EDA on the quality of White Wine

EDA on the quality of White Wine

Merzu K Belete

21 April 2018

White wine, enjoy with your dinner

Introduction: White wine has believed to be existed for at least 2500 years. The sommelier - an expert on wine taster learns and practices for many years to understand the qualities of wines. Artificial intelligence is an ideal area to explore and solve this problem. Here. I'll go through an exploratory data analysis (EDA) and a simple Machine learning using random forest on the white wine dataset. This dataset contains 11 physicochemical (some are physical and others are chemical properties) properties and a quality of the wine, which is a sensor data from the sommelier.

Loading the dataset:

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.0             0.27        0.36           20.7     0.045
## 2 2           6.3             0.30        0.34            1.6     0.049
##   free.sulfur.dioxide total.sulfur.dioxide density  pH sulphates alcohol
## 1                  45                  170   1.001 3.0      0.45     8.8
## 2                  14                  132   0.994 3.3      0.49     9.5
##   quality
## 1       6
## 2       6
```

Short data Summary and descriptions: The white wine dataset contains close to 5000 observations with 12 physicochemical properties and one observation ID. The target variable in this EDA is the sensor data wine quality. The median wine qualities is 6 with mean value 5.878. Moreover, the distribution of wine quality is nearly a normal centred on the median. Therefore, I will create a categorical variable, rate, with a low below the median, high above the median and medium with median value.

Lets look a descriptive statistics of the dataset

Dimension of the dataset

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Column names

```
##  [1] "X"                   "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"         "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide" "total.sulfur.dioxide" "density"
## [10] "pH"                  "sulphates"            "alcohol"
```

```
## [13] "quality"
```

Descriptive statistics

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

White Wine Attributes Information

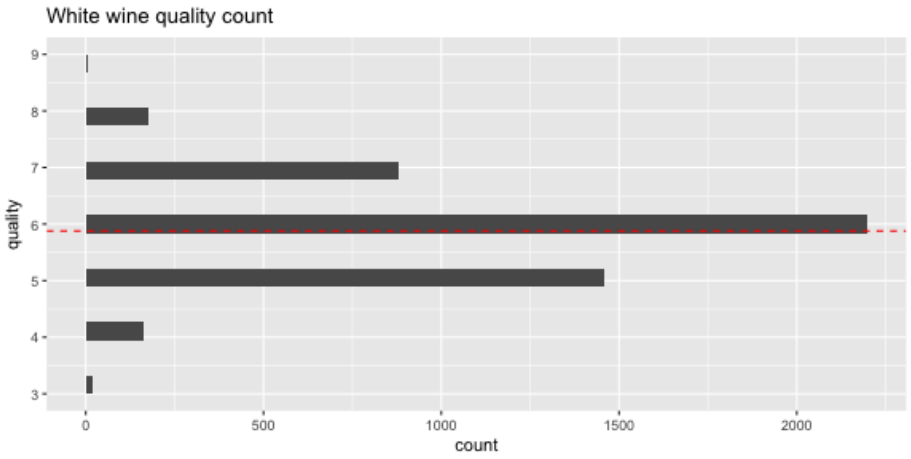
Input variables (based on physicochemical tests) & attributes description:

1. Fixed acidity (mostly tartaric acid - g / $dm^{3)}$: imparting sourness and resistance to microbial infection.
2. Volatile acidity (acetic acid - g / $dm^{3)}$: gives vinegar its characteristic flavour and aroma.
3. Citric acid (g / $dm^{3)}$: found in small quantities, citric acid can add 'freshness' and flavour to wines

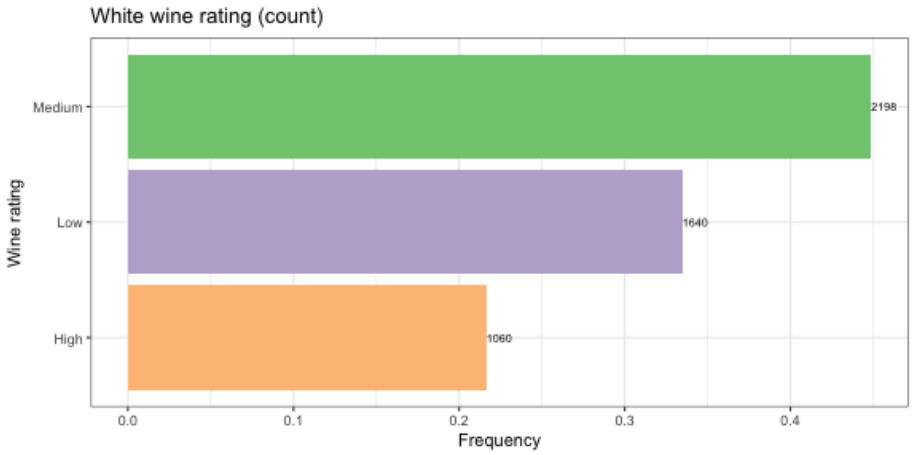
4. Residual sugar (g / $dm^{3)}$: the amount of sugar remaining after fermentation stops.
5. Chlorides (sodium chloride - g / $dm^{3)}$: the amount of salt in the wine.
6. Free sulfur dioxide (mg / $dm^{3)}$: the free form of SO2 exists which prevents microbial growth and the oxidation of wine.
7. Total sulfur dioxide (mg / $dm^{3)}$: the amount of free and bound forms of SO2.
8. Density (g / $cm^{3)}$: the density of the wine.
9. pH: describes how acidic or basic a wine is on a scale from <7 (acid range) to >7 (base range) and 7 neutral. Wines are mostly acidic.
10. Sulphates (potassium sulphate - g / $dm^{3)}$: which prevents microbial spoilage and fight oxygenation.
11. Alcohol (% by volume): the percent alcohol content of the wine
12. Quality: A sensor data from a sommelier with (0 to 10).

Univariate Plots Section

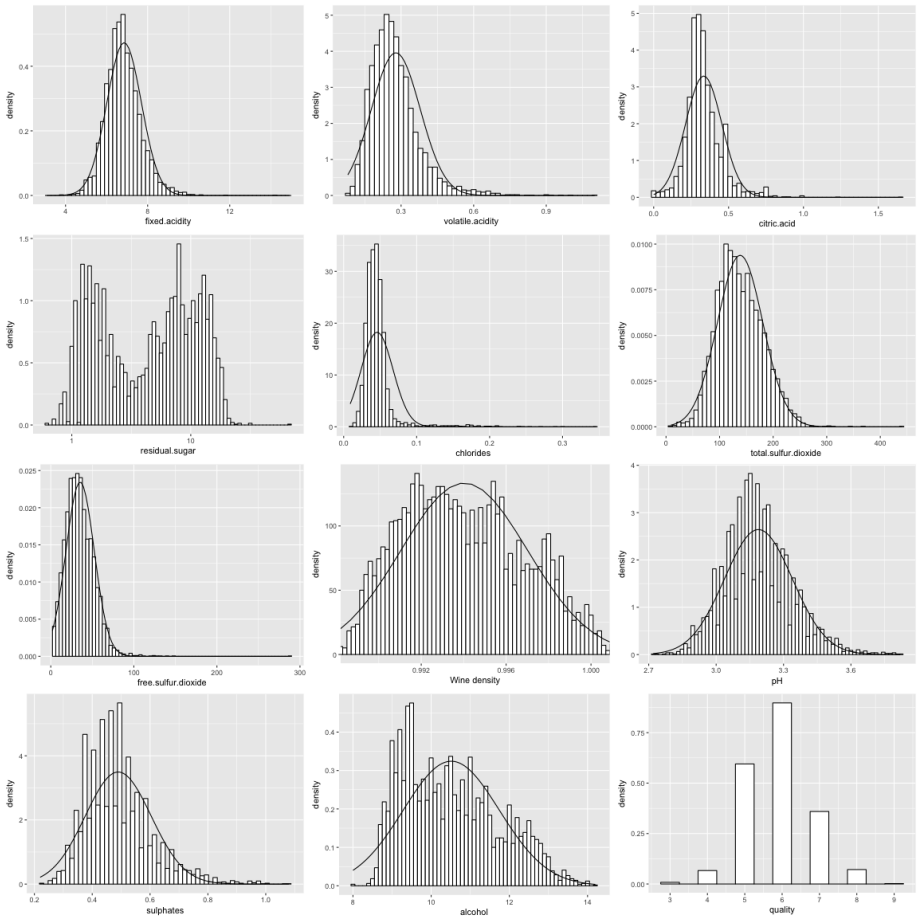
Lets plot the distribution of white wine quality. The red dashed line is the mean of the quality of wine. Majority of the wine qualites are of rating 6, very few are rating 3 and 9.



Let us group the wine quality to low (quality < 6), medium (quality $= 6$) and high (quality > 6). Now the distribution of wines rating is comparable and a simple classifier with these categories can work reasonably well.



Let us explore the distribution of each feature variables using a histogram. Most of the variables follow nearly a Gaussian distribution. Alcohol has high observations at low alcohol and citric acid has high observation around 0.495. However, residual sugar after log transformation, it follows bimodal distribution.



Univariate Analysis

What is the structure of your dataset?

The white wine dataset contains 11 variables, one ID and one a sensor variable with 4898 observations. Some of them are chemical and others are physical properties of a wine.

What is/are the main feature(s) of interest in your dataset?

In this EDA, wine quality is target variable while others like alcohol, density , residual sugars are important variables for that correlate with the wine quality.

**What other features in the dataset do you think will help support your **

investigation into your feature(s) of interest?

I think all the other variables are also important because the wine quality is not completely described by the few features. That means that the wine quality does not have a strong correlation with any of the other variables. In fact, machine learning may predict wine quality with very low accuracy and more features may require to accurately predict the wine quality.

Did you create any new variables from existing variables in the dataset?

Yes, I created (i) rate variable that classifies wines quality to low, medium and high categories, and (ii) bound sulfur dioxide that is the difference between the total and free sulfur dioxide.

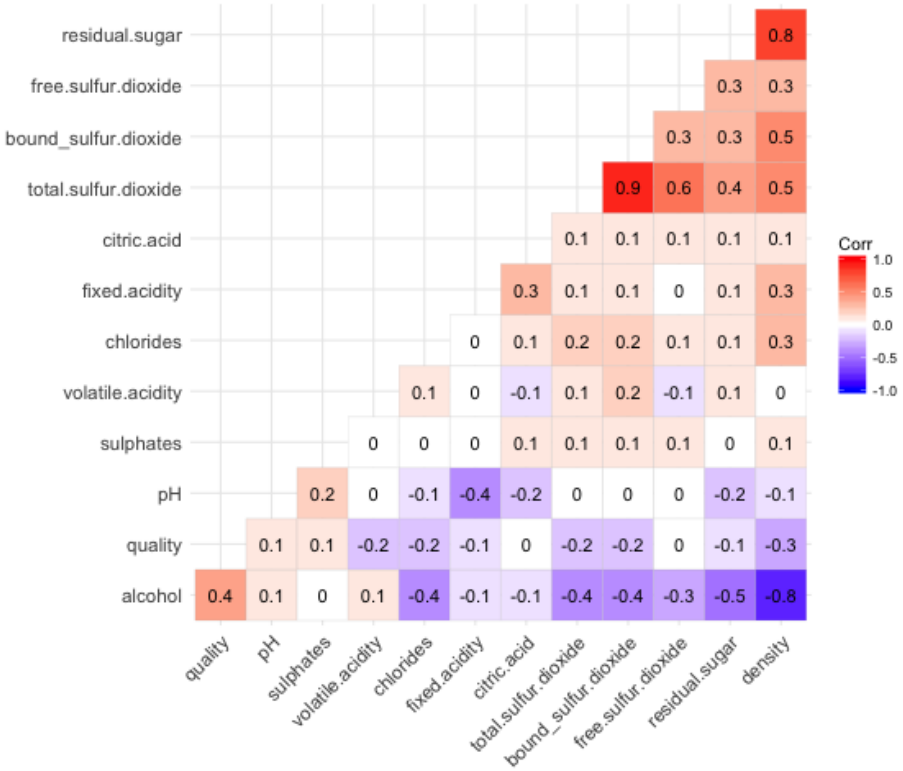
**Of the features you investigated, were there any unusual distributions? **

Did you perform any operations on the data to tidy, adjust, or change the form \ of the data? If so, why did you do this?

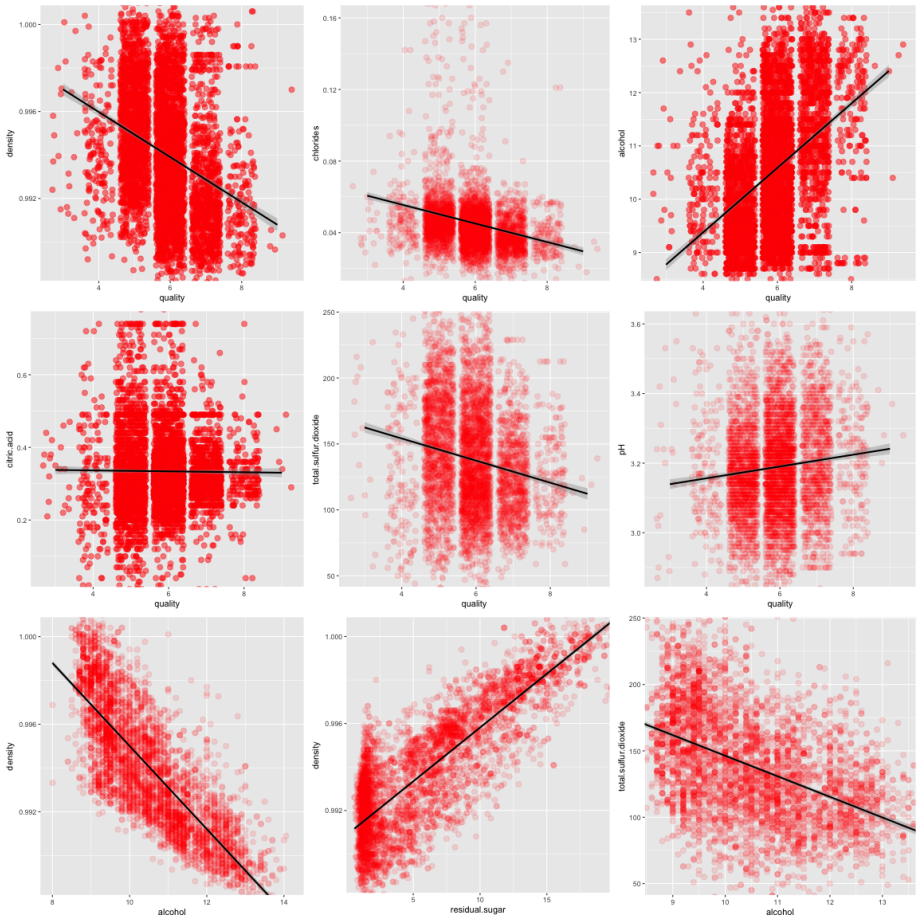
Most of the features have a normal distribution with a few outliers to the right . However, residual sugar after log transformation is a bimodal distribution. Citric acid distribution also follows with a normal distribution, but at 0.5 there is a high observation that deviates the normal curve fitted on it. This dataset is tidy but I think it is not enough to completely describe/predict the quality of wine with a high degree of accuracy unless it is bucketed with a few classes. The distribution of the wine quality is not uniform, very few with high quality as well very few with low quality. Any machine learning will struggle to correctly predict those outlier wine qualities.

Bivariate Plots Section

Below I used correlation matrix to visualize the pair-wise relationship between two variables and will guide us which variables to look deeper. Alcohol has the highest positive correlation with the quality of the wine and density has the highest anti-correlation with quality of a wine.



let us look deeper into the relationship between two variables using scatter plots. As shown in the figure below, the wine quality has a positive correlation with alcohol and pH, but it has also a negative correlation with density, chlorides total sulfur dioxide. However, it has little or no relationship with citric acid. Alcohol also has a negative correlation with density and total sulfur dioxide. Density and residual sugar have a positive correlation with higher residual sugar.



Bivariate Analysis

**Talk about some of the relationships you observed in this part of the **

investigation. How did the feature(s) of interest vary with other features in the dataset?

Without converting the quality of wine to a few buckets, the quality of a wine is weakly correlated with pH and sulphates and negatively correlated with all variables except citric acid and free sulphates dioxide with no correlation for those variables. But after, bucket the wine quality into low, medium and high based the quality scores with low less than 6, medium with quality 6 and high above 6, it is clear that wine quality is positively correlated with alcohol, and negatively correlated with density.

**Did you observe any interesting relationships between the other features **

(not the main feature(s) of interest)?

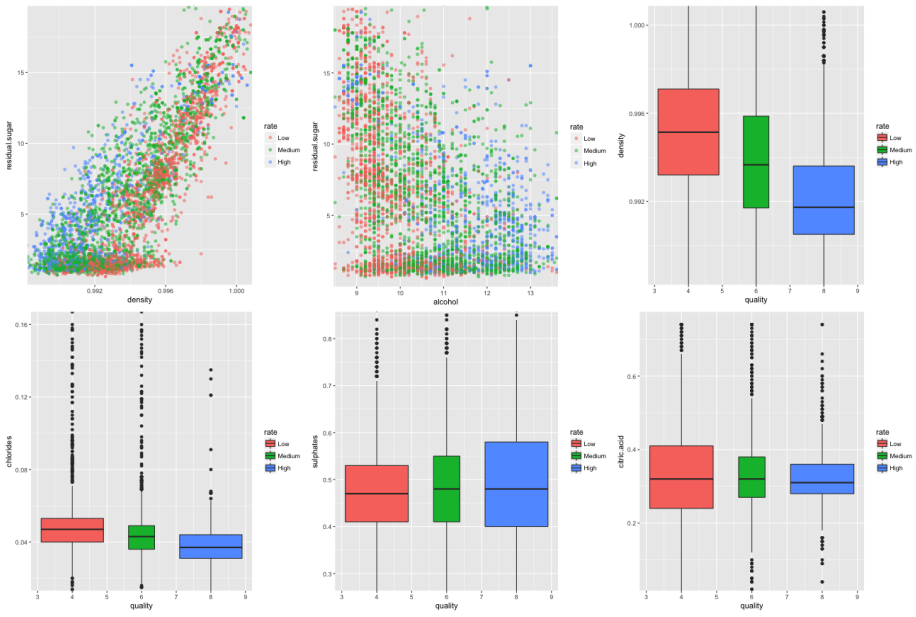
The density of wine is positively correlated with residual sugar (not surprising since residual sugars (mostly solid) have a higher density) but it is also negatively correlated with the alcohol content of the wine.

What was the strongest relationship you found?

Density with alcohol and density with residual sugar has a higher correlation.

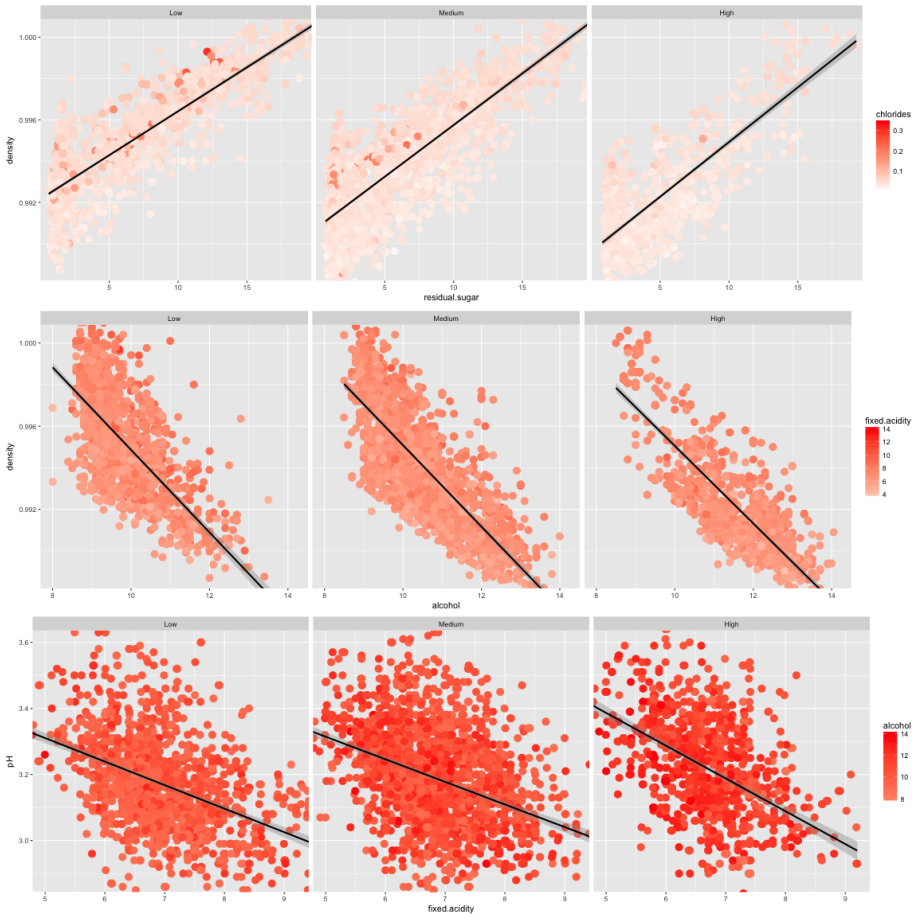
Multivariate Plots Section

Let us dive into three or more variables in one figure with scatter and box plots. Again it is apparent that the quality of a wine is directly correlated with alcohol and sulphate. However, it is negatively correlated with density, residual sugar and chlorides. Here the box plot also shows that citric acid has little negative correlation with quality of a wine.



Diving into more relationship with a scatter plots adding smoothing. Quality of a wine has a negative relationship with density, chlorides and residual sugar. It also shows that the density and residual sugar has a linear relationship at higher

residual sugar. Similarly, density has a linear relationship at higher alcohol contents of the wine. pH has a negative correlation with fixed acidity (not surprising that higher pH are bases than acid).



Multivariate Analysis

Talk about some of the relationships you observed in this part of the

investigation. Were there features that strengthened each other in terms of

looking at your feature(s) of interest?

High density, high residual sugar and high citric acid are the quality of a low rated wine. High alcohol, low density and low fixed acidity are the quality of a highly rated wine. Many variables also have very weak but positive correlation.

Were there any interesting or surprising interactions between features?

I think most of the correlations I observed are what I expected. However, the density of highly rated wines is less than one (less than water at room temperature). I know we don't like to drink a wine that is too heavy in our mouth but that is a surprise for me. What if we can reduce the density of the wine with half of the water and will that be still a good wine? Maybe or it may not go down through our oesophagus?

Modeling with random forest classifier

Let us use a random forest regressor to classify the wines quality in the three class.

```
## ntree      OOB      1      2      3
##   100:  29.26% 28.71% 24.63% 39.53%
##   200:  28.97% 28.53% 23.73% 40.32%
##   300:  28.82% 28.53% 23.79% 39.53%
##   400:  28.53% 28.53% 23.53% 38.74%
##   500:  28.62% 28.26% 23.53% 39.53%
```

```
##
## Call:
##  randomForest(formula = rate ~ ., data = train, importance = TRUE,      do.trace = 100)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 28.62%
## Confusion matrix:
```

```
##        Low Medium High class.error
## Low    802    299   17   0.2826476
## Medium 240   1186  125   0.2353320
## High    18    282  459   0.3952569
```

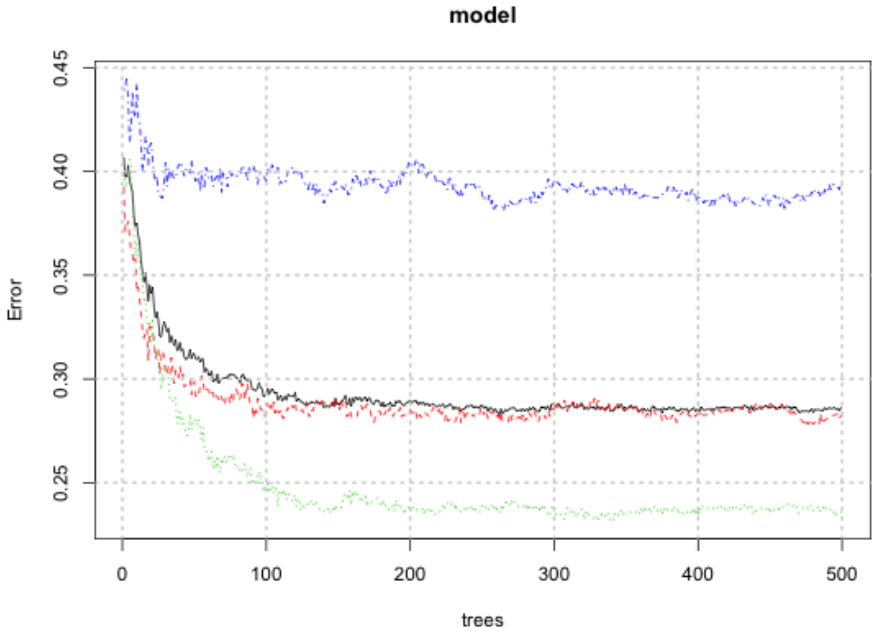
Confusion matrix on test data

```
##
## pred     Low Medium High
##   Low    368     83    3
##   Medium 144    511  102
##   High    10     53  196
```

**OPTIONAL: Did you create any models with your dataset? Discuss the **

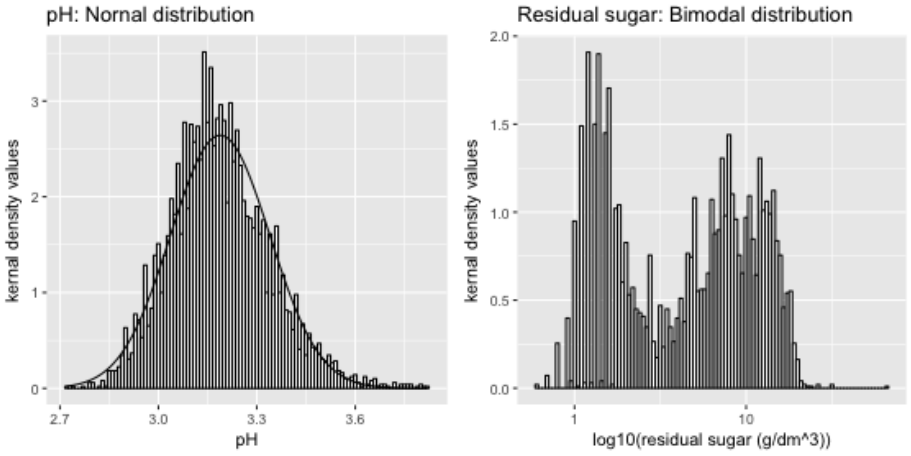
strengths and limitations of your model.

I created a random forest regressor that classifies the wine qualities into three groups (low, medium or high) with an out-of-bag (OOB) of 28.27%. OOB is a method of measuring the prediction error of random forests. This model is powerful toclassify in those quality types. However, if we use the model on the original dataset, the OOB surely increases. Any model will struggle to predict the lower and higher qualities of the wine.



Final Plots and Summary

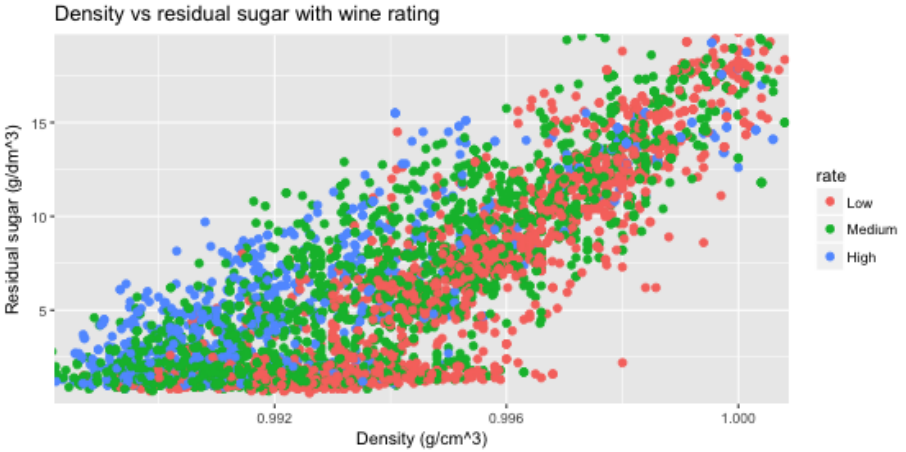
Plot One



Description One

Most of the feature variables have an almost normal distribution with a few outliers. The pH distributions are the one with a perfect uniform distribution, while the residual sugar after log10 transformation it is a bimodal distribution.

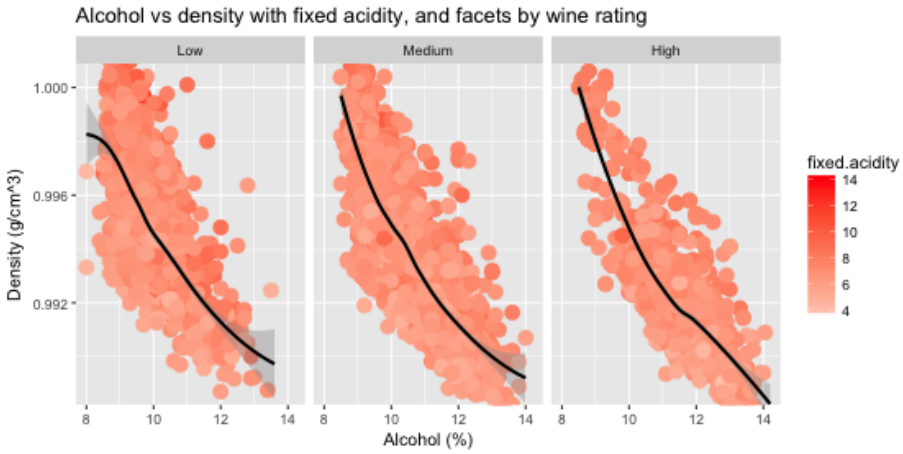
Plot Two



Description Two

Here I showed three variables in one plot: residual sugar as the function of density with a colour of wine rating. From this plot, residual sugar is not a strictly linear relationship. In fact, there are many ratings of low, medium and high with very low residual sugar with a continuous increasing density and majority of this wines are low rating class. The other class of wines are where residual sugar increases linearly with density, and the majority of this classes of wine are high rating with low density and low residual sugars.

Plot Three



Description Three

In this last figure, there are four variables explored. The density of the function of alcohol colour by fixed acidity and facets with wine rating. First, highest quality of a wine is with low acidity (light red colour) and low density. Second, the density and alcohol are not linearly correlated especially for low rating wine, most probably the correlation is a power law.

Reflection

In this report, first, I explored the distribution of each variable with a histogram and fitting with a normal distribution. Second, I examined in detail on the relationship between two variables using correlation table, scatter and box plots. Finally, three or more variable is studied in detail to give us insight into the target variable, the wine quality. In the end, a simple random forest classifier was used to predict the wine qualities into three categories.

I convinced in this dataset, there is no strong predictor of the quality of the wine. In fact, simple linear models will predict very poorly. Some of the variables are also not a simple linear relationship. For example, density with residual sugar is not a simple linear relationship for high rated wine.

The white wine dataset is tidy. However, the dataset is skewed for wine quality observations. There are very few high or low wine quality observations. This will be very difficult to train a machine learning to predict all qualities of wine (3-9) accurately.