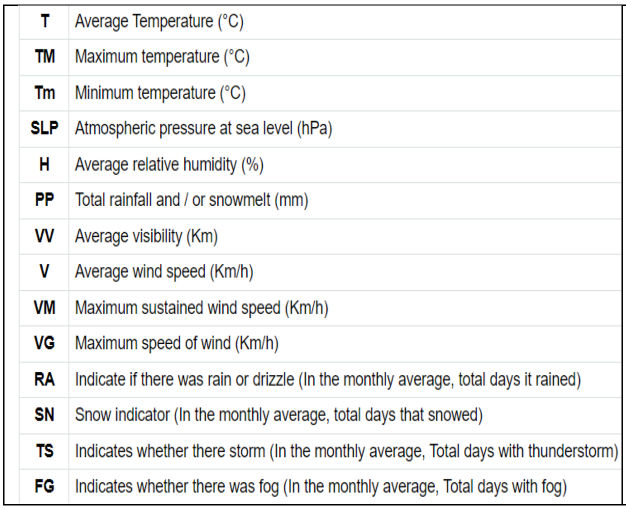
Big Data Analysis and Project – Part B

Air Aware: AQI Prediction and Analysis

Implementing data-driven techniques to predict air quality index and inform public health decisions

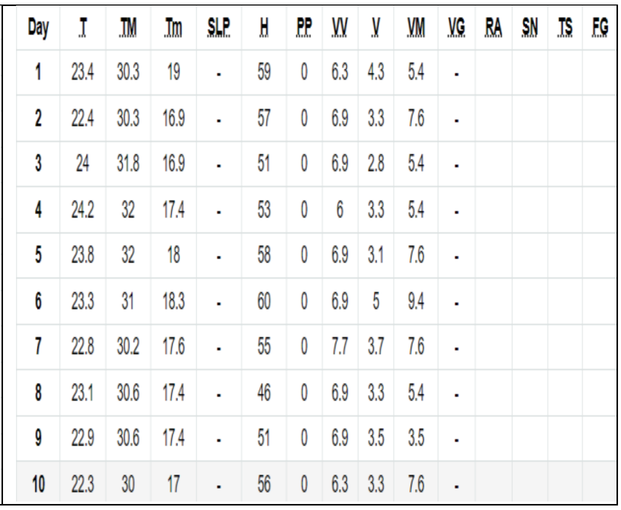
Combining datasets and Initial analysis

On further advancement of analysis, I decided to combine the two datasets (Tutiempo and Weather Map). I needed to gather data containing other variables that I would use for my predictive analysis. From my first dataset, I had the PM2.5 values of Bangalore city. These values were recorded on an hourly basis, on each day from the years 2013 to 2018. To extract only the PM2.5 feature and combine it with the other features of my second dataset, I computed the average PM2.5 value for each day of every month, from 2013 to 2018.



T	Average Temperature (°C)
TM	Maximum temperature (°C)
Tm	Minimum temperature (°C)
SLP	Atmospheric pressure at sea level (hPa)
H	Average relative humidity (%)
PP	Total rainfall and / or snowmelt (mm)
VV	Average visibility (Km)
V	Average wind speed (Km/h)
VM	Maximum sustained wind speed (Km/h)
VG	Maximum speed of wind (Km/h)
RA	Indicate if there was rain or drizzle (In the monthly average, total days it rained)
SN	Snow indicator (In the monthly average, total days that snowed)
TS	Indicates whether there storm (In the monthly average, Total days with thunderstorm)
FG	Indicates whether there was fog (In the monthly average, Total days with fog)

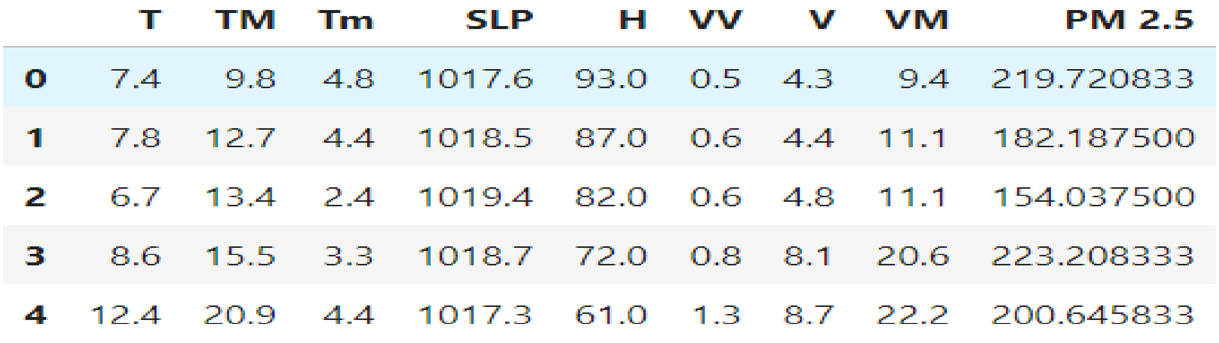
Fig 1. Description of the independent variables



Day	T	TM	Tm	SLP	H	PP	VV	V	VM	VG	RA	SN	TS	FG
1	23.4	30.3	19	-	59	0	6.3	4.3	5.4	-				
2	22.4	30.3	16.9	-	57	0	6.9	3.3	7.6	-				
3	24	31.8	16.9	-	51	0	6.9	2.8	5.4	-				
4	24.2	32	17.4	-	53	0	6	3.3	5.4	-				
5	23.8	32	18	-	58	0	6.9	3.1	7.6	-				
6	23.3	31	18.3	-	60	0	6.9	5	9.4	-				
7	22.8	30.2	17.6	-	55	0	7.7	3.7	7.6	-				
8	23.1	30.6	17.4	-	46	0	6.9	3.3	5.4	-				
9	22.9	30.6	17.4	-	51	0	6.9	3.5	3.5	-				
10	22.3	30	17	-	56	0	6.3	3.3	7.6	-				

Fig 2. Head of the second dataset – original form

Now, there are 9 variables in total in my final dataset. I dropped 'Date, Time, PM2.5 AQI' from the first dataset (Weather Map) and 'Day, PP, VG, RA, SN, TN, FG' from the second dataset (Tutiempo), because these variables aren't contributing any significant information to the modeling.



	T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
0	7.4	9.8	4.8	1017.6	93.0	0.5	4.3	9.4	219.720833
1	7.8	12.7	4.4	1018.5	87.0	0.6	4.4	11.1	182.187500
2	6.7	13.4	2.4	1019.4	82.0	0.6	4.8	11.1	154.037500
3	8.6	15.5	3.3	1018.7	72.0	0.8	8.1	20.6	223.208333
4	12.4	20.9	4.4	1017.3	61.0	1.3	8.7	22.2	200.645833

Fig 3. The first few rows of the final dataset

It can be observed that all the features in the final dataset are quantitative variables and there are no categorical features. The response variable (PM2.5) and the other independent variables are quantitative variables.

Exploratory data analysis

After dropping the unnecessary variables, I went ahead with checking of null values and missing values in my dataset. To visualize the null values, I used a heatmap from the Seaborn library. In the first diagram, we see that the null values are shown by the yellow line. In the second diagram, we see that the grid is completely purple after dropping the rows containing null values.

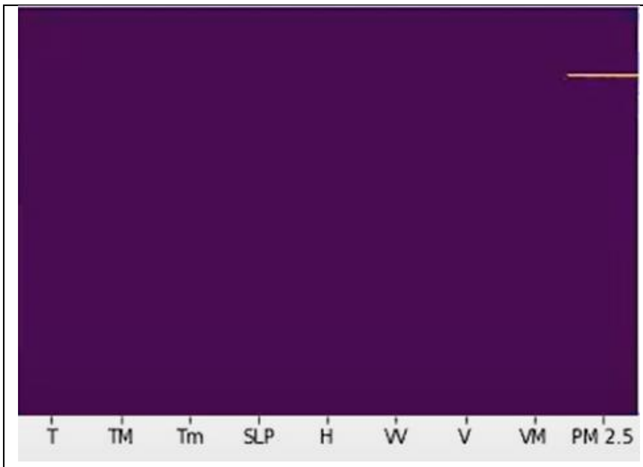


Fig 4. Heatmap before dropping null rows

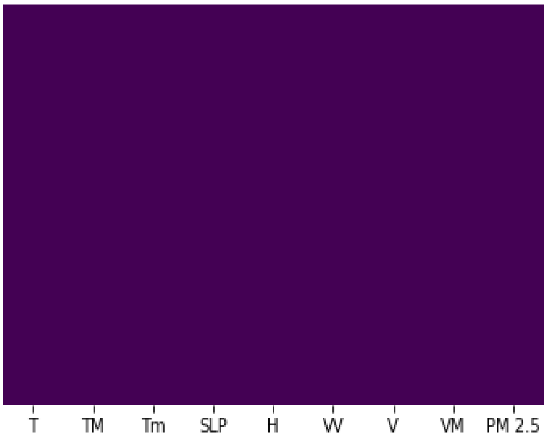
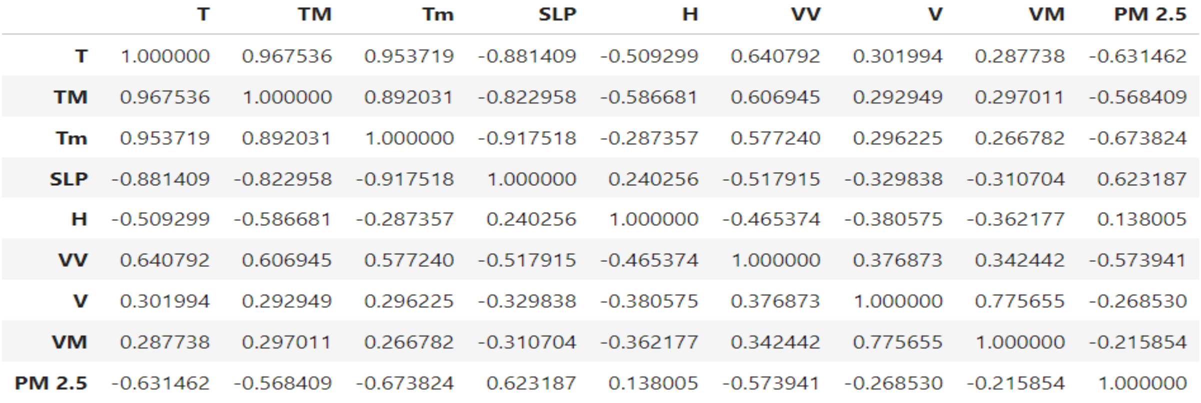


Fig 5. Heatmap after dropping null rows

After this, I divided my dataset into dependent and independent features. 'PM2.5' is the dependent feature and the rest of the features are independent. With the help of the Seaborn Library, I performed a multivariate analysis. This allowed me to evaluate all the variables and identify correlations between them. I used the pairplot() method to plot multiple pairwise, bivariate distributions. We can observe the correlation of every variable with every other feature of the dataset. For example, the first pair plot tells us that as T increases, the TM and Tm also increase. They are linearly growing, and it is easy to see that. But our focus is on the target feature PM2.5 viz. the last row of the pair plot distribution diagram. The plots are not very clearly linear. This is an indication that using Linear Regression might not necessarily be the best idea. Furthermore, I got the numerical values of the pairwise correlations between each feature using the 'corr()' method. For example, when VV increases, there is a 57.39% probability that PM2.5 decreases.



	T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
T	1.000000	0.967536	0.953719	-0.881409	-0.509299	0.640792	0.301994	0.287738	-0.631462
TM	0.967536	1.000000	0.892031	-0.822958	-0.586681	0.606945	0.292949	0.297011	-0.568409
Tm	0.953719	0.892031	1.000000	-0.917518	-0.287357	0.577240	0.296225	0.266782	-0.673824
SLP	-0.881409	-0.822958	-0.917518	1.000000	0.240256	-0.517915	-0.329838	-0.310704	0.623187
H	-0.509299	-0.586681	-0.287357	0.240256	1.000000	-0.465374	-0.380575	-0.362177	0.138005
VV	0.640792	0.606945	0.577240	-0.517915	-0.465374	1.000000	0.376873	0.342442	-0.573941
V	0.301994	0.292949	0.296225	-0.329838	-0.380575	0.376873	1.000000	0.775655	-0.268530
VM	0.287738	0.297011	0.266782	-0.310704	-0.362177	0.342442	0.775655	1.000000	-0.215854
PM 2.5	-0.631462	-0.568409	-0.673824	0.623187	0.138005	-0.573941	-0.268530	-0.215854	1.000000

Fig 6. Pair-wise correlations

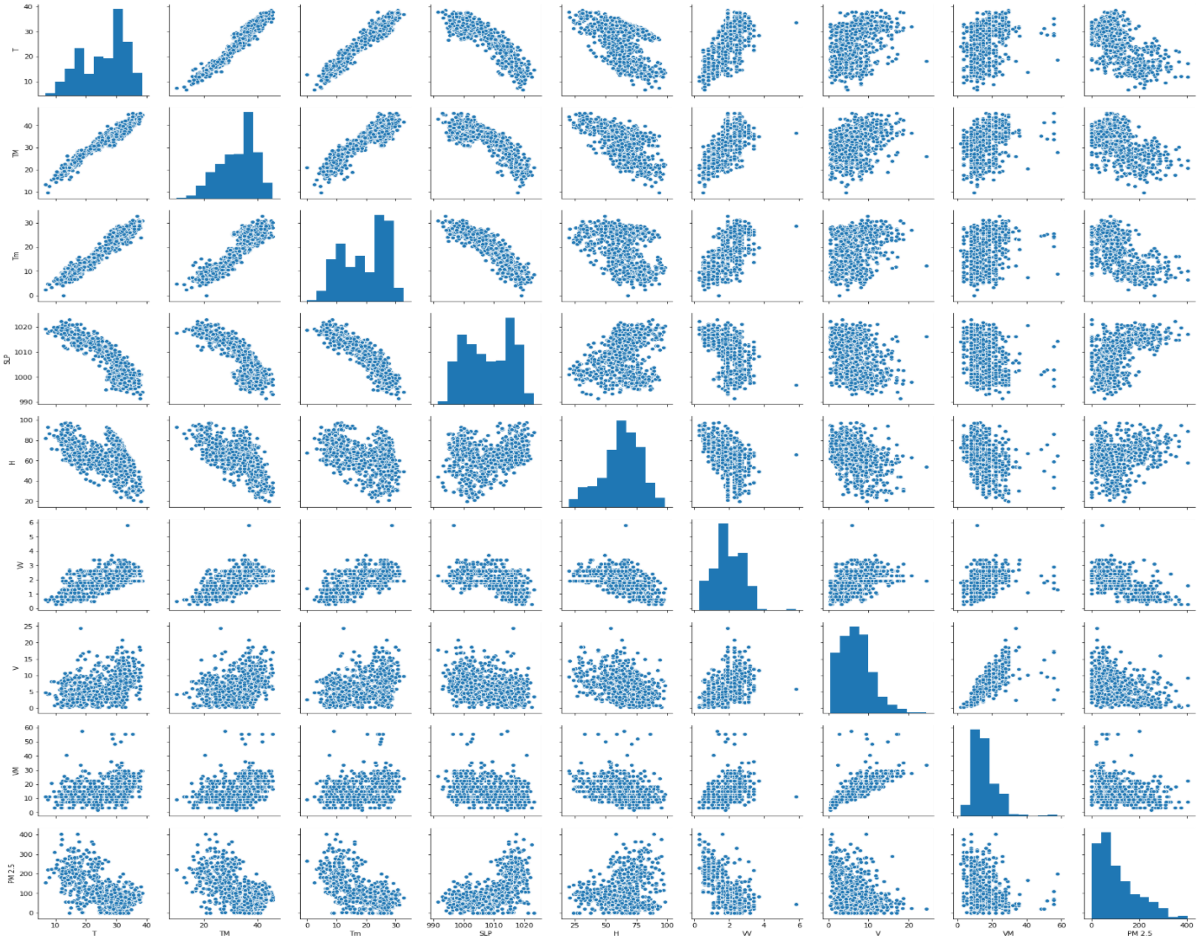


Fig 7. Pair-wise bivariate distributions in the dataset

A heatmap makes it easy to identify which features are most related to the target variable. This further helped me in deciding which feature is important and which one is not. The higher the value of the correlation of PM2.5 with every independent feature (it can be positive or negative), the higher is it's importance. For example, the value between PM2.5 and T is -0.63 but the value between PM2.5 and H is only 0.14. So, we can say that PM2.5 has a high correlation with T. Similarly, 'T, TM, Tm, VV' have a high correlation with PM2.5.

We can also determine the significance of each feature in our dataset using the model's feature importance property. This provides a score indicating each feature's relevance to the output variable. Tree-based regressors, like the Extra Trees Regressor, have this built-in capability.

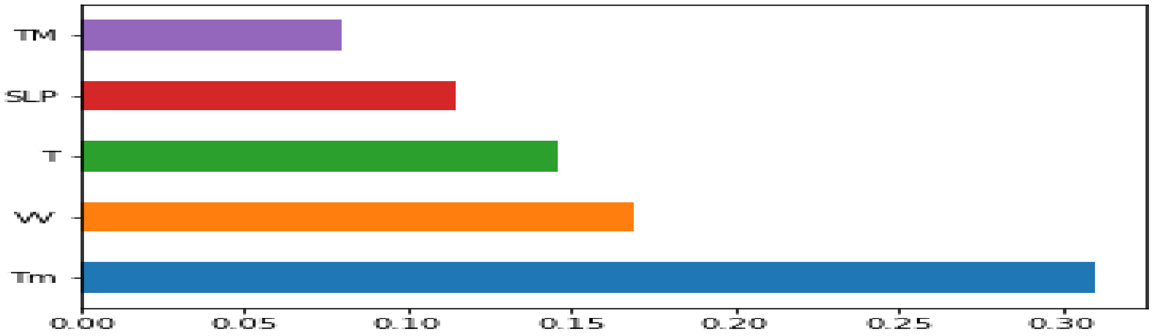
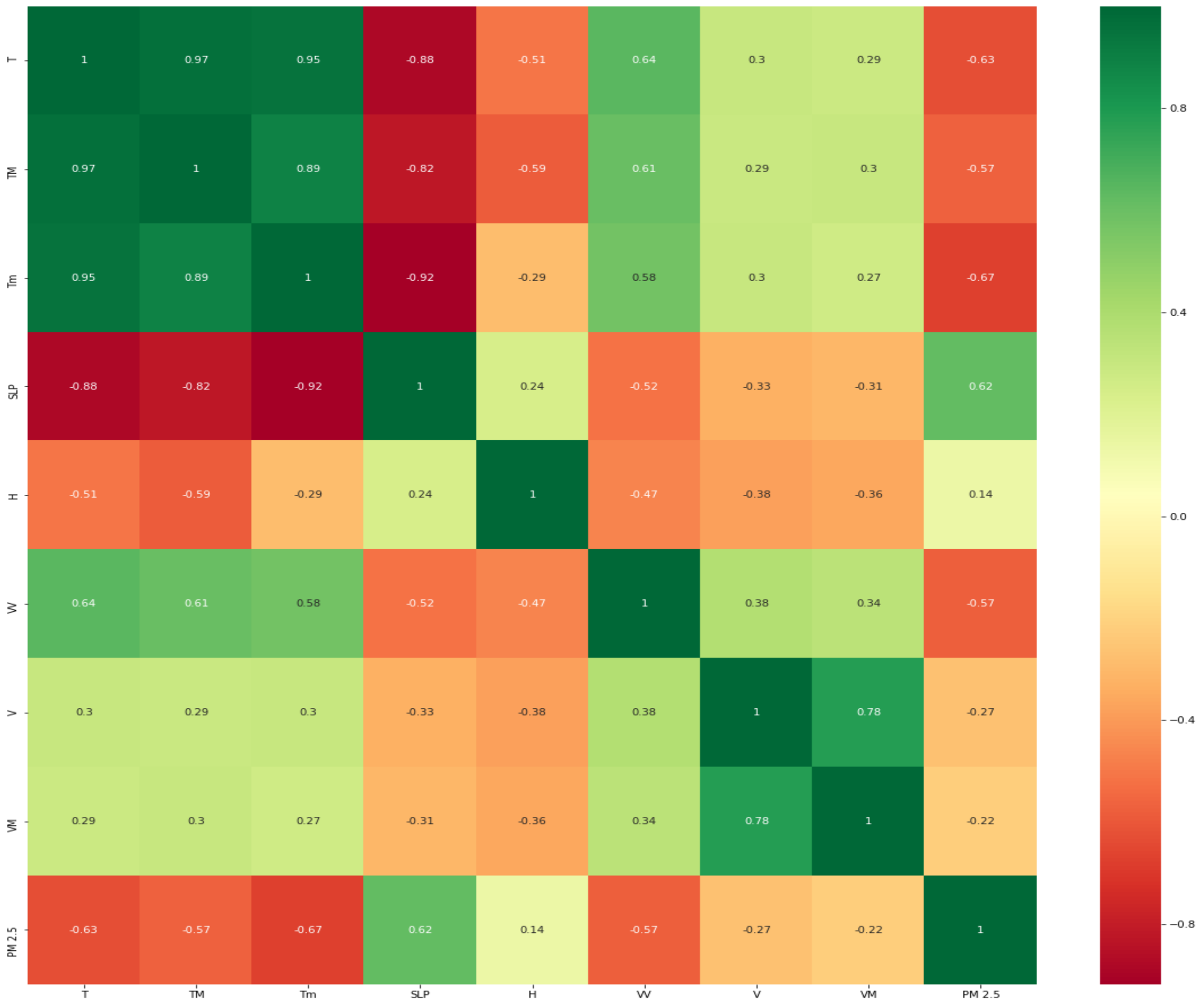


Fig 8. Feature importance bar plot



	T	TM	Tm	SLP	H	VV	V	VM	PM 2.5
T	1	0.97	0.95	-0.88	-0.51	0.64	0.3	0.29	-0.63
TM	0.97	1	0.89	-0.82	-0.59	0.61	0.29	0.3	-0.57
Tm	0.95	0.89	1	-0.92	-0.29	0.58	0.3	0.27	-0.67
SLP	-0.88	-0.82	-0.92	1	0.24	-0.52	-0.33	-0.31	0.62
H	-0.51	-0.59	-0.29	0.24	1	-0.47	-0.38	-0.36	0.14
VV	0.64	0.61	0.58	-0.52	-0.47	1	0.38	0.34	-0.57
V	0.3	0.29	0.3	-0.33	-0.38	0.38	1	0.78	-0.27
VM	0.29	0.3	0.27	-0.31	-0.36	0.34	0.78	1	-0.22
PM 2.5	-0.63	-0.57	-0.67	0.62	0.14	-0.57	-0.27	-0.22	1

Fig 9. Correlation matrix with heatmap

Both techniques gave us the same answer.

<u>Linear Regression</u>

Since both my target and predictor variables are quantitative, I decided to try Linear Regression. I split the dataset into training and testing sets, trained the model using the Sklearn library, and computed the R^2 score on both the training and testing sets. It was found that the R^2 score for the train set was 0.5515 and for the test set, it was 0.4852. Additionally, the mean score of cross-validation was 0.4710.

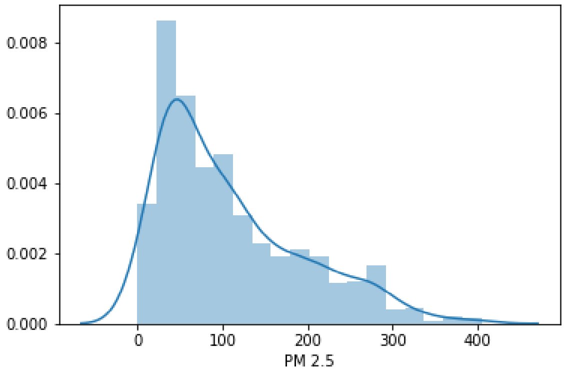


Fig 10. Distribution of response variable

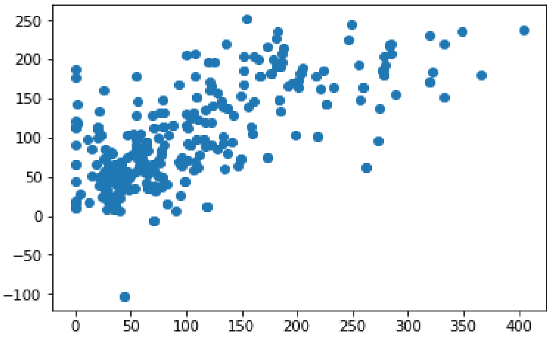


Fig 11. Scatter plot between y-test and x-test

Conclusion: The linear Regression model gives us the line of best fit, but it may not be the best choice for modeling quantitative data. The evaluation metrics did not give us a very high score. Hence, the next step would be to try the Lasso and Ridge Regression. If not successful in getting a good evaluation score, I will try other models like Random Forest Regressor and Decision Tree Regressor.

References:

Tutiempo 2024, World Weather, Tutiempo 2024, viewed 7 June 2024, <https://en.tutiempo.net/>

Aman Kharwal 2023, *Air Quality Index Analysis Using Python*, The Clever Programmer, viewed 7 June 2024, <https://thecleverprogrammer.com/2023/09/18/air-quality-index-analysis-using-python/>

World Health Organization 2022, WHO Ambient Air quality database 2022 v5, viewed 7 June 2024, <https://data.who.int/countries/036>

Air Now n.d., Air Quality Index (AQI) basics, viewed 7 June 2024, <https://www.airnow.gov/aqi/aqi-basics/#:~:text=Think%20of%20the%20AQI%20as,300%20represents%20hazardous%20air%20quality.>