MA4710

FINAL PROJECT

VAISHNAVI JANGILI

12/13/2022

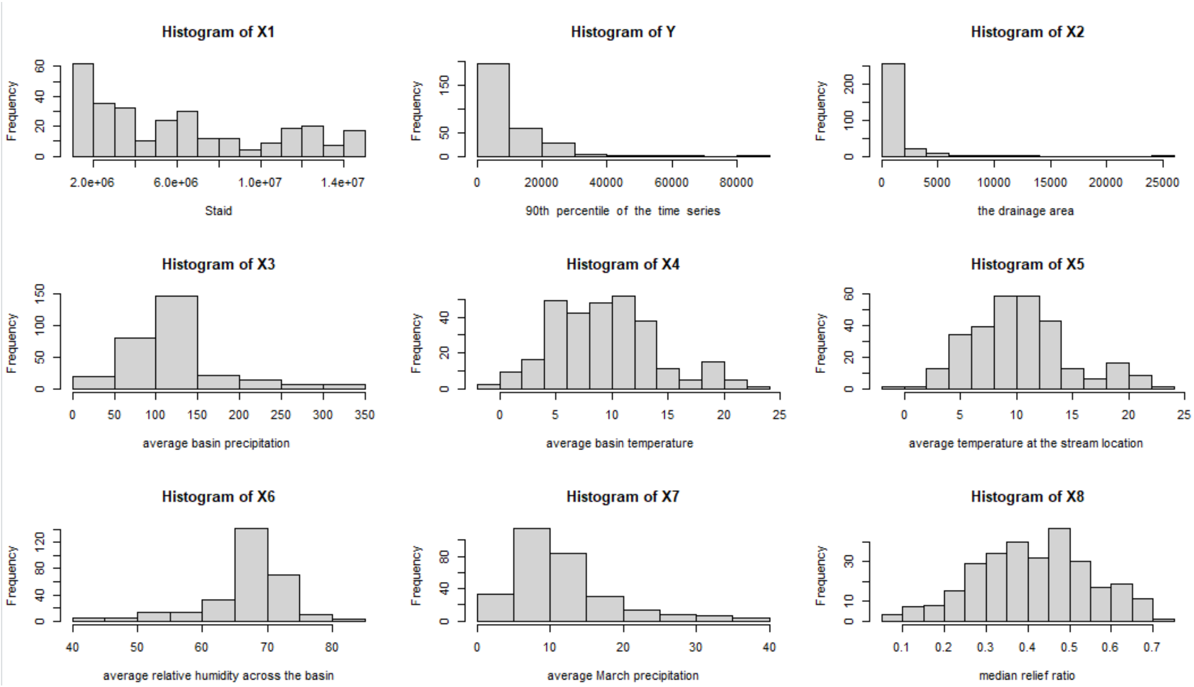
1. INTRODUCTION:

We were provided with a dataset called **streamflow.csv** which contains the 90th percentile maximum streamflow. It has 294 variables and 9 variables. The following is the description of the variables of the dataset:

- The response variable, Y (max90), is the 90th percentile of the time series of annual daily maxima.
- X1 (STAID) is, the stream identification number.
- X2 (DRAIN_SQKM) the drainage area.
- X3(PPTAVG_BASIN) the average basin precipitation.
- X4(T_AVG_BASIN) the average basin temperature.
- X5(T_AVG_SITE) the average temperature at the stream location.
- X6(RH_BASIN) the average relative humidity across the basin.
- X7(MAR_PPT7100_CM) the average March precipitation.
- X8(RRMEDIAN) the median relief ratio.

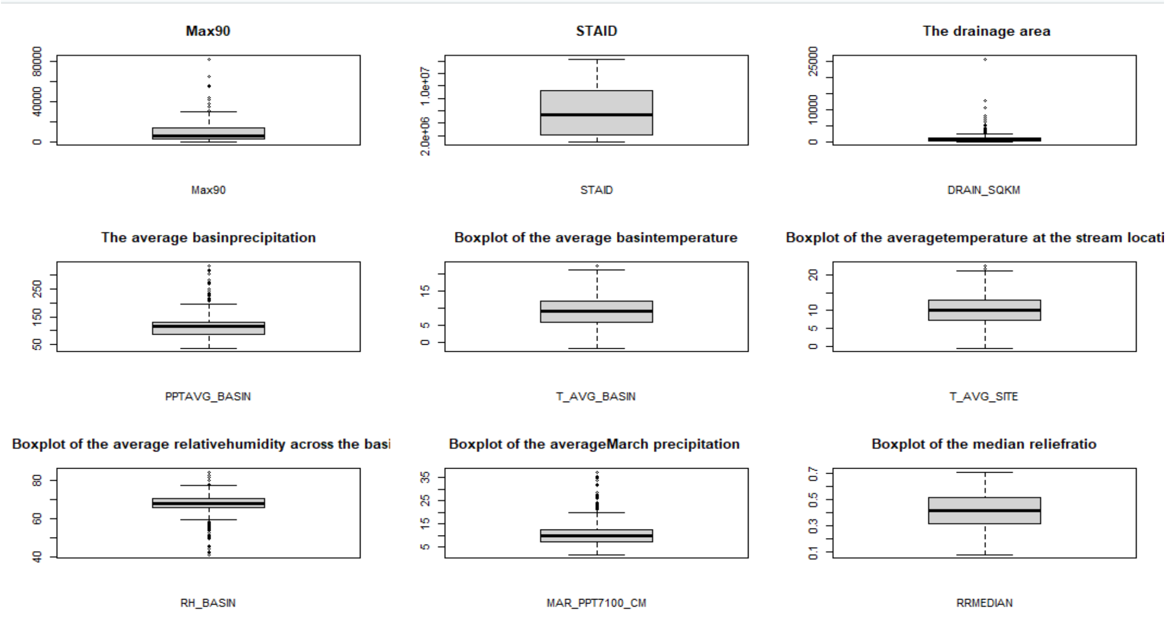
```
> summary(streamflow)
       X1                  Y                   X2                   X3
 Min.   : 1013500   Min.   :   16.03   Min.   :    5.377   Min.   : 37.78
 1st Qu.: 2065500   1st Qu.: 2231.00   1st Qu.:  208.686   1st Qu.: 88.46
 Median : 5362000   Median : 5646.00   Median :  450.199   Median :114.68
 Mean   : 5940630   Mean   : 9272.69   Mean   : 1102.691   Mean   :120.17
 3rd Qu.: 9223000   3rd Qu.:13670.00   3rd Qu.: 1151.567   3rd Qu.:131.41
 Max.   :14325000   Max.   :81900.00   Max.   :25791.040   Max.   :334.17
       X4                X5                X6                X7
 Min.   :-1.580   Min.   :-0.40   Min.   :41.11   Min.   : 1.739
 1st Qu.: 5.908   1st Qu.: 7.30   1st Qu.:65.74   1st Qu.: 7.304
 Median : 9.044   Median :10.00   Median :67.79   Median : 9.876
 Mean   : 9.415   Mean   :10.34   Mean   :66.69   Mean   :11.408
 3rd Qu.:12.189   3rd Qu.:12.90   3rd Qu.:70.24   3rd Qu.:12.261
 Max.   :22.500   Max.   :22.50   Max.   :84.20   Max.   :37.370
       X8
 Min.   :0.08042
 1st Qu.:0.31652
 Median :0.41379
 Mean   :0.41466
 3rd Qu.:0.51370
 Max.   :0.71084
>
```

Above is the summary of all the numeric variables (Y and X1 through X8) provided in the data set streamflow.

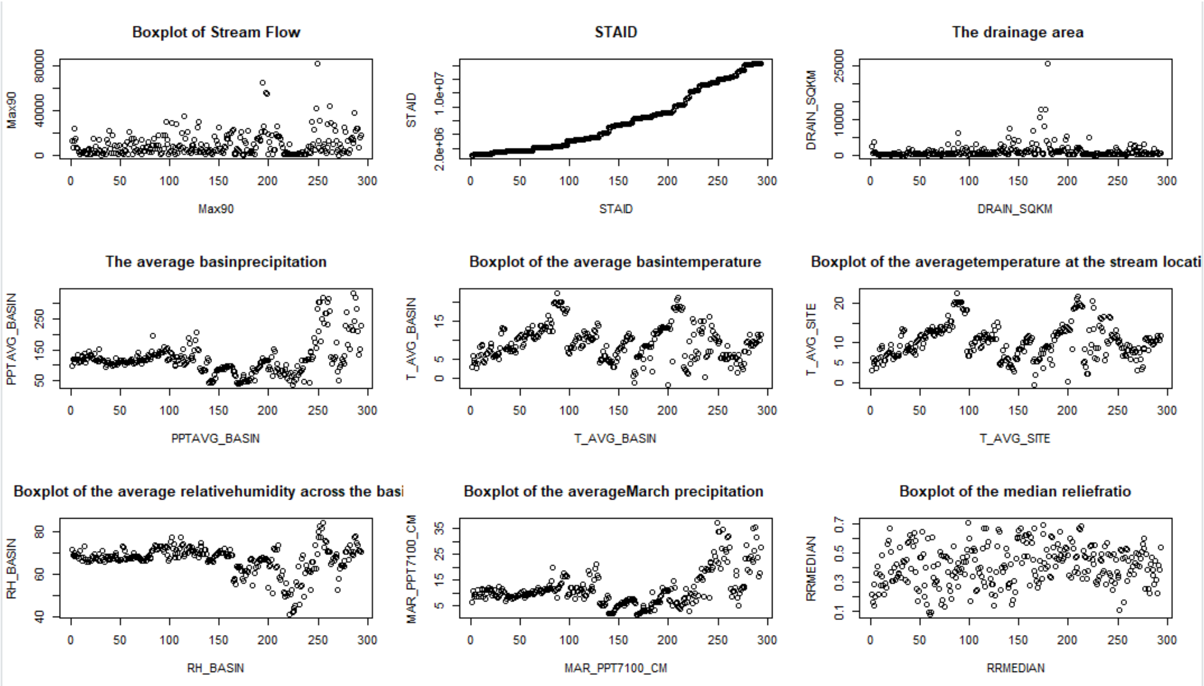


From the above figures we can interpret that,

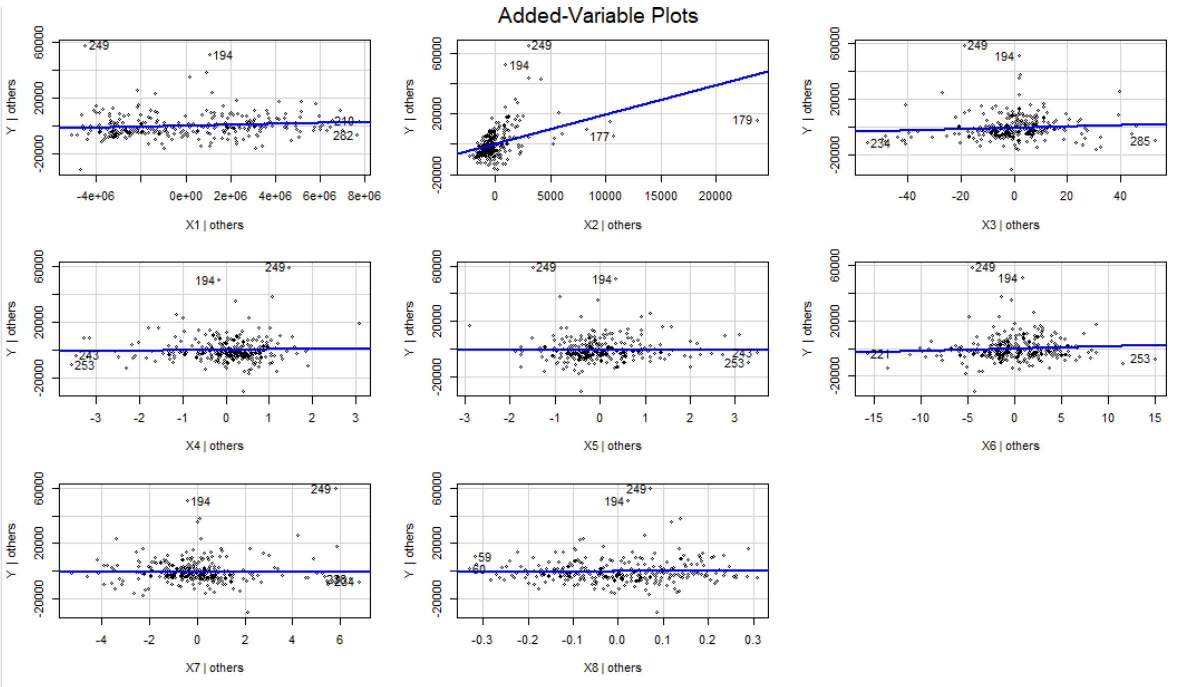
- The response variable Y has a distinct left skew.
- The predictor variable X1, has a uniform skew.
- The predictor variable X2, has a strong left skew.
- The predictor variable X3, has a left skew.
- The predictor variable X4, has a normal distribution.
- The predictor variable X5, has a symmetric distribution.
- The predictor variable X6, has a strong right skew.
- The predictor variable X7, has a left skew.
- The predictor variable X8 has a normal distribution.



Above are the box plots of all the numeric variables (Y and X1 through X8) provided in the data set streamflow.

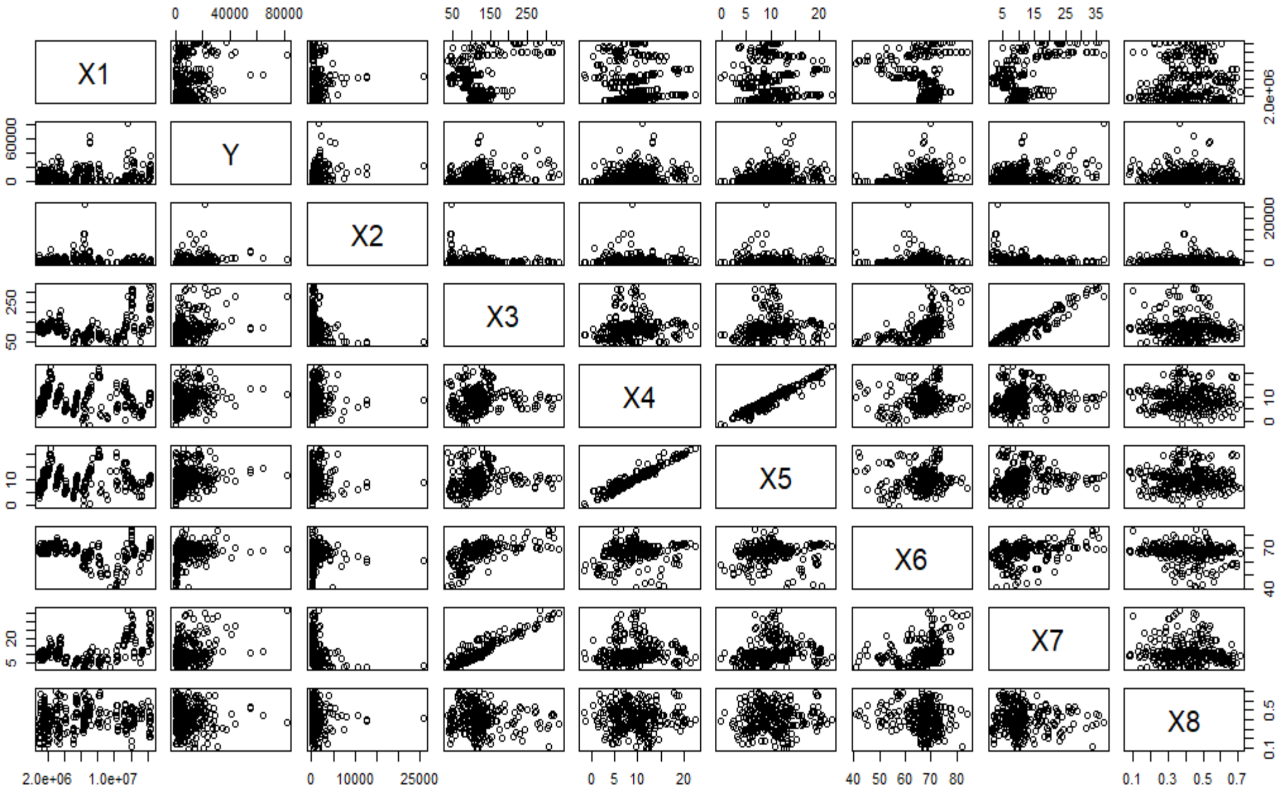


Above are the scatter plots of all the numeric variables (Y and X1 through X8) provided in the data set streamflow.



Above are the added-variable plots of all the numeric variables (Y and X1 through X8) provided in the data set streamflow.

```
> cor(streamflow)
            X1          Y          X2          X3          X4          X5          X6
X1  1.00000000  0.1926035  0.02097545  0.30978548 -0.15557066 -0.04759166 -0.21430059
Y   0.19260350  1.0000000  0.31807850  0.29602977  0.20550773  0.21113865  0.21802869
X2  0.02097545  0.3180785  1.00000000 -0.24704239 -0.03020605 -0.04769574 -0.08566400
X3  0.30978548  0.2960298 -0.24704239  1.00000000  0.07752031  0.10881444  0.55728901
X4 -0.15557066  0.2055077 -0.03020605  0.07752031  1.00000000  0.96818515  0.19074913
X5 -0.04759166  0.2111386 -0.04769574  0.10881444  0.96818515  1.00000000  0.09235669
X6 -0.21430059  0.2180287 -0.08566400  0.55728901  0.19074913  0.09235669  1.00000000
X7  0.48388068  0.2829461 -0.25355037  0.92688029  0.08247133  0.14754361  0.35554378
X8  0.11320656 -0.0116222 -0.02146808 -0.11035338 -0.01089296  0.02374341 -0.16804586
            X7          X8
X1  0.48388068  0.11320656
Y   0.28294614 -0.01162220
X2 -0.25355037 -0.02146808
X3  0.92688029 -0.11035338
X4  0.08247133 -0.01089296
X5  0.14754361  0.02374341
X6  0.35554378 -0.16804586
X7  1.00000000 -0.10848844
X8 -0.10848844  1.00000000
```



From the above correlation matrix and the plot, we can observe that there is not extreme multicollinearity problem between the variables.

2. MODELS AND METHODS:

Now we can fit our preliminary model. Our preliminary model will simply be our response variable Y (max90) regressed against all the predictor variables in our data set.

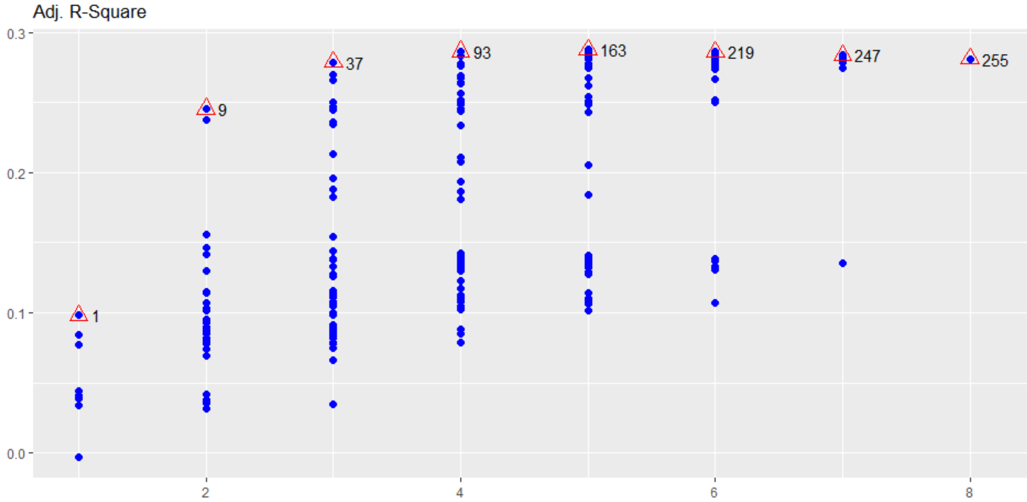
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8, data = streamflow)

Residuals:
   Min     1Q Median     3Q    Max 
-29981  -4579  -1538   2868  59389 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.765e+04  8.178e+03  -2.159   0.0317 *  
X1           3.323e-04  1.745e-04   1.904   0.0579 .  
X2           1.942e+00  2.532e-01   7.671 2.75e-13 ***
X3           4.969e+01  3.527e+01   1.409   0.1600    
X4           3.440e+02  5.819e+02   0.591   0.5549    
X5           1.287e+02  6.068e+02   0.212   0.8322    
X6           1.603e+02  1.305e+02   1.228   0.2203    
X7           5.115e+01  2.751e+02   0.186   0.8526    
X8           2.405e+03  4.039e+03   0.595   0.5521    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8999 on 284 degrees of freedom
Multiple R-squared:  0.3009,	Adjusted R-squared:  0.2812 
F-statistic: 15.28 on 8 and 284 DF,  p-value: < 2.2e-16
```

```
> anova(fitstream)
Analysis of Variance Table

Response: Y
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
X1          1 1.2203e+09 1220348895 15.0689  0.000129 ***
X2          1 3.2457e+09 3245732542 40.0783 9.495e-10 ***
X3          1 3.9125e+09 3912466049 48.3112 2.494e-11 ***
X4          1 1.3622e+09 1362215144 16.8206 5.370e-05 ***
X5          1 6.7370e+06    6736980  0.0832  0.773233    
X6          1 1.2057e+08  120570374  1.4888  0.223414    
X7          1 5.1717e+05     517174  0.0064  0.936363    
X8          1 2.8703e+07   28703189  0.3544  0.552092    
Residuals 284 2.3000e+10   80984724                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that the model is significant, but some of the individual predictors are not significant. Running a best subset and stepwise regression on this full model results in the following. First, we will look at the best subset for each number of predictor variables selected based on the highest R2 adj.

R2 Adjusted:

To get a better idea of where the R2 adj peaks we can look at a plot of the R2 adj against the number of predictors.

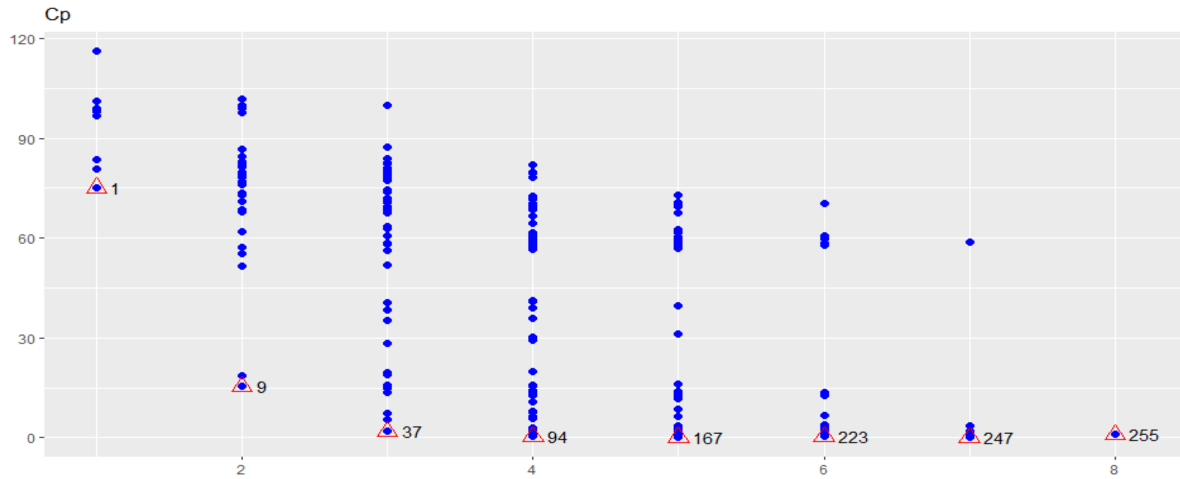



```
    n             predictors      adjr
93  4            x1 x2 x3 x4 0.2863228
163 5         x1 x2 x3 x4 x6 0.2875855
219 6      x1 x2 x3 x4 x6 x8 0.2859723
247 7   x1 x2 x3 x4 x5 x6 x8 0.2835982
255 8 x1 x2 x3 x4 x5 x6 x7 x8 0.2811632
> |
```

Now, lets perform the same by using CP, AIC, BIC.

CP:

Then plotting the CP against the number of predictors.

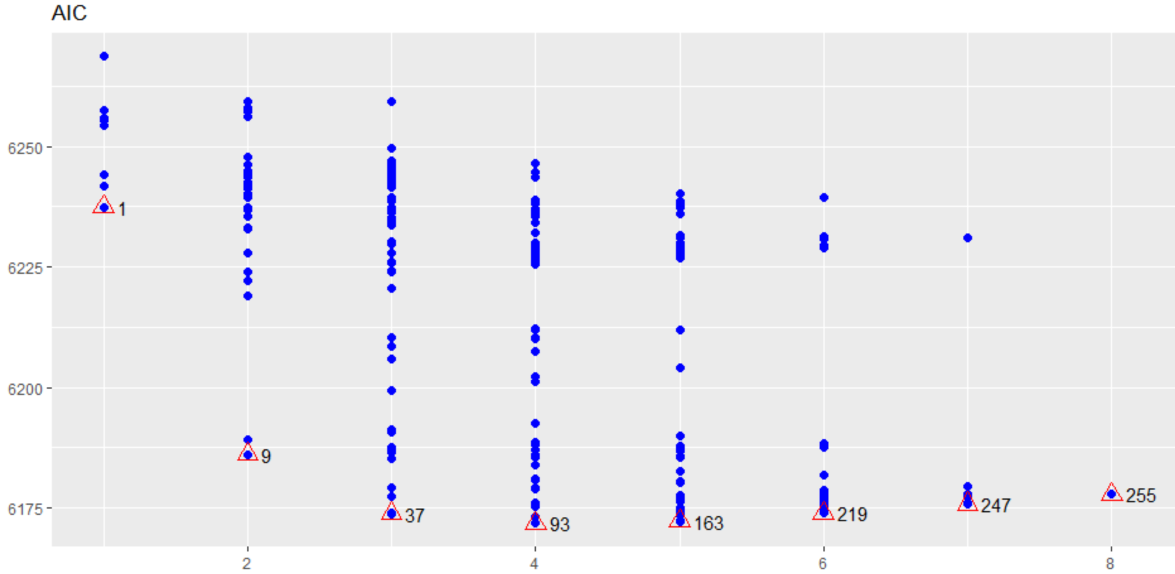


Then getting our best subset.

```
    n              predictors       cp
93  4             X1 X2 X3 X4 2.932805
163 5          X1 X2 X3 X4 X6 3.435856
219 6       X1 X2 X3 X4 X6 X8 5.086642
247 7    X1 X2 X3 X4 X5 X6 X8 7.034584
255 8 X1 X2 X3 X4 X5 X6 X7 X8 9.000000
```

AIC:

Plotting AIC against number of predictors.

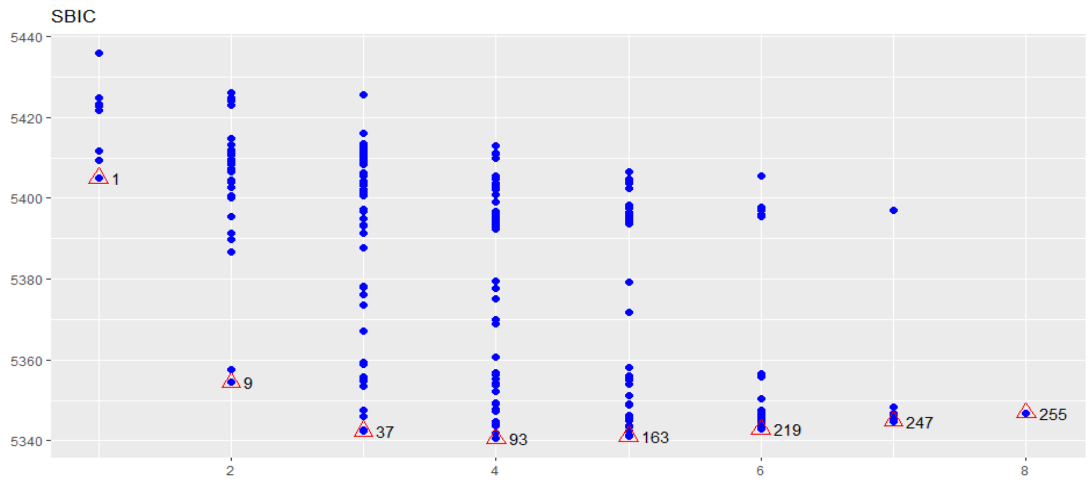


Then getting our best subset.

```
    n              predictors      aic
93  4             X1 X2 X3 X4 6171.807
163 5          X1 X2 X3 X4 X6 6172.269
219 6       X1 X2 X3 X4 X6 X8 6173.909
247 7    X1 X2 X3 X4 X5 X6 X8 6175.855
255 8 X1 X2 X3 X4 X5 X6 X7 X8 6177.820
>
```

BIC:

Plotting BIC against number of predictors.



Then getting our best subset.

```
    n              predictors      bic
93  4             X1 X2 X3 X4 5340.555
163 5          X1 X2 X3 X4 X6 5341.131
219 6       X1 X2 X3 X4 X6 X8 5342.848
247 7    X1 X2 X3 X4 X5 X6 X8 5344.861
255 8 X1 X2 X3 X4 X5 X6 X7 X8 5346.890
>
```

And finally, a stepwise regression

We can see that all our procedures agree on a model. The best subset model based on R2 adj, CP, AIC, and BIC contain predictors X2, X3 and X5.

```
We are selecting variables based on p value...

Stepwise Selection: Step 1

+ X2

                              Model Summary
-----------------------------------------------------------------------
R                        0.318       RMSE                 10080.200
R-Squared                0.101       Coef. Var              108.708
Adj. R-Squared           0.098       MSE                101610439.565
Pred R-Squared           0.048       MAE                   6897.294
-----------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                    ANOVA
---------------------------------------------------------------------------------
                    Sum of
                   Squares         DF       Mean Square       F          Sig.
---------------------------------------------------------------------------------
Regression     3328314064.255        1    3328314064.255    32.756     0.0000
Residual      29568637913.443      291     101610439.565
Total         32896951977.699      292
---------------------------------------------------------------------------------

                                  Parameter Estimates
-----------------------------------------------------------------------------------------------
      model        Beta     Std. Error     Std. Beta        t        Sig        lower      upper
-----------------------------------------------------------------------------------------------
(Intercept)    7581.312       658.885                    11.506    0.000     6284.527   8878.097
         X2       1.534         0.268         0.318       5.723    0.000        1.006      2.061
-----------------------------------------------------------------------------------------------



Stepwise Selection: Step 2

+ X3

                              Model Summary
-----------------------------------------------------------------------
R                        0.501       RMSE                  9219.937
R-Squared                0.251       Coef. Var               99.431
Adj. R-Squared           0.245       MSE                 85007236.614
Pred R-Squared           0.161       MAE                   6158.189
-----------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                    ANOVA
---------------------------------------------------------------------------------
                    Sum of
                   Squares         DF       Mean Square       F          Sig.
---------------------------------------------------------------------------------
Regression     8244853359.704        2    4122426679.852    48.495     0.0000
Residual      24652098617.994      290      85007236.614
Total         32896951977.699      292
---------------------------------------------------------------------------------

                                  Parameter Estimates
-----------------------------------------------------------------------------------------------
      model        Beta     Std. Error     Std. Beta        t        Sig        lower      upper
-----------------------------------------------------------------------------------------------
(Intercept)   -2286.243      1430.630                    -1.598    0.111    -5101.977    529.492
         X2       2.009         0.253         0.417       7.942    0.000        1.511      2.507
         X3      77.753        10.224         0.399       7.605    0.000       57.631     97.876
-----------------------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 3

+ X5

                          Model Summary
-------------------------------------------------------------------
R                      0.535       RMSE                   9013.625
R-Squared              0.286       Coef. Var                97.206
Adj. R-Squared         0.279       MSE                81245443.753
Pred R-Squared         0.193       MAE                    5926.228
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                    ANOVA
------------------------------------------------------------------------------
                    Sum of
                   Squares         DF       Mean Square      F         Sig.
------------------------------------------------------------------------------
Regression    9417018733.194         3    3139006244.398    38.636    0.0000
Residual     23479933244.504       289      81245443.753
Total        32896951977.699       292
------------------------------------------------------------------------------

                                     Parameter Estimates
-----------------------------------------------------------------------------------------------
      model          Beta     Std. Error     Std. Beta       t         Sig         lower        upper
-----------------------------------------------------------------------------------------------
(Intercept)     -6708.252      1819.746                  -3.686     0.000    -10289.887    -3126.617
         X2         2.029         0.247         0.421     8.204     0.000         1.543        2.516
         X3        73.928        10.046         0.379     7.359     0.000        54.156       93.700
         X5       470.125       123.771         0.190     3.798     0.000       226.518      713.731
-----------------------------------------------------------------------------------------------
```

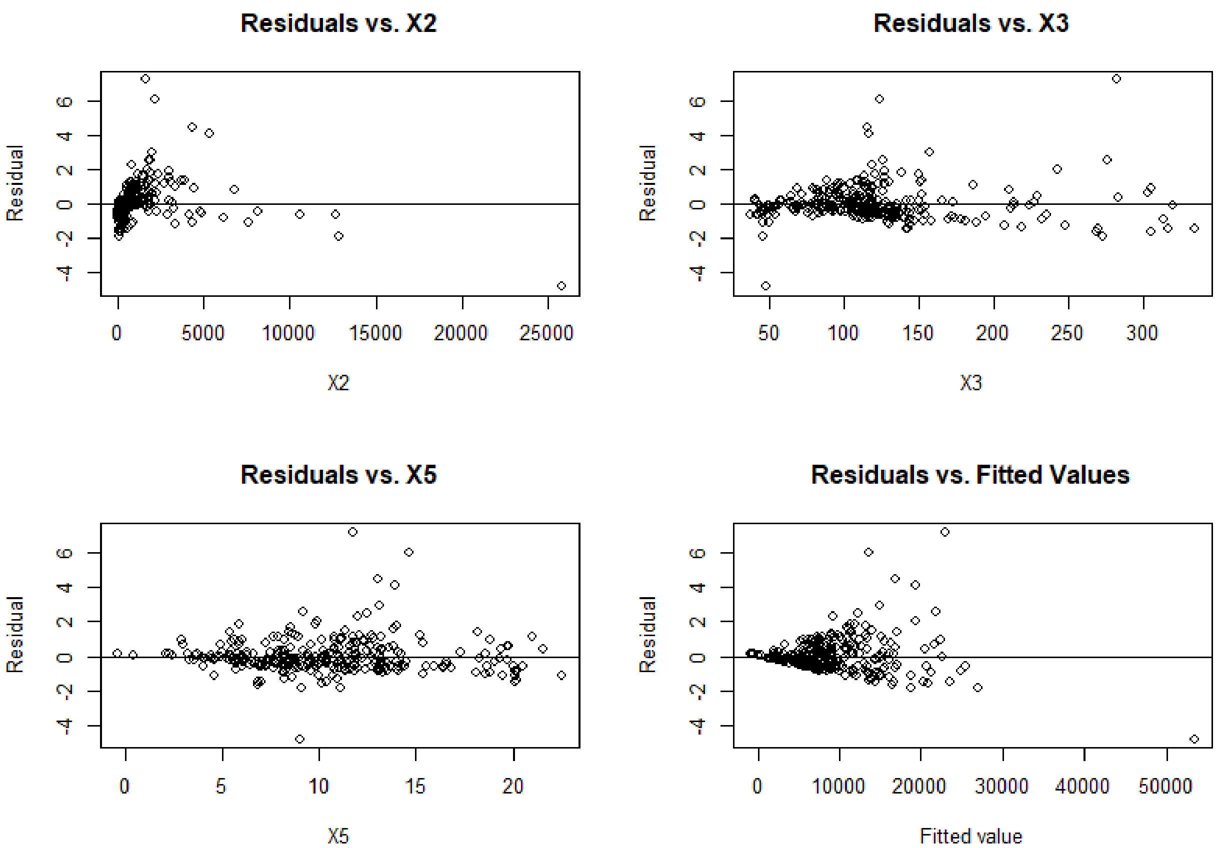
Since, we do not have any multicollinearity problem with the variables, we can proceed with the final model.

```
No more variables to be added/removed.


Final Model Output
------------------

                          Model Summary
-------------------------------------------------------------------
R                      0.535       RMSE                   9013.625
R-Squared              0.286       Coef. Var                97.206
Adj. R-Squared         0.279       MSE                81245443.753
Pred R-Squared         0.193       MAE                    5926.228
-------------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                                    ANOVA
------------------------------------------------------------------------------
                    Sum of
                   Squares         DF       Mean Square      F         Sig.
------------------------------------------------------------------------------
Regression    9417018733.194         3    3139006244.398    38.636    0.0000
Residual     23479933244.504       289      81245443.753
Total        32896951977.699       292
------------------------------------------------------------------------------

                                     Parameter Estimates
-----------------------------------------------------------------------------------------------
      model          Beta     Std. Error     Std. Beta       t         Sig         lower        upper
-----------------------------------------------------------------------------------------------
(Intercept)     -6708.252      1819.746                  -3.686     0.000    -10289.887    -3126.617
         X2         2.029         0.247         0.421     8.204     0.000         1.543        2.516
         X3        73.928        10.046         0.379     7.359     0.000        54.156       93.700
         X5       470.125       123.771         0.190     3.798     0.000       226.518      713.731
-----------------------------------------------------------------------------------------------
```

```
Call:
lm(formula = Y ~ X2 + X3 + X5, data = streamflow)

Residuals:
   Min     1Q Median     3Q    Max 
-31394  -4965  -1404   2619  59010 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6708.2520  1819.7455  -3.686 0.000272 ***
X2              2.0295     0.2474   8.204 7.73e-15 ***
X3             73.9282    10.0458   7.359 1.93e-12 ***
X5            470.1248   123.7708   3.798 0.000178 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9014 on 289 degrees of freedom
Multiple R-squared:  0.2863,	Adjusted R-squared:  0.2788 
F-statistic: 38.64 on 3 and 289 DF,  p-value: < 2.2e-16
```

There is not much multicollinearity in the final model as we can observe.

Now, lets perform regression diagnostics and get the residual plots.



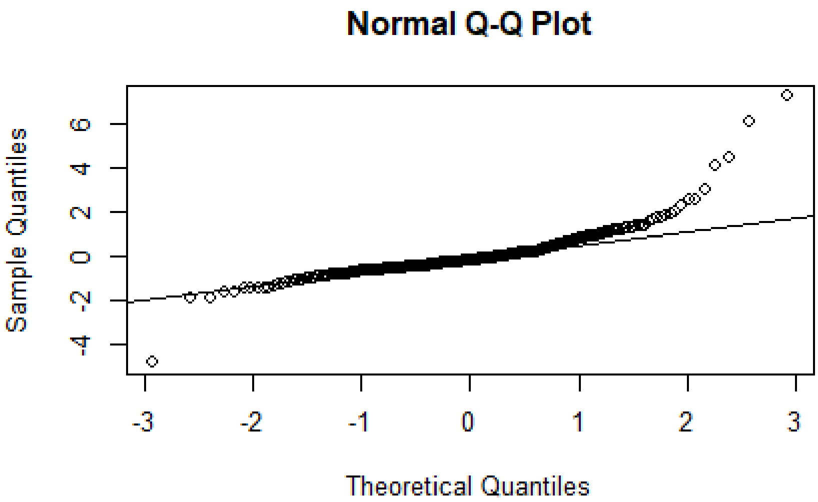
By using the vif built-in function we can confirm that there is not multicollinearity in the reduced model (X2+X3+X5) as well

```
      X2       X3       X5
1.065494 1.075808 1.012455
```

To confirm what we see in the above plots we can run a Breusch-Pagan test and Durbin-Watson test, where the p value in both cases is not greater than 0.05 so, we cannot retain our null assumption of independence.

```
> bptest(reduced.lmfit)

        studentized Breusch-Pagan test

data:  reduced.lmfit
BP = 39.762, df = 3, p-value = 1.197e-08


        Durbin-Watson test

data:  fitstream
DW = 1.3968, p-value = 4.424e-08
alternative hypothesis: true autocorrelation is not 0
```

To assess normality we will start by looking a at Q-Q Plot of our residuals.



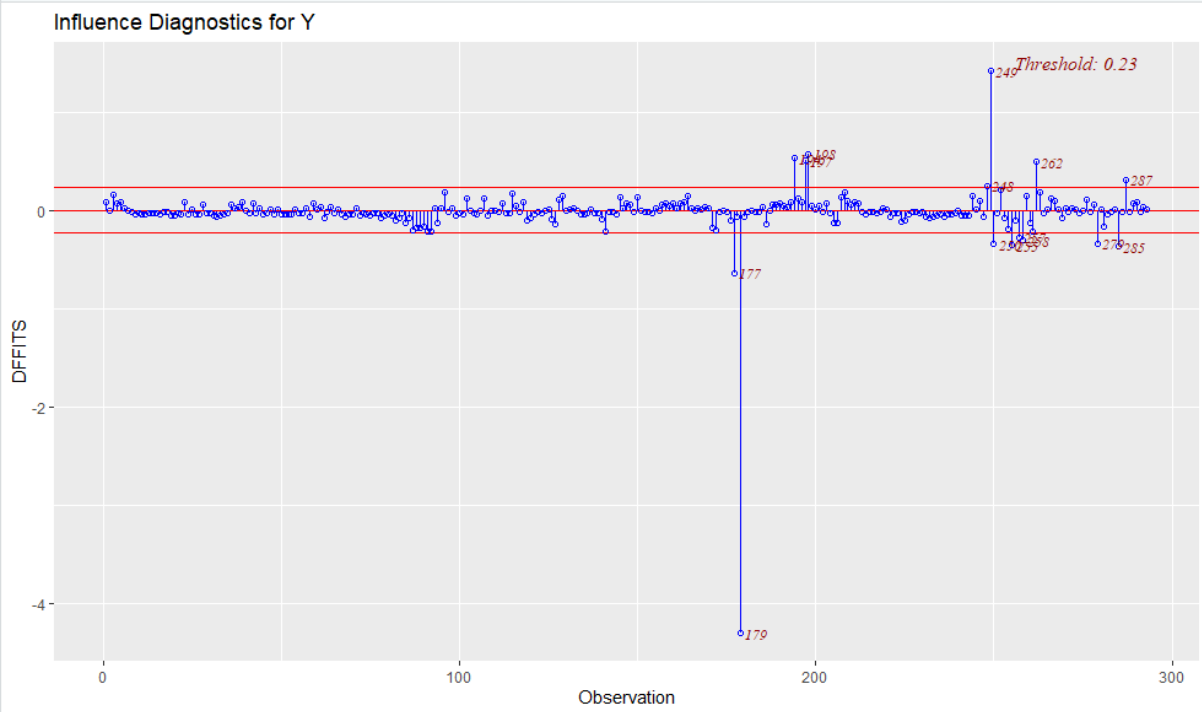
We can clearly see that there is a right skew primarily on the right side of the distribution. By using a Shapiro-Wilk test we can test for the likelihood of this distribution under the assumption of normality.

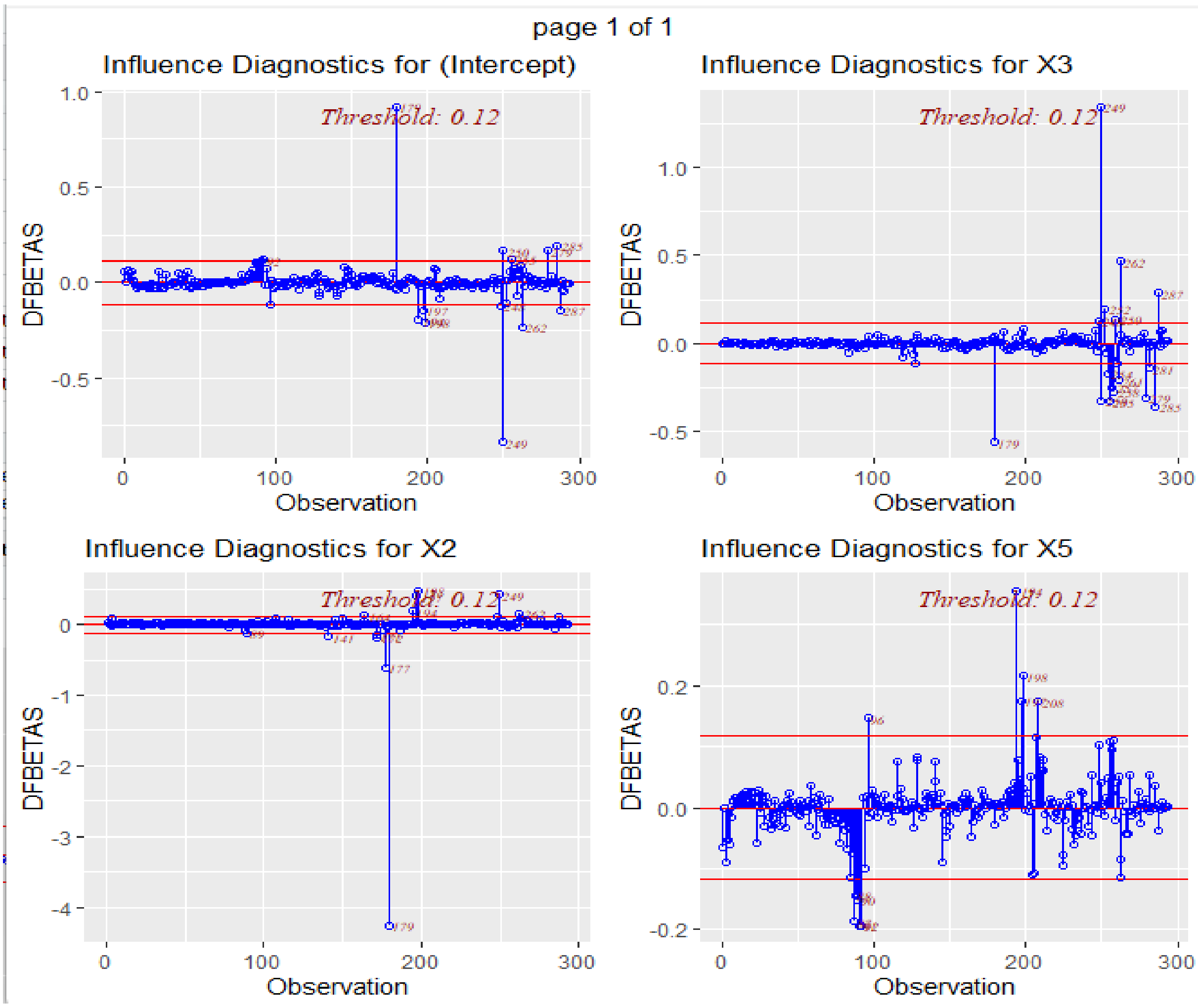
```
        Shapiro-Wilk normality test

data:  res
W = 0.80758, p-value < 2.2e-16
```

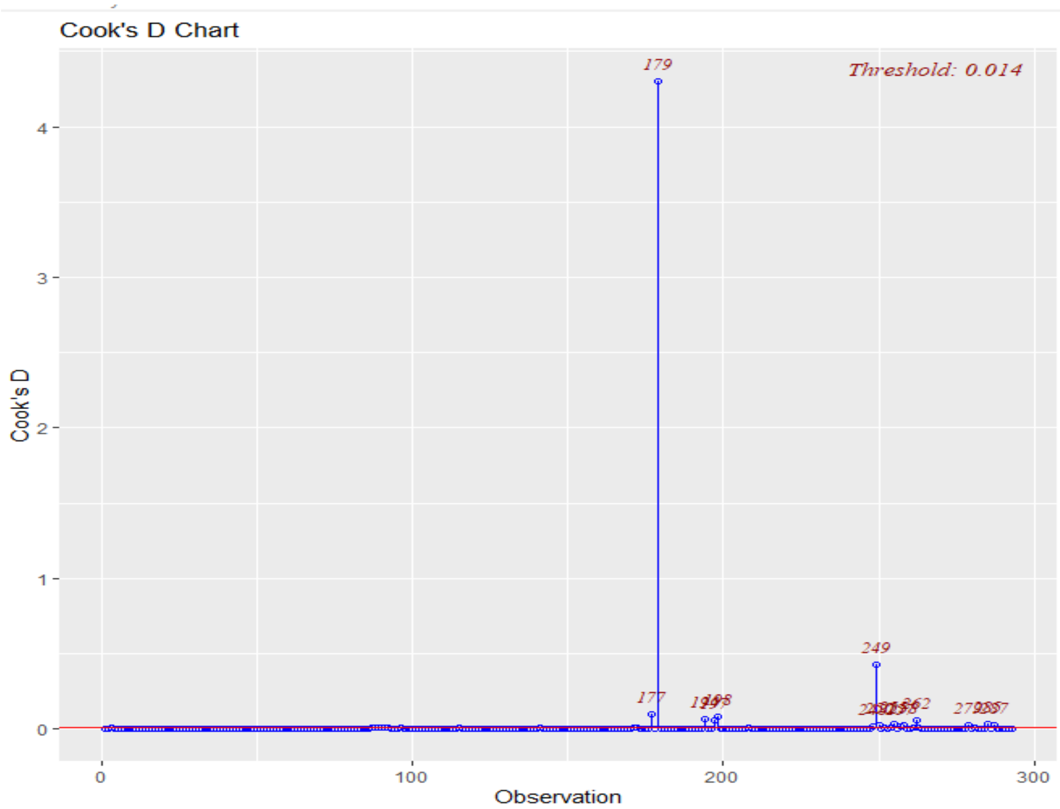
The p-value of 0.00000000000000022 means that this distribution has a 0.0000000000000022% chance of occurring if the population is normally distributed.

There is clearly a problem with our assumption of normality.

Finally, we will look at outliers using DFBETAS, DFFITS, and Cook's Distance plots.







Within our outliers we have observations 179 and 249 which are also two of the points that heavily influence the right skew previously seen in our Q-Q plot.

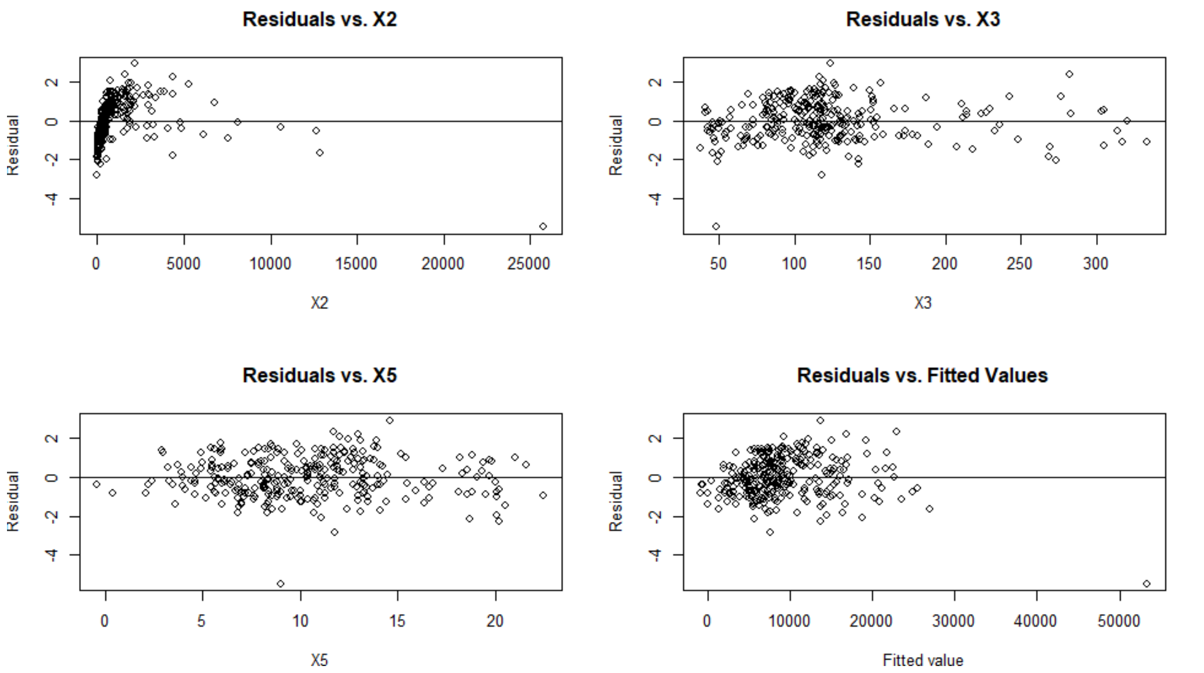
We have one assumption to remedy with our model, which is our assumption of normality. We will attempt 23 to resolve this by transforming our model with a **Box-Cox Transformation.**

```
> lambda
[1] 0.2792849
```

We get that our optimized model has $\lambda = 0.2792849$. We then raise our response variable Y to λ and fit our model with our transformed Y.

```
Call:
lm(formula = trans.Y ~ X2 + X3 + X5, data = streamflow)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0998  -2.2829  -0.1921   2.1780   8.9177 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.043e+00  6.227e-01   8.098 1.58e-14 ***
X2          7.833e-04  8.466e-05   9.253  < 2e-16 ***
X3          2.766e-02  3.438e-03   8.046 2.24e-14 ***
X5          2.058e-01  4.236e-02   4.860 1.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.085 on 289 degrees of freedom
Multiple R-squared:  0.3414,	Adjusted R-squared:  0.3346 
F-statistic: 49.94 on 3 and 289 DF,  p-value: < 2.2e-16
```

The R2 adj went up, but if our model satisfies all necessary assumptions it is a better model. Let's take a look at assumption diagnostics again, starting with constancy of variance.

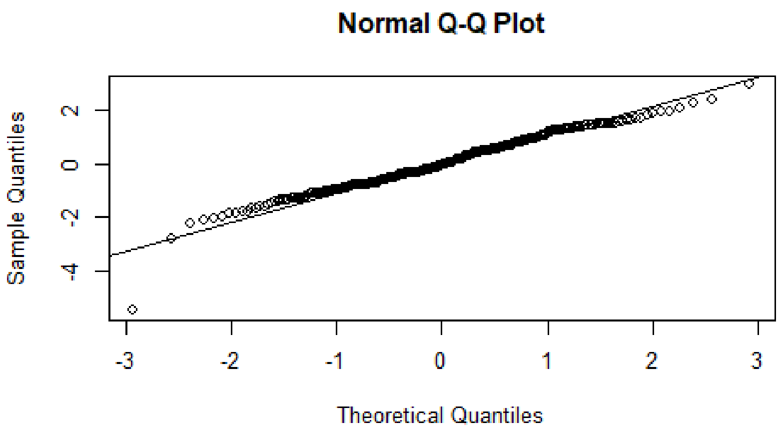



```
        studentized Breusch-Pagan test

data:  boxcox.lmfit
BP = 67.678, df = 3, p-value = 1.341e-14
```

Our graphs show fairly even spread. But a quick Breusch-Pagan test shows that our assumption of constancy of variance error is reasonable.

```
        Durbin-Watson test

data:  boxcox.lmfit
DW = 1.3658, p-value = 2.635e-08
alternative hypothesis: true autocorrelation is not 0
```

Even spread of residuals against an index and a high p-value in a Durbin-Watson test shows that our assumption of independence is reasonable.



```
        Shapiro-Wilk normality test

data:  boxcox.res
W = 0.97508, p-value = 5.551e-05
```

Our assumption of normality, which was violated in our previous model, is satisfied here. The p-value from the Shapiro-Wilk test is now much higher, and our Q-Q plot shows that the assumption of normality is reasonable.

3. RESULTS:

The summary of the final optimal model is:

Y = 5.043e+00 + 7.833e-04X2 + 2.766e-02X3 + 2.058e-01X5

The predictor variables of the model are DRAIN_SQKM, PPTAVG_BASIN, T_AVG_SITE.

The value of the **F-statistic** is 49.94 on six predictor variables, and the **p-value** is 2.2e-16; Therefore, there is an overall significant relationship between the response variable and the predictor variables.

```
Call:
lm(formula = trans.Y ~ X2 + X3 + X5, data = streamflow)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.0998  -2.2829  -0.1921   2.1780   8.9177 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.043e+00  6.227e-01   8.098 1.58e-14 ***
X2          7.833e-04  8.466e-05   9.253  < 2e-16 ***
X3          2.766e-02  3.438e-03   8.046 2.24e-14 ***
X5          2.058e-01  4.236e-02   4.860 1.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.085 on 289 degrees of freedom
Multiple R-squared:  0.3414,	Adjusted R-squared:  0.3346 
F-statistic: 49.94 on 3 and 289 DF,  p-value: < 2.2e-16

> anova(boxcox.lmfit)
Analysis of Variance Table

Response: trans.Y
           Df  Sum Sq Mean Sq F value    Pr(>F)    
X2          1  501.01  501.01  52.655 3.678e-12 ***
X3          1  699.79  699.79  73.547 6.074e-16 ***
X5          1  224.71  224.71  23.616 1.932e-06 ***
Residuals 289 2749.81    9.51                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the summary statistics and ANOVA table, the results of each p-value are less than the significant level 0.05.

4. CONCLUSION:

To find the optimal model for the STREAMFLOW dataset, we performed exploratory data analysis and reduced the model based on the significant level of 0.05. To validate it, we chose the method for model selection and applied diagnostic measures to ensure the model is fitted perfectly. Later, to improvise it further, we transformed the model, and there is a significant improvement in the Adjusted R-squared of the final model. We chose the transformed model, which has the better improvement.

5. APPENDIX:

```
#Data Summarization
library(readr)
streamflow <- read_csv("streamflow.csv")
View(streamflow)
colnames(streamflow) <- c("X1","Y","X2","X3","X4","X5","X6","X7","X8")
summary(streamflow)

## HISTOGRAMS
par(mfrow=c(3,3))
hist(streamflow$X1,main="Histogram of X1",xlab="Staid")
hist(streamflow$Y,main="Histogram of Y",xlab="90th percentile of the time series")
hist(streamflow$X2,main="Histogram of X2",xlab="the drainage area ")
hist(streamflow$X3,main="Histogram of X3",xlab="average basin precipitation ")
hist(streamflow$X4,main="Histogram of X4",xlab="average basin temperature  ")
hist(streamflow$X5,main="Histogram of X5",xlab="average temperature at the stream
location ")
hist(streamflow$X6,main="Histogram of X6",xlab="average relative humidity across
the basin")
hist(streamflow$X7,main="Histogram of X7",xlab="average March precipitation")
hist(streamflow$X8,main="Histogram of X8",xlab="median relief ratio")

## BOX PLOTS
boxplot(streamflow$Y, xlab="Max90", main="Max90")
boxplot(streamflow$X1, xlab="STAID", main="STAID")
boxplot(streamflow$X2, xlab="DRAIN_SQKM", main=" The drainage area")
```

boxplot(streamflow$X3, xlab="PPTAVG_BASIN", main="The average basinprecipitation")

boxplot(streamflow$X4, xlab="T_AVG_BASIN", main="Boxplot of the average basintemperature")

boxplot(streamflow$X5, xlab="T_AVG_SITE", main="Boxplot of the averagetemperature at the stream location")

boxplot(streamflow$X6, xlab="RH_BASIN", main="Boxplot of the average relativehumidity across the basin ")

boxplot(streamflow$X7, xlab="MAR_PPT7100_CM", main="Boxplot of the averageMarch precipitation ")

boxplot(streamflow$X8, xlab="RRMEDIAN", main="Boxplot of the median reliefratio")

SCATTER PLOTS

plot(streamflow$Y, xlab="Max90",ylab="Max90", main="Boxplot of Stream Flow")

plot(streamflow$X1, xlab="STAID",ylab="STAID", main="STAID")

plot(streamflow$X2, xlab="DRAIN_SQKM", ylab="DRAIN_SQKM", main=" The drainage area")

plot(streamflow$X3, xlab="PPTAVG_BASIN",ylab="PPTAVG_BASIN", main="The average basinprecipitation")

plot(streamflow$X4, xlab="T_AVG_BASIN",ylab="T_AVG_BASIN", main="Boxplot of the average basintemperature")

plot(streamflow$X5, xlab="T_AVG_SITE", ylab="T_AVG_SITE", main="Boxplot of the averagetemperature at the stream location")

plot(streamflow$X6, xlab="RH_BASIN",ylab="RH_BASIN", main="Boxplot of the average relativehumidity across the basin ")

plot(streamflow$X7, xlab="MAR_PPT7100_CM",ylab="MAR_PPT7100_CM", main="Boxplot of the averageMarch precipitation ")

plot(streamflow$X8, xlab="RRMEDIAN",ylab="RRMEDIAN", main="Boxplot of the median reliefratio")

ADDED VARIABLE PLOTS

library(car)

avPlots(fitstream)

CORRELATION MATRIX

cor(streamflow)

```
pairs(streamflow)
##checking for multicollinearity.
eigen(cor(streamflow))$values

## MODELS AND METHODS

fitstream<-lm(Y ~X1+X2+X3+X4+X5+X6+X7+X8,data=streamflow)
fitstream
summary(fitstream)

##ANOVA t-test
anova(fitstream)

library(MASS)
## Model Selection
library(olsrr)

#### Print all possible regression models in terms of adjr, Cp, AIC, and BIC.

par(mfrow=c(1,1))
b <- ols_step_all_possible(fitstream)
plot(b)
#### Adjusted R2 ####

b.adjr = data.frame(n=b$n,predictors=b$predictors,adjr=b$adjr)
print(b.adjr)
print(b.adjr[c(93,163,219,247,255),])

#### Cp ####

b.cp = data.frame(n=b$n,predictors=b$predictors,cp=b$cp)
print(b.cp)
```

```
print(b.cp[c(93,163,219,247,255),])

#### AIC ####

b.aic = data.frame(n=b$n,predictors=b$predictors,aic=b$aic)
print(b.aic)
print(b.aic[c(93,163,219,247,255),])

#### BIC ####

b.bic = data.frame(n=b$n,predictors=b$predictors,bic=b$sbic)
print(b.bic)
print(b.bic[c(93,163,219,247,255),])

#### PRESS ####

b.press = data.frame(n=b$n,predictors=b$predictors,press=b$msep)
print(b.press)
print(b.press[c(93,163,219,247,255),])

#### Stepwise Regression ####

k <- ols_step_both_p(fitstream,pent=0.10,prem=0.1,details=TRUE)
plot(k)

#### Final Model? ####

reduced.lmfit <- lm(Y ~ X2 + X3+X5, data=streamflow)
summary(reduced.lmfit)

######## Regression Diagnostics ############
```

```
res <- rstudent(reduced.lmfit)
fitted.y <- fitted(reduced.lmfit)

######## Residual Plots ##########

par(mfrow=c(2,2))

plot(res ~ streamflow$X2, xlab="X2", ylab="Residual", main="Residuals vs. X2")
abline(h=0)
plot(res ~ streamflow$X3, xlab="X3", ylab="Residual", main="Residuals vs. X3")
abline(h=0)
plot(res ~ streamflow$X5, xlab="X5", ylab="Residual", main="Residuals vs. X5")
abline(h=0)

plot(res ~ fitted.y, xlab="Fitted value", ylab="Residual", main="Residuals vs. Fitted Values")
abline(h=0)

######### Multicollinearity ##########

vif(reduced.lmfit)



######### Constancy of Error Variances #########
library(lmtest)
bptest(reduced.lmfit)



#Durbin-Watson
#install lmtest
library(lmtest)
dwtest(fitstream, alternative="two.sided")
```

```
######### Normality ###########

qqnorm(res);qqline(res)
########Shapiro test########
shapiro.test(res)

#DFFITS values
library(olsrr)

ols_plot_dffits(reduced.lmfit)

#DFBETAS values
ols_plot_dfbetas(reduced.lmfit)

#Cook's distance values
ols_plot_cooksd_chart(reduced.lmfit)

######### Transformation #########
library(EnvStats)

boxcox.summary <- boxcox(reduced.lmfit, optimize=TRUE)
lambda <- boxcox.summary$lambda
lambda
trans.Y <- streamflow$Y^lambda

streamflow <- cbind(streamflow,trans.Y)
streamflow

######### Re-fitting a model using the transformed response variable. ##########

boxcox.lmfit <- lm(trans.Y ~ X2 + X3 + X5, data=streamflow)
```

```
summary(boxcox.lmfit)
anova(boxcox.lmfit)

boxcox.res <- rstudent(boxcox.lmfit)

boxcox.fitted.y <- fitted(boxcox.lmfit)
######## Residual Plots ##########
par(mfrow=c(2,2))

plot(boxcox.res ~ streamflow$X2, xlab="X2", ylab="Residual", main="Residuals vs. X2")
abline(h=0)
plot(boxcox.res ~ streamflow$X3, xlab="X3", ylab="Residual", main="Residuals vs. X3")
abline(h=0)
plot(boxcox.res ~ streamflow$X5, xlab="X5", ylab="Residual", main="Residuals vs. X5")
abline(h=0)

plot(boxcox.res ~ fitted.y, xlab="Fitted value", ylab="Residual", main="Residuals vs. Fitted Values")
abline(h=0)

######### Multicollinearity ##########
library(HH)
vif(boxcox.lmfit)

######### Constancy of Error Variances #########

bptest(boxcox.lmfit)

dwtest(boxcox.lmfit, alternative="two.sided")
```

```
######### Normality ###########

qqnorm(boxcox.res);qqline(boxcox.res)
shapiro.test(boxcox.res)

######### Final Model ##########

final.lmfit <- boxcox.lmfit
summary(final.lmfit)

##Obtain DFFITS, DFBETAS, and Cook’s distance values
library(olsrr)

#DFFITS values
ols_plot_dffits(final.lmfit)

#DFBETAS values
ols_plot_dfbetas(final.lmfit)

#Cook's distance values
ols_plot_cooksd_chart(final.lmfit)

################ Fit a regression model with interaction terms ##################

streamflow.lmfit <- lm(Y ~ X2 + X3 + X5 + X2*X3 + X2*X5 + X3*X5, data=streamflow)
summary(streamflow.lmfit)
anova(streamflow.lmfit)

################ Fit a regression model with no interaction terms ###############

streamflow.reduced <- lm(Y ~ X2 + X3 + X5, data=streamflow)
```

################ Test for significance of the interaction terms #################

anova(streamflow.reduced,streamflow.lmfit)