Vertical Institute Capstone Project (Data Analytics Bootcamp)

File created on: 7/27/2024 9:41:23 AM

Leong Wei Shin

Preamble

This data set was approved on Jul 20 2024 via Telegram: https://www.kaggle.com/datasets/goyaladi/fraud-detection-dataset

The csv files used in the visualisation and analysis are as follows:

1. Customer ID, Balance
2. Customer ID, Age
3. Transaction ID, Fraudulent Indicator (1/0)
4. Customer ID, Suspicious Activity (1/0)
5. Merchant ID
6. Transaction ID, Category
7. Transaction ID, Amount
8. Transaction ID, Anomaly Score
9. Transaction ID, Merchant ID 1
10. Transaction ID, Amount, Customer ID

The data is clean and can be easily connected to a Licensed Tableau Version 2022.2.

The worksheets can be found here:

Tableau Packaged Workbook

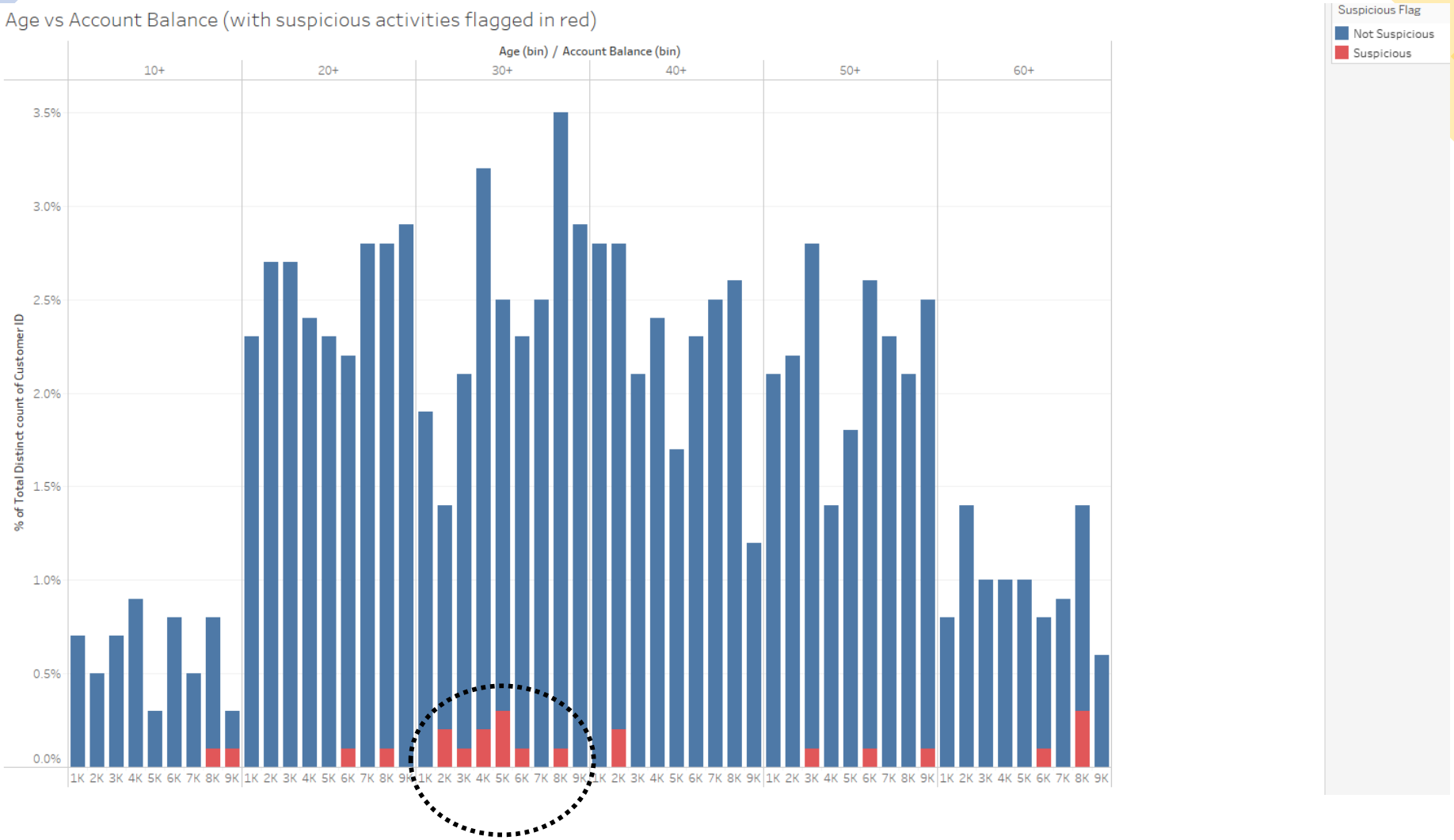
Inquiry Questions

These are the approved inquiry questions:

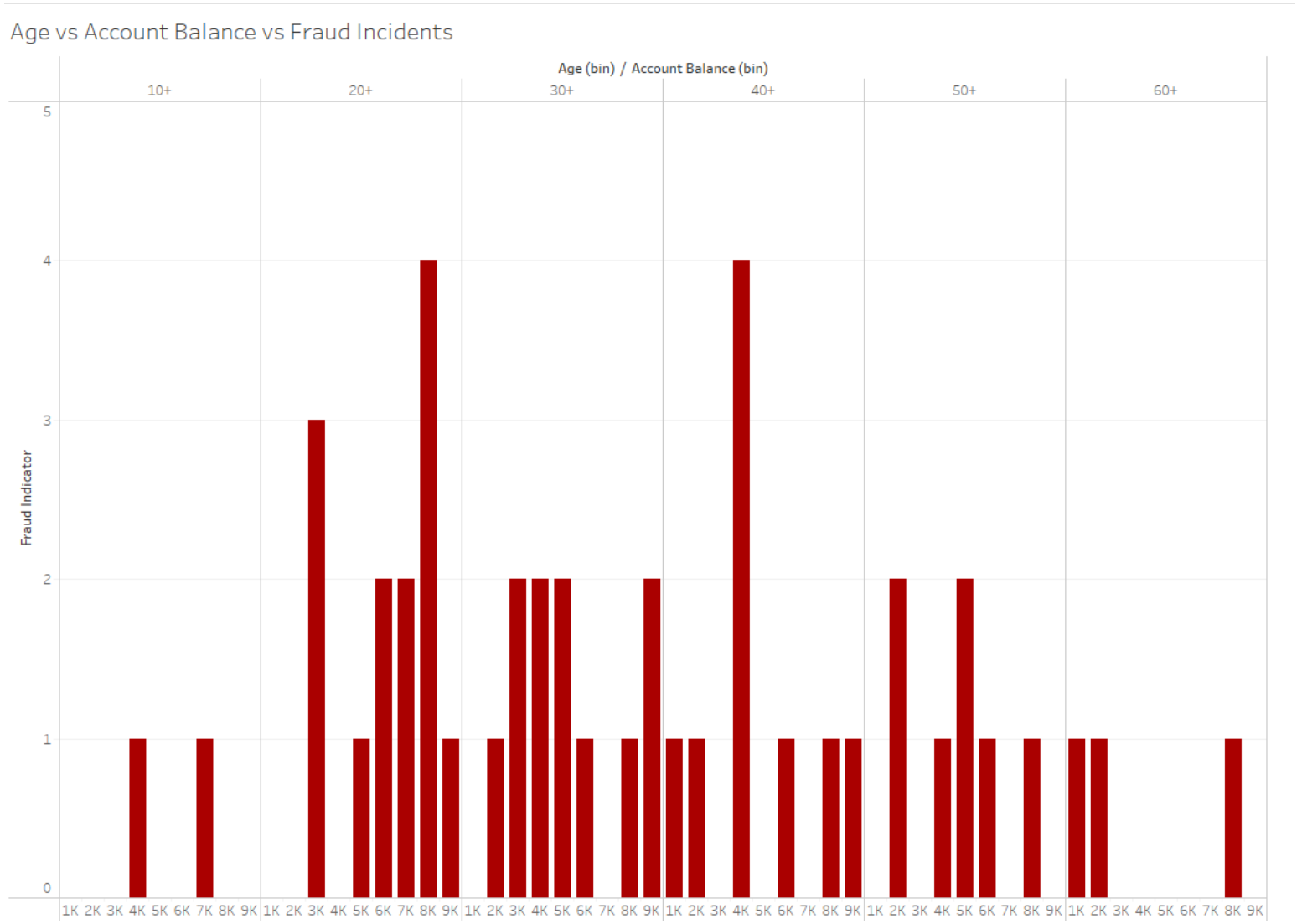
a. Which profile of customer (age and balance) is prone to experience fraudulent or flagged up as having suspicious activity?

b. Which category of transaction has higher/lower fraudulent incidents?

c. What is the relationship between transaction anomaly score and presence/absence of fraudulence or flagged up as having suspicious activity?

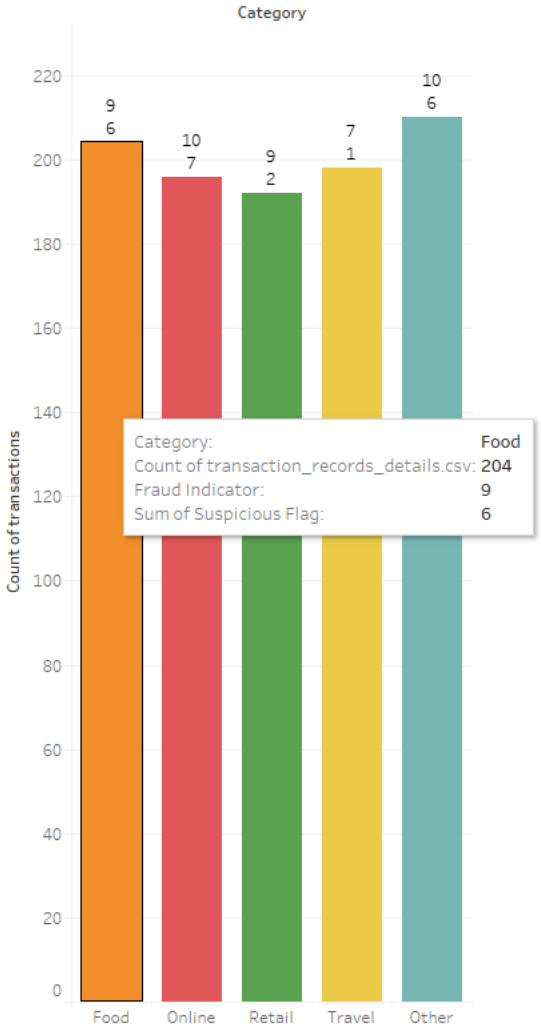


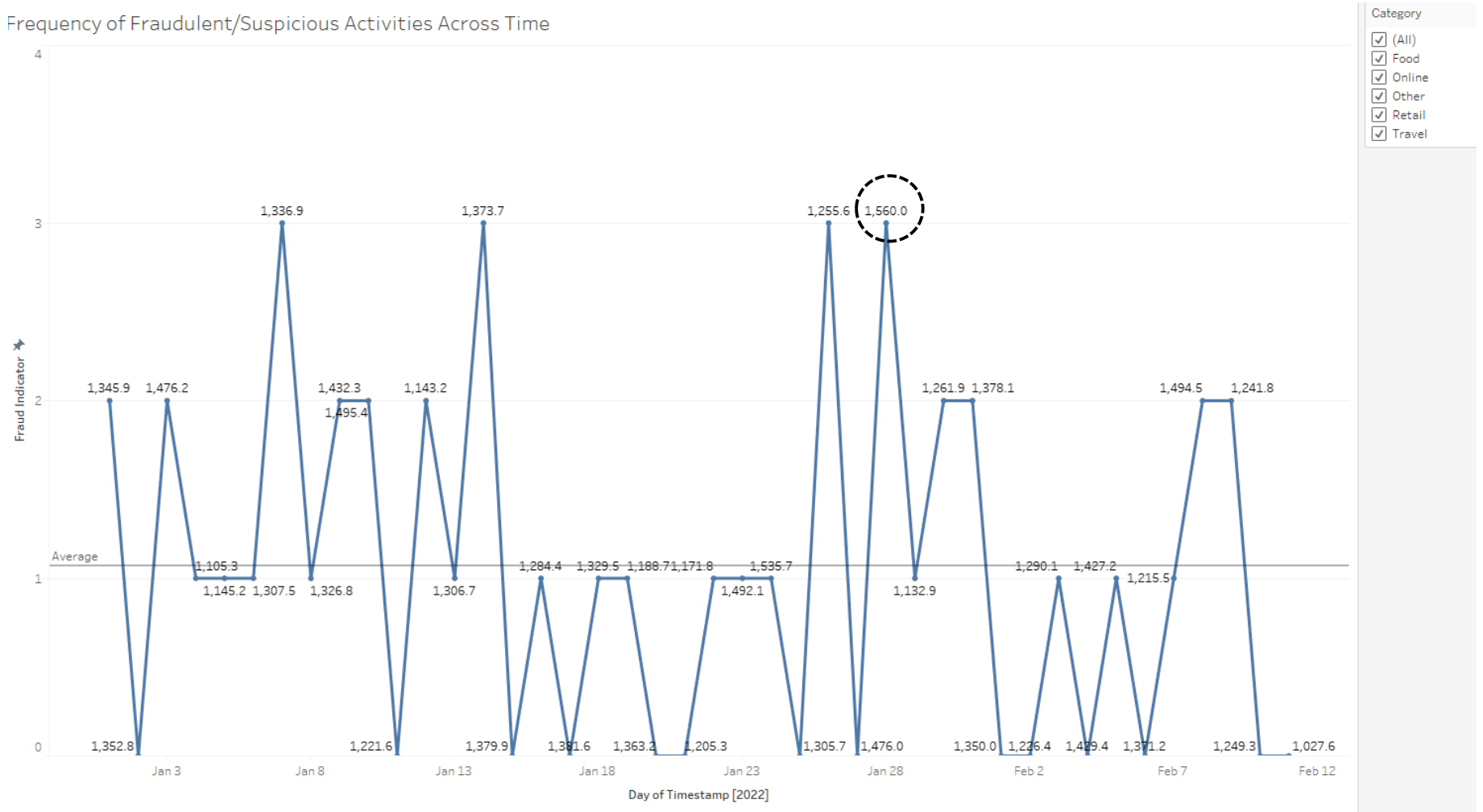
Interpretation: Suspicious transactions are detected across age group and size of account, though the more prevalent occurrence seem to happen at age group 30+.



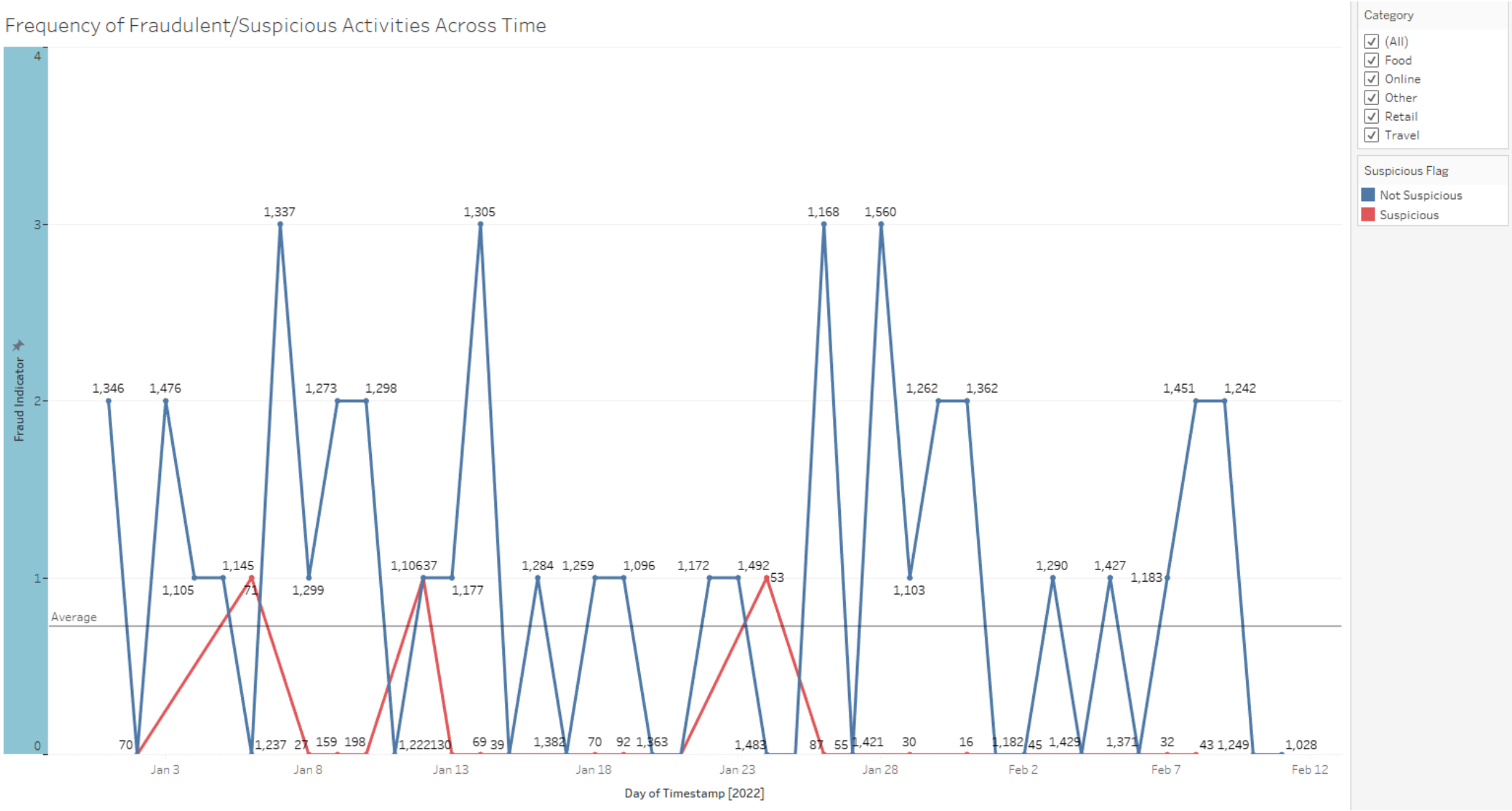
Interpretation: Frauduent transactions are detected across age group and size of account, though the more prevalent occurrence seem to happen at age group 20+ and 40+.

Category of Transactions and Count of Transactions (with labels of frequency of fraudulent and suspicious activities)

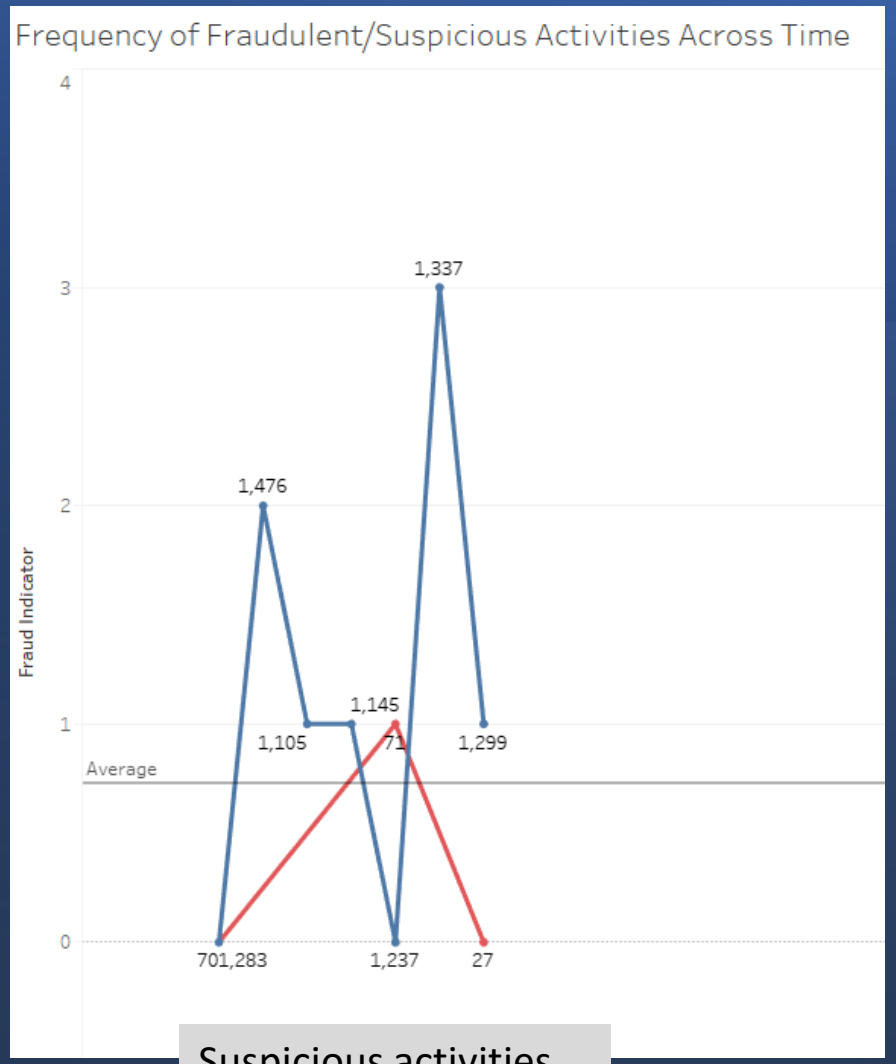




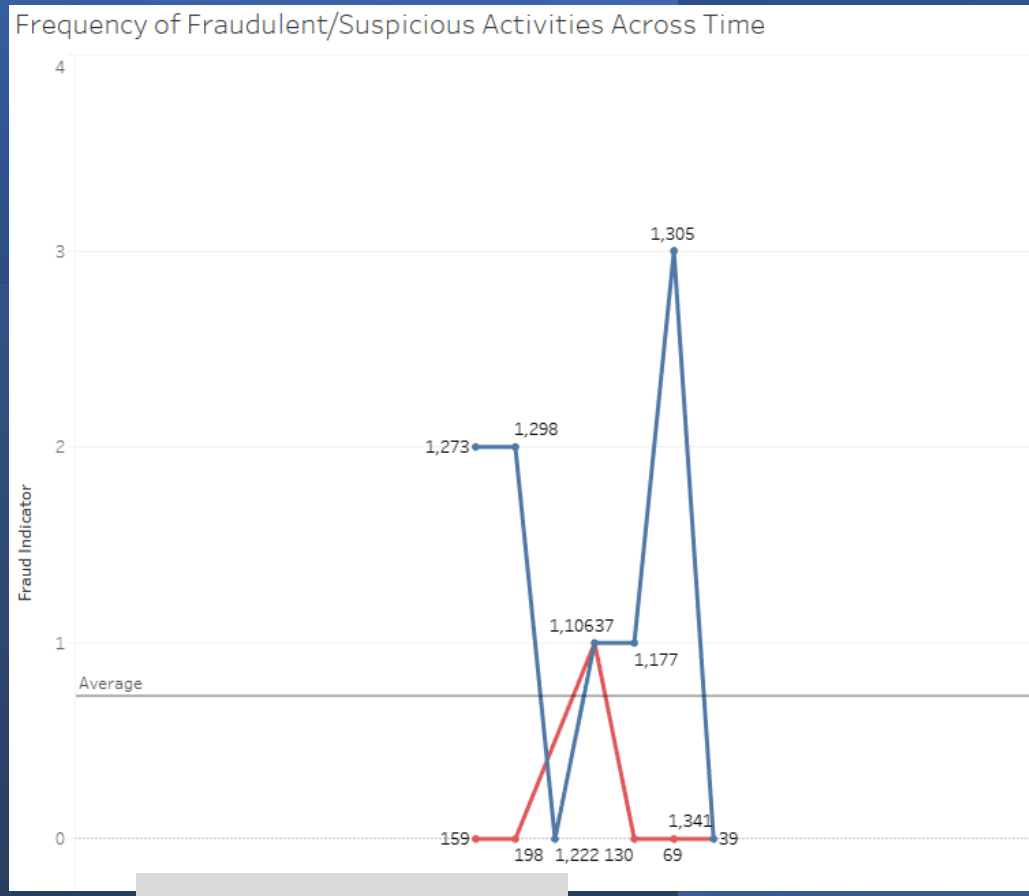
Interpretation: Fraudulent transactions are detected across all categories of transactions ranging from 0-3 per day (average of about 1), without any seeming pattern of day it occur (within the span of available data of two months). Max fraudulent transactions amount size is 1.56k on Jan 28 2022.



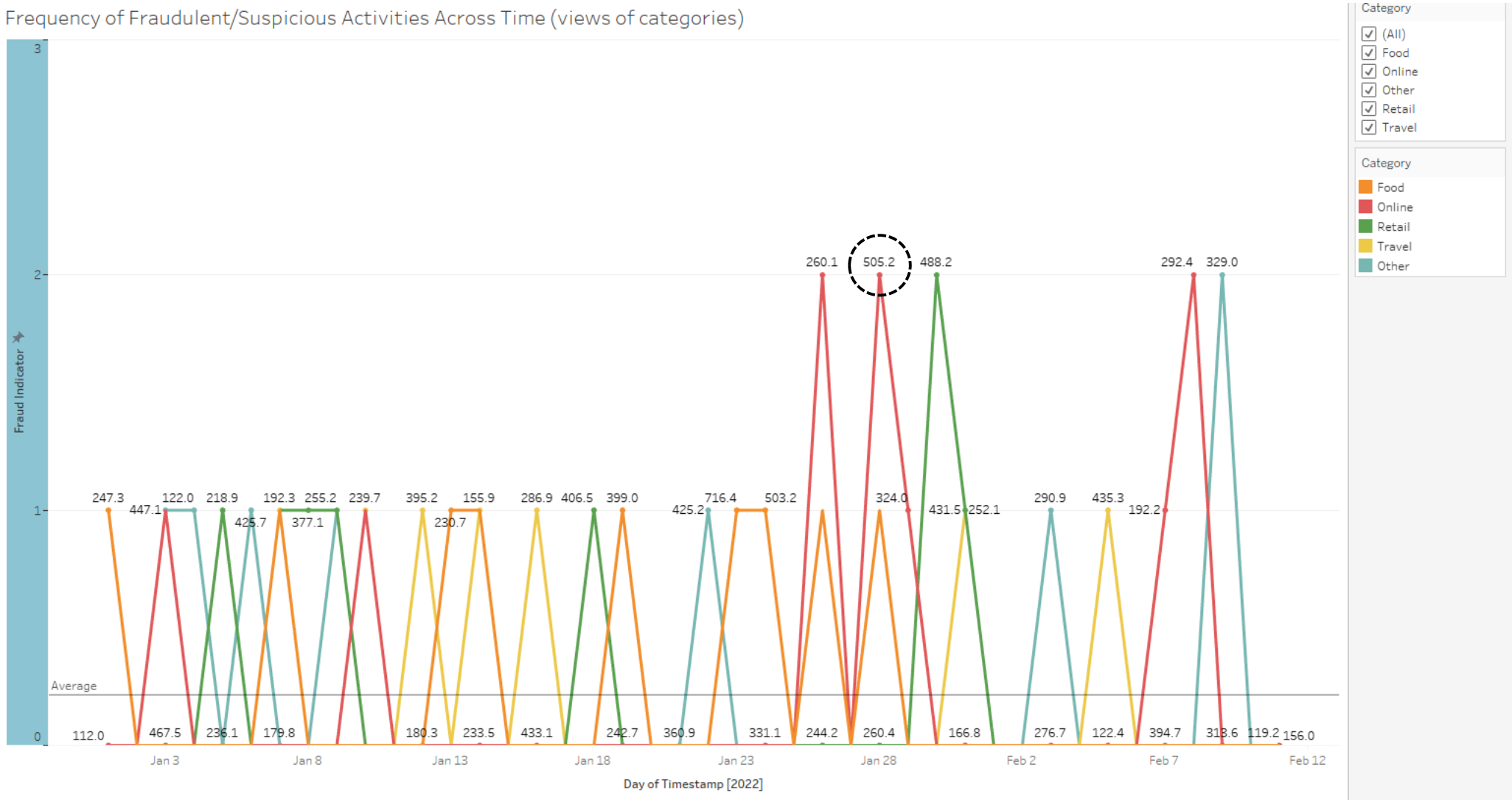
Interpretation: Superimposing the trend data for flagging of 'suspicious activity' (red line), it does not necessarily flag out fraudulent activities on the day itself. However, it seems to predict a possible spike of fraudulent activities **in the next day/next few days (see next slide).**



Suspicious activities flagged on Jan 6 and spike of fraudulent activities on Jan 7.



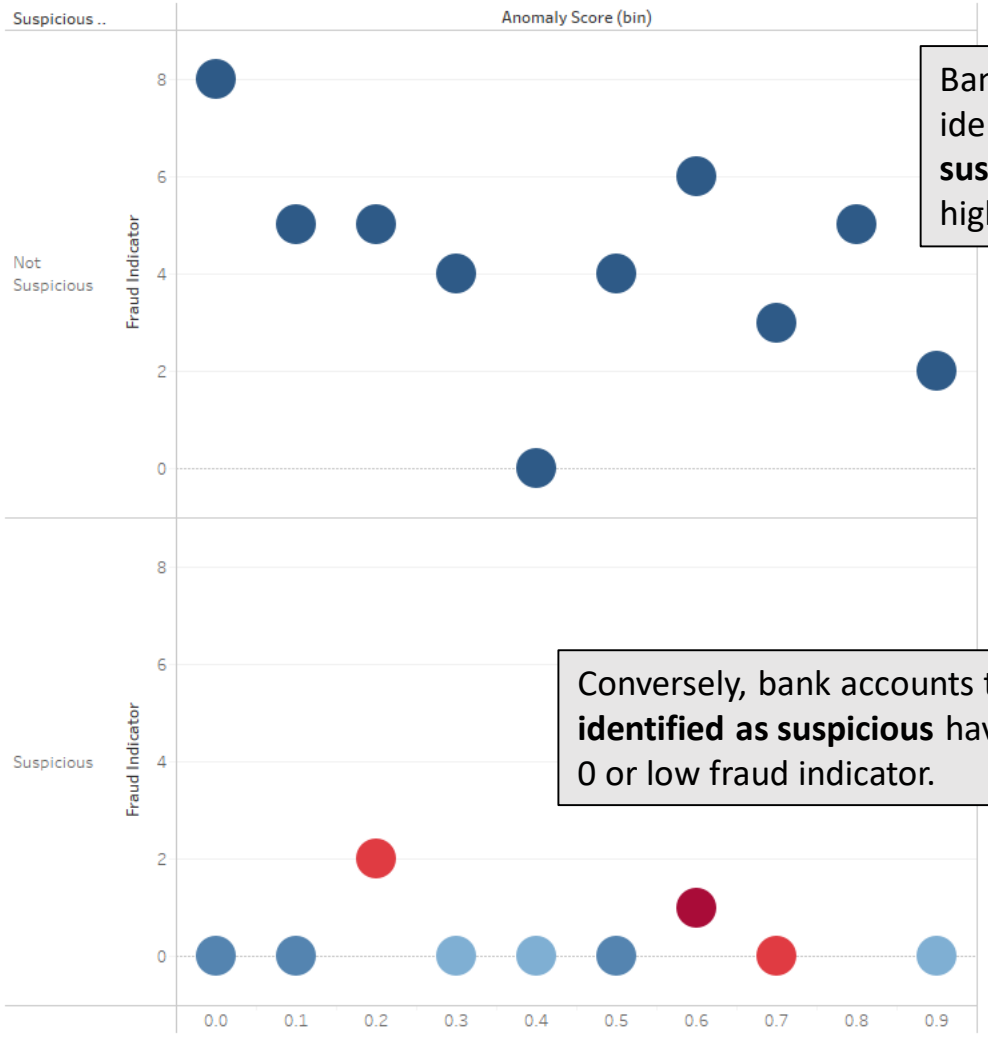
Suspicious activities flagged on Jan 9 and spike of fraudulent activities on Jan 14.



Interpretation: Digging into the category of transactions that attract the more/most fraudulent activities (that coincide with the spike on Jan 26 and Jan 28), **online category** seems to jump out more than other categories. Max single fraudulent transaction amount size is on Jan 28 2022 is an online transaction.

Caveat: Not sure what is the 'other' category referring to in the data set.

Relationship between Anomaly Score of Transaction with Fraudulent Indicator and Flagging of Suspicious Activity.



Bank accounts that were identified as **NOT suspicious** did have higher fraud indicator.

Conversely, bank accounts that were **identified as suspicious** have mostly 0 or low fraud indicator.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.048197753
R Square	0.002323023
Adjusted R Square	0.001323347
Standard Error	0.207270573
Observations	1000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.099832	0.099832	2.323776	0.127727252
Residual	998	42.87517	0.042961		
Total	999	42.975			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.062062241	0.012971	4.784789	1.97E-06	0.036609196	0.087515	0.036609	0.087515
AnomalyScore	-0.034659506	0.022737	-1.52439	0.127727	-0.079276508	0.009957	-0.07928	0.009957

	AnomalyScore	FraudIndicator
AnomalyScore	1	
Fraud Indicator	-0.048197753	1

No significant relationship between current Anomaly Score and Fraud Indicator

Interpretation: Analysing the relationship between the anomaly score, fraudulent indicator and flagging of suspicious activity – nothing conclusive can be stated now. For instance, for transactions flagged as highly anomalous (> 0.9), only two incidents of fraudulent and 0 suspicious account were identified. Bank accounts that were identified as suspicious did not have high incidents of fraud (<2). **Further regression analysis also shows no significant correlation between anomaly score and fraudulent indicators which suggest improvements need to enhancing the former.**

Recommendations

- Although based on the current data set suggests the occurrence of fraudulent activities is low (45 cases out of the 1000 transactions = **4.5%**), **one is too many.** Any case(s) can cause severe financial hardships for individuals.
- Based on the analysis, more can be done to **sharpen the identification** of suspicious activities and identification of anomalous activities.
- If online transactions are likely to grow, it is probably where fraudulent activities will be more concentrated in the future. **MORE need to be done** to alert both clients and banks on likely suspicious and possibly anomalous activities.

Specific Recommendations for Banks

- Explore ways to sharpen/fine tune the derivation of anomaly score and flagging of suspicious account. Perhaps these data/variable need to be considered:
  - ✓ Data from other banks that have the same identity including record of filed Suspicious Transaction Report ("STR")
  - ✓ Relevant 'Know Your Customer' ("KYC") data (can AI help?) such as expected transaction volume and frequency, purpose of account opening, source of fund

- Based on KYC data, would there be some intelligent way for Banks to alert customers on one or more of these:
  - ✓ Unusual frequency/volume of transactions of recipient accounts
  - ✓ Warn clients *NEVER* send money to recipients who are unknown to them previously
  - ✓ Constant education to clients on the latest fraudulent typologies

Acknowledgements

- Clarence Tan and Tay Xun Yang for their excellent teaching and support throughout the course.
- Putri Astika for the efficient communications and support.
- Seike Sei for his invaluable suggestions from a bank’s perspective.
- Christabel Soh for her feedback on improving the visualisations.