****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modeling (SCMA 632)**

**A4: Multivariate Analysis and Business Analytics applications**

**SAMPREETH S SHETTY**

**V01107634**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objectives | **1** |
| **3.** | Business significance | **1-2** |
| **4.** | Results and Interpretations (Python) | **3-15** |
| **5.** | Results and Interpretations (R) | **16-23** |

**Introduction**

The provided dataset, "icecream.csv," contains information about ice cream sales, capturing attributes such as flavors, sales figures, geographic locations, and periods. This dataset aims to help understand consumer preferences, sales trends, and market dynamics within the ice cream industry. By analyzing this data, businesses can gain valuable insights into popular products, identify seasonal trends, and optimize marketing strategies to boost sales and customer satisfaction.

The second dataset, "Survey.csv," includes survey responses, capturing demographic information, consumer opinions, and preferences. This data is invaluable for businesses looking to understand their customer base, gather feedback on products or services, and make data-driven decisions to enhance customer experience and drive business growth.

Principal Component Analysis (PCA) is a dimensionality reduction technique that simplifies large datasets by transforming them into a smaller set of uncorrelated variables called principal components. Factor Analysis is a statistical method used to identify underlying relationships between variables by grouping them into factors, assuming that observed variables are influenced by a few underlying unobserved variables (factors). Conjoint Analysis is a survey-based statistical technique used in market research to determine how people value different attributes that make up a product or service.

**Objectives**

a. Perform Principal Component Analysis and Factor Analysis to identify data dimensions.

b. Conduct Cluster Analysis to characterize respondents based on background variables.

c. Apply Multidimensional Scaling and interpret the results.

d. Conjoint Analysis

**Business Significance**

Principal Component Analysis (PCA) reduces the dimensionality of data while preserving as much variability as possible by identifying the directions (principal components) along which the data's variance is maximized. This process produces fewer uncorrelated variables (principal components) that represent the most significant trends in the data. In contrast, Factor Analysis is a technique used to uncover underlying relationships between measured variables by identifying and modeling the factors that explain the pattern of correlations within a set of observed variables.

In business, PCA and Factor Analysis are invaluable for simplifying complex datasets and uncovering hidden patterns and relationships that inform strategic decision-making. These techniques enable companies to:

**Reduce Complexity**: Simplify large datasets by reducing the number of variables to a more manageable set without losing significant information.

**Identify** **Key** **Drivers**: Discover the main factors that drive consumer behaviour, product preferences, or market trends.

**Enhance** **Predictive** **Models**: Improve the performance of predictive models by focusing on the most influential variables.

**Inform** **Strategy**: Aid in developing targeted marketing strategies, product development, and customer segmentation.

Conjoint Analysis helps understand consumer preferences and the trade-offs they are willing to make between different product features. Respondents are presented with a set of products or services described by varying levels of attributes and are asked to choose or rank them. The analysis then decomposes their preferences to estimate each attribute level's value (utility). This provides quantitative measures of the value consumers place on each attribute and the optimal combination of features for a product.

Conjoint Analysis is crucial for businesses as it provides detailed insights into consumer preferences, which can be directly translated into actionable strategies for product development and marketing. By revealing which features are most valued by consumers and the trade-offs they are willing to make, businesses can:

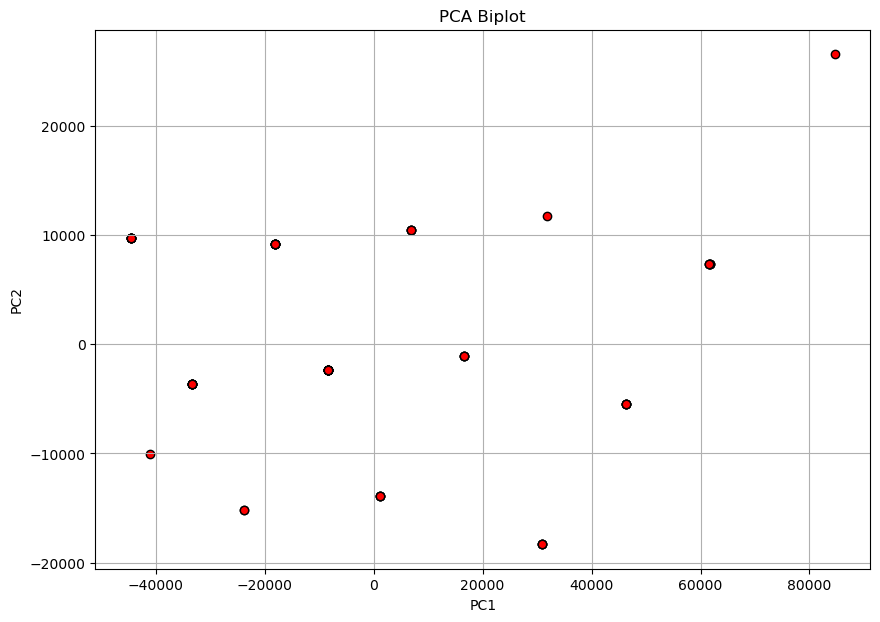
**Optimize** **Product** **Design**: Create products that better meet consumer needs and preferences.

**Enhance** **Pricing** **Strategy**: Determine the optimal price points and feature bundles.

**Improve** **Market** **Segmentation**: Identify consumer segments based on their preferences and tailor marketing efforts accordingly.

**Results and Interpretation (Python)**

1. **Principle component analysis**



The PCA biplot visually represents the data in the space defined by the first two principal components (PC1 and PC2). Here's a summary of the interpretation:

**Observations**: The points represent the data points projected onto the new principal component axes. The spread of these points indicates the variance captured by the principal components. A widespread suggests significant variance captured by PC1 and PC2.

**Axes**: The horizontal axis (PC1) and the vertical axis (PC2) are the first two principal components. PC1 captures the most variance in the data, while PC2 captures the second most, orthogonal to PC1.

**Factor** **Loadings** indicate how much each original variable contributes to the principal components. High absolute values suggest a substantial contribution:

Variables with high positive loadings (e.g., variables corresponding to 0.601550625, 0.498941183, etc.) contribute strongly and positively to PC1.

Variables with high negative loadings (e.g., variables corresponding to -0.184447937) contribute strongly and negatively to PC2.

**Explained Variance**: This shows how much of the total variance is captured by each principal component:

PC1 captures approximately 30.58% of the total variance.

PC2 captures approximately 7.84% of the total variance.

PC3 captures approximately 5.07% of the total variance.

PC4 captures approximately 4.63% of the total variance.

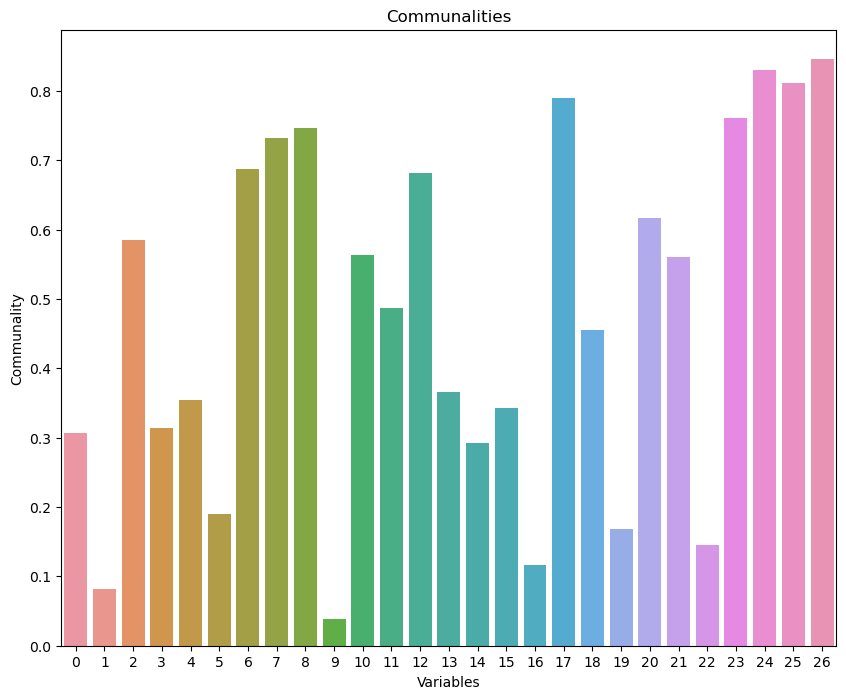
PC5 captures approximately 4.26% of the total variance.

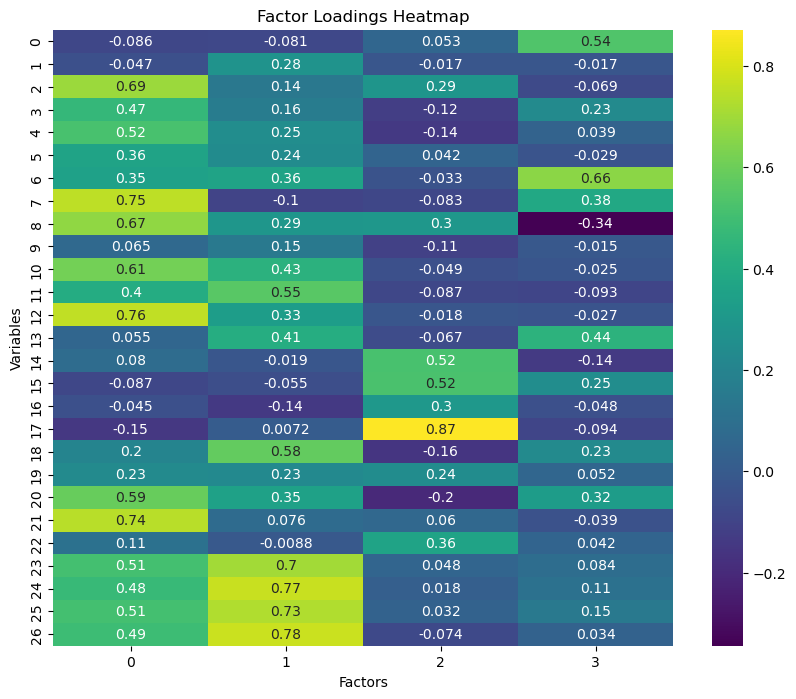
Cumulative Variance:

The first two principal components capture approximately 38.42% of the total variance.

The first five principal components capture approximately 52.38% of the total variance.

1. **Factor analysis**
2. Factor Loadings with Factor Names:
3. Factor3 Factor2 Factor1 fds
4. 3. Proximity to transport 0.053156 -0.080974 -0.086236
5. 4. Proximity to work place -0.016544 0.281713 -0.047079
6. 5. Proximity to shopping 0.288108 0.142647 0.690591
7. 1. Gym/Pool/Sports facility -0.124854 0.163524 0.466517
8. 2. Parking space -0.142834 0.248550 0.519622
9. 3.Power back-up 0.042454 0.238051 0.362113
10. 4.Water supply -0.033121 0.360861 0.347408
11. 5.Security -0.083320 -0.100945 0.752826
12. 1. Exterior look 0.301715 0.294399 0.671084
13. 2. Unit size -0.108469 0.149637 0.065008
14. 3. Interior design and branded components -0.049326 0.432131 0.611537
15. 4. Layout plan (Integrated etc.) -0.086777 0.554334 0.404877
16. 5. View from apartment -0.017909 0.328739 0.756467
17. 1. Price -0.067204 0.406913 0.054603
18. 2. Booking amount 0.516262 -0.019179 0.080124
19. 3. Equated Monthly Instalment (EMI) 0.520293 -0.054892 -0.086762
20. 4. Maintenance charges 0.302818 -0.141021 -0.045122
21. 5. Availability of loan 0.871707 0.007222 -0.145551
22. 1. Builder reputation -0.157026 0.577809 0.203555
23. 2. Appreciation potential 0.243617 0.228441 0.231016
24. 3. Profile of neighbourhood -0.203621 0.351803 0.590418
25. 4. Availability of domestic help 0.060175 0.075905 0.741031
26. Time 0.362378 -0.008787 0.110709
27. Size 0.048069 0.701323 0.510278
28. Budgets 0.018482 0.769450 0.475594
29. Maintainances 0.031953 0.727605 0.508890
30. EMI.1 -0.074451 0.775484 0.487657
31. Factor4
32. 3. Proximity to transport 0.538652
33. 4. Proximity to work place -0.016725
34. 5. Proximity to shopping -0.069206
35. 1. Gym/Pool/Sports facility 0.232471
36. 2. Parking space 0.038646
37. 3.Power back-up -0.029130
38. 4.Water supply 0.660016
39. 5.Security 0.384690
40. 1. Exterior look -0.344246
41. 2. Unit size -0.014709
42. 3. Interior design and branded components -0.024710
43. 4. Layout plan (Integrated etc.) -0.093437
44. 5. View from apartment -0.027259
45. 1. Price 0.438339
46. 2. Booking amount -0.138105
47. 3. Equated Monthly Instalment (EMI) 0.249079
48. 4. Maintenance charges -0.048089
49. 5. Availability of loan -0.094296
50. 1. Builder reputation 0.234381
51. 2. Appreciation potential 0.051778
52. 3. Profile of neighbourhood 0.321576
53. 4. Availability of domestic help -0.038846
54. Time 0.041916
55. Size 0.083502
56. Budgets 0.109056
57. Maintainances 0.145793
58. EMI.1 0.033923





Factor analysis aims to identify underlying relationships between observed variables. The provided factor loadings and visualization suggest that the dataset can be effectively reduced to a few underlying factors. The visualization is a network graph showing the relationships between variables (Var1, Var2, ..., Var27) and factors (Factor1, Factor2, Factor3, Factor4). The edges (lines) between variables and factors are labeled with factor loadings, indicating the strength and direction of the relationships. Thicker lines represent stronger relationships.

Factor loadings indicate the correlation between the variables and the factors. Higher absolute values signify stronger relationships.

**Factor1 (Amenities):**

Security: 0.752826

Proximity to shopping: 0.690591

Gym/Pool/Sports facility: 0.466517

Parking space: 0.519622

Availability of domestic help: 0.741031

This factor represents variables related to amenities and security, highlighting their importance in the dataset.

**Factor2 (Financial Considerations):**

Budgets: 0.769451

Maintenance: 0.727605

EMI: 0.775484

Size: 0.701323

This factor emphasizes financial considerations, including budget, maintenance, and loan-related aspects.

**Factor3 (Availability):**

Availability of loan: 0.871707

Booking amount: 0.516262

Equated Monthly Instalment (EMI): 0.520293

This factor represents practical aspects of purchasing, such as loan availability and booking amounts.

**Factor4 (Proximity):**

Proximity to transport: 0.538652

Water supply: 0.660016

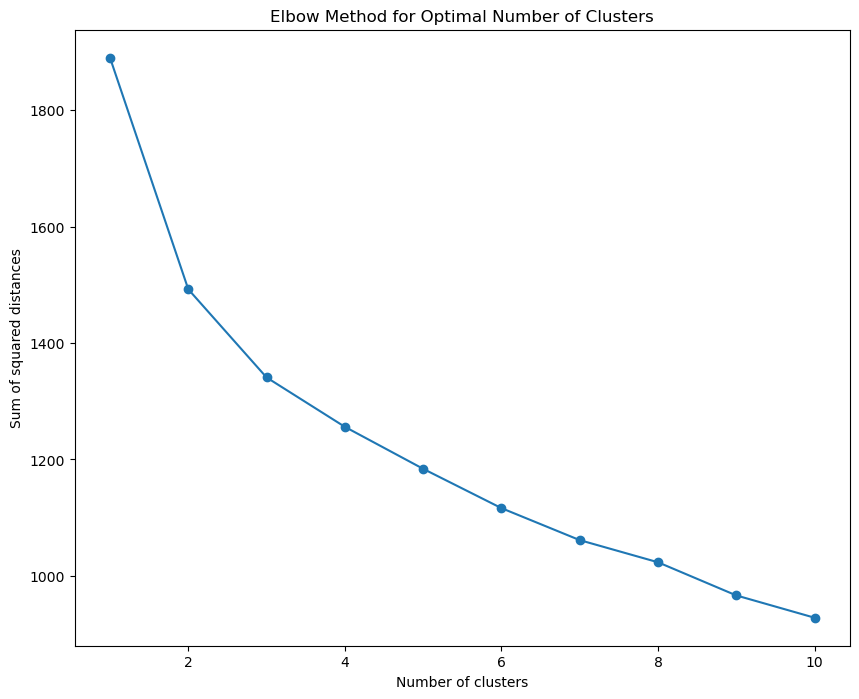
Exterior look: 0.344246

Profile of neighbourhood: 0.321576

This factor underscores the importance of proximity to essential services and infrastructure, like transport and water supply.

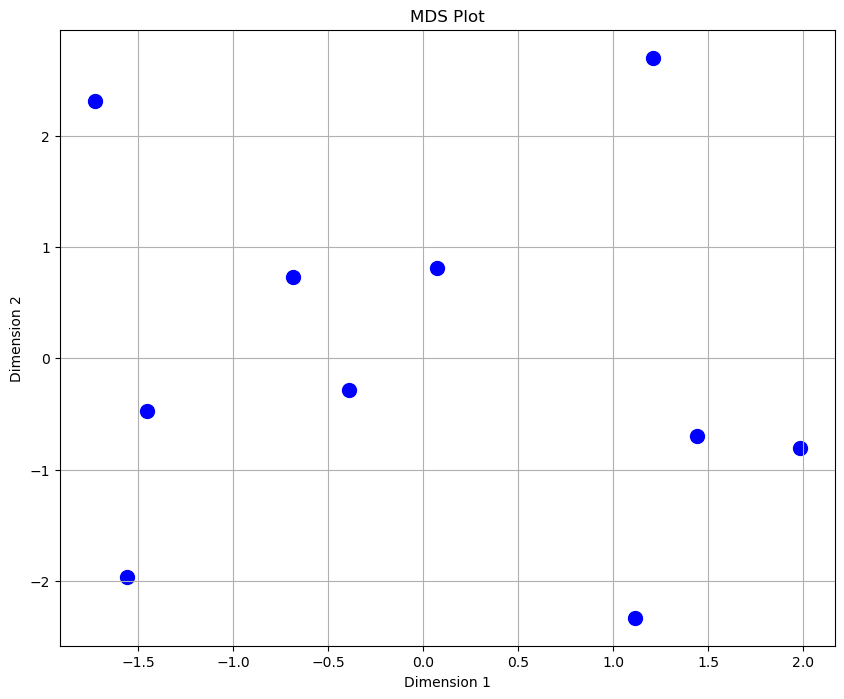
The factor analysis reveals four vital underlying factors in the dataset: Amenities, Financial Considerations, Availability, and Proximity. These factors provide a comprehensive understanding of the variables and their interrelationships, offering valuable insights for further analysis and decision-making.

**c. Cluster Analysis**



The silhouette score plot you've provided determines the optimal number of clusters for a dataset by evaluating how well each data point fits within its assigned cluster compared to other clusters. The silhouette score measures how similar an object is to its cluster (cohesion) compared to different clusters (separation). The score ranges from -1 to 1. A higher silhouette score indicates that the data points are well-matched to their cluster and poorly matched to neighboring clusters. Further, increasing clusters generally result in lower silhouette scores, with some fluctuations.

**d. Multi-dimensional scaling (icecream.csv)**



The MDS plot displays ice cream brands in a two-dimensional space defined by MDS Dimension 1 and MDS Dimension 2. Each point represents a brand, with the distances between points reflecting the similarity or dissimilarity of the brands based on scaled numerical features used in the MDS analysis.

**Clusters and Proximity:**

**Kwality, Arun, and Vadilal**: These brands are clustered closely in the lower left quadrant, indicating similar feature profiles.

**Nandini, KVAFSU, and Joy:** These brands are located in the upper right quadrant, forming another group with features that are similar to those of the first cluster.

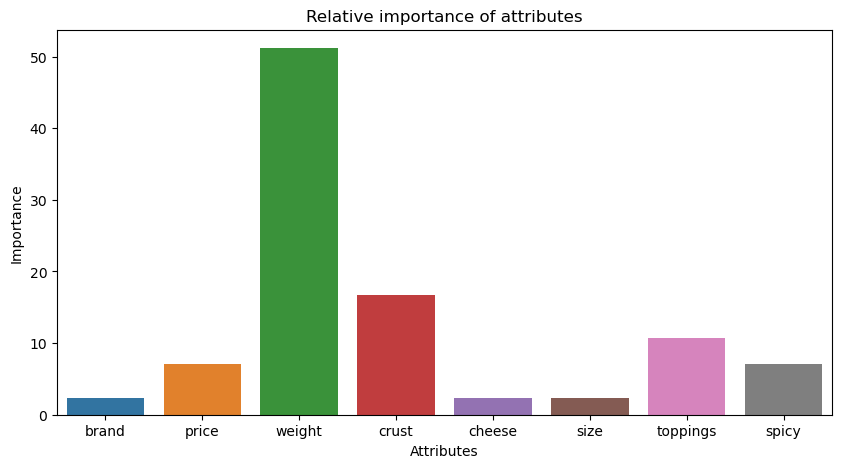
**Hatson**: Positioned near the center-left, this brand appears similar to the Kwality-Arun-Vadilal cluster but is distinct enough to occupy its unique position.

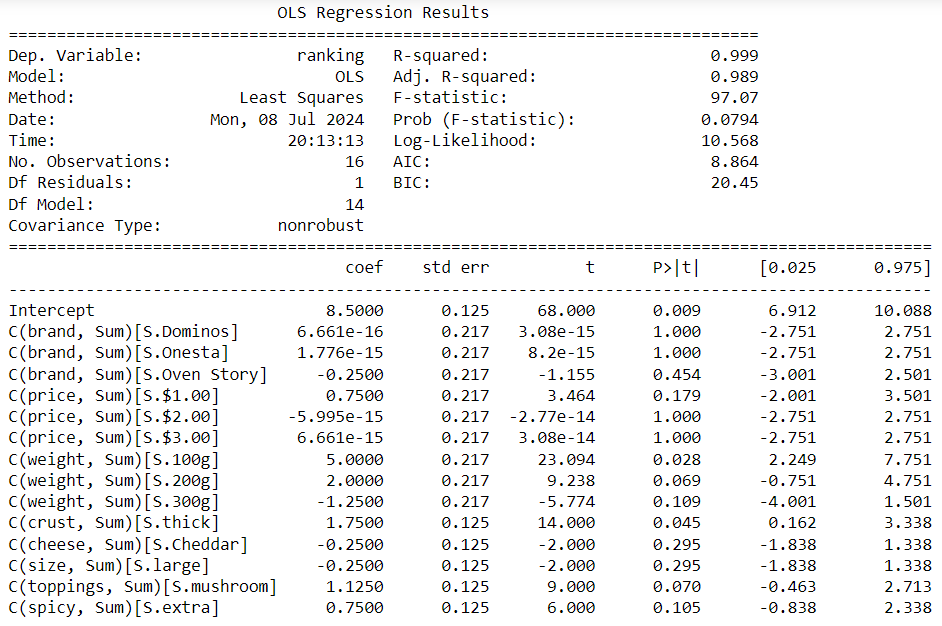
**Vijaya:** Situated at the top center, Vijaya stands alone, indicating a unique feature profile compared to the other brands.

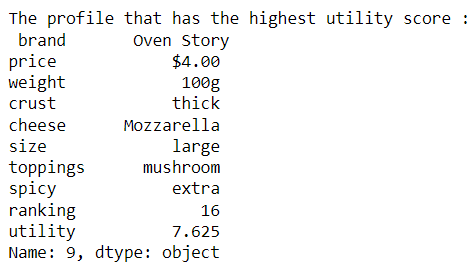
**Dodla and Amul:** These brands are located in the lower right quadrant, suggesting they share similar characteristics but are distinct from the other clusters.

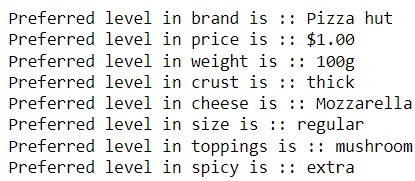
Brands like Vijaya and Joy are positioned away from other brands, indicating unique feature sets that differentiate them from others in the market. Dodla and Amul, while close to each other, are also distant from different clusters, indicating a shared but distinct profile. Brands that are close together in the MDS plot might compete directly in the market, as their product features are similar. Conversely, farther apart brands may cater to different market segments or have distinct unique selling propositions.

**e. Conjoint Analysis (pizza\_data.csv)**



****

****

****

Conjoint analysis is a statistical technique used to understand how consumers value different product or service attributes. In this case, the analysis was conducted on various characteristics of pizzas.

**Regression Analysis:**

The OLS regression model has an exceptionally high R-squared value of 0.999, indicating that the model explains nearly all variability in the ranking. However, the adjusted R-squared value drops to 0.989, suggesting potential overfitting due to the many predictors relative to observations. With a p-value of 0.0794, the overall model is not statistically significant at the conventional 0.05 level, raising concerns about the robustness of the model's predictive power. Considerable predictors include the intercept (p=0.009), weight of 100g (p=0.028), and thick crust (p=0.045). These attributes positively influence the ranking.

**Highest Utility Profile:**

The profile with the highest utility score is:

Brand: Oven Story

Price: $4.00

Weight: 100g

Crust: Thick

Cheese: Mozzarella

Size: Large

Toppings: Mushroom

Spicy: Extra

Ranking: 16

Utility: 7.625

This profile has the highest utility score of 7.625, meaning it is the most preferred combination of attributes among the profiles evaluated in the analysis.

**Comparison and Insights:**

**Brand**:

Highest Utility Profile: Oven Story

Preferred: Pizza Hut

Insight: Although Pizza Hut is the preferred brand, Oven Story was found to have the highest utility. This suggests that other attributes in the Oven Story profile may have contributed significantly to its overall preference.

**Price**:

Highest Utility Profile: $4.00

Preferred: $1.00

Insight: There is a discrepancy here. Despite the preferred price being $1.00, the highest utility profile had a price of $4.00. This indicates that price might not be the dominant factor in determining the overall utility for this consumer.

**Weight**:

Both the highest utility profile and the preferred profile have a weight of 100g, indicating this is a consistent and preferred attribute.

**Crust**:

Both profiles prefer a thick crust, showing consistency and preference for this attribute.

**Cheese**:

Both profiles prefer Mozzarella cheese, indicating this is a preferred cheese type.

**Size**:

Highest Utility Profile: Large

**Preferred**: Regular

Insight: There is a discrepancy here, suggesting that while the consumer may prefer a regular size in general, the large size combined with other attributes led to the highest utility.

**Toppings**:

Both profiles prefer mushroom toppings, showing a consistent preference.

**Spicy**:

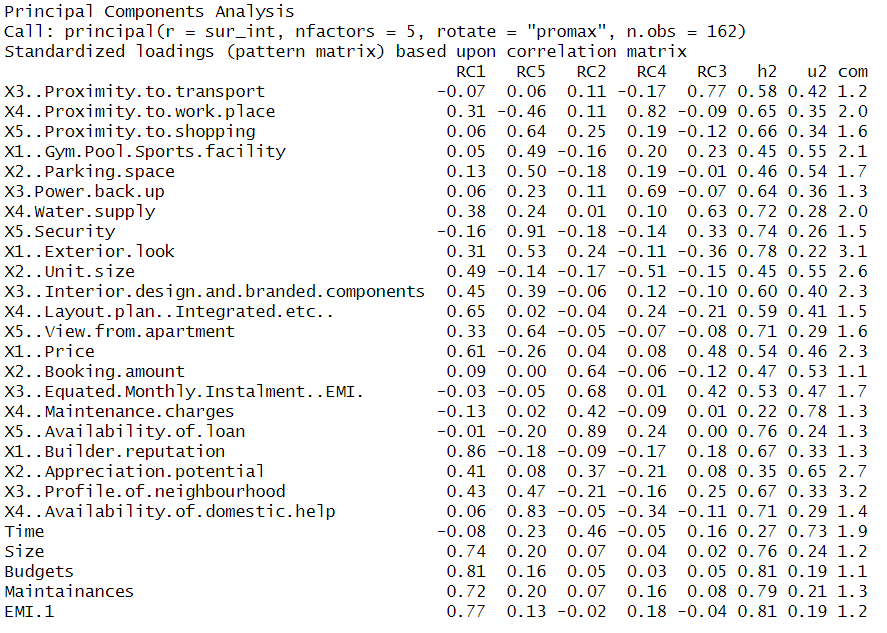
Both profiles prefer extra spicy, indicating a preference for spiciness.

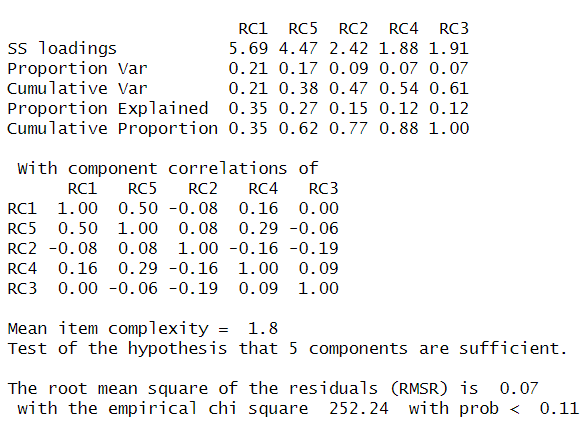
**Conclusion**:

The conjoint analysis reveals that the profile with the highest utility score (Oven Story, $4.00, 100g, thick crust, Mozzarella cheese, large size, mushroom toppings, extra spicy) is not entirely aligned with the preferred levels for some attributes (brand, price, and size). However, the consistency in preferences for weight, crust, cheese, toppings, and spiciness suggests these attributes are highly valued by the consumer. The discrepancy between the highest utility profile and the preferred levels of brand, price, and size indicates that the consumer values the overall combination of attributes more than individual preferred levels.

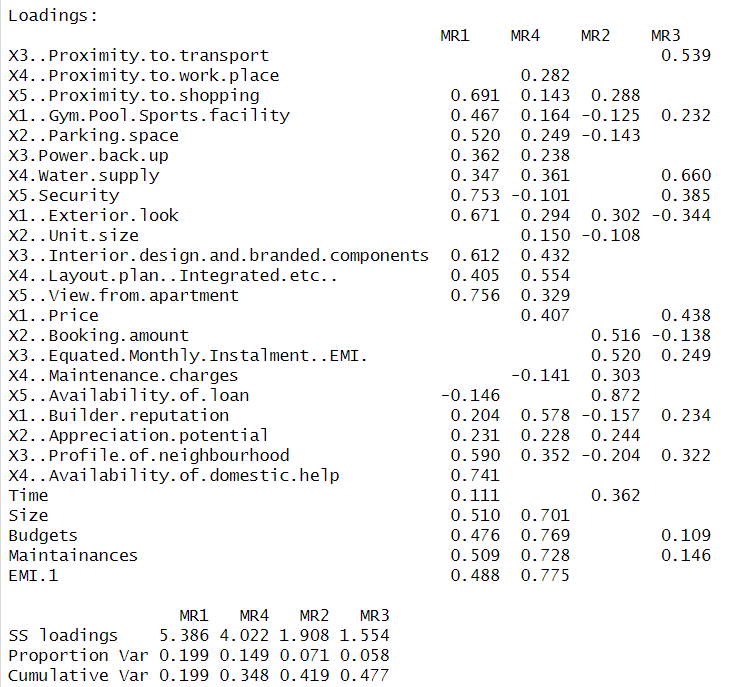
**Results and interpretation (R)**

1. **Principle factor analysis**

****

****

1. RC1: The variables X1 (Builder reputation), Size, Budgets, Maintenance, and EMI 1 have high loadings, indicating a strong relationship with the first principal component.
2. RC2: The variables X2 (Booking amount), X3 (Equated Monthly installment - EMI), and X5 (Availability of loan) have high loadings, suggesting this component is primarily associated with financial aspects.
3. RC3: The variables X3 (Proximity to transport) and X4 (Water supply) have high loadings, indicating a focus on infrastructure and utilities in this component.
4. RC4: The variables X4 (Proximity to the workplace) and X3 (Power backup) have high loadings, indicating a relationship between proximity to work and backup facilities.
5. RC5: The variables X5 (Security) and X1 (Exterior look) have high loadings, suggesting this component is related to security and aesthetics.
6. **Factor analysis**

****

**Factor Loadings:**

Factor loadings reflect the correlation between the observed variables and the underlying factors. The

four factors (MR1, MR2, MR3, MR4) can be interpreted as follows based on their loadings:

**MR1:**

• High loadings on "X5.Security," "X1..Exterior look," "X5..View from the apartment,"

"X4..Availability of domestic help," "Size," "Budgets," and "Maintenance."

• This factor can be interpreted as representing "Safety and Aesthetic Appeal."

**MR2:**

• High loadings on "X5. Availability of loan," "X2. Booking amount," and "X3.Equated Monthly Instalment (EMI)."

• This factor appears to capture "Financial Aspects."

**MR3:**

• High loadings on "X4. Water supply," "X5..Security," and "X1..Price."

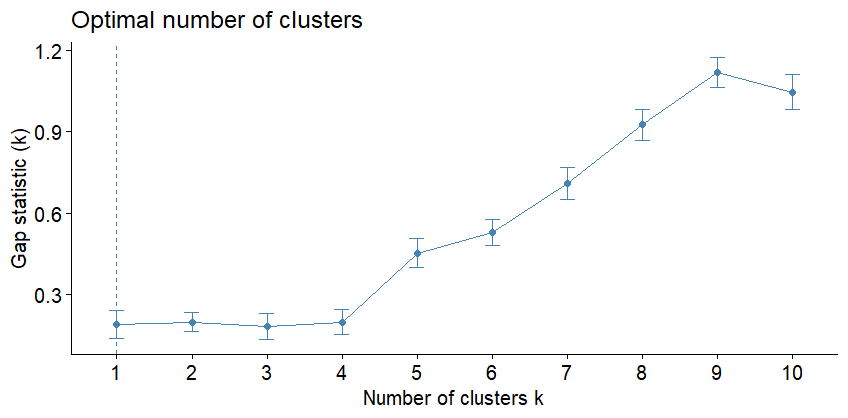
• This factor could represent "Basic Amenities and Security."

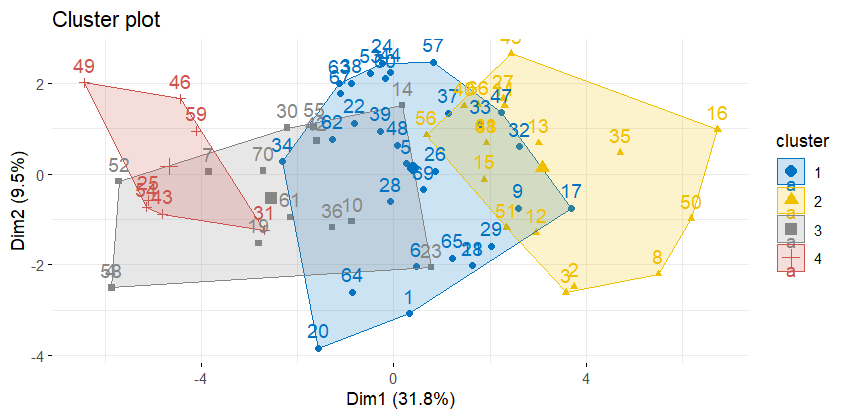
**MR4:**

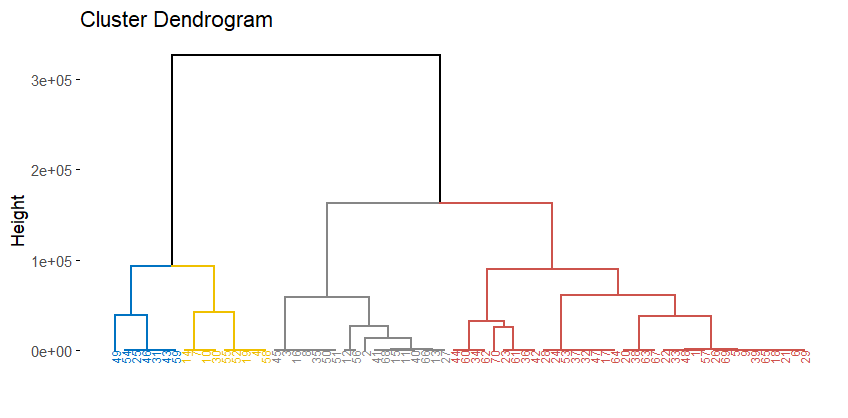
• High loadings on "Size," "Budgets," "Maintenance," and "EMI.1."

• This factor might represent "Space and Maintenance Costs."

1. **Cluster analysis**

****

****

****

This plot determines the optimal number of clusters (k) for k-means clustering. The x-axis represents the number of clusters, while the y-axis shows the gap statistic. The optimal number of clusters is typically the one with the highest gap statistic value, which seems to be k = 8. This plot combines a heatmap and a dendrogram to show the hierarchical clustering results. The dendrogram on the left indicates the hierarchical structure of the observations, grouped into eight clusters. The heatmap shows the values of different variables for each observation, with color intensity representing the magnitude of the values.

• Optimal number of clusters (k): 8

• Cluster visualization: The clusters are well-separated in the cluster plot and the

heatmap/dendrogram provides a detailed hierarchical structure. The plot shows the clusters in different colors and shapes. Each cluster is represented by a polygon that encloses the data points belonging to that cluster. Here's a breakdown:

• Cluster 1: Represented in dark blue with circles. This cluster is distinct and separated from

the others, indicating well-defined data points.

• Cluster 2: Represented in yellow with triangles. It spans a large area, indicating variability

within the cluster.

• Cluster 3: Represented in light grey with squares. This cluster overlaps significantly with other

clusters, suggesting less distinct separation.

• Cluster 4: Represented in red with crosses. This cluster is relatively small and overlaps with

clusters 3, 5, and 6.

• Cluster 5: Represented in light blue with diamonds. This cluster also overlaps significantly with other clusters.

• Cluster 6: Represented in dark grey with upward triangles. This cluster is similar to cluster 4 in size and position.

• Cluster 7: Represented in black with X marks. This cluster overlaps with clusters 3 and 5,

indicating less distinct boundaries.

• Cluster 8: Represented in light grey with asterisks. This cluster overlaps significantly with

others, showing less distinct separation.

Overlapping Clusters

• Clusters 3, 4, 5, 6, and 7 significantly overlap, indicating that these data points are not well separated and may share similarities.

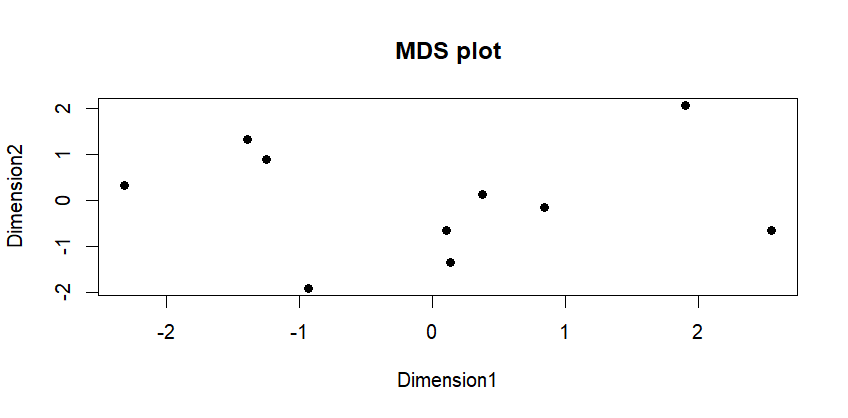
• Cluster 1 is the most distinct, with minimal overlap, indicating well-defined separation from

other clusters.

• Cluster 2 has some overlap but still maintains a relatively distinct boundary.

The plot provides insights into the clustering structure of the data. Cluster 1 is the most well-separated, while other clusters show significant overlap, indicating potential challenges in distinguishing these clusters based on the first two principal components alone.

1. **Multidimensional scaling (icecream.csv)**

****

The MDS plot provides a visual representation of the similarities and differences among ice cream brands based on scaled numerical features:

**Bottom Left Quadrant:** A small cluster of points in this quadrant indicates brands with similar feature profiles. These brands may compete closely in the market or share common characteristics.

**Upper Left Quadrant**: A single point here represents a brand with a unique feature profile compared to others. This brand is distinct and may cater to a niche market.

**Center to Center-Right Quadrant:** The spread of points from the center to the right suggests several brands with similar but distinct features. These brands are not closely clustered, indicating moderate differences in their features.

**Upper Right Quadrant:** A single point in this quadrant signifies another brand with unique characteristics that set it apart.

**Bottom Right Quadrant**: A small cluster of points here indicates brands with similar features that may compete directly with each other.

Brands positioned away from clusters, such as those in the upper left and upper right quadrants, likely have unique feature sets. These brands differentiate themselves in the market with distinct, unique selling propositions. The spread of points across the plot suggests that brands cater to different market segments or emphasize different features to appeal to various customer preferences.

1. **Conjoint Analysis (pizza\_data.csv)**

**Brand**: The most important attribute, with an importance value of 30. This indicates that the brand of the pizza significantly influences consumer choice.

**Weight**: The second most important attribute, with an importance value close to 30. This suggests that the pizza's weight is almost as crucial as the brand.

**Crust**: The third most important attribute, with an importance value of around 20. The type of crust is also a significant factor in consumer preference.

**Cheese**: This attribute has a moderate importance value of around 10. The type of cheese used affects consumer choice, but not as much as brand, weight, or crust.

**Size, Spicy, and Toppings**: These attributes have relatively low importance values, indicating they are less critical in determining consumer preferences than the other attributes.

**Price**: Surprisingly, price has the lowest importance value, suggesting that consumers may prioritize other attributes over the cost when choosing a pizza.

The part-worth utilities (attribute levels) reflect the relative preferences for different levels of each attribute. These utilities help understand each attribute level's impact on the overall preference for a product.

**Dominos**: Has the highest positive utility (17.375), indicating a strong preference.

**Pizza Hut:** Serves as the base level for brand comparison.

**Onesta and Oven Story**: They have damaging utilities, indicating a lesser preference than the base level (Pizza Hut).

**Lowest Price Level ($1.00):** Has a positive utility, showing a preference for cheaper options.

**Heavier Pizzas (300g and 400g)** Have significantly negative utilities, indicating a strong preference for lighter pizzas.

**Thin Crust**: Has a negative utility, suggesting a preference for thick crust (since thick crust serves as the base level).

Mozzarella Cheese, Regular Size, Paneer Topping, and Normal Spiciness Levels: Have specific utilities indicating varying preferences among these attributes. These utilities provide insights into which attributes and levels are most and least preferred by consumers, guiding product development and marketing strategies.