PROJECT PHASE 2 EXPLANATION ON DATA PREPROCESSING

We processed missing data on our dataset. First using the chosen columns , we find out the missing values in each column using the “isnull()” method. All columns and rows in our dataset returned false. In addition, some columns had negative values only. In this case, we opted to ignore since this was the only best option. Choosing to remove the column was not important since that was losing large part of data.

Outliers are extreme values within a dataset. That is, they are either extremely larger or significantly small. In our data preparation steps, we opted to use two ways of finding out the outliers in the dataset columns. We used pandas describe() as method one and box plotting as method two. The pandas describe() presents the mean value of each column of our dataset plus other relevant values such as the maximum value, the minimum value and so on. With these values provided, we could tell, if there is outlier in each column of our dataset. The same case for box plot. The plots clear tell how values of a dataset column have been distributed. Hence, easy to find out an outlier value.

In response to the outliers found, we used the method called “imputation technique” whereby we perform the replacement of outliers using imputations as if they were missing values”.

We also checked the redundant of attributes in our dataset. We found out that 'EVNTIDX', 'PANEL', 'DENTPROX', 'GENDENT\_M18', 'GENDENT\_M18' variables of our dataset had correlation of over 0.95. Thus, we dropped these variables from our selected dataset.

Data normalization was not left behind. This was an important step in preprocessing our dataset. With this step applied we were able to scale down our dataset columns a process called feature scaling. By doing this, we are making our dataset ready for use.