BSAN 450 Assignment 5

1) The purpose of this example is to develop a regression model to predict the price of Bordeaux wines based on the following variables: Temp = the average temperature between April and September in the year the grapes were grown, Rain = the total rainfall in the months of August and September in the year the grapes were grown, PRain = the total rainfall in the months October to March in the year the grapes were grown, and Age = the age of the wine in years in the year 1983. The variable we wish to predict is Price = the fraction of the price of the 1961 vintage. The prices were obtained from auctions in 1990 and 1991 in the London market. The data are in a file named "wine.csv".

a) Read in the data and plot scatter plots of the variable Price versus all the independent variables.

Comment on these plots. For each plot, is there a relationship between the independent variable and Price? What is the nature of this relationship?

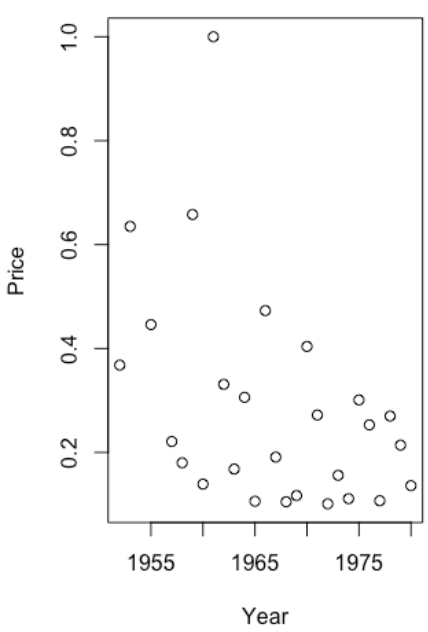
First plot is price vs year. Seems to be a non-linear relationship between these two variables. This is a monotonic plot. We could transform the Y variable "Price" since we have a non-linear relationship. You can tell the variance is not consistent and it starting to spread out.

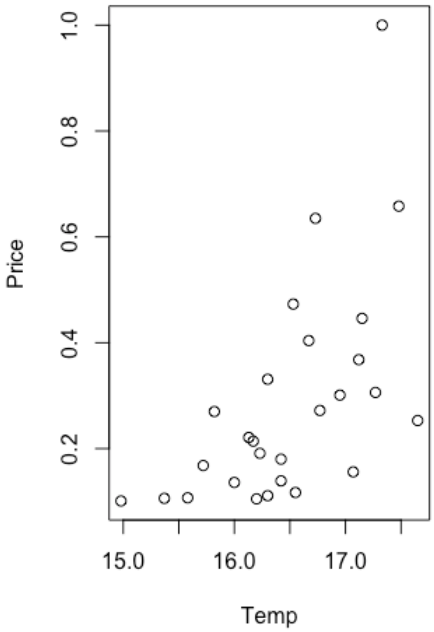
Second plot is price vs temp. There is a relationship between these two variables. This is a non-linear relationship and a monotonic plot. We could transform the Y variable "Price" to try and fix this. The variance is not consistent. As temp increases, the variance of price spreads further.

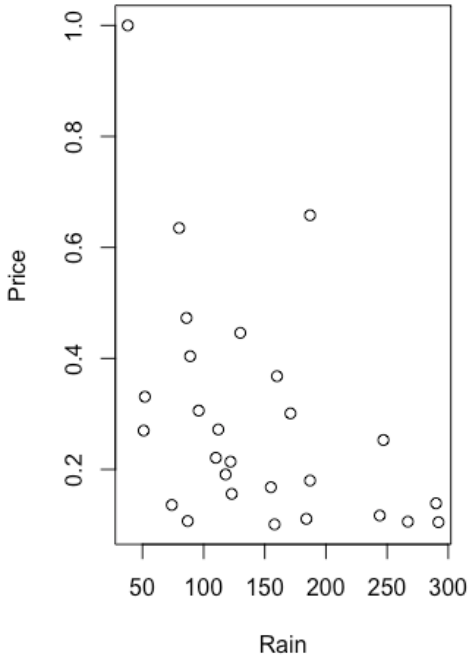
Third plot is price vs rain. There is a relationship between these two variables. Non-linear relationship.

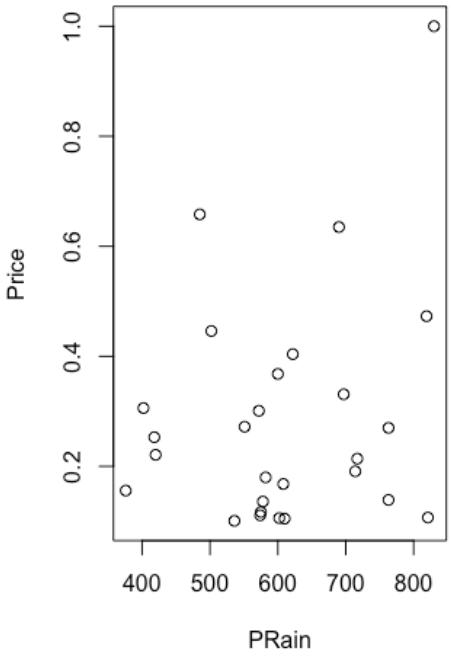
Fourth plot is Price vs PRain. There may be a relationship between these two variables. It's hard to tell from the plot, I will leave this variable in the model for now.

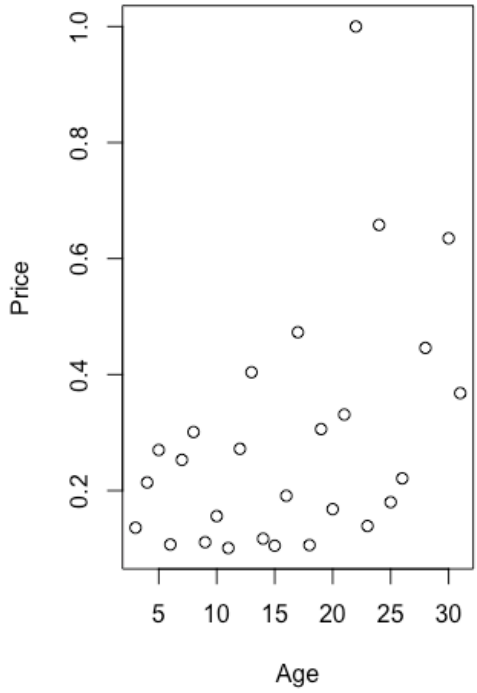
The fifth and final plot is price vs age. There is a relationship between these two variables.


Price
0.2
0.4
0.6
0.8
1.0
1955
1965
1975
Year


Price
0.2
0.4
0.6
0.8
1.0
15.0
16.0
17.0
Temp


Price
0.2
0.4
0.6
0.8
1.0
50
100
150
200
250
300
Rain


Price
0.2
0.4
0.6
0.8
1.0
400
500
600
700
800
PRain



b) Based on the plots in part a) propose a regression model for this data and fit this model. Perform the diagnostic checks for this model. Are there any problems with your model? Does your model need to be modified?

I took the variable Year out of the model because Year and Age are the same thing. That's why I was getting NA values for Age. I refit the model and got the following output.

Summary output for the second model is good with a 73.56% on multiple R-Squared and 0.1176 for RSE.

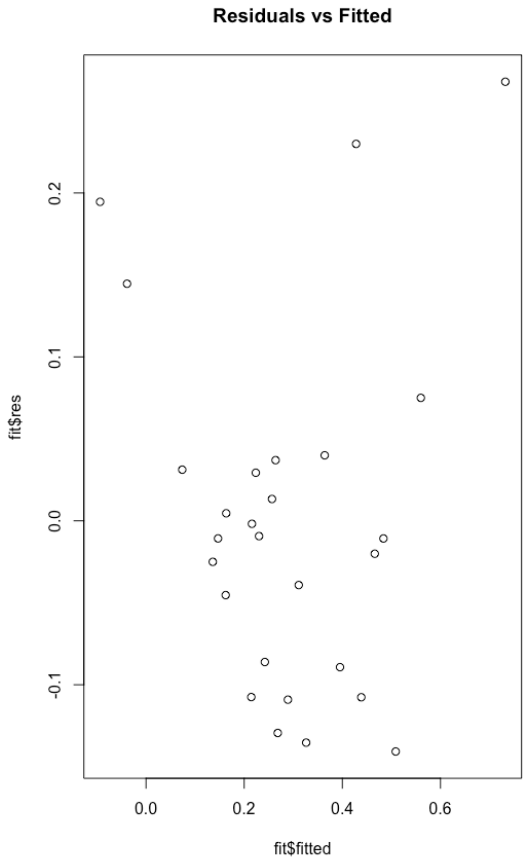
Residuals vs fitted plot doesn't look the best and suggests that we need a transformation.

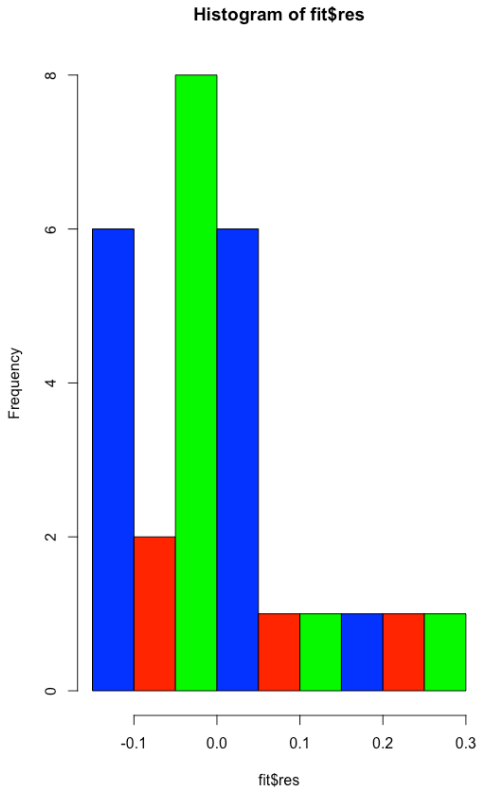
The histogram is a dead giveaway that something is wrong with the data.

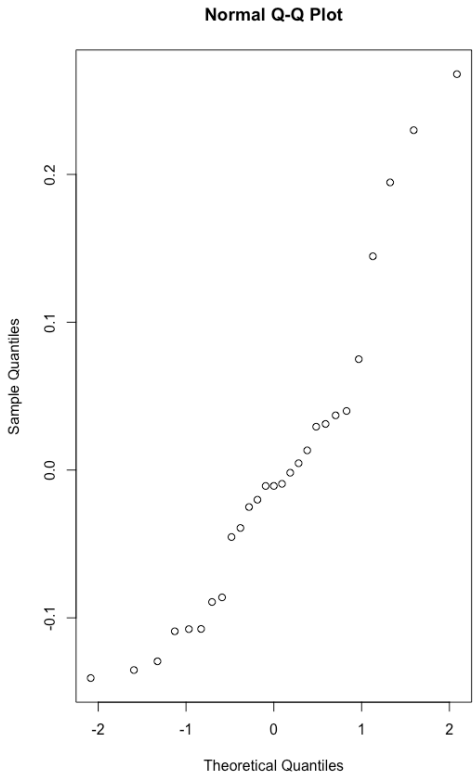
Normal qq plot is somewhat of a linear straight line but it's weak.

```
Call:
lm(formula = Price ~ +Year + Temp + Rain + PRain + Age, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14072 -0.08770 -0.01074  0.03410  0.26783

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.7951953  6.0297018   2.122  0.04533 *
Year        -0.0080519  0.0029410  -2.738  0.01201 *
Temp         0.1903096  0.0390606   4.872 7.18e-05 ***
Rain        -0.0010351  0.0003314  -3.123  0.00495 **
PRain        0.0005638  0.0001979   2.849  0.00934 **
Age                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1176 on 22 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7356,    Adjusted R-squared:  0.6875
F-statistic:  15.3 on 4 and 22 DF,  p-value: 4.017e-06
```

```
Call:
lm(formula = Price ~ +Temp + Rain + PRain + Age, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14072 -0.08770 -0.01074  0.03410  0.26783

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.1716289  0.6928899  -4.577 0.000147 ***
Temp         0.1903096  0.0390606   4.872 7.18e-05 ***
Rain        -0.0010351  0.0003314  -3.123 0.004947 **
PRain        0.0005638  0.0001979   2.849 0.009338 **
Age          0.0080519  0.0029410   2.738 0.012013 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1176 on 22 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7356,    Adjusted R-squared:  0.6875
F-statistic:  15.3 on 4 and 22 DF,  p-value: 4.017e-06
```








```
        Shapiro-Wilk normality test

data:  (fit$res)
W = 0.90843, p-value = 0.02096
```

c) If the model you fit in part a) needs to be modified, suggest a new model and fit that model. Repeat the diagnostic checks. Is there a need for further modification? If so propose a new model, fit that new model and check that model. Continue until you have a model that you are satisfied with.

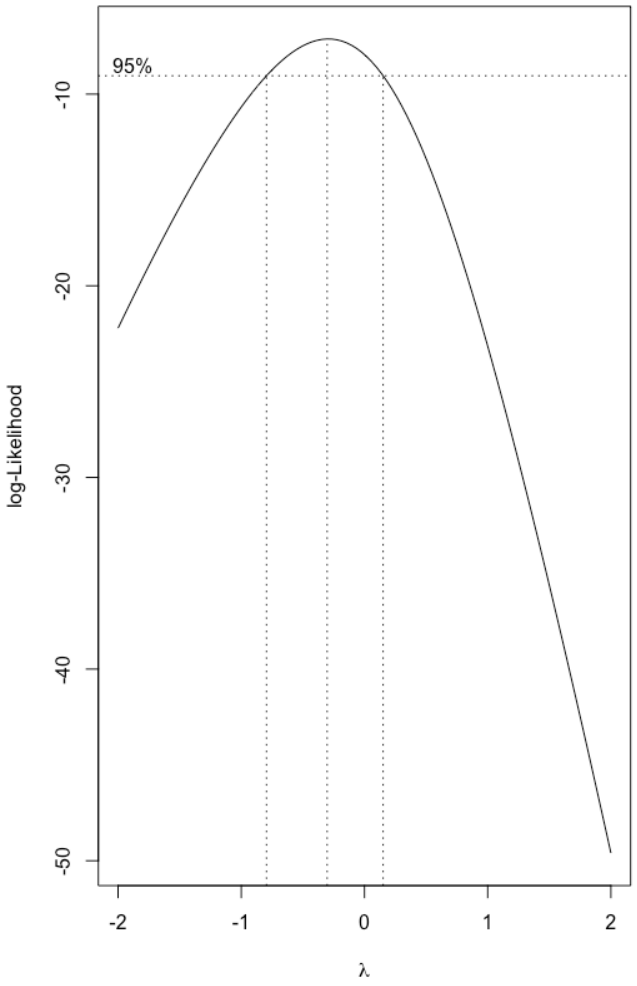
I decided to run a boxcox on my model to see if I can fix the Y variable with a transformation. The boxcox plot said we need to take the log of price since the results gave us -.5.

Then I ran a diagnostic test over the new fit model with logprice and got the following output:

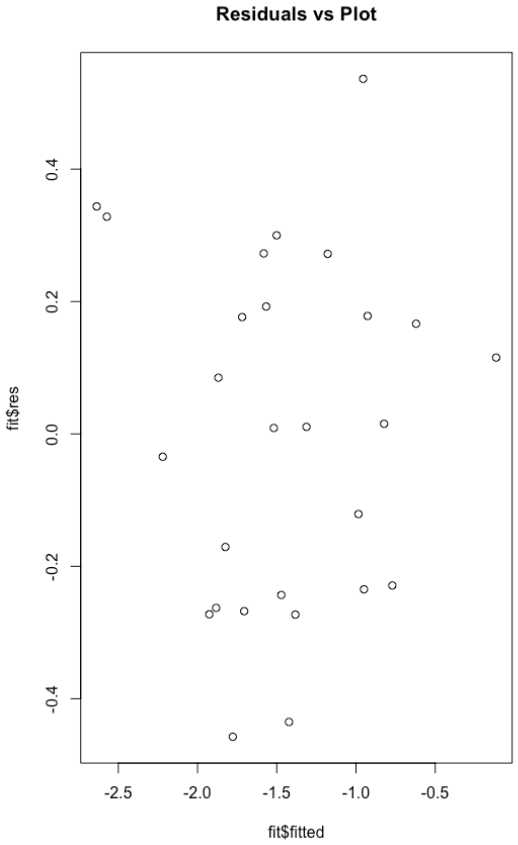
Our summary output is better with a 82% for multiple R-Squared. Our RSE increased a little bit. The Ressiduals vs Fitted plot looks a lot better.

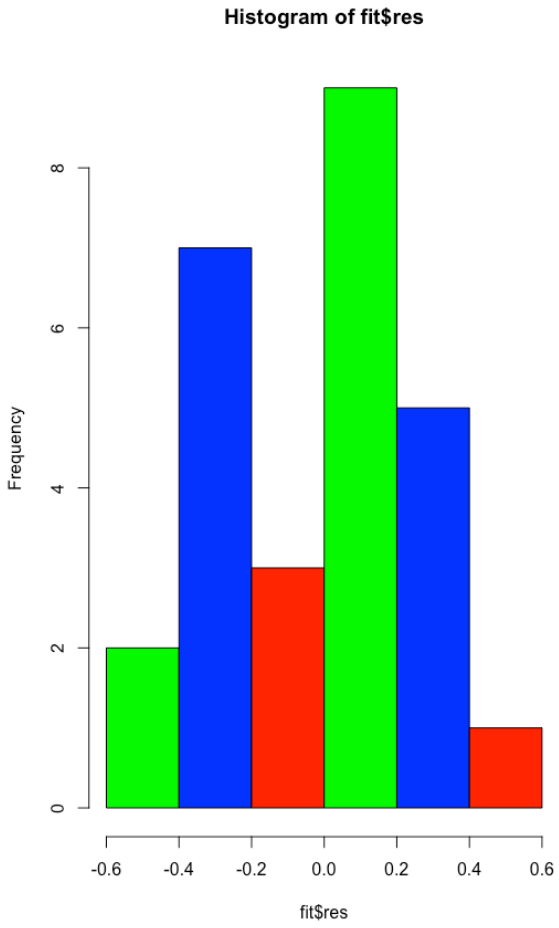
Our Histogram still isn't good and doesn't show normal distribution.

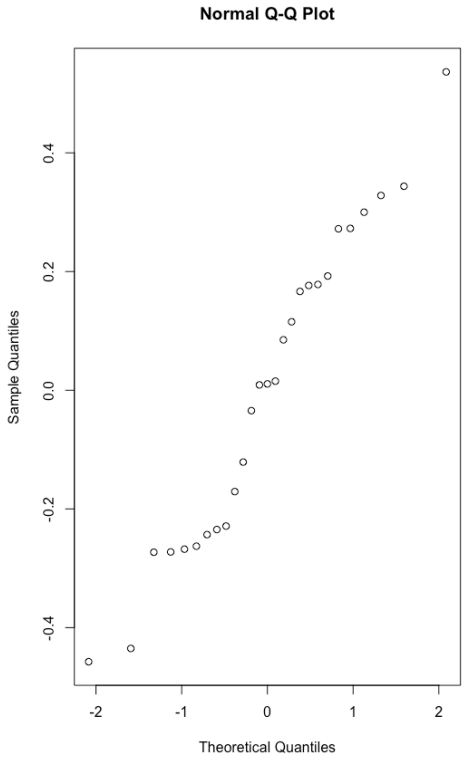
QQ-Plot for normality looks about the same.


95%
log-Likelihood
-10
-20
-30
-40
-50
-2
-1
0
1
2
λ

```
Call:
lm(formula = logprice ~ +Temp + Rain + PRain + Age, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45748 -0.23902  0.01067  0.18533  0.53642

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.216e+01  1.686e+00  -7.213 3.15e-07 ***
Temp         6.170e-01  9.502e-02   6.493 1.57e-06 ***
Rain        -3.866e-03  8.062e-04  -4.795 8.66e-05 ***
PRain        1.171e-03  4.814e-04   2.432  0.02359 *
Age          2.390e-02  7.155e-03   3.341  0.00296 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2861 on 22 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8282,    Adjusted R-squared:  0.797
F-statistic: 26.51 on 4 and 22 DF,  p-value: 3.89e-08
```

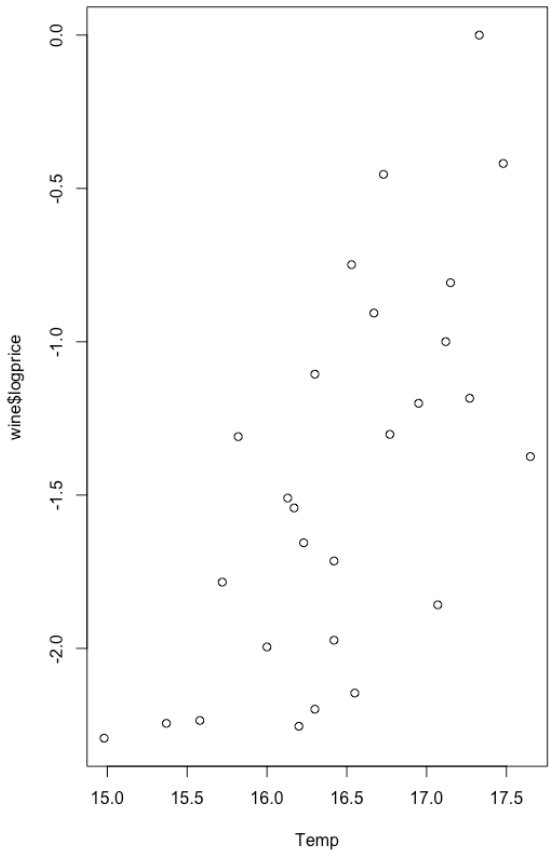


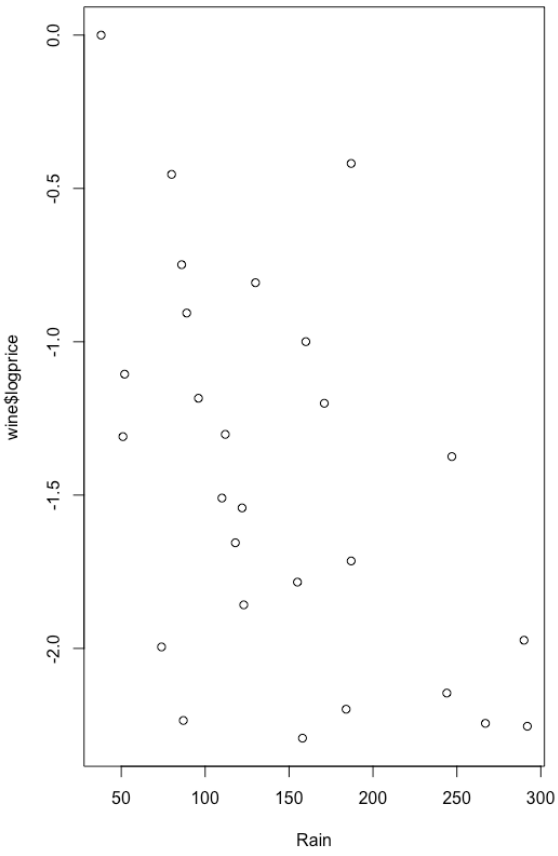


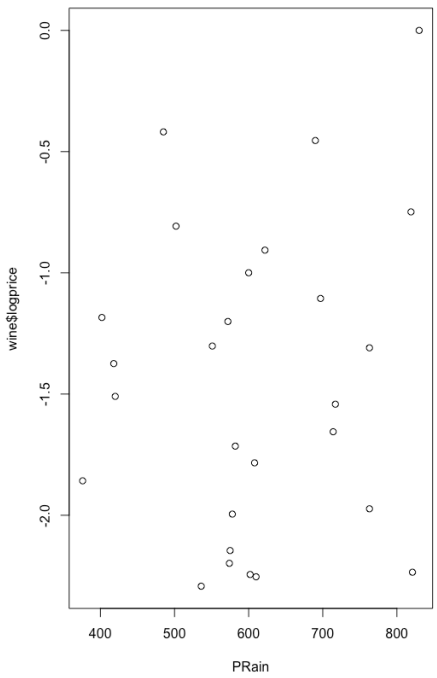


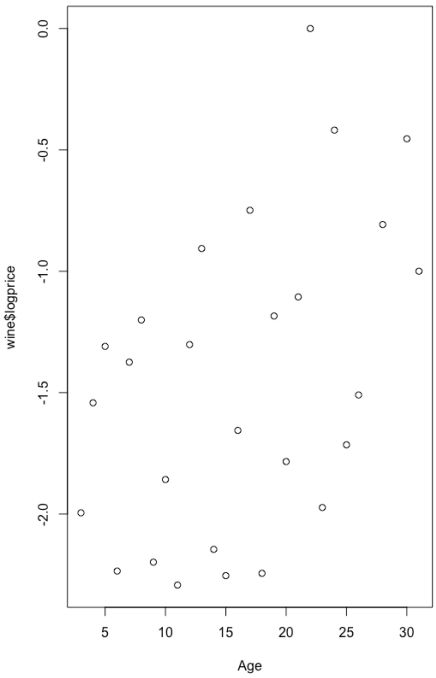
I am going to plot the logprice with each independent variable individually, so I can see what the relationship is between them. Hopefully this will show us which variable should be removed from our model.

Based on the graphs below, it looks like we can remove PRain from our model as there is no relationship when comparing the logPrice vs PRain.


wine$logprice
0.0
-0.5
-1.0
-1.5
-2.0
15.0
15.5
16.0
16.5
17.0
17.5
Temp


wine$logprice
0.0
-0.5
-1.0
-1.5
-2.0
50
100
150
200
250
300
Rain

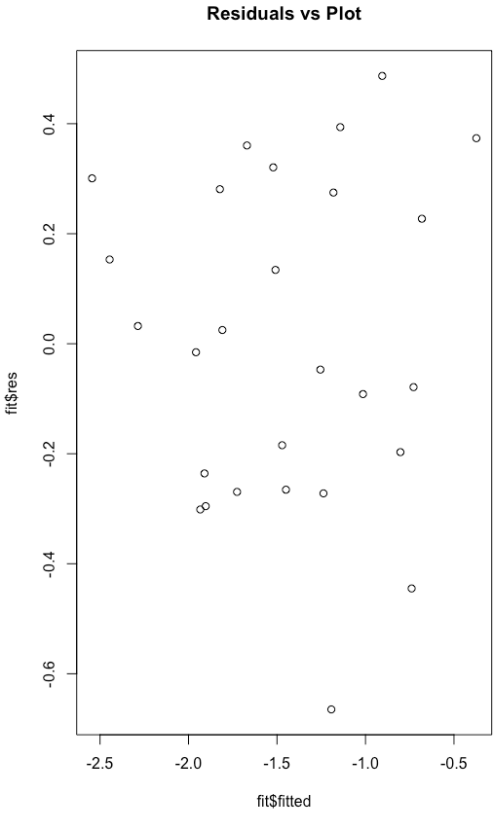

wine$logprice
0.0
-0.5
-1.0
-1.5
-2.0
400
500
600
700
800
PRain

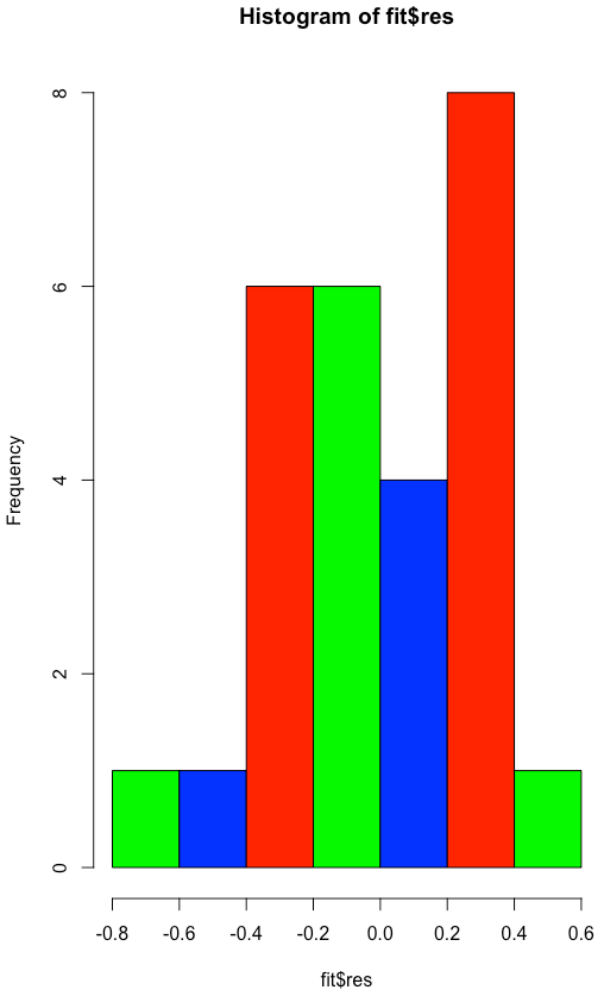

wine$logprice
0.0
-0.5
-1.0
-1.5
-2.0
5
10
15
20
25
30
Age

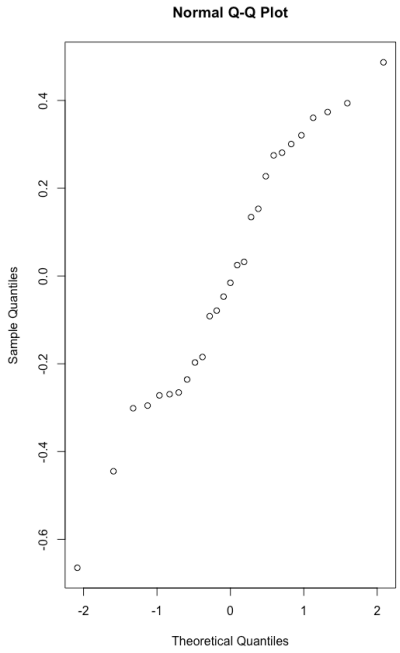
The Following is the Diagonostics for the new model without PRain:

I think this is our best model. Even though our R-Squared went down, it's till above 70% which is good. The Residuals vs Plot model looks a lot better, and our histogram looks a little better. Still not competely normal distribution. Normal QQ-Plot looks about the same and and our Shapir-Wilk Test P-Value is 35% which is a little better than our original value. While the model is not perfect, I believe that this is the best model. The R-Squared is still a good value and you always want to have the least amount of variables possible when fitting a model, "Have the simpliest model."

```
Call:
lm(formula = wine$logprice ~ +Temp + Rain + Age, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66483 -0.25059 -0.01549  0.27791  0.48689

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.006e+01  1.596e+00  -6.305 1.96e-06 ***
Temp         5.369e-01  9.820e-02   5.467 1.47e-05 ***
Rain        -4.447e-03  8.483e-04  -5.243 2.56e-05 ***
Age          2.515e-02  7.862e-03   3.199  0.00399 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3152 on 23 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.782,     Adjusted R-squared:  0.7536
F-statistic:  27.5 on 3 and 23 DF,  p-value: 8.712e-08
```








```
        Shapiro-Wilk normality test

data:  fit$res
W = 0.95907, p-value = 0.3519
```