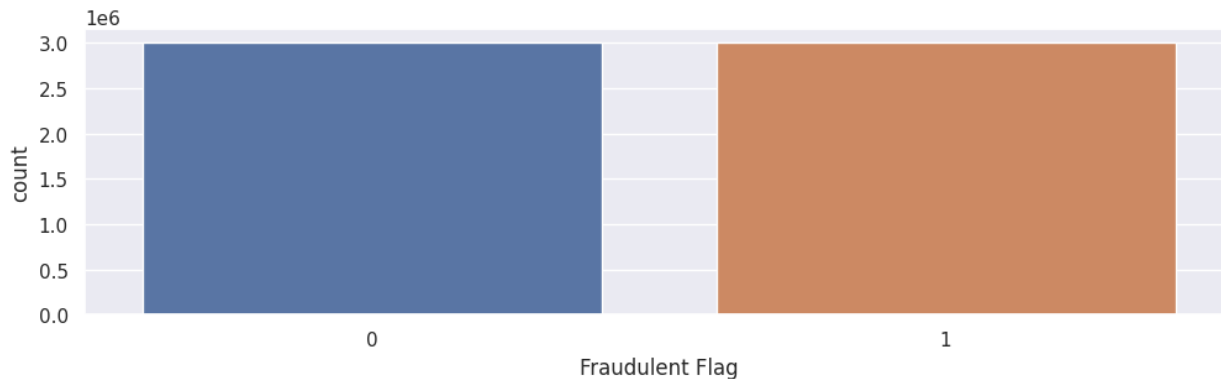


Oluwafemi's Executive Report On DataFest23' Datathon

Assessing Data:

Started my analysis by assessing the dataset provided which contain 6 million rows and 32 columns out of which was 18 categorical column, 13 numerical column and 1 datetime column. Went ahead to investigate the data by checking the data size, duplicates, missing values, data types, summary statistics, and other observation as seen in the code.

I plotted a histogram plot of my target variable which was quite balanced

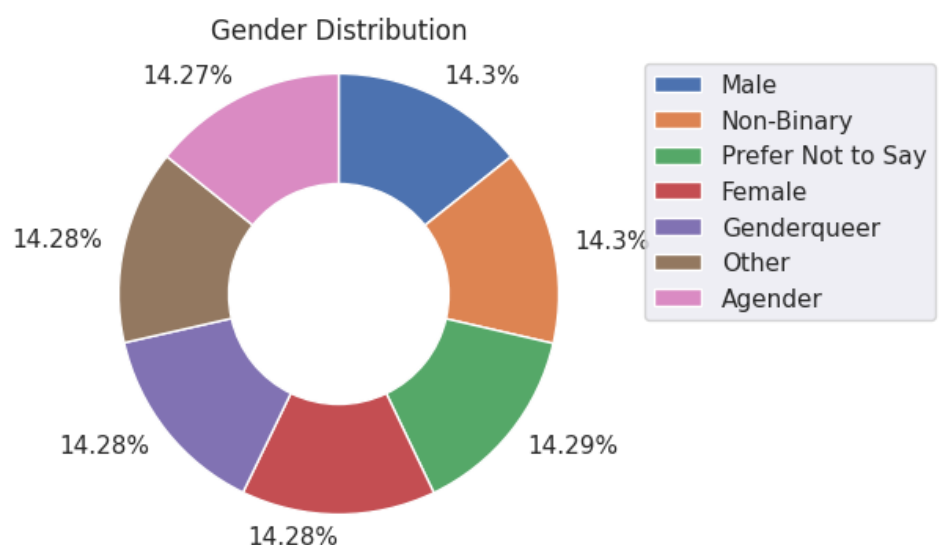


Data Exploration:

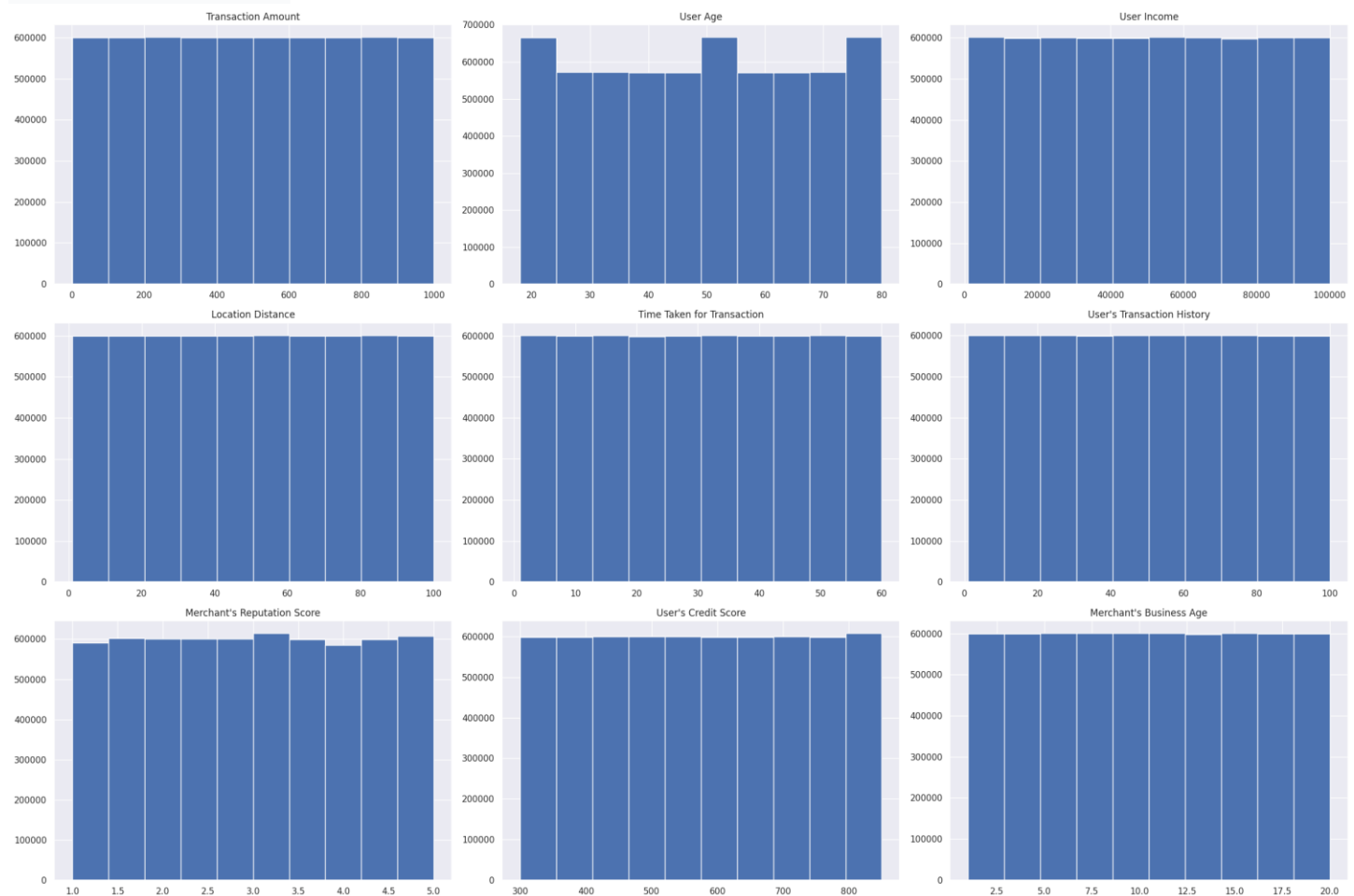
Univariate Analysis

Carried out Univariate Analysis by Checking Frequency for categorical variables and distribution for numerical variables.

Categorical Variable



Numerical Variable:



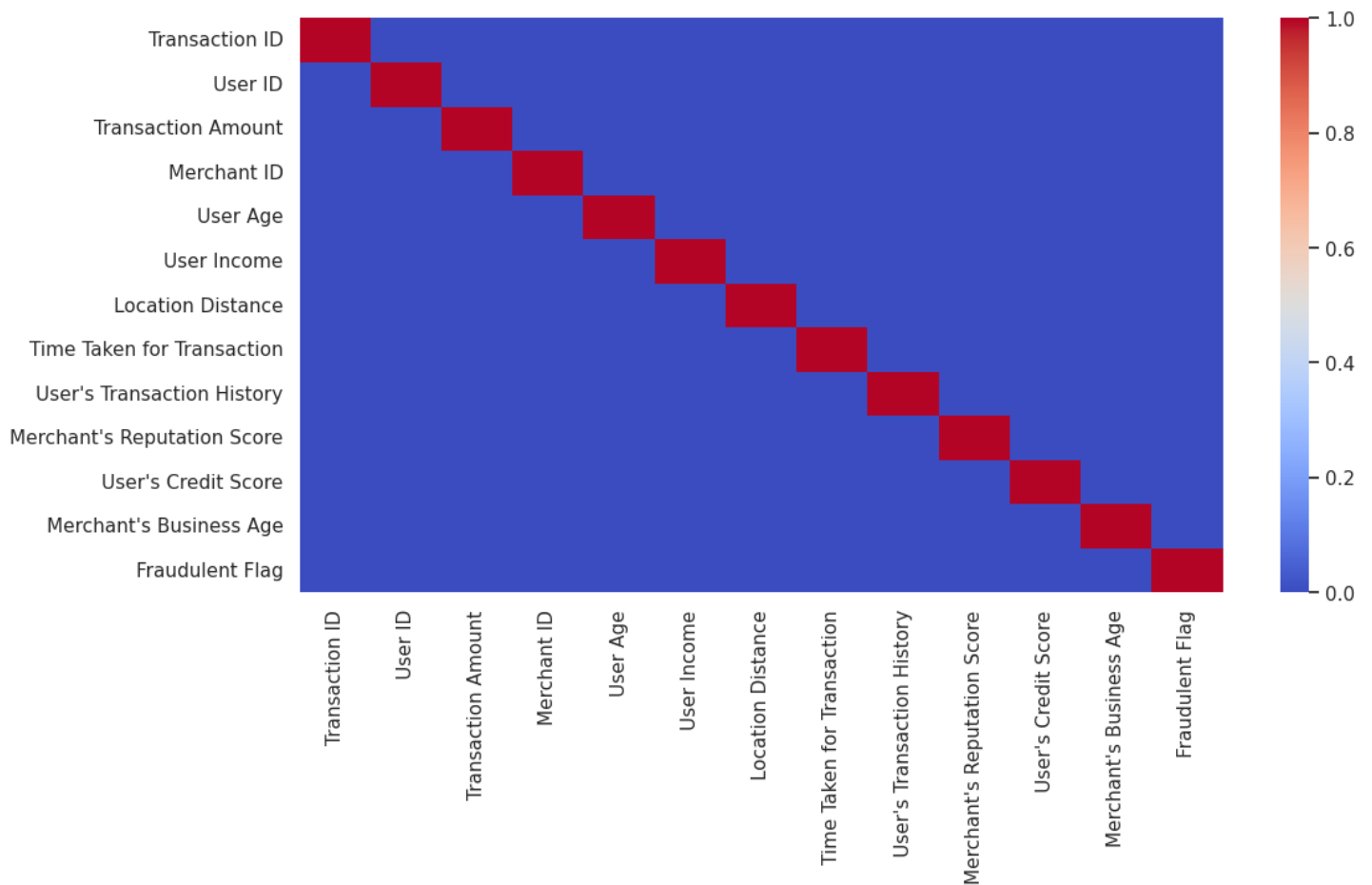
There seem not to be any pattern in the numerical distribution as well.

Bivariate Analysis

I proceeded with Bivariate analysis, where I compared each variable with the target variable (Fraudulent Flag), which didn't provide much insight as there seems to be a uniform distribution for the target

Multivariate Analysis

I proceeded to check the relationship with all categorical variable with fewer cardinality (less than 41) to check for correlation and multicollinearity as seen below.



Feature Engineering:

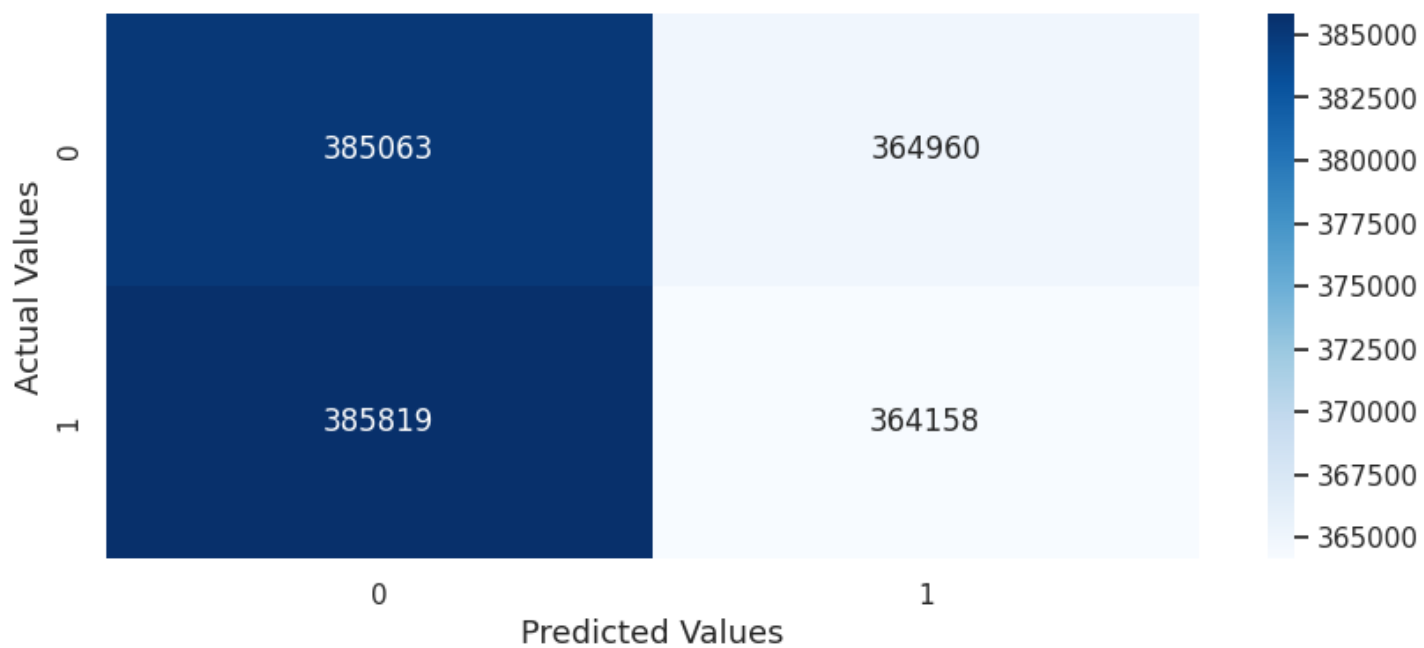
I proceeded to extract new variables from Transaction date and time including, Day of the week, Month of the year, and Day of the month and conducted bivariate analysis to observe any pattern.

Model Training and Evaluation:

Used Random Forest Classifier model and stratified Kfold (folds=4) for data splitting and set some parameters. Performed the prediction with an accuracy of slightly above the baseline accuracy of 0.5

Communicating Result:

A confusion matrix



Feature Importance

