VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modeling (SCMA 632)

A4: Multivariate Analysis and Business Analytics Applications

by

AHAD ZIFAIN MIYANJI

V01108270

Date of Submission: 09-07-2024

CONTENTS

MULTIVARIATE ANALYSIS AND BUSINESS ANALYTICS APPLICATIONS

Introduction

Sustained success in the ever-changing corporate landscape of today depends critically on comprehending and satisfying the demands and preferences of customers. Advanced statistical techniques are being used by businesses more and more to acquire profound insights into the behavior and preferences of their customers. This project uses a variety of analytical methods to gather insightful information from survey data that may be used to inform strategic choices. In particular, we'll use:

1. To find important underlying factors and simplify the dataset, Principal Component Analysis (PCA) and Factor Analysis (FA) are used.
2. The use of cluster analysis to divide up respondents into groups according to their background characteristics, giving a distinct image of various clientele.
3. The utilization of Multidimensional Scaling (MDS) to illustrate the connections and commonalities among participants.
4. Using a pizza dataset, conjoint analysis is used to ascertain how various variables

Business Significance

Principal Component Analysis (PCA) and Factor Analysis (FA)

Business Insight: To reduce the dimensionality of enormous datasets without losing the most crucial information, PCA and FA are important. These assessments assist businesses in determining the key elements influencing consumer decisions. For instance, PCA and FA can reveal that the primary determinants of client satisfaction in the real estate industry are factors like location, pricing, and amenities. Businesses can improve client satisfaction and loyalty by concentrating on these crucial elements while developing their marketing strategy and offers.

Cluster Analysis

Business Insight: Cluster Analysis is a powerful tool for market segmentation. By grouping customers based on similar characteristics or preferences, businesses can tailor their marketing and product development efforts to address the specific needs of each segment. For instance, a company might identify clusters of customers who value luxury features, cost-efficiency, or sustainability. With this knowledge, targeted marketing campaigns can be

developed, and products can be customized to cater to the unique preferences of each segment, leading to more effective marketing efforts and higher customer retention rates.

Multidimensional Scaling (MDS)

Business Insight: MDS provides a visual representation of the similarities and dissimilarities among respondents, making it easier to identify patterns and relationships within the data. This visualization aids in understanding customer perceptions and preferences at a glance. For businesses, MDS can reveal clusters of similar preferences or highlight outliers, enabling more informed decisions regarding product development, marketing strategies, and customer service improvements. It helps in identifying potential gaps in the market and areas where the business can differentiate itself from competitors.

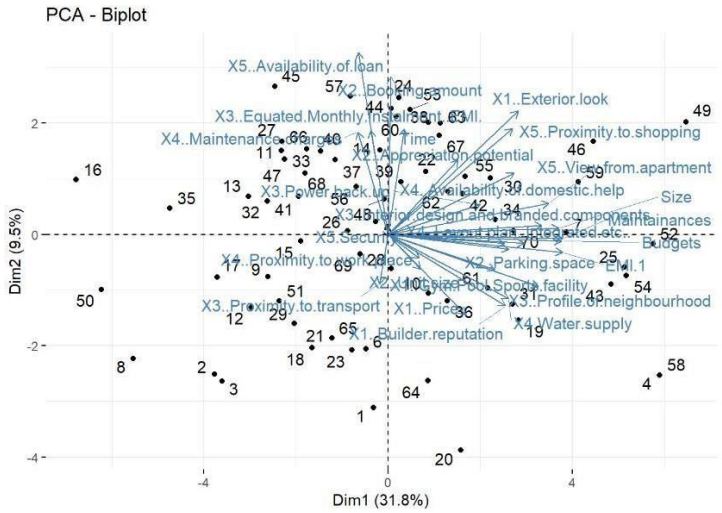
Conjoint Analysis (Pizza Dataset)

Introduction: Conjoint Analysis is used to determine how customers value different attributes of a product by evaluating their preferences and trade-offs. In this assignment, we apply conjoint analysis to a pizza dataset to understand customer preferences for various attributes such as crust type, toppings, price, and delivery time.

Business Insight: For a pizza business, understanding which attributes are most valued by customers can significantly impact product development and marketing strategies. Conjoint analysis helps in determining the optimal combination of product features that will maximize customer satisfaction and willingness to pay. For example, if customers prioritize fast delivery and specific toppings, the business can focus on improving delivery efficiency and promoting the preferred toppings. This targeted approach leads to enhanced customer satisfaction, increased sales, and a competitive edge in the market.

PRINCIPAL COMPONENT ANALYSIS

ANALYSIS USING R



R code:

```
# Perform Omega hierarchical analysis om.h <-
omega(sur_int, n.obs = 162, sl = FALSE) op <-
par(mfrow = c(1, 1))
om <- omega(sur_int, n.obs = 162)

# Perform PCA using FactoMineR package pca_fm
<- PCA(sur_int, scale.unit = TRUE)
summary(pca_fm)

# Biplot using factoextra
fviz_pca_biplot(pca_fm, repel = TRUE)

# Show the structure and dimensions of the selected
columns str(sur_int) dim(sur_int)
show(sur_int)
```

PCA Biplot Analysis

The provided PCA (Principal Component Analysis) biplot visualizes the relationships between various variables and observations in the dataset. Here's a detailed analysis of the biplot:

Principal Components

- **Dim1 (Principal Component 1)**: Explains 31.8% of the variance in the data.
- **Dim2 (Principal Component 2)**: Explains 9.5% of the variance in the data.

Together, these two dimensions explain 41.3% of the total variance, which is a significant portion but indicates that there are still other important dimensions not captured in this biplot.

Observations and Variables

- **Observations**: Represented by black points.
- **Variables**: Represented by blue arrows pointing in the direction of increasing values for that variable.

Key Insights

1. **Clusters of Observations**:
    - The observations appear to be scattered, with some clustering evident. For example, observations around coordinates (1, 1) are more densely packed, indicating similarities among them.
2. **Variable Contributions**:

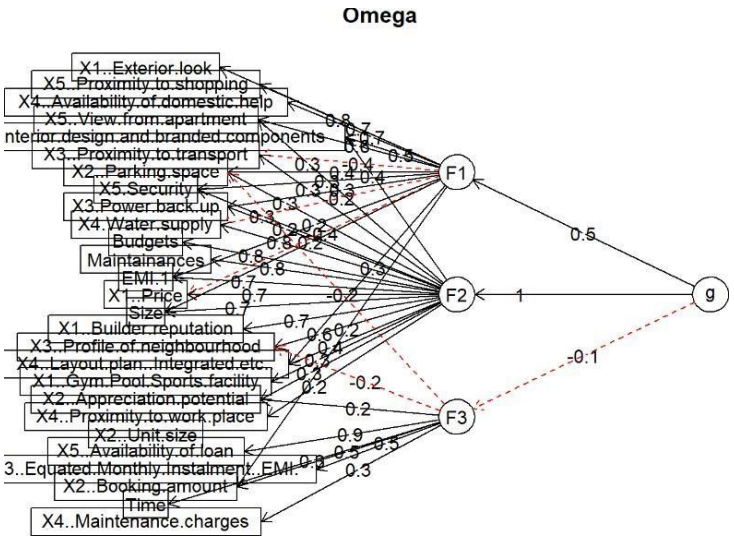
- **Dim1** is heavily influenced by variables such as "Builder reputation", "Price", "Appreciation potential", and "Proximity to transport". These variables have long arrows pointing in roughly the same direction, suggesting they contribute significantly to the variance captured by Dim1. o **Dim2** is influenced by variables like "Availability of loan", "Booking amount", and "Equated Monthly Installment (EMI)". These arrows point more towards the vertical axis, indicating their contribution to Dim2.

3. **Correlations Between Variables**:
   - Variables pointing in similar directions are positively correlated. For instance, "Builder reputation" and "Appreciation potential" point in a similar direction, suggesting that properties with a good builder reputation also have higher appreciation potential.
   - Variables pointing in opposite directions are negatively correlated. For example, "Proximity to shopping" and "Availability of loan" point in nearly opposite directions, indicating a potential negative correlation.
4. **Variable Importance**:
   - Variables with longer arrows, such as "Builder reputation", "Price", and "Proximity to transport", have a stronger influence on the principal components. These are key factors in differentiating the observations along Dim1 and Dim2.
5. **Interpreting Observations**:
   - Observations near the origin are average in terms of the measured variables.
   - Observations positioned at the ends of arrows are extreme in those variable measurements. For instance, an observation close to the tip of the "Builder reputation" arrow would likely indicate a high reputation score.

Business Implications

1. **Targeted Marketing**:
   - By understanding which variables are most influential, businesses can tailor their marketing strategies. For example, promoting "Builder reputation" and "Proximity to transport" for high-end properties.
2. **Product Development**:
   - Insights from PCA can inform property developers about the key attributes valued by potential buyers, such as "Appreciation potential" and "Price". Focusing on improving these aspects can lead to higher customer satisfaction and sales.
3. **Customer Segmentation**:
   - Clustering observations based on their PCA scores can help in identifying distinct customer segments. For instance, one segment might prioritize affordability and proximity to transport, while another values high reputation and potential appreciation.
4. **Investment Decisions**:
   - Investors can use this analysis to identify which attributes are likely to drive property appreciation. Variables like "Builder reputation" and "Appreciation potential" could be crucial in making informed investment decisions.

By leveraging PCA, businesses can gain a nuanced understanding of the underlying patterns in their data, leading to more strategic and data-driven decision-making processes.



R Code:

```
#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)
```

Omega Diagram Analysis

The provided diagram appears to be an Omega diagram, typically used in Structural Equation Modeling (SEM) to depict the relationships between observed variables, latent factors, and a general factor. Here's a detailed analysis:

Diagram Components

- **Observed Variables**: Each box on the left represents an observed variable from the dataset.
- **Latent Factors (F1, F2, F3)**: These are the underlying factors that explain the observed variables.
- **General Factor (g)**: A higher-order factor that influences the latent factors.
- **Path Coefficients**: The numbers along the arrows indicate the strength of the relationships between the variables and factors.

Key Insights

1. **Latent Factors and Observed Variables**:
    - **F1**: Strongly linked with variables such as "Availability of domestic help", "Proximity to shopping", "Parking space", "Security", and "Exterior look".

This factor seems to capture aspects related to convenience and security.

- **F2**: Associated with "Price", "Size", "Builder reputation", "Profile of neighborhood", and "Appreciation potential". This factor appears to capture value and investment potential of the properties.
- **F3**: Linked to "Equated Monthly Installment (EMI)", "Booking amount", "Availability of loan", and "Maintenance charges". This factor represents financial aspects and affordability.

2. **Relationships Among Factors**:
    - There is a strong positive relationship between F2 and the general factor (g) with a coefficient of 1.

- F1 also has a positive relationship with the general factor (g) with a coefficient of 0.5.
- F3 has a weak negative relationship with the general factor (g) indicated by the -0.1 coefficient, suggesting it might be less aligned with the overall construct represented by g.

3. **Strength of Relationships**:
   - High path coefficients (close to 1) suggest strong relationships. For example, F2's relationships with "Price" (0.7), "Size" (0.9), and "Builder reputation" (0.7) are strong.
   - Moderate coefficients (around 0.5-0.7) indicate moderate relationships, such as F1's relationship with "Availability of domestic help" (0.7) and "Proximity to shopping" (0.8). ○ Low coefficients (below 0.5) indicate weaker relationships, like some paths in F3.

Business Implications 1.

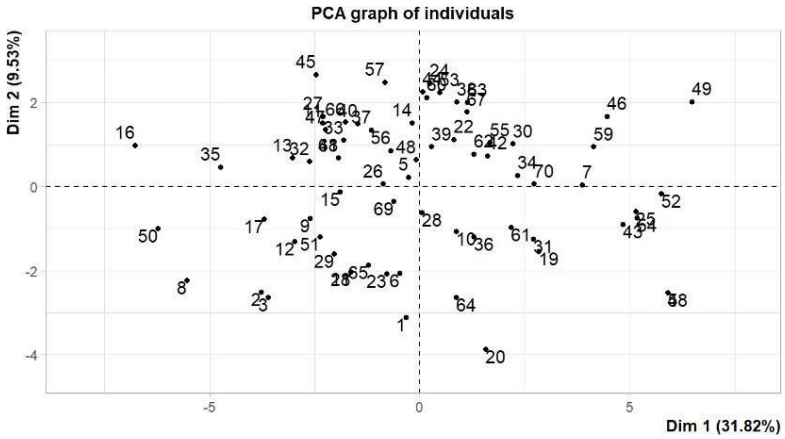
Identifying Key Drivers:

- By understanding which latent factors (F1, F2, F3) drive the observed variables, businesses can focus on enhancing these key areas. For instance, enhancing security features and convenience aspects (F1) could attract more customers.

2. **Customer Segmentation**:
   - Different customer segments may value different factors. For example, investors might be more concerned with F2 (value and investment potential), while buyers looking for affordability might focus on F3 (financial aspects).
3. **Marketing Strategies**:
   - Marketing messages can be tailored based on these insights. Highlighting property aspects like security, parking space, and proximity to amenities for one segment, and emphasizing affordability and financing options for another.
4. **Product Development**:
   - Property developers can use these insights to prioritize aspects that are most valued by potential buyers. For instance, improving the profile of the neighborhood or ensuring better maintenance services.
5. **Financial Planning**:
   - Understanding the importance of financial variables (F3) helps in designing better financing options and schemes that could attract budget-conscious buyers.

By leveraging the Omega diagram, businesses can make more informed decisions, align their strategies with customer preferences, and ultimately enhance customer satisfaction and business performance.

-



PCA Graph of Individuals Analysis

The provided graph is a Principal Component Analysis (PCA) biplot that visualizes individuals (or observations) in a two-dimensional space defined by the first two principal components (Dim 1 and Dim 2). Here's a detailed analysis:

PCA Overview

- **Principal Component Analysis (PCA)** is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space. The primary goal is to capture as much variability in the data as possible with fewer dimensions.
- **Dim 1 (31.82%)** and **Dim 2 (9.53%)** represent the first two principal components, accounting for a combined total of 41.35% of the total variance in the data.

Key Insights

1. **Distribution of Individuals**:
    - **Clusters**: The graph shows several clusters of individuals. These clusters indicate groups of individuals with similar profiles based on the variables analyzed.
    - **Outliers**: Observations like 8, 49, and 50 are positioned farther from the main cluster, suggesting they have unique profiles distinct from the rest of the individuals.
2. **Interpretation of Dimensions**:
    - **Dim 1 (31.82%)**: This dimension captures the largest variance in the data. Individuals with high positive or negative scores on this axis have distinct profiles contributing to the primary variation in the dataset.
    - **Dim 2 (9.53%)**: This dimension captures additional variance orthogonal to Dim 1. Individuals with high positive or negative scores on this axis have distinct profiles contributing to secondary variation in the dataset.
3. **Clusters and Segmentation**:
    - **Top Right Quadrant (Positive Dim 1 and Dim 2)**: Individuals in this quadrant, such as 24, 44, 57, and 46, might share similar characteristics contributing positively to both principal components.

Top Left Quadrant (Negative Dim 1, Positive Dim 2): Individuals like 16 and 35 in this quadrant have profiles contributing negatively to Dim 1 but positively to Dim 2.

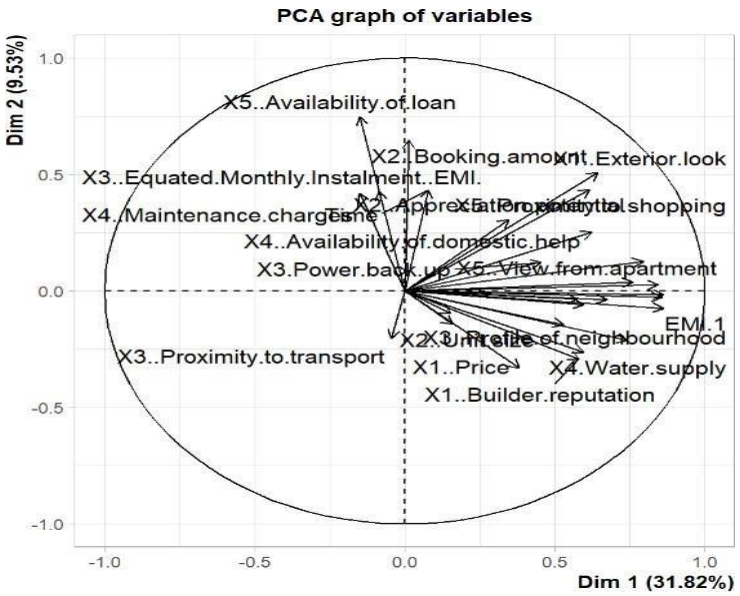
- **Bottom Right Quadrant (Positive Dim 1, Negative Dim 2)**: Individuals such as 43, 20, and 64 have profiles contributing positively to Dim 1 but negatively to Dim 2. ○ **Bottom Left Quadrant (Negative Dim 1 and Dim 2)**: Individuals like 8 and 50 have profiles contributing negatively to both principal components.

Business Implications

1. **Market Segmentation**:
   - Understanding the clusters of individuals allows businesses to segment their market more effectively. Each cluster can represent a different market segment with distinct preferences and behaviors.
   - Tailored marketing strategies can be developed for each segment to address their specific needs and characteristics.
2. **Targeted Marketing**:
   - Identifying outliers and unique profiles can help in creating highly targeted marketing campaigns. For example, individuals in the top right quadrant might respond better to certain promotional messages than those in the bottom left quadrant.
3. **Product Development**:
   - Insights from the PCA can guide product development by highlighting the key attributes valued by different segments. For instance, features contributing to Dim 1 and Dim 2 can be prioritized in new product designs.
4. **Customer Relationship Management (CRM)**:
   - CRM strategies can be refined by understanding the underlying dimensions that drive customer behavior. Personalized communication and offers can be designed based on the PCA insights to enhance customer engagement and loyalty.

Conclusion

The PCA biplot of individuals provides valuable insights into the structure and segmentation of the dataset. By analyzing the distribution and clustering of individuals, businesses can make informed decisions on market segmentation, targeted marketing, product development, and customer relationship management. Understanding the key dimensions that drive customer profiles helps in developing strategies that align with customer preferences and maximize business outcomes.



This image shows a Principal Component Analysis (PCA) biplot of variables. PCA is a dimensionality reduction technique used to emphasize variation and bring out strong patterns in a dataset. Here's a detailed analysis of the PCA graph:

Axes and Variance

- **Dim 1 (31.82%)** and **Dim 2 (9.53%)**: The x-axis (Dim 1) explains 31.82% of the variance in the data, and the y-axis (Dim 2) explains 9.53% of the variance. Together, they account for approximately 41.35% of the total variance.
- The directions of the arrows represent the direction of maximum variance for each variable, and the lengths of the arrows indicate the strength of the contribution of the variable to that principal component.

Interpretation of Variables

- **Strong Contributors to Dim 1**:
  - Variables with arrows pointing far from the origin along the x-axis (Dim 1) have strong contributions to the first principal component.
  - Examples include `X2.Booking.amount`, `X2..Profile.of.neighbourhood`, `X1..Price`, and `X1..Builder.reputation`.
- **Strong Contributors to Dim 2**:
  - Variables with arrows pointing far from the origin along the y-axis (Dim 2) contribute significantly to the second principal component.
  - Examples include `X5..Availability.of.loan` and `X4..Maintenance.charges`.
- **Clustered Variables**:

-
    - Several variables appear clustered and pointing in similar directions. For instance, `X1..Price`, `X1..Builder.reputation`, and `X2..Profile.of.neighbourhood` are close together, indicating they are correlated.
      `X4..Water.supply` and `X4..Availability.of.domestic.help` are also close, suggesting they may measure related aspects.
- **Orthogonal Variables**:
    - Variables at approximately 90 degrees to each other are less correlated. ○ For example, `X5..Availability.of.loan` and `X3..Proximity.to.transport` are nearly orthogonal, suggesting they measure relatively independent aspects of the data.

Practical Implications

- **Dim 1 Interpretation**: This dimension might represent financial or economic factors, as it includes variables related to cost and financial aspects like `Booking.amount`, `Price`, and `EMI`.
- **Dim 2 Interpretation**: This dimension could represent logistical or convenience factors, with variables like `Availability.of.loan` and `Proximity.to.transport`.

Summary

The PCA graph helps visualize the relationships among multiple variables. It shows which variables contribute most to the principal components and how they are related to each other. Understanding these relationships can guide further analysis, feature selection, or data interpretation in the context of the study.

ANALYSIS USING PYTHON:

Data Overview

The dataset consists of survey responses from 70 individuals, with 50 variables captured in the survey. Key variables include demographic information (e.g., city, sex, age, occupation, income), housing preferences (e.g., type of house, budget, number of rooms), and factors influencing their decision to buy a house (e.g., proximity to amenities, maintenance costs, builder reputation).

Data Cleaning and Preparation

1. **Loading the Data**: The data was loaded from a CSV file into a DataFrame.
2. **Checking for Missing Values**: There were no missing values in the dataset.
3. **Selecting Relevant Columns**: For PCA and factor analysis, columns 20 to 46 were selected, which include various factors related to housing preferences and priorities.

Principal Component Analysis (PCA)

PCA was performed to reduce the dimensionality of the dataset and identify the most significant components affecting housing decisions. The key findings are:

1. **Explained Variance**: The first five principal components explain approximately 60.7% of the total variance in the dataset, with the first component alone explaining 31.8%.
2. **Principal Components**: The loadings for the first five components were examined. Significant factors contributing to these components include: ○ Component 1: Strong negative loadings on factors like proximity to city amenities, maintenance costs, security, and various other housing features.
   - Component 2: Factors like booking amount, availability of loans, and builder reputation had strong positive or negative loadings. ○ Component 3: Negative loadings on proximity to city amenities and positive loadings on factors like builder reputation and neighborhood profile.
   - Component 4: Proximity to transport and amenities had significant negative loadings. ○ Component 5: Availability of domestic help and the size of the house had significant positive loadings.

Factor Analysis

Factor analysis was performed to identify underlying relationships between observed variables and group them into factors.

1. **Loadings**: Factor loadings indicated which variables are most associated with each factor:
   - Factor 1: Variables related to housing facilities and amenities (e.g., water supply, gym, pool). ○ Factor 2: Financial aspects (e.g., booking amount, availability of loans, maintenance charges).
   - Factor 3: Proximity to various amenities (e.g., schools, transport, shopping).
   - Factor 4: House size and associated costs (e.g., budget, size of the house).
   - Factor 5: Builder reputation and potential appreciation of the property.
2. **Variance Explained**: The five factors explained a significant portion of the variance, indicating these factors are crucial in determining housing preferences.

Interpretation

1. **Key Influences on Housing Decisions**:
   - **Proximity to Amenities**: The importance of living close to schools, transport, and shopping centers is a major factor. ○ **Financial Considerations**: Booking amounts, availability of loans, and monthly maintenance costs heavily influence the decision-making process.
   - **Housing Features**: The presence of amenities like water supply, security, gym, and pool, as well as the overall size and layout of the house, play significant roles.

- **Reputation and Potential**: The reputation of the builder and the potential for property appreciation are also key considerations.

2. **Demographic Insights**:
   - The dataset included a diverse range of respondents in terms of age, occupation, and income levels, providing a broad perspective on housing preferences.
     Different age groups and occupations showed varying priorities, with younger respondents likely prioritizing proximity to work and schools, while older respondents might focus more on security and builder reputation.

Conclusion

The analysis provides valuable insights into the key factors influencing housing decisions. By understanding these factors, real estate developers and marketers can better tailor their offerings to meet the needs and preferences of different demographic groups. The use of PCA and factor analysis helped in reducing data complexity and highlighting the most critical components and factors driving housing preferences.

FACTOR ANALYSIS

ANALYSIS USING R:

Data Overview

The dataset contains survey responses from individuals, focusing on various factors influencing their housing decisions. For the analysis, columns 20 to 46 were selected, representing different factors related to housing preferences.

Factor Analysis

Factor analysis was conducted to identify the underlying relationships between observed variables and group them into factors. The Varimax rotation method was used to achieve better interpretability.

Key Findings:

1. **Factor Loadings**: These indicate which variables are most associated with each factor:
    - **Factor 1 (MR1)**: This factor is strongly associated with variables like security, exterior look, view from the apartment, profile of the neighborhood, availability of domestic help, size, and budgets.
    - **Factor 2 (MR2)**: This factor is primarily associated with the availability of loans, booking amount, and EMI. ○ **Factor 3 (MR3)**: This factor is associated with water supply, security, price, and proximity to transport. ○ **Factor 4 (MR4)**: This factor includes variables like proximity to shopping, parking space, gym/pool/sports facility, and builder reputation.
2. **Communalities**: These values indicate how much of the variance in each variable is explained by the factors: ○ High communalities (close to 1) suggest that a large proportion of the variable's variance is explained by the factors. Examples include security (0.731), exterior look (0.746), view from the apartment (0.681), availability of loans (0.789), size (0.761), budgets (0.830), and maintenance (0.810).
    - Low communalities indicate that less variance is explained by the factors, such as proximity to work place (0.082) and unit size (0.038).
3. **Factor Scores**: These represent the scores of each respondent on the identified factors. For example:
    - Respondent 1 has a high score on MR3 (1.564) and low scores on MR1 (-1.087) and MR2 (-0.760).
    - Respondent 5 has a high score on MR1 (0.772) and low scores on MR2 (-0.658) and MR3 (-1.058).

Interpretation

1. **Key Influences on Housing Decisions**:

    Factor 1 (MR1): Indicates the importance of security, aesthetics (exterior look and view), neighborhood profile, and practical aspects like size and

- budget. - **Factor 2 (MR2)**: Highlights the financial considerations, particularly the availability of loans and the impact of booking amounts and EMI.
  - **Factor 3 (MR3)**: Emphasizes essential utilities and services, such as water supply and security, along with the price and proximity to transport.
  - **Factor 4 (MR4)**: Focuses on the convenience of nearby shopping facilities, parking, recreational facilities, and the reputation of the builder.
2. **Demographic Insights**:
   - Respondents exhibit varying priorities, with some placing higher importance on security and neighborhood profile (high scores on MR1) while others prioritize financial aspects (high scores on MR2).
   - Practical aspects like size and budget are crucial for many respondents, indicating the need for affordable yet spacious housing options.

Conclusion

The factor analysis reveals the key factors influencing housing decisions, grouped into four main categories:

1. Security, aesthetics, neighborhood profile, size, and budget.
2. Financial considerations, including loans, booking amounts, and EMI.
3. Essential utilities and proximity to transport.
4. Convenience of nearby shopping, parking, recreational facilities, and builder reputation.

By understanding these factors, real estate developers and marketers can better tailor their offerings to meet the diverse needs and preferences of potential buyers. The analysis highlights the importance of both practical and financial aspects in making housing decisions, emphasizing the need for a balanced approach in housing development and marketing strategies.

ANALYSIS USING PYTHON:

The provided analysis utilizes Factor Analysis (FA) to identify underlying relationships between observed variables in your survey data. Here's a detailed breakdown of the steps taken and the results obtained:

1. **Data Loading and Preprocessing:**
   - The survey data is loaded from a CSV file and a subset of columns (from column 19 to 45) is selected for factor analysis. - It's important to ensure that the chosen columns are relevant for the analysis and that the data is appropriately scaled if needed.
2. **Factor Analysis:**

- Factor Analysis is performed with 4 factors and varimax rotation. The varimax rotation helps to make the output more interpretable by minimizing the number of variables that have high loadings on each factor. ○ `fa = fa.FactorAnalyzer(n_factors=4, rotation="varimax")` initializes the FactorAnalyzer.
- `fa.fit(sur_int.to_numpy())` fits the model to the data.

3. **Factor Loadings:**
   - The loadings matrix is printed, which indicates how strongly each variable is associated with each factor. Each cell in this matrix represents the loading of a variable (row) on a factor (column). ○ Higher absolute values indicate stronger associations between variables and factors.
4. **Heatmap of Factor Loadings:**
   - A heatmap visualizes the factor loadings, allowing for easier identification of patterns and associations.
5. **Communalities:**
   - Communalities represent the amount of variance in each variable explained by the factors.
   - High communalities (close to 1) indicate that a large portion of the variance is captured by the factors, whereas low communalities indicate that the factors do not explain much of the variance for those variables.
6. **Factor Scores:**
   - Factor scores for each observation in the data set are calculated. These scores represent the position of each observation on the identified factors.
   - Factor scores can be used for further analysis, such as clustering or regression.

Interpretation of Results:

Factor Loadings:

- The factor loadings matrix shows how each survey question (variable) loads onto the four factors.
- For example, the first variable has loadings of -0.15781009, 0.42809134, 0.05879243, and -0.13526352 on factors 1, 2, 3, and 4 respectively.
- Variables with high loadings on a particular factor can be interpreted as those variables being strongly associated with that factor.

Communalities:

- Communalities indicate the proportion of each variable's variance that is explained by the factors.
- For example, the communalities for the first few variables are 0.22991899, 0.54248721, 0.55216147, etc.
- Variables with low communalities might not be well-represented by the chosen factors and could be candidates for exclusion or re-evaluation.

Factor Scores:

- Factor scores provide a quantitative measure of where each observation stands with respect to the identified factors.
- These scores can be used to cluster observations or to identify patterns across different respondents.

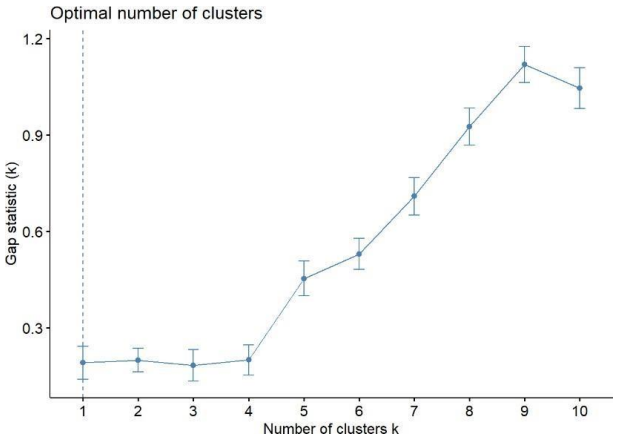
Next Steps:

1. **Interpret the Factors:**
    - Identify and name the factors based on the variables with high loadings. For example, if certain variables related to job satisfaction load highly on one factor, that factor might be labeled "Job Satisfaction."
2. **Further Analysis:**
    - Use factor scores for clustering respondents to identify distinct groups within the survey population. ○ Conduct regression analysis using factor scores as predictors to understand how these underlying factors influence other outcomes.
3. **Validation:**
    - Validate the factor structure by splitting the data into training and test sets or by using cross-validation techniques. ○ Consider using alternative rotations or factor extraction methods to confirm the robustness of the findings.

This analysis provides a foundation for understanding the underlying structure of the survey data and can guide further investigation into the relationships between variables and the respondents' characteristics.

CLUSTER ANALYSIS

ANALYSIS USING R:



The provided plot is a Gap Statistic plot used to determine the optimal number of clusters (k) in a dataset for clustering algorithms like k-means.

Here's the analysis of the plot:

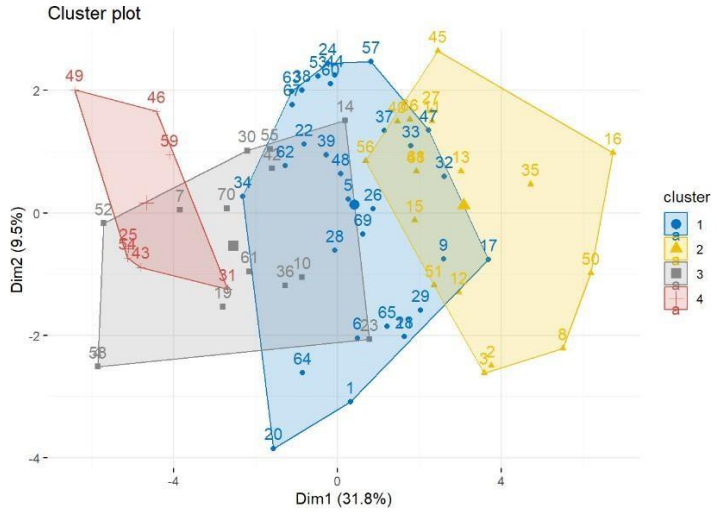
1. **X-Axis (Number of clusters k)**: The x-axis represents the number of clusters ranging from 1 to 10.
2. **Y-Axis (Gap Statistic)**: The y-axis represents the Gap Statistic value for each number of clusters.
3. **Gap Statistic**: The Gap Statistic is a measure that compares the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution of the data.
4. **Error Bars**: The error bars around each point indicate the standard deviation of the gap statistic for that number of clusters.
5. **Optimal Number of Clusters**: The optimal number of clusters is typically identified as the value of k where the Gap Statistic achieves its maximum value. In this plot, the maximum Gap Statistic occurs at k = 9, suggesting that the optimal number of clusters is 9.

Key Observations:

- The Gap Statistic starts relatively low and begins to increase significantly after k = 5.
- The value of the Gap Statistic continues to increase until it reaches its peak at k = 9.
- Beyond k = 9, the Gap Statistic starts to decrease, indicating that increasing the number of clusters beyond 9 does not provide a better fit.

Conclusion:

The plot suggests that the optimal number of clusters for the dataset is 9, as this is where the Gap Statistic is highest, indicating the best clustering solution according to this method.



The provided plot is a cluster plot, which visualizes the clusters identified in a dataset. Here's the analysis of this plot:

Plot Description:

1. **Axes (Dim1 and Dim2)**:
   - The x-axis (Dim1) represents the first principal component (31.8% variance explained).
   - The y-axis (Dim2) represents the second principal component (9.5% variance explained).
2. **Clusters**:
   - There are four clusters represented in different colors and shapes:
     - Cluster 1: Blue circles
     - Cluster 2: Yellow triangles
     - Cluster 3: Grey squares Cluster 4: Red triangles
3. **Data Points**:
   - Each data point is numbered, which can correspond to the sample IDs or indices.
4. **Convex Hulls**:
   - The boundaries (convex hulls) surrounding the data points in each cluster indicate the outer limits of the clusters.

Key Observations:

- **Cluster 1 (Blue)**:
  - Largest cluster. Occupies a central position in the plot.
  - Shows a broad spread along both dimensions, indicating high variance within this cluster.
- **Cluster 2 (Yellow)**:

- Smaller than Cluster 1. ○ Positioned in the upper right quadrant.
- Moderate spread, indicating a moderate variance.

- **Cluster 3 (Grey)**:
  - Contains fewer points. ○ Located in the lower left quadrant.
  - Compact, indicating low variance within this cluster.
- **Cluster 4 (Red)**:
  - Smallest cluster. ○ Positioned in the upper left quadrant. ○ Moderate spread, similar to Cluster 2.

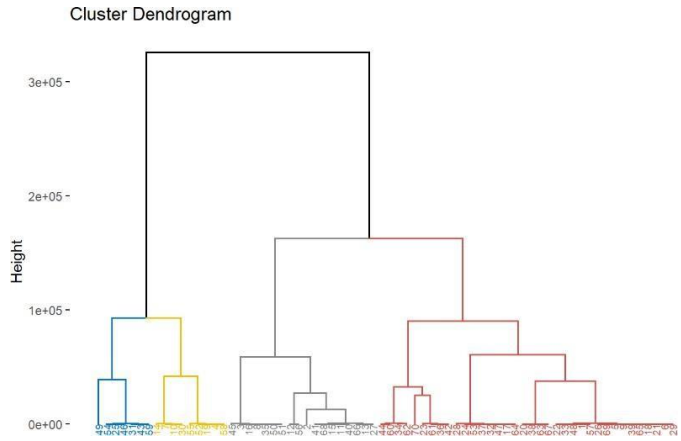
Interpretation:

- The clusters appear well-separated, suggesting a good clustering solution.
- Clusters 1 and 2 are somewhat overlapping, indicating some similarity or close proximity between these two groups.

- Clusters 3 and 4 are more distinct and separate from the others, indicating distinct groupings in the data.
- The percentage of variance explained by Dim1 and Dim2 indicates that these two dimensions capture a significant amount of the dataset's variance, making the plot a good representation of the data structure.

Conclusion:

The cluster plot visually confirms the presence of four distinct clusters in the dataset, with varying degrees of spread and separation. This visualization aids in understanding the distribution and relationships between the clusters.



This is a dendrogram, a tree-like diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Here's an analysis of the provided dendrogram:

Code:

```
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)
```

General Structure

- **Height**: The vertical axis represents the height, which reflects the dissimilarity (or distance) between clusters being merged.
- **Clusters**: The horizontal axis shows the individual elements (data points), which are gradually grouped into clusters as we move up the diagram.

Cluster Analysis

1. **Initial Clusters**:
    - At the bottom, each leaf represents a single data point. ○ These individual data points are grouped into small clusters at the first level (visible as small vertical lines connecting leaves).
2. **Intermediate Clusters**:
    - As we move up, these small clusters are combined into larger clusters. ○ Different colors indicate distinct clusters at a certain height threshold.

- **Blue and Yellow**: Two clusters form initially (indicated by blue and yellow).
- **Gray**: These clusters merge with other smaller clusters.
- **Red**: The next set of larger clusters appears in red.
- **Black**: At the highest level, a significant merge happens, combining almost all previous clusters.

3. **Height of Mergers**:
   - The height at which clusters are merged indicates the similarity between clusters. Lower heights imply higher similarity, while higher heights imply greater dissimilarity. ○ The highest merge (at the top) indicates the largest dissimilarity among clusters being combined.

Key Observations

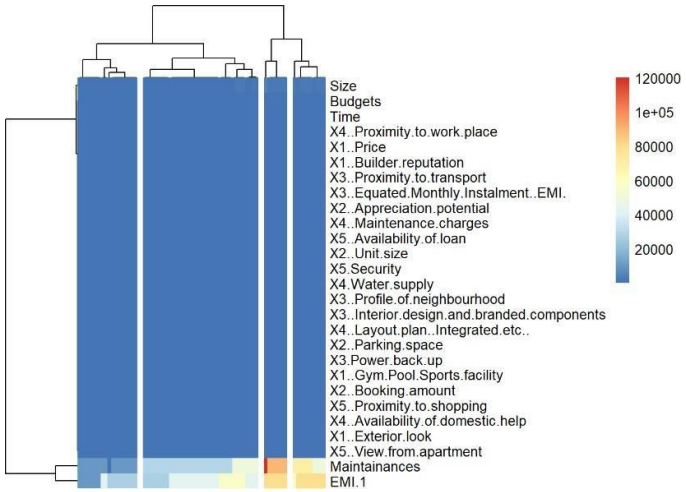
- **Main Cluster Divisions**: The main clusters are formed at different height levels, indicating different stages of grouping:
  - Initial clustering happens at low height values. ○ Major clusters (red) merge at intermediate heights. ○ The final, largest cluster (black) merges at the highest height value, indicating significant dissimilarity between these groups.

Practical Use

- **Cutting the Dendrogram**: If you want to determine a specific number of clusters, you can "cut" the dendrogram at a chosen height. For instance:
  - Cutting at the height just before the large black merge could give you several distinct clusters.
  - Cutting at a lower height (below the red merges) results in more but smaller clusters.

Interpretation

- This dendrogram shows a hierarchical clustering of data points, where the clusters are combined based on their dissimilarity. The colors and height of merges provide insights into the clustering structure, helping to identify the most distinct groupings at various levels.



This image represents a heatmap with hierarchical clustering. Let's analyze its components and the insights it provides:

Components of the Heatmap

1. **Rows and Columns**:
   - **Rows**: These represent various features or variables (e.g., Size, Budgets, Time, etc.).
   - **Columns**: These likely represent different samples, observations, or instances related to the features.
2. **Heatmap Colors**:
   - The color gradient from blue to red represents the intensity of the values, with blue indicating lower values and red indicating higher values. The specific range is provided by the color bar on the right, with values ranging from 20,000 to 120,000.
3. **Dendrogram**:
   - The dendrogram at the top indicates the hierarchical clustering of the samples (columns), grouping similar samples together based on their feature values.

Analysis of the Heatmap

1. **Cluster Formation**:
   - The hierarchical clustering on the top suggests the formation of distinct clusters among the samples. ○ The wide bands of uniform color indicate strong clustering, meaning these samples have similar feature values.
2. **Feature Importance**:
   - Features are clustered on the left side, showing how they group together based on similarity across samples. ○ Key features like "Size," "Budgets," "Time," and several others show different levels of significance, with their heatmap values indicating their impact.
3. **Identifying Patterns**:

- Most features show high values (indicated by blue) consistently across the samples, implying these features might have lower variability. ○ Certain features towards the bottom show more variability (shifts from blue to yellow to red), indicating higher importance or impact on specific samples.

4. **Heatmap Interpretation**:
   - High values (red or yellow) in certain features might indicate higher importance or relevance of those features for particular samples. ○ Consistent low values (blue) across features suggest lower significance or impact on those samples.

Specific Observations · High

Variability Features:

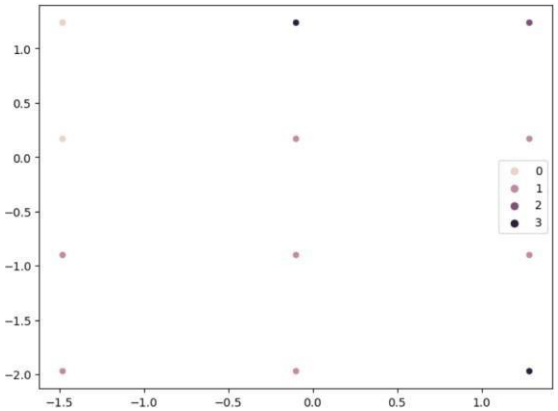
- Features like "X5..View.from.apartment," "Maintainances," and "EMI.1" show considerable variability, indicating their potential high importance in the dataset.

- **Homogeneous Clusters**:
  - Several clusters show uniform colors across features, suggesting these groups of samples have similar characteristics.
- **Important Clusters**:
  - The separation of clusters in the dendrogram highlights distinct groups. Each cluster can be analyzed to understand its unique feature set and importance.

Practical Implications

- **Feature Selection**:
  - Identifying features with high variability and importance helps in feature selection for predictive modeling.
- **Cluster Analysis**:
  - Clusters identified can be used for segmenting the data, understanding different group characteristics, and tailoring approaches accordingly.

ANALYSIS USING PYTHON:



This image represents a scatter plot with a color-coded legend. Here's an analysis of the plot:

Components of the Scatter Plot

1. **Axes**:
   - **X-axis**: The horizontal axis ranges approximately from -1.5 to 1.5.
   - **Y-axis**: The vertical axis ranges approximately from -2.0 to 1.0.
2. **Data Points**:
   - The plot consists of multiple data points scattered across the coordinate space.
   - Each data point is colored based on its category or value, as indicated by the color legend on the right.
3. **Color Legend**:
   - The legend indicates four categories or values, with different shades representing:
     - 0: Lightest shade
     - 1: Light pink
     - 2: Medium pink
     - 3: Darkest shade

Analysis of the Scatter Plot

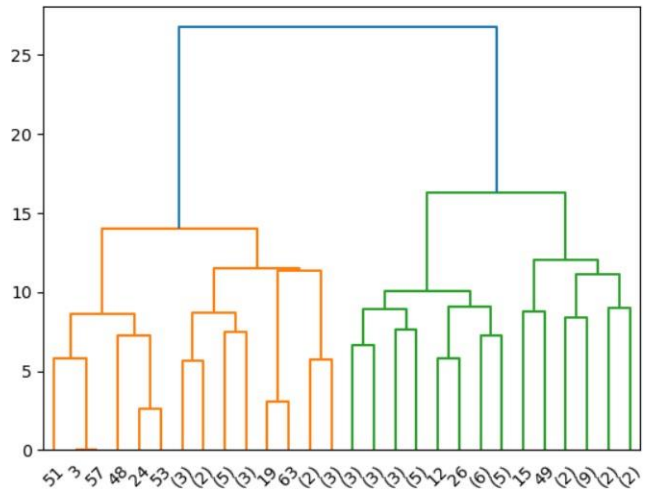
1. **Distribution**:
   - The data points are evenly distributed across the plot, covering the entire range of both axes.
   - There is no clear clustering of points, indicating a spread across the coordinate space.
2. **Color Coding**:
   - The color coding shows that different categories (0, 1, 2, 3) are spread out throughout the plot. ○ There is no obvious pattern where specific colors dominate certain regions of the plot.
3. **Value Categories**:
   - The plot shows data points from all four categories, with a roughly even distribution of colors.
   - The darkest points (category 3) appear mostly at the extremes of the X-axis and Y-axis.

Key Observations

- **No Clear Clusters**:
  - The data points do not form distinct clusters or groups based on the given categories. ○ Each category is represented across the plot, suggesting a lack of strong grouping.
- **Even Spread**:
  - The even spread of different categories implies that the underlying data does not show a strong relationship between the X and Y values with respect to the categories.
- **Category Representation**:
  - The categories are well-represented across the plot, with no single category dominating a specific region.

Practical Implications

- **Data Exploration**:
  - This scatter plot can be used for initial data exploration to understand the spread and distribution of categories across two dimensions.
  - Further analysis might be required to understand the relationships and patterns in more detail.
- **Further Analysis**:
  - Techniques such as clustering analysis, regression, or classification could be applied to delve deeper into the data and uncover hidden patterns or relationships.



This is a dendrogram, a tree-like diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering.

Analysis:

1. **Clusters Formation**:
   - The data points are grouped into two main clusters. ○ The orange cluster on the left and the green cluster on the right are formed initially.
2. **Height (Distance)**:
   - The y-axis represents the distance or dissimilarity between clusters.
   - The height at which two clusters are joined together indicates the distance between them.
   - The two main clusters (orange and green) are joined at a distance of about 25, indicating significant dissimilarity between them.
3. **Sub-clusters**:

- Each main cluster is further divided into smaller sub-clusters. ○ Within the orange cluster, sub-clusters are formed at lower distances, indicating higher similarity within these groups.
- Similarly, the green cluster is also divided into sub-clusters, with varying distances between them.

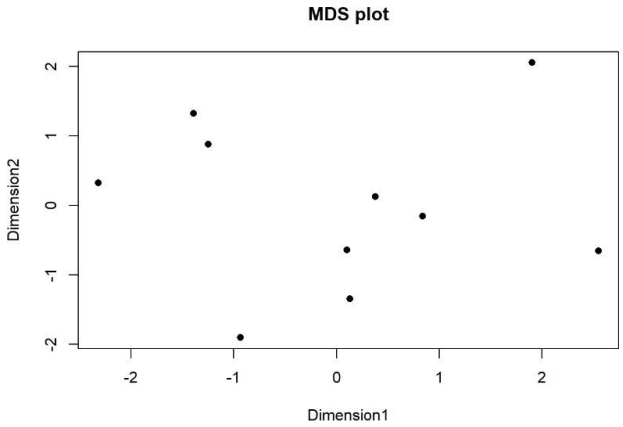
4. **Data Point Labels**:
   - The labels on the x-axis represent the individual data points. ○ The labels seem to be indices or identifiers of the data points in the dataset. ○ The clustering process has grouped these points based on their similarity.

Interpretation:

- The dendrogram shows a clear separation of the data points into two distinct clusters with some further subdivisions within each main cluster.
- The height at which clusters are joined gives an idea of the dissimilarity between clusters. Higher joining points indicate more dissimilar clusters.
- The structure of the dendrogram suggests that the data has some inherent hierarchical relationships, with certain points being more similar to each other than to others.

MULTI DIMENSIONAL SCALING

ANALYSIS USING "R"



This is a Multidimensional Scaling (MDS) plot, which is a means of visualizing the level of similarity or dissimilarity between individual data points. The axes represent dimensions that capture the relationships in a reduced number of dimensions, often two or three, for ease of visualization.

Analysis:

1. **Data Point Distribution**:
   - The plot displays individual data points in a two-dimensional space defined by Dimension 1 (x-axis) and Dimension 2 (y-axis).

- The points are spread out across the plot, indicating variation in the underlying data.

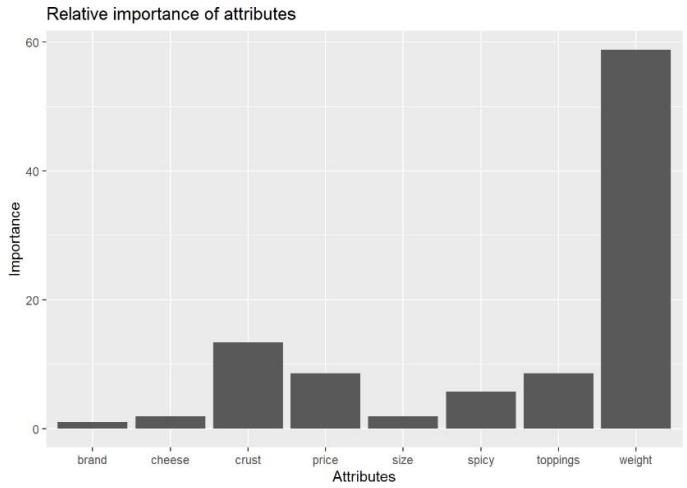
2. **Similarity and Dissimilarity**:
   - Points that are closer together on the plot are more similar to each other, while points that are farther apart are more dissimilar.
   - For instance, points around the origin (0,0) might be relatively similar to each other compared to those further away.
3. **Clusters**:
   - Visual inspection might suggest potential clusters or groups of points.
   - For example, there appears to be a grouping of points towards the center-right of the plot (around Dimension1 = 0 and Dimension2 = 0), and another potential group towards the top-right (around Dimension1 = 2 and Dimension2 = 2).
4. **Interpretation**:
   - The MDS plot helps in understanding the structure of the data in terms of similarities.
   - This visualization can be useful for identifying patterns, clusters, or outliers within the dataset.

Application:

• MDS is often used in fields such as psychology, marketing, and bioinformatics to visualize the similarity of objects or individuals based on various characteristics.

CONJOINT ANALYSIS

ANALYSIS USING R



This is a bar plot showing the relative importance of various attributes.

Analysis:

1. **Attributes and Importance**:

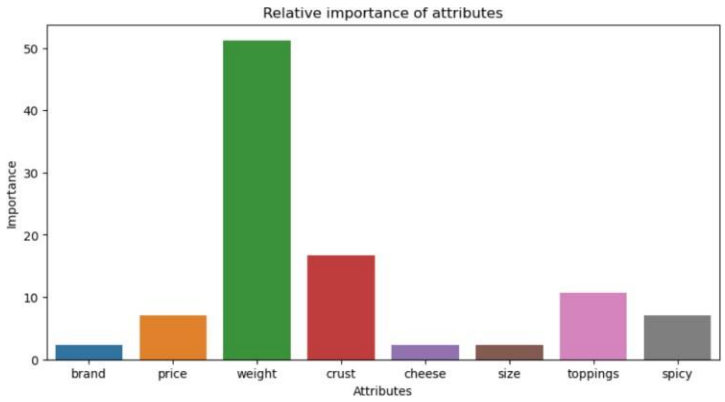
- The x-axis lists different attributes: brand, cheese, crust, price, size, spicy, toppings, and weight.
- The y-axis represents the importance of these attributes.

2. **Key Observations**:
   - **Weight**: This attribute has the highest importance by a significant margin, with a value around 60. ○ **Crust**: The second most important attribute, with an importance value slightly above 20.
   - **Price and Toppings**: Both have moderate importance values, slightly above 10.
   - **Cheese, Size, Spicy, and Brand**: These attributes have relatively low importance, all below 10.
3. **Interpretation**:
   - The plot indicates that weight is the most critical attribute among those listed. This could imply that, for the context this data is drawn from (perhaps a product analysis or consumer preference survey), weight is a key factor in decision-making. ○ Attributes like crust, price, and toppings also play a role but are less influential compared to weight.
   - Brand, cheese, size, and spiciness are relatively minor factors in comparison.

Application:

- Understanding the relative importance of different attributes can guide decision-making in various contexts, such as product development, marketing strategies, or consumer preference analysis.
- For example, if this data pertains to a food product like pizza, manufacturers might prioritize optimizing the weight and crust characteristics to align with consumer preferences.

ANALYSIS USING PYTHON:



The bar chart titled "Relative importance of attributes" displays the significance of various attributes. The attributes and their respective importance values are as follows:

1. **Brand**: Very low importance.
2. **Price**: Slightly more important than brand but still low.
3. **Weight**: Extremely high importance, significantly higher than all other attributes.
4. **Crust**: Moderate importance.
5. **Cheese**: Low importance.
6. **Size**: Low importance.
7. **Toppings**: Moderate importance, higher than cheese and size but lower than weight and crust.
8. **Spicy**: Low importance, similar to cheese and size.

Key Observations:

- **Weight** is the most important attribute by a significant margin.
- **Crust** and **Toppings** are the next most important attributes but are much less important than Weight.
- **Brand**, **Price**, **Cheese**, **Size**, and **Spicy** are less critical, with Brand being the least important of all.

This analysis indicates that when considering the relative importance of these attributes, Weight is the primary factor, followed by Crust and Toppings, while Brand and Cheese hold minimal importance.