Credit Card Fraud Detection

Lighthouse Labs Capstone Project

Yibing Kong (Josie)

Overview:

The primary objective of the project is to gain insights into credit card fraud through exploratory data analysis, which includes creating visualizations and dashboards to observe and identify trends in various aspects. This involves:

1. Analyzing the distribution of transaction amounts between fraudulent and legitimate transactions.
2. Examining transactions based on the time of day.
3. Identifying geographical patterns of transactions.
4. Investigating the correlation between certain variables and the occurrence of fraud.
5. Assessing whether specific merchants are more susceptible to fraud than others.
6. Exploring activities of individuals who are noticeable for fraudulent activities, including analyzing their Date of Birth (DOB), job information, and genders.

Additionally, a secondary objective is to explore and develop a regression model based on the available dataset.

Dataset Info/Exploratory Data Analysis:

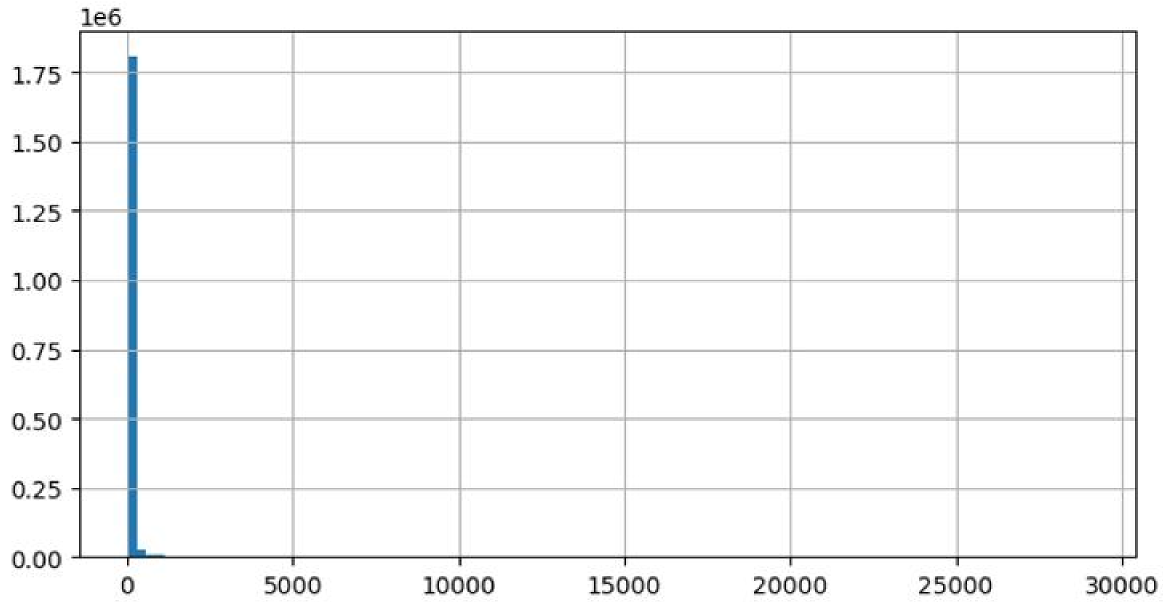
This dataset simulates credit card transactions, encompassing both legitimate and fraudulent activities, spanning from January 1, 2019, to December 31, 2020. It involves transactions conducted by 1000 customers with a diverse pool of 800 merchants.

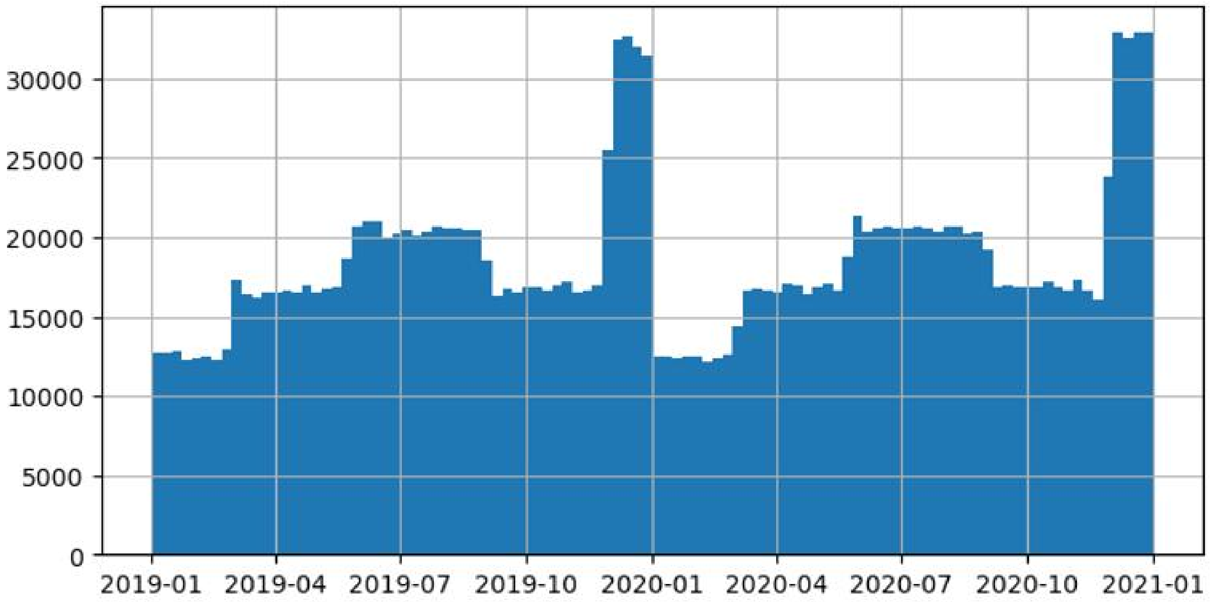
Transactions are classified into 14 different categories, with only 9651 out of the 1.85 million transactions (0.52%) were identified as fraudulent.

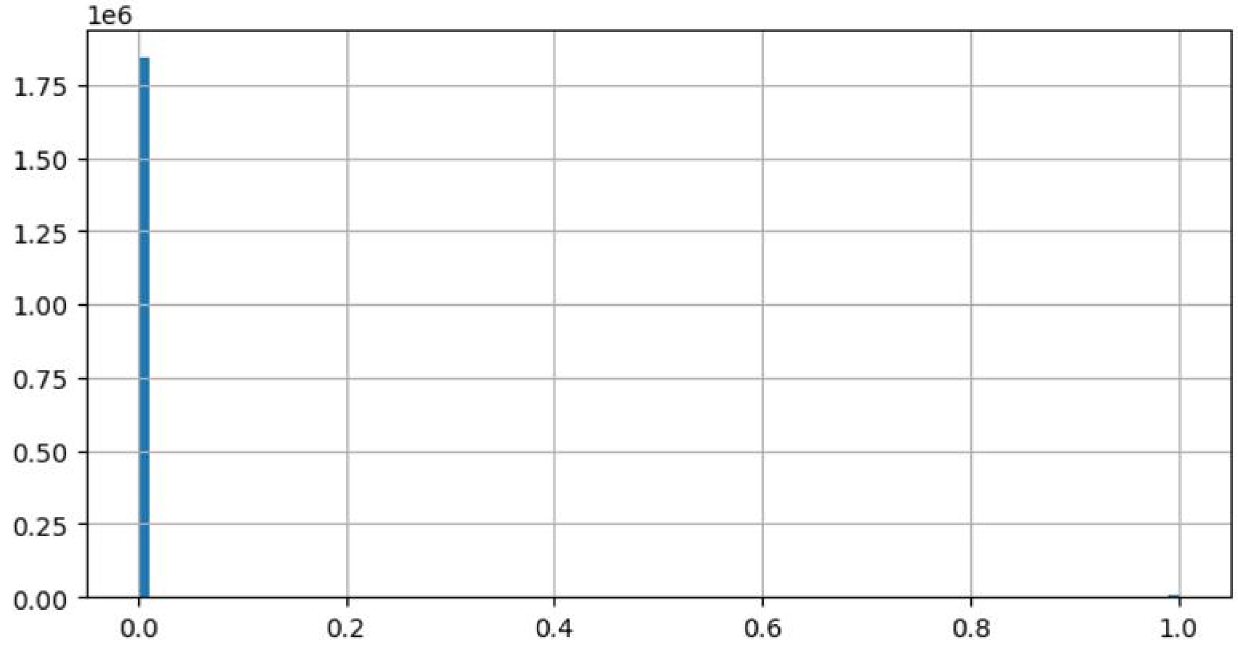
The columns within dataset:

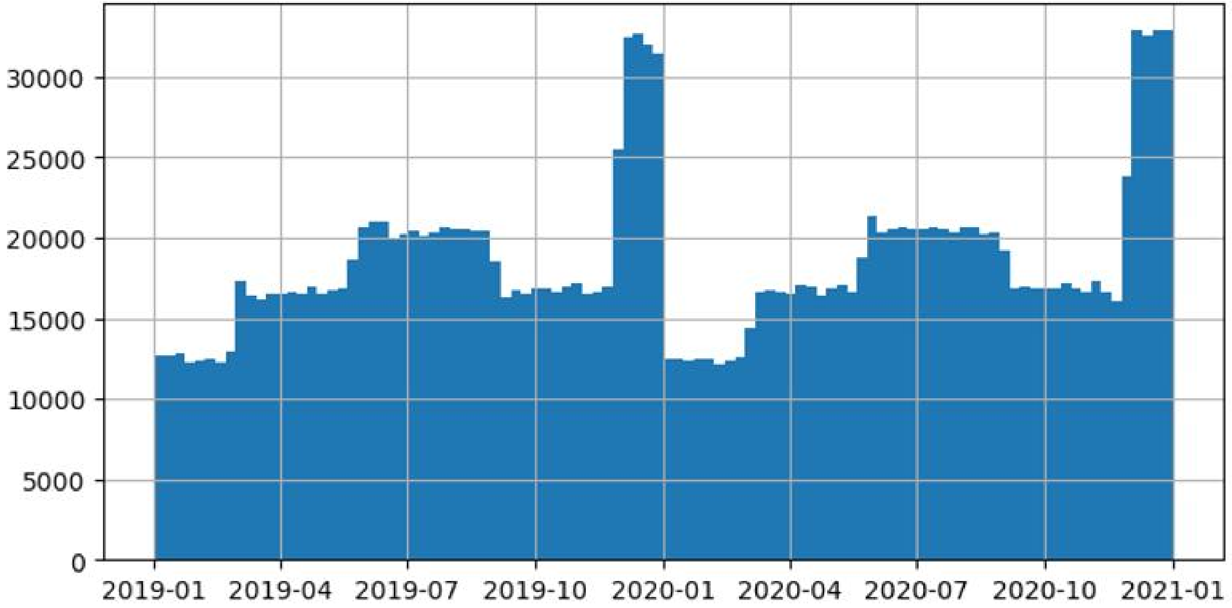
- **Timestamps**: Both trans_date_trans_time and unix_time allowing for time-based pattern analysis.
- **Merchant Details**: Includes the name of the merchant and geolocations (latitude & longitude).
- **Transaction Details**: Indicates the category of the purchased product and the amount of the transactions.
- **Credit Card Holder Information**: Encompasses names, gender, dates of birth, addresses, and credit card numbers.
- **Fraud Indicator (is_fraud)**: 1 for fraudulent transactions, 0 for legitimate transactions.

EDA

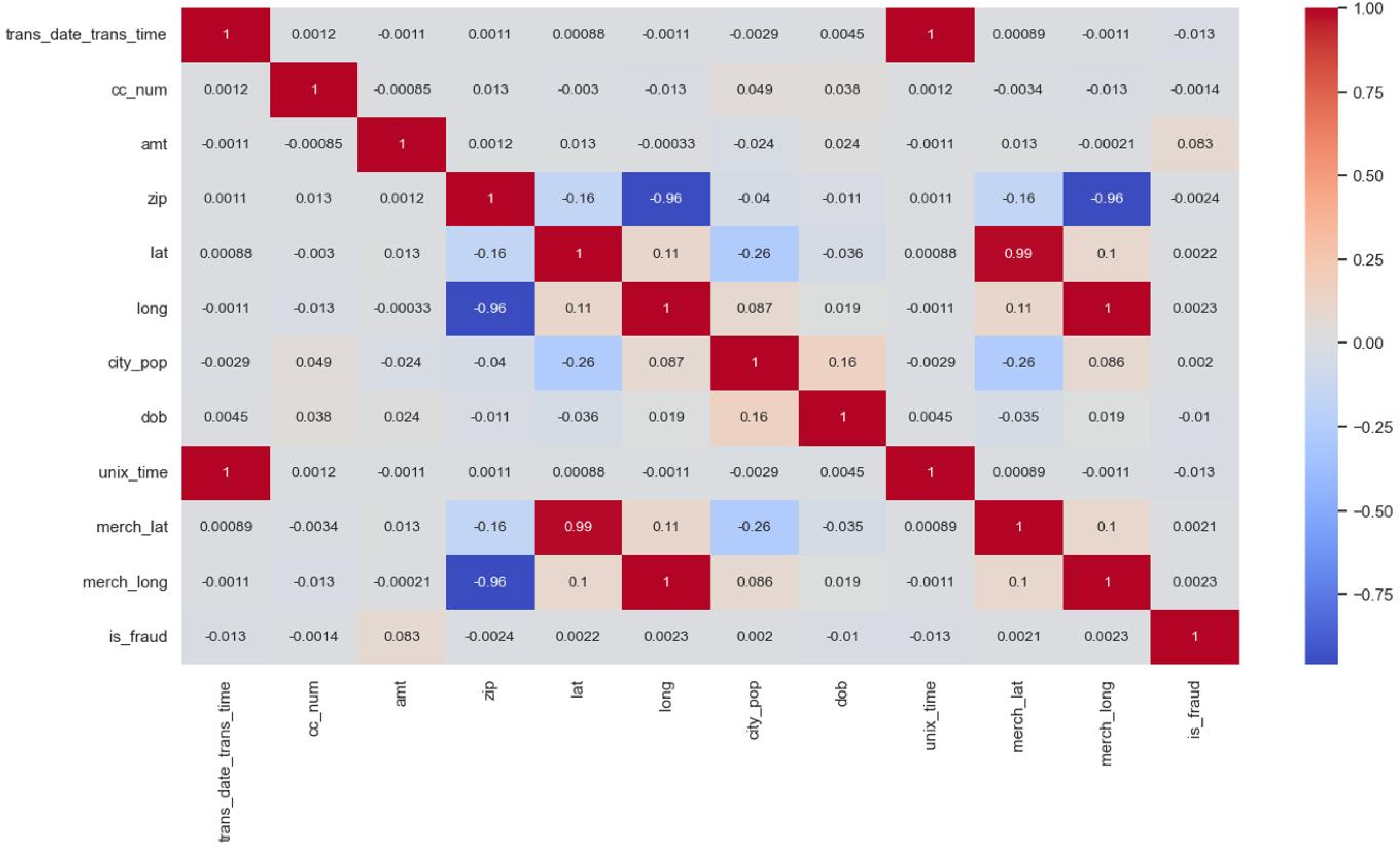

1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
5000
10000
15000
20000
25000
30000


30000
25000
20000
15000
10000
5000
0
2019-01
2019-04
2019-07
2019-10
2020-01
2020-04
2020-07
2020-10
2021-01

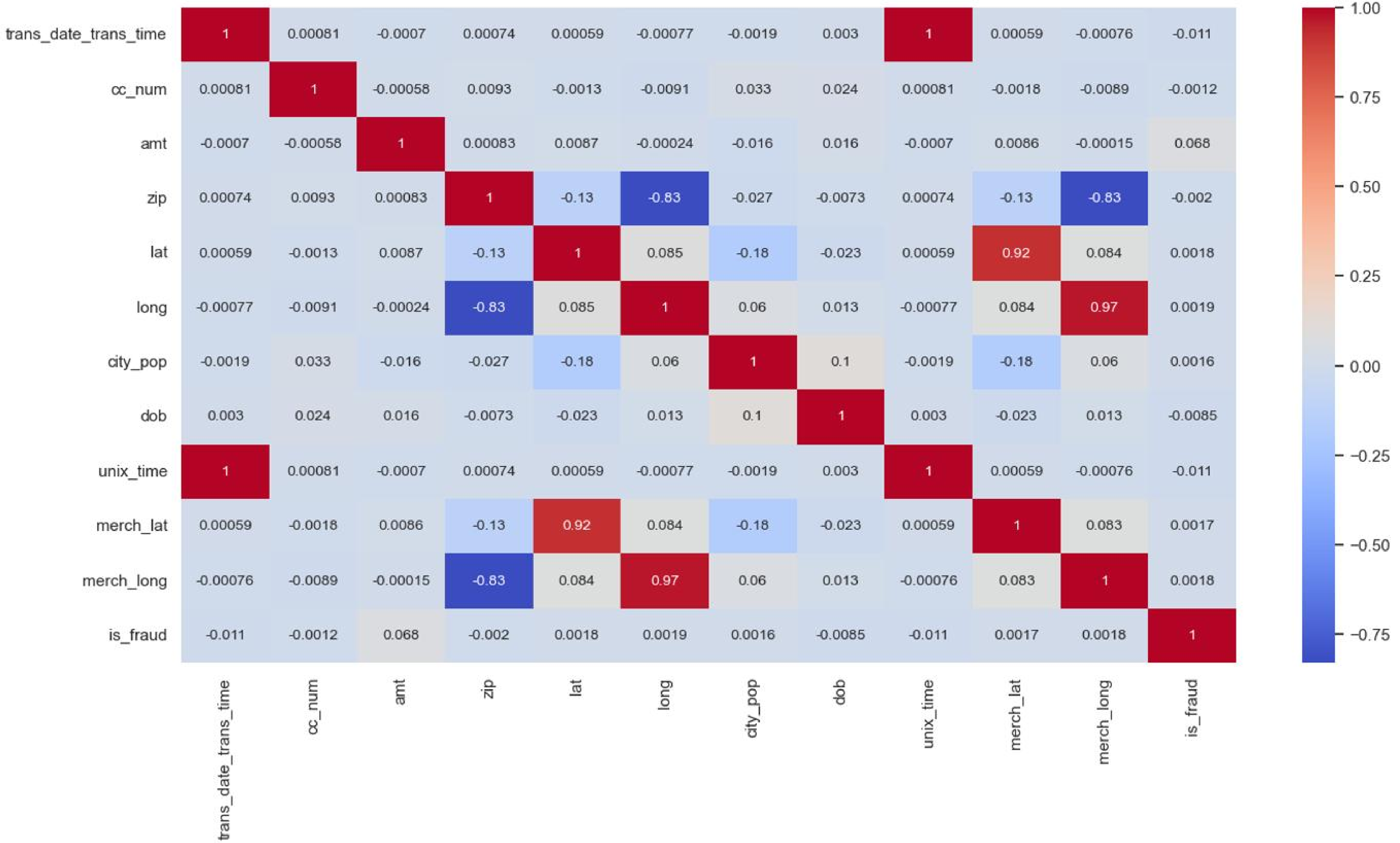

1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
1.0


30000
25000
20000
15000
10000
5000
0
2019-01
2019-04
2019-07
2019-10
2020-01
2020-04
2020-07
2020-10
2021-01

Heatmap - Spearman



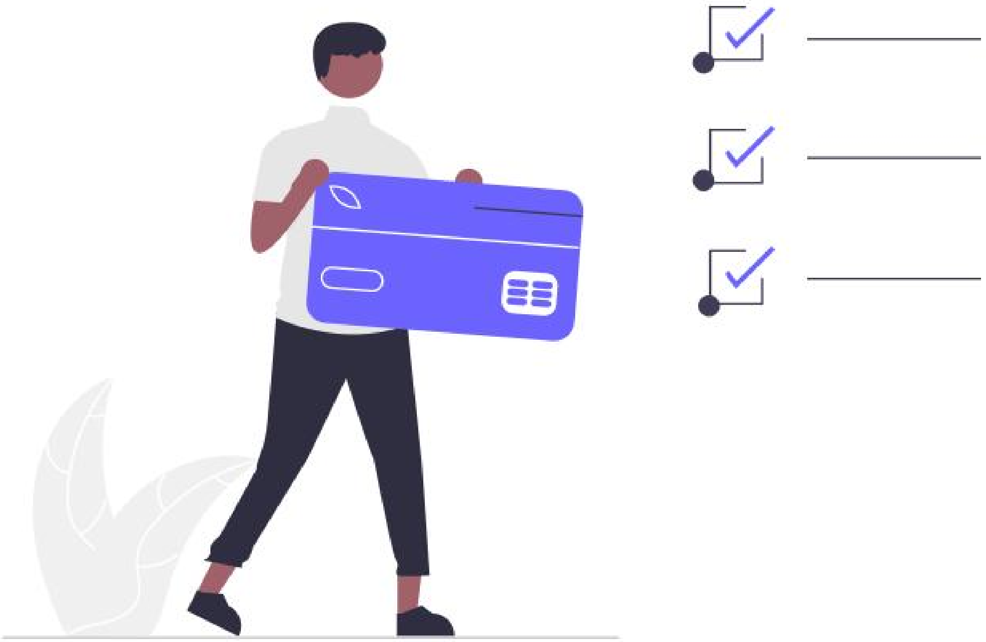
Heatmap - Kendall

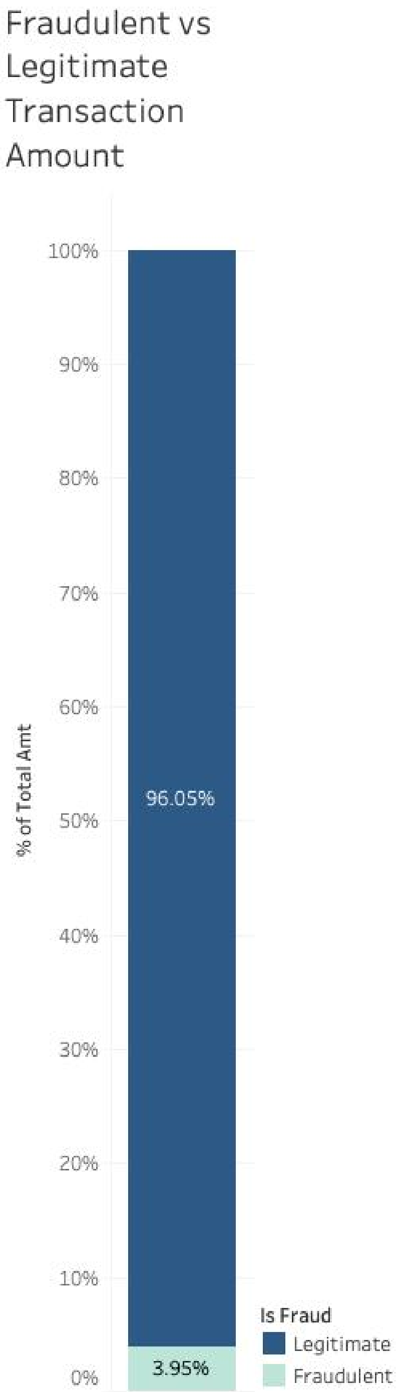


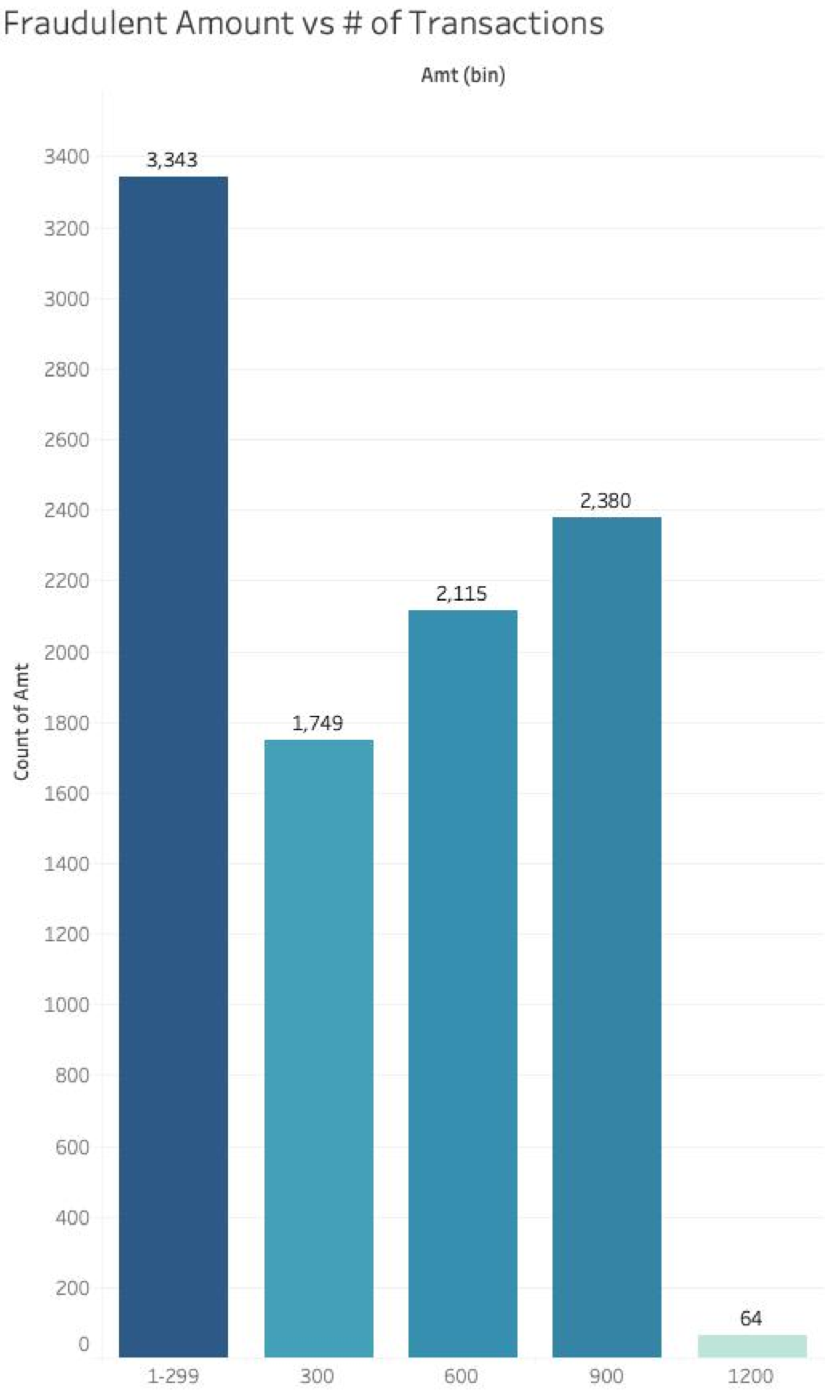
Observation:

Majority of the transactions in the dataset is legitimate.

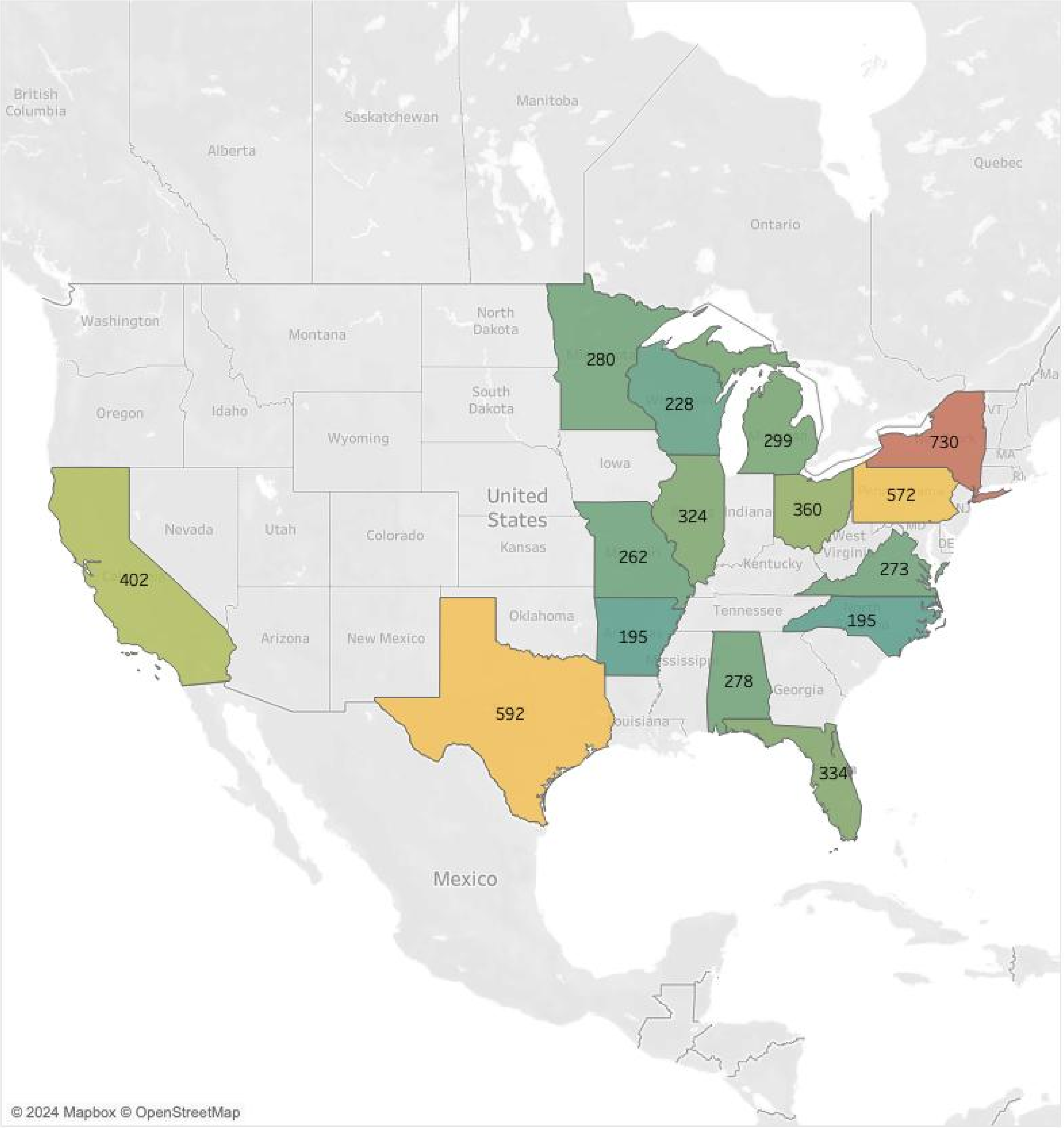
Majority of the fraudulent transactions are below 1200 USD per transaction.



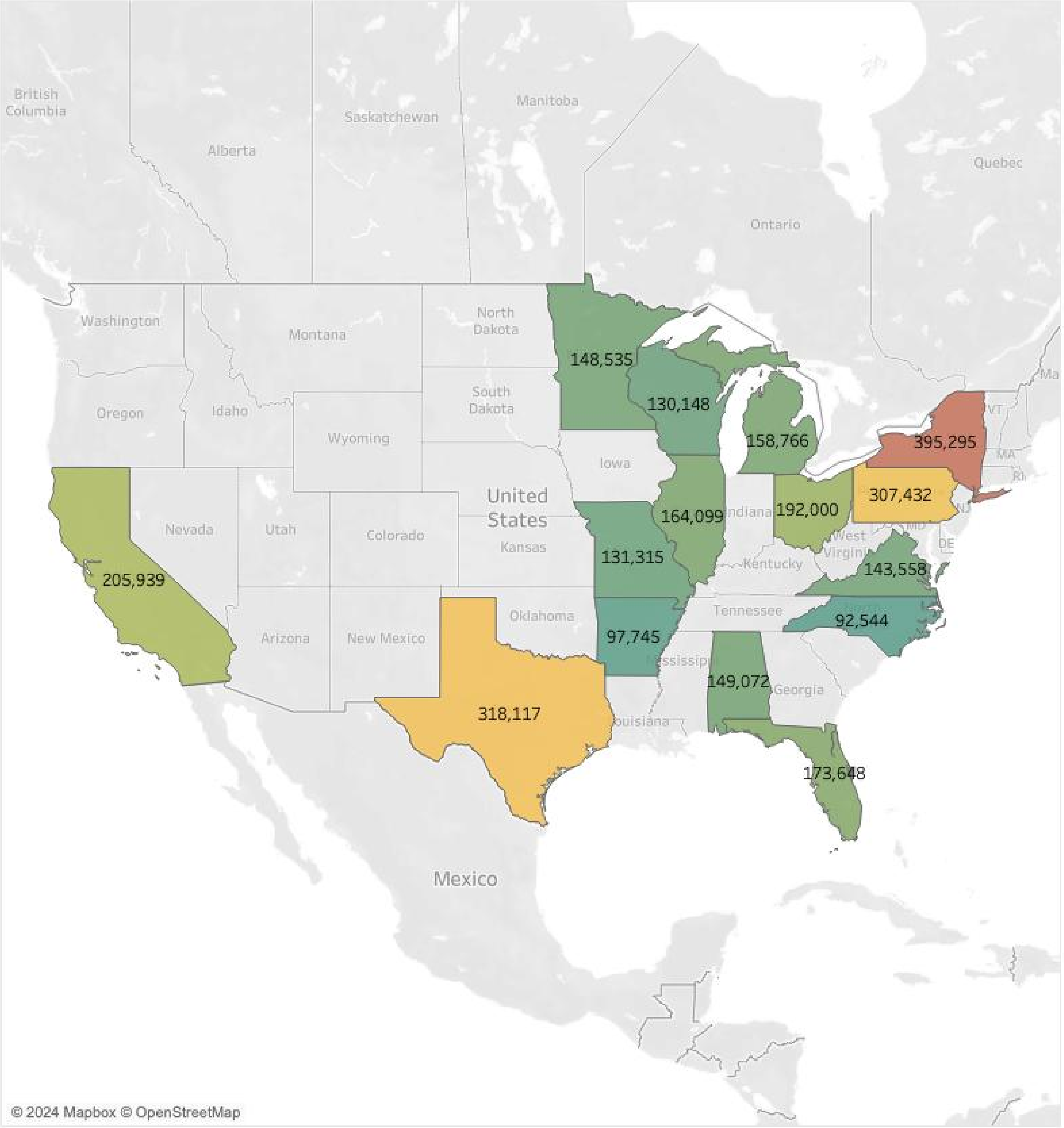




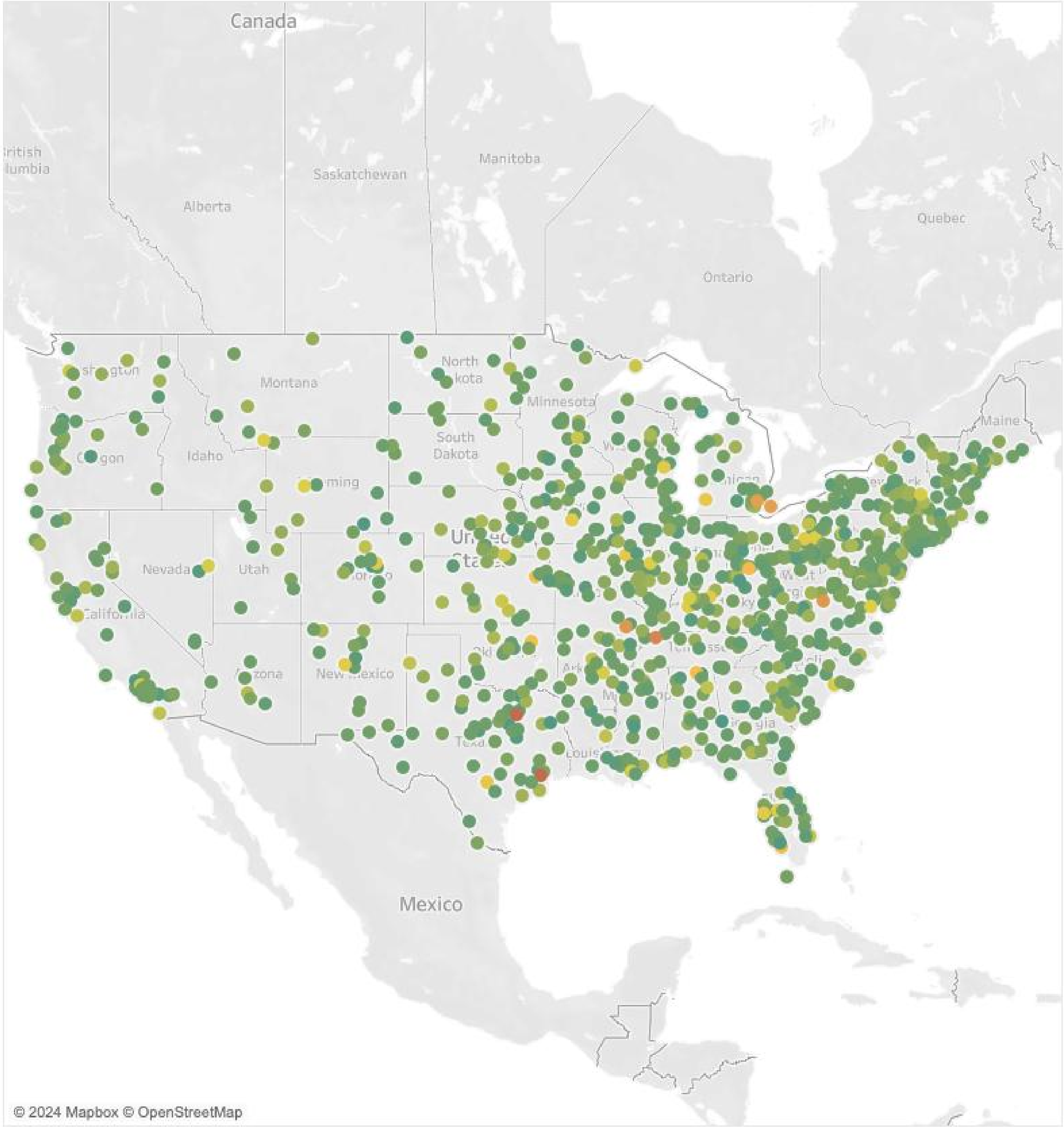
Map - Fraudulent Transaction



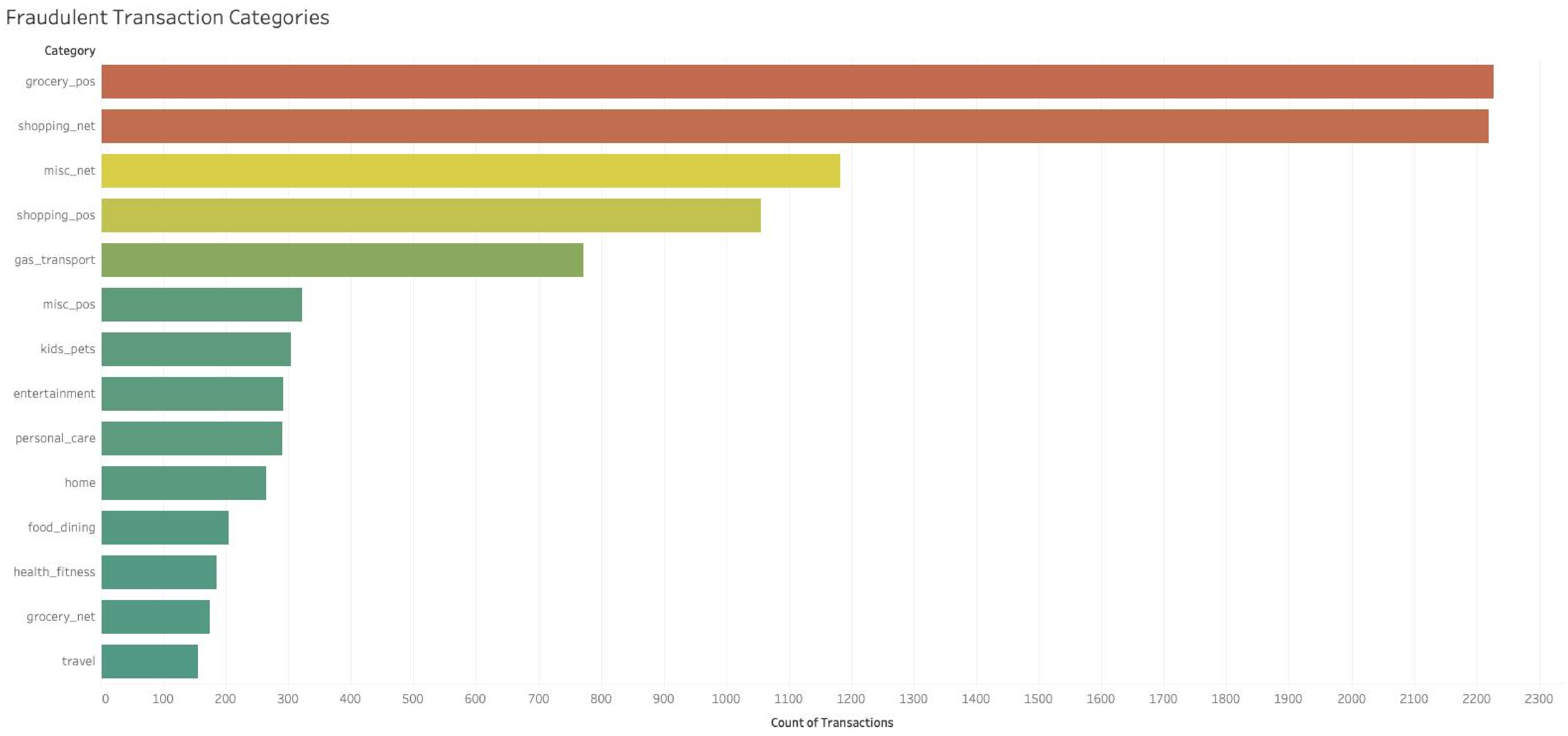
Map - Fraudulent Transaction Amt



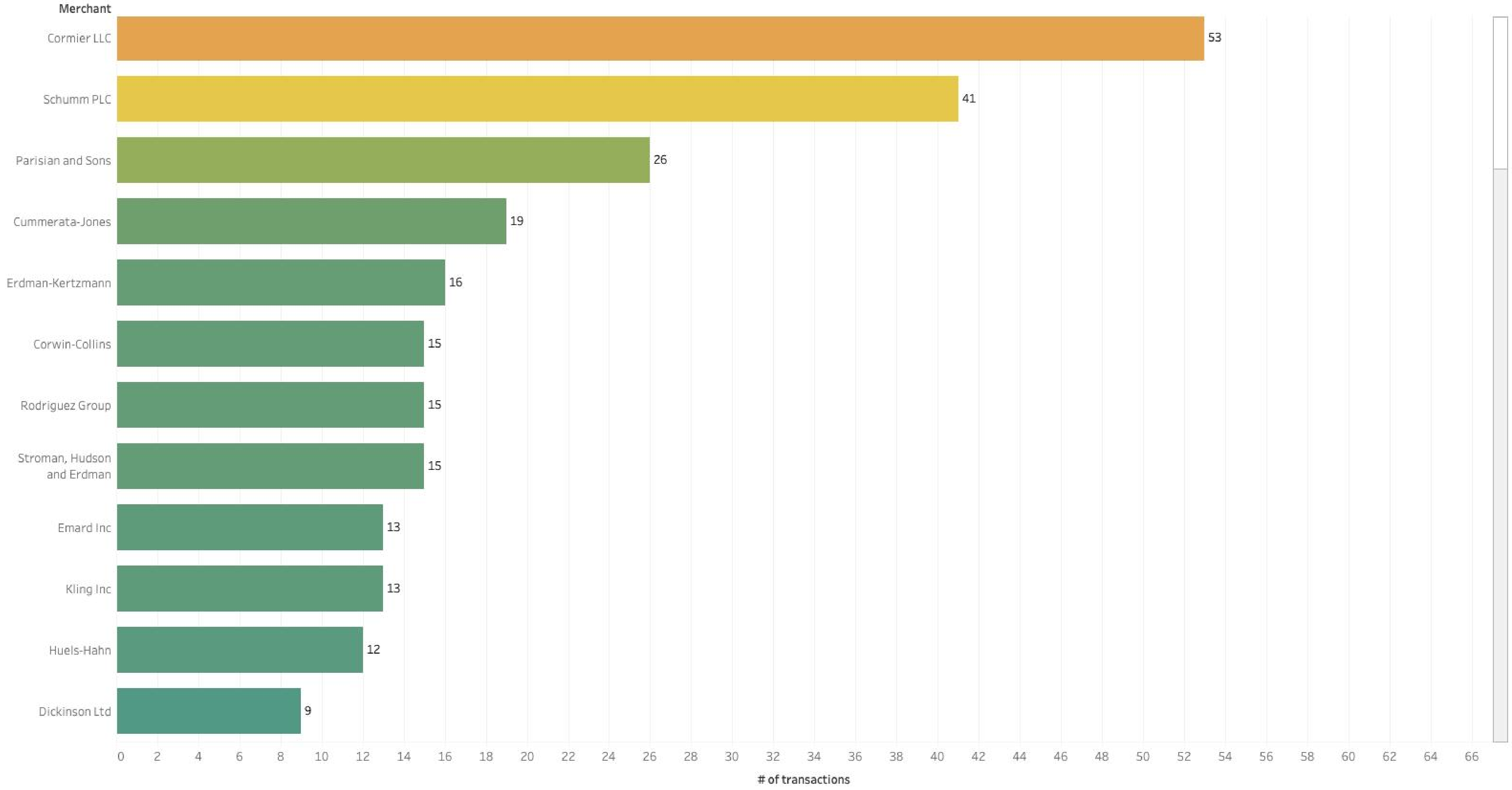
Fraudulent Transactions by Cities Map View



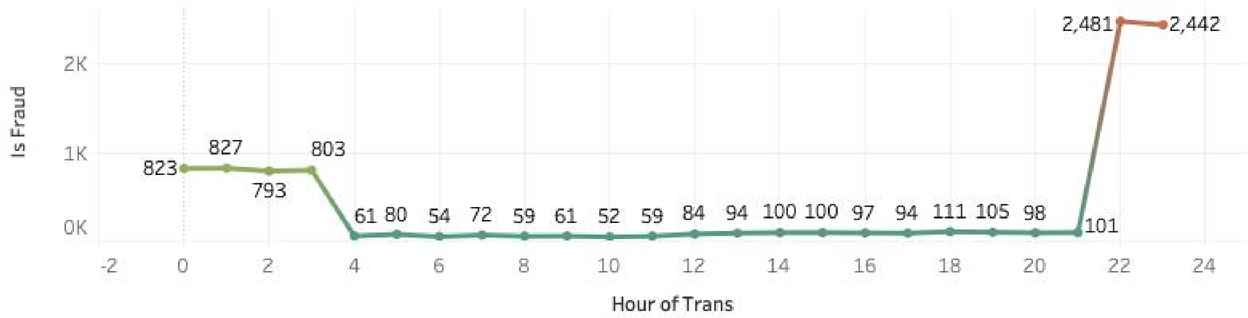
Fraudulent Transactions by State/City


Fraudulent Transaction Categories
Category
grocery_pos
shopping_net
misc_net
shopping_pos
gas_transport
misc_pos
kids_pets
entertainment
personal_care
home
food_dining
health_fitness
grocery_net
travel
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
1400
1500
1600
1700
1800
1900
2000
2100
2200
2300
Count of Transactions

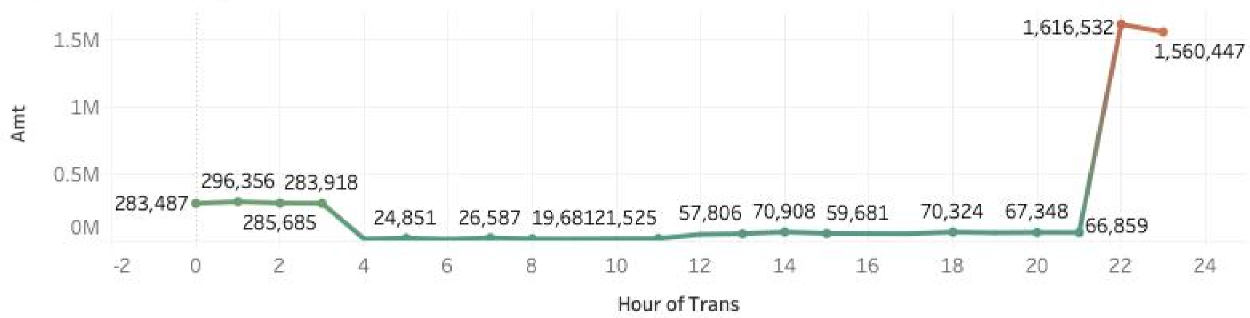
Top 15 Merchants with Highest Total Fraudulent Transactions



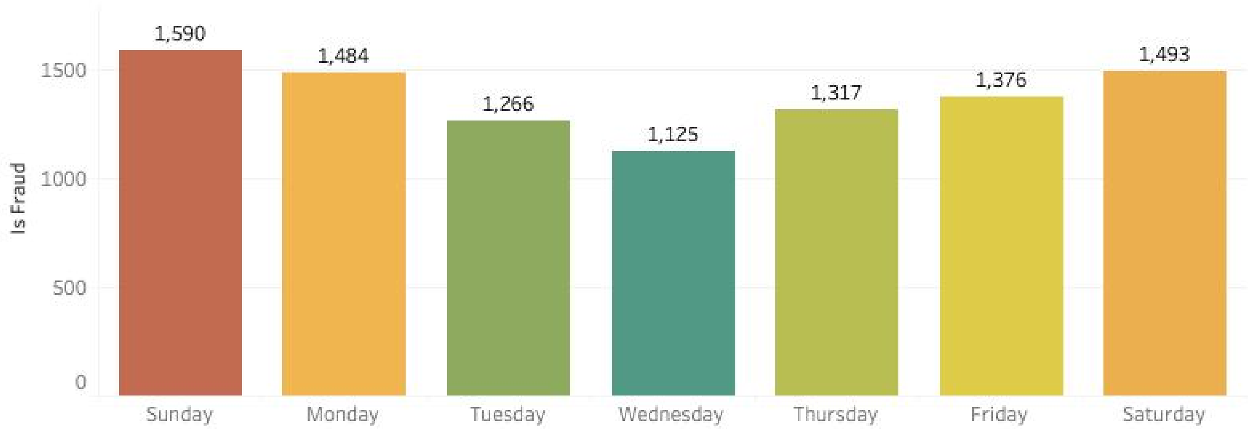
Fraud Transaction # by Hours



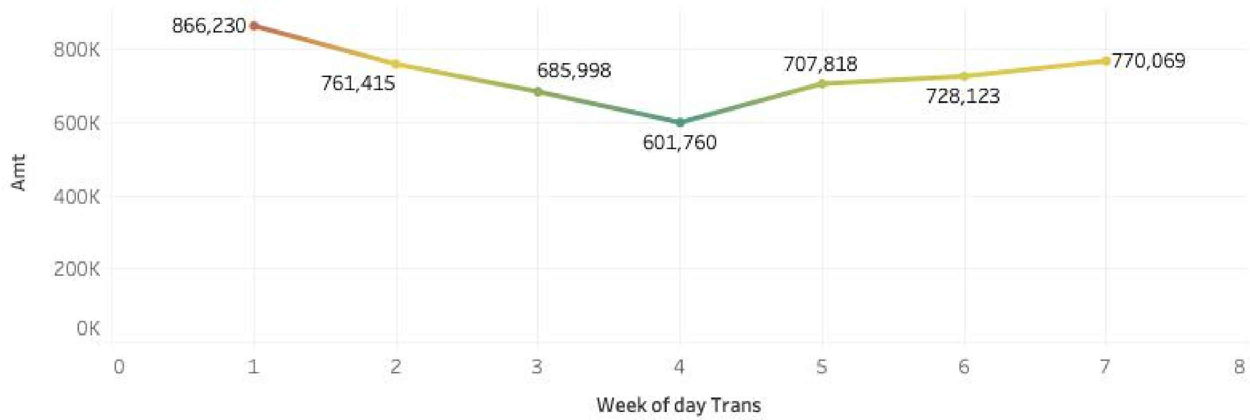
Fraud Transaction Amt by Hours



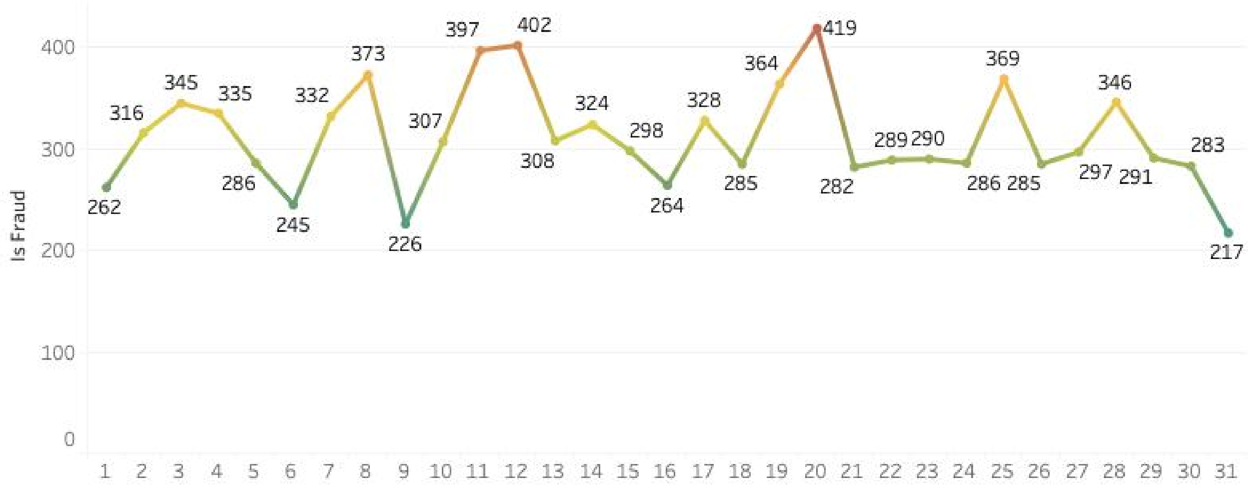
Fraud Transaction # by Day of Week



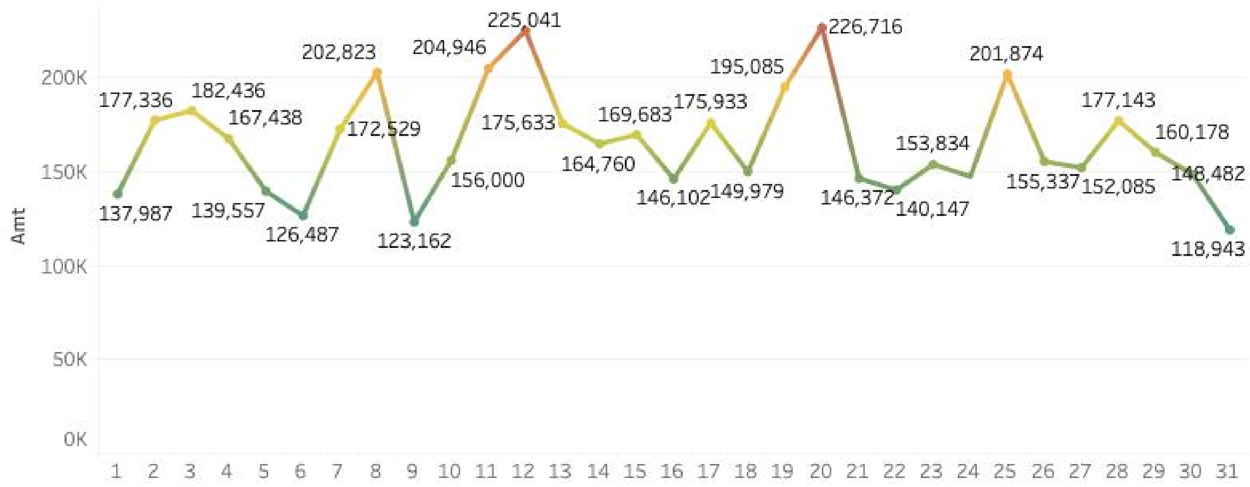
Fraud Transaction Amt by Day of Week



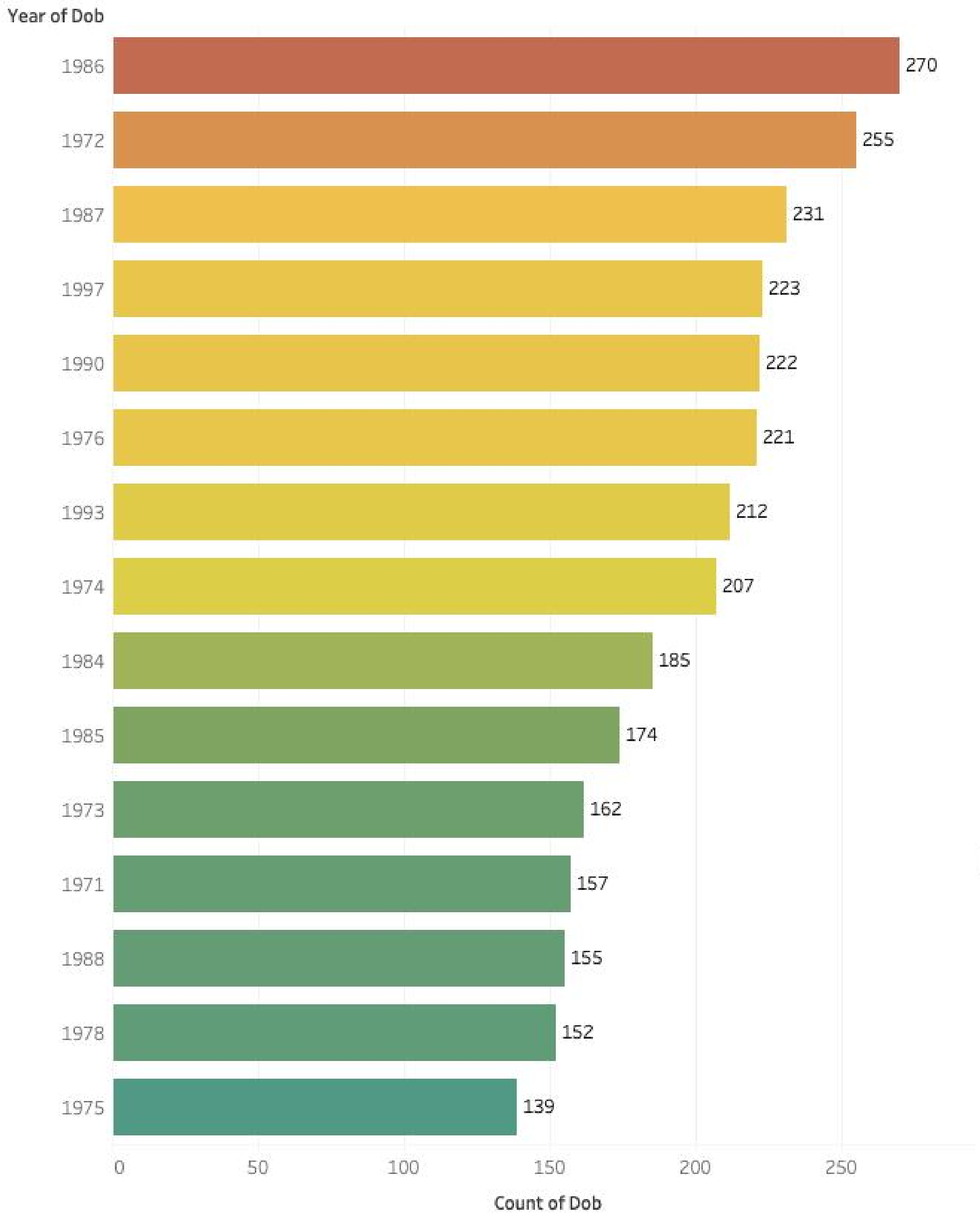
Fraud Transaction # Day of Month



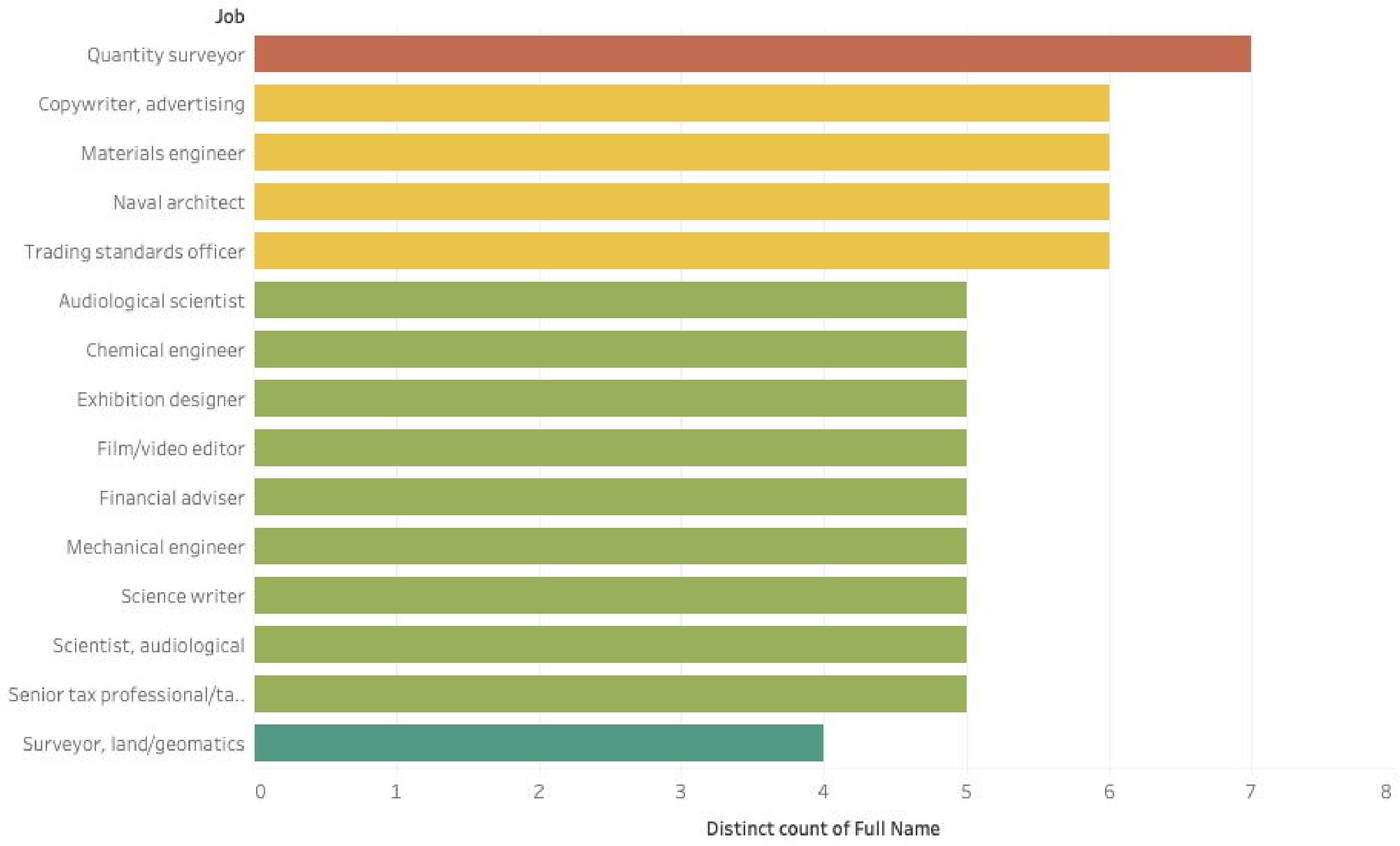
Fraud Transaction Amt Day of Month



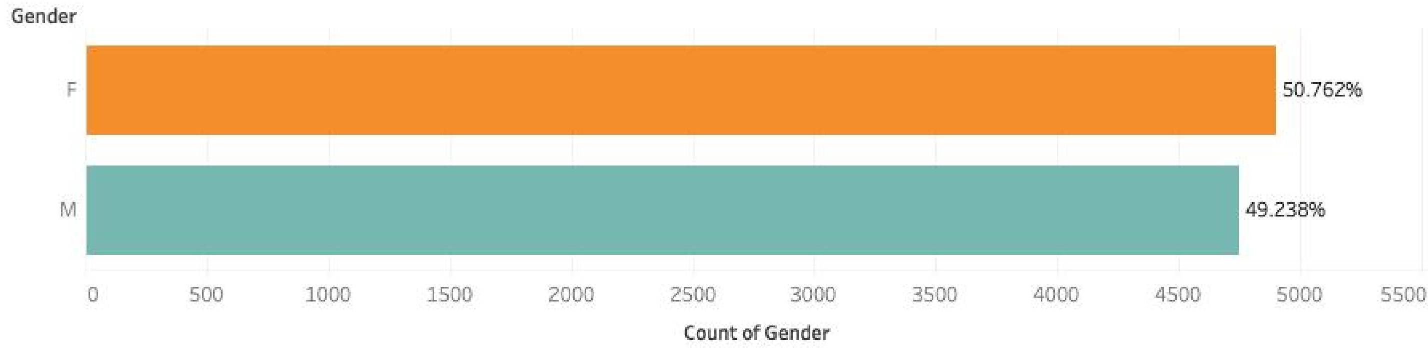
Top 15 Birth Year with Highest Fraudulent Activties



Job Title with Headcount of Individuals Impacted by Fraudulent Activities



Fraudulent Activities - Gender

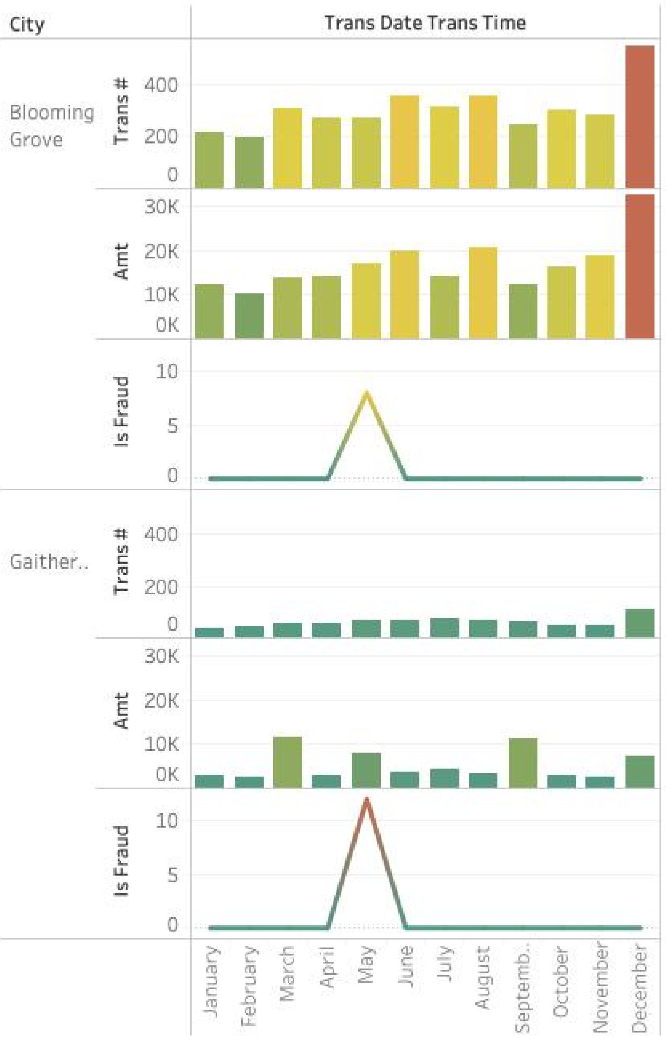


Individual with highest Fraudulent Transactions

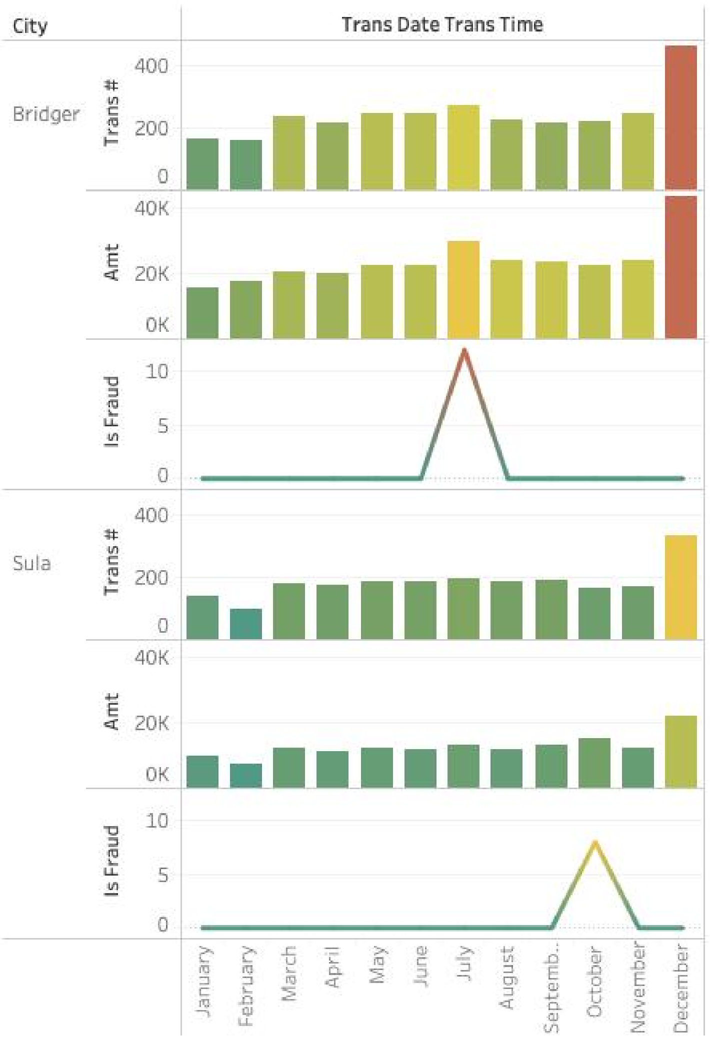
Full Name	Cc Num	Year of Dob	Job	Street	City	Zip	State	# of transacti..	Amt
Jeffrey Smith	3534330126107879	1978	Chartered loss adjuster	713 Scott Pike Apt. 712	Bridger	59014	MT	2,922	286,343
	4292902571056973207	1995	Therapist, horticultural	135 Joseph Mountains	Sula	59871	MT	2,196	152,386
Linda Davis	4433091568498503	1936	Clinical biochemist	493 Todd Views	Gaithersburg	20882	MD	740	62,993
	4452366298769043	1978	Financial adviser	6602 Ortiz Pine Apt. 179	Blooming Grove	76626	TX	3,649	201,944
Scott Martin	3502088871723054	1976	Operations geologist	31472 Cody Place Suite 740	Kensington	20895	MD	3,656	218,185
	4334230547694630	1967	Education officer, museum	7483 Navarro Flats	Freedom	83120	WY	2,927	203,200

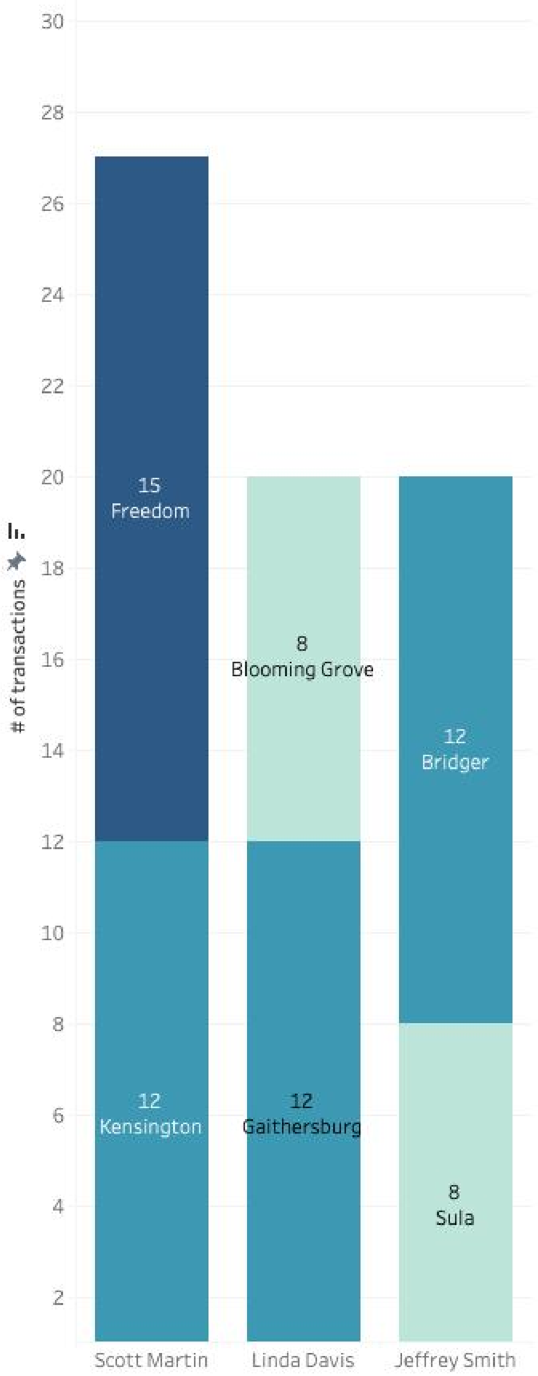
Scott Martin

Linda Davis



Jeffrey Smith





```
                            OLS Regression Results
==============================================================================
Dep. Variable:               is_fraud   R-squared:                      -0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    -13.64
Date:                Wed, 24 Jan 2024   Prob (F-statistic):               1.00
Time:                        20:31:22   Log-Likelihood:             2.2456e+06
No. Observations:             1852394   AIC:                        -4.491e+06
Df Residuals:                 1852390   BIC:                        -4.491e+06
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -5.615e-17   2.25e-17     -2.497      0.013      -1e-16   -1.21e-17
cc_num     -5.404e-23   4.04e-23     -1.336      0.181   -1.33e-22    2.52e-23
amt        -4.999e-14   2.01e-14     -2.486      0.013   -8.94e-14   -1.06e-14
zip        -4.933e-09   1.97e-09     -2.498      0.012    -8.8e-09   -1.06e-09
lat         9.127e-14   3.66e-14      2.492      0.013    1.95e-14    1.63e-13
long        2.302e-12   9.21e-13      2.498      0.012    4.96e-13    4.11e-12
city_pop    1.235e-10   1.76e-10      0.702      0.483   -2.21e-10    4.68e-10
unix_time    4.01e-12   8.15e-14     49.230      0.000    3.85e-12    4.17e-12
merch_lat    9.14e-14   3.67e-14      2.492      0.013    1.95e-14    1.63e-13
merch_long  2.302e-12   9.21e-13      2.498      0.012    4.96e-13    4.11e-12
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               is_fraud   R-squared:                      -0.003
Model:                            OLS   Adj. R-squared:                 -0.003
Method:                 Least Squares   F-statistic:                    -9.144
Date:                Wed, 24 Jan 2024   Prob (F-statistic):               1.00
Time:                        20:37:26   Log-Likelihood:                -7271.7
No. Observations:               10000   AIC:                         1.455e+04
Df Residuals:                    9996   BIC:                         1.458e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.734e-15   1.39e-15     -1.247      0.212   -4.46e-15     9.9e-16
cc_num     -4.195e-21   3.91e-21     -1.073      0.283   -1.19e-20    3.47e-21
amt         7.689e-11   6.18e-11      1.245      0.213   -4.42e-11    1.98e-10
zip        -2.318e-07   1.86e-07     -1.246      0.213   -5.97e-07    1.33e-07
lat         2.644e-12   2.12e-12      1.248      0.212   -1.51e-12     6.8e-12
long        1.098e-10   8.82e-11      1.246      0.213    -6.3e-11    2.83e-10
city_pop   -7.539e-10   1.64e-08     -0.046      0.963   -3.29e-08    3.14e-08
unix_time   3.775e-10   7.67e-12     49.206      0.000    3.62e-10    3.93e-10
merch_lat   2.685e-12   2.15e-12      1.248      0.212   -1.53e-12     6.9e-12
merch_long  1.099e-10   8.82e-11      1.246      0.213    -6.3e-11    2.83e-10
==============================================================================
```

Regression Model 1

Regression Model 2

Initially created regression model based on the available dataset.

The negative R-squared values and unusual F-statistic might indicate issues with the model, we can tell that it might because of imbalance of dataset.

The steps we took next is to resample with sample size 10000 with 50% fraudulent activities and 50% non-fraud activities.

The negative R-squared values, unusual F-statistic, and the fact that many coefficients are very close to zero with high p-values suggest that the model and the individual coefficients are not statistically significant.

Conclusion:

Fraudulent transactions exhibited a discernible pattern, predominantly occurring during specific periods, usually a few hours before and after midnight.

These fraudulent transactions were distinguished by significantly lower spending amounts.

While many other transaction patterns mirrored those of legitimate transactions, the scarcity of fraudulent instances in the dataset (i.e., an imbalanced dataset) prevents us from creating a regression model that could offer valuable insights.