USING MACHINE LEARNING TO DETERMINE QUALITY OF RED WINE

IBM DATA SCIENCE CAPSTONE PROJECT

PREPARED BY
KATHRYN DARK
SEPTEMBER 2020

Section 1: Introduction and Business Problem

The purpose of this report is to predict the quality of red wine based on measures of eleven different variables in the wine. The target audience for this report is Vineyards, Winemakers and Restaurants. This report will determine which variables of wine have a stronger impact on its quality rating. This report will also use Machine Learning Algorithms to predict whether a wine will be rated 'good' (a quality rating 7 and above) or 'bad' (quality rating less than 7) based on its feature variables. With this report, a vineyard can adjust different variables of its wine production which can increase their quality of wine. An increase in their quality of wine will lead to more sales and profit. Restaurants can use this machine learning algorithm to determine if a wine will have a good quality rating or a bad quality rating. They can then pick the best wines to sell to their guests.

Section 2: Data

The dataset used for this project revolves around red wine. This dataset is related to red variants of the Portuguese "Vinho Verde" wine.

The data includes eleven variables that can be tested in the wine:

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3 - citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
4 - residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter
5 - chlorides: the amount of salt in the wine
6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion;
7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine
8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content
9 - 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
10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial
11 - alcohol.

The data also includes a 12th column, which is Quality, a score between 0 and 10. The purpose of this report is to determine if the quality of a wine can be predicted based on the eleven qualities listed above.

Relevant publication: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties.In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Figure 1. Head of Dataset

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Section 3: Methodology

Exploratory Data Analysis

Exploratory data analysis shows how many rows (different kinds of wines) are in the dataset, what type of data we have, and if we have any missing values.

Figure 2. Exploratory Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
fixed acidity           1599 non-null float64
volatile acidity        1599 non-null float64
citric acid             1599 non-null float64
residual sugar          1599 non-null float64
chlorides               1599 non-null float64
free sulfur dioxide     1599 non-null float64
total sulfur dioxide    1599 non-null float64
density                 1599 non-null float64
pH                      1599 non-null float64
sulphates               1599 non-null float64
alcohol                 1599 non-null float64
quality                 1599 non-null int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Figure 2 shows that every column except the quality column is of type float64. The quality column is of type int64 since the quality is a number given from 0 to 10. This figure also shows that there are 1599 entries and all 12 columns have 1599 non-null entries, which means the dataframe is not missing any values. The dataset does not require any cleaning since all values are present.

Determining Correlation

In this next section, two different methods were used to quickly see if any of the feature measures of wine determine the quality rating. First, the different attributes of wine and their correlation with quality was ordered from highest to lowest. Attributes with a correlation coefficient close to +1 and -1 show the strongest correlation.

Figure 3. Correlation of Eleven Different Wine Features with Quality Rating

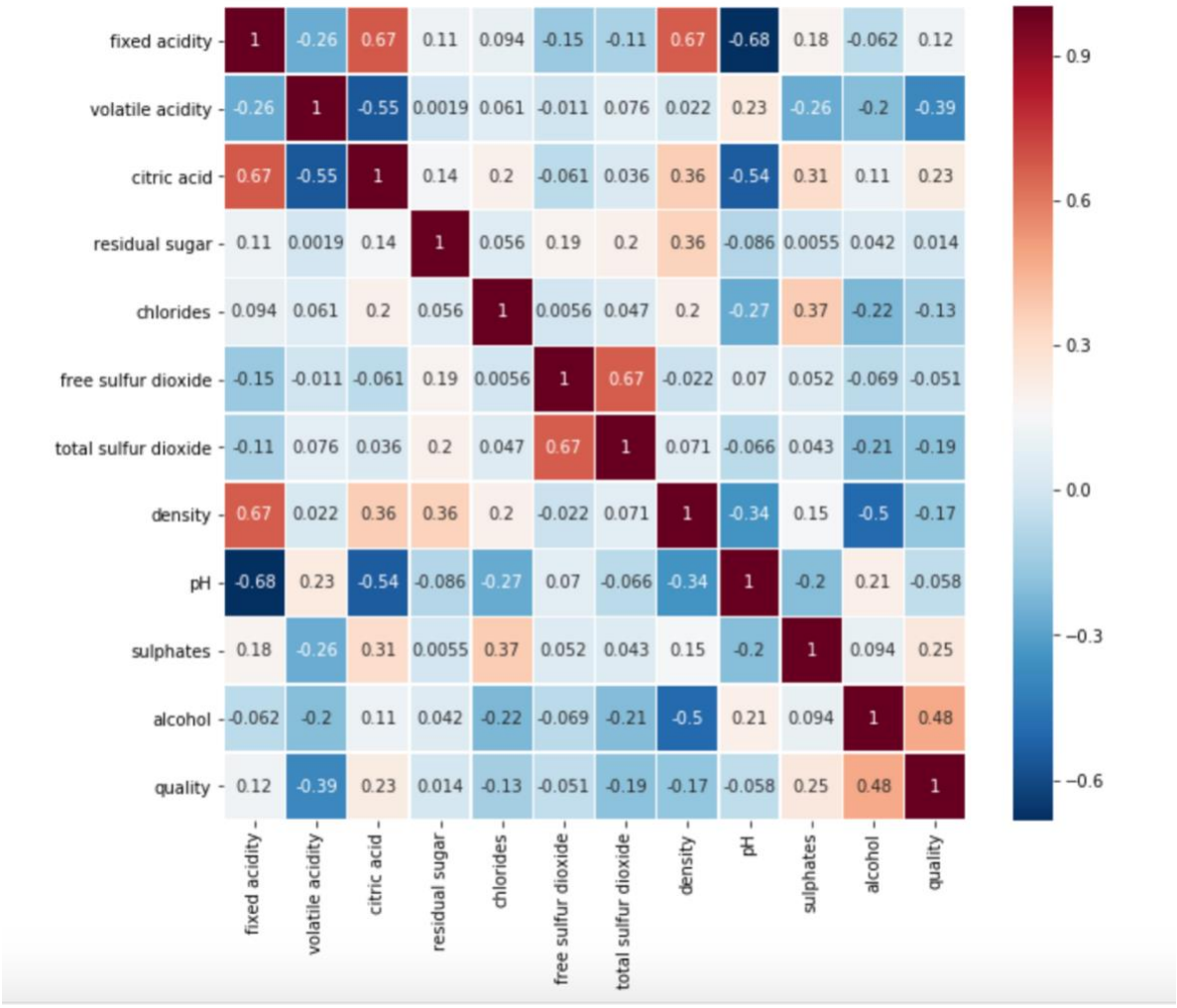
```
alcohol                0.476166
sulphates              0.251397
citric acid            0.226373
fixed acidity          0.124052
residual sugar         0.013732
free sulfur dioxide   -0.050656
pH                    -0.057731
chlorides             -0.128907
density               -0.174919
total sulfur dioxide  -0.185100
volatile acidity      -0.390558
Name: quality, dtype: float64
```

In Figure 3 Alcohol, Volatile Acidity, Sulphates and Citric Acid have the strongest correlation coefficients to determine the quality rating of the wine.

Data Visualization: Heatmap

The second method used to determine correlation was visual: a heatmap.

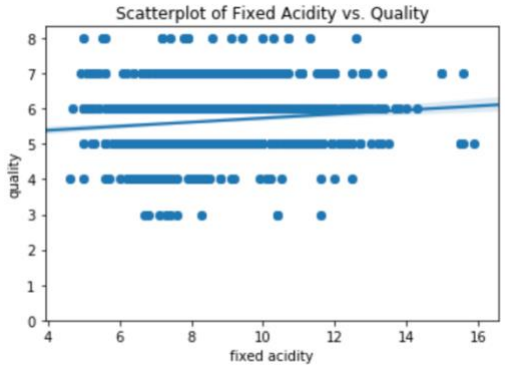
Figure 4. Heatmap Showing Correlation of Eleven Different Wine Features with Quality

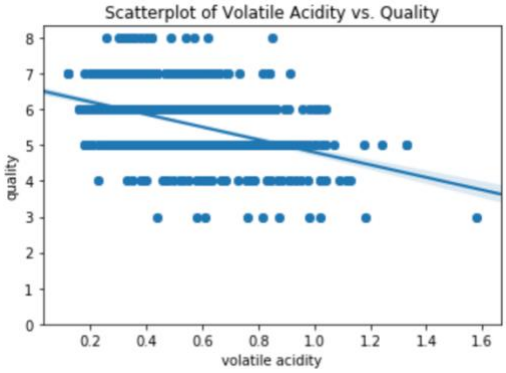


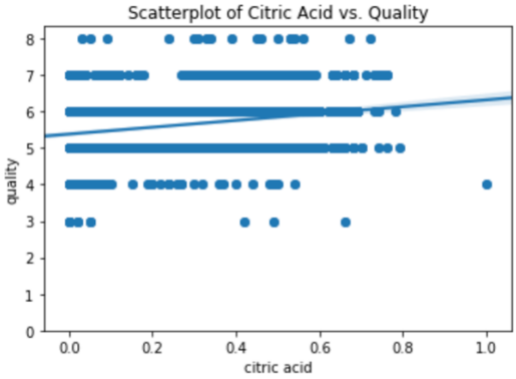
A heatmap is a statistical metric to measure to what extent different variables are interdependent. Here the boxes with darker red and darker blue have the strongest correlations.

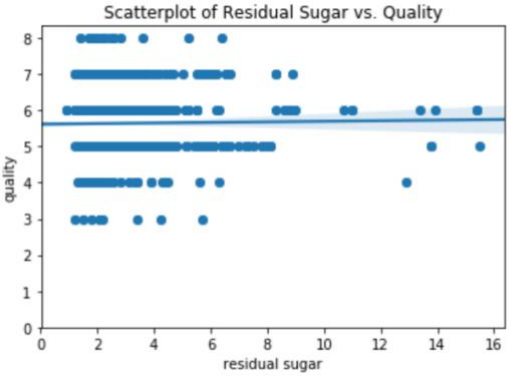
Data Visualization: Scatterplots

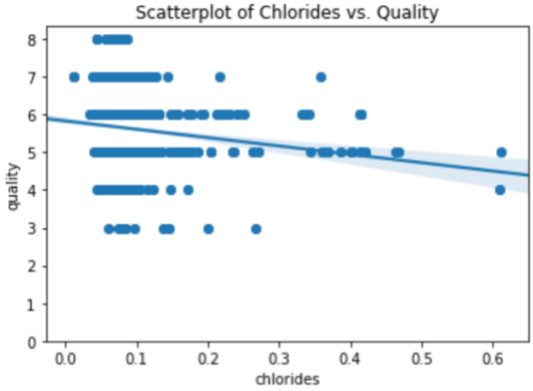
Next each measured feature variable was plotted against quality to determine if there are variables that are more strongly correlated to quality. Scatter plots are used to begin. The predictor or independent variable will be on the x-axis, and the target or dependent variable (quality) will be on the y-axis.

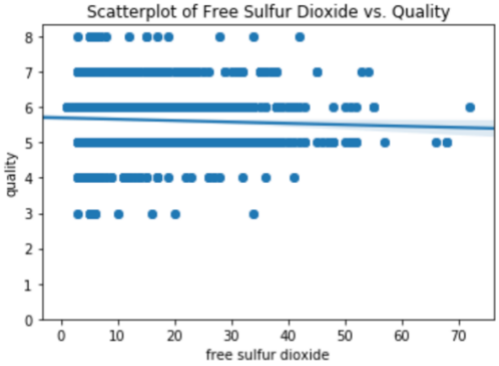


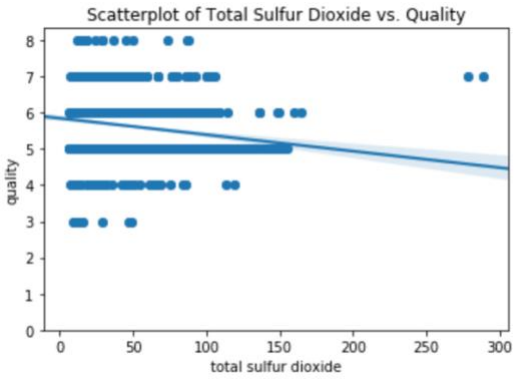


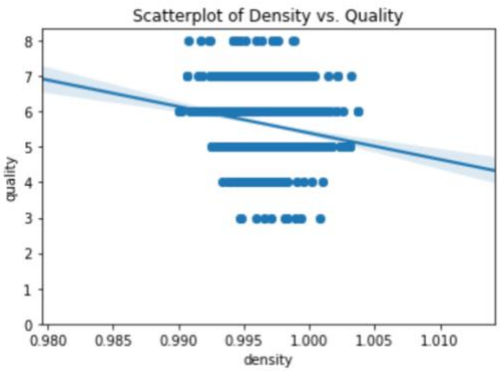


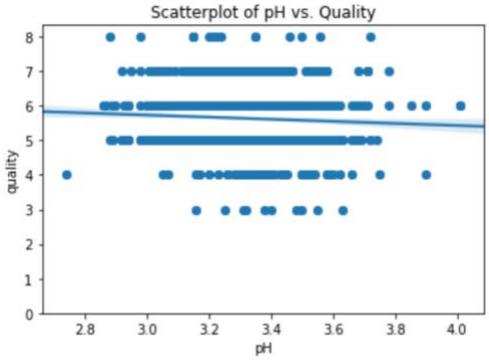


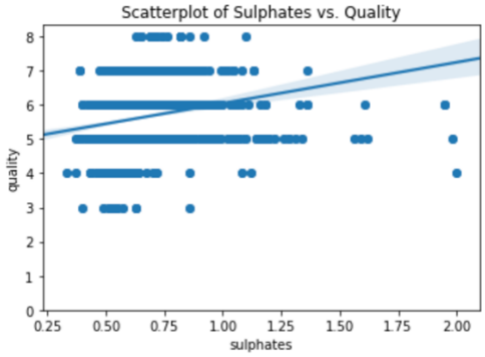


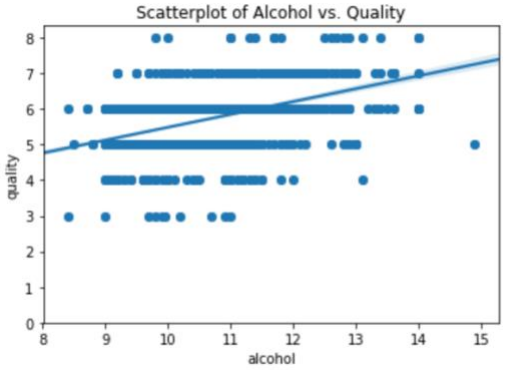












From these scatterplots, a visual idea of which features of wine have a stronger correlation with the quality rating is present. One can also see if the features have a positive linear relationship or a negative linear relationship based on the slope of the line of best fit. The R-squared of the line of best fit can also be determined. The R-squared value is a measure to determine how close the data is to the line of best fit. See figure 5 for equations of each line of best fit as well as the R-squared value.

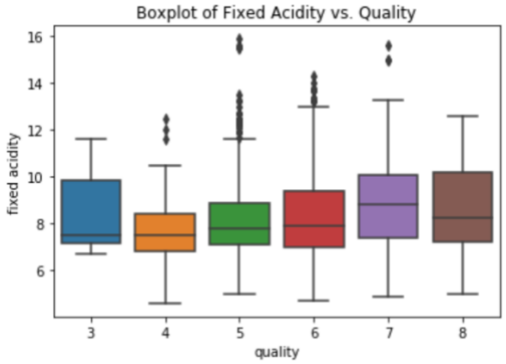
Figure 5. Table of Line of Best Fit and R-Squared Value for Each Feature of Wine

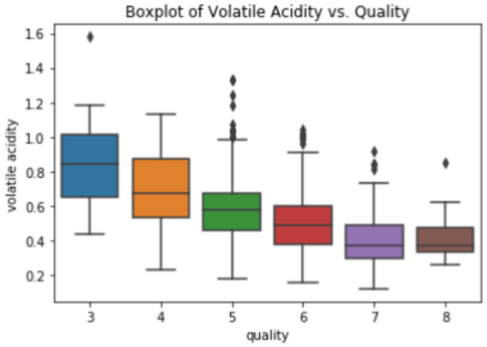
Feature Variable	Line of Best Fit	R-Squared
Fixed Acidity	y=0.057(x) + 5.15	0.015388811647710552
Volatile Acidity	y=-1.76x + 6.56	0.1525353797247485
Citric Acid	y=0.938x + 5.38	0.05124451523867202
Residual Sugar	y=0.0078x + 5.616	0.00018855786403915165
Chlorides	y=-2.21x + 5.829	0.0166169011930003
Free Sulphur Dioxide	y=-3.91e-03x + 5.69	0.00256603613553541
Total Sulphur Dioxide	y=-4.544e-03x + 5.847	0.03426211696068793
Density	y=-74.846x + 80.238	0.030596736248323153
pH	y=-0.302x + 6.636.	0.0033329135305089075
Sulphates	y=1.197x + 4.847	0.06320049136455663
Alcohol	y=0.36x + 1.874	0.22673436811275482

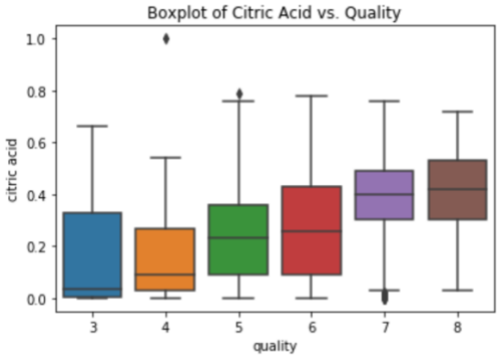
An R-squared value closer to 1 indicates that the line is a good fit for the data. Here the R-squared values are quite small, and the closest value to 1 is for the feature alcohol vs. quality, with an r-squared value of 0.2267. The next highest feature, Volatile Acidity, has an R-squared value of 0.1525. These two features have the strongest correlation to quality rating as well as the strongest R-squared values for predicting quality of wine.

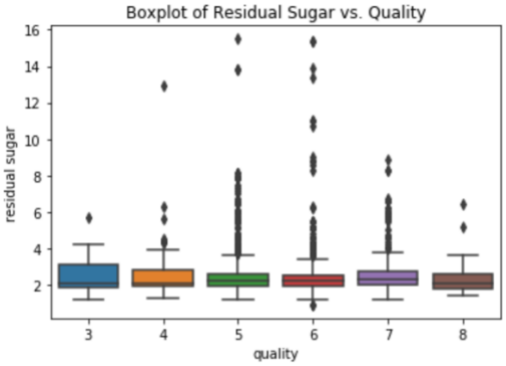
Data Visualization: Box Plots

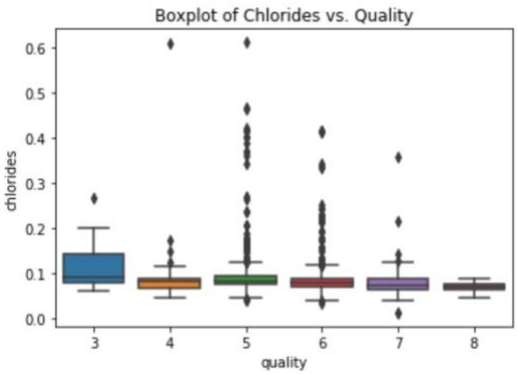
The second method for Data Visualization used in this project was box plots. Box plots can show any outliers in the data and help understand trends in correlation. Box plots statistically represent the distribution of the data and contain many important points in the data including the minimum, the first quartile (25% through data: a quarter of the data points are less than this value), the median, the third quartile (75% through data: three quarters are less than this value), and the maximum. Outliers lay outside the boxed area.

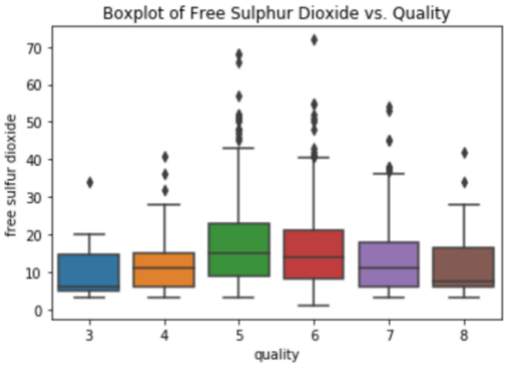


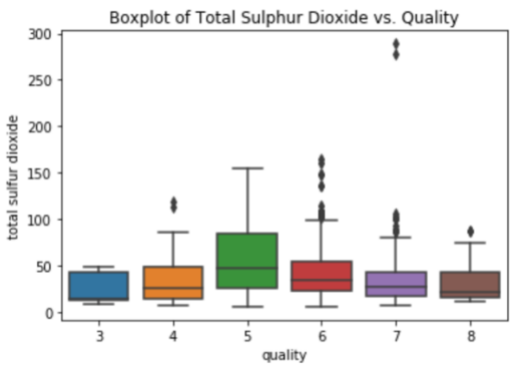


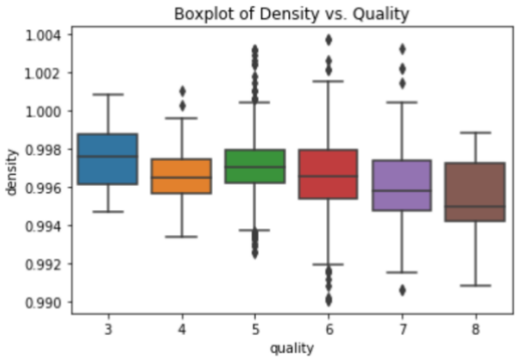


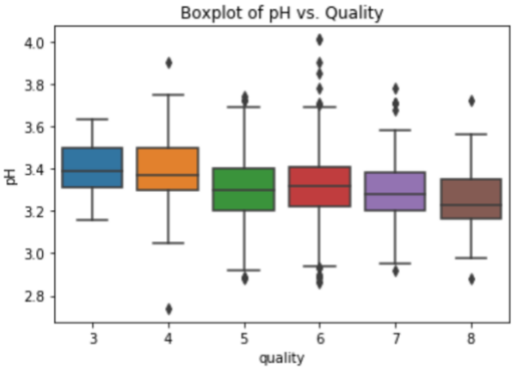


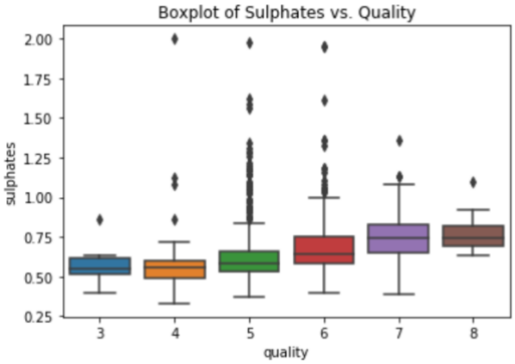


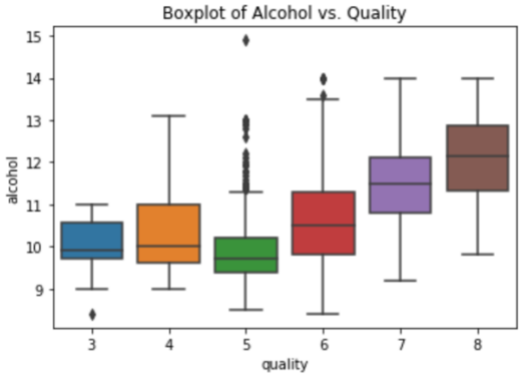












Boxplots are a great visual tool because they show the outliers for each feature metric. Here it can be seen that Residual Sugar, Chlorides and Sulphates have many outliers and these might be affecting the results. Boxplots also make trends easer to visualize. For example, it can be seen in the boxplot, Volatile Acidity vs. Quality, that as the quality rating increases, the volatile acidity level decreases. Also, the Alcohol vs. Quality boxplot shows that the quality rating increases as the alcohol level also increases. A positive correlation can be seen in the boxplot Sulphates vs. Quality and Citric Acid vs. Quality. A negative correlation can be seen in the boxplots Density vs. Quality and pH vs. Quality. A negative correlation does not mean the two features are not correlated, it means that as the quality rating increases, the feature level decreases.

Preparing the Data for Machine Learning

In order to use the data on machine learning algorithms, it first had to be prepared. First, the data was separated into two bins based on the quality rating. Wines rated 3-6 were considered "bad" while wines rated 7-8 were considered "good". After the data was sorted into these bins, the labels were changed from "bad" and "good" to 0 and 1, respectively.

Figure 6. Sum of Data in Two Bins: 0 and 1

```
0    1382
1     217
Name: quality, dtype: int64
```

There are 1382 wines with a quality rating of 'bad' and 217 wines with a quality rating of 'good'. This unbalanced data will be fixed by scaling in a few steps.

Next, the data was separated into a feature set and a target set. The feature set included all columns except quality and the target set included only the quality column. After, the data was split into training and testing using train_test_split from sklearn. The testing size was 20% and the training size was 80%.

Finally, the unbalanced aspect of the dataset is addressed by scaling. This dataset has a significant amount more wines considered 'bad' than 'good', so StandardScaler was used to scale and transform the data.

Machine Learning Algorithms

Next two different machine learning algorithms were tested and their accuracy level was determined. These algorithms could be used to predict a quality rating for a new wine given all the features of the current dataset. These algorithms would be able to predict if the new wine receives a quality rating of 0 or 1: 'bad' or 'good'.

Two different machine learning algorithms were used on this dataset. First, K-Nearest Neighbor (KNN). This machine learning algorithm is a classification algorithm and classifies cases based on their similarity to other cases. Cases that are near each other are said to be neighbors. The quality score of a new wine could be predicted based on the quality score of wines with similar feature values.

The algorithm started with a k=4, and produced an accuracy of 87.8%. Then, it was determined that the best value for k was k=1, which produced an accuracy of 89.1%.

The second machine learning algorithm used for this dataset was logistic regression. Logistic regression is another classification algorithm for categorical variables and uses one or more

independent variables to predict the outcome or dependent variable. In this case 11 feature variables were use to predict the outcome of the dependent variable, the target, quality. Since the target variable was binary: 0 or 1, logistic regression is a good option. The training and testing data were modeled and tested and produced a Jaccard index score of 87.5%.

Figure 7. Machine Learning Algorithms and their Accuracy

Machine Learning Algorithm	Accuracy
K-Nearest Neighbor	89.1%
Logistic Regression	87.5%

Section 4: Results

The goal of this report is to determine if a machine learning algorithm can be used in order to determine the quality rating of red wine given the eleven feature points. With an 89.1% accuracy, the K-Nearest Neighbor machine learning algorithm can determine if a red wine will have a quality rating of 0 'bad' or 1 'good', given the eleven feature points.

It was also determined that Alcohol (correlation coefficient 0.47), Volatile Acidity (correlation coefficient -0.39), Sulphates (correlation coefficient 0.25) and Citric Acid (correlation coefficient 0.23) have the strongest correlation coefficients to determine the quality rating of the wine. That means these features have the most impact on whether or not a wine is rated 0 'bad' or 1 'good'. Winemakers and Vineyards should focus on these four features as they have the strongest impact on the quality rating of red wine.

From the data visualization tools one could see that volatile acidity level decreases as the quality rating increases. Also the Alcohol vs. Quality boxplot shows that the quality rating increases as the alcohol level also increases. A positive correlation was evident in the boxplot Sulphates vs. Quality and Citric Acid vs. Quality. A positive correlation means that as the independent feature or variable increases, so does the target feature. A negative correlation was evident in the boxplots Density vs. Quality and pH vs. Quality. A negative correlation does not mean the two features are not correlated, it means that as the quality rating increases, the feature level decreases.

In the case of this dataset, the boxplots were more valuable visual tools than the scatterplots because of the number of outliers. The boxplots visually showed the immense number of outliers in the residual sugars, chlorides and sulphates vs. quality graphs. These outliers are not immediately visually apparent in the scatterplots and affect the line of best fit as well as the R-squared value.

The scatterplots however were useful for finding a line of best fit and a corresponding R-squared value. An R-squared value closer to 1 indicates that the line is a good fit for the data. Here the R-squared values are quite small, and the closest value to 1 for the line of best fit for the feature alcohol vs. quality, with an r-squared value of 0.2267. The next highest R-squared feature, Volatile

Acidity, has an R-squared value of 0.1525. These two features have the strongest correlation to quality rating as well as the strongest R-squared values for predicting quality of wine.

Section 5: Discussion

Based on the data analysis, it is recommended to use the K-Nearest Neighbor Algorithm to determine if a new wine will have a quality rating of 0 'bad' or 1 'good'. The K-Nearest Neighbor Algorithm gives a result with 89.1% accuracy.

Also, the four feature measures with the strongest impact on the quality rating of red wine were alcohol, volatile acidity, sulphates and citric acid. All except volatile acidity had a positive correlation, meaning the quality rating was increased with an increase in those features. The quality rating would increase with a decrease in volatile acidity. Winemakers and vineyards should pay extra attention to these four features in order to increase their quality rating of red wine.

Section 6: Conclusion

In this analysis, the data was cleaned to determine correlation of eleven feature variables with the quality rating of red wine. Then three data visualization tools; a heatmap, scatterplots and boxplots; were used to visualize the correlation of the eleven feature variables with red wine quality. Lines of best fit and R-squared values were determined from the scatterplots. The data was then prepped for machine learning by binning the target variable to 0: any quality rating below 7, and 1: any quality rating 7 and above. The data was unbalanced (much more data in the 0 'bad' category), so scaling was applied. The data was separated into feature set (every column except quality) and target set (only the quality column). The data was split into training and test sets, with a testing size of 20% and a training size of 80%. Two machine learning algorithms were modeled, trained and tested for accuracy on the data: K-Nearest Neighbor and Logistic Regression.

In conclusion, four features of wine have the strongest impact on its quality rating: alcohol, volatile acidity, sulphates and citric acid. The K-Nearest Neighbors machine learning algorithm should be used to determine if a new wine will have a good rating (7 and above) or a bad rating (below 7), with a 89.1% accuracy.