Jing Li (S5374553)

A report about Linear Regression Modeling on Red Wine Quality

In this report, I will analyze the wine quality dataset to predict the quality of the wine. I will explore the dataset and select two predictor variables and build a linear regression model using them to predict the wine quality.

Data Exploration

I started by importing the dataset using pandas and plotted histograms of all variables except quality to visualize the distribution of the data. We observed that the variables have different scales and ranges. Next, I checked the unique values in the quality column and found that there are 6 unique values (3, 4, 5, 6, 7, 8).

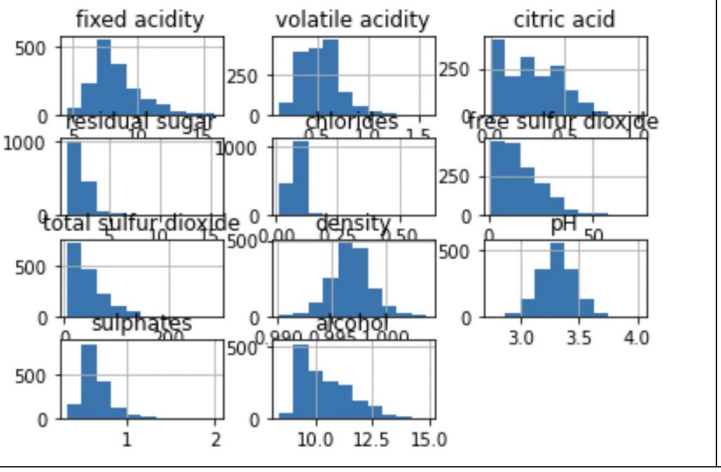

Figure 1. Distribution of the data

Predictor Selection

To predict the quality of wine, I selected two variables, fixed acidity, and pH. I plotted the correlation of each variable with quality, and we found that fixed acidity has a negative correlation with quality, whereas pH has a positive correlation with quality.

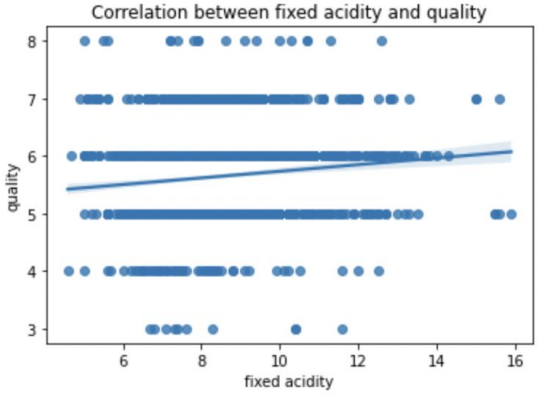

Figure 2. Correlation between fixed acidity and quality

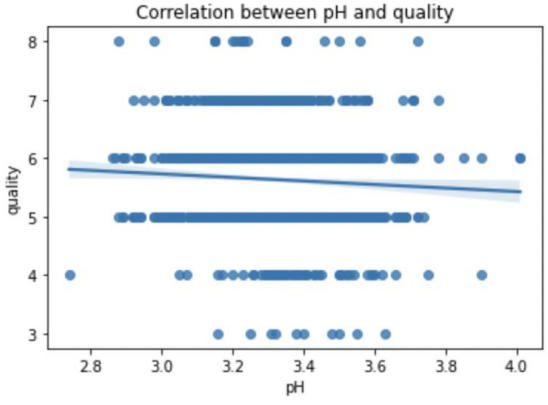

Figure 3. Correlation between pH and quality

Linear regression experiments

In this part, I split the dataset into train and test using the train_test_split() function. I built two linear regression models, one with fixed acidity as a predictor and another with pH as a predictor. I trained the models using the training set and saved them for later use. I evaluated the performance of the models using mean squared error on the test set.

The mean squared error for the fixed acidity model was 0.69, and for the pH, the model was 0.70. This indicates that the fixed acidity model performed slightly better than the pH model.