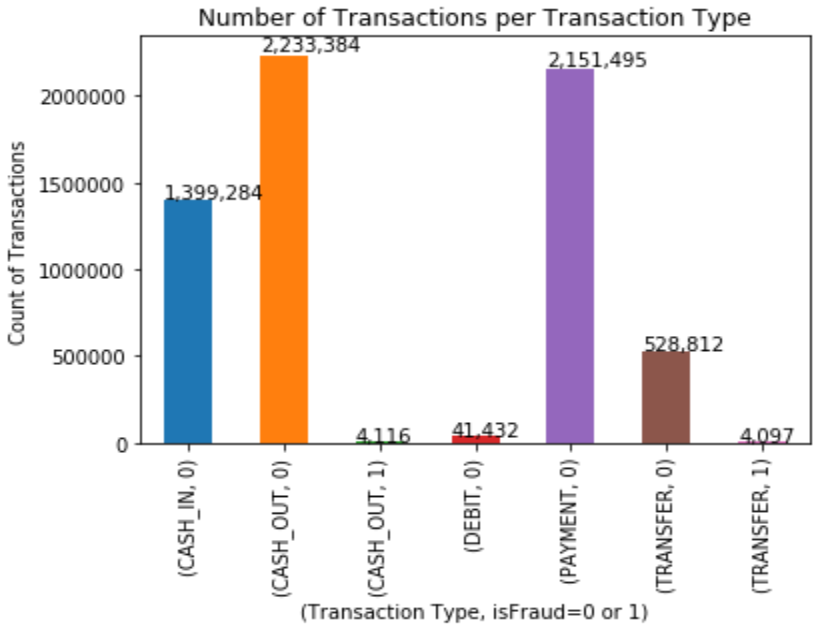
Capstone 1 Project : Data Storytelling
Fraud Detection in Mobile Payment Data
Springboard Data Science Career Track
Eric Cruz

To recap the Data Wrangling exercise from earlier, note there was no special processing required with respect to cleaning up missing data, zeroes, or outliers. However, the data was found to be highly imbalanced, with only 8,213 out of 6,362,620, or 0.13% identified with isFraud=1, which is the target prediction value. There are 2 sets of 16 observations which may require special attention:

- The value isFlaggedFraud =1 only occurs 16 times out of 6,362,620 and shows a clear relationship between the amounts and balances for these records. All of these records coincide with isFraud=1, so it isn't clear whether that makes isFlaggedFraud redundant.
- There are 16 observations with Amount=0, and isFraud=1. These may need to be excluded, as it is not clear whether it makes sense to consider a transaction as legitimate with a zero amount.
- The type distribution where isFraud=1 is limited to only CASH_OUT and TRANSFER types. See the bar chart which includes the fraud transactions in separate categories.

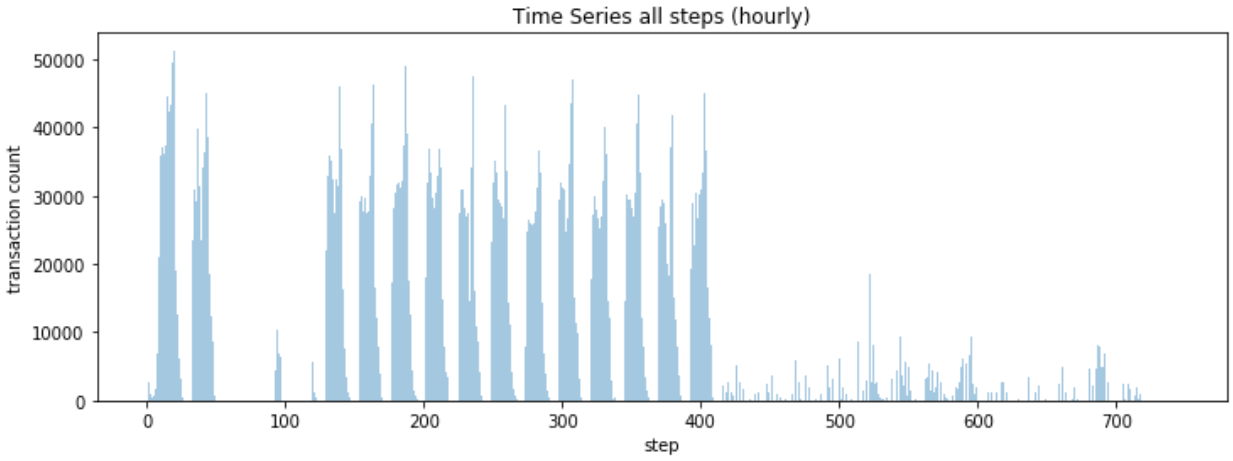


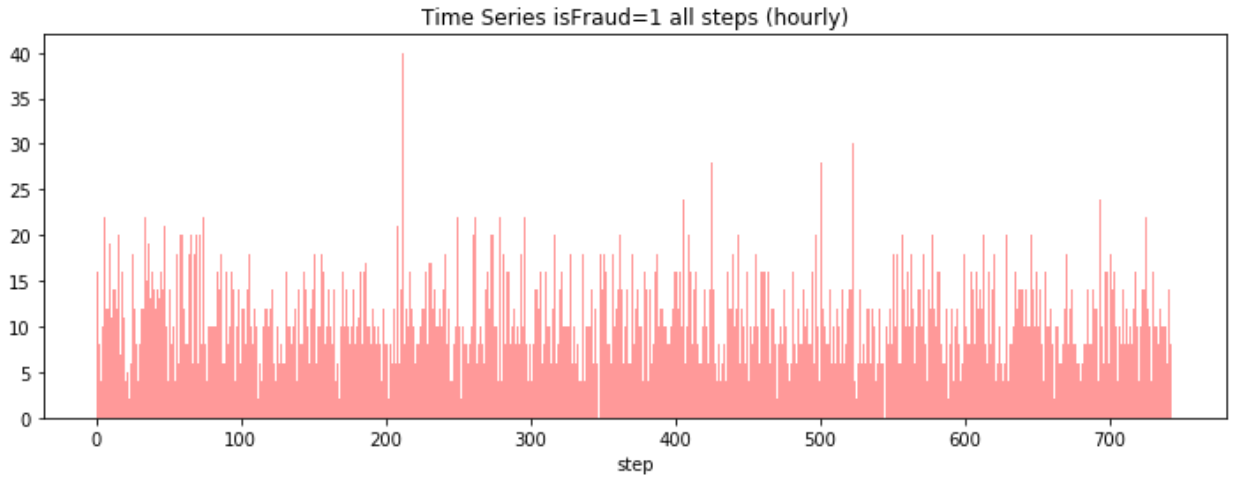
Some additional data stories can be observed, which lead to a few predictions about features which could help detect fraud.

1. **The timing of fraud transactions appears to be steady and consistent, contrary to other activity which contains gaps of varying sizes between active time slots.**
2. **The transaction "Amount" is highly correlated with the "OldBalanceOrig" for fraud transactions but not correlated for other transactions**
3. **The Amount on fraud transactions is in a smaller, tighter range than the distribution of amounts for all transactions.**

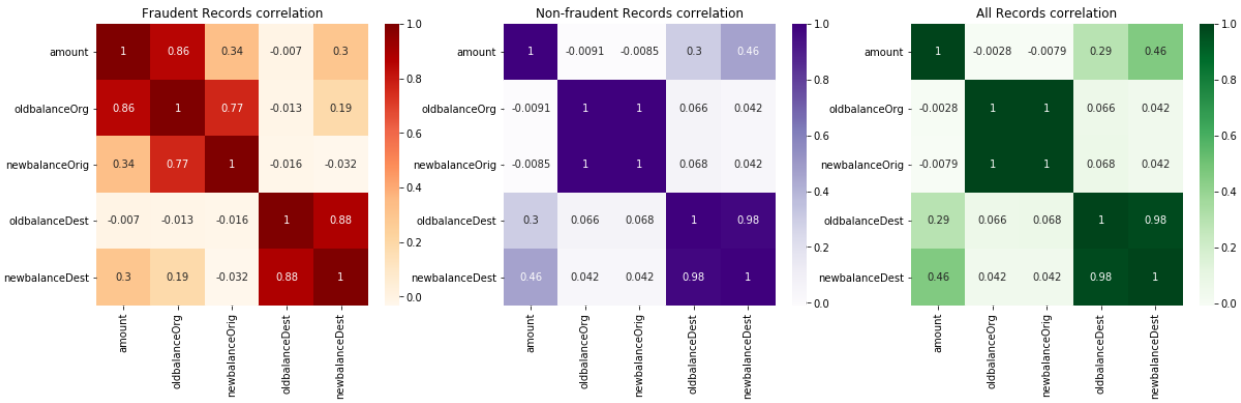
Note the following descriptions of the visualizations for the summary points above:

- The frequency distribution of transactions over time (743 steps, which is 1 hour less than 24 hours * 31 days) shows no more activity after step 718 (the final 24 hours or full last day of the data period). The frequency distribution where isFraud=1 is pretty even across the entire timeframe, which does not correspond to the non-fraud transaction distribution which contains gaps and tails off after about 400 out of 743 steps.

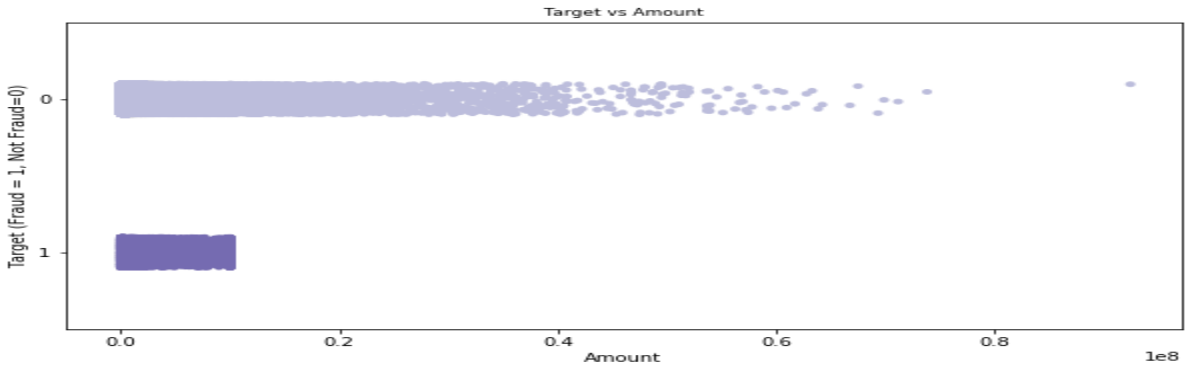




- The correlation heatmaps showing the fraud transactions, non-fraud, and combined results are plotted side-by-side. Note that fraud transactions show significantly higher correlation between “Amount” and “OldBalanceOrig” and slightly higher correlations between other features, as highlighted by the grid shading.



- See the stripplot of Amount for fraud vs non-fraud transactions. It suggests that amounts for fraud are generally on a smaller scale than other transactions.



- The step value was converted to date/time format, setting the first step to the default value of the first hour of 1/1/1970. Although the assumption that the step starts on a specific day or hour may be wrong, any patterns detected could easily be shifted to different starting point. It turns out that day of week does not appear to reveal a different pattern for fraud transactions, but the hourly activity of fraudulent transaction remains steady througout the period.