CS 4375

ASSIGNMENT 1

Names of students in your group: Kylie Quinney (krq210000)

Number of free late days used: 0  
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

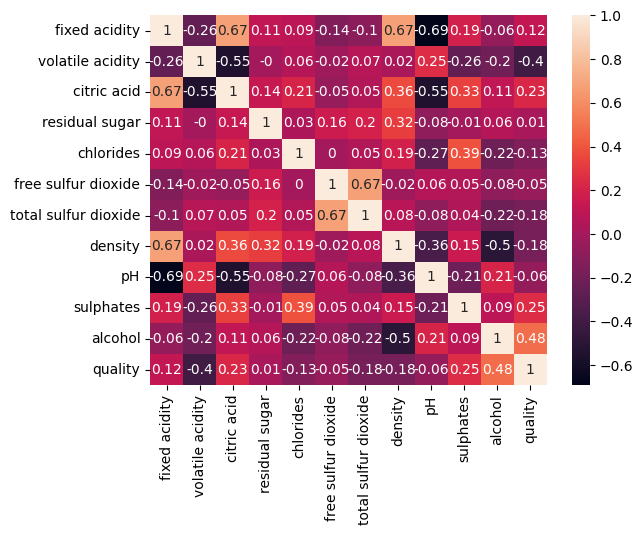
Please list clearly all the sources/references that you have used in this assignment.

**Overview:**

For this assignment, I chose data about red wine with the intention to design a model that will predict the quality of redwine based on chosen features of the wine.

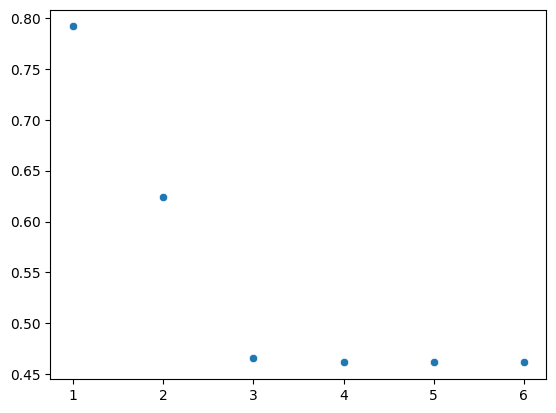
**Plots:**

**Heatmap of Correlation Matrix**

****

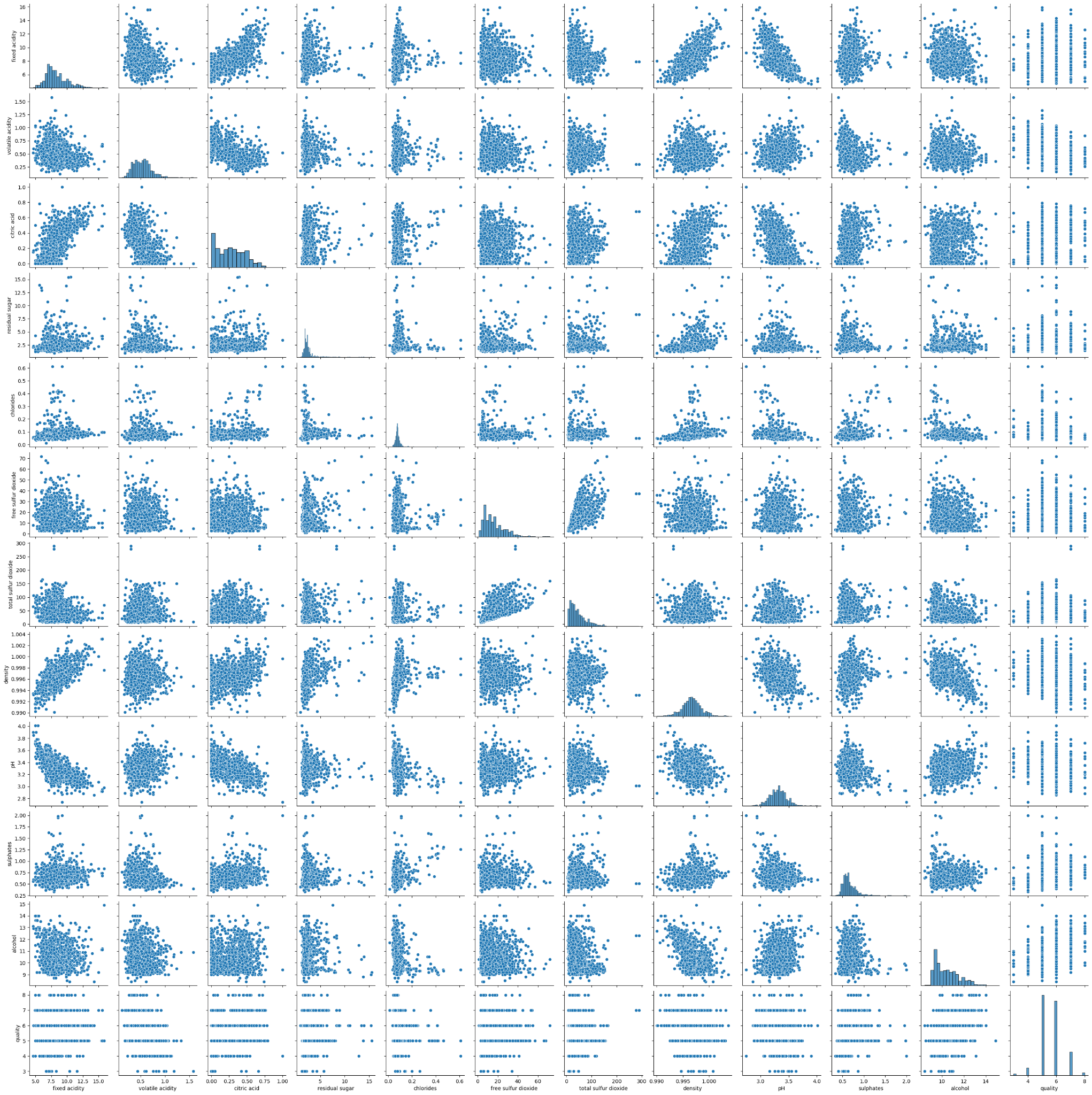
I used the heatmap correlation matrix to help determine wich features would be most useful in our model. I did this by choosing the features with the highest absolute value correlation with the target variable (quality).

**Scatterplot of MSE v Trials #**

****

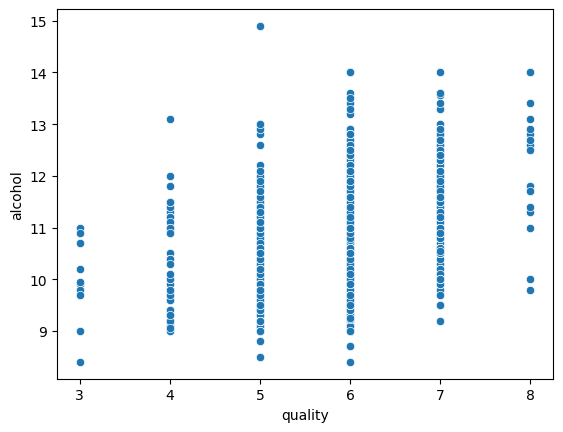
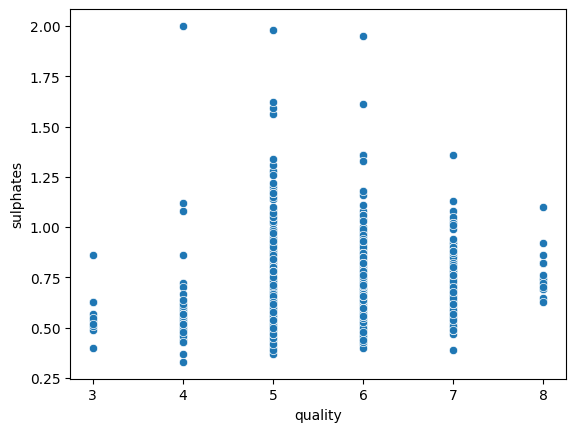
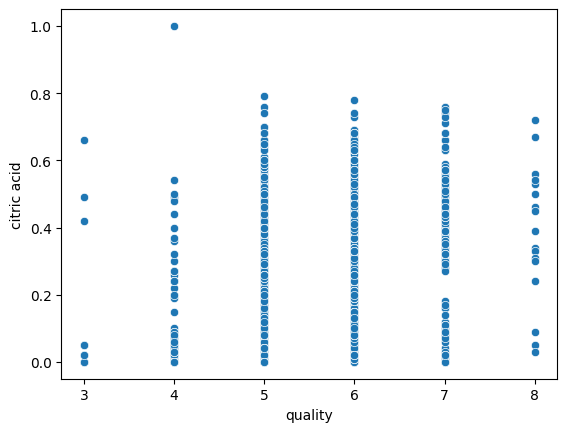
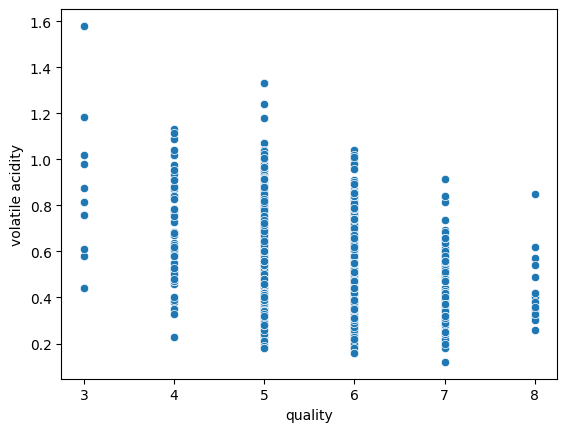
This plot shows the MSE of training predictions v. the trial #, I am using this to demonstrate how my model improved as I tinkered with parameters and features.

**Pairplot**

****

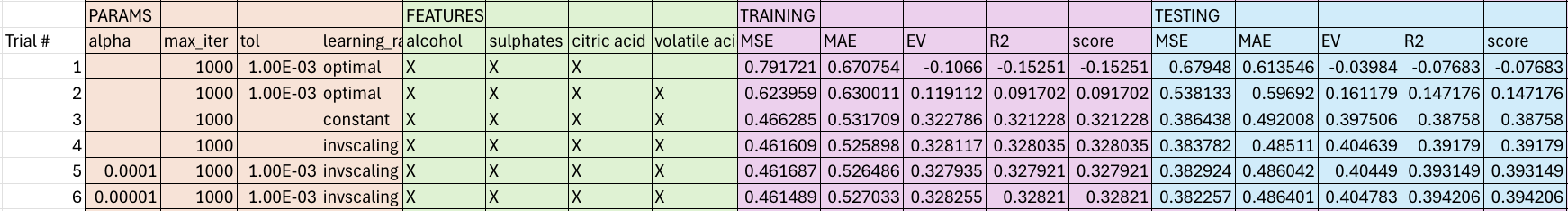
The pairplot above shows how every feature is correlated to one another.

**Scatterplots of Chosen Features v Quality (target variable):**



Here is a closer look at the correlation between the chosen features and the target variable. You can see in the plots above that the relationship between the features and the target are not very linear. Although it is clear that the volatile acidity has an inverse relationship with quality, which is confirmed in the heatmap correlation matrix.

**Trial Data:**



For my result metrics, I used the avg of three runs for each trial, the result for each run can be found in the additional excel file that I submitted. I did this because I noticed that each time I ran the code I would get a slightly different result. I have since learned that I can use the parameter “random\_state” in my regression model to get the same result for each run, however, I did not know that at the time, so I will just have to stick with the average of three runs for this assignment.

**Answer this question: Are you satisfied that you have found the best**

**solution? Explain. How can you check?**

I am satisfied that I have found *almost* the best solution. I have come to this conclusion because as I have continued to tinker with my features and hyperparameters, I have hit a point where my R2 and MSE are no longer improving my any significant amount (as seen in the scatterplot labeled “Scatterplot of MSE v Trials #” above). To further test if I have reached the best solution, I found the OLS solution to use as a comparison. The OLS solution gave an R2 of 0.343, compared to my best R2 result of 0.328 for training and 0.394 for testing, I would say that my model is close to the best it can be.