# DATA ANALYTICS CAPSTONE PROJECT

ZAID FAKHRUDIN

# TABLE OF CONTENTS

01 Problem Statement

02 Data Cleaning & Preparation

03 Exploratory Analysis

04 Findings & Recommendation

# CREDIT CARD FRAUD ANALYSIS FOR UNITED STATES IN 2020

## Problem Statement:

With the advancements of technology, digital and online banking has now been made possible for the efficiency and convenience of its users to make transactions anywhere and anytime. However, the downside of having your credit card information stored in the cloud is that it can be easily accessed by third parties for credit card fraud. This project aims to investigate and examine how various factors relate to fraud.

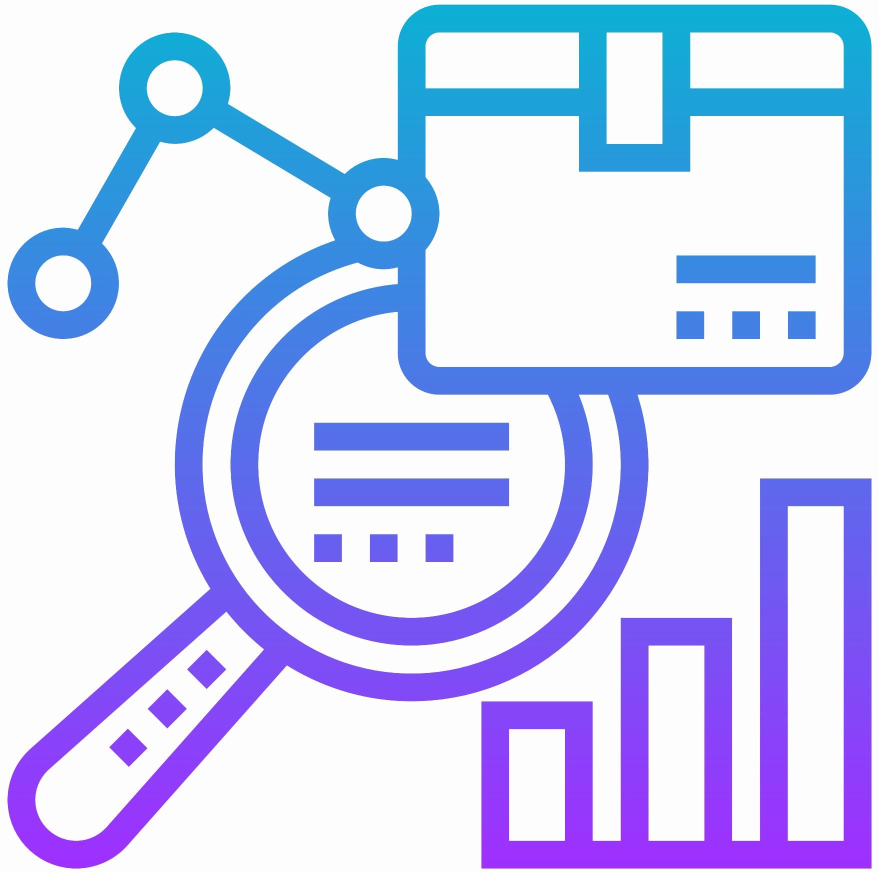
# DATASET

This dataset is downloaded from Kaggle

https://www.kaggle.com/datasets/kartik2112/fraud-detection

It contains data of credit card transactions with 22 columns that include:

- trans_date_trans_time
- cc_num
- merchant
- category
- amt
- first_name
- last_name
- gender
- street
- city
- state
- zip
- lat
- long
- city_pop
- job
- date_of_birth
- transaction_number
- unix_time
- merch_lat
- merch_long
- is_fraud



# DATA TYPE & DESCRIPTION

| Variable | Type | Description |
|---|---|---|
| trans_date_trans_time | date & time | Transaction date and time |
| cc_num | num | Credit card number |
| Merchant | string | Merchant name |
| Category | string | Category of product sold by merchant |
| Amt | num | Transaction amount |
| First_name | String | Credit card holder’s first name |
| Last_name | String | Credit card holder’s last name |
| Gender | Char | Credit card holder’s gender |
| Street | String | Credit card holder’s street address |
| City | String | City credit card holder lives in |
| State | String | State credit card holder lives in |

| Variable | Type | Description |
|---|---|---|
| Zip | Num | Zip code/postal code |
| Lat | String | Address Latitude |
| Long | string | Address Longitude |
| City_pop | Int | City population |
| Job | String | Credit card holder’s profession |
| Date_of_birth | Date | Credit card holder’s date of birth |
| Transaction_number | String | Transaction number |
| Unix_time | string | Unix time stamp |
| Merch_lat | String | Merchant latitude |
| Merch_long | string | Merchant longitude |
| Is_fraud | Boolean | Transaction legitimacy |

# DATA CLEANING & PREPARATION

**In Excel**

- checked for duplicates. None were found.
- ‘first’ and ‘last’ columns renamed to ‘***first_name***’ and ‘***last_name***’
- s/no column reformatted to start with ‘**1**’ instead of ‘**0**’
- cc_num column formatted from ***text*** to ***number***
- using ‘Find & Replace’, removed ‘fraud_’ from all merchant names in ‘merchant’ column
- Replaced ‘0’ = ‘**No**’ and ‘1’ = ‘**Yes**’ in ‘is_fraud’ column for easier visualisation
- ‘unix_time’ column has been excluded as it will not be used for this project

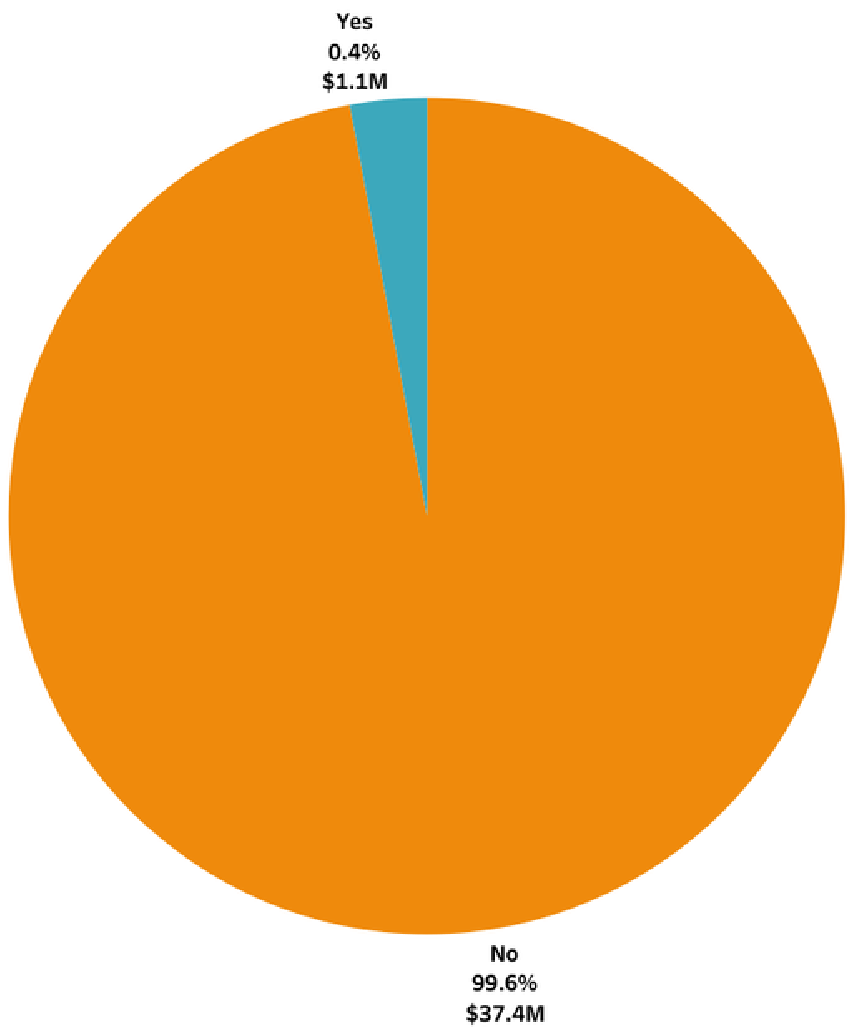
# EXPLORATORY ANALYSIS

https://public.tableau.com/app/profile/zaid.fakhrudin/viz/CreditCardFraudAnalysis_16667032867570/CreditCardFraudAnalysis?publish=yes

# OVERVIEW

As seen from the pie chart, fraudulent transactions only make up 0.4% of the total transactions.

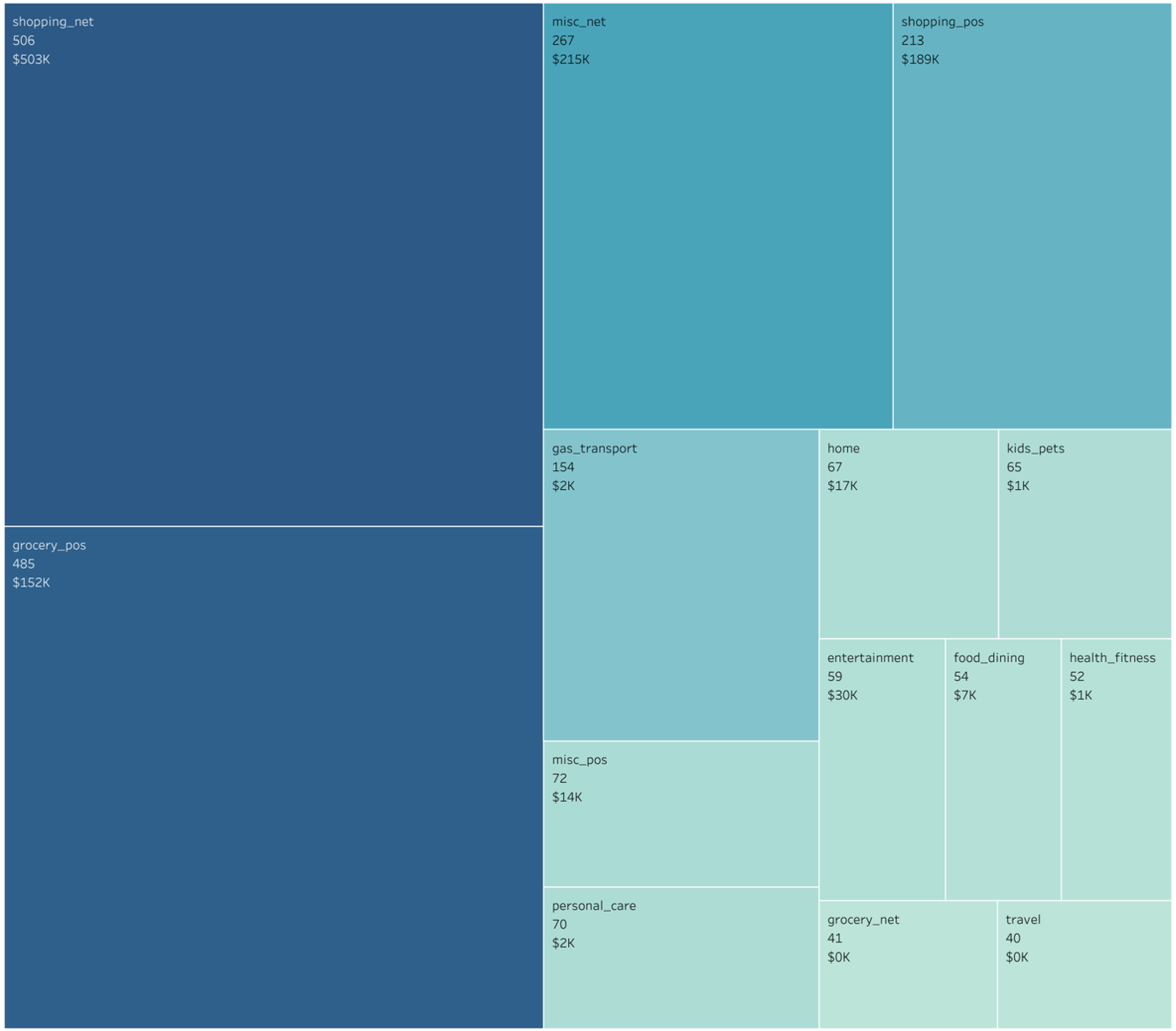
However despite it's miniscule percentage, it is still quite a sizeable amount; totaling up to $1.1M



## **Tree Map of fraud by category & amount**

Upon closer inspection of the tree map, it is discovered that 'shopping_net' takes up the bulk of fraud cases - 506 cases amounting to $503k ; making it the highest category, followed by groceries

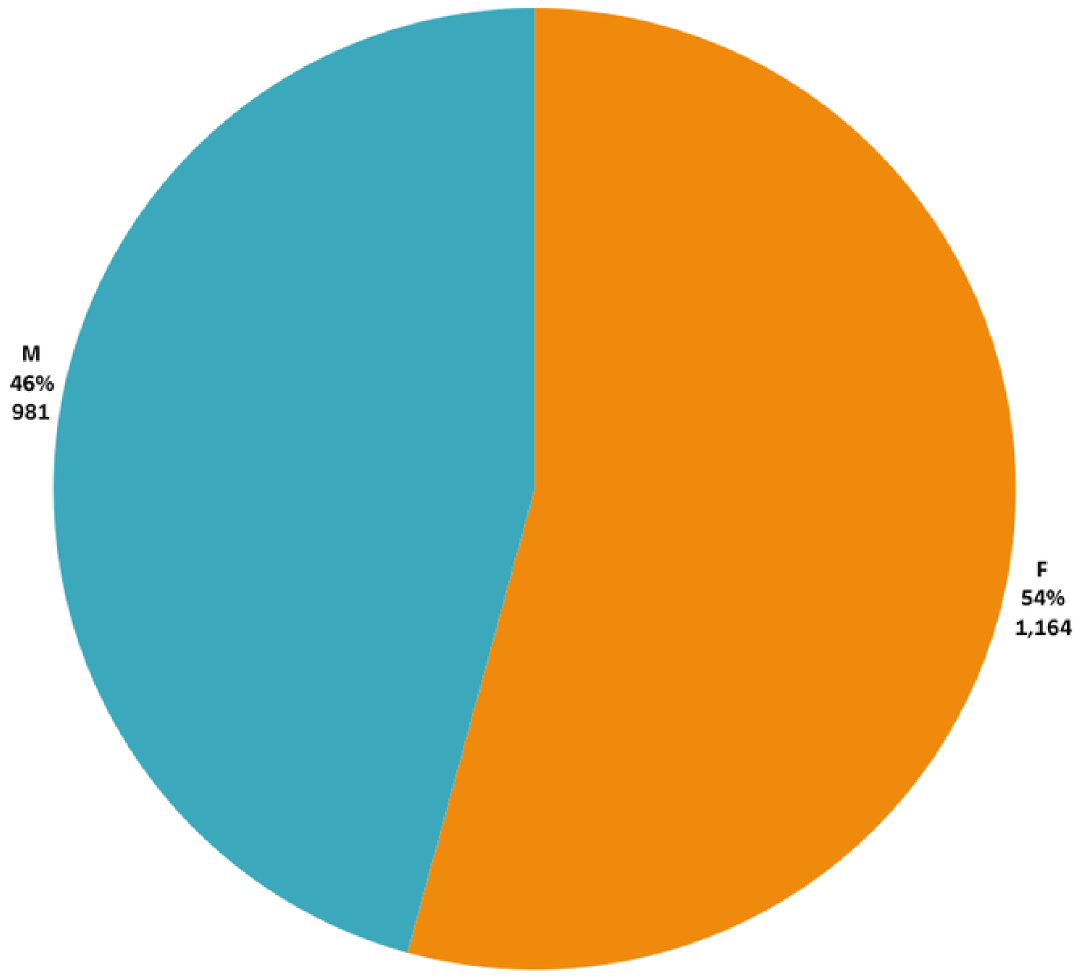
Very interesting!



## Gender pie chart

Next, let's take a look at the gender statistics

Numbers are almost evenly distributed; indicating gender has no direct correlation to a fraudulent transaction



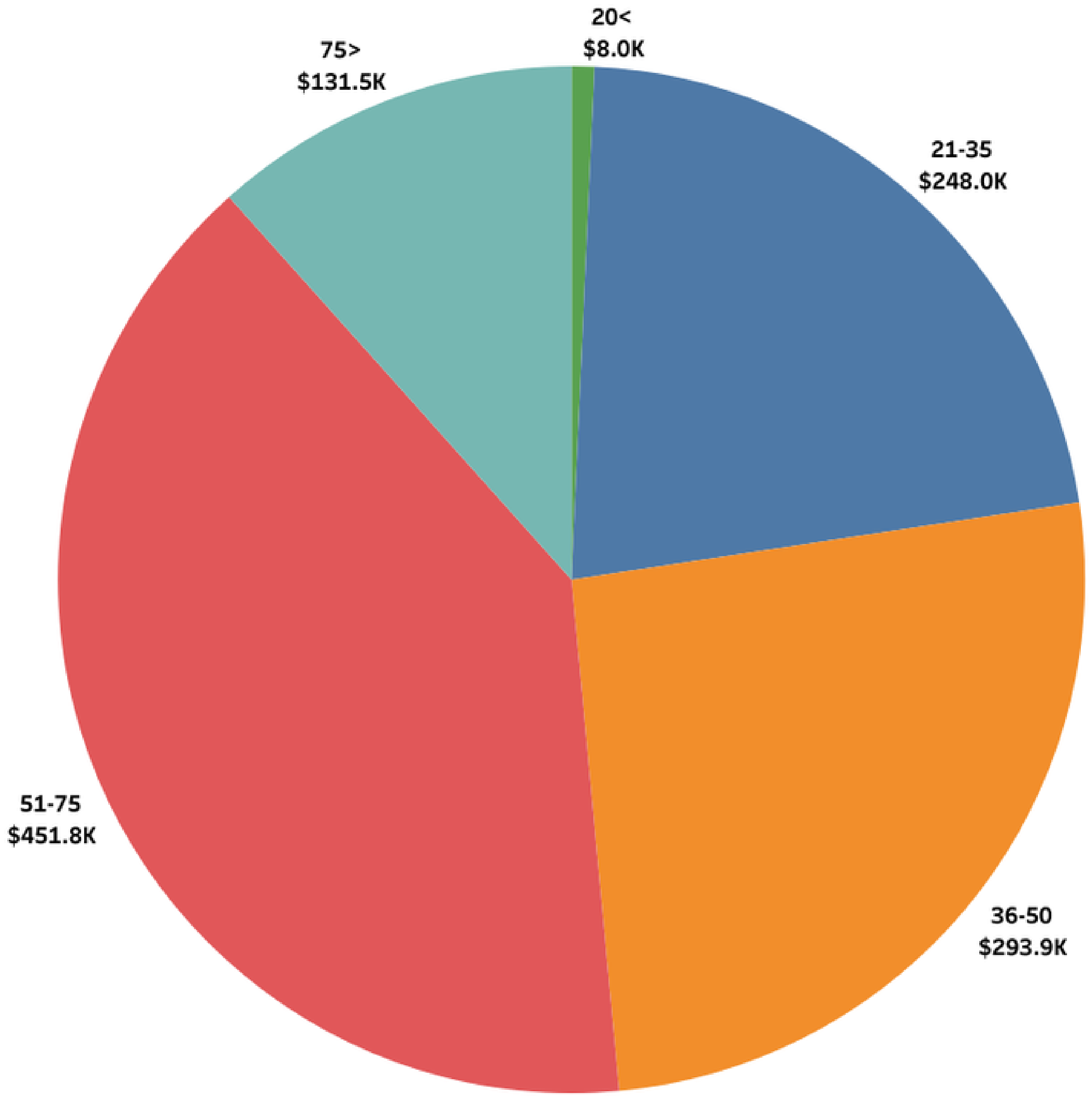
## Fraud transactions between the different age groups

Interesting discovery!

If we take a look at the age groups below 20 and that of 21-35, we can see that there is a big gap in amount ; $8k and $248k respectively

This could suggest that it is likely to happen to those 21yrs and above (min age requirement)

In contrast, highest fraud cases seem to happen to elderly aged between 51-75 at $451k

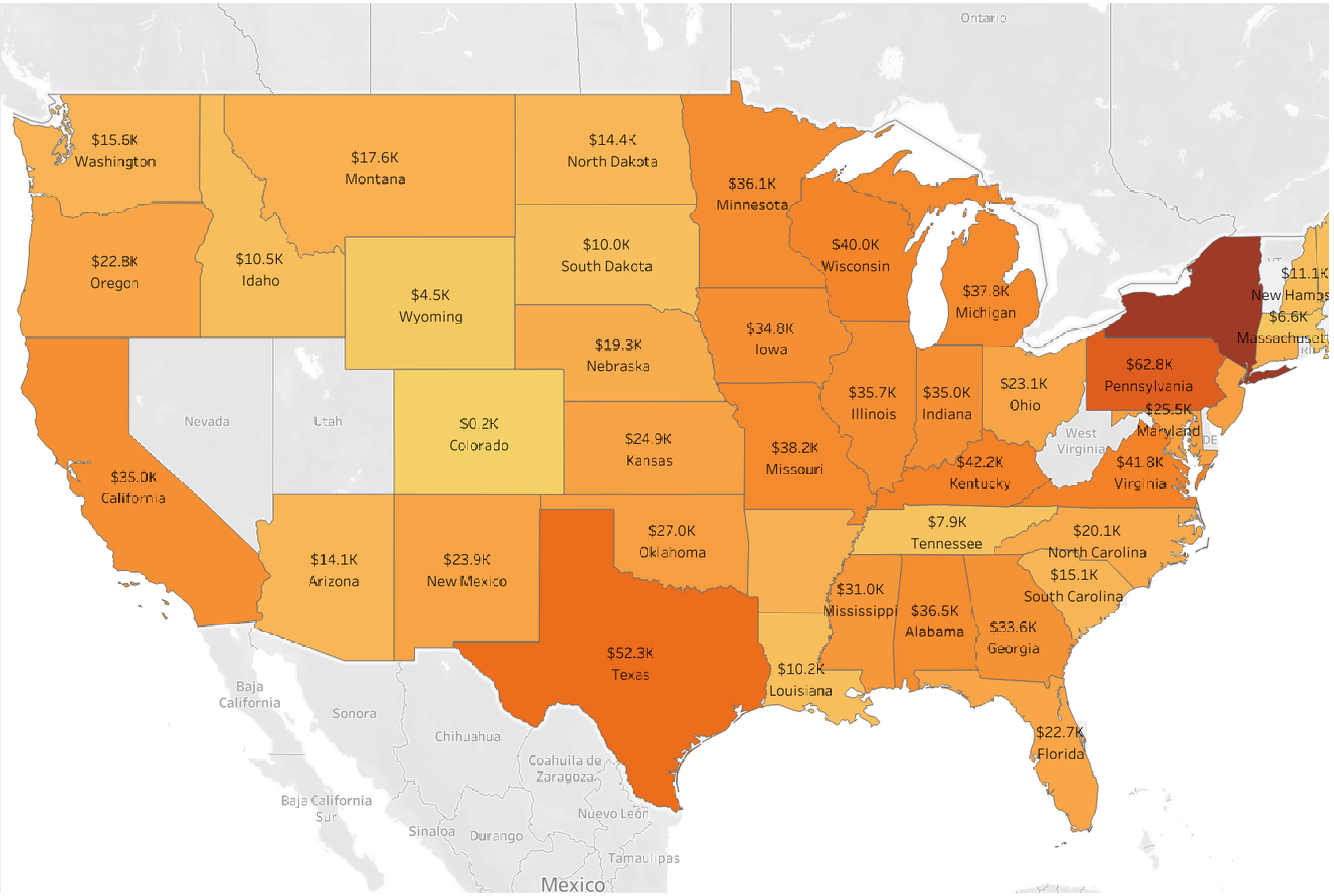


## Total Fraud by State

Now let's see if location is a factor in fraud

The map highlights the Top 3 States with highest fraudulent transactions as such:

**New York - $99.7k**
**Pennsylvania - $62.8k**
**Texas - $52.3k**



As it turns out, the states with highest fraud transactions are also one of the most populated states; **Texas**, **New York**, followed by **Pennsylvania**

*Fig 2.* calls attention to 'shopping' as the top category yet again. This is synonymous with the tree map shown previously in which 'shopping' also comes up on top. There is clearly a pattern here.

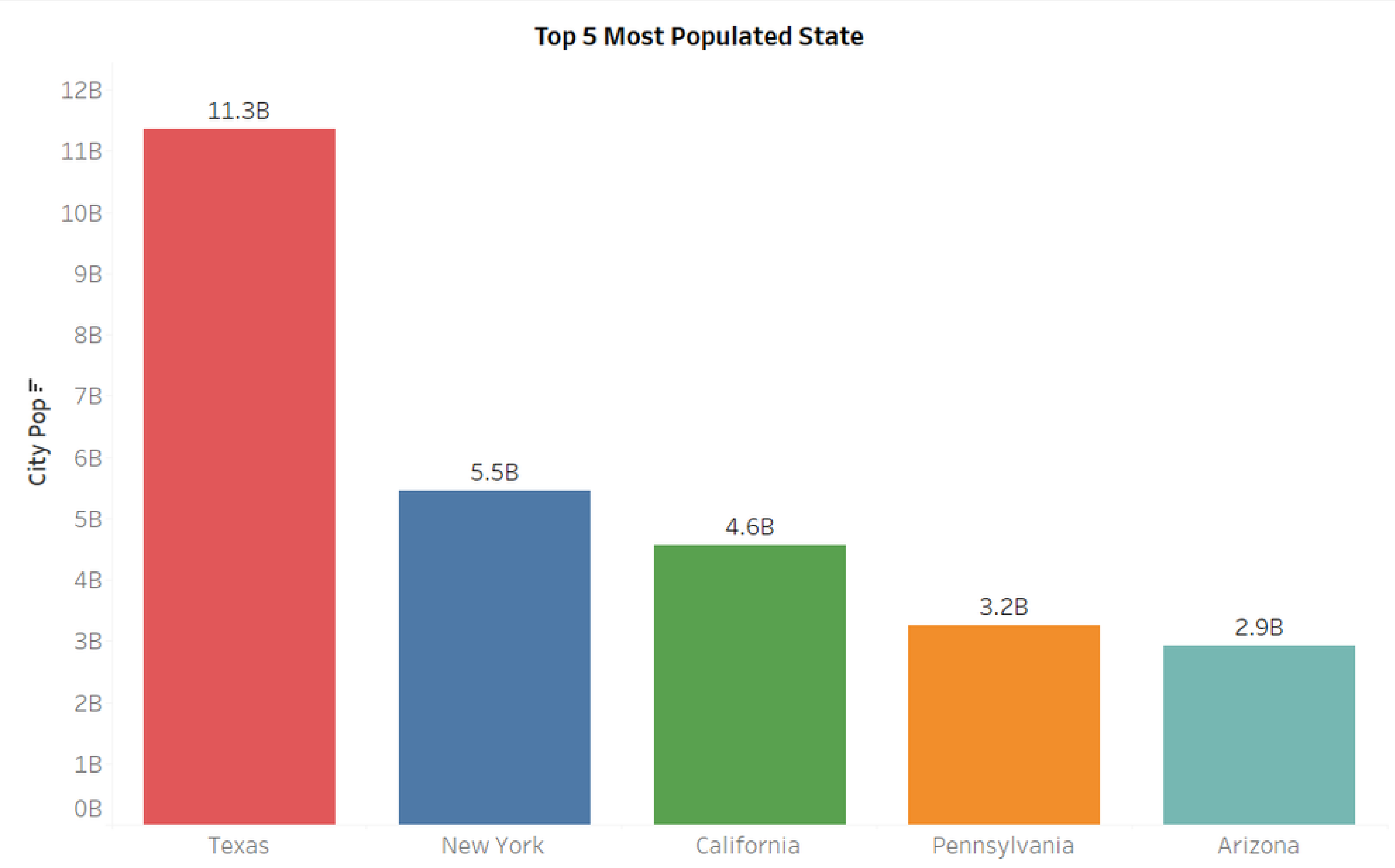

Top 5 Most Populated State

*Fig 1*

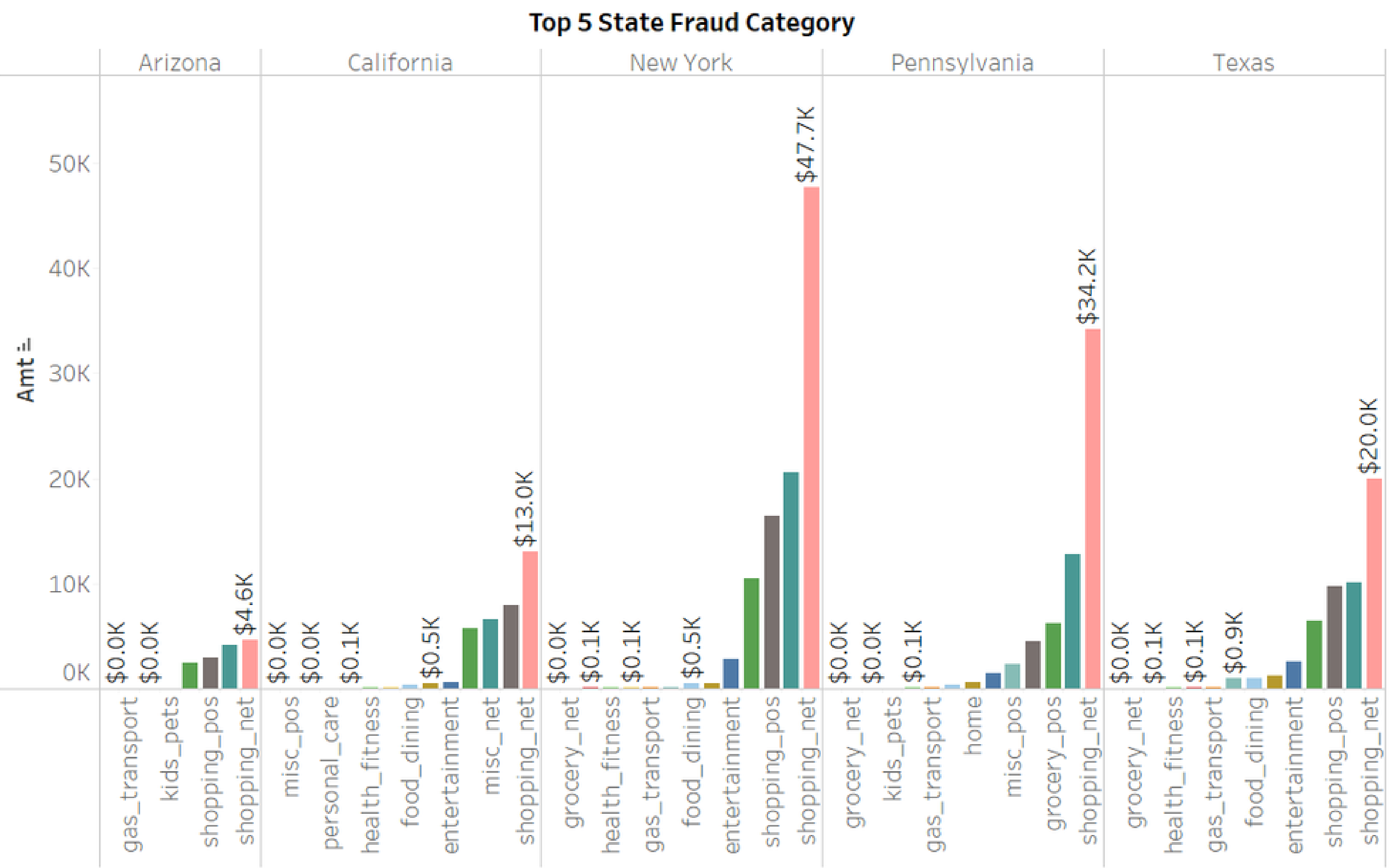
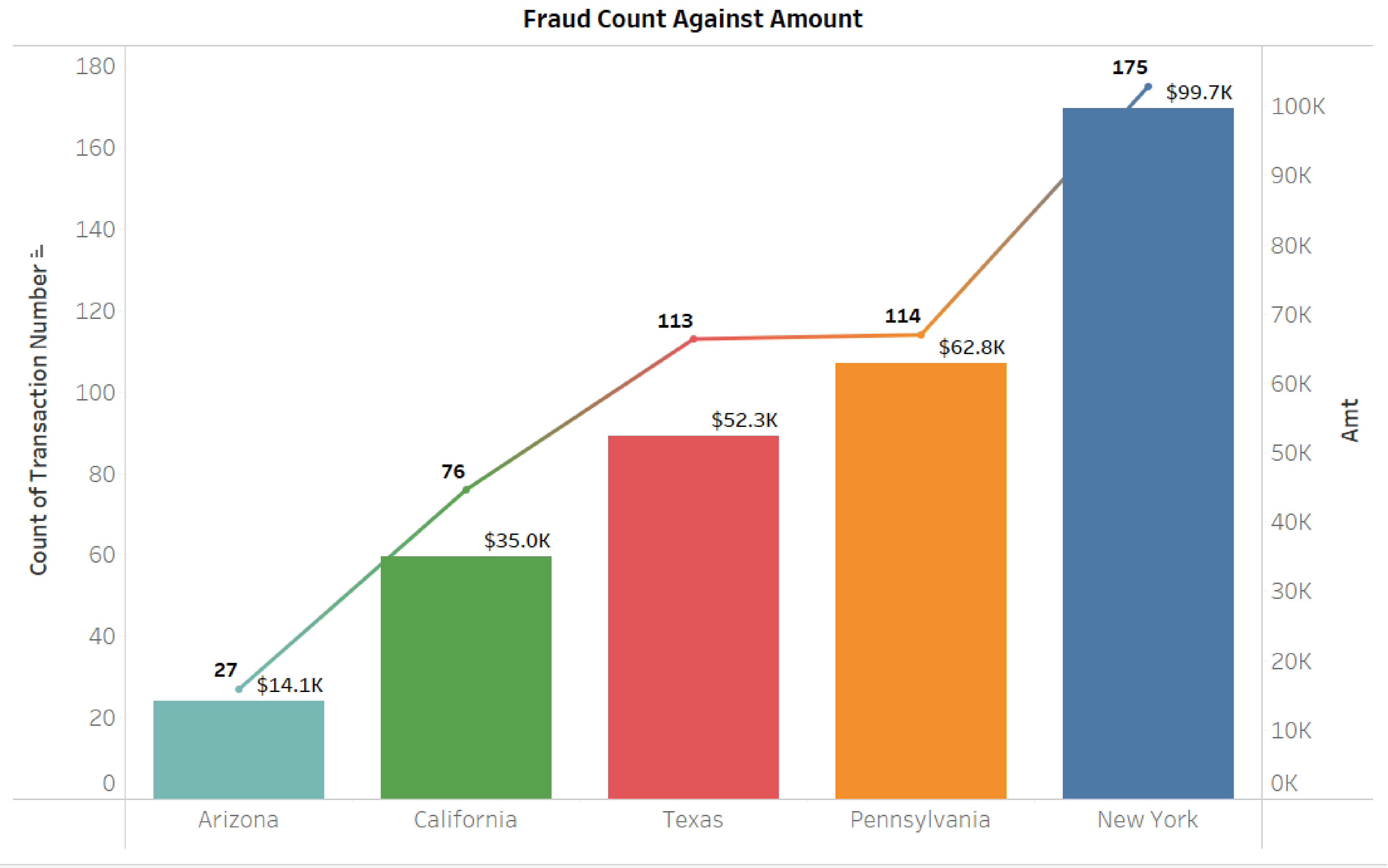

Top 5 State Fraud Category

*Fig 2*

Even so, higher population does not indicate higher fraud cases.

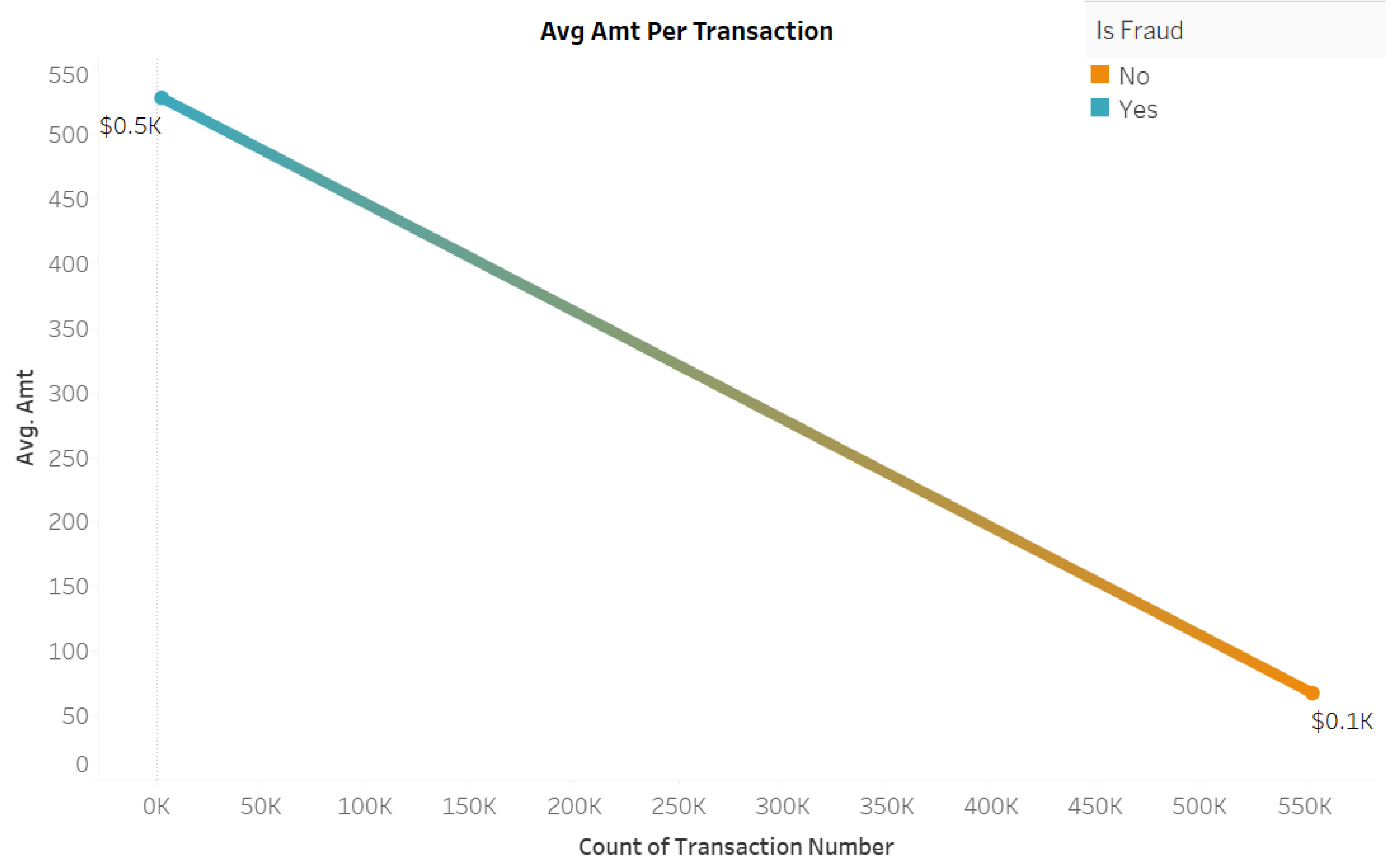
As observed from the graph, the number of cases in each state are below 200 yet the total amount are above $10k and goes up to almost $100k



## Average Amount Per Transaction

Instead, let's have a look at the average amount per transaction.

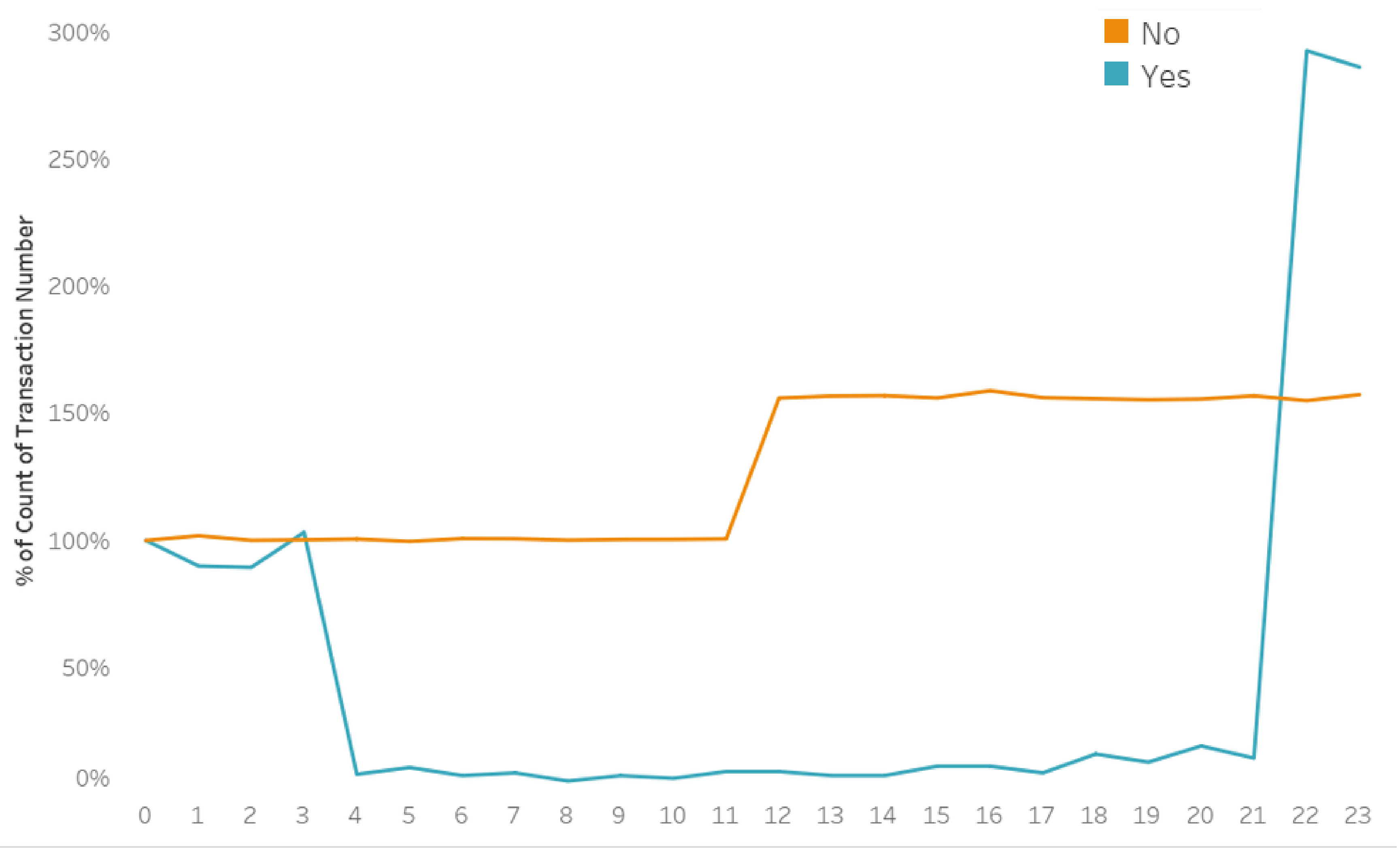
Normal transactions have a higher case count with an average of **$100** per transaction whereas fraudulent transactions tend to have a lower case count with an average amount of **$500** per transaction



## Hourly Fraud Trend

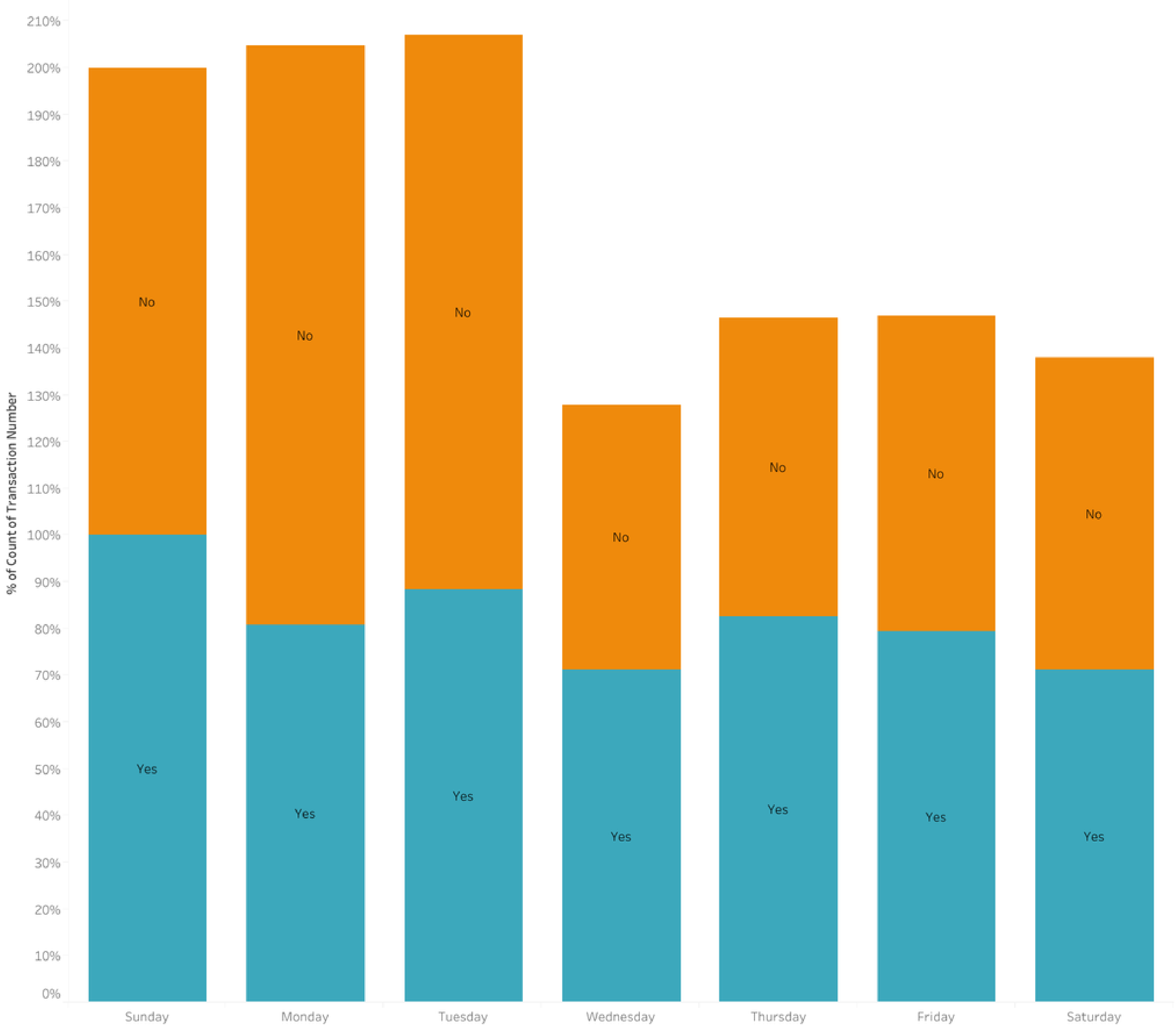
Going into the fraud cycle trend, the line graph is showing a very sharp contrast between normal and fraudulent transaction!

Normal transactions seem to distribute more or less equally throughout the day whereas fraud transactions are disproportionate and are highest between midnight and early morning when people are asleep



## Weekly Fraud Trend

Bar graph shows fraudulent transactions tend to occur at a consistent rate daily throughout the week while normal transactions peak on ***Sunday-Tuesday***



# FINDINGS & RECOMMENDATIONS

## FINDING

As seen from the pie chart previously, the percentage and count of fraud cases between Male and Female are almost evenly distributed at 46% and 54% respectively. What we can gather from the statistics is that both genders are equally susceptible and are not very indicative of a fraudulent transaction.

Inversely, there is a huge disparity between the different age groups

## RECOMMENDATION

The age group with the lowest total fraud amount is from those below 20 years old. This could also be due to the minimum age requirement to own a credit card. On the other hand, the highest total fraud amount came from the age group 51-75 at $451k, indicating that they are indeed more susceptible to fraud. It's recommended to look into what makes older people more susceptible to fraud.

## FINDING

Trend of Fraudulent Activity – Upon closer inspection of the trend charts, we can predict high fraudulent activity occurring in late hours between 2300hrs-0500hrs, spread across evenly throughout the week.

Additionally, normal transactions have an average of $100 per transaction while fraudulent transactions tend to have a higher average amount of $500 per transaction

## RECOMMENDATION

It might be useful to consider including additional forms of security for transactions that exceed a certain amount or occurring during certain periods of the day. Perhaps a review of the daily transaction limit could also aid in curbing fraudulent transactions.

# CONCLUSION

While gender is not indicative of a fraudulent transaction, age, on the other hand, is an important factor when looking into these cases. As such, it is important to look further into what makes users from this particular age group more susceptible: does technology and computer literacy play a part?

Interestingly, 'shopping' seems to be the category that comes up highest in fraud cases among all age groups, regardless young or old. It is also notable that fraudulent transactions have a higher average than normal transactions. It would be worthwhile to study the consumer behaviour and their spending habits to understand better the consumers' motivations.

THANK YOU