upGrad

BRIDGING THE GAP: CUSTOMER SEGMENTATION FOR ENHANCED RETENTION IN E-COMMERCE

This project focuses on identifying distinct customer segments in online retail data to better understand purchasing behavior, support targeted marketing, and improve retention strategies.

NAVNEET KUMAR

August 2025

snapdeal

EXECUTIVE SUMMARY

- Retention is cheaper than acquisition:
  - Our analysis confirms that a smaller segment of customers drives a major portion of revenue. Retaining these is more cost-efficient than marketing to acquire new customers in a competitive e-commerce market.

- K-Means Clustering is utilized to group customers into segments based on Recency, Frequency, and Monetary (RFM) metrics.

- Results reveal actionable profiles (VIP, Loyal, Regular, At-risk) that can guide personalized engagement strategies for each group.

PROBLEM STATEMENT & CONTEXT

- In e-commerce, failing to retain repeat customers leaves significant revenue on the table, especially when acquisition costs are rising.

- A data-driven approach to understanding customer behavior enables personalized strategies that foster loyalty.

- By applying unsupervised learning (K-Means), we can create distinct behavioral segments that form the foundation of a targeted CRM strategy.

DATA DESCRIPTION

- Source: Online Retail Dataset, covering Dec 2010 – Dec 2011, with over 540,000 transaction records.
- Key variables: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country.
- Majority (~85%) of customers are UK-based, with sales distributed across multiple countries.
- Transactions include both purchases and returns; data cleaning was essential to remove anomalies (e.g., negative quantities).

METHODOLOGY

Data Cleaning: Removed nulls in CustomerID, excluded canceled orders (negative quantities), removed duplicates, corrected monetary anomalies.

- Feature Engineering (RFM):
  - Recency: Days since customer’s most recent purchase.
  - Frequency: Total distinct purchase occasions.
  - Monetary: Total spend in the period.

- Scaling: StandardScaler applied to normalize RFM values for clustering.

- Segmentation: K-Means clustering run for k=3 and k=4, evaluating both for interpretability. k=3 gave clearer segment differentiation.

EXPLORATORY DATA ANALYSIS

- Recency Distribution: Large proportion of customers had very high recency (i.e., they hadn’t purchased in months), indicating churn risk.

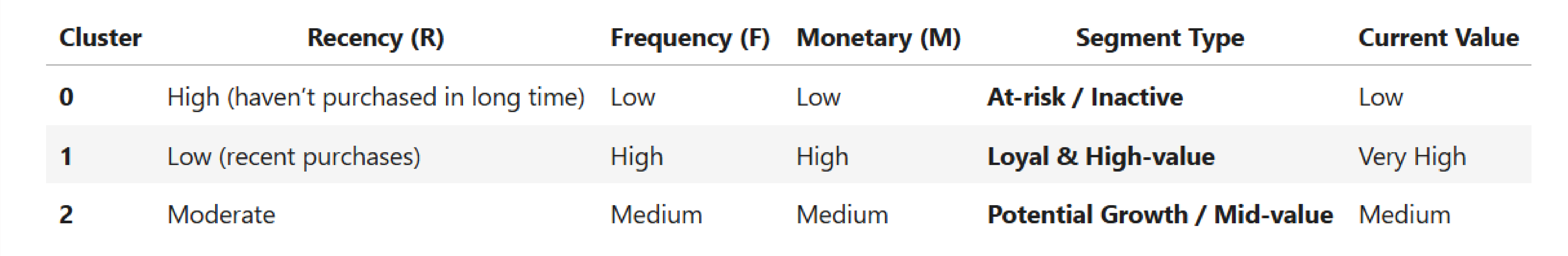
- Frequency Distribution: Highly skewed — most customers had low purchase counts, with a small group purchasing frequently.

- Monetary Distribution: Similar skew — a few high spenders contribute heavily to revenue.

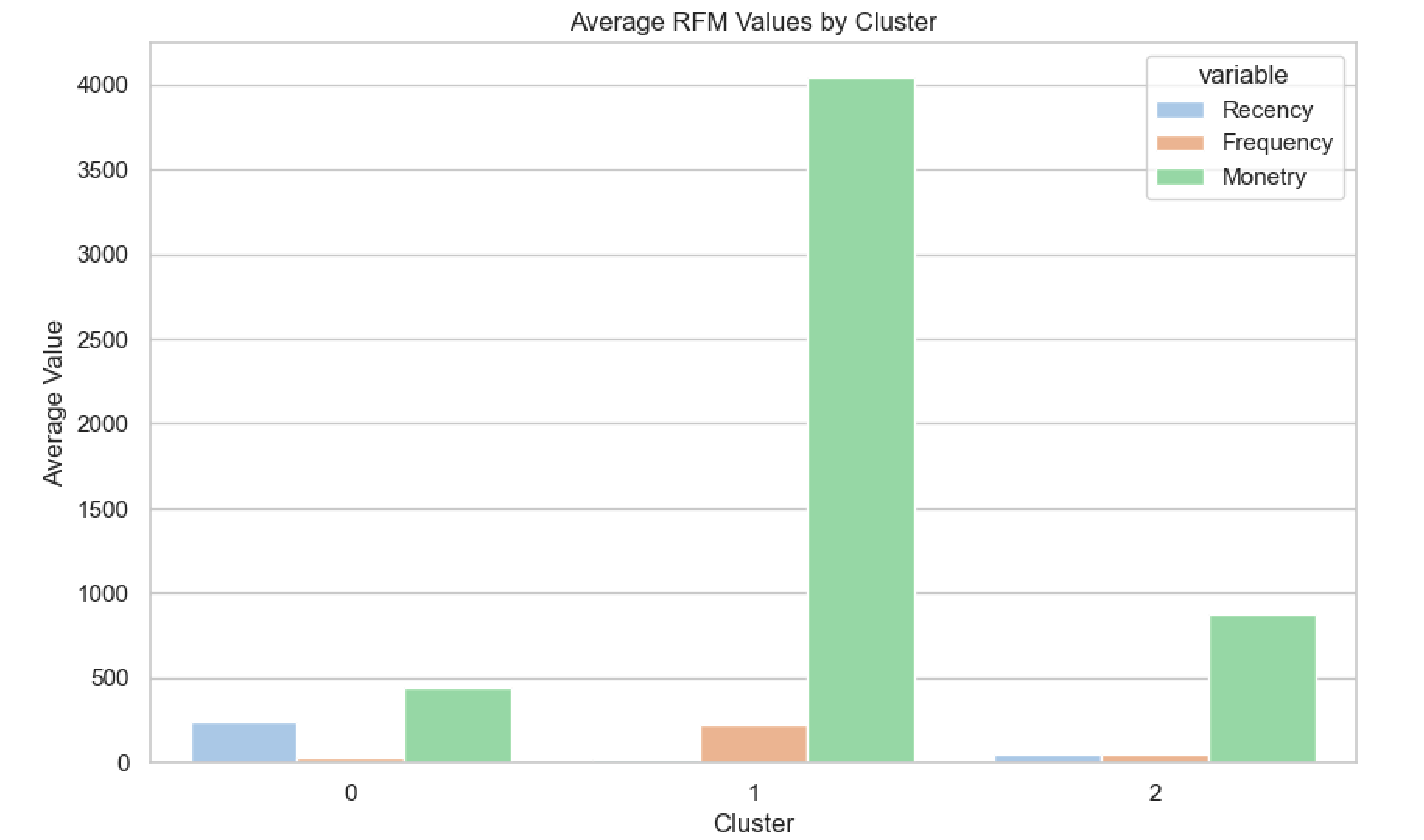
- RFM Correlation: Higher frequency often relates to higher spend, but recent buys don’t always mean the highest spenders.

SEGMENT OVERVIEW

Based on the analysis of the Recency, Frequency, and Monetary (RFM) metrics across three customer clusters:



Cluster	Recency (R)	Frequency (F)	Monetary (M)	Segment Type	Current Value
0	High (haven't purchased in long time)	Low	Low	**At-risk / Inactive**	Low
1	Low (recent purchases)	High	High	**Loyal & High-value**	Very High
2	Moderate	Medium	Medium	**Potential Growth / Mid-value**	Medium



CUSTOMER SEGMENTATION INSIGHTS (K=3)

Cluster 1 – High-Value

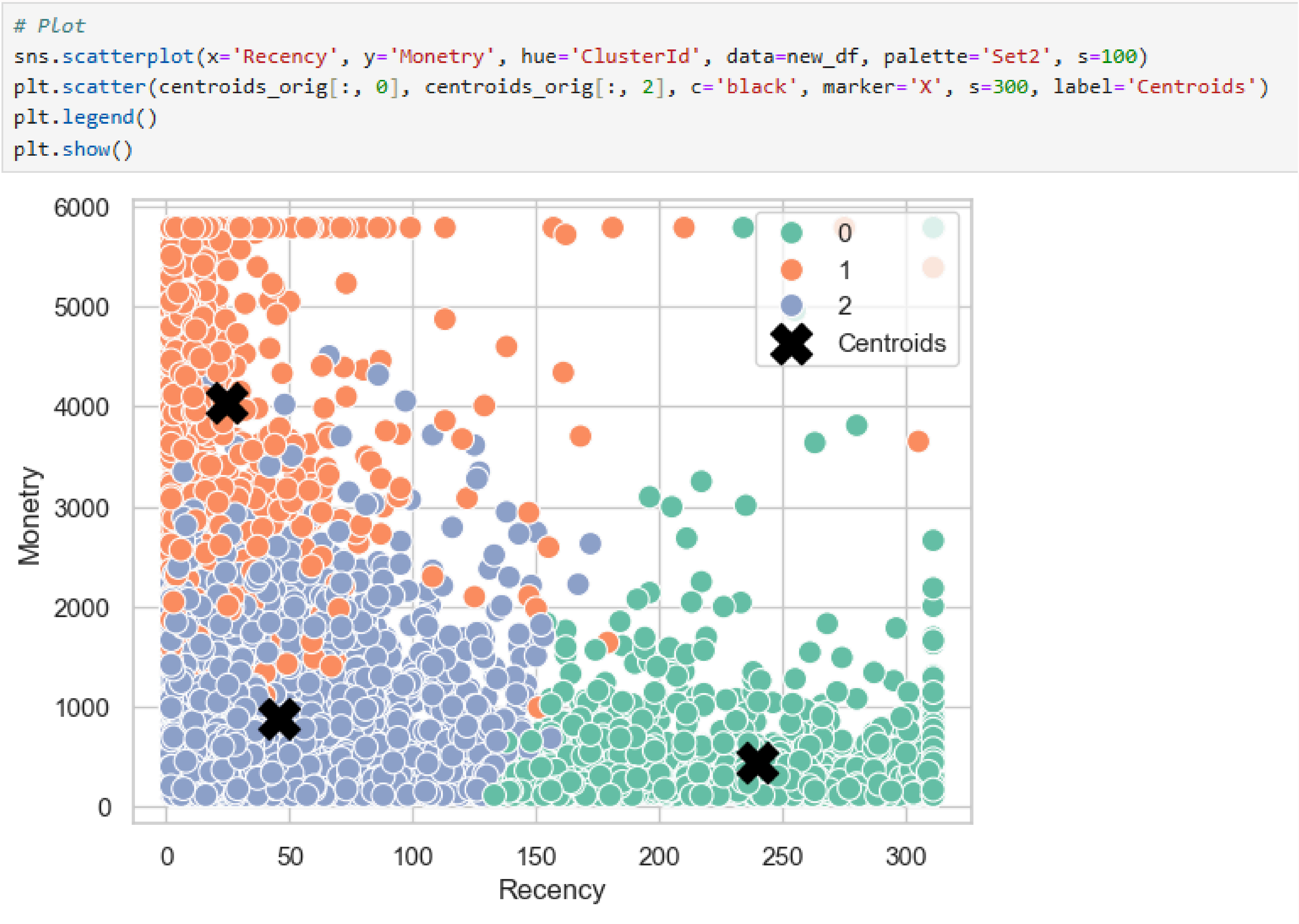
- Most engaged, frequent shoppers with high spend.
- Action: Reward loyalty with exclusives, upsell/cross-sell, and make them brand advocates.

Cluster 2 – Mid-Value

- Moderate spend, semi-regular buyers with growth potential.
- Action: Use targeted offers, personalized recommendations, and remarketing to move them up.

Cluster 0 – Low-Value

- Low spend, infrequent, long since last purchase.
- Action: Try cost-effective win-back campaigns, but focus resources on more valuable clusters.



```
# Plot
sns.scatterplot(x='Recency', y='Monetry', hue='ClusterId', data=new_df, palette='Set2', s=100)
plt.scatter(centroids_orig[:, 0], centroids_orig[:, 2], c='black', marker='X', s=300, label='Centroids')
plt.legend()
plt.show()
```

KEY INSIGHTS

- Most revenue (by far) is generated from a small, very active VIP cluster—these are the customers whose retention and satisfaction are crucial.

- A sizable segment has become inactive but had previously spent a fair amount (Cluster 2): they represent potential for revenue recovery if approached thoughtfully.
- The regular, moderate-spending cluster offers stable business but less dramatic opportunity for growth.

RECOMMENDATIONS

- **Cluster 1 – Loyal & High-Value Customers (Keep & Grow)**
  - **Importance:** These are the most engaged, frequent shoppers who spend the most.
  - **Opportunities:**
    - Strengthen brand loyalty with exclusive benefits (early access to sales, VIP tiers, personalized offers).
    - Implement cross-sell and up-sell tactics since they already trust the brand.
    - Gather customer feedback to improve product/service offerings; their opinion carries strong influence.
  - **Key Action:** Make them brand advocates by sustaining their engagement and rewarding loyalty.

- **Cluster 2 – Mid-Value Customers (Nurture & Convert)**
  - **Importance:** They buy somewhat regularly and spend moderately—they could be moved toward Cluster 1.
  - **Opportunities:**
    - Offer targeted promotions to encourage more frequent purchases (e.g., bundle deals, limited-time offers).
    - Introduce personalized recommendations using previous purchase data.
    - Use remarketing campaigns to bring them back before engagement drops.
  - **Key Action:** Push them toward high-value behavior through incentives and personalized experiences.

- **Cluster 0 – At-risk / Inactive Customers (Win-back or Let Go)**
  - **Importance:** Low frequency, low spend, purchased long ago—least valuable segment today.
  - **Opportunities:**
    - Win-back campaigns: Special “We miss you” discounts, time-sensitive offers, or new product announcements.
    - Identify why they disengaged (price, product variety, competition, poor service).
    - If acquisition and retention cost outweighs potential return, focus resources on higher-value clusters.
  - **Key Action:** Attempt recovery – but prioritize cost-efficient reactivation strategies.

RECOMMENDATIONS FOR BUSINESS TO IMPROVE EXPERIENCE & REPEAT PURCHASES

- Personalized Engagement Strategies
  - Use segmentation to design cluster-specific campaigns.
  - Communicate via preferred channels (email, SMS, app notifications) for each segment.

- Loyalty & Retention Programs
  - Introduce a tiered loyalty program with rewards increasing based on spending and engagement.
  - Offer Cluster 1 top-tier perks; allow Cluster 2 an attainable pathway to reach them.

- Reactivation Plans
  - For Cluster 0: Time-limited discounts, product trial offers, and re-engagement email campaigns.
  - Consider seasonal campaigns to pull them back into shopping cycles.

- Upsell & Cross-sell Initiatives
  - For Cluster 1 and Cluster 2: Use recommendation engines and curated bundles.
  - Position premium products or add-ons during checkout.

- Customer Feedback & Experience Enhancement
  - Routinely collect satisfaction feedback from Cluster 1 (loyal customers) and Cluster 2 (growth customers) to improve offerings.
  - Address pain points that might be causing churn in Cluster 0.

- Monitor Movement Between Clusters
  - Track how customers transition between clusters over time to measure marketing effectiveness.
  - Adjust strategies dynamically as segments evolve.

Conclusion

- K-Means clustering provided clear, actionable customer segments that can directly enhance CRM retention strategies.
- Addressing the needs of each segment differently maximizes revenue while reducing acquisition costs.
- This analysis showcases data science impact on business outcomes — turning transaction logs into targeted marketing advantages.

upGrad