Credit Card Fraud – Data Analysis

Blake Babikian

MA – 346 – Kim

03/08/24

Table of Contents

Introduction:

This report delves into a comprehensive dataset that covers financial transactions over a two-year period, including the transactional data, personal information, merchant data, and geographical data. The primary research question aims to understand the underlying patterns within these transactions, focusing on aspects such as spending behaviors across different categories, the relationship between demographic factors and fraud incidence, and the geographical distribution of transactions.

The project will explore data in various ways: statistical summaries to describe general trends, visualizations to detect patterns and outliers, and machine learning models to predict fraudulent transactions, by looking at transaction attributes that may lead to a higher likelihood of a transaction to be fraudulent.

A pronounced disparity between the average amounts of fraudulent and non-fraudulent transactions, suggests that attackers are drawn to transactions that offer higher financial yields.

Despite initial assumptions, merchant category—not specific merchants—emerged as a more significant factor in fraud prevalence, particularly in Grocery POS and online shopping.

The demographic analysis refutes the stereotype that elderly customers are disproportionately targeted by fraud, showing only a slight age difference between fraud victims and other customers.

Contrary to common beliefs, the distance between customers and merchants does not significantly differ between fraudulent and non-fraudulent transactions.

Using machine learning models, uncovered the complexities of fraud detection. The challenge of balancing the dataset revealed that an increase in the identification of fraudulent transactions could lead to a higher rate of false positives, illustrating the delicate equilibrium required for effective fraud detection systems.

Throughout the statistical analyses, visual aids, and predictive modeling are used to offer a granular view of transactional integrity, laying the groundwork for more informed fraud prevention strategies and financial practices.

Methods/Data:

I. Data Summary

i. Date and Time is used to analyze transaction trends over time, detect seasonal patterns, and identify unusual activity outside of normal hours which could indicate fraud.

ii. The data set includes a column for credit card number and will only be used in modeling part of this analysis.
iii. The merchant's name will be used to identify any suspicious merchants.
iv. The merchant's category will be used to identify any suspicious connections to.
v. Transaction amount will be used to identify any trends in fraudulent transactions that differ from non-fraudulent transactions.
vi. The data set includes a column for the customer's name and will only be used in modeling part of this analysis.
vii. Gender will be used to see if there is an imbalance of fraud cases based on gender.
viii. The data set includes a column for customer address info and will only be used in modeling part of this analysis.
ix. Customer latitude and longitude will be used to calculate distance from merchant.
x. The data set includes a column for city population and will only be used in modeling part of this analysis.
xi. The customers job will be used to analyze any connection to
xii. Customer date of birth will be used to calculate age, at the time of transaction.
xiii. The data set includes a column for transaction numbers and will only be used in modeling part of this analysis.
xiv. The data set includes a column for universal time and will only be used in modeling part of this analysis.
xv. Merchant latitude and longitude will be used to calculate distance from customer.
xvi. The data set includes a column to identify it is a fraudulent case or not.

II. Column Manipulation

i. **Age:** To make the data easy to analyze the demographics in the data, a new column was created by taking the difference in the transaction time and the customers date of birth, this was then used to create a new column, for age at the time of transaction.
ii. **Distance:** The dataset had two sets of coordinates, so it was natural to find the distance between the two, the first method used to find this distance was taking the Euclidean distance, this returned a rather similar number between the fraud and not fraud cases. To ensure this accuracy, a Python library was implemented to calculate the kilometer distance between these two points, making it into a new column.

III. Data Imbalance

Overall, the data set was comprised of almost one point three million rows of data, each row representing a unique transaction. Breaking the data into whether it was fraudulent or not, it contains 1,289,169 non fraudulent transaction compared to 7,506 fraudulent

transactions. With such an imbalance in data, it can be hard to see trends in fraudulent transactions when skewed by the imbalance. Here are some tactics used to combat this:

i. **Ratios:** Most of the charts comparing fraudulent to non-fraudulent cases, a ratio system was implemented. Where the number of fraudulent transactions was divided by the number of non-fraudulent transactions to get a percentage ratio.
ii. **Sampling Data Model:** For the model, several iterations were implemented where the model was running unbalances, slightly balance, and completely balanced to compare the performance based on the manipulation.

Results

i. Transaction Amount (Appendix a)

i. On average, fraud transactions are significantly higher in amount ($531.32) compared to non-fraudulent transactions ($67.67). This significant difference suggests that fraudsters tend to target higher-value transactions, possibly due to the higher potential payout.
ii. Fraud transactions also show a higher standard deviation ($390.56) compared to non-fraudulent transactions ($154.01), indicating a wider range of transaction amounts in fraud cases.
iii. The maximum fraud transactions are just 5% of the largest non-fraud transactions. On the flip side, the top non-fraud transactions only reach 9% of what fraud transactions hit at their 75th percentile. The big transactions in the fraud category are much smaller compared to the big ones in the non-fraud pool. Meanwhile, most of the above average non-fraud transactions are just a small slice of the average fraud ones, showing there's a pretty clear difference in how much money is moving in fraud versus regular transactions.

ii. Merchant Information (Appendix d, g)

i. **Merchant Name:** Certain merchants like Kilback LLC and Schumm PLC have a higher representation in fraud compared to their presence in non-fraud transactions (0.65% and 0.64%). This might indicate specific vulnerabilities or potential targeting in these merchant establishments.
ii. **Category:** The categories with the highest percentage of representation in fraud transactions are Grocery POS and Shopping Online (both 23%), which are significantly higher than their representation in non-fraud transactions (9% and 7%, respectively). This suggests that these categories are particularly vulnerable to fraud.

iii. **Category x Transaction Amount (Appendix j):** When breaking up Transaction amount by category, there are a few general trends, one of those being that, for most categories, the average amount is higher in the fraud cases than the non-fraudulent. Additionally, there tends to be a trend where there far more extremely high outliers in the non-fraudulent boxplots than the fraudulent. Enforcing the findings from the section above

iii. Customer – Merchant Distance (Appendix l)

i. The mean distance for fraudulent and non-fraudulent transactions is remarkably similar, at 76.27 km and 76.11 km, respectively. There is a similar trend with all other statistics. This indicates that the customer-merchant distance does not significantly differ between fraudulent and non-fraudulent transactions.

iv. Date & Time (Appendix b, c, n)

i. **Time:** Fraudulent transactions have notable peaks during late night and early morning hours, particularly at 11:00 pm, 12:00 am, and 10:00 pm and other early morning hours, suggesting fraudsters prefer times with potentially lower transaction monitoring or when customers are less likely to notice unauthorized activities.

ii. **Time vs. Amount:** When subsetting rows where transaction time is from 11:00pm until 4:00am, and then take the average amount, the fraudulent average is $529, compared to the non-fraudulent average of $69.

iii. **Date:** The analysis shows a consistent pattern of fraudulent transactions across the timeline, with a noticeable orange layer visible at the bottom, indicating a persistent occurrence and dollar amount range of fraud throughout the period. Non-fraudulent transactions, represented by the blue, are much more variable in amount, with several peaks that suggest higher transaction volumes or amounts on specific dates. Please note that the orange line is on top of the blue, otherwise the sheer number of non-fraudulent transactions would completely cover the fraudulent. Notably, there are a few exceptionally high spikes in the non-fraudulent transactions, which could represent seasonal shopping trends, or other transaction volumes events.

v. Customer Demographics (Appendix f, h, i)

i. **Age:** The mean age for fraud victims is slightly higher at 48.87 compared to 46.01 for non-fraud cases. This may imply a targeted preference or vulnerability among slightly older demographics. Though this trend is not absolute, as the graph shows only a slight upward trend and the pointbiserialr correlation between the two is 0.012, with a very low p-value, showing significant weak connection between the two.

ii. **Gender:** Both genders are almost equally targeted in fraud cases, with a negligible difference in the number of fraud cases between males and females.
iii. **Job:** Specific jobs are more associated with fraud, but vague job titles or industries may provide a more generalized picture of a connection to occupation and likelihood of being a victim of fraudulent activity.

vi. Modeling (Appendix m)

i. **Unbalanced:** Had a low precision of 0.07, and a recall of 0.51, indicating that while the model could identify over half of the fraudulent transactions, it did so with a high rate of false positives. The overall accuracy of this model stood impressively at 97%, demonstrating its effectiveness in identifying non-fraudulent transactions but struggling significantly with accurately predicting fraud.
ii. **Slightly Balanced:** There was a slight improvement in fraud detection precision to 0.08, while maintaining the recall rate. This marginal increase in precision underscores the delicate balance when modeling a complex system such as this.
iii. **Completely Balanced**: The precision for fraud detection dipped further to 0.03, but the recall surged to 0.88. This significant increase in recall suggests the model became much better at identifying fraudulent transactions, capturing a vast majority of them. However, the low precision indicates a high number of legitimate transactions were incorrectly flagged as fraud, highlighting a trade-off between detecting as many fraud cases as possible and minimizing incorrect fraud classifications.

Discussion

This season's personal encounters with fraud underscore the practical implications of the findings from the data analysis. One instance involved an online shopping scam: a Facebook ad appeared to offer a perfect gift, which it was purchased, only to have the product never arrived. Upon further research, there were many similar stories tied to this company. Similarly, an experience at a local grocery store, where a fraudulent card scanner intercepted payment information. Conveniently, both online shopping and grocery store POS were shown in the analysis as the most common fraud merchant categories.

Unfortunately encounters with fraud is common, going into this analysis some may have some preconceived notions. Initially thinking that fraud primarily stems from a few dubious merchants, but the data showed a broader distribution, with certain merchant categories being more prone to fraudulent activity rather than specific merchants. While older individuals are often stereotypically viewed as prime targets for fraud, the data revealed only a slight age difference between fraud and non-fraud cases, suggesting that susceptibility to fraud is more

subtle across age demographics. As for the role of distance, it's a common belief that transactions occurring far from a customer's usual location, such as during vacations, are more likely to be flagged as fraudulent. However, the data contradicted this assumption, indicating that the mean distance for fraudulent and non-fraudulent transactions was remarkably similar, which suggests that distance alone isn't a reliable predictor of fraud. These insights from real-world experiences and data analysis can help to reshape understanding and expectations about the nature of fraud, pointing towards more complex patterns and indicators.

The data summary illustrates the value of various variables in detecting fraud, such as time, merchant categories, and amounts for spotting anomalous activities and identifying potential risk points. It's telling that while the mean distance between customers and merchants does not differ markedly between fraudulent and non-fraudulent transactions, combinations of factors like the time of the transaction and the transaction amount are significant indicators of fraud.

In the dataset, the high average transaction amount in fraud cases, and its greater variability, suggest that attackers aim for more lucrative payouts, however, there seems to be a cap to how much they are willing to steal. This may have to do with legislation, as in many states, theft doesn’t get elevated to a felony, unless it exceeds $1200. Conveniently, most of the fraudulent transactions never exceed that value. Additionally, attackers seem to attack mostly during time.

Customer demographics add another layer of insight, revealing that fraud does not discriminate much by gender but does show a slight preference for older demographics, as suggested by the average ages of fraud victims.

The modeling part of the analysis revealed the inherent challenges in fraud detection. Even with a completely balanced dataset, while recall increased significantly, indicating better detection of fraudulent transactions, precision suffered, leading to a higher rate of false positives. Which for card user’s and credit card fraud detectors is not ideal, as it is an extreme inconvenience to customers to flag a card for a legitimate charge. This illustrates the delicate balance necessary in fraud detection systems to avoid misclassifying legitimate transactions as fraudulent.

Conclusion

The findings of this data analysis offer a multifaceted view of fraud, dismantling some widespread assumptions while affirming others. The earlier belief that fraud mainly originates from certain dubious merchants was rebuffed, as the data depicted a wider berth of merchant categories prone to fraudulent activities. The subtle age differences between fraud and non-fraud cases debunk the stereotype of the elderly as prime targets. Furthermore, the customer-merchant distance was similar for both fraudulent and non-fraudulent transactions, debunking

the myth that distance is a primary fraud determinant. From a modeling perspective, while efforts to balance the dataset improved the identification of fraudulent transactions, they also inflated the rate of false positives, underscoring the intricate balance necessary in fraud detection systems. These insights not only reshape our understanding of fraud but also emphasize the importance of sophisticated, multifactorial approaches to fraud detection that can adeptly navigate between accuracy and user convenience.

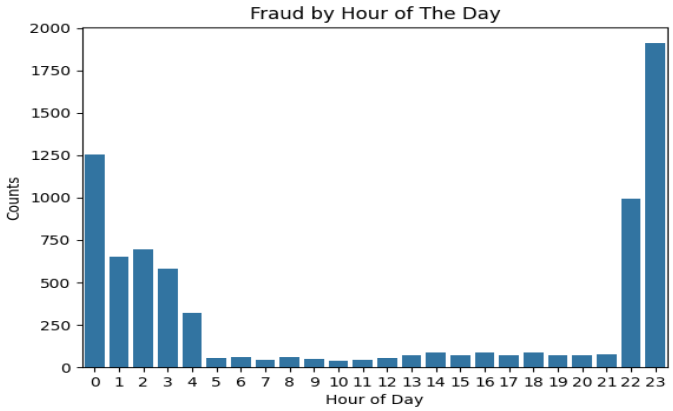
Appendix

a. Transaction Amount

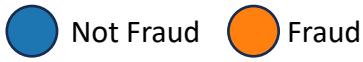
Describe Amount		
Statistic	**Fraud Cases**	**Not Fraud**
Count	7,506	1,289,169
Mean (average)	$531.32	$67.67
Std (std dev)	$390.56	$154.01
Min	$1.06	$1.00
25% percentile	$245.66	$9.61
50% percentile (Median)	$396.51	$47.28
75% percentile	$900.88	$82.54
Max	$1,376.04	$28,948.90

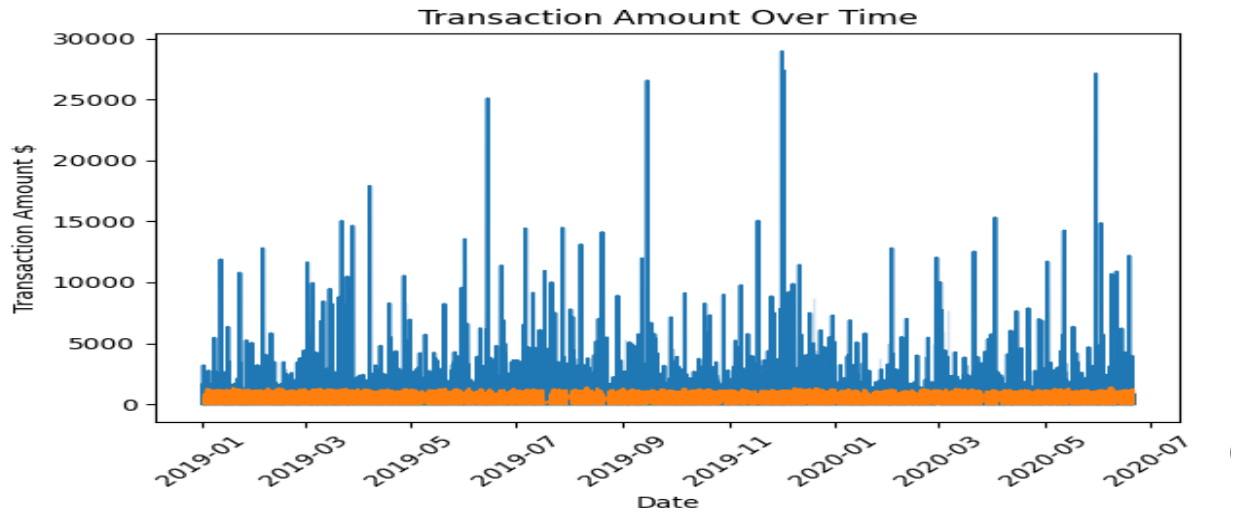
b. Fraud Transaction Hour

Top Fraud Hours	
Hour of Transaction	**Count**
11:00pm	1910
12:00am	1255
10:00pm	993
02:00am	694
01:00am	654
03:00am	584



c. Transaction Amount over time

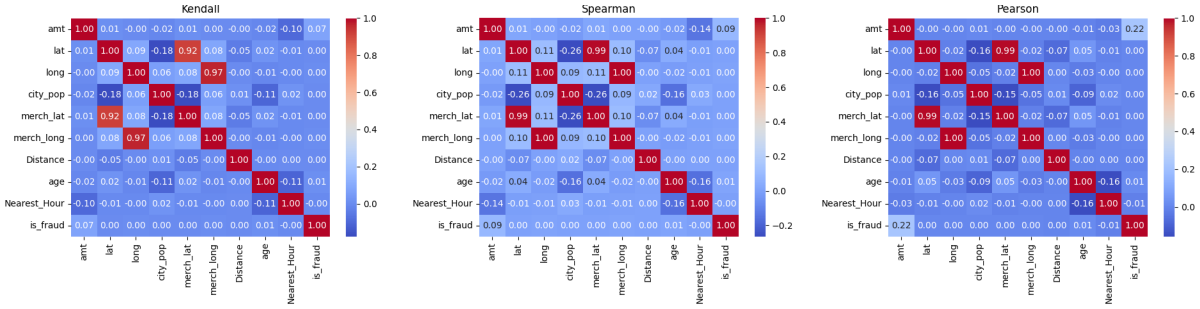




d. Transaction Merchant Category (num in category / total num * 100)

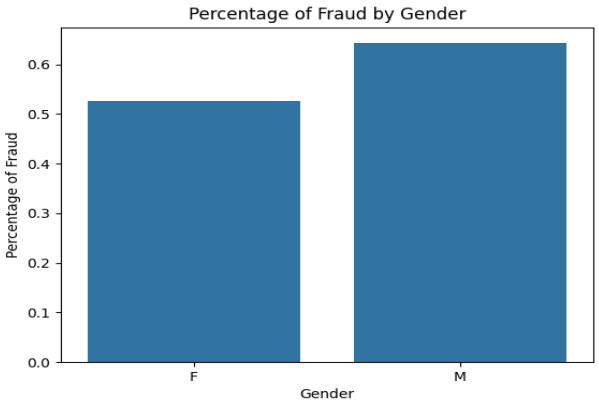
Categorical Percent Representation in Data		
Category	**Percent Represented in Fraud**	**Percent Represented in Not Fraud**
Grocery POS	23%	9%
Shopping Online	23%	7%
Misc. Online	12%	5%
Shopping POS	11%	9%
Gas & Transport	8%	10%
Misc. POS	3%	6%
Kids & Pets	3%	9%
Entertainment	3%	7%
Personal Care	3%	7%
Home	3%	10%
Food & Dining	2%	7%
Grocery Online	2%	4%
Health & Fitness	2%	7%
Travel	2%	3%

e. Correlation Heat Maps



f. Fraud by Gender (num fraud / num not fraud * 100)

Number Male Not Fraud	583,041
Number Female Not Fraud	706,128
Number Male Fraud	3771
Number Female Fraud	3735

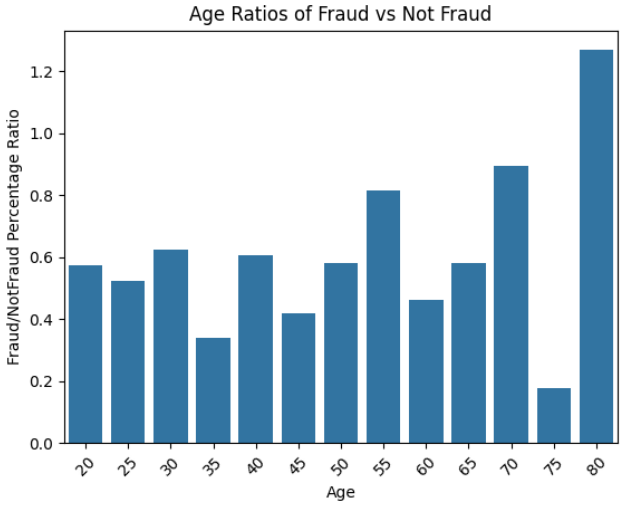


g. Fraud by Merchant (num is merchant / num isn't merchant * 100)

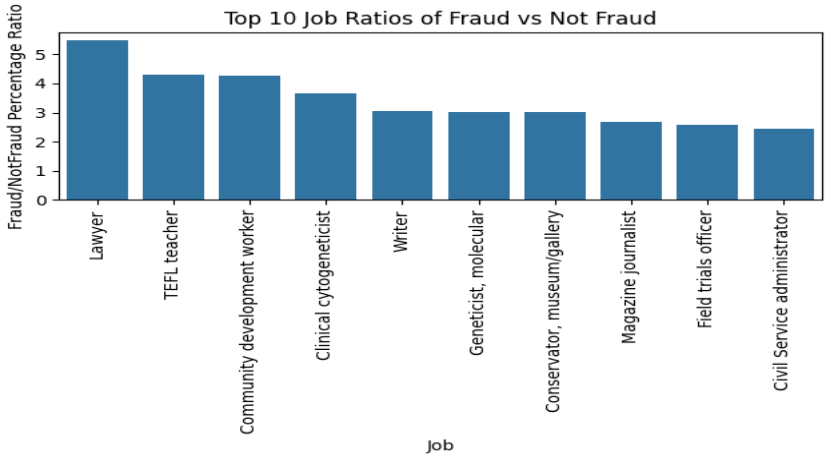
Top 10 Merchant Proportions			
Merchant	**Percent Represented in Fraud**	**Merchant**	**Percent Represented in Not Fraud**
Rau and Sons	0.65%	Kilback LLC	0.34%
Cormier LLC	0.64%	Schumm PLC	0.28%
Kozey-Boehm	0.64%	Cormier LLC	0.28%
Doyle Ltd	0.63%	Kuhn LLC	0.27%
Vandervort-Funk	0.63%	Boyer PLC	0.27%
Kilback LLC	0.63%	Dickinson Ltd	0.27%
Padberg-Welch	0.59%	Cummerata-Jones	0.21%
Kuhn LLC	0.59%	Kutch LLC	0.21%
Terry-Huel	0.57%	Olson, Becker and Koch	0.21%
Koepp-Witting	0.56%	Stroman, Hudson & Edan	0.21%

h. Fraud by Age (num fraud / num not fraud * 100)

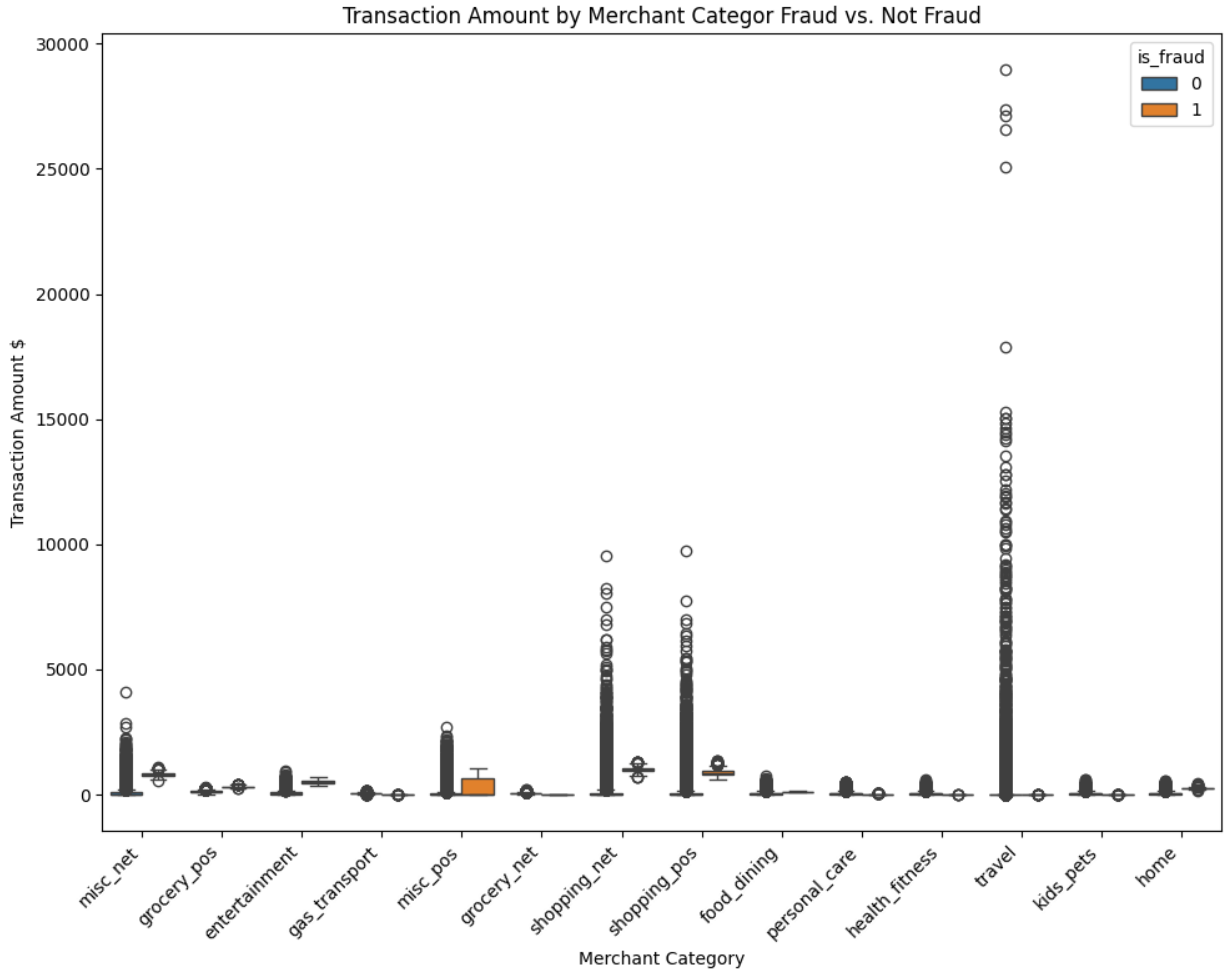
Describe Age by Fraud		
Metric	**Fraud Age**	**Not Fraud Age**
Count	7506.00	1289169.00
Mean	48.87	46.01
STD	18.86	17.37
Min	14.00	14.00
25%	33.00	33.00
50%	48.00	44.00
75%	61.00	57.00
Max	94.00	96.00



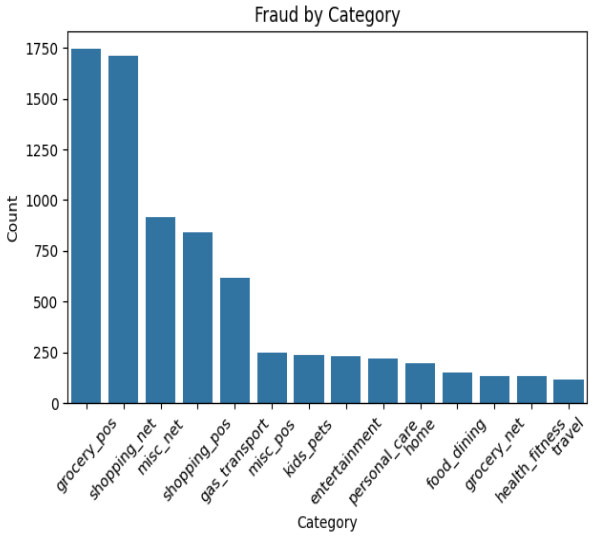
i. Top 10 Jobs Fraud Ratios (num fraud / num not fraud * 100)

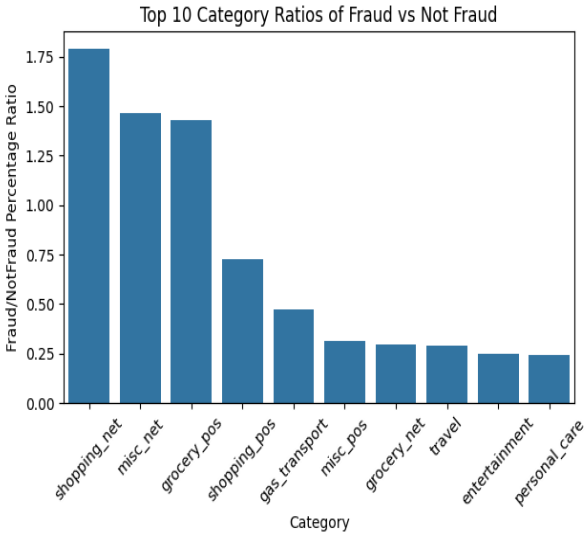


j. Transaction Amount by Category (not fraud on left fraud on right)



k. Fraud by Category & Top Ten Category Ratios (num fraud / num not fraud * 100)





l. Fraud by Distance

Describe Customer Merchant Distance		
Metric	**Fraud Distance**	**Not Fraud Dist**
Count	7,506	1,289,169
Mean	76.27	76.11
Std	28.73	29.09
Min	0.74	0.02
25%	55.66	55.36
50%	78.00	78.26
75%	98.34	98.47
Max	144.36	151.87

m. Machine Learning Models

Model Performance with The Original Unbalanced Data				
	Precision	**Recall**	**F1-Score**	**Support**
Not Fraud	1.00	0.97	0.99	553574
Fraud	0.07	0.51	0.12	2145
Accuracy			0.97	555719
Macro AVG	0.53	0.74	0.56	555719
Weighted AVG	0.99	0.97	0.98	555719

Model Performance with 1/10th of The Original Not Fraud Data				
	Precision	**Recall**	**F1-Score**	**Support**
Not Fraud	1.00	0.97	0.99	553574
Fraud	0.08	0.51	0.12	2145
Accuracy			0.97	555719
Macro AVG	0.54	0.79	0.56	555719
Weighted AVG	0.99	0.97	0.98	555719

Model Performance with Completely Balanced				
	Precision	**Recall**	**F1-Score**	**Support**
Not Fraud	1.00	0.90	0.95	553574
Fraud	0.03	0.88	0.06	2145
Accuracy			0.90	555719
Macro AVG	0.52	0.89	0.51	555719
Weighted AVG	0.99	0.90	0.94	555719

n.

Describe Transaction Amount (11pm-4am)		
Statistic	**Fraud Cases**	**Not Fraud**
Count	6,409	352,269
Mean	$529.22	$69.81
Std	$389.62	$145.74
Min	$1.06	$1.00
25%	$247.47	$10.71
50% (Median)	$381.39	$51.43
75%	$898.72	$85.91
Max	$1,376.04	$26,544.12