Exploring Narration for Better Understanding & Insights

Dataset

The dataset shows the Masked Accounts Transactions recorded from `2020-06-08` to `2022-05-31`. There are 10 columns and a total of 7000 records. We extracted most information from narrations column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7000 entries, 0 to 6999
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Masked_Account_Number  7000 non-null   int64
 1   SEX                    5422 non-null   object
 2   NATIONALITY            6969 non-null   object
 3   DATE_OF_BIRTH          5422 non-null   object
 4   BRANCH_CODE            7000 non-null   int64
 5   TRN_DT                 7000 non-null   object
 6   TRN_CODE               7000 non-null   object
 7   LCY_AMOUNT             7000 non-null   float64
 8   CCY                    7000 non-null   object
 9   Masked_Narration       7000 non-null   object
dtypes: float64(1), int64(2), object(7)
memory usage: 547.0+ KB
```

Task 1: Masking Account Numbers

Generated masked account numbers using random function and mapped them onto real account numbers both in Account number column and the Narration column.

Task 2: Extracting Information from Narration

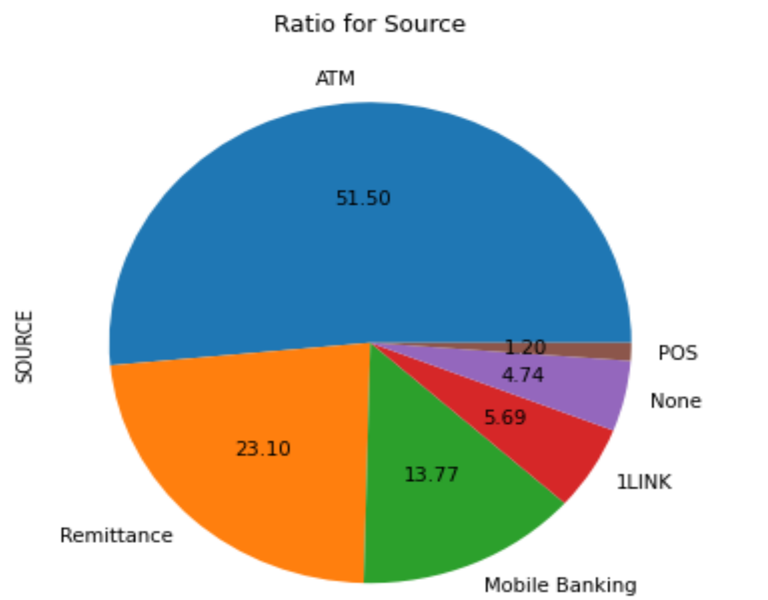
Extracted the following from narration and added it into the dataset as separate columns.

- Account Title
- Bank Name
- Type of Transaction
  - Cash Withdrawal
  - ATM Charges
  - Bill Payments
  - IBFT
  - Purchase
- Source of Transaction
  - ATM
  - 1-Link
  - Mobile Banking
  - POS
  - Remittance
  - None (The ones where nothing could be identified)
- Money Flow
  - Credit or Debit

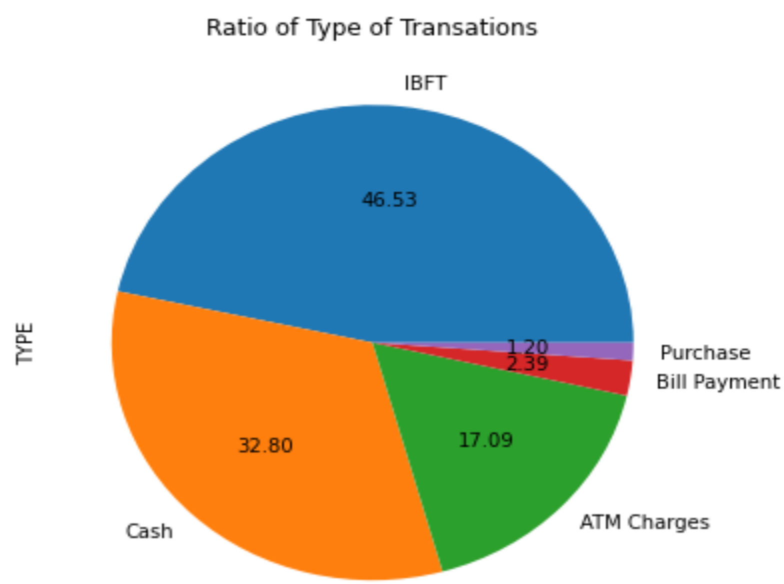
Task 3: Visualizing these Results

Summary

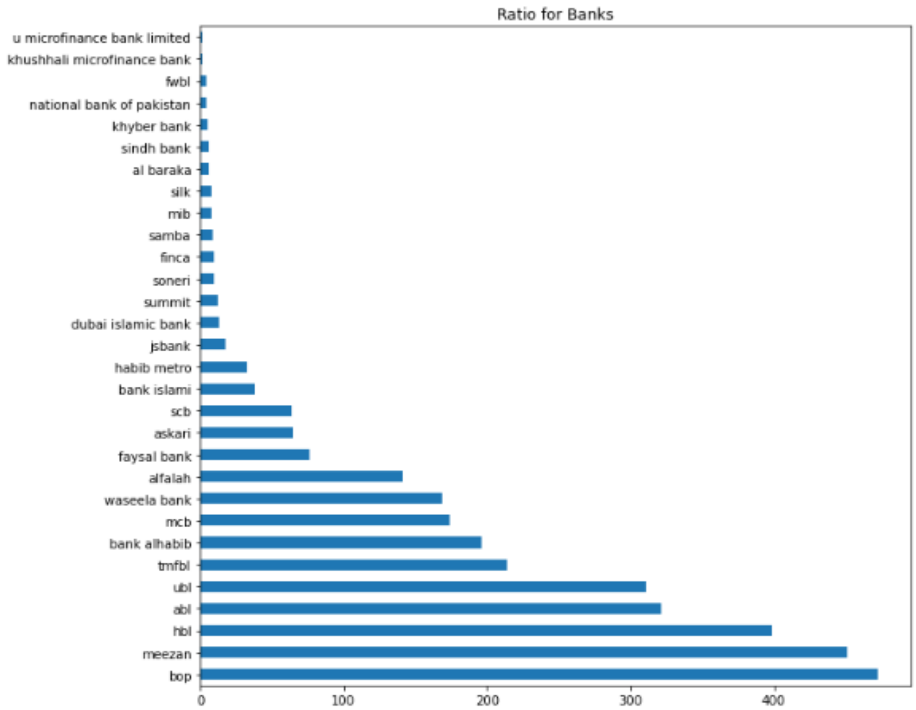
```
Money Flow Summary ['Credit' 'Debit'] [ 821 6179]
Source Summary ['1LINK' 'ATM' 'Mobile Banking' 'None' 'POS' 'Remittance'] [ 398 3605  964  332   84 1617]
Type Summary ['ATM Charges' 'Bill Payment' 'Cash' 'IBFT' 'Purchase'] [1196  167 2296 3257   84]
Nationality Summary ['PK' 'UK' 'nan'] [5475 1494   31]
```



As we can see from the pie chart above ATM is still the most popular source of transaction



As we can see from the pie chart above most of the transaction were IBFT or cash withdrawals.



Most of our customers had made transactions in BOP, Meezan, HBL, ABL and UBL, compared to the rest.

Task 4: Creating a Graph to better understand relation between Accounts

The network created gives us a better understanding of:

- The frequency of transactions
- The weight of transactions
- The direction of transactions
- The connectivity level between all accounts

The approach used to show visualization of funds transfer in our dataset is as follows:

. Accounts are represented by nodes

- Masked account nodes are blue
- Narration account nodes are yellow
- Intersecting account nodes are white(larger in size so easily spotted)

. Edges b/w nodes represents IBFT from masked accounts perspective

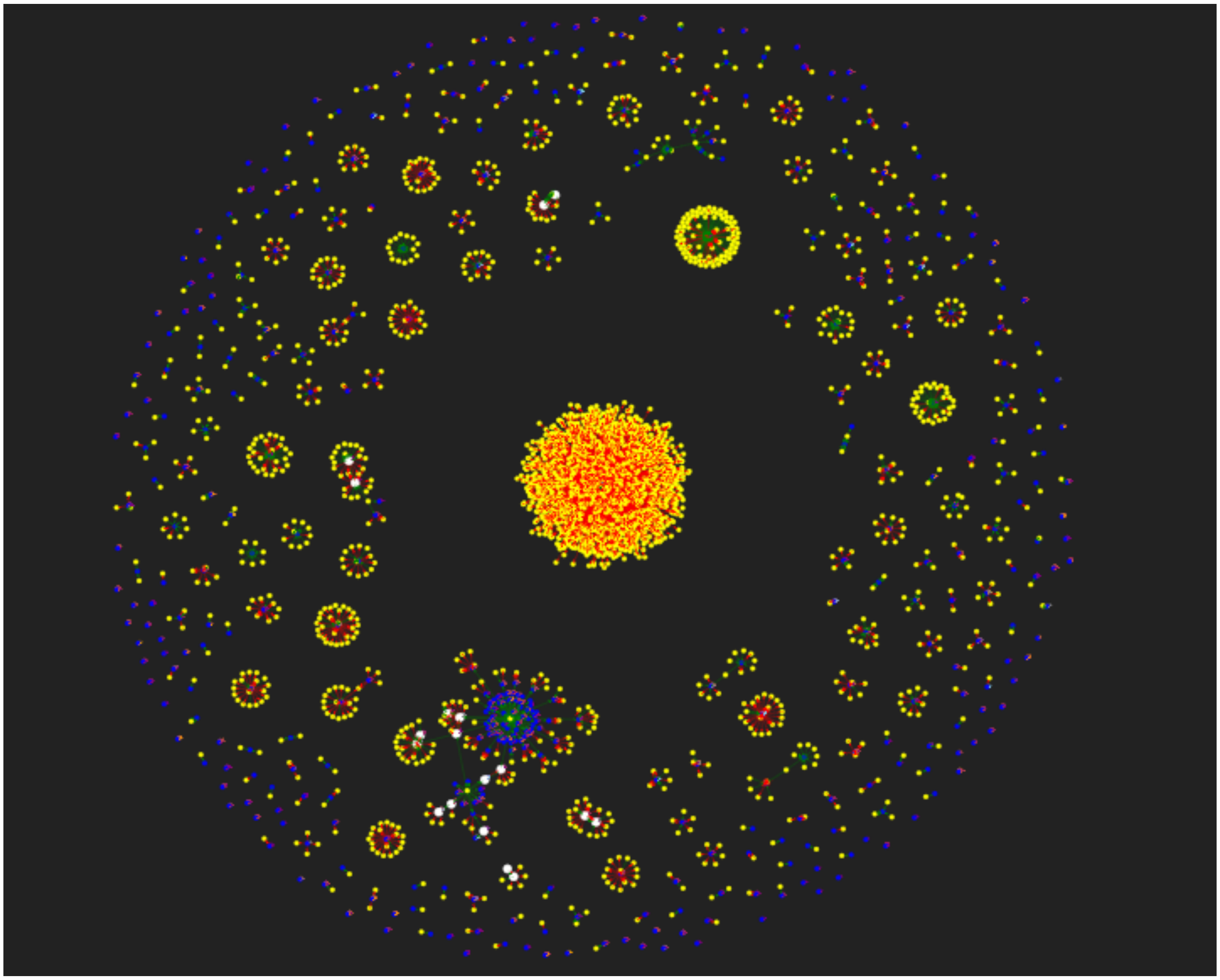
- Width of edges is directly proportional to amount of transaction
- Red edges represent debit
- Green edges represent credit

. Self-loops on nodes represent the following:

- Purple loop represents cash type transaction
- Orange loop represents atm charges type transaction
- Sky-blue loop represents bill type transaction
- Grey loop represents pos type transaction

. Zooming into the graph shows us:

- Account numbers
- Transaction amount
- Type of transaction on self-loops



Installations

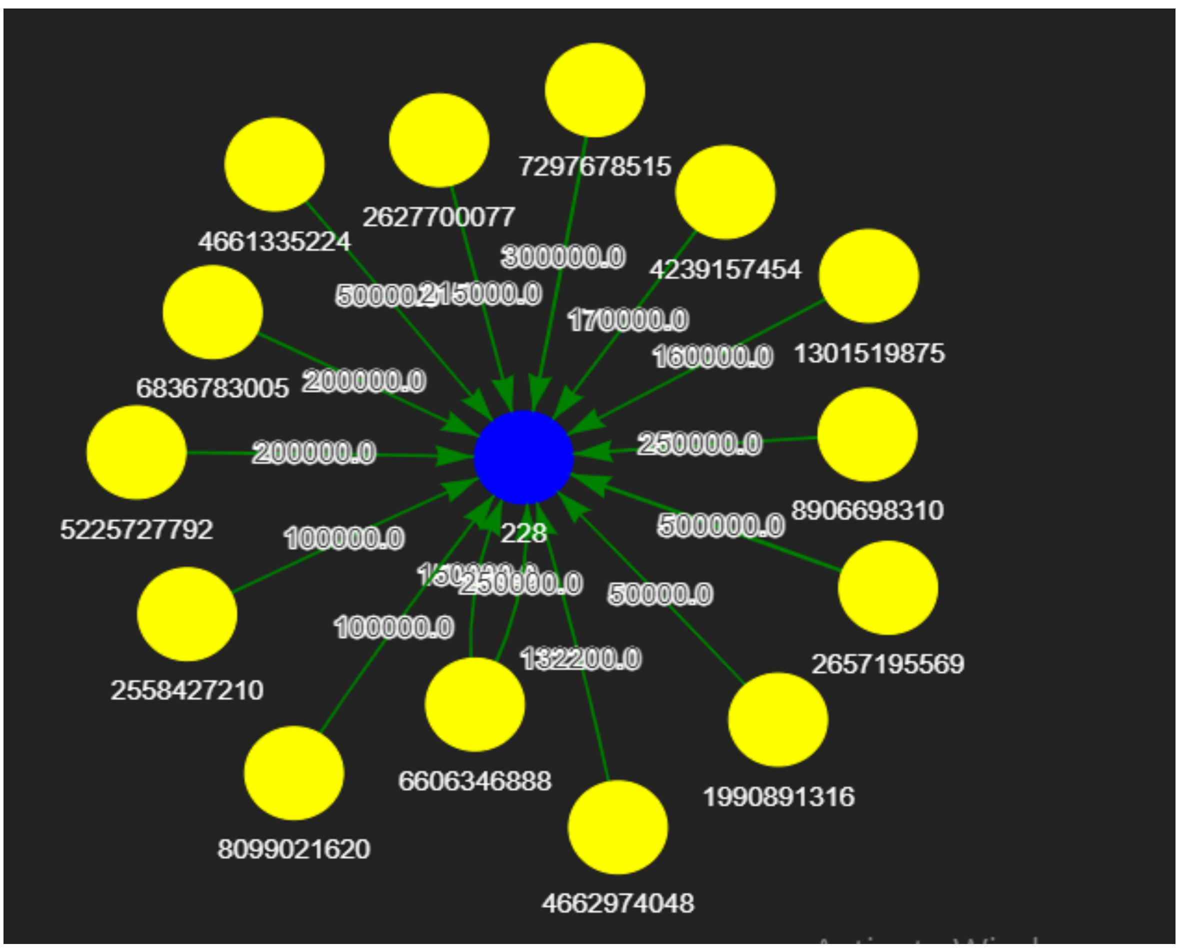
For data processing, data evaluating and data visualizing.

- pip install NumPy
- pip install pandas
- pip install matplotlib
- install pyvis

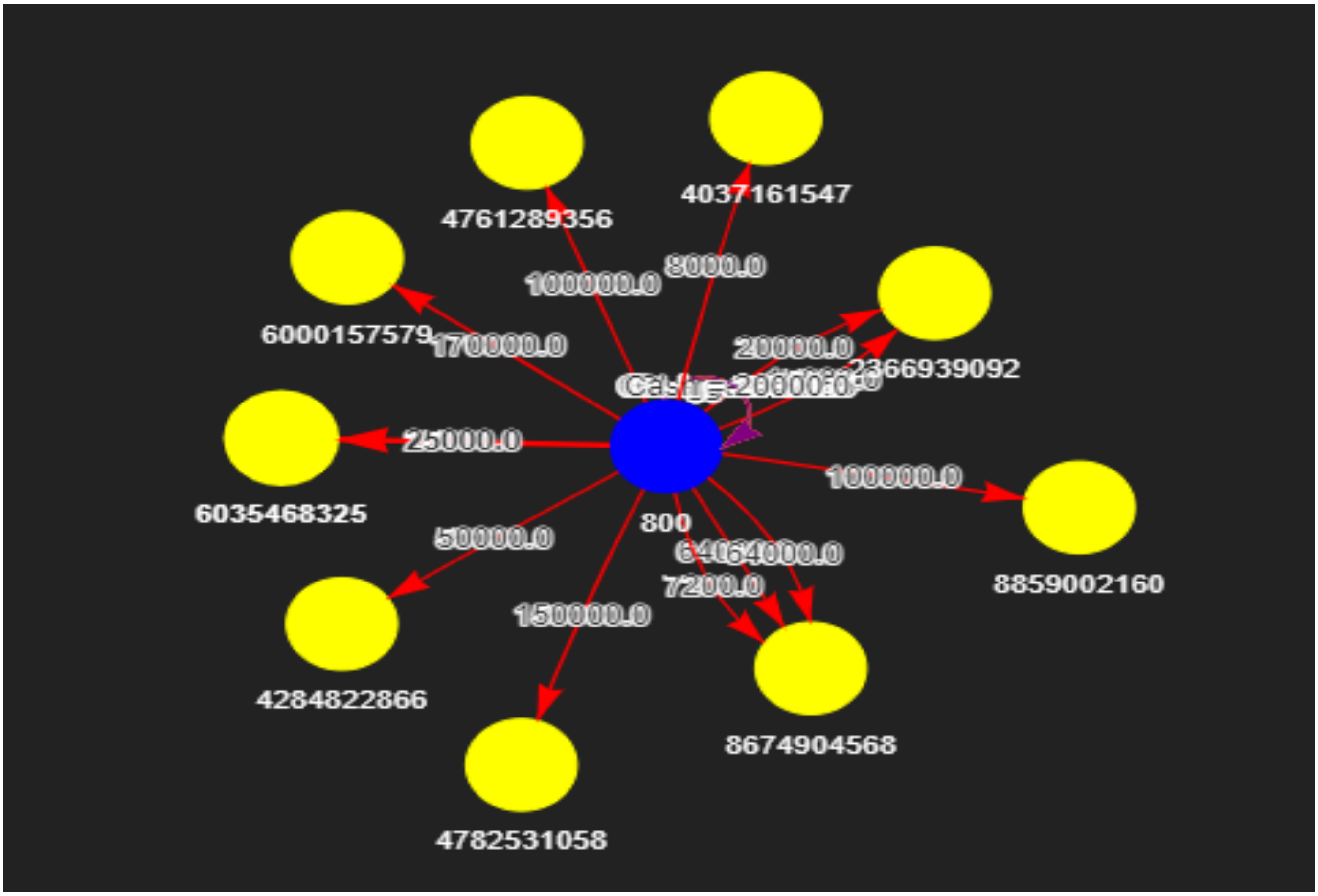
Results: Analyzing the graph

- Nodes at the circumference represent accounts that have a small network, they made very few transactions
- Moving towards the center shows more dense networks
- As we can see that there are more red edges, thus more transactions were debit in nature
- The mesh in the center shows a single blue node making many debit transactions (~1500) to different yellow nodes, further exploration told us that the nationality of this account is UK

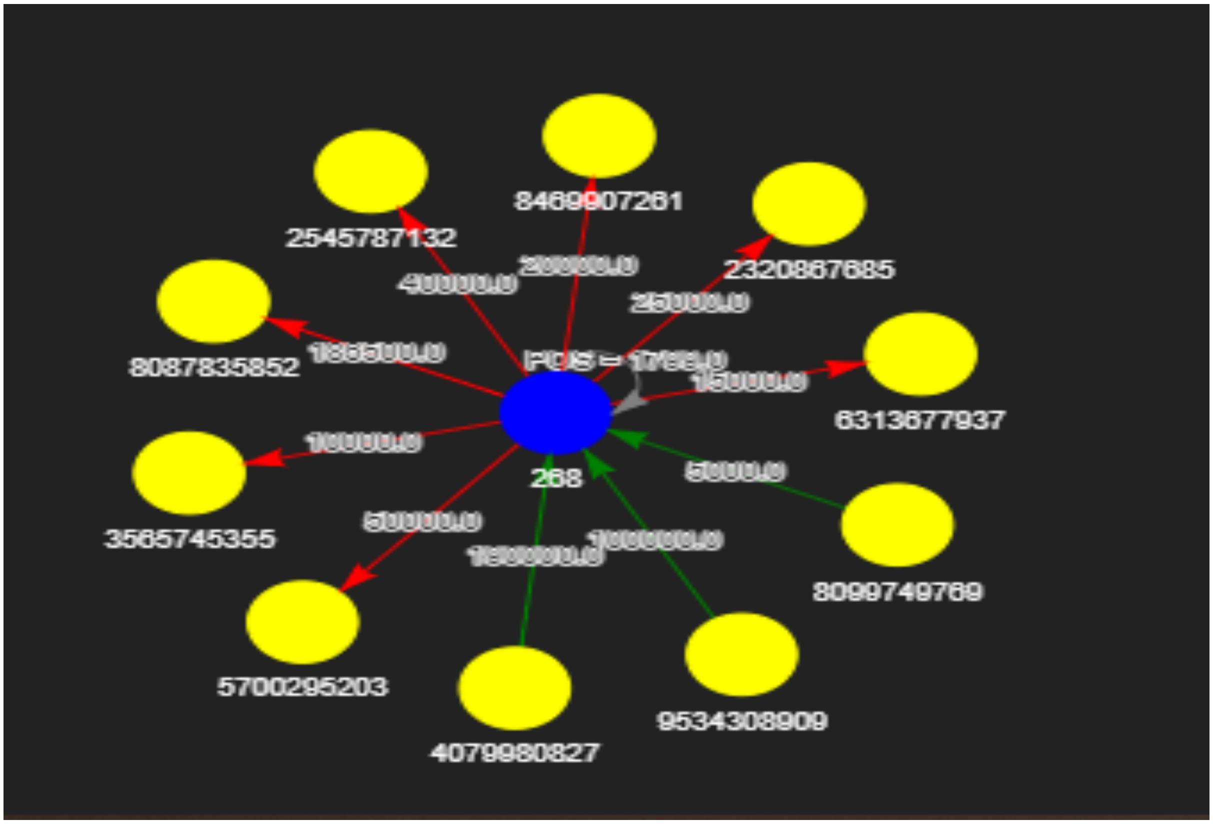
Looking into some examples:



In the above mesh, we can see that the blue account (228) was credited amounts from many different yellow accounts, and it did not make any debit transaction.



Whereas, in this mesh we can see that the blue account (800) made only debit transactions to many different yellow accounts, to one specific account it even made three debit transactions and it did not have any credit transaction.



In this example we can see that both debit and credit transactions were made.

Key Performance Indicator’s

The KPI’s have been calculated for each relationship of a masked account number and they have been measured upon monthly basis, taking a relationship as a single entity.

- Original data of all transactions of masked account number 668

	Masked_Account_Number	Sent-To	Money Flow	Date	LCY_AMOUNT2
815	668	4520789110	Debit	2020-06	-2500.00
816	668	3991706867	Debit	2021-02	-1000.00
817	668	3712840425	Debit	2021-02	-7500.00
818	668	4605821414	Debit	2021-02	-5000.00
819	668	2	Debit	2021-06	-18.75
820	668	2	Debit	2021-06	-2.50
821	668	1	Debit	2021-06	-25000.00
822	668	3991706867	Debit	2021-06	-5000.00
823	668	3991706867	Debit	2021-09	-2000.00
824	668	3991706867	Debit	2021-10	-5000.00
825	668	2	Debit	2021-12	-18.75
826	668	1	Debit	2021-12	-10000.00
827	668	9257549257	Credit	2021-12	40000.00
828	668	6077502749	Credit	2022-02	50000.00
829	668	8254781479	Credit	2022-03	100000.00
830	668	9220880966	Debit	2022-03	-50000.00
831	668	9220880966	Credit	2022-03	1500.00
832	668	1	Debit	2022-03	-20000.00
833	668	1	Debit	2022-03	-10000.00
834	668	1	Debit	2022-03	-20000.00
835	668	7165085954	Credit	2022-04	320276.00

- Summing up all transaction having the same month

	Masked_Account_Number	Sent-To	Date	Money Flow	LCY_AMOUNT2
3974	668	1	2021-06	Debit	-25000.00
3975	668	1	2021-12	Debit	-10000.00
3976	668	1	2022-03	Debit	-50000.00
3977	668	2	2021-06	Debit	-21.25
3978	668	2	2021-12	Debit	-18.75
3979	668	3712840425	2021-02	Debit	-7500.00
3980	668	3991706867	2021-02	Debit	-1000.00
3981	668	3991706867	2021-06	Debit	-5000.00
3982	668	3991706867	2021-09	Debit	-2000.00
3983	668	3991706867	2021-10	Debit	-5000.00
3984	668	4520789110	2020-06	Debit	-2500.00
3985	668	4605821414	2021-02	Debit	-5000.00
3986	668	6077502749	2022-02	Credit	50000.00
3987	668	7165085954	2022-04	Credit	320276.00
3988	668	8254781479	2022-03	Credit	100000.00
3989	668	9220880966	2022-03	Credit	1500.00
3990	668	9220880966	2022-03	Debit	-50000.00
3991	668	9257549257	2021-12	Credit	40000.00

- Representing each relationship as a single entity

	Masked_Account_Number	Sent-To	Money Flow
2545	668	1	Debit
2546	668	2	Debit
2547	668	3712840425	Debit
2548	668	3991706867	Debit
2549	668	4520789110	Debit
2550	668	4605821414	Debit
2551	668	6077502749	Credit
2552	668	7165085954	Credit
2553	668	8254781479	Credit
2554	668	9220880966	Credit
2555	668	9220880966	Debit
2556	668	9257549257	Credit

Bucket

The transaction amount has been divided into ranges. Each transaction amount falls into a bucket from 1-10 depending upon the amount. Largest bucket is 10. The average has been taken for all the transactions within a month. Large bucket size is good indicator, as it represents the transactions were of higher amounts.

Using individual amounts had the drawback of similar amounts even where the difference isn't much to be treated differently.

Longevity

The number of months transactions had taken place divided by the total months. Note that the number of transactions within a month are not considered but all transactions within a month act as a single unit. The value being closer to 1 will be a good indicator as it represents the relationship was frequent.

Currently we are dividing by total months it has potential drawbacks such as an account that was opened after the starting date of our data or an account being closed before the ending date of our data. Waiting to get data with accounts' starting and closing dates and then using the difference between them to calculate longevity.

Variance

Buckets had been categorized depending upon the range that the transaction amount falls into. The median for the range that a particular month falls into has been considered and variance has been calculated from all the months that the transaction has taken place. 0 or closer to zero is a good indicator as it represents that the transactions belonged to the same bucket or closer buckets.

Variance has been calculated upon the median of the range instead of bucket values, as they are based upon real amounts whereas buckets are a made-up feature.

Frequency

Variance has been calculated between the difference in number of months after which a transaction takes place. If a transaction takes place every month and another takes place every three months both will have the same value as both are predictable and show stability. 0 or closer to zero is a good indicator as it represents stability.

For now, frequency has been calculated using the time-period of the first transaction and the last transaction. And another has been calculated using the time-period of the first transaction till the last date of the dataset.

Representation of all KPI's

	Masked_Account_Number	Sent-To	C/D	Bucket	Longevity	Variance	Frequency-End Date	Frequency- Transaction Date
2545	668	1	Debit	5.0	0.125000	8.520833e+08	2.0	2.250000
2546	668	2	Debit	1.0	0.083333	0.000000e+00	0.0	0.000000
2547	668	3712840425	Debit	3.0	0.041667	NaN	NaN	NaN
2548	668	3991706867	Debit	3.0	0.166667	0.000000e+00	6.5	1.555556
2549	668	4520789110	Debit	3.0	0.041667	NaN	NaN	NaN
2550	668	4605821414	Debit	3.0	0.041667	NaN	NaN	NaN
2551	668	6077502749	Credit	6.0	0.041667	NaN	NaN	NaN
2552	668	7165085954	Credit	8.0	0.041667	NaN	NaN	NaN
2553	668	8254781479	Credit	7.0	0.041667	NaN	NaN	NaN
2554	668	9220880966	Credit	3.0	0.041667	NaN	NaN	NaN
2555	668	9220880966	Debit	6.0	0.041667	NaN	NaN	NaN
2556	668	9257549257	Credit	5.0	0.041667	NaN	NaN	NaN