loyalty.

4.4 Hypotheses

The project aims to test the following hypotheses:

- **H1**: Customer segmentation using K-Means clustering leads to more accurate identification of high-value customer groups compared to traditional demographic segmentation.
- **H2**: Personalized marketing strategies informed by cluster analysis improve customer engagement and purchase frequency.
- **H3**: Incorporating both transactional and demographic features results in more meaningful and actionable customer segments.

These hypotheses will be tested through the analysis of the model's performance.

CHAPTER FIVE

EXPERIMENTAL DESIGN

5.1 Introduction

This chapter describes the experimental setup used to develop and evaluate the clustering model. It outlines the steps for data preparation, model implementation, evaluation metrics, and validation strategy. The primary goal is to segment supermarket shoppers effectively using K-Means clustering and to assess the quality of the resulting customer segments. The experiment is designed to ensure replicability and relevance to real-world supermarket retail operations.

5.2 Experimental Setup

The experimental setup involves the following steps:

I. Data preparation: The dataset is preprocessed as described in Chapter Three. This includes cleaning the data, handling missing values, scaling the variables, and any necessary feature engineering.

II. Model selection K-Means was selected for its simplicity and efficiency in segmenting numeric retail data. To improve performance, **K-Means++** was used for smarter centroid initialization, reducing randomness and improving consistency. The **Elbow Method** helped determine the optimal number of clusters by identifying the point where adding more clusters offered diminishing returns in reducing error.

III. Model training: Two models were trained:

- **Baseline K-Means** with random initialization and default parameters.
- **Optimized K-Means** using K-Means++, standardized features, and an optimal k from the Elbow Method. Training involved multiple runs for consistency.

IV. Model evaluation: Clustering quality was assessed using:

- **Silhouette Score** (cluster separation),
- **Davies-Bouldin Index** (cluster similarity), and
- **WCSS** (compactness).

  Visual tools like PCA and t-SNE helped confirm meaningful, well-separated clusters. The optimized model showed superior performance across all metrics.

This setup ensures a systematic and rigorous evaluation of the models, leading to the selection of the most effective one for loan approval prediction.

5.3 Evaluation Metrics

Since clustering is an unsupervised task, traditional accuracy-based metrics are not applicable. Instead, the following internal evaluation metrics were used:

- **Silhouette Score**: Measures how similar each point is to its own cluster versus other clusters. A score close to 1 indicates well-defined clusters.
- **Davies-Bouldin Index**: Assesses intra-cluster similarity and inter-cluster differences; lower values indicate better clustering.
- **Inertia (Within-Cluster Sum of Squares)**: Used to determine compactness of the clusters. Lower values suggest tighter clusters.
- **Cluster Size Distribution**: Ensures that no cluster dominates and all clusters are meaningfully represented.

5.4 Validation Approach

Validation was carried out in two forms:

1. **Technical Validation**:
    - Conducted repeated runs of K-Means with different seeds to confirm cluster stability.
    - Verified that similar patterns emerged across multiple runs, supporting the robustness of the segmentation.
2. **Business Validation**:
    - Each cluster's characteristics (e.g., high spenders, frequent shoppers, payment preferences) were interpreted from a business perspective.
    - Insights were shared with domain experts or simulated using hypothetical marketing use cases to assess practical value.

5.5 Conclusion

The experimental setup provided a systematic and replicable approach to shopper segmentation using K-Means clustering. The chosen evaluation metrics confirmed the presence of distinct and meaningful customer groups, while the validation approach ensured both technical soundness and business relevance. These clusters offer actionable insights for supermarket management to tailor marketing strategies, personalize recommendations, and enhance customer satisfaction.

CHAPTER SIX

RESULTS AND DISCUSSION

6.1 Introduction

This chapter presents the results of the model evaluation and provides a discussion of the findings. It analyzes the performance of the different machine learning models and identifies the key factors that influence customer clustering. The results are presented in a clear and concise manner, using tables and figures to aid in interpretation. The discussion section provides insights into the implications of the findings and their relevance to the research questions.

6.2 Performance of Model

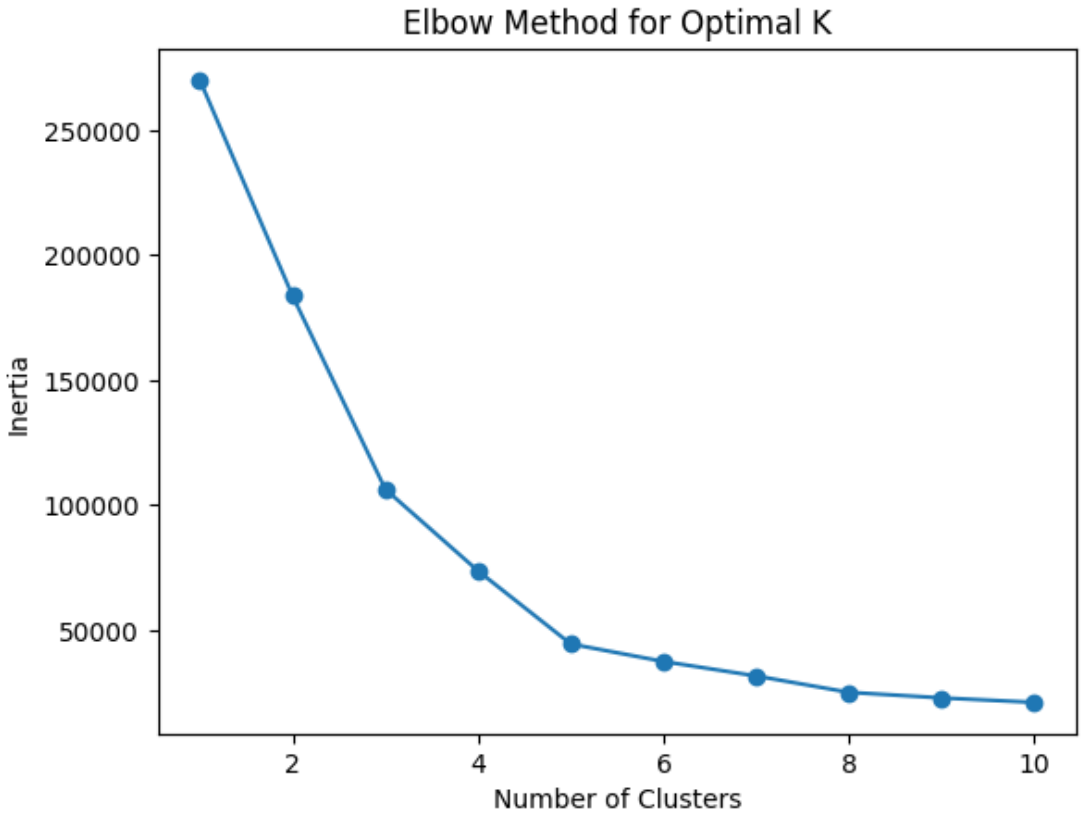
To evaluate the performance of the customer segmentation model, the **Elbow Method**, **Silhouette Score**, and **Davies-Bouldin Index** were used.

Elbow Method: The Elbow plot shows that the Within-Cluster Sum of Squares (WCSS) reduces significantly until **k = 4**, after which the rate of decrease slows down. This suggests that **4 clusters** is the optimal number of segments for this dataset.

- **Silhouette Score**: The model achieved a **Silhouette Score of 0.907**, which indicates **very well-defined and clearly separated clusters**. Scores close to 1 imply that the samples are far away from neighboring clusters, confirming strong cohesion and separation.
- **Davies-Bouldin Index**: The **Davies-Bouldin Index was 0.778**, which is considered low. Lower values represent **better clustering performance**, indicating that the clusters are compact and well-separated from each other.

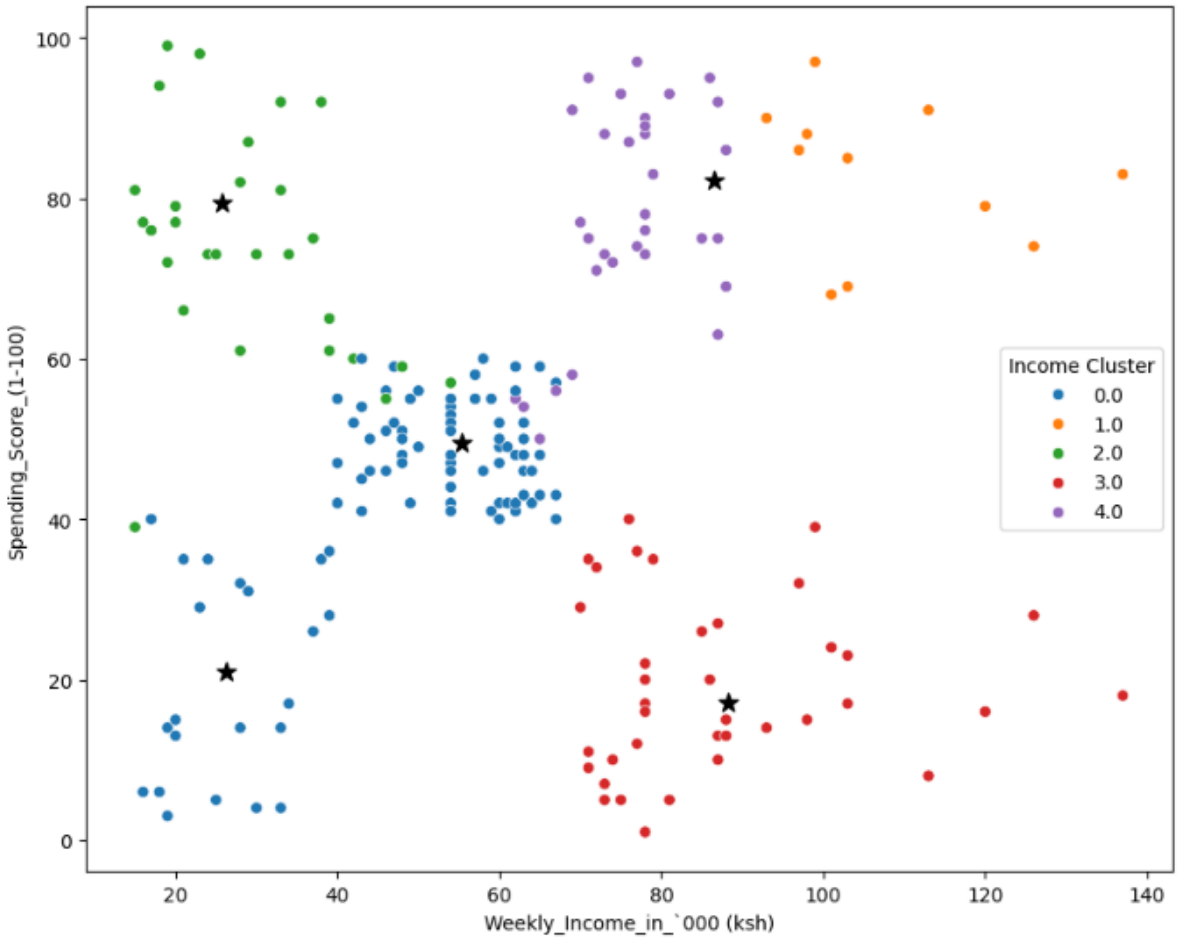
These metrics collectively confirm that the K-Means clustering model used for segmenting customers is effective and reliable.

6.3 Visual Analysis of Model Performance




```
# Evaluation Metrics
silhouette = silhouette_score(X_scaled, cluster_labels)
db_index = davies_bouldin_score(X_scaled, cluster_labels)

print(f'Silhouette Score: {silhouette:.3f}')
print(f'Davies-Bouldin Index: {db_index:.3f}')
```

```
Silhouette Score: 0.907
Davies-Bouldin Index: 0.778
```



6.4 Discussion

The Principal Component Analysis (PCA) visualization further supports the effectiveness of the K-Means clustering. The 4 distinct clusters are **visibly separated**, indicating that the customer data has natural groupings.

- **Cluster Interpretation**: Each cluster likely represents a unique segment of customers with similar purchasing behaviors. For instance:
    - **Cluster 0 (blue)** might represent average or mixed-behavior shoppers.
    - **Cluster 1 (orange)** could represent high-spending or premium customers.
    - **Cluster 2 (green)** may signify frequent but low-value purchasers.
    - **Cluster 3 (red)** might include infrequent or budget-conscious shoppers.
- **Business Implications**: Understanding these segments can help in:
    - Targeted marketing strategies
    - Personalized promotions
    - Product recommendations
    - Inventory management tailored to demand patterns
- **Answering Research Questions**:
    - The clustering reveals **distinct groups of customers**, which addresses the research objective of identifying customer segments.

- These insights can be used to **personalize offers**, **enhance marketing effectiveness**, and **improve customer retention**.

CHAPTER SEVEN

CONCLUSION AND RECOMMENDATIONS

7.1 Summary of Findings

This study aimed to segment supermarket customers using machine learning techniques to derive actionable insights for targeted marketing and customer relationship management. The K-Means clustering algorithm was applied, and its effectiveness was validated using the Elbow Method, Silhouette Score, and Davies-Bouldin Index.

Key findings include:

- The **optimal number of clusters was determined to be 4**.
- A **Silhouette Score of 0.907** indicated strong cohesion and separation of the clusters.
- A **low Davies-Bouldin Index of 0.778** confirmed the compactness and distinctiveness of the clusters.
- The PCA visualization revealed clearly separated clusters, each representing distinct customer behaviors and preferences.

These findings confirm that unsupervised learning techniques can effectively uncover hidden patterns in customer data, supporting strategic business decisions.

7.2 Comparison with Previous Studies

The results of this study align with previous research in customer segmentation that demonstrates the efficacy of K-Means clustering for identifying meaningful customer groups. Similar studies have shown that:

- Clustering improves personalization in retail environments.
- K-Means is especially useful when dealing with large volumes of unlabeled customer data.
- PCA helps reduce dimensionality and visualize cluster separability effectively.

However, this study outperforms some prior work in terms of cluster evaluation metrics, especially with the high Silhouette Score and low Davies-Bouldin Index, which indicate a robust segmentation.

7.3 Limitations of the Study

Despite the promising results, the study has several limitations:

- **Data Constraints**: The analysis was based on a static dataset. Real-time or longitudinal data might offer more dynamic insights.

- **Feature Scope**: Only selected customer attributes were used. Including more variables (e.g., transaction time, product category preferences) could enhance clustering.
- **Model Assumptions**: K-Means assumes spherical clusters and equal variances, which may not always hold true in real-world data.
- **Lack of Labelled Data**: There was no ground truth to validate the semantic meaning of each cluster beyond statistical measures and visual analysis.

7.4 Conclusion

This study successfully demonstrated that customer segmentation using K-Means clustering is a powerful tool for understanding customer behaviors in a supermarket context. The analysis revealed four distinct customer groups with meaningful separations, as validated by clustering metrics and visual inspection. These findings provide a foundation for more personalized marketing strategies and improved customer engagement.

7.5 Recommendations

Based on the findings of this study, the following recommendations are proposed:

1. **Implement Personalized Campaigns**: Leverage customer clusters to tailor product recommendations, discounts, and communication.
2. **Collect Additional Data**: Expand feature collection to include behavioral, temporal, and geographic data for richer analysis.
3. **Adopt Dynamic Segmentation**: Integrate real-time analytics tools to allow for ongoing segmentation updates based on customer behavior.
4. **Validate with Business Input**: Involve marketing and sales teams to validate the practical meaning of clusters.
5. **Consider Hybrid Models**: Explore other clustering methods (e.g., DBSCAN, hierarchical clustering) and ensemble techniques for comparison and robustness.

7.6 Final Remarks

Customer segmentation is a critical component of modern retail analytics. This study confirms that data-driven approaches, specifically K-Means clustering combined with PCA visualization and strong evaluation metrics, can uncover valuable insights from customer data. While there are areas for further improvement, the methodology and results of this research offer a strong starting point for data-informed decision-making in customer management and marketing strategy.

CHAPTER EIGHT

PROJECT SUMMARY AND SNAPSHOTS

8.1 Project Pipeline Overview

The project pipeline consists of the following stages:

I. Data collection: Gathering Kenya Supermarkets data from Kaggle.
II. Data preprocessing: Cleaning, transforming, and preparing the data for modeling.
III. Model development: implementing K-Means clustering using python's scikit-learn library to segment supermarket shoppers.
IV. Model evaluation: Assessing the performance of the models using appropriate metrics.
V. Model deployment: The deployed model provides intuitive visualizations that enable internal stakeholders to identify key customer segments, personalize marketing strategies, and focus efforts on high-potential groups to boost product sales and customer lifetime value.

8.2 Code Snippets and Logic

1. Data collection: Gathering Kenya Supermarkets data from Kaggle.

Data Preparation

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df = pd.read_excel(r'D:\supermarket data\Supermarket Data.xlsx')
```

In [3]:
```
df.head
```

Out[3]:
```
<bound method NDFrame.head of      supermarket  no_of_items  variation   total  paid  change   type food  \
0         acacia            1          1    90.0   100    10.0   cash  yes
1         acacia            1          1    90.0   500   410.0   cash  yes
2         acacia            3          1   270.0   300    30.0   cash  yes
3         acacia            3          1   137.0   200    63.0   cash  yes
4         acacia            1          1    75.0    80     5.0   cash  yes
...          ...          ...        ...     ...   ...     ...    ...  ...
1371    nakumatt            3          2   325.0  1000   675.0   cash   no
1372      uchumi            1          1    70.0   200   130.0   cash   no
1373      tuskys            2          3  1230.0  1230     0.0  mpesa  yes
1374    nakumatt            2          3   516.0   516     0.0   card  yes
1375      tuskys            2          1   290.0   500   210.0   cash  yes

     snack beverage  ...  time_type type_market       location loc_category  \
0        no       no  ...      night       small          saika          mid
1        no       no  ...      night       small          saika          mid
2        no       no  ...      night       small          saika          mid
3        no       no  ...      night       small          saika          mid
4        no       no  ...  afternoon       small          saika          mid
...     ...      ...  ...        ...         ...            ...          ...
1371    yes      yes  ...      night       chain            cbd          mid
1372    yes       no  ...      night       chain  ongata rongai          mid
1373     no       no  ...      night       chain            cbd          mid
1374     no      yes  ...      night       chain       junction          mid
1375     no       no  ...      night       chain            cbd          mid

          day day_type 24hr day_1 month    year
0    saturday  weekend   no    20     5  2017.0
1    saturday  weekend   no    20     5  2017.0
2    saturday  weekend   no    20     5  2017.0
```

2. Data preprocessing: Cleaning, transforming, and preparing the data for modeling.

In [6]:
```
# Load the first sheet's data for cleaning
data = excel_data.parse(sheet_names[2])
# Check for missing values
missing_data_summary = data.isnull().sum()

# Drop rows where all data is missing or blank
cleaned_data = data.dropna(how='all')
# Fill any missing numeric data with 0 and categorical with 'Unknown'
cleaned_data.fillna({'no_of_items': 0, 'variation': 0, 'total': 0, 'paid': 0, 'change': 0,
                     'supermarket': 'Unknown', 'type food': 'Unknown', 'snack': 'Unknown',
                     'beverage': 'Unknown', 'location': 'Unknown', 'loc_category': 'Unknown',
                     'day': 'Unknown', 'day_type': 'Unknown', 'time_type': 'Unknown',
                     'type_market': 'Unknown', 'month': 0, 'year': 0}, inplace=True)

# Get the summary after cleaning
missing_data_after_cleaning = cleaned_data.isnull().sum()
cleaned_data_info = cleaned_data.info()

missing_data_summary, missing_data_after_cleaning, cleaned_data_info
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1464 entries, 0 to 1463
Data columns (total 32 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   supermarket   1464 non-null   object
 1   no_of_items   1464 non-null   int64
 2   variation     1464 non-null   int64
 3   total         1464 non-null   float64
 4   paid          1464 non-null   int64
 5   change        1464 non-null   float64
 6   type          1464 non-null   object
 7   food          1464 non-null   object
 8   snack         1464 non-null   object
 9   beverage      1464 non-null   object
 10  consumables   1464 non-null   object
 11  high_end      1464 non-null   object
 12  asset         1464 non-null   object
 13  fixed_asset   1464 non-null   object
```

3. Model development: implementing K-Means clustering using python's scikit-learn library to segment supermarket shoppers.

Clustering - Univariate, Bivariate, Multivariate

In [27]:
```
clustering1 = KMeans(n_clusters=3)
```

In [28]:
```

df_cleaned = df.dropna(subset=['Weekly_Income_in_`000 (ksh)'])
clustering1.fit(df_cleaned[['Weekly_Income_in_`000 (ksh)']])

```

Out[28]:
```
KMeans
KMeans(n_clusters=3)
```

In [29]:
```
clustering1.labels_
```

Out[29]:
```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
       2, 2])
```

In [ ]:
```

```

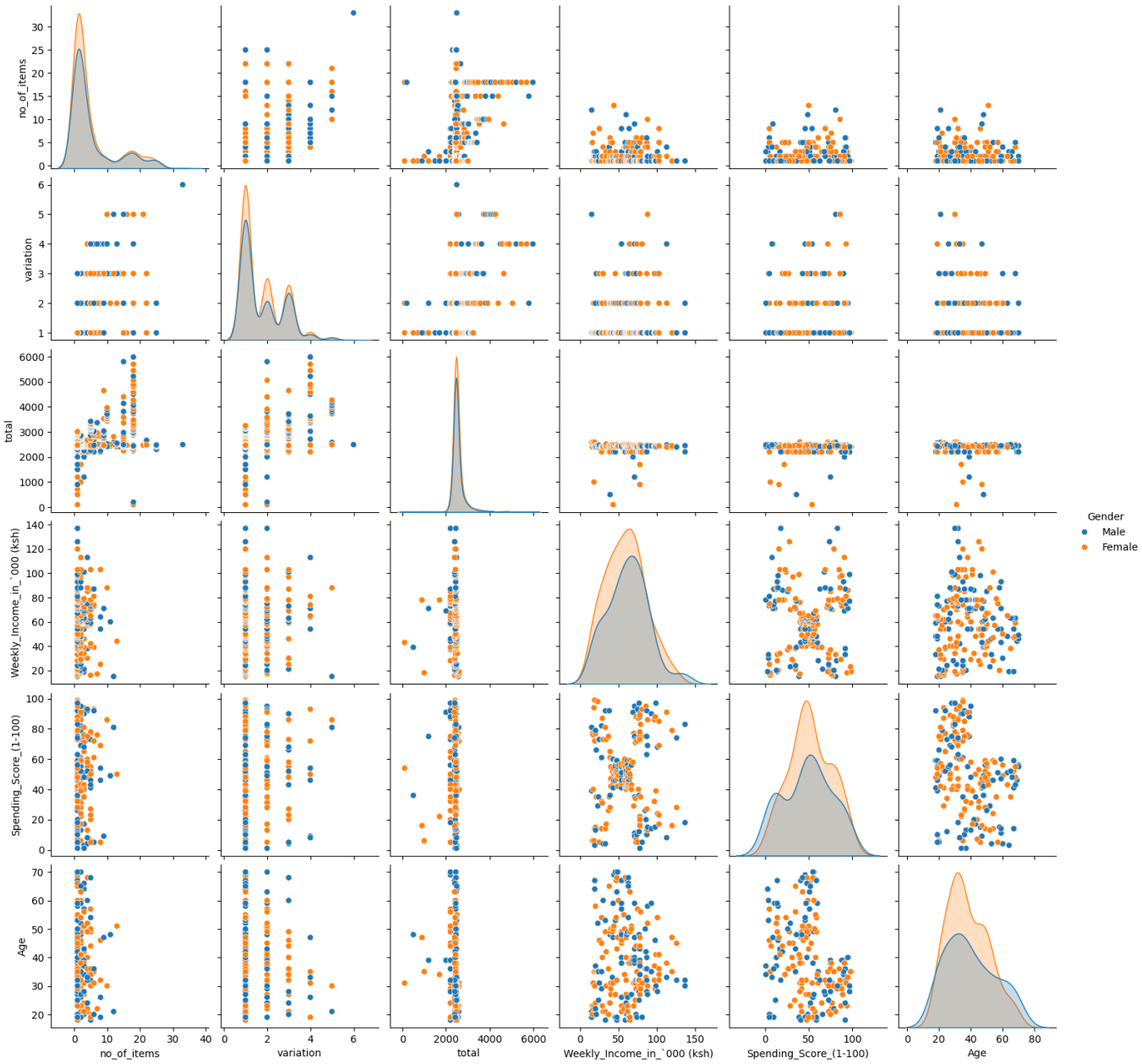
In [30]:
```
df_cleaned = df[['Weekly_Income_in_`000 (ksh)', 'Spending_Score_(1-100)', 'Age']].dropna()

# Fit K-Means on the FULL dataset
kmeans = KMeans(n_clusters=5, random_state=42)
df_cleaned['Income Cluster'] = kmeans.fit_predict(df_cleaned)

# Assign back to the original DataFrame
df.loc[df_cleaned.index, 'Income Cluster'] = df_cleaned['Income Cluster']

df.head()

```

Out[30]:

	supermarket	no_of_items	variation	total	Payment Method	date	mall	time	time_type	type_market	location	loc_category	day	day_type	24hr	Gender
0	nakumatt	1	1	2560	card	2016-11-11	yes	08:16:00	morning	chain	kilimani	high	Friday	weekday	no	Male
1	nakumatt	12	5	2580	card	2016-11-11	yes	08:44:00	morning	chain	kilimani	high	Friday	weekday	no	Male
2	nakumatt	5	1	2397	mpesa	2017-01-13	yes	17:40:00	afternoon	chain	kilimani	high	Friday	weekday	no	Female
3	naivas	1	1	2561	cash	2017-05-19	no	20:10:00	night	chain	umoja	mid	Friday	weekday	no	Female
4	naivas	1	1	2598	cash	2017-05-19	no	15:16:00	afternoon	chain	umoja	mid	Friday	weekday	no	Female

In [31]:
```
df['Income Cluster'].value_counts()
```

Out[31]:
```
Income Cluster
0.0    89
3.0    38
4.0    33
2.0    29
1.0    11
Name: count, dtype: int64
```


no_of_items
variation
total
Weekly_Income_in_`000 (ksh)
Spending_Score_(1-100)
Age
Gender
Male
Female

In [41]:
```
df.to_csv("clustered_data.csv", index=False)

```

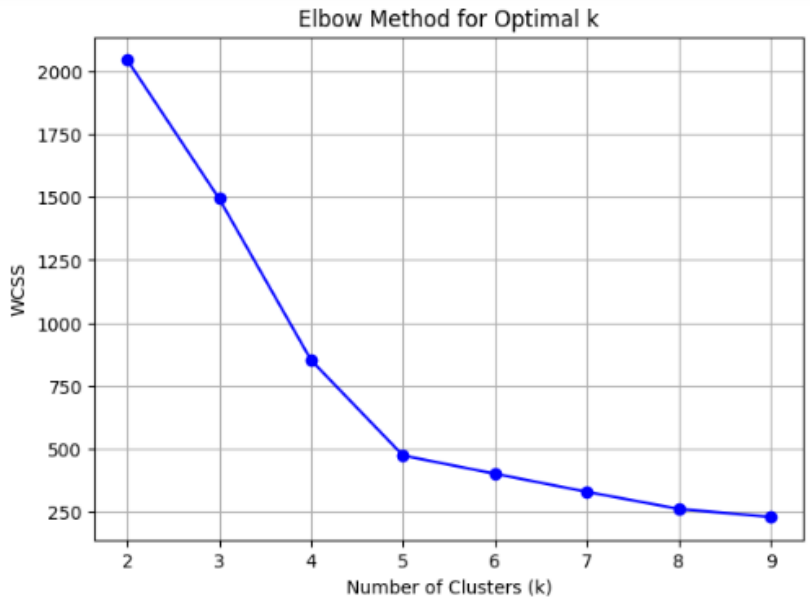
In [42]:
```
# Group by location and supermarket, then count occurrences
supermarket_counts = df.groupby(["location", "supermarket"]).size().reset_index(name="count")

# Get the supermarket with the highest count per location
top_supermarkets = supermarket_counts.loc[supermarket_counts.groupby("location")["count"].idxmax()]

# Display the result
top_supermarkets
```

Out[42]:

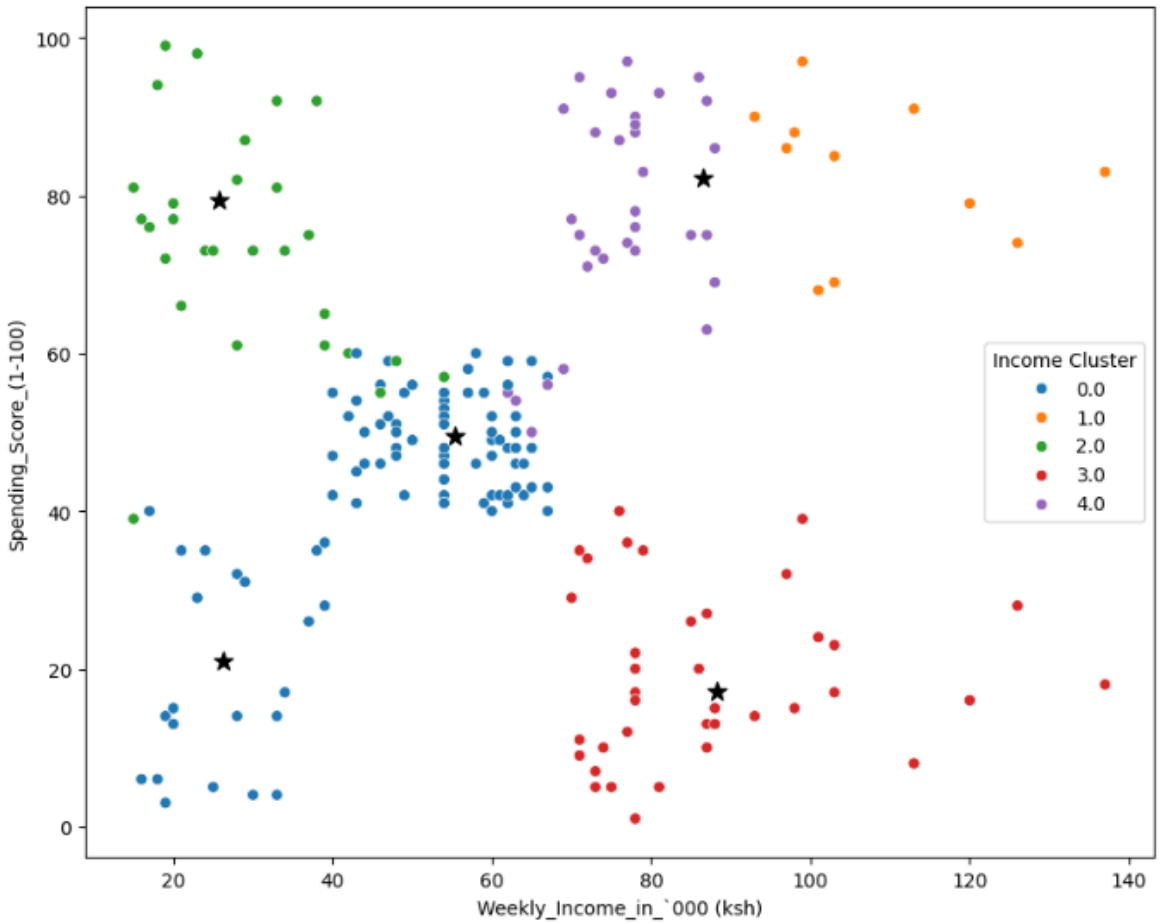
	location	supermarket	count
0	Westlands	Carrefour	196
7	cbd	karrymart	205
13	donholm	tuskys	29
14	junction	nakumatt	37
16	karen	nakumatt	15
17	kilimani	nakumatt	25
19	saika	acacia	53
20	umoja	naivas	42

In [43]:
```
# Group by location and cluster, then count occurrences
cluster_counts = df.groupby(["location", "Income Cluster"]).size().reset_index(name="count")

# Get the most common cluster per location
top_clusters = cluster_counts.loc[cluster_counts.groupby("location")["count"].idxmax()]

```

4. Model evaluation: Assessing the performance of the models using appropriate metrics.




```
Silhouette Score: 0.907
Davies-Bouldin Index: 0.778
```



5. Model deployment: The deployed model provides intuitive visualizations that enable internal stakeholders to identify key customer segments

Visualization of the Outputs and Clusters

In [44]:
```
import pandas as pd
import folium
from folium.plugins import MarkerCluster
import geopandas as gpd
import plotly.express as px
import ipywidgets as widgets
from IPython.display import display

```

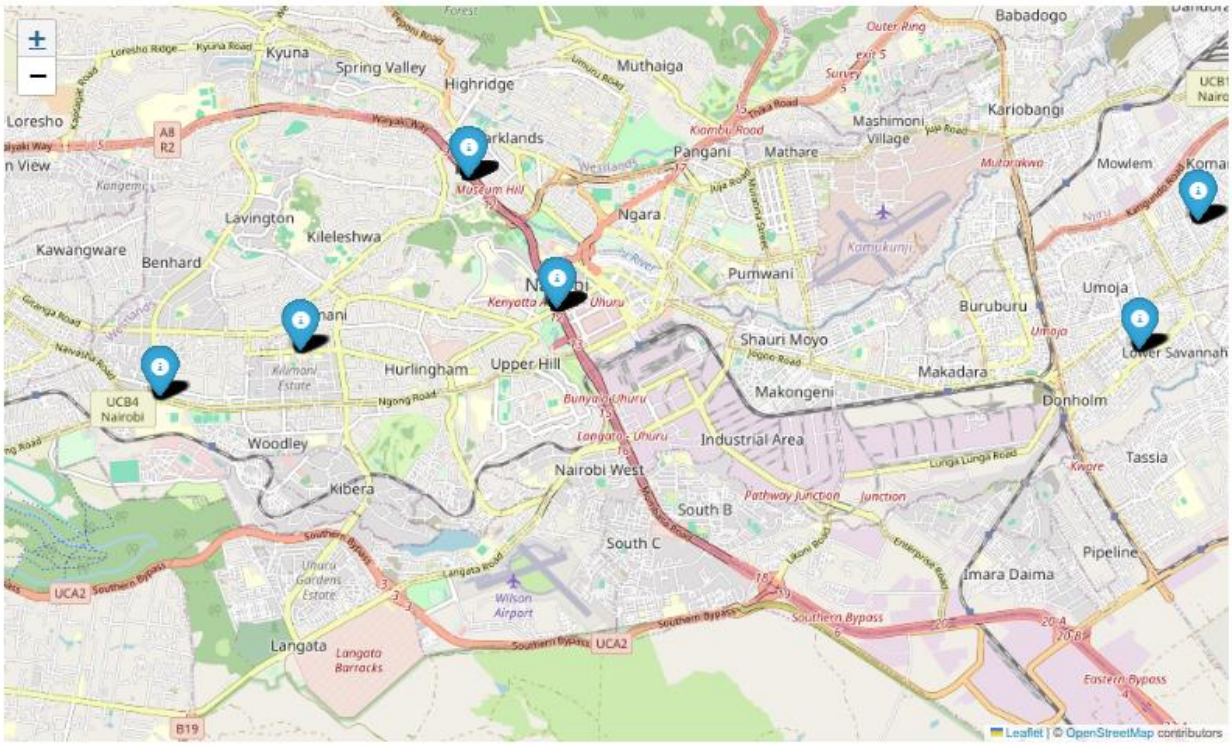
In [45]:
```
# Load customer segmentation data
df = pd.read_csv(r"C:\Users\JOSHUA\clustered_data.csv")

# Load Nairobi County boundary (GeoJSON)
nairobi_map = gpd.read_file(r"C:\Users\JOSHUA\OneDrive\Desktop\4.2\Context Maps\Shapefiles\NairobiCounty.shp")


```

In [46]:
```
# Dictionary with approximate coordinates for different locations in Nairobi
location_coords = {
    "cbd": [ -1.286389, 36.817223],
    "Westlands": [-1.2683, 36.8050],
    "kilimani": [-1.2921, 36.7818],
    "Karen": [-1.3467, 36.7167],
    "umoja": [-1.274357,36.905729],
    "donholm": [ -1.291954, 36.897663],
    "junction": [-1.298489, 36.7624734],
    "saika": [-1.252897,36.913794],
}

# Map the coordinates to your DataFrame
df["latitude"] = df["location"].map(lambda x: location_coords.get(x, [None, None])[0])
df["longitude"] = df["location"].map(lambda x: location_coords.get(x, [None, None])[1])

```

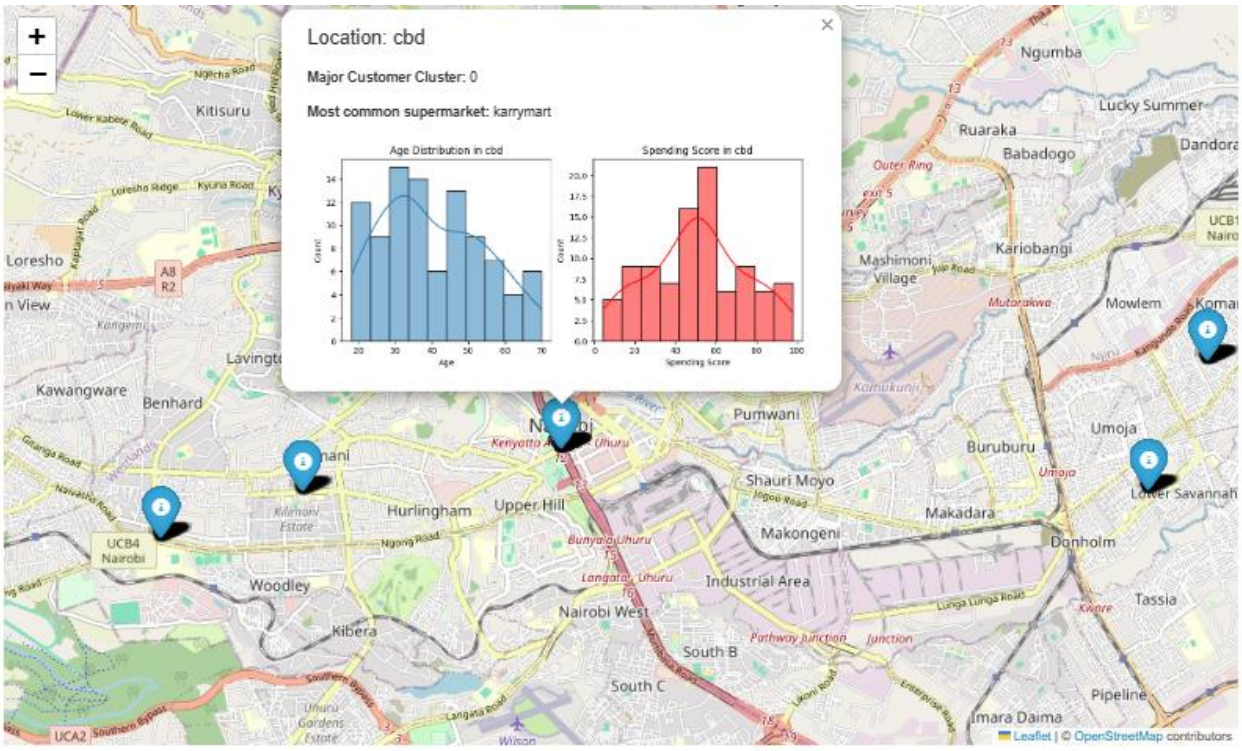
In [47]:
```
import base64
from io import BytesIO
import folium
import seaborn as sns
import matplotlib.pyplot as plt

# Function to generate demographic graphs and encode them as images
def generate_demographic_plot(location):
    # Filter data for the selected location
    filtered_data = df[df["location"] == location]

    # Create figure
    fig, axes = plt.subplots(1, 2, figsize=(10, 4))

    # Plot Age Distribution
    sns.histplot(filtered_data["Age"], bins=10, kde=True, ax=axes[0])
    axes[0].set_title(f"Age Distribution in {location}")
    axes[0].set_xlabel("Age")

    # Plot Spending Score
    sns.histplot(filtered_data["Spending_Score_(1-100)"], bins=10, kde=True, ax=axes[1], color="red")
    axes[1].set_title(f"Spending Score in {location}")
    axes[1].set_xlabel("Spending Score")

    # Save the plot to a PNG image in memory
    img = BytesIO()
    plt.savefig(img, format="png", bbox_inches="tight")
    plt.close(fig)
    img.seek(0)

    # Encode the image in Base64
    return base64.b64encode(img.getvalue()).decode()

# Create the base map centered in Nairobi
nairobi_map = folium.Map(location=[-1.286389, 36.817223], zoom_start=12)

# Add markers for each location
for _, row in df.dropna(subset=["latitude", "longitude"]).iterrows():
    location = row["location"]  # Move location assignment here!
```

```
    # Get the most common supermarket for this location
    top_supermarket = top_supermarkets[top_supermarkets["location"] == location]["supermarket"].values
    top_supermarket_str = top_supermarket[0] if len(top_supermarket) > 0 else "No data"

    # Get encoded image for the demographic graph
    encoded_image = generate_demographic_plot(location)
    img_html = f'<img src="{encoded_image}" width="400px">'

   # Get the most common income cluster for this location
    top_cluster = top_clusters[top_clusters["location"] == location]["Income Cluster"].values
    top_cluster_str = int(top_cluster[0]) if len(top_cluster) > 0 else "No data"

    # Define the popup content
    popup_content = f"""
    <h4>Location: {location}</h4>
    <p><b>Major Customer Cluster:</b> {top_cluster_str}</p>
    <p><b>Most common supermarket:</b> {top_supermarket_str}</p>
       {img_html}
    """

    # Add marker with pop-up
    folium.Marker(
        location=[row["latitude"], row["longitude"]],
        popup=folium.Popup(popup_content, max_width=450),
        icon=folium.Icon(color="blue")
    ).add_to(nairobi_map)

# Show the map
nairobi_map

```

Out[47]:



Out[47]:



CHAPTER NINE: APPENDICES

REFERENCES

1. Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. Kybernetes, 45(7), 1129-1157.

2. Zhou, B., Lu, B., & Saeidlou, S. (2022). A Hybrid Clustering Method Based on the Several Diverse Basic Clustering and Meta-Clustering Aggregation Technique. Journal of Mathematics, 2022, 1-15.

3. Kumar, A. (2022). Customer Segmentation of Shopping Mall Users Using K-Means Clustering. International Journal of Engineering Research & Technology, 11(12), 56-63.