GROUPING SIMILAR CUSTOMERS BASED ON PURCHASE HABITS USING K-MEANS

A PROJECT REPORT

Submitted by

SANTHIYA S(2303811724322096)

in partial fulfillment of requirements for the award of the course

AGI1252 - FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE- 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report on **“GROUPING SIMILAR CUSTOMERS BASED ON PURCHASE HABITS USING K-MEANS”** is the bonafide work of **SANTHIYA S(2303811724322096)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE

Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,

HEAD OF THE DEPARTMENT

PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

SIGNATURE

Ms.S.Murugavalli., M.E.,(Ph.D).,

SUPERVISOR

ASSISTANT PROFESSOR

Department of Artificial Intelligence

K.Ramakrishnan College of Technology (Autonomous)

Samayapuram–621112.

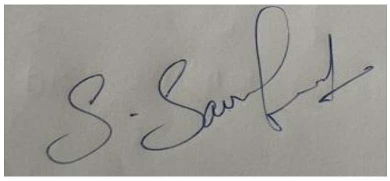
Submitted for the viva-voce examination held on **02.06.2025**.

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I declare that the project report on **“GROUPING SIMILAR CUSTOMERS BASED ON PURCHASE HABITS USING K-MEANS”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the completion of the course **AGI1252 - FUNDAMENTALS OF DATA SCIENCE USING R**



SANTHIYA S

Place: Samayapuram

Date:02.06.2025

ACKNOWLEDGEMENT

It is with great pride that I express my gratitude and in-debt to my institution "**K.Ramakrishnan College of Technology (Autonomous)**", for providing me with the opportunity to do this project.

I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of my study in college.

I would like to express my sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to my project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express my deep expression and sincere gratitude to our project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render my sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express my special thanks to the officials and Lab Technicians of my departments who rendered their help during the period of the work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals, and an engineering specialization to develop solutions to complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development.
3. **Design/development of solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required.
4. **Conduct investigations of complex problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions.
5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.
6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.

7. **Ethics:** Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws.

8. **Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.

9. **Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences.

10. **Project management and finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

11. **Life-long learning:** Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change.

ABSTRACT

Customer segmentation plays a pivotal role in modern data-driven marketing strategies. With the exponential growth of customer data, businesses must leverage this information to derive meaningful insights and offer tailored services. This project presents a practical implementation of customer segmentation using the K-Means clustering algorithm in R programming. By analyzing purchase patterns, we can group customers into distinct clusters with similar buying habits. These clusters serve as the foundation for targeted marketing, product recommendations, and service customization. Through various data preprocessing, normalization, and visualization techniques, this project highlights how unsupervised machine learning models like K-Means can uncover hidden patterns in purchase behavior. The outcomes provide valuable business intelligence, enabling organizations to improve customer satisfaction, loyalty, and profitability. The project follows a structured data science pipeline: from data acquisition and preprocessing, to exploratory data analysis, clustering, and result interpretation.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD DATA SCIENCE USING R PROGRAMMING FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
Customer segmentation plays a pivotal role in modern data-driven marketing strategies. With the exponential growth of customer data, businesses must leverage this information to derive meaningful insights and offer tailored services. This project presents a practical implementation of customer segmentation using the K-Means clustering algorithm in R programming. By analyzing purchase patterns, we can group customers into distinct clusters with similar buying habits.	**PO1 -3** **PO2 -3** **PO3 -3** **PO5 -3** **PO7 -2** **PO8 -2** **PO10 -3** **PO11-2**	**PSO1 -3** **PSO2 -3**

Note: 1- Low, 2-Medium, 3- High

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Objective

The objective of this project is to implement the K-Means clustering algorithm using R programming to group customers based on their purchase habits. The primary aim is to analyze real-world customer data and identify patterns that allow the segmentation of customers into distinct and meaningful clusters. This segmentation helps businesses gain deeper insights into customer behavior, enabling them to design personalized marketing strategies, optimize inventory, and enhance customer satisfaction.

The project involves collecting relevant customer purchase data, preprocessing it to handle inconsistencies, and applying K-Means clustering to detect groups with similar purchasing characteristics, such as frequency of purchases, spending patterns, and product preferences. By identifying the optimal number of clusters through methods like the Elbow Method and Silhouette Score, the model ensures that each cluster is both well-defined and useful for interpretation.

Additionally, this project aims to showcase the practical application of R as a robust tool for data analytics, data visualization, and machine learning. It highlights how unsupervised learning can be applied to real-time problems in the business domain, particularly in customer relationship management and strategic planning. Through this, the project not only enhances the understanding of data-driven decision-making but also supports the development of intelligent systems that can adapt to customer needs and behaviors.

1.2 Overview

In the evolving landscape of modern business, data has become a central asset for shaping strategic decisions, particularly in understanding customer behavior. Organizations today collect vast volumes of data through various channels such as online purchases, point-of-sale systems, loyalty programs, and customer feedback platforms. The ability to transform this raw data into meaningful insights is what gives businesses a competitive edge in the digital economy. One such powerful approach to extracting value from data is **customer segmentation**, which involves categorizing customers into distinct groups based on shared characteristics or purchasing behavior. This enables businesses to move from a one-size-fits-all approach to more personalized and effective engagement strategies. This project centers on customer segmentation using the **K-Means clustering algorithm** in **R programming**. The aim is to group similar customers based on their purchase habits in order to identify distinct segments that can be analyzed and targeted accordingly. The project begins with the acquisition and exploration of a relevant customer dataset. This dataset is then subjected to thorough preprocessing steps which include cleaning missing or anomalous values,transforming and normalizing numerical features, and selecting key variables that significantly influence customer behavior. These processes ensure that the dataset is consistent, scalable, and suitable for unsupervised learning models.

After data preparation, the **K-Means clustering algorithm** is applied. As a widely used unsupervised learning technique, K-Means aims to partition the dataset into 'k' groups in which each observation belongs to the cluster with the nearest mean. The project uses techniques such as the **Elbow Method** and **Silhouette Coefficient** to determine the optimal number of clusters, thereby enhancing the reliability and interpretability of the clustering output. The resulting customer segments are then analyzed to identify defining features of each cluster, such as average purchase value, frequency of transactions, preferred product categories, or seasonal buying patterns.

In addition to numerical outputs, the project utilizes **data visualization tools** within R—such as `ggplot2`, `factoextra`, and `cluster`—to graphically represent the cluster distribution and customer profiles. Visual plots not only aid in interpretation but also make the results more accessible to non-technical stakeholders. These tools allow us to gain a clearer understanding of how customer groups differ and how these insights can be translated into business strategies.

The final stage of the project involves interpreting the clusters to derive **actionable business insights**. For instance, a particular cluster might represent loyal customers with high lifetime value, while another could identify price-sensitive or infrequent shoppers. These segments can inform targeted marketing campaigns, customized product recommendations, dynamic pricing strategies, inventory stocking decisions, and customer relationship management. The insights generated can also help in developing loyalty programs or promotional offers tailored to specific customer types.

By applying K-Means clustering through R, this project not only demonstrates the practical utility of machine learning in solving real-world business problems but also emphasizes the importance of data preprocessing, visualization, and interpretation in the overall analytical pipeline. Moreover, it reflects how unsupervised learning can be employed to discover structure in unlabelled datasets and to support informed, data- driven decision-making. The methodology adopted in this project is scalable and can be replicated across various sectors including retail, banking, e-commerce, and telecommunications, wherever customer behavior analysis is critical. Ultimately, this project contributes to a deeper understanding of customer diversity and provides a blueprint for how businesses can leverage data science for competitive advantage.

1.3 R library used

R is a robust statistical programming language ideal for data analysis, machine learning, and data visualization. Its open-source nature, extensive community support, and vast collection of packages make it a preferred tool for projects involving clustering and unsupervised learning. For the purpose of grouping similar customers based on their purchase habits using the K-Means algorithm, several key R packages have been utilized to enable seamless data preprocessing, model execution, and result interpretation.

One of the foundational packages employed in this project is **`dplyr`**, part of the Tidyverse collection. `dplyr` offers a powerful set of functions for data manipulation and transformation. It provides a consistent grammar for performing tasks such as filtering rows, selecting columns, summarizing data, and creating new variables. In the context of this project, `dplyr` is essential for preparing the customer data, ensuring that all variables are cleaned, formatted, and ready for clustering. It enables operations like removing duplicates, handling missing values, grouping customer transactions, and summarizing purchase frequency and spending patterns, which are critical for building meaningful clusters.

The **`ggplot2`** package, also from the Tidyverse suite, is used to create high-quality visualizations. It is based on the Grammar of Graphics, which allows users to build plots in layers. For this project, ggplot2 is instrumental in visualizing the clustering results. Scatter plots are created to show how customers are grouped into clusters based on their purchase behavior. These visuals help in identifying cluster patterns, overlaps, and outliers. Additionally, bar plots and histograms generated through ggplot2 aid in exploring the distribution of variables like purchase frequency, total amount spent, and product categories across different customer segments.

The **cluster** package plays a central role in the application of clustering algorithms. It provides the core functionality needed to execute the K-Means algorithm and assess the quality of the resulting clusters. Specifically, it offers tools for calculating **silhouette scores**, which measure how similar a data point is to its own cluster compared to others. This helps in validating the effectiveness of the clustering process and choosing the optimal number of clusters. The package also supports other clustering techniques, allowing flexibility if alternative methods such as hierarchical clustering need to be considered.
The **factoextra** package enhances the interpretability of multivariate data analysis by simplifying the extraction and visualization of clustering results. It is used to generate intuitive and informative cluster plots, including those that display the **Elbow Method** and **Silhouette Method**, both of which are used to determine the optimal number of clusters. factoextra integrates seamlessly with the output from the cluster and stats packages, allowing users to visualize cluster centroids, intra- cluster distances, and the spread of data points within each group. This visual clarity is essential for drawing actionable insights from the model results.Together, these packages—dplyr, ggplot2, cluster, and factoextra—form a powerful ecosystem that supports the full lifecycle of clustering analysis in R.

CHAPTER 2
PROJECT METHODOLOGY

2.1 Proposed Work

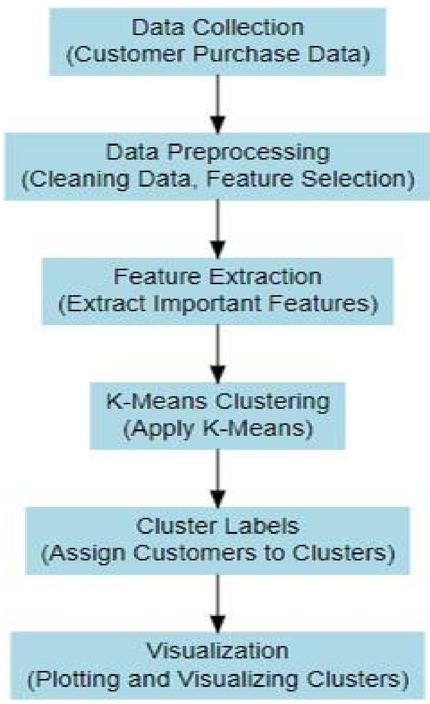
The proposed work aims to implement a customer segmentation model using the K- Means clustering algorithm to identify groups of customers with similar purchase habits. This segmentation will assist businesses in better understanding their customer base, enabling targeted marketing, improved customer service, and optimized resource allocation. The project begins with the acquisition of customer purchase data, which includes relevant features such as transaction frequency, amount spent, product categories, and purchase timelines. The dataset undergoes a rigorous preprocessing phase involving data cleaning, handling of missing values, normalization of numerical variables, and transformation of categorical data where necessary. These steps ensure that the data is structured appropriately for effective clustering.

Following preprocessing, the next phase involves applying the K- Means algorithm. K- Means is an unsupervised learning technique that partitions the dataset into k distinct, non-overlapping clusters based on similarity metrics, typically using Euclidean distance. To determine the optimal number of clusters, internal validation techniques such as the Elbow Method and Silhouette Analysis are employed. Once the ideal value of k is identified, the algorithm groups customers such that intra-cluster similarity is maximized while inter-cluster similarity is minimized.

The resulting clusters are then analyzed to understand the behavioral patterns and characteristics that distinguish each segment. This may include high-value customers, frequent shoppers, occasional buyers, or price-sensitive users. Visualizations such as cluster plots, distribution graphs, and centroid analyses are used to present the results in a clear and interpretable manner. The insights derived from this analysis can be directly applied to real-world business strategies like personalized marketing, loyalty programs, and inventory management.

This project proposes not just a theoretical framework but a practical, data-driven solution for businesses aiming to improve their customer relationship management through meaningful segmentation. By utilizing R programming and its robust libraries, the proposed work enables a fully integrated approach—from raw data ingestion to insight-driven decision-making—thus demonstrating how clustering can solve real- time business problems efficiently.

2.2 Block Diagram



CHAPTER 3
MODULE DESCRIPTION

3.1 generate_data()

This module is responsible for creating or importing the dataset that contains customer purchase information. It may include synthetic data generation or real-world data import using functions like read.csv() or read_excel(). Once generated or loaded, this data serves as the foundation for further analysis and clustering. The module ensures the data is structured in a usable format with appropriate column names and data types for subsequent processing.Additionally, this module may include initial preprocessing tasks such as handling missing values, encoding categorical variables, and basic normalization to prepare the data for analytical workflows. It can also include logging mechanisms to track the source and time of data generation or import, thereby enhancing reproducibility and traceability. Proper data validation checks are applied to ensure consistency and reliability before proceeding to the clustering steps.

3.2 fviz_cluster()

The fviz_cluster() function from the **factoextra** package is used to visually display the results of clustering. After applying the K-Means algorithm, this module graphically represents the formed clusters in two-dimensional space. Each cluster is color-coded, and centroids are highlighted to make the segmentation more interpretable. This helps in visually validating the grouping and understanding how well-separated the clusters are. It plays a key role in interpreting the results and communicating the structure of customer groups.

3.3 group_by() + summarise()

These two functions from the **dplyr** package are used together for aggregating and analyzing data within groups. After clustering, customers are grouped based on their cluster assignment using group_by(). Then, summary statistics such as average spending, total purchases, or frequency per cluster are calculated using summarise(). This module provides behavior of each customer segment and helps identify patterns that can inform targeted marketing or business decisions. It can also be used to generate comparative insights across different clusters by aggregating multiple metrics simultaneously. Visual representations such as bar charts or tables can be created based on the summarized data to make the findings more accessible and easier to interpret. Additionally, these grouped summaries serve as a foundation for advanced segmentation reporting and can be exported for use in dashboards or presentations. This module plays a key role in turning raw clustering results into actionable business intelligence.

3.4 observeEvent()

The observeEvent() function is a reactive function used in Shiny applications (R's web app framework). In this context, it monitors user inputs or interactions in the interface (e.g., selecting the number of clusters or pressing a button to run the model). When an event is triggered, this function updates the server-side logic to reflect new outputs or visualizations accordingly. It enables interactive exploration of clustering results, allowing users to dynamically analyze different scenarios or cluster configurations in real-time.

CHAPTER 4

CONCLUSION & FUTURE SCOPE

In this project, the implementation of the K-Means clustering algorithm to group similar customers based on their purchasing habits has successfully demonstrated the potential of data-driven insights in understanding customer behavior. By collecting and preprocessing relevant customer data, performing exploratory data analysis (EDA), and applying unsupervised machine learning through K-Means, we were able to identify distinct customer segments with shared characteristics. These clusters were analyzed both statistically and visually to derive meaningful interpretations, which can significantly aid businesses in decision-making processes.

The project highlighted the importance of data preparation, such as normalization and handling missing values, which directly influence clustering accuracy. Additionally, the use of R and its powerful ecosystem of packages—including `dplyr`, `ggplot2`, `cluster`, and `factoextra`—enabled efficient data manipulation, model implementation, and insightful visualization. Each cluster revealed unique patterns, such as high-frequency buyers, occasional spenders, or seasonal customers, providing a strategic advantage for targeted marketing, personalized service offerings, and inventory management.

From a business perspective, the clustering results offer a foundation for creating customer-centric strategies that can improve satisfaction, boost retention rates, and drive profitability. By tailoring campaigns and promotions to the needs and preferences of each cluster, companies can optimize resource allocation and enhance the effectiveness of their engagement efforts.Looking ahead, there is ample scope to build on this work. Future enhancements can include the integration of more complex data, such as behavioral logs, product reviews, or demographic information, to form richer customer profiles. Other clustering techniques like DBSCAN or hierarchical clustering can be explored for comparative performance. Moreover, the model can be embedded in a real-time Shiny dashboard for dynamic segmentation, allowing businesses to interactively explore clusters as new data arrives. Machine learning can also be extended to supervised methods to predict customer churn, lifetime value, or conversion probabilities.

In conclusion, this project lays a strong foundation for customer segmentation using K-Means clustering in R. It not only demonstrates technical feasibility but also emphasizes its real-world applicability and impact. As the availability of customer data continues to grow, such analytical approaches will become increasingly vital in enabling data-driven, customer-focused strategies in competitive business environments.

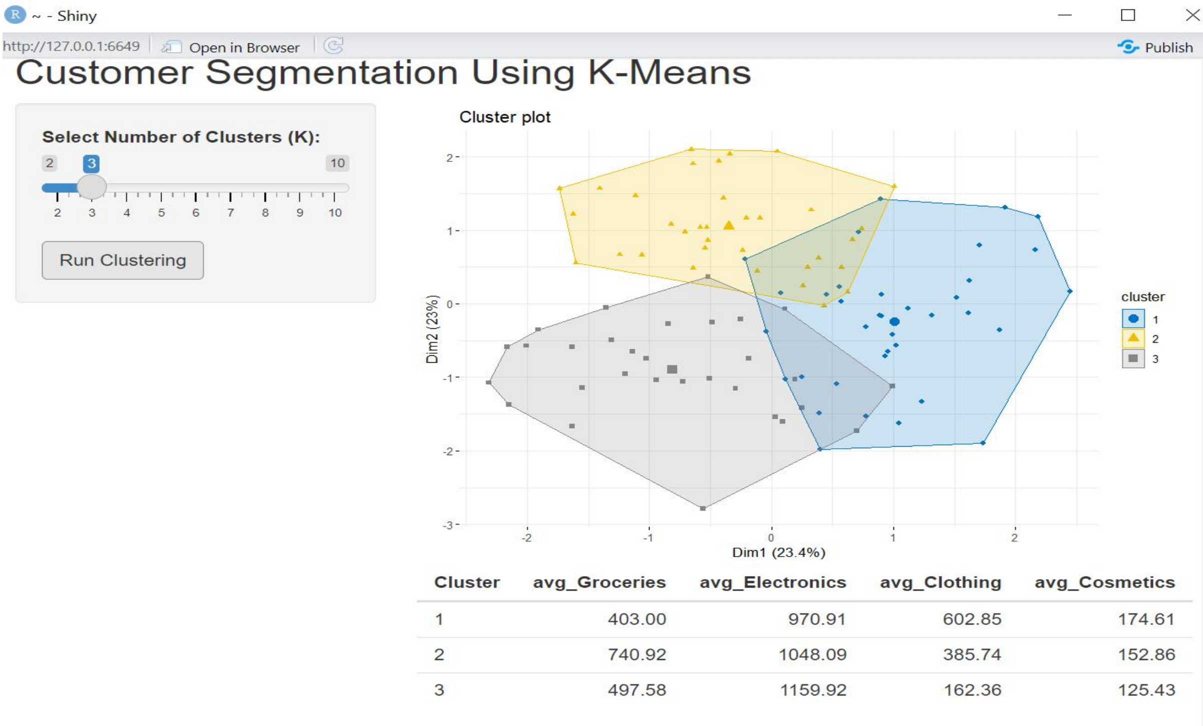
APPENDICES

APPENDIX A SOURCE CODE

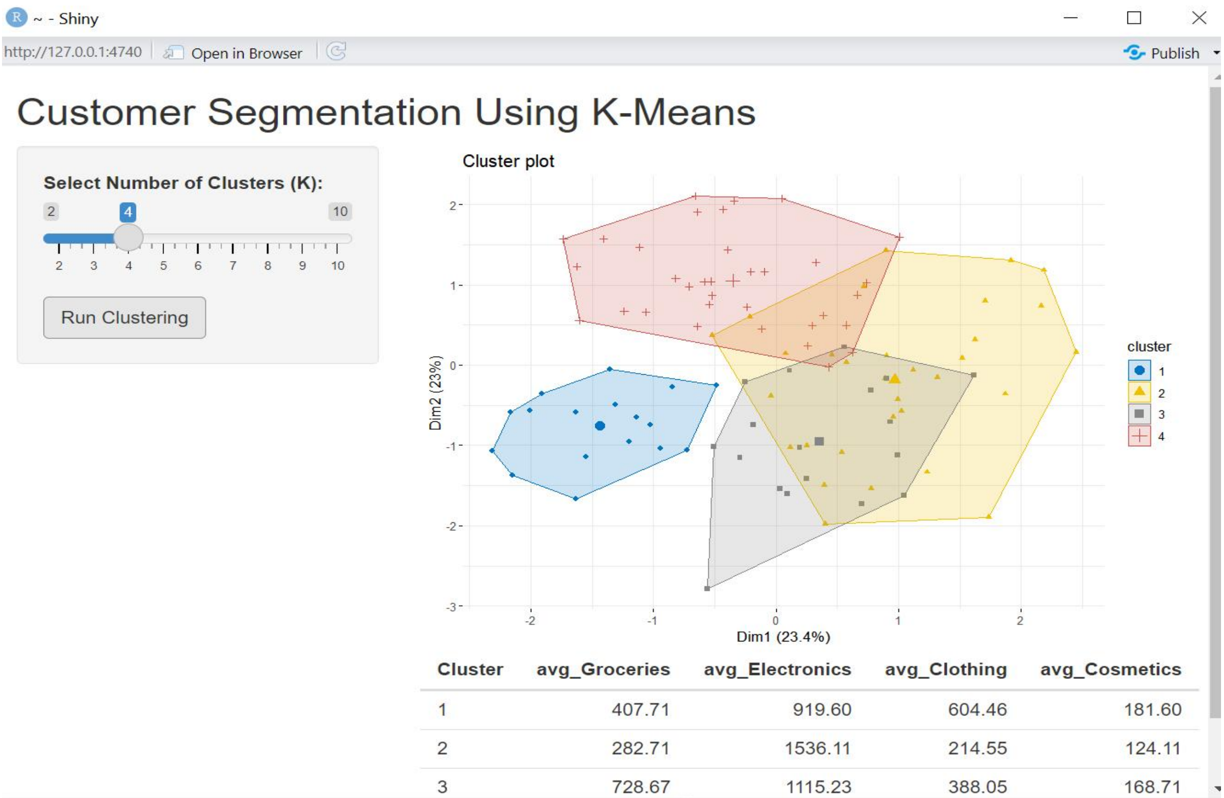
```
library(shiny)
library(ggplot2)
library(factoextra)
library(dplyr)
generate_data <- function(n = 100) {
set.seed(123)
data.frame(
CustomerID = 1:n,
Groceries = runif(n, 100, 1000),
Electronics = runif(n, 50, 2000),
Clothing = runif(n, 20, 800),
Cosmetics = runif(n, 10, 300))}
ui <- fluidPage(
titlePanel("Customer Segmentation Using K-Means"),
sidebarLayout(
sidebarPanel(
sliderInput("k", "Select Number of Clusters (K):", min = 2, max = 10, value =3),
actionButton("run", "Run Clustering")),
mainPanel(
plotOutput("clusterPlot"),
tableOutput("clusterSummary") ) ))
server <- function(input, output) {
```

```
data <- reactiveVal(generate_data())
observeEvent(input$run, {
df <- data()
df_numeric <- df[, sapply(df, is.numeric)]
df_scaled <- scale(df_numeric)
set.seed(123)
km <- kmeans(df_scaled, centers = input$k, nstart = 25)
df$Cluster <- as.factor(km$cluster)
data(df) })
output$clusterPlot <- renderPlot({ req(data()$Cluster) df <- data()
df_numeric <- df[, sapply(df, is.numeric)]
df_scaled <- scale(df_numeric)
km <- kmeans(df_scaled, centers = input$k, nstart = 25)
fviz_cluster(km, data = df_scaled,
palette = "jco",
geom = "point",
ellipse.type = "convex",
ggtheme = theme_minimal())  })
output$clusterSummary <- renderTable({
req(data()$Cluster)
data() %>%
group_by(Cluster) %>%
summarise(across(Groceries:Cosmetics, mean, .names =
"avg_{.col}"))})}shinyApp(ui = ui, server = server)
```

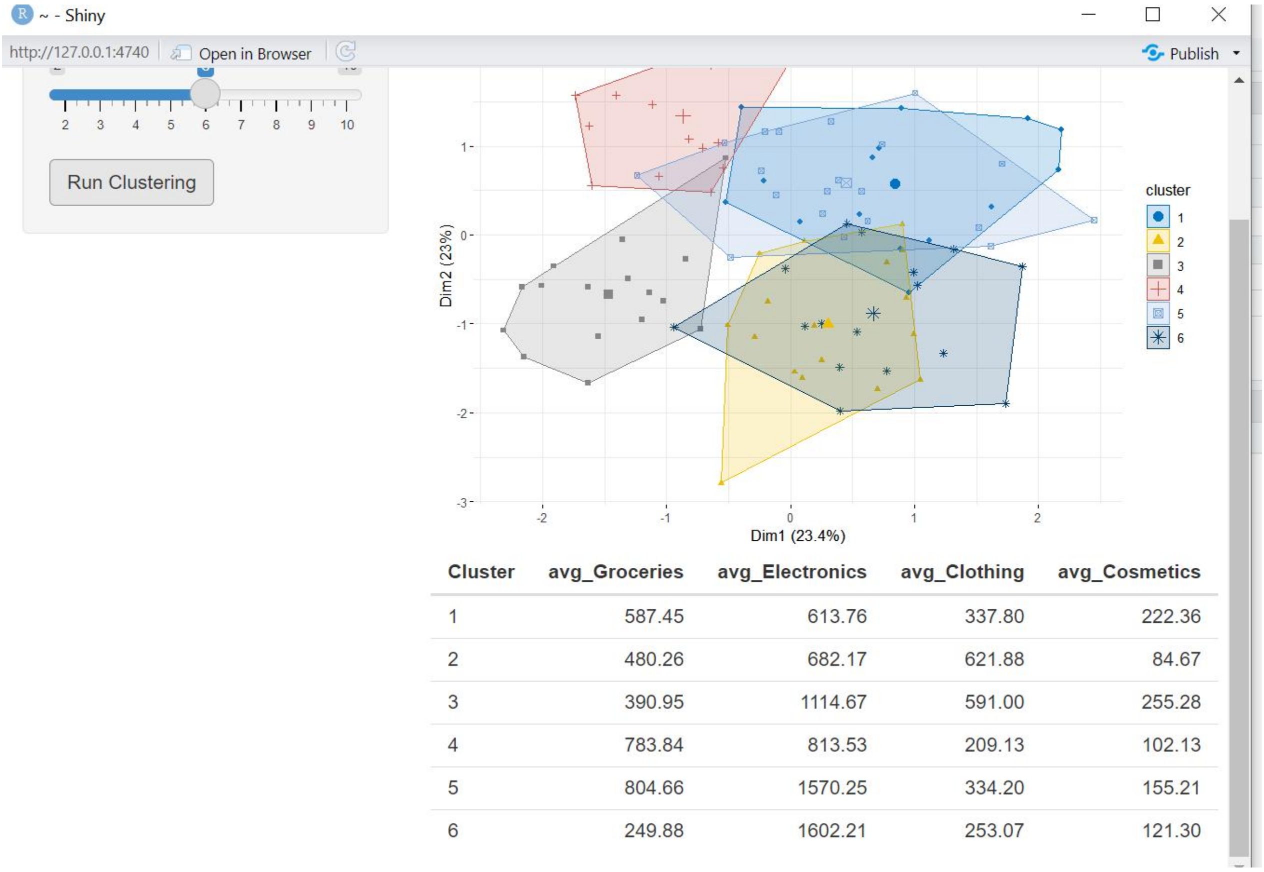
APPENDIX B SCREEN SHOT



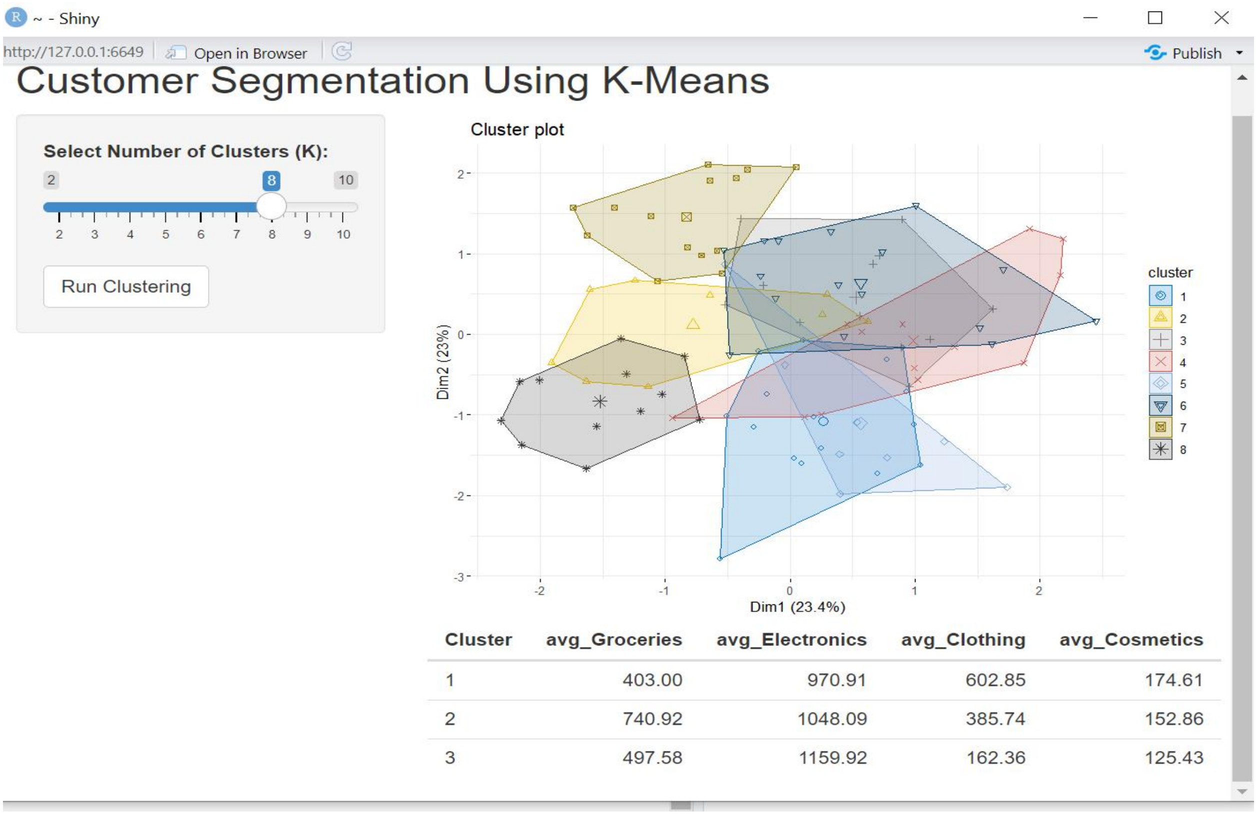
Cluster	avg_Groceries	avg_Electronics	avg_Clothing	avg_Cosmetics
1	403.00	970.91	602.85	174.61
2	740.92	1048.09	385.74	152.86
3	497.58	1159.92	162.36	125.43



Cluster	avg_Groceries	avg_Electronics	avg_Clothing	avg_Cosmetics
1	407.71	919.60	604.46	181.60
2	282.71	1536.11	214.55	124.11
3	728.67	1115.23	388.05	168.71



Cluster	avg_Groceries	avg_Electronics	avg_Clothing	avg_Cosmetics
1	587.45	613.76	337.80	222.36
2	480.26	682.17	621.88	84.67
3	390.95	1114.67	591.00	255.28
4	783.84	813.53	209.13	102.13
5	804.66	1570.25	334.20	155.21
6	249.88	1602.21	253.07	121.30



Cluster	avg_Groceries	avg_Electronics	avg_Clothing	avg_Cosmetics
1	403.00	970.91	602.85	174.61
2	740.92	1048.09	385.74	152.86
3	497.58	1159.92	162.36	125.43

REFERENCES

1. **Han, J., Kamber, M., & Pei, J. (2011).**
   Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
   Core concepts of clustering, including K-Means and customer data analysis.
2. **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).**
   An Introduction to Statistical Learning: With Applications in R. Springer.
   Provides practical K-Means applications and examples in R.
3. **Kassambara, A. (2017).**
   Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. STHDA.
   Focused on hands-on K-Means implementation and customer segmentation in R.
4. **Tan, P.-N., Steinbach, M., & Kumar, V. (2005).**
   Introduction to Data Mining. Pearson Education.
   Covers K-Means clustering and segmentation logic in marketing contexts.
5. **Wedel, M., & Kamakura, W. A. (2000).**
   Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Springer.
   Advanced insights into customer segmentation based on behavioral data.