CS 5010 Final Project
3 August 2017
Tyler Lewis: tal3fj
Adrian Mead: atm4rf
Court Haworth: ach2wd

Final Project

Introduction:

The wine industry is enormous. Each year sees at least $30BN a year in US revenues alone. There are thousands of wineries around the world with the average American drinking nine liters of wine a year.[1] Many adults enjoy relaxing with a glass of wine after a long day, ourselves included. With the huge numbers involved, it makes sense that any knowledge that can be discerned from wine data is going to be worth a ton of money. In particular, we were interested in finding the sort of variables that are strong predictors of wine quality so that we could produce an accurate model to predict a wine's popularity.

Utilizing UC Irvine's convenient machine learning repository, we discovered two data sets containing information on both red and white wines with various input variables and one output variable, "quality". These data sets were taken knowing full well that a wine's quality is not absolute: it depends on who is doing the judging. The data claims that quality is provided as the mean of scores from at least three professional wine tasters.

The Data:

The two data sets used for our analysis were obtained from UC Irvine's machine learning repository. White wines were contained in one csv and red wines in another. Each data set is related to red and white wine variants of the Portuguese "Vinho Verde" wine. There are 4,898 observations in the white wine data set and 1,599 observations in the red wine data set. The attributes in our data sets include:

Input Variables:

1	Fixed acidity
2	Volatile acidity
3	Citric acid
4	Residual sugar
5	Chlorides
6	Free Sulfur Dioxide

7	Total Sulfur Dioxide
8	Density
9	pH
10	Sulphates
11	Alcohol

Output Variable:

12	Quality

Note: Unfortunately, due to privacy and logistic concerns, the data sets do not include any information regarding grape types, wine brand, selling price, and several other important characteristics. In hindsight, this information would have proved valuable in terms of creating an end product. “Quality” is based on a score between 0 and 10, where 0 represents the poorest quality and 10 represents the highest quality. Additionally, the data set classes are ordered and not balanced. That is, there are much more normal wines than excellent or poor ones.

We wanted to work with these data sets as they were the only wine data sets from UC Irvine that had “quality” as a variable. It was important to us to try and uncover useful information that could be applied in the real world and make an impact. If we were able to create a model to predict a wine’s quality, or uncover certain attributes that lead to a particular rating, we could remove some of the subjectiveness out of the rating itself by working with many observations. This information could prove very useful to those in the wine and restaurant industry as features such as “taste” are difficult to measure. However, we also favored these two data sets as it made separating red and white wines into classes much easier. Thus, we could provide insight into what factors encompass each type of wine. Perhaps it would be beneficial to certain individuals in the wine industry to see exactly what makes up a red wine and how it differs from its white counterpart.

Data Preprocessing:

To begin with, the data was fairly clean. We came in and didn’t have to do much work at all as there was no complicated text-parsing or missing values to worry about. The only real work that need to be done was changing the column names to replace spaces with underscores (_), and adding a column for whether the wine was red or white.

We did look into outlier detection of our own design. The general idea was that every column might have data that looks like an outlier, but that by itself might not indicate that the entire wine row should be removed from the dataset. Rather, we’d want to identify true outliers by looking for outlier-esque behavior in at least two variables. To create this effect, we defined outlier-esque behavior as any value that fell outside of the mean +/- 3 * stddev for that column. Then we found all the rows where this condition was fulfilled for at least two columns. We did this for the two wine colors separately, so that white wine outliers were handled apart from red wine outliers. There were 79 white wine outliers and 71 red wine outliers. At this point, we removed these entries from our dataset and ran our analyses.

Data Structure:

Extracted data was stored immediately as a Pandas DataFrame object and remained in this format for the remainder of the analysis. This made it easy to manipulate and graph on the fly, giving us quick insights into the structure of our data.

We made use of matplotlib for plotting, numpy for manipulating the Pandas, and sklearn for modeling.

Data Analysis / Data Processing:

Our first step of data analysis, once we had finished the pre-processing, was to get a handle on our data. It seemed useful to try some of the built-in Pandas and MatPlotLib functionality to produce univariate and bivariate information about our columns. Some initial findings are in Fig. 1

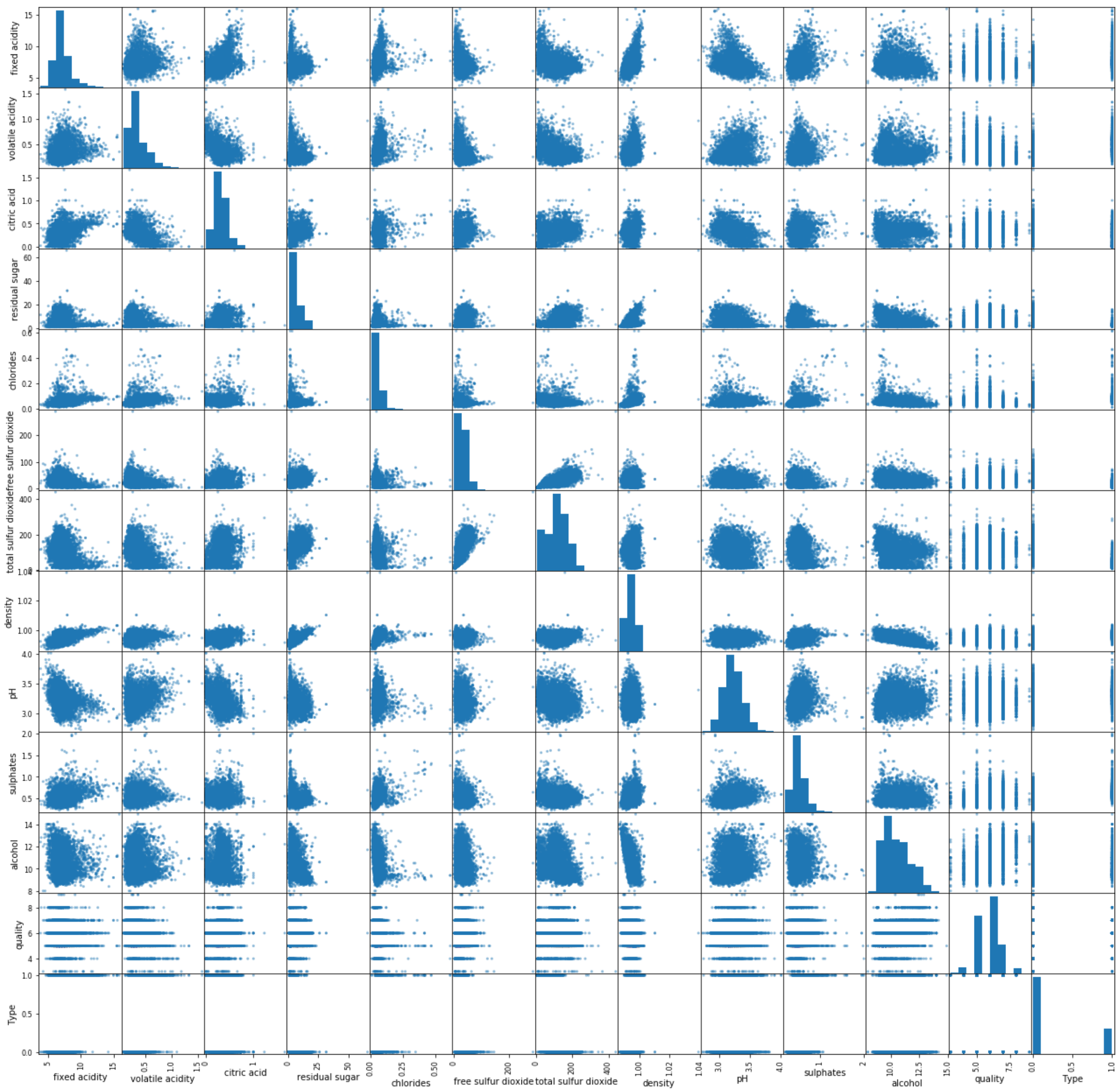


Fig 1.

It's a lot to look at, but this view helped us decide how to move forward with our analysis. Clearly it was going to be difficult to find variables that predict quality well given the cloud of points for the quality row/column . However, it does open some other interesting doors. We decided it would be worthwhile to try and predict if the wine is red/white (red = 1, white = 0 in the 'type' column) given the information from all of the other variables. It was also at this point that, based on some of the skewed and wide histograms, we decided it would be worth looking at the data for outliers.

Before approaching a model to predict the quality, it made sense to take a look at what the quality actually looks like with four distinct queries of the data. First, I wanted to better understand what were the most critically acclaimed wines, those with ratings 9 or greater. There were few of these, just 5 wines in total. We wrote these wines to another csv as a reminder to look at them later to figure out what made them so good. In our second query, the same work was done with the very worst wines, those with a score of 3 or less, in order to find what to avoid (there were 30 of these). The third query featured a histogram of quality across all wines. This is viewable in Fig 2.

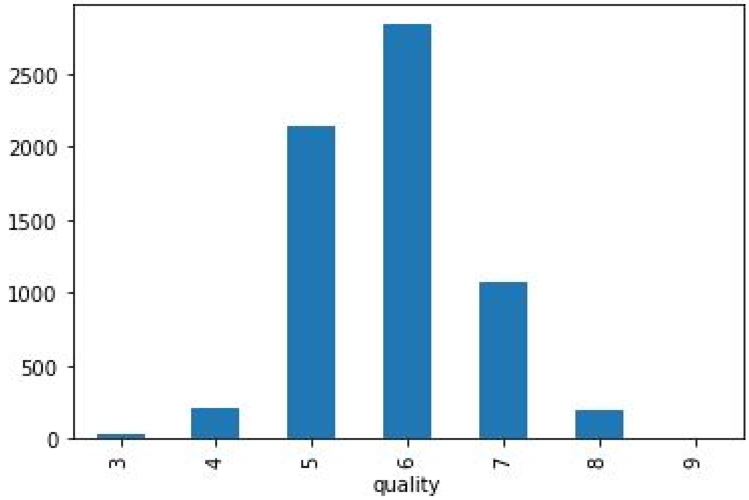


Fig 2.

As we can see, the distribution of ratings is actually pretty normally distributed with the mean around 6. Looking forward based on this shape, it's possible that the linear model that we try to fit to this data will have trouble predicting the particularly large or small ratings given the smaller number of observations.

The final query told us about the quality according to wine color in another histogram, this time with the y-axis normalized since the red and white wines have a different number of observations. This is available in Fig 3.

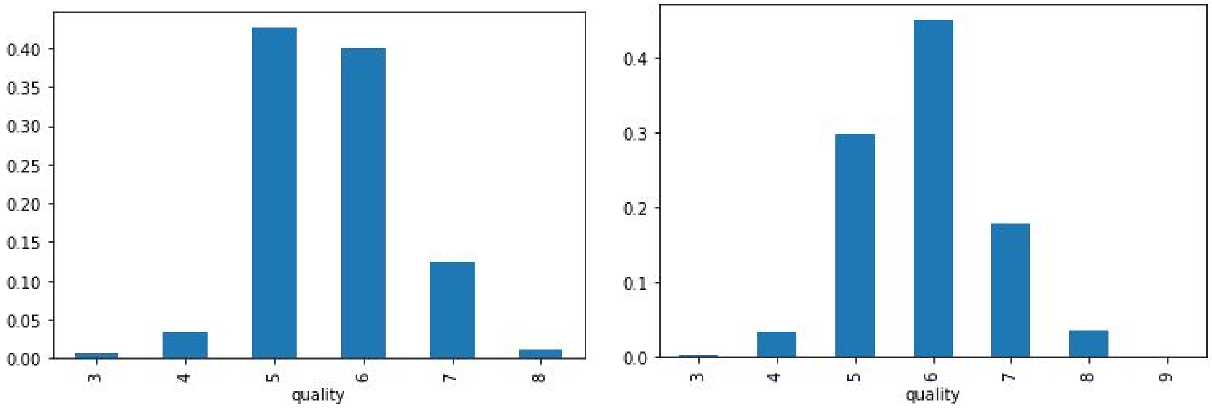


Fig 3. Red Wine (on left) and White Wine qualities

Now we begin to see that the distribution of ratings across wines is actually slightly different from white to red. We can see that on average, the white wines receive a higher rating than the red wines.

It was at this point that we decided it would make more sense to fit separate linear models to each color independently rather than one linear models to all wines. The next step was to actually build our models and visualize the results.

In order to determine the most interesting features in our data set, we performed cross validation. After building a linear regression model which included all features and testing, we ran 10-fold cross validation on models using only one input variable as predictors of quality in order to determine which variables were the strongest. This was done for both red and white wine separately and the findings are discussed in the Results section. A similar process was done for the logistic regression with a model first being trained and tested using all the variables to predict whether the wine was red or white. Then the models were created using single variables to determine which variables had the largest separation between white and red wines and were thus strong predictors in a logistic regression model. These results are also discussed in the following section.

Results:

To begin with we looked to create a linear regression model based on the features in order to predict the quality of a wine. We separated the wines into red and white wines and built separate models on their data. When we ran the White Wine Regression with All Variables we obtained an R-Squared of 0.275842827992. When we ran the Red Wine Regression with All Variables we got an R-Squared of 0.336723210472. Neither of these are particularly promising models but we decided to look at individual variables to determine if any of them had predictive power on quality. Below are two graphs, Figures 4 and 5, Figure 4 showing alcohol predicting quality, which was the best predictor for red wines, with an R-squared of 0.219981817953. The other, Figure 5, was free sulfur dioxide which was the worst predictor for red wine quality with an R-squared of -0.0040901520918. For brevity’s sake we have not included the white wine regression graphs in this section of the report but they are present as well as more red wine graphs in the appendix.

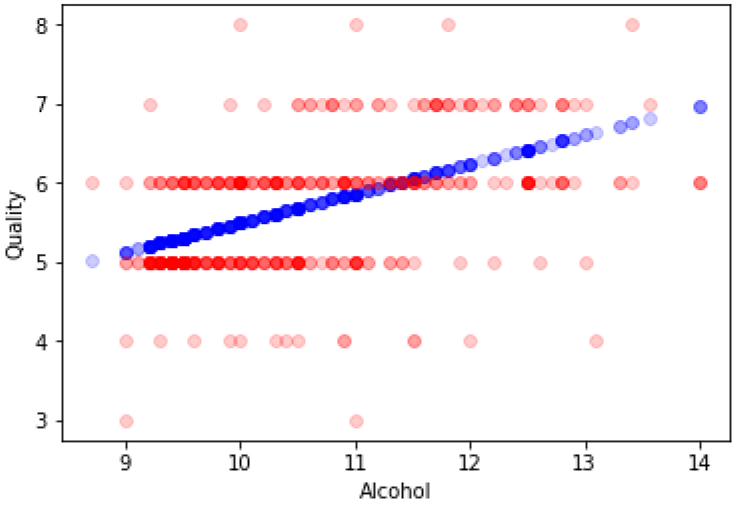


Figure 4.

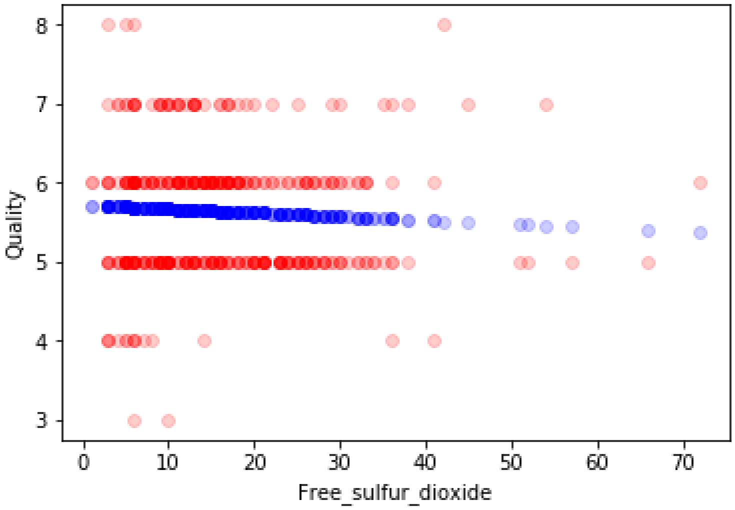


Figure 5.

After realizing that the regression model was not sufficient given the limitations, (our limited range of models, and the objectivity of the features compared with the subjectivity of a drinker's quality rating), we decided to investigate the feasibility of classification. We built a logistic model to classify the wine into red and white based on their features. As there were more white wines than red wines, we randomly subsetted 1,599 white wines so that there were an equal number of data points of white wines and red wines. A model that included all twelve features gave us an accuracy of 0.983384615385. This means that the classification is remarkably accurate, but on a holistic level it's relatively useless as the difference between red and white wine is pretty obvious so the need for classification is not prevalent (ie: just look at it). However, there is still interesting information to be gleaned when looking at what variables are useful for classification. That would tell us which variables have a distinction in distribution between red and white wines and consequently give information about the physicochemical differences between the two types. As individuals who do not know much about the chemical properties of wine, this information was interesting to us. Total sulfur dioxide turned out to be the best predictor, with an accuracy of .921230769231 when a model was built using it as the only feature. As you can see in Figure 6, the range and distribution among total sulfur dioxide values were quite different between red wines(1.0) and white wines(0.0).

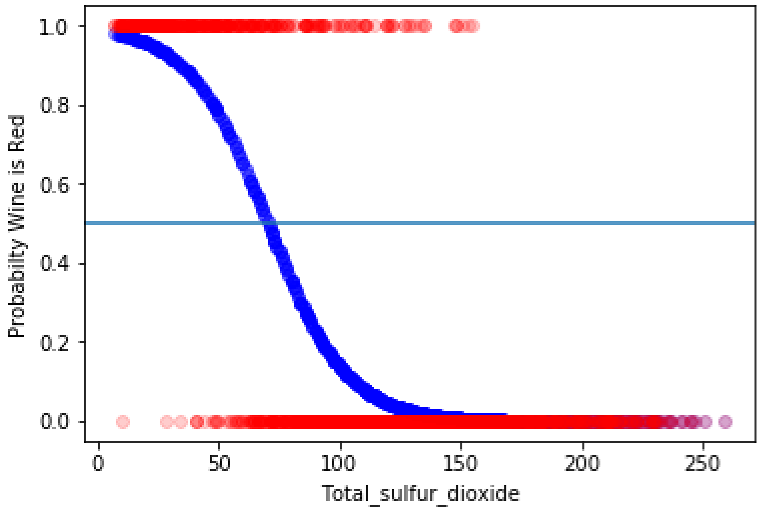


Figure 6.

On the other hand, as you can see in Figure 7 below, residual sugar was the worst predictor between the two types of wine, with an accuracy of 0.753230769231. This is really not that poor of a predictor, lying about halfway between a model that just predicts 1 always and is correct half the time (accuracy .5) and a perfect model with accuracy of 1.0. All in all the logistic regression was useful in determining which variables differ strongly between the two types, and histograms are included in the appendix that offer a visual representation in the difference of spread and distribution amongst these variables for the two wine categories.

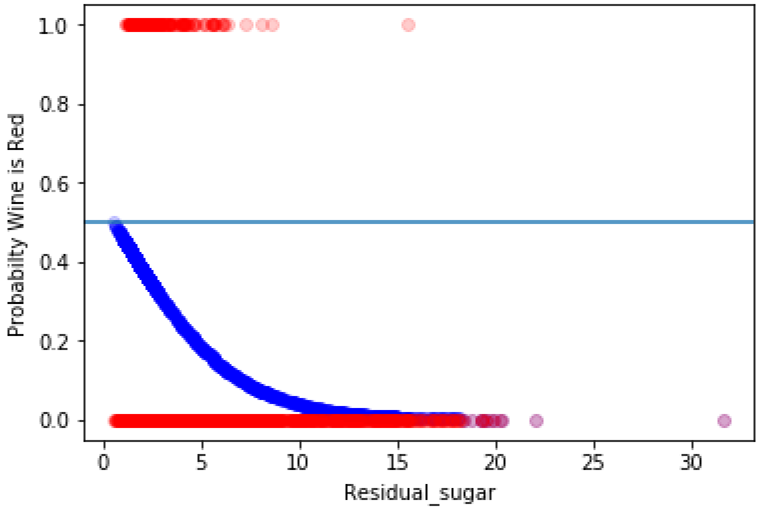


Figure 7.

Testing:

The first unit test we performed was right at the very beginning -- scraping the wine data from the web. We wrote a function named funcGetDataFromURLAndFormat(). It's a bit evident from the name, but the purpose of this function was to take a URL as input and to return a Pandas DataFrame of that page after scraping it and changing the spaces in the column names to underscores. I wrote a test to ensure that it behaves correctly when querying an empty URL (so returns the string 'empty URL'). I wrote another test to check that we receive predictable input when cURLing the wine datasets.

Conclusions (Explanation of Results):

After extensive analysis of the data, our findings returned a promising result: alcohol and density appear to be correlated to quality for white wines and alcohol and volatile acidity were the two best predictors of quality for the red wines. These results can be used by individuals in the wine industry, or even the casual wine drinker, to separate high quality and low quality wines. With our data, perhaps an individual could compare the alcohol content and density (or volatile acidity) levels of two types of wine and make an educated buying decision to select the wine that most likely has the higher quality. Furthermore, the application of our findings translates to businesses or individuals who create wine from scratch. To increase the relative alcohol content and density levels would improve the wine's quality rating and thus make for a more sought-after wine. To stay competitive, businesses are always looking for an edge. This information could provide them with that edge and enable that business to improve its product and potentially increase sales.

After finding a relationship between density, alcohol, and quality, we shifted our focus to analyzing what factors comprise a red vs a white wine. By using total sulfur dioxides and volatile acidity, we are able to build a model that correctly predicts whether the wine is white or red with an accuracy of 95.2%. Clearly those two variables are powerful for determining whether or not the wine is red. Perhaps this finding could be useful for those deep in the wine industry looking to identify what factors contribute to a red wine. Furthermore, if there is an instance where an individual is buying wine online with only a description of attributes and no identifying picture, our data could help identify whether or not those wines are red.

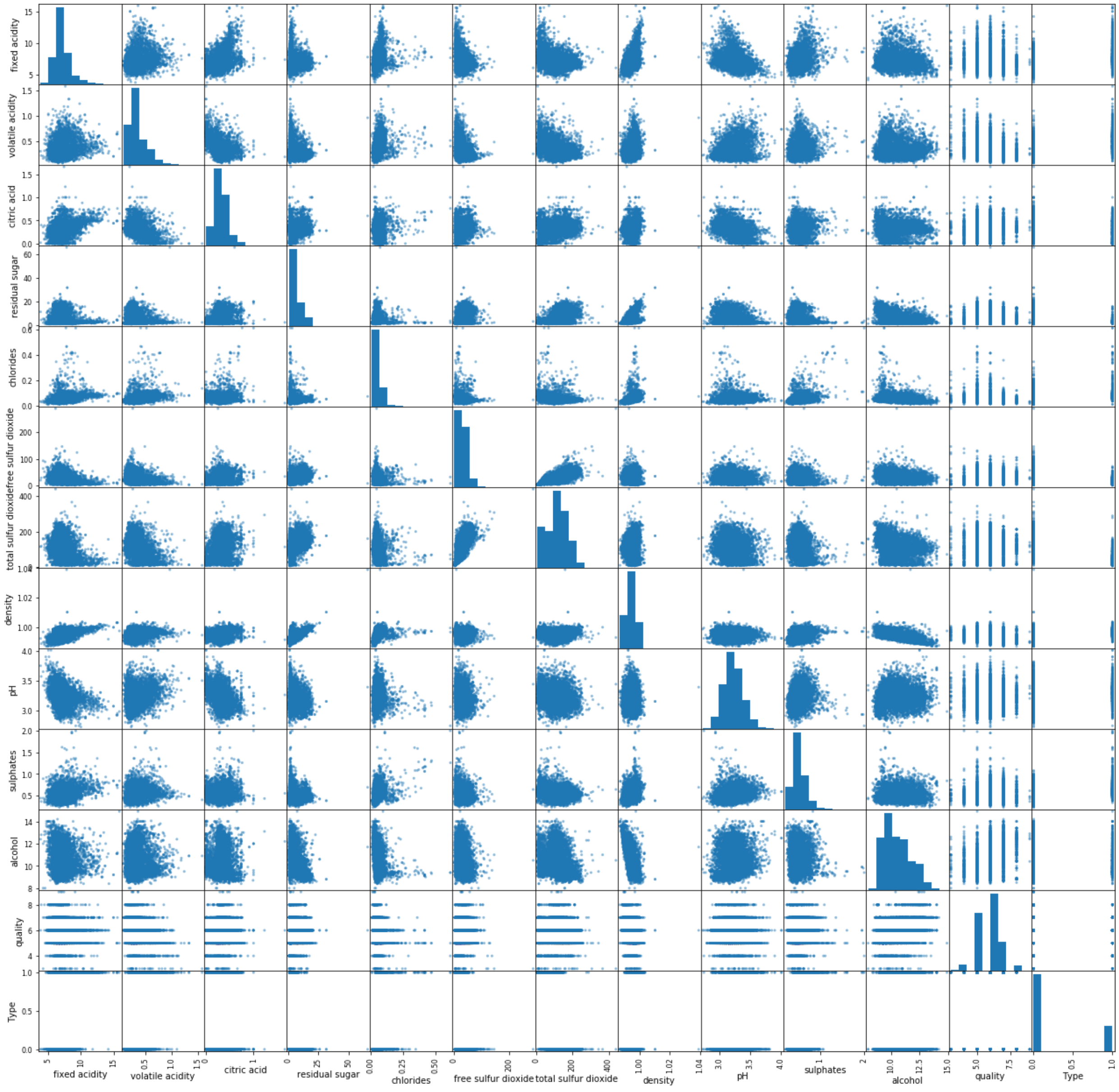
Although our results were promising, gathering more data would greatly increase both the functionality and applicability of our program. Examples of the types of data that would help improve our program include: age, grape types, wine brand, selling price, descriptions of the wine, geological information, ways the wine was stored, and any other identifying characteristics. It would be exciting to expand upon this project and provide a client-oriented product with greater quality accuracy. The hidden power inside wine data could revolutionize the production of wine and perhaps translate to other industries (restaurant, entertainment, fashion) where subjective factors like "taste" play a significant role. Furthermore, if we were to look at combinations of variables, rather than single variables themselves, perhaps we could improve the performance of our model. In the scope of this project, we are excited to have learned a great deal about wine and the industry as a whole and look forward to exploring other data.

Extra Credit:

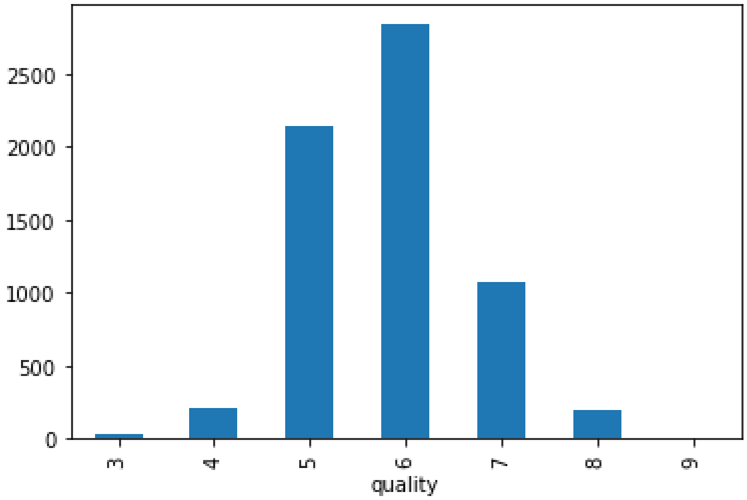
To enable collaborative coding and to go above and beyond, we created a function named funcGetDataFromURLAndFormat() that takes a URL as input and returns a Pandas DataFrame from that page after scraping it and changes the spaces in the column names to underscores. Additional testing was performed to make sure the function worked as intended. Without an explicit working directory, utilizing a webscraper to grab the data from UC Irvine's website made it easier and more efficient to code on different computers. Webscraping is an incredibly useful tool and one that we plan on using in future projects and in the professional world.

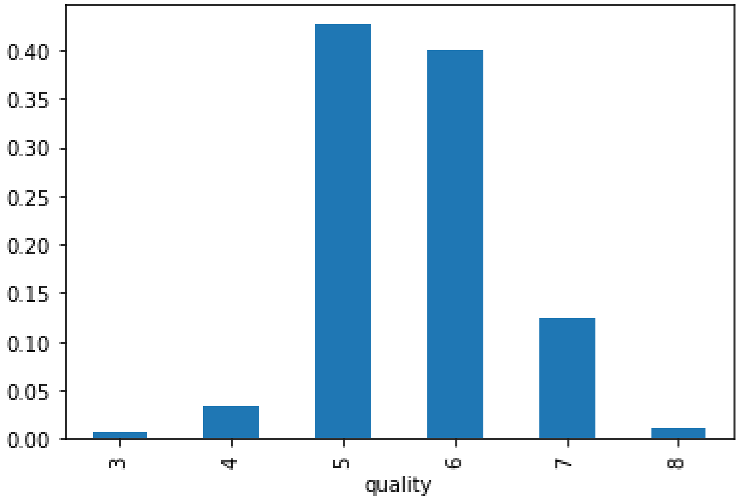
Appendix

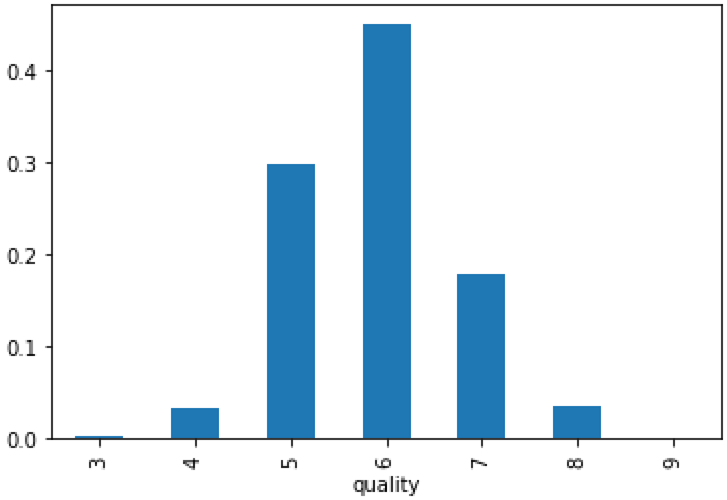
Scatter Matrix



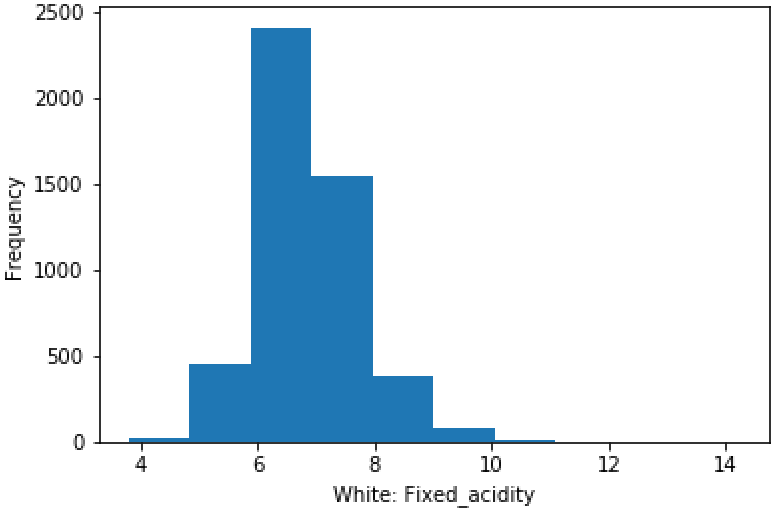
Histogram - Quality

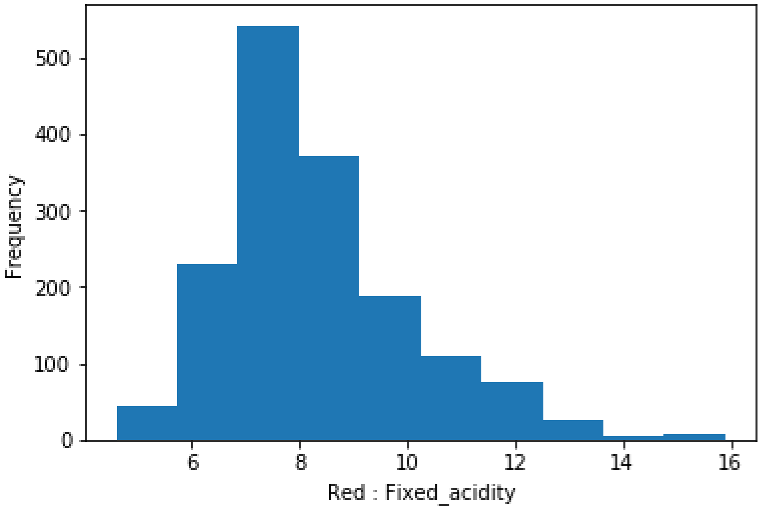


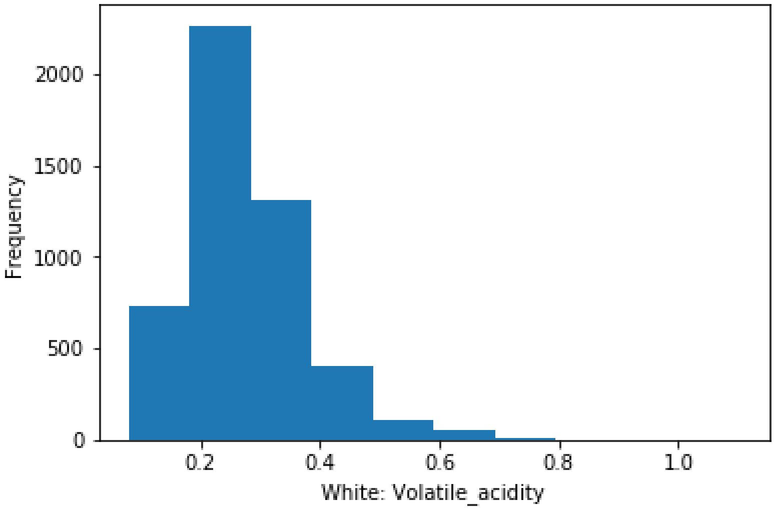


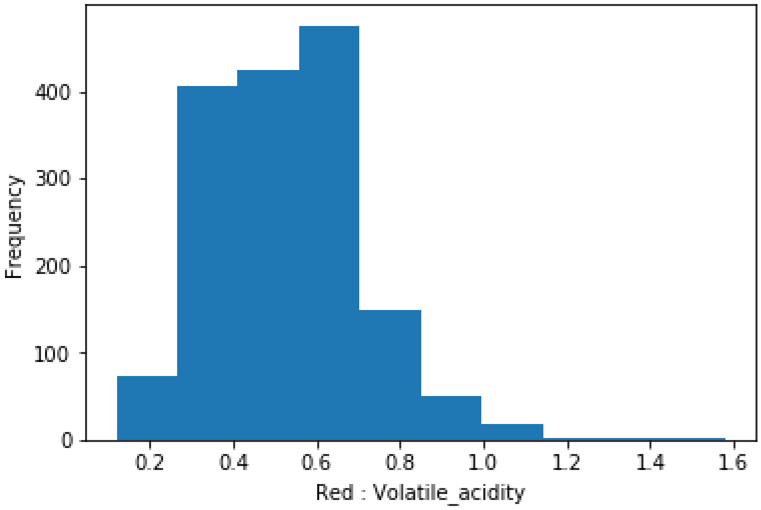


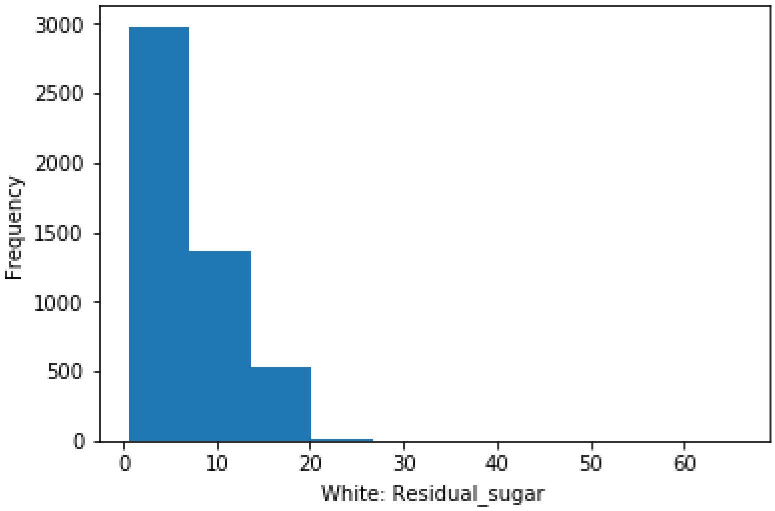
```
-------
White Wine Regression with All Variables:
R-Squared: 0.27496725871
-------
Red Wine Regression with All Variables:
R-Squared: 0.329238124008
-------
One Feature Linear Regression
-------
White Wines:
citric_acid: -0.005831 (0.005512)*
density: 0.088528 (0.073801)*
sulphates: -0.003663 (0.005272)*
alcohol: 0.217882 (0.053309)*
-------
Red Wines:
volatile_acidity: 0.155060 (0.091357)*
free_sulfur_dioxide: -0.002968 (0.010960)*
pH: -0.001156 (0.018044)*
alcohol: 0.214596 (0.087871)*
-------
```

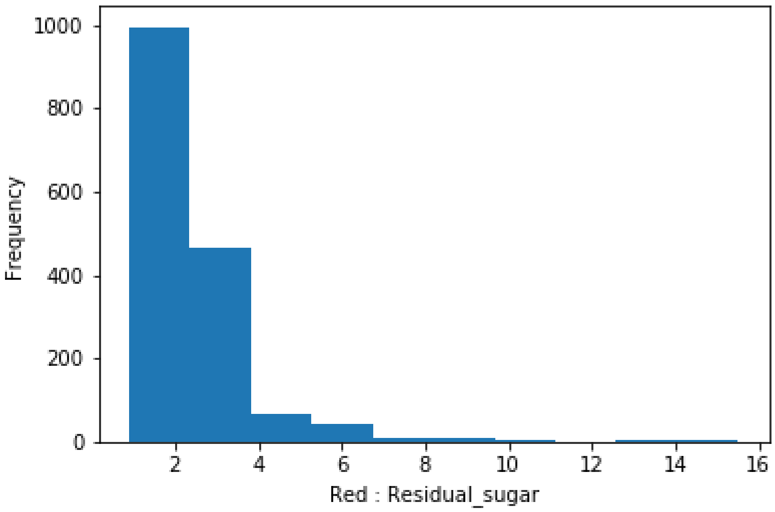

2500
2000
1500
1000
500
0
Frequency
4
6
8
10
12
14
White: Fixed_acidity

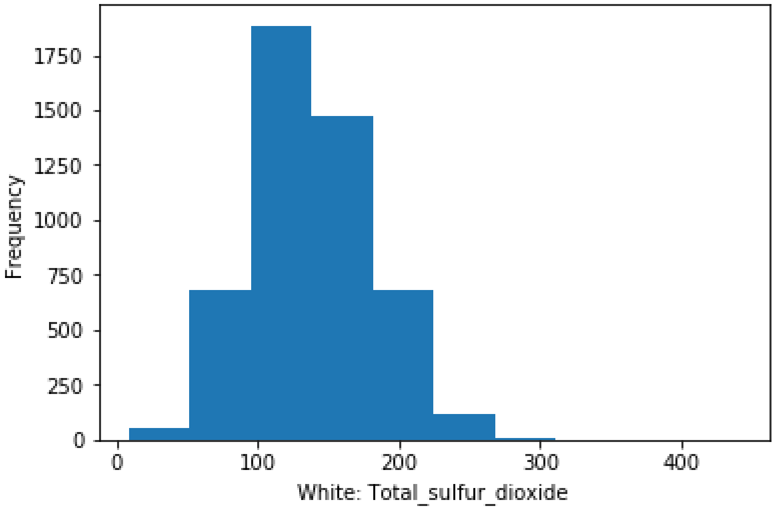

500
400
300
200
100
0
Frequency
6
8
10
12
14
16
Red : Fixed_acidity

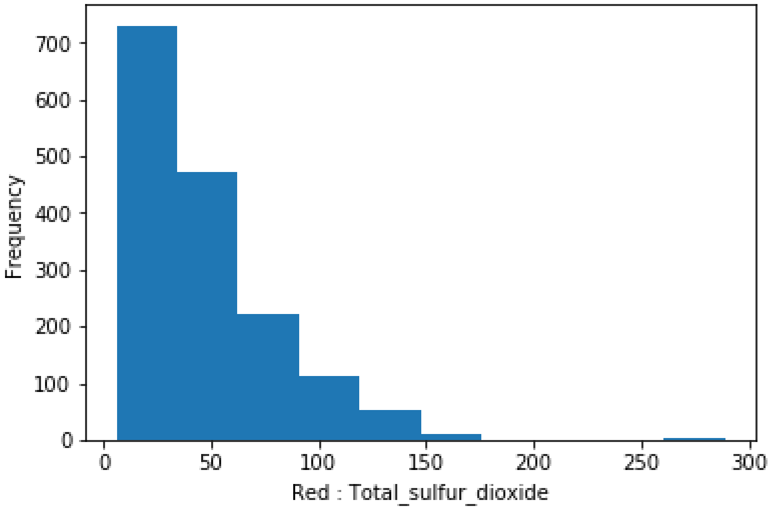

Frequency
0
500
1000
1500
2000
0.2
0.4
0.6
0.8
1.0
White: Volatile_acidity

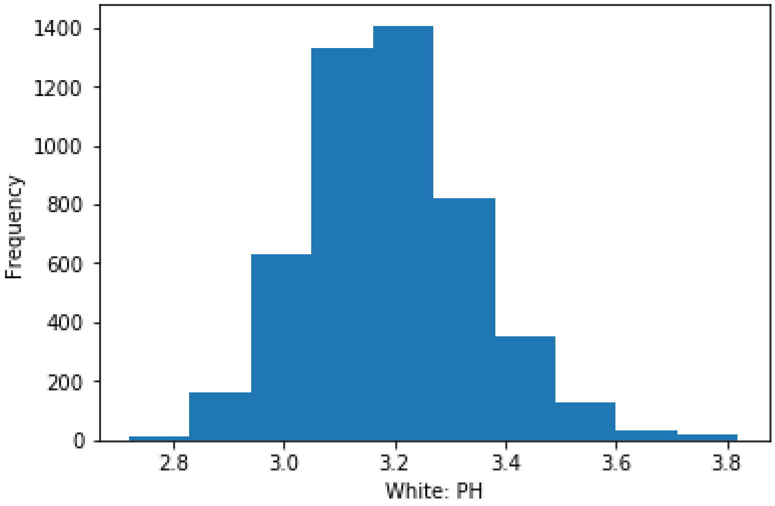

Frequency
0
100
200
300
400
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Red : Volatile_acidity

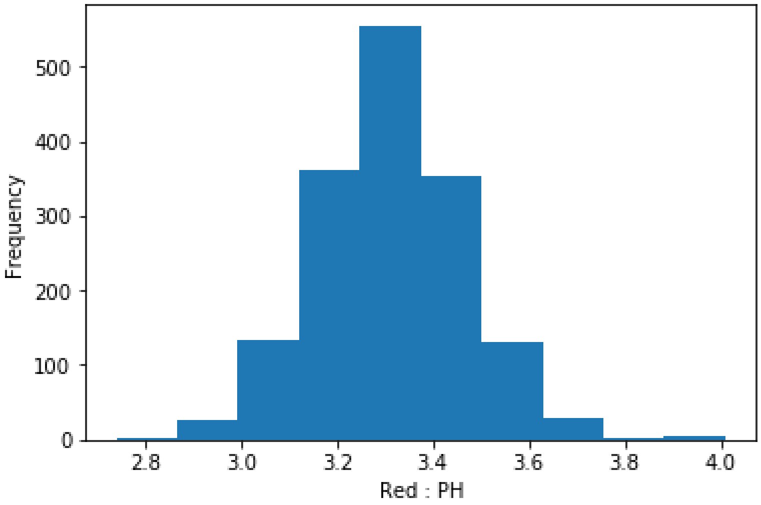

3000
2500
2000
1500
1000
500
0
Frequency
0
10
20
30
40
50
60
White: Residual_sugar


1000
800
600
400
200
0
Frequency
2
4
6
8
10
12
14
16
Red : Residual_sugar


Frequency
1750
1500
1250
1000
750
500
250
0
0
100
200
300
400
White: Total_sulfur_dioxide

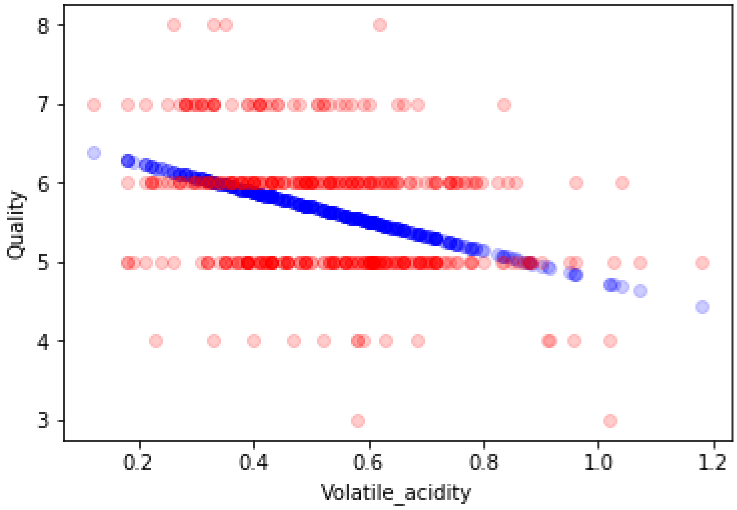

Frequency
700
600
500
400
300
200
100
0
0
50
100
150
200
250
300
Red : Total_sulfur_dioxide


1400
1200
1000
800
600
400
200
0
Frequency
2.8
3.0
3.2
3.4
3.6
3.8
White: PH


500
400
300
200
100
0
Frequency
2.8
3.0
3.2
3.4
3.6
3.8
4.0
Red : PH

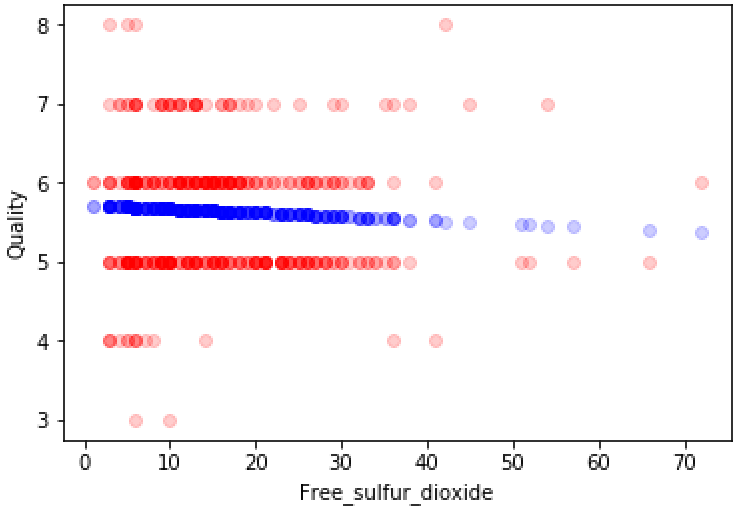
Red Wine

Linear Regression: Volatile Acidity predictor



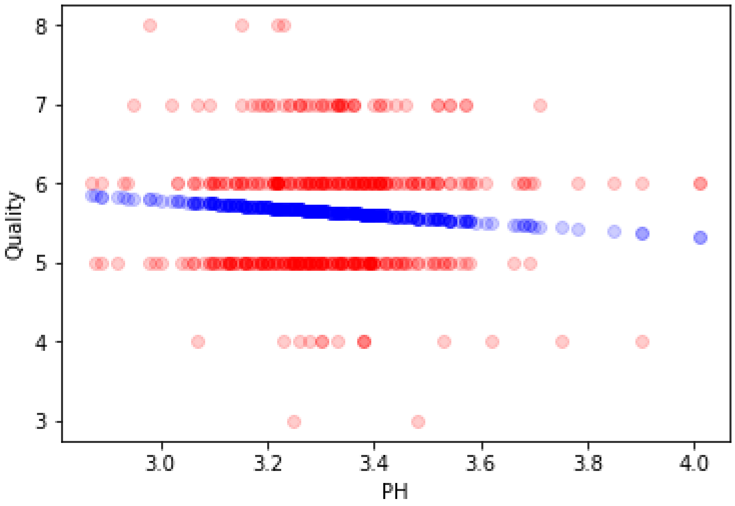
R-Squared: 0.100539289054

Linear Regression: Free Sulfur Dioxide predictor



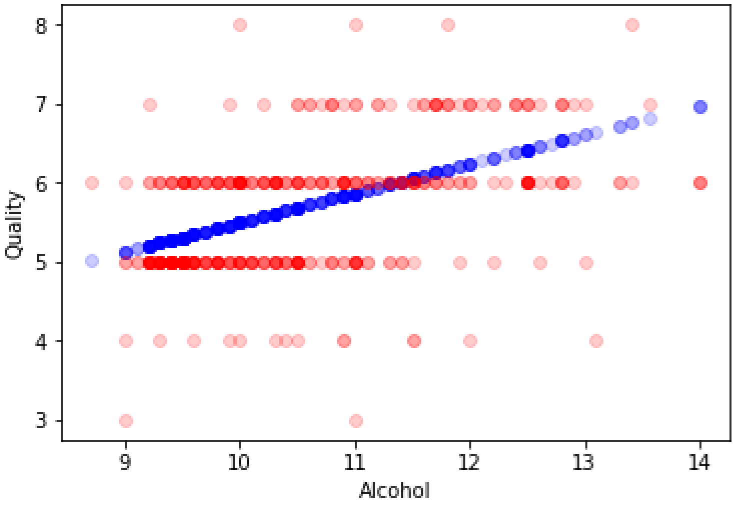
R-Squared: -0.0040901520918

Linear Regression: pH predictor



R-Squared: -0.0149111034526

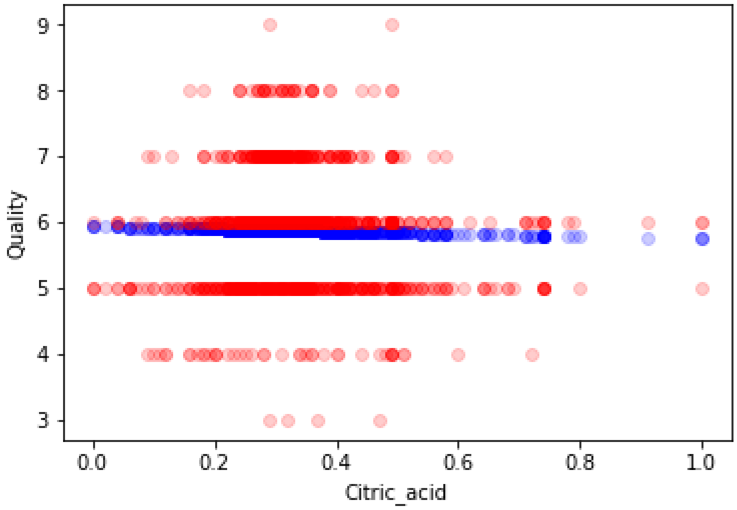
Linear Regression: Alcohol predictor



R-Squared: 0.219981817953

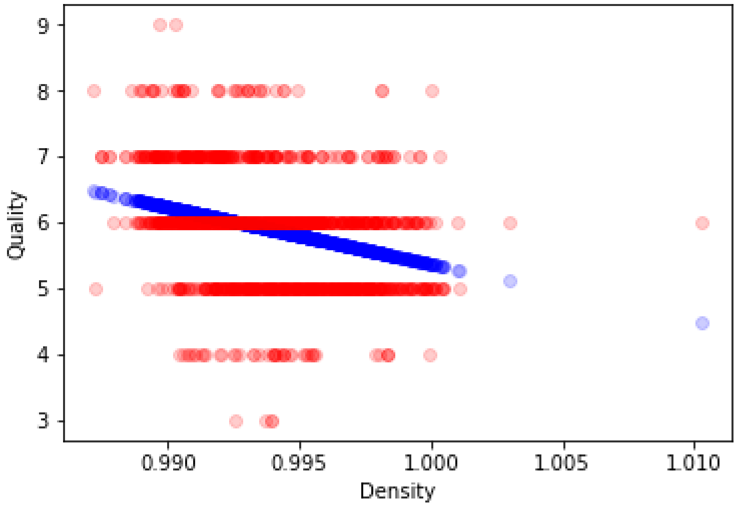
White Wine

Linear Regression: Citric Acid predictor



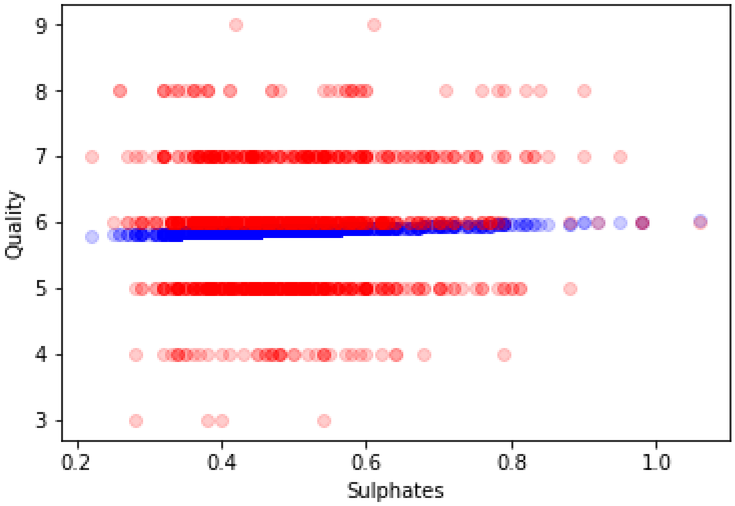
R-Squared: 0.00100341835694

Linear Regression -- Density predictor



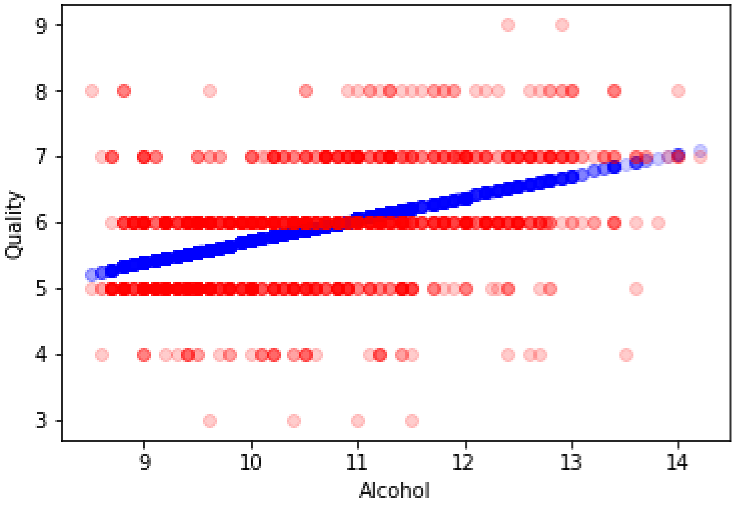
R-Squared: 0.110121892702

Linear Regression -- Sulfates predictor



R-Squared: 0.00459399800165

Linear Regression -- Alcohol predictor



R-Squared: 0.19264217974

Accuracy of Logistic Regression using 12 features: 0.982769230769

Categorization: Red or White

One Feature Logistic Models:

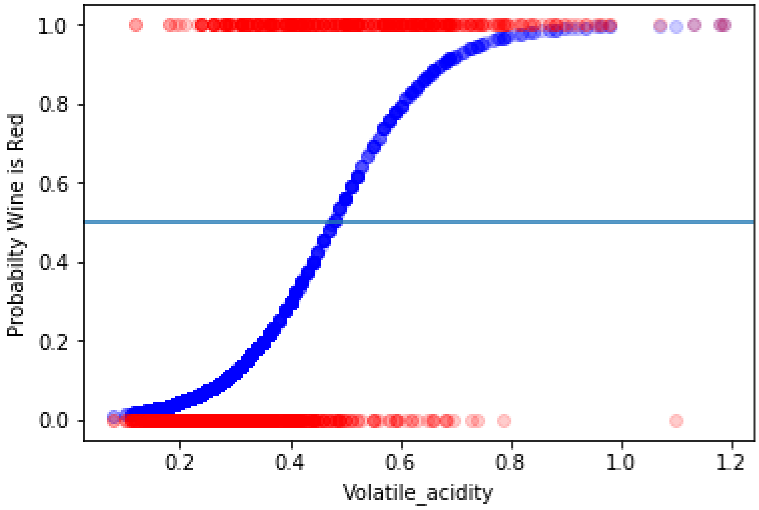
volatile_acidity: 0.842935 (0.180171)*

residual_sugar: 0.731427 (0.387315)*

total_sulfur_dioxide: 0.913158 (0.082779)*

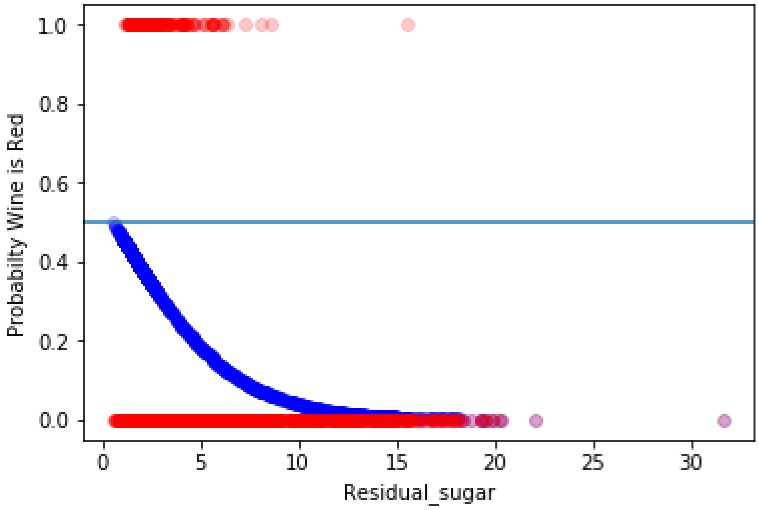
pH: 0.742927 (0.392344)*

Logistic Regression -- Volatile Acidity predictor



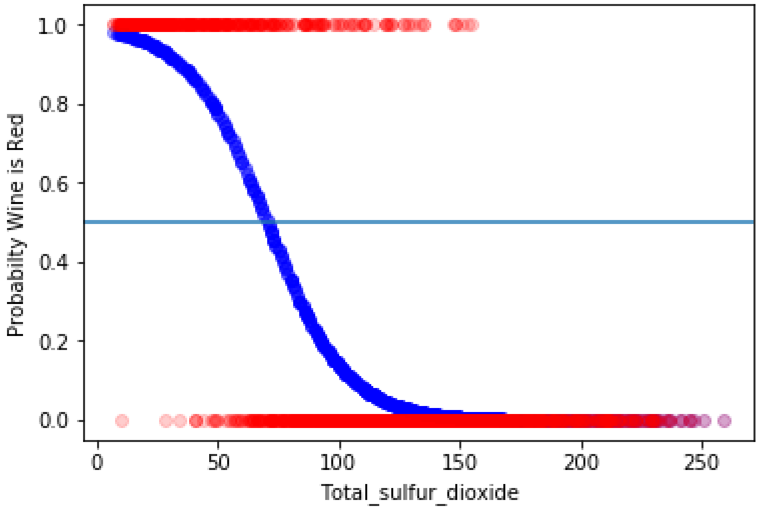
Accuracy: 0.863384615385

Logistic Regression -- Residual Sugar predictor



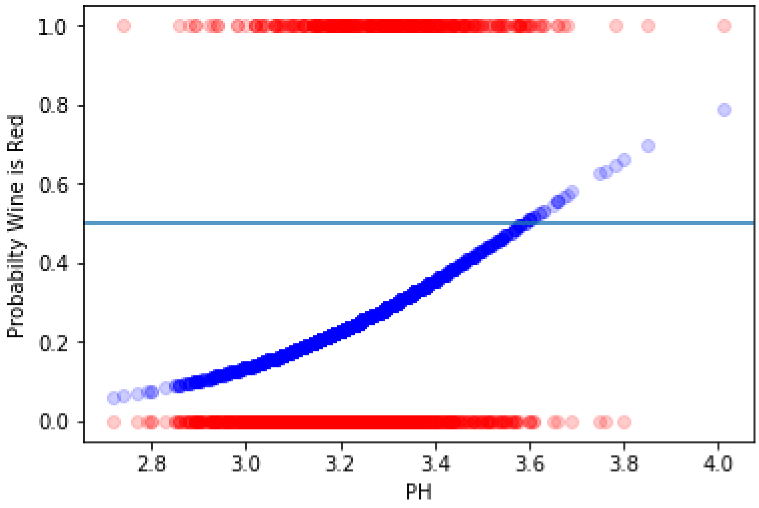
Accuracy: 0.753230769231

Logistic Regression -- Total Sulfur Dioxide predictor



Accuracy: 0.921230769231

Logistic Regression -- pH predictor



Accuracy: 0.758153846154

Accuracy of Logistic Regression using 2 features: 0.952