**Methodology**

**Dataset Collection**: The two datasets we found are:

<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?datasetId=310&searchQuery=eda>

<https://www.datacamp.com/workspace/datasets/dataset-python-credit-card-fraud>

* **Data Simulation:** Due to the unavailability of datasets, data simulation can be used to simulate fraudulent and non-fraudulent transactions with basic features for predictive modeling. Reference: <https://fraud-detection-handbook.github.io/fraud-detection-handbook/Chapter_3_GettingStarted/SimulatedDataset.html>
* **Dataset used**: For the idea submission phase, we are analyzing the widely used kaggle dataset named Credit Card Fraud Detection. The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

**Building Hypothesis Set:**

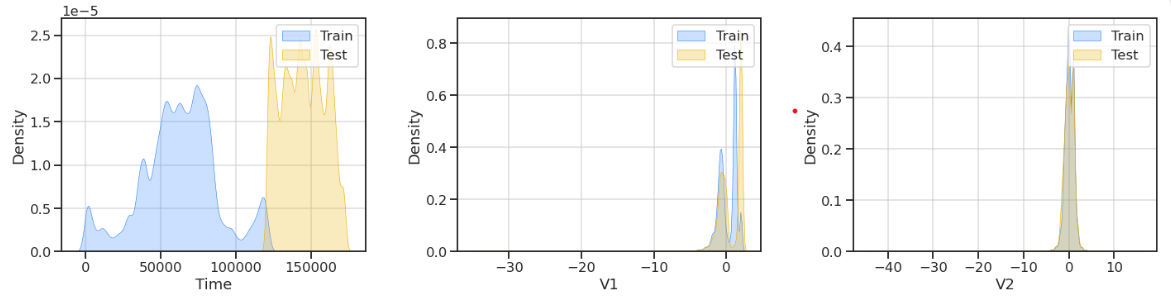
Hypothesis set for Credit Card Fraud Detection-

1. **Online transactions**: Card-not-present fraud is more likely to occur when making purchases online, as the fraudster can use stolen credit card information without the need to present a physical card.
2. **High-value transactions**: Transactions with high purchase amounts or high-value items are more likely to be targeted by fraudsters.
3. **Unusual transactions**: Transactions made in unusual locations or outside of the cardholder's typical purchasing patterns may be flagged as suspicious and targeted by fraudsters.
4. **Travel**: Transactions made in foreign countries or regions where the cardholder does not normally travel to are more likely to be targeted by fraudsters.
5. **Cash withdrawals**: Criminals may use stolen credit card information to withdraw cash from ATMs or make cash advances, which are more difficult to trace.
6. **E-commerce sites with weak security measures**: Fraudsters can target e-commerce sites with weak security measures that make it easier to steal credit card information.

**Exploratory Data Analysis of the kaggle dataset ‘Credit Card Fraud Detection’:**

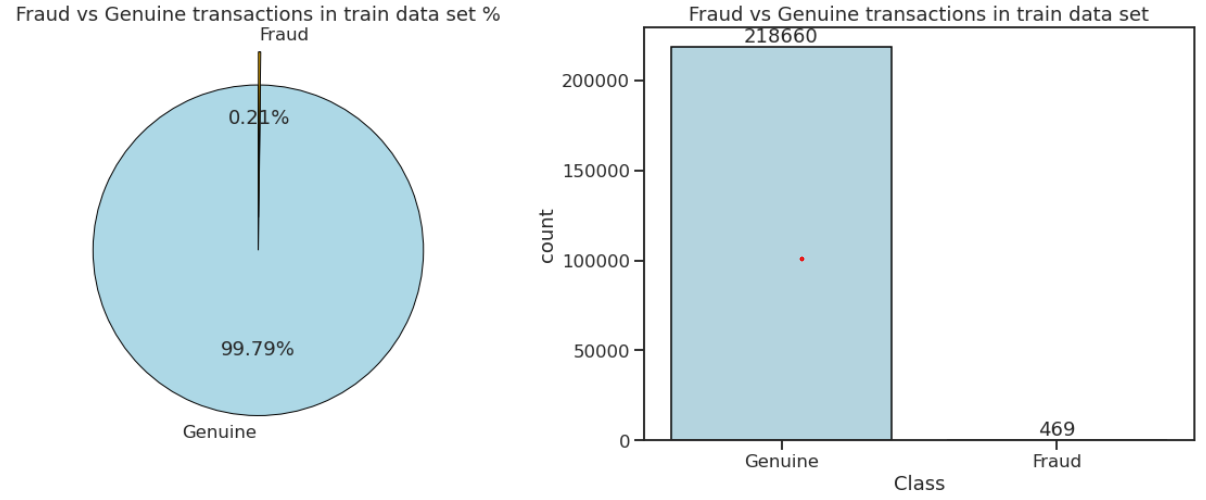
Due to confidentiality issues, the original features have not been provided. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

* **Feature Distribution:**



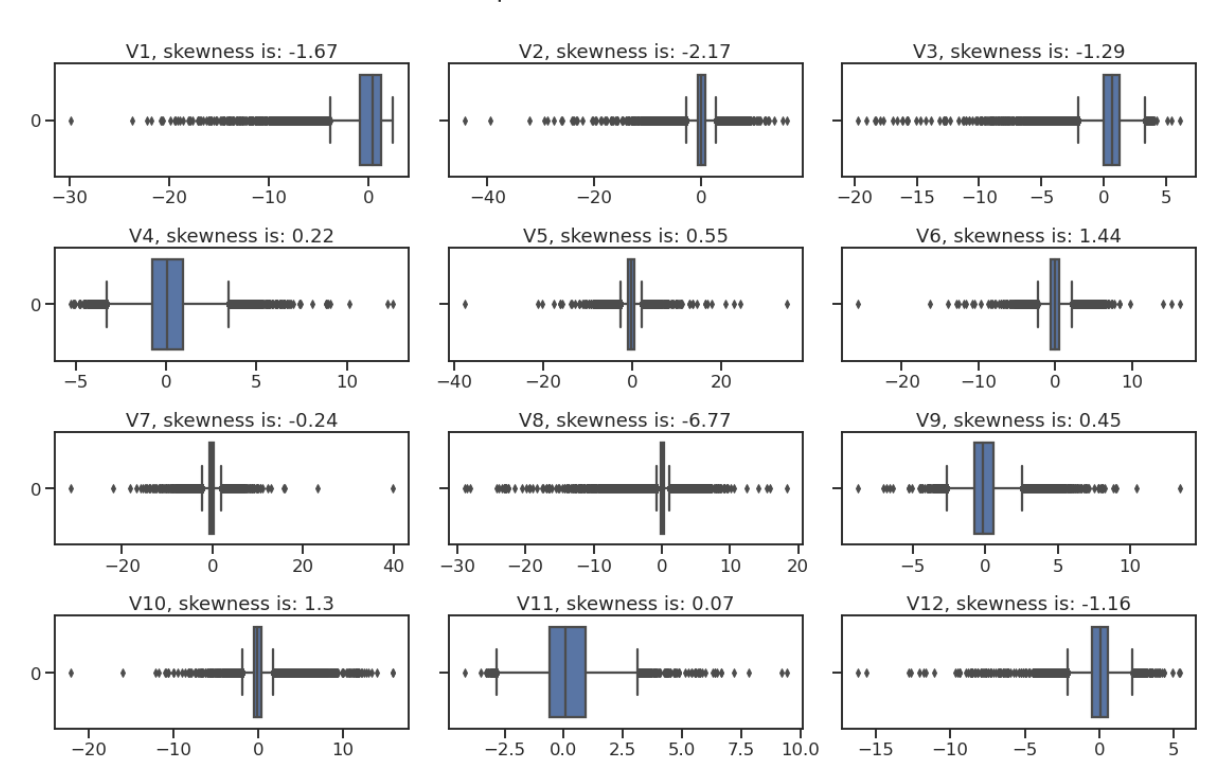
The feature distribution of all the features was analyzed with the plots shown. We can see that the distributions of 'Time' are very different for train and test set, therefore, the time feature can be dropped.

* **Data Imbalance Check:**



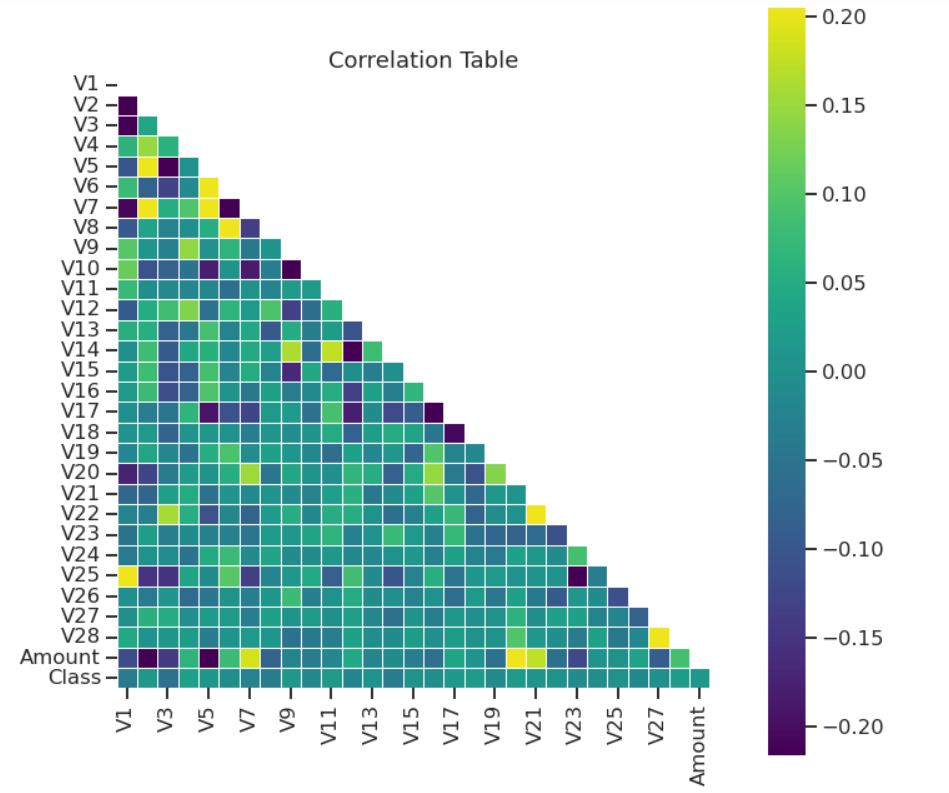
There are 99,8% of Genuine transactions (218,660) and only 0,214% (469) of fraud transactions! This means that a blind guess (bet on Genuine) would give us accuracy of 99,8%. Accuracy score as a metric cannot be used with imbalanced datasets - it will be usually very high and misleading. Instead AUC and F1 Score are used.

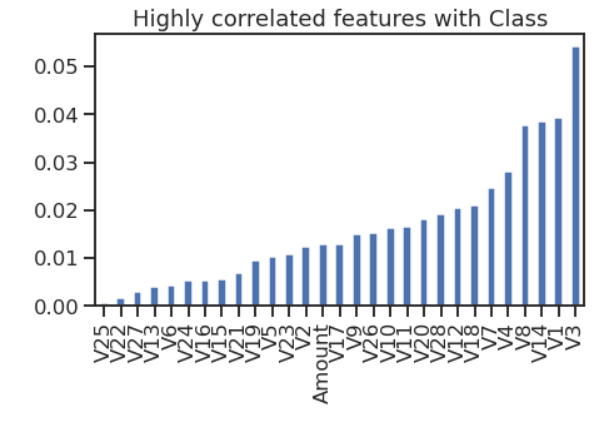
* **Outlier Detection:**



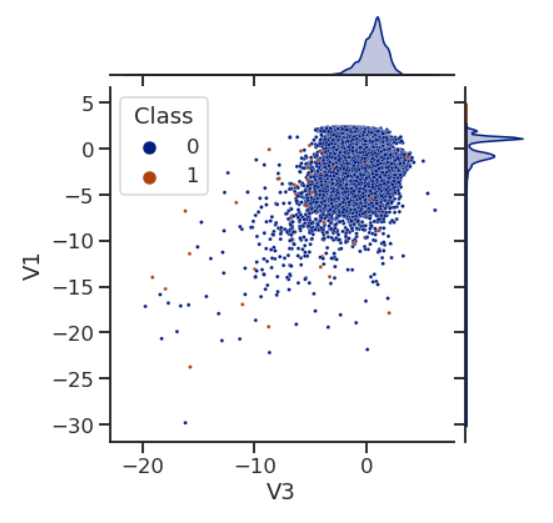
We have a significant problems with outliers: huge outliers; highly skewed data; a lot of outliers.

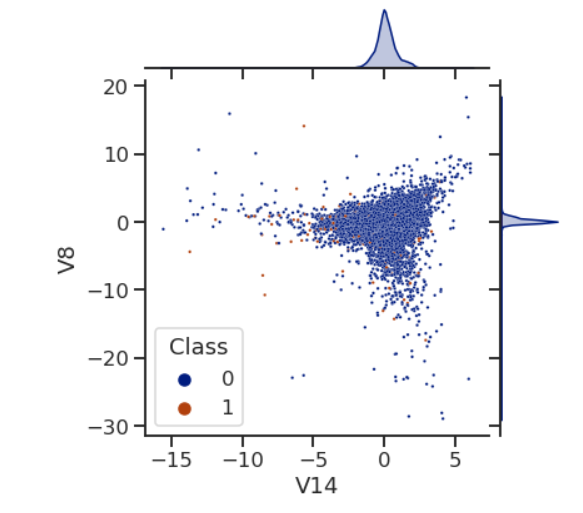
* **Correlation:**





* **Jointplots**





**Data Sampling (handling imbalanced dataset):**

* SMOTE
* SMOTE Tomek
* RUS

**Feature Selection:**

* EDA
* Genetic Algorithm(GA)

**Feature Extraction:**

* PCA
* CAE

**Model Training:**

Ensemble models-

* Random Forest
* XGBoost
* CatBoost
* LightGBM

**Model Testing:**

* AUC
* F1 Score