> B.Sc. in Artificial Intelligence - A.Y. 2023/24
> **Fundamentals of Data Science Project**.
>
> Project B assigned to the group composed of:
> *Roberto Bruno, Alessandro Carosia, Emmanuele Crifasi, Mirko Speciale, Martina Taormina*.

Introduction

Market segmentation analysis is a powerful tool that helps companies target their marketing efforts more effectively and efficiently. By dividing a market through data analysis into distinct groups (or clusters) with similar needs, preferences, and characteristics, companies can tailor their products, services, and marketing strategies to better meet customer needs.

This project, written in *Python* in a *Jupyter environment*, aims to build an in-depth analysis of the tastes of the clientele of one of the famous restaurants in the McDonald's chain.

We used the following data analysis and modeling techniques:

1. **PCA** (*Principal Component Analysis*)
2. **K-Means Clustering**
3. **Hierarchical Clustering**

Objectives

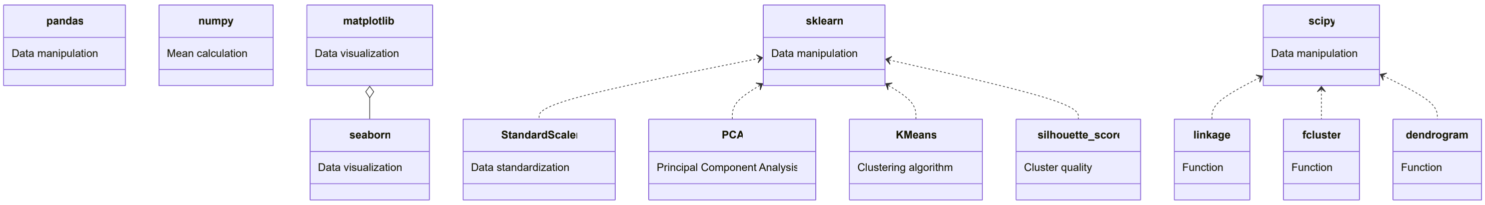
- **Customer Segmentation:** The goal is to divide McDonald's customers into distinct groups based on their preferences and characteristics.
  - This is done using cluster analysis techniques, such as *K-Means* and *Hierarchical Clustering*.
- **Detection of correlations between features:** The analysis seeks to discover if there are significant relationships between different customer characteristics and their preferences for McDonald's products.
  - This is done primarily through *Principal Component Analysis* (PCA) and *correlation matrix analysis*.

✓ **Data science to build a marketing strategy**

The idea is that by identifying groups of customers with similar preferences, McDonald's can more effectively tailor its marketing strategies and products to meet the needs of each segment.

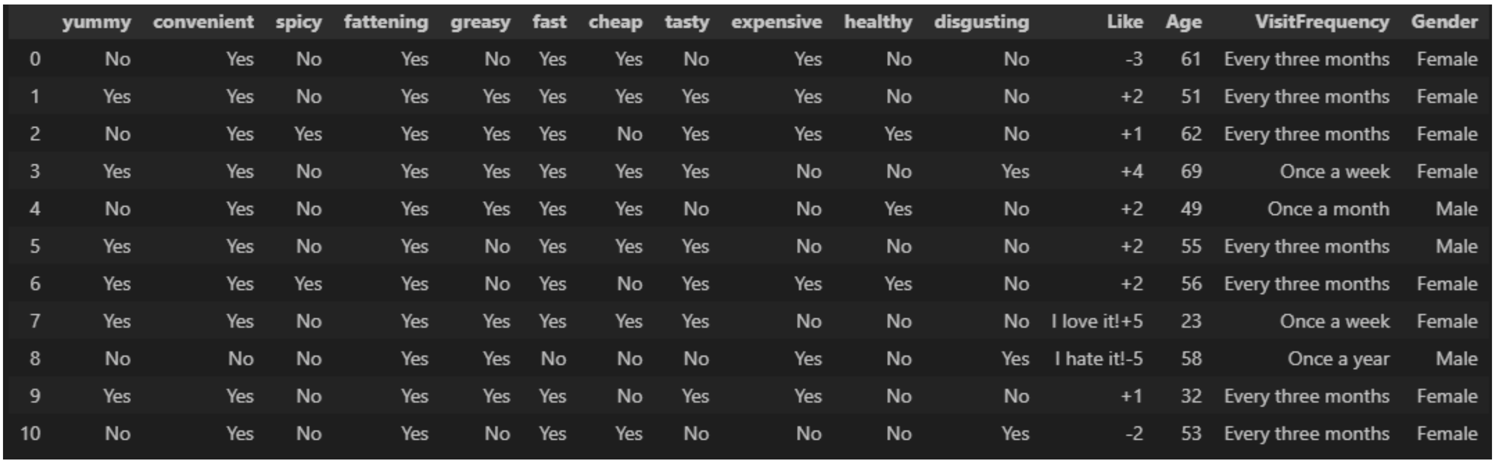
Workflow

Libraries used



First look at the data

1. **Reading the dataset ("mcdonalds.csv"):** The code uses the pandas library (`pd.read_csv`) to read the "mcdonalds.csv" CSV file and store it in a DataFrame called `data` .
   - The DataFrame is a tabular data structure that allows for efficient data manipulation and analysis.



	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No	No	-3	61	Every three months	Female
1	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	+2	51	Every three months	Female
2	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	No	+1	62	Every three months	Female
3	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	Yes	+4	69	Once a week	Female
4	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes	No	+2	49	Once a month	Male
5	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	No	No	+2	55	Every three months	Male
6	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	+2	56	Every three months	Female
7	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	I love it!+5	23	Once a week	Female
8	No	No	No	Yes	Yes	No	No	No	Yes	No	Yes	I hate it!-5	58	Once a year	Male
9	Yes	Yes	No	Yes	Yes	Yes	No	Yes	Yes	No	No	+1	32	Every three months	Female
10	No	Yes	No	Yes	No	Yes	Yes	No	No	No	Yes	-2	53	Every three months	Female

2. **Initial data inspection:**
   - `data.info()` : Provides a summary of the columns in the DataFrame, including their data types (e.g., object, int64) and the number of non-null values. This function is useful for quickly identifying any missing values and understanding the overall data structure.

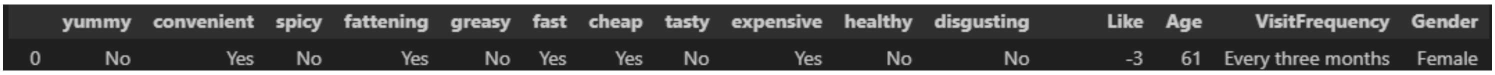
```
>>>data.info()
```

```
<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
```

Dataset Description

The analyzed dataset contains the results of a survey conducted on McDonald's customers. It includes both **demographic data** (gender and age) and **preferences** regarding various aspects of the McDonald's experience.



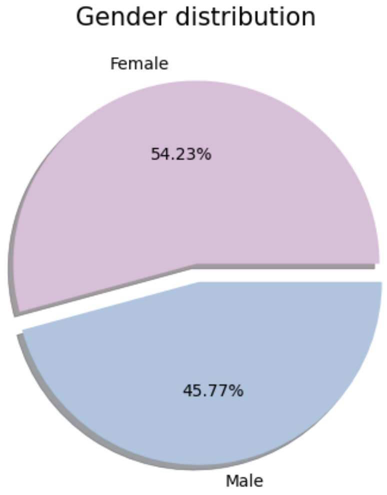
	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	No	Yes	No	Yes	No	Yes	Yes	No	Yes	No	No	-3	61	Every three months	Female

Overview of features

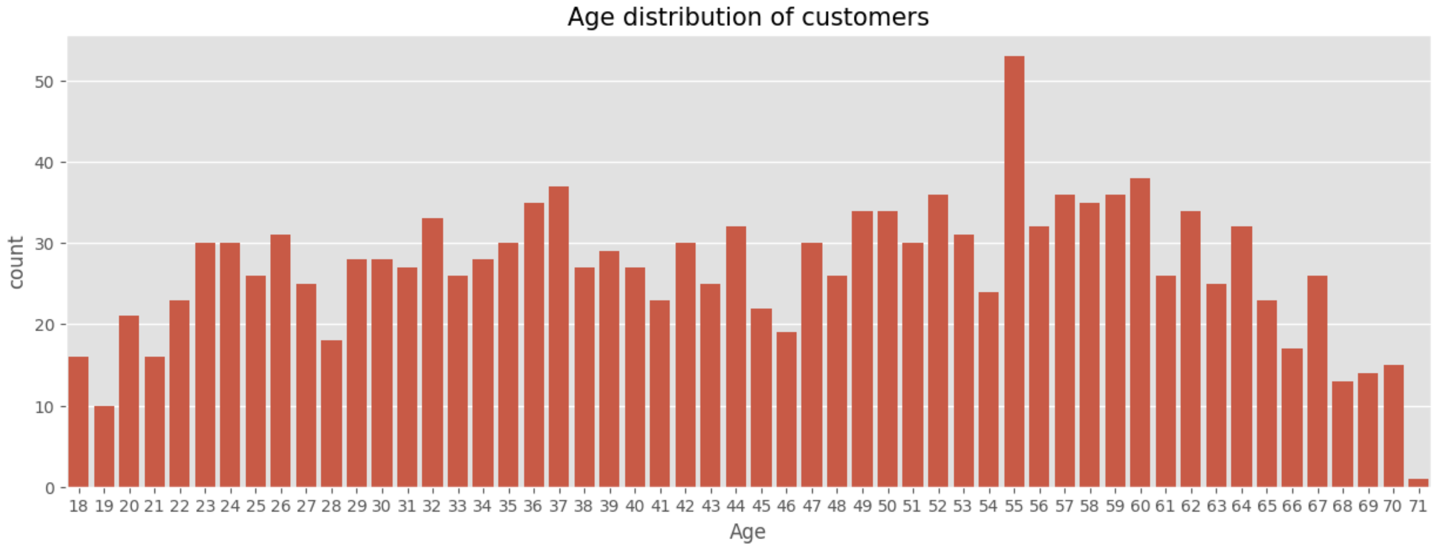
The variables can be classified into two main categories:

1. **Demographic Data:**
   - **Gender:** Male or Female.
   - **Age:** The age of the survey participants was collected and later categorized into age groups for analysis.
2. **Preferences and Opinions:**
   - **Ratings of McDonald's food characteristics:** Customers were asked to rate specific aspects of the food such as "tasty," "convenient," "healthy," "fattening," etc.
   - **Visit frequency:** This variable indicates how often a customer visits McDonald's (e.g., several times a week, once a week, once a month, etc.).
   - **Overall liking:** An overall rating of the McDonald's experience expressed on a scale from -5 to +5.

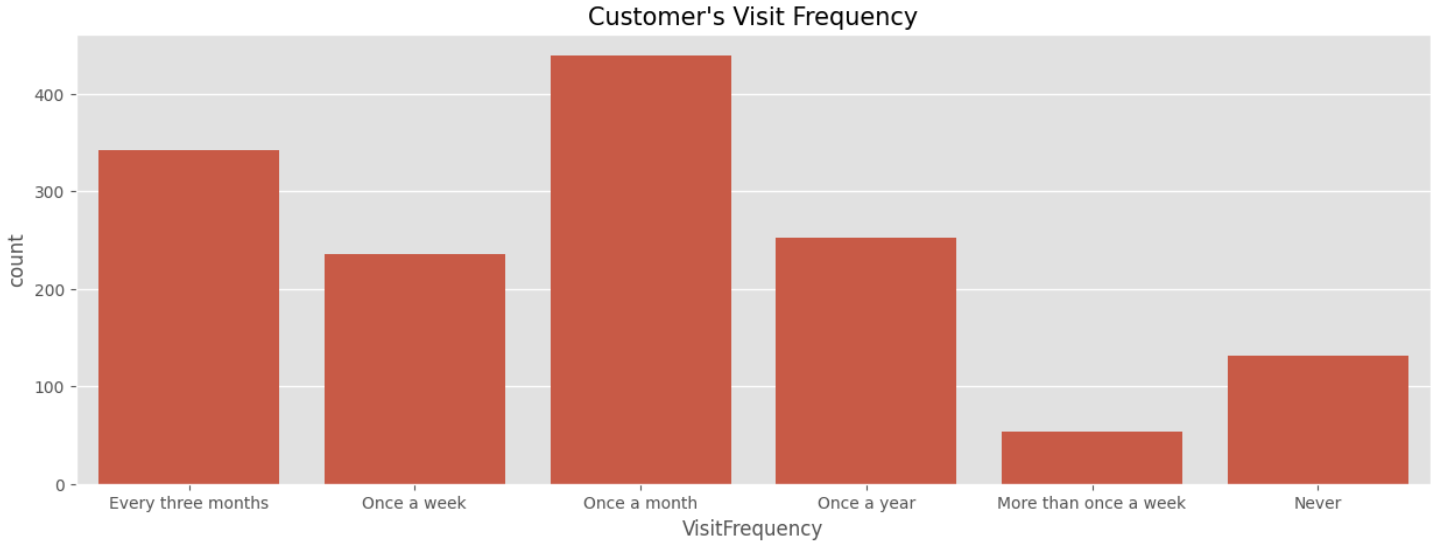
Graphical visualizations with observations



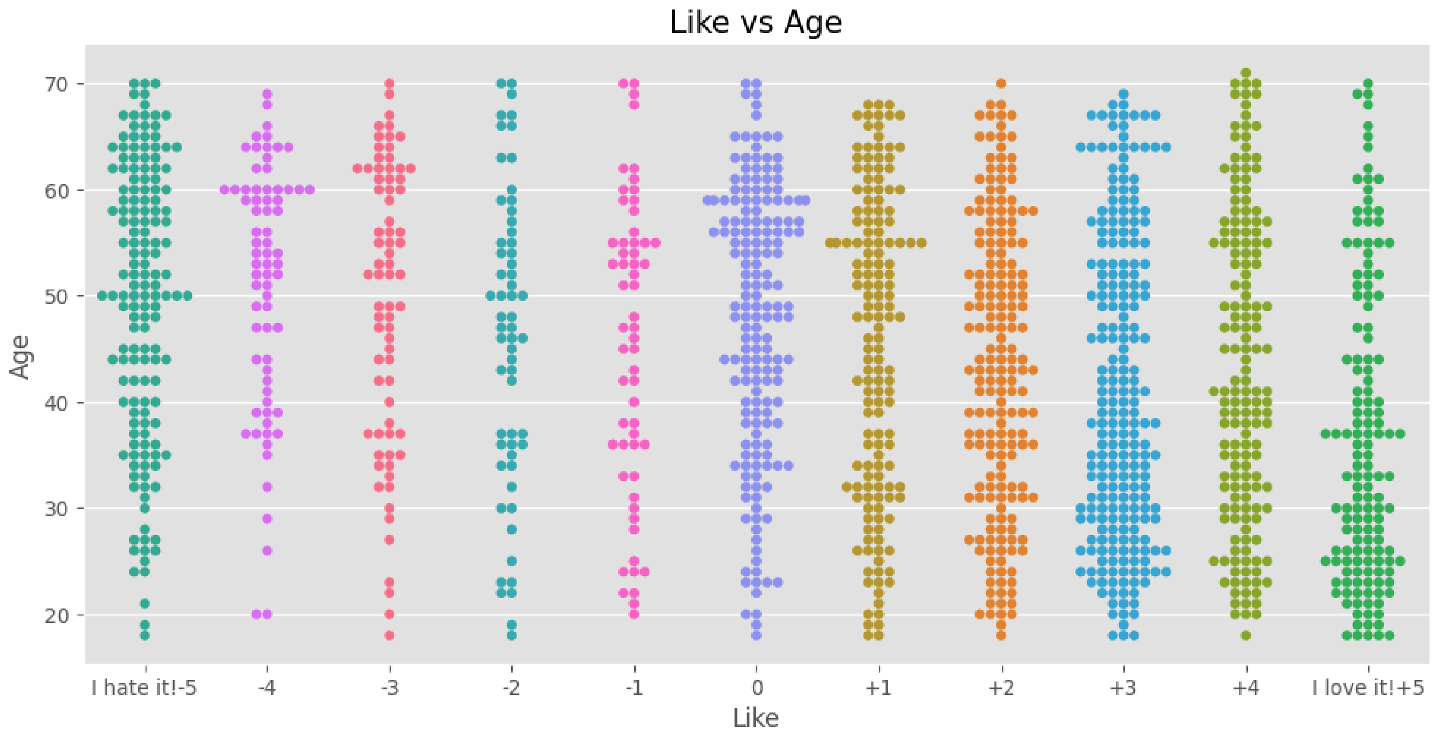
- The pie chart on gender distribution shows a slight **prevalence of female customers (54.23%)** compared to male customers (45.77%).
  - *The difference is not very marked, suggesting that McDonald's attracts both genders fairly evenly.*



- **Majority between 30 and 60 years old**: Most customers are aged between 30 and 60, with peaks around 35-40 years and 55-60 years.
    - The chart suggests that McDonald's might want to focus its marketing campaigns on the 30 to 60 age groups.
- **Few young and elderly customers**: There are relatively few customers under 25 and over 65.
    - The company might also want to develop products or promotions that appeal more to younger or older customers.



- **"Once a month" is the highest frequency**: The chart clearly shows that most customers tend to visit McDonald's once a month, while very few visit more than once a week.
    - The company could design campaigns to encourage customers to use the service more frequently, bridging this gap between the two visit frequencies.



- **General appreciation**: Customers tend to give more positive ratings than negative ones.
- **Older age groups**: The customer age group in the (50, 70) range is the one that did not particularly appreciate McDonald's, although there are still customers who do.
- **Younger age groups**: Almost all younger customers appreciate the service.

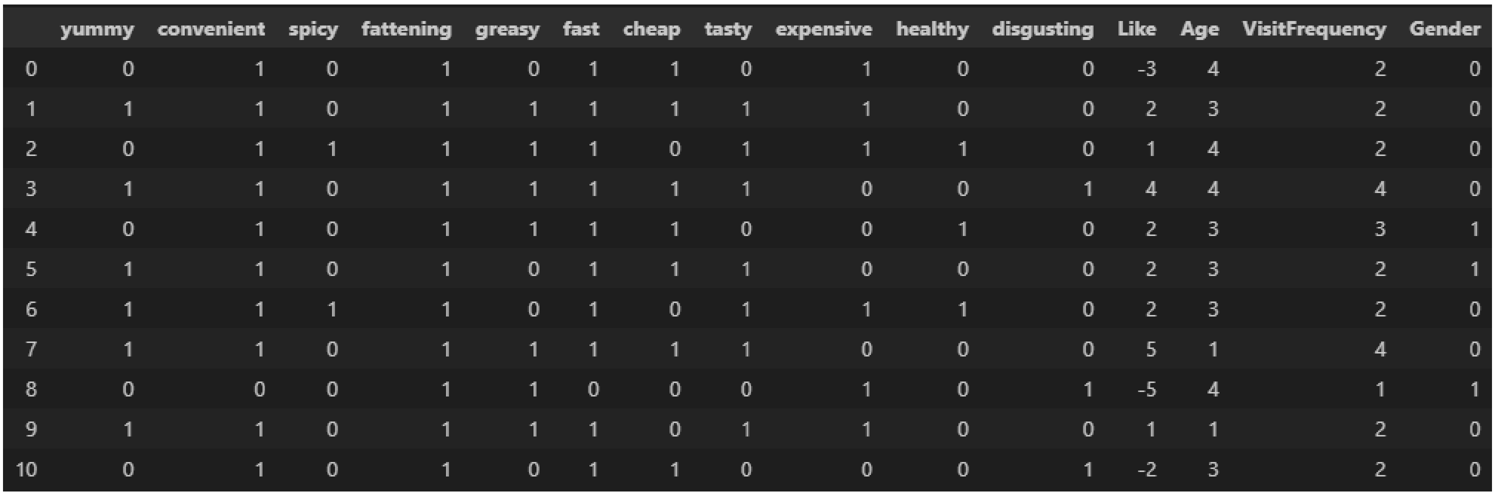
Data Problems or Limitations

1. **No missing data:** There are no missing values in the dataset, which simplifies the analysis and does not require missing data imputation techniques.
   - `data.isna().sum()` : Calculates and prints the number of missing (NaN) values for each column.
2. **Lack of context for some variables:** Some variables, such as the overall rating of the McDonald's experience ("Like"), were collected on a numerical scale without providing clear context to the participants. This could introduce a degree of subjectivity into the responses.
3. **Absence of temporal information:** The dataset does not include information on the period when the survey was conducted. This could be relevant if customer preferences or the characteristics of McDonald's food have changed over time.

Data Discretization

Data discretization is a fundamental step in this project to adapt the variables to the type of analysis to be performed.

- Python dictionaries were used to replace categorical values with numerical values.



	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	0	1	0	1	0	1	1	0	1	0	0	-3	4	2	0
1	1	1	0	1	1	1	1	1	1	0	0	2	3	2	0
2	0	1	1	1	1	1	0	1	1	1	0	1	4	2	0
3	1	1	0	1	1	1	1	1	0	0	1	4	4	4	0
4	0	1	0	1	1	1	1	0	0	1	0	2	3	3	1
5	1	1	0	1	0	1	1	1	0	0	0	2	3	2	1
6	1	1	1	1	0	1	0	1	1	1	0	2	3	2	0
7	1	1	0	1	1	1	1	1	0	0	0	5	1	4	0
8	0	0	0	1	1	0	0	0	1	0	1	-5	4	1	1
9	1	1	0	1	1	1	0	1	1	0	0	1	1	2	0
10	0	1	0	1	0	1	1	0	0	0	1	-2	3	2	0

This process was applied to:

1. **Binary columns (Yes/No, Male/Female):**

- Columns like `yummy`, `convenient`, `spicy`, etc., which originally contained "*Yes*" or "*No*" values, were converted into 0/1 binary variables.
  - The `replacement_dict` dictionary maps "*Yes*" to 1 and "*No*" to 0. Similarly, the `Gender` column was transformed by mapping "*Male*" to 1 and "*Female*" to 0.

2. **"Like" column:**
   - The `Like` column, which represented a liking on a scale from "*I hate it! -5*" to "*I love it! +5*", was converted into numerical values from -5 to 5 using the same `replacement_dict`.
3. **"VisitFrequency" column:**
   - The `VisitFrequency` column, which indicated the frequency of visits with textual values (e.g., "*More than once a week*"), was transformed into an ordinal scale from 0 to 5.
     - The `VisitSostitution` dictionary maps each visit frequency to a corresponding numerical value, where 0 represents "*Never*" and 5 represents "*More than once a week*".
4. **"Age" column:**
   - The `Age` column, which contained the customers' ages, was discretized into age groups. Bins (`age_bins`) representing age intervals and labels (`age_labels`) corresponding to each interval were defined.
     - The `pd.cut` function was used to assign the appropriate age group label to each customer.

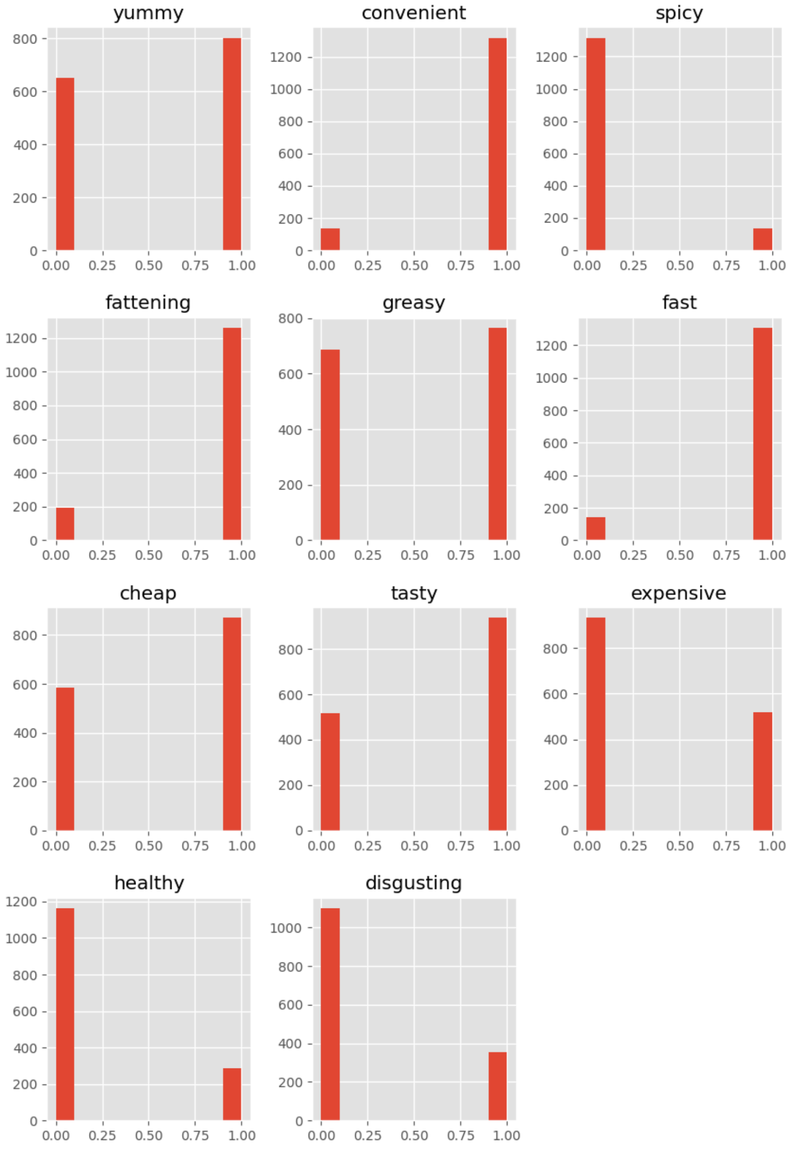
Limitations

- **Loss of information:** Discretization leads to a loss of granularity in the data. For example, in the `Age` column, there is no longer a distinction between a 25-year-old and a 30-year-old customer, as both belong to the same age group.
- **Choice of bins:** The definition of bins in the age discretization is arbitrary and could influence the results of the analysis.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was a crucial phase in the data analysis project. It is an initial investigation process where statistical techniques and visualizations are used to:

- **Understand the data structure:** What are the variables present? What are their types and distributions?
- **Identify relationships and patterns:** Are there correlations between variables? Are there outliers or interesting trends?
- **Check data quality:** Are there missing values or errors that need to be addressed?



yummy

convenient

spicy

fattening

greasy

fast

cheap

tasty

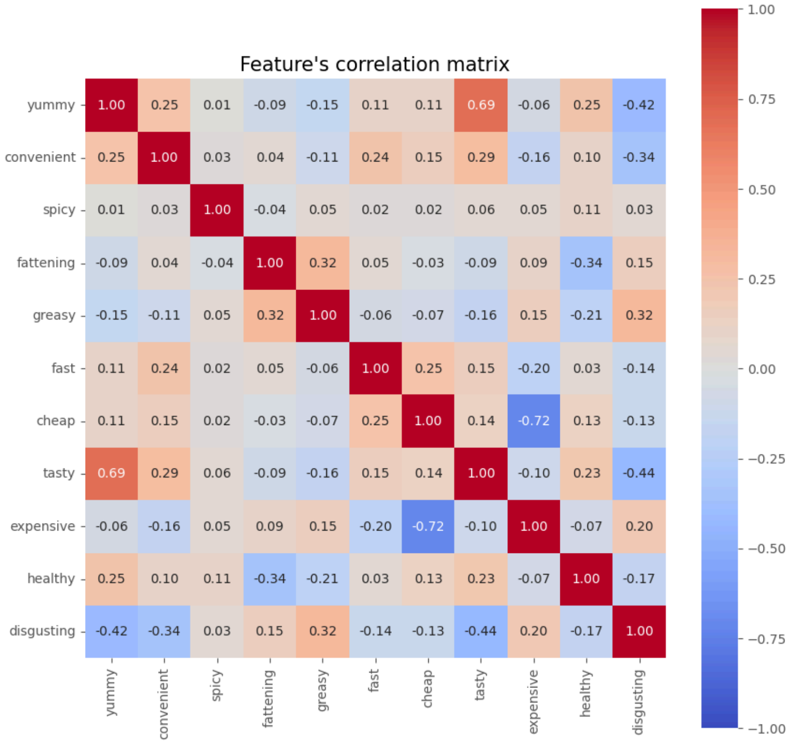
expensive

healthy

disgusting

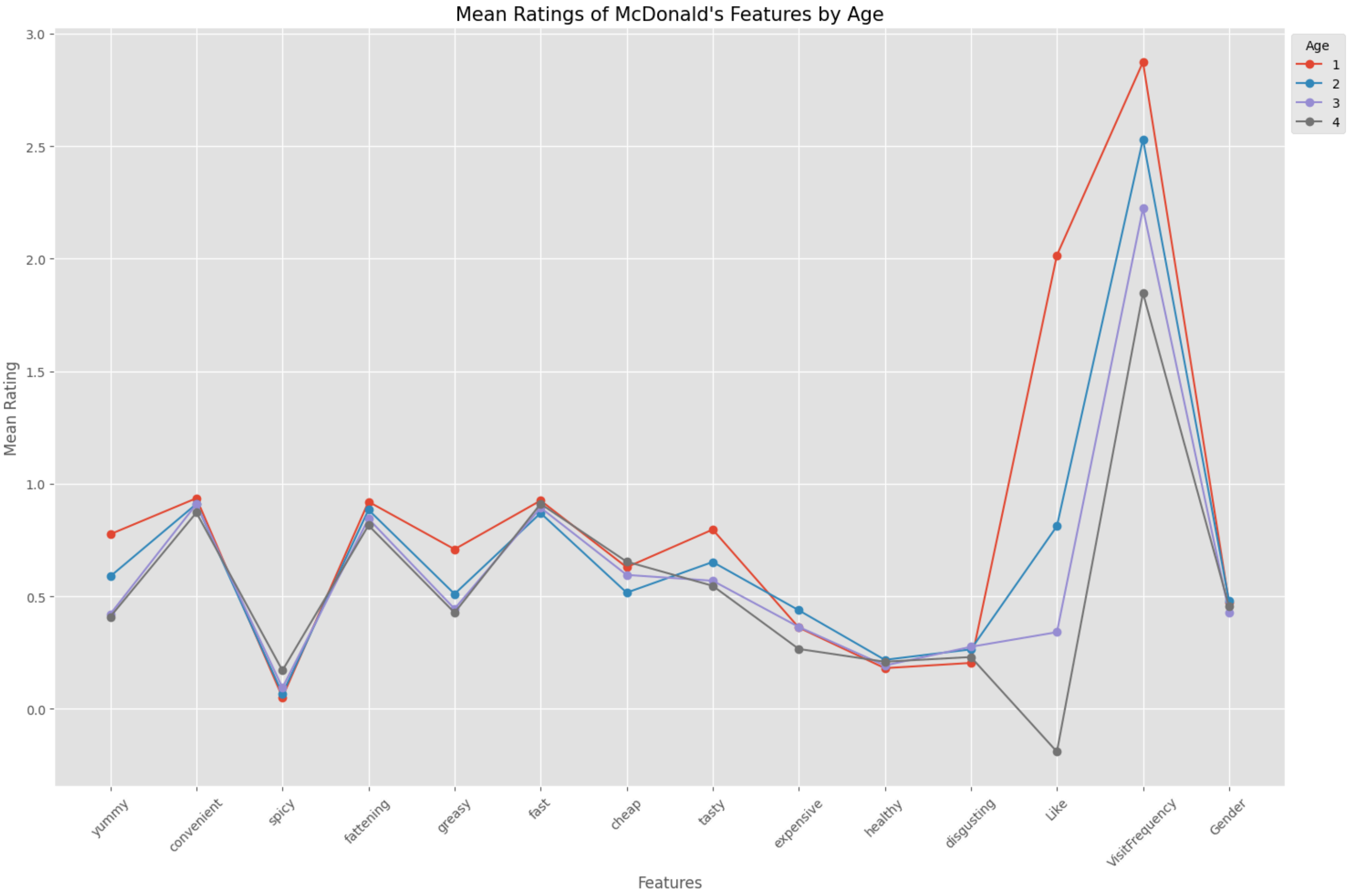
The food is perceived mainly as delicious, convenient, spicy, fattening, greasy, fast, cheap, tasty, but not healthy. It is also considered expensive, but not disgusting.

- **Yummy, Convenient, Spicy, Fattening, Greasy, Fast, Cheap, Tasty**: For these attributes, the majority of ratings is 1.
- **Expensive**: For this attribute as well, the majority of ratings is 1, suggesting that many consider the fast-food chain's food expensive.
- **Healthy**: The majority of ratings is 0, indicating that the food is not considered healthy by most of the sample population.
- **Disgusting**: The majority of ratings is 0, indicating that the food is not considered disgusting by most of the respondents.



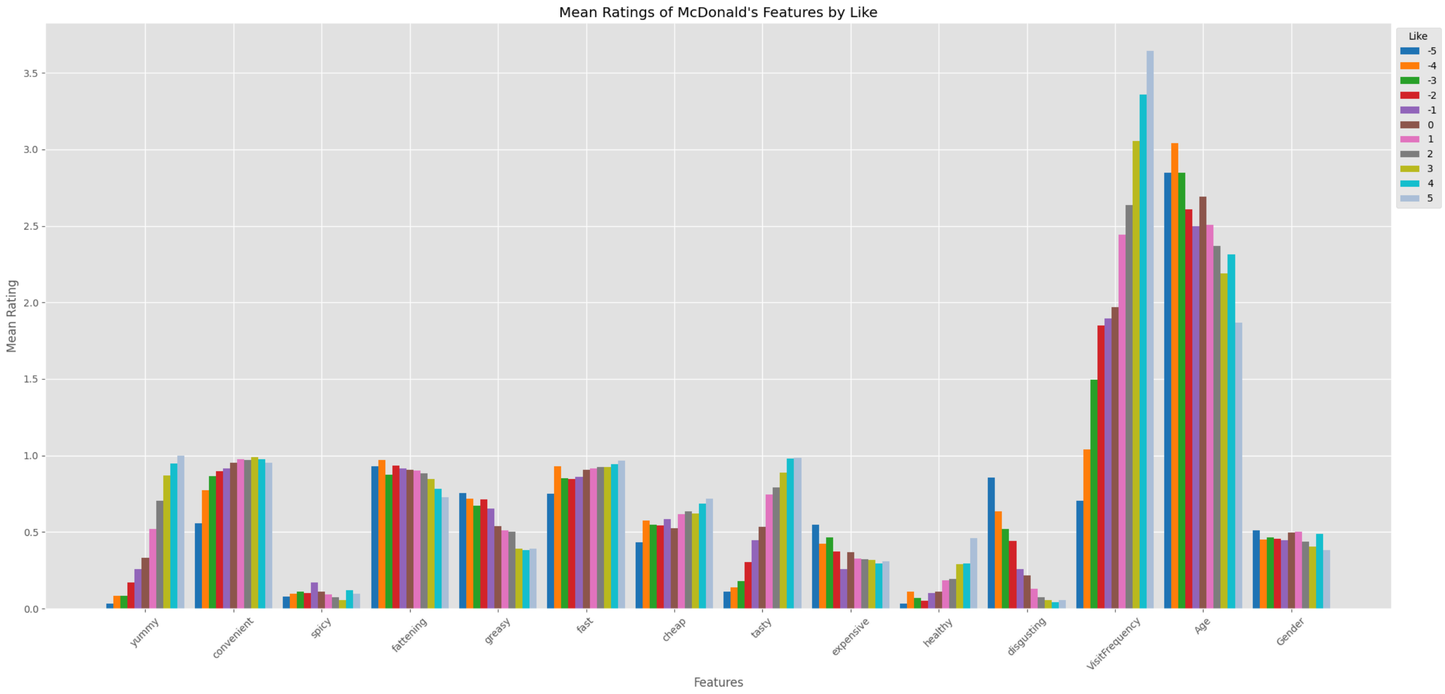
The graph shows that most features are not correlated with each other, despite some exceptions.

- **Positive correlations (red):** "*yummy*" tends to be the feature with the most positive correlations, in particular, "*yummy*" and "*tasty*" have the highest positive value, which indicates that customers tend to consider a good product also tasty.
- **Negative correlations (blue):** In general, "*disgusting*" is negatively correlated with many features, including "*yummy*", "*convenient*", and "*tasty*", as is intuitive. The features, however, with the most negative value are "*expensive*" and "*cheap*", so customers who find McDonald's expensive certainly do not also call it cheap.



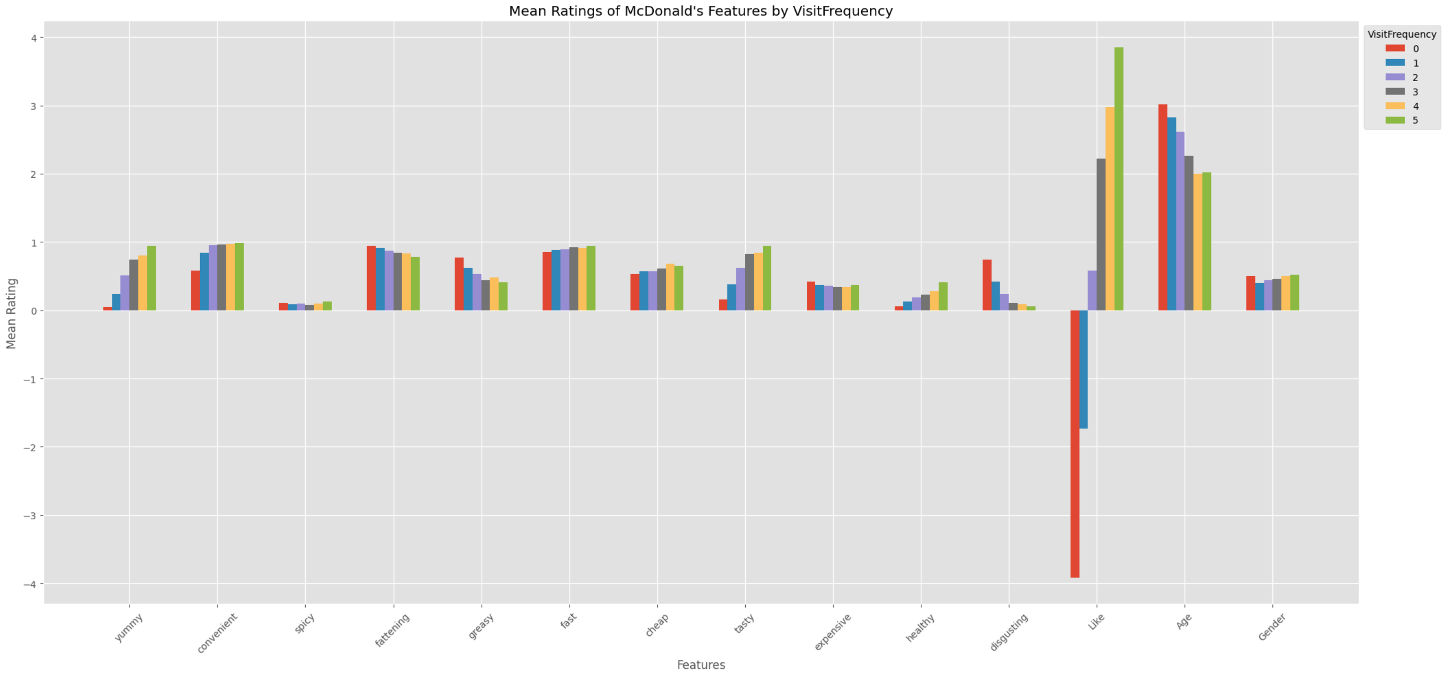
The graph shows 4 lines representing the bins into which we have divided the customers' age groups. *Although some features are shared by all age groups or, in the case of gender, represent an equal distribution*, some specific trends can be observed among the various groups.

- **Younger age groups (red)**: Younger customers are those who appreciate the service the most, which is why they tend to visit more often and consider it tasty, despite being the ones who consider it more "*greasy*".
- **Age groups from 33 to 45 (blue)**: This customer group is in the middle; they have no particular opinions about the service except for the cost: this age group considers McDonald's expensive.
- **Age groups from 45 to 71 (purple and gray)**: These two age groups tend to have shared opinions on almost everything. The graph shows that from age 45 onwards, people tend to appreciate McDonald's food less and less, so they avoid visiting despite considering it convenient. This highlights how customers in this age group prefer quality over savings, unlike younger people.



The graph shows the trend of appreciation for each feature: For some of these, the opinion does not create particularly evident gaps in the level of liking, as for "*spicy*" for example. The last feature shows that appreciation does not change based on gender but has a significant impact based on the frequency of visiting McDonald's.

- **General appreciation**: Customers tend to give positive/negative ratings consistent with their opinions on features that express liking/disliking of the service. This can be clearly seen with "*yummy*" and "*disgusting*", as the trend of one is the opposite of the other.



The graph shows the frequency trend for each feature: The values are organized so that smaller values indicate lower frequency (0 = Never) and larger values indicate higher frequency (5 = More than once a week).

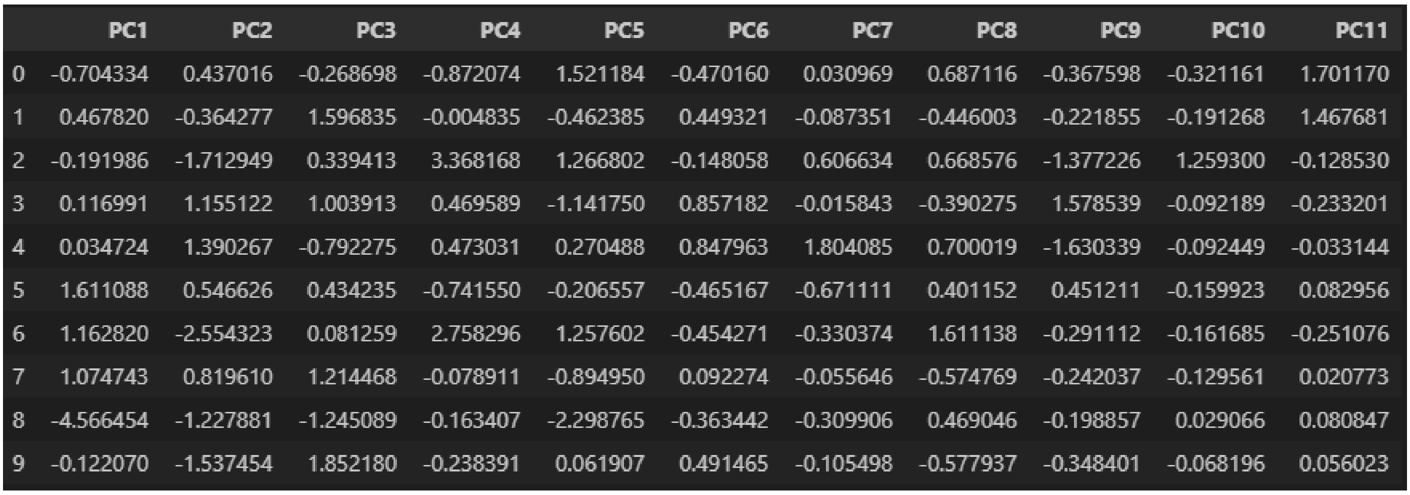
> - **General trend**: It can be noted that the most regular customers are those who appreciate McDonald's the most, unlike those who visit less frequently.
> - **People who have never been to McDonald's (red)**: The graph highlights that those who have never eaten at McDonald's do not have a positive opinion of it.
>   - The company should pay attention to the opinion of these people to encourage them to try the service or at least not to negatively influence the general opinion.

PCA

- **The problem:** We are faced with a large number of variables. These variables can be correlated with each other, making the analysis very complex and difficult to interpret.
- **The solution:** PCA is a statistical technique that aims to reduce the dimensionality of the data: it tries to combine the original variables into a smaller number of new variables, called principal components (PCs).
  - These PCs are constructed to capture most of the variance present in the original data, eliminating redundancies and correlations.

Dimensionality reduction was a crucial step to simplify our data and prepare it for clustering. *We used Principal Component Analysis (PCA) to achieve these goals:*

1. **Feature standardization:** We standardized the numerical variables to ensure they had a mean of zero and a standard deviation of one. This step is important to prevent variables with different scales from dominating the PCA analysis.
2. **Application of PCA:** We applied PCA to reduce the dimensions of our dataset, keeping the principal components that explain most of the variance in the data (PC1 and PC2).
3. **Analysis of weights:** Each PC is a linear combination of the original variables. The coefficients of this combination, called weights, indicate the importance of each original variable in defining the PC.
4. **Visualization of results:** We created a simple matrix, a *scree plot* to evaluate the variance explained by each principal component, and a biplot to visualize the relationship between the original variables and the principal components.



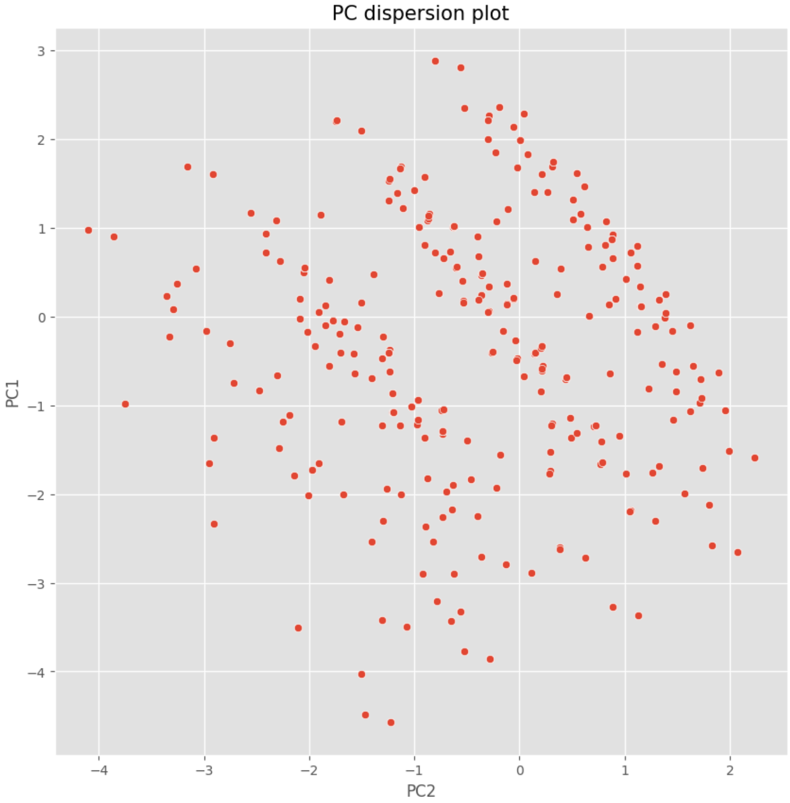
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
0	-0.704334	0.437016	-0.268698	-0.872074	1.521184	-0.470160	0.030969	0.687116	-0.367598	-0.321161	1.701170
1	0.467820	-0.364277	1.596835	-0.004835	-0.462385	0.449321	-0.087351	-0.446003	-0.221855	-0.191268	1.467681
2	-0.191986	-1.712949	0.339413	3.368168	1.266802	-0.148058	0.606634	0.668576	-1.377226	1.259300	-0.128530
3	0.116991	1.155122	1.003913	0.469589	-1.141750	0.857182	-0.015843	-0.390275	1.578539	-0.092189	-0.233201
4	0.034724	1.390267	-0.792275	0.473031	0.270488	0.847963	1.804085	0.700019	-1.630339	-0.092449	-0.033144
5	1.611088	0.546626	0.434235	-0.741550	-0.206557	-0.465167	-0.671111	0.401152	0.451211	-0.159923	0.082956
6	1.162820	-2.554323	0.081259	2.758296	1.257602	-0.454271	-0.330374	1.611138	-0.291112	-0.161685	-0.251076
7	1.074743	0.819610	1.214468	-0.078911	-0.894950	0.092274	-0.055646	-0.574769	-0.242037	-0.129561	0.020773
8	-4.566454	-1.227881	-1.245089	-0.163407	-2.298765	-0.363442	-0.309906	0.469046	-0.198857	0.029066	0.080847
9	-0.122070	-1.537454	1.852180	-0.238391	0.061907	0.491465	-0.105498	-0.577937	-0.348401	-0.068196	0.056023

> After applying PCA, a table is obtained that shows the percentage of variance explained by each principal component.
>
> - *Typically, the first few PCs explain most of the total variance, while the last ones explain a very small portion.*

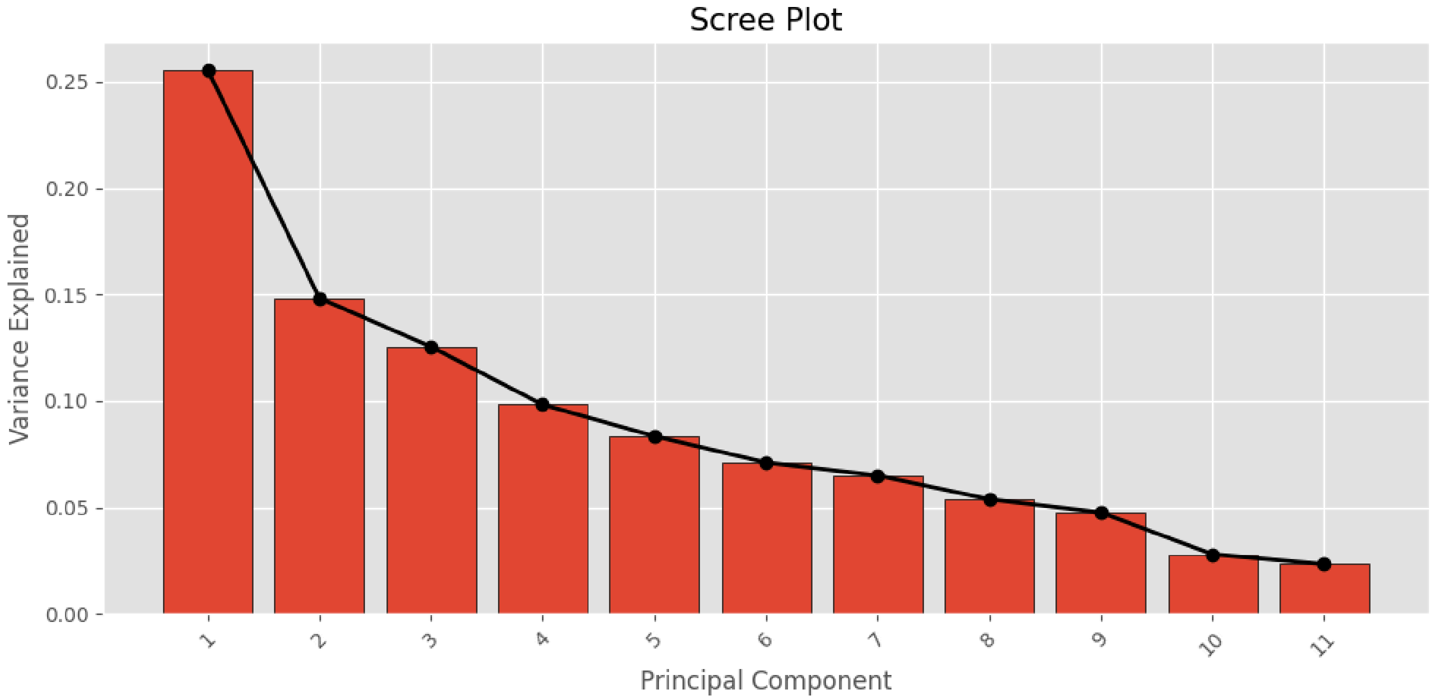
In summary..

- PCA allowed us to reduce the complexity of the data, making it easier to identify patterns and customer groups in the subsequent phases.
- The analysis found that most of the variance in the data can be explained by two principal components. The first principal component is positively correlated with characteristics like "tasty," "convenient," and "fast," and negatively with "unhealthy" and "disgusting." This suggests that most customers appreciate speed and convenience but are also health-conscious.

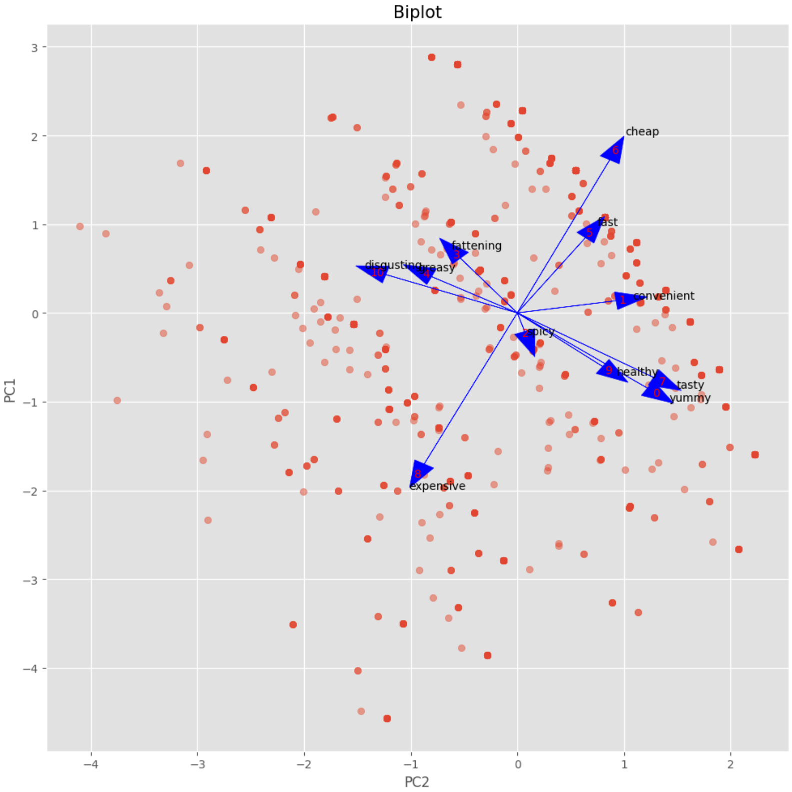
Graphical visualizations with observations



The graph shows a relatively homogeneous distribution of samples: as seen by the arrangement of points around the center, without distinct clusters. The higher variance of PC1 suggests that the major differences in the data are captured by the first principal component.

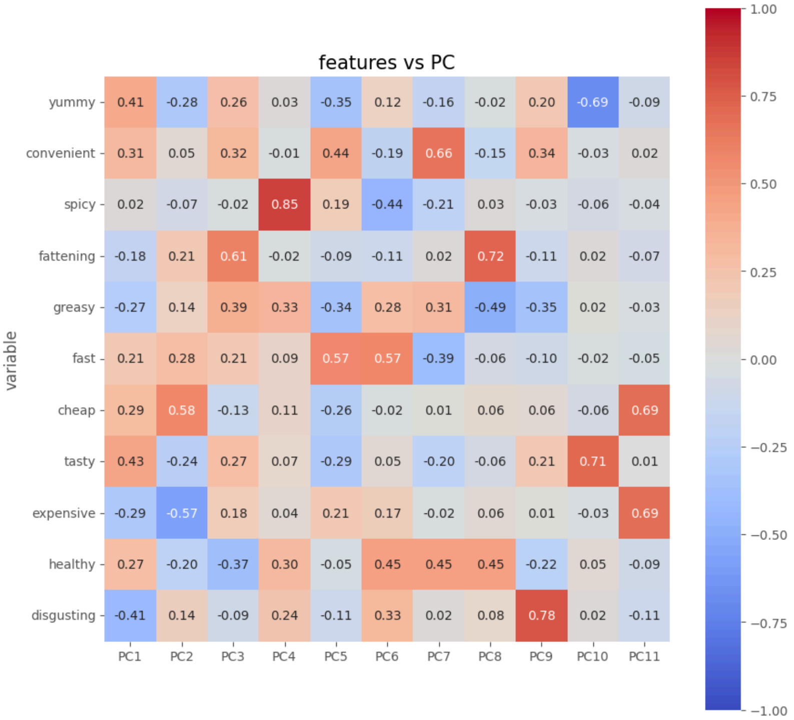


The graph shows the proportion of total variance for each principal component: Specifically, the first component is the one that best represents the data's variance, with a clear gap from the second component, where we can find the so-called elbow, which is the point on the curve that describes the optimal number of principal components.



The graph is an overlay of scores and loadings and specifically, you can see how each variable contributes to the two principal components and with the other variables.

- **Correlations between variables**: The variables with negative correlations are those that immediately stand out as they form an angle of about 180°, like cheap and expensive, as well as the features that constitute a negative opinion of the company ("*fattening*", "*greasy*", and "*disgusting*"), which indeed have a positive correlation due to the proximity of the arrows and their direction, and those that create a positive one ("*tasty*", "*yummy*", "*healthy*"), also with a positive correlation.
- **Correlations between variables and PCs**: In general, the level of parallelism of the features with respect to the PC axes determines their contribution to that component.
  - It is noted that "*convenient*", being parallel to the x-axis, has a greater contribution to the second principal component, while "*spicy*", on the contrary, is more influential in the first principal component, although its contribution is minimal due to the length of the arrow.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
yummy	0.41	-0.28	0.26	0.03	-0.35	0.12	-0.16	-0.02	0.20	-0.69	-0.09
convenient	0.31	0.05	0.32	-0.01	0.44	-0.19	0.66	-0.15	0.34	-0.03	0.02
spicy	0.02	-0.07	-0.02	0.85	0.19	-0.44	-0.21	0.03	-0.03	-0.06	-0.04
fattening	-0.18	0.21	0.61	-0.02	-0.09	-0.11	0.02	0.72	-0.11	0.02	-0.07
greasy	-0.27	0.14	0.39	0.33	-0.34	0.28	0.31	-0.49	-0.35	0.02	-0.03
fast	0.21	0.28	0.21	0.09	0.57	0.57	-0.39	-0.06	-0.10	-0.02	-0.05
cheap	0.29	0.58	-0.13	0.11	-0.26	-0.02	0.01	0.06	0.06	-0.06	0.69
tasty	0.43	-0.24	0.27	0.07	-0.29	0.05	-0.20	-0.06	0.21	0.71	0.01
expensive	-0.29	-0.57	0.18	0.04	0.21	0.17	-0.02	0.06	0.01	-0.03	0.69
healthy	0.27	-0.20	-0.37	0.30	-0.05	0.45	0.45	0.45	-0.22	0.05	-0.09
disgusting	-0.41	0.14	-0.09	0.24	-0.11	0.33	0.02	0.08	0.78	0.02	-0.11

The heatmap between features and PCs allows us to see the contribution of each single feature to all principal components. The minimal contribution of "*spicy*" to almost all components except for PC4 is confirmed, while some features show their highest contribution in the final PCs, even though the percentage of variance in those is minimal, as shown by the Scree Plot.

> - **The choice of the first two PCs is therefore confirmed to be the most accurate** according to the heatmap, as it clearly shows that these are the components where the features contribute more or less all together.

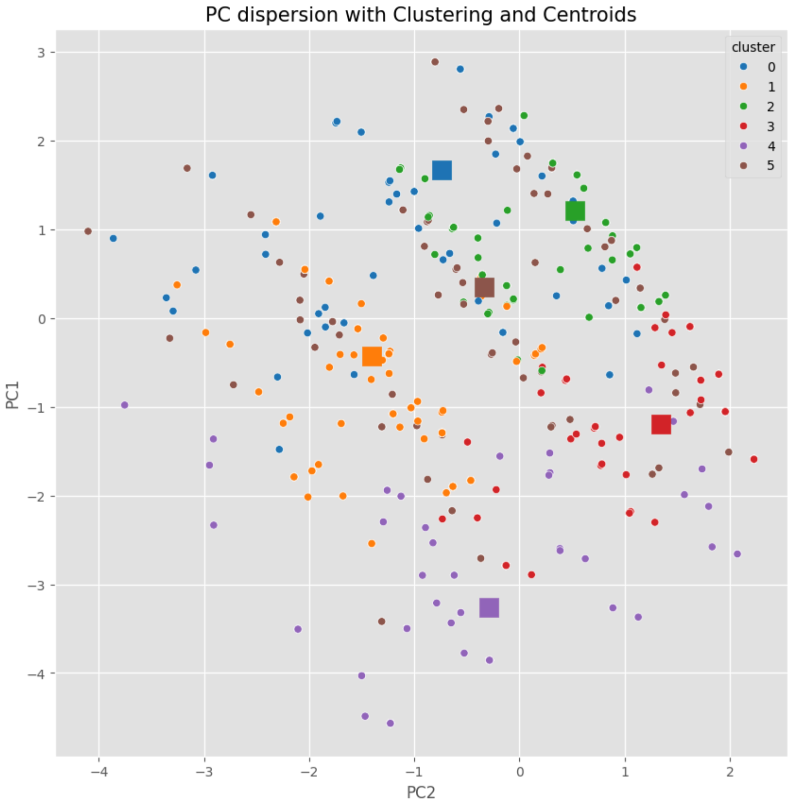
Clustering

Clustering was the core of our project, allowing us to identify groups of customers with similar preferences. We used two main techniques:

K-Means Clustering

K-Means clustering is a partitional clustering algorithm. The goal is to divide customers into a predefined number of groups (K), so that customers within each group are as similar as possible to each other and as different as possible from customers in other groups.

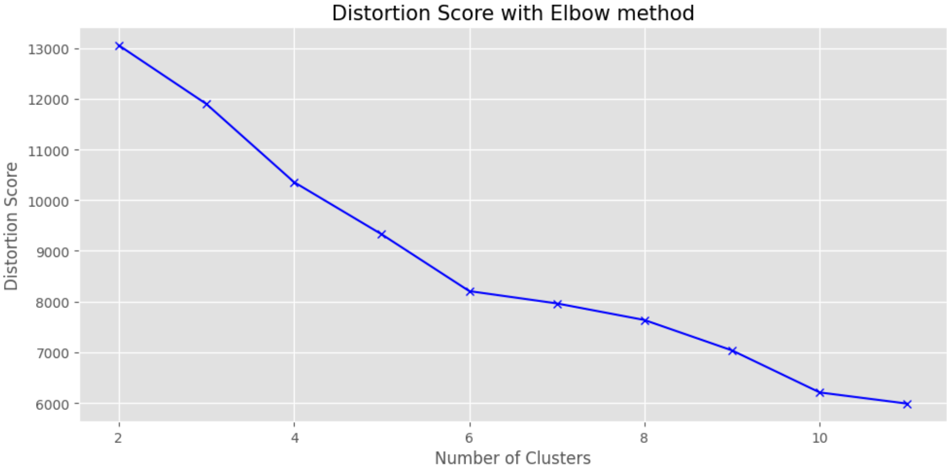
- K-means is computationally efficient but requires specifying the number of clusters in advance.
- Use of the **Elbow method** and **Silhouette Score** to determine the optimal number of clusters.

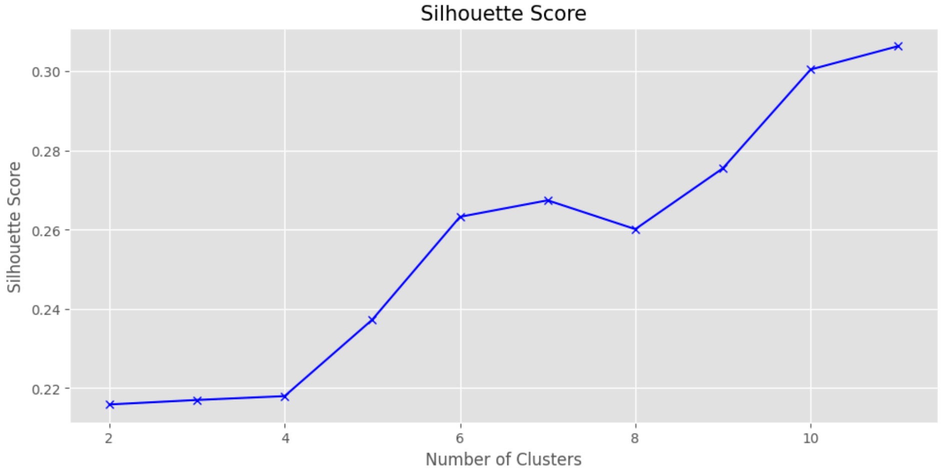


> The graphs show how the points are projected onto the 2 components and divided into clusters.
>
> - From the way the clusters are distributed, we can see that **although the centroids are quite far apart, the shape of the clusters is not very cohesive** and the data points "invade" other clusters, in some cases overlapping. This is likely due to the similarity of the data points, which does not allow for a clearer separation between the clusters and the data points themselves.

Determining the optimal number of clusters

- **Variance-based methods:** These methods aim to find the number of clusters that minimizes the within-cluster variance and maximizes the between-cluster variance. An example is the elbow method, which involves plotting a graph of the explained variance as a function of the number of clusters and looking for the point where the curve "bends" (like an elbow).
- **Silhouette-based methods:** These methods evaluate the quality of the clusters based on the silhouette, a measure of how well a customer fits into their own cluster compared to other clusters. The goal is to find the number of clusters that maximizes the average silhouette.



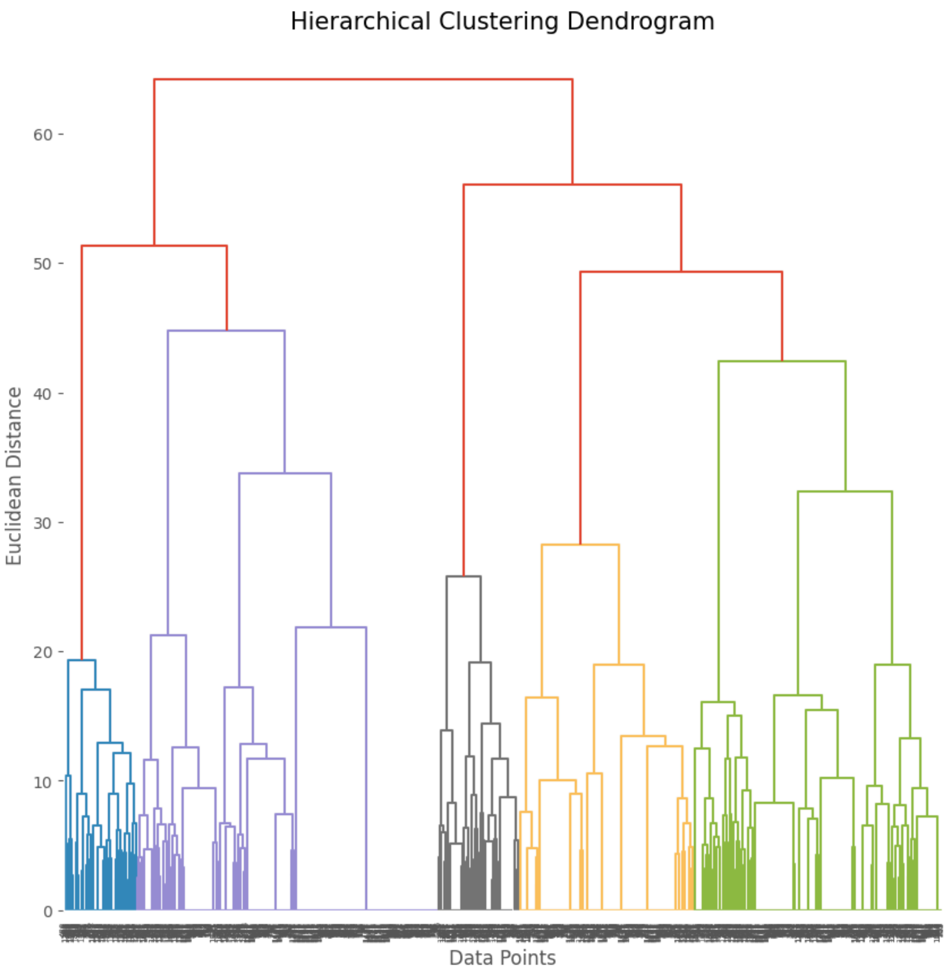


> - **Elbow method**: In the graph, it is clear that the elbow point where the sum of squared distances becomes marginal is at 6 clusters.
> - **Silhouette Score**: The silhouette coefficient, for example, is a metric used to evaluate the quality of a clustering, measuring cohesion within clusters and separation between clusters.
>   - Unlike the *elbow method*, which focuses on the overall evaluation of the clustering based on the within-cluster sum of squares (WSS), the silhouette coefficient provides information on the quality of the clustering at the individual point and cluster level.
>   - In this case, although the maximum Silhouette Score is given by 11 clusters, we opted for a number of clusters that represented the silhouette score well without ignoring what we had seen through the Elbow Method.

- **6 customer clusters were identified**, each with distinct preferences and demographic characteristics.
  - For example, one cluster is composed mainly of young people who appreciate tasty and convenient food, while another consists of older, health-conscious customers.

Hierarchical Clustering

- *Hierarchical Clustering* is an agglomerative clustering algorithm. It starts by considering each customer as a separate cluster and then progressively merges the most similar clusters until a single large cluster is obtained using the Ward method.
  - Hierarchical clustering does not require specifying the number of clusters in advance but can be less computationally efficient than K-means.



The dendrogram shows how the data was grouped into 6 main clusters with different hierarchical structures and varying levels of similarity. The Euclidean distance gives an idea of the similarity between clusters and within the clusters themselves.

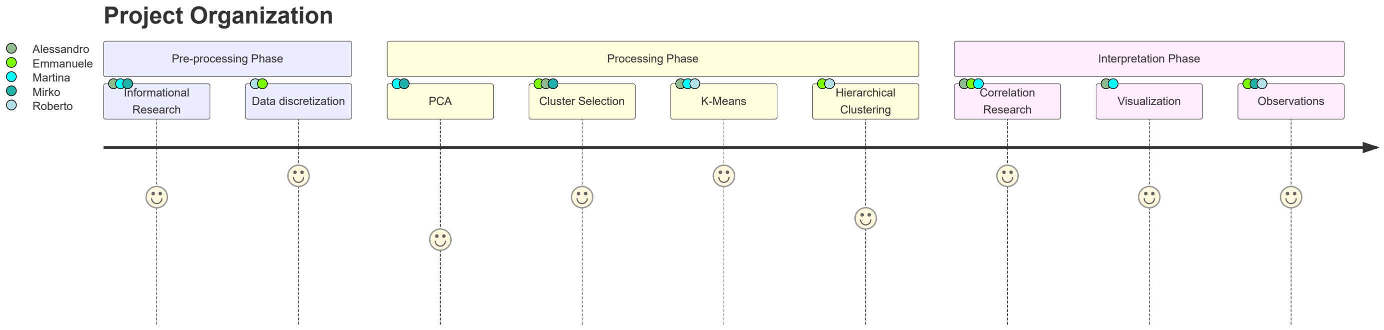
- The clusters at lower levels represent sub-clusters that merge into larger clusters as you move up the tree.

Deep dive into dendrograms

It should be noted that in a dendrogram, the Y-axis indicates the distance or dissimilarity between clusters; in our project, the axis was named "Euclidean distance".

- We call *branches* the vertical lines that connect the clusters, representing the fusions between them. These meet at points called nodes, which represent the moments when the clusters merged.
  - The height of a node with respect to the Y-axis indicates the distance between the clusters merged at that point. At the bottom, we have the leaves describing the terminal branches of the dendrogram and thus the original observations.
- The creation of a dendrogram allowed us to facilitate the understanding of the underlying structure of the data and therefore the similarity that exists between the people in our study.
  - The dendrogram also highlights the presence of outliers and anomalies in the data, which could be indicative of measurement errors or unusual phenomena.

Conclusions



In this data analysis project, we explored the preferences of McDonald's customers using techniques of exploratory data analysis (EDA), dimensionality reduction (PCA), and clustering (K-means and hierarchical).

1. Through **EDA**, we gained a deep understanding of customer characteristics and their opinions on McDonald's food and service. We identified the most and least appreciated features, as well as the correlations between different variables.
2. **PCA** allowed us to reduce the complexity of the data by identifying the principal components that explain most of the variance. This helped us to visualize the data more simply and to prepare the ground for cluster analysis.
3. **Clustering**, both K-means and hierarchical, allowed us to identify distinct groups of customers with similar preferences and characteristics. We described the profiles of these clusters, highlighting their demographic characteristics, purchasing behaviors, and specific preferences.

Although the fast-food clientele has a good opinion of the food served, the company should pay more attention to the older age groups, investing in quality and not just in competitive pricing.

This project offered us a unique opportunity to apply theoretical and technical knowledge to a real and tangible case. We learned how to tackle the different phases of a data science project, from data cleaning and preparation to the interpretation of results.

Working on this project has opened our eyes to the potential that machine learning techniques can offer, showing us how these methods can be used to solve real-world problems and to obtain valuable information.