# 🧠 Customer Segmentation & Clustering – Business Insights Report

## Project Title:

Customer Segmentation and Clustering using RFM Analysis and Unsupervised Learning

## Objective:

To segment customers based on their purchasing behavior using RFM (Recency, Frequency, Monetary) metrics and apply clustering algorithms to derive actionable business insights.

## 1. 📊 Summary of Data and Preprocessing

- Raw Data Shape: (541,909 rows × 8 columns)

- Cleaned Data Shape: (397,884 rows)

- Issues Resolved:  
 - Missing CustomerID (~25% of rows)  
 - Negative/zero Quantity or UnitPrice  
 - Canceled transactions (InvoiceNo starts with 'C')

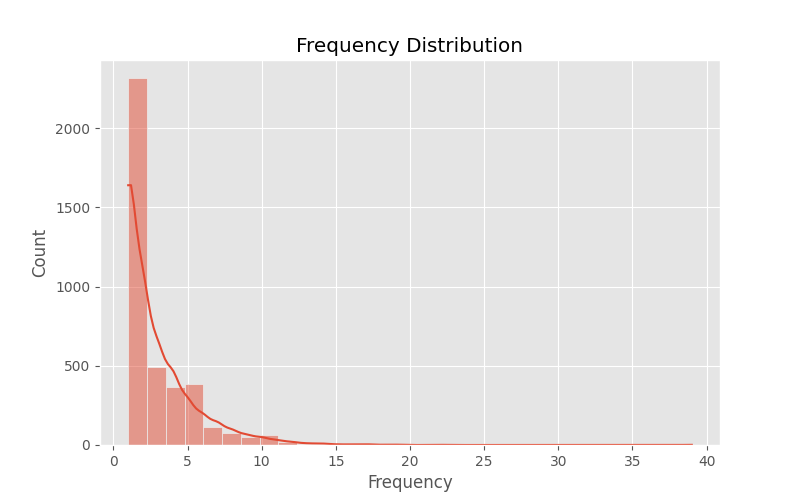
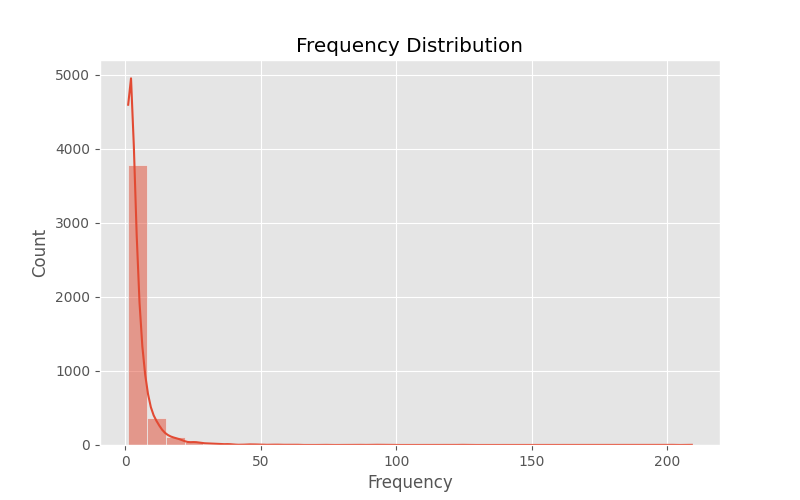
✅ Clean dataset ensured reliable analysis and insights.

## 2. 🔍 RFM Feature Engineering

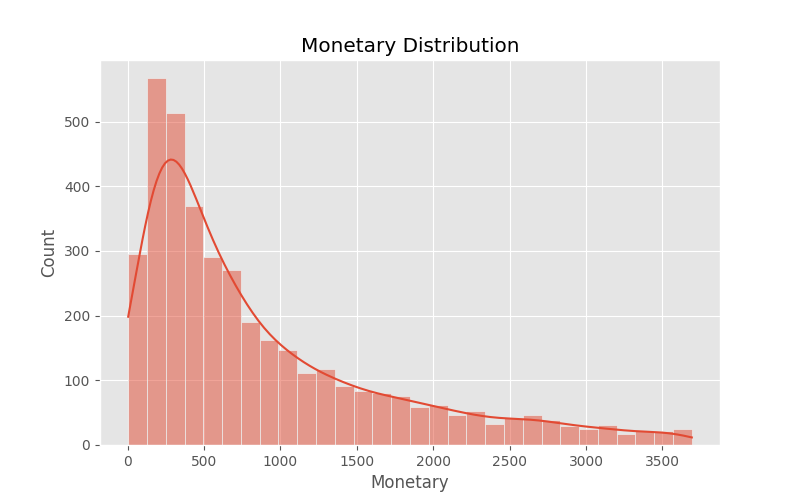
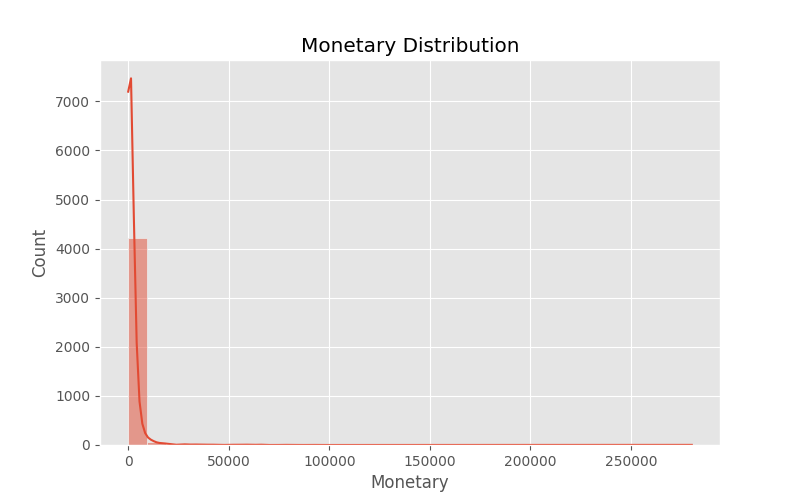
- Customers analyzed: 4,338 unique customers

- Features:  
 - Recency – Days since last purchase  
 - Frequency – Number of purchases  
 - Monetary – Total spend

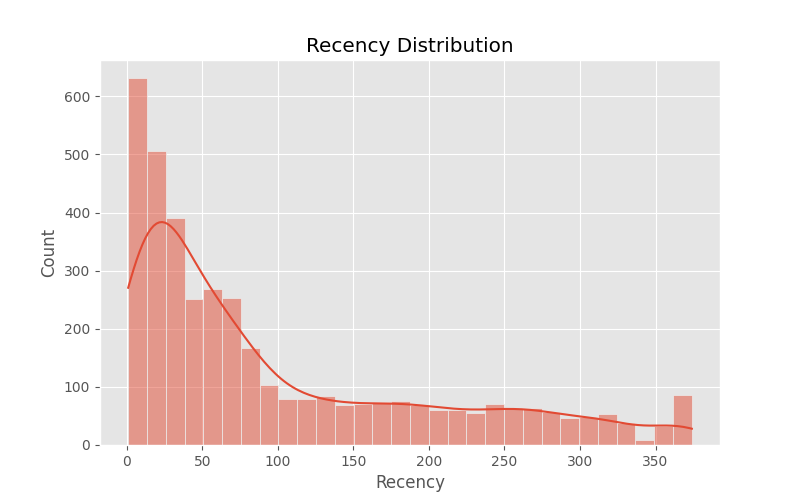
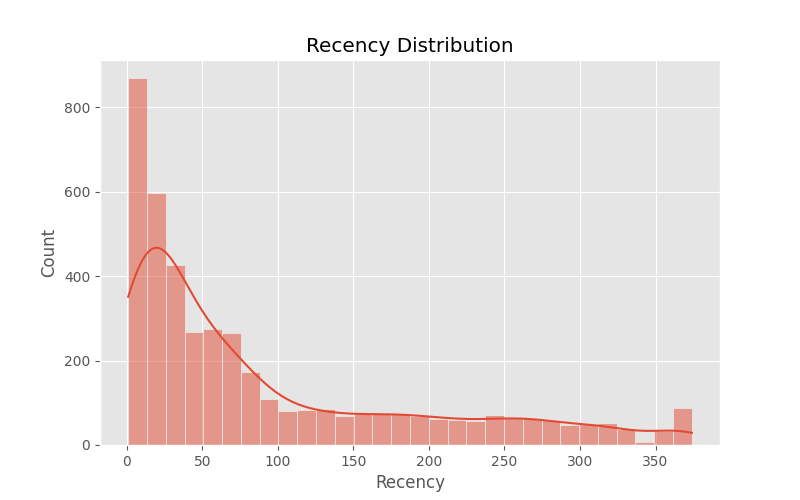
Filtered for outliers, resulting in 3,911 customers for clustering.  
  
Frequency distribution Before / After:



Monitory distribution Before / After:



Recency distribution Before / After:

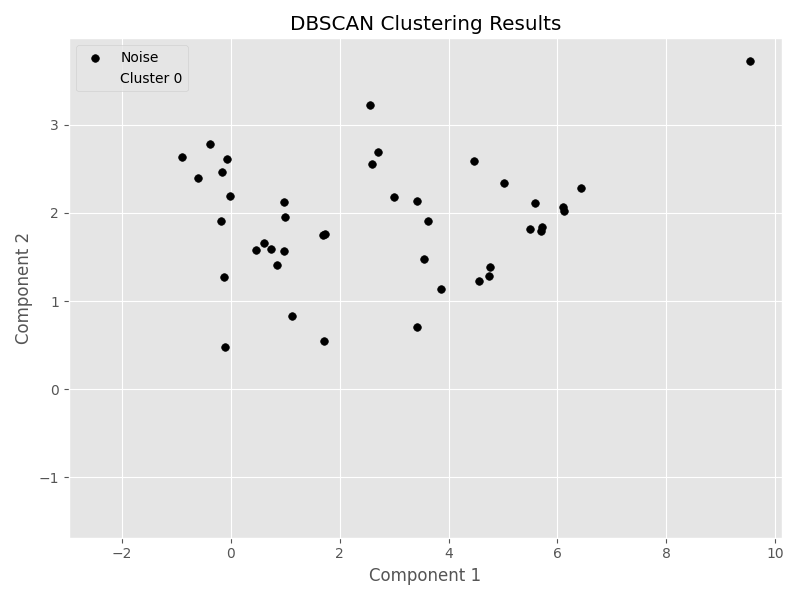


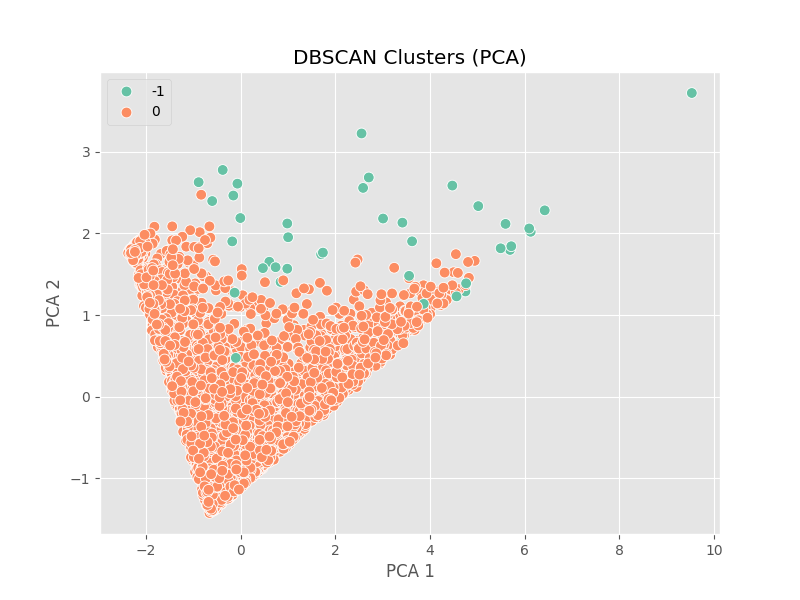
## 3. 🔗 Clustering Results & Interpretation

### A. DBSCAN Clustering

- Cluster 0 (Core): 3,869 customers  
- Noise (Outliers): 42 customers

DBSCAN plot:



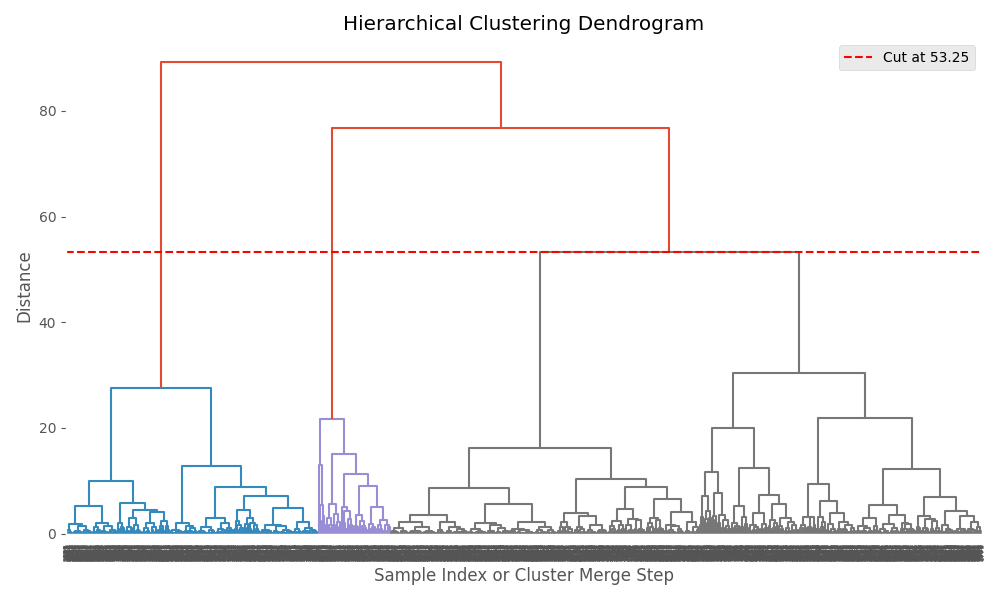
PCA plot :

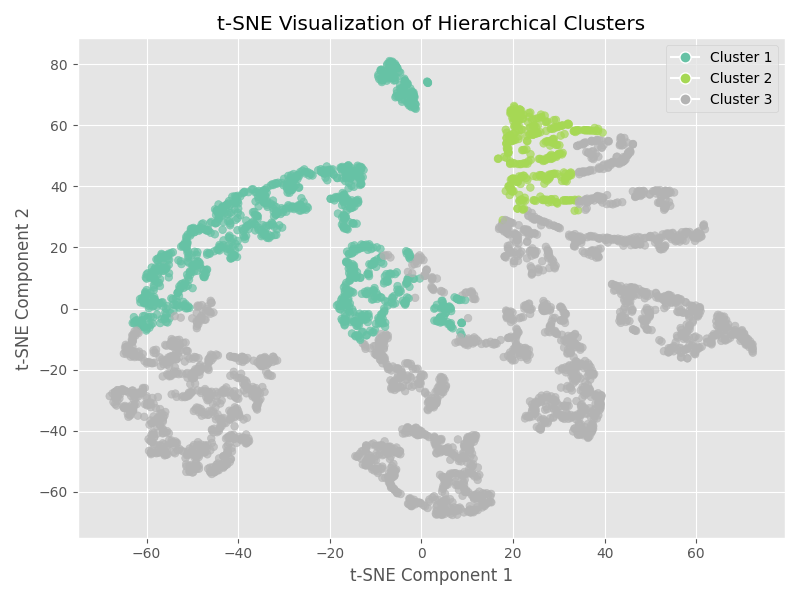
### B. Hierarchical Clustering

- Silhouette Score: 0.3975  
- DB Index: 0.7988

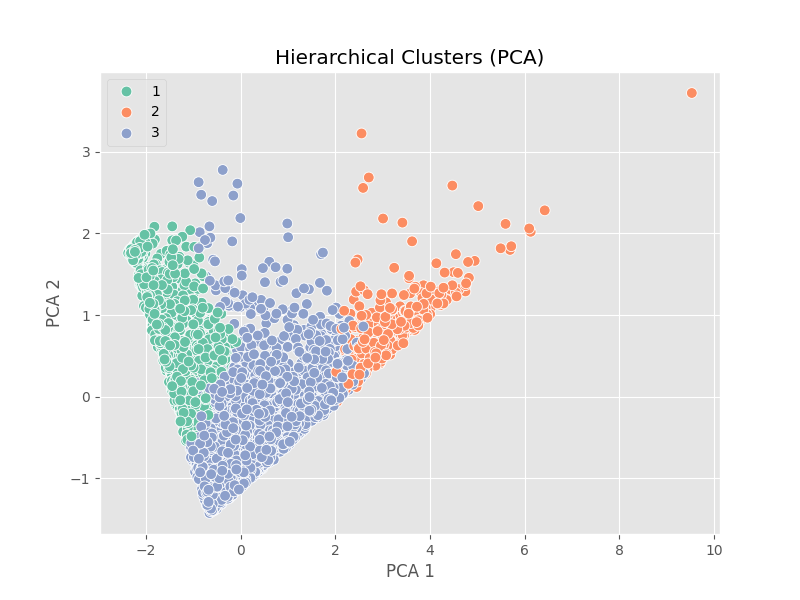
Formed 3 clusters:  
- Large base of moderate customers  
- A few premium or outlier groups

Dendrogram:



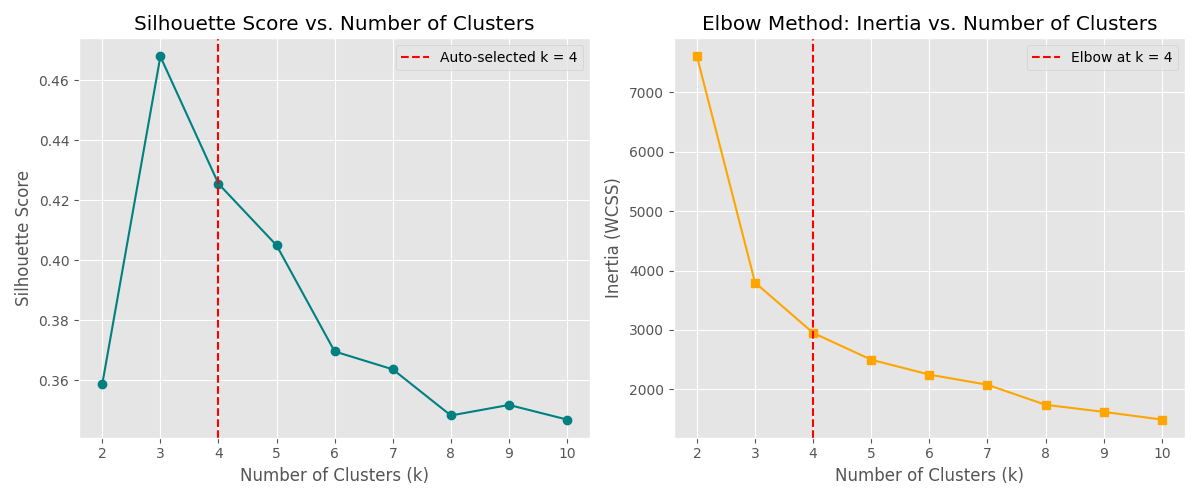
t-SNE plot:  


PCA plot:

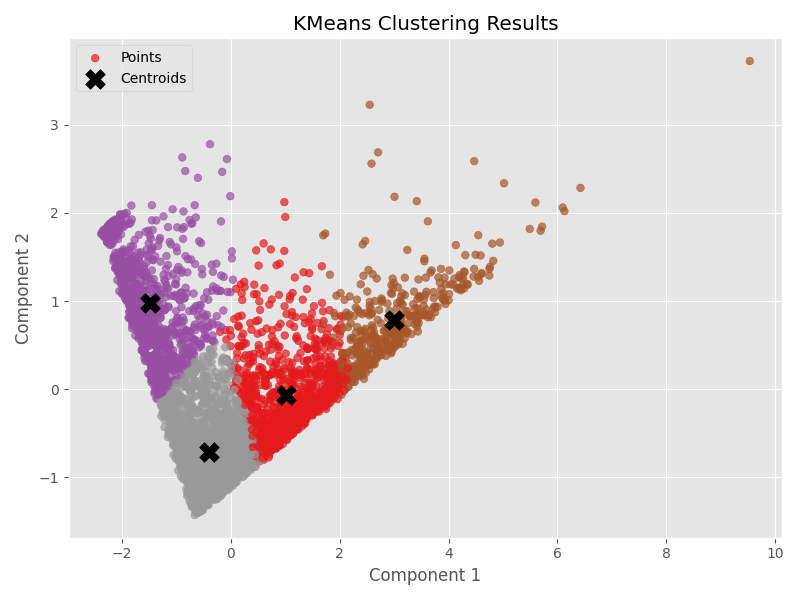


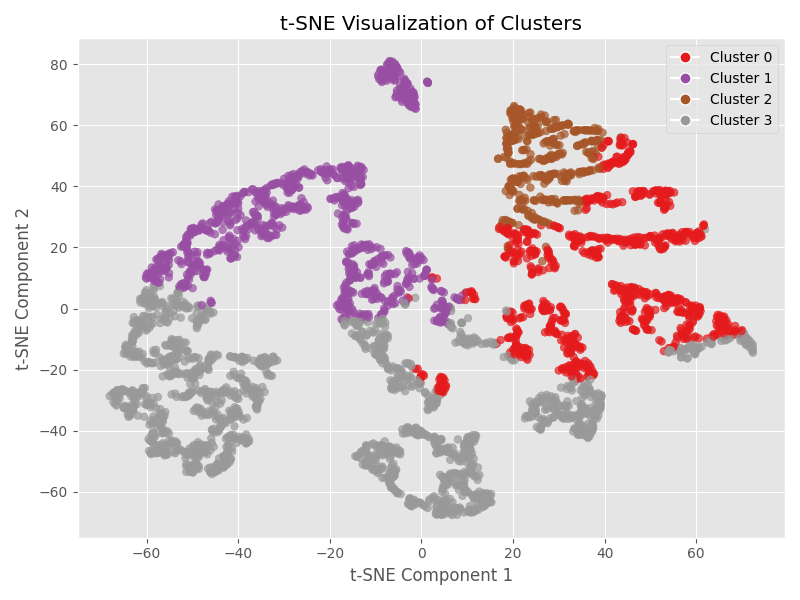
### C. KMeans Clustering (Best Performance)

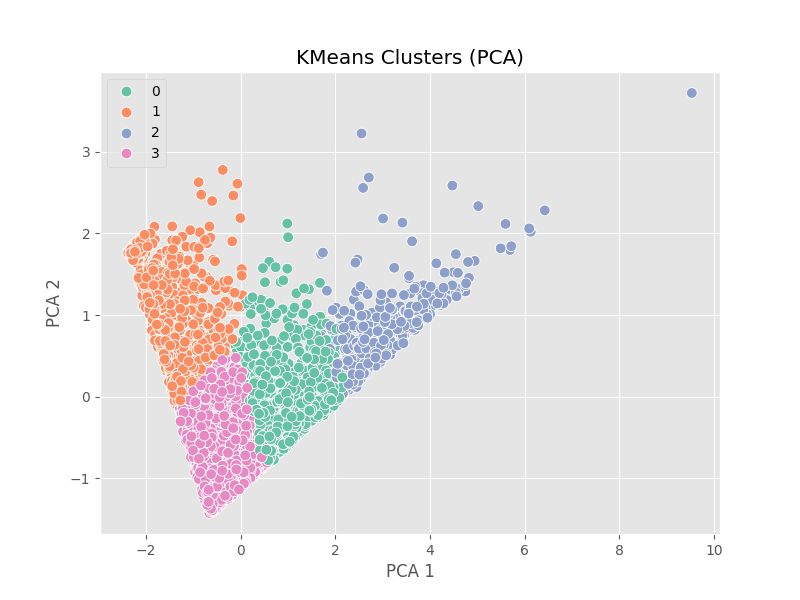
- Optimal Clusters: 4  
- Silhouette Score: 0.4255  
- Davies-Bouldin Index: 0.8735

Elbow plot:  


Kmeans Cluster plot:



t-SNE plot:  


PCA plot:  


Cluster Breakdown:  
Cluster 3: 1669 – High frequency & spend, recent buyers – Top Customers  
Cluster 1: 952 – Good spenders, moderately recent – Loyal Regulars  
Cluster 0: 904 – Inactive customers, low spend – At Risk  
Cluster 2: 386 – High spend, low frequency – Big Spenders

## 4. 💡 Business Insights

### Customer Behavior:

- Cluster 3 (Best Customers): High revenue, prioritize with loyalty programs  
- Cluster 0 (At Risk): Run reactivation campaigns  
- Cluster 2 (Big Spenders): Incentivize regular purchases  
- Cluster 1 (Loyal Regulars): Engage with personalized offers

### RFM Metric Learnings:

- Many customers haven’t purchased recently  
- Monetary values are highly skewed  
- Frequency shows a sharp drop after 3 purchases

## 5. 📦 Strategic Recommendations

- Retention Programs for Cluster 3 and 2  
- Targeted Email Campaigns for Cluster 0 and 1  
- VIP & Premium Segments for high-Monetary clusters  
- Churn Monitoring using Recency  
- Segment Tracking monthly

## 6. 🧾 Final Artifacts and Storage

- Cleaned Data: data/cleaned\_data.csv  
- RFM Data: data/rfm\_data\_filtered.csv  
- Final Labeled Data: data/clustered\_rfm.csv

- Models saved:  
 - models/kmeans\_model.pkl  
 - models/hierarchical\_model.pkl  
 - models/dbscan\_model.pkl  
 - models/scaler.pkl

- Visualizations: All plots stored in plots/

✅ Conclusion:  
KMeans clustering outperformed other methods, offering meaningful customer segments for marketing and operational strategy.