Course Name: Machine Learning

Course Code: DS8002

Project 3

Md. Khaled Hyder

ID: 500800613

Objectives:

Exploratory Analysis:

- Visualize variance of each column
- Calculate and visualize the correlation matrix and comment on dependencies between the features.
- Looking at the correlation matrix suggest a set of features, which may be removed from the experiment.

Analysis:

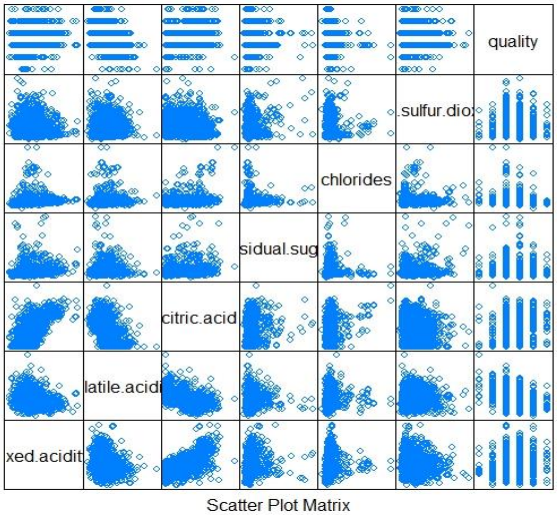
- Implement multivariate linear regression
- Perform model selection by testing higher order polynomials with the regression (multivariate polynomial regression)
- Plot a graph similar to the one in Figure 4.6 and show the polynomial order, which is giving the lowest error.

Dataset Description:

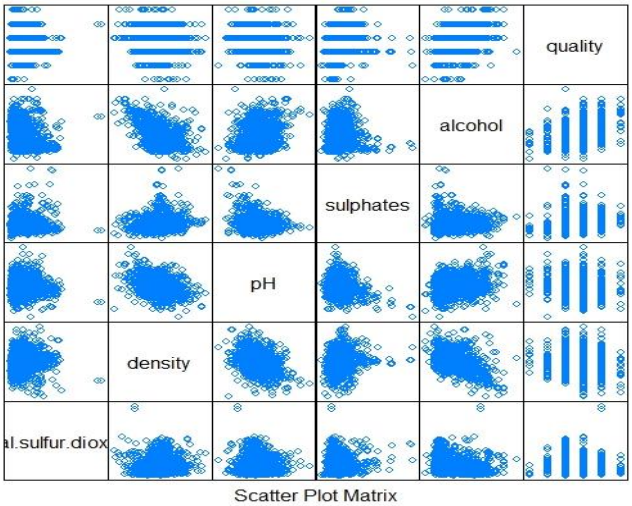
We will work with UCI wine quality dataset, which contains two CSV files. One for red wine and the other for white wine. Both files have 12 variables where red wine dataset has 1599 observation, and white wine data set has 4898 observations. We will consider wine quality as our output variable.

Initial data analysis:

Red wine dataset:

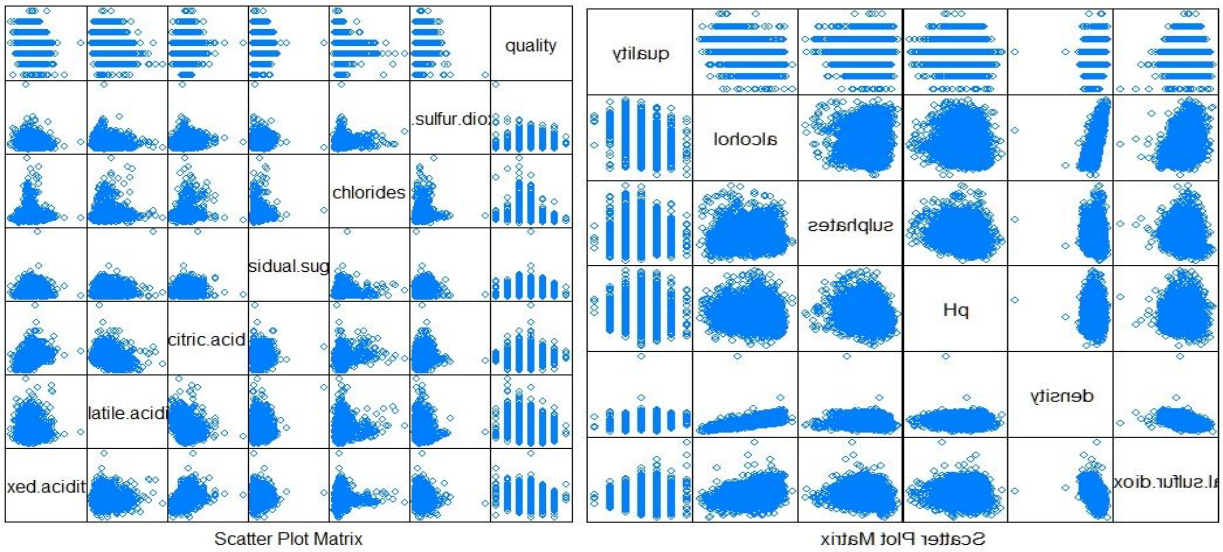


Scatter Plot Matrix



Scatter Plot Matrix

White wine data set:

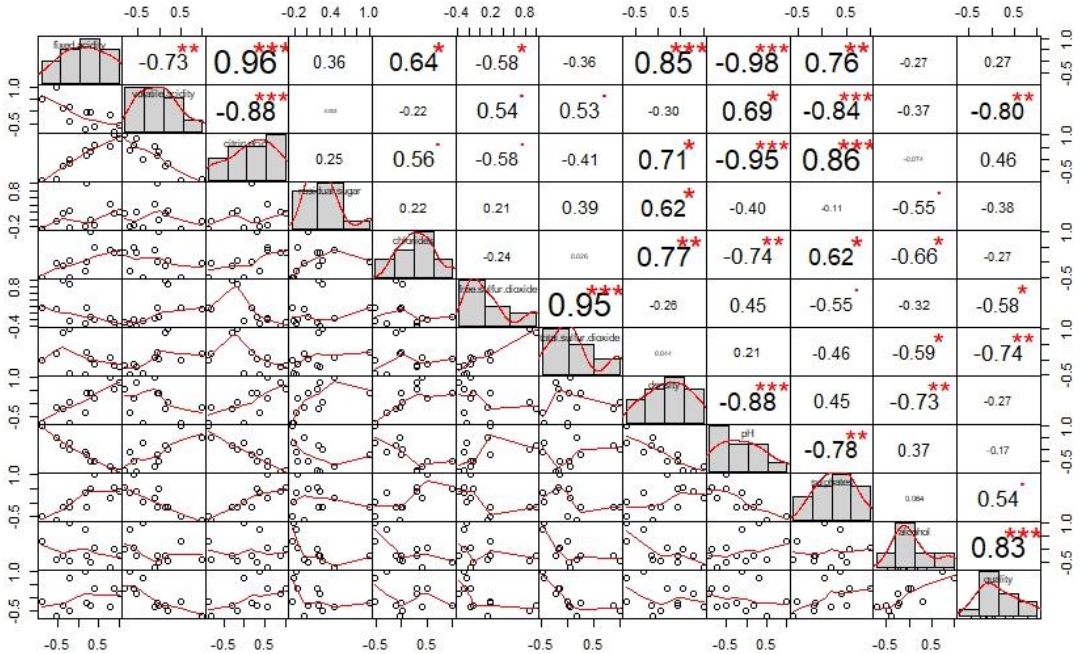


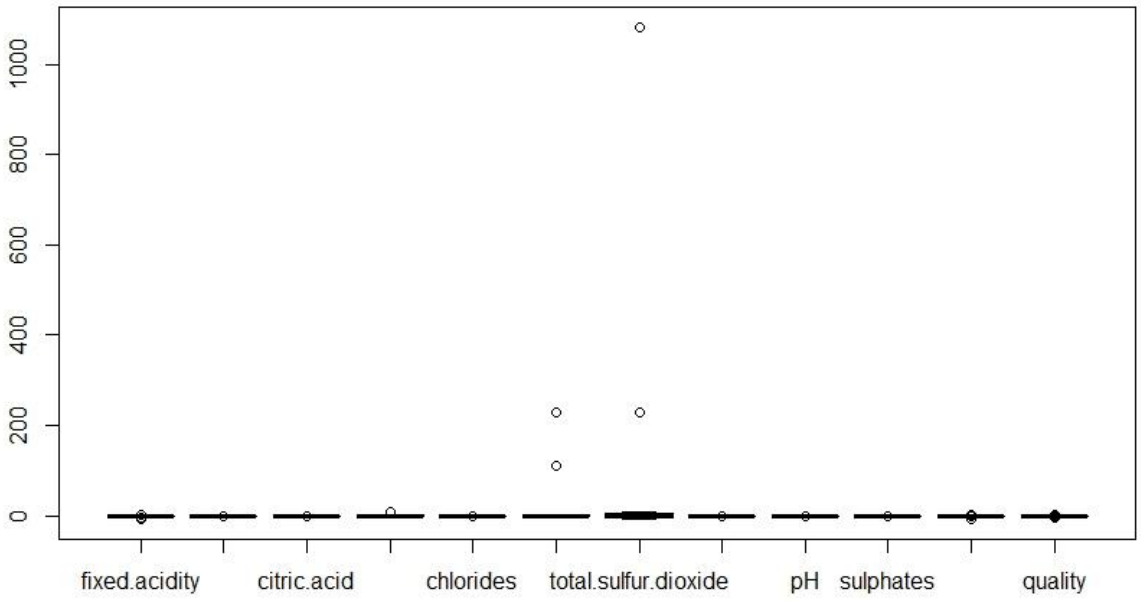
At initial observation, we can see the output variable displaying categorical data for both cases. Thus we can see low accuracy rate in the model.

Exploratory Analysis:

Visualize variance of red and white wine

For Red wine data:





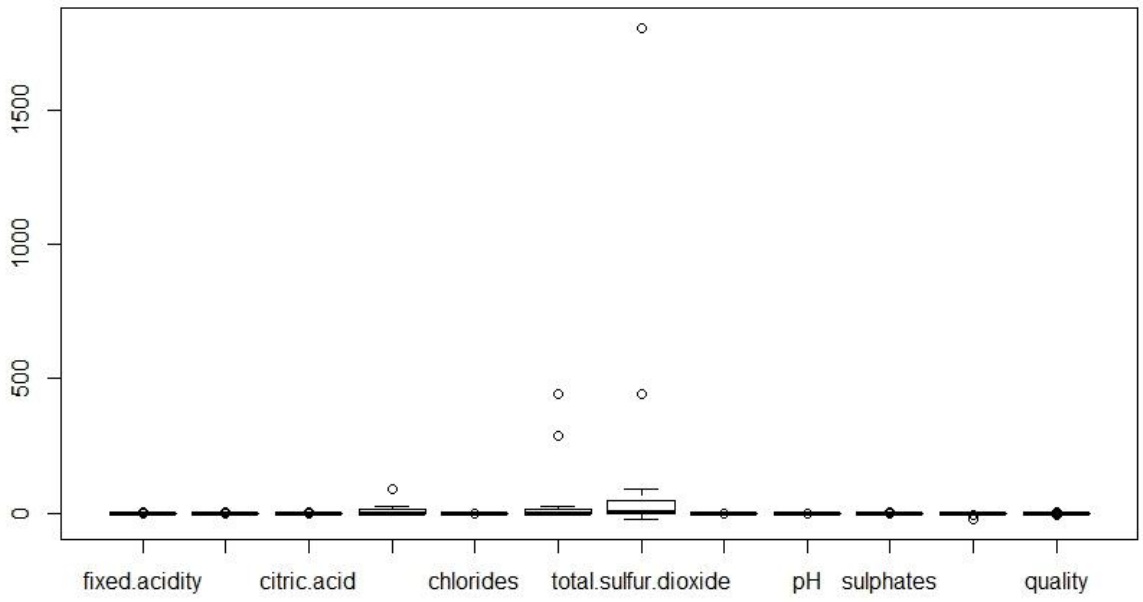
Console Terminal

```
> splom(~white_wine_df[c(7:12)], groups=NULL,data=white_wine_df,axis.line.tck=0,axis.text.alpha=0)
> ###### View Variance#####
> var_red <- var(red_wine_df)
> plot(var_red)
> barplot(var_red)
> boxplot(var_red)
> var_red
                      fixed.acidity volatile.acidity   citric.acid residual.sugar     chlorides free.sulfur.dioxide
fixed.acidity          3.031416389    -7.985142e-02  0.2278200037   0.2817562623  7.678692e-03       -2.800921e+00
volatile.acidity      -0.079851417     3.206238e-02 -0.0192716208   0.0004841910  5.165869e-04       -1.967359e-02
citric.acid            0.227820004    -1.927162e-02  0.0379474831   0.0394342700  1.868725e-03       -1.242521e-01
residual.sugar         0.281756262     4.841910e-04  0.0394342700   1.9878971330  3.690176e-03        2.758611e+00
chlorides              0.007678692     5.165869e-04  0.0018687248   0.0036901759  2.215143e-03        2.738303e-03
free.sulfur.dioxide   -2.800921493    -1.967359e-02 -0.1242521139   2.7586114522  2.738303e-03        1.094149e+02
total.sulfur.dioxide  -6.482345858     4.504257e-01  0.2276972740   9.4164414790  7.338675e-02        2.297375e+02
density                0.002195224     7.443665e-06  0.0001341746   0.0009454109  1.782176e-05       -4.332504e-04
pH                    -0.183585704     6.494699e-03 -0.0162975823  -0.0186442890 -1.925745e-03        1.136531e-01
sulphates              0.054010092    -7.921434e-03  0.0103277145   0.0013209414  2.961878e-03        9.159247e-02
alcohol               -0.114421153    -3.860022e-02  0.0228151729   0.0632189598 -1.109152e-02       -7.736984e-01
quality                0.174423588    -5.647588e-02  0.0356118929   0.0156350457 -4.899545e-03       -4.279071e-01
                      total.sulfur.dioxide       density            pH     sulphates       alcohol       quality
fixed.acidity               -6.482346e+00  2.195224e-03 -1.835857e-01  5.401009e-02 -0.1144211534  0.1744235876
volatile.acidity             4.504257e-01  7.443665e-06  6.494699e-03 -7.921434e-03 -0.0386002214 -0.0564758833
citric.acid                  2.276973e-01  1.341746e-04 -1.629758e-02  1.032771e-02  0.0228151729  0.0356118929
residual.sugar               9.416441e+00  9.454109e-04 -1.864429e-02  1.320941e-03  0.0632189598  0.0156350457
chlorides                    7.338675e-02  1.782176e-05 -1.925745e-03  2.961878e-03 -0.0110915178 -0.0048995449
free.sulfur.dioxide          2.297375e+02 -4.332504e-04  1.136531e-01  9.159247e-02 -0.7736984003 -0.4279070696
total.sulfur.dioxide         1.082102e+03  4.424727e-03 -3.376988e-01  2.394710e-01 -7.2092978948 -4.9172370717
density                      4.424727e-03  3.562029e-06 -9.956395e-05  4.750962e-05 -0.0009979518 -0.0002666037
pH                          -3.376988e-01 -9.956395e-05  2.383518e-02 -5.146186e-03  0.0338316166 -0.0071978223
sulphates                    2.394710e-01  4.750962e-05 -5.146186e-03  2.873262e-02  0.0169067772  0.0344134084
alcohol                     -7.209298e+00 -9.979518e-04  3.383162e-02  1.690678e-02  1.1356473950  0.4097890108
quality                     -4.917237e+00 -2.666037e-04 -7.197822e-03  3.441341e-02  0.4097890108  0.6521684000
>
```

Type here to search

Observing the boxplot and value table, we have found that Free sulfur dioxide and total sulfur dioxide has the high variance(positive) with outliers. Chlorides and residual sugar also have high positive variance.

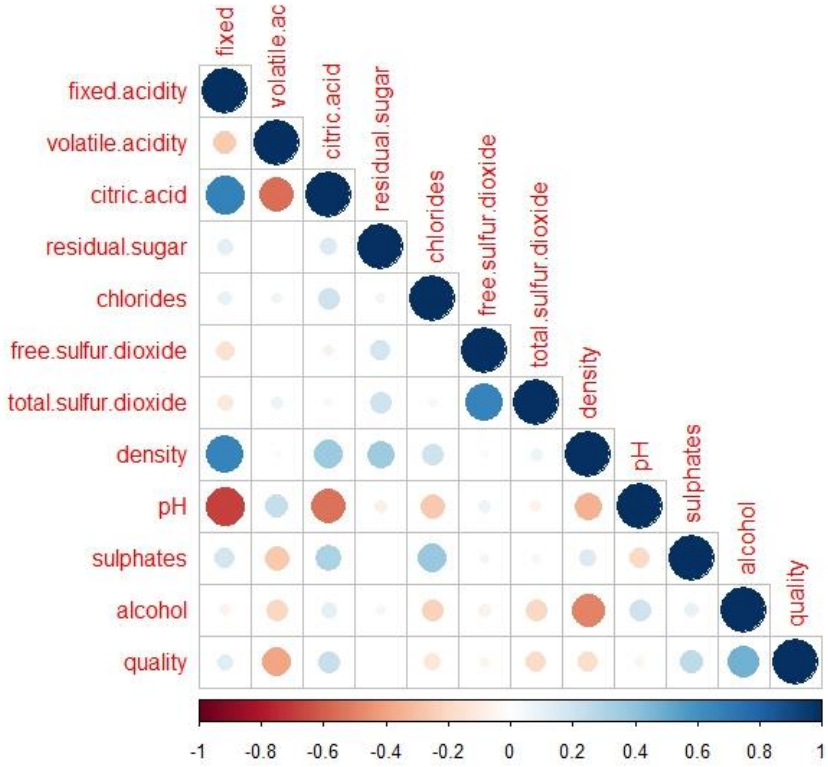
For white wine Data:




```
Console  Terminal
~/
alcohol                -7.209298e+00 -9.979518e-04  3.383162e-02  1.690678e-02  1.1356473950  0.4097890108
quality                -4.917237e+00 -2.666037e-04 -7.197822e-03  3.441341e-02  0.4097890108  0.6521684000
> #Variance study of White wine
> var_white <- var(white_wine_df)
> barplot(var_white)
> boxplot(var_white)
> var_white
                      fixed.acidity volatile.acidity   citric.acid residual.sugar     chlorides free.sulfur.dioxide
fixed.acidity         0.7121135857    -1.930571e-03  0.0295325116    0.381021814  4.256255e-04        -0.708918642
volatile.acidity     -0.0019305706     1.015954e-02 -0.0018232776    0.032865334  1.552775e-04        -0.166300459
citric.acid           0.0295325116    -1.823278e-03  0.0146457930    0.057828926  3.023838e-04         0.193629777
residual.sugar        0.3810218137     3.286533e-02  0.0578289265   25.725770164  9.827502e-03        25.800577899
chlorides             0.0004256255     1.552775e-04  0.0003023838    0.009827502  4.773337e-04         0.037674498
free.sulfur.dioxide  -0.7089186424    -1.663005e-01  0.1936297767   25.800577899  3.767450e-02       289.242719999
total.sulfur.dioxide  3.2660133926     3.823539e-01  0.6229887081   86.531302970  1.846875e-01       444.865890947
density               0.0006696773     8.173933e-06  0.0000541138    0.012727165  1.680754e-05         0.014965532
pH                   -0.0542648260    -4.857531e-04 -0.0029923451   -0.148683661 -2.983649e-04        -0.001586555
sulphates            -0.0016509923    -4.109902e-04  0.0008608829   -0.015434743  4.179687e-05         0.114937934
alcohol              -0.1255328219     8.399723e-03 -0.0112782389   -2.812740332 -9.684235e-03        -5.234508674
quality              -0.0849473094    -1.738244e-02 -0.0009870286   -0.438316094 -4.062106e-03         0.122878250
                      total.sulfur.dioxide       density            pH     sulphates       alcohol       quality
fixed.acidity                 3.26601339  6.696773e-04 -5.426483e-02 -1.650992e-03  -0.125532822 -0.0849473094
volatile.acidity              0.38235390  8.173933e-06 -4.857531e-04 -4.109902e-04   0.008399723 -0.0173824405
citric.acid                   0.62298871  5.411380e-05 -2.992345e-03  8.608829e-04  -0.011278239 -0.0009870286
residual.sugar               86.53130297  1.272717e-02 -1.486837e-01 -1.543474e-02  -2.812740332 -0.4383160939
chlorides                     0.18468749  1.680754e-05 -2.983649e-04  4.179687e-05  -0.009684235 -0.0040621056
free.sulfur.dioxide         444.86589095  1.496553e-02 -1.586555e-03  1.149379e-01  -5.234508674  0.1228782499
total.sulfur.dioxide       1806.08549085  6.735203e-02  1.489422e-02  6.526446e-01 -23.476604603 -6.5767463484
density                       0.06735203  8.945524e-06 -4.226861e-05  2.542747e-05  -0.002871430 -0.0008135274
pH                            0.01489422 -4.226861e-05  2.280118e-02  2.687523e-03   0.022565052  0.0132966000
sulphates                     0.65264458  2.542747e-05  2.687523e-03  1.302471e-02  -0.002448356  0.0054254507
alcohol                     -23.47660460 -2.871430e-03  2.256505e-02 -2.448356e-03   1.514426982  0.4747263688
quality                      -6.57674635 -8.135274e-04  1.329660e-02  5.425451e-03   0.474726369  0.7843556855
>
```

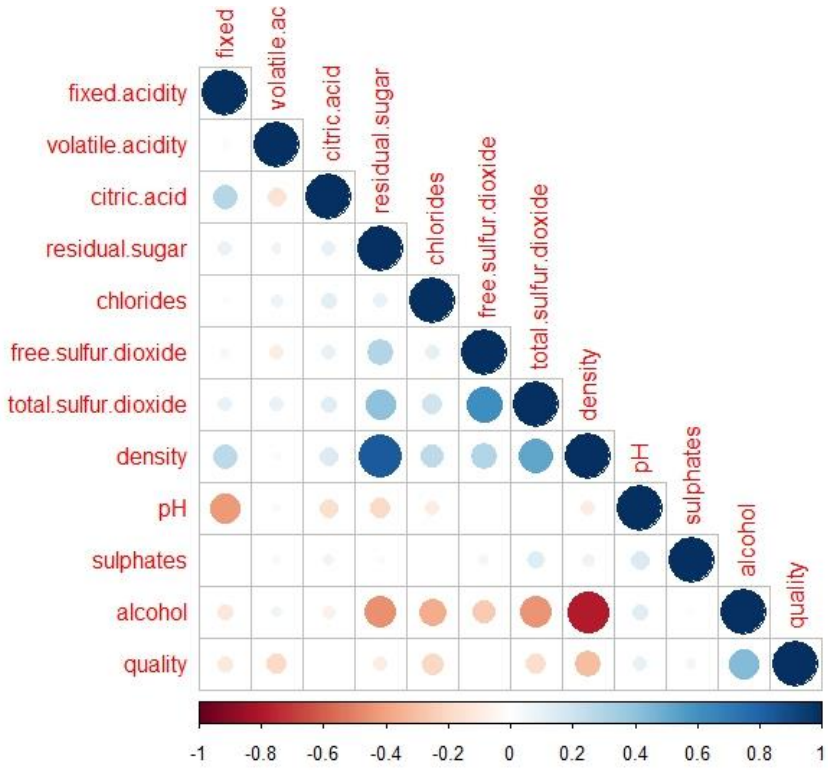
From the above table and boxplot, we can see free and total sulfur have significantly high variance. Residual sugar has also observed high variance.

Correlation among features:



Correlation among variables – red wine dataset:

We can see there is a strong positive relation between density and citric acid with fixed acidity Dark Blue and also with free and total sulfur dioxide Dark Blue. On the other hand, pH has a strong negative correlation with fixed acidity Dark Red and citric acid Dark red.



Correlation among variables – white wine dataset:

In case of white wine data, density and residual sugar have a solid positive correlation(84%), side by side free and total sulfur dioxide have also positive 62% correlation.

Feature Selection based on correlation:

For both red and white wine dataset, we could remove free and total sulfur dioxide as they have similarity in distribution and value. Side by side for red wine dataset we can also omit citric acid and density.

Data Analysis:

Implement multivariate linear regression

Split data frame to training and test dataset :

To build up multivariate linear regression, I have split the dataset into 65:35 ration, where 65% data is for training and 35% for the test.

Since the output feature contains categorical data, so the R^2 value is significantly low. To improve the model performance, I have considered p-value and excluded features with least p values.

```
> all_fet_red <- lm(red_wine_df$quality ~ red_wine_df$fixed.acidity + red_wine_df$volatile.acidit
y + red_wine_df$citric.acid + red_wine_df$residual.sugar + red_wine_df$chlorides + red_wine_df$de
nsity + red_wine_df$alcohol + red_wine_df$total.sulfur.dioxide + red_wine_df$pH +red_wine_df$sulp
hates + red_wine_df$free.sulfur.dioxide)
> summary(all_fet_red)

Call:
lm(formula = red_wine_df$quality ~ red_wine_df$fixed.acidity +
    red_wine_df$volatile.acidity + red_wine_df$citric.acid +
    red_wine_df$residual.sugar + red_wine_df$chlorides + red_wine_df$density +
    red_wine_df$alcohol + red_wine_df$total.sulfur.dioxide +
    red_wine_df$pH + red_wine_df$sulphates + red_wine_df$free.sulfur.dioxide)

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)                       2.197e+01  2.119e+01   1.036   0.3002
red_wine_df$fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
red_wine_df$volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
red_wine_df$citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
red_wine_df$residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
red_wine_df$chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
red_wine_df$density              -1.788e+01  2.163e+01  -0.827   0.4086
red_wine_df$alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
red_wine_df$total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
red_wine_df$pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
red_wine_df$sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
red_wine_df$free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
---
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: < 2.2e-16



> red_mod_trn <- lm(quality~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + ch
lorides + total.sulfur.dioxide + density + pH + sulphates + alcohol, data=tr_red)
> summary(red_mod_trn)

Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + total.sulfur.dioxide + density +
    pH + sulphates + alcohol, data = tr_red)

Residuals:
     Min       1Q   Median       3Q      Max
-2.63714 -0.37010 -0.03815  0.43921  1.90117

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.752926  26.459744  -0.028 0.977304
fixed.acidity         0.005738   0.032746   0.175 0.860931
```

```
volatile.acidity     -1.154237   0.147478  -7.826 1.24e-14 ***
citric.acid          -0.181971   0.181775  -1.001 0.317025
residual.sugar        0.004585   0.018664   0.246 0.806003
chlorides            -1.404892   0.535143  -2.625 0.008787 **
total.sulfur.dioxide -0.002354   0.000695  -3.387 0.000732 ***
density               5.004544  27.012321   0.185 0.853055
pH                   -0.455589   0.239084  -1.906 0.056986 .
sulphates             0.899218   0.143567   6.263 5.53e-10 ***
alcohol               0.302520   0.033065   9.149  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6427 on 1028 degrees of freedom
Multiple R-squared:  0.3699, Adjusted R-squared:  0.3638
F-statistic: 60.36 on 10 and 1028 DF,  p-value: < 2.2e-16
```

Considering P value, after deducting less valued p we got a similar result.

```
> summary(red_tr_revised)

Call:
lm(formula = quality ~ volatile.acidity + chlorides + total.sulfur.dioxide +
    pH + sulphates + alcohol, data = tr_red)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6416 -0.3563 -0.0372  0.4419  1.8914

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.1889225  0.5045970   8.302 3.19e-16 ***
volatile.acidity     -1.0716653  0.1227820  -8.728  < 2e-16 ***
chlorides            -1.5158208  0.5115426  -2.963 0.003114 **
total.sulfur.dioxide -0.0024147  0.0006555  -3.684 0.000242 ***
pH                   -0.4247343  0.1434173  -2.962 0.003131 **
sulphates             0.8982842  0.1383752   6.492 1.32e-10 ***
alcohol               0.2952481  0.0209147  14.117  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6418 on 1032 degrees of freedom
Multiple R-squared:  0.3692, Adjusted R-squared:  0.3655
F-statistic: 100.7 on 6 and 1032 DF,  p-value: < 2.2e-16
```

For white wine data:

Considering p-value, I have deducted citric acid and chlorides, but model accuracy was decreased thus we considered the previous result.

```
> all_fet_white <-lm(white_wine_df$quality ~ white_wine_df$fixed.acidity + white_wine_df$volatile
.acidity + white_wine_df$citric.acid + white_wine_df$residual.sugar + white_wine_df$chlorides + w
```

```
hite_wine_df$density + white_wine_df$alcohol + white_wine_df$total.sulfur.dioxide + white_wine_df
$pH + white_wine_df$sulphates + white_wine_df$free.sulfur.dioxide)
> summary(all_fet_white)

Call:
lm(formula = white_wine_df$quality ~ white_wine_df$fixed.acidity +
    white_wine_df$volatile.acidity + white_wine_df$citric.acid +
    white_wine_df$residual.sugar + white_wine_df$chlorides +
    white_wine_df$density + white_wine_df$alcohol + white_wine_df$total.sulfur.dioxide +
    white_wine_df$pH + white_wine_df$sulphates + white_wine_df$free.sulfur.dioxide)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8348 -0.4934 -0.0379  0.4637  3.1143

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         1.502e+02  1.880e+01   7.987 1.71e-15 ***
white_wine_df$fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
white_wine_df$volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
white_wine_df$citric.acid           2.209e-02  9.577e-02   0.231  0.81759
white_wine_df$residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
white_wine_df$chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
white_wine_df$density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
white_wine_df$alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
white_wine_df$total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
white_wine_df$pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
white_wine_df$sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
white_wine_df$free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 4886 degrees of freedom
Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

After deducting less valued p, we have got model with less accuracy again,

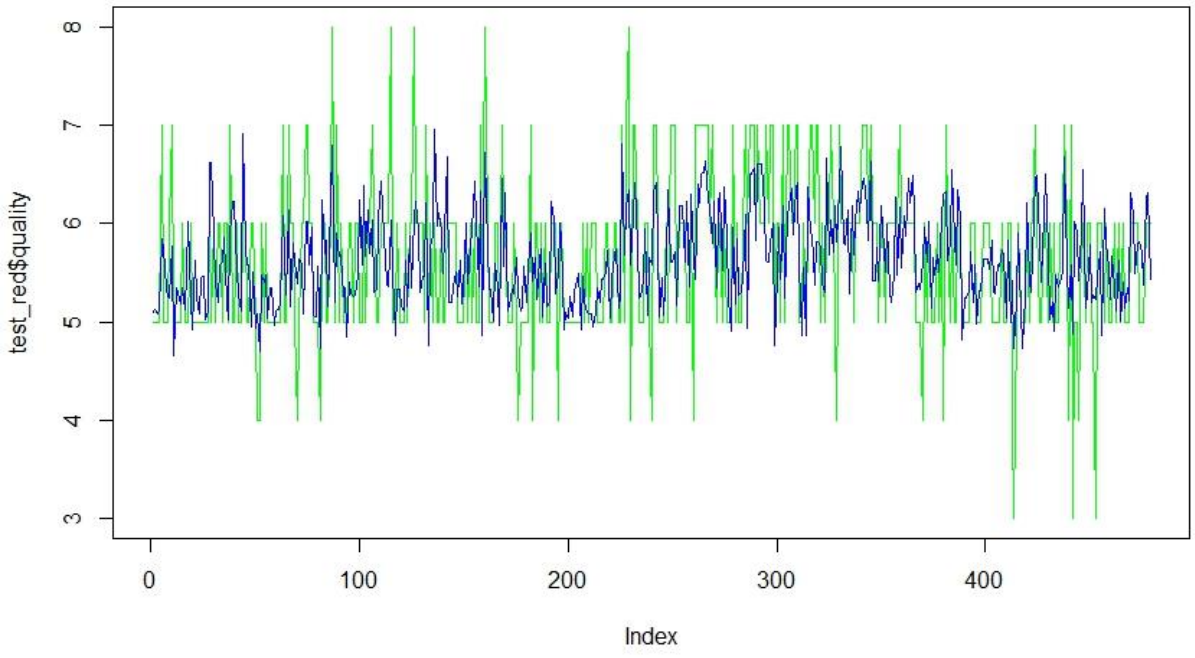
```
> summary(white_tr_revised)

Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
    total.sulfur.dioxide + pH + sulphates + alcohol, data = tr_white)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4154 -0.4901 -0.0556  0.4647  3.1908

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.8813421  0.4304298   4.371 1.28e-05 ***
fixed.acidity      -0.0579665  0.0180074  -3.219   0.0013 **
volatile.acidity   -2.1385542  0.1377722 -15.522  < 2e-16 ***
residual.sugar      0.0267030  0.0032091   8.321  < 2e-16 ***
total.sulfur.dioxide  0.0004596  0.0003720   1.235   0.2168
pH                  0.1922128  0.1043599   1.842   0.0656 .
```

```
sulphates             0.2962601  0.1219519   2.429   0.0152 *
alcohol               0.3789802  0.0131334  28.856  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7606 on 3174 degrees of freedom
Multiple R-squared:  0.2677, Adjusted R-squared:  0.2661
F-statistic: 165.8 on 7 and 3174 DF,  p-value: < 2.2e-16
```

Model performance on Test Data:

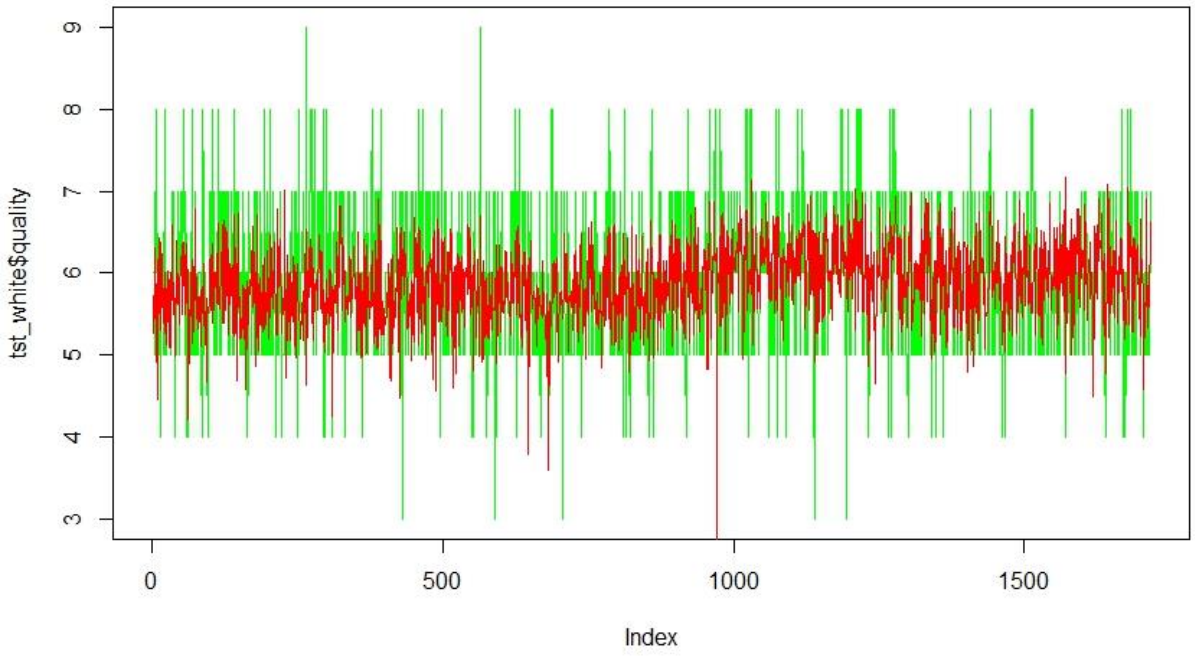
For Red wine data:

The green line indicating actual data points and the blue line shows the predicted data points. Thought the predicted points are within range but in most cases the distance between actual and predicted points are high.



For white wine data:

The green line is indicating actual and red line indicating predicted data points.



As the R^2 value is not significantly high so in both cases (for red and white wine dataset) we are observing the large distance between actual and predicted value.

Perform model selection by testing higher order polynomials with the regression

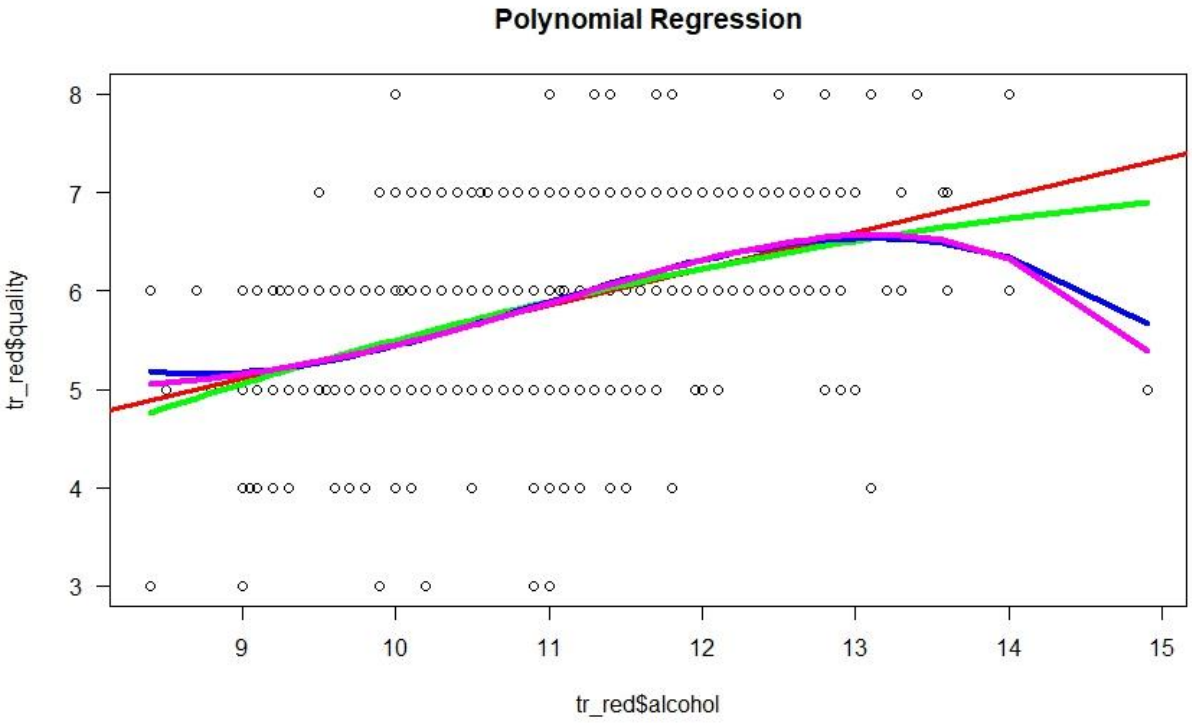
Polynomial Regression:

The particular case of linear regression where the relation between 2 variable(suppose x and y) is modeled using a polynomial rather than a line.

Model with polynomial regression:

For red wine data

Here in below pictures will display different r^2 value hence the accuracy of the different model for a different order of polynomial for feature alcohols and outcome quality. The feature^4 gives comparatively highest accuracy where I have tried feature^3 and feature^2. The polynomial regression plot displays the “magenta” line which is in polynomial order and achieved through alcohol^2+alcohol^3+alcohol^4.

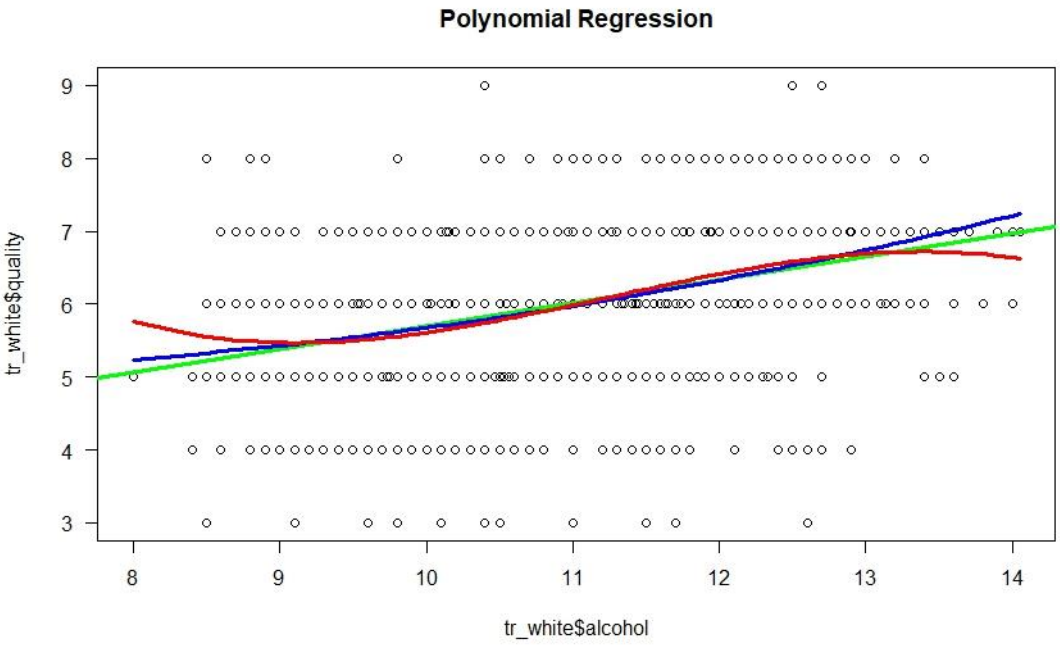



```
> plot(ts_red$quality, type = "l", lty=1.8, col="purple")
> lines(predic_red, type="l", col="blue")
> #Compare predict and actual value
> plot(ts_red$quality, type = "l", lty=1.8, col="green")
> lines(predic_red, type="l", col="blue")
> #Compare predict and actual value
> plot(ts_red$quality, type = "l", lty=1.8, col="yellow")
> lines(predic_red, type="l", col="red")
> plot(tst_white$quality, type = "l", lty=1.8, col="green")
> lines(predic_white, type="l", col="red")
> plot(tr_red$alcohol, tr_red$quality,  main="Polynomial Regression", las=1)
> model1<- lm(tr_red$quality ~tr_red$alcohol)
> summary(model1)

Call:
lm(formula = tr_red$quality ~ tr_red$alcohol)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8540 -0.4090 -0.1506  0.5168  2.5168

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.77538    0.21746   8.164 9.34e-16 ***
tr_red$alcohol  0.37078    0.02077  17.853  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7051 on 1037 degrees of freedom
Multiple R-squared:  0.2351,    Adjusted R-squared:  0.2344
F-statistic: 318.7 on 1 and 1037 DF,  p-value: < 2.2e-16

> abline(model1, lwd=3, col="red")
> model2 <-lm(tr_red$quality ~ tr_red$alcohol + I(tr_red$alcohol^2))
> summary(mod2)
```

```
I(tr_red$alcohol^3)  -0.03239    0.00988  -3.278  0.00108 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7011 on 1035 degrees of freedom
Multiple R-squared:  0.2451,    Adjusted R-squared:  0.2429
F-statistic:   112 on 3 and 1035 DF,  p-value: < 2.2e-16

> lines(smooth.spline(tr_red$alcohol, predict(model3)), col="blue", lwd=4)
> model4 <-lm(tr_red$quality ~ tr_red$alcohol + I(tr_red$alcohol^2)+I(tr_red$alcohol^3)+I(tr_red$alcohol^4))
> summary(model4)

Call:
lm(formula = tr_red$quality ~ tr_red$alcohol + I(tr_red$alcohol^2) +
    I(tr_red$alcohol^3) + I(tr_red$alcohol^4))

Residuals:
    Min      1Q  Median      3Q     Max
-2.8743 -0.3778 -0.1995  0.5216  2.5505

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -17.749535  86.840837  -0.204    0.838
tr_red$alcohol       10.664508  31.049949   0.343    0.731
I(tr_red$alcohol^2)  -1.851167   4.137614  -0.447    0.655
I(tr_red$alcohol^3)   0.139372   0.243522   0.572    0.567
I(tr_red$alcohol^4)  -0.003770   0.005341  -0.706    0.480

Residual standard error: 0.7013 on 1034 degrees of freedom
Multiple R-squared:  0.2455,    Adjusted R-squared:  0.2426
F-statistic:  84.1 on 4 and 1034 DF,  p-value: < 2.2e-16

> lines(smooth.spline(tr_red$alcohol, predict(model4)), col="magenta", lwd=4)
>
```

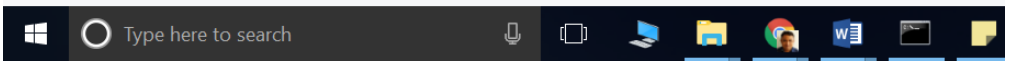
For White wine Data:

Analyzing the below scatter plot and R^2 values, I have found that the red polynomial line which represents (feature + feature^2 + feature^3) has the highest accuracy rate. The more we increase the polynomial order, the more the accuracy increasing. Which also indicating model is performing better with a higher number of the polynomial.

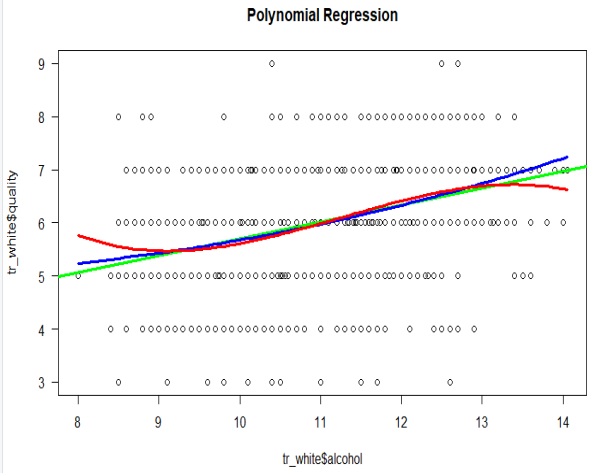


```
F-statistic:  84.1 on 4 and 1034 DF,  p-value: < 2.2e-16

> lines(smooth.spline(tr_red$alcohol, predict(model4)), col="magenta", lwd=4)
> #For white wine data
> plot(tr_white$alcohol, tr_white$quality,  main="Polynomial Regression", las=1)
> model5<- lm(tr_white$quality ~tr_white$alcohol)
> summary(model5)

Call:
lm(formula = tr_white$quality ~ tr_white$alcohol)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5253 -0.5094 -0.0174  0.4906  3.1731

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.52495    0.12145   20.79   <2e-16 ***
tr_white$alcohol  0.31749    0.01148   27.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7972 on 3180 degrees of freedom
Multiple R-squared:  0.194,     Adjusted R-squared:  0.1938
F-statistic: 765.5 on 1 and 3180 DF,  p-value: < 2.2e-16

> abline(mod5, lwd=3, col="green")
Error in abline(mod5, lwd = 3, col = "green") : object 'mod5' not found
> abline(model5, lwd=3, col="green")
> model6 <-lm(tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2))
> summary(model6)

Call:
lm(formula = tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2))
```


```
Console  Terminal
~/
> abline(model5, lwd=3, col="green")
> model6 <-lm(tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2))
> summary(model6)

Call:
lm(formula = tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2))

Residuals:
    Min      1Q  Median      3Q     Max
-3.5696 -0.5203  0.0266  0.4797  3.2133

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.748981   1.051010   5.470 4.85e-08 ***
tr_white$alcohol    -0.287123   0.196118  -1.464  0.14328
I(tr_white$alcohol^2)  0.027957   0.009053   3.088  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7962 on 3179 degrees of freedom
Multiple R-squared:  0.1964,    Adjusted R-squared:  0.1959
F-statistic: 388.5 on 2 and 3179 DF,  p-value: < 2.2e-16

> lines(smooth.spline(tr_white$alcohol, predict(model6)), col="blue", lwd=3)
> model7 <-lm(tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2)+I(tr_white$alcohol^3))
> summary(model7)

Call:
lm(formula = tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2) +
    I(tr_white$alcohol^3))

Residuals:
    Min      1Q  Median      3Q     Max
-3.6071 -0.4996  0.0136  0.5004  3.2569
```



```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7962 on 3179 degrees of freedom
Multiple R-squared:  0.1964,    Adjusted R-squared:  0.1959
F-statistic: 388.5 on 2 and 3179 DF,  p-value: < 2.2e-16

> lines(smooth.spline(tr_white$alcohol, predict(model6)), col="blue", lwd=3)
> model7 <-lm(tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2)+I(tr_white$alcohol^3))
> summary(model7)

Call:
lm(formula = tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2) +
    I(tr_white$alcohol^3))

Residuals:
    Min      1Q  Median      3Q     Max
-3.6071 -0.4996  0.0136  0.5004  3.2569

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           46.161482   8.425670   5.479 4.62e-08 ***
tr_white$alcohol     -11.559416   2.340127  -4.940 8.23e-07 ***
I(tr_white$alcohol^2)   1.066450   0.215027   4.960 7.43e-07 ***
I(tr_white$alcohol^3)  -0.031602   0.006538  -4.834 1.40e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7934 on 3178 degrees of freedom
Multiple R-squared:  0.2023,    Adjusted R-squared:  0.2015
F-statistic: 268.6 on 3 and 3178 DF,  p-value: < 2.2e-16

> lines(smooth.spline(tr_white$alcohol, predict(model7)), col="red", lwd=3)
> lines(smooth.spline(tr_white$alcohol, predict(model7)), col="red", lwd=3)
```



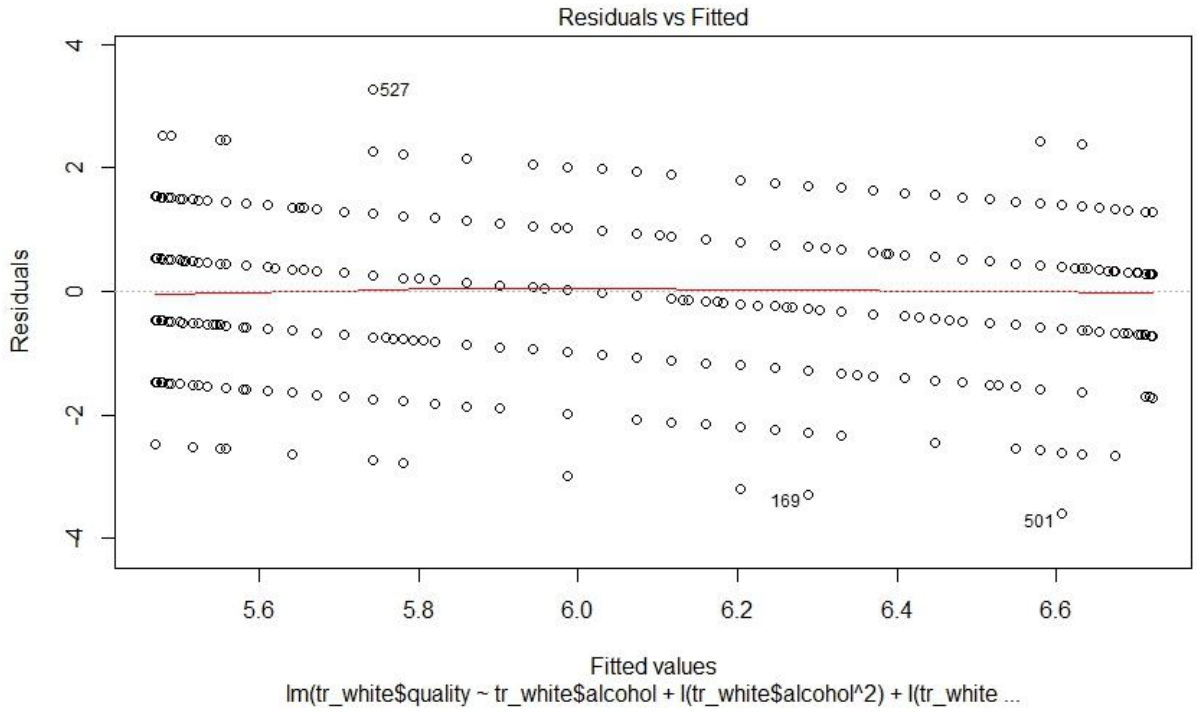
Analysis of Variance Test:

```
> anova(model5, model7)
Analysis of Variance Table

Model 1: tr_white$quality ~ tr_white$alcohol
Model 2: tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2) +
    I(tr_white$alcohol^3)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3180 2021.2
2   3178 2000.4  2    20.753 16.485 7.543e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

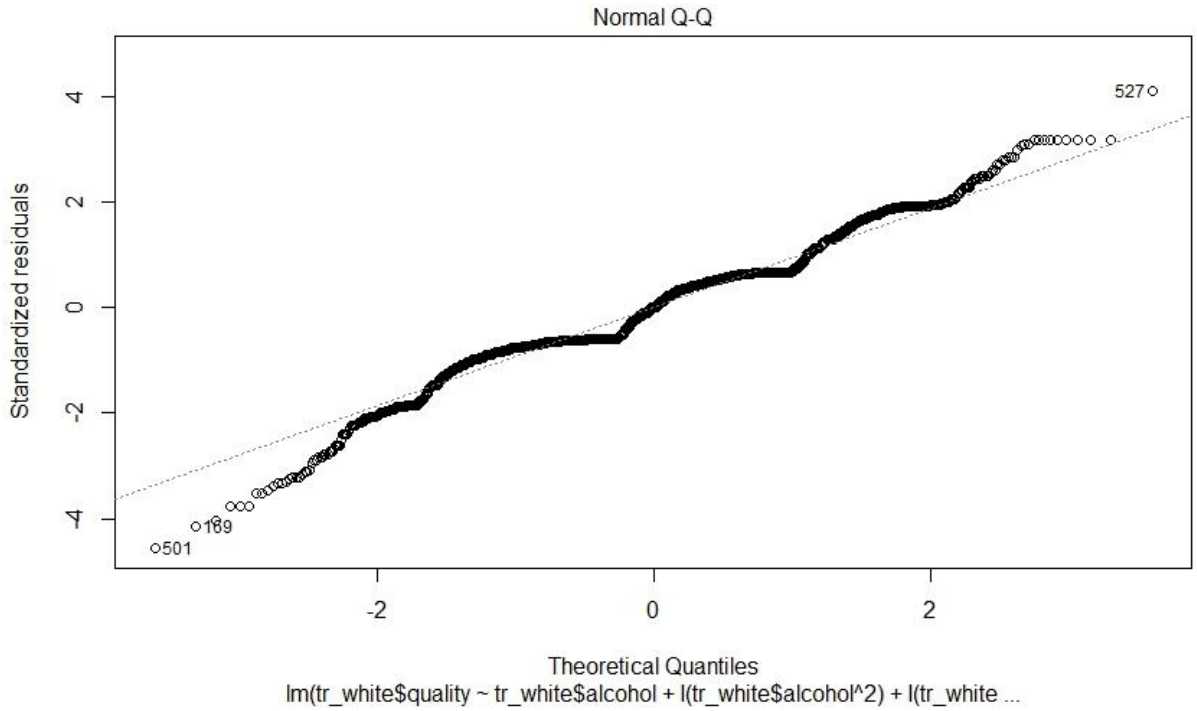
ANOVA is also indicating model is developing regarding p-value.

Plotting error indicating the lowest error:

Here we have drawn a plot based on model 7, which has three polynomial variables and the accuracy was increased due to apply for a polynomial order.



Here the x-axis is the predicted or fitted Y value and the y-axis the residuals/errors €, as the red line is indicating that the linearity assumption has been met here although the distribution is high.



The second plot is called quantile-quantile plot (QQ Plot) where the y-axis is ordered standardized residuals, and the x-axis is ordered theoretical residuals. Here we can see the error/residual is normally distributed because the data points are to follow the diagonal line mostly.

R- Code:

```
library(lattice)

library(caTools)

red_wine_df <- read.csv("C:\\Users\\Amara\\Desktop\\data_set\\winequality-red.csv")

white_wine_df <- read.csv("C:\\Users\\Amara\\Desktop\\data_set\\winequality-white.csv")


#Split dataset for training and testing
```

```
set.seed(2)

split <- sample.split(red_wine_df$quality, SplitRatio = 0.65)

split


tr_red <- subset(red_wine_df,split=="TRUE")

ts_red <- subset(red_wine_df, split=="FALSE")


#for whitewine dataset

set.seed(2)

split2 <- sample.split(white_wine_df$quality, SplitRatio = 0.65)

split2


tr_white <- subset(white_wine_df,split=="TRUE")

tst_white <- subset(white_wine_df, split=="FALSE")




#Create scatterplot to see the trend


splom(~red_wine_df[c(1:6,12)], groups=NULL,data=red_wine_df,axis.line.tck=0,axis.text.alpha=0)

splom(~red_wine_df[c(7:12)], groups=NULL,data=red_wine_df,axis.line.tck=0,axis.text.alpha=0)



splom(~white_wine_df[c(1:6,12)], groups=NULL,data=white_wine_df,axis.line.tck=0,axis.text.alpha=0)
```

```
splom(~white_wine_df[c(7:12)], groups=NULL,data=white_wine_df,axis.line.tck=0,axis.text.alpha=0)


###### Variance Study#####

#Variance study of red wine

var_red <- var(red_wine_df)

barplot(var_red)

boxplot(var_red)

boxplot(red_wine_data)

summary(red_wine_data)

var_red

#Variance study of White wine

var_white <- var(white_wine_df)

barplot(var_white)

boxplot(var_white)

var_white

#### Calcualte and visualise COrrelation Matrix#####

library(corrplot)


cor_red <- cor(red_wine_df)

cor_white <- cor(white_wine_df)

corrplot(cor_red,type = "lower")


corrplot(cor_white,type = "lower")
```

```
#Study alcohol and wine from red_wine data

plot(red_wine_df$alcohol,red_wine_df$quality)

abline(lm(red_wine_df$quality~red_wine_df$alcohol),col="red")



#Implement multivariate linear regression

all_fet_red <- lm(red_wine_df$quality ~ red_wine_df$fixed.acidity + red_wine_df$volatile.acidity +
red_wine_df$citric.acid + red_wine_df$residual.sugar + red_wine_df$chlorides + red_wine_df$density +
red_wine_df$alcohol + red_wine_df$total.sulfur.dioxide + red_wine_df$pH +red_wine_df$sulphates +
red_wine_df$free.sulfur.dioxide)

summary(all_fet_red)



red_mod_trn <- lm(quality~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides +
total.sulfur.dioxide + density + pH + sulphates + alcohol, data=tr_red)

summary(red_mod_trn)



all_fet_white <-lm(white_wine_df$quality ~ white_wine_df$fixed.acidity +
white_wine_df$volatile.acidity + white_wine_df$citric.acid + white_wine_df$residual.sugar +
white_wine_df$chlorides + white_wine_df$density + white_wine_df$alcohol +
white_wine_df$total.sulfur.dioxide + white_wine_df$pH + white_wine_df$sulphates +
white_wine_df$free.sulfur.dioxide)

summary(all_fet_white)



white_mod_tr <- lm(quality~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides +
total.sulfur.dioxide + density + pH + sulphates + alcohol, data=tr_white)

summary(white_mod_tr)
```

```
#Remove variable with lowest p values.

red_tr_revised <- lm(quality~ volatile.acidity  + chlorides + total.sulfur.dioxide +  pH + sulphates + alcohol, data=tr_red)

summary(red_tr_revised)



white_tr_revised <- lm(quality~ fixed.acidity + volatile.acidity  + residual.sugar + total.sulfur.dioxide + pH + sulphates + alcohol, data=tr_white)

summary(white_tr_revised)







#Model prediction for both red and white wine test data

predic_red <- predict(red_mod_trn,ts_red)

predic_red



predic_white <- predict(white_mod_tr,tst_white)

predic_white



#Visualize prediction over test data

plot(ts_red$quality, type = "l", lty=1.8, col="yellow")

lines(predic_red, type="l", col="red")



plot(tst_white$quality, type = "l", lty=1.8, col="green")

lines(predic_white, type="l", col="red")
```

```
##### Visualize Multivariate Polynomial Regression #####


#FOr red wine data set:

plot(tr_red$alcohol, tr_red$quality,  main="Polynomial Regression", las=1)



model1<- lm(tr_red$quality ~tr_red$alcohol)

summary(model1)

abline(model1, lwd=3, col="red")


model2 <-lm(tr_red$quality ~ tr_red$alcohol + I(tr_red$alcohol^2))

summary(model2)

lines(smooth.spline(tr_red$alcohol, predict(model2)), col="green", lwd=4)

model3 <-lm(tr_red$quality ~ tr_red$alcohol + I(tr_red$alcohol^2)+I(tr_red$alcohol^3))

summary(model3)

lines(smooth.spline(tr_red$alcohol, predict(model3)), col="blue", lwd=4)

model4 <-lm(tr_red$quality ~ tr_red$alcohol +
I(tr_red$alcohol^2)+I(tr_red$alcohol^3)+I(tr_red$alcohol^4))

summary(model4)

lines(smooth.spline(tr_red$alcohol, predict(model4)), col="magenta", lwd=4)


#For white wine data

plot(tr_white$alcohol, tr_white$quality,  main="Polynomial Regression", las=1)
```

```
model5<- lm(tr_white$quality ~tr_white$alcohol)

summary(model5)

abline(model5, lwd=3, col="green")


model6 <-lm(tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2))

summary(model6)

lines(smooth.spline(tr_white$alcohol, predict(model6)), col="blue", lwd=3)



model7 <-lm(tr_white$quality ~ tr_white$alcohol + I(tr_white$alcohol^2)+I(tr_white$alcohol^3))

summary(model7)

lines(smooth.spline(tr_white$alcohol, predict(model7)), col="red", lwd=3)



# using the partial F-test

ANOVA(model5, model7)



####Error assumption


plot(model7)

#Hit enter four times to get error plot
```