Customer Segmentation Analysis

1. Executive Summary

This report details a comprehensive customer segmentation analysis performed on a retail transaction dataset, utilizing RFM (Recency, Frequency, Monetary) analysis and K-Means clustering. The objective was to identify distinct customer groups based on their purchasing behavior to enable targeted marketing strategies.

Through rigorous data preprocessing and clustering, **four key customer segments** were identified:

1. **Loyal Champions:** Highly recent, frequent, and high-spending customers.
2. **New & Promising:** Recent but less frequent, lower-spending customers with potential for growth.
3. **At-Risk Loyalists:** Customers who were once valuable but have become less recent in their purchases, indicating a risk of churn.
4. **Churned/Lost:** Infrequent, low-spending customers who have not purchased recently, indicating potential inactivity.

Actionable recommendations are provided for each segment to optimize customer retention, engagement, and revenue generation.

2. Introduction

In today's competitive retail landscape, understanding customer behavior is paramount for business growth. This project aims to segment customer base into distinct groups based on their transactional history. By identifying these segments, Indolike can tailor marketing campaigns, personalize communications, and allocate resources more effectively, ultimately enhancing customer lifetime value and overall profitability.

The analysis utilizes a comprehensive online retail transaction dataset spanning from December 2009 to December 2011, sourced from Kaggle (online_retail_II.xlsx).

3. Data Acquisition & Initial Exploration

The dataset was loaded and combined from two separate Excel sheets (Year 2009-2010 and Year 2010-2011). The combined dataset initially contained **1,067,371 rows** and 8 columns: Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country.

Initial exploration revealed key data quality issues:

- **Missing Customer IDs:** Approximately 22.7% (243,007 rows) of records had missing Customer IDs, which are essential for customer-level analysis.
- **Negative Quantities:** Presence of negative Quantity values, indicating product returns or cancellations rather than purchases.
- **Negative Prices:** Isolated instances of negative Price values, which are illogical for sales transactions.
- **Missing Descriptions:** A smaller number of missing Description values.

4. Data Cleaning & Preprocessing

To ensure the integrity and relevance of the analysis, the following cleaning steps were performed:

- **Handling Missing Customer IDs:** Rows with missing Customer IDs were removed, as these transactions could not be attributed to specific customers, reducing the dataset to **824,364 rows**. The Customer ID column was then converted to an integer type.
- **Handling Non-Positive Quantities and Prices:** Rows with Quantity less than or equal to 0 or Price less than or equal to 0 were removed, focusing the analysis solely on actual purchase transactions. This further reduced the dataset to **805,549 rows**.
- **Calculating Total Price:** A new column, TotalPrice, was created by multiplying Quantity and Price for each transaction line item.

5. Feature Engineering (RFM Analysis)

To characterize customer behavior effectively, RFM (Recency, Frequency, Monetary) features were engineered for each unique customer:

- **Recency (R):** Calculated as the number of days since a customer's last purchase. A snapshot date (one day after the latest transaction in the dataset) was used as the reference point. **Lower Recency indicates a more recently active customer.**
- **Frequency (F):** Calculated as the total number of unique invoices for each customer. **Higher Frequency indicates a more engaged and repeat customer.**
- **Monetary (M):** Calculated as the sum of all TotalPrice for each customer. **Higher Monetary value indicates a more valuable customer in terms of spending.**

The resulting RFM DataFrame rfm_df contains **5,878 unique customer entries**.

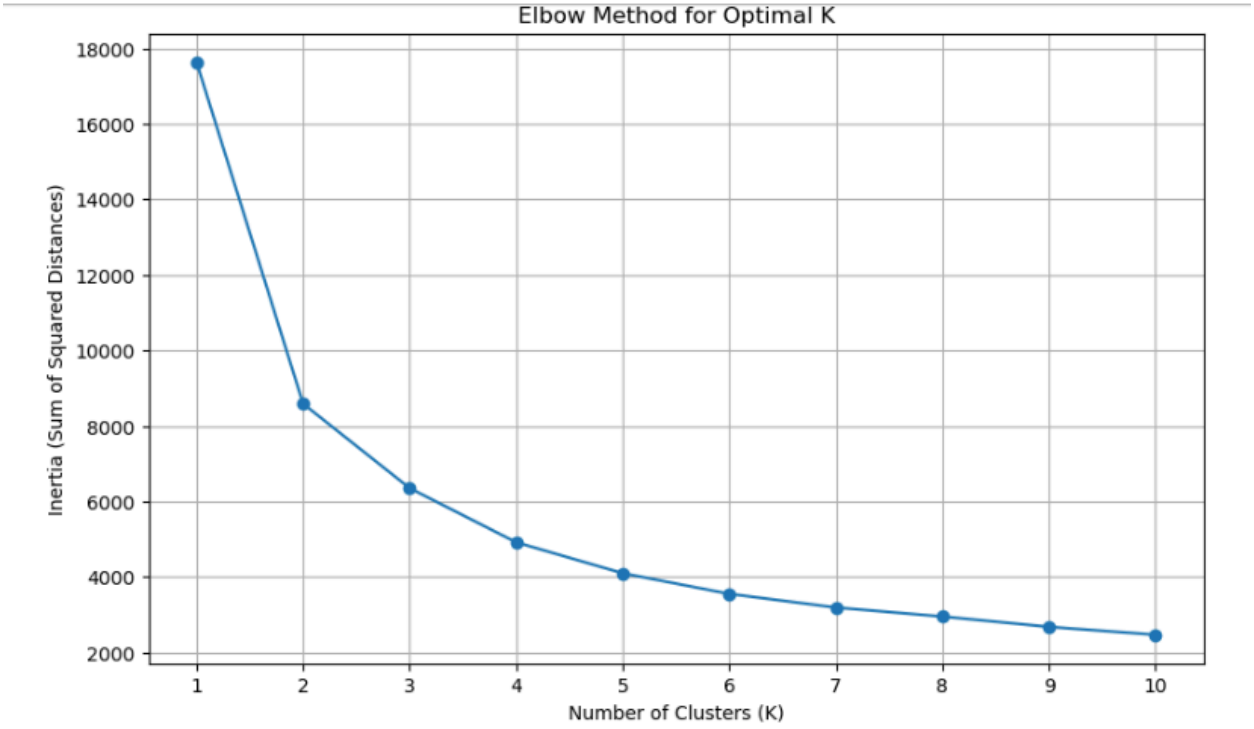
6. Data Transformation & Scaling for Clustering

The raw RFM values were highly skewed, with a few customers exhibiting extremely high frequency and monetary values. K-Means clustering performs best with features that are normally distributed and on a similar scale. Therefore, the following transformations were applied:

- **Log Transformation (np.log1p):** Applied to Recency, Frequency, and Monetary to reduce skewness and mitigate the impact of outliers. This brings the distributions closer to a Gaussian shape.
- **Standard Scaling (StandardScaler):** Applied to the log-transformed RFM values. This standardizes the features to have a mean of 0 and a standard deviation of 1, ensuring that each feature contributes equally to the distance calculations in K-Means.

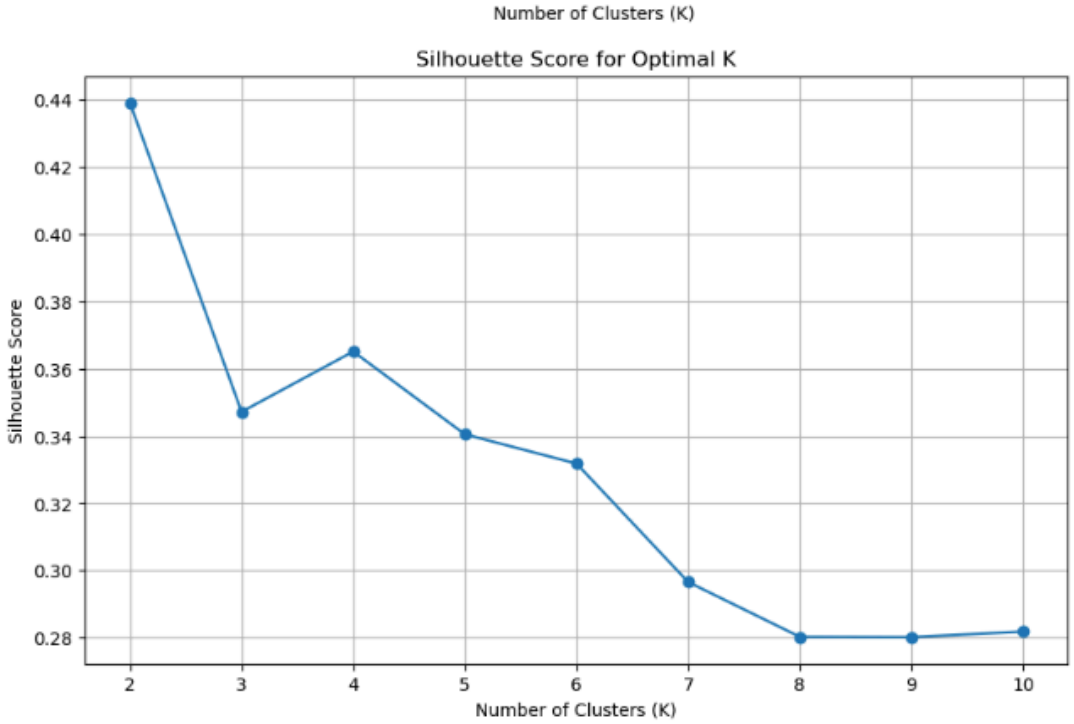
7. K-Means Clustering

To determine the optimal number of clusters (k), both the Elbow Method and Silhouette Score were employed.

- **Elbow Method:** The plot of inertia (sum of squared distances) against the number of clusters indicated an "elbow" at K=3 or K=4, suggesting that adding more clusters beyond this point yielded diminishing returns in reducing within-cluster variance.
  - 

- **Silhouette Score:** This metric measures how similar an object is to its own cluster compared to others. While K=2 yielded the highest silhouette score, K=4 provided

a balance, being the second highest and offering a more granular, yet well-separated, segmentation for business actionability.



Based on a combined analysis of both methods, **K=4 was selected** as the optimal number of clusters, offering a balance between cluster compactness/separation and business interpretability.

The K-Means algorithm was then applied to the scaled RFM data, assigning each of the 5,878 customers to one of the four identified clusters.

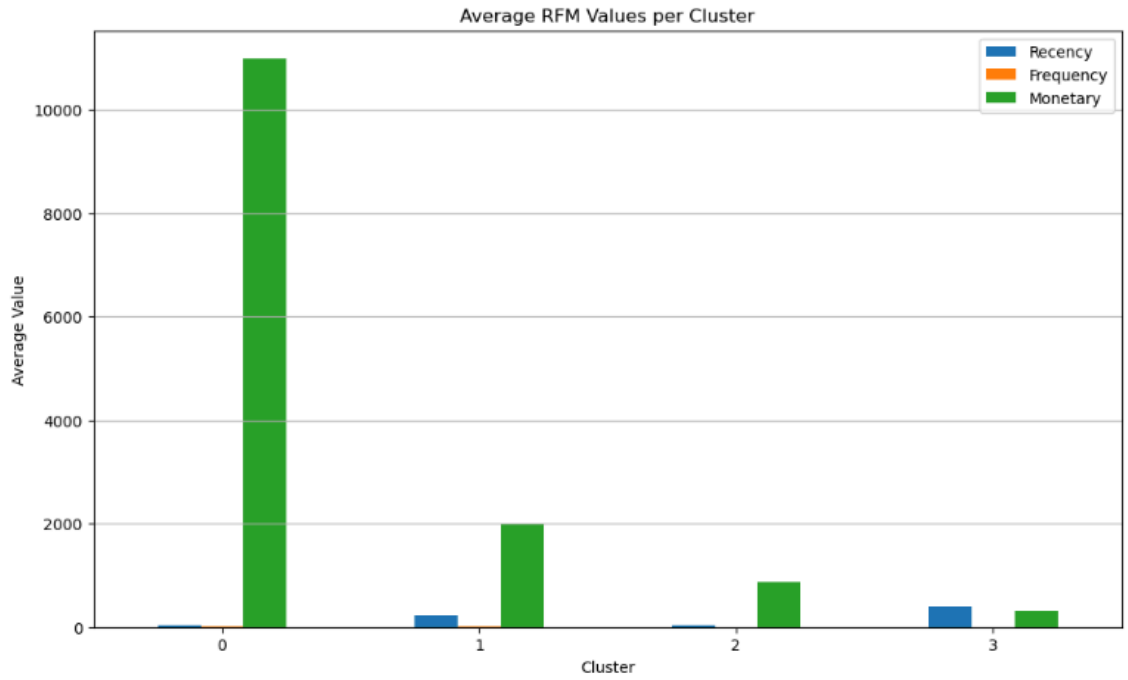
8. Customer Segment Characterization & Naming

The average Recency, Frequency, and Monetary values for each cluster (using the original, unscaled RFM values) were calculated to understand their distinct behavioral patterns:

Cluster	Recency (Days)	Frequency (Purchases)	Monetary (Total Spend)	Customer Count
0	27.60	19.29	10979.16	1194
1	229.73	5.06	1984.21	1469
2	28.41	3.04	867.89	1253

3	396.23	1.38	322.15	1962

```
3         396.229867    1.375637    322.149980
```




- ```
  Mean RFM Values (Transposed for easy comparison)
  ```

Based on these characteristics, the four segments were named and interpreted as follows:

- **Cluster 0: "Loyal Champions" (Count: 1194)**
  - **Characteristics:** These customers are highly recent (low Recency), very frequent purchasers (high Frequency), and generate the highest revenue (very high Monetary). They represent the most valuable and engaged segment.
  - **Strategic Value:** Core customer base, brand advocates, and key revenue drivers.
- **Cluster 2: "New & Promising" (Count: 1253)**
  - **Characteristics:** Highly recent purchasers (low Recency) but with lower frequency and monetary values compared to 'Loyal Champions'. These are likely newer customers who have recently engaged.
  - **Strategic Value:** Untapped potential for growth; can be nurtured into more loyal segments.
- **Cluster 1: "At-Risk Loyalists" (Count: 1469)**
  - **Characteristics:** Moderately frequent and high-spending customers (moderate Frequency, good Monetary), but with a relatively high Recency, indicating they haven't purchased recently.
  - **Strategic Value:** High potential for re-engagement; preventing churn is
  ```

critical.

- **Cluster 3: "Churned/Lost" (Count: 1962)**
  - **Characteristics:** Very high Recency (haven't purchased in a long time), very low Frequency, and very low Monetary value. This is the least engaged and least profitable segment.
  - **Strategic Value:** Low re-engagement potential; focus on low-cost win-back efforts or minimal resource allocation.

9. Actionable Business Recommendations

Based on the identified customer segments, the following tailored strategies are recommended :

- **For "Loyal Champions":**
  - **Retention & Reward:** Implement exclusive loyalty programs, VIP access to new products or sales, and personalized thank-you gestures.
  - **Advocacy Programs:** Encourage referrals and reviews, leveraging their brand loyalty.
  - **Premium Offerings:** Introduce higher-value products or services that align with their demonstrated spending habits.
- **For "New & Promising":**
  - **Onboarding & Nurturing:** Develop a personalized welcome series introducing product categories, benefits, and usage tips.
  - **Incentivize Repeat Purchases:** Offer small discounts on their next purchase or free shipping to encourage continued engagement.
  - **Cross-Selling:** Recommend complementary products based on their initial purchases.
- **For "At-Risk Loyalists":**
  - **Win-Back Campaigns:** Design targeted campaigns with compelling personalized offers (e.g., discounts on past favorite items, exclusive access).
  - **Re-engagement Communications:** Send "we miss you" emails, highlighting recent product updates or platform improvements.
  - **Feedback Collection:** Consider surveys to understand reasons for decreased activity and address pain points.
- **For "Churned/Lost":**
  - **Minimal Effort Re-engagement:** Limit investment to low-cost, broad re-engagement efforts like major seasonal sales announcements.
  - **Focus on Acquisition:** Prioritize resources towards acquiring new customers, as the cost of re-engaging this segment is likely high for minimal return.
  - **Long-Term Monitoring:** Keep them on a low-frequency mailing list in case

they naturally re-engage.

10. Conclusion & Future Work

This customer segmentation analysis provides a clear, actionable framework for understanding and targeting diverse customer base. By moving beyond a one-size-fits-all approach, Indolike can significantly improve customer satisfaction, retention, and overall revenue.

Future Work:

- **Integrate Demographic Data:** If available, incorporate demographic information (Country, age, etc.) to enrich segment profiles.
- **Product-Level Insights:** Analyze which product categories are favored by specific segments to further refine recommendations.
- **Dynamic Segmentation:** Explore more advanced clustering techniques or time-series analysis to observe how customers migrate between segments over time.
- **A/B Testing:** Implement and A/B test the recommended strategies to measure their effectiveness and optimize future campaigns.