Using Linear Regression to Predict Daily Air Quality Index (AQI) of Chittagong Using Weather Parameters

CSE 837 : Machine Learning

Submitted by

Mahdee Hasan - BSSE 1010
Ahmed Ryan - BSSE 1011
Ashraful Gafur - BSSE 1021
Junaid Mansur Ifti - BSSE 1027

Submitted to

B M Mainul Hossain
Professor
Institute of Information Technology
University of Dhaka

Date of Submission: November 10, 2022

1. Introduction **2**

2. Dataset Preparation **3**

2.1 Data Source 3

2.2 Dataset Merging and Preprocessing 5

3. Methodology **6**

3.1 Linear Regression 6

3.2 Assumptions of Linear Regression 6

3.3 Analysis (step by step) 14

3.4 Data Transformations 23

4. Results **23**

4.1: Result Analysis after Transformation 23

4.2 Final Equation 25

5. Conclusion **25**

6. References **26**

1. Introduction

Weather is something all humans in the world constantly experience through their senses, at least while being outside. Extreme weather events have caused smaller scale population movements and intruded directly on historical events. And weather data is readily available in most cases Hence, we took interest in the weather data and tried to use it for predicting a feature which may be related to weather.

Air Quality Index (AQI) is a measure of how polluted the air currently is. In this era of global warming, AQI is one of the most discussed measures as it is directly related to the health of the general people. Keeping the importance in mind, we tried to predict the Air Quality Index (AQI) of Chittagong division from the features mainly related to weather.

Since we need to predict data, we have to go back to statistics. Linear regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression. If the goal is prediction, forecasting, or reducing error, linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables.

We have used linear regression prediction techniques to find a relationship between weather and AQI data. We planned to perform ordinary linear regression (OLS) on the raw dataset and later on the preprocessed dataset so that we can get better results.

In the later sections of the report, detailed description has been provided on how we predicted AQI from weather and climate features. It also explains the relationship we found among them.

2. Dataset Preparation

For our project, we incorporated two datasets from two distinct sources. We first tried to incorporate weather parameters with agricultural data from Chittagong division, but there was insufficient dataset which made it not possible to do so. Then we tried to incorporate armed crime violence with the weather dataset parameters though it seems a different genre from the weather dataset but we thought if there had been any relationship. But again the data we got was insufficient. Finally, we thought what if AQI had any relationship with the weather parameters. So we searched for it and finally got an archived dataset which was used in an US embassy research and taken from CASE (Clean Air and Sustainable Environment) Bangladesh.

Firstly we collected a weather dataset of the Chittagong division from Bangladesh. This information contained several weather parameters of the Chittagong division and rainfall from 1 January 2012 to 31 December 2017. The dataset was referred to as“data.gov.bd” as source and at the time of doing this project the govt website was down for some intrusion activity. So we collected the dataset from GitHub which was used by the author in rainfall prediction projects. The dataset consisted of 2191 instances that had been collected from the Chittagong division of the country. It has 20 columns containing year, month, day, tempLow, tempAvg, tempHigh, DPLow, DPAvg, DPHigh, humidityLow, humidityAvg, humidityHigh, SLPLow, SLPAvg, SLPHigh, visibilityLow, visibilityAvg, visibilityHigh, windAvg, rainfall. Here, DP = Dew Point , SLP = Sea Level Point.

Then we collected an Air Quality Index (AQI) dataset all around Bangladesh based on Division. This information contains several weather parameters of the Chittagong division and rainfall from 3 January 2014 to 31 December 2021. But there are some dates missing in the datasets. This dataset was collected from CASE BD and is stored on the GitHub website. The dataset consists of 24044 instances that had been collected from all the divisions of the country. It has 6 columns containing Date, Location, AQI, Category, Range, and CAMS tag. In this dataset, our point of interest is the AQI values of the Chittagong division in Bangladesh.

2.1 Data Source

1. Weather Data: taken from archived resource of https://live.bmd.gov.bd/ . csv taken from github.

	year	month	day	tempHigh	tempAvg	tempLow	DPHigh	DPAvg	DPLow	humidityHigh	humidityAvg	humidityLow	SLPHigh	SLPAvg	SLPLow	visibilityHigh	visibilityAvg	visibilityLow	windAvg	Ra
2	2012	1	1	28	23	19	18	14	10	88	57	34	1015	1012	1010	6	5	4	5	0
3	2012	1	2	26	22	18	17	15	13	88	65	47	1015	1013	1012	4	3	2	2	0
4	2012	1	3	27	22	17	20	17	16	94	71	54	1014	1012	1010	5	4	4	6	0
5	2012	1	4	26	23	20	18	17	15	83	66	51	1015	1013	1010	4	3	1	3	0
6	2012	1	5	26	23	19	18	17	16	88	71	57	1016	1014	1012	4	3	1	3	0
7	2012	1	6	27	22	18	19	17	15	94	75	48	1016	1015	1013	4	3	2	2	0
8	2012	1	7	26	22	18	18	17	15	94	73	51	1018	1016	1015	4	2	1	2	0
9	2012	1	8	26	22	18	19	18	16	94	79	61	1018	1016	1014	4	3	2	3	0
10	2012	1	9	26	22	18	18	18	16	94	75	54	1015	1014	1013	4	3	0	8	0

Figure: Snippet of weather data

2. AQI Data (scrapped from archived daily report of http://case.doe.gov.bd/) csv taken from github.

Department of Environment
E-16, Agargaon, Sher-e- Bangla Nagar, Dhaka-1207

Daily Air Quality Index (AQI) Report

Date: 09/11/2022

LOCATION	AQI	CATEGORY	RANGE
Dhaka[b]	116	CAUTION	
Chittagong[b]	144	CAUTION	
Gazipur[c]	114	CAUTION	
Narayanganj[c]	DNA	DNA	
Sylhet[c]	DNA	DNA	
Khulna[c]	158	UNHEALTHY	
Rajshahi[c]	168	UNHEALTHY	
Barisal[c]	148	UNHEALTHY	
Savar[c]	163	UNHEALTHY	
Mymensingh[c]	168	UNHEALTHY	
Rangpur[c]	167	UNHEALTHY	
Cumilla[c]	120	CAUTION	
Narsingdi[c]	163	UNHEALTHY	

Figure: Daily Report of AQI from CASE BD

2.2 Dataset Merging and Preprocessing

We used linear regression to predict the AQI based on weather parameters of Chittagong so we only selected the Chittagong division AQI values from the AQI dataset. The AQI datasets had some errors in the date column as they didn't maintain proper date formatting. So we process those data into a formal date format. After that, we created three new columns: year, month, and day using the old date column and removed the old date column from the AQI dataset. We needed those three columns to map the weather datasets to evaluate the AQI on the same date. After processing we took those AQI values which are between 3 Jan 2014 to 31 Dec 2017 as our weather dataset ends in 2017. After that, we had 1283 instances and four columns. We deleted the empty AQI values and got 1275 instances finally.

On the other hand, we selected data between 3 Jan 2014 to 31 Dec 2017 from the weather dataset to match with the AQI dataset. After that, we have 1459 instances. The reason for the difference between instances between these two datasets was because of the AQI dataset. There are some dates of AQI missing. Also, some data have no AQI data.

Then we merged the two datasets into one final dataset based on the day, month and year. After merging those two datasets we had 1459 data. But as the AQI dataset had less value, we removed those data from the merged dataset which didn't have AQI value. Finally, we had 1275 instances and 21 columns. This is how, after a long and thorough process of preprocessing, we got the final merged dataset of "Weather Parameter and AQI of Chittagong division from 2014 to 2017".

Our final dataset link:
https://github.com/bsse1027/Weather-and-AQI-dataset-of-Chittagong/blob/main/final_data_set.csv

3. Methodology

In this chapter we will talk about the processes or methodologies we used for our project.

3.1 Linear Regression

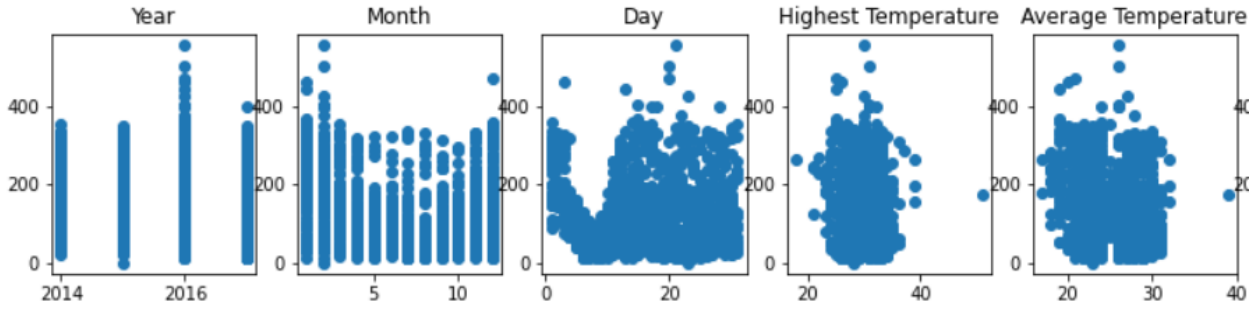
We used linear regression to predict AQI value based on weather parameters. After preprocessing we created a preliminary model using the final dataset. We measured the model R Squared and Adjusted R Square value and other model-related values.

3.2 Assumptions of Linear Regression

We have tested the following assumptions -

a. Linearity
b. Multicollinearity
c. Normality of Residual
d. Homoscedasticity and
e. Autocorrelation of Error

Linearity is a measure of the maximum deviation of any reading from a straight line calibration line. If we want to fit a model with linear regression if we must ensure that the features maintain linearity.
We initially checked the linearity of all 20 features of the dataset. The resultant plots are provided below.



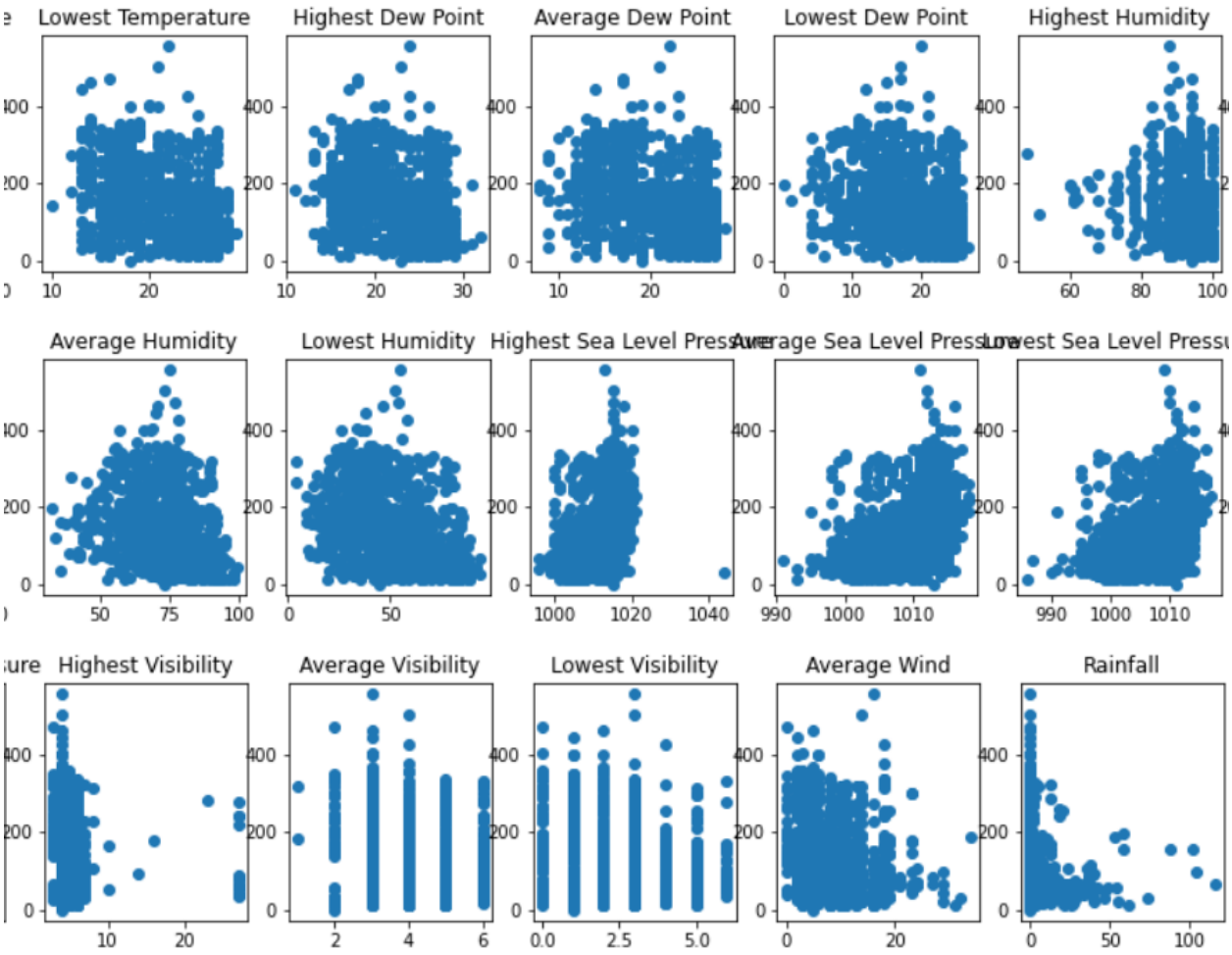


Figure: Linearity Check on Raw Data

After checking the linearity of raw data, we found that the results were not satisfactory. So, we performed the following transformations on all the features.

a. Inverse
b. Exponential
c. Logarithm

Some transformations are illustrated below -

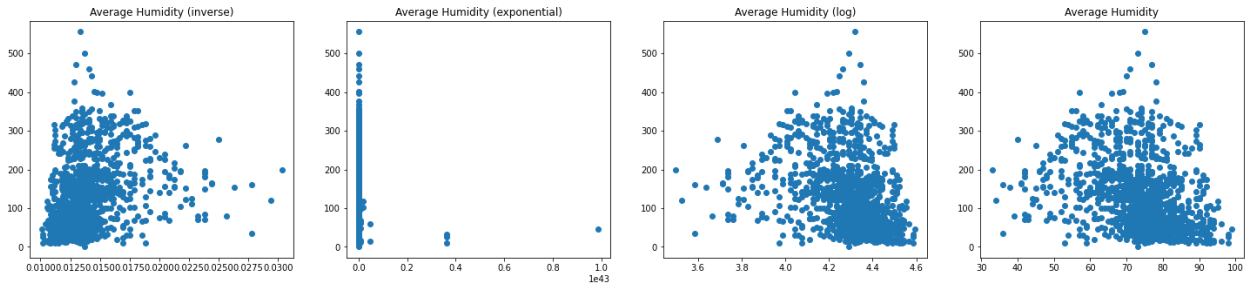


Figure: Average Humidity Transformations

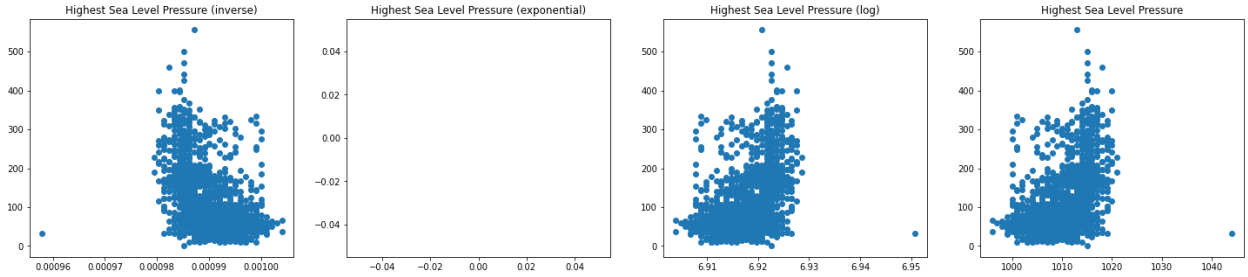


Figure: Highest Sea Level Pressure Transformations

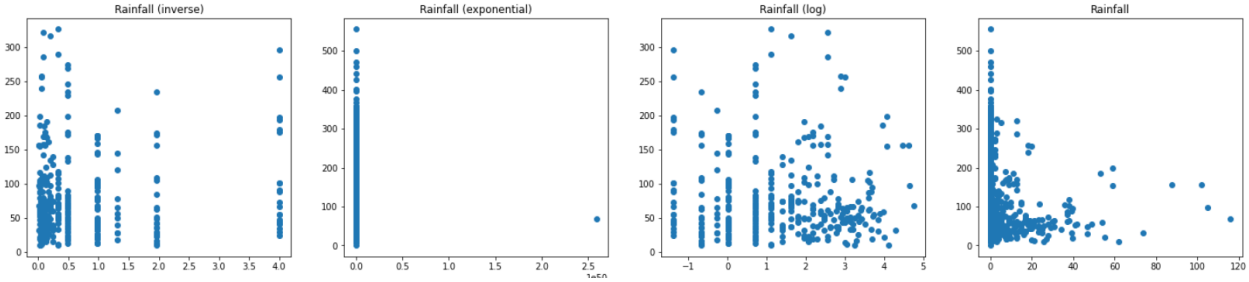


Figure: Rainfall Transformations

Multicollinearity is a statistical concept where several independent variables in a model are correlated. Before fitting the model with linear regression, we must ensure that no multicollinearity exists between the features. We find out the variance inflation factor of the 20 features. In an ideal case, a VIF value > 5 implies that multicollinearity exists in the system. The VIF values of our features have been mentioned below.

Feature	Variance Inflation Factor (VIF)
year	1.435312e+05
month	6.246926e+00
day	4.477352e+00
tempHigh	2.428866e+03
tempAvg	5.316877e+03
tempLow	1.322532e+03
DPHigh	6.866738e+02
DPAvg	1.538626e+03
DPLow	3.298488e+02
humidityHigh	7.151311e+02
humidityAvg	9.598228e+02
humidityLow	1.353104e+02
SLPHigh	6.114305e+05
SLPAvg	2.340801e+06
SLPLow	1.442319e+06
visibilityHigh	7.219848e+00
visibilityAvg	7.317831e+01
visibilityLow	1.312779e+01
windAvg	5.836690e+00
Rainfall	1.343543e+00

Furthermore, a correlation heatmap has been used to get an overall understanding of the relationship among all the features. The heatmap has been attached below.

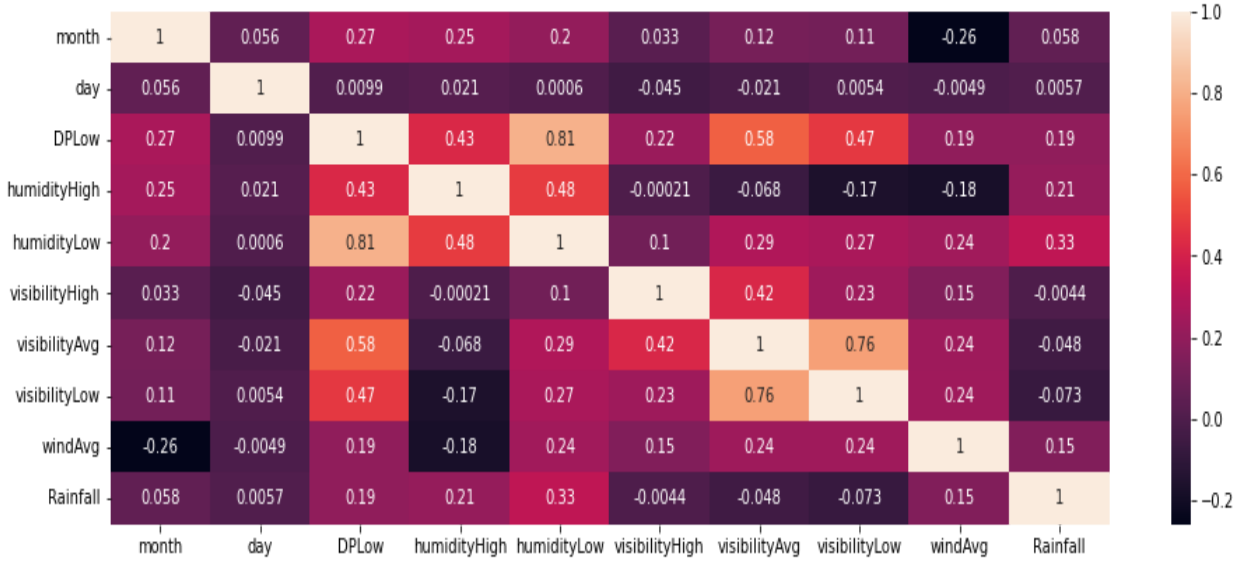


Figure: Correlation Heatmap

Normality of Residual is a measure of checking if the residuals are normally distributed. If the residuals are not normally distributed, then model inference (i.e. model predictions and confidence intervals) might be invalid. Therefore, it is crucial to check this assumption. We checked the normality of residuals of all the features and it provided a pretty decent output. The graph is provided below -

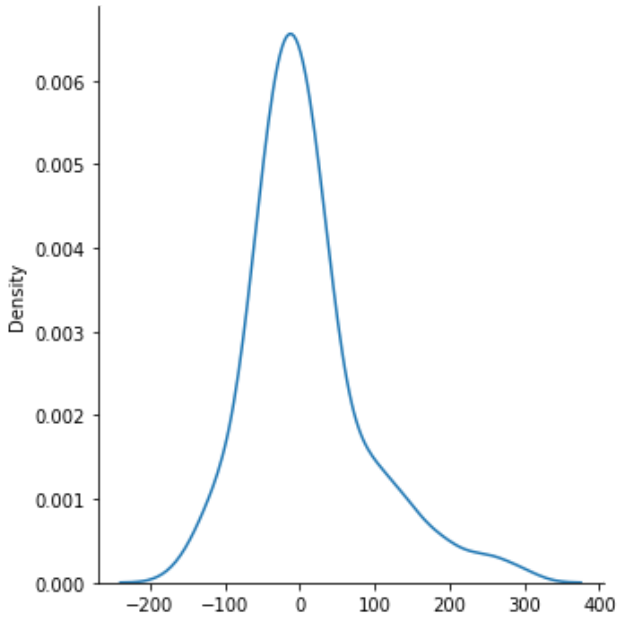


Figure: Normality of Residuals

Furthermore, we have used a Q-Q plot which also implies that the residuals are normally distributed. The graph of the Q-Q plot has been provided below.

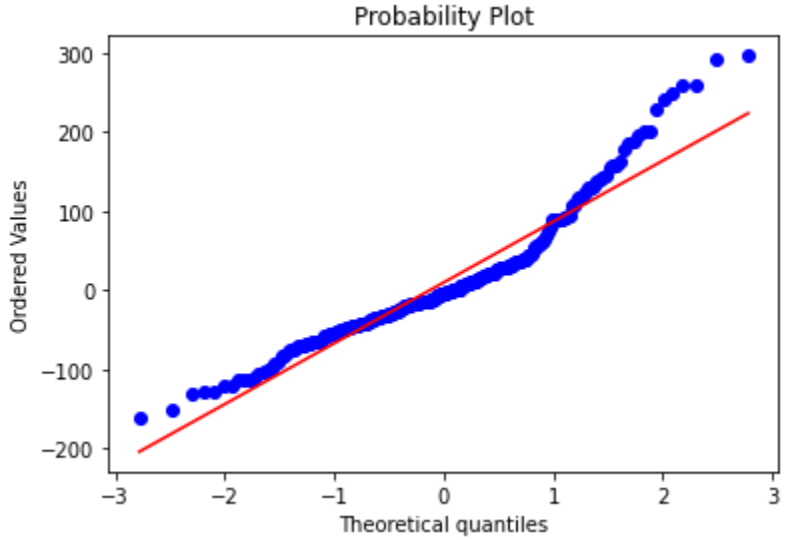


Figure: Q-Q Plot

Homoscedasticity means to be of “The same Variance”. In Linear Regression, one of the main assumptions is that there is a Homoscedasticity present in the errors or the residual terms (Y_Pred – Y_actual). In other words, Linear Regression assumes that for all the instances, the error terms will be the same and of very little variance.

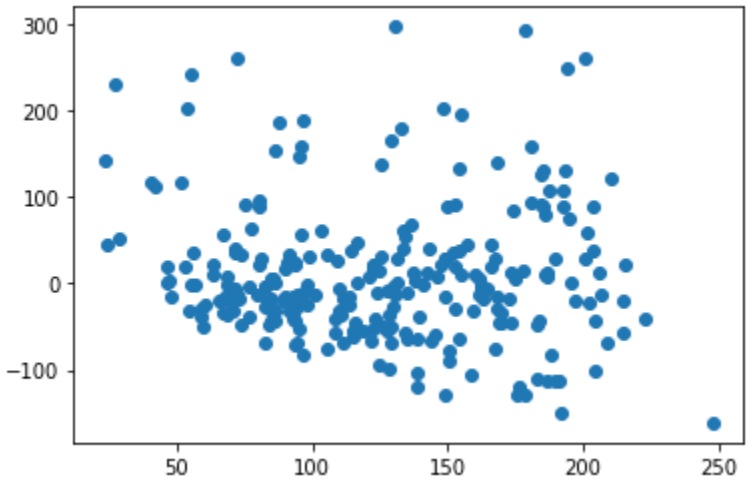


Figure: Scatter Plot

From the above scatter plot, we can imply that the raw dataset maintains homoscedasticity to a certain extent. Further transformation of data will lead to better output which is explained in the later part of the report.

Autocorrelation in the model means that the error terms are correlated. For example, previous trading days (e.g. the stock price is more prone to fall after a huge price hike) influence the current stock price. In the cross-section data, the neighboring units tend to have similar characteristics.

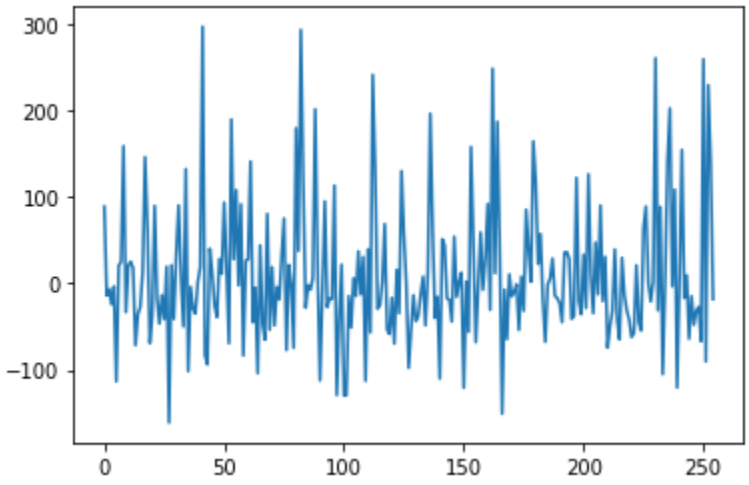


Figure: Scatter Plot

The above plot can be translated in the following way:
More hikes imply that there is lesser autocorrelation of errors.
Fewer hikes imply that there is more autocorrelation of errors.

So, from the above plot, we can conclude that there is almost no autocorrelation of error in the current dataset.

3.3 Analysis (step by step)

From the assumptions section we saw that our dataset has a very high collinearity between features which is a major setback. So, through variance inflation factor(VIF) we iteratively dropped columns which had the highest VIF score and every time ran the OLS regression on the newly modified dataset. We ran that until all of the columns remaining had a VIF score nearly 5. But there should also be a concern that the R square value should not drastically fall as soon as we drop columns. So, having the R square value almost the same or better, we tried to drop as much as the column we can so that we can diminish the multicollinearity impact. And as we had 20 columns as features so the iteration was quite long.

First Iteration:

First, we ran OLS regression for the raw merged dataset having all the features intact. Then ran VIF and analyzed which variable had the highest factor. We dropped that variable from the features in the next iteration and ran the process again.

OLS: For original dataset

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.336
Model:	OLS	**Adj. R-squared:**	0.323
Method:	Least Squares	**F-statistic:**	25.26
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	5.55e-75
Time:	09:51:47	**Log-Likelihood:**	-5827.9
No. Observations:	1020	**AIC:**	1.170e+04
Df Residuals:	999	**BIC:**	1.180e+04
Df Model:	20		
Covariance Type:	nonrobust		

VIF_1 (original dataset):

12	SLPHigh	6.114305e+05
13	SLPAvg	2.340801e+06
14	SLPLow	1.442319e+06
15	visibilityHigh	7.219848e+00

2nd Iteration:

Deducting SLPAvg from dataset :

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.335
Model:	OLS	**Adj. R-squared:**	0.322
Method:	Least Squares	**F-statistic:**	26.46
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	2.62e-75
Time:	09:56:33	**Log-Likelihood:**	-5828.9
No. Observations:	1020	**AIC:**	1.170e+04
Df Residuals:	1000	**BIC:**	1.180e+04
Df Model:	19		
Covariance Type:	nonrobust		

VIF_2 (Deducting SLPAvg):

7	DPAvg	1537.004603
8	DPLow	329.846800
9	humidityHigh	715.129910
10	humidityAvg	957.933163
11	humidityLow	135.205048
12	SLPHigh	398626.192933
13	SLPLow	426930.072202
14	visibilityHigh	7.217098

3rd Iteration:

Deducting SLPLow from dataset :

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.
  x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.330
Model:	OLS	**Adj. R-squared:**	0.318
Method:	Least Squares	**F-statistic:**	27.37
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	1.52e-74
Time:	11:35:18	**Log-Likelihood:**	-5832.5
No. Observations:	1020	**AIC:**	1.170e+04
Df Residuals:	1001	**BIC:**	1.180e+04
Df Model:	18		
Covariance Type:	nonrobust		

VIF_3 (Deducting SLPLow):

123464.84544380633

	feature	VIF
0	year	123464.845444
1	month	6.234526
2	day	4.469796
3	tempHigh	2424.961288
4	tempAvg	5316.876599
5	tempLow	1321.268630
6	DPHigh	684.177315
7	DPAvg	1536.817753
8	DPLow	329.845975

4th Iteration:

Deducting Year from dataset :

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.308
Model:	OLS	**Adj. R-squared:**	0.297
Method:	Least Squares	**F-statistic:**	26.28
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	1.20e-68
Time:	11:36:53	**Log-Likelihood:**	-5848.6
No. Observations:	1020	**AIC:**	1.173e+04
Df Residuals:	1002	**BIC:**	1.182e+04
Df Model:	17		
Covariance Type:	nonrobust		

VIF_4 (Deducting Year):

5287.933177713449

	feature	VIF
0	month	6.229309
1	day	4.469191
2	tempHigh	2366.593229
3	tempAvg	5287.933178
4	tempLow	1297.559456
5	DPHigh	684.106880

5th Iteration:

Deducting tempAvg from dataset :

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.308
Model:	OLS	**Adj. R-squared:**	0.297
Method:	Least Squares	**F-statistic:**	27.85
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	3.84e-69
Time:	11:38:12	**Log-Likelihood:**	-5849.2
No. Observations:	1020	**AIC:**	1.173e+04
Df Residuals:	1003	**BIC:**	1.182e+04
Df Model:	16		
Covariance Type:	nonrobust		

VIF_5 (Deducting tempAvg):

1532.8668258217447

	feature	VIF
0	month	6.224606
1	day	4.469054
2	tempHigh	881.912722
3	tempLow	346.352310
4	DPHigh	681.539252
5	DPAvg	1532.866826
6	DPLow	328.340129

6th Iteration:

Deducting DPAvg from dataset :

```
x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.306
Model:	OLS	**Adj. R-squared:**	0.295
Method:	Least Squares	**F-statistic:**	29.46
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	2.78e-69
Time:	11:39:29	**Log-Likelihood:**	-5850.6
No. Observations:	1020	**AIC:**	1.173e+04
Df Residuals:	1004	**BIC:**	1.181e+04
Df Model:	15		
Covariance Type:	nonrobust		

VIF_6 (Deducting DPAvg):

989.7439551328116

	feature	VIF
0	month	6.175010
1	day	4.466216
2	tempHigh	796.669568
3	tempLow	313.873958
4	DPHigh	470.308703
5	DPLow	203.237094
6	humidityHigh	709.723702
7	humidityAvg	826.817052
8	humidityLow	132.949229
9	SLPHigh	989.743955

7th Iteration:

Deducting SLPHigh from dataset :

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatool
  x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.299
Model:	OLS	**Adj. R-squared:**	0.290
Method:	Least Squares	**F-statistic:**	30.66
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	4.10e-68
Time:	11:40:39	**Log-Likelihood:**	-5855.2
No. Observations:	1020	**AIC:**	1.174e+04
Df Residuals:	1005	**BIC:**	1.181e+04
Df Model:	14		
Covariance Type:	nonrobust		

VIF_7 (Deducting SLPHigh):

746.8970566185434

	feature	VIF
0	month	6.171189
1	day	4.411282
2	tempHigh	384.355211
3	tempLow	311.837449
4	DPHigh	410.828796
5	DPLow	140.935091
6	humidityHigh	552.368212
7	humidityAvg	746.897057

8th Iteration:

Deducting humidityAvg from dataset :

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools.py:1
  x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.298
Model:	OLS	**Adj. R-squared:**	0.289
Method:	Least Squares	**F-statistic:**	32.83
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	1.93e-68
Time:	11:41:46	**Log-Likelihood:**	-5856.3
No. Observations:	1020	**AIC:**	1.174e+04
Df Residuals:	1006	**BIC:**	1.181e+04
Df Model:	13		
Covariance Type:	nonrobust		

VIF_8 (Deducting humidityAvg):

377.97412915989497

	feature	VIF
0	month	5.993183
1	day	4.410826
2	tempHigh	373.278581
3	tempLow	307.926237
4	DPHigh	377.974129
5	DPLow	139.507037

9th Iteration:

Deducting DPHigh from dataset :

```
x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.298
Model:	OLS	**Adj. R-squared:**	0.289
Method:	Least Squares	**F-statistic:**	35.55
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	3.83e-69
Time:	11:42:38	**Log-Likelihood:**	-5856.5
No. Observations:	1020	**AIC:**	1.174e+04
Df Residuals:	1007	**BIC:**	1.180e+04
Df Model:	12		
Covariance Type:	nonrobust		

VIF_9 (Deducting DPHigh):

351.8468249266671

	feature	VIF
0	month	5.931117
1	day	4.401971
2	tempHigh	351.846825
3	tempLow	259.676512
4	DPLow	117.447706

10th Iteration:

Deducting TempHigh from dataset :

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatool
  x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.297
Model:	OLS	**Adj. R-squared:**	0.289
Method:	Least Squares	**F-statistic:**	38.66
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	1.10e-69
Time:	11:43:36	**Log-Likelihood:**	-5857.1
No. Observations:	1020	**AIC:**	1.174e+04
Df Residuals:	1008	**BIC:**	1.180e+04
Df Model:	11		

VIF_10 (Deducting TempHigh):

167.21620846463867

	feature	VIF
0	month	5.839095
1	day	4.401968
2	tempLow	167.216208
3	DPLow	117.048004

11th Iteration:

Deducting TempLow from dataset :

```
x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.295
Model:	OLS	**Adj. R-squared:**	0.288
Method:	Least Squares	**F-statistic:**	42.27
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	4.66e-70
Time:	11:44:29	**Log-Likelihood:**	-5858.2
No. Observations:	1020	**AIC:**	1.174e+04
Df Residuals:	1009	**BIC:**	1.179e+04
Df Model:	10		
Covariance Type:	nonrobust		

VIF_11 (Deducting TempLow):

	feature	VIF
0	month	5.836677
1	day	4.401148
2	DPLow	66.213781
3	humidityHigh	35.916466
4	humidityLow	43.578842
5	visibilityHigh	7.172449
6	visibilityAvg	62.202414
7	visibilityLow	12.528385
8	windAvg	4.847437
9	Rainfall	1.275978

11th Iteration:

Deducting DPLow from dataset :

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tsa/tsatools
  x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.291
Model:	OLS	**Adj. R-squared:**	0.285
Method:	Least Squares	**F-statistic:**	46.13
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	1.05e-69
Time:	11:45:29	**Log-Likelihood:**	-5861.0
No. Observations:	1020	**AIC:**	1.174e+04
Df Residuals:	1010	**BIC:**	1.179e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

VIF_12 (Deducting DPLow):

	feature	VIF
0	month	5.761638
1	day	4.401126
2	humidityHigh	35.794547
3	humidityLow	16.991388
4	visibilityHigh	7.170747
5	visibilityAvg	51.258435
6	visibilityLow	12.521603
7	windAvg	4.785185
8	Rainfall	1.274488

12th Iteration:

Deducting visibilityAvg from dataset :

```
x = pd.concat(x[::order], 1)
```

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.265
Model:	OLS	**Adj. R-squared:**	0.259
Method:	Least Squares	**F-statistic:**	45.58
Date:	Mon, 07 Nov 2022	**Prob (F-statistic):**	1.01e-62
Time:	11:46:49	**Log-Likelihood:**	-5879.6
No. Observations:	1020	**AIC:**	1.178e+04
Df Residuals:	1011	**BIC:**	1.182e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

After dropping “visibilityAvg” from the dataset and running OLS regression we saw the R square value drastically drop by nearly 15% from 0.30-ish value. So we stopped dropping more columns. We tested that dropping more columns based on VIF reduced the R square value more. So we took the 9 features we got after the 11th iteration and we then tried to do different transformations on them.

3.4 Data Transformations

We then tried different combinations of below transformations on our final deducted dataset.

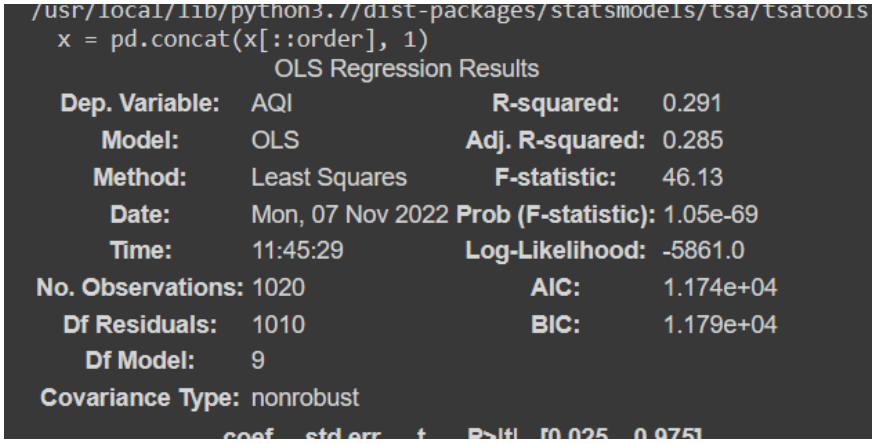
- Natural Logarithm
- Exponential
- Square Root
- Cubic Root
- Inverse

4. Results

4.1: Result Analysis after Transformation

So, before transformation on the final dataset we had the result followingly:

Before Transformation:



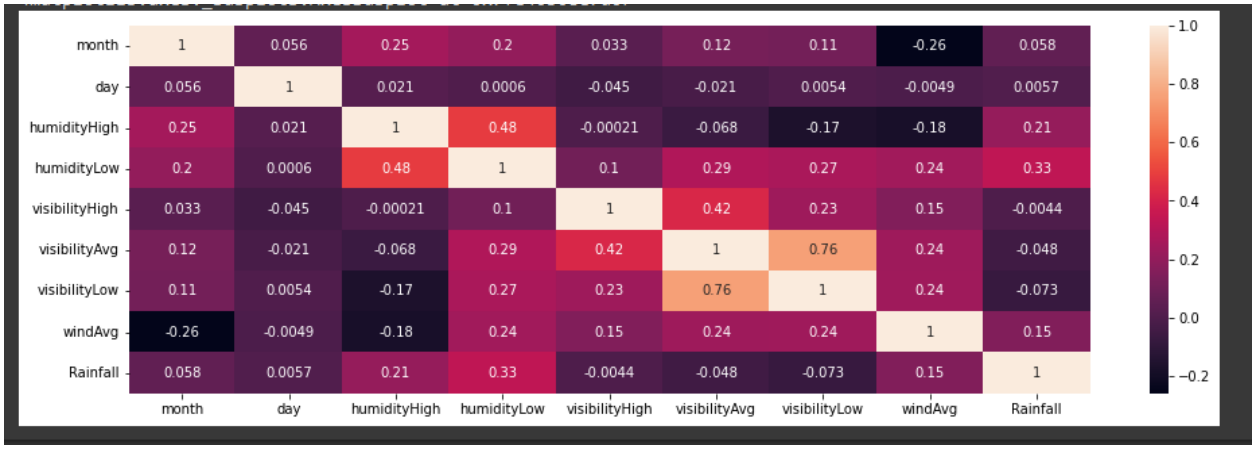
After running various combinations of transformation and running regression everytime, we tried to see which combination gave us the best R square value. Then we found out that after log transformation of the non zero columns and then square rooting the whole 8 features gave us the best R square result which is a nearly 9% increase from the raw result.

After Transformation:

OLS Regression Results

Dep. Variable:	AQI	**R-squared:**	0.319
Model:	OLS	**Adj. R-squared:**	0.313
Method:	Least Squares	**F-statistic:**	52.67
Date:	Thu, 10 Nov 2022	**Prob (F-statistic):**	1.89e-78
Time:	15:17:56	**Log-Likelihood:**	-5840.4
No. Observations:	1020	**AIC:**	1.170e+04
Df Residuals:	1010	**BIC:**	1.175e+04
Df Model:	9		
Covariance Type:	nonrobust		

So, finally the MultiColinearity Correlation heatmap after feature selection and transformation :



4.2 Final Equation

Linear Regression Equation for our model:

$$AQI = \beta_0 + \beta_1\sqrt{log(month)} + \beta_2\sqrt{log(day)} + \beta_3\sqrt{log(humidityHigh)} + \beta_4\sqrt{log(humidityLow)} + \beta_5\sqrt{log(visibilityHigh)} + \beta_6\sqrt{log(visibilityAvg)} + \beta_7\sqrt{log(visibilityLow)} + \beta_8\sqrt{log(windAvg)} + \beta_9\sqrt{log(Rainfall)}$$

Coefficients: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9$

After performing the linear regression the final equation was:

$$AQI = 1348.23 - 25.15\sqrt{log(month)} - 44.74\sqrt{log(day)} - 231.06\sqrt{log(humidityHigh)} - 172.68\sqrt{log(humidityLow)} - 39.58\sqrt{log(visibilityHigh)} - 148.58\sqrt{log(visibilityAvg)} - 10.60\sqrt{log(visibilityLow)} - 14.24\sqrt{log(windAvg)} - 6.09\sqrt{log(Rainfall)}$$

RMSE value is: 70. 04 and Normalized RMSE value is : 0.126

5. Conclusion

We had our weather dataset of Chittagong which was used for rainfall prediction. But we thought about how we can incorporate the feature variables as well as the rainfall prediction variable to infer something else. We already thought about what can be connected with weather variables and then AQI seemed a very interesting thing to be connected to weather eventually. As we merged our dataset and ran linear regression on this, we didn't get much promising result. Then we tried to do feature selection through many iterations and then did various transformations on the variables. But we saw that in the end, the linearity between the weather variables and AQI was not so evident. Although there is some impact of the weather parameters on the AQI value, we cannot seal complete causation or strong correlation between AQI and other weather parameters. This can also be a reason that the data on which we worked were real life raw data and linear regression is sometimes naive to be fit on real life scattered data. Any other prediction model other than linear regression might predict AQI based on the weather parameter more accurately. But, as a base, as there is not much work of

weather parameters and AQI relationship especially in the context of Bangladesh, this can be a good starting point for further research. Lastly, according to our linear regression model, we may assume that AQI is not much moved by the weather parameters, that's why the linear regression prediction model might not have given a stellar result.

6. References

1. http://case.doe.gov.bd/index.php?option=com_content&view=category&id=8&Itemid=32
2. http://data.gov.bd/
3. https://github.com/TSGreen/bangladesh-air-quality
4. https://github.com/TanvirMahmudEmon/Rainfall-Prediction