Wine – Data Analysis

Group 31

Yajiao Lei B00729064

June Wang B00806032

Stat 3340

Dalhousie University

Abstract

The ingredients of red wine have been researched in the study. The variables that affect the quality of red wine have been obtained using multiple linear regression analysis. A new data point named alcohol level has been added to the dataset. The variable quantity has been visualized using a pie chart. A moderate positive association between alcohol and pH has been found using a scatter plot. There is a negative association between pH value and density. A large number of medium quality wine is collected in the sample. A high amount of sulphate, chloride, and alcohol increases the quality of the wine. Red wine has less pH value. ANOVA found that there is at least one of the means is different.

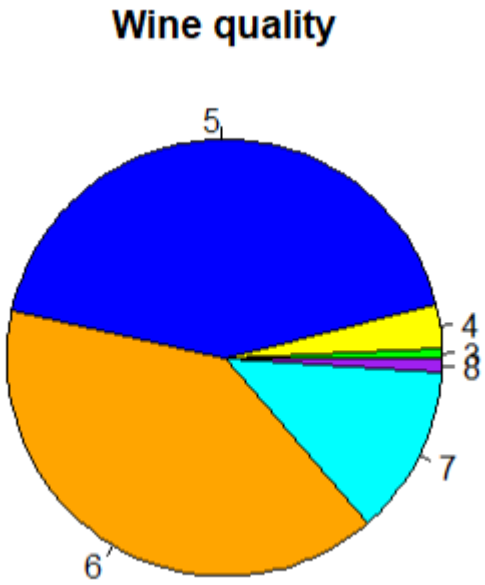
Introduction

Numerous factors affect the quality of the wine. The study is determined to research what are the variables that significantly affect the quality of red wine. So, the dependent variable in this dataset is quality. The study has focused on analyzing the variance of a significant variable that impacts the wine's quality. The correlation between variables will be studied. Another variable will be added to the dataset, which will be based on alcohol. The regression analysis and ANOVA are used as a statistical tool to analyze the dataset.

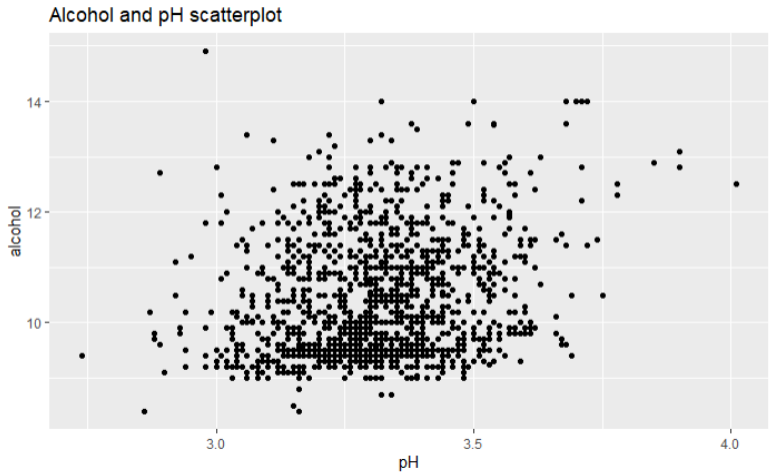
Data Description

The data is obtained from UCI. The information consists of 12 variables as different characteristics of the wine. The variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. A

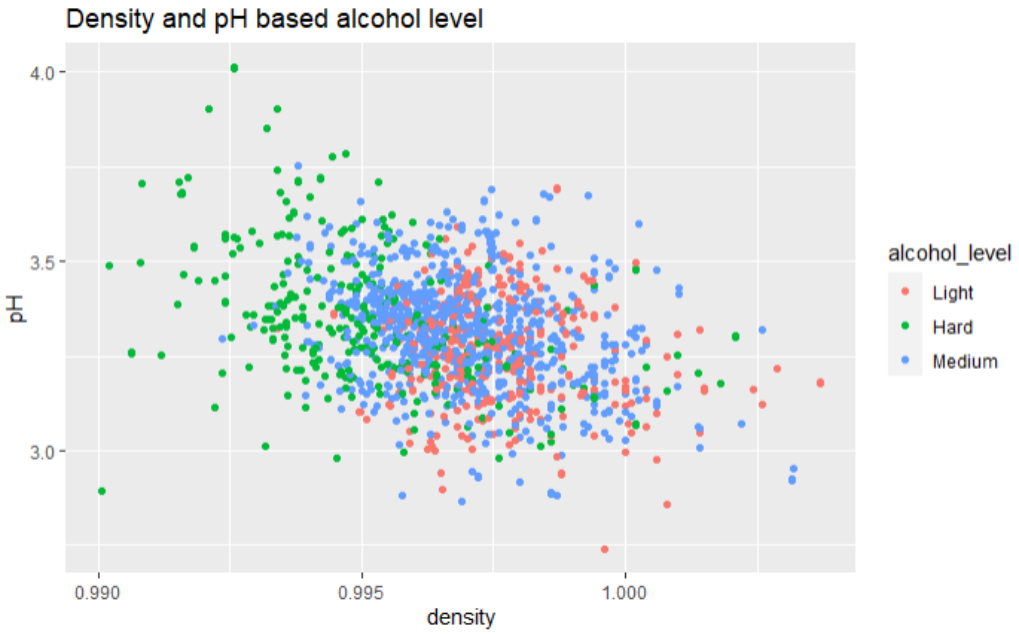
new point has been added in the data based on the quantity of alcohol in the wine. A wine contains less than 9.50 units of alcohol will be categorized as light red wine. A wine having alcohol quantity of between 9.50 and 11.4 will be categorized as medium wine. The wine with an alcohol level of more than 11.4 will be categorized as hard.



There are six different levels of quality of wine ranged from 3 to 8. The 3 indicates the low quality of the wine, whereas 8 indicates the supreme quality of the wine. The above figure shows that most of the dataset samples have the approximately medium quality of the wine, from 5 to 6. Poor and supreme quality wine are very few in numbers. The sample wines with a quality rating of 7 are also quite significant.



The above figure depicts the scatter plot of pH and alcohol of wine. It can be observed that there is a medium positive association between the alcohol level in the wine and pH level. As the pH level increases, the alcohol level also increases.



The above figure shows that there is a negative association between density and pH of red wine. The figure indicates that both the variable pH and density are moderately associated with each other. An increase in pH level is associated with a moderate decrease in density of red wine.

Methods

The multiple linear regression analysis and analysis of variance (ANOVA) will be used to analyze the dependent variable quality of wine and all other numeric variable as independent variables. The result of regression analysis will show which of the given 11 variables impact the quality of wine significantly. The analysis of variance will show whether means of all the variable that significantly impact the quality of wine has equal means or at least one of the means is different.

Results

The obtained result of multiple linear regression shows that volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphate, and alcohol significantly affect the wine's quality. Free sulfate dioxide, sulphates, and alcohol negatively significantly impacts positively to the quality of wine when their quantity increases whereas, volatile acidity, chlorides, total sulfur dioxide, and pH decrease the quality of wine when their quantity increases. The ANOVA shows that at least one mean different as the p-value is less than 0.05.

Conclusion

The quality of the wine is significantly dependent upon the pH, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, sulphate, and alcohol. This implies that red wine producing

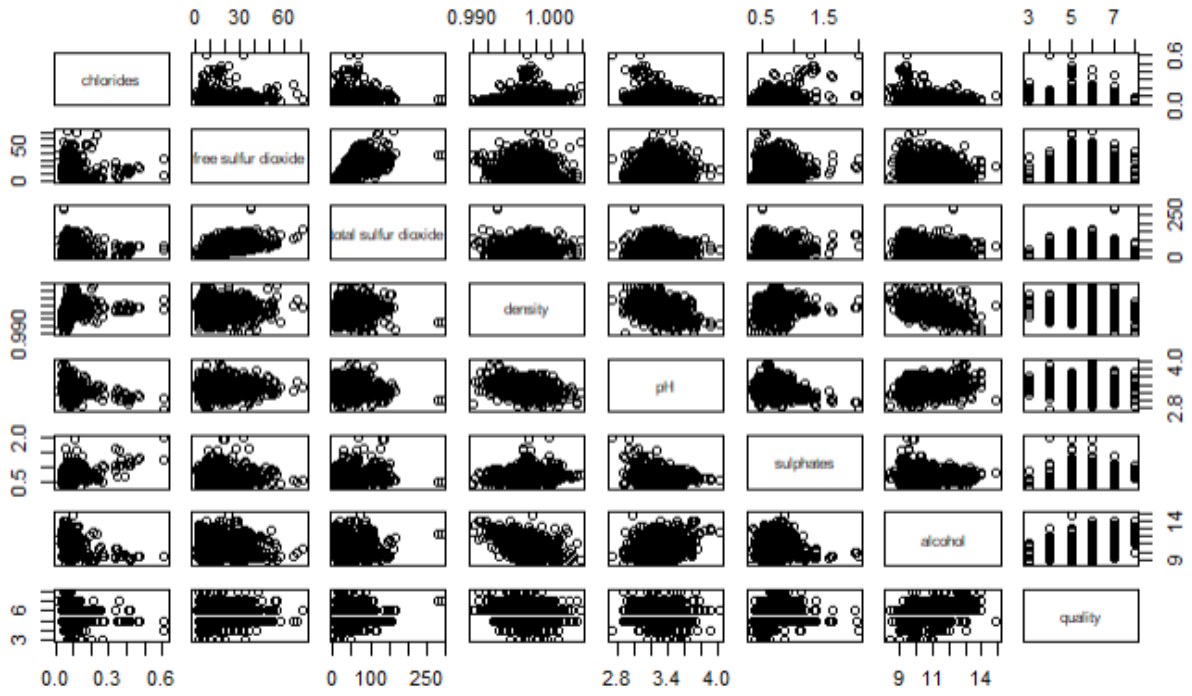
firms should maintain the quality and quantity of these commodities in the wine to keep the wine's quality. More free sulfur dioxide, sulphates, and alcohol are better for keeping quality up. Less volatile acidity, chlorides, total sulfur dioxide, and pH is better for keeping quality high.

Appendix

Descriptive statistics

```
fixed acidity   volatile acidity  citric acid    residual sugar     chlorides       free sulfur dioxide
Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900   Min.    :0.01200   Min.    : 1.00
1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00
Mean    : 8.32   Mean    :0.5278   Mean    :0.271   Mean    : 2.539   Mean    :0.08747   Mean    :15.87
3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
Max.    :15.90   Max.    :1.5800   Max.    :1.000   Max.    :15.500   Max.    :0.61100   Max.    :72.00
total sulfur dioxide    density             pH          sulphates         alcohol        quality
Min.    :  6.00       Min.    :0.9901   Min.    :2.740   Min.    :0.3300   Min.    : 8.40   Min.    :3.000
1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
Mean    : 46.47       Mean    :0.9967   Mean    :3.311   Mean    :0.6581   Mean    :10.42   Mean    :5.636
3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
Max.    :289.00       Max.    :1.0037   Max.    :4.010   Max.    :2.0000   Max.    :14.90   Max.    :8.000
```

Wine Quality Correlation



New Variable

```
 Light   Hard Medium
   436    312    851
```

Linear Regression

```
Call:
lm(formula = quality ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68911 -0.36652 -0.04699  0.45202  2.02498

Coefficients:
                       Estimate  Std. Error t value             Pr(>|t|)
(Intercept)          21.9652085  21.1945750   1.036                0.3002
`fixed acidity`       0.0249906   0.0259485   0.963                0.3357
`volatile acidity`   -1.0835903   0.1211013  -8.948 < 0.0000000000000002 ***
`citric acid`        -0.1825639   0.1471762  -1.240                0.2150
`residual sugar`      0.0163313   0.0150021   1.089                0.2765
chlorides            -1.8742252   0.4192832  -4.470  0.00000837395338495 ***
`free sulfur dioxide` 0.0043613   0.0021713   2.009                0.0447 *
`total sulfur dioxide` -0.0032646  0.0007287  -4.480  0.00000800460981496 ***
density             -17.8811638  21.6330999  -0.827                0.4086
pH                   -0.4136531   0.1915974  -2.159                0.0310 *
sulphates             0.9163344   0.1143375   8.014  0.00000000000000213 ***
alcohol               0.2761977   0.0264836  10.429 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.648 on 1587 degrees of freedom
Multiple R-squared:  0.3606,    Adjusted R-squared:  0.3561
F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 0.00000000000000022
```

Linear regression with significant variables

```
Call:
lm(formula = quality ~ df$`volatile acidity` + chlorides + df$`free sulfur dioxide` +
    df$`total sulfur dioxide` + pH + sulphates + alcohol, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                           Estimate Std. Error t value             Pr(>|t|)
(Intercept)                4.4300987  0.4029168  10.995 < 0.0000000000000002 ***
df$`volatile acidity`     -1.0127527  0.1008429 -10.043 < 0.0000000000000002 ***
chlorides                 -2.0178138  0.3975417  -5.076  0.00000043137165892 ***
df$`free sulfur dioxide`   0.0050774  0.0021255   2.389                0.017 *
df$`total sulfur dioxide` -0.0034822  0.0006868  -5.070  0.00000044348335492 ***
pH                        -0.4826614  0.1175581  -4.106  0.00004234962312451 ***
sulphates                  0.8826651  0.1099084   8.031  0.00000000000000186 ***
alcohol                    0.2893028  0.0167958  17.225 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 0.00000000000000022
```