Statistics for Data Science

White Wine Quality Analysis Report

Sumanth M

S20150010033

CSE, Final Year, IIIT Sri City

Problem Statement

The goal is to model Wine Quality based on physicochemical tests. Quality is based on scores where each was graded with the quality between 3 (very bad) and 9 (very excellent). In short, Exploration and Analysis of Wine Quality.

Data Set Description

The given data set is a Multivariate Data Set consists of 4898 Observations along with 12 Variables. Among 12 variables, there are 11 Independent Variables and 1 Dependent Variable. It is given that there are no Missing Values. The list of Independent Variables is as follows :

X_1 : Fixed Acidity X_2 :Volatile Acidity X_3 :Citric Acid X_4 Residual Sugar

X_5 :Chlorides X_6 :Free Sulfur Dioxide X_7 :Total Sulfur Dioxide X_8 :Density

X_9 :pH X_{10} :Sulphates X_{11} :Alcohol

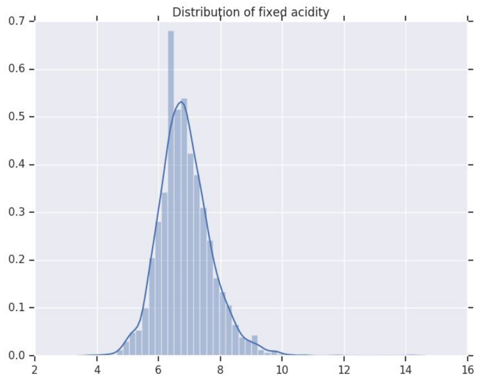
Dependent Variable : Quality

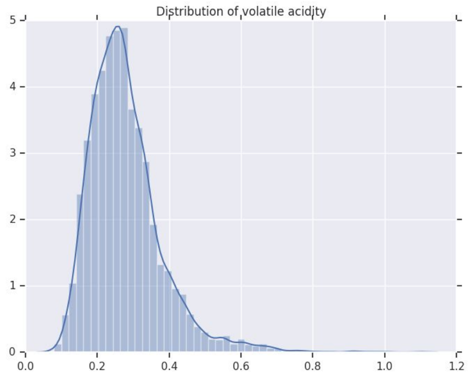
Objectives

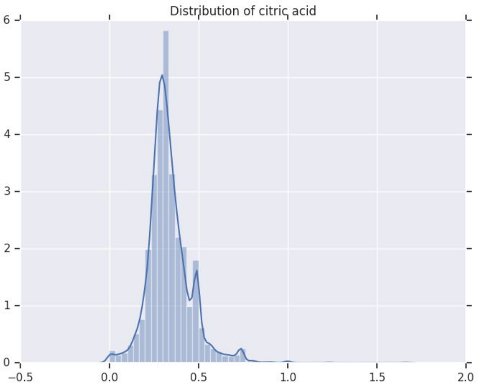
- Descriptive Statistics
- Performing Multiple Linear Regression
- Model Adequacy Tests
- Model Diagnostics
- Principal Component Analysis
- 2D Clustering , Plots from other aspects and Confidence Regions for β's

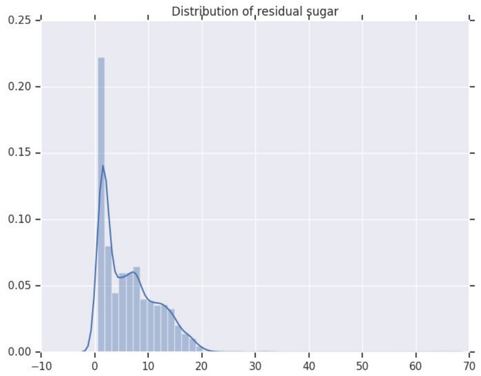
Descriptive Statistics

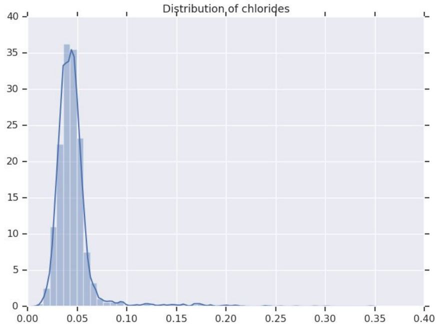
Distribution of Variables : Each variable in the dataset is distributed normally. Below are the distribution figures for variables in the mentioned order above.

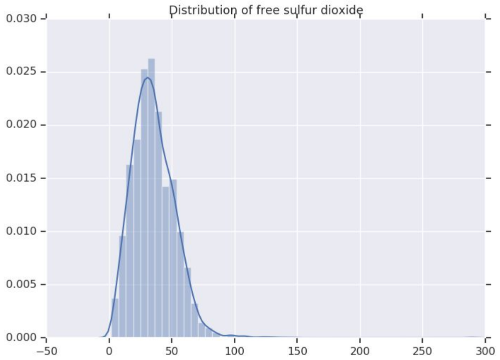


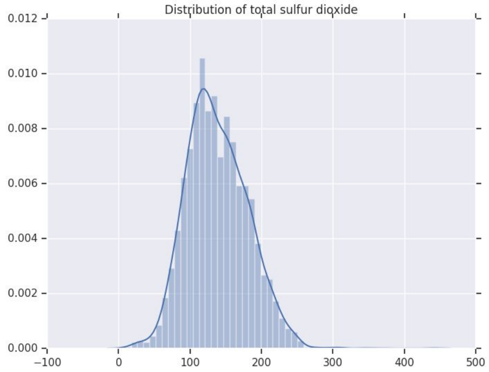


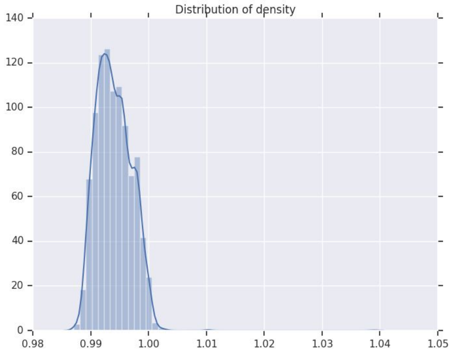


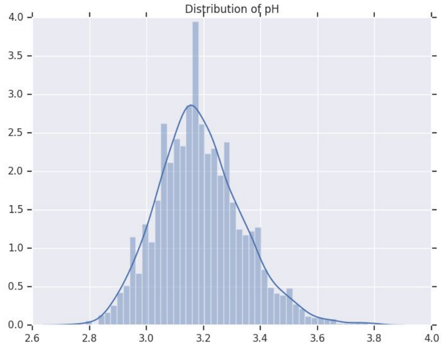


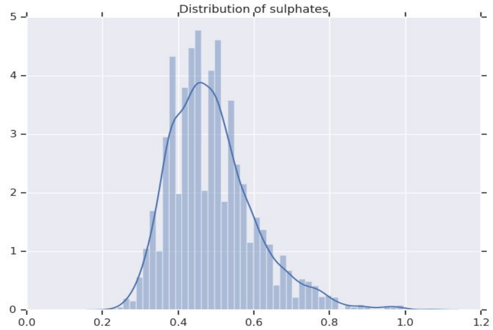

Distribution of chlorides
40
35
30
25
20
15
10
5
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

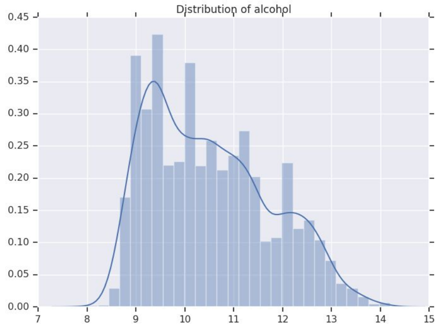

Distribution of free sulfur dioxide
0.030
0.025
0.020
0.015
0.010
0.005
0.000
−50
0
50
100
150
200
250
300

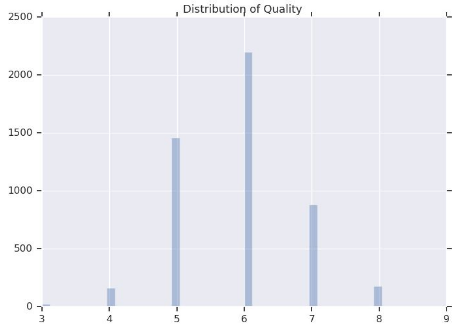

Distribution of total sulfur dioxide
0.012
0.010
0.008
0.006
0.004
0.002
0.000
−100
0
100
200
300
400
500


Distribution of density
140
120
100
80
60
40
20
0
0.98
0.99
1.00
1.01
1.02
1.03
1.04
1.05

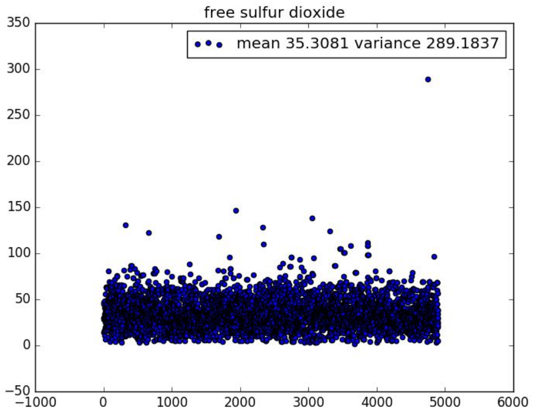

Distribution of pH
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0

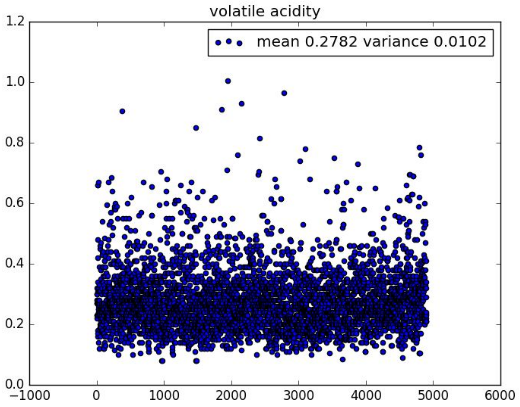

Distribution of sulphates
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2

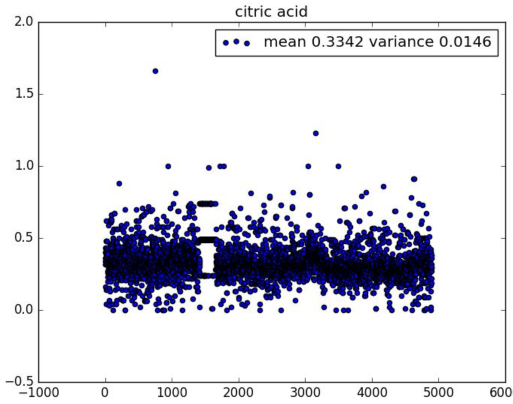


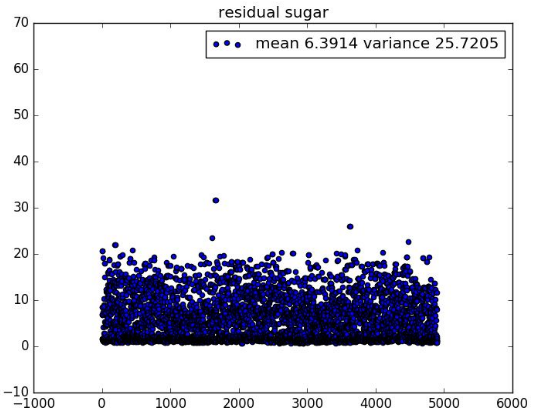


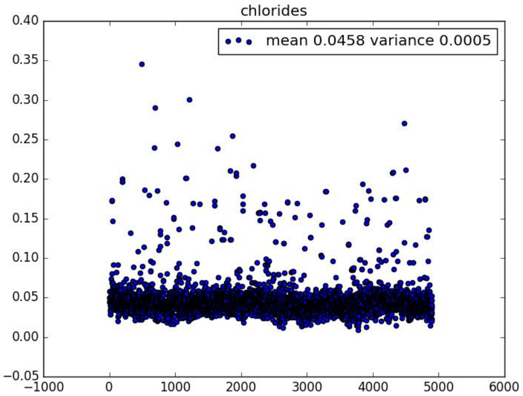
Scatter Plot : Displaying the location of each observation in the feature space of the independent variable in the given dataset.

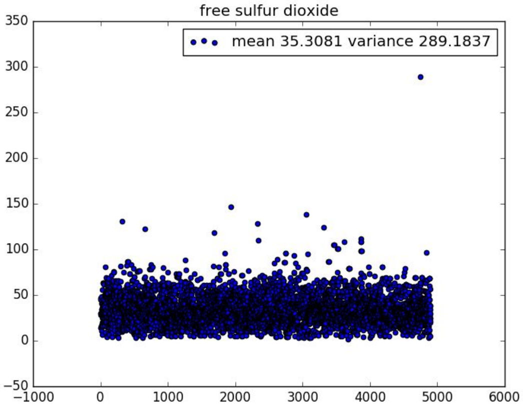


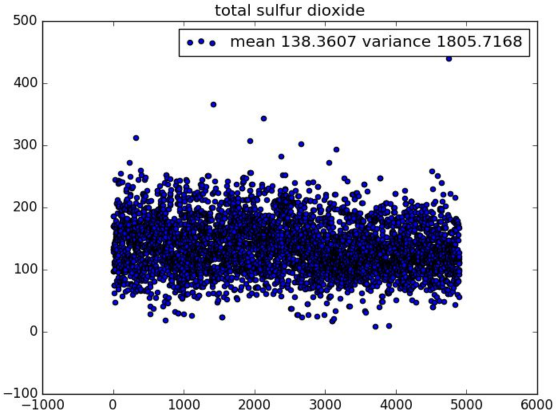


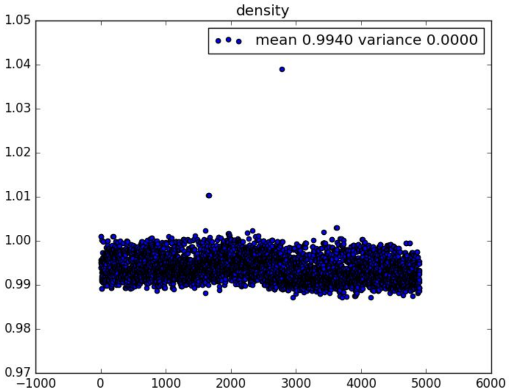


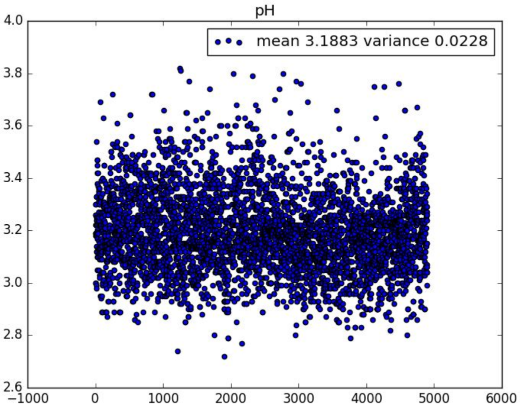


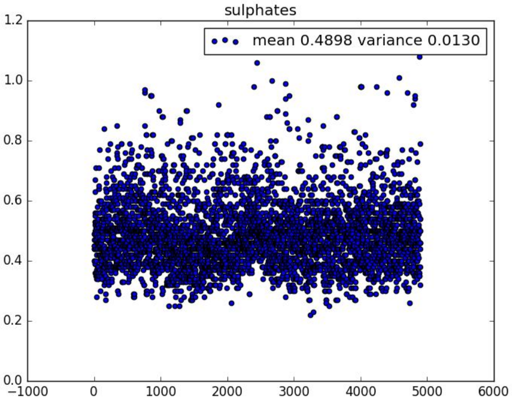

chlorides

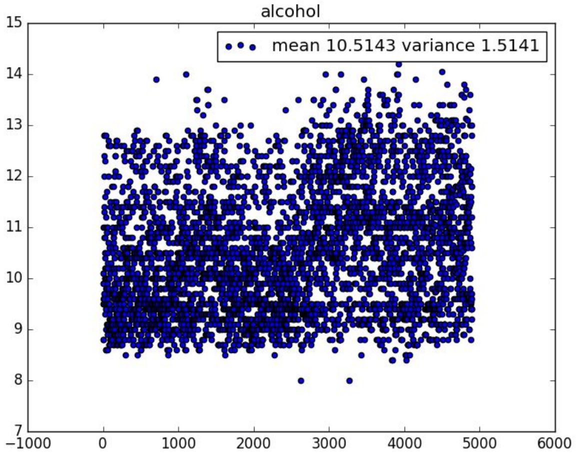

free sulfur dioxide


total sulfur dioxide


density

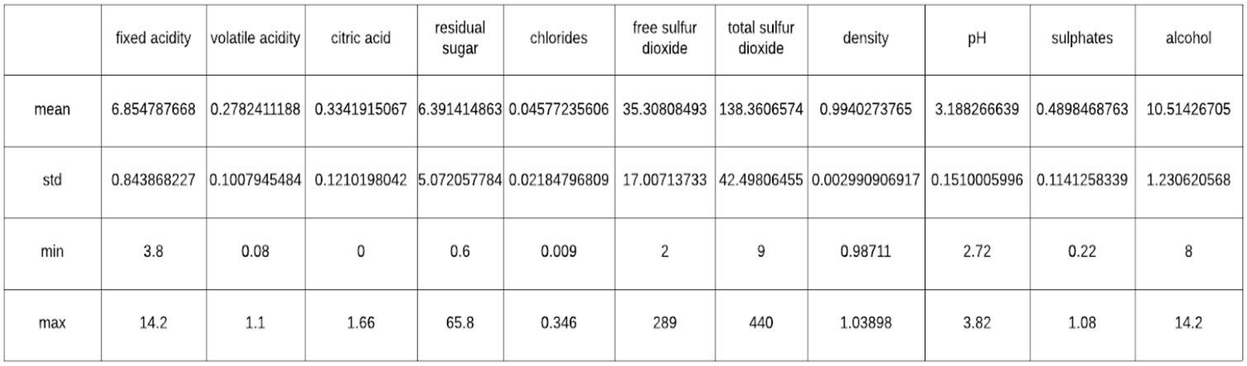

pH


sulphates



In the above graphs, it is observed that more points of each observation are densely connected.

The table shows the mean, standard deviation, min and max of each independent variable.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
mean	6.854787668	0.2782411188	0.3341915067	6.391414863	0.04577235606	35.30808493	138.3606574	0.9940273765	3.188266639	0.4898468763	10.51426705
std	0.843868227	0.1007945484	0.1210198042	5.072057784	0.02184796809	17.00713733	42.49806455	0.002990906917	0.1510005996	0.1141258339	1.230620568
min	3.8	0.08	0	0.6	0.009	2	9	0.98711	2.72	0.22	8
max	14.2	1.1	1.66	65.8	0.346	289	440	1.03898	3.82	1.08	14.2

Multiple Linear Regression

The MLR Equation is $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$

X is Data Matrix of Independent Variables, $\boldsymbol{\beta}$ is Regression Coefficient, $\boldsymbol{\epsilon}$ is the Residual Error and Y is Dependent Variable.

β can be found using the formula : $(X^T X)^{-1} X^T Y$

ϵ can be found using the formula : $Y_{Estimated} - Y_{Actual}$

In the below shown graphs, we can observe the distribution of errors is Normal.

I have shown the scatter plot of residual beside the distribution graph.

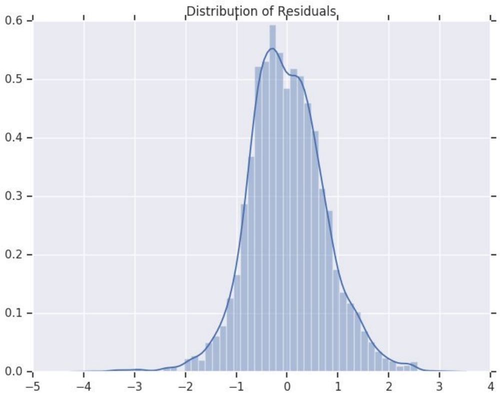
$Y_{Estimated} Vs\ Y_{Actual}$ is shown here in the below graph.

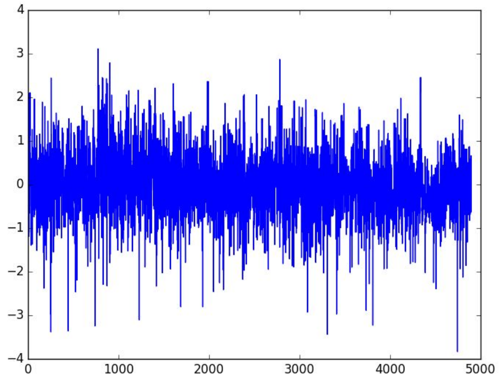
Confidence Intervals for each regression coefficient is shown below.

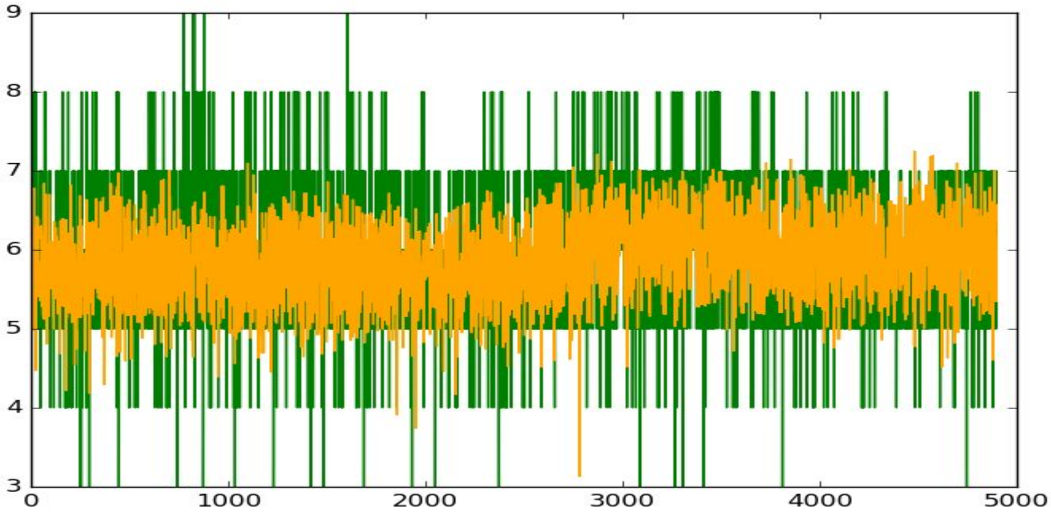
The final equation of Multiple Linear Regression is :

$$Y = 150.19X_0 + 0.065X_1 - 1.863X_2 + 0.0221X_3 + 0.081X_4 - 0.2473X_5 + 0.0037X_6$$

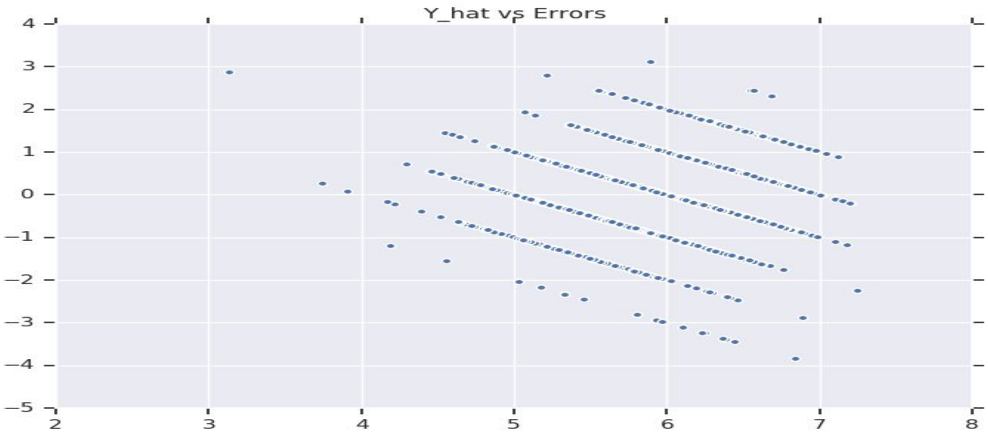
$$- 0.0003X_7 - 150X_8 + 0.686X_9 + 0.6315X_{10} + 0.1935X_{11}$$

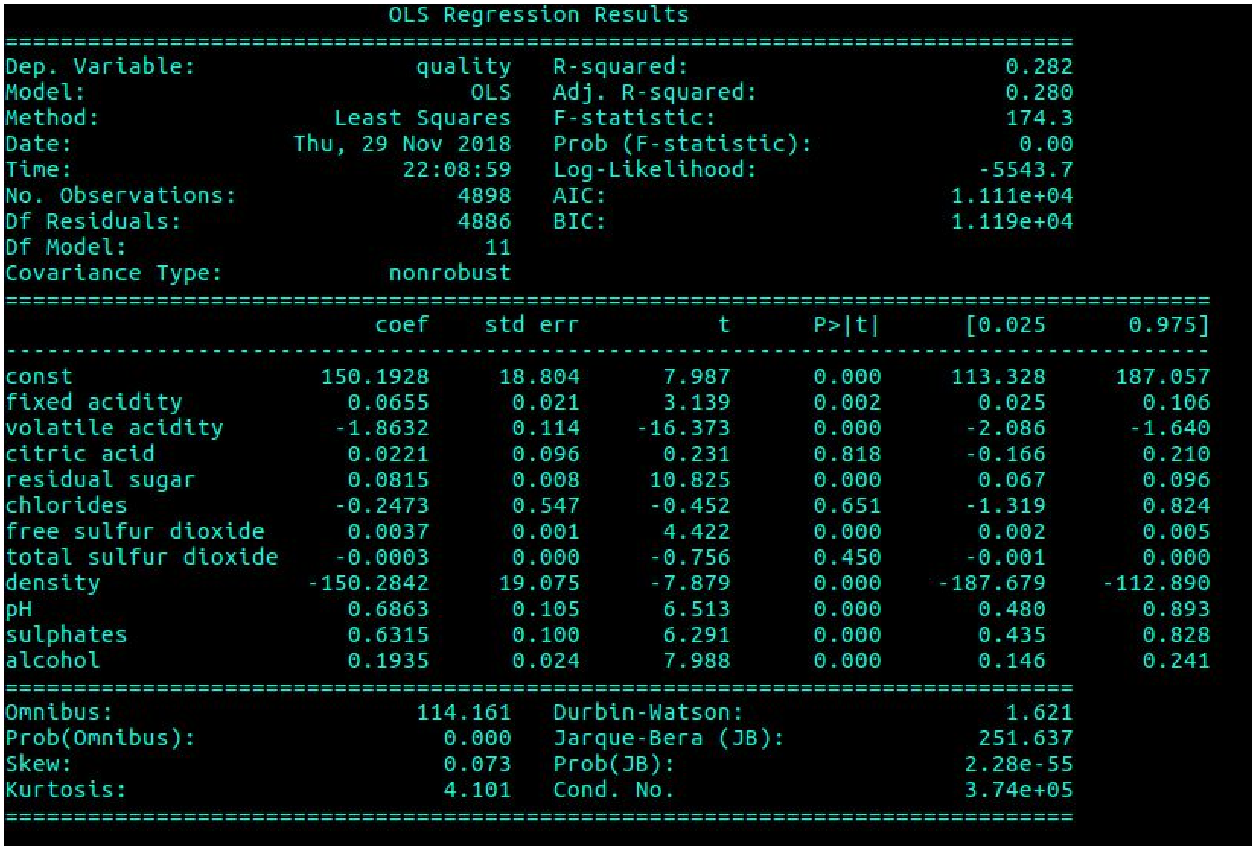






$Y_{Estimated}\, Vs\, Y_{Actual}$





```
                            OLS Regression Results
==============================================================================
Dep. Variable:                quality   R-squared:                       0.282
Model:                            OLS   Adj. R-squared:                  0.280
Method:                 Least Squares   F-statistic:                     174.3
Date:                Thu, 29 Nov 2018   Prob (F-statistic):               0.00
Time:                        22:08:59   Log-Likelihood:                -5543.7
No. Observations:                4898   AIC:                         1.111e+04
Df Residuals:                    4886   BIC:                         1.119e+04
Df Model:                          11
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                  150.1928     18.804      7.987      0.000     113.328     187.057
fixed acidity            0.0655      0.021      3.139      0.002       0.025       0.106
volatile acidity        -1.8632      0.114    -16.373      0.000      -2.086      -1.640
citric acid              0.0221      0.096      0.231      0.818      -0.166       0.210
residual sugar           0.0815      0.008     10.825      0.000       0.067       0.096
chlorides               -0.2473      0.547     -0.452      0.651      -1.319       0.824
free sulfur dioxide      0.0037      0.001      4.422      0.000       0.002       0.005
total sulfur dioxide    -0.0003      0.000     -0.756      0.450      -0.001       0.000
density               -150.2842     19.075     -7.879      0.000    -187.679    -112.890
pH                       0.6863      0.105      6.513      0.000       0.480       0.893
sulphates                0.6315      0.100      6.291      0.000       0.435       0.828
alcohol                  0.1935      0.024      7.988      0.000       0.146       0.241
==============================================================================
Omnibus:                      114.161   Durbin-Watson:                   1.621
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              251.637
Skew:                           0.073   Prob(JB):                     2.28e-55
Kurtosis:                       4.101   Cond. No.                     3.74e+05
==============================================================================
```

Model Adequacy Tests

The above Model shown in the figure is not good fit model as one can observe R^2 Value is very less.

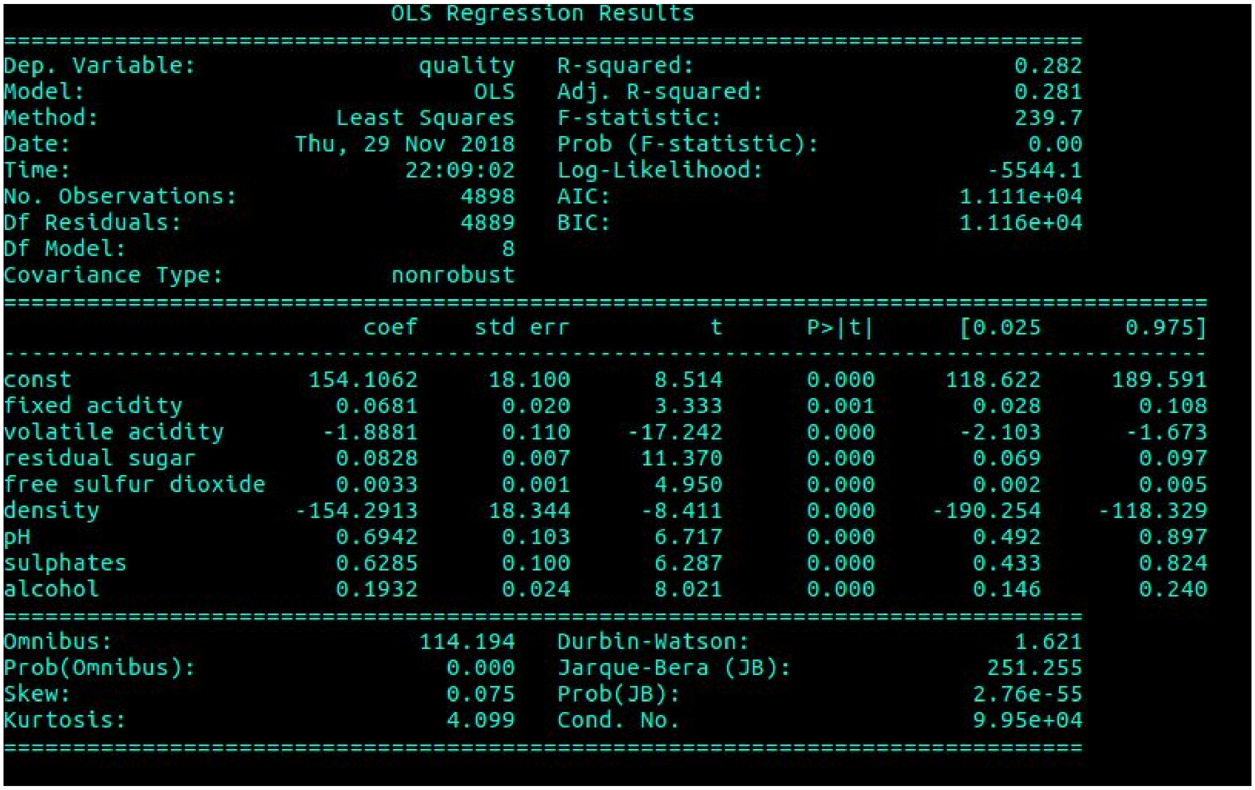
Test of Individual Parameter :

H_0 : Any of the regression coefficients, = 0

H_1 : ≠ 0 for all values of **β**.

The p-value for variables such as citric acid, chlorides, total sulfur dioxide are higher than 0.05, which states that we fail to reject the Hypothesis Test.

So Performed MLR again after the removal of above mentioned variables.

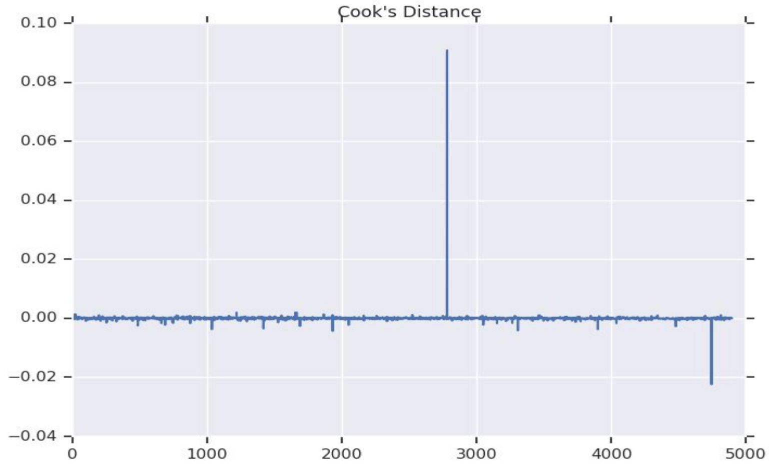


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                quality   R-squared:                       0.282
Model:                            OLS   Adj. R-squared:                  0.281
Method:                 Least Squares   F-statistic:                     239.7
Date:                Thu, 29 Nov 2018   Prob (F-statistic):               0.00
Time:                        22:09:02   Log-Likelihood:                -5544.1
No. Observations:                4898   AIC:                         1.111e+04
Df Residuals:                    4889   BIC:                         1.116e+04
Df Model:                           8
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                 154.1062     18.100      8.514      0.000     118.622     189.591
fixed acidity           0.0681      0.020      3.333      0.001       0.028       0.108
volatile acidity       -1.8881      0.110    -17.242      0.000      -2.103      -1.673
residual sugar          0.0828      0.007     11.370      0.000       0.069       0.097
free sulfur dioxide     0.0033      0.001      4.950      0.000       0.002       0.005
density              -154.2913     18.344     -8.411      0.000    -190.254    -118.329
pH                      0.6942      0.103      6.717      0.000       0.492       0.897
sulphates               0.6285      0.100      6.287      0.000       0.433       0.824
alcohol                 0.1932      0.024      8.021      0.000       0.146       0.240
==============================================================================
Omnibus:                      114.194   Durbin-Watson:                   1.621
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              251.255
Skew:                           0.075   Prob(JB):                     2.76e-55
Kurtosis:                       4.099   Cond. No.                     9.95e+04
==============================================================================
```

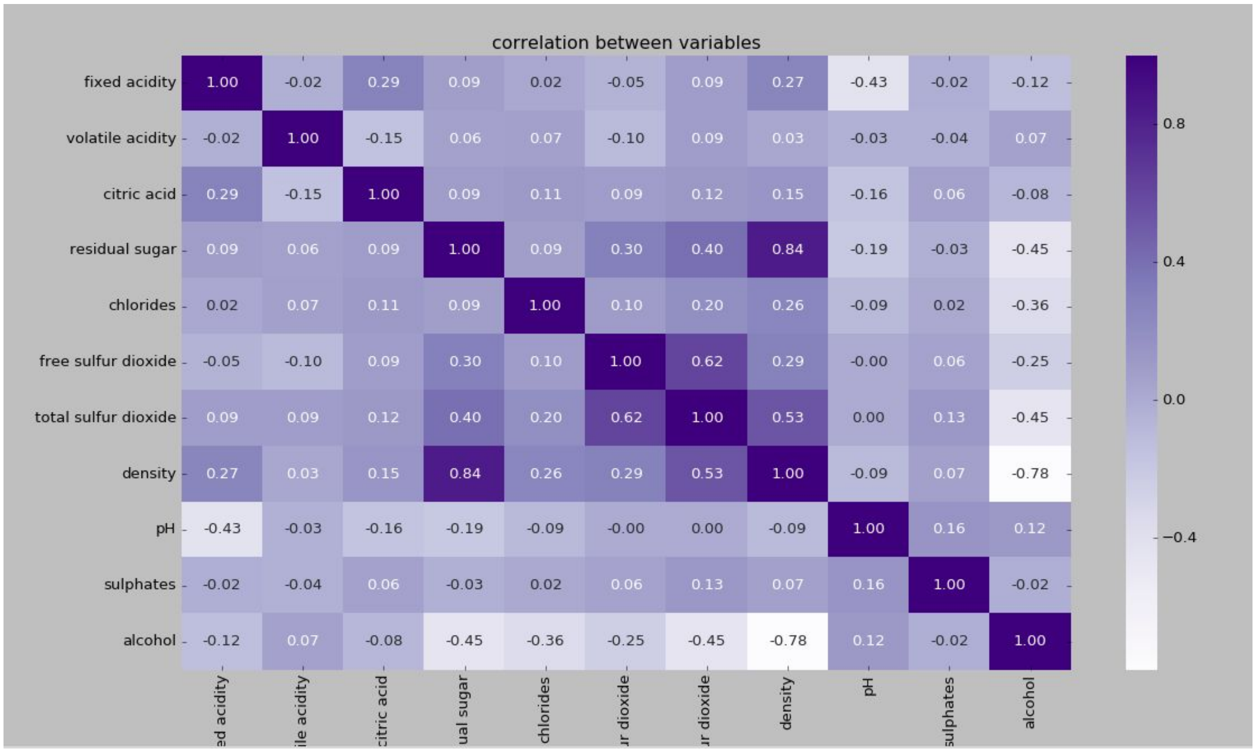
Even though We removed less p-value variables, R^2 has not improved. It is still same with value 0.282

Model Diagnostics

Cook’s Distance :

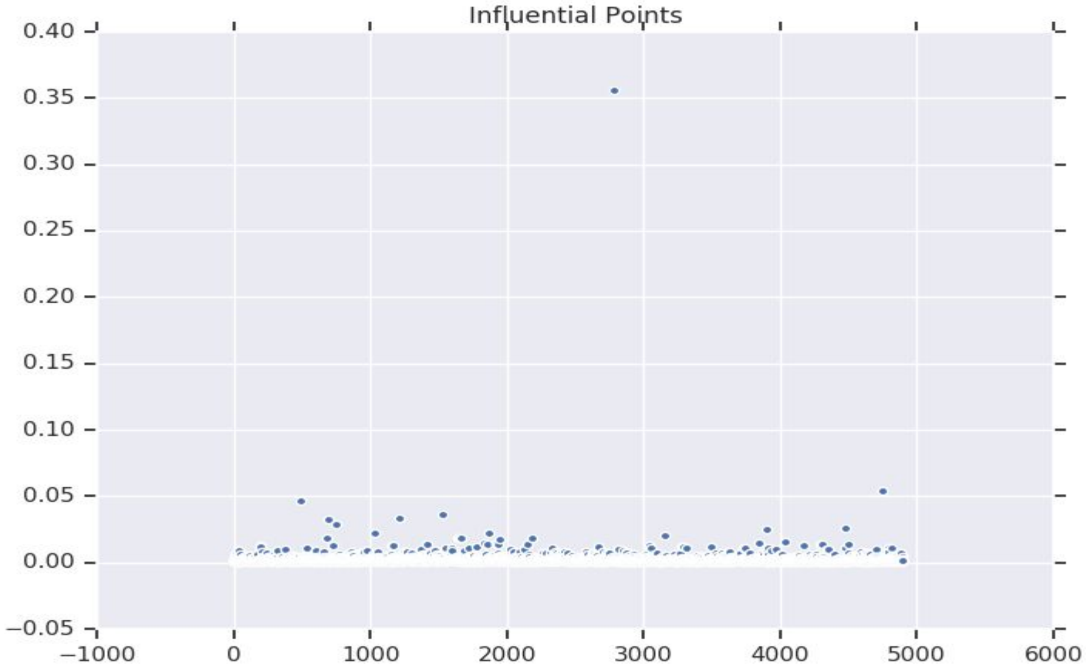


Multi-Collinearity :



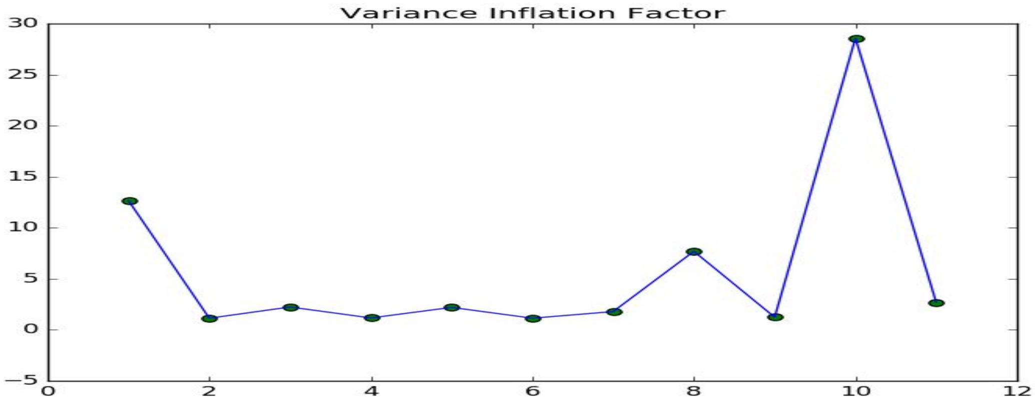
	ed acidity	ile acidity	citric acid	ual sugar	chlorides	r dioxide	r dioxide	density	pH	sulphates	alcohol
fixed acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12
volatile acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07
citric acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08
residual sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36
free sulfur dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	-0.00	0.06	-0.25
total sulfur dioxide	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78
pH	-0.43	-0.03	-0.16	-0.19	-0.09	-0.00	0.00	-0.09	1.00	0.16	0.12
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00

Influential Points :



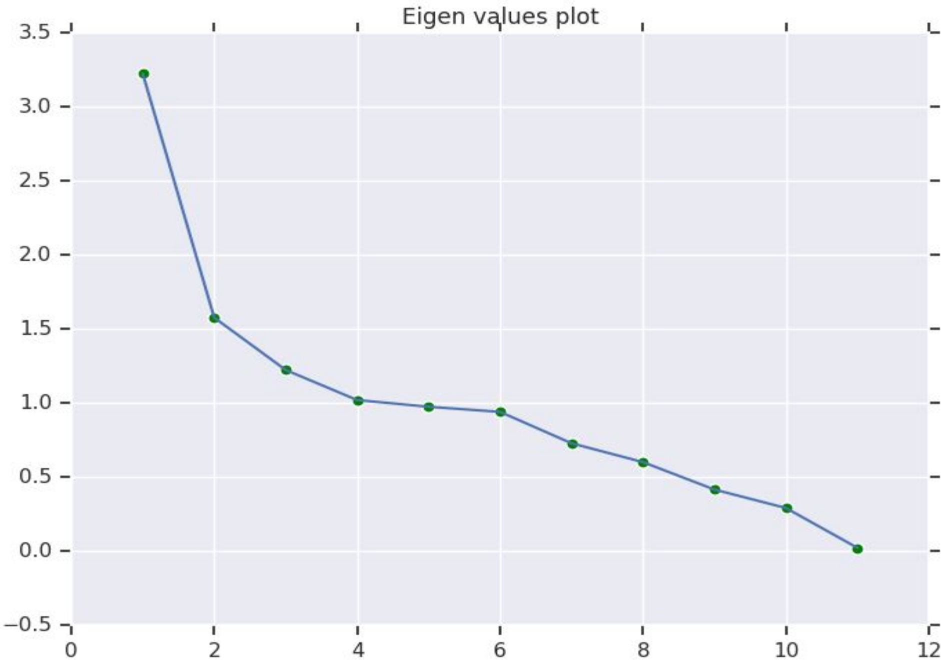
There are some influential points which need to be removed but due to less number of observations, I have not removed.

Variance Inflation Factor : Here, I considered only the variables whose VIF < 10



PCA

To know which variables are affecting the dependent variable more

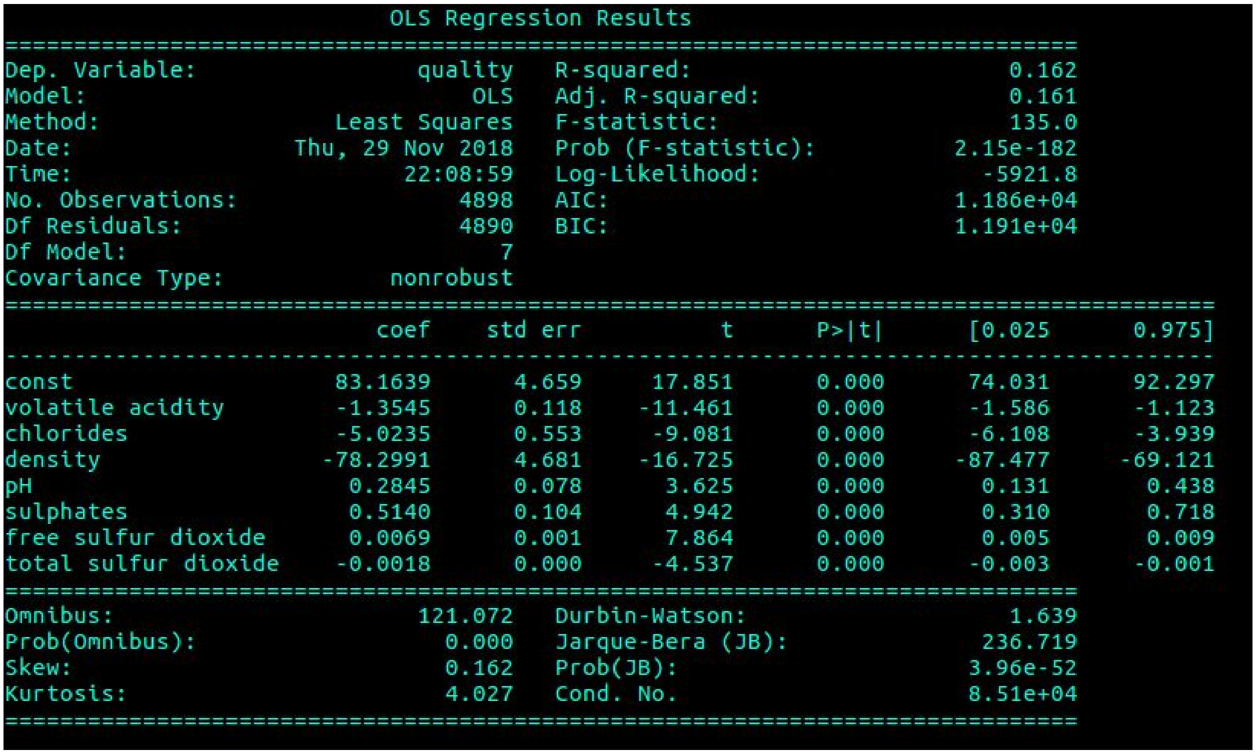


In the above figure, it is shown that the Eigen-values of independent variables after ordering them in descending order.

I have found that Eigen-value of

volatile acidity > chlorides > density >pH > sulphates > free sulfur dioxide > total sulfur dioxide.

After reducing the dimensionality to 7 i,e Top 7 Eigen-value Independent Variables, MLR results are



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                quality   R-squared:                       0.162
Model:                            OLS   Adj. R-squared:                  0.161
Method:                 Least Squares   F-statistic:                     135.0
Date:                Thu, 29 Nov 2018   Prob (F-statistic):          2.15e-182
Time:                        22:08:59   Log-Likelihood:                -5921.8
No. Observations:                4898   AIC:                         1.186e+04
Df Residuals:                    4890   BIC:                         1.191e+04
Df Model:                           7
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   83.1639      4.659     17.851      0.000      74.031      92.297
volatile acidity        -1.3545      0.118    -11.461      0.000      -1.586      -1.123
chlorides               -5.0235      0.553     -9.081      0.000      -6.108      -3.939
density                -78.2991      4.681    -16.725      0.000     -87.477     -69.121
pH                       0.2845      0.078      3.625      0.000       0.131       0.438
sulphates                0.5140      0.104      4.942      0.000       0.310       0.718
free sulfur dioxide      0.0069      0.001      7.864      0.000       0.005       0.009
total sulfur dioxide    -0.0018      0.000     -4.537      0.000      -0.003      -0.001
==============================================================================
Omnibus:                      121.072   Durbin-Watson:                   1.639
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              236.719
Skew:                           0.162   Prob(JB):                     3.96e-52
Kurtosis:                       4.027   Cond. No.                     8.51e+04
==============================================================================
```

Multi Collinearlity Number : 156.04 which is not a great serious problem.

Supervised Learning :

Here, I used Supervised Learning with Linear Regression to each the Mean Squared Errors. Split ratio - 80, 20

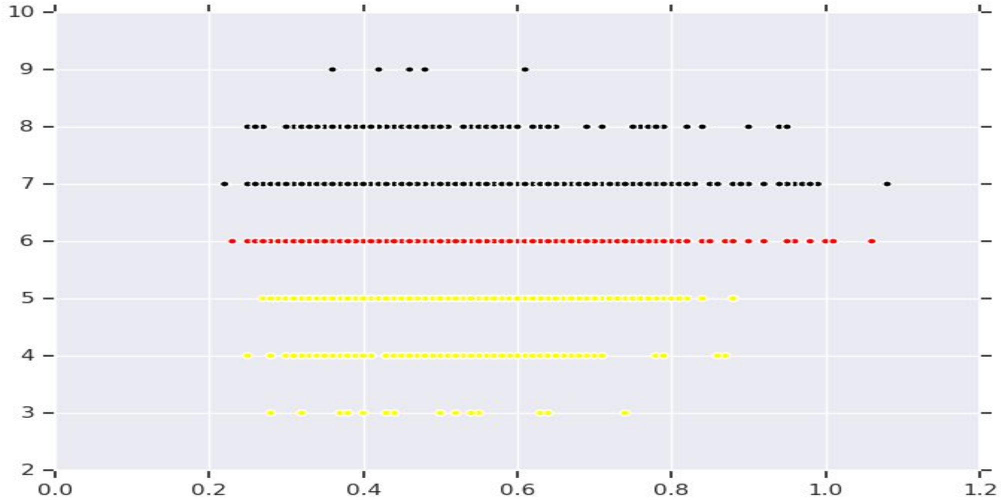
MSE for In Sample(Training) : 0.56

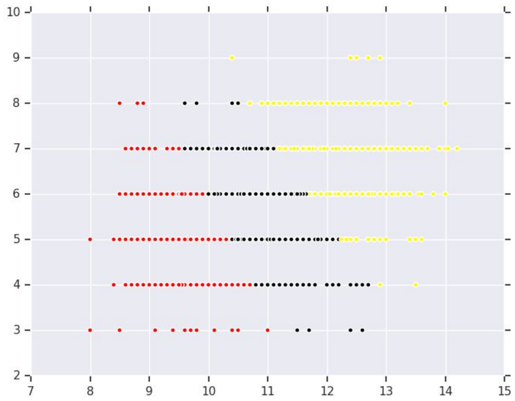
MSE for Out Sample(Testing) : 0.57

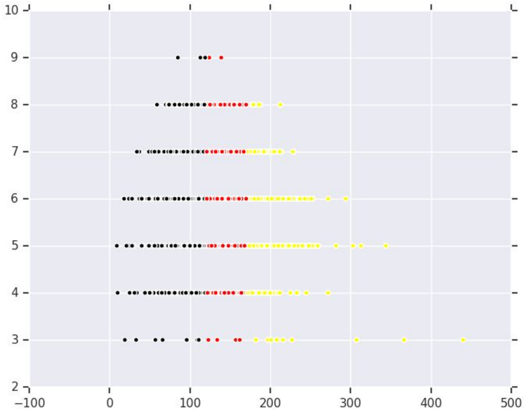
No much difference between them.

Classification - Clustering

Here, I considered each independent variable along with target Y to from clusters and observed the clustering process. Some of them shown here. In every case Y is on Y-axis.

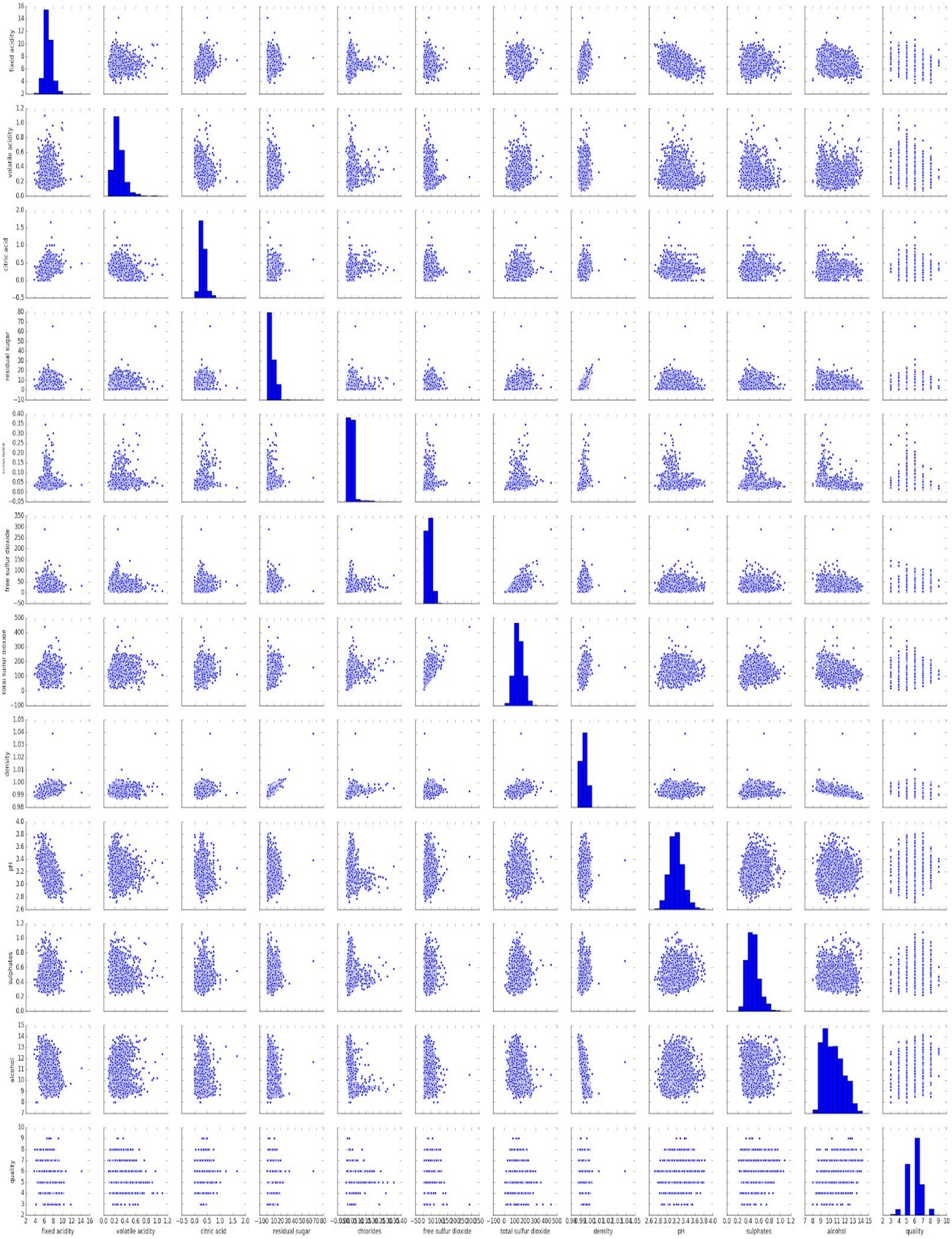




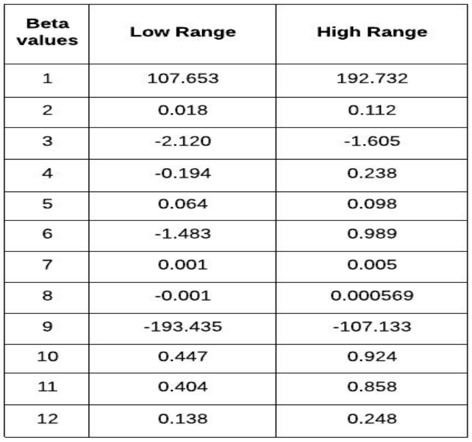


Plots from other aspects

Pair wise plot of each variable along with another variable is shown below:


fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality

Confidence Regions of Regression Coefficients :



Beta values	Low Range	High Range
1	107.653	192.732
2	0.018	0.112
3	-2.120	-1.605
4	-0.194	0.238
5	0.064	0.098
6	-1.483	0.989
7	0.001	0.005
8	-0.001	0.000569
9	-193.435	-107.133
10	0.447	0.924
11	0.404	0.858
12	0.138	0.248

My personal views on this data is, There are some other variables like Grape Riping, Temperature, Brand etc are impacting this quality variable.

Only 28.2% variance of Quality is explained by this independent variables.