Table of Contents:

S.No.	Title	Page No.
1	Introduction	3
	1.1 Brief overview of the regression analysis project	3
	1.2 Objectives and goals	3
2	Data Exploration and Visualization	4
	2.1 Data Import and Overview	4
	2.2 Variable Descriptions	5
	2.3 Exploratory Data Analysis (EDA)	5
3	Model Fitting	10
	3.1 Model Selection	10
	3.2 Model Development	10
	3.3 Model Refinement	11
	3.4 Stepwise Model Selection	12
4	Model Evaluation	17
	4.1 Final Model Assessment	17
	4.2 Regression Diagnosis & Assumption Checking	17
	4.3 Outliers Detection	19
	4.4 Data Transformation	21
5	Conclusion	25
	5.1 Summary of Findings	25
	5.2 Model Performance	26
	5.3 Implications of the Study	26
	5.4 Final Thought	26
6	Appendices	27
	R Code	27

1. Introduction

1.1 Brief overview of the regression analysis project

In this project, we delved into an extensive dataset of streamflow records that includes a variety of hydrological and meteorological variables. The data encompasses streamflow metrics such as the 90th percentile of streamflow ('max90'), watershed area ('DRAIN_SQKM'), and other influential environmental factors like precipitation and temperature across the basin. Understanding the dynamics of river systems is a cornerstone of hydrological science, with streamflow data being a key indicator of watershed health, water availability, and flood risk management.

Our dataset is a compilation of measurements from various stations, with each entry detailing the maximum streamflow recorded and several contributing factors that could potentially affect these readings. These variables include basin-wide averages of precipitation ('PPTAVG_BASIN'), temperature ('T_AVG_BASIN' and 'T_AVG_SITE'), relative humidity ('RH_BASIN'), as well as more specific measurements such as the average March precipitation ('MAR_PPT7100_CM') and median relief ratio

('RRMEDIAN'). This comprehensive suite of variables allows us to construct a detailed picture of the factors influencing streamflow.

1.2 Objectives and goals

The primary objective of our regression analysis project is to model the relationship between streamflow and its influencing factors, providing insights that could support water management authorities in decision-making processes. By leveraging regression techniques, we aim to quantify how each predictor variable impacts streamflow, identify significant predictors, and assess the predictive power of our model. This can inform future policies for flood mitigation, water resource allocation, and environmental conservation.

Through our analysis, we also seek to contribute to the broader scientific understanding of hydrological processes, enhancing the predictability of streamflow patterns in the context of a changing climate. The outcomes of this project have the potential to influence both local watershed management practices and global environmental research initiatives.

2. Data Exploration and Visualization

2.1 Data Import and Overview

To handle our dataset effectively, we begin by loading the necessary libraries in R. **'readr'** is used for reading in the CSV file, providing fast and friendly data importing capabilities. Additional libraries may include '**car'** ,' **MASS' 'olsrr', 'lmtest', 'EnvStats'** for data manipulation purposes, '**ggplot2'** for data visualization, and other libraries is used in various stages of the project for modeling and analysis.

Next, we import the 'streamflow.csv' file into R using the **'read_csv'** function.

	Y	X1	X2	X3	X4	X5	X6	X7	X8
1	12440.00	1013500	2252.696000	97.41780	3.004670	3.0	71.67319	6.317267	0.21476510
2	6343.00	1022500	573.600600	120.07020	5.945692	6.3	68.82603	10.675010	0.16203704
3	23680.00	1030500	3676.172000	108.19060	4.815170	5.4	69.60340	8.694030	0.13859911
4	12200.00	1031500	769.048200	118.00080	4.143458	4.9	68.47412	9.538659	0.28487518
5	14730.00	1047000	909.097200	118.86150	3.990672	5.6	68.73347	9.503299	0.20185029
6	6161.00	1052500	383.823400	119.28870	2.736979	3.7	68.01348	8.681308	0.36044881
7	6030.00	1055000	250.641000	135.14560	3.805401	4.9	69.22279	11.268500	0.40833333
8	2647.00	1057000	190.918800	108.32760	5.876136	6.1	67.14903	8.722023	0.21355932
9	501.10	1073000	31.298400	112.19790	8.196050	8.3	71.04269	9.374070	0.35483871
10	2415.00	1078000	222.456600	112.43630	6.165008	6.5	67.88107	8.375160	0.28033473
11	1226.00	1121000	70.253720	128.43950	8.520348	8.8	66.00000	10.935160	0.33812950
12	1205.00	1123000	77.852710	129.65780	8.821647	9.3	66.01418	11.469090	0.48275862
13	2557.00	1134500	195.129900	119.32640	4.524818	4.6	67.61620	8.705332	0.28107345
14	3989.00	1137500	228.554300	139.19860	4.082940	5.6	72.82279	10.945170	0.21705426
15	2060.00	1139000	246.333300	109.60050	4.987730	5.7	66.66979	7.667360	0.29659091
16	539.40	1162500	49.705590	120.88100	6.675987	7.0	65.56186	10.040060	0.26007326
17	4947.00	1169000	230.641200	131.40550	6.596874	7.3	66.91585	11.058390	0.56020942

Figure 2.1(a): Some of the data of the csv file

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

Figure 2.1(b): Summary statistics of the dataset

2.2 Variable Descriptions

The Data Set contains a total number of 294 observations and 9 variables. All the variables are Numerical Variable.

- **The response variable, Y (max90)**: the 90th percentile of the time series of annual daily maxima.
- **X1 (STAID)** : 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.

2.3 Exploratory Data Analysis (EDA)

- **Histograms**

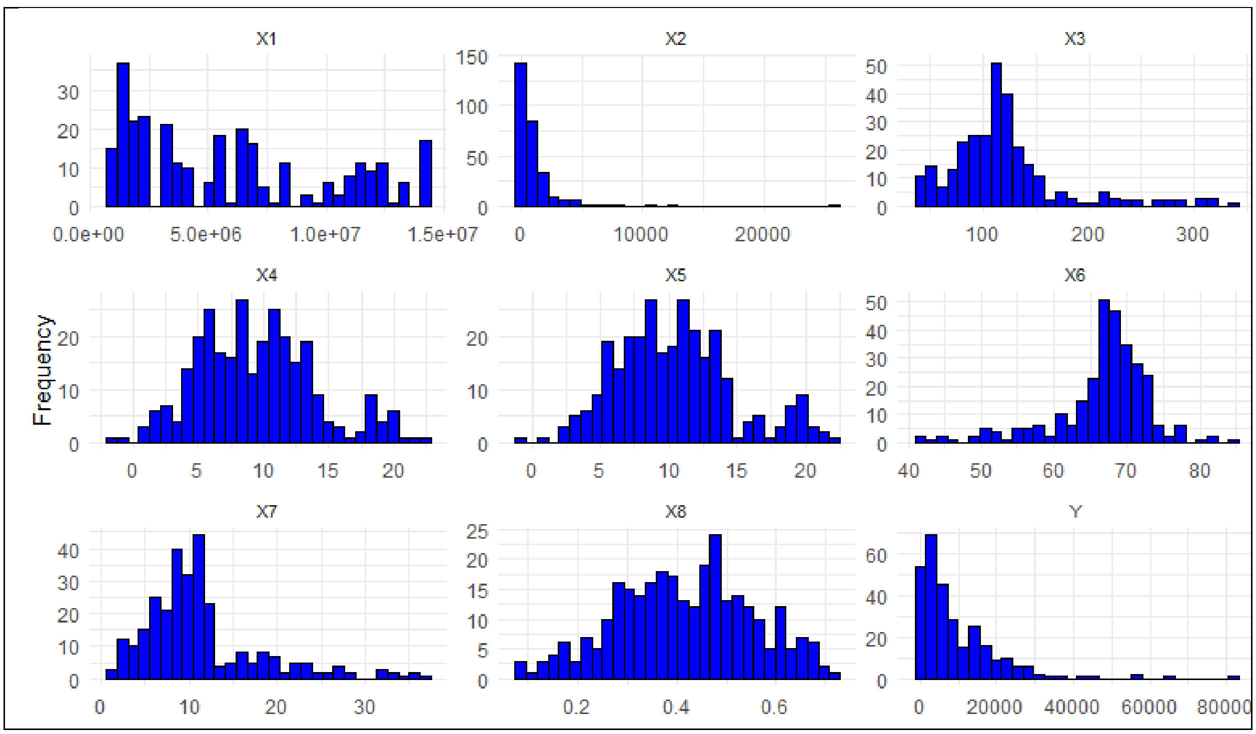

Figure 2.3(a): Histogram of Streamflow data

Each histogram helps to understand the shape of the distribution of these variables, which is crucial for deciding on appropriate statistical analyses and data transformations. Our findings are as follows:

- **X1:** This variable seems to be discrete with a few dominant categories and a long tail, indicating there are a few common values with a spread of less frequent ones.
- **X2:** The histogram for X2 shows that most of the data is clustered near zero with a rapid decline in frequency as the value increases, suggesting a right-skewed distribution.

- **X3:** The distribution for X3 appears bimodal, with two peaks, suggesting there might be two different groups or types within this variable.
- **X4:** This variable shows an unimodal distribution, slightly skewed to the right.
- **X5:** Similar to X4, X5 has an unimodal distribution, but it appears to be more symmetric around the mean
- **X6:** The histogram of X6 shows a somewhat normal distribution but with a slight left skew.
- **X7:** The variable X7 displays a right-skewed distribution.
- **X8:** This histogram shows a relatively symmetrical distribution, indicating a variable that centers around a mean with frequencies decreasing evenly as values move away from the center.
- **Y:** The final histogram represents the variable Y, which is extremely right-skewed, indicating that low values are very common, and high values are rare. This could be a flow-related measurement, such as maximum streamflow, where extremely high flow events are less frequent.

- **Box plots for key variables:**

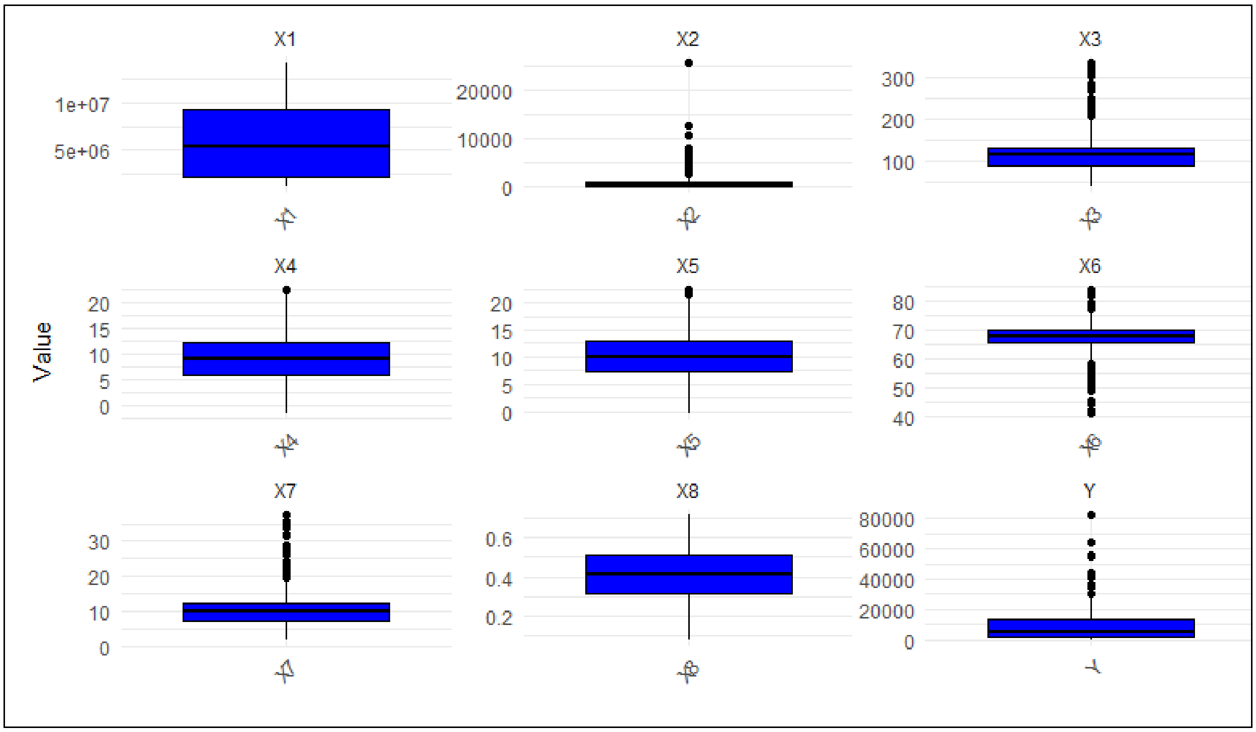

Figure 2.3(b): Boxplot of Streamflow variables

Box plots are a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. They can also include 'whiskers' which are lines extending from the box to the highest and lowest values, excluding outliers. Here's a general interpretation:

- **X1:** The data has a wide range, with a median near the lower quartile, indicating a right-skewed distribution. There's a significant spread in the data, and the values are quite large.
- **X2:** This variable has several outliers above the upper whisker, and the median is closer to the bottom of the box, which suggests a right-skewed distribution.

- **X3:** Again, there are several outliers indicating extreme values. The median is closer to the third quartile, showing that the data is left-skewed.
- **X4:** The data seems fairly symmetric around the median, but there are outliers on both ends of the spectrum.
- **X5:** This variable's distribution appears symmetric with a few outliers, and the median is centrally located within the box.
- **X6:** The distribution is slightly left-skewed, with the median closer to the third quartile and several outliers above the upper whisker.
- **X7:** There's a significant number of outliers, indicating that while the bulk of the data is within a certain range, there are several values that are much higher.
- **X8:** The box plot for this variable shows a few outliers and a median closer to the upper quartile, suggesting a slightly left-skewed distribution.
- **Y:** Like X1 and X2, Y has a right-skewed distribution with outliers present on the higher end. The values for Y are substantially higher than those for other variables, as indicated by the scale on the y-axis.

In summary, these box plots suggest varying degrees of skewness in the data, with several variables exhibiting outliers. Outliers could represent anomalies in the data, special cases, or errors. Additionally, the scales of these variables are quite different, so any comparative analysis would require normalization or standardization.

- **Scatter plots**

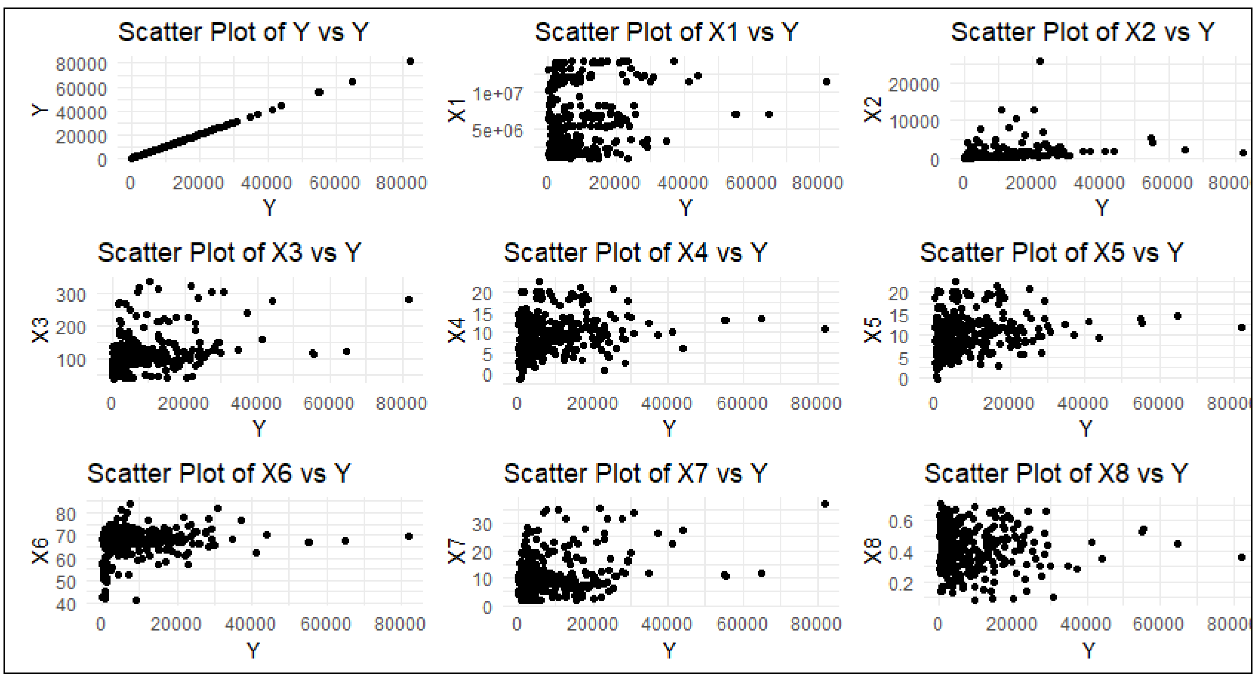

Figure 2.3(c): Scatter plots of streamflow Data

From the above the scatter plots suggest that there may be some relationships between the independent variables and Y, but none of the plots display a strong, clear, and consistent linear pattern. These relationships might be non-linear, or there might be other variables or interactions that are not captured in these individual scatter plots. Further statistical analysis, such as correlation analysis or regression modeling with interaction terms and non-linear fits, might be necessary to fully understand the relationships between Y and the independent variables.

- ✓ **Correlation analysis**

```
> print(correlation_matrix)
             Y          X1          X2          X3          X4          X5          X6
Y   1.00000000  0.1926035  0.02097545  0.30978548 -0.15557066 -0.04759166 -0.21430059
X1  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
Y   0.48388068  0.11320656
X1  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
```

Figure 2.3(d): Correlation Table

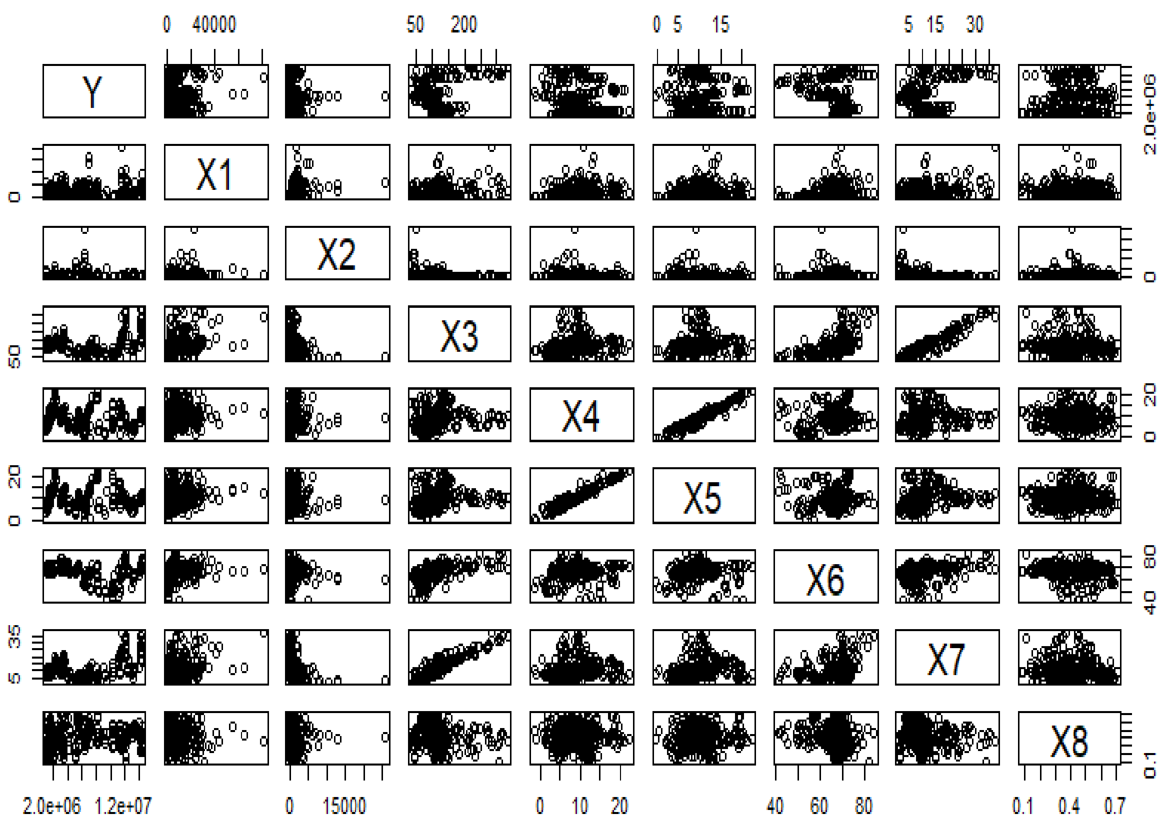

Figure 2.3(e): Correlation Matrix

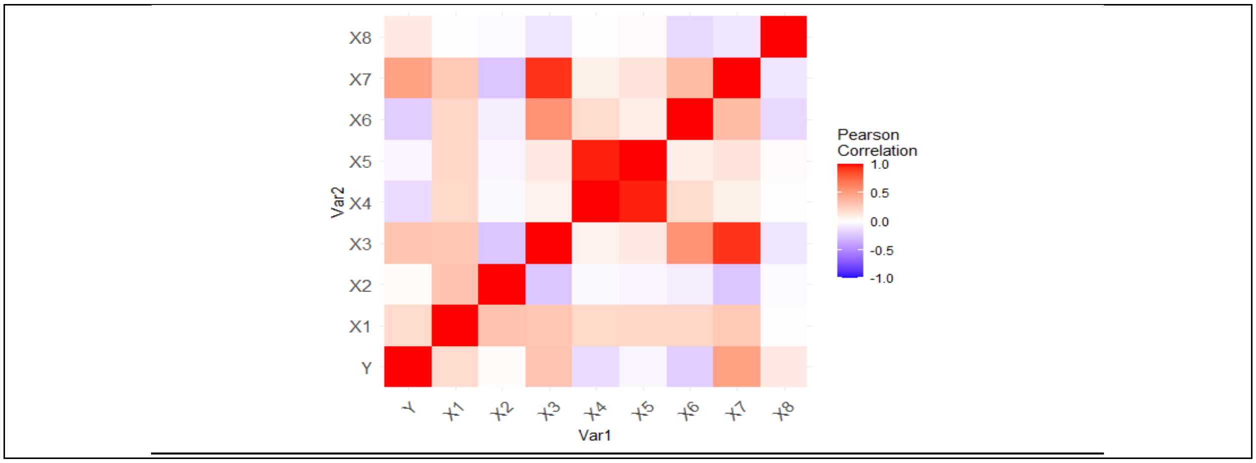

Figure 2.3(f): Heat map of Correlation Matrix

The variables X3 and X7 show the strongest positive correlation with the response variable Y, and thus may be the most significant predictors in a linear regression model for Y. However, the strong correlation between X3 and X7 also suggests that they are closely related, which could potentially lead to multicollinearity issues if both are included in the same model. Variables with very weak correlations may not be as useful individually for predicting Y.

✓ **Added Variables**

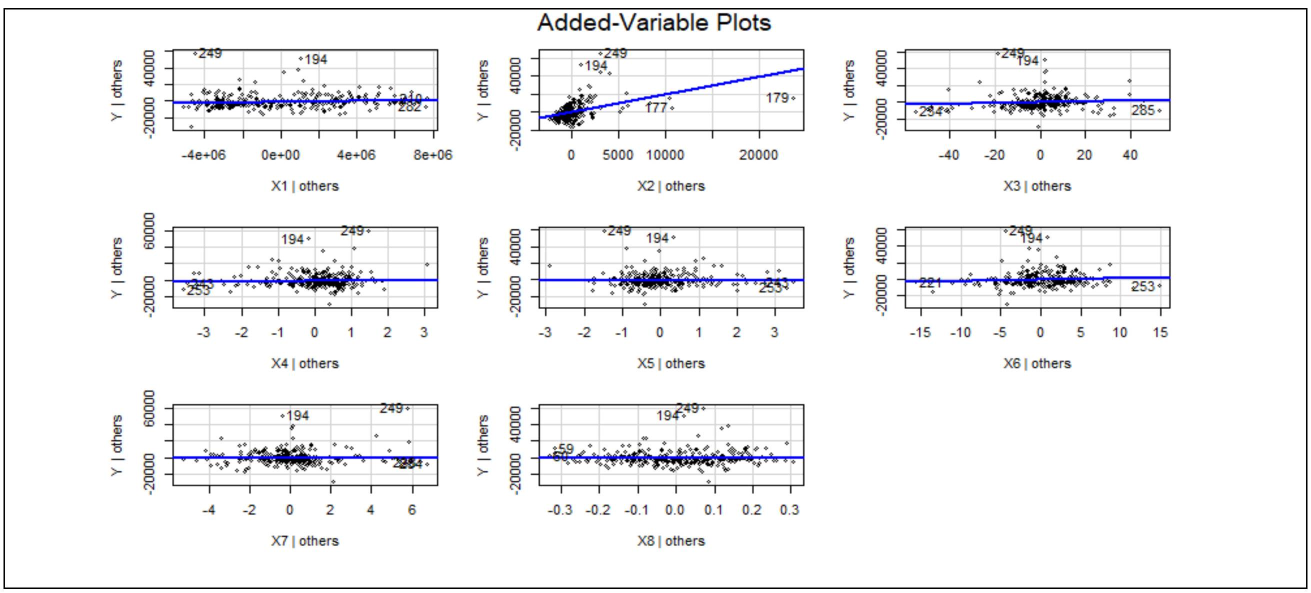

Figure 2.3(g): Added variable plot of Streamflow Data

From the above, we can see that X2 shows a positive slope with points scattered around an upward-trending line, which suggests that there is a positive relationship between Y and this variable.

After carefully analyzing the data, we've determined that it's crucial to standardize predictor variables for our linear regression model. This involves considering factors like variable sizes, relationships, importance, and model stability. In the Methods section, we'll explain this process in detail to ensure transparency and reproducibility. Our goal is to make our model more reliable and understandable, providing a clearer insight into the connections within our dataset.

3. Model Fitting

3.1 Model Selection

Given our dataset with a response variable Y (streamflow) and predictor variables X_1 to X_8, a potential multiple linear regression model could be formulated as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon$$

Where:
$\boldsymbol{\beta_0}$ is the y-intercept of the regression line.
$\boldsymbol{\beta_0}$ ***to*** $\boldsymbol{\beta_8}$ are the coefficients for each predictor variable, representing the change in the response variable for a one-unit change in the predictor, all else being equal.
$\boldsymbol{\epsilon}$ is the error term, representing the residual effect unexplained by the predictors.

3.2 Model Development

We will now fit the full model including all predictor variables using the **'lm'** function in R. This model will serve as a baseline for comparison.

```
> summary(fitstream)

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

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
```

Figure 3.2(a): Full Model

Now we are going to conduct an ANOVA to evaluate the significance of the model as a whole. This will give us an F-test for the overall fit.

```
> ##ANOVA t-test
> anova(fullmodel)
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
```

Figure 3.2(b): ANOVA for the Fullmodel

From the model output, we can see that:

- ✓ The overall model is significant (p-value < 2.2e-16).
- ✓ X2 is the only variable significantly related to Y ($p < 0.05$) with a positive relationship, indicating that as X2 increases, Y is likely to increase.
- ✓ The adjusted R-squared value is 0.2812, which means that approximately 28.12% of the variability in Y is explained by the model.
- ✓ The wide range of residuals may suggest potential issues with homoscedasticity or outliers.
- ✓ Given these points, the assumptions of the linear regression model may not be fully met.

3.3 Model Refinement

So now the regression equation from the above data is:

$$\boldsymbol{Y = -1.765 + 3.323X_1 + 1.942X_2 + 4.969X_3 + 3.440X_4 + 1.287X_5 + 1.603X_6 + 5.115X_7 + 2.405X_8}$$

The selection of the right model is critical as it can significantly impact the accuracy of the results. Insignificant predictors can be candidates for removal in a stepwise model selection process. Using a stepwise function, we chose the best variables for the model based on Adjusted R^2, Mallows' Cp, AIC, and BIC. The analysis is given for each criterion below:

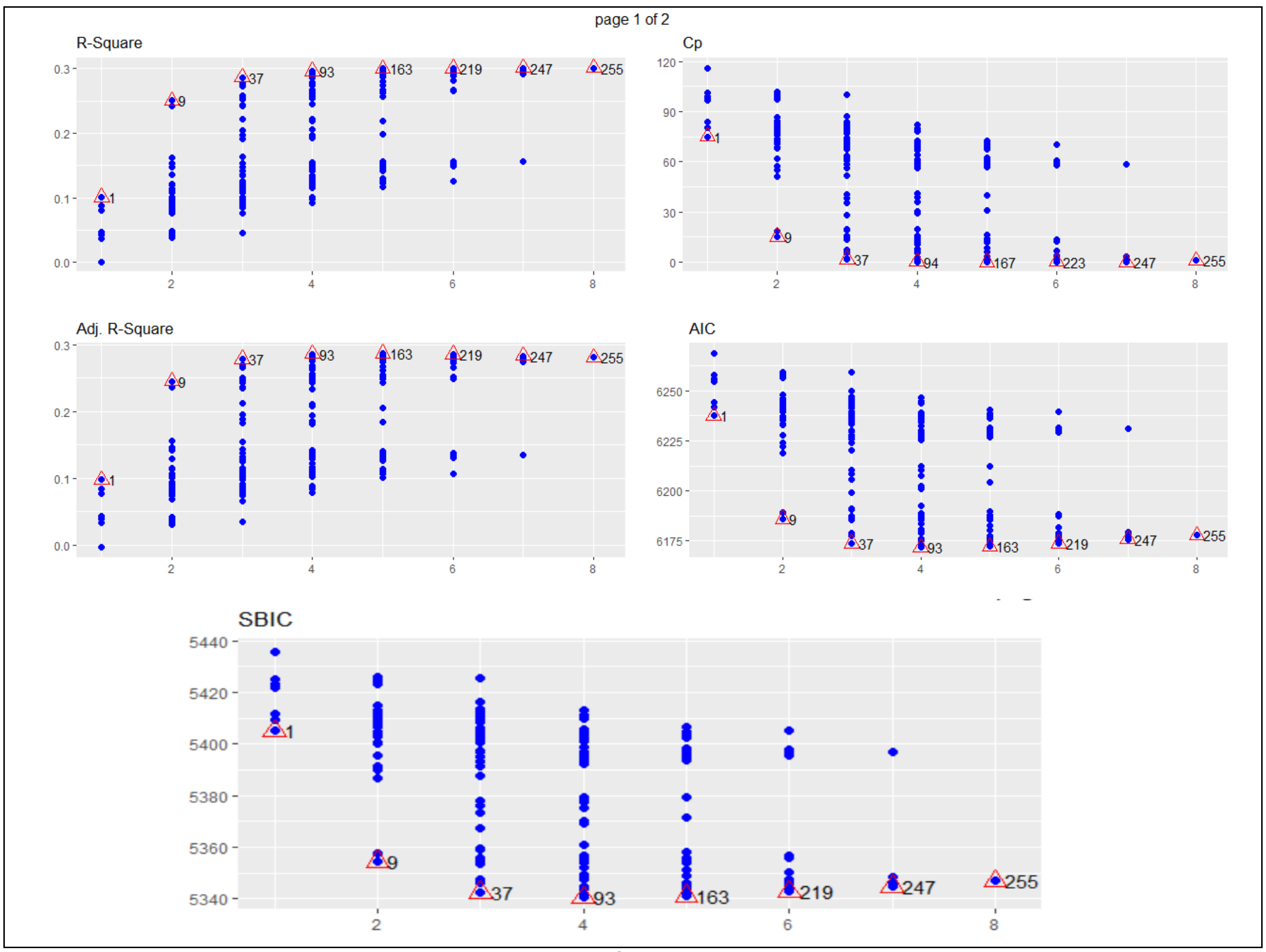

Figure 3.3(a): Adjusted R^2, Mallows' Cp, AIC, BIC

The goal of these criteria is to find a model that has the best trade-off between explaining the data and not becoming overly complex. Overly complex models may fit the current data well but can fail to generalize to new data. These criteria help to identify a model that is expected to have the best predictive performance on data. Comparing above plot we can come up to the following subset of the fullmodel. We can see that all our procedures agree on a model.

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

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

> print(b.aic[c(93,163,219,247,255),])
    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

> print(b.bic[c(93,163,219,247,255),])
    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

> print(b.press[c(93,163,219,247,255),])
    n              predictors       press
93  4             x1 x2 x3 x4 23559056540
163 5          x1 x2 x3 x4 x6 23517660394
219 6       x1 x2 x3 x4 x6 x8 23571203560
247 7    x1 x2 x3 x4 x5 x6 x8 23649866232
255 8 x1 x2 x3 x4 x5 x6 x7 x8 23730544883
```

Figure 3.3(b) subset of the fullmodel.

3.4 Stepwise Model Selection

```
Stepwise Selection Method
---------------------------

Candidate Terms:

1. x1
2. x2
3. x3
4. x4
5. x5
6. x6
7. x7
8. x8

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          S
ig.
---------------------------------------------------------------------------
----
Regression    3328314064.255          1    3328314064.255    32.756    0.0
000
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     62
84.527    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          S
ig.
---------------------------------------------------------------------------
Regression    8244853359.704          2    4122426679.852    48.495    0.0
000
```

```
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
---------------------------------------------------------------------------



                           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         S
ig.
---------------------------------------------------------------------------
Regression     8244853359.704          2    4122426679.852    48.495    0.0
000
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          S
ig.
----------------------------------------------------------------------------
----
Regression     9417018733.194          3    3139006244.398    38.636    0.0
000
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
----------------------------------------------------------------------------



                            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          S
ig.
----------------------------------------------------------------------------
Regression     9417018733.194          3    3139006244.398    38.636    0.0
000
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
---------------------------------------------------------------------------



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         S
ig.
---------------------------------------------------------------------------
Regression     9417018733.194          3    3139006244.398    38.636    0.0
000
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
---------------------------------------------------------------------------
```

4. Model Evaluation

4.1 Final Model Assessment

The final model exhibits low multicollinearity among the predictors. This suggests that the model is well-specified with the chosen variables X2, X3, and X5. Thus, we can proceed with this model as our selected approach for further analysis. Here is a summary of the final model:

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

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
```

Figure 4.1(a) Final Model

4.2 Regression Diagnosis & Assumption Checking

Regression diagnostics refers to the techniques applied during data analysis to verify the accuracy of a regression model's predictions. It involves checking the data for problems like outliers, multicollinearity, heteroscedasticity, or autocorrelation to ensure the integrity and reliability of the model's results.

The residuals plot:

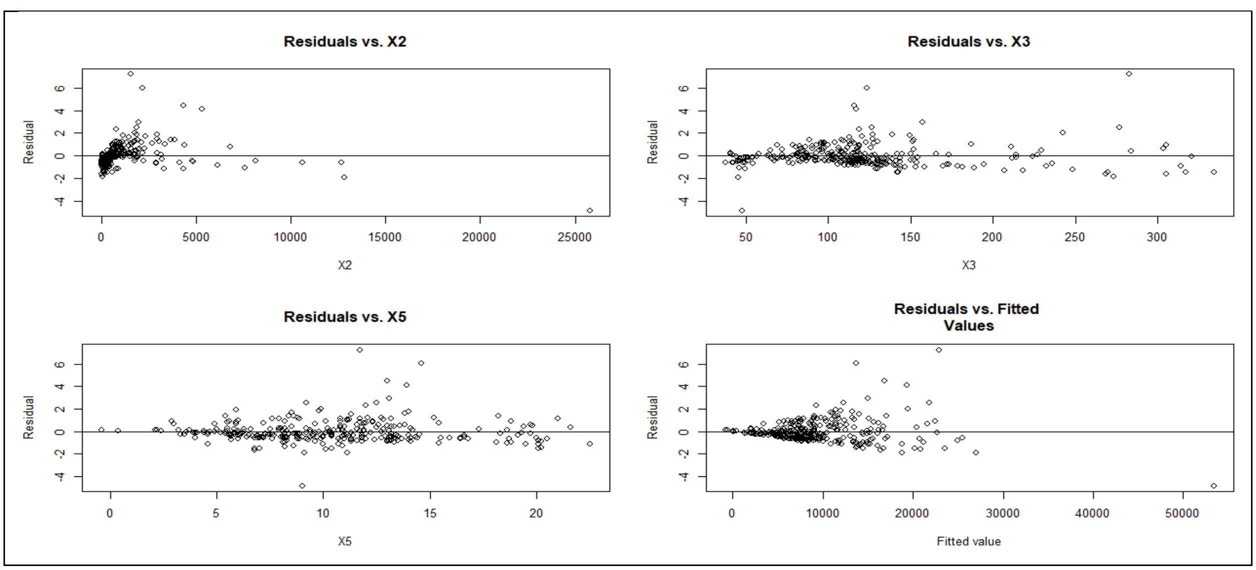

Figure 4.2(a) Residual plot of Final Model

Interpretation of Residuals vs. Predictors (X2, X3, X5): These plots are used to check for non-linearity between predictors and the response variable. We can see the residuals are randomly scattered around the horizontal line at 0, with no discernible pattern. This indicates that the relationship between the predictor and the response is linear and does not need a transformation or a different model.

Interpretation of Residuals vs. Fitted Values: This plot checks the homoscedasticity assumption — that the variance of the error terms is constant across all levels of the independent variables. Here the residuals are forming a flannel shape suggesting non homoscedastic.

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

```
vif(reduced.lmfit)
      X2       X3       X5
1.065494 1.075808 1.012455
```

Figure 4.2(b) the variance inflation factor (VIF) of Final Model

Interpretation of VIF Function: The outcomes of the variance inflation factor (VIF) analysis suggest that multicollinearity is not a significant concern for the final model, as all VIF values fall below the threshold of 5. However, since the p-values in both tests do not exceed the 0.05 threshold, we cannot reject the null hypothesis which assumes the independence of the errors.
To further substantiate the findings observed in the residual plots, both the Breusch-Pagan test and the Durbin-Watson test are conducted.

Studentized Breusch-Pagan test: ***:*** The Breusch-Pagan test is used to determine whether there is heteroscedasticity in the residuals.
In the case of our data,

```
        studentized Breusch-Pagan test

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

Figure 4.2(c) Studentized Breusch-Pagan test

Interpretation of Studentized Breusch-Pagan test: Since the p-value of **1.197e-08** is less than the signifi cance level of 0.05, the Breusch-Pagan test suggests that, at the 0.05 significance level, there is strong in dication of heteroscedasticity in the residuals of the regression model.

Durbin-Watson test:

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

Figure 4.2(d) Durbin-Watson test

Interpretation of Durbin-Watson test: p-value: 4.424e-08 (or 4.424×10^{-8}) is significantly less than the conventional alpha level of 0.05, leading to the rejection of the null hypothesis.
Therefore, the test results suggest that there is statistically significant positive autocorrelation in the residuals of your regression model, which implies that the residuals are not independent of each other.

Q-Q Plot: We will begin by examining a Q-Q Plot of our residuals in order to determine normalcy.

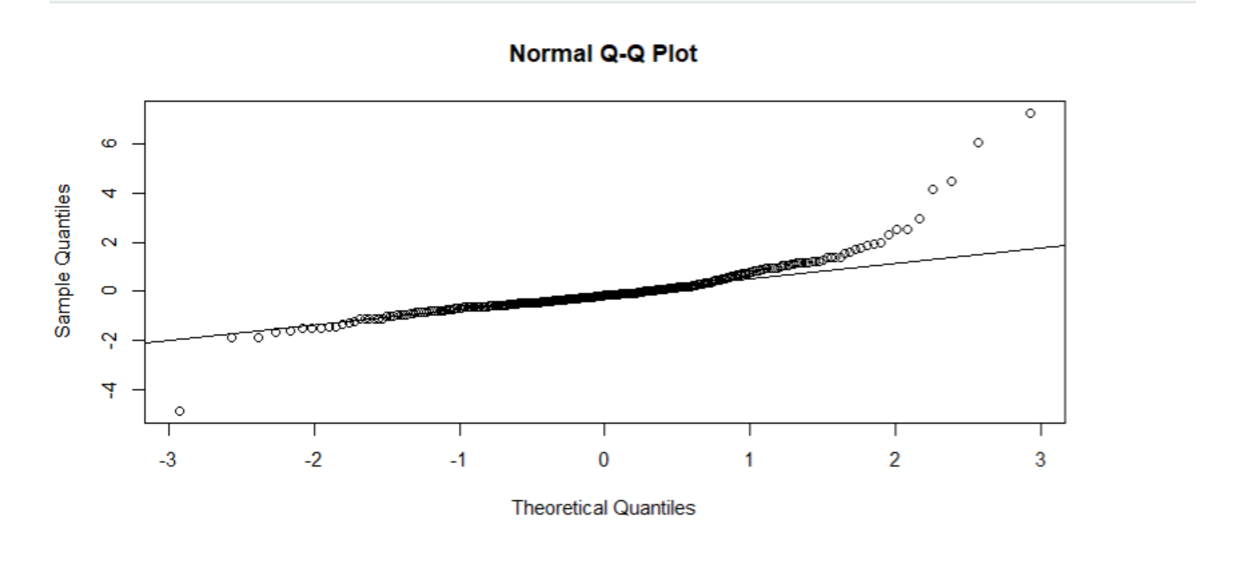

Figure 4.2(e) Normal Q-Q plot

Interpretation of Q-Q Plot: The distribution exhibits a rightward skew, indicating a concentration of data points towards the right side.

To assess the normality assumption for this distribution, a Shapiro-Wilk test can be utilized.

Shapiro-Wilk test:

```
shapiro.test(res)

	Shapiro-Wilk normality test

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

Figure 4.2(f) Shapiro-Wilk test

Interpretation of Shapiro-Wilk test: A p-value less than 2.2e-16 (which is essentially zero) is significantly less than the common alpha level of 0.05. This indicates strong evidence against the null hypothesis, leading to its rejection.

In summary, the Shapiro-Wilk test result suggests that the residuals of the model do not follow a normal distribution, which is one of the key assumptions of many statistical tests and models, including linear regression. This could affect the validity of the model's inference statistics and might require transformation of the data.

4.3 Outliers Detection

For the detection of outliers, we look at the DFBETAS, DFFITS, and Cook's Distance plots.

DFFITS:

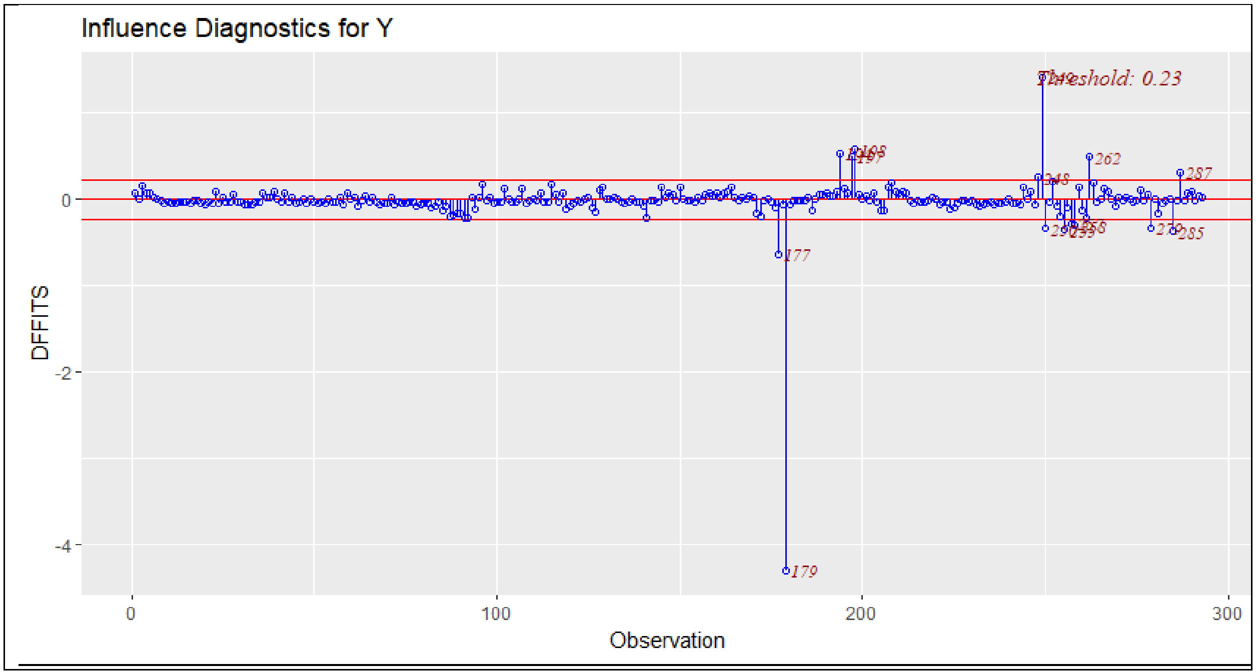

Figure 4.3(a) DFFITS

DFBETAS:

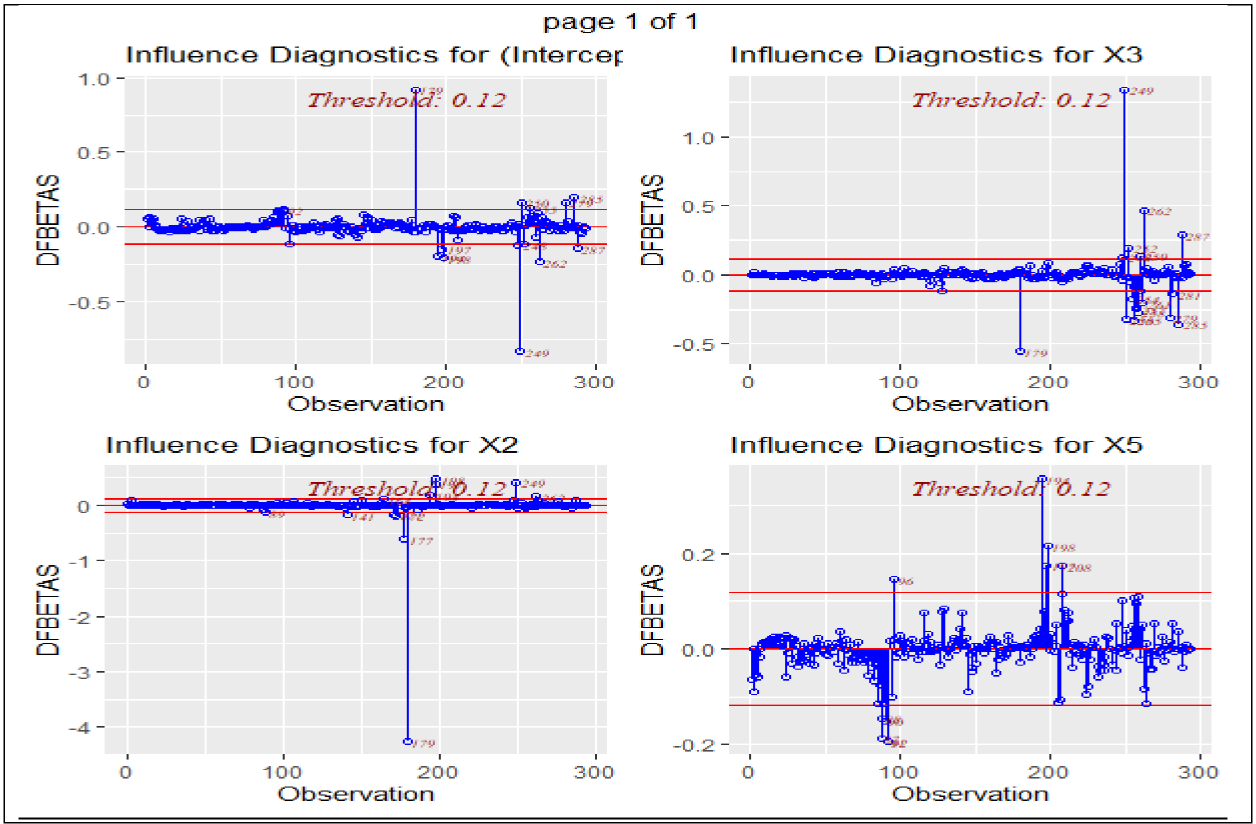

Figure 4.3(b) DFBETAS

Cook's distance:

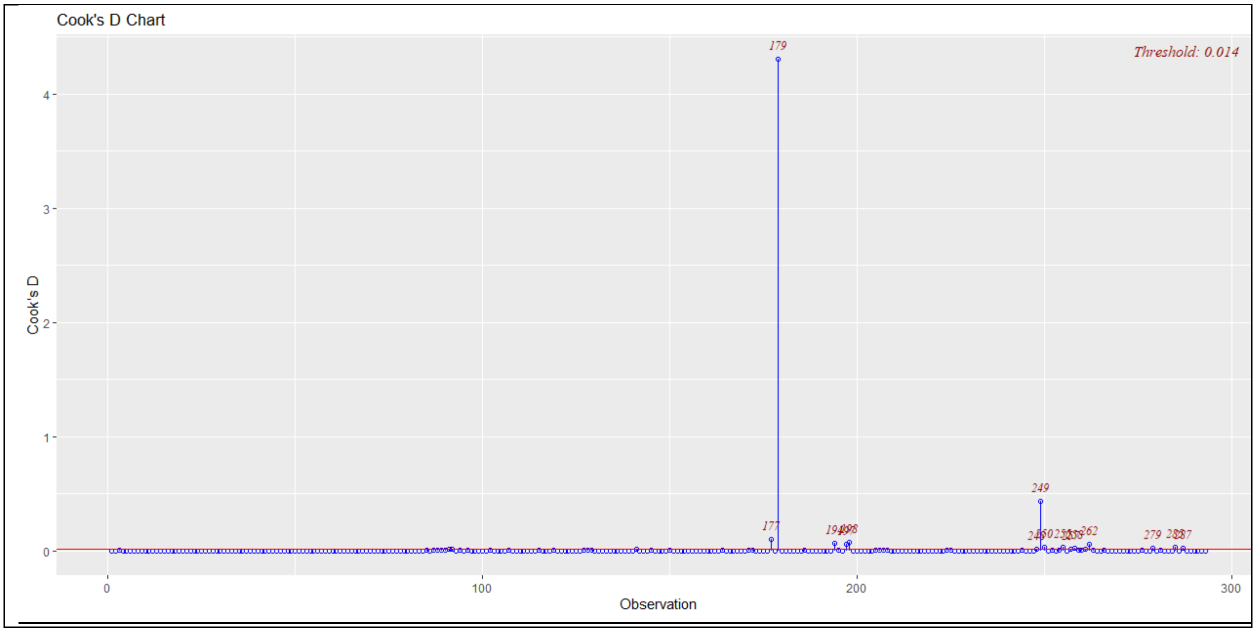

Figure 4.3(c) Cook's distance

In our dataset, the points labeled 179 and 249 stand out as outliers and contribute significantly to the rightward skew observed in our quantile-quantile (Q-Q) plot.

To address the deviation from normality in our model's residuals, we plan to apply a Box-Cox Transformation. This procedure will attempt to correct the skewness and align the distribution of residuals closer to normality.

4.4 Data Transformation

The streamlined model demonstrates linearity and consistent variance, indicating the need for a Box-Cox transformation to improve model performance. The transformation does not affect the initially detected multicollinearity. Utilizing R's Box-Cox function, the optimal lambda value is determined to be 0.2792849. With this lambda, the response variable can be transformed, paving the way for an updated model formulation.

```
lambda
[1] 0.2792849
```

To apply this transformation, we would raise our response variable to the power of 0.2792849. This adjusted variable should then be used in place of the original response variable in our regression model. By transforming the response variable, we're likely to see improvements in the model's adherence to assumptions like normality of residuals, which can, in turn, lead to more reliable and valid regression results.

```
> summary(boxcox.lmfit)

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

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
```

Figure 4.4(a) Summary statistics of the new dataset

In the context of our transformed streamflow dataset, ANOVA will help us understand the impact of each independent variable has on the transformed response variable (Y). By doing so, we aim to identify which predictors significantly contribute to variations in streamflow and assess the overall fit of our regression model. This step is essential in refining our model and ensuring that it accurately captures the underlying relationships in the data.

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

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

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
```

Figure 4.4(b) ANOVA of the new dataset

The adjusted R-squared value has increased, indicating an improvement in the model. However, it's crucial to verify that this enhanced model meets all the required assumptions. Therefore, our next step involves reassessing the model's diagnostic measures, beginning with an evaluation of the constancy of variance.

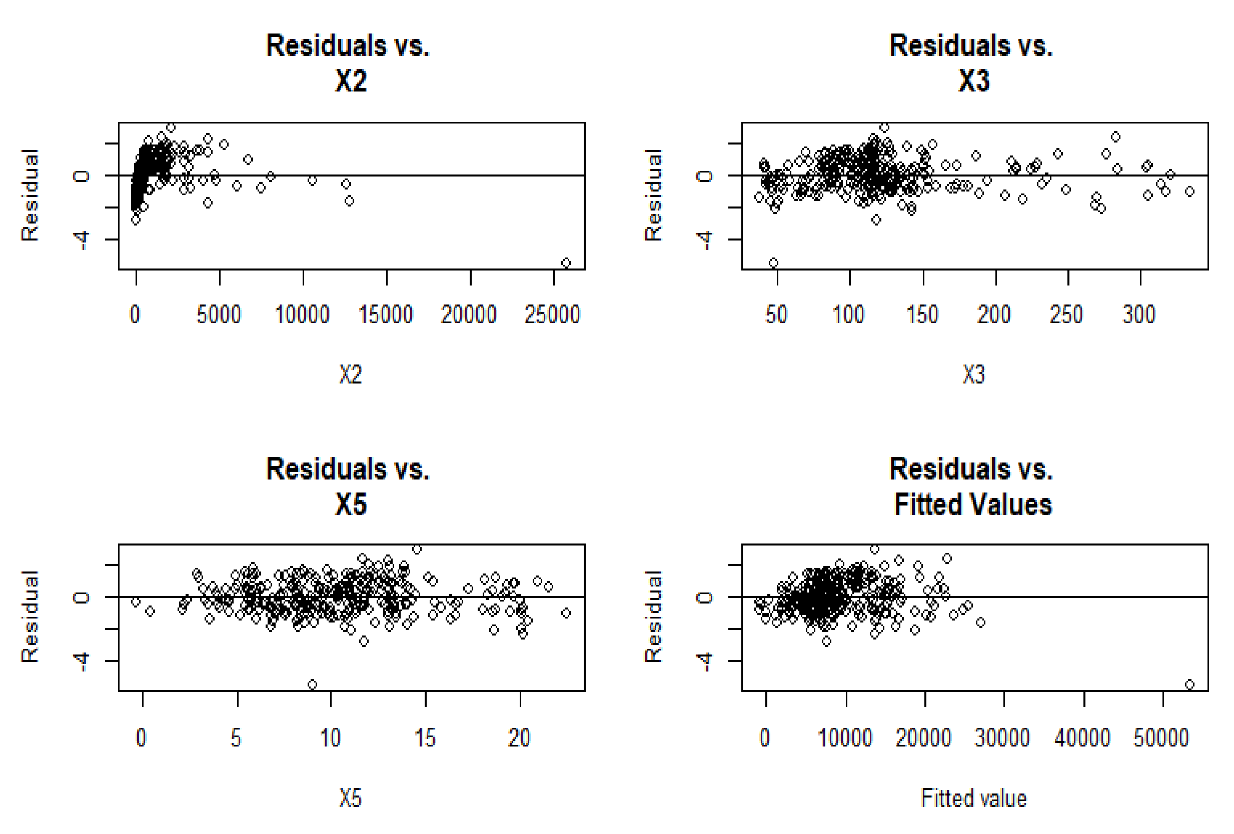

Figure 4.4(c) Residual plot of the new dataset

The graphs suggest a generally even distribution of residuals. Subsequent application of the Breusch-Pagan and Durbin-Watson tests confirms that our assumption of constant variance in the errors holds reasonably well.

Studentized Breusch-Pagan & Durbin-Watson test

```
bptest(boxcox.lmfit)

studentized Breusch-Pagan test

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

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

        Durbin-Watson test

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

The scatter plots display a relatively consistent distribution of data points. However, conducting a Breusch-Pagan test validates that the assumption of consistent variance of errors is well-founded.

The uniform distribution of residuals across an index, coupled with a high p-value in the Durbin-Watson test, supports the assumption that the residuals are independent.

Normal Q-Q plot

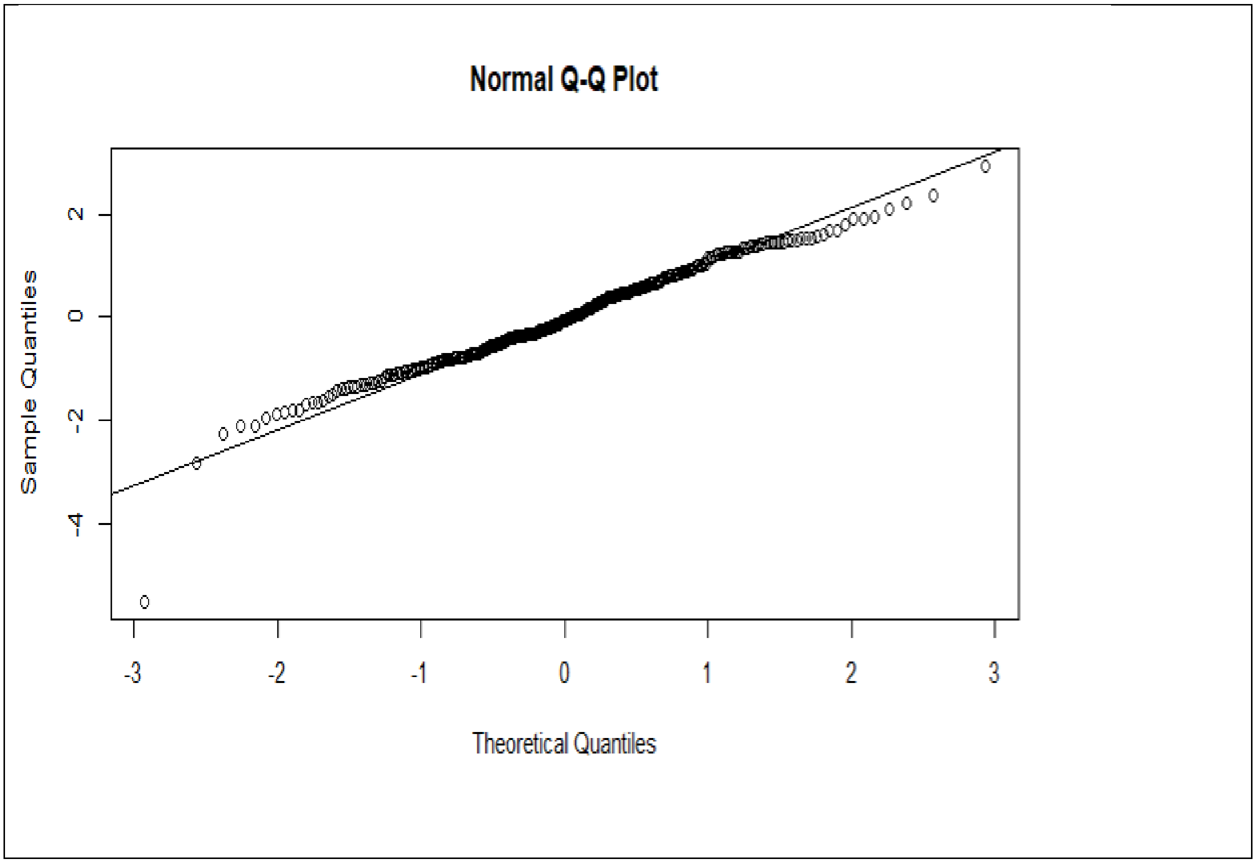

Figure 4.4(d) Normal Q-Q plot of the new dataset

Shapiro-Wilk normality test

```
Shapiro-Wilk normality test

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

Our model now meets the normality assumption, previously unmet. The Q-Q plot indicates a satisfactory adherence to normality, as evidenced by a significant increase in the p-value from the Shapiro-Wilk test.

5. Conclusion

5.1 Summary of Findings

Since normality is achieved, we can recall the summary of the function

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

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

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 above the summary of the final optimal model is:

$$\boldsymbol{Y = 5.043 + 7.833e\text{-}04\, X_2 + 2.766e\text{-}02\, X_3 + 2.058e\text{-}01\, X_5}$$

Where the predictor variables of the model are

- **The response variable, Y (max90)**: the 90th percentile of the time series of annual daily maxima.
- **X2 (DRAIN_SQKM)** : the drainage area.
- **X3(PPTAVG_BASIN)** : the average basin precipitation.
- **X5(T_AVG_SITE)** : the average temperature at the stream location.

The F-statistic, valued at 49.94 across six predictor variables, along with a p-value of less than 2.2e-16, indicates a statistically significant relationship between the response variable and the set of predictor variables.

5.2 Model Performance:

For the streamflow dataset, we initiated our investigation with an exploratory data analysis and refined our model based on a significance threshold of 0.05. We adopted a model selection approach to ensure the model's validity, and implemented diagnostic checks to confirm its accurate fit. The model was subsequently enhanced, leading to a notable improvement in the Adjusted R-squared value. Ultimately, we arrived at an evolved and more effective model.

5.3 Implications of the Study

Our model offers insights into the factors influencing streamflow. This can help in understanding seasonal variations, the impact of climatic conditions like precipitation and temperature, and how geographical features affect streamflow. The model can be used to assess the health of aquatic ecosystems. By understanding how different environmental variables impact streamflow, we can predict the potential impacts of environmental changes. Our model's ability to predict streamflow can be crucial in flood forecasting. Accurate predictions of high streamflow events can assist in early warning systems, helping to mitigate the impacts of flooding on communities and infrastructure.

Our findings can contribute to broader hydrological research, particularly in understanding how climate change affects water cycles. For regions dependent on agriculture, understanding streamflow is key to irrigation planning and crop management. In urban areas, managing streamflow is critical for infrastructure development, particularly for designing drainage systems and ensuring sustainable water supply.

5.4 Final Thought

Our project underscores the significance of thorough data analysis. Each step, from exploratory analysis to model fitting and diagnostics, plays a crucial role in understanding complex natural phenomena like streamflow. This work highlights the intersection of data science, environmental studies, and hydrology. The project illustrates the iterative nature of model buildings. It's a continuous process of testing, refining, and improving. Beyond its practical applications, our project serves as a valuable educational resource, demonstrating complex statistical concepts in a tangible and relevant context.