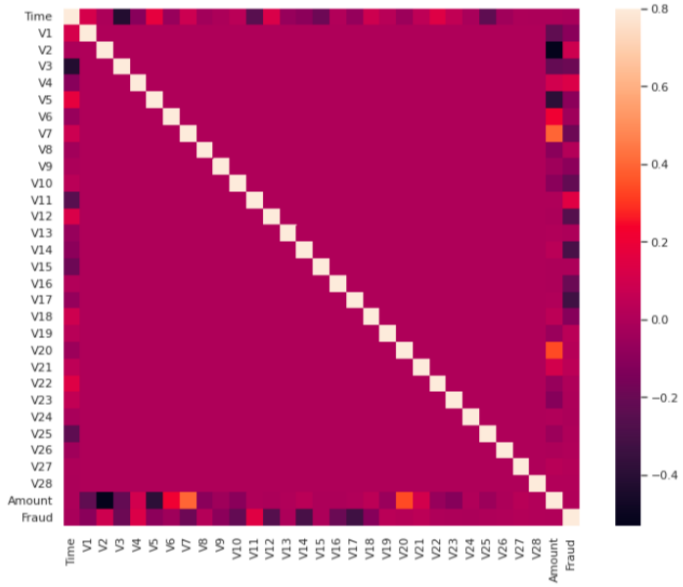
# EDA Report for Credit Card Fraud Detection

| # Time | # V1 | # V2 | # V3 | # V4 | # V26 | # V27 | # V28 | # Amount | # Class |
|---|---|---|---|---|---|---|---|---|---|
| 0 | -1.3598071336738 | -0.07278117330984497 | 2.536346737969144 | 1.378155224274443 | -0.189114843888824 | 0.13355837674038787 | -0.021053053453821 5 | 149.62 | 0 |
| 0 | 1.191857111314846 | 0.2661507120596 3 | 0.166480113353 21 | 0.448154078460911 | 0.125894532368176 | -0.008983099143 22813 | 0.014724169192 4927 | 2.69 | 0 |
| 1 | -1.358354061598 23 | -1.340163074736 09 | 1.773209342631 9 | 0.379779593034328 | -0.139096571514 147 | -0.055352794038 4261 | -0.059751840592 9204 | 378.66 | 0 |
| 1 | -0.966271711572 087 | -0.185226008082 898 | 1.792993339578 72 | -0.863291275036 453 | -0.221928844458 407 | 0.0627228487293 033 | 0.0614576285006 353 | 123.5 | 0 |
| 2 | -1.158233093495 23 | 0.8777367548484 51 | 1.548717846511 | 0.403033933955121 | 0.502292224181569 | 0.219422229513348 | 0.215153147499206 | 69.99 | 0 |

This report analyzes data taken from Kaggle, which contains credit card data from September 2013 European card holders' transactions. Only 0.172% of all transactions that occurred on the two days of the dataset were reported as frauds. However, the dataset only contains numerical input values, which are the result of a PCA transformation. Due to confidentiality reasons, we are unable to access the original features. Twenty-eight columns are listed as V1, V2, V3, all the way up to V28. The only named columns are the transaction amount, the time elapsed from the start of the dataset, and classification of the transaction (fraud or no fraud). The dataset contains 284,807 rows and 31 columns. This report aims to use this dataset to understand the relationship between variables in order to eventually train a ML model which will classify credit card fraud, given the aforementioned variables.

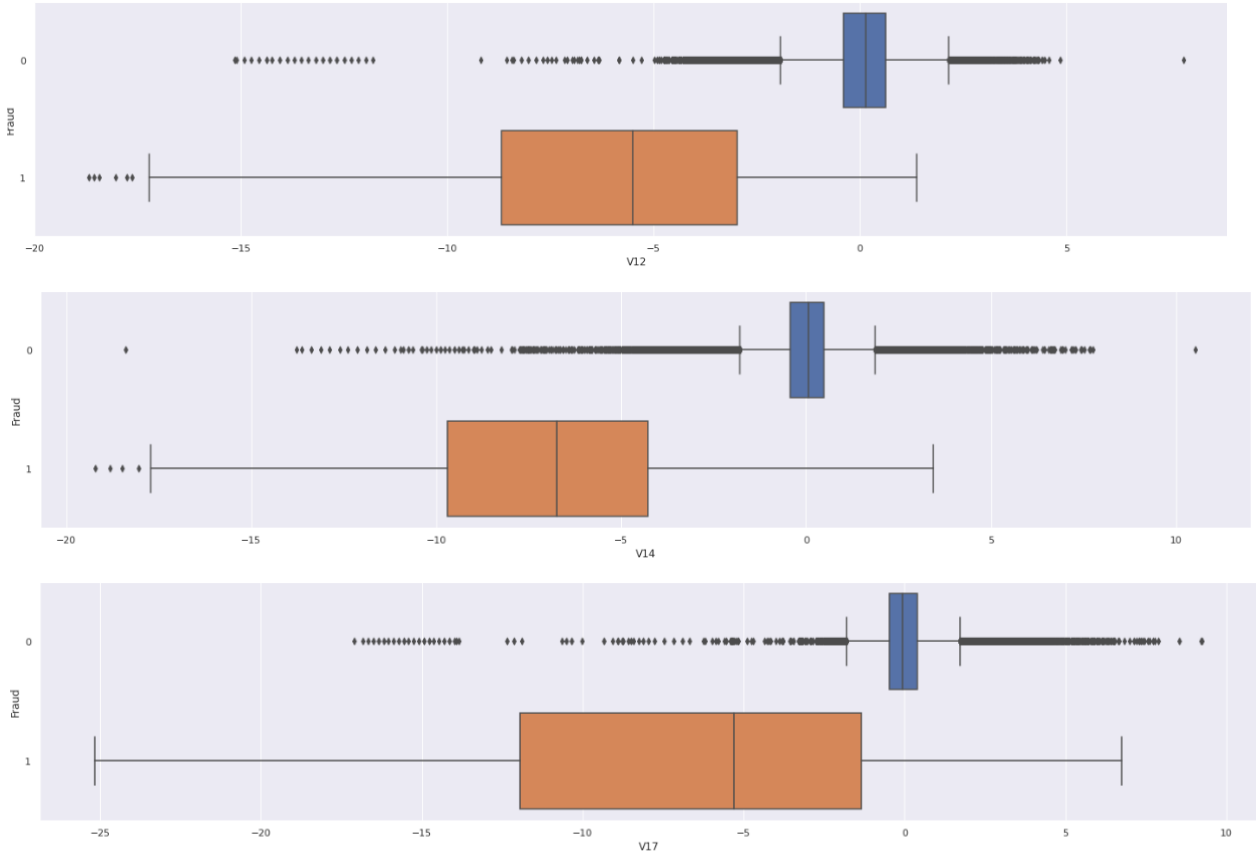



Firstly, it is important to understand the *correlation* among all the columns in the dataset. Correlation is measured by the *correlation coefficient*, a value from 1 to -1 that describes the direction and strength of a 2-variable linear relationship. As the correlation coefficient's absolute value approaches 1, the linear relationship between the 2 variables becomes stronger; by contrast, as the correlation approaches 0, the 2 variables have a weaker (or even nonexistent) linear relationship. The *correlation matrix* above illustrates the correlation coefficients of each pair of variables from the credit card fraud dataset.

Since no information about V1-V28 was disclosed, we hypothesized that they would have a correlation with each other. However, according to the correlation matrix, V1-V28 actually

have no correlation with each other; rather, they only have an apparent linear relationship with time, amount, and/or fraud.

Additionally, the column headers/variables were expected to have a more distinct relationship with fraud; however, only V12, V14, and V17 had strong negative correlations with fraud, far fewer than we were expecting.




These box plots further prove that higher values of V12, V14, and V17 were directly related to less fraudulent transactions, since the blue boxplot (which signified no fraud) generally had larger values.

Ultimately, now that relationships between variables have been determined and explored, the next steps are creating a machine learning algorithm that is trained on this dataset and predicts credit card fraud.