808: Harnessing Data Analytics for Pollution level forecasting using the air quality attributes.

First Name	Last Name	Email (hawk.iit.edu)	Student ID
Srujan Shekar	Shetty	Sshetty16@hawk.iit.edu	**A20529733**
Senthilvel	Rajakarthihan		A20555853

Table of Contents

1. Introduction

Introduce the background and motivations

Background & Motivation : Concern over air quality and its effects on public health is growing along with urbanization. It has been determined that one of the main causes of air pollution and its detrimental impacts on health is the existence of fine particulate matter, specifically PM2.5. With an emphasis on the Beijing PM2.5 dataset, the goal of this project is to create a pollution level forecasting model using air quality parameters. The goal of this project is to give precise and timely PM2.5 concentration predictions so that preventative actions can be taken to lessen the health risks related to poor air quality. Our goal is to promote public health and contribute to environmental monitoring by utilizing data analysis techniques.

2. Data

Introduce your data, such as where did you get it (provide the URL if possible), how large it is, what are the variables/features, what are the variable types, etc

The data is from datasciencedojo- https://code.datasciencedojo.com/datasciencedojo/datasets/tree/master/Beijing%20PM2.5

The size of the data -Row Size- 43825; Column Size-13.

Data Dictionary

Column Position	Atrribute Name	Definition	Data Type
1	No	No: row number	Quantitative
2	Year	Year: year of data in this row	Quantitative
3	Month	Month: month of data in this row	Quantitative
4	Day	Day: day of data in this row	Quantitative
5	Hour	Hour: hour of data in this row	Quantitative
6	PM2.5	PM2.5: PM2.5 concentration (ug/m$^{3)}$	Quantitative
7	DEWP	DEWP: Dew Point (â„ƒ)	Quantitative
8	TEMP	TEMP: Temperature (â„ƒ)	Quantitative
9	PRES	PRES: Pressure (hPa)	Quantitative
10	cbwd	cbwd: Combined wind direction	Quantitative
11	lws	lws: Cumulated wind speed (m/s)	Quantitative
12	Ir	Is: Cumulated hours of snow	Quantitative
13	Is	Ir: Cumulated hours of rain	Quantitative

3. Problems to be Solved

List the problems you want to solve

To estimate the PM2.5 which is air quality index(dependent variable) of the Beijing city based on independent variables such as DEWP: Dew Point (â„ƒ),TEMP: Temperature (â„ƒ), PRES: Pressure (hPa),cbwd: Combined wind direction,Iws: Cumulated wind speed (m/s),Is: Cumulated hours of snow,Ir: Cumulated hours of rain,Hour,Day,Year,Month.

Dataset also deals with following problems:

- Missing values in Dataset
- Duplicate rows in Dataset

4. Solutions

You can use linear regression to predict air quality(Pm2.5),
Independent variables such as DEWP: Dew Point (â„ƒ),TEMP: Temperature (â„ƒ),
PRES: Pressure (hPa),cbwd: Combined wind direction,Iws: Cumulated wind speed (m/s),Is:
Cumulated hours of snow,Ir: Cumulated hours of rain,Hour,Day,Year,Month.
and air quality acting as the dependent variable(Pm2.5),.

The get to the solution ,problem is split into 3 parts:

1.Data Pre-Processing

1.1 Exploring Data-Set
1.2 Filling Missing Values in Dataset with mean
1.3 Duplicate Rows Check
1.4 Correlation of Dataset
1.5 Using hold-out evaluation only, 80% as training

2.Linear Regression

2.1 Full model
2.2 Backward method using p-value in t-test as metric.
2.3 Backward method using AIC as metric
2.4 Forward method using AIC as metric
2.5 Stepwise method using ACI as a metric

3.Post Processing

3.1 Best Model from Linear Regression(Skipped RMSE Output Slide in PPT)
3.2 Model Diagnosis
3.3 Improving Model(Multi-Collinearity output slide,5 Cross Validation)

The best fit linear regression model can be used to predict air quality, with Dew Point (â„ƒ),TEMP: Temperature (â„ƒ), and other property qualities acting as independent factors and air quality acting as the dependent variable.

5. Experiments and Results

5.1. Methods and Process

1.Data Pre-Processing

1.1 Exploring Data-Set

To understand the dataset better we explored the data

The Code(comments explain the code):

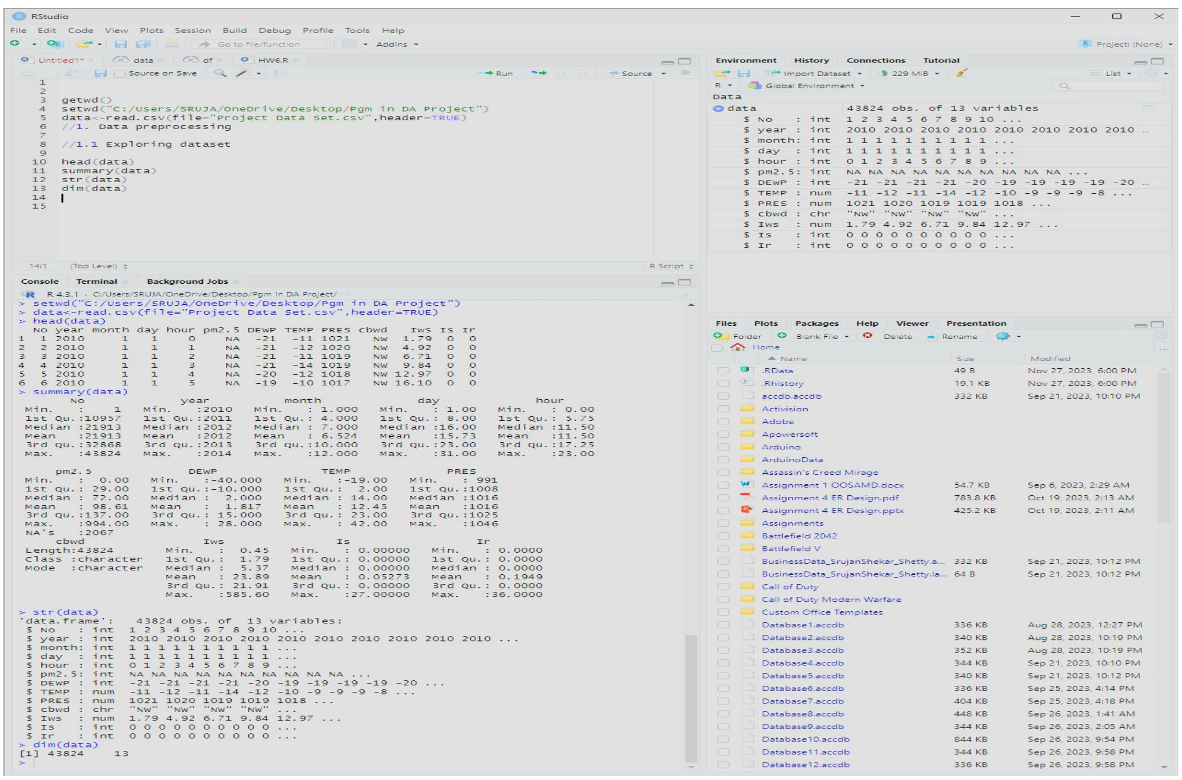
head(data) *# displays the top few rows of your dataset.*

summary(data) *# summary of the dataset's variables' respective statistical data.*

str(data) *#displays data types giving the dataset's structure*

dim(data) *#displays dataset's number of rows and columns*

The Snapshot:



1.2 Filling Missing Values in Dataset with mean

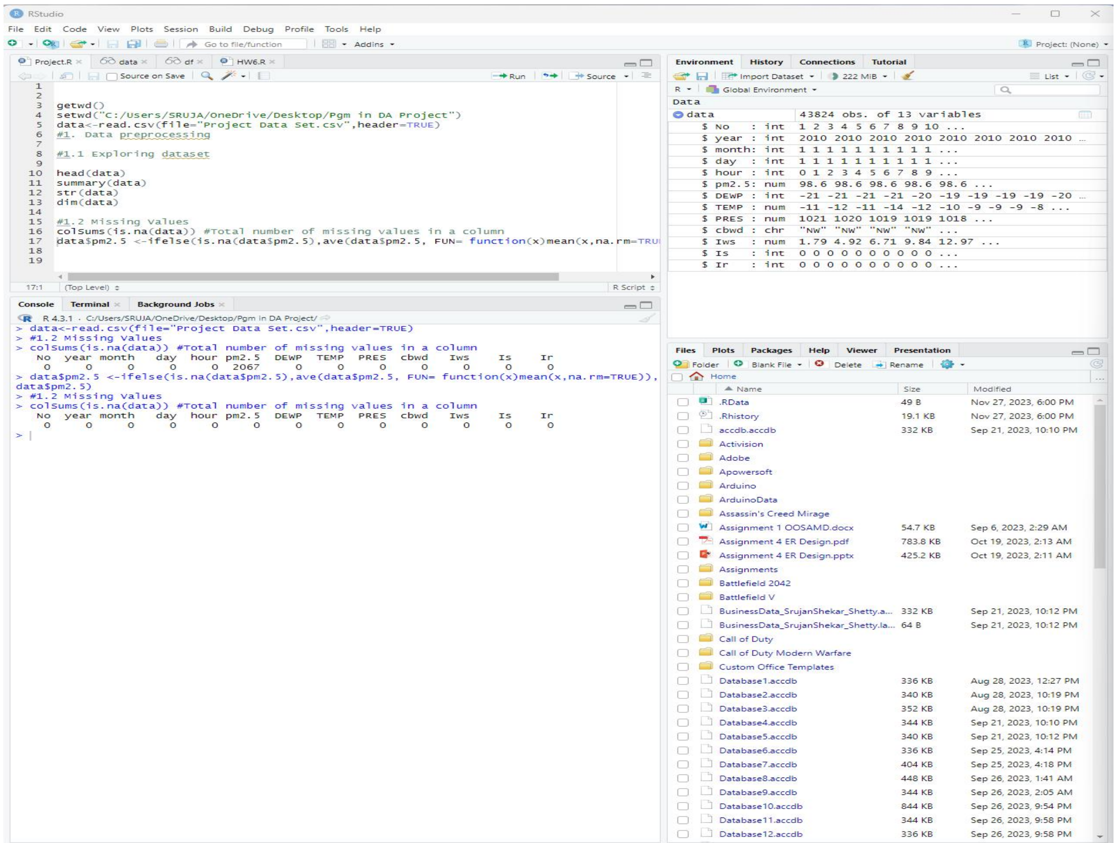
There were missing values in the data set,we filled them all with mean values.

The Code(comments explain the code):

colSums(is.na(data)) *#Total number of missing values in a column*

data$pm2.5 <-ifelse(is.na(data$pm2.5),ave(data$pm2.5, FUN= function(x)mean(x,na.rm=TRUE)),data$pm2.5) *#Fill missing values with mean values*

<u>The Snapshot:</u>



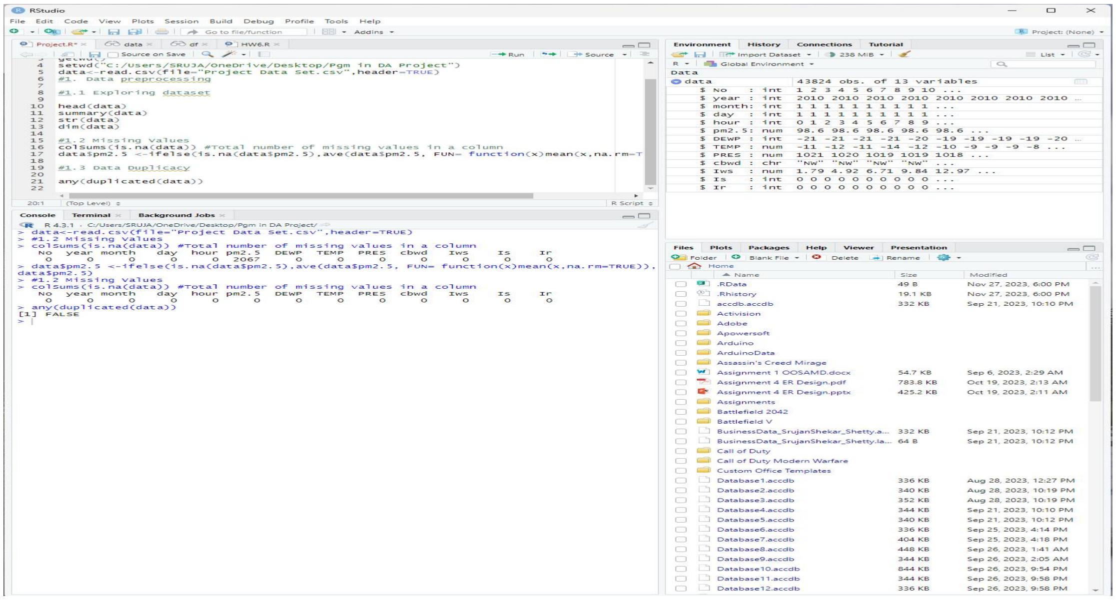
1.3 Duplicate Rows Check

In order to avoid error Duplicate rows check was made

The Code(comments explain the code):

any(duplicated(data)) *#check duplicacy*

The Snapshot:

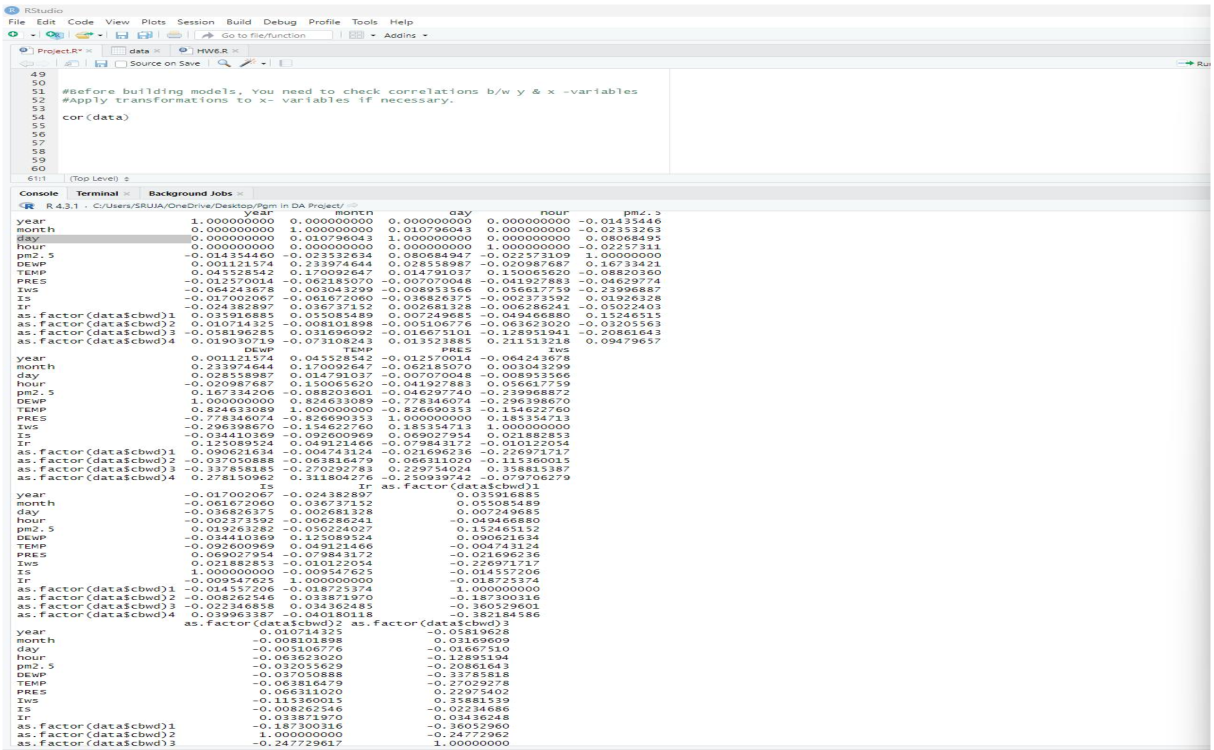


1.4 Correlation of Dataset:

We check correlation of datasets before building a model to avoid highly correlated independent variables.

The Code(comments explain the code):

Cor(data)



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Project.R* data HW6.R

Source on Save

```r
#Before building models, You need to check correlations b/w y & x -variables
#Apply transformations to x- variables if necessary.

cor(data)
```

61:1 (Top Level)

Console Terminal Background Jobs

R 4.3.1 · C:/Users/SRUJA/OneDrive/Desktop/Pgm in DA Project/

```
                             year        month          day         hour       pm2.5
year                  1.000000000  0.000000000  0.000000000  0.000000000 -0.01435446
month                 0.000000000  1.000000000  0.010796043  0.000000000 -0.02353263
day                   0.000000000  0.010796043  1.000000000  0.000000000  0.08068495
hour                  0.000000000  0.000000000  0.000000000  1.000000000 -0.02257311
pm2.5                -0.014354460 -0.023532634  0.080684947 -0.022573109  1.00000000
DEWP                  0.001121574  0.233974644  0.028558987 -0.020987687  0.16733421
TEMP                  0.045528542  0.170092647  0.014791037  0.150065620 -0.08820360
PRES                 -0.012570014 -0.062185070 -0.007070048 -0.041927883 -0.04629774
Iws                  -0.064243678  0.003043299 -0.008953566  0.056617759 -0.23996887
Is                   -0.017002067 -0.061672060 -0.036826375 -0.002373592  0.01926328
Ir                   -0.024382897  0.036737152  0.002681328 -0.006286241 -0.05022403
as.factor(data$cbwd)1  0.035916885  0.055085489  0.007249685 -0.049466880  0.15246515
as.factor(data$cbwd)2  0.010714325 -0.008101898 -0.005106776 -0.063623020 -0.03205563
as.factor(data$cbwd)3 -0.058196285  0.031696092 -0.016675101 -0.128951941 -0.20861643
as.factor(data$cbwd)4  0.019030719 -0.073108243  0.013523885  0.211513218  0.09479657
                             DEWP         TEMP         PRES          Iws
year                  0.001121574  0.045528542 -0.012570014 -0.064243678
month                 0.233974644  0.170092647 -0.062185070  0.003043299
day                   0.028558987  0.014791037 -0.007070048 -0.008953566
hour                 -0.020987687  0.150065620 -0.041927883  0.056617759
pm2.5                 0.167334206 -0.088203601 -0.046297740 -0.239968872
DEWP                  1.000000000  0.824633089 -0.778346074 -0.296398670
TEMP                  0.824633089  1.000000000 -0.826690353 -0.154622760
PRES                 -0.778346074 -0.826690353  1.000000000  0.185354713
Iws                  -0.296398670 -0.154622760  0.185354713  1.000000000
Is                   -0.034410369 -0.092600969  0.069027954  0.021882853
Ir                    0.125089524  0.049121466 -0.079843172 -0.010122054
as.factor(data$cbwd)1  0.090621634 -0.004743124 -0.021696236 -0.226971717
as.factor(data$cbwd)2 -0.037050888 -0.063816479  0.066311020 -0.115360015
as.factor(data$cbwd)3 -0.337858185 -0.270292783  0.229754024  0.358815387
as.factor(data$cbwd)4  0.278150962  0.311804276 -0.250939742 -0.079706279
                               Is           Ir as.factor(data$cbwd)1
year                 -0.017002067 -0.024382897           0.035916885
month                -0.061672060  0.036737152           0.055085489
day                  -0.036826375  0.002681328           0.007249685
hour                 -0.002373592 -0.006286241          -0.049466880
pm2.5                 0.019263282 -0.050224027           0.152465152
DEWP                 -0.034410369  0.125089524           0.090621634
TEMP                 -0.092600969  0.049121466          -0.004743124
PRES                  0.069027954 -0.079843172          -0.021696236
Iws                   0.021882853 -0.010122054          -0.226971717
Is                    1.000000000 -0.009547625          -0.014557206
Ir                   -0.009547625  1.000000000          -0.018725374
as.factor(data$cbwd)1 -0.014557206 -0.018725374           1.000000000
as.factor(data$cbwd)2 -0.008262546  0.033871970          -0.187300316
as.factor(data$cbwd)3 -0.022346858  0.034362485          -0.360529601
as.factor(data$cbwd)4  0.039963387 -0.040180118          -0.382184586
                      as.factor(data$cbwd)2 as.factor(data$cbwd)3
year                            0.010714325           -0.05819628
month                          -0.008101898            0.03169609
day                            -0.005106776           -0.01667510
hour                           -0.063623020           -0.12895194
pm2.5                          -0.032055629           -0.20861643
DEWP                           -0.037050888           -0.33785818
TEMP                           -0.063816479           -0.27029278
PRES                            0.066311020            0.22975402
Iws                            -0.115360015            0.35881539
Is                             -0.008262546           -0.02234686
Ir                              0.033871970            0.03436248
as.factor(data$cbwd)1          -0.187300316           -0.36052960
as.factor(data$cbwd)2           1.000000000           -0.24772962
as.factor(data$cbwd)3          -0.247729617            1.00000000
```

1.5 Using hold-out evaluation only, 80% as training

Estimating a machine learning model's performance on fresh, untested data is made easier with the aid of holdout evaluation.

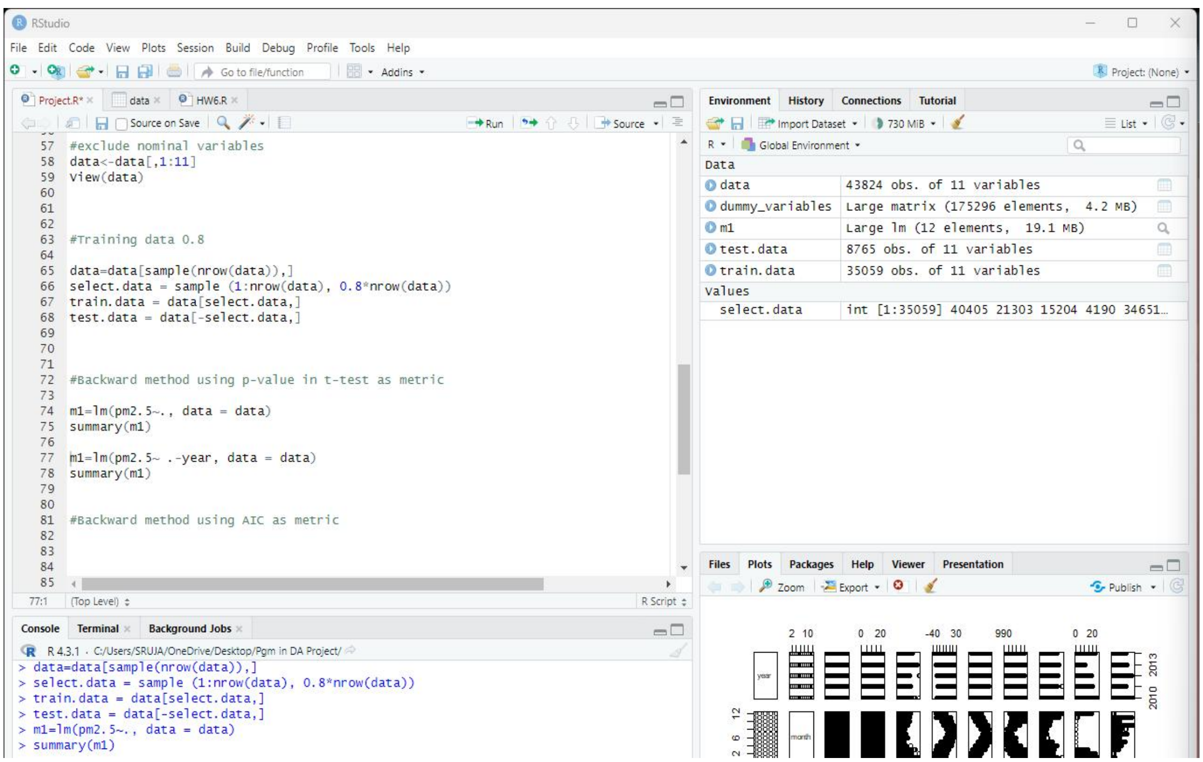
The Code(comments explain the code):

data=data[sample(nrow(data)),]

select.data = sample (1:nrow(data), 0.8*nrow(data)) *#80% as select data*

train.data = data[select.data,] *#Select data as train data*

test.data = data[-select.data,]*#Select data as*



Hypothesis Testing for Individual Coefficients:

Null Hypothesis (H0): There is no significant relationship between variables like DEWP, TEMP, etc., and PM2.5 concentration.

Alternative Hypothesis (H1): There is a significant relationship between variables like DEWP, TEMP, etc., and PM2.5 concentration.

Test Method: Perform z-tests for individual coefficients in a multiple linear regression model. The null hypothesis is rejected if the p-value is below a predetermined significance level (e.g., 0.05).

Conclusion:We reject Null Hypothesis (H0) based on the test results

2.Linear Regression

2.1 Full model
2.2 Backward method using p-value in t-test as metric.
2.3 Backward method using AIC as metric
2.4 Forward method using AIC as metric
2.5 Stepwise method using ACI as a metric

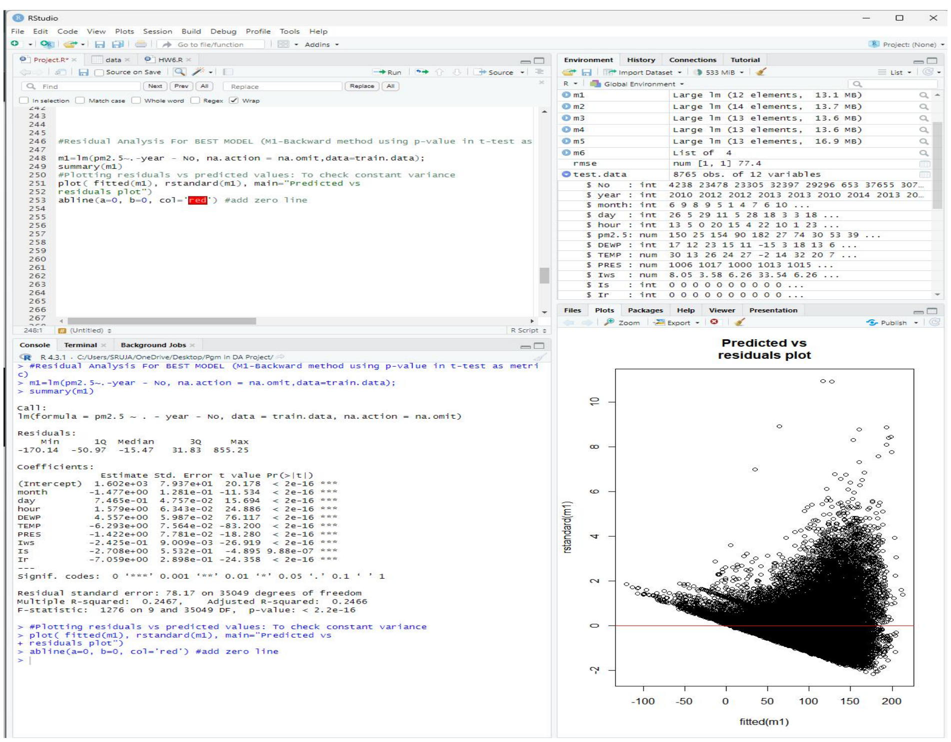
The Above Linear regression models are built in order to best fit the data in them and predict the unknown values of PM2.5.

The Best Model comes out to be Backward method using p-value in t-test as metric.

3.Post Processing

3.1 Best Model from Linear Regression
3.2 Model Diagnosis
Residual Analysis is considered for the following :

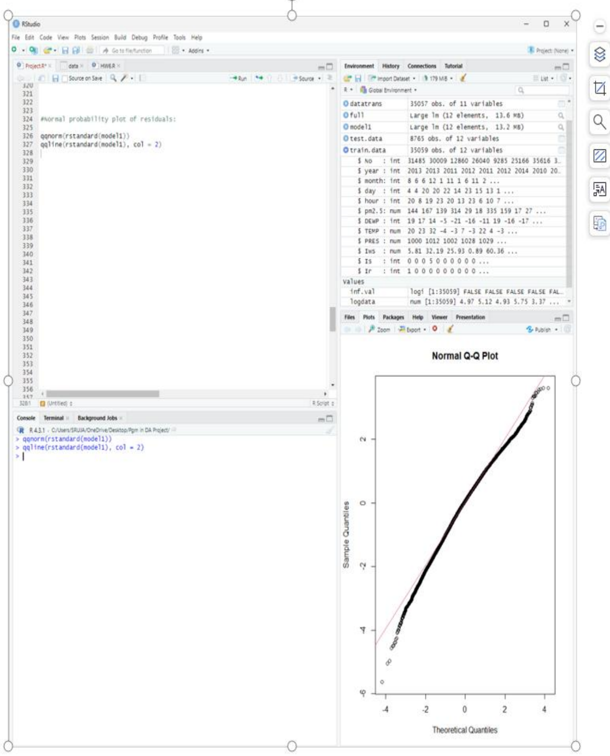
a) Plot residuals vs predicted values: To Validate the constant variance



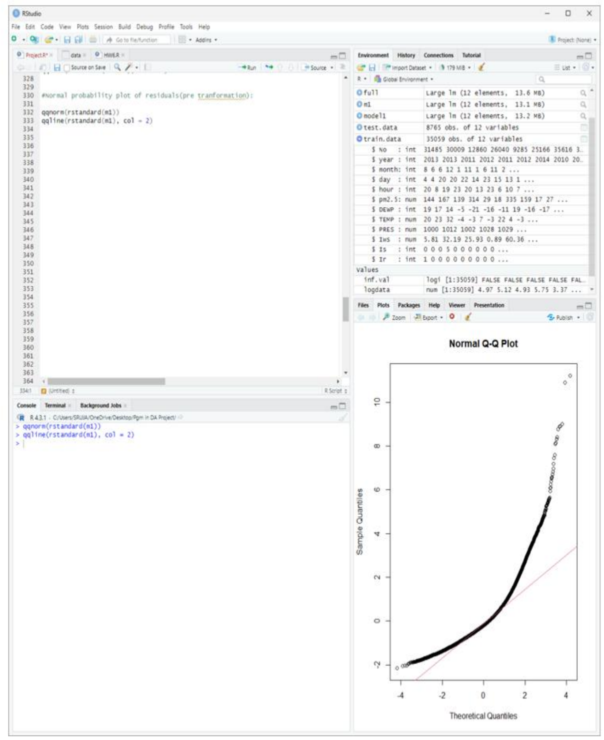
Transformation of PM2.5(Y variable):
Transformation is done as the graph predicts a problem(pattern)

b) Plot residuals vs each x-variable:To Validate the linearity relationship

c)Normal probability plot of residuals:To check normality assumption for the error terms



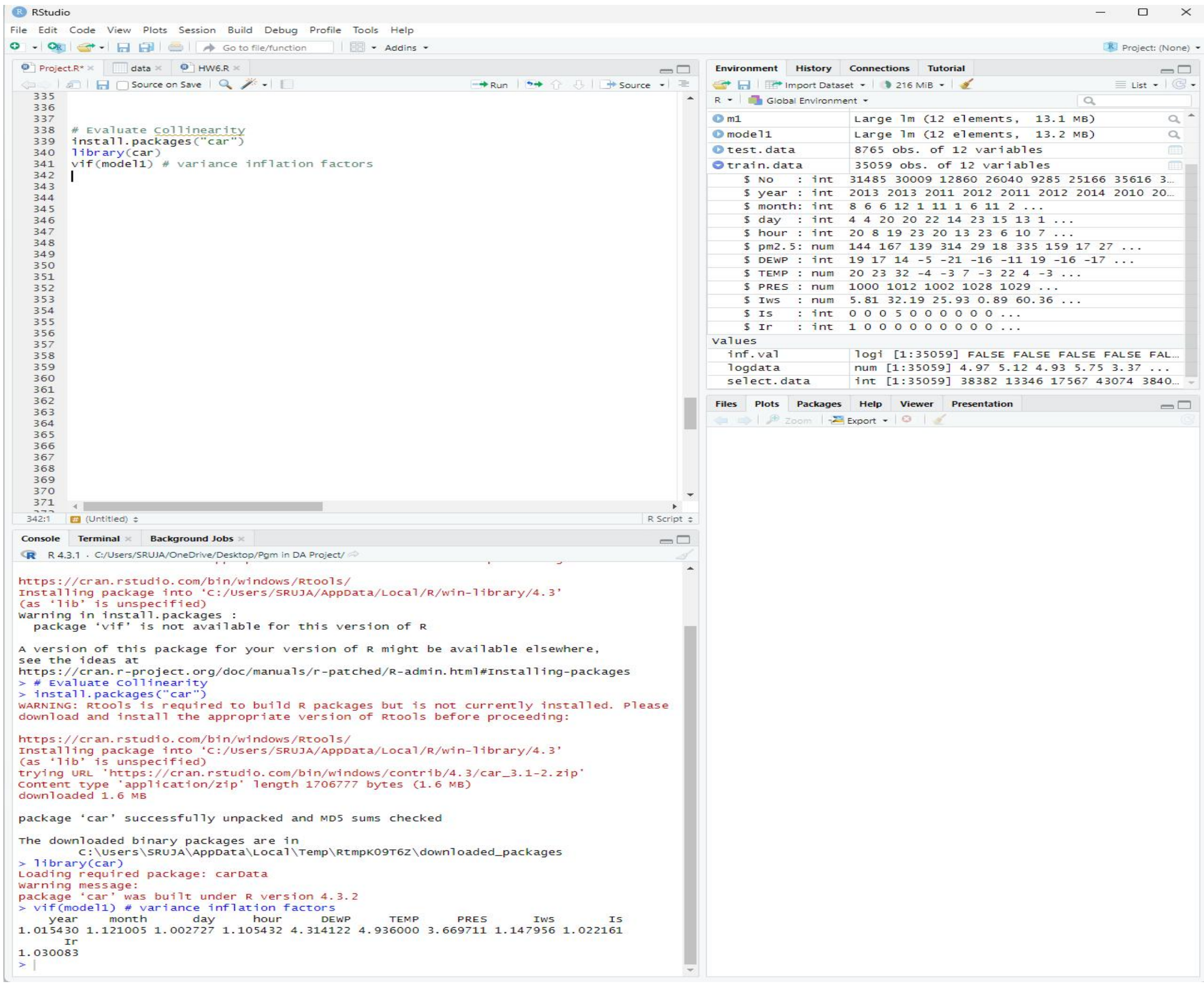
QQ plot after Tranforming Pm2.5 variable.



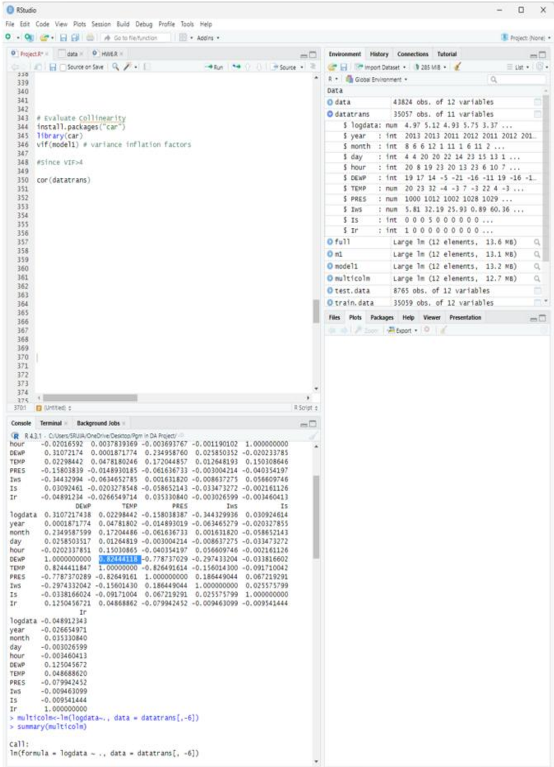
QQ plot **before** Tranforming Pm2.5 variable.

3.3 Improving Model

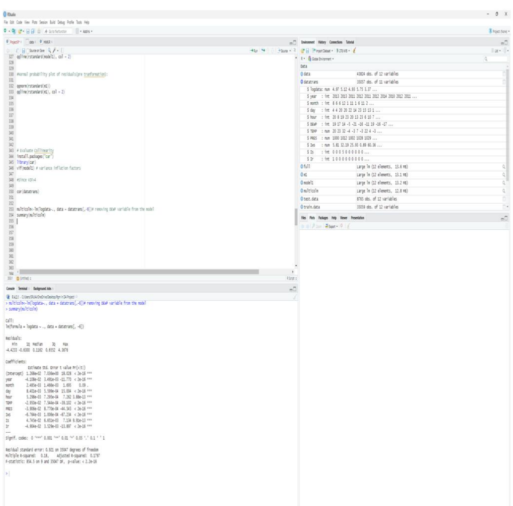
a)Multicollinearity problem
Computing Variance Inflation Factor Statistics



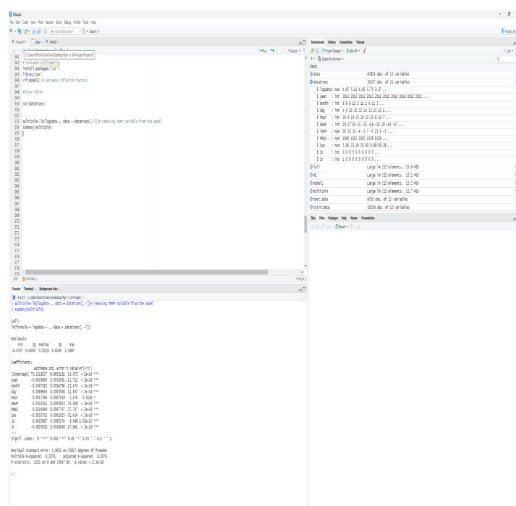
Since VIF > 4 for DEWP ,TEMP, we are testing corr to find pair of independent variables



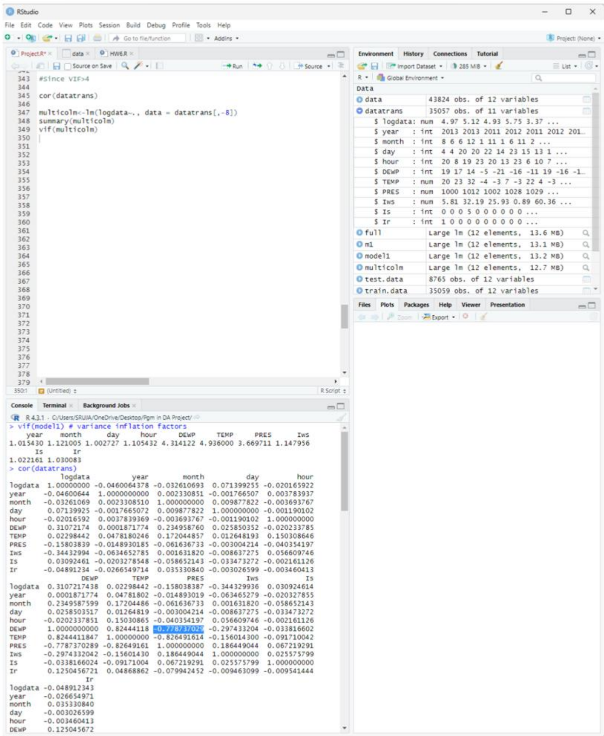
The Correaltion between DEWP and TEMP is **0.8244411847**



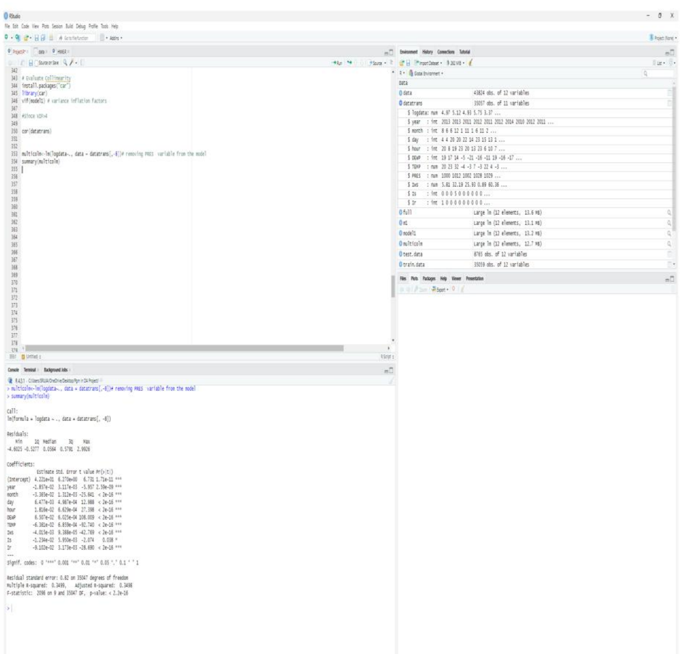
Option 1: After removing DEWP in the model AdjR2 Dropped to 0.1797



Option 2: After removing TEMPin the model AdjR2 Dropped to 0.1797



Option 3: Let us remove second highest correlated variable PRES with DEWP from the model



After removing PRES in the model
AdjR2 remained almost same i.e., 0.3498

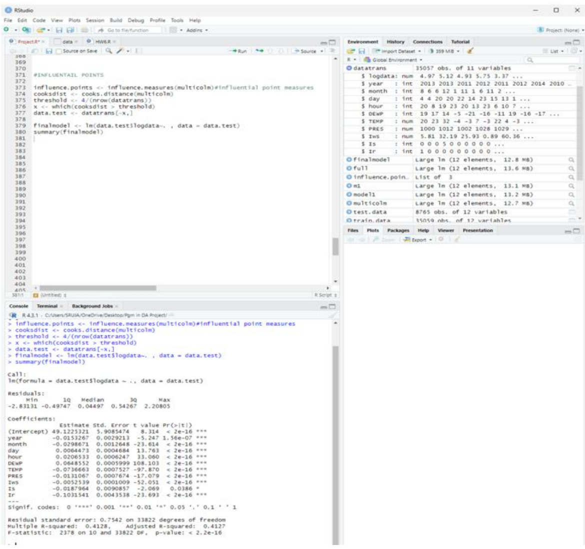
Conclusion :

As we know when two independent variables are strongly correlated there is no need to keep both of them in the model! They don't add predictive value to the model.
Therefore we can remove PRES variable from the model.

3.3IMPROVING MODEL

b)Influential points(removing)

Influential points are observations that significantly affect the fitted model; these are usually outliers. In the event that they are eliminated, the parameter estimations change

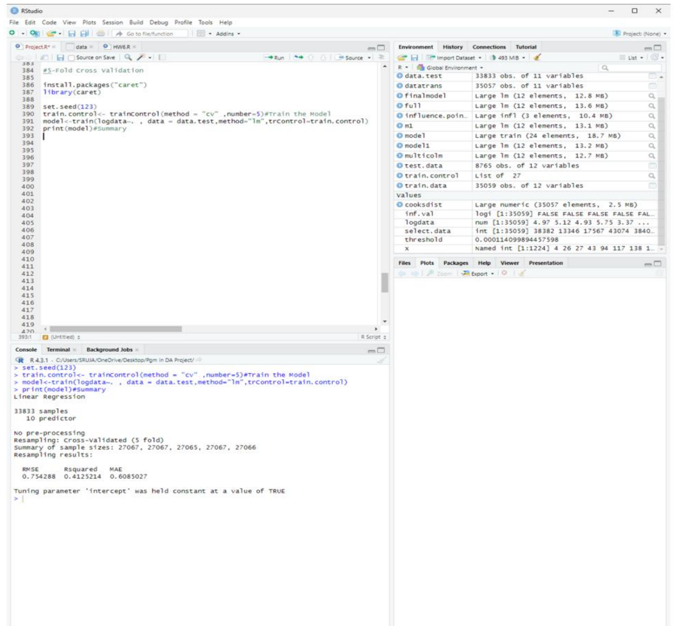


After removing infulential points the Adjusted R-squared increased to: 0.4127

EVALUATION STRATEGIES

5-Crossfold evaluation

RMSE -0.754288
RSquared - 0.4125214
MAE 0.6085027



5.2. Evaluations and Results

We concluded M2 is the best by looking at evaluation metrics RMSE and proceeded to improve the model further

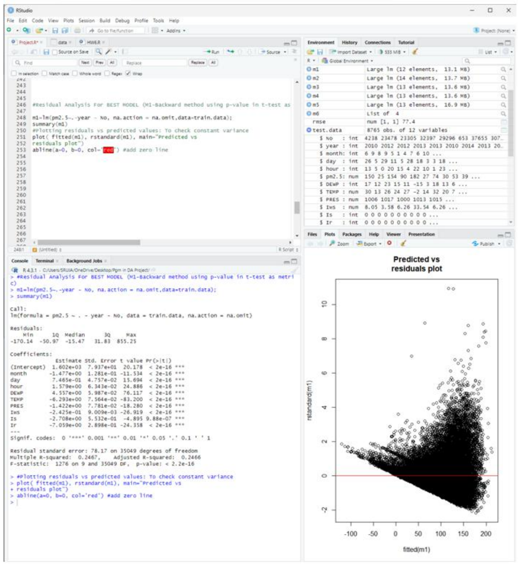
3. POST PROCESSING
3.1BEST MODEL
To Check Predictive performance we have conisdered Train.data in each model.

- Comparing RMSE of all the 5 Models.
- M1.Full Model- RMSE :77.36838
- M2**.Backward method using p-value in t-test as metric- RMSE :77.41664**
- M3.Backward method using AIC as metric: RMSE :77.36838
- M4.Forward method using AIC as metric: RMSE :77.36838
- M5.Stepwise method using ACI as metric: RMSE :77.36838

- Therefore M2 is Better Model compared to other models having better RMSE value.

Residual Analysis on the Best Model

a)Plot residuals vs predicted values:



We plotted residuals vs predicted values to validate the constant variance for the residuals

In the given screenshot we can confirm a **pattern** and tell there is a **PROBLEM TO FIX**

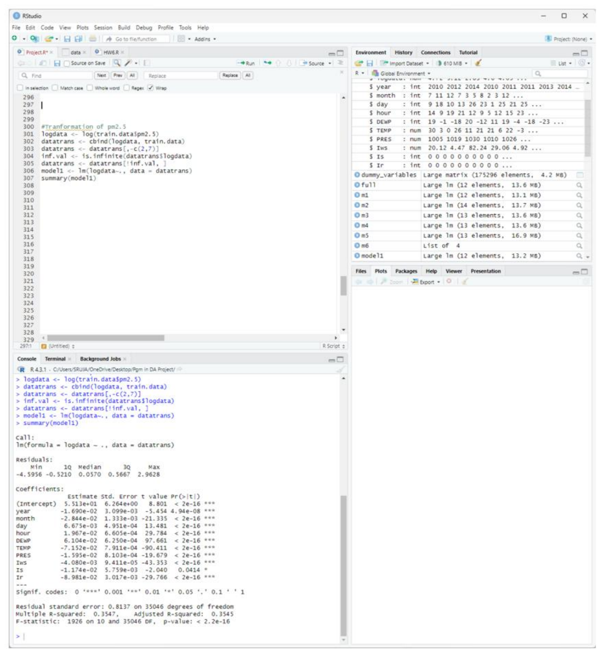
Solution : We need to apply a transformation on y, such as log transformation, and then re-fit the regression model

Transformation of PM2.5(Y variable)

Improved Adj R2 Value after Logarithmic transformation

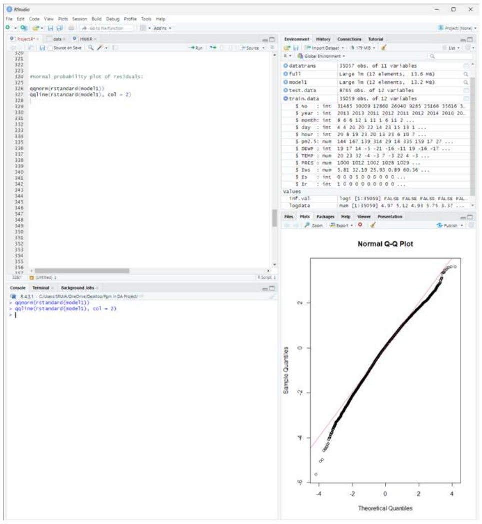
Adj R2 Previously : 0.246

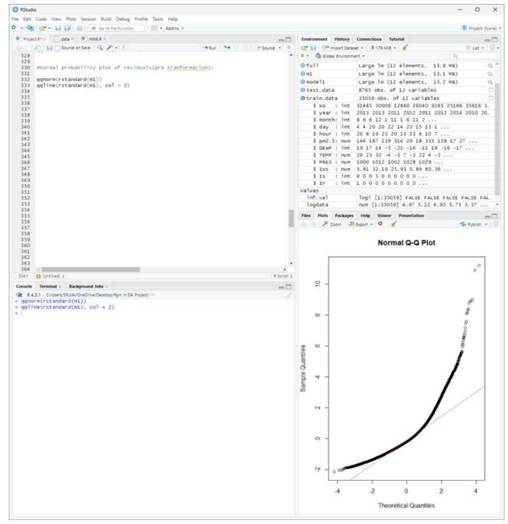
Adj R2 After Transformation : 0.3545



3.2MODEL DIAGNOSIS

c)Normal probability plot of residuals:To check normality assumption for the error terms





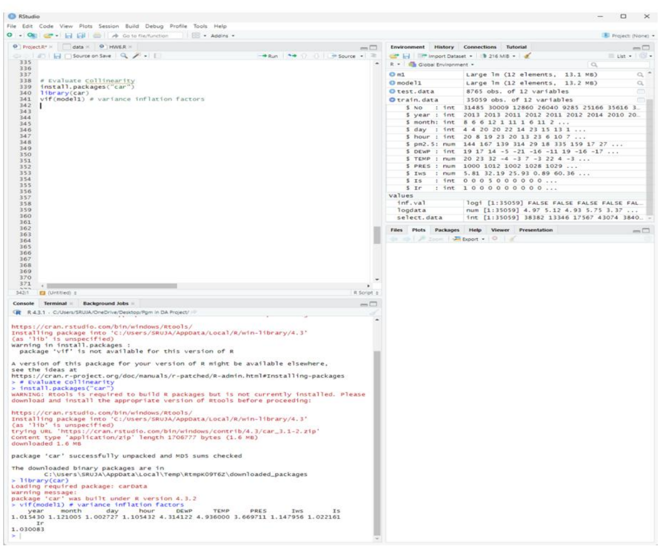
QQ plot after Tranforming Pm2.5 variable. QQ plot **before** Tranforming Pm2.5 variable.

Looking at the QQ Plot for after Transforming PM2.5 variable there is a straight line which means the data is normally distributed

While Improving the model

a)Multicollinearity problem

Computing Variance Inflation Factor Statistics



Since VIF > 4 for DEWP ,TEMP, we are testing corr to find pair of independent variables

5.3. Findings

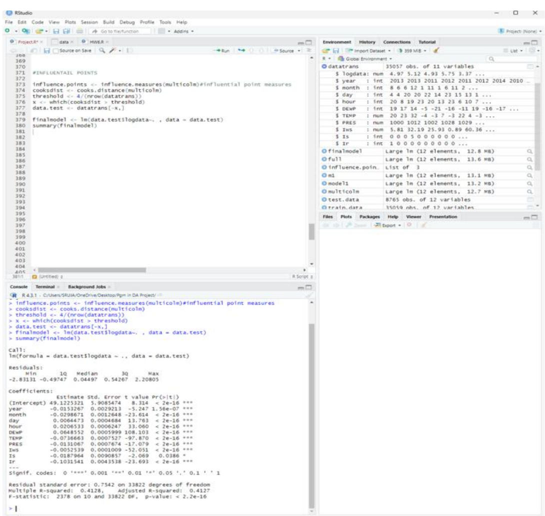
The best model we could offer for this data set

Applying 5 regression models

Selecting the best model using best RMSE value and Adj R2 value.

Doing Residual analysis and fixing problem by transforming Y variable which is PM2.5

Improving the model checking multi collinearity problem and dropping inferential points



After removing infulential points the Adjusted R-squared increased to: <u>0.4127</u>

ALL THE regression model gives below average accuracy even after transformation and Model improvements techniques like Multicollinearity,dropping Inferential points and dealing with missing values.

6. Conclusions and Future Work

6.1. Conclusions

We have performed multiple tests and analyzed the data set at various levels. We can conclude that although all the approaches or tests show relatively less accuracy, ALL THE regression model gives below average accuracy even after transformation and Model improvements techniques like Multicollinearity,dropping Inferential points and dealing with missing values. Also, the number of data points is very small, and hence, an increase in the data volume with respect to the number of rows may make the analysis easier and more meaningful with better accuracy.We might also learn and apply different approach to tackle such datasets.

6.2. Limitations

- We need to find out more approaches to improve the model.Domain knowledge is lacking
- Data Quality
- Model complexity
- Correlation of Data
- Feature Quality

6.3. Potential Improvements or Future Work

1.Data Pre Processing

- Better Data Cleaning can help
- To drop nominal variable
- Cross Validation strategies
- Continuous learning to best fit the data to use in the model

2. Linear regression model

Increase domain knowledge of various Linear regression model to implement it to the right data set

3. Post processing

Adapt enhanced Post processing techniques like better smoothing techniques,calibration etc,.