ROME BUSINESS SCHOOL
By:
Planeta Formación y Universidades

BETTER MANAGERS FOR A BETTER WORLD

## Master in Data Science

October 2023

**Capstone project by:**

Jeswin Sebi Puthussery

Oscar Hernandez

Maroua Marouani

Wafa khadher

# Oracle
# "ORA-FASHION Ltd Customer segmentation and marketing mission"

Rome, date of submission
September 2024

# Contents

## Abbreviations

1. **RFM** - Recency, Frequency, and Monetary
2. **CLV** - Customer Lifetime Value
3. **EDA** - Exploratory Data Analysis
4. **K-means** - K-means Clustering (a method for partitioning data into clusters)
5. **DBS** - Davies-Bouldin Score (a measure used to evaluate the quality of clusters)
6. **SS** - Silhouette Score (a measure used to determine the appropriateness of clusters)
7. **SKU** - Stock Keeping Unit (a unique identifier for each product type)

## Symbols

1. **Σ** - Summation (used in various statistical formulas, e.g., sum of values)
2. **μ** - Mean (average value of a dataset)
3. **σ** - Standard Deviation (measure of the amount of variation in a set of values)
4. **R** - Recency (number of days since the last purchase)
5. **F** - Frequency (number of purchases made by a customer)
6. **M** - Monetary Value (total amount spent by a customer)
7. **k** - Number of Clusters (in K-means clustering)
8. **n** - Number of data points (e.g., customers or transactions)
9. **d** - Distance (often used in clustering algorithms to measure similarity)

# List of Figures



# List of Tables

# I. Executive Summary

This capstone project focuses on segmenting ORA-FASHION's customer base to enhance loyalty and retention through targeted marketing strategies. By employing Recency, Frequency, and Monetary (RFM) analysis in conjunction with unsupervised clustering algorithms, customers were effectively segmented into distinct groups based on their purchasing behaviors. Additionally, an Exploratory Data Analysis (EDA) was conducted, and Customer Lifetime Value (CLV) was calculated to enrich the segmentation analysis. The clustering validity was confirmed using metrics like the Silhouette Score and Davies-Bouldin Index. These insights will drive personalized marketing efforts, optimizing customer engagement and driving long-term business growth.

**Keywords:** Customer Segmentation, RFM Analysis, k-Means, Customer Lifetime Value (CLV), Targeted Marketing

# II. Introduction

## A. Personal Part

We have chosen to work on the ORA-FASHION Ltd business challenge due to our shared passion for the fashion industry and our interest in using data analytics to address real-world issues. Our team members—Wafa, Oscar, Maroua, and Jeswin—bring diverse skills: Wafa and Jeswin excel in exploratory data analysis (EDA) and dashboard development, Oscar specializes in clustering algorithms and RFM analysis, and Maroua focuses on crafting innovative marketing strategies along with Business recommendation. This project is an excellent opportunity for us to apply our academic knowledge and professional skills to a practical and impactful business challenge.

## B. Scientific Part

This capstone project aims to segment ORA-FASHION's customer base and develop targeted marketing strategies to enhance customer loyalty and retention. The approach utilizes Recency, Frequency, and Monetary (RFM) analysis and unsupervised clustering algorithms, effective techniques in customer segmentation.

This project holds personal, professional, academic, and social relevance. Personally, it deepens our expertise in data analytics and marketing strategy. Professionally, it provides practical experience in addressing real business challenges. Academically, it bridges theoretical knowledge and practical application. Socially, it supports ORA-FASHION in delivering quality and transparency, thereby contributing to its growth and sustainability.

- The methodology comprises:

- Data Collection and Preprocessing**: Using anonymous purchase data from 22,626 customers.
- Exploratory Data Analysis (EDA): Uncovering initial insights and patterns.
- RFM Analysis**: Segmenting customers based on purchase behavior.
- Clustering Algorithms**: Identifying natural customer groupings.
- Customer Lifetime Value (CLV) Calculation**: Estimating the value of each segment.
- Marketing Strategies Development**: Proposing personalized marketing actions.
- Dashboard Development**: Creating a tool to monitor customer segmentation and data for better decision making.
- Conclusion and Business recommendation

By following the above structure, we aim to provide ORA-FASHION with actionable insights and tools to enhance their marketing strategies, foster stronger customer relationships, and drive sustainable business growth.

# III. Materials and Methodology

## A. Data Collection and Preparation

In the experimental phase of the project, two datasets were utilized. The first dataset contained the transaction records made by customers. Table 1 – Attributes descriptions of the eight fields contained in the Orders dataset.

| NOME | DESCRIZIONE | ESEMPIO | NOTE |
|---|---|---|---|
| id | Order Unique Identifier | 1 | |
| Date | Purchase Date | 2/1/2021 | |
| Customer_ID | Customer ID | 2547 | |
| Transaction_ID | Transaction ID | 1 | Multiple orders can be merged into one transaction |
| SKU_Category | Macro Category | X52 | Macro Aggregazione di SKU |
| SKU | Product Type | PUBUK | The SKU (Stock Keeping Unit) code is the unique number used by companies to track their inventory. Specifically, it is a sequence of alphanumeric characters that refer to key product details such as price, color, style, brand, gender, type, and size. |
| Quantity | Quantity | 1 | |
| Sales_Amount | Valore Monetario Ordine | 3.13 | |

*Table 1 - Attributes descriptions of dataset Orders*

It is from some variables found in this dataset that the RFM Analysis can be done as it has the Date attribute that represents the purchase date, the Quantity attribute which stands for the number of items bought and the Sale amount attribute with the monetary value of an order.

The second dataset comprised relevant customer information. ***Error! Reference source not found.*** provides descriptions of the attributes in the Customers dataset.

These two datasets were joined using the key attribute Customer_ID, ensuring that each transaction could be linked to the respective customer information. It is important to note that merging the two datasets was not strictly necessary for performing the RFM analysis, as all the relevant attributes for this analysis were already present in the first dataset. However, consolidating all information into a single dataset allows for deeper insights and more comprehensive analysis once customer segmentation is complete.

| NOME | DESCRIZIONE | ESEMPIO | NOTE |
|---|---|---|---|
| Customer_ID | Customer ID | 2547 | |
| GENDER | Customer Gender | F | |
| AGE | Customer Age | 48 | |
| GEOGRAPHY | Customer Nation | Spain | |

*Table 2 - Attributes descriptions of dataset Customers*

The final dataset comprises 22,625 customers whose age ranges from 18 to 62 years, providing a diverse demographic spread. These customers hail from seven different European countries, encompassing both men and women. The data spans a full calendar year, from January 2, 2021, to December 31, 2021, offering a comprehensive view of customer behavior and trends throughout the year.

## B. RFM Analysis

To effectively segment customers, we employed RFM (Recency, Frequency, Monetary) analysis, which is a proven method in marketing to evaluate customer value and behavior. As John R. Miglautsch explains, "The purpose of RFM is to provide a simple framework for quantifying that customer behavior. Once customers are assigned RFM behavior scores, they can be grouped into segments and their subsequent profitability analyzed. This profitability analysis then forms the basis for future customer contact frequency decisions" (Miglautsch, 2000).

The three variables required for this analysis were created as follows:

- **Recency (R):** The number of days since the last purchase.
- **Frequency (F):** The total number of purchases made by the customer.
- **Monetary (M):** The total monetary value of the purchases made by the customer.

### 1. Recency (R)

The recency variable in the RFM analysis addresses the question: How many days ago was the last purchase made by customers? *Figure 1* illustrates the distribution of the number of days since the last purchase for customers.

As shown, the distribution of the recency variable is right-skewed. This skewness is further highlighted by the mean being higher than the median, as detailed in *Table 3* which contains the main statistics for the recency variable. It is clear that while some customers made their last purchase within the past 2 months, with a significant peak in this range indicating a fairly active customer base, a significant number have not made a purchase in a long time.

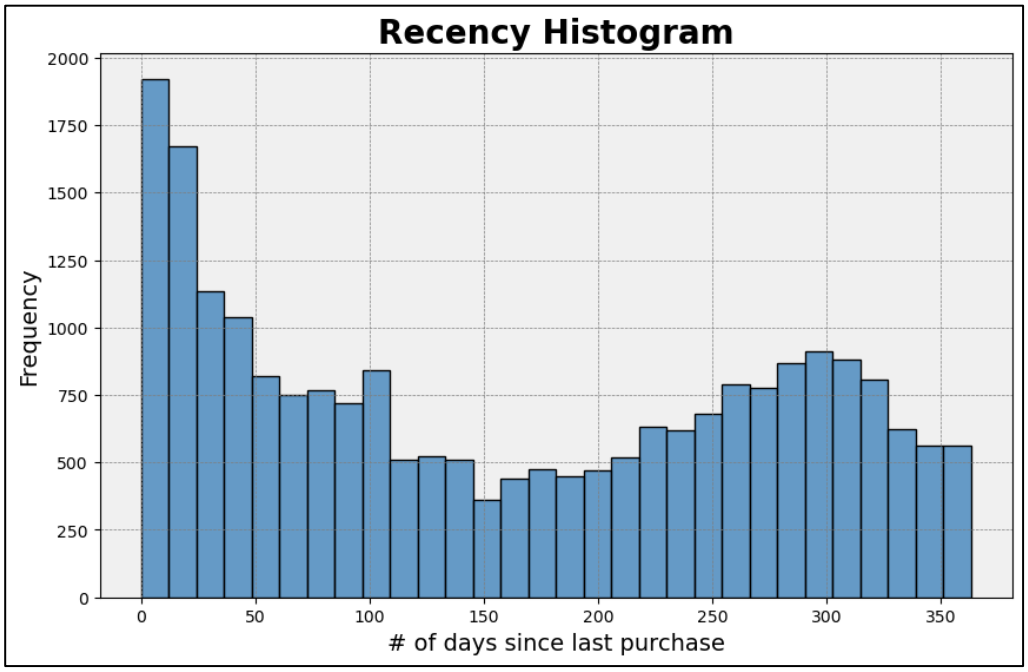


*Figure 1 - Recency Histogram*

The quartiles contain relevant information for an initial approach to customer segmentation. The lower quartile indicates that at least 25% of the customers made a purchase within the last 47 days. These are highly engaged customers who are currently active. The interquartile range (from 47 to 271 days) shows where the middle 50% of the data lies, suggesting a significant portion of customers have made a purchase within approximately 9 months. These customers are moderately engaged and could potentially be reactivated. Finally, the top 25% of customers made their last purchase more than 271 days ago, with the maximum recency reaching 363 days. These customers are at risk of becoming inactive or already inactive.

| | |
|---|---|
| **Count** | 22625 |
| **Mean** | 160.96 |
| **Standard deviation** | 115.72 |
| **Minimum** | 0 |
| **Lower quartile** | 47 |
| **Middle quartile (median)** | 149 |
| **Upper quartile** | 271 |
| **Maximum** | 363 |

*Table 3 - Main Statistics for Recency Variable*

## 2. Frequency (F)

The frequency variable in the RFM analysis addresses the question: How many purchases have customers made at the company since registering? *Figure 2* illustrates the distribution of the number of purchases made by customers since registration.

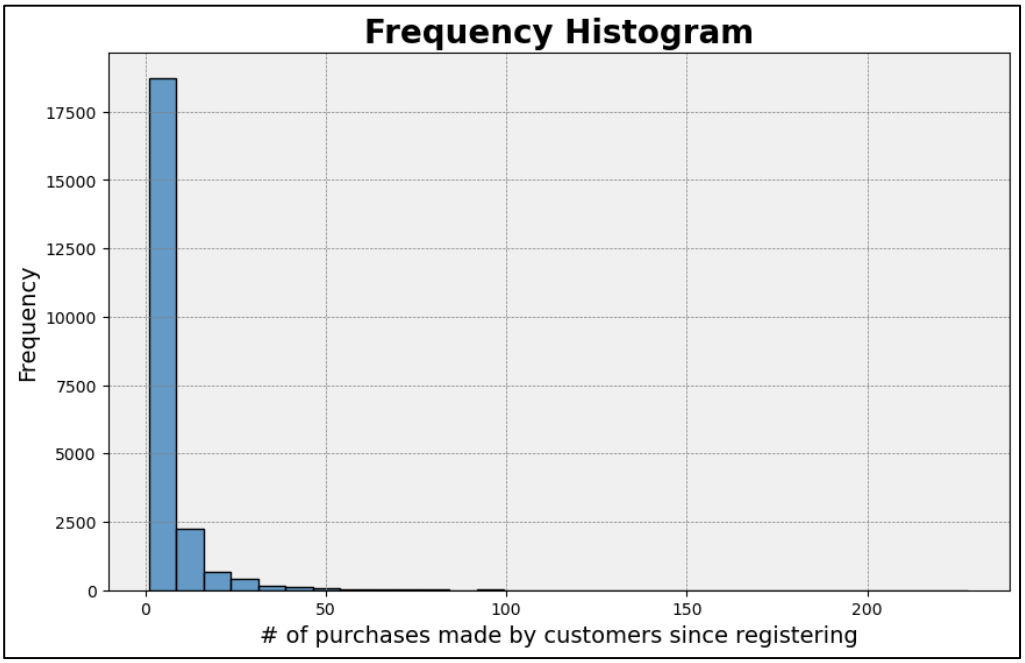


*Figure 2 - Frequency Histogram*

As shown, the distribution of the frequency variable is positively skewed, with a long tail towards higher purchase frequencies. The large standard deviation (9.88) found in *Table 4* reinforces this, indicating a wide range of purchase frequencies among customers.

| | |
|---|---|
| **Count** | 22625 |
| **Mean** | 5.82 |
| **Standard deviation** | 9.88 |
| **Minimum** | 1 |
| **Lower quartile** | 1 |
| **Middle quartile (median)** | 3 |
| **Upper quartile** | 6 |
| **Maximum** | 228 |

*Table 4 - Main Statistics for Frequency Variable*

The quartile values provide a detailed view of the distribution. The first quartile (1 purchase) indicates that a quarter of the customers are one-time buyers, representing a significant group with potentially lower engagement. The interquartile range (from 1 to 6 purchases) shows where the middle 50% of the data lies, suggesting that most customers made between 1 and 6 purchases. The upper quartile (6 purchases) indicates that while a substantial number of customers have made multiple purchases, very frequent buyers are relatively few.

## 3. Monetary Value (F)

The monetary value variable in the RFM analysis addresses the question: How much money have customers already spent at the company? *Figure 3* shows the distribution of the amount of money spent by customers since registering.

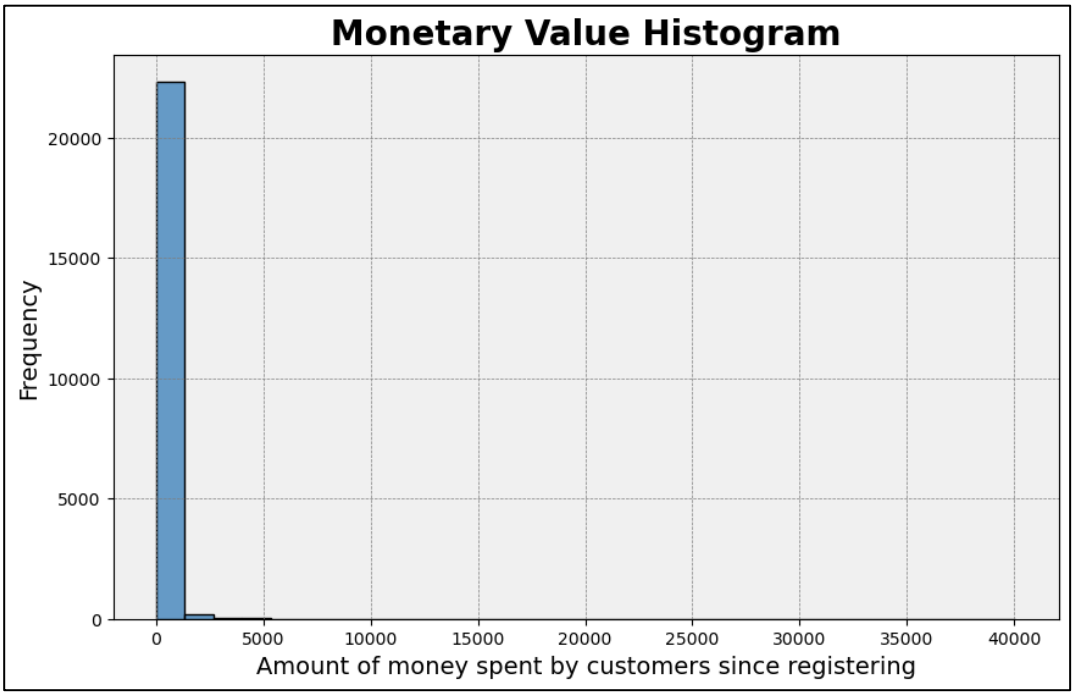


*Figure 3 - Monetary Value Histogram*

As shown, the distribution of the monetary value is positively skewed with a long tail towards higher spending. The high standard deviation (482.1) found in *Table 5* indicates substantial variability in spending behavior.

The quartile values provide a detailed view of the distribution. The first quartile (10.81) indicates that a quarter of the customers spent very little, up to $10.81. The interquartile range (from 10.81 to 83.08) shows where the middle 50% of the data lies, suggesting that most customers spent between $10.81 and $83.08. The upper quartile (83.08) suggests that while a substantial number of customers have moderate spending, very high spenders are relatively few.

| | |
|---|---|
| **Count** | 22625 |
| **Mean** | 122.01 |
| **Standard deviation** | 482.10 |
| **Minimum** | 0.14 |
| **Lower quartile** | 10.81 |
| **Middle quartile (median)** | 27.26 |
| **Upper quartile** | 83.08 |
| **Maximum** | 40070.49 |

*Table 5 - Main Statistics for Monetary Value Variable*

## C. Data Preprocessing

Prior to clustering, the RFM variables were standardized to ensure that each variable contributed equally to the clustering process, as Mohamad and Usman pointed out: “standardization before

clustering algorithm leads to obtain a better quality, efficient and accurate cluster result" (Mohamad & Usman, 2013)

For the purpose of this research, StandardScaler[1] was used to standardize the variables of the RFM analysis. StandardScaler is a feature scaling tool provided by the scikit-learn library in python that standardizes features by removing the means and scaling to unit variance.

## D. Clustering with K-means

K-means clustering[2] was selected as the method for segmenting customers based on their RFM values. K-means is used to partition a dataset into a predefined number of clusters. Each cluster is defined by its centroid, which is the mean of the points in that cluster.

The reason why this algorithm was chosen is well explained by Analytics Vidhya, which states, "K-Means Clustering is a simple yet powerful algorithm in data science. The algorithm works to group together or form clusters of data points from an unlabeled dataset. It is used in a plethora of real-world situations across various domains, such as banking and image segmentation, to document clustering and business decision-making" (Analytics Vidhya, 2019).

In order to make a more informed decision about the optimal number of clusters for the KMeans algorithm based on the structure and distribution of the RFM variables and given that the number of clusters is not known beforehand, the clustering performance was evaluated across a range of cluster counts (from 2 to 10 in this case) by using multiple evaluation metrics taken from the library scikit-learn in python.[3]

- **Silhouette Score[4]:** It measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a higher value indicates better-defined clusters. *Figure 4* contains the silhouette scores for different cluster counts of RFM variables.

  *Figure 4* illustrates that the highest silhouette score is at 6 clusters (0.5596), indicating that 6 clusters might provide the best clustering solution in terms of cohesion within clusters and separation between them. After 7 clusters, the silhouette scores slightly increase but do not reach the peak observed with 6 clusters. This suggests that adding more clusters beyond the optimal number of 6 does not significantly improve the clustering quality and may lead to over-segmentation.

[1] All documentation and a deeper understanding about this feature can be found here https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[2] All documentation and a deeper understanding about this clustering algorithm can be found here https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#kmeans

[3] The part of the code concerning the algorithm with the pre-processing of the data, the selection of the parameters and the use of the evaluation metrics can be found in the annexes section.

[4] All documentation and a deeper understanding about this score can be found here https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#silhouette-score

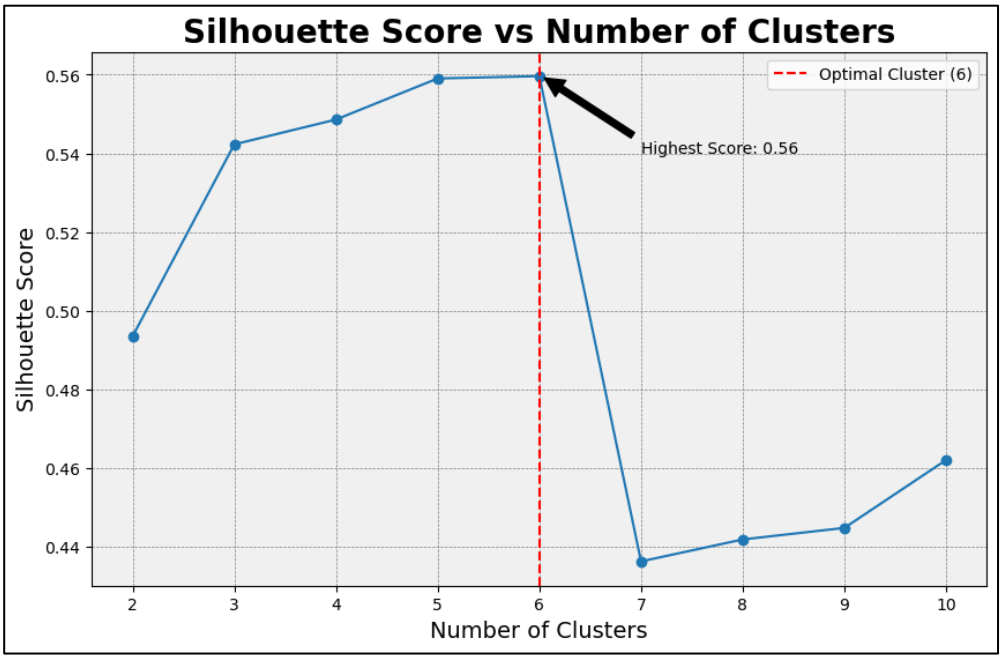


*Figure 4 - Line plot Silhouette Score vs Number of Clusters*

- **Davies-Bouldin Score[5]:** It evaluates the average similarity ratio of each cluster with the cluster that is most similar to it. Lower values indicate better clustering, as it means clusters are compact and well-separated.

  The decrease trend from 2 clusters (0.85) to 6 clusters (0.62) shown in *Figure 5*, it indicates an improvement in cluster separation and cohesion as the number of clusters increases up to 6. The lowest Davies-Bouldin score is at 6 clusters (0.62), suggesting that this configuration provides the best clustering solution in terms of minimizing the average ratio of intra-cluster distances to inter-cluster distances. The increase in the score after 6 clusters indicates that adding more clusters beyond this point reduces clustering quality.

  According to both the silhouette score and the Davies-Bouldin score Cluster 6 is identified as the optimal choice. The highest silhouette score and the lowest Davies-Bouldin score both suggest that clusters are well-separated and cohesive at 6 clusters and beyond 6 clusters, both scores indicate a drop in clustering quality, suggesting that additional clusters may lead to over-segmentation and reduced clustering effectiveness.

[5] All documentation and a deeper understanding about this score can be found here https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html#sklearn.metrics.davies_bouldin_score

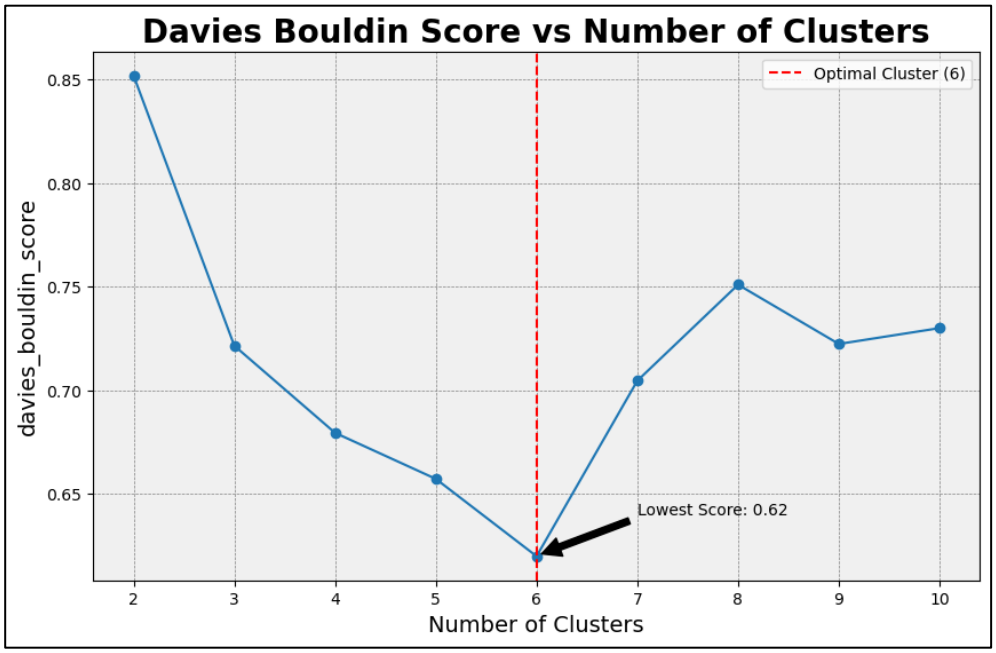


*Figure 5 - Line plot Davies Bouldin Score vs Number of Clusters*

## E. Cluster Validation

### 1. Structure Validation

To validate the structure of the clusters, a comparison was made with a randomly generated dataset that had the same number of attributes of the RFM analysis (In this case 3 attributes). The same K-means algorithm with six clusters was applied to this random dataset.

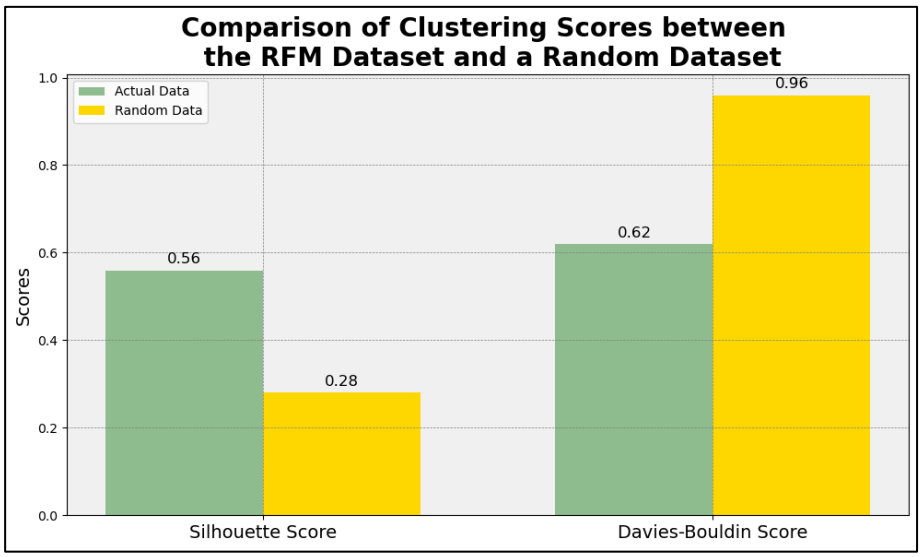


*Figure 6 - Bar plot Comparison of Silhouette and Davies-Bouldin Scores for Different Cluster Counts in RFM Analysis and a Random Dataset*

When comparing the clustering performance on the actual dataset versus the randomly generated dataset, it is observed significant differences in both the silhouette score and the Davies-Bouldin score. The actual dataset achieved a higher silhouette score (0.5596) compared to the random dataset (0.28), indicating better-defined and more distinct clusters. Additionally, the actual dataset had a lower Davies-Bouldin score (0.62) compared to the random dataset (0.96), suggesting higher cluster quality with less overlap between clusters. These metrics collectively demonstrate that the clustering structure in the actual dataset is meaningful and not a result of random chance.

### 2. Stability Validation

To ensure the stability of the clusters, a cross-validation approach was used. The original dataset was split into five different subsets. The K-means clustering algorithm with six clusters was applied to each subset, and both of the metrics were calculated for each subset. The mean and standard deviation of each metric were then computed. F*igure 7* displays the overall mean and standard deviation of the silhouette and Davies-Bouldin scores across the five subsets. The green dots represent the scores calculated from the entire dataset.

As shown, the scores calculated from the entire dataset fall into the confidence interval. This visual comparison demonstrates the stability and reliability of our clustering solution.

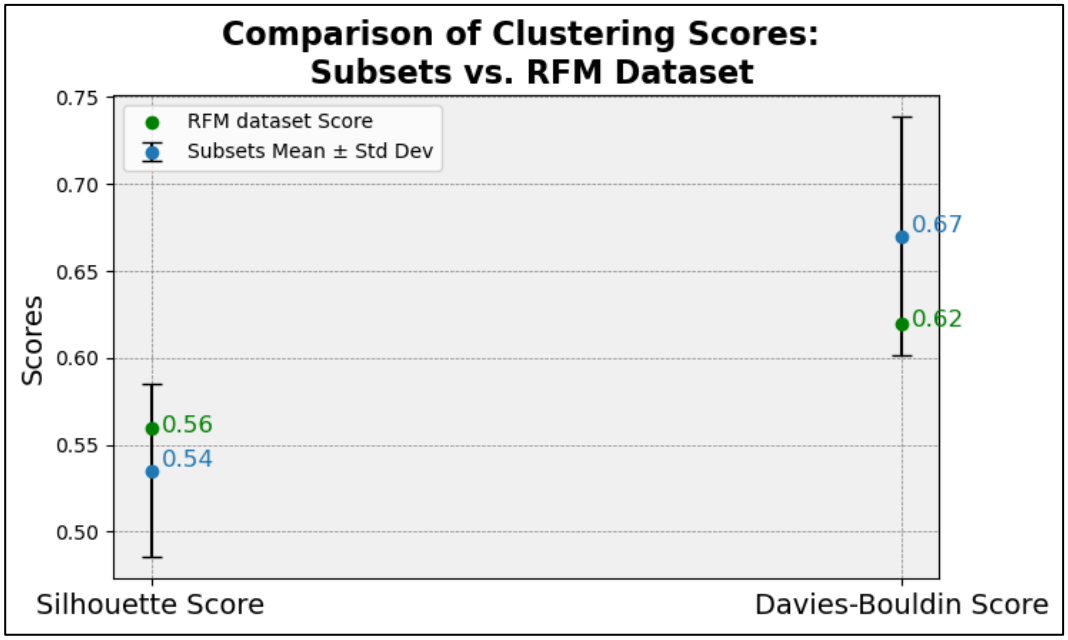


*Figure 7 - Comparison of Clustering Scores: Subsets vs. RFM Dataset*

## F. Exploratory Data Analysis (EDA) and Statistical Analysis insights

The previous sections successfully established a robust methodology for customer segmentation using RFM analysis and K-means clustering. By validating both the structure and stability of the clusters, we ensured that the segmentation was reliable and could effectively support targeted marketing strategies. This segmentation provides a foundation for developing personalized marketing campaigns, improving customer retention, and increasing overall customer satisfaction.

At this stage of the project, Exploratory Data Analysis (EDA) is a critical step in analyzing the data at hand. It allows us to understand the distributions of the data, identify outliers, visualize relationships among variables, and develop hypotheses for further testing. "EDA is detective work – numerical detective work – or counting detective work – or graphical detective work" (Tukey, 1977), therefore, performing EDA the following steps is crucial:

1. **Data Cleaning and Preparation:**
   - Check for missing values and handle them appropriately.
   - Convert data types if necessary (e.g., dates).

2. **Descriptive Statistics:**
   - Summary statistics for numerical columns.
   - Frequency counts for categorical columns.

3. **Visualization:**
   - Distributions of key variables
   - Relationships between variables
   - Clustering information visualization.
   - Identify any interesting patterns.

## 1. Data Cleaning and Preparation:

The dataset contains customer transaction data, including information such as customer demographics, transaction details, age, and no missing values detected at this stage. The dataset contains 131,706 entries with 18 columns Their description as follow:

1. **Customer ID**: Unique identifier for each customer.
2. **GENDER**: Gender of the customer.
3. **AGE**: Age of the customer.
4. **GEOGRAPHY**: Geographic location of the customer.
5. **ID**: Unique order identifier
6. **Date**: Date of the transaction.
7. **Transaction ID**: Unique identifier for each transaction.
8. **Subcategory**: Category of the SKU (Stock Keeping Unit).
9. **SKU**: Unique identifier for each SKU.
10. **Quantity**: Quantity of items purchased.
11. **Sales Amount**: order monetary value
12. **Total payment**: Total amount paid by the customer.
13. **Month name**: Month in which the transaction took place.
14. **Recency**: Recency of the customer's transaction (likely in days).
15. **Frequency**: Frequency of transactions made by the customer.
16. **Monetary value**: Monetary value of the customer (possibly total spend).
17. **Cluster**: Cluster assignment for each customer.
18. **Age range**: Age range classification.

## 2. Descriptive Statistics and Visualization

Summarizing data using statistical measures such as mean, median, count, and standard deviation is essential for gaining deeper insights into the dataset. *Table 6* contains the main descriptive statistics of the dataset concerned. Here is an overview of the meaning of each statistic:

- **Count**: The number of non-missing values for each variable.
- **Mean**: The average value of the variable.
- **Standard Deviation (std)**: The amount of variation or dispersion of the values.
- **Minimum (min)**: The smallest value in the dataset.
- **25th Percentile (25%)**: The value below which 25% of the data falls (first quartile).
- **Median (50%)**: The middle value of the dataset (second quartile).
- **75th Percentile (75%)**: The value below which 75% of the data falls (third quartile).
- **Maximum (max)**: The largest value in the dataset.

| | Customer ID | AGE | ID | Transaction ID | Quantity | Sales Amount | total payment | recency | frecuency | monetary value | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 131706 | 131706 | 131706 | 131706 | 131706 | 131706 | 131706 | 131706 | 131706 | 131706 | 131706 |
| mean | 12386.4504 | 35.42382 | 65853.5 | 32389.60419 | 1.485318 | 11.98152 | 20.95966 | 108.4077 | 22.61299 | 567.01677 | 1.315939 |
| median | 13496 | 36 | 65853.5 | 32620 | 1 | 6.92 | 7.64 | 52 | 12 | 167.69 | 1 |
| std | 6086.44755 | 7.06437 | 38020.39 | 18709.90124 | 3.872666 | 19.3597 | 101.324 | 112.7976 | 30.28018 | 1657.7311 | 1.037904 |
| min | 1 | 18 | 1 | 1 | 0.01 | 0.02 | 0.0006 | 0 | 1 | 0.14 | 0 |
| 25% | 7349 | 31 | 32927.25 | 16134 | 1 | 4.23 | 4.5 | 15 | 5 | 48.27 | 0 |
| 50% | 13496 | 36 | 65853.5 | 32620 | 1 | 6.92 | 7.64 | 52 | 12 | 167.69 | 1 |
| 75% | 17306 | 40 | 98779.75 | 48548 | 1 | 12.33 | 16.44 | 210 | 28 | 552.49 | 2 |
| max | 22625 | 62 | 131706 | 64682 | 400 | 707.73 | 13164.8 | 363 | 228 | 40070.491 | 5 |

*Table 6 – Main Statistics for the Data set*

As show in *Table 6*, The total number of customers as per the unique identifier is 22,625, indicating a large customer base. On average, each sale amounts to $11.98, with a standard deviation of $19.36, indicating variability in purchase amounts. The total payment made by customers averages $20.96, suggesting some high-value transactions. Recency, measured as the days since the last purchase, averages 108.4 days, while the frequency of purchases is 22.6 on average. The monetary value, representing the total amount spent by a customer, varies significantly with an average of $567.02.

The next part of the Exploratory Data Analysis will proceed by variable.

### a) Gender

- As *Figure 8* illustrates, there are 12,636 unique female customers and 9,989 unique male customers.

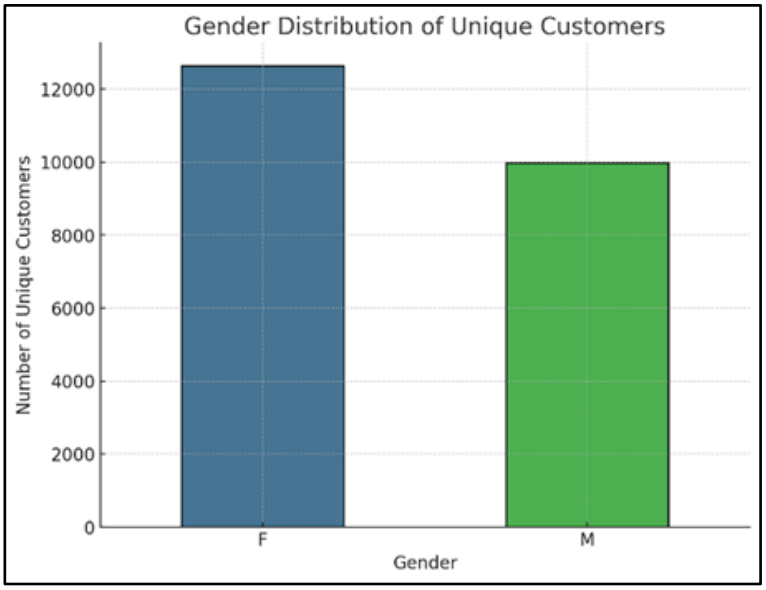


*Figure 8 - Gender Distribution based on the EDA analysis and taking Customer ID unique identifier count*

### b) Geography

As shown in *Figure 10* and *table 7*, the majority of customers are from Germany (6590), followed by Italy (5657) and France (3942).

| Country | Germany | Italy | France | Greece | Spain | Netherlands | UK |
|---|---|---|---|---|---|---|---|
| Customer Count | 6590 | 5657 | 3942 | 2272 | 1654 | 1351 | 1159 |

*Table 7 - Main Count of Customer per Country*

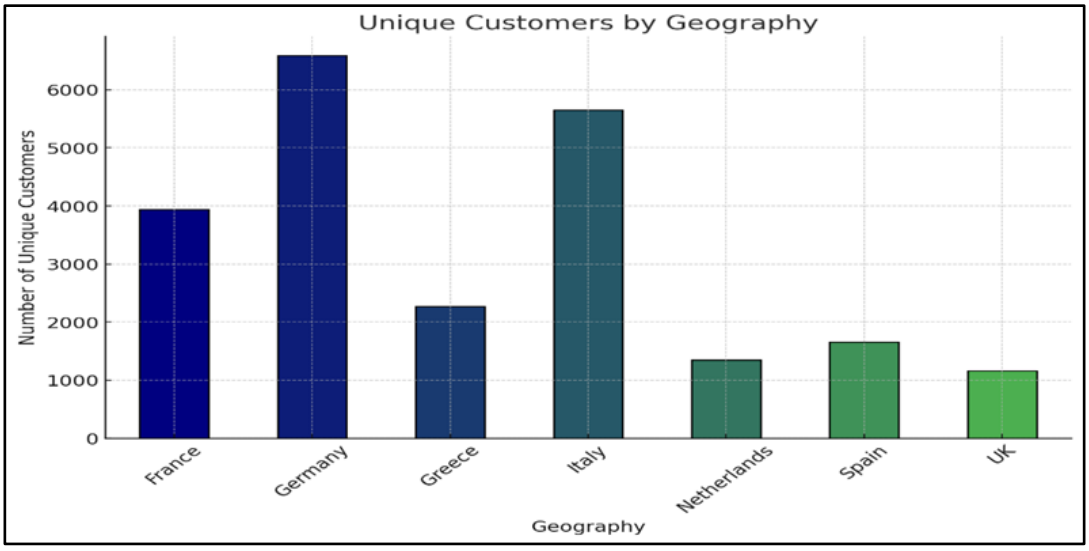


*Figure 9 - Geography Distribution*

### c) Monthly transaction distribution

Analyzing monthly purchase trends helps in identifying patterns and seasonal variations in customer behavior. This overview provides a detailed overview of sales, quantity purchased, and unique customers for each month of the year 2021.

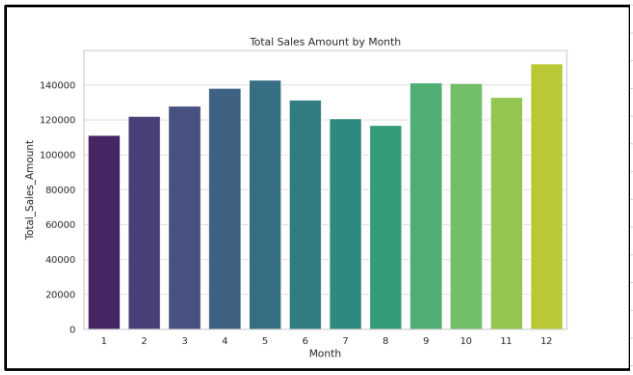


*Figure 10 - Total sales amount on monthly distribution*

| Year | Month | Total_Sales_Amount | Total_Quantity | Unique_Customers |
|---|---|---|---|---|
| 2021 | 1 | 111200.28 | 14875.27 | 3396 |
| 2021 | 2 | 122114.61 | 15253.71 | 3527 |
| 2021 | 3 | 127924.54 | 15686.1 | 4145 |
| 2021 | 4 | 138172.11 | 16633.486 | 4161 |
| 2021 | 5 | 142719.86 | 18295.257 | 4418 |
| 2021 | 6 | 131305.35 | 16120.596 | 3720 |
| 2021 | 7 | 120591.97 | 14296.96 | 3353 |
| 2021 | 8 | 116908.98 | 13901.52 | 3303 |
| 2021 | 9 | 141239.89 | 17399.23 | 4089 |
| 2021 | 10 | 140853.51 | 17801.98 | 3980 |
| 2021 | 11 | 132883.33 | 16683.92 | 3901 |
| 2021 | 12 | 152124.19 | 18677.3 | 4186 |

*Table 8 - Monthly analysis overview*

**Monthly Insights:**

- January typically starts with lower sales possibly due to the post-holiday season slowdown.
- February shows a slight increase in sales and unique customers, indicating a gradual recovery after the holiday season.
- Sales and customer numbers continue to rise in March, reflecting increased consumer spending.
- April sees a consistent increase in sales, quantity, and customer count, possibly due to spring promotions and events.
- May records high sales and customer engagement, potentially driven by Mother's Day and spring sales.
- June shows a slight decrease in sales and customer numbers, indicating the beginning of the summer slowdown.
- July continues the downward trend, which is typical for mid-summer months.
- September sees a significant rebound in sales and customer engagement, likely driven by back-to-school promotions.
- November's sales are solid, driven by early holiday shopping and Black Friday promotions.
- December records the highest sales, quantity, and customer count, attributed to the holiday season and year-end purchases.

### d) SKU insights

When analyzing SKU column, we identified 10 top-performing products in which they are generating $114,468.85 in revenue. However, characteristically varying, in high demand by women aged 25-40, which make them critical to the business. Consequently, they present product purchased denoting customer preference and a business opportunity for bundling or cross-selling strategies to enhance overall sales when further combined with marketing strategy.

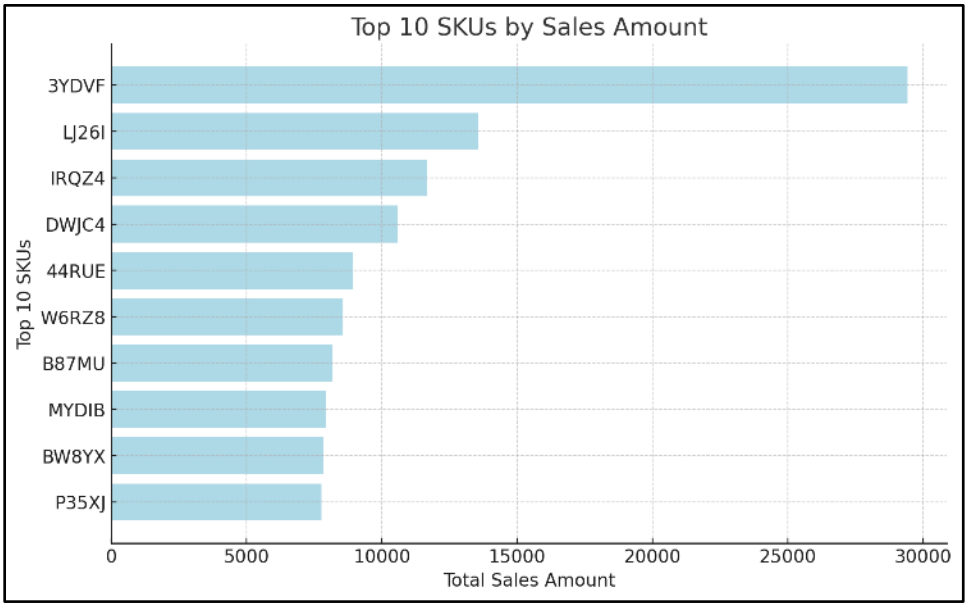


*Figure 11 - Top 10 SKU based on sales amount*

### *e) Age*

Figure 12 and 13 show that customers are relatively young to middle-aged (18 to 62 years) lead by Female customers, with the average age being about 35 years. Most customers fall between the ages of 31 and 40 (based on the interquartile range), which suggests that this age group might be a key demographic for the business.

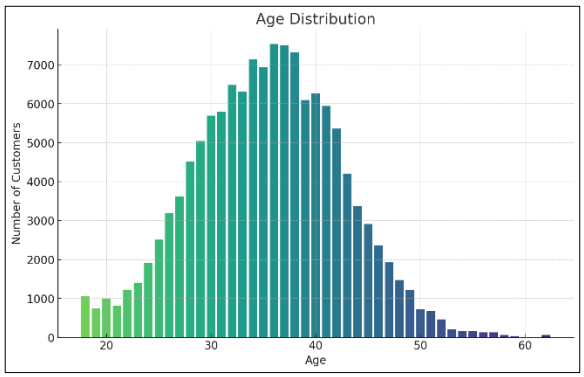


*Figure 12 - Age distribution*

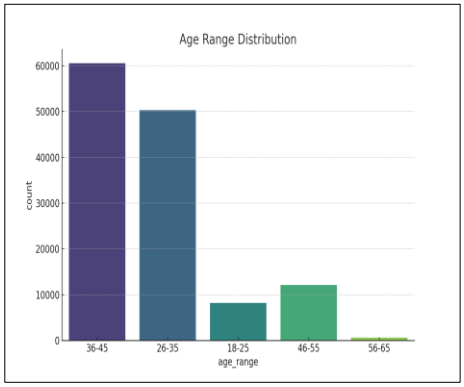


*Figure 13 - Age bar plot by ranges*

### *f) Quantity*

Most transactions involve small quantities, typically around 1 item per transaction. The fact that the maximum quantity purchased in a single transaction is 400 indicates that some transactions may involve bulk purchases, but these are likely outliers.

- **Count:** 131,706 (Total number of transactions recorded for quantity)

- **Mean:** 1.49 (Average number of items purchased per transaction)
- **Standard Deviation (std):** 3.87 (Variation in quantity of items purchased)
- **Min:** 0.01 (Smallest quantity purchased)
- **25th Percentile (25%):** 1 (25% of transactions involved purchasing 1 or fewer items)
- **Median (50%):** 1 (Middle value, meaning 50% of transactions involve purchasing 1 item)
- **75th Percentile (75%):** 1 (75% of transactions involve purchasing 1 item, indicating most purchases are for 1 item)
- **Max:** 400 (Largest quantity purchased in a single transaction)

### g) Sales Amount

The average sales amount per transaction is around $12, which suggests that the typical transaction value is relatively low. However, the range indicates that some transactions are worth significantly more, possibly due to bulk purchases or high-value items.

- **Count:** 131,706 (Total number of transactions with recorded sales amounts)
- **Mean:** $11.98 (Average sales amount per transaction)
- **Standard Deviation (std):** $19.36 (Variation in sales amounts)
- **Min:** $0.02 (Smallest transaction amount)
- **25th Percentile (25%):** $4.23 (25% of transactions are less than $4.23)
- **Median (50%):** $6.92 (Middle transaction amount)
- **75th Percentile (75%):** $12.33 (75% of transactions are less than $12.33)
- **Max:** $707.73 (Largest transaction amount)

### h) Total Payment

The average total payment per transaction is higher than the average sales amount, likely due to the inclusion of additional costs such as taxes, shipping, or other fees. The wide range and high standard deviation suggest significant variation in transaction values.

- **Count:** 131,706 (Total number of transactions with recorded payments)
- **Mean:** $20.96 (Average total payment per transaction)
- **Standard Deviation (std):** $101.32 (Variation in total payments)
- **Min:** $0.0006 (Smallest payment)
- **25th Percentile (25%):** $4.5 (25% of transactions involve payments less than $4.5)
- **Median (50%):** $7.64 (Middle payment value)
- **75th Percentile (75%):** $16.44 (75% of transactions involve payments less than $16.44)
- **Max:** $13,164.80 (Largest payment)

## G. CLV calculation

(Shapiro, 2007)Customer Lifetime Value (CLV) is *a crucial metric that helps businesses understand the total value a customer brings over their entire relationship with the company."* This metric is particularly valuable for informing marketing strategies, resource allocation, and overall business planning. By calculating CLV, businesses can identify their most profitable customers, because

(Gupta, 2005)"Investing in customers should be viewed as an investment decision, similar to other long-term investments in assets", and ultimately drive sustainable growth.

The calculation of CLV involves several key metrics, each providing insights into different aspects of customer behavior and value along with the business such as total revenue generated, total number of purchases, churn and repeat rate etc. In order for us to calculate the CLV, we will be using the following basic CLV calculation formula:[6]

1. **Avg. Purchase Value** (APV): Calculate the total revenue divided by the total number of purchases.
2. **Avg. Purchase Frequency** (APF): Calculate the total number of purchases divided by the total number of customers.
3. **Repeat Rate**: Calculate the percentage of customers with at least 2 transactions.
4. **Churn Rate**: Calculate as $1 - \text{Repeat Rate}$.
5. **Avg. Customer Lifespan** (ACL): Calculate as $\frac{1}{\text{Churn Rate}}$.
6. **Customer Lifetime Value** (CLV): Finally, calculate using the formula $\text{CLV} = \text{APV} \times \text{APF} \times \text{ACL}$.

*Table 9* illustrates the results of the Calculation based of the data at hand.

| Total Revenue | Total Number of Purchases | Total Customers Number | Average Purchase Value (APV) | Average Purchase Frequency (APF) | Repeat Rate | Churn Rate | Average Customer Lifespan (ACL) | Customer Lifetime Value (CLV) |
|---|---|---|---|---|---|---|---|---|
| 1578039 | 131706 | 22625 | 11.98152 | 5.8212597 | 0.721503 | 0.278497 | 3.590699889 | 250.442568 |

*Table 9 - CLV calculation based on the Data set*

**Explanation of key metrics and data Insights:**

1. **Average Purchase Value (APV)**: This is the total revenue generated by all purchases divided by the total number of purchases. It provides an understanding of how much revenue is typically generated from a single purchase per customer. And based on the data, average amount of money a customer spends per purchase is $11.98.
2. **Average Purchase Frequency (APF)**: This metric indicates how often a customer makes a purchase. It is calculated by dividing the total number of purchases by the total number of customers. And based on data results, each customer made nearly 6 purchases during the observed period.
3. **Repeat Rate**: The repeat rate measures the proportion of customers who make more than one purchase. For our data set, 72% of the customers made more than one purchase which suggests strong customer loyalty, and consequently it is essential for a high CLV.

[6] All the formulas were provided by the mentor of ORA-FASHION Ltd

4. **Churn Rate**: The churn rate represents the percentage of customers who do not return for a repeat purchase and for our data, we have 28% means that nearly a third of customers did not make more than one purchase.
5. **Average Customer Lifespan (ACL)**: this metric estimates the duration of a customer's relationship with the business. It is calculated at 3.59, which reflects how long customers are active and making purchases with the business before churning.
6. **Overall CLV:** ORA Fashion can expect to generate approximately $250.44 from each customer over the time they remain active until they churn.

To Sum up, the CLV analysis helps businesses understand their customer value and can guide decisions on marketing, customer acquisition, and retention strategies.

## H. Dashboard development

to ensure that ORA-FASHION Ltd project is showing its full potential ,an engaging dashboard using Google Looker Studio has been created because (Few, 2006)"A well-designed dashboard is not just a collection of charts but a carefully crafted tool that allows for quick and effective decision-making".

The dashboard is structured with various charts and scorecards, each designed to provide a specific view of the data. The core and main elements include:

- **Customer Age Distribution:** A column chart is used to illustrate the distribution of customers across different age groups. This visualization helps in identifying which age segments are most prominent among the customer base, offering valuable insights for targeted marketing.
- **Sales Trends by Month:** A line chart captures the sales performance over time, displaying monthly sales trends. This visualization is essential for identifying seasonal patterns, peak sales periods, and any anomalies in sales data.
- **Customer Count by Geography:** A bar chart shows the distribution of customers across different geographical regions. By visualizing this data, the company can assess market penetration and tailor regional marketing strategies accordingly.
- **Gender and Sales Distribution:** A donut chart is utilized to depict the gender distribution of customers and their corresponding sales figures. This dual-purpose chart provides a quick snapshot of gender-specific engagement and purchasing patterns, which can inform gender-targeted marketing efforts.
- **Key Performance Indicators:** Two scorecards are prominently featured on the dashboard. The first displays the Customer Lifetime Value (CLV) and its component, a critical metrics for understanding the long-term value generated by customers. The second scorecard shows the total sales amount, providing an instant overview of revenue performance.

In addition to the primary dashboard, a second and third dashboard focuses on Exploratory Data Analysis (EDA). This interactive EDA dashboard digs deeper into the data, revealing underlying trends, correlations, and insights that might not be immediately visible in the summary views.

Overall, the dashboards designed not only present data in a clear and aesthetically pleasing manner but also provide ORA-FASHION Ltd with insights into each cluster. By bringing together various data points into a visual story, these dashboards play a crucial role in empowering ORA-FASHION Ltd strategic business decisions.

# IV. Results, Solution, and Discussion

## A. Cluster Analysis insights

Clusters analysis combined with some statistical insights is essential, and it offers significant benefits for businesses especially in developing personalized Marketing strategy. Table 10 and Table 11 show a cluster statistical overview, which provides some insights.

| Cluster | Gender | Geography | total customers | total_revenue | total_purchases |
|---|---|---|---|---|---|
| 0 | F | Germany | 10420 | 375213.65 | 33850 |
| 1 | F | Germany | 10366 | 448369.83 | 42633 |
| 2 | F | Germany | 1563 | 489518.79 | 37466 |
| 3 | F | Italy | 226 | 217703.28 | 15414 |
| 4 | F | Italy | 49 | 43247.13 | 2205 |
| 5 | F | Germany | 1 | 3985.94 | 138 |

*Table 10 - 1st Cluster statistical overview*

| | Age | | | Recency | | | Frecuency | | | Monetary Value | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cluster** | *mean* | *median* | *std* | *mean* | *median* | *std* | *mean* | *median* | *std* | *mean* | *median* | *std* |
| 0 | 35.49353 | 36 | 6.912406 | 272.2736 | 278 | 47.33331 | 6.316632 | 5 | 4.789721 | 114.12631 | 51.99 | 171.0373 |
| 1 | 35.5511 | 36 | 6.963859 | 60.71862 | 48 | 47.75435 | 6.885629 | 6 | 3.942482 | 96.546845 | 66.54 | 102.412 |
| 2 | 35.42561 | 36 | 6.837429 | 55.06112 | 22 | 77.48828 | 27.06507 | 25 | 9.216563 | 551.6795 | 451.77 | 403.8085 |
| 3 | 34.86408 | 35 | 8.206273 | 19.6679 | 11 | 35.85455 | 79.80654 | 67 | 39.10263 | 1717.5618 | 1553.75 | 791.3148 |
| 4 | 35.92698 | 36 | 6.441882 | 47.42857 | 14 | 81.42693 | 94.19093 | 60 | 62.96453 | 6361.3619 | 5233.66 | 2901.946 |
| 5 | 33 | 33 | 0 | 16 | 16 | 0 | 138 | 138 | 0 | 40070.491 | 40070.491 | 0 |

*Table 11 - 2nd Cluster statistical overview*

Female customers dominate the customer base, with Germany being the leading country in customer representation. Among the clusters, Cluster 0 stands out with the highest customer count of 10,420 and the highest monetary value at $40,000. Cluster 2 generates the highest total revenue, amounting to $44,000, while Cluster 1 leads in purchase activity with a total count of 42,000 purchases.

### 1. Cluster 0

**Count:** This cluster contains 25% of the customers, as indicated by the 25th percentile value of 0. (based on Table 6 - Main Statistics for the Data set).

**Mean:** The average cluster value of 1.32 indicates that Cluster 0, along with Cluster 1, is one of the most populated clusters. (based on Table 6 - Main Statistics for the Data set).

**High Recency:** These customers might not have made a purchase recently, as their recency value is higher on average (272 days). This suggests that they might be at risk of churn.

**Low Frequency:** Frequency at around 6.3 meaning that Customers in this cluster tend to make fewer purchases or infrequent purchases.

**Low Monetary Value:** The overall spending of these customers is relatively low. With an average monetary value of $114.13, indicating infrequent low-value purchases.

Customers in Cluster 0 are likely less engaged or inactive with the business. They might need targeted marketing efforts, such as re-engagement campaigns or special offers, to encourage more frequent purchases and increase their lifetime value.

### 2. Cluster 1

**Count:** Represents another 25% of the customers, as indicated by the 50th percentile value of 1. (based on Table 6 - Main Statistics for the Data set).

**Mean:** Cluster 1 has an average value of around 1.32, indicating it is also a populous cluster. (based on Table 6 - Main Statistics for the Data set).

**Moderate Recency and Frequency:** These customers make purchases slightly more frequently and more recently than those in Cluster 0.

**Moderate Monetary Value:** Spending patterns are somewhat higher than Cluster 0 but still not very high.

Cluster 1 customers are a bit more active than Cluster 0 but still somewhat more recent but still infrequent and low spenders, potentially on the verge of becoming inactive and represent an area where engagement could be improved. A marketing strategy that focusses on increasing purchase frequency, such as loyalty programs or incentives for repeat purchases, could be effective.

### 3. Cluster 2

**Count:** These customers are in the 50th to 75th percentile range, representing about 25% of the customer base. (based on Table 6 - Main Statistics for the Data set)

**Mean:** With a value of around 2, Cluster 2 customers show more positive behaviors than those in Clusters 0 and 1.

**Better Recency and Frequency:** Customers in this cluster tend to make purchases more frequently and more recently, indicating higher engagement.

**Higher Monetary Value:** The average monetary value is $551.68, with a median of $451.77. This is a considerable increase compared to Clusters 0 and 1, suggesting higher spending behavior.

Cluster 2 represents a group of customers that are a more active and higher-value customer group, and valuable to the business. Retention strategies, such as personalized recommendations or exclusive offers, could help maintain and even increase their loyalty and value.

### 4. Cluster 3

**Count:** A smaller cluster, representing customers in the upper 25% to 75% range of purchase behavior.

**Mean:** The characteristics of this cluster suggest it is populated by customers who are more active and valuable.

**Frequent and Recent Purchases:** This cluster makes purchases often, with an average recency of about 19.7 days and extremely high Frequency at about 79.8 which shows strong engagement and very frequent purchasing behavior

**High Monetary Value:** Spending is also higher as the average monetary value is $1717.56, with a median of $1553.75, making these customers particularly important to the business.

Cluster 3 customers represent a very valuable customer segment, with high engagement, frequent purchases, and hefty spending. As these customers highly important to the business due to their loyalty, they should be nurtured with special attention, such as VIP programs or early access to new products, to ensure continued satisfaction and retention.

### 5. Cluster 4

**Count:** A very small, specialized segment with only a few customers only 49.

**Mean:** This cluster represents a very high-value group, as indicated by its characteristics.

**Very Frequent Purchases and Low Recency:** These customers make purchases very frequently with the highest average Frequency at 94.2 among all clusters, with low recency at 47.4 days, meaning they are constantly engaging with the business and making very frequent purchases.

**High Spending:** Their monetary value is exceptionally high at $6361.36 among all clusters, reflecting their frequent and large purchases.

Cluster 4 customers are the most valuable and should receive the highest level of personalized service. Ensuring these customers have an excellent experience is critical, as they contribute significantly to the business's revenue. Personalized communication and exclusive benefits could be key strategies for this group.

### 6. Cluster 5

**Count:** The smallest cluster only 1 customer.

**Mean:** Cluster 5 is an outlier, representing customers with extraordinarily high engagement and spending.

**Very High Frequency:** the Customer in this cluster is highly engaged, making frequent and high-value purchases.

**Extremely High Monetary Value:** This customer spent significantly more than others with purchase worth of $40,070.

Cluster 5 customer is extremely rare but exceptionally valuable and considered as VIP customer, both in terms of average and median spending. This customer has a potential for other future high-value purchases. In order to maintain their loyalty, we need to understand their needs and preferences by focusing on personalized engagement and premium VIP customer experience. This can also give insights into how to attract and retain other high-value customers.

## B. Marketing Strategies

In this section, a comprehensive marketing strategy will be outlining for ORA-FASHION Ltd., designed to maximize customer engagement and boost sales through a targeted approach. This strategy is built on insights derived from geographical segmentation and customer clustering, allowing us to tailor our efforts to the unique needs of each market segment.

Geographical Segmentation Insights will highlight the sales performance across key regions, helping us understand where our strongest markets are and where there's potential for growth. On the other hand, customer Clusters Insights will dig into the distinct customer groups based on their purchasing behavior, providing a clear picture of the value each cluster brings to the business.
With these insights in mind, a Marketing Strategy will be present for each cluster. The strategy for High-Value Customers (Cluster 5) focuses on deepening loyalty and offering premium experiences. For Medium-High Value Customers (Cluster 4), the strategy aims to maintain engagement and encourage higher spending through regular communication and upselling tactics in key regions including Germany, Italy, and France. Medium-Value Customers (Cluster 3) will be targeted with promotional offers and educational content to boost their purchasing levels in the same regions. Finally, the strategy for Low-Value

Customers (Clusters 0, 1 & 2) involves re-engagement and awareness campaigns, particularly in regions like Germany, Italy, and Spain, where there is potential to grow their contribution.

The choice to adopt these specific strategies for each cluster stems from the need to optimize our marketing resources by focusing on the areas with the highest return on investment while also nurturing the potential of lower-value customers.

### 1. Geographical Segmentation Insights

- **High Sales Countries:** Germany, Italy
- **Medium Sales Countries:** France, Greece
- **Low Sales Countries:** Spain, Netherlands, UK

### 2. Customer Clusters Insights

The values taken into account correspond to the median of the monetary value found in *Table 12 - 2nd Cluster statistical overview*.

- **Cluster 5:** High-value customers (~$40K in sales)
- **Cluster 4:** Medium-high value customers (~$5K in sales)
- **Cluster 3:** Medium-value customers (~$1.5K in sales)
- **Clusters 0, 1 & 2:** Low-value customers (~$500 combined in sales)

### 3. Combining Insights for a Marketing Strategy

Given the diverse nature of our customer base and the varying levels of engagement across different regions, it's essential to adopt tailored strategies for each customer cluster. The following sections outline the specific strategies designed for each customer segment, considering their unique behaviors, value to the business, and geographical distribution.

#### *a) High-Value Customers (Cluster 5)*

High-value customers are our most significant contributors, primarily located in Germany. To maintain and enhance their loyalty, our strategy focuses on rewarding their repeat business and providing them with exclusive experiences. This approach ensures that we continue to nurture these top-tier customers, securing their long-term commitment to ORA-FASHION Ltd**.**

- **Primary Regions:** Germany

- **Strategy:**

1. **Loyalty Programs:**

- **Exclusive Rewards:** Offer loyalty points, discounts, or free products for repeat purchases.
- **VIP Events:** Invite them to exclusive product launches or special events.

2. **Personalized Communication:**

   - **Email Campaigns:** Send personalized emails with exclusive offers and product recommendations.
   - **Direct Mail:** Use direct mail for special promotions or holiday greetings.

3. **Premium Services:**

   - **Priority Support:** Provide dedicated customer support lines.
   - **Early Access:** Give early access to new products or sales events.

### *b) Medium-High Value Customers (Cluster 4)*

Customers in Cluster 4, spread across Germany, Italy, and France, show considerable potential to increase their spending. By regularly engaging with them and offering up selling opportunities, we aim to transition them into high-value customers. The strategy for this group is designed to strengthen their connection with our brand and encourage higher purchase frequencies.

- **Primary Regions:** Germany, Italy, France

- **Strategy:**

1. **Regular Engagement:**

   - **Newsletters:** Send regular newsletters with updates on new products and company news.

   - **Social media:** Engage with these customers on social media platforms with targeted content.

2. **Upselling/Cross-Selling:**

   - **Product Bundles:** Create bundles of related products to encourage higher spending.
   - **Recommendations:** Use email and website recommendations based on previous purchases.

3. **Feedback Requests:**

   - **Surveys:** Conduct surveys to gather feedback on products and services.

- **Incentives:** Offer small incentives (e.g., discount on next purchase) for completing surveys.

### *c) Medium-Value Customers (Cluster 3)*

Medium-value customers represent a solid customer base with room for growth. The strategy here is to entice them with targeted promotions and educational content, primarily in regions like Germany, Italy, and France. By addressing potential barriers to higher spending, we aim to gradually elevate their value to the company.

- **Primary Regions:** Germany, Italy, France
- **Strategy:**

1. **Targeted Promotions:**
   - **Discount Campaigns:** Offer time-limited discounts to encourage purchases.
   - **First-Time Buyer Discounts:** Provide special discounts for their first purchase.
2. **Educational Content:**
   - **Blog Posts:** Share blog posts that highlight the benefits and uses of your products.
   - **Tutorials:** Create video tutorials or how-to guides to increase product understanding and usage.
3. **Customer Surveys:**
   - **Understand Barriers:** Use surveys to understand why their purchase amounts are lower.
   - **Address Concerns:** Identify common issues or barriers and address them through FAQs or customer support.

### *d) Low-Value Customers (Clusters 0, 1 & 2)*

Low-value customers, particularly those in Germany, Italy, and Spain, offer untapped potential. The strategy for these clusters is focused on re-engaging them through win-back campaigns and creating brand awareness. By introducing them to new products and keeping them informed about promotions, we aim to increase their activity and contribution.

- **Primary Regions:** Germany, Italy, Spain
- **Strategy:**

1. **Re-engagement Campaigns:**

    - **Win-Back Emails:** Send emails with special offers to re-engage inactive customers.
    - **Abandoned Cart Reminders:** Send reminders for abandoned carts with additional discounts.

2. **Basic Communication:**

    - **General Promotions:** Include them in general promotions and holiday sales.
    - **Seasonal Offers:** Offer discounts during major holidays or shopping seasons.

3. **Awareness Campaigns:**

    - **Introduce Campaigns:** Run campaigns that introduce new products and their unique benefits.
    - **Content Marketing:** Use content marketing strategies to create awareness about your brand.

### 4. Implementation Plan

To effectively execute our marketing strategies, an implementation plan was developed plan to prioritizes data integration, automation, and continuous monitoring. This plan will ensure that our campaigns are precisely targeted and adaptable to evolving customer needs, setting the stage for successful outcomes.

- **Data Integration:** Integrate customer and sales data into a CRM system to track and manage segmentation.
- **Marketing Automation:** Use marketing automation tools (e.g., Mailchimp, HubSpot) to schedule and personalize campaigns.
- **Regular Monitoring:** Continuously monitor campaign performance and adjust strategies based on data insights.
- **Customer Feedback Loop:** Implement a feedback loop to gather customer insights and improve segmentation strategies continuously.

## C. Conclusion

The statistical overview and the RFM analysis of the dataset provides a clear picture of customer behavior and value distribution. By understanding these metrics, the business can develop more effective strategies for customer engagement, retention, and growth

opportunities. The insights gained from the cluster analysis, in particular, highlight the importance of well-developed and crafted marketing strategies for better customer experience approaches tailored based on different customer segments behavior.

The marketing strategy leverages data-driven insights to deliver personalized and targeted campaigns. By addressing the unique needs and behaviors of different customer segments, ORA-FASHION Ltd can enhance customer engagement, boost sales, and foster long-term loyalty.

## V. Conclusion and recommendations

The paper demonstrates that Recency, Frequency, and Monetary (RFM) analysis, along with K-means clustering, effectively assigned ORA-FASHION Ltd's client base into various groups. This segmentation enabled a more in-depth understanding of client habits, allowing for the creation of tailored marketing tactics. The statistical overview and analysis revealed substantial differences in customer engagement, spending habits, and loyalty, all of which are critical for refining customer relationship management and marketing strategies.

**Final Recommendations:**

**Expand Data Sources:** Incorporate additional data sources:

- Net promotor score or Customer satisfaction score
- Customer feedback or comments
- Browsing behavior
- Product categorization in data sources

This will help gaining a more comprehensive understanding of customer preferences and refining product bundling, product suggestion and pricing strategy.

**Data-Driven Personalization actions:**

- **AB testing** for marketing strategies and use data for informed business decision as we uncover how different user segments respond to various changes and allow for immediate marketing intervention.

- **Use customer data:** to personalize marketing content dynamically. For example, using mail merge in tailoring emails, recommendations, and promotions based on individual customer data such as past purchases, browsing history, and demographics.

- **Focus on Best-Selling Products:** Leverage the insights from SKU analysis to prioritize marketing efforts on top-selling products that appeal to key customer segments, especially women aged 25-40, as they represent a critical demographic for ORA-FASHION.

- **Seasonal and Event-Based Promotions:** Capitalize on monthly basis how sales are trending and aligning them with marketing campaigns (seasonal and event-based shopping behaviors.) For example, intensify marketing efforts around back-to-school promotions, Black Friday, and the holiday season to maximize sales during peak periods.

These recommendations derived from the insights gained from the analysis to drive targeted, efficient, and effective marketing strategies, ultimately enhancing customer engagement, loyalty, and business growth for ORA-FASHION Ltd.

## VI. References

**Miglautsch, J.R.**, (2000). *Thoughts on RFM Scoring*. Available at: https://rfm.migmar.com/2000/05/28/thoughts-on-rfm-scoring-john-miglautsch/

**Mohamad, I. B., & Usman, D**. (2013). *Standardization and Its Effects on K-Means Clustering Algorithm*. *Research Journal of Applied Sciences, Engineering and Technology*, 6(17), pp.3299-3303.

**Analytics Vidhya**. (2019, August 5). *Comprehensive Guide to K-Means Clustering. Available at:* https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Scikit-learn, n.d. *sklearn.preprocessing.StandardScaler*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [Accessed 10 Aug. 2024].
Scikit-learn, n.d. *sklearn.cluster.KMeans*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#kmeans [Accessed 10 Aug. 2024].
Scikit-learn, n.d. *sklearn.metrics.silhouette_score*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#silhouette-score [Accessed 10 Aug. 2024].
Scikit-learn, n.d. *sklearn.metrics.davies_bouldin_score*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html#sklearn.metrics.davies_bouldin_score [Accessed 10 Aug. 2024].

**Tukey, John W.** (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company.

**Gupta, Sunil, and Donald R. Lehmann.** (2005). *Managing Customers as Investments: The Strategic Value of Customers in the Long Run*. Wharton School Publishing.

**Few, Stephen.** (2006). *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring*. Analytics Press.

**Customer Lifetime Value: Marketing Models and Applications**, edited by David Bejou and Barry T. Shapiro. Nova Science Publishers, 2007.

# VII. Annexes/Appendices

The following Python libraries were used across all code snippets in this Annexes:Pandas, numpy, seaborn, matplotlib.pyplot, from sklearn.cluster KMeans, from sklearn.metrics silhouette_score, davies_bouldin_score, from sklearn.preprocessing StandardScaler and from sklearn.model_selection KFold.

## A. AnnexA: Standardizing RFM Data Using StandardScaler

This code snippet defines a function, **StandardScaler_dataset**, which standardizes a given DataFrame using StandardScaler from scikit-learn. The function takes a DataFrame as input, scales the numerical features, and returns a new DataFrame with the standardized values. In this case, the function is applied to the 'recency,' 'frequency,' and 'monetary_value' columns of the RFM (Recency, Frequency, Monetary) dataset, creating a standardized version of these features for further analysis or clustering tasks.

```
def StandardScaler_dataset(df):

 scaler = StandardScaler()

 return pd.DataFrame(scaler.fit_transform(df), columns=[df.columns])

df_rfm_standard = StandardScaler_dataset(df_rfm[['recency','frecuency','monetary_value']])
```

## B. Annex B: Python Script for Calculating Clustering Scores

This code snippet defines the function **calculate_scores** calculates the Silhouette Score and Davies-Bouldin Score for clustering results by using k-means algorithm within a specified range of clusters. The function returns a DataFrame with the number of clusters and corresponding scores.

```
def calculate_scores(datos, min_clusters, max_clusters):
   scores = {'silhouette_score': [], 'davies_bouldin_score': []}
   cluster_dict = {}
   clusters_range = range(min_clusters, max_clusters)

   for k in clusters_range:
      kmeans = KMeans(n_clusters = k, n_init=20, max_iter = 300, random_state=42)

      clusters = kmeans.fit_predict(X=datos)
      #cluster_dict[k] = [clusters]

      scores['silhouette_score'].append(silhouette_score(datos, clusters))
      scores['davies_bouldin_score'].append(davies_bouldin_score(datos, clusters))
```

```
    df_scores = pd.DataFrame({'clusters': clusters_range, **scores})

    return df_scores
```

## C. Annex C: Generating Random Data for Baseline Comparison

This code snippet generates a random dataset with the same dimensions as the standardized RFM dataset using np.random.rand. The calculate_scores function is then applied to this random dataset to calculate clustering metrics (Silhouette and Davies-Bouldin scores) for comparison with the actual data. The results (scores_random_data) provide a baseline to assess the significance of the clustering structure in the original dataset by comparing it to randomly generated data.

```
random_data = np.random.rand(df_rfm_standard.shape[0],df_rfm_standard.shape[1])
scores_random_data = calculate_scores(random_data, 6, 7)
scores_random_data
```

## D. Annex D: Cross- Validation for Clustering Stability

This function, **cross_validate_clustering**, performs cross-validation to assess the stability and consistency of K-Means clustering results. It splits the dataset into n_splits subsets using K-Fold cross-validation. For each subset, K-Means is applied with n_clusters clusters, and the Silhouette and Davies-Bouldin scores are computed. The function returns the mean and standard deviation of these scores, providing insight into the reliability of the clustering model across different data splits. This method helps ensure that the clustering results are robust and not dependent on a specific subset of the data.

```
def cross_validate_clustering(data, n_splits, n_clusters):
    kf = KFold(n_splits=n_splits, shuffle=True, random_state=42)
    silhouette_scores = []
    davies_bouldin_scores = []

    for train_index, _ in kf.split(data):
        train_data = data[train_index]

        kmeans = KMeans(n_clusters=n_clusters, n_init=20, max_iter = 300, random_state=42)
        kmeans.fit(train_data)
        labels = kmeans.predict(train_data)

        silhouette_scores.append(silhouette_score(train_data, labels))
        davies_bouldin_scores.append(davies_bouldin_score(train_data, labels))

    return {
        'silhouette': (np.mean(silhouette_scores), np.std(silhouette_scores)),
        'davies_bouldin': (np.mean(davies_bouldin_scores), np.std(davies_bouldin_scores)) }
```