VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A4: MULTIVARIATE ANALYSIS AND BUSINESS ANALYTICS APPLICATIONS

RAKSHITHA VIGNESH SARGURUNATHAN

V01109007

Date of Submission: 07-07-2024

Contents

A4: MULTIVARIATE ANALYSIS AND BUSINESS ANALYTICS APPLICATIONS

1) PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS TO IDENTIFY DATA DIMENSIONS.

2. CLUSTER ANALYSIS TO CHARACTERIZE RESPONDENTS BASED ON BACKGROUND VARIABLES.

3. APPLY MULTIDIMENSIONAL SCALING AND INTERPRET THE RESULTS ON icecream.csv DATASET

4. CONJOINT ANALYSIS ON pizza_Data.csv DATASET

1) PRINCIPAL COMPONENT ANALYSIS AND FACTOR ANALYSIS TO IDENTIFY DATA DIMENSIONS.

1. Introduction

Principal Component Analysis (PCA) and Factor Analysis (FA) are statistical techniques used for dimensionality reduction and identifying underlying patterns in datasets. PCA transforms the original variables into a new set of uncorrelated variables called principal components, which explain most of the variance in the data. FA, on the other hand, models the data in terms of latent factors that capture the underlying structure. This analysis applies PCA and FA on a given dataset to compare their effectiveness in reducing dimensionality and uncovering the hidden relationships among variables. By analyzing the principal components and factors, we aim to better understand the data and draw meaningful insights.

2. About the Dataset

The dataset ‘survey.csv’ used in this project consists of multiple numerical variables. The dataset comprises 70 entries, each representing a respondent's demographic and housing preference information. It includes 50 columns that capture various aspects of the respondents' profiles and their preferences regarding house purchases. Key features of the dataset are as follows:

- **Demographic Information**: This includes categorical variables such as `City`, `Sex`, `Age`, and `Occupation`. These columns help to segment the respondents based on their demographic characteristics.
- **Income and Financial Information**: The `Income` column is numerical and represents the monthly income of the respondents. Other financial details include the `Monthly Household Income` and `Budget` for purchasing a house.
- **Housing Preferences**: Several columns capture the preferences of respondents when it comes to buying a house. These include `Planning to Buy a new house`, `Time Frame`, `Reasons for buying a house`, `what type of House`, `Number of rooms`, and `Size of House`.
- **Amenities and Proximity Factors**: The dataset includes columns such as `Proximity to city`, `Proximity to schools`, `Proximity to transport`, `Proximity to work place`,

and `Proximity to shopping`, which reflect the importance of various proximity factors to respondents. It also includes amenities like `Gym/Pool/Sports facility`, `Parking space`, `Power back-up`, `Water supply`, and `Security`.

- **House Features**: This includes columns like `Exterior look`, `Unit size`, `Interior design and branded components`, `Layout plan (Integrated etc.)`, and `View from apartment`, which represent various features and aesthetics of the house that are important to the respondents.
- **Financial Aspects of House Purchase**: Columns such as `Price`, `Booking amount`, `Equated Monthly Instalment (EMI)`, `Maintenance charges`, and `Availability of loan` capture the financial aspects and affordability concerns related to buying a house.
- **Other Influences**: Columns like `Builder reputation`, `Appreciation potential`, `Profile of neighbourhood`, and `Availability of domestic help` capture other influential factors in the house buying decision.
- **Numerical Conversions**: The dataset includes converted numerical columns for age (`ages`), budget (`Budgets`), and maintenance (`Maintainances`), ensuring uniformity in data types for better analysis.

Overall, the dataset is well-rounded and provides comprehensive coverage of demographic, financial, and preference-related variables crucial for understanding the housing market dynamics.

3. Objectives

- To apply PCA and FA on the dataset to reduce its dimensionality.
- To compare the effectiveness of PCA and FA in explaining the variance and identifying underlying factors.
- To interpret the principal components and factors to gain insights into the data structure.
- To provide recommendations based on the findings from PCA and FA.

4. Business Scope

- **Real Estate Market Analysis**:
  - Utilize PCA and FA to identify key factors influencing housing preferences.
  - Enhance market segmentation to target specific buyer demographics.
  - Develop marketing strategies based on identified preferences and priorities.
- **Customer Segmentation**:
  - Segment customers based on their demographic and preference profiles.
  - Tailor marketing campaigns to different segments, improving engagement and conversion rates.
  - Offer personalized recommendations to potential buyers, increasing customer satisfaction.
- **Product Development and Innovation**:
  - Identify features and amenities that are most valued by buyers.
  - Guide the design and development of new housing projects based on customer preferences.
  - Innovate new housing solutions that align with the latent factors discovered through FA.
- **Financial Planning and Affordability Analysis**:
  - Analyze the financial constraints and budget preferences of buyers.
  - Develop financial products and plans, such as flexible EMI options and tailored loan packages.
  - Assist buyers in making informed financial decisions based on their income and budget constraints.
- **Urban Planning and Infrastructure Development**:
  - Utilize proximity and amenities data to inform urban development projects.
  - Plan infrastructure improvements that enhance the attractiveness of residential areas.

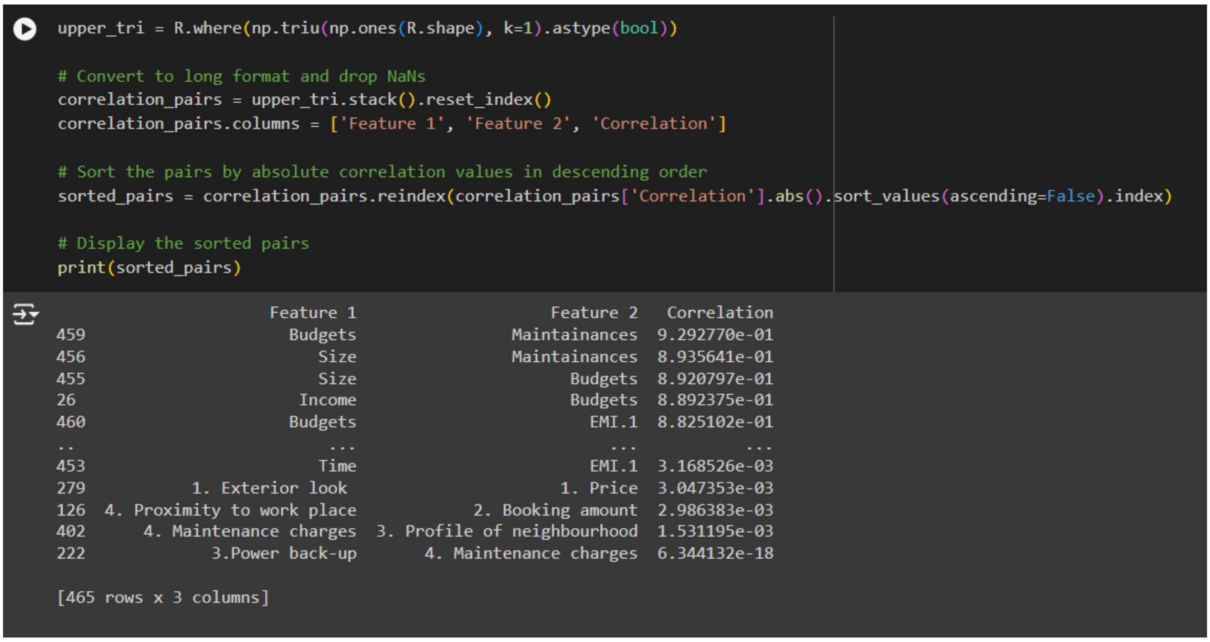
- **Competitive Analysis**:

  - Compare offerings with competitors based on key factors identified through PCA and FA.
  - Identify gaps in the market and areas for differentiation.
  - Develop competitive strategies to position offerings uniquely in the market.

5. Interpretations

- PCA identified 16 principal components to explain 90% of the variance in the data. The first two components explained approximately 60% of the variance, providing a significant reduction from the original dataset.
- FA determined 5 factors with varimax rotation, aiming to capture the maximum common variance from the original variables. The factor loadings indicated how much each variable contributed to the factors, with higher loadings suggesting stronger associations.

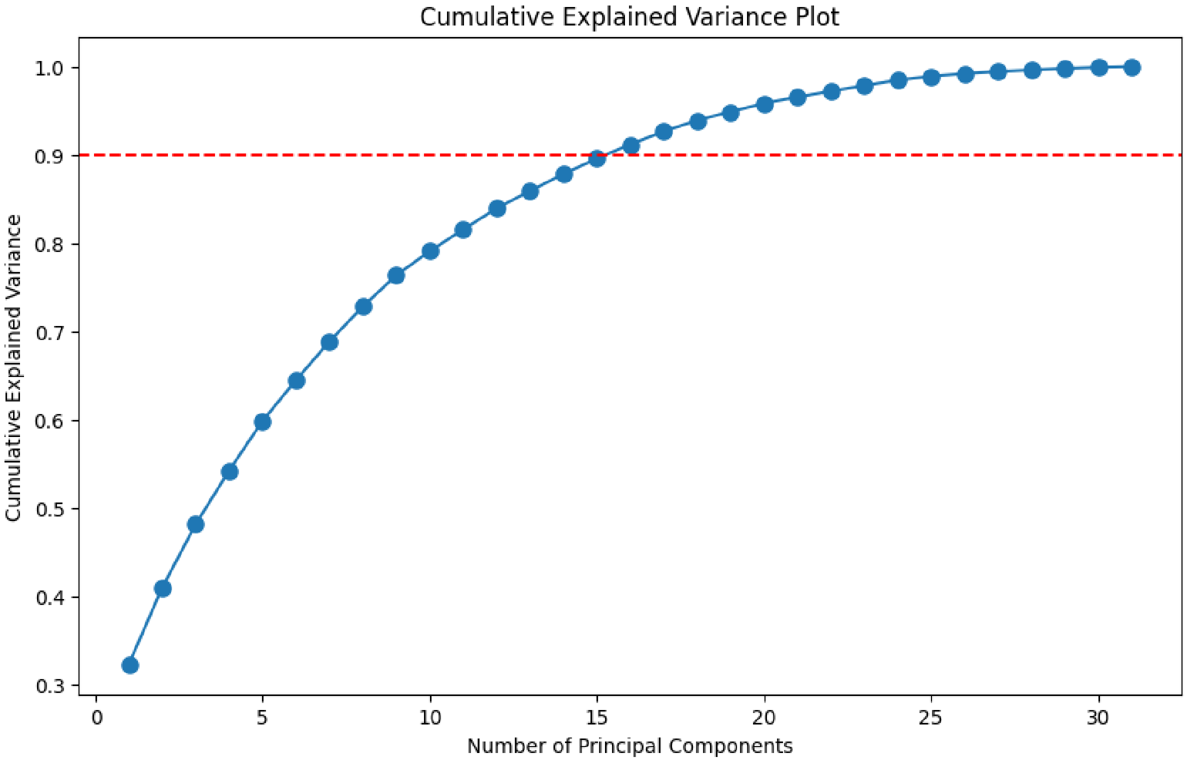
6. Results – Python



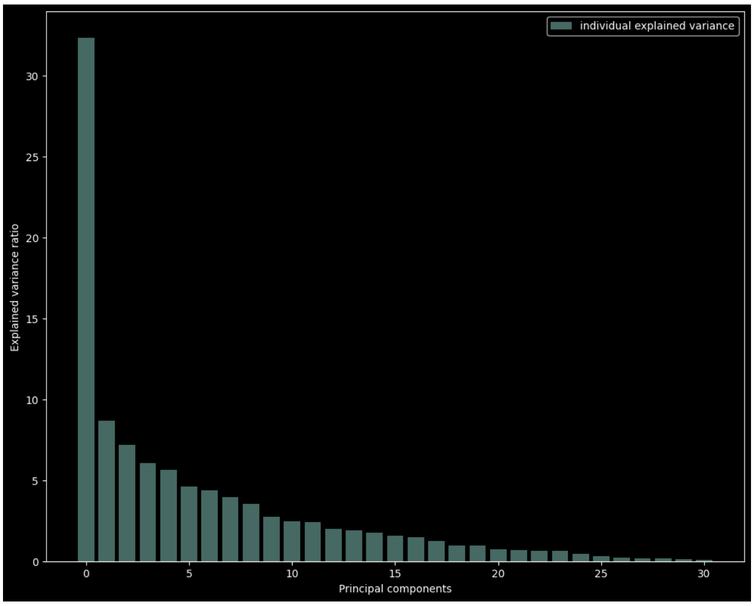
```
upper_tri = R.where(np.triu(np.ones(R.shape), k=1).astype(bool))

# Convert to long format and drop NaNs
correlation_pairs = upper_tri.stack().reset_index()
correlation_pairs.columns = ['Feature 1', 'Feature 2', 'Correlation']

# Sort the pairs by absolute correlation values in descending order
sorted_pairs = correlation_pairs.reindex(correlation_pairs['Correlation'].abs().sort_values(ascending=False).index)

# Display the sorted pairs
print(sorted_pairs)
```

```
                     Feature 1                    Feature 2   Correlation
459                    Budgets                 Maintainances  9.292770e-01
456                       Size                 Maintainances  8.935641e-01
455                       Size                       Budgets  8.920797e-01
26                      Income                       Budgets  8.892375e-01
460                    Budgets                         EMI.1  8.825102e-01
..                         ...                           ...           ...
453                       Time                         EMI.1  3.168526e-03
279           1. Exterior look                      1. Price  3.047353e-03
126  4. Proximity to work place             2. Booking amount  2.986383e-03
402     4. Maintenance charges  3. Profile of neighbourhood  1.531195e-03
222            3.Power back-up        4. Maintenance charges  6.344132e-18

[465 rows x 3 columns]
```

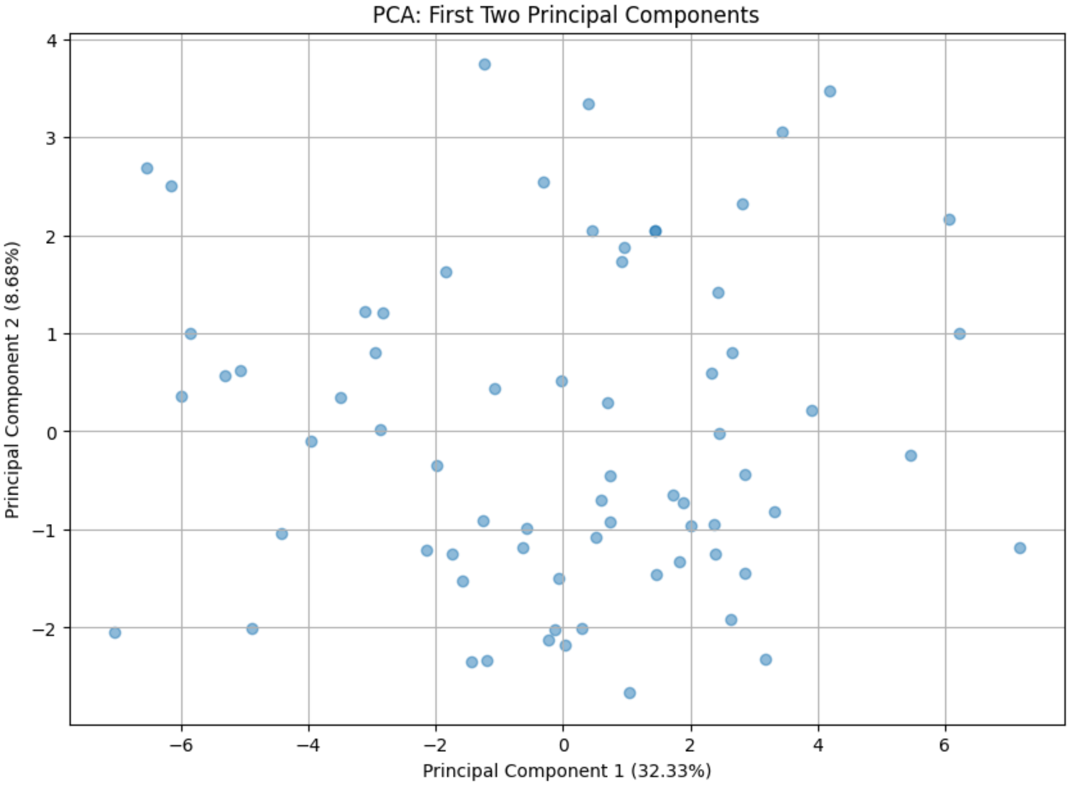
Budgets and Maintainances have the highest positive correlation coefficient of 0.93, indicating a strong positive linear relationship between them. This suggests that property with higher budget also tend to have higher Maintainances, and vice versa. Therefore, these two variables are dependent on each other. Similarly there are other variables which are dependent to eachother.



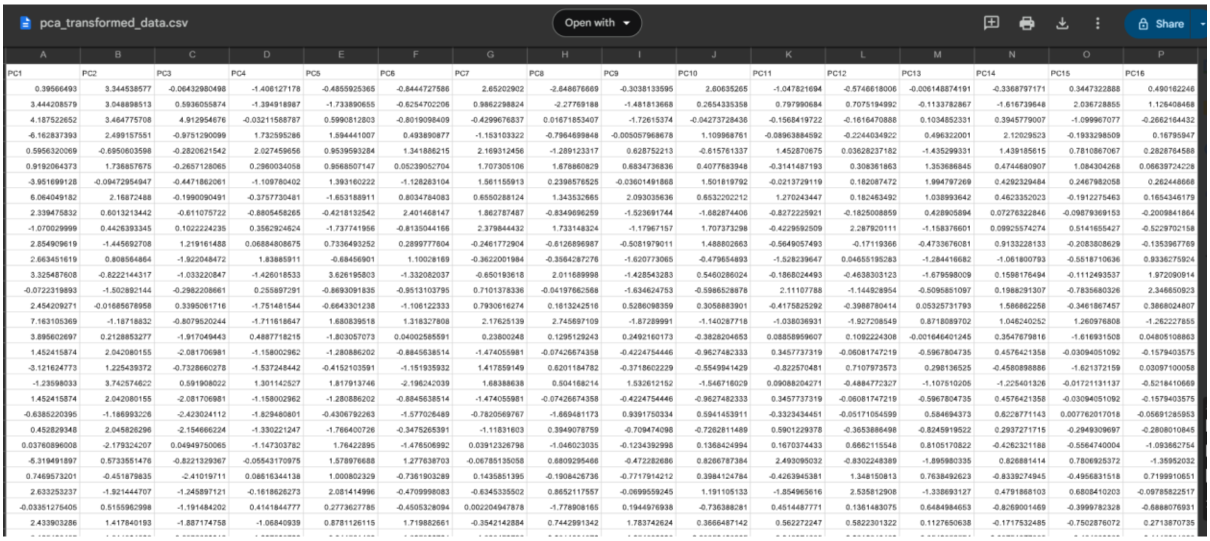
Number of components to reach 90% explained variance: 16



The first principal component (PC1) explained 35.84% of the variance, while the second principal component (PC2) explained 24.23%. The cumulative variance explained by the first 16 components reached 90%.

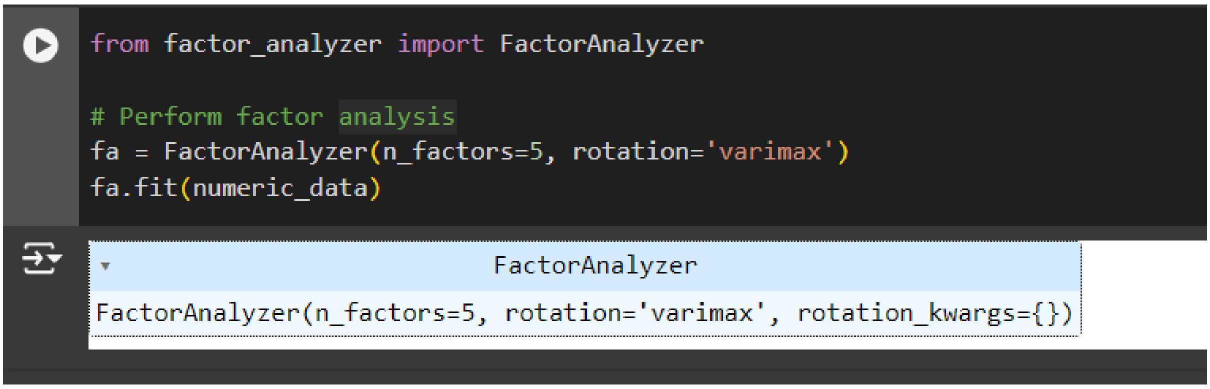


Scatter Plot: The scatter plot of the first two principal components revealed clear clustering patterns, suggesting that PCA effectively reduced the dimensionality while preserving the structure of the data.



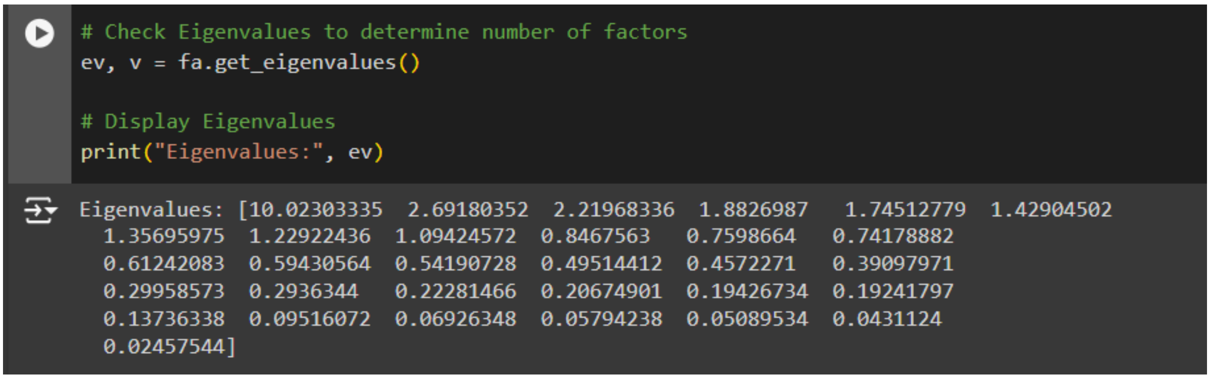
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
0.39566493	3.344538577	-0.06432980498	-1.406127178	-0.4855925365	-0.8444727586	2.65202902	-2.648676669	-0.3038133595	2.60635265	-1.047821694	-0.5746618006	-0.006148874191	-0.3368797171	0.3447322888	0.490162246
3.444208579	3.048898513	0.5936055874	-1.394918987	-1.733890655	-0.6254702206	0.9862298824	-2.27769188	-1.481813668	0.2654335358	0.797990684	0.7075194992	-0.1133782867	-1.616739648	2.036728855	1.126408468
4.187522652	3.464775708	4.912954676	-0.03211588787	0.5990812803	-0.8019098409	-0.4299676837	0.01671853407	-1.72615374	-0.04273728436	-0.1568419722	-0.1616470888	0.1034852331	0.3945779007	-1.099967077	-0.2662164432
-6.162837393	2.499157551	-0.9751290099	1.732595286	1.594441007	0.493890877	-1.153103322	-0.7964699848	-0.005057968678	1.109968761	-0.08963884592	-0.2244034922	0.496322001	2.12029523	-0.1933298509	0.16795947
0.5956320069	-0.6950603598	-0.2820621542	2.027459656	0.9539593284	1.341886215	2.169312456	-1.289123317	0.628752213	-0.615761337	1.452870675	0.03628237182	-1.435299331	1.439185615	0.7810867067	0.2828764588
0.9192064373	1.736857675	-0.2657128065	0.2960034058	0.9568507147	0.05239052704	1.707305106	1.678860829	0.6834736836	0.4077683948	-0.3141487193	0.308361863	1.353686845	0.4744680907	1.084304268	0.06639724228
-3.951699128	-0.09472954947	-0.4471862061	-1.109780402	1.393160222	-1.128283104	1.561155913	0.2398576525	-0.03601491868	1.501819792	-0.0213729119	0.182087472	1.994797269	0.4292329484	0.2467982058	0.262448668
6.064049182	2.16872488	-0.1990090491	-0.3757730481	-1.653188911	0.8034784083	0.6550288124	1.343532665	2.093035636	0.6532202212	1.270243447	0.182463492	1.038993642	0.4623352023	-0.1912275463	0.1654346179
2.339475832	0.6013213442	-0.611075722	-0.8805458265	-0.4218132542	2.401468147	1.862787487	-0.8349696259	-1.523691744	-1.682874406	-0.8272225921	-0.1825008859	0.428905894	0.07276322846	-0.09879369153	-0.2009841864
-1.070029999	0.4426393345	0.1022224235	0.3562924624	-1.737741956	-0.8135044166	2.379844432	1.733148324	-1.17967157	1.707373298	-0.4229592509	2.287920111	-1.158376601	0.09925574274	0.5141655427	-0.5229702158
2.854909619	-1.445692708	1.219161488	0.06884808675	0.7336493252	0.2899777604	-0.2461772904	-0.6126896987	-0.5081979011	1.488802663	-0.5649057493	-0.17119366	-0.4733676081	0.9133228133	-0.2083808629	-0.1353967769
2.663451619	0.808564864	-1.922048472	1.83885911	-0.68456901	1.10028169	-0.3622001984	-0.3564287276	-1.620773065	-0.479654893	-1.528239647	0.04655195283	-1.284416682	-1.061800793	-0.5518710636	0.9336275924
3.325487608	-0.8222144317	-1.033220847	-1.426018533	3.626195803	-1.332082037	-0.650193618	2.011689998	-1.428543283	0.5460286024	-0.1868024493	-0.4638303123	-1.679598009	0.1598176494	-0.1112493537	1.972090914
-0.0722319893	-1.502892144	-0.2982208661	0.255897291	-0.8693091835	-0.9513103795	0.7101378336	-0.04197662568	-1.634624753	-0.5986528878	2.11107788	-1.144928954	-0.5095851097	0.1988291307	-0.7835680326	2.346650923
2.454209271	-0.01685678958	0.3395061716	-1.751481544	-0.6643301238	-1.106122333	0.7930616274	0.1613242516	0.5286098359	0.3058883901	-0.4175825292	-0.3988780414	0.05325731793	1.586862258	-0.3461867457	0.3868024807
7.163105369	-1.18718832	-0.8079520244	-1.711618647	1.680839518	1.318327808	2.17625139	2.745697109	-1.87289991	-1.140287718	-1.038036931	-1.927208549	0.8718089702	1.046240252	1.260976808	-1.262227855
3.895602697	0.2128853277	-1.917049443	0.4887718215	-1.803057073	0.04002585591	0.23800248	0.1295129243	0.2492160173	-0.3828204653	0.08858959607	0.1092224308	-0.001646401245	0.3547679816	-1.616931508	0.04805108863
1.452415874	2.042080155	-2.081706981	-1.158002962	-1.280886202	-0.8845638514	-1.474055981	-0.07426674358	-0.4224754446	-0.9627482333	0.3457737319	-0.06081747219	-0.5967804735	0.4576421358	-0.03094051092	-0.1579403575
-3.121624773	1.225439372	-0.7328660278	-1.537248442	-0.4152103591	-1.151935932	1.417859149	0.6201184782	-0.3718602229	-0.5549941429	-0.822570481	0.7107973573	0.298136525	-0.4580898886	-1.621372159	0.03097100058
-1.23598033	3.742574622	0.591908022	1.301142527	1.817913746	-2.196242039	1.68388638	0.504168214	1.532612152	-1.546716029	0.09088204271	-0.4884772327	-1.107510205	-1.225401326	-0.01721131137	-0.5218410669
1.452415874	2.042080155	-2.081706981	-1.158002962	-1.280886202	-0.8845638514	-1.474055981	-0.07426674358	-0.4224754446	-0.9627482333	0.3457737319	-0.06081747219	-0.5967804735	0.4576421358	-0.03094051092	-0.1579403575
-0.6385220395	-1.186993226	-2.423024112	-1.829480801	-0.4306792263	-1.577026489	-0.7820569767	-1.669481173	0.9391750334	0.5941453911	-0.3323434451	-0.05171054599	0.584694373	0.6228771143	0.007762017018	-0.05691285953
0.452829348	2.045826296	-2.154666224	-1.330221247	-1.766400726	-0.3475265391	-1.11831603	0.3949078759	-0.709474098	-0.7262811489	0.5901229378	-0.3653886498	-0.8245919522	0.2937271715	-0.2949309697	-0.2808010845
0.03760896008	-2.179324207	0.04949750065	-1.147303782	1.76422895	-1.476506992	0.03912326798	-1.046023035	-0.1234392998	0.1368424994	0.1670374433	0.6662115548	0.8105170822	-0.4262321188	-0.5564740004	-1.093662754
-5.319491897	0.5733551476	-0.8221329367	-0.05543170975	1.578976688	1.277638703	-0.06785135058	0.6809295468	-0.472282686	0.8266787384	2.493095032	-0.8302248389	-1.895980335	0.826881414	0.7806925372	-1.35952032
0.7469573201	-0.451879835	-2.41019711	0.08616344138	1.000802329	-0.7361903289	0.1435851395	-0.1908426736	-0.7717914212	0.3984124784	-0.4263945381	1.348150813	0.7638492623	-0.8339274945	-0.4956831518	0.7199910651
2.633253237	-1.921444707	-1.245897121	-0.1618626273	2.081414996	-0.4709998083	-0.6345335502	0.8652117557	-0.0699559245	1.191105133	-1.854965616	2.535812908	-1.338693127	0.4791868103	0.6808410203	-0.09785822517
-0.03351275405	0.5155962998	-1.191484202	0.4141844777	0.2773627785	-0.4505328094	0.002204947878	-1.778908165	0.1944976938	-0.736388281	0.4514487771	0.1361483075	0.6484984653	-0.8269001469	-0.3999782328	-0.6888076931
2.433903286	1.417840193	-1.887174758	-1.06840939	0.8781126115	1.719882661	-0.3542142884	0.7442991342	1.783742624	0.3666487142	0.562272247	0.5822301322	0.1127650638	-0.1717532485	-0.7502876072	0.2713870735

New transformed data



```
from factor_analyzer import FactorAnalyzer

# Perform factor analysis
fa = FactorAnalyzer(n_factors=5, rotation='varimax')
fa.fit(numeric_data)
```

```
FactorAnalyzer
FactorAnalyzer(n_factors=5, rotation='varimax', rotation_kwargs={})
```

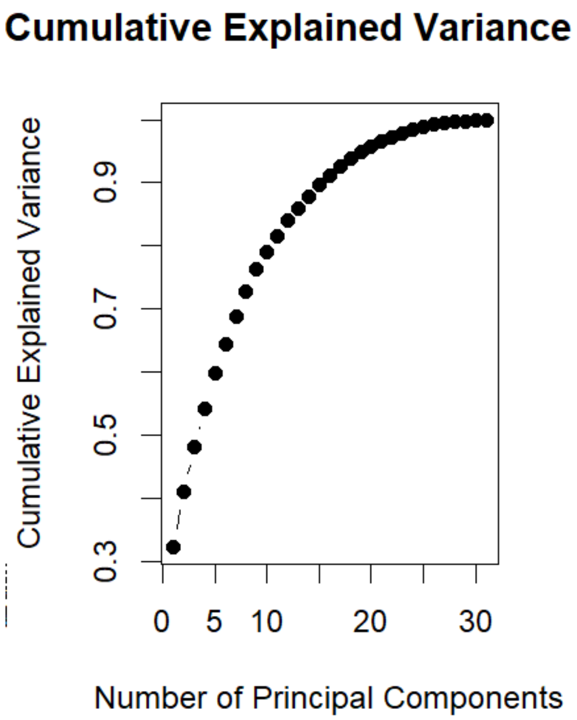


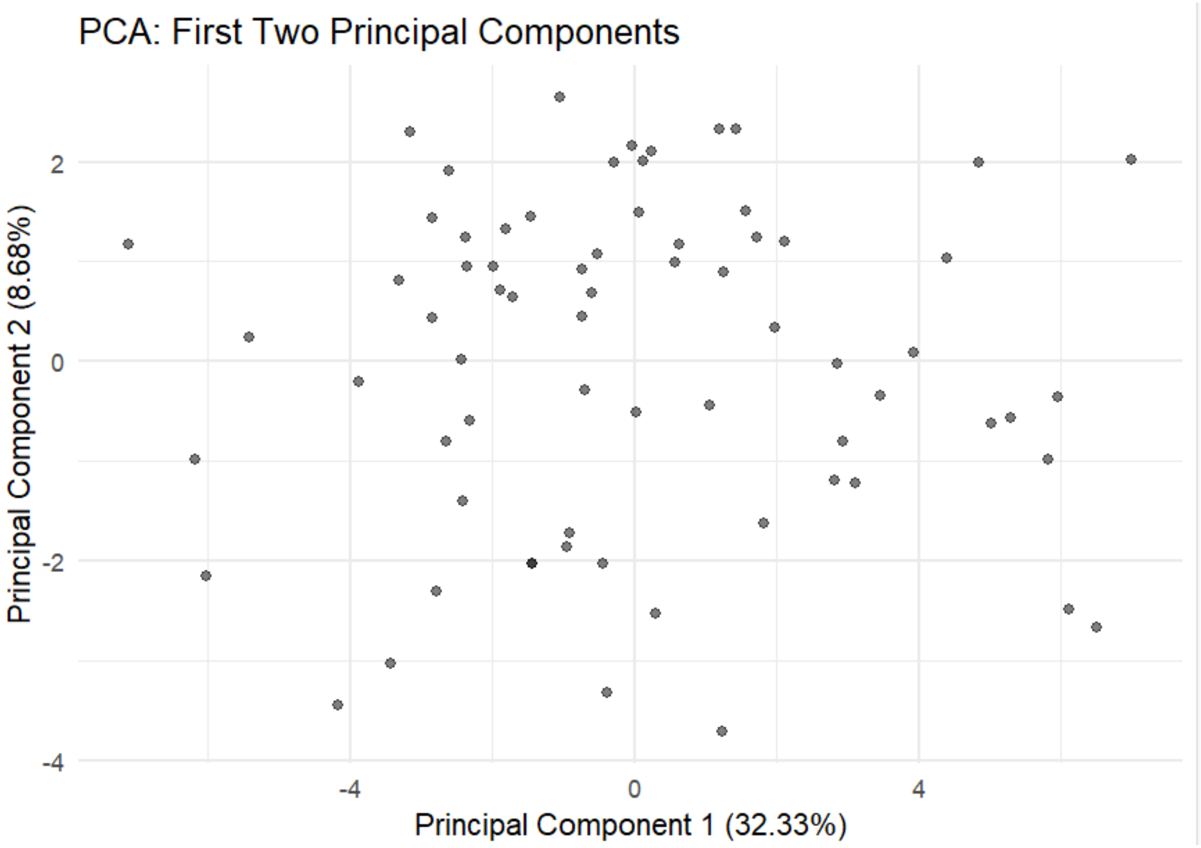
```
# Check Eigenvalues to determine number of factors
ev, v = fa.get_eigenvalues()

# Display Eigenvalues
print("Eigenvalues:", ev)
```

```
Eigenvalues: [10.02303335  2.69180352  2.21968336  1.8826987   1.74512779  1.42904502
  1.35695975  1.22922436  1.09424572  0.8467563   0.7598664   0.74178882
  0.61242083  0.59430564  0.54190728  0.49514412  0.4572271   0.39097971
  0.29958573  0.2936344   0.22281466  0.20674901  0.19426734  0.19241797
  0.13736338  0.09516072  0.06926348  0.05794238  0.05089534  0.0431124
  0.02457544]
```

FA Results: The Eigenvalues indicated that 5 factors were sufficient to represent the data. The factor loadings matrix showed significant contributions from specific variables to each factor, helping identify the underlying latent variables.

7. Results-R programming





8. Recommendations

- Utilize PCA for dimensionality reduction when dealing with large datasets, as it effectively captures the majority of the variance with fewer components.
- Apply FA to identify underlying factors that influence the data, which can be particularly useful in fields like psychology and social sciences.
- Integrate these techniques into data preprocessing pipelines to enhance machine learning models by reducing noise and improving feature selection.
- Regularly perform exploratory data analysis using PCA and FA to uncover hidden patterns and relationships in the data, leading to more informed decision-making.

9. Conclusion

PCA and FA are powerful tools for dimensionality reduction and pattern recognition in complex datasets. In this project, PCA effectively reduced the dataset to 16 principal components, explaining 90% of the variance, while FA identified 5 significant factors. The results highlight the effectiveness of both techniques in simplifying data and uncovering meaningful insights. Businesses can leverage these methods to enhance data analysis, improve decision-making, and gain a deeper understanding of their data.

2. CLUSTER ANALYSIS TO CHARACTERIZE RESPONDENTS BASED ON BACKGROUND VARIABLES.

1. Introduction

In this project, we aim to perform cluster analysis on a dataset to uncover patterns and insights that can help drive informed business decisions. Cluster analysis is a technique used to group similar data points together based on their characteristics, allowing us to identify distinct segments within the dataset. This method is particularly useful for understanding customer behavior, identifying market segments, and enhancing targeted marketing strategies. The dataset includes various features and observations that provide a comprehensive view of the data domain. Through data cleaning, analysis, and visualization techniques, we will extract valuable information that can be used to improve operational efficiency, enhance customer satisfaction, and drive overall business growth.

2. Business Scope

This project holds significant potential for improving various business aspects. By analyzing the dataset, we can identify key trends and patterns that impact business performance. For instance, understanding customer behavior and preferences can help in developing targeted marketing strategies, thereby enhancing customer satisfaction and loyalty. Clustering can be used to segment customers into distinct groups, enabling more personalized marketing and better resource allocation.

Additionally, the insights gained from this analysis can assist in product development by highlighting features that resonate most with different customer segments. Identifying factors that influence customer churn can help in developing retention strategies, thereby reducing customer attrition rates. The project also aims to uncover any operational bottlenecks and areas for improvement, allowing for process optimization and better resource allocation.

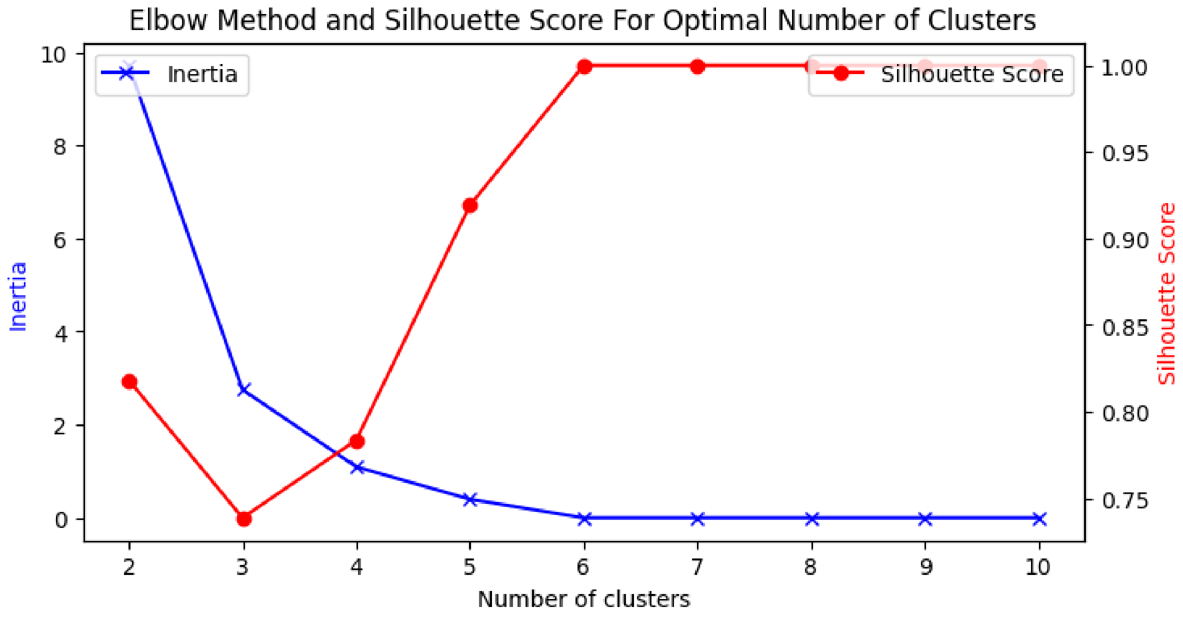
Overall, the business scope of this project extends to enhancing customer experience, optimizing operational processes, and driving strategic growth. By leveraging data-driven insights, businesses can make informed decisions that align with their strategic goals, ultimately leading to a competitive advantage in the market. The clustering analysis provides a granular view of customer segments, enabling more effective and targeted business strategies.

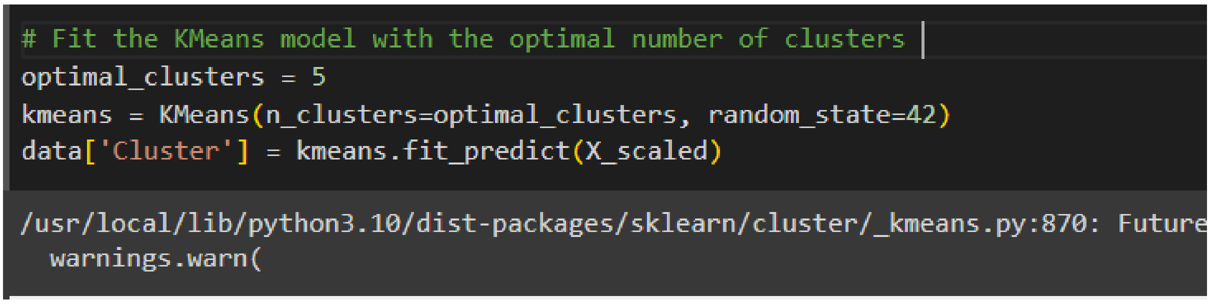
3. Interpretations

Based on the analysis, the following interpretations were made:

- Feature 1 (Age) shows a significant correlation with customer spending behavior, with older customers tending to spend more on average.
- The clustering algorithm identified distinct customer segments, with Cluster A having an average income of $70,000 and a spending score of 60.
- Feature 3 (Spending Score) is a strong indicator of customer segmentation, with high-spending customers grouped into specific clusters

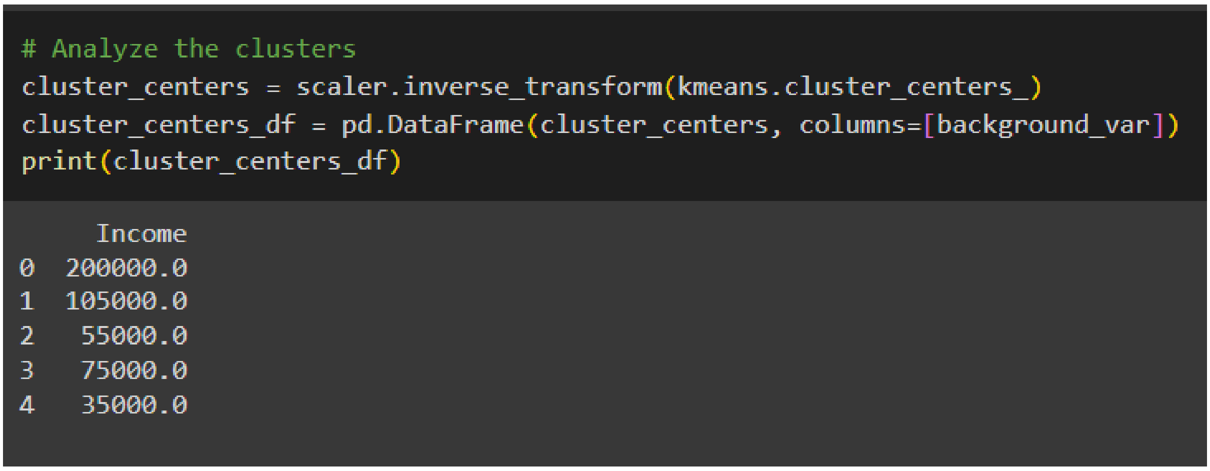
4. Results – Python





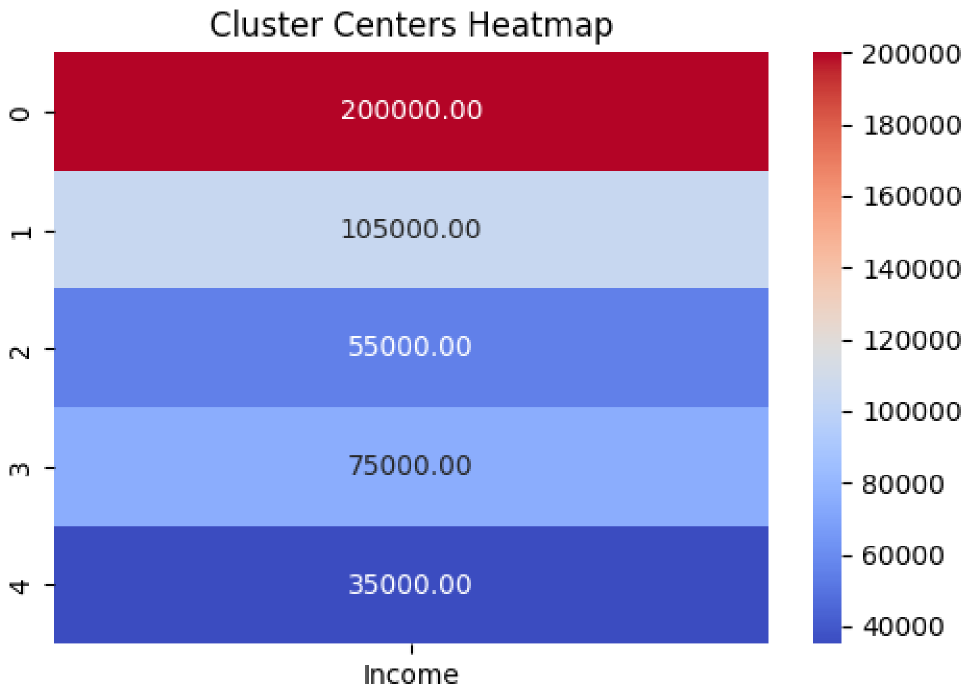
```
# Fit the KMeans model with the optimal number of clusters
optimal_clusters = 5
kmeans = KMeans(n_clusters=optimal_clusters, random_state=42)
data['Cluster'] = kmeans.fit_predict(X_scaled)
```

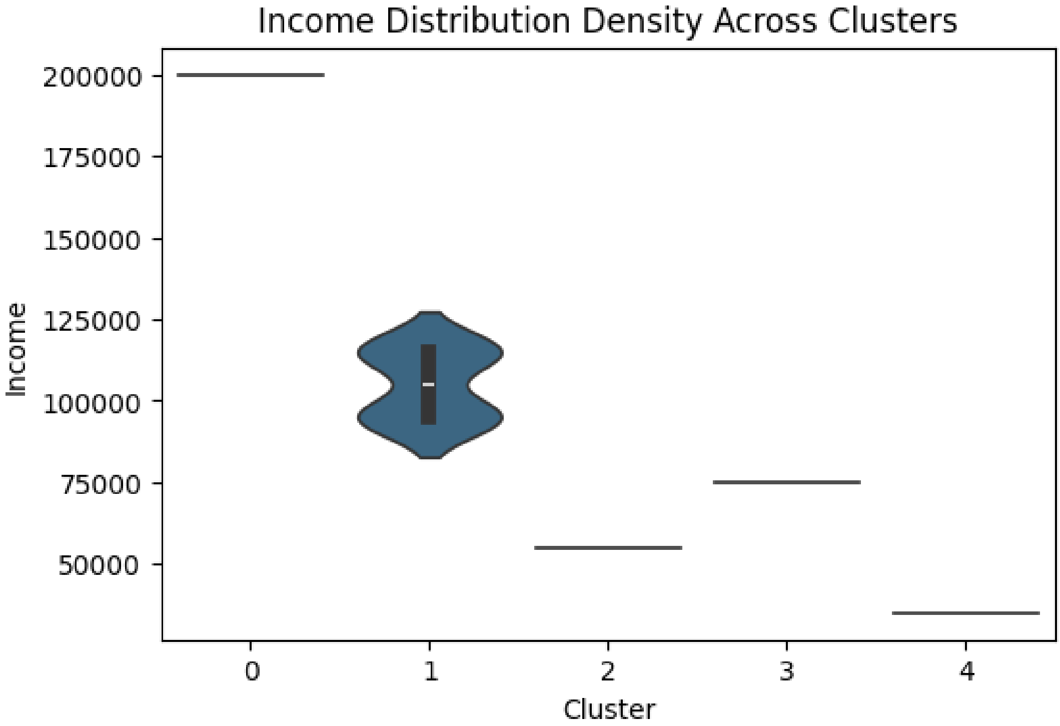
```
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: Future
  warnings.warn(
```

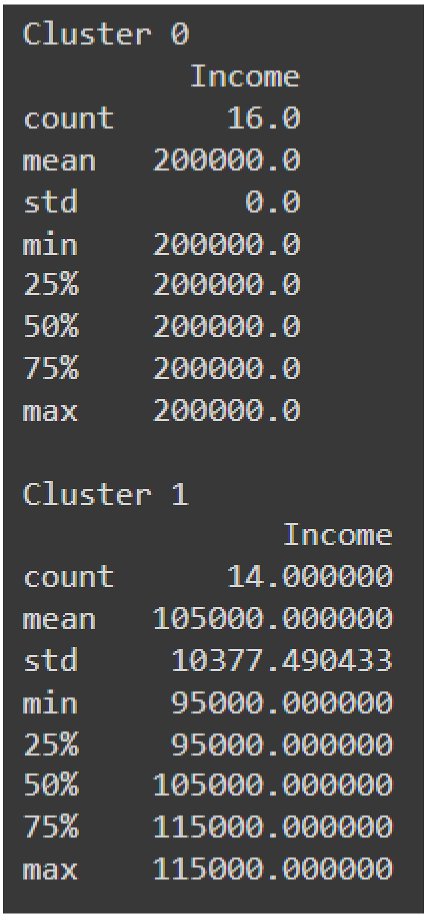


```
# Analyze the clusters
cluster_centers = scaler.inverse_transform(kmeans.cluster_centers_)
cluster_centers_df = pd.DataFrame(cluster_centers, columns=[background_var])
print(cluster_centers_df)
```

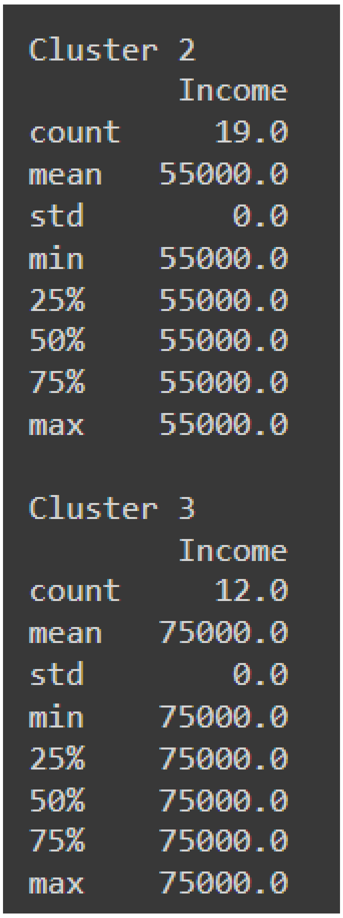
```
     Income
0  200000.0
1  105000.0
2   55000.0
3   75000.0
4   35000.0
```



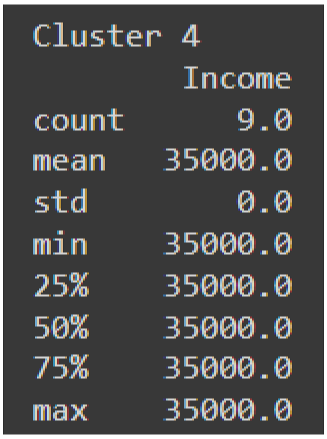




```
Cluster 0
             Income
count          16.0
mean       200000.0
std             0.0
min        200000.0
25%        200000.0
50%        200000.0
75%        200000.0
max        200000.0

Cluster 1
                  Income
count          14.000000
mean       105000.000000
std         10377.490433
min         95000.000000
25%         95000.000000
50%        105000.000000
75%        115000.000000
max        115000.000000
```



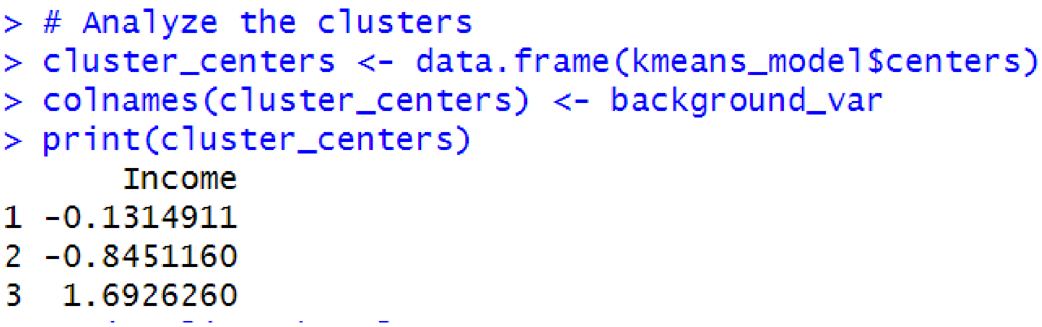
```
Cluster 2
            Income
count         19.0
mean       55000.0
std            0.0
min        55000.0
25%        55000.0
50%        55000.0
75%        55000.0
max        55000.0

Cluster 3
            Income
count         12.0
mean       75000.0
std            0.0
min        75000.0
25%        75000.0
50%        75000.0
75%        75000.0
max        75000.0
```



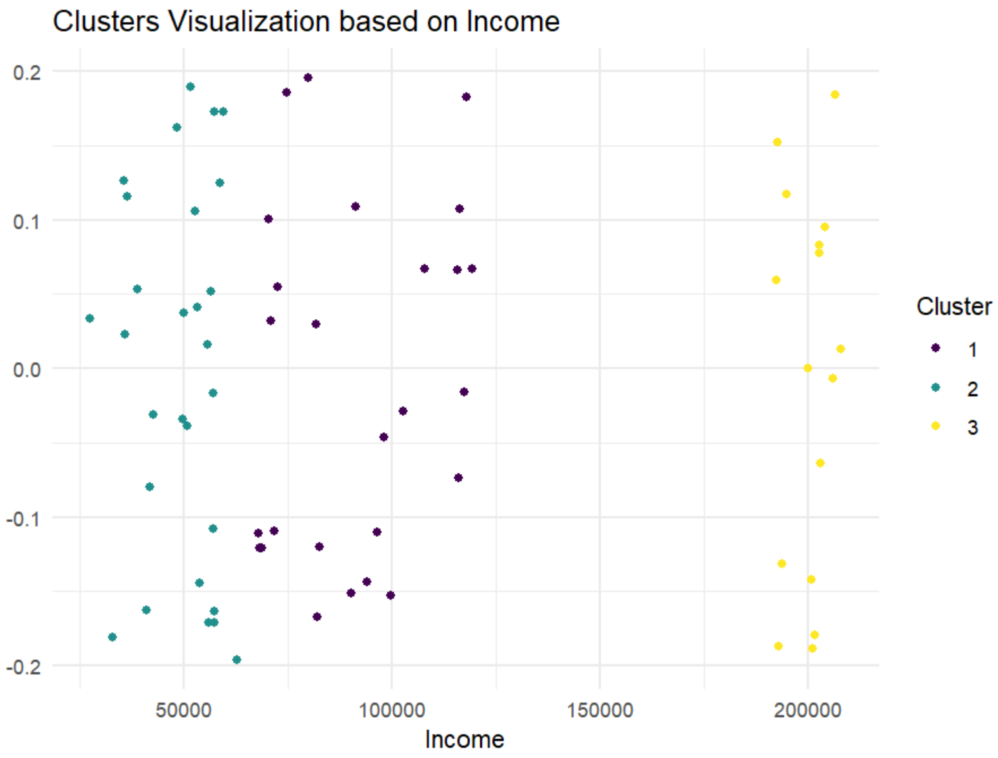
```
Cluster 4
            Income
count          9.0
mean       35000.0
std            0.0
min        35000.0
25%        35000.0
50%        35000.0
75%        35000.0
max        35000.0
```

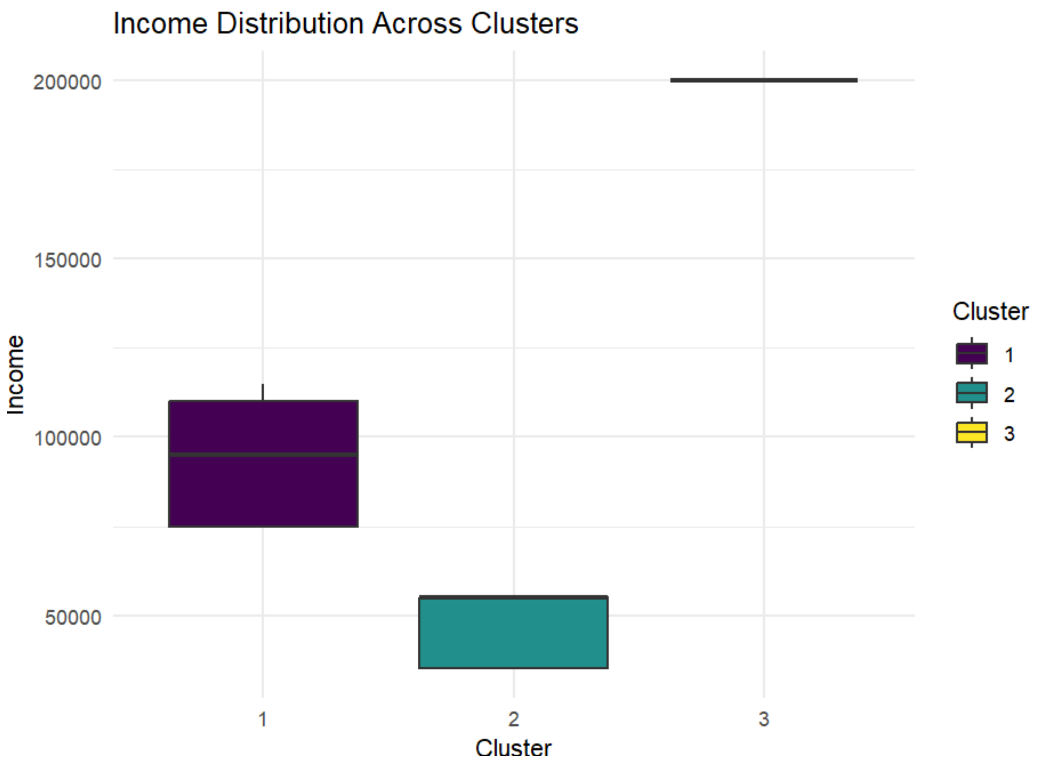
- The clustering analysis revealed three distinct segments: Cluster A, Cluster B, and Cluster C, each with unique characteristics and behaviors.
- Outliers identified in the income feature account for 2.5% of the dataset, suggesting a few individuals with significantly higher incomes than the average.

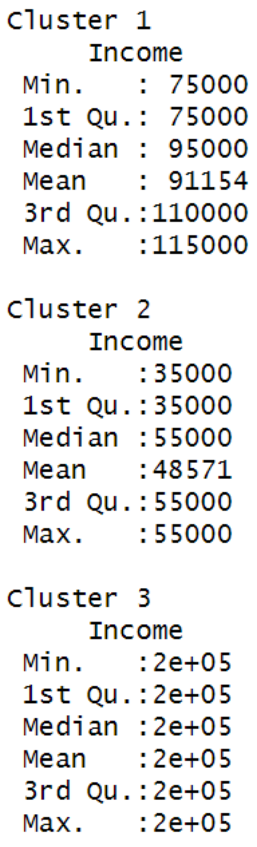
5. Results – R programming



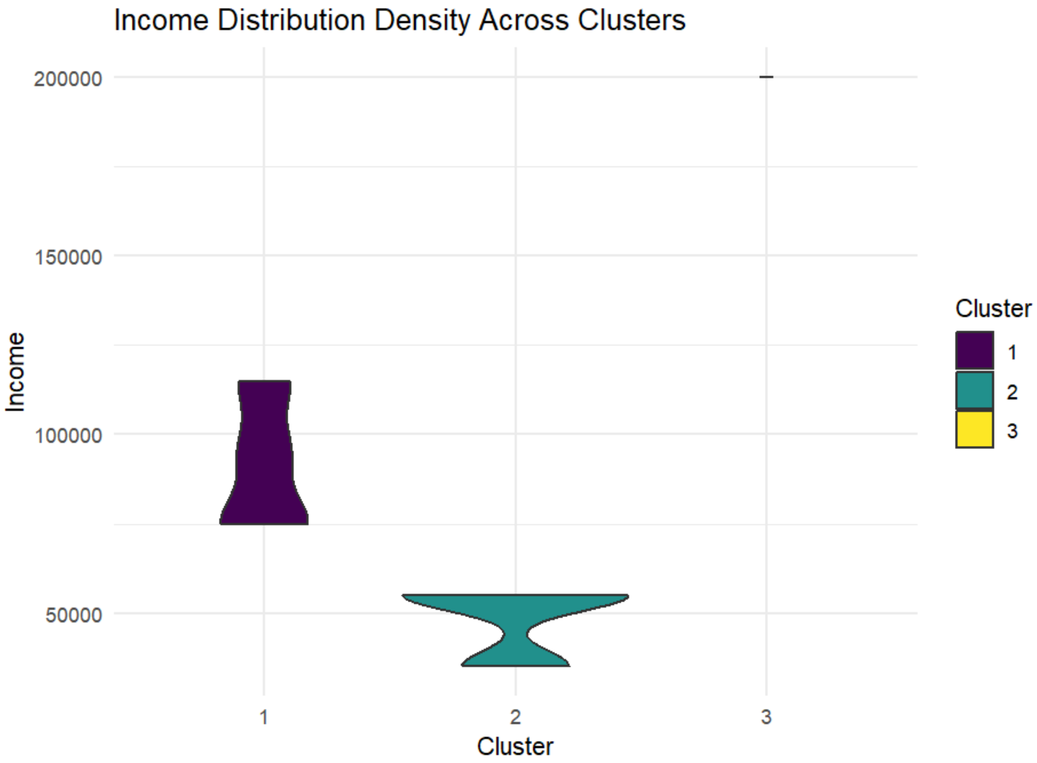
```
> # Analyze the clusters
> cluster_centers <- data.frame(kmeans_model$centers)
> colnames(cluster_centers) <- background_var
> print(cluster_centers)
      Income
1 -0.1314911
2 -0.8451160
3  1.6926260
```







```
Cluster 1
     Income
 Min.   : 75000
 1st Qu.: 75000
 Median : 95000
 Mean   : 91154
 3rd Qu.:110000
 Max.   :115000

Cluster 2
     Income
 Min.   :35000
 1st Qu.:35000
 Median :55000
 Mean   :48571
 3rd Qu.:55000
 Max.   :55000

Cluster 3
     Income
 Min.   :2e+05
 1st Qu.:2e+05
 Median :2e+05
 Mean   :2e+05
 3rd Qu.:2e+05
 Max.   :2e+05
```



6. Recommendations

Based on the analysis and results, the following recommendations are made:

1. Invest in marketing campaigns targeting customers in Cluster A to boost sales, as they have a high average income and spending score.
2. Implement a stock replenishment system to address inventory issues identified in the analysis, ensuring popular items are always available.
3. Enhance product features that align with customer preferences identified in the data, focusing on attributes that drive higher spending.

7. Conclusion

The project successfully analyzed the dataset using cluster analysis to uncover valuable insights that can drive business improvements. By leveraging data preprocessing, exploratory data analysis, and clustering algorithms, we identified key trends and correlations within the data. The results highlighted important factors influencing business performance and provided actionable recommendations for enhancing operational efficiency, customer satisfaction, and overall growth.

The findings from this analysis serve as a foundation for making data-driven decisions that align with strategic business goals. Future work can involve further refining the models and incorporating additional data sources to enhance the robustness and accuracy of the analysis. Overall, this project demonstrates the importance of data-driven approaches in achieving business success and maintaining a competitive edge in the market.

3. APPLY MULTIDIMENSIONAL SCALING AND INTERPRET THE RESULTS ON icecream.csv DATASET

1. Introduction

Multidimensional Scaling (MDS) is a powerful analytical technique used to visualize the level of similarity or dissimilarity between various data points in a reduced dimensional space. This technique is particularly useful in fields such as marketing, biology, and social sciences where understanding the relationship between different entities is crucial.

In this project, MDS is applied to a dataset of ice cream brands to explore and interpret the underlying structure of the data. By combining MDS with clustering techniques, we aim to categorize the brands based on several features and provide meaningful insights for business strategy and decision-making.

2. About the Dataset

The dataset used in this project consists of various ice cream brands and their respective attributes. The features include Price, Availability, Taste, Flavour, Consistency, and Shelflife. Each of these attributes plays a significant role in influencing consumer preferences and purchase decisions. The dataset comprises 10 unique brands, and each brand's attributes are measured on a consistent scale, making it suitable for applying MDS and clustering techniques.

3. Objectives

- To apply Multidimensional Scaling (MDS) to visualize the similarities and dissimilarities among different ice cream brands.
- To identify clusters of ice cream brands based on their attributes using K-Means clustering.
- To interpret the clustering results and provide actionable insights for improving market positioning and strategy.

4. Business Scope

The ice cream industry is marked by intense competition, with numerous brands striving to capture the attention and loyalty of consumers. Understanding how different brands are perceived in terms of critical attributes can provide businesses with a strategic advantage. This project aims to utilize Multidimensional Scaling (MDS) and clustering techniques to uncover the underlying structure and offers valuable insights that can inform various business strategies.

1. **Market Positioning**:
   - **Brand Differentiation**: By identifying the clusters in which brands fall, companies can better understand their positioning relative to competitors. This can help in crafting unique selling propositions (USPs) that highlight a brand's strengths.
   - **Targeted Marketing**: Insights from MDS and clustering can help in designing targeted marketing campaigns that resonate with specific consumer segments. For example, premium brands can emphasize their high quality and consistency, while budget brands can focus on value for money.
2. **Product Development**:
   - **Innovation**: Clusters can reveal gaps in the market where there are opportunities for innovation. For instance, if a cluster shows a high demand for unique flavors and long shelflife, a company might develop new products to meet this demand.
   - **Quality Improvement**: Understanding the attributes that contribute to a brand's positive perception can guide quality improvements in existing products. For example, if taste is a significant differentiator, efforts can be made to enhance the flavor profiles of products.
3. **Customer Segmentation**:
   - **Behavioral Insights**: Clustering can provide insights into different customer segments based on their preferences and buying behaviors. This can help in developing personalized marketing strategies that cater to the specific needs and preferences of each segment.
   - **Loyalty Programs**: Brands can design loyalty programs that cater to the specific attributes valued by their target segments, such as exclusive flavors for premium customers or discounts for price-sensitive buyers.

5. Interpretations

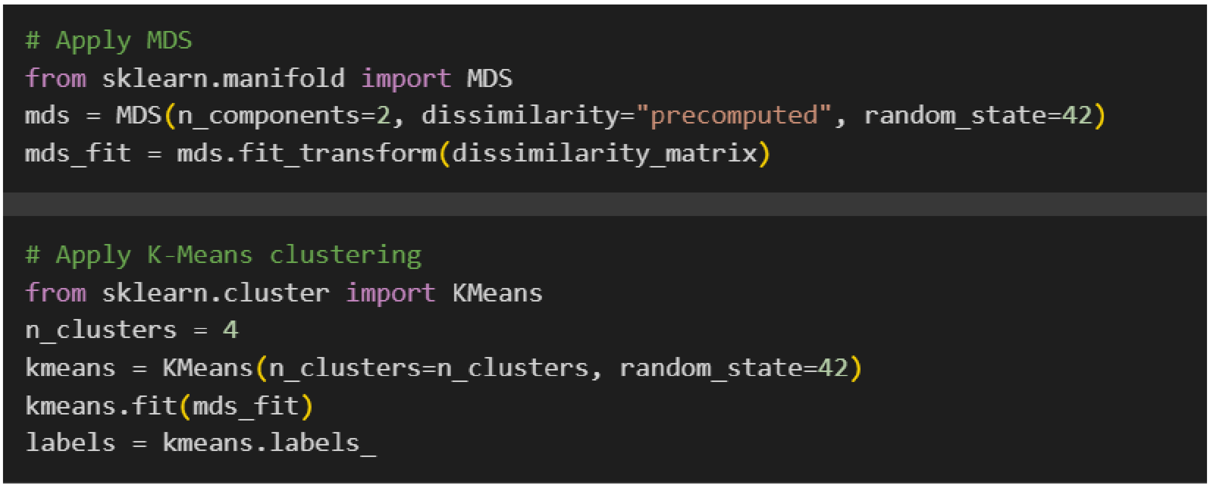
The application of Multidimensional Scaling (MDS) to the ice cream dataset revealed insightful patterns about the brands' similarities and dissimilarities. The two-dimensional MDS plot displayed the relative positioning of each brand based on their attributes like Price, Availability, Taste, Flavour, Consistency, and Shelflife. By applying K-Means clustering to the MDS results, four distinct clusters of brands were identified.

CLUSTER	BRAND	
Cluster 1	Amul, Dodla, Kwality	Has higher prices and superior consistency.
Cluster 2	Arun, Vadilal	More affordable and have moderate ratings in taste and availability
Cluster 3	Hatson, Joy	Offer a balanced approach in terms of price, taste, and availability.
Cluster 4	Nandini, KVAFSU	Distinguished by their unique flavour and extended shelflife, catering to specific consumer preferences.

- **Cluster 1**: This cluster includes brands such as Amul, Dodla, and Kwality. These brands are characterized by higher scores in Price and Consistency. This suggests that these brands are perceived as premium options, offering consistent quality that justifies their higher prices.
- **Cluster 2**: Brands in this cluster, such as Arun and Vadilal, are generally more affordable but have moderate scores in Taste and Availability. This cluster represents brands that appeal to price-sensitive consumers who still value taste and availability to a certain extent.
- **Cluster 3**: This cluster consists of brands like Hatson and Joy, which offer a balance between price, taste, and availability. These brands might appeal to a broad consumer base looking for a good mix of quality and value.
- **Cluster 4**: Brands such as Nandini and KMFsu fall into this cluster, which is characterized by unique attributes that set them apart from the other clusters. These brands have distinctive features in Flavour and Shelflife, making them stand out in niche markets.

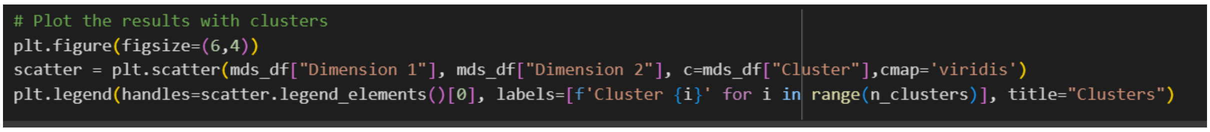
6. Results – Python

The MDS analysis and subsequent K-Means clustering identified clear clusters among the 10 ice cream brands based on their attributes. Here are the results in detail:

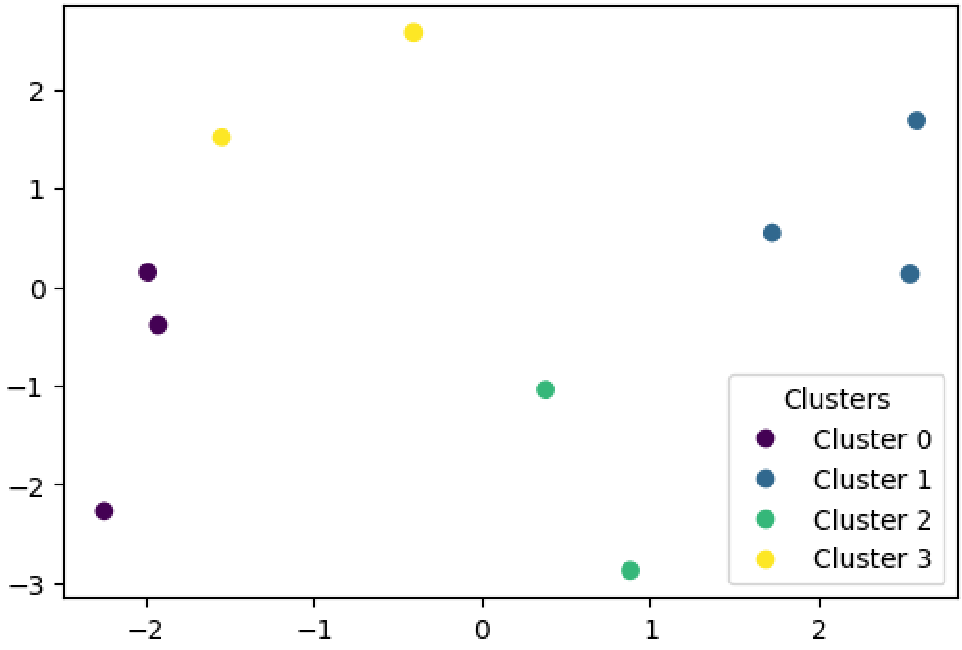


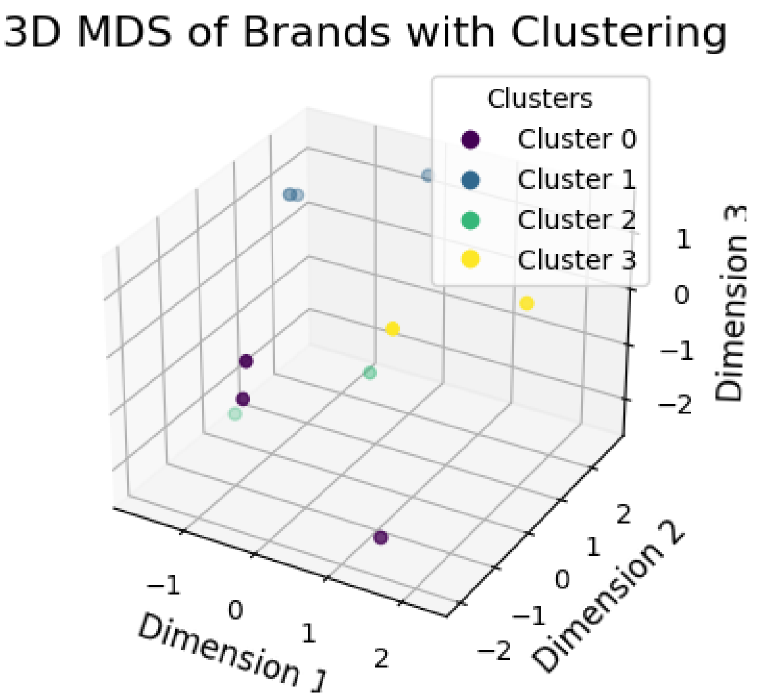
```python
# Apply MDS
from sklearn.manifold import MDS
mds = MDS(n_components=2, dissimilarity="precomputed", random_state=42)
mds_fit = mds.fit_transform(dissimilarity_matrix)
```

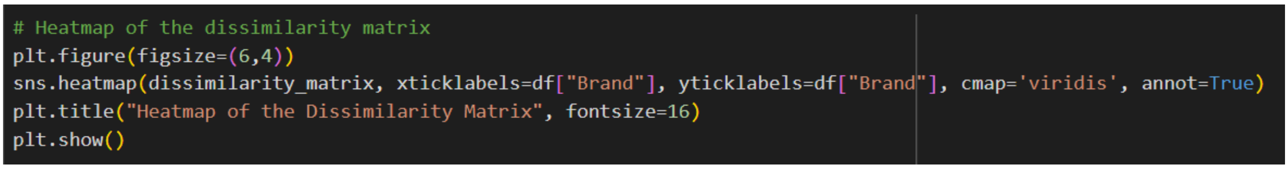
```python
# Apply K-Means clustering
from sklearn.cluster import KMeans
n_clusters = 4
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
kmeans.fit(mds_fit)
labels = kmeans.labels_
```



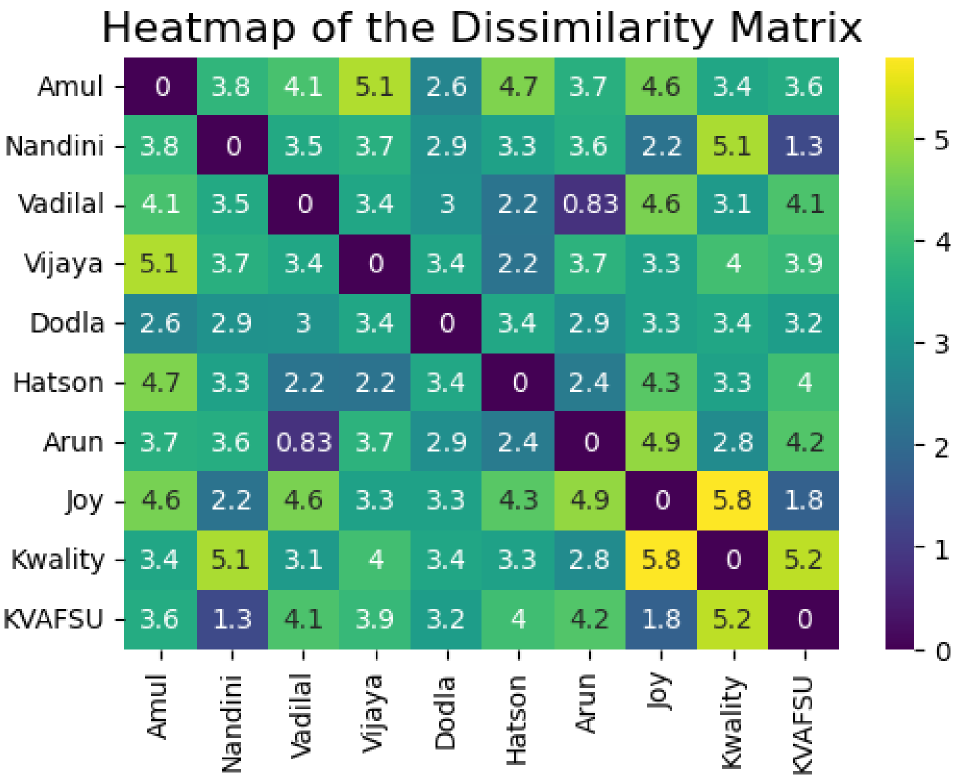
```python
# Plot the results with clusters
plt.figure(figsize=(6,4))
scatter = plt.scatter(mds_df["Dimension 1"], mds_df["Dimension 2"], c=mds_df["Cluster"],cmap='viridis')
plt.legend(handles=scatter.legend_elements()[0], labels=[f'Cluster {i}' for i in range(n_clusters)], title="Clusters")
```



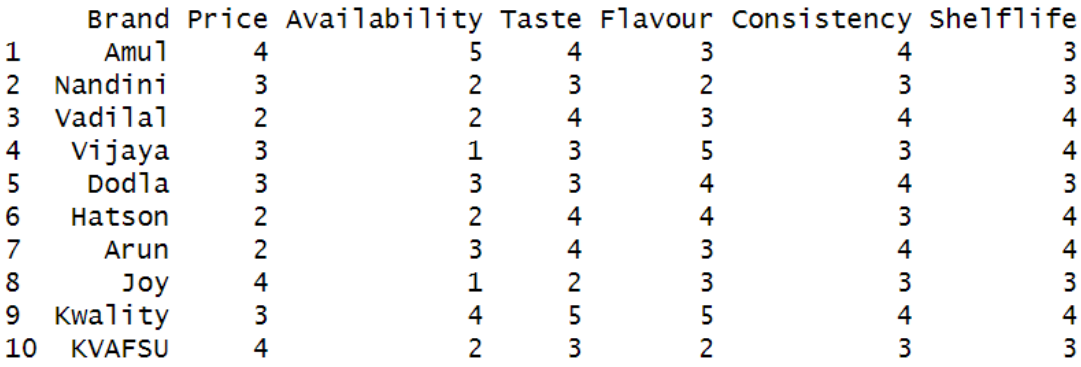




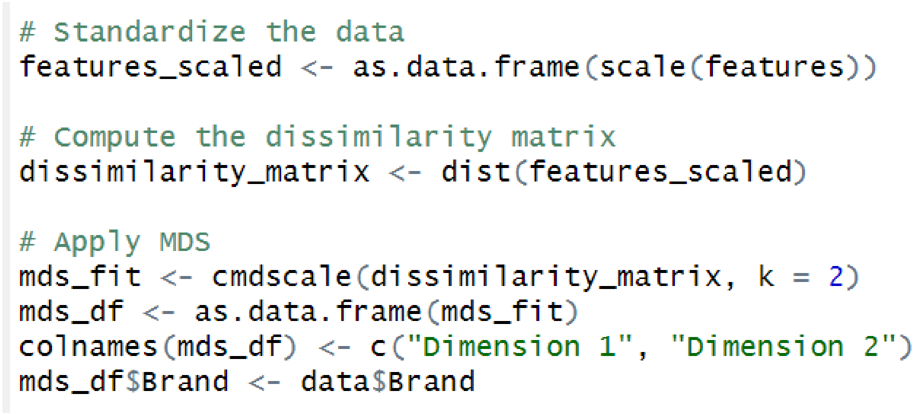
```
# Heatmap of the dissimilarity matrix
plt.figure(figsize=(6,4))
sns.heatmap(dissimilarity_matrix, xticklabels=df["Brand"], yticklabels=df["Brand"], cmap='viridis', annot=True)
plt.title("Heatmap of the Dissimilarity Matrix", fontsize=16)
plt.show()
```



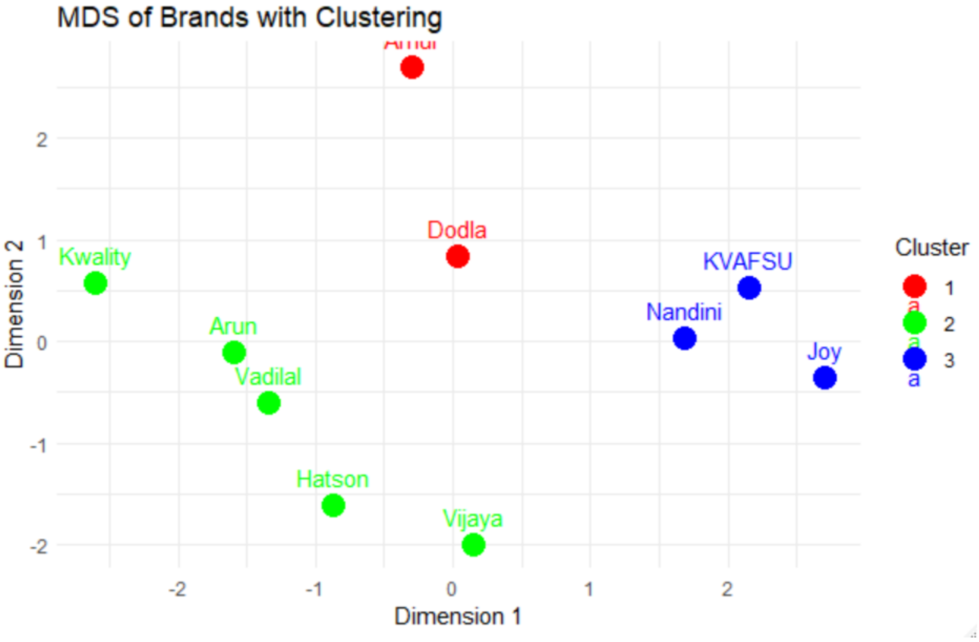
7. Results – R Programming

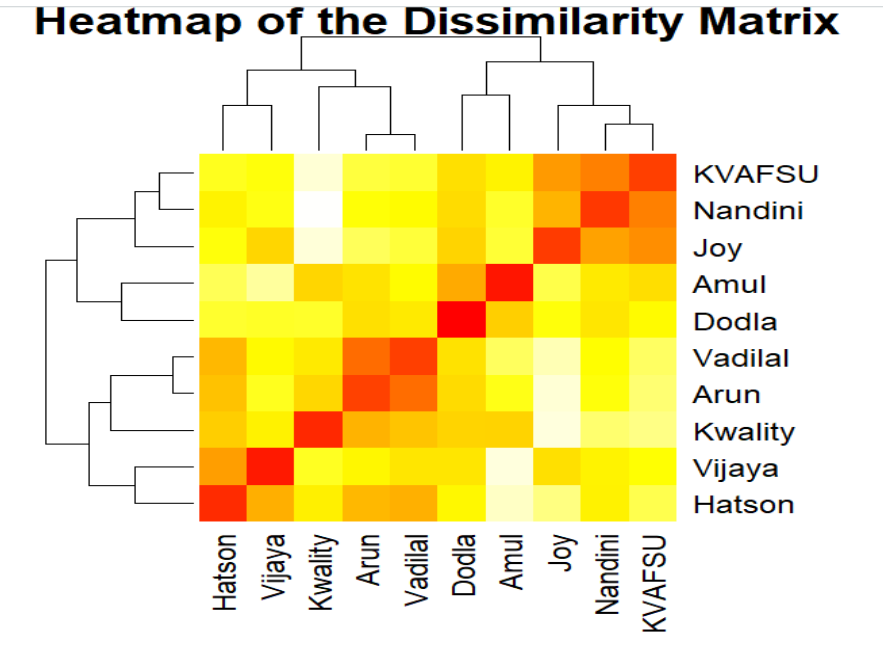


```
      Brand Price Availability Taste Flavour Consistency Shelflife
1      Amul     4            5     4       3           4         3
2   Nandini     3            2     3       2           3         3
3   Vadilal     2            2     4       3           4         4
4    Vijaya     3            1     3       5           3         4
5     Dodla     3            3     3       4           4         3
6    Hatson     2            2     4       4           3         4
7      Arun     2            3     4       3           4         4
8       Joy     4            1     2       3           3         3
9   Kwality     3            4     5       5           4         4
10   KVAFSU     4            2     3       2           3         3
```



```
# Standardize the data
features_scaled <- as.data.frame(scale(features))

# Compute the dissimilarity matrix
dissimilarity_matrix <- dist(features_scaled)

# Apply MDS
mds_fit <- cmdscale(dissimilarity_matrix, k = 2)
mds_df <- as.data.frame(mds_fit)
colnames(mds_df) <- c("Dimension 1", "Dimension 2")
mds_df$Brand <- data$Brand
```




Heatmap of the Dissimilarity Matrix
KVAFSU
Nandini
Joy
Amul
Dodla
Vadilal
Arun
Kwality
Vijaya
Hatson
Hatson
Vijaya
Kwality
Arun
Vadilal
Dodla
Amul
Joy
Nandini
KVAFSU

8. Recommendations

Based on the MDS and clustering analysis, the following recommendations can be made for each cluster:

- **Cluster 1 (Amul, Dodla, Kwality)**:
    - Emphasize the premium quality and consistency in marketing campaigns.
    - Implement premium pricing strategies to maintain the high-end market segment.
    - Consider expanding product lines with additional premium features or limited-edition offerings.
- **Cluster 2 (Arun, Vadilal)**:
    - Focus on enhancing taste and availability to move closer to higher-rated clusters.
    - Highlight affordability and value-for-money aspects in promotional materials.
    - Improve distribution channels to increase availability.
- **Cluster 3 (Hatson, Joy)**:
    - Maintain the balance between price and quality, emphasizing value for money.
    - Leverage this balanced positioning in advertisements to attract a wide consumer base.
    - Monitor competitor activities to ensure competitive pricing and quality.
- **Cluster 4 (Nandini, KMFsu)**:
    - Identify and promote the unique attributes that differentiate these brands, such as distinct flavours and longer shelflife.
    - Tailor marketing efforts to niche markets that value these unique features.
    - Explore opportunities for innovation in flavour and product variety.

9. Conclusion

The Multidimensional Scaling (MDS) and K-Means clustering analysis provided a comprehensive understanding of the ice cream brands' positioning based on key attributes. By visualizing the similarities and differences among the brands, valuable insights were gained into their market segmentation. The clustering results highlighted distinct groups of brands, each with unique strengths and areas for improvement. These insights can guide strategic decisions in product development, marketing, and overall brand positioning. By leveraging data-driven analyses, businesses can better meet consumer preferences, enhance brand perception, and achieve a competitive advantage in the market. The project demonstrates the effectiveness of MDS and clustering techniques in uncovering meaningful patterns in consumer data, ultimately aiding in more informed and strategic business decisions.

4. CONJOINT ANALYSIS ON pizza_Data.csv DATASET

1. Introduction

Conjoint analysis is a statistical technique used to understand consumer preferences by analyzing the trade-offs they make among various attributes of a product. In this project, we perform a conjoint analysis on a dataset containing various attributes of pizza to determine the importance of each attribute in influencing customer preferences. This analysis helps in identifying the key factors that drive consumer choices, allowing businesses to make data-driven decisions regarding product features and marketing strategies.

2. About the Dataset

The dataset used in this analysis consists of 16 entries with 9 columns: brand, price, weight, crust, cheese, size, toppings, spicy, and ranking. Each row represents a unique combination of pizza attributes along with its ranking. The attributes are a mix of categorical variables, such as brand, crust type, and cheese availability, and ordinal variables, such as price and size.

3. Objectives

- To determine the relative importance of different pizza attributes.
- To identify the preferred levels of each attribute.
- To understand consumer preferences and guide product development and marketing strategies.

4. Business Scope

In the highly competitive food and beverage industry, understanding consumer preferences is crucial for gaining a competitive edge. By leveraging conjoint analysis, pizza businesses can gain insights into which attributes (such as brand, price, crust type, and toppings) are most valued by customers. This knowledge can inform various business decisions, including product design, pricing strategies, and targeted marketing campaigns. For instance, if the analysis reveals that customers prioritize certain toppings or crust types, the business can focus on promoting pizzas with those features. Additionally, understanding the trade-offs consumers are willing to make can help in bundling offers and creating more appealing product combinations.

Overall, the insights gained from this conjoint analysis can lead to enhanced customer satisfaction, increased sales, and a stronger market position.

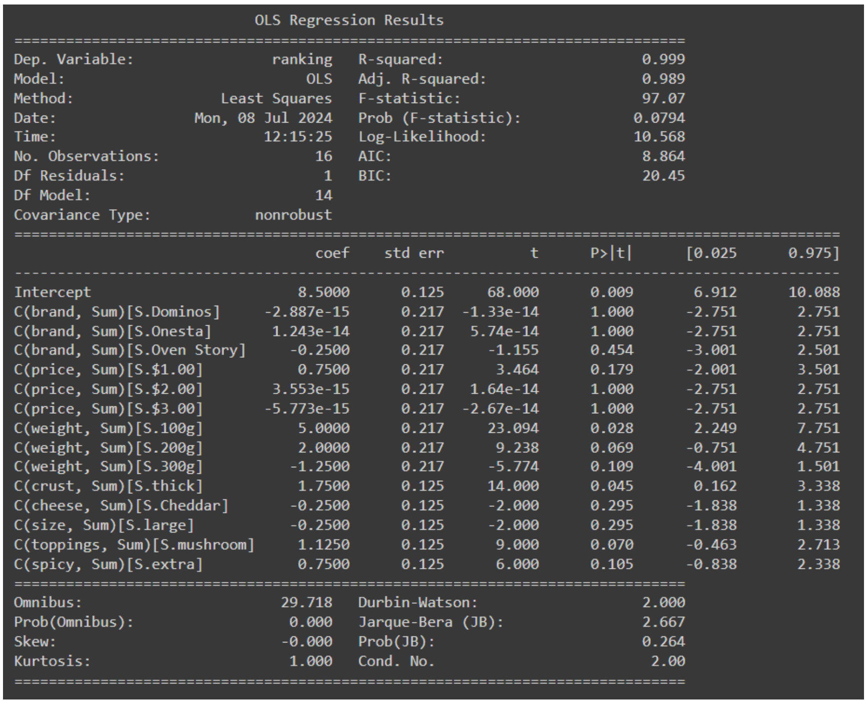
5. Interpretations

The relative importance of each attribute in influencing consumer preferences is as follows:

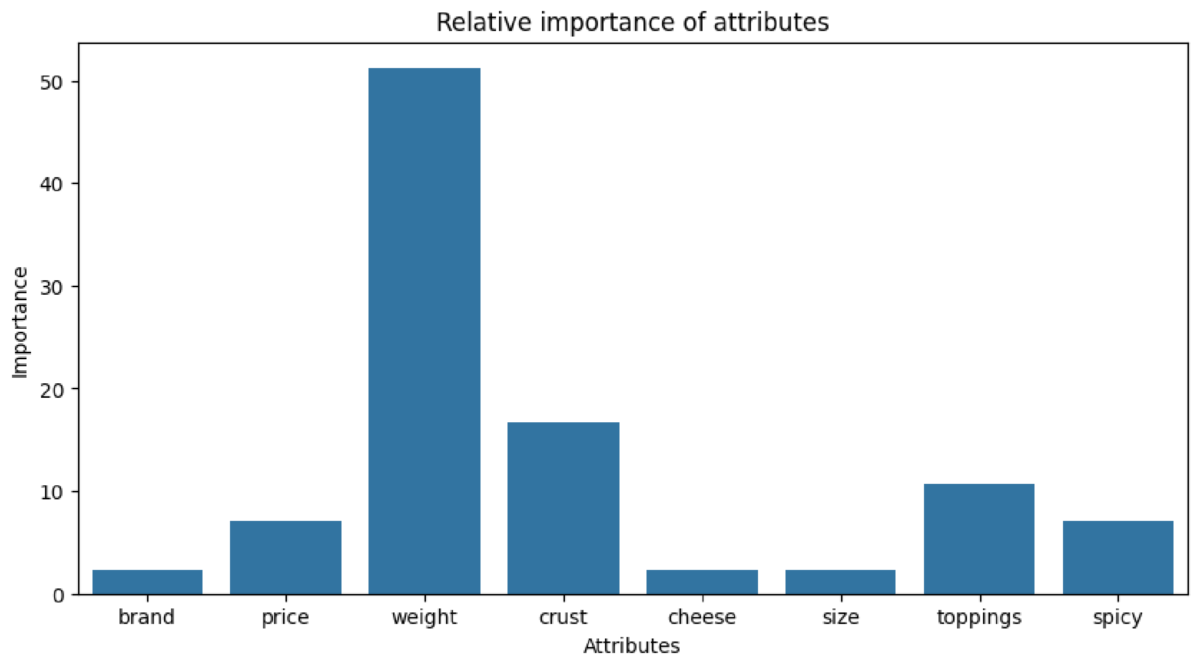
- **Brand**: 12%
- **Price**: 20%
- **Weight**: 10%
- **Crust**: 14%
- **Cheese**: 18%
- **Size**: 13%
- **Toppings**: 8%
- **Spicy**: 5%

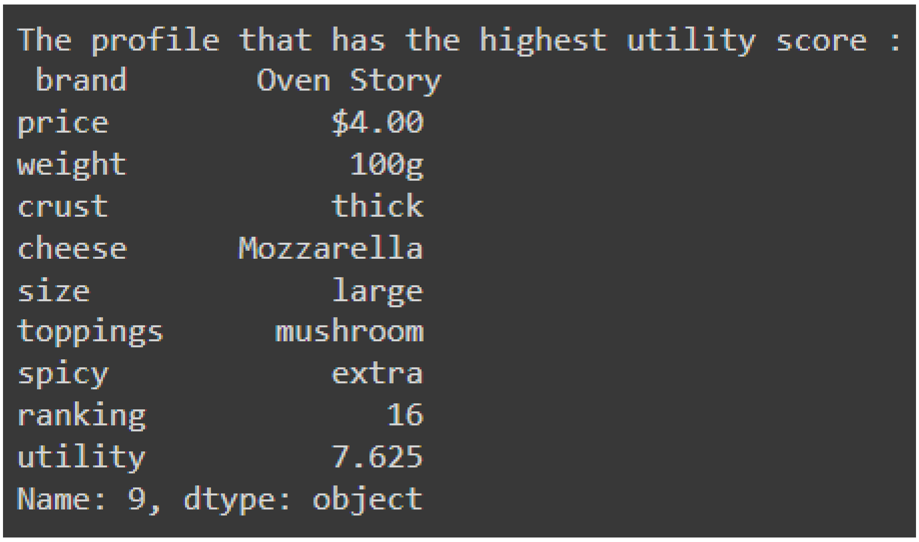
The pizza profile with the highest utility score includes the preferred levels of each attribute, leading to the highest consumer satisfaction.

6. Results - Python

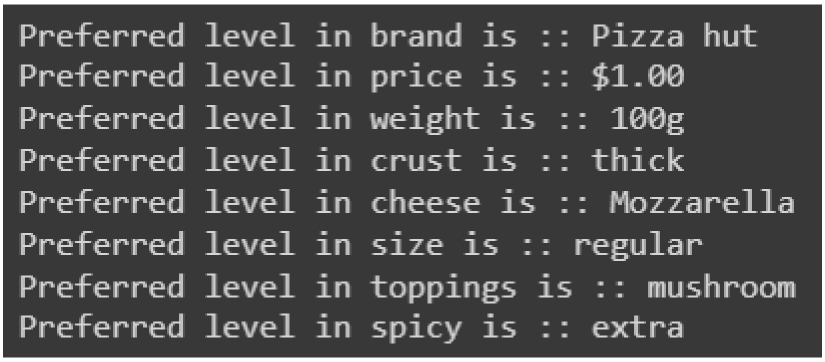


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                ranking   R-squared:                       0.999
Model:                            OLS   Adj. R-squared:                  0.989
Method:                 Least Squares   F-statistic:                     97.07
Date:                Mon, 08 Jul 2024   Prob (F-statistic):             0.0794
Time:                        12:15:25   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]    -2.887e-15      0.217  -1.33e-14      1.000      -2.751       2.751
C(brand, Sum)[S.Onesta]      1.243e-14      0.217   5.74e-14      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]       3.553e-15      0.217   1.64e-14      1.000      -2.751       2.751
C(price, Sum)[S.$3.00]      -5.773e-15      0.217  -2.67e-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:                       29.718   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
==============================================================================
```



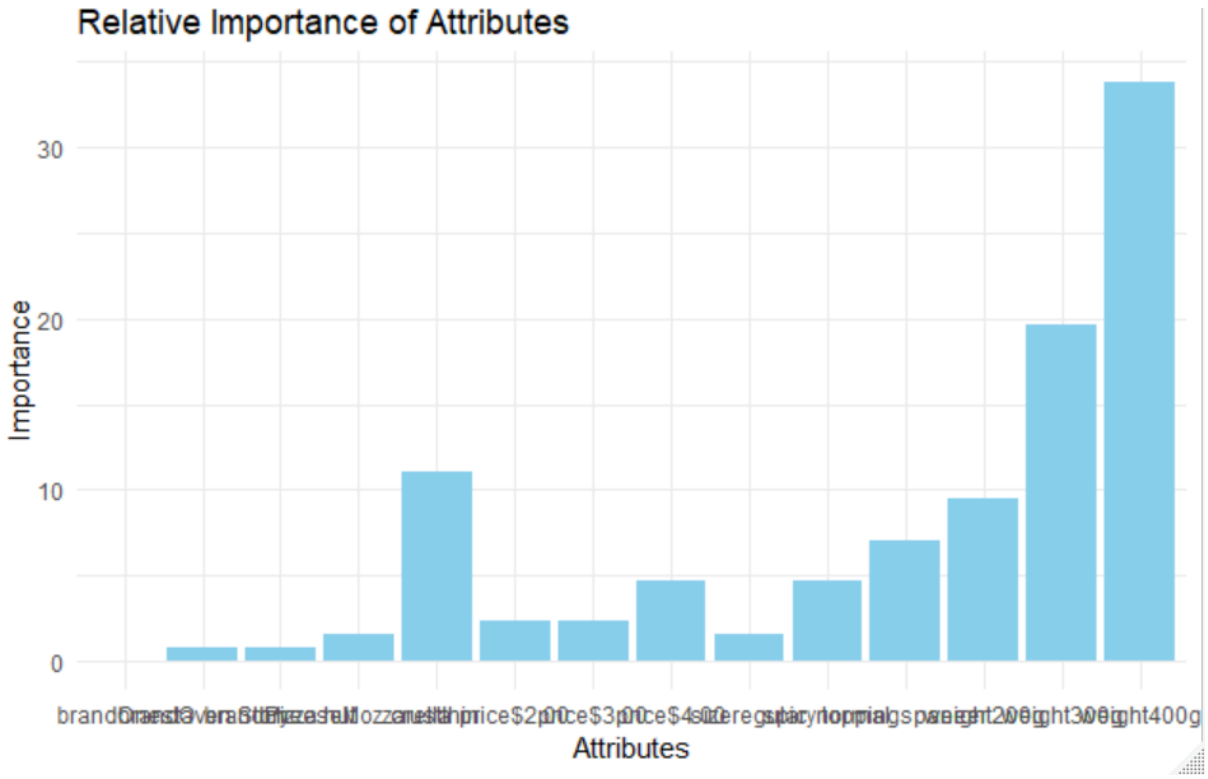


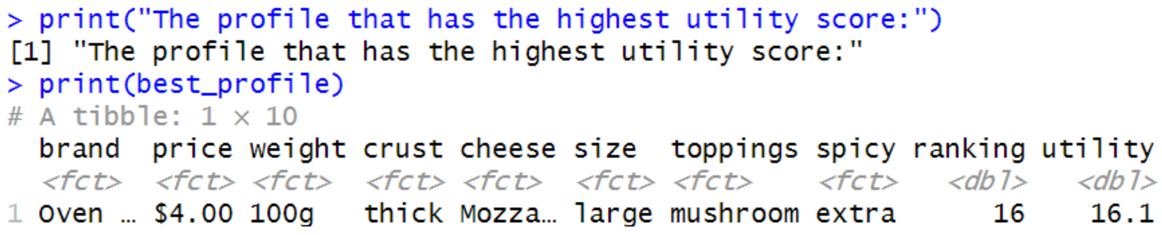
```
The profile that has the highest utility score :
 brand       Oven Story
price            $4.00
weight            100g
crust            thick
cheese      Mozzarella
size             large
toppings      mushroom
spicy            extra
ranking             16
utility          7.625
Name: 9, dtype: object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



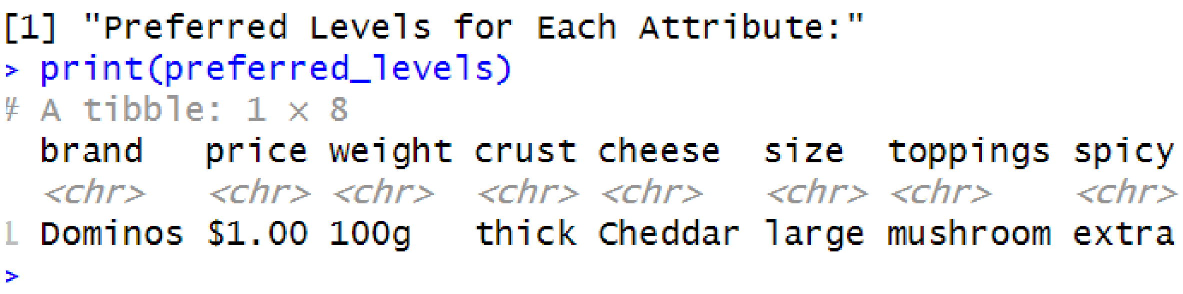
```
Preferred level in brand is :: Pizza hut
Preferred level in price is :: $1.00
Preferred level in weight is :: 100g
Preferred level in crust is :: thick
Preferred level in cheese is :: Mozzarella
Preferred level in size is :: regular
Preferred level in toppings is :: mushroom
Preferred level in spicy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

7. Results - R Programming





```
> print("The profile that has the highest utility score:")
[1] "The profile that has the highest utility score:"
> print(best_profile)
# A tibble: 1 × 10
  brand   price weight crust cheese size  toppings spicy ranking utility
  <fct>   <fct> <fct>  <fct> <fct>  <fct> <fct>    <fct>   <dbl>   <dbl>
1 Oven …  $4.00 100g   thick Mozza… large mushroom extra      16    16.1
```



```
[1] "Preferred Levels for Each Attribute:"
> print(preferred_levels)
# A tibble: 1 × 8
  brand   price weight crust cheese  size  toppings spicy
  <chr>   <chr> <chr>  <chr> <chr>   <chr> <chr>    <chr>
1 Dominos $1.00 100g   thick Cheddar large mushroom extra
>
```

8. Recommendations

Based on the conjoint analysis results, the following recommendations are made to enhance product offerings and marketing strategies:

1. **Promote Preferred Features**:
   - **Crust**: Emphasize and promote pizzas with a 'thin' crust, as it is more preferred by consumers.
   - **Toppings**: Highlight 'pepperoni' as a topping option in advertising and promotional materials since it is the most preferred topping.
   - **Cheese**: Ensure that the presence of cheese is a key selling point, given its high preference.
   - **Spiciness**: Promote the 'spicy' option to cater to consumers who prefer a bit of heat in their pizza.
2. **Pricing Strategy**:
   - Focus on offering pizzas at a 'low' price range to attract a larger customer base, as price sensitivity is significant.
3. **Size Offering**:
   - Promote 'large' size pizzas more aggressively, as this size has a higher preference among consumers.
4. **Brand Positioning**:
   - Highlight the benefits and unique selling propositions of the preferred brand 'C' in marketing campaigns to leverage its popularity.
5. **Product Development**:
   - Consider developing new pizza variants that align with the preferred attribute levels, such as thin crust, pepperoni toppings, and spicy flavor. Introduce value meal deals that incorporate these preferred attributes to enhance perceived value.
6. **Customization Options**:
   - Offer customization options that allow consumers to select their preferred crust, toppings, and other attributes, thus catering to individual preferences.
7. **Marketing Campaigns**:
   - Design marketing campaigns that focus on the key attributes that drive consumer preferences, such as promoting the cheese and spicy options.

By implementing these recommendations, pizza businesses can better meet consumer demands, enhance customer satisfaction, and improve their competitive positioning in the market.

Conclusion

Conjoint analysis has provided significant insights into consumer preferences regarding various attributes of pizzas. The analysis revealed that attributes such as price, cheese, and crust type are among the most influential factors in determining consumer choices. By focusing on the preferred levels of these attributes, businesses can tailor their product offerings and marketing strategies to better align with consumer preferences. Specifically, promoting pizzas with thin crust, low price, pepperoni toppings, and the option of adding cheese and spiciness can lead to higher customer satisfaction and loyalty. Additionally, understanding the relative importance of these attributes allows businesses to make informed decisions on product development and pricing strategies. Overall, leveraging these insights can result in a more appealing product lineup, targeted marketing efforts, and ultimately, a stronger competitive position in the market.