Tech gadgets Brand Classification

**ABSTRACT**

This project aims to analyze and discover the common standard features shared by various tech gadgets from various brands that help the brand stand out in the market.

**PROBLEM STATEMENT**

Each of the technological gadgets is available in a variety of brands. In this modern era, it is difficult to determine which brand a customer should purchase based on their preferences, so clustering analysis and PCA analysis are used to classify the brand based on the characteristics and standard features shared by each brand.

**INTRODUCTION**

Smartphone brand data is analyzed and clustered based on feature similarities, and the most important features that contribute the most to the brand are identified using dimensionality reduction using PCA.

**METHODOLOGY USED**

**Clustering Analysis using K- Means Algorithm**

**Principle Component Analysis**

Clustering- data classification based on characteristic similarity

PCA –to identify the variables that have the greatest impact on all brands

**SOFTWARE USED**

Data Collection - **Typeform**

Data Analysis - **R- Studio**

**TECH GADGETS BRAND ANALYSIS**

**DATASET :**

**Different brands of tech gadgets with various features are chosen for analysis and collected via typeform.**

**Sample –** 129 (Respondents)

**Gadgets names –** Smartphone, Mouse Keyboard, Headphone Camera and Laptop

**Gadget:** Smartphone

### **Source Code**

library("factoextra") – extracts and visualizes the result of Multivariate data

library("NbClust") – determine the best number of clusters

library("dplyr") – resolves the data manipulation hurdles

library("cluster") - perform the cluster analysis with the k-means algorithm

library(ggbiplot) – to visualize the PCA components in 2D

library("rstatix") – helper package of the univariate and multivariate data

library("FactoMineR") – to perform the principle component analysis

library(parameters) - Utilities for processing the parameters of various statistical models

### **Data Preprocessing**

smartphone <-

read.table(

file = "D:/Files/CIT/M.Sc.DCS/4th Semester/17MDC46 - PA Lab/PA Sem IV project/smartphone.csv",

sep = ",",

dec = ".",

header = TRUE,

)

sm = smartphone[2:11]

#preprocessing the data by replacing the NA values with zero

sm[is.na(sm)] = 0

# sm = scale(sm)

**A picture containing text

Description automatically generated**sm

The above smartphone data is preprocessed by selecting numeric columns and are scaled to normalize the data to perform the clustering analysis.

### **Optimal number of clusters and Quality of a k-means partition**

#Elbow method

fviz\_nbclust(sm, kmeans, method = "wss") +

geom\_vline(xintercept = 4, linetype = 2) +

labs(subtitle = "Elbow method")

#silhouette method

fviz\_nbclust(sm, kmeans , method = "silhouette")

#gap statistics method

gap\_stat = clusGap(

sm ,

FUN = kmeans ,

nstart = 25 ,

K.max = 10 ,

B = 50

)

fviz\_gap\_stat(gap\_stat)

# Consensus Based Algorithm

n\_clust <- n\_clusters(sm,

package = c("easystats", "NbClust", "mclust"),

standardize = TRUE,fast = TRUE,

nbclust\_method = "kmeans")

n\_clust

**Graphical user interface, text

Description automatically generated with medium confidence**

The **Elbow method** looks at the total within-cluster sum of square (WSS) as a function of the number of clusters. Here the optimal number is **4.** The Elbow method is sometimes ambiguous and an alternative is the average silhouette method.

The **Silhouette method** measures the Quality of a clustering and determines how well each point lies within its cluster. The Silhouette method suggests **7** clusters.

The optimal number of clusters is the one that maximizes the **gap statistic.** This method suggests only **9** clusters.

Here, the 3 approaches suggest a different number of clusters.

**Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated**Chart, line chart

Description automatically generatedBecause no method is clearly better, a fourth alternative is to run many methods and take the number of clusters that is the most agreed upon (i.e., find the consensus).

Based on all indices, most methods suggest to retain **3 clusters**, followed by a 4-clusters solution.

### **Chart, bar chart Description automatically generated**

### **K- Means Clustering**

# Elbow Method - 4 , silhouette = 7 , gap statistic - 9 , Consense based algorithm -3

smart\_result\_e = kmeans(sm , 4 , nstart = 25)

smart\_result\_e

fviz\_cluster(

smart\_result\_e,

data = sm,

palette = c("#2E9FDF", "#00AFBB", "#E7B800", "violet"),

geom = "point",

ellipse.type = "convex",

ggtheme = theme\_bw()

)

# silhoutte = 7

smart\_result\_s = kmeans(sm , 7)

smart\_result\_s

fviz\_cluster(

smart\_result\_s,

data = sm,

palette = c(

"#2E9FDF",

"#00AFBB",

"#E7B800" ,

"violet",

"red",

"pink",

"green"

),

geom = "point",

ellipse.type = "convex",

ggtheme = theme\_bw()

)

# gap statistic - 9

smart\_result\_g = kmeans(sm , 9)

smart\_result\_g

fviz\_cluster(

smart\_result\_g,

data = sm,

palette = c(

"#2E9FDF",

"#00AFBB",

"#E7B800" ,

"violet",

"red",

"pink",

"green",

"brown",

"yellow"

),

geom = "point",

ellipse.type = "convex",

ggtheme = theme\_bw()

)

# Final Optimal Cluster #Consense based algorithm is taken with the optimal clusters of 3

optimal\_cluster = kmeans(sm , 3 , nstart = 25)

optimal\_cluster

fviz\_cluster(

optimal\_cluster,

data = sm,

palette = c("#2E9FDF",

"#00AFBB",

"#E7B800"),

geom = "point",

ellipse.type = "convex",

ggtheme = theme\_bw()

)

#quality of k-means partition

BSS = optimal\_cluster$betweenss

TSS = optimal\_cluster$totss

# We calculate the quality of the partition

BSS / TSS \* 100

**Elbow Method**

Chart

Description automatically generated

The WSS value of cluster 4 in Elbow method is high 1907.1750 which means that there is dissimilarity in the clusters members

**Silhouette Method**

Chart

Description automatically generated

Graphical user interface, text

Description automatically generated

Graphical user interface, text

Description automatically generated

The WSS value of cluster 2 in Silhouette method is high 1907.1705 which means that there is dissimilarity in the cluster’s members

**Gap Statistics**

Chart

Description automatically generated

**A picture containing graphical user interface

Description automatically generated**The WSS value of cluster 4 in Gap statistic method is high 867.04545 which means that there is dissimilarity in the clusters members

Chart

Description automatically generated**Optimal Clusters**

The WSS value of cluster 3 in Consensus based Algorithm method is high 867.04545 which means that there is dissimilarity in the clusters members

Calendar

Description automatically generated

### **QUALITY TESTING**

**Centers= 3 (consensus based algorithm)**

**With scaling Without scaling**

Text

Description automatically generatedText

Description automatically generated

**Centers= 4**

**Text

Description automatically generatedWith scaling Without scaling**

**Text

Description automatically generated**

The classification into **four or more** groups allows for a higher explained percentage and a higher quality.

This will always be the case: with more classes, the partition will be finer, and the BSS contribution will be higher. On the other hand, the “model” will be more complex, requiring more classes. In the extreme case where k = n (each observation is a singleton class), we have BSS = TSS, but the partition has lost all interest.

This is the reason we compare partitions via their Quality only for partitions that have the same number of clusters.

So we choose the **3 clusters (without scaling)** based on the optimality using the **consensus based algorithm** The nstart() argument in the  function also allows to run the algorithm several times with different initial centers, in order to obtain a potentially better partition

### **Visualizations**

To confirm that your number of classes is indeed optimal, there is a way to evaluate the Quality of your clustering via the silhouette plot (which shows the silhouette coefficient on the y axis).

We draw the silhouette plot for 3 clusters, as suggested by the average silhouette method:

Chart, histogram

Description automatically generated

The silhouette plot above and the average silhouette coefficient say that the clustering is good and the clusters are optimal as the value is greater than zero means that the observation is well grouped. The closer the coefficient is to 1, the better the observation is grouped.

### **Principle Component Analysis**

res.pca = PCA(sm)

res.pca

Graphical user interface, text, application

Description automatically generated

Diagram

Description automatically generated

Chart, scatter chart

Description automatically generated

The PCA() function performs PCA analysis for the smartphone data and plots the necessary variables in that contributes the most in the dataset

### **Eigen Values /Variances**

#Extract and visualize the eigen values

get\_eig(res.pca)

#Visualize the eigen values/variances

fviz\_screeplot(res.pca , addlabels = TRUE, ylim = c(0, 50))

Chart, histogram

Description automatically generated

The sum of all the eigenvalues gives a total variance of 10.

**Text

Description automatically generated with low confidence**The proportion of variation explained by each eigenvalue is given in the second column .The PC1 and PC2 has 36% variability over the data

### **Contribution of the Variables**

# Extract the results from the variables

var = get\_pca\_var(res.pca)

var

#Coordinates of variables

head(var$coord)

#Contribution of variables

head(var$contrib)

sm.pca <- prcomp(sm, center = TRUE, scale. = TRUE)

summary(sm.pca)

A picture containing graphical user interface

Description automatically generated

The variables **Performance, Quality, and Design** contribute the most to the PC1 component, while **Price and Performance** contribute the most to the PC2 component. This means that the greater the contribution value, the more the variable contributes to the component.

### **Validating the variables with clustering**

# Contribution of the variables with the

res.km <- kmeans(var$coord, centers = 3, nstart = 25)

grp <- as.factor(res.km$cluster)

# Color variables by groups

fviz\_pca\_var(res.pca, col.var = grp,

palette = c("#0073C2FF", "#EFC000FF", "#868686FF"),

legend.title = "Cluster",addEllipses = TRUE ,repel = TRUE)

ggbiplot(

sm.pca,

ellipse = TRUE,

obs.scale = 1,

var.scale = 1,

var.axes = T,

groups = smartphone$Smartphone

) + theme\_minimal()

A picture containing diagram

Description automatically generated

In the above graph, the variables **Get. Used to. Habit, Brand** contributes more to the cluster 1

**Services, Reputation, Value, and Quality** contribute more to the cluster 2

**Operating Platform/System, Design, Performance, and Price** contribute more to the cluster 3

Because the above-mentioned variables contributed more to a single cluster, this does not imply that it does not contribute to other groups, but their contribution is less than that of its cluster.

Clusters are formed based on the above criteria. When comparing the clusters with the brands, the above characteristic best fits the top three brands, namely **Samsung, Xiaomi, and Apple** and we can say that customers buys smartphones from these brands based on the characteristic depicted by each of the three clusters regarding the variables.

**For example :**

The person who seeks for best design, price, and performance can buy Apple Smartphones

The person who is Brand-specific will buy Samsung smartphones

The person who strives for the best Quality goes for Xiaomi smartphones

Chart, diagram

Description automatically generatedThe above graph represents the overall brands and the variables that contributes for each of the brand on a 2 D plane.

**CONCLUSION**

The above clustering analysis creates clusters based on customers' preferences for brands, and through PCA analysis, the features that contribute the most to each brand of smartphone are identified and validated with the clusters, making it easier for consumers to select brands based on their preferences. For each brand, the output is verified and compared with survey respondents. The techniques described above can also be applied to other types of technology to determine brand categorization based on features.

* **Smartphone - Samsung**
* **Mouse - Logitech**
* **Headphone - Oneplus**
* **Camera – Canon**
* **Keyboard - Apple**
* **Laptop - HP**

**Timeline

Description automatically generated with low confidence**

**REFERENCE**

**Clustering -** [**https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/**](https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/)

[**https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/**](https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/)

**PCA -**

[**http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/**](http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/)