IE 6400 – FOUNDATIONS OF DATA ANALYTICS
SUSHMITHA SUDHARSAN (+1 857 565 8800)
HARINI PRASAD VASISHT (+1 984 374 4836)
TANMAYI SHURPALI (+1 848 205 5511)

Project 2
E-commerce Data Analysis

1. Introduction

The project analyses e-commerce transaction data to provide actionable insights into customer purchasing behaviour. Using techniques such as Recency, Frequency, and Monetary (RFM) analysis, clustering, and visualization, the goal is to identify customer segments and recommend targeted marketing strategies.

2. Project Workflow

2.1 Data Acquisition and Inspection

- Data Source: E-commerce dataset from Kaggle.
- Steps:
  1. Loaded data using `pandas` and inspected structure using `.info()` and `.describe()`.
  2. Identified missing and duplicate values and addressed them appropriately.

2.2 Data Preprocessing

- Converted `InvoiceDate` to datetime format for temporal analysis.
- Added a `TotalPrice` column (`Quantity * UnitPrice`) to calculate purchase values.
- Removed duplicates and missing values for cleaner analysis.

2.3 Exploratory Data Analysis (EDA)

- Conducted exploratory analysis to understand the data distribution and identify trends.
- Visualized data distributions (e.g., histograms, scatter plots) to gain insights into customer behaviour.

2.4 RFM Analysis

- Recency: Time since the last purchase.
- Frequency: Number of unique transactions.
- Monetary: Total spending.
- Used quartiles to assign RFM scores, creating a combined `RFM_Score`.

3. Clustering and Segment Profiling

3.1 K-Means Clustering

- Applied K-Means clustering to RFM data to segment customers into meaningful groups.
- Elbow method used to determine the optimal number of clusters.

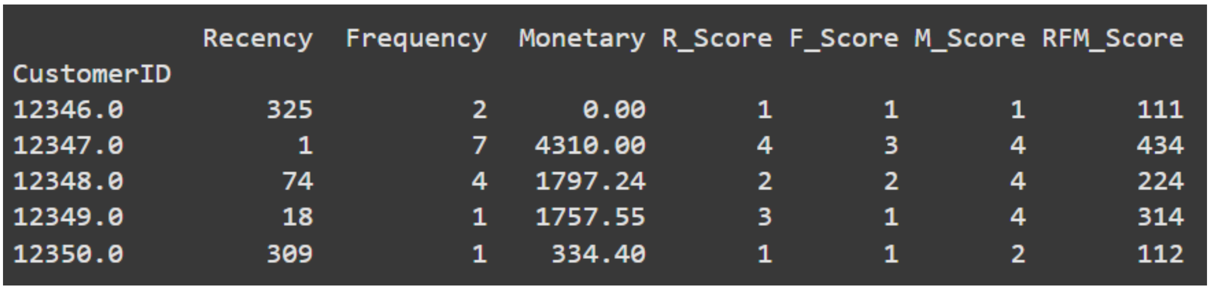
3.2 Cluster Profiles

Clusters were identified with the following key traits:

Cluster	Recency	Frequency	Monetary	Description
0	Low	High	High	Champions: High-value and frequent customers.
1	High	Low	Low	Hibernating: Rarely active customers.
2	Moderate	High	High	Loyal customers with consistent purchases.
3	High	Moderate	Low	At-risk customers with reduced activity.

4. Visualizations and Interpretations

4.1 RFM Distribution



```
            Recency  Frequency  Monetary  R_Score  F_Score  M_Score  RFM_Score
CustomerID
12346.0         325          2      0.00        1        1        1        111
12347.0           1          7   4310.00        4        3        4        434
12348.0          74          4   1797.24        2        2        4        224
12349.0          18          1   1757.55        3        1        4        314
12350.0         309          1    334.40        1        1        2        112
```

Interpretation:

This table represents a sample of customers' **Recency, Frequency, and Monetary (RFM)** metrics, along with their respective RFM scores. Here's a breakdown of the columns and their significance:

1. **Recency**: The number of days since the customer's last purchase.
   - Lower values indicate more recent activity (e.g., Customer 12347.0 has a Recency of 1, showing very recent engagement).
   - Higher values suggest inactivity (e.g., Customer 12346.0 has a Recency of 325, indicating a long gap since the last purchase).
2. **Frequency**: The total number of unique purchases made by the customer.
   - Higher values indicate frequent purchasing behavior (e.g., Customer 12347.0 has a Frequency of 7).
   - Lower values suggest infrequent engagement (e.g., Customers 12349.0 and 12350.0 have a Frequency of 1).
3. **Monetary**: The total amount spent by the customer.
   - Higher values indicate significant spending (e.g., Customer 12347.0 has spent $4310.00).
   - Lower values (or $0.00, as seen for Customer 12346.0) suggest minimal or no monetary contribution.
4. **R_Score, F_Score, M_Score**: Scores assigned to each RFM metric based on quartiles (1 = low, 4 = high).
   - **R_Score**: Reflects recency, with higher scores indicating more recent activity.
   - **F_Score**: Reflects frequency, with higher scores indicating frequent purchases.

   - **M_Score**: Reflects monetary value, with higher scores indicating greater spending.
5. **RFM_Score**: A combined score (R_Score, F_Score, M_Score) to classify customer segments.
   - Example: Customer 12347.0 has an RFM_Score of 434 (high recency, high frequency, and high monetary value), categorizing them as a top-tier customer (e.g., "Champion").
   - Customer 12346.0, with an RFM_Score of 111 (low in all metrics), would fall into a less engaged segment, such as "Hibernating."

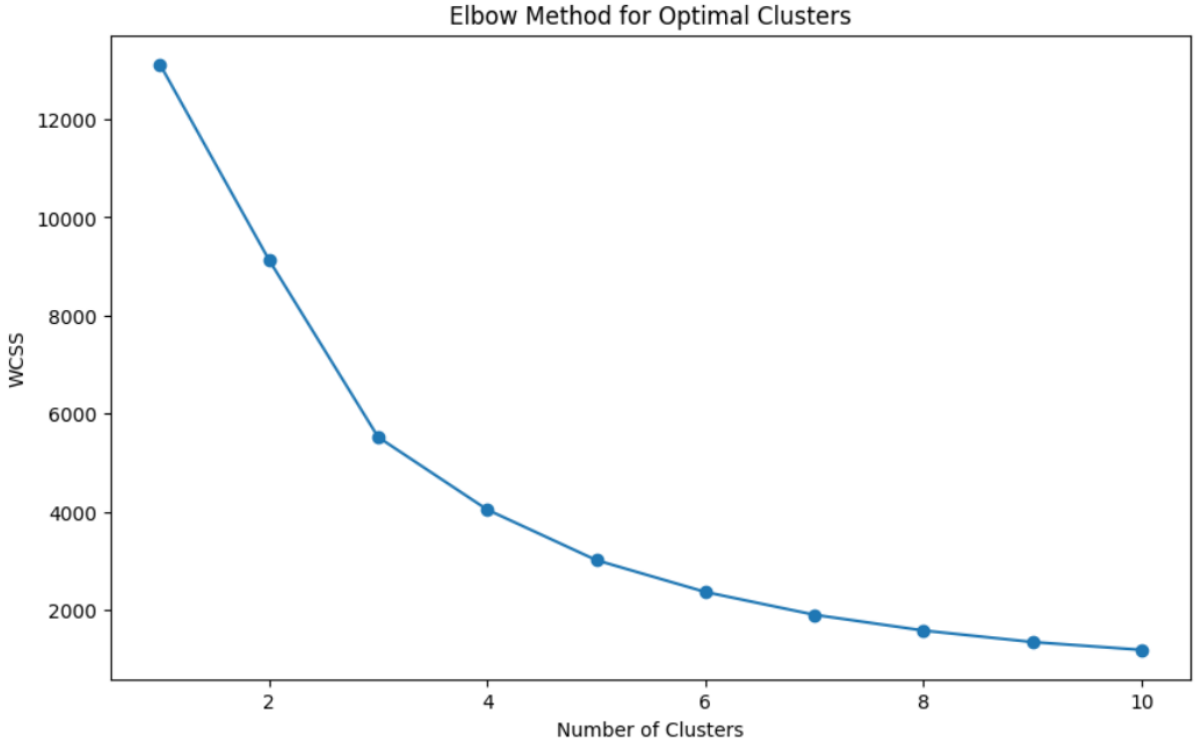
Key Insights:
Customer 12347.0, a high-value customer with an RFM_Score of 434, has very recent activity, frequent purchases, and high spending, making them a "Champion" segment ideal for loyalty programs and personalized offers. In contrast, Customer 12346.0, with an RFM_Score of 111, is inactive with a long gap since their last purchase, low engagement, and no significant spending, categorizing them as "Hibernating" and in need of retargeting campaigns. Customer 12348.0, with a moderate RFM_Score of 224, shows somewhat recent activity, moderate frequency, and significant spending, fitting the "Potential Loyalist" profile suitable for upselling or cross-selling strategies. Customer 12349.0, with an RFM_Score of 314, recently made a purchase but exhibits infrequent activity, placing them in the "At-Risk" segment that requires re-engagement campaigns. Lastly, Customer 12350.0, a low-engagement customer with an RFM_Score of 112, has infrequent activity and minimal spending, aligning with the "Hibernating" segment and needing strategies like promoting trending products to reignite interest.

Summary:

- **Top Priority** is for the customer 12347.0 (Champions) for retention and upselling. The **Focus Area** is by re-engaging customers like 12346.0 and 12350.0 (Hibernating) and retaining 12349.0 (At-Risk).

4.2 Elbow Method for Optimal Clusters

- The Line chart showing WCSS (Within-Cluster Sum of Squares) for different cluster counts.

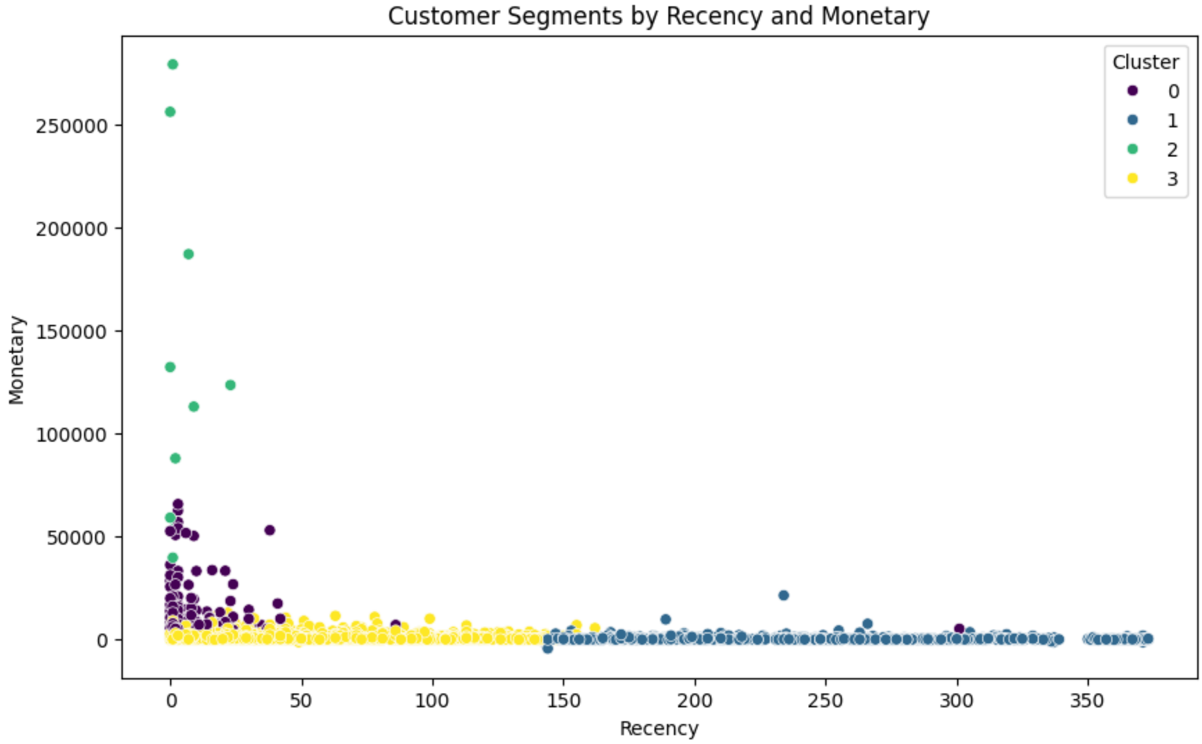


Interpretation:

The elbow point (4 clusters) suggests the best balance between simplicity and interpretability, ensuring meaningful segmentation.

4.3 Cluster Scatter Plots

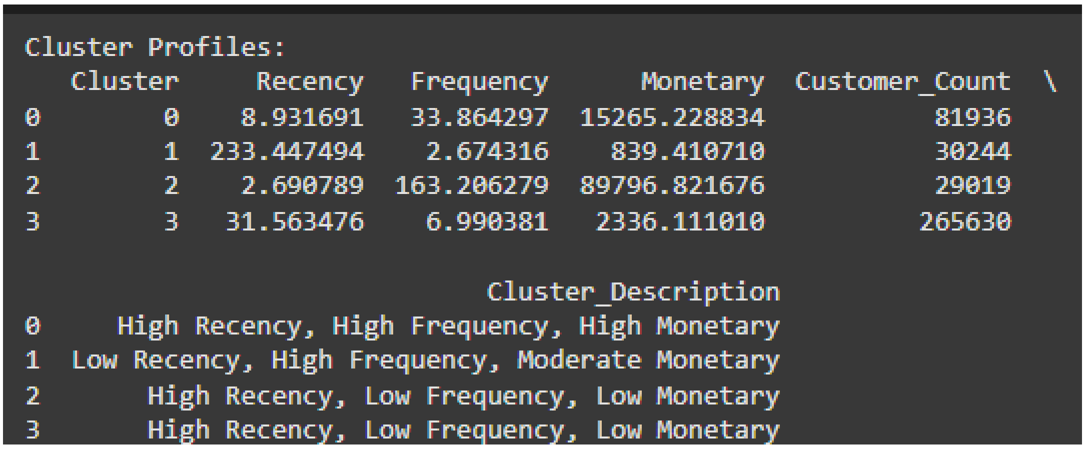
1. Interpretation of the Customer Segments by Recency and Monetary:



1. **Cluster 2 (Green Points - Champions):**
   - These customers exhibit **low recency (recent purchases)** and **high monetary value**, indicating they are frequent and high-spending customers. This segment represents the most valuable group for retention and targeted loyalty strategies.
2. **Cluster 0 (Purple Points - Potential Loyalists):**
   - These customers have **moderate recency** and **moderate-to-high monetary value**. They are promising customers who may be nurtured into champions through personalized marketing and offers.
3. **Cluster 1 (Blue Points - Hibernating):**
   - Customers in this cluster show **high recency (long gap since last purchase)** and **low monetary value**, indicating they are inactive and have contributed minimally to revenue. Re-engagement campaigns are critical for this segment.
4. **Cluster 3 (Yellow Points - At-Risk):**
   - This group has **moderate recency** and **low monetary value**, suggesting they might have been moderately active previously but are now less engaged. Targeted campaigns can help revive their activity and spending.

This visualization highlights the relationship between how recently customers have purchased (recency) and their overall contribution to revenue (monetary), enabling clear identification of priority segments.

2. Cluster Profile:



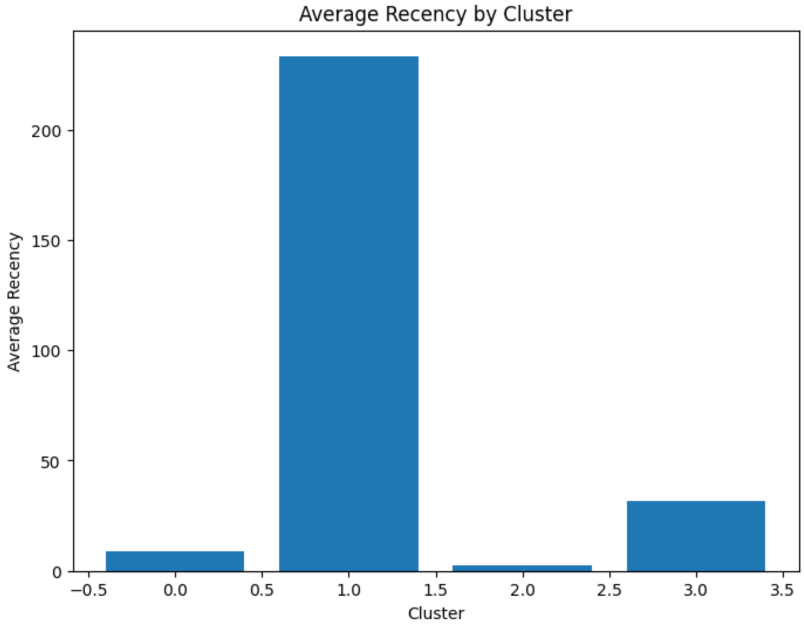
```
Cluster Profiles:
   Cluster     Recency   Frequency      Monetary  Customer_Count  \
0        0    8.931691   33.864297  15265.228834           81936
1        1  233.447494    2.674316    839.410710           30244
2        2    2.690789  163.206279  89796.821676           29019
3        3   31.563476    6.990381   2336.111010          265630

                              Cluster_Description
0     High Recency, High Frequency, High Monetary
1  Low Recency, High Frequency, Moderate Monetary
2       High Recency, Low Frequency, Low Monetary
3       High Recency, Low Frequency, Low Monetary
```

This table outlines customer profiles based on their purchasing behavior:

1. **Cluster 0**: High-value, active customers with **low recency**, **high frequency**, and **high monetary value** (ideal for loyalty programs).
2. **Cluster 1**: Moderately active customers with **low recency**, **high frequency**, and **moderate spending**, suitable for upselling and referrals.
3. **Cluster 2**: Inactive but high-spending customers with **high recency**, **low frequency**, and **high monetary value**, needing re-engagement.
4. **Cluster 3**: Majority of customers have **high recency**, **low frequency**, and **low monetary value**, requiring retargeting campaigns.

The segmentation helps prioritize marketing strategies to maximize engagement and revenue.

3. Interpretation of the Bar Chart (Average Recency by Cluster):

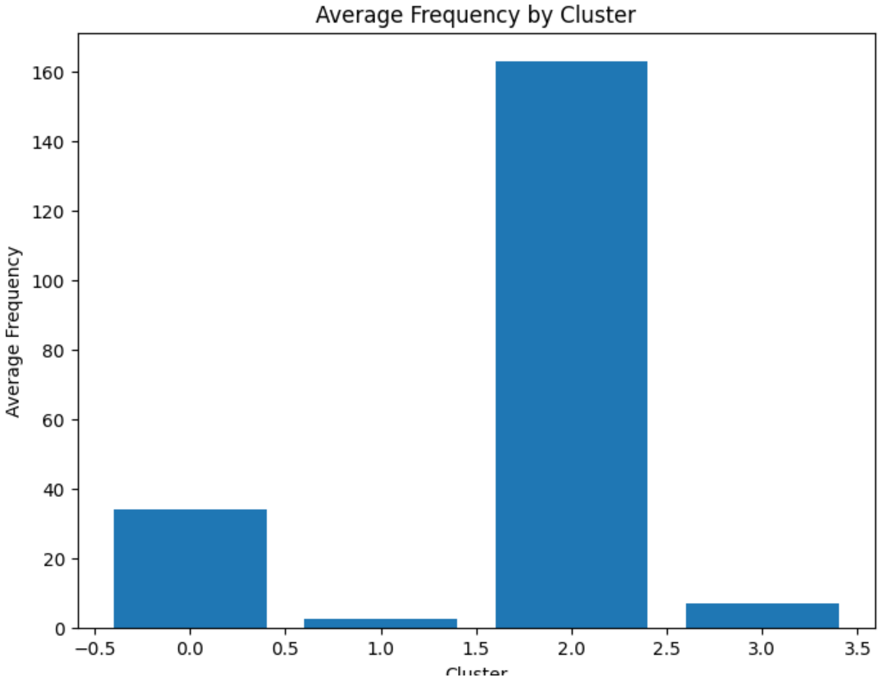


1. **Cluster 1**: This cluster has the **highest average recency** (over 200), meaning these customers have not made a purchase in a long time. They are likely inactive or "Hibernating" customers, requiring strong re-engagement efforts.

2. **Cluster 0**: This group has a **very low average recency**, indicating they have made purchases very recently. These are "Champion" customers who are highly engaged and should be nurtured with loyalty programs.
3. **Cluster 3**: Moderate recency suggests these customers made purchases somewhat recently but are not as engaged as Cluster 0. They may fall into the "At-Risk" category and need targeted retention strategies.
4. **Cluster 2**: This group also shows **very low average recency**, similar to Cluster 0, suggesting they have recently interacted but may be categorized differently based on other metrics like frequency or monetary value.

The chart highlights the recency behavior across clusters, helping identify which segments require immediate attention to boost engagement.

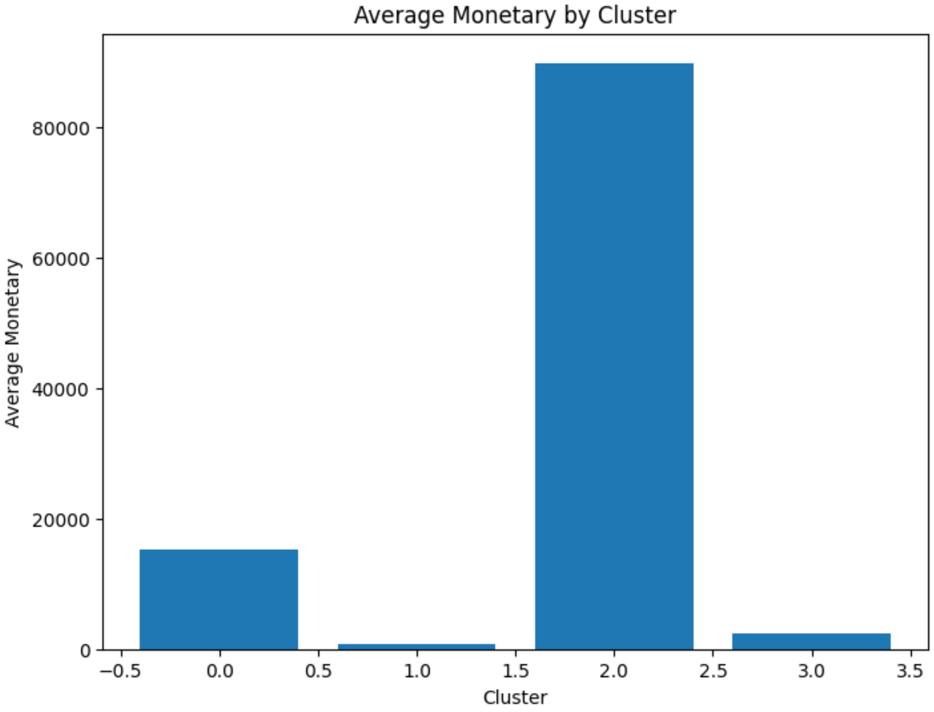
4.Interpretation of the Bar Chart (Average Frequency by Cluster):



1. **Cluster 2**: This cluster has the **highest average frequency** (above 160), indicating that these customers make purchases very frequently. They are likely "Champions" or "Loyal Customers," representing a critical group for maintaining strong engagement.
2. **Cluster 0**: This group shows **moderate frequency** (around 30-40), suggesting they are regular buyers but less frequent than Cluster 2. These could be "Potential Loyalists" who might be encouraged to buy more frequently with targeted offers.
3. **Cluster 3**: This cluster has a **low frequency**, indicating infrequent purchases. These customers may be in the "At-Risk" category and require retention strategies to increase their engagement.
4. **Cluster 1**: With the **lowest frequency**, this group shows minimal purchasing activity, identifying them as "Hibernating" customers. They need reactivation campaigns to bring them back into the buying cycle.

The chart emphasizes the varying purchase frequencies across clusters, highlighting the most engaged groups (Cluster 2) and those needing intervention (Clusters 1 and 3).

5. Interpretation of the Bar Chart (Average Monetary by Cluster):

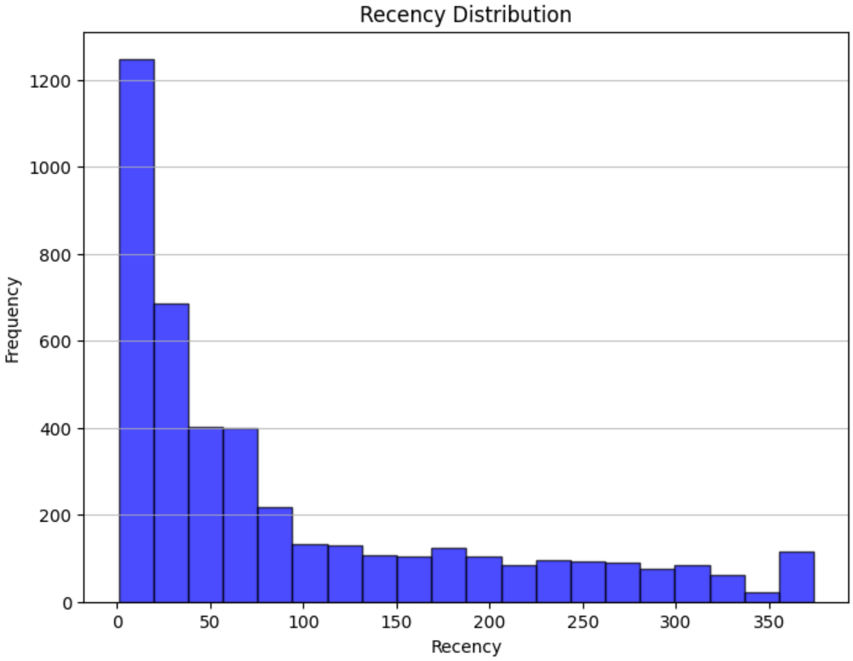


1. **Cluster 2**: This cluster shows the **highest average monetary value** (above 90,000), indicating that these customers contribute significantly to revenue. They are the most valuable segment, likely "Champions," and should be prioritized for retention through exclusive offers and loyalty programs.
2. **Cluster 0**: This group has a **moderate monetary value** (around 15,000), reflecting steady spending. These customers might be "Loyal Customers" or "Potential Loyalists," and targeted promotions could increase their spending.
3. **Cluster 3**: Customers in this cluster show **low monetary value** (around 2,000), suggesting infrequent spending. They could be "At-Risk" or "Hibernating" customers requiring re-engagement campaigns to boost their activity.
4. **Cluster 1**: With the **lowest monetary value**, this group contributes minimally to revenue. They fall into the "Hibernating" category and need strategies like discounts or trending product promotions to encourage purchases.

This chart highlights the monetary contribution across clusters, emphasizing the need to nurture high-value segments (Cluster 2) while re-engaging low-value segments (Clusters 1 and 3).

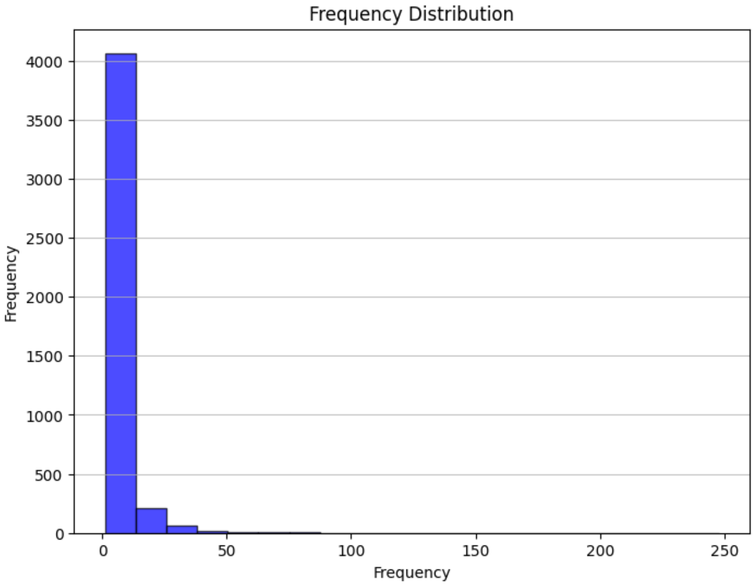
Interpretation of Distributions:

1. Recency Distribution:



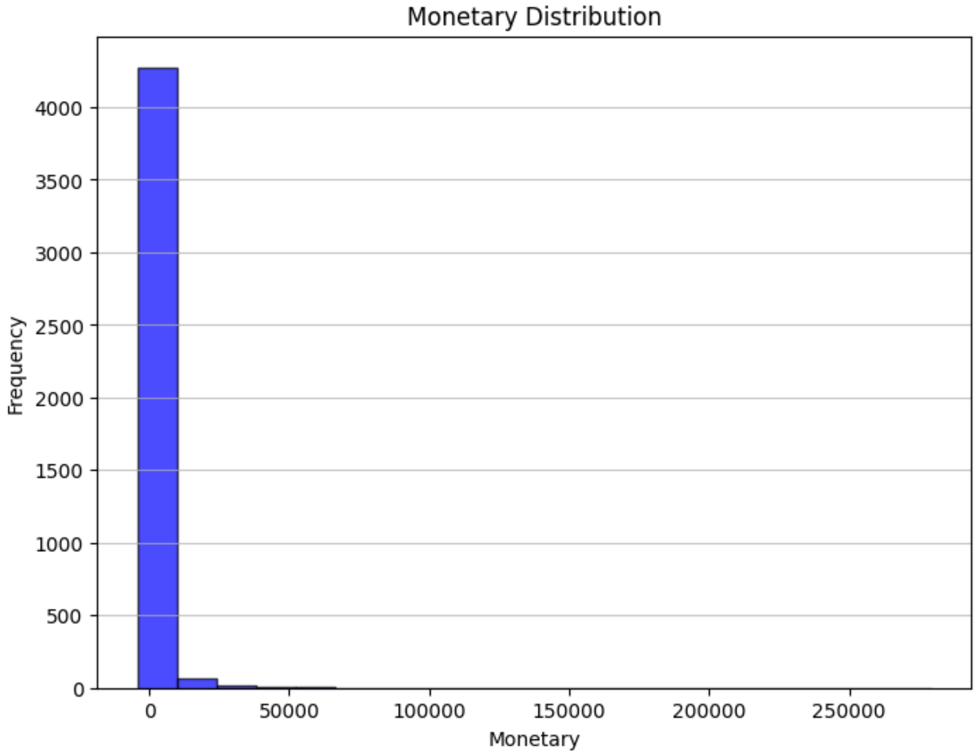
- Most customers have very low recency, meaning they have made purchases recently.
- The frequency decreases significantly as recency increases, indicating fewer customers have been inactive for a long time.
- This highlights the need to engage customers with higher recency values (longer inactivity).

2. Frequency Distribution:



- The majority of customers have very low frequency, indicating they have made only a few purchases.
- A small subset of customers shows high frequency, representing the loyal or repeat buyers.
- Focus can be given to converting low-frequency customers into regular buyers through promotional strategies.

3. Monetary Distribution:



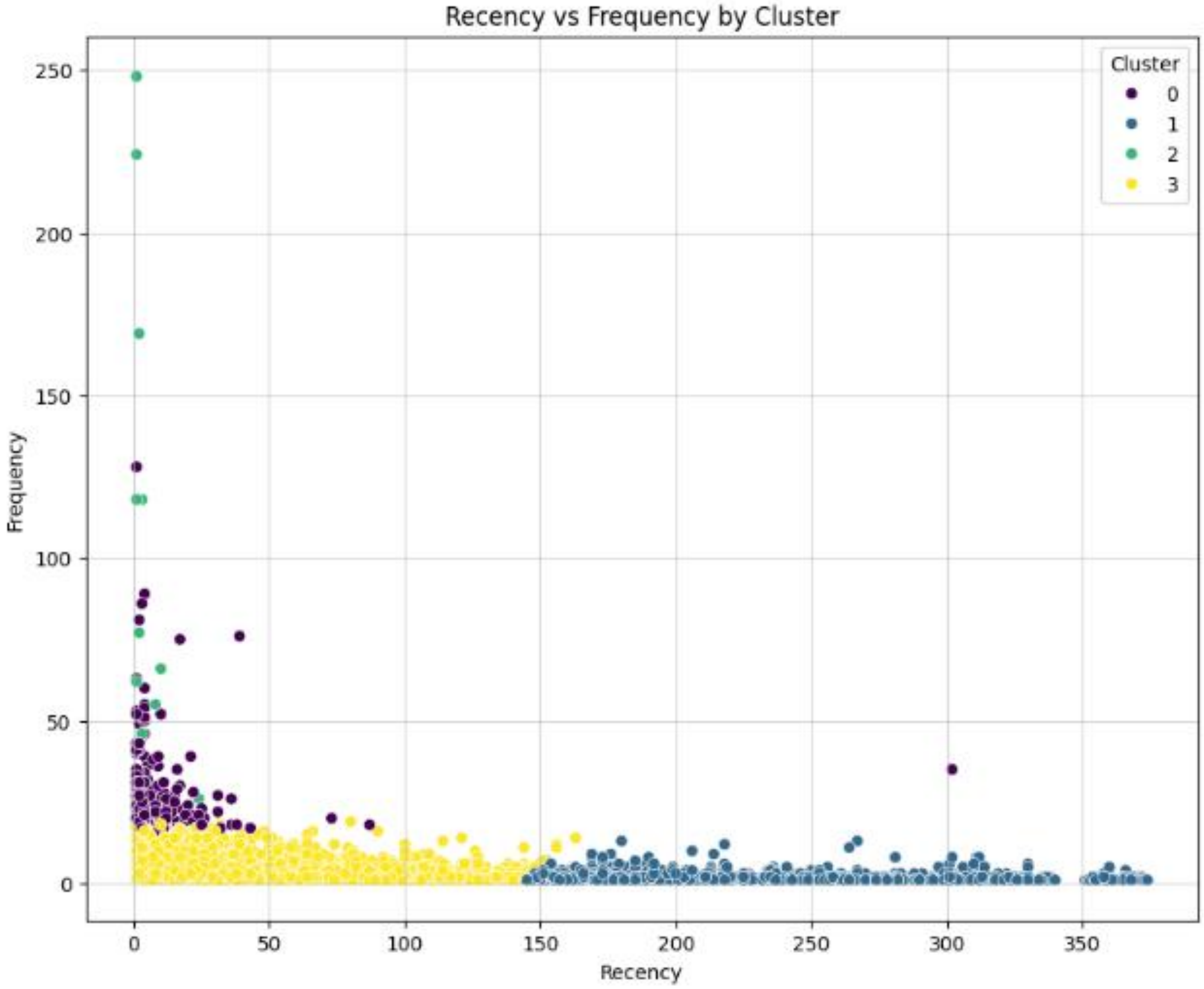
- Most customers contribute a small monetary value, with only a few contributing significantly to revenue.
- A small number of customers (outliers) account for very high monetary values, making them critical for retention and loyalty programs.
- This skewed distribution suggests that a small percentage of customers drives the majority of revenue.

Key Insights:

- Engage and retain high-value and frequent customers while devising strategies to boost engagement for less active segments.
- Focus on outliers in the frequency and monetary distributions to understand and replicate their behavior in other customer segments.

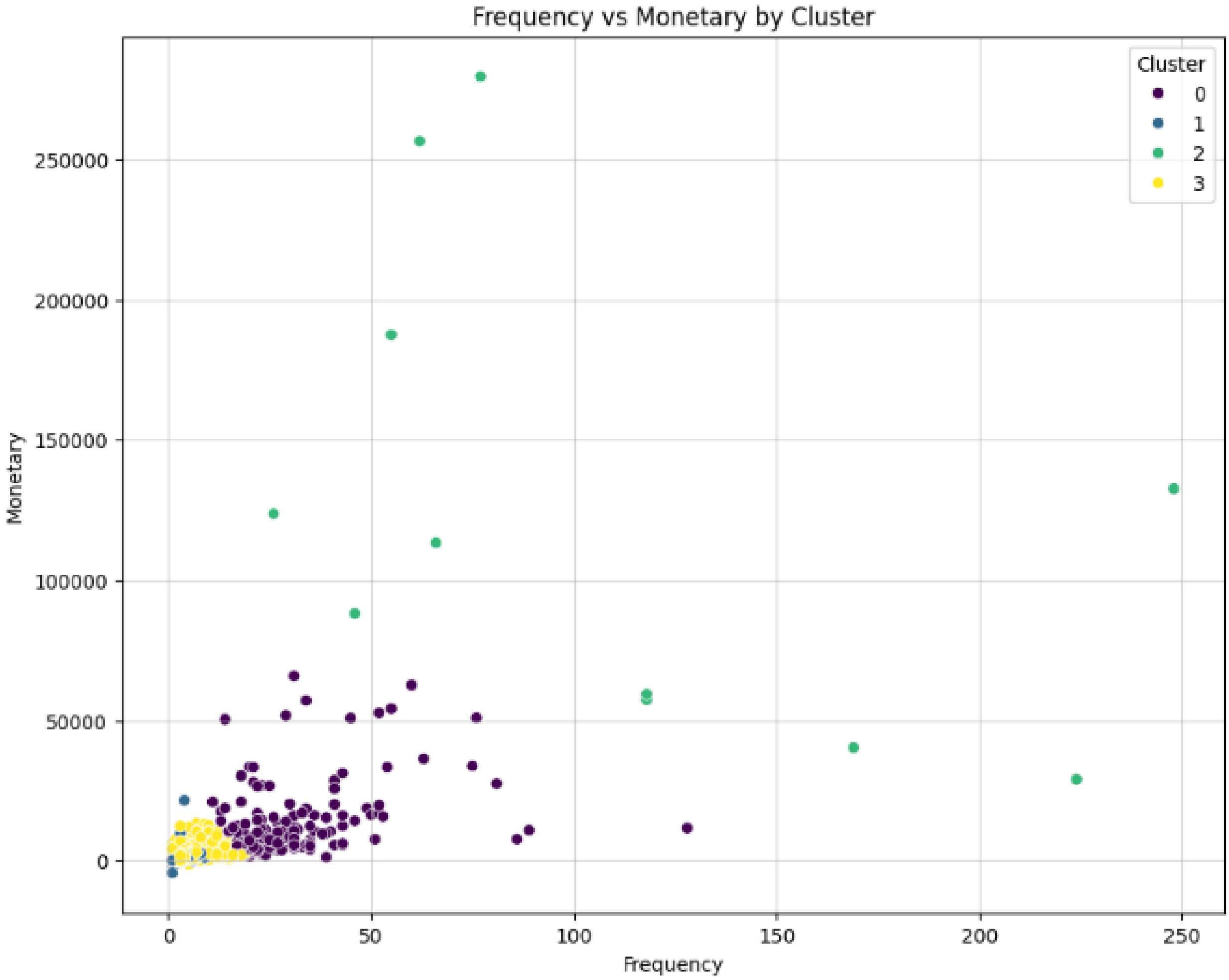
Interpretation of the Scatter Plots:

1. Recency vs Frequency by Cluster:



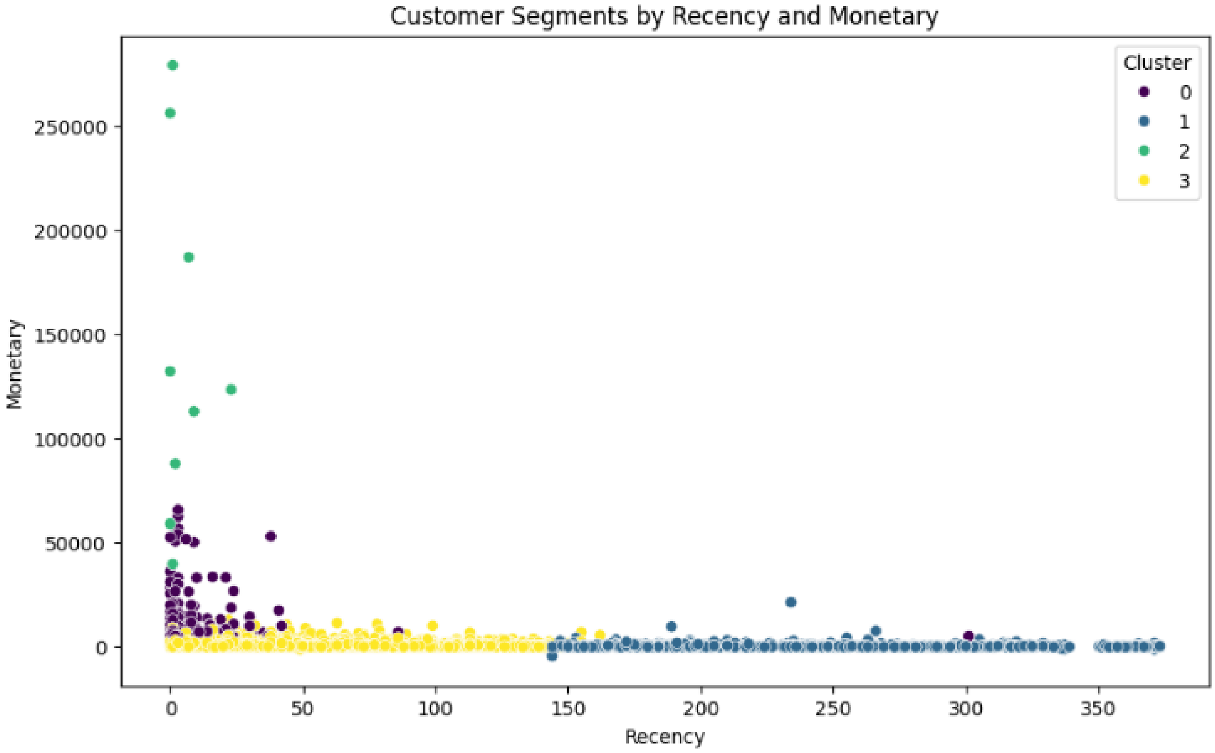
- **Cluster 2 (Green):** Customers in this cluster exhibit very low recency and very high frequency, indicating they are highly engaged and frequent buyers ("Champions").
- **Cluster 0 (Purple):** Moderate frequency and low recency, representing steady buyers who are moderately engaged ("Loyal Customers" or "Potential Loyalists").
- **Cluster 1 (Blue):** High recency and very low frequency, indicating inactivity and infrequent purchases ("Hibernating Customers").
- **Cluster 3 (Yellow):** Moderate recency and low frequency, possibly representing customers at risk of churn ("At-Risk Customers").

2. Frequency vs Monetary by Cluster:



- **Cluster 2 (Green):** Customers here show very high frequency and monetary value, contributing the most to revenue. They are prime for retention strategies.
- **Cluster 0 (Purple):** Moderate frequency and monetary value, representing regular spenders who could be incentivized for upselling.
- **Cluster 3 (Yellow) and Cluster 1 (Blue):** Both show low frequency and low monetary contributions, indicating disengaged customers who need re-engagement strategies.

3. Recency vs Monetary by Cluster:

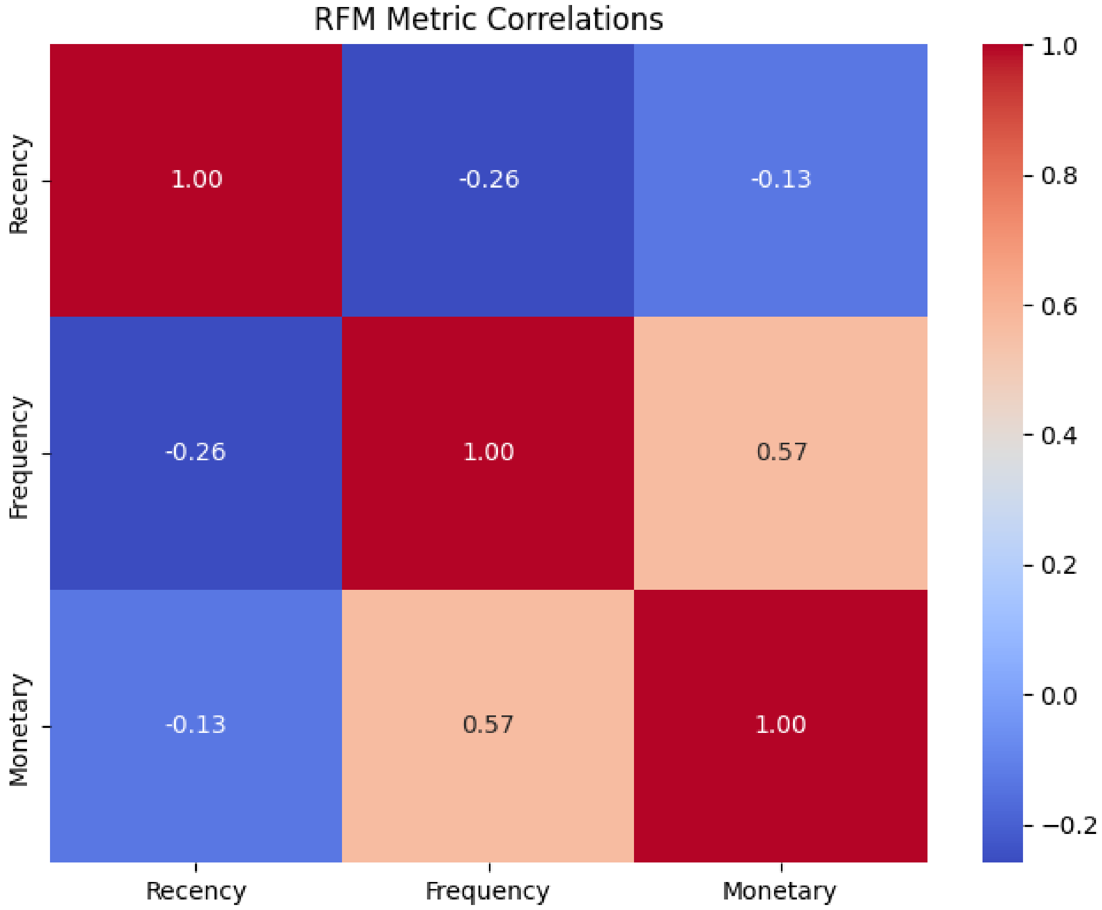


- **Cluster 2 (Green):** Very low recency and very high monetary value highlight these customers as recent and high spenders ("Champions").
- **Cluster 0 (Purple):** Moderate recency and monetary value suggest consistent spending with a potential for growth ("Loyal Customers").
- **Cluster 1 (Blue):** High recency and low monetary contributions, indicating inactive customers requiring reactivation efforts.
- **Cluster 3 (Yellow):** Moderate recency and low monetary contributions, indicating customers at risk of complete inactivity.

Key Insights:

- **Priority Segments:** Focus on Clusters 2 and 0 (Green and Purple) for retention and loyalty-building.
- **Re-engagement Focus:** Clusters 1 and 3 (Blue and Yellow) need targeted campaigns to revive their engagement and spending.

Interpretation of the RFM Metric Correlation Heatmap:



1. **Recency vs. Frequency (-0.26)**:

○ A slight negative correlation indicates that customers who purchase more frequently tend to make more recent purchases. This is expected, as frequent buyers are more likely to stay engaged.

2. **Recency vs. Monetary (-0.13)**:

○ A weak negative correlation suggests that customers with recent purchases tend to spend slightly more, but the relationship is not strong.

3. **Frequency vs. Monetary (0.57)**:

○ A moderate positive correlation shows that customers who purchase more frequently tend to spend more overall. This is a key insight for identifying high-value customers.

4. **Key Insight**:

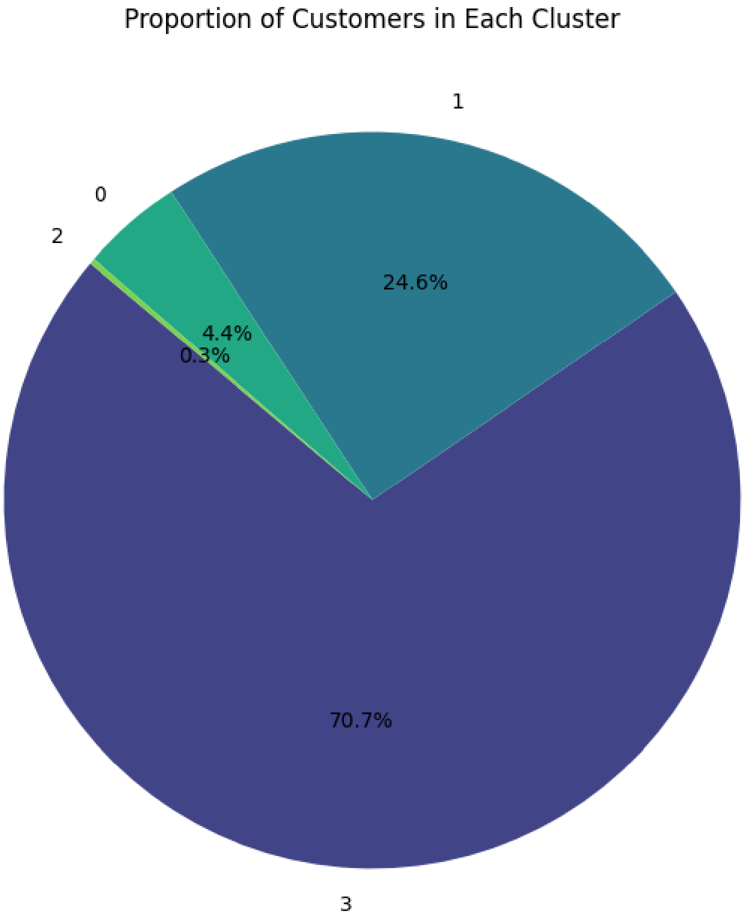
○ **Frequency and Monetary** are the most strongly correlated metrics, indicating that focusing on increasing purchase frequency could lead to higher monetary contributions.

○ **Recency** has weaker correlations, but it is still important to monitor for identifying customer engagement trends.

This correlation analysis highlights how the RFM metrics interrelate, aiding in the identification of strategies for retention and revenue maximization.

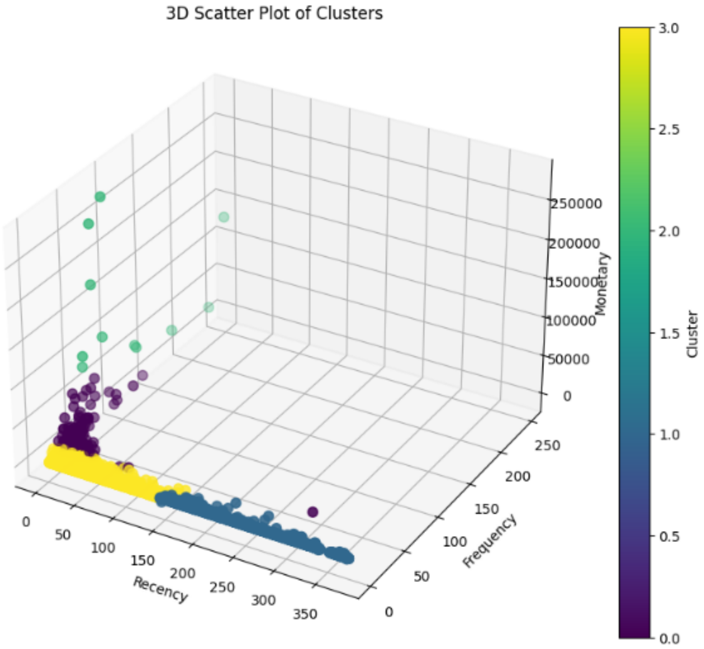
Interpretation of the Visualizations:

1. **Proportion of Customers in Each Cluster (Pie Chart):**



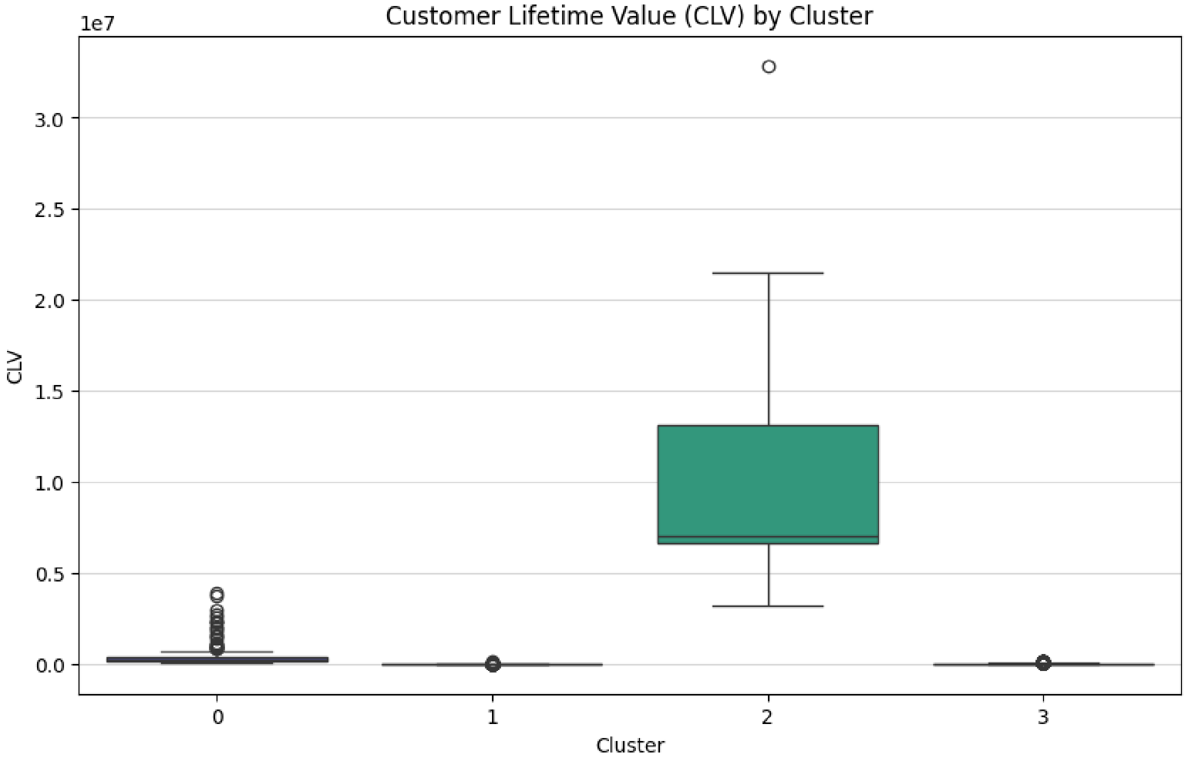
• **Cluster 3 (Yellow):** Represents 70.7% of the customers, the largest group. These are likely inactive or "Hibernating" customers requiring re-engagement strategies to reintroduce them to purchasing.
• **Cluster 1 (Blue):** Covers 24.6% of the customers, moderately engaged but contributing less to revenue. These customers could be "At-Risk" and need targeted retention strategies.
• **Cluster 0 (Purple):** Makes up 4.4%, representing steady buyers ("Loyal Customers" or "Potential Loyalists").
• **Cluster 2 (Green):** Comprising only 0.3%, this is the smallest but most valuable segment ("Champions"), contributing significantly to revenue and requiring personalized attention.

2. 3D Scatter Plot of Clusters:



- **Cluster 2 (Green):** Customers in this cluster are concentrated at low recency and high frequency/monetary values, indicating frequent, high-value transactions.
- **Cluster 0 (Purple):** Displays moderate frequency and monetary contributions, reflecting consistent but less frequent buyers.
- **Cluster 1 (Blue) and Cluster 3 (Yellow):** Spread across high recency with low frequency and monetary contributions, showing inactivity and disengagement.

3. Customer Lifetime Value (CLV) by Cluster (Boxplot):



- **Cluster 2 (Green):** Has the highest CLV, with a wide range and significant outliers, demonstrating that these customers drive the majority of the revenue. Retention is critical.
- **Cluster 0 (Purple):** Displays moderate CLV, suggesting steady spenders who could be incentivized for growth.
- **Cluster 1 (Blue) and Cluster 3 (Yellow):** Show very low CLV, highlighting the need for strategies to boost their value.

Key Insights:

1. **Cluster 2 (Champions):** Focus on retaining these high-value customers with loyalty programs and personalized offers.
2. **Cluster 0 (Loyal Customers):** Encourage upselling or cross-selling to increase their CLV.
3. **Cluster 1 and 3 (At-Risk and Hibernating):** Implement re-engagement campaigns and promotions to revitalize activity in these large but low-value groups.

MARKETING RECOMMENDATIONS

Define General Recommendations for Common RFM Segments:

Typical RFM segmentation divides customers into groups based on their purchasing behavior. The following are common segments and tailored strategies for them:

1. **Champions** (High Recency, High Frequency, High Monetary): Characteristics: Loyal and high-spending customers.
   Recommendations: Offer exclusive loyalty programs or VIP memberships. Provide early access to new products or premium services. Send personalized thank-you notes or gifts.
2. **Loyal Customers** (Moderate Recency, High Frequency, Moderate Monetary): Characteristics: Regular customers with consistent spending.
   Recommendations: Use targeted upselling or cross-selling strategies. Introduce referral incentives to attract similar customers. Offer discounts on bulk purchases or subscription plans.
3. **Potential Loyalists** (High Recency, Moderate Frequency, Moderate Monetary) Characteristics: Recently acquired customers with promising behavior.
   Recommendations: Send welcome emails and introductory offers. Provide onboarding materials or tutorials for product use. Offer limited-time discounts to encourage repeat purchases.
4. **At-Risk Customers**(Low Recency, High Frequency, High Monetary): Characteristics: Previously loyal customers showing signs of disengagement.
   Recommendations: Send re-engagement campaigns with exclusive offers. Offer surveys to understand their concerns or challenges. Highlight your brand's unique value proposition through email or social media.
5. **Hibernating Customers**(Low Recency, Low Frequency, Low Monetary): Characteristics: Infrequent and low-spending customers.
   Recommendations: Use retargeting ads or emails to bring them back. Bundle products with discounts to increase their order size. Highlight trending or new products to spark interest.
   6.**Lost Customers** (Low Recency, Low Frequency, High Monetary): Characteristics: Previously high-value customers who are now inactive.

Recommendations: Send win-back campaigns with personalized messages. Offer exclusive, high-value promotions. Engage through surveys to understand their reasons for leaving.

5. Insights and Recommendations

5.1 Key Insights

1. Champions (Cluster 0) represent a significant opportunity for upselling and loyalty programs.
2. Loyal Customers (Cluster 2) show steady purchasing patterns, ideal for bulk purchase discounts.
3. At-Risk (Cluster 3) and Hibernating Customers (Cluster 1) need re-engagement campaigns to revive activity.

5.2 Recommendations

- Champions: Offer loyalty rewards and VIP access to new products.
- Loyal Customers: Implement targeted promotions and referral programs.
- Hibernating: Use retargeting ads and incentives to re-engage.
- At-Risk: Address concerns via surveys and exclusive offers.

6. Conclusion

This project successfully analysed customer purchasing behaviour through RFM segmentation and clustering. The actionable insights derived can guide marketing strategies, improve customer retention, and maximize revenue.