**Name:** Saahil Deshpande

**INFOSYS 722 Assignment 4 – BDAS**

**Association of Air Quality with increase in Global Temperature**

**1) Business and/or Situation understanding**

* 1. **Identify the objectives of the business/ situation**

Global warming has been an issue of concern for a long time. The global temperatures have been increasing for centuries, but the rate of increase has doubled in the past 40 years. The average temperature has increase by C since the 1880’s but this rate has increased to C since 1981 (Climate Change: Global Temperature, Rebecca Lindsey and LuAnn Dahlman, 2020). There are several factors which contribute to this rise in the temperature, of which one factor is the quality of the air. The presence of pollutants in the air affect the distribution of heat, hence entrapping more heat and causing an increase in the temperatures. The aim of this research is to figure out the presence of which particles in the air is most associated with this increase in the temperature. Once we figure out this association, we can look at the factors which contribute to emitting these particles and suggest an alternative in order to decrease the particle concentration. The objectives include:

1. Identify particles associated with change in temperature
   * Which particles affect the change in temperature?
   * Do all particles affect the temperature equally?
2. Provide a potential solution to temperature change
   * Advise ways in which emission of the particles can be decreased
   * Which particles should be focused on as an immediate threat?
   1. **Assess the situation**

Increase in global temperatures have an everlasting impact on our day to day life. It makes it difficult to grow crops. The frequency of droughts increases which leads to less availability of water for irrigation. Higher temperatures combined with irregular rainfall lead to increased growth of weeds and pests in previously unaffected areas.

Increase in temperatures causing irregular rainfall will lead to places receiving less water. Water is the major source of generating electricity. Without water, we would have to reply on fossil fuels to generate electricity. This will not only lead to depletion of natural resources, but also lead to increase in greenhouse gas emissions which will in turn lead to further rises in temperature. Increased temperatures lead to people making more use of air conditioners to keep themselves cool. This leads to consumption of electricity and release of greenhouse gases.

Lakes and rivers are the major sources of drinking water for humans. Frequent droughts and unpredictable rainfall deplete the water levels at the sources. This leads to scarcity of drinking water. People will require to cut usage of water for many household activities such as watering their lawns

Humans are not the only ones who experience the effects of rising temperatures. Plants and animals used to colder habitats are unable to find a place to live. Many ecosystems are destroyed to this change in the temperature. This in turn lead to the extinction of many species. The effects of rising temperatures are felt by live under the ocean as well. The corals in the ocean create the ecosystem for many fishes. Higher temperatures destroy the corals and they are on the path of being lost. This loss of habitat will affect all the life under water.

The dry conditions caused by higher temperatures create favorable conditions for wildfires. Wildfires have become a very common phenomenon. The recent Bushfires in Australia are an example of the devastative power of the calamity. Increase in temperature causes reduction in rainfall which can put out these fires.

Looking at all the effects of increasing temperatures, it is imperative that we find a solution to decrease the rate of increasing temperature and finally find a method to maintain it.

Resources:

1. Hardware/ Software requirement:
   * The project would require a machine with at least 8GB of RAM and enough memory to download data and require software
   * The software required to complete the project is Tableau, Anaconda to run Python, and Weca.
2. Data sources:
   * The data is freely available on UCI machine learning repository for the year 2004-2005
   * The data is stored in CSV format
   * Current temperature data may be needed to be purchased if model provides promising results

Requirements:

1. There are no security or legal restrictions on the data or the project results
2. There are no special requirements for result deployment

Assumptions:

1. Data quality assumptions:
   * The data has been collected in two different methods.
   * We will assume the data has been collected fairly and has not be tampered with.
   * We will assume the data has been merged correctly.
2. Results:
   * The results will be provided as an explanation of the data mining model as well as a potential solution to the problem in focus
3. Constraints:
   * The data is freely available and is not password protected
   * There are no legal constraints for the data

Risks and contingencies:

1. Data quality:
   * The risk here is that the data quality may not be adequate to achieve the desired results
   * In such a case, other data sources will have to be identified which would be consistent with the business objective
2. Result expectation:
   * This risk is that the result may not be as expected
   * In this situation, we would have to look into other factors, not considered in the project, which could affect the temperature
   * Addressing a problem as huge as temperature change, we may not come across a result where controlling a single particle in the air could fix the problem. There will be other confounding factors which would have not been considered. The main objective is to find out which of the particles present in the air in major proportions, affect the temperature.
   1. **Determine Data Mining objectives**

The objectives of the project are:

1. To use the data to build a data mining model:
   * The model will allow us to observe patterns in the data leading to temperature changes.
   * It will allow us to check associations between the pollutants in the air and the temperature at that instant.
   * We can figure out which pollutants affect the temperature the most.
2. Advise solution based on findings:
   * Identifying significant pollutants will allow us to device future actions
   * Decisions to controls pollutant can be made after identifying associations
   1. **Produce a project plan**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Time** | **Action** |
| Business Understanding | 1 week | Define the problem and requirements |
| Data Understanding | 1 week | Explore the raw data |
| Data Preparation | 1.5 weeks | Basic cleaning of raw data |
| Data Transformation | 0.5 weeks | Convert variables into required formats |
| Data Mining Method Selection | 1 week | Decide on regression/ categorization/ clustering |
| Data Mining Algorithm Selection | 1.5 weeks | Select the algorithm for building the model |
| Data Mining | 1.5 weeks | Build the model |
| Interpretation | 2 weeks | Explain role of variables in the model |

The project will be completed in 8 steps. The first step is the Business understanding. This includes understanding the use of the results of the project in actual scenarios. We assess the existing situation here, figure out our objective and explain how we can achieve a solution for the situation. This step is completed at the moment

The second step is Data understanding. The process of collecting the data is explained here. The raw data is explored and the fields in the data set are explained. The data is checked as to make sure there will be some significant results returned. This step is completed as well

The third step is Data preparation. In this step we start processing the data to make it suitable for analysis. The data is cleaned and combined to create a fresh and structured dataset which is useful for the analysis. This is the most important step as the result of the further step will depend on the quality of the data achieved from this step. This step will be completed by the end of a week and half.

The fourth step is data transformation. This step will include combining of features to create more useful features. We may also perform transformations to normalize the data. The process in this step will depend upon the results from the previous step. This step will be completed in another half week

The fifth step is selecting the method of data mining. After preparing and transforming data, we will decide to perform regression or classification depending on the format of the target variable. This step will be completed in one week’s time after data transformation

The sixth step is selecting the data mining algorithm. Once we have finalized the method for data mining, we will select the most relevant algorithm for the data. We will use multiple algorithms to compare the results of each one. This step will require one and a half week.

The next step is data mining. After preparing the data and finalizing the method and algorithm, we will finally run the algorithm on the data. We will try finding patterns and significant results from the analysis. This step will require one and a half week.

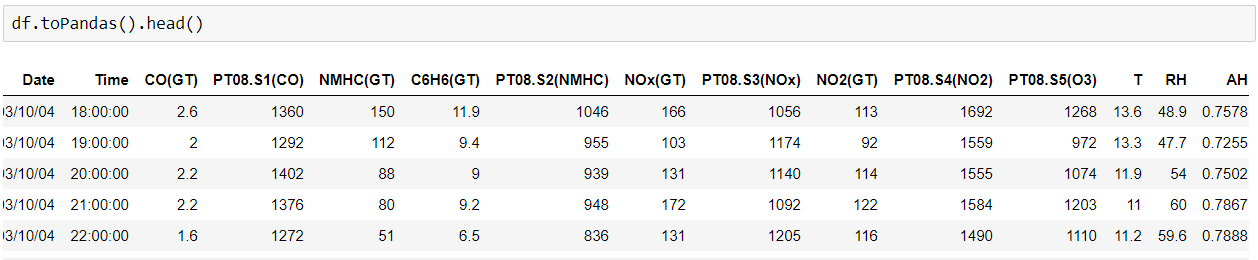
The final step is interpretation. This will be the most useful step as this is where we will be able to explain the results and suggest a solution to the issue being addressed. This step will require two weeks.

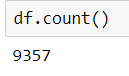
**2) Data understanding**

**2.1) Collect initial data**

The data is available from the UCI machine learning repository (Air Quality Data set, [4]). The data was collected by sensors embedded in an Air Quality Multisensor Device. The device was installed on the field, in a city in Italy, at road level, in a significantly polluted area. The data was recorded for a year between March 2004 and February 2005. The device recorded the concentration of 5 metal oxides. Ground truth averaged concentrations for CO, Non-Metallic Hydrocarbons, Benzene, Total Nitrogen Oxides and Nitrogen dioxide were provided through a co-located reference certified analyzer. The missing values were tagged as -200.

**2.2) Describe the data**





The raw data set contains 15 columns and 9357 rows. The columns of the data set are explained below.

Date: The first column is the date column. It is in the MM/DD/YY format. The dates start from 03/10/04 and end at 04/04/05.

Time: The next column is the time column. It is in the HH:MM:SS format. The is recorded periodically at every hour.

CO(GT): The next column is the true hourly averaged concentration of carbon monoxide in . This reading is collected from the reference analyzer

PT08.S1(CO): The next column is the hourly averaged sensor response for tix oxide. This targets the CO concentration. This is measure by the installed sensor

NHMD(GT): The next column is the true hourly averaged overall Non-Metallic Hydrocarbons concentration in . This is measured by the reference analyzer.

C6H6(GT): The next column is true hourly averaged Benzene concentration in . This is measure by the reference analyzer

PT08.S2(NMHC): The next column is the hourly averaged sensor response for Titania. This targets the NMHC concentration. This is collected by the installed sensor

NOx(GT): The next column is the true hourly averaged NOx concentration in ppb. NOx is the total Nitrogen Oxides. This is measured by the reference analyzer

PT08.S3(NOx): The next column is the hourly averaged sensor response for tungsten oxide. This targets the NOx concentration. This is measured by the installed sensor

NO2(GT): The next column is the true hourly averaged Nitrogen dioxide concentration in . This is collected by the reference analyzer

PT08.S4(NO2): The next column is the hourly averaged sensor response for tungsten oxide. This targets the NO2 concentration. This is measured by the installed sensor.

PT08.S5(O3): The next column is the hourly averaged sensor response for indium oxide. This targets the Ozone concentration. This is measured by the installed sensor.

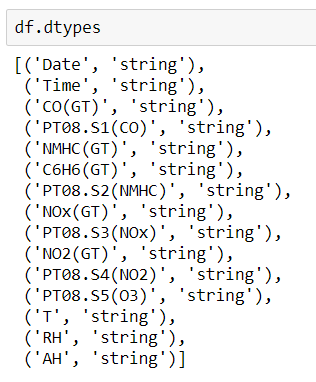
T: The next column is the temperature in

RH: The next column is the Relative humidity expressed as a percentage.

AH: The final column is the absolute humidity

**2.3) Explore the data**

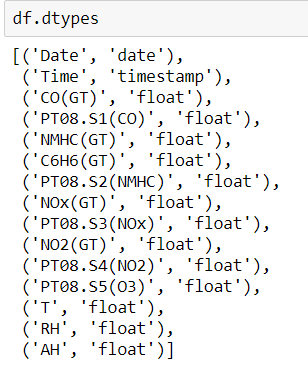
We will now look at each variable separately and try to understand the distribution of each of them. This exploration will be done on the raw data to identify errors and irregularities in the data.



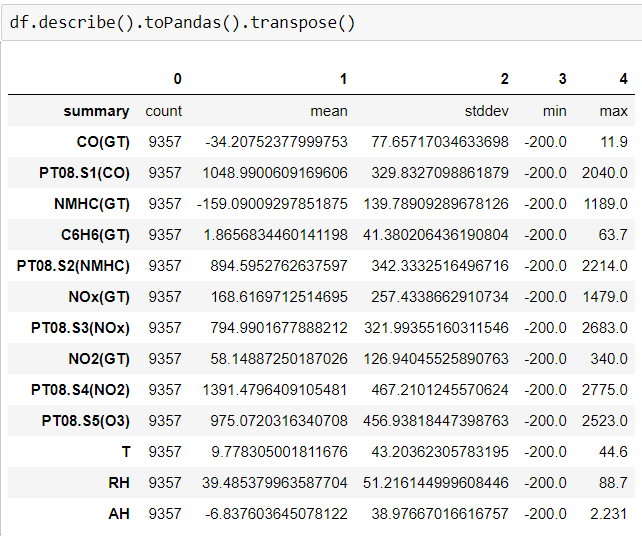
The date variable is in the string type variable. We will be treating it as a categorical variable and each date represent a category.

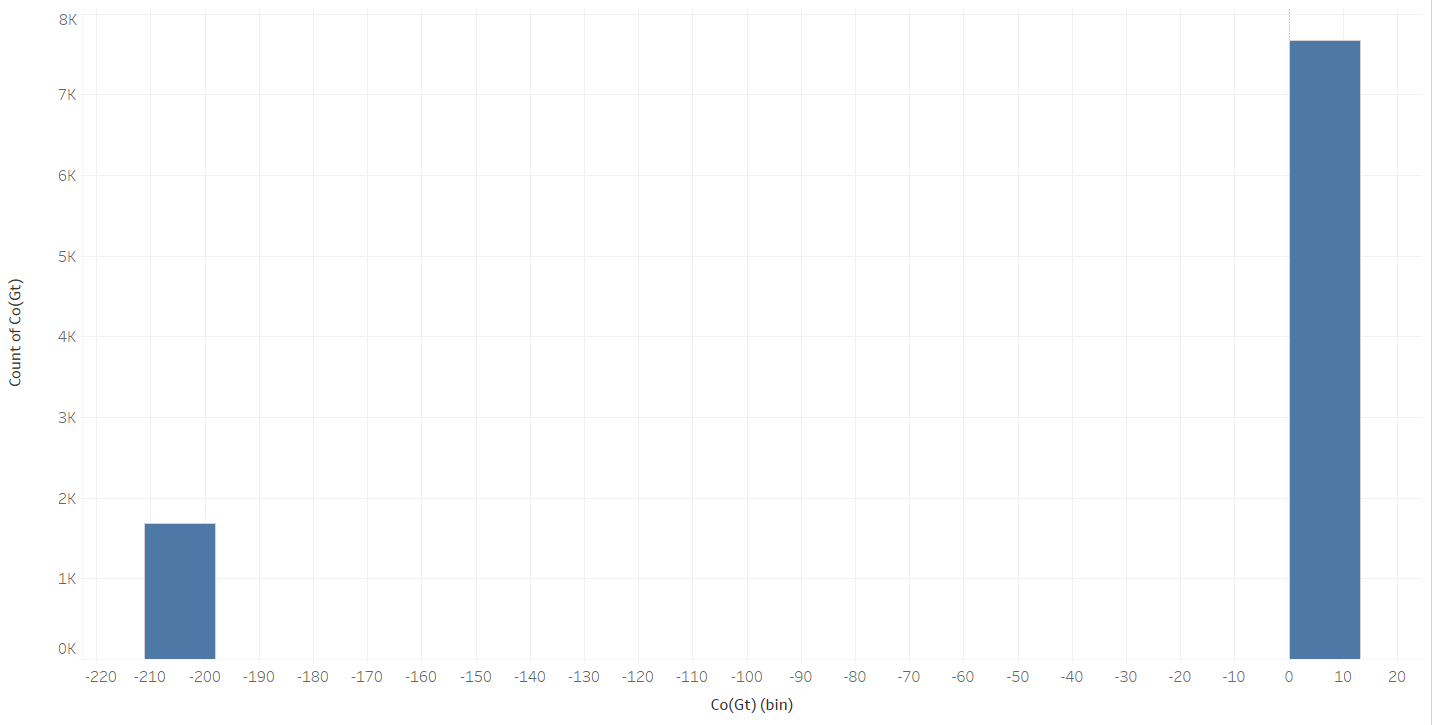
The time variable is again string type variable. We will treat this as a categorical variable as well. Since the data is recorded hourly, there will be 24 different categories for this variable.

The remaining all variables must be continuous and we will perform the correction for the data types

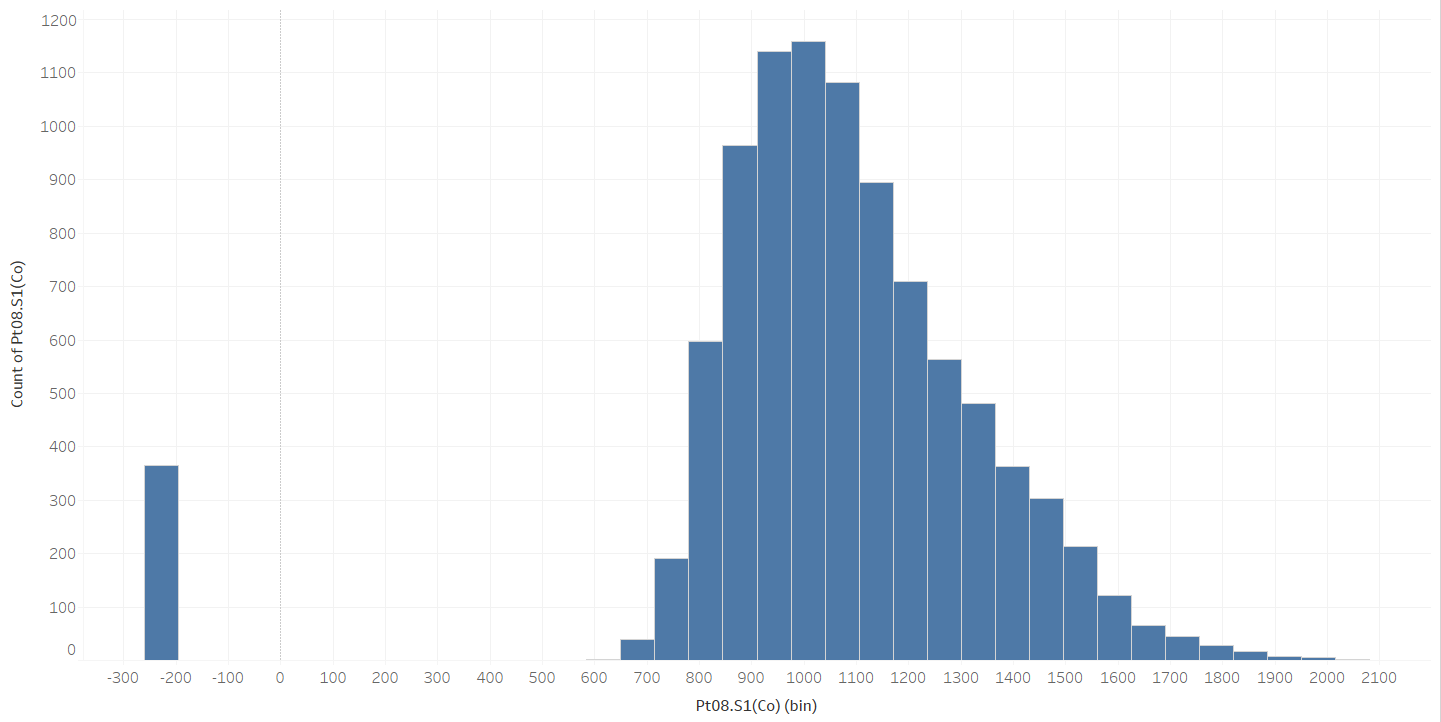


We can show some statistics for the continuous variables

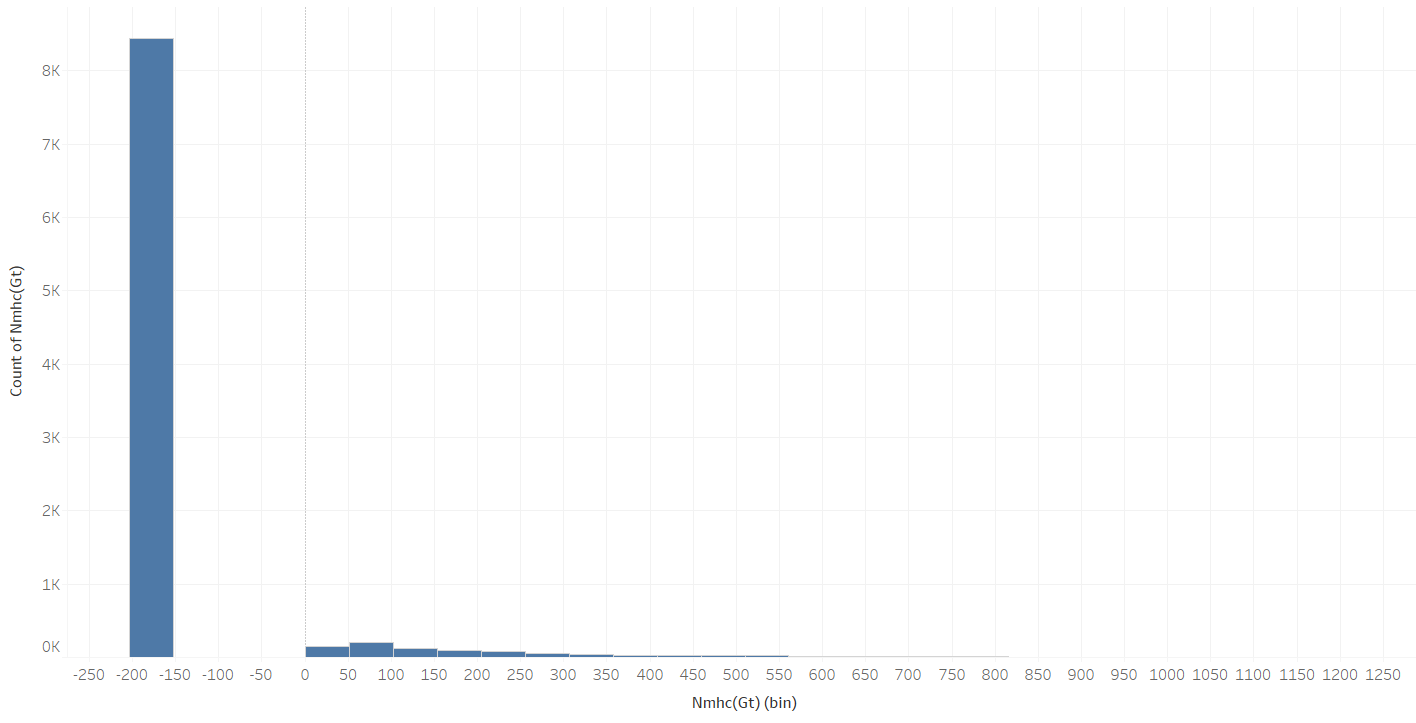




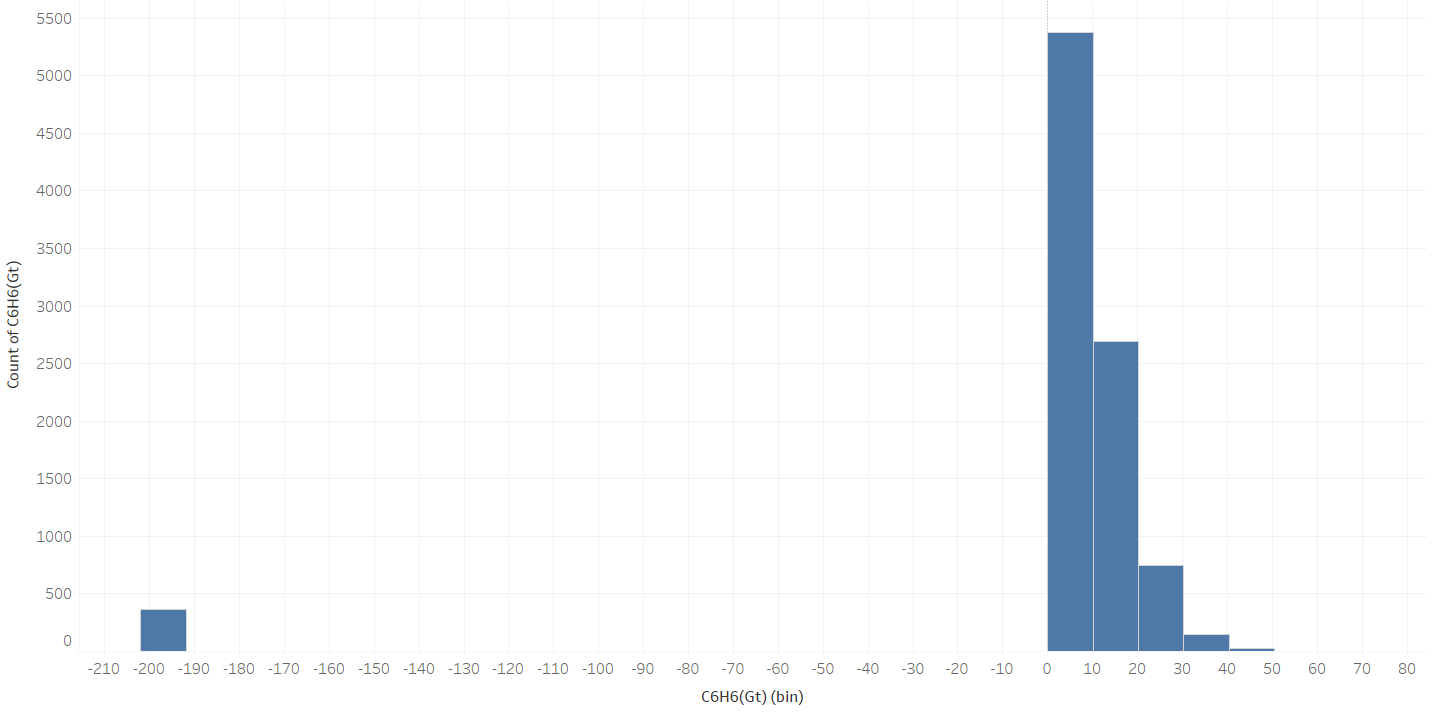
The CO(GT) variable has most of the values concentrated between 0 and 10. There are a significant number of missing values at -200, which will have to be addressed. The presence of the -200 values n the data set makes the mean of the variable negative and the mean should come back to the positive scale once the missing values are addressed.



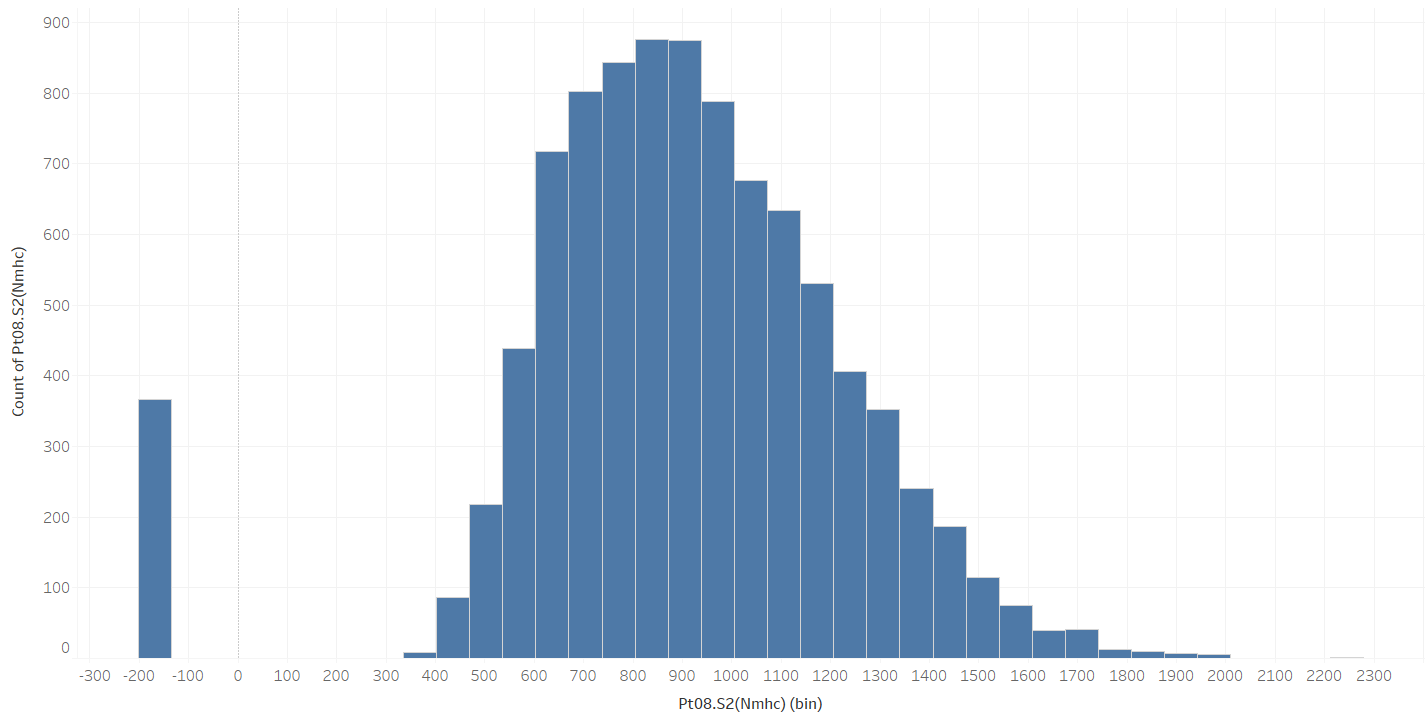
The PT08.S1(CO) variable is fairly normally distributed about 1000. There are few missing values marked by -200 which will be needed to be addressed. The missing values are not affecting the distribution significantly and we can expect the mean to stay about the same close to 1000.



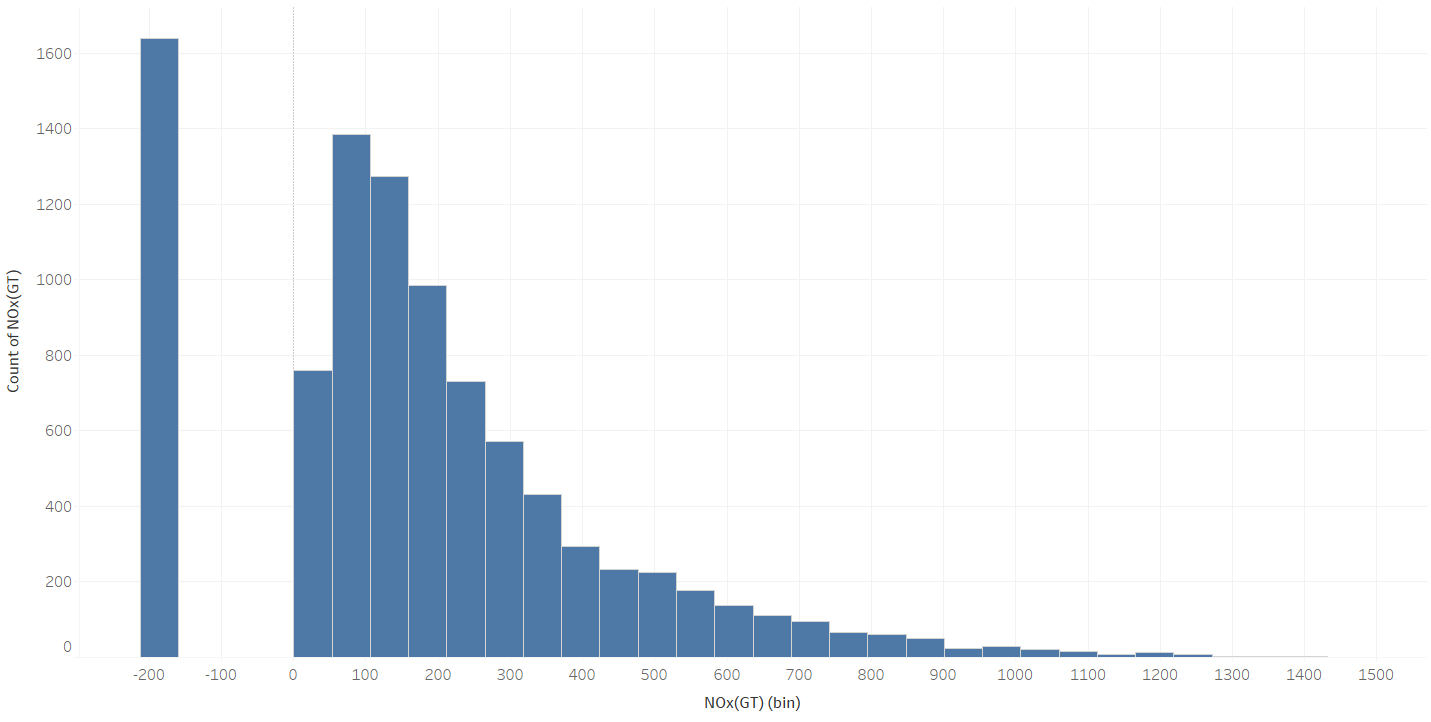
The NMHC(GT) variable in in the first instance does not come across to be very useful. There are too many values missing and too few values recorded with data. This variable may have to be dropped for the final modelling.



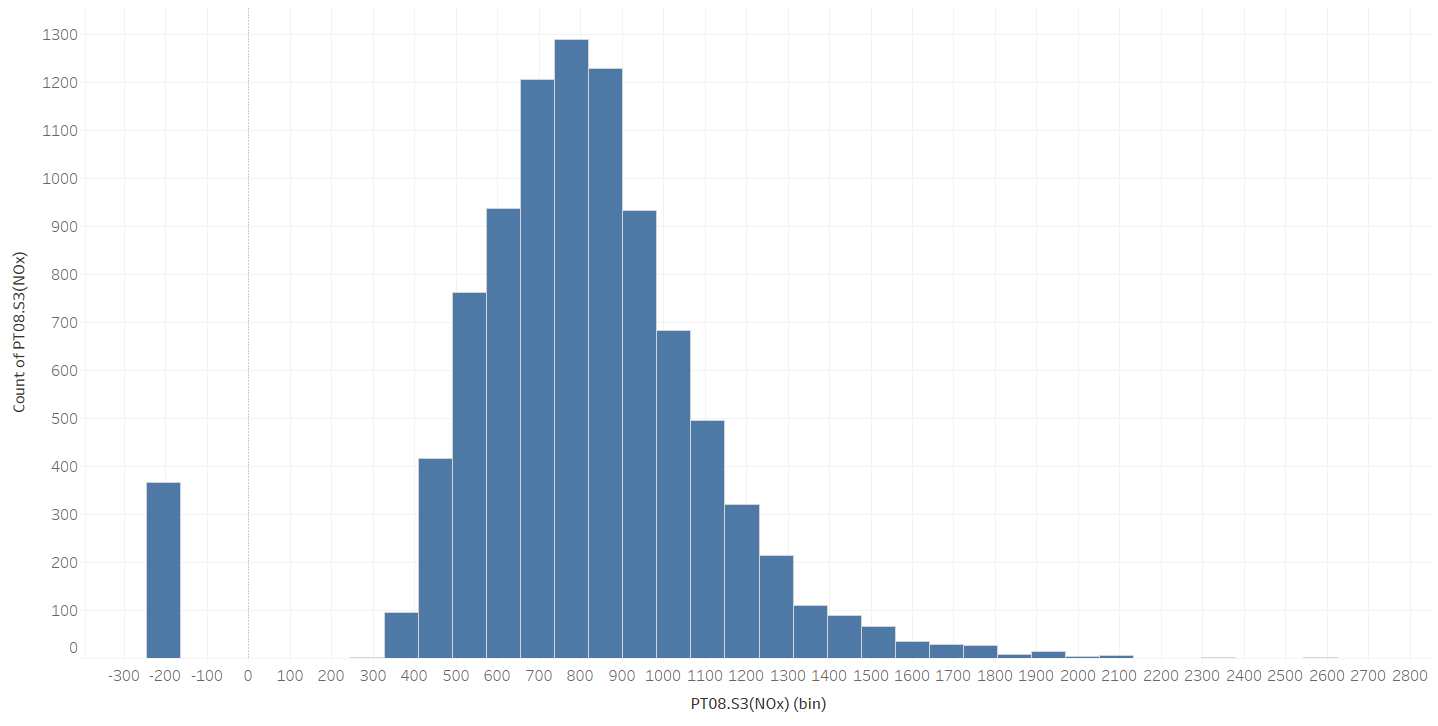
The C6H6(GT) variable has most of the data between 0 and 50. There are a few missing values that will need to be addressed. The mean I currently close to 2 but may move further ahead after addressing the missing values. This data is fairly normal but may need transformation to get it more appropriate.



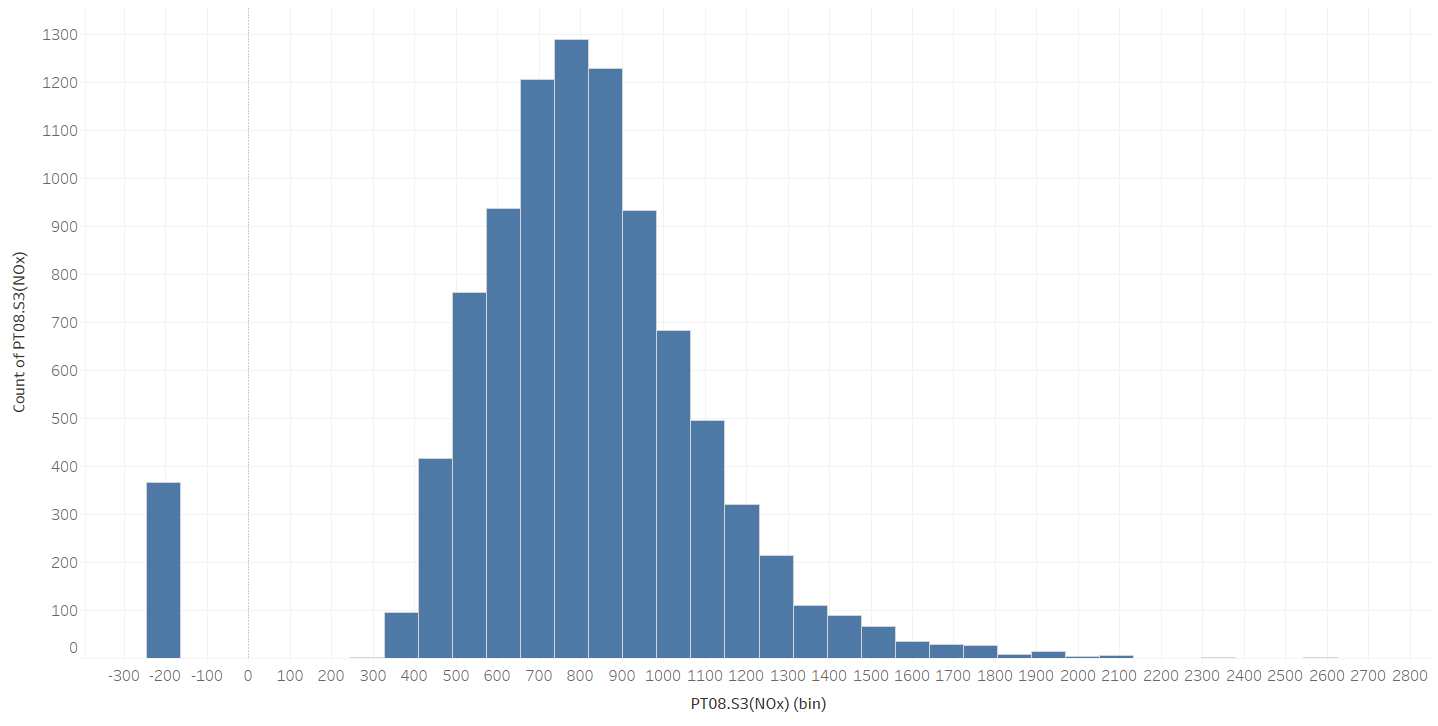
The PT08.S2(NMHC) variable is normally distributed. There are a few missing values which shall be addressed further. The mean which is close to 900 shall not be affected much after addressing the missing values.



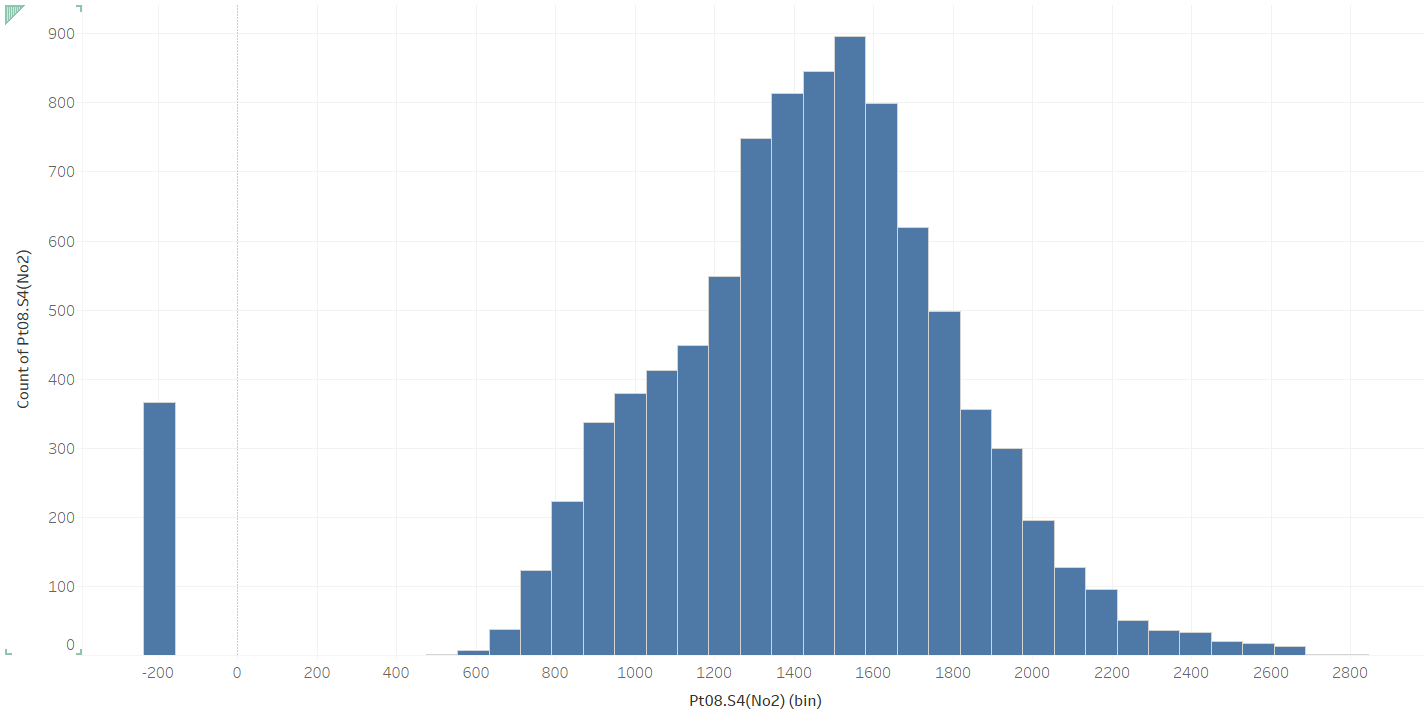
The NOx(GT) variable is fairly normal. There are quite a few missing values that will have to be addressed. The number of missing values is fairly high so addressing them would increase the mean of the data. The data is left skewed and some transformation may be needed to be performed.



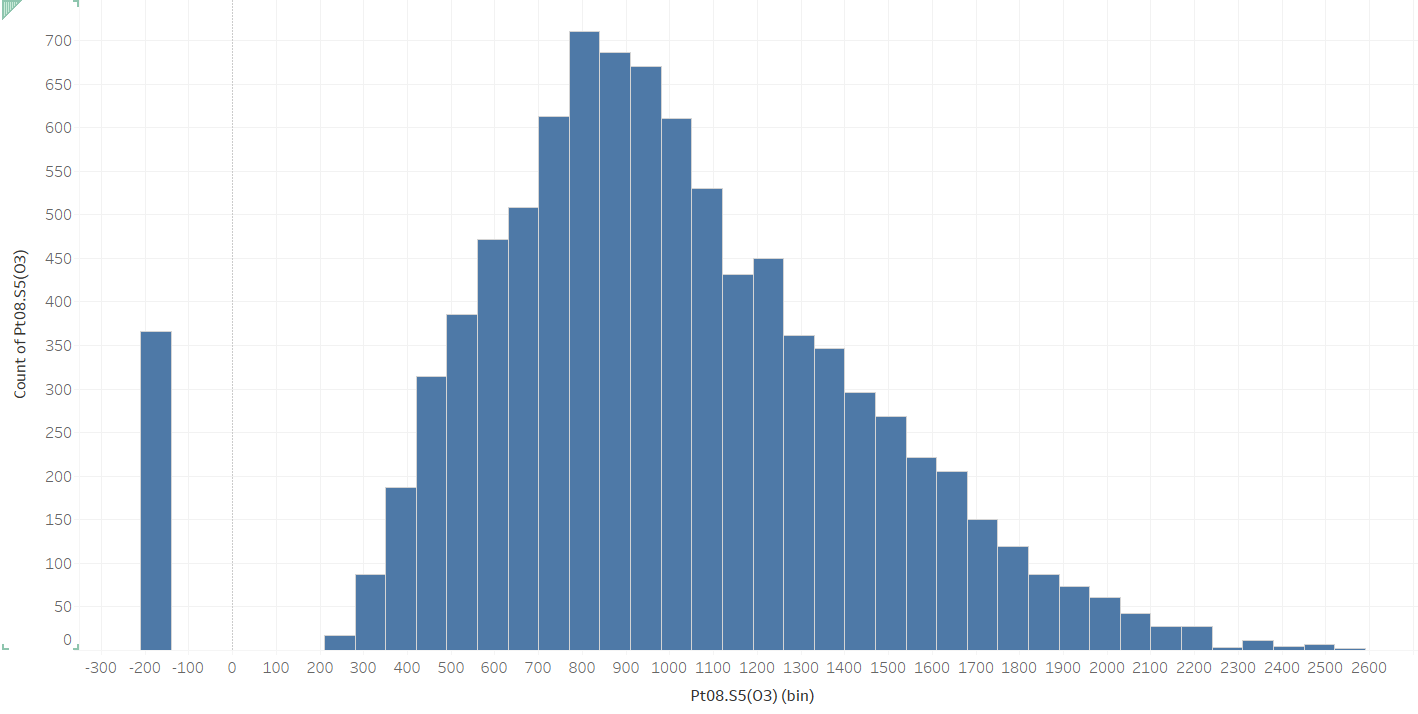
The PT08.S3(NOx) variable is normally distributed with some missing values at -200. These values will have to be addressed. The mean is close to 800 and is expected to stay so after addressing the missing values.



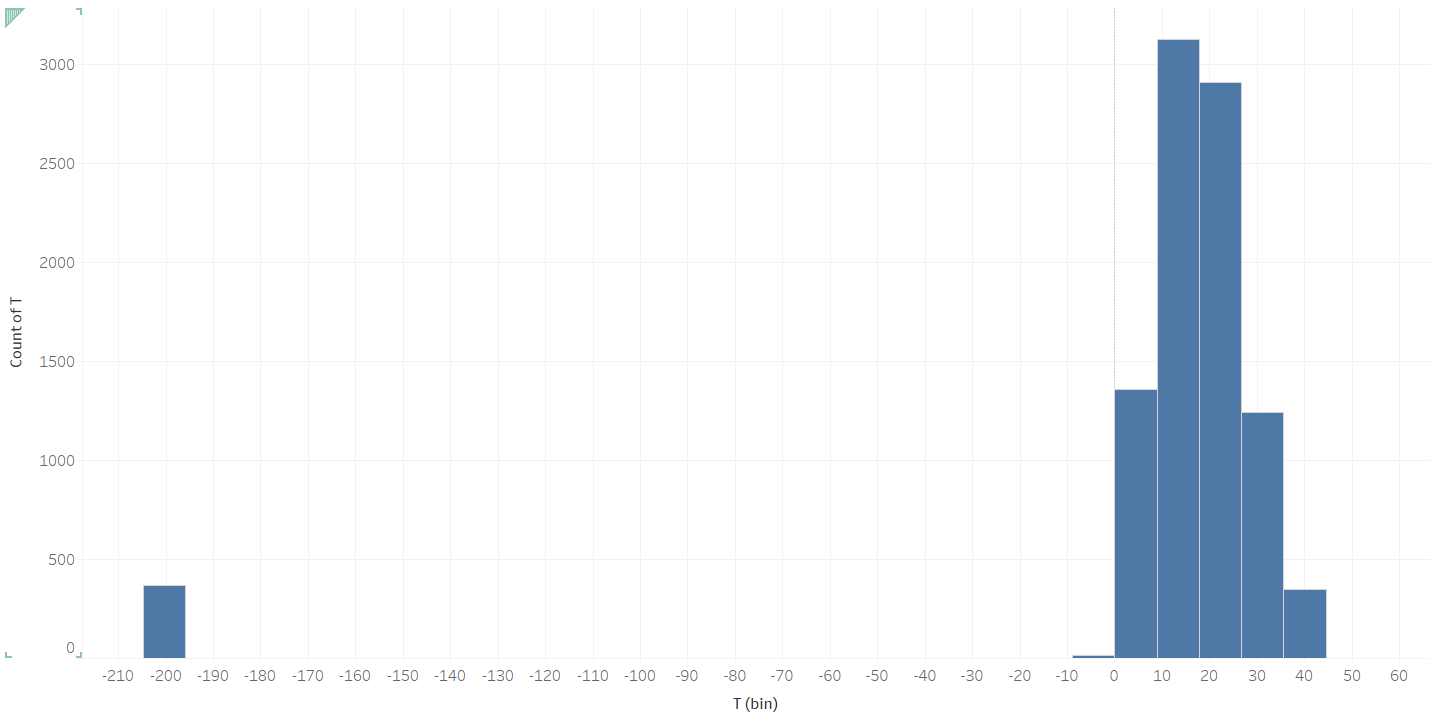
The NO2(GT) variable is normally distributed with quite a few missing values at -200. The mean is close to 60 but should move closer to 100 after addressing the missing values



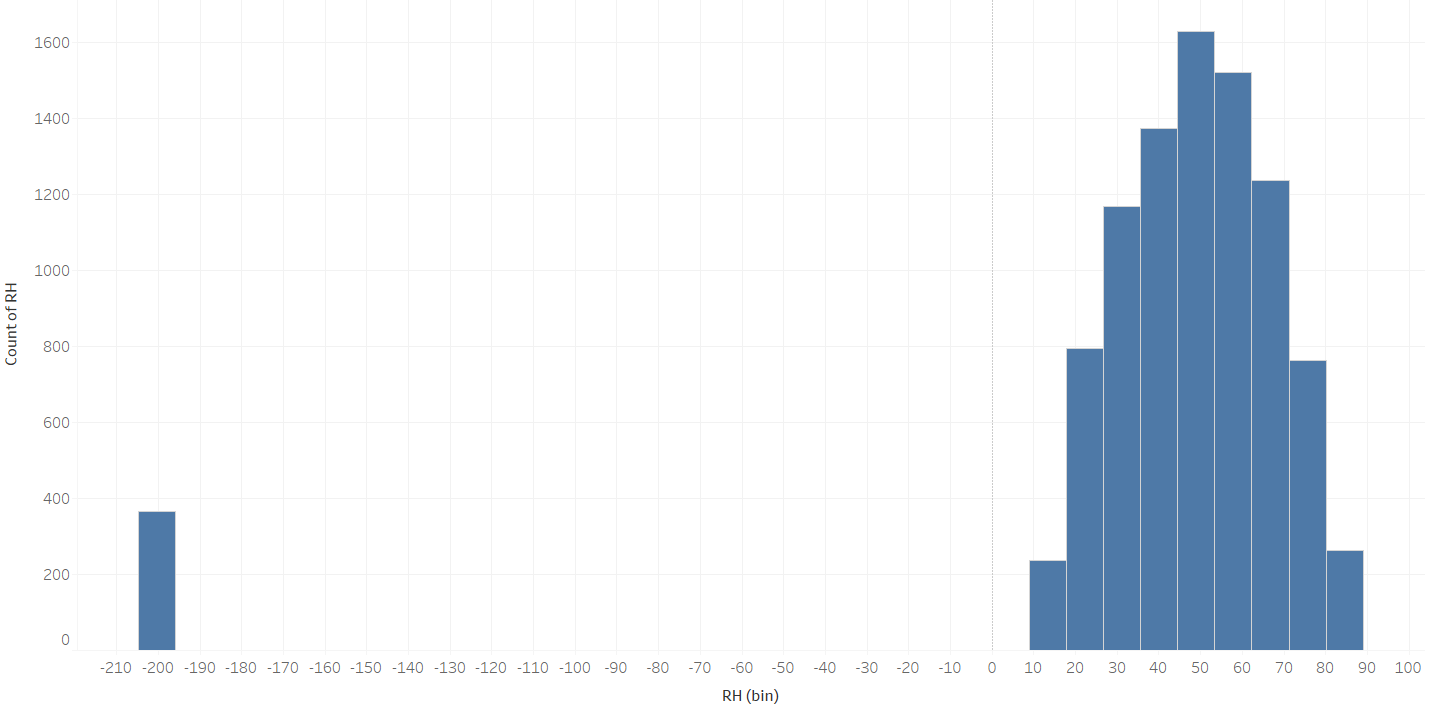
The PT08.S4(NO2) variable is normally distributed. There are a few missing values. The mean is currently at 1391 and should increase slightly without the missing values.



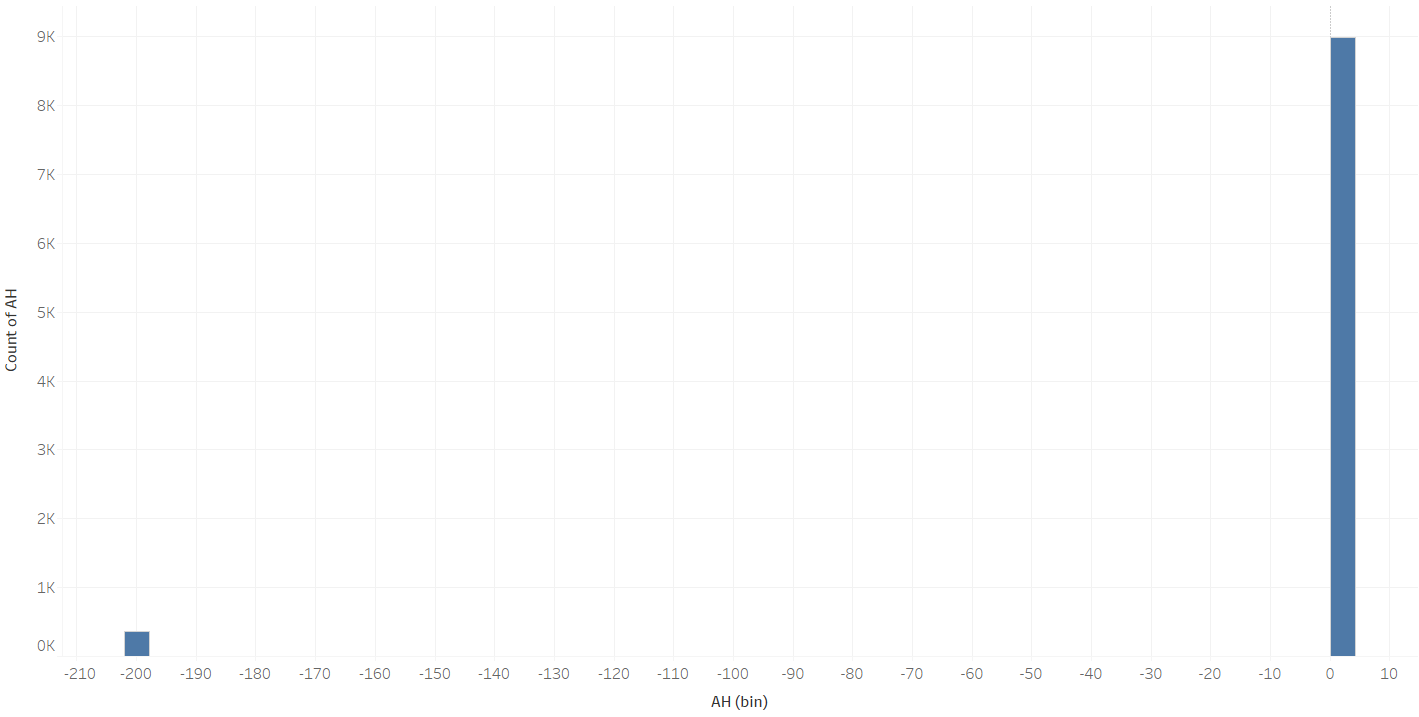
The PT08.S5(O3) variable is normally distributed with some left skewness. There are a few missing values. A transformation should help in making the data more normal.



T is out target varaible. It is normally distributed. There are few missing values.

` 

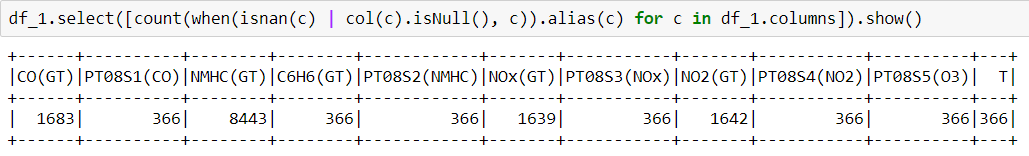
The RH variable is normally distributed as well. There are very few missing values for this variable



The AH variable is distributed very close to 0. There are very few missing values.

**2.4) Verify the data quality**

The histograms of the variables have provided evidence that most of the variables are normally distributed. The other variables can be normalized by performing some transformation. There are -200 values in the data which are basically missing values in the data, that will have to be addressed. The number of values missing in the data set is not too large. We can either impute these or decide to removed these from the data. A decision will have to be taken for the NMHC(GT) variable. This variable has most values missing and will need to be dropped.



After exploring the data, we can verify that the data quality is good enough to proceed to the next step. The data seems to be good enough to provide us a result for the issue being addressed.

**3) Data preparation**

**3.1) Select the Data**

For the purpose of easy identification, the -200 values in the data set implying missing values have been converted to ‘Nans’. Our software can now identify these as missing values.

Selecting rows:

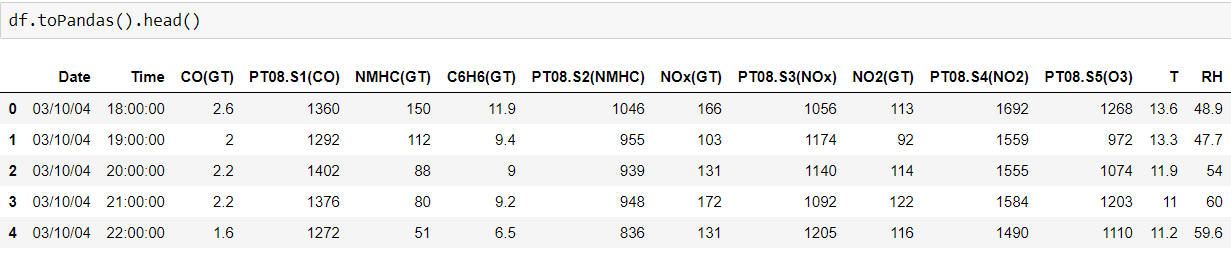
The raw data consists of 9357 entries. We will be using all the entries from the data set. There are some rows which have most of the columns with data missing and this will be addresses. We will finally have a shorter data set once the data cleaning is done.

Selecting Columns:

The raw data set has 15 columns. The temperature column is set as the target. There are columns for absolute humidity and relative humidity. These factors may contribute to the change in temperature but cannot be controlled by humans as effectively to control temperature changes. Hence, these columns will not be useful for the purpose of our project. The columns have been dropped from the data frame and would not play any further role.

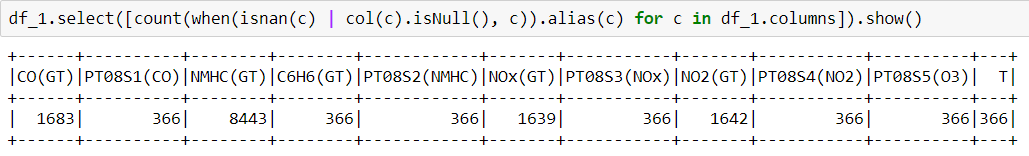
All other columns are essential in predicting the temperature so we will include the columns in further analysis.



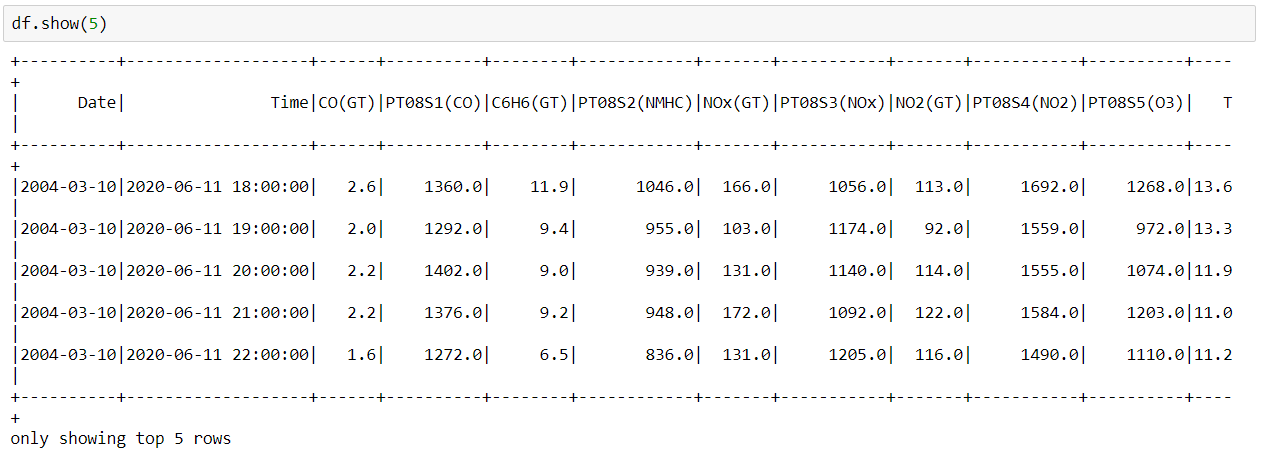
**3.2) Clean the Data**

Missing Data:

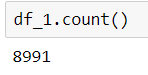
The first issue we will look at is that of the missing data. Checking the data frame for null values we can see that there are a considerable number of missing values.



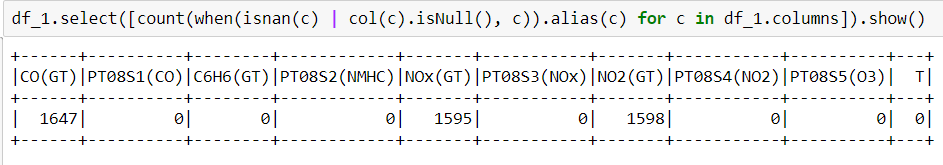
The NMHC(GT) variable has majority of the records missing. Imputing values to the variable, is not a good idea since it will introduce a lot of bias in the data set. With so many missing values in a single variable, it is not going to be very useful in predicting the temperature. It is better idea to remove the variable. We will just drop the variable.



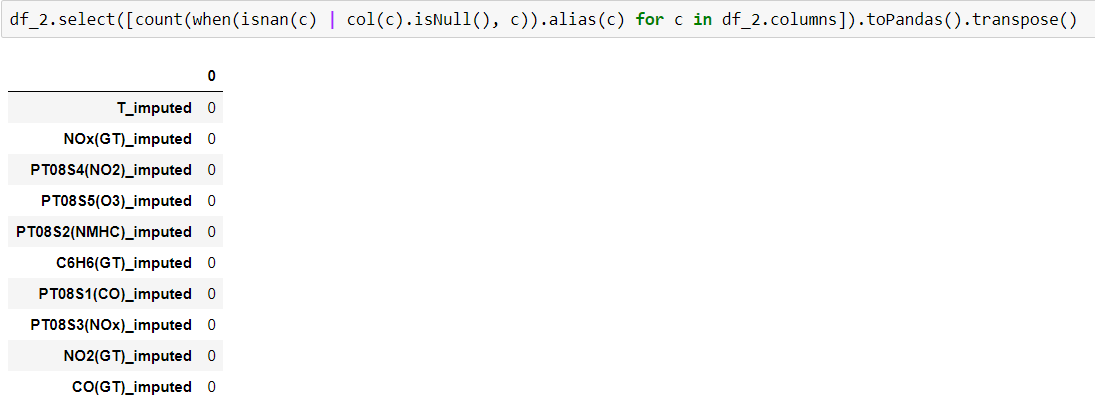
Next, there are other variables that have missing values that need to be addressed. Majority of the variables have missing values for the same rows. There are 366 rows were all the variables are missing entries. Imputing values at these rows will again introduce bias in the data set. We will drop these values by selectively dropping rows with missing values in only one variable.



The number of observations have now reduced to 8991. This has shortened the overall data set but has removed majority of the missing values.



There are only 3 variables now which have missing values. We will be imputing these values into the variables. We will be using the means of the columns for the imputation. The data after imputation has no missing values.

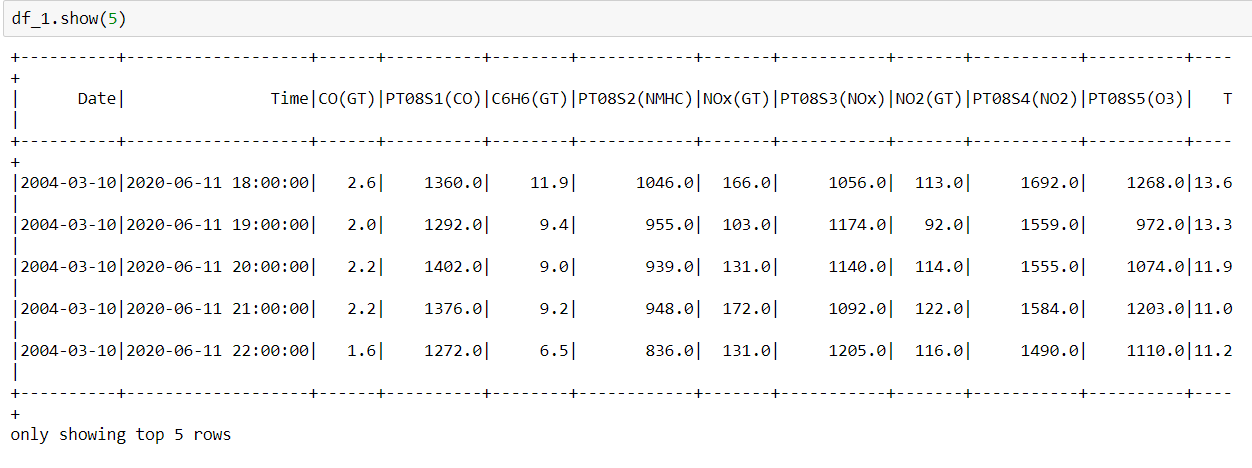


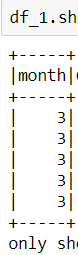
The rest of the data is good enough. There are no errors in the data. The final data set has 8991 observations. After performing all these steps, the data set has been cleaned. This means that there are no missing values. In the current form, the data is already usable for analysis, but we will perform some further feature engineering to the make the data more suitable for our purpose

**3.3) Construct the Data**

Deriving New attributes:

The Date field by itself is not very effective. It cannot be treated as a continuous variable and as a categorical variable, it has far too many categories. A better approach would be to use the month from the date variable. The temperature of a place is more or less similar during a particular month. The traits in temperature change can be tracked easily on a monthly basis. Hence, we have extracted the month from the date.

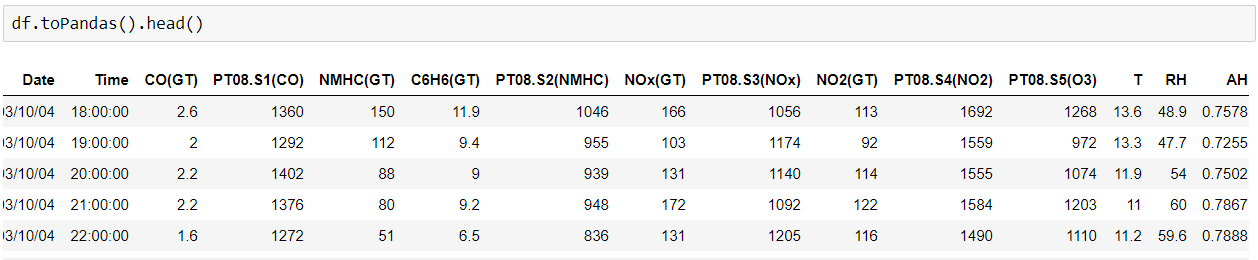


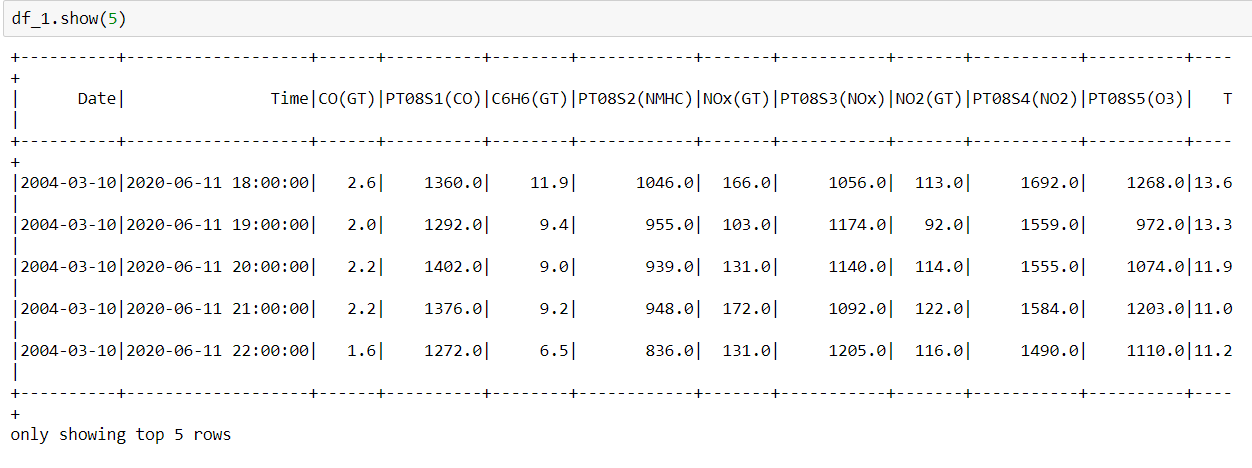


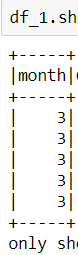
A Month variable with the months extracted and ordinal type with 12 levels is created in this step. We shall use this attribute for further analysis.

**3.4) Integrate various data sources**

The original data set provided on the UCI machine learning repository had used two different sources. One source was the data from the sensors installed in the city in Italy while the other was a certified analyzer co-located to the sensors. The data provided was already merged and has enough row and columns to build an effective model and answer our questions of interest. No further merging or appending of data is necessary. Although, we have created new variables which have been added to the data frame. The new data set contains one variable more than the number of variables in the original data set.

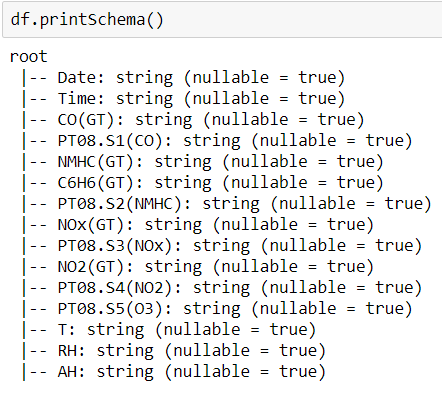




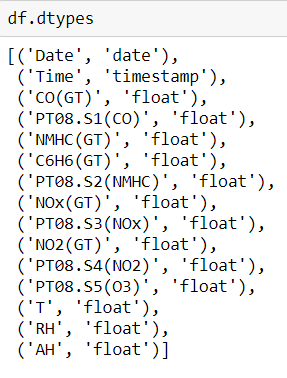


**3.5) Format the data as required**

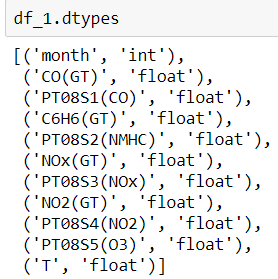
The data when read into Python had all the columns of the type string.



Here there are multiple variable that should be numeric for the purpose of our analysis. We will change the type of the variables to numeric in order to further analyze the data.



Once the month variable was added, it was of numeric type and had to be converted to categorical.

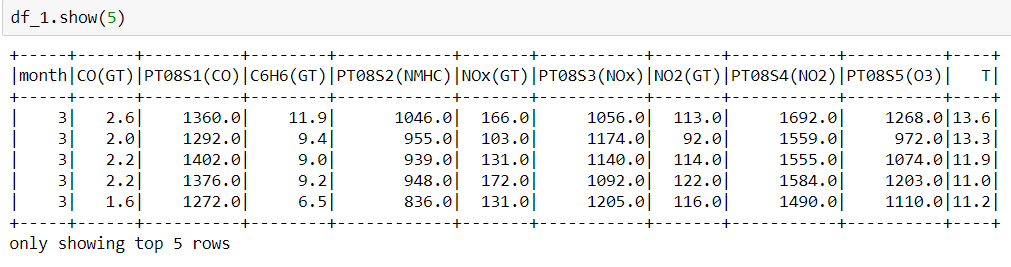




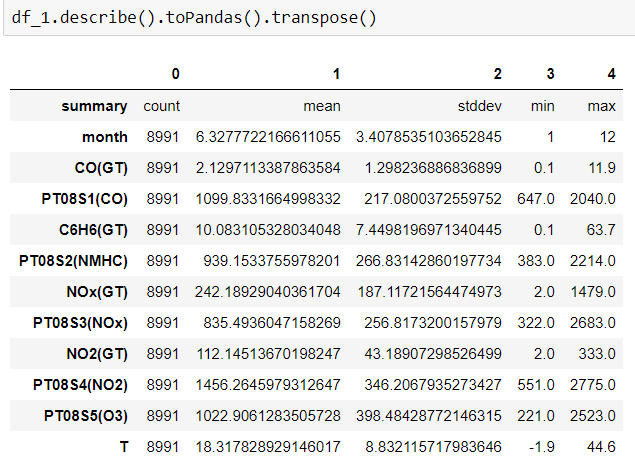
**4) Data Transformation**

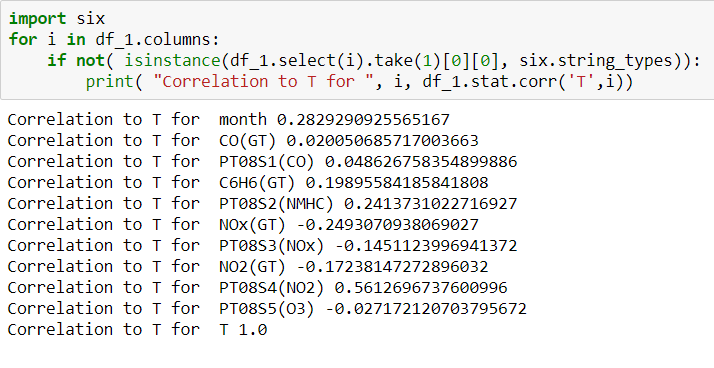
**4.1) Reduce the Data**

In the previous steps, we have created the month variable from the date variable. The months would be playing a larger part in the temperature prediction than the date. Once we have the month variable the date variable is not required and including the date variable will also cause variance inflation since the month is derived from the date variable. Hence the date variable will be needed to be removed. The time variable as well is not important since we are not interested in the effect the time has on the temperature. This variable will be removed as well.



The data set after reduction has 11 columns and 8967 rows.

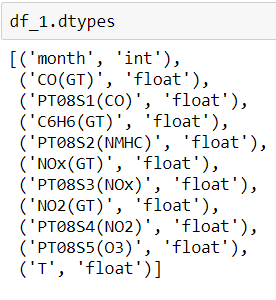
We will now check the correlation of each variable with the temperature to perform feature selection



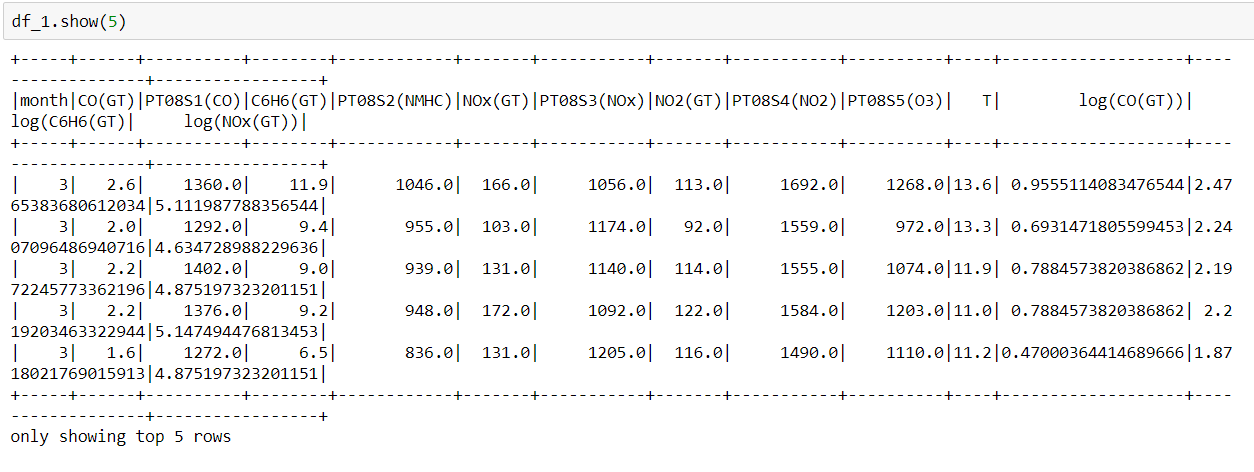
Looking at the values above there is no strong evidence to eliminate any variable for the purpose of model building. All the current variables contribute to the target which is temperature. So, we will not reduce the data further.

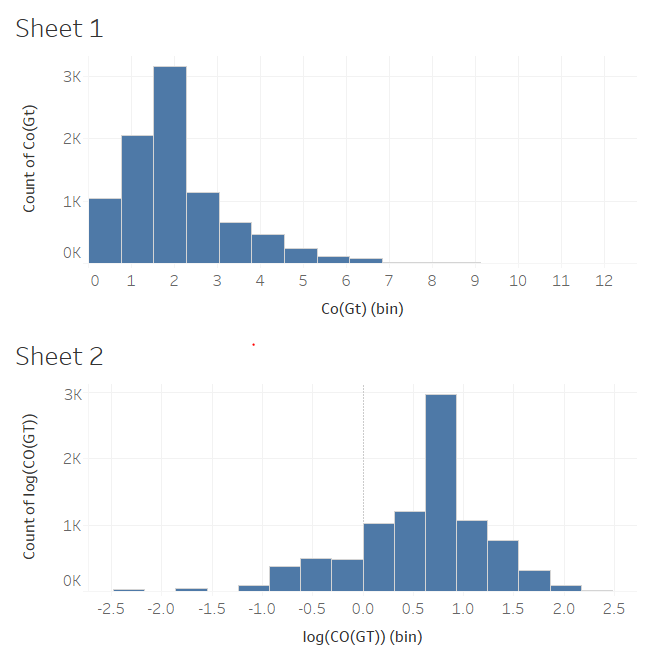
Next, we will look at the distributions of the variables in the cleaned data set and decide whether any transformation is necessary

**4.2) Project the Data**

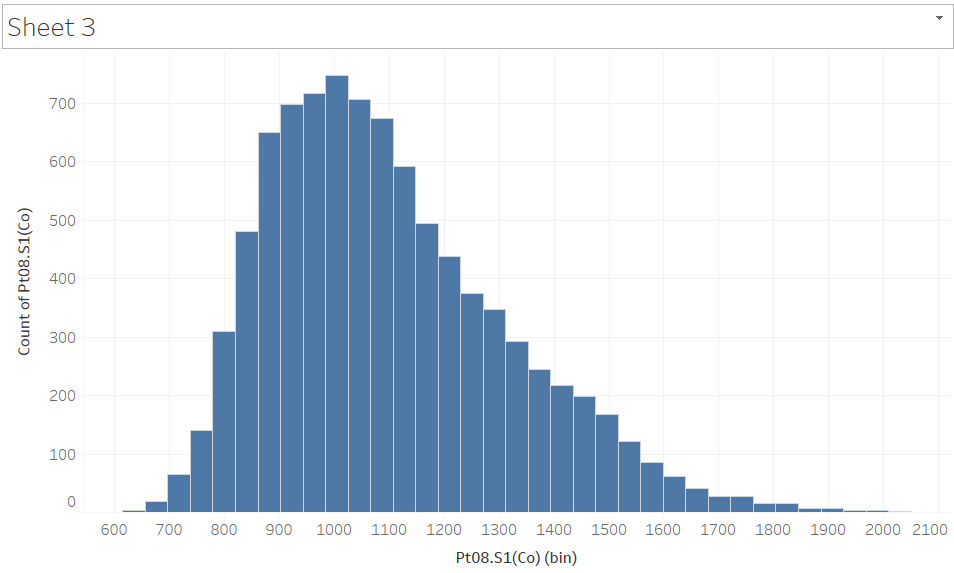


The data set now has 10 continuous variable and one ordinal variable. Of the 10, temperature is our target and the others are out input. We will see every variable individually to decide if transformations are necessary. The aim is to bring the distribution of the variable as close to the normal distribution as possible. Achieving this will satisfy assumptions needed for many data mining models.

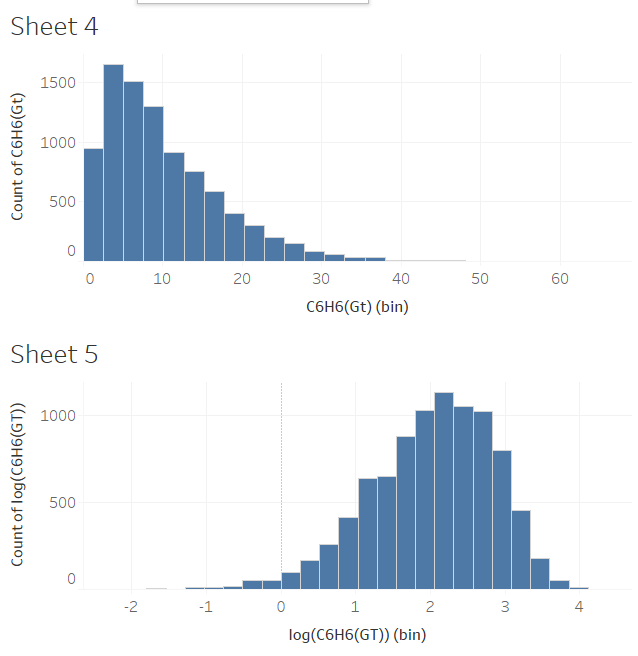




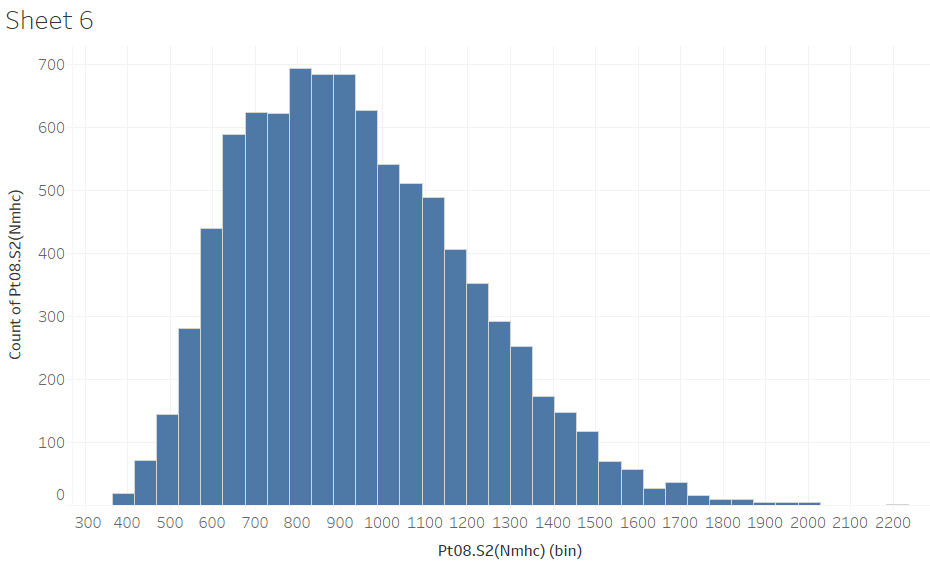
The CO(GT) variable is not exactlydistributed normally. The data is left skewed. Performing a log transformation on the variable makes the data much more normally distributed. The log transformed data would provide us better results that the original data. We will use this variable in the log transformed form.



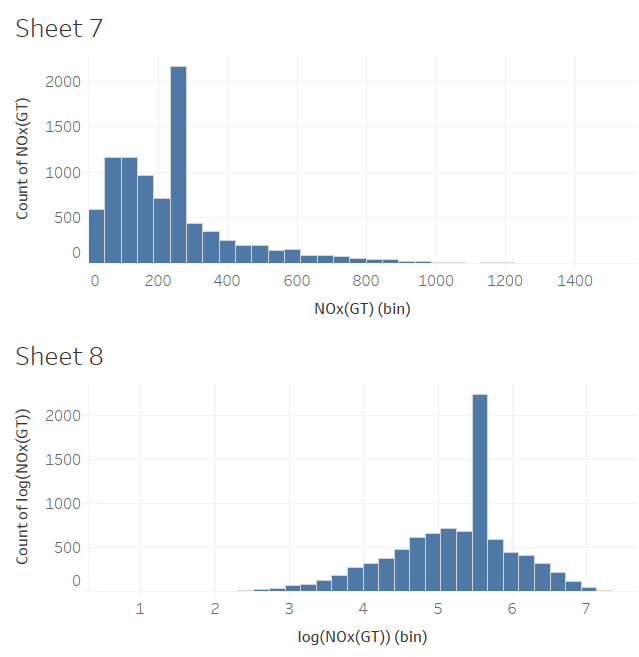
The PT08.S1(CO) variable is already distributed normally. The distribution does not vary too much. No transformation is necessary.



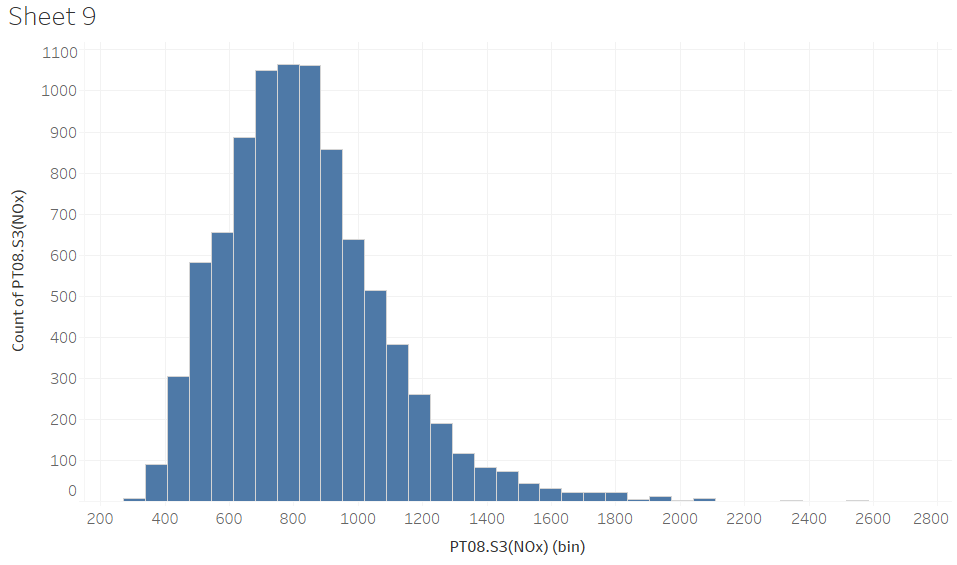
The C6H6(GT) variable in its original form is skewed. The distribution can be brought close to a normal distibution by performing a log transformation. This variable will be used in the log transformed form as well.

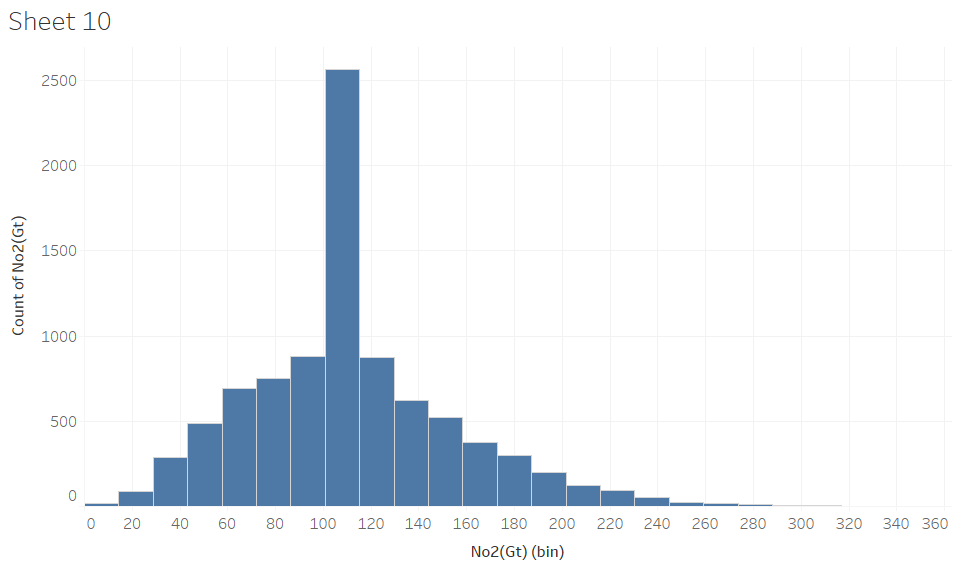


The PT08.S2(NMHC) variable is very close to the normal distribution in the original form and no transformation is necessary.

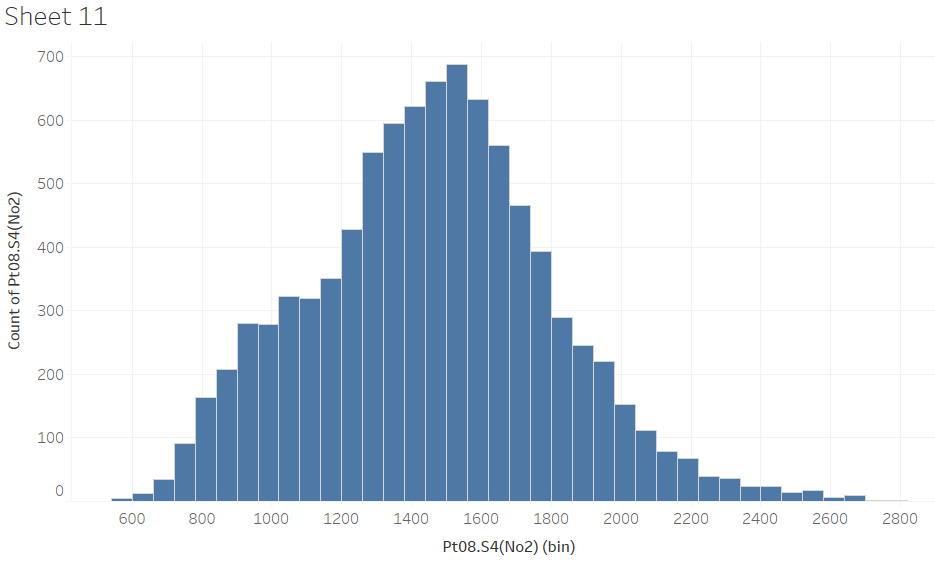


The NOx(GT) variable is not normally distributed. There is skewness in the data. This can be fixed with a log transformation. The transformation brings the distribution very close to the normal distribution.

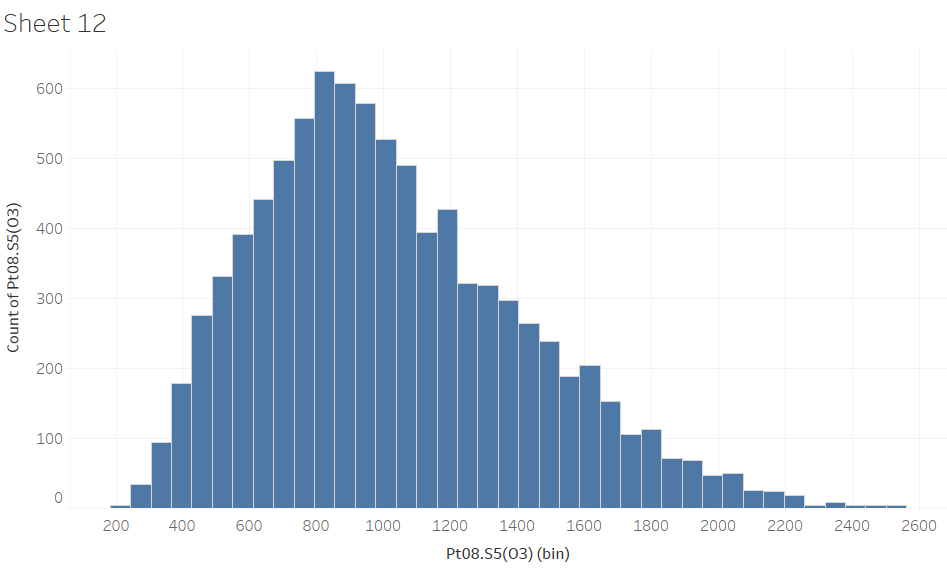
 The PT08.S3(NOx) variable is normally distributed in the original form and no transformation is necessary.



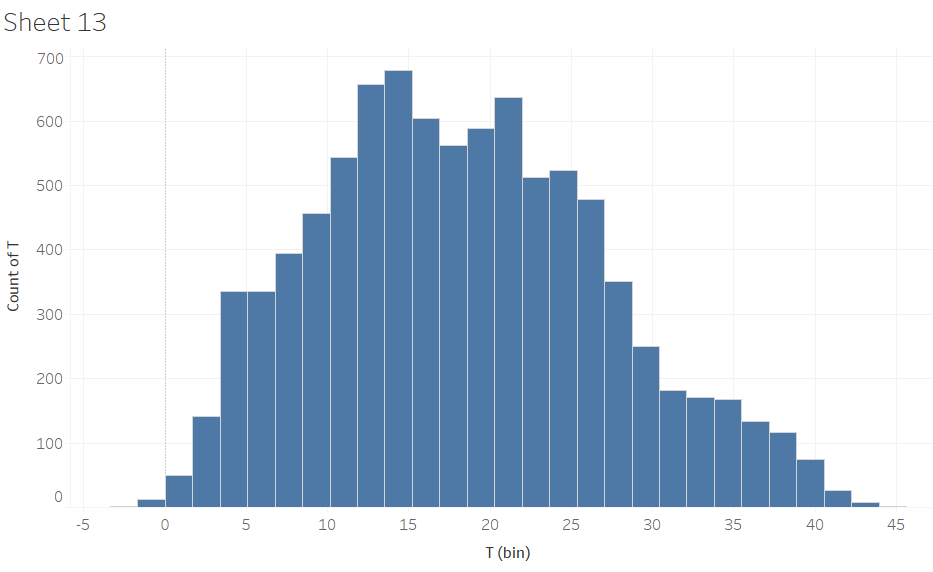
The NO2(GT) variable is close to the normal distribution and there is not strong evidence for deviation from normality. No transformation is necessary for the variable.



The PT08.S4(NO2) variable is normally distributed and no transformation is necessary.

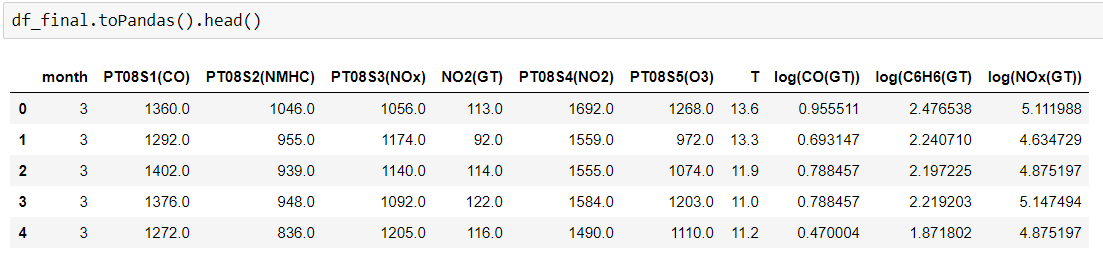


The PT08.S5(O3) variable is normally distributed and no transformation is necessary. There isn’t strong evidence of deviation from normality



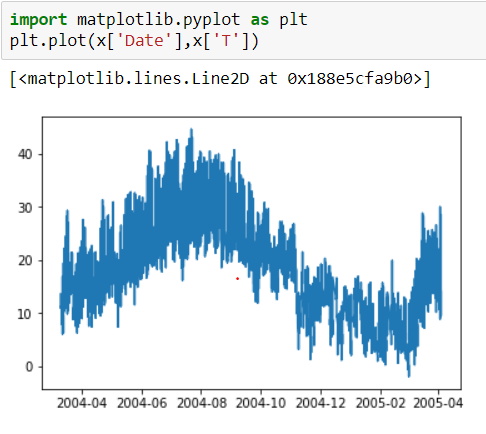
The target variable T is normally distributed, and no transformation is necessary.

In this step 3 new variables have been created and 3 existing variables will be dropped. The final data set still has 11 columns and 8991 rows.



**5) Data Mining Model Selection**

**5.1) Match and discuss the objectives of data mining (1.1) to data mining methods**



The target variable is the temperature variable (T). T is a continuous variable. We are interested in building a data mining model which will allow us to find the association between temperature and the pollutants. We want to see how the presence of a pollutant in the air affects the temperature of that place. We intend to build an interpretive data mining model. This model should help us see which pollutant affects the temperature the most and in which direction.

1. Clustering:

* This is a unsupervised learning method. This method is effective when the data is unlabeled. This means that the target is unknown (Altman, N., & Krzywinski, [8]).
* This method can be used to group similar data to find trends and patterns.
* The only use of this method is to gain insights from the data. There can be no associations made or any predictions done.
* Since we have a target variable in our data set and our aim is to derive association, this method is not useful for us.

1. Classification:
   * This is a supervised learning method. This method is effective when the data is labeled. This means the target variable is known (Dejaegher, [7]).
   * This method can be used when the target variable has limited discrete outcomes.
   * We can derive associations and make predictions based on this method
   * This method is used for risk analysis and decision analytics.
   * Our target variable is continuous, so this method will not be useful for us.
2. Regression:
   * This is the final data mining method and is a supervised learning method as well. The data has to be labeled (Fitzmaurice, Garrett M, [6]).
   * This method is used when the target variable is continuous.
   * This method can be used to derive associations and make predictions.
   * This method is used in predictive analytics applications.

This method looks to be the most appropriate for our purpose.

**5.2) Select the appropriate data-mining methods based on discussion**

We have discussed the various data mining methods available to us and how each one can be used. As per our requirements, regression would be the best fit for us (Fitzmaurice, Garrett M, [6]). Regression will help us derive the associations between the pollutants and the temperature. We can find out how a pollutant affects the temperature and check the collective effect of the pollutants as well. Hence, we will be moving forward with regression and discuss various algorithms available to us.

There are various factors which we can look at to decide the success of the model. We can look at the r-squared values, area under the ROC, or the mean squared error (Torres-Barrán, A., Alonso, &., & Dorronsoro, J, [9]). All these parameters can be used to judge the models. We shall select the mean squared error as the judging criteria for the model’s success. The lower the mean square error, the better the model will be.

**6) Data Mining Algorithm selection:**

**6.1) Conduct exploratory analysis and discuss**

In the previous step we have established that we will be using regression for the purpose of our project. The next step is the identify the algorithms available to us. There are various algorithms used for regression ranging for simple linear regression to Neural networks. We will discuss the algorithms and then decide which ones could be the best ones for us.

1. Multiple Linear regression:
   * Multiple linear regression uses several independent variables to predict a single dependent variable.
   * Multiple linear regression is the form:

* This the easiest model to interpret as it provides the direct relationship between the independent and dependent variables.
* However, in order to use this method, there are certain assumptions the data must satisfy.
* The assumptions are a linear relation between dependent and independent variables, Normality of the residuals, constant variance of the residual and independence of the observations.
* Another requirement is that the dependent variables should not be correlated. This would cause variance inflation.
* Multiple linear regression has various variable selection methods. The most basic selection methods are forward, backward, stepwise and all subset selection.
* While all subset selection is the method that can provide the best possible linear regression solution, it requires an enormous amount of computing power.
* Multiple linear regression also uses regularization to improve model performance. These regularization models are ridge regression, lasso and elastic net.

1. Classification and Regression trees:
   * Regression trees are algorithms based on decision trees
   * Decision trees are like if-else statements. Depending on a condition for a variable a split will occur, and the tree will expand.
   * Regression trees occur when the target is continuous as in our case.
2. Random Forest:
   * Random forest is a method derived from Bagging (Bootstrap Aggregation)
   * In this method multiple data sets are created by bootstrapping. This is basically sampling from our existing data set with replacement. Hence, we have multiple data sets with the same length as the original data set.
   * A separate decision tree is formed for every data set created and the results for all the trees are aggregated.
   * The issue here is that since all variables are used at the nodes, in case of a single influential variable, the first split always occurs at this variable. This causes the trees to be correlated and increases the variance
   * In case of Random forest, at every node ‘m’ number of variables are randomly selected at each split. Now the trees are different from each other and there is less correlation.
   * This algorithm requires the data to be independent and is not concerned with the linearity of the data.
3. Gradient Boosting:
   * This algorithm is an ensemble algorithm which has is similar to the CART trees.
   * The idea of boosting is to make a weak learner into a better learner.
   * Gradient boosting has a loss function to optimize, a weak learner to make predictions and an additive model to add weak learners to minimize the loss function.
   * Gradient boosting can be enhanced by introducing tree constraints, updating the weights, random sampling of data and penalized learning.
   * The assumptions of gradient boosting are similar to that of random forest.

Without actually building a model with every algorithm, it is impossible to tell which one will work the best. There are many criteria or judging a model, but we will go ahead with the mean absolute error for the model on a test data set. This implies the error in prediction a model has when it comes across previously unseen data. This criterion is extremely useful in the real world scenario since it is always impossible to tell the distribution of the test data before it is actually collected. Building a model which has the ability to perform well on previously unseen data helps us to deal with this issue.

**6.2) Select Data Mining algorithm based on Discussion**

We will go forward with the above-mentioned models as each model has its own advantage.

1. The Multiple Linear Regression model will be extremely easy to interpret and in case of a linear relationship, this would prove to be the best model of all 4. We will apply the stepwise selection algorithm to check if any of the variables can be dropped
2. The CART is a simpler tree-based algorithm, easy to interpret and needs lesser assumptions satisfied. This will be more robust than Linear regression
3. The Random Forest, being an ensemble algorithm, has the ability to minimize the error and at the same time a relatively easy to interpret. We will set the algorithm to build and combine 10 trees.
4. The Neural network, being the most difficult to interpret, would be our only choice in case all other algorithms fail. This algorithm is free of any assumption requirement and the hyperparameter tuning makes it robust to different types of data. We will build a multilayer perceptron with one hidden layer.

We will assess the models based on the average absolute error on the testing data. This will give us an estimate of the rick of a particular model. The model that has the lowest risk will be our final choice.

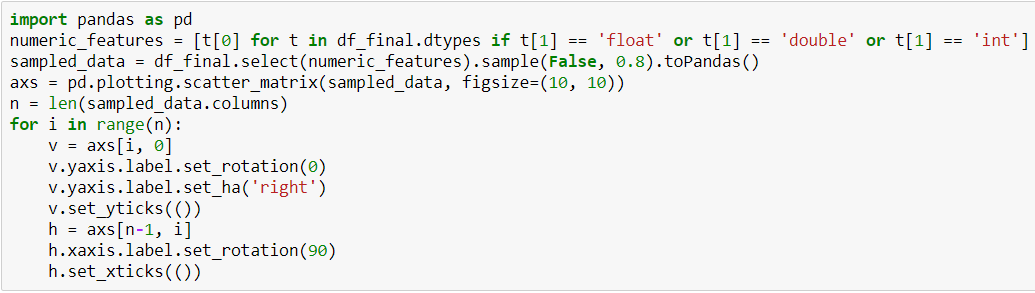
**6.3) Build/select appropriate models and choose relevant parameters**

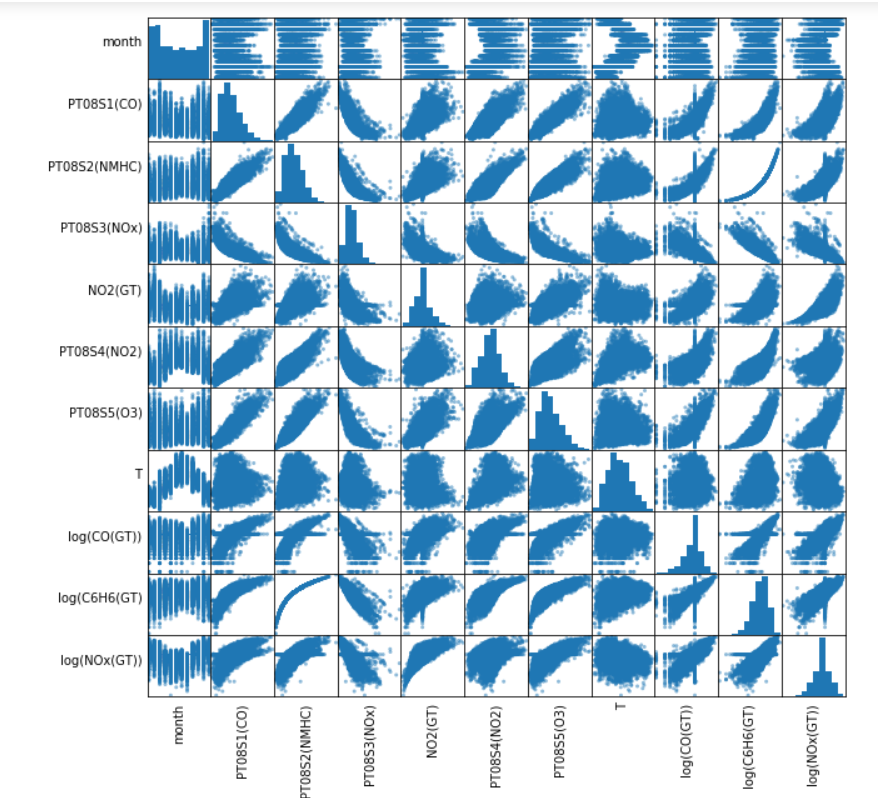
We will now proceed to check the requirements and results from each of the about discussed algorithm. This will then allow us to select the best algorithm for our purpose.

1) Multiple Linear regression:

The first criteria to be satisfied to proceed to multiple linear regression is that the assumptions have to be satisfied. We will take a look at the assumptions.

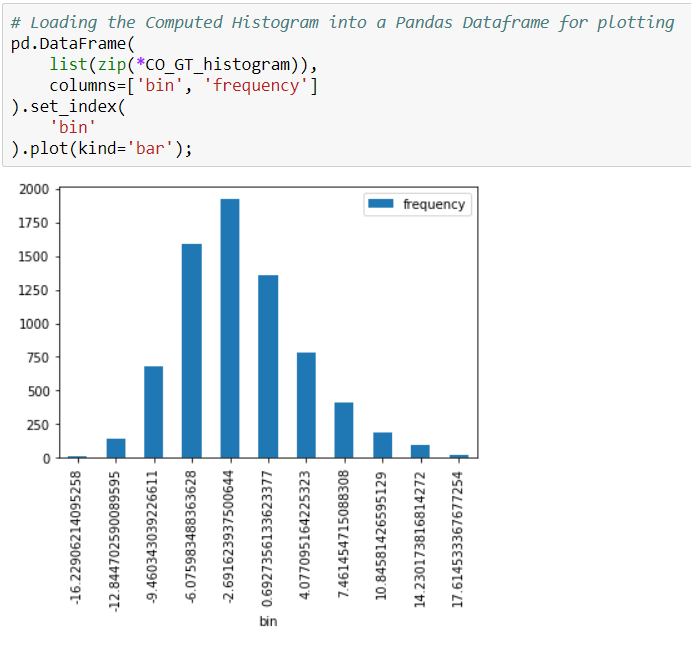
Linearity:





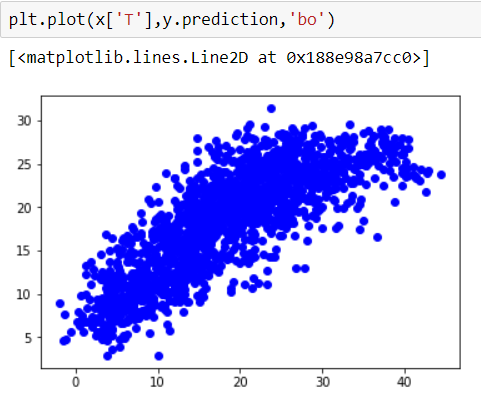
Looking at the scatterplot matrix, there is very little evidence that the dependent variables have a linear relationship with temperature. This would imply a very small correlation coefficient. Linear regression can be performed on this data since the data is not spread as a random cloud, but one should not expect to see extraordinary results from the model.

Normality of Residuals:



From the graph above, there is strong evidence that the residuals are normally distributed.

Constant variance:

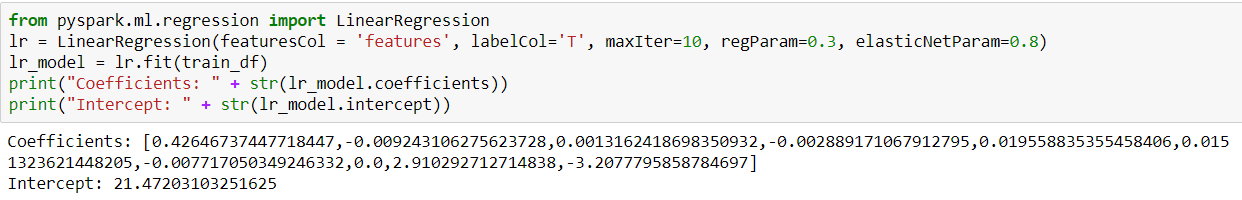


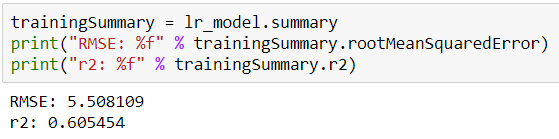
From the graph above, it is evident that the residuals are spread as a random cloud. There is no pattern to the residuals and the variance remains constant.

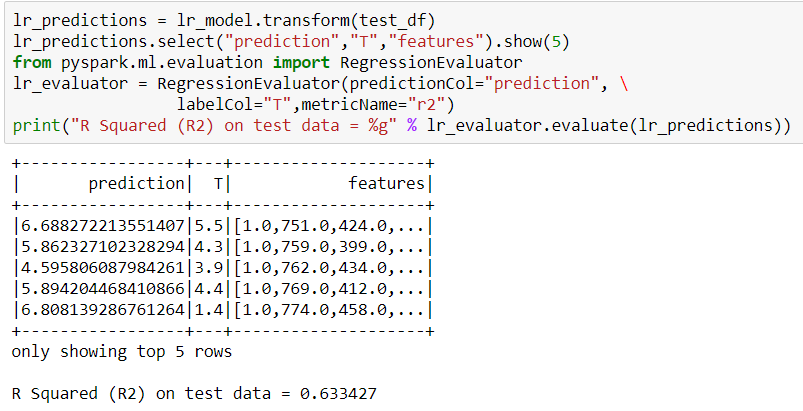
Independence:

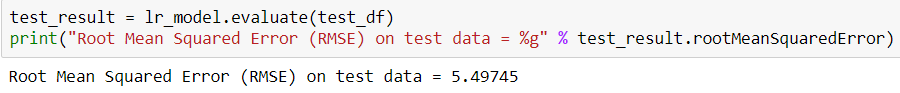
We will assume that the data is independent as this can be ensured only at the time of collection.

The assumptions are satisfied but not as strongly as we would like. A weak evidence on linearity would be a weaker model.





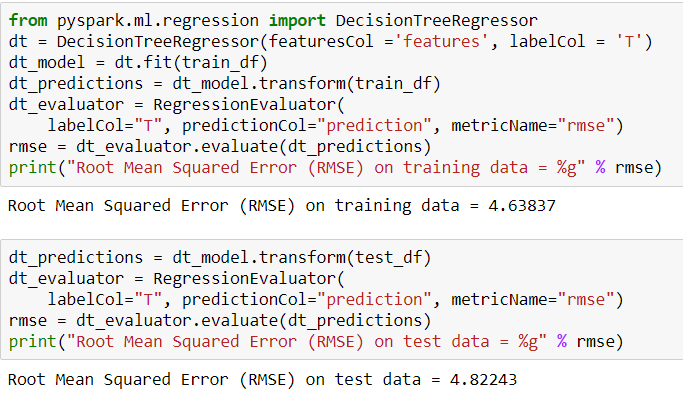




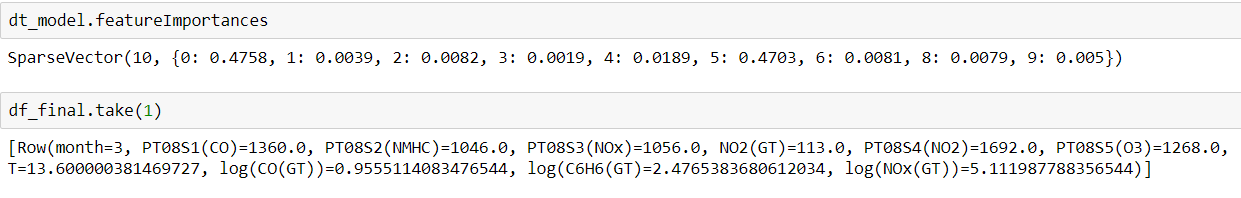
The error is quite high as was assumed previously. The training error is 5.5081 and the testing error is 5.4974. It is good to see the model performing better on the test data, but the error is still very high. We will move on to the next model and see if we can improve the performance.

2) CART

CART requires the target to be normally distributed, which we have already made sure in the data cleaning step. It also needs the observations to be independent, which will be assumed as this can be ensured only when data is collected. We will check the model performance and compare with the previous model



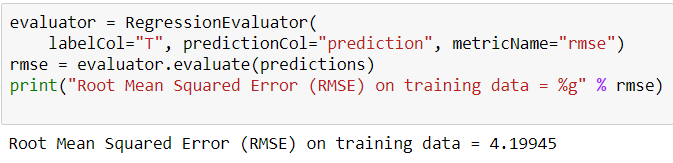
These results are an improvement over the multiple linear regression model. The testing error has come down to 4.8224. This model would perform better than the previous one. It is safe to say that at this stage we would go ahead with CART.

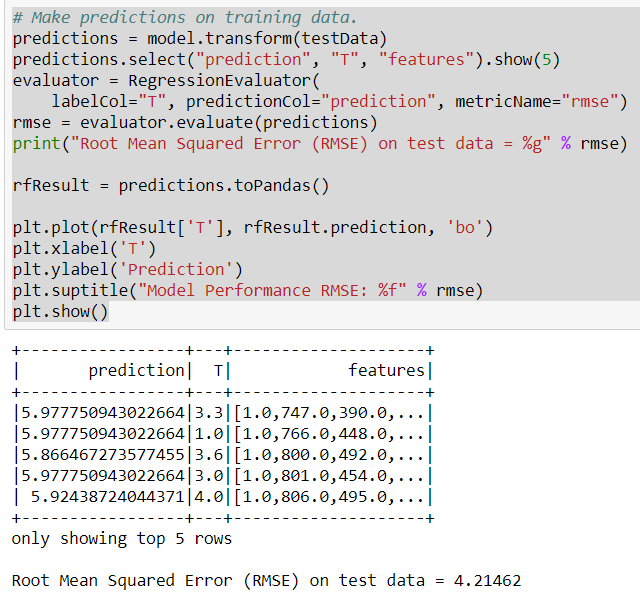


Using cart, month and PT08.S4(NO2) are the most important variables affecting the temperature

3) Random Forests:

The assumptions for Random Forests are the same as CART as this is a tree-based algorithm as well. We have already addressed the assumptions of normality and independence and we can proceed to compare the results of the model.

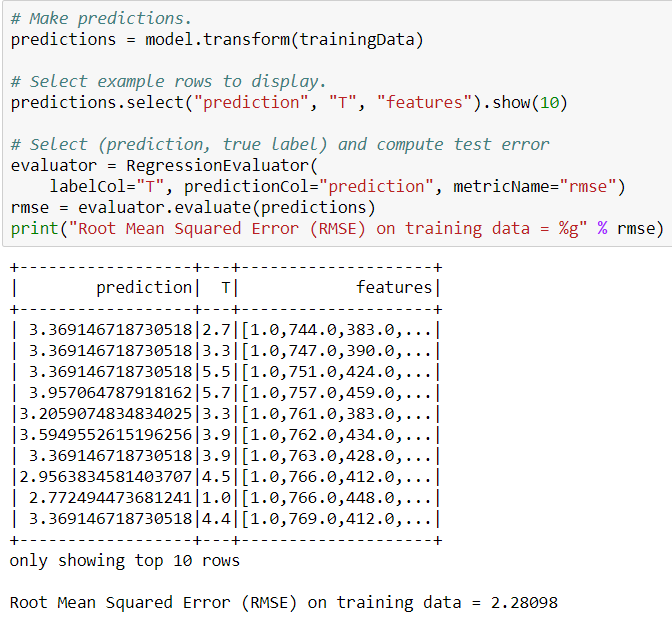


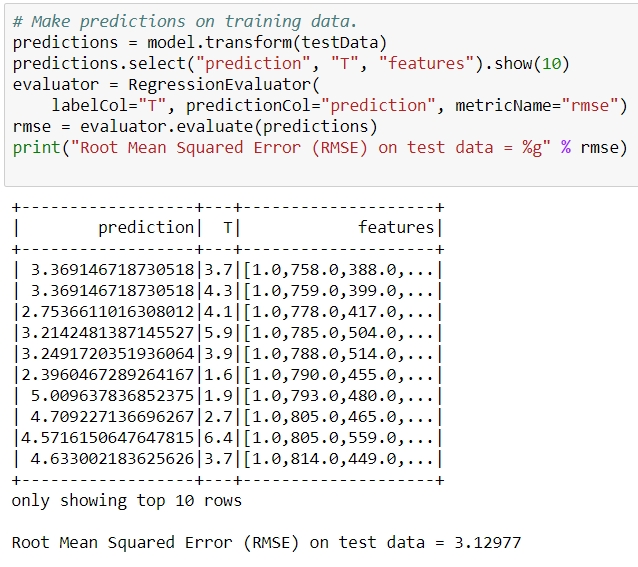


The results from Random Forest are even better than before. This makes sense as Random forest is ensemble method where many trees are averaged into one. This remarkably improves the performance of the tree. The model works extremely well on the training data with an error of 4.1994, and also works fairly well on the test data with an error of 4.2146. This result is much lower than the previous two methods.

4) Gradient Boosting:

Gradient boosting has the same assumptions as that of random forest that the data is independent, and the target is normally distributed. These assumptions have already been addressed before and we can compare the results





The results of the gradient boosting regressor is better than all other models. The training data has and error of 2.2809 while the test data has an error of 3.1297. This error is smaller than the other models already and by hyper tuning the parameters, it can be brough even lower.

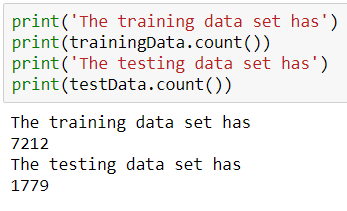
After reviewing the performance of all 4 models, we have come to the conclusion that Gradient Boosting would serve us the best. It has a very good performance on the test data, and it is comparatively easier to explain the behavior of every variable to the temperature. One could argue that the error is still too high for this to be a good model but, that is expected as we are not taking into account many other important factors that would affect the temperature. The aim of the project is not to build a model to predict the temperature based on certain factors but, to find the dependence of the temperature on the presence of various pollutants. In this case even if the error is a little higher, by establishing association we can reach the answer to our business objective. Hence, we will go forward with this model

We will tune the parameters of the model. We will build and perform 100. No hyperparameter tuning is performed at this stage.

**7) Data Mining**

**7.1) Create and Justify test designs:**

Testing data- Our data set consists of 8991 observations. We will be using a partition node to split the data into training and testing data sets. We will be using a 80-20 split. The aim is to leave a large chunk for training, which is the primary aim at this stage, and use a smaller chunk of data for testing the model. We would like the model to train on as much data as possible in order to generalize it and still leave out enough data to assess the performance of the model. An 80-20 split will leave 7205 observations for training and 1786 observations for testing. There are enough observations to train the data for and still leaving out enough data to validate the model.

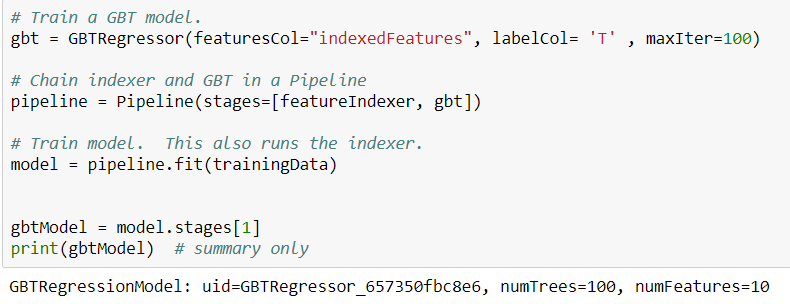


Setting success criteria- We will be assessing the model based on the root mean squared error on the test data. The parameters for the training set are not a good indicator of the model performance as they are based on the data that the model has already seen before. In a real world scenario, the model will always be working on data that it has never seen before. It is impossible to know how well the model will work on data it has never seen before. This is known as the risk of the model. It is never possible to know the exact value of the risk. We can only estimate the risk of the model. The mean absolute error is a good estimate for the risk. The lower the error, the better the model.

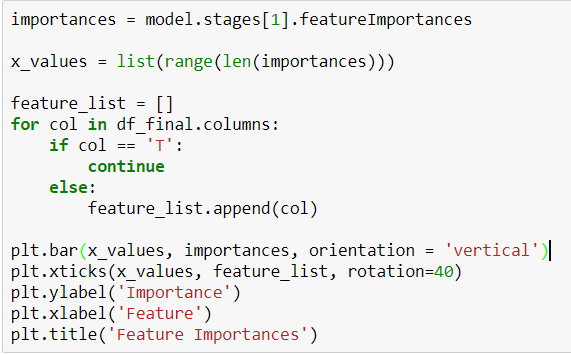
Multiple Iterations- Tuning the hyper parameters will affect the performance of the model. This will in turn affect the estimate of risk. Changing factors like the number of trees and number of iterations will improve performance up to a certain point after which the model will start overfitting and affecting the model performance. We will perform at least 3 iterations to figure out which parameters will result in the best model.

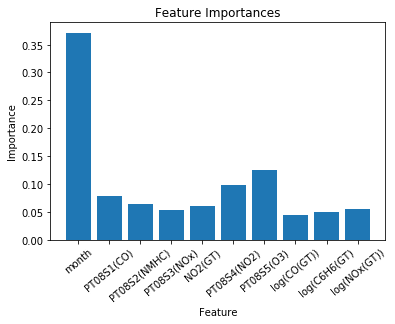
**7.2) Conduct Data Mining:**

The model executes successfully. We have built a Gradient boosting regression tree. Let us take a look at few of the model parameters.

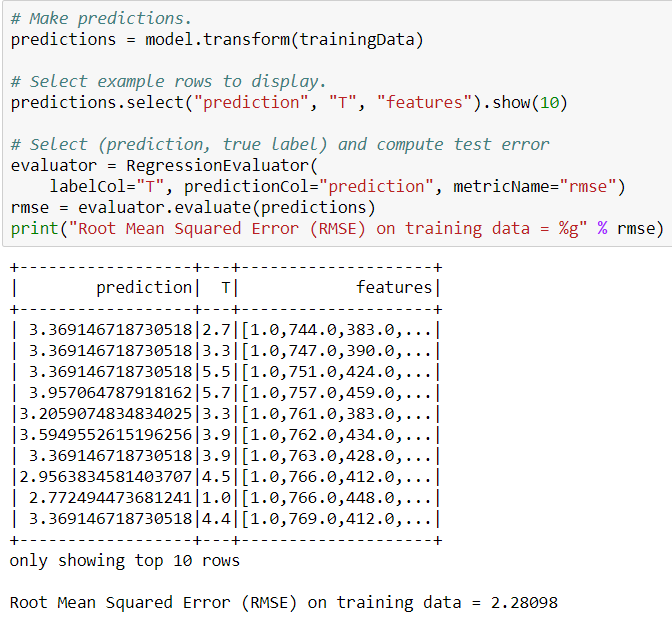


The output shows us the default hyperparameters of the model. We have only set the number of trees to 100 in this iteration

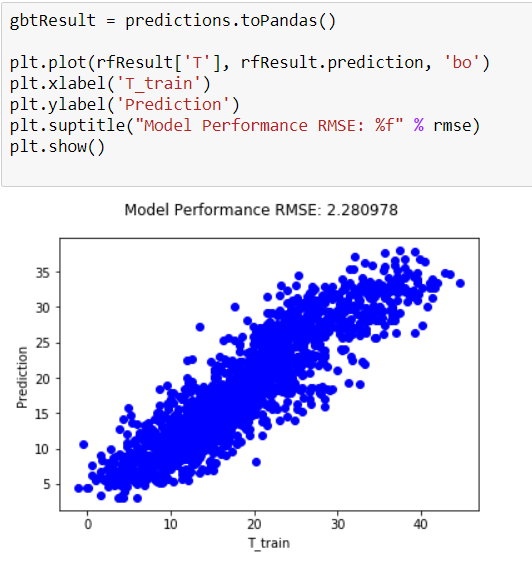


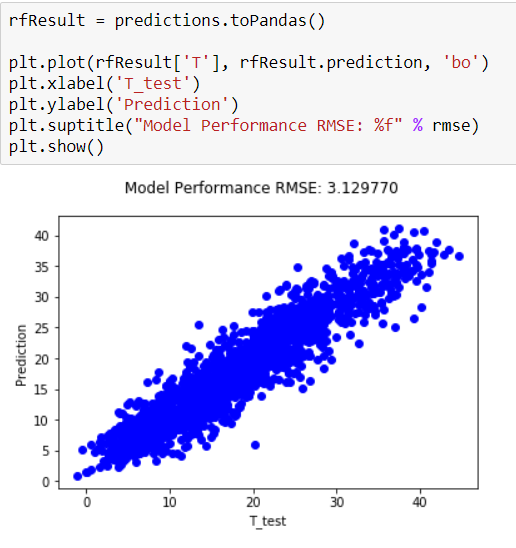


We will discuss the output from the model in further steps and look at many other statistics for the model.

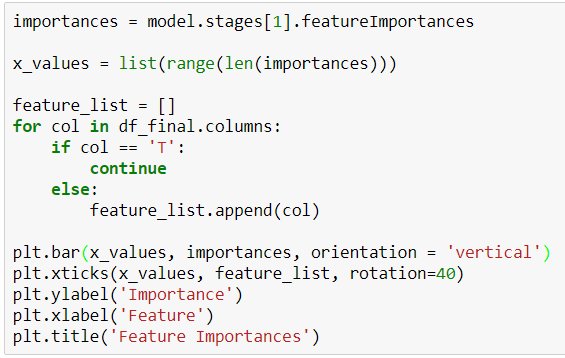


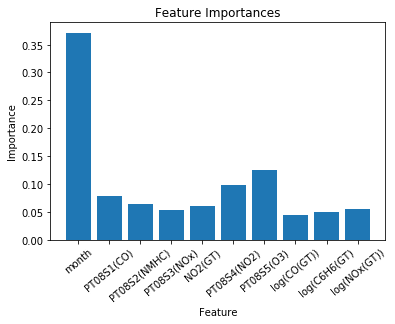
We run the model in the training data to get predictions. This is the predicted value of the temperature for the training data. The predicted values seem close to the original values. It is difficult to understand how well the model is built just by looking at the output and making visualizations will be more helpful in identifying patterns





The scatterplot of the residuals shows strong evidence that the predicted values are very similar to the observed values. This is good indicator that the model has been trained well on the training data set. Although this might also mean that the model has started overfitting since it is explaining the training data too much but, this can be verified only by estimating the risk.





The Predictor importance table and graph show that the month variable is significantly more important than the rest. Although every variable does contribute in predicting the temperature, the effect of every variable is very different.



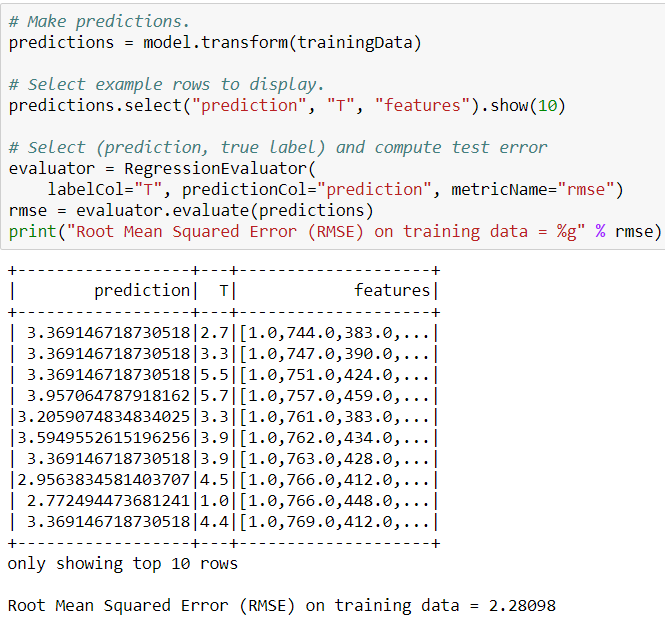


The analysis node gives us various evaluation metrics for the model on the training and testing data. Our metric of interest is the root mean squared error which is 2.2809 for the training data and 3.1297 for the testing data.

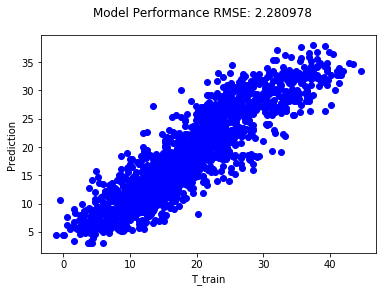
**8) Interpretation**

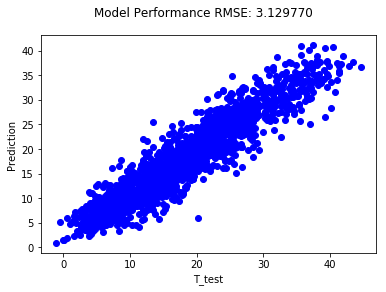
**8.1) Study and discuss the mined patterns**

In the previous step we have looked at the output from the model and searched for patterns in the data. Now we shall discuss more about patterns and explain the results.



The predicted column generated by the us for the training data is very close to the original data. This implies that the model works very well in predicting the training data. This is an expected trend as the model has seen the data before. The model uses the existing data to find the association between the dependent and independent variables and using this association, it makes the prediction for the temperature.





The scatterplot is a better visualization of model prediction. For an ideal model the data would lie along a straight line and the predicted and observed value would be exactly the same for every observation. But this would imply that the model is explaining the training data too much. This indicates overfitting. Although the model would work extremely well for the training data, the performance on the testing data will be extremely poor. Our model seems to explain the training data quite well without strong evidence for overfitting. It may not work with the exact accuracy on the test data, but the performance will be good enough. In our case the test data also shows evidence that the data is close along the line





The analysis node provides us a better metric to compare the performance of the model on the training and test set. As we have been seeing, the model has trained really well on the training data and the root mean squared error of 2.2993 is the evidence to support the same. The performance is not replicated on the test data. The performance is a little worse but is still good enough for our purpose. The model generates and error of 3.1433 on the test data.

**8.2) Visualize the data, results, models, and patterns**

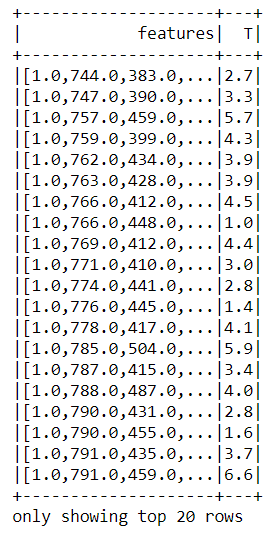
We have already visualized the model, data, results and the patterns in the previous steps so we will take a look at the visualizations again and summarize our findings

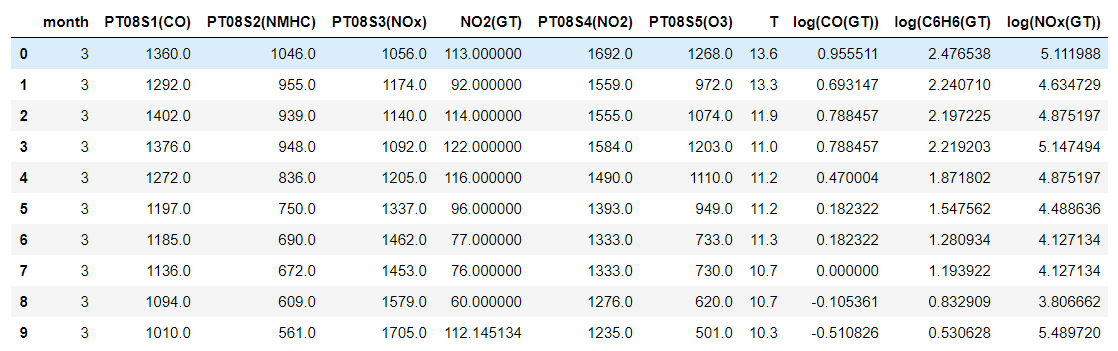
Model:



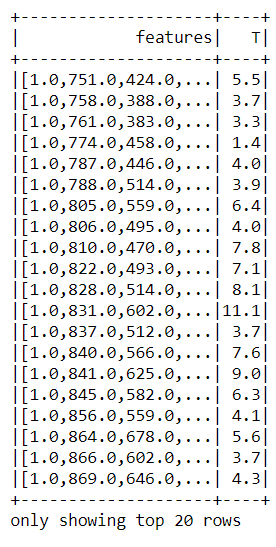
Data:

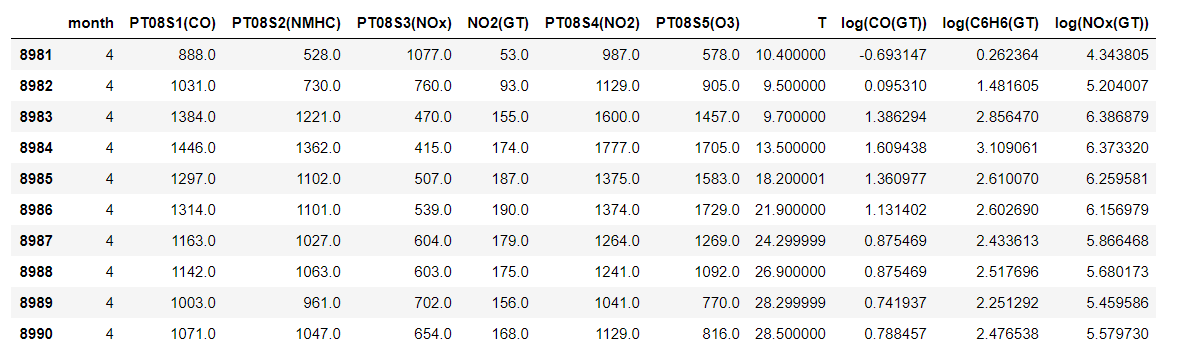
Training data:





Test data:

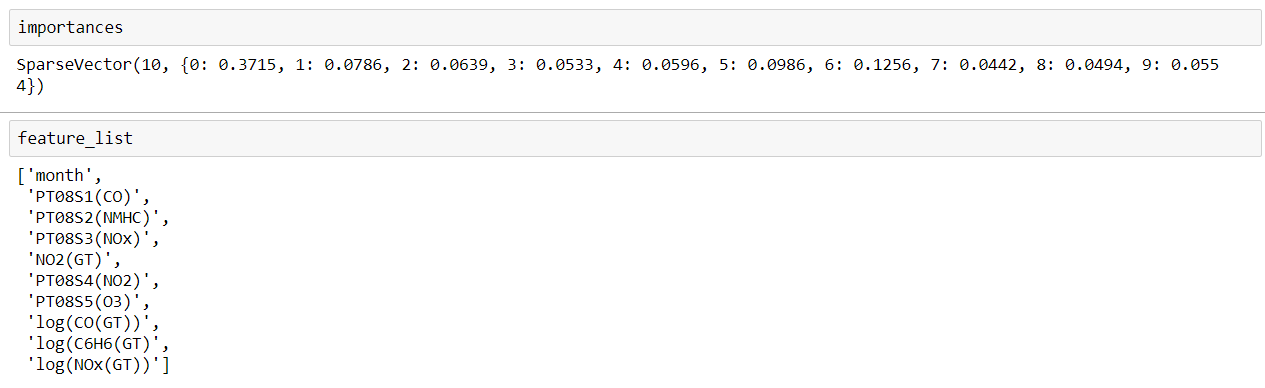




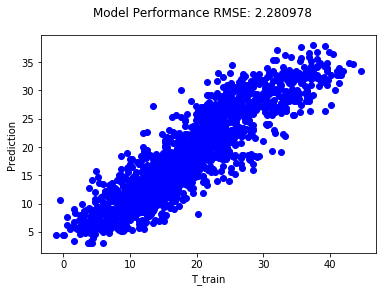
Results:

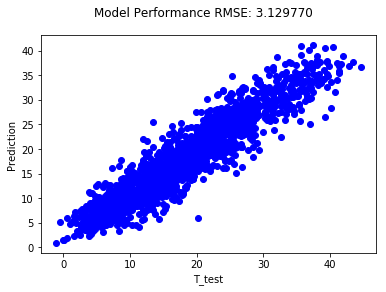


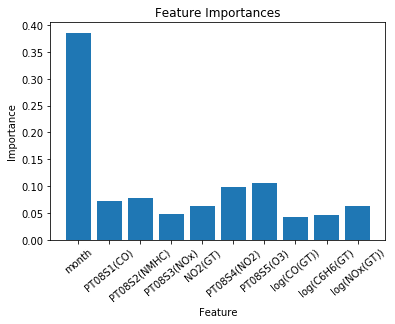




Patterns:



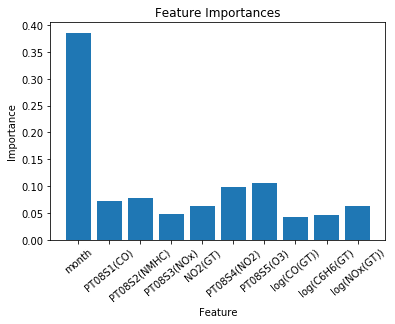


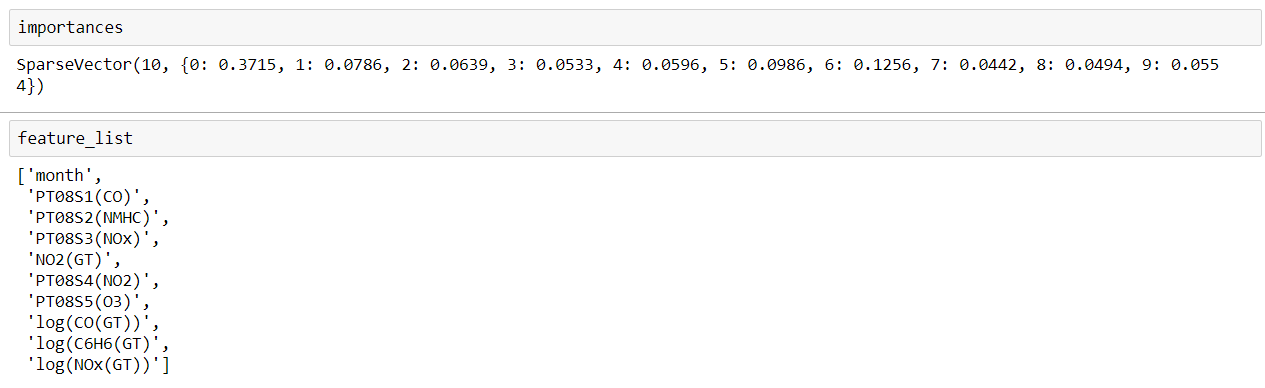


From the visualizations, we can see that the model worked very well for the training data and good enough on the test data too. There is enough evidence to infer that the model is adequate for our purpose. The predictor importance table will be extremely useful when deriving associations. We will interpret the results in the next step.

**8.3) Interpret the results, models, and patterns:**

The most important thing we would look here is for the predictor importance table. Since the aim of the project is to identify the pollutant, which is the most important in predicting temperature, this table will provide us with an answer to the question.





The most important variable is month. This is expected as the time of the year will have a great influence on the temperature. Month explains about 37.15% of the variation in the data. This variable is not controllable by us so we will look at the rest of the variables.

The 2nd important variable is PT08S5(O3). This variable explains about 12.56% of the variation in the data. This along with month explain 50% of the variation in the data. This means that controlling this variable will have the most effect on the temperature.

The 3rd important variable is the PT08S4(NO2). This variable explains only about 9.86% of the variation in the data

The 4th important variable is PT08S1(CO). This variable explains about 7.86% of the variation in the data

The 5th important variable is PT08S2(NMHC). This variable explains about 6.39% of the variation in the data

The 6th important variable is NO2(GT). This variable explains about 5.96% of the variation in the data

The 7th important variable is Log(NOx(GT)). This variable explains about 5.54% of the variation in the data

The 8th important variable is PT08.S3(NOx). This variable explains about 5.33% of the variation in the data

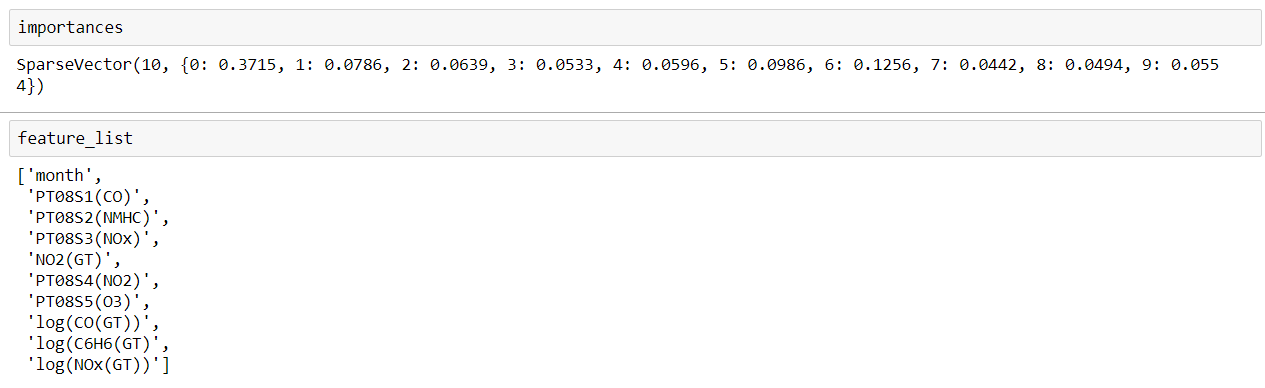
The 9th important variable is log(‘C6H6(GT)’). This variable explains about 4.94% of the variation in the data

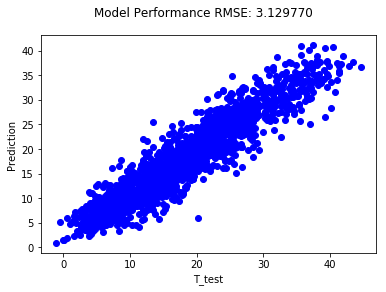
The 10th important variable is log(‘CO(GT)’). This variable explains about 4.42% of the variation in the data

This implies that PT08S5(O3) and the Month together explain about 50% of the variation. The other variables explain much less variation than the other two. The Month variable is not of much use to us as we cannot control it but finding the effect of PT08S5(O3) is very interesting. This is a variable that can be looked further into and can be controlled to change the temperature.

**8.4) Assess and evaluate the results, models and patterns**

The model is assessed based on the performance on the test data







From the about outputs it can be clearly seen that the model works quite well on the test data. The error on the testing data is quite small. Based on our findings this model is successful for our purpose.

The results of the model are assessed based on if we can answer the business objectives. In the previous step we had interpreted the model and that would help us to answer the business questions.

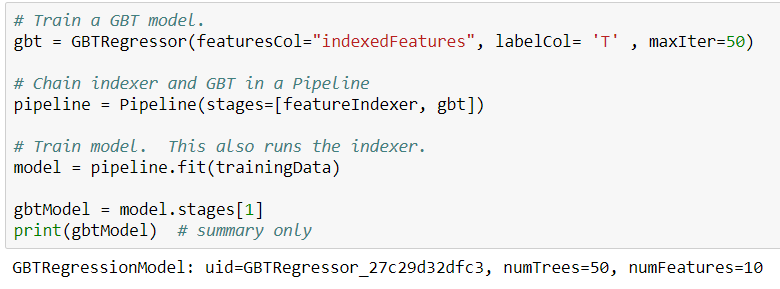
* We were interested to know which pollutants are associated with the temperature change and we have found that all the variables that we analyzed played some role in affecting the temperature.
* We wanted to find out if all the pollutants affect the temperature by the same amount or does the association vary and looking the predictor importance, we can say the some pollutants are more significantly associated with the temperature than the rest. PT08.S5(O3) is the most important variable affecting the temperature while the CO(GT) affects the temperature the least.

Since the results help us answer our business questions, we can consider this model successful.

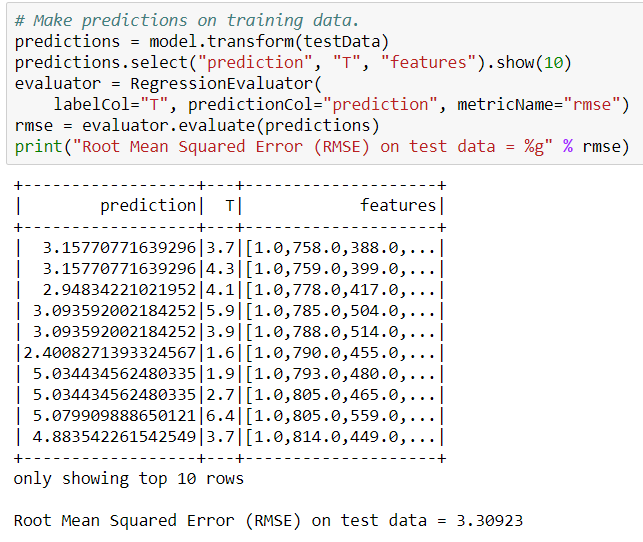
**8.5) Iterate prior steps**

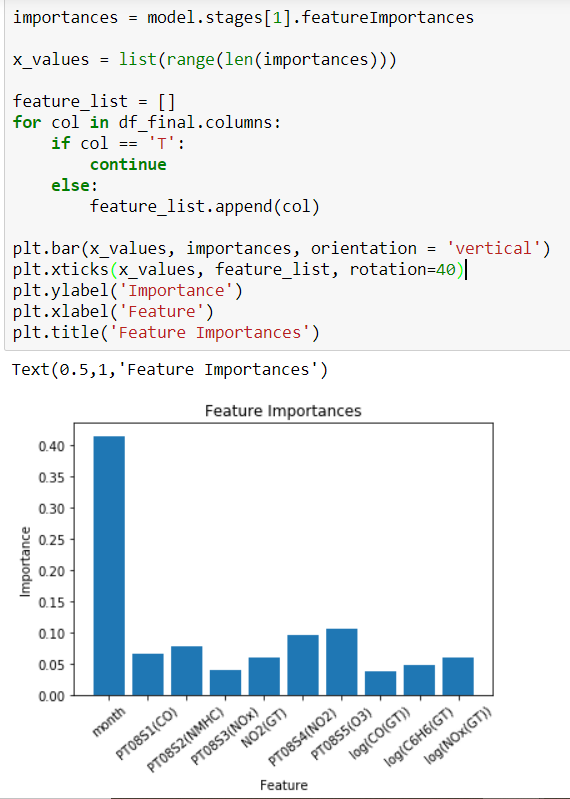
Building just one single model is not an indicator of the consistency of the results. It is a good idea to build multiple models with an algorithm by hyperparameter tuning in order to check if the model is robust and if the results are consistent. In doing so we can come across a better solution to the problem.

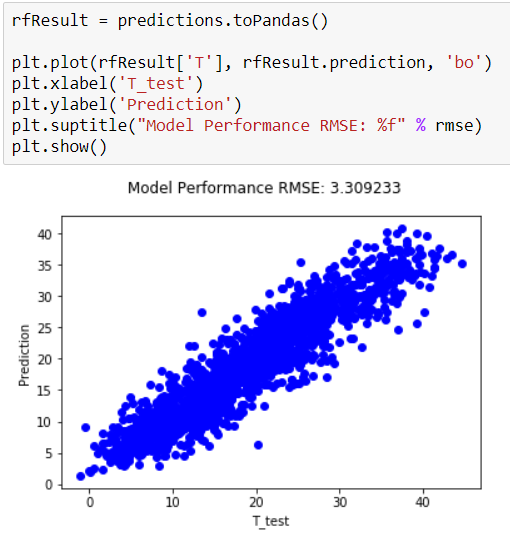
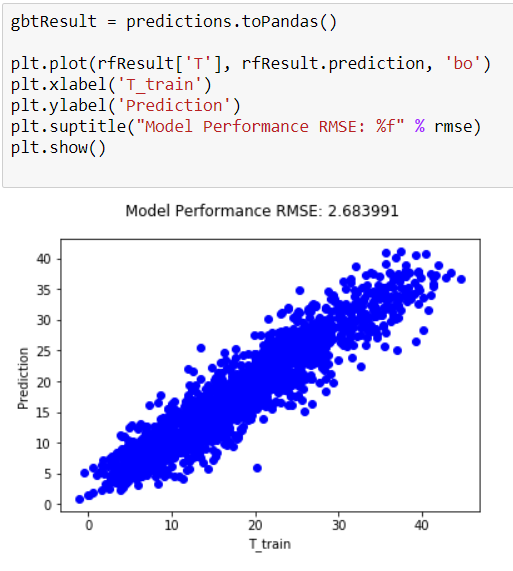
In the previous steps we have built 100 regression trees and have not performed hyperparameter tuning for the random forest. We will try with 50 trees.



The results of the model are:







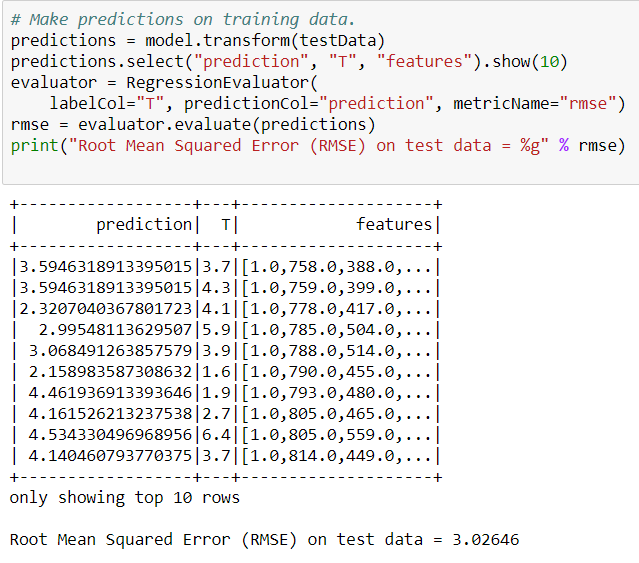
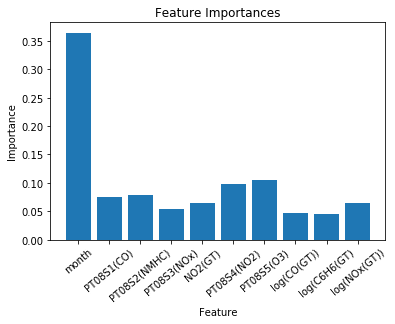


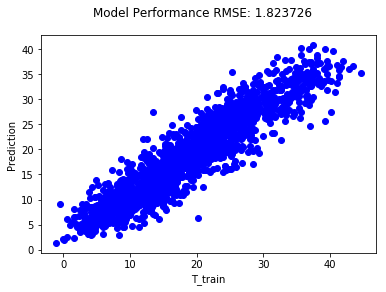


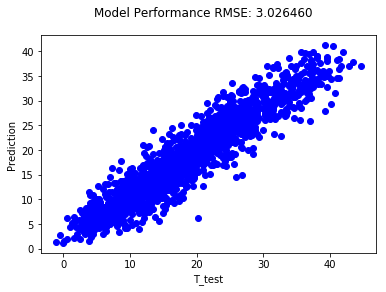
In this case the results are consistent with our findings. The results are still very similar. The error is slightly higher but the predictor importance is still the same. Next we will try with 200 trees.



The resutls of this model are:









Again the result are similar so we can say that our model is robust and shows consitent results. The error has reduced further indicating that increasing the number of tress might result into a better model. Hence we have built as susccessful data mining model.

**References:**

[1] James Hansen, Makiko Sato, Reto Ruedy, Ken Lo, David W. Lea, & Martin Medina–Elizade,

Global temperature change PNAS September 26, 2006 103 (39) 14288–14293;

<https://doi.org/10.1073/pnas.0606291103>

[2] S. De Vito, E. Massera, M. Piga, L. Martinotto & G. Di Francia, On field calibration of an electronic

nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators

B: Chemical, Volume 129, Issue 2, 22 February 2008, Pages 750–757, ISSN 0925–4005

<https://www.sciencedirect.com/science/article/abs/pii/S0925400507007691?via%3Dihub>

[3] Ekwurzel, B., Boneham, J., Dalton, M.W. *et al.* The rise in global atmospheric CO2, surface

temperature, and sea level from emissions traced to major carbon producers. *Climatic*

*Change* **144,**579–590 (2017). <https://doi.org/10.1007/s10584-017-1978-0>

[4] Air Quality Dataset

[https://archive.ics.uci.edu/ml/datasets/Air+Quality#](https://archive.ics.uci.edu/ml/datasets/Air+Quality)

[5] Breiman, L. Random Forests. *Machine Learning* **45,**5–32 (2001).

<https://doi.org/10.1023/A:1010933404324>

[6] Fitzmaurice, Garrett M. "Regression." *Diagnostic Histopathology* 22.7 (2016): 271-78. Web.

https://www-sciencedirect- com.ezproxy.auckland.ac.nz/science/article/pii/S1756231716300627

[7] Dejaegher, B., Dhooghe, Goodarzi, Apers, Pieters, and Vander Heyden. "Classification

Models for Neocryptolepine Derivatives as Inhibitors of the β-haematin Formation."

Analytica Chimica Acta 705.1-2 (2011): 98-110. Web.

[https://www-sciencedirect-](https://www-sciencedirect-  com.ezproxy.auckland.ac.nz/science/article/pii/S0003267011005411)

[com.ezproxy.auckland.ac.nz/science/article/pii/S0003267011005411](https://www-sciencedirect-  com.ezproxy.auckland.ac.nz/science/article/pii/S0003267011005411)

[8] Altman, N., & Krzywinski, M. (2017). Clustering. *Nature Methods,* *14*(6), 545-546.

https://catalogue.library.auckland.ac.nz/permalink/f/ss4o5v/TN\_wos000402291800004

[9] Torres-Barrán, A., Alonso, &., & Dorronsoro, J. (2019). Regression tree ensembles for

wind energy and solar radiation prediction. Neurocomputing, 326-327, 151-160.

[https://www-sciencedirect](https://www-sciencedirect  -com.ezproxy.auckland.ac.nz/science/article/pii/S0925231217315229)

[-com.ezproxy.auckland.ac.nz/science/article/pii/S0925231217315229](https://www-sciencedirect  -com.ezproxy.auckland.ac.nz/science/article/pii/S0925231217315229)

"I acknowledge that the submitted work is my own original work in accordance with the University of Auckland guidelines and policies on academic integrity and copyright.

I also acknowledge that I have appropriate permission to use the data that I have utilised in this project. (For example, if the data belongs to an organisation and the data has not been published in the public domain then the data must be approved by the rights holder.) This includes permission to upload the data file to Canvas. The University of Auckland bears no responsibility for the student's misuse of data."