**Proposal Wine Company: Iteration 4**

INFOSYS 722   
Colby Carrillo 984745460  
Github: <https://github.com/ColbyCarrillo/Iteration4>

# Business and/or Situation Understanding

### 1.1:

The company has hired wine experts to evaluate the quality of their wines. They hope to find out which of the different attributes collected contribute to the level of quality so they can make better more successful wine.

With the data collected by the company from the wine expert’s evaluation, the company would like to find out which inputs to the wine determine the overall quality of the wine (Parmar, R., 2018 / P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis). They hope to leverage these insights to increase sales and quality of their product.

They expect to have an increase in quality of their wines, increase in sales, also potential reduction in supply chain costs by determining which components of the wine are most useful.

Our success criteria are to find attributes that contribute the most and the least to the overall quality of a wine data collected. The ability to classify the type of wine, red or white, based off inputs with at least 80% accuracy. Also, to determine which attributes contribute to higher levels of alcohol contents as they plan to release a new product based on its high level of alcohol.

### 1.2:

Accessing the situation, we see that the company has funded research evaluating several thousand wines, both red and white, by wine experts to determine its overall quality.

Some of the biggest risks are the improper use of data mining techniques could lead to poor business decisions and potential loss of money. If the attributes that determine the quality of the wine in the dataset do not hold for the whole population, we could potentially lead to loss resources creating new products.

Our contingency plan is to use best practices in our data mining techniques. There will also be a business review and more discussions to mitigate the risk of new product creation based off our insights. Luckily, we have all the resources required to complete this data mining project.

Our team has several assumptions we are making to complete this data mining project. We assume that the experts gave accurate analysis of the wine and there is no malicious intent. Our sensors that collected the data for each wine are accurate and none of the data is corrupt. Some of our constraints working with this data involves not having any data on the grapes used to make the wine. Data such as weather, geographic location, soil quality, etc. could be very beneficial to the analysis of the determinants of the quality of wine. As well we only have just over six thousand data points with white wine data points roughly three times larger in size than red.

Some of the risks involved could potentially be the financial downside for the company if our evaluation of the sample wine does not correctly depict the population. There is also some level of bias involved in the expert’s opinions on what makes the wine quality increase and decrease. These biases could not reflect the population of our consumers. The company will decide what actions to take with data mining our evaluation and make sure to mitigate risk as much as possible.

### 1.3:

The data mining problem is one of prediction, classification, and potentially clustering. We plan to answer multiple data mining questions in our evaluation of the data and report back to the business.

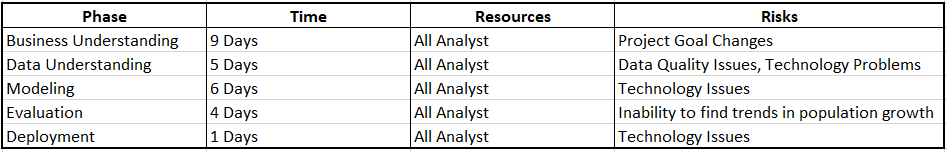
While our technical goals are to complete the project before the specified due date given by the course. The validity of our findings should last for an unknown amount of time given that these variables are useful in predicting the quality of wine.

Several goals have been drawn up to evaluate the success using concrete levels. One goal is to be able to classify the type of wine, red or white, with at least 90% accuracy. Evaluate what variables are important and not important in the overall quality of the wine. Predict the quality of the wine with 80% accuracy. Evaluate what variables are important and not important in the level of alcohol content in the wine. Predict the alcohol content of the wine with 80% accuracy.

For our data mining goal to be successful we will evaluate all the models using accuracy. This accuracy of predictions will be done by one of several techniques Root Mean Squared Error, Mean Squared Error, Mean Absolute Error, R2, etc. While the accuracy of the classification will be a normal percentage of correct compared to total predicted.

### 1.4:

Below is our project plan for the project. The time frame is from a start of when the project was assigned to the due date. All parts of the project have been completed in full and as specified.



# Data Understanding

### 2.1:

The data was collected from Kaggle, which is a known repository of data sets, at the location (Parmar, R., 2018 / P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis). It was collected by downloading all the csv files, which were broken into and white, included in the Wine Dataset project. There were no issues encountered when collecting the data as the two datasets were self-explanatory and were complete.

After collecting the initial data, our team thinks that all variables will be valuable to insights into the quality and predicting the color of the wine (Parmar, R., 2018 / P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis). It is to note that we will use the quality to determine the color and alcohol levels but not the color to determine the quality or alcohol levels of the wines. This is due to our research goal of finding the attributes that determine quality wine and should be independent of color.

We do not believe any variables are irrelevant and all will be used in our analysis. We do believe with just under 6500 datapoints in total, and over 1500 in each wine type, that there is enough data to draw generalizable conclusions. There are no concerns about there being too many variables, but there could be some key missing ingredients variables to explain the quality of the wine.

We are merging two different data sheets, one being white wines and the other being red. There are no concerns of problems when merging as both sheets contain the same variables, so we will just be combining the rows together. We do not have to work with any missing values, so this is not of concern with this dataset.

### 2.2:

Our team completed a basic data overview with the following outcomes. We used pyspark package, in Jupyter Notebooks hosted on AWS, for our analysis with and found that there are 6497 observations, 13 variables (two being target variables: output and color) and found that there are no missing values in any of the variables. The value types are shown in Figure 1 below. We were required to specify the input data types, using a schema that we provided, making all variables either double or integer, depending whether they contained decimals.

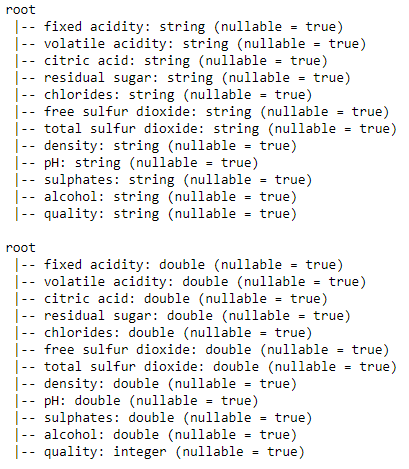


Figure : Variable Data Types

The data collected is in the format of two csv files which we have combined into a data frame after bringing it into pyspark. We looked at some basic statistics, which include count, min/max, mean, standard deviation, and quartiles shown in Figure 2 below. One interesting variable that is sticking out is that residual sugars mean, and standard deviation are almost the same. This means that the data is very wide spread out and we have a lot of variability in the data. We will have to investigate this further with distribution graphs such as box plots but will keep our eye on it for potential issues. Note that the color of the wine was is included but has null for mean and standard deviation because it is a categorical variable..

Each of the eleven double variables represent a different numeric value attribute associated with the wine. In example, the residual sugar represents the amount of sugar found in the specific wine being tested, while density and pH attribute to the density and pH values of the same wine.

### 

Figure : Basic Statistics

### 2.3:

While exploring the data we derived a few hypotheses of what is to be expected in our future analysis. We believe that the variables with the highest variability in the data may be the most insightful in the quality of the wine. These are variables such as residual sugars, fixed acidity, chlorides, sulphates, and volatile acidity. Though, we cannot be certain by looking at the basic statistics of the data which will contribute most to the quality, alcohol content, or color.

All attributes seem promising in our analysis. As we have deduced that variability will be an indicator of variable importance that looking at the basic statistics have given us the variables with the highest variability.

After exploring our data our goals remain the same, in that we have not found anything concerning in our data exploration which is covered in depth in section2.ipynb files. We have seen that, given our assumptions hold on the data, the data will potentially provide use in answering our data mining questions.

We see by the below diagrams, Figures 3 and 4, the distribution of both the quality of wines and color of wines.

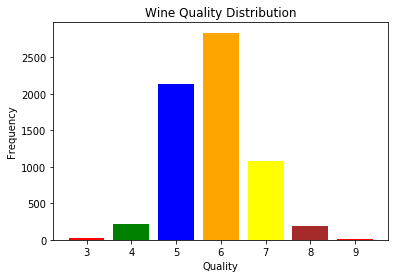


Figure : Quality Distribution

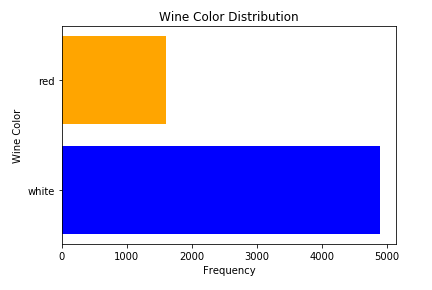
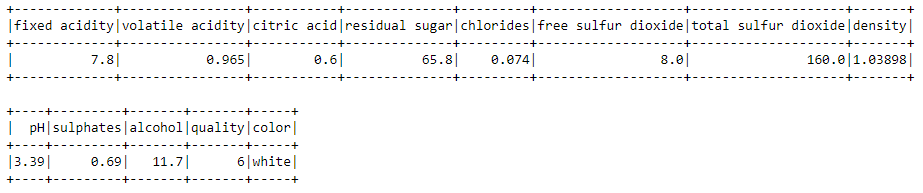


Figure : Color Distribution

### 2.4:

Our initial data quality inspection shows that we do not have any missing data (Figure 6), data errors, coding inconsistencies, or bad metadata. There could be a potential measurement error on one observation, which we will keep in mind in further analysis, but we do not believe that this is true currently. This observation is number 4380 which is showing a very large value for residual sugar in the wine. We believe this is a viable value for the content of sugar but wanted to address the issue with the distribution and values shown in Figure 5.



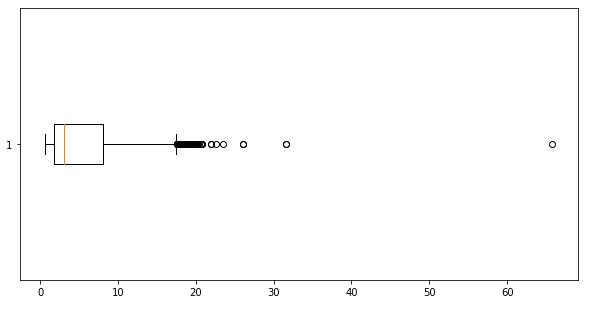


Figure : Potential Measurement Error

We see that there are no missing attributes (Figure 7), blank fields, or spelling inconsistencies that will cause issues in our future data analysis.

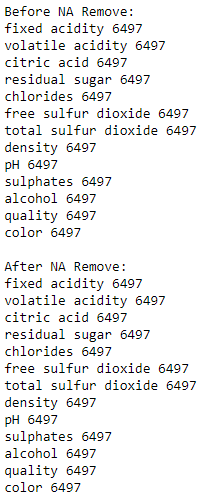


Figure : Null or Missing Values

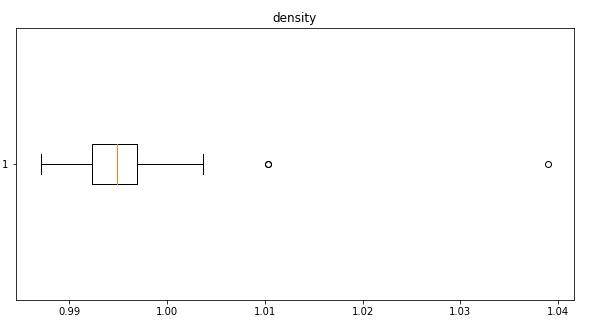


Figure : Density Distribution

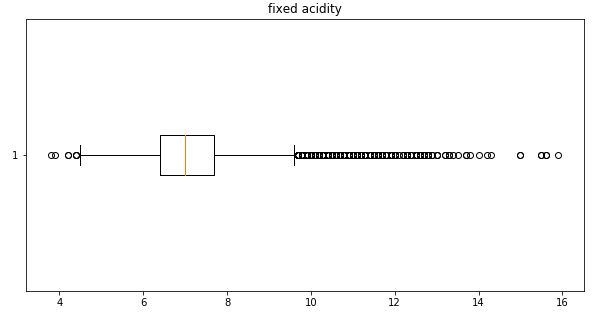


Figure : Fixed Acidity Distribution

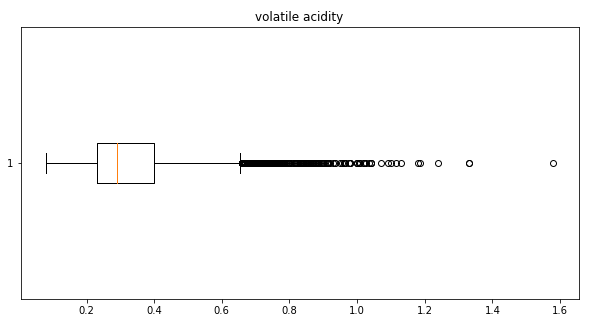


Figure : Volatile Acidity Distribution

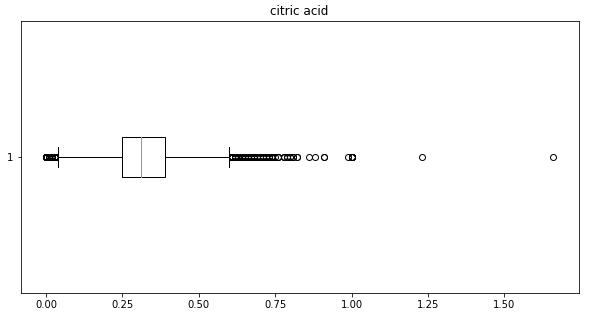


Figure : Citric Acid Distribution

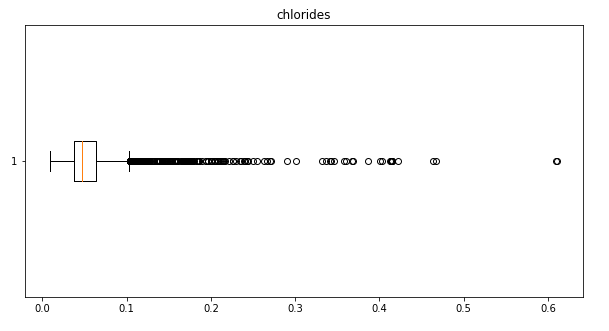


Figure : Chlorides Distribution

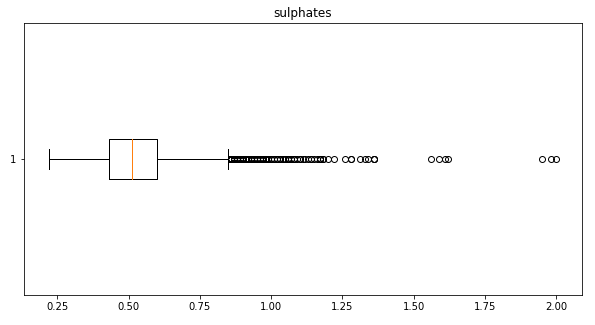


Figure : Sulphates Distribution

Above are a few diagrams of the distributions of some of the variables, Figures 7-12, that are not ideally normal. By doing tests of six standard deviations, as referenced in [9] we see that there are only a few outliers in the data (Figure 13). We choose not to modify the data in any way, as these distributions could be normal for wines and we do not want to overfit any future models. We assume that the measurements collected are accurate and there were no issues or bias with any devices or personnel in its collection. We will not be excluding any data as we find that all of it may be beneficial to our analysis.

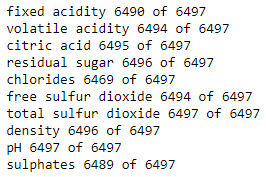


Figure : Potential Outliers

# Data Preparation

### 3.1:

When selecting our data there will be two different subsets to consider:

* First, the dataset including all rows will be used to predict/evaluate the quality of wine, as well as predict/evaluate the color of the wine and alcohol content.
* Second, will be the subset of white and red wine which models will be applied to separately to see if there are difference by the color of the wine.

We will choose to exclude the color attribute of the wine when determining the quality and alcohol content of the wine, code shown in Figure 13. We do this as we want to fit a model that will be able to predict quality regardless of the color. All other variables are valid and will be used in our analysis of the data.

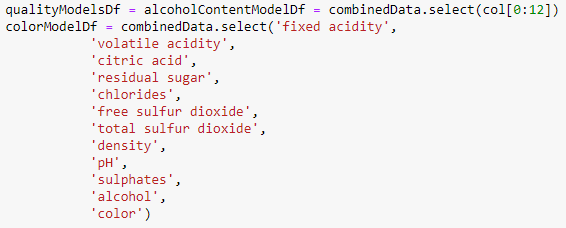


Figure : Sub setting Data Frames for Two Models

For data cleaning purposes, I have created two other data files that have had some issues with the data collection process. First there are some missing values in the Fixed Acidity and some data entry errors of the Residual Sugars. I will go through the cleaning process below on how we would handle such an issue in python.

### 3.2:

We have created a box plot of every variable, just to verify that there is nothing of concern, some of these are shown above in Figures 7-12. Even though these box plots show some potential outliers, they still do not pose any threat to our analysis and all will be kept in the dataset. If we were to take the route of removing all values with outliers, 576 in total some of which probably overlap, we would run the issue of overfitting our models to non-realistic situation. If interested in the box plots that were not included in our report, you can reference the section3.ipynb Jupyter Notebooks which hold all the code and outputs.

Fortunately, after reviewing the data we will not be required to clean any of the data. We do this as we already have a limited number of values and none of these values seem unreasonable, as far as our knowledge. There are no missing values (Figure 6) and all outliers will be included in the dataset.

In our second (dirty) dataset, we have run all diagnostics on the dataset as if this was our first time analyzing the data. We first check for null values and find that there are 91 records which contain null values. We then subset the data frame to a new variable removing the null values shown in Figure 15. We verify that the new data frame now has 6,406 values compared to the 6,497 at the start.

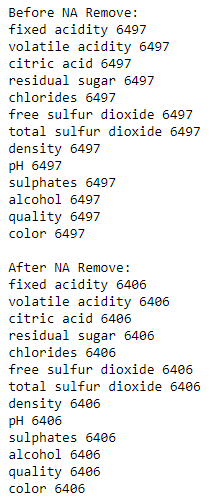


Figure : Checking Dataset for Null Values

Now that the nulls are removed, we proceed to check if there are any concerning with the basic statistics and distribution of our variables, Figure 17. One thing immediately catches my eye and that is a max value of 999, and standard deviation much larger than the mean, which seems like a data entry or measurement error. We further investigate this by looking at the visual distribution of the data using a box plot, Figure 18.

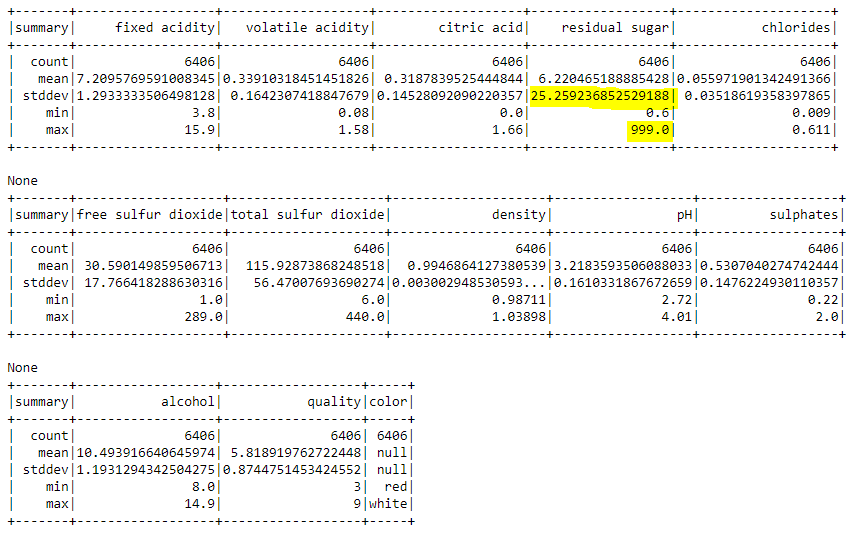


Figure : Dirty Data Summary Statistics

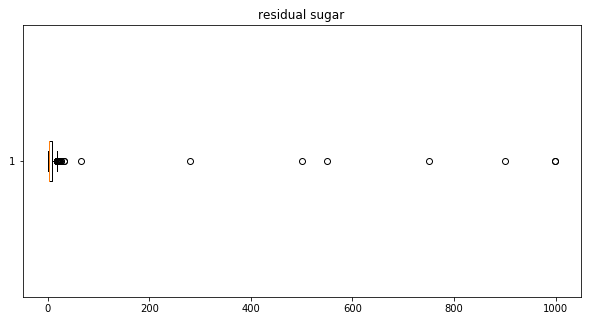


Figure : Residual Sugar (Dirty) Box Plot

Clearly there is something wrong with this variable with very large outliers which seeming to start after the value of approximately 200. We know this as almost all the box plot is squeezed to the point of almost non-visibility. Also, the value of 1000 sugar is not feasible amount of sugar to be in wine. We will go ahead and remove these observations from the data frame shown in Figure 19. To verify our issue is gone, we then plot the distribution again shown in Figure 20 below. All other variables distribution seems to be okay. If a reader would like to dig deeper into the code it can be found in the section3.ipynb Jupyter Notebook.

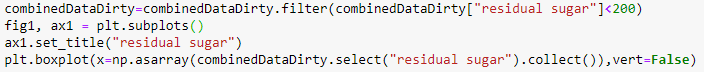


Figure : Removing Residual Sugar Large Outliers

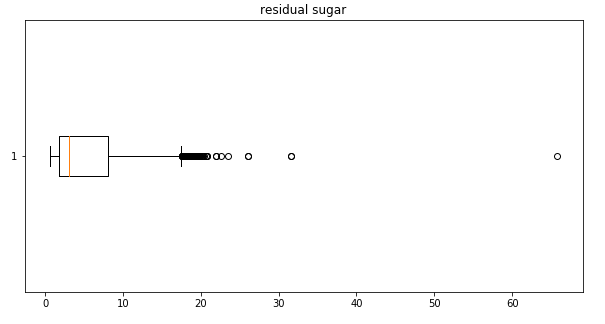


Figure : Fixed Residual Sugar (Dirty) Distribution

We have decided not to impute and values of the outliers or null values as this could add undesired bias, specifically if imputed around the mean, into the data. The best option is to simply remove the data as there was only a total of 98 data points with issues. If we found that a larger proportion of the data was impacted, we would have considered dropping a variable or imputing the values.

### 3.3-3.4:

For our analysis we were provided with two csv files each containing a specific color of wine. Our team decided to construct new data by add a column to each imported data frame for the color of the wine, making them distinguishable when combining the two datasets. We did this as it allows us to include additional hypothesis on the data and further divide our data if necessary. We were then able to integrate our two data sources, since our data has all the same columns, by concatenating or combining the datasets one on top of the other, shown in Figure 20.

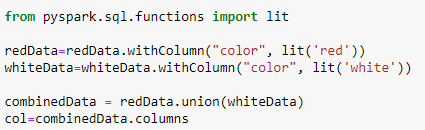


Figure : Adding Column and Combining Datasets

### 3.5:

As seen above in Figure 1, all our variables are either double, integer, or string. Reviewing our target variables: quality, color, and alcohol content we see that each variable is in a data type that will fit will with most of the models supervised models. We will have to fit all the target variables and input variables to feature columns with no column headers to fit models with pyspark modules. This process for all the classification models is shown in Figure 22 and 23 below. It is what is required for the data to be fit with many of the supervised methods we use in our project. We are also required to alter our color variable from string to a binary or integer to fit models with pyspark machine learning to fit Classification models as shown in Figure 21 and 22 below.



Figure : Altering Color to Numeric Format



Figure : Alter Quality to Numeric Format

The string variable, color, and integer variable, quality, can both be seen as categorical and ordinal values respectively. After our one alteration, all input and output variables currently are in the desired state, and if any issues occur it will be addressed later in the paper. If you are interested in viewing how the data was fit for the alcohol level model, please reference the section3.ipynb Jupyter Notebook.

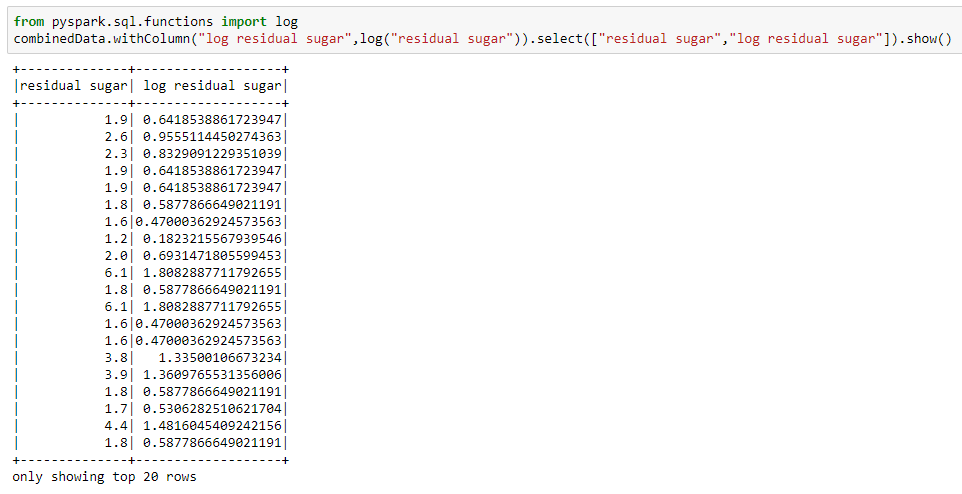
# Data Transformation

### 4.1:

We do not want to reduce any of our data as we already have a limited number of records, 6500. We do reduce/categorize the data into the red and white groupings as specified earlier. This reduction of color attribute is a horizontal reduction that will be used solely in fitting our quality model. If we wanted to do further color, quality, or etc. specific analysis we could select the based on the desired values.

### 4.2:

Due to the multiple dimensions of our data (more than three) it is hard to visualize all the data at once, as humans have a hard time visualizing above three dimensions. The only variable that potentially needs transformation would be Residual Sugars as the data seems to be slightly right skewed. I have logged the variable and shown the new box plot in Figure 21. This seems to have fixed any concerns of right skew, but I do not believe this is necessary and will not be used in model assessment. The distributions of the different variables are covered earlier in section three for reference.



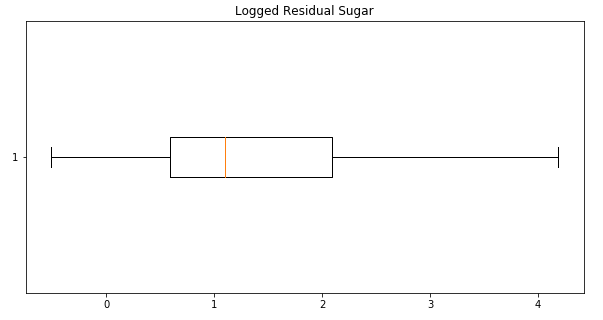


Figure : Logged Residual Sugar and Box Plot

# Data-mining method(s) selection

### 5.1:

We were provided by the business the following goals referenced earlier in the paper:

1. Find attributes that contribute the most and the least to the overall quality of a wine.
2. Ability to classify the color of wine, red or white.
3. Determine which attributes contribute to higher levels of alcohol content.

These three goals have a similar success criterion for classification, or prediction, of model accuracy of at least 80%. That being said, if a model is not found to have high enough accuracy, it can still be used for explanation, or description, of variables that go into determining the target variable. That is, according to (Fayyad,, U., Piatetsky-Shapiro,, G., & Smyth, P. (1996).), prediction which involves using input variables to determine future values of a target variable, and description which focuses on finding interoperability of the variables to describe the data. Each of these can be found using several data-mining methods such as: Classification, Regression, Clustering, Summarization, Dependency Modeling, and Change and Deviation Detection.

For our first criteria, it will be a prediction and description as the method goal to not only predict the quality but find which input variables determine the quality. To accomplish both tasks we will mainly explore classification methods as the quality of the wine is an ordinal variable. We are unable to use popular regression techniques as these quality values are an ordinal grouping, in example, it can only be in one of the groupings, one through ten, and not between. We have decided not to attempt Clustering as it has not proven to be successful for this goal in previous iterations.

Our second criteria will again be one of prediction and description. For both tasks it will be a straight forward classification problem as there are only two groups to classify between, red and white. We will however be able to dive into fit some regression method techniques, as the number of groupings is binary.

Our last criteria will require again, both prediction and description methods. We will use the subcategory of Regression method as the target variable is a continuous numeric value that can take the value of any real number.

The plan for each of these is to fit a few models that were found in our previous iterations in this course. We will fit less but more proven useful models, due to the difficulty of the pyspark tool. I plan to take some reoccurring models found by in previous work for each objective and fit them in pyspark. We will then test with specific evaluation criteria specific to each group of methods, which we will use to choose the best models for each goal.

### 5.2:

As stated previously, we have figured the high-level method for each of the three goals for this project. These include both prediction and description methods such as Classification and Regression which apply to all three of our goals. We choose these due to the target variable that we are using in the model. With Color and Quality both being categorical, only a Classification method fits. While Alcohol Content is a continuous variable, we choose regression as it fits best for this type of target. This does not mean however we are not open to using Clustering techniques, to help in the prediction and description process of each goals if the client desires them, but we have found them not useful in previous work.

Since we have all our variables allocated correctly, as completed in section 3, and don’t have any issues with data quality the assumptions are very limited. One assumption or restriction being that quality variable is a multi-group classification issues, we will be limited with the number of methods we are able to use to only Classification. On top of that, it being ordinal means that even fewer models are justified, but our technique will be to fit it as a categorical, not counting it as ordinal. Another assumption is that the we will use the default settings of the model function unless the model is poorly fitting, and then will try a few different parameters.

# Data-mining algorithm(s) selection

### 6.1:

There are numerous types of algorithms for each of the three different Data-Mining methods we have covered so far. It is useful to note that all algorithms are selected based on the type of variable of the output or target variable. This helps us determine which models can be applied and which can be discarded.

For Classification there are several categories of models that should be addressed, note this is only a subsection of the total number of models out there. The target variable is either a binary (1/0) or a category of limited, preset values. The main categories I want to cover are Tree models, Regression Models, and Neural Networks which are popular methods used for Classification. Tree models, such as CART and Random Forest are not statistical models, but instead just split the models on nodes (attributes) based off some measurement specific to each type of tree, usually impurity or information gain. A more statistic specific models, Regression Models, in example Logistic Regression and Linear regression, will provide estimated coefficients of the effect of the input variables on the output variable. Lastly, Neural Networks, are sometimes “black box” methods, that uses exponential growing amount of weights or lines in-between nodes that help predict the output.

For Clustering, there are several methods used to find clusters within the data, or unseen groupings. This is mainly an unsupervised method of learning and does not have a specific output variable it is trying to predict. Some of the popular methods are KNN (K neared neighbor), which tries to group data into a predefined number of clusters using the K nearest data points in the data set. Another being Hierarchical Cluster which will treat individual data points as their own cluster and proceed to group them together one.

For Numeric, or continuous output or target variables the types of models are similar to classification. The main categories being Regression Models, Tree Models, and Neural Networks, the only difference is that the output is a continuous value and not a prediction of grouping. We will fit a lot of the similar models as stated and explained earlier in the Classification model section.

### 6.2:

For our first run through fitting models to the wine data we have had to select only a few options of the many models available. From supervised techniques, we have chosen to go with Random Forest, Decision Trees, Gradient Boosting Tree, and Multilayer Perceptron (MLP) for each of our three research questions where they are able to be fit. The reader should note that Gradient Boosting Tree in pyspark currently only has the ability to fit binary classification so fitting for the Quality model is not applicable. Also, we have chosen not to fit the MLP for our Color research question as it has proven to not be successful in the past. These models were chosen for four reasons, first they are reoccurring in our previous iteration using another tool SPSS. Second, these are all very predominant machine learning algorithms that we have also learned in our other course COMPSCI 760 on Machine learning. Third, each of these algorithms can be adopted to do both Classification and Regression so we are able to apply the same model on most of our three of our research questions. Lastly, they have the lowest learning curve for adaptation into pyspark, which is a complicated tool.

We partition the data into a training and testing set with an 80% and 20% split respectively for each of our research questions. This split allows for our model to train on a larger amount of data, allowing it to fit better, then a 70 30 split as shown in our lab (Scikit-learn Tutorial: Machine Learning in Python – Dataquest. (2019).) and is what we have used in other courses in our education.

For evaluating and predicting the Quality of the wine, a classification problem, alcohol content of the wine, a continuous variable problem, and color of the wine, a binary classification problem, we fit Random Forest and Decision Trees. These were slightly modified in model in that Quality and Color had to use a Classification package while Alcohol Content had to use a Regression package.

Note that each default model was fit in some cases of our work, if we found that its prediction did not suffice, we then used it as baseline model to compare to other parameter tuned models. This only happened in the case of Quality, where we did a deeper dive into different variations in Random Forest models and which fit best. This will be covered mainly in section 8 of the paper.

### 6.3:

When using pyspark, they provide default settings to parameters for each of the models that are going to be fit. This is done for ease of use so coders can just run the model, and have it work without having to specify loads of parameters they are unsure about. Upon reviewing the specific parameters, most of these are fine for what we are trying to achieve. That is, fit the model and see how it performs on prediction for our use cases.

Once these models were fit, we found that the research question for Quality required much more work as we were only getting predictions topping 66%. Since our previous knowledge of success with Random Forest we decided to only evaluate this model due to constraints on time.

* Random Forest: We attempted to alter the numTrees and maxDepth parameters to the model which fits how many trees and their depth are built in the model. We fitted values of default, 10, 50, and 100 while also altering the maximum depth of the trees which had a max of 30 supported by pyspark. We assumed by having more trees, it allows for more accuracy predictions as it uses a majority voting strategy amongst the “forest”, shown in Figure 24. Surprisingly, what we learned is that having very wide trees was not as useful as having a hybrid of wide and deep trees. The combination of the two variables causes for more accurate prediction better than just altering either of the two alone.

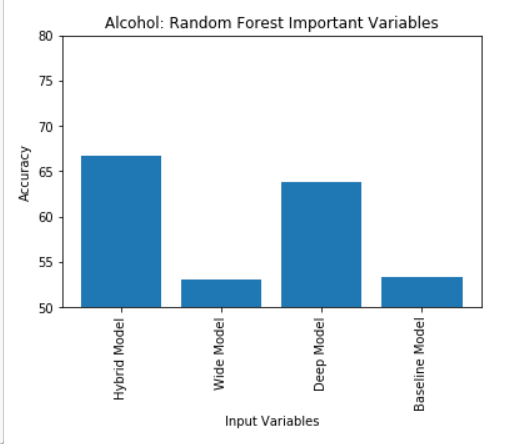


Figure : Random Forest Models vs Baseline

# Data Mining

### 7.1:

We will be partitioning our data into an 80% and 20% split between the training and testing set respectively for each of the three questions. We do this as in previous courses in computer science and statistics when fitting complicated models, it is a good rule of thumb to split in this distinction. Another option would be to add in a validation set but for this analysis I have decided to stick with the basic split. We see in other literature such as our lab we completed in this course (Scikit-learn Tutorial: Machine Learning in Python – Dataquest. (2019).) that they use a 70% and 30% test. As I only have 6500 records, I want more values to go to training the models, so I have altered this split by 10%.

For each of our models we will be fitting evaluation criteria based on prediction and important variables. This output will be provided below in section 7.2 for each respective model. For our Classification models we are able to get an accuracy of their predictions, while for our Regression models we will evaluate the root mean squared error and R2 of the model’s predictions and observed values.

### 7.2:

The below section is split into three sections, one for each of the three questions we are researching in our data mining analysis.

* + 1. Quality Analysis:

We decided to fit three supervised models, Random Forest, Decision Tree, and Multilayer Perceptron. We see the output of the Random forest with 50 trees and a max depth of 20 is the best at prediction with a value of 66.74%. We show below in Figure 25 accuracy of the four models that we fit for random forest. Above this we discuss the different parameters tuned in 6.3 shows our results of finding our best model but further detail of the process will be found in section 8 below covering iteration 2 of this project.

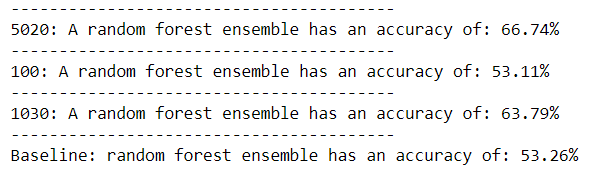


Figure : Quality Random Forest Evaluation

We will now review our other the other, less accurate, model fitted. We see that the decision tree is slightly less affective at prediction than the random forest, given default parameters (52.28% to 54.25%), found in Figure 26, while the MLP is much less useful at predicting wine Quality, with accuracy of 49.22%. The baseline for these models were the default settings of each model, besides Neural Network fitted based on (Singh, D., & Scikit-learn, M. (2019)), and were used for further interpretation of fine tuning the models seen in section 8 below for our best of the two models.

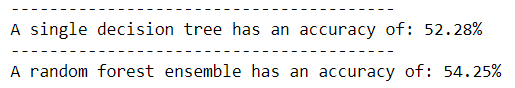
  


Figure : Baseline Random Forest, Decision Tree, and MLP

We find that our prediction rates are not very high for this objective. Some potential reasons being that the input variables into the model are not indicative of the quality of the wine. I assume that if we had data on the grape data that we used to create this wine, we would be able to have a much higher prediction rate.

Another important takeaway to our question on the quality of wine is the important variables in determining the quality. Even though our prediction rate is not high, these variables could be insightful. In Random Forest, as seen in Figure 27 below, Alcohol has the highest influence on the model with the approximate value of .12. All other variables are quite similar in importance with a range of just over .7 to .1. Other potential important variables could be volatile acidity, density, and chlorides. As seen in Figure 28, Alcohol is the main importance variable with a value of over .3. This is followed by Volatile Acidity at just under .15 and the rest are similar under .1. We are unable to derive important features from MLP as this is not a built-in function in pyspark.

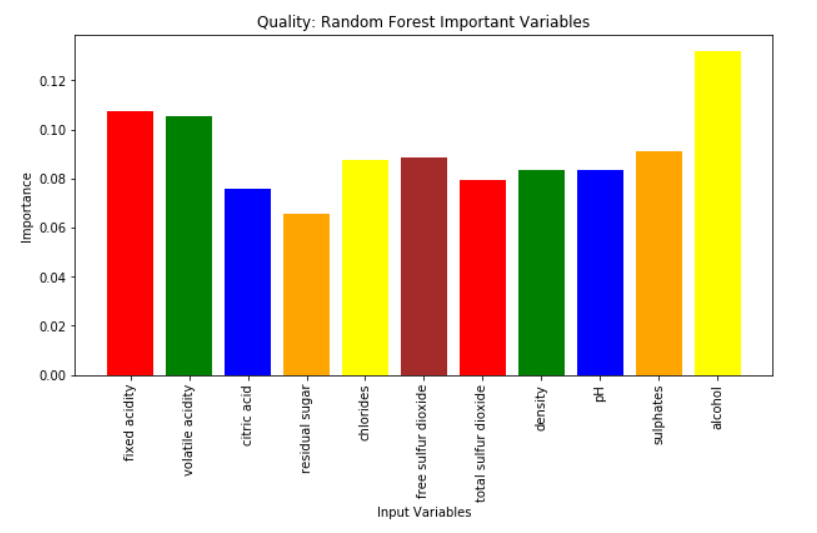


Figure : Quality Random Forest Important Variable

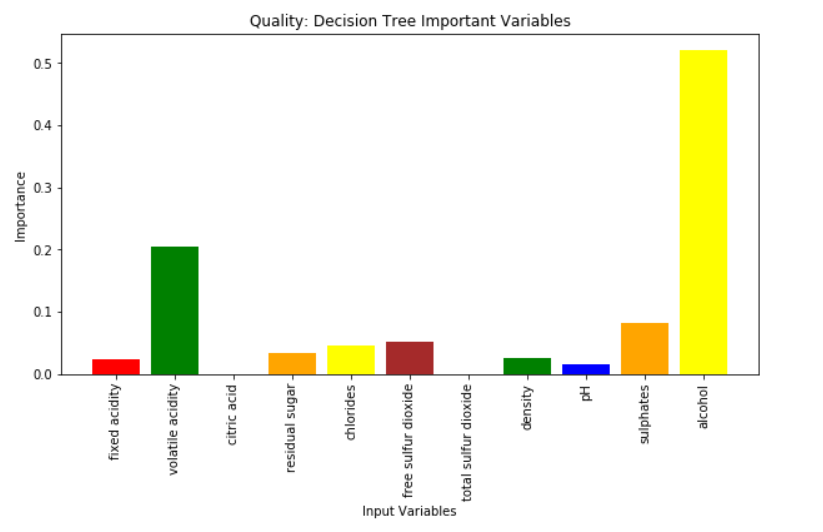


Figure : Quality Decision Tree Baseline Important Variable

* + 1. Color Analysis:

Since our Color variable only has two options, it is seen as binary which is a straight forward classification problem. Note that we did fit a partition of training and testing data into an 80% and 20% split respectively. We went ahead Random Forest, Decision Tree, and Gradient Boosting Tree. We see that the prediction of all the models is quite high, all being over 94%. We believe that this is because the color of a wine is heavily dependent on some of the variables that are captured in our data. We have placed all the model’s accuracies in Figures 30 below.

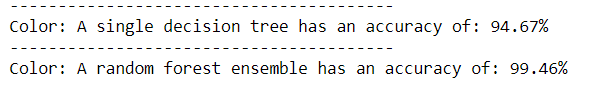
  


Figure : Color Random Forest vs Decision Tree Accuracy

We then decided to investigate the variable importance of the two whose variables importance is easily interpretable. We show the output of both Random Forest and Decision Tree below in Figures 30, 31, and 32 respectively. We see from the charts Chlorides and Total Sulfur Dioxide are important in both models. Though these values vary in importance and scale, as seen in the Random Forest total sulfur dioxide is just under .35 and Chlorides is just under .3, while in the Decision Tree Chlorides is just above .2 while Total Sulfur Dioxide is just under .7, and in the GB Tree Total Sulfur Dioxide is just under .5 while Chlorides are approximately .3. I assume that this is due to the nature of how the trees are built in the model. With Decision Tree not randomly selecting variable such as Random Forest is created, and the Gradient Boosting tree is altering weights of each variable instead of randomly selecting variables. In summary, it is hard to decipher exactly about these values, but from my knowledge I would deduce that due to this output, Chloride and Total Sulfur Dioxide hold large importance in determining the color of the wine.

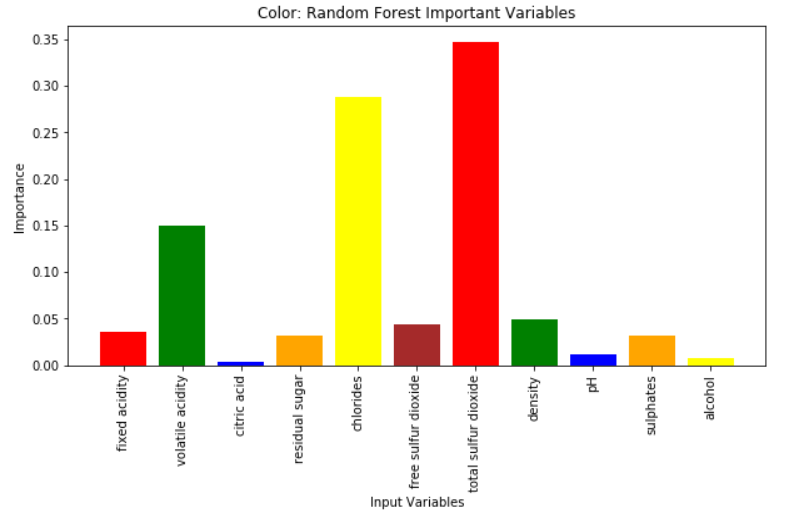


Figure : Color Random Forest Important Variables

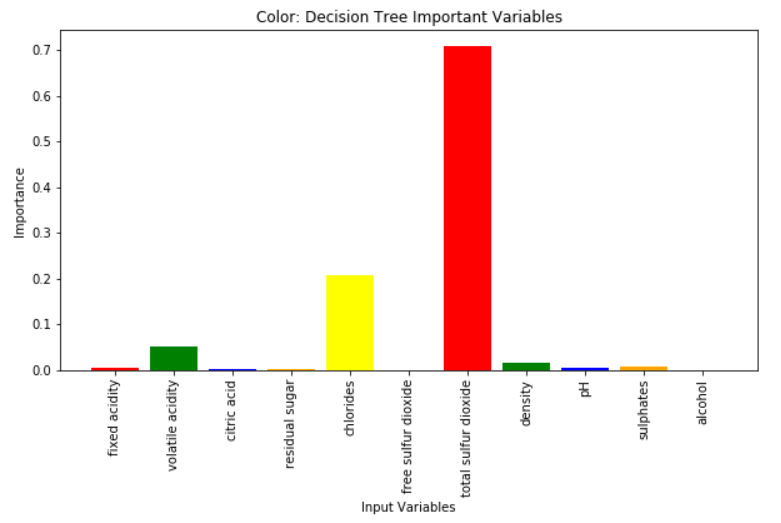


Figure : Color Gradient Boosting Important Variables

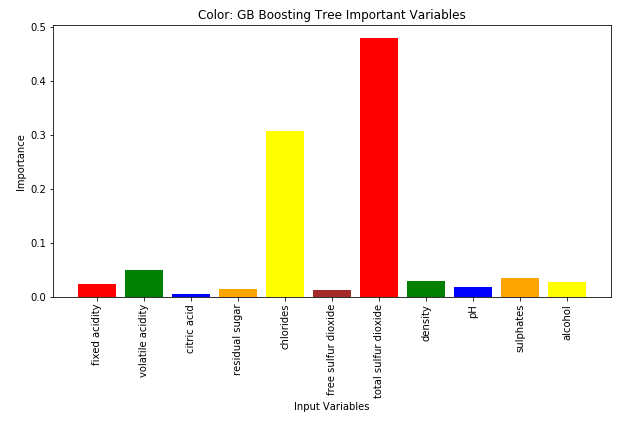


Figure : Color Gradient Boosting Important Variables

* + 1. Alcohol Content Analysis:

Since our Alcohol Content variable is continuous, we will fit all the same models as previously shown, but the regression form of these. This only takes a slight change in the naming of the package used, as seen in section7.ipynb, all other parameters remain the same. We use this because the alcohol variable can take any real number value so the same Random Forest and Decision Tree models will be fit to predict this value. We see from our output, which is not as straight forward as it has been in previous iterations, looks at both the R2 and RMSE. We desire lower values which indicates less error for RMSE while higher values for Random Forest with R2. Where R2 is stating how much of the variability in the data is explained by the model. We see that we have values of .678, .683, .446 for the Random Forest, Decision tree, and Gradient Boosting (100 iterations) respectively for RMSE seen in Figure 33. This is showing that the Gradient Boosting is better at prediction than the other two trees. Similarly, in R2 with values of .677, .673, and .859 the Gradient Boosting Tree is much better.

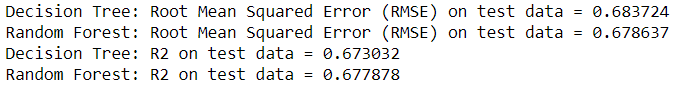
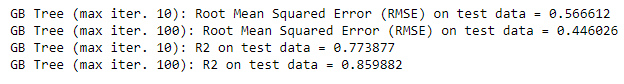
  


Figure : Alcohol Content Random Forest, Decision Tree, and Gradient Boosting Accuracy

We then investigate the variable importance of the three models whose variables importance is easily interpretable, again Random Forest, Decision Tree, and Gradient Boosting. We see the output of both below seen in Figure 34, 35, and 36 respectively. It is to note that it is clear in the output below that density is a very important factor with a value over 50% of the total in both Random Forest and Decision Tree, while many more are important in the Gradient Boosting Tree. This is interesting to see as previous iterations have shown similar output to Random Forest and Decision Tree, but this Gradient Boosting has many important variables and is the best predictor of the three. We see Density and Residual Sugar are the most important with values of .175 and approximately .14 respectively.

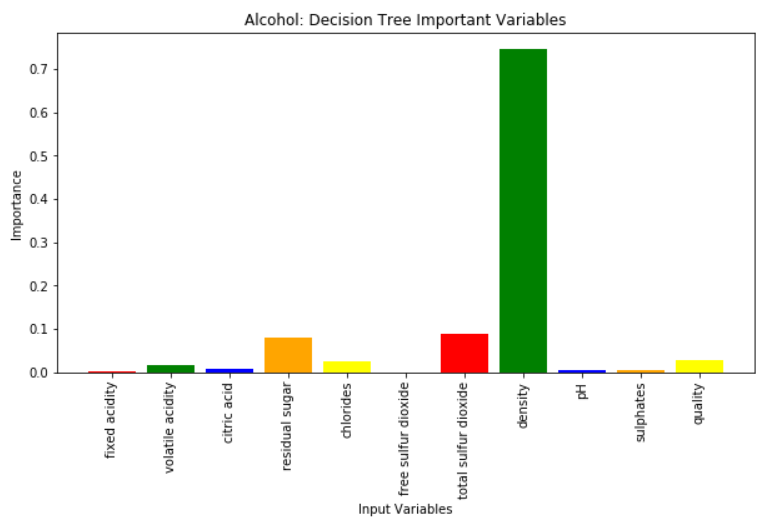


Figure : Alcohol Content Important Variables: Decision Tree

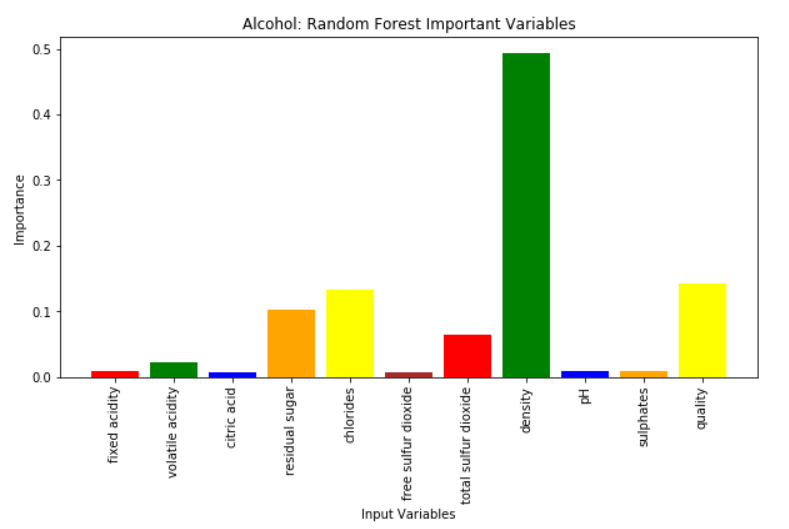


Figure : Alcohol Important Variables: Random Forest

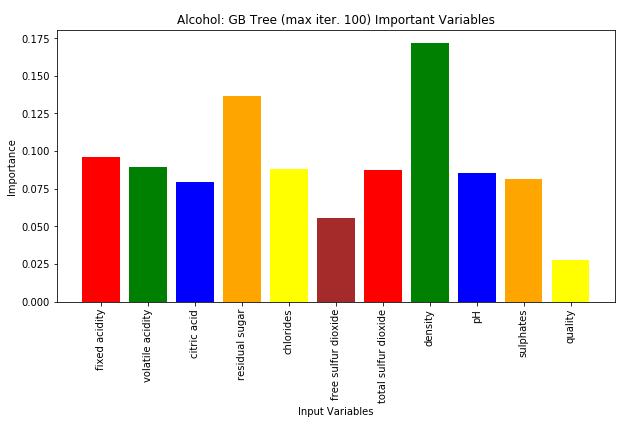


Figure : Alcohol Important Variables: Gradient Boosting

### 7.3:

We see some patterns in our results, ranging from relatively weak to very strong. Looking at the Random Forest output for our Quality data we see that all the variables are roughly the same important with one or two slightly more important in justifying/predicting the target variable. Similarly, in our Color variable analysis we see two variables, total sulfur dioxide and chlorides, with high important in the prediction of the target variable. While in our Alcohol Content analysis we see that the Density variable is important, but so are many others, followed closely by residual sugars. It is unfortunate that there were not some outlying factors that would be more influential on the prediction of Quality, as this was the main business objective desired.

It is to note, as shown above in the Color variable as a target analysis, that depending on what method we choose to use, the variables important will change. Though there should be some sort of pattern of which are most and least important in each of the methodologies used.

# Interpretation

### 8.1:

We have found two patterns in our data that provide direct results to your business objectives. Those being the prediction ability and important variables in both the Color of the wine, as well as the Alcohol Content.

When we look at the variables that determine the Color, total sulfur dioxide and chlorides, we can see that there is a pattern between the two variables and their ability to determine Color. This analysis is shown below.

When we look at the variables that determine the Alcohol Content, density, we see that Density has a clear linear relationship with Alcohol Content. Further analysis is shown below

Unfortunately, we were not able to find a model that predicted the quality of the wine accurately enough for us to use, with an ideal percentage over 90% but 80% being acceptable, and our closest model being the Random Forest at 66.74%.

### 8.2:

Most of our data visualization for the data, results, and models is done in the above modeling section. We will review in depth plots, created in Tableau, on how the individual variables interact with the target variable directly to show the patterns that were found.

First, we will look at Alcohol Content compared to Density to see the correlation that was picked up by our modeling technique. In Figure 37 below, we see a clear negative linear trend between Alcohol Content and Density. As the Density of the wine decreases, we see that there is an increase in the Alcohol Content of the wine.

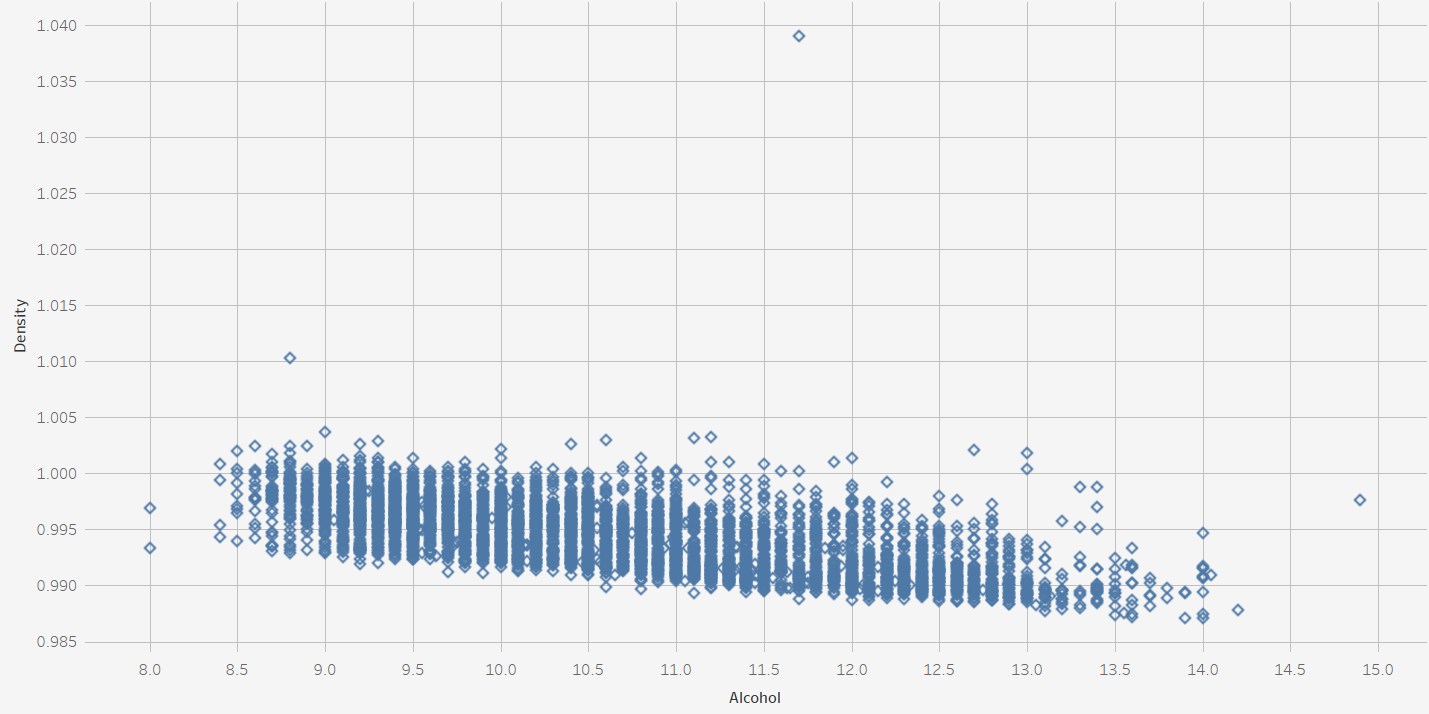


Figure : Relationship Between Density and Alcohol

Second, we will look at how Total Sulfur Dioxide and Chlorides relate to the Color of the Wine. Below in Figure 38, we see how the Chlorides level interacts with the color of the wine. We see that both wines have a lot of observations with values between 0 and approximately .3. When values go over this mark that is when we start to see only red wines with the highest value being just over 0.6 for Chlorides. This clear linear trend explains why we can derive such a strong pattern in our models to determine the Alcohol Content level.

Below in Figure 39, we see how the Total Sulfur Dioxide levels interact with the color of the wine. We see that both wines have observations ranging from 0 to approximately 175. Past this point we see most white wines up until a maximum value approximately just under 450. The combination of these two patterns in the data are what allows us to make such accurate predictions of the Wine Color.

The lack of these sort of patterns in the data for quality make it relatively impossible, without more observations or further variables, to find a strong enough signal in the data to predict at a high rate.

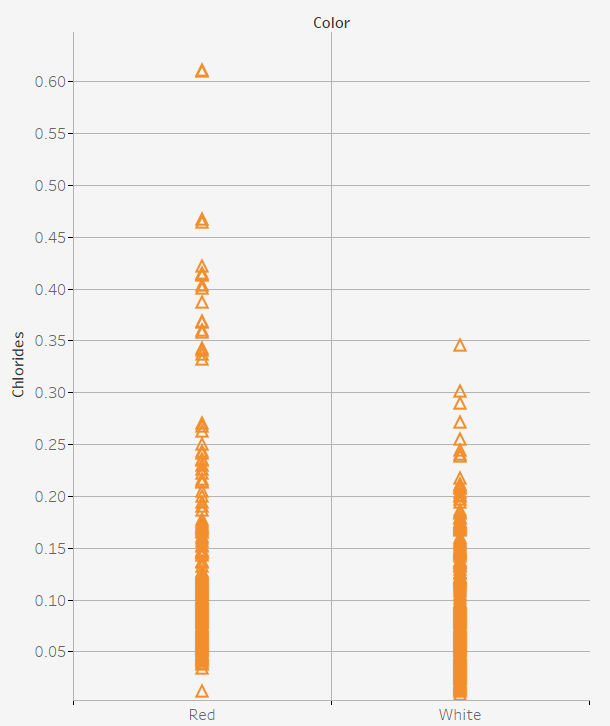


Figure : Chlorides vs Wine Color Plot

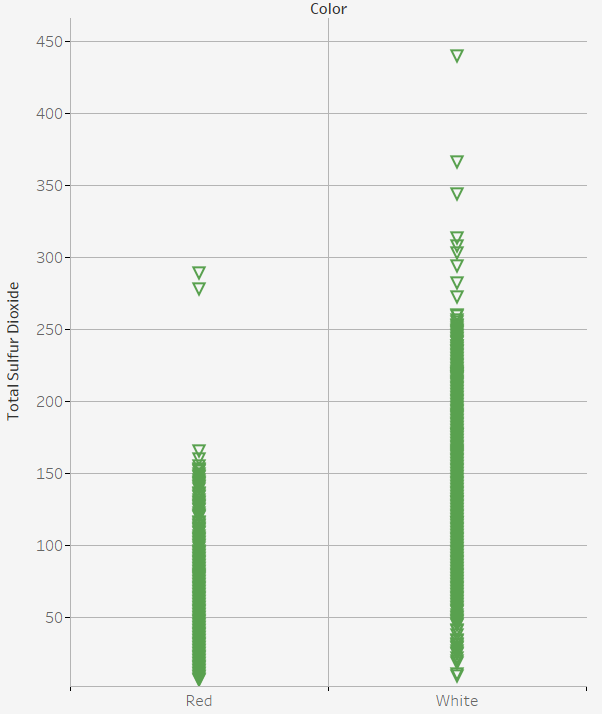


Figure : Total Sulfur Dioxide vs Wine Color Plot

### 8.3 + 8.4:

We will take some time to interpret and access what this all means as far as the models that we fitted for each of our three research questions. Please note that the patterns have been discussed in the section above.

For our Quality question, we have fit multiple Classification models specified above in our modeling section. Our highest predicting model was a Random Forest model with a prediction accuracy of 66.74% on our testing set. This means that when the model was shown data that it had never seen before, it predicted the correct Quantity of the wine roughly two thirds of the time. Unfortunately, in our statistics teachings, we learn that it is ideal to have model that has at least 80% accuracy to use for future predictions. This is of utmost importance as we would not want to predict a wine of high quality with low accuracy, and have the business pursue said wine for it to do poorly.

We have identified a few things we would need to create a better model which are out of our control at this time, though we will attempt to fine tune some of our models in the second iteration. In summary this included more observations and more attributes on the wine, such as the data on the grapes used to make this wine. The model can still be used for evaluation and is still insightful into what factors determine the quality of the wine. It did not find a specific variable that was significantly more influential on the quality of the wine, justifying that almost all the variables used in the model are useful.

For our Color research question, we have fit multiple Classification models that were covered in earlier sections of the paper. We found that our highest predicting model was a tie between Gradient Boosting and Random Forest with 99.46% accuracy on the test set. We recommend that if the company would like to pursue using a model for predicting the wine color, to use the Gradient boosting or Random Forest as Neural Networks are rather hard to interpret. We have also found what variables are important to determining the Color of the wine with Total Sulfur Dioxide and Chlorides. When looking at the relationship between these variables directly against the Color of the wine, you can see clear trends in the data, as shown in the above section.

For our last model, evaluating and predicting the Alcohol Content of the wine, we have fit the same models as the other but in regression form, specified earlier in the paper. We have found that our best model, was Gradient Boosting, has lower RMSE and R2 than the Decision Tree and Random Forest. Due to these evaluations, we would the customer use these models, as we saw that the R2 of the Gradient Boosting Tree was over 80%. This is worrisome though as we found that models in previous iterations fit much better and would recommend for our customers to potentially use those models instead. When evaluating the important variables, we see that Density is important, followed by residual sugars, in evaluating and predicting the Alcohol Content. Upon graphing the most important variable we see a negative linear trend, which justifies its ability to determine the impact of change in Alcohol Content.

### 8.5:

As our model for Quality was so too poor for prediction, I have attempted to investigate a few different types of Random Forest parameters to fine tune the model. We were interested in whether increases in the number of trees or the depth of the trees would better affect the prediction rate. We found that between a wide model, many trees, and a deep model, more depth, that the deeper model proved to be more efficient. We then combined the two into a hybrid between the two and found that the hybrid fit the best resulting in our top model for Quality. There is the potential to create a very large model for this research question to see if the performance increase levels off but was out of the scope for this project.

Though we were not able to derive a model that has a prediction over the 80% threshold, we were able to improve our models by ten percent each by fine tuning and learned more about how the different parameters affect the trees.

It seems that our prediction ability for quality to increase we would need more features, or input data, that better describes Quality as specified above. I attempted to fit a Random Forest with 500 trees and 30 depth but pyspark kept throwing errors, so I commented out the code. Another note is that Quality is a subjective value produced by “experts” of wine. There may be bias in these results and may explain why we are having issues predicting its value, unlike Color and Alcohol Content which are objective values.

# References

1. Parmar, R. (2018, July 09). Wine Quality. Retrieved August 01, 2019, from <https://www.kaggle.com/rajyellow46/wine-quality>
2. 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. ISSN: 0167-9236.
3. Fayyad,, U., Piatetsky-Shapiro,, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37-54.Auto Numeric Node Model Options. (n.d.). Retrieved August 05, 2019, from
4. Gradient Boosting Explained - The Coolest Kid on The Machine Learning Block | Displayr. (2019). Retrieved 22 September 2019, from <https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/>
5. KNN Classification using Scikit-learn. (2019). Retrieved 22 September 2019, from <https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn>
6. Scikit-learn Tutorial: Machine Learning in Python – Dataquest. (2019). Retrieved 22 September 2019, from <https://www.dataquest.io/blog/sci-kit-learn-tutorial/>
7. Singh, D., & Scikit-learn, M. (2019). Machine Learning with Neural Networks Using Scikit-learn | Pluralsight. Retrieved 22 September 2019, from <https://www.pluralsight.com/guides/machine-learning-neural-networks-scikit-learn>
8. Hyperparameter Tuning the Random Forest in Python. (2019). Retrieved 22 September 2019, from <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>
9. Outlier Detection: An ETL Tutorial with Spark. Retrieved 19 October 2019, from <https://towardsdatascience.com/outlier-detection-an-etl-tutorial-with-spark-f4197eb84a7b>

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright. (See: <https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html>).

I also acknowledge that I have appropriate permission to use the data that I have utilized in this project. (For example, if the data belongs to an organization and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."