

# Customer Segmentation in Shopping Malls: Analysis and Marketing Strategies


Sebastián David Pinzón Zambrano

Data Scientist

## Resumen

The study analyzes customer segmentation in shopping malls by means of the K-means algorithm, using a Kaggle dataset with demographic (age, gender, annual income) and behavioral (spending score) variables. After rigorous preprocessing (standardization and winsorization) and validation with indices such as the elbow method and the silhouette coefficient, six clusters were identified in a unisex analysis, highlighting young people with high income and high spending, while segmentation by gender revealed 11 clusters in men and 6 in women, with profiles such as young men with high spending and young women with high predisposition to consume despite low income. The results validate the usefulness of K-means for designing personalized marketing strategies, although limitations such as the size of the dataset and the absence of psychographic variables are pointed out, proposing future research with these variables and other algorithms to delve deeper into consumer behavior.

## 1 Introduction

N In the competitive retail environment, a deep understanding of customer characteristics and behaviors is essential to designing effective marketing strategies. The present research is based on a dataset provided by Kaggle, available at <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>, which includes demographic (age, gender, annual income) and shopping behavior (spending score) information of 200 customers of a shopping mall, obtained through membership cards. The variables analyzed include the customer's unique identifier, age, gender, annual income expressed in thousands of dollars, and a spending score assigned on a scale of 1 to 100, based on consumption patterns.

The main objective of this study is to profile customers, identifying those with the greatest buying potential, and to show that, even within the same gender, there are subgroups with differentiated behaviors. To achieve this, we use the K-means clustering algorithm, a robust unsupervised learning tool that segments the clientele into homogeneous and detailed profiles. This segmentation facilitates the detection of high-value customers, improves the understanding of diversity in consumption patterns and enables the formulation of personalized marketing strategies, thus optimizing resources in the management of the shopping center.

The analysis is developed at two levels: a global unisex approach and a specific segmentation by gender. In the global analysis, six optimal clusters are identified using a multimetric approach combining validation indices such as the elbow method, the silhouette coefficient, the Dunn index, the Calinski-Harabasz index and the Davies-Bouldin index. For men, 11 clusters are determined, reflecting greater heterogeneity, while for women 6 clusters are identified, indicating a greater concentration in consumption patterns. These segmentations are supported by advanced preprocessing techniques, including standardization to homogenize the scales of the variables and winsorization to treat outliers, ensuring the robustness of the results.

The findings reveal high-value segments, such as young men with high income and high spending, and young women with a high predisposition to consume despite low income. In addition, significant correlations are observed, such as an inverse relationship between age and expenditure in women (-0.4), absent in men, which underlines the relevance of considering gender in segmentation. These results not only enrich the

understanding of purchase dynamics, but also provide a solid basis for the application of personalized marketing strategies in real retail environments.

Furthermore, the study pursues a dual purpose: to demonstrate the practical applicability of machine learning techniques in solving real-world problems and to serve as a methodological guide for researchers and practitioners interested in data science. Through a rigorous methodology and a detailed interpretation of emerging patterns, it seeks to provide a comprehensive view of consumer behavior, highlighting the importance of accurate segmentation that recognizes the inherent heterogeneity of the customer base.

Finally, although the project is developed in an academic and experimental context, the results lay a solid foundation for business applications. Accurate identification of profiles with high conversion probability, supported by multi-metric analysis and advanced techniques, can make a significant difference in competitiveness and business success, opening new perspectives for resource optimization and personalization in retail.

## 2 State of the Art



Review of the literature in the field of customer segmentation in shopping malls reveals a constant evolution in the application of quantitative and qualitative methodologies to understand consumer behavior. Several studies have approached this topic from multiple perspectives, offering a broad overview of the problems to be solved, the methodologies employed, the results obtained and the limitations encountered.

### 2.1 Problems addressed in previous studies

The analyzed scientific works have identified several key issues, such as:

- **Understanding and defining customer segments:** Research such as Gilboa (2009) has focused on developing typologies that allow the identification of homogeneous groups based on behaviors and demographic characteristics. Similarly, recent studies have attempted to classify customers not only by their demographic data but also by integrating psychographic and behavioral variables (for example, “Mall Enthusiasts” or “Deal Hunters”), which is essential for directing personalized marketing strategies.
- **Adaptation to changing economic and technological contexts:** Some works have highlighted how changes in the economic environment, especially during periods of crisis, modify purchasing behavior. Studies focused on “profiling shopping mall customers during hard times” emphasize the need for dynamic segmentations that respond to the evolution of market conditions and digital transformation.
- **Limitations of traditional methods:** The literature shows that classical approaches based solely on demographic data or traditional clustering models (such as k-means) may lack interpretability or fail to fully capture the complexity of customer behaviors. This issue has motivated the integration of deep learning methods and explainable analysis (for example, through models like DeepLimeSEG) to improve both the accuracy and interpretation of the obtained segments.

### 2.2 Methodologies and Applied Approaches

The reviewed studies have used various techniques to address segmentation, among which the following stand out:

- **Traditional clustering techniques:** Many authors have employed algorithms such as k-means and hierarchical clustering to identify segments based on variables such as age, annual income, and spending score. These methods allow for an initial classification of customers and have been used in both classical and recent research to delineate consumption profiles in shopping centers.
- **Integration of RFM analysis and advanced techniques:** The combination of RFM analysis (Recency, Frequency, and Monetary Value) with clustering algorithms has been recurrent, as it allows for capturing behavioral patterns more precisely. Some studies have complemented this methodology

with association rules and conjoint analysis to identify product bundles and the specific preferences of each segment.

- **Application of deep learning and explainable models:** In view of the limitations of traditional methods, innovative approaches have been proposed that integrate neural networks with explainable artificial intelligence techniques (XAI). These models not only improve segmentation accuracy but also facilitate the interpretation of the key factors in purchasing behavior, providing a more robust framework for targeted marketing strategies.
- **Multi-method approaches and comparative analysis:** Several studies have opted to combine quantitative and qualitative methods to obtain a holistic view of the consumer. This includes the integration of statistical analyses (for example, ANOVA, regressions) with case studies and international comparisons, which enrich the understanding of cultural and contextual particularities in customer segmentation.

## 2.3 Main Findings and Limitations

Among the most relevant results, the following stand out:

- **Identification of differentiated segments:** The literature has shown the existence of customer groups with very diverse behaviors and characteristics, ranging from those with a high propensity to spend and high purchasing power to segments with more moderate consumption patterns or those resistant to economic changes.
- **Importance of psychographic factors:** Recent studies emphasize that the incorporation of psychographic variables (attitudes, perceptions, and motivations) significantly improves the quality of segmentation, allowing for a classification that goes beyond demographics and better reflects consumer preferences.
- **Challenges in model interpretability and generalization:** Despite methodological advances, it is recognized that some approaches, especially those based on deep learning, may present difficulties in interpreting the results. Moreover, the generalization of the findings to different contexts and formats of shopping centers continues to be a challenge, which opens the door to future research to adapt and validate these models in varied environments.

## 2.4 Connection with the Present Research

In this context, the current research positions itself as an innovative contribution, since it:

- Proposes the integration of advanced machine learning techniques and explainable methods for customer segmentation, seeking to overcome the limitations identified in previous studies.
- Provides a differentiated perspective by combining demographic and psychographic variables, which allows for generating a more robust and contextually adapted profile of potential customers in shopping centers.
- Focuses on validating the applicability of these methodologies in both academic and business environments, with the aim of providing practical tools that optimize marketing strategies in the sector.

This state of the art not only demonstrates a deep knowledge of the topic and the diversity of existing approaches, but also justifies the relevance and originality of the present study. It highlights the need to continue exploring integrated models that allow for obtaining more interpretable and generalizable segmentations, responding to current market demands and adding value to strategic decision-making.

## 3 Methodology



This section provides a detailed description of the methodological approach used to analyze and segment customer behavior. Starting with a dataset obtained from Kaggle, which includes demographic and

spending variables, the process begins with rigorous preprocessing: cleaning the data, identifying and treating outliers through winsorization, and standardizing to ensure the integrity and homogeneity of the information.

Subsequently, an exploratory analysis is performed to identify key patterns and relationships among the variables, supported by visualization techniques and statistical validation. The determination of the optimal number of clusters is carried out by integrating various validation indices—such as the elbow method, silhouette coefficient, Dunn index, Calinski-Harabasz, and Davies-Bouldin indices—ensuring robust segmentation.

Finally, the KMeans algorithm is implemented to group customers into homogeneous segments, facilitating the identification of profiles with high conversion potential and providing a solid foundation for applying personalized marketing strategies and machine learning techniques in real-world commercial environments.

### 3.1 Dataset

The present study is based on the analysis of a dataset obtained from Kaggle, which is available at the following link: <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>. This dataset consists of 200 observations and 5 variables: **CustomerID**, **Age**, **Annual Income (k\$)**, and **Spending Score (1-100)**.

The structure of the dataset is described in Table 1, which details the name of each variable, its data type, and its respective explanation. This resource has been developed for educational and experimental purposes, aimed at teaching customer segmentation techniques and market analysis using the KMeans clustering algorithm. Although it is assumed that the data comes from a real environment, likely a shopping mall or supermarket, specific details such as the collection date or exact location are not provided.

The dataset is shared through a GitHub repository, which highlights its educational purpose and the intention to provide researchers and students with a practical tool for applying machine learning methods in customer segmentation.

Table 1: Column Format, Data Types, and Explanations

| Column                 | Data Type | Explanation  |
|------------------------|-----------|--|
| CustomerID             | int64     | Unique identifier assigned to each customer.   |
| Gender                 | object    | Customer’s gender.   |
| Age                    | int64     | Customer’s age.  |
| Annual Income (k\$)    | int64     | Customer’s annual income, expressed in thousands of dollars.                                 |
| Spending Score (1-100) | int64     | Score assigned by the shopping center based on the customer’s spending behavior and pattern. |

### 3.2 Preprocessing

Data preprocessing is a fundamental stage in any analysis, as it ensures that the information to be used is of the highest quality and free from errors. In this process, techniques using **pandas** were implemented to identify and remove erratic data and outliers, such as **NaN**, **inf**, **-inf**, and completely empty rows. In our case, no inconsistencies or empty rows were detected, so the dataset remained intact.

Below is a comparison of the dataset’s status before and after preprocessing (see Table 2):

Subsequently, a detailed analysis was conducted to detect outliers in the **Annual Income (k\$)**, **Age**, and **Spending Score (0-100)** columns. The distribution of these data was explored using a histogram and a boxplot, providing a comprehensive understanding of their behavior (see Figure 1).

The analysis identified and removed two outlier values present in the **Annual Income (k\$)** column. It is important to note that no outliers were found in the **Age** and **Spending Score (0-100)** columns. A summary

Table 2: Comparison of Data Status

| Status | Rows | Columns |
|--------|------|---------|
| Before | 200  | 5       |
| After  | 200  | 5       |

of this process is detailed in Table 3.

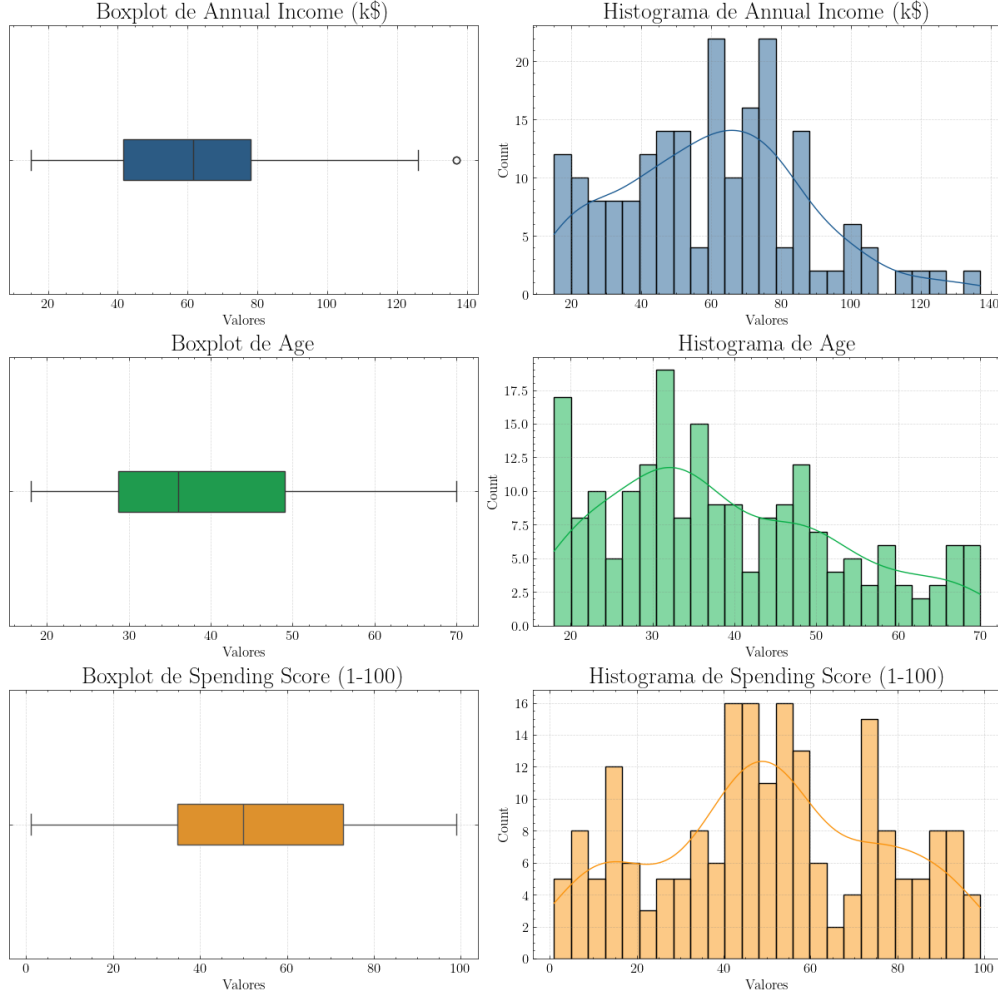


Figure 1: Data distribution represented through a histogram and a boxplot.

Finally, after an exhaustive process of data cleaning and analysis, standardization was performed using the **StandardScaler** technique. This procedure ensures that all variables are on the same scale, which not only facilitates subsequent analysis but also optimizes the performance of the customer segmentation models. In this way, a rigorous and consistent treatment of the data is ensured, laying the groundwork for future studies and analyses with a high level of precision and reliability.

Additionally, a comparative analysis of the modifications applied to the **Age**, **Annual Income (k\$)**, **Spending Score (0-100)**, and **Gender** variables was performed. Through detailed visualization, the changes experienced by these variables before and after standardization are illustrated, providing a clear representation of the impact of this process on the data structure (see Figure 2).

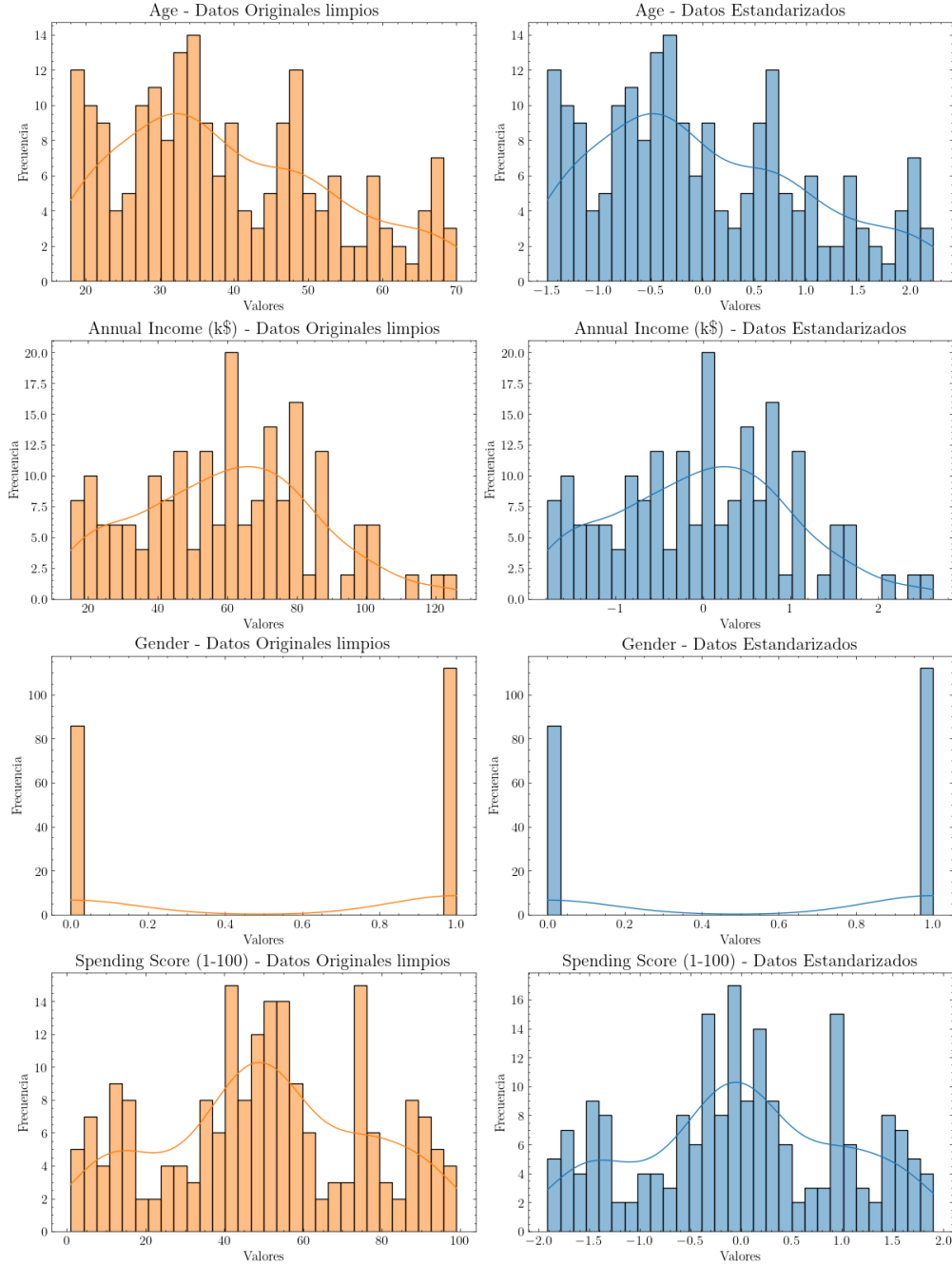


Figure 2: Comparative visualization of data before and after standardization.

Table 3: Eliminated Outliers

| Column                 | Number of Outliers |
|------------------------|--------------------|
| Annual Income (k\$)    | 2                  |
| Age                    | 0                  |
| Spending Score (0-100) | 0                  |
| Total rows eliminated  | 2                  |

### 3.3 Exploratory Analysis

We will begin exploring the data using a correlation map or "Heatmap" as an initial tool. This instrument will allow us to identify significant correlations that could reveal patterns not immediately evident, as well as help us avoid collinearity issues among variables.

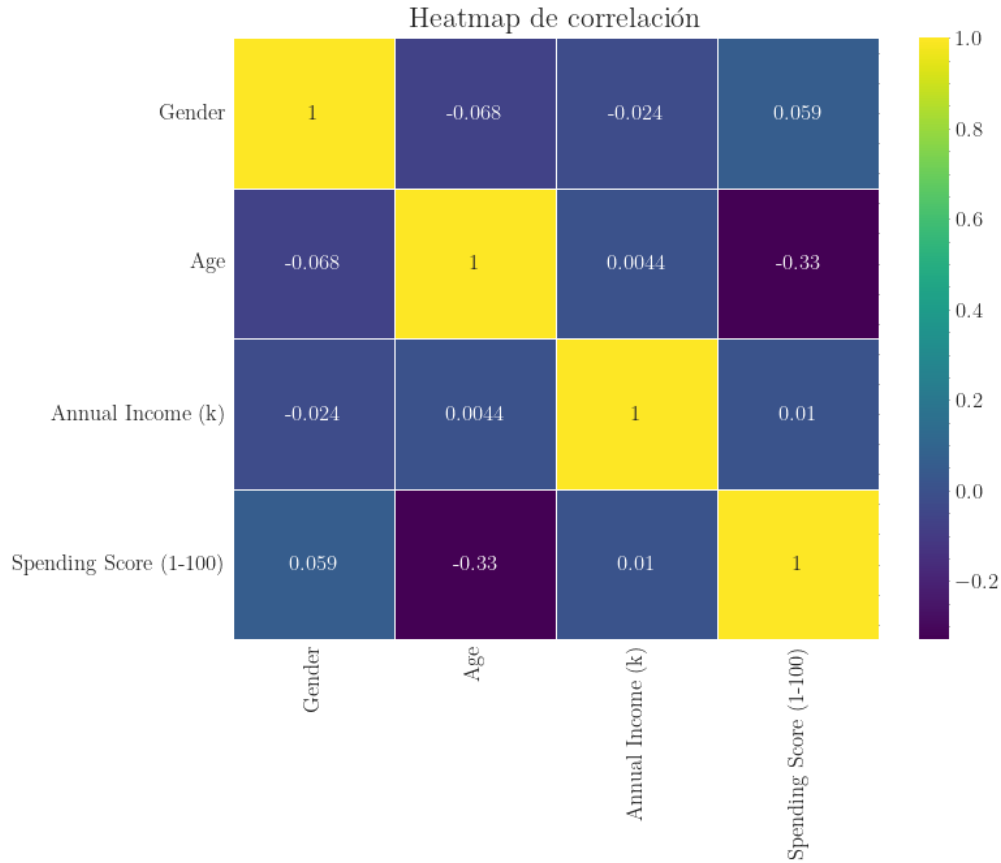


Figure 3: Visualization of the heatmap for the analysis of correlations among variables.

Figure 3 presents the resulting heatmap. Through this analysis, no collinearity problems were detected among the studied variables. However, a negative correlation between "Age" and "Spending Score (0-100)" is notable. Although this correlation is relatively weak, it suggests that as a customer's age increases, their spending score tends to decrease.

It is important to note that this study is conducted on the entire sample, including both men and women, so the exploratory analysis considers the complete dataset.

To examine this hypothesis in greater detail, we calculated the average Spending Score for each age group, which allows us to visualize the spending pattern in relation to the consumer's age. In this analysis, we implemented a linear regression to determine the overall trend of the relationship, complemented with a

LOWESS (Locally Weighted Scatterplot Smoothing) smoother to reduce noise in the data.

Once the LOWESS smoothing was applied, we complemented our analysis with a histogram that represents the percentage change relative to the previous point, calculated from the smoothed values. This approach allows us to identify trend changes more clearly, as shown in Figure 4.

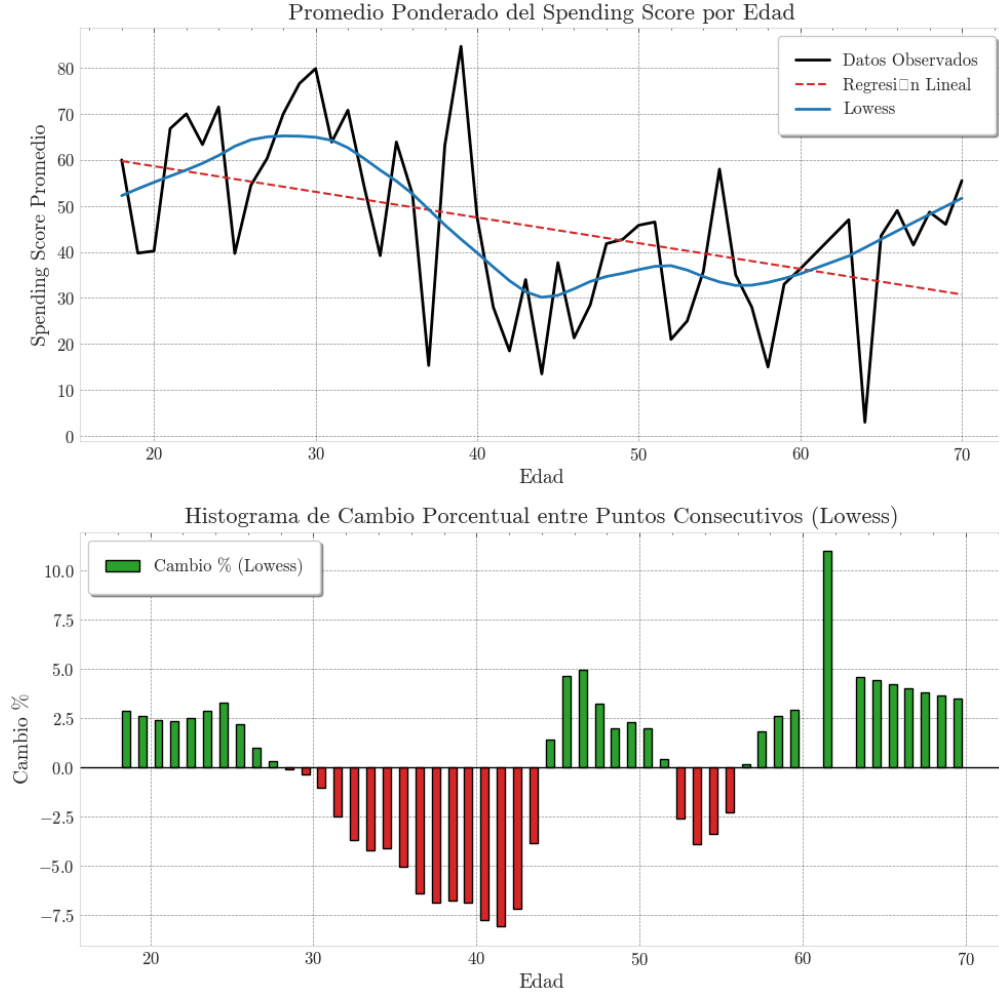


Figure 4: Variation of the Spending Score by age with LOWESS smoothing and percentage change histogram.

The initial hypothesis suggesting an inversely proportional relationship between age and spending is supported by the overall trend observed in the linear regression. However, when examining the data from a fractal perspective, the percentage change histogram reveals that this relationship is not uniform across the entire age spectrum. On the contrary, segments with both positive and negative trends are identified.

These trend ranges allow us to segment the age spectrum and draw conclusions about the age groups with the highest potential as customers, as summarized in Table 4.



Table 4: Spending Score Trend by Age Range

| Trend    | Age Range   | Weighted Average Spending Score |
|----------|-------------|---------------------------------|
| Upward   | 18.0 – 29.0 | 59.39                           |
| Downward | 28.0 – 45.0 | 50.74                           |
| Upward   | 44.0 – 53.0 | 32.38                           |
| Downward | 52.0 – 57.0 | 33.79                           |
| Upward   | 56.0 – 70.0 | 37.04                           |

From this initial exploration, we can infer that, despite the complexity revealed by the fractal analysis, there is clear evidence of age-related patterns in spending behavior. If the goal is to optimize resources by focusing on segments with a higher potential to convert into frequent customers, the data suggests prioritizing the age range between 18 and 29, while not neglecting the segment between 28 and 45. The recommended strategy would be to diversify the product offerings targeted at these age groups.

Regarding the age segment between 45 and 70, which exhibits a significantly lower spending behavior, a deeper study would be required to determine the underlying causes of this trend. However, such an analysis is beyond the scope of the present study due to the limitations of the available dataset.

This is only a preliminary observation; now we will proceed to use the K-means technique to uncover patterns in the dataset.

### 3.3.1 Global Unisex Study

Initially, we will examine the three-dimensional distribution of the data using the variables “Annual Income (k\$)”, “Age”, and “Spending Score (1-100)”. The exclusion of the “Gender” variable is aligned with our goal of conducting an integrated analysis without gender distinction, which will allow us to identify universal patterns in customer behavior. Including this variable would not only be unnecessary but would also add additional complexity without substantial analytical benefit.

Figure 5 shows the three-dimensional distribution of the analyzed dataset.

Prior to applying the K-means algorithm, it is essential to determine the optimal number of clusters that best represent the intrinsic structure of the data.

The scientific literature offers various methodologies for determining the optimal number of clusters. In this study, we have chosen a comprehensive approach that combines five validation indices, each with complementary strengths. This multimetric approach, in addition to providing a holistic view, allows us to avoid tie situations by using an odd number of indicators.

Figure 6 presents the evolution of the five validation indices as a function of the number of clusters. It is important to note that an iterative analysis was performed, evaluating from 2 up to 101 clusters to obtain a complete perspective of the indices’ behavior.

The indices used were: the elbow method (Inertia), the Silhouette Coefficient, the Dunn Index, the Calinski-Harabasz Index, and the Davies-Bouldin Index.

To delimit the search area for optimal clusters, the elbow index was used to identify the starting point from the inflection of its curve, which indicated that the optimal initial value was  $k = 4$ .

Subsequently, the search range was extended up to  $k = 15$ . The endpoint of this area was determined by the convergence observed in the following indices:

- **Dunn Index.**
- **Calinski-Harabasz Index.**
- **Davies-Bouldin Index.**

### Grafico de dispersion (Dataframe Unisex)

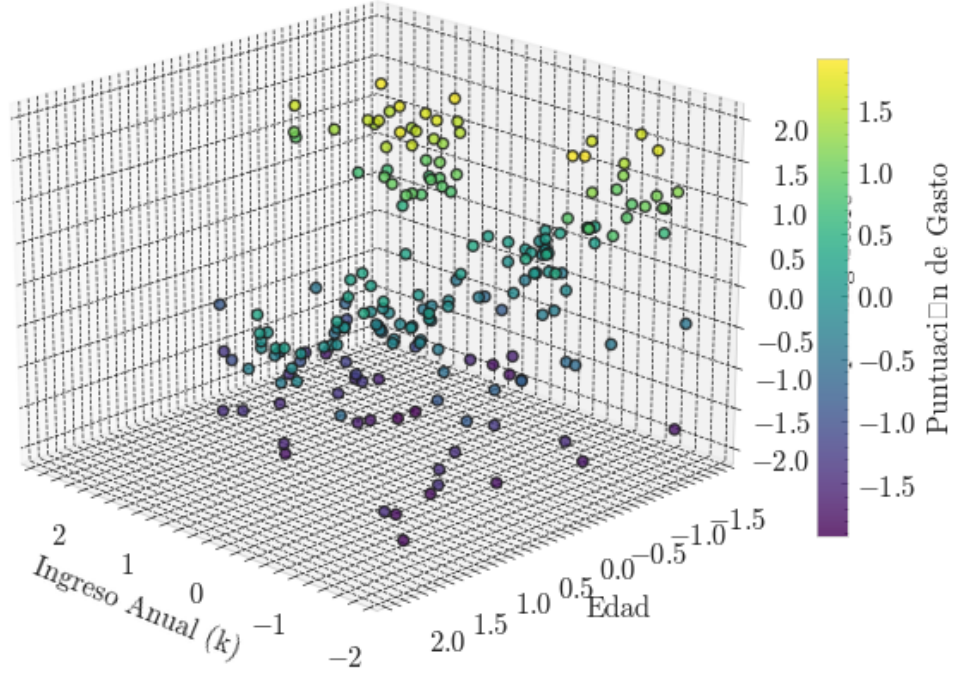


Figure 5: Three-dimensional visualization of the data distribution according to annual income, age, and spending score.

With the increase in  $k$ , a marked drawdown was observed in these indicators, which allowed defining  $k = 15$  as the upper limit of the area of interest.

It is important to clarify that this procedure was based on a visual and intuitive interpretation of the graphs, and not on a rigorous statistical calculation.

Figure 7 illustrates this refined analysis. Table 15 summarizes the results obtained for each metric, presenting the three optimal clusters identified by each index.

Table 5: **Optimal Clusters by Metric**

| Metric                 | Optimal Cluster 1 | Optimal Cluster 2 | Optimal Cluster 3 |
|------------------------|-------------------|-------------------|-------------------|
| Silhouette Coefficient | 6                 | 7                 | 10                |
| Dunn Index             | 4                 | 5                 | 6                 |
| Calinski-Harabasz      | 6                 | 9                 | 11                |
| Davies-Bouldin         | 6                 | 7                 | 9                 |

The convergence analysis of the obtained results suggests that the optimal number of clusters for the dataset is six ( $k = 6$ ). Once this value was determined, the  $k$ -means algorithm was implemented.

Figure 8 shows the resulting classification of the data into the six identified clusters.

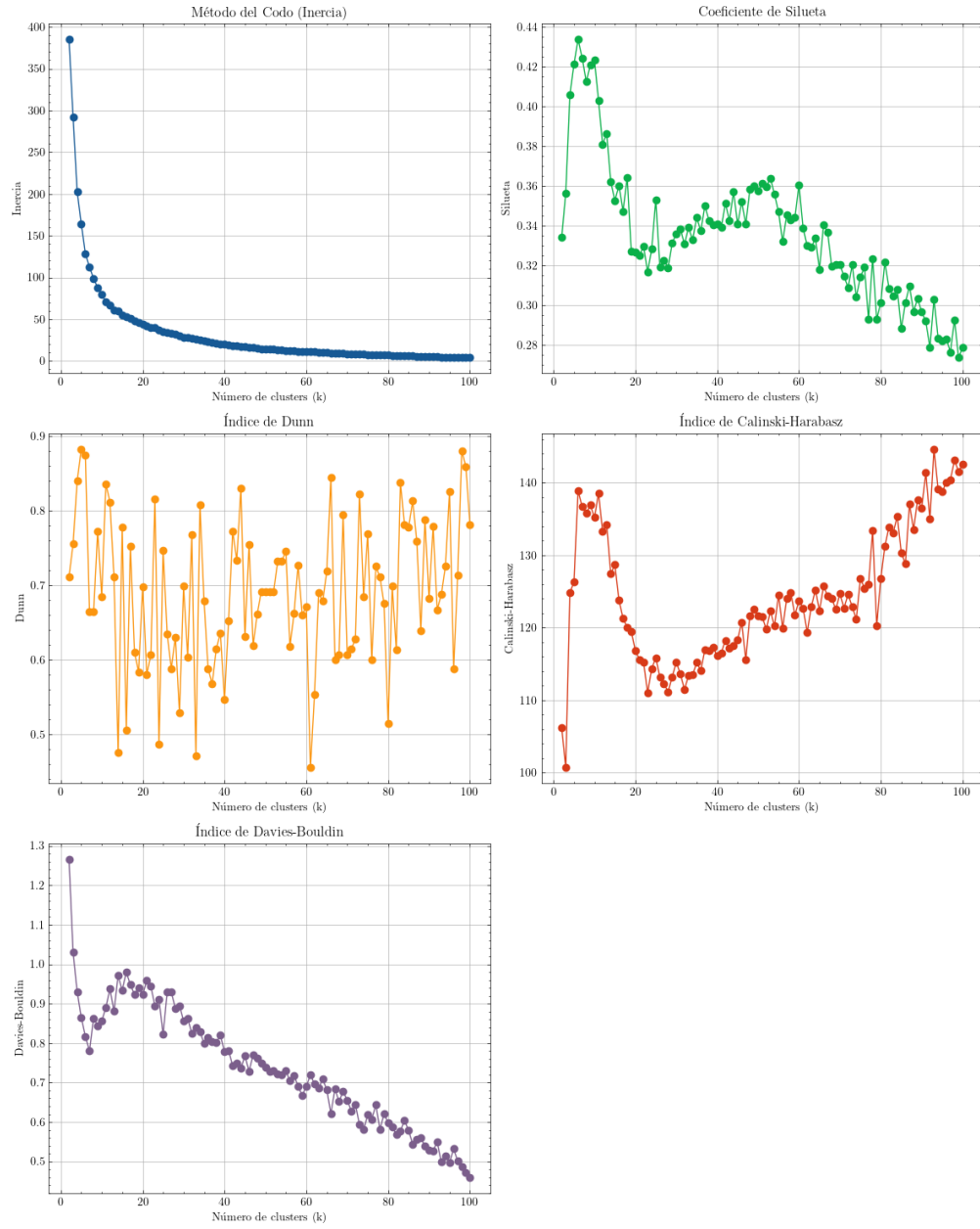


Figure 6: Evolution of the cluster validation indices as a function of the number of groupings.

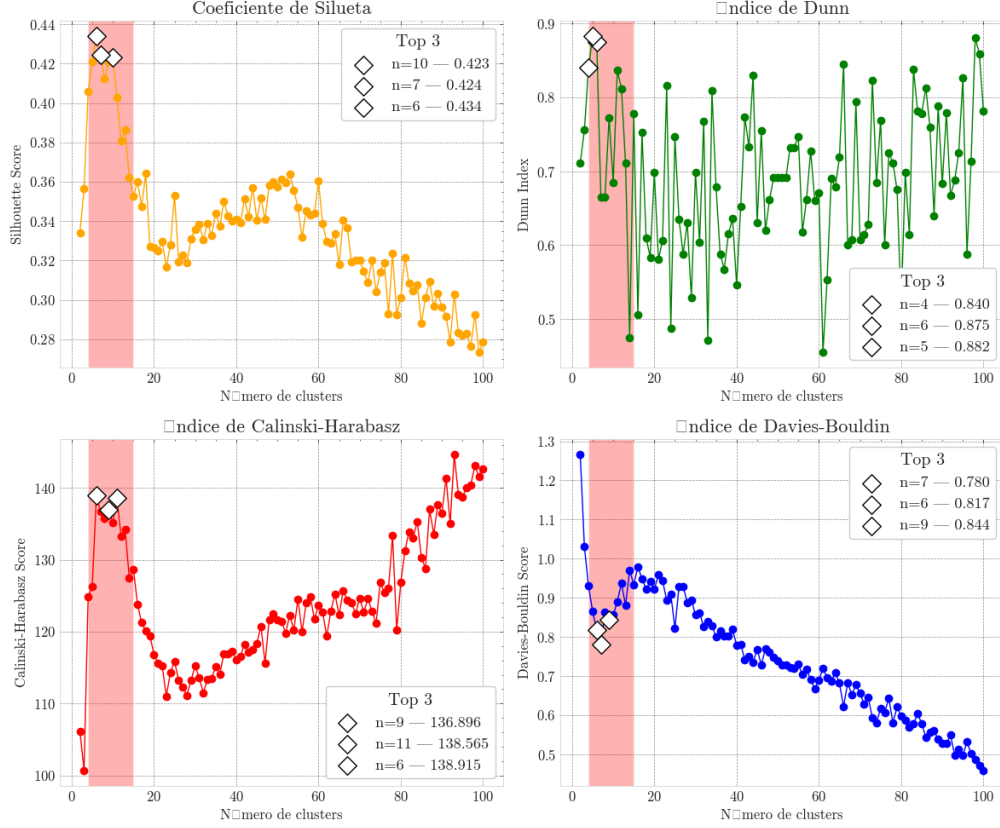


Figure 7: Identification of optimal values for each validation index in the area of interest.

To precisely characterize each cluster, a descriptive statistical analysis was developed covering measures of central tendency, such as the mean and median, as well as the standard deviation. In addition, the skewness coefficient was included to assess the symmetry of the data distribution in each grouping.

The inclusion of the skewness analysis is based on several methodologically significant considerations, allowing for a more complete and robust interpretation of the clusters' structure.

Table 6: Statistics by Unisex Cluster

| Cluster | Age          |             |              | Annual Income (k\$) |              |              | Spend Score  |              |              | Age (skew)  | Income (skew) | Spend Score (skew) |
|---------|--------------|-------------|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|-------------|---------------|--------------------|
|         | mean         | std         | median       | mean                | std          | median       | mean         | std          | median       | skew        | skew          | skew               |
| 0       | 56.34        | 8.55        | 54.00        | 53.70               | 8.24         | 54.00        | 49.39        | 5.99         | 49.00        | 0.27        | -0.06         | -0.02              |
| 1       | <b>32.76</b> | <b>3.75</b> | <b>32.00</b> | <b>85.21</b>        | <b>14.24</b> | <b>78.50</b> | <b>82.11</b> | <b>9.49</b>  | <b>84.00</b> | <b>0.38</b> | <b>1.23</b>   | <b>-0.10</b>       |
| 2       | 25.56        | 5.44        | 24.00        | 26.48               | 8.53         | 25.00        | 76.24        | 13.56        | 76.00        | 0.56        | 0.30          | -0.52              |
| 3       | <b>26.12</b> | <b>7.03</b> | <b>25.00</b> | <b>59.42</b>        | <b>10.59</b> | <b>60.00</b> | <b>44.45</b> | <b>14.28</b> | <b>48.00</b> | <b>0.71</b> | <b>-0.12</b>  | <b>-1.65</b>       |
| 4       | <b>44.80</b> | <b>8.04</b> | <b>44.00</b> | <b>88.20</b>        | <b>14.52</b> | <b>86.50</b> | <b>18.50</b> | <b>10.37</b> | <b>16.50</b> | <b>0.33</b> | <b>1.00</b>   | <b>0.31</b>        |
| 5       | 45.52        | 11.77       | 46.00        | 26.29               | 7.44         | 25.00        | 19.38        | 12.56        | 15.00        | -0.08       | 0.34          | 0.17               |

The detailed results of this statistical characterization of the clusters are presented in Table 6, offering a comprehensive view of the distinctive properties of each grouping identified by the  $k$ -means algorithm.

As can be seen, there are inconsistencies in the skewness values in clusters 1, 3, and 4, since these values exceed the threshold of 0.7 or fall below the threshold of -0.7, which could indicate asymmetric distributions.

Since skewness is a factor that can affect the quality of the analysis, and considering that eliminating these data would be methodologically inappropriate, the winsorization technique was applied to adjust the outlier values. Figure 9 illustrates the comparison between the original data and those modified after applying

## Dataframe clasificado mediante K-means

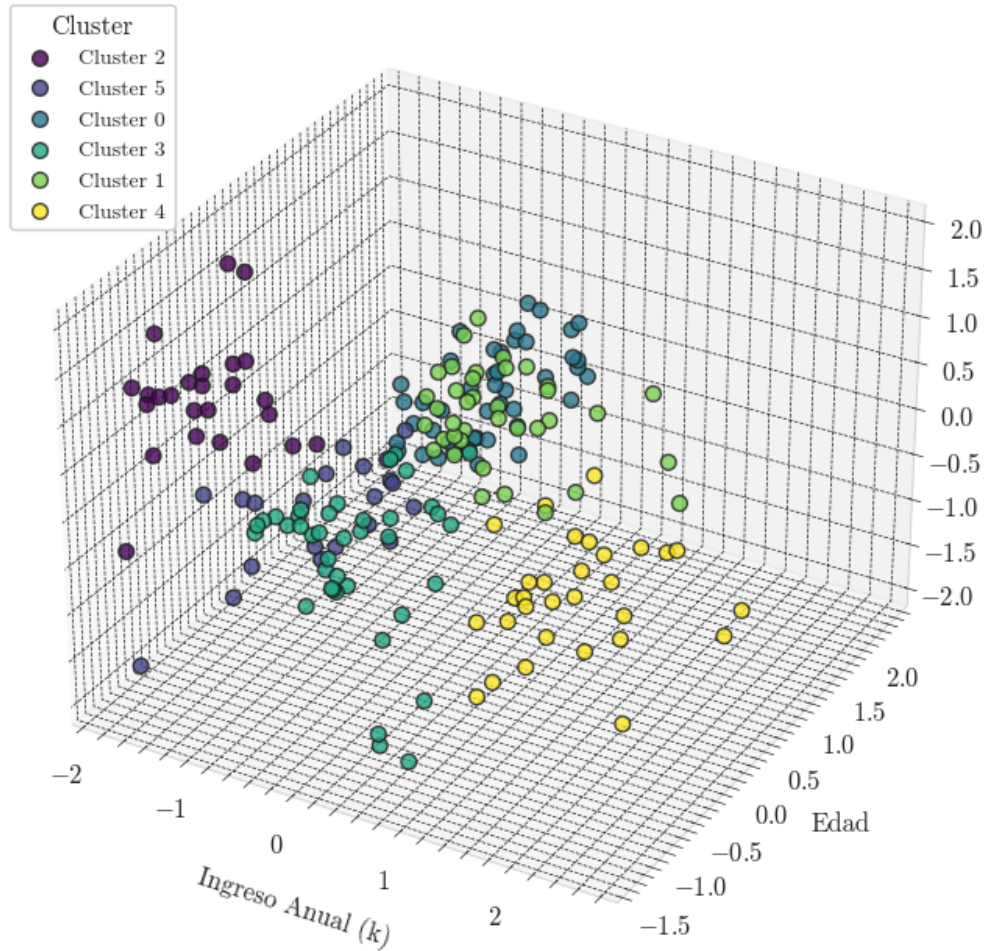


Figure 8: Visualization of the data classification into six clusters using the  $k$ -means algorithm.

winsorization.

Table 7 details the transformations applied to each cluster and variable. As can be seen, winsorization and outlier detection were employed to correct the identified asymmetries. Table 8 shows the statistical results after applying these transformations.

With the data now transformed and presenting more appropriate skewness indices, we proceeded to generate a heatmap for each cluster to identify any additional patterns, as shown in Figure 10.

The heatmap analysis reveals two particularly relevant negative correlations, located in clusters 3 and 5:

1. Cluster 3 shows a negative correlation of -0.55 between “Spending Score” and “Annual Income”, indicating that within this group, individuals with incomes between \$49,000 and \$69,000 and ages between 18 and 32 tend to exhibit purchasing behavior that is inversely proportional to their age.
2. Cluster 5 exhibits a negative correlation of -0.33 between “Spending Score” and “Age”, suggesting that individuals aged between 45 and 57 with incomes between \$19,000 and \$33,000 tend to decrease their purchasing behavior as they age.

Our initial conclusion was that our target audience was between 18 and 45 years old, combining both ranges

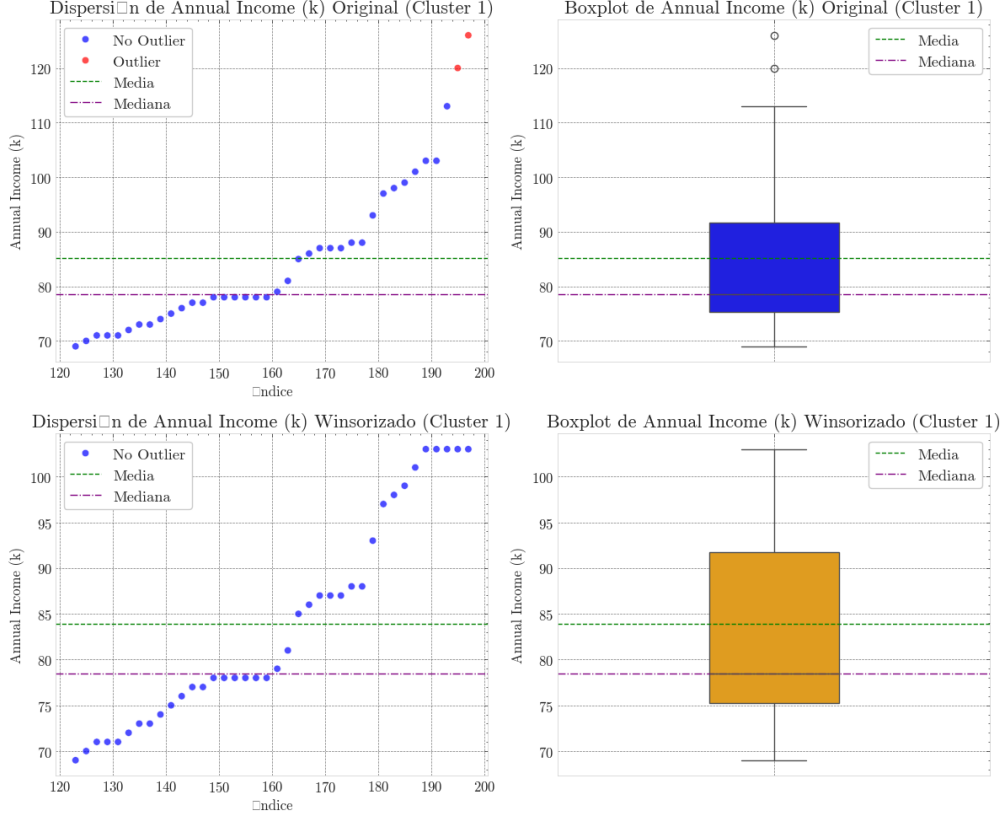


Figure 9: Comparison of distributions before and after winsorization.

with a spending score greater than 50. However, the income range required for these individuals to be considered high-potential customers was never specified. Thanks to the observed correlations (see Table 9), it is now possible to evidence the significant influence of *Annual Income* in this context.

The new hypothesis establishes that such individuals must have an income below \$49,000 or above \$69,000. To reach this conclusion, it will be necessary to verify it by exploring those individuals who register higher levels of spending relative to their incomes (see Table 9).

We can conclude, therefore, that the hypothesis — according to which individuals earning less than \$49,000 or more than \$69,000, and who are between 18 and 45 years old, are more likely to be potential customers — is valid. Thus, the overall profile of a potential customer is defined as individuals between 18 and 45 years old who earn an annual income below \$49,000 or above \$69,000.

It should be noted that this profile will only be useful in cases where the individual’s gender is unknown, but information on age and annual income is available.

### 3.3.2 Dataset Division by Gender: Rationale and Justification

Dividing the dataset into two subgroups—one for male customers and one for female customers—is based on the need to capture the particularities and nuances that each segment presents in terms of behavior and consumption characteristics. The following provides a detailed and well-founded explanation for this segmentation:

1. **Differences in Behavior and Preferences:** Previous studies and empirical evidence indicate that, in many consumption contexts, men and women may exhibit different spending patterns and decision-making processes. By separating the dataset, the analysis of specific variables (such as *Spending Score* and *Annual Income*) is facilitated, allowing for the identification of correlations and trends particular to

Table 7: Summary of Transformations Applied to Each Cluster and Variable

| Cluster | Variable               | Winsorization | IQR Factor | Change Description   |
|---------|------------------------|---------------|------------|--|
| 1       | Annual Income (k)      | [0, 0.1]      | 1.5        | The top 10% of extreme values was removed and outliers were detected with an IQR factor of 1.5 (Cluster 1).    |
| 3       | Spending Score (1-100) | [0.1, 0]      | 1.5        | The bottom 10% of extreme values was removed and outliers were detected with an IQR factor of 1.5 (Cluster 3). |
| 3       | Age                    | [0, 0.2]      | 0.5        | The top 20% of extreme values was removed and outliers were detected with an IQR factor of 0.5 (Cluster 3).    |
| 4       | Annual Income (k)      | [0, 0.1]      | 1.0        | The top 10% of extreme values was removed and outliers were detected with an IQR factor of 1 (Cluster 4).      |

Table 8: Statistics by Unisex Cluster (after Winsorization)

| Cluster | Age   |       |        | Annual Income (k\$) |       |        | Spend Score |       |        | Age (skew) | Income (skew) | Spend Score (skew) |
|---------|-------|-------|--------|---------------------|-------|--------|-------------|-------|--------|------------|---------------|--------------------|
|         | mean  | std   | median | mean                | std   | median | mean        | std   | median | skew       | skew          | skew               |
| 0       | 56.34 | 8.55  | 54.00  | 53.70               | 8.24  | 54.00  | 49.39       | 5.99  | 49.00  | 0.27       | -0.06         | -0.02              |
| 1       | 32.76 | 3.75  | 32.00  | 83.89               | 11.31 | 78.50  | 82.11       | 9.49  | 84.00  | 0.38       | 0.56          | -0.10              |
| 2       | 25.56 | 5.44  | 24.00  | 26.48               | 8.53  | 25.00  | 76.24       | 13.56 | 76.00  | 0.56       | 0.30          | -0.52              |
| 3       | 25.02 | 5.21  | 25.00  | 59.42               | 10.59 | 60.00  | 46.55       | 9.31  | 48.00  | 0.11       | -0.12         | -0.56              |
| 4       | 44.80 | 8.04  | 44.00  | 86.53               | 11.20 | 86.50  | 18.50       | 10.37 | 16.50  | 0.33       | 0.26          | 0.31               |
| 5       | 45.52 | 11.77 | 46.00  | 26.29               | 7.44  | 25.00  | 19.38       | 12.56 | 15.00  | -0.08      | 0.34          | 0.17               |

each gender. This enriches the analysis and enables the development of more tailored and personalized marketing and segmentation strategies.

- Improved Profiling Accuracy:** The inherent heterogeneity in consumption behaviors is reinforced by the diversity of socioeconomic and cultural factors that influence each gender differently. By analyzing men and women separately, a more accurate profiling of potential customers is achieved, leading to the identification of segments with a high probability of conversion and the optimization of resources by focusing strategies on the specific characteristics of each subgroup.
- Methodological and Academic Considerations:** From a methodological standpoint, dividing the dataset helps reduce internal variability and potential confounding effects that arise when mixing heterogeneous groups. This practice is consistent with the application of advanced clustering techniques, such as K-means, as a homogeneous analysis within each subgroup ensures that the validation indices and the cohesion and separation metrics are more representative of each gender’s behavior. Moreover, this segmentation supports the applicability of winsorization methods and more robust statistical analysis, strengthening the interpretability of the results obtained.
- Impact on Strategic Decision-Making:** Identifying specific patterns in each gender facilitates the development of targeted marketing strategies. For instance, by observing that certain age ranges or income levels are associated differently with spending in each segment, personalized campaigns can be designed to maximize customer acquisition and retention. This segmented approach is crucial in a competitive environment, where the adaptation and personalization of strategies become key differentiators for commercial success.

In summary, dividing the dataset addresses a dual need: to improve the quality and accuracy of the analysis by reducing heterogeneity, and to provide a solid basis for generating potential customer profiles that accurately reflect the behavioral differences between men and women. This approach not only enhances the academic rigor of the study but also offers a practical tool for implementing effective and personalized marketing strategies.

Heatmap de Correlación por Cluster (Dataframe Unisex)

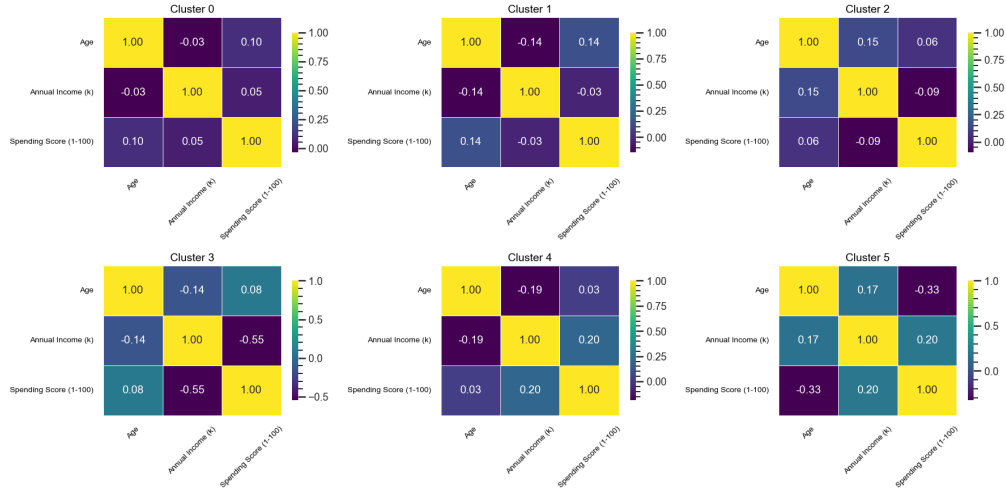


Figure 10: Heatmap of correlations between variables for each cluster.

| Income Range | Average Spending |
|--------------|------------------|
| 0–49         | 58.96            |
| 50–68        | 49.25            |
| 69+          | 57.67            |

Table 9: Simple average spending by income group (ages 18–32)

### 3.3.3 Profiling of Male Customers

After segmenting the dataset, we proceed with the analysis of the subset corresponding to male customers, which is composed of 86 records. Although this number does not represent half of the total, it is sufficient to conduct a robust and meaningful study.

In the first instance, the correlations between the variables are examined to identify possible collinearities and relationships of interest, as illustrated in Figure 11.

Although a correlation of -0.28 was observed between the *Spending Score (1-100)* variable and other variables, this value is not considered significant since it does not reach the threshold for a moderate correlation.

Next, the k-means method is applied to uncover underlying patterns in the data. Initially, the optimal number of clusters is determined using various statistical indices, the results of which are summarized in Table 10.

In Figure 12 these indices are shown, highlighting the delimited area and the positions of the optimal clusters according to each criterion.

Table 10: Optimal Clusters by Metric (range [4,15])

| Metric                 | Optimal Cluster 1 | Optimal Cluster 2 | Optimal Cluster 3 |
|------------------------|-------------------|-------------------|-------------------|
| Silhouette Coefficient | 5                 | 9                 | 11                |
| Dunn Index             | 4                 | 5                 | 11                |
| Calinski-Harabasz      | 10                | 11                | 12                |
| Davies-Bouldin         | 8                 | 11                | 14                |



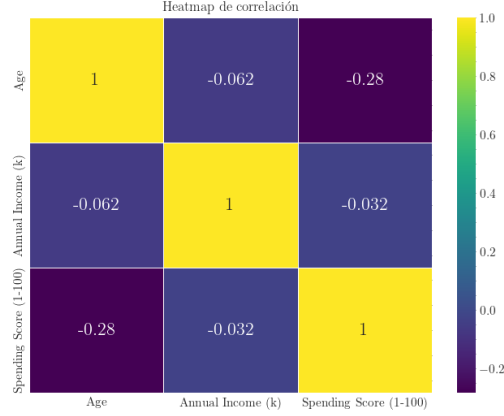


Figure 11: Heatmap of correlations between variables (male customers).

According to the various indices, it was determined that the optimal number of clusters is 11. With this parameter, the k-means model was trained and descriptive statistics were extracted for each group, as detailed in Table 11.

| Cluster | Age          |              |              | Annual Income (k\$) |              |              | Spend Score  |              |              | Skew         |               |                |
|---------|--------------|--------------|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|----------------|
|         | Mean         | Std          | Median       | Mean                | Std          | Median       | Mean         | Std          | Median       | Age          | Annual Income | Spending Score |
| 0       | <b>52.78</b> | <b>4.74</b>  | <b>53.00</b> | <b>55.22</b>        | <b>7.08</b>  | <b>54.00</b> | <b>49.22</b> | <b>5.63</b>  | <b>47.00</b> | <b>0.27</b>  | <b>-0.69</b>  | <b>0.97</b>    |
| 1       | <b>35.17</b> | <b>4.00</b>  | <b>35.50</b> | <b>78.50</b>        | <b>9.57</b>  | <b>76.00</b> | <b>88.00</b> | <b>8.70</b>  | <b>90.50</b> | <b>-0.45</b> | <b>1.24</b>   | <b>-0.95</b>   |
| 2       | <b>39.78</b> | <b>5.09</b>  | <b>40.00</b> | <b>86.22</b>        | <b>12.71</b> | <b>86.00</b> | <b>10.44</b> | <b>6.60</b>  | <b>10.00</b> | <b>0.10</b>  | <b>1.23</b>   | <b>-0.24</b>   |
| 3       | <b>21.60</b> | <b>3.69</b>  | <b>19.50</b> | <b>56.90</b>        | <b>7.42</b>  | <b>59.50</b> | <b>52.60</b> | <b>5.76</b>  | <b>54.50</b> | <b>0.53</b>  | <b>-0.37</b>  | <b>-0.95</b>   |
| 4       | <b>24.40</b> | <b>5.62</b>  | <b>23.00</b> | <b>24.70</b>        | <b>7.39</b>  | <b>24.50</b> | <b>73.80</b> | <b>15.80</b> | <b>76.00</b> | <b>0.80</b>  | <b>0.35</b>   | <b>-1.11</b>   |
| 5       | 40.57        | 5.74         | 40.00        | 46.71               | 14.76        | 43.00        | 44.29        | 10.23        | 41.00        | 0.08         | 0.20          | 0.65           |
| 6       | <b>55.50</b> | <b>10.82</b> | <b>56.50</b> | <b>24.00</b>        | <b>6.07</b>  | <b>21.50</b> | <b>11.17</b> | <b>9.99</b>  | <b>8.50</b>  | <b>-0.99</b> | <b>0.85</b>   | <b>1.33</b>    |
| 7       | <b>29.40</b> | <b>2.41</b>  | <b>28.00</b> | <b>97.80</b>        | <b>16.84</b> | <b>88.00</b> | <b>69.80</b> | <b>4.87</b>  | <b>69.00</b> | <b>0.47</b>  | <b>1.66</b>   | <b>-0.38</b>   |
| 8       | <b>20.75</b> | <b>2.87</b>  | <b>19.50</b> | <b>76.25</b>        | <b>3.59</b>  | <b>75.50</b> | <b>8.00</b>  | <b>3.56</b>  | <b>7.50</b>  | <b>1.85</b>  | <b>0.89</b>   | <b>0.27</b>    |
| 9       | <b>54.80</b> | <b>5.36</b>  | <b>58.00</b> | <b>82.80</b>        | <b>8.79</b>  | <b>85.00</b> | <b>20.40</b> | <b>10.41</b> | <b>15.00</b> | <b>-0.66</b> | <b>-0.38</b>  | <b>1.01</b>    |
| 10      | 67.22        | 2.33         | 67.00        | 54.67               | 8.12         | 54.00        | 50.11        | 6.09         | 51.00        | -0.50        | -0.06         | -0.12          |

Table 11: Statistics by cluster (male clients). Rows where any skew value exceeds 0.7 or is below -0.7 are highlighted in bold.

Table 12 meticulously details the modifications made and the IQR factors used to detect outliers in each cluster. Subsequently, Table 14 presents the data after winsorization.

Knowing the distribution of the data, the correlations between the variables within each cluster are examined below to further explore the classification conditions, as illustrated in Figure 13.

Based on the correlations shown in Figure 13 and the descriptive statistics, it is possible to draw conclusions of special interest:

#### Clusters with high income and high spending:

- **Cluster 1** (31.2–39.2 years): It exhibits high incomes (70.3–83.7) and very high spending (88.0–94.8).
- **Cluster 7** (27.0–31.8 years): It records even higher incomes (85.3–100.3) and moderately high spending (64.9–74.7).

These groups represent consumers with high purchasing power and intense consumption patterns.

#### Clusters with low income and low spending:

- **Cluster 6** (51.4–64.6 years): It is characterized by reduced incomes (17.9–30.1) and limited spending (3.2–14.2).
- **Cluster 2** (34.7–44.9 years): Although its incomes are high (75.3–93.8), the spending is extremely low (3.8–17.0).

While Cluster 6 might represent people nearing retirement with certain economic limitations, Cluster 2 suggests consumers with a high capacity to save and a low propensity to spend.

Table 12: Variables and parameters used for outlier elimination (Clusters 0, 1, 2, 3, 4, 6, 7, 8, and 9)

| Index | Variable               | Interval | Factor | Cluster | Description   |
|-------|------------------------|----------|--------|---------|---|
| 0     | Spending Score (1-100) | [0, 0.2] | 1.5    | 0       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 1.5. (Cluster 0)    |
| 1     | Annual Income (k)      | [0, 0.2] | 1      | 1       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 1. (Cluster 1)      |
| 2     | Spending Score (1-100) | [0.3, 0] | 1      | 1       | The bottom 30% of extreme values was removed and outliers were detected using an IQR factor of 1. (Cluster 1)   |
| 3     | Annual Income (k)      | [0, 0.2] | 1.5    | 2       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 1.5. (Cluster 2)    |
| 4     | Spending Score (1-100) | [0.1, 0] | 1      | 3       | The bottom 10% of extreme values was removed and outliers were detected using an IQR factor of 1. (Cluster 3)   |
| 5     | Spending Score (1-100) | [0.1, 0] | 1.5    | 4       | The bottom 10% of extreme values was removed and outliers were detected using an IQR factor of 1.5. (Cluster 4) |
| 6     | Age                    | [0, 0.1] | 1.5    | 4       | The top 10% of extreme values was removed and outliers were detected using an IQR factor of 1.5. (Cluster 4)    |
| 7     | Age                    | [0.2, 0] | 1      | 6       | The bottom 20% of extreme values was removed and outliers were detected using an IQR factor of 1. (Cluster 6)   |
| 8     | Spending Score (1-100) | [0, 0.2] | 1.5    | 6       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 1.5. (Cluster 6)    |
| 9     | Annual Income (k)      | [0, 0.2] | 0.5    | 6       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 0.5. (Cluster 6)    |
| 10    | Annual Income (k)      | [0, 0.2] | 1.5    | 7       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 1.5. (Cluster 7)    |
| 11    | Age                    | [0, 0.3] | 1      | 8       | The top 30% of extreme values was removed and outliers were detected using an IQR factor of 1. (Cluster 8)      |
| 12    | Annual Income (k)      | [0, 0.3] | 0.5    | 8       | The top 30% of extreme values was removed and outliers were detected using an IQR factor of 0.5. (Cluster 8)    |
| 13    | Spending Score (1-100) | [0, 0.2] | 0.5    | 9       | The top 20% of extreme values was removed and outliers were detected using an IQR factor of 0.5. (Cluster 9)    |

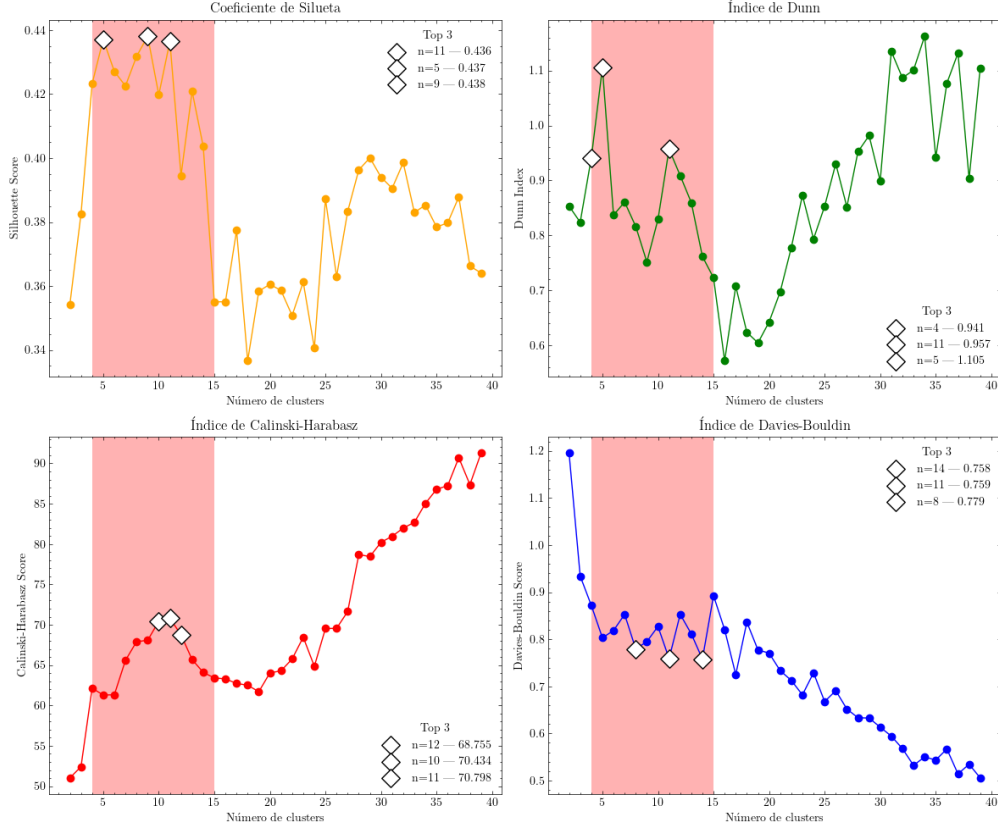


Figure 12: Indices for determining the optimal number of clusters.

- **Cluster 10** (64.9–69.6 years): It is characterized by having both income and spending at moderate levels.

These groups are associated with mature individuals exhibiting more conservative financial behaviors.

#### Notable correlations:

- **Positive (green):** These are observed in Cluster 2 (age and income are positively correlated, despite low spending) and in Clusters 3 and 4 (young people with high spending).
- **Negative (red):** These are evident in Clusters 6 and 9, suggesting that in older ages there tends to be lower spending.
- **Auto-correlation (orange):** This is noticeable in Clusters 1 and 3, indicating consistent internal patterns within these segments.

#### Segments of greatest commercial interest: Cluster 1 (31.2–39.2 years):

- High incomes: 70.3–83.7.
- Very high spending: 88.0–94.8.
- Positive correlation between income and spending.

This segment is especially attractive due to its high purchasing power and marked propensity to consume. It likely comprises young adults with financial stability and a strong desire to acquire goods and services.

#### Cluster 4 (18.8–30.0 years):

Heatmap de Correlación por Cluster

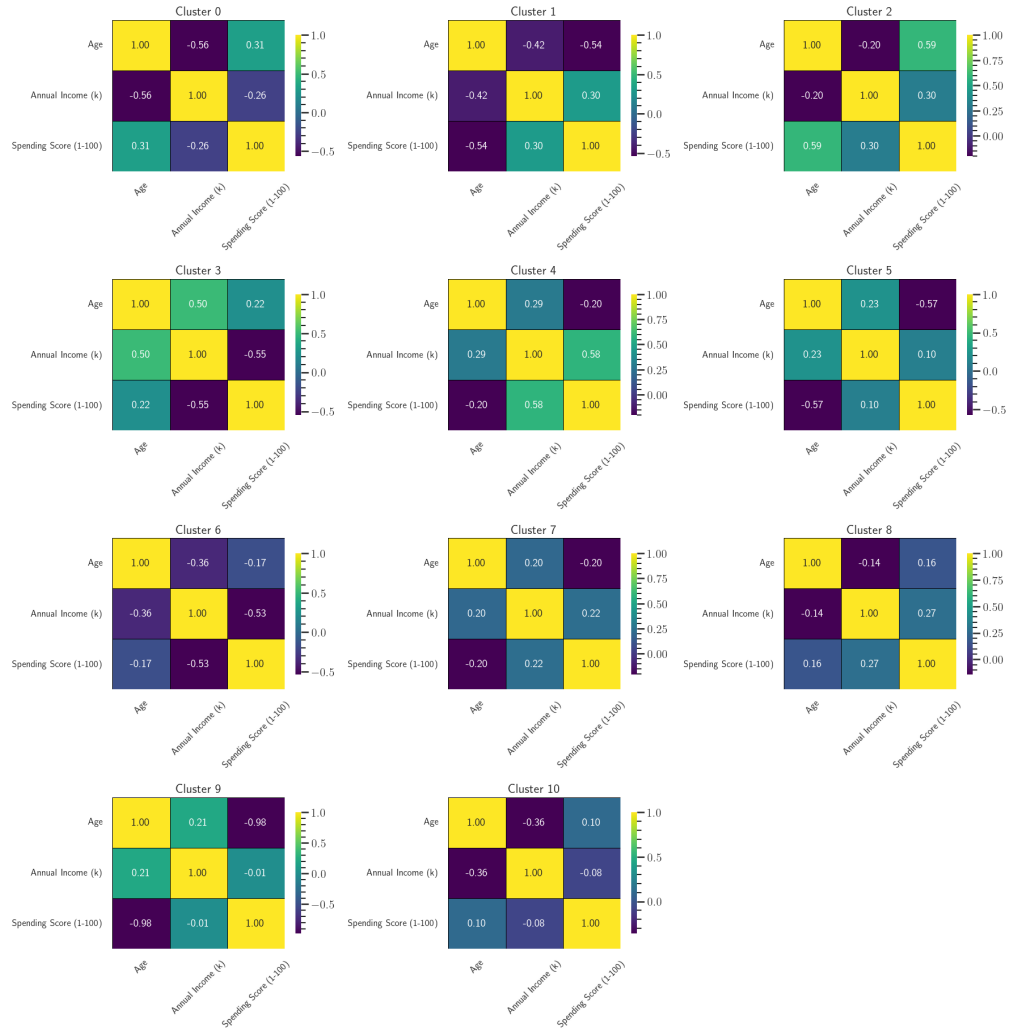


Figure 13: Heat map of correlations between variables for each cluster.

| Cluster | Age   |      |        | Annual Income (k\$) |       |        | Spend Score |       |        | Skew  |               |             |
|---------|-------|------|--------|---------------------|-------|--------|-------------|-------|--------|-------|---------------|-------------|
|         | Mean  | Std  | Median | Mean                | Std   | Median | Mean        | Std   | Median | Age   | Annual Income | Spend Score |
| 0       | 52.78 | 4.74 | 53.00  | 55.22               | 7.08  | 54.00  | 48.78       | 4.76  | 47.00  | 0.27  | -0.69         | 0.57        |
| 1       | 35.17 | 4.00 | 35.50  | 77.00               | 6.70  | 76.00  | 91.42       | 3.42  | 90.50  | -0.45 | 0.64          | 0.63        |
| 2       | 39.78 | 5.09 | 40.00  | 84.56               | 9.28  | 86.00  | 10.44       | 6.60  | 10.00  | 0.10  | 0.31          | -0.24       |
| 3       | 21.60 | 3.69 | 19.50  | 56.90               | 7.42  | 59.50  | 53.10       | 4.77  | 54.50  | 0.53  | -0.37         | -0.43       |
| 4       | 24.00 | 4.88 | 23.00  | 24.70               | 7.39  | 24.50  | 76.00       | 11.30 | 76.00  | 0.45  | 0.35          | 0.07        |
| 5       | 40.57 | 5.74 | 40.00  | 46.71               | 14.76 | 43.00  | 44.29       | 10.23 | 41.00  | 0.08  | 0.20          | 0.65        |
| 6       | 58.00 | 6.60 | 56.50  | 23.50               | 5.24  | 21.50  | 8.67        | 5.50  | 8.50   | 0.41  | 0.67          | -0.00       |
| 7       | 29.40 | 2.41 | 28.00  | 92.80               | 7.50  | 88.00  | 69.80       | 4.87  | 69.00  | 0.47  | 0.60          | -0.38       |
| 8       | 19.50 | 0.58 | 19.50  | 75.25               | 2.06  | 75.50  | 8.00        | 3.56  | 7.50   | 0.00  | -0.20         | 0.27        |
| 9       | 54.80 | 5.36 | 58.00  | 82.80               | 8.79  | 85.00  | 18.40       | 7.09  | 15.00  | -0.66 | -0.38         | 0.41        |
| 10      | 67.22 | 2.33 | 67.00  | 54.67               | 8.12  | 54.00  | 50.11       | 6.09  | 51.00  | -0.50 | -0.06         | -0.12       |

Table 14: Cluster statistics (male clients) after winsorization.

- Low incomes: 17.3–32.1.
- High spending: 64.7–87.3.
- A positive correlation in spending and a negative correlation in income are observed.

Despite their limited incomes, this group shows a notable inclination to consume, which could indicate the presence of students or young professionals willing to prioritize spending over saving. Thus, it constitutes a potential market for aspirational or affordable luxury products.

**Cluster 3** (17.9–25.3 years):

- Medium incomes: 49.5–64.3.
- Medium-high spending: 48.3–57.9.
- High correlation between income and spending.

This segment probably groups young individuals with stable employment, inclined to invest in technology, fashion, or entertainment. With incomes higher than those in Cluster 4, their dependency on credit is potentially lower.

### 3.3.4 Profiling of Female Customers

After profiling the potential customers, a specific analysis of the female gender is carried out, in which the passion for science and reason merges with the art of understanding human behaviors. The method employed is analogous to that used for the male gender; however, a particular phenomenon has been identified: the correlation between the *spending score* and age reaches approximately  $-0.4$  (see Figure 14). This finding, which contrasts with both the analysis of the male gender and the overall heatmap, suggests that as age increases, women tend to show less interest in products. Far from being a mere number, this result invites further analysis by incorporating new variables and using techniques such as *k-means*, in a sincere homage to science and the power of reasoning.

The determination of the optimal number of clusters was carried out based on various indicators (see Figure 15). The convergence of these indicators suggests the formation of 6 clusters, which contrasts with the segmentation observed in the male gender. This lower diversity in the female dataset implies a greater concentration of information in each group, thereby strengthening the robustness and accuracy of the applied classification rules.

Table 15 summarizes the optimal points selected according to each metric used:

With the *k-means* algorithms trained for the optimal number of clusters, the classification condition statistics were examined (see Table 16). As expected, clusters with skewed distributions were identified, with the

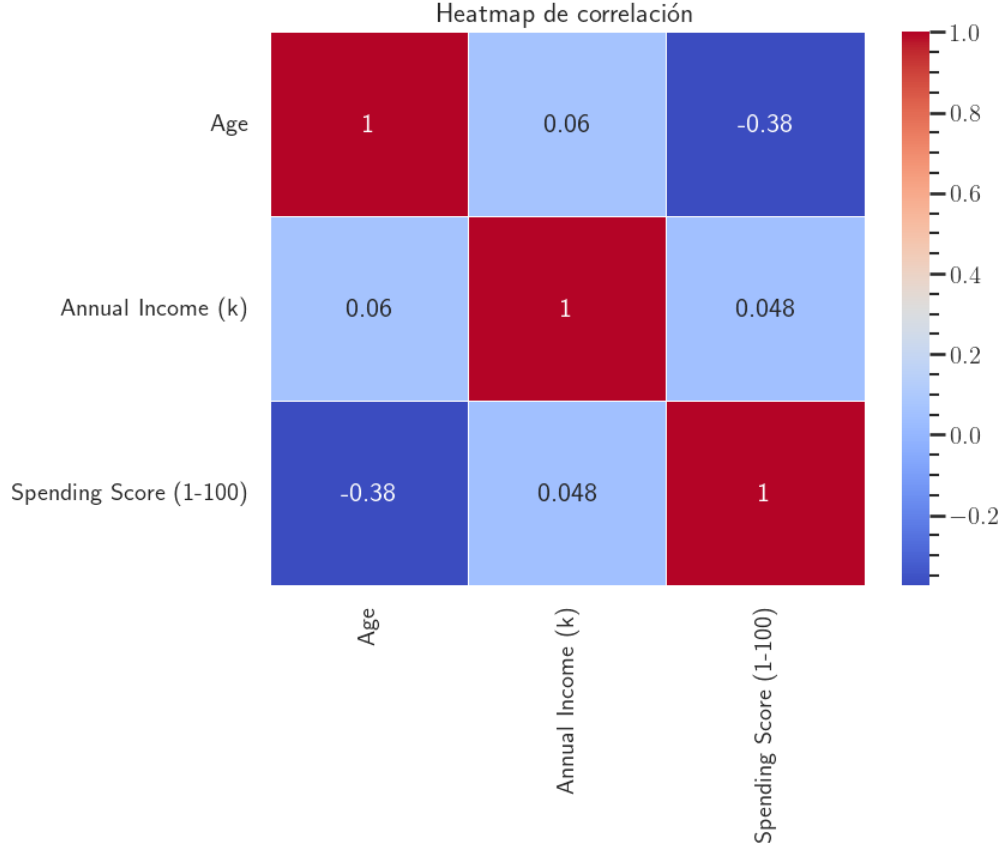


Figure 14: Heat map of correlations between variables for each cluster.

results corresponding to Clusters 2 and 3 highlighted in bold. To improve the symmetry of the groups, the winsorization technique was applied, whose results are shown in Table 17.

The correlation analysis using the heatmap (Figure 16) provides enriching conclusions about the behavior of the clusters based on age, annual income, and the *spending score*. In particular:

- **Differentiation of clusters by ranges:** Each cluster groups individuals who share specific ranges of age, income, and spending. For example, Cluster 2 brings together young women (20.8–28.6 years) with low annual incomes (17.5–33.9) and high spending (69.8–91.2), while Cluster 0 groups older women (46.1–62.1 years) with moderate incomes and spending.
- **Significant correlations:** In Cluster 2, a strong negative correlation (less than  $-0.4$ ) between annual income and the *spending score* is evident, suggesting that in this group, as income increases, spending tends to decrease. In contrast, Cluster 4 shows a significant positive correlation (greater than  $0.4$ ) between age and annual income, indicating that in this group, age is associated with a slight increase in

Table 15: Optimal Clusters by Metric

| Metric                 | Optimal Cluster 1 | Optimal Cluster 2 | Optimal Cluster 3 |
|------------------------|-------------------|-------------------|-------------------|
| Silhouette Coefficient | 6                 | 7                 | 10                |
| Dunn Index             | 4                 | 5                 | 6                 |
| Calinski-Harabasz      | 6                 | 9                 | 11                |
| Davies-Bouldin         | 6                 | 7                 | 9                 |

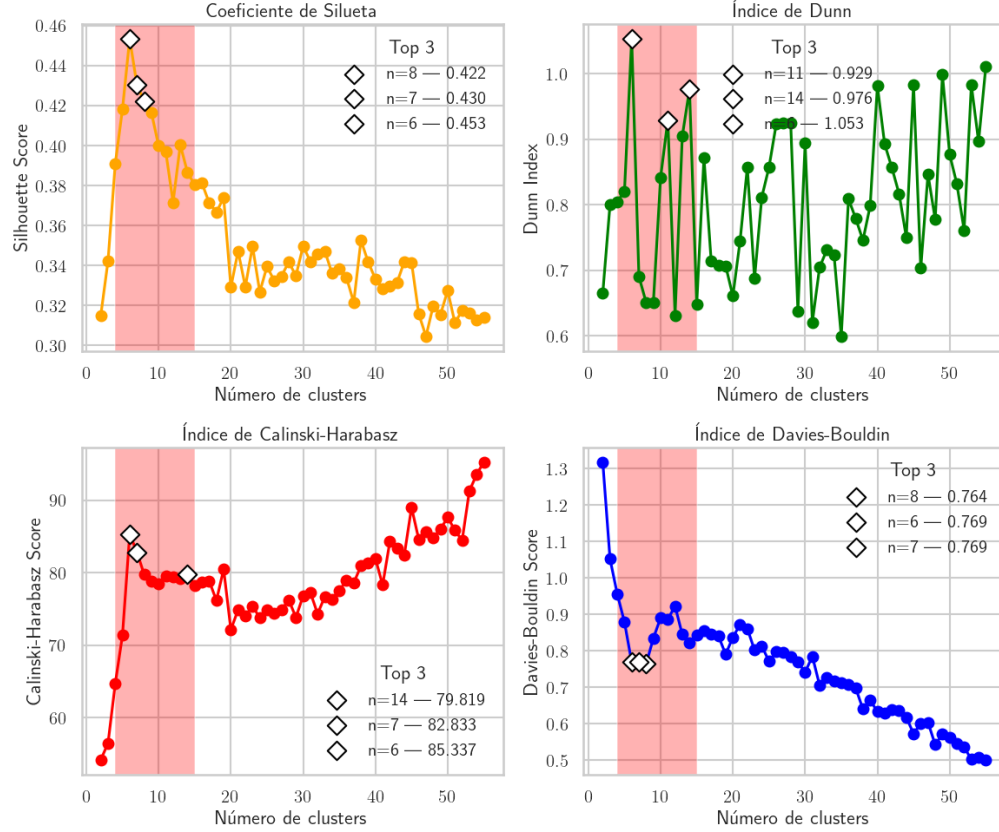


Figure 15: Heat map of correlations between variables for each cluster.

income. Clusters 0, 1, 3, and 5 do not present statistically relevant correlations among the analyzed variables.

- **Interpretation of correlations in orange:** The correlations highlighted in orange indicate that, within the same cluster, one variable maintains strong links with two different relationships. This phenomenon can be interpreted as a sign of consistency in the behavior of that variable, reflecting the underlying harmony in the analysis.
- **Practical applications:** These findings allow for the design of segmented marketing strategies. For example, in Cluster 2, campaigns could be developed targeted at young women with low incomes but a high propensity to spend, incentivizing consumption through offers or promotions. On the other hand, Cluster 4, with its positive association between age and income, could be the focus of strategies oriented towards experience and financial stability.

Finally, two groups of special interest for the commercial strategy are highlighted:

| Cluster | Age          |             |              | Annual Income (k\$) |              |              | Spend Score  |              |              | Skew        |               |              |
|---------|--------------|-------------|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|-------------|---------------|--------------|
|         | Mean         | Std         | Median       | Mean                | Std          | Median       | Mean         | Std          | Median       | Age         | Annual Income | Spend Score  |
| 0       | 54.08        | 7.97        | 50.00        | 53.24               | 8.76         | 54.00        | 49.52        | 6.19         | 50.00        | 0.67        | 0.04          | -0.27        |
| 1       | <b>44.60</b> | <b>7.66</b> | <b>44.00</b> | <b>92.33</b>        | <b>16.44</b> | <b>88.00</b> | <b>21.60</b> | <b>9.70</b>  | <b>22.00</b> | <b>0.20</b> | <b>0.72</b>   | <b>0.04</b>  |
| 2       | <b>25.46</b> | <b>5.22</b> | <b>23.00</b> | <b>25.69</b>        | <b>8.24</b>  | <b>23.00</b> | <b>80.54</b> | <b>10.70</b> | <b>77.00</b> | <b>1.03</b> | <b>0.36</b>   | <b>0.70</b>  |
| 3       | <b>32.19</b> | <b>3.08</b> | <b>32.00</b> | <b>86.05</b>        | <b>14.16</b> | <b>79.00</b> | <b>81.67</b> | <b>7.83</b>  | <b>83.00</b> | <b>0.52</b> | <b>1.10</b>   | <b>-0.00</b> |
| 4       | 41.54        | 10.26       | 42.00        | 26.54               | 7.76         | 28.00        | 20.69        | 11.93        | 17.00        | -0.41       | 0.14          | 0.13         |
| 5       | 27.96        | 6.67        | 27.00        | 57.36               | 10.98        | 60.00        | 47.12        | 8.19         | 47.00        | 0.34        | -0.39         | -0.15        |

Table 16: Cluster Statistics (female customers)

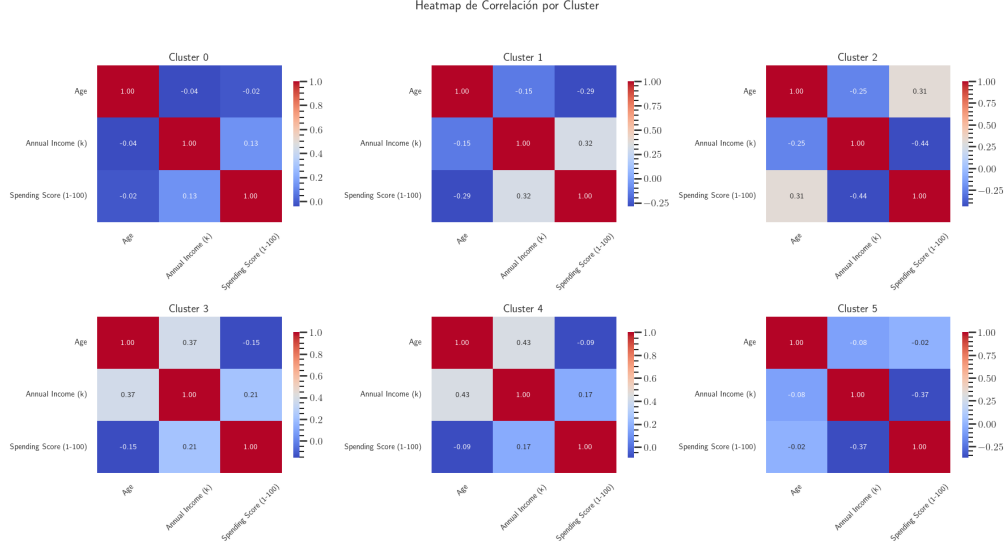


Figure 16: Heat map of correlations between variables for each cluster.

- **Cluster 3 (29.1–35.3 years):** This segment exhibits high incomes (73.2–96.3) and a high *spending score* (73.8–89.5), positioning it as an ideal candidate for premium products or services.
- **Cluster 2 (20.8–28.6 years):** Although its incomes are relatively low (17.5–33.9), this group shows notable spending behavior (69.8–91.2), evidencing a high propensity to consume. This makes it an attractive opportunity for promotional campaigns and affordable products, targeted at a young and dynamic audience.

In summary, the segmentation of the female gender not only differentiates customers according to ranges of age, income, and spending but also reveals significant relationships among these variables. This analysis, inspired by scientific rigor and a love for reason, paves the way for designing marketing strategies tailored to the particularities of each segment, merging technical precision with a deep sensitivity to human behavior.

### 3.4 Applied Methods

**I**N this section the applied methods are addressed, which constitute the backbone of the analysis. Techniques for clustering – with the K-means algorithm as the central axis – are combined with validation and optimization strategies, such as the elbow method, the silhouette coefficient, and the Dunn index, to ensure that the data partition reflects real and significant patterns. Additionally, winsorization is incorporated to mitigate the impact of outliers and improve the robustness of the analysis. This methodological integration not only guarantees precise segmentation but also establishes a smooth transition to the detailed explanation of each technique employed.

| Cluster | Age   |       |        | Annual Income (k\$) |       |        | Spend Score |       |        | Skew  |               |             |
|---------|-------|-------|--------|---------------------|-------|--------|-------------|-------|--------|-------|---------------|-------------|
|         | Mean  | Std   | Median | Mean                | Std   | Median | Mean        | Std   | Median | Age   | Annual Income | Spend Score |
| 0       | 54.08 | 7.97  | 50.00  | 53.24               | 8.76  | 54.00  | 49.52       | 6.19  | 50.00  | 0.63  | 0.04          | -0.25       |
| 1       | 44.60 | 7.66  | 44.00  | 92.33               | 16.44 | 88.00  | 21.60       | 9.70  | 22.00  | 0.18  | 0.65          | 0.04        |
| 2       | 24.69 | 3.86  | 23.00  | 25.69               | 8.24  | 23.00  | 80.54       | 10.70 | 77.00  | 0.54  | 0.32          | 0.62        |
| 3       | 32.19 | 3.08  | 32.00  | 84.76               | 11.55 | 79.00  | 81.67       | 7.83  | 83.00  | 0.48  | 0.59          | -0.00       |
| 4       | 41.54 | 10.26 | 42.00  | 26.54               | 7.76  | 28.00  | 20.69       | 11.93 | 17.00  | -0.36 | 0.12          | 0.11        |
| 5       | 27.96 | 6.67  | 27.00  | 57.36               | 10.98 | 60.00  | 47.12       | 8.19  | 47.00  | 0.32  | -0.37         | -0.14       |

Table 17: Cluster Statistics (female customers) after winsorization



### 3.4.1 K-means Algorithm

The K-means algorithm is a clustering technique used to partition a dataset into  $K$  groups or clusters, so that the data within each cluster are as similar as possible to each other while maximizing the differences with respect to data in other clusters. From an academic perspective, its functioning can be described in the following terms:

**Theoretical Foundation** K-means falls within the field of unsupervised learning and is based on minimizing intra-cluster variation. The objective is to find the centers (centroids) of the clusters that minimize the total sum of squared distances between each data point and the centroid of the cluster to which it belongs. This metric, commonly defined using the Euclidean distance, is formally expressed as:

$$\arg \min_C \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2,$$

where  $C = \{C_1, C_2, \dots, C_K\}$  represents the partition of the dataset into  $K$  clusters, and  $\mu_i$  is the centroid vector of cluster  $C_i$ .

**Algorithmic Procedure** The algorithm is structured in the following steps:

1. **Initialization:**  $K$  initial centroids are selected. The choice can be made randomly or through heuristic methods, such as the K-means++ algorithm, to improve convergence and avoid local minima.
2. **Cluster Assignment:** Each data point in the dataset is assigned to the cluster whose centroid is closest in terms of Euclidean distance, thus establishing the current partition of the dataset.
3. **Centroid Update:** For each cluster, the centroid is recalculated as the mean of the points assigned to that cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x.$$

4. **Convergence:** The assignment and update steps are repeated iteratively until the cluster assignments do not change significantly between iterations or a predefined maximum number of iterations is reached. Convergence is generally defined when the reduction in the sum of squared distances is less than a set threshold.

### Methodological Considerations

- **Choice of the Number of Clusters ( $K$ ):** Selecting  $K$  is a critical parameter that significantly influences the quality of the segmentation. Techniques such as the elbow method, silhouette analysis, and information criteria (AIC, BIC) are used to determine an appropriate value.
- **Sensitivity to Initialization:** Since K-means can converge to suboptimal local solutions, it is common to run the algorithm multiple times with different initializations and select the partition that minimizes the objective function.
- **Assumption of Cluster Distribution and Shape:** K-means assumes that clusters are spherical and of similar size, which may limit its effectiveness in scenarios where clusters exhibit more complex geometries or heterogeneous densities.

**Applications in Customer Segmentation** In contexts such as customer segmentation, K-means is particularly well-suited for:

- Identifying groups of customers with similar behaviors and demographic or consumption characteristics.
- Facilitating the design of segmented marketing strategies, allowing targeted campaigns to be focused on clusters with high conversion potential.

- Reducing the dimensionality of the information, making it easier to interpret large volumes of data in a structured manner.

In conclusion, the K-means algorithm presents itself as a robust and efficient tool in exploratory data analysis, especially in applications where the identification of natural patterns and the segmentation of homogeneous populations are essential for developing strategies based on the knowledge extracted from the data.

### 3.4.2 Winsorization

Winsorization is a statistical method designed to mitigate the effect of extreme or outlier values in a dataset without eliminating them. This process preserves the integrity of the original sample and allows for more robust estimates of statistical parameters, such as the mean and variance. The procedure is based on the following steps:

1. **Selection of the Winsorization Level:** A percentage  $\alpha$  (for example, 0.05 or 0.10) is determined, which establishes the lower and upper thresholds for the transformation of the data.
2. **Calculation of the Quantiles:** Given a sample  $\{x_1, x_2, \dots, x_n\}$  ordered from smallest to largest, the following are calculated:
  - The lower quantile  $Q(\alpha)$ , defined as the value such that  $100\alpha\%$  of the observations are less than or equal to it.
  - The upper quantile  $Q(1 - \alpha)$ , defined as the value such that  $100(1 - \alpha)\%$  of the observations are less than or equal to it.
3. **Application of the Winsorized Transformation:** Each observation  $x_i$  is adjusted according to the following function:

$$x_i^{(w)} = \begin{cases} Q(\alpha) & \text{if } x_i < Q(\alpha), \\ x_i & \text{if } Q(\alpha) \leq x_i \leq Q(1 - \alpha), \\ Q(1 - \alpha) & \text{if } x_i > Q(1 - \alpha). \end{cases}$$

This implies:

- Replacing values below  $Q(\alpha)$  with  $Q(\alpha)$ .
  - Leaving unaltered those values that fall between  $Q(\alpha)$  and  $Q(1 - \alpha)$ .
  - Replacing values above  $Q(1 - \alpha)$  with  $Q(1 - \alpha)$ .
4. **Impact on Statistical Analysis:** Winsorization reduces the influence of outliers, leading to more stable and representative estimates of the central tendency and dispersion of the distribution. This method is especially useful in contexts where the presence of outliers could significantly distort the analysis and the conclusions drawn from the dataset.

### 3.4.3 Silhouette Index

The silhouette coefficient is a fundamental metric in cluster analysis, allowing for an integrated evaluation of the quality of a dataset's partition. Its value is especially useful for quantitatively determining the internal cohesion of clusters and the separation between them, both of which are essential aspects in any segmentation study.

**Theoretical Foundation and Calculation** For each observation  $i$ , two terms are defined:

- $a(i)$ : represents the average distance between observation  $i$  and all other observations belonging to the same cluster, reflecting the compactness or internal cohesion of the group.
- $b(i)$ : is defined as the average distance between observation  $i$  and all observations in the nearest cluster to which it does not belong, quantifying the separation between clusters.

The silhouette coefficient for the data point  $i$  is expressed by the formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

The value of  $s(i)$  is bounded within the interval  $[-1, 1]$ , where:

- **Values close to 1:** indicate that the observation is well-clustered, as the distance to its own cluster is significantly smaller than the distance to the nearest cluster.
- **Values close to 0:** suggest that the observation lies on the boundary between two clusters.
- **Negative values:** reveal that the observation might have been misassigned, since the distance to the nearest cluster is lower than that to its own cluster.

**Optimization of Clusters in the K-Means Algorithm** The silhouette coefficient is especially useful in optimizing the K-means algorithm, where the number of clusters  $k$  must be appropriately determined to obtain a meaningful partition of the data. Its application in this context allows:

- **Internal validation:** by averaging  $s(i)$  over all observations, a global measure of the clustering quality is obtained, facilitating comparisons between different partitions.
- **Detection of the inherent structure of the data:** a high average value suggests that the generated clusters are homogeneous and well-differentiated, while low values indicate the need to reconsider the number of clusters or the clustering method.
- **Iterative adjustment:** the metric can serve as a stopping criterion or adjustment parameter in iterative optimization processes, allowing fine-tuning of the K-means configuration and improving the interpretation of the results.

**Interpretation and Applications** The silhouette coefficient is used to evaluate both the individual placement of each observation and the overall quality of the clustering when averaged over the entire dataset. A high average value implies that, in general, the clusters exhibit a well-defined and separated structure, which is desirable in areas such as data mining, market segmentation, image analysis, among others. In contrast, low or negative values may indicate the need to adjust the number of clusters or review the method employed.

**Conclusion** In summary, the silhouette coefficient stands as a robust and versatile tool in cluster analysis. Its ability to synthesize information related to internal cohesion and separation between groups makes it an essential resource for both researchers specializing in statistics and data analysis, as well as for those who are new to the field and seek an intuitive and effective measure to validate their clustering models. Moreover, its application in the optimization of algorithms such as K-means allows for an objective determination of the optimal number of clusters, ensuring that the partitioning of the data is representative and meaningful.

#### 3.4.4 Elbow Method

The K-means algorithm seeks to partition a dataset into  $k$  clusters, optimizing the internal representation by minimizing the inertia or the sum of squared errors (WSS). The inertia is defined as:

$$WSS(k) = \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2,$$

where  $C_i$  denotes the set of points in the  $i$ -th cluster and  $\mu_i$  is its centroid. As  $k$  increases, the value of  $WSS(k)$  decreases, reflecting a better fit of the clusters to the structure of the data; however, beyond a certain threshold, the improvement becomes marginal.

**Procedure** The following scheme is used to apply the elbow method:

1. **Implementation and Computation:** K-means is executed for various values of  $k$  (for example, from 1 up to a reasonable upper limit), evaluating the value of  $WSS(k)$  in each case.
2. **Analysis of the Decrease in Inertia:** The evolution of  $WSS(k)$  is plotted as a function of  $k$ . Initially, a steep reduction in inertia is observed, followed by a plateau where the benefit of increasing  $k$  is marginal.
3. **Identification of the Inflexion Point:** The “elbow” is identified as the point at which the slope of the curve changes dramatically, indicating the optimal number of clusters. Beyond this point, the decrease in  $WSS(k)$  does not justify the additional complexity of the model.

**Optimization and Validation of the Segmentation** The elbow method is particularly useful in optimizing segmentation for the following reasons:

- **Balance between Accuracy and Complexity:** It allows the determination of a value of  $k$  that balances the reduction in inertia with model complexity, avoiding both overfitting (too many clusters capturing noise) and underfitting (insufficient representation of the data structure).
- **Computational Efficiency:** Identifying the optimal number of clusters prevents unnecessary computations, optimizing processing time and resource usage.
- **Robustness and Complementarity:** The joint application of the elbow method with other validation indices (such as the silhouette coefficient or the Calinski-Harabasz index) strengthens the reliability of the segmentation, offering particularly useful cross-validation in high-dimensional data analyses.

**Conclusion** The elbow index constitutes an essential quantitative and visual tool in optimizing K-means. Its ability to identify the equilibrium point between error reduction and model complexity facilitates robust and efficient segmentation, fundamental both in high-impact scientific research and in practical applications of data analysis.

### 3.4.5 Dunn Index

The Dunn index is a validation measure for the quality of clustering segmentation, designed to identify optimal partitions by simultaneously evaluating the separation between clusters and the internal compactness of each one. It is mathematically defined as:

$$D = \frac{\min_{1 \leq i < j \leq k} d(C_i, C_j)}{\max_{1 \leq l \leq k} \delta(C_l)},$$

where:

- $d(C_i, C_j)$  is the distance between clusters  $C_i$  and  $C_j$ , which can be defined, for example, as the minimum Euclidean distance between their points or the distance between their centroids.
- $\delta(C_l)$  represents the diameter of cluster  $C_l$ , that is, the maximum distance between any pair of points within that cluster.

**Calculation Procedure** To apply the Dunn index, the following steps are followed:

1. **Calculation of Inter-Cluster Distances:** Compute the distance between all pairs of clusters using an appropriate distance function.
2. **Calculation of Intra-Cluster Diameter:** For each cluster, determine the diameter by evaluating the maximum distance between any pair of points within the same cluster.
3. **Determination of the Index:** Identify the minimum inter-cluster distance and the maximum intra-cluster diameter. The Dunn index is obtained by dividing these two values, providing a measure that reflects both the external separation and the internal compactness of the clusters.

**Interpretation and Application in Optimization** A high value of the Dunn index indicates that the clusters are well-separated and exhibit high internal cohesion, translating into high-quality segmentation. This index is especially useful for:

- **Evaluating Clustering Quality:** It provides a single metric that summarizes the internal coherence and external separation of the clusters.
- **Comparing Different Configurations:** It allows for the comparison of various partitions obtained using different algorithms or parameter variations, facilitating the selection of the optimal number of clusters.
- **Segmentation Optimization:** It helps identify configurations that maximize the separation between clusters while minimizing internal dispersion, thereby optimizing both the interpretation and robustness of the clustering model.

**Conclusion** The Dunn index is an essential quantitative tool for validating clustering techniques. Its ability to simultaneously assess both separation and cohesion of clusters makes it a valuable instrument for optimizing segmentation, particularly in studies where the accurate identification of homogeneous groups is crucial for data interpretation and strategic decision-making.

### 3.4.6 Calinski-Harabasz Index

The Calinski-Harabasz index, also known as the variance ratio criterion, is a clustering validation measure that evaluates the separation between groups relative to the internal compactness of each group. It is defined as follows:

$$CH(k) = \frac{SSB/(k-1)}{SSW/(n-k)},$$

where:

- $SSB$  is the sum of squares between clusters, which measures the variability between the cluster centroids and the global centroid.
- $SSW$  is the sum of squares within clusters, which quantifies the internal variability of each cluster.
- $k$  is the number of clusters.
- $n$  is the total number of observations.

### Calculation Procedure

1. **Calculation of  $SSB$ :** Compute the variability between each cluster centroid and the global centroid, weighted by the number of observations in each cluster.
2. **Calculation of  $SSW$ :** Sum the internal variability for each cluster, i.e., the sum of the squared distances of the points from their respective cluster centroid.
3. **Index Evaluation:** The index is obtained by dividing  $SSB/(k-1)$  by  $SSW/(n-k)$ . A higher value of  $CH(k)$  indicates greater separation between clusters relative to the internal dispersion, suggesting an optimal segmentation.

**Interpretation and Application in Optimization** The Calinski-Harabasz index is an effective tool for evaluating clustering quality because:

- **Balance between Separation and Cohesion:** A high value implies that the variability between clusters significantly exceeds the internal variability, indicating robust segmentation.
- **Comparison of Configurations:** It allows for comparing different partitions and selecting the number of clusters that maximizes this index, thus optimizing the model structure.

- **Application in Various Contexts:** It is widely used in exploratory analysis and studies requiring quantitative validation of segmentation, thereby supporting data-driven strategic decisions.

**Conclusion** The Calinski-Harabasz index is a key measure in validating and optimizing clustering algorithms. Its ability to reflect both inter-cluster separation and intra-cluster homogeneity makes it a robust and efficient tool for determining the optimal number of clusters, proving highly relevant in both scientific studies and practical data analysis applications.

### 3.4.7 Davies-Bouldin Index

The Davies-Bouldin (DB) index is an internal validation metric that quantifies the quality of a clustering segmentation by evaluating the ratio between the internal dispersion of each cluster and the separation between clusters. It is defined as follows:

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left( \frac{S_i + S_j}{d(\mu_i, \mu_j)} \right),$$

where:

- $k$  is the number of clusters.
- $S_i$  represents the dispersion or internal inertia of cluster  $i$ , calculated, for example, as the average distance between each point in the cluster and its centroid  $\mu_i$ .
- $d(\mu_i, \mu_j)$  is the distance between the centroids of clusters  $i$  and  $j$ .

#### Calculation Procedure

1. **Calculation of Internal Dispersion:** For each cluster, calculate  $S_i$ , which measures the cluster's compactness.
2. **Calculation of Inter-Cluster Separation:** Determine the distance  $d(\mu_i, \mu_j)$  between the centroids of each pair of clusters.
3. **Determination of the Ratio:** For each cluster  $i$ , identify the maximum value of the ratio  $\frac{S_i + S_j}{d(\mu_i, \mu_j)}$  evaluated for all clusters  $j \neq i$ .
4. **Averaging:** The Davies-Bouldin index is obtained by averaging these maximum values over all clusters.

**Interpretation and Application in Optimization** A lower Davies-Bouldin index value indicates better clustering quality, as it suggests that the clusters exhibit high internal cohesion and adequate separation. In this sense, the index is useful for:

- **Comparing Configurations:** Evaluating and selecting the partition that minimizes the ratio between internal dispersion and inter-cluster separation.
- **Model Optimization:** Serving as a criterion for determining the optimal number of clusters, thereby facilitating the choice of the model that best fits the underlying data structure.
- **Internal Validation:** Complementing other validation metrics to provide a comprehensive evaluation of the segmentation.

**Conclusion** The Davies-Bouldin index is a robust tool for evaluating clustering quality, providing a direct measure of the compactness and separation of clusters. Its application is fundamental for optimizing and validating segmentation models, making it particularly relevant in studies and applications that require a detailed analysis of the internal structure of the data.

### 3.4.8 Lowess Smoother

The Lowess smoother (LOcally WEighted Scatterplot Smoothing) is a nonparametric technique for fitting regression models, designed to capture the underlying relationship between variables without imposing a global functional form. It is based on fitting locally weighted regressions using windows or "neighborhoods" that encompass a subset of the data.

**Theoretical Foundation** The Lowess method estimates the function value at a point  $x_0$  by performing a local linear (or polynomial) regression in the neighborhood of  $x_0$ . The estimate is obtained by minimizing the weighted sum of residuals:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n w_i(x_0) (y_i - \beta_0 - \beta_1(x_i - x_0))^2,$$

where  $w_i(x_0)$  are the weights assigned to each observation  $x_i$ , which decrease as the distance between  $x_i$  and  $x_0$  increases. A commonly used weight function is the tricubic function, defined by:

$$w_i(x_0) = \left(1 - \left(\frac{|x_i - x_0|}{d(x_0)}\right)^3\right)^3,$$

for  $|x_i - x_0| < d(x_0)$ , and  $w_i(x_0) = 0$  otherwise. Here,  $d(x_0)$  represents the maximum distance between  $x_0$  and the points that form the neighborhood.

#### Calculation Procedure

1. **Neighborhood Selection:** A smoothing parameter or bandwidth is defined, which determines the fraction of the data to include in the neighborhood around each point  $x_0$ .
2. **Weight Assignment:** The weights  $w_i(x_0)$  are computed for each observation, based on the distance to  $x_0$  and the chosen weight function.
3. **Local Fitting:** A regression (usually linear or of low order) is performed within the neighborhood, weighting the data by  $w_i(x_0)$ , to obtain a local estimate of the function at  $x_0$ .
4. **Repetition:** This process is repeated for each point in the domain of interest, thus constructing a smooth curve that follows the trend of the data without imposing a rigid global structure.

**Interpretation and Application** The Lowess smoother is especially useful in contexts where the relationship between variables is complex or nonlinear, allowing one to:

- **Capture Local Variations:** Its ability to fit locally enables the identification of patterns and changes in the trend that might be missed with global models.
- **Flexibility without Global Assumptions:** Without requiring a predetermined functional form, Lowess adapts to the intrinsic structure of the data, making it useful in exploratory analysis and in visualizing complex relationships.
- **Robustness to Outliers:** With the incorporation of robust techniques in the weight calculation, the influence of outliers can be mitigated, thereby improving the estimation of the central trend.

**Conclusion** The Lowess smoother offers a versatile and robust approach for modeling relationships between variables through local regression fitting. Its implementation provides a smooth curve that flexibly reflects the structure of the data, making it a valuable tool in both exploratory analysis and practical applications where specifying a parametric model is inadequate.

### 3.4.9 Linear Regression

Linear regression is a fundamental statistical method used to model and analyze the relationship between a dependent variable and one or more independent variables. In its simplest form, the simple linear regression model is expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where:

- $y$  is the dependent variable,
- $x$  is the independent variable,
- $\beta_0$  is the intercept or constant term,
- $\beta_1$  is the regression coefficient quantifying the effect of  $x$  on  $y$ ,
- $\epsilon$  represents the error term, which captures the variability not explained by the model.

**Estimation Procedure** The parameters  $\beta_0$  and  $\beta_1$  are commonly estimated using the least squares method, which seeks to minimize the sum of the squared errors between the observed values and those predicted by the model:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

The analytical solution is obtained using the following formulas:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x},$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means of  $x$  and  $y$ , respectively.

**Interpretation and Application** Linear regression allows one to:

- **Interpret Relationships:** The estimated coefficients indicate both the magnitude and direction of the effect of the independent variables on the dependent variable.
- **Prediction:** Once the parameters are estimated, the model can be used to predict future values of  $y$  based on new observations of  $x$ .
- **Model Evaluation:** Tools such as the coefficient of determination ( $R^2$ ) and residual analysis help evaluate the goodness-of-fit and the validity of the model assumptions.

**Conclusion** Linear regression is an essential tool in data analysis and statistical modeling, providing a simple and effective framework for understanding and predicting relationships between variables. Its implementation through the least squares method ensures optimal estimation under the assumptions of homoscedasticity and independent errors, making it a cornerstone in both scientific studies and practical applications across various fields.

### 3.4.10 Standardization and Its Importance in K-Means

**Standardization** is a fundamental data preprocessing step that normalizes variables to have a common scale. Mathematically, it is defined as:


$$X' = \frac{X - \mu}{\sigma}, \tag{1}$$

where  $X$  is the original value,  $\mu$  is the mean of the variable, and  $\sigma$  is its standard deviation. This transformation ensures that the data have a mean of zero and a unit standard deviation, which improves the numerical stability of many machine learning algorithms.



In the case of the **K-Means algorithm**, standardization is particularly beneficial because this method uses Euclidean distances to assign points to centroids. If the variables are on different scales, those with larger values will dominate the distance calculations, biasing the formation of clusters. By applying standardization, it is ensured that all features contribute equally to the clustering process, resulting in a more representative and stable segmentation.

## 4 Results and Discussion

HE results obtained in this study offer a detailed view of customer segmentation in shopping centers, highlighting the value of employing advanced data analysis and machine learning techniques to identify profiles with high conversion potential. Using the K-means algorithm on a Kaggle dataset that includes demographic variables (age, gender, annual income) and behavioral variables (spending score), homogeneous groups of customers were identified that allow for the effective optimization of marketing strategies.

### Overall Unisex Analysis

In the initial analysis without gender distinction, it was determined that the optimal number of clusters is six, established through the convergence of multiple validation indices: the elbow method, the silhouette coefficient, the Dunn index, the Calinski-Harabasz index, and the Davies-Bouldin index. This multi-metric approach ensures a robust segmentation that is representative of the underlying data structure.

The identified clusters show differentiated patterns in terms of age, annual income, and spending score. For example:

- **Cluster 1:** Young customers (mean age: 32.76 years) with high incomes (mean: 85.21 k\$) and a high spending score (mean: 82.11), constituting a high-value segment for targeted marketing strategies.
- **Cluster 5:** Older customers (mean age: 45.52 years) with low incomes (mean: 20.29 k\$) and a low spending score (mean: 19.38), suggesting a lower conversion potential.

Winsorization played a crucial role in this analysis by correcting skewness in the data distribution, particularly in clusters with outliers in annual income. This methodological step minimized biases, improving the interpretability and accuracy of the results.

### Gender-based Analysis

The segmentation by gender enriched the analysis by capturing specific nuances in consumer behavior between men and women, facilitating more personalized marketing strategies.

**Male Customers** For male customers, 11 optimal clusters were identified, reflecting greater heterogeneity in this group. Among the highlighted segments are:

- **Cluster 1:** Young men (age range: 31.2–39.2 years) with high incomes (70.3–83.7 k\$) and a very high spending score (88.0–94.8), demonstrating a strong propensity to consume.
- **Cluster 4:** Very young men (age range: 18.8–30.0 years) with low incomes (17.3–32.1 k\$) but high spending score (64.7–87.3), suggesting potential for aspirational or affordable luxury products.
- **Cluster 3:** Young men (age range: 17.9–25.3 years) with medium incomes (49.5–64.3 k\$) and a medium-high spending score (48.3–57.9), representing a stable market for categories such as technology, fashion, or entertainment.

These findings indicate that men exhibit subgroups with markedly distinct consumption behaviors, justifying differentiated strategies.

**Female Customers** In the case of female customers, 6 optimal clusters were identified, suggesting less diversity but greater concentration within each group. The most relevant segments include:

- **Cluster 3:** Young women (age range: 29.1–35.3 years) with high incomes (73.2–96.3 k\$) and a high spending score (73.8–89.5), ideal for premium products or services.
- **Cluster 2:** Very young women (age range: 20.8–28.6 years) with low incomes (17.5–33.9 k\$) but a high propensity to spend (69.8–91.2), representing an opportunity for promotional campaigns and affordable products.

In addition, a significant negative correlation between age and spending score ( $-0.4$ ) was observed in the overall female analysis, indicating that older women tend to spend less—a pattern that contrasts with that observed in men and underscores the importance of considering gender in segmentation.

## 4.1 Discussion

The results confirm the initial hypothesis that segmentation using K-means allows for the identification of profiles with a high likelihood of conversion. The combination of demographic and behavioral variables, along with rigorous preprocessing (standardization and winsorization) and multi-metric validation, ensures the robustness and applicability of the findings.

Compared to previous studies, such as Gilboa (2009), which focused on demographic typologies, this work incorporates implicit psychographic variables through the spending score, offering a more comprehensive view of consumer behavior. Additionally, the application of advanced techniques such as winsorization and cluster-based correlation analysis overcomes the limitations of traditional methods, aligning with the recommendations of Talnat et al. (2023) regarding the need for explainable models in segmentation.

A notable finding is the identification of clusters with atypical spending patterns, such as male Cluster 4 and female Cluster 2, which, despite having low incomes, show a high propensity to consume. This phenomenon, possibly influenced by cultural or aspirational factors, suggests new lines of research to integrate explicit psychographic variables and understand the motivations behind these behaviors.

From a practical perspective, the results advocate for highly personalized marketing strategies. For example:

- For male Cluster 1, campaigns focused on high-value products and premium experiences would be effective.
- For male Cluster 4 and female Cluster 2, strategies based on offers and promotions could capitalize on their aspirational consumption tendencies.

## Limitations and Future Work


Despite these advances, this study presents certain limitations. The dataset, although representative, is of moderate size (200 observations) and lacks explicit psychographic variables, which could restrict the depth of the analysis. Moreover, the absence of contextual information about the shopping center and the data collection period limits the generalization of the results.

For future research, the following lines are proposed:

- **Incorporation of Additional Variables:** Include psychographic data (preferences, motivations, lifestyles) to enrich the segmentation.
- **Temporal Analysis:** Conduct longitudinal studies to evaluate the evolution of consumption patterns in response to economic or technological changes.
- **Validation in Different Contexts:** Apply the methodology in various shopping centers and markets to test its adaptability.
- **Advanced Techniques:** Explore algorithms such as DBSCAN or hierarchical clustering to capture more complex data structures.

In conclusion, this study demonstrates the effectiveness of K-means in customer segmentation, providing a solid foundation for optimizing marketing strategies in shopping centers. The results highlight the importance of a detailed, multi-metric analysis, opening new perspectives for research and practical applications in data analysis and marketing science.

## 5 Conclusion and Future Work

HE present study has addressed customer segmentation in shopping centers using advanced data analysis and machine learning techniques, with a particular focus on the K-means algorithm. Based on a dataset obtained from Kaggle, which includes demographic variables (age, gender, annual income) and behavioral variables (spending score), homogeneous groups of customers with high conversion potential have been identified. This analysis provides a solid foundation for optimizing personalized marketing strategies, aligning with the main objective of profiling customers with the highest likelihood of purchase and improving resource allocation in competitive commercial environments.

The results highlight the effectiveness of the K-means algorithm in revealing differentiated customer profiles. In the overall unisex analysis, six clusters were identified that reveal varied consumption patterns, from young people with high incomes and high spending propensity to older segments with more conservative behaviors. The segmentation by gender deepened these findings, showing 11 clusters in men and 6 in women, which reflects greater heterogeneity among the former and greater concentration among the latter. Among the most valuable segments are young men with high incomes and high spending (male Cluster 1) and young women with a high propensity to consume despite low incomes (female Cluster 2). These groups underscore the importance of considering both demographic and behavioral variables for effective segmentation.

A key methodological aspect was the rigorous preprocessing, which included standardization to homogenize the scales of the variables and winsorization to mitigate the impact of outliers. Complemented by a multi-method validation — which integrated indices such as the elbow method, the silhouette coefficient, and the Dunn, Calinski-Harabasz, and Davies-Bouldin indices — this approach ensured the robustness and reliability of the results. Furthermore, the exploratory analysis revealed significant correlations, such as the inverse relationship between age and spending in women (-0.4), which highlights the relevance of gender as a differentiating factor in the segmentation.

From a practical perspective, this study demonstrates how machine learning-based segmentation can be translated into actionable marketing strategies. For example, the high-value segments identified justify campaigns focused on premium products, while clusters with high spending propensity and low incomes suggest opportunities for promotions and affordable products. These findings not only validate the applicability of K-means in real-world contexts, but also offer a guide for marketing professionals and data scientists interested in optimizing strategic decision-making.

However, the study presents limitations that must be acknowledged. The moderate size of the dataset (200 observations) and the absence of explicit psychographic variables restrict the depth of the profiles obtained. Likewise, the lack of contextual information about the shopping center and the data collection period limits the generalization of the results to other settings. These constraints open opportunities for future research that can overcome these challenges and broaden the scope of the analysis.

Based on the above, the following lines of future work are proposed:

- **Incorporation of Psychographic Variables:** Enrich the dataset with data on preferences, motivations, and lifestyles to capture more complex dimensions of consumer behavior and improve the precision of the segmentation.
- **Temporal Analysis:** Develop longitudinal studies that evaluate the evolution of consumption patterns in response to economic, social, or technological changes, enabling dynamic and adaptive marketing strategies.
- **Validation in Diverse Contexts:** Replicate the methodology in different shopping centers and markets to test its robustness and adaptability, strengthening the generalization of the findings.

- **Exploration of Advanced Techniques:** Investigate alternative algorithms such as DBSCAN or hierarchical clustering, which may capture more complex or non-spherical data structures, complementing the capabilities of K-means.

In conclusion, this work not only confirms the potential of machine learning techniques for customer segmentation in shopping centers, but also sets a precedent for future research. By addressing the limitations noted and exploring the proposed lines of work, it will be possible to deepen the understanding of consumer behavior and enhance the practical application of these methods in real commercial settings, thereby contributing to the advancement of data science and strategic marketing.

- **Incorporación de Variables Psicográficas:** Enriquecer el dataset con datos sobre preferencias, motivaciones y estilos de vida para capturar dimensiones más complejas del comportamiento del consumidor y mejorar la precisión de la segmentación.
- **Análisis Temporal:** Desarrollar estudios longitudinales que evalúen la evolución de los patrones de consumo frente a cambios económicos, sociales o tecnológicos, permitiendo estrategias de marketing dinámicas y adaptativas.
- **Validación en Diversos Contextos:** Replicar la metodología en diferentes centros comerciales y mercados para probar su robustez y adaptabilidad, fortaleciendo la generalización de los hallazgos.
- **Exploración de Técnicas Avanzadas:** Investigar algoritmos alternativos como DBSCAN o clustering jerárquico, que puedan capturar estructuras de datos más complejas o no esféricas, complementando las capacidades de K-means.

En conclusión, este trabajo no solo confirma el potencial de las técnicas de machine learning para la segmentación de clientes en centros comerciales, sino que también establece un precedente para investigaciones futuras. Al abordar las limitaciones señaladas y explorar las líneas propuestas, será posible profundizar en la comprensión del comportamiento del consumidor y potenciar la aplicación práctica de estos métodos en entornos comerciales reales, contribuyendo así al avance de la ciencia de datos y el marketing estratégico.

## References

- [1] Gilboa, S. (2009). *A segmentation study of Israeli mall customers*. Journal of Retailing and Consumer Services, 16, 135–144. doi:10.1016/J.JRETCONSER.2008.11.001.
- [2] Calvo-Porrá, C. & Lévy-Mangin, J. (2019). *Profiling shopping mall customers during hard times*. Journal of Retailing and Consumer Services, 48, 238–246. doi:10.1016/j.jretconser.2019.02.023.
- [3] Ruiz, J., Chébat, J. & Hansen, P. (2004). *Another trip to the mall: a segmentation study of customers based on their activities*. Journal of Retailing and Consumer Services, 11, 333–350. doi:10.1016/J.JRETCONSER.2003.12.002.
- [4] Dey, D. & Banerjee, K. (2023). *AI Driven Customer Segmentation and Recommendation of Product for Super Mall*. Journal of Mines, Metals and Fuels. doi:10.18311/jmmf/2023/34166.
- [5] Talaat, F., et al. (2023). *A Mathematical Model for Customer Segmentation Leveraging Deep Learning, Explainable AI, and RFM Analysis in Targeted Marketing*. Mathematics. doi:10.3390/math11183930.
- [6] Afzal, A., et al. (2024). *Customer Segmentation Using Hierarchical Clustering*. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1–6). doi:10.1109/I2CT61223.2024.10543349.
- [7] Yuldasheva, O., et al. (2024). *Shopping value and mall attributes: Generational and gender differences among Russian customers*. Upravlenets. doi:10.29141/2218-5003-2024-15-3-1.
- [8] Lee, D., Kim, S. & Ahn, B. (2000). *A conjoint model for Internet shopping malls using customer's purchasing data*. Expert Systems With Applications, 19, 59–66. doi:10.1016/S0957-4174(00)00020-8.

- [9] Bhatia, T., et al. (2022). *Analysis of Customer Segmentation Model through K-Means Clustering*. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 1–6). doi:10.1109/ICRITO56286.2022.9965157.
- [10] Agarwal, S., et al. (2024). *Harnessing Machine Learning for Effective Customer Segmentation*. In *2024 International Conference on Signal Processing and Advance Research in Computing (SPARC)*, 1, 1–6. doi:10.1109/SPARC61891.2024.10829373.
- [11] pantakanch (n.d.). *Customer-Segmentation-using-K-Means-Clustering* [Repositorio de GitHub]. Recuperado el 21 de marzo de 2025, de <https://github.com/pantakanch/Customer-Segmentation-using-K-Means-Clustering>.
- [12] Lathifah, S. N. & Azzahra, Z. F. (2025). *AI-Driven Customers Segmentation Using K-Means Clustering*. G-Tech: Jurnal Teknologi Terapan, 9(1), 320–329. doi:10.70609/gtech.v9i1.6202.
- [13] Wang, G. (2025). *Customer segmentation in the digital marketing using a Q-learning based differential evolution algorithm integrated with K-means clustering*. PLOS One, 20(2), e0318519. doi:10.1371/journal.pone.0318519.
- [14] Omol, E., et al. (2024). *Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya*. International Journal of Science, Technology & Management, 5(1), 192–200. doi:10.46729/ijstm.v5i1.1024.
- [15] saniya-k (n.d.). *Market-Basket-Analysis* [Repositorio de GitHub]. Recuperado el 21 de marzo de 2025, de <https://github.com/saniya-k/Market-Basket-Analysis>.
- [16] Rizky AL, M., et al. (2022). *Product Recommendations Using Market Basket Analysis with FP-Growth and Clustering Techniques*. En *Proceedings of the First Australian International Conference on Industrial Engineering and Operations Management*, Sydney, Australia, 20–21 de diciembre de 2022. Recuperado de <https://ieomsociety.org/proceedings/2022australia/85.pdf>.
- [17] Metilda, R. M., et al. (2023). *A Study on Customer Segmentation Using K-Means Clustering for Online Shoppers*. Rifanalitica Journal, 15 de agosto de 2023. Recuperado de <https://rifanalitica.it/index.php/journal/article/view/339>.