Project Based Learning Report

On

Build an application for customer segmentation with machine learning using R Language

Submitted in the partial fulfilment of the requirement

For the Project Based Learning in **ITC-II: Essentials of Data Science**

In

Electronics & Communication Engineering

By

2214110410 Parikshit Bhoyar

2214110416 Isha Saini

2214110430 Goutam Sahu

Under the guidance of **V.P. Kaduskar**

Department of Electronics & Communication Engineering

Bharati Vidyapeeth
(Deemed to be University)
College Of Engineering
Pune – 410043

Academic Year : 2023-24

Bharati Vidyapeeth
(Deemed to be University)
College of Engineering,
Pune – 411043

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

CERTIFICATE

Certified that the Project Based Learning report entitled, **“Build an application for customer segmentation with machine learning using R Languange”**

is work done by

2214110410	**Parikshit Bhoyar**
2214110416	**Isha Saini**
2214110430	**Goutam Sahu**

in partial fulfilment of the requirements for the award of credits for Project Based Learning (PBL) in of **ITC-II: Essentials of Data Science** Bachelor of Technology Semester IV, in Electronics & Communication Engineering.

Date:

Prof. V.P. Kaduskar
Course In-charge

Dr. Arundhati A. Shinde
Professor & Head

Index

Introduction

Customer segmentation is the process of dividing a company's customer base into distinct groups based on certain characteristics or behaviors they share. These characteristics can include demographic factors (such as age, gender, income), geographic location, psychographic factors (such as lifestyle, values, personality), behavioral patterns (such as purchasing frequency, brand loyalty, usage habits), or even firmographic factors (for business-to-business segmentation, such as industry, company size, revenue).

The purpose of customer segmentation is to better understand customers' needs, preferences, and behaviors so that businesses can tailor their marketing efforts, products, and services to effectively meet the specific needs of each segment. By identifying different segments within their customer base, companies can develop targeted marketing strategies, improve customer satisfaction, enhance customer retention, and ultimately increase profitability.

R is a programming language and software environment primarily used for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques, making it popular among statisticians, data miners, and data analysts for data analysis, visualization, and statistical modeling.

Originally developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, in the early 1990s, R has since become one of the most widely used languages for statistical computing and data analysis, particularly in academia, research, and industries such as finance, healthcare, and technology.

Problem Statement

Customer Segmentation with K-means Clustering using R Language

Objective:

Utilize K-means clustering in R to effectively segment customers based on relevant features, facilitating targeted marketing strategies and personalized customer experiences.

R language

R is a programming language and environment primarily used for statistical computing and graphics. It was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, in the early 1990s. Since then, it has grown into one of the most widely used languages for statistical analysis, data visualization, and machine learning. Here's a detailed overview of R:

Programming IDE:

R Studio is the go-to integrated development environment (IDE) for R programming, a language favored for statistical computing and data analysis. It offers a robust suite of features tailored to meet the needs of data scientists and researchers. From its intuitive interface to its powerful coding tools, R Studio streamlines the process of data manipulation, visualization, and statistical modeling. With built-in support for version control systems like Git, it facilitates seamless collaboration among teams. Its extensive collection of packages covers a wide range of domains, including machine learning and time series analysis, ensuring versatility and efficiency in analytical tasks. In short, R Studio provides a comprehensive environment that empowers users to explore, analyze, and visualize data effectively.

Customer Segmentation

Customer segmentation involves categorizing customers into groups based on similar characteristics or behaviors. This strategy enables businesses to better understand their customers and tailor marketing efforts, products, and services accordingly. Here's an overview:

1. **Demographic Segmentation**:

- Age, gender, income, education, and occupation are used to group customers.

2. **Geographic Segmentation**:

- Location and climate are considered, dividing customers based on where they reside and the prevailing weather conditions.

3. **Psychographic Segmentation**:

- Lifestyle, personality, and social class are factors used to categorize customers based on their interests, values, and social status.

4. **Behavioral Segmentation**:

- Purchase behavior, brand loyalty, occasion, and usage rate help segment customers based on their buying habits and product/service usage patterns.

5. **Technographic Segmentation**:

- Technology usage and online behavior are analyzed, categorizing customers by the devices they use and their online activities.

6. **Benefit Segmentation**:

- Needs, preferences, and problem-solving approaches are considered to identify the specific benefits or solutions customers seek.

7. **Usage-Based Segmentation**:

- Heavy vs. light users and first-time vs. repeat buyers are distinctions made based on the intensity and frequency of product/service usage.

8. **Custom Segmentation**:

- RFM Analysis categorizes customers based on Recency, Frequency, and Monetary value.
- Predictive Segmentation employs machine learning algorithms to predict customer segments based on various attributes and behaviors.

By utilizing these segmentation methods, businesses can effectively target their customer base and enhance their marketing strategies and offerings.

Customer segmentation using machine learning.

Customer segmentation through machine learning utilizes algorithms to automatically detect patterns and similarities in customer data, resulting in the creation of distinct customer segments. Here's an overview of the process:

1. **Data Collection and Preprocessing**:
    - Gather relevant customer data from various sources like CRM systems, transaction records, and social media.
    - Clean and preprocess the data, handling missing values and encoding categorical variables.
    - Feature engineering may be applied to create or transform features for better model performance.

2. **Feature Selection**:
    - Select the most pertinent features likely to contribute to segmentation, such as demographic info and purchase history.
    - Use techniques like correlation analysis or feature importance ranking to identify predictive features.

3. Choosing an Algorithm:

- Select a suitable machine learning algorithm for segmentation, like:
- K-means Clustering
- Hierarchical Clustering
- DBSCAN
- Gaussian Mixture Models (GMM)
- Self-Organizing Maps (SOM)

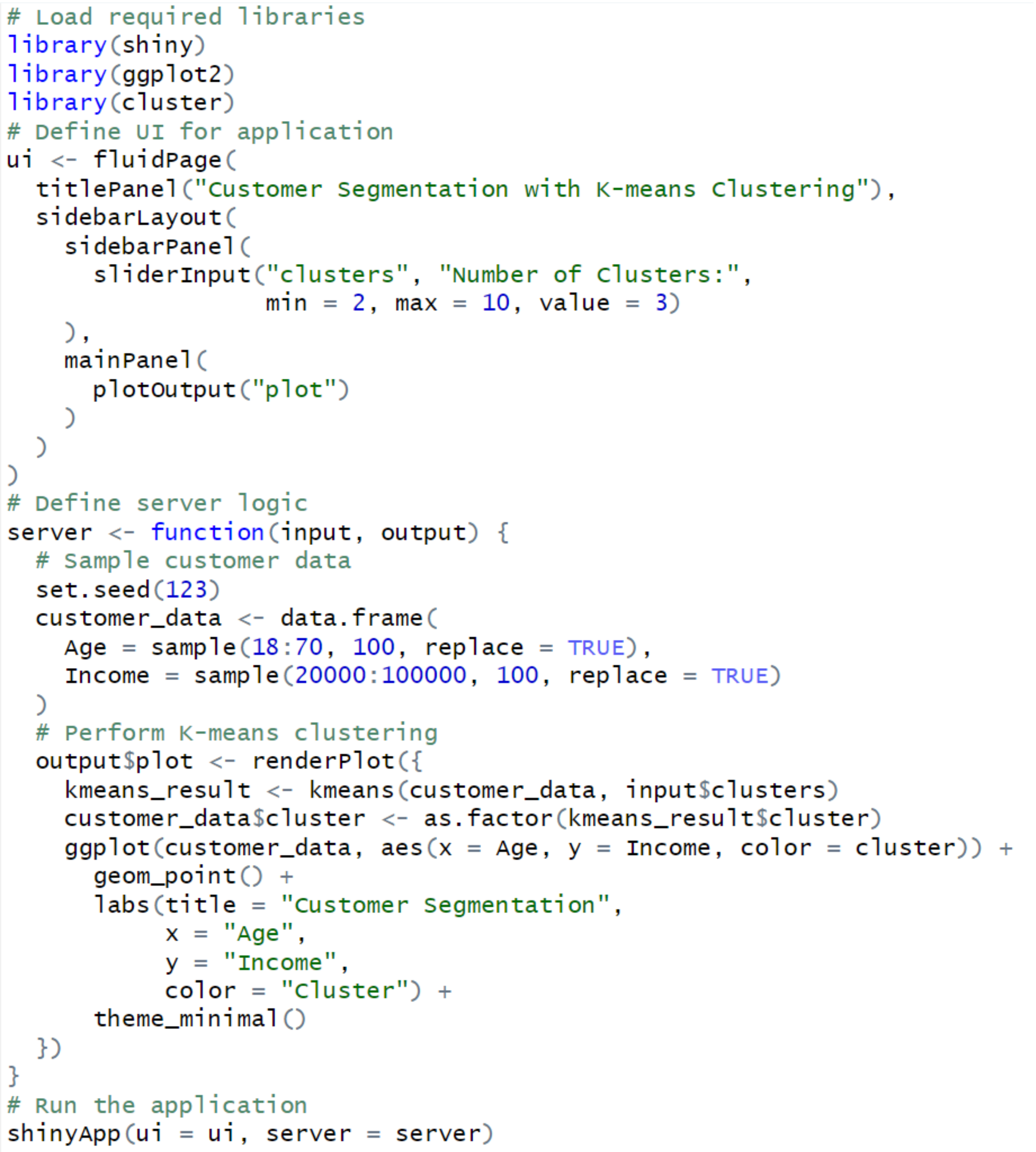
4. **Model Training**:
    - Split data into training and validation sets.
    - Train the chosen algorithm on the training data, tuning hyperparameters for optimal performance.

5. **Customer Segmentation**:
    - Apply the trained model to segment customers into clusters based on their characteristics and behavior.
    - Evaluate segmentation quality using metrics like silhouette score or Davies–Bouldin index.

6. **Interpretation and Validation**:
    - Understand segment characteristics, preferences, and needs.
    - Validate results by assessing cluster coherence and distinctiveness.

7. **Implementation and Monitoring**:
    - Implement targeted marketing campaigns and personalized experiences based on identified segments.
    - Continuously monitor customer behavior to refine the segmentation model and adapt strategies.

R Code



```r
# Load required libraries
library(shiny)
library(ggplot2)
library(cluster)
# Define UI for application
ui <- fluidPage(
  titlePanel("Customer Segmentation with K-means Clustering"),
  sidebarLayout(
    sidebarPanel(
      sliderInput("clusters", "Number of Clusters:",
                  min = 2, max = 10, value = 3)
    ),
    mainPanel(
      plotOutput("plot")
    )
  )
)
# Define server logic
server <- function(input, output) {
  # Sample customer data
  set.seed(123)
  customer_data <- data.frame(
    Age = sample(18:70, 100, replace = TRUE),
    Income = sample(20000:100000, 100, replace = TRUE)
  )
  # Perform K-means clustering
  output$plot <- renderPlot({
    kmeans_result <- kmeans(customer_data, input$clusters)
    customer_data$cluster <- as.factor(kmeans_result$cluster)
    ggplot(customer_data, aes(x = Age, y = Income, color = cluster)) +
      geom_point() +
      labs(title = "Customer Segmentation",
           x = "Age",
           y = "Income",
           color = "Cluster") +
      theme_minimal()
  })
}
# Run the application
shinyApp(ui = ui, server = server)
```

Code Explanation:

This code creates a Shiny web application for performing customer segmentation using K-means clustering. Here's a breakdown of how it works:

1. **Librarie**s: It loads necessary libraries, including `shiny`, `ggplot2`, and `cluster`.

2. **UI Definition**:

- It defines the user interface (UI) for the Shiny app using `fluidPage`.
- The UI consists of a title panel displaying "Customer Segmentation with K-means Clustering" and a sidebar layout.
- In the sidebar panel, there's a slider input allowing the user to select the number of clusters (from 2 to 10).

3. **Server Logic**:

- It defines the server logic for the Shiny app.
- It generates sample customer data consisting of age and income.
- Inside the `renderPlot` function, it performs K-means clustering on the customer data based on the input provided by the user (number of clusters).
- It assigns each customer to a cluster and creates a plot using ggplot, where each point represents a customer, colored by their cluster assignment.
- The plot has labels and a title for better visualization.

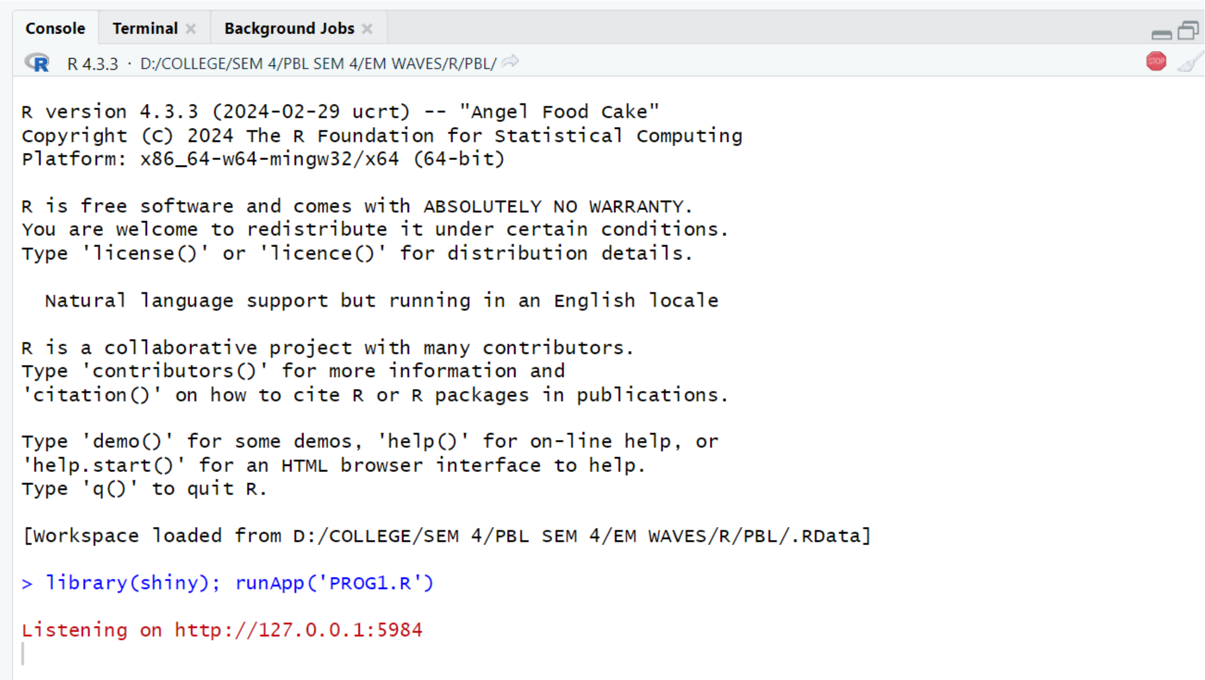
4. **Running the Application**:

- It runs the Shiny application using `shinyApp` function, with UI and server logic specified.\

K-means clustering algorithm:

1. Initialization: Randomly select K data points from the dataset as initial cluster centroids.
2. Assignment: Assign each data point to the nearest centroid based on Euclidean distance or another chosen distance metric.
3. Update Centroids: Recalculate the centroid of each cluster by taking the mean of all data points assigned to that cluster.
4. Repeat: Iteratively repeat steps 2 and 3 until convergence criteria are met. Convergence can be achieved when centroids no longer change significantly or when a specified number of iterations is reached.
5. Output: The final clusters are formed by the data points assigned to each centroid.

Console Output:

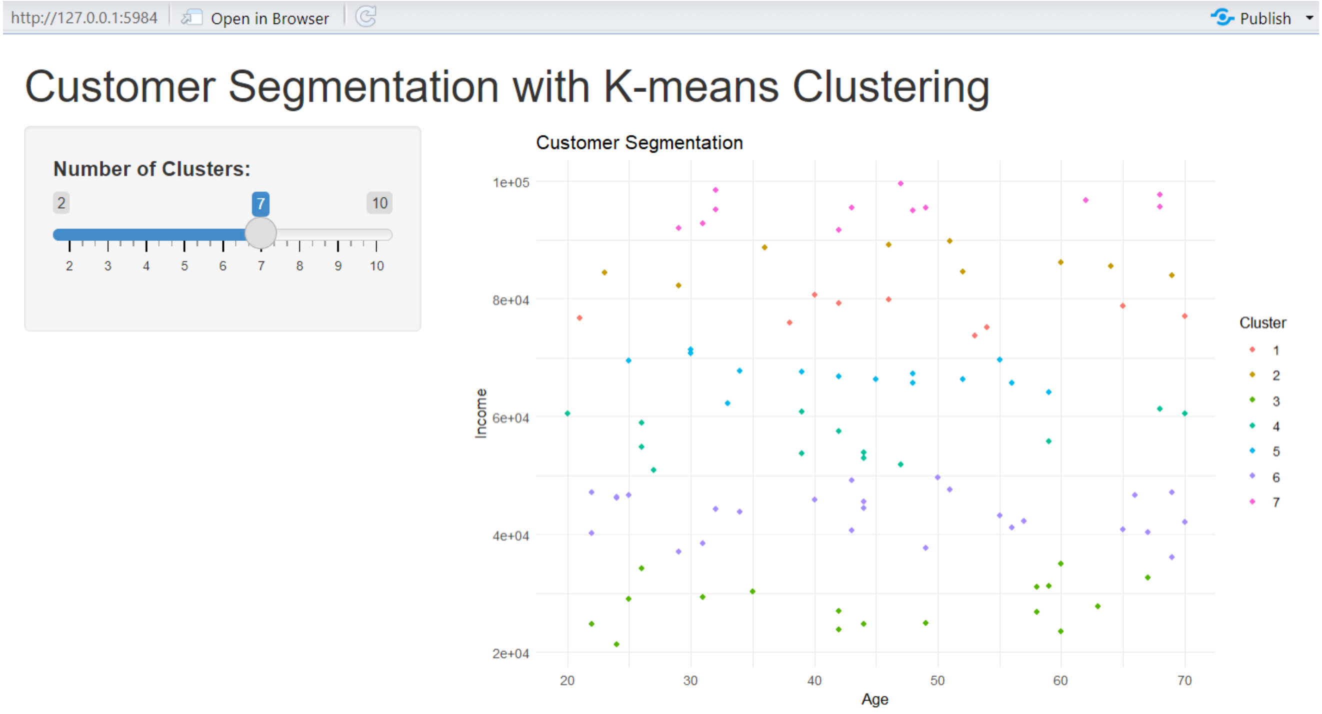


```
R version 4.3.3 (2024-02-29 ucrt) -- "Angel Food Cake"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from D:/COLLEGE/SEM 4/PBL SEM 4/EM WAVES/R/PBL/.RData]

> library(shiny); runApp('PROG1.R')

Listening on http://127.0.0.1:5984
```

Output Window:

With Number of clusters 7:

The customer data is divided into 7 clusters using the K-means algorithm, with each cluster represented by a distinct color in the output plot.

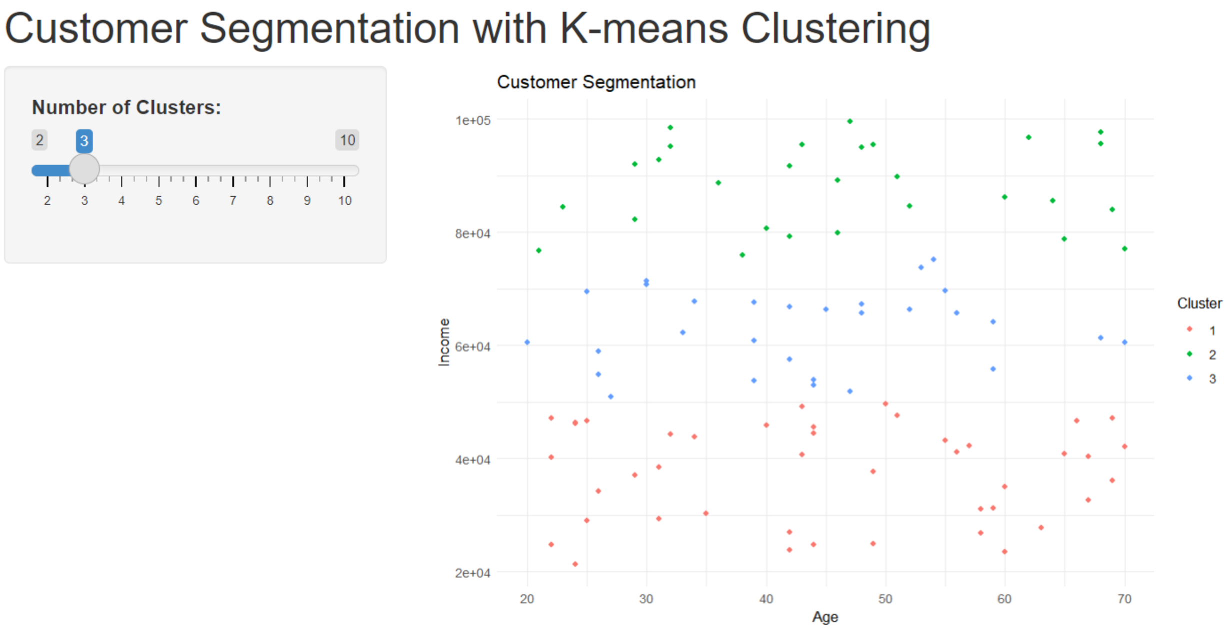


Customer Segmentation with K-means Clustering

Number of Clusters: 7

With the Number of clusters 3:

The customer data is segmented into 3 clusters where the K-mean algorithm divides the plots in different group with the specific color as show in output.

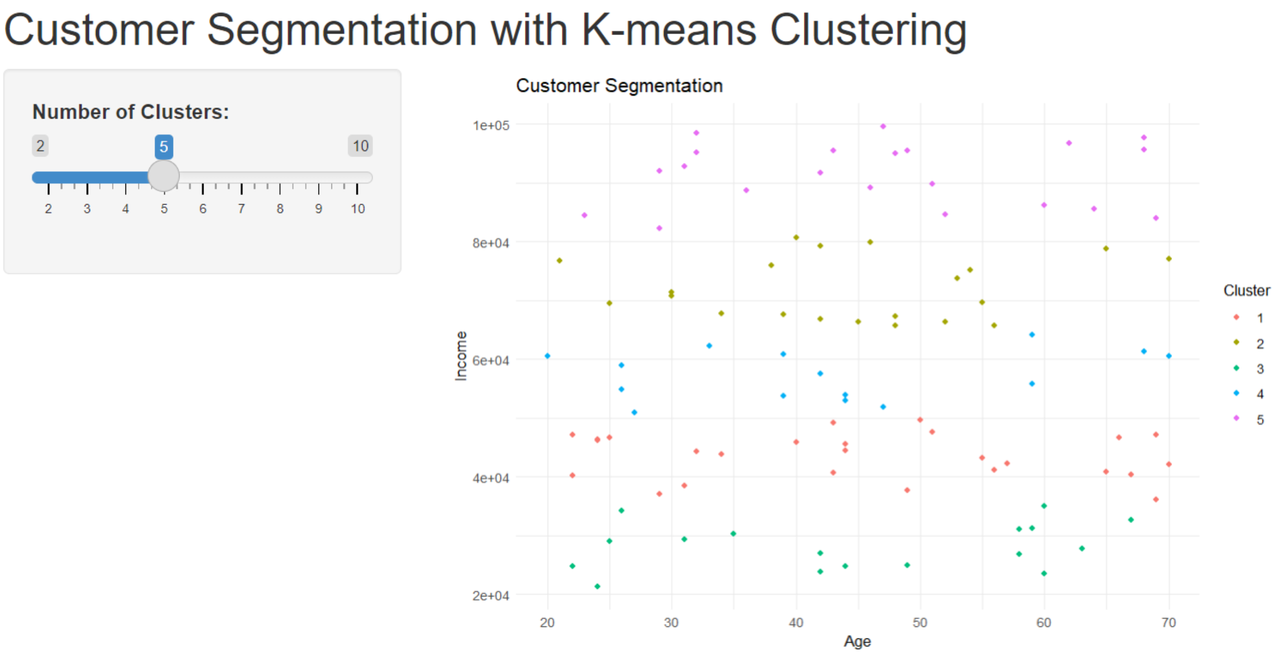


Customer Segmentation with K-means Clustering

Number of Clusters: 3

With the Number of clusters 5:

The customer data is segmented into 5 clusters where the K-mean algorithm divide the plots in different group with the specific color as show in output.



Customer Segmentation with K-means Clustering

Number of Clusters: 5

Conclusion

In conclusion, employing machine learning for customer segmentation offers a multitude of benefits that surpass traditional methods. Firstly, these algorithms bolster accuracy by deciphering intricate data patterns, resulting in more refined segmentation outcomes. Additionally, the dynamic nature of machine learning allows for real-time adjustments, ensuring marketing strategies remain pertinent amidst evolving consumer behavior. Personalized marketing initiatives, enabled by machine learning, further amplify engagement and satisfaction by catering to individual customer segments.

Moreover, automated segmentation processes streamline operations, reducing manual labor and resource expenditures. By unveiling latent customer segments, machine learning unveils previously unseen growth opportunities. Predictive insights derived from machine learning models empower strategic decision-making, enabling businesses to anticipate and respond adeptly to future customer demands. Furthermore, the scalability and integrative capabilities of machine learning render it accessible and effective across businesses of varying sizes and industries.

In essence, harnessing machine learning for customer segmentation equips businesses with deeper insights into their customer base, facilitating more impactful marketing endeavors, and ultimately fostering heightened customer satisfaction and profitability.