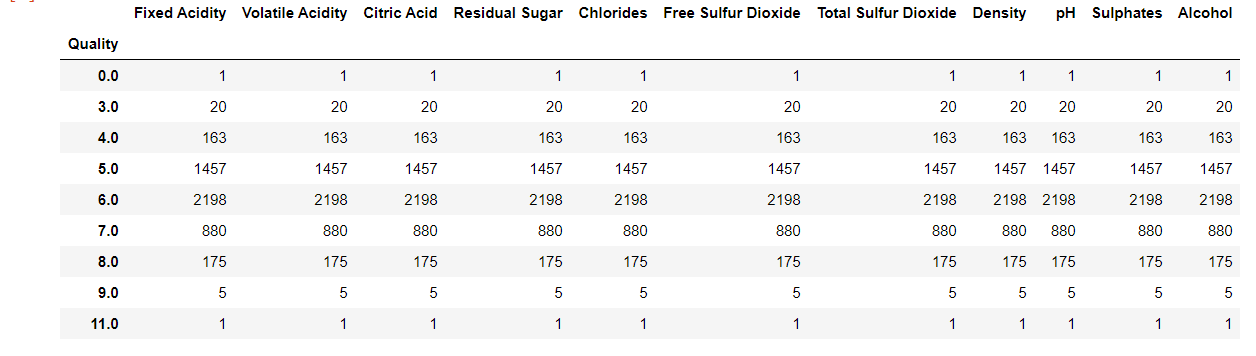
Victor Castellanos

Ian Schenck

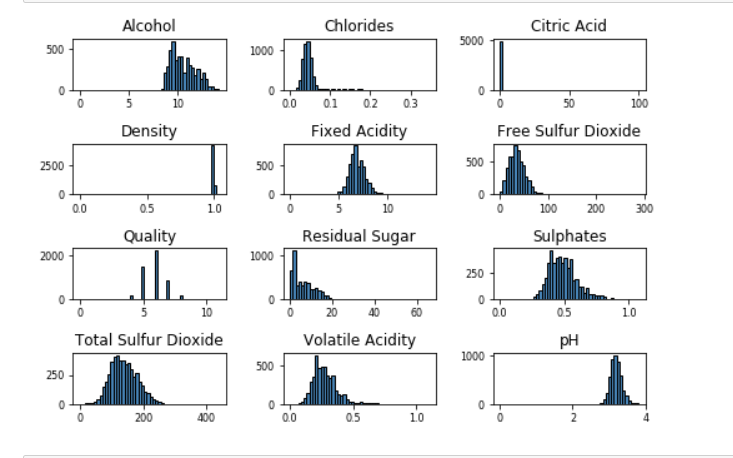
CECS 550 Project 3 - kNN

**Part 1 – Data Wrangling**

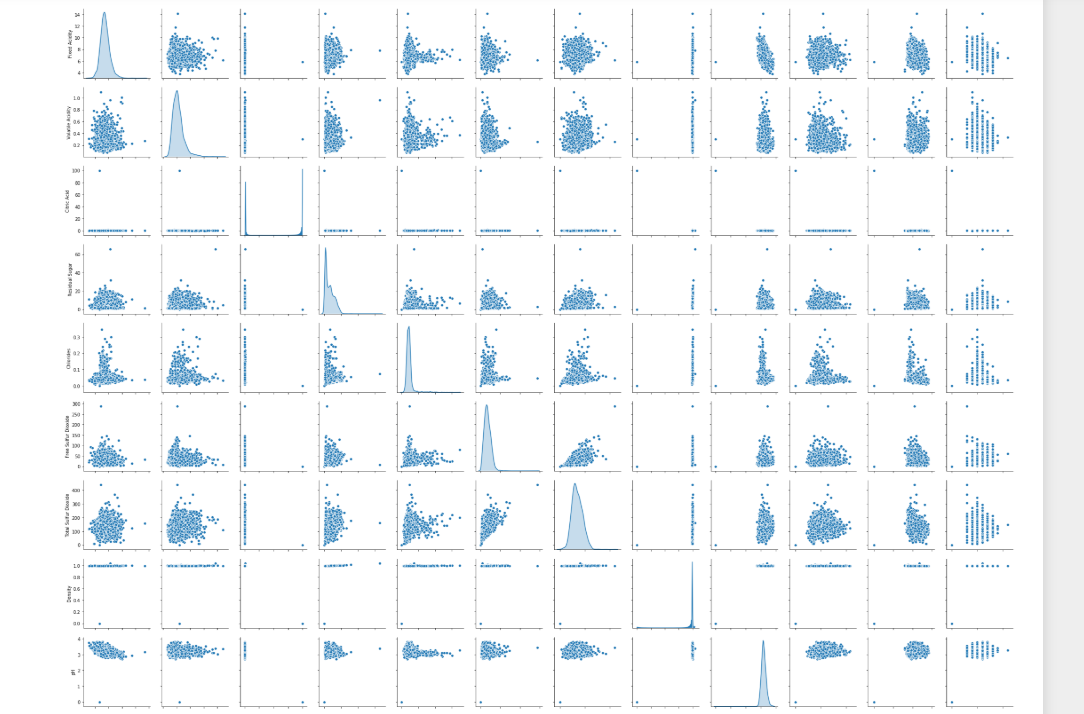
Our first task was to preprocess the data, The data shows that there are 9 classes. There is one data point for quality 0 and one data point at quality 11. There is very few high quality 9’s and very few quality 3’s.

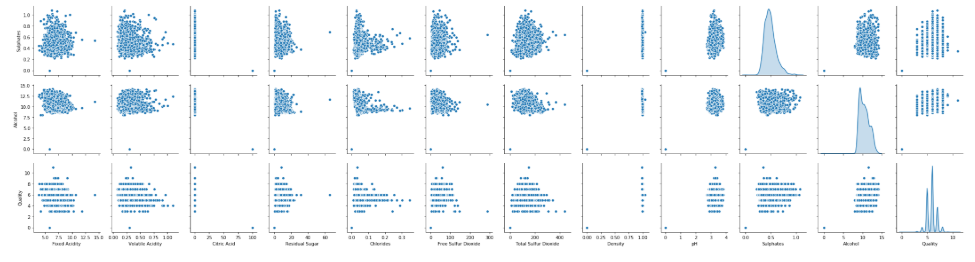


Looking at the histogram of each of the features we can see that most of the data for each of the features are around the same region. There are some exceptions, for example free sulfur dioxide has most values from 0 to 100 but has some values up to 300. Total Sulfur Dioxide has most values ranging from 0 to a little over 200 and few values going to about 400. This showed us that we need to investigate for possible outliers.



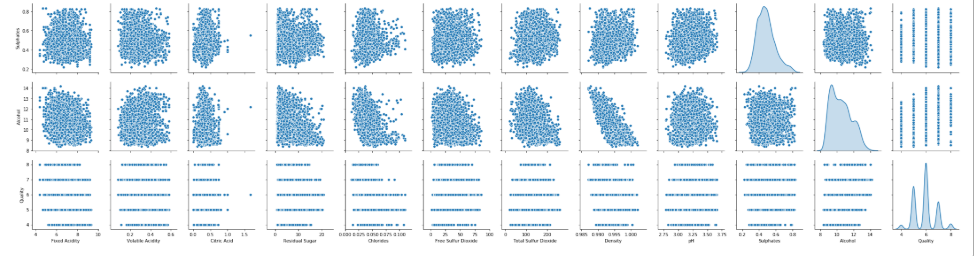
We can also look at a 2d scatter plot comparing pairs of features. The graph going left to right look at Fixed Acidity, Volatile Acidity,Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density,pH, Sulphates, Alcohol , Quality. Then top to bottom in the exact same order. Most of the graphs have their features concentrated in one area, but it is easy to see that there are points that are outliers.



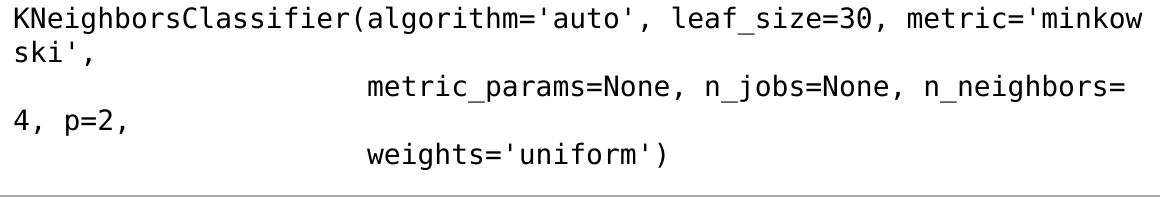


We opted for taking out outliers by removing data points that were larger than 3 standard deviations. This lowered are sample size from 4901 with 9 classes to a sample size of 4558 with 5 classes. We had tried cleaning the data by using the Interquartile rule but found that by doing this we would be cutting out close to a quarter of the data, which seemed excessive.

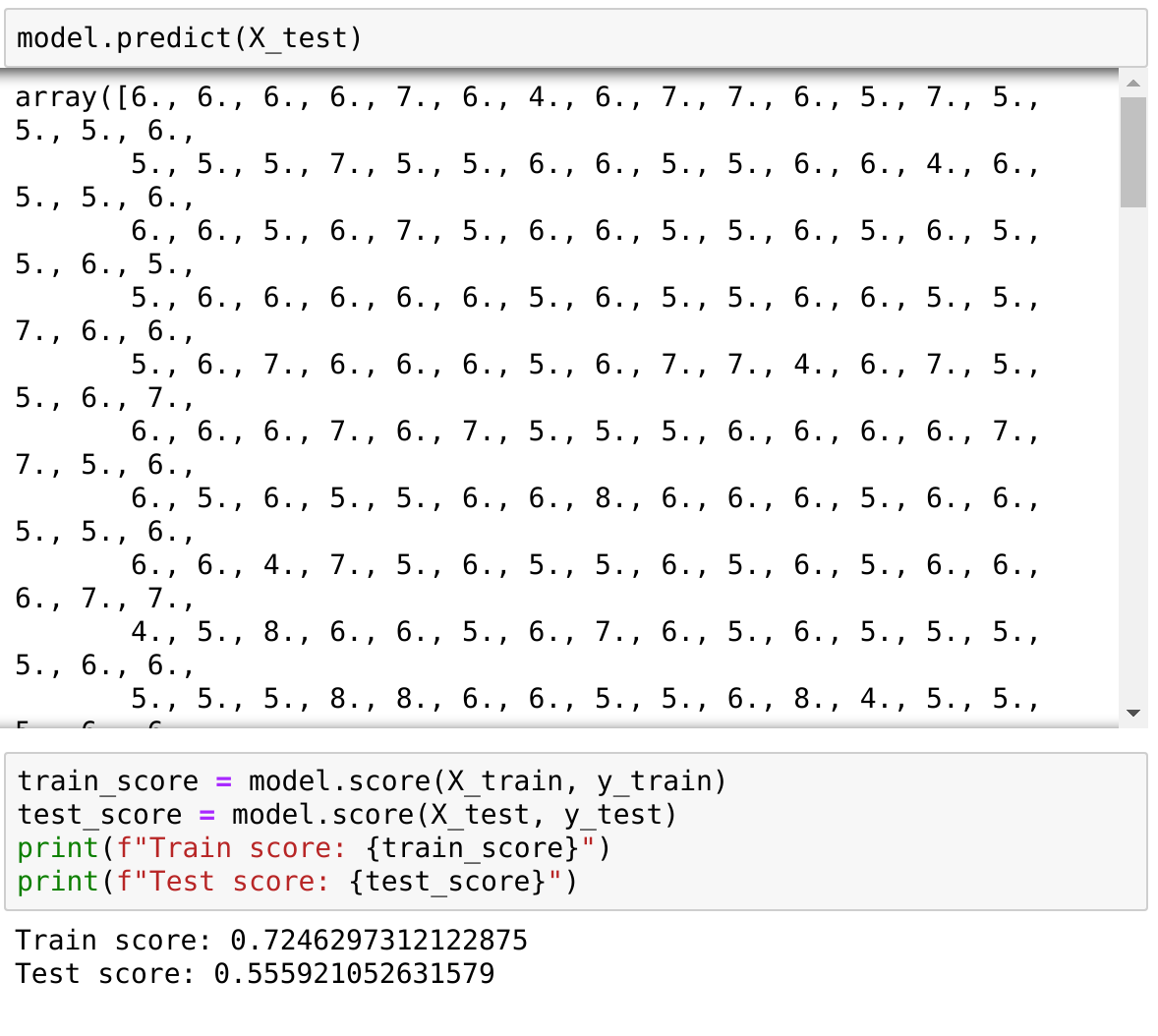


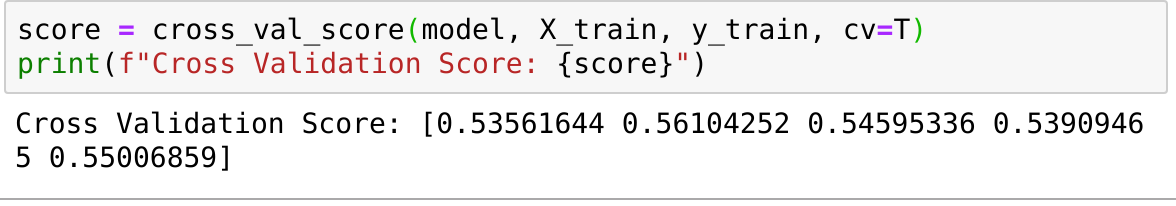


**Part 2 – Building and training the kNN model**

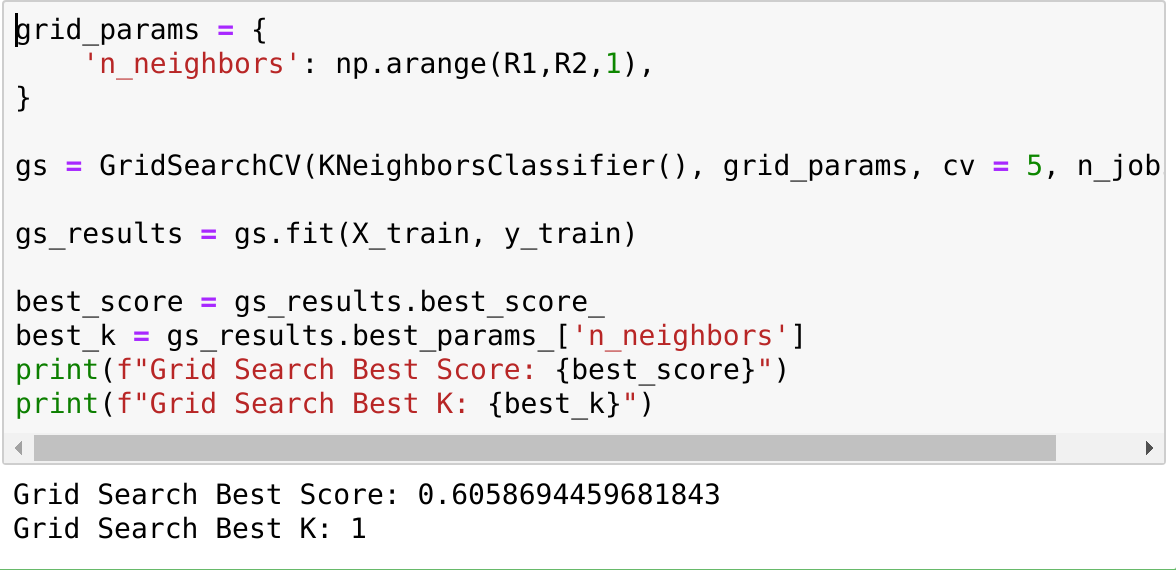
****

**Part 3 – Testing kNN Model**

****

**Part 4 – Cross Validation**

**Part 5 – Optimizing n-neighbors paramater**

****