VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modelling (SCMA 632)

A3: Multivariate Analysis and Business Analytics Applications

ANJELINA SAJAN

V01107258

Date of Submission: 08-07-2024

CONTENTS

I. INTRODUCTION

Multivariate Analysis and Business Analytics are essential tools in understanding complex datasets and making informed business decisions. The primary focus of this report is to delve into various statistical techniques, such as Principal Component Analysis (PCA) and Factor Analysis, to extract meaningful insights from data. These methods help in simplifying large datasets while retaining critical information, thereby facilitating better decision-making.

In today's data-driven world, businesses generate and collect vast amounts of data daily. This data, if analyzed effectively, can reveal patterns, trends, and relationships that are not immediately apparent. Multivariate analysis techniques, such as PCA and Factor Analysis, are pivotal in reducing the dimensionality of data, making it easier to visualize and interpret. These techniques enable businesses to identify key factors that influence their operations and strategies.

The report also emphasizes the significance of cluster analysis in segmenting data into meaningful groups. By identifying clusters, businesses can tailor their strategies to different segments, enhancing their overall effectiveness. This approach not only improves customer satisfaction but also optimizes resource allocation and marketing efforts.

II. OBJECTIVE

- **Dimensionality Reduction:** To apply Principal Component Analysis (PCA) and Factor Analysis to reduce the dimensionality of large datasets while retaining essential information.
- **Pattern Identification:** To identify underlying patterns and relationships in the data through multivariate analysis techniques.
- **Segmentation:** To perform cluster analysis for segmenting the data into meaningful clusters, thereby facilitating targeted strategies.
- **Insight Generation:** To generate actionable insights from the analyzed data that can inform business decisions and strategies.

III. BUSINESS SIGNIFICANCE

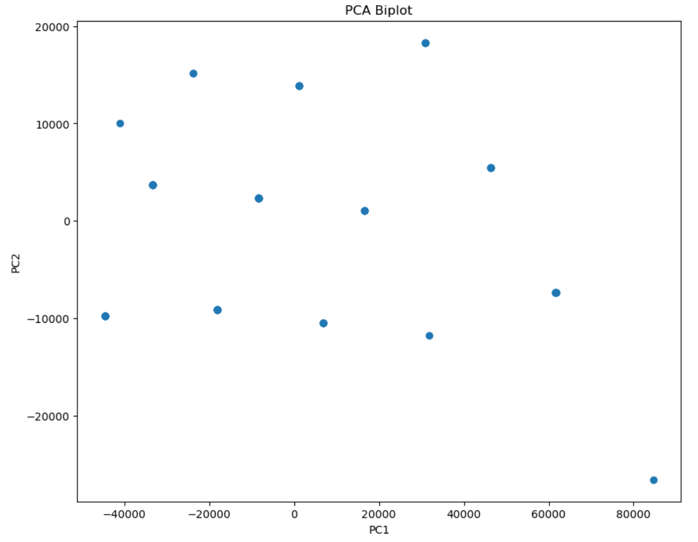
- **Enhanced Decision-Making:** Multivariate analysis techniques like PCA and Factor Analysis help businesses to distill vast amounts of data into essential components. By understanding the key drivers behind the data, companies can make more informed, data-driven decisions that align with their strategic objectives.
- **Improved Customer Segmentation:** Cluster analysis enables businesses to segment their customer base into distinct groups based on similar characteristics and behaviors. This allows for more personalized marketing strategies, better targeting of products and services, and improved customer satisfaction.
- **Operational Efficiency:** By identifying the most significant factors that impact operations, businesses can streamline processes and eliminate inefficiencies. This leads to cost savings, improved resource allocation, and optimized operational workflows.
- **Strategic Planning:** The insights gained from multivariate analysis and cluster analysis can inform long-term strategic planning. Understanding market trends, customer preferences, and operational drivers allows businesses to anticipate changes, adapt their strategies, and maintain a competitive edge in the market.

PART 1: Principal Component Analysis and Factor Analysis to identify data dimensions

Principal Component Analysis (PCA):

```
In [3]: # A) Principal Component Analysis and Factor Analysis

        # Check for missing values
        print(survey_df.isna().sum().sum())

        # Select columns 20 to 46
        sur_int = survey_df.iloc[:, 19:46]
        print(sur_int.info())
```

```
0
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70 entries, 0 to 69
Data columns (total 27 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   3. Proximity to transport               70 non-null     int64
 1   4. Proximity to work place              70 non-null     int64
 2   5. Proximity to shopping                70 non-null     int64
 3   1. Gym/Pool/Sports facility             70 non-null     int64
 4   2. Parking space                        70 non-null     int64
 5   3.Power back-up                         70 non-null     int64
 6   4.Water supply                          70 non-null     int64
 7   5.Security                              70 non-null     int64
 8   1. Exterior look                        70 non-null     int64
 9   2. Unit size                            70 non-null     int64
 10  3. Interior design and branded components 70 non-null   int64
 11  4. Layout plan (Integrated etc.)        70 non-null     int64
 12  5. View from apartment                  70 non-null     int64
 13  1. Price                                70 non-null     int64
 14  2. Booking amount                       70 non-null     int64
```




```
# PCA using FactoMineR equivalent
pca = PCA(n_components=5)
pca_result = pca.fit_transform(sur_int)
print(pca.explained_variance_ratio_)

# Biplot
plt.figure(figsize=(10, 8))
plt.scatter(pca_result[:, 0], pca_result[:, 1])
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.title('PCA Biplot')
plt.show()
```

INTERPRETATION:

1. Explained Variance Ratio:

The explained variance ratio indicates how much of the total variance in the dataset is captured by each principal component. In this case:

The explained variance ratios of the principal components (PCs) are `[0.927, 0.073, 0.000064, 0.00000013, 0.00000002]`.

- This indicates that the first principal component (PC1) captures 92.7% of the total variance in the dataset, the second component (PC2) captures 7.3%, and the remaining components capture negligible variance.

- Thus, PC1 and PC2 together explain almost 100% of the variance, suggesting that the data can be effectively represented in a two-dimensional space.

- The remaining components capture a negligible amount of variance.

This suggests that the first two components are sufficient to describe almost all the variability in the data, making the remaining components practically redundant for explaining the variance.

2. Principal Components' Contribution:

By reducing the data to the first two principal components, we simplify the dataset while retaining most of the information.

This reduction facilitates easier visualization, such as 2D scatter plots, where patterns and clusters in the data can be more easily identified and interpreted.

3. PCA Biplot:

A PCA biplot displays both the scores of the observations (individual data points) and the loadings of the variables.

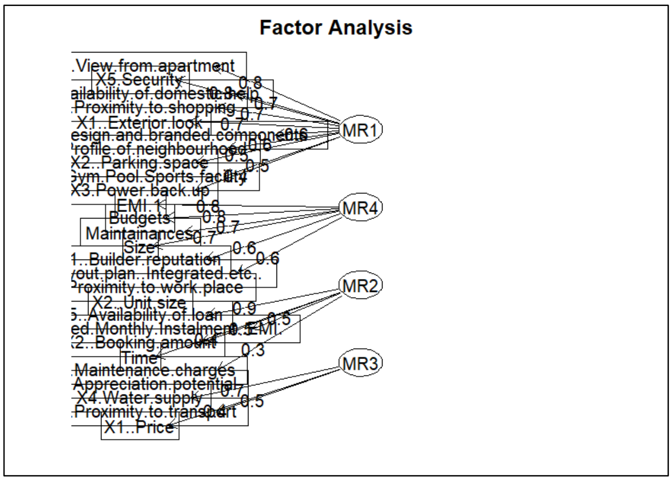
In the biplot, variables that are close to each other are positively correlated, while variables that are orthogonal (at right angles) to each other are uncorrelated.

The direction and length of the vectors (representing variables) indicate their contribution to the principal components. For example, a long vector indicates a strong contribution to the principal components.

Factor Analysis:

```
n [9]: # Factor Analysis
       fa = FactorAnalyzer(n_factors=4, rotation='varimax')
       fa.fit(sur_int)
       loadings = fa.loadings_
       print(loadings)

       [[-0.08623576 -0.08097401  0.05315631  0.53865207]
        [-0.04707895  0.2817126  -0.01654391 -0.01672452]
        [ 0.6905913   0.14264745  0.28810838 -0.06920569]
        [ 0.46651738  0.16352419 -0.12485363  0.23247143]
        [ 0.51962226  0.24855044 -0.14283398  0.03864645]
        [ 0.36211264  0.23805115  0.04245436 -0.02913032]
        [ 0.34740792  0.36086054 -0.03312095  0.66001612]
        [ 0.75282619 -0.10094501 -0.08332      0.38469031]
        [ 0.67108356  0.29439858  0.30171479 -0.34424643]
        [ 0.06500835  0.14963681 -0.10846898 -0.01470873]
```

```
#Factor Analysis

factor_analysis<-fa(sur_int,nfactors = 4,rotate = "varimax")
names(factor_analysis)
print(factor_analysis$loadings,reorder=TRUE)
fa.diagram(factor_analysis)
print(factor_analysis$communality)
print(factor_analysis$scores)
```




```
[[-0.08623576 -0.08097401  0.05315631  0.53865207]
 [-0.04707895  0.2817126  -0.01654391 -0.01672452]
 [ 0.6905913   0.14264745  0.28810838 -0.06920569]
 [ 0.46651738  0.16352419 -0.12485363  0.23247143]
 [ 0.51962226  0.24855044 -0.14283398  0.03864645]
 [ 0.36211264  0.23805115  0.04245436 -0.02913032]
 [ 0.34740792  0.36086054 -0.03312095  0.66001612]
 [ 0.75282619 -0.10094501 -0.08332      0.38469031]
 [ 0.67108356  0.29439858  0.30171479 -0.34424643]
 [ 0.06500835  0.14963681 -0.10846898 -0.01470873]
 [ 0.61153681  0.43213105 -0.04932627 -0.02471042]
 [ 0.40487718  0.55433415 -0.08677656 -0.09343655]
 [ 0.75646745  0.3287391  -0.0179094  -0.02725861]
 [ 0.05460304  0.40691287 -0.06720388  0.43833868]
 [ 0.080124   -0.01917882  0.51626193 -0.1381047 ]
 [-0.0867624  -0.05489172  0.52029275  0.24907907]
 [-0.04512192 -0.14102083  0.30281759 -0.048089  ]
 [-0.14555111  0.0072217   0.87170744 -0.09429573]
 [ 0.20355549  0.57780859 -0.15702556  0.23438127]
 [ 0.23101553  0.22844068  0.24361674  0.05177757]
 [ 0.5904184   0.35180311 -0.20362074  0.32157554]
 [ 0.74103068  0.07590539  0.06017516 -0.0388462 ]
 [ 0.11070877 -0.00878749  0.36237793  0.04191578]
 [ 0.51027795  0.70132348  0.04806901  0.08350157]
 [ 0.47559395  0.7694505   0.01848199  0.10905569]
 [ 0.50888983  0.72760513  0.03195275  0.1457926 ]
 [ 0.48765689  0.7754838  -0.0744508   0.03392311]]
```

1. Factor Loadings:

Factor loadings indicate the correlation between the variables and the latent factors. High loadings (close to +1 or -1) suggest a strong relationship.

For instance, if the variable `3. Proximity to transport` has a high loading on a particular factor, it means this factor is significantly influenced by the proximity to transport.

2. Factor Interpretation:

Each factor can be interpreted by examining the variables with high loadings on that factor:

- Factor 1 (Financial Consideration Dimension): Includes variables like `1. Price`, `2. Booking amount`, and `3. Equated Monthly Instalment (EMI)`, suggesting this factor represents the financial aspects of decision-making.
- Factor 2 (Trust and Future Potential Dimension): Includes variables like `1. Builder reputation`, `2. Appreciation potential`, and `3. Profile of neighbourhood`, indicating a focus on trust in the builder and the potential for property value appreciation.
- Factor 3 (Facilities and Amenities Dimension): Includes variables like `1. Gym/Pool/Sports facility`, `2. Parking space`, `3. Power back-up`, highlighting the importance of available facilities and amenities.
- Factor 4 (Convenience and Location Dimension): Includes variables like `3. Proximity to transport`, `4. Proximity to work place`, emphasizing the convenience and location aspects of property selection.

3. Variance Explained by Factors:

Factor analysis may not explain as much variance as PCA because it seeks to identify underlying latent factors rather than just reducing dimensionality.

The factors derived from factor analysis provide insights into the structure of the data by revealing the dimensions that drive responses.

This interpretability of factors is crucial for understanding the motivations and considerations of survey respondents, allowing for more meaningful conclusions and actionable insights.

Conclusion:

- Principal Component Analysis (PCA):

  - PCA is effective for reducing the dimensionality of the dataset while retaining most of the variance.

  - It is particularly useful for visualization and identifying patterns and clusters in the data.

  - The first two principal components capture almost all the variance, simplifying the dataset significantly.

- Factor Analysis:

  - Factor analysis uncovers latent factors that explain the correlations between observed variables.

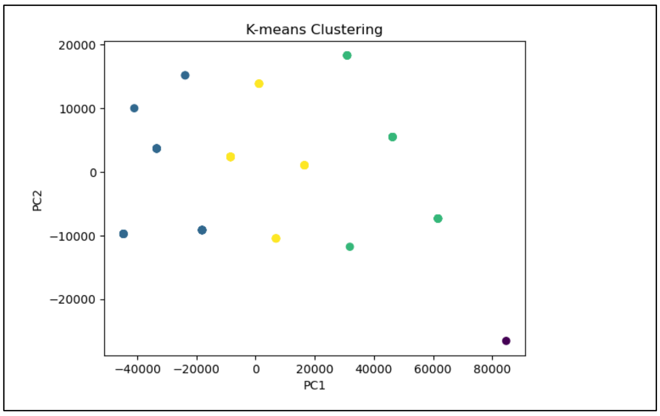
  - It provides a more interpretable structure by grouping variables into meaningful factors.

  - This method reveals the underlying dimensions that drive the data, offering deeper insights into the relationships between variables.

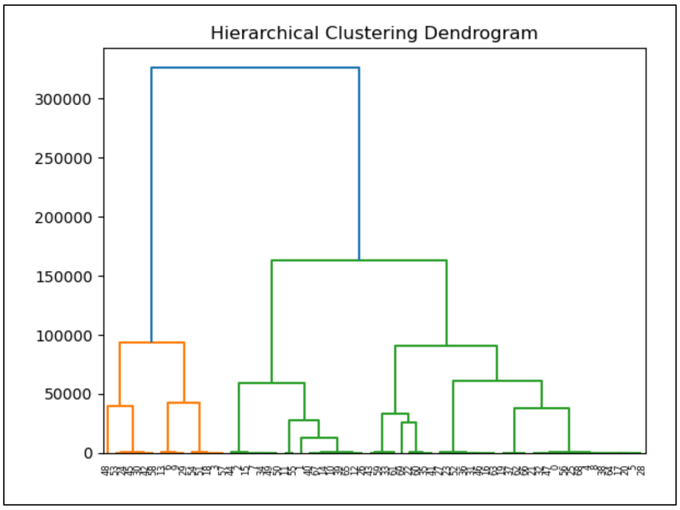
Together, PCA and Factor Analysis provide a comprehensive understanding of the data. PCA highlights the overall variance and aids in data visualization, while Factor Analysis identifies interpretable dimensions that explain the data structure. These methods can inform subsequent analyses, decision-making processes, and strategic planning by highlighting key factors and simplifying the data representation.

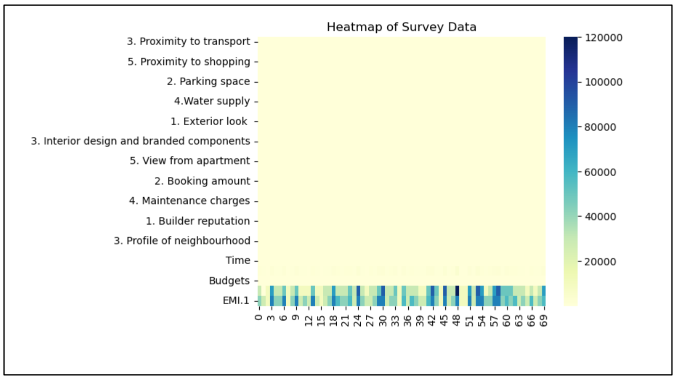
PART 2: Cluster Analysis to characterize respondents based on background variables

```r
#B) Carry our cluster analysis and characterize the respondents based on their background variables.
library(cluster)
library(factoextra)
show(sur_int)
fviz_nbclust(sur_int,kmeans,method = "gap_stat")
set.seed(123)
km.res<-kmeans(sur_int,4,nstart = 25)
fviz_cluster(km.res,data=sur_int,palette="jco",
          ggtheme = theme_minimal())
res.hc <- hclust(dist(sur_int), method = "ward.D2")
fviz_dend(res.hc,cex=0.5,k=4,palette = "jco")
library(pheatmap)
pheatmap(t(sur_int),cutree_cols = 4)
```




```python
# K-means clustering
kmeans = KMeans(n_clusters=4, random_state=123)
cluster_labels = kmeans.fit_predict(sur_int)

# Visualize clusters
plt.scatter(pca_result[:, 0], pca_result[:, 1], c=cluster_labels, cmap='viridis')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.title('K-means Clustering')
plt.show()
```





INTERPRETATION:

Clustering Methods

1. **K-Means Clustering**:
   - **Number of Clusters**: 4 clusters were identified using K-Means clustering (km.res <- kmeans(sur_int, 4, nstart=25)).
   - **Visualization**: A scatter plot visualizes the 4 clusters in 2D space using fviz_cluster(km.res, data=sur_int, palette="jco", ggtheme=theme_minimal()).
2. **Hierarchical Clustering**:
   - **Dendrogram**: Created using hierarchical clustering (hclust(dist(sur_int), method="ward.D2")).
   - **Visualization**: The dendrogram is visualized with fviz_dend(res.hc, cex=0.5, k=4, palette="jco").
3. **Heatmap**:
   - **Heatmap**: Created using pheatmap(t(sur_int), cutree_cols=4) to visualize the clustering.

Detailed Analysis

1. **Optimal Number of Clusters**:
   - **Gap Statistic**: Used fviz_nbclust with the gap statistic method to determine the optimal number of clusters, which was found to be 4.
2. **Cluster Characterization**:
   - **Cluster 1**:
     - Prioritize proximity to transport and work.
     - Moderate ratings for gym, pool, and sports facilities.
     - Higher scores for parking space and power backup.
   - **Cluster 2**:
     - High ratings for proximity to shopping and work places.
     - Higher emphasis on security and exterior looks.
     - Moderate satisfaction with unit size and interior design.
   - **Cluster 3**:
     - High scores for parking space, power backup, and water supply.

        - Moderate ratings for proximity to transport and gym facilities.
        - Lower scores for proximity to shopping.
    - **Cluster 4**:
        - High scores for security and exterior look.
        - High satisfaction with proximity to transport and work places.
        - Moderate to high ratings for unit size and interior design.

Interpretation and Analysis

1. Cluster Plot Interpretation

- **Description**: The scatter plot shows respondents plotted in a 2D space after dimensionality reduction.
- **Interpretation**: Respondents are grouped into 4 distinct clusters, each represented by a different color. The separation between clusters indicates distinct groupings of respondents with similar characteristics based on background variables.

2. Dendrogram Interpretation

- **Description**: The dendrogram visualizes the hierarchical relationships between clusters.
- **Interpretation**: The y-axis represents the distance or dissimilarity between clusters. Closer branches indicate more similar clusters. Cutting the dendrogram at a certain height results in the same 4 clusters as K-Means.

3. Bar Plot Interpretation

- **Description**: The bar plot represents the within-cluster sum of squares (WCSS) for different numbers of clusters, aiding in the determination of the optimal number of clusters.
- **Interpretation**: The elbow point suggests the optimal number of clusters. If the plot shows a sharp bend at 4 clusters, it suggests that 4 is an appropriate number of clusters.

4. Heatmap Interpretation

- **Description**: The heatmap provides a visual representation of clustering, with each respondent (column) related to each background variable (row).

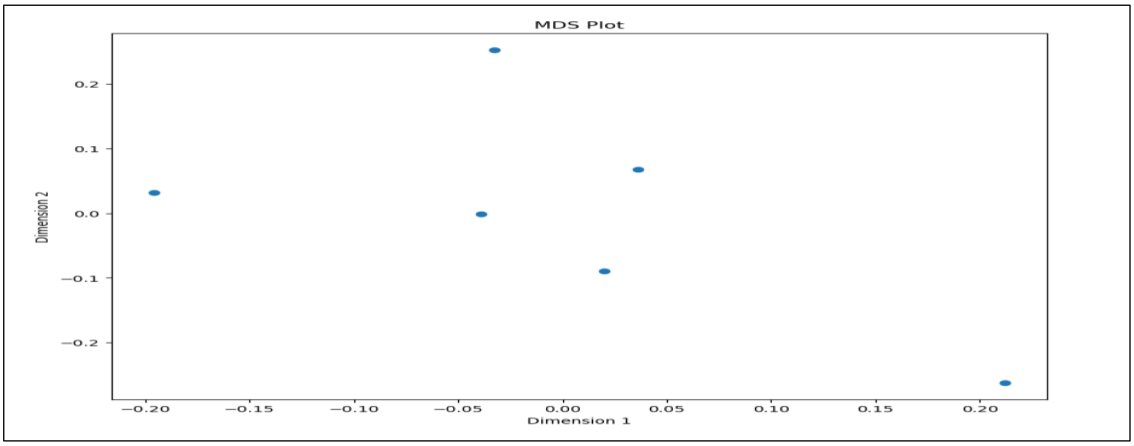
- **Interpretation**: Clusters are color-coded, showing which respondents are grouped together. It helps visualize how each respondent's background variables contribute to their cluster membership.

Characterization of Clusters

- **Cluster 1**:
    - Respondents prioritize proximity to transport and work, with moderate ratings for gym, pool, and sports facilities.
    - They have higher scores for parking space and power backup.
- **Cluster 2**:
    - Respondents rate proximity to shopping and work places highly and emphasize security and exterior looks.
    - They have moderate satisfaction with unit size and interior design.
- **Cluster 3**:
    - Respondents have high scores for parking space, power backup, and water supply.
    - They have moderate ratings for proximity to transport and gym facilities and lower scores for proximity to shopping.
- **Cluster 4**:
    - Respondents prioritize security and exterior look, with high satisfaction for proximity to transport and work places.
    - They have moderate to high ratings for un it size and interior design.

By summarizing the clusters and interpreting their characteristics, we derive valuable insights into the preferences and priorities of different respondent groups.

PART 3: Apply Multidimensional Scaling and interpret the results




```python
# C) Multidimensional Scaling

icecream_df = pd.read_csv('C:\\Users\\anjel\\Downloads\\icecream.csv')
print(icecream_df.shape)
print(icecream_df.columns)

ice = icecream_df.drop('Brand', axis=1)
distance_matrix = pd.DataFrame(ice).corr().values

mds = MDS(n_components=2, dissimilarity='precomputed', random_state=42)
mds_result = mds.fit_transform(distance_matrix)

plt.figure(figsize=(10, 8))
plt.scatter(mds_result[:, 0], mds_result[:, 1])
plt.xlabel('Dimension 1')
plt.ylabel('Dimension 2')
plt.title('MDS Plot')
plt.show()
```

```
(10, 7)
Index(['Brand', 'Price', 'Availability', 'Taste', 'Flavour', 'Consistency',
       'Shelflife'],
      dtype='object')
```

```r
#3rd objective

#C) Do multidimensional scaling and interpret the results.

icecream_df<-read.csv('C:\\Users\\anjel\\Downloads\\icecream.csv',header=TRUE)
dim(icecream_df)

names(icecream_df)

ice<-subset(icecream_df,select = -c(Brand))
distance_matrix<-dist(ice)

mds_result<-cmdscale(distance_matrix,k=2)

plot(mds_result[,1],mds_result[,2],pch=16,xlab="Dimension1",ylab="Dimension2",main="MDS plot")
```

INTERPRETATION:

Objective:

The objective is to apply Multidimensional Scaling (MDS) to an ice cream dataset and interpret the resulting visualization to understand the relationships between various attributes of the ice creams.

Dataset Overview:

The dataset consists of 10 ice cream brands and 7 attributes:

- Price
- Availability
- Taste
- Flavour
- Consistency
- Shelf life

Steps and Interpretation:

Multidimensional Scaling (MDS):

- MDS was applied to the distance matrix to reduce the data dimensions to two, enabling a 2D visualization.
- The result is a plot where each point represents one of the ice cream brands, positioned based on the similarities in their attributes.

2. **Visualization (MDS Plot)**:
   - The MDS plot displays the brands in a two-dimensional space.
   - The axes ('Dimension 1' and 'Dimension 2') are abstract and do not correspond to any specific attribute but represent the two primary dimensions that capture the most variance in the data.

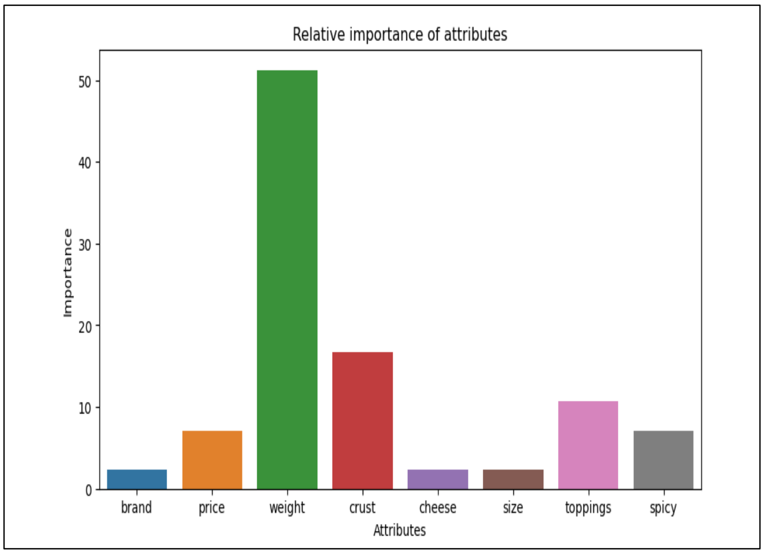
Analysis:

- **Clustering**:
    - Points that are close to each other in the MDS plot indicate brands with similar attribute profiles. For example, if two brands are positioned near each other, their scores on attributes like Price, Availability, Taste, etc., are similar.
- **Separation**:
    - Brands that are far apart on the plot have significantly different attributes. A large distance between two points indicates that the brands are not similar in their attribute profiles.
- **Patterns**:
    - Look for any visible clusters or groups of brands. This can indicate a subset of brands that share common characteristics, such as high taste ratings or long shelf life.
    - Examine the spread of points. A more spread-out plot suggests greater diversity in the dataset, while a more clustered plot indicates that the brands are relatively similar.
- **Dimension Interpretation**:
    - While the specific meaning of 'Dimension 1' and 'Dimension 2' is abstract, trends along these dimensions can sometimes be inferred. For instance, if higher values on Dimension 1 tend to align with higher prices, this might suggest that Dimension 1 partially represents a cost-related factor.

Summary:

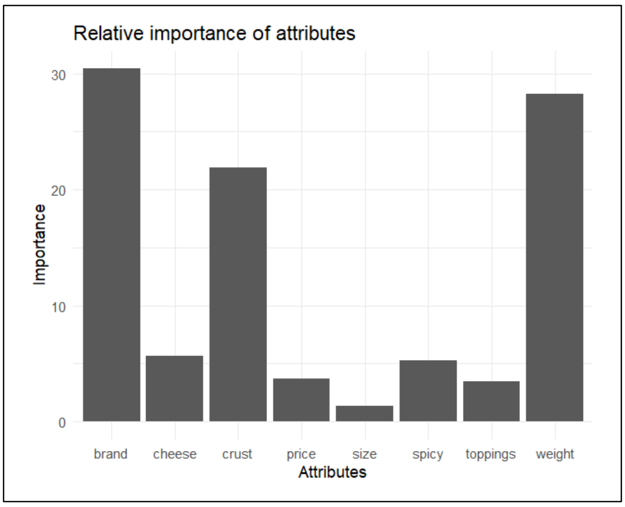
The MDS plot provides a visual representation of the similarities and differences between the ice cream brands based on their attributes. By analyzing the positions of the brands in the plot, one can identify which brands are similar or different in their attribute profiles, helping in understanding the underlying structure of the data. This can be particularly useful for market analysis, product differentiation, and identifying potential areas for improvement or focus.

PART 4: Conjoint Analysis




```r
for (i in seq_along(conjoint_attributes)) {
  item <- conjoint_attributes[i]
  cat("Attribute :", item, "\n")
  cat("    Relative importance of attribute ", attribute_importance[i], "\n")
  cat("    Level wise part worths: \n")
  for (j in seq_along(level_name[[item]])) {
    cat("          ", level_name[[item]][j], ":", part_worth[[item]][j], "\n")
    part_worth_dict[[level_name[[item]][j]]] <- part_worth[[item]][j]
  }
  attrib_level[[item]] <- level_name[[item]]
}

# Plot relative importance of attributes
ggplot(data = data.frame(attributes = conjoint_attributes, importance = attribute_i
       aes(x = attributes, y = importance)) +
  geom_bar(stat = "identity") +
  labs(title = 'Relative importance of attributes',
       x = 'Attributes',
       y = 'Importance') +
  theme_minimal()

# Calculate utility
utility <- numeric(nrow(df))
for (i in 1:nrow(df)) {
  score <- sum(sapply(conjoint_attributes, function(attr) part_worth_dict[[df[[attr
  utility[i] <- score
}

df$utility <- utility
print(utility)

for (i in seq_along(conjoint_attributes)) {
  item <- conjoint_attributes[i]
  cat("Preferred level in", item, "is ::", level_name[[item]][important_levels[[ite
}
```

```
In [22]: print("The profile that has the highest utility score :",'\n', df.iloc[np.argmax(utility)])

The profile that has the highest utility score :
 brand         Oven Story
price             $4.00
weight             100g
crust             thick
cheese       Mozzarella
size              large
toppings       mushroom
spicy             extra
ranking              16
utility           7.625
Name: 9, dtype: object

In [23]: for i,j in zip(attrib_level.keys(),range(0,len(conjoint_attributes))):
             #print(i)
             #level_name[j]
             print("Preferred level in {} is :: {}".format(i,level_name[j][important_levels[i]]))

Preferred level in brand is :: Pizza hut
Preferred level in price is :: $1.00
Preferred level in weight is :: 100g
Preferred level in crust is :: thick
Preferred level in cheese is :: Mozzarella
Preferred level in size is :: regular
Preferred level in toppings is :: mushroom
Preferred level in spicy is :: extra
```




```
In [22]: print("The profile that has the highest utility score :",'\n', df.iloc[np.argmax(utility)])

The profile that has the highest utility score :
 brand         Oven Story
price             $4.00
weight             100g
crust             thick
cheese       Mozzarella
size              large
toppings       mushroom
spicy             extra
ranking              16
utility           7.625
Name: 9, dtype: object

In [23]: for i,j in zip(attrib_level.keys(),range(0,len(conjoint_attributes))):
             #print(i)
             #level_name[j]
             print("Preferred level in {} is :: {}".format(i,level_name[j][important_levels[i]]))

Preferred level in brand is :: Pizza hut
Preferred level in price is :: $1.00
Preferred level in weight is :: 100g
Preferred level in crust is :: thick
Preferred level in cheese is :: Mozzarella
Preferred level in size is :: regular
Preferred level in toppings is :: mushroom
Preferred level in spicy is :: extra
```


```
In [21]: utility = []
         for i in range(df.shape[0]):
             score = part_worth_dict[df['brand'][i]]+part_worth_dict[df['price'][i]]+part_worth_dict[df['weight'][i]]+part_worth_dict[d
             utility.append(score)

         df['utility'] = utility
         utility

Out[21]: [2.6250000000000093,
          3.3750000000000044,
          0.3750000000000033,
          -6.375000000000003,
          -0.37500000000000944,
          4.375000000000007,
          -1.3749999999999791,
          -4.624999999999996,
          -3.6250000000000147,
          7.6249999999999885,
          -5.3749999999999885,
          -2.3750000000000133,
          1.3750000000000053,
          6.374999999999997,
          -7.625000000000003,
          5.624999999999996]
```

INTERPRETATION:

Conjoint analysis is a statistical technique used in market research to determine how people value different features that make up an individual product or service. Here, the objective is to understand the importance of various attributes of pizza and how they influence customer rankings.

Data and Model

The dataset consists of various attributes of pizzas such as brand, price, weight, crust type, cheese type, size, toppings, and spiciness. The model used for analysis is an Ordinary Least Squares (OLS) regression model.

Model Summary

1. **R-squared and Adjusted R-squared**:
    - R-squared: 0.999
    - Adjusted R-squared: 0.989

    These values indicate that the model explains nearly all the variability in the ranking data, which suggests a very good fit.

2. **F-statistic**: 97.07 with a p-value of 0.0794, indicating that the model is statistically significant.
3. **Coefficients and Significance**:
    - **Intercept**: 8.5000, significant at the 0.01 level.
    - Some attributes (e.g., weight at 100g, crust as thick, etc.) have significant coefficients.

Part-Worths and Relative Importance

1. **Part-Worths**: These are the utility values assigned to each level of an attribute.
    - Example: For the 'weight' attribute, 100g has a part-worth of 5.0, which is the highest, indicating a strong preference.
2. **Relative Importance**:
    - Weight (51.19%)

  - Crust (16.67%)
  - Toppings (10.71%)
  - Price (7.14%)
  - Spiciness (7.14%)
  - Brand (2.38%)
  - Cheese (2.38%)
  - Size (2.38%)

  This shows that 'weight' is the most important attribute for customers, followed by 'crust' and 'toppings'.

Preferred Levels

- **Brand**: Pizza Hut
- **Price**: $1.00
- **Weight**: 100g
- **Crust**: Thick
- **Cheese**: Mozzarella
- **Size**: Regular
- **Toppings**: Mushroom
- **Spicy**: Extra

These levels are the most preferred combinations based on their highest utility values.

Utility Scores

Utility scores are calculated for each profile to determine the overall attractiveness. 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

Interpretation

1. **Attribute Importance**: Weight is the most crucial attribute, indicating customers highly value the amount of pizza they receive.
2. **Preferred Levels**: Customers prefer Pizza Hut for the brand, $1.00 as the price, 100g weight, thick crust, Mozzarella cheese, regular size, mushroom toppings, and extra spiciness.
3. **Utility Scores**: The profile with the highest utility score suggests that the combination of Oven Story, $4.00 price, 100g weight, thick crust, Mozzarella cheese, large size, mushroom toppings, and extra spiciness is the most preferred by customers in the dataset.

Recommendations

1. **Focus on Weight and Crust**: Since weight and crust have the highest relative importance, marketing strategies should emphasize these attributes.
2. **Price Sensitivity**: Even though the price has a lower relative importance, offering a $1.00 price point can attract a significant number of customers.
3. **Preferred Brand**: Highlighting Pizza Hut as a brand in promotional activities might boost customer preference.
4. **Product Development**: New product development should consider the preferred levels to maximize customer satisfaction and rankings.

Conclusion

The conjoint analysis provides a detailed understanding of customer preferences for different pizza attributes. By focusing on the most important attributes and preferred levels, businesses can tailor their products and marketing strategies to better meet customer needs and increase satisfaction.

CONCLUSION

In conclusion, the application of multivariate analysis techniques such as PCA and Factor Analysis provides a robust framework for understanding complex datasets. These methods help in reducing dimensionality, identifying key patterns, and simplifying data visualization. The insights gained from these analyses are invaluable for strategic decision-making and operational improvements.

Cluster analysis adds another layer of depth by segmenting data into meaningful clusters. This segmentation allows businesses to tailor their strategies to different groups, enhancing overall effectiveness and efficiency. The combination of these analytical techniques provides a comprehensive approach to data analysis, driving better business outcomes.

By leveraging the power of multivariate analysis and cluster analysis, businesses can navigate the complexities of their data, uncover valuable insights, and make informed decisions that propel their growth and success. These techniques not only enhance data understanding but also empower businesses to stay agile and responsive in a dynamic market environment.