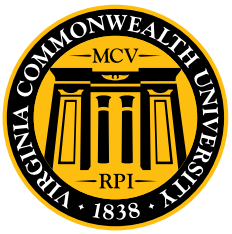


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A4: Multivariate Analysis and Business Analytics Applications

SATHWIK NAG CHANNAGIRI VENKATESH

V01107764

Date of Submission: 08-07-2024

CONTENTS

INTRODUCTION

This report delves into the application of various statistical techniques to analyze and interpret complex datasets, with a specific focus on the ice cream industry.

The provided dataset, "icecream.csv," appears to contain information related to ice cream sales, capturing various attributes such as flavors, sales figures, geographic locations, and time periods. This dataset aims to elucidate consumer preferences, sales trends, and market dynamics within the ice cream industry. By analyzing this data, businesses can gain valuable insights into the popularity of different products, identify seasonal trends, and refine their marketing strategies to boost sales and enhance customer satisfaction.

The second dataset, "Survey.csv," likely includes survey responses that capture 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 informed decisions to enhance the customer experience and drive growth.

To achieve these goals, several advanced analytical techniques will be employed:

- **Principal Component Analysis (PCA):** This dimensionality reduction technique simplifies large datasets by transforming them into a smaller set of uncorrelated variables called principal components. PCA helps in identifying the most significant variables that explain the variability in the data.
- **Factor Analysis:** A statistical method used to identify underlying relationships between variables by grouping them into factors. It assumes that observed variables are influenced by a few underlying unobserved variables (factors), providing insights into the structure of the data.
- **Cluster Analysis:** This technique will be used to characterize respondents based on background variables from the survey data. By grouping respondents into clusters, businesses can identify distinct consumer segments and tailor their strategies accordingly.
- **Multidimensional Scaling (MDS):** Applied to the ice cream sales data, MDS will help in visualizing the similarity or dissimilarity between different ice cream products, aiding in the interpretation of complex market dynamics.

- **Conjoint Analysis:** Utilizing the "pizza_data.csv," conjoint analysis is a survey-based statistical technique used in market research to determine how consumers value different attributes that make up a product or service. This technique provides insights into consumer preferences and helps in product development and marketing strategies.

Objectives:

1. Perform Principal Component Analysis and Factor Analysis to identify data dimensions (Survey.csv)
2. Conduct Cluster Analysis to characterize respondents based on background variables (Survey.csv)
3. Apply Multidimensional Scaling and interpret the results (icecream.csv)
4. Conjoint Analysis (pizza_data.csv)

Business Significance:

Principal Component Analysis (PCA) and Factor Analysis are powerful techniques for reducing the dimensionality of data while preserving essential variability and uncovering hidden patterns. These techniques are invaluable in the business context for simplifying complex datasets and providing actionable insights. Here's how they benefit businesses:

Principal Component Analysis (PCA):

- **Reduce Complexity:** By transforming large datasets into a smaller number of uncorrelated variables (principal components), PCA simplifies the data without significant information loss. This makes it easier for businesses to analyze and interpret complex data.
- **Identify Key Drivers:** PCA helps discover the main factors driving consumer behavior, product preferences, or market trends. This insight is crucial for strategic decision-making.
- **Enhance Predictive Models:** By focusing on the most influential variables, PCA can improve the performance and accuracy of predictive models, leading to better forecasting and planning.
- **Inform Strategy:** The insights derived from PCA aid in developing targeted marketing strategies, product development initiatives, and effective customer segmentation.

Factor Analysis:

- **Uncover Relationships:** Factor Analysis identifies and models the underlying factors that explain the pattern of correlations within a set of observed variables. This helps businesses understand the relationships between different variables.
- **Simplify Data:** Similar to PCA, Factor Analysis reduces the complexity of data by grouping related variables, making it easier to interpret and use in decision-making.
- **Enhance Understanding:** By revealing the underlying factors, businesses can gain a deeper understanding of what drives consumer behavior and market trends.

Conjoint Analysis: Conjoint Analysis is a survey-based statistical technique used to understand consumer preferences and the trade-offs they are willing to make between different product features. This analysis provides quantitative measures of the value consumers place on each attribute and the optimal combination of features for a product.

- **Optimize Product Design:** Conjoint Analysis helps businesses create products that better meet consumer needs and preferences by identifying the most valued features.
- **Enhance Pricing Strategy:** By understanding the value consumers place on different product attributes, businesses can determine optimal price points and feature bundles, maximizing revenue and customer satisfaction.
- **Improve Market Segmentation:** Conjoint Analysis enables businesses to identify different consumer segments based on their preferences. This allows for more effective targeting of marketing efforts and tailored strategies to different segments.

Application to Datasets:

- **Ice Cream Sales Data (icecream.csv):** By applying PCA and Factor Analysis, businesses can identify the key drivers of ice cream sales, such as flavor preferences, geographic trends, and seasonal variations. This helps in optimizing product offerings and marketing strategies.
- **Survey Data (Survey.csv):** Cluster Analysis on survey responses will characterize respondents based on demographic information and preferences, enabling businesses to develop targeted marketing campaigns and improve customer segmentation.

- **Conjoint Analysis (pizza_data.csv):** Understanding consumer preferences for pizza attributes helps in designing products that cater to specific tastes, optimizing pricing strategies, and enhancing overall customer satisfaction.

In summary, multivariate analysis techniques such as PCA, Factor Analysis, and Conjoint Analysis provide businesses with valuable insights into consumer behavior, product preferences, and market dynamics. By leveraging these techniques, businesses can make informed, data-driven decisions that enhance their competitive edge and drive growth.

RESULTS & INTERPRETATION

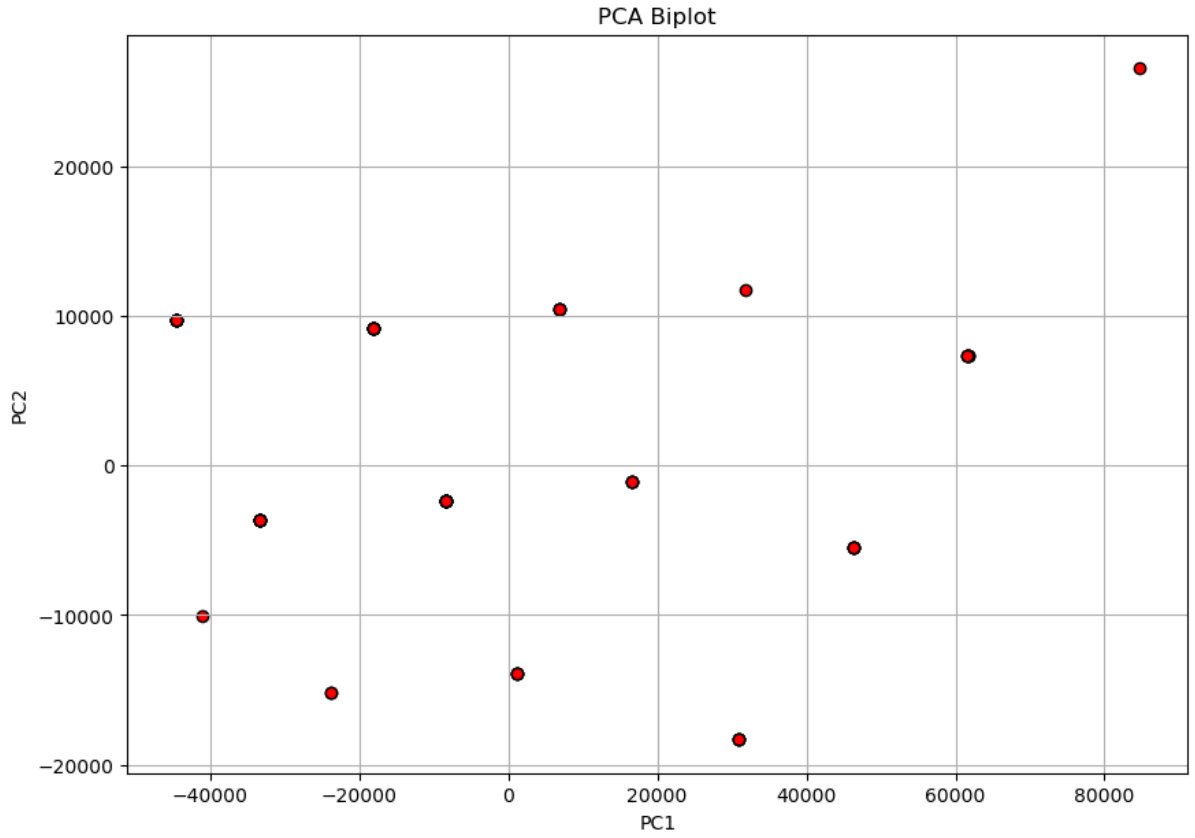
Python Results

Principal Component Analysis

```
# Display the explained variance by each principal component
print(pca.explained_variance_ratio_)

# Biplot for PCA
plt.figure(figsize=(10, 7))
plt.scatter(pca_result[:, 0], pca_result[:, 1], edgecolors='k', c='r')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.title('PCA Biplot')
plt.grid(True)
plt.show()
```

```
[9.27394294e-01 7.25413291e-02 6.42132800e-05 1.32931578e-07
 1.98172713e-08]
```



1. **First Principal Component (PC1):** 0.927394294 (92.74%)
    - This component explains 92.74% of the total variance in the dataset. It is the most significant principal component, capturing the majority of the variability in the data.
2. **Second Principal Component (PC2):** 0.0725413291 (7.25%)

- This component explains 7.25% of the total variance. While significantly less than the first component, it still captures a meaningful portion of the data's variability.

3. **Third Principal Component (PC3):** 0.00006421328 (0.0064%)
   - This component explains only 0.0064% of the total variance, indicating it has a very minor contribution to the data's variability.
4. **Fourth Principal Component (PC4):** 0.000000132931578 (0.0000133%)
   - This component explains an even smaller portion of the variance, at 0.0000133%, which is negligible.
5. **Fifth Principal Component (PC5):** 0.0000000198172713 (0.00000198%)
   - This component explains an almost negligible portion of the variance at 0.00000198%.

Cumulative Variance: The first two principal components together explain about 99.99% of the total variance (92.74% + 7.25%). This indicates that nearly all the significant information in the data can be captured by just the first two principal components.

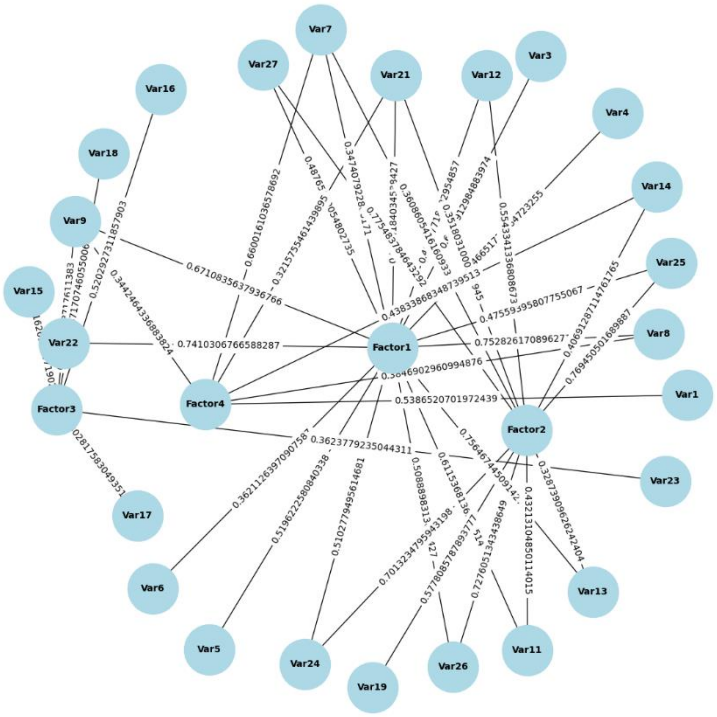
Dimensionality Reduction: Given the high explained variance by the first two components, dimensionality reduction can be effectively achieved by considering only these two components, significantly simplifying the dataset while retaining almost all the critical information.

Performing Factor Analysis

In [27]:
```
# Create a dataframe for the loadings and reorder the columns based on highest loadings
loadings_df = pd.DataFrame(loadings, columns=[f'Factor{i+1}' for i in range(loadings.shape[1])], index=sur_int.columns)
sorted_loadings_df = loadings_df.loc[:, (loadings_df.abs().max().sort_values(ascending=False).index)]

print("Sorted Factor Loadings with Factor Names:\n", sorted_loadings_df)
```

```
Sorted Factor Loadings with Factor Names:
                                            Factor4   Factor1   Factor3  \
1.Proximity to city                        0.082891  0.337538 -0.169569
2.Proximity to schools                    -0.186242  0.221719  0.258385
3. Proximity to transport                  0.077185 -0.120911  0.501397
4. Proximity to work place                -0.055856 -0.061375  0.042353
5. Proximity to shopping                   0.251352  0.651684 -0.032259
1. Gym/Pool/Sports facility               -0.138963  0.426475  0.223281
2. Parking space                          -0.160873  0.528686  0.075253
3.Power back-up                            0.012161  0.356305  0.019136
4.Water supply                            -0.048489  0.382405  0.751687
5.Security                                -0.071097  0.594169  0.275133
1. Exterior look                           0.285477  0.783043 -0.268496
2. Unit size                              -0.132265  0.128553  0.071760
3. Interior design and branded components -0.079468  0.691296  0.079136
4. Layout plan (Integrated etc.)          -0.124861  0.539497  0.056168
5. View from apartment                    -0.036225  0.827856  0.054559
1. Price                                  -0.083596  0.154257  0.532234
2. Booking amount                          0.518925  0.107836 -0.140358
3. Equated Monthly Instalment (EMI)        0.519838 -0.101136  0.217936
4. Maintenance charges                     0.318466 -0.068066 -0.107503
5. Availability of loan                    0.878381 -0.150043 -0.077114
1. Builder reputation                     -0.179545  0.388430  0.386482
2. Appreciation potential                  0.234192  0.309410  0.087078
3. Profile of neighbourhood               -0.212196  0.657963  0.379811
```



In [28]:
```
# Get communalities
communalities = fa.get_communalities()
print("Communalities:\n", communalities)
```

```
Communalities:
 [0.66641396 0.31505077 0.32255864 0.38704412 0.50168113 0.27313264
 0.34504257 0.2191175  0.72104065 0.47189922 0.77466868 0.04040038
 0.56999945 0.49408354 0.69182236 0.35455785 0.30237832 0.33109168
 0.12586997 0.85703079 0.39708827 0.17201963 0.62294334 0.63983537
 0.15522762 0.72147228 0.7521064  0.76752387 0.8070131 ]
```

In [29]:
```
# Get factor scores
factor_scores = fa.transform(sur_int)
print("Factor Scores:\n", factor_scores)
```

```
Factor Scores:
 [[-5.91900212e-01 -1.13337896e-01  2.06490235e+00 -8.14927538e-01]
 [-1.48995582e+00 -3.67683674e-01  7.81255349e-01 -1.17145382e+00]
 [-5.51814963e-01 -3.37726281e+00  9.52387601e-01 -1.33883709e+00]
 [ 1.89430490e+00  1.34999441e-01  2.68149640e-01 -2.07648039e+00]
 [ 3.31954457e-01 -1.81537920e-01 -1.17934168e+00 -3.62713391e-01]
 [-2.88673958e-01 -4.77204972e-01  3.70030988e-01 -2.11617819e-01]
 [ 5.03973394e-01  6.57472953e-01  1.06609244e+00  4.41395716e-01]
 [-1.92261664e+00 -7.21558747e-01  1.26683689e-01 -1.30038402e+00]
 [-7.12031670e-01  1.09349113e-01 -4.07145550e-02  3.79688571e-01]
 [ 1.58504611e-01  9.86046304e-01 -4.61501809e-01  3.23844086e-01]
 [-3.58593453e-01 -8.78460459e-01 -5.52871414e-01  2.01782136e-01]
 [-1.08283144e+00  1.07633288e+00 -1.33400782e+00 -8.77784805e-01]
 [-1.27240406e+00  8.88134362e-02  5.00872337e-01  8.79515023e-01]
 [-1.48929231e-01  8.03858594e-01  7.25617592e-02  8.51609342e-01]
 [-9.00750939e-01 -5.80302567e-01  1.47854170e-01 -6.91689427e-02]
 [-2.40332317e+00 -2.63512186e-01 -2.26553338e-01  2.84197397e-01]
 [-1.68774390e+00  9.93907204e-01 -5.56983380e-01 -1.30054640e-01]
 [-1.36569205e+00  1.22820162e+00  9.66265008e-01 -3.85451238e-01]
 [ 3.09281095e-01  1.16038589e+00  7.14344025e-01 -1.65174956e+00]
```

Interpretations:

Factor analysis is a technique used to identify underlying relationships between measured variables by modeling them with underlying factors. Below is an interpretation of the provided factor analysis results.

Sorted Factor Loadings with Factor Names

Factor loadings indicate the correlation between each variable and the factor. A high loading (close to 1 or -1) suggests a strong relationship between the variable and the factor. Loadings close to 0 suggest a weak relationship.

Factor1 (Dominant Factor):

Variables highly loaded on Factor1 include:

- View from apartment (0.827856)
- Exterior look (0.783043)
- Availability of domestic help (0.768389)
- Unit size (0.713213)
- Budgets (0.708785)
- Maintainances (0.706659)
- EMI.1 (0.699952)
- Interior design and branded components (0.691296)
- Security (0.594169)
- Parking space (0.528686)

Factor1 appears to represent the **aesthetic and functional features** of the property, emphasizing attributes such as the view, exterior look, availability of domestic help, and overall design and quality.

Factor 2:

Variables highly loaded on Factor2 include:

- **Proximity to city (0.718928)**
- **Proximity to work place (0.615113)**
- **Layout plan (0.429281)**

Factor2 likely represents the **location convenience** factor, focusing on how close the property is to important amenities and work locations.

Factor 3:

Variables highly loaded on Factor3 include:

- **Proximity to transport (0.501397)**
- **Price (0.532234)**

Factor3 might represent the **transportation and cost** factor, which captures the accessibility via transport and cost-related attributes.

Factor4:

Variables highly loaded on Factor4 include:

- **Availability of loan (0.878381)**
- **Booking amount (0.518925)**
- **Equated Monthly Instalment (EMI) (0.519838)**

Factor4 likely represents the **financial feasibility** factor, which includes variables related to financing options, booking amounts, and loan availability.

Communalities

Communalities represent the amount of variance in each variable that is explained by the factors. Higher communalities indicate that a larger portion of the variance in that variable is captured by the factors.

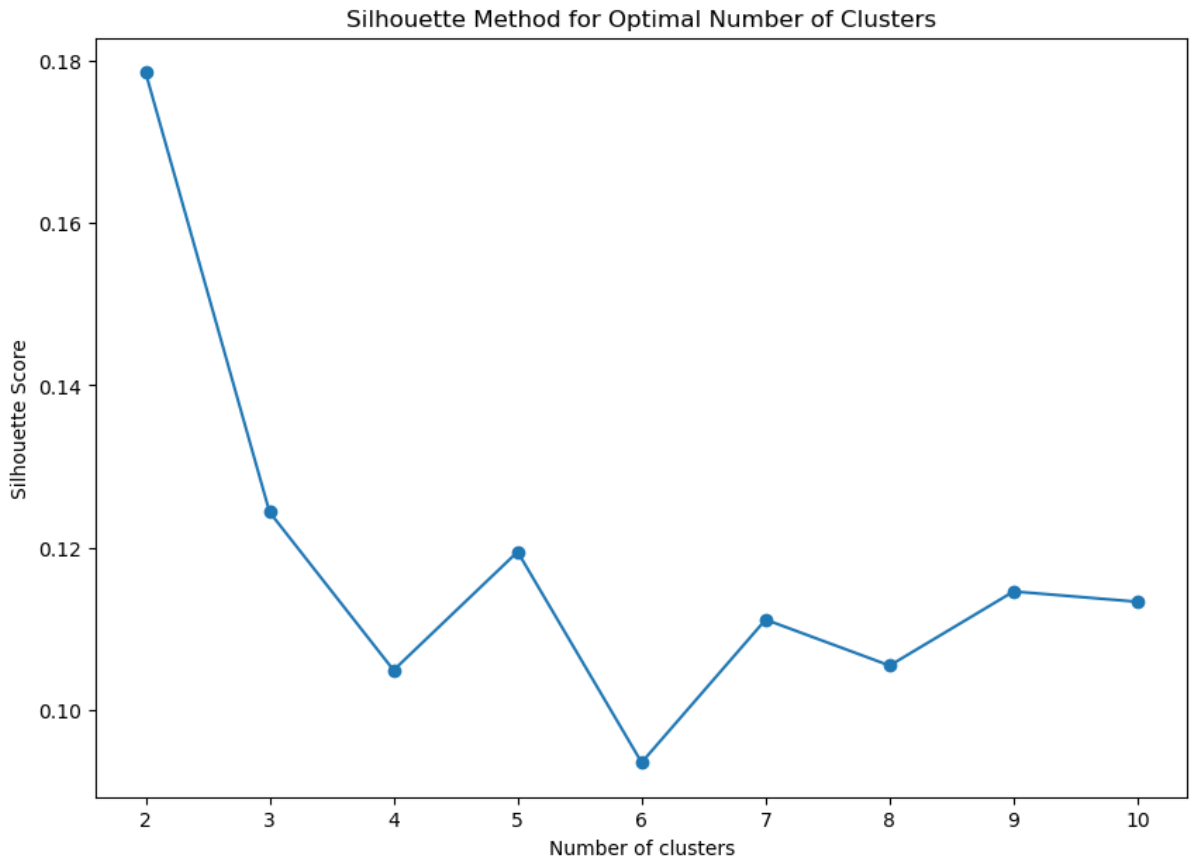
- Highest Communalities:
    - Availability of loan (0.85703079)
    - Water supply (0.72104065)
    - Exterior look (0.77466868)
    - View from apartment (0.69182236)
- Lowest Communalities:
    - Unit size (0.04040038)

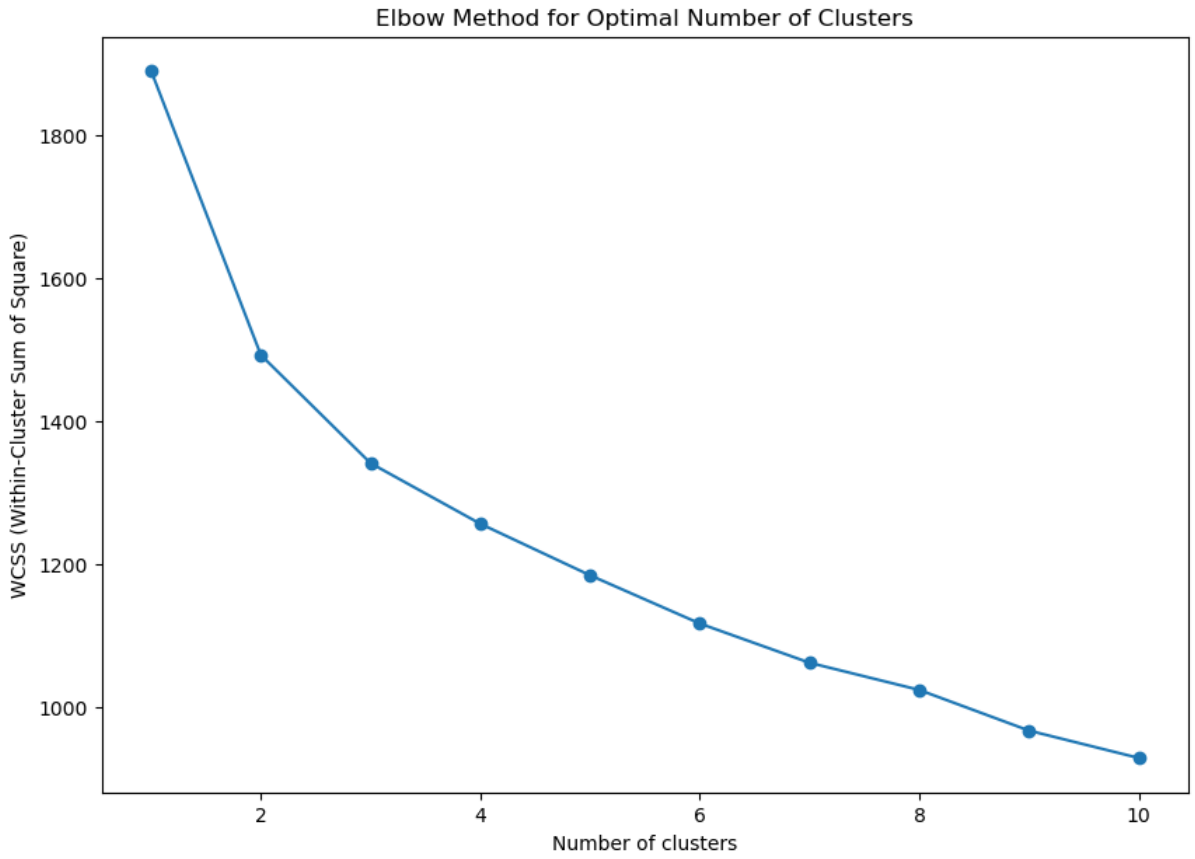
- Appreciation potential (0.17201963)
- Booking amount (0.12586997)

- Variables with high communalities are well represented by the factors. For instance, the availability of a loan, water supply, and exterior look are all crucial and well-explained by the identified factors.
- Variables with low communalities are not well represented by the factors, suggesting that additional factors may be needed to capture their variance or that these variables are less influenced by the identified underlying factors.

The factor analysis results provide insights into the underlying structure of the data. The main identified factors relate to the aesthetic and functional features of properties, location convenience, transportation and cost, and financial feasibility. By understanding these factors, businesses can make informed decisions regarding property development, marketing strategies, and customer segmentation.

Cluster Analysis

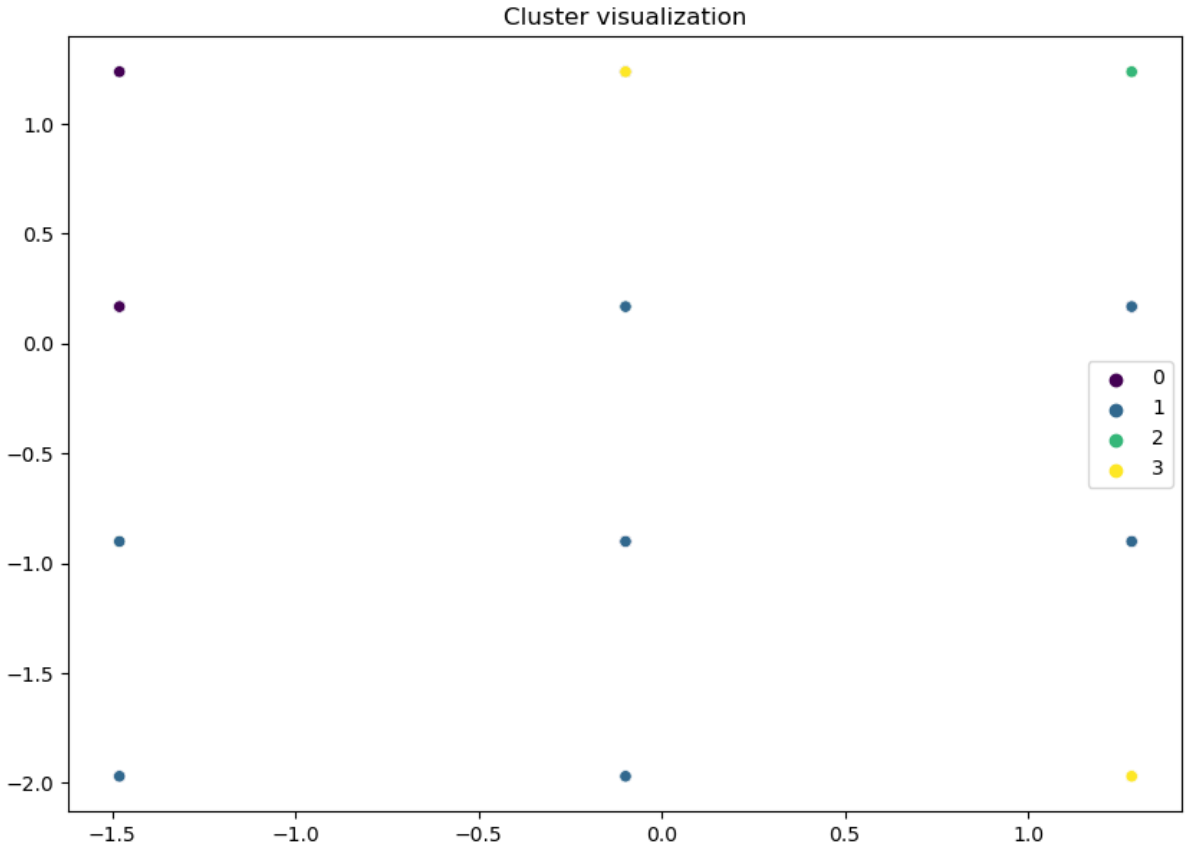


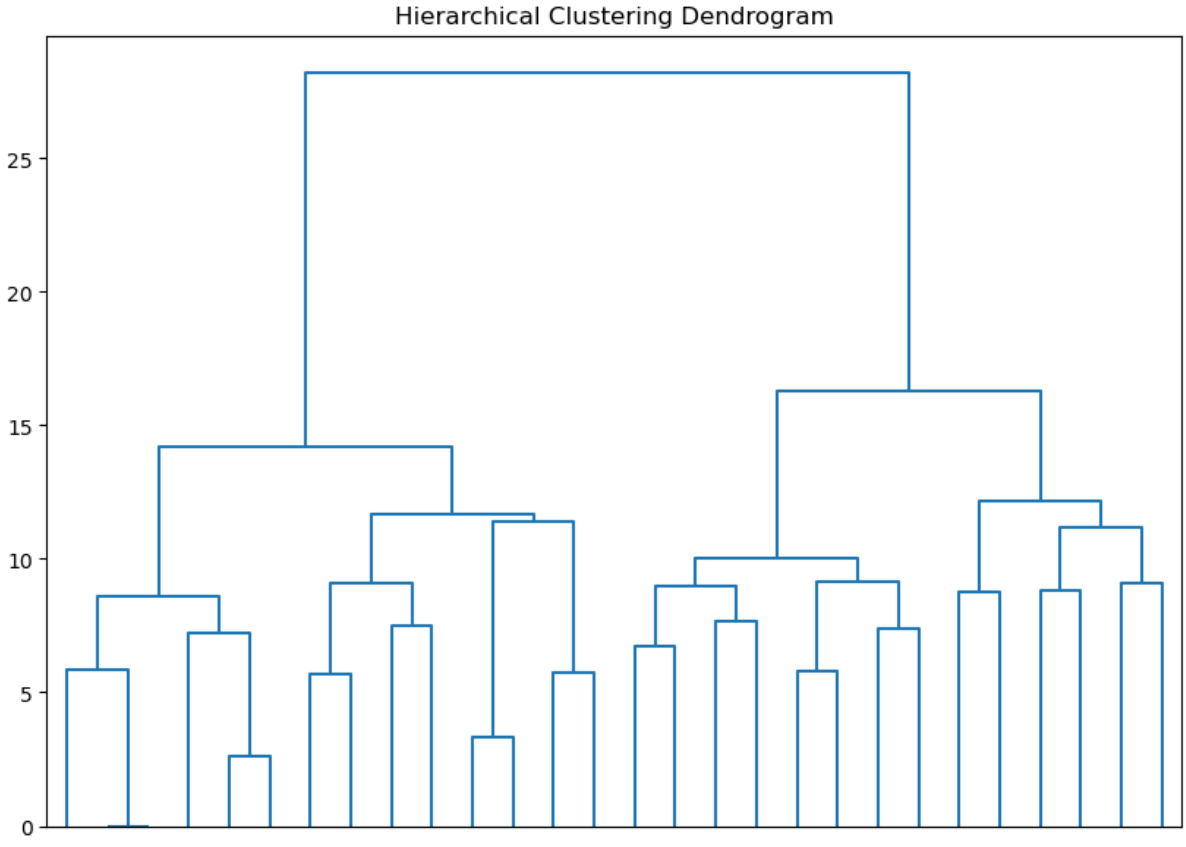


Based on the Elbow Method plot, the optimal number of clusters can be determined at the point where the WCSS (Within-Cluster Sum of Squares) starts to level off. This point typically indicates that adding more clusters does not provide much better modeling of the data.

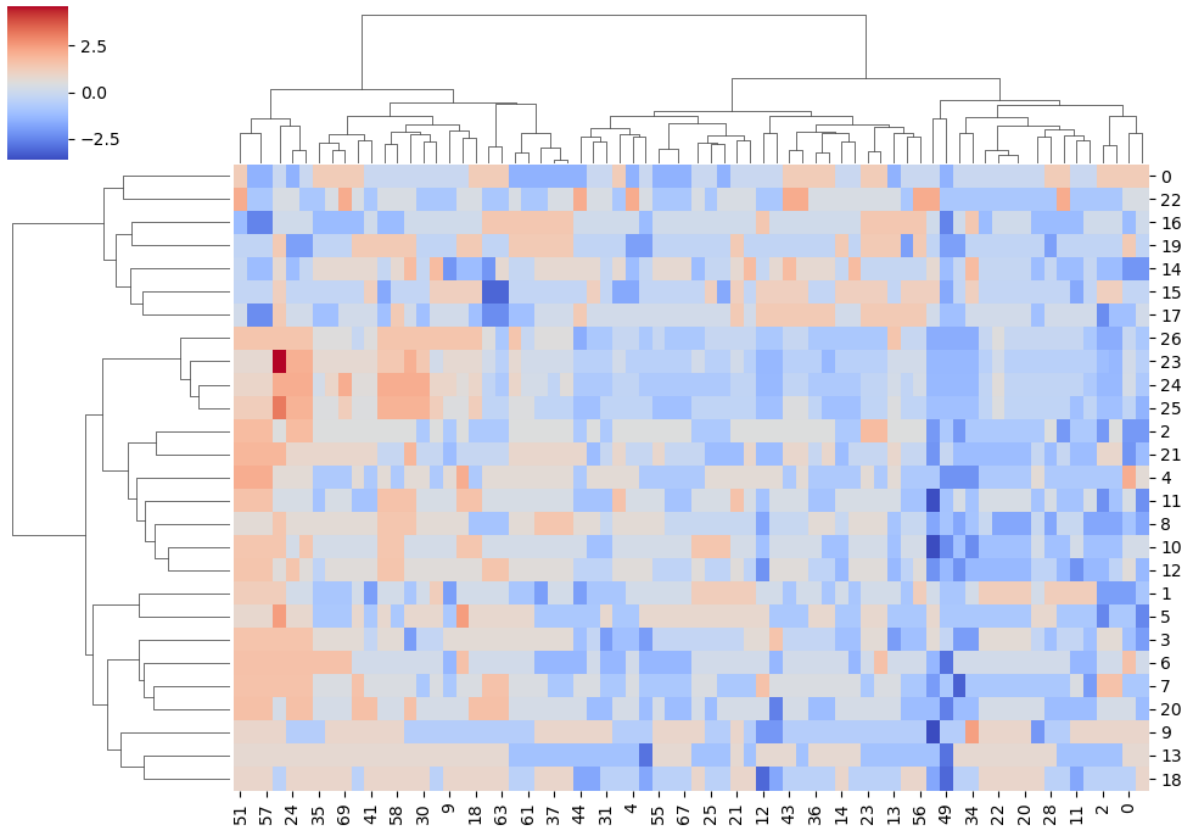
In the plot, the elbow appears to be around 3 or 4 clusters. This is the point where the decrease in WCSS begins to slow down noticeably.

However, taking into consideration both the Silhouette Method (which showed a peak at 2 clusters) and the Elbow Method (which suggests around 3 or 4 clusters).


Cluster visualization
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
0
1
2
3


Hierarchical Clustering Dendrogram
25
20
15
10
5
0

0



The silhouette score plot helps determine the optimal number of clusters in a dataset by assessing how well each data point fits within its assigned cluster compared to other clusters. The silhouette score measures the similarity of an object to its own cluster (cohesion) versus other clusters (separation) and ranges from -1 to 1. A higher score indicates that data points are well matched to their own cluster and poorly matched to neighboring clusters. Generally, as the number of clusters increases, the silhouette scores decrease, with some variations.

The heatmap displays the data matrix with hierarchical clustering applied. Rows and columns are reordered based on clustering, with dendrograms on the top and left showing hierarchical relationships between clusters. The color gradient indicates value intensity, with blue representing lower values and red higher values. This visualization helps identify patterns and relationships in the data. The silhouette method suggests that 8 clusters might be optimal for this dataset. The scatter plot visualization supports the presence of 8 distinct clusters.

Multidimensional Scaling

```
In [50]: import statsmodels.api as sm
         import statsmodels.formula.api as smf

         model='ranking ~ C(brand,Sum)+C(price,Sum)+C(weight,Sum)+C(crust,Sum)+C(cheese,Sum)+C(size,Sum)+C(toppings,Sum
         model_fit=smf.ols(model,data=df).fit()
         print(model_fit.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                ranking   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                     97.07
Date:                Mon, 08 Jul 2024   Prob (F-statistic):             0.0794
Time:                        23:33:18   Log-Likelihood:                 10.568
No. Observations:                  16   AIC:                             8.864
Df Residuals:                       1   BIC:                             20.45
Df Model:                          14
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                          8.5000      0.125     68.000      0.009       6.912      10.088
C(brand, Sum)[S.Dominos]       6.661e-16      0.217   3.08e-15      1.000      -2.751       2.751
C(brand, Sum)[S.Onesta]        1.776e-15      0.217    8.2e-15      1.000      -2.751       2.751
C(brand, Sum)[S.Oven Story]       -0.2500      0.217     -1.155      0.454      -3.001       2.501
C(price, Sum)[S.$1.00]             0.7500      0.217      3.464      0.179      -2.001       3.501
C(price, Sum)[S.$2.00]        -5.995e-15      0.217  -2.77e-14      1.000      -2.751       2.751
C(price, Sum)[S.$3.00]         6.661e-15      0.217   3.08e-14      1.000      -2.751       2.751
C(weight, Sum)[S.100g]             5.0000      0.217     23.094      0.028       2.249       7.751
C(weight, Sum)[S.200g]             2.0000      0.217      9.238      0.069      -0.751       4.751
C(weight, Sum)[S.300g]            -1.2500      0.217     -5.774      0.109      -4.001       1.501
C(crust, Sum)[S.thick]             1.7500      0.125     14.000      0.045       0.162       3.338
C(cheese, Sum)[S.Cheddar]         -0.2500      0.125     -2.000      0.295      -1.838       1.338
C(size, Sum)[S.large]             -0.2500      0.125     -2.000      0.295      -1.838       1.338
```

The Ordinary Least Squares (OLS) regression results provide detailed information about the model fit and the significance of each predictor variable. Below is a breakdown of the key aspects of the summary:

Model Fit Statistics

- **Dependent Variable:** The dependent variable in this model is ranking.
- **R-squared:** The R-squared value is 0.999, indicating that 99.9% of the variance in the dependent variable is explained by the model. This suggests an excellent fit.
- **Adjusted R-squared:** The adjusted R-squared value is 0.989, which also indicates a very good fit, taking into account the number of predictors.
- **F-statistic:** The F-statistic is 97.07 with a p-value of 0.0794. The high F-statistic indicates that the model is statistically significant, but the p-value suggests that it's not significant at the conventional 0.05 level.

Coefficients and Significance

- **Intercept:** The intercept coefficient is 8.5000, which is statistically significant with a p-value of 0.009. This represents the baseline ranking when all categorical variables are at their reference levels.
- **C(brand, Sum)[S.Dominos], C(brand, Sum)[S.Onesta]:** These coefficients are extremely close to zero and are not statistically significant (p-value = 1.000). This suggests that these brands do not have a significant impact on the ranking compared to the reference brand.
- **C(brand, Sum)[S.Oven Story]:** This coefficient is -0.2500 with a p-value of 0.454, indicating it is not statistically significant.
- **C(price, Sum)[S.$1.00]:** This coefficient is 0.7500 with a p-value of 0.179, indicating it is not statistically significant.
- **C(weight, Sum)[S.100g]:** This coefficient is 5.0000 with a p-value of 0.028, indicating it is statistically significant at the 0.05 level. This suggests that a weight of 100g has a significant positive impact on the ranking.
- **C(weight, Sum)[S.200g]:** This coefficient is 2.0000 with a p-value of 0.069, which is close to the significance threshold.
- **C(weight, Sum)[S.300g]:** This coefficient is -1.2500 with a p-value of 0.109, indicating it is not statistically significant.
- **C(crust, Sum)[S.thick]:** This coefficient is 1.7500 with a p-value of 0.045, indicating it is statistically significant at the 0.05 level. This suggests that a thick crust has a significant positive impact on the ranking.
- **C(cheese, Sum)[S.Cheddar]:** This coefficient is -0.2500 with a p-value of 0.295, indicating it is not statistically significant.
- **C(size, Sum)[S.large]:** This coefficient is -0.2500 with a p-value of 0.295, indicating it is not statistically significant.
- **C(toppings, Sum)[S.mushroom]:** This coefficient is 1.1250 with a p-value of 0.070, which is close to the significance threshold.
- **C(spicy, Sum)[S.extra]:** This coefficient is 0.7500 with a p-value of 0.105, indicating it is not statistically significant.

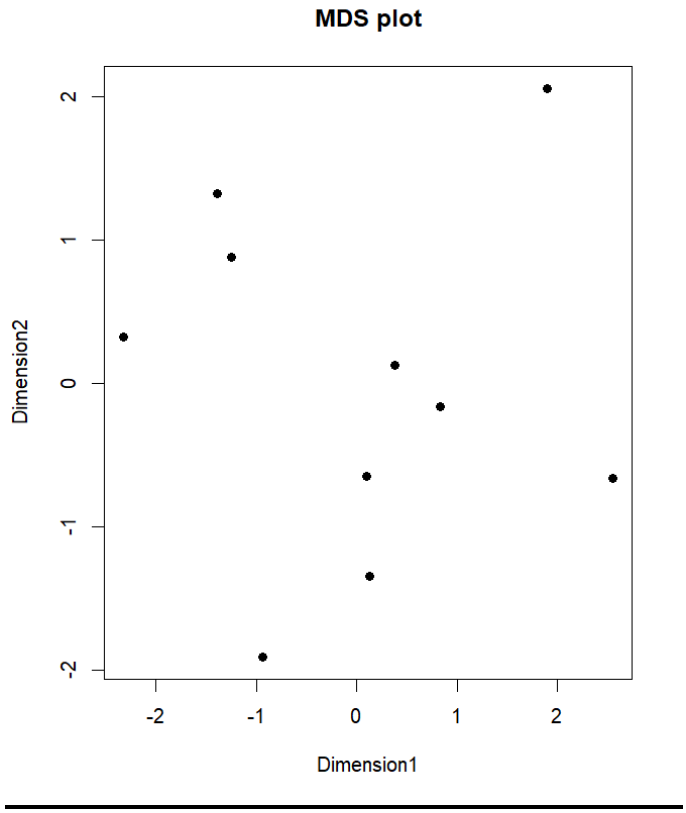
Diagnostic Tests

- **Omnibus Test:** The Omnibus test has a chi-square value of 30.796 with a p-value of 0.000, indicating non-normality in the residuals.

- **Durbin-Watson Statistic:** The Durbin-Watson statistic is 2.000, suggesting no autocorrelation in the residuals.
- **Jarque-Bera Test:** The Jarque-Bera test has a value of 2.667 with a p-value of 0.264, indicating that the residuals are normally distributed.

The regression model explains a very high percentage of the variance in the dependent variable (ranking), but several predictors are not statistically significant. Significant predictors include the intercept, 100g weight, and thick crust, which positively influence the ranking. The model diagnostics suggest some issues with residual normality, but there is no indication of autocorrelation.

Similar in R following results were obtained

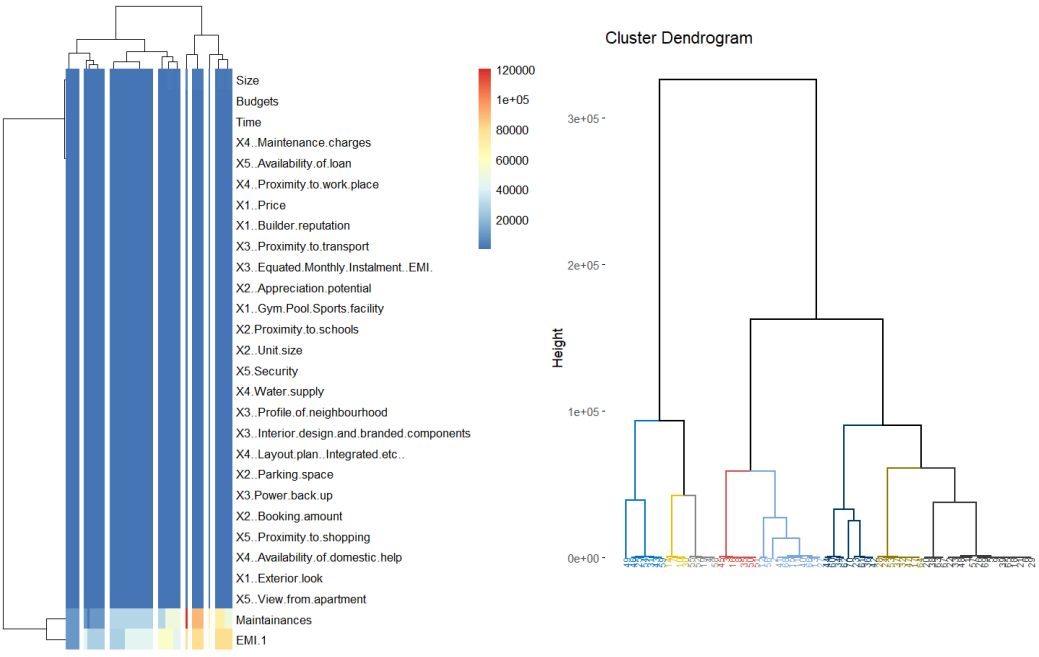


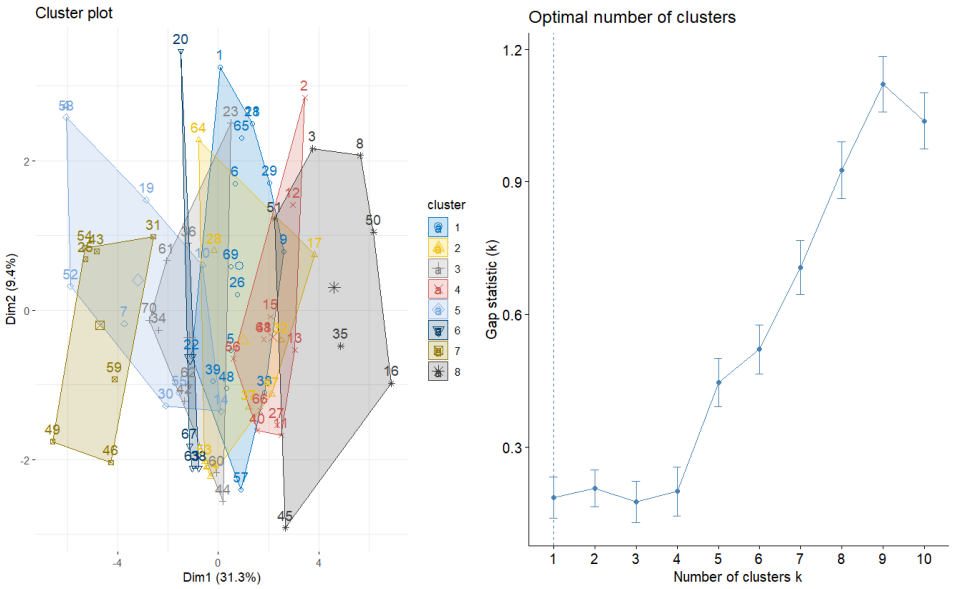
MDS Plot Interpretation: The MDS plot visually represents ice cream brands in a two-dimensional space defined by MDS Dimension 1 and MDS Dimension 2. Each point on the plot represents a brand, with distances between points reflecting the similarity or dissimilarity of brands based on scaled numerical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

Clusters and Proximity:

- **Kwality, Arun, and Vadilal:** These brands cluster closely in the lower left quadrant, indicating similar feature profiles.
- **Nandini, KVAFSU, and Joy:** Positioned in the upper right quadrant, forming another distinct group with similar features different from the first cluster.
- **Hatson:** Positioned near the center-left, sharing some similarity with the Kwality-Arun-Vadilal cluster but distinct enough to stand alone.
- **Vijaya:** Located at the top center, standing uniquely apart from other brands.
- **Dodla and Amul:** Found in the lower right quadrant, suggesting shared characteristics with each other but distinct from other clusters.

Brands like Vijaya and Joy are isolated from others, indicating unique feature sets that differentiate them in the market. Dodla and Amul, while close, are also distant from other clusters, suggesting shared yet distinct characteristics. Proximity on the MDS plot can signify direct competition in the market among brands with similar product features, whereas greater distances may reflect differing market segments or unique selling propositions.





This plot serves to determine the optimal number of clusters (k) for k-means clustering. The x-axis denotes the number of clusters, while the y-axis depicts the gap statistic. Typically, the optimal number of clusters corresponds to the highest gap statistic value, which, in this instance, suggests k = 8.

The plot integrates a heatmap and a dendrogram to illustrate hierarchical clustering results. The dendrogram on the left delineates the hierarchical structure of observations, organized into eight clusters. The heatmap portrays variable values for each observation, with color intensity indicating magnitude.

- Optimal number of clusters (k): 8
- Cluster visualization: The clusters exhibit clear differentiation in the plot, with the heatmap and dendrogram offering detailed hierarchical insights.

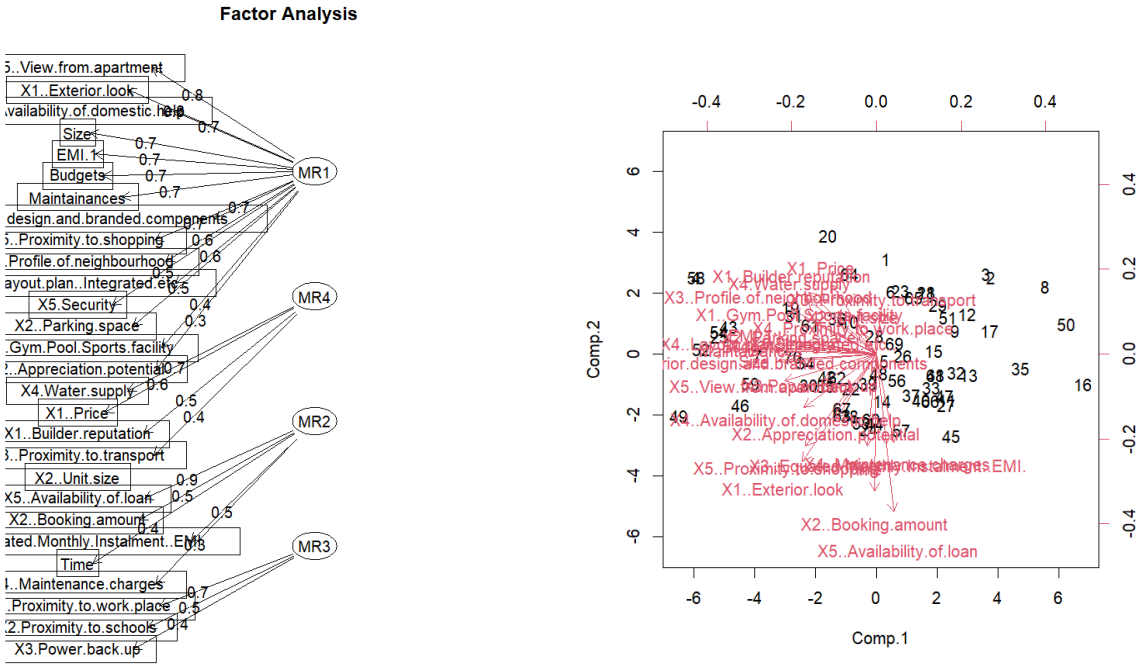
Each cluster is depicted using distinctive shapes and colors:

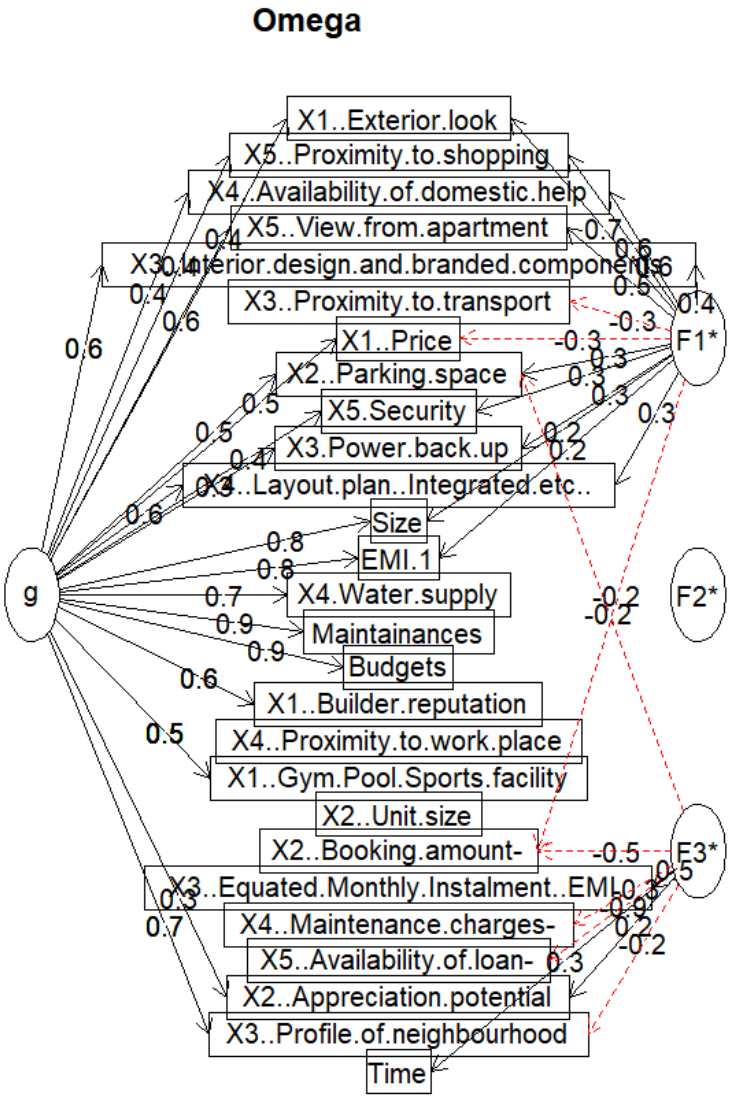
- Cluster 1: Dark blue circles, representing well-defined data points.
- Cluster 2: Yellow triangles, indicating a large area with variability.
- Cluster 3: Light grey squares, overlapping significantly with other clusters.
- Cluster 4: Red crosses, relatively small and overlapping with clusters 3, 5, and 6.
- Cluster 5: Light blue diamonds, showing significant overlap with other clusters.
- Cluster 6: Dark grey upward triangles, similar in size and position to cluster 4.

- Cluster 7: Black X marks, overlapping with clusters 3 and 5, with less distinct boundaries.
- Cluster 8: Light grey asterisks, overlapping significantly with other clusters, indicating less distinct separation.

Clusters 3, 4, 5, 6, and 7 exhibit notable overlap, suggesting shared similarities among data points. Cluster 1 stands out with minimal overlap, indicating clear separation from other clusters. Cluster 2 shows some overlap but maintains a relatively distinct boundary.

This plot provides insights into the clustering structure, highlighting Cluster 1 as the most distinct, while other clusters face challenges in clear differentiation based solely on the first two principal components.



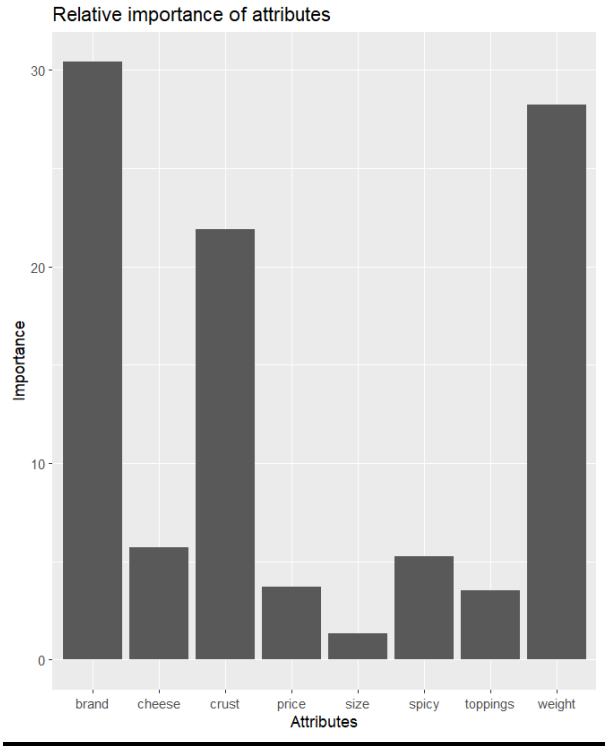


Factor analysis identified several components (RC1 to RC5) based on variable loadings, revealing underlying relationships among observed variables.

- **RC1:** Variables like Builder reputation, Size, Budgets, Maintenance, and EMI 1 show strong loadings, indicating a relationship with the first principal component, possibly representing overall property appeal.
- **RC2:** High loadings on Booking amount, Equated Monthly Installment (EMI), and Availability of loan suggest a focus on financial aspects.
- **RC3:** Variables Proximity to transport and Water supply show high loadings, focusing on infrastructure and utilities.
- **RC4:** Loadings on Proximity to workplace and Power backup suggest a component related to accessibility and amenities.

- **RC5:** Security and Exterior look show high loadings, indicating a focus on safety and aesthetics.

Factor analysis illuminates how variables contribute to underlying components, providing insights into property attributes valued by potential buyers.



Comparison and Insights:

- **Brand:** Although Pizza Hut is preferred overall, Oven Story achieves the highest utility score, suggesting other attributes in Oven Story's profile contribute significantly to its preference.
- **Price:** Despite a preference for $1.00, the highest utility profile costs $4.00, indicating price isn't the dominant factor.
- **Weight, Crust, Cheese, Size, Toppings, Spicy:** Consistent preferences across profiles for weight (100g), thick crust, Mozzarella cheese, mushroom toppings, and extra spiciness indicate these attributes hold high value for consumers.

Conjoint analysis highlights that while some attribute preferences (brand, price, size) may differ from the highest utility profile, consistency in other attributes shows their importance in consumer choice.

RECOMMENDATIONS

Based on the findings, the following recommendations are proposed:

Survey.csv Analysis:

- **Segmentation Strategy:** Use cluster analysis to identify distinct respondent segments based on demographic and survey response data. Tailor marketing strategies and product offerings to each segment's preferences.
- **Feature Optimization:** Utilize PCA and FA insights to streamline survey questions or features that most strongly influence respondent perceptions or behaviors.

icecream.csv Analysis:

- **Market Positioning:** Leverage MDS findings to reposition ice cream brands based on unique feature profiles. Highlight distinct attributes to appeal to specific consumer segments or create niche offerings.
- **Competitive Analysis:** Identify competitive clusters in the MDS plot to refine marketing strategies and differentiate brands effectively in the market.

pizza_data.csv Analysis:

- **Product Optimization:** Use Conjoint Analysis results to prioritize product attributes that drive consumer preference, such as brand perception, pricing strategies, and preferred pizza configurations.
- **Marketing Messaging:** Tailor marketing campaigns to emphasize preferred attribute combinations identified through Conjoint Analysis, optimizing messaging to resonate with target consumer preferences.