BEHAVIOURAL SEGMENTS EXPLORER: A K-MEANS AND PCA APPROACH

NAME :- SHUBHAM ROHIT HAUDGE

REGISTERATION NO:-202200124

SYMBIOSIS CENTRE FOR DISTANCE LEARNING (SCDL), PUNE (2024-25)

Contents

Motivation:

In today’s highly competitive and customer-centric market, businesses must go beyond traditional metrics and develop a deep understanding of their customers’ behaviours. However, customer bases are often large, diverse, and complex making it difficult to design one-size-fits-all strategies. This is where **clustering** becomes an invaluable tool.

- ➢ Discover hidden patterns in customer behaviour.
- ➢ Enable personalized marketing strategies.
- ➢ Improve customer retention and engagement.
- ➢ Allocate resources more effectively across segments.
- ➢ Enhance strategic decision-making in areas like loyalty programs, product recommendations, and customer support.

The core motivation behind this project is to use **unsupervised learning techniques** to segment customers based on behavioural patterns, particularly through **RFM (Recency, Frequency, and Monetary)** analysis. By grouping similar customers together, we aim to:

Rather than treating all customers equally, clustering allows us to target the right customers with the right messages at the right time, leading to more efficient marketing and higher customer lifetime value.

This project is motivated by the need to transition from generic customer treatment to data-driven segmentation and strategic personalization, helping businesses drive long-term growth through better customer understanding.

Background Research:

Customer segmentation is a well-established strategy in marketing and customer relationship management (CRM) aimed at identifying groups of customers with similar behaviours or characteristics. Traditional segmentation methods often rely on predefined business rules (e.g., age groups, location, or income levels). However, these methods lack flexibility and may fail to capture the nuanced behaviours found in modern customer data.

With the rise of data-driven decision-making, unsupervised machine learning techniques—particularly clustering algorithms have become increasingly popular for uncovering natural groupings within customer datasets. Among these, K-Means and Agglomerative

(Hierarchical) Clustering are widely used due to their simplicity, interpretability, and effectiveness in partitioning large datasets.

A common approach in customer segmentation is RFM analysis, which evaluates:

- **Recency** – How recently a customer made a purchase.
- **Frequency** – How often they purchase.
- **Monetary** – How much they spend.

This RFM framework, combined with clustering techniques, enables businesses to uncover patterns such as:

- High-value, loyal customers
- At-risk or inactive customers
- New or low-engaged users

Studies in marketing analytics, including work by Hughes (1994) and subsequent industry applications, have shown that RFM-based clustering can significantly improve customer targeting, reduce churn, and increase revenue by enabling personalized marketing strategies.

This project builds on this foundation by applying modern clustering techniques to real-world transactional data, aiming to uncover actionable customer segments that drive smarter business decisions.

Data Dictionary

Invoice: Unique identifier for each transaction or bill.

StockCode: Unique product/item code.

Description: Name or description of the product.

Quantity: Number of units sold in the transaction.

InvoiceDate: Date and time when the transaction occurred.

Price: Price per unit of the product.

Customer ID: Unique identifier for each customer (can be missing).

Country: Country where the customer is located.

Description of Key Fields

1. **Invoice:**

- **Type**: Categorical (String/ID)
- **Description**: This is the unique identifier for each individual transaction or invoice issued by the system. Every invoice represents a distinct sales event. In some cases, invoice numbers starting with "C" may indicate cancelled or credit transactions (e.g., returns or corrections). This field helps track and differentiate between valid purchases and reversed transactions.

2. **StockCode:**

- **Type**: Categorical (String/ID)
- **Description**: A unique code assigned to each product or item available for sale. This field enables product-level tracking across transactions. It plays a vital role in identifying which items are most frequently purchased, returned, or generate the most revenue.

3. **Description:**

- **Type**: Text (String).
- **Description**: Provides the textual name or description of the product identified by the StockCode. While this is a non-numeric field, it is helpful for interpreting clustering results or profiling customer behavior by product category. Data cleaning might be necessary as it can contain inconsistencies or missing values.

4. **Quantity:**

- **Type**: Numeric (Integer)
- **Description**: Represents the number of units of a particular product purchased in a single transaction. Positive values indicate purchases, while negative values often indicate product returns. Analysing this field can help understand purchase volume and detect unusual behaviour (e.g., very high or negative quantities).

5. **InvoiceDate:**

- **Type**: DateTime
- **Description**: Records the date and exact time the invoice was generated. This field is essential for time series analysis, including trend identification, seasonal purchasing patterns, and Recency calculation in RFM analysis. It can also be used to derive new features such as "Month," "Day of Week," or "Hour of Purchase."

6. **Price:**

- **Type**: Numeric (Float)
- **Description**: Indicates the unit price (typically in British Pounds) of the product at the time of sale. Combined with Quantity, it can be used to compute the total transaction value (Quantity × Price). Outliers in price can suggest premium items, data entry errors, or promotional discounts.

7. **Customer ID:**

- **Type**: Categorical (String/ID)
- **Description**: A unique identifier assigned to each customer. This field allows for customer-level aggregation, such as total spending, frequency of purchases, or customer segmentation. Some records may have missing Customer IDs, especially for anonymous or guest checkouts such rows might be excluded from clustering.

8. **Country:**

- **Type**: Categorical (String)
- **Description**: Specifies the country where the customer is based or where the transaction originated. This is useful for geographic segmentation, identifying international demand patterns, and understanding market behaviour across regions.

Variable categorization

1. **Categorical Variables:**
   - **Invoice:** Unique identifier for each transaction. Not used for modelling but useful for grouping or deduplication.
   - **StockCode:** Identifies individual products. Often encoded or used for item-based segmentation.
   - **Description**: Textual name of the product. Helpful for human interpretation but not directly used in clustering.
   - **Customer ID:** Unique identifier for each customer. Essential for customer-level aggregation and segmentation.
   - **Country:** Indicates customer location. Useful for geographic segmentation and regional insights.
2. **Numerical Variables:**

These contain measurable quantities and can be directly used in analysis, aggregation, and clustering (often after scaling).

- **Quantity:** Number of units purchased. Used to compute frequency or detect abnormal purchase behaviour.
- **Price:** Unit price of the product. Used in total transaction value or customer monetary value.

3. **Date/Time Variables:**

These contain temporal information and are used to derive time-based features or patterns.

- **InvoiceDate:** Timestamp of the transaction. Used to calculate:
  - Recency (days since last purchase),
  - Time-based trends (monthly, weekly patterns),

Seasonal or hourly purchasing behaviour.

4. **Derived/Engineered Variables (not originally present but usually created during analysis):**
   - **Total Price (Quantity × Price):** Represents the total value of a line item.
   - **Recency:** Days since the customer's last purchase. Used in RFM analysis.
   - **Frequency:** Number of transactions by a customer. Captures how often they shop.

- **Monetary:** Total amount spent by a customer. Reflects customer value.

Data Pre-Processing & Cleaning

Data preprocessing and cleaning are essential to prepare the dataset for effective clustering. These steps ensure high-quality data that supports meaningful segmentation and customer insights. The following steps were implemented:

1. **Data Inspection:**

   - Initially reviewed the dataset using .info() and .describe() to understand data types, missing values, and basic statistics.
   - Examined unique values in key columns such as Invoice, StockCode, Customer ID, and Country to detect inconsistencies or irrelevant entries (e.g., test records or non-purchase invoices).

2. **Handling Missing Values:**

   - Detected missing values using df.isnull().sum() and visual tools (e.g., heatmaps).
   - Customer ID field had a significant number of missing entries; these rows were removed, as customer-level clustering requires complete IDs.
   - For Description, missing values were rare and dropped to maintain data quality.

3. **Removing Duplicates:**

   - Checked for duplicates using df.duplicated().sum().
   - Removed duplicate transactions (entire duplicate rows), if any, to avoid bias in clustering.

4. **Filtering Irrelevant Records:**

   - **Removed:** Negative or zero values in Quantity or Price, which may indicate returns or data entry errors. Non-regular invoices (e.g., credit memos) by excluding entries with Invoice codes starting with 'C'.

5. **Feature Engineering:**

- Created a new variable TotalPrice = Quantity × Price to capture the monetary value of each transaction.
- Aggregated transactional data at the customer level to compute RFM features:

  - **Recency**: Days since the customer’s last purchase.
  - **Frequency**: Number of transactions made by the customer.
  - **Monetary**: Total amount spent by the customer.

6. **Scaling / Normalization of Numerical Features:**

- Applied scaling techniques to the numerical features (Recency, Frequency, Monetary) to ensure uniform contribution to clustering algorithms like K-Means, which are sensitive to magnitude.
- Two different methods were considered:
  - **Standardization (Z-score normalization)** using StandardScaler: Transforms the data to have a mean of 0 and standard deviation of 1. Suitable when data has a normal distribution or when relative distance is more important than absolute scale.

  - **Min-Max Scaling** using MinMaxScaler: Scales all feature values to a range between 0 and 1. Useful when the goal is to preserve the original distribution shape but bring all values into a fixed range.

The choice of scaler was validated by comparing clustering results (e.g., silhouette score) to ensure optimal performance.

7. **Final Data Review:**

- Performed a sanity check using .info() and .describe() again to ensure: No missing values remain. Numerical variables are scaled. Categorical variables are encoded (if used).
- Verified the shape and structure of the final dataset (rfm_scaled) before passing it to clustering algorithms.

Project Justification:

In today's highly competitive business environment, understanding customer behaviour is vital for long-term profitability and growth. With the rise of data-driven decision-making, businesses now possess vast volumes of transaction-level data but often lack the tools and strategies to extract actionable insights from it.

This project leverages unsupervised machine learning (clustering) to segment customers based on their purchasing behavior, using the Recency, Frequency, and Monetary (RFM) model. By grouping similar customers, businesses can:

- Identify high-value and at-risk customers.
- Personalize marketing strategies.
- Optimize resource allocation for retention and acquisition.
- Improve customer satisfaction and loyalty.

Clustering provides a data-backed approach to customer segmentation, eliminating the guesswork from traditional demographic-based marketing strategies. This not only enhances business efficiency but also enables better strategic planning.

Problem Statement:

The business currently treats all customers uniformly, without understanding the varying levels of customer value and engagement. As a result, marketing efforts may be misaligned, inefficient, and cost-ineffective.

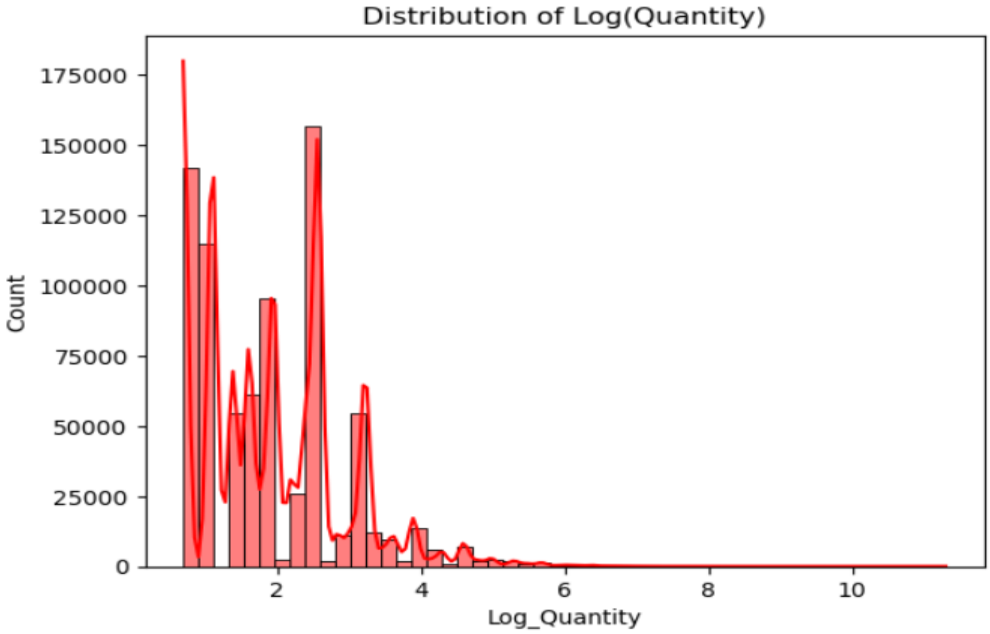
Segment customers into distinct behavioural groups using clustering algorithms on RFM features, enabling the business to tailor marketing strategies, improve customer retention, and enhance revenue generation.

Specifically, the project aims to:

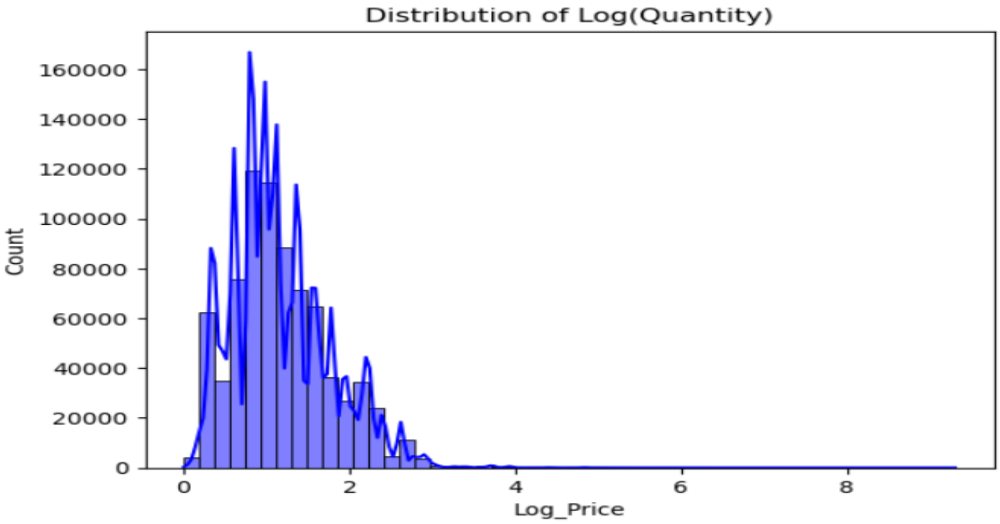
- Analyse customer purchase history using RFM metrics.
- Apply clustering algorithms (K-Means, Hierarchical Clustering) to discover natural customer groupings.
- Profile each cluster to derive actionable business insights.
- Recommend targeted strategies for each customer segment.

Exploratory Data Analysis

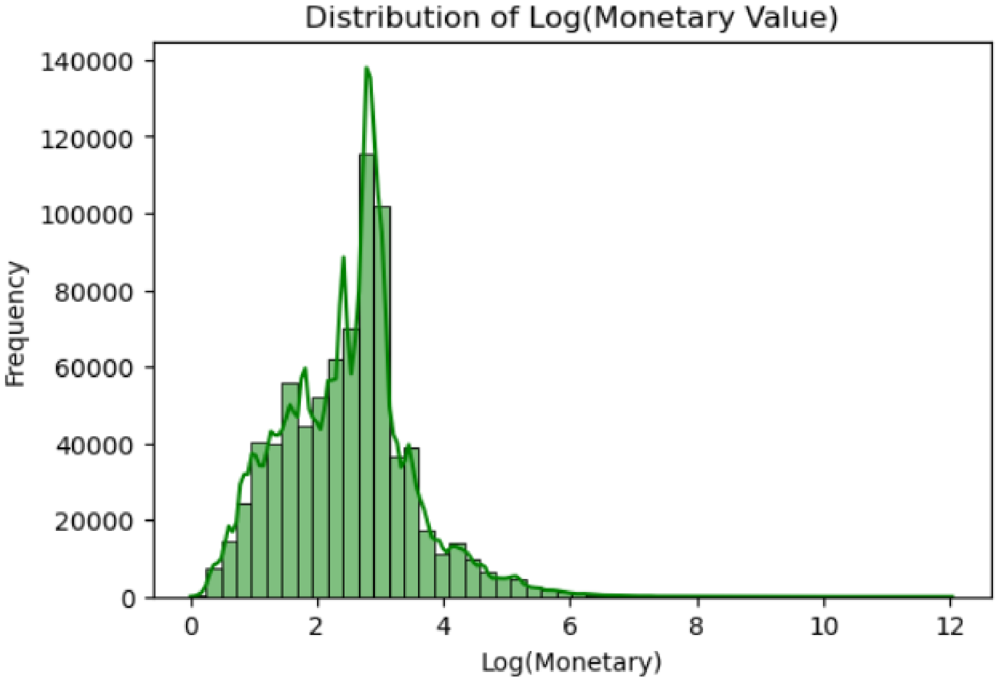
1. Distribution Plots of Quantity, Price and Monetary Value:



The histogram shows the log-transformed distribution of Quantity, where most values lie between log(1) and log(4), indicating small purchase quantities are most common. The distribution is right-skewed, with very few large quantity values (log > 6) appearing as outliers. The red outline likely represents a density curve, highlighting the overall trend. Log transformation helps reduce skewness and makes the data more interpretable for analysis.

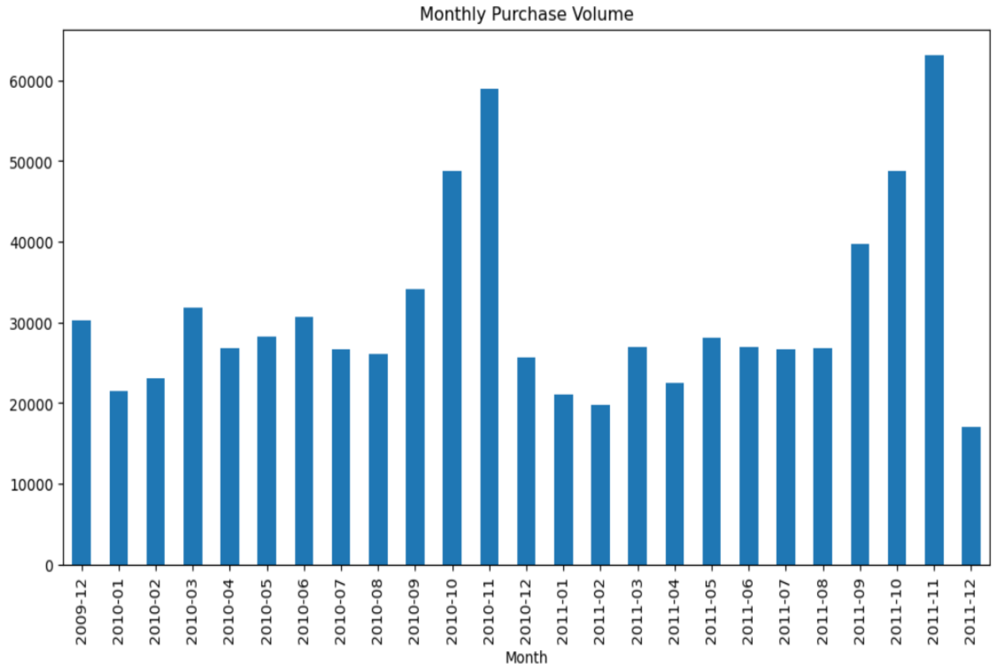


The histogram shows the log-transformed distribution of Price, where most values fall between log(0) and log(2), indicating that low prices are most frequent. The distribution is strongly right-skewed, with few high-price values beyond log(4). The blue curve outlines the overall trend, peaking just above log(1). Log transformation helps normalize skewed data for better analysis.



The graph is a histogram titled "Distribution of Log(Monetary Value)", which visualizes the frequency of log-transformed monetary values in a dataset. The x-axis represents the logarithm of monetary values, while the y-axis shows the frequency of those values. The distribution is right-skewed, indicating that lower monetary values are much more frequent than higher ones. Most of the data points are concentrated between log values 1.5 and 4, with a clear peak around log value 3, where the frequency exceeds 130,000. Beyond a log value of 4, the frequency steadily declines, and values greater than 6 are extremely rare. The smooth green line overlaying the bars suggests a kernel density estimate that outlines the overall shape of the distribution. This kind of log transformation is typically applied to reduce skewness in financial or transactional data, making the distribution more manageable for statistical analysis and machine learning models.

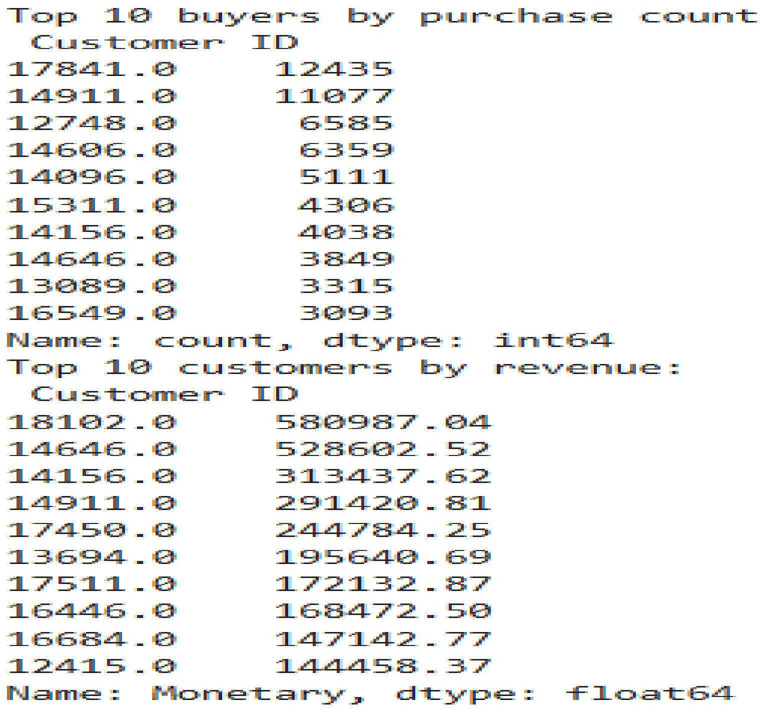
2. Monthly Purchase Trend Analysis:



The graph shows the monthly purchase volume from December 2009 to December 2011. The x-axis represents each month, while the y-axis indicates the total purchase volume. Overall, purchase activity fluctuates throughout the period. Notably, there are significant spikes in November 2010 and November 2011, suggesting seasonal peaks, possibly due to holiday sales or promotional events. In contrast, months like February 2011 and December 2011 show relatively lower purchase volumes. This pattern highlights a recurring surge in consumer activity towards the end of each year.

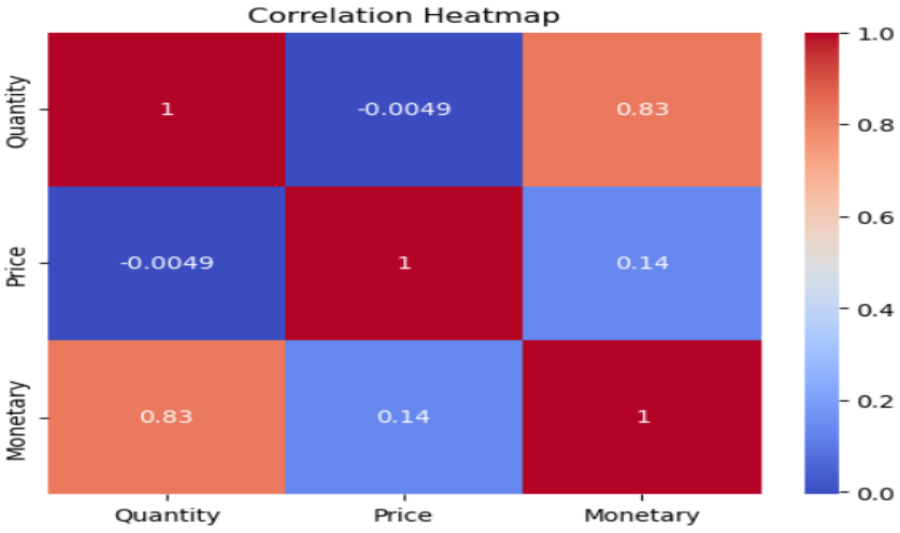
3. Customer Behavior Analysis:

The image displays two tables listing top customers based on different metrics. The first section shows the top 10 buyers by purchase count, where Customer ID 17841.0 leads with 12,435 purchases, followed by Customer IDs 14911.0 and 12748.0 with 11,077 and 6,585 purchases respectively. This list highlights the most frequent shoppers, indicating high engagement or loyalty. The second section shows the top 10 customers by revenue, with Customer ID 18102.0 generating the highest revenue of 588,987.04, followed by 14646.0 and 14156.0 with revenues of 528,062.52 and 313,407.62. This segment identifies the most valuable customers in terms of monetary contribution, which is crucial for revenue-focused strategies. Interestingly, some customers appear in both lists, suggesting they are not only frequent buyers but also high spenders.



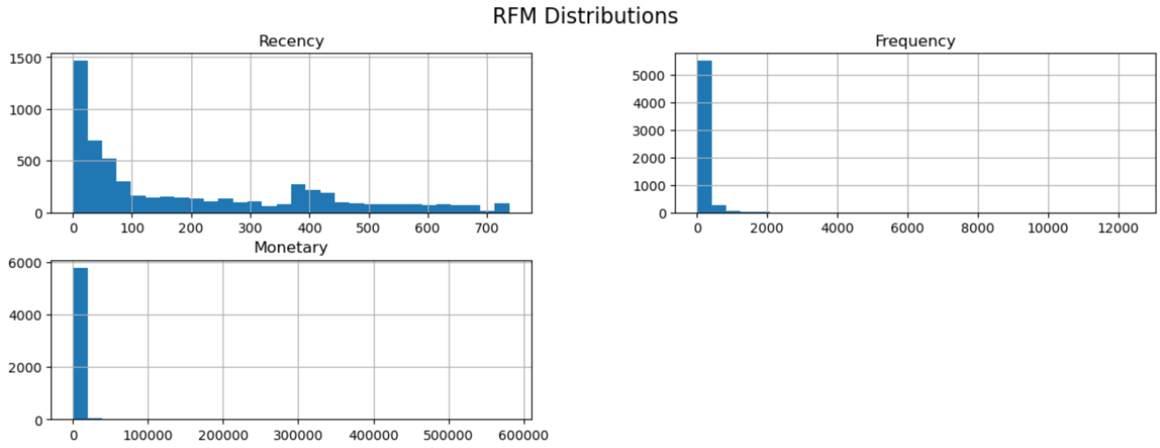
```
Top 10 buyers by purchase count:
 Customer ID
17841.0    12435
14911.0    11077
12748.0     6585
14606.0     6359
14096.0     5111
15311.0     4306
14156.0     4038
14646.0     3849
13089.0     3315
16549.0     3093
Name: count, dtype: int64
Top 10 customers by revenue:
 Customer ID
18102.0    580987.04
14646.0    528602.52
14156.0    313437.62
14911.0    291420.81
17450.0    244784.25
13694.0    195640.69
17511.0    172132.87
16446.0    168472.50
16684.0    147142.77
12415.0    144458.37
Name: Monetary, dtype: float64
```

4. **Feature Correlation Heatmap:**



The image is a correlation heatmap that visually represents the relationships between three numerical features: Quantity, Price, and Monetary. The values in the heatmap range from -1 to 1, where 1 indicates a perfect positive correlation, -1 a perfect negative correlation, and 0 no correlation. In this heatmap, Quantity and Monetary show a strong positive correlation of 0.83, suggesting that as quantity increases, the total monetary value also tends to increase significantly. Price has a very weak positive correlation with Monetary (0.14) and virtually no correlation with Quantity (-0.0049), indicating that price variations have minimal influence on the other two variables in this dataset. The diagonal elements all show a correlation of 1, as each variable is perfectly correlated with itself. This heatmap is useful for identifying how different variables relate to each other, which is valuable in feature selection and exploratory data analysis.

5. Customer RFM Feature Distributions:



The image shows RFM (Recency, Frequency, Monetary) Distributions, which are commonly used in customer segmentation to analyze purchasing behavior. Each subplot displays the distribution of one RFM component across customers:

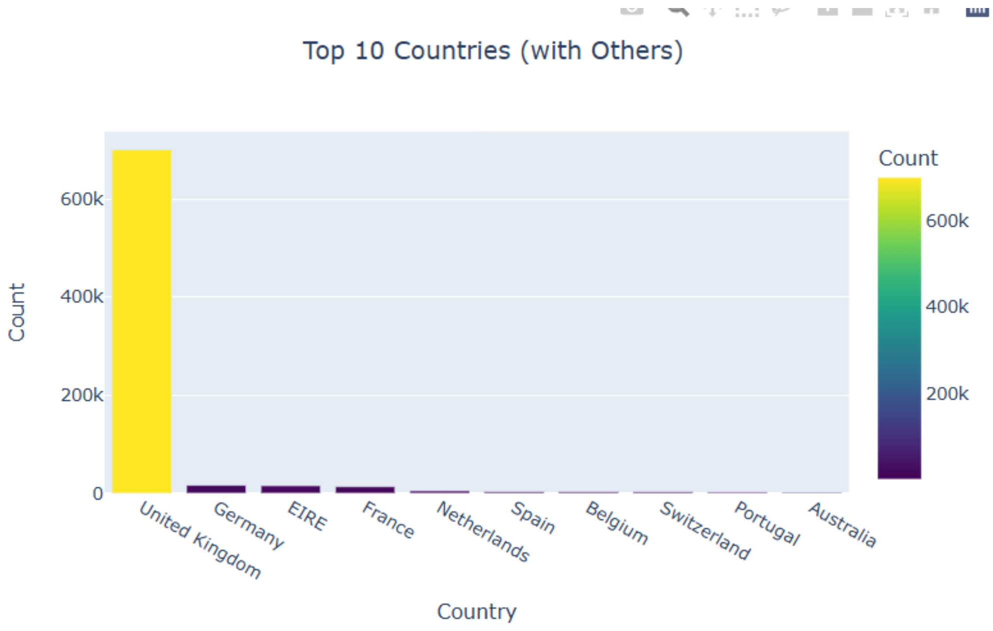
Recency (top-left) indicates how recently a customer made a purchase. Most values are clustered towards the lower end (0–100), suggesting that many customers purchased recently, while fewer have been inactive for longer periods.

Frequency (top-right) reflects how often customers made purchases. The distribution is heavily right-skewed, with most customers having a low number of purchases and a few having very high frequencies, up to around 12,000.

Monetary (bottom-left) shows the total amount spent by customers. Again, the data is extremely skewed, with the majority of customers having relatively low spending, while a few outliers spent significantly more (approaching 600,000).

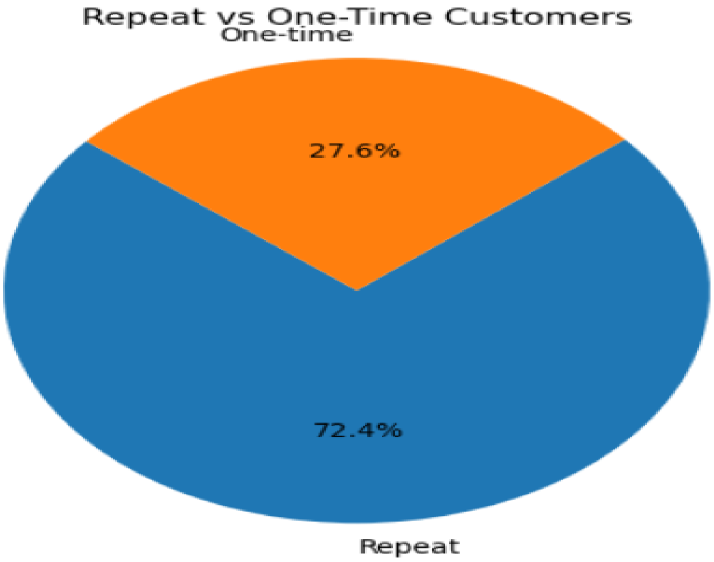
These skewed distributions are typical in real-world customer data and highlight the importance of handling outliers in customer analytics or segmentation tasks.

6. Top 10 Countries by Transaction Volume:



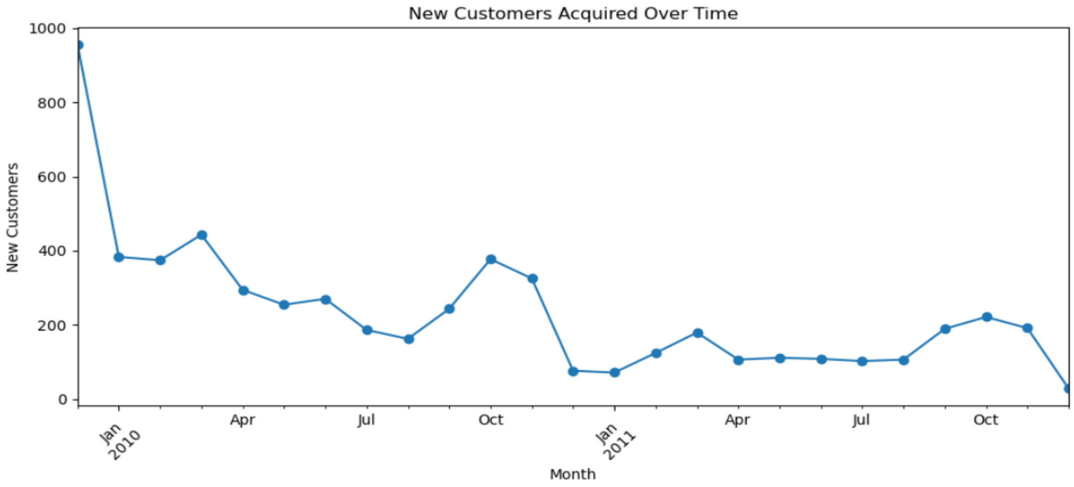
The bar chart displays the top 10 countries by transaction count, including a category for "Others" not explicitly listed. The x-axis represents different countries, and the y-axis shows the count of transactions or purchases. It is immediately evident that the United Kingdom overwhelmingly dominates the dataset, with over 600,000 transactions, while all other countries such as Germany, EIRE (Ireland), France, and others contribute only a small fraction by comparison. This suggests that the majority of business activity or sales occurred in the UK, possibly indicating it as the primary market or customer base. A color gradient is also used to visually emphasize transaction volumes, further highlighting the UK's dominance in the dataset.

7. Customer Retention Breakdown: Repeat vs One-Time Buyers:



The pie chart titled "Repeat vs One-Time Customers" illustrates the distribution of customer types based on their purchasing behavior. It shows that 72.4% of the customers are repeat buyers, indicating they have made more than one purchase. In contrast, 27.6% of the customers are one-time buyers, meaning they made only a single purchase. This distribution suggests a strong base of loyal or returning customers, which is a positive indicator for business sustainability and customer satisfaction. Maintaining and nurturing this repeat customer segment can be more cost-effective than acquiring new customers.

8. Customer Acquisition Over Time: Peaks and Patterns:

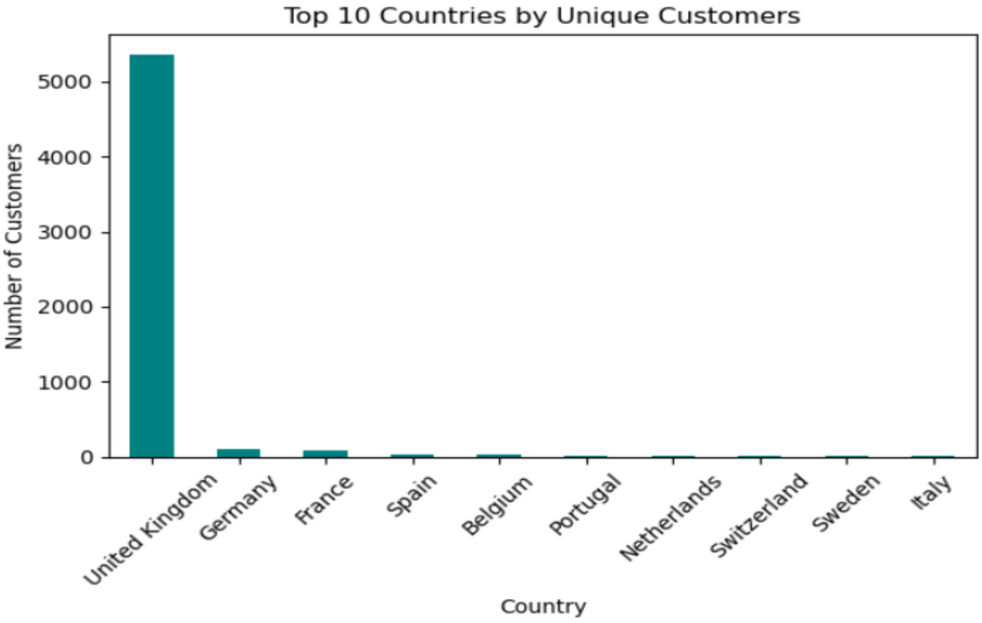


The line chart presents a temporal analysis of new customer acquisition based on their first recorded purchase month. It allows us to observe how the company has performed over time in attracting fresh customers. Each data point on the line represents the count of new customers who made their first purchase in that particular month, helping identify periods of high growth or stagnation.

A notable observation is the significant surge in customer acquisition during January 2010. This spike may reflect the impact of a major promotional campaign, a new product launch, seasonal shopping (e.g., post-holiday or New Year sales), or expansion into new markets. However, following this peak, there’s a marked decline in new customer acquisitions, which continues with fluctuations throughout 2011. This could indicate reduced marketing efforts, saturation of the target market, or diminishing customer interest.

Such patterns highlight the importance of consistently maintaining customer outreach and engagement strategies. The sharp decline after a successful month emphasizes the need for sustainable marketing plans and retention programs to maintain steady growth in customer acquisition over time. Analyzing the reasons behind these trends can help improve future customer acquisition efforts.

9. Geographic Analysis of Customer Base:



The bar chart highlights the top 10 countries in terms of the number of unique customers, offering insight into the geographic distribution of the company's customer base. This visualization is crucial for understanding market concentration and potential expansion opportunities across regions.

The United Kingdom stands out as the dominant market, with a significantly higher number of customers compared to other countries. This suggests that the business is either based in the UK or has heavily focused its operations, marketing, and logistics in this region. The strong customer presence may be due to better brand recognition, stronger distribution networks, or localized offerings that resonate with UK consumers.

In contrast, other countries like Germany, France, and Spain show a much smaller share of unique customers. This imbalance indicates a relatively limited international footprint or targeted focus on fewer regions. While these countries are part of the top 10, the disparity suggests that there may be untapped potential in international markets. Expanding marketing efforts, tailoring offerings to local preferences, or improving delivery logistics could help grow the customer base outside the UK.

Understanding this geographic skew allows decision-makers to evaluate whether to consolidate strength in existing markets or strategically expand into underpenetrated regions for broader growth.

Feature Engineering – Scaling:

This step standardizes the RFM features using two different scaling techniques to prepare the data for clustering:

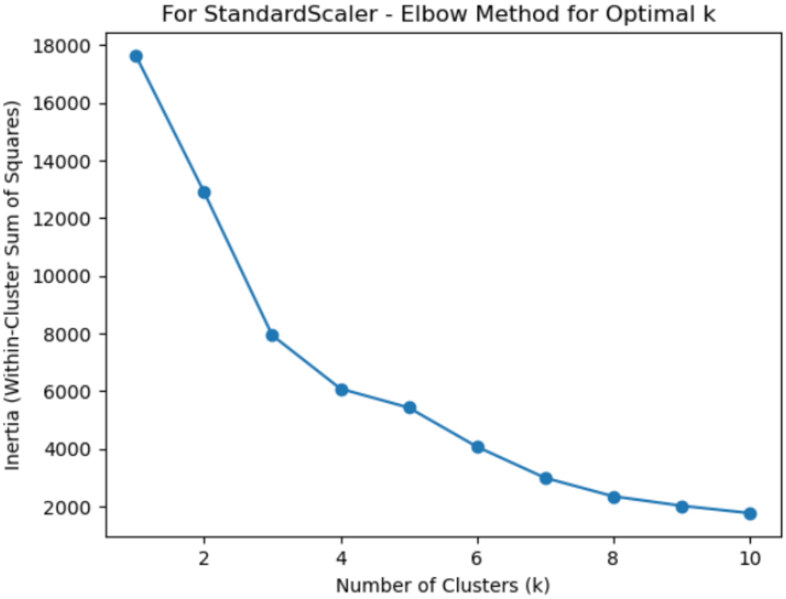
- **StandardScaler** transforms the Recency, Frequency, and Monetary values to have a mean of 0 and standard deviation of 1, which is useful when features have different units or ranges.
- **MinMaxScaler** rescales the same features to a fixed range between 0 and 1, preserving the original distribution shape while making the features directly comparable.

Using both allows us to test which scaling method gives better clustering performance in terms of compactness and separation (e.g., via silhouette score).

Finding Optimal number of Clusters (k-Means) Using Elbow Method:

1. **Using StandardScaler:**

   The Elbow Method is a graphical technique used to determine the optimal number of clusters (k) for K-Means clustering. In the case of the StandardScaler-transformed data, the plot of inertia (within-cluster sum of squares) against the number of clusters reveals a steep decline in inertia as the number of clusters increases from 1 to 4. This indicates that each additional cluster up to k = 4 significantly improves the model by reducing the variance within each cluster meaning the data points within clusters become more similar to each other.

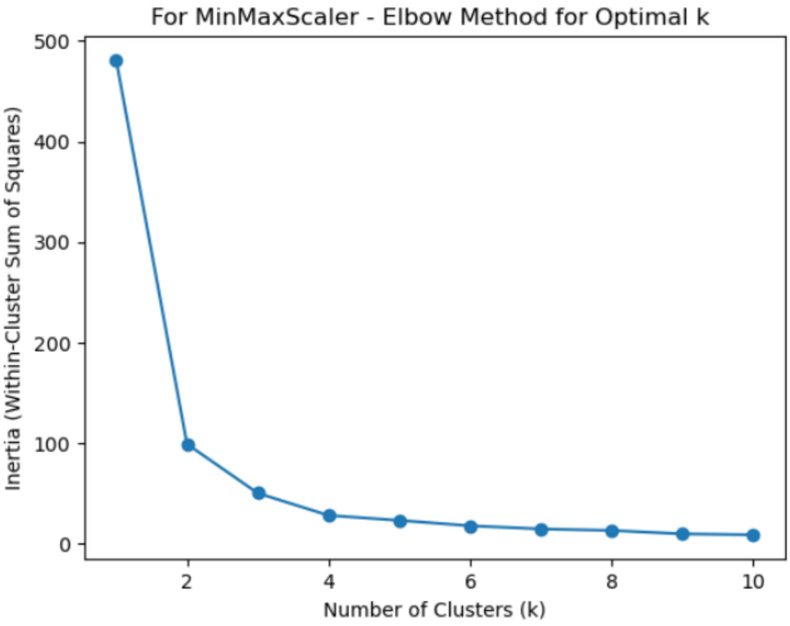


However, beyond k = 4, the curve begins to flatten, meaning the reduction in inertia becomes marginal with the addition of more clusters. This flattening indicates diminishing returns, where adding more clusters does not lead to a substantial improvement in the clustering quality. Therefore, choosing k=4 strikes a balance between model simplicity and performance, preventing overfitting while still capturing meaningful groupings in the customer behavior based on RFM features.

2. **Using MinMaxScaler:**

The Elbow Plot generated using MinMaxScaler-transformed data reveals a rapid decline in inertia when increasing the number of clusters from 1 to 2. This indicates that a significant portion of the within-cluster variance is captured by simply splitting the data into two clusters. Inertia, which measures how internally coherent the clusters are, drops sharply, suggesting that customers can be meaningfully divided into two distinct groups based on their RFM (Recency, Frequency, and Monetary) characteristics when scaled between 0 and 1.

Beyond k = 2, the inertia curve begins to flatten out, meaning the additional clusters bring only marginal improvement in reducing the within-cluster sum of squares. This flattening implies diminishing returns in clustering performance and suggests that increasing the number of clusters adds unnecessary complexity without significant gains in segmentation quality. Therefore, k = 2 is likely the most efficient and interpretable choice for identifying major customer segments in the dataset when using MinMax scaling.



3. **Enhanced Clustering Accuracy with MinMaxScaler: A Silhouette Score Perspective:**

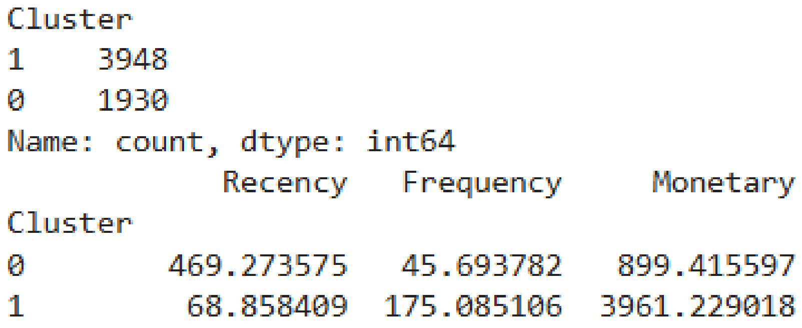
   In clustering analysis, the Silhouette Score is a crucial metric that measures how well each data point fits within its assigned cluster relative to other clusters. It ranges from -1 to 1, with higher values indicating better cluster cohesion and separation. In this analysis, applying MinMaxScaler followed by K-Means clustering with 2 clusters resulted in a Silhouette Score of 0.720, which is substantially higher than the 0.579 score obtained using StandardScaler with 4 clusters. This implies that the clusters formed under the MinMaxScaler approach are more compact and better separated from each other, leading to more reliable and interpretable groupings.

The higher score with fewer clusters suggests that the data scaled with MinMaxScaler is more amenable to clean segmentation with a simple structure. By rescaling all values between 0 and 1, MinMaxScaler preserves the relative distances and shape of the original distribution, which seems to enhance the performance of distance-based algorithms like K-Means in this case. In contrast, StandardScaler centers the data but may distort feature distributions if they are not normally distributed, potentially leading to suboptimal cluster boundaries.

From a business perspective, this finding is particularly valuable. It means that a simpler model with just two segments possibly high-value vs. low-value customers can capture most of the variation in customer behavior effectively. Such clarity in segmentation can simplify marketing strategies and decision-making, allowing targeted actions for each customer group with greater confidence.

4. **Cluster Distribution and Behavioural Profiling Using RFM Segmentation:**

The cluster size output is a crucial step in evaluating the results of customer segmentation. It shows how many customers fall into each identified cluster, giving insight into the relative size of each segment. For example, if one cluster contains a large number of customers while others are significantly smaller, it may represent a dominant or general customer group, whereas smaller clusters might highlight more specialized or niche segments. Understanding this distribution helps businesses prioritize their focus—whether it's retaining a high-volume segment or nurturing smaller but high-value groups.



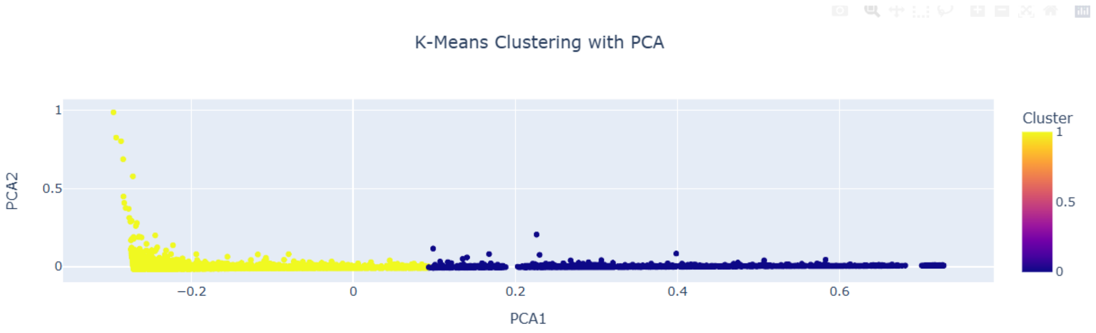
```
Cluster
1    3948
0    1930
Name: count, dtype: int64
            Recency   Frequency     Monetary
Cluster
0        469.273575   45.693782   899.415597
1         68.858409  175.085106  3961.229018
```

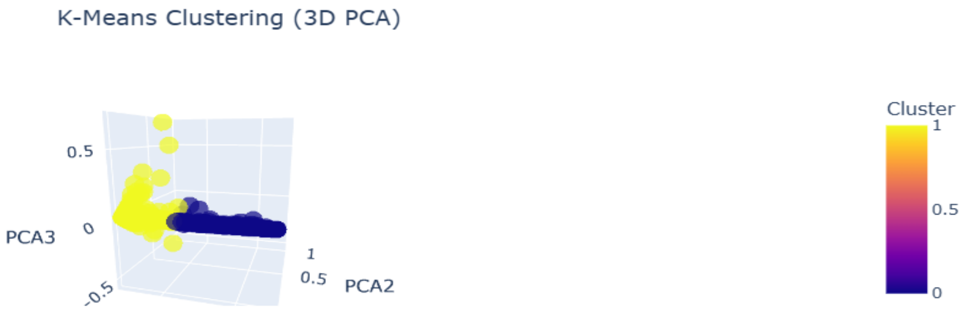
Additionally, the mean values of Recency, Frequency, and Monetary (RFM) for each cluster provide a behavioural profile of the customers within each group. For instance, a cluster with low Recency (meaning recent purchases), high Frequency (frequent transactions), and high Monetary value represents loyal, engaged, and profitable customers. In contrast, a cluster with high Recency (long time since last purchase), low Frequency, and low Monetary spending may represent inactive or at-risk customers.

These insights are instrumental in crafting targeted marketing strategies. High-value clusters can be rewarded with loyalty programs or exclusive offers, while low-engagement groups can be approached with reactivation campaigns or personalized incentives. By aligning marketing efforts with the specific characteristics of each cluster, businesses can improve customer retention, increase lifetime value, and enhance overall marketing ROI.

5. **Dimensionality Reduction for Enhanced Cluster Interpretation:**

   The PCA scatter plot serves as a powerful tool for visualizing customer segmentation outcomes derived from K-Means clustering. By reducing the three-dimensional RFM feature space (Recency, Frequency, Monetary) into two principal components, the plot highlights the separation between clusters. In this case, two distinct clusters emerge clearly along the PCA axes. Cluster 1, represented in yellow, is visibly more compact and aligned with customers who are recent purchasers, frequent buyers, and high spenders. Conversely, Cluster 0, shown in dark blue, is more spread out, indicating a more diverse group of customers with generally lower engagement and spending behavior.



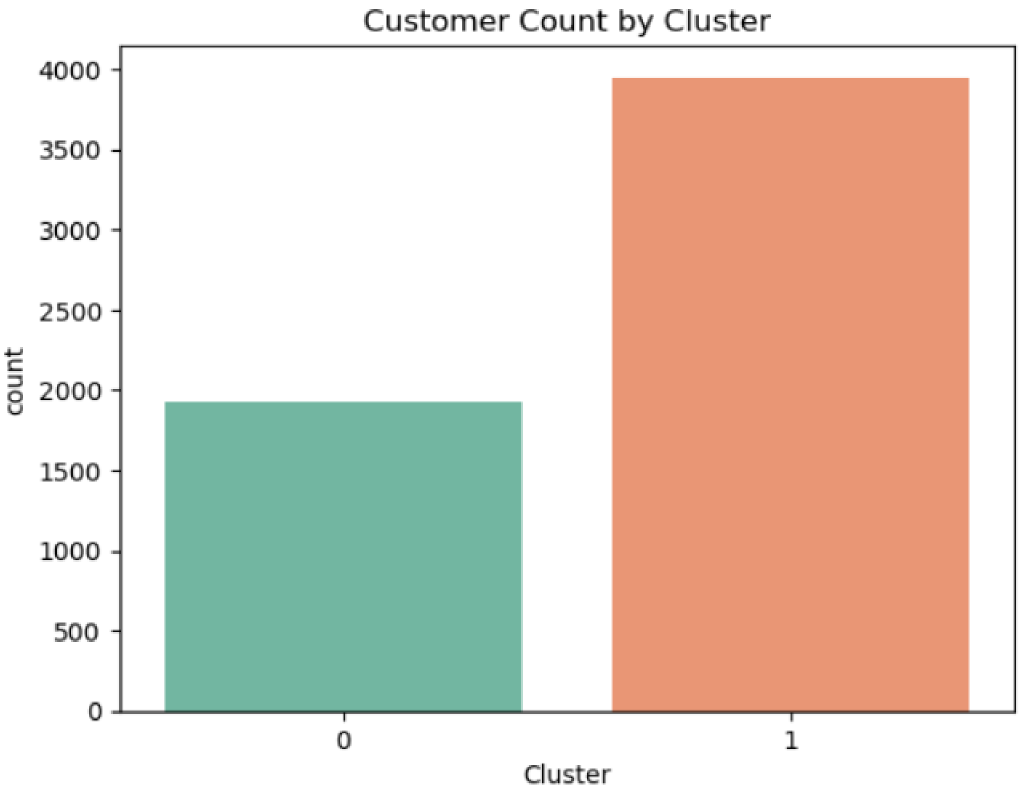


To enhance this interpretation, the 3D PCA plot incorporates a third principal component, capturing even more variance from the original dataset. This added dimension provides a deeper perspective on the clustering structure, making the separation between clusters even more apparent. The spatial distribution shows that the yellow cluster maintains a distinct and compact shape in the 3D space, reinforcing its identity as a high-value customer segment. Meanwhile, the dark blue cluster remains broader and less cohesive, consistent with its classification as a low-engagement group. These visual insights not only confirm the effectiveness of the K-Means model but also provide valuable guidance for targeted customer strategies.

Overall, the combination of 2D and 3D PCA plots validates the quality of the clustering outcome and supports strategic segmentation decisions. Businesses can use this information to prioritize retention efforts for the high-value cluster while designing re-engagement campaigns for the less active group.

Cluster profiling and interpretation:

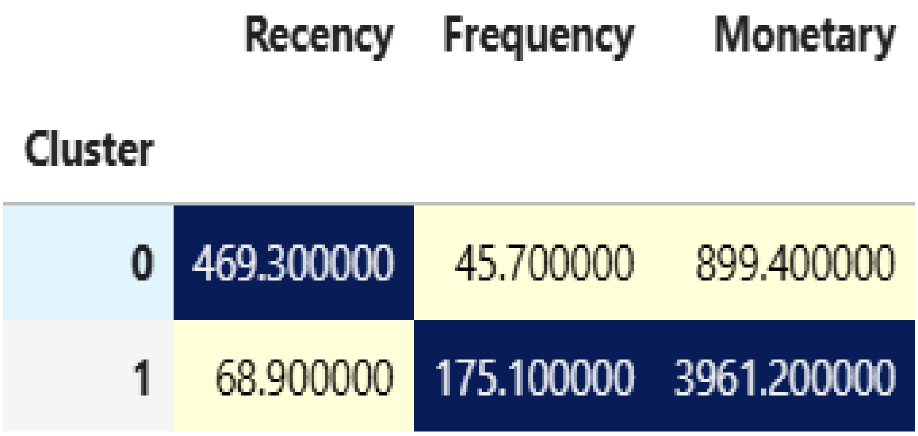
1. **Visualizing Customer Segmentation: A View into Cluster Populations:**



The value_counts().sort_index() output presents a numerical breakdown of how many customers fall into each cluster, organized by cluster label. This statistical insight provides a quick snapshot of the overall segmentation structure. A well-balanced distribution would indicate that the clustering algorithm has segmented the customer base into similarly sized groups. However, if the distribution is skewed—where one cluster has a much higher count than others—it may reflect a dominant customer behavior pattern or an overrepresentation of certain traits.

To complement this, the count plot offers a clear visual representation of these numbers. In this case, it becomes evident that Cluster 1 comprises significantly more customers than Cluster 0. This visual imbalance highlights the fact that a majority of customers exhibit similar behavior characteristics grouped into Cluster 1. Such findings are essential when crafting targeted marketing strategies, as they reveal not only the size of each segment but also help identify which group may require special attention—either for retention (if valuable) or re-engagement (if less active).

2. **Targeted Marketing Through RFM Cluster Summaries:**



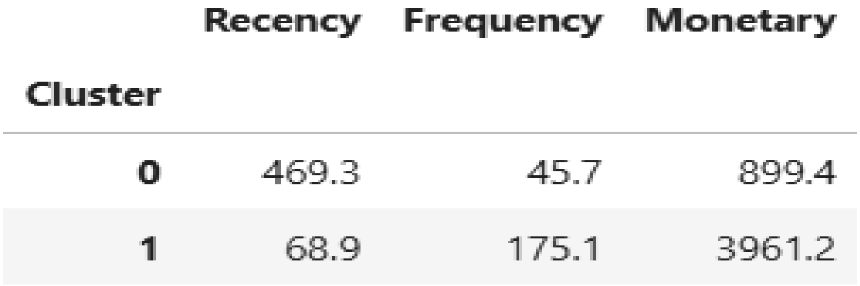
	Recency	Frequency	Monetary
Cluster			
0	469.300000	45.700000	899.400000
1	68.900000	175.100000	3961.200000

The cluster summary table presents a clear picture of customer behavior by analysing the average Recency, Frequency, and Monetary (RFM) scores within each cluster. These metrics provide a structured way to understand customer engagement and value. By grouping customers based on these aggregated values, businesses can identify patterns and form distinct marketing strategies for each segment. In this case, we observe two distinct clusters, each representing a unique customer profile.

Cluster 1 exhibits low Recency, meaning customers have made recent purchases, combined with high Frequency and high Monetary value, indicating they buy often and spend more. This group can be classified as the company's most loyal and profitable segment, deserving focused retention efforts, personalized offers, and loyalty rewards to further strengthen their relationship with the brand.

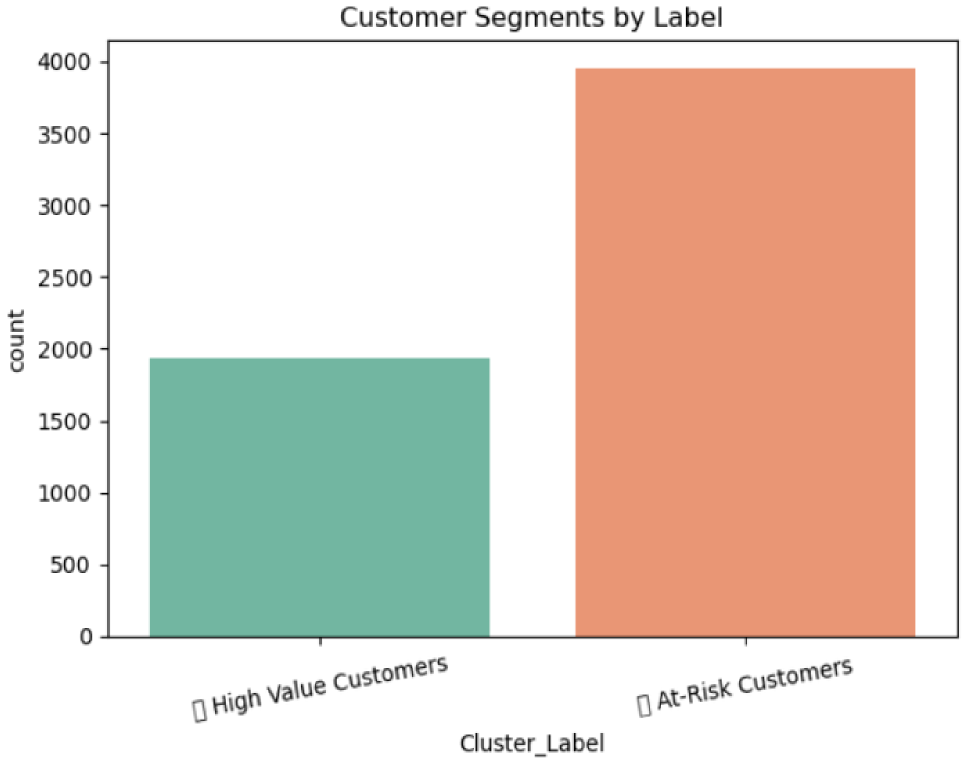
In contrast, Cluster 0 reflects high Recency (purchases made long ago), low Frequency, and low Monetary value, suggesting a group of inactive or low-value customers. These customers may have churned or lost interest in the company's offerings. As a result, they are prime candidates for re-engagement campaigns, such as reactivation emails, exclusive discounts, or surveys to understand their disengagement. This differentiation through clustering enables smarter decision-making in customer relationship management.

3. RFM-Based Cluster Interpretation and Segment Volume Analysis:



	Recency	Frequency	Monetary
Cluster			
0	469.3	45.7	899.4
1	68.9	175.1	3961.2

The RFM cluster summary table provides a concise yet powerful view of how customer behaviours differ across segments. By analyzing the average Recency, Frequency, and Monetary values for each cluster, businesses can decode the typical behavioural traits of each group. For instance, a cluster with low Recency, high Frequency, and high Monetary value reflects customers who shop regularly, spend more, and have purchased recently making them ideal targets for loyalty and upsell campaigns. On the other hand, clusters with high Recency (indicating long gaps since last purchase), low Frequency, and low Monetary value are indicative of disengaged or low-spending customers who may need reactivation strategies or could be de-prioritized in resource allocation.



The count plot of customer segments by cluster label visually highlights the distribution of customers across predefined labels such as High-Value, Low-Value, Occasional, or Churned. This visual aid is critical in understanding which customer types dominate the base. For example, if the majority of customers fall into the *Low-Value* segment, it may prompt a business to re-evaluate its acquisition or retention strategies. Conversely, a balanced distribution can signal a healthy mix of customer types, allowing for diversified targeting strategies tailored to the needs and value of each segment. This count-based visualization complements the cluster summary table by providing a sense of scale and strategic focus for customer engagement efforts.