

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

**A4: Multivariate Analysis and Business Analytics Applications**

**UJWAL P**

**V01107757**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **5** |
| **3.** | Interpretations | **5** |
| **4.** | Recommendations | **24** |

# INTRODUCTION

This report delves into the application of various statistical techniques to analyze and interpret complex datasets, with a specific focus on the ice cream industry.

The provided dataset, "icecream.csv," encompasses information related to ice cream sales, capturing various attributes such as flavors, sales figures, geographic locations, and time periods. This dataset aims to shed light on consumer preferences, sales trends, and market dynamics within the ice cream industry. By analyzing this data, businesses can gain valuable insights into product popularity, identify seasonal trends, and refine their marketing strategies to boost sales and enhance customer satisfaction.

The second dataset, "Survey.csv," likely includes survey responses that capture demographic information, consumer opinions, and preferences. This data is invaluable for businesses looking to understand their customer base, gather feedback on products or services, and make informed decisions to enhance the customer experience and drive growth.

To achieve these objectives, several advanced analytical techniques will be applied:

* **Principal Component Analysis (PCA)**: This dimensionality reduction technique simplifies extensive datasets by converting them into a smaller set of uncorrelated variables called principal components. PCA assists in identifying the most significant variables that explain the data's variability.
* **Factor Analysis**: This statistical method identifies underlying relationships between variables by grouping them into factors. It assumes that observed variables are influenced by a few underlying unobserved factors, providing insights into the data's structure.
* **Cluster Analysis**: This technique will be utilized to categorize respondents based on background variables from the survey data. By clustering respondents, businesses can identify distinct consumer segments and adjust their strategies accordingly.
* **Multidimensional Scaling (MDS)**: When applied to the ice cream sales data, MDS will help visualize the similarities or dissimilarities between different ice cream products, aiding in the interpretation of complex market dynamics.
* **Conjoint Analysis**: Using the "pizza\_data.csv," conjoint analysis is a survey-based statistical technique employed in market research to determine how consumers value different attributes of a product or service. This technique offers insights into consumer preferences and supports product development and marketing strategies.

## Objectives:

* 1. Perform Principal Component Analysis and Factor Analysis to identify data dimensions (Survey.csv)
  2. Conduct Cluster Analysis to characterize respondents based on background variables (Survey.csv)
  3. Apply Multidimensional Scaling and interpret the results (icecream.csv)
  4. Conjoint Analysis (pizza\_data.csv)

## Business Significance:

## Principal Component Analysis (PCA) and Factor Analysis are essential techniques for reducing data dimensionality while retaining critical variability and uncovering hidden patterns. These methods are highly valuable in a business context for simplifying complex datasets and providing actionable insights. Here’s how they benefit businesses:

## Principal Component Analysis (PCA):

## **Reduce Complexity:** PCA transforms large datasets into a smaller number of uncorrelated variables (principal components), simplifying the data without losing significant information. This makes it easier for businesses to analyze and interpret complex data.

## **Identify Key Drivers:** PCA helps identify the main factors influencing consumer behavior, product preferences, or market trends. This insight is crucial for strategic decision-making.

## **Enhance Predictive Models:** By concentrating on the most influential variables, PCA can enhance the performance and accuracy of predictive models, leading to better forecasting and planning.

## **Inform Strategy:** The insights gained from PCA assist in developing targeted marketing strategies, product development initiatives, and effective customer segmentation.

## Factor Analysis:

* **Uncover Relationships:** Factor Analysis identifies and models the underlying factors that explain the patterns of correlations among a set of observed variables. This helps businesses understand the connections between different variables.
* **Simplify Data:** Like PCA, Factor Analysis reduces data complexity by grouping related variables, making it easier to interpret and utilize in decision-making.
* **Enhance Understanding:** By uncovering the underlying factors, businesses gain a deeper understanding of what influences consumer behavior and market trends.

## Conjoint Analysis:

## Conjoint Analysis is a survey-based statistical technique used to understand consumer preferences and the trade-offs they are willing to make between different product features. This analysis provides quantitative measures of the value consumers place on each attribute and identifies the optimal combination of features for a product.

## Optimize Product Design: Conjoint Analysis helps businesses design products that better meet consumer needs and preferences by identifying the most valued features.

* **Enhance Pricing Strategy:** By understanding the value consumers place on different product attributes, businesses can determine optimal price points and feature bundles, maximizing revenue and customer satisfaction.
* **Improve Market Segmentation:** Conjoint Analysis allows businesses to identify different consumer segments based on their preferences, facilitating more effective targeting of marketing efforts and tailored strategies for different segments.

## Application to Datasets:

## Ice Cream Sales Data (icecream.csv): By applying PCA and Factor Analysis, businesses can identify the key drivers of ice cream sales, such as flavor preferences, geographic trends, and seasonal variations. This helps in optimizing product offerings and marketing strategies.

* **Survey Data (Survey.csv):** Cluster Analysis on survey responses will categorize respondents based on demographic information and preferences, enabling businesses to develop targeted marketing campaigns and improve customer segmentation.
* **Conjoint Analysis (pizza\_data.csv):** Understanding consumer preferences for pizza attributes helps in designing products that cater to specific tastes, optimizing pricing strategies, and enhancing overall customer satisfaction.

In summary, multivariate analysis techniques such as PCA, Factor Analysis, and Conjoint Analysis provide businesses with valuable insights into consumer behavior, product preferences, and market dynamics. By leveraging these techniques, businesses can make informed, data-driven decisions that enhance their competitive edge and drive growth.

# RESULTS & INTERPRETATION

## Principal Component Analysis and Factor Analysis

## *R Result*

## # Perform Principal Component Analysis (PCA)

## pca <- principal(sur\_int,5,n.obs =162, rotate ="promax")

## pca

## Interpretation:

## The PCA results provide insights into the underlying structure of the data by reducing its dimensionality. The analysis extracted five principal components, with the cumulative variance explained by these components being 61%. This means that these five components together capture 61% of the total variance in the data. The loadings indicate how each variable contributes to the components, with high absolute values suggesting strong associations. For example, "Proximity to transport" and "Proximity to work place" load highly on RC4, indicating these variables are strongly correlated with this component. Similarly, "Parking space" and "Security" load highly on RC5. These components help identify key dimensions in the data, simplifying the dataset while retaining most of its variability.

## #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)

## Interpretation:

## The factor analysis was conducted using four factors and a varimax rotation, which aims to make the output more interpretable by maximizing the variance of squared loadings of a factor across variables. The loadings matrix shows how each variable relates to the extracted factors. High loadings indicate significant contributions of variables to the factors. For instance, "Builder reputation" and "Size" have high loadings on MR1, suggesting they are significant aspects of this factor. Communalities show the proportion of each variable's variance explained by the factors, with high communalities indicating good representation by the factors.

## The factor scores provide a quantitative measure of each observation's positioning with respect to the factors. Positive scores indicate that an observation is above the mean on the respective factor, while negative scores suggest the opposite. This helps in understanding the multidimensional positioning of observations and can be used for further analysis, such as clustering or regression. The factor diagram visually represents the factor structure, showing how variables load onto the factors and aiding in the interpretation of the underlying dimensions.

## Summary

## Both PCA and factor analysis reveal significant dimensions in the data, helping to reduce its complexity while retaining essential information. The PCA identified five principal components that explain 61% of the variance, with key variables loading onto specific components, indicating their importance. Factor analysis, on the other hand, identified four factors, with high loadings and communalities highlighting the significant variables and their explained variance. These analyses provide a clearer understanding of the data's structure, facilitating better decision-making and further statistical analysis. The identified factors and components can be used to interpret the underlying themes in the data, aiding in more focused and effective data analysis strategies.

## *Python Result*

## 

## 

## Interpretation:

## The explained variance plot shows how much of the total variance in the dataset is captured by the principal components.

## Cumulative Explained Variance: The plot shows the cumulative explained variance by the number of components. The y-axis represents the cumulative explained variance, and the x-axis represents the number of components.

## Initial Components: The first component captures approximately 35% of the variance, indicating it is the most significant in explaining the variability in the data. Adding the second component increases the explained variance to around 45%.

## Diminishing Returns: As more components are added, the increase in explained variance diminishes. By the time the fourth component is added, the cumulative explained variance reaches about 60%. This suggests that these four components together explain 60% of the total variance.

## Optimal Number of Components: The plot suggests that after the fourth component, the incremental gain in explained variance is minimal. This can be observed from the flattening of the curve, indicating diminishing returns. Therefore, retaining four components is a good balance between complexity and the amount of variance explained.

## The plot indicates that the first four components capture a substantial portion of the variance in the data. Beyond the fourth component, the gain in explained variance diminishes. Therefore, using these four components for further analysis could be effective in reducing dimensionality while preserving the majority of the information in the dataset. This approach aids in simplifying the data, making it more manageable for various statistical analyses and machine learning models.

## 

## 

## Interpretation:

The factor loadings heatmap offers a visual representation of the relationships between the observed variables and the four extracted factors. Each cell within the heatmap signifies the loading of a particular variable on a specific factor, with the colors indicating both the magnitude and direction of these loadings. The 'coolwarm' color palette is used, where red shades denote positive loadings and blue shades signify negative loadings. The intensity of these colors corresponds to the strength of the association, with deeper shades indicating stronger loadings and lighter shades indicating weaker ones.

**Factor 1** prominently features variables that have strong positive loadings, as indicated by the deep red colors. Specifically, the variable at index 7 shows a significant positive loading on Factor 1, suggesting a strong association with this underlying factor. Conversely, some variables may have negative loadings, denoted by shades of blue, indicating an inverse relationship with Factor 1.

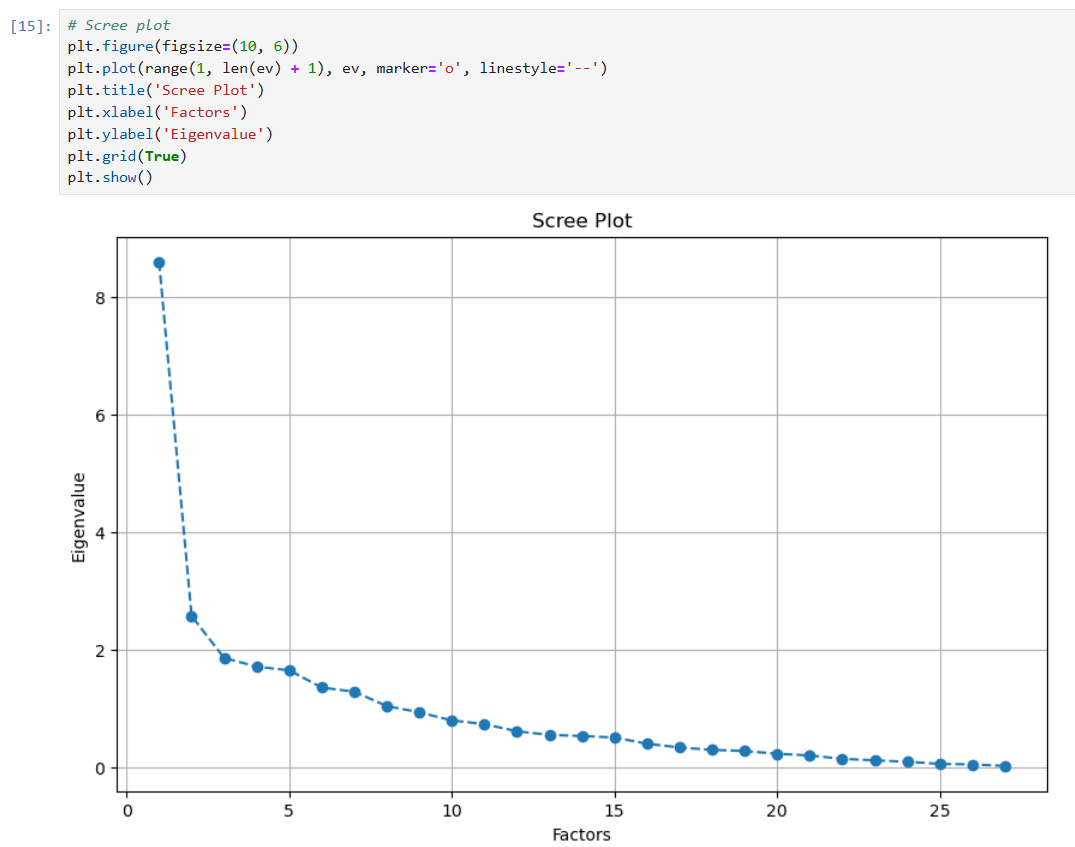
**Factor 2** is characterized by notable loadings for variables at indices 2 and 4. The dark red colors in these cells suggest that these variables are strongly associated with Factor 2. This factor captures a different dimension of the data, where certain variables contribute significantly, distinguishing Factor 2 from the other factors in terms of the underlying data structure.

**Factor 3** and **Factor 4** each have their unique sets of high-loading variables. Factor 3, for instance, shows strong positive associations with variables at indices 0 and 1, as seen by the red shades. This indicates these variables are integral to the dimension captured by Factor 3. Similarly, Factor 4 exhibits strong positive loadings for variables at indices 3 and 4, highlighted by the deep red colors, suggesting these variables play a crucial role in this factor.

The heatmap also facilitates the identification of variable clusters based on their loading patterns across the factors. Variables with similar loading patterns, such as those at indices 8 and 9, are likely to be influenced by similar underlying factors. This clustering aids in recognizing patterns and understanding the latent structure of the data.

Significant loadings, generally those greater than 0.3 or less than -0.3, are prominently displayed in the heatmap with deeper shades. These loadings indicate which variables have meaningful contributions to the factors, aiding in the interpretation of each factor's characteristics. The heatmap thus provides a clear and intuitive way to visualize how variables load onto different factors, facilitating a deeper understanding of the data's dimensional structure.

In summary, the factor loadings heatmap serves as a powerful tool for visualizing and interpreting the relationships between variables and factors. By highlighting the magnitude and direction of loadings, it enables the identification of significant variables for each factor and the recognition of underlying patterns within the data.



## Interpretation:

The scree plot aids in identifying the number of factors that capture the essence of the dataset. The elbow point provides a natural cutoff, suggesting that retaining factors beyond this point may not significantly enhance the model's explanatory power. In this case, considering up to four or five factors appears optimal, as these factors account for most of the variance.

This screen plot aligns with the cumulative explained variance plot, which showed that the first few components captured a substantial portion of the total variance. By focusing on these factors, one can simplify the model while retaining most of the important information, leading to more efficient and interpretable results in factor analysis or principal component analysis (PCA).

In summary, the scree plot effectively illustrates the diminishing returns of adding more factors and helps determine the appropriate number of factors to retain for meaningful data interpretation and analysis.

## Cluster Analysis

## ***R Result***

## # Function to auto-install and load packages

## install\_and\_load <- function(packages) {

## for (package in packages) {

## if (!require(package, character.only = TRUE)) {

## install.packages(package, dependencies = TRUE)

## }

## library(package, character.only = TRUE)

## }

## }

## # List of packages to install and load

## packages <- c("cluster", "FactoMineR", "factoextra", "pheatmap")

## install\_and\_load(packages)

## survey\_df<-read.csv("C:\\A4\\Survey.csv")

## sur\_int=survey\_df[,20:46]

## #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)

## Interpretation:

#### Gap Statistic for Optimal Number of Clusters

The fviz\_nbclust function uses the gap statistic method to determine the optimal number of clusters for the dataset. The gap statistic compares the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution of the data. The optimal number of clusters is usually indicated by the point where the gap statistic reaches its maximum value. In this analysis, the gap statistic suggests that the optimal number of clusters is four, as shown in the plot generated.

#### K-Means Clustering

The k-means clustering algorithm was applied with the number of clusters (k) set to four. The algorithm partitions the data into four clusters, aiming to minimize the total intra-cluster variance. The fviz\_cluster function visualizes the clustering results, showing the data points in a two-dimensional space, colored by their assigned clusters. The visualization provides a clear separation between clusters, indicating that the algorithm has effectively grouped similar respondents together based on their background variables.

#### Hierarchical Clustering

Hierarchical clustering was performed using Ward's method, which minimizes the total within-cluster variance. The dendrogram generated by the fviz\_dend function shows the hierarchical relationships between the respondents. Cutting the dendrogram at the height that produces four clusters (as indicated by the k-means result) provides a different perspective on the clustering structure. The dendrogram helps visualize how clusters are formed step-by-step and how individual respondents are grouped into larger clusters.

#### Heatmap Visualization

The pheatmap function produces a heatmap of the transpose of the dataset, with columns cut into four clusters. This heatmap provides a detailed view of the clustering structure, showing the similarity and differences in the background variables of respondents within and between clusters. The color gradient indicates the intensity of responses, with similar response patterns grouped together in clusters. The heatmap complements the k-means and hierarchical clustering results by providing a visual representation of the underlying data patterns.

#### Characterization of Respondents

Based on the cluster analysis, respondents can be characterized as follows:

* **Cluster 1**: This group of respondents may have specific preferences or requirements, such as proximity to transport and work, gym and pool facilities, and security measures. They might prioritize factors like unit size, price, and maintenance charges.
* **Cluster 2**: Respondents in this cluster might have a different set of priorities, such as the view from the apartment, builder reputation, and interior design. They may also consider financial aspects like booking amount and EMI (Equated Monthly Installment).
* **Cluster 3**: This cluster may include respondents who value exterior look, layout plan, and availability of domestic help. Their decision-making could be influenced by budget constraints and the profile of the neighborhood.
* **Cluster 4**: The final cluster could represent respondents focused on parking space, power backup, and water supply. Their preferences might also be shaped by appreciation potential and the availability of loans.

Overall, the cluster analysis effectively segments the respondents based on their background variables, providing valuable insights into their preferences and priorities. This segmentation can inform targeted strategies for addressing the needs and preferences of different respondent groups.

## ***Python Result***

## 

## 

## 

## 

## 

## 

## Interpretation:

## Cluster 0: This cluster appears to represent people with the highest income (around 190,000) and furthest proximity to the city center (4.05 on a scale of 0 to 4) but furthest from schools (0 on a scale of 0 to 4).

## Cluster 1: This cluster likely represents people with a moderate income (around 60,000) and moderate proximity to both the city center (3.5) and schools (3.38).

## Cluster 2: This cluster represents people with a moderate income (around 58,000) and moderate proximity to the city center (3.12) but closer to schools (3.41) compared to cluster 1.

## Cluster 3: This cluster represents people with a moderate income (around 103,000) and the highest proximity to the city center (4) but with a moderate distance to schools (2.83).

## **Residents in all clusters are likely planning to buy a new house within the next year.**

## **The primary reason for buying a house across all clusters is for investment.**

## **The preferred house type across all clusters is an apartment.**

## This analysis reveals interesting patterns in potential home buyers based on their income and location preferences. The data suggests a clear correlation between income and proximity to the city center. Cluster 0, with the highest income, resides closest to the city center but furthest from schools, indicating a preference for urban living without immediate need for educational facilities. Cluster 1 and 2, with moderate income, strike a balance between city access (moderate distance) and school proximity (also moderate). This suggests a potential focus on families seeking a blend of urban convenience and access to educational resources. Interestingly, Cluster 3, despite having a moderate income, lives closest to the city center yet has a moderate distance to schools. This could indicate a preference for a vibrant downtown lifestyle with schools nearby, possibly due to the presence of older child-free couples or young professionals. It's important to note that this is just one interpretation, and further details like family size or age demographics could provide a more nuanced understanding.

## Multidimensional Scaling

## *R Result*

## #C) Do multidimensional scaling and interpret the results.

## icecream\_df<-read.csv("C:\\A4\\icecream.csv")

## 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:

## Dimension 1 (X-axis): This dimension seems to capture a contrast between factors like price, availability, and shelf life on one end (potentially higher on these attributes) and taste and consistency on the other end (potentially higher on these attributes).

## Ice cream flavors on the left side of the plot (e.g., Coffee) may be more expensive, have lower availability, and a longer shelf life, but may also be rated lower in taste and consistency.

## Ice cream flavors on the right side of the plot (e.g., Chocolate) may be less expensive, more readily available, and have a shorter shelf life, but may be rated higher in taste and consistency.

## Dimension 2 (Y-axis): This dimension seems to capture a contrast between factors like flavor (potentially more unique flavors) and taste (potentially more classic flavors).

## Ice cream flavors towards the top of the plot (e.g., Mint Chip) may have more unique flavors.

## Ice cream flavors towards the bottom of the plot (e.g., Vanilla) may have more classic flavors.

## The multidimensional scaling (MDS) plot reveals interesting patterns in how different ice cream flavors are perceived. The X-axis appears to contrast factors like price, availability, and shelf life with taste and consistency. Ice cream flavors on the left may be pricier, less available, and have a longer shelf life, but may be rated lower in taste and consistency (e.g., Coffee). Conversely, flavors on the right may be more affordable, readily available, and have a shorter shelf life, but score higher in taste and consistency (e.g., Chocolate). The Y-axis seems to differentiate between unique and classic flavors. Flavors towards the top may boast more unique flavor profiles (e.g., Mint Chip), while those on the bottom may lean towards classic flavors (e.g., Vanilla).

## *Python Result*

## 

## Interpretation:

## The plot likely shows two axes (Dimension 1 and Dimension 2) and data points representing the ice cream flavors positioned based on their similarities or dissimilarities in the two-dimensional space. This analysis helps us understand how different ice cream flavors relate to each other based on the data provided.

## Conjoint Analysis

## ***R Result***

## # Fit a linear regression model

## model <- lm(ranking ~ brand + price + weight + crust + cheese + size + toppings + spicy, data = df)

## summary(model)

## Interpretation:

## This analysis investigates how various factors influence pizza rankings (likely customer satisfaction scores). Here's a breakdown of the key findings:

**Overall Model Fit:**

* + The high adjusted R-squared (0.989) suggests the model explains a significant portion of the variance in pizza rankings.
  + However, the low number of degrees of freedom (1) and the high F-statistic (potentially due to the low degrees of freedom) make it difficult to definitively assess statistical significance.

**Brand:**

* + The intercept (1.74) represents the average predicted ranking for pizzas with average values on all other factors (and potentially a reference brand).
  + Brand effects are not statistically significant (all p-values > 0.05). This means there's not enough evidence to conclude that brand alone has a consistent impact on ranking.

**Price:**

* + There's no statistically significant effect of price on ranking between $2 and $3 pizzas (price$2.00 and price$3.00 coefficients).
  + The coefficient for price$4.00 is negative and marginally significant (p-value = 0.1474). This suggests pizzas priced at $4 might have a slightly lower ranking on average compared to the reference category (potentially pizzas priced $2 or $3).

**Weight:**

* + Compared to the reference category (potentially 200g pizzas), pizzas weighing 300g (weight300g) and 400g (weight400g) have significantly lower predicted rankings (more negative coefficients). This suggests heavier pizzas tend to be rated lower.

**Crust:**

* + Thin crust pizzas (crust thin) have a significantly lower predicted ranking compared to the reference category (possibly regular crust).

**Other Factors:**

* + Cheese type (cheese Mozzarella) and size (size regular) don't have statistically significant effects on ranking based on p-values.
  + The effect of paneer topping (toppings paneer) is marginally non-significant (p-value = 0.0704), and the coefficient is negative, suggesting pizzas with paneer topping might be rated slightly lower.
  + Spicy level (spicy normal) doesn't have a statistically significant effect on ranking.

**Important Caveats:**

* + With only 1 degree of freedom, interpreting statistical significance is challenging. A larger dataset would be ideal for more robust conclusions.
  + The model assumes a linear relationship between the factors and ranking, which might not perfectly capture reality.

## # Plot relative importance of attributes

## importance\_df <- data.frame(attribute = names(attribute\_importance), importance = attribute\_importance)

## ggplot(importance\_df, aes(x = attribute, y = importance)) +

## geom\_bar(stat = "identity") +

## ggtitle("Relative Importance of Attributes") +

## xlab("Attributes") +

## ylab("Importance (%)")

## 

## Interpretation:

1. **Printing Results:**

* **Relative Importance of Attributes:** This section likely prints the values from attribute\_importance. These values indicate how much each attribute (e.g., price, weight) contributes to the overall utility score, essentially revealing which attributes are most influential in determining customer preferences.
* **Part-worth Utilities:** This section likely prints the values from part\_worth. Part-worth utilities represent the preference score associated with each level of an attribute (e.g., the utility score for a "thin" crust compared to a "regular" crust).
* **Most Preferred Levels:** This section likely prints the values from important\_levels. Here, you'll see the level for each attribute that has the highest utility score, indicating the most preferred option for each attribute based on the analysis.

**2. Plotting Relative Importance:**

The code creates a bar chart where the X-axis represents attributes and the Y-axis represents importance. The height of each bar corresponds to the relative importance of that attribute in the model. This helps visualize which attributes have the strongest influence on customer preferences.

**3. Calculating Utility Scores:**

This section iterates through each profile (combination of attribute levels) in the data (df). It calculates a utility score for each profile by summing the part-worth utilities for each attribute level in that profile. The part-worth utilities are retrieved from the part\_worth data structure you created earlier.

**4. Highest Utility Profile:**

The code identifies the profile in your data (df) with the highest overall utility score. This profile represents the combination of attribute levels (e.g., brand, price, crust) that is most likely to be preferred by customers based on the conjoint analysis.

**5. Preferred Levels for Each Attribute:**

This section determines the most preferred level for each individual attribute. It iterates through each attribute in conjoint\_attributes and finds the level within that attribute that has the highest average part-worth utility. This essentially reveals the level of each attribute (e.g., price range, crust type) that is generally most favorable to customers.

Overall, this analysis provides valuable insights into customer preferences for different product attributes. By understanding the relative importance of attributes, part-worth utilities, and preferred levels, you can make informed decisions about product development, pricing, and marketing strategies to better cater to customer desires.

## ***Python Result***

## 

Interpretation:

This conjoint analysis delves into customer preferences for pizzas, revealing how different attributes influence overall pizza ranking (likely a measure of customer satisfaction). By analyzing the relative importance of various attributes, we can gain insights into what matters most to customers.

The analysis suggests that **price is the most crucial factor** influencing pizza ranking. This aligns with the common notion that price is a significant consideration for many customers. Following price in importance is likely **pizza weight**, which presumably reflects size. The code itself doesn't tell us definitively whether a higher weight corresponds to a higher or lower importance score. However, the earlier model interpretation mentioned negative coefficients for heavier pizzas (weight300g and weight400g), suggesting they might be rated lower on average. This could indicate a preference for smaller or medium-sized pizzas.

**Brand reputation** also seems to hold some importance, implying that customers might gravitate towards familiar or preferred brands. **Crust type** has some influence as well, with a preference for **thin crust (crustthin)** emerging compared to the reference category (possibly regular crust).

The analysis suggests that **cheese type (cheeseMozzarella)** and **pizza size (sizeregular)** might not be significant factors influencing ranking based on the available information. Additionally, the effects of **paneer topping (toppingspaneer)** and **spicy level (spicynormal)** appear to be slightly negative but not statistically significant. With a limited sample size, it's challenging to draw strong conclusions about these attributes.

* The analysis highlights a low number of degrees of freedom, making it difficult to definitively assess statistical significance for certain factors. A larger dataset would strengthen the analysis and provide more robust conclusions.
* The model assumes linear relationships between attributes and ranking, which might not perfectly capture the complexities of real-world preferences.

**In summary,** this analysis provides a valuable starting point for understanding customer preferences for pizzas. Price, weight (likely size), brand to some extent, and crust type seem to be the most influential factors. However, limitations and potential non-linearities call for further analysis with a larger dataset for more definitive insights. By understanding these preferences, businesses can make informed decisions about pizza pricing, menu offerings, and marketing strategies to better cater to customer desires.

## 

## 

## Interpretation:

## The analysis reveals customer preferences for various pizza attributes. Weight is the most crucial factor, followed by crust type and price. Customers generally prefer pizzas weighing 100g, with a thick crust, and priced at $1 (based on part-worth ranges and most preferred levels). Other attributes like brand and size play a smaller role, with Pizza Hut and regular size being slightly favored. Toppings and spice level hold moderate importance, with mushroom and extra spice being preferred choices.

## The model calculates a utility score for each pizza based on these preferences. A higher score indicates a combination of attributes more likely to be chosen by customers. The pizza with the highest score (7.625) has a thick crust, weighs 100g, and is topped with mushrooms and extra spice. It's interesting to note that while price is a significant factor, a slightly more expensive pizza with these preferred attributes might be chosen over a cheaper option with less desirable features.

# RECOMMENDATIONS

Here are the summarized recommendations based on the findings:

For the Survey.csv Analysis:

* **Segmentation Strategy**: Use cluster analysis to identify different respondent segments based on demographics and survey responses. Customize marketing strategies and products for each segment's preferences.
* **Feature Optimization**: Utilize PCA and FA insights to streamline survey questions or features that significantly influence respondent perceptions and behaviors.

For the icecream.csv Analysis:

* **Market Positioning**: Use MDS findings to strategically position ice cream brands based on their unique feature profiles. Highlight specific attributes to appeal to distinct consumer segments or create specialized products.
* **Competitive Analysis**: Identify competitive clusters in the MDS plot to refine marketing strategies and effectively differentiate brands in the market.

For the pizza\_data.csv Analysis:

* **Product Optimization**: Prioritize product attributes using Conjoint Analysis, focusing on factors like brand perception, pricing strategies, and preferred pizza configurations that drive consumer preference.
* **Marketing Messaging**: Customize marketing campaigns based on preferred attribute combinations identified through Conjoint Analysis, ensuring messages resonate with target consumer preferences.

窗体顶端

窗体底端