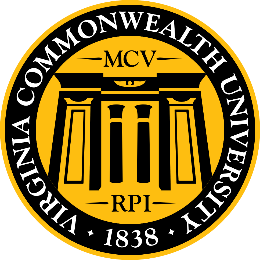
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

# A4: Multivariate Analysis and Business Analytics Applications

**KATHRYN SHAJU**

**V01111327**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Objectivise | **2** |
| **3.** | Business Significance | **2** |
| **4.** | Results & Interpretations | **3-8** |

**INTRODUCTION**

Our main focus for this study is to conduct Multivariate analysis and the business applications on data sets of Surveys.csv, pizza\_data.csv and icecream.csv. These data sets are used to conduct various multivariate analysis such as Principal Component Analysis and Factor Analysis to identify data dimensions, similarly to Conduct Cluster Analysis to characterize respondents based on background variables and apply Multidimensional Scaling and also conduct conjoint analysis. We will be testing the analysis in both R and Python soft wares, to get more accurate results.

A strong statistical method for lowering a dataset's dimensionality while maintaining as much variability as possible is principal component analysis (PCA). The original variables are converted into a new collection of uncorrelated variables known as principle components. These are arranged so that the majority of the variation found in the original dataset is retained in the first few variables.

Similarly Factor Analysis seeks to identify "factors," or latent variables, that account for the patterns of correlations between the variables that are observed. These variables are influenced by underlying dimensions, or factors, that are not directly measurable.

When a set of items can be grouped using a technique called cluster analysis, sometimes referred to as clustering, so that the objects in the same group (referred to as a cluster) are more similar to one another than to the objects in other groups. Using similarity or distance measures, this unsupervised learning technique aids in the discovery of naturally occurring groupings within the data.

Multidimensional Scaling (MDS) is used to represent the degree of similarity or dissimilarity between data as distances in a geometric space. Putting each object in a low-dimensional space—usually 2D or 3D—so that the distances between points there as nearly resemble the original dissimilarities or similarities as feasible is the aim of multidimensional scaling (MDS).

Conjoint analysis is employed in market research to comprehend consumer preferences and the relative significance of various product qualities. By calculating the value or utility that customers place on each feature, it assists firms in designing goods and services that are more closely in line with what customers want.

**OBJECTIVES**

1. Perform Principal Component Analysis and Factor Analysis to identify data dimensions
2. Conduct Cluster Analysis to characterize respondents based on background variables

3. Apply Multidimensional Scaling and interpret the results

4. Conjoint Analysis

**BUSINESS SIGNIFICANCE**

Multivariate analysis that has been conducted on this study has different sets of Business applications, all of which allow the business to make competent and accurate decision.

In a dataset comprising several variables that gauge various features of a product, including size, weight, cost, like the survey data set. To gain a deeper understanding of the underlying patterns, we should consider reducing the number of variables. By lowering the number of variables, PCA simplifies the dataset, making it simpler to view and analyze. PCA is a flexible method that is frequently used for data exploration and preparation in domains including bioinformatics, machine learning, and data science.

To determine the underlying personality traits (such as agreeableness and extraversion) that account for the answers to the questionnaire items, factor analysis can be employed. In order to minimize the dimensionality of data, enhance interpretability, and reveal the underlying structure of data, factor analysis is frequently utilized in the domains of psychology, sociology, marketing, and finance.

Imagine a marketing scenario in which a business want to divide up its clientele into groups according to the Survey data set's purchasing patterns. Customers with comparable buying habits can be grouped together using cluster analysis, which enables the business to customize its marketing tactics for various clientele groups. Many disciplines, including marketing, biology, finance, and the social sciences, employ cluster analysis extensively to examine data, spot trends, and come to well-informed conclusions based on the data's inherent groupings.

Considering the survey where rating the brands of ice cream that you favor, based on taste, price etc. Based on these ratings, MDS can be used to show how similar ice-creams are thought to be. MDS is an adaptable method that is used in many domains, including bioinformatics, psychology, and market research, to reveal the structure in intricate datasets and to display correlations in a more comprehensible way.

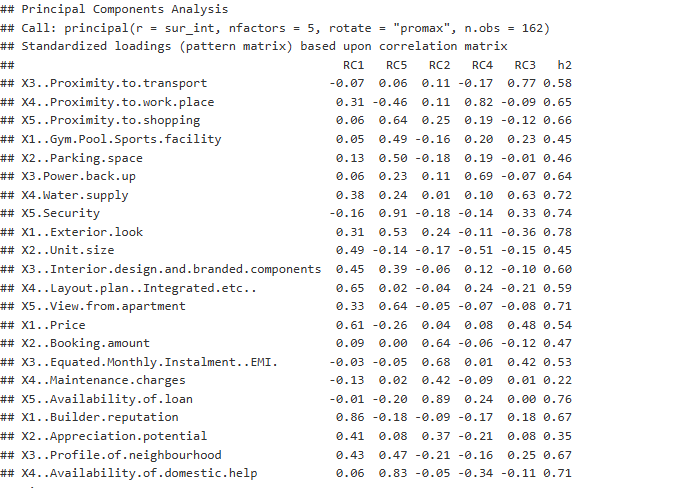
Assume about a business that want to start a new pizza delivery service. Pizza size, quality, price, brand, and quality are examples of attributes. The business can use conjoint analysis to ascertain which features customers value the most and how much they are ready to pay for each one. Conjoint analysis is frequently used in a variety of industries, such as technology, consumer goods, healthcare, and automotive, to better design goods and services that cater to the needs and preferences of customers.

**RESULTS AND INTERPRETATION**

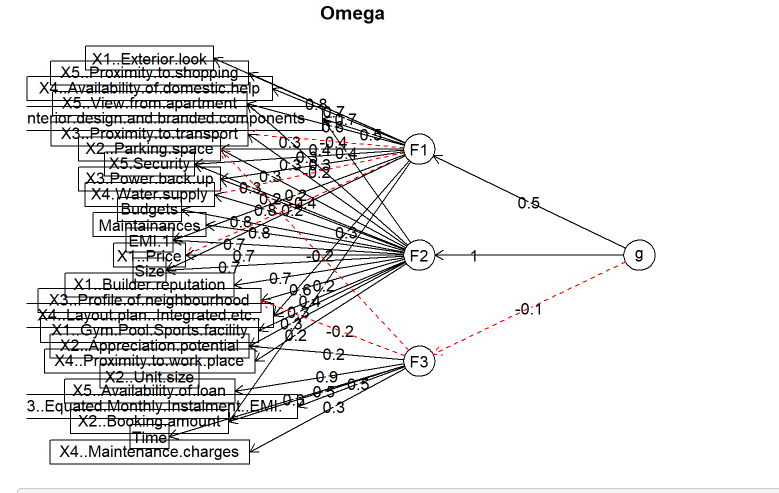
1. Perform Principal Component Analysis and Factor Analysis to identify data dimensions

Firstly to conduct the analysis, we load the data set Survey.csv. We ten proceed to clean the data, which we can use codes such as dim () and name () to identify the variables present. After which we then proceed to find the missing values, so that we may be easily be able to run our analysis and get our results.

Running the code for PCA we get,

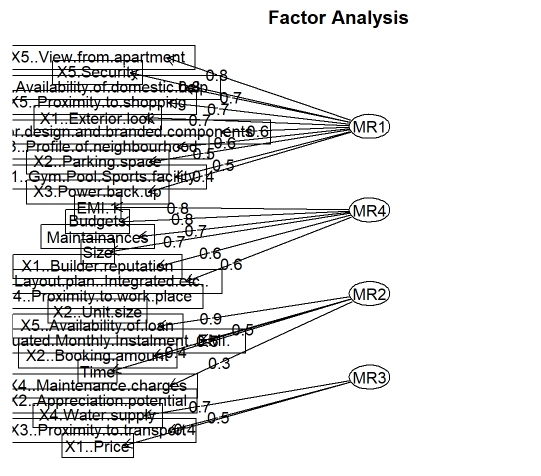


This is how we sort the variables in different categories. For our Interpretation we can also view:



Due to there being many variable we might see them being overlapped. But we can now easily identify that they have been grouped into 3 category’s F1, F2, and F3.

Similarly for factor analysis, we see:



Here we interpret that the latent factor variables can be observed. They can be dubbed as these,

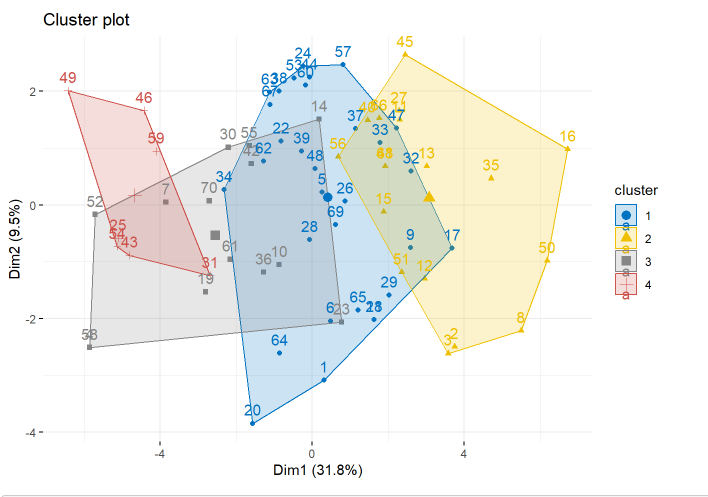
**Table 1**

|  |  |
| --- | --- |
| FACTOR ANALYSIS | LATENT VARIABLE |
| MR1 | Amenities |
| MR2 | Payments and Time Management |
| MR3 | Utilities Supply |
| MR4 | Builder Reputation |

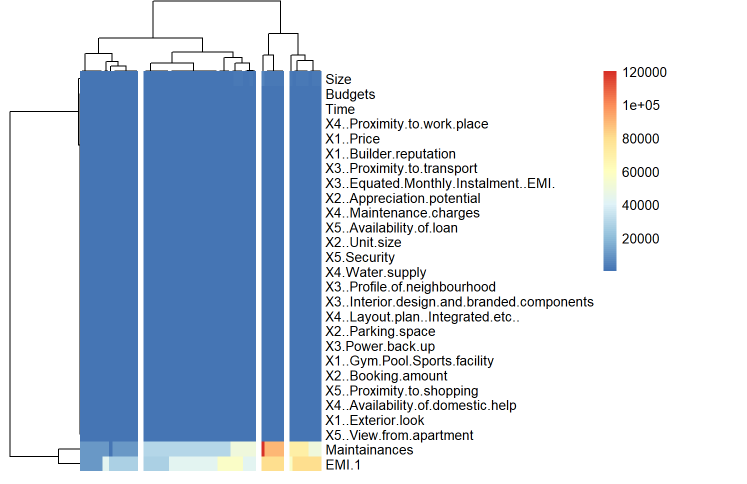
1. Conduct Cluster Analysis to characterize respondents based on background variables

Similar to PCA, here also we have cleaned and manipulated the data to reveal more accurate results.

Running the clustering code, we see that,



Here we have clustered the different variables into 4 groups, where we see red, yellow, blue and grey mapped out. Here we can figure out which variables in a group tend to have more impact on the analysis and understand the pattern that begins to emerge.

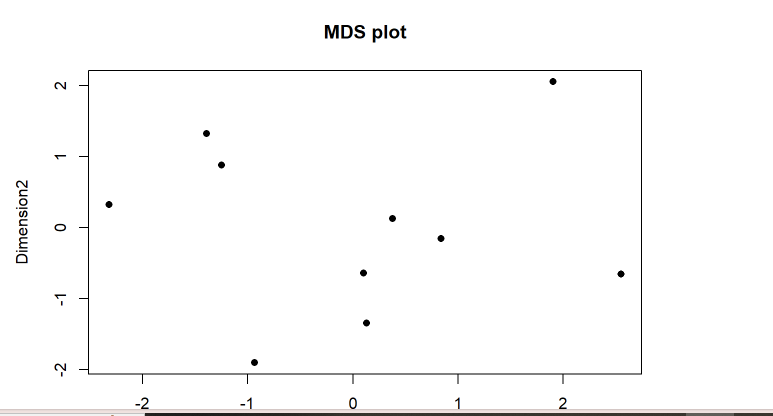


Using a heat map we can easily identify that the set of variable which are more of the latent variable Amenities seem to be more clustered and spread over, indicating the most influential group out of all.

1. Apply Multidimensional Scaling and interpret the results

Over here we focus on the Ice=cream dataset. We firstly load the data set and clean the data to further our analysis without having outliers.

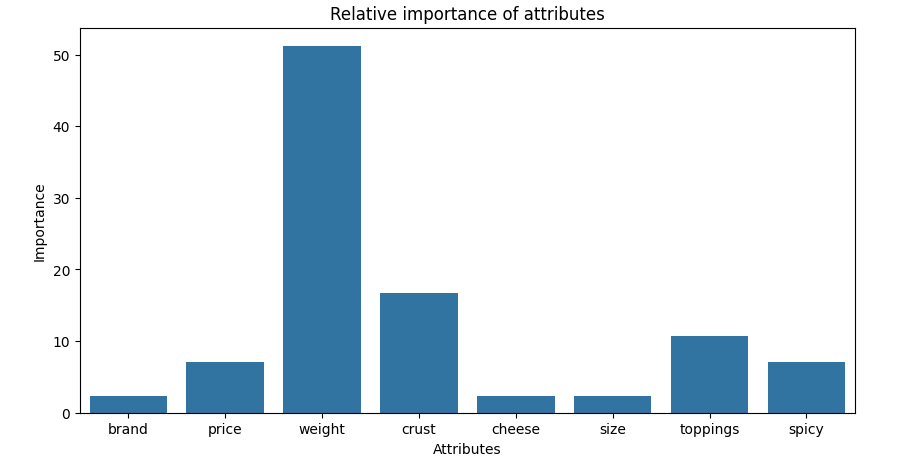
Running the Multidimensional scaling code, we see



The plot above helps us to scale the data and allows us to rate the different dimensions of the dataset which has many aspects like brand, pricing etc. to it.

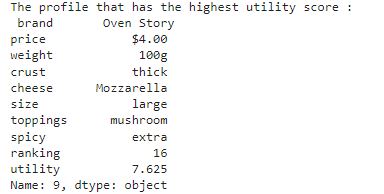
4. Conjoint Analysis

For Conjoint Analysis, we make use of the pizza\_data, dataset. Similar to the others we load and clean the data. When attribute the data such as brand, pricing, quality etc., following which we are able to conjoint the data set. We get



We can see that weight is the attribute with the highest relative importance at 51%, followed by crust at 16% and toppings at 10%. Brand, cheese, and size are the least important attributes, each at 2.38%.

We then calculate the utility profile of each score, we get,



We observe that the most utility goes to the brand Oven Story.

In this way we can analyses the company or brand making the most out of the given dataset.