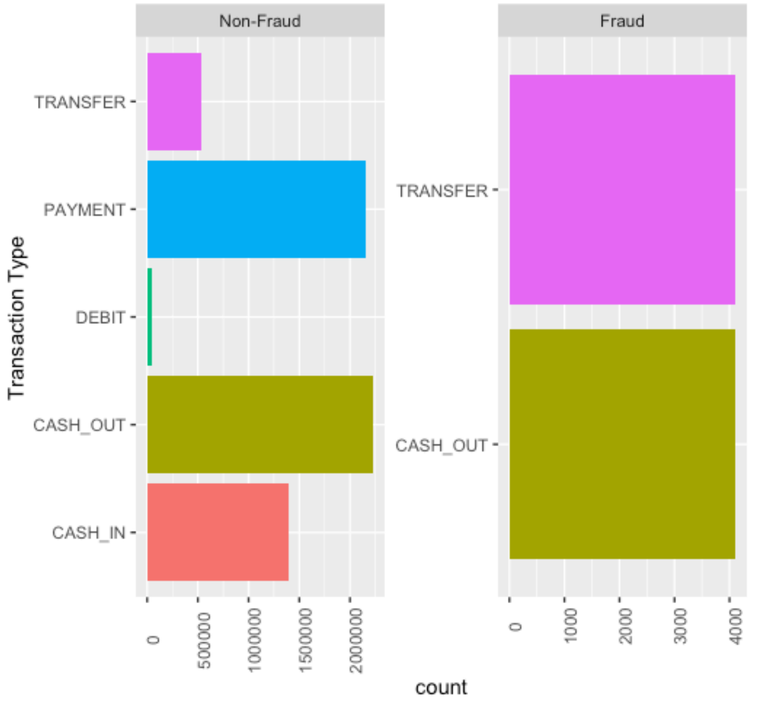
Preliminary Analysis of a Synthetic Financial Dataset

What does the dataset have?

It has simulated mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country.

What kind of transactions are fraudulent?

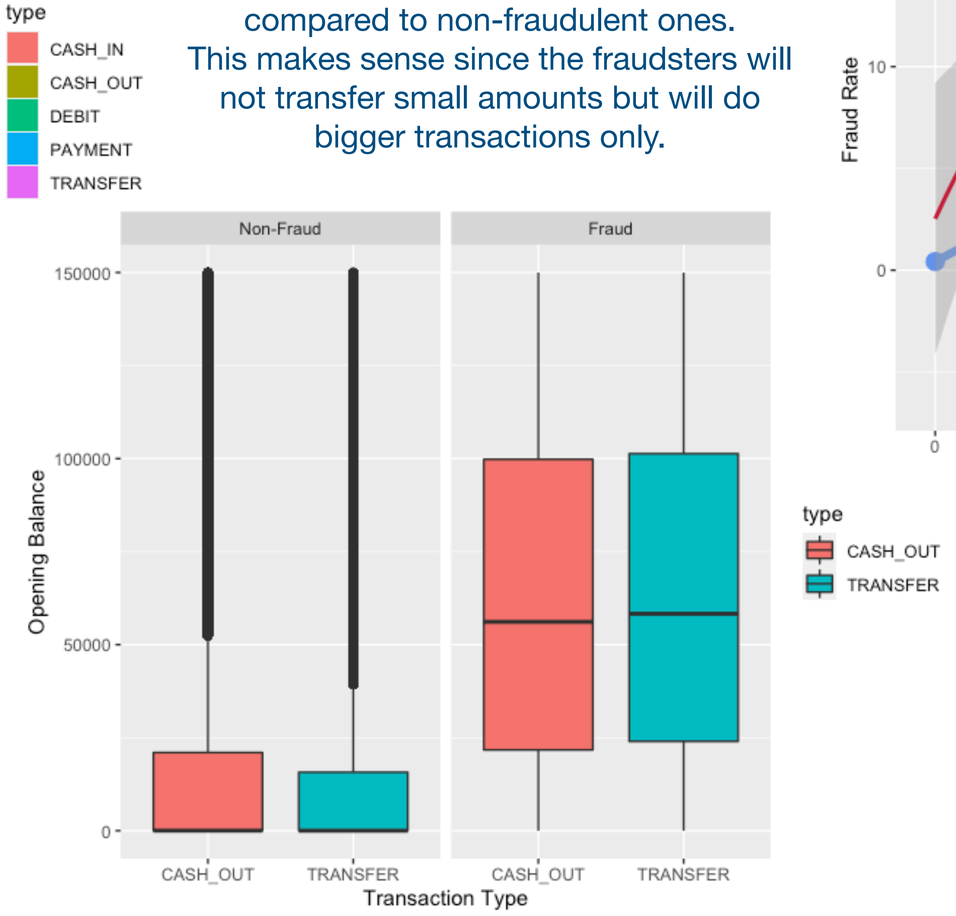


If we look at the plot above then we can see that the type of fraudulent transactions belong to only two categories: Transfer and Cash-Out

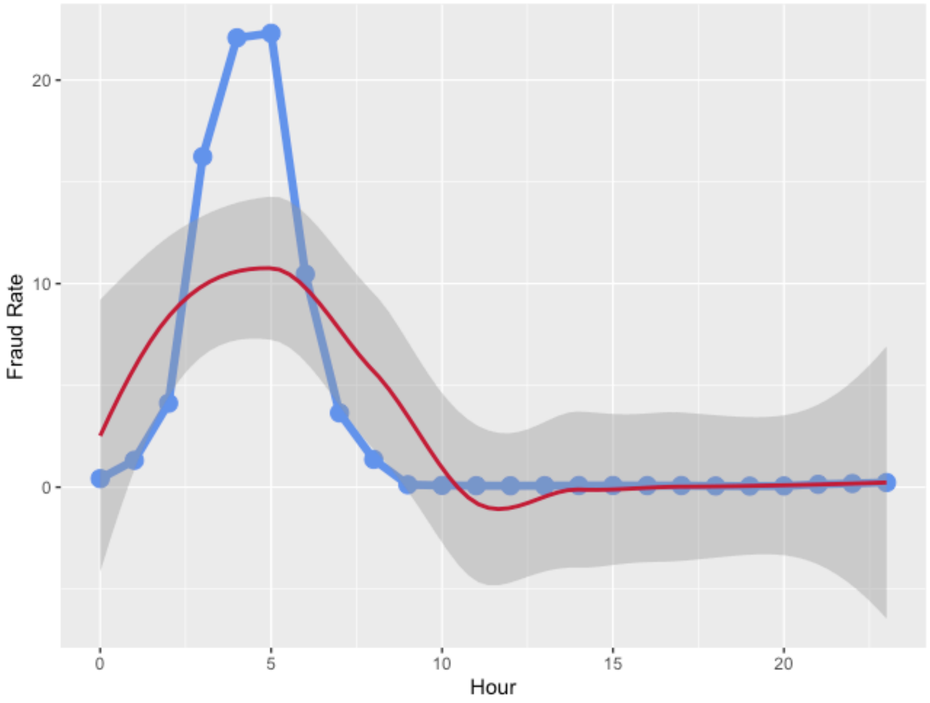
In fact, a close look at the scale in both fraudulent and non-fraudulent transactions tells use that the amount of fraudulent transactions are very less compared to the non-fraudulent. In fact, only 8213 observations out of 6362620 are fraudulent.

Is there some visible difference between the opening balance of fraudulent and non-fraudulent transactions?

If we look at the plot below then we can see that the median of opening balance is much higher in fraudulent transactions as compared to non-fraudulent ones.
This makes sense since the fraudsters will not transfer small amounts but will do bigger transactions only.



Does the fraud happen during a specific time of the day?



The data simulation has been done for 30 days and has 744 steps (hours).
On checking the rate of frauds done every 24 hours a pattern was seen where most frauds occur around 5 AM, as can be seen in the plot above. This seems to make sense since a fraudster would pick a time when most people are offline.

Author: Anshul Rao
Data Source: https://www.kaggle.com/ealaxi/paysim1