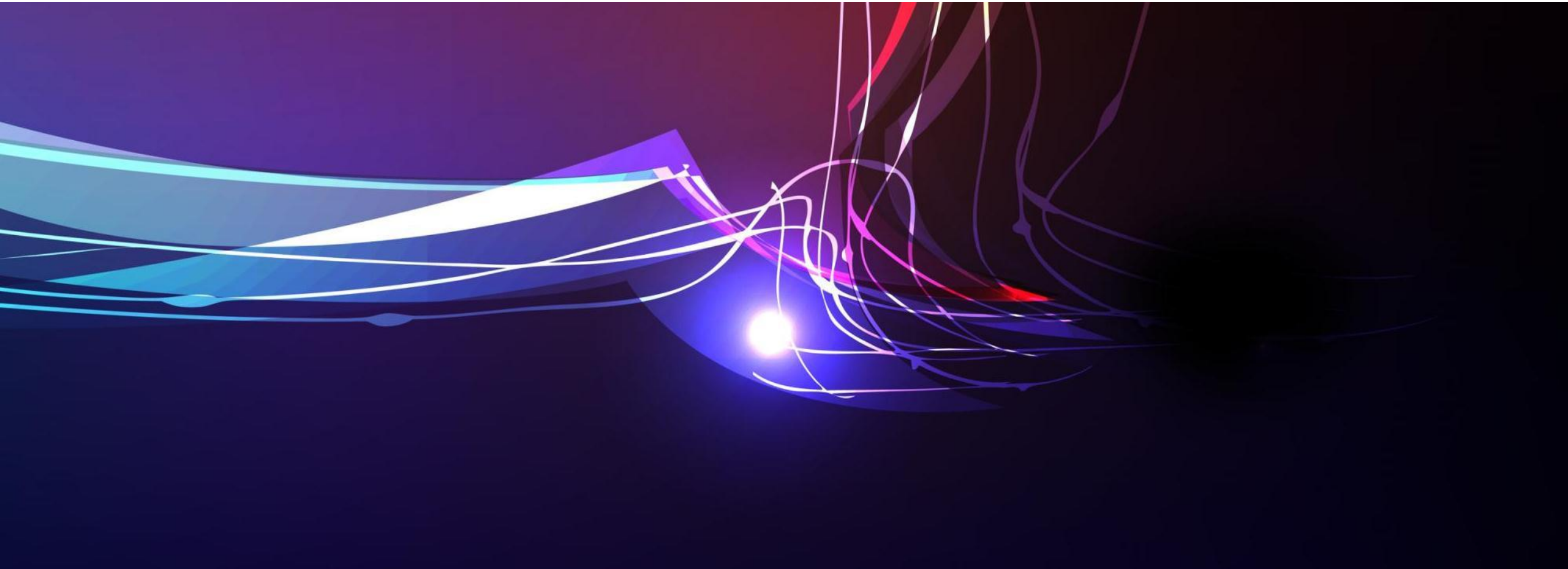
Air Quality – Machine Learning Project Report

By Shikhar Raj



Agenda

- About the Project
- Dataset description
- Business questions identified
- Hidden insights of the dataset
- Approach to the solution
- Solution Details
- Conclusion

Air Quality – Project Description

The Dataset

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device

Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer.

The objective is to predict the Relative Humidity at a given point of time based on all other attributes affecting the change in RH

Dataset Attributes

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. CO Concentration
4. PT08.S1 (tin oxide)
5. Non Metanic Hydro Carbons concentration
6. Benzene concentration
7. Titania hourly averaged
8. NOx concentration
9. Tungsten Oxide concentration
10. NO2 concentration
11. Tungsten oxide
12. Indium oxide
13. Temperature in °C
14. Relative Humidity (%) - Target Variable
15. AH Absolute Humidity

Air Quality – Business Questions to be addressed

1. How good is the quality of data collected? Are there any major discrepancies?
2. What is the pattern of air quality during the day?
3. Which factors has +ve relationship with Relative Humidity and which have –ve relation?
4. What is the monthly pattern of air quality?
5. How is air quality on a weekday Vs weekend?
6. Which are the most important features that determine Relative Humidity?
7. Build a prediction model to forecast Relative Humidity on any given day.

Air Quality – How good is the quality of data collected?

1. The dataset had 9358 rows. There were considerable NaN values
2. Strategy for filling NaN values
   - The CSV file was read replacing -200 with NaN and Date was read as DateTime data type
   - As the column name Date and Time are python pre defined data types, the column names were changed
   - As each subsequent row is contiguous date and hours, the best strategy to fill NaN values is to interpolate
   - Record_Date missing value was filled with Forward Fill method

Missing Values

Record_Date	114
Record_Hour	114
CO(GT)	1797
PT08.S1(CO)	480
NMHC(GT)	8557
C6H6(GT)	480
PT08.S2(NMHC)	480
NOx(GT)	1753
PT08.S3(NOx)	480
NO2(GT)	1756
PT08.S4(NO2)	480
PT08.S5(O3)	480
T	480
RH	480
AH	480

dtype: int64

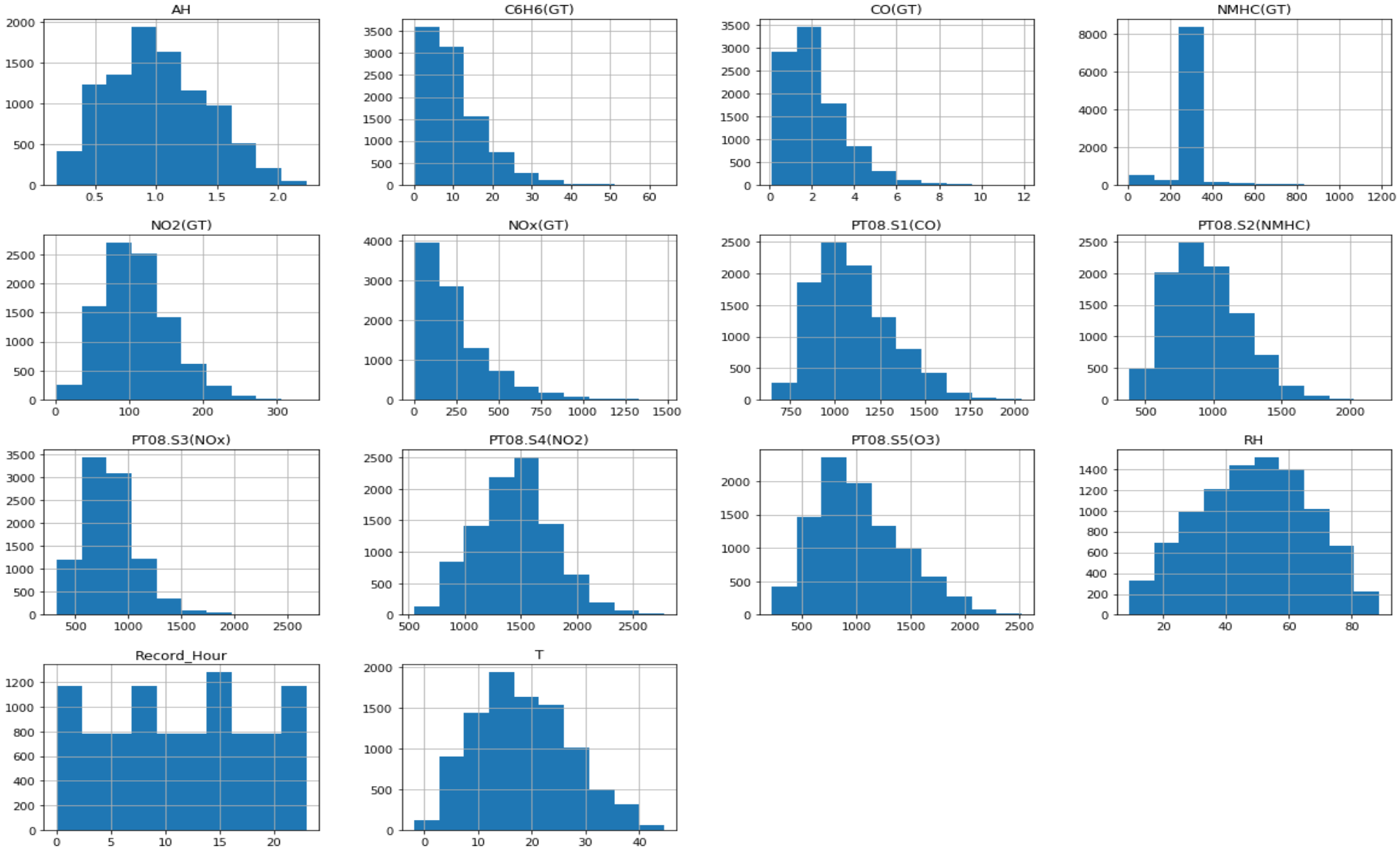
Air Quality – How good is the quality of data collected?

- NMHC(GT) has very high max value as compared to mean but IQR values look OK. It appears, there could be outliers

- Some of the input variables have very high std deviation but IQR range is OK: CO(GT), C6H6(GT) and NOx(GT). There could be some skewness in the data

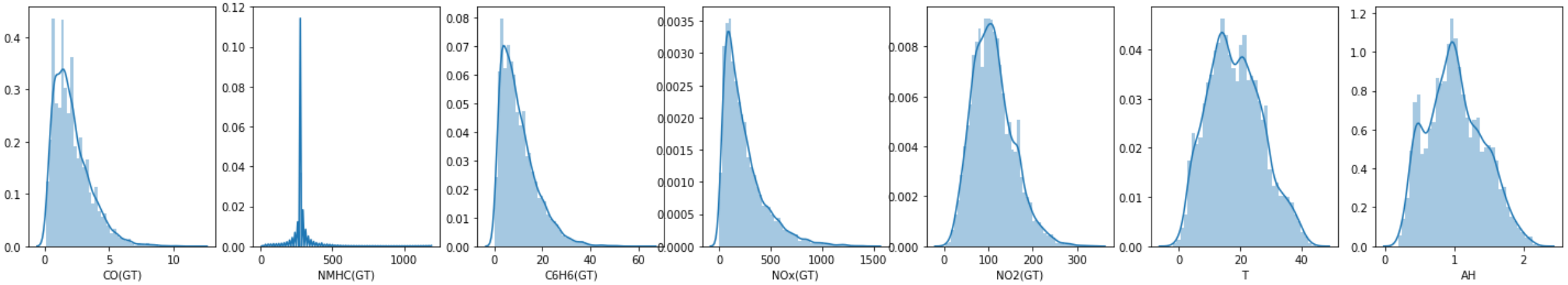
	Record_Hour	CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3
count	9471.000000	9471.000000	9471.000000	9471.000000	9471.000000	9471.000000	9471.000000	9471.000000	9471.000000	9471.000000	9471.0000
mean	11.528666	2.131438	1102.673846	269.896526	10.199868	943.404762	242.199979	830.607222	110.334653	1449.395312	1029.9378
std	6.886790	1.423113	216.907224	73.805872	7.460870	266.494895	203.097172	254.912582	46.618419	342.962507	402.6989
min	0.000000	0.100000	647.000000	7.000000	0.100000	383.000000	2.000000	322.000000	2.000000	551.000000	221.0000
25%	6.000000	1.100000	939.000000	275.000000	4.500000	739.000000	96.057823	654.000000	76.000000	1215.000000	737.0000
50%	12.000000	1.800000	1071.000000	275.000000	8.400000	914.000000	181.727273	800.000000	105.000000	1455.000000	964.0000
75%	17.000000	2.900000	1236.000000	275.000000	14.000000	1117.000000	323.000000	965.000000	138.000000	1664.000000	1288.5000
max	23.000000	11.900000	2040.000000	1189.000000	63.700000	2214.000000	1479.000000	2683.000000	340.000000	2775.000000	2523.0000

Air Quality – How good is the quality of data collected?



Air Quality – How good is the quality of data collected?

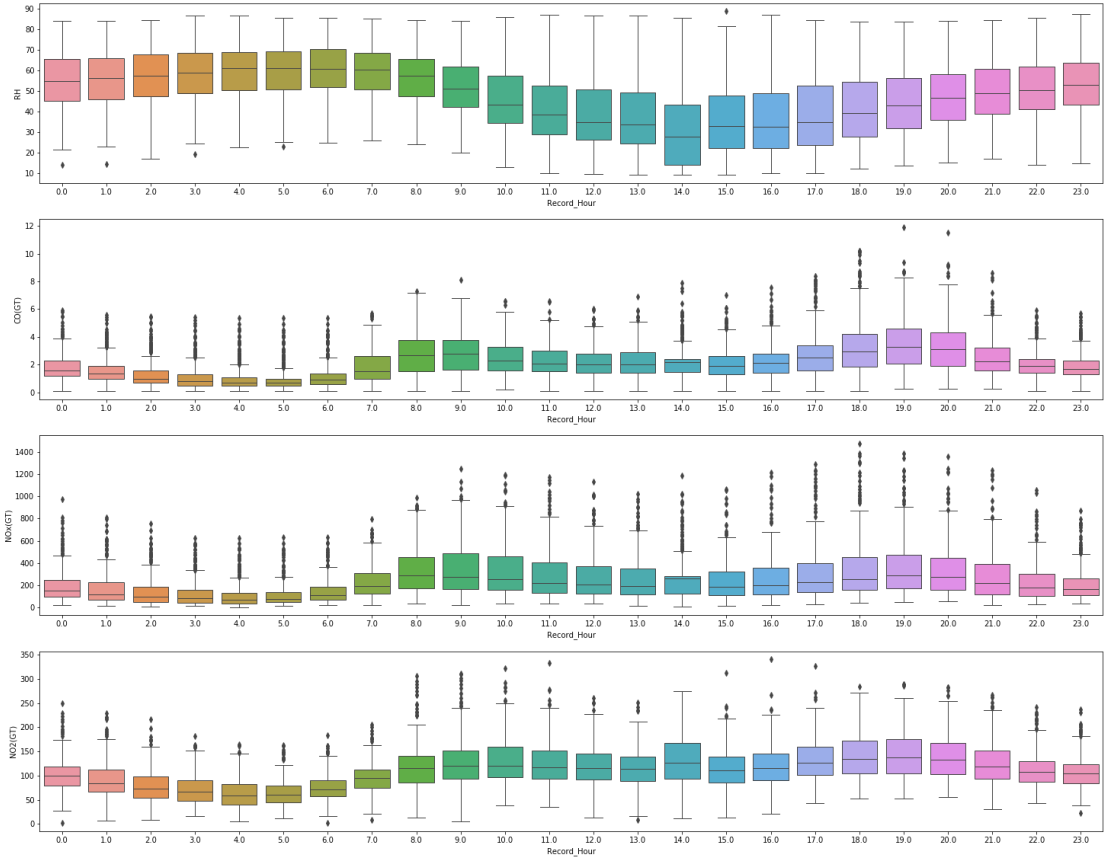
The distribution graph confirms the observation recorded in statistical observation recorded earlier on high standard deviation and skewness



Air Quality – What is the pattern of air quality during the day?

- CO(GT), NOx(GT) and NO2(GT) seems to follow similar pattern during the day. They peak during daytime or office hours(8 am to 8 pm)

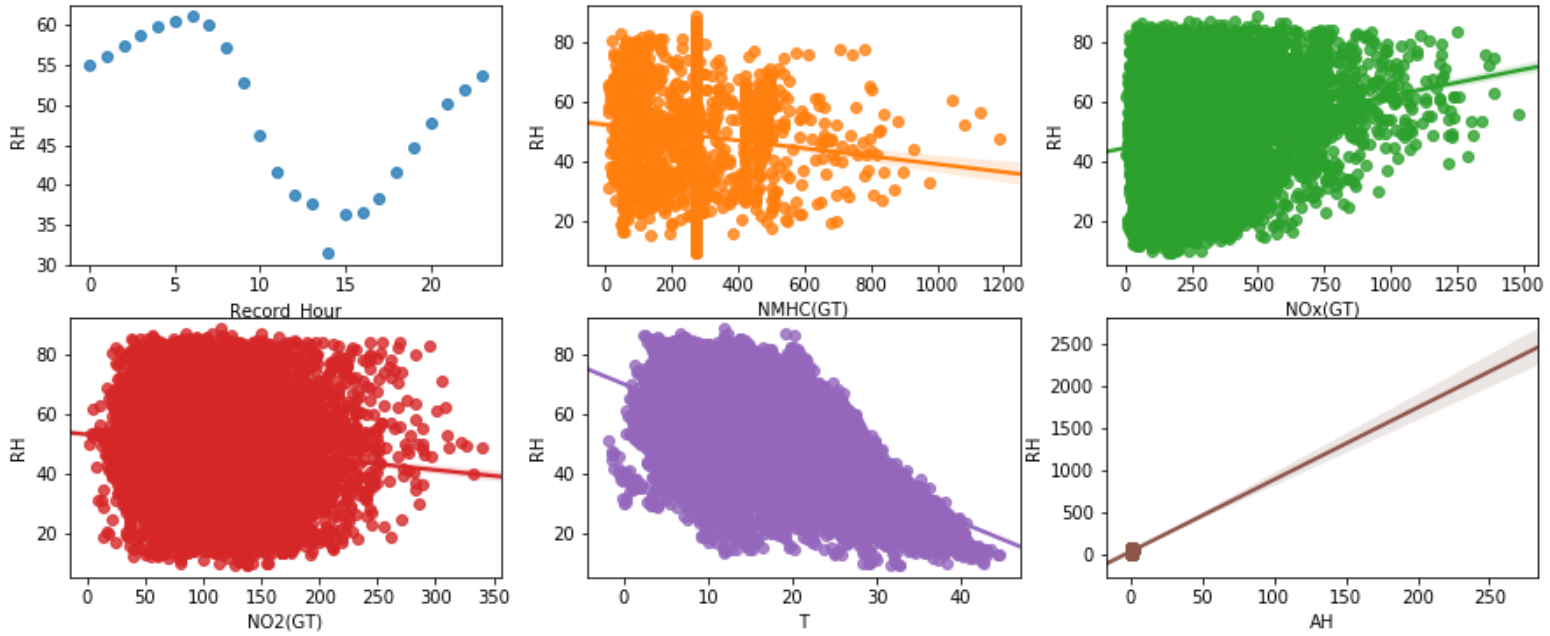
- RH(Output Variable) seems to follow a different pattern during the day, it shows higher values during night(8 pm to 8 am) than daytime



Air Quality – Positive and Negative Factors for Relative Humidity

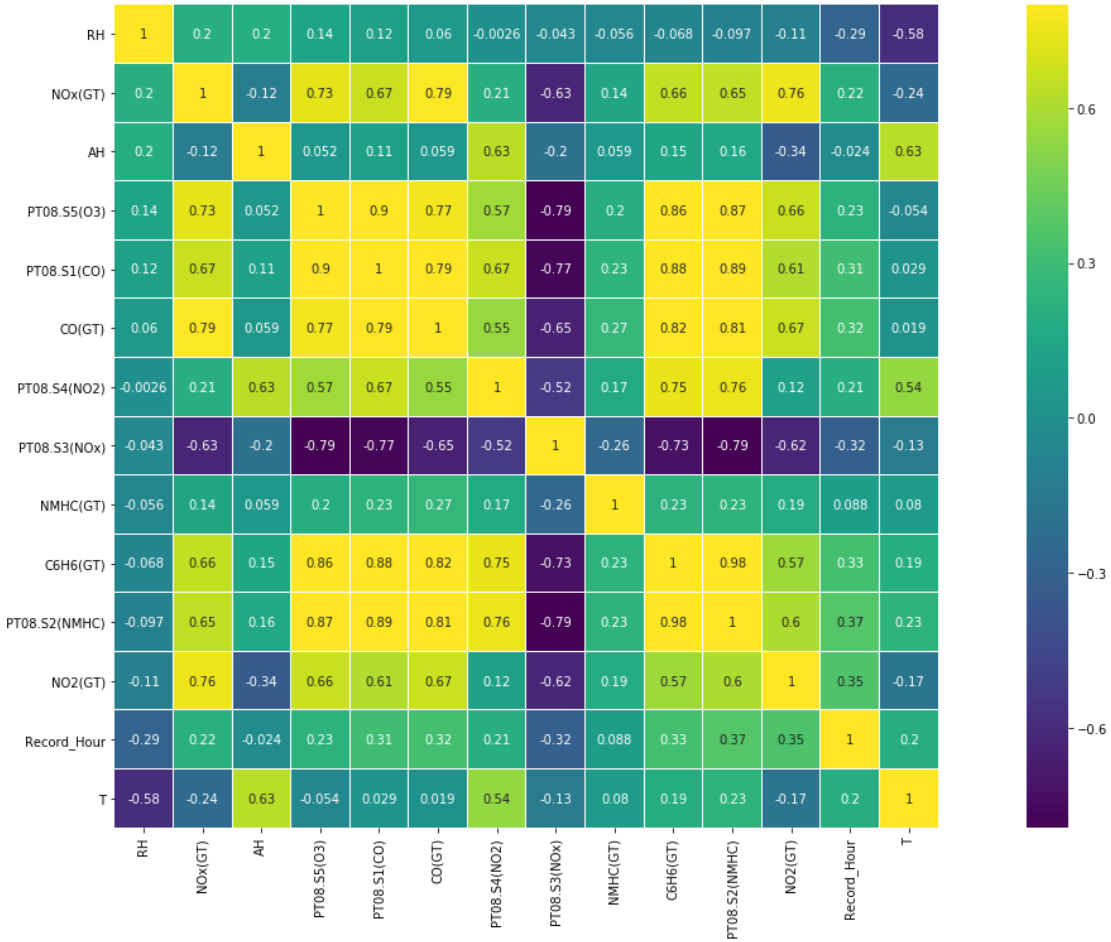
- NOx seems to be positively correlated to RH(o/p variable) whereas temp and NMHC(GT) appear to be -vely correlated with RH

- Relative Humidity varies cyclically with Hour of the day



Air Quality – Positive and Negative Factors for Relative Humidity

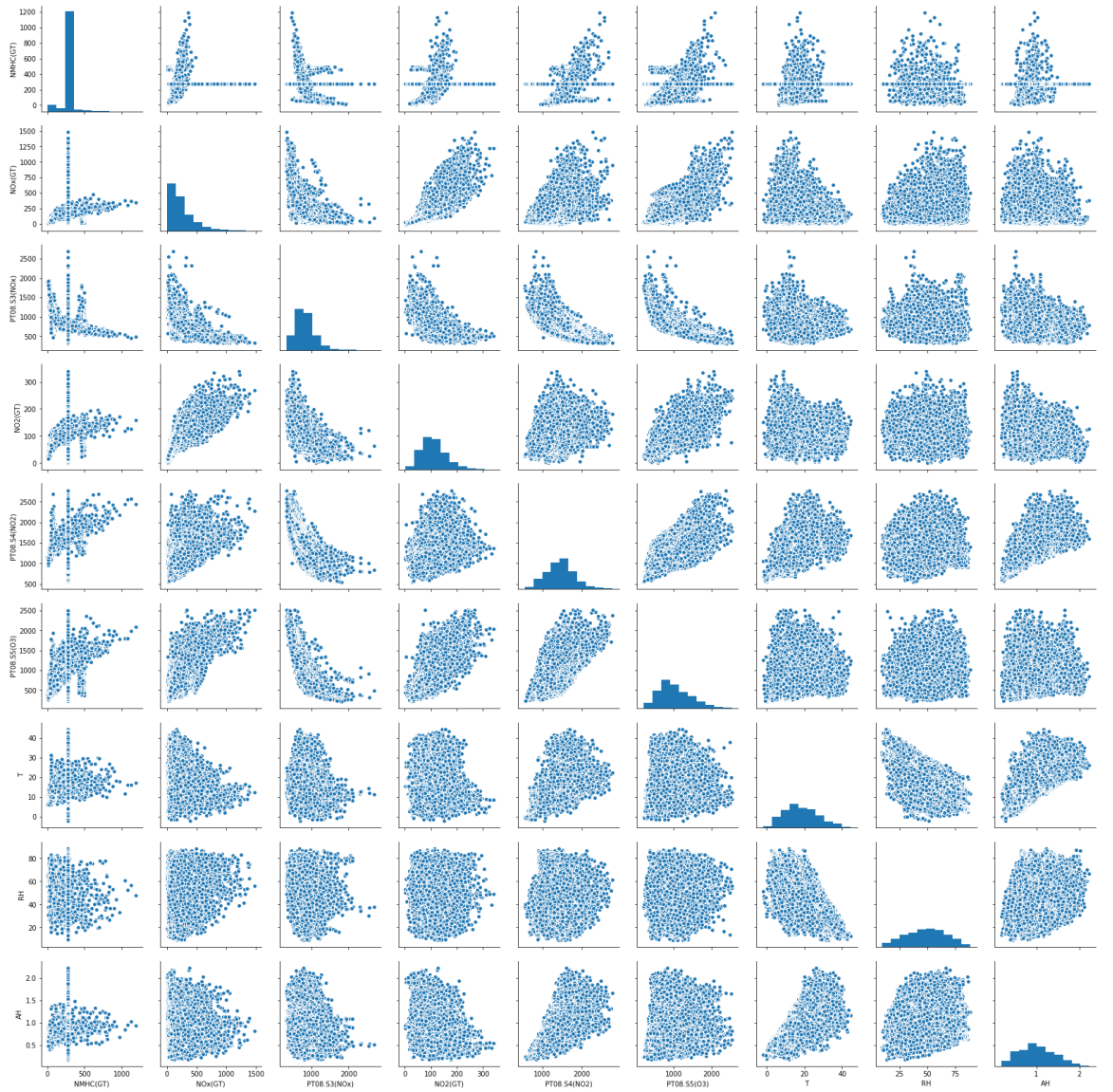
- There are some very high correlated predictor variables. We must clean this up
- NOx(GT) is highly correlated with CO(GT), hence, we shall take only 1 of them. Since, NOx(GT) has higher correlation with target variable, we keep it and drop CO(GT)
- PT08.S5(O3) is highly correlated with PT08.S1(CO), hence, we shall take only 1 of them. Since, PT08.S5(O3) has higher correlation with target variable, we keep it and drop PT08.S1(CO)
- PT08.S5(O3) and CO(GT) are again twins. We have already dropped CO(GT)
- PT08.S4(NO2), C6H6(GT) and PT08.S2(NMHC) are triplets. We drop C6H6(GT) and PT08.S2(NMHC)



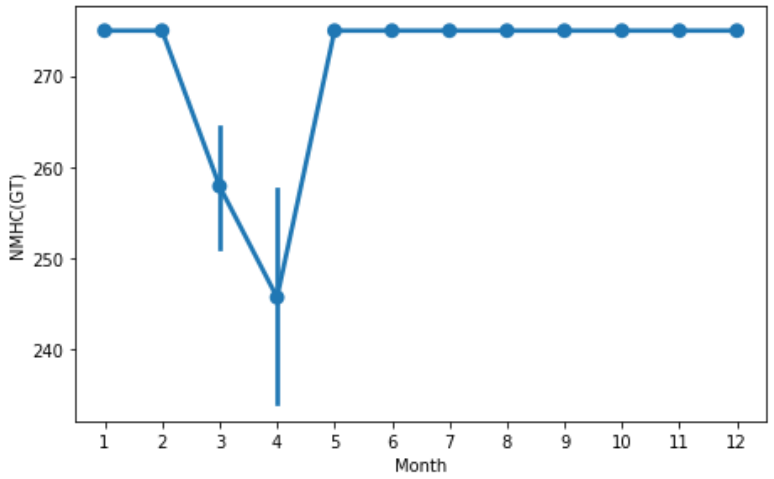
Air Quality – Positive and Negative Factors for Relative Humidity

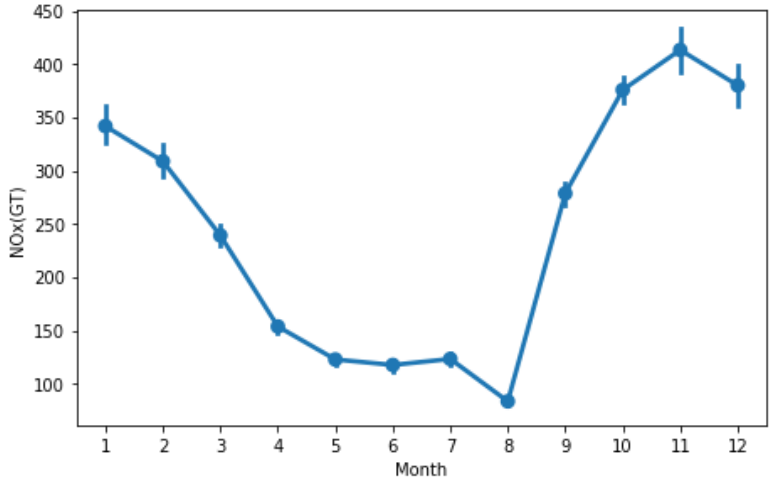
- The Pair Plot shows +ve correlation between some variable, -ve correlation with some variables.

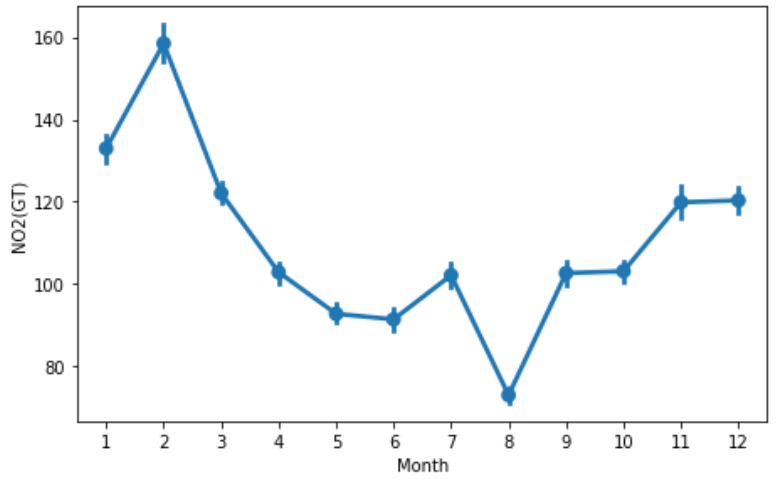
- The output variable(RH) does not shows either +ve or -ve correlation with some variables - NO2(GT), PT08.S4(NO2), PT08.S5(O3)

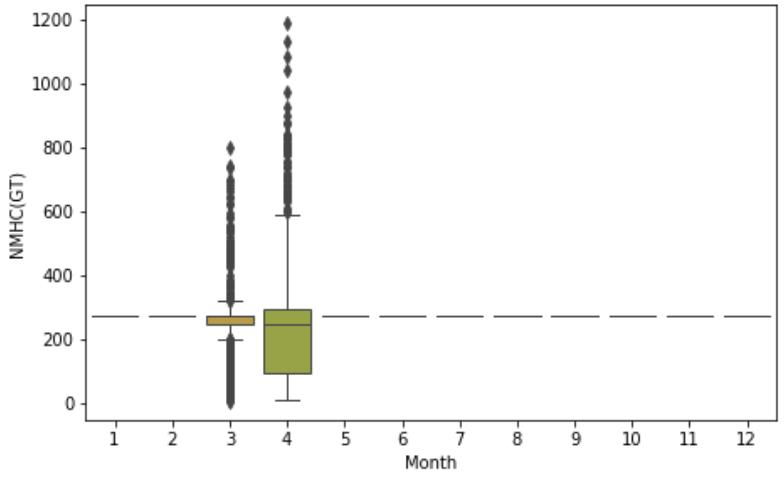


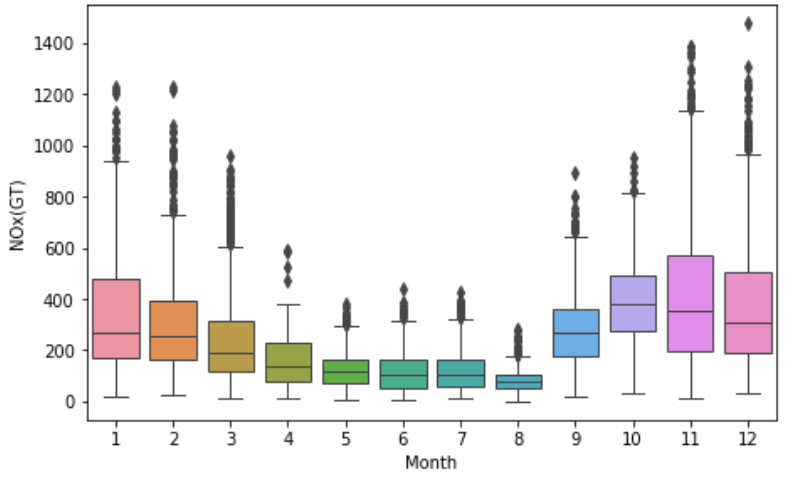
Air Quality – How is the monthly pattern of air quality?

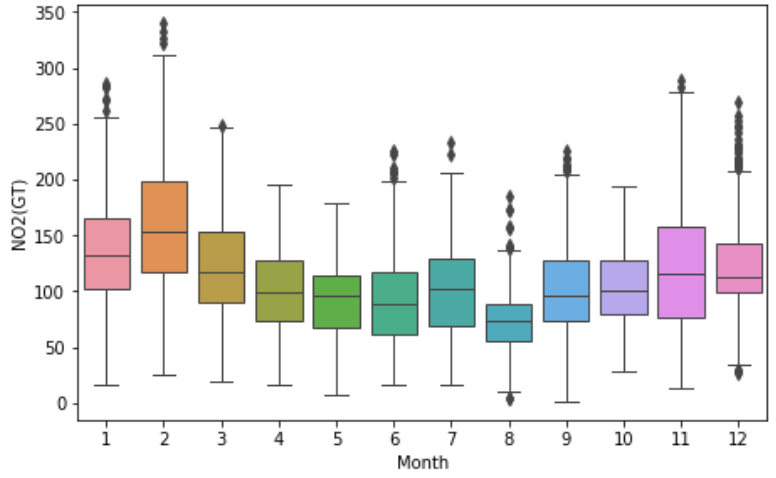




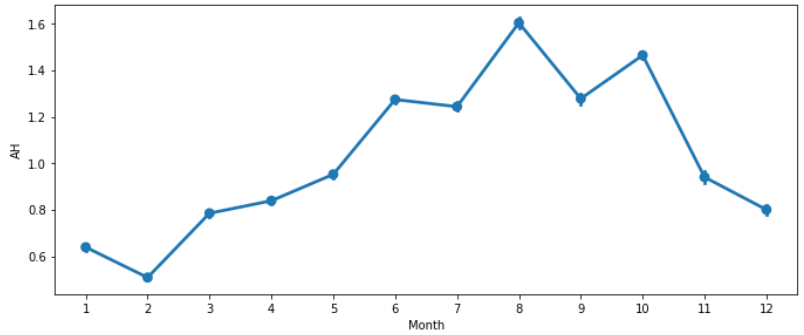


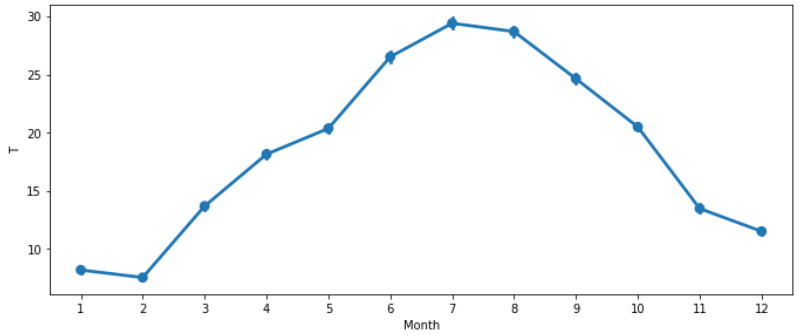


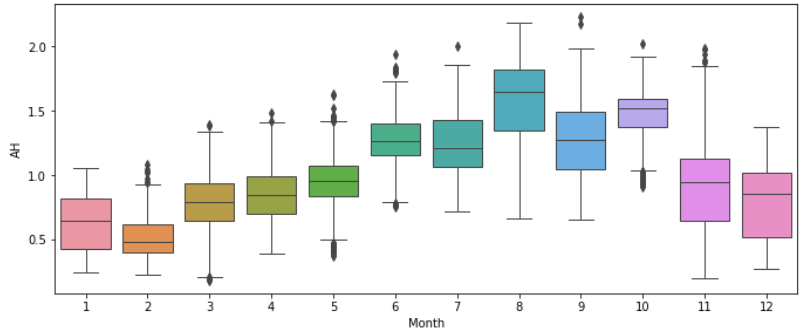


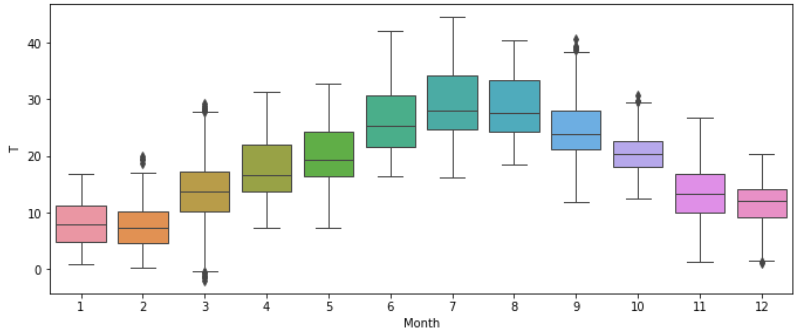


Air Quality – How is the monthly pattern of air quality?



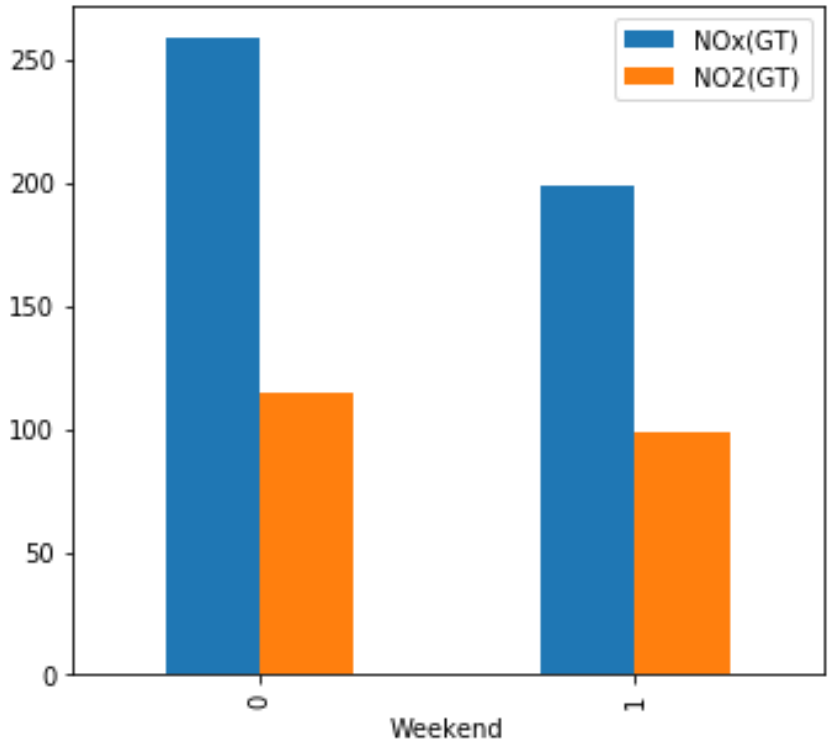


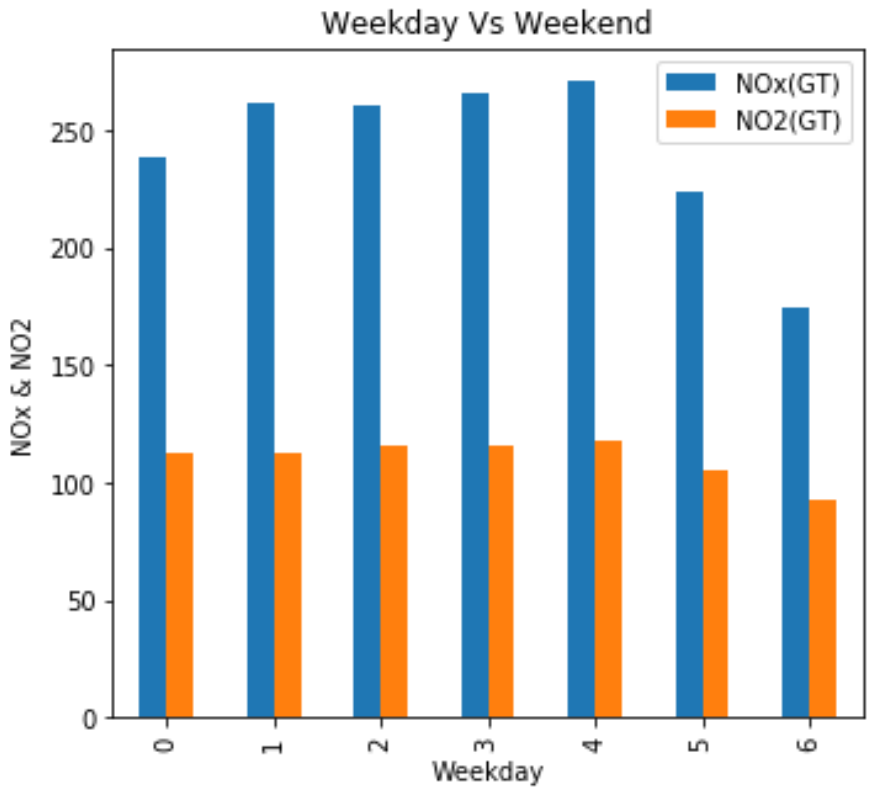




- Colder months(September to February) show higher concentration of NOx
- NMHC shows sudden dip in level in March and April. For other months, its nearly the same
- NO2 shows high value in Jan and Feb compared to other months
- Absolute Humidity goes higher during rainy season(June to October) where as the temperature goes higher during summer

Air Quality – How is air quality on a weekday Vs the weekend?





- The NOx and NO2 levels on weekends are considerably lower than on Weekdays
- The NOx levels is considerably lower on Monday among weekdays

Model Building

Air Quality – Solution Approach

- Different input variable have different ranges, this may lead to bias in the model. We first normalise the data
- We then split the dataset into train and test data
- We consider 6 different algorithms (Linear Regression, Lasso, Elastic Net, K Nearest Neighbour, Decision Tree/Random Forest and SVM) to find out best performing models
- Next we shall reduce dimensionality of the model. We may use Principal Component Analysis(PCA) or Recursive Feature Elimination(RFE). RFE gave a good result, hence, I used RFE
- Now, we have best performing model, best performing features and normalised predictor variables. We now do detailed analysis with top 3 algorithms and compare result

Air Quality – Comparison of Algorithm

- The K Nearest Neighbour, SVM and Decision Tree provides best accuracy on test data and lowest RMSE value.

- Thus we shall do detailed implementation of these three algorithms – KNN, SVM and Decision Tree and then select our final algorithm

	Algorithm Name	Train Accuracy	Test Accuracy	RMSE
3	KNeighborsRegressor	0.9993	0.9987	0.629172
5	SVR	0.9980	0.9981	0.770554
4	DecisionTreeRegressor	1.0000	0.9964	1.063886
0	LinearRegression	0.8663	0.8655	6.470510
1	Lasso	0.8486	0.8478	6.883080
2	ElasticNet	0.5618	0.5620	11.675846

Air Quality – Selecting best predictors variables

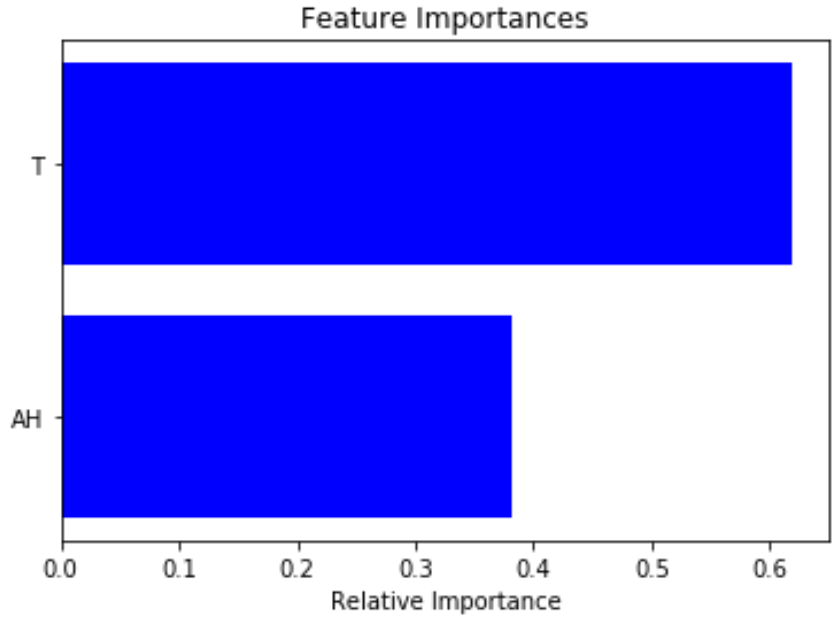
Iteration	No of Predictors	Best Predictors	Test Accuracy	RMSE	Adjusted R^2
1	5	T, AH, Record_Hour, dNOx, PT08.S4(NO2)	98.46%	1.08	0.87
2	10	T, AH, Record_Hour, NOx(GT), NO2(GT), PT08.S4(NO2), DateOfMonth, Weekday, dNOx, dNO2	99.61%	1.10	0.873
3	**2**	**T, AH**	**99.63%**	**1.065**	**0.866**
4	3	T, AH, Record_Hour	99.63%	1.070	0.867
5	4	T, AH, Record_Hour, PT08.S4(NO2)	99.62%	1.07	0.870

Conclusion

- The test accuracy is best with 2 features – Temp and Absolute Humidity. It has lowest RMSE and Adjusted R Squared values.
- Thus, I go ahead with 2 features for detailed implementation in the 3 algorithms – KNN, SVM and Decision Tree

Air Quality – Model Development – Decision Tree/Random Forest

- Optimized Grid Search Parameters:
  - max_depth: 9
  - min_samples_split: 5
- Accuracy
  - Test Accuracy: 0.9963519859689669
  - RMSE: 1.065577926789993
  - Adjusted R Squared: 0.8662557093052974



Air Quality – Model Development – K Nearest Neighbours

- Optimized Grid Search Parameters:
  - n_neighbors: 4
  - p: 2
  - Weights: distance
- Accuracy
  - Test Accuracy: 0.998954729019120 6
  - RMSE: 0.5703893557115431
  - Adjusted R Squared: 0.8662557093052974

Air Quality – Model Development – SVM

- Optimized Grid Search Parameters:
  - gamma : 1
  - kernel : rbf
- Accuracy
  - Test Accuracy: 0.9982761915411439
  - RMSE: 0.7324896108613005
  - Adjusted R Squared: 0.8662557093052974

Air Quality – Model Development – Comparison of 3 Algorithms

Algorithm	Test Accuracy	RMSE	Adjusted R^2	
Decision Tree/ Random Forest	0.9963519859689669	1.065577926789993	0.8662557093052974	
KNN	0.9989547290191206	0.5703893557115431	0.8662557093052974	
SVM	0.9982761915411439	0.7324896108613005	0.8662557093052974	

Conclusion

KNN provides best accuracy, lowest RMSE value. Thus we choose KNN as our final model.

- **Test Accuracy: 99.895%**
- **RMSE: 0.57**
- **Adjusted R^2: 0.866**

Thank You