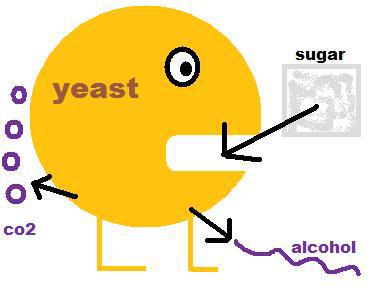
**Dataset**

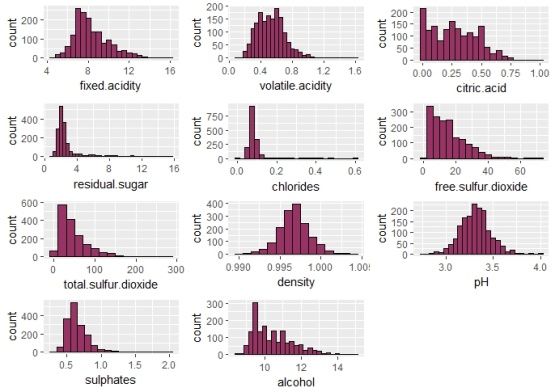
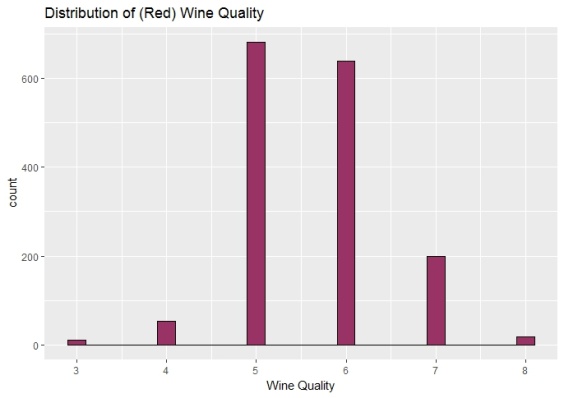


The dataset consists of 12 measurements on 6,497 samples of white and red styles of *Vinho Verde* wines. There are NO missing data. Descriptions of each variable are below.

|  |  |
| --- | --- |
| Variable | Description |
| fixed\_acidity | Naturally occurring acid, higher levels yield fresher or more tart wines |
| volatile\_acidity | Acid produced during fermentation, leads to unpleasant aromas |
| citric\_acid | Inexpensive supplement to boost total acidity |
| residual\_sugar | Remaining sugar after fermentation or sometimes added, higher levels yield sweeter wines |
| chlorides | Saltiness, often just barely detectable |
| pH | Higher levels make wines taste softer, inversely related to acidity |
| alcohol | Produced from yeast and sugar during fermentation, higher levels associated with more complex wines |
| density | Lower for "dry" wines, higher for sweet wines |
| free\_sulfur\_dioxide,  total\_sulfur\_dioxide | Antibiotics/antioxidants to protect wine and sanitize wineries; detectable as pungent odor at higher levels and responsible for label “contains sulfites” |
| sulphates | Potassium sulphate used to lower pH, raise acidity and intensify color in red wines |
| quality | Median of ratings on a 0 to 10 (poor to excellent) scale from three or more experts who participated in blind tastings |
| style | Red or white |

**Exploratory Results:**

**Redwine-**



**Fig – 1:** Redwine – Distruibution of wine quality and the distribution of individual componenets

Red wine data has 1599 observations with 11 variables + quality (12th) as target variable.

Wine quality is a discrete variable. Its value ranging from 3 to 8. Median value is at 6.

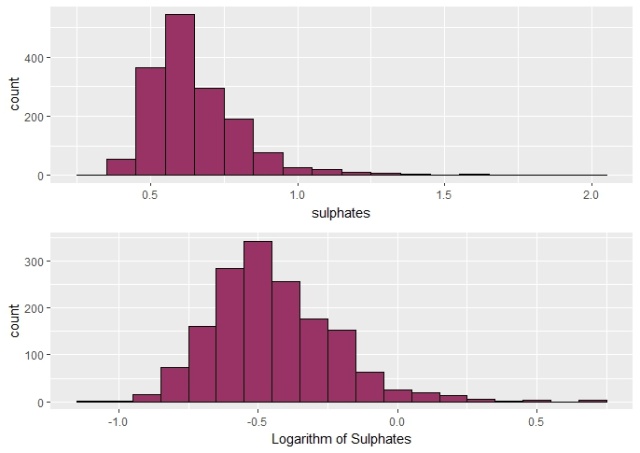
From the individual distribution of data we observe that-

1. Normal distribution: volatile.acidity, density, pH

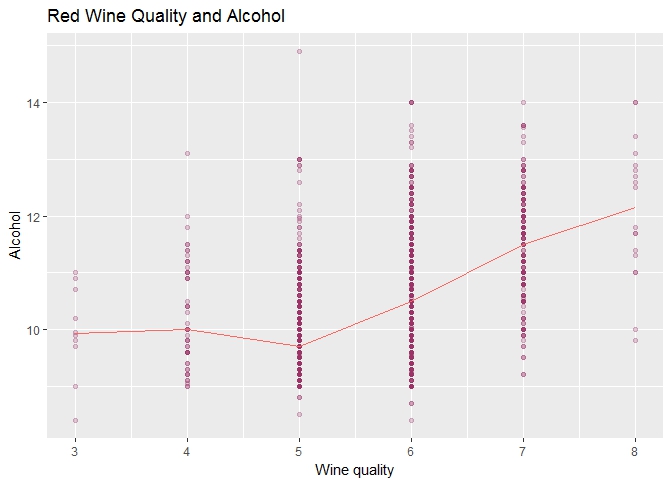
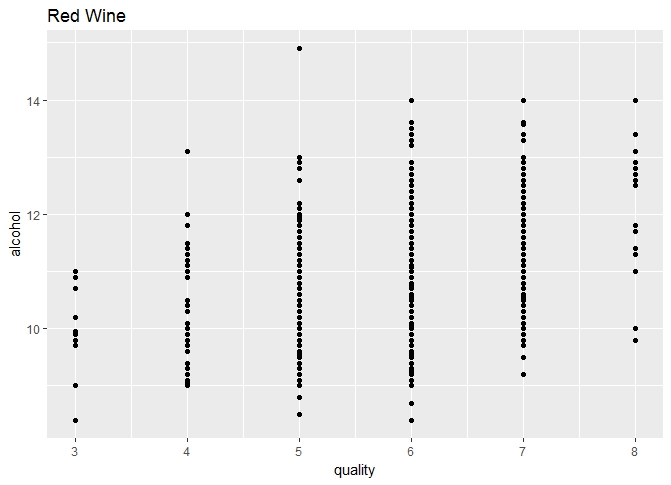
2. Positive Skewed: fixed.acidity, citric.acid, free.sulfur.dioxide, total.sulfur.dioxide,sulphates,alcohol

3. Long Tail: choloride, residual.sugar

Now rescale variable the sulphate variable by taking "log" of the data-



**Fig – 2:** Redwine –Distribution of rescaled variable sulphate



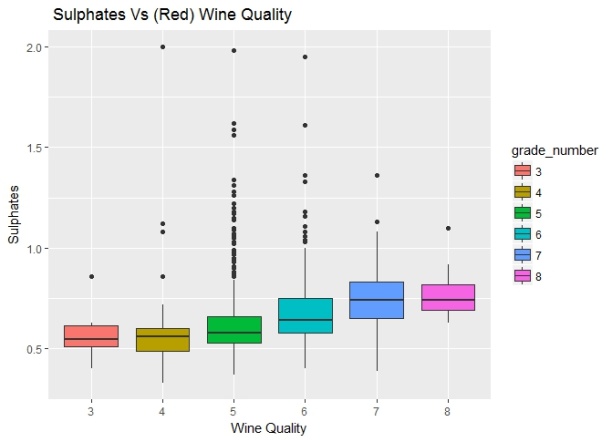
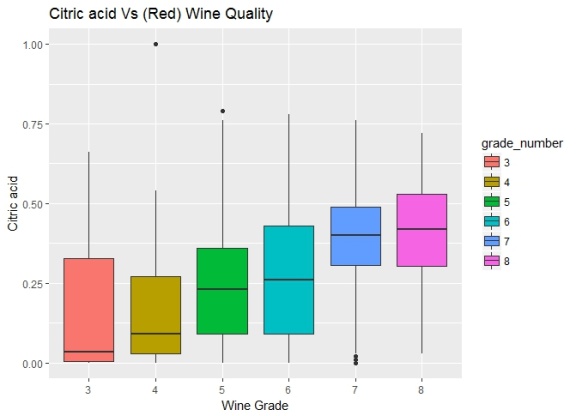
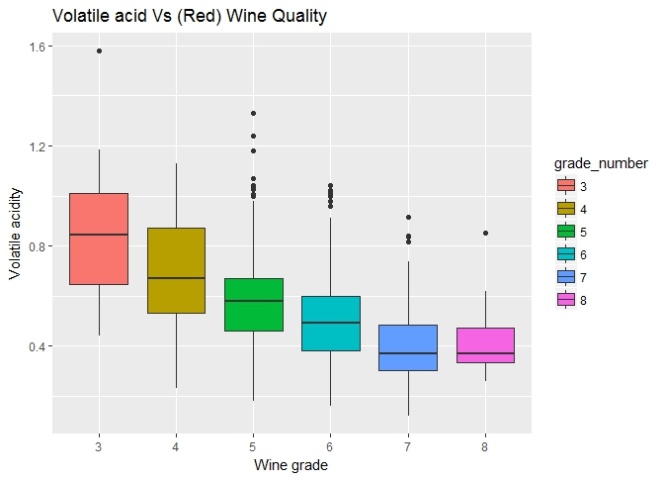
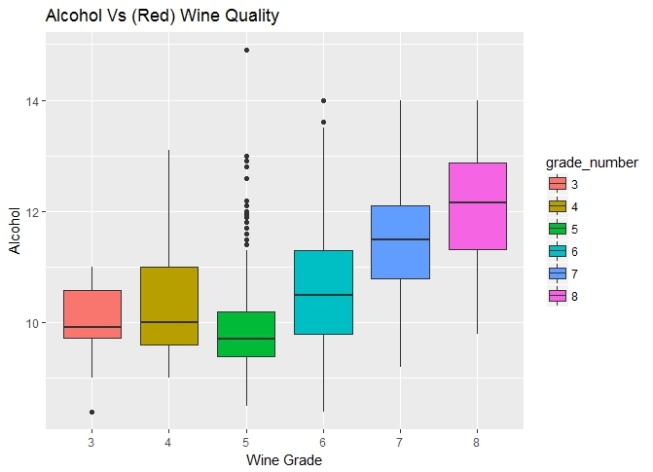
**Fig – 3:** Redwine – Scatterplot of Wine Quality Vs Alcohol with the median value

As you can see, higher quality wines have higher value of alcohol content.

**Correlation matrix-** Below is the table of correlation between the wine quality and the other chemical components in the decreasing order.

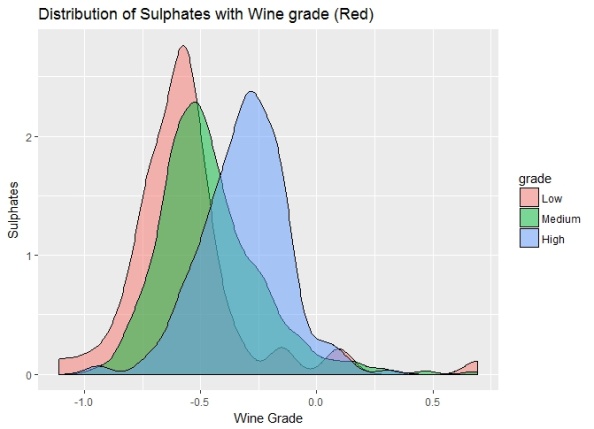
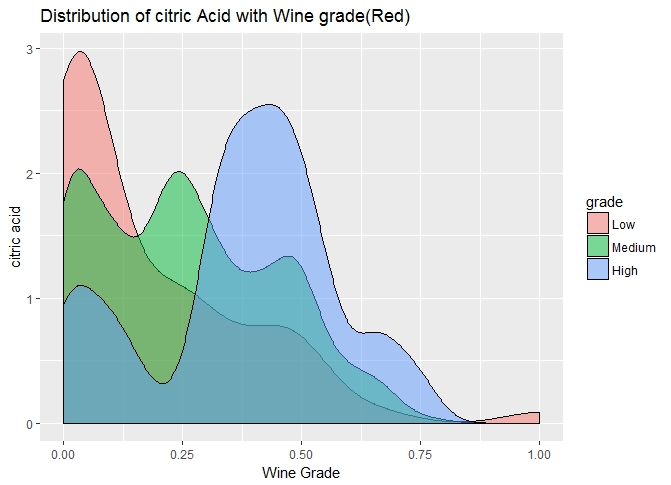
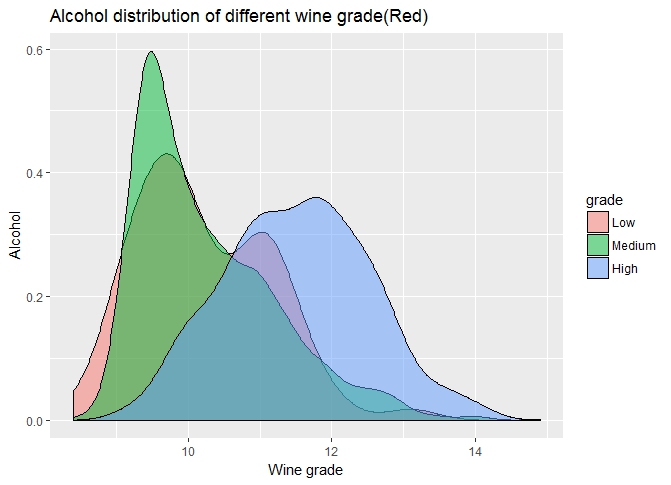
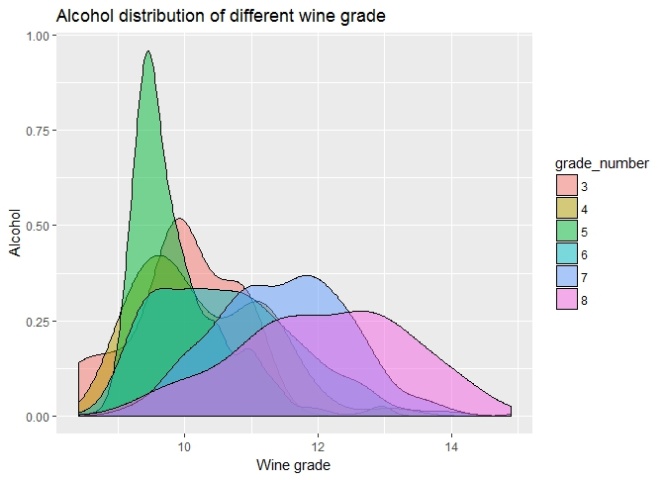
So the alcohol, volatile acidity, sulphates and citric acid are the most correlated variable with the wine quality.

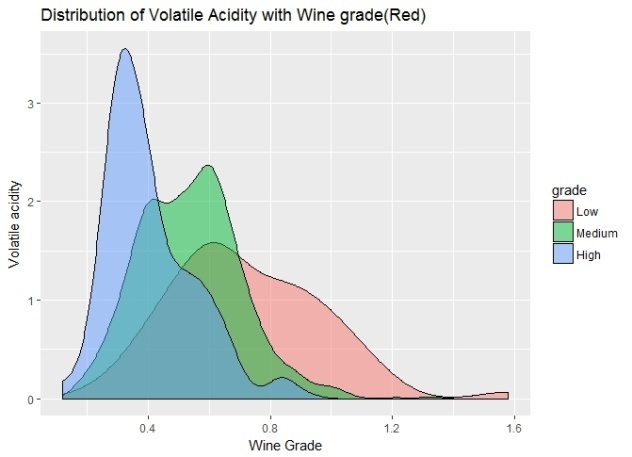
|  |  |
| --- | --- |
| **Chemical components** | **Correlation coefficient value with Red wine quality** |
| alcohol | **0.47616632** |
| volatile.acidity | **-0.39055778** |
| sulphates | **0.25139708** |
| citric.acid | **0.22637251** |
| total.sulfur.dioxide | -0.18510029 |
| density | -0.17491923 |
| chlorides | -0.12890656 |
| fixed.acidity | 0.12405165 |
| pH | -0.05773139 |
| free.sulfur.dioxide | 0.05065606 |
| residual.sugar | 0.01373164 |



**Fig – 4:** Redwine – Boxplot of highly correlated variables with Wine Quality

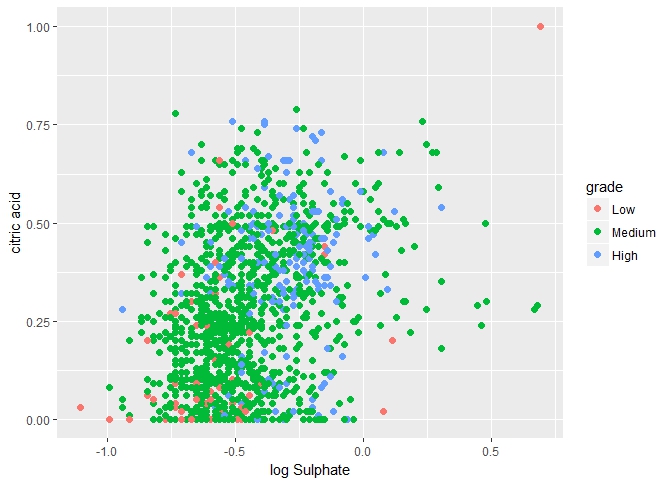
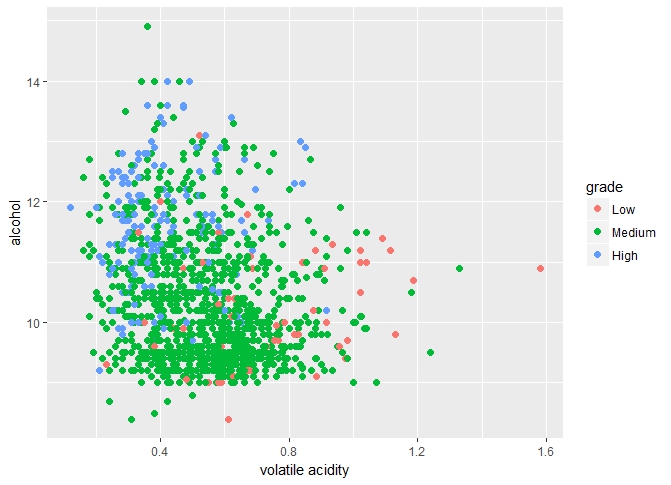
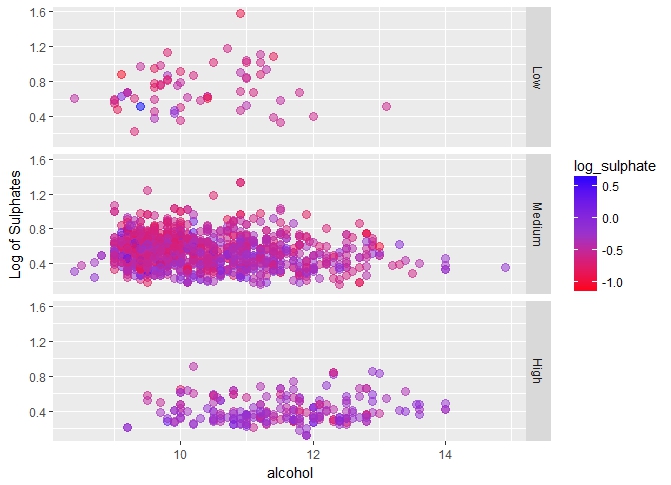
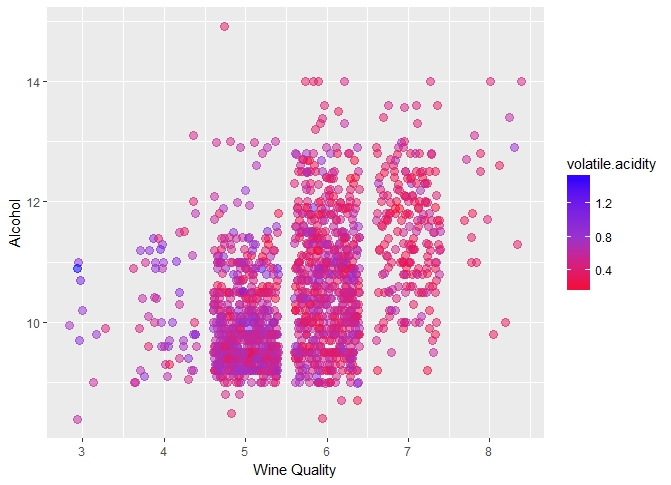
It is clear from the above graphs that the higher quality wine have relatively higher alcohol content, higher value of citric acid and higher values if sulphates with the lower values of volatile acidity.





**Fig – 5:** Redwine – Density plot of highly correlated variables with Wine Quality

Similar observations can be deduced from here as well.



**Fig – 6:** Redwine – Scatter plot between highly correlated variables and Wine Quality with multi dimensional approach

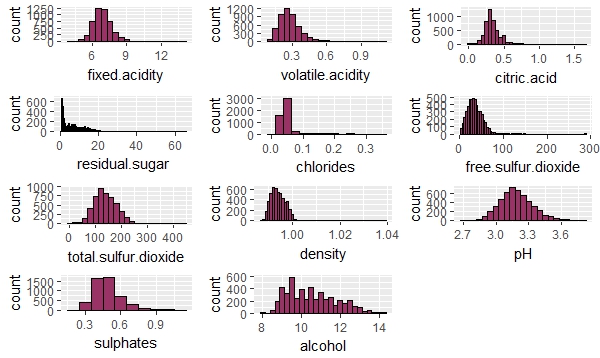
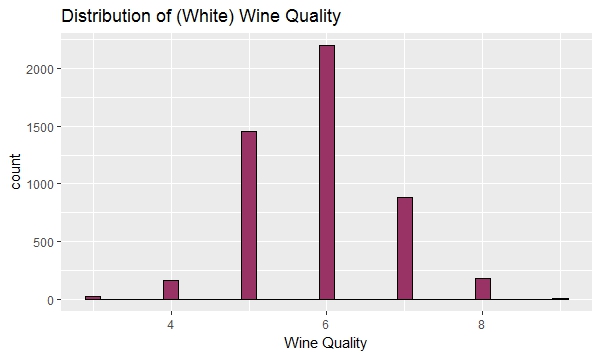
**Modeling -**

Linear model is prepared to get the quality of wine – The model with 6 features (m6) has the lowest AIC (akaike information criterion) number. As the number of feature increase AIC become higher. Also the parameter of the predictor also changed.

RedWine\_quality = 2.985 - 1.104\*volatile.acidity + 0.276\*alcohol + 0.908\*sulphates + 0.065\*citric.acid - 1.763\*chlorides - 0.002\*total.sulfur.dioxide

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Outcomes of Linear model for RedWine** | | | | | | | |
|  | **m1** | **m2** | **m3** | **m4** | **m5** | **m6** | **m7** |
| **(Intercept)** | 6.566\*\*\* | 3.095\*\*\* | 2.611\*\*\* | 2.646\*\*\* | 2.769\*\*\* | 2.985\*\*\* | -0.953 |
| **(std error)** | -0.058 | -0.184 | -0.196 | -0.201 | -0.202 | -0.206 | -11.99 |
| **volatile.acidity** | -1.761\*\*\* | -1.384\*\*\* | -1.221\*\*\* | -1.265\*\*\* | -1.155\*\*\* | -1.104\*\*\* | -1.114\*\*\* |
|  | -0.104 | -0.095 | -0.097 | -0.113 | -0.115 | -0.115 | -0.12 |
| **alcohol** |  | 0.314\*\*\* | 0.309\*\*\* | 0.309\*\*\* | 0.292\*\*\* | 0.276\*\*\* | 0.280\*\*\* |
|  |  | -0.016 | -0.016 | -0.016 | -0.016 | -0.017 | -0.02 |
| **sulphates** |  |  | 0.679\*\*\* | 0.696\*\*\* | 0.871\*\*\* | 0.908\*\*\* | 0.903\*\*\* |
|  |  |  | -0.101 | -0.103 | -0.111 | -0.111 | -0.112 |
| **citric.acid** |  |  |  | -0.079 | 0.021 | 0.065 | 0.044 |
|  |  |  |  | -0.104 | -0.106 | -0.106 | -0.124 |
| **chlorides** |  |  |  |  | -1.663\*\*\* | -1.763\*\*\* | -1.747\*\*\* |
|  |  |  |  |  | -0.405 | -0.403 | -0.406 |
| **total.sulfur.dioxide** |  |  |  |  |  | -0.002\*\*\* | -0.002\*\*\* |
|  |  |  |  |  |  | -0.001 | -0.001 |
| **density** |  |  |  |  |  |  | 3.923 |
|  |  |  |  |  |  |  | -11.944 |
|  |  |  |  |  |  |  |  |
| **R-squared** | 0.153 | 0.317 | 0.336 | 0.336 | 0.343 | **0.352** | 0.352 |
| **adj. R-squared** | 0.152 | 0.316 | 0.335 | 0.334 | 0.341 | **0.349** | 0.349 |
| **sigma** | 0.744 | 0.668 | 0.659 | 0.659 | 0.656 | 0.651 | 0.652 |
| **F** | 287.444 | 370.379 | 268.912 | 201.777 | 166.407 | 143.91 | 123.298 |
| **p** | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Log-likelihood** | -1794.31 | -1621.814 | -1599.38 | -1599.09 | -1590.662 | -1580.192 | -1580.138 |
| **Deviance** | 883.198 | 711.796 | 692.105 | 691.852 | 684.595 | 675.689 | 675.643 |
| **AIC** | 3594.624 | 3251.628 | 3208.768 | 3210.186 | 3195.324 | **3176.384** | 3178.276 |
| **BIC** | 3610.756 | 3273.136 | 3235.654 | 3242.448 | 3232.964 | 3219.401 | 3226.67 |
| **N** | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 | 1599 |

**WhiteWine-**

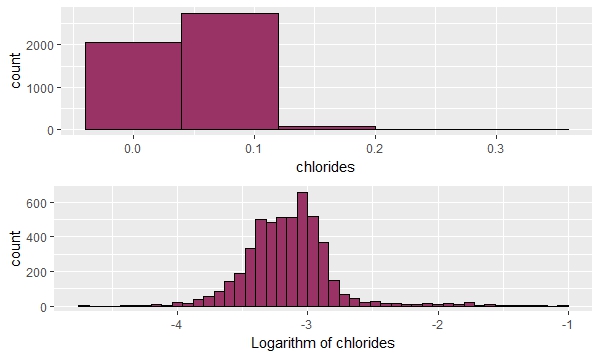


**Fig – 7:** Whitewine – Distruibution of wine quality and the distribution of individual componenets

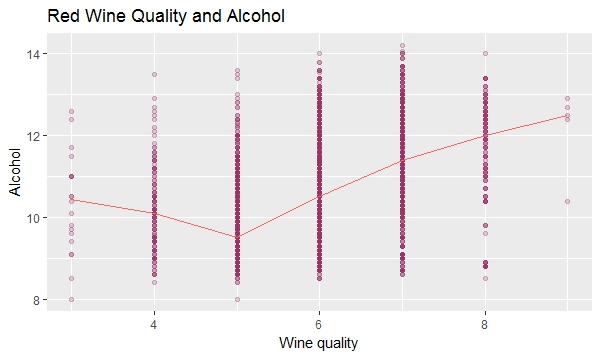
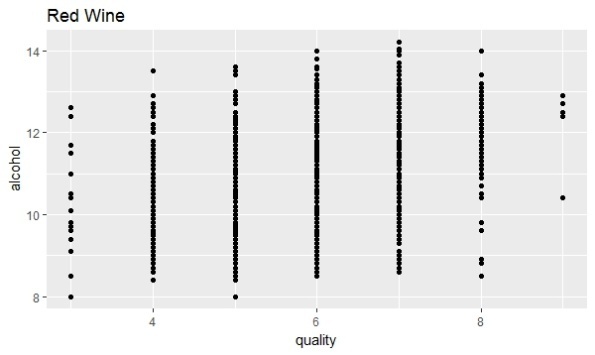
White wine data has 4898 observations with 11 variables + quality (12th) as target variable.

Wine quality is a discrete variable. Its value ranging from 3 to 9. Median value is at 6.

Now rescale variable - chlorides by taking "log" of the data-



**Fig – 8:** WhiteWine –Distribution of rescaled variable chlorides



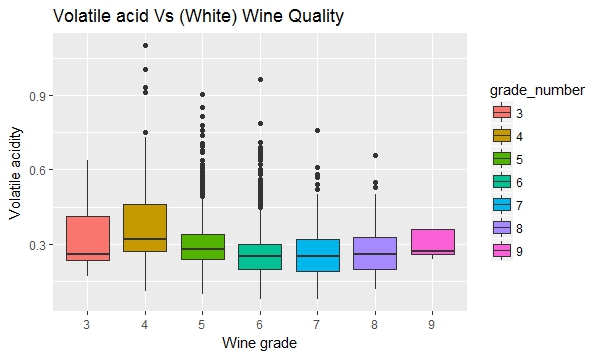
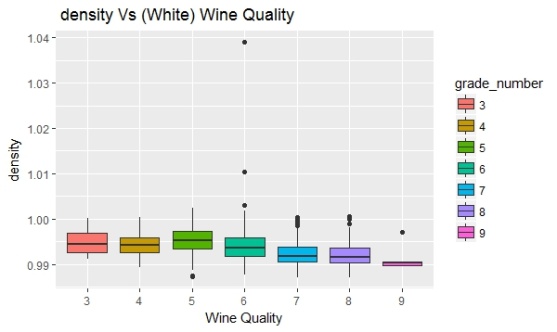
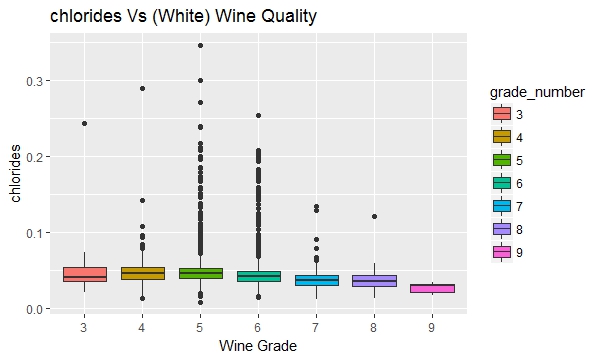
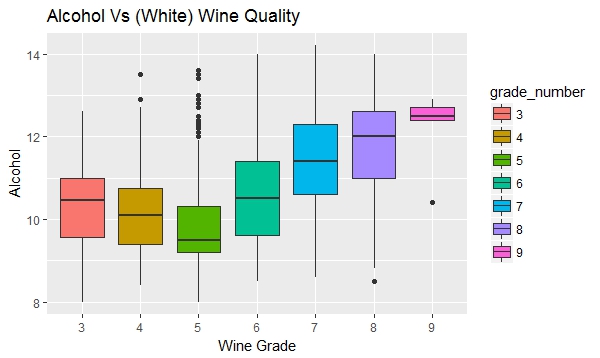
**Fig – 9:** Whitewine – Scatterplot between Wine Quality and Alcohol with the median value

As you can see, higher quality wines have higher value of alcohol content.

**Correlation matrix-** Below is the table of correlation between the wine quality and the other chemical components in the decreasing order.

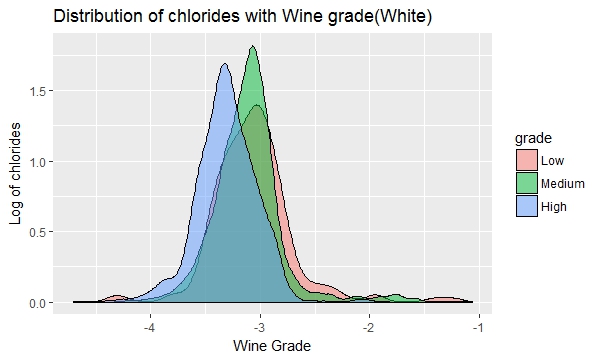
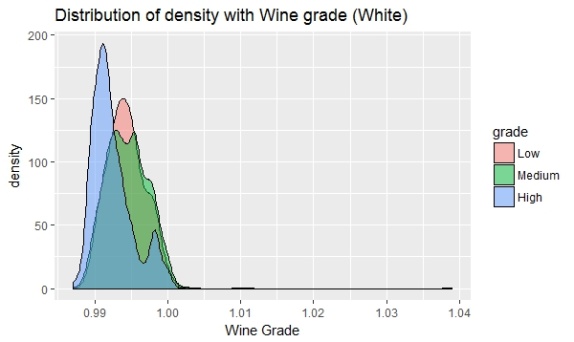
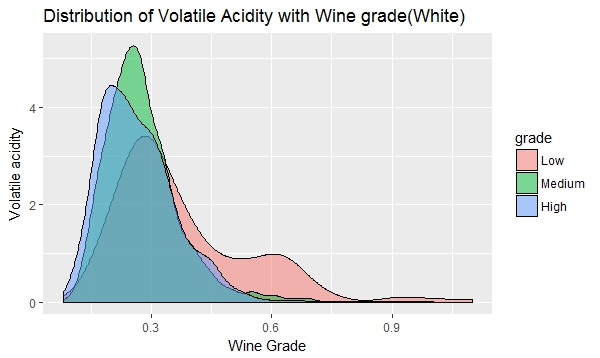
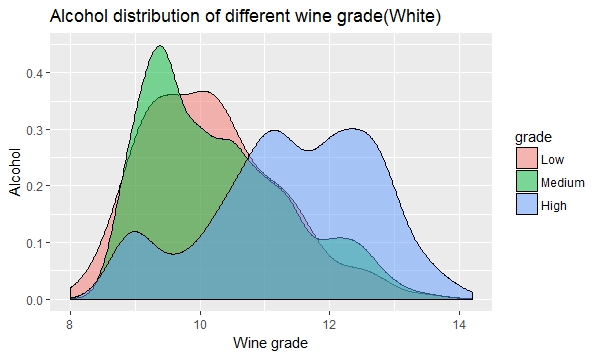
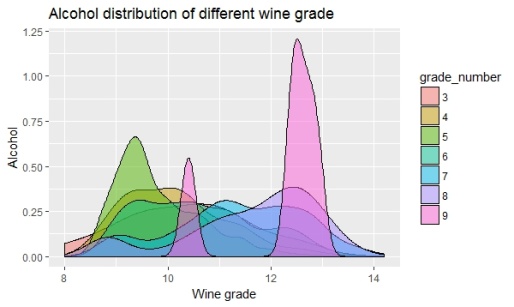
So the alcohol, volatile acidity, chlorides and density are the most correlated variable with the wine quality.

|  |  |
| --- | --- |
| **Chemical componets** | **Correlation coefficient value with White wine quality** |
| Alcohol | **0.435574715** |
| Density | **-0.307123313** |
| Chlorides | **-0.209934411** |
| volatile.acidity | **-0.194722969** |
| total.sulfur.dioxide | -0.174737218 |
| fixed.acidity | -0.113662831 |
| pH | 0.099427246 |
| residual.sugar | -0.097576829 |
| Sulphates | 0.053677877 |
| citric.acid | -0.009209091 |
| free.sulfur.dioxide | 0.008158067 |



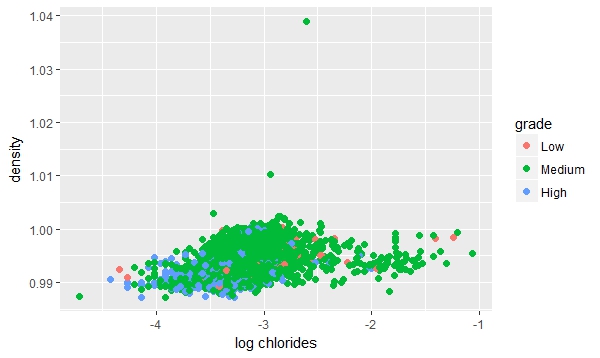
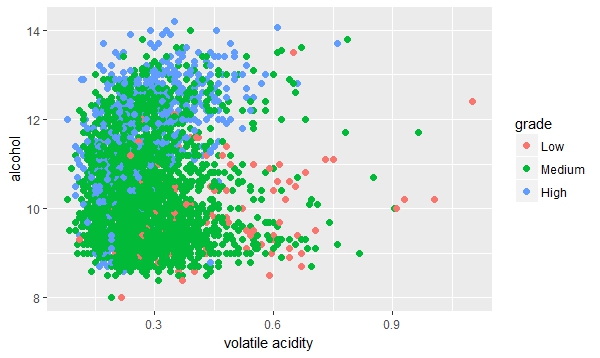
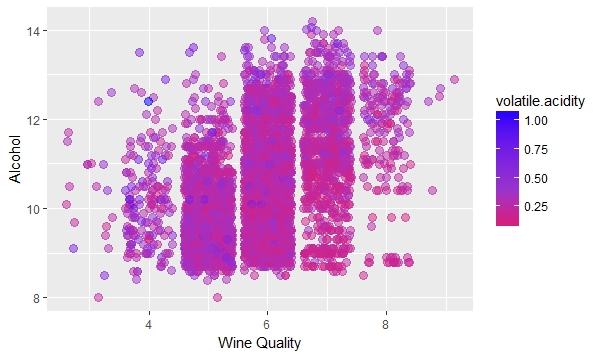
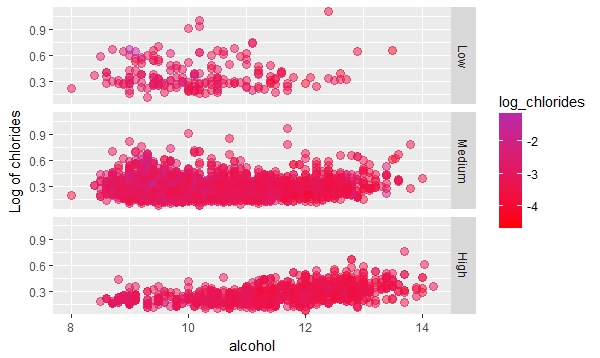
**Fig – 10:** Whitewine – Boxplot of highly correlated variables with Wine Quality

It is clear from the above graphs that the higher quality wine have relatively higher alcohol content with lower density, lower chloride and lower values of volatile acidity.



**Fig – 11:** Whitewine – Density plot of highly correlated variables with Wine Quality

Similar observations can be deduced from here as well.



**Fig – 12:** Whitewine – Scatter plot between highly correlated variables and Wine Quality with multi dimensional approach

**Modeling -**

Linear model is prepared to get the quality of wine – The model with 8 features (m8) has the lowest AIC (akaike information criterion) number. As the number of feature increase AIC become higher. Also the parameter of the predictor also changed.

WhiteWine\_quality = 99.527 -2.047 \*volatile.acidity + 0.261 \*alcohol + 0.629 \*sulphates -0.056\*citric.acid -0.693 \*chlorides + 0.001 \*total.sulfur.dioxide – 97.153 density +0.060 \* residual.sugar

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Outcomes of Linear model for WhiteWine** | | | | | | | | |
|  | **m1** | **m2** | **m3** | **m4** | **m5** | **m6** | **m7** | **m8** |
| **(Intercept)** | 6.354\*\*\* | 3.017\*\*\* | 2.803\*\*\* | 2.840\*\*\* | 2.988\*\*\* | 2.699\*\*\* | -31.200\*\*\* | 99.527\*\*\* |
| **(std error)** | -0.036 | -0.098 | -0.11 | -0.116 | -0.128 | -0.148 | -6.359 | -12.743 |
| **volatile.acidity** | -1.711\*\*\* | -1.979\*\*\* | -1.963\*\*\* | -1.978\*\*\* | -1.942\*\*\* | -2.010\*\*\* | -2.085\*\*\* | -2.047\*\*\* |
|  | -0.123 | -0.11 | -0.11 | -0.111 | -0.111 | -0.113 | -0.113 | -0.112 |
| **alcohol** |  | 0.324\*\*\* | 0.325\*\*\* | 0.324\*\*\* | 0.315\*\*\* | 0.333\*\*\* | 0.391\*\*\* | 0.261\*\*\* |
|  |  | -0.009 | -0.009 | -0.009 | -0.01 | -0.011 | -0.015 | -0.019 |
| **sulphates** |  |  | 0.416\*\*\* | 0.421\*\*\* | 0.423\*\*\* | 0.369\*\*\* | 0.340\*\*\* | 0.629\*\*\* |
|  |  |  | -0.097 | -0.097 | -0.097 | -0.098 | -0.097 | -0.099 |
| **citric.acid** |  |  |  | -0.089 | -0.061 | -0.101 | -0.173 | -0.056 |
|  |  |  |  | -0.092 | -0.093 | -0.093 | -0.094 | -0.093 |
| **chlorides** |  |  |  |  | -1.486\*\* | -1.524\*\* | -1.279\* | -0.693 |
|  |  |  |  |  | -0.546 | -0.545 | -0.546 | -0.541 |
| **total.sulfur.dioxide** |  |  |  |  |  | 0.001\*\*\* | 0.001\* | 0.001\* |
|  |  |  |  |  |  | 0 | 0 | 0 |
| **density** |  |  |  |  |  |  | 33.607\*\*\* | -97.153\*\*\* |
|  |  |  |  |  |  |  | -6.303 | -12.718 |
| **residual.sugar** |  |  |  |  |  |  |  | 0.060\*\*\* |
|  |  |  |  |  |  |  |  | -0.005 |
|  |  |  |  |  |  |  |  |  |
| **R-squared** | 0.038 | 0.24 | 0.243 | 0.243 | 0.244 | 0.247 | 0.251 | **0.272** |
| **adj. R-squared** | 0.038 | 0.24 | 0.243 | 0.243 | 0.244 | 0.246 | 0.25 | **0.271** |
| **sigma** | 0.869 | 0.772 | 0.771 | 0.771 | 0.77 | 0.769 | 0.767 | 0.756 |
| **F** | 192.958 | 773.875 | 523.939 | 393.178 | 316.435 | 267.046 | 234.243 | 228.101 |
| **p** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Log-likelihood** | -6259.95 | -5681.776 | -5672.52 | -5672.06 | -5668.358 | -5660.642 | -5646.443 | -5577.85 |
| **Deviance** | 3695.351 | 2918.264 | 2907.258 | 2906.713 | 2902.319 | 2893.189 | 2876.464 | 2797.013 |
| **AIC** | 12525.9 | 11371.552 | 11355.04 | 11356.13 | 11350.716 | 11337.284 | 11310.886 | **11175.7** |
| **BIC** | 12545.39 | 11397.538 | 11387.53 | 11395.11 | 11396.192 | 11389.256 | 11369.355 | 11240.66 |
| **N** | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 | 4898 |