Final Project

Mariah Diaz, Faaz Arshad, Nathan Holmes-King, Jeerapaj Daithanawong, Cassia Magsie

Make sure you click yes when a window pops up while you're running final()

Annotations of our steps and output is in the function file.

#A summary is included down below.

```
source("WineQuality.R")
```

```
## Warning: package 'glmnet' was built under R version 4.1.2
```

```
## Loaded glmnet 4.1-4
```

```
## Warning: package 'faraway' was built under R version 4.1.2
```

```
## ── Attaching packages ─────────────────────────────────────
tidyverse 1.3.1 ──
```

```
## ✔ ggplot2 3.3.5     ✔ purrr   0.3.4
## ✔ tibble  3.1.8     ✔ dplyr   1.0.7
## ✔ tidyr   1.1.3     ✔ stringr 1.4.0
## ✔ readr   2.0.1     ✔ forcats 0.5.1
```

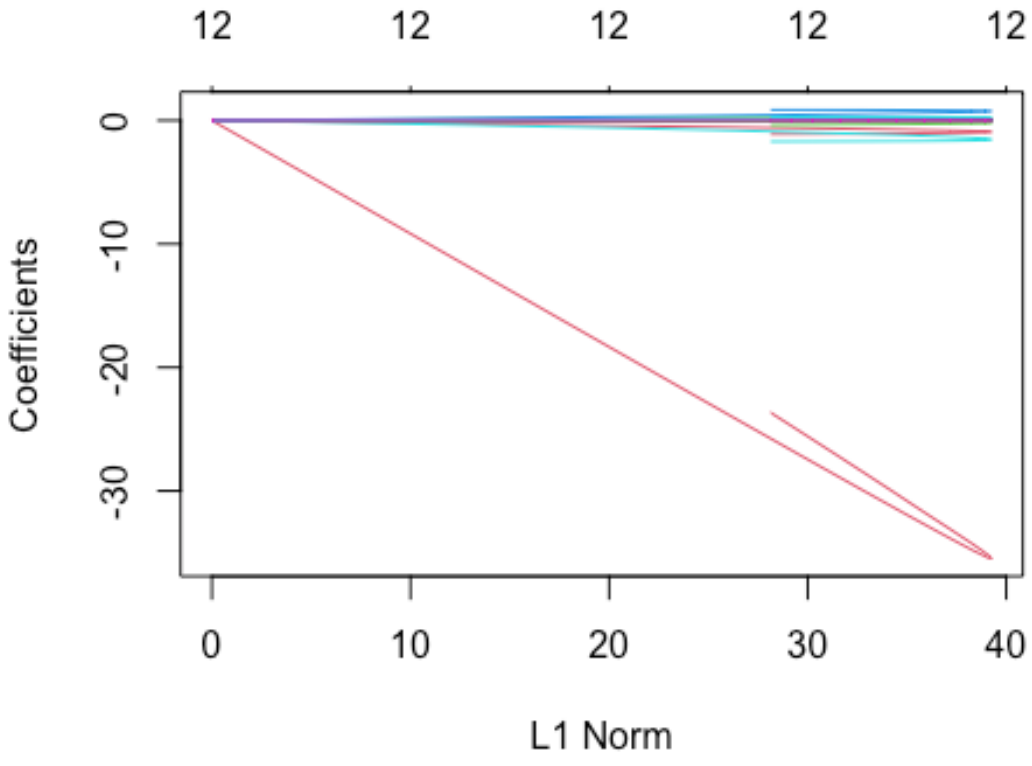
```
## Warning: package 'tibble' was built under R version 4.1.2
```

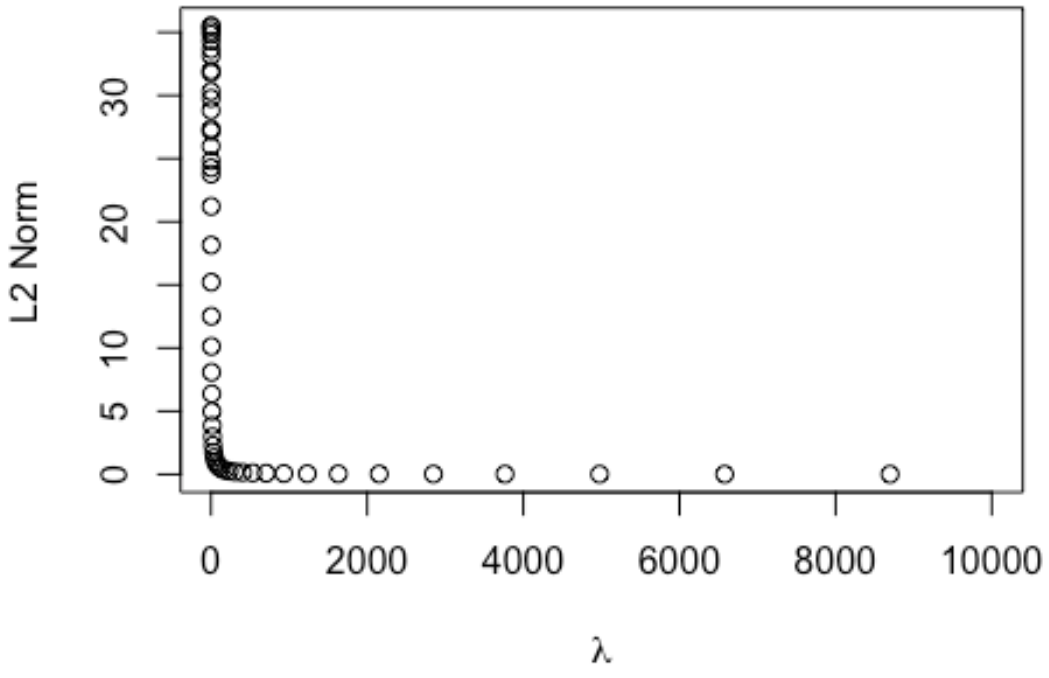
```
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ tidyr::expand() masks Matrix::expand()
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ tidyr::pack()   masks Matrix::pack()
## ✖ tidyr::unpack() masks Matrix::unpack()
```

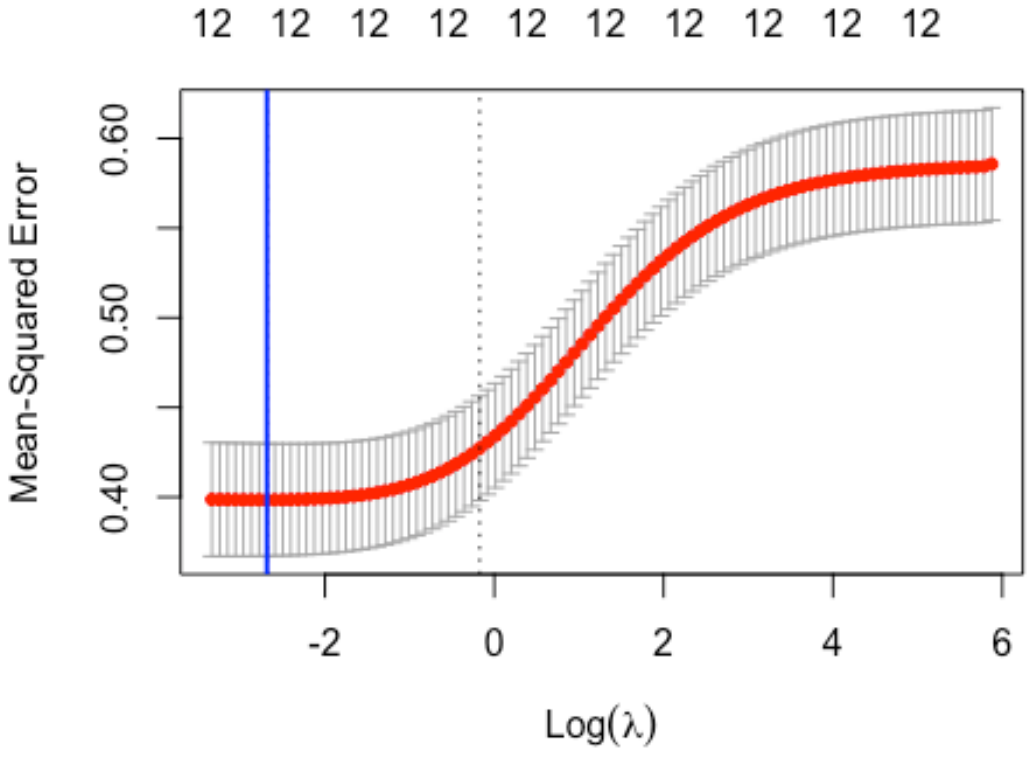
```
## Warning: package 'performance' was built under R version 4.1.2
```

```
## Warning: package 'insight' was built under R version 4.1.2
```

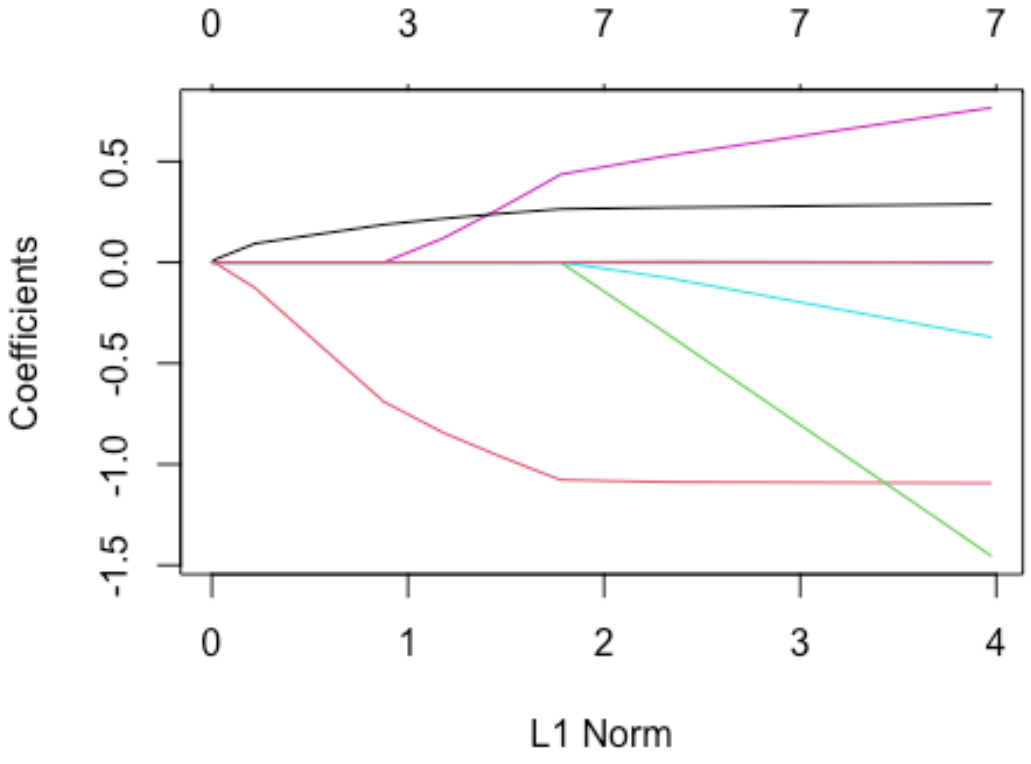
```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract

## Warning: package 'ISLR2' was built under R version 4.1.2

Final()

## [1] "Step 1: LASSO, RIDGE, OLS"
## [1] "plot l1 norm of the coefficients"
```

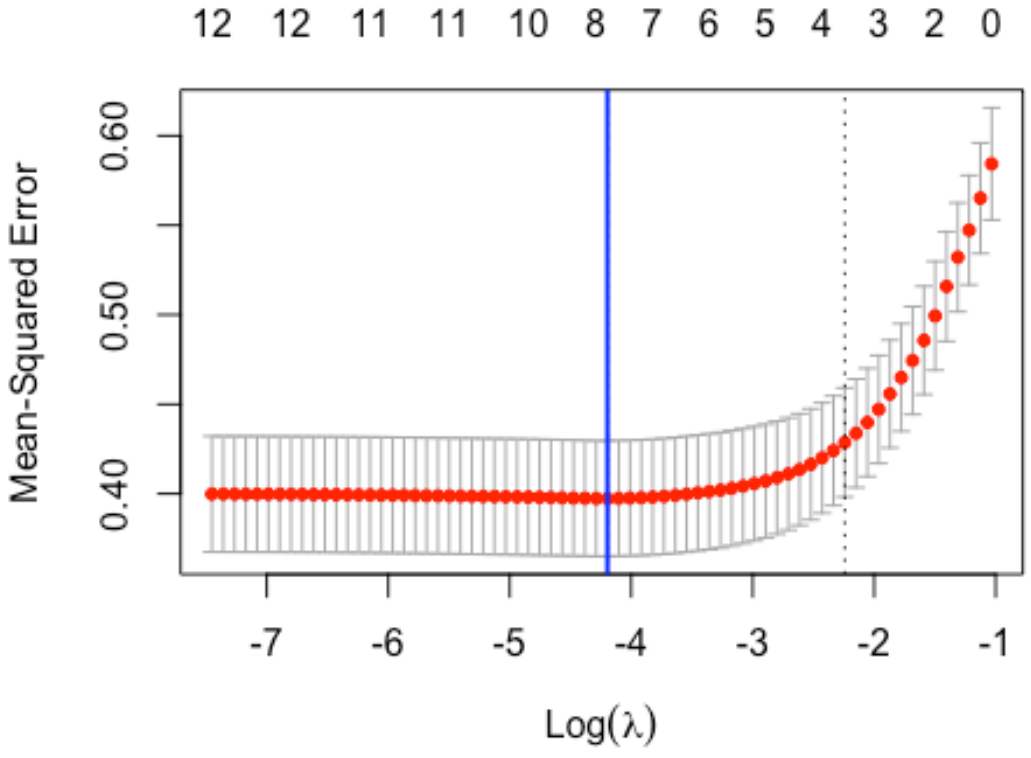

12
12
12
12
12
0
-10
-20
-30
Coefficients
0
10
20
30
40
L1 Norm


L2 Norm
0
5
10
20
30
0
2000
4000
6000
8000
10000
λ

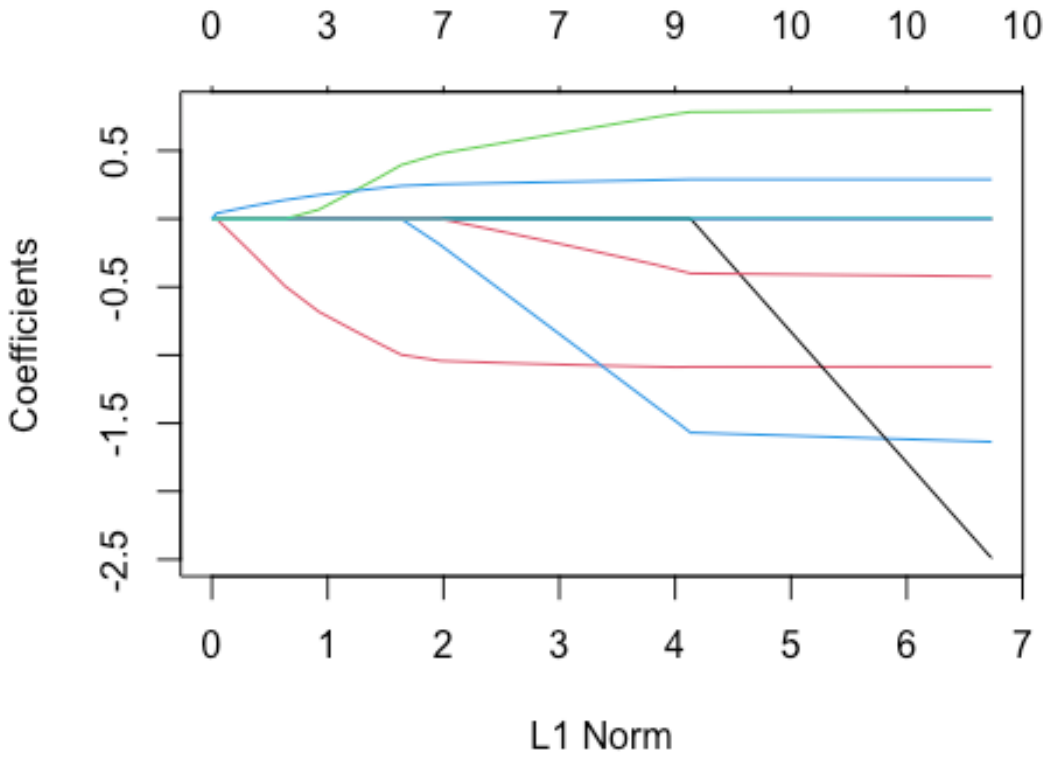


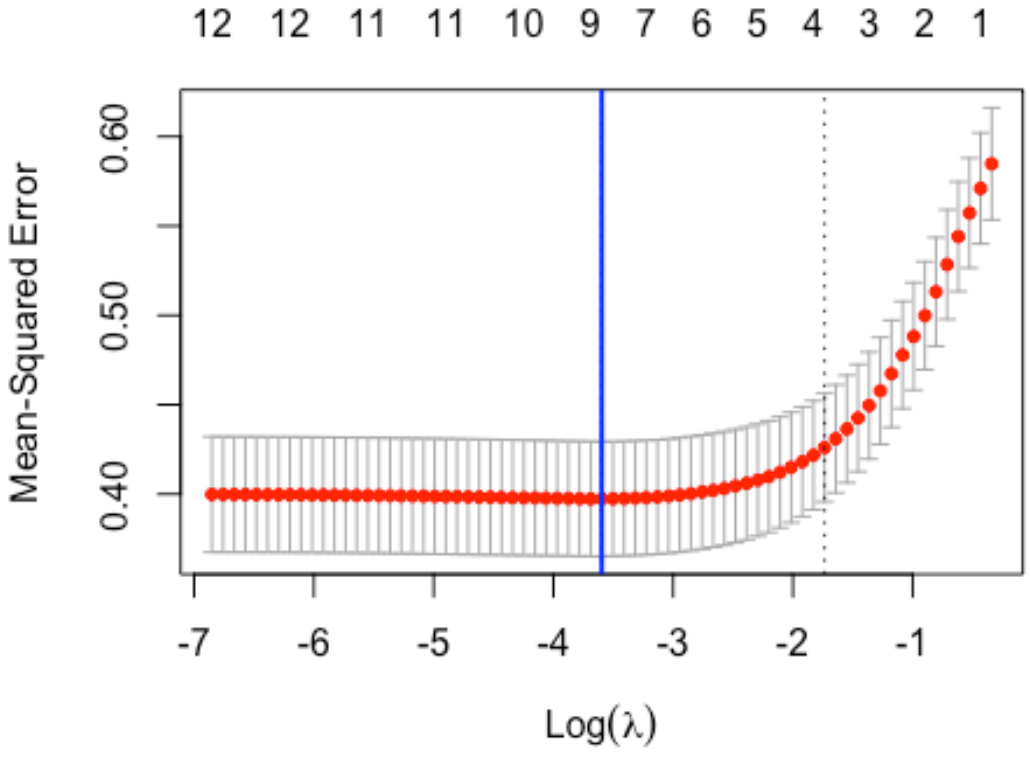

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm =
na.rm):
## collapsing to unique 'x' values
```


0
3
7
7
7
Coefficients
0.5
0.0
-0.5
-1.0
-1.5
0
1
2
3
4
L1 Norm

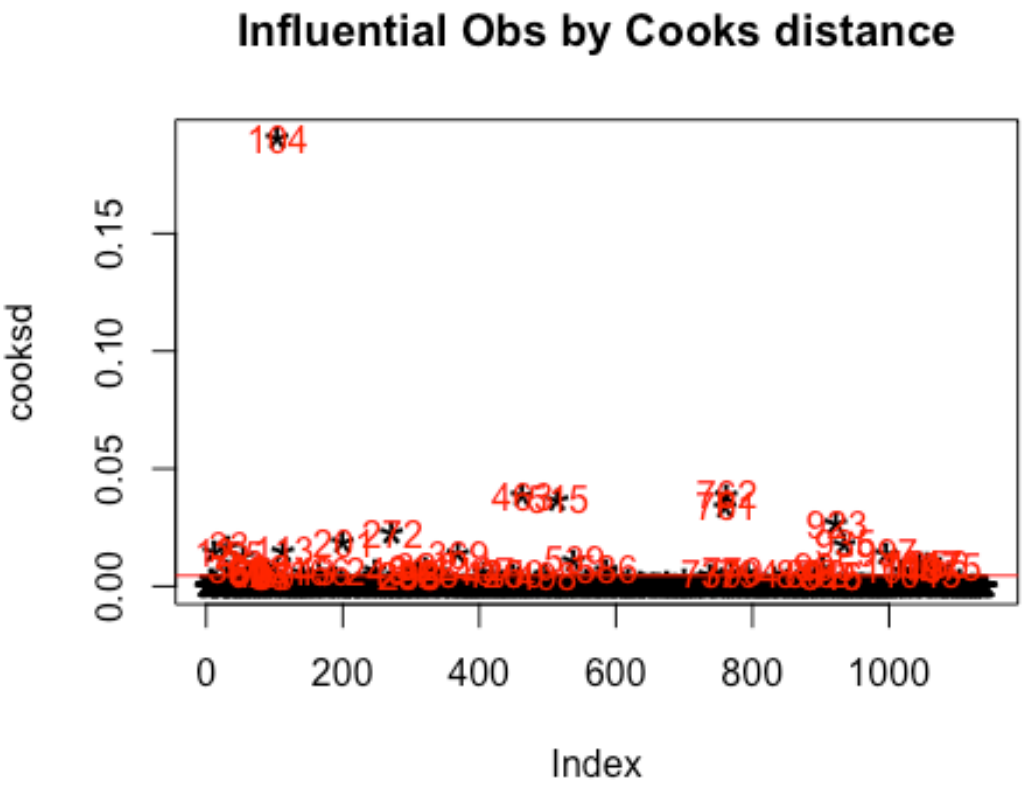



```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm =
na.rm):
## collapsing to unique 'x' values
```

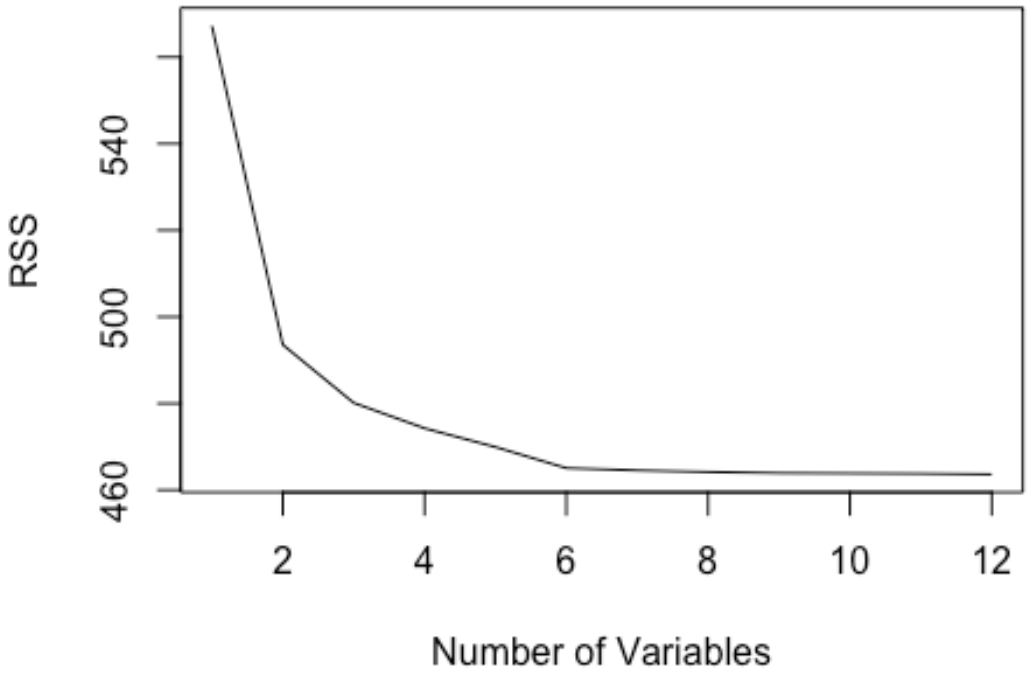

0 3 7 7 9 10 10 10
Coefficients
0.5
-0.5
-1.5
-2.5
0 1 2 3 4 5 6 7
L1 Norm



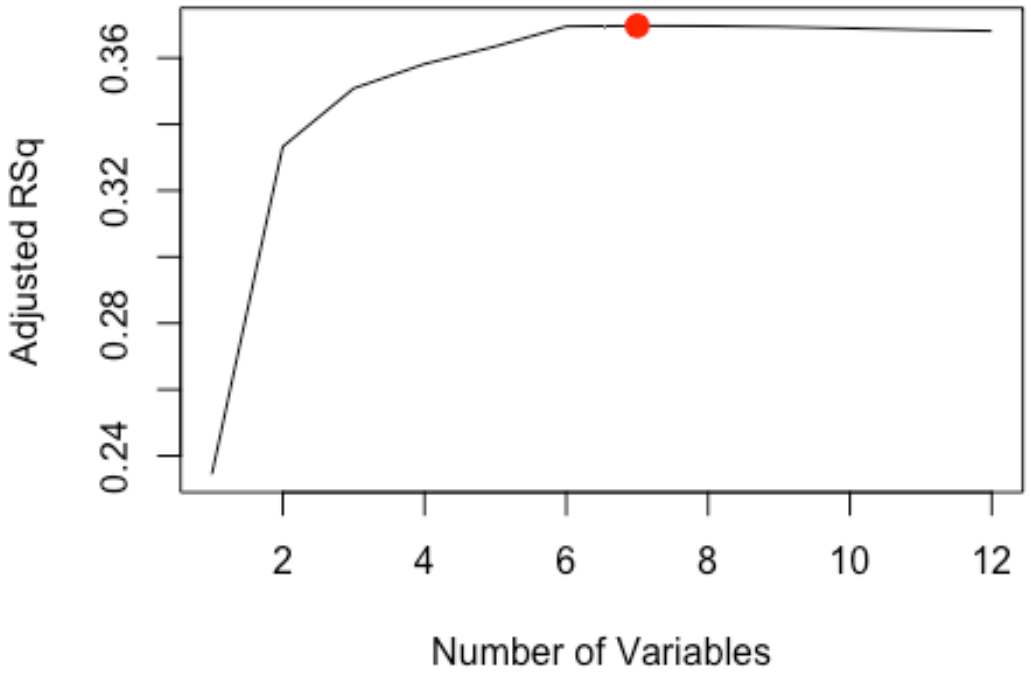

```
## [1] "Step 2: Dealing with Outliers"
```



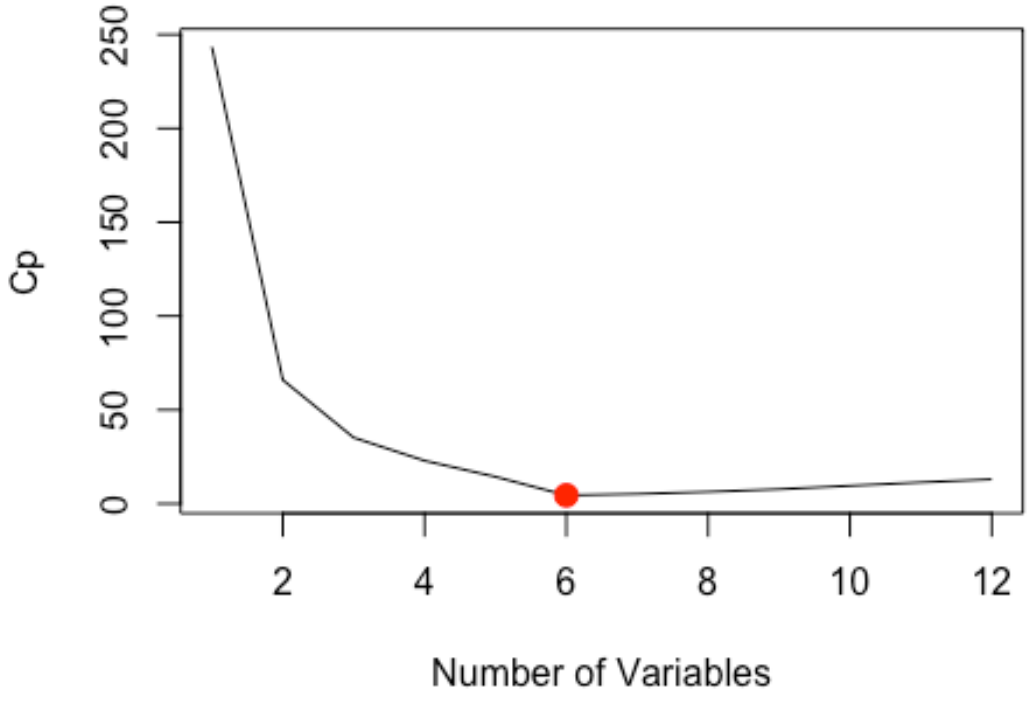

```
## [1] "Outliers/influential points can be observed in the plot
highlighted with red labels."
## [1] "In our data set we have decided to removes all points for
which Cook's distance"
## [1] "for that point is at least 4 times the mean Cook's distance
for all the points."
## [1] "Step 3:Performing Model Selection"
##
## The downloaded binary packages are in
##  /var/folders/2x/7tbbl2m17z3fyl7wylk2t4qc0000gn/T//RtmppKoRhH/
downloaded_packages
```



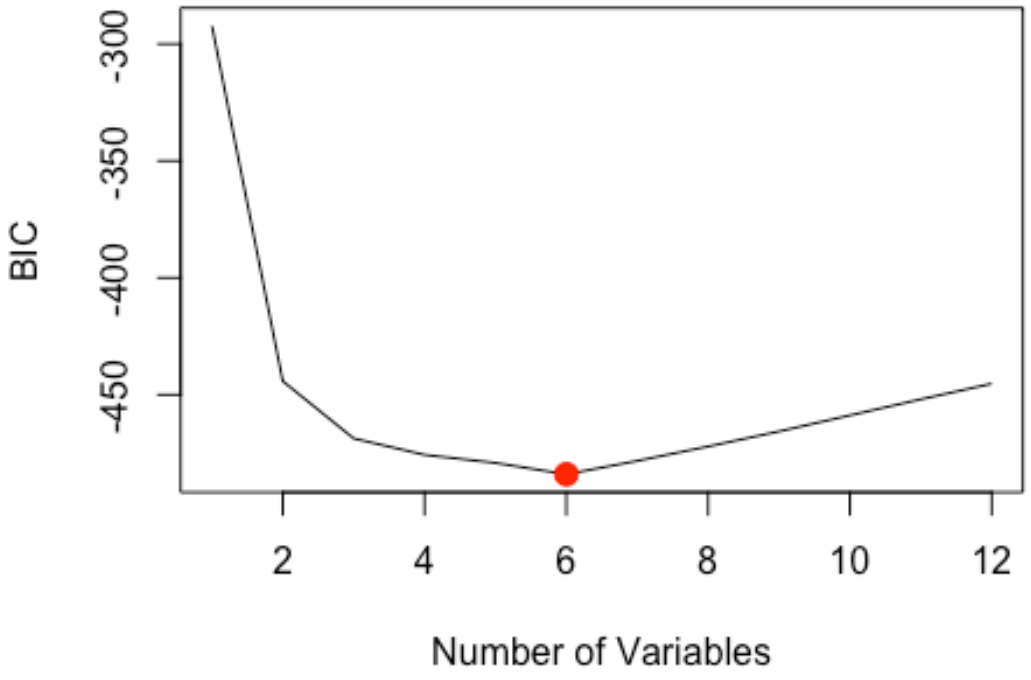

```
## [1] "Looking to minimize the RSS"
```



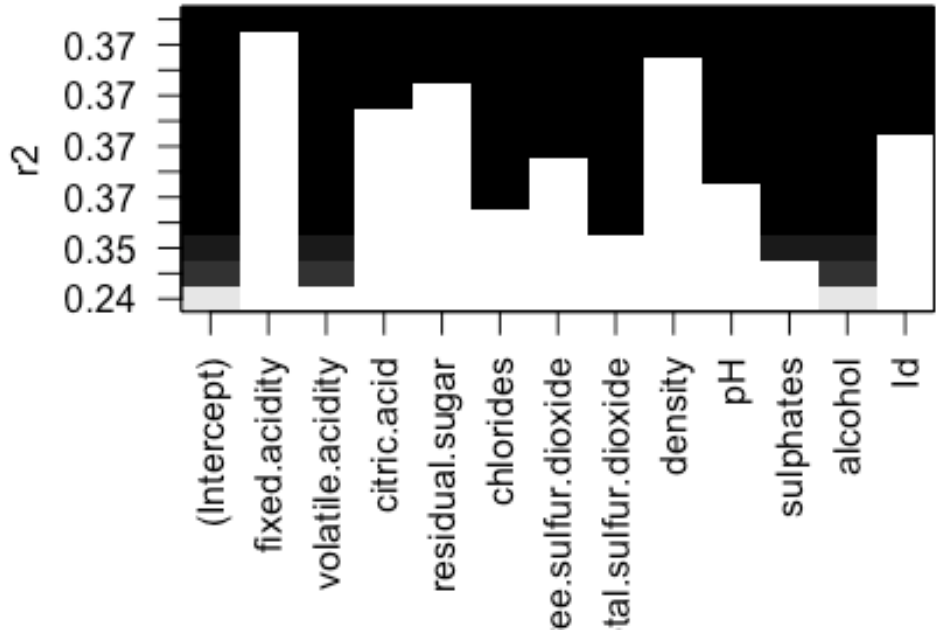

```
## [1] "Looking for highest R squared"
```



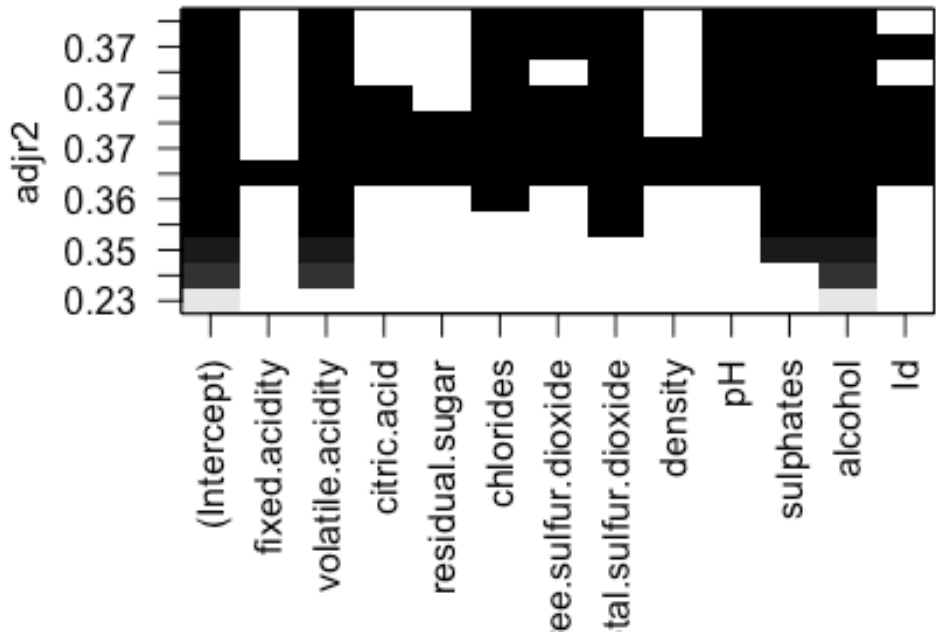

```
## [1] "Looking for lowest CP"
```



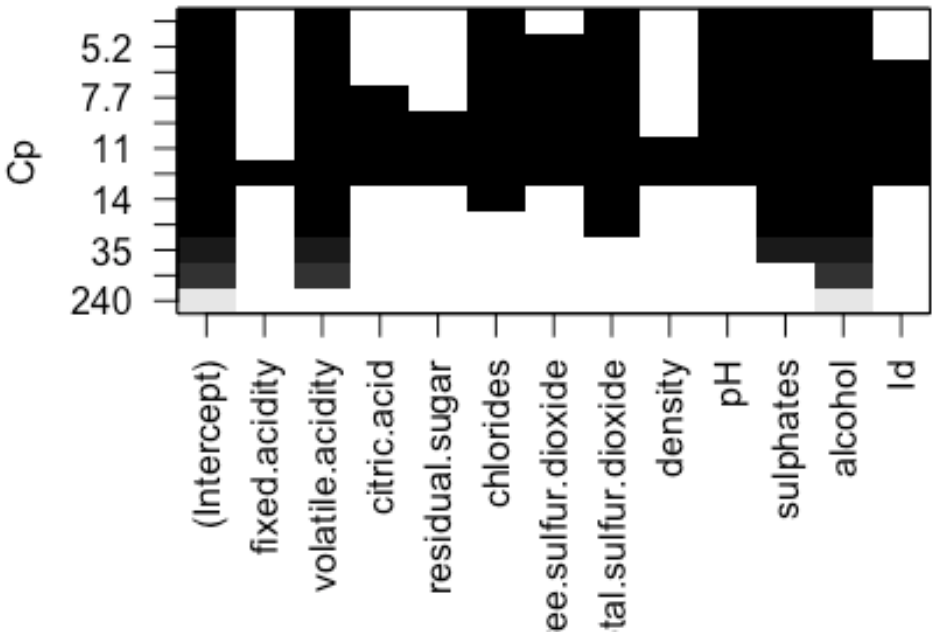

```
## [1] "Want smallest BIC value"
## [1] "BIC Plot"
```



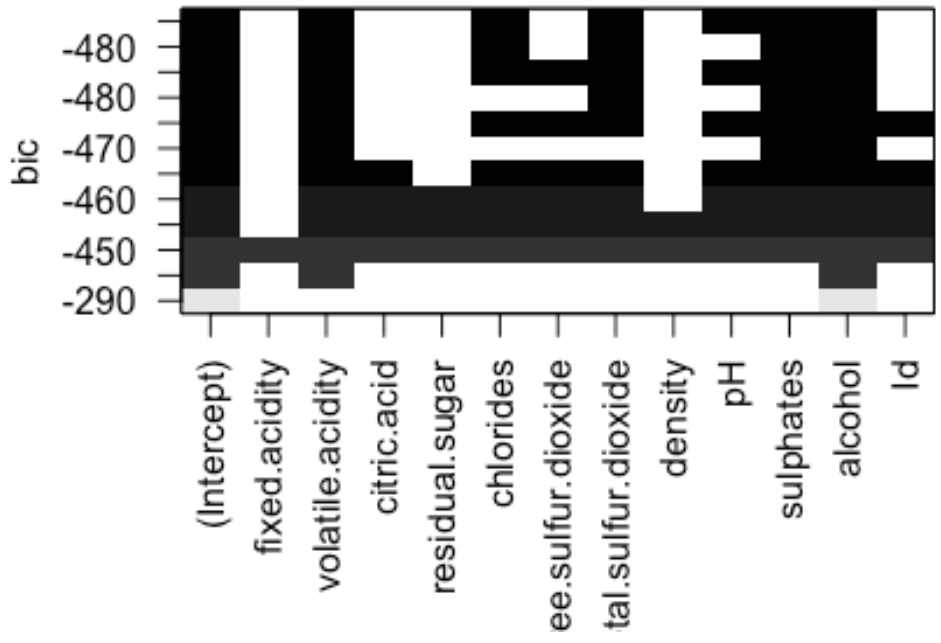

```
## [1] "Adjusted R squared plot"
```


adjr2
0.37
0.37
0.37
0.36
0.35
0.23
(Intercept)
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
ee.sulfur.dioxide
tal.sulfur.dioxide
density
pH
sulphates
alcohol
Id

```
## [1] "Cp plot"
```


Cp
5.2
7.7
11
14
35
240
(Intercept)
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
ee.sulfur.dioxide
tal.sulfur.dioxide
density
pH
sulphates
alcohol
Id

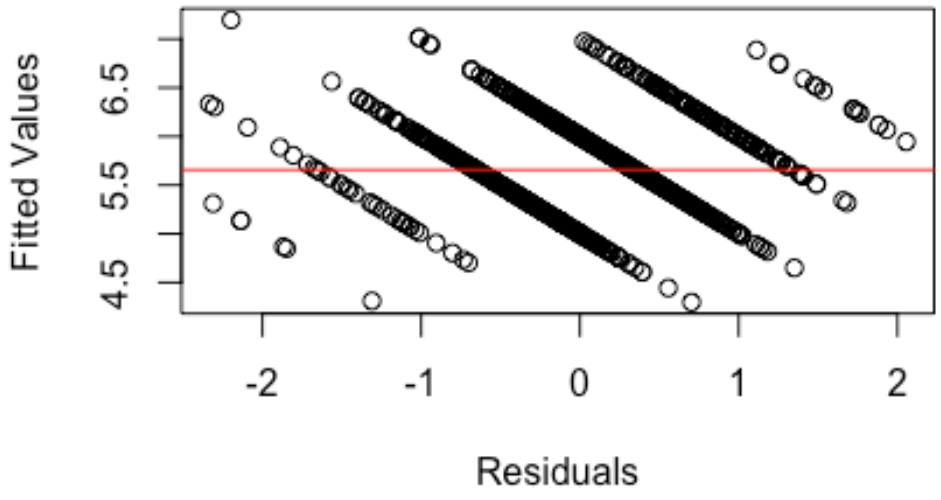
```
## [1] "Bic plot"
```



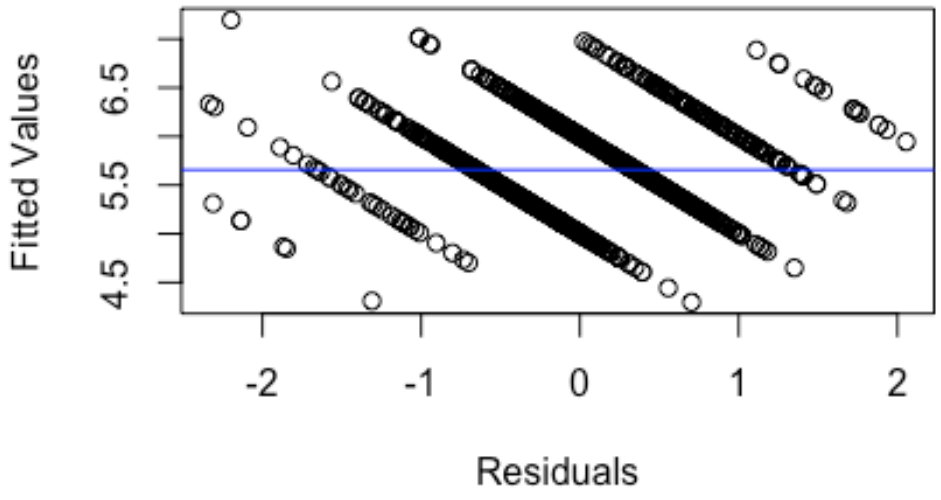

```
##
## Call:
## lm(formula = d$quality ~ ., data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49730 -0.37125 -0.04815  0.44220  1.97744
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.303e+01  2.482e+01   0.928 0.353614
## fixed.acidity         1.882e-02  3.054e-02   0.616 0.537799
## volatile.acidity     -1.125e+00  1.408e-01  -7.994 3.20e-15 ***
## citric.acid          -1.221e-01  1.733e-01  -0.704 0.481357
## residual.sugar        1.400e-02  1.846e-02   0.758 0.448432
```

```
## chlorides            -1.721e+00  4.976e-01  -3.458 0.000564 ***
## free.sulfur.dioxide   2.890e-03  2.607e-03   1.109 0.267723
## total.sulfur.dioxide -2.977e-03  8.608e-04  -3.458 0.000564 ***
## density              -1.880e+01  2.532e+01  -0.743 0.457880
## pH                   -4.342e-01  2.244e-01  -1.935 0.053255 .
## sulphates             8.643e-01  1.340e-01   6.452 1.64e-10 ***
## alcohol               2.830e-01  3.139e-02   9.016  < 2e-16 ***
## Id                   -4.528e-05  4.576e-05  -0.990 0.322604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6405 on 1130 degrees of freedom
## Multiple R-squared:  0.3748, Adjusted R-squared:  0.3681
## F-statistic: 56.45 on 12 and 1130 DF,  p-value: < 2.2e-16
##
## [1] "We can identify which variables are significant by looking at
the"
## [1] "codes in the summary. For our wine set this would include,
volatile-"
## [1] "acidity, chlorides, total sulfer, sulphates, and alcohol"
## [1] "These are the variables we have decided to include in our
model."
## [1] "5 of the most signifcant variables were chosen based on our"
## [1] "exploratory plots and summaries."
## [1] "F-Test:"
## Analysis of Variance Table
##
## Model 1: d$quality ~ d$volatile.acidity + d$chlorides +
d$total.sulfur.dioxide +
##     d$sulphates + d$alcohol
## Model 2: d$quality ~ fixed.acidity + volatile.acidity + citric.acid
+
##     residual.sugar + chlorides + free.sulfur.dioxide +
total.sulfur.dioxide +
##     density + pH + sulphates + alcohol + Id
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   1137 469.94
## 2   1130 463.64  7    6.3015 2.1941 0.03247 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "H0: No significant difference in these two models: P-
value>0.05"
## [1] "Reduced model is significantly better: p-value<0.05"
## [1] "When we run the f-test on our chosen Wine data set, we are
comparing the"
## [1] "full model with the reduced model with a select few variables.
From the"
## [1] "output we can see our p-value is 0.03247 which is less than
our"
## [1] "significance level of 0.05 and we reject the null hypothesis
that there"
## [1] "is no significant difference between the two models. From this
output we"
## [1] "have more evidence to go with the reduced chosen model. "
## [1] "Step 5:"
## [1] "Observing linearity"
## [1] "We are using linearity to see if there is any noticeable trend
between"
## [1] "the residuals."
```

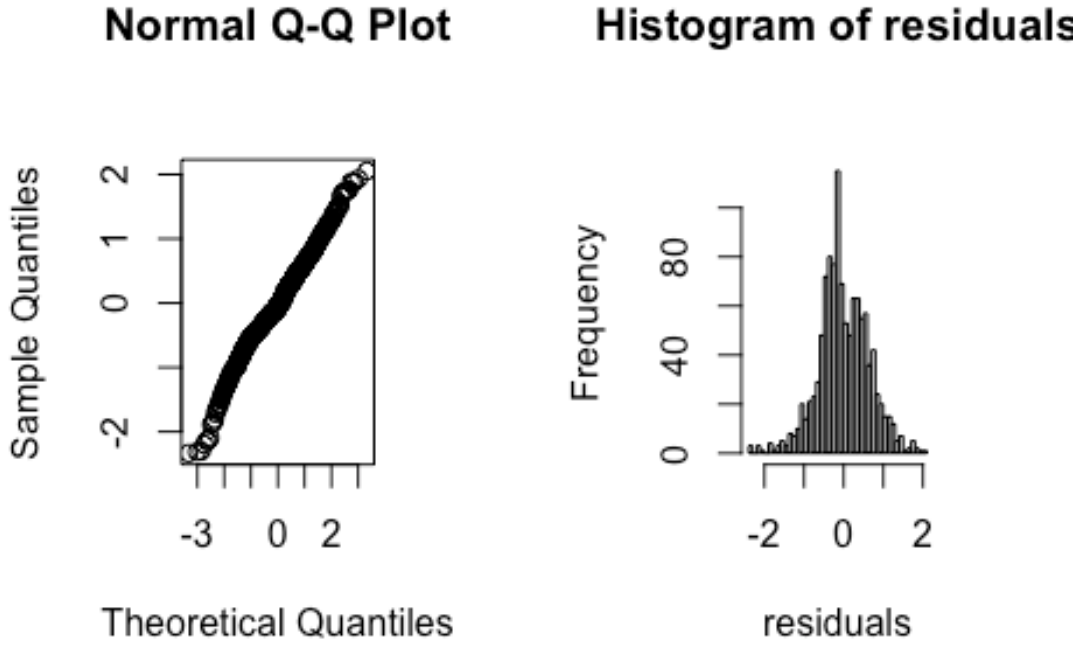



```
## [1] "Homoscedasticity:"
## [1] "If the data is homoscedastic, the points on the plot should be
evenly"
## [1] "distributed #around the zero line."
## [1] "We use this to check whether the spread of the residuals
change as a"
## [1] "function of the predictors and/or fitted values."
```

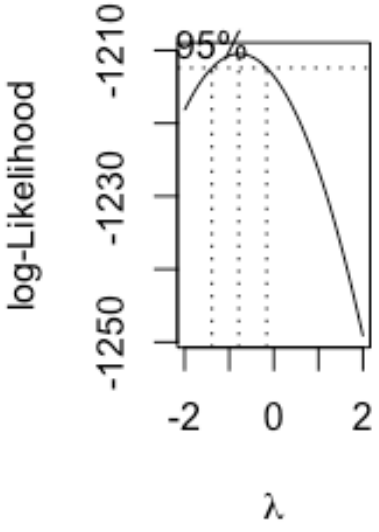



```
## [1] "Testing for Normality"
## [1] "We use normality to check if the error terms are iid normal."

## [1] "Plots appear to show normality."

## Warning in ks.test(residuals, normal.sample): p-value will be
approximate in the
## presence of ties

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  residuals and normal.sample
## D = 0.078148, p-value = 0.002963
## alternative hypothesis: two-sided
##
```

```
## [1] "with p-value<0.05 we fail to reject the Null Hypothesis"
## [1] "Step 6: Transformations"

## Warning: package 'MASS' was built under R version 4.1.2

##
## Attaching package: 'MASS'

## The following object is masked from 'package:ISLR2':
##
##     Boston

## The following object is masked from 'package:dplyr':
##
##     select
```

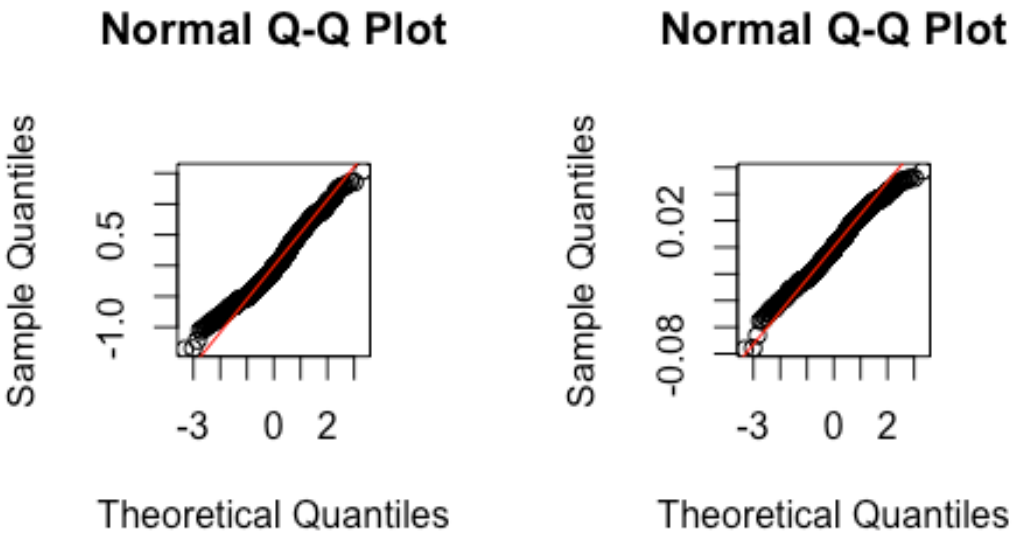



```
## [1] "Find optimal lambda for Box-Cox transformation"
```




```
## [1] "Q-Q plot for orignial Model:Right"
```

```
## [1] "Q-Q plot for Box-Cox transformed model:Left"
```




```
## [1] "Data appears to be linear."
```

The first step of the project compares MSE's of LASSO, Ridge and OLS techniques. We are looking for the technique that gives us the lowest MSE. Important to note Ridge regression tends to have a model that is less interpretable, however lower variance with higher bias. Lasso is preferred for prediction rather than inference so it is important to choose a method with outcome in mind. In step two were are testing for outliers and influential points/ points of high leverage using Cook's distance. We then remove all outliers for which cooks distance is 4 times the mean of the other points. Visually a plot is added with outliers highlighted in red. The third step is where we perform model selection using various metrics(MSE,AIC,BIC,Mallow's CP, and Adjusted R2). We used our results to determine which variables among our data set were the most significant. When it came to R-squared and adjusted R-squared we looked for a higher value, while with CP and BIC we were looking for variables that gave us a lower value. In the end the model we chose included the variables volatile-acidity, chlorides, total sulfur dioxide, sulphates, and alcohol. Which through our analysis we found to be the most significant. (The 5 variables were present at

the top of all metric plots.) Step four was performing an F-test.With the F-test we are trying to determine whether the linear regression model we have chosen is a better fit to the data than the full model. Our H0 is there is no significant difference between the two models. While the HA: says our chosen model is significantly better. At a significance level of 0.05 we are looking for the p-value of our f-statistic to be less than 0.05 in order to reject the null hypothesis and have evidence to support the new model we have chosen. When we run the f-test on our chosen Wine data set, we are comparing the full model with the reduced model with a select few variables. From the output we can see our p-value is 0.03247 which is less than our significance level of 0.05 and we reject the null hypothesis that there is no significant difference between the two models. From this output we have more evidence to go with the reduced chosen model. In the fifth step we perform diagnostics tests. We want to get the expected behavior from the residuals. We have chosen our 5 selected models to see if there is any random fluctuations around 0, if there is any noticeable trend between the residuals, if the spread of the residuals change as a function of the predictors, and if the error terms are iid normal using the 3 diagnostic methods which are linearity, homoscedasticity, and normality. Using linearity, we see a pattern trend on the plot. Homoscedasticity method shows that it is somewhat mostly around 0, and lastly the normality shows the K-S test has p-value 0.04086 that is considered weak evidence for not rejecting the null hypothesis. In the last test we determine which transformations if any are appropriate. We apply Box Cox to our data and find for the Wine data set the best lambda value to be -1. We then plot the data to show pre and post transformation.