Project Update 1

As of 11/17/2020, the project is still on schedule.

Dataset was explored and modified to reduce outlier data. After checking the data entries, I found that under column “residual sugar” the value seems abnormal. I did a box plot and found that the average value of all data is 6.39, the max is 65.9 and the min is 0.6. After further investigation, I found that about 25% of data is smaller than 1.7 and about 75% of data is under 9.9. I applied the Z-Score method which eliminates outlier. After doing some research, it is recommended that z-score greater than 3 or less than -3 should be considered as outliers. After removing outliers, the data size gets reduced to 4487 entries which are about 92% of the original dataset.

The first classification algorithm was also implemented--the linear regression algorithm. Dataset is randomly divided into a training set (90% of all data entries) and validation set(10% of all data entries). After running a degree 1 polynomial regression, the accuracy is about 0.5040816326530613.

Degree 2 and degree 3 polynomial regression will be evaluated later on.

At this point, I am planning on making some changes to my plan. Explore K-means clustering will be postponed for 1 week, and it will be done by 11/24.