Introduction

The dataset I have analyzed includes quantitative information about wine. Specifically, various attributes related to Portuguese "Vinho Verde" red wine, and the associated quality rating (on a scale of 0-10). This quality rating is precisely what I will be predicting.

The dataset is located within University of California Irvine's Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/wine+quality), and is also hosted on Kaggle (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). Please refer to the appendix section for a listing of all the features associated with this dataset.

Exploratory Analysis

My first insight into this project was performing some simple exploratory data analysis about the target attribute (quality). As the below tables indicate, the average quality rating ranges between 5 to 6. No wine rated lower than a 3, or higher than an 8.

Initial Metrics of Target Attribute

Statistic	Value
Mean	5.67
Standard Deviation	0.81
Maximum value	8
Minimum value	3

Percentage of Target Attribute Values

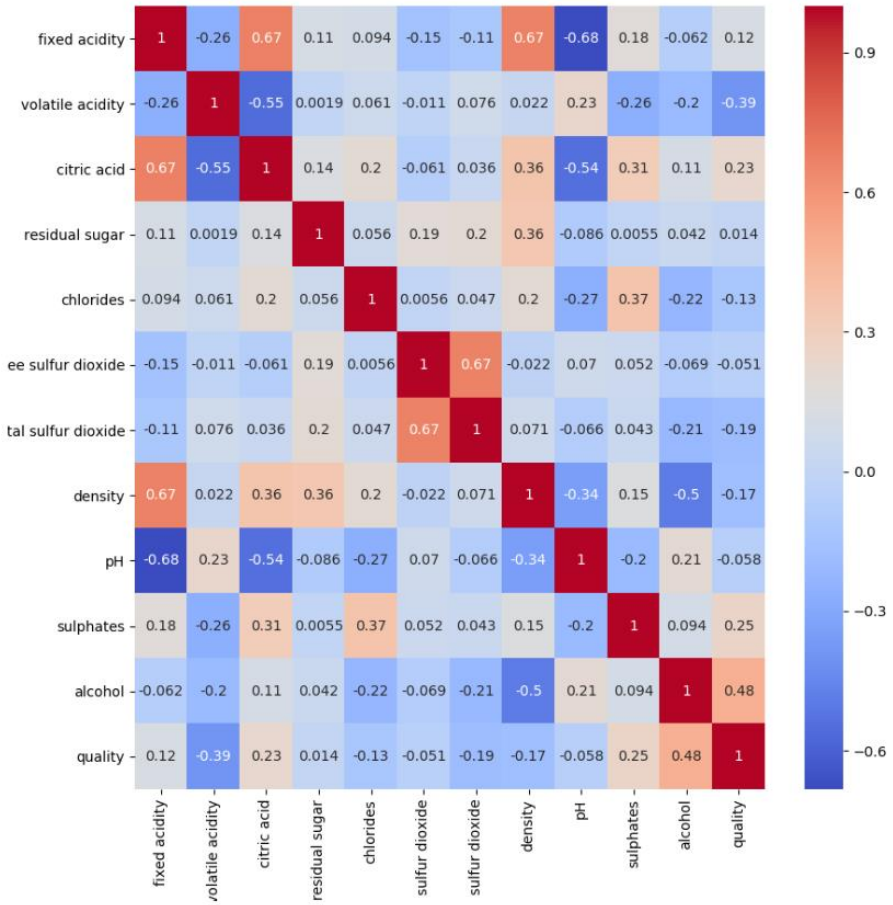
Quality Value	Percentage
3	0.63%
4	3.31%
5	42.59%
6	39.90%
7	12.45%
8	1.13%

Algorithm Selection

Linear regression was the algorithm of choice for predicting the quality values of the wine. This was due to a few reasons. First, all of the features associated with the dataset are numeric, including the target feature, quality. Because the quality of wine is non-binary (ex. Good Wine/Bad Wine), a linear regression model is a more appropriate approach than logistic regression.

Data Preparation

In looking at the features associated with this dataset, the correlation amongst each feature and the target feature (quality) was computed. This was done in an effort to determine how each feature related to each other. Below is a heatmap showing the correlation values:



In looking at the above heatmap, the strongest positive relationships (as one feature increases, so does the other) are identified as:

- fixed acidity & density
- fixed acidity & citric acid
- free sulfur dioxide & total sulfur dioxide

While the strongest negative relationships (as one feature increases, the other decreases) are identified as:

- fixed acidity & ph
- volatile acidity & citric acid
- citric acid & ph

To determine which feature columns to use, I felt *0.20* was an appropriate correlation threshold. A correlation value of 0.20 typically indicates a small to medium strength relationship. Using this filter produced four feature columns for the linear regression model:

```
###DETERMINE NUMBER OF FEATURES TO USE###
# find absolute values of correlation values
corr = abs(corr)
# 0.20 for small to medium strength correlation
featureColumns = corr[(corr >= 0.20)].index.values.tolist()
```

```
Feature columns: ['volatile acidity', 'citric acid', 'sulphates', 'alcohol']
```

Once the appropriate features were separated from the target attribute, the linear regression model was created. Prior to injesting the data into the linear regression model and performing the analysis, the data was split between training and testing data, utilizing scikit-learn's *train_test_split* method. When splitting the data between training and testing, I opted for a 75% training/25% testing breakdown:

```
x_train,x_test,y_train,y_test = train_test_split(x, y, train_size=0.75, test_size= 0.25, random_state=1)
```

Regression Model Analysis

Below are the metrics associated with my linear regression model:

Metric	Value
Score	0.33
Mean Absolute Error	0.50
Mean Squared Error	0.39
Root Mean Square Error (Training Data)	0.67
Root Mean Square Error (Testing Data)	0.63

Looking at the above metrics, my linear regression model's predictions were only 33% accurate. However, the root mean square error (RMSE) for the test data (0.63) was similar to the RMSE for the training data (0.67), which indicated my model was fit correctly.

Retrospective Analysis

One computation which I altered, to view the difference in results, was the correlation threshold. First, I lowered it to 0, which produced an accuracy score of 0.34. I subsequently increased it to 0.40, which decreased the accuracy score to 0.23. Given that the model was fit correctly, a correlation threshold between 0.00 - 0.25 produced the optimal results.

Appendix

Dataset Features and Their Respective Descriptions

Feature	Description
fixed acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
citric acid	found in small quantities, citric acid can add 'freshness' and flavor to wines
residual sugar	the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
chlorides	the amount of salt in the wine
free sulfur dioxide	the free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
total sulfur dioxide	amount of free and bound forms of $S0_2$; in low concentrations, SO_2 is mostly undetectable in wine, but at free SO_2 concentrations over 50 ppm, SO_2 becomes evident in the nose and taste of wine
density	the density of water is close to that of water depending on the percent alcohol and sugar content
pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
sulphates	a wine additive which can contribute to sulfur dioxide gas ($S0_2$) levels, which acts as an antimicrobial and antioxidant
alcohol	the percent alcohol content of the wine
quality	**output variable (based on sensory data, score between 0 and 10)**