MAST5954 Assessment 2: Final

Project Report

**Introduction**

The aim of this report is to use data analysis techniques to gain an understanding of the data set and predict a response variable from the seven predictors C1 to C7.

In this report, the data set provided is untitled with one response variables and seven dependent variables which are all numeric values including one variable (C7) which has a binary of 0 and 1. Due to the nature of the data set, the best approach decided was a through explanatory data analysis to explore the relationships each dependent variable had with the response (Y) variables. As well as the relationship between each of the dependent variables.

After each section of the data set is explored, the data analysis then on is determined by previous results. Once results are analysed, we can interpret them and use the information gained to infer what our Y variable could be in response to the other predictor variables.

Firstly, to achieve our aim a general explanatory data analysis (EDA) is conducted followed by a further EDA that allows exploration of the relationships between variables. After, results can influence any changes to our statistical methods, and this is leads to building hypotheses and using linear regression models and clustering to further breakdown the data. The final step after building an appropriate model is to use that model to assess if it can predict response variables using new data.

**Exploratory Analysis through Descriptive statistics and Graphical summaries**

First, the data set was loaded into R and checked for any missing values. Variables C1, C2, C3, C6, and C7 were found to have one missing value each. Variable C4 had two missing values and C5 had none. This was then corrected using the average value for each variable while maintaining the binary values in variable C7. After the data is cleaned and a data summary is shown, it is ready for our initial EDA where graphs are plotted to show the distribution of data for each variable. These plots were followed by a boxplot of all the dependent variables for direct comparison.

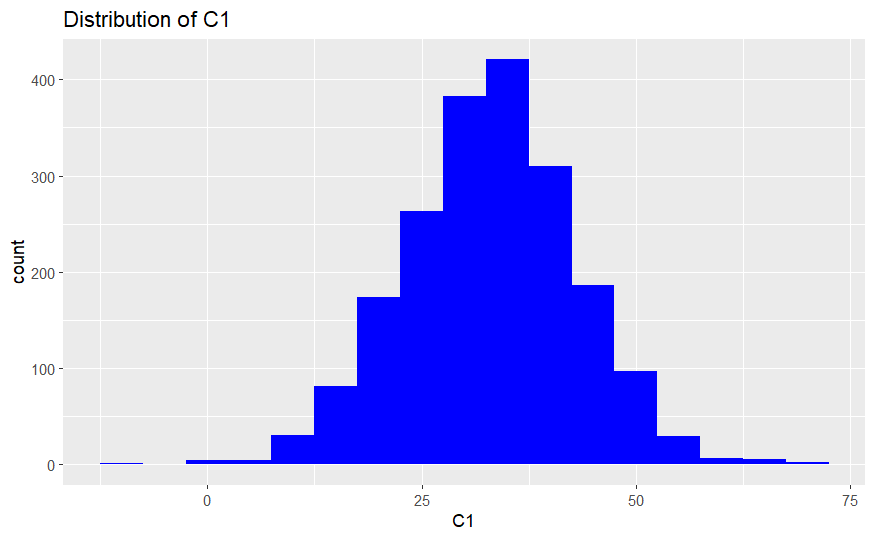
|  |  |
| --- | --- |
| Variable | Average |
| Y | 854.4126 |
| C1 | 32.99384 |
| C2 | 103.4392 |
| C3 | 115.8035 |
| C4 | 22.80039 |
| C5 | -0.502065 |
| C6 | 11.35267 |

**Table 1: table of the average value for each variable**

A graph of a number of numbers

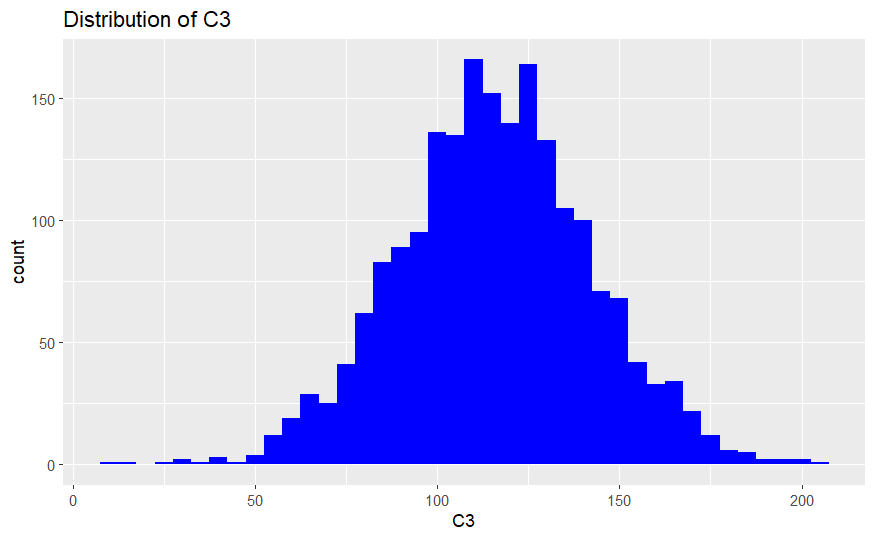
Description automatically generated Figure 1.1: Distribution of the response variable Y.

The histogram presents a normal distribution of the Y variable across the data set so we can infer that majority of the data points fall around the mean and there are few outliers.

A graph of a graph

Description automatically generated

**Figure 1.2: Distribution of variable C1 Figure 1.3: Distribution of variable C2**

A graph of a number of data

Description automatically generated

**Figure 1.4: Distribution of variable C3 Figure 1.5: Distribution of variable C4**

A graph of a number of data

Description automatically generatedA graph of a graph

Description automatically generated

**Figure 1.6: Distribution of variable C5 Figure 1.7: Distribution of variable C6**

A graph showing a distribution of c7 variable

Description automatically generated

**Figure 1.8: Distribution of variable C7**

This variable only had data points of 1 or 0. Even though the variables are untitled, from this we can tell that this variable is a binary as opposed to the other variables. For example, it might represent sex or a yes or no question.

A graph of a box plot

Description automatically generated

**Figure 1.9: Box plot of variables C1 to C6.**

Variable C7 was left out of the box plot due to the nature of the values. This boxplot tells us that all these variables contain outliers. The medians for C1, C4, and C6 are relatively similar, so are the medians for C2,C3. While this cannot directly tell us about the relationship between the variables, we can still gather that they have similar values compared to other variables indicating a closer relationship.

If we look at the histograms above we can clearly see that they all have a relatively similar normal distribution. This suggests that the data set has few outliers and the variables behave in a predictable manner which is important as we want to verify that our dependent variables are able to predict our Y values. It allows for a more accurate analysis and modeling of the data.

**Further EDA**

This section will explore the correlation between the response variable and each dependent variable.

A graph with a line and a dotted line

Description automatically generatedA graph with a line and a dotted line

Description automatically generated

Figure 2.1: Scatter plot of C1 against Y Figure 2.2: Scatter plot of C2 against Y

A graph with a line and a dotted line

Description automatically generatedA graph with a line and dotted line

Description automatically generated

**Figure 2.3: Scatter plot of C3 against Y Figure 2.4: Scatter plot of C4 against Y**

A graph with a line and a dotted line

Description automatically generatedA graph with a line and dotted line

Description automatically generated

**Figure 2.5: Scatter plot of C5 against Y Figure 2.6: Scatter plot of C6 against Y**

At first glance, these scatter plots above seem very similar to which can make it difficult to interpret, however, adding a line of best fit allows us to confirm correlations between variables.

All variables C1 to C3, and C5 to C7 have a postive correlation to the Y variable. C4 has a negative line of best fit indicating a negative correlation between the two.

Frequency Polygons breakdown of each variable against C7

A graph of a graph

Description automatically generatedA graph of a graph

Description automatically generated

**Figure 3.1: Frequency polygon of C1 by C7 Figure 3.2: Frequency polygon of C2 by C7**

A graph of a graph

Description automatically generatedA graph of a graph

Description automatically generated

**Figure 3.3: Frequency polygon of C3 by C7 Figure 3.4: Frequency polygon of C4 by C7**

A graph of a graph

Description automatically generatedA graph with a line graph

Description automatically generated

**Figure 3.5: Frequency polygon of C5 by C7 Figure 3.6: Frequency polygon of C6 by C7**

A graph of a graph

Description automatically generated

**Figure 3.7: Frequency polygon of Y by C7**

The frequency polygon above were plotted for each variable and seperated by the C7 variable with responses either 0 or 1. For all the variables, the value of 1 was more frequent than 0. So, for example, if C7 was by gender and 1 = female. There would be a higher proprtion of women in this dataset than men.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

**Scatter matrix of variables C1 to C6. Correlation Heatmap of variables Y to C6.**

The correlation matrix an dthe heatmap present values indicating the correlation between each of the variables for C1 to C6. Variables C1 and C2 have the highest correlation value of 0.765 whilst variables C4 and C5 have negative correlations. However, it is important to acknowledge that a positive correlation between these variables it not enough to confirm if there truly is a relation. Therefore, further analysis is needed to confirm this. In the correlation matrix we can see that C4 indeed has a negative correlation with response variable Y.

**Statistical Methods**

Hypothesis: There are distinct patterns in the data set which if explored can reveal the relationship between the Y variable and variables C1 to C7.

Variable C7 is a binary variable and could be based on a demographic of gender, for example. Other variables might also represent demographic and behavioural characteristics such as age, height, or weight. The remaining variables represent values like price increasing over a certain period of time or average time spent on an activity.

Y is the response variable, and the predictor variables (C1 to C7) have distributions and correlations that should be understood through exploratory data analysis.   
By evaluating the model's predictive capacity and its capacity to correctly forecast consumer behaviour based on behavioural and demographic traits, we may validate our hypothesis.

The main methods of statistical analysis chosen were the clustering method and linear regression. The clustering method divides unlabelled data points into diverse groups (clusters) where the data points fall into corresponding categories. I plan to use clustering because based on similarities in their characteristics, or variables, clustering can be used to find organic groupings or clusters of observations. By highlighting observations that do not fit into any of the detected groups. It can also be used to find an outliers in the data.

Linear Regression is type of analysis technique used to predict the response variable based on the behaviours and patterns of the dependent variables. I opted to use a linear regression model because it can be an appropriate technique to model a relationship if you gain a solid understanding of our hypothesis about, the link between the response variable (Y) and the predictor variables (C1 to C7).  
The direction and strength of each predictor variable's effect on the respondent variable are represented by easily interpreted coefficients in linear regression. Making decisions that are practical and comprehending the relationship between factors can both benefit from this.

**Results and Conclusions**

* Clearly display the results from your analyses and state what conclusions can be drawn. Discuss the results. Do they correspond to the impression given by the graphical and numerical summaries? Can you draw new graphs that convincingly tell the same story as your statistical analysis? Can the variable Y be predicted accurately from the predictors? [30% of marks]

Using the prediction model in the previous section, new data was generated for each variable C1 to C7 to create predictions for the value of Y.

A screenshot of a computer screen

Description automatically generated

Based on the prediction values above which appear quite random, this would be in line with response variable Y as the value of Y just increased suggesting a more complex relationship between the independent and dependent variables for this data set. However, due to the nature of the data set, the model does seem to accurately predict Y values.

**Linear regression model**

lm\_model <- lm(Y ~ C1 + C2 + C3 + C4 + C4 + C5 + C6 + C7, data = train\_data)

To validate the model, a residuals graph was plotted.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

Figure 4.1: The nature of this plot indicates homoscedasticity. Residuals in this plot show a standard trend with most of the clusters being plotted in the middle. However, they are quite evenly distributed indicating a linear relationship between the response and dependent variables. No clear pattern in the residuals plot indicates that it is suitable as a prediction model.

The plot indicates that our linear regression model is suitable for prediction of the response variables. A linear relationship indicates that there is a proportionate change in the response variable for every change in the predictor variables.  
Therefore, there is confidence in the accuracy of the predictions made for the response variable Y. The model's predictive reliability is increased by the even distribution of data points surrounding the regression line, which indicates that the model's predictions are probably accurate over a wide range of predictor variable values.

A diagram of a number of dots

Description automatically generated

summary(lm\_model)

A screenshot of a computer

Description automatically generated

Image 1: Presents the summary of the prediction model.

If we look at the estimate values, we are able to interpret each of the variables affect on the response variable.

C1 – the estimate value for C1 is 1.282 which suggests that is has a positive correlation with the response variable because the value is above 0.

C2-the estimate value is 5.0702 suggesting that C2 variable has a positive effect on the response variable as it is above 0 by quite a bit.

C3- the estimate value is -0.3082 which indicates a negative correlation between the response variable and C3. This conclusion is reached because the estimate value is below 0.

C4- The estimate value is -1.3116 which again indicates a negative correlation between C4 and the response variable which means that C4 has a negative effect on Y.

C5 – The estimate value is 3.6391 tells us that this variable has a strongly positive correlation with the Y variable.

C6 – The estimate value for C6 is -0.1092 suggesting a negatively correlated relationship between C6 and the response variable Y.

The results from the linear regression method align with previous conclusions drawn about the data set by the initial graphical summaries. The distribution of each variable was normal, however, the scatter plots of each variable showed minimal variation in trend lines. Adding a line to the scatter plots allowed us to see where the positive and negative correlations for each variable happened.

The linear regression model confirmed a negative correlation between Y and C4, however this was not the case for other variables if we compare this to the scatter plots in the general EDA section. For example, figure 2.3 show a positive correlation between the C3 variable and Y. This could be a result of the negative estimate value being closer to 0 compared to the C4 variable.

Even then, the statistical analysis does not support the initial hypothesis as the Y variable continually increases so there is no identifiable pattern between this and the dependent variables. However, the prediction model could still be used to make valid predictions according to the linear regression.

Most of the dependent variables for this data set are above 1 indicating the positive relationship with the response variable. This supports the initial hypothesis for spending and purchase would typically increase as variables such as age or salary increase.

After many attempts at the prediction model, the model including all the variable was found to be the best for predict as shown by the residual plots. From this we can conclude that is suitable to accurately predict Y values.

**Reflection**

**First Clustering with k = 8**

A graph of a number of clusters

Description automatically generatedA graph showing different colored shapes

Description automatically generated

The first clustering conducted used eight clusters because the results using the elbow method indicated that it would be optimal. However, when the cluster plot was created there were large amounts of overlapping groups, specifically in the centre of the graph which could cause confusion when interpreting results. After this revelation, I opted to repeat the cluster plot using K = 6 instead. As a results, the six cluster groups on the plot were clearer and more defined making it easier to understand.

**Second Clustering with K = 6**

A graph of a number of clusters

Description automatically generatedA graph of different colored dots

Description automatically generated

The graphs shown above present clearer clusters which allow for better interpretation. This suggests that using 6 clusters was more appropriate than 8.

Due to the analysis's heavy emphasis on just a handful of variables, significant patterns or relationships may be missed. Using a predictive model can be jeopardised by both overfitting and underfitting, even after the cluster plots have been adjusted. Confounding variables and plausible alternative explanations must be carefully considered when extrapolating causal deductions from observational data. While more sophisticated models may lack transparency, making it difficult to evaluate and trust the results, linear regression models are interpretable.

The main answer/conclusion to this analysis remains partially unanswered because we cannot actually be sure of what the values in the data set represent.

Appendix

title: "MAST5954 Assignment 2"

author: "Anisha Njuki"

date: "2024-03-30"

output: html\_document

```{r setup, include=FALSE}

if (!("knitr" %in% installed.packages())) {

install.packages('knitr', repos='http://cran.rstudio.org')}

library(knitr)

if (!("formatR" %in% installed.packages())) {

install.packages("formatR")}

library(formatR)

{r working directory}

setwd("~/MAST5954")

{r packages and libraries}

if (!("tidyverse" %in% installed.packages())) {

install.packages("tidyverse")}

library(tidyverse)

if (!("dplyr" %in% installed.packages())) {

install.packages("dplyr")}

library(dplyr)

if (!("ggplot2" %in% installed.packages())) {

install.packages("ggplot2")}

library(ggplot2)

if (!("DescTools" %in% installed.packages())) {

install.packages("DescTools")}

library(DescTools)

if (!("reshape2" %in% installed.packages())) {

install.packages("reshape2")}

library(reshape2)

if (!("tidyr" %in% installed.packages())) {

install.packages("tidyr")}

library(tidyr)

if (!("cluster" %in% installed.packages())) {

install.packages("cluster")}

library(cluster)

if (!("gridExtra" %in% installed.packages())) {

install.packages("gridExtra")}

library(gridExtra)

if (!("rpart" %in% installed.packages())) {

install.packages("rpart")}

library(rpart)

if (!("rpart.plot" %in% installed.packages())) {

install.packages("rpart.plot")}

library(rpart.plot)

if (!("corrplot" %in% installed.packages())) {

install.packages("corrplot")}

library(corrplot)

if (!("GGally" %in% installed.packages())) {

install.packages("GGally")}

library(GGally)

if (!("stats" %in% installed.packages())) {

install.packages("stats")}

if (!("factoextra" %in% installed.packages())) {

install.packages("factoextra")}

{r data checking}

data <- read.csv("AsstDat2024cut.csv")

missing\_values <- data %>%

summarise\_all(~sum(is.na(.)))

numerical\_columns <- sapply(data, is.numeric)

# Impute missing values for numerical variables with mean

data[, numerical\_columns] <- lapply(data[, numerical\_columns], function(x) {

ifelse(is.na(x), mean(x, na.rm = TRUE), x)

})

data$C7 <- ifelse(data$C7 != 0 & data$C7 != 1, 1, data$C7)

data$C7 <- ifelse(data$C7 == 0 | data$C7 == 1, data$C7, 0)

summary(data)

Y <- data$Y

C1<- data$C1

C2<-data$C2

C3<-data$C3

C4<-data$C4

C5<-data$C5

C6<-data$C6

C7<-data$C7

```{r General EDA}

# general plots to present the distribution of each variable in the data set.

ggplot(data, aes(x = Y)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of y")

ggplot(data, aes(x = C1)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of C1")

ggplot(data, aes(x = C2)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of C2")

ggplot(data, aes(x = C3)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of C3")

ggplot(data, aes(x = C4)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of C4")

ggplot(data, aes(x = C5)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of C5")

ggplot(data, aes(x = C6)) +

geom\_histogram(fill = 'blue', binwidth = 5) +

labs(title = "Distribution of C6")

# Ensure missing value entries do not affect the binary variable

data$C7 <- ifelse(data$C7 != 0 & data$C7 != 1, 1, data$C7)

data$C7 <- ifelse(data$C7 == 0 | data$C7 == 1, data$C7, 0)

binary\_counts <- table(data$C7)

# Create bar chart

barplot(binary\_counts,

names.arg = names(binary\_counts),

xlab = "C7",

ylab = "Frequency",

main = "Distribution of C7 Variable")

# Create box plot

selected\_data <- c("C1","C2","C3","C4","C5","C6")

subset\_data <- data[, c("C1","C2","C3","C4","C5","C6")]

boxplot(subset\_data, main = "Boxplot")

{r EDA}

#Create scatter plots to compare each of the variables to the response variable

ggplot(data, aes(x = C1, y = Y, )) +

geom\_point() +

geom\_smooth(aes(x = C1, y = Y), method = "lm") +

labs(x = "C1", y = "Y", title = "Scatter Plot with Smooth Line")

ggplot(data, aes(x = C2, y = Y, )) +

geom\_point() +

geom\_smooth(aes(x = C2, y = Y), method = "lm") +

labs(x = "C2", y = "Y", title = "Scatter Plot with Smooth Line")

ggplot(data, aes(x = C3, y = Y, )) +

geom\_point() +

geom\_smooth(aes(x = C3, y = Y), method = "lm") +

labs(x = "C3", y = "Y", title = "Scatter Plot with Smooth Line")

ggplot(data, aes(x = C4, y = Y, )) +

geom\_point() +

geom\_smooth(aes(x = C4, y = Y), method = "lm") +

labs(x = "C4", y = "Y", title = "Scatter Plot with Smooth Line")

ggplot(data, aes(x = C5, y = Y, )) +

geom\_point() +

geom\_smooth(aes(x = C5, y = Y), method = "lm") +

labs(x = "C5", y = "Y", title = "Scatter Plot with Smooth Line")

ggplot(data, aes(x = C6, y = Y, )) +

geom\_point() +

geom\_smooth(aes(x = C6, y = Y), method = "lm") +

labs(x = "C6", y = "Y", title = "Scatter Plot with Smooth Line")

# Calculate averages for all the variables

mean(Y)

mean(C1)

mean(C2)

mean(C3)

mean(C4)

mean(C5)

mean(C6)

# Analysis comparison to binary C7

data$C7\_group <- factor(ifelse(data$C7 == 0, "0", "1"))

ggplot(data, aes(x = C1, color = C7\_group)) +

geom\_freqpoly(binwidth = 10) +

labs(title = "Frequency Polygon of C1 by C7 binary",

x = "C1",

y = "Frequency")

ggplot(data, aes(x = C2, color = C7\_group)) +

geom\_freqpoly(binwidth = 10) +

labs(title = "Frequency Polygon of C2 by C7 binary",

x = "C2",

y = "Frequency")

ggplot(data, aes(x = C3, color = C7\_group)) +

geom\_freqpoly(binwidth = 10) +

labs(title = "Frequency Polygon of C3 by C7 binary",

x = "C3",

y = "Frequency")

ggplot(data, aes(x = C4, color = C7\_group)) +

geom\_freqpoly(binwidth = 10) +

labs(title = "Frequency Polygon of C4 by C7 binary",

x = "C4",

y = "Frequency")

ggplot(data, aes(x = C5, color = C7\_group)) +

geom\_freqpoly(binwidth = 10) +

labs(title = "Frequency Polygon of C5 by C7 binary",

x = "C5",

y = "Frequency")

ggplot(data, aes(x = C6, color = C7\_group)) +

geom\_freqpoly(binwidth = 10) +

labs(title = "Frequency Polygon of C6 by C7 binary",

x = "C6",

y = "Frequency")

ggplot(data, aes(x = Y, color = C7\_group)) +

geom\_freqpoly(binwidth = 30) +

labs(title = "Frequency Polygon of Y by C7 binary",

x = "Y",

y = "Frequency")

# Correlation analysis

num\_vars\_subset <- c("C1","C2","C3","C4","C5","C6")

pairs(data[, num\_vars\_subset])

num\_vars <- c("C1","C2","C3","C4","C5","C6")

scatter\_matrix <- ggpairs(data, columns = num\_vars)

print(scatter\_matrix)

# Select the relevant variables

selected\_vars <- data[, c("Y","C1","C2","C3","C4","C5","C6")]

# Calculate the correlation matrix

correlation\_matrix <- cor(selected\_vars, method = "pearson")

ggplot(data = reshape2::melt(correlation\_matrix), aes(Var1, Var2, fill = value)) +

geom\_tile(color = "white") +

scale\_fill\_gradient2(low = "purple", mid = "white", high = "pink", midpoint = 0) +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

labs(title = "Correlation Heatmap",

x = "Variables",

y = "Variables")

```

# Table of average value for each variable

Variable | Average

------------- | -------------

Y | 854.4126

C1 | 32.99384

C2 | 103.4392

C3 | 115.8035

C4 | 22.80039

C5 | -0.502065

C6 | 11.35267

```{r clustering}

# Load additional libraries

library(cluster)

library(factoextra)

library(stats)

# Clustering for variable C1 to check

# Select relevant variables for clustering (adjust as needed)

clustering\_data <- data[, c("C1", "Y")]

# Standardize the data

scaled\_data <- scale(clustering\_data)

# Determine the optimal number of clusters using, for example, the elbow method

wss <- numeric(10)

for (i in 1:10) {

kmeans\_model <- kmeans(scaled\_data, centers = i, nstart = 10)

wss[i] <- sum(kmeans\_model$tot.withinss)

}

plot(1:10, wss, type = "b", main = "Elbow Method", xlab = "Number of Clusters", ylab = "Within Sum of Squares")

# Based on the elbow method, choose the optimal number of clusters

# Replace 'optimal\_clusters' with the selected number

optimal\_clusters <- 8

# Perform k-means clustering

kmeans\_model <- kmeans(scaled\_data, centers = optimal\_clusters, nstart = 10)

# Visualize the clustering results

fviz\_cluster(kmeans\_model, data = scaled\_data, geom = "point")

# Perform k-means clustering with the optimal number of clusters

final <- kmeans(scaled\_data, optimal\_clusters, nstart = 25)

# Visualize the clustering results

fviz\_cluster(final, data = scaled\_data)

# Add clustering information to the original data

data\_with\_clusters <- data %>%

mutate(Cluster = final$cluster)

# Select relevant variables for summarisation

variables\_of\_interest <- c("C1", "Y")

# Summarize selected variables by cluster

summary\_by\_cluster <- data\_with\_clusters %>%

select(Cluster, all\_of(variables\_of\_interest)) %>%

group\_by(Cluster) %>%

summarise\_all("mean")

# Print summary

print(summary\_by\_cluster)

# Clustering for all variables

# Select relevant variables for clustering (adjust as needed)

clustering\_data2 <- data[, c("Y","C1","C2","C3","C4","C5","C6")]

# Standardize the data

scaled\_data <- scale(clustering\_data2)

# Determine the optimal number of clusters using, for example, the elbow method

wss <- numeric(10)

for (i in 1:10) {

kmeans\_model <- kmeans(scaled\_data, centers = i, nstart = 10)

wss[i] <- sum(kmeans\_model$tot.withinss)

}

plot(1:10, wss, type = "b", main = "Elbow Method", xlab = "Number of Clusters", ylab = "Within Sum of Squares")

# Based on the elbow method, choose the optimal number of clusters

# Replace 'optimal\_clusters' with the selected number

optimal\_clusters <- 8

# Perform k-means clustering

kmeans\_model <- kmeans(scaled\_data, centers = optimal\_clusters, nstart = 10)

# Visualize the clustering results

fviz\_cluster(kmeans\_model, data = scaled\_data, geom = "point")

# Perform k-means clustering with the optimal number of clusters

final <- kmeans(scaled\_data, optimal\_clusters, nstart = 25)

# Visualize the clustering results

fviz\_cluster(final, data = scaled\_data)

# Add clustering information to the original data

data\_with\_clusters <- data %>%

mutate(Cluster = final$cluster)

# Select relevant variables for summarisation

variables\_of\_interest <- c("Y","C1","C2","C3","C4","C5","C6")

# Summarize selected variables by cluster

summary\_by\_cluster <- data\_with\_clusters %>%

select(Cluster, all\_of(variables\_of\_interest)) %>%

group\_by(Cluster) %>%

summarise\_all("mean")

# Print summary

print(summary\_by\_cluster)

# Clustering 2, refined after first clustering

# Select relevant variables for clustering (adjust as needed)

clustering\_data3 <- data[, c("Y","C1","C2","C3","C4","C5","C6")]

# Standardize the data

scaled\_data <- scale(clustering\_data3)

# Determine the optimal number of clusters using, for example, the elbow method

wss <- numeric(10)

for (i in 1:10) {

kmeans\_model <- kmeans(scaled\_data, centers = i, nstart = 10)

wss[i] <- sum(kmeans\_model$tot.withinss)

}

plot(1:10, wss, type = "b", main = "Elbow Method", xlab = "Number of Clusters", ylab = "Within Sum of Squares")

# Based on the elbow method, choose the optimal number of clusters

# Replace 'optimal\_clusters' with the selected number

optimal\_clusters <- 6

# Perform k-means clustering

kmeans\_model <- kmeans(scaled\_data, centers = optimal\_clusters, nstart = 10)

# Visualize the clustering results

fviz\_cluster(kmeans\_model, data = scaled\_data, geom = "point")

# Perform k-means clustering with the optimal number of clusters

final <- kmeans(scaled\_data, optimal\_clusters, nstart = 25)

# Visualize the clustering results

fviz\_cluster(final, data = scaled\_data)

# Add clustering information to the original data

data\_with\_clusters <- data %>%

mutate(Cluster = final$cluster)

# Select relevant variables for summarisation

variables\_of\_interest <- c("Y","C1","C2","C3","C4","C5","C6")

# Summarize selected variables by cluster

summary\_by\_cluster <- data\_with\_clusters %>%

select(Cluster, all\_of(variables\_of\_interest)) %>%

group\_by(Cluster) %>%

summarise\_all("mean")

# Print summary

print(summary\_by\_cluster)

# This print is much cleaner after being refined.

{r dimnsionality reduction}

# Group clustering 1

# Choose the number of components (e.g., number of principal components)

num\_components <- 7

# Perform PCA

pca\_model <- prcomp(scaled\_data, center = TRUE, scale. = TRUE)

# Extract the principal components

principal\_components <- pca\_model$x[, 1:num\_components]

# Visualize the clustered data in the reduced space

plot(principal\_components, col = "blue",

main = "Clustered Data in Reduced Space (PCA)", xlab = "PC1", ylab = "PC2")

{r modelling}

# Split the data into training and testing sets

set.seed(123) # For reproducibility

train\_indices <- sample(nrow(data), 0.7 \* nrow(data)) # 70% train, 30% test

train\_data <- data[train\_indices, ]

test\_data <- data[-train\_indices, ]

# Fit a linear regression model

lm\_model <- lm(Y ~ C1 + C2 + C3 + C4 + C5 + C6 + C7, data = train\_data)

# Make predictions on the test set

predictions <- predict(lm\_model, newdata = test\_data)

# Evaluate the model

mse <- mean((test\_data$response - predictions)^2)

r\_squared <- summary(lm\_model)$r.squared

par(mfrow = c(1, 2))

plot(lm\_model, which = c(1, 2))

plot(lm\_model, which = 3)

new\_data1 <- data.frame(

C1 = rnorm(100), # Generate 100 random values for C1

C2 = rnorm(100), # Generate 100 random values for C2

C3 = rnorm(100), # Generate 100 random values for C3

C4 = rnorm(100), # Generate 100 random values for C4

C5 = rnorm(100), # Generate 100 random values for C5

C6 = rnorm(100), # Generate 100 random values for C6

C7 = rnorm(100) # Generate 100 random values for C7

)

# Predict the response variable Y using the linear regression model

predictions <- predict(lm\_model, newdata = new\_data1)

# View the predicted values

print(predictions)

mean(predictions)

# Model predictions for different combinations

# Split the data into training and testing sets

set.seed(123) # For reproducibility

train\_indices1 <- sample(nrow(data), 0.7 \* nrow(data)) # 70% train, 30% test

train\_data1 <- data[train\_indices1, ]

test\_data1 <- data[-train\_indices1, ]

# Fit a linear regression model

lm\_model1 <- lm(Y ~ C7, data = train\_data1)

# Make predictions on the test set

predictions <- predict(lm\_model1, newdata = test\_data1)

# Evaluate the model

mse <- mean((test\_data1$response - predictions)^2)

r\_squared <- summary(lm\_model1)$r.squared

par(mfrow = c(1, 2))

plot(lm\_model1, which = c(1, 2))

plot(lm\_model1, which = 3)

summary(lm\_model)