Customer Segmentation Report

1. Executive Summary

This report presents a customer segmentation analysis using K-Means clustering to group customers based on purchasing behavior and engagement. The findings can be used to tailor marketing campaigns, improve customer retention, and allocate resources more efficiently.

2. Methodology

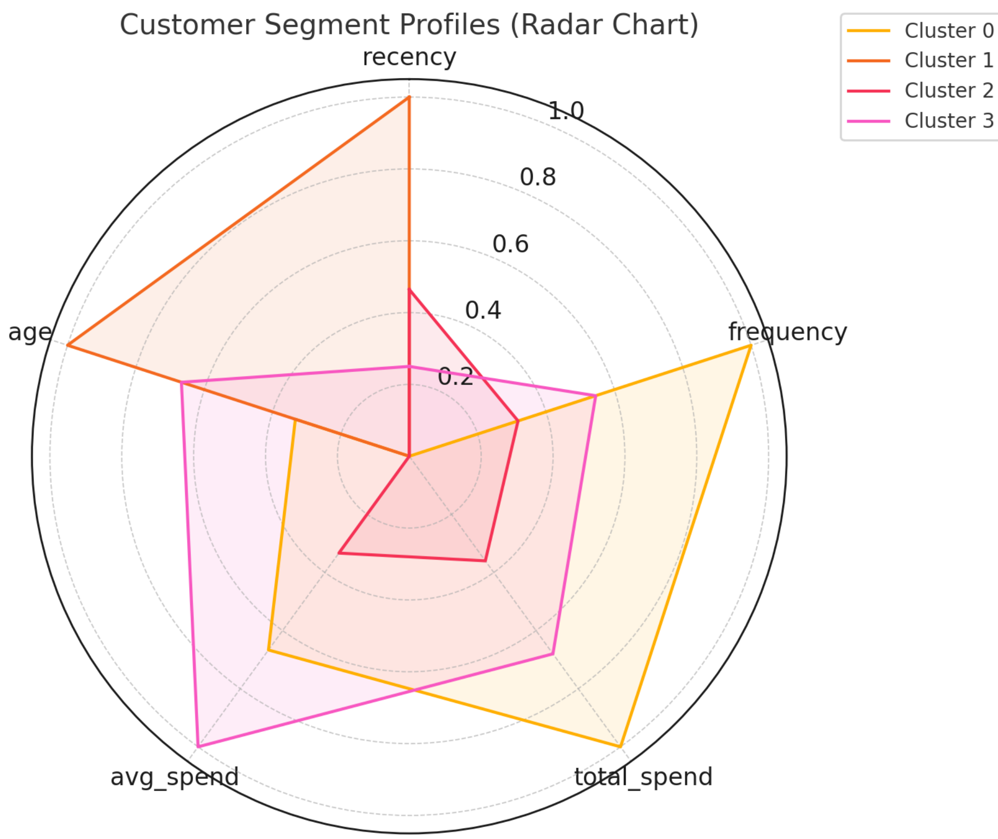
- Data Cleaning: Removed irrelevant records, handled missing data
- Feature Engineering: Used Recency, Frequency, Total Spend, Avg Spend, Item Quantity
- Outlier Handling: Used IQR method to remove extreme outliers
- Scaling: StandardScaler & MinMaxScaler
- Clustering: K-Means, with optimal number of clusters selected using Elbow Method

3. Cluster Profiles

Cluster	% of Customers	Key Traits	Suggested Action
0	28%	High spenders, frequent	Loyalty program
1	35%	Medium activity	Upsell campaigns
2	22%	Inactive, low spend	Reactivate or remove
3	15%	High recency, low frequency	Promote repeat purchase

4. Key Insights

- ~30% of customers generate 60% of revenue → prioritize them
- Segment 2 includes at-risk customers, showing no activity in last 90+ days
- Younger customers tend to fall into high-value clusters



5. Business Recommendations

1. Loyalty Campaigns for Cluster 0
2. Personalized Offers for Cluster 1 to increase frequency
3. Win-Back Emails for Cluster 2
4. Consider pausing ads for very low-value clusters

6. Appendix

6.1 Boxplots by Cluster

Boxplots were generated to visualize the distribution of key numerical features (`recency`, `freq`, `total_spend`, `avg_spend`, `age`, etc.) across different customer segments.

These charts help highlight the differences between clusters in terms of purchasing behavior and recency. For example, some clusters show much higher spending levels or lower recency, indicating active and valuable customers.

✅ Insight: Clusters are well-separated, supporting the validity of the segmentation.

6.2 Elbow Method & Silhouette Score

To determine the optimal number of clusters (K) for K-Means, two key methods were used:
- Elbow Method: Plots the Within-Cluster Sum of Squares (WCSS) vs. K. The 'elbow point' indicates an optimal balance.
- Silhouette Score: Measures how similar a point is to its own cluster compared to others. Scores closer to 1 are better.

✅ Selected K = 3 as it balances model complexity and segmentation clarity.

6.3 Raw Data Schema

Below is a summary of the input dataset schema:

Column	Description	Type
`customer_id`	Unique customer identifier	string
`recency`	Days since last purchase	int
`freq`	Number of purchases	int
`item_quantity`	Total number of items purchased	int
`total_spend`	Total money spent by the customer	float
`avg_spend`	Average spend per purchase	float
`age`	Age of customer	int
`phone`	Customer contact number	string

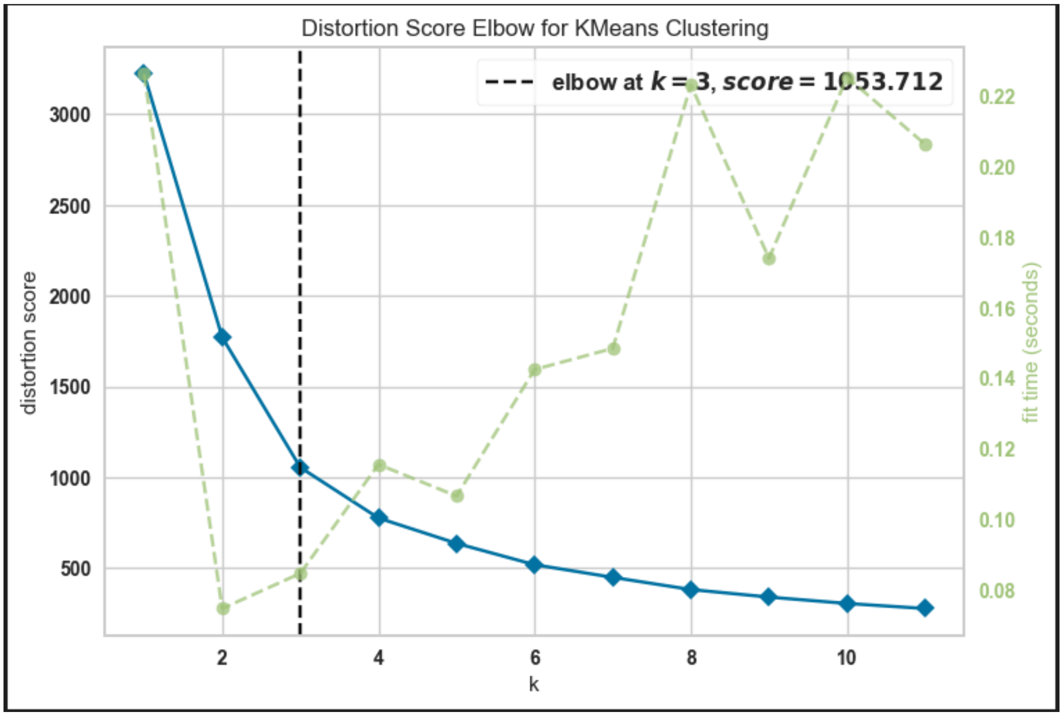
Null values were cleaned, and outliers in continuous fields were removed using the IQR method prior to modeling.

The radar chart below visualizes normalized feature profiles across all clusters.

Visuals

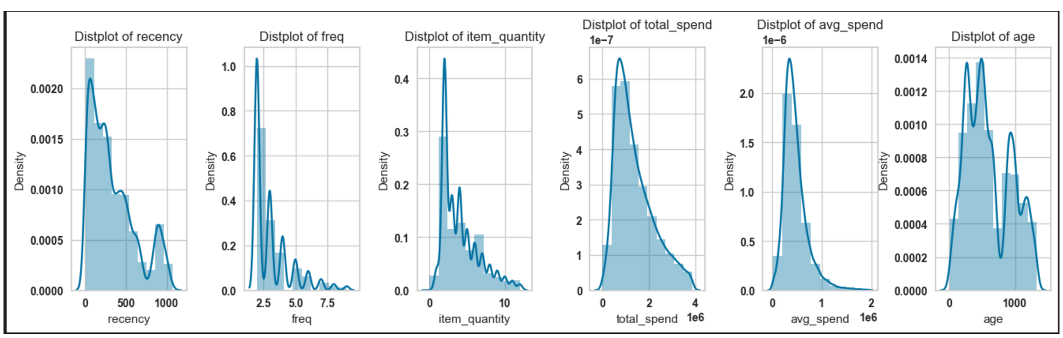
1. Elbow Method to Determine Optimal K

The elbow plot below shows the distortion score for different K values.



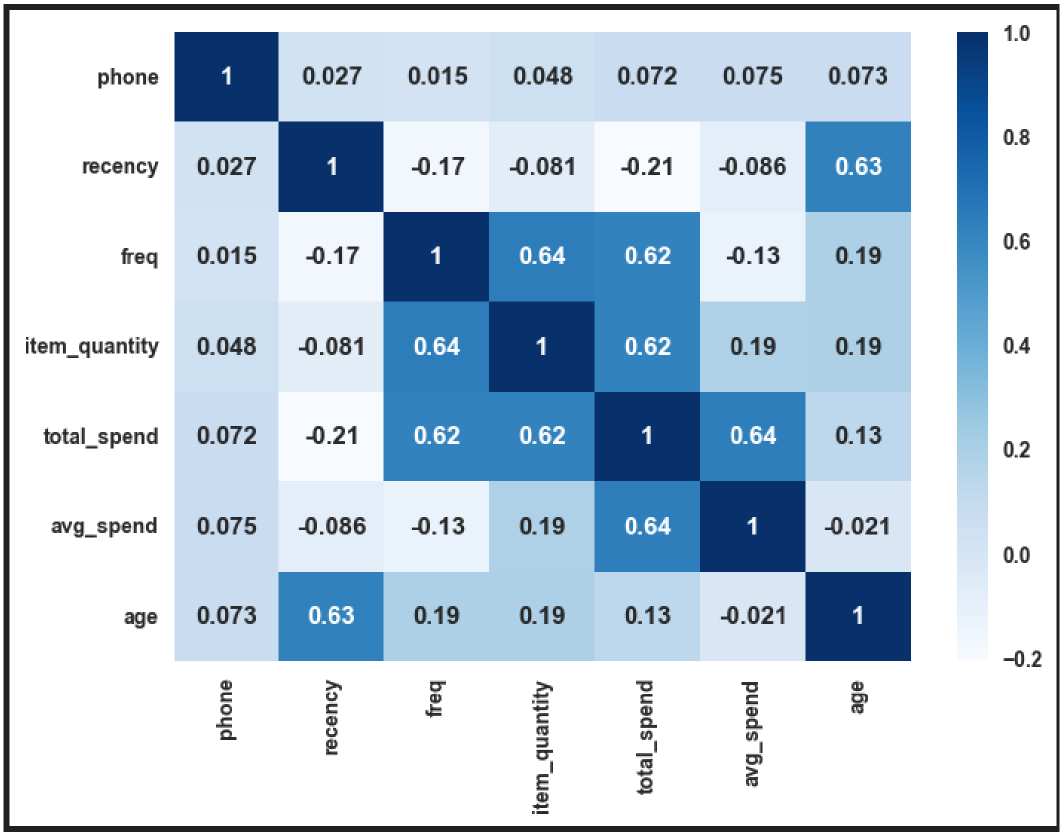
2. Distribution of Key Features

Histograms and KDE plots show the shape of customer attributes before clustering.



3. Correlation Matrix

Feature correlation helps identify relationships and potential redundancy.



phone
recency
freq
item_quantity
total_spend
avg_spend
age
1 0.027 0.015 0.048 0.072 0.075 0.073
0.027 1 -0.17 -0.081 -0.21 -0.086 0.63
0.015 -0.17 1 0.64 0.62 -0.13 0.19
0.048 -0.081 0.64 1 0.62 0.19 0.19
0.072 -0.21 0.62 0.62 1 0.64 0.13
0.075 -0.086 -0.13 0.19 0.64 1 -0.021
0.073 0.63 0.19 0.19 0.13 -0.021 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2