MT5758 – Assignment 3

Project Report

230012908

Group 1

University of St Andrews

2024

Introduction

Social media's influence on marketing strategies is a dynamic field and understanding these dynamics can offer valuable insights for effective content generation and posting. Platforms like Facebook have become essential for brands to reach and engage with their audience. The Facebook Metrics dataset offers a detailed look into how users interact with the brand posts, providing valuable information for marketers. Understanding user behaviour within this dataset is key to crafting effective marketing strategies.

The primary motivation behind this project is to understand how various factors influence user engagement with posts from a well-known cosmetic brand on Facebook and to make predictions to optimise future content strategy, improve reach, and enhance overall user interaction with the cosmetics brand Facebook page.

The main objective for the analysis is to explore and understand the impact of the cosmetics brand posts on the audience engagement through the analysis of the Facebook Metrics dataset. To achieve this goal, it is necessary to address the following aims:

1. Prepare the Facebook Metrics data for analysis which includes data exploration and data cleaning (including feature selection).

2. Perform an exploratory analysis of the interactions with the cosmetics' Facebook page which involves the following steps:

a) Conducting Principal Component Analysis (PCA) – a dimensionality reduction method – to simplify the analysis while preserving key information in the data.

b) Conducting Clustering Analysis (K-Means and Hierarchical clustering) – to see if there are any groups which could help identify different engagement patterns within social media users.

We propose the following two solutions to solve the problem. First, PCA could be applied to the pre-processed dataset to reduce dimensionality of the data and identify the number of principal components that explain the most variance while retaining most of the information. These components may represent patterns within the data making it easier to understand the underlying structure. We will attempt interpreting the principal components to understand the relationships between variables and identify which factors contribute most to user engagement with the cosmetic brand's Facebook page. This method would make the Facebook metrics data more suitable for further analysis.

Second, we assume that the proposed solution in terms of the K-means and hierarchical clustering analyses could involve using techniques like the elbow method and plotting silhouette to determine the optimal number of clusters for K-means algorithm. For hierarchical clustering, we attempt visualising dendrograms to understand cluster relationships. We expect each cluster to represent a segment -with similar engagement patterns.

Pre-Processing

Data Exploration

The Data Exploration section is split into two parts: Data Exploration where we look at the Facebook Metrics data in details, checking number of observations (500), number of variables (19) and examining data types – 1 binary, 5 categories, and the rest are integers. The second part of the section – Further Data Exploration - is performed after Data Cleaning section and includes creating scatterplots, inspecting correlations (see Figure 1), plotting parallel coordinates (see Figure 2), stars and Chernoff faces. Next, we look into Euclidean and Manhattan distances.

From the Facebook Data Correlogram (see Figure 1) we can see that almost all the variables are highly correlated though there are still some variables with 0 correlation. This means that we can still try applying Principal Component Analysis to this data.

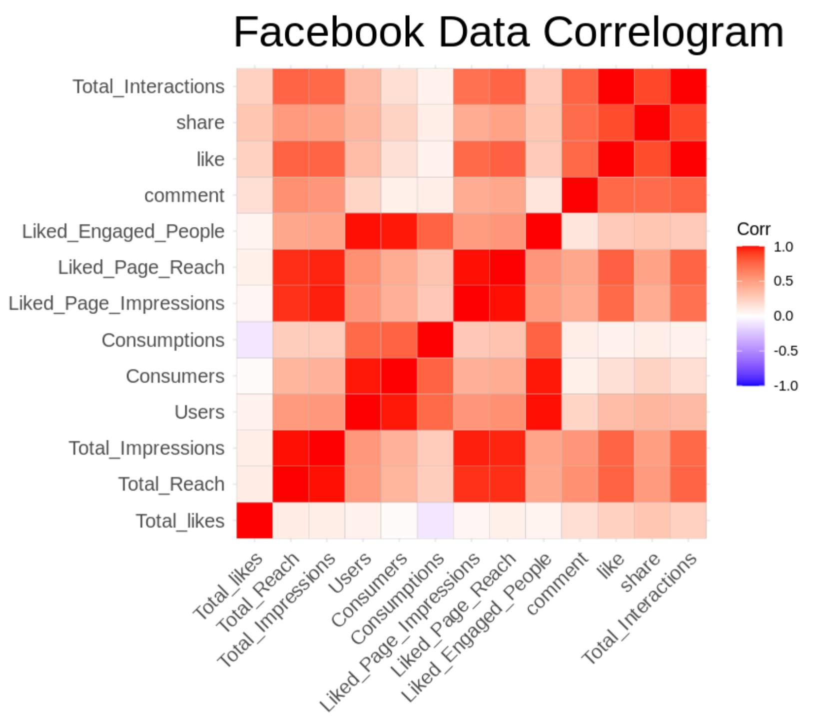


Figure 1. Facebook Data Correlogram

Next step taken is plotting Parallel Coordinate Plot (see Figure 2) which shows that there are some outliers in the data, the scales of which differs from the lines that have distinct patterns. Later in the analysis we scale the data as scaling ensures that all features contribute equally to the analysis by bringing them to a similar scale, which is crucial for the accurate performance of many algorithms including PCA and Clustering analysis.

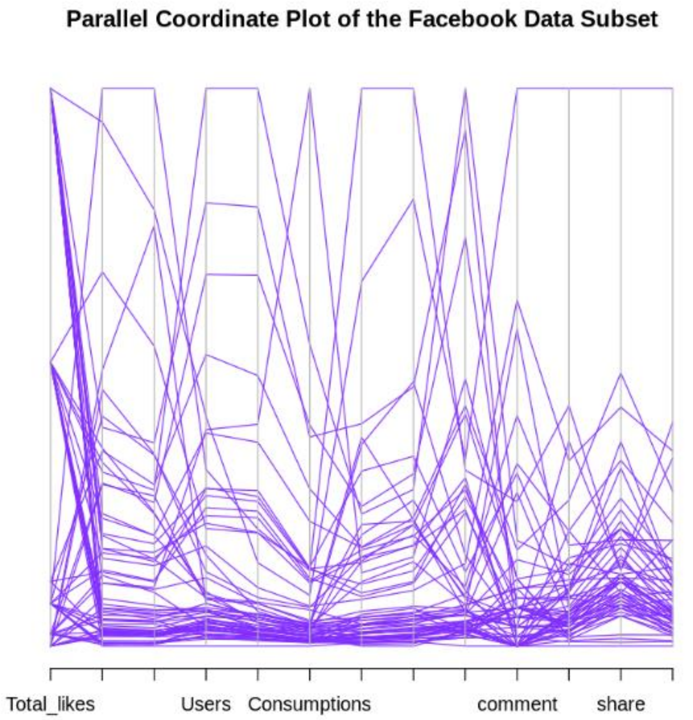


Figure 2.Parallel Coordinate Plot of the Facebook Data Subset

Data Cleaning and Feature Selection

In the Data Cleaning section, first thing we do is checking the Facebook data for missing values and deal with them by eliminating rows with missing values from the following columns "Paid", "like", and "share". To conduct PCA and Clustering analyses we should keep in mind that all data types in the dataset must be numeric. For this reason, we drop categorical and binary variables from the Facebook data, namely, "Type", "Category", "Monthly", "Weekday", "Paid", and "Hour" and then we convert all remaining variables to numeric data type. Overall, Facebook dataset contains 500 observations which is quite a big amount for PCA and clustering analysis that is why we get a small subset of data (1:50) from the transformed Facebook_numeric dataset for more convenient visualisation. In addition, some variable names turned out to be too long to fit the plots nicely and to fix this we change the names of variables to shorter versions.

Analysis

Principal Component Analysis (PCA)

We start the Principal Component Analysis by applying PCA to the data_subset using a built-in function and then plotting Biplots to explore the variability explained. After that is done, we proceed to creating Elbow plot/Scree plot (see Figure 3) that help identify the optimal number of clusters in the dataset by plotting the within-cluster sum of squares against the number of clusters and observing the point where the rate of decrease sharply changes, resembling an "elbow". Based on the Scree Plot (see Figure 3) we can see that it suggests that 2 PCs are enough to capture the most important information in the Facebook Metrics data.

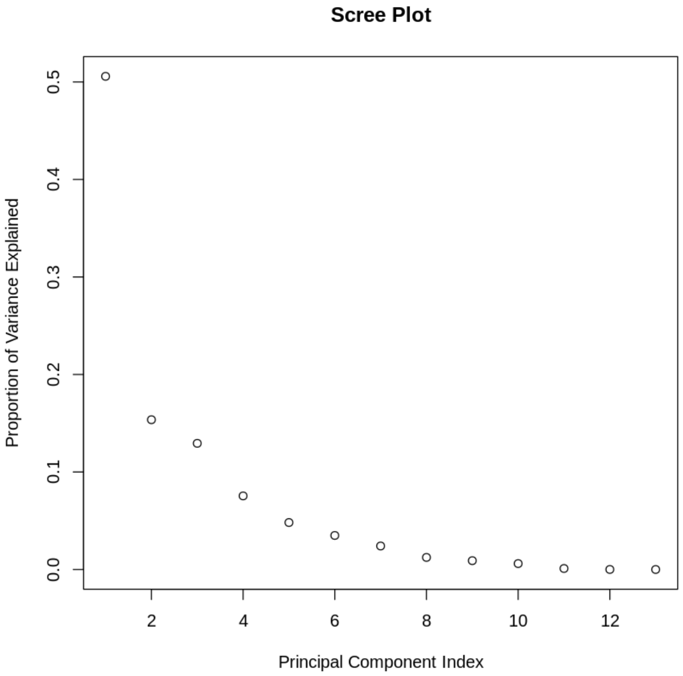


Figure 3.Scree Plot/Elbow Plot

Finally, we create PCA Biplot. The direction of each variable vector represents the direction of highest variance for that variable in the original feature space. The length of the vector indicates the magnitude of the variable's contribution to the principal component.

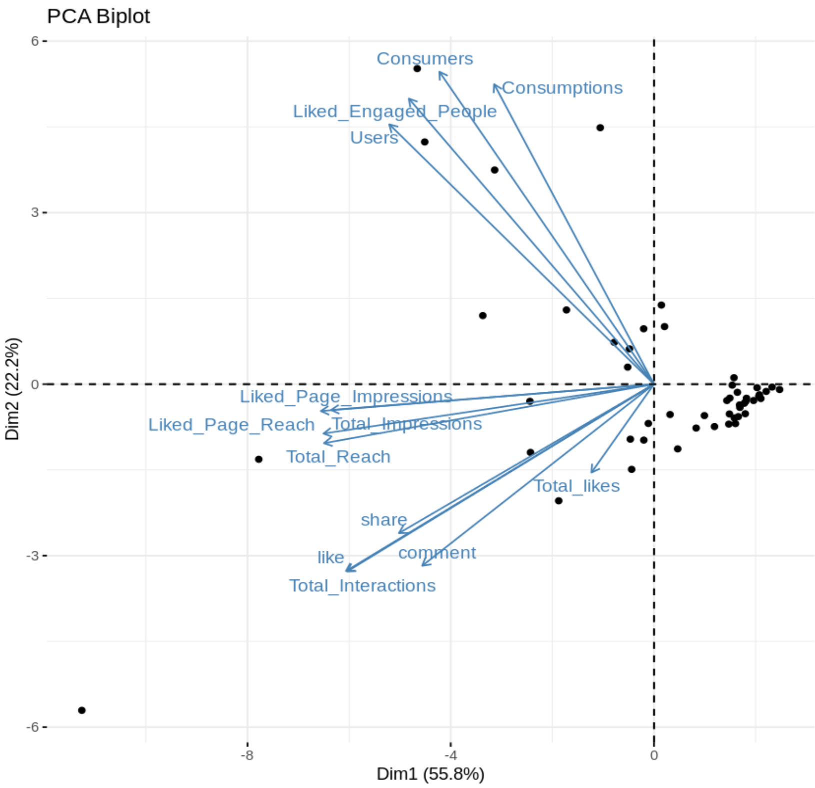


Figure 4.PCA Biplot

For example, Total likes has the smallest contribution to the PC2 while Consumers contribute greatly to the PC2. The angle between vectors reflects the correlation between variables. Small angles indicate high positive correlation, while large angles suggest low or negative correlation. For instance, from PCA Biplot (see Figure 4) we can observe that first principal component is negatively correlated with Like Page Impressions and Liked Page Reach, while the second principal component is positively correlated with Consumers and Consumptions.

Clustering Analysis

For clustering analysis, we use both K-means and hierarchical clustering algorithms to observe grouping patterns in the Facebook page engagement. K-means clustering algorithm partitions the data into distinct clusters, while hierarchical clustering creates a tree-like structure of clusters based on similarity between different interaction patterns. We use both approaches in this analysis since relying on a single clustering algorithm may not capture all the nuances and structures present in the data. By using both K-means and hierarchical clustering, we can cross-validate the results and ensure robustness in the analysis. If both algorithms produce similar results, it gives more confidence in the clustering outcomes.

K-Means Clustering

To begin with, we apply K-means clustering algorithm to the data_subset we derived earlier in the data cleaning section and assign K-means with 2, 3 and 4 centres which are then used to create three scatterplot matrices coloured by clusters (see in Appendix). After that, we use a criteria function to select and visualise the number of clusters which conveyed that K=2 is the optimal number of clusters (see Figure 5).

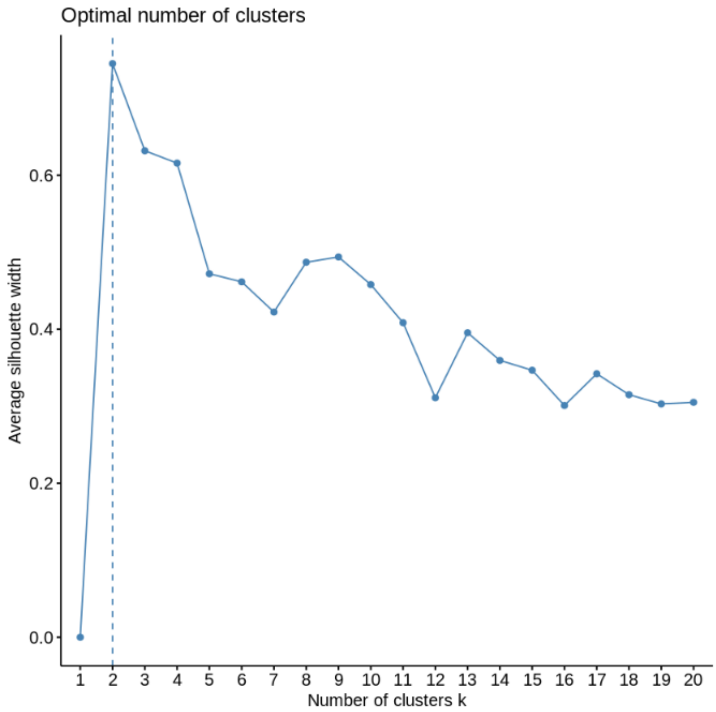


Figure 5.Optimal number of clusters – K-means clustering

Furthermore, we create three silhouette plots to visualise 2, 3 and 4 clusters of the Facebook data (see Appendix). Silhouette plots are used to assess the quality of clustering by measuring how similar an object is to its own cluster compared to other clusters. Initially, the analysis considered dividing the data into two clusters, as suggested by the Optimal Number of Clusters plot (Figure 5). However, upon examining the silhouette plot for two clusters, it was observed that cluster 1 significantly dominates over cluster 2. This suggests that the data might not be well-separated into just two clusters. Additionally, the silhouette coefficient for some data points in cluster 2 drops below zero, indicating that these points might be better assigned to a different cluster. This implies that the silhouette plot with 2 clusters is not suitable for the analysis, as it does not provide clear and distinct clusters. The analysis then considered dividing the data into three clusters. Upon examining the silhouette plot

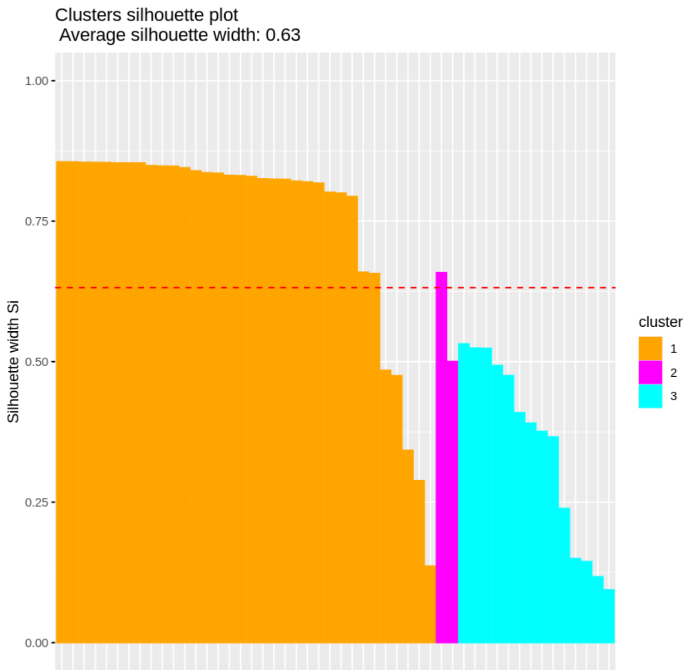


Figure 6.Clusters silhouette plot – 3 clusters

for three clusters, it was observed that none of the silhouette coefficients drop below zero, indicating a better separation and cohesion among the clusters compared to the plot with two clusters. The silhouette plot with 3 clusters appears to be more proportional and relevant in terms of grouping the data, suggesting that dividing the data into three clusters provides a better representation of the underlying structure in the Facebook data.

Hierarchical Clustering

Hierarchical clustering is another commonly used method for grouping data into clusters. Unlike K-means clustering, hierarchical clustering does not require the number of clusters to be specified beforehand. Instead, it creates a tree-like structure known as a dendrogram that illustrates the arrangement of clusters at different levels of similarity. Hierarchical clustering analysis also suggests that it is optimal to choose K=2 clusters (see Figure 7), similarly to what we found for K-means clustering (see Figure 5). This would strengthen our confidence in the identified clustering solution, demonstrating consistency and reliability of our findings. However, the Clusters silhouette plot (see Figure 6) resonates with both of the Optimal number of clusters plots (see Figures 5 and 8) which leaves us with a solid ground for further research and analysis.

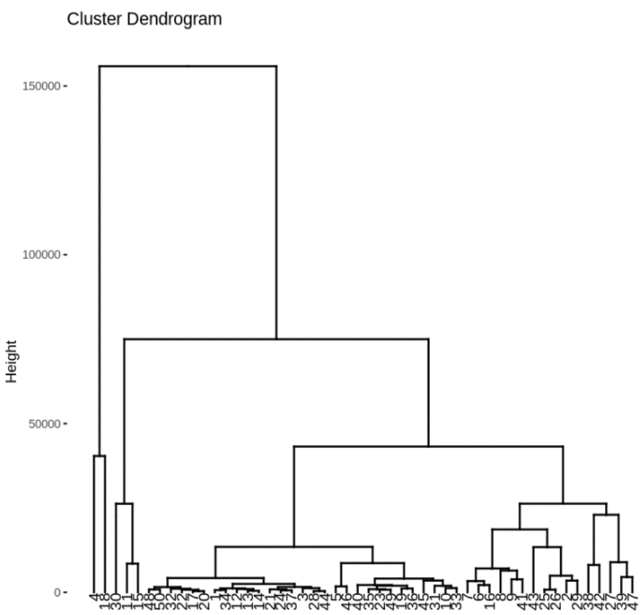


Figure 7.Cluster Dendrogram, method = 'complete'

Conclusion

To sum it all up, the Exploratory Facebook Metrics Data Analysis of user engagement with a cosmetic brand's Facebook page has highlighted key variables such as total likes, like page impressions, liked page reach, consumers, and consumptions, which significantly influence engagement metrics. However, the clustering analysis presents an interesting dilemma: while the elbow plot indicates that 2 clusters may be optimal, the silhouette plot for K-means suggests that 3 clusters could offer a better representation of the data's underlying structure. This confusing result underscores the complexity of user behaviour on social media platforms and the importance of employing multiple analytical techniques for a comprehensive understanding. Moving forward, further exploration and refinement of clustering algorithms, com with a deeper dive into the variables identified through PCA, will be crucial in fine-tuning marketing strategies

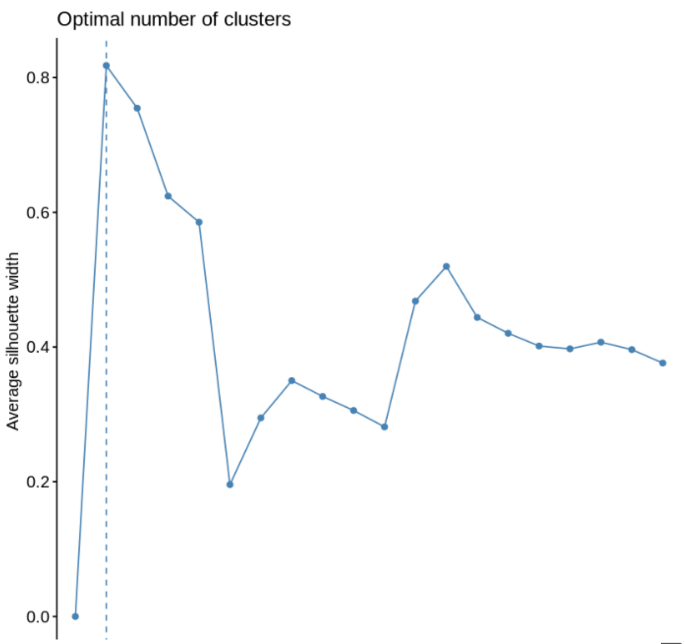


Figure 8.Optimal number of clusters – Hierarchical clustering

to better resonate with the diverse audience segments identified in our analysis.

#

Exploratory Clustering Analysis of Facebook Metrics Data

Table of Contents

Pre-Processing

Data Exploration

- Checking number of observations
- Checking data types

Data Cleaning

- Dealing with Missing Values
- Dropping categorical and binary variables from the Facebook data
- Converting all data types to numeric
- Getting a small subset from the Facebook dataset
- Changing the names of variables to shorter ones

Further Data Exploration

- Scatterplots
- Correlations
- Parallel coordinates, stars and faces
- Distance and Similarity:

1. Euclidean Distance
2. Manhattan Distance

Analysis

PCA

- Applying PCA
- Biplots
- Covariance Matrix, Eigenvalues and Eigenvectors
- PC1 and PC2 Scatterplot
- Elbow Plot & Cumulative Proportion of Variance Explained
- Final PCA plot

Clustering Analysis

- K-Means Clustering Analysis
- Hierarchical Clustering Analysis

```
# Reading in the Facebook data
Facebook <- read.table('/content/dataset_Facebook.csv', sep=";",
header=TRUE)
head(Facebook)

  Page.total.likes Type   Category Post.Month Post.Weekday Post.Hour
Paid
1 139441           Photo  2        12         4                3         0

2 139441           Status 2        12         3               10         0

3 139441           Photo  3        12         3                3         0

4 139441           Photo  2        12         2               10         1
```

```
5 139441           Photo  2        12         2             3        0

6 139441           Status 2        12         1             9        0

  Lifetime.Post.Total.Reach Lifetime.Post.Total.Impressions
1  2752                      5091
2 10460                     19057
3  2413                      4373
4 50128                     87991
5  7244                     13594
6 10472                     20849
  Lifetime.Engaged.Users Lifetime.Post.Consumers
Lifetime.Post.Consumptions
1  178                    109                        159

2 1457                   1361                       1674

3  177                    113                        154

4 2211                    790                       1119

5  671                    410                        580

6 1191                   1073                       1389

  Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
1  3078
2 11710
3  2812
4 61027
5  6228
6 16034
  Lifetime.Post.reach.by.people.who.like.your.Page
1  1640
2  6112
3  1503
4 32048
5  3200
6  7852
  Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
comment
1  119
4
2 1108
5
3  132
0
4 1386
58
```

```
5  396
19
6 1016
1
  like share Total.Interactions
1   79  17    100
2  130  29    164
3   66  14     80
4 1572 147   1777
5  325  49    393
6  152  33    186
```

```r
# Installing necessary packages
install.packages('corrplot')
install.packages('ggcorrplot')
install.packages('factoextra')
install.packages('plotly')
install.packages('ape')
install.packages('usmap')
install.packages('silhouette')
install.packages("GGally")
install.packages("MASS")
install.packages('aplpack')
install.packages('usmap')
install.packages("scatterplot3d")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Warning message:
“package ‘silhouette’ is not available for this version of R
```

```
A version of this package for your version of R might be available
elsewhere,
see the ideas at
https://cran.r-project.org/doc/manuals/r-patched/R-
admin.html#Installing-packages”
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
library(corrplot)
library(ggcorrplot)
library(factoextra)
library(plotly)
library(ape)
library(usmap)
library(dplyr)
library(GGally)
library(MASS)
library(aplpack)
library(cluster)
library(scatterplot3d)
```

```
corrplot 0.92 loaded

Loading required package: ggplot2

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


Attaching package: ‘plotly’


The following object is masked from ‘package:ggplot2’:

    last_plot


The following object is masked from ‘package:stats’:
```

```
    filter


The following object is masked from ‘package:graphics’:

    layout



Attaching package: ‘dplyr’


The following object is masked from ‘package:ape’:

    where


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2


Attaching package: ‘MASS’


The following object is masked from ‘package:dplyr’:

    select


The following object is masked from ‘package:plotly’:

    select


Warning message:
“no DISPLAY variable so Tk is not available”
```

Pre-Processing

Data Exploration

Checking number of observations and Checking data types

```
# Data Exploration
str(Facebook)
summary(Facebook)
head(Facebook)
```

```
'data.frame':	500 obs. of  19 variables:
 $
Page.total.likes                                                            :
int  139441 139441 139441 139441 139441 139441 139441 139441 139441
139441 ...
 $
Type                                                                        :
chr  "Photo" "Status" "Photo" "Photo" ...
 $
Category                                                                    :
int  2 2 3 2 2 2 3 3 2 3 ...
 $
Post.Month                                                                  :
int  12 12 12 12 12 12 12 12 12 12 ...
 $
Post.Weekday                                                                :
int  4 3 3 2 2 1 1 7 7 6 ...
 $
Post.Hour                                                                   :
int  3 10 3 10 3 9 3 9 3 10 ...
 $
Paid                                                                        :
int  0 0 0 1 0 0 1 1 0 0 ...
 $
Lifetime.Post.Total.Reach                                                   :
int  2752 10460 2413 50128 7244 10472 11692 13720 11844 4694 ...
 $
Lifetime.Post.Total.Impressions                                             :
int  5091 19057 4373 87991 13594 20849 19479 24137 22538 8668 ...
 $
Lifetime.Engaged.Users                                                      :
int  178 1457 177 2211 671 1191 481 537 1530 280 ...
 $
Lifetime.Post.Consumers                                                     :
int  109 1361 113 790 410 1073 265 232 1407 183 ...
 $
```

```
Lifetime.Post.Consumptions                                          :
int  159 1674 154 1119 580 1389 364 305 1692 250 ...
 $
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page       :
int  3078 11710 2812 61027 6228 16034 15432 19728 15220 4309 ...
 $
Lifetime.Post.reach.by.people.who.like.your.Page                    :
int  1640 6112 1503 32048 3200 7852 9328 11056 7912 2324 ...
 $
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post:
int  119 1108 132 1386 396 1016 379 422 1250 199 ...
 $
comment                                                             :
int  4 5 0 58 19 1 3 0 0 3 ...
 $
like                                                                :
int  79 130 66 1572 325 152 249 325 161 113 ...
 $
share                                                               :
int  17 29 14 147 49 33 27 14 31 26 ...
 $
Total.Interactions                                                  :
int  100 164 80 1777 393 186 279 339 192 142 ...

 Page.total.likes     Type              Category      Post.Month
 Min.   : 81370   Length:500         Min.   :1.00   Min.   : 1.000
 1st Qu.:112676   Class :character   1st Qu.:1.00   1st Qu.: 4.000
 Median :129600   Mode  :character   Median :2.00   Median : 7.000
 Mean   :123194                      Mean   :1.88   Mean   : 7.038
 3rd Qu.:136393                      3rd Qu.:3.00   3rd Qu.:10.000
 Max.   :139441                      Max.   :3.00   Max.   :12.000

  Post.Weekday    Post.Hour          Paid
Lifetime.Post.Total.Reach
 Min.   :1.00   Min.   : 1.00   Min.   :0.0000   Min.   :   238

 1st Qu.:2.00   1st Qu.: 3.00   1st Qu.:0.0000   1st Qu.:  3315

 Median :4.00   Median : 9.00   Median :0.0000   Median :  5281

 Mean   :4.15   Mean   : 7.84   Mean   :0.2786   Mean   : 13903

 3rd Qu.:6.00   3rd Qu.:11.00   3rd Qu.:1.0000   3rd Qu.: 13168

 Max.   :7.00   Max.   :23.00   Max.   :1.0000   Max.   :180480

                                NA's   :1

 Lifetime.Post.Total.Impressions Lifetime.Engaged.Users
Lifetime.Post.Consumers
```

```
 Min.   :    570                  Min.   :    9.0        Min.   :
9.0
 1st Qu.:   5695                  1st Qu.:  393.8        1st Qu.:
332.5
 Median :   9051                  Median :  625.5        Median :
551.5
 Mean   :  29586                  Mean   :  920.3        Mean   :
798.8
 3rd Qu.:  22086                  3rd Qu.: 1062.0        3rd Qu.:
955.5
 Max.   :1110282                  Max.   :11452.0
Max.   :11328.0


 Lifetime.Post.Consumptions
 Min.   :    9.0
 1st Qu.:  509.2
 Median :  851.0
 Mean   : 1415.1
 3rd Qu.: 1463.0
 Max.   :19779.0

 Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
 Min.   :    567
 1st Qu.:   3970
 Median :   6256
 Mean   :  16766
 3rd Qu.:  14860
 Max.   :1107833

 Lifetime.Post.reach.by.people.who.like.your.Page
 Min.   :  236
 1st Qu.: 2182
 Median : 3417
 Mean   : 6585
 3rd Qu.: 7989
 Max.   :51456

 Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
 Min.   :   9.0
 1st Qu.: 291.0
 Median : 412.0
 Mean   : 610.0
 3rd Qu.: 656.2
 Max.   :4376.0

    comment              like             share
Total.Interactions
 Min.   :  0.000   Min.   :   0.0   Min.   :  0.00   Min.   :   0.0
```

```
 1st Qu.:  1.000   1st Qu.:  56.5   1st Qu.: 10.00   1st Qu.:  71.0
 Median :  3.000   Median : 101.0   Median : 19.00   Median : 123.5
 Mean   :  7.482   Mean   : 177.9   Mean   : 27.27   Mean   : 212.1
 3rd Qu.:  7.000   3rd Qu.: 187.5   3rd Qu.: 32.25   3rd Qu.: 228.5
 Max.   :372.000   Max.   :5172.0   Max.   :790.00   Max.   :6334.0
                   NA's   :1        NA's   :4
  Page.total.likes Type   Category Post.Month Post.Weekday Post.Hour
Paid
1 139441           Photo  2        12         4             3        0
2 139441           Status 2        12         3            10        0
3 139441           Photo  3        12         3             3        0
4 139441           Photo  2        12         2            10        1
5 139441           Photo  2        12         2             3        0
6 139441           Status 2        12         1             9        0
  Lifetime.Post.Total.Reach Lifetime.Post.Total.Impressions
1  2752                      5091
2 10460                     19057
3  2413                      4373
4 50128                     87991
5  7244                     13594
6 10472                     20849
  Lifetime.Engaged.Users Lifetime.Post.Consumers
Lifetime.Post.Consumptions
1  178                   109                        159
2 1457                  1361                       1674
3  177                   113                        154
4 2211                   790                       1119
5  671                   410                        580
6 1191                  1073                       1389
  Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
1  3078
2 11710
```

```
3  2812
4 61027
5  6228
6 16034
  Lifetime.Post.reach.by.people.who.like.your.Page
1  1640
2  6112
3  1503
4 32048
5  3200
6  7852
  Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
comment
1  119
4
2 1108
5
3  132
0
4 1386
58
5  396
19
6 1016
1
  like share Total.Interactions
1   79  17    100
2  130  29    164
3   66  14     80
4 1572 147   1777
5  325  49    393
6  152  33    186
```

Data Cleaning

Dealing with missing values

```
# Summarising the number of missing values in each column
col_missing <- colSums(is.na(Facebook))

# Printing the number of missing values in each column
print(col_missing)
```

```
                                                    Page.total.likes
                                                                   0
                                                                Type
                                                                   0
                                                            Category
```

```
                                                                  0
                                                         Post.Month
                                                                  0
                                                       Post.Weekday
                                                                  0
                                                          Post.Hour
                                                                  0
                                                               Paid
                                                                  1
                                          Lifetime.Post.Total.Reach
                                                                  0
                                    Lifetime.Post.Total.Impressions
                                                                  0
                                              Lifetime.Engaged.Users
                                                                  0
                                            Lifetime.Post.Consumers
                                                                  0
                                         Lifetime.Post.Consumptions
                                                                  0
       Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
                                                                  0
                   Lifetime.Post.reach.by.people.who.like.your.Page
                                                                  0
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
                                                                  0
                                                            comment
                                                                  0
                                                               like
                                                                  1
                                                              share
                                                                  4
                                                 Total.Interactions
                                                                  0
```

```
# Verifying the dimensions of the cleaned dataset
dim(na.omit(Facebook))
```

```
[1] 495  19
```

```
# Summarizing the number of missing values in each column
col_missing <- colSums(is.na(na.omit(Facebook)))

# Printing the number of missing values in each column
print(col_missing)
```

```
                                                   Page.total.likes
                                                                  0
                                                               Type
                                                                  0
                                                           Category
```

```
                                                                           0
                                                                  Post.Month
                                                                           0
                                                                Post.Weekday
                                                                           0
                                                                   Post.Hour
                                                                           0
                                                                        Paid
                                                                           0
                                                   Lifetime.Post.Total.Reach
                                                                           0
                                             Lifetime.Post.Total.Impressions
                                                                           0
                                                    Lifetime.Engaged.Users
                                                                           0
                                                     Lifetime.Post.Consumers
                                                                           0
                                                  Lifetime.Post.Consumptions
                                                                           0
       Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
                                                                           0
                   Lifetime.Post.reach.by.people.who.like.your.Page
                                                                           0
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
                                                                           0
                                                                     comment
                                                                           0
                                                                        like
                                                                           0
                                                                       share
                                                                           0
                                                          Total.Interactions
                                                                           0
```

Dropping categorical and binary variables from the Facebook data

```
 # Using negative indices to drop categorical and binary variables
from the dataset
Facebook_numeric <- subset(Facebook, select = -c(Type, Category,
Post.Month, Post.Weekday, Paid, Post.Hour))
```

Converting all remaining data to numeric

```
# Converting all remaining data to numeric
Facebook_numeric <- as.data.frame(lapply(Facebook_numeric,
as.numeric))
```

Getting a small subset from the Facebook dataset

```
# Getting a small subset from the Facebook dataset
data_subset <- Facebook_numeric[1:50, ]
```

Changing the names of variables to shorter ones

```
# Defining new column names
new_column_names <- c("Total_likes", "Total_Reach",
"Total_Impressions", "Users", "Consumers", "Consumptions",
"Liked_Page_Impressions", "Liked_Page_Reach", "Liked_Engaged_People",
"comment", "like", "share", "Total_Interactions")

# Assigning new column names to the data frame
names(data_subset) <- new_column_names

head(data_subset)
summary(data_subset)
str(data_subset)

  Total_likes Total_Reach Total_Impressions Users Consumers
Consumptions
1 139441       2752        5091              178   109       159

2 139441      10460       19057             1457  1361      1674

3 139441       2413        4373              177   113       154

4 139441      50128       87991             2211   790      1119

5 139441       7244       13594              671   410       580

6 139441      10472       20849             1191  1073      1389

  Liked_Page_Impressions Liked_Page_Reach Liked_Engaged_People comment
like
1  3078                   1640             119                        4
79
2 11710                   6112            1108                        5
130
3  2812                   1503             132                        0
66
4 61027                  32048            1386                       58
1572
5  6228                   3200             396                       19
325
6 16034                   7852            1016                        1
152
  share Total_Interactions
1  17    100
2  29    164
```

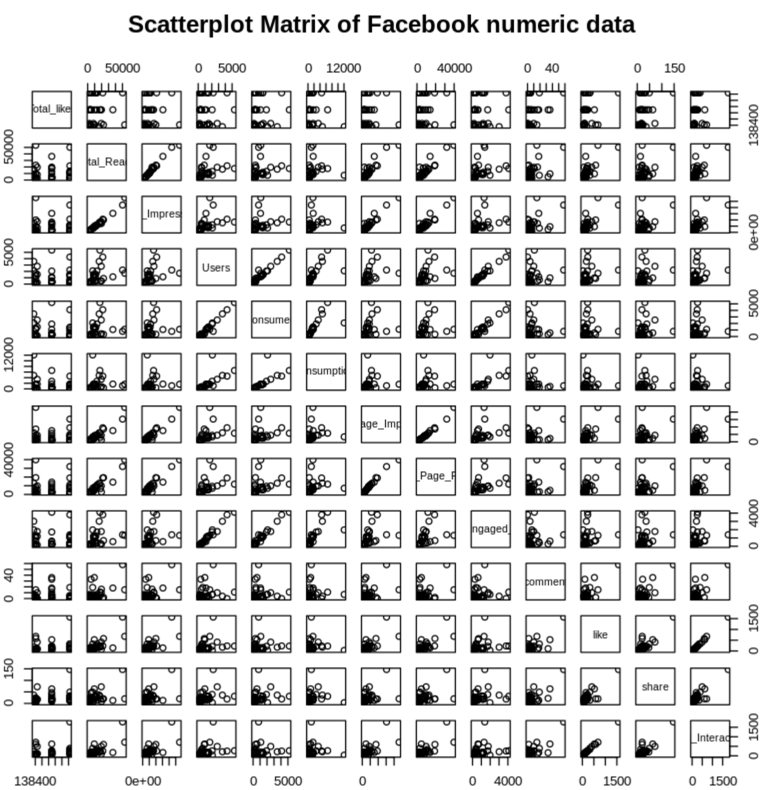
```
3  14     80
4 147   1777
5  49    393
6  33    186

  Total_likes      Total_Reach    Total_Impressions     Users
 Min.   :138353   Min.   : 1384   Min.   :  2467    Min.   :  15.0
 1st Qu.:138414   1st Qu.: 2776   1st Qu.:  5072    1st Qu.: 194.8
 Median :138895   Median : 4817   Median :  9029    Median : 361.5
 Mean   :138829   Mean   : 9766   Mean   : 17750    Mean   : 883.3
 3rd Qu.:139441   3rd Qu.:11806   3rd Qu.: 22116    3rd Qu.:1245.8
 Max.   :139441   Max.   :53264   Max.   :111785    Max.   :5352.0
   Consumers       Consumptions     Liked_Page_Impressions
Liked_Page_Reach
 Min.   :  15.0   Min.   :   20.0   Min.   : 1585          Min.   :
858
 1st Qu.: 124.8   1st Qu.:  161.2   1st Qu.: 3199          1st Qu.:
1774
 Median : 274.0   Median :  409.5   Median : 6044          Median :
3027
 Mean   : 742.5   Mean   : 1163.8   Mean   :12207          Mean   :
6239
 3rd Qu.:1071.8   3rd Qu.: 1416.0   3rd Qu.:15379          3rd Qu.:
7897
 Max.   :5202.0   Max.   :12074.0   Max.   :92512
Max.   :39776
 Liked_Engaged_People    comment           like             share

 Min.   :  15.0       Min.   : 0.00   Min.   :   0.00   Min.   :  0.0

 1st Qu.: 145.2       1st Qu.: 0.00   1st Qu.:  56.25   1st Qu.: 12.0

 Median : 270.5       Median : 3.00   Median :  97.00   Median : 17.5

 Mean   : 682.7       Mean   : 6.26   Mean   : 172.18   Mean   : 22.9

 3rd Qu.:1010.8       3rd Qu.: 6.00   3rd Qu.: 173.50   3rd Qu.: 25.5

 Max.   :4104.0       Max.   :58.00   Max.   :1572.00   Max.   :147.0

 Total_Interactions
 Min.   :   0.0
 1st Qu.:  75.0
 Median : 115.0
 Mean   : 201.3
 3rd Qu.: 209.2
 Max.   :1777.0

'data.frame':	50 obs. of  13 variables:
 $ Total_likes           : num  139441 139441 139441 139441 139441 ...
```

```
 $ Total_Reach            : num  2752 10460 2413 50128 7244 ...
 $ Total_Impressions      : num  5091 19057 4373 87991 13594 ...
 $ Users                  : num  178 1457 177 2211 671 ...
 $ Consumers              : num  109 1361 113 790 410 ...
 $ Consumptions           : num  159 1674 154 1119 580 ...
 $ Liked_Page_Impressions: num  3078 11710 2812 61027 6228 ...
 $ Liked_Page_Reach       : num  1640 6112 1503 32048 3200 ...
 $ Liked_Engaged_People   : num  119 1108 132 1386 396 ...
 $ comment                : num  4 5 0 58 19 1 3 0 0 3 ...
 $ like                   : num  79 130 66 1572 325 ...
 $ share                  : num  17 29 14 147 49 33 27 14 31 26 ...
 $ Total_Interactions     : num  100 164 80 1777 393 ...
```

Further Data Exploration

Scatterplots

```
# Creating a scatterplot matrix of all variables using the function
pairs
pairs(data_subset, main="Scatterplot Matrix of Facebook numeric data")
```

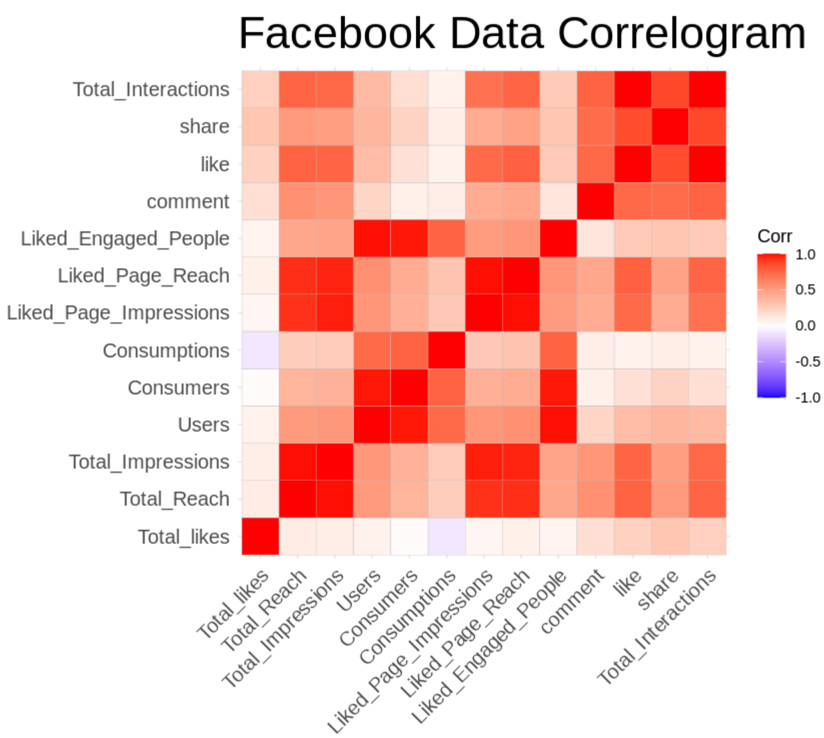


Correlations

```
# Getting correlation matrix
cor_mat <- cor(data_subset)
cor_mat
```

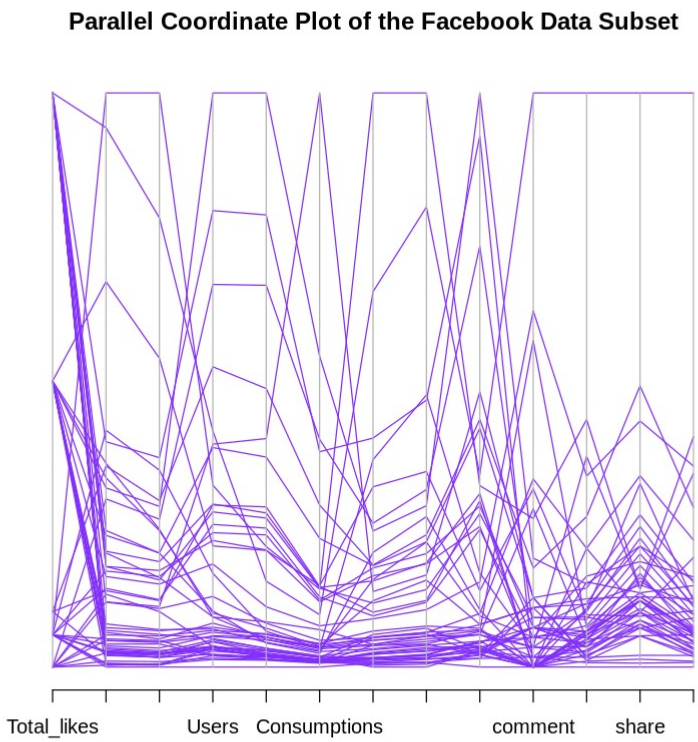
```
                       Total_likes Total_Reach Total_Impressions Users

Total_likes             1.00000000 0.1032159   0.08541053
0.06979444
Total_Reach             0.10321587 1.0000000   0.99199220
0.52224652
Total_Impressions       0.08541053 0.9919922   1.00000000
0.52972585
Users                   0.06979444 0.5222465   0.52972585
1.00000000
Consumers               0.02422572 0.3887539   0.39912677
0.98045374
Consumptions           -0.09816336 0.2574930   0.26688894
0.74398523
Liked_Page_Impressions  0.04996372 0.9323903   0.96587434
0.53717492
Liked_Page_Reach        0.07713628 0.9400969   0.96328443
0.56863515
Liked_Engaged_People    0.05795605 0.4584301   0.47353395
0.98913638
comment                 0.16564093 0.5696069   0.53863649
0.22291628
like                    0.23794894 0.7743234   0.76090545
0.35269717
share                   0.30301150 0.5245607   0.49639224
0.38151990
Total_Interactions      0.24546129 0.7616623   0.74606131
0.35734378
                       Consumers  Consumptions Liked_Page_Impressions
Total_likes            0.02422572 -0.09816336  0.04996372
Total_Reach            0.38875391  0.25749296  0.93239032
Total_Impressions      0.39912677  0.26688894  0.96587434
Users                  0.98045374  0.74398523  0.53717492
Consumers              1.00000000  0.76965513  0.41001330
Consumptions           0.76965513  1.00000000  0.29024134
Liked_Page_Impressions 0.41001330  0.29024134  1.00000000
Liked_Page_Reach       0.43357852  0.30745448  0.99244011
Liked_Engaged_People   0.98333701  0.77289564  0.50712919
comment                0.08242092  0.08706430  0.43394511
like                   0.16219654  0.06588357  0.73676847
share                  0.22814725  0.08614353  0.42914648
Total_Interactions     0.16792606  0.06974966  0.71472022
                       Liked_Page_Reach Liked_Engaged_People comment
```

```
Total_likes             0.07713628       0.05795605
0.16564093
Total_Reach             0.94009688       0.45843012
0.56960687
Total_Impressions       0.96328443       0.47353395
0.53863649
Users                   0.56863515       0.98913638
0.22291628
Consumers               0.43357852       0.98333701
0.08242092
Consumptions            0.30745448       0.77289564
0.08706430
Liked_Page_Impressions  0.99244011       0.50712919
0.43394511
Liked_Page_Reach        1.00000000       0.53611348
0.45630090
Liked_Engaged_People    0.53611348       1.00000000
0.14069073
comment                 0.45630090       0.14069073
1.00000000
like                    0.78160934       0.27954679
0.74797994
share                   0.48488177       0.29888260
0.73373186
Total_Interactions      0.76051942       0.28154102
0.77205127
                        like       share      Total_Interactions
Total_likes             0.23794894 0.30301150 0.24546129
Total_Reach             0.77432336 0.52456074 0.76166225
Total_Impressions       0.76090545 0.49639224 0.74606131
Users                   0.35269717 0.38151990 0.35734378
Consumers               0.16219654 0.22814725 0.16792606
Consumptions            0.06588357 0.08614353 0.06974966
Liked_Page_Impressions  0.73676847 0.42914648 0.71472022
Liked_Page_Reach        0.78160934 0.48488177 0.76051942
Liked_Engaged_People    0.27954679 0.29888260 0.28154102
comment                 0.74797994 0.73373186 0.77205127
like                    1.00000000 0.84930218 0.99838986
share                   0.84930218 1.00000000 0.87488389
Total_Interactions      0.99838986 0.87488389 1.00000000

# Plotting Facebook data correlogram
ggcorrplot(cor_mat, title = "Facebook Data Correlogram") +
  theme(plot.title = element_text(size = 27, face = "plain"))
```

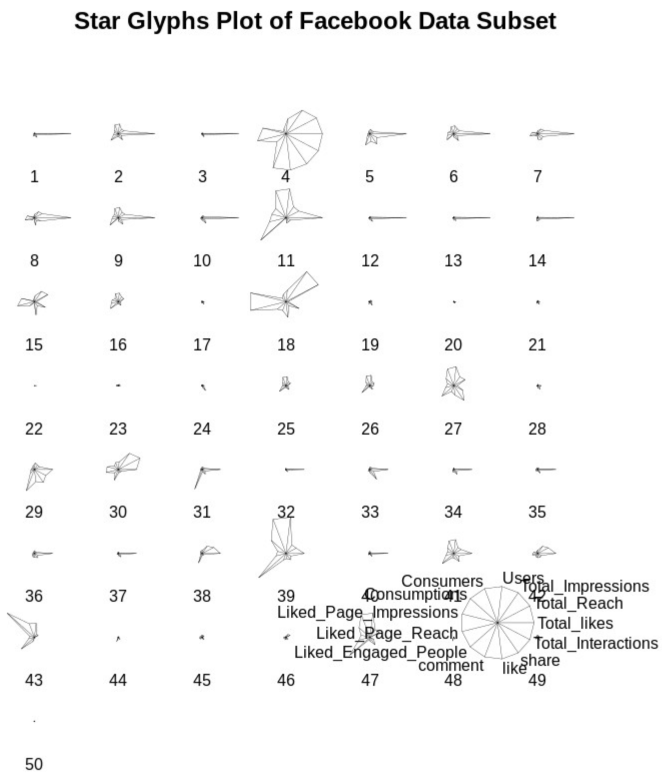


Parallel coordinates, stars and faces

```
# Creating parallel coordinate plot
parcoord(data_subset, col = "#822EFF", main = "Parallel Coordinate
Plot of the Facebook Data Subset")
```

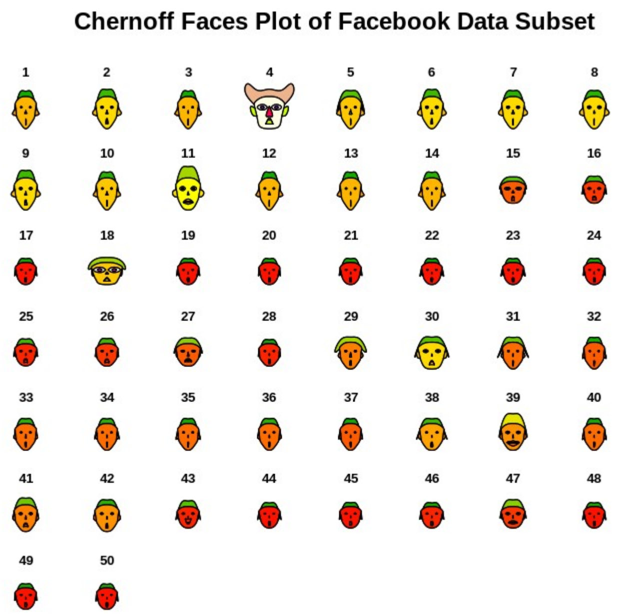



```r
# Creating star glyphs using the function stars
stars(data_subset, xlim = c(0, 20), key.loc = c(15, 5), main = "Star Glyphs Plot of Facebook Data Subset")
```




```r
# Creating Chernoff faces
faces(data_subset, cex = 1, main = "Chernoff Faces Plot of Facebook
Data Subset")
```

```
effect of variables:
 modified item       Var
 "height of face   " "Total_likes"
 "width of face    " "Total_Reach"
 "structure of face" "Total_Impressions"
 "height of mouth  " "Users"
 "width of mouth   " "Consumers"
 "smiling          " "Consumptions"
 "height of eyes   " "Liked_Page_Impressions"
 "width of eyes    " "Liked_Page_Reach"
 "height of hair   " "Liked_Engaged_People"
 "width of hair   "  "comment"
 "style of hair   "  "like"
 "height of nose  "  "share"
 "width of nose   "  "Total_Interactions"
 "width of ear    "  "Total_likes"
 "height of ear   "  "Total_Reach"
```



Distance and Similarity

```r
# Creating a new data set where the variables are scaled
data_scaled <- scale(data_subset, center = TRUE, scale = TRUE)
```

Euclidean Distance

```r
D <- dist(data_scaled)

# Converting to regular matrix object
D_mat <- as.matrix(D)
D_mat[1:5, 1:5]
```

	1	2	3	4	5
1	0.0000000	2.494900	0.4088489	14.07190	2.610501
2	2.4949001	0.000000	2.5969560	12.78626	2.513045
3	0.4088489	2.596956	0.0000000	14.34147	2.950376
4	14.0718975	12.786257	14.3414749	0.00000	11.731736
5	2.6105011	2.513045	2.9503762	11.73174	0.000000

```r
# Minimum distance between categories
min_dist <- min(D_mat[which(D_mat > 0)])

# Finding indices of minimum element in distance matrix
which_min <- which(D_mat == min_dist, arr.ind = TRUE)
```

```
# Minimum distance between categories
print('Minimum distance between categories: ')
min_dist

# Indices of minimum element in distance matrix
print('Indices of minimum element in distance matrix: ')
which_min
```

```
[1] "Minimum distance between categories: "
```

```
[1] 0.1719778
```

```
[1] "Indices of minimum element in distance matrix: "
```

```
   row col
13 13   3
3   3  13
```

```
# Maximum distance
max_dist <- max(D_mat)

# Finding indices of maximum element in distance matrix
which_max <- which(D_mat == max_dist, arr.ind = TRUE)


# Maximum distance between categories
print('Maximum distance between categories: ')
max_dist

# Indices of maximum element in distance matrix
print('Indices of maximum element in distance matrix: ')
which_max
```

```
[1] "Maximum distance between categories: "
```

```
[1] 15.12548
```

```
[1] "Indices of maximum element in distance matrix: "
```

```
   row col
22 22   4
4   4  22
```

Manhattan Distance

```
# Manhattan distance matrix
D_man <- dist(data_scaled, method = "manhattan")

# Converting to regular matrix object
D_man_mat <- as.matrix(D_man)
```

```r
# Minimum distance between categories
min_man_dist <- min(D_man_mat[which(D_man_mat > 0)])

# Finding indices of minimum element in distance matrix
which_min_man <- which(D_man_mat == min_man_dist, arr.ind = TRUE)

# Minimum distance between categories
print('Minimum distance between categories: ')
which_min_man

# Indices of minimum element in distance matrix
print('Indices of minimum element in distance matrix: ')
which_min_man
```

```
[1] "Minimum distance between categories: "
```

```
   row col
13 13  12
12 12  13
```

```
[1] "Indices of minimum element in distance matrix: "
```

```
   row col
13 13  12
12 12  13
```

```r
# Maximum distance between categories
max_man_dist <- max(D_man_mat)

# Finding indices of maximum element in distance matrix
which_max_man <- which(D_man_mat == max_man_dist, arr.ind = TRUE)

# Maximum distance between categories
print('Maximum distance between categories: ')
max_man_dist

# Indices of maximum element in distance matrix
print('Indices of maximum element in distance matrix: ')
which_max_man
```

```
[1] "Maximum distance between categories: "
```

```
[1] 47.66058
```

```
[1] "Indices of maximum element in distance matrix: "
```

```
   row col
22 22   4
4   4  22
```

Analysis

#PCA

Applying PCA

```
# Applying PCA using a built-in function
pca <- prcomp(data_subset)

# Eigenvectors returned by prcomp
pca$rotation
```

```
PC1
Page.total.likes                                                   -
0.0011506736
Lifetime.Post.Total.Reach                                          -
0.3735278893
Lifetime.Post.Total.Impressions                                    -
0.7121597789
Lifetime.Engaged.Users                                             -
0.0206049632
Lifetime.Post.Consumers                                            -
0.0150017660
Lifetime.Post.Consumptions                                         -
0.0195533943
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page       -
0.5367667788
Lifetime.Post.reach.by.people.who.like.your.Page                   -
0.2526515425
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post -
0.0152010229
comment                                                            -
0.0001852963
like                                                               -
0.0064022268
share                                                              -
0.0003750343
Total.Interactions                                                 -
0.0069625574

PC2
Page.total.likes                                                   -
0.0138126650
Lifetime.Post.Total.Reach                                          -
0.4849189347
Lifetime.Post.Total.Impressions                                    -
0.3942209929
Lifetime.Engaged.Users
0.0301600677
```

```
Lifetime.Post.Consumers
0.0320092909
Lifetime.Post.Consumptions
0.0807859877
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
0.7136710153
Lifetime.Post.reach.by.people.who.like.your.Page
0.2988562042
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
0.0458390570
comment                                                              -
0.0009529192
like                                                                 -
0.0036540424
share                                                                -
0.0012040422
Total.Interactions                                                   -
0.0058110038

PC3
Page.total.likes
9.135542e-03
Lifetime.Post.Total.Reach                                            -
6.969468e-02
Lifetime.Post.Total.Impressions
1.004978e-02
Lifetime.Engaged.Users                                               -
3.520677e-01
Lifetime.Post.Consumers                                              -
3.693505e-01
Lifetime.Post.Consumptions                                           -
7.936668e-01
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
1.214985e-01
Lifetime.Post.reach.by.people.who.like.your.Page                     -
5.436407e-02
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post -
2.943268e-01
comment
6.611536e-05
like
1.313435e-02
share                                                                -
4.956113e-04
Total.Interactions
1.270486e-02

PC4
Page.total.likes                                                     -
```

```
0.0949335892
Lifetime.Post.Total.Reach                                                 -
0.4800926435
Lifetime.Post.Total.Impressions
0.3805218011
Lifetime.Engaged.Users                                                    -
0.1228242428
Lifetime.Post.Consumers                                                   -
0.0439924070
Lifetime.Post.Consumptions
0.2266160809
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
0.1635097610
Lifetime.Post.reach.by.people.who.like.your.Page                          -
0.7042670732
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post -
0.0953353480
comment                                                                   -
0.0008174156
like                                                                      -
0.0835719822
share                                                                     -
0.0076236762
Total.Interactions                                                        -
0.0920130740

PC5
Page.total.likes                                                          -
0.091809859
Lifetime.Post.Total.Reach
0.209162374
Lifetime.Post.Total.Impressions                                           -
0.145887599
Lifetime.Engaged.Users                                                    -
0.473747552
Lifetime.Post.Consumers                                                   -
0.503197804
Lifetime.Post.Consumptions
0.549811826
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
0.009873503
Lifetime.Post.reach.by.people.who.like.your.Page
0.127979999
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post -
0.359612776
comment
0.001196575
like
0.020660785
```

```
share                                                                  -
0.002677735
Total.Interactions
0.019179625

PC6
Page.total.likes
0.951319549
Lifetime.Post.Total.Reach                                              -
0.171356070
Lifetime.Post.Total.Impressions
0.113094016
Lifetime.Engaged.Users                                                 -
0.027301305
Lifetime.Post.Consumers                                                -
0.115633955
Lifetime.Post.Consumptions
0.091923248
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page          -
0.059099064
Lifetime.Post.reach.by.people.who.like.your.Page
0.053288206
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post -
0.020975261
comment
0.004109404
like
0.098670278
share
0.013666534
Total.Interactions
0.116446216

PC7
Page.total.likes
0.27227378
Lifetime.Post.Total.Reach
0.44277118
Lifetime.Post.Total.Impressions                                        -
0.30712122
Lifetime.Engaged.Users                                                 -
0.16961788
Lifetime.Post.Consumers
0.17146868
Lifetime.Post.Consumptions                                             -
0.02861269
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
0.26786990
Lifetime.Post.reach.by.people.who.like.your.Page                      -
```

```
0.33579847
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
0.09259406
comment                                                               -
0.01904347
like                                                                  -
0.39789057
share                                                                 -
0.05404173
Total.Interactions                                                    -
0.47097577

PC8
Page.total.likes                                                      -
0.008822031
Lifetime.Post.Total.Reach                                             -
0.325267180
Lifetime.Post.Total.Impressions
0.232251030
Lifetime.Engaged.Users                                                -
0.179794921
Lifetime.Post.Consumers
0.231928651
Lifetime.Post.Consumptions
0.008860553
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page          -
0.275722499
Lifetime.Post.reach.by.people.who.like.your.Page
0.451407059
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post -
0.251881800
comment                                                               -
0.031092967
like                                                                  -
0.418826596
share                                                                 -
0.033066674
Total.Interactions                                                    -
0.482986237

PC9
Page.total.likes                                                      -
0.055062701
Lifetime.Post.Total.Reach                                             -
0.119603918
Lifetime.Post.Total.Impressions
0.098674175
Lifetime.Engaged.Users                                                -
0.396891108
```

```
Lifetime.Post.Consumers                                                -
0.287281697
Lifetime.Post.Consumptions                                             -
0.013708972
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page           -
0.096596311
Lifetime.Post.reach.by.people.who.like.your.Page
0.108423954
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
0.835389979
comment
0.009913028
like                                                                   -
0.096702837
share
0.027485967
Total.Interactions                                                     -
0.059303842

PC10
Page.total.likes                                                       -
0.0054919294
Lifetime.Post.Total.Reach                                              -
0.0258187351
Lifetime.Post.Total.Impressions
0.0205346982
Lifetime.Engaged.Users
0.4473413003
Lifetime.Post.Consumers                                                -
0.4743670778
Lifetime.Post.Consumptions                                             -
0.0003327954
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page           -
0.0179071026
Lifetime.Post.reach.by.people.who.like.your.Page
0.0166732549
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
0.0457659293
comment                                                                -
0.2004084221
like
0.2025661655
share                                                                  -
0.4959783497
Total.Interactions                                                     -
0.4938206063

PC11
Page.total.likes
```

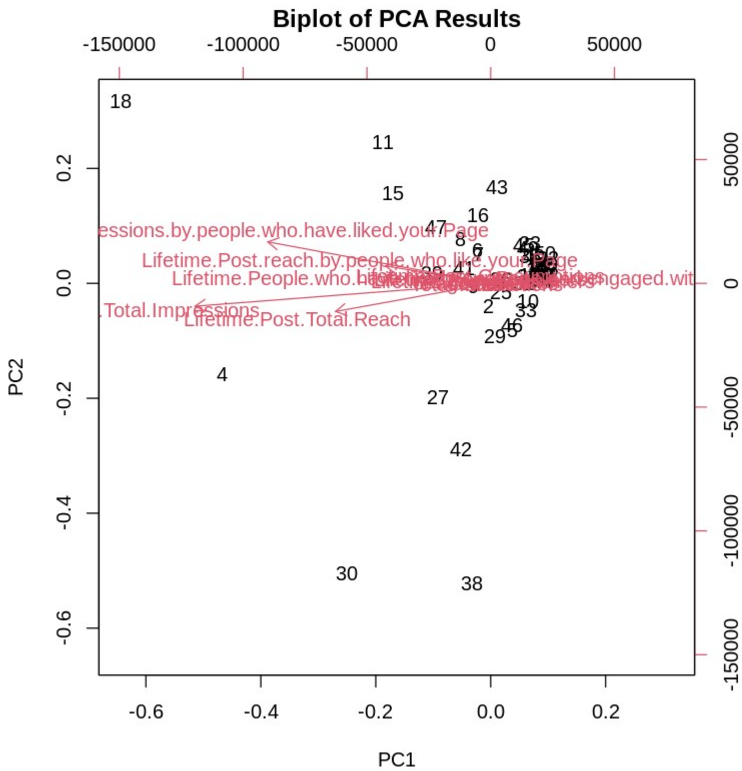
```
0.0010382339
Lifetime.Post.Total.Reach                                             -
0.0044763820
Lifetime.Post.Total.Impressions
0.0027229842
Lifetime.Engaged.Users                                                -
0.2192725081
Lifetime.Post.Consumers
0.2151072371
Lifetime.Post.Consumptions                                            -
0.0009918193
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page          -
0.0039195170
Lifetime.Post.reach.by.people.who.like.your.Page
0.0069902413
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
0.0042924295
comment
0.6372343438
like
0.1355447030
share                                                                 -
0.6884678042
Total.Interactions
0.0843112426

PC12
Page.total.likes
0.0003578024
Lifetime.Post.Total.Reach
0.0059708701
Lifetime.Post.Total.Impressions                                       -
0.0041599236
Lifetime.Engaged.Users                                                -
0.4094295936
Lifetime.Post.Consumers
0.3959767598
Lifetime.Post.Consumptions
0.0005323909
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page
0.0048525229
Lifetime.Post.reach.by.people.who.like.your.Page                      -
0.0086331161
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
0.0134843403
comment                                                               -
0.5498374551
like
0.5745752325
```

```
share                                                              -
0.1580957515
Total.Interactions                                                 -
0.1333579741

PC13
Page.total.likes                                                   -
3.483365e-17
Lifetime.Post.Total.Reach                                          -
7.351263e-17
Lifetime.Post.Total.Impressions
5.214977e-17
Lifetime.Engaged.Users
6.082988e-16
Lifetime.Post.Consumers                                            -
5.163641e-16
Lifetime.Post.Consumptions                                         -
1.810952e-17
Lifetime.Post.Impressions.by.people.who.have.liked.your.Page       -
1.113212e-17
Lifetime.Post.reach.by.people.who.like.your.Page                   -
1.538640e-17
Lifetime.People.who.have.liked.your.Page.and.engaged.with.your.post
5.568491e-18
comment
5.000000e-01
like
5.000000e-01
share
5.000000e-01
Total.Interactions                                                 -
5.000000e-01
```

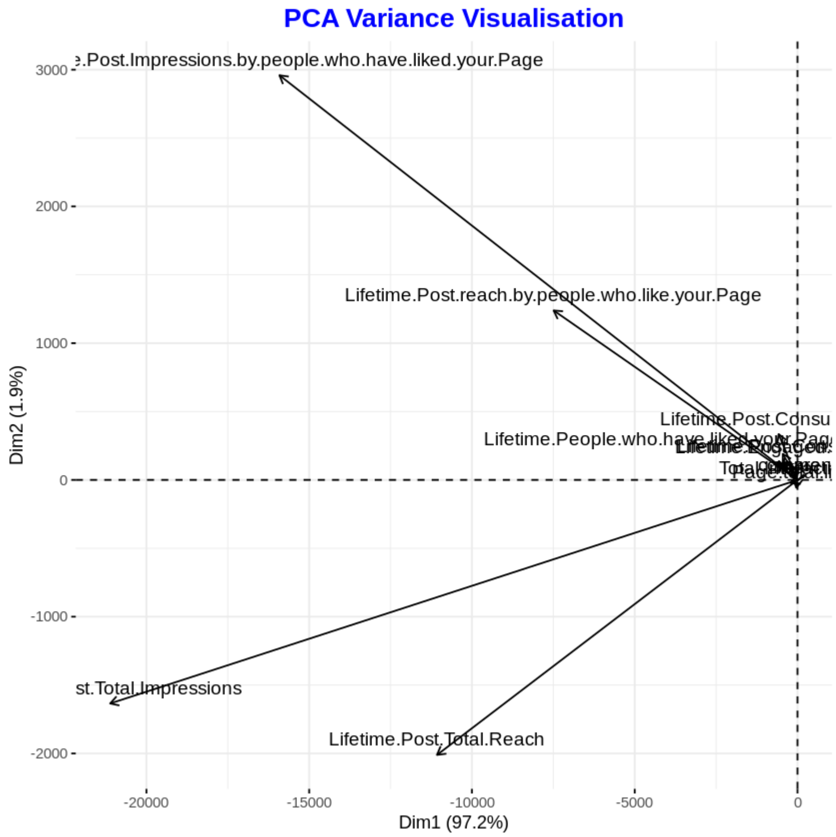
Biplots

```r
# Biplots
biplot(pca, main = "Biplot of PCA Results")
```

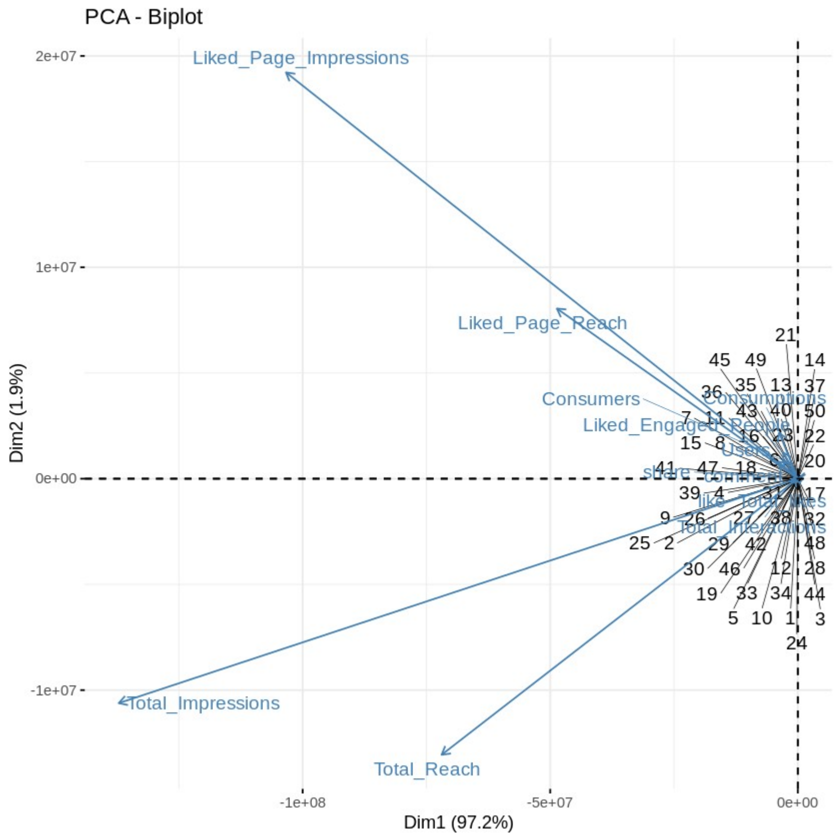
```
Warning message in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col =
col[2L], length = arrow.len):
“zero-length arrow is of indeterminate angle and so skipped”
```



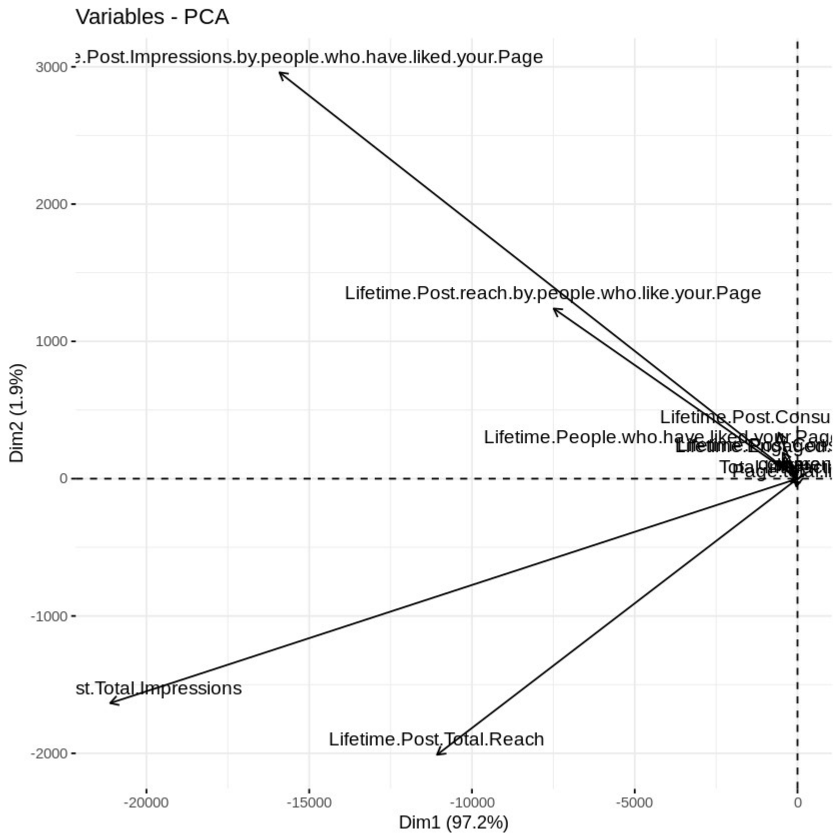

```
# Visualising variable contributions to PCA
pca_var_plot <- fviz_pca_var(pca, ggtheme = theme_minimal() +
theme(plot.title = element_text(hjust = 0.5, size = 16, face = "bold",
color = "blue")))
plot_title <- "PCA Variance Visualisation"
plot <- pca_var_plot + ggtitle(plot_title)
plot
```




```
factoextra::fviz_pca_biplot(pca, repel = TRUE)
```




```
factoextra::fviz_pca_var(pca)
```




```
# Proportion of variance explained
summary(pca)

Importance of components:
                            PC1       PC2       PC3       PC4
PC5
Standard deviation     2.965e+04 4.146e+03 2.371e+03 1.111e+03
864.10794
Proportion of Variance 9.723e-01 1.902e-02 6.220e-03 1.370e-03
0.00083
Cumulative Proportion  9.723e-01 9.913e-01 9.975e-01 9.989e-01
0.99970
                            PC6       PC7       PC8   PC9  PC10  PC11
PC12
Standard deviation     426.6690 265.75921 118.29092 51.31 9.787 5.878
4.781
Proportion of Variance   0.0002   0.00008   0.00002  0.00 0.000 0.000
0.000
Cumulative Proportion    0.9999   0.99998   1.00000  1.00 1.000 1.000
1.000
                            PC13
Standard deviation     2.291e-14
Proportion of Variance 0.000e+00
Cumulative Proportion  1.000e+00
```

Centring and Scaling Variables

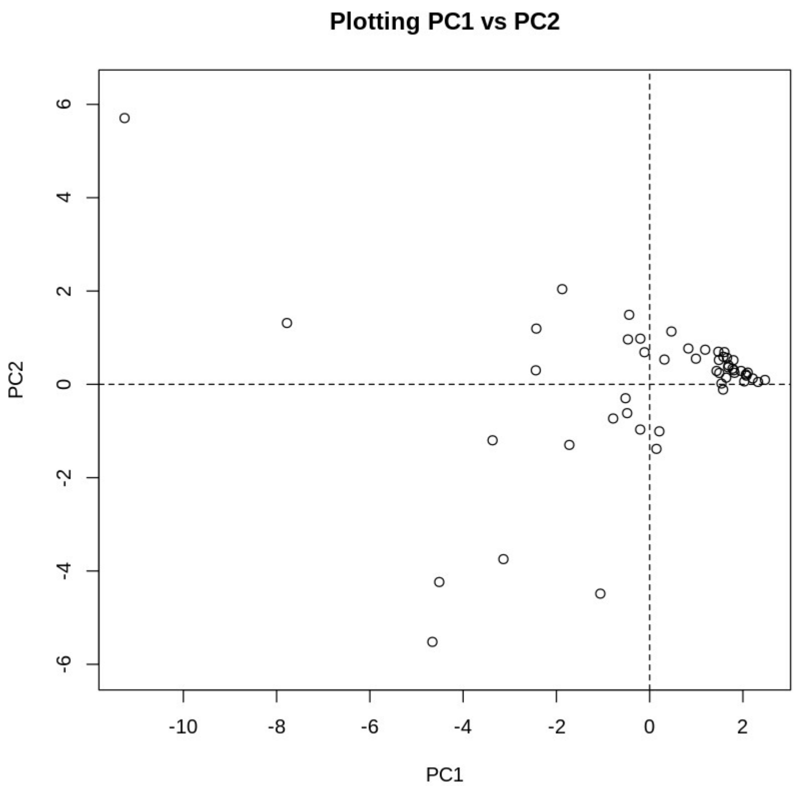
```
# Centring and scaling variables
X <- scale(data_subset, center = TRUE, scale = TRUE)
```

Covariance Matrix, Eigenvalues and Eigenvectors

```
# Getting covariance matrix
Sigma <- cov(X)

# Getting eigenvalues and eigenvectors
eig <- eigen(Sigma)

# Computing scores
Z <- X %*% eig$vectors
```

PC1 and PC2 Scatterplot

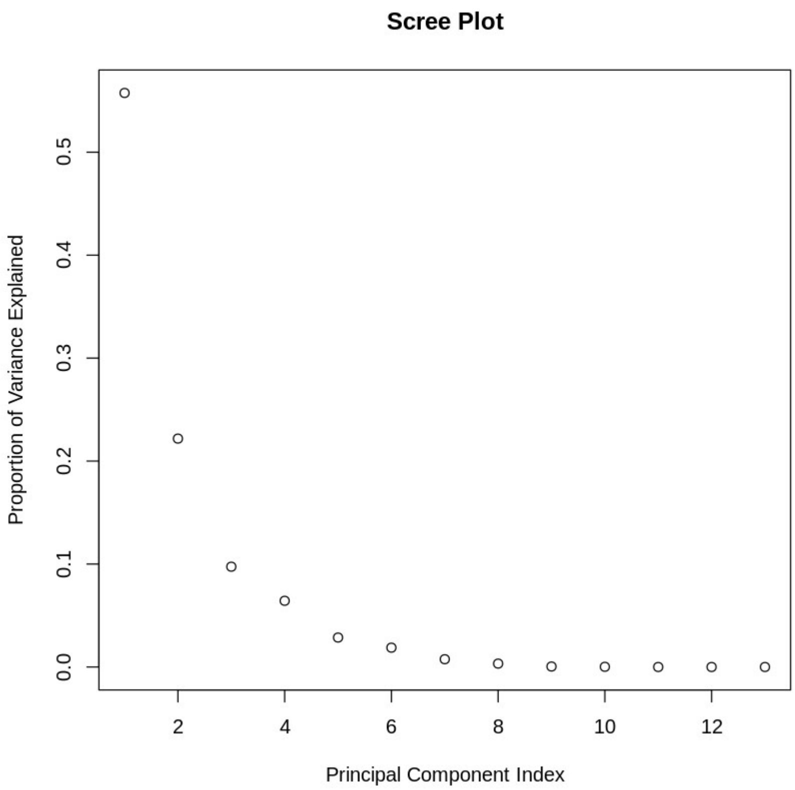
```
# Defining a function to plot PC1 vs PC2
plot_PC1_vs_PC2 <- function(Z) {
  # Plotting PC1 vs PC2
  plot(Z[,1:2], asp = 1, xlab = "PC1", ylab = "PC2")
  abline(v = 0, lty = 2)
  abline(h = 0, lty = 2)

  # Adding title
  title("Plotting PC1 vs PC2")
}

# Calling the function with your data Z
plot_PC1_vs_PC2(Z)
```



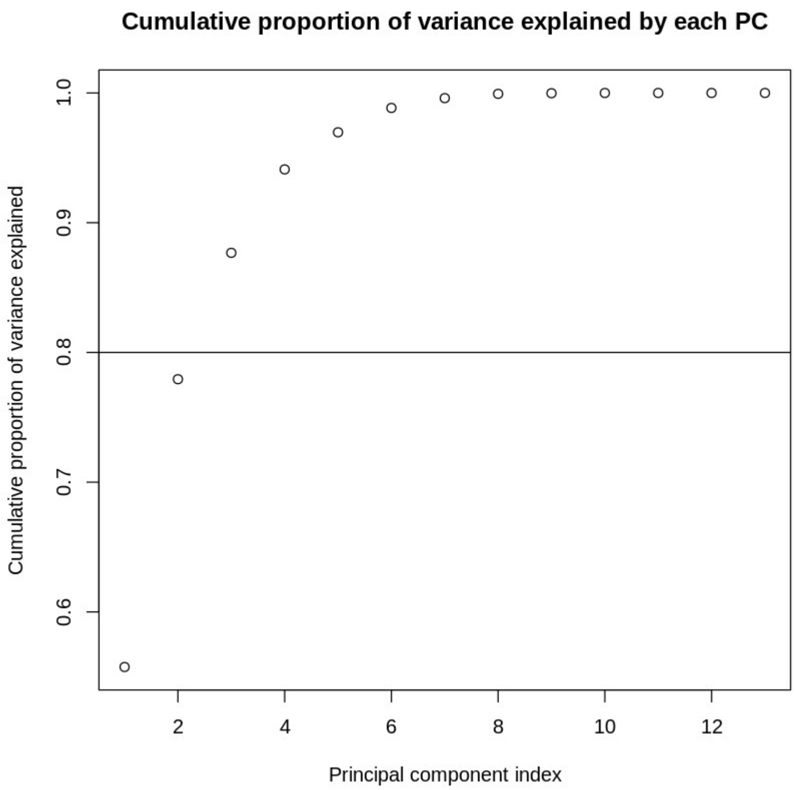

```r
# Principal component variances
pc_var <- eig$values
# Proportion of variance explained
pc_prop_var <- pc_var/sum(pc_var)
# Cumulative proportion of variance explained
pc_cumul_prop_var <- cumsum(pc_prop_var)
```

Elbow Plot & Cumulative Proportion of Variance Explained

```r
# Plotting the proportion of variance explained by each principal
component using Elbow Plot
plot(pc_prop_var, xlab = "Principal Component Index",
     ylab = "Proportion of Variance Explained",
     main = "Scree Plot")
```

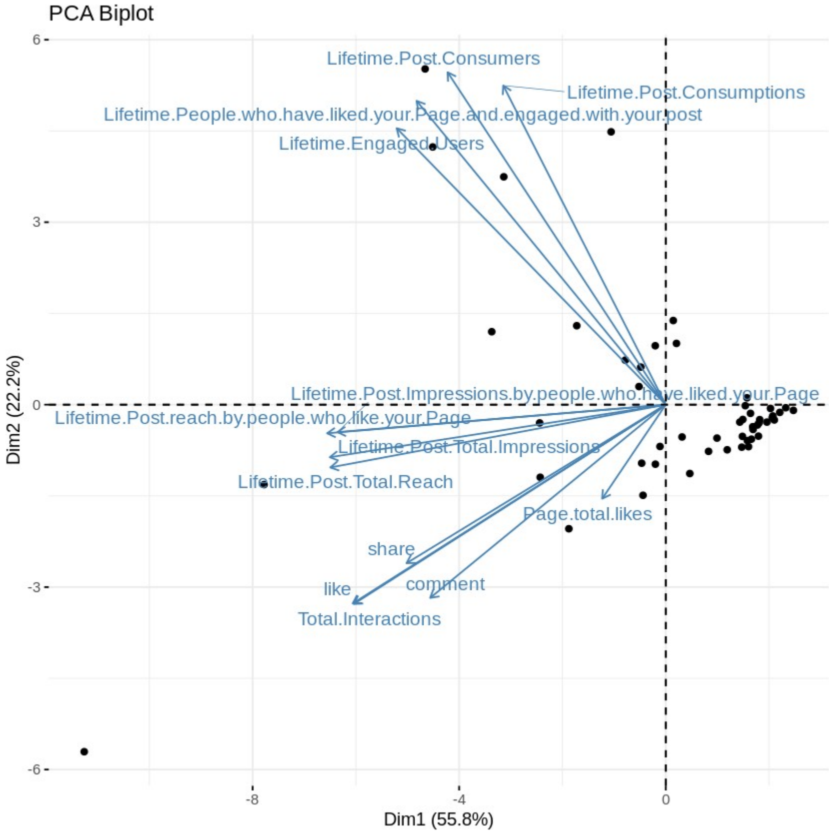



```
# Plotting the cumulative proportion of variance explained by each
principal component
plot(pc_cumul_prop_var, xlab = "Principal component index",
     ylab = "Cumulative proportion of variance explained",
     main = "Cumulative proportion of variance explained by each PC")
abline(h = 0.8)
```



Final PCA Plot

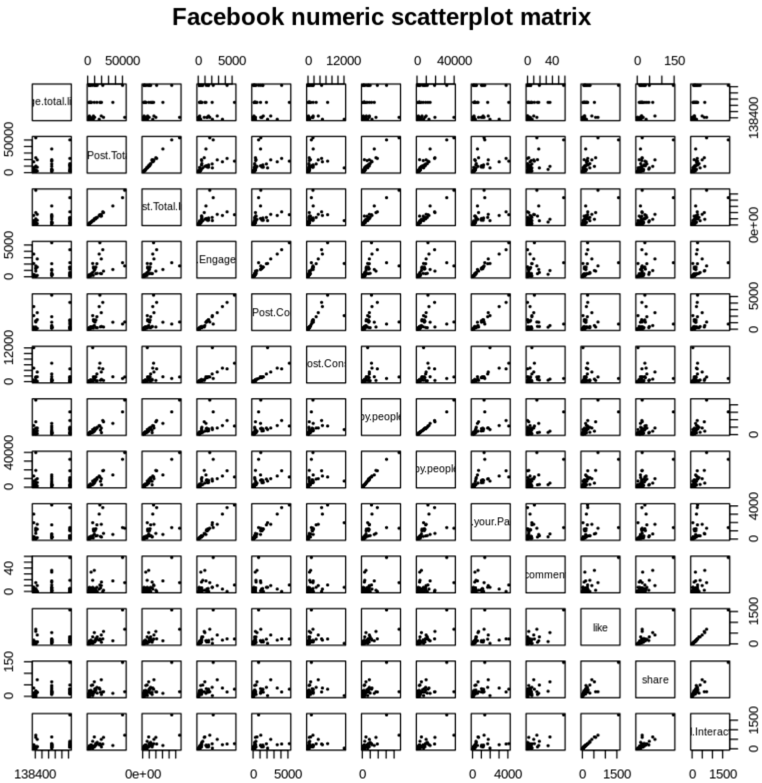
```r
# Applying PCA using a built-in function
pca <- prcomp(X)
factoextra::fviz_pca_biplot(pca, label = "var", repel = TRUE) +
  ggtitle("PCA Biplot")
```



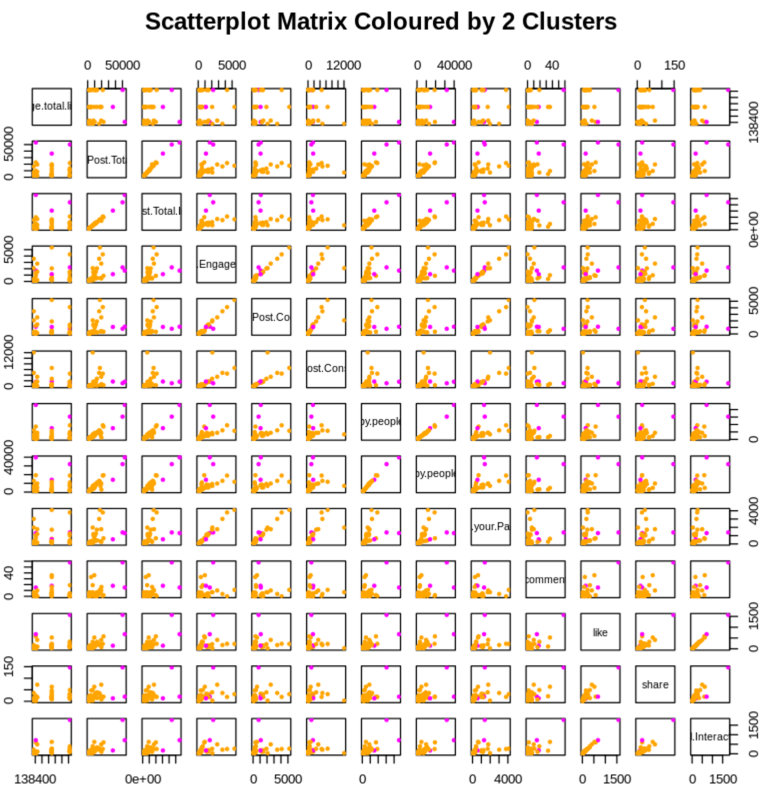
Clustering Analysis

K-Means Clustering Analysis

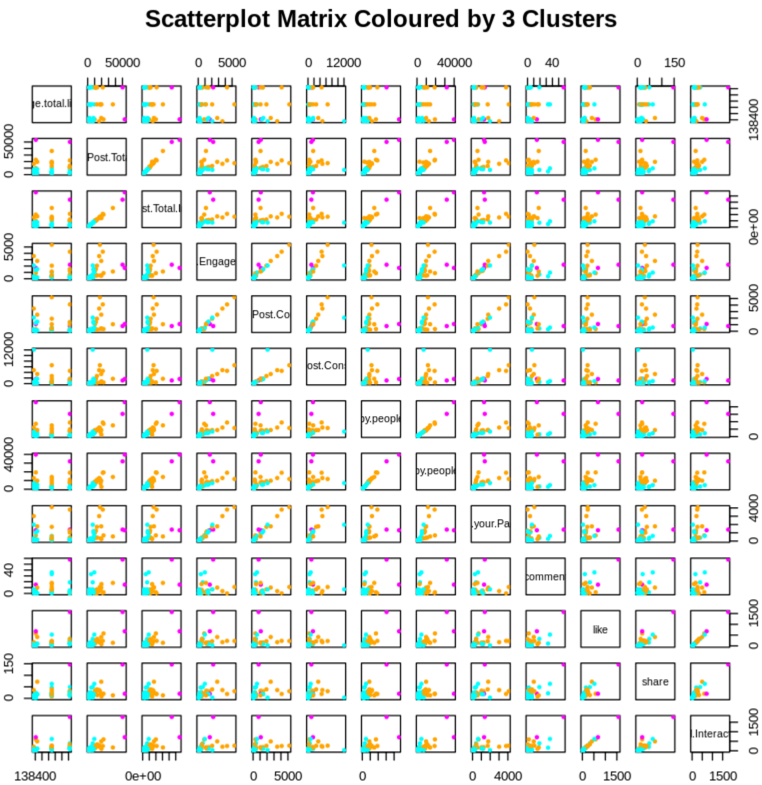
```
# Scatterplot matrix
pairs(data_subset, pch = 20, cex = 0.5, main = "Facebook numeric
scatterplot matrix")
```



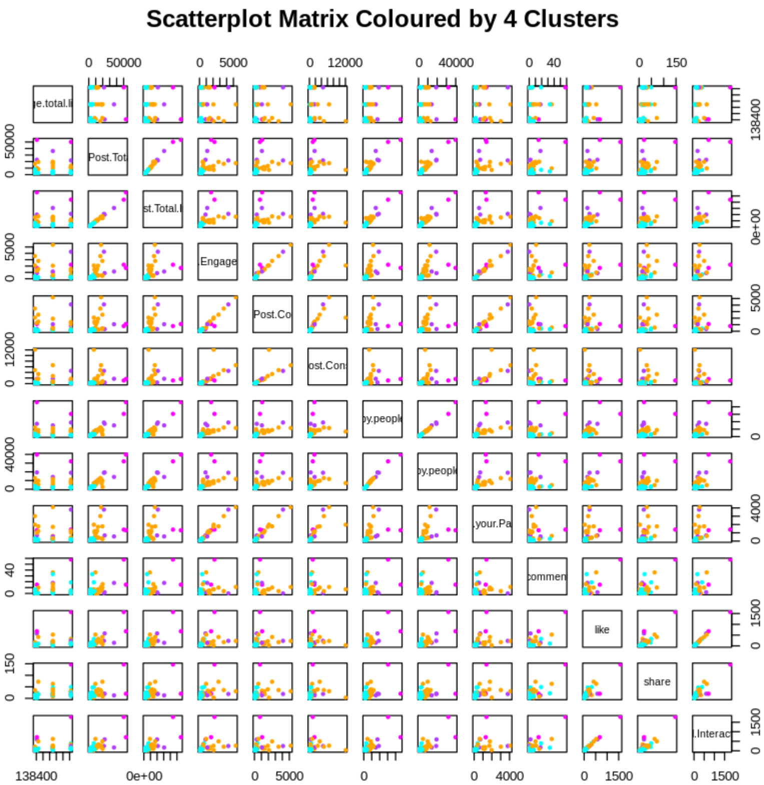

```
# K-means with 2, 3, and 4 clusters
km2 <- kmeans(data_subset, centers = 2, nstart = 50)
km3 <- kmeans(data_subset, centers = 3, nstart = 50)
km4 <- kmeans(data_subset, centers = 4, nstart = 50)

# Colour palette
pal <- c("#FFA500", "#FF00FF", "#00FFFF", "#B041FF")

# Scatterplot matrices coloured by clusters
pairs(data_subset, pch = 20, cex = 0.8, col = pal[km2$cluster], main =
"Scatterplot Matrix Coloured by 2 Clusters")
```



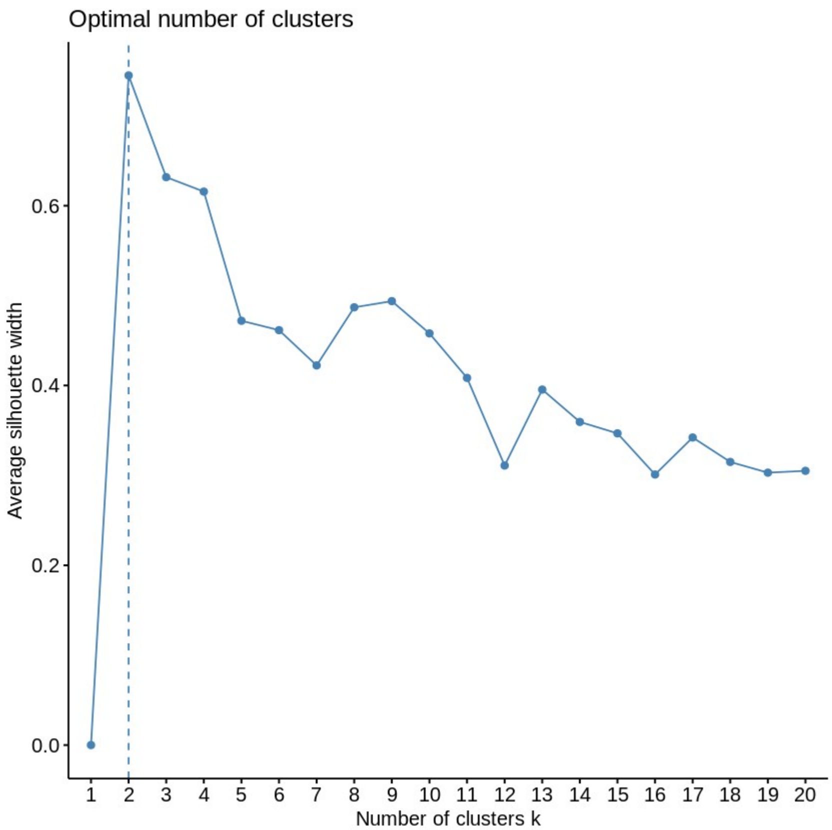

```r
# Scatterplot matrices coloured by clusters (with different clusters)
pairs(data_subset, pch = 20, cex = 0.8, col = pal[km3$cluster], main =
"Scatterplot Matrix Coloured by 3 Clusters")
```

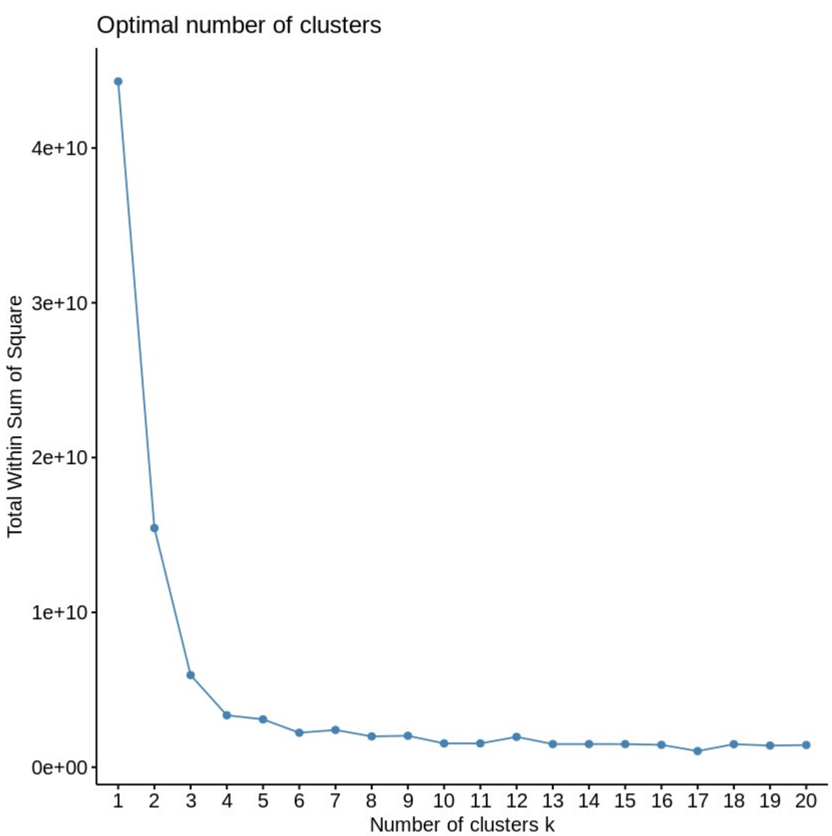



```
# Scatterplot matrices coloured by clusters (with different clusters)
pairs(data_subset, pch = 20, cex = 0.8, col = pal[km4$cluster], main =
"Scatterplot Matrix Coloured by 4 Clusters")
```



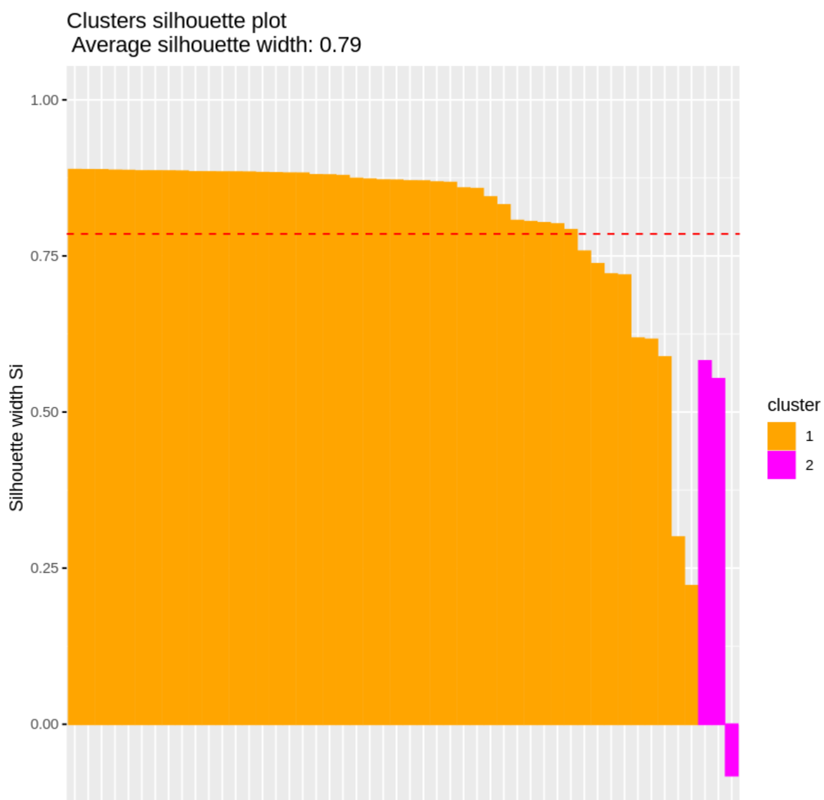

```r
# Criteria to select number of clusters
fviz_nbclust(x = data_subset, FUNcluster = kmeans, method =
"silhouette", k.max = 20)
fviz_nbclust(x = data_subset, FUNcluster = kmeans, method = "wss",
k.max = 20)
```


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of clusters k


Optimal number of clusters
Total Within Sum of Square
0e+00
1e+10
2e+10
3e+10
4e+10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of clusters k

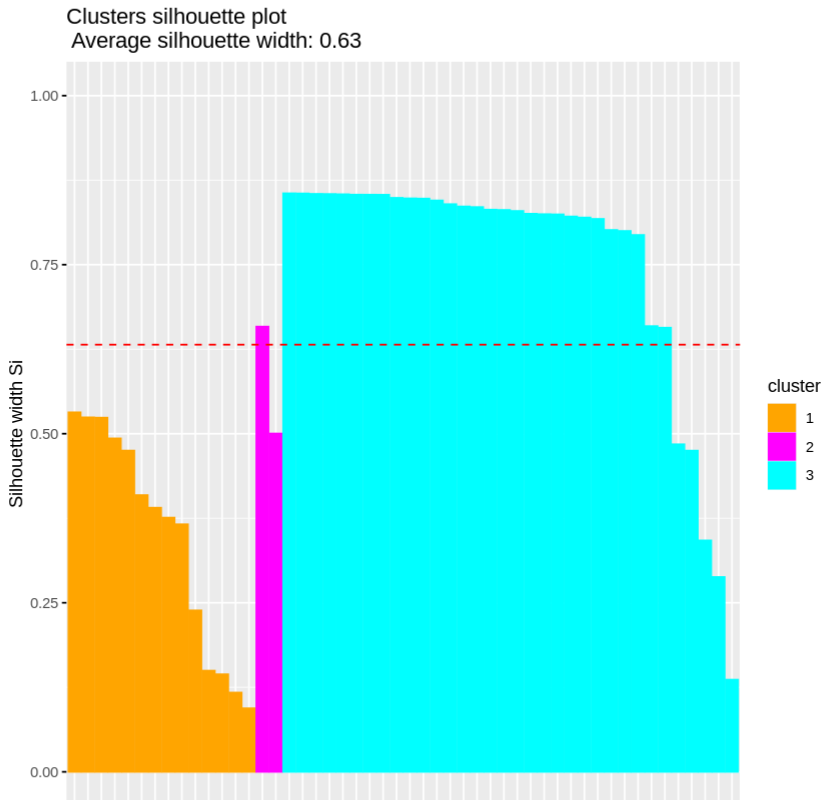
```
# Creating silhouette objects
sil2 <- silhouette(x = km2$cluster, dist = dist(data_subset))
sil3 <- silhouette(x = km3$cluster, dist = dist(data_subset))
sil4 <- silhouette(x = km4$cluster, dist = dist(data_subset))

# Silhouette plot with 2 clusters
fviz_silhouette(sil2) +
  scale_fill_manual(values = pal) +
  scale_color_manual(values = pal)
```

```
  cluster size ave.sil.width
1       1   47          0.81
2       2    3          0.35
```



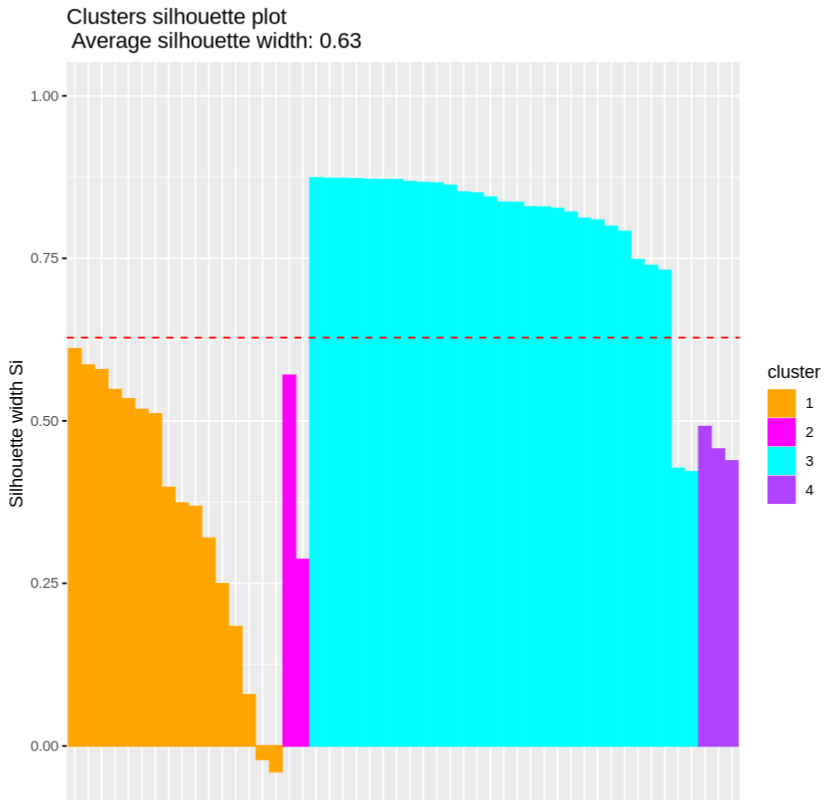

```
# Silhouette plot with 3 clusters
fviz_silhouette(sil3) +
  scale_fill_manual(values = pal) +
  scale_color_manual(values = pal)
```

```
  cluster size ave.sil.width
1       1   14          0.35
2       2    2          0.58
3       3   34          0.75
```



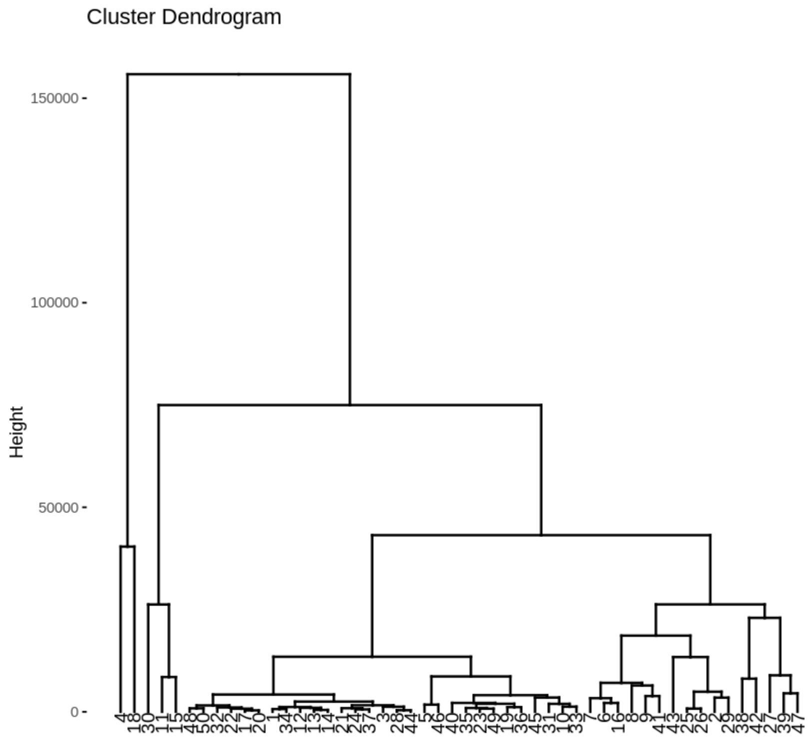

```
# Silhouette plot with 4 clusters
fviz_silhouette(sil4) +
  scale_fill_manual(values = pal) +
  scale_color_manual(values = pal)
```

	cluster	size	ave.sil.width
1	1	16	0.36
2	2	2	0.43
3	3	29	0.81
4	4	3	0.46

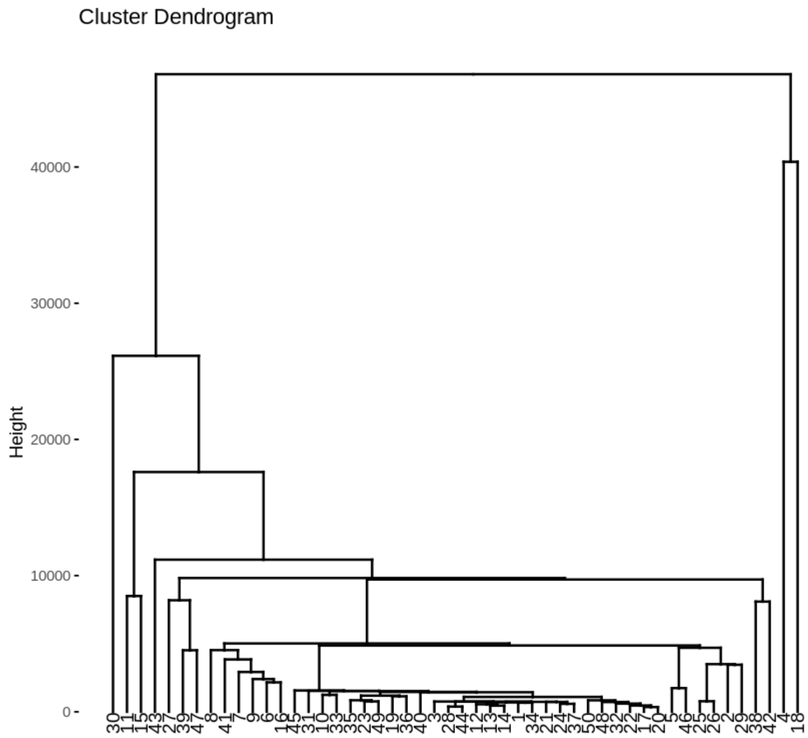


Hierarchical Clustering Analysis

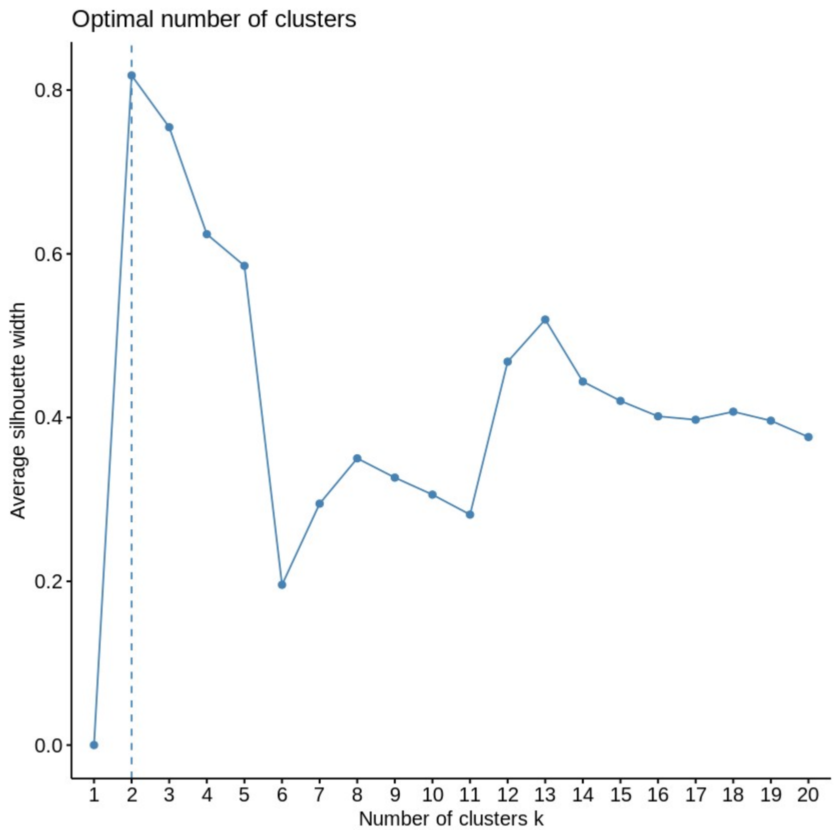
```
D <- dist(data_subset)

# Applying complete linkage
hc1 <- hclust(D, method = "complete")

# Plotting dendrogram
fviz_dend(hc1)

Warning message:
“The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none"
instead as
of ggplot2 3.3.4.
ℹ The deprecated feature was likely used in the factoextra package.
  Please report the issue at
<https://github.com/kassambara/factoextra/issues>.”
```




```
# Applying single linkage
hc2 <- hclust(D, method = "single")

# Plotting dendrogram
fviz_dend(hc2)
```




```r
hclust_custom <- function(x, k = 2) {
  hc <- hclust(dist(x), method = "single")
  clust <- cutree(hc, k = k)
  return(list(cluster = clust))
}

fviz_nbclust(x = data_subset, FUNcluster = hclust_custom,
             method = "silhouette", k.max = 20)
```


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
0.8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of clusters k