CSE 4062 – Introduction to Data Science and Analytics
Spring 2021
Delivery #3 - Exploring Data part 2

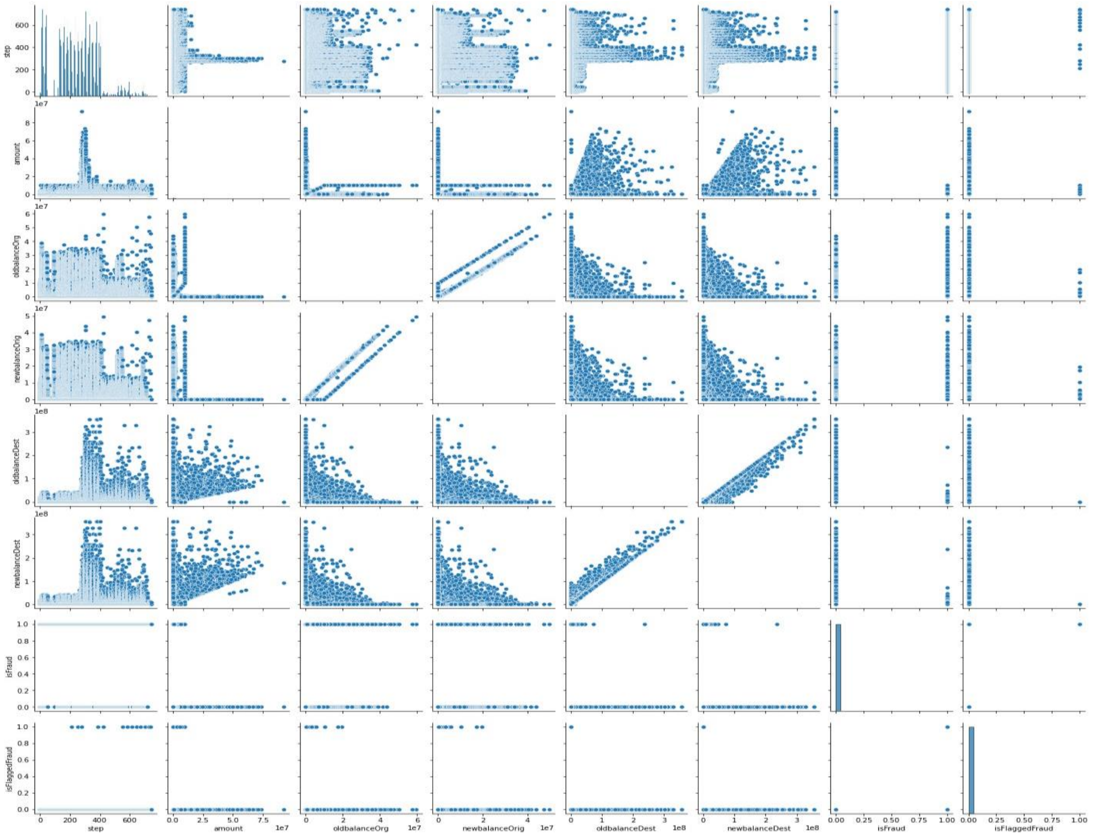
Project Report - Group #8

Group Members:
Caner Dağdaş | 150716001 | canerdagdas@hotmail.com
Ceyhun Vardar | 150317022 | vardarceyhun13@gmail.com
Büşra Gökmen | 150116027 | busragokmen67@gmail.com
Cem Güleç | 150117828 | cem.ggulecc@gmail.com
Ömer Faruk Çakı | 150117821 | omerfarukcaki@gmail.com

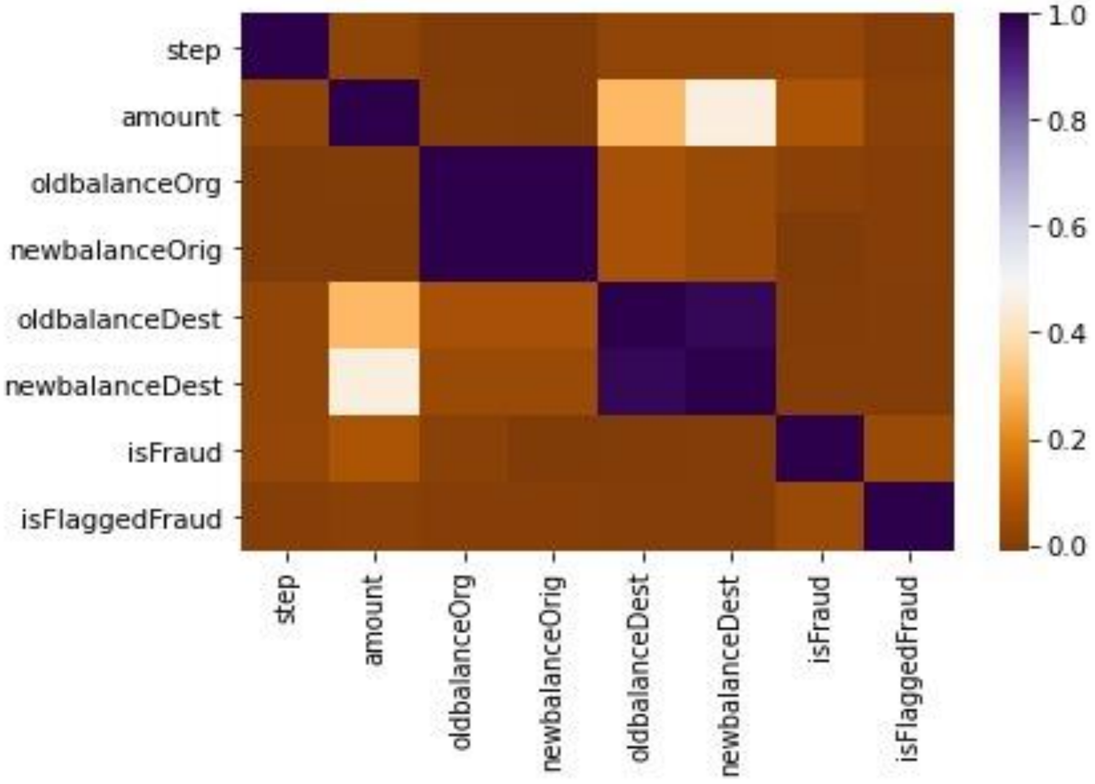
Project Title: Fraud Detection on Financial Data

Lecturer: Assoc. Prof. Murat Can Ganiz

Scatter plot matrix

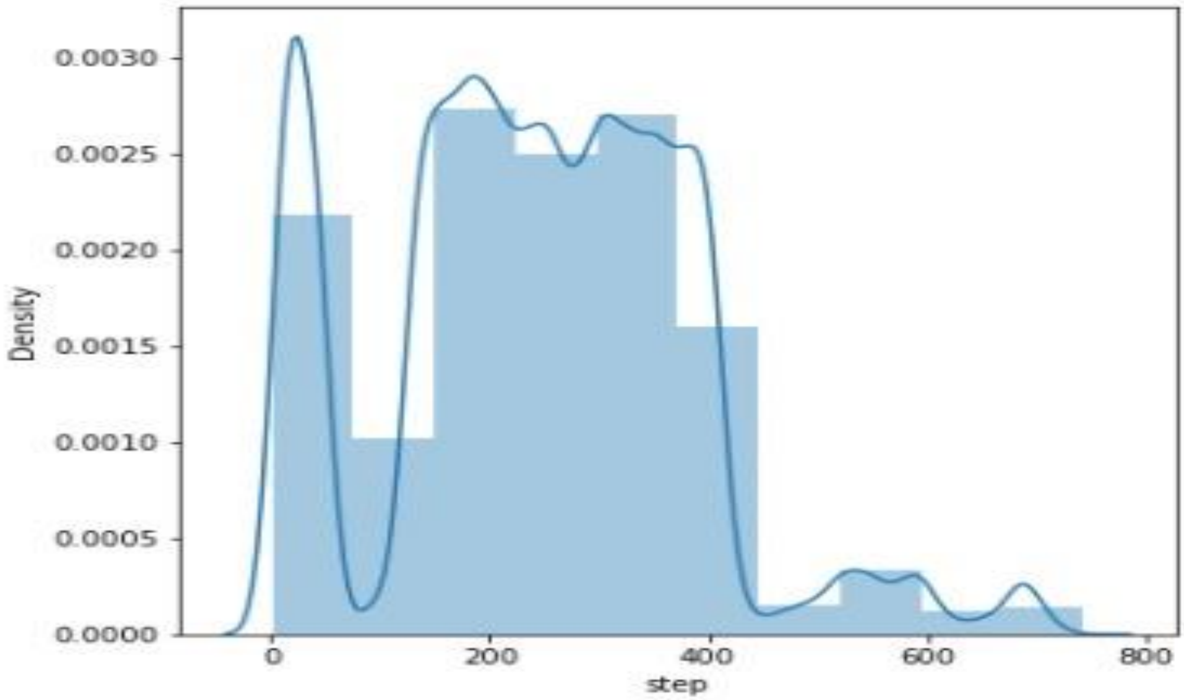


Heat map

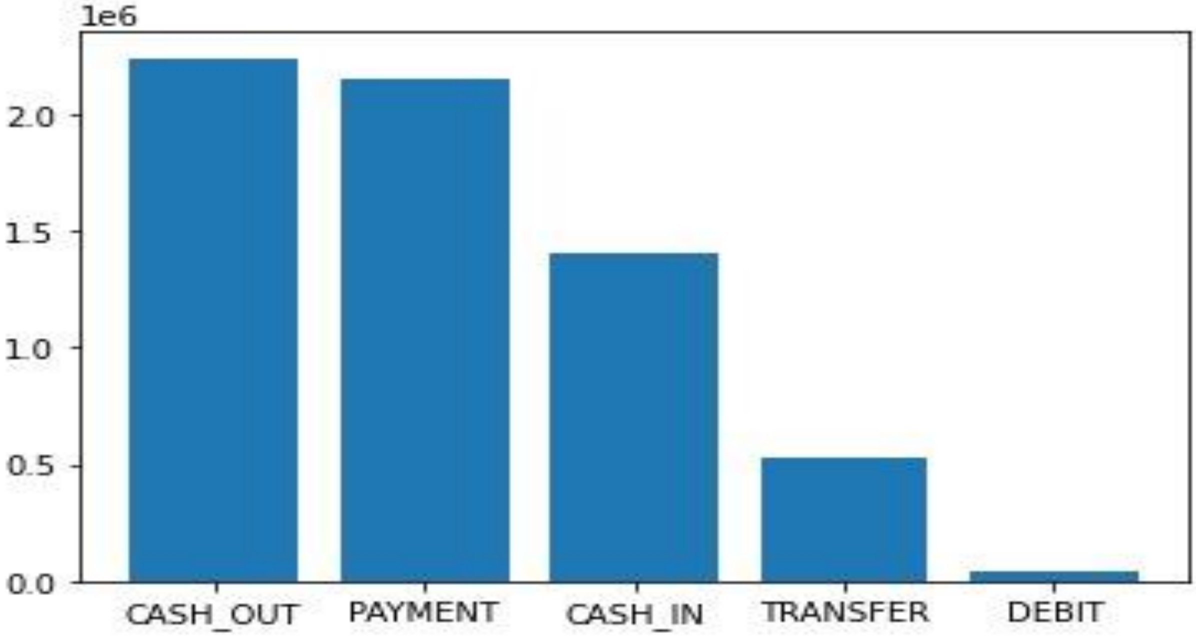


Charts of Attributes

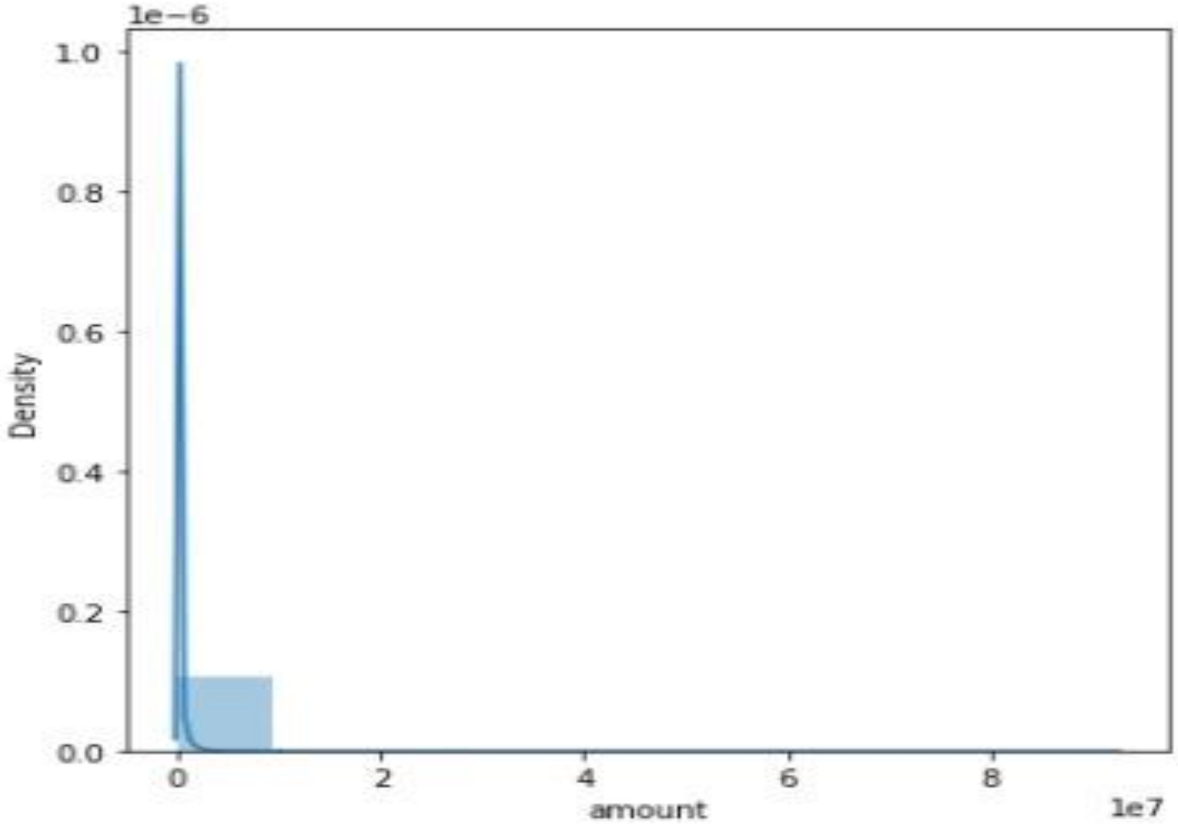
Step



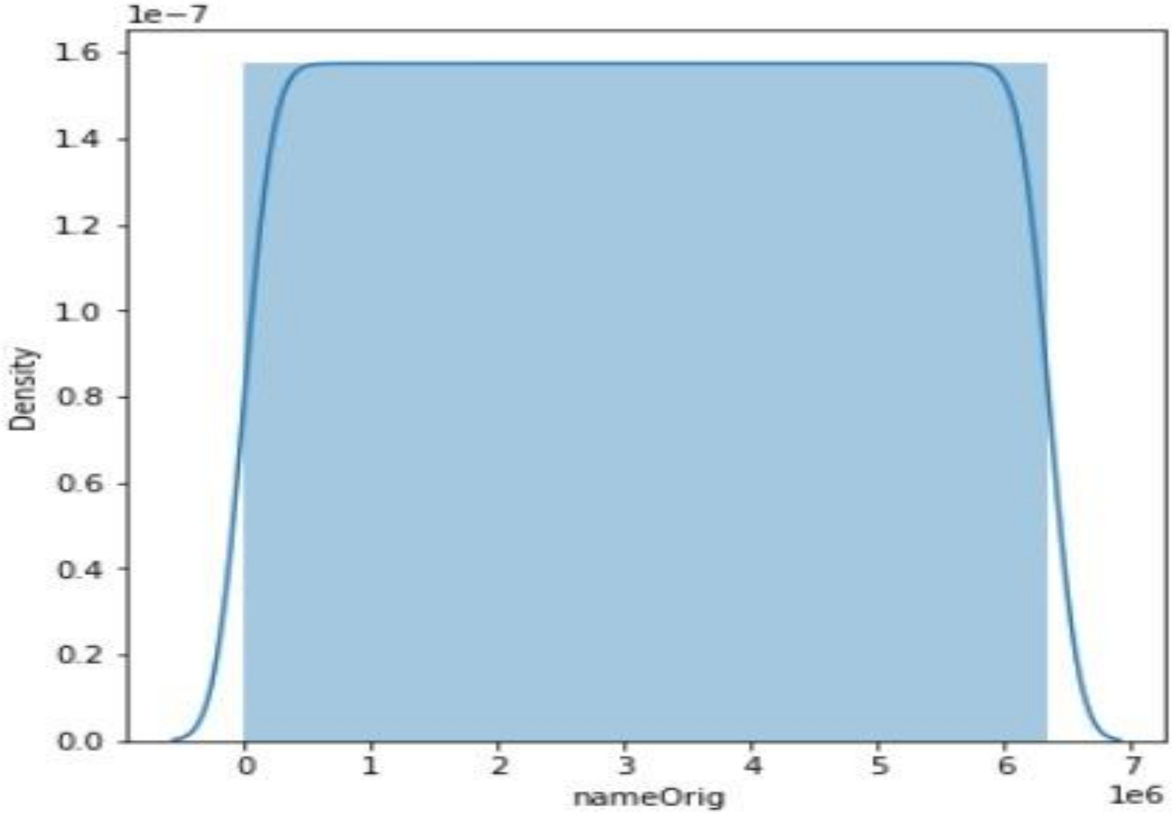
Type



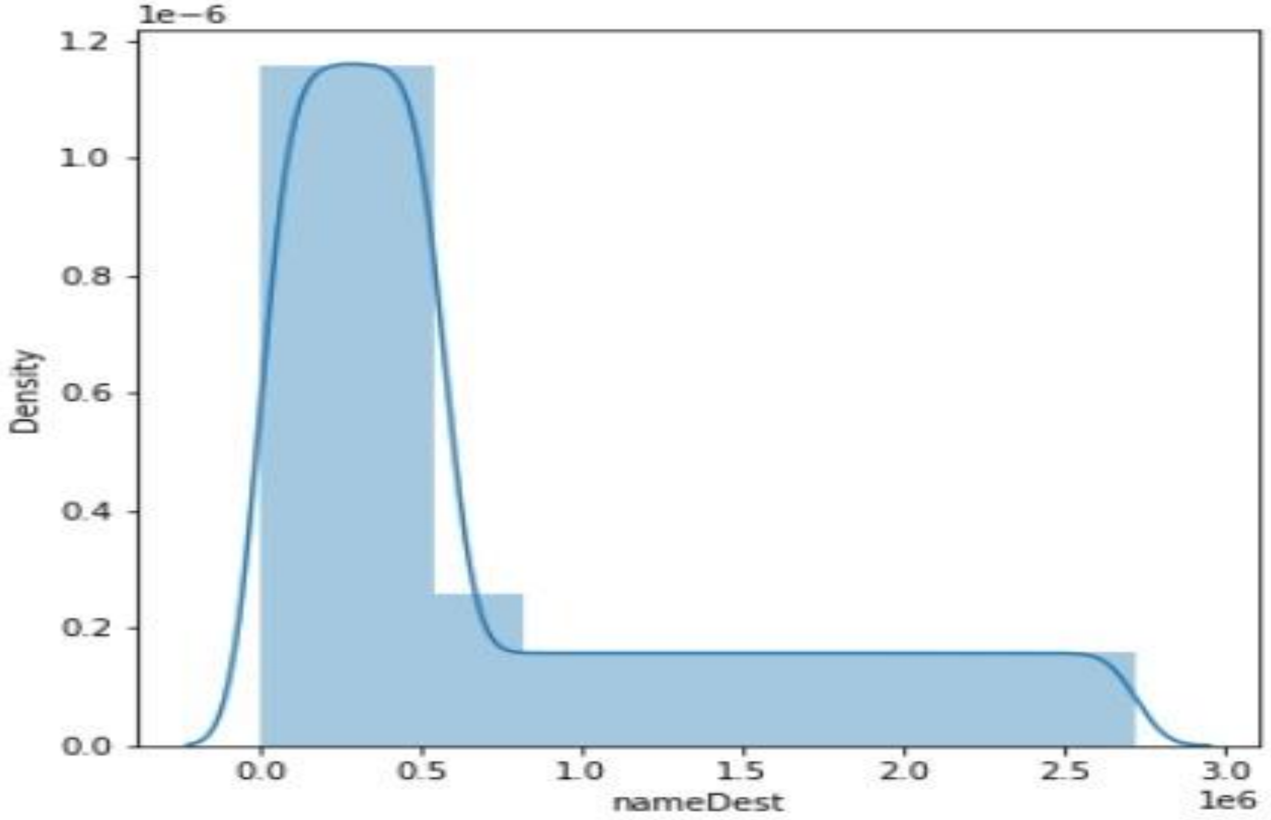
Amount



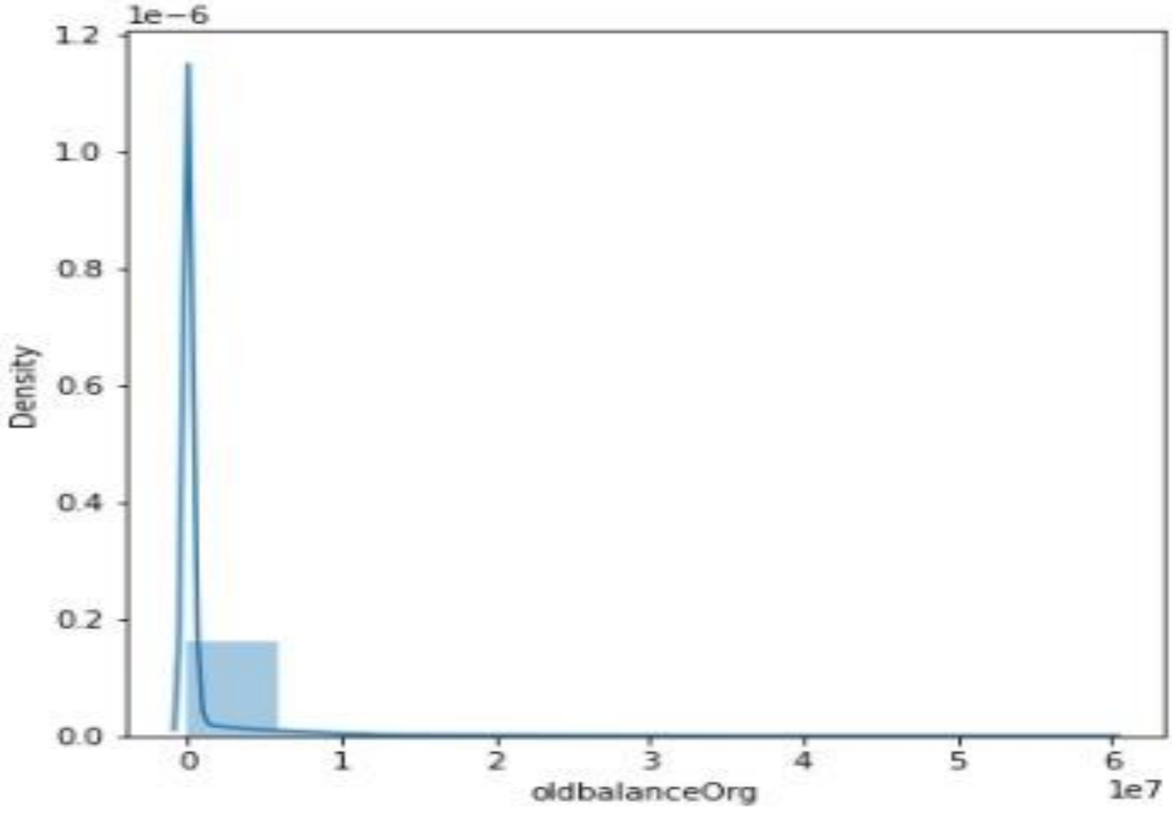
NameOrig



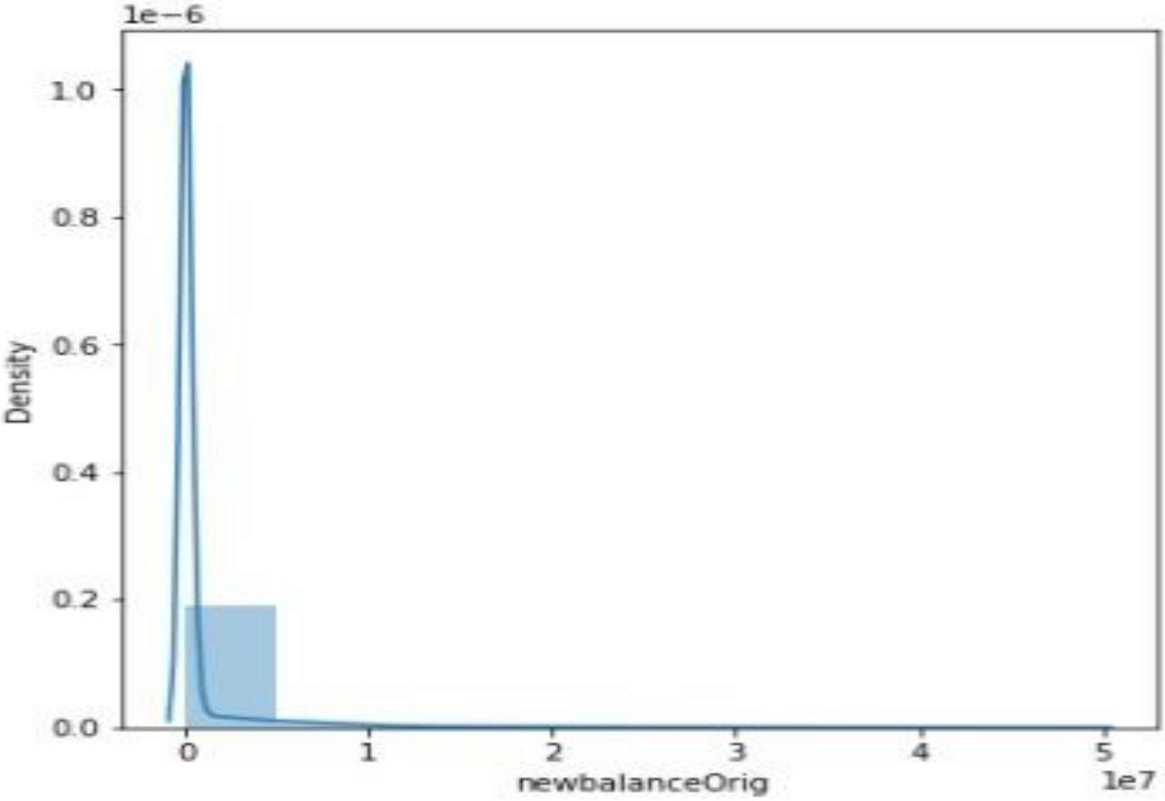
NameDest



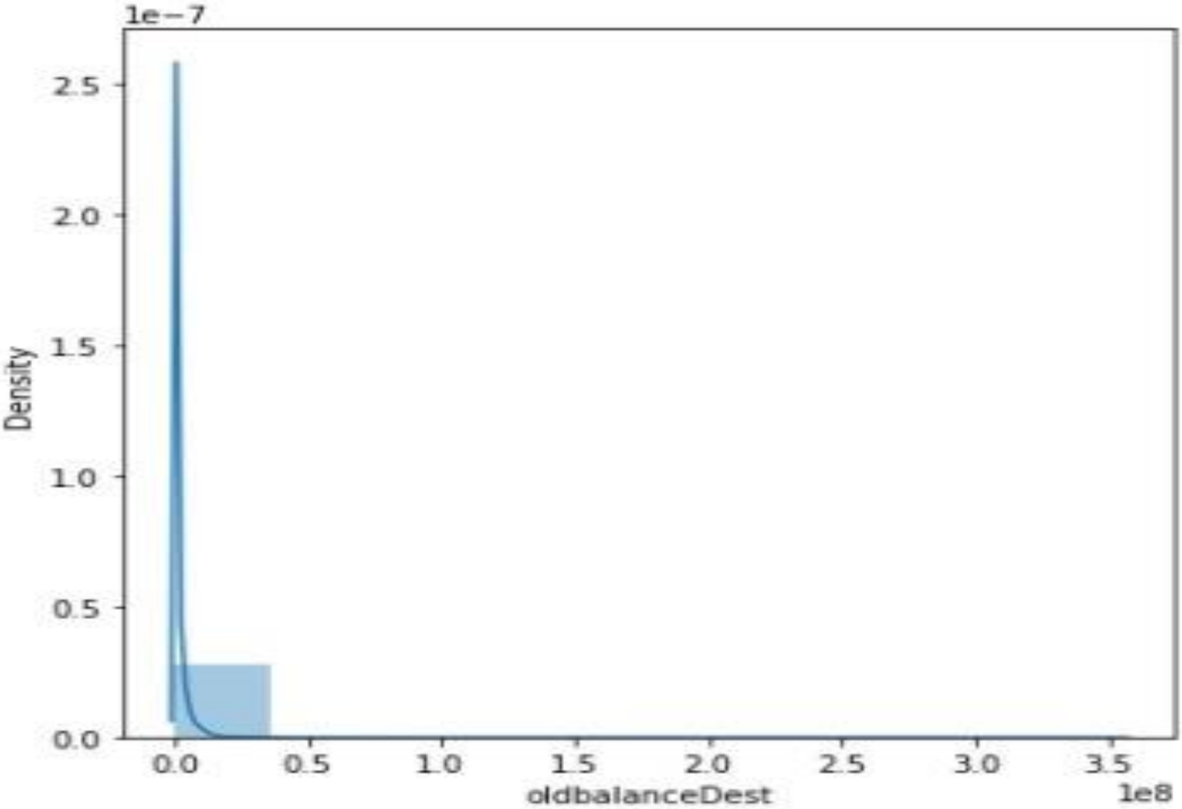
OldBalanceOrg



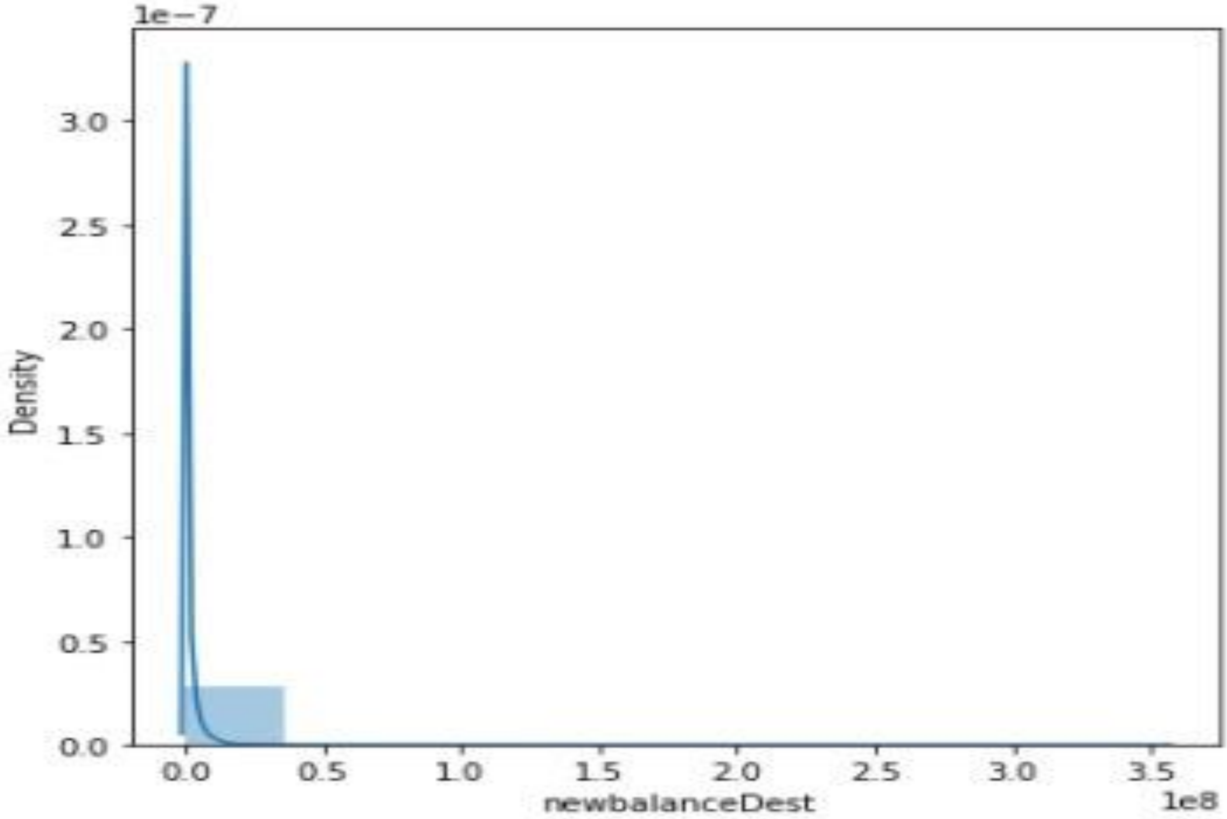
NewBalanceOrig



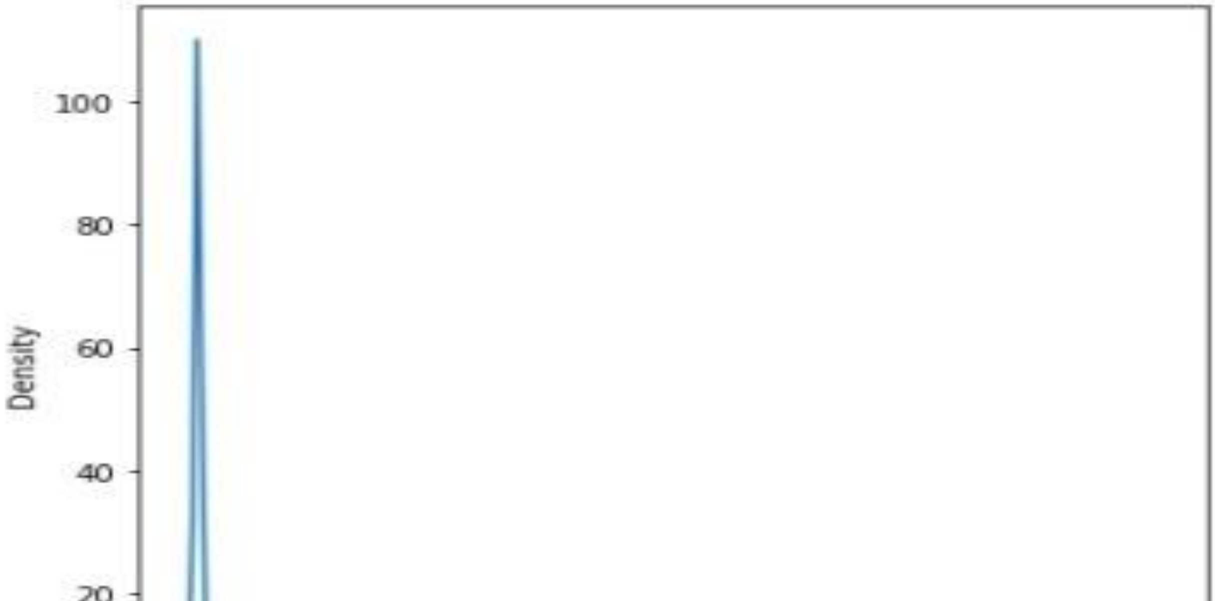
OldBalanceDest



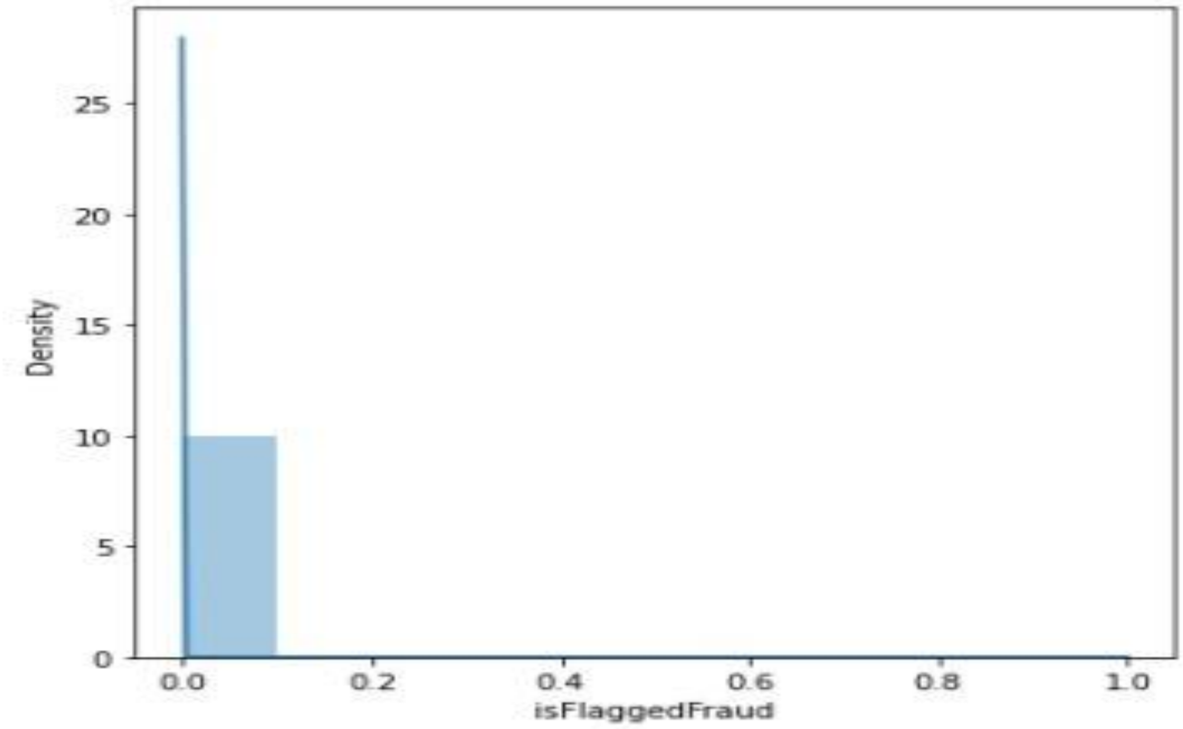
NewBalanceDest



IsFraud



IsFlaggedFraud



Results

- R values lower than 0.20 and near zero indicate no relationship or very weak relationship.
- Weak relationship between 0.20-0.39
- Medium level relationship between 0.40-0.59
- Strong relationship between 0.6-1.00

According to our scatter plot and correlation matrix, there is strong positive correlation between newBalanceOrig-oldBalanceOrg and newBalanceDest-oldBalanceDest. There are Medium level positive correlation between amount- oldBalanceDest and amount-newBalanceDest. Other relationships between attributes has weak positive or no correlation.

According to our attribute charts:

- Most of the transactions occurred between step 0 and 400. After 400 transaction counts has decreased significantly
- Most of the transactions has occurred as a “CASH_OUT, PAYMENT and CASH_IN” type. TRANSFER and DEBIT’s counts are lower than the other 3 type.
- More than 99% of the amount’s instances are lower than 100,000 $. That’s why Amount chart is right-skewed
- Most of the new and old balances of original and destination accounts’ values are lower than 200,000$. That’s why these charts are right skewed too.
- There are 6.3 million instances in our dataset and there are only 8213 fraud instances in our dataset. That’s why there is accumulation of non-fraud labeled data.
- In nameOrig and nameDest, we encoded the unique values to numbers. One unique value occurred at most 1 time in nameOrig chart and at most 6 times in nameDest chart.