SwissBorg

Tech Challenge - Fraud Data Analyst

Candidate name : Júlio Silva

Notes:

This document presents the answers to the questions in the tech challenge and all graphs that were created with a brief analysis.

Most of the graphs are in a bar **stack** format with the y axis representing the number of transactions for a given scenario. In this format multiple bars can be stacked upon each other, if a colored bar starts at y1 = 20 and ends at y2 = 30 the total transactions for that bar will be 10 (y2-y1) and not 30.

The stacked bar chart seemed to me the best graphical option to use since there are so many categorical variables for this dataset.

The best 6 graphs that I suggest for an imaginary dashboard will be stored as an svg in a separate folder.

All code is stored in code_notebook.ipynb for the graphs and sql_notebook.ipynb for the sql queries. This ipynb format is basically a python jupyter notebook and should automatically display all results when opened.

As a final note, I would like to thank Swissborg for giving me the opportunity of working in such an entertaining dataset. I had fun exploring it in my free time!

Question 1:

Create a set of up to 6 charts that best describe the evolution of fraud cases over time in the dataset. Imagine that these charts would form a dashboard.

Answer:

The 6 graphs are stored in the folder **imag_dashboard**

Question 2:

What are the 3 main differences between fraud and legitimate users?

Answer:

1. Legitimate users normally have fewer transactions on Friday and Thursday when compared with fraudulent users. Fraudulent users also do more transactions towards the morning and afternoon while legitimate users do more towards the end of the night.
2. Fraudulent users seem to utilise more fiat currencies (EUR and USDT) and BTC in fraudulent transactions that are not of the event_kind **exchange** when compared to legitimate users.
3. Legitimate users very rarely make transactions in the first 9 days after onboarding, while fraudulent users are the most active in this time.

Question 3 :

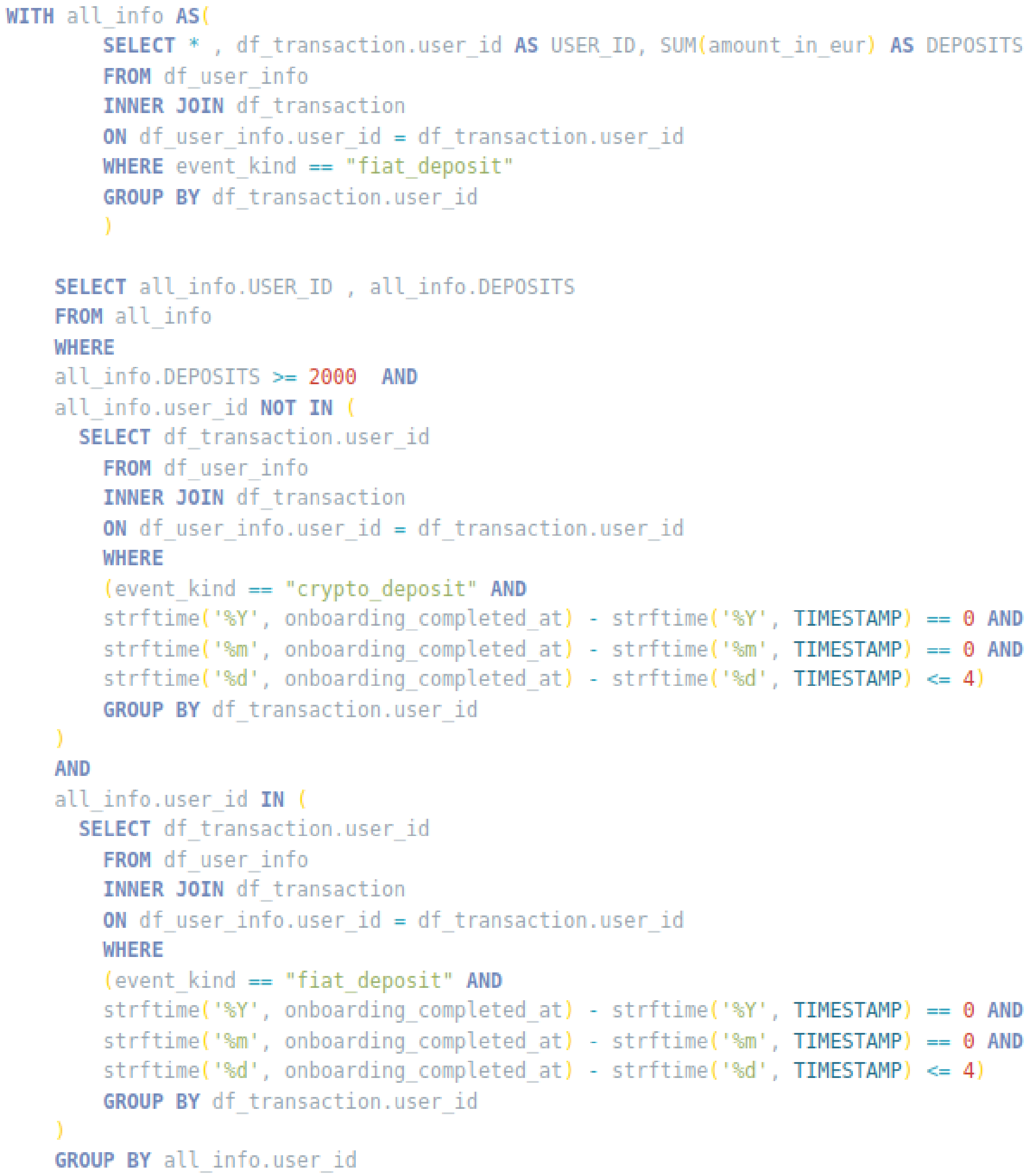
Is there a distinctive transaction behaviour of fraud users?

Answer:

Transactions that are of the event_kind of “exchange” have a high probability of being fraudulent when the pair of currency used is GBP_BTC. Legit users normally do not use this conversion in exchanges.

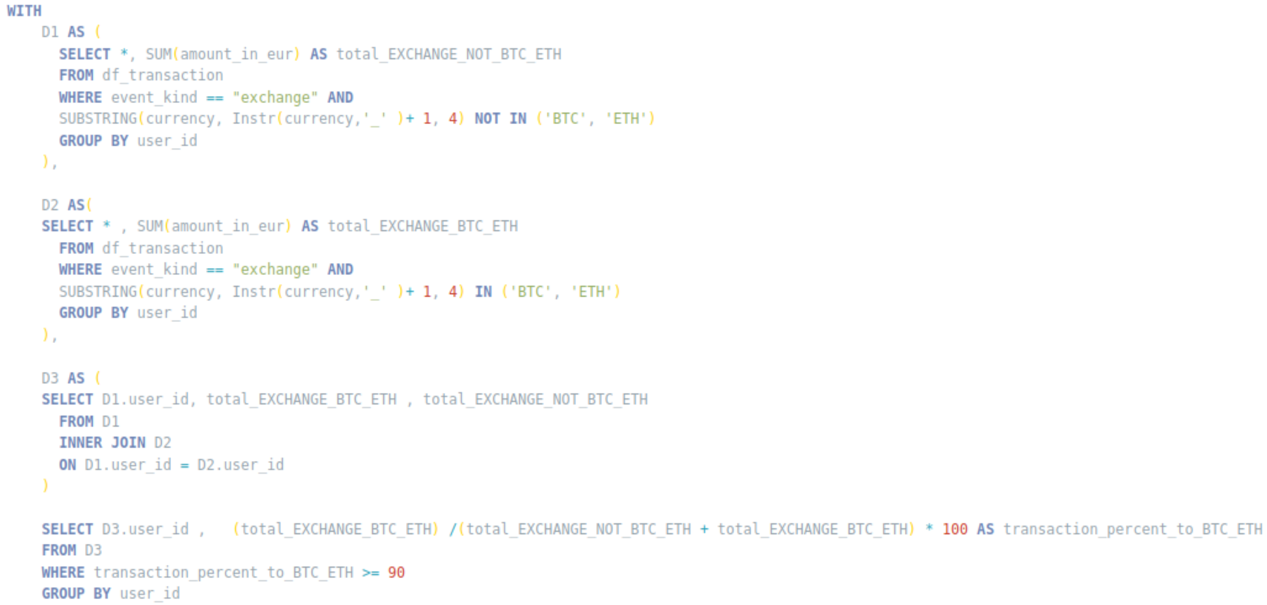
In addition, fraud users seem to be overwhelming from the country_code GB and they are very rarely a tier T3 or a premium user.

Question 4 a)



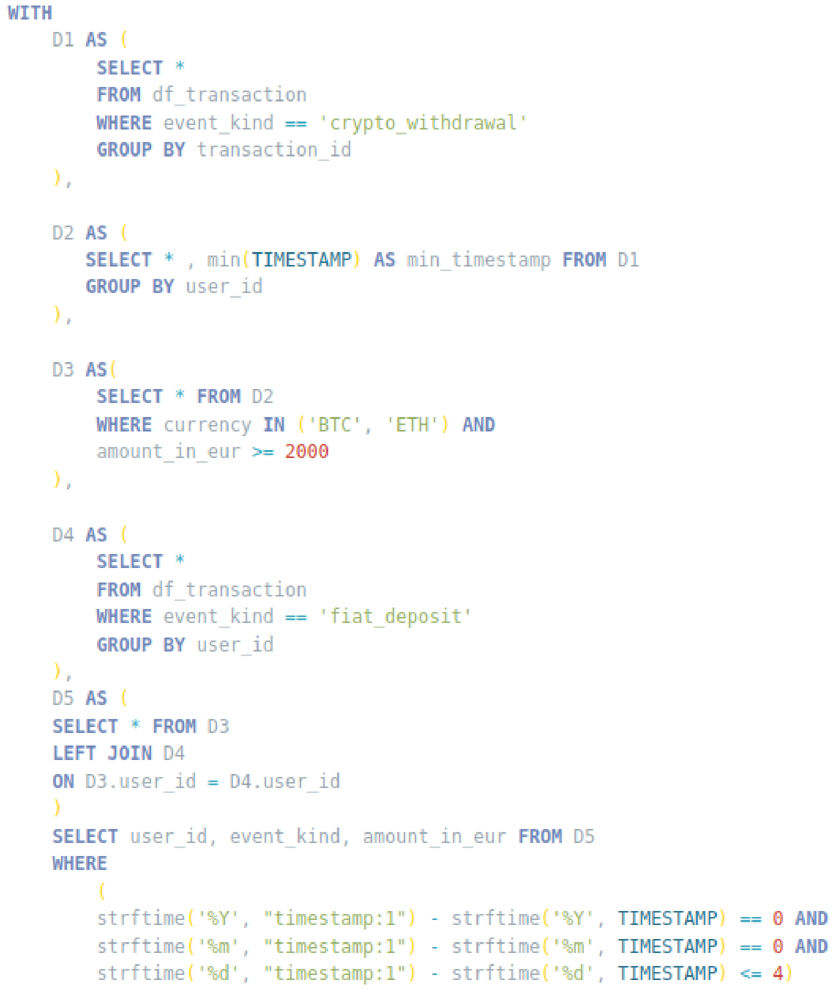
```
WITH all_info AS(
        SELECT * , df_transaction.user_id AS USER_ID, SUM(amount_in_eur) AS DEPOSITS
        FROM df_user_info
        INNER JOIN df_transaction
        ON df_user_info.user_id = df_transaction.user_id
        WHERE event_kind == "fiat_deposit"
        GROUP BY df_transaction.user_id
        )

    SELECT all_info.USER_ID , all_info.DEPOSITS
    FROM all_info
    WHERE
    all_info.DEPOSITS >= 2000  AND
    all_info.user_id NOT IN (
      SELECT df_transaction.user_id
        FROM df_user_info
        INNER JOIN df_transaction
        ON df_user_info.user_id = df_transaction.user_id
        WHERE
        (event_kind == "crypto_deposit" AND
        strftime('%Y', onboarding_completed_at) - strftime('%Y', TIMESTAMP) == 0 AND
        strftime('%m', onboarding_completed_at) - strftime('%m', TIMESTAMP) == 0 AND
        strftime('%d', onboarding_completed_at) - strftime('%d', TIMESTAMP) <= 4)
        GROUP BY df_transaction.user_id
    )
    AND
    all_info.user_id IN (
      SELECT df_transaction.user_id
        FROM df_user_info
        INNER JOIN df_transaction
        ON df_user_info.user_id = df_transaction.user_id
        WHERE
        (event_kind == "fiat_deposit" AND
        strftime('%Y', onboarding_completed_at) - strftime('%Y', TIMESTAMP) == 0 AND
        strftime('%m', onboarding_completed_at) - strftime('%m', TIMESTAMP) == 0 AND
        strftime('%d', onboarding_completed_at) - strftime('%d', TIMESTAMP) <= 4)
        GROUP BY df_transaction.user_id
    )
    GROUP BY all_info.user_id
```

Question 4 b)



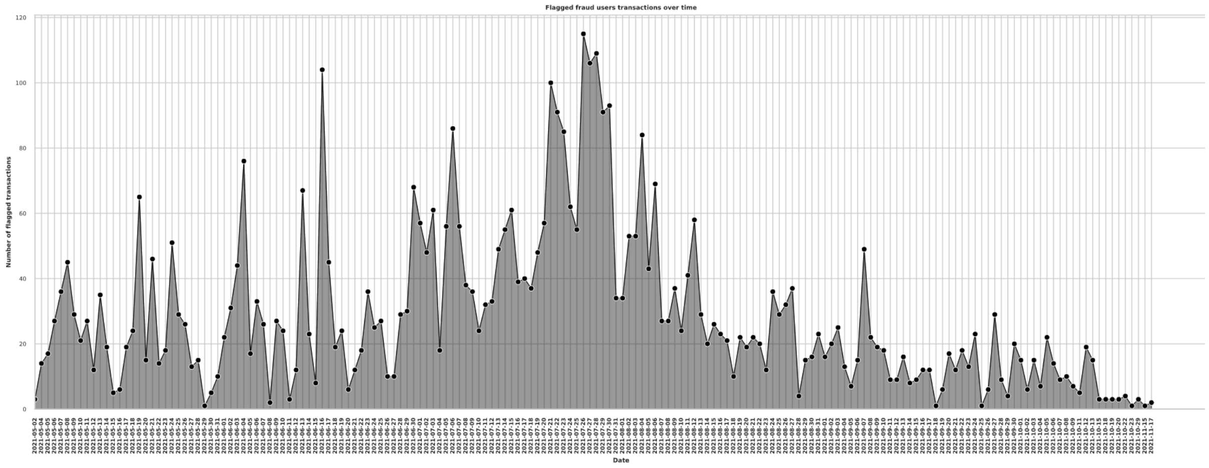
```
WITH
    D1 AS (
      SELECT *, SUM(amount_in_eur) AS total_EXCHANGE_NOT_BTC_ETH
      FROM df_transaction
      WHERE event_kind == "exchange" AND
      SUBSTRING(currency, Instr(currency,'_' )+ 1, 4) NOT IN ('BTC', 'ETH')
      GROUP BY user_id
    ),

    D2 AS(
    SELECT * , SUM(amount_in_eur) AS total_EXCHANGE_BTC_ETH
      FROM df_transaction
      WHERE event_kind == "exchange" AND
      SUBSTRING(currency, Instr(currency,'_' )+ 1, 4) IN ('BTC', 'ETH')
      GROUP BY user_id
    ),

    D3 AS (
    SELECT D1.user_id, total_EXCHANGE_BTC_ETH , total_EXCHANGE_NOT_BTC_ETH
      FROM D1
      INNER JOIN D2
      ON D1.user_id = D2.user_id
    )

    SELECT D3.user_id ,   (total_EXCHANGE_BTC_ETH) /(total_EXCHANGE_NOT_BTC_ETH + total_EXCHANGE_BTC_ETH) * 100 AS transaction_percent_to_BTC_ETH
    FROM D3
    WHERE transaction_percent_to_BTC_ETH >= 90
    GROUP BY user_id
```

Question 4 c)



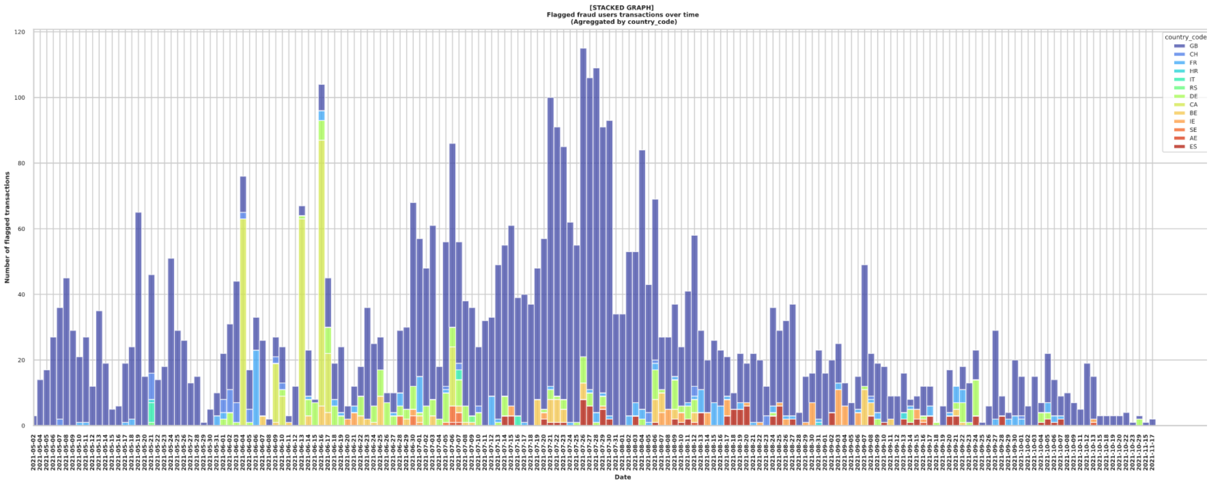
```
WITH
    D1 AS (
        SELECT *
        FROM df_transaction
        WHERE event_kind == 'crypto_withdrawal'
        GROUP BY transaction_id
    ),

    D2 AS (
       SELECT * , min(TIMESTAMP) AS min_timestamp FROM D1
       GROUP BY user_id
    ),

    D3 AS(
        SELECT * FROM D2
        WHERE currency IN ('BTC', 'ETH') AND
        amount_in_eur >= 2000
    ),

    D4 AS (
        SELECT *
        FROM df_transaction
        WHERE event_kind == 'fiat_deposit'
        GROUP BY user_id
    ),
    D5 AS (
    SELECT * FROM D3
    LEFT JOIN D4
    ON D3.user_id = D4.user_id
    )
    SELECT user_id, event_kind, amount_in_eur FROM D5
    WHERE
        (
        strftime('%Y', "timestamp:1") - strftime('%Y', TIMESTAMP) == 0 AND
        strftime('%m', "timestamp:1") - strftime('%m', TIMESTAMP) == 0 AND
        strftime('%d', "timestamp:1") - strftime('%d', TIMESTAMP) <= 4)
```

Flagged Fraud Transactions over time.



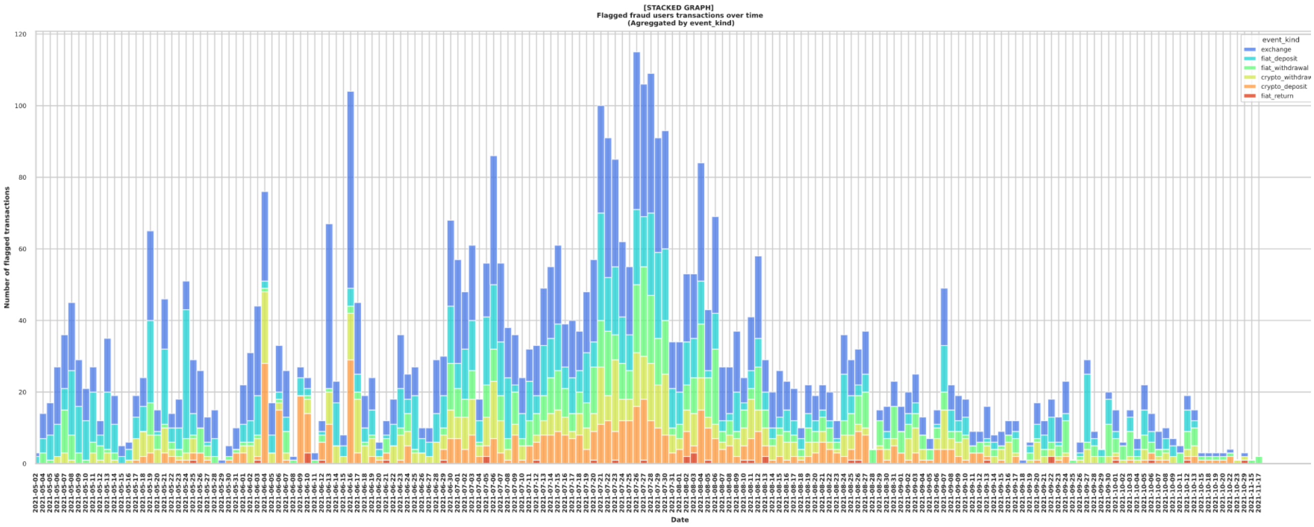
Analysis : The month of july seemed to be the most active for fraudulent transactions

Country Code



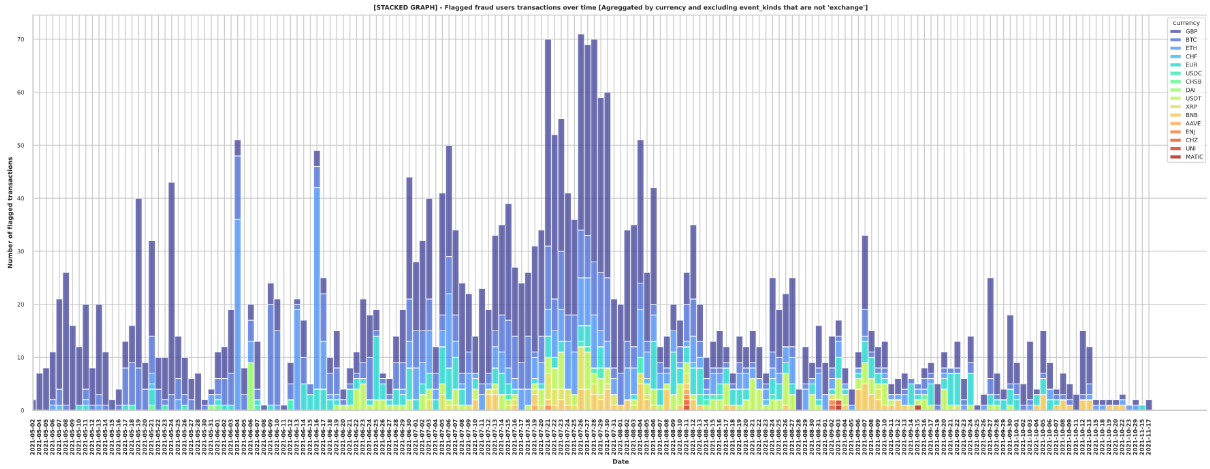
Analysis : Most flagged transactions were done by users from GB and CA.

Event_kind



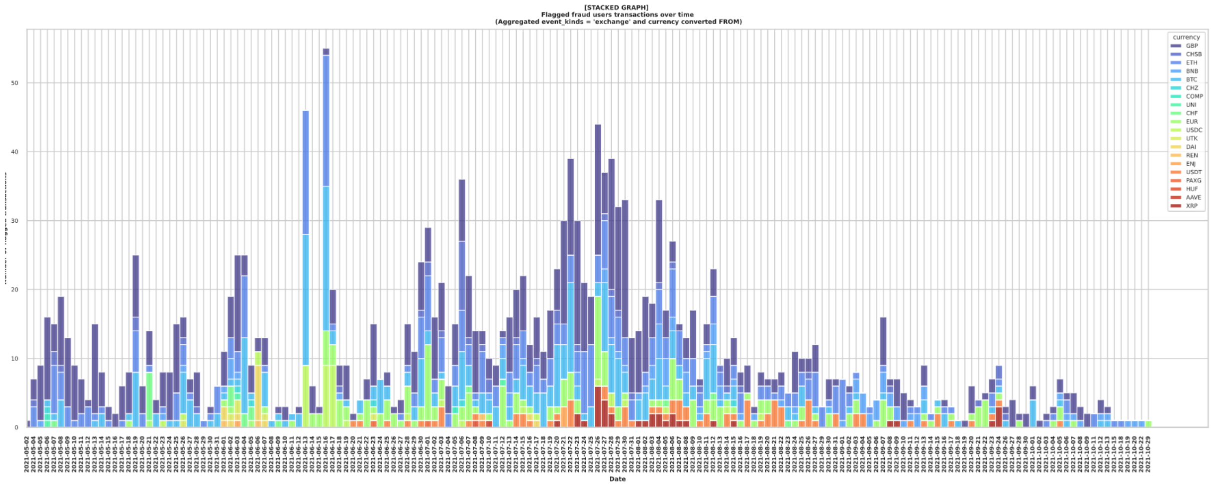
Analysis : Regarding flagged transactions events, exchanges seem to be the most common for fraudulent activity while fiat_return seems to be the rarest.

Currency USED in Event_kind that is not “Exchange”



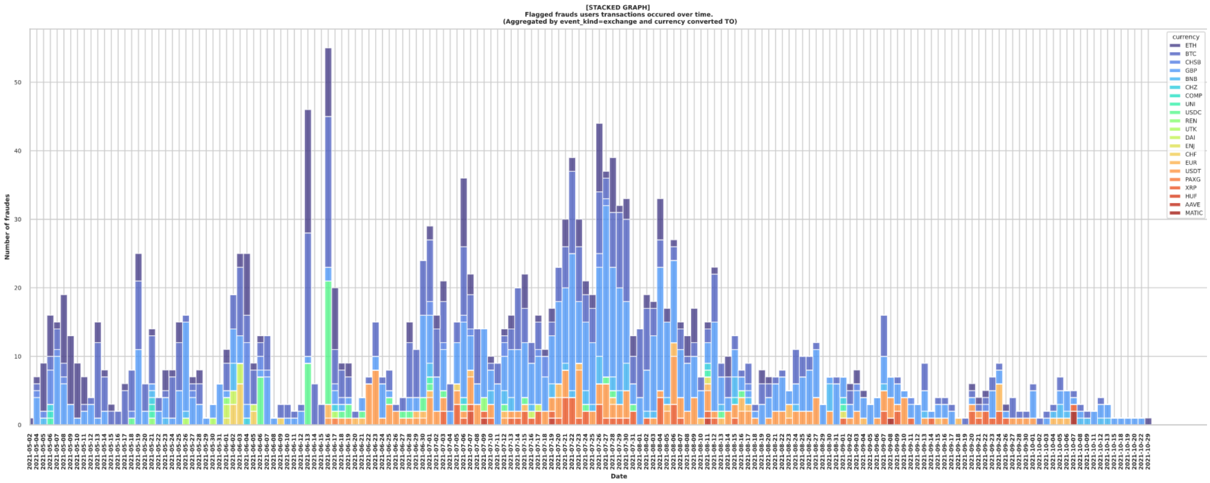
Analysis : Most flagged transactions that occurred for every event_kind != “Exchange” were done with GBP, BTC and ETH.

FromCurrency Currency most used in Event_kind that is “Exchange”



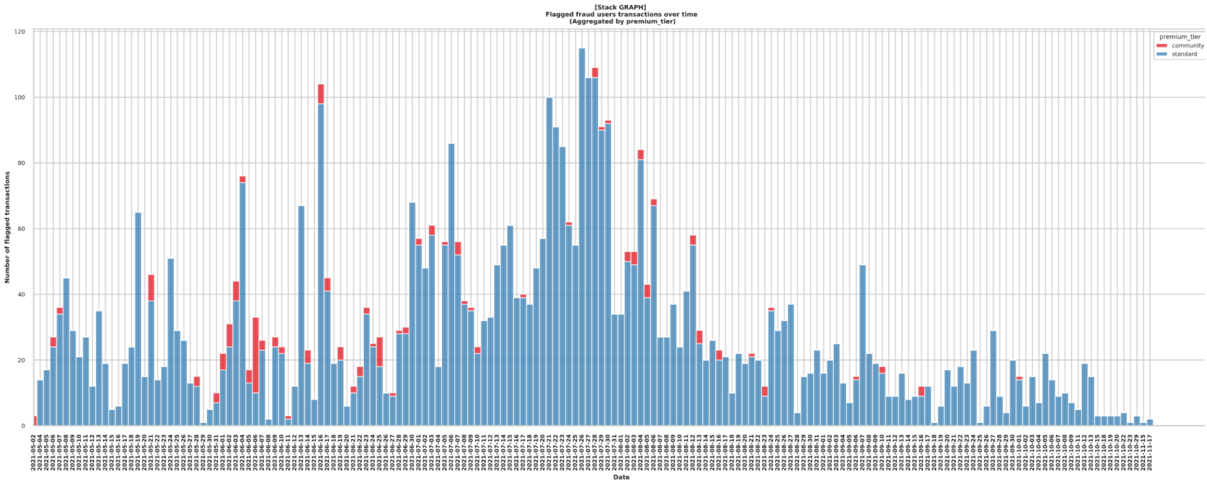
Analysis : Most flagged transactions that occurred for every event_kind == “Exchange” were done with GBP, CHSB and ETH. (FromCurrency MOST USED)

ToCurrency most used in Event_kind that is “Exchange”



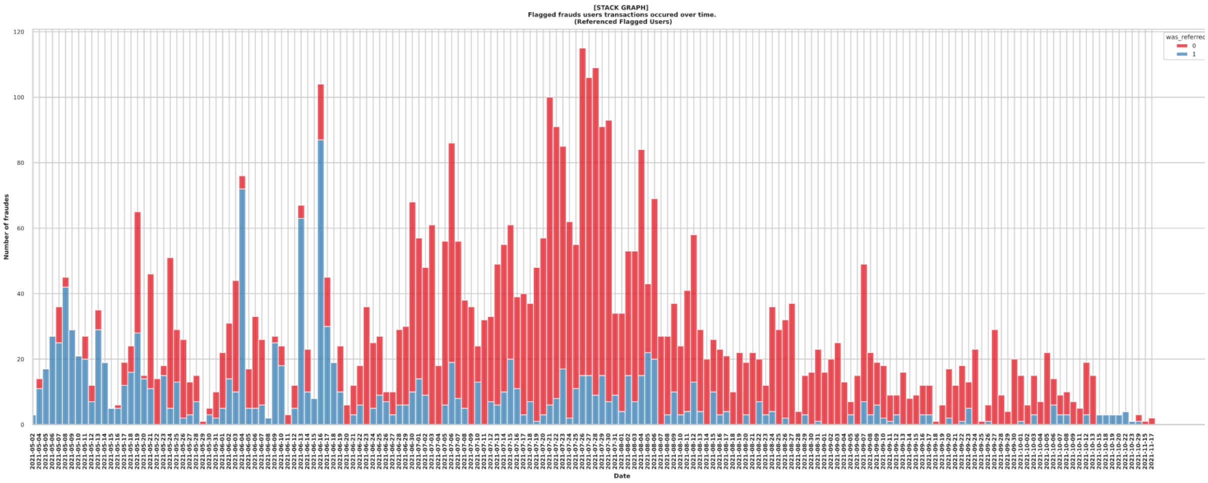
Analysis : Most flagged transactions that occurred for every event_kind == “Exchange” were done with ETH, BTC, CHSB. (ToCurrency MOST USED)

Premium tier



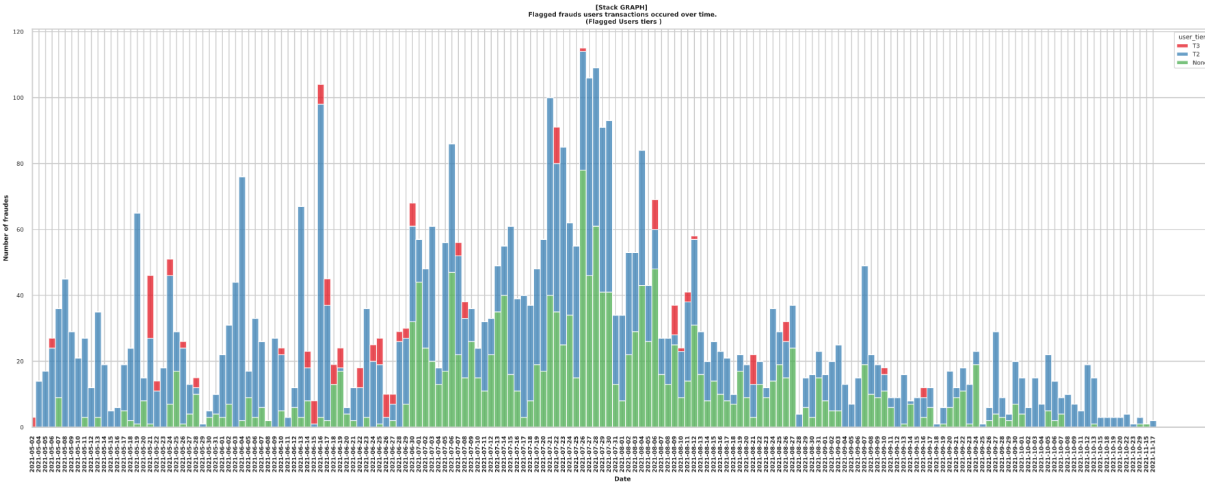
Analysis : Most flagged transactions that occurred were done by users with a standard profile.

Was_referred



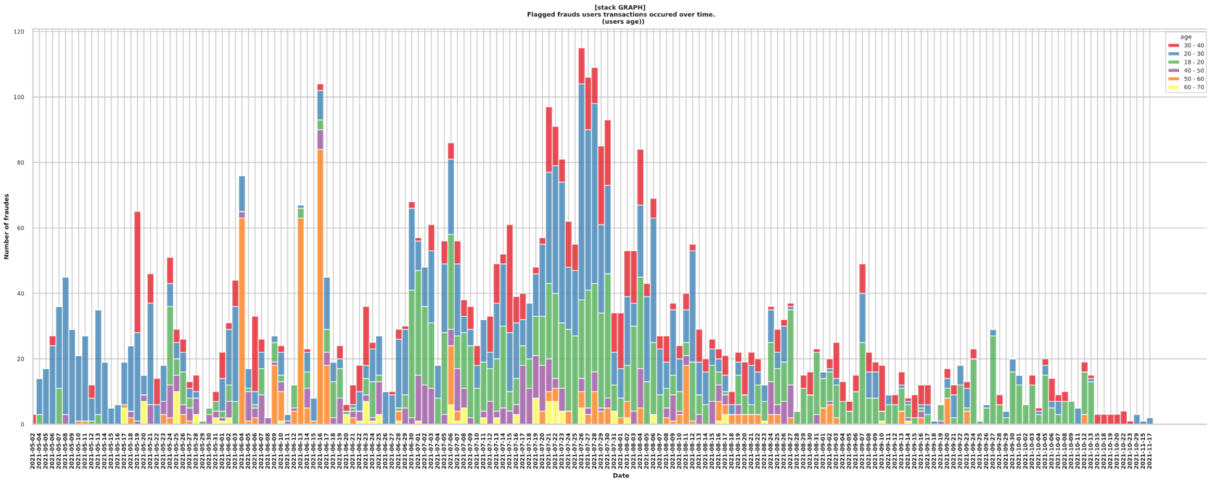
Analysis : Most flagged transactions that occurred were done by users that were not referred.

User Tier



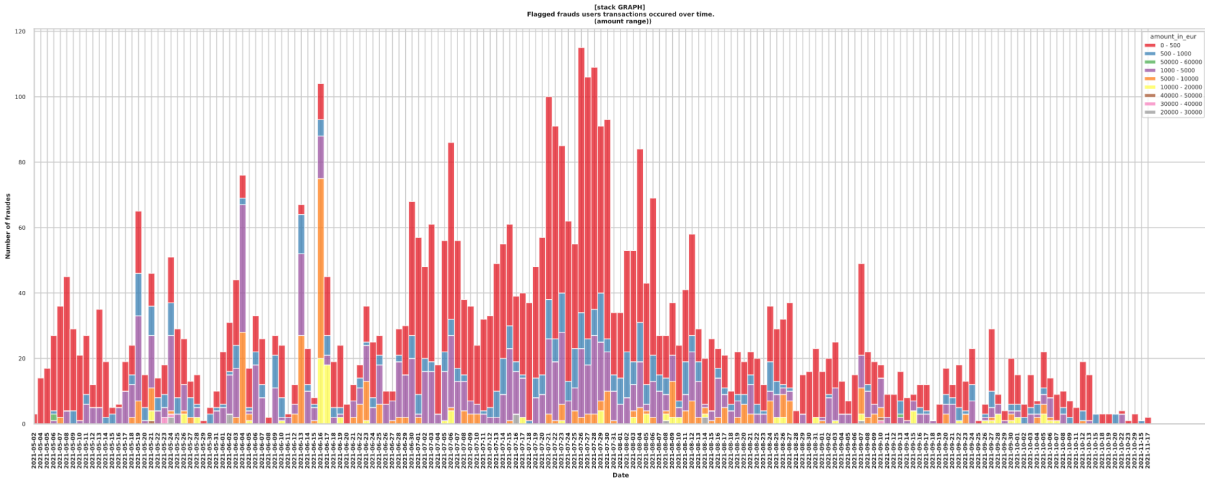
Analysis : Most flagged transactions that occurred were done by users that had no tier or Tier 2.

User Age



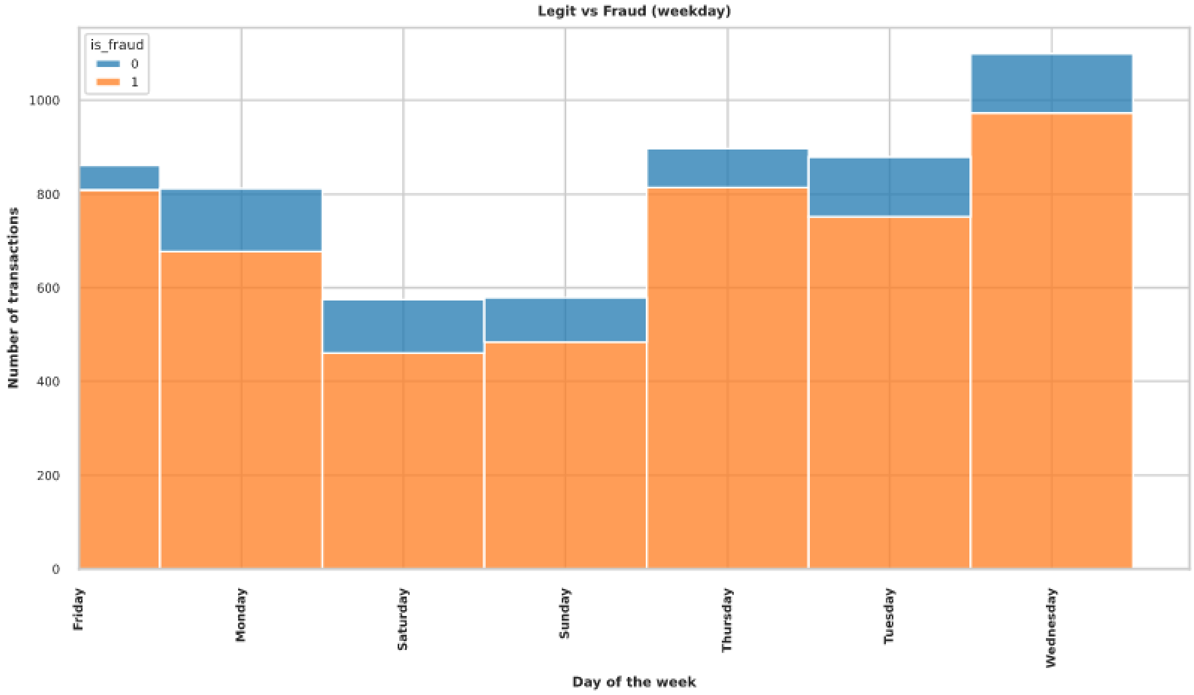
Analysis : Most flagged transactions that occurred were done by users had age between 18-20 and 20-30

Transaction amount range



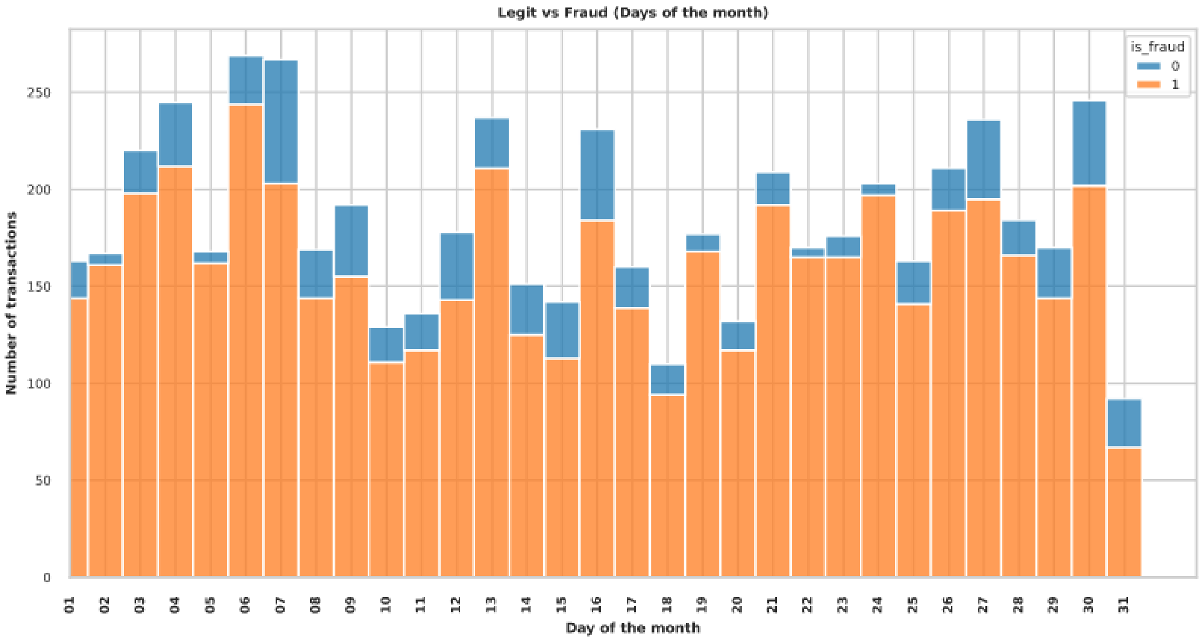
Analysis : Most flagged transactions that occurred were in the range of 0-500.

Days of the week (fraud vs legitimate)



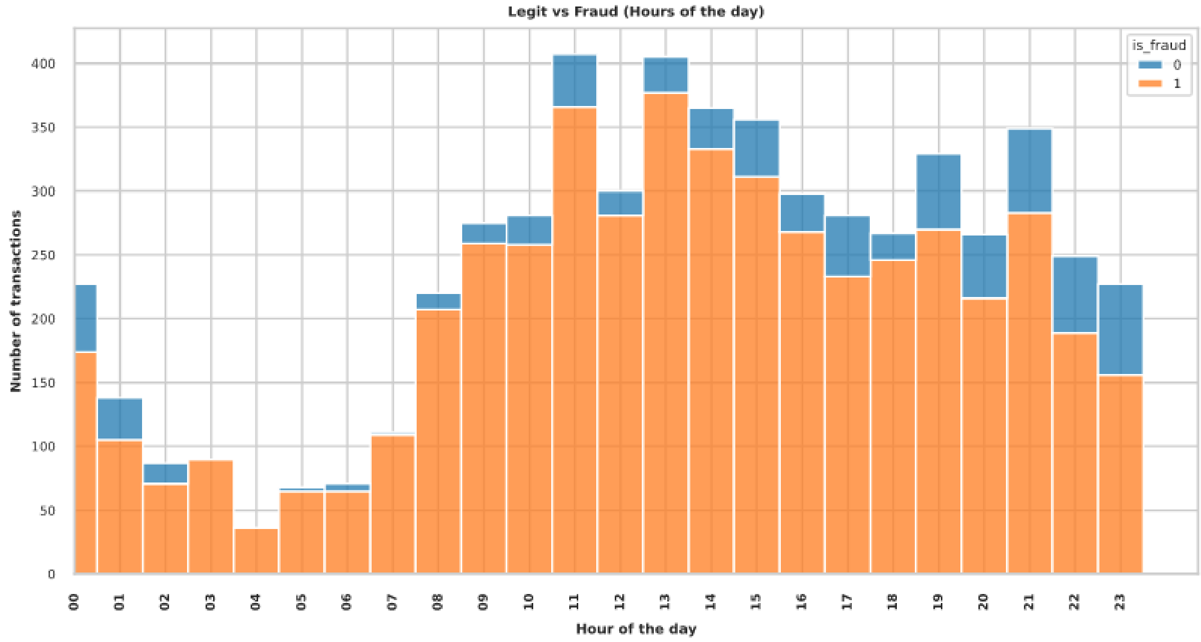
Analysis : Legitimate users normally have fewer transactions on Friday and Thursday when compared with fraudulent users. Wednesday is when more fraudulent behaviour is prevalent.

Days of the month (fraud vs legitimate)



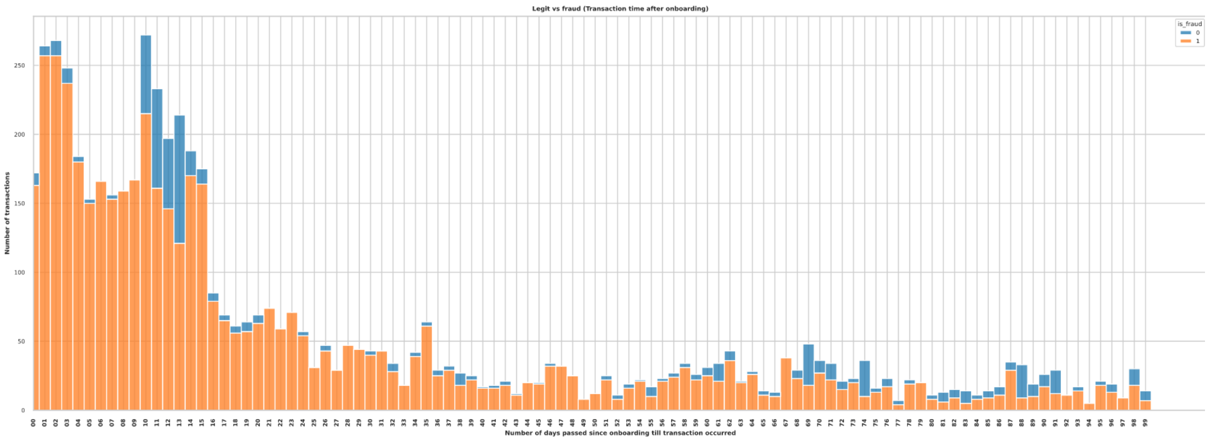
Analysis : Legitimate users transactions have lower amounts in the following ranges of days of every month : 01-06 and 17-26. The 6th of every month has the highest probability of fraud.

Transaction HOURS (fraud vs legitimate)



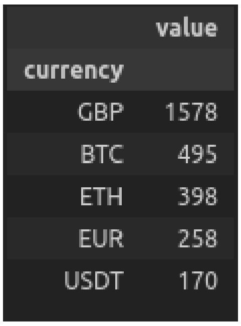
Analysis: Legitimate users are more active towards the night, while fraudulent users have more activity during the morning and afternoon. The peak hours for fraud is at 11h and 13h.

Time after onboarding (fraud vs legitimate)

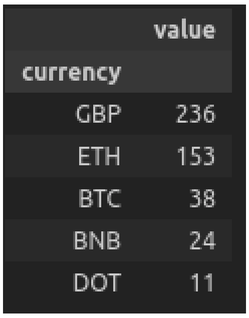


Analysis: Normally fraudulent users have very high activity during the first 9 days after their onboarding is accepted, while legit users usually have higher activity after the first 9 days. This seems to be the best way to distinguish them.

5th most used currencies outside of exchange (fraud vs legitimate)



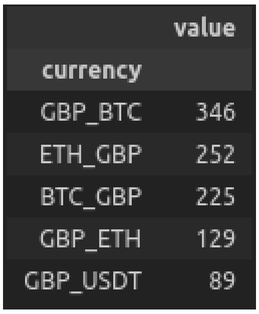
	value
currency	
GBP	1578
BTC	495
ETH	398
EUR	258
USDT	170



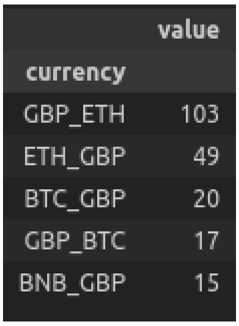
	value
currency	
GBP	236
ETH	153
BTC	38
BNB	24
DOT	11

Analysis: Fraudulent users have a high rate of fiat currencies (eur and usdt) usage when compared with legitimate users.

5th most used conversions in exchange (fraud vs legitimate)



	value
currency	
GBP_BTC	346
ETH_GBP	252
BTC_GBP	225
GBP_ETH	129
GBP_USDT	89



	value
currency	
GBP_ETH	103
ETH_GBP	49
BTC_GBP	20
GBP_BTC	17
BNB_GBP	15

Analysis: Fraudulent users have a high rate of GBP_BTC usage when compared with legitimate users.