Market Segmentation – A Case Study

Harsh Sinha- 21BLC1419
Ayan panda 22BCY10173
Divya Agarwal 21BAI10159
Daksh Mehta 21BCT0277

STEP 1: CHECKLIST

The provided checklist outlines the initial steps and considerations for deciding whether to proceed with market segmentation. It includes a series of tasks and questions that serve as knock-out criteria. The summary of the checklist is as follows:

The checklist begins by assessing the organization's culture and willingness to change. If the organization is market-oriented and open to new ideas, the process can proceed. Long-term perspective, good communication across units, and the ability to make structural changes are also important factors.

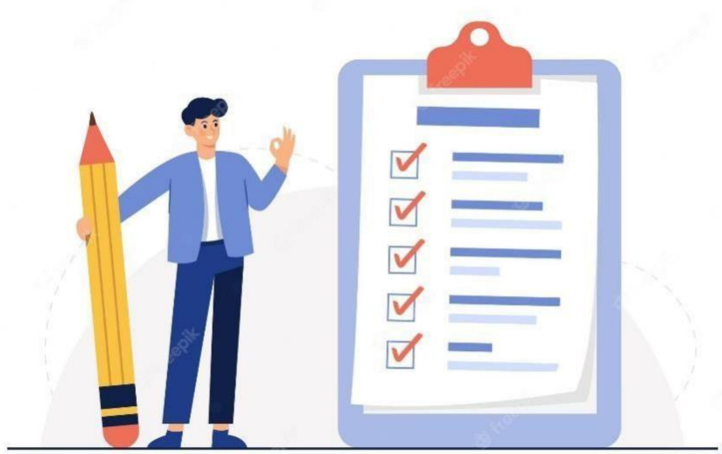
Financial resources play a crucial role, as the organization should have sufficient funds to support a market segmentation strategy. Visible commitment from senior management is necessary, along with their active involvement and required financial support.

Understanding the market segmentation concept and its implications is essential. Training should be conducted until these aspects are fully grasped. Forming a segmentation team with expertise in marketing, data analysis, and data management is important.

An advisory committee representing all affected organizational units should be set up. Clear objectives for the market segmentation analysis should be established, and a structured process should be developed and followed. Responsibilities should be assigned to team members according to the process.

Sufficient time should be allocated for the market segmentation analysis, without time pressure.

In summary, the checklist emphasizes the importance of a market-oriented culture, willingness to change, long-term perspective, open communication, financial resources, and senior management commitment. It also highlights the need for a knowledgeable team, clear objectives, and a structured process to conduct the market segmentation analysis.



STEP 2: SPECIFYING THE IDEAL TARGET SEGMENT

In Step 2 of market segmentation analysis, the focus is on specifying the ideal target segment by establishing segment evaluation criteria. User input plays a crucial role throughout the process, as their involvement is needed to ensure the usefulness of the analysis. The organization must contribute conceptually to the market segmentation analysis, guiding subsequent steps.

Two sets of segment evaluation criteria are defined: knock-out criteria and attractiveness criteria. Knock-out criteria are essential and non-negotiable features that segments must possess to be considered as potential targets. They include factors such as homogeneity, distinctiveness, size, match with organizational strengths, identifiability, and reachability.

On the other hand, attractiveness criteria are used to assess the relative attractiveness of segments that comply with the knock-out criteria. There is a wide range of proposed criteria in the literature, and the segmentation team selects a subset of no more than six criteria that are most relevant to their specific situation.

A structured process is recommended for evaluating market segments. A popular approach involves using a segment evaluation plot, which shows segment attractiveness and organizational competitiveness. The criteria for both attractiveness and competitiveness need to be negotiated and agreed upon by the segmentation team, preferably with input from representatives of different organizational units in the advisory committee.

During this step, the team determines the segment attractiveness criteria and assigns weights to each criterion, indicating its relative importance. The weights are typically decided through team discussions and negotiations, ensuring agreement among team members. The proposed criteria and weights are then presented to the advisory committee for further discussion and adjustment if necessary.

By the end of Step 2, the segmentation team should have a list of segment attractiveness criteria, each with its assigned weight. This groundwork facilitates data collection in Step 3 and simplifies the selection of a target segment in Step 8.

The checklist for Step 2 includes tasks such as convening a segmentation team meeting, discussing and agreeing on knock-out criteria, presenting them to the advisory committee, studying and selecting attractiveness criteria, distributing weights, and presenting the selected criteria and weights to the advisory committee for further discussion and adjustment if needed.

STEP 3: DATA COLLECTION

Step 3 of market segmentation involves collecting data to identify and describe market segments. Empirical data is essential for both commonsense and data-driven segmentation. In commonsense segmentation, a single characteristic, such as gender, is used as the segmentation variable to divide the sample into segments. Other personal characteristics serve as descriptor variables to describe the segments in detail.

Data-driven segmentation, on the other hand, utilizes multiple segmentation variables to identify naturally existing or artificially created market segments. These variables can be characteristics or benefits sought by consumers. The quality of empirical data is crucial in developing valid segmentation solutions and accurately describing the segments.

Data for segmentation studies can be obtained from various sources such as surveys, observations (e.g., scanner data), or experimental studies. Survey data is commonly used but may have limitations in reflecting actual behavior, especially for socially desirable actions. Therefore, alternative data sources should be explored to reflect consumer behavior accurately.

Segmentation criteria are the basis for market segmentation and can be geographic, socio-demographic, psychographic, or behavioral. Geographic segmentation uses location as the criterion, while socio-demographic segmentation considers age, gender, income, and education. Psychographic segmentation focuses on psychological criteria, such as beliefs, interests, preferences, and benefits sought. Behavioral segmentation analyzes actual behavior or reported behavior, such as purchase history or information search behavior.

When collecting data through surveys, the response options provided to respondents play a crucial role. Binary or metric response options are preferred for segmentation analysis as they enable distance measures and statistical procedures. Ordinal response options have an ordered scale but lack a clearly defined distance between adjacent options.

Survey data is susceptible to response biases and styles, which can impact the segmentation results. Response biases include extreme or midpoint responses, while response styles are consistent biases shown by respondents over time. Researchers should minimize these biases to ensure accurate segment identification.

Sample size is another important consideration. Larger sample sizes improve the accuracy of segment extraction. The recommended sample size depends on the number of segmentation variables and should be sufficient to enable correct identification of segments.

Internal data from organizations, such as scanner data or online purchase data, can be valuable for segmentation analysis as they reflect actual consumer behavior. However, caution should be exercised to avoid bias towards existing customers and ensure representation of potential future customers.

Experimental studies can provide data for segmentation analysis, particularly through tests on consumer responses to advertisements or choice experiments that assess preferences for specific product attributes.

In summary, collecting high-quality data is crucial for effective market segmentation. Careful consideration should be given to the selection of segmentation variables, data sources, response options, sample size, and potential biases to ensure accurate segment identification and description.

Step 4: Exploring Data

After data collection, this step involves the following steps:

1) Dataset and Features
2) Data Cleaning
3) Data Analysis
4) Data Pre-Processing and Preparation
5) Principal Component Analysis

This step first cleans the data and then pre-process if necessary and then provides insights about the suitability of different segmentation methods for extracting the market segments.

Data Description:

To illustrate this step, we use McDonald’s dataset.

```
df = pd.read_csv("mcdonalds.csv")
df.head(10)
```

[6]

...

```
df.shape
```

[7]

```
(1453, 15)
```

```
# checking for datatypes and missing values
df.info()
df.isnull().sum()
```

[8]

```
...   <class 'pandas.core.frame.DataFrame'>
      RangeIndex: 1453 entries, 0 to 1452
      Data columns (total 15 columns):
       #   Column          Non-Null Count  Dtype
      ---  ------          --------------  -----
       0   yummy           1453 non-null   object
       1   convenient      1453 non-null   object
       2   spicy           1453 non-null   object
       3   fattening       1453 non-null   object
       4   greasy          1453 non-null   object
       5   fast            1453 non-null   object
       6   cheap           1453 non-null   object
       7   tasty           1453 non-null   object
       8   expensive       1453 non-null   object
       9   healthy         1453 non-null   object
       10  disgusting      1453 non-null   object
       11  Like            1453 non-null   object
       12  Age             1453 non-null   int64
       13  VisitFrequency  1453 non-null   object
       14  Gender          1453 non-null   object
      dtypes: int64(1), object(14)
      memory usage: 170.4+ KB

...   yummy             0
      convenient        0
      spicy             0
      fattening         0
      greasy            0
      fast              0
      cheap             0
      tasty             0
      expensive         0
      healthy           0
      disgusting        0
      Like              0
      Age               0
      VisitFrequency    0
      Gender            0
      dtype: int64
```

Data Cleaning: Before we conduct data analysis, we have to go through the data cleaning step. This step involves to clean unnecessary data and to check the consistent labels for the levels of categorical variable have been used. From the summery of the dataset, we can say that there is no cleaning required for the variable Age and Gender. But categories of Income2 variable are not sorted properly. So, we have sorted them by the following approach.

[11]

```
df['Gender'].value_counts()
```

```
Gender
Female    788
Male      665
Name: count, dtype: int64
```

[12]

```
df['VisitFrequency'].value_counts()
```

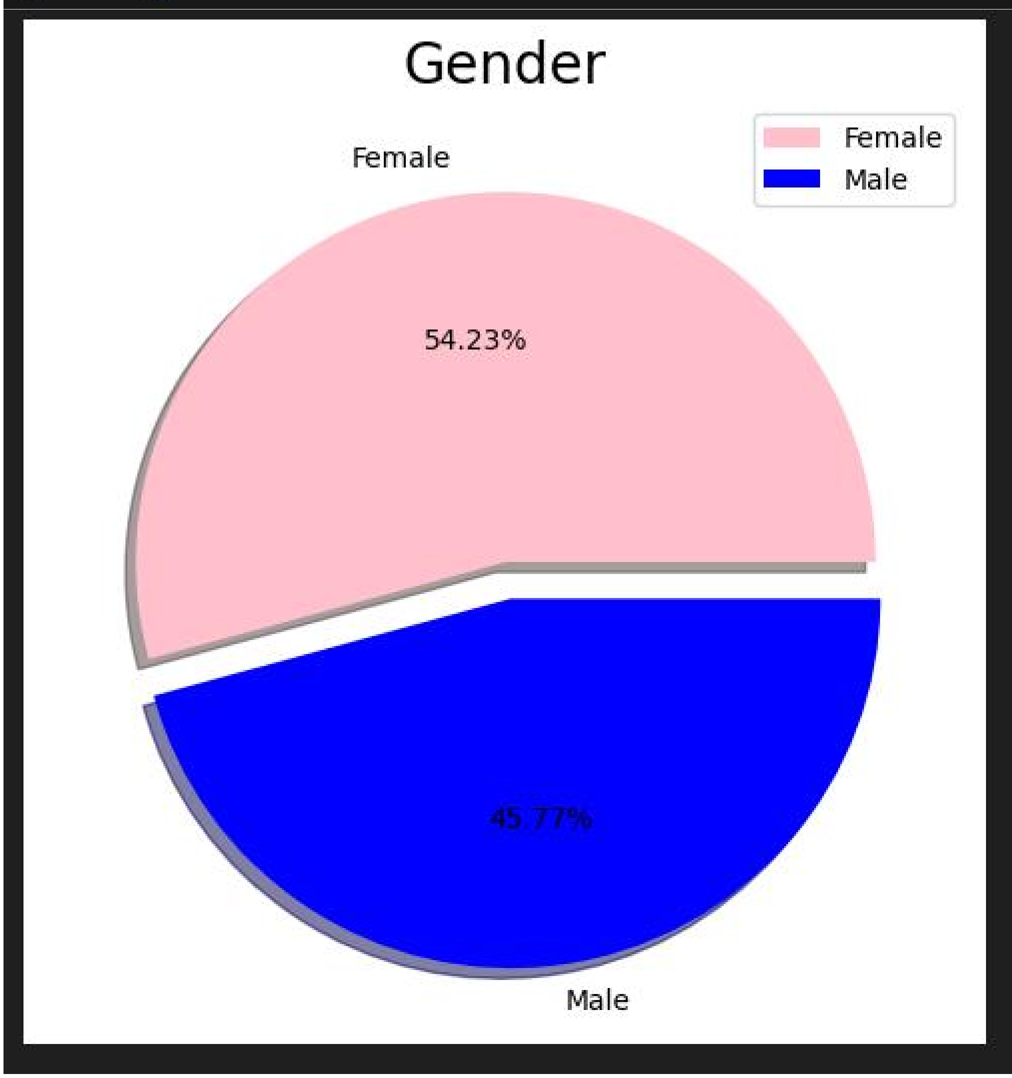
```
VisitFrequency
Once a month             439
Every three months       342
Once a year              252
Once a week              235
Never                    131
More than once a week     54
Name: count, dtype: int64
```

[13]

```
df['Like'].value_counts()
```

```
Like
+3              229
+2              187
0               169
+4              160
+1              152
I hate it!-5    152
I love it!+5    143
-3               73
-4               71
-2               59
-1               58
Name: count, dtype: int64
```

```
# Distribution using gender
labels = ['Female', 'Male']
size = df['Gender'].value_counts()
colors = ['pink', 'blue']
explode = [0, 0.1]
plt.rcParams['figure.figsize'] = (6, 6)
plt.pie(size, colors = colors, explode = explode, labels = labels, shadow = True, autopct = '%.2f%%')
plt.title('Gender', fontsize = 20)
plt.axis('off')
plt.legend()
plt.show()
```



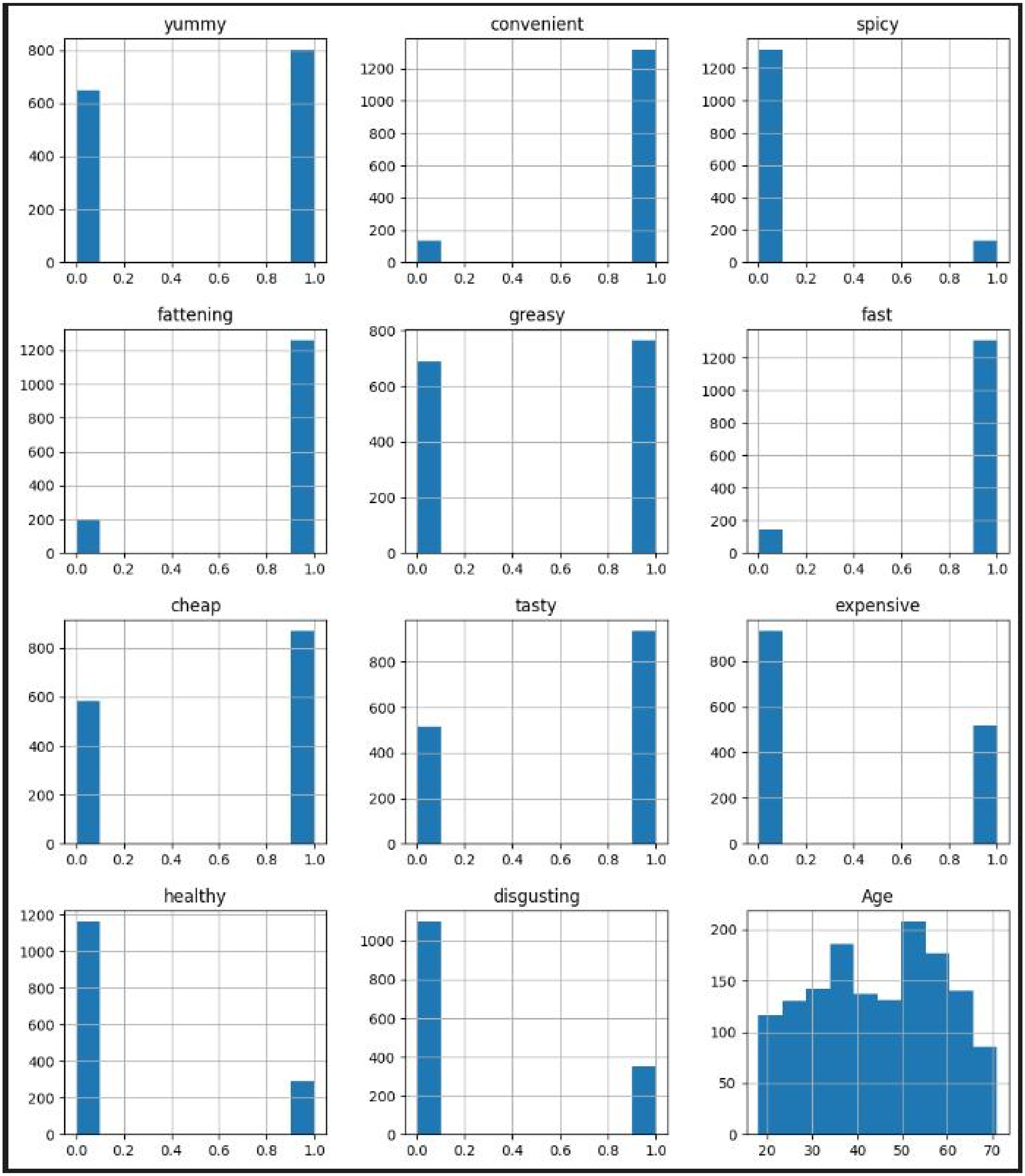
Descriptive Analysis:

Descriptive analysis helps to get meaning insights about the data through analysis. There are various plot which can help us to analyse data like histogram, box-plot, scatter plot etc. Histogram is a graphical representation of the distribution of the data. It helps us to visualize the distribution of the numeric variable. It also helps us to visualize the frequency of the observations within a certain range. We can check weather a distribution of a variable is symmetric or skewed using histogram. To create histogram, first we have to create bins (i.e, categories of values).

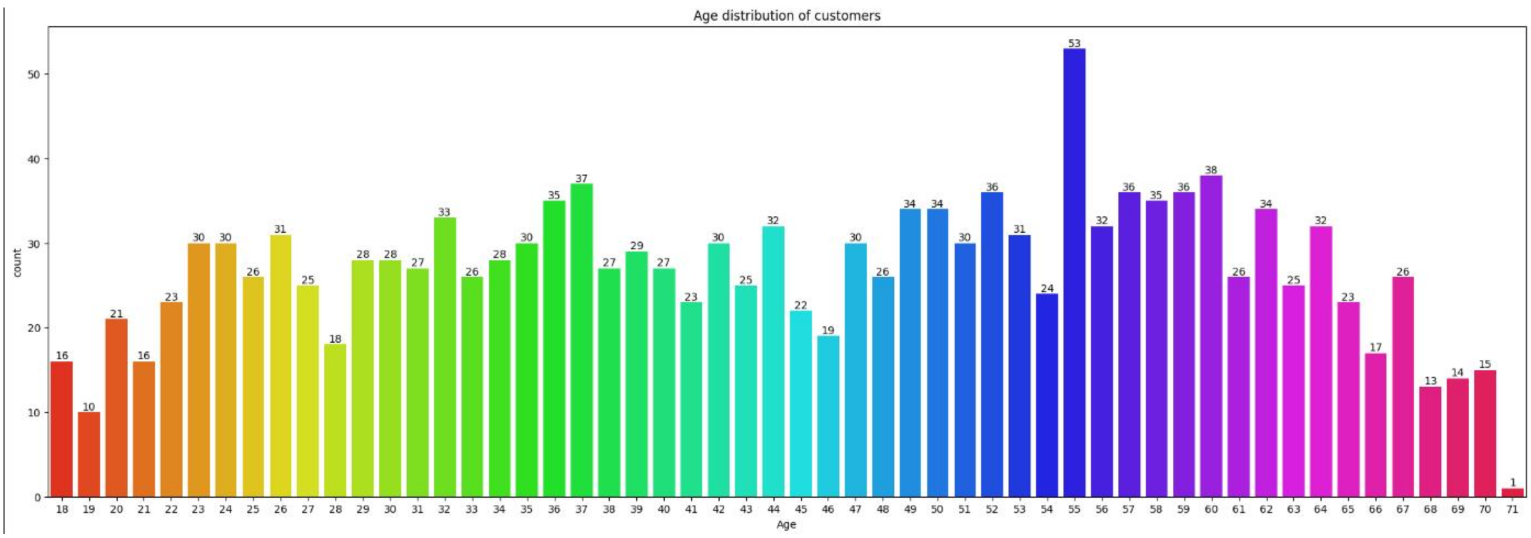
```
#Label encoding for categorical - Converting 11 cols with yes/no

from sklearn.preprocessing import LabelEncoder
def labelling(x):
    df[x] = LabelEncoder().fit_transform(df[x])
    return df

cat = ['yummy', 'convenient', 'spicy', 'fattening', 'greasy', 'fast', 'cheap',
        'tasty', 'expensive', 'healthy', 'disgusting']

for i in cat:
    labelling(i)
df
```

```
0 - no
1 - yes

one thing that is important in this dataset all the field is import and relevant to get the conclusion of segmenatation so we can not neglect any field of dataset
```

```
#Histogram of the each attributes
plt.rcParams['figure.figsize'] = (12,14)
df.hist()
plt.show()
```


yummy
convenient
spicy
fattening
greasy
fast
cheap
tasty
expensive
healthy
disgusting
Age

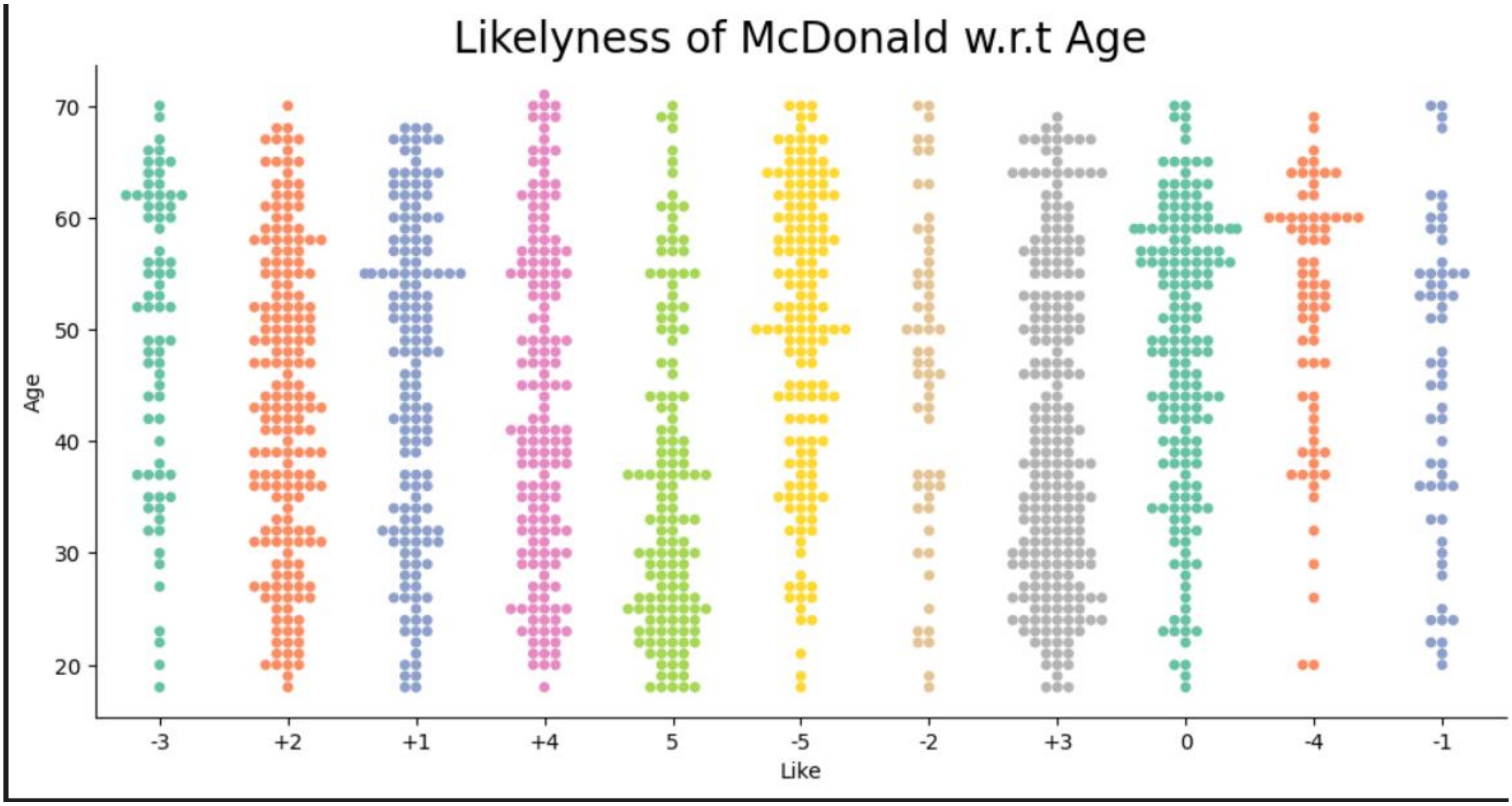
```python
#distribution on age
import seaborn as sns
plt.rcParams['figure.figsize'] = (25, 8)
f = sns.countplot(x=df['Age'],palette = 'hsv')
f.bar_label(f.containers[0])
plt.title('Age distribution of customers')
plt.show()
```



We created 50 bins for the above histogram of the Age variable. From the above figure, we can say that the distribution is bimodal with many respondents aged around 35-40 and around 60 years.

Another useful visualization tool is boxplot. This plot compresses a data into minimum, first quartile, median, third quartile and maximum. These five numbers often called five-point summary.

```python
# Pyschographic segmentation using 'Like'

# renaming the category for convinence
df['Like'] = df['Like'].replace({'I hate it!-5': '-5','I love it!+5':'+5'})

# plotting the results
sns.catplot(data=df, x="Like", y="Age", orient="v", height=5, aspect=2, palette="Set2",kind="swarm")
plt.title('Likelyness of McDonald w.r.t Age', fontsize=20)
plt.show()
```



Pre-Processing:

```
considering first 11 fields of dataset so we find the principal axis it helful to reduce the dimension because here all the field is important
```

```
df_imp = df.loc[:,cat]
df_imp
```

```
#converting these 11 colomns into arrays
x = df.loc[:,cat].values
x
```

Principal Components Analysis:

Principal component analysis is a dimensionality reduction technique that is often used to reduce the dimensionality of the large dataset into the smaller one that still contains the most information about the large dataset. Since principal component analysis reduces the number of variables, it naturally comes at the expense of accuracy. For conducting PCA, first we need to standardize the range of the continuous variable. Mathematically, this can be done by,

(Value-mean)/standard deviation

```
#creating a principal component axis along these colomns
from sklearn.decomposition import PCA
from sklearn import preprocessing

pca_data = preprocessing.scale(x)

pca = PCA(n_components=11)
pc = pca.fit_transform(x)
names = ['pc1','pc2','pc3','pc4','pc5','pc6','pc7','pc8','pc9','pc10','pc11']
pf = pd.DataFrame(data = pc, columns = names)
pf
```

	pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8	pc9	pc10	pc11
0	0.425367	-0.219079	0.663255	-0.401300	0.201705	-0.389767	-0.211982	0.163235	0.181007	0.515706	-0.567074
1	-0.218638	0.388190	-0.730827	-0.094724	0.044669	-0.086596	-0.095877	-0.034756	0.111476	0.493313	-0.500440
2	0.375415	0.730435	-0.122040	0.692262	0.839643	-0.687406	0.583112	0.364379	-0.322288	0.061759	0.242741
3	-0.172926	-0.352752	-0.843795	0.206998	-0.681415	-0.036133	-0.054284	-0.231477	-0.028003	-0.250678	-0.051034
4	0.187057	-0.807610	0.028537	0.548332	0.854074	-0.097305	-0.457043	0.171758	-0.074409	0.031897	0.082245
...	...	...	...	...	...	...	...	...	...	...	...
1448	1.550242	0.275031	-0.013737	0.200604	-0.145063	0.306575	-0.075308	0.345552	-0.136589	-0.432798	-0.456076
1449	-0.957339	0.014308	0.303843	0.444350	-0.133690	0.381804	-0.326432	0.878047	-0.304441	-0.247443	-0.193671
1450	-0.185894	1.062662	0.220857	-0.467643	-0.187757	-0.192703	-0.091597	-0.036576	0.038255	0.056518	-0.012800
1451	-1.182064	-0.038570	0.561561	0.701126	0.047645	0.193687	-0.027335	-0.339374	0.022267	-0.002573	-0.105316
1452	1.550242	0.275031	-0.013737	0.200604	-0.145063	0.306575	-0.075308	0.345552	-0.136589	-0.432798	-0.456076

1453 rows × 11 columns

Next, we must calculate covariance matrix to see if there is any relationship between the variables. Principal components are new variables that are constructed as linear combination of the initial variables. These combinations are done in such a way new variables (principal components) are uncorrelated and most of the information within the initial variables is compressed into the first components. So, the idea is 10-dimensional data gives us 10 principal components, but PCA puts most of the information into the component, then maximum remaining information into the second principal component and so on.

```
#Proportion of Variance (from PC1 to PC11)
pd.Series(pca.explained_variance_ratio_)
```

```
0     0.299447
1     0.192797
2     0.133045
3     0.083096
4     0.059481
5     0.050300
6     0.043849
7     0.039548
8     0.036761
9     0.032353
10    0.029323
dtype: float64
```

We can conduct further analysis on the fitted object with the summary function.

```
#so after finding the PCA axis we can see last three PCA axis does not give more variance
np.cumsum(pca.explained_variance_ratio_)
```

```
array([0.29944723, 0.49224445, 0.6252898 , 0.70838558, 0.7678661 ,
       0.81816566, 0.86201476, 0.90156255, 0.93832345, 0.97067674,
       1.        ])
```

first seven PCA axis are capable to give 85% of information
Now we find Corelation between original component and PCA components

```python
loadings = pca.components_
num_pc = pca.n_features_
pc_list = ["PC"+str(i) for i in list(range(1, num_pc+1))]
loadings_df = pd.DataFrame.from_dict(dict(zip(pc_list, loadings)))
loadings_df['variable'] = df_imp.columns.values
loadings_df = loadings_df.set_index('variable')
loadings_df
```

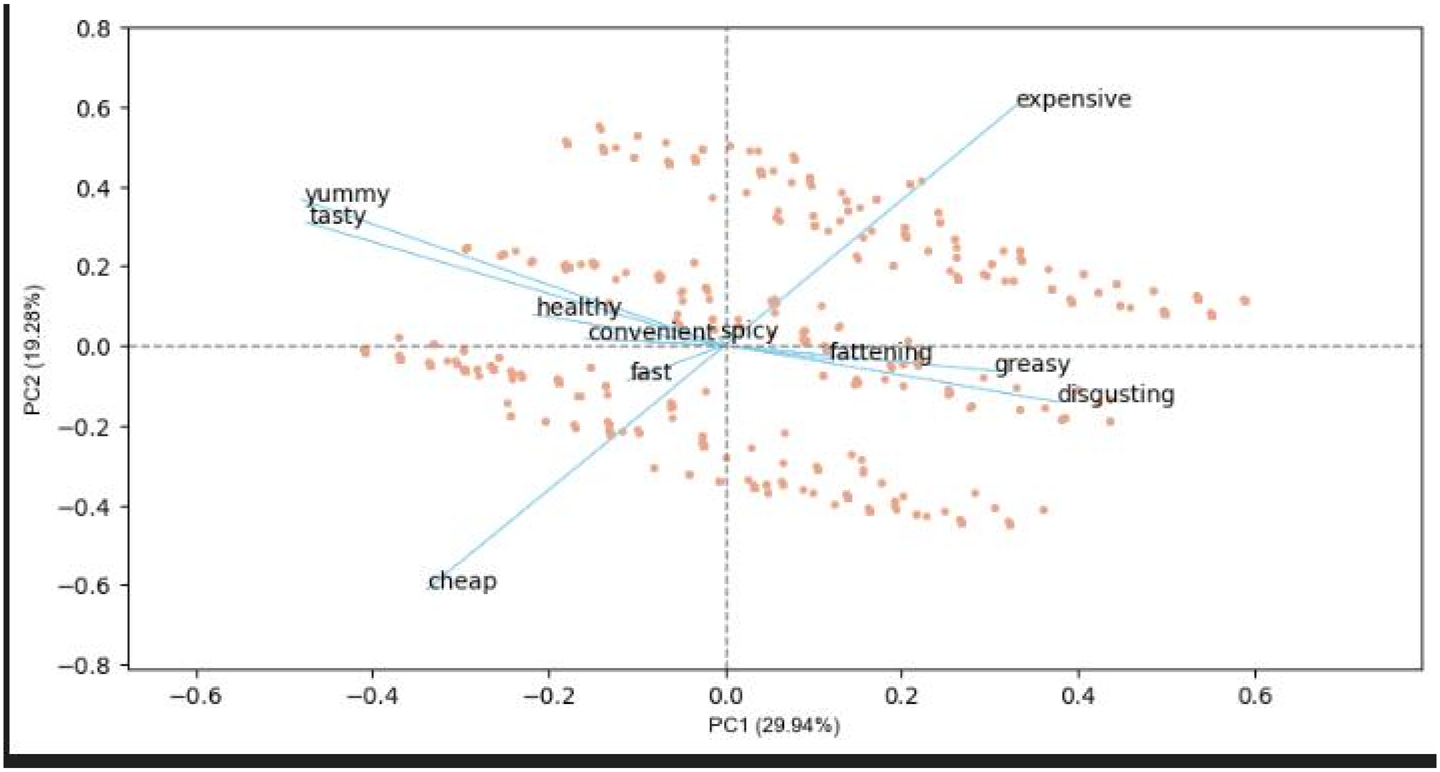
```
D:\ana\Lib\site-packages\sklearn\utils\deprecation.py:101: FutureWarning: Attribute `n_features_` was deprecated in version 1.2 and will be removed in 1.4. Use `n_features_in_` instead.
  warnings.warn(msg, category=FutureWarning)
```

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
yummy	-0.476933	0.363790	-0.304444	0.055162	-0.307535	0.170738	-0.280519	0.013041	0.572403	-0.110284	0.045439
convenient	-0.155332	0.016414	-0.062515	-0.142425	0.277608	-0.347830	-0.059738	-0.113079	-0.018465	-0.665818	-0.541616
spicy	-0.006356	0.018809	-0.037019	0.197619	0.070620	-0.355087	0.707637	0.375934	0.400280	-0.075634	0.141730
fattening	0.116232	-0.034094	-0.322359	-0.354139	-0.073405	-0.406515	-0.385943	0.589622	-0.160512	-0.005338	0.250910
greasy	0.304443	-0.063839	-0.802373	0.253960	0.361399	0.209347	0.036170	-0.138241	-0.002847	0.008707	0.001642
fast	-0.108493	-0.086972	-0.064642	-0.097363	0.107930	-0.594632	-0.086846	-0.627799	0.166197	0.239532	0.339265
cheap	-0.337186	-0.610633	-0.149310	0.118958	-0.128973	-0.103241	-0.040449	0.140060	0.076069	0.428087	-0.489283
tasty	-0.471514	0.307318	-0.287265	-0.002547	-0.210899	-0.076914	0.360453	-0.072792	-0.639086	0.079184	0.019552
expensive	0.329042	0.601286	0.024397	0.067816	-0.003125	-0.261342	-0.068385	0.029539	0.066996	0.454399	-0.490069
healthy	-0.213711	0.076593	0.192051	0.763488	0.287846	-0.178226	-0.349616	0.176303	-0.185572	-0.038117	0.157608
disgusting	0.374753	-0.139656	-0.088571	0.369539	-0.729209	-0.210878	-0.026792	-0.167181	-0.072483	-0.289592	-0.040662

Interpretation: Principal component 1 explains 18% of the variation in the original data. Principal component 2 explains 9% of the variation in the original data.

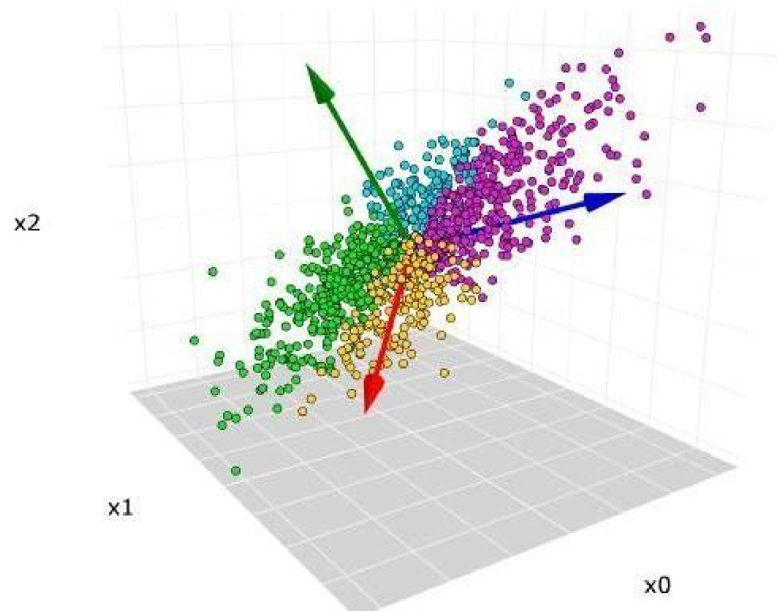
Now, we want to plot the data into two-dimensional space. Inspecting the rotation matrix reveals that the first principal component does not differentiate well between motives because all motives load on it negatively. So, we consider principal component 2 and principal component 3 to create perceptual map. Function projAxes plots how the principal components are composed of the original variables and visualises the rotation matrix.

```
pca_scores = PCA().fit_transform(x)

cluster.biplot(cscore=pca_scores, loadings=loadings, labels=df.columns.values, var1=round(pca.explained_variance_ratio_[0]*100, 2),
    var2=round(pca.explained_variance_ratio_[1]*100, 2),show=True,dim=(10,5))
```



Interpretation: NOT EXCEEDING THE PLANNED BUDGET (represented by the arrow pointing in the top slightly left direction) is a travel motive that is quite unique. On the other hand, LIFESTYLE OF LOCAL PEOPLE, and interest in CULTURAL OFFERS available at destinations often occur simultaneously (as indicated by the two arrows both pointing to the left bottom).

We can use principal component analysis to reduce the number of segmentation variable before extracting the market segment from the consumer data. It is helpful because after reducing the number of variables it provides meaningful insights which is easy handle.



Step 5: Extracting Segments

Step 5 is where we extract segments. To illustrate a range of extraction techniques, we subdivide this step into sections. In the first section, we will use standard k-means analysis.

5.1 Grouping Consumers:

For extracting market segments from data mostly clustering analysis is used with a variety of approaches. Selecting a suitable clustering method requires matching the data analytic featuresof the resulting clustering with the context-dependent requirements that are desired by the researcher of the market.

Some of Data set and segment characteristics informing extraction algorithm selection given:

Data set characteristics: – Size (number of consumers, number of segmentation variables) – Scale level of segmentation variables (nominal, ordinal, metric, mixed) – Special structure, additional information.

Segment characteristics: – Similarities of consumers in the same segment – Differences between consumers from different segments – Number and size of segments

None of these methods outperform other methods in all situations. Rather, each method has pros and cons.

Distance-based methods use a particular notion of similarity or distance between observations (consumers), and try to find groups of similar observations (market segments). For distance- based methods, the choice of the distance measure depends on the scale level of the data.

Model-based methods formulate a concise stochastic model for the market segments where mainly distributions and probabilistic approach is used. Model based approaches are more probability based considering parameters of segment size and characteristics consumer's probability of getting fit into segment is derived and best solution is provided in end. If the data set contains repeated measurements of consumers over time, for example, an algorithm that takes this longitudinal natureof the data into account is needed. Such data generally requires a model-

based approach. If the data contains purchase histories and price information, and market segments are based on similar price sensitivity levels, regression models are needed. This, in turn, calls for theuse of a model-based segment extraction algorithm.

In the case of binary segmentation variables, another aspect needs to be considered. We may want consumers in the same segments to have both the presence and absence of segmentation variables in common, here these variables would be symmetrical (with 0s and 1s treated equally).

Alternatively, we may be concerned about segmentation variables consumers have in common, herethese variables would be symmetrical (with only common 1s being of interest). Biclustering uses binary information asymmetrically. Distance-based methods can use distance measures that accountfor this asymmetry, and extract segments characterized by common 1s.

Data-driven market segmentation analysis is exploratory by nature. Consumer data sets are typically not well structured. Consumers come in all shapes and forms a two-dimensional plot of consumers' product preferences typically does not contain clear groups of consumers. Rather, consumer preferences are spread across the entire plot. The combination of exploratory methods andunstructured consumer data that results from any method used to extract market segments from suchdata will strongly depend on the assumptions made on the structure of the segments implied by the method. The result of a market segmentation analysis, therefore, is determined as much by the underlying data as it is by the extraction algorithm chosen.

There are few types of **Distance-based Methods:**

5.2 Distance Methods:

1) Euclidean distance

Euclidean Distance represents the shortest distance between two points. Most machine learning algorithms including K-Means use this distance metric to measure the similarity between observations

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

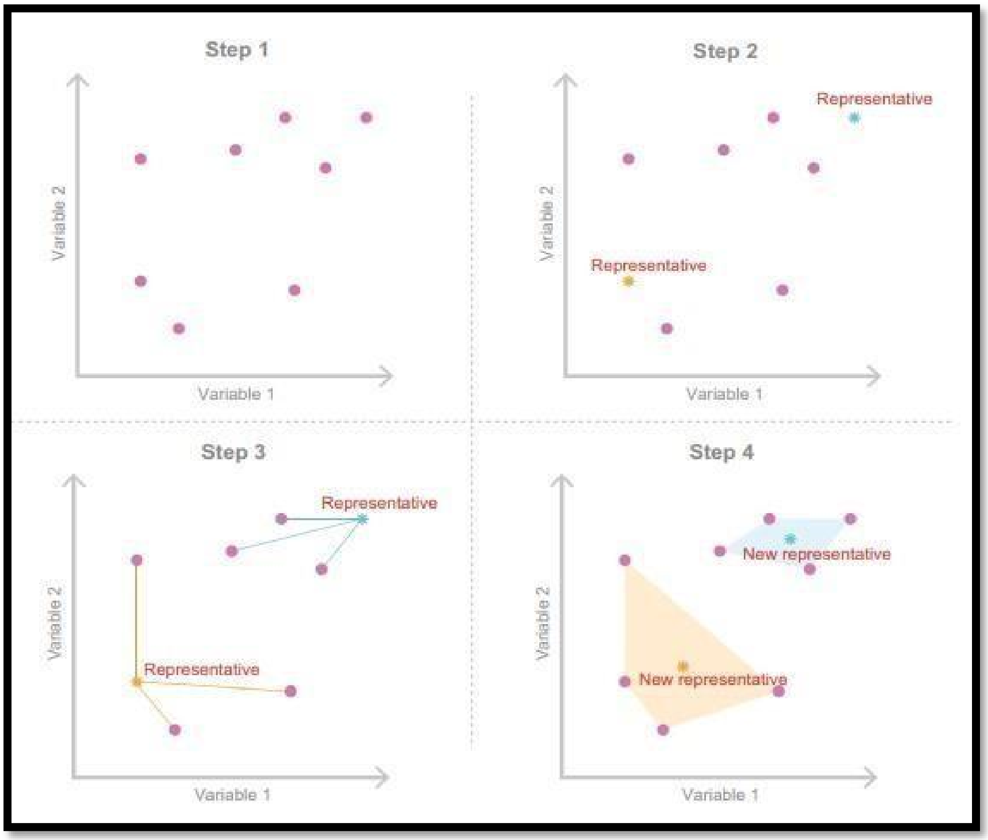
2) Hierarchical measures

Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations(consumers) into k groups (segments). If the aim is to have one large market segment (k = 1), the only possible solution is one big market segment containing all consumers in data X.

$$l(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} d(\mathbf{x}, \mathbf{y}),$$

3) Partitioning method

Hierarchical clustering methods are particularly well suited for the analysis of small data sets with up to a few hundred observations. For larger data sets, dendrograms are hard to read, and the matrix of pairwise distances usually does not fit into computer memory**.** For data sets containing more than 1000 observations (consumers), clustering methods creating a single partition are more suitable thana nested sequence of partitions. This means that – instead of computing all distances between all pairsof observations in the data set.

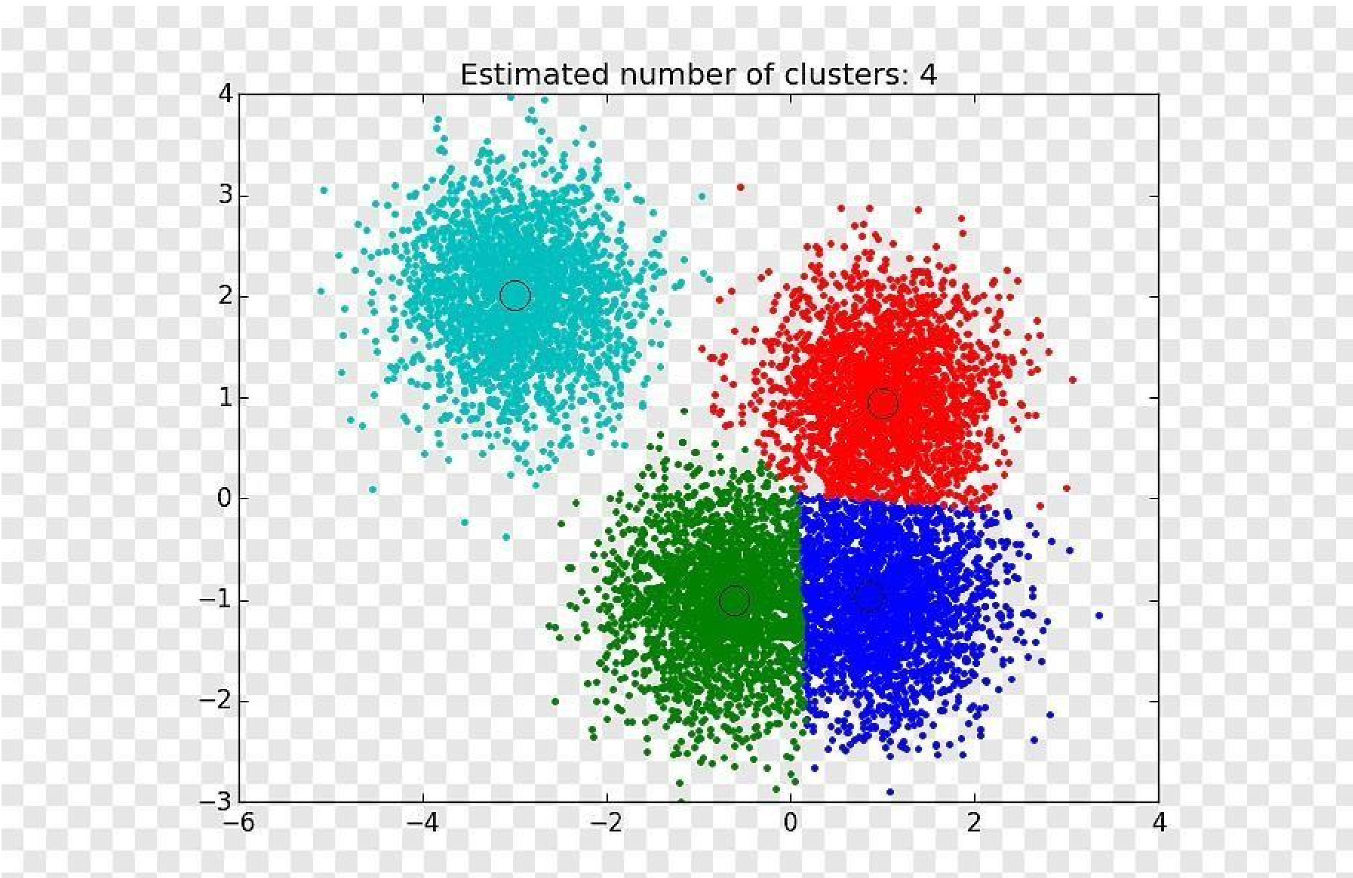


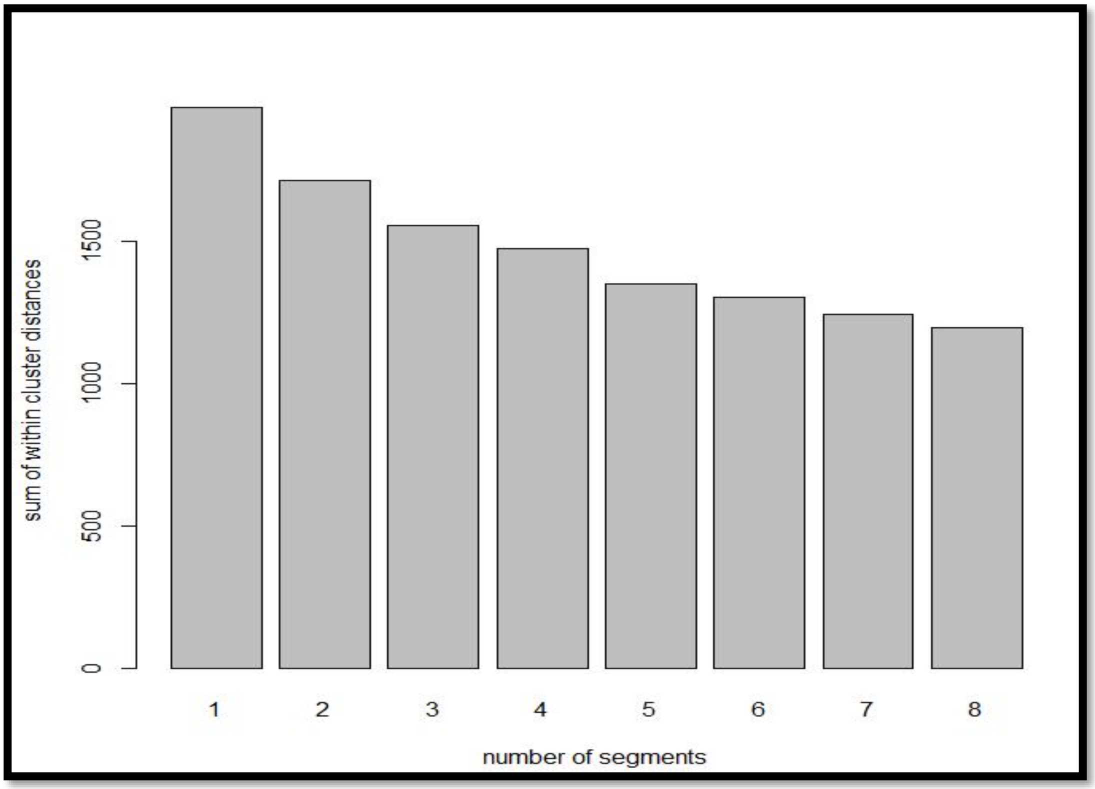
Simplified visualisation of the k-means clustering algorithm

5.3 Using k-Means

We calculate solutions for two to eight market segments using standard k-means analysis with ten random restarts. We then relabel segment numbers such that they are consistent across segmentations.

We extract between two and eight segments because we do not know in advance what the best numberof market segments is. If we calculate a range of solutions, we can compare them and choose the onewhich extracts segments containing similar consumers which are distinctly different from members of other segments. We compare different solutions using a scree plot:

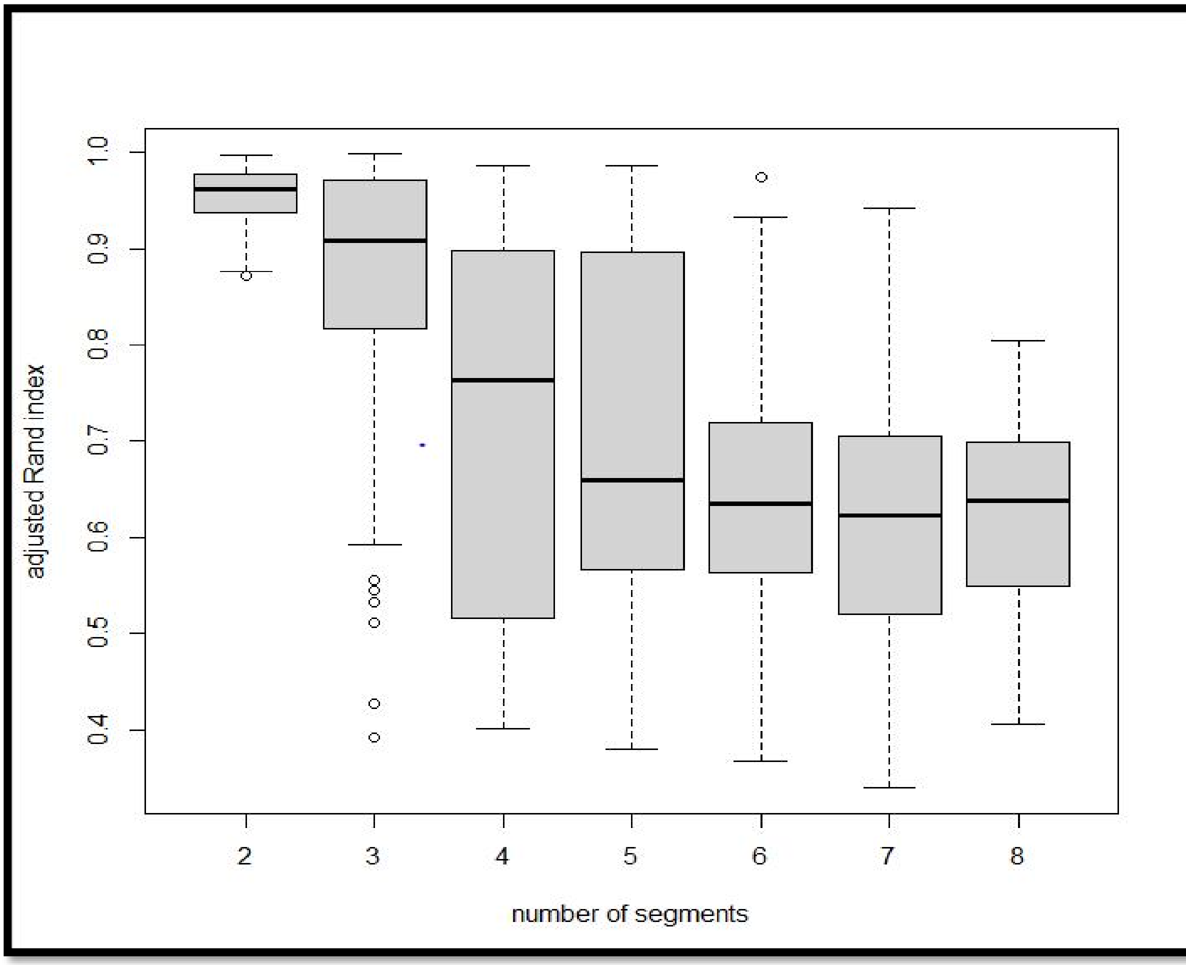




Scree plot for the McDonald's data set

The scree plot has no distinct elbow: the sum of distances within market segments drops slowly as the number of market segments increases. We expect the values to decrease because more market segments automatically mean that the segments are smaller and, as a consequence, that segment members are more similar to one another. But the much-anticipated point where the sum of distancesdrops dramatically is not visible. This scree plot does not provide useful guidance on the number of market segments to extract.

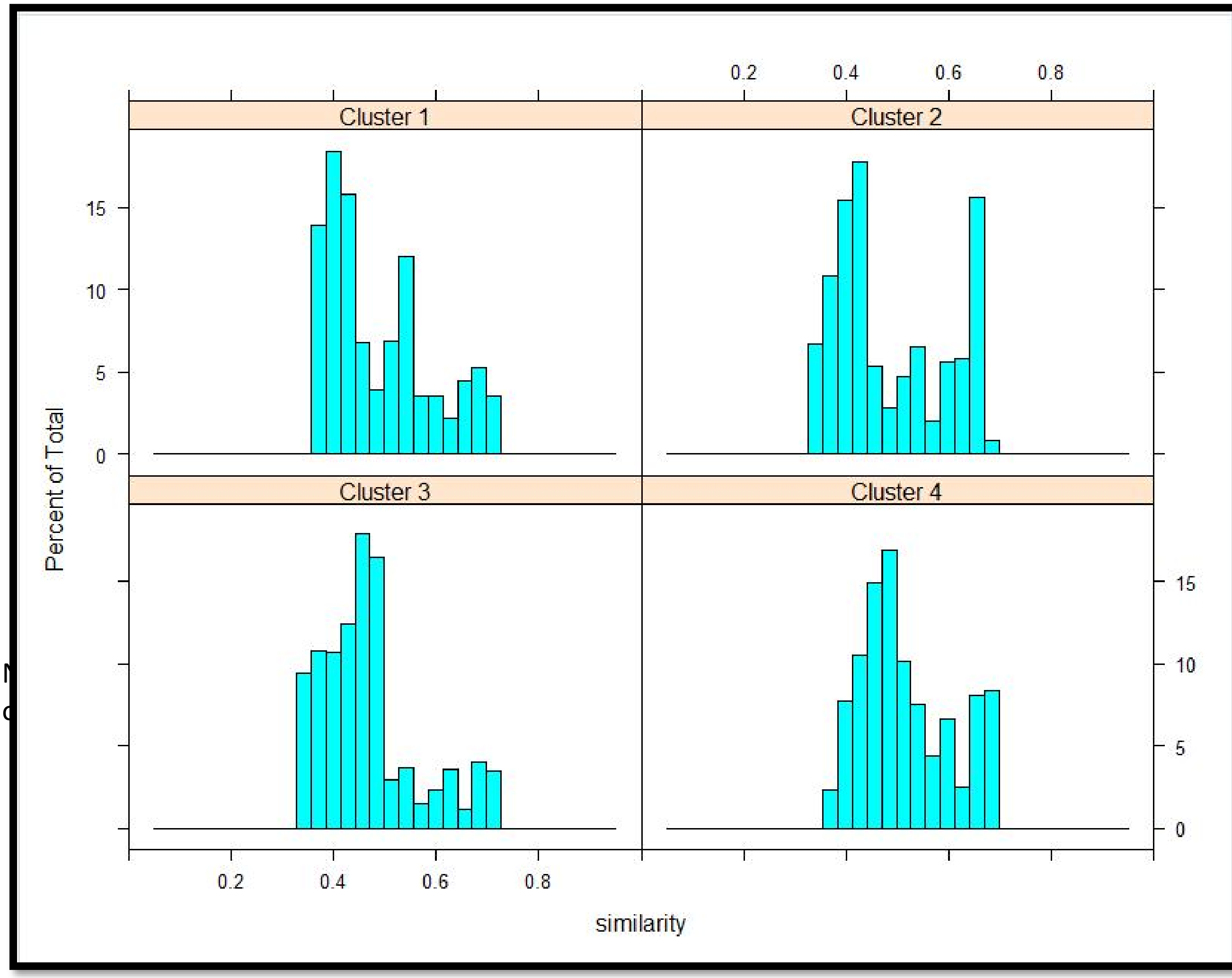
A second approach to determining a good number of segments is to use stability-based data structureanalysis. Stability-based data structure analysis also indicates whether market segments occurnaturally in the data, or if they have to be artificially constructed. Stability-based data structure analysis uses stability across replications as criterion to offer this guidance. Imagine using a market segmentation solution which cannot be reproduced. Such a solution would give McDonald's management little confidence in terms of investing substantial resources into a market segmentationstrategy.

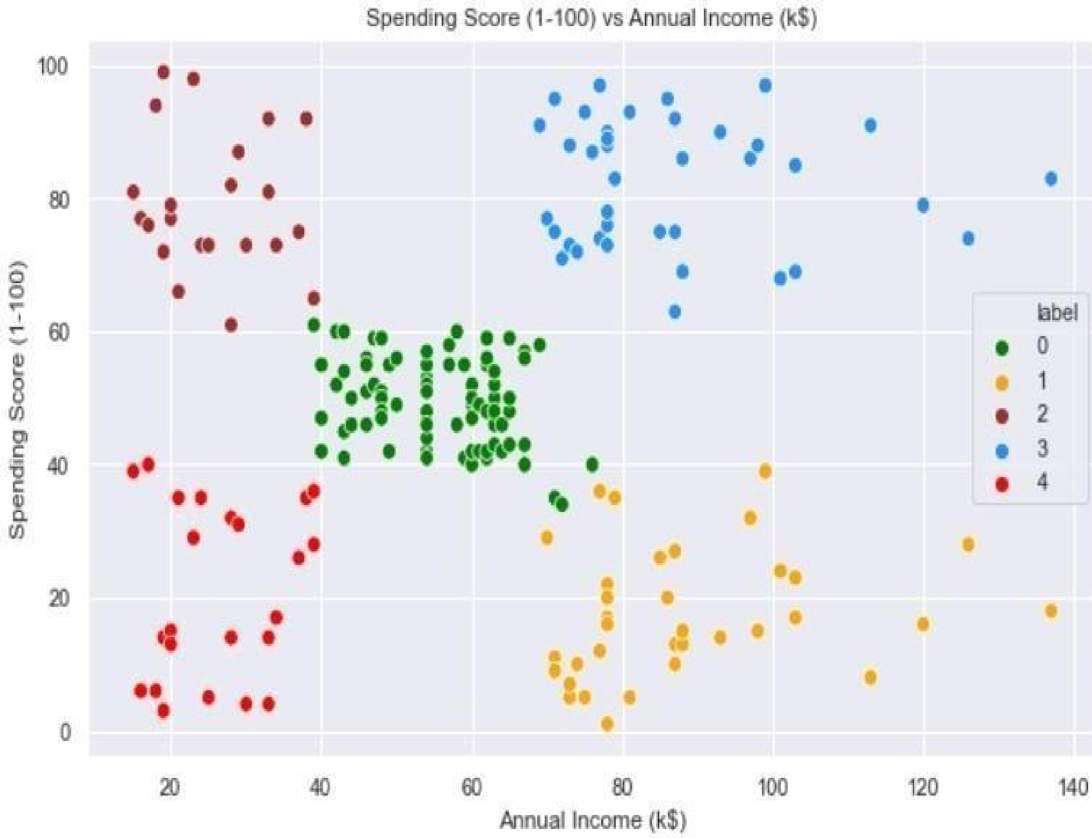


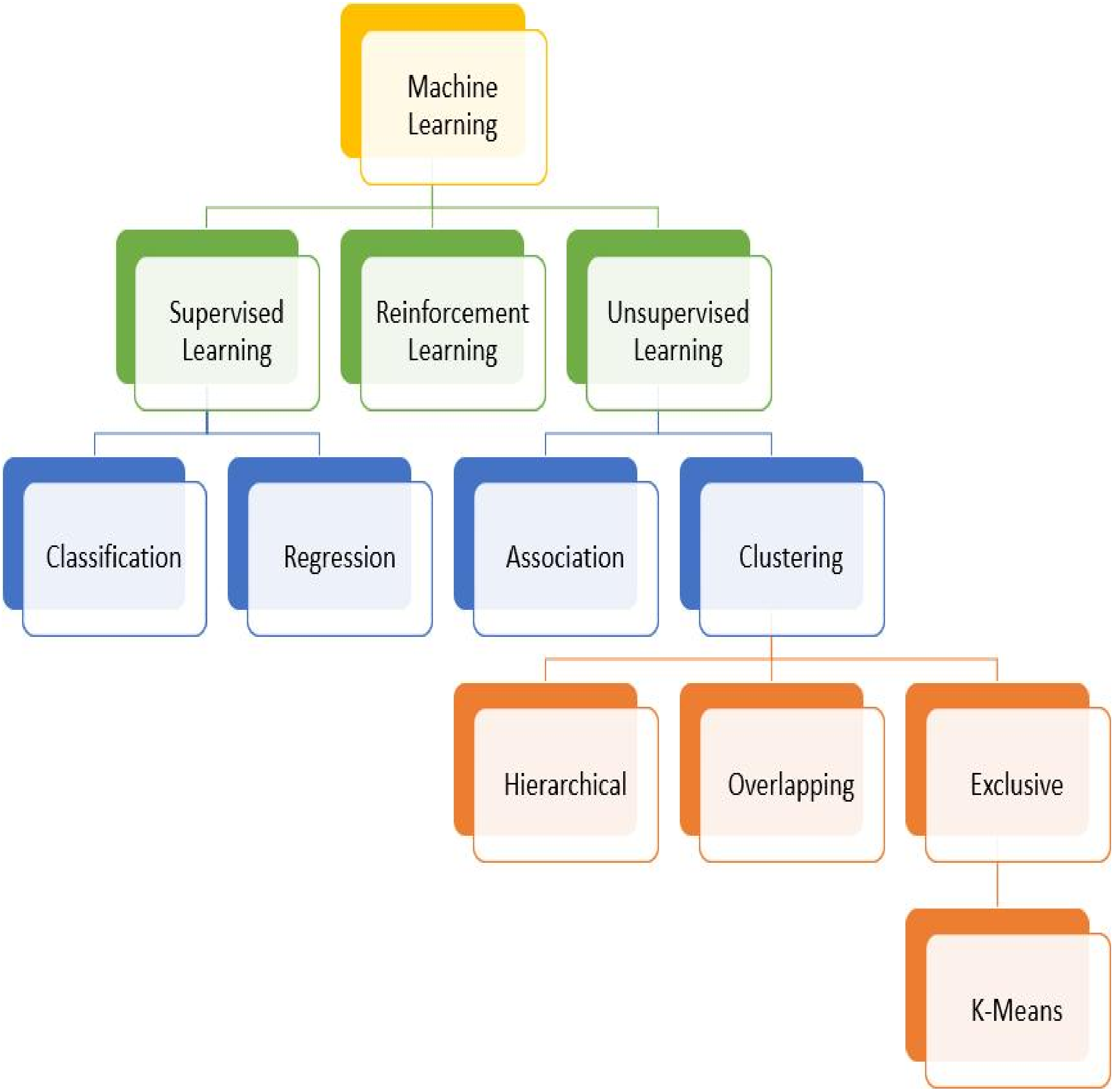
Global stability of k-means segmentation solutions for the McDonald's data set

The vertical boxplots show the distribution of stability for each number of segments. The median is indicated by the fat black horizontal line in the middle of the box. Higher stability is better. Inspecting points to the two-, three- and four-segment solutions as being quite stable. However, the two- and three-segment solutions do not offer a very differentiated view of the market.

We gain further insights into the structure of the four-segment solution with a gorge plot:



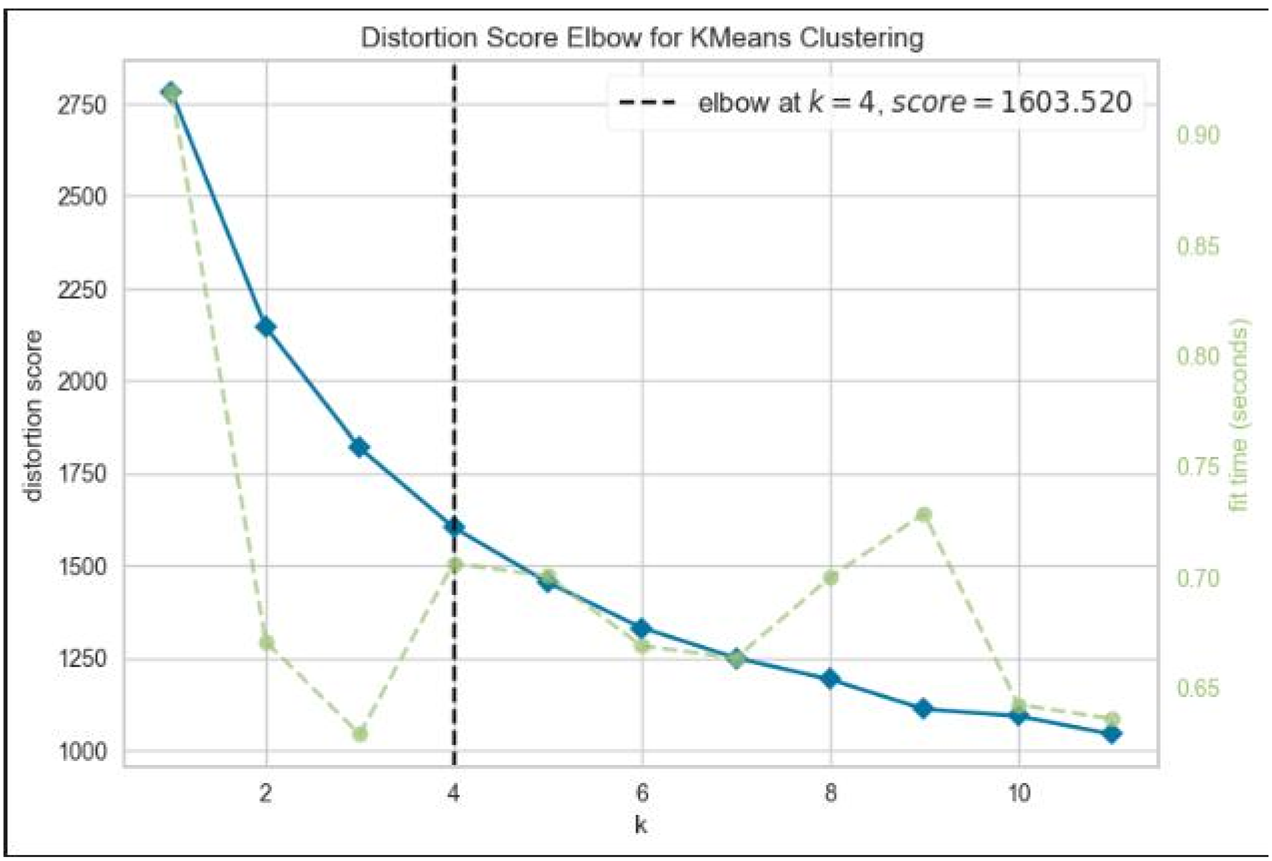

Spending Score (1-100) vs Annual Income (k$)
Spending Score (1-100)
Annual Income (k$)
label
0
1
2
3
4
0
20
40
60
80
100
20
40
60
80
100
120
140


Machine Learning
Supervised Learning
Reinforcement Learning
Unsupervised Learning
Classification
Regression
Association
Clustering
Hierarchical
Overlapping
Exclusive
K-Means

5.4 Using Mixtures of Distributions:

We calculate latent class analysis using a finite mixture of binary distributions. The mixture model maximizes the likelihood to extract segments (as opposed to minimizing squared Euclidean distance, as is the case for k-means).

```
#Using k-means clustering analysis
from sklearn.cluster import KMeans
from yellowbrick.cluster import KElbowVisualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(1,12)).fit(df_imp)
visualizer.show()
plt.show()
```



Information criteria for the mixture models of binary distributions with 2 to 8 components (segments) for the McDonald's data set

K	$=$	number of mixture components
N	$=$	number of observations
$\theta_{i=1\dots K}$	$=$	parameter of distribution of observation associated with component i
$\phi_{i=1\dots K}$	$=$	mixture weight, i.e., prior probability of a particular component i
$\boldsymbol{\phi}$	$=$	K-dimensional vector composed of all the individual $\phi_{1\dots K}$; must sum to 1
$z_{i=1\dots N}$	$=$	component of observation i
$x_{i=1\dots N}$	$=$	observation i
$F(x\|\theta)$	$=$	probability distribution of an observation, parametrized on θ
$z_{i=1\dots N}$	$\sim$	$\operatorname{Categorical}(\boldsymbol{\phi})$
$x_{i=1\dots N}\|z_{i=1\dots N}$	$\sim$	$F(\theta_{z_i})$

Step 6: Describing Segments

9.1 Developing a Complete Picture Market Segments:

In this section, the goal is to gain a comprehensive understanding of market segments. This involves gathering data and conducting analysis to identify key characteristics and attributes that define each segment.

9.2 Using Visualizations to Describe Market Segments:

Visualizations are powerful tools for conveying information about market segments. This subsection explores the use of visual representations, such as charts and graphs, to effectively communicate the characteristics and differences between various segments.

9.2.1 Nominal and Ordinal Descriptor Variables:

This subsection discusses the use of nominal and ordinal descriptor variables in visualizations. Nominal variables represent categories without a specific order, while ordinal variables have a predetermined order. Visualizations help in presenting these variables in a clear and understandable manner.

9.2.2 Metric Descriptor Variables:

Metric descriptor variables are discussed in this section. These variables are quantitative and can be measured on a continuous scale. Visualizations are utilized to showcase relationships, trends, and differences among market segments based on these metric variables.

9.3 Testing for Segment Differences in Descriptor Variables:

To determine if there are significant differences between market segments, statistical tests are performed on descriptor variables. This subsection explains various statistical methods that can be employed to test for differences and establish the uniqueness of each segment.

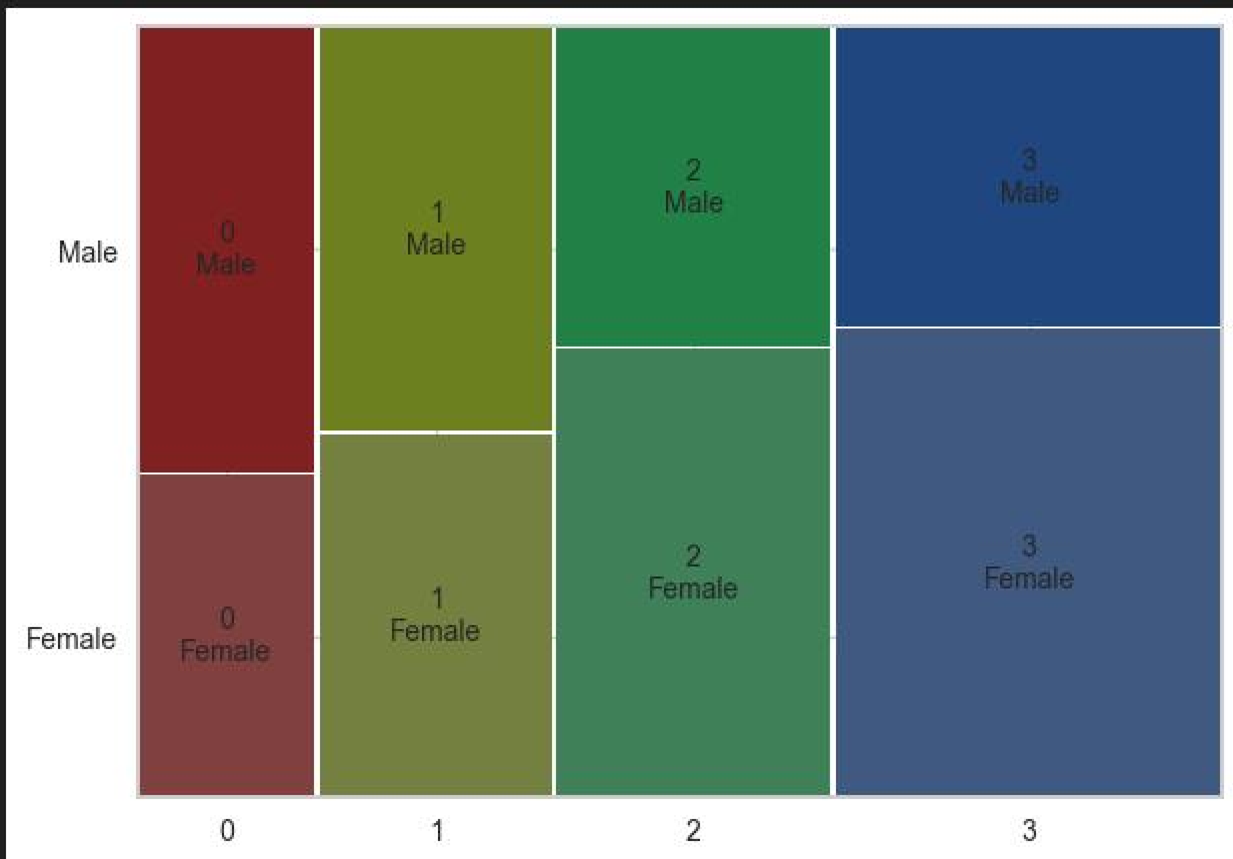
```
from statsmodels.graphics.mosaicplot import mosaic
from itertools import product

crosstab =pd.crosstab(df['cluster_num'],df['Like'])
#Reordering cols
crosstab = crosstab[['-4','-3','-2','-1','0','+1','+2','+3','+4']]
crosstab
```

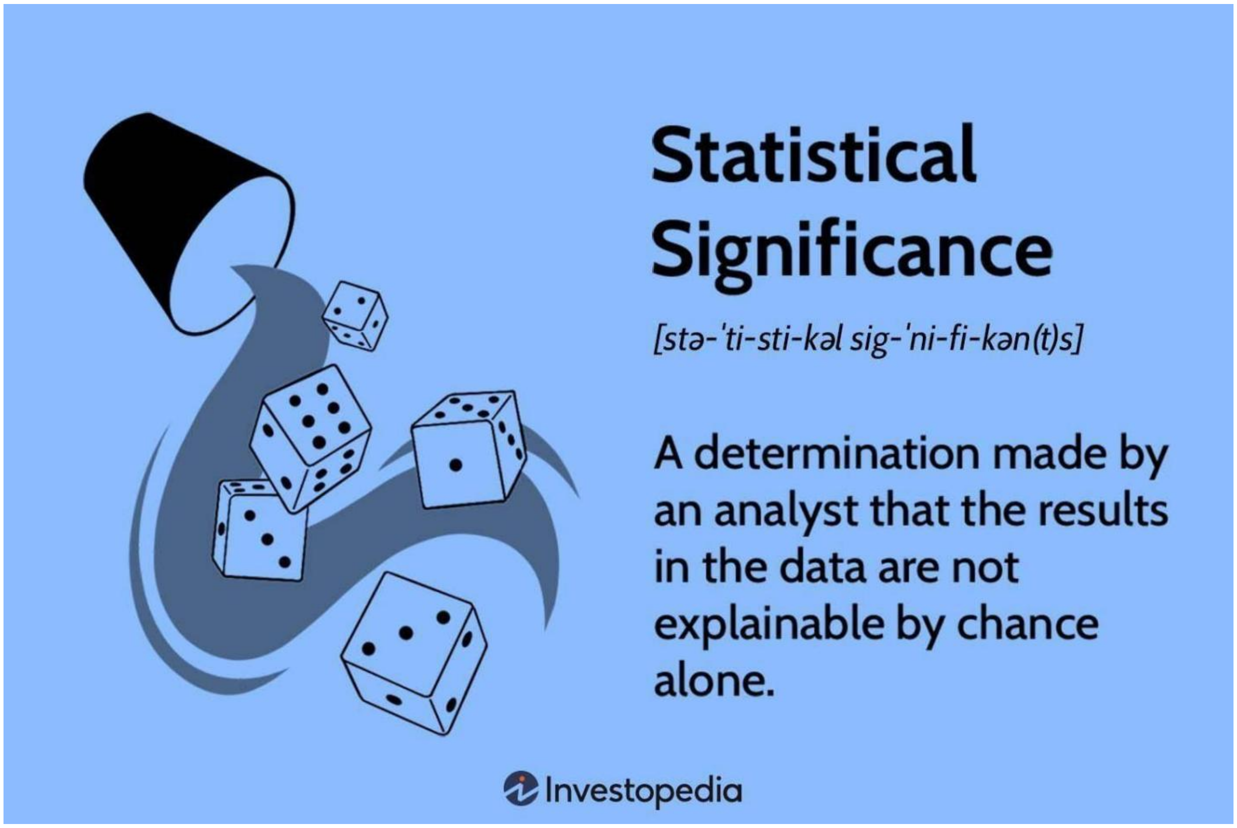
Like	-4	-3	-2	-1	0	+1	+2	+3	+4
cluster_num									
0	27	30	18	12	36	14	6	8	0
1	3	7	6	6	32	41	58	66	47
2	38	36	30	30	68	48	45	17	4
3	3	0	5	10	33	49	78	138	109

```
crosstab_gender =pd.crosstab(df['cluster_num'],df['Gender'])
crosstab_gender
```

Gender	Female	Male
cluster_num		
0	101	139
1	149	166
2	217	155
3	321	205

```
plt.rcParams['figure.figsize'] = (7,5)
mosaic(crosstab_gender.stack())
plt.show()
```





9.4 Predicting Segments from Descriptor Variables:

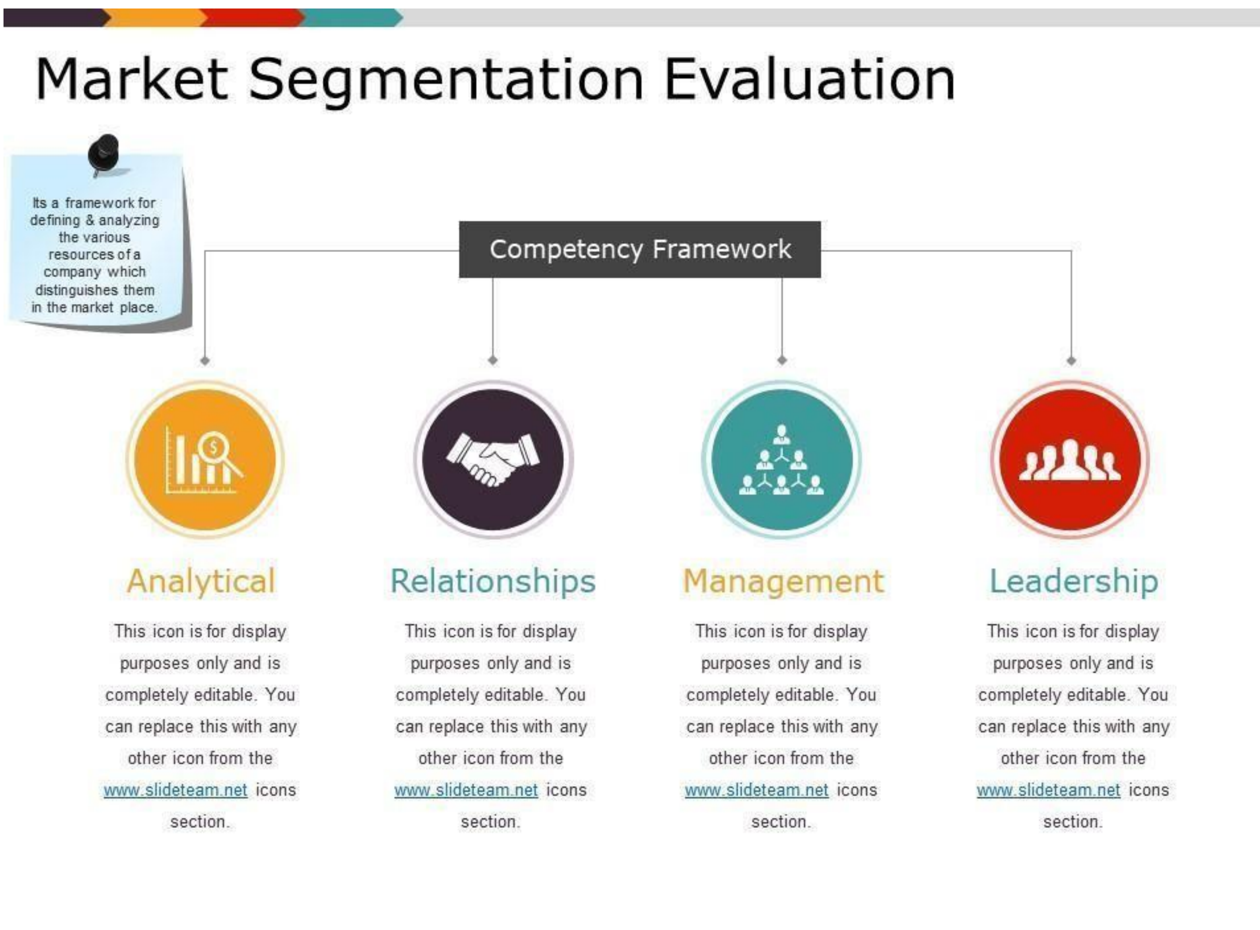
This section focuses on predicting market segments using descriptor variables. It introduces binary logistic regression, multinomial logistic regression, and tree-based methods as techniques to predict the likelihood of an individual belonging to a particular segment based on the descriptor variables.

9.4.1 Binary Logistic Regression:

Binary logistic regression is a statistical technique used to predict the probability of an event occurring. In the context of market segmentation, it can be used to predict which segment an individual is likely to belong to based on specific descriptor variables.

Market Segment Evaluation:

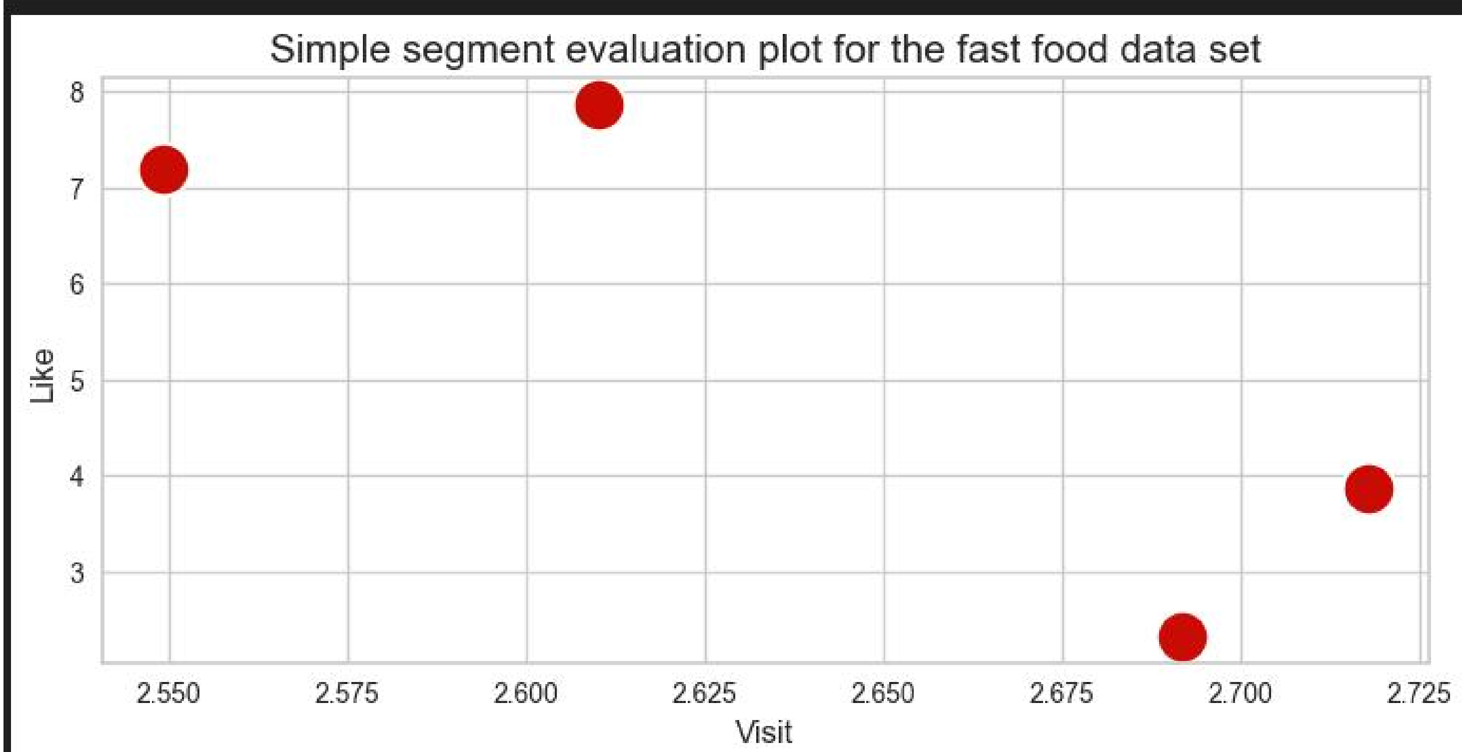
The subsection delves into the evaluation of market segments to assess their suitability as target segments. It outlines various criteria and methods to evaluate segments, including market size, growth potential, competitive intensity, segment accessibility, and compatibility with the company's offerings. By thoroughly evaluating each segment, marketers can identify the most promising opportunities.



```python
#Combining into 1 segment we get
segment = Gender.merge(Like, on='cluster_num', how='left').merge(visit, on='cluster_num', how='left')
segment
```

	cluster_num	Gender	Like	VisitFrequency
0	0	0.579167	2.329167	2.691667
1	1	0.526984	7.180952	2.549206
2	2	0.416667	3.857527	2.717742
3	3	0.389734	7.866920	2.610266

```python
plt.figure(figsize = (9,4))
sns.scatterplot(x = "VisitFrequency", y = "Like",data=segment,s=400, color="r")
plt.title("Simple segment evaluation plot for the fast food data set",
          fontsize = 15)
plt.xlabel("Visit", fontsize = 12)
plt.ylabel("Like", fontsize = 12)
plt.show()
```



STEP 9 – CUSTOMIZING THE MARKETING MIX

In the past, marketing was seen as a toolbox of different strategies toachieve sales results. This toolbox included things like product planning pricing advertising distribution

One common model for the marketing mix is the 4Ps: Product, Price,Promotion, and Place.

Market segmentation is not a standalone strategy but works together withother strategic areas like competition and positioning. The segmentation- targeting-positioning (STP) approach is often used, where segmentation is

the first step, followed by targeting a specific segment, and then positioningthe product in a distinct way.

When selecting a target segment, it is important to customize the marketingmix accordingly. This means adjusting the product, price, promotion, and

place to meet the needs and preferences of the chosen segment. For example, if a segment is interested in cultural activities, a company maydesign a product specifically tailored to their interests, offer relevant promotions, and choose appropriate distribution channels.

Each element of the marketing mix can be influenced by the target segment.

For instance, product design may be modified to better meet customer

needs, pricing decisions can be adjusted based on segment preferences, andpromotional messages can be tailored to resonate with the target segment.

Overall, the content emphasizes the importance of aligning the marketingmix with the chosen target segment to effectively meet their needs and increase the chances of success in the market.

Importance of Bi Clustering the Price aspect of segmentation:

Identify Price Sensitive Segments: Identify those segments which respondto the market in a similar way

1. Customize Pricing Strategies: Make changes in the pricing strategies by knowing/getting the unique patterns of buying of the customers
2. Optimize Pricing Structures
3. Bi Clustering can reveal the customer buying patterns and price preferences
4. Price Positioning

Benefits:

1) Distribution Channel Insight

2) Decision Making-For Direct Sales or Intermediaries

PROMOTION:

In simple terms, the content explains the importance of promotion decisionsin the marketing mix. It discusses the need to develop an advertising message that resonates with the target market and identifies effective waysof communicating this message. Other promotion tools, such as public relations, personal selling, and sponsorship, are also mentioned.

The goal is to determine the best information sources to reach these customers and inform them about the "MUSEUMS, MONUMENTS & MUCH, MUCH MORE" product. This is done by comparing the information sources they used for their last domestic holiday and investigating their preferred TVstations.

To visualize the use of different information sources, a plot is generated using the same command as before, but this time using variables starting with "info". The resulting plot (Fig. 11.4) shows that members of segment 3rely more on information provided by tourist centers when deciding whereto spend their vacation compared to other tourists. This insight can be used to design the promotion component of the marketing mix, such as creatingspecific information packs for the product available in hard copy at local tourist information centers and online on the tourist information center's website.

Additionally, a mosaic plot (Fig. 11.5) is used to display TV channel preferences. This plot helps understand the preferred TV channels ofcustomers in segment 3.

Benefits of Promotion in Marketing Segmentation with respect to the mentioned code:

1. Targeted Advertising: By understanding the preferred information sources and TV channel preferences of customers in segment 3, businesses can tailor their advertising messages to resonate with this specific market segment.
This increases the effectiveness of promotional efforts and enhances the chances of reaching the target audience with the right message.

2. Customized Information Packs: The insight gained from the informationsources analysis allows businesses to create customized information packsfor the "MUSEUMS, MONUMENTS & MUCH, MUCH MORE" product.
Providing these packs both in hard copy at local tourist information centersand online on the tourist information center's website caters to the
preferences of segment 3 customers, ensuring they have access to relevantinformation through their preferred channels.

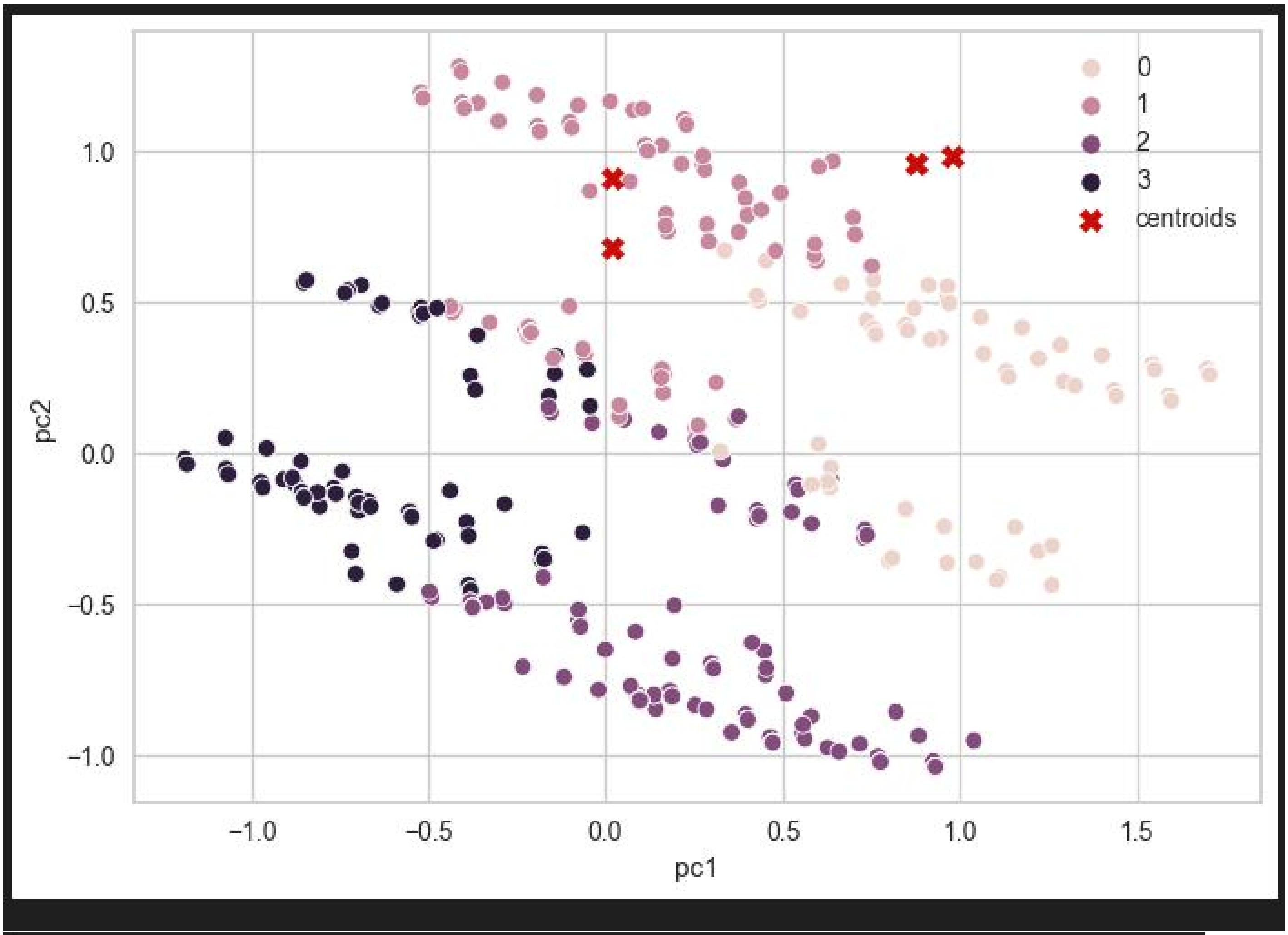
3. Improved Communication Channels: Knowing the preferred TV channelsof customers in segment 3 helps in selecting the most effective communication channels for promotional activities. Businesses can allocateresources towards advertising on these preferred channels, maximizing thereach and impact of their promotional messages.

By leveraging promotion strategies based on market segmentation insights, businesses can effectively communicate their message, engage with the target audience, and increase the likelihood of customer engagement,

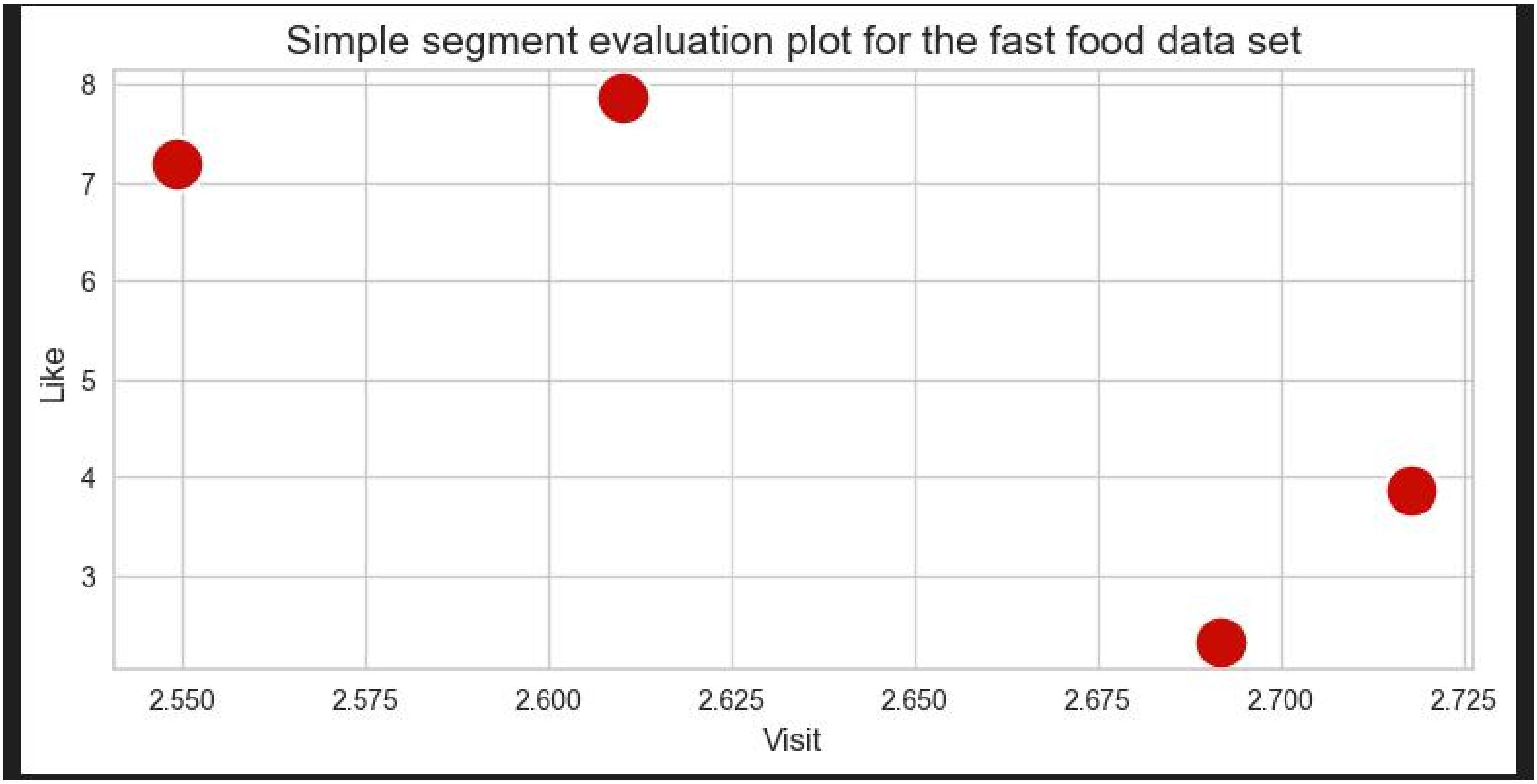
ultimately leading to higher conversions and sales.

Code Used:

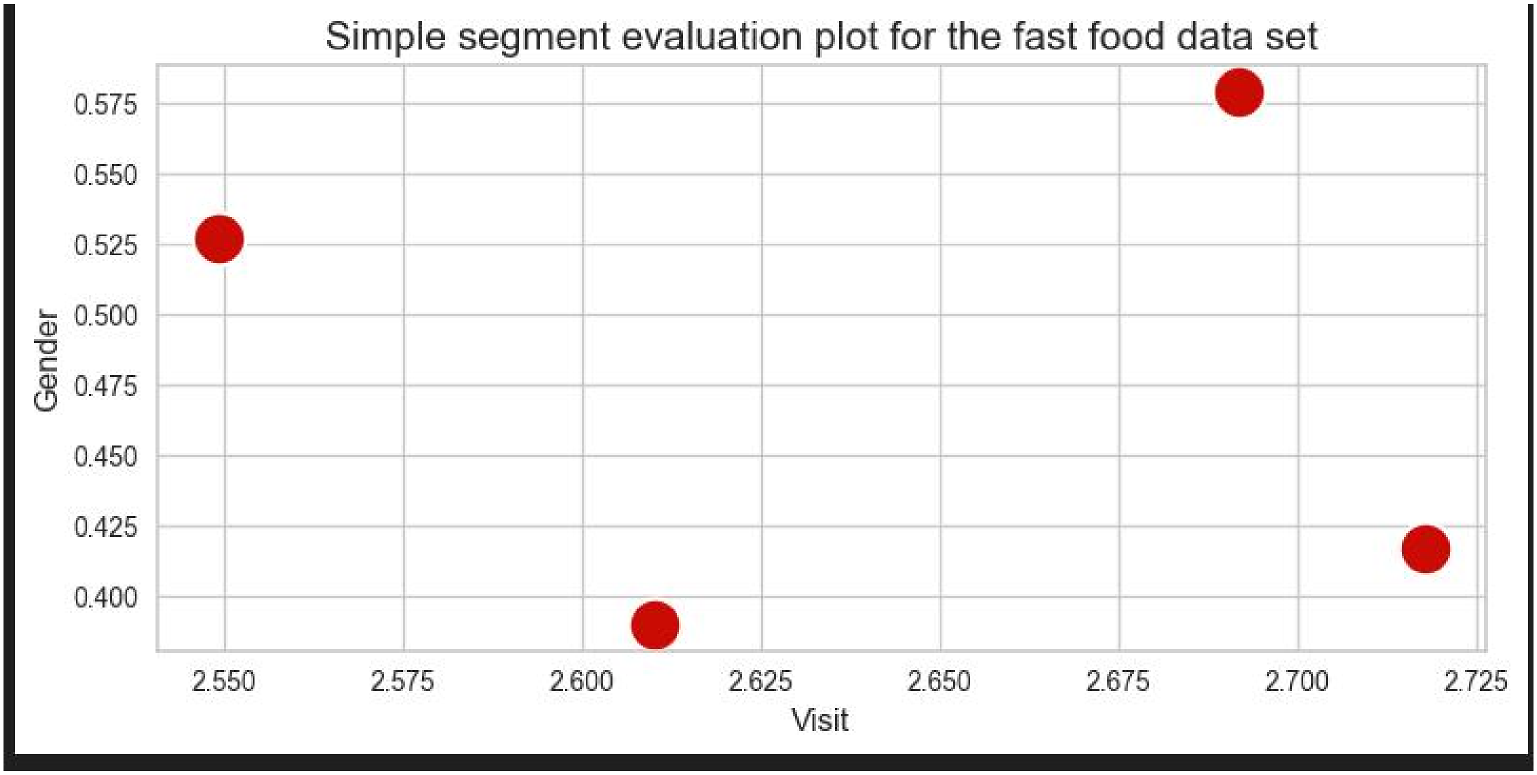
```
Cluster Visualization:


sns.scatterplot(data=pf, x="pc1", y="pc2", hue=kmeans.labels_)
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1],
            marker="X", c="r", s=80, label="centroids")
plt.legend()
plt.show()
```




```
plt.figure(figsize = (9,4))
sns.scatterplot(x = "VisitFrequency", y = "Like",data=segment,s=400, color="r")
plt.title("Simple segment evaluation plot for the fast food data set",
          fontsize = 15)
plt.xlabel("Visit", fontsize = 12)
plt.ylabel("Like", fontsize = 12)
plt.show()
```




```
plt.figure(figsize = (9,4))
sns.scatterplot(x = "VisitFrequency", y = "Gender",data=segment,s=400, color="r")
plt.title("Simple segment evaluation plot for the fast food data set",
          fontsize = 15)
plt.xlabel("Visit", fontsize = 12)
plt.ylabel("Gender", fontsize = 12)
plt.show()
```



Market Segmentation Case Study on McDonalds Dataset