Credit Card Fraud

Project 9

Daniel Molina Valencia
Michael Elkabas
Mohammed Ali
Olamide Usman
Yi Wen

MAIN PROBLEM STATEMENT

What is it?

- Examine and analyze if different variables such as **date, time, cardholder and merchant locations, cardholder’s age, amount, category and merchant** are determining factors and can be used to predict credit card fraudulent transactions

Who it benefits?

- Banks, consumers, anyone who owns a credit card.

Which domain or industry?

- Banking/Financial Services

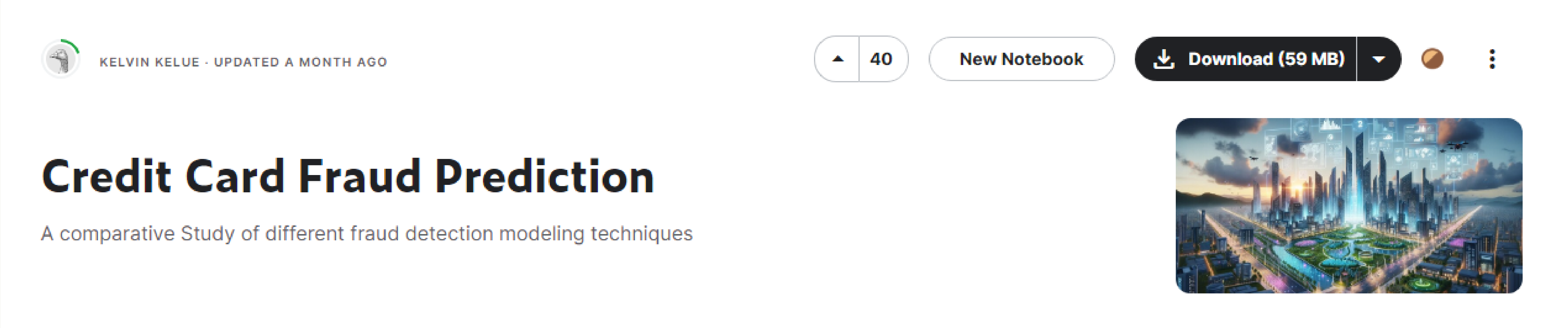
ANALYSIS QUESTIONS

- What time or month is fraud occurring more
- Relation between cardholder and merchant locations in fraudulent transactions
- Distribution of the fraudulent transaction amounts
- Average age of credit card fraud victim
- The highest number of frauds occurring in each category
- Compare the amount of fraud per merchant.

Why these questions

Credit Card fraud is a problem that we all face in our day to day, so we were interested in identifying and understanding insights and factors behind this phenomenon that can help us predict future fraud and protect ourselves better against it.

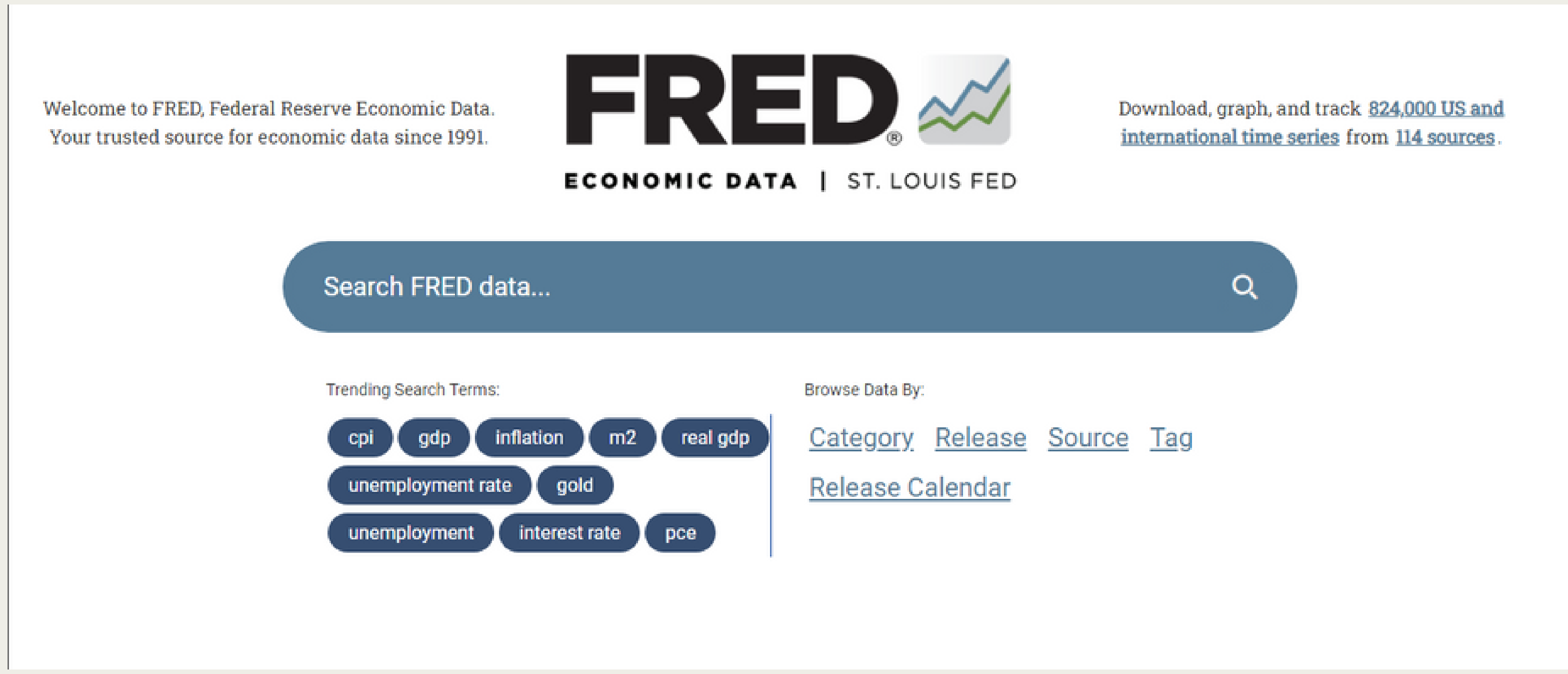
DATA SUMMARY


KELVIN KELUE · UPDATED A MONTH AGO

Credit Card Fraud Prediction

A comparative Study of different fraud detection modeling techniques

Source of data: Kaggle


Welcome to FRED, Federal Reserve Economic Data.
Your trusted source for economic data since 1991.

FRED ECONOMIC DATA | ST. LOUIS FED

Download, graph, and track 824,000 US and international time series from 114 sources.

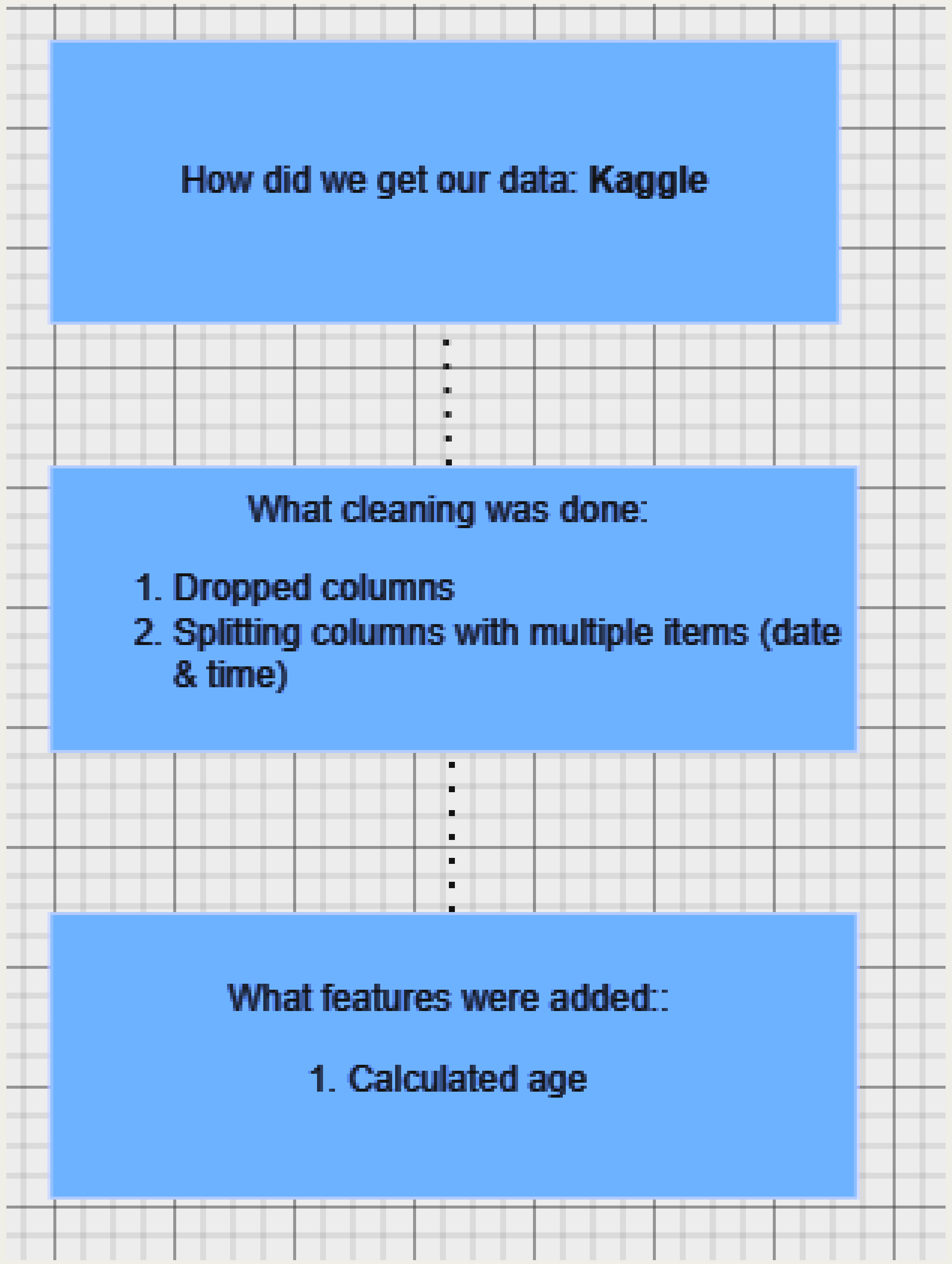
Source of data: FRED

<u>Why we chose this dataset</u>

- Many metrics to break down
- Easy to follow by our audience
- Relates to all of us
- Recency: 2020
- Over 500,000 rows, 22 columns

DATA PREPARATION PROCESS

Initial cleaning



How did we get our data: **Kaggle**

What cleaning was done:

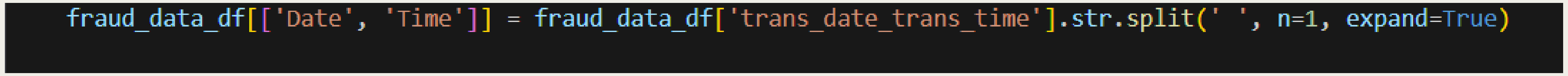
1. Dropped columns
2. Splitting columns with multiple items (date & time)

What features were added::

1. Calculated age

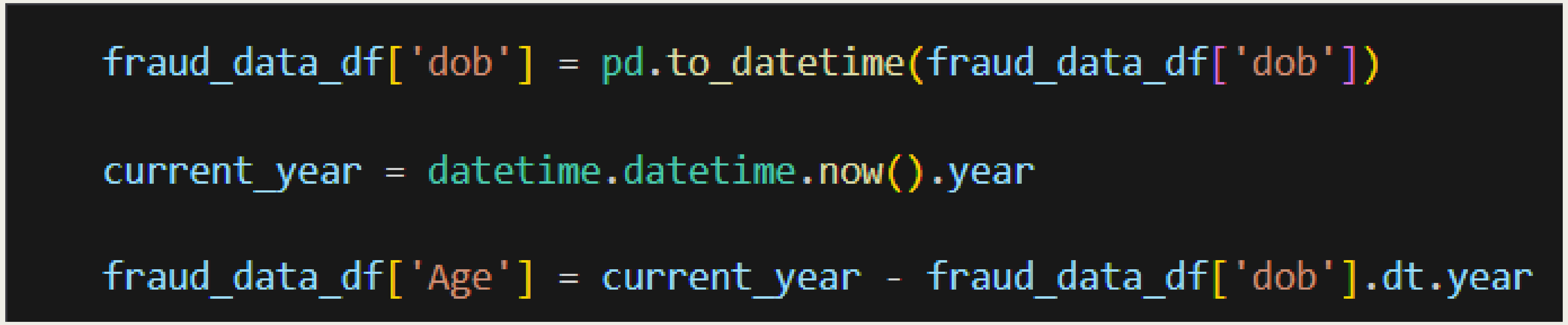
Code utilized in the initial cleaning

split trans_date_time



```
fraud_data_df[['Date', 'Time']] = fraud_data_df['trans_date_trans_time'].str.split(' ', n=1, expand=True)
```

calculate age based on current year



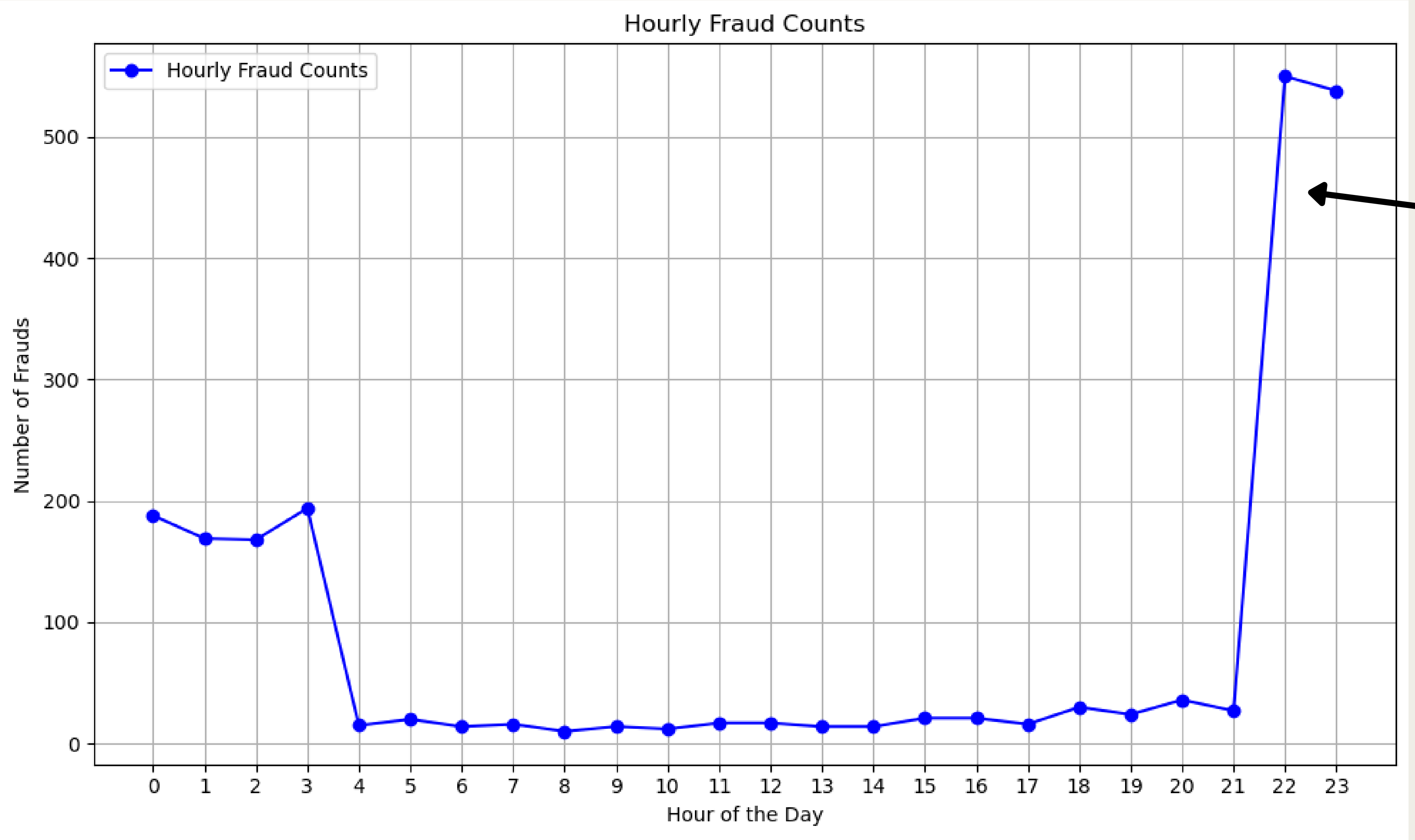
```
fraud_data_df['dob'] = pd.to_datetime(fraud_data_df['dob'])

current_year = datetime.datetime.now().year

fraud_data_df['Age'] = current_year - fraud_data_df['dob'].dt.year
```

Further data preparation

- Geoapify used to obtain merchant State and Country
- Create columns to calculate percentages
- Create column to compare the values of two other columns

RESULTS (DATE AND TIME)

What time is fraud occurring more at (morning or night)



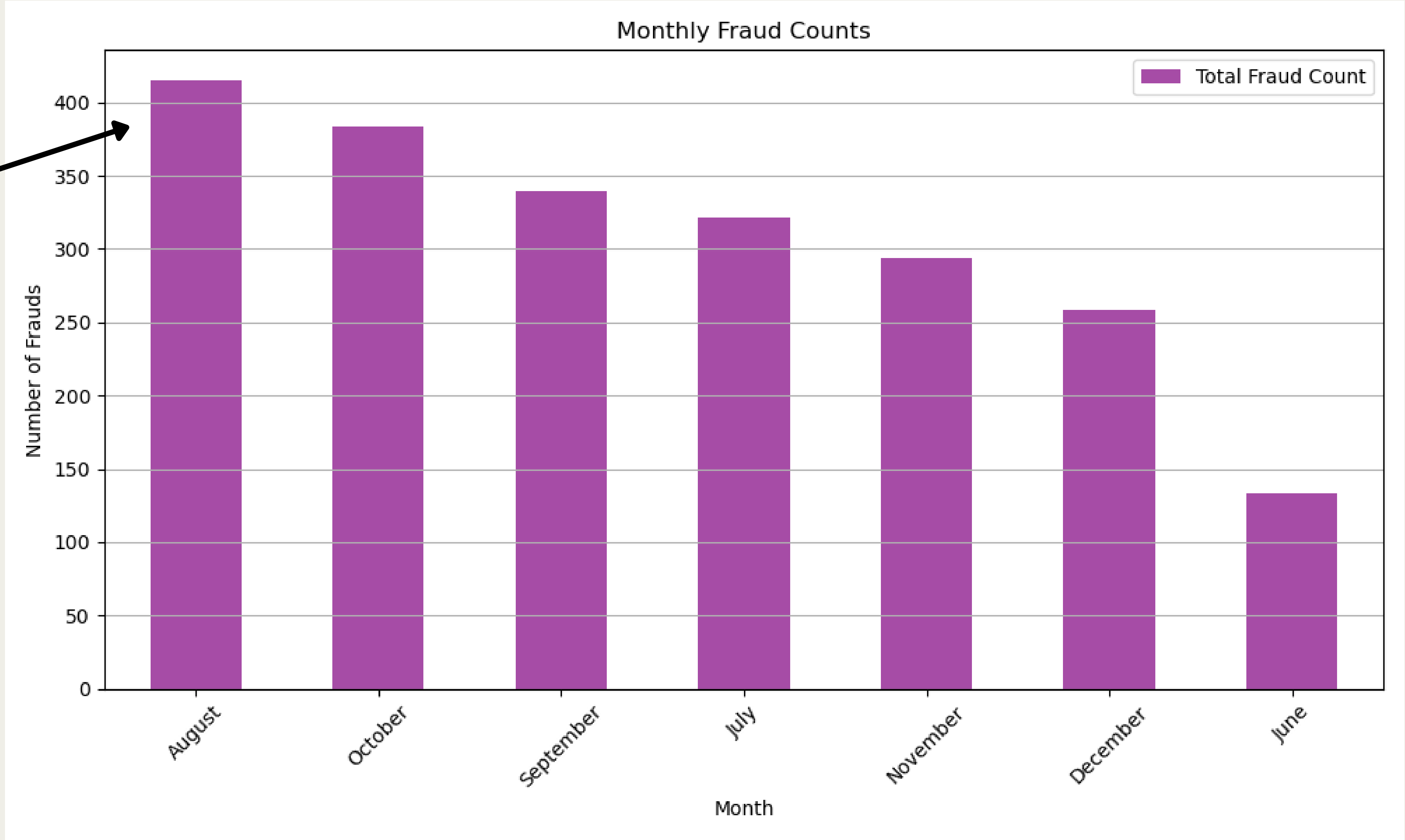
Highest

Line chart: Understand whether there is a trend.

Hours 21 - 23 have the most fraud. This seems like an accurate statement as fraudulent transactions take place when people cannot react as quickly to them.

What month is fraud happening the most in

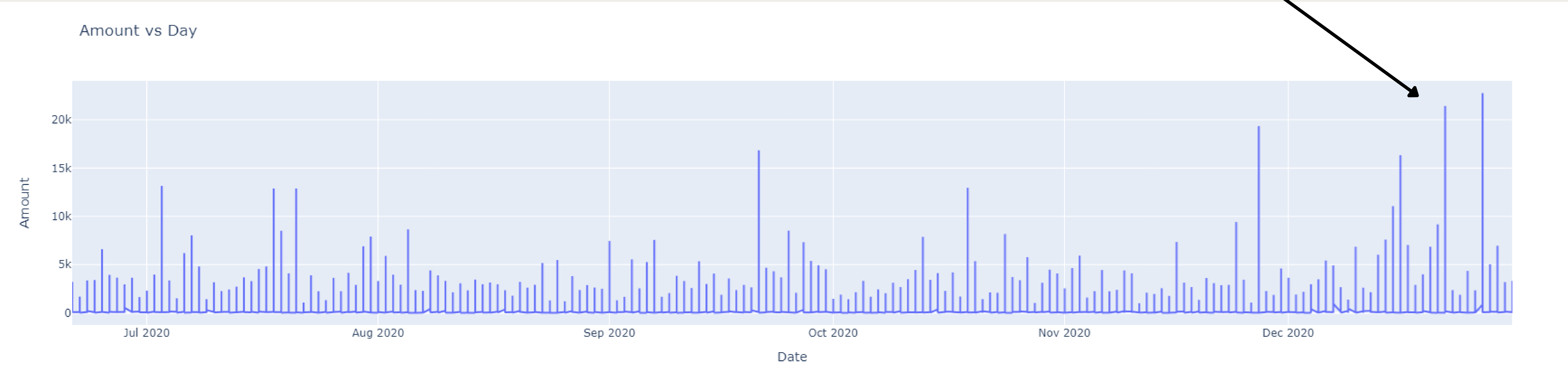
Highest



After comparing the graphs we can say that August has the most fraud. With over 400+ fraud counts during that amount.

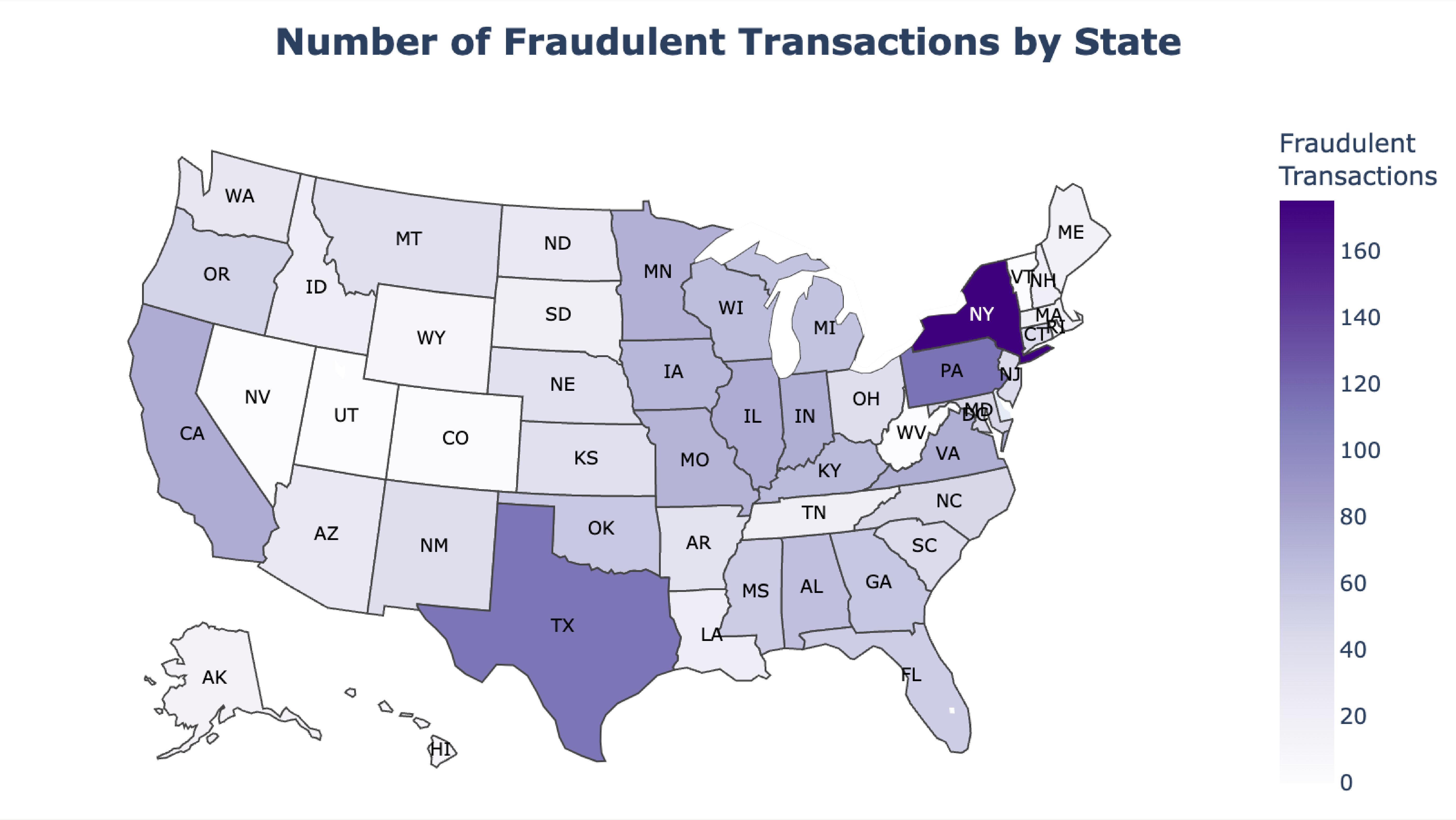
Highest

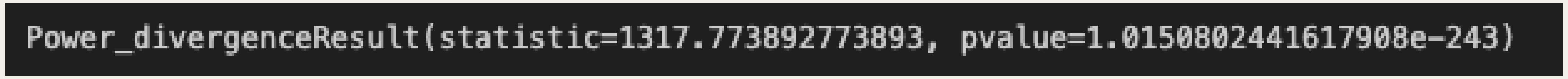
The amount($) of fraud per day



Although August has the months with the most fraud December has the most amount fraud.

RESULTS (LOCATION)

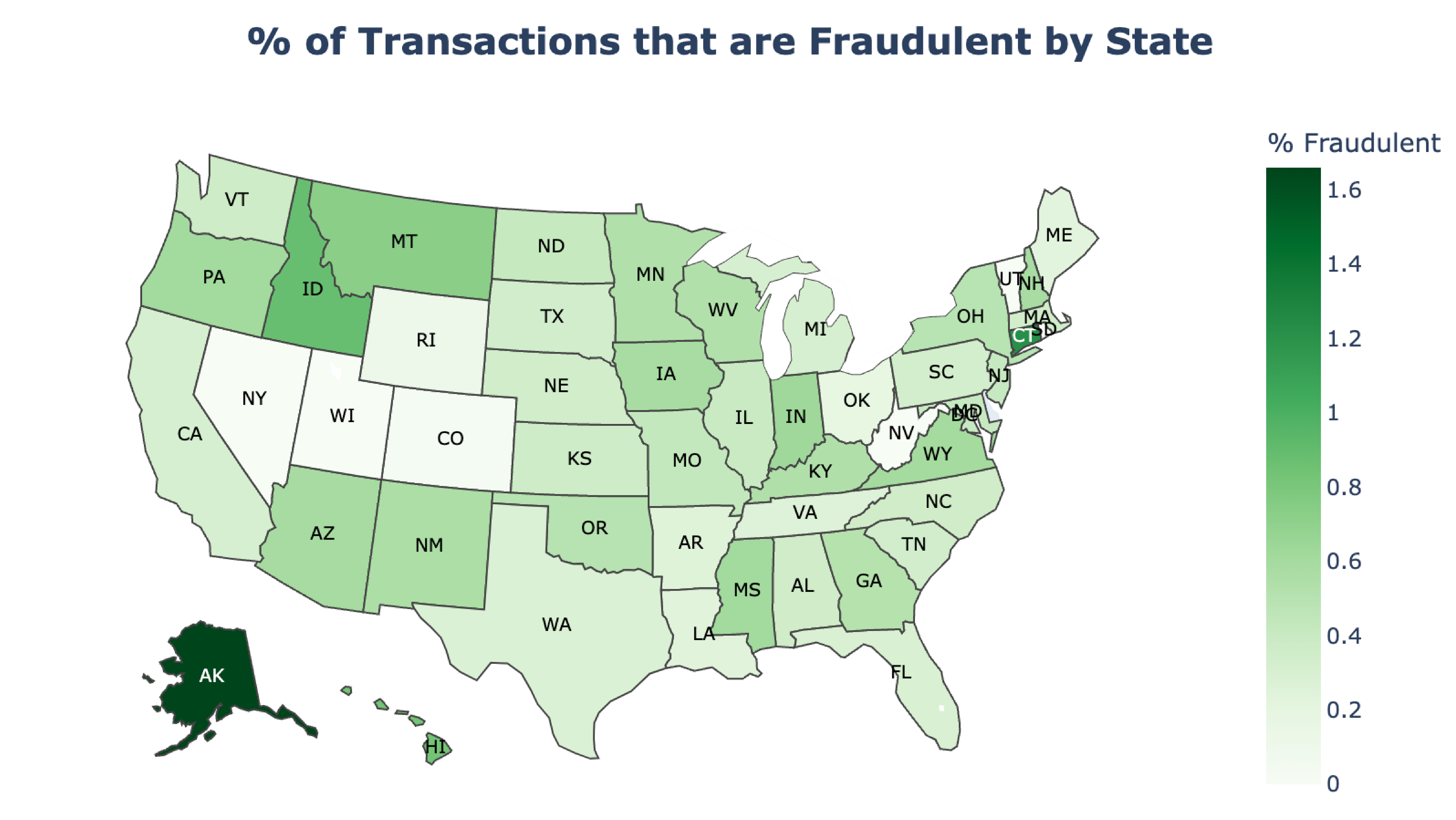




```
Power_divergenceResult(statistic=1317.773892773893, pvalue=1.0150802441617908e-243)
```

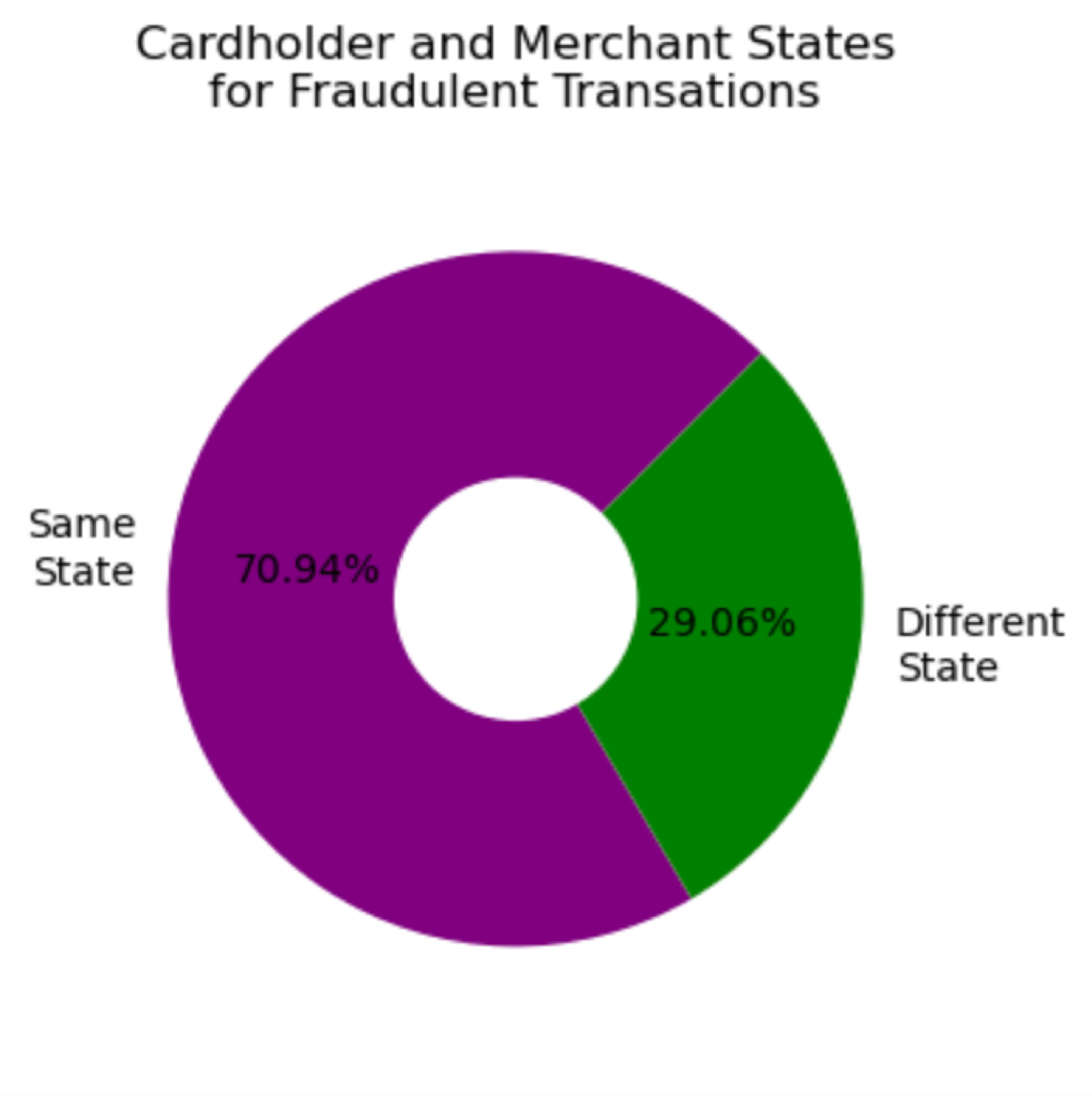
- New York (175), followed by Pennsylvania (114) and Texas (113) are the states with the highest number of fraudulent transactions.
- The p-value of the chi-square test confirmed that the number of fraudulent transactions are higher in specific states and not distributed equally among them.

RESULTS (LOCATION) CONT.



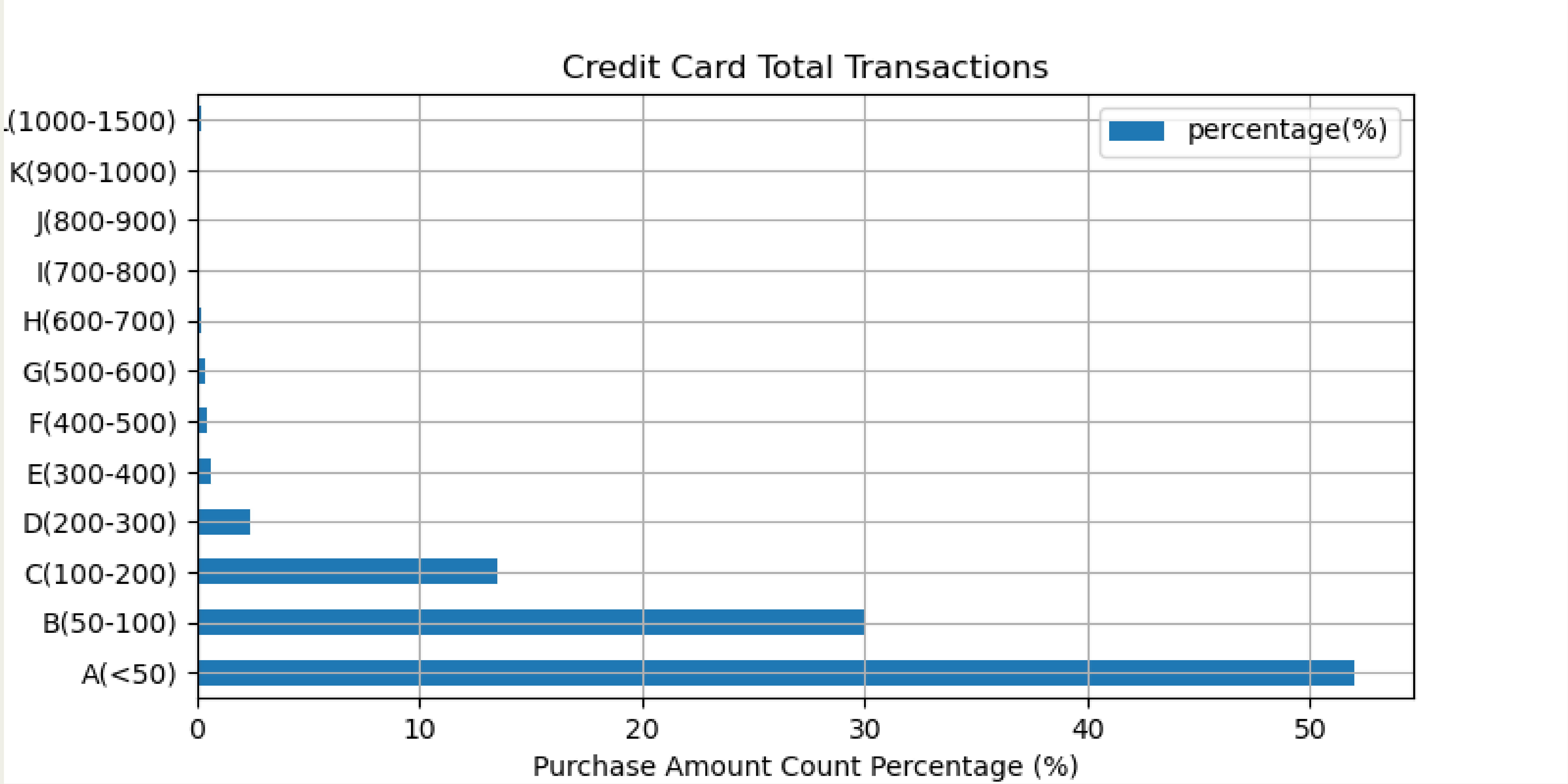
- In absolute terms, Alaska and Connecticut may not have many fraudulent transactions, but the fraudulent transactions represent more than 1% of their total number of transactions.

RESULTS (LOCATION) CONT.



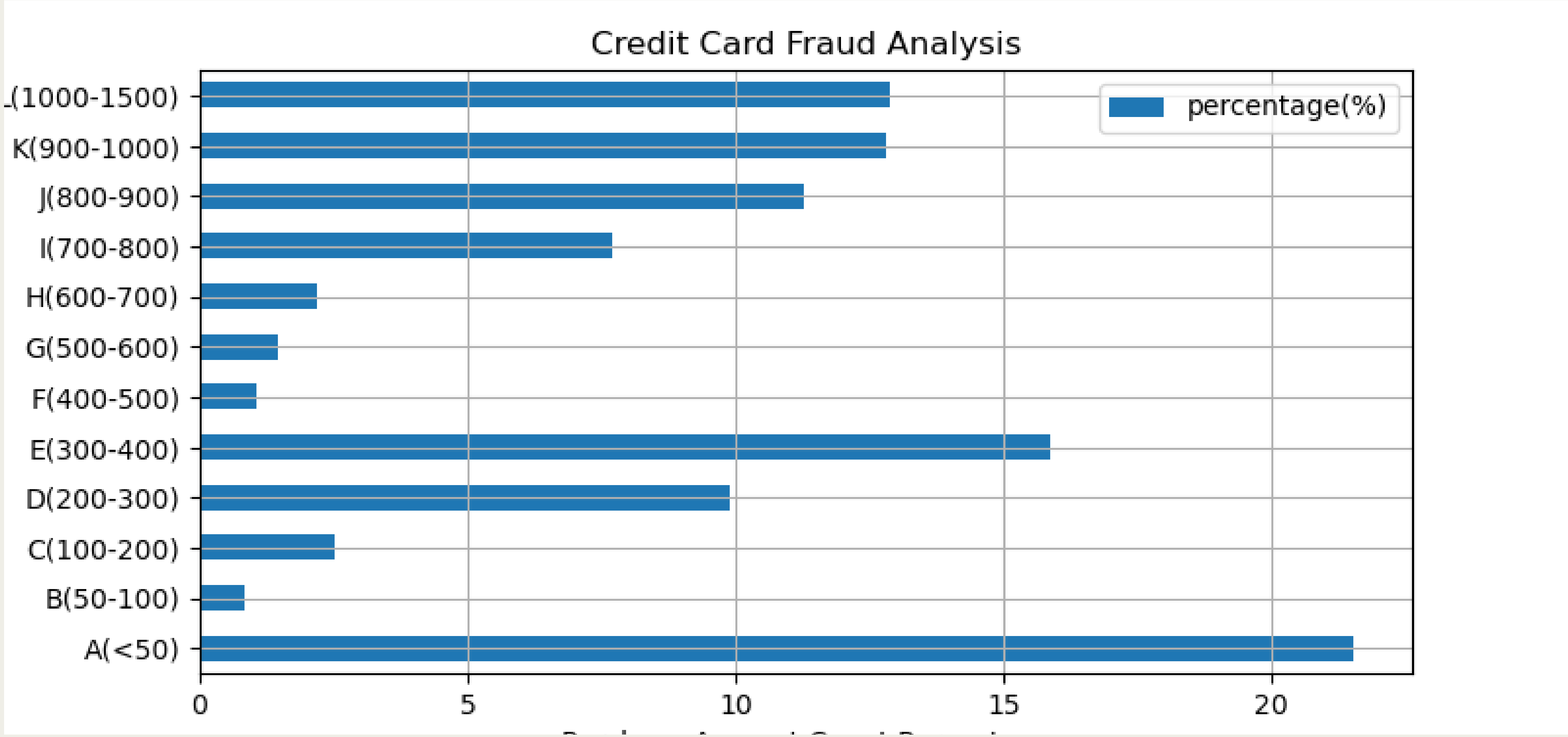
- Around 71% of the fraudulent transactions are performed with merchants located in the same State where the cardholder is located.

RESULTS (AMOUNT)



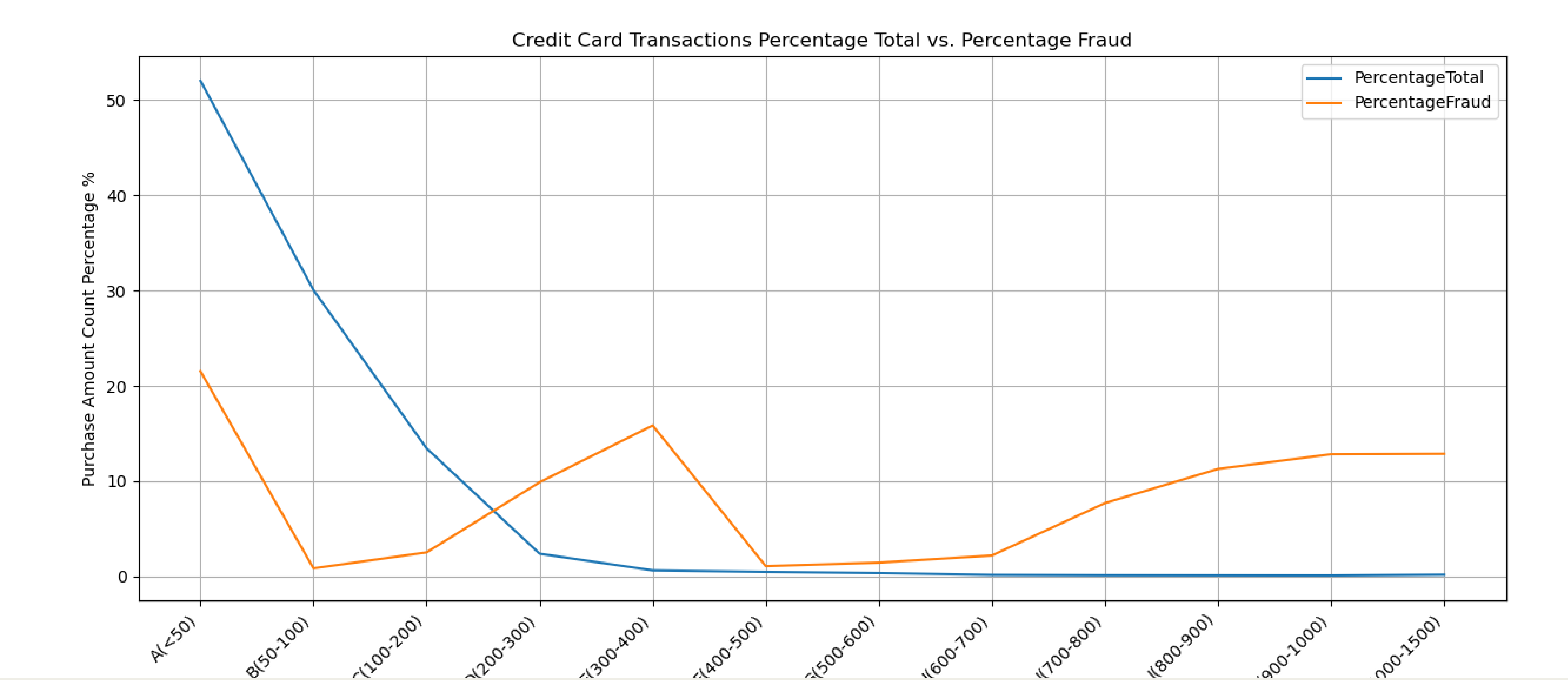
For total transactions, the transaction count percentage decreases with purchase amount increases.

RESULTS (AMOUNT)



For fraudulent transactions, the transaction count percentage increases comparing to total transactions

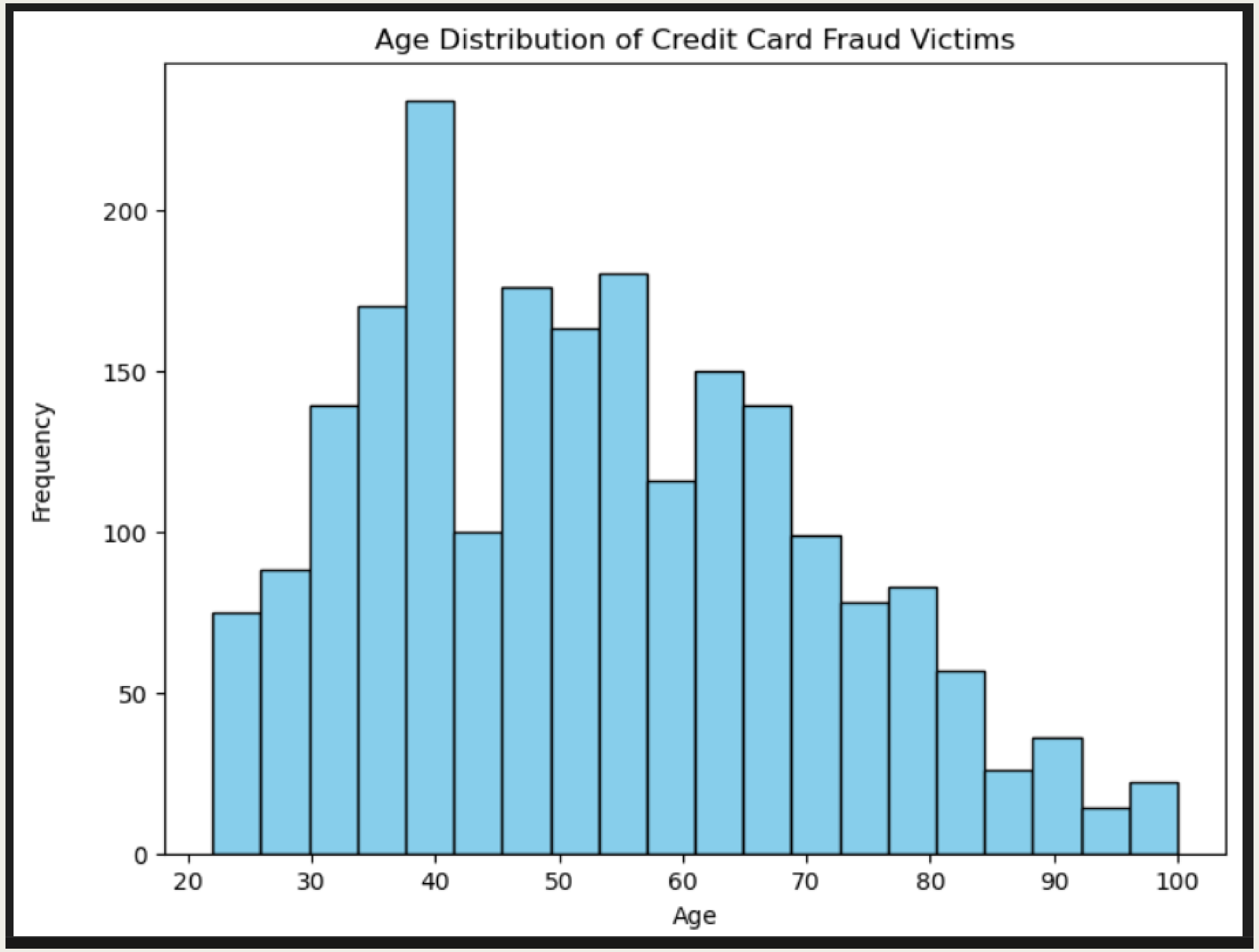
RESULTS (AMOUNT)



1. The percentage of purchase amount count of total transactions decreases when the purchase amount increases especially greater than $400
2. In contrast to total transactions, the percentage of purchase amount count of fraudulent transactions increases with the the purchase amount increases especially at $300-400 and greater than $700

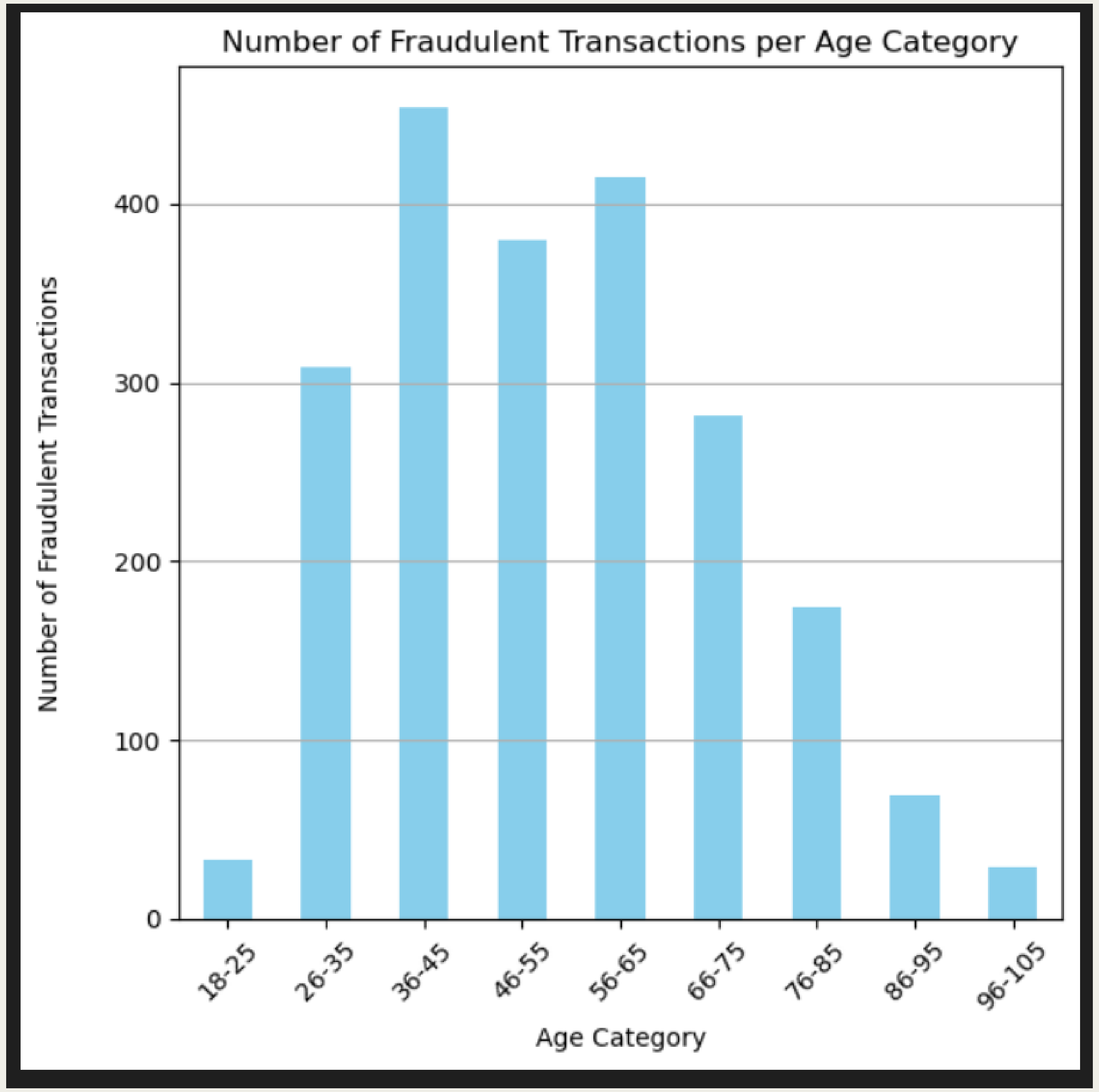
RESULTS (CARDHOLDER'S AGE)

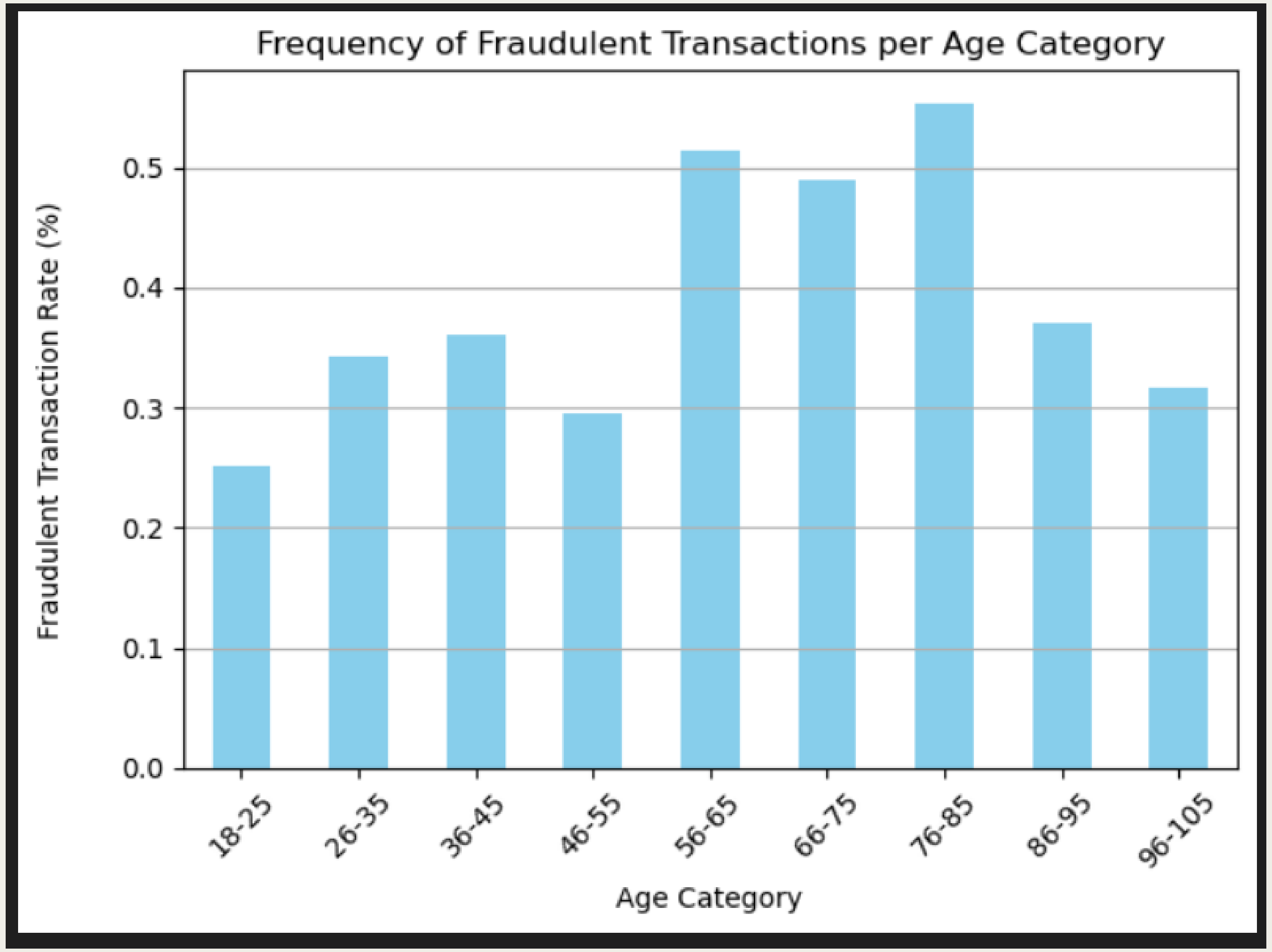
Average age of credit card fraud victim



Analyzing the average age of credit card fraud victims, shows a left-skewed distribution which suggests that there are relatively fewer instances of older individuals being victims of credit card fraud, as most victims being younger

Number of fraud transactions and frequency based on Age Category

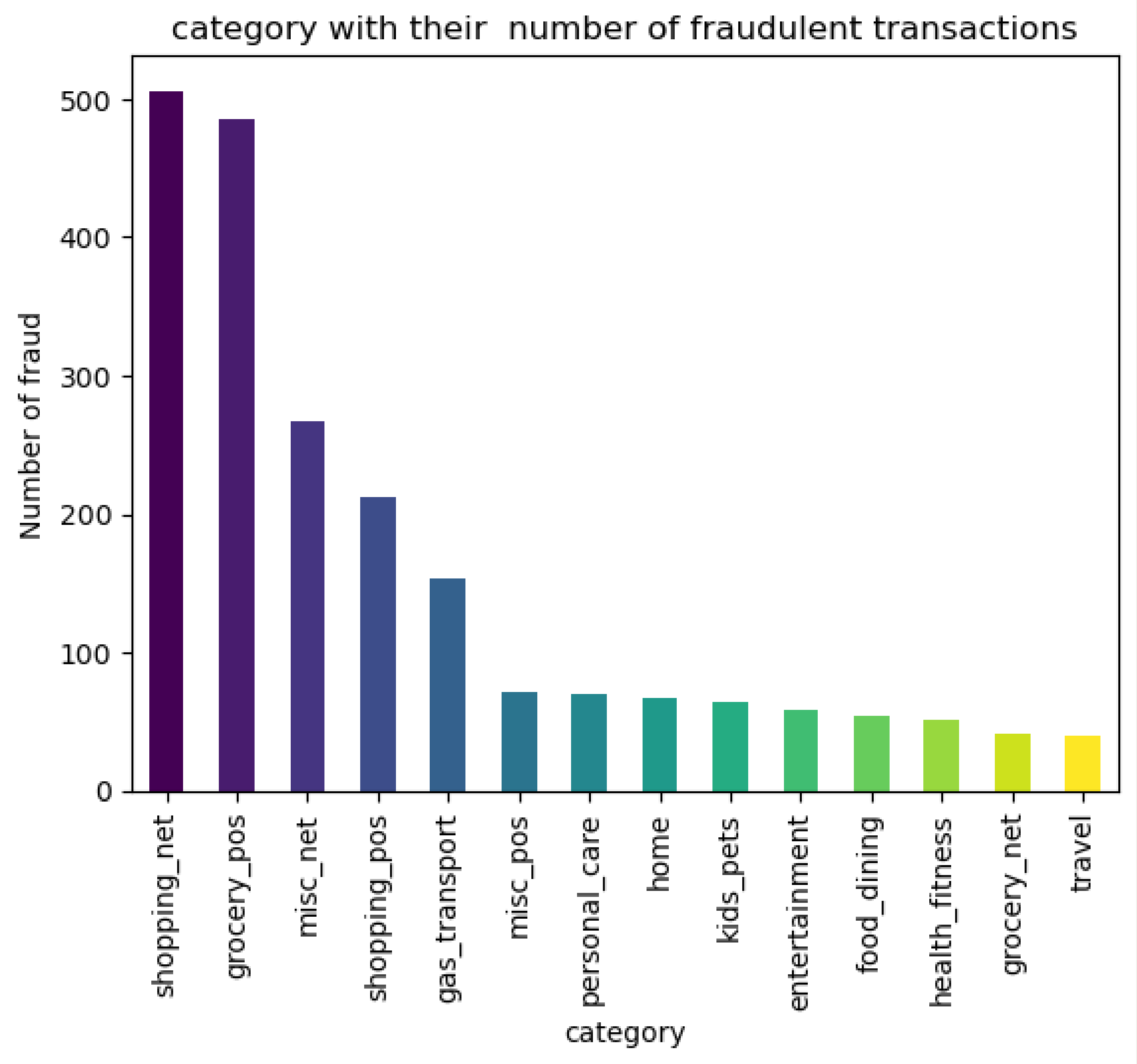




Even though a left-skewed distribution shows higher amounts of ***fraud transactions occur in yonger age catagories***, **the frequency is higher in elderly** suggesting that they are more likely targets.

When running a chi-square hypothesis test, chi-square value of **(115.096) indicated a higher deviation** in the data, however when looking at the **p-value (3.404172531724408e-21)** it is unlikely to be due to chance and therefore, we reject the null hypothesis

RESULTS (CATEGORY)



Bar chart showing number of fraud per category. the top 3 category with the highest incident of fraud is shopping net, grocery pos, misc net as represented in the bar chart . Pie chart is not applicable in this situation because it is less effective in communicating the difference between each category in relation to the slice size and theres is too many variables to be represented.

MAIN OUTCOMES

- We were able to identify how the majority of the fraudulent transactions fall under specific characteristics among the different variables analyzed. Therefore these variables can be used, with further analysis, to create a predictive model.

WHAT'S NEXT

- Analyze data set from previous or following years to confirm the trends found.
- Compare findings with findings of data sets from different countries (Amount of fraud, locations, age, etc.).
- Identify correlations with other variables.
- Combine the different factors and add it to the predictive model.

LIMITATIONS

- Data was limited to the year 2020. This data can be affected by the pandemic.
- Data did not include a whole year of information.
- Did not count with other characteristics of the cardholders such as income or education level.
- There was no data about the purchased item.

LEARNINGS

What went well:

- Experimented with new libraries and charts
- Use of APIs
- Coordinate with team members in a short time.

What we want to do better:

- Improve our merging abilities with Github
- Don’t bite more than we can chew
- Use a dataset that is less clean