K-MEANS: Clustering of Loan Applicants

Assignment 5

Alex Enriquez

OCTOBER 2024

PROBLEM RESTATEMENT

The purpose of this analysis is to segment loan applicants based on demographic and socio-economic data. The objective is to group customers with similar characteristics to improve customer targeting and tailor loan offers according to their profiles.

By applying data mining techniques, specifically K-means clustering, the aim is to identify patterns and group customers with similar characteristics. This segmentation will allow the loan company to better understand its customer base, enabling it to make informed decisions about loan offerings, marketing strategies, and customer engagement.

Loan companies often need to assess various customer attributes such as age, gender, education level, income, and occupation to predict customer behavior, including loan repayment likelihood and customer lifetime value. By clustering customers into different segments, the company can target these segments with tailored loan products, enhancing customer satisfaction and optimizing its marketing efforts.

DATA MINING TECHNIQUES – STEPS FOLLOWED

1. Data Preprocessing

The first step was to import the customer data from a CSV file called *CustomerProfileData.csv*. The dataset contained several features such as gender, age, marital status, education, income, occupation, and size of the community. To ensure accurate clustering, irrelevant information, such as the unique customer identifier (CustomerNo), was removed from the analysis.

This step also involved handling any missing or inaccurate data, though the dataset provided was already clean and ready for analysis.

2. Scaling the Data

Since K-means clustering relies on distance measures such as Euclidean distance, it is sensitive to the scale of the variables. The features of age and income had different ranges and units, which could skew the clustering process. To address this, all the features were standardized (scaled) to have a mean of 0 and a standard deviation of 1. This ensured that all variables contributed equally to the clustering algorithm.

The *scale()* function in R was used to transform the data before applying the clustering algorithm.

3. Application of K-means Clustering

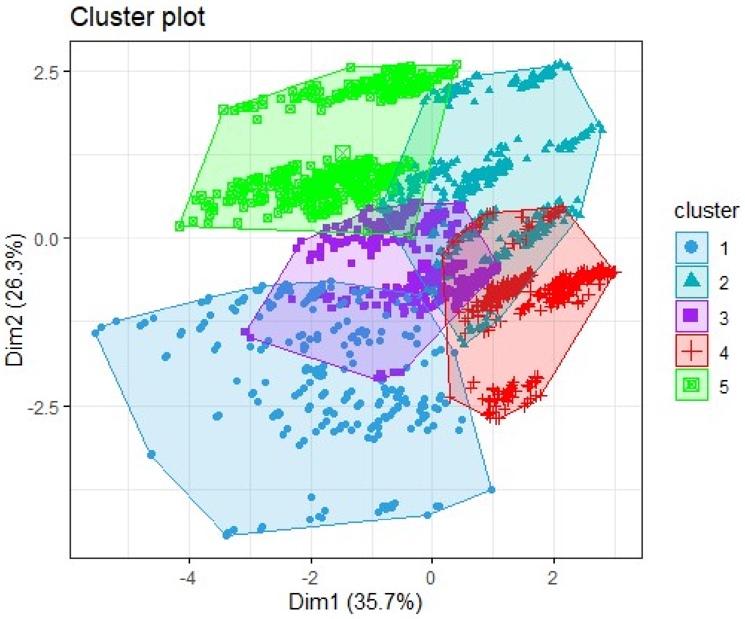
Once the data was prepared and scaled, the K-means clustering algorithm was applied in R. The algorithm partitions the dataset into a predefined number of clusters by minimizing the within-cluster variance. For this analysis, a number of five clusters was initially chosen based on the nature of the customer data and initial evaluations.

The following steps were used to apply K-means clustering in R:

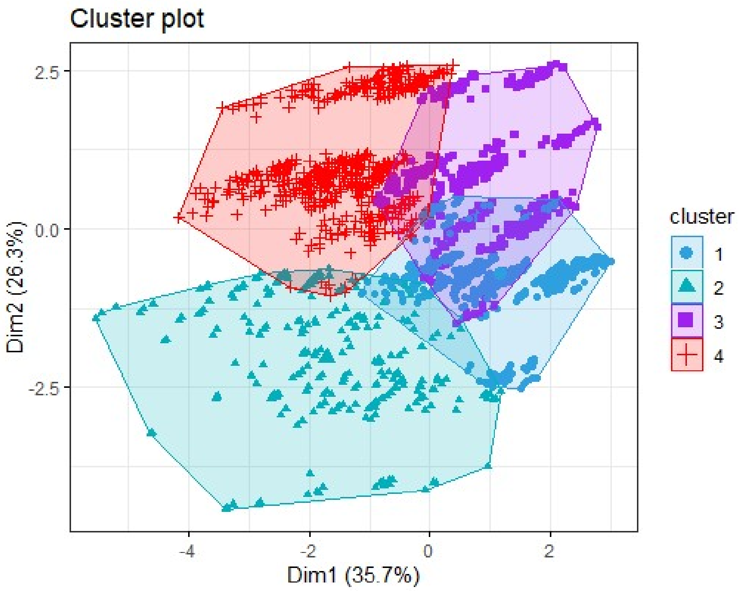
- The data was divided into 5 clusters (later changed to 4) using the *kmeans()* function.
- The *nstart* parameter was set to 25, ensuring that the algorithm ran 25 different initial configurations and chose the best clustering solution.
- After running the algorithm, each customer was assigned to a specific cluster, with the cluster centers representing the “average” customer in each group.

4. Evaluation and Visualization

After clustering the customers, the next step involved evaluating the results by visualizing the clusters. Using the *fviz_cluster()* function from the *factoextra* package in R, I was able to generate a scatter plot to visualize the clusters. This plot helped identify how separated the clusters were and if there were any overlapping clusters that required further refinement.



From the initial plot, it shows there are some overlaps between clusters, particularly between clusters 3 and 4. Refinements such as adjusting the number of clusters was then applied to see if separation between clusters improved. The clusters were reduced to **4** as the main selection to analyze.



4a). Cluster Characteristics

To analyze the clusters in more detail, the clustered data was exported to Excel to review the customer characteristics in each group. The mean values of the key demographic and socio-economic variables for each cluster were calculated, revealing the following insights:

Cluster 1:

- Sex: Majority are female
- Marital Status: Nearly all are not single, meaning they're married, divorced, separated, or widowed.
- Age: Young (average age ~29).
- Education: High school only
- Income: Moderate income (Mean $105,759).
- Occupation: Primarily skilled workers
- Size of Community: Mostly from small to mid-sized cities

Cluster 2:

- Sex: Fairly balanced in terms of gender (~0.50), meaning an even split between those identifying as female and those who do not.
- Marital Status: Most are not single

- Age: Older group (~56 years old).
- Education: Primarily university graduates
- Income: The highest income group (~$158,338).
- Occupation: Primarily skilled workers and management
- Size of Community: Living in mid to large-sized cities

Cluster 3:

- Sex: Predominantly male or non-binary
- Marital Status: Almost all are single
- Age: Younger (~36 years old).
- Education: Some high school graduates but mostly unknown education levels (~0.75).
- Income: The lowest income group (~$97,860).
- Occupation: Predominantly unemployed or in unskilled jobs
- Size of Community: Primarily from small cities or rural areas

Cluster 4:

- Sex: Almost entirely male or non-binary
- Marital Status: Predominantly single
- Age: Middle-aged (~36 years old).
- Education: Mostly high school graduates
- Income: Moderate to high income (~$141,218).
- Occupation: Primarily in management roles.
- Size of Community: Predominantly from large cities.

ANALYSIS OF BEST SEGMENTATION

Cluster 1 Appear to represent young, mostly female, skilled workers with moderate education and income levels, living in smaller or mid-sized communities. These individuals are likely in a more stable life stage but may be a moderate credit risk.

Cluster 2 appears to be an older, well-educated group with the highest income levels, primarily in skilled and management roles in mid-to-large cities. This group would be considered highly creditworthy and could be targeted with premium loan products.

Cluster 3 consists of younger, predominantly male or non-binary individuals, with lower education and income levels. Most are unemployed and live in rural areas. This cluster represents the highest credit risk due to low education and employment levels.

Cluster 4 is a middle-aged, mostly male or non-binary group, primarily in management roles with higher income levels, living in large cities. They are likely good candidates for high-value loan products but may still have some credit risk depending on their other financial circumstances

CLUSTERING DESCRIPTION & OTHER TECHNIQUES

Clustering is a technique used to group similar data points based on their features. Popular clustering methods include:

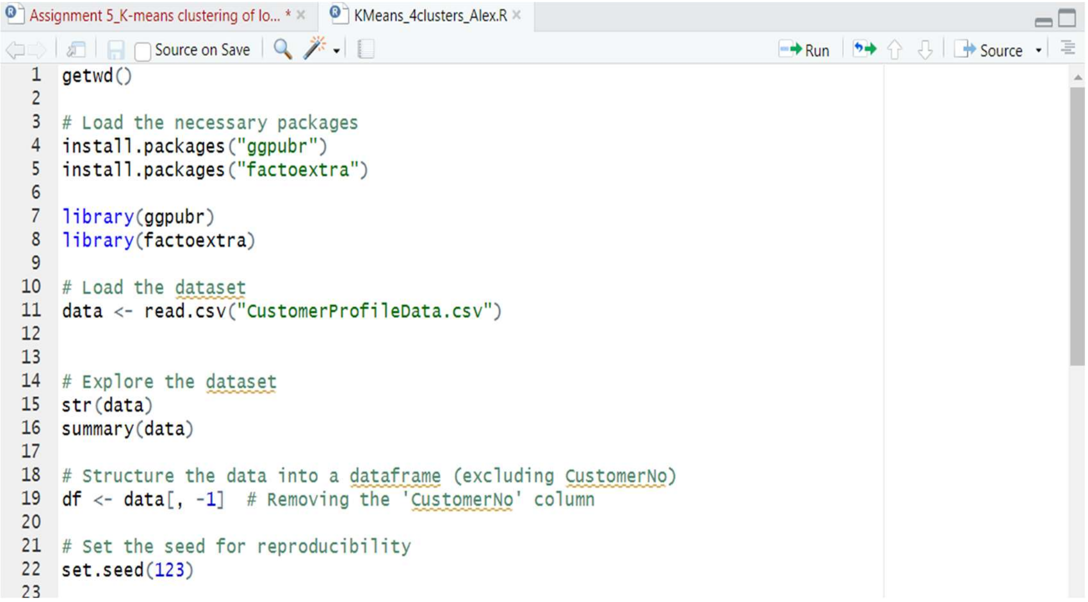
- ***Hierarchical Clustering:*** Groups data based on a nested hierarchy, suitable for small datasets.
- ***DBSCAN****:* Identifies clusters based on density, useful for data with noise.
- ***K-means Clustering***: Assigns data points into a pre-defined number of clusters, minimizing intra-cluster variance. K-means was selected for this analysis due to its simplicity and effectiveness in handling large datasets like customer data.

APPENDIX

1) Excel File Output (4 Clusters – Link)
   ClusteredData_4Clusters.xlsx

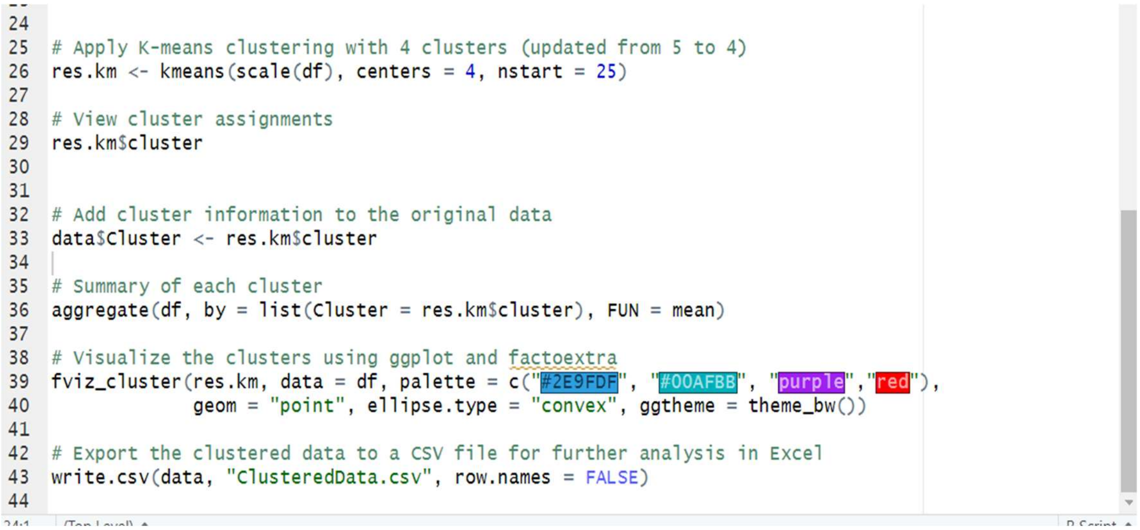
2) *R CODE SCRIPT (snippets)*

a). Loading Data & Prepping Data



```r
getwd()

# Load the necessary packages
install.packages("ggpubr")
install.packages("factoextra")

library(ggpubr)
library(factoextra)

# Load the dataset
data <- read.csv("CustomerProfileData.csv")


# Explore the dataset
str(data)
summary(data)

# Structure the data into a dataframe (excluding CustomerNo)
df <- data[, -1]  # Removing the 'CustomerNo' column

# Set the seed for reproducibility
set.seed(123)

```

b.) K-Means Application Steps



```r

# Apply K-means clustering with 4 clusters (updated from 5 to 4)
res.km <- kmeans(scale(df), centers = 4, nstart = 25)

# View cluster assignments
res.km$cluster


# Add cluster information to the original data
data$Cluster <- res.km$cluster

# Summary of each cluster
aggregate(df, by = list(Cluster = res.km$cluster), FUN = mean)

# Visualize the clusters using ggplot and factoextra
fviz_cluster(res.km, data = df, palette = c("#2E9FDF", "#00AFBB", "purple","red"),
             geom = "point", ellipse.type = "convex", ggtheme = theme_bw())

# Export the clustered data to a CSV file for further analysis in Excel
write.csv(data, "ClusteredData.csv", row.names = FALSE)

```

.