# Assignment 1 - EDA Report

Banking Transactions & Fraud Detection

## Summary

From the dataset, we found a few main findings that call for next steps:

- Exploring the transactions with >1 LoginAttempts
- Exploring the accounts with > 9 transactions
- Exploring the accounts with > $1500 across all age groups
- Exploring the Devices and IP Address with > 9 accounts
- Exploring the transactions with > 50% of Transaction Amount is Account Balance
- Explore the transactions tagged “Potential Frauds” from K-means Clustering

## Data Overview

We analyzed a dataset of 2512 transaction records and 16 features, with no missing values. From the initial descriptive statistics, we have a few findings:

- We see ~75% of TransactionAmounts <$500, however we have a max of $1919. This could be someone doing a large transaction, however an outlier to keep in mind.
- Similarly, ~75% of TransactionDuration <161, however we have a max of 300.
- Similarly, ~75% of transactions have 1 LoginAttempts, however we have a max of 5

Table below.

```
Missing Values:
 TransactionID             0
AccountID                  0
TransactionAmount          0
TransactionDate            0
TransactionType            0
Location                   0
DeviceID                   0
IP Address                 0
MerchantID                 0
Channel                    0
CustomerAge                0
CustomerOccupation         0
TransactionDuration        0
LoginAttempts              0
AccountBalance             0
PreviousTransactionDate    0
dtype: int64
```

```
Descriptive Statistics:
       TransactionAmount                TransactionDate  CustomerAge  \
count        2512.000000                           2512  2512.000000
mean          297.593778  2023-07-05 20:32:10.826433024    44.673965
min             0.260000            2023-01-02 16:00:06    18.000000
25%            81.885000  2023-04-03 16:22:05.750000128    27.000000
50%           211.140000     2023-07-07 17:49:43.500000    45.000000
75%           414.527500     2023-10-06 18:40:53.500000    59.000000
max          1919.110000            2024-01-01 18:21:50    80.000000
std           291.946243                            NaN    17.792198

       TransactionDuration  LoginAttempts  AccountBalance  \
count          2512.000000    2512.000000     2512.000000
mean            119.643312       1.124602     5114.302966
min              10.000000       1.000000      101.250000
25%              63.000000       1.000000     1504.370000
50%             112.500000       1.000000     4735.510000
75%             161.000000       1.000000     7678.820000
max             300.000000       5.000000    14977.990000
std              69.963757       0.602662     3900.942499

             PreviousTransactionDate
count                           2512
mean   2024-11-04 08:09:22.219745024
min              2024-11-04 08:06:23
25%              2024-11-04 08:07:53
50%              2024-11-04 08:09:22
75%    2024-11-04 08:10:53.249999872
max              2024-11-04 08:12:23
std                              NaN
```

| | |
|---|---|
| TransactionID | 2512 |
| AccountID | 495 |
| TransactionAmount | 2455 |
| TransactionDate | 2512 |
| TransactionType | 2 |
| Location | 43 |

We also have expected uniqueness across Transaction IDs. We see 43 locations, and 100 Merchants. We have 681 Devices, while having 592 IP Addresses meaning we have some IPs with multiple devices. We also have 2510 AccountBalance, meaning 2 accounts have the same value. Interesting.

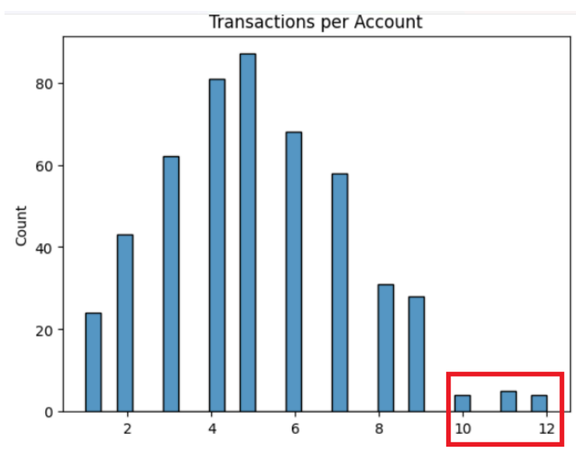
When looking at the distribution of Debit vs Credit, we see a strong preference for Debit (1944 vs 568). However, the spread between Channels (Branch, ATM, Online) is even.

| TransactionType | count |
|---|---|
| Debit | 1944 |
| Credit | 568 |

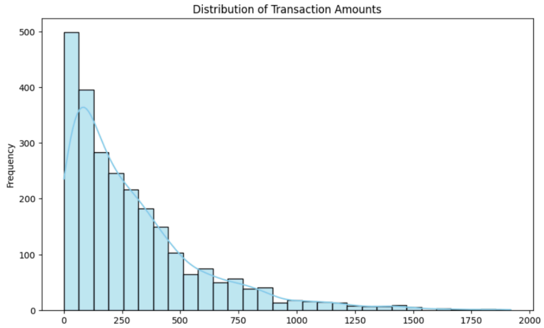
| Channel | count |
|---|---|
| Branch | 868 |
| ATM | 833 |
| Online | 811 |

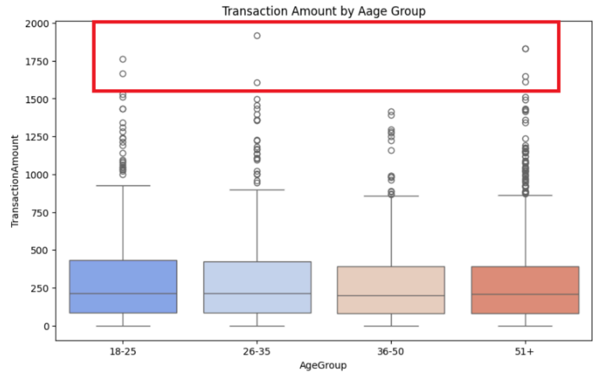
## Business Oriented

When we look at the Transactions per Account, we saw some outliers of a few accounts making > 9 transactions.

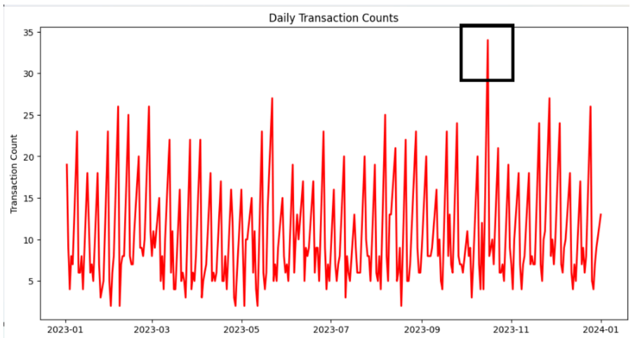


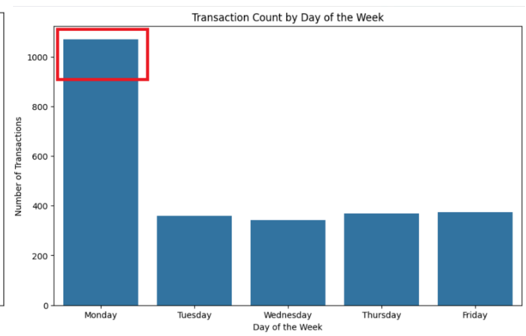

When looking at the distribution of amounts and ages, we see very few accounts with > $1500. Especially in the lower age groups.



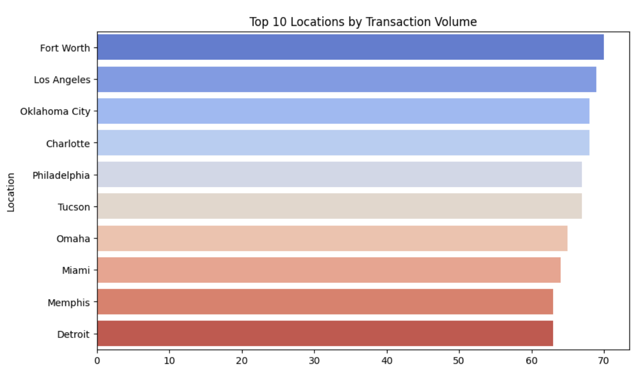



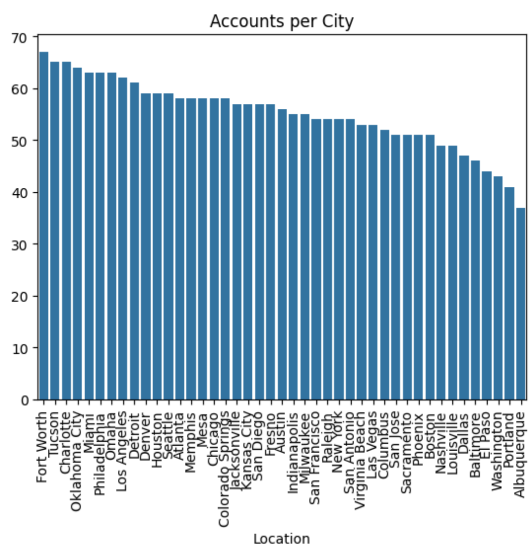

Looking at the data across time (days/yearly), we see a spike of transactions usually on Mondays. We should confirm if these are not outliers. Also there is a day in Late October with a very high and unusual amount of transactions. We should confirm this is not due to holiday traffic and expected.



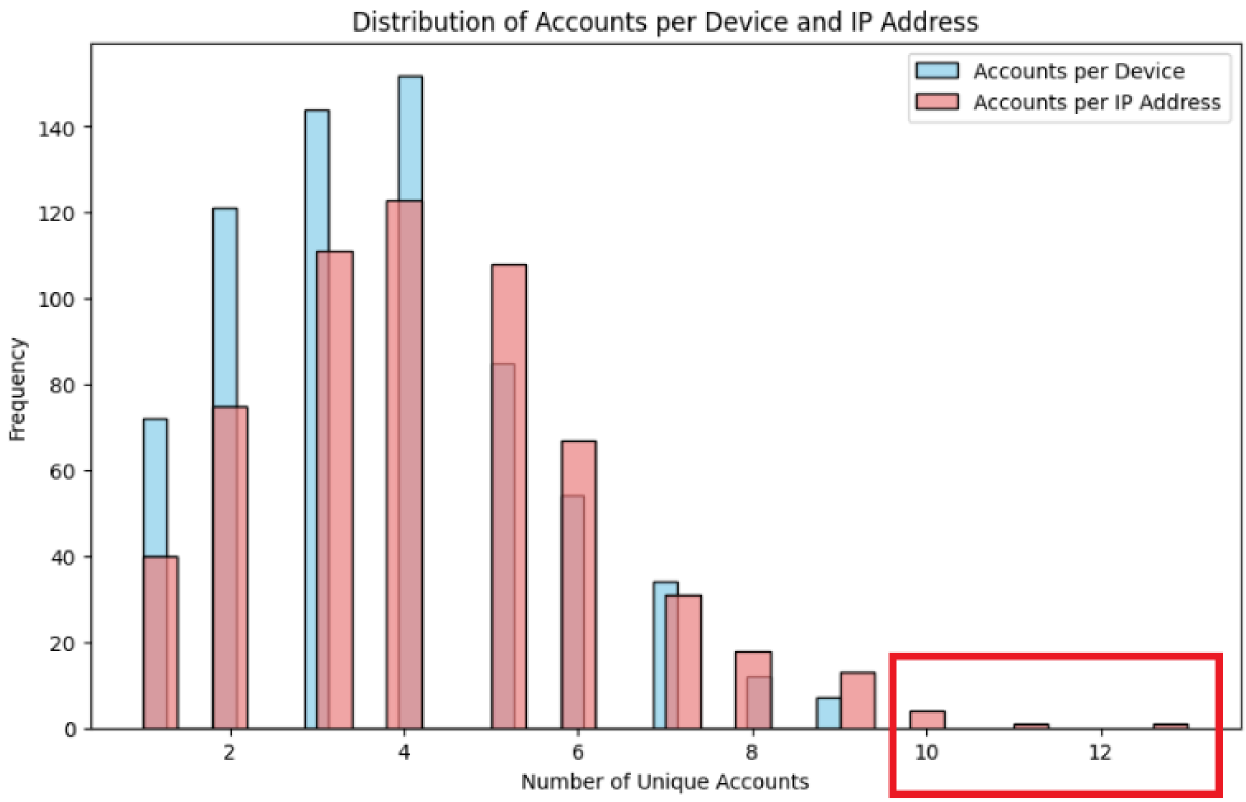




When looking at the top cities and locations, we see the same cities when looking at transaction volumes and # of accounts. There is some spread but overall Fort Worth, LA, Charlotte, OKC are in the top 10 for both.

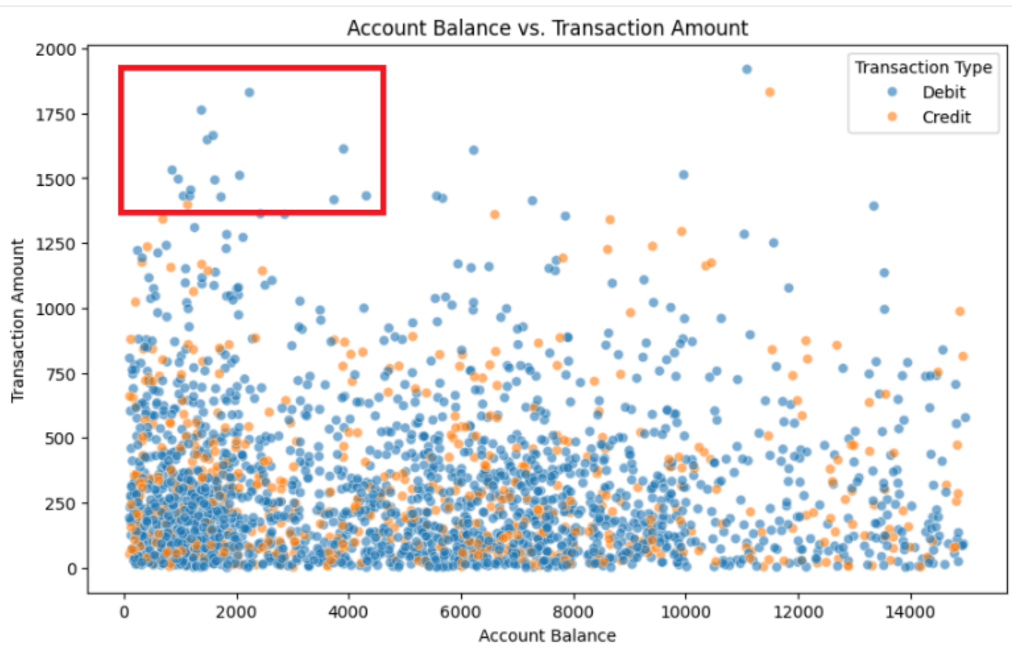


When looking at the distribution of Devices and IP Addresses, we see outliers with a few accounts 10 or greater.

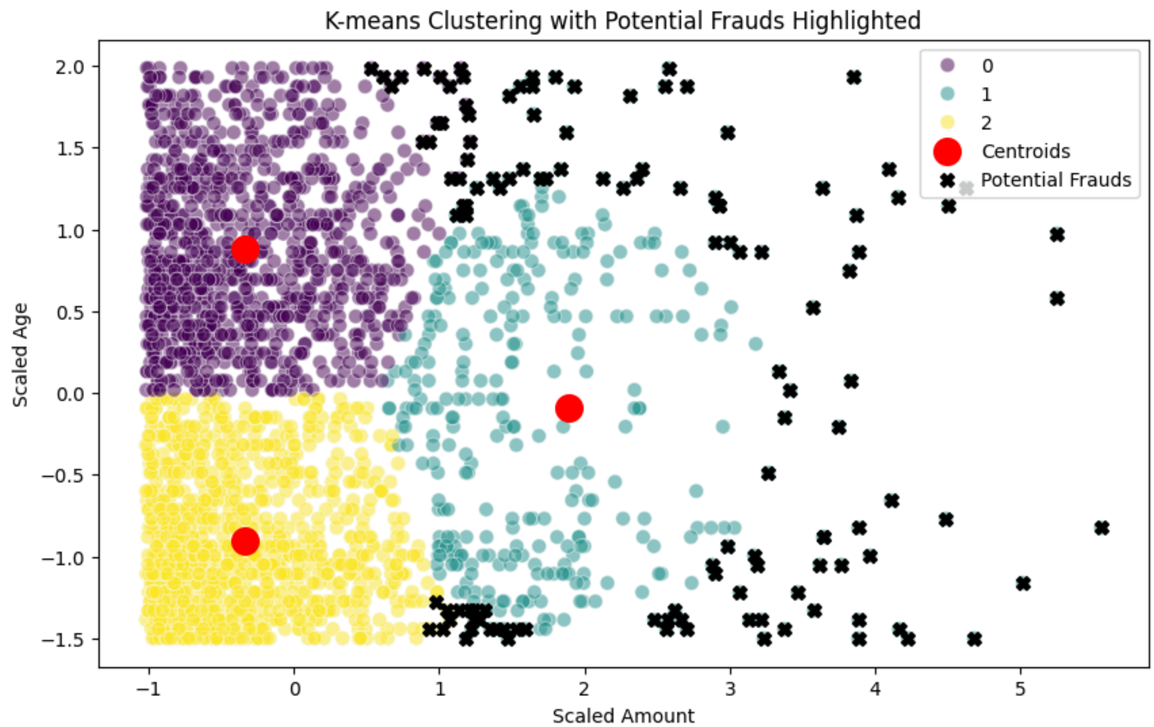



If we compare the Account Balance vs Transaction Amount, we see a few transactions with > 50% of the Account Balance is transacted away. These could be cases to explore unless they are moving banking accounts or companies.



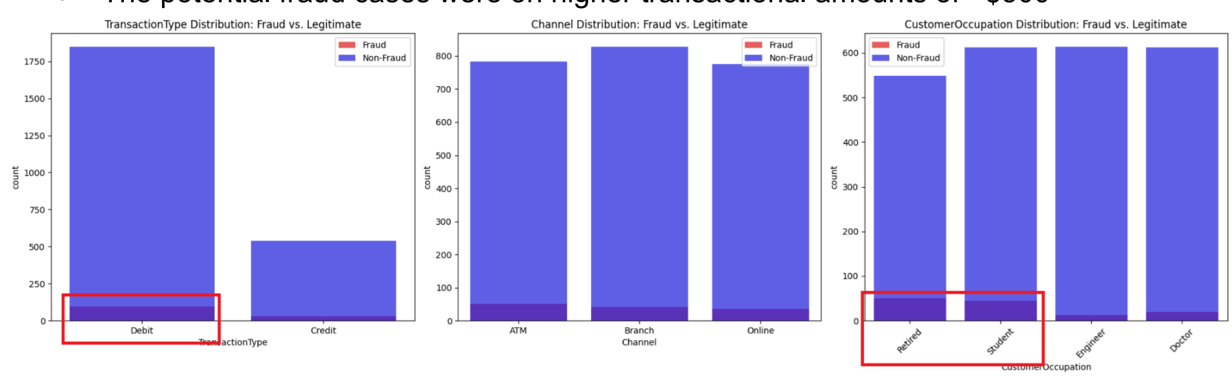
# K-means Clustering

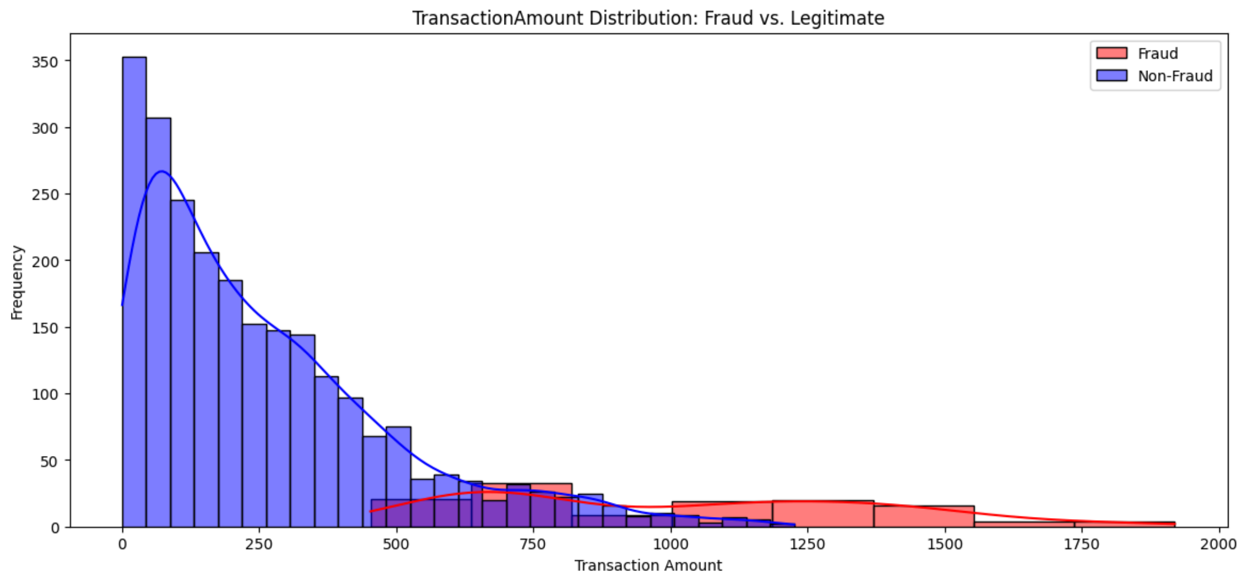
Finally, I leverage a k-means clustering from Kaggle to group transactions by similarity and identify transactions with unique attributes. From this we identified 126 potential fraudulent cases.

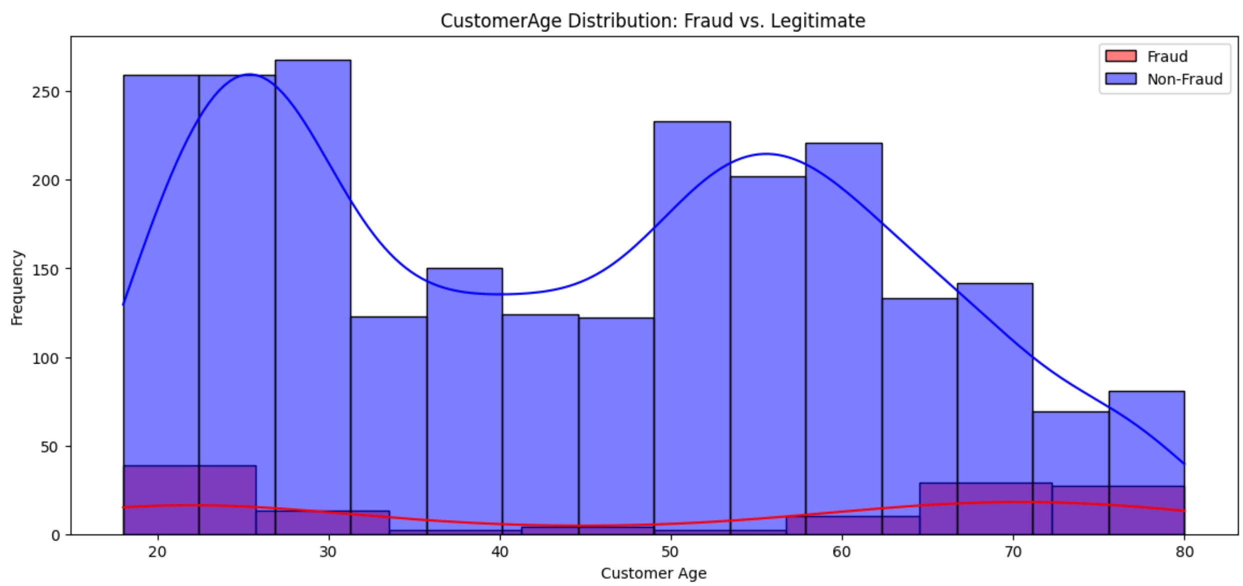



When comparing the statistics of the “potential fraud” vs “not fraud” tagged cases. We found:

- The potential fraud cases were seen more often in Debit cases
- Retired and Student people were more impacted. The customer age also supported this for most cases around 20s and 60+ (this intuitively makes sense as well)
- The potential fraud cases were on higher transactional amounts of >$500




TransactionAmount Distribution: Fraud vs. Legitimate
Fraud
Non-Fraud
Frequency
Transaction Amount
0
50
100
150
200
250
300
350
0
250
500
750
1000
1250
1500
1750
2000


CustomerAge Distribution: Fraud vs. Legitimate
Fraud
Non-Fraud
Frequency
Customer Age
0
50
100
150
200
250
20
30
40
50
60
70
80