Exploratory Data Analysis Report on Transaction Data

extensoData

unleash your x factor

Submitted By:

Bishesh Kafle

6th June, 2024

1. Overview

The following report provides a summary of the Exploratory Data Analysis (EDA) performed on a transactional dataset. The dataset contains various transactions made across different branches and products within a specified period. The main objective of the report is to uncover patterns, detect anomalies, and understand the distribution of the data.

The dataset contains 45031 rows and 9 columns. The columns are:

- **tran_date** (DateTime) : The date of transaction.
- **account_number** (string): Account number associated with the transaction.
- **branch** (int): Branch id of the bank's branch.
- **product** (string): The type of account scheme.
- **lcy_amount** (float): The transaction amount is local currency.
- **transaction_code** (string): Type of transaction.
- **description1** (string): The description of the transaction.
- **dc_indicator** (string): Indicates whether the transaction is debit or credit.
- **is_salary** (int): Indicates whether the transaction is associated with salary or not.

Among all the columns, only description1 has few missing values i.e. the associated records are the transactions with no any descriptions.

The dataset contains information of 1838 accounts and the transactions made in those accounts. It incorporates the records from 73 different days. The transactions are spread across 13 different branches of the given bank.

2. Data preprocessing

In the data preprocessing phase, several steps were undertaken to ensure the dataset was clean, consistent, and ready for analysis. The preprocessing steps applied are as follows:

1. The transaction date column, tran_date, was converted from its original format to a datetime format.

2. The dataset was checked for any missing values. Only the description column had few of the missing values that will be dealt with later on during the analysis.

3. Duplicate rows in the dataset were identified. However, since a user can perform the same transaction more than once a day and there is no unique identifier provided for each transaction, the duplicated values are left as it is.

4. A new feature, desc, was created by extracting the first two words from the description1 column. This was achieved by processing the text by removing punctuation, splitting the text into words, and returning the first two words. If the text contains less than two words, it returns the available word or an empty string. Then to standardize the column, all values were converted to lowercase.

3. Descriptive Statistics

In the given dataset, the following columns contain categorical values:

- account_number
- branch
- product
- transaction_code
- dc_indicator
- is_salary

and the numerical values are:

- lcy_amount

The following is an overview of the column **lcy_amount.**

Statistic	**Value**
count	45031
mean	39,810.24
std	227,260.60
min	0.01
25%	944.74
50%	5,000
75%	20,000
max	1,02,80,810

The dataset consists of 45,031 observations with a mean value of 39,810.24 and a high standard deviation of 227,260.60, indicating significant variability in the data. The values range from a minimum of 0.01 to a maximum of 10,280,810. The data distribution is skewed, as evidenced by the large difference between the mean and the median (5,000). Additionally, 25% of the data points are below 944.74 (first quartile), and 75% are below 20,000 (third quartile), highlighting the presence of outliers or extreme values that greatly influence the mean.

4. Data Distribution and Visualization

4.1 Distribution of lcy_amount

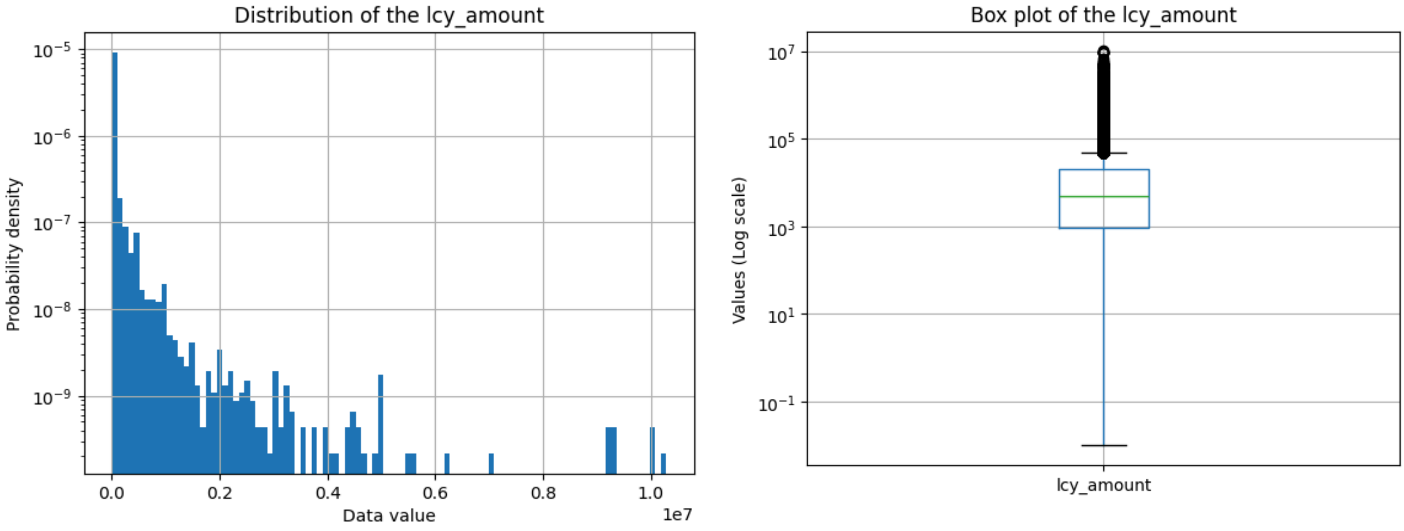

Fig 1: Distribution plot and Box plot of lcy_amount

A normal linear scale would compress most of the data points into a tiny region near the bottom of the graph, making it difficult to see any trends or patterns therefore we use log scaling on the y-axis. Now, we can see that the data is positively skewed.

From section 3 above and fig1, we can confirm that there are presence of outliers, that too only upper tail outliers. Calculating their count, we get the following result:

- Inter Quartile Range Outliers (IQR outliers) : 4741
- Z-score Outliers (Z outliers) : 426

However, since the Z-Score method assumes a normal distribution, and in the case of skewed data, it might be sensitive to the skewness. On the other hand, the IQR method is more robust and less affected by the skewness, making it a better choice for identifying outliers in skewed distributions.

4.2 Transaction by Branch and by Transaction Code

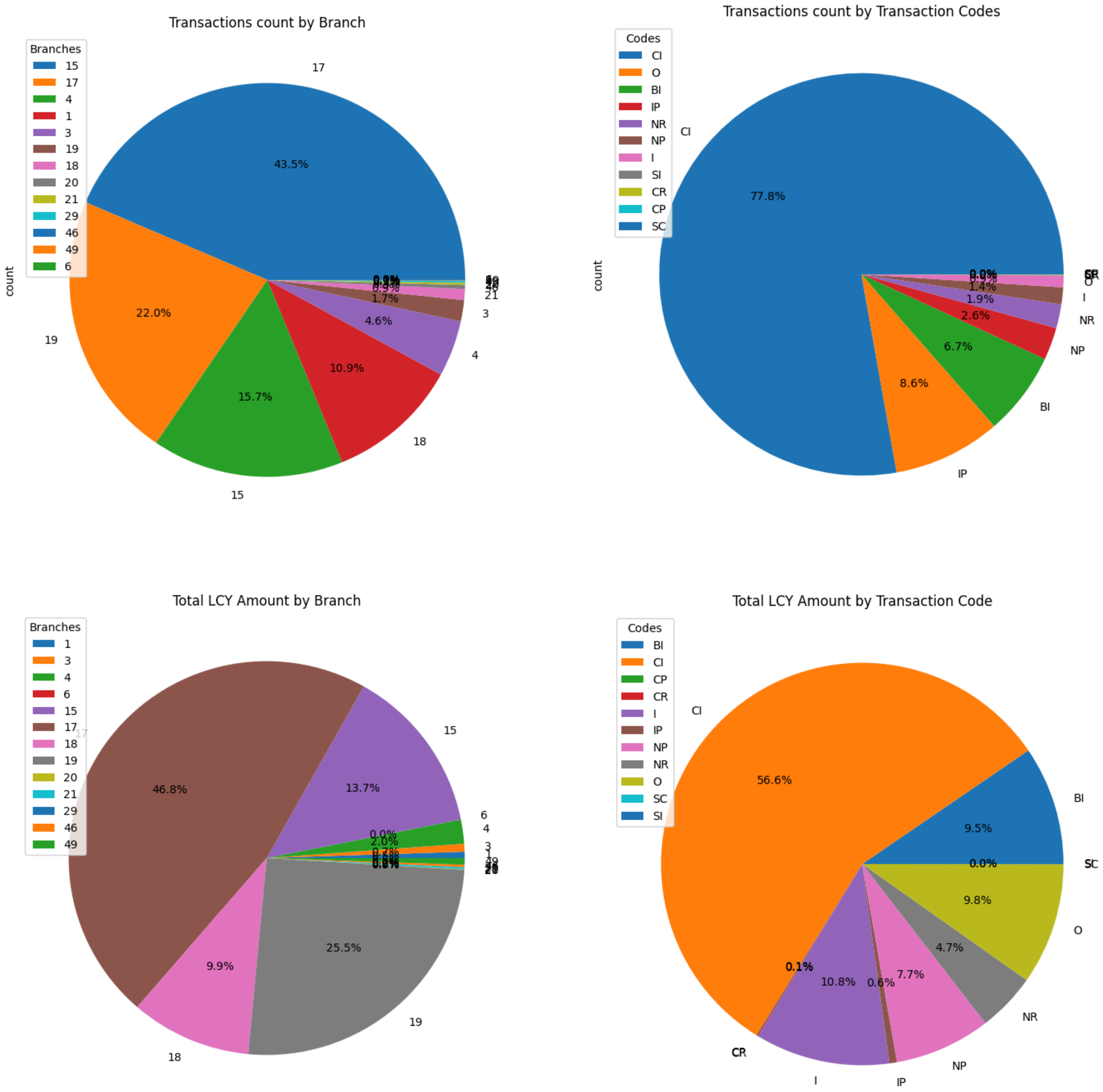

Fig 2: Pie chart of transactions by branch and transaction code

Here we can see the highest number of transactions were at branch 17. Similarly, branch 19, branch 15 and branch 18 also take up a high percentage of total transaction counts suggesting the 4 most active branches. Implied by this claim, we can also see that branch 19 has the highest total transaction amount followed by branches 19,15 and 18.

Again, we can see from the right hand side plot above that the majority of transactions are represented by the code CI. The second-largest portion corresponds to the code IP. Other significant codes include BI and NP. The remaining transaction codes each account for less than 2% of the total transaction counts. CI also has the highest total transaction amount followed by I, O and BI.

Now, plotting the distribution of lcy_amount by branches and transaction code using box plot we get the following plot,

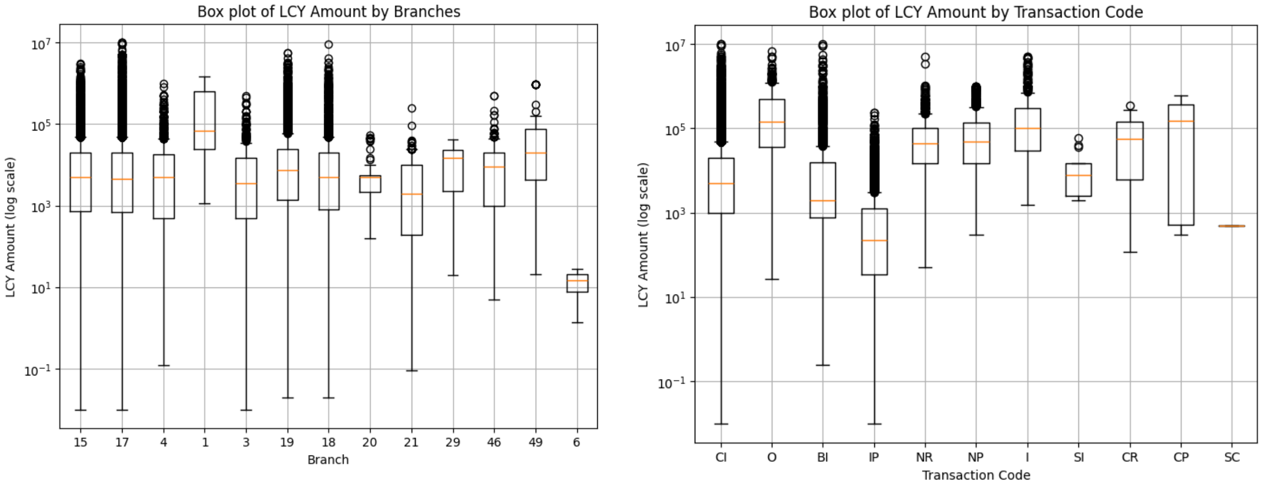

Fig: Box plot of lcy_amount based on branches and transaction codes

Overall, there is a fair amount of variability in the amount of LCY by branch as well as transaction codes. We can see frequent presence of outliers in almost every category. Transaction code of CP and SR seem to have no IQR outliers. SC contains only one value, thus it is trivially given that it has no outliers. CR and SI also have relatively few outliers. Looking from a broad perspective, this might be related to SI, CR, CP and SC having very few transaction counts to express enough variability. Similar case on branches 1,6 and 29. There seem to be no presence of outliers and this might be associated with these branches having very few transaction counts.

4.3 Transaction by dc_indicator

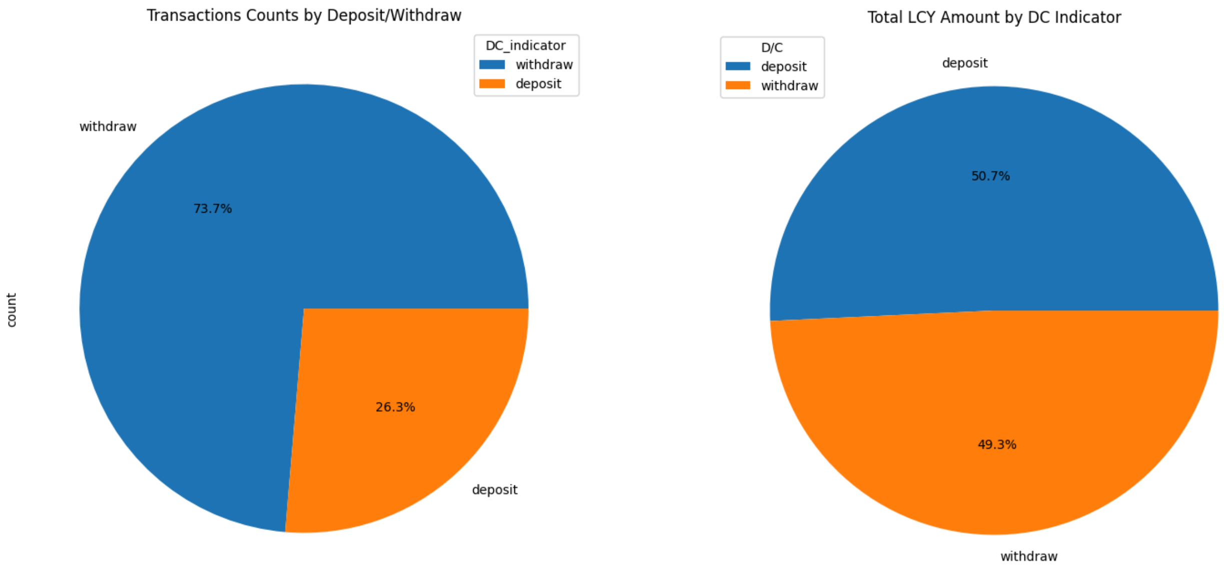

Fig 3: Pie chart of transactions by dc_indicator

From the above pie chart we can see that among all transactions, more than ¾ th of the transactions are withdrawals. However, the total amount deposited is greater than the total amount withdrawn.

Therefore, from this significant difference, it can be concluded that the amounts being deposited must be greater than the amounts being withdrawn from the bank accounts.

5. Findings

5.1 Salary Analysis

Let us first analyze the distribution of overall transaction amounts with salary amounts i.e. the amounts with **is_salary** flag active.

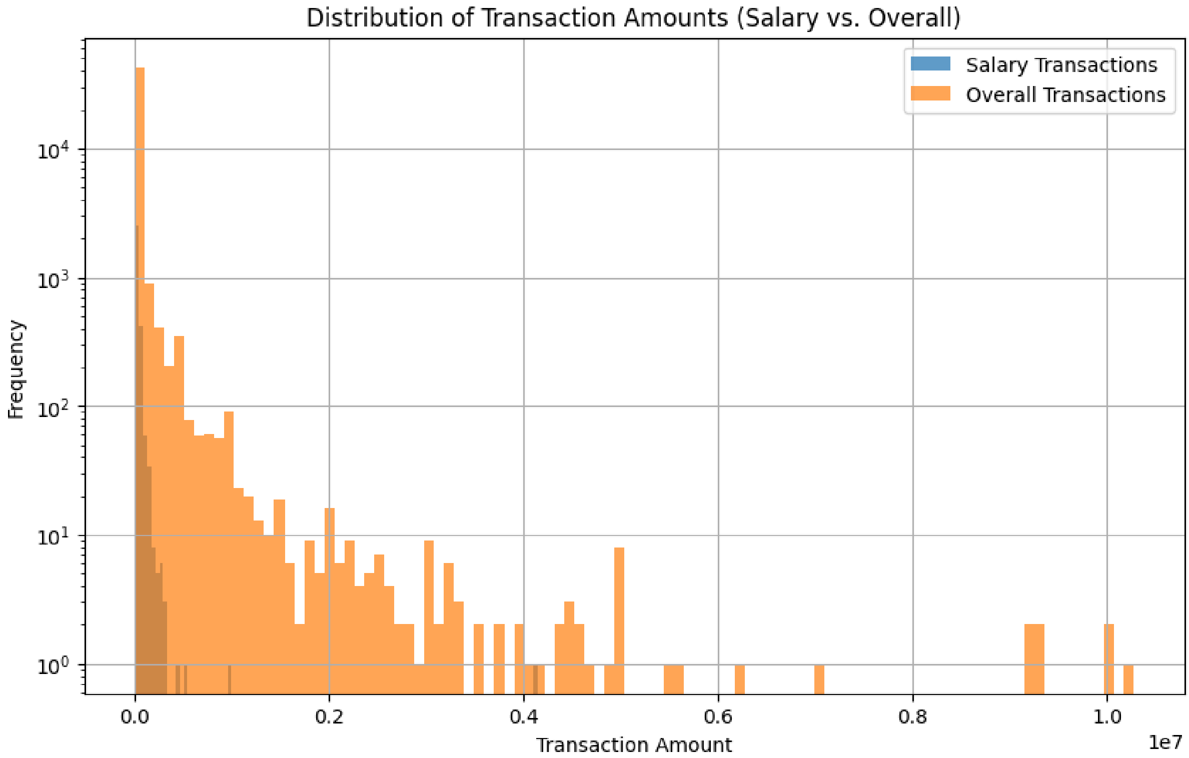

Fig: Distribution of Salary amounts vs total transaction amounts

From the above distribution plot, it can be seen that the salary amounts are also positively skewed. For transaction amounts on the lower side, a high percentage are likely salary deposits. However, as transaction amounts increase, the proportion of salary deposits seems to significantly decrease. This suggests that most salaries fall within a specific range, and larger transactions are more likely to be for other purposes.

All salaries are correctly designated as deposits, as they should be.

Now checking on the transaction codes for given salaries, we get the following count:

Transaction Code	Count
CI	2748

BI	297
NR	13
IP	7
O	4

A significantly large portion of salary transactions lie in the CI category. Here, we can also see that there are 7 records of IP transactions, which likely is the deposit of the interest from a fixed deposit that is almost similar to the regular salary amount being deposited in the respective account.

Now, looking at the **description1** column, we could see that almost 94.95% of transactions that were flagged as salary have the inclusion of words like ‘salary’, ‘sal’ or ‘pension’.

5.2 Relationship between description, dc_indicator and transaction_code

Next, we generate a word cloud from the **desc** column that was engineered during the preprocessing. The word cloud generated is shown in the figure below:

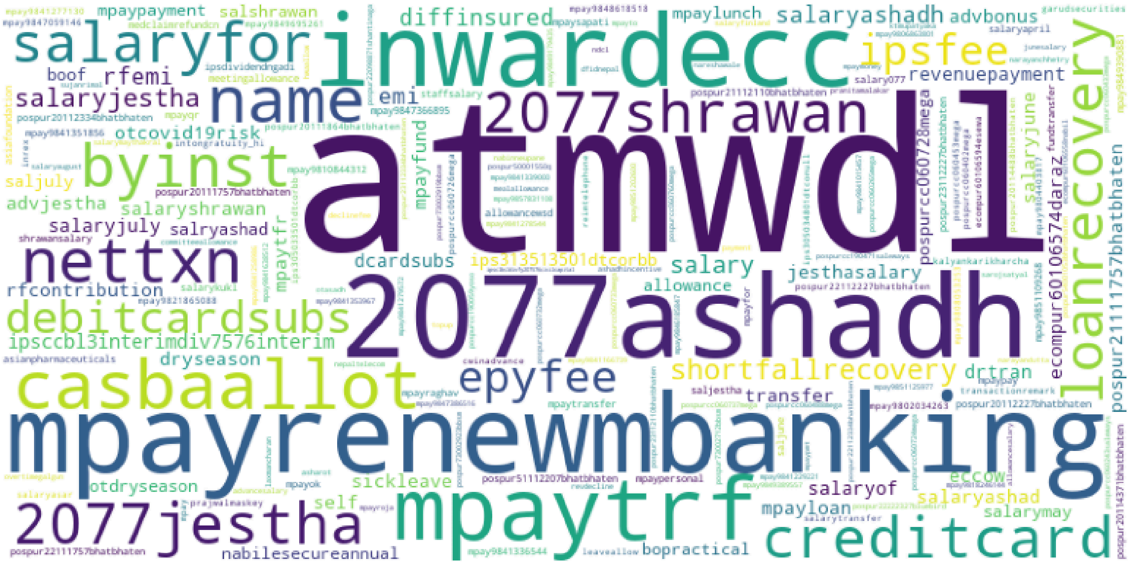

Fig: Word cloud generated from the desc column.

Let us take a few of the highly recurring words from the cloud and see how they compare with **dc_indicator** and **transaction_codes**.

1. **atmwdl**: On checking the values in the dc_indicator column, there are 12589 counts of withdraws. However we could also see 73 deposits associated with

atmwdl. Since this was an unusual finding, manually inspecting few of description1 columns associated with atmwdl and deposit, it was seen that the 73 records were actually due to revert transaction from the atm. Similarly, the transaction code for every ATM withdrawal transaction is CI.

2. **mpayrenewbanking:** All the records are withdrawal transactions and are associated with the transaction code of CI.

3. **inwardecc:** All the records are withdrawal transactions and are associated with the transaction code of I.

4. **creditcard:** All the records are withdrawal transactions and are associated with the transaction code of BI.

5. **loanrecovery:** All the records are withdrawal transactions and are associated with the transaction code of CI.

6. **casbaallot:** All the records are withdrawal transactions and are associated with the transaction code of BI.

7. **mpaytrf:** About 70% of records associated are withdrawal transactions. The transaction code is CI for all.

5.3 Deposit to Withdraw Ratio Analysis

For every account number, deposit to withdraw ratio was calculated on the basis of value as well as count for the given total time frame. There were no accounts with 0 deposits, but there were a few accounts with no withdrawals. Ignoring those accounts in the analysis, the results of the ratio are as follows:

1. **By Value:**

   Based on individual ratio we have the following counts,

   - Ratio > 1 : 924
   - Ratio < 1: 904
   - Ratio = 1 : 0

   On calculating the overall ratio, we get a value of 1.03. This indicates that the amounts being deposited in the accounts exceed the amounts being withdrawn.

2. **By Count:**

   Based on individual ratio we have the following counts,

- Ratio > 1 : 220
- Ratio < 1: 1445
- Ratio = 1 : 0

On calculating the overall ratio, we get a value of 0.35. This indicates that the number of withdrawal transactions are way higher than the number of deposit transactions.

The primary cause of this difference in count and values might be the small magnitude service fees charged by the bank.

5.3.1 Wilcoxon Signed Rank Test

Now, let us perform the Wilcoxon Signed Rank test on the results above to see the tendency of deposit by account holders in their SBA account. We use Wilcoxon signed rank test because it also makes use of magnitude rather than only the signs.

The results of performing Wilcoxon test on deposit to withdraw ratio by value we get,

Null Hypothesis (H_0): There is no significant difference between the deposit-to-withdraw ratio and 1.

Alternative Hypothesis(H_1): deposit-to-withdraw ratio is significantly higher than 1.

- Test statistic : 959992.0
- P-Value: 1.9 x 10^{-8}

Here, since the P-Value is significantly small, we can reject the null hypothesis on the significance level of 1%. Therefore it can be concluded that account holders tend to deposit larger sums in their bank account than the amounts being withdrawn.

5.4 Transaction Buckets vs Branch

Firstly, transaction amounts were categorized into quartiles (buckets) and then a contingency table was created with Bank Branch on rows and buckets as columns.

The table is shown below:

Branches \ Buckets	Q1	Q2	Q3	Q4
1	0	2	2	24
3	240	230	167	148
4	621	539	485	426
6	2	0	0	0
15	1839	1923	1715	1580
17	5199	6046	4017	4345
18	1255	1421	1056	1155
19	1901	2538	2680	2769
20	10	49	11	9
21	147	122	98	39
29	3	0	4	3
46	30	34	48	34
49	11	9	23	22

5.4.1 Chi-Square Test of Independence

Chi-square test of independence was used to determine whether there is a significant association between a bank's branch and transaction buckets. The results of the test are as follows,

Null Hypothesis (H_0): There is no significant association between branches and transaction amount buckets (quartiles).

Alternative Hypothesis(H_1): There is a significant association between branches and transaction amount buckets (quartiles) i.e. the distribution of transaction amounts is different across at least two branches.

The results of the test are as follows:

- Chi-Square statistic : 705.8012884551438
- P-Value : 3.29 x 10^{-125}

Here, since the P-Value is significantly small, we can reject the null hypothesis on the significance level of 1%. Therefore it can be concluded that there is association between branches and the transaction buckets.

5.5 Time series Analysis

Let us plot the **lcy_amount** column from the initial date to the last date. The result of the plot is shown below:

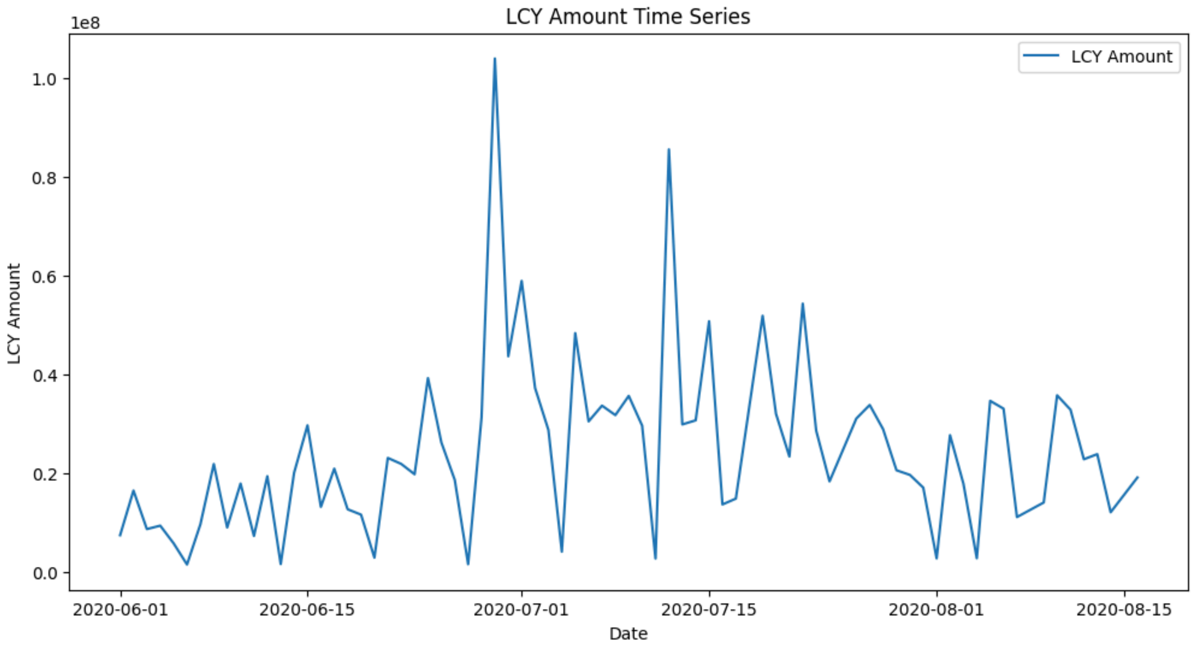

Fig: Time series plot of lcy_amount

This plot can be further explained by the time series decomposition. However, to choose the model as additive or multiplicative let us look at the rolling standard deviation plot.

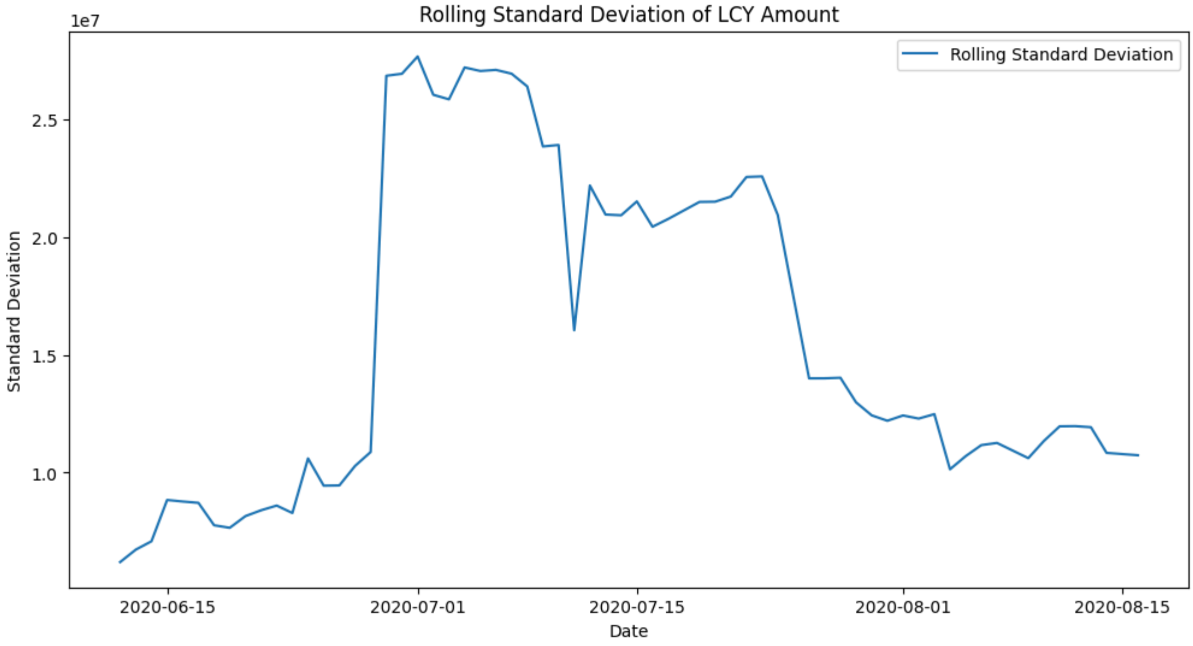

Fig: Rolling Standard Deviation Plot for lcy_amount

Here, the plot shows an increasing or decreasing trend (non-constant variance). Therefore we opt for the multiplicative model. Now decomposing the time series data, we get the following result.

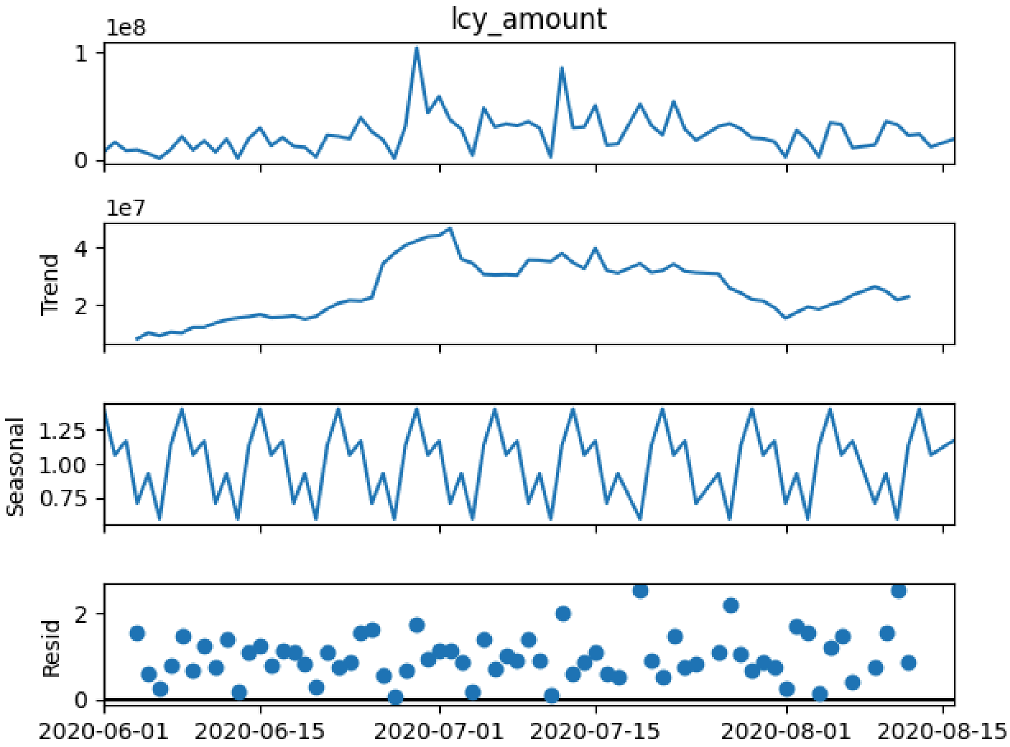

Fig: Time series decomposition of lcy_amount

From the above plot, we can see that there is a noticeable increase in lcy_amount up to a peak, followed by a slight decline and then some fluctuations. This plot also shows the periodic fluctuations that repeat at regular weekly intervals. The values oscillate around a fixed level, indicating the presence of a weekly pattern in the transaction amounts. The residual component represents the remaining variability in the data after removing the trend and seasonal components.

5.6 Account Segmentation

In this study, accounts were segmented based on two key metrics: total transaction counts and deposit-to-withdraw ratio. These metrics were chosen to provide insights into the frequency of transactions and the behavior of funds moving into and out of accounts.

To identify distinct segments of accounts based on their transaction behavior, K-means clustering was applied. The data was standardized to ensure all features contribute equally to the clustering process.

Initially, an elbow plot was used to determine the optimal number of clusters. The result was as follows:

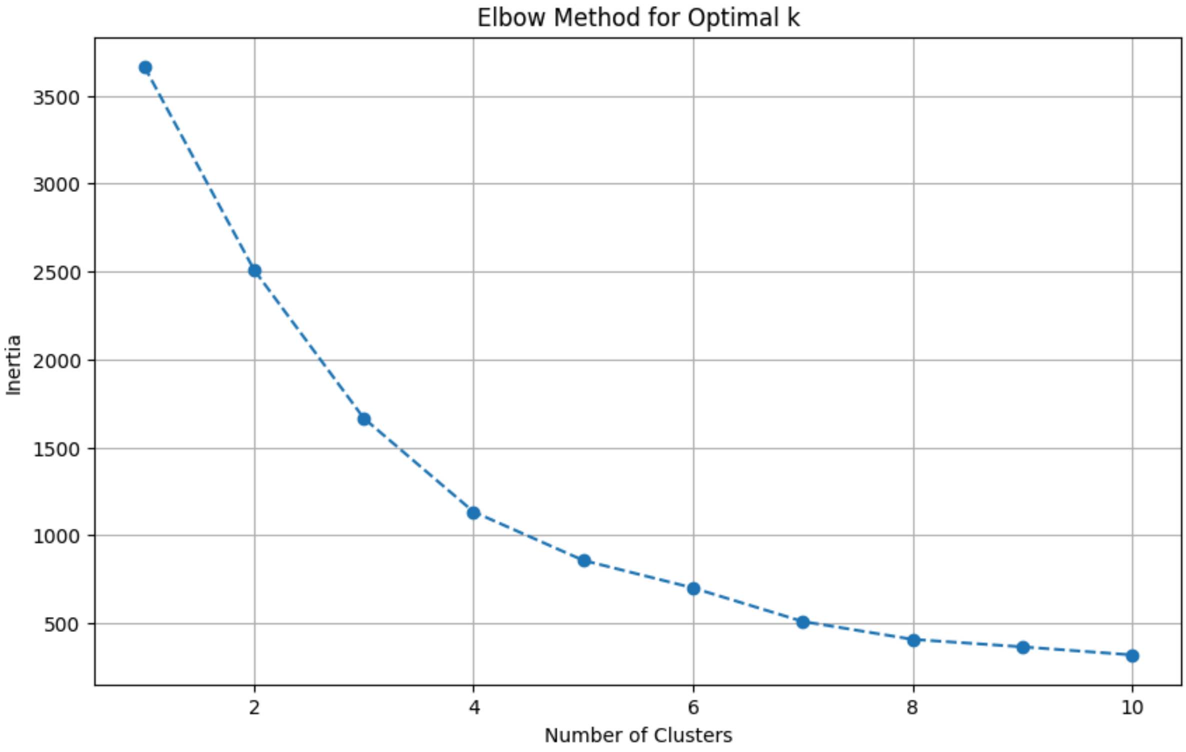

Fig: Elbow plot for account segmentation

However, as there was no distinct elbow observed, the silhouette coefficient method was employed. The silhouette coefficient plot is shown below.

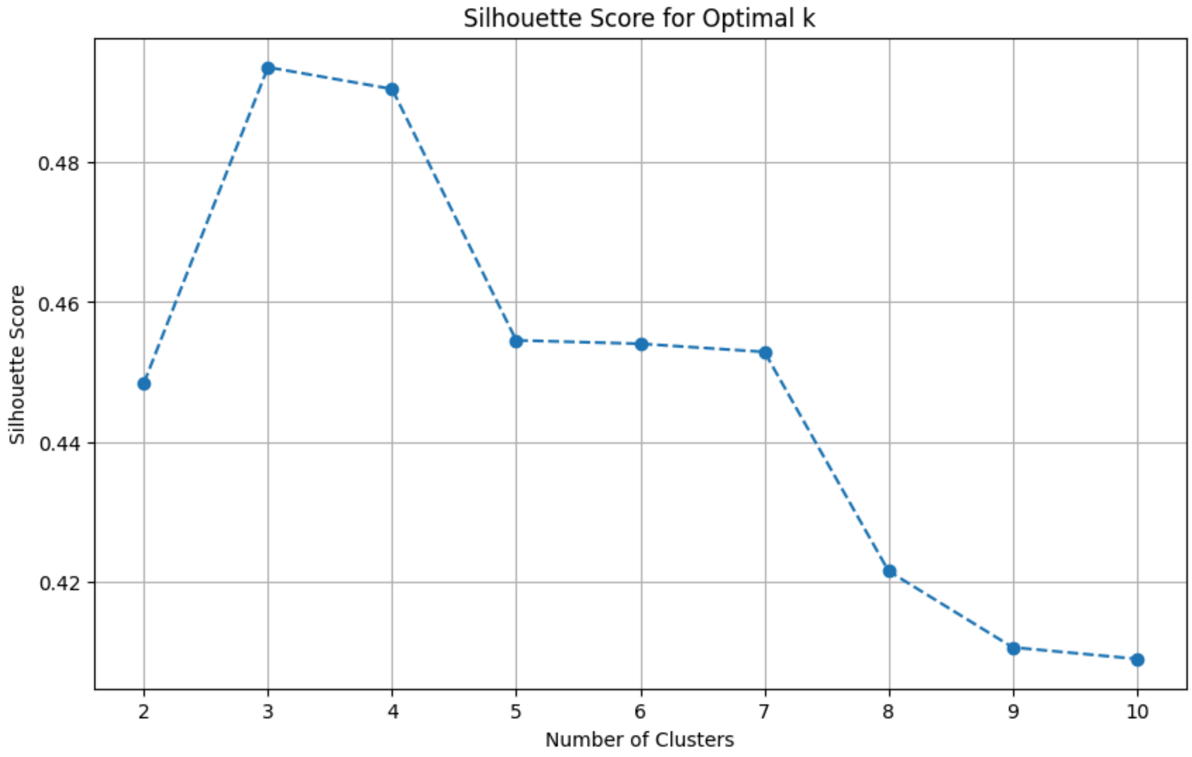

Fig: Fig: Elbow plot for account segmentation

The result of clustering is shown below.

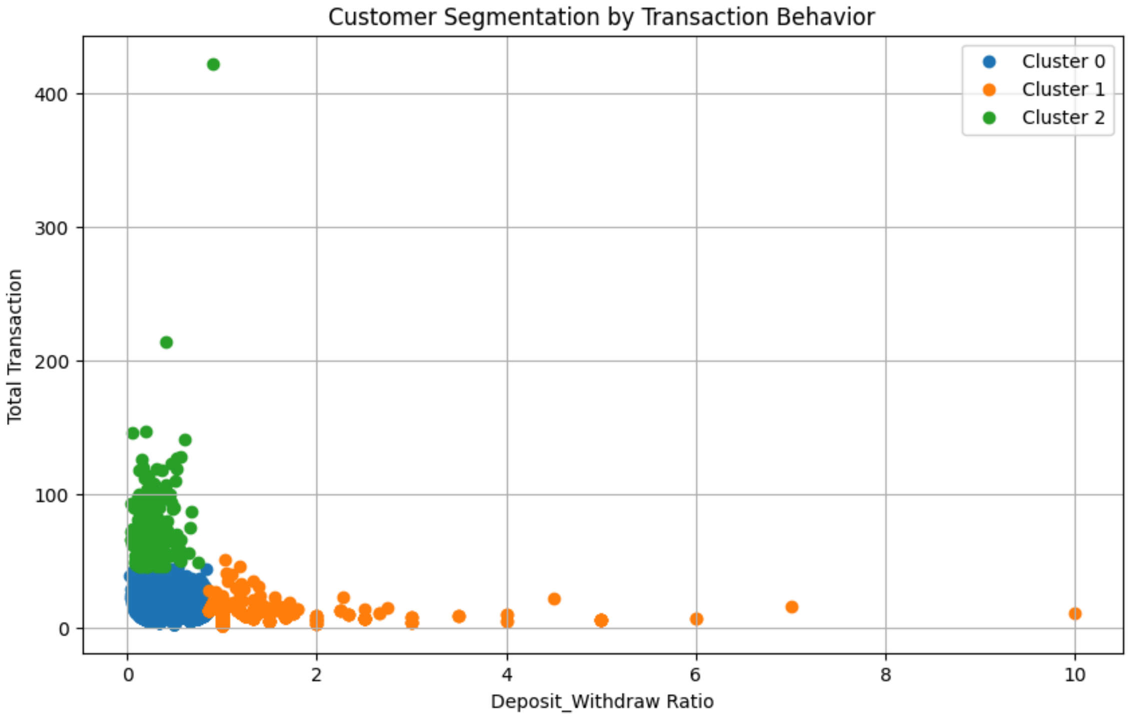

Fig: Clusters of accounts

However since this plot was difficult to interpret visually due to large variations in deposit to withdraw ratio as well as total transactions we perform log scaling on the x and y axis. The resulting figure is shown below.

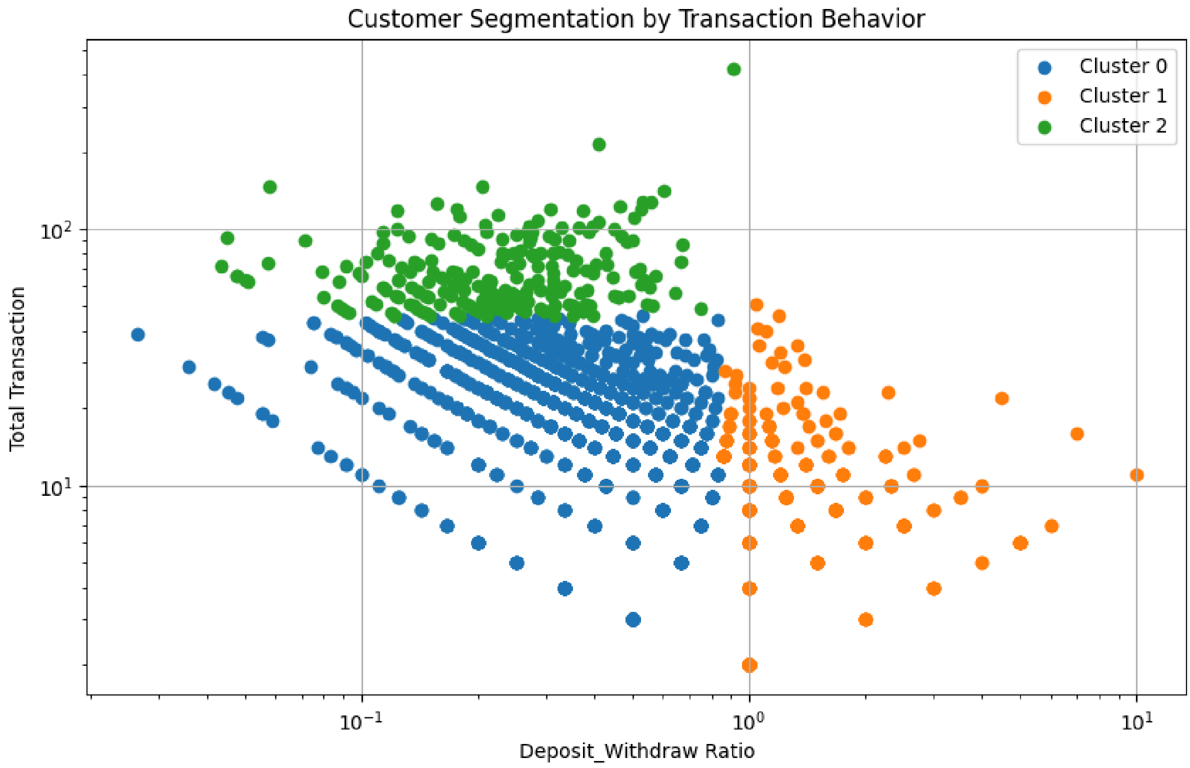

Fig: Clusters of accounts (log scaled)

Here the accounts are classified into 3 clusters:

- **Cluster 0 :** Customers in this cluster have a low Deposit_Withdraw Ratio and a relatively low number of total transactions. These customers might have a balanced transaction behavior with a slight inclination towards withdrawals, and they generally perform fewer transactions overall.

- **Cluster 1 :** Customers in this cluster have a wide range of Deposit_Withdraw Ratios, including some very high ratios, but they tend to have a lower number of total transactions. These customers might include those who either deposit a lot relative to their withdrawals or have very high deposit activity but do not transact frequently.

- **Cluster 2 :** Customers in this cluster have a low to moderate Deposit_Withdraw Ratio but a high number of total transactions. These customers are very active in terms of transaction volume. They are likely to be regular users of the service.

6. Conclusion

Exploratory Data Analysis on this transactional dataset brought several insights related to patterns, trends, and anomalies. The number of transactions in the dataset is 45,031, but there is excellent skewness toward lower amounts of transactions. This leads to a high number of transactions observed at lower amounts. This skewness has brought the necessity of log scaling for more precise visualization and analysis.

The analysis has also shown which branches are more active regarding transaction volume and value. The decomposition of the transaction amounts in the time series disclosed a clear pattern of weekly-based periodic fluctuation i.e. steady weekly cycles in transaction behavior. K-means clustering was employed to identify three groups of customers based on transaction counts and deposit-to-withdrawal ratios. The Chi-Square test of independence has also confirmed that there is a significant association existing between the bank branches and the transaction amount buckets, hence meaning that the transaction behaviors are different between bank branches.