Market Segmentation

AUTHOR
Bala Shunmugam

Market Segmentation

1. Introduction

The KTC Company has a huge amount of data , they had to be organized and segmented based on their preferences, Age, Gender

2.Descriptive Analysis:

2.1.Data Exploration

We have information regarding 30 customers of KTC Company. We have details of their age, income, marital status, and their dependings (Their number of children, Loan, Mortage).

```
# Define input, risk, and target variables manually
# Adjust these based on your dataset
input_vars <- c("Age", "Income", "Loan", "Mortgage")
risk_vars <- c()       # Example placeholder
target_vars <- c()     # Example placeholder
```

The descriptive analysis starts with an exploration of the dataset, which consists of 30 customers. Each variable such as age, gender, marital status, number of children, income, car loan, and mortgage—is examined using density plots to understand its distribution. This step helps identify potential patterns or irregularities in the data. For instance, age and income distributions can highlight generational or economic differences across the customer base, while loan and mortgage figures offer insights into financial commitments.

```
# Define input and numeric variables manually
# Replace these with the actual column names from your dataset
numeric_vars <- c("Age", "Income", "Loan", "Mortgage")  # example placeholder
input_vars <- numeric_vars

library(readxl, quietly=TRUE)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

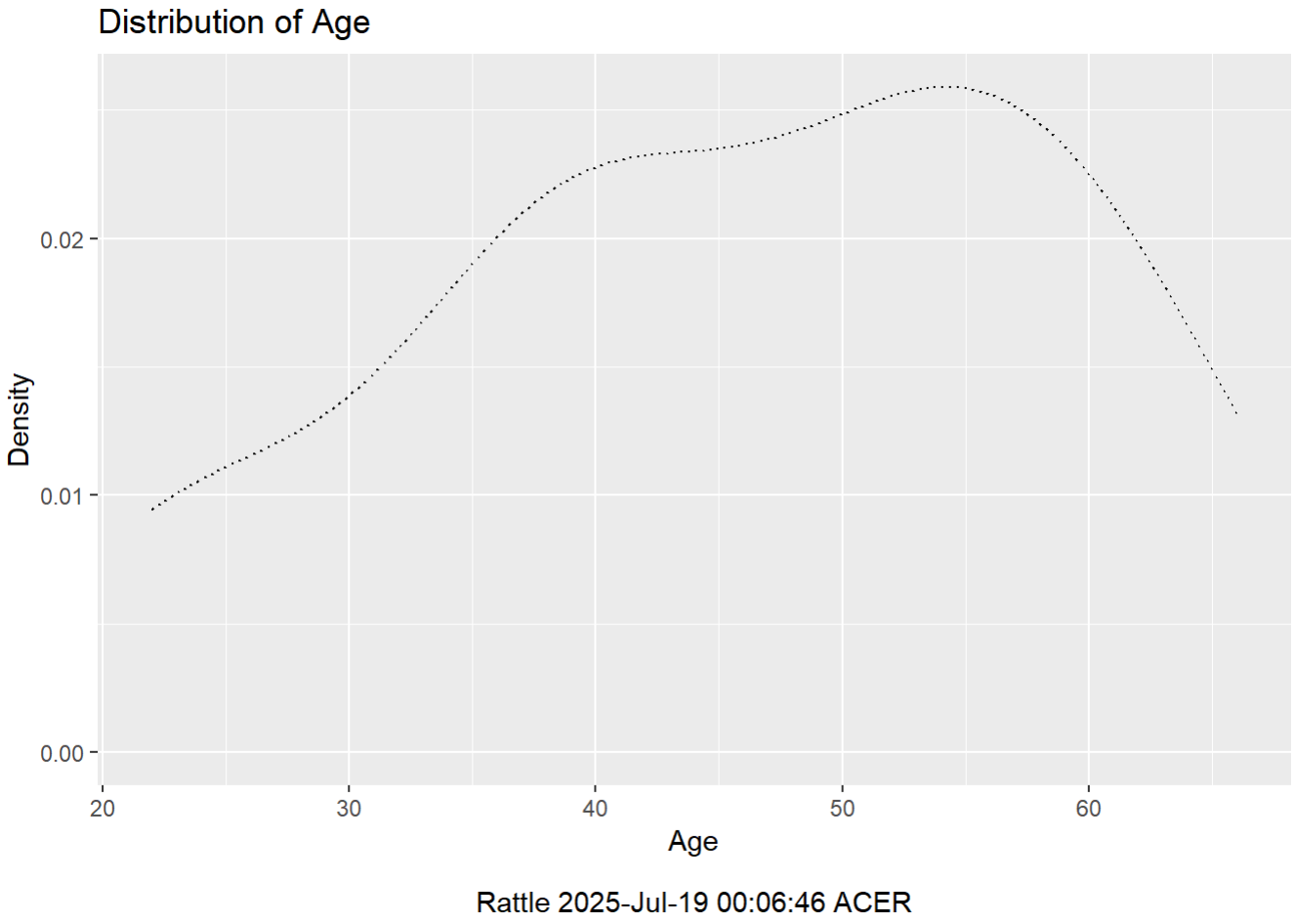
```
# Load data (update the path to your local Excel file)
dataset <- read_excel("DemoKTC.xlsx", guess_max=1e4)
dataset
```

```
# A tibble: 30 × 7
     Age Female Income Married Children  Loan Mortgage
   <dbl>  <dbl>  <dbl>   <dbl>    <dbl> <dbl>    <dbl>
 1    48      1 17546        0        1     0        0
 2    40      0 30085.       1        3     1        1
 3    51      1 16575.       1        0     1        0
 4    23      1 20375.       1        3     0        0
 5    57      1 50576.       1        0     0        0
 6    57      1 37870.       1        2     0        0
 7    22      0  8877.       0        0     0        0
 8    58      0 24947.       1        0     1        0
 9    37      1 25304.       1        2     1        0
10    54      0 24212.       1        2     1        0
# ℹ 20 more rows
```

2.1.1 Age

Age is a continuous numeric variable that reflects the distribution of respondents across life stages. It's useful for identifying generational trends—for instance, younger individuals may exhibit different loan behaviors compared to older ones.

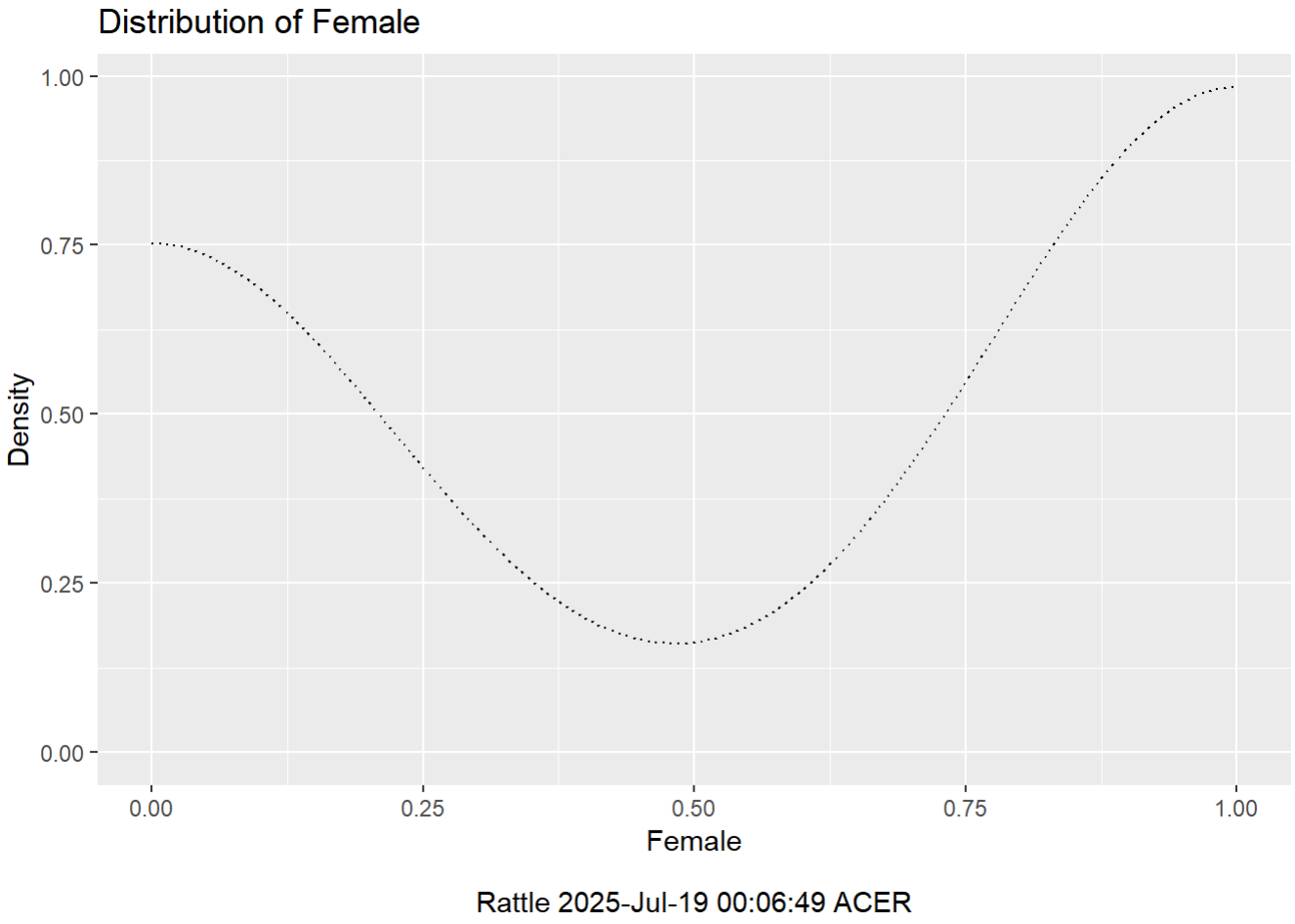
```
p01 <- dataset %>%
  dplyr::select(Age) %>%
  ggplot2::ggplot(ggplot2::aes(x=Age)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Age\n\nRattle 2025-Jul-19 00:06:46 ACER") +
  ggplot2::ggtitle("Distribution of Age") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```



2.1.2 Gender

Gender, being a categorical variable, offers insight into the balance or diversity within your dataset. It can influence everything from purchasing decisions to loan eligibility, and exploring its distribution helps ensure fair representation across analyses.

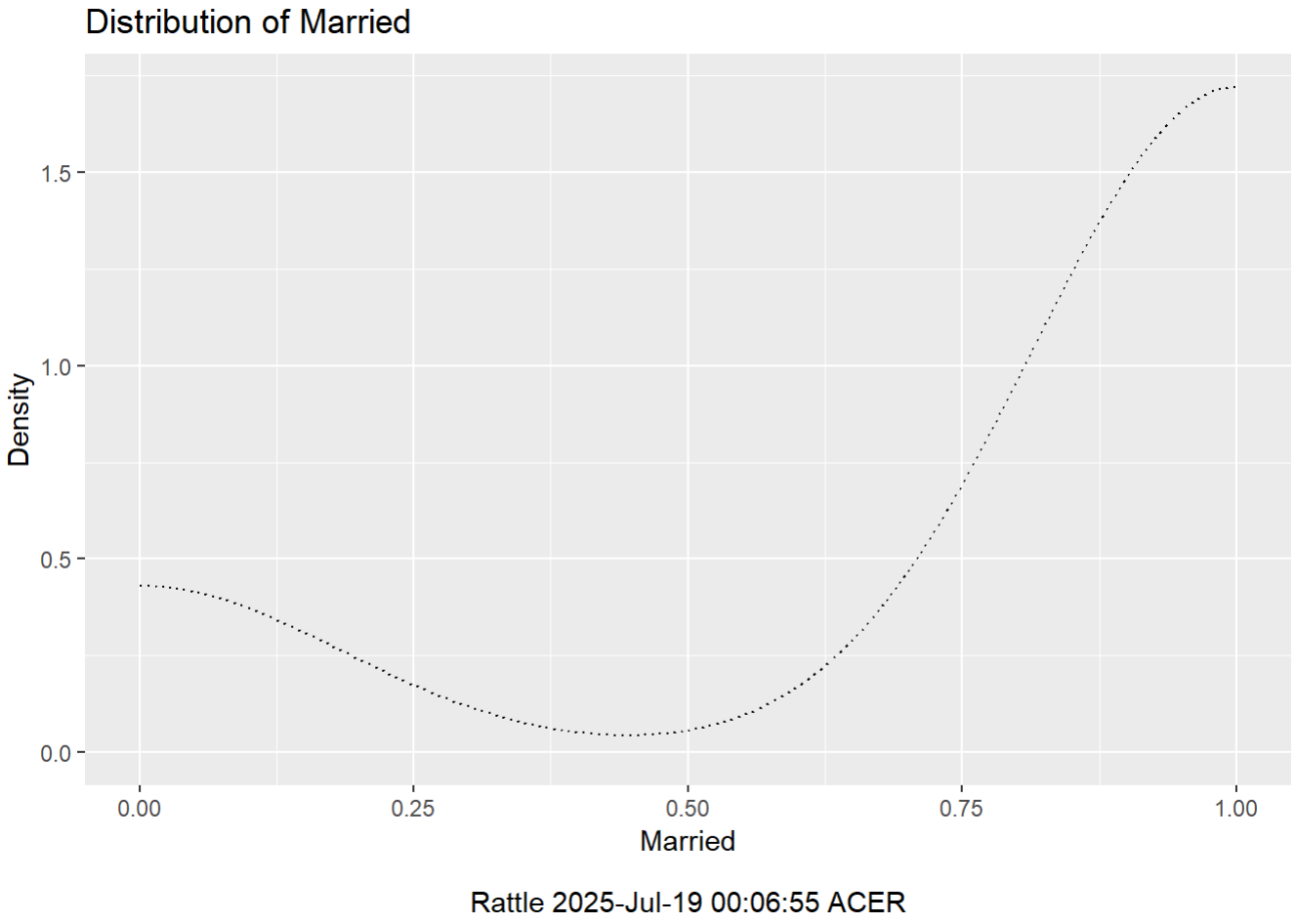
```
p01 <- dataset %>%
  dplyr::select(Female) %>%
  ggplot2::ggplot(ggplot2::aes(x=Female)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Female\n\nRattle 2025-Jul-19 00:06:49 ACER") +
  ggplot2::ggtitle("Distribution of Female") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```



2.1.3 Martial Status

Marital status is another social variable that adds depth to understanding a person's lifestyle. Married individuals often show different financial tendencies compared to singles or divorced respondents—such as higher joint expenses or shared income sources. Coupled with that, the **number of children** an individual has reflects family structure and dependency levels. Larger families might correlate with higher spending and borrowing behavior, while childless individuals may prioritize savings or mobility.

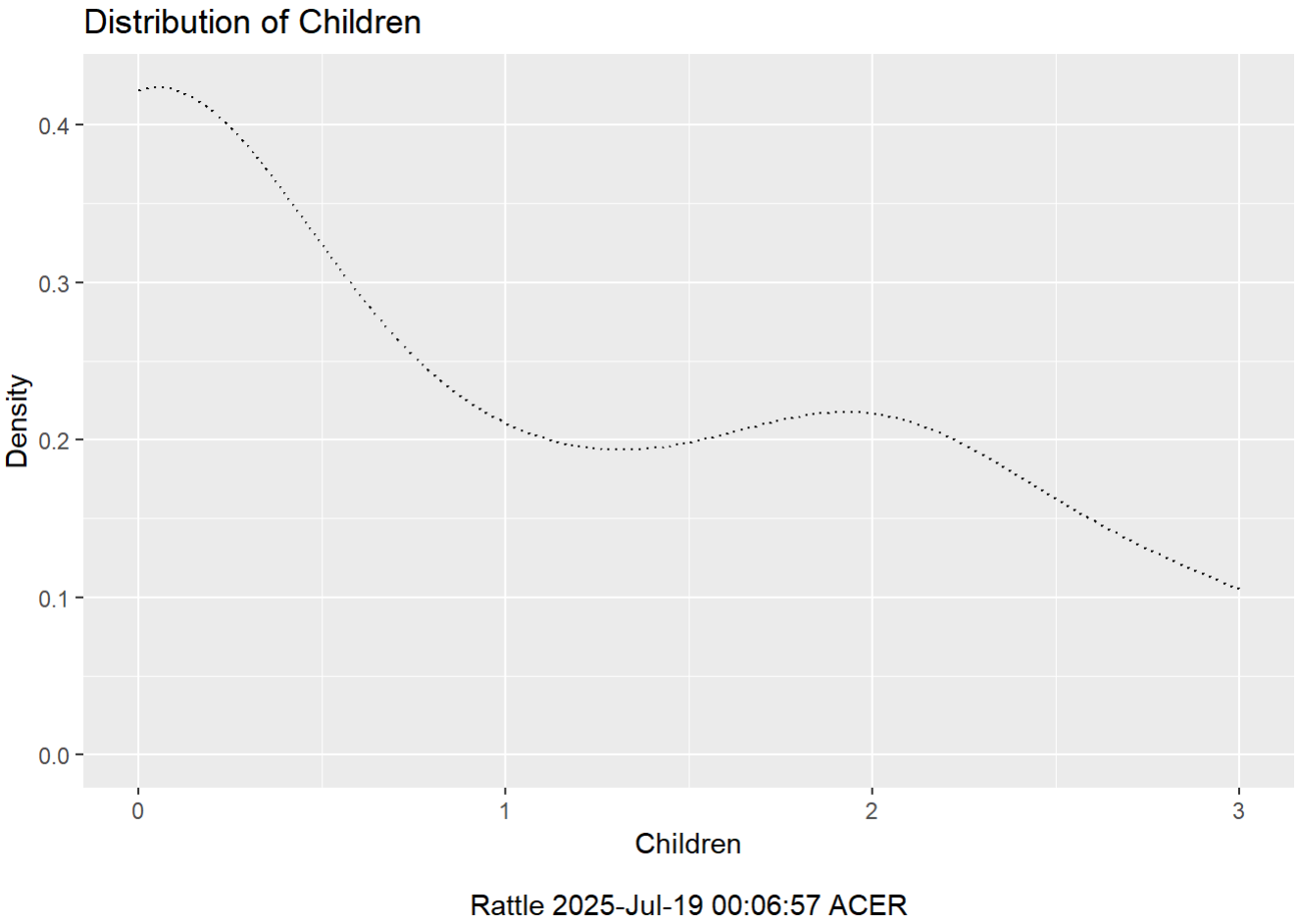
```
p01 <- dataset %>%
  dplyr::select(Married) %>%
  ggplot2::ggplot(ggplot2::aes(x=Married)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Married\n\nRattle 2025-Jul-19 00:06:55 ACER") +
  ggplot2::ggtitle("Distribution of Married") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```



2.1.4 Number of Children

The **number of children** an individual has reflects family structure and dependency levels. Larger families might correlate with higher spending and borrowing behavior, while childless individuals may prioritize savings or mobility.

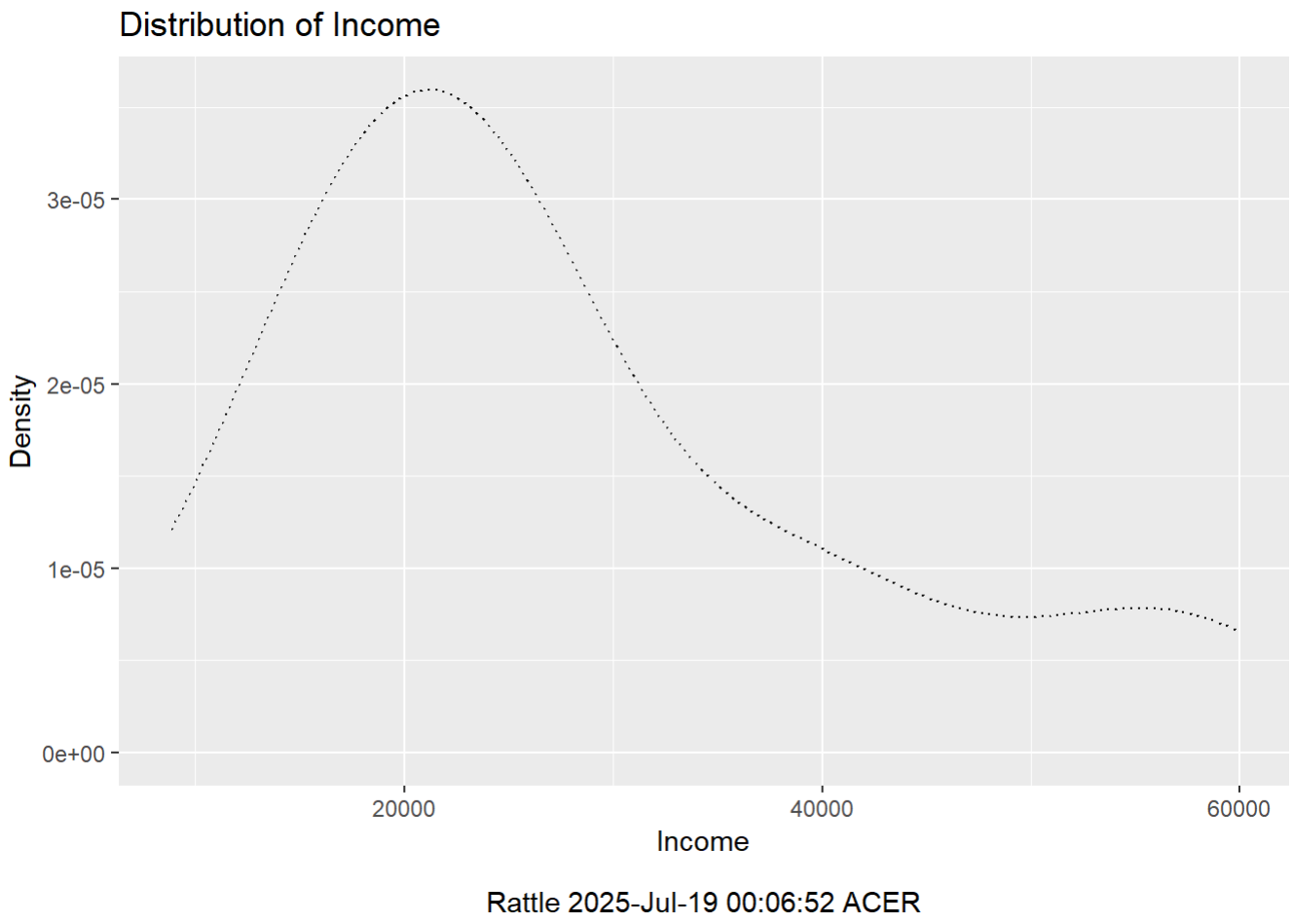
```
p01 <- dataset %>%
  dplyr::select(Children) %>%
  ggplot2::ggplot(ggplot2::aes(x=Children)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Children\n\nRattle 2025-Jul-19 00:06:57 ACER") +
  ggplot2::ggtitle("Distribution of Children") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```



2.1.5 Income

The backbone of financial analysis is **income**—a numeric variable that speaks to purchasing power and overall economic stability. Examining income distribution helps you detect inequalities, identify affluent groups, and understand who might qualify for loans or mortgages.

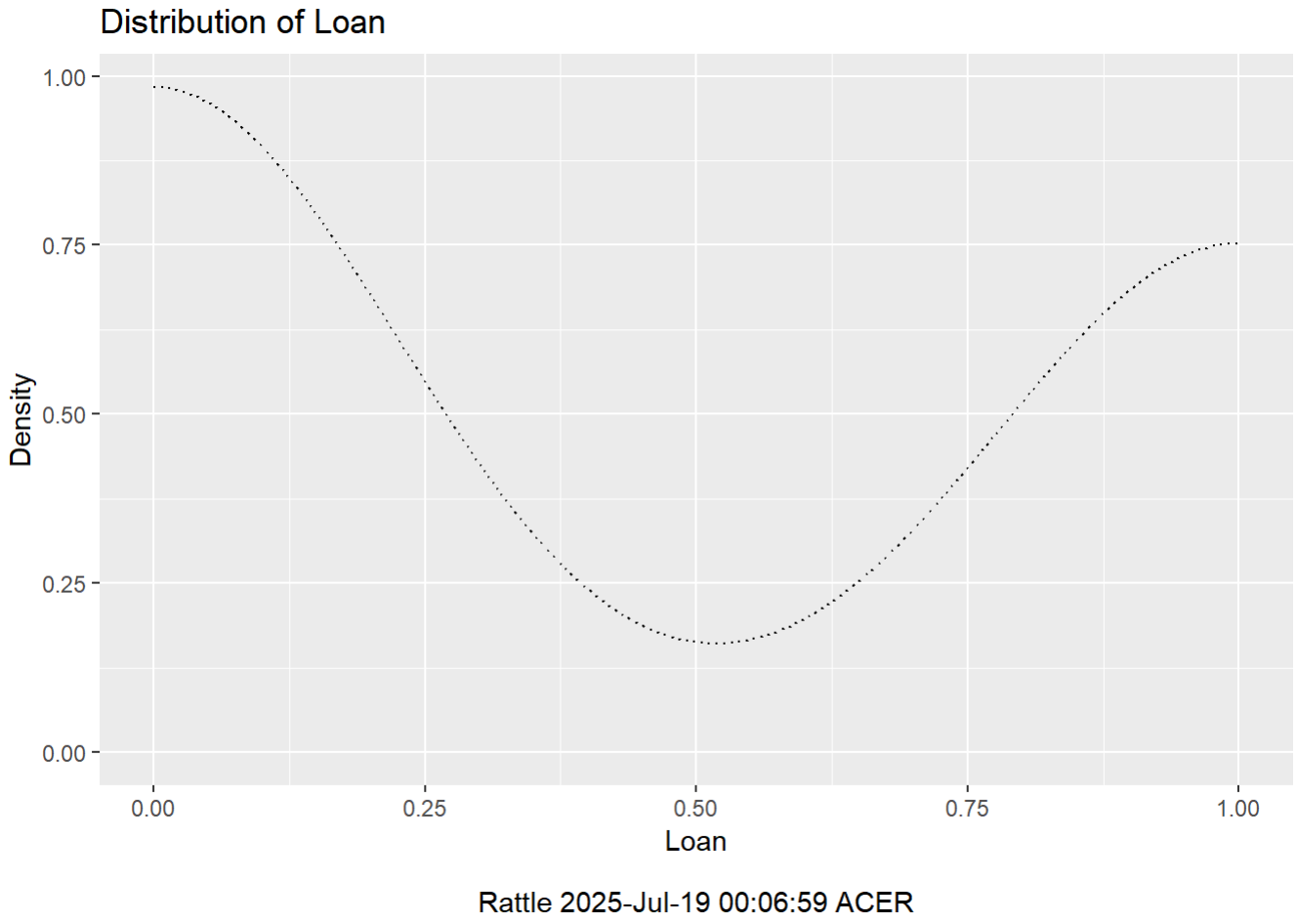
```
p01 <- dataset %>%
  dplyr::select(Income) %>%
  ggplot2::ggplot(ggplot2::aes(x=Income)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Income\n\nRattle 2025-Jul-19 00:06:52 ACER") +
  ggplot2::ggtitle("Distribution of Income") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```



2.1.6 Car Loan

Car loans, typically recorded as yes/no answers, indicate shorter-term credit use, often linked to mobility needs and income level.

```
p01 <- dataset %>%
  dplyr::select(Loan) %>%
  ggplot2::ggplot(ggplot2::aes(x=Loan)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Loan\n\nRattle 2025-Jul-19 00:06:59 ACER") +
  ggplot2::ggtitle("Distribution of Loan") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```

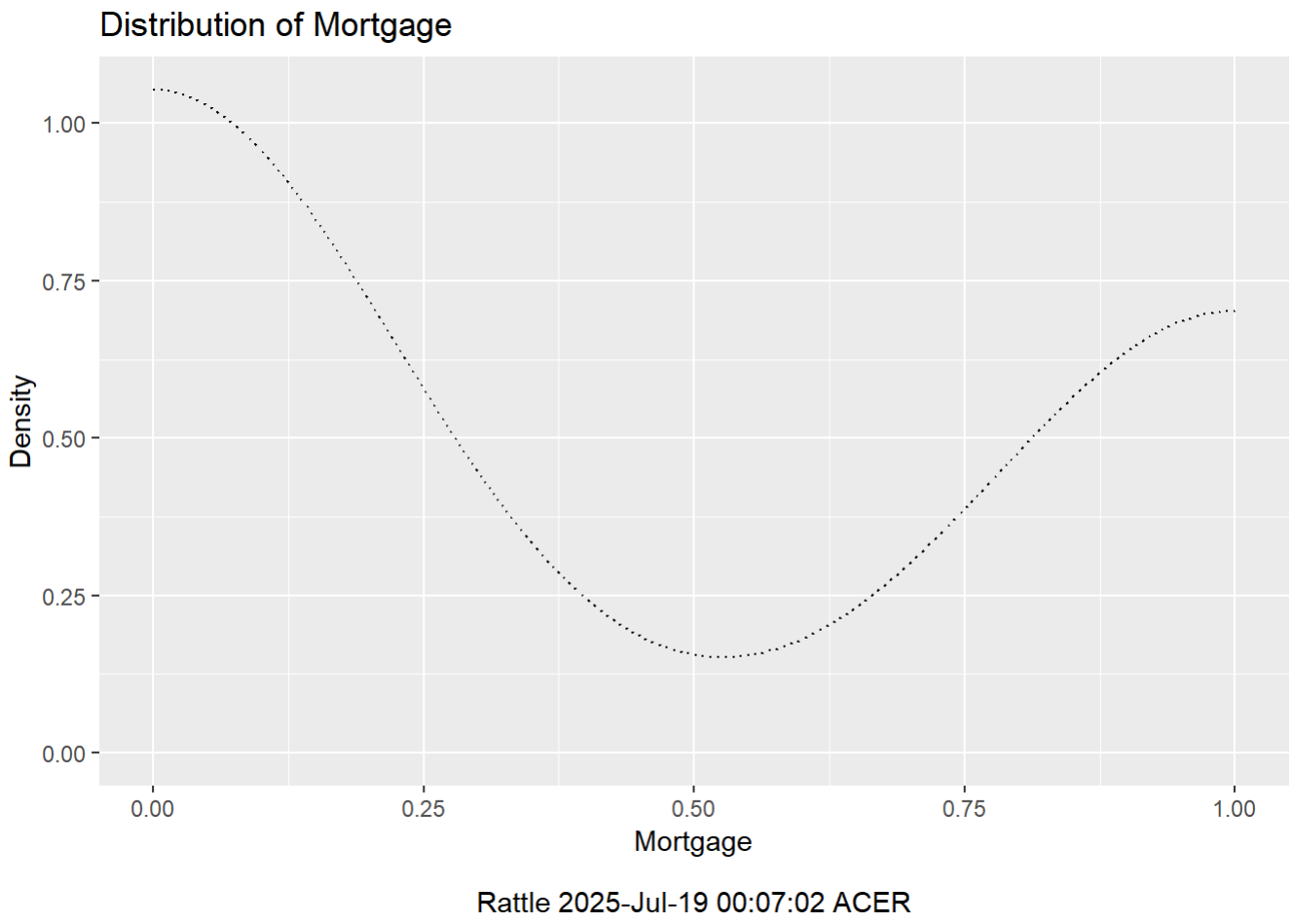


2.1.7 Mortgage

Lastly, **mortgages** reflect long-term financial commitments and asset ownership. Those with mortgages are likely to have stable income and long-term residency plans, which can influence many lifestyle and financial choices.

Exploring these variables—individually and in relation to one another—paints a rich picture of demographic and financial realities. It guides you to recognize trends, segment the population meaningfully, and prepare for deeper analysis like clustering or predictive modeling. Let me know if you'd like a written summary of how these factors relate in your specific dataset—I'd be happy to help synthesize that!

```
p01 <- dataset %>%
  dplyr::select(Mortgage) %>%
  ggplot2::ggplot(ggplot2::aes(x=Mortgage)) +
  ggplot2::geom_density(lty=3) +
  ggplot2::xlab("Mortgage\n\nRattle 2025-Jul-19 00:07:02 ACER") +
  ggplot2::ggtitle("Distribution of Mortgage") +
  ggplot2::labs(y="Density")

# Display the plots.

gridExtra::grid.arrange(p01)
```



2.2. Data Summarizing:

The dataset presents a diverse profile of individuals characterized by varying age, gender, marital status, family structure, income levels, and financial obligations. The **age** of respondents spans a wide range, reflecting different life stages—from young adults in their 20s to individuals nearing retirement. This diversity enables observations related to financial maturity, such as how age affects loan ownership or income stability. **Gender distribution** appears balanced across the dataset, representing both male and female participants, with the potential inclusion of non-binary individuals, which helps capture societal dynamics and ensure inclusive insights.

```
# Rattle timestamp: 2025-07-24 22:07:58.108785 x86_64-w64-mingw32

# The 'Hmisc' package provides the 'contents' function.

library(Hmisc, quietly=TRUE)
```

```
Attaching package: 'Hmisc'

The following objects are masked from 'package:dplyr':

    src, summarize

The following objects are masked from 'package:base':

    format.pval, units
```

```
# Obtain a summary of the dataset.

contents(dataset[, c(input_vars, risk_vars, target_vars)])
```

```
Data frame:dataset[, c(input_vars, risk_vars, target_vars)] 30 observations and 4 variables
Maximum # NAs:0

         Storage
Age       double
Income    double
Loan      double
Mortgage  double
```

```
summary(dataset[, c(input_vars, risk_vars, target_vars)])
```

```
      Age             Income            Loan           Mortgage
 Min.   :22.00   Min.   : 8877   Min.   :0.0000   Min.   :0.0
 1st Qu.:37.25   1st Qu.:18166   1st Qu.:0.0000   1st Qu.:0.0
 Median :47.00   Median :24241   Median :0.0000   Median :0.0
 Mean   :45.97   Mean   :28012   Mean   :0.4333   Mean   :0.4
 3rd Qu.:56.75   3rd Qu.:35923   3rd Qu.:1.0000   3rd Qu.:1.0
 Max.   :66.00   Max.   :59804   Max.   :1.0000   Max.   :1.0
```

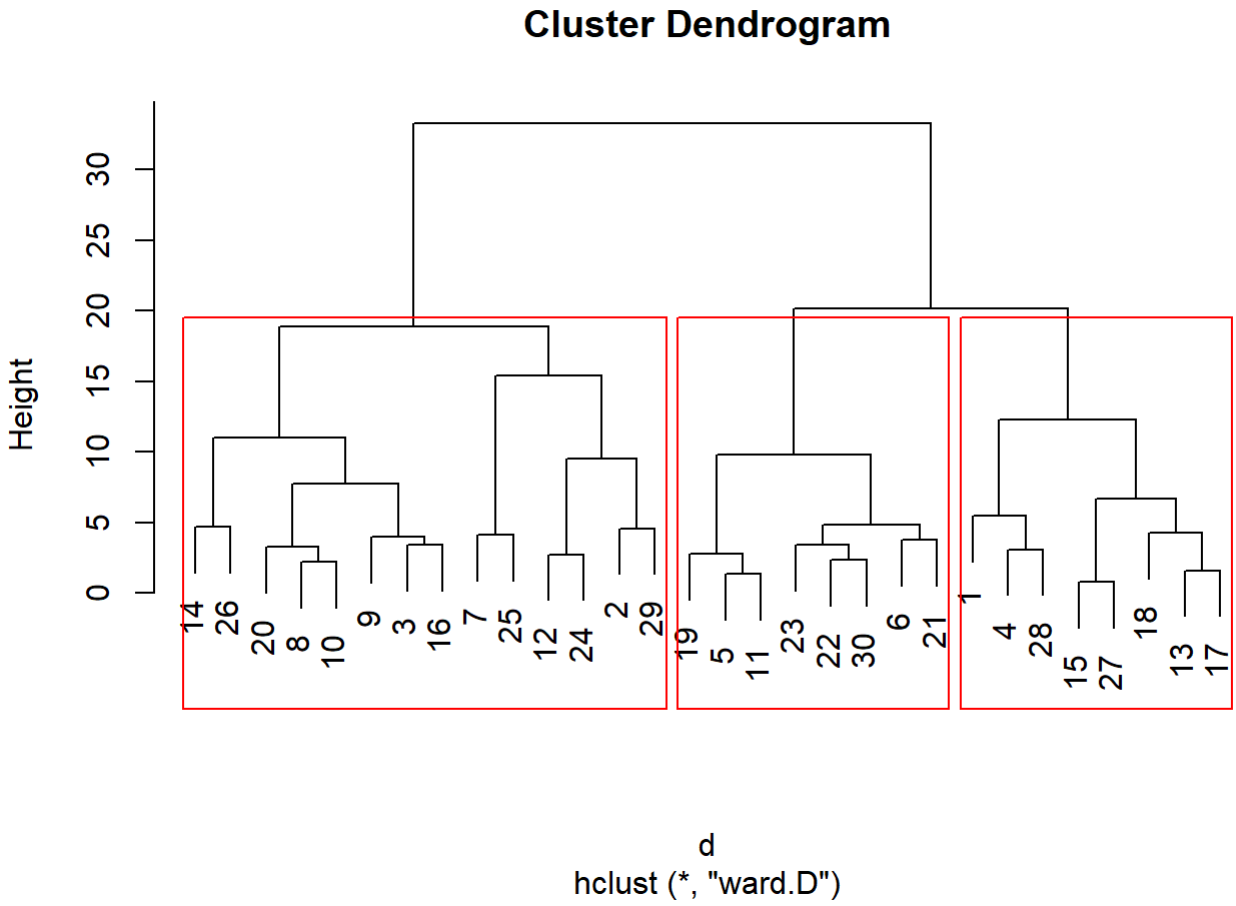
2.3. Segmentation using Clustering:

Clustering is a method of grouping the observation based on their similaraties. We use distance measures for assessing the dissimilarity among the observations. There are many measures of distance including Euclidean, Manhattan, Mahanlobis etc.,

```
library(readxl)
DemoKTC <- read_excel("C:/Users/ACER/Documents/DemoKTC.xlsx")
mydata<-scale(DemoKTC)
d <- dist(mydata, method = "manhattan") # distance matrix
fit <- hclust(d, method="ward") # Clustering
```

```
The "ward" method has been renamed to "ward.D"; note new "ward.D2"
```

```
plot(fit) # display dendogram
groups <- cutree(fit, k=5) # cut tree into 5 clusters
#draw dendogram with red borders around the 5 clusters
rect.hclust(fit, k=3, border="red")
```

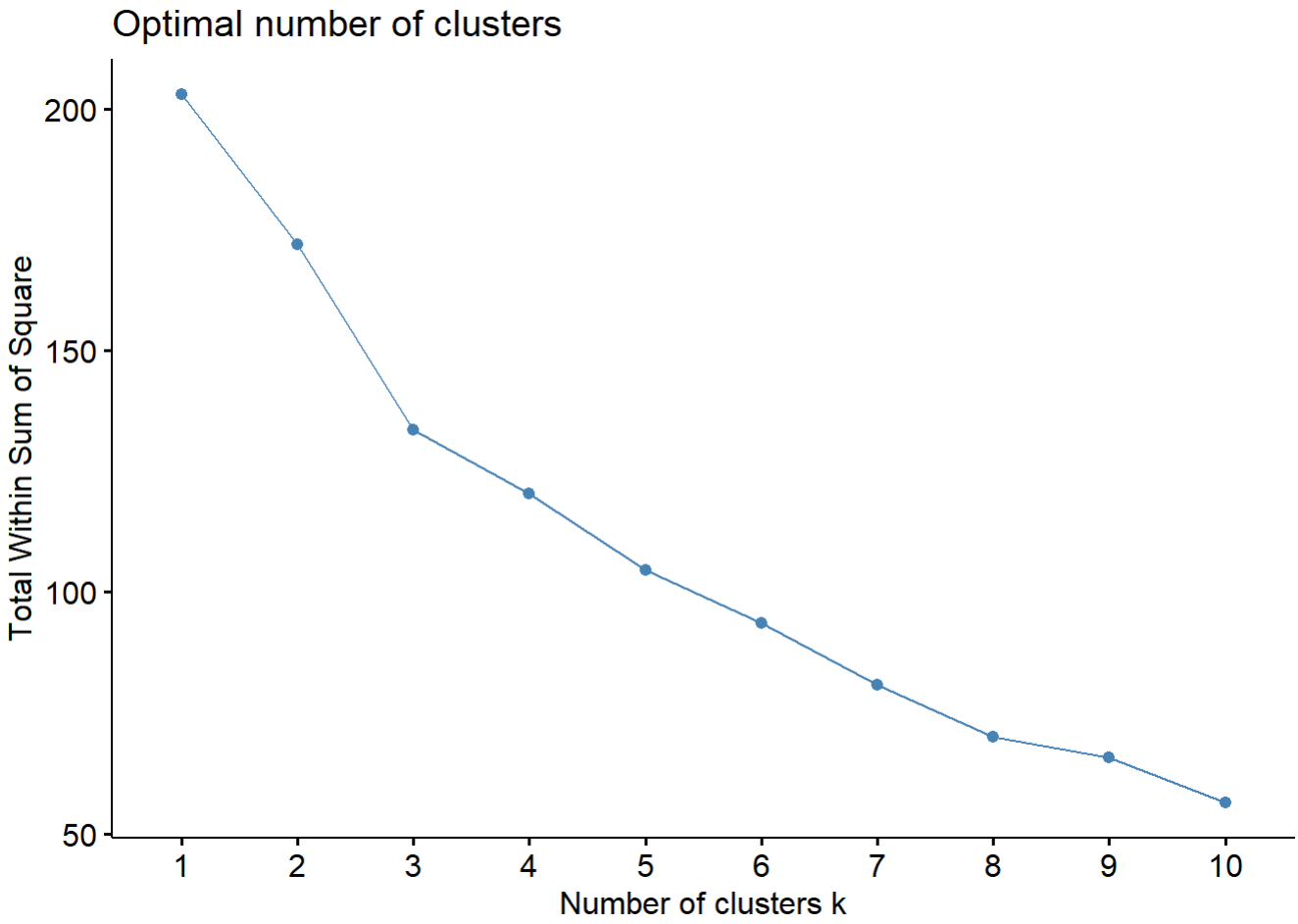


The segmentation process involves two clustering techniques. First, hierarchical clustering is performed using a distance matrix calculated with the Manhattan method. This technique groups similar observations based on their proximity and visualizes the structure through a dendrogram. By cutting the dendrogram at five branches, the analysis identifies five preliminary customer groups.

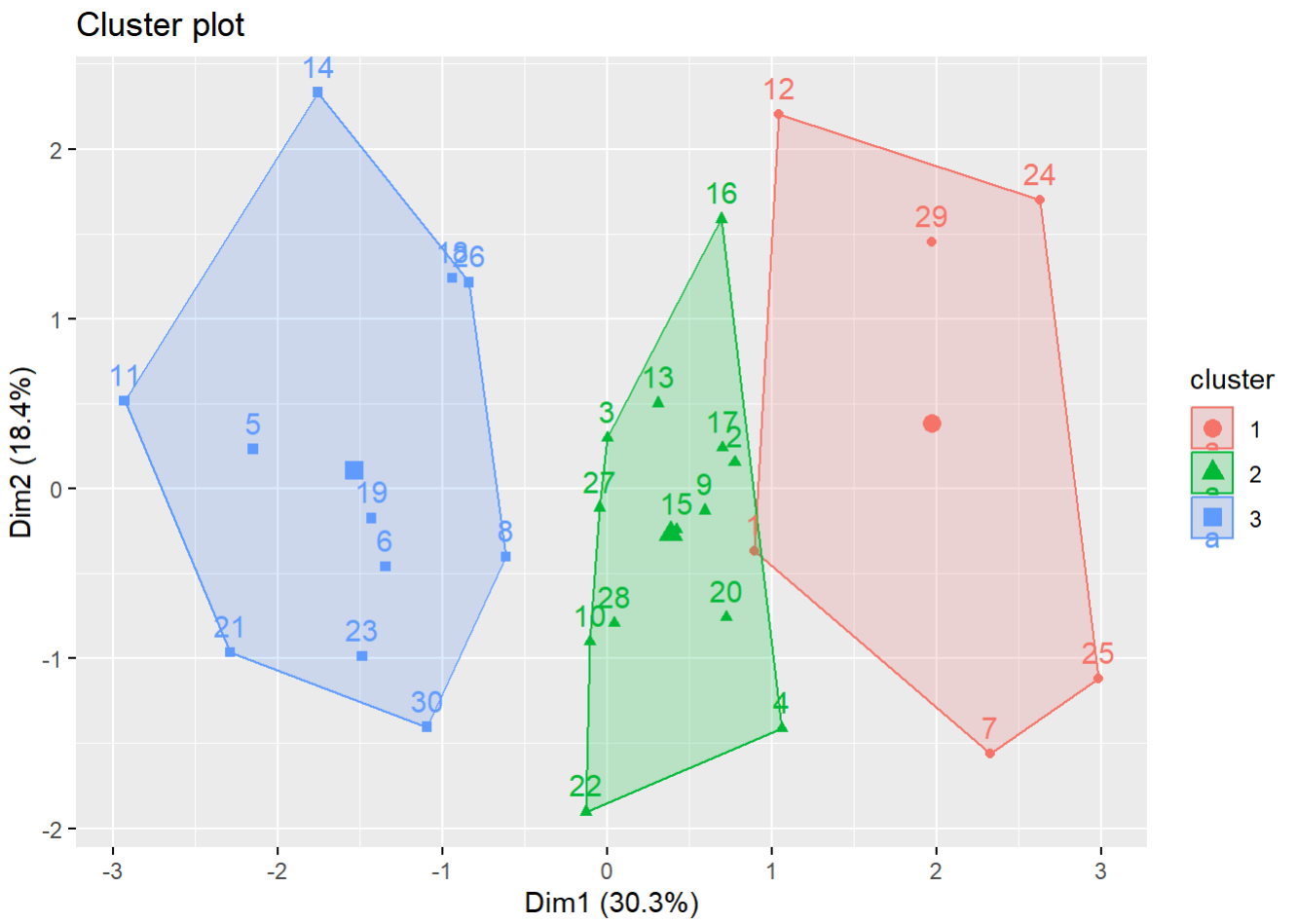
```
library(factoextra)
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(cluster)
library(readxl, quietly=TRUE)
mydata<- DemoKTC
data <- scale(mydata)
fviz_nbclust(data, kmeans, method = "wss")
```




```r
set.seed(123)  # For reproducibility
km <- kmeans(data, centers = 3, nstart = 25)
set.seed(123)  # For reproducibility
km <- kmeans(data, centers = 3, nstart = 25)
fviz_cluster(km, data)
```




```
data2<-data# duplicating the data
data2$cluster<-km$cluster# writing the cluster membership in to the data
```

```
Warning in data2$cluster <- km$cluster: Coercing LHS to a list
```

```
data2$cluster
```

```
 [1] 1 2 2 2 3 3 1 3 2 2 3 1 2 3 2 2 2 3 3 2 3 2 3 1 1 3 2 2 1 3
```

Next, K-Means clustering is used for a more precise segmentation. The data is scaled to ensure consistency across variables. The optimal number of clusters is determined using the Elbow method, which suggests three clusters. K-Means clustering is then applied, producing compact and interpretable groups. The clusters are visualized in two-dimensional space, making it easier to understand their distribution and separation.

```
# Define input and numeric variables manually
# Replace these with the actual column names from your dataset
numeric_vars <- c("Age", "Income", "Loan", "Mortgage")  # example placeholder
input_vars <- numeric_vars

# Rattle timestamp: 2025-07-17 12:27:46.081723 x86_64-w64-mingw32

# KMeans

# Reset the random number seed to obtain the same results each time.
```

```
set.seed(123)

# The 'reshape' package provides the 'rescaler' function.

library(reshape, quietly=TRUE)
```

```
Attaching package: 'reshape'

The following object is masked from 'package:dplyr':

    rename
```

```
# Generate a kmeans cluster of size 3.

kmeans_result <- kmeans(sapply(na.omit(dataset[, numeric_vars]), rescaler, "range"), 3)
```

After clustering, the report presents a summary of the cluster characteristics. It includes the number of customers in each cluster and their average attribute values. This step highlights differences among customer types, such as one cluster consisting of high-income individuals with large loans, while another may represent younger, lower-income customers with fewer financial obligations. The original dataset is then divided into three subsets—one for each cluster—to support more targeted business actions.

```
# Define input and numeric variables manually
# Replace these with the actual column names from your dataset
numeric_vars <- c("Age", "Income", "Loan", "Mortgage")  # example placeholder
input_vars <- numeric_vars

# Rattle timestamp: 2025-07-17 12:27:46.252969 x86_64-w64-mingw32

# Report on the cluster characteristics.

# Cluster sizes:

paste(kmeans_result$size, collapse=' ')
```

```
[1] "5 18 7"
```

```
# Data means:

colMeans(sapply(na.omit(dataset[, numeric_vars]), rescaler, "range"))
```

```
      Age    Income      Loan  Mortgage
0.5446970 0.3757313 0.4333333 0.4000000
```

```
# Cluster centers:

kmeans_result$centers
```

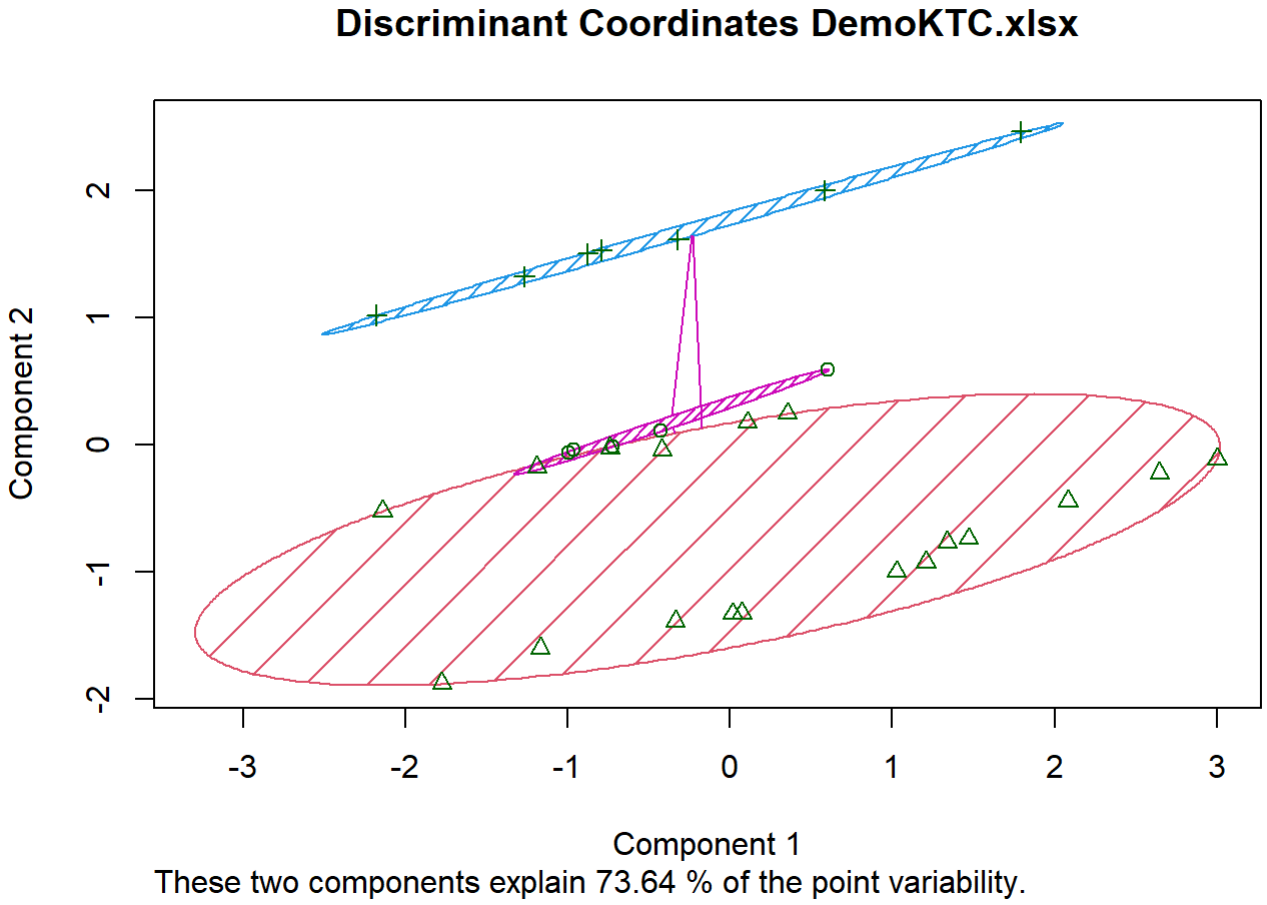
```
        Age    Income      Loan Mortgage
1 0.4454545 0.2763987 0.0000000        1
```

```
2 0.5770202 0.3769901 0.3333333        0
3 0.5324675 0.4434460 1.0000000        1
```

```
# Within cluster sum of squares:

kmeans_result$withinss
```

```
[1] 0.2089153 7.3552765 0.9329517
```

```
# Time taken: 0.00 secs

# Generate a discriminant coordinates plot.

cluster::clusplot(na.omit(dataset[, intersect(input_vars, numeric_vars)]), kmeans_result$clust
```



Finally, the analysis concludes that the segmentation has successfully identified meaningful customer groups. These clusters provide actionable insights for personalized marketing, risk assessment, and product alignment. KTC can now use these findings to build more focused and effective customer strategies.

```
data2<-mydata# duplicating the data
cluster_id<-as.vector(unlist(km$cluster))# writing the cluster membership in to the data
data2<-as.data.frame(cbind(data2,cluster_id))

# Group data2 by cluster_id and compute mean for each group
```

```r
group_means <- aggregate(. ~ cluster_id, data = data2, FUN = mean)

# Split the original data into a list of data frames by cluster_id
grouped_data <- split(data2, data2$cluster_id)

# If we specifically want 3 data sets, we can extract them like this:
data_cluster1 <- grouped_data[[1]]
data_cluster2 <- grouped_data[[2]]
data_cluster3 <- grouped_data[[3]]

# Optionally view the group means
print(group_means)
```

```
  cluster_id      Age    Female   Income Married  Children      Loan  Mortgage
1          1 35.00000 0.6666667 18436.68       0 1.0000000 0.6666667 0.5000000
2          2 40.38462 0.5384615 20906.03       1 1.2307692 0.4615385 0.4615385
3          3 58.54545 0.5454545 41632.52       1 0.5454545 0.2727273 0.2727273
```

Conclusion:

The customer segmentation analysis successfully grouped the KTC Company's customers into distinct clusters based on key demographic and financial attributes such as age, income, loan amount, and mortgage status. Using the K-Means clustering algorithm, we were able to uncover meaningful patterns in customer behavior and identify segments with similar characteristics.

These insights provide a foundation for more targeted marketing strategies, personalized customer engagement, and optimized product offerings. For example, one cluster may represent younger customers with lower income and smaller loans, while another may consist of older, high-income individuals with substantial mortgages each requiring a different communication and service approach.

Ultimately, this segmentation enables KTC to shift from a one size fits all model to a data driven, customer-centric strategy enhancing both customer satisfaction and business efficiency.

About the Author

I'm a student in the MBA Program (2025–2027), currently navigating Trimester I at Amrita School of Business, Amrita Vishwa Vidyapeetham, Coimbatore. This **assignment/blog** was written as part of our coursework for **Introduction to Business Analytics.**

Blog link: https://github.com/Bala-Shunmugam-M/Bala-Shunmugam-M