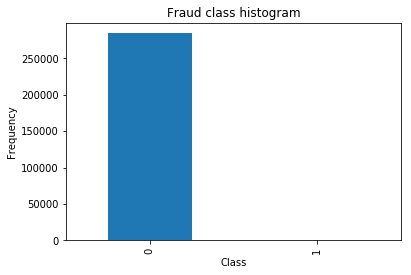
About the Data...from Kaggle

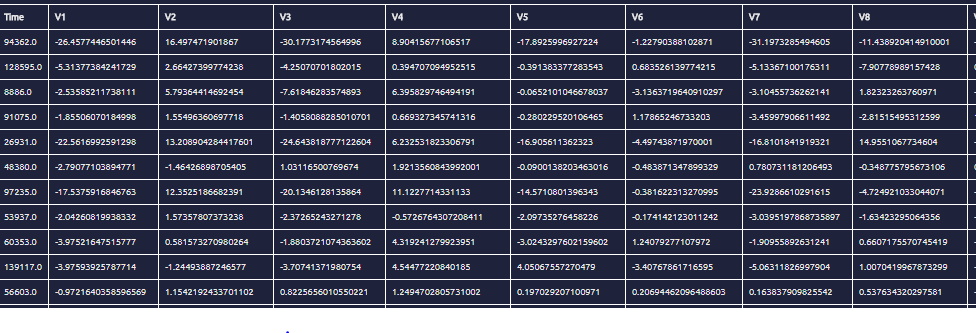
**Is Labeled (**outlier is 1... normal is 0)

**Is Skewed (**toward normal)



Only .17% of the an outlier (Class = 1) so the data is overwhelmingly normal (Class = 0)

**and where’s the credit card information?**



The data from Kaggle was run through a PCA principal component analysis model in order preprocess the data using complex vector mathematics to feature reduction (fewer independent variables to account for)… It also has your basic dehumanizing effect.

**1. Logistic Regression with undersampling**:

Undersampling is reducing the size of the dominate dataset so that it matches the subordinate dataset and then training your model on it.

Positive: emphasizes what makes up the subordinate dataset to the model

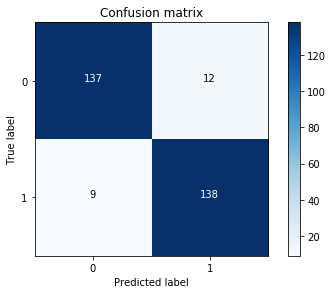
Negative: loses emphasis on what makes up the dominate dataset

* Accuracy = (TP+TN)/total
* Precision = TP/(TP+FP)
* Recall = TP/(TP+FN)

Number transactions train undersample dataset: 688  
Number transactions test undersample dataset: 296  
Total number of undersample transactions: 984 (442 - 0 and 442 - 1)

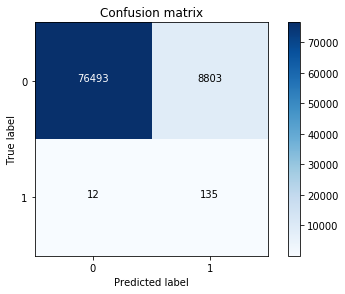
Train over undersampled training and Predict over undersampled test:

Recall: **0.938775510204**

****

Train over undersampled training and Predict over skewed test:

Recall: **0.918367346939**



**Logistic Regression with undersampling**: (continued)

Train over skewed training and Predict over skewed test:

Recall: 0.619047619048