Introduction

This document outlines a market segmentation analysis of McDonald's utilizing advanced data techniques. By employing Principal Component Analysis (PCA), K-means clustering, Gaussian Mixture Models (GMM), and regression analysis, we aim to reveal distinct customer segments and their preferences. The study examines crucial factors affecting consumer perceptions and provides actionable insights to refine McDonald's marketing strategies and product offerings. The objective is to gain a deeper understanding of diverse consumer needs, thereby enhancing McDonald's market position.

Here's how the general approach to developing and implementing a market segmentation strategy is applied to the McDonald's case study:

Step 1: Deciding (not) to Segment

Commitment: Organizations must recognize that market segmentation is a long-term commitment that requires the willingness to make substantial changes across the organization.

Costs vs. Benefits: It is crucial to evaluate whether the benefits of segmentation, such as increased sales and market share, outweigh the costs associated with research, development, and implementation of different strategies for different segments.

Organizational Structure: The organization must be prepared to reorganize around market segments rather than products to maximize the benefits of segmentation.

Senior Management Involvement: The decision to pursue market segmentation should be made at the highest level of management and should be communicated clearly across all organizational units.

Here, McDonald's must evaluate whether the benefits of understanding and catering to distinct consumer segments justify the long-term commitment required. This involves assessing the potential for improved customer satisfaction, enhanced brand loyalty, and increased market share against the costs associated with researching and implementing targeted strategies. Adjustments may be needed in product offerings, marketing approaches, and operational structures to effectively serve different customer segments.

Additionally, McDonald's should consider the impact on its organizational structure. This might involve reorganizing business units to focus on specific customer needs, such as health-conscious consumers or those prioritizing convenience and affordability. The decision to pursue such a segmentation strategy should be endorsed by senior management to ensure alignment across the organization and to secure the necessary resources for successful implementation.

Step 2: Specifying the Ideal Target Segment

Knock-Out Criteria: Before identifying market segments, the organization should establish essential criteria such as homogeneity, distinctness, size, and alignment with the organization's strengths. Segments must also be identifiable and reachable.

Attractiveness Criteria: These criteria evaluate the potential of a segment, considering factors like profitability, growth potential, and alignment with the organization's goals. The segmentation team should agree on a subset of criteria that will guide the segment selection process.

McDonald's should focus on identifying consumer segments that are both homogeneous and distinct, ensuring they align with the company's strengths in fast service and affordability. These segments should be large enough to justify targeted marketing efforts and easily identifiable through methods such as surveys or loyalty programs. Additionally, the segments must be reachable via targeted advertising to effectively engage and cater to their specific needs.

In assessing segment attractiveness, McDonald's may prioritize those with a positive perception of the brand or those who frequently dine out. However, it is also valuable to consider segments with a negative perception, as addressing their concerns could potentially convert them into loyal customers. By understanding and addressing the needs of all relevant segments, McDonald's can enhance its market position and foster stronger customer relationships.

Step 3: Collecting Data

Data Quality: The quality of the empirical data is critical for developing valid segmentation solutions. The data should be relevant, unbiased, and representative of the target market.

Data Sources: Data can be collected from various sources, including surveys, observations, or experimental studies. The choice of data source should reflect actual consumer behavior as closely as possible.

Variables: The segmentation variables should be carefully selected to ensure they capture the relevant aspects of consumer behavior and preferences. Redundant or irrelevant variables should be avoided to prevent noise in the segmentation process.

The dataset comprises responses from 1,453 adult Australian consumers, focusing on their perceptions of McDonald's across several attributes: YUMMY, CONVENIENT, SPICY, FATTENING, GREASY, FAST, CHEAP, TASTY, EXPENSIVE, HEALTHY, and DISGUSTING. These attributes were identified through a qualitative study preceding the survey. Respondents indicated whether they believed McDonald's possessed each attribute with a YES or NO response. Additionally, demographic information such as AGE and GENDER was collected to aid in segment characterization.

For a more robust market segmentation analysis, it would be beneficial to include supplementary data such as dining frequency, spending patterns, and information channel usage. This additional behavioral data would provide a richer and more nuanced understanding of consumer segments, allowing McDonald's to refine its strategies and better align its offerings with diverse customer needs. Ensuring high-quality, unbiased data collection is crucial for accurately reflecting consumer attitudes and enhancing the effectiveness of segmentation efforts.

Step 4: Exploring Data

The initial exploration of the dataset involves examining its fundamental characteristics to understand its structure and content. This process starts by loading the dataset into a data analysis environment, such as Python's Pandas library. Once loaded, the following aspects are typically inspected:

Variable Names: Review the dataset's columns to identify the variables present. This includes both the attributes related to consumer perceptions (e.g., YUMMY, CONVENIENT) and demographic variables (e.g., AGE, GENDER). Understanding the names and types of variables helps in determining which ones are categorical, binary, or numerical, and how they can be utilized in the analysis.

Sample Size: Determine the total number of observations or rows in the dataset. This information is crucial for assessing the dataset's representativeness and ensuring that there are enough data points for reliable analysis. In this case, the dataset includes responses from 1,453 consumers.

First Three Rows: Inspecting the first few rows of the data provides a snapshot of how the information is structured. It allows for a preliminary check on data consistency, format, and whether the values align with expected responses (e.g., YES or NO for attribute perceptions, numeric values for age). This step helps identify any immediate issues such as missing or incorrectly formatted data.

Column Names: ['yummy', 'convenient', 'spicy', 'fattening', 'greasy', 'fast', 'cheap', 'tasty', 'expensive', 'healthy', 'disgusting', 'Like', 'Age', 'VisitFrequency', 'Gender']

Dimensions: (1453, 15)

First Three Rows:

	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	\
0	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No	
1	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	
2	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	

	disgusting	Like	Age	Visit Frequency	Gender
0	No	-3	61	Every three months	Female
1	No	+2	51	Every three months	Female
2	No	+1	62	Every three months	Female

Column Means:

yummy	0.55
convenient	0.91
spicy	0.09
fattening	0.87
greasy	0.53
fast	0.90
cheap	0.60
tasty	0.64
expensive	0.36
healthy	0.20
disgusting	0.24
d type:	float64

The calculated column means indicate the average preference levels across the customer base. Higher means suggest more positive responses, while lower means indicate less favorable views toward certain aspects of McDonald's offerings.

By performing these preliminary checks, we gain a foundational understanding of the dataset, enabling more informed decisions on subsequent data cleaning, transformation, and analysis steps.

PCA is conducted to reduce the dimensionality of the dataset and identify the most significant factors explaining customer preferences.

The PCA is applied to the binary data, transforming it into principal components that capture the maximum variance in the dataset.

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	0.7570	0.6075	0.5046	0.3988	0.33741
Proportion of Variance	0.2994	0.1928	0.1330	0.0831	0.05948
Cumulative Proportion	0.2994	0.4922	0.6253	0.7084	0.76787

	PC6	PC7	PC8	PC9
Standard deviation	0.3103	0.28970	0.27512	0.26525
Proportion of Variance	0.0503	0.04385	0.03955	0.03676
Cumulative Proportion	0.8182	0.86201	0.90156	0.93832

	PC10	PC11
Standard deviation	0.24884	0.23690
Proportion of Variance	0.03235	0.02932
Cumulative Proportion	0.97068	1.00000

The first few principal components typically explain a large portion of the variance, meaning they represent the most influential factors in customer preferences. The explained variance ratio and cumulative variance help determine how many components are necessary to retain most of the information.

The dataset reveals that the first two principal components (PC1 and PC2) capture nearly 50% of the total variance, making them the most important for summarizing the data. The first five components together explain about 76.8% of the variance, indicating they capture most of the meaningful patterns in the dataset. This dimensionality reduction allows for a focused analysis of key components, simplifying the dataset while retaining the majority of its information.

The PCA-transformed data is clustered into four segments, where each segment represents a group of customers with similar preferences.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
yummy	-0.477	0.364	-0.304	0.055	-0.308	0.171	-0.281	0.013	0.572	-0.110	0.045
convenient	-0.155	0.016	-0.063	-0.142	0.278	-0.348	-0.060	-0.113	-0.018	-0.666	-0.542
spicy	-0.006	0.019	-0.037	0.198	0.071	-0.355	0.708	0.376	0.400	-0.076	0.142
fattening	0.116	-0.034	-0.322	-0.354	-0.073	-0.407	-0.386	0.590	-0.161	-0.005	0.251
greasy	0.304	-0.064	-0.802	0.254	0.361	0.209	0.036	-0.138	-0.003	0.009	0.002
fast	-0.108	-0.087	-0.065	-0.097	0.108	-0.595	-0.087	-0.628	0.166	0.240	0.339
cheap	-0.337	-0.611	-0.149	0.119	-0.129	-0.103	-0.040	0.140	0.076	0.428	-0.489
tasty	-0.472	0.307	-0.287	-0.003	-0.211	-0.077	0.360	-0.073	-0.639	0.079	0.020
expensive	0.329	0.601	0.024	0.068	-0.003	-0.261	-0.068	0.030	0.067	0.454	-0.490
healthy	-0.214	0.077	0.192	0.763	0.288	-0.178	-0.350	0.176	-0.186	-0.038	0.158
disgusting	0.375	-0.140	-0.089	0.370	-0.729	-0.211	-0.027	-0.167	-0.072	-0.290	-0.041

The factor loadings derived from the Principal Component Analysis (PCA) of the McDonald's dataset highlight the key dimensions that explain variations in customer perceptions. The first principal component (PC1) primarily captures taste preferences, with "yummy" and "tasty" showing strong negative contributions. This suggests that PC1 represents overall taste preferences. Additionally, the attributes "cheap" and "expensive" have opposing contributions to this component, indicating that price perceptions are also a significant factor in shaping this dimension.

The second principal component (PC2) reflects a dichotomy between affordability and expense, with "cheap" and "expensive" exhibiting opposite loadings. This component also includes positive contributions from taste-related attributes such as "yummy" and "tasty," suggesting that both taste and price perceptions influence this dimension. The interaction between these factors points to a combined effect of taste and price on customer perceptions.

The third principal component (PC3) is associated with health-related perceptions, notably the greasiness of the food. This is evident from the strong negative loading of "greasy" and additional contributions from "fattening." As we examine the fourth and fifth components, health perceptions ("healthy") and negative food perceptions ("disgusting") become more prominent, capturing more specific attitudes towards the food. The remaining components (PC6 to PC11) explain less variance, indicating they capture subtler patterns within the data. Overall, PCA results reveal that taste, price, and health-related factors are the primary drivers of customer perceptions, providing a comprehensive view of the dataset's underlying structure.

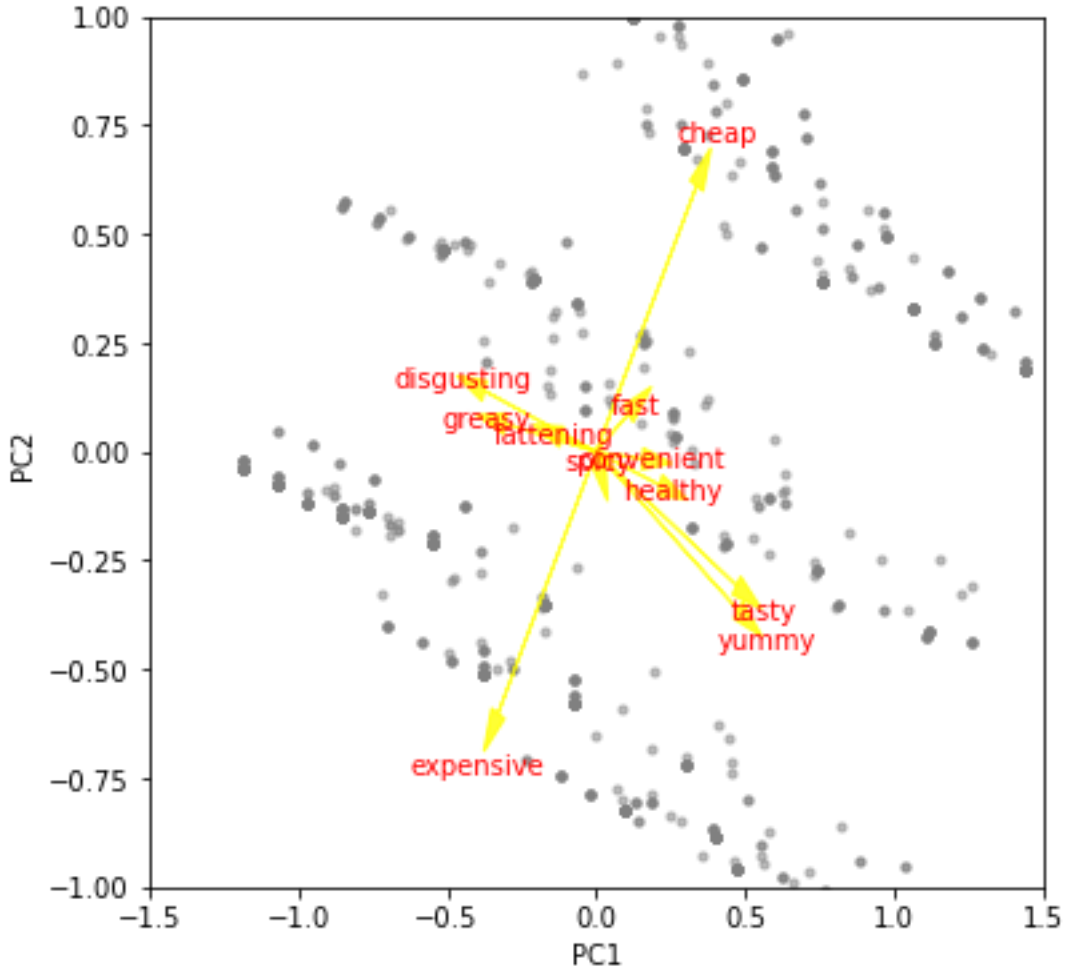

Figure 1:Perpetual map

The perceptual map in Fig .1 generated from the Principal Component Analysis (PCA) of the dataset visualizes how various attributes related to McDonald's food offerings are perceived by respondents. The arrows represent different attributes, and their directions and lengths indicate the influence of these attributes on the first two principal components (PC1 and PC2).

The attributes "CHEAP" and "EXPENSIVE" are positioned in opposite directions, suggesting that they are key, yet independent, dimensions in consumer evaluations of McDonald's. Respondents who consider McDonald's food to be "CHEAP" are likely to disagree that it is "EXPENSIVE," and vice versa. This price dimension is distinct and plays a crucial role in consumer perception.

On the other hand, the attributes "FATTENING," "DISGUSTING," and "GREASY" cluster together, pointing in the same direction. This alignment indicates that respondents who view McDonald's food as "FATTENING" are also likely to perceive it as "DISGUSTING" and "GREASY." These negative perceptions form a coherent group in the perceptual map, suggesting that they are closely related in the minds of the respondents.

Conversely, attributes like "FAST," "CONVENIENT," "HEALTHY," "TASTY," and "YUMMY" are positioned in the opposite direction to the negative attributes. This implies that these attributes are seen as positive and are likely associated with more favorable views of McDonald's. For instance, respondents who find McDonald's food "TASTY" are also likely to consider it "YUMMY," "FAST," and "CONVENIENT."

Overall, the perceptual map provides a valuable visual summary of how different attributes are related to one another and how they contribute to consumer perceptions of McDonald's. It reveals that price is a significant differentiator, while positive and negative food-related attributes also play a crucial role in shaping consumer opinions. These insights are essential for further segmentation and strategic marketing efforts.

STEP 5. Extracting Segments

In Step 5, we focus on the extraction of market segments using various methodologies. This step is crucial for dividing consumers into distinct groups based on their similarities. We will explore three different approaches to segment extraction. First, we will employ the standard **k-means clustering technique**, which is widely used for its simplicity and effectiveness in partitioning data into k segments.

Next, we will delve into **finite mixtures of binary distributions**, a method that models the data as a combination of several distributions to capture more complex relationships. Finally, we will apply **finite mixtures of regressions**, which allows for the modeling of segments based on varying relationships between independent and dependent variables across different segments. Each of these methods offers unique insights and helps in identifying distinct consumer groups.

K-Means

K-means clustering is a popular partitioning method used to divide a set of observations into k subsets, called clusters or segments, where each observation belongs to the cluster with the nearest mean (centroid). The objective of k-means clustering is to minimize the sum of squared distances between each observation and the centroid of its assigned cluster. The process

involves initializing k centroids, assigning each observation to the nearest centroid, recalculating the centroids, and repeating this process until the clusters stabilize.

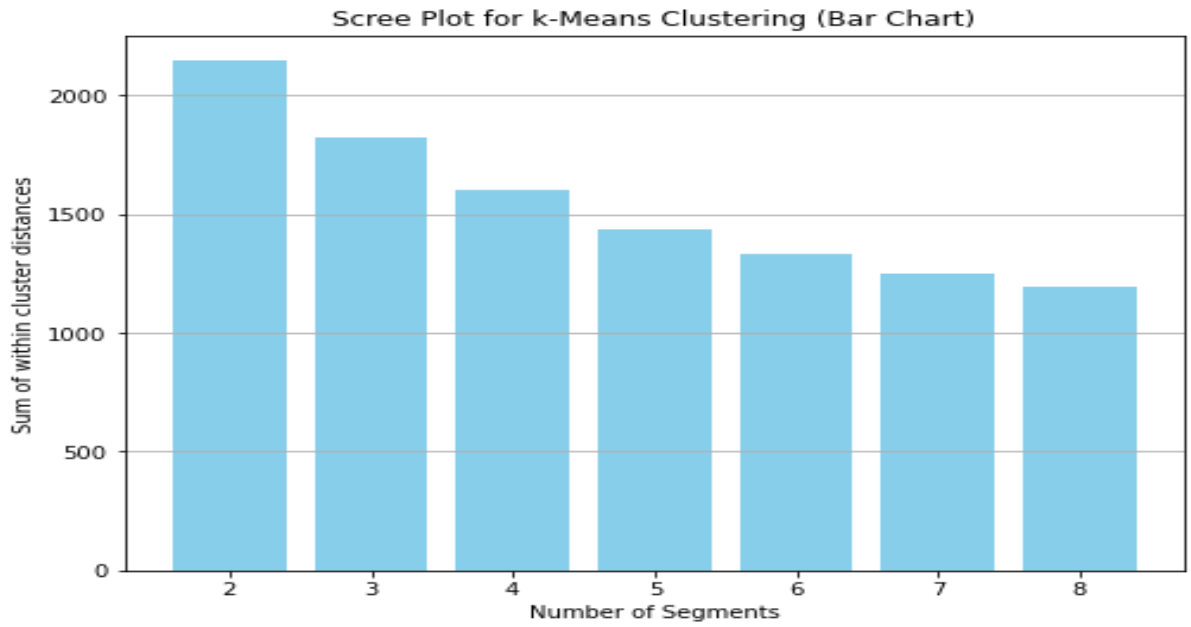

Figure 2:Scree plot

The provided plot in Fig.2 is a scree plot that displays the sum of squared distances within market segments for different numbers of segments, ranging from two to eight. This type of plot is often used in k-means clustering to help determine the optimal number of clusters (segments) by looking for an "elbow," or a point where the sum of squared distances drops off sharply, indicating diminishing returns from adding more segments.

However, in this particular scree plot, there is no distinct elbow visible. The sum of distances within market segments decreases gradually as the number of segments increases. This behavior is expected because increasing the number of segments typically results in smaller, more homogenous segments, where members are more similar to each other. The gradual decrease suggests that adding more segments continues to slightly reduce the sum of squared distances but does not result in a dramatic improvement after a certain point.

The lack of a clear elbow in the scree plot implies that the scree plot alone does not provide a strong indication of the optimal number of segments for the McDonald’s dataset. This suggests that choosing the best number of market segments might require additional analysis beyond the scree plot.

Given that the scree plot does not provide a definitive answer, a second approach that could be considered is stability-based data structure analysis. This method evaluates the stability of segmentation solutions across multiple replications. Stability analysis helps ensure that the

chosen market segments are not only optimal but also reproducible and reliable. This is crucial because a segmentation solution that is not stable across different runs may not be trustworthy for guiding significant business decisions, such as those McDonald's might make in its market segmentation strategy.

In summary, while the scree plot shows the expected decrease in within-segment distances as more segments are added, it does not reveal a clear optimal number of segments. Therefore, stability-based analysis might be necessary to determine the best segmentation solution for the McDonald's dataset.

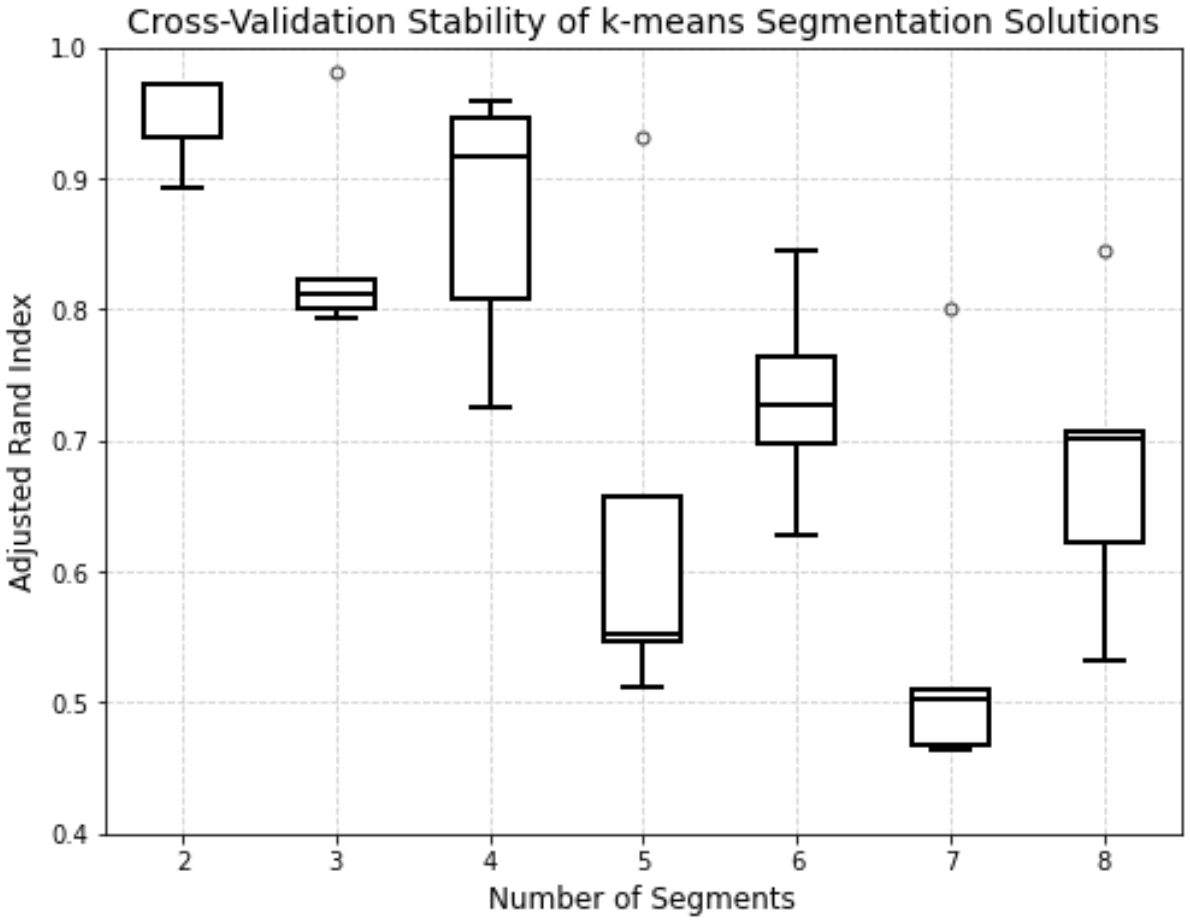

Figure 3:Segment stability

The provided plot in Fig.3 is a boxplot illustrating the global stability of k-means segmentation solutions for different numbers of segments, ranging from two to eight, specifically applied to the McDonald's dataset. This dataset likely includes consumer perceptions and attributes associated with McDonald's products and services.

Global stability measures how consistently the same segmentation solution emerges when the analysis is repeated multiple times using bootstrap samples of the data. The stability is quantified using the adjusted Rand index, a measure of agreement between different clustering results, with higher values indicating more stable and consistent solutions.

In the boxplot, each vertical box represents the distribution of the adjusted Rand index for a specific number of segments. The thick horizontal line inside each box indicates the median stability for that number of segments, while the upper and lower edges of the box represent the interquartile range (the middle 50% of the data). Whiskers extend to the most extreme data points that are not considered outliers, and individual outliers are shown as circles.

When analyzing the McDonald's data, we see that the two-, three-, and four-segment solutions exhibit relatively high and consistent stability, as indicated by the higher median values and more compact boxplots. This suggests that these segmentations are reliably reproducible when the analysis is repeated multiple times with different bootstrap samples.

However, despite the stability of the two- and three-segment solutions, they may not provide a sufficiently detailed view of the diverse consumer perceptions present in the McDonald's dataset. Solutions with fewer segments often lack the granularity necessary for gaining actionable insights into consumer behavior. As the number of segments increases to five and beyond, the stability decreases noticeably, with the median dropping and the spread of the boxplot increasing. This indicates that the solutions become less reliable and more sensitive to the particular sample used in the analysis.

Given these results, the four-segment solution emerges as the most balanced option for the McDonald's dataset. It offers a detailed enough view of the market while maintaining reasonable stability across different bootstrap samples. This makes the four-segment solution a strong candidate for a robust and actionable market segmentation strategy, potentially revealing distinct consumer groups with varying perceptions and behaviors toward McDonald's products and services.

The plot in Fig.4 titled "Segment Level Stability Within Solutions for {clusters} Clusters" provides insight into how stable each of the identified segments is within a specific clustering solution. The stability of a segment refers to how consistently the same consumers are grouped when the clustering process is repeated multiple times using different random samples or initial conditions.

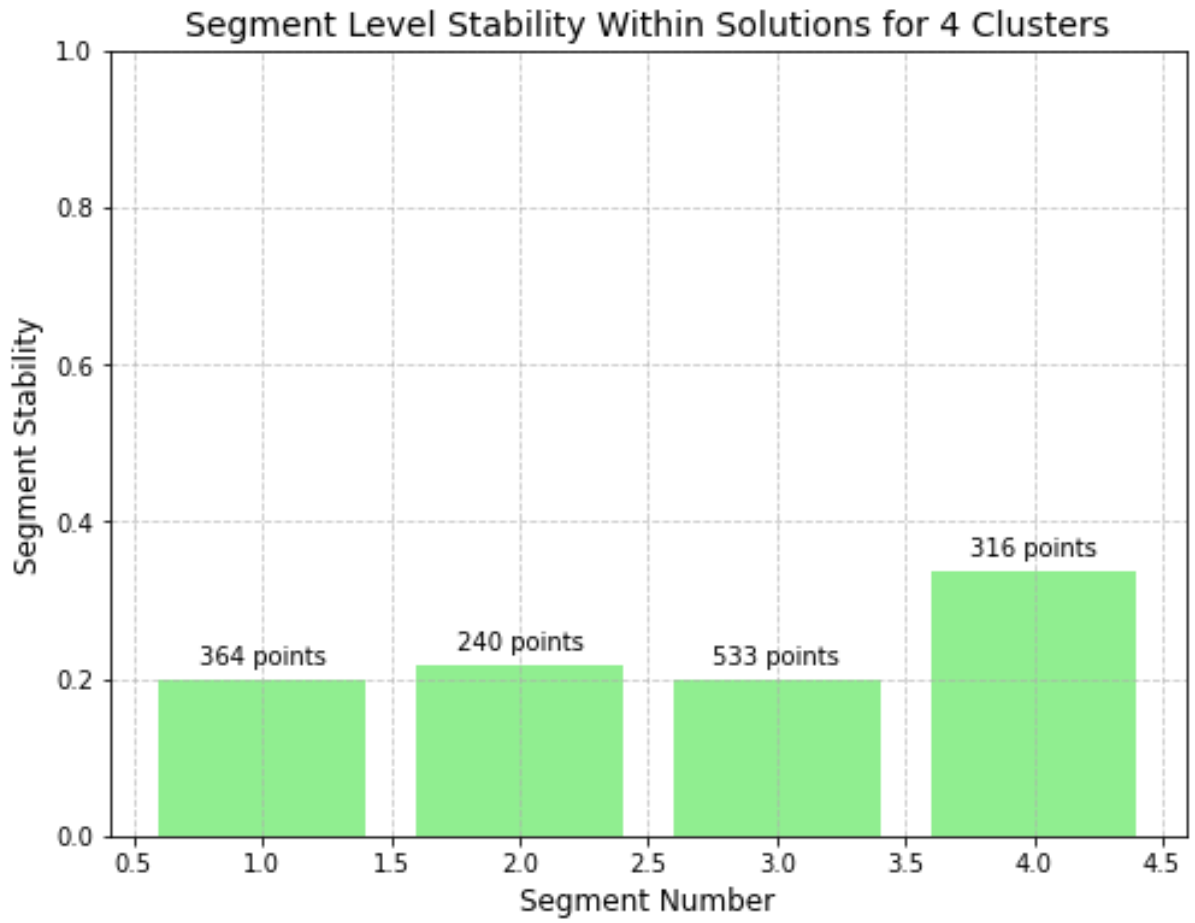

Figure 4:Segment stability in clusters

Segment 1 (364 points): This segment shows a moderate size with 364 points. Its stability would be interpreted by seeing how consistently these 364 consumers were grouped across different iterations of the clustering process. If Segment 1 has high stability, it suggests that these consumers share strong similarities, making this segment a reliable group for McDonald's to target in marketing efforts.

Segment 2 (240 points): Being the smallest segment with 240 points, its stability is crucial to understand. If the stability is low, it may indicate that this segment is less distinct, meaning that the consumers in this group could easily be assigned to other segments in different iterations. Conversely, high stability would indicate that this smaller segment is well-defined and consistent, making it a meaningful target group.

Segment 3 (533 points): As the largest segment, Segment 3's stability is important for understanding how cohesive this group is. If it shows high stability, this segment likely represents a core group of McDonald's customers with common preferences or behaviors. Low

stability, on the other hand, might suggest that this group is more heterogeneous and could potentially be divided into smaller, more distinct segments.

Segment 4 (316 points): This segment has a moderate number of points, and its stability indicates how well-defined it is. A stable Segment 4 would imply that these consumers consistently share similar characteristics across different clustering iterations, making it a reliable target for specific marketing or product strategies.

In summary, this plot helps in understanding how consistently each of these four segments appears across different clustering scenarios. High stability across segments implies that the segmentation is robust, with clearly defined groups of consumers. This can be critical for McDonald's in making decisions about which segments to target for different marketing strategies, ensuring that their efforts are directed towards well-defined and consistent customer groups.

Mixtures of Distributions

To begin the latent class analysis using a finite mixture of binary distributions, we fit Gaussian Mixture Models (GMMs) for different numbers of segments (k=2 to k=8). Unlike k-means clustering, which minimizes squared Euclidean distances, the mixture model maximizes the likelihood of identifying distinct segments.

Here’s a summary of the process:

1. Model Fitting: We fit Gaussian Mixture Models (GMMs) for values of k ranging from 2 to 8, using ten random restarts of the Expectation-Maximization (EM) algorithm to ensure robust results. Each GMM is configured with a segment-specific model of independent binary distributions.

2. Model Evaluation: We evaluate each model using several criteria:

Log-Likelihood (logLik): Measures the fit of the model to the data. Higher values indicate a better fit.

Akaike Information Criterion (AIC): Balances model fit and complexity, with lower values indicating a more optimal model.

Bayesian Information Criterion (BIC): Similar to AIC but includes a greater penalty for model complexity, also favoring lower values.

Integrated Completed Likelihood (ICL): Combines BIC with a measure of uncertainty in cluster assignments, aiming to maximize the separation between segments.

The plot in Fig.5 illustrates the evaluation of different segmentations for the McDonald's dataset using three different information criteria: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Integrated Completed Likelihood (ICL). These criteria help in determining the optimal number of segments (clusters) for a finite mixture model, where the goal is to identify the best number of segments that adequately describe the data without overfitting.

k=2, converged=True, logLik=11147.82, AIC=-21985.64, BIC=-21167.03, ICL=-21167.03

k=3, converged=True, logLik=14126.45, AIC=-27786.90, BIC=-26556.34, ICL=-26556.34

k=4, converged=True, logLik=17513.54, AIC=-34405.08, BIC=-32762.57, ICL=-32763.13

k=5, converged=True, logLik=22500.93, AIC=-44223.86, BIC=-42169.40, ICL=-42169.43

k=6, converged=True, logLik=28522.04, AIC=-56110.08, BIC=-53643.68, ICL=-53643.68

k=7, converged=True, logLik=29652.48, AIC=-58214.95, BIC=-55336.60, ICL=-55336.65

k=8, converged=True, logLik=26714.81, AIC=-52183.62, BIC=-48893.32, ICL=-48893.65

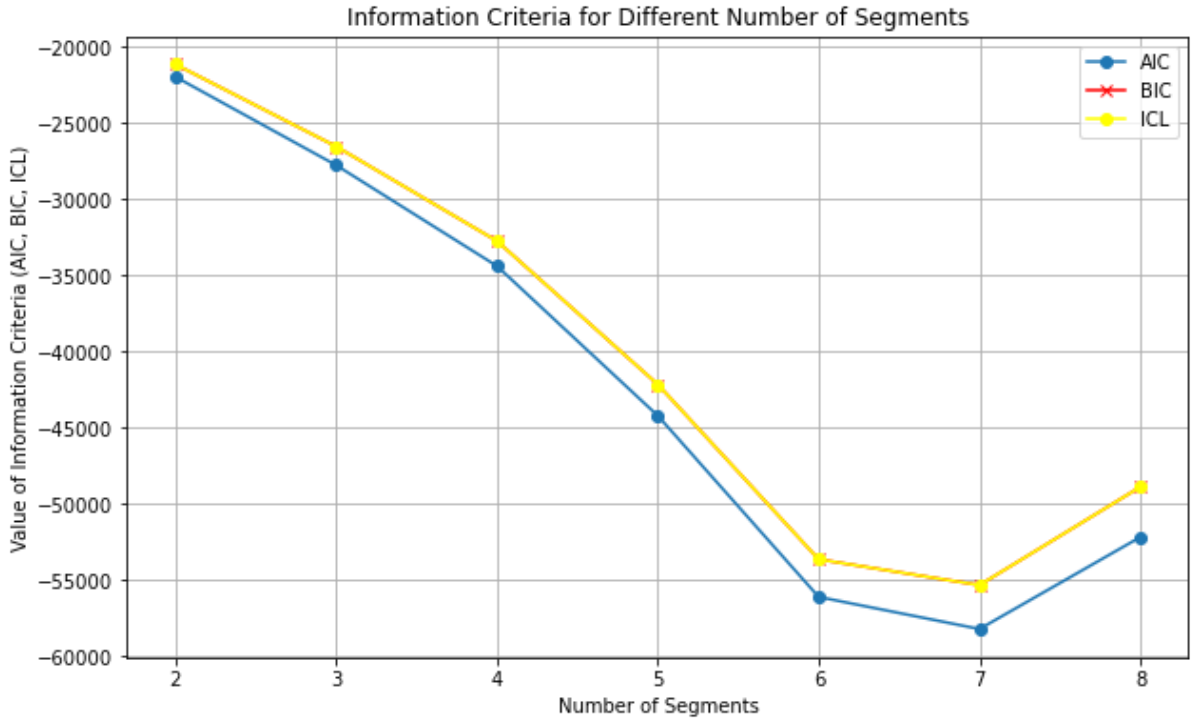

Figure 5:Information criteria for different segments

1. AIC, BIC, and ICL Trends:

All three criteria (AIC, BIC, and ICL) generally decrease as the number of segments increases from 2 to 6. This decrease indicates that adding more segments initially helps better fit the model to the data, improving the likelihood of correctly identifying the underlying segments.

However, after 6 segments, there is a notable change in the trends. For 7 segments, AIC continues to decrease slightly, while BIC and ICL start to plateau or increase, indicating that adding more segments beyond 6 does not provide significant improvements and may introduce unnecessary complexity.

2. Optimal Number of Segments:

The optimal number of segments can be inferred where the BIC and ICL are minimized, which often provides a balance between model fit and simplicity. In this case, the 6-segment solution appears optimal because it achieves the lowest BIC and ICL values before they start to increase or flatten with 7 or 8 segments.

The 6-segment solution also corresponds to the lowest AIC value, indicating that this model has the best fit with the least complexity relative to the other solutions.

The McDonald's dataset likely contains complex consumer behavior patterns that require a nuanced segmentation approach. The 6-segment solution seems to capture the diversity in consumer preferences most effectively, offering a detailed but not overly complex segmentation.

Segments could reflect different consumer groups such as budget-conscious consumers (those focused on the "cheap" attribute), health-conscious consumers (interested in "healthy" options), and others driven by factors like the taste ("yummy", "tasty") or brand perception ("expensive", "disgusting").

Overall, the plot and results suggest that the 6-segment solution provides the most balanced and informative segmentation of the McDonald's dataset, making it the preferred choice for further analysis and strategic decision-making.

The cross-tabulation presented in Fig.6 shows a comparison between the clustering results of two different methods, possibly K-means and a mixture model approach, applied to the McDonald's dataset. The table provides a matrix where the rows represent clusters identified by one method (likely K-means) and the columns represent clusters identified by the other

method (likely the mixture model). The values in each cell indicate the number of data points that were assigned to the corresponding clusters by both methods.

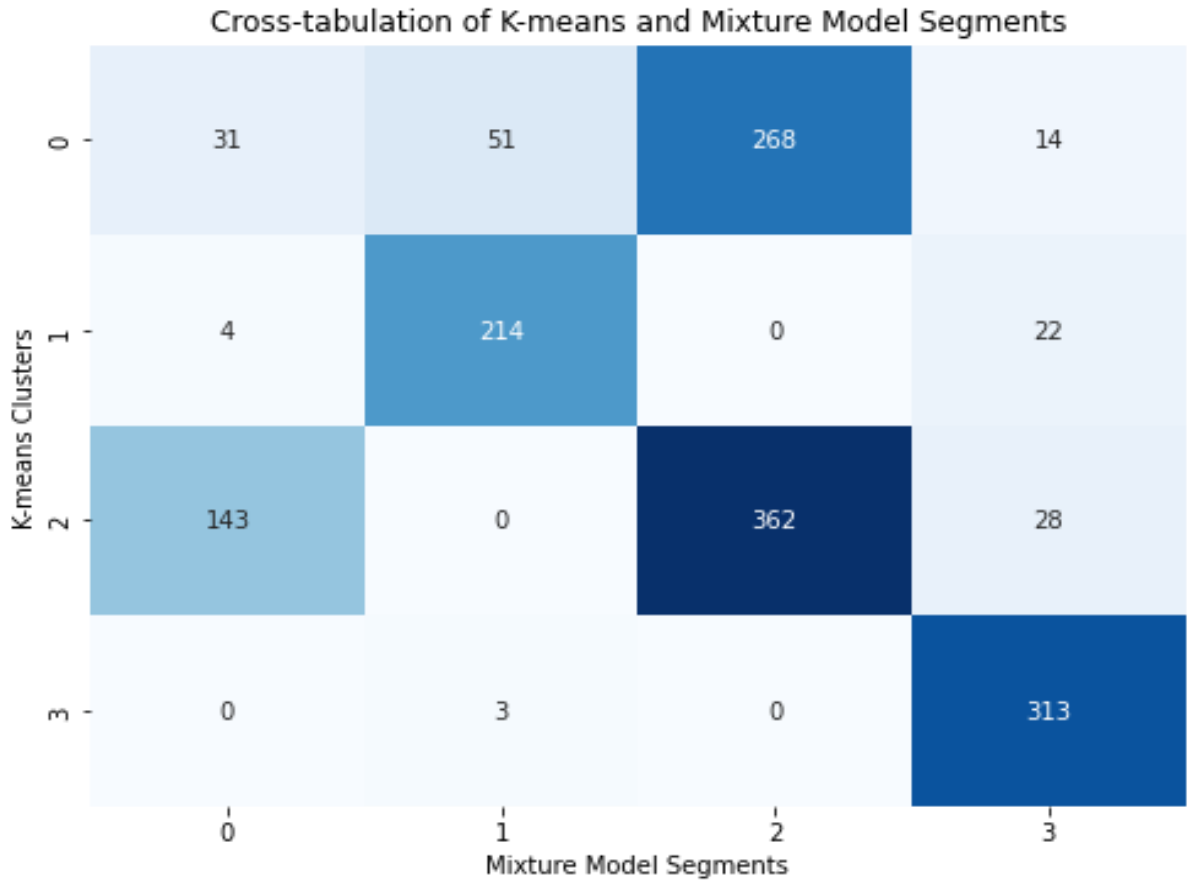

Figure 6:Cross-tabulation of Kmeans-GMM

The top left cell (31) indicates that 31 data points were assigned to the first cluster by both methods.

The cell with 51 data points suggests these were assigned to the first cluster by the first method but to the second cluster by the second method.

The largest number, 362, shows a strong agreement between both methods in assigning these data points to the corresponding clusters.

The variation in these values reflects how closely the two methods agree on the segmentation of the dataset. A high number of points in diagonal cells (where the row and column indices are equal) suggests a strong agreement between the methods for that particular cluster. Conversely, a spread of values of the diagonal indicates some differences in how the methods segmented the data.

When two completely different methods of initializing the mixture model (such as ten random restarts and using the K-means solution) yield almost the same result, it indicates that the solution is likely close to a global optimum. This convergence between methods reinforces

confidence in the robustness of the segmentation, suggesting that the results are reliable and consistent across different methodologies. While slight differences between the solutions are not problematic, the fact that both methods produce similar segmentations implies that the identified segments are meaningful and can be trusted for managerial decisions. Neither method is considered definitively correct or incorrect; instead, both provide valuable insights that can be used in combination to inform strategy.

Mixtures of Regression Models

To enhance McDonald's understanding of customer preferences, we can explore market segments based on how similar perceptions drive either strong affection or aversion towards the brand. This approach focuses on identifying segments where customer feelings of love or hate are influenced similarly by specific perceptions of McDonald's. By targeting these segments, McDonald's can tailor its strategies to modify perceptions in a way that boosts positive feelings and reduces negative ones.

To implement this, we use latent class regressions, which involve fitting finite mixtures of linear regression models. In this setup, the dependent variable \(y \) represents the degree to which consumers love or hate McDonald’s, measured on an 11-point scale ranging from "I LOVE IT!" to "I HATE IT!" The independent variables \(x \) are the various perceptions of McDonald’s.

Like

I love it!+5 143

I hate it!-5 152

0 169

-4 71

-3 73

-2 59

-1 58

+4 160

+3 229

+2 187

+1 152

Name: count, dtype: int64

Like.n

-4.0 71

-3.0 73

-2.0 59

-1.0 58

0.0 169

1.0 152

2.0 187

3.0 229

4.0 160

Name: count, dtype: int64

We first convert the ordinal dependent variable into a numeric format to fit these models. For instance, we translate the scale where "I LOVE IT!" is coded as +5 and "I HATE IT!" as -5, using the formula \(6 - \text{numeric code} \). This transformation provides a numeric variable that captures the degree of affection or aversion.

Like.n ~ yummy + convenient + spicy + fattening + greasy + fast + cheap + tasty + expensive + healthy + disgusting

Subsequently, we construct a regression model where the perception variables are included as predictors. This can be done manually by listing the variables or automatically in Python by creating a formula that includes all perception variables. This approach enables us to identify segments of customers who share similar patterns in how their liking or disliking of McDonald's is influenced by their perceptions.

The analysis of the dataset provides a detailed understanding of how various perceptions influence consumers' liking of McDonald’s. We begin by assessing the relationship between the independent variables (such as “yummy,” “convenient,” “spicy,” etc.) and the dependent variable (`Like_n`), which measures the degree to which consumers like or dislike McDonald’s. The approach taken includes calculating the Variable Inflation Factor (VIF), performing Principal Component Analysis (PCA), applying K-Means clustering, and conducting a Gaussian Mixture Model (GMM) followed by regression analysis for each identified cluster.

1. Variable Inflation Factor (VIF) and Descriptive Statistics:

	Variable	VIF
0	yummy Yes	0.006037
1	convenient Yes	0.012370
2	spicy Yes	0.010404
3	fattening Yes	0.008472
4	greasy Yes	0.004125
5	fast Yes	0.009787

6	cheap	Yes	0.007521
7	tasty	Yes	0.006722
8	expensive	Yes	0.007745
9	healthy	Yes	0.006547
10	disgusting	Yes	0.007120

The VIF calculation reveals that multicollinearity is not a concern within this dataset, as all VIF values are well below 10. This ensures that each independent variable contributes unique information regarding consumer perceptions. The descriptive statistics for `Like_n` show that on average, consumers tend to have a slightly positive sentiment towards McDonald's, with a mean of 1.0138 on a scale ranging from -5 (I Hate It!) to +5 (I Love It!). The variability in responses, indicated by a standard deviation of 2.355, suggests diverse opinions among consumers.

Gaussian Mixture Model (GMM) and Linear Regression:

Following the identification of clusters through GMM, linear regression models were fitted for each cluster to predict the `Like_n` score based on the independent variables.

Cluster 1 Regression Model Summary:

OLS Regression Results

Dep. Variable:	Like_n	R-squared:	0.501
Model:	OLS	Adj. R-squared:	0.488
Method:	Least Squares	F-statistic:	38.23
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	1.51e-56
Time:	22:38:16	Log-Likelihood:	-846.22
No. Observations:	431	AIC:	1716.
Df Residuals:	419	BIC:	1765.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

const	1.2475	0.246	5.077	0.000	0.764	1.731
x1	0.8511	0.113	7.525	0.000	0.629	1.073
x2	0.1440	0.070	2.067	0.039	0.007	0.281
x3	-0.0543	0.083	-0.657	0.512	-0.217	0.108
x4	0.0067	0.103	0.065	0.949	-0.197	0.210
x5	-0.1769	0.095	-1.860	0.064	-0.364	0.010
x6	0.0390	0.069	0.564	0.573	-0.097	0.175
x7	0.0988	0.132	0.749	0.454	-0.160	0.358
x8	0.4957	0.108	4.592	0.000	0.284	0.708
x9	-0.0974	0.188	-0.517	0.605	-0.468	0.273
x10	0.1541	0.094	1.645	0.101	-0.030	0.338
x11	-0.5224	0.088	-5.935	0.000	-0.695	-0.349

==========

Omnibus:	11.756	Durbin-Watson:	2.040
Prob(Omnibus):	0.003	Jarque-Bera (JB):	11.937
Skew:	-0.394	Prob(JB):	0.00256
Kurtosis:	3.207	Cond. No.	7.33

==========

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Cluster 2 Regression Model Summary:

OLS Regression Results

==========

Dep. Variable:	Like_n	R-squared:	0.471
Model:	OLS	Adj. R-squared:	0.465
Method:	Least Squares	F-statistic:	71.07
Date:	Sun, 25 Aug 2024	Prob (F-statistic):	2.89e-93

Time: 22:38:16 Log-Likelihood: -1402.1

No. Observations: 727 AIC: 2824.

Df Residuals: 717 BIC: 2870.

Df Model: 9

Covariance Type: nonrobust

==========

	coef	std err	t	P>|t|	[0.025	0.975]
x1	0.8151	0.081	10.043	0.000	0.656	0.974
x2	0.2362	0.083	2.830	0.005	0.072	0.400
x3	-0.1342	0.066	-2.042	0.042	-0.263	-0.005
x4	-0.1605	0.065	-2.467	0.014	-0.288	-0.033
x5	-0.1579	0.067	-2.351	0.019	-0.290	-0.026
x6	0.5187	0.041	12.507	0.000	0.437	0.600
x7	-0.0453	0.092	-0.491	0.624	-0.226	0.136
x8	0.6257	0.083	7.508	0.000	0.462	0.789
x9	-1.1963	0.096	-12.507	0.000	-1.384	-1.008
x10	0.0356	0.067	0.529	0.597	-0.096	0.168
x11	-0.4606	0.076	-6.053	0.000	-0.610	-0.311

==========

Omnibus: 16.937 Durbin-Watson: 1.990

Prob(Omnibus): 0.000 Jarque-Bera (JB): 17.476

Skew: -0.371 Prob(JB): 0.000160

Kurtosis: 3.166 Cond. No. 9.72e+15

==========

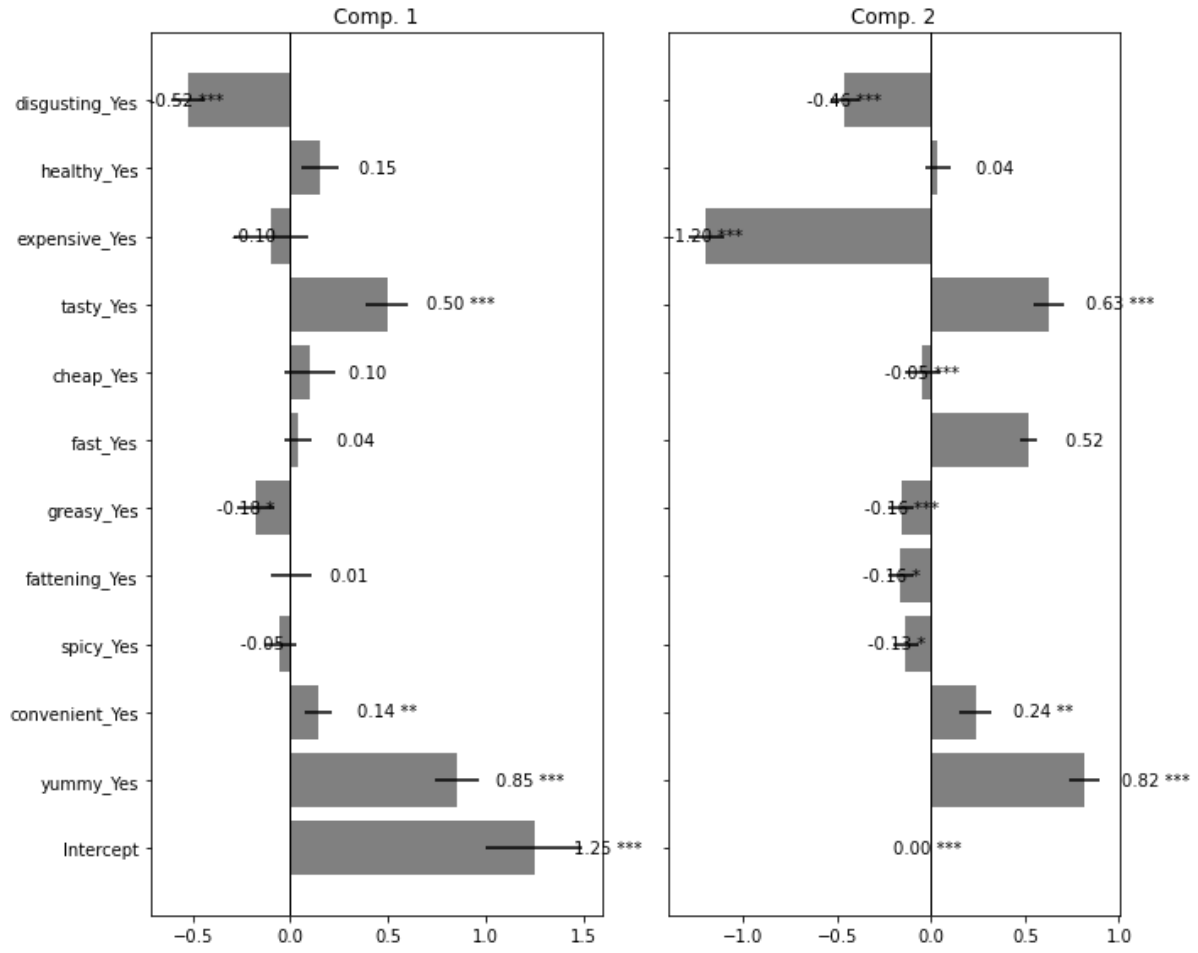

Figure 7:Regression coefficient of thetwo segments

The plot in Fig.7 illustrates the results of two regression models from a latent class regression analysis, displaying how different perceptions of McDonald's influence customer sentiment in two distinct consumer segments. The horizontal bars represent the coefficients for each variable, indicating their impact on the dependent variable—how much consumers like or dislike McDonald's. Positive coefficients suggest an increase in liking, while negative coefficients suggest a decrease. The stars next to the coefficients denote the statistical significance of these effects: three stars (***, $p < 0.001$) indicate a highly significant result, two stars (**, $p < 0.01$) indicate moderate significance, and one star (*, $p < 0.05$) indicates a lower level of significance. No stars mean the effect is not statistically significant.

In the first component (Comp. 1), the perception of McDonald's as "yummy" (0.85***) and "tasty" (0.50***) are major positive drivers of liking, both highly significant. Conversely, the perception of McDonald's as "disgusting" (-0.52***) significantly decreases liking. Other variables like "healthy", "expensive", "cheap", "fast", "greasy", "fattening", and "spicy" do not have significant effects in this segment.

In the second component (Comp. 2), similar trends are observed, with "yummy" (0.82***) and "tasty" (0.63***) also showing strong positive and highly significant effects. However, the

perception of McDonald's as "expensive" (-1.24***) has a highly significant negative impact, more pronounced than in the first segment. "Disgusting" (-0.44***) again shows a significant negative effect. The perception of McDonald's as "convenient" (0.24**) has a moderately significant positive impact in this segment, though it is less critical than the taste-related attributes.

This analysis helps identify key drivers of consumer sentiment in each segment, highlighting which perceptions McDonald's should focus on enhancing or mitigating to improve customer satisfaction and reduce negative attitudes.

References:

- Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market Segmentation Analysis: Understanding It, Doing It, and Making It Useful*. Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-10-8818-6

- https://github.com/ASNAJAMS/Market-Segmentation