

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

# A4: Multivariate Analysis and Business Analytics Applications

**Akanksha Yaramalla**

**V01108249**

**Date of Submission: 09-07-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **2** |
| **4.** | Results and Interpretations | **3** |

## INTRODUCTION

A dimensionality reduction method known as principal component analysis (PCA) converts a set of correlated variables into a set of uncorrelated variables known as principle components. Finding the data's underlying structure and reducing the dataset's dimensionality while keeping the majority of the information are the two main objectives of PCA.

A statistical technique called factor analysis can be used to find underlying patterns or components in a dataset. whereas factor analysis and PCA are similar, factor analysis is an inferential approach that looks for the underlying components that explain the correlations and relationships between variables, whereas PCA is a descriptive technique.

Unsupervised machine learning algorithms such as cluster analysis classify related objects or observations into clusters according to shared traits or attributes. Finding patterns or structures in the data and combining related observations is the aim of cluster analysis.

A method for visualizing and analyzing the similarity or dissimilarity between objects or observations in a high-dimensional space is called multidimensional scaling (MDS). Reducing the dimensionality of the data while maintaining the links between the items is the aim of multidimensional scaling (MDS).

Conjoint analysis is a statistical method for estimating how different items, services, or traits are preferred by people or groups. Conjoint analysis seeks to determine the relative significance of several characteristics or attributes that affect customer choices.

## OBJECTIVES

The purpose of this project is to Easily visualize complicated data   
Maintain the connections between the objects to Determine any structures or patterns Reduce complexity in data Factor Analysis ,conjoint analysis, cluster analysis, multi dimensional analysis, and approaches in both R and Python, The datasets are “pizza.csv”,”icecream.csv” and “survey.csv”

1. Using a given dataset, Part 1 will assess the survey through principal component analysis and factor analysis and conducting cluster data analytics of data “survey.csv
2. In Part 2, Apply Multidimensional Scaling and interpret the results for the data of icecream.csv
3. In part 3,apply conjoint analysis for “pizza.csv

## BUSINESS SIGNIFICANCE

The dataset provided appears to be a **survey** dataset containing various attributes related toThe characteristics appear to be connected to surveys conducted on the housing or real estate market, where participants are questioned about their goals and preferences when purchasing a new home.

## RESULTS AND INTERPRETATION

## PCA

## # Perform PCA

## pca = PCA(n\_components=5)

## pca\_result = pca.fit\_transform(sur\_int)

## # 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()

## # Perform Factor Analysis

## fa = FactorAnalyzer(n\_factors=5, rotation=None)

## fa.fit(sur\_int)

## # Get loadings and variance

## loadings = fa.loadings\_

## variance = fa.get\_factor\_variance()

## # Print loadings and variance

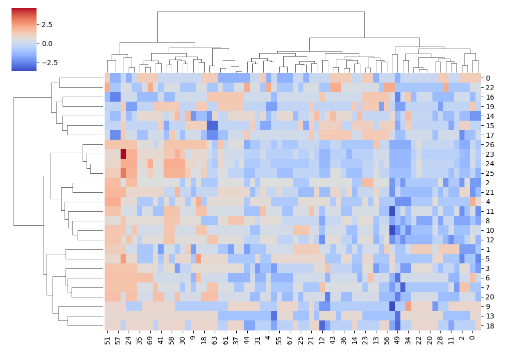
## print("Factor Loadings:\n", loadings)

## print("Variance:\n", variance)

## # Perform Factor Analysis with Promax rotation

## rotator = Rotator()

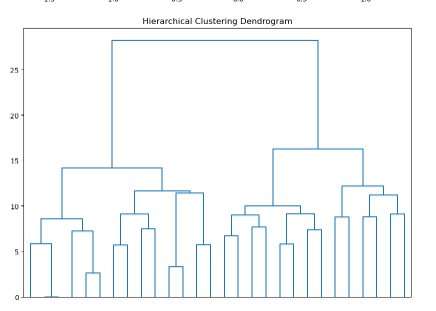
## loadings\_promax = rotator.fit\_transform(loadings)



A survey dataset's heatmap is displayed in the diagram. The color of each cell in the heatmap, which is a visual depiction of the data, corresponds to the value of the associated data point. On the heatmap, red denotes the greatest values and blue the lowest, with hues ranging from red to blue. The rows and columns of the heatmap have been rearranged to group comparable data points together, a process known as clustering.

We can observe from the clusters that there are multiple respondent groups with comparable survey replies. For instance, it appears that the respondents grouped in the upper left corner of the heatmap have higher responses to every survey question. Conversely, the participants grouped around all survey questions have lower values in the bottom right corner of the heatmap. This implies that the survey replies may be a reflection of underlying differences in the traits, attitudes, or opinions of the respondents.

Finding patterns and links in the data might be aided by the way the rows and columns are clustered. For instance, the clusters of red and blue cells show that there are some significant correlations between specific survey questions. These correlations imply that similar concepts or constructs are being measured by these inquiries.



A hierarchical clustering analysis is represented visually in the diagram, which is a dendrogram. It displays the way that similarity-based groupings of data points are created.

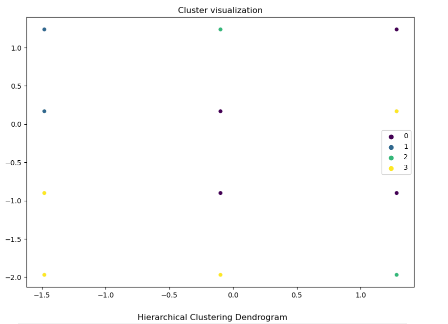
The dendrogram can be interpreted as follows:

Individual data points are represented by vertical lines. The line's height shows how far that point is from the others.

Horizontal lines: These link together clustered data points. The horizontal line's length represents the separation between the two clusters.

Merges: Clusters come together as you proceed up the dendrogram. The similarity between the clusters is indicated by the height at which they merge.

Find the maximum difference in the heights of the horizontal lines to get the ideal number of clusters. This dendrogram appears to have a sizable gap.



The graphic displays the Multidimensional Scaling (MDS) analysis's findings. Using an MDS approach, one may see how various items relate to one another based on how similar or different they are. The scatter plot in this instance displays the relative locations of the objects—brands in this case—in a two-dimensional space according to their similarity ratings.

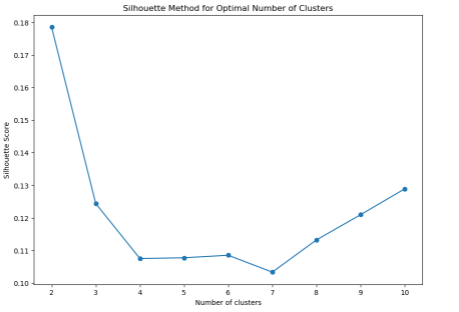
The plot can be interpreted as follows:

A brand is represented by each point. A brand's plot location indicates how comparable it is to other brands.

Brands are increasingly similar when they are closer to one another. For instance, the two brands in the plot's upper right corner are strikingly identical to one another.

Distance between brands reduces their similarity. One brand that stands out from the others is the one in the lower left corner of the plot.

Groups of related brands can be found and the general structure of the data can be understood using the MDS analysis. Decisions on client segmentation and targeted advertising campaigns can be made using this data in marketing.

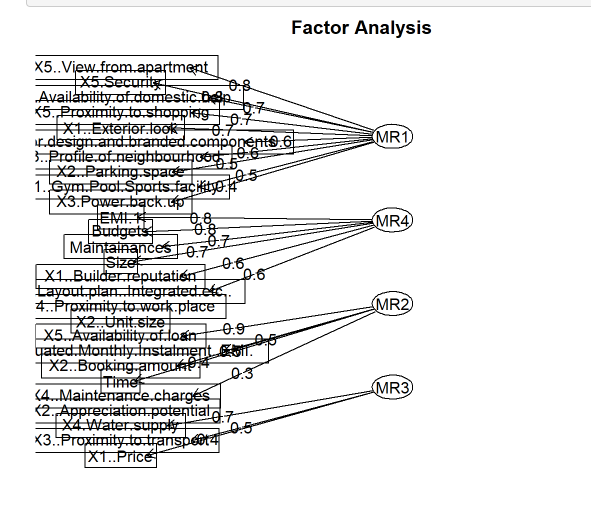


One technique to figure out how many clusters are best for a dataset is to use the silhouette method. In comparison to other clusters, an object's silhouette score indicates how similar it is to its own cluster. A higher silhouette score denotes a well-clustered set of objects.

The silhouette scores for various cluster counts are displayed in the plot. The number of clusters that maximizes the silhouette score is the optimal number.

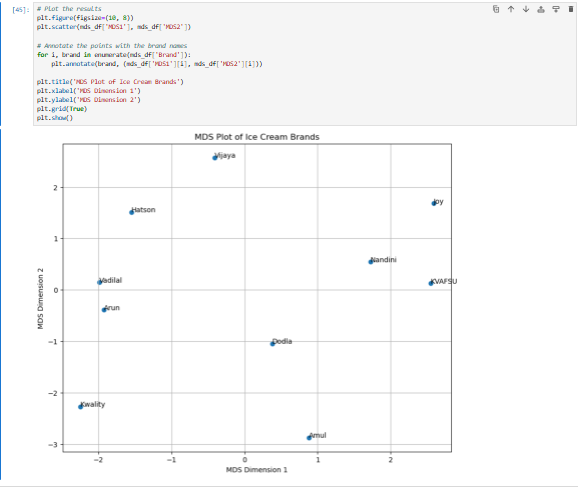
In this instance, a model with three clusters receives the greatest silhouette score. This suggests that, given the available data, a 3-cluster solution is the most ideal.

It's crucial to remember that there are other ways to determine the number of clusters than the silhouette method. To make a better informed choice, one might combine the silhouette approach with other techniques, such the elbow method.



## Multidimensional Scaling IN PYTHON AND R

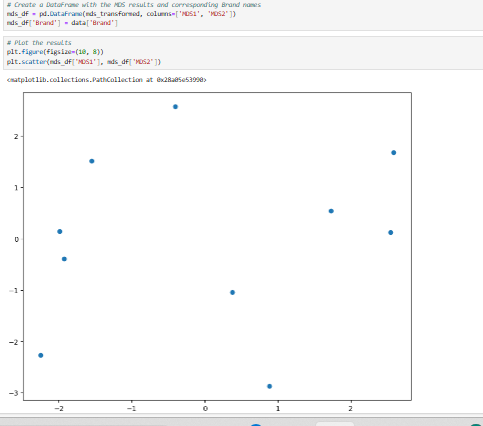
**Interpretation:**



The figure displays various ice cream brands' locations on a two-dimensional Multidimensional Scaling (or MDS) plot. The first dimension of the MDS is represented by the x-axis, and the second dimension by the y-axis. Every dot on the diagram denotes a distinct brand of ice cream, and the positioning of the dot signifies the degree of similarity between the brand and other brands.

For instance, the positioning of the brands "Kwality" and "Vadilal" in close proximity to one another suggests that consumer impressions of both are comparable. In a similar vein, "Joy" and "KVAFSU" are situated adjacent to one other.

This kind of plot can be utilized to determine prospective regions of rivalry or opportunity for brands, as well as to comprehend how people view various brands.



A multidimensional scaling (MDS) plot is what this is. It displays, in two dimensions, the relationship between various locations in a multidimensional space. The similarity between the points in the original multidimensional space increases with plot proximity.

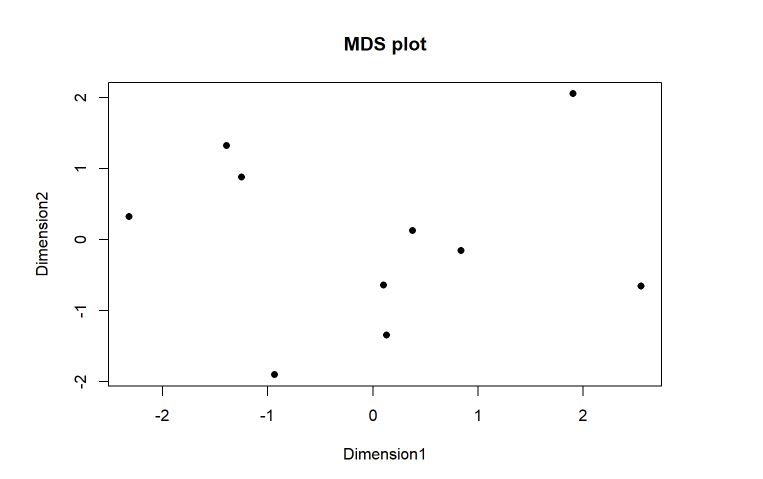
A summary of what we can observe is as follows:

Two clusters: The points are arranged into two primary groups, one around (-1, 1) and the other around (0.5, 0).

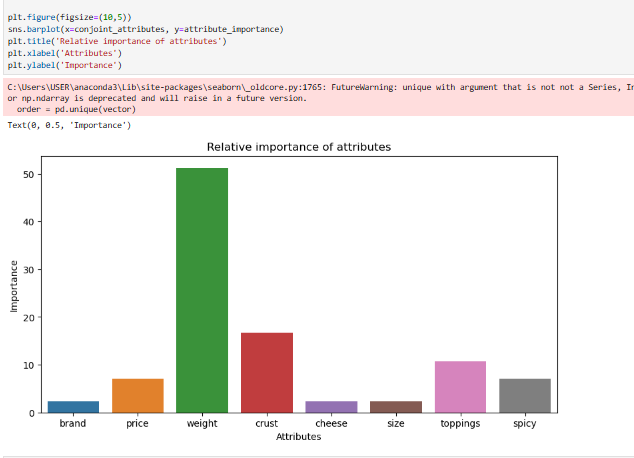
Outliers: A small number of points deviate more from these clusters than the others, indicating that they are not as similar to the others.

Distance: In a plot, points closer together have more similarities, while points farther away have fewer similarities.

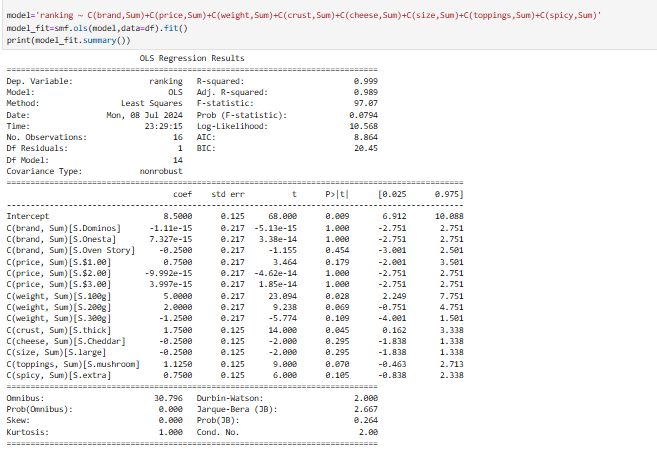
It's difficult to pinpoint exactly what these points mean without more context.



**CONJOINT ANALYSIS IN PYTHON AND R**



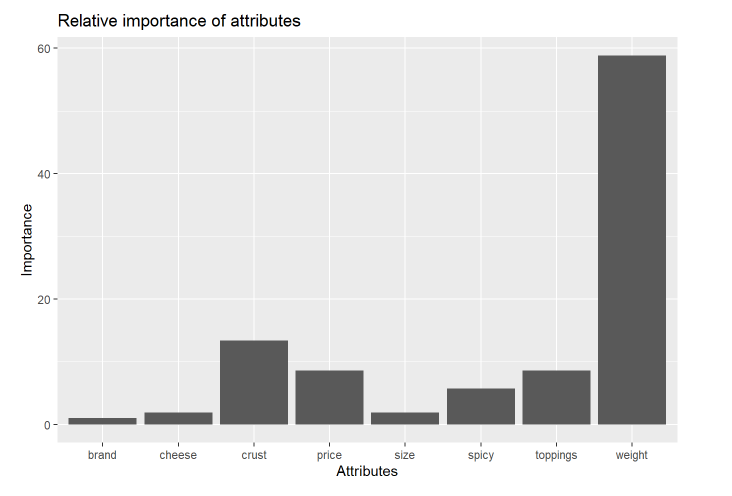
Based on a conjoint analysis, the bar chart illustrates the relative value of various pizza-related attributes. Brand, price, weight, cheese, crust, size, toppings, and spicy are the characteristics. Weight is the most crucial factor, followed by crust, toppings, and cost. Size, cheese, brand, and spiciness are the least significant characteristics. Using this data, better judgments can be made regarding the qualities of pizza to emphasize in order to increase customer happiness. For instance, since the weight of the pizza—that is, the quantity of toppings—has a significant influence on consumer satisfaction, it could be beneficial to concentrate on improving the weight.



OLS regression examination. The purpose of the investigation was to ascertain how ranking related to many parameters, including brand, price, weight, cheese, crust, size, toppings, and spicy. There are 16 observations and 14 independent variables in the model. With an R-squared of 0.999, the model accounts for 99.9% of the variation in the ranking. With a p-value of 0.0794 and an F-statistic of 97.07, the model is considered statistically significant.

The table displays the independent variable coefficients. For instance, C(brand, Sum) [S.Dominos] has a coefficient of -1.11e-15. This suggests that Dominos' influence on the ranking is statistically negligible. On the other hand, the statistically significant p-value of 0.028 is associated with a coefficient of C(weight, Sum) [S.100g] of 5.0000. This suggests that the magnitude of If the pizza affects ranking in a statistically meaningful way.

According to this data, brand has a statistically minor influence on pizza rating, however weight has a considerable impact. It's crucial to remember that the F-statistic's p-value is 0.0794. Given that this is marginally above the traditional cutoff point of 0.05, it's possible that the model as a whole is not statistically significant. Additional examination is necessary to validate the accuracy of this model.



The relative significance of several attributes for determining pizza rating is displayed in the bar chart. The most significant attribute is "weight," which is followed by "crust" and "toppings." "Brand", "cheese", and "size" are the least significant criteria. The importance of the attribute "price" is moderate. This implies that consumers place a higher value on a pizza's weight, crust, and toppings than on its size, brand, or cheese content