Report of Q6
Bank Customer Clustering

50015627 JIANG Zhuoyang

1. Question Description:

Banks classify or cluster customers for several reasons, aiming to enhance their perational efficiency, tailor services, and manage risks effectively. I was required to offer insights as a data analyst. Furthermore, I was expected to employ at least three distinct clustering algorithms to categorize customers. This report encompass the following two components:

- Explore the data table using visualization techniques and provide at least 10 pictures and at least 10 business insights
- Use at least three different clustering algorithms to cluster customers and explain the common characteristics shared by customers within the same cluster after your clustering, as well as the differences among customers in different clusters. Please provide evidence..

2. Load Data and Do Preprocessing

Load data and do Data Preprocessing as below:

- Handle non-finite values in 'Quantity' column before conversion
- Convert 'Quantity' column to integer type after handling non-finite values
- Fill missing values in 'Description' column with 'Unknown'
- Fill missing values in 'Customer ID' column using forward fill method
- Convert 'Date' column to datetime type
- Remove duplicate records and outliers

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048567 entries, 0 to 1048566
Data columns (total 9 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   TransactionID           1048567 non-null  object
 1   CustomerID              1048567 non-null  object
 2   CustomerDateOfBirth     1045170 non-null  object
 3   CustGender              1047467 non-null  object
 4   CustLocation            1048416 non-null  object
 5   CustAccountBalance      1046198 non-null  float64
 6   TransactionDate         1048567 non-null  object
 7   TransactionTime         1048567 non-null  int64
 8   TransactionAmount (INR) 1048567 non-null  float64
dtypes: float64(2), int64(1), object(6)
memory usage: 72.0+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048567 entries, 0 to 1048566
Data columns (total 9 columns):
 #   Column                  Non-Null Count    Dtype
---  ------                  --------------    -----
 0   TransactionID           1048567 non-null  object
 1   CustomerID              1048567 non-null  object
 2   CustomerDateOfBirth     1048567 non-null  datetime64[ns]
 3   CustGender              1048567 non-null  object
 4   CustLocation            1048567 non-null  object
 5   CustAccountBalance      1048198 non-null  float64
 6   TransactionDate         1048567 non-null  datetime64[ns]
 7   TransactionTime         1048567 non-null  int64
 8   TransactionAmount (INR) 1048567 non-null  float64
dtypes: datetime64[ns](2), float64(2), int64(1), object(4)
memory usage: 72.0+ MB
```

Figure.1 Load data and its information after data preprocessing

3. EDA:

Exploration of Amount Data

(1) Exploration of Amount Data:

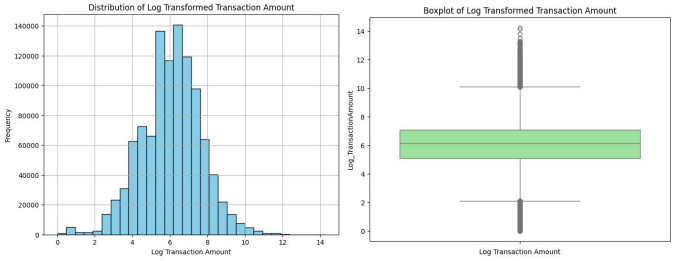


Figure.2 Log Distribution of Amount Data

Insight 1: Due to the long-tail effect observed in the direct distribution analysis, I examined the logarithmic distribution of transaction amounts. When plotted on a logarithmic scale, it demonstrated a normal distribution pattern. From this, we can draw the following analysis:

- The logarithmic transformation brings the data closer to a normal distribution, suggesting increased symmetry and alignment with common statistical assumptions on a logarithmic scale. This situation indicates that using statistical methods based on the normal distribution for analysis and modeling, particularly on "transaction amounts in logarithmic scale," might yield more reliable and clearer results compared to applying these methods directly on the original data.。
- The presentation of a normal distribution after the logarithmic transformation implies that the data largely exhibits characteristics of exponential growth or a long-tail distribution. This suggests the existence of extreme or outlier values that are diminished after the logarithmic transformation, thereby bringing the overall distribution closer to a normal curve.。

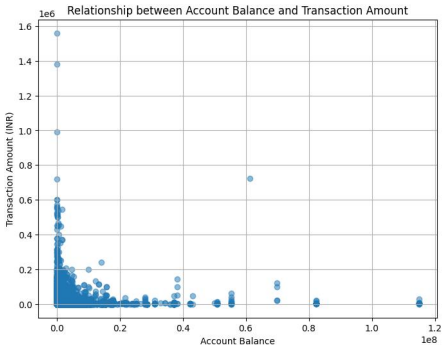


Figure.3 Relationship between Account Balance and Transaction Amount

Insight 2: The clustering of points in the scatter plot, concentrated in the bottom-left corner and spreading along both axes, often indicates a certain constraint or relationship between two variables. In our dataset, lower account balances tend to correspond with lower transaction amounts. This clustering might reflect a pattern in user behavior, suggesting that when account balances are lower, users tend to engage in smaller transactions, while users with higher account balances participate in larger transactions.

(2) Exploration of customer profiles:

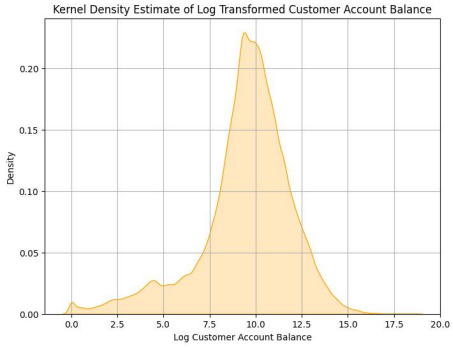


Figure.4 Kernel Density Estimate of Log Transformed Customer Account Balance

Insight 3: Due to the long-tail effect observed in the direct statistical distribution, I analyzed the logarithmic distribution of customer account balances. Similarly to transaction amounts, the logarithmic plot of account balances also displays a Gaussian-like distribution when viewed on a logarithmic scale. The peak density is observed around 22025.47.

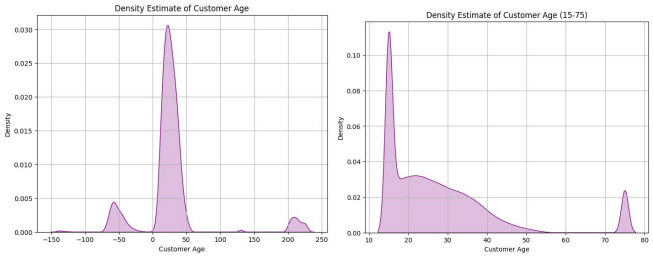


Figure.5 'Density Estimate of Customer Age

Insight 4: Due to the sensitivity and voluntary nature of age disclosure by customers, the derived age values from customer-provided information might lack accuracy. Moreover, the observed age distribution does not consistently align with real societal demographics. Therefore, age should not be considered a significant factor in subsequent analyses.

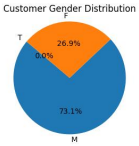


Figure.6 Customer Gender Distribution

Insight 5: From the gender distribution data, it's evident that the customer base is predominantly male. This insight suggests that the bank may need to offer more tailored services to minority gender groups to attract a more diverse customer base.

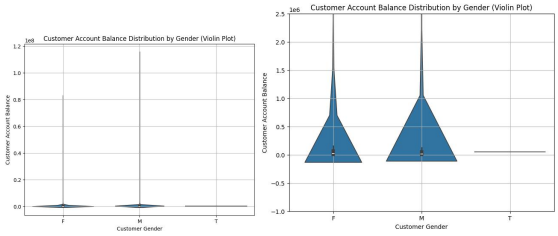


Figure.7 Customer Account Balance Distribution by Gender

Insight 6: Although the effectiveness of the Customer Account Balance Distribution by Gender (Violin Plot) isn't significant, it still reveals that both male and female customers tend to concentrate their deposits within 1,000,000 units. While both genders include some high-value customers, males generally maintain a slightly higher account balance compared to females. Thus, insights from the Account Balance Distribution, as discussed in Insight 6, can be instrumental in devising strategies to attract minority customers based on their deposit patterns.

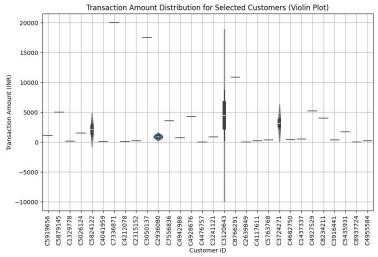


Figure.8 Transaction Amount Distribution for Selected Customers

Insight 7: We can clearly observe varying transaction distributions among customers. Most customers have only one transaction recorded, while those with multiple transactions exhibit distinct transaction amount distributions. Leveraging this transaction distribution information to create features could facilitate personalized customer services.

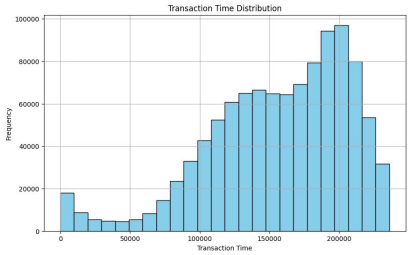


Figure.9 Transaction Time Distribution

Insight 8: About 200,000 transactions are the most frequently occurring, and this information not only helps to understand the current overall market situation, but also aids the bank in optimizing its transaction capacity. It allows for a more rational allocation of resources and targeted personnel arrangements based on the corresponding transaction handling capabilities.

(3) Exploration about Pricing Strategy and Stock Selection Strategy:

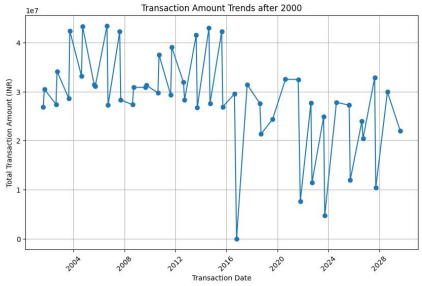


Figure.10 Transaction Amount Trends after 2000

Insight 9: Transactions before 2016 were generally higher in volume compared to those after 2016. There seems to be a cyclical pattern in the fluctuation of transaction volumes, with a trend of rising one year and falling the next (occasionally, there are instances of two consecutive years of increase followed by a decrease, or vice versa). Leveraging this cyclicality in future transaction data could enable the development of strategies for customer service and risk control to adapt to the fluctuations in transaction volumes.

(4) Exploration about the Customers:

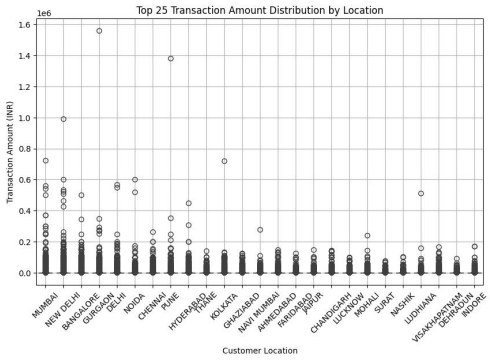


Figure11 'Top 25 Transaction Amount Distribution by Location

Insight 10: From the distribution of transaction amounts across the 25 example locations mentioned, it's evident that certain locations have concentrated transaction amounts (such as SURAT and KOLKATA), while others exhibit a higher dispersion across larger amounts and a substantial overall variance (such as MUMBAI and NEW DELHI). This indicates the potential for devising varied resource allocation strategies and management plans tailored to different transactional locations.

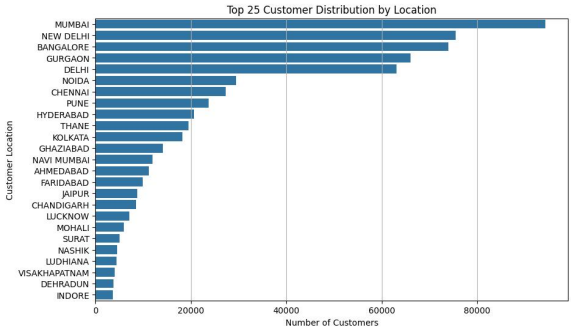


Figure.12 Top 25 Customer Distribution by Location

Insight 11: Based on the customer distribution across the mentioned 25 example locations, clear distinctions can be observed between densely populated customer areas and sparsely populated ones. This allows for the formulation of distinct resource allocation strategies and customer service plans tailored to different transactional locations.

4. Categorize customers - Clustering Analysis

(1) Algorithm:

- KMeans: Clustering algorithm that partitions data into 'k' clusters based on centroids.
- Birch: Hierarchical clustering method for large datasets using a tree-based structure.
- MiniBatchKMeans: Variation of KMeans that processes subsets of data in batches to reduce

(2) Enhanced Feature:

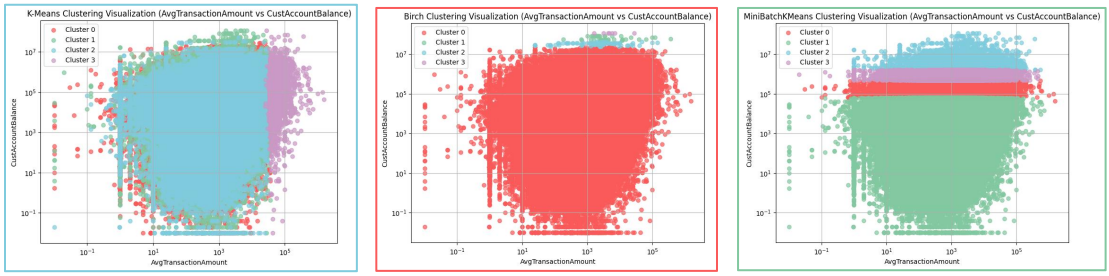
- Calculate transaction counts for each CustomerID
- Compute TransactionAmount-related static features based on CustomerID
- Perform one-hot encoding on CustGender
- Classify CustLocation by transaction counts and Perform one-hot encoding

(3) Experiment:

- **Experiment 1:** KMeans Algorithm Clustering with **FULL Feature Vector**
- **Experiment 2:** Birch Algorithm Clustering with ['Middle 5 Locations', 'Other Locations', 'Top 5 Locations', 'CustAccountBalance']
- **Experiment 3:** MiniBatchKMeans Algorithm Clustering with ['AvgTransactionAmount ', 'CustAccountBalance']

(4) Visualizing the distribution of features across different clusters for comparison.

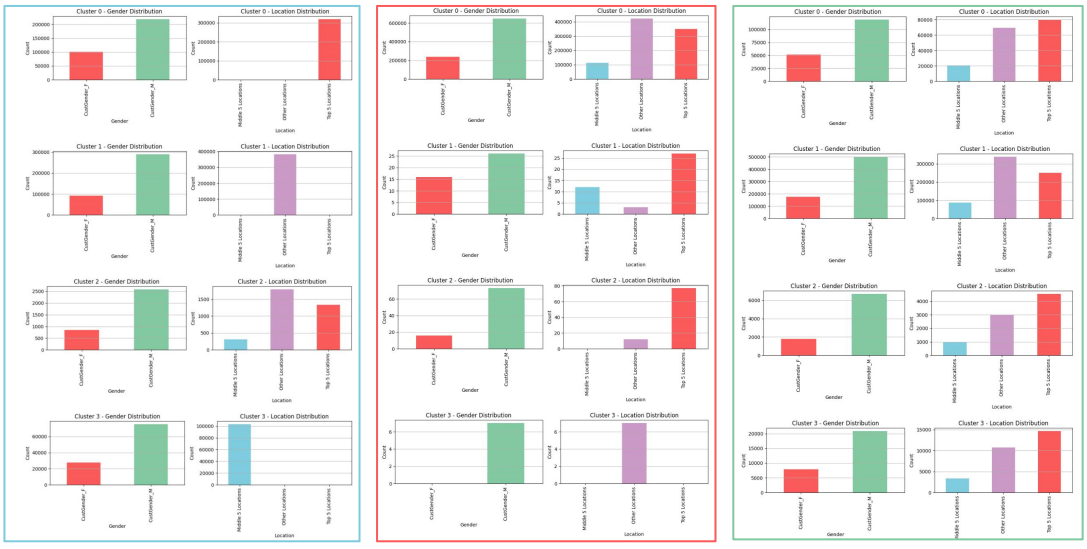
- Visualize Important Static Feature in Different Cluster:



Experiment 1 **Experiment 2** **Experiment 3**

Figure.13 Visualization of AvgTransactionAmount vs. 'CustAccountBalance'

- Visualize One-hot Feature Distribution (Gender and Location) in Different Cluster:



Experiment 1 **Experiment 2** **Experiment 3**

Figure.14 Visualization of the distribution of CustGender and 'CustLocation'

(5) Clustering Analysis Conclusion:

In the three experiments, I categorized samples into four groups using different algorithms and selected distinct features for each. I visualized the clustering for both "statistical continuous features" and "categorical one-hot encoded features," leading to the following conclusions:

- The gender feature didn't significantly contribute to the clustering.
- Location features were meaningful and contributed positively to the clustering.
- Two crucial statistical features also had a noticeable impact on the clustering process.