GR5291 Project Wine Quality Analysis

Saier Gong (sg3772)

5/2/2020

Part 1 Data Introduction

1.1 Data Introduction

This wine quality data set is downloaded from Kaggle, it has 6497 observations and 13 columns, including one categorical explanatory variable "type", 11 quantitative explanatory variables, and one response variable. The reponse variable is the quality of wine, from 1 to 10, which represents from low quality to high quality.

After cleaning the data set, removing some rows with missing values, the data set has 6463 observations. I also modify the response variable:

- if the quality score is higher than 5, set it as 1, representing high quality;
- if the quality score is lower than 5, set it as 0, representing low quality.

So now, the data set has 6463 observation rows, 1 categorical variable, 11 quantitative variables and a $(0, 1)$ response variable.

```
## # A tibble: 6 x 2
##   variable         feature
##   <chr>            <chr>
## 1 type             categorical
## 2 fixed.acidity    quantitative
## 3 volatile.acidity quantitative
## 4 citric.acid      quantitative
## 5 residual.sugar   quantitative
## 6 chlorides        quantitative
```

```
## # A tibble: 7 x 2
##   variable             feature
##   <chr>                <chr>
## 1 free.sulfur.dioxide  quantitative
## 2 total.sulfur.dioxide quantitative
## 3 density              quantitative
## 4 pH                   quantitative
## 5 sulphates            quantitative
## 6 alcohol              quantitative
## 7 quality              dependent
```

1.2 Data Collection

Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

1.3 Interesting Question

Through analyzing this wine quality data, I want to show the following interesting topics:

- which feature or combination of several features may contribute most to the taste of wine;
- how the type of wine may influence the taste of wine;
- only according to the provided variables, how to make and choose good wine.

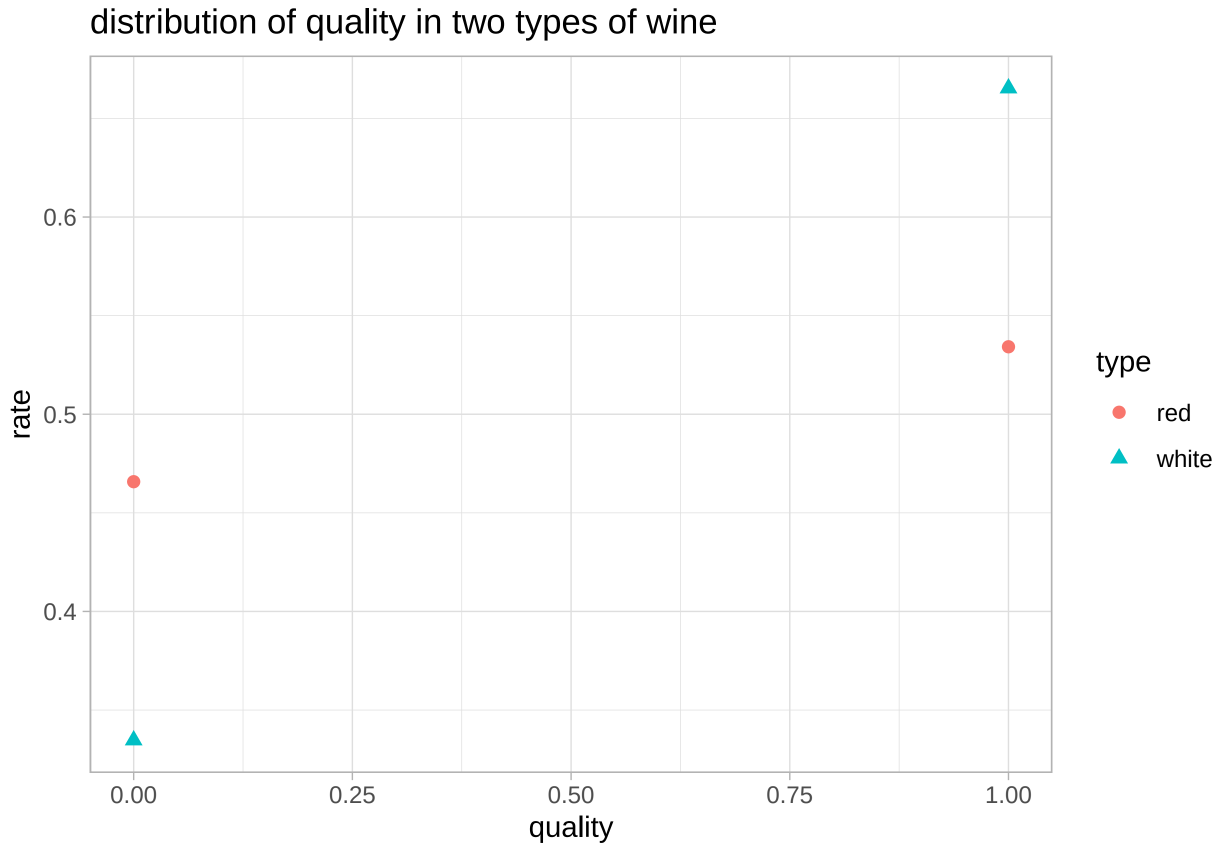
Part 2 Conclusions

According to my analysis, I have the following conclusions about the interesting topics:

- The volatile acidity variable has the greatest impact on the quanlity of wines;
- Given other quantitative variables the same, the estimated odds of high quality in white wine is about 6% of the odds of high quality in red wine. Therefore, the type of wine has an influence on quality, and it is more likely to buy fake white wine.
- With more alcohol and sulphates and with less volatile acidity, the quality of wine is increasing.

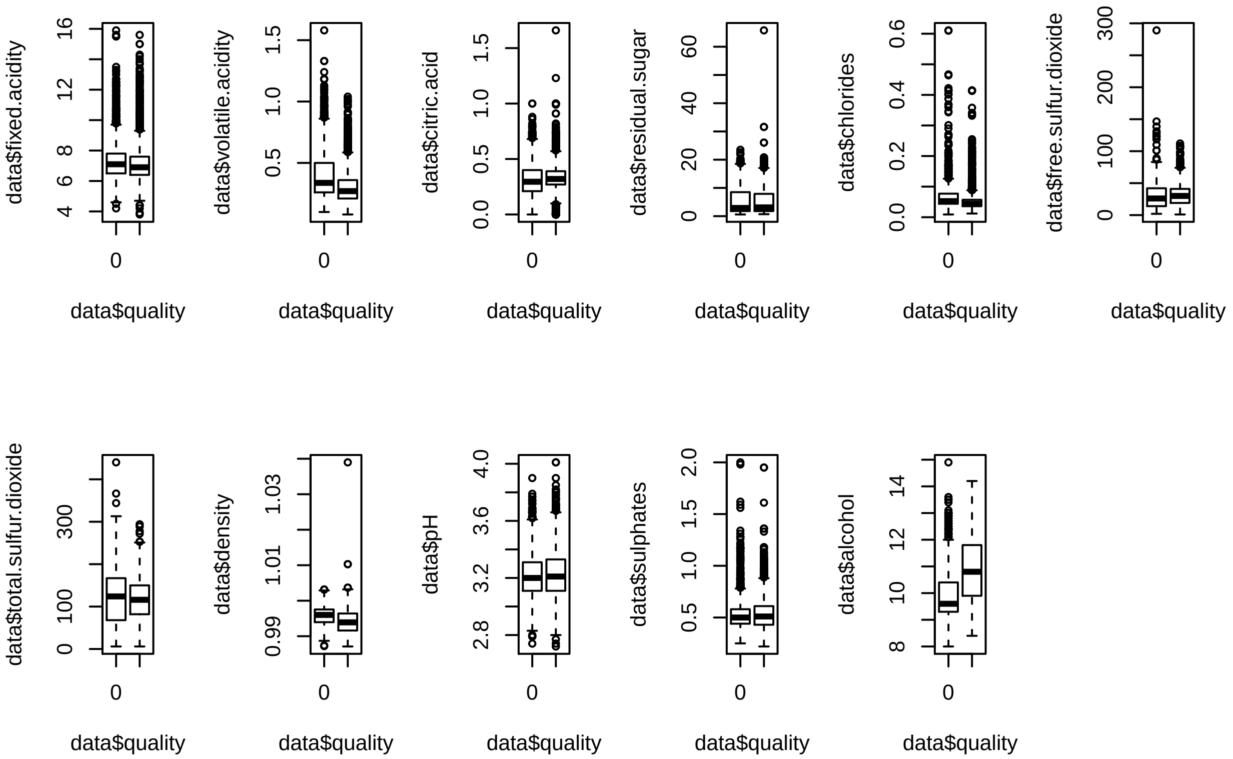
Part 3 Wine Quality Analysis

3.1 Basic Understanding of the Data



From the plot, we can see that in this data set:

- both red and white wine group have more high quality wine than low quality wine;
- the proportion of high quality wine in white wine group is higher than that in red wine group;



As we can see from the plots, some quantitative variables seem not having an absolute influence on the quality, so in the model building process, we should drop some variables.

3.2 Only Consider Quantitative Variables

First of all, we only consider the influence of quantitave variables on the quality of wines

At the beginning, when we do not consider any explanatory variables in the model, the estimated probability of high quality wine is 0.54506.

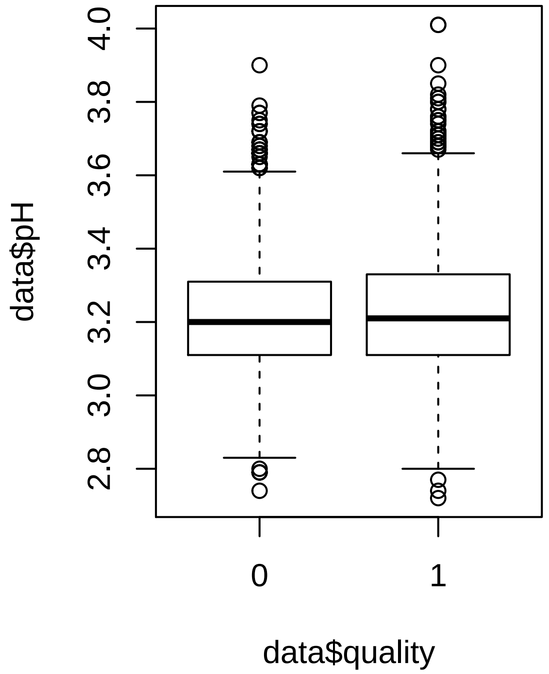
When we consider all the quantitative variables in the model, from the p-value of each parameter, we get that the $p-value_{fixed.acidity} = 0.1055, p-value_{chlorides} = 0.3747, p-value_{density} = 0.1621$, which are much more bigger than 0.01, so it is greatly likely that we can remove these variables.

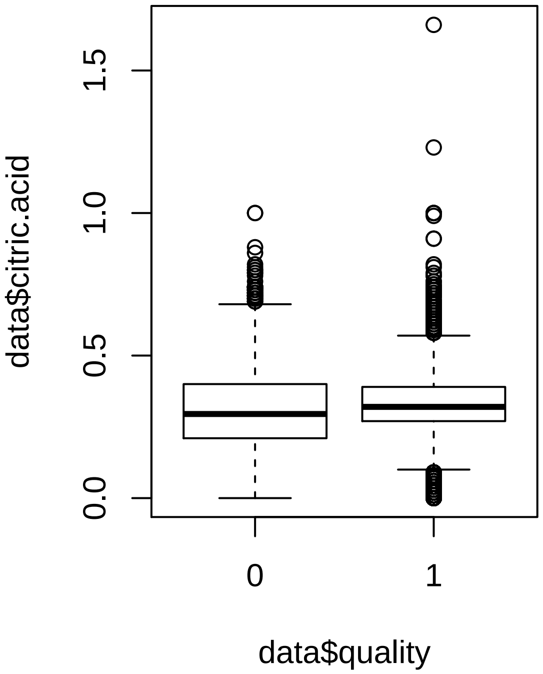
To test the significance of these 3 variables, we set the null hypothesis $H_0 : \beta_{fixed.acidity} = \beta_{chlorides} = \beta_{density} = 0$.

From the ANOVA table, we can see the $p-value = 0.2561$, which is much greater than 0.01, so there is not enough evidence to reject the null hypothesis, which means the reduced model can explain the data well.

However, the p-value of parameter citric.acid and pH is 0.0106 and 0.0858, respectively. So we are going to test the significance of these 2 variables. $H_0 : \beta_{citric.acid} = \beta_{pH} = 0$

When removing both two variables,the p-value is 0.0015, much smaller than 0.01, so we can reject the null hypothesis, and know that we should at least keep one variable.





According to their p-value and the boxplots, we can keep the variable citric.acid.

When only removing pH variable, the p-values of all the parameters in the new model are smaller than 0.01.

We are 95% confindence that:

- for every 1 increase in alcohol variable, the odds of the wine quality increases by a factor between 2.4 and 2.7;
- for every 1 increase in volatile acidity, the odds of the wine quality decreases by a factor between $6.3 * 10^{-3}$ and $1.6 * 10^{-2}$.

3.3 Consider the type variable

In this data set, it has 1593 red wine and 4870 white wine observations.

Is type of wine effect significant?

After adding the type variable into the model, we set the null hypothesis $H_0 : \beta_{type} = 0$.

From the ANOVA table, he p-value of β_{type} is 0.01246, a little bigger than 0.01, although not strong enough to reject the null hypothesis, I decide to keep this variable.

So now, our model has 7 quantitative variables (alcohol, volatile.acidity, sulphates, residual.sugar, total.sulfur.dioxide, free.sulfur.dioxide and citric.acid) and one categorical variable (type).

3.4 Consider the Interaction

Let's consider a bigger model.

```
## # A tibble: 7 x 2
##   variable                    p.value
##   <chr>                         <dbl>
## 1 type*alcohol              0.875
## 2 type*volatile.acidity     0.0000000276
## 3 type*sulphates            0.827
## 4 type*citric.acid          0.00112
## 5 type*residual.sugar       0.0458
## 6 type*total.sulfur.dioxide 0.0000000194
## 7 type*free.sulfur.dioxide  0.000657
```

```
## # A tibble: 4 x 2
##   variable                    p.value
##   <chr>                         <dbl>
## 1 type*volatile.acidity     0.0000000276
## 2 type*citric.acid          0.00112
## 3 type*total.sulfur.dioxide 0.0000000194
## 4 type*free.sulfur.dioxide  0.000657
```

So, based on the p-value of each interaction term, I choose to add

$$type * volatile.acidity, type * citric.acid, type * total.sulfur.dioxide, type * free.sulfur.dioxide$$

into the model as interaction terms.

3.5 Final Model

Conclusions:

The final model I decide is $m_{final} = -9.253 + 1.016\beta_{alcohol} - 3.494\beta_{volatile.acidity} + 1.848\beta_{sulphates} + 0.0670\beta_{residual.sugar} - 1.064\beta_{total.SO_2} + 0.0194\beta_{free.SO_2} - 0.623\beta_{citric.acid} + 0.057\beta_{type.white} - 3.055\beta_{type.white*volatile.acidity} + 0.244\beta_{type.white*citric.acid} + 0.0132\beta_{type.white*total.SO_2} - 0.006\beta_{type.white*free.SO_2}$.

- Given other quantitative variables the same, the estimated odds of high quality of white wine is about 6% of the estimated odds of high quality of red wine.
- Given other quantitative variables the same, when the alcohol increases 0.1, the estimated log odds of high quality of white wine increases 0.2122, red increases 0.1061.
- Given other quantitative variables the same, when the sulphates increases 0.01, the estimated log odds of high quality of white wine increases 0.03696, red increases 0.01848.
- Given other quantitative variables the same, when the residual sugar increases 1, the estimated log odds of high quality of white wine increases 0.134, red increases 0.067.
- Given other quantitative variables the same, when the volatile acidity increases 0.01, the estimated log odds of high quality of white wine decreases 0.1, red decreases 0.035.
- Given other quantitative variables the same, when the citric.acid increases 0.01, the estimated log odds of high quality of white wine decreases 0.01, red decreases 0.006.
- Given other quantitative variables the same, when the total SO_2 increases 10, the estimated log odds of high quality of white wine decreases 0.195, red decreases 0.164.
- Given other quantitative variables the same, when the free SO_2 increases 10, the estimated log odds of high quality of white wine inreases 0.328, red increases 0.194.

Basic on this analysis, if you are not an expert of wine but want to choose a good wine based on this data set, I will give the following suggestions:

- If you don't have a preferrence of red wine or white wine, I suggest to choose red wine with a little bit higher alcohol, sulphates residual sugar and free SO_2 and with a little bit lower volatile acidity, citric.acid and total.SO_2.
- If you prefer white wine, I suggest to choose he white wine with a little bit higher alcohol, sulphates, residual sugar and free SO_2, and with lower volatile acidity, citric.acid and total.SO_2.

If you want to open a wine factory, the suggestions are the same as the above.

Other shortage

The variables are not contain a wide range, since there are many other valuable variables not considered such the type of grades, the design of bottle and the wine brand.