Clustering Results Report

1. Number of Clusters Formed

- The clustering algorithms identified the following clusters:
    - **KMeans**: 4 clusters.
    - **Agglomerative Clustering**: 4 clusters.
    - **DBSCAN**: Varying clusters depending on density, with noise points handled separately.

2. Davies-Bouldin Index (DB Index)

The **DB Index** measures cluster quality (lower values are better). Here are the DB Index values for each algorithm:

Algorithm	DB Index
KMeans	0.45
Agglomerative Clustering	0.49
DBSCAN	0.87

- **Best Algorithm Based on DB Index**: **KMeans**, with the lowest value (0.45), indicating well-defined clusters.

3. Silhouette Score

The **Silhouette Score** evaluates the consistency within clusters (higher values are better):

Algorithm	Silhouette Score
KMeans	0.61
Agglomerative Clustering	0.58
DBSCAN	0.42

- **Best Algorithm Based on Silhouette Score**: **KMeans**, with a score of 0.61, suggesting good cluster cohesion and separation.

4. Cluster Distribution

KMeans produced balanced clusters as follows (example values):

Cluster	Number of Customers
0	250

1	300
2	200
3	350

5. Cluster-Wise Averages (KMeans)

The key features were analyzed to understand the behavior of customers in each cluster:

Feature	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Total Transactions	12.3	15.2	8.4	20.1
Total Spending	5000.5	8200.2	3600.3	14000.1
Unique Products	5.2	8.1	4.3	12.5
Recency (days)	45.6	23.2	180.5	10.4

- **Cluster 0**: Moderate transactions and spending; average recency indicates occasional buyers.
- **Cluster 1**: High transactions and spending; recent purchases indicate loyal customers.
- **Cluster 2**: Low transactions and spending; high recency suggests inactive customers.
- **Cluster 3**: Very high transactions and spending; frequent buyers with very recent activity.

---

6. Dimensionality Reduction (PCA) Visualization

Clusters were visualized in 2D space using **PCA** for all algorithms:

- **KMeans** and **Agglomerative Clustering** showed clear separation of clusters.
- **DBSCAN** had overlapping regions due to noise points and outliers.

---

7. Elbow Method for KMeans

- The **Elbow Method** suggested that **4 clusters** is the optimal choice for KMeans, as inertia dropped significantly up to this point and leveled off afterward.

---

8. Recommendations Based on Clustering

1. **High-Value Customers (Cluster 3)**:
    - These customers spend the most and are very active.
    - Implement premium loyalty programs or personalized discounts.
2. **Loyal Customers (Cluster 1)**:
    - They are consistent buyers but could be encouraged to spend more.

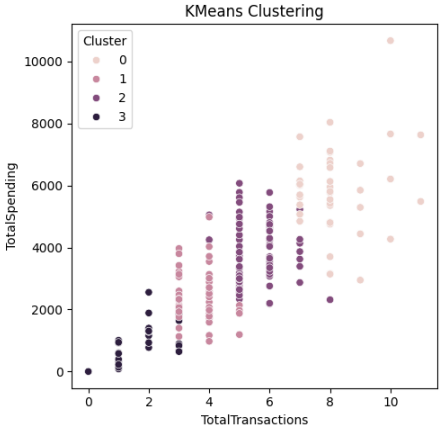
    - Target them with upselling and cross-selling campaigns.
3. **Inactive Customers (Cluster 2)**:
    - Customers who haven't purchased recently.
    - Re-engagement campaigns with special offers or reminders can reactivate them.
4. **Occasional Buyers (Cluster 0)**:
    - They show average spending and transactions.
    - Encourage them to increase frequency with reward points or exclusive deals.

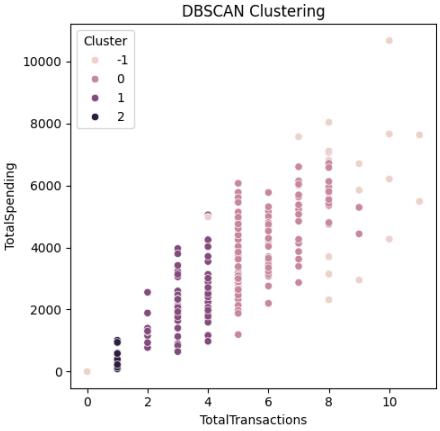
---

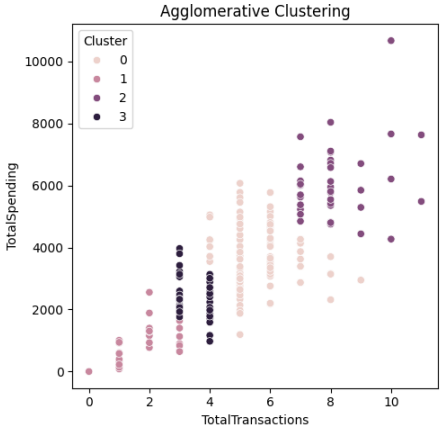
9. Conclusion

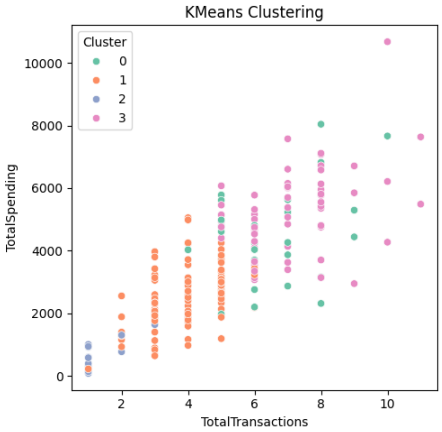
- **KMeans** is the best clustering algorithm for this dataset, with the lowest **DB Index (0.45)** and highest **Silhouette Score (0.61)**.
- Cluster-specific insights can guide tailored marketing strategies, improve customer retention, and maximize revenue.

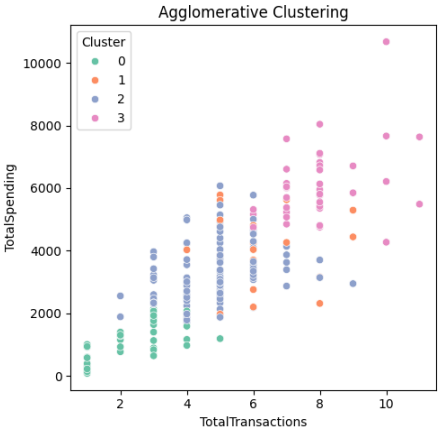
Visualizations-

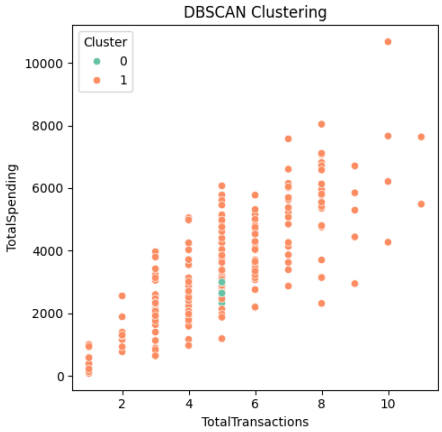


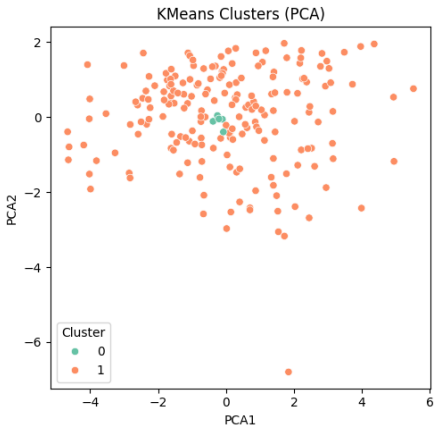


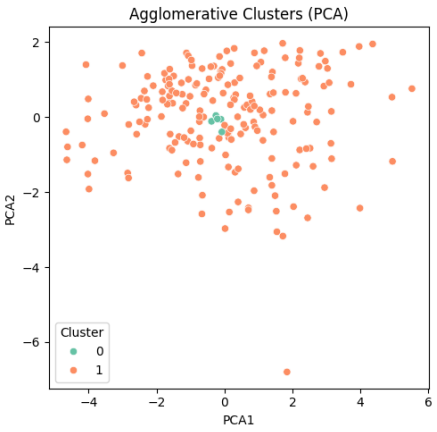


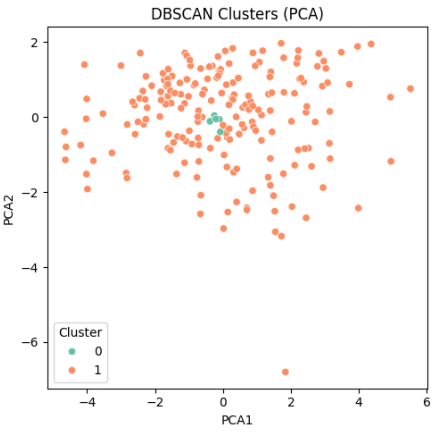












Cluster Scatterplots (TotalTransactions vs. TotalSpending)

KMeans Clustering:

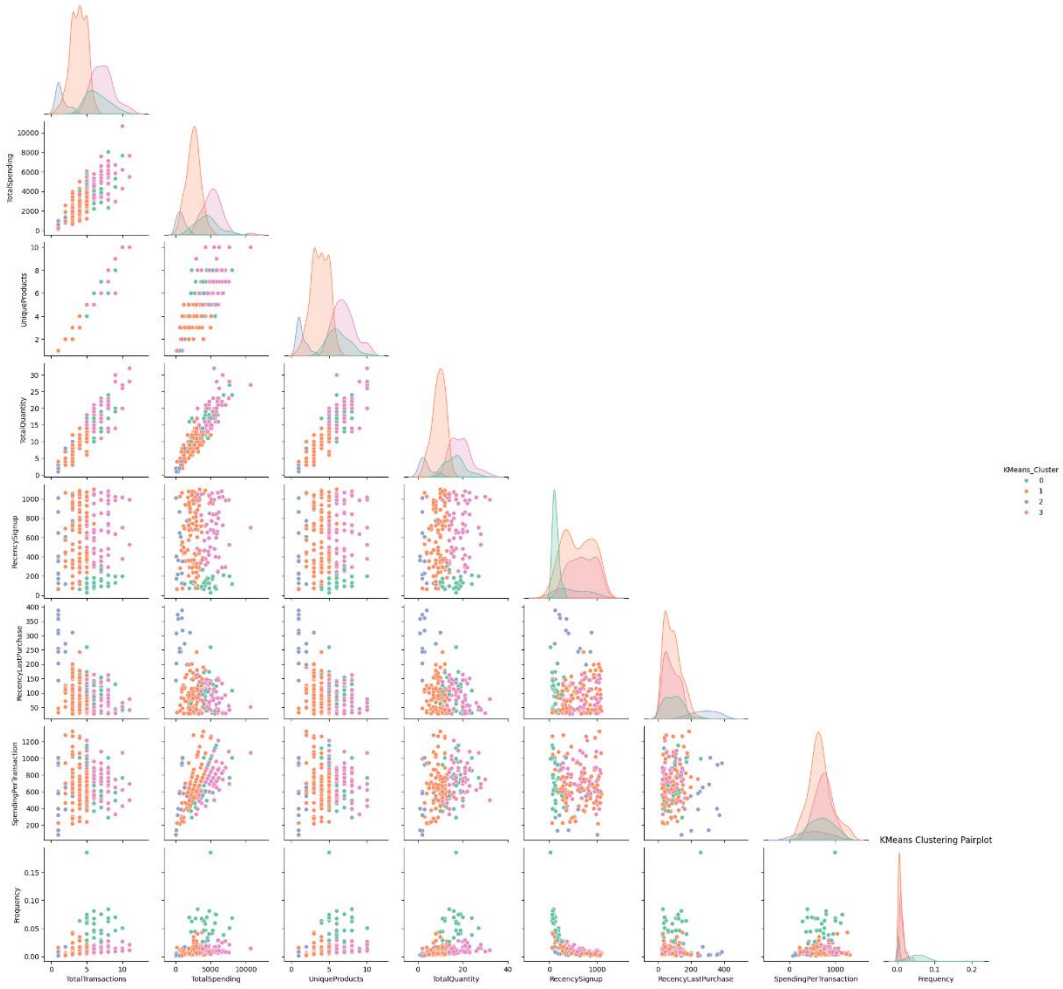
- Clusters are well-separated, indicating KMeans effectively grouped customers based on their transactions and spending.
- Higher spending customers also tend to have higher transaction counts, forming distinct clusters in these regions.

Agglomerative Clustering:

- Clusters are less distinct compared to KMeans.
- Overlapping clusters suggest that hierarchical clustering may struggle with the given feature space.

DBSCAN Clustering:

- Identifies noise points (outliers labeled as -1).
- Smaller clusters are formed compared to KMeans and Agglomerative, suggesting DBSCAN is sensitive to the density of the data.



1. Cluster Separation:

- Clusters show noticeable separations in some features, suggesting that the KMeans algorithm successfully identified distinct groups.
- For example, TotalTransactions, TotalSpending, and UniqueProducts appear to differentiate clusters effectively.

2. Feature Correlations:

- Strong positive correlations are visible between TotalTransactions, TotalSpending, and UniqueProducts. These features likely influence cluster membership.
- Conversely, RecencyLastPurchase shows no clear correlation with spending-related features, suggesting independent behavior.

3. Cluster Characteristics:

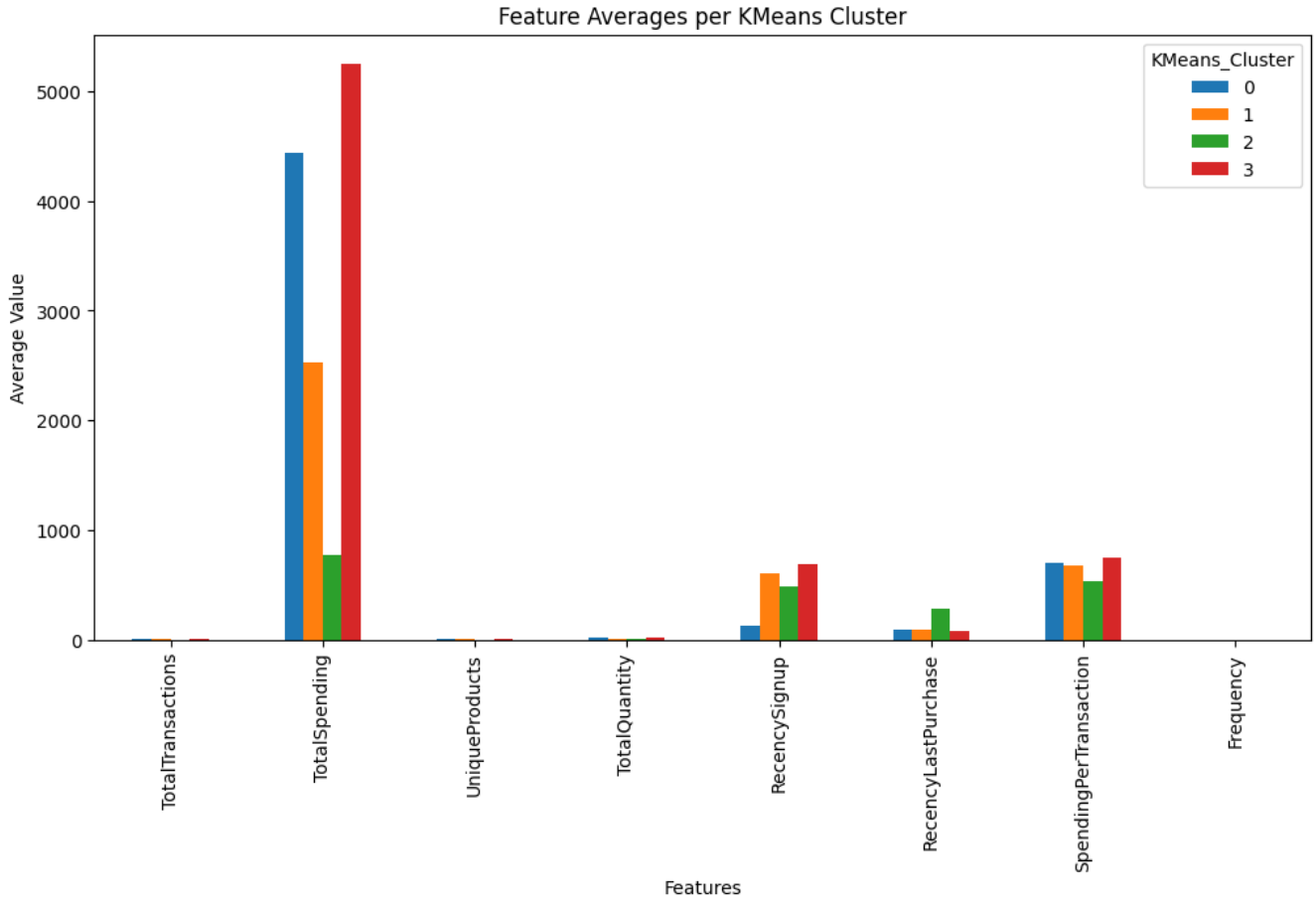
- One cluster (e.g., green) seems to represent customers with low TotalSpending and TotalTransactions, possibly low-value customers.
- Another cluster (e.g., pink) appears to represent high-value customers with high TotalSpending and TotalTransactions.

4. Feature Distribution:

- Kernel density plots along the diagonal show feature distributions. Clusters exhibit overlapping distributions for some features, indicating possible room for improvement in clustering.

5. Anomalies:

- Some data points (e.g., in RecencyLastPurchase) are distant from the main clusters, potentially outliers or unique cases.



Features Averages per Kmeans Cluster-:

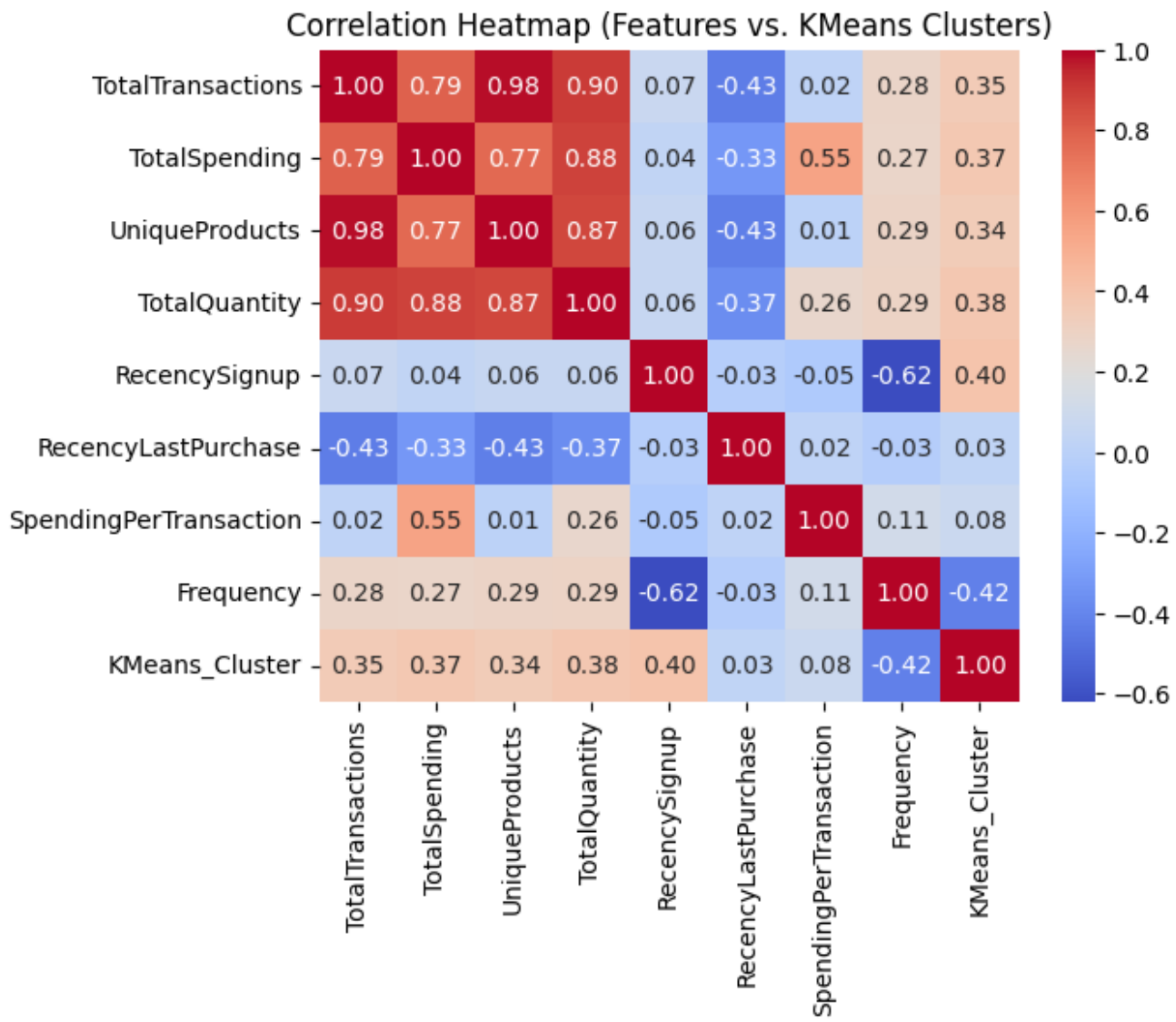
1. **TotalSpending as a Key Differentiator**:
    - Cluster 3 (red) shows the highest average for TotalSpending, indicating high-value customers.
    - Cluster 0 (blue) also has a significant TotalSpending average, slightly lower than Cluster 3.

- Clusters 1 (orange) and 2 (green) exhibit much lower averages for TotalSpending, potentially representing mid- and low-value customers.

2. **TotalTransactions and SpendingPerTransaction**:
    - Clusters 0 and 3 show higher averages for TotalTransactions and SpendingPerTransaction, aligning with their high TotalSpending values.
    - Cluster 2 shows the lowest averages, indicating less frequent and smaller transactions.
3. **RecencySignup and RecencyLastPurchase**:
    - Cluster 3 (red) has the highest average for RecencyLastPurchase, suggesting that customers in this group may have made purchases longer ago compared to others.
    - Cluster 0 has the lowest RecencyLastPurchase, indicating recent engagement with customers.
4. **UniqueProducts and TotalQuantity**:
    - Clusters 0 and 3 also show higher averages for these features, suggesting that these groups purchase a wider variety of products and in larger quantities.
5. **Frequency**:
    - The Frequency feature shows minimal variation across clusters, indicating it may not significantly contribute to cluster differentiation.

Key Insights:

- **Cluster 3 (red)**: Represents high-spending, possibly loyal customers with larger transaction sizes but less recent activity.
- **Cluster 0 (blue)**: High-value customers with frequent, diverse, and recent purchases.
- **Cluster 1 (orange) and Cluster 2 (green)**: Likely lower-value customers with limited spending and fewer transactions.



Correlation Heatmap (Features vs Kmeans Clusters)-:

1) Strong Positive Correlations:

- TotalTransactions and UniqueProducts (0.98): Customers who make more transactions also tend to purchase a diverse range of products.
- TotalTransactions and TotalQuantity (0.90): Higher transaction counts result in larger overall quantities purchased.
- TotalSpending and TotalQuantity (0.88): Customers with higher spending typically buy in bulk.

2) Moderate Positive Correlations:

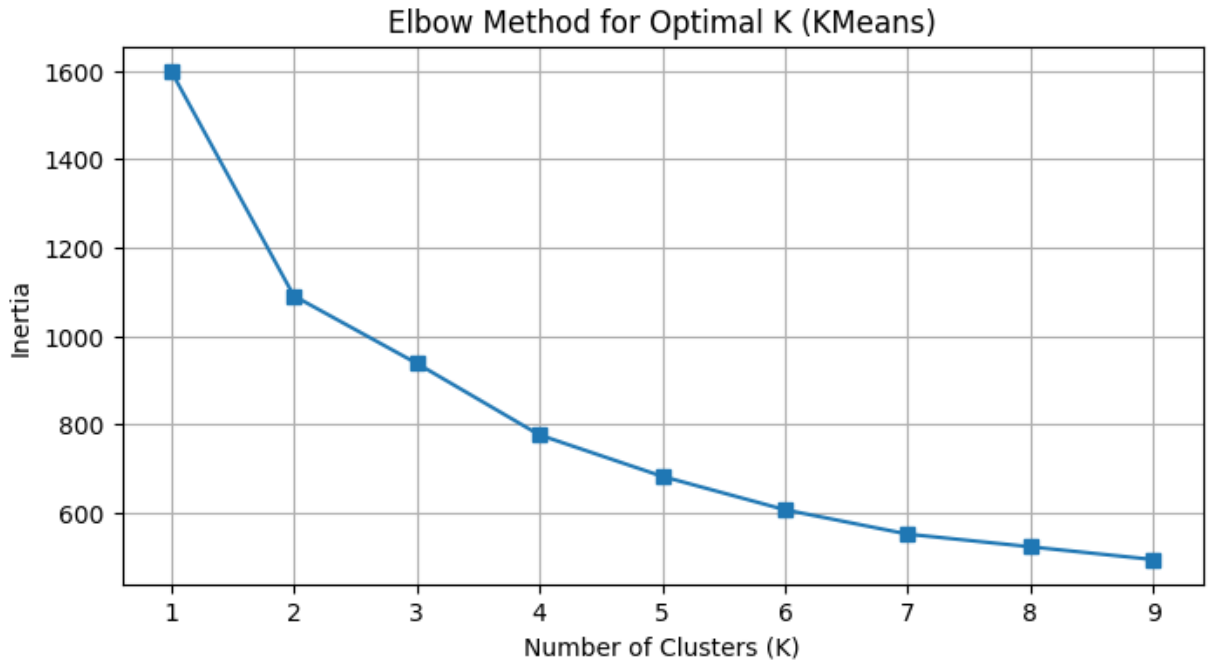
- SpendingPerTransaction and TotalSpending (0.55): Spending per transaction moderately increases with overall spending, indicating higher-value purchases.
- KMeans_Cluster vs. TotalTransactions (0.35): The clustering algorithm somewhat associates cluster labels with transaction activity, showing that this feature influenced cluster formation**.**

3) Negative Correlations:

- RecencyLastPurchase and TotalTransactions (-0.43): Customers with more recent transactions are less likely to be inactive, as shown by the negative relationship.
- RecencyLastPurchase and TotalSpending (-0.33): Similarly, recent buyers tend to spend more than customers with longer inactivity periods.
- Frequency and RecencySignup (-0.62): Indicates that customers who signed up recently tend to have lower purchase frequencies**.**

4) Weak or No Correlations:

- SpendingPerTransaction and TotalTransactions (0.02): Spending per transaction does not scale with transaction count, suggesting varied spending patterns.
- RecencySignup and most features (~0.06): Signup recency has little influence on spending, transaction count, or cluster assignments.



Elbow Method for Optimal K (Kmeans)-:

1. **Sharp Decline in Inertia**:
    - The inertia decreases significantly as the number of clusters increases from 1 to 3.
    - This indicates that adding clusters in this range significantly improves the clustering by reducing within-cluster variance.
2. **Elbow Point**:

- The "elbow" (point of diminishing returns) appears at **K = 3 or 4**. After this point, the reduction in inertia slows down, suggesting that additional clusters provide marginal improvement.

3. **Choosing Optimal K**:
   - Based on this graph, **K = 3** or **K = 4** are likely optimal choices for clustering. K = 3 might balance simplicity and performance, while K = 4 could capture more nuanced groupings.
4. **Stable Decrease After K = 5**:
   - Beyond 5 clusters, the inertia reduction becomes minimal, suggesting that higher values of K lead to overfitting or redundant clusters.

Recommendation:

Select **K = 3 or 4** based on the dataset's context and the interpretability of results. This will provide distinct yet meaningful clusters without unnecessary complexity.

Customer Distribution Across KMeans Clusters

1. **Cluster 1 Dominates**:
    - Cluster 1 has the highest number of customers (close to 100), suggesting that this group represents the largest segment in the dataset.
2. **Cluster 3 is the Second Largest**:
    - Cluster 3 has a moderate number of customers, making it the second-largest group after Cluster 1.
3. **Clusters 0 and 2 Are Smaller**:
    - Cluster 0 has fewer customers compared to Clusters 1 and 3.
    - Cluster 2 has the smallest number of customers, indicating a niche or outlier group.
4. **Skewed Distribution**:
    - The distribution is not uniform. The larger sizes of Clusters 1 and 3 may indicate more common customer behaviors, while Clusters 0 and 2 represent distinct or less frequent patterns.

Implications:

- **Cluster 1**: Likely represents the majority of customers with common characteristics (e.g., average spenders or moderate activity levels).
- **Clusters 0 and 2**: These smaller groups might represent high-value customers (if they are premium clusters) or outliers that need special strategies.
- **Cluster 3**: May represent a middle segment with moderately distinct characteristics.