VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis & Modelling

A4 – Multivariate Analysis

Using R

Submitted by

MONIKA SARADHA

#V01068784

Date of Submission: 07/05/2023

Table of Contents

1. Introduction

The datasets provided, Survey.csv and icecream.csv, provide useful information about consumer preferences and characteristics in two distinct domains. These datasets allow you to investigate and analyze consumer behavior, opinions, and choices in a variety of contexts. Both datasets provide opportunities for statistical techniques such as principal component analysis (PCA), factor analysis, cluster analysis, and multidimensional scaling to be applied. We can use these techniques to uncover underlying dimensions, group respondents based on their background variables, and analyze the relationships and similarities between variables or objects in datasets.

These datasets provide a means to investigate consumer preferences, comprehend market dynamics, and inform strategic decision-making in the domains of dairy products and ice cream consumption. These datasets can be analyzed to provide valuable insights into consumer behavior, assisting businesses in developing effective marketing strategies and meeting the changing needs of their target audience.

1.1. About the Data

Survey Dataset: This dataset provides insights into the preferences and factors influencing the decision-making process of Bangalore home buyers. It includes factors such as city, gender, age, occupation, monthly household income, desired house features, budget, and influencing factors. Analyzing this dataset provides useful information about potential homebuyers' demographics, income levels, and desired characteristics. It also reveals the significance of proximity to amenities, preferences for specific features, and the impact of price, builder reputation, and neighborhood profile. This dataset is a valuable resource for researchers, real estate professionals, and policymakers interested in understanding the Bangalore housing market and making informed decisions to meet the changing needs of homebuyers.

Ice-cream Dataset: The dataset provided provides information about the attributes and ratings of various brands in the dairy products market. Brand, price, availability, taste, flavor, consistency, and shelf life are all factors to consider. We can learn about consumer perceptions and preferences for various dairy brands by analyzing this dataset. The dataset allows users to evaluate and compare the performance of various brands based on a variety of criteria, providing useful information for market research, brand positioning, and decision-making in the highly competitive dairy industry.

1.2. Objective

The goal of this analysis is to apply various statistical techniques to the provided datasets (Survey.csv and icecream.csv), such as principal component analysis (PCA), factor analysis, cluster analysis, and multidimensional scaling.

- Principal Component Analysis (PCA) and Factor Analysis: Use PCA and factor analysis to identify the underlying dimensions or latent factors in the data. Extract the main components or factors that explain the most variance in the data and interpret their meaning in relation to the variables in the datasets.
- Cluster Analysis: Use cluster analysis to categorize respondents based on their demographics. Identify distinct segments or clusters within the data using appropriate clustering algorithms, characterizing respondents with similar characteristics or preferences.
- Multidimensional Scaling: Use multidimensional scaling to determine the similarity or dissimilarity of variables or objects in a dataset. Interpret multidimensional scaling results to gain insight into the relationships and proximity of variables or objects.

This way one can uncover patterns, structures, and relationships within the datasets, which will provide valuable insights for decision-making, market segmentation, and understanding consumer preferences in the given context.

1.3. Business Significance

Survey Dataset: This dataset has significant business implications for various real estate stakeholders. Analyzing this dataset can provide valuable insights into the preferences and priorities of potential homebuyers in Bangalore for developers and builders. Understanding the desired features, budget ranges, and decision-making factors enables developers to align their construction and marketing strategies accordingly, ensuring that their projects cater to the needs and preferences of their target market. Real estate agents can use this data to provide personalized recommendations and guidance to clients, increasing customer satisfaction and conversion rates. Furthermore, policymakers and financial institutions can use this dataset to assess market trends, identify areas of potential growth, and tailor loan products and incentives to meet the population's housing needs. Overall, the dataset's commercial significance stems from its ability to inform

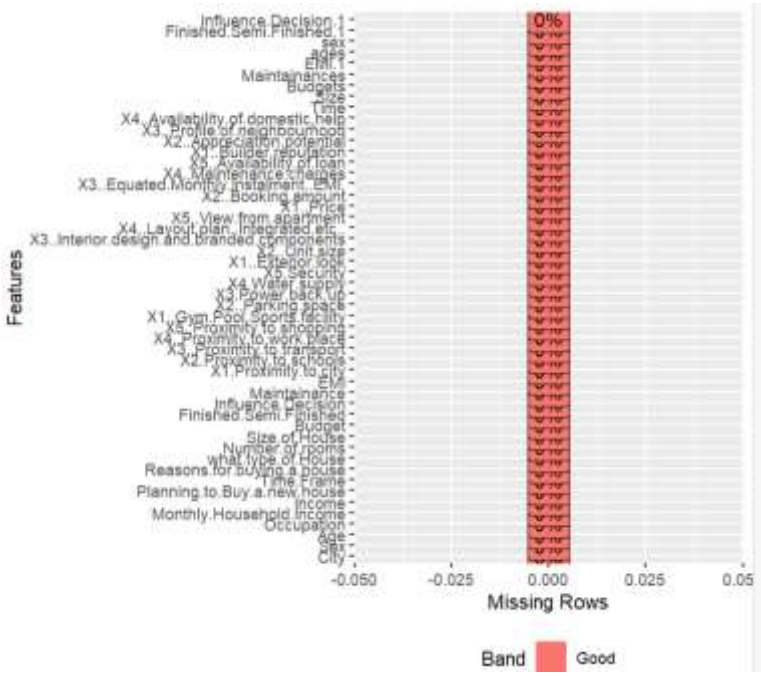
strategic decision-making, customer targeting, and market positioning in Bangalore's highly competitive real estate market.

Ice-cream Dataset: This dataset is extremely important for players in the dairy products industry. Companies can gain insights into consumer perceptions and preferences by analysing the attributes and ratings of various brands. Businesses can make informed decisions about product development, pricing strategies, and marketing campaigns by understanding the factors that influence consumer decisions, such as price, taste, flavour, consistency, and shelf life. By aligning their offerings with consumer preferences, companies can identify areas for improvement and develop competitive advantages. This dataset can also be used by market research teams to identify market trends, assess brand positioning, and make data-driven decisions to increase their market share. Overall, the commercial significance of this dataset stems from its ability to inform brand management strategies, product innovation, and marketing efforts in the competitive dairy products market.

2. Results

2.1. Data Preprocessing

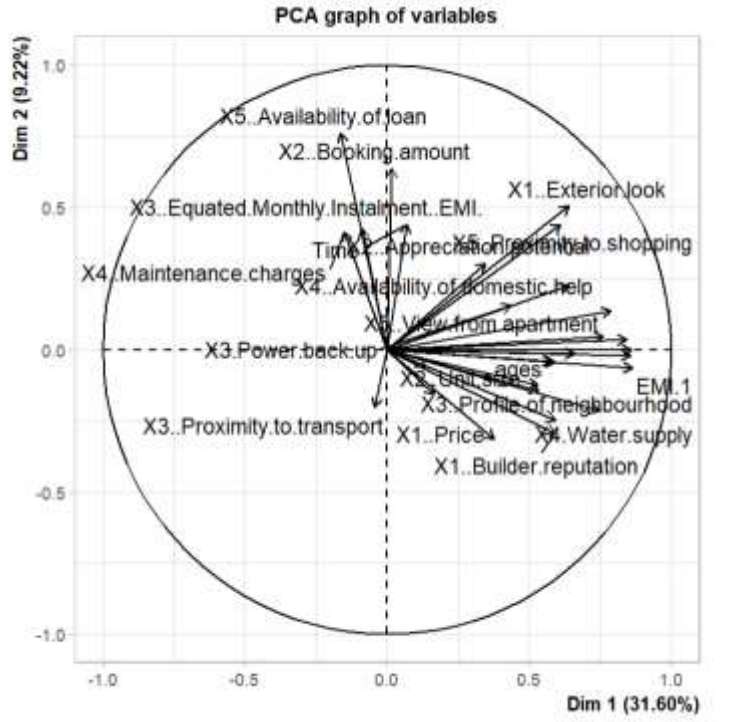
Missing Value Plot:

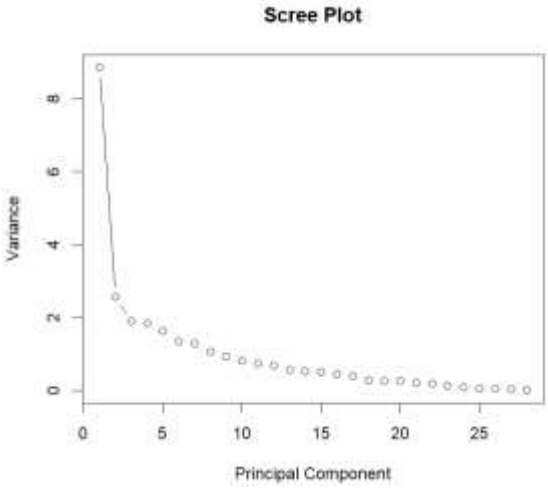


Inference:

The dataset was looked in for missing values and had null. Hence no processing steps was required.

2.2. Principal Component Analysis





Inference:

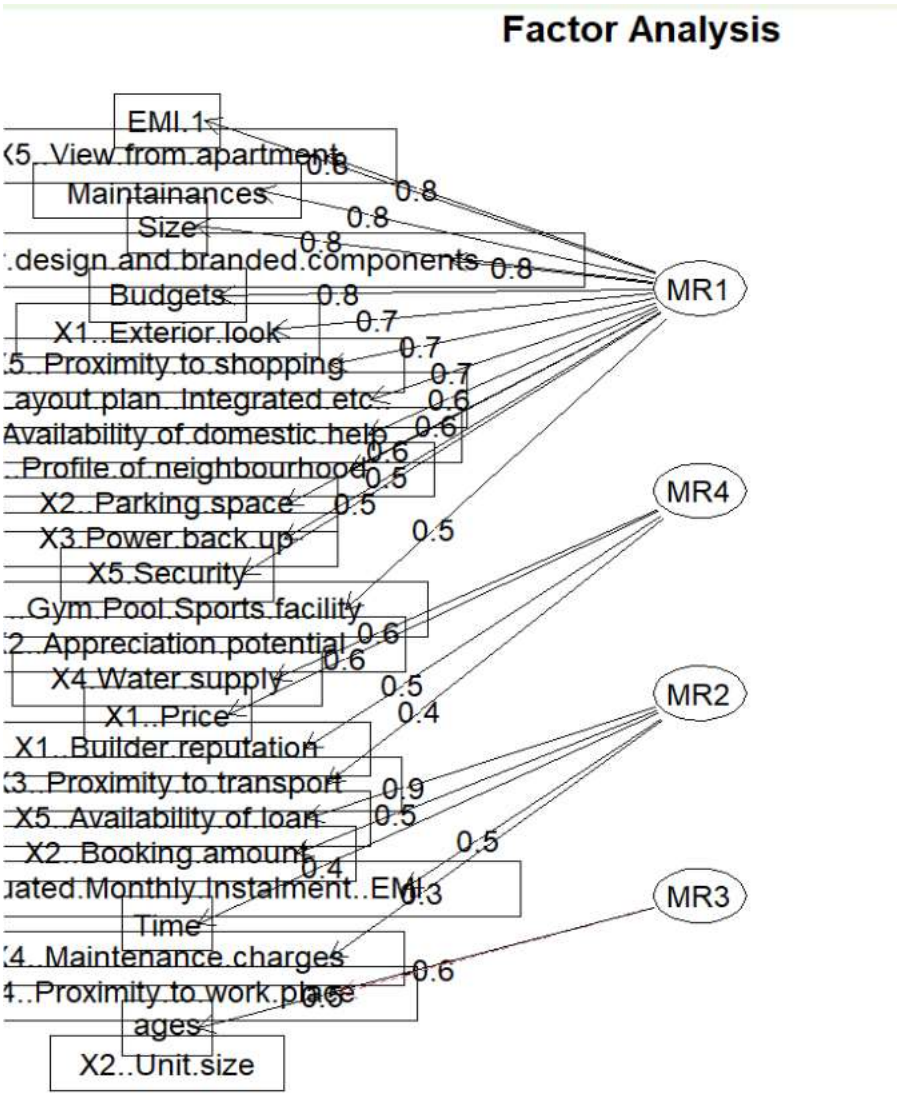
The dimensions in the dataset are represented by the following components based on the obtained Principal Component Analysis (PCA) output: Dim.1, Dim.2, Dim.3, and so on. These dimensions

are derived from the dataset's variable analysis and represent orthogonal linear combinations of the original variables. Each dimension captures a different pattern or structure in the data and represents a different amount of variance. Dim.1, for example, explains 31.6% of the variance in the data, Dim.2, 9.221% of the variance, Dim.3, 6.801% of the variance, and so on.

The identified dimensions in the data are labelled as follows based on the PCA output, these dimensions are identified based on the loadings of the original variables on each dimension, indicating the variables that contribute most to the variation in each dimension.

- Dimension 1 (Dim.1): This dimension captures the variance in variables related to proximity to shopping, gym/pool/sports facilities, parking space, exterior look, and unit size. It represents factors related to amenities and physical features of the property.
- Dimension 2 (Dim.2): This dimension is associated with variables related to proximity to transport, proximity to work, and security. It captures factors related to convenience and safety.
- Dimension 3 (Dim.3): This dimension primarily represents the variance in variables related to proximity to work and exterior look. It captures factors related to commuting convenience and the aesthetic appeal of the property.
- Dimension 4 (Dim.4): This dimension is primarily associated with variables related to water supply, power backup, and availability of domestic help. It represents factors related to basic amenities and infrastructure.
- Dimension 5 (Dim.5): This dimension captures the variance in variables related to security, interior design, and layout plan. It represents factors related to the overall quality and design of the property.
- Dimension 6 (Dim.6): This dimension is primarily associated with variables related to proximity to shopping and power backup. It captures factors related to convenience and reliability.
- Dimension 7 (Dim.7): This dimension represents the variance in variables related to water supply and availability of loan. It captures factors related to financial aspects and utility services.

2.3. Factor Analysis



Inference:

Based on the loadings obtained, we can infer these dimensions:

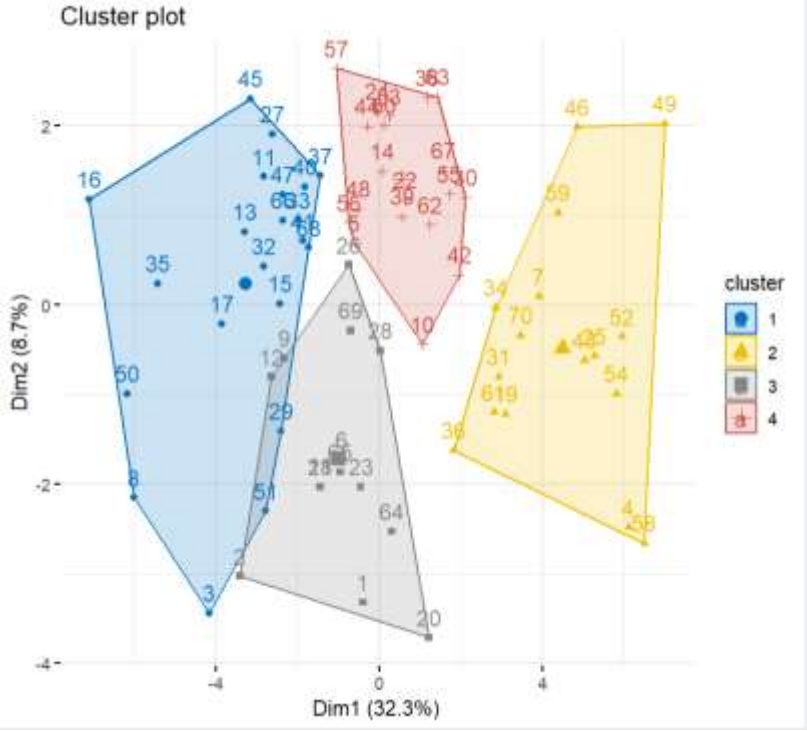
- Dimension 1 (MR1): This dimension is characterized by high loadings for variables such as proximity to shopping (0.680), gym/pool/sports facility (0.459), parking space (0.584), power backup (0.528), water supply (0.399), exterior look (0.729), interior design and branded components (0.753), layout plan (0.660), view from apartment (0.788), price (0.190), builder reputation (0.397), and appreciation potential (0.276). It represents factors related to the quality and amenities of the property.

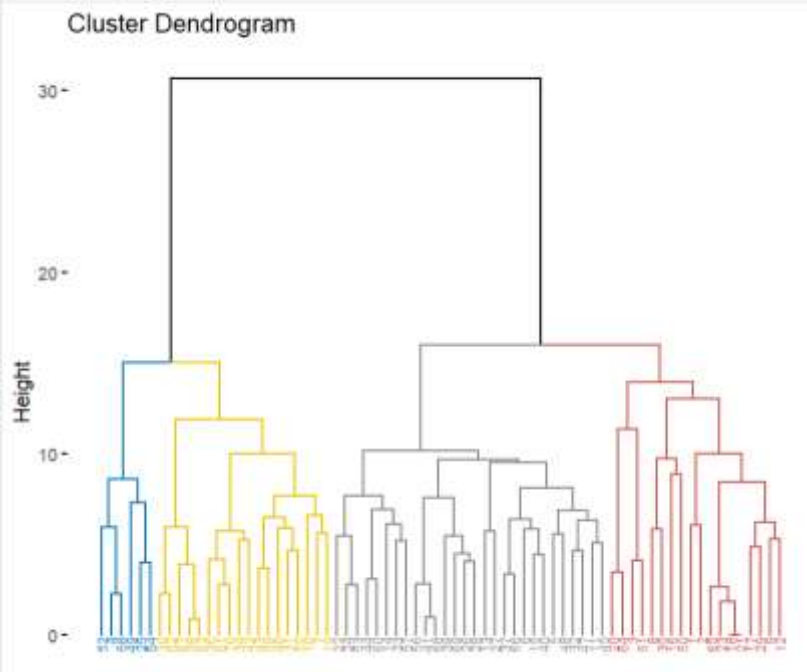
- Dimension 2 (MR4): This dimension has high loadings for variables related to proximity to work (0.210), availability of domestic help (0.624), and booking amount (0.527). It represents factors related to the convenience of commuting and household support services.
- Dimension 3 (MR2): This dimension is primarily associated with variables related to proximity to transport (-0.191), security (0.120), and availability of loan (0.873). It represents factors related to transportation convenience, security measures, and financing options.
- Dimension 4 (MR3): This dimension is characterized by high loadings for variables related to proximity to work place (-0.588), interior design and branded components (0.266), layout plan (-0.155), view from apartment (0.183), and appreciation potential (0.271). It represents factors related to the location and layout of the property, as well as its potential for future value appreciation.

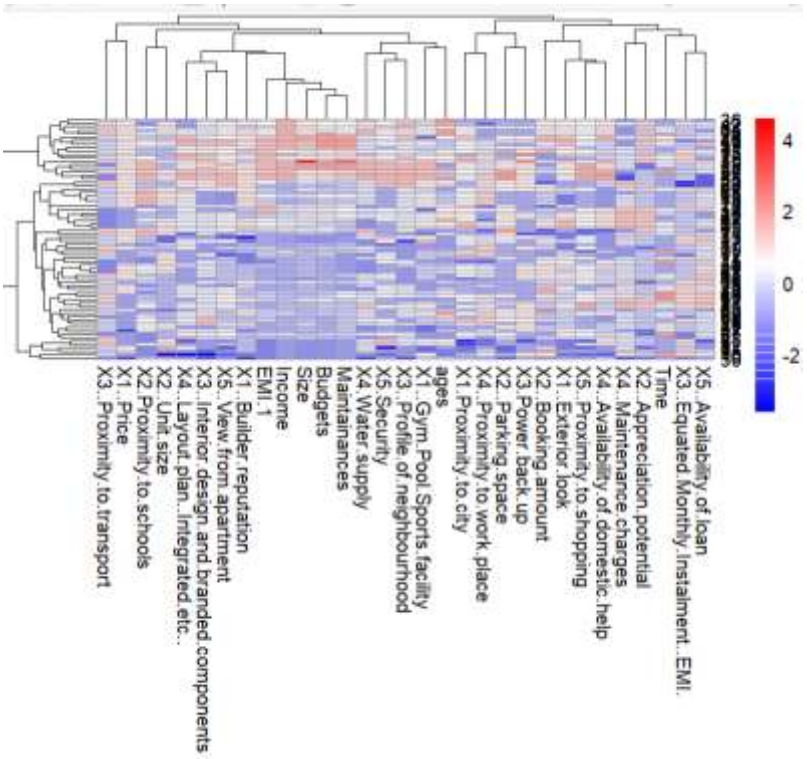
The communalities represent the proportion of variance in each variable that is accounted for by the common factors in the factor analysis. The scores represent the factor scores for each individual or observation in the dataset. These scores indicate the position of each individual along the extracted factors.

Overall, the extracted dimensions from PCA and FA are similar in terms of the underlying constructs they represent. Both methods emphasize similar themes, such as location-related factors, amenities, financial aspects, and property characteristics. However, due to differences in their underlying principles and assumptions, the specific variables and their loadings may differ slightly between the two methods.

2.4. Cluster Analysis







Inference:

The gap statistic method was used to determine the optimal number of clusters for the given dataset. With a seed of 123, the k-means algorithm was applied to cluster the data into 4 distinct groups. The cluster visualization using fviz_cluster and fviz_dend revealed meaningful patterns and structures within the data. Additionally, a heatmap was generated to visualize the scaled variables in the dataset with a color gradient representing their values.

Based on the provided cluster summary, we can characterize the respondents based on their background variables as follows:

Cluster 1:

- Proximity to city, schools, and transportation: Moderate
- Proximity to work place: Moderate
- Proximity to shopping: Low
- Amenities and facilities (gym/pool/sports facility, parking space, power back-up, water supply, security): Moderate
- Exterior look and unit size: Low

- Interior design and branded components, layout plan and integration: Moderate
- View from apartment: Low
- Price, booking amount, EMI, and maintenance charges: Moderate
- Loan availability: Moderate
- Builder reputation and appreciation potential: Moderate
- Profile of neighborhood and availability of domestic help: Moderate
- Time spent: Moderate
- Size, budget, maintenance, EMI, and age: Moderate

Cluster 2:

- Proximity to city, schools, transportation, and shopping: High
- Amenities and facilities (gym/pool/sports facility, parking space, power back-up, water supply, security): High
- Exterior look and unit size: High
- Interior design and branded components, layout plan and integration: High
- View from apartment: High
- Price, booking amount, EMI, and maintenance charges: High
- Loan availability: Moderate
- Builder reputation and appreciation potential: High
- Profile of neighborhood and availability of domestic help: High
- Time spent: High
- Size, budget, maintenance, EMI, and age: High

Cluster 3:

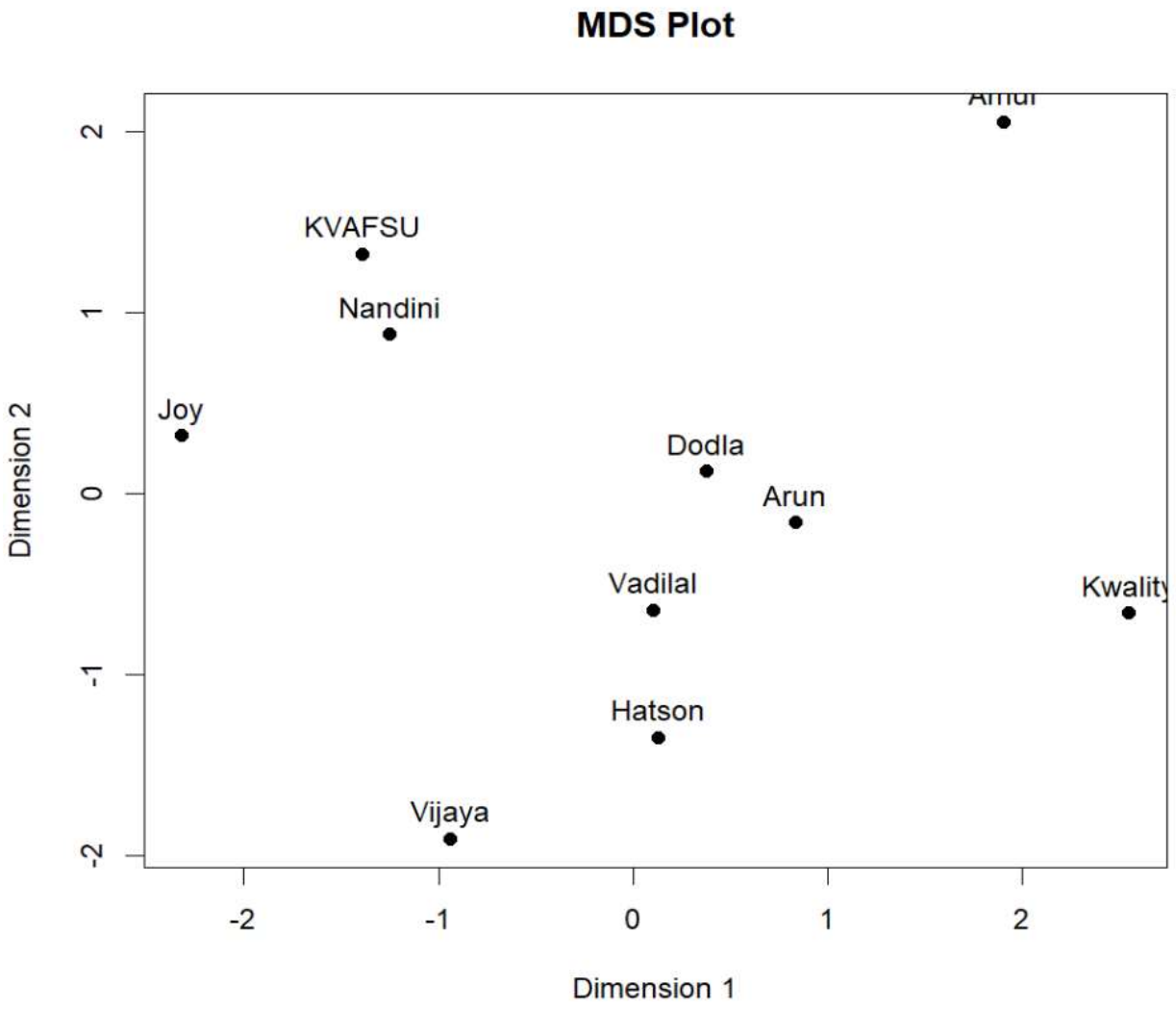
- Proximity to city, schools, transportation, and work place: Moderate to High
- Proximity to shopping: Low
- Amenities and facilities (gym/pool/sports facility, parking space, power back-up, water supply, security): Moderate
- Exterior look: Low
- Unit size: Moderate
- Interior design and branded components, layout plan and integration: Moderate

- View from apartment: Moderate
- Price, booking amount, EMI, and maintenance charges: Moderate
- Loan availability: Moderate
- Builder reputation and appreciation potential: Moderate
- Profile of neighborhood: Moderate
- Availability of domestic help: Moderate to Low
- Time spent: Moderate
- Size, budget, maintenance, EMI, and age: Moderate

Cluster 4:

- Proximity to city: High
- Proximity to schools: Moderate
- Proximity to transportation, work place, and shopping: Moderate to High
- Amenities and facilities (gym/pool/sports facility, parking space, power back-up, water supply, security): Moderate
- Exterior look and unit size: High
- Interior design and branded components, layout plan and integration: High
- View from apartment: High
- Price, booking amount, EMI, and maintenance charges: High
- Loan availability: High
- Builder reputation and appreciation potential: High
- Profile of neighborhood: High
- Availability of domestic help: Moderate
- Time spent: High
- Size, budget, maintenance, EMI, and age: High

2.5. Multidimensional Scaling



Inference:

Multidimensional scaling (MDS) is a technique used to visualize the similarity or dissimilarity between objects or cases based on a set of variables. By reducing the dimensionality of the data, MDS allows us to represent complex relationships in a lower-dimensional space, typically two or three dimensions. In the provided dataset, the MDS analysis has been performed, resulting in two dimensions: Dimension 1 and Dimension 2. Each item in the dataset is represented by its coordinates on these dimensions.

Based on the analysis, the variable "Taste" is strongly associated with Dimension 1, while the variable "Shelf life" is strongly associated with Dimension 2. This suggests that Dimension 1 captures the variability in taste preferences among the different ice cream brands, while Dimension 2 captures the variability in shelf-life ratings.

Items that are closer to each other in the plot are more similar based on the variables used in the analysis, while items that are farther apart are more dissimilar.

- KVAFSU and Nandini are closely associated to Dimension 2 and similar to each other in terms of variables employed mostly with shelf-life.
- Hatson, Vadilal, Dodla and Arun are similarly associated with dimension 1 which is taste related variables.
- Arun is associated positively with both the dimensions of taste and shelf life.

3. Recommendation

3.1. Business Implications

- **Taste Differentiation:** According to the MDS analysis, taste is a significant factor in consumers' perceptions of ice cream brands. To differentiate themselves in the market, businesses should focus on developing unique and appealing flavours. This can be accomplished through continuous product innovation, consumer preference research, and feedback gathering to create flavors that cater to a wide range of tastes.
- **Shelf-Life Optimization:** The analysis emphasizes the importance of shelf life in the decision-making process of consumers. Brands with longer shelf lives, such as KVAFSU and Nandini, have an advantage. To extend the shelf life of their products, businesses must prioritize quality control measures, packaging innovation, and distribution strategies. Marketing efforts that communicate the longer shelf life to consumers can help build trust and loyalty.
- **Competitive Positioning:** The MDS plot provides information about the market positioning of ice cream brands based on taste and shelf life. Businesses can use this data to evaluate their competitive position in comparison to other brands and identify areas for improvement. Companies can improve their brand positioning strategies to attract target customers and gain a competitive advantage by leveraging their strengths and addressing any weaknesses.

3.2. Business Recommendations

- **Product Development:** Based on the MDS analysis, businesses should invest in R&D to introduce new and innovative flavors that cater to a variety of taste preferences. Consumer surveys, taste tests, and market trends can all provide useful information for flavor development. Updating the product portfolio on a regular basis with new and exciting flavors can help to attract a larger customer base and boost brand loyalty.
- **Quality Assurance and Packaging:** In order to increase shelf life and maintain product freshness, businesses should prioritize the implementation of robust quality assurance processes. This includes proper storage, temperature control monitoring, and optimizing

packaging materials. Investing in advanced packaging technologies that maintain product quality and extend shelf life can boost brand reputation and consumer satisfaction.

- **Marketing and communication:** Businesses should incorporate the MDS analysis findings into their marketing and communication strategies. To distinguish the brand from competitors, emphasize the distinct taste profiles and flavor varieties. Highlight the longer shelf life as a quality assurance measure that provides consumers with convenience. Engage with target audiences and raise brand awareness through various marketing channels such as social media, influencer collaborations, and interactive campaigns.
- **Customer Feedback and Continuous Improvement:** Seek customer feedback on a regular basis to better understand their preferences and gain insights into taste preferences and shelf-life expectations. To gather valuable feedback, conduct surveys, engage in social listening, and encourage customer reviews. This data can be used to drive product improvements, address quality concerns, and continuously improve the overall customer experience.

Businesses in the ice cream industry can improve their market positioning, meet customer expectations, and drive growth by offering appealing flavors, ensuring product quality, and differentiating themselves based on taste and shelf life by implementing these recommendations.

4. Codes - R Studio:

```
survey<-(read.csv("C:\\Users\\monis\\OneDrive\\Desktop\\SCM\\Survey.csv"))
getwd()
library(dplyr)
library(tidyr)
library(readr)
library(GPArotation)
library(FactoMineR)
library(factoextra)
library(DataExplorer)
library(psych)
dim(survey)
names(survey)
is.na(survey)
sum(is.na(survey))
plot_missing(survey)
summary(survey)

#a) Do principal component analysis and factor analysis and identify the dimensions in the data,
data=survey[,c(20:50)]

#Principal Component Analysis

pca_result1 <- PCA(numericdata, scale = TRUE)
summary(pca_result1)
fviz_pca_biplot(pca_result, label = "var", repel = TRUE)

#Factor Analysis
factor_analysis <- fa(numericdata, nfactors = 4, rotate = "varimax")
print(fac_analysis$loadings, reorder = TRUE)
```

```
print(factor_analysis$communalities)

print(factor_analysis$scores)

fa.diagram(fac_analysis)


#b) Carry our cluster analysis and characterize the respondents based on their background variables.

#Cluster analysis

install.packages("cluster")

install.packages("factoextra")

library(cluster)

library(factoextra)

library(pheatmap)

library(dplyr)



survey_cleaned <- na.omit(survey)

numeric_survey <- select_if(survey_cleaned, is.numeric)

data2 <- scale(numeric_survey)

set.seed(123)

gap_stat <- fviz_nbclust(data2, FUNcluster = kmeans, method = "gap_stat")

set.seed(123)

kmeans_res <- kmeans(data2, 4, nstart = 25)

fviz_cluster(kmeans_res, data = data2, palette = "jco", ggtheme = theme_minimal())

hclust_res <- hclust(dist(data2), method = "ward.D2")

fviz_dend(hclust_res, cex = 0.5, k = 4, palette = "jco")

pheatmap(data2, color = colorRampPalette(c("blue", "white", "red"))(100))


numeric_logical_survey <- survey_clustered[, sapply(survey_clustered, is.numeric) | sapply(survey_clustered, is.logical)]


cluster_summary <- aggregate(numeric_logical_survey[, -1], by = list(Cluster = survey_clustered$Cluster), FUN = function(x) mean(x, na.rm = TRUE))
```

```
print(cluster_summary)

#-----------------------------------------------------------------------

#c) Do multidimensional scaling and interpret the results.

ic<-read.csv("C:\\Users\\monis\\OneDrive\\Desktop\\SCM\\icecream.csv")
ic
ds <- subset(ic, select = -c(Brand))
dist_matrix <- dist(ds)
mds <- cmdscale(dist_matrix, k = 2)

plot(mds[, 1], mds[, 2], pch = 16, xlab = "Dimension 1", ylab = "Dimension 2", main = "MDS Plot")
text(mds[, 1], mds[, 2], labels = ic$Brand, pos = 3)

mds_val <- data.frame(Item = rownames(mds), Dimension1 = mds [, 1], Dimension2 = mds [, 2])
print(mds_val)


ds <- subset(ic, select = -c(Brand))
dist_matrix <- dist(ds)
mds <- cmdscale(dist_matrix, k = 2)


correlations <- cor(ds, mds)
dim1_vars <- colnames(ds)[which.max(abs(correlations[, 1]))]
dim2_vars <- colnames(ds)[which.max(abs(correlations[, 2]))]
cat("Variables associated with Dimension 1:", dim1_vars, "\n")
cat("Variables associated with Dimension 2:", dim2_vars, "\n")
```