Project Goal:

1. Description: For our project, we are investigating the "Red Wine Quality" dataset (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009) which includes variables that represent different components of red wine. With this dataset, we are trying to find the elements which constitute a higher quality red wine. There were many more chemical components in the original dataset. In order to minimize the situation the researcher, Cortez, only adopted the most common components and made them available. In addition, the response variable is the quality of a wine which was tasted by at least three evaluators and the median had been adopted for the final quality. That's why we cannot observe very poor (scored 1 or 2) and very excellent (scored 9 or 10) wines.

2. Importance of the problem: Many think what determines the quality of wine is just taste. We have approached this issue through a little bit different point of view, as making a good wine only through controlling the whole production chains might not be economically reasonable. What if we can increase the quality and consumer satisfaction by just adding some acidity? In order to answer this, one question should be processed before: Could the chemical compounds determine the quality of a wine? By investigating whether a chemical element has a statistically significant effect on the quality of wine, we are able to replicate a recipe for higher quality red wines. With this, red wine makers will be able to understand what components they should focus on during wine production. If the result shows little to no correlation between the variables and the wine quality, other aspects of red wine might be more important such as scent, brand values, or pricing.

Exploratory Analysis:

To begin, we explored all of the predictor variables of the data set, as well as the response variable (wine quality). We noticed how the range of wine qualities only went from 3 -- 8, instead of 0 -- 10, as mentioned on the Kaggle website from which this data set was pulled from.

Initially, we attempted creating scatterplots for each of the 11 predictor variables, where the x-axis is the predictor, and the y-axis is wine quality. Unfortunately, these did not turn out to be as pretty and neat as we had hoped for, since 'Wine Quality' is a categorical variable with 6 classes (qualities 3 -- 8).

One of the first ideas we had was to categorize the wine qualities into 3 groups: 'low', 'medium', and 'high'. This would allow us to pick up on some general trends associated with each of the 3 groups. With this, the lowest wine qualities (such as 3) would be placed into the 'low' group, while the highest wine qualities (such as 8) would be placed into the 'high' group. We accomplished this by using a Pandas method called pd.qcut. **[Reference Table 1]**

An interesting observation we noticed is that there were about 3x as many 'Low' and "Medium observations as there were "High" observations. For our data set, 46.5% of the observations were "low" quality wines, 40% were "medium" quality wines, and only 13.5% were "high" quality wines.**[Reference Table 2, 3, & Figure 1]** This means that it is very uncommon for a wine to have a "high quality", which makes sense since high quality wines are rare.

Then, we created box plots for all of the predictors to take a closer look at the range of those variables.**[Reference Figure 2]**Based on the plots, most of the predictors showed outliers that could potentially skew patterns or trends, therefore we decided to use the median function to aggregate our pivot table.

We looked at the average predictor values for the different "Low", "Medium", and "High" quality wines, which were defined by pd.qcut.**[Reference Table 4 & 5]**. We noticed some general trends such as how larger values of (alcohol, citric acid, fixed acidity, and sulphates) led to higher quality wines. Additionally, lower values of (chlorides, free sulfur dioxide, total sulfur dioxide, and volatile acidity) also led to higher quality wines. Lastly, the values of pH, density, and residual sugar seemed to barely change between the different qualities of wines. Just based on this pivoted_quartile dataframe, we can expect those predictor variables (which barely changed) to be the least significant in determining the quality of wines.

Lastly, to take a closer look at the relationship between the response and predictors, we kept the quality variable in its original form (3,4,5,6,7,8) without grouping them into categories and created another pivot table based on the median of the predictors.**[Reference Table 5]** From here on, we graphed a barchart for each predictor against the response variable. **[Reference Figure 3]** Based on the barcharts, we found that both volatile acidity (negative) and citric acid (positive) have the strongest correlations with wine quality, whereas residual sugar, density, and pH shows little to no correlation. We also found some interesting shapes with total sulfur dioxide and free sulfur dioxide where the two show a normal distribution with the quality variable. After comparing these barcharts to the trends we observed from the grouped (Low, Medium, High) pivoted table values, we realized that by not assigning the different qualities into groups, we are able to define a clearer trend as there are more values on both the x and y axis. Therefore, we based our hypothesis solely on the barchart and expect the predictors that show the strongest correlation to also be statistically significant in our models.

Solution and Insights:

1. Solution: In the beginning, we decided to experiment with a linear regression model. Although the value of R-squared from the multiple linear regression model is around 0.36, It is hard to say that this value makes our prediction meaningless. Considering that there are so many potential factors excluding our predictors (volatile acidity, PH, density, and so on) and that this is a kind of exploratory approach with only few variables, the relatively low value in the R-squared can be considered to be located in an acceptable range. Based on this, we can conclude that not only the scent, mood, and pricing are important factors in determining the quality of a red wine, but also the chemical composition is able to contribute in accounting for the quality.

2. Insight: In addition, we have conducted three logistic regression models with 1.) All the variables; 2.) All except three variables thought to be less meaningful based on EDA: (density, PH, and residual sugar); 3.) With only two variables thought of as important through the result of EDA: (volatile acidity and citric acid).

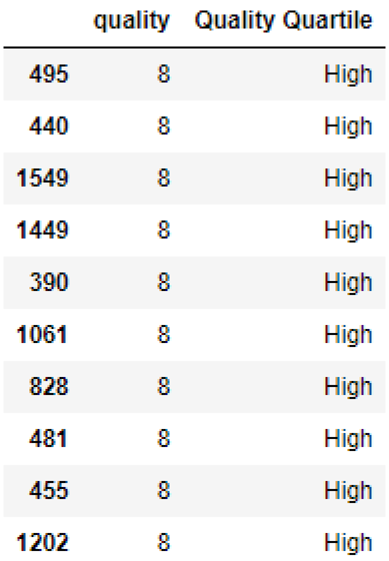
   The test accuracy for the first model with all the variables was 71.46%. For the model excluding unimportant variables, the test accuracy was also the same: 71.46%. For the last model: 61.67%.The results of the weights of the coefficients for each of the three logistic regression models we ran are shown below. The weights indicate the significance of each of the predictors. The greater the absolute value of the weight for a predictor variable, the more significance/relevance the predictor variable has.**[Reference Table 6, 7, & 8]**

Across all three models, predictors (volatile acidity, sulphate, chloride and alcohol) had always come out with the highest absolute coefficients, which indicates that they are the most important variables. Though this differs a little with our hypothesis from the EDA, we were able to correctly identify volatile acidity as an important predictor according to the model coefficients. We were also able to correctly identify the variables (density, pH, and residual sugar) as not important since the removal of those variables did not change testing or training accuracy.

Based on the result, we and winemakers would know the extent to which some variables harm the accuracy score and quality. This can also be thought of as a tradeoff between including variables and the costs that the associated chemicals bring. Winemakers then can form educated decisions during the grape raising and winemaking process, and find out which component will give more economic competitiveness.

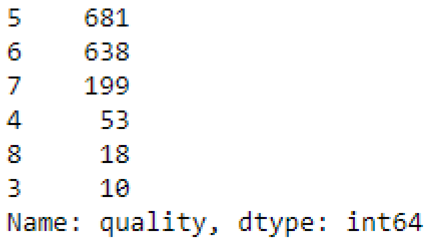
*Référence: [Cortez et al., 2009]

Appendix



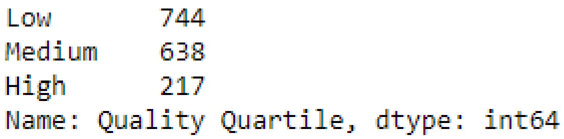
	quality	Quality Quartile
495	8	High
440	8	High
1549	8	High
1449	8	High
390	8	High
1061	8	High
828	8	High
481	8	High
455	8	High
1202	8	High

***Table 1*: Quick pull of data to verify successful separation of wines by High, Medium, and Low**



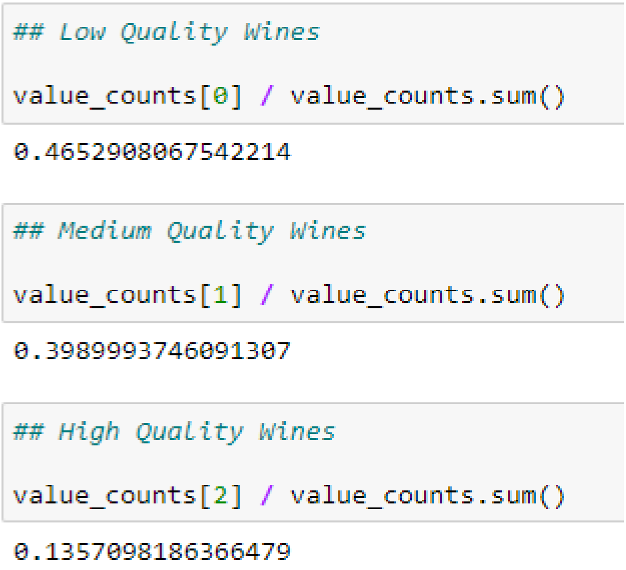
```
5    681
6    638
7    199
4     53
8     18
3     10
Name: quality, dtype: int64
```

***Table 2*: Number of wines in each numerical quality category**



```
Low       744
Medium    638
High      217
Name: Quality Quartile, dtype: int64
```

***Table 3*: The number of wines in each of our self defined categories**



```
## Low Quality Wines

value_counts[0] / value_counts.sum()
```

```
0.4652908067542214
```

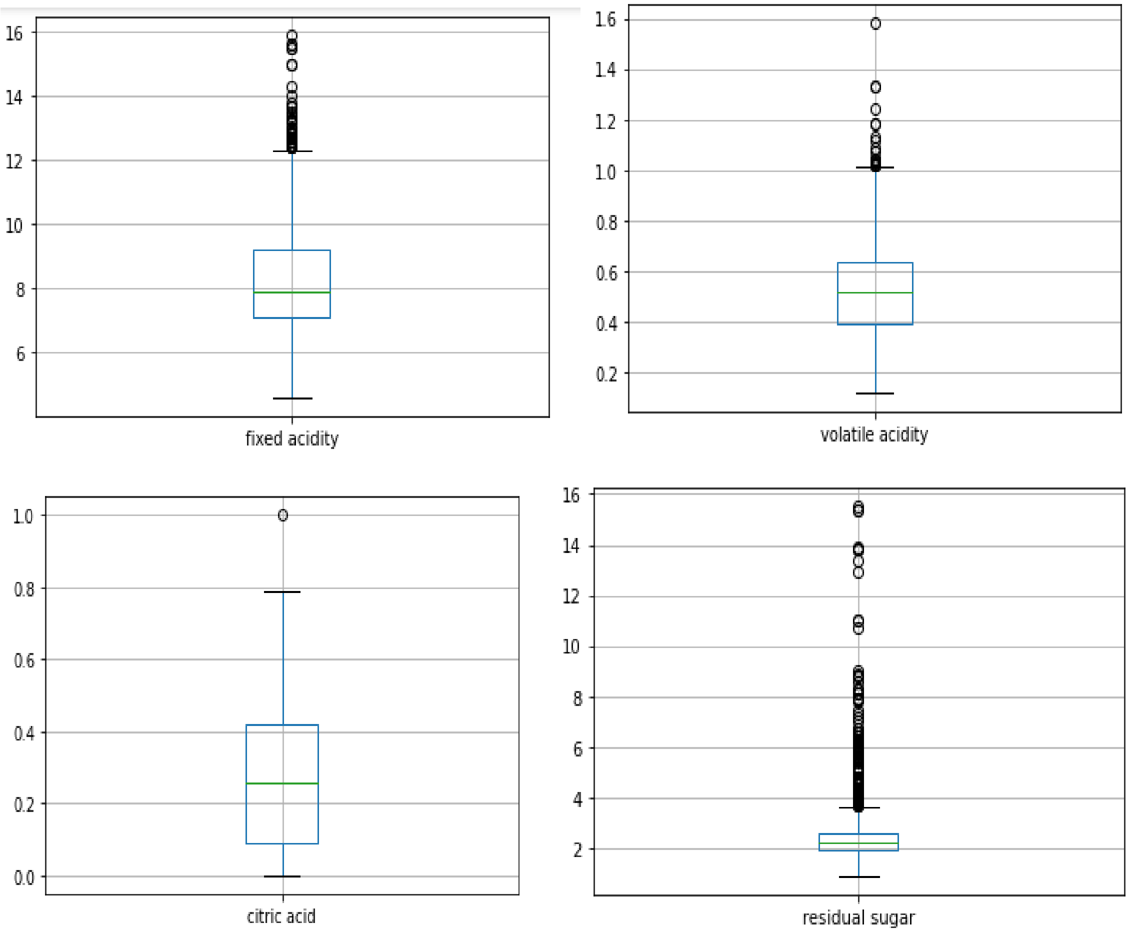
```
## Medium Quality Wines

value_counts[1] / value_counts.sum()
```

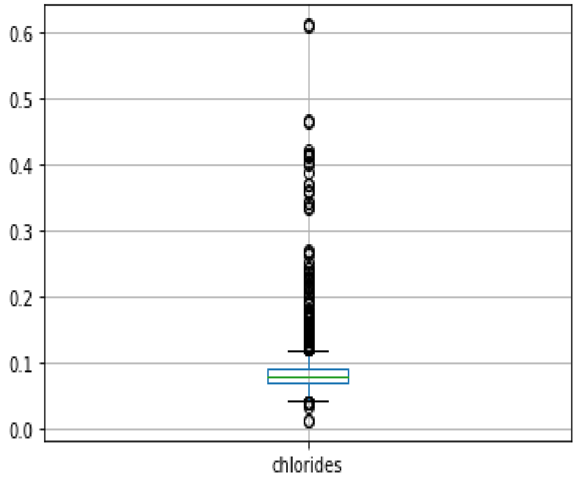
```
0.3989993746091307
```

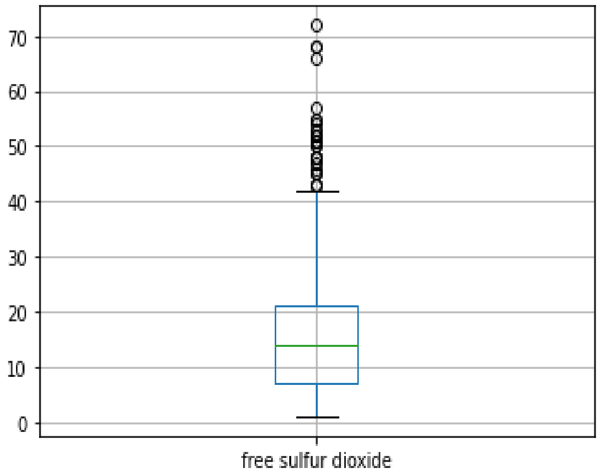
```
## High Quality Wines

value_counts[2] / value_counts.sum()
```

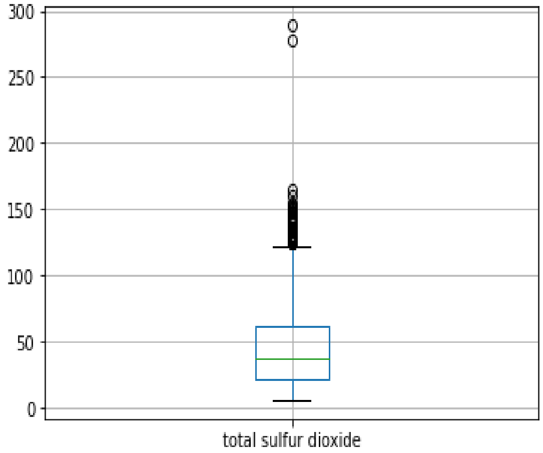
```
0.1357098186366479
```

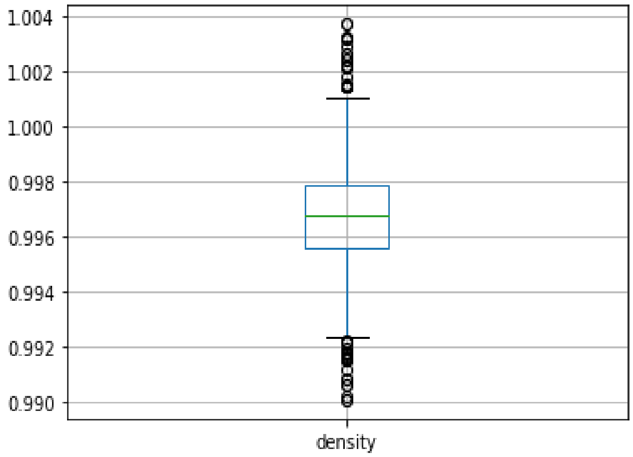
***Figure 1*: Percentage of each quality category over all the wines**

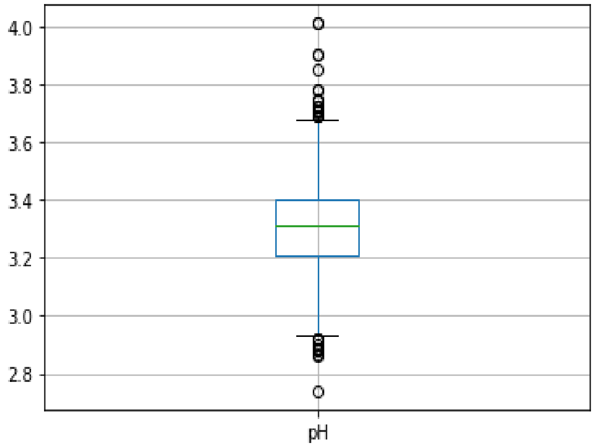


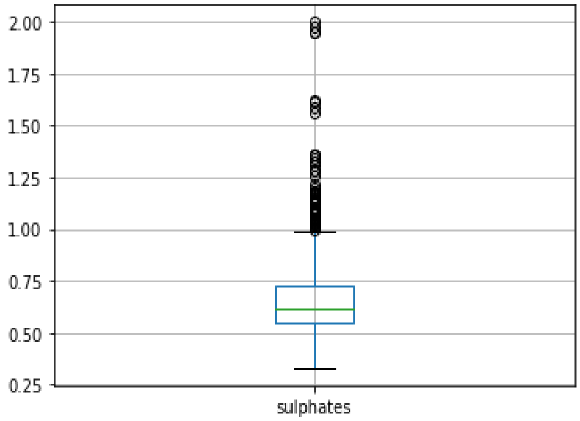

0.6
0.5
0.4
0.3
0.2
0.1
0.0
chlorides


70
60
50
40
30
20
10
0
free sulfur dioxide


300
250
200
150
100
50
0
total sulfur dioxide


1.004
1.002
1.000
0.998
0.996
0.994
0.992
0.990
density


4.0
3.8
3.6
3.4
3.2
3.0
2.8
pH


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
sulphates

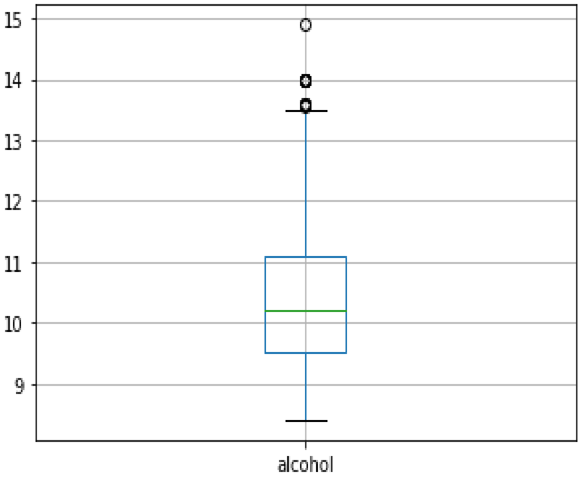
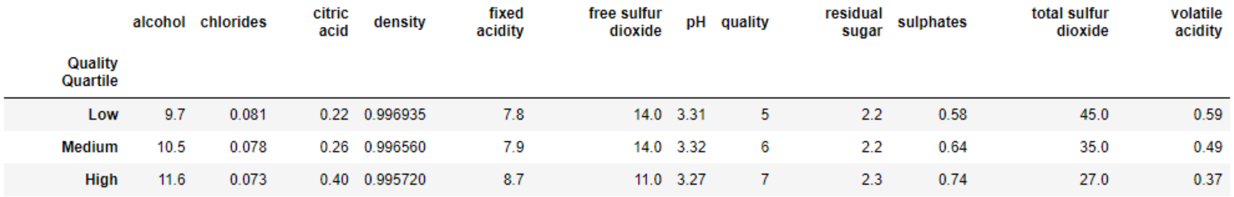
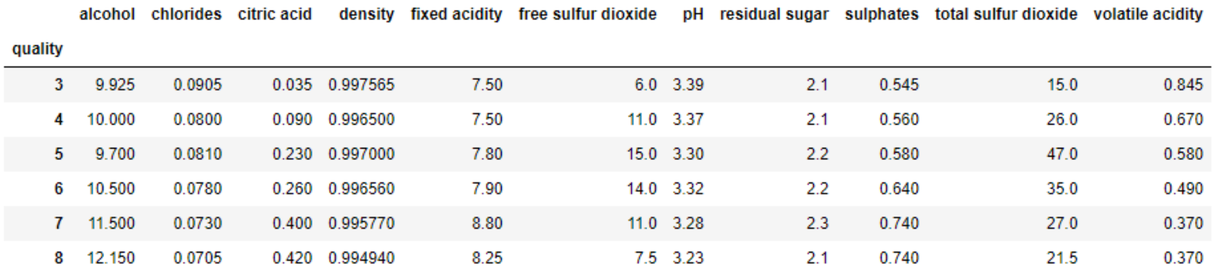


***Figure 2*: Box plots for all predictors**



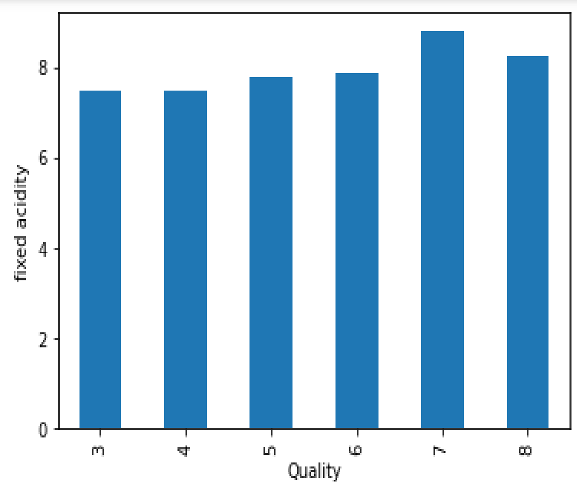
Quality Quartile	alcohol	chlorides	citric acid	density	fixed acidity	free sulfur dioxide	pH	quality	residual sugar	sulphates	total sulfur dioxide	volatile acidity
Low	9.7	0.081	0.22	0.996935	7.8	14.0	3.31	5	2.2	0.58	45.0	0.59
Medium	10.5	0.078	0.26	0.996560	7.9	14.0	3.32	6	2.2	0.64	35.0	0.49
High	11.6	0.073	0.40	0.995720	8.7	11.0	3.27	7	2.3	0.74	27.0	0.37

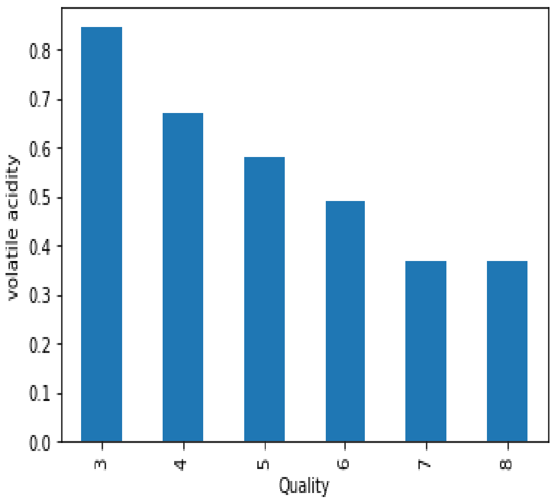
***Table 4*: Average predictor values for High, Medium, and Low categories**

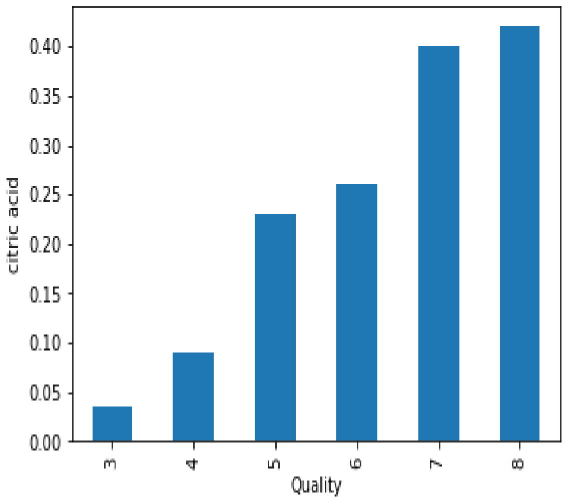


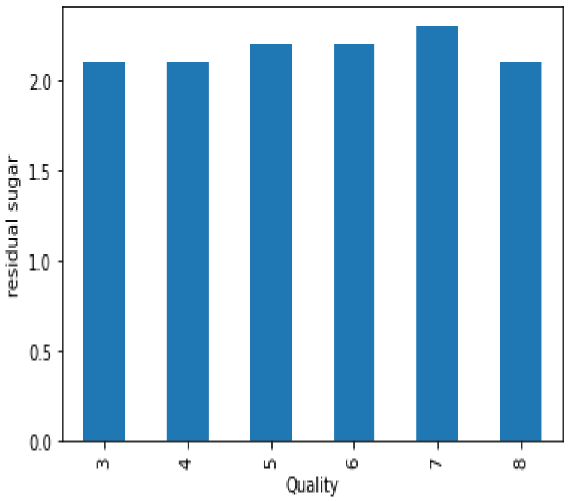
quality	alcohol	chlorides	citric acid	density	fixed acidity	free sulfur dioxide	pH	residual sugar	sulphates	total sulfur dioxide	volatile acidity
3	9.925	0.0905	0.035	0.997565	7.50	6.0	3.39	2.1	0.545	15.0	0.845
4	10.000	0.0800	0.090	0.996500	7.50	11.0	3.37	2.1	0.560	26.0	0.670
5	9.700	0.0810	0.230	0.997000	7.80	15.0	3.30	2.2	0.580	47.0	0.580
6	10.500	0.0780	0.260	0.996560	7.90	14.0	3.32	2.2	0.640	35.0	0.490
7	11.500	0.0730	0.400	0.995770	8.80	11.0	3.28	2.3	0.740	27.0	0.370
8	12.150	0.0705	0.420	0.994940	8.25	7.5	3.23	2.1	0.740	21.5	0.370

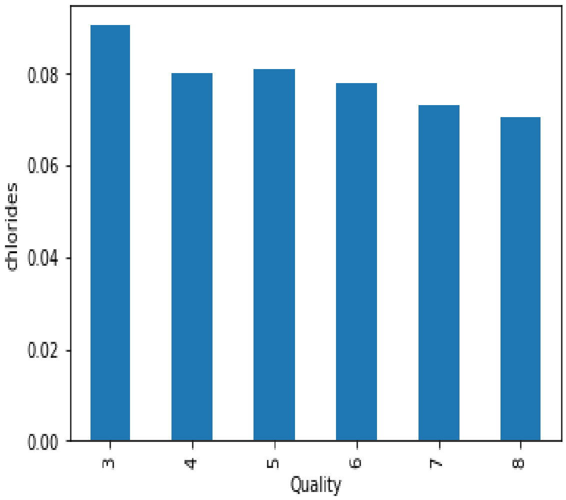
***Table 5*: Median values for each predictor variable in each numerical quality category**

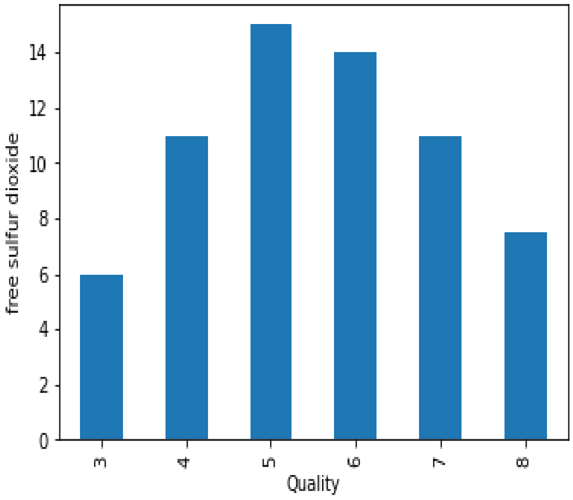

fixed acidity
0
2
4
6
8
3
4
5
6
7
8
Quality

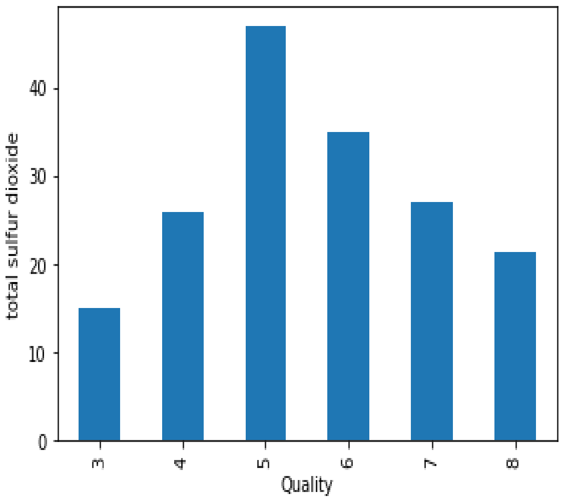

volatile acidity
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
3
4
5
6
7
8
Quality

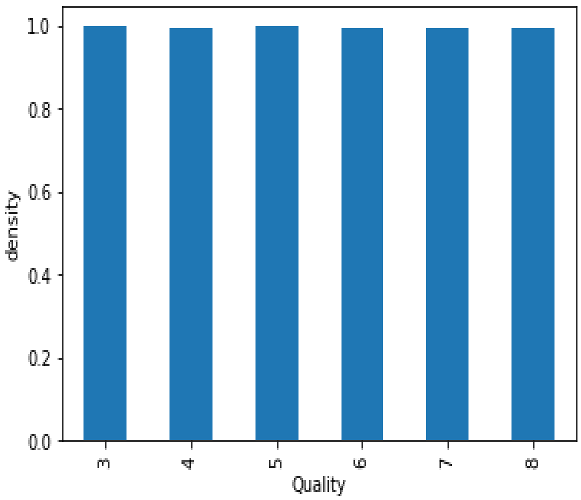

citric acid
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
3
4
5
6
7
8
Quality

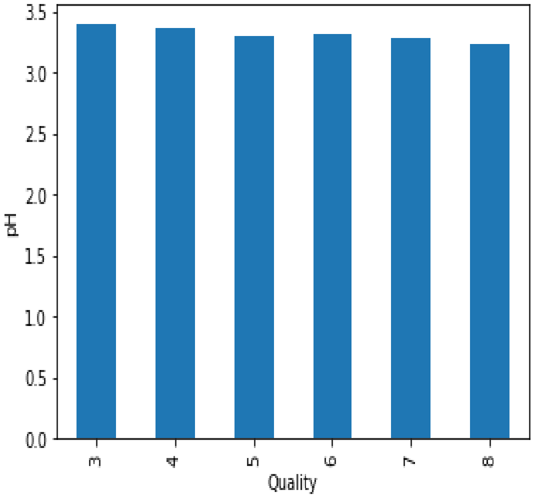

residual sugar
0.0
0.5
1.0
1.5
2.0
3
4
5
6
7
8
Quality

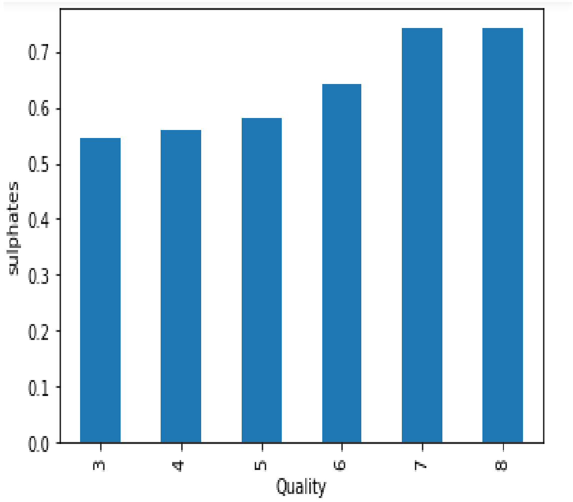

chlorides
0.00
0.02
0.04
0.06
0.08
3
4
5
6
7
8
Quality


free sulfur dioxide
0
2
4
6
8
10
12
14
3
4
5
6
7
8
Quality


total sulfur dioxide
0
10
20
30
40
3
4
5
6
7
8
Quality


density
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
Quality


pH
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
3
4
5
6
7
8
Quality


sulphates
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
3
4
5
6
7
8
Quality

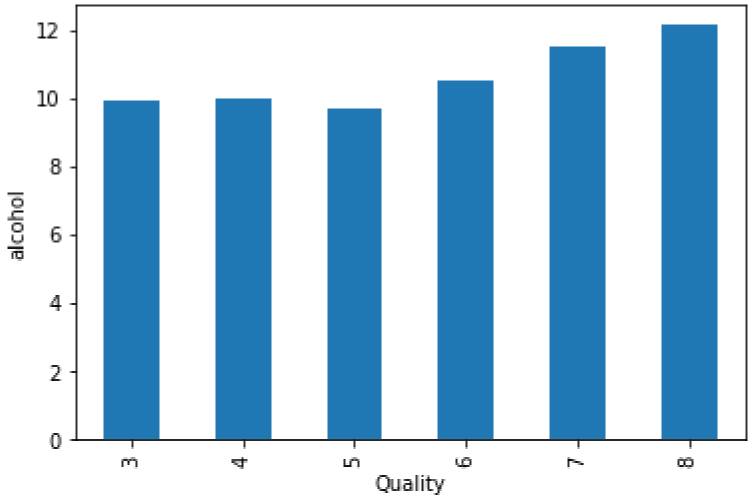
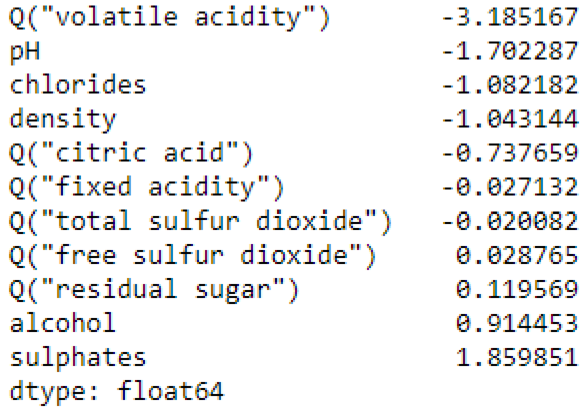
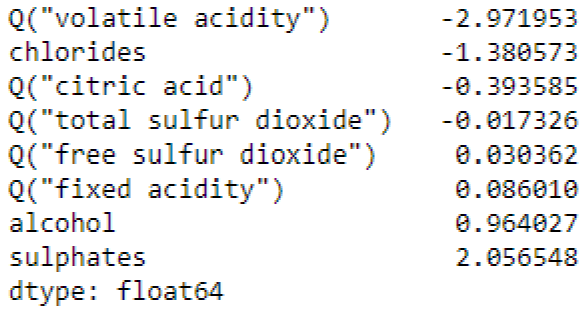

alcohol
0
2
4
6
8
10
12
3
4
5
6
7
8
Quality

***Figure 3*: Comparing the relationship between the predictor variables and the wine quality**



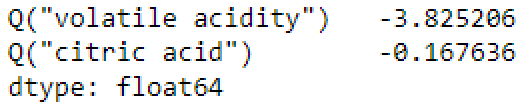
```
Q("volatile acidity")       -3.185167
pH                          -1.702287
chlorides                   -1.082182
density                     -1.043144
Q("citric acid")            -0.737659
Q("fixed acidity")          -0.027132
Q("total sulfur dioxide")   -0.020082
Q("free sulfur dioxide")     0.028765
Q("residual sugar")          0.119569
alcohol                      0.914453
sulphates                    1.859851
dtype: float64
```

***Table 6*: Weight associated with each predictor variable**



```
Q("volatile acidity")       -2.971953
chlorides                   -1.380573
Q("citric acid")            -0.393585
Q("total sulfur dioxide")   -0.017326
Q("free sulfur dioxide")     0.030362
Q("fixed acidity")           0.086010
alcohol                      0.964027
sulphates                    2.056548
dtype: float64
```

***Table 7*: Weight associated with each predictor variable (without Density, pH, & Residual Sugars)**



```
Q("volatile acidity")   -3.825206
Q("citric acid")        -0.167636
dtype: float64
```

***Table 8*: Weight for only Volatile Acidity and Citric Acid**