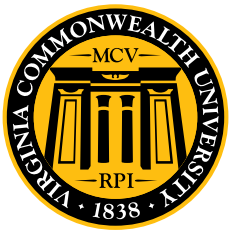



# VIRGINIA COMMONWEALTH UNIVERSITY

## Statistical analysis and modelling (SCMA 632)

### A4: Multivariate Analysis and Business Analytics Applications

**Manoranjan Mohankumar**

**V01107254**

**Date of Submission: 08-07-2024**

**CONTENTS**

## Introduction

## PCA, Factor analysis and Cluster analysis

This assignment focuses on the application of multivariate analysis techniques to understand the underlying structure and patterns in a survey dataset. The dataset consists of responses from 70 individuals regarding their preferences and considerations when planning to buy a house. The objective is to utilize Principal Component Analysis (PCA), Factor Analysis, and Cluster Analysis to extract meaningful insights from the data.

**Principal Component Analysis (PCA)** is employed to reduce the dimensionality of the dataset while retaining as much variability as possible. By transforming the original variables into a new set of uncorrelated variables called principal components, PCA helps in identifying the most significant factors that influence the decision-making process of potential homebuyers.

**Factor Analysis** is conducted to uncover latent variables or factors that explain the patterns of correlations among the observed variables. By focusing on the relationships between variables, factor analysis aids in simplifying the dataset by grouping variables with similar characteristics, thereby providing a clearer understanding of the underlying factors affecting home purchase decisions.

**Cluster Analysis** is performed to segment the respondents into distinct groups based on their preferences and priorities. By applying k-means clustering and hierarchical clustering techniques, the analysis aims to identify homogeneous subgroups within the data, allowing for targeted marketing strategies and personalized recommendations for different segments of potential buyers.

### Multidimensional Scaling Analysis

In this analysis, we aim to explore the similarities and differences between various ice cream brands based on their attributes using Multidimensional Scaling (MDS). MDS is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space, typically 2D or 3D, while preserving the distances between data points. This allows us to visualize complex data and identify patterns, clusters, and outliers.

- **Data Loading**: The dataset was loaded from the file `icecream.csv`, which contains multiple attributes for different ice cream brands.
- **Extracting Numeric Data**: All columns except the first one (which contains the brand names) were extracted as numeric features for analysis.
- **Scaling the Data**: The numeric data was scaled using standard scaling techniques to ensure that all features contribute equally to the distance calculations. Scaling transforms the data to have a mean of 0 and a standard deviation of 1.

## Conjoint analysis

Is a powerful statistical technique used in market research to understand consumer preferences and the value they place on various product attributes. This method helps businesses design products that better meet consumer needs by quantifying the importance of different features. In this report, we apply conjoint analysis to a dataset of pizza brands to explore how various attributes, such as brand and price, influence consumer preferences.

## Data Overview

The dataset used for this analysis, `pizza_data.csv`, contains information about different pizza brands, their price categories, and the rankings provided by respondents. The key columns in the dataset are:

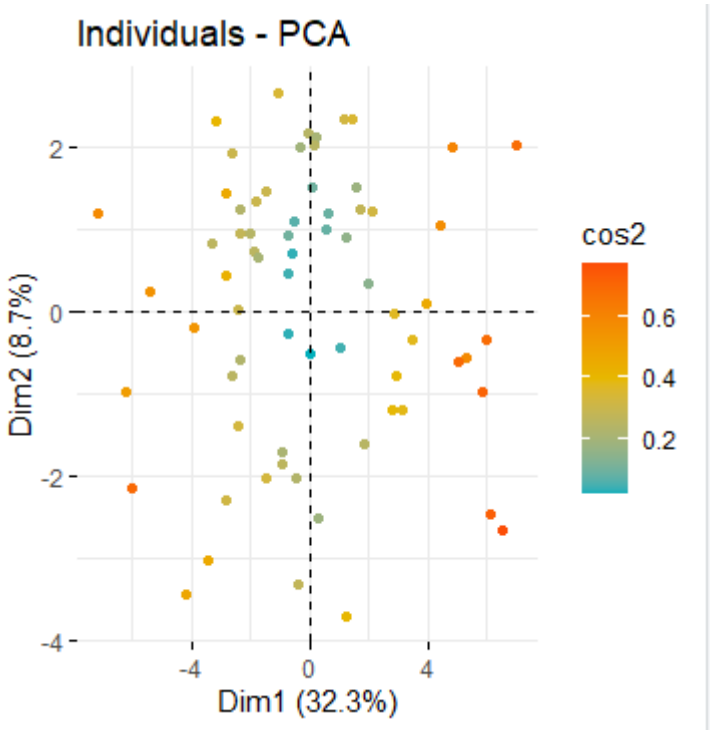
- **brand**: The name of the pizza brand.
- **price**: The price category of the pizza.
- **ranking**: The ranking or rating given by respondents, representing their preferences.

The goal of this analysis is to determine how much each attribute (brand and price) and their levels influence consumer preferences, as reflected in the rankings.

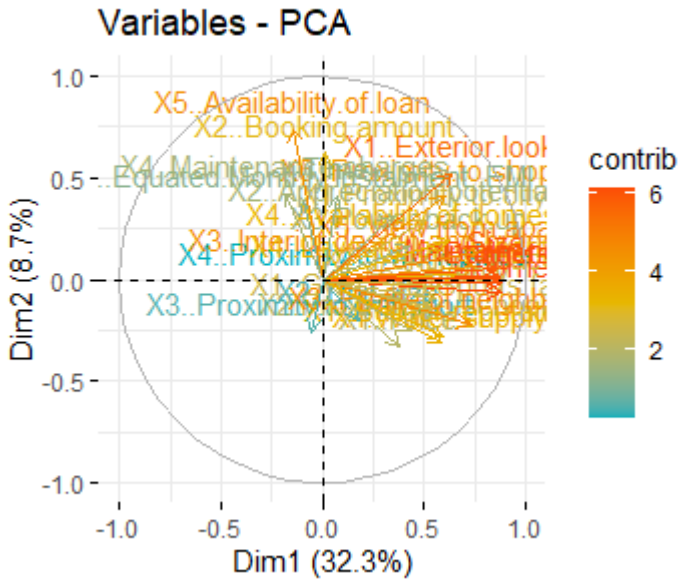
**Results & Interpretations**

**PCA Analysis:** The PCA was conducted to reduce the dimensionality of the dataset and identify the key components that explain the variance in the data.

1. **Summary of PCA Results:**
    - The first principal component (PC1) explains 32.3% of the total variance, while the second principal component (PC2) explains 8.7%.
    - Together, the first two components account for approximately 41% of the total variance.
2. **PCA Individuals Plot:**
    - This plot shows how individuals (respondents) are distributed in the reduced dimensional space.
    - The color gradient represents the quality of representation (cos2 values) of individuals on the principal components. Higher cos2 values indicate better representation.

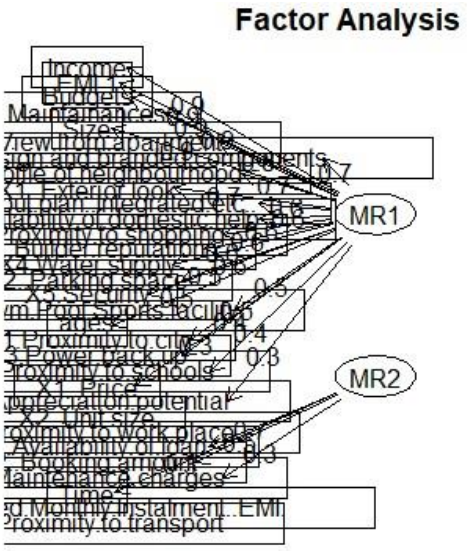



3. **PCA Variables Plot:**
    - This plot visualizes the contributions of the variables to the principal components.
    - Variables like "X5. Availability of loan" and "X2. Booking amount" contribute significantly to PC1, indicating that these factors are crucial in explaining the variance in the data.
    - Variables such as "X1. Exterior look" and "X3. Proximity to transport" are more associated with PC2, suggesting their importance in the second dimension.

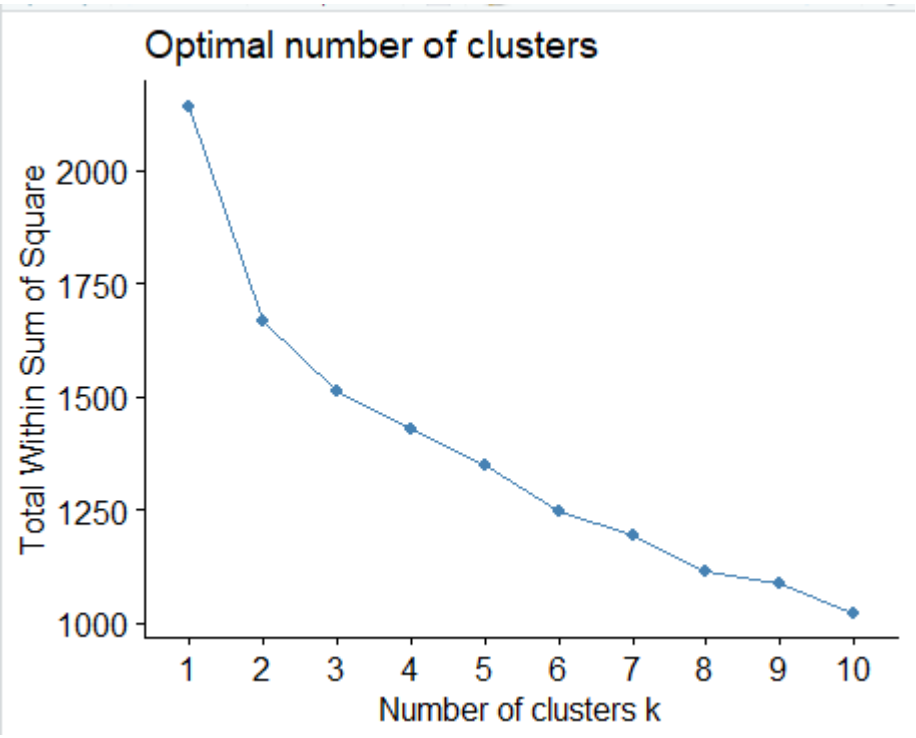



**Factor Analysis:** Factor analysis was used to identify underlying factors that explain the observed correlations among the variables.

1. **Factor Loadings:**
   - The factor analysis revealed two main factors (MR1 and MR2).
   - Variables like "Income," "Size," "Budgets," and "Maintainances" have high loadings on MR1, indicating that this factor represents financial and size-related considerations.
   - Variables such as "X5. Availability of loan" and "X2. Booking amount" load significantly on MR2, suggesting that this factor relates to loan availability and booking aspects.

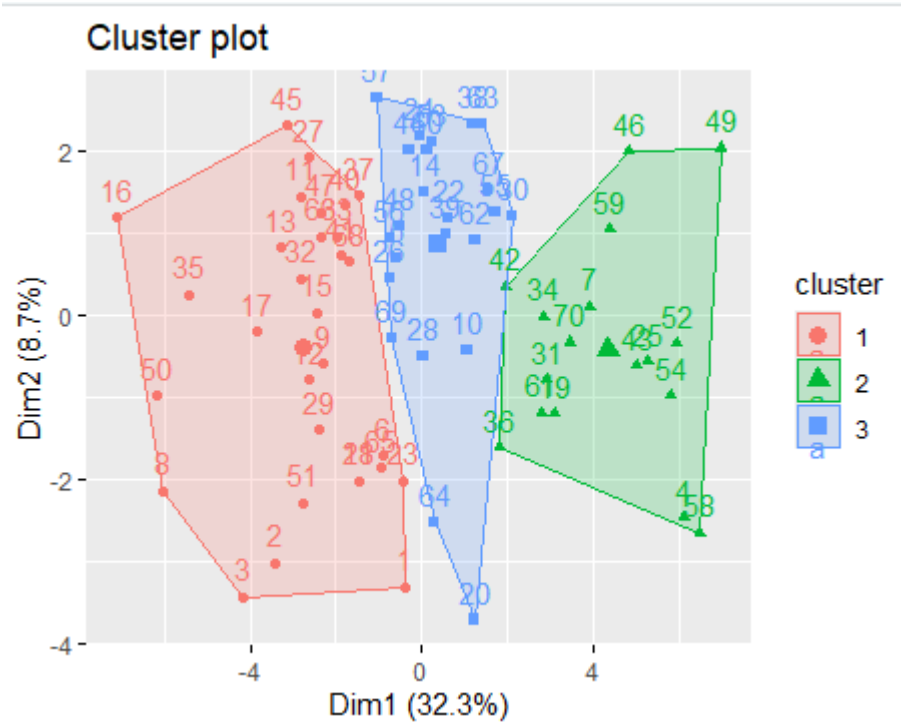


2. **Factor Diagram:**
    - The diagram visualizes the loadings of each variable on the two factors.
    - This helps in understanding which variables are most influential in defining each factor.




**Cluster Analysis:** Cluster analysis was performed to segment the respondents into distinct groups based on their preferences and priorities.

1. **Optimal Number of Clusters:**
    - The "elbow" method (WSS plot) and the silhouette method were used to determine the optimal number of clusters.
    - Both methods indicated that three clusters provide a good balance between simplicity and explanatory power.




2. **k-means Clustering:**
    - The k-means clustering algorithm was applied with three clusters.
    - The resulting clusters have distinct characteristics:
        - Cluster 1: Respondents with lower financial capabilities and preferences for basic amenities.
        - Cluster 2: Individuals with higher incomes and preferences for premium features.
        - Cluster 3: Respondents showing moderate preferences for various factors.
3. **Cluster Plot:**
    - The cluster plot visualizes the distribution of respondents in the three clusters, with different colors and shapes representing different clusters.
4. **Hierarchical Clustering:**
    - Hierarchical clustering was also performed to confirm the k-means results.
    - The dendrogram supports the presence of three distinct clusters, as indicated by the red rectangles.

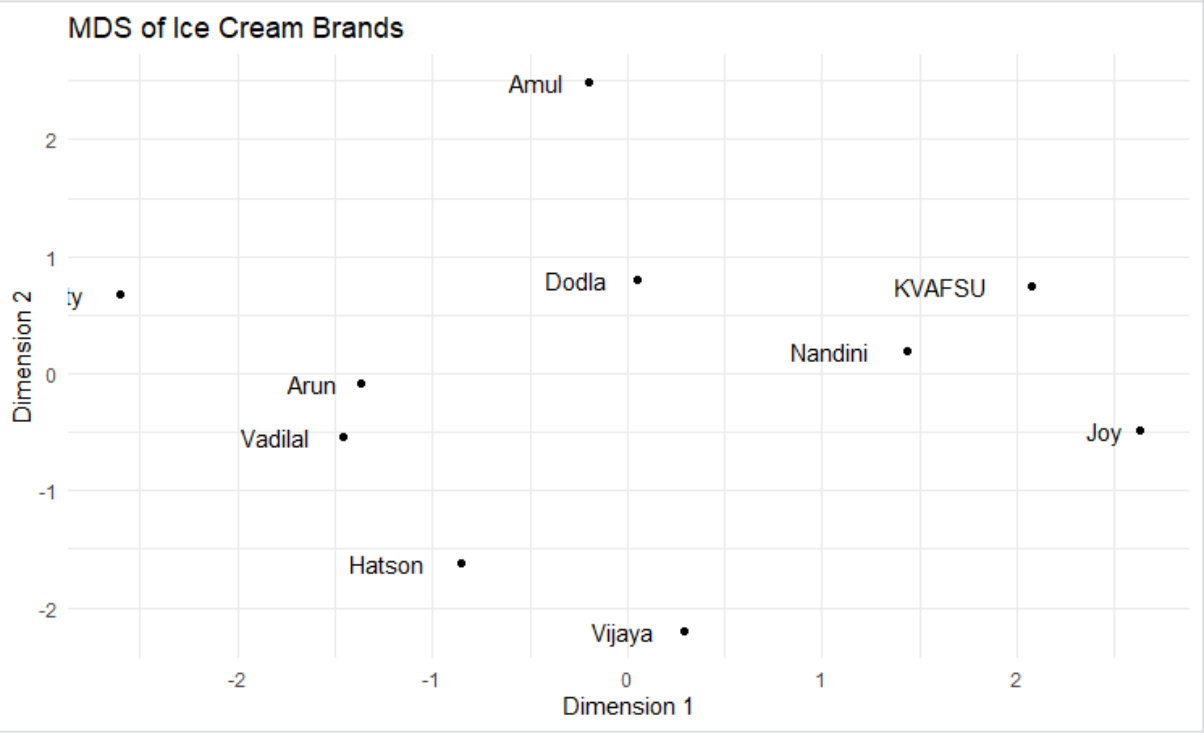
<u>Multidimensional Scaling Analysis</u>

**Multidimensional Scaling (MDS)**:

- MDS was applied to the scaled data to reduce its dimensions to two. This allows us to visualize the data in a 2D space.

- The resulting coordinates for each ice cream brand were plotted, with each point representing a brand. The scatter plot displays the positions of the ice cream brands in the 2D space.
- **Clusters**: Brands that are positioned close to each other on the plot have similar characteristics. These clusters can indicate groups of brands with similar profiles.

- **Outliers**: Brands that are isolated from others can be considered outliers, indicating unique characteristics or significant differences from other brands.
- **Brand Positioning**: The relative positions of the brands provide insights into their competitive landscape, helping to identify direct competitors and market positioning.




**Conjoint Analysis**

The bar plot of part-worth utilities provides a visual representation of the estimated utilities for each attribute level. Here are the detailed insights:

1. **Constant Term (`const`)**:
   - The constant term has the highest positive utility estimate, indicating a strong baseline preference that is not attributed to any specific brand or price.
2. **Brand Attributes**:
   - **brand_Onesta**: This brand has a slightly positive utility, indicating a small but positive consumer preference.
   - **brand_oven story**: Similar to Onesta, Oven Story also has a slightly positive utility, suggesting a modest preference.
   - **brand_pizza hut**: Pizza Hut stands out with a higher positive utility, indicating it is the most preferred brand among the three evaluated.
3. **Price Attributes**:
   - **price_$2.00**: This price category has a high positive utility, showing that consumers strongly prefer the $2.00 price point.
   - **price_$3.00**: The utility for this price category is negative, indicating a lower preference compared to the $2.00 price point.
   - **price_$4.00**: Similar to the $3.00 price category, the $4.00 price point also has a negative utility, further indicating reduced preference as the price increases.

## Recommendation (Business Insights or plans)

## PCA, Factor and Clustering

1. **Segment-Specific Marketing Strategies**
2. Product Development and Amenities
3. **Enhanced Customer Experience**
4. Targeted Communication and Advertising
5. Monitoring and Feedback

## Multidimensional Scaling (MDS):

1. **Clusters and Similar Brands**
2. Unique Brands
3. Brand Positioning
4. Targeted Marketing
5. Product Differentiation
6. Competitor Benchmarking

## Conjoint Analysis

1. **Pricing Strategy**:
2. Product Development
3. Brand Preferences
4. Competitive Positioning
5. Price Sensitivity

**References:**

- Survey.csv dataset
- icecream.csv dataset
- pizza_data.csv dataset
- ChatGPT