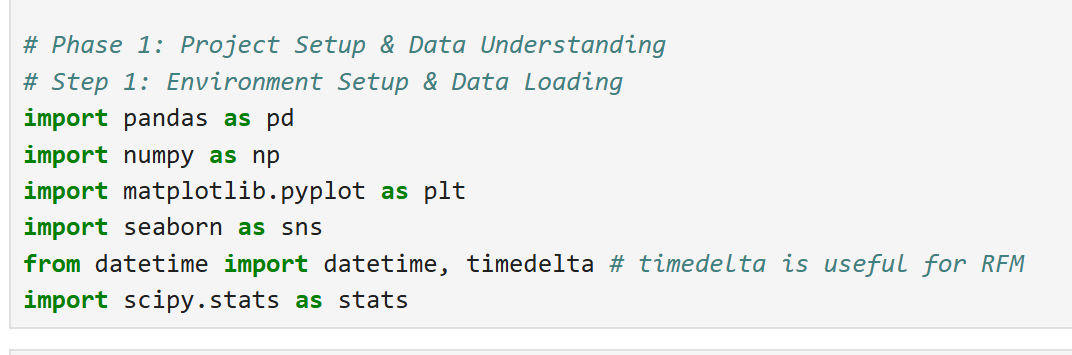
1. **Initial Python Environment Setup for Data Analysis**

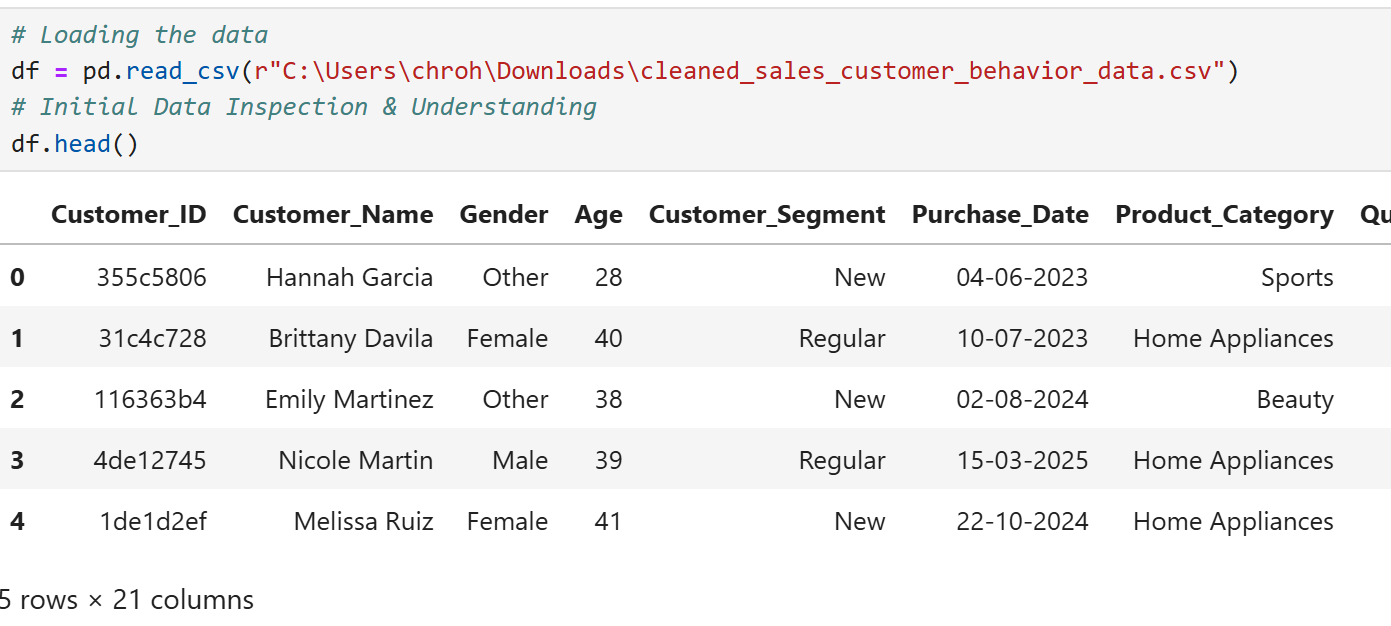
To import core Python libraries necessary for data loading, manipulation, numerical computation, visualization, and statistical analysis. This is the foundational step before any data work begins.



**Observation:**  
The code imports standard libraries: pandas (for data structures/analysis), numpy (numerical operations), matplotlib and seaborn (plotting), datetime (date/time handling, with a note for RFM), and scipy.stats (statistical functions). This prepares the workspace for a typical data analysis project.

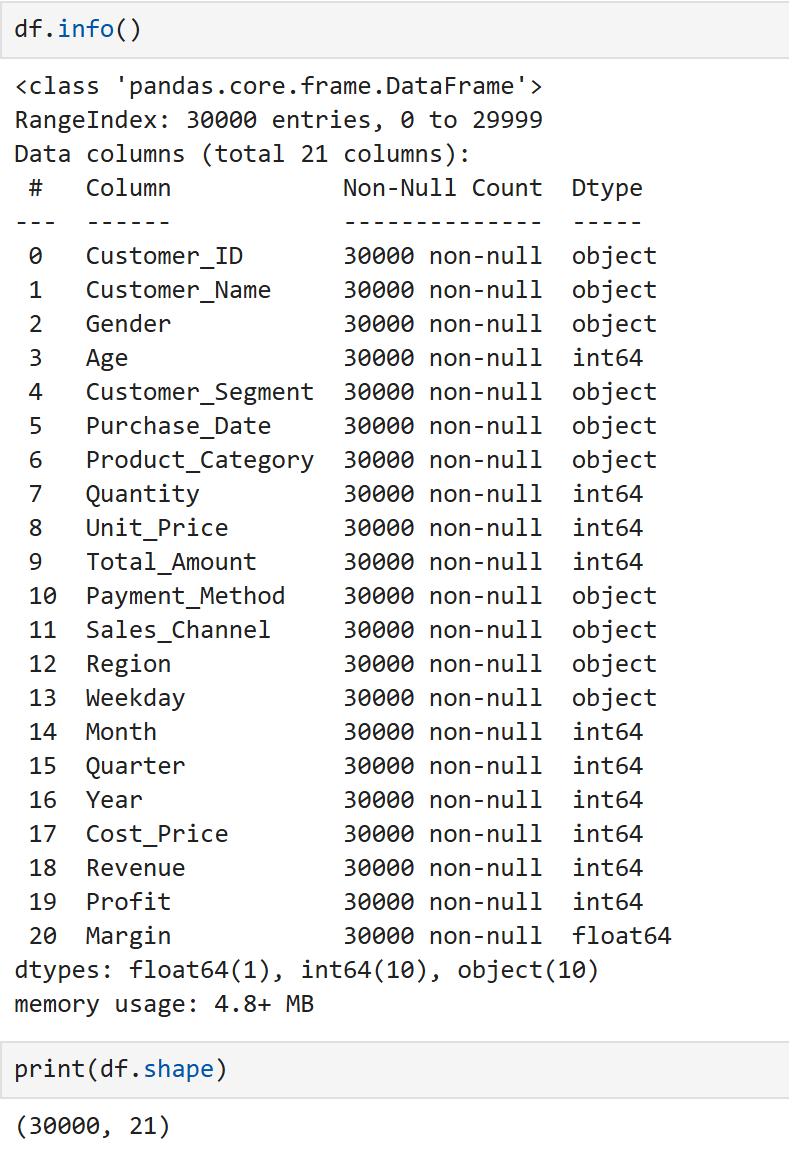
**2. Data Loading and Initial Inspection**

To import the 'cleaned\_sales\_customer\_behavior\_data.csv' dataset into a pandas DataFrame and display its first few rows. This allows for a quick verification of successful loading and a preliminary understanding of the data's structure, column names, and sample values.



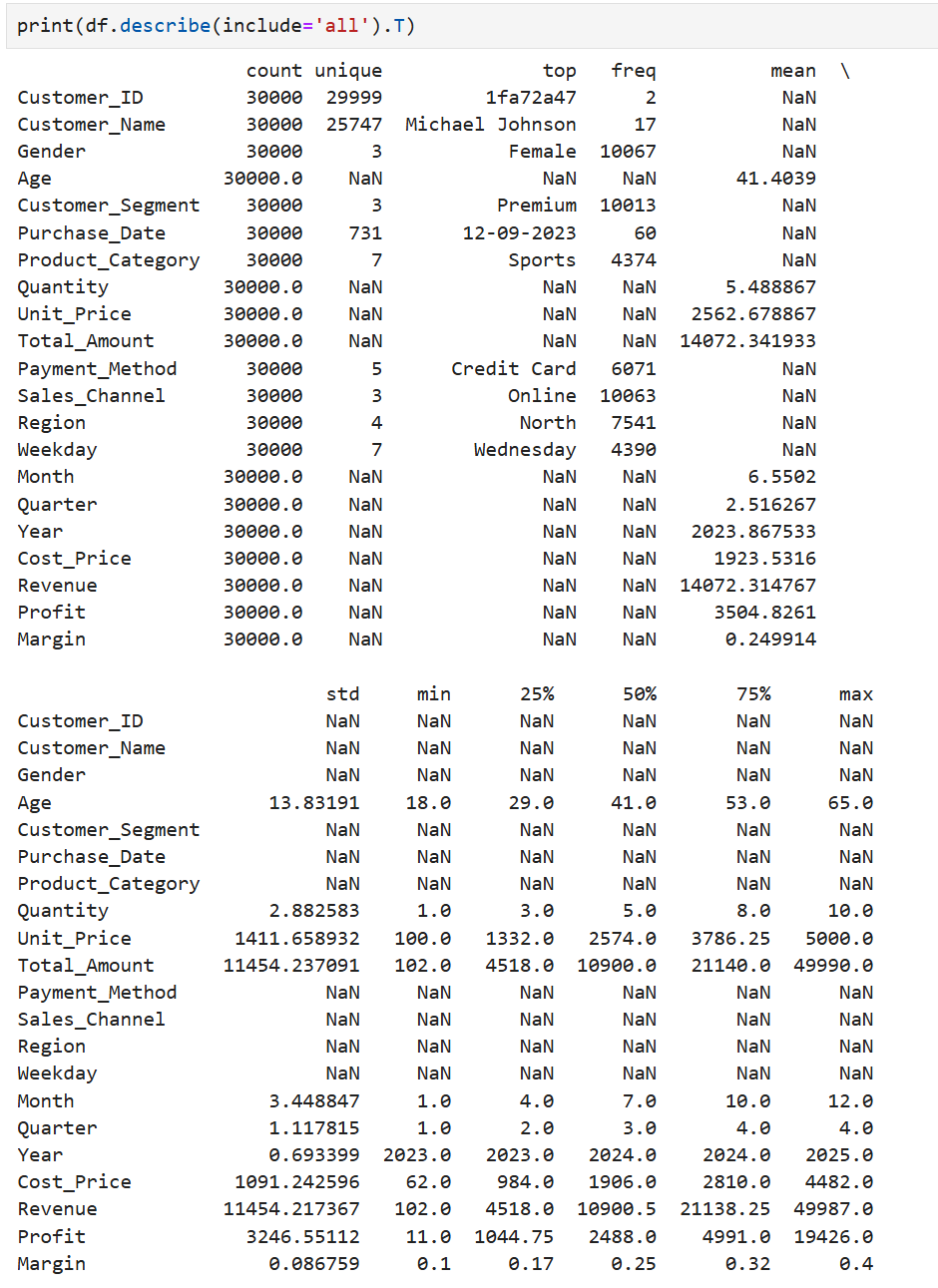
**Observation:**  
The code successfully loads the CSV file. The df.head() output shows the first 5 rows of the dataset, revealing columns like Customer\_ID, Customer\_Name, Gender, Age, Customer\_Segment, Purchase\_Date, and Product\_Category. The dataset contains 21 columns in total, providing customer demographic and transaction information.

**3. DataFrame Structure and Data Type Inspection**

To obtain a summary of the DataFrame, including its dimensions (rows, columns), the data type of each column, and the count of non-null values. This is crucial for understanding the dataset's structure, identifying missing data, and planning data type conversions.

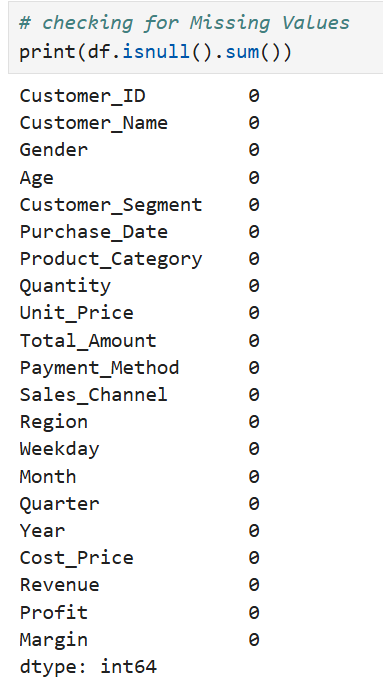
**Observation:**  
The dataset comprises 30,000 entries (rows) and 21 columns. All columns are fully populated, showing 30,000 non-null values, indicating no missing data. The data types are a mix: 10 object columns (e.g., Customer\_ID, Purchase\_Date), 10 int64 columns (e.g., Age, Quantity), and 1 float64 column (Margin). The Purchase\_Date being an object suggests it might need conversion to a datetime type.

**4. Comprehensive Descriptive Statistics**  
To generate a detailed statistical summary for all columns in the DataFrame. This includes measures of central tendency, dispersion, and range for numerical columns, and counts of unique values, the most frequent value (top), and its frequency for categorical columns. The .T transposes the output for easier column-wise reading.

****

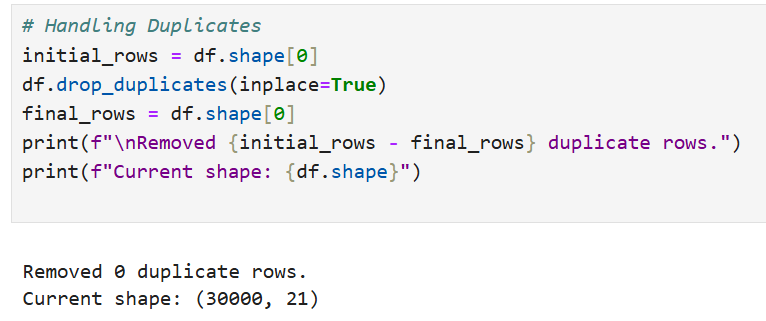
**Observation:**  
The output provides a statistical profile for each of the 21 columns.  
For numerical columns like Age and Total\_Amount, it details mean, standard deviation, min/max, and quartiles (e.g., average age is ~41.4).  
For categorical columns like Gender and Product\_Category, it shows the number of unique values, the most frequent category, and its count (e.g., "Female" is the most common gender, "Sports" the most common product category).  
A Customer\_ID (1fa72a47) appears twice, which might warrant investigation.

**5. Missing Value Check**  
To identify and quantify missing (null) values in each column of the DataFrame. This is a critical data quality check, as missing values can impact analysis and model performance.

****

**Observation:**  
The output shows a count of zero missing values for every column in the DataFrame. This indicates that the dataset is complete and does not require any immediate missing value imputation or handling strategies.

**6. Duplicate Row Removal and Verification**To identify and eliminate any completely identical rows within the DataFrame, then confirm the number of rows removed and the resulting DataFrame dimensions. This is a data cleaning step to ensure data integrity.



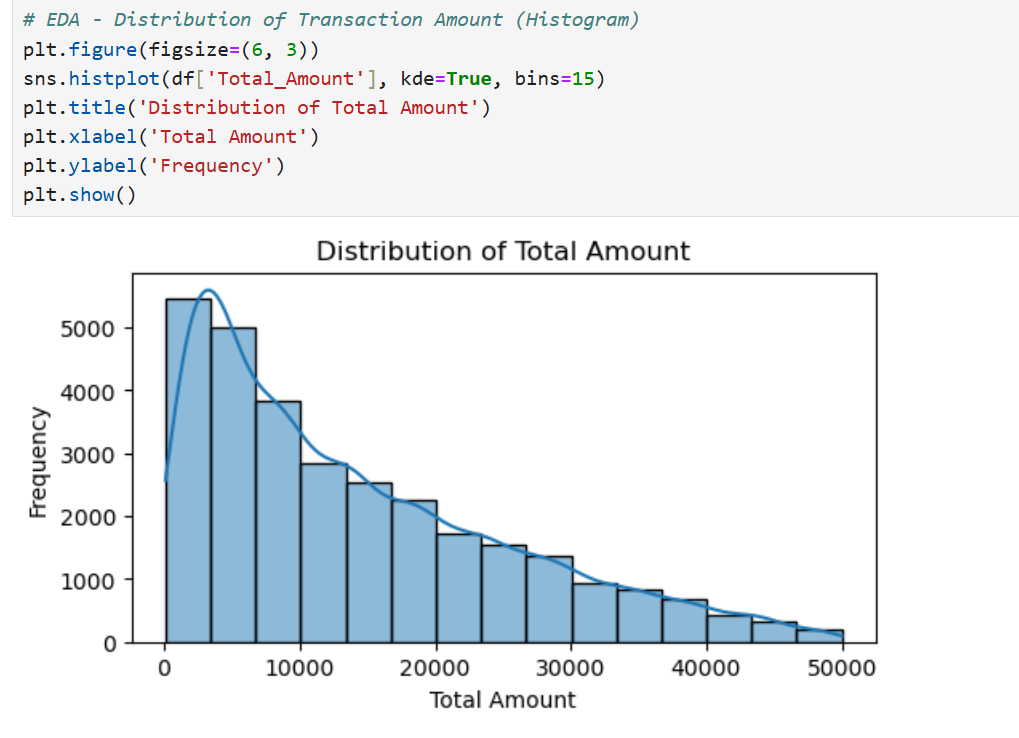
**Observation:**  
The code successfully executed the duplicate removal process. The output "Removed 0 duplicate rows" and the "Current shape: (30000, 21)" indicate that no duplicate rows were found in the dataset, so the DataFrame's dimensions remained unchanged.

**7. Visual Outlier Detection for Numerical Features**  
To generate boxplots for selected numerical columns (Age, Quantity, Unit\_Price, Total\_Amount, Revenue, Profit) to visually inspect their distributions and identify potential outliers that might require further investigation or specific handling.



**Observation:**  
The boxplots reveal that Age and Quantity have relatively symmetrical distributions with few outliers. Unit\_Price, Total\_Amount, Revenue, and Profit exhibit significant positive skewness with numerous high-value outliers, suggesting further investigation or specific handling for these extreme values. Profit also shows some negative outliers.

**8. Distribution of Total Transaction Amount (Histogram & KDE)**  
To visualize the frequency distribution of the 'Total\_Amount' column using a histogram and Kernel Density Estimate (KDE). This helps understand the shape, central tendency, and spread of transaction values.

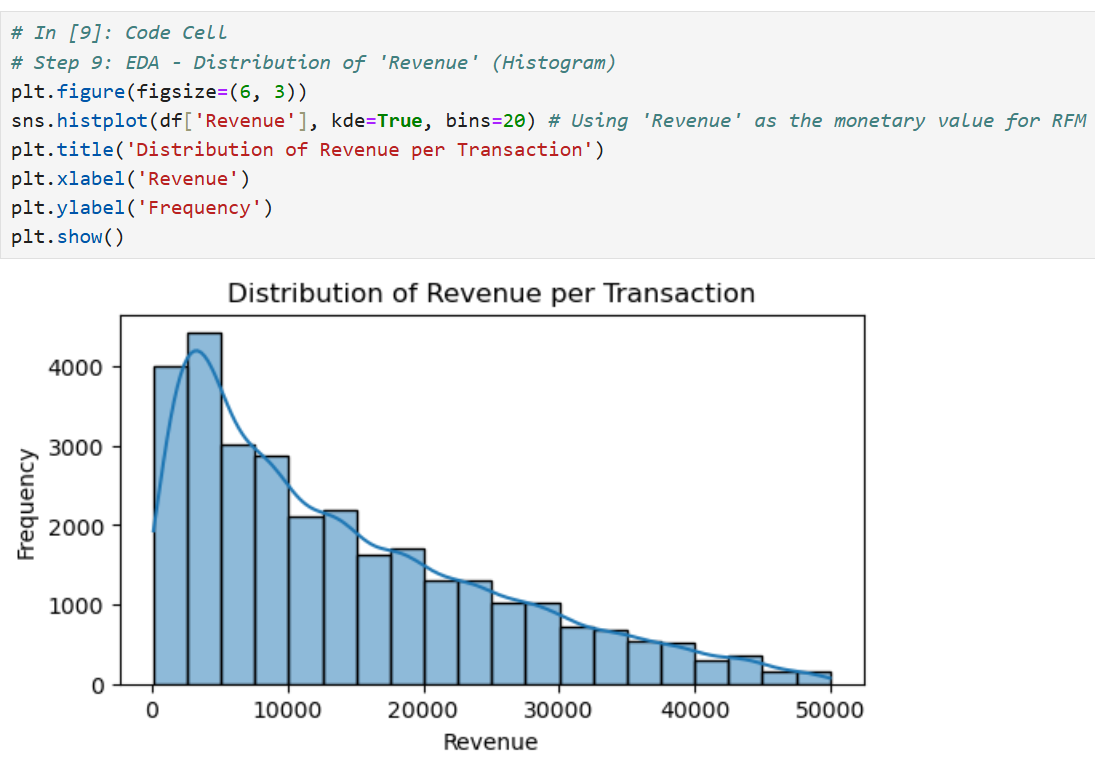
****

**Observation:**  
The 'Total\_Amount' distribution is right-skewed (positively skewed), with most transactions having lower values and a tail of fewer, higher-value transactions.

**Conclusion:**  
The skewness suggests that while most purchases are of modest value, a smaller number of high-value transactions significantly influence the overall spend. This pattern is common in sales data and may warrant further investigation into these high-value customers or transactions.

**9. Distribution of Revenue per Transaction (Histogram & KDE)**

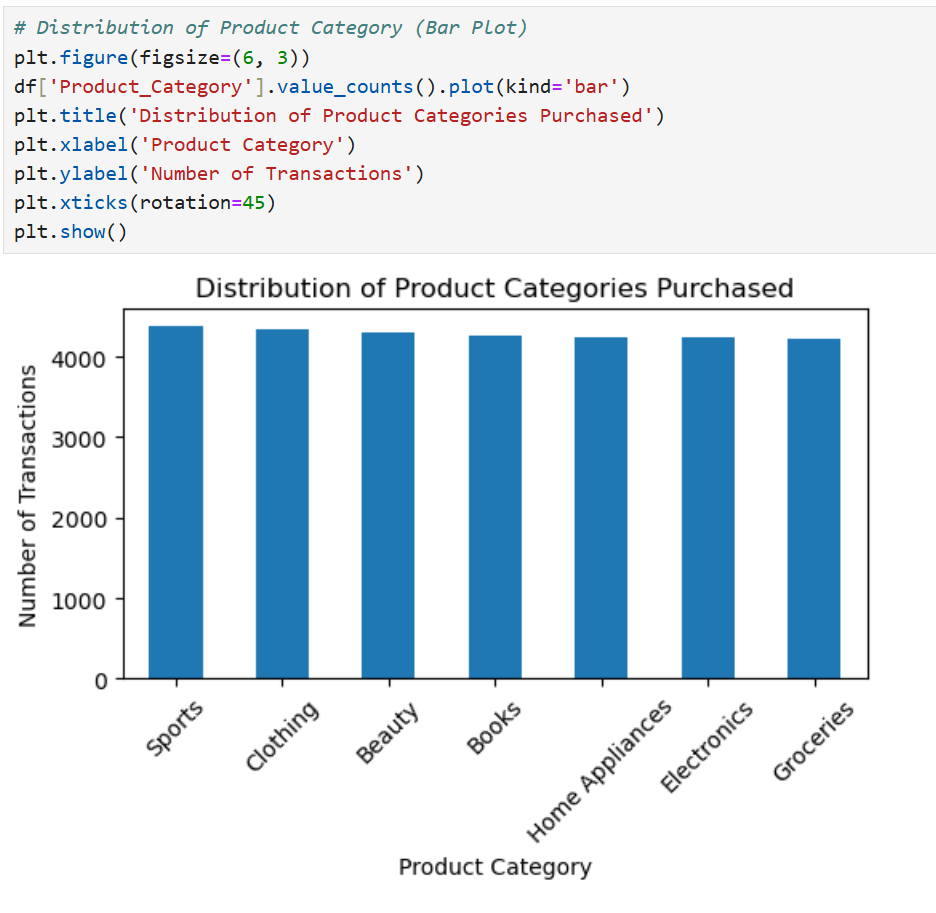
To visualize the frequency distribution of the 'Revenue' column using a histogram and Kernel Density Estimate (KDE). This helps understand its shape, central tendency, and spread, particularly as 'Revenue' is noted for use as the monetary value in RFM (Recency, Frequency, Monetary) analysis.



**Observation:**  
The 'Revenue' distribution is right-skewed (positively skewed). The majority of transactions generate lower revenue values, with a decreasing frequency of transactions as revenue increases, forming a long tail.

**Conclusion:**  
The pronounced right skew indicates that a small number of high-revenue transactions contribute significantly to the total revenue, while most transactions are of lower value. This pattern is typical for sales data and crucial for understanding customer value in RFM segmentation.

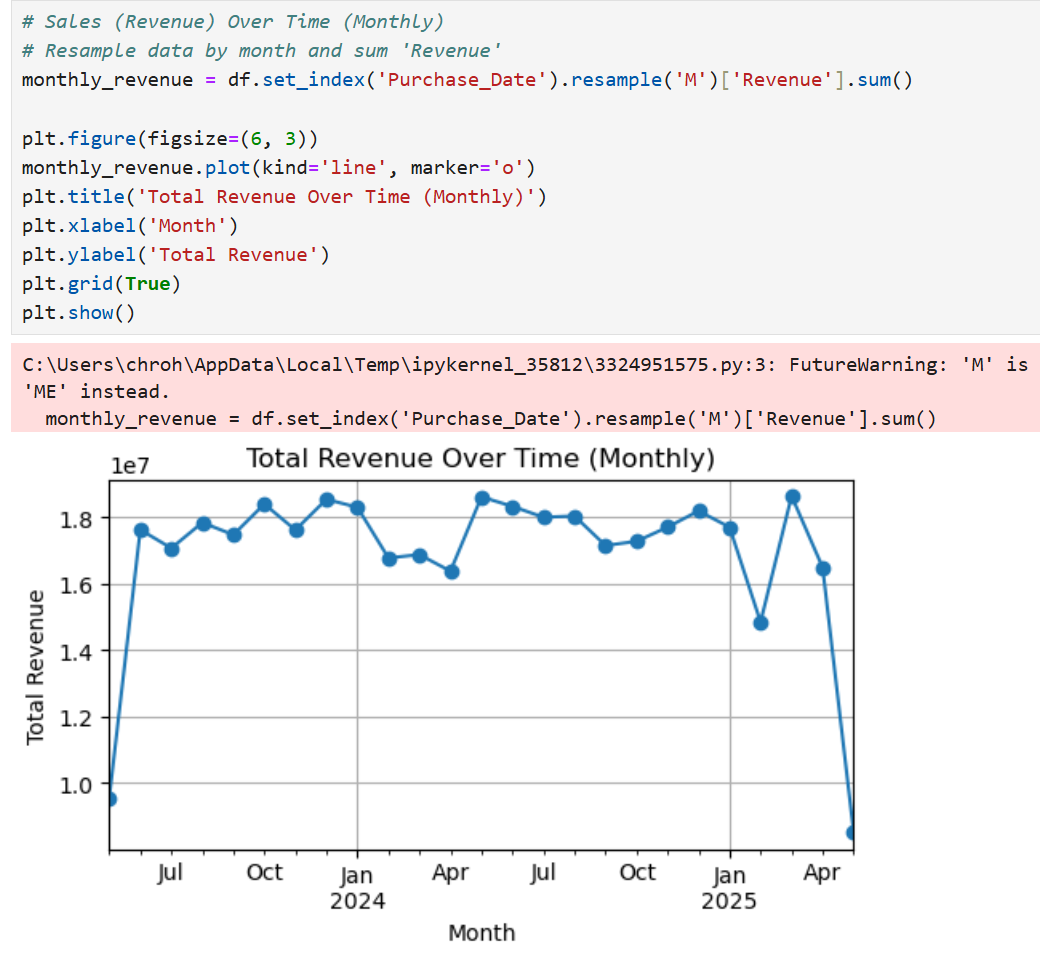
**10. Distribution of Product Categories Purchased (Bar Plot)**To visualize the frequency of transactions for each 'Product\_Category' using a bar chart. This helps identify which product categories are most or least frequently purchased.



**Observation:**  
The bar chart shows that all listed product categories (Sports, Clothing, Beauty, Books, Home Appliances, Electronics, Groceries) have a remarkably similar number of transactions, each hovering just above 4000.

**Conclusion:**  
There is a fairly even distribution of purchase frequency across all product categories within this dataset. No single category significantly dominates in terms of the number of transactions, suggesting diverse customer purchasing habits or a balanced product offering.

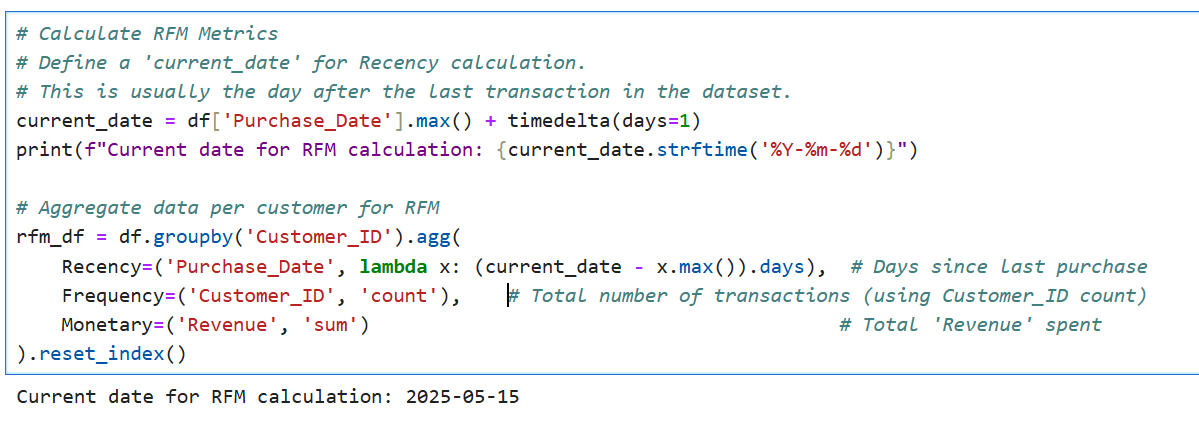
**11. Total Monthly Revenue Trend**  
To aggregate 'Revenue' data by month and visualize the resulting trend over time using a line plot. This helps in understanding sales performance, seasonality, and identifying significant changes in revenue.



**Observation:**  
Monthly revenue generally fluctuates between approximately 1.6e7 and 1.8e7, showing some peaks (e.g., around Oct-Jan 2024, and again around Nov 2024-Jan 2025). There is a very sharp and significant drop in revenue in the last plotted month (April 2025). A FutureWarning indicates the resampling alias 'M' should be updated to 'ME'.

**Conclusion:**  
While revenue exhibits some cyclical behavior with periodic peaks, the drastic decline in April 2025 is the most prominent feature. This could signify incomplete data for that final month, a data error, or a genuine, severe downturn in business that warrants immediate investigation. The code warning should also be addressed for future compatibility.

**12. RFM Metric Calculation for Customer Segmentation**  
To transform raw transactional data into Recency (R), Frequency (F), and Monetary (M) values for each unique customer. These metrics are fundamental for understanding customer behavior and performing value-based segmentation.

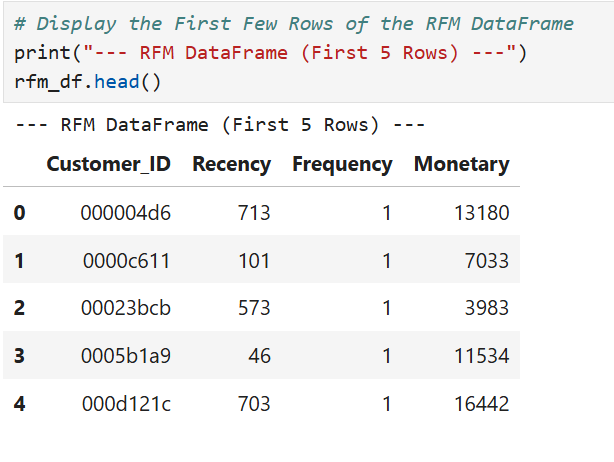
****

**Observation:**  
The code first establishes a current\_date (2025-05-15), which is one day after the last transaction date in the dataset, to serve as a reference for Recency calculations. It then groups the data by Customer\_ID and calculates:

* **Recency:** The number of days between the current\_date and the customer's last purchase date.
* **Frequency:** The total number of transactions made by the customer.
* **Monetary:** The total revenue generated by the customer.  
  These aggregated metrics are stored in a new DataFrame called rfm\_df.

**Conclusion:**  
The rfm\_df DataFrame now contains the core RFM values for each customer. This structured data is ready for the next steps in RFM analysis, which typically involve scoring these metrics and segmenting customers into distinct groups (e.g., champions, loyal customers, at-risk, etc.) to tailor marketing strategies.

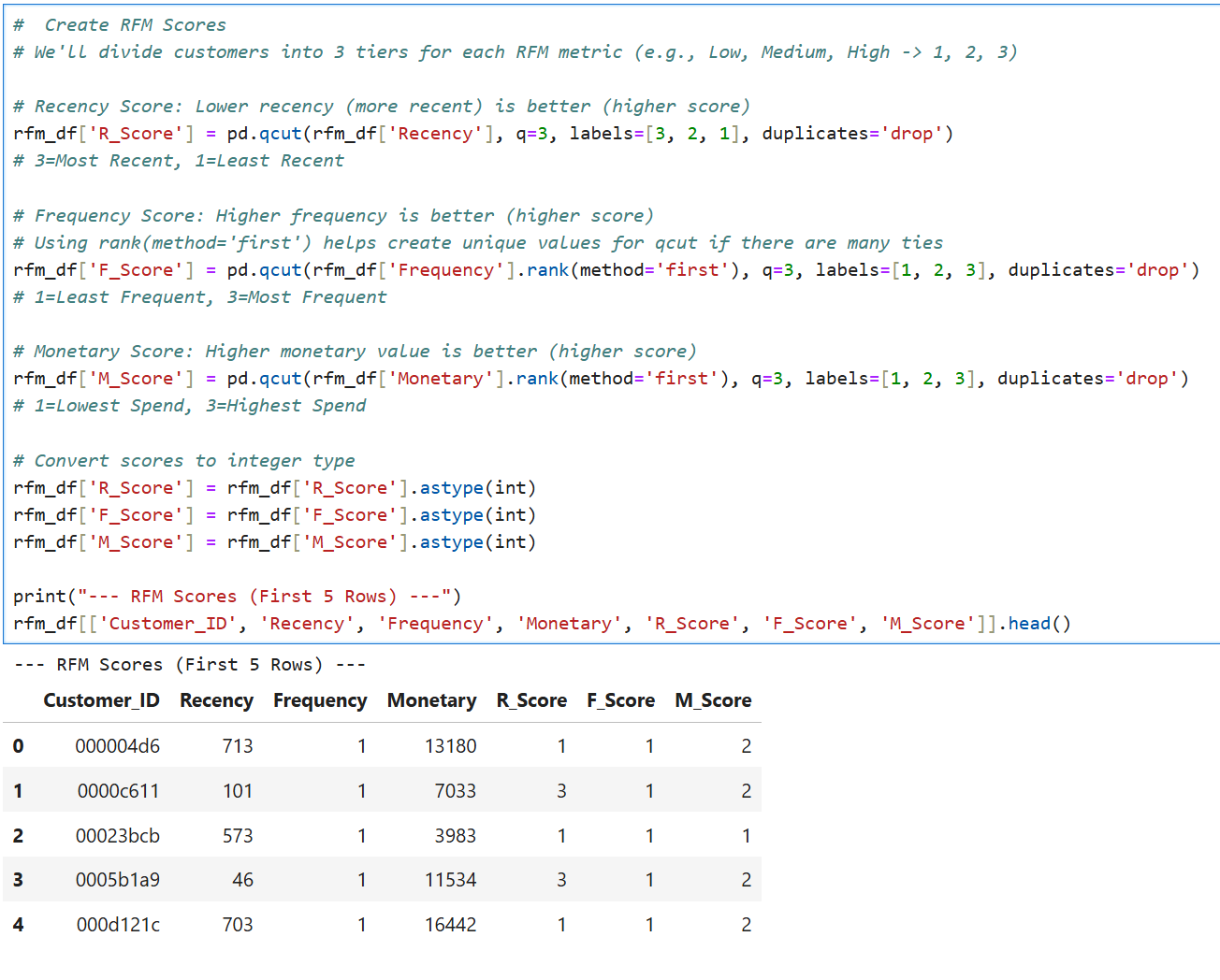
**13. RFM DataFrame Preview**  
To display the first few rows of the newly created rfm\_df DataFrame. This allows for a quick inspection and verification of the calculated Recency, Frequency, and Monetary values for a sample of customers.

****

**Observation:**  
The output shows the first 5 rows of the rfm\_df, with columns Customer\_ID, Recency, Frequency, and Monetary. For these initial 5 customers, Recency varies (e.g., 713 days, 46 days), Frequency is consistently 1, and Monetary values also differ.

**Conclusion:**  
The RFM metrics have been successfully computed and are available in the rfm\_df. The preview indicates that at least the first few customers displayed are one-time purchasers (Frequency=1). Further analysis on the entire rfm\_df will reveal the broader distribution of these metrics.

**14. RFM Score Calculation and DataFrame Augmentation**  
To convert the continuous Recency, Frequency, and Monetary values into discrete scores (1, 2, or 3) by dividing customers into three tiers (quantiles) for each metric. This simplifies RFM analysis and allows for easier customer segmentation based on these scores.

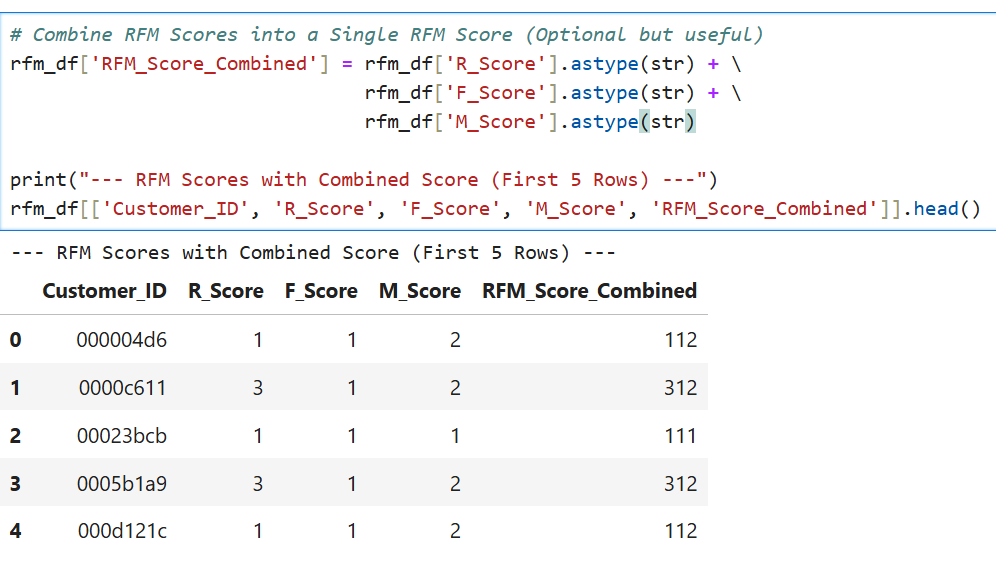
****

**Observation:**  
The code uses pd.qcut to create R\_Score, F\_Score, and M\_Score.

* **Recency (R\_Score):** Lower recency (more recent) gets a higher score (3=Most Recent, 1=Least Recent).
* **Frequency (F\_Score):** Higher frequency gets a higher score (3=Most Frequent, 1=Least Frequent). rank(method='first') is used to handle ties in frequency before qcut.
* **Monetary (M\_Score):** Higher monetary value gets a higher score (3=Highest Spend, 1=Lowest Spend). rank(method='first') is used for monetary values as well.  
  The scores are then converted to integer type. The rfm\_df.head() output shows the new R\_Score, F\_Score, and M\_Score columns alongside the original RFM values. For example, customer 000004d6 with high Recency (713) gets R\_Score=1, Frequency=1 gets F\_Score=1, and Monetary=13180 gets M\_Score=2.

**Conclusion:**  
The RFM DataFrame has been successfully augmented with R, F, and M scores (1-3). This tiered scoring system provides a standardized way to compare and group customers, paving the way for creating combined RFM segments and targeted marketing strategies. The use of rank before qcut for Frequency and Monetary helps in managing potential issues with many tied values, which is common for these metrics.

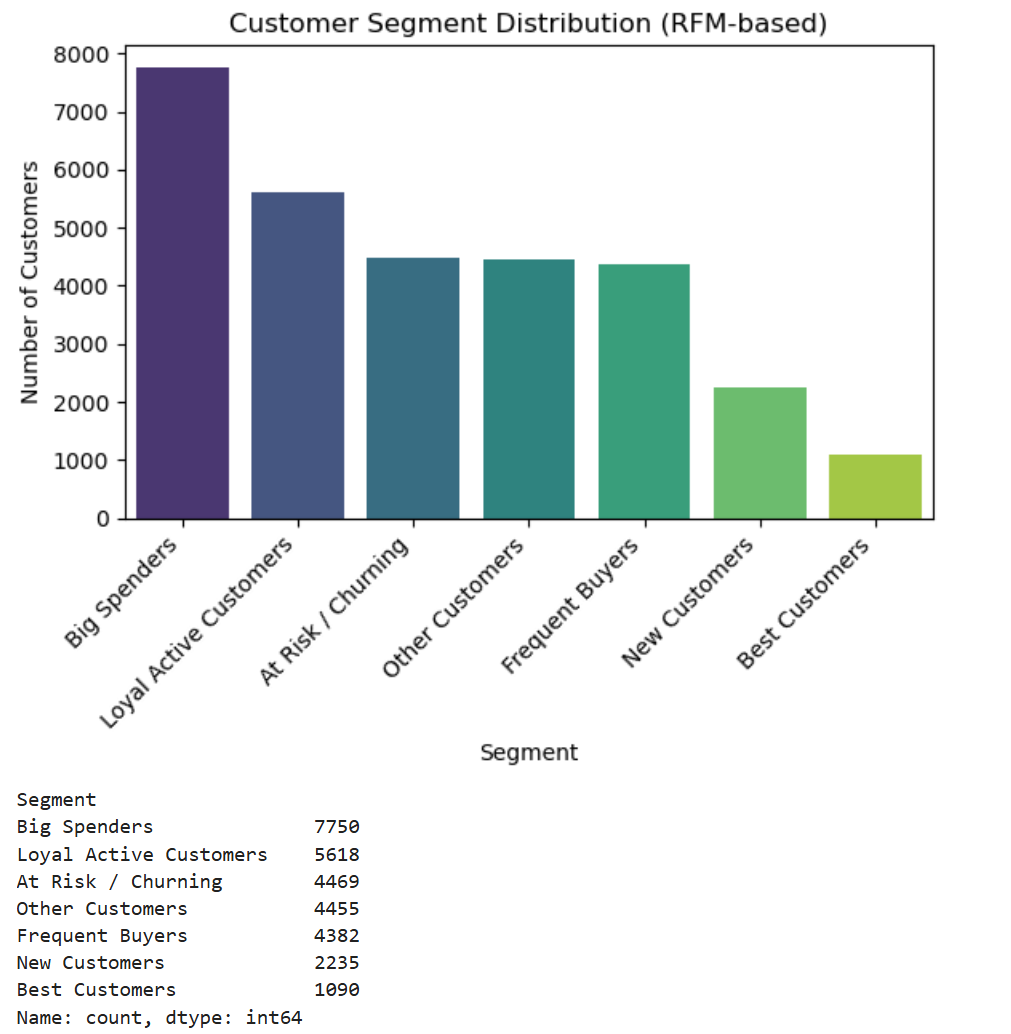
**15. Creation of Combined RFM Score**  
To concatenate the individual R\_Score, F\_Score, and M\_Score into a single string representation (e.g., "312"). This creates a composite RFM score that provides a quick, at-a-glance summary of a customer's RFM profile, facilitating easier high-level segmentation and analysis.

****

**Observation:**  
The code converts R\_Score, F\_Score, and M\_Score to strings and then concatenates them to form the RFM\_Score\_Combined column. The head() output shows this new column, where, for instance, a customer with R\_Score=1, F\_Score=1, and M\_Score=2 now has an RFM\_Score\_Combined of "112".

**Conclusion:**  
A combined RFM score has been successfully added to the DataFrame. This composite score allows for a more holistic view of customer value and behavior based on all three RFM dimensions. It can be directly used to define customer segments or as a basis for more sophisticated segmentation logic.

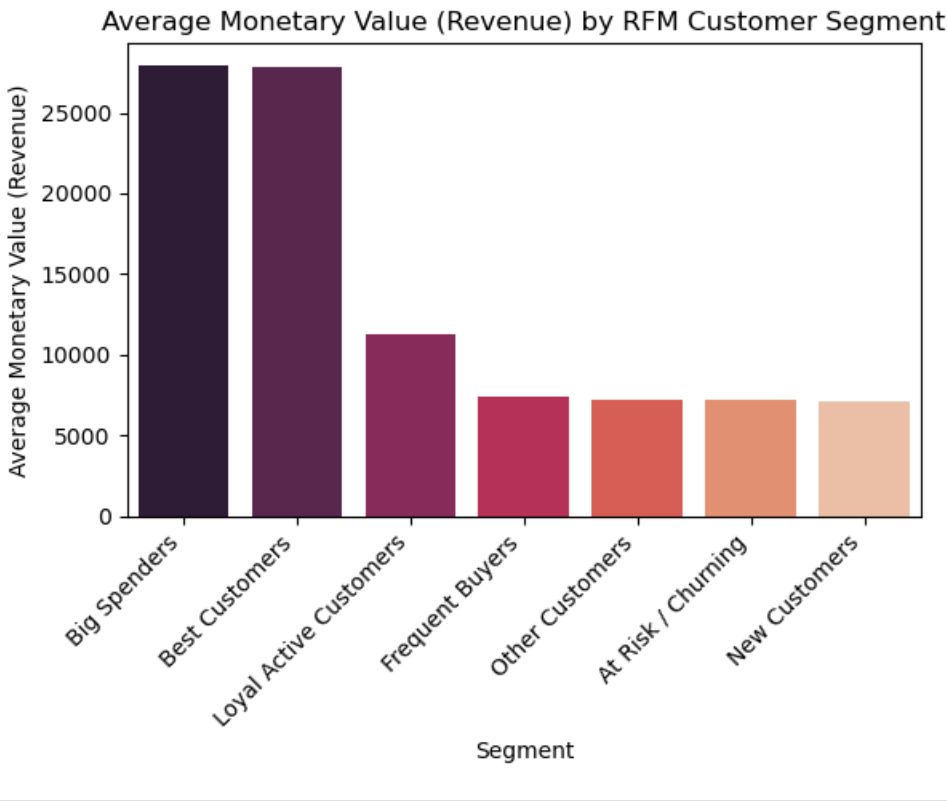
**16. Customer Segment Distribution (RFM-based)**  
To visualize the number of customers falling into each predefined RFM segment. This provides an overview of the customer base structure according to their value and behavior.

****

**Observation:**  
The bar chart displays the count of customers in each segment. "Big Spenders" is the largest segment (7750 customers), followed by "Loyal Active Customers" (5618). Segments like "At Risk / Churning," "Other Customers," and "Frequent Buyers" have similar, moderate sizes (around 4400-4500 each). "New Customers" (2235) and "Best Customers" (1090) are the smallest segments.

**Conclusion:**  
The segmentation reveals a substantial group of "Big Spenders." While there's a healthy number of "Loyal Active Customers," the "Best Customers" segment is relatively small, indicating potential for nurturing more customers into this high-value group. The size of the "At Risk / Churning" segment also warrants attention for retention strategies.

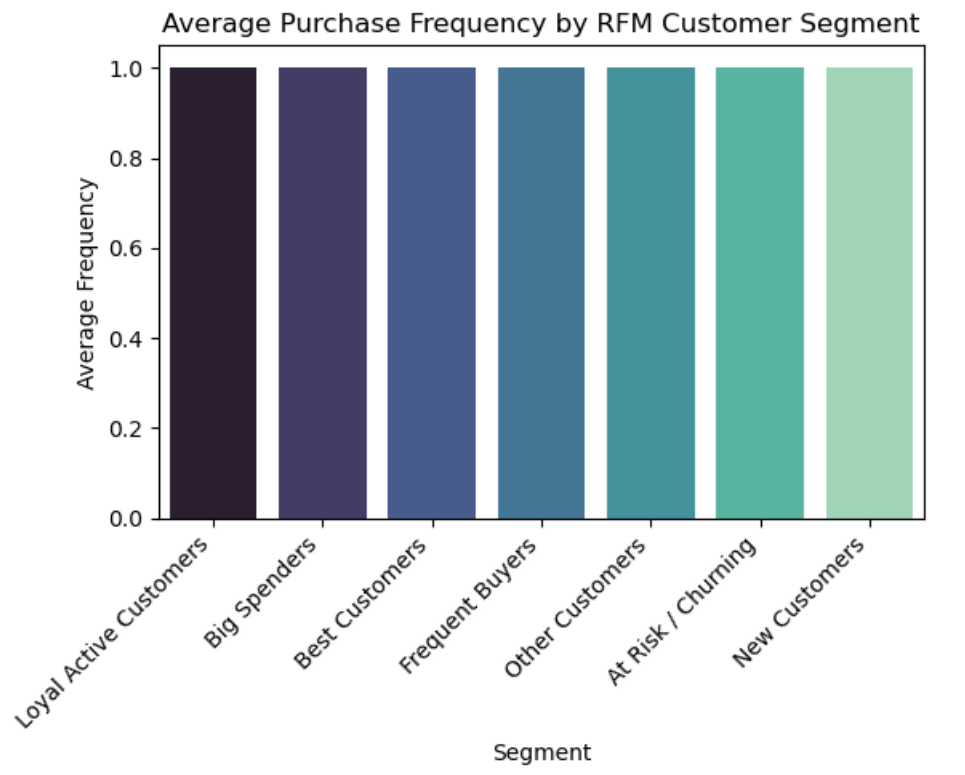
**17. Average Monetary Value (Revenue) by RFM Customer Segment**  
To visualize and compare the average revenue contributed by customers within each defined RFM segment. This helps in understanding the typical spending power associated with each segment.

****

**Observation:**  
The "Big Spenders" and "Best Customers" segments exhibit by far the highest average monetary value (revenue), both exceeding 25,000. "Loyal Active Customers" have the next highest average, around 11,000. The remaining segments ("Frequent Buyers," "Other Customers," "At Risk / Churning," and "New Customers") show considerably lower and relatively similar average monetary values, all around 7,000.

**Conclusion:**  
The RFM segmentation effectively differentiates customer groups by their average spending. "Big Spenders" and "Best Customers" are significantly more valuable on an average per-customer basis. This reinforces the importance of focusing retention and high-value offerings on these top segments, while strategies for other segments might aim to increase their average spending.

**18. Average Purchase Frequency by RFM Customer Segment**  
To visualize and compare the average purchase frequency of customers within each defined RFM segment. This helps assess the repeat purchase behavior characteristic of each segment.



**Observation:**  
The bar chart indicates that the average purchase frequency is remarkably consistent across all RFM segments, hovering very close to 1.0 for each. Segments like "Loyal Active Customers," "Big Spenders," and "Best Customers" show an average frequency at or just slightly above 1, while others are also at 1.

**Conclusion:**  
The average purchase frequency of approximately 1 across all segments suggests that, on average, customers in this dataset (within these defined segments) tend to be single-purchase customers. This implies that the "Frequency" dimension, when averaged at the segment level, does not strongly differentiate these particular customer groups in this dataset.

**Findings and Strategic Recommendations for Sales Optimization**

**Findings**

By looking at the sales information, we found out some important things:

1. **Most Customers Buy Only Once:**
   * A really big discovery was that almost all customers, no matter how we grouped them, usually only made one purchase. Even those we called "Loyal" or "Frequent Buyers" mostly just bought something a single time during the period we looked at.
2. **Some Customers Spend a Lot in That One Purchase:**
   * We found a large group of "Big Spenders." These customers spend a lot of money when they do buy, even if it's just once. There's also a smaller group of "Best Customers" who also spend a lot.
3. **Understanding Our Customer Groups:**
   * Our "Best Customers" (who buy recently, often, and spend a lot) are very valuable, but there aren't many of them. This means there's a chance to help more customers become like them.
   * We also have a group of customers "At Risk" of leaving because they haven't bought anything in a while. If we don't do something, we might lose their business.
   * Customers we called "Loyal" are buying recently and spending a good amount, but like everyone else, usually just once.
4. **People Like All Kinds of Products:**
   * It was interesting to see that customers buy pretty evenly from all product types (like Sports, Clothing, Beauty products, etc.). No single type of product is way more popular than others in terms of how many times it's bought.
5. **A Few Big Sales Make a Big Difference:**
   * When we looked at how much money each sale brought in, we saw that most sales are for smaller amounts. But, a few sales for very large amounts of money really add up and make a big impact on the total revenue.
6. **A Sudden Drop in Sales Last Month:**
   * When looking at sales month by month, we saw a big, unexpected drop in sales in the very last month of our information (April 2025).
7. **The Information We Used Was Good:**
   * The sales information we started with was clean and complete, which means we can trust what we found from looking at it. We just had to make sure dates were in the right format.

**Recommendations for Sales Optimization**

Based on what we found, here are some ideas to help the business improve sales:

1. **Focus on Getting Customers to Buy Again (This is Key!):**
   * **Problem:** Most people only buy once.
   * **What to do:**
     + After someone buys something, send them a friendly follow-up (like an email) suggesting other things they might like, or maybe a small discount on their next purchase.
     + Start a simple rewards program where customers get points or benefits for buying more than once.
     + For new customers, have a special welcome that encourages them to make a second purchase soon.
   * **Why:** If even a few more people buy a second time, it will make a big difference to sales.
2. **Treat Your Biggest Spenders Well:**
   * **Problem:** "Best Customers" and "Big Spenders" are valuable but might not come back often.
   * **What to do:**
     + For "Best Customers": Make them feel special with VIP treatment, like early news about new items or special deals.
     + For "Big Spenders": Look at what they bought and suggest other expensive items they might like.
   * **Why:** These customers bring in a lot of money, so it's important to keep them happy and interested.
3. **Try to Win Back Customers Who Might Leave:**
   * **Problem:** Some customers are "At Risk" of not buying again.
   * **What to do:** Send them a special offer (like a "We Miss You!" discount) or ask them why they haven't shopped lately.
   * **Why:** It's often easier to get an old customer back than to find a brand new one.
4. **Encourage People to Buy More Different Things:**
   * **Problem:** People like all sorts of products, but maybe they only buy one at a time.
   * **What to do:**
     + When they're buying something, suggest other useful items that go well with it, even from different product areas.
     + Offer deals if they buy a few different types of products together.
   * **Why:** If customers buy more items each time they shop, sales will go up.
5. **Figure Out That Sudden Sales Drop:**
   * **Problem:** Sales dropped sharply last month.
   * **What to do:** Check right away if all the sales information for April 2025 was entered correctly. If it was, then try to find out why sales went down (like, did something change in the market or with the business?).
   * **Why:** It's important to know if this is a data mistake or a real problem that needs fixing.
6. **Keep Using Information to Make Smart Choices:**
   * **Problem:** Customer habits can change.
   * **What to do:**
     + Keep an eye on how these customer groups change over time and see if your new ideas are working.
     + Think about other ways to understand customers, like trying to guess who might stop buying before they actually do.
     + Look for tools that can help offer even more personal suggestions to each customer.
   * **Why:** The more you understand your customers, the better you can meet their needs and improve sales.

By taking these steps, the business can use what it's learned about its customers to make better decisions, keep customers coming back, and sell more products.