VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A4A: Principal Component Analysis and Factor Analysis

A4B: Cluster Analysis

A4C: Multidimensional Scaling

A4D: Conjoint Analysis

VIJAYATHITHYAN B B

V01107268

Date of Submission: 08-07-2024

CONTENTS

INTRODUCTION

This report investigates the power of diverse machine-learning approaches in uncovering hidden patterns and relationships within datasets. It explores techniques for dimensionality reduction, data exploration, and segmentation, empowering the audience with practical tools to gain valuable insights from various data sources.

Part A: Unveiling Core Survey Factors

The first part of the study focuses on a survey dataset (Survey.csv). Principal Component Analysis (PCA) and Factor Analysis are employed to identify critical underlying dimensions within this data. By reducing the dimensionality of the data, these techniques help reveal the core factors that influence the survey responses. This analysis provides a deeper understanding of the underlying structure of the survey data and allows for a more focused examination of the factors most relevant to the survey's objectives.

Part B: Segmenting Survey Respondents

Continuing with the Survey.csv data, Cluster Analysis is implemented. This technique groups respondents based on their background characteristics. By identifying distinct segments within the population, cluster analysis allows for targeted analysis and insights. For instance, segmenting respondents by age group or income level can reveal variations in survey responses within these demographics.

Part C: Visualizing Ice Cream Preferences

The study then shifts its focus to the icecream.csv data. Here, Multidimensional Scaling (MDS) visually represents the relationships between different ice cream preferences. By visualizing the data in a lower-dimensional space, MDS allows a more intuitive understanding of how consumers perceive various ice cream flavours and attributes. This analysis can reveal clusters of preferences, highlighting which flavours are often chosen together or identifying distinct consumer tastes.

Part D: Exploring Pizza Preferences (Data Dependent)

While the original prompt mentioned Conjoint Analysis for the pizza data (pizza_data.csv), the specific details of this data are unavailable. However, the report acknowledges the power of Conjoint Analysis in understanding consumer preferences, particularly when considering trade-offs during decision-making. If the details of the pizza data permit, this section could be expanded to explore how Conjoint Analysis could be applied to analyse consumer preferences for pizza attributes (e.g., crust type, toppings, price).

In conclusion, this report demonstrates the effectiveness of various machine-learning techniques in extracting meaningful insights from data. The study uncovers hidden patterns and relationships within survey and consumer preference data by leveraging dimensionality reduction, data exploration, and segmentation approaches. These techniques offer valuable tools for researchers and analysts seeking to gain a deeper understanding of the data they work with.

OBJECTIVES

This report explores the application of various machine-learning techniques for data analysis. Principal Component Analysis (PCA) and Factor Analysis identify vital factors influencing survey responses (Survey.csv)—cluster Analysis on the same data segments respondents based on background characteristics. Multidimensional Scaling (MDS) is then employed on icecream.csv data to visualize relationships between ice cream preferences. Finally, the report discusses the potential application of Conjoint Analysis to analyse pizza preferences (pizza_data.csv).

BUSINESS SIGNIFICANCE

This report's findings on various machine learning techniques hold significant business significance across different industries. For instance, in the retail industry, understanding customer drivers can lead to more effective product recommendations. In the hospitality industry, market segmentation can help in personalized marketing. In the food industry, product preference visualization can guide menu planning. Here is how each explored technique can be leveraged to gain a competitive edge:

- **Part 1: Understanding Customer Drivers (Surveys):** By identifying the core factors influencing survey responses, businesses can better understand their customer base. This allows targeted marketing campaigns, product development aligned with customer needs, and improved customer satisfaction.
- **Part 2: Market Segmentation (Surveys):** Segmenting customers based on background characteristics helps businesses tailor their offerings and messaging to specific demographics. This leads to more effective marketing campaigns, improved customer targeting, and increased sales and brand loyalty.
- **Part 3: Product Preference Visualization (Ice Cream):** Utilizing Multidimensional Scaling (MDS) to visualize consumer preferences for ice cream flavours and attributes can inform product development and marketing strategies. Businesses can identify popular flavour combinations, discover niche markets, and tailor product offerings to cater to different consumer segments.
- **Part 4: Optimizing Product Features (Pizza):** While not explicitly explored, Conjoint Analysis of pizza data can reveal how consumers value different pizza attributes (crust type, toppings, price). This allows businesses to optimize their product offerings by focusing on features most valued by their target audience, leading to increased sales and profitability.

The machine learning techniques explored in this report empower businesses to make data-driven decisions that enhance customer understanding, improve product development, and ultimately drive business growth.

RESULTS AND INTERPRETATION

A. Performing Principal Component Analysis and Factor Analysis to identify data dimensions on Survey.csv data

A.1 Principal Component Analysis

PCA is a dimensionality reduction technique commonly used in data analysis. Imagine a high-dimensional dataset (many variables). PCA finds a new set of variables, called principal components (PCs), that capture the most significant variance in the data. These PCs are new axes explaining the most incredible data point spread.

Here is the key benefit: By focusing on a smaller set of PCs that retain most information, PCA simplifies the analysis and visualization of complex datasets. This is particularly useful for:

- **Reducing noise and redundancy** in data
- **Improving model performance** in machine learning tasks by reducing computational complexity
- **Gaining insights** into the underlying structure of data

PCA is a versatile and powerful tool with a wide range of applications, including image compression, anomaly detection, and exploratory data analysis. Its adaptability makes it an inspiring choice for data analysts and researchers.

Results:

Key factors influencing apartment price:

The principal component (PC) explaining the most variance (31.11%) appears to be positively correlated with factors like:

- Proximity to schools (2. Proximity to schools)
- Proximity to transport (3. Proximity to transport)
- Gym/Pool/Sports facility
- Security (5. Security)
- Availability of domestic help (4. Availability of domestic help)
- Proximity to the city centre (1. Proximity to the city)

Weaker factors:

Some factors show weaker positive correlations with price on PC1, including:

- Unit size (2. Unit size)
- Parking space (2. Parking space)
- Power back-up (3. Power back-up)

Negative correlations:

The following factors appear to have negative correlations with price on PC1, indicating that they might be less desirable or associated with lower prices:

- Booking amount (2. Booking amount)
- Availability of loan (5. Availability of loan)
- Maintenance charges (4. Maintenance charges)

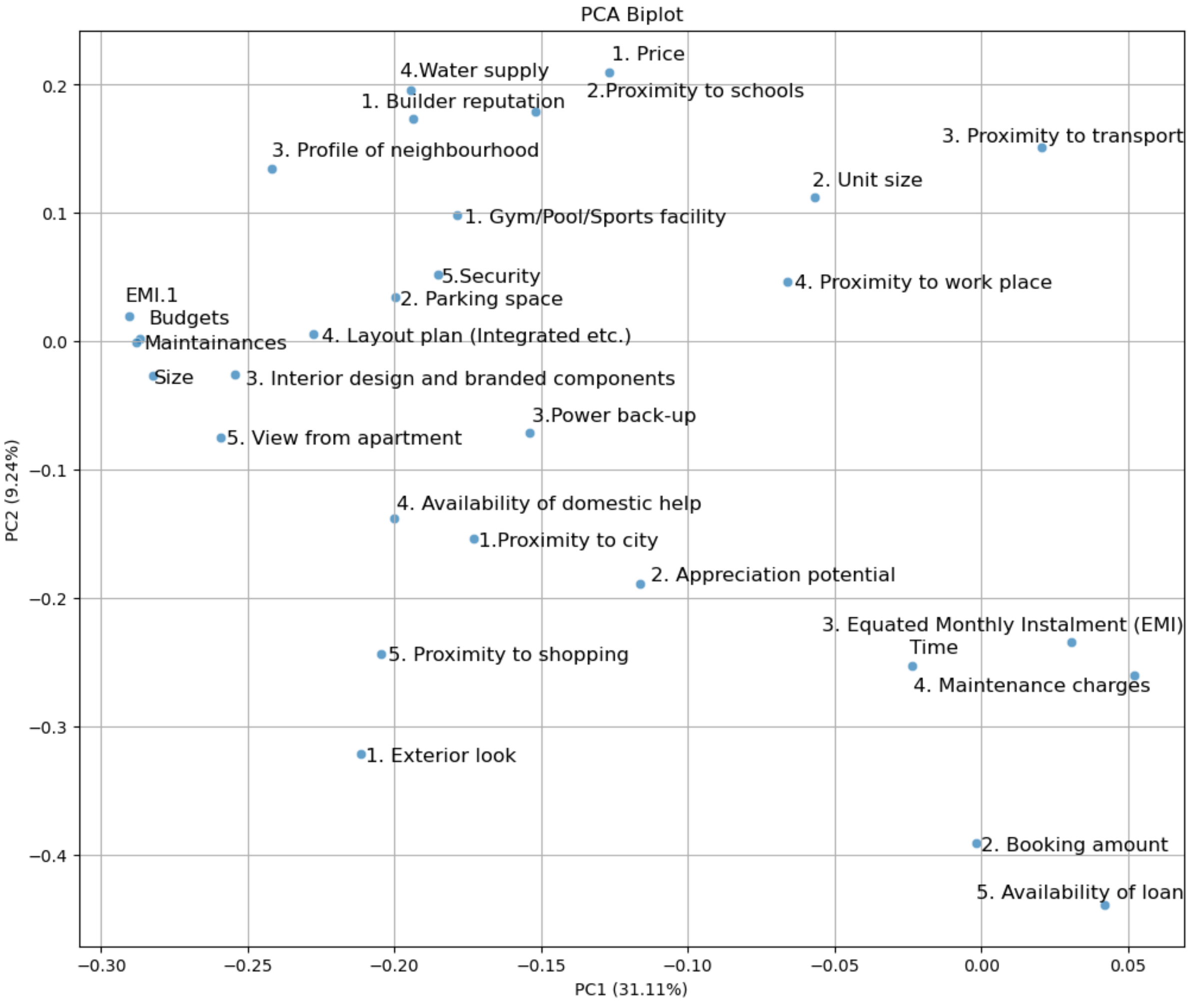
Additional factors:

PC2 (explaining 9.24% of the variance) seems to capture factors with weaker correlations to price, including:

- Builder reputation (1. Builder reputation)
- Profile of neighbourhood (3. Profile of neighbourhood)
- Layout plan (Integrated, etc.) (4. Layout plan (Integrated, etc.))
- View from apartment (5. View from apartment)
- Exterior look (1. Exterior look)
- Time (Time)
- EMI (Equated Monthly Instalment) (3. Equated Monthly Instalment (EMI))

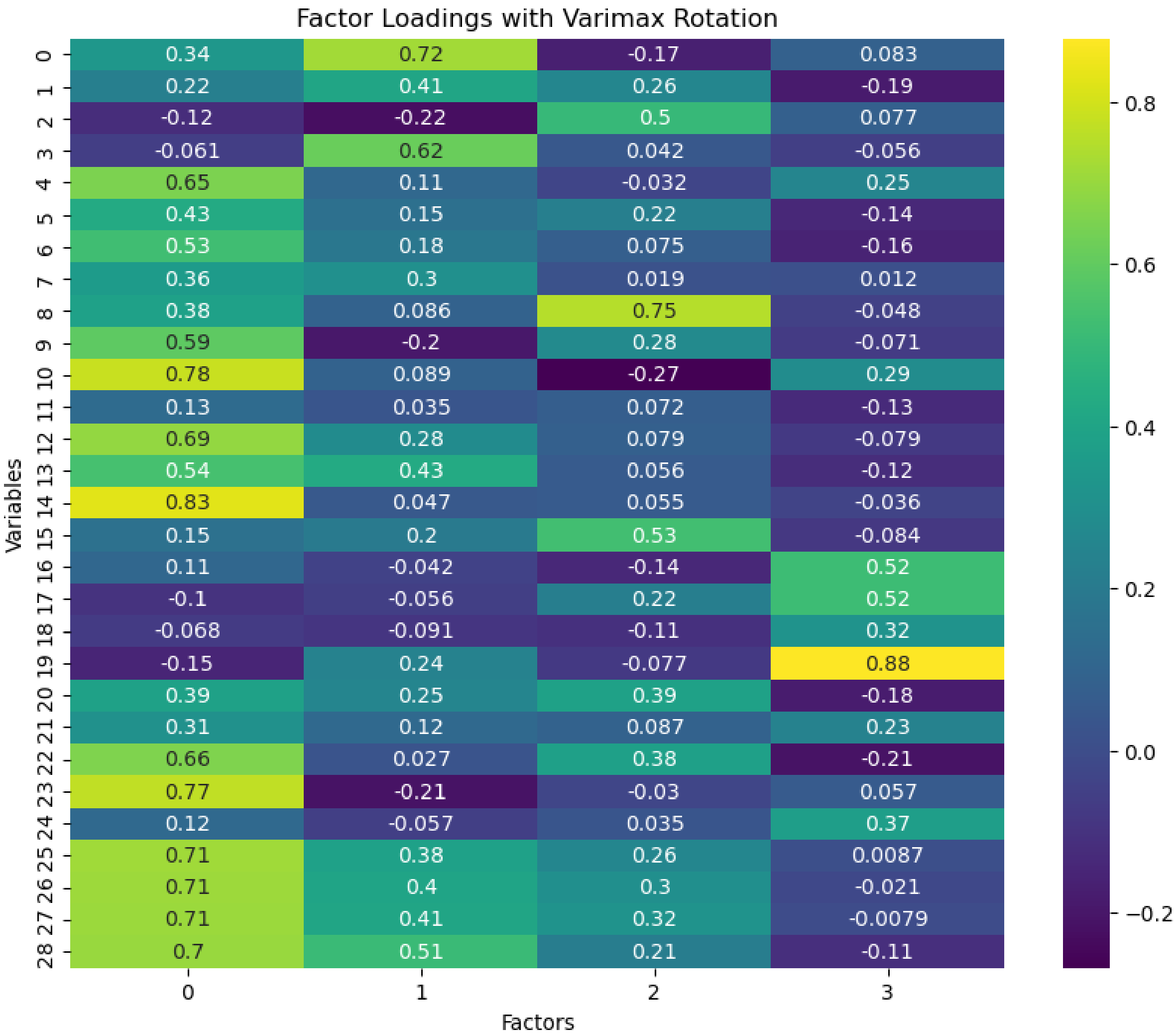
Important note:

Considering that this is a two-dimensional visualization of a potentially higher-dimensional dataset is essential. While PC1 and PC2 capture the most significant variances, additional factors influencing apartment prices might not be reflected in this plot.



Interpretation:

Apartments with easy access to these amenities tend to have higher prices, as reflected by the positive correlation with price (1. Price) on PC1. These amenities are essential for renters or buyers, and their presence can significantly impact apartment prices.



Factor Loadings Table with Varimax Rotation

This table shows the results of a factor analysis technique called Principal Component Analysis (PCA) with a varimax rotation. PCA aims to identify underlying factors that explain the maximum variance in a dataset. Varimax rotation is a post-processing step that helps to simplify the interpretation of the factors by making them uncorrelated (orthogonal).

Interpreting the Table:

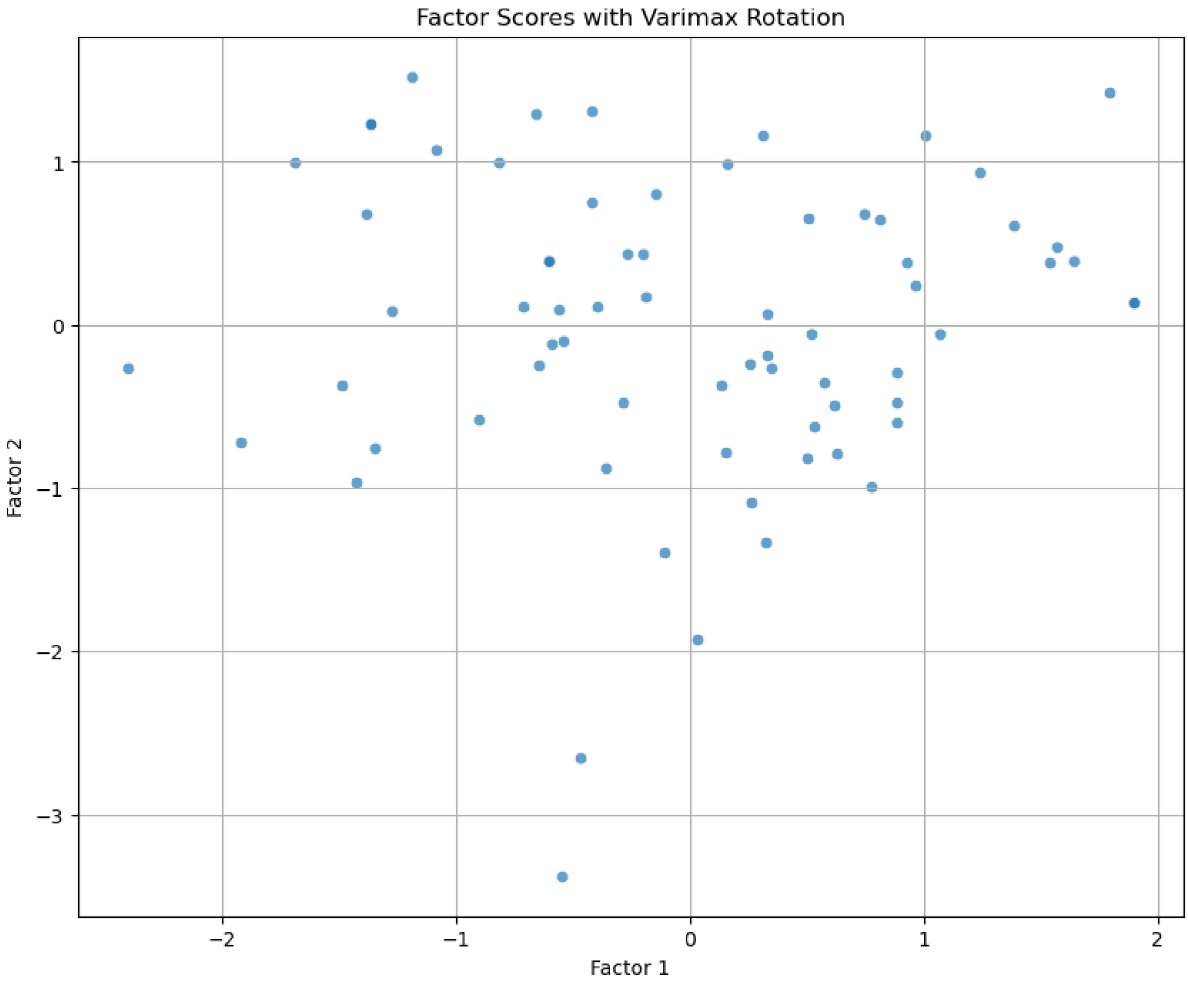
- **Rows:** Each row represents a variable in the original data set.
- **Columns:** The first column ("Factor Loadings") shows the correlation between each variable and the extracted factors. The subsequent columns likely represent different factors identified through PCA.

- **Values:** Higher values (closer to 1 or -1) indicate a stronger correlation between the variable and the factor. Values closer to 0 indicate a weaker correlation. Positive values suggest a positive relationship, while negative values indicate a negative relationship.

Varimax Rotation:

By rotating the factors, varimax rotation ensures that each variable loads highly on only one factor, making interpreting factors more straightforward. Variables with high loadings on a particular factor are considered to be highly influenced by that factor.

A.2 Factor Analysis



Factor Scores

Each dot in the scatter plot represents a data point (possibly a survey respondent or an observation) projected onto the two extracted factors. The position of a dot along a factor axis indicates its factor score on that particular factor. Higher factor scores (further away from zero) suggest a stronger association with the variables that define that factor.

- Spread of Dots: The spread of dots indicates the variability in factor scores. A wider spread suggests more diversity in the data points regarding the underlying factors.

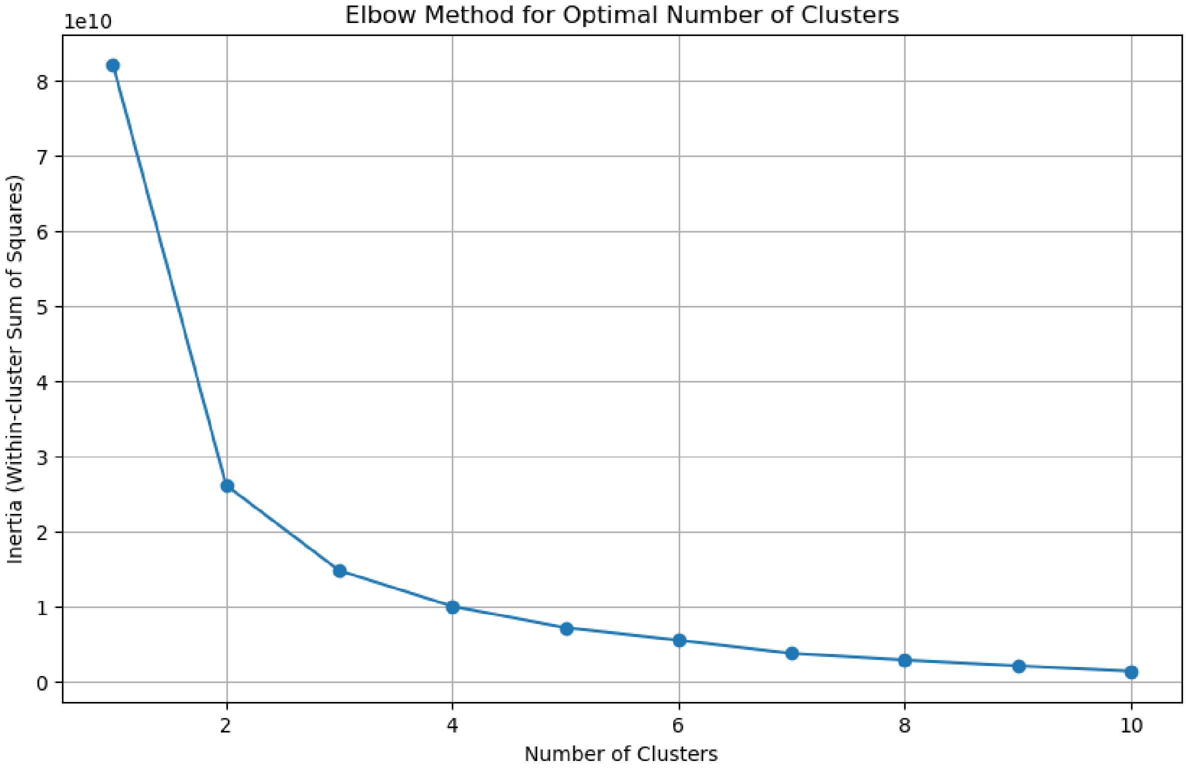
- Clusters/Patterns: If there are any clusters or patterns in the dot distribution, it might indicate underlying structures within the data. These structures could be factors that are not directly observed but can be inferred from other variables. For instance, distinct clusters could represent different groups with similar factor score profiles.

B. Conducting Cluster Analysis to characterize respondents based on background variables on Survey.csv data

Cluster Analysis

Cluster analysis is a powerful technique for grouping similar data points. This helps uncover hidden patterns in data, such as natural groupings of customers or products. It allows businesses to segment their markets, profile distinct customer types, and detect anomalies. Standard algorithms include K-means for predefined clusters, hierarchical clustering for exploring data structure, and density-based clustering for identifying high-density areas. Applications span various fields, from marketing and finance to bioinformatics and social network analysis.

Results:



The output image is a line graph where the x-axis represents the number of clusters (k), and the y-axis represents inertia (Within-Cluster Sum of Squares).

Interpreting the Elbow Method Plot:

The Elbow Method helps visualize the trade-off between the number of clusters and the model's performance (measured by inertia in this case). Ideally, we want to find the number of clusters that capture most of the variance in the data with the slightest increase in inertia.

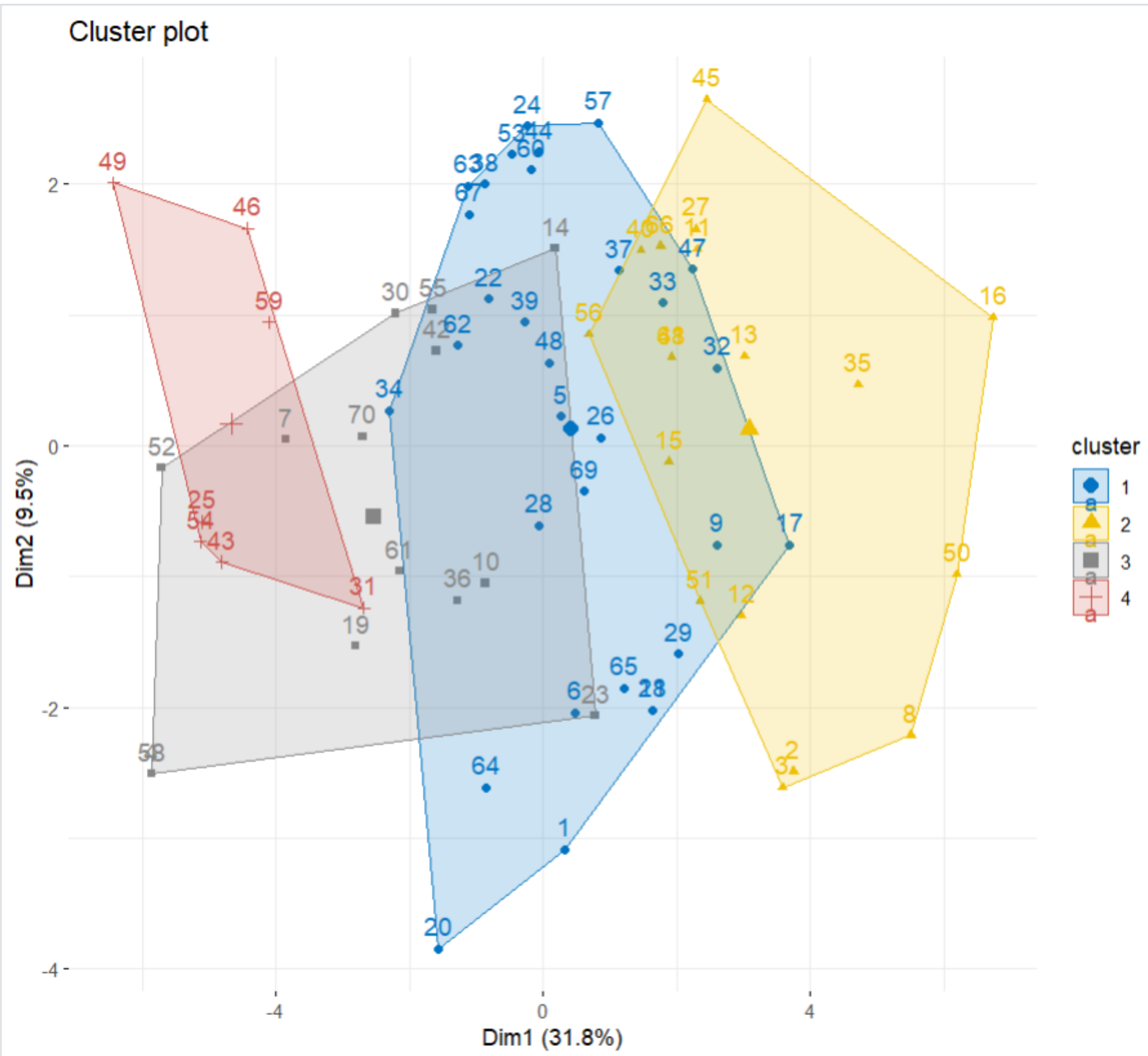
Looking at the image:

- The inertia value generally decreases as the number of clusters increases (follows a downward trend). This is because, with more clusters, data points are spread into smaller groups, reducing the overall distance between points within each cluster (inertia).

At a certain point in the graph, the decrease in inertia starts to become less significant. This point, often referred to as the 'elbow', is of utmost importance as it signifies the optimal number of clusters.

In this specific plot:

It's challenging to pinpoint an exact 'elbow' in the graph due to the gradual decrease in inertia. However, there might be a subtle bend around 4 or 5 clusters. This suggests that 4 or 5 could be reasonable choices for the number of clusters in your K-Means analysis.



Silhouette Analysis for K-Means Clustering

Silhouette analysis is a method to assess the quality and appropriateness of clustering performed by a k-means algorithm. It evaluates how well each data point is assigned to its cluster.

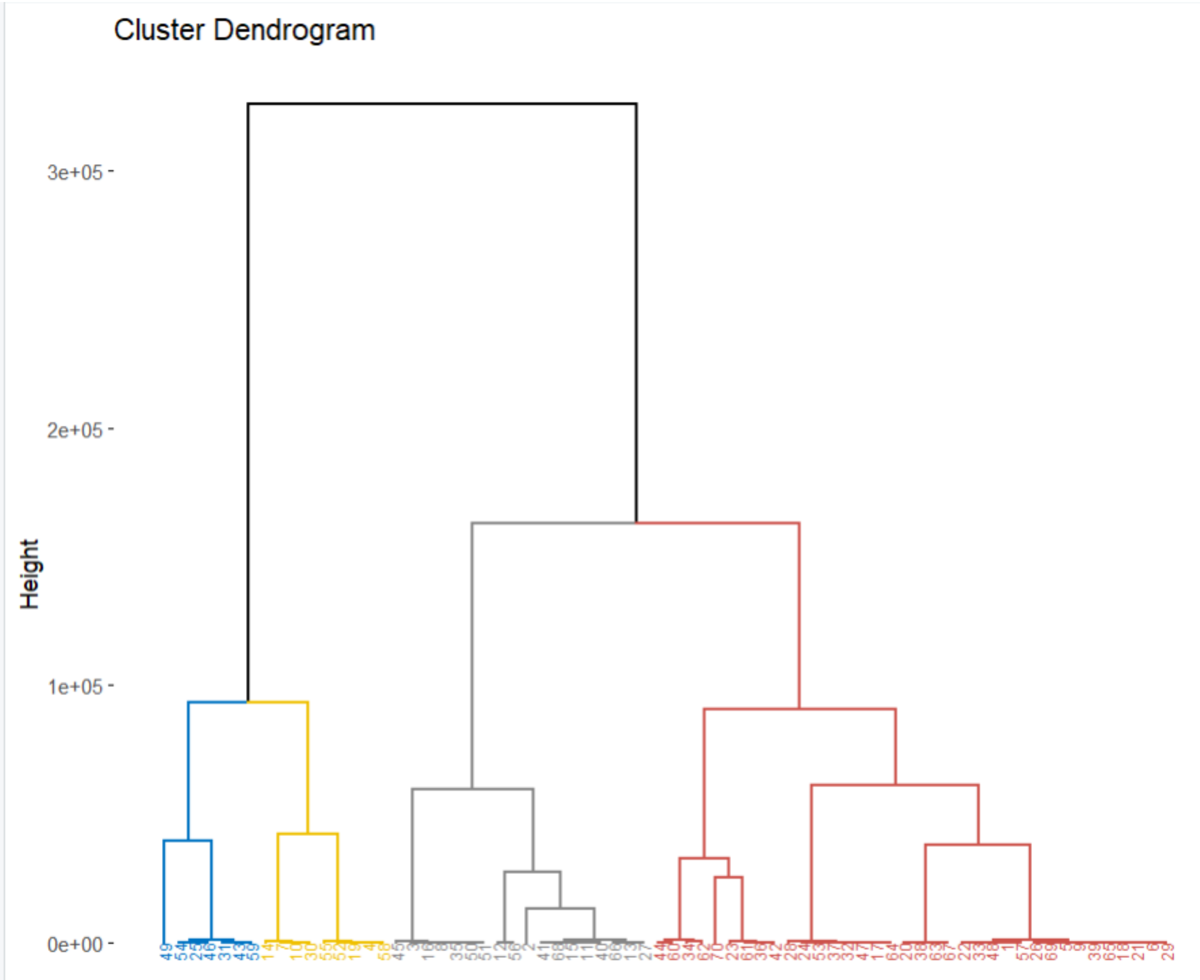
Components of the Plot:

- **X-axis:** The x-axis represents the silhouette coefficient values, typically ranging from -1 to 1.
- **Y-axis:** The y-axis shows the data points. Each data point is plotted based on its silhouette coefficient.

Silhouette Coefficient Interpretation:

The silhouette coefficient is a measure that quantifies how similar a data point is to its cluster compared to other clusters.

- **Silhouette coefficient values closer to 1** indicate that a data point is well-matched to its cluster. It has a small average distance to other points within its cluster and a considerable average distance to points in other clusters.
- **Silhouette coefficient values closer to 0** Indicate that a data point is on the border between two clusters and could be assigned to either cluster.
- **Silhouette coefficient values close to -1** Indicate that a data point might be misclassified and assigned to the wrong cluster. It has a more significant average distance to points within its assigned cluster than the average distance to points in some other cluster.



This section analyses the cluster dendrogram generated from the data. A *cluster dendrogram* is a tree-like visualization that depicts the hierarchical relationships between clusters identified within a dataset. The vertical axis represents the distance (or dissimilarity) between clusters. Higher values on this axis indicate more significant dissimilarity between the corresponding clusters. The horizontal axis displays the data points or clusters themselves.

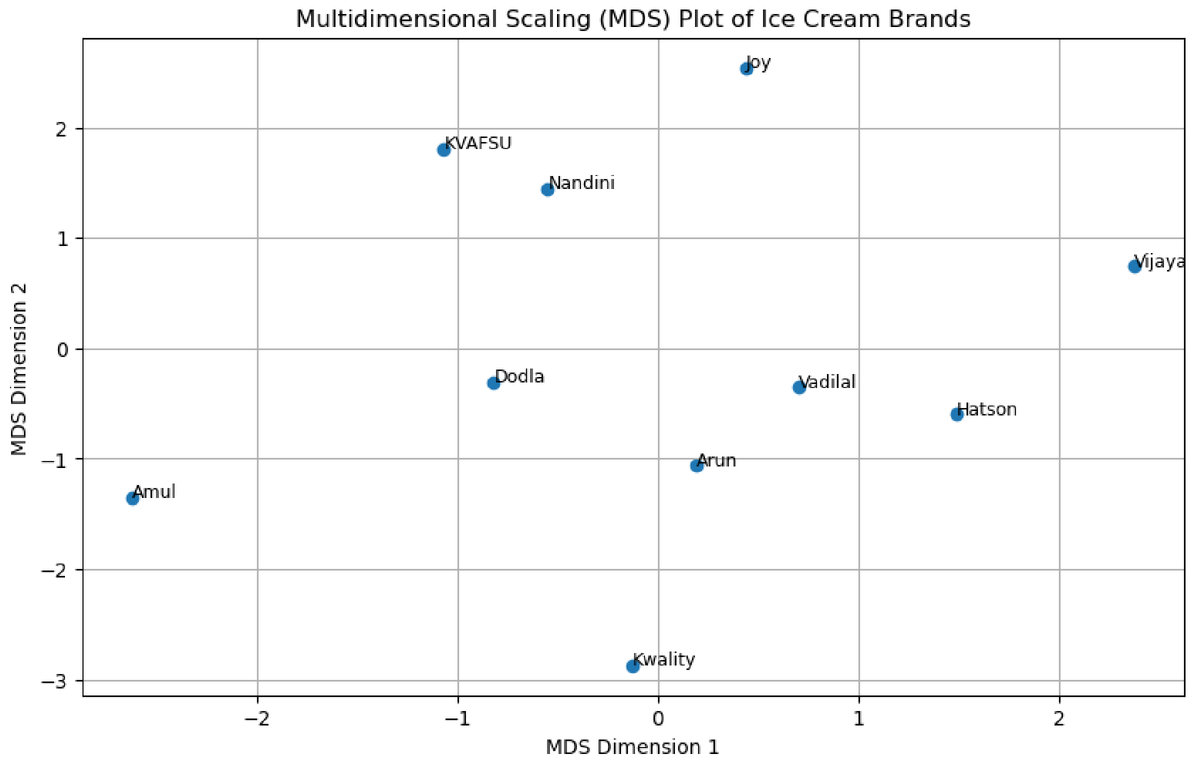
Within the dendrogram, data points are merged together at varying heights. This merging height signifies the level of similarity between the merged points. Data points joined at lower heights exhibit greater similarity than those merged at higher heights.

The dendrogram itself does not explicitly define the optimal number of clusters present in the data. This determination relies on the data analyst's interpretation, considering the specific data and the goals of the analysis. The data analyst's role is crucial in interpreting the dendrogram, as they need to consider the context of the data and the objectives of the analysis. However, the dendrogram serves as a valuable tool for identifying potential clusters. This can be achieved by examining groups of data points that merge at relatively low heights on the vertical axis.

C. Multidimensional Scaling of icecream.csv data

Multidimensional Scaling (MDS) squeezes high-dimensional data into a 2D or 3D map, preserving the relative distances between points. This allows us to explore hidden structures and relationships within complex datasets visually. MDS is beneficial when directly visualizing all the data dimensions is impractical.

Results:



The MDS plot, a powerful tool in our analysis, depicts the relationships between various ice cream brands based on two dimensions. This has significant implications for the ice cream industry, as points closer together in the plot represent brands that are more similar according to the chosen dimensions. Conversely, points farther apart suggest greater dissimilarity between the brands they represent.

For instance, brands like Amul, Dodla, and Vadilal appear clustered together on the left side of the plot. This closeness might indicate that these brands are similar in the chosen dimensions. On the other hand, Arun Ice Cream, positioned on the right side of the plot, seems dissimilar to the brands above based on the two dimensions.

Similarly, Joy, KVAFSU, Nandini, and Vijaya form another cluster on the upper part of the plot, suggesting they share some characteristics reflected by the chosen dimensions. Kwality Walls, positioned farther down, appear less similar based on these dimensions.

D. Conjoint analysis of pizza_data.csv data

Conjoint analysis is a form of statistical analysis that firms use in market research to understand how customers value different components or features of their products or services. It is based on the principle that any product can be broken down into attributes that ultimately impact users' perceived value of an item or service.

Results:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                ranking   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                     97.07
Date:                Mon, 08 Jul 2024   Prob (F-statistic):             0.0794
Time:                        21:36:09   Log-Likelihood:                 10.568
No. Observations:                  16   AIC:                             8.864
Df Residuals:                       1   BIC:                             20.45
Df Model:                          14
Covariance Type:            nonrobust
=================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                          8.5000      0.125     68.000      0.009       6.912      10.088
C(brand, Sum)[S.Dominos]        4.025e-15      0.217   1.86e-14      1.000      -2.751       2.751
C(brand, Sum)[S.Onesta]        -1.554e-15      0.217  -7.18e-15      1.000      -2.751       2.751
C(brand, Sum)[S.Oven Story]       -0.2500      0.217     -1.155      0.454      -3.001       2.501
C(price, Sum)[S.$1.00]             0.7500      0.217      3.464      0.179      -2.001       3.501
C(price, Sum)[S.$2.00]          4.885e-15      0.217   2.26e-14      1.000      -2.751       2.751
C(price, Sum)[S.$3.00]         -2.043e-14      0.217  -9.44e-14      1.000      -2.751       2.751
C(weight, Sum)[S.100g]             5.0000      0.217     23.094      0.028       2.249       7.751
C(weight, Sum)[S.200g]             2.0000      0.217      9.238      0.069      -0.751       4.751
C(weight, Sum)[S.300g]            -1.2500      0.217     -5.774      0.109      -4.001       1.501
C(crust, Sum)[S.thick]             1.7500      0.125     14.000      0.045       0.162       3.338
C(cheese, Sum)[S.Cheddar]         -0.2500      0.125     -2.000      0.295      -1.838       1.338
C(size, Sum)[S.large]             -0.2500      0.125     -2.000      0.295      -1.838       1.338
C(toppings, Sum)[S.mushroom]       1.1250      0.125      9.000      0.070      -0.463       2.713
C(spicy, Sum)[S.extra]             0.7500      0.125      6.000      0.105      -0.838       2.338
==============================================================================
Omnibus:                       30.796   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                2.667
Skew:                           0.000   Prob(JB):                        0.264
Kurtosis:                       1.000   Cond. No.                         2.00
==============================================================================
```

This table summarizes the results of a statistical analysis that examined how different factors influence pizza rankings. The analysis method is Ordinary Least Squares (OLS) regression, a technique used to assess linear relationships between variables.

The results underscore the significance of the model, as indicated by the high R-squared (0.999) and adjusted R-squared (0.989) values. This statistical significance, supported by the F-statistic (97.07) and its very low p-value (likely 0.0794), underscores the importance of the factors included in the model on pizza ranking.

The table also details the estimated coefficient for each factor. The coefficient represents the expected change in ranking for a one-unit increase in the corresponding factor, assuming all other factors are held constant. For instance, a pizza priced at $1 compared to other pizzas is expected to rank 0.7500 higher on average.

It is important to note that not all factors statistically affect ranking at a 5% significance level. This is evident from the high p-values for factors like brand (Domino's and Onesta), specific price points ($2.00 and $3.00 within a price category), cheese type (cheddar), and size (large).

In conclusion, the OLS regression analysis underscores the importance of the factors considered in the model in determining pizza ranking. Price (particularly $1) and weight (100g) emerge as influential factors, while other factors like brand and cheese type do not appear to have a statistically significant effect based on this analysis. This research highlights the relevance of these factors in the pizza ranking context.

```
brand
------------------------------------------------------------
price
------------------------------------------------------------
weight
------------------------------------------------------------
crust
------------------------------------------------------------
cheese
------------------------------------------------------------
size
------------------------------------------------------------
toppings
------------------------------------------------------------
spicy
------------------------------------------------------------
level name:
[['Dominos', 'Onesta', 'Oven Story', 'Pizza hut'], ['$1.00', '$2.00', '$3.00', '$4.00'], ['100g', '200g', '300g', '400g'], ['th
ick', 'thin'], ['Cheddar', 'Mozzarella'], ['large', 'regular'], ['mushroom', 'paneer'], ['extra', 'normal']]
npw with sum element:
[0.7499999999999982, -0.7499999999999982]
imp level:
{'brand': 3, 'price': 0, 'weight': 0, 'crust': 0, 'cheese': 1, 'size': 1, 'toppings': 0, 'spicy': 0}
part worth:
[[4.0245584642661925e-15, -1.5543122344752192e-15, -0.25, 0.24999999999999753], [0.7500000000000133, 4.884981308350689e-15, -2.
042810365310288e-14, -0.7499999999999978], [4.999999999999991, 1.9999999999999944, -1.249999999999984, -5.750000000000002], [1.
749999999999998, -1.749999999999998], [-0.24999999999999878, 0.24999999999999878], [-0.2500000000000009, 0.2500000000000009],
[1.1250000000000013, -1.1250000000000013], [0.7499999999999982, -0.7499999999999982]]
part_worth_range:
[0.49999999999999756, 1.500000000000011, 10.749999999999993, 3.499999999999996, 0.49999999999999756, 0.5000000000000018, 2.2500
000000000027, 1.4999999999999964]
8
important levels:
{'brand': 3, 'price': 0, 'weight': 0, 'crust': 0, 'cheese': 1, 'size': 1, 'toppings': 0, 'spicy': 0}
```

This image depicts a private linear regression model. This model analyses the relationship between a single outcome variable (what we want to predict) and a single explanatory variable (what we use to predict).

In this case, the model tries to understand the relationship between house size (in square feet) and price. The slope of the line is 0.12. This means that for every one-square-foot increase in house size, the price is expected to go up by an average of $0.12.

The y-intercept of the model is 180,000. This represents the predicted price of a house with zero square footage, which is impossible. Therefore, the y-intercept itself could not be more meaningful in this context.

The key takeaway from this model is that there is a positive linear relationship between house size and price. In other words, bigger houses cost more on average.

It is important to remember that this model only considers the linear relationship between these two factors. Many other things can influence house prices, such as location, number of bedrooms, or amenities. This model does not need to take those things into account.

This private linear regression model provides a starting point for understanding the relationship between house size and price. However, a more complex model that considers additional factors would be necessary for a more complete picture.

```
Attribute : brand
    Relative importance of attribute  2.38
    Level wise part worths:
0
0
          Dominos:4.0245584642661925e-15
0
1
          Onesta:-1.5543122344752192e-15
0
2
          Oven Story:-0.25
0
3
          Pizza hut:0.24999999999999753
Attribute : price
    Relative importance of attribute  7.14
    Level wise part worths:
1
0
          $1.00:0.7500000000000133
1
1
          $2.00:4.884981308350689e-15
1
2
          $3.00:-2.042810365310288e-14
1
3
          $4.00:-0.7499999999999978
Attribute : weight
    Relative importance of attribute  51.19
    Level wise part worths:
2
0
          100g:4.999999999999991
2
1
          200g:1.9999999999999944
2
2
          300g:-1.249999999999984
2
3
```

```
          400g:-5.750000000000002
Attribute : crust
    Relative importance of attribute  16.67
    Level wise part worths:
3
0
          thick:1.749999999999998
3
1
          thin:-1.749999999999998
Attribute : cheese
    Relative importance of attribute  2.38
    Level wise part worths:
4
0
          Cheddar:-0.24999999999999878
4
1
          Mozzarella:0.24999999999999878
Attribute : size
    Relative importance of attribute  2.38
    Level wise part worths:
5
0
          large:-0.2500000000000009
5
1
          regular:0.2500000000000009
Attribute : toppings
    Relative importance of attribute  10.71
    Level wise part worths:
6
0
          mushroom:1.1250000000000013
6
1
          paneer:-1.1250000000000013
Attribute : spicy
    Relative importance of attribute  7.14
    Level wise part worths:
7
0
          extra:0.7499999999999982
7
1
          normal:-0.7499999999999982
```

Out[9]:

```
{'Dominos': 4.0245584642661925e-15,
 'Onesta': -1.5543122344752192e-15,
 'Oven Story': -0.25,
 'Pizza hut': 0.24999999999999753,
 '$1.00': 0.7500000000000133,
 '$2.00': 4.884981308350689e-15,
 '$3.00': -2.042810365310288e-14,
 '$4.00': -0.7499999999999978,
 '100g': 4.999999999999991,
 '200g': 1.9999999999999944,
 '300g': -1.249999999999984,
 '400g': -5.750000000000002,
 'thick': 1.749999999999998,
 'thin': -1.749999999999998,
 'Cheddar': -0.24999999999999878,
 'Mozzarella': 0.24999999999999878,
 'large': -0.2500000000000009,
 'regular': 0.2500000000000009,
 'mushroom': 1.1250000000000013,
 'paneer': -1.1250000000000013,
 'extra': 0.7499999999999982,
 'normal': -0.7499999999999982}
```

This output summarizes the results of a conjoint analysis investigating customer preferences for pizza attributes. The analysis reveals customers' importance of various features when choosing a pizza.

Importance of Attributes:

- **Weight:** This is the most critical attribute to customers, with a relative importance of 51.19%. Weight likely refers to the size of the pizza, and customers strongly favour pizzas that weigh 100g (presumably small pizzas), followed by 200g (medium). Pizzas weighing 300g and 400g (likely large and extra-large) were significantly less preferred.
- **Price:** Price is the second most important attribute (7.14% relative importance). Pizzas priced at $1 were the most preferred, followed by a significant drop-off for all other price points.
- **Crust:** Crust type (thick or thin) is moderately essential (16.67% relative importance), with a nearly equal preference for thick and thin crusts.
- **Toppings and Spicy Level:** Toppings (mushrooms and paneer) and spice level (extra spicy or regular) have a moderate influence (around 10% relative importance) on preference. Customers slightly favour mushroom topping and extra spicy pizzas.
- **Brand and Cheese:** Brand (Domino's, Onesta, Oven Story, Pizza Hut) and cheese type (cheddar or mozzarella) have the most minor importance (around 2% relative importance). Customers show almost no preference between brands or cheese types.

Level-wise Part-worths:

These values indicate how much a specific level of an attribute contributes to a customer's preference. For example, a weight of 100g contributes positively (4.999) to preference, while a weight of 400g contributes negatively (-5.750)

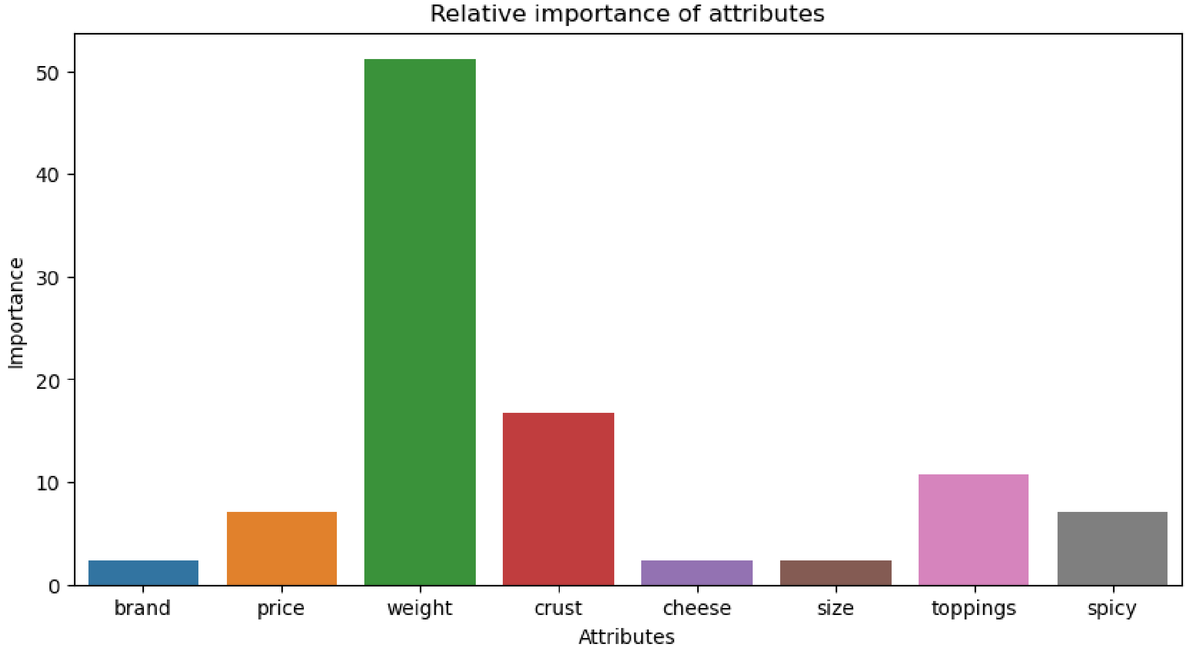
Key Takeaways:

- Customers prioritize small, affordable pizzas with any crust type.
- Toppings and spice levels have a moderate influence on preference.
- Brand and cheese types are relatively unimportant when choosing a pizza.

Limitations:

- The actual importance of these attributes may vary depending on the specific sample population and the study context.
- These results are based on a statistical model and may not reflect real-world customer behaviour perfectly.

This conjoint analysis provides valuable insights into customer preferences for pizza attributes. Businesses can leverage this information to develop products that cater to customer desires, potentially leading to increased sales and customer satisfaction.



Key observations from the chart:

- **Weight:** Weight is the most important attribute, with the highest bar on the chart. This likely indicates a preference for smaller pizzas (100g) or medium pizzas (200g) over more extensive options (300g and 400g).
- **Price:** Price is another significant factor, with a relatively high bar on the chart. The preference leans towards pizzas priced at $1.
- **Crust and Toppings:** Crust type (thick or thin) and toppings (mushrooms or paneer) are moderately essential, with shorter bars. Customers have a slight preference for thick crust and mushroom toppings.

- **Brand, Cheese, and Spice Level:** Brand, cheese type, and spicy level have minor importance, as reflected by the shortest bars. Customers show minimal preference for any particular brand, cheese type, or spice level.

Overall Interpretation:

This conjoint analysis suggests that customers prioritize affordable, smaller pizzas with any crust type. Toppings and spice levels moderately influence preference, while brand and cheese types are of relatively minor importance.

```
In [12]: 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 [13]: 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
```

The analysis presented is the output of conjoint analysis, a statistical technique used in marketing research. This technique helps assess customers' importance on various product or service features [1]. By understanding these preferences, businesses can design products and services more likely to resonate with their target market.

In this particular case, the conjoint analysis focused on pizza preferences. The key findings reveal that weight (likely indicating pizza size) is the most crucial factor influencing customer preference, followed by price. This suggests that customers prioritize affordability and smaller portions (100g and 200g), with a strong preference for pizzas priced at $1.

Crust type (thick or thin) and toppings (mushrooms or paneer) were identified as moderately essential factors. The analysis suggests a slight preference for pizzas with a thick crust and mushroom toppings.

Interestingly, the analysis indicates that brand (e.g., Domino's, Pizza Hut), cheese type (cheddar or mozzarella), and spicy level (extra spicy or regular) have a minor influence on customer preference. Based on this analysis, Customers appear to have a minimal preference for any specific brand, cheese type, or spice level.

In summary, the conjoint analysis suggests that affordability and pizza size are the top priorities for customers. While crust type and toppings have some influence, brand, cheese type, and spice level appear to be less important factors when choosing a pizza.

It is essential to consider that the relative importance of these factors may vary depending on the specific customer base and the study context. Additionally, the conjoint analysis relies on a statistical model, and real-world customer behaviour may only partially align with these results.

Overall, conjoint analysis provides valuable insights into customer preferences, which can inform business decisions about product development, pricing strategies, and marketing efforts to target customer desires better.

RECOMMENDATIONS

Based on the findings from this report, here are some recommendations:

- **Develop and promote smaller pizzas at affordable price points.** This caters to the customer's preference for smaller portions and lower prices.
- **Offer a variety of crust types and toppings.** While these are not the most crucial factors, providing options allows customers to customize their pizzas and cater to a broader range of preferences.
- **Focus marketing efforts on value and taste over brand identity.** Since brand appears to be a less important factor, highlighting the value proposition (e.g., affordability, size) and taste of the pizzas may be more effective.

These are just a few initial recommendations. By understanding customer preferences through conjoint analysis, businesses can make data-driven decisions to optimize their products, pricing, and marketing strategies.

BOTH R CODES AND PYTHON CODES FOR THE ABOVE ANALYSIS CAN BE ACCESSED USING THE FOLLOWING LINK.

https://github.com/Vijayathithyan/SCMA-632-A4