Milestone 3: Exploratory Data Analysis

Martin L. Gonzalez

Boston University

DX699_02 AI for Leaders (summer 1)

Joshua Von Korff

June 29, 2025

This project is related to the Banking, Financial Services, and Insurance sector. I decided to go with a Credit Card Fraud Detection dataset from Kaggle. The data provides real world scenarios relating to credit card transactions, with our target variable being “Is_fraud”, we are trying to determine if a transaction is a fraudulent transaction or not. First, what is credit card fraud? Credit card fraud occurs when a person uses someone else’s card or card information to make unauthorized purchases or withdrawals (Office of the Comptroller of the Currency, n.d.).

Credit card fraud can happen by either stealing the physical card and going into stores to use it or by stealing the card information such as the card number and CVV code to make transactions online. How can this happen? Someone can either steal your physical card and use it that way or fraudsters can use a skimming device to capture your credit/debit cards information. Skimming devices are usually small instruments fraudsters attach to chip readers, similar to the ones you see at gas stations or ATM machines, by attaching that small tool to the card reader it will scan your information and store it in that database.

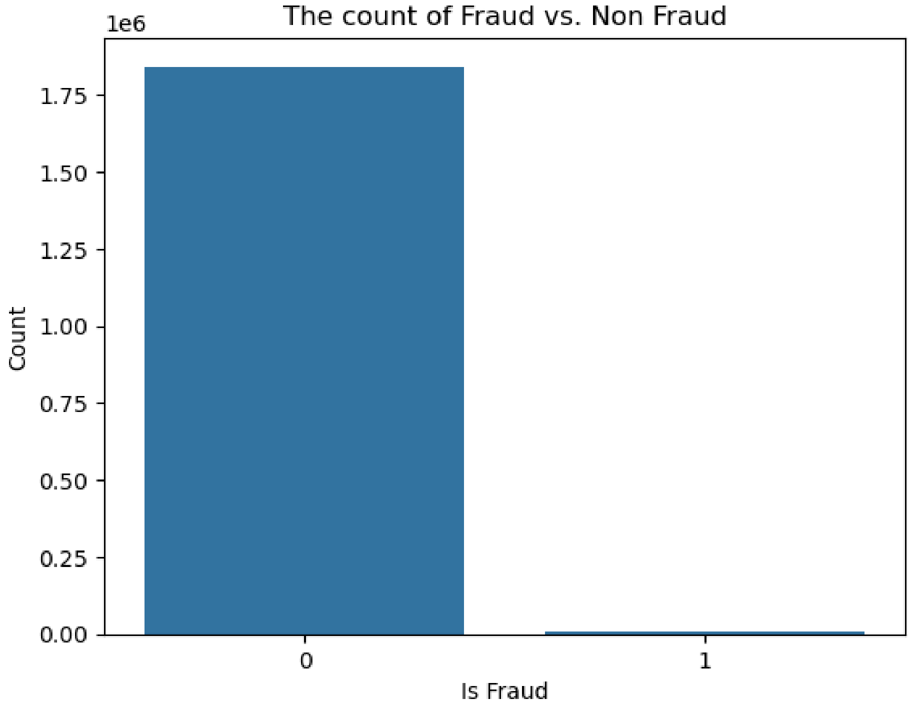
By analyzing this dataset we hope to gain more information into what goes on behind the scenes of a fraudulent transaction. By the end of this program we can hopefully answer the questions: Is credit card fraud targeted? Do fraudsters prey on women, men, certain age demographics? Is there something in common between fraudulent transactions? Will a specific merchant or area get more fraud? And most importantly, can we do anything to prevent a fraudster from getting my card information?

This dataset has 1,852,394 transactions (rows) and 23 features (columns) with the target column being a binary feature where 1 is a fraudulent transaction and 0 being a non-fraudulent transaction. See columns and what they are below:

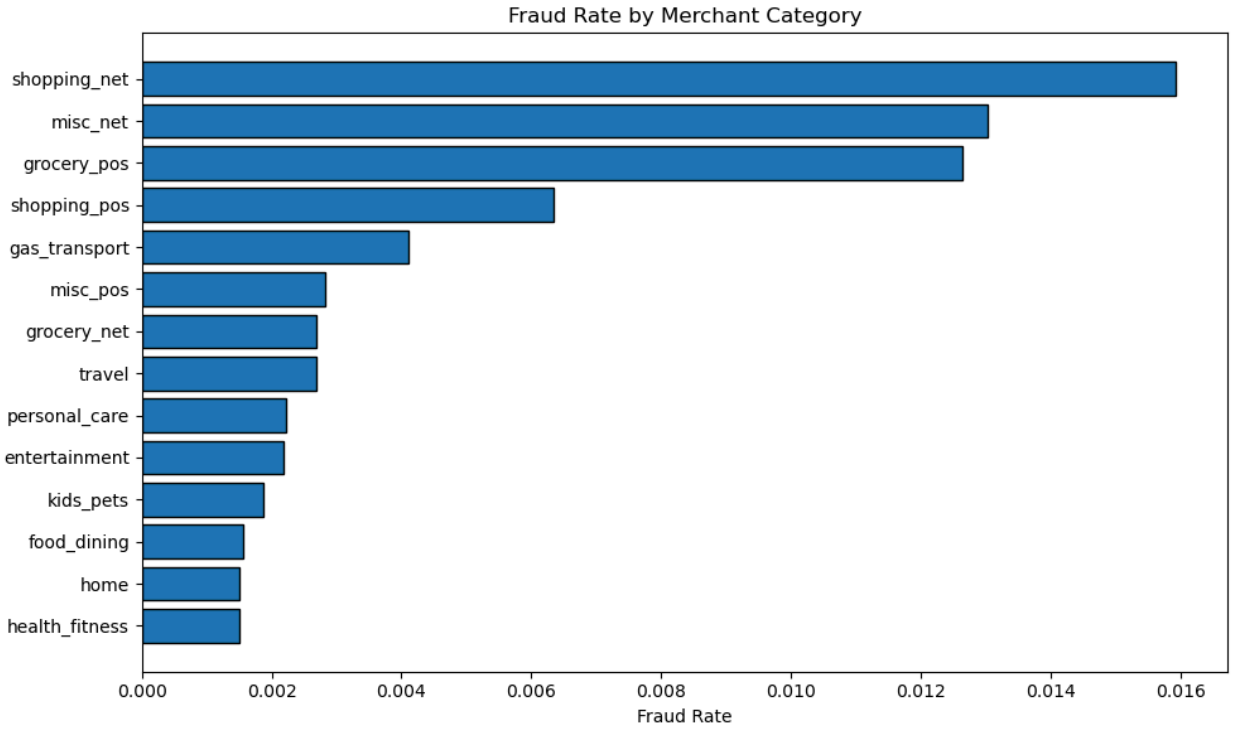
- 1: trans_date_trans_time; This is the time the transaction took place.

- 2: cc_num; The credit card number.
- 3: merchant; This is where the transaction occurred.
- 4: Category; Shows the category type (groceries, gas, misc, etc)
- 5: amt; The amount of the transaction in dollars.
- 6: first; First name on the card.
- 7: last; Last name on the card.
- 8: gender; Gender of the cardholder.
- 9: street; Street the card holder lives on.
- 10: lat: Geographical location of the cardholder (latitude)
- 11: long; Geographical location of the cardholder (longitude)
- 13: city; City the cardholder lives in
- 14: state; State the cardholder lives in.
- 15: zip; zipcode of the cardholder's address.
- 16: city_pop; Population of the cardholder’s city.
- 17: job: profession of the cardholder.
- 18: dob; Date of birth of the cardholder.
- 19: trans_num; The unique transaction identifier.
- 20: unix_time: Timestamp of the transaction in Unix Format
- 21: merch_lat: Merchant’s location (latitude)
- 22: merch_long; Merchant’s location (longitude)
- 23: is_fraud; Target Variable (1 = Fraud, 0 = Legit Transaction)

From using explanatory data analysis we are able to determine there are no null values in this data. From the graph below we are able to see the comparison of the fraud transactions and the number of legit transactions. By finding the value counts we are able to determine that 99.48% of the data has legit transactions and 0.52% of the data has fraudulent transactions. Approximately 1,842,742 rows are non fraudulent and 9,652 rows are fraudulent. This tells us that the dataset is highly imbalanced. If we do not handle the imbalance in the data when we train a model it can potentially always predict the majority class (not fraud) which would make it achieve a very high accuracy score by not doing anything useful to properly calculate the prediction.

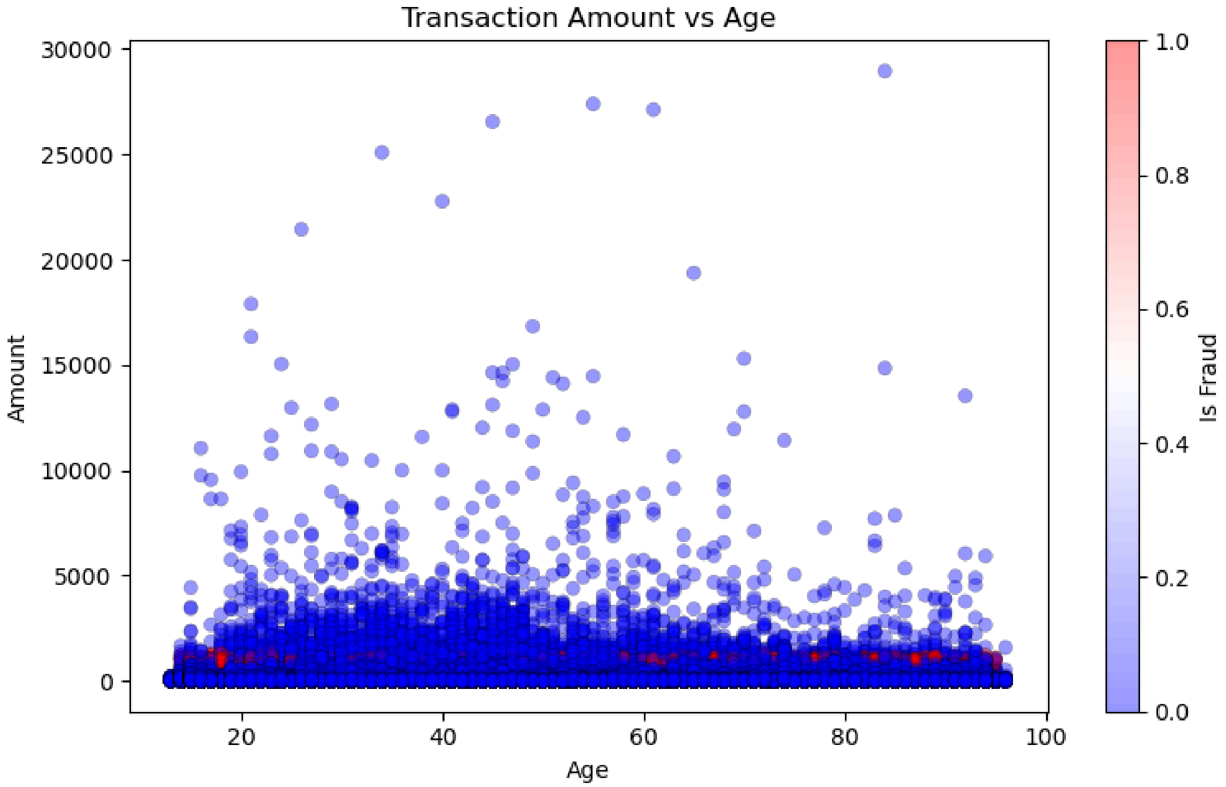


The graph below shows the fraud rate by merchant category. It’s expected that most of the transactions come from shopping as that’s what people spend their money on most of the time.



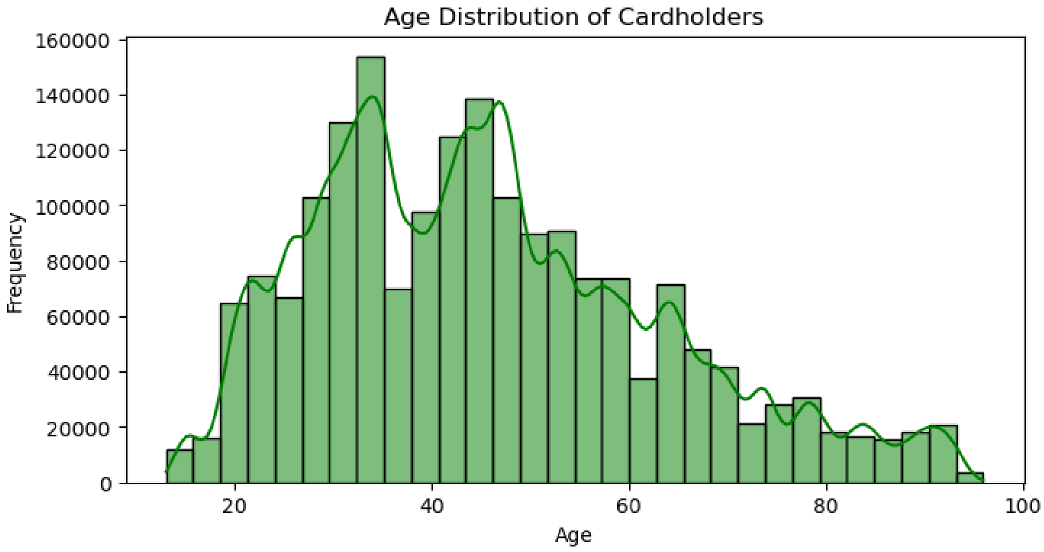
Aspects of the data that were expected/unexpected:

In this data I did not expect such a high imbalance of the target variable. I would have assumed there would have been a lot more transactions (rows) regarding fraud. From the graph below (Transaction Amount vs Age) I expected a higher amount of purchases regarding individuals in their 30s-40s. In the US the average person buys a home at 32 years old and has a kid around 30 years old so I was very surprised to see most of the transactions were below $10,000. However, for the age range 20-30 I expected it to be below $10,000 since that's usually when people graduate college and get their first job, so it's expected that there are smaller purchases rather than bigger purchases for the smaller age range.

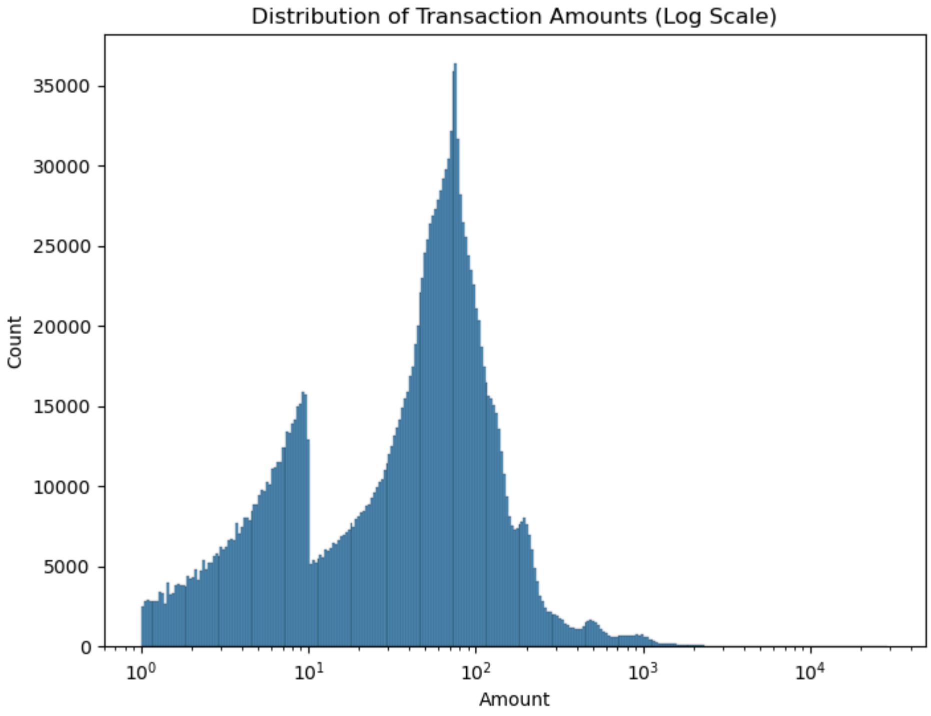


Explanation of univariate analysis:

By performing univariate analysis I was able to find the age distribution of the cardholders. Where the average age is 46 years old, the minimum is 13 years old, and the oldest is 96.

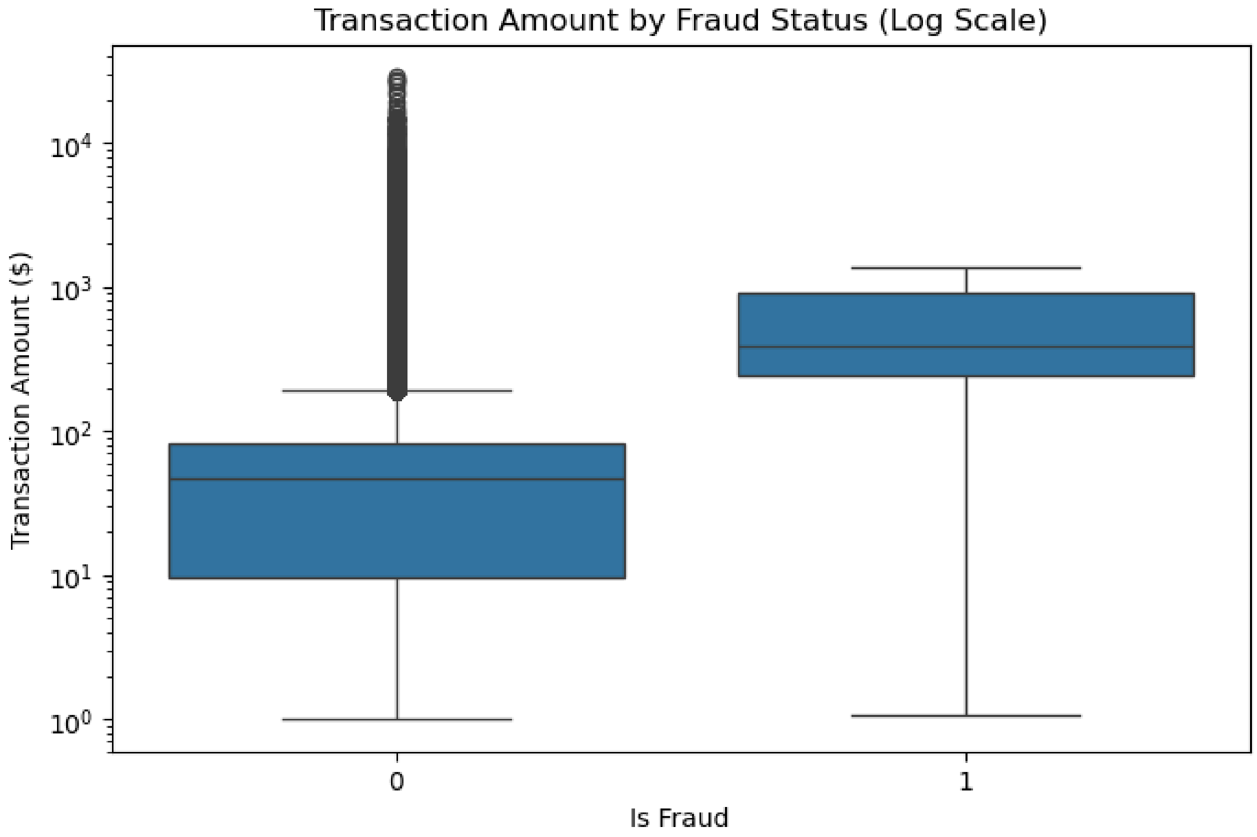


Another thing I was able to find is the distribution of the transaction amounts. Where the x-axis is 10^0, 10^1, 10^2, etc. It shows the amount of the transaction, $1, $10, $100, etc.

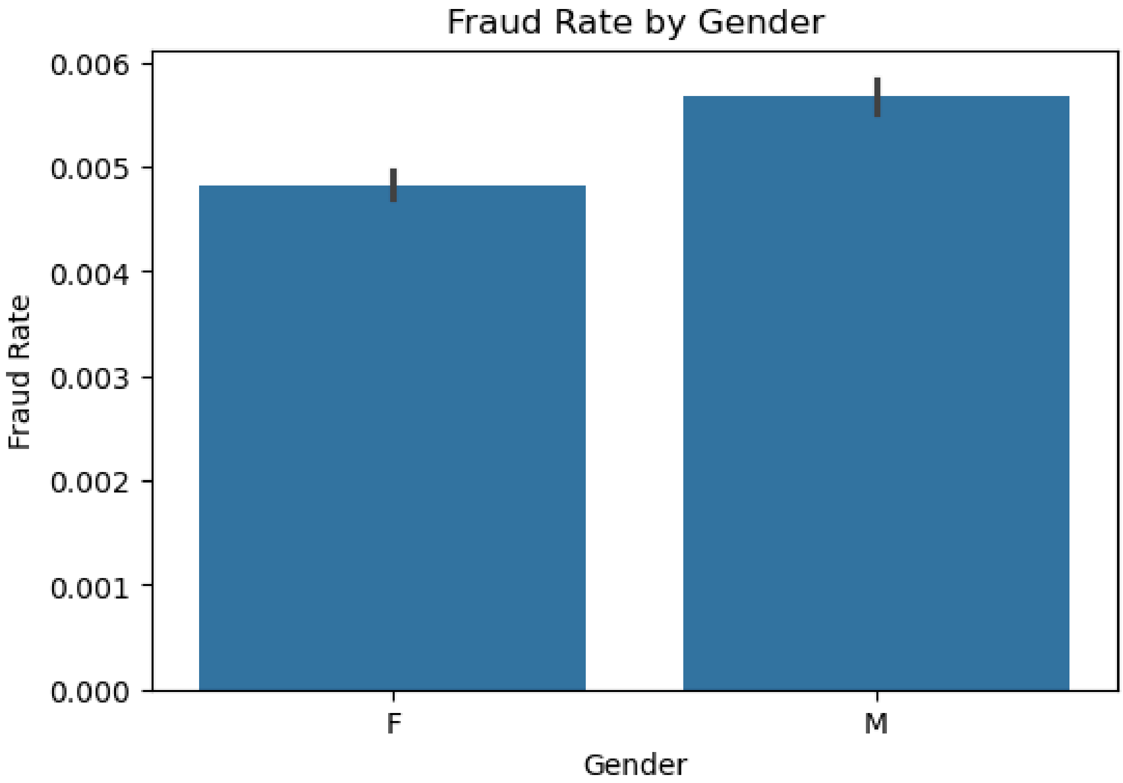


Explanation of bivariate analysis:

What we can tell from bivariate analysis is the comparison between multiple features. For example, these boxplots show the “is_fraud” column with the Transaction amount.



The graph below shows the fraud rate by gender. What we can gather from this plot is Male’s will often get hit more with a fraudulent transaction than Females.



Summary of results 2, 4, 6:

From week 2 I was able to gather that there are no duplicated values, there are no missing values, and I was able to see the percentage of the "is_fraud" column. From week 4 I was able to perform more in depth analysis into how often a name appeared in the dataset; the average amount of money an age group spent; and a summary based on age, transaction amount, account balance, and the target variable. Lastly, on week 6 I was able to show more graphs on the data. Specifically histograms showing the distributions of certain features on a different dataset. I decided to do a different dataset on week 6 because I wasn't sure if I was supposed to keep using the same dataset but mainly I wanted to broaden my skillset and test my data analysis skills into different datasets.

Going forward, I will continue to perform explanatory data analysis on my focused dataset for the project. Hoping to utilize regression models to dig deeper into the data, potentially find more correlated values and find the weight of which features contribute the most to our prediction.

References

Office of the Comptroller of the Currency. (n.d.). *Credit Card and Debit Card Fraud*. OCC.gov. https://www.occ.gov/topics/consumers-and-communities/consumer-protection/fraud-resources/credit-card-and-debit-card-fraud.html