Data Wrangling

Overview

This section describes the various data cleaning and data wrangling methods applied to the “Credit Card Fraud Detection” dataset.

Summary Files

I took a creditcards.csv file where Class is the target variable in there already prelabeled as 1 and 0 to easier classify as Fraud or Non-Fraud transaction and do not have to find them first. Dataset is ready for training a model but very imbalanced to make accurate classifications.
The dataset that is used for credit card fraud detection is derived from the Kaggle https://www.kaggle.com/mlg-ulb/creditcardfraud The dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

Duplicates and Missing Values

For the analysis it was necessary to check for duplicates.
After analysis, no missing values and duplicates were found and no columns to drop.

Outliers and handling them

After making plots of Amount Distribution of an individual V, it is easy to see outliers. (unusual pattern of data that is really different or far from normal). What I did is I annotated a point which is far from normal visually. So nobody will miss it and will notice. But are those really outliers? To decide outliers or not, I can perform an IQR test.
This is my approach at this point.