Iteration 3 – MSAS

1. **Situation Understanding**
   1. **Situation objectives identification:**

With the rapid development of our world, the quality of human beings’ lives, and living standards are improving to a quite high level. However, the truth is that there are still existing various kinds of problems that influence the world we are living in. In order to solve those problems, the United Nations proposed ’17 Sustainable Development Goals.’ In this case, we focus on the air quality problem, which belongs to multiple goals, such as the 3rd ‘Good health and well-being’ because of the low-quality air is harmful to us; also, the 13th ‘Climate action’ as the impurity in the air may cause the climate change – in a bad way. The objective of this case is trying to discover, find and summarize the possible pattern of the air quality data sets, and even predict the future quality of air – in other words, by utilizing those data sets and mining the underlying pattern, we may find a better way to help the world and the public to understand the change of the air that we breathe every second, and take actions to protect the air.

* 1. **Situation Assessment:**

**Sources of Task:**

**Data source**: The datasets are available from Kaggle and the Taiwan Environmental Protection Administration Executive Yuan.

**Hardware source**: Personal laptop, Amazon Web Service.

**Software source**: GitHub Desktop, Putty.

The current situation our air quality is not that optimal. The air quality has always been a serious problem for us, especially for Some developing country. In Taiwan, the air quality indicators, no matter the PM2.5 or PM10 or other chemicals (nitric oxide), all of them indicate that the people are breathing impurity such as ozone, different respirable particulate matters, and so on. Those impurities are no doubt harmful to human beings.

For the purpose of analysis, the air quality monitoring data is indispensable. Usually, the data file data has various kinds of type - the numerical such as numbers and dates, the categorical like locations and weather-type. In general, all the data can be used during the analysis process. However, the risk of this is, the data we used is not always ‘clean,’ and this unclean situation may lead to the inaccurate model generation which is used to make prediction – if the model we have is not accurate, consequently, the predicted results are also not precise, in some special occasion, this inaccuracy may cause serious consequence. The possible solution is, at the very beginning we will execute the operation ‘data cleaning,’ and we will try several different algorithms to make sure that we can generate the optimistic model in a certain way.

* 1. **Data Mining Objectives:**

In general, the goal is to find the potential pattern among the attributes and try to find a way to make a prediction and even get a way to control the air pollution.

To help our society to understand the air quality data, we need to analyze the oceans of past monitoring air data. The data mining objectives are mainly focused on finding the different types or kinds of relations between the numerous data attributes. By analyzing those data, such as the different numbers of different indicators in various time slots. We may find and summary a pattern that when the air pollution is most serious, and which kind of pollutant contributes the most. During the mining process, several different algorithms may be utilized to test for some diverse model, and finally, choose the one who can describe the past data and optimally predict the future air quality.

* 1. **Project Plan:**
     1. **Project Background**

As above, to achieve the United Nations’ sustainable development goals – ‘Good health and well-being,’ and ‘Climate action,’ the society need to pay more attention to the air that is crucial and indispensable for us. Based on this, we need to collect, process and then analyze the air quality monitoring data about the past. After analysis, expectable we will get a pattern of these data, and ideally, we can utilize this model/pattern to predict the future air quality and also, take action on air quality control to provide a better environment for human beings.

* + 1. **Tasks**

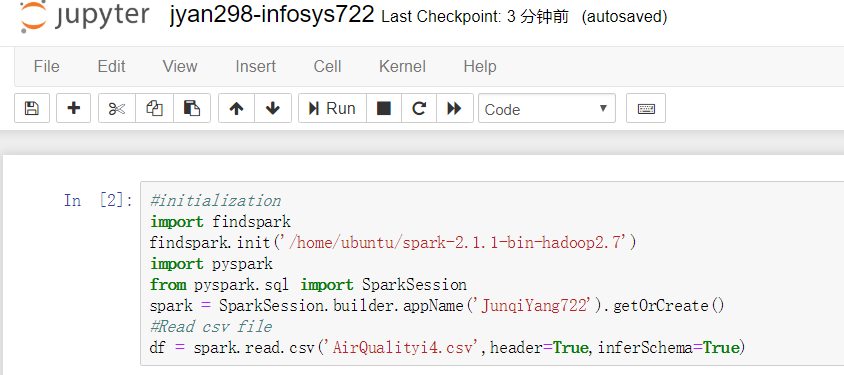
|  |  |  |  |
| --- | --- | --- | --- |
| **Phase** | **Time (ETA)** | **Resources** | **Risks** |
| Situation understanding | 6 hrs. | Data analysts | Environment change, criteria change |
| Data understanding | 12 hrs. | Data analysts | Data problems, technology problems |
| Data preparation | 1 day. | Data analysts, and specific software/tools | Data problems, technology problems |
| Modeling | 1 day. | Data analysts, specific software/tools | No model can fit the requirements |
| Evaluation | 6 hrs. | Data analysts | Environment change, criteria change |
| Deployment | 3 hrs. | Data analysts | Environment change, criteria change |

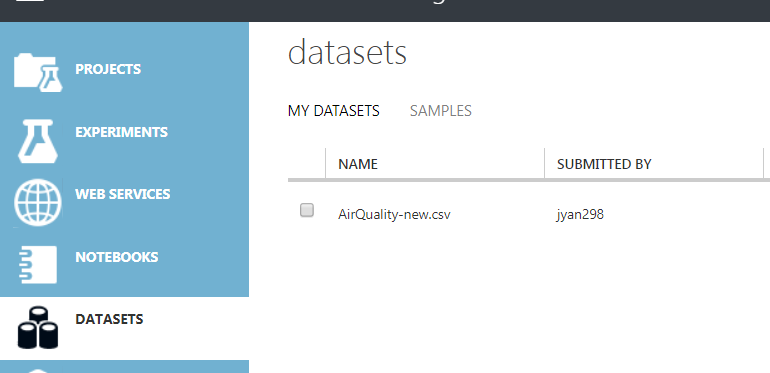
1. **Data Understanding** 
   1. **Initial Data Collection:**

In this phase, we will collect the data sets that we need to perform data mining operation. There are plenty of ways for us to gather data – for an organization, the data within the organization should be collected in terms of the current data; also, sometimes we may need to purchase additional data from organizations because our data set is incomplete. If still there are not enough data sets, we need to research and add more data to the data sets.

There are thousands of websites (governments, organizations and so on) provide the data sets to download. In this case, the data can be collected from the official websites or other specialized data sets websites such as Kaggle. For this task, the data sets we collected in this case, it was downloaded via Kaggle.

In this case, we can import the dataset by using pandas or PySpark.





* 1. **Data Description:**

Dataset Scale: 23 attributes, nearly 220k records.

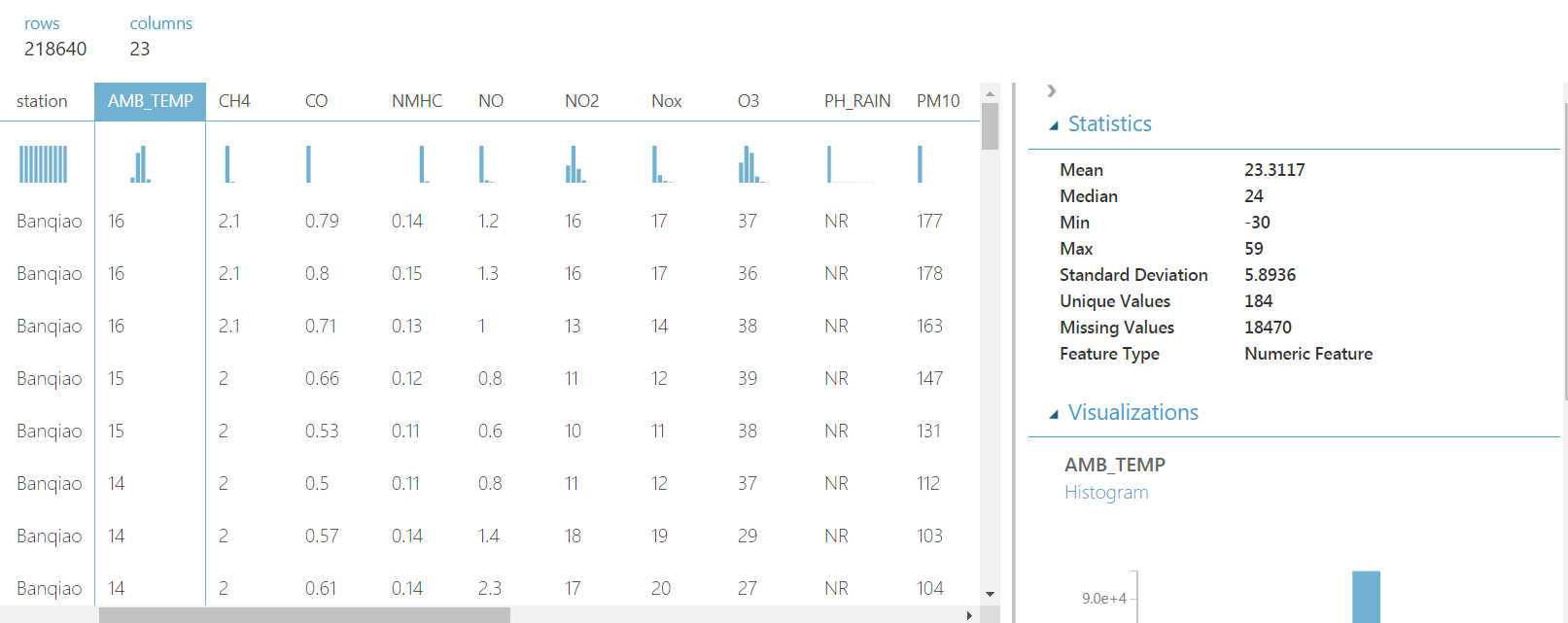
|  |  |  |  |
| --- | --- | --- | --- |
| Attributes | Description | Type | Role |
| Time | Time records | Continuous | None |
| Station | The station that recorded the data | Categorical | Input |
| AMB\_TEMP | Temperature | Continuous | Input |
| CH4 | CH4 index | Continuous | Input |
| CO | CO index | Continuous | Input |
| NMHC | Non-methane hydrocarbons index | Continuous | Input |
| NO | NO index | Continuous | Input |
| NO2 | NO2 index | Continuous | Input |
| NOx | NOx index | Continuous | Input |
| O3 | O3 index | Continuous | Input |
| PH\_Rain | The PH value | Categorical | Input |
| PM10 | PM10 index | Continuous | None |
| PM2.5 | PM2.5 index | Continuous | Target |
| Rainfall | Rainfall index | Continuous | Input |
| Rain\_Cond | Conductivity index | Continuous | Input |
| RH | Relative-humidity index | Continuous | Input |
| SO2 | SO2 index | Continuous | Input |
| THC | Total-Hydrocarbon index | Continuous | Input |
| UVB | UVB index | Continuous | Input |
| WD\_HR | Wind direction/hour | Continuous | Input |
| Wind\_Direction | Wind direction | Continuous | Input |
| Wind\_Speed | Wind speed | Continuous | Input |
| WS\_HR | Wind speed/hour | Continuous | Input |

Considering the methods of air monitoring, the air quality records will be a large number. The scale of our data set has 23 columns and approximately 220k records, and it stored in a CSV file, so I can upload this file to the ‘Dataset’ module on the Azure Machine Learning Studio, and drag it on the canvas of experiment. The majority of the value types are numerical and categorical, such as date, time, numbers, location, and so forth.

The Azure machine learning studio and Power BI can be used to discover the dataset.

* 1. **Data Exploration:**

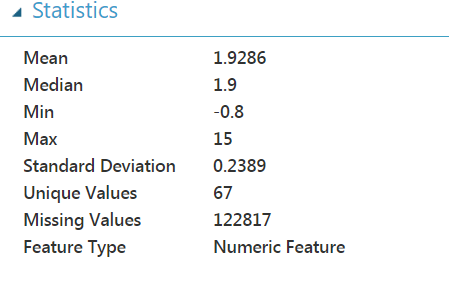
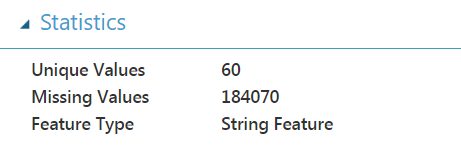
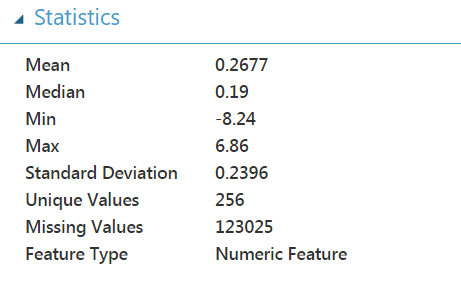
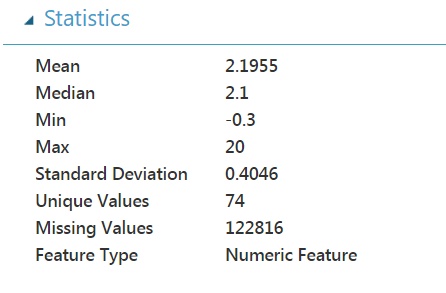
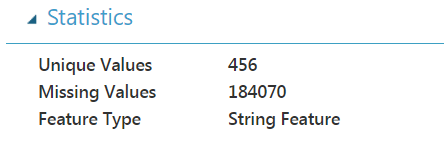
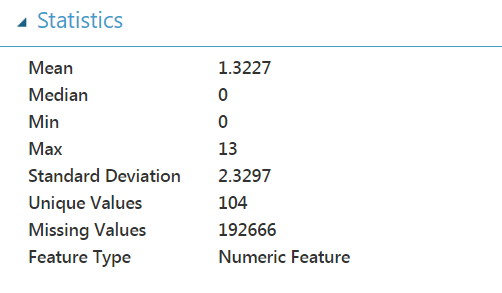
The Microsoft Azure Studio and Power BI can be used to initially discover the dataset. The Visualization function can help us to discover the dataset information and data quality. Below is the screenshot of result.



MSAS Initial View of Data

Here we can notice, overall we got 218,640 records and 23 columns. The data set is not very ‘clean’ – for the ‘AMP\_TEMP’ attribute, we can see there are some missing values and unique values, which means we have to do the ‘clean’ later.

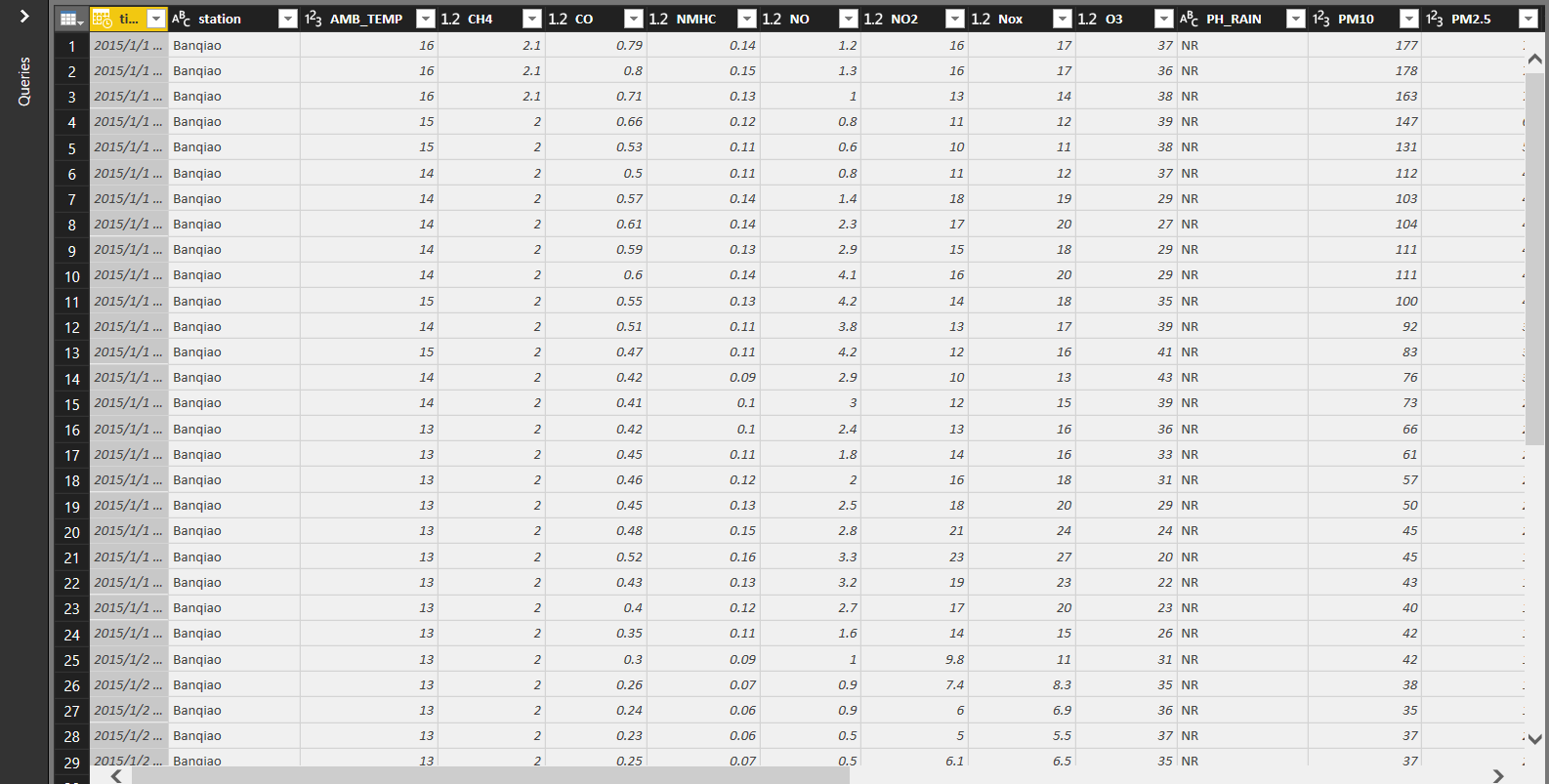
The ‘Statistics’ module provides a convenient way to initially discover the quality of our dataset, we can click each attributes to check the detail information.

Initially check of some attributes

After the initial check of the dataset, I found that some attributes have too many missing values that can hardly be ‘processed,’ so I consider to drop those attributes (CH4, NMHC, PH\_Rain, THC, Rain\_Cond, UVB).

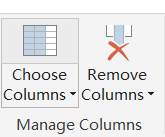
In order to explore more about our dataset, I decide to use the visualization function in the Power BI to discover. To begin with, I need to build a connection between the dataset and Power BI.



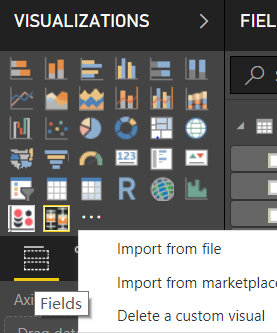
Import dataset into Power BI

In order to further explore the dataset, I removed some columns.

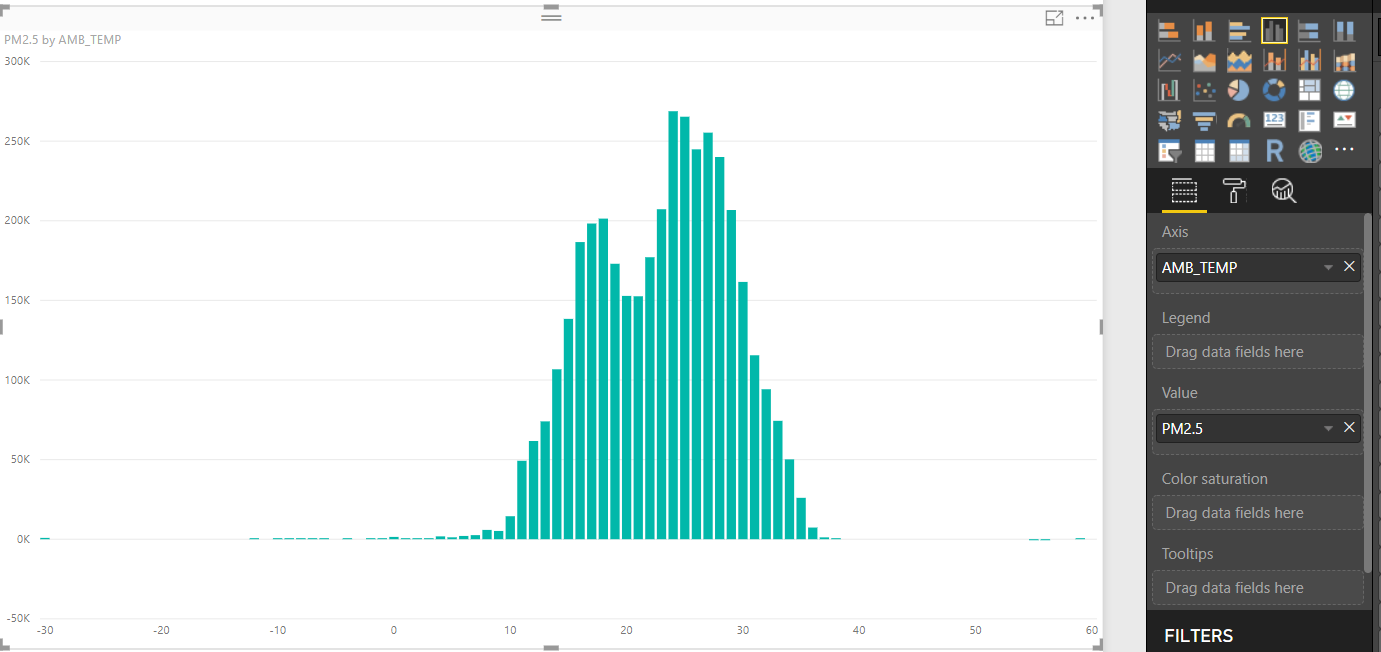




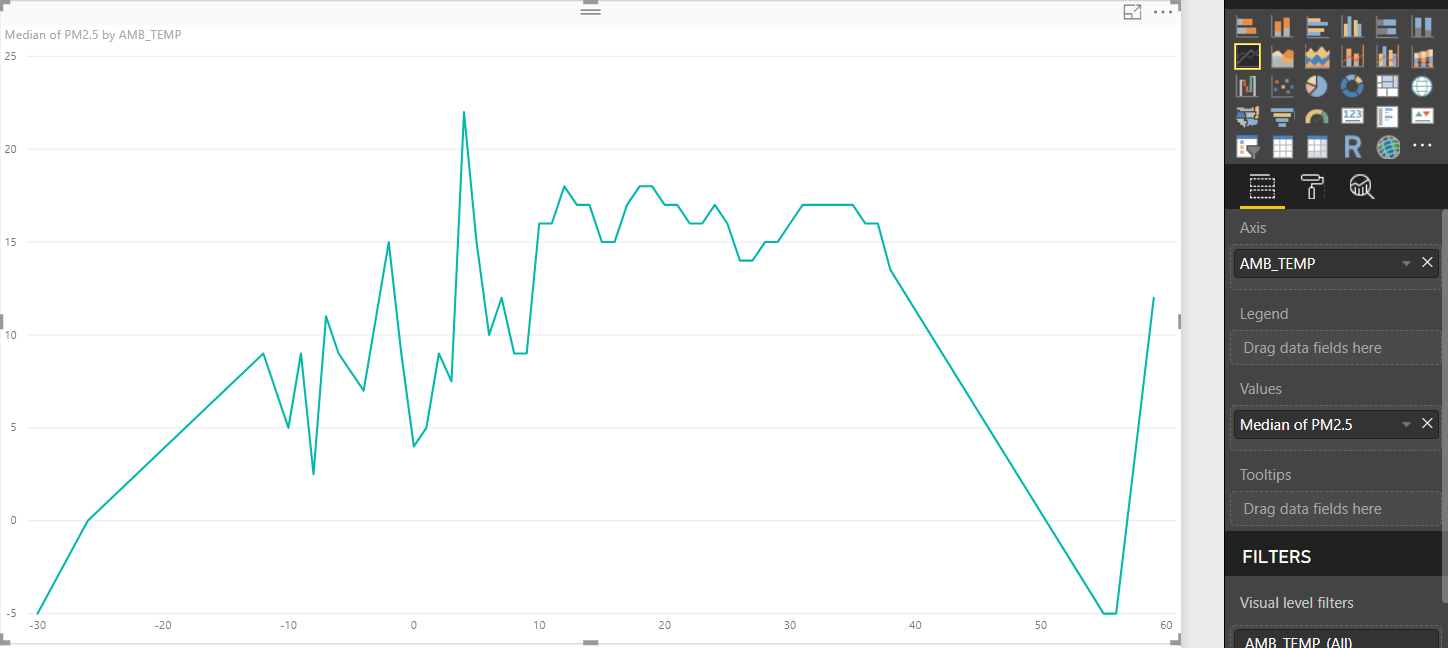
Also, Microsoft provides a lot of customized visualizations, we can import them from the official website (https://appsource.microsoft.com/en-us/marketplace/) to enhance the visualization function.



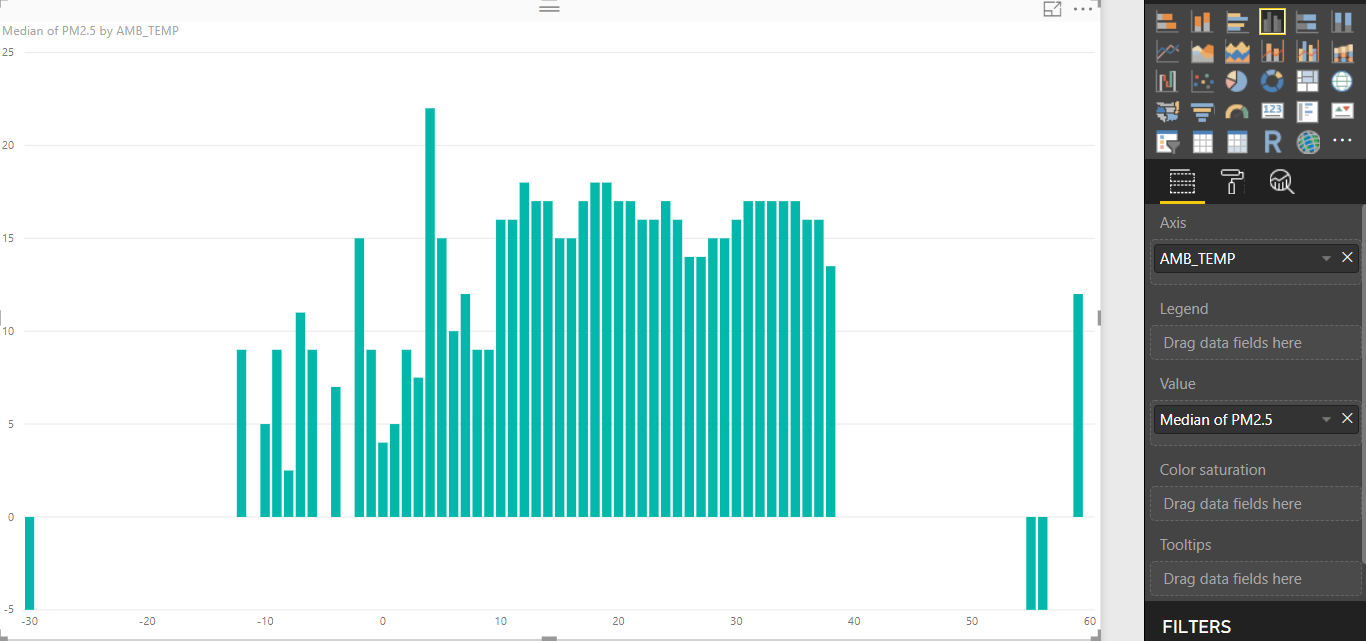
I assume that the temperature and the PM values may have some relationships, so I choose the AMP\_Temp and PM2.5 attributes and make some graphs in the Power BI.



PM2.5 (Sum) – Temperature Graph



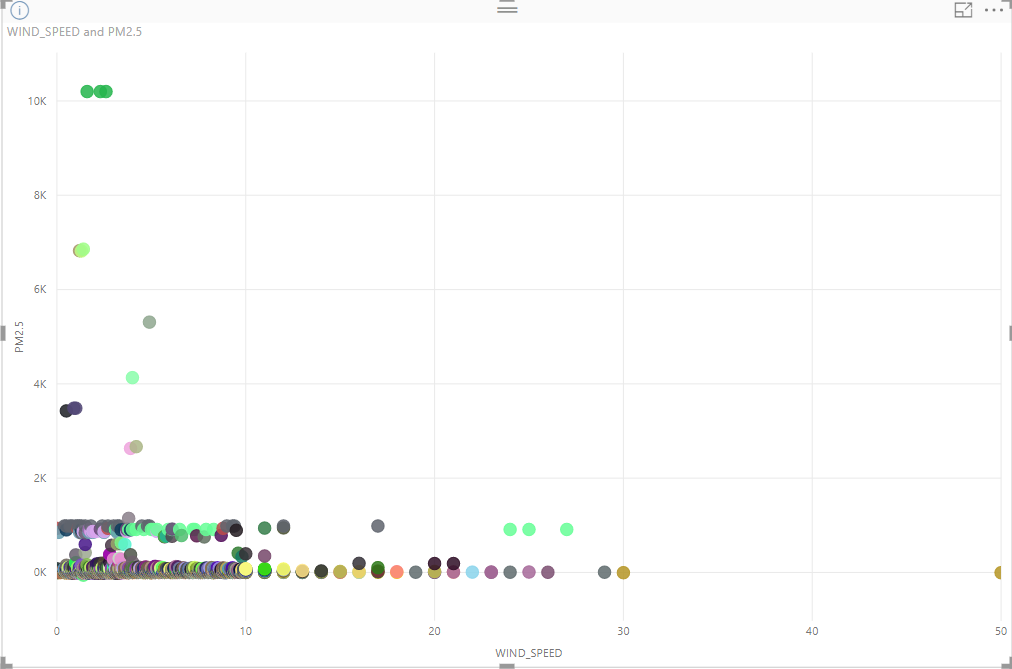
PM2.5 (Median) – Temperature Line Graph

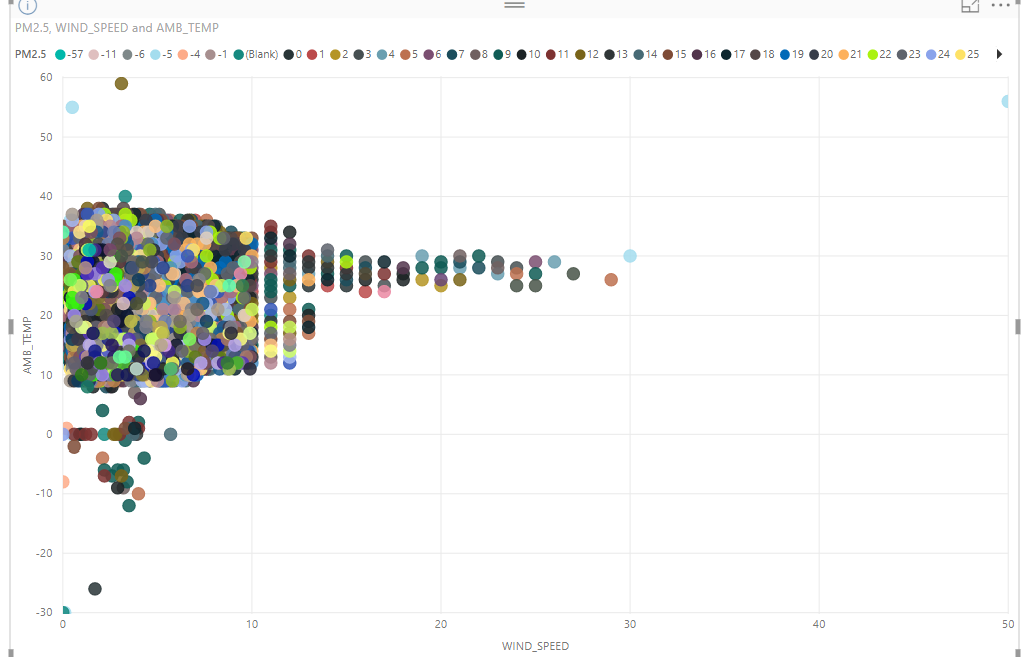


PM2.5 (Median) – Temperature Graph

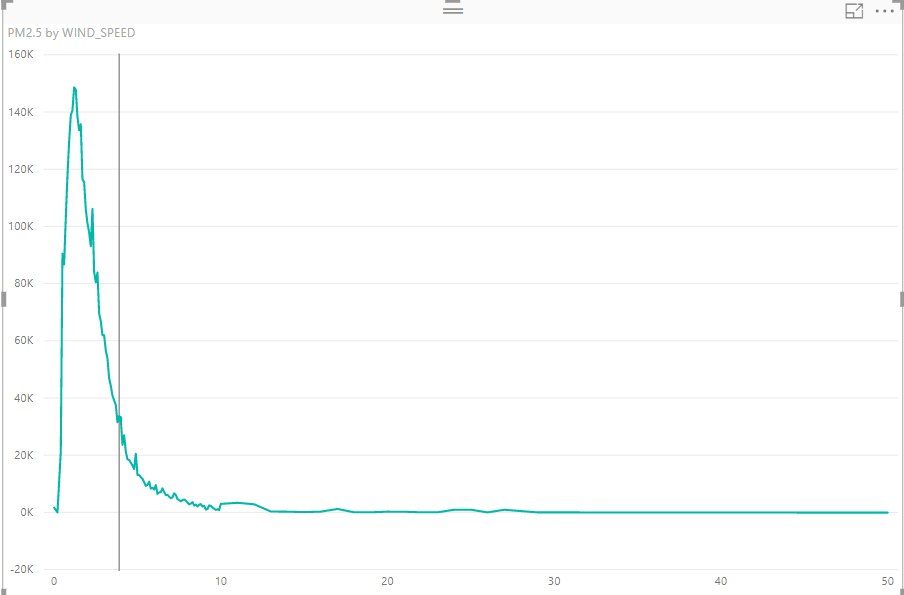
From this graph, we can see there are some values are not correct as they are negative values of PM2.5 values. The reason why chooses the Median as it could mitigate the influence of outliers to a certain degree. We cannot have a clearly view about the relationship between temperature and PM value.

Additionally, I consider that the wind speed might also influence the value of PM, so we made another graph.



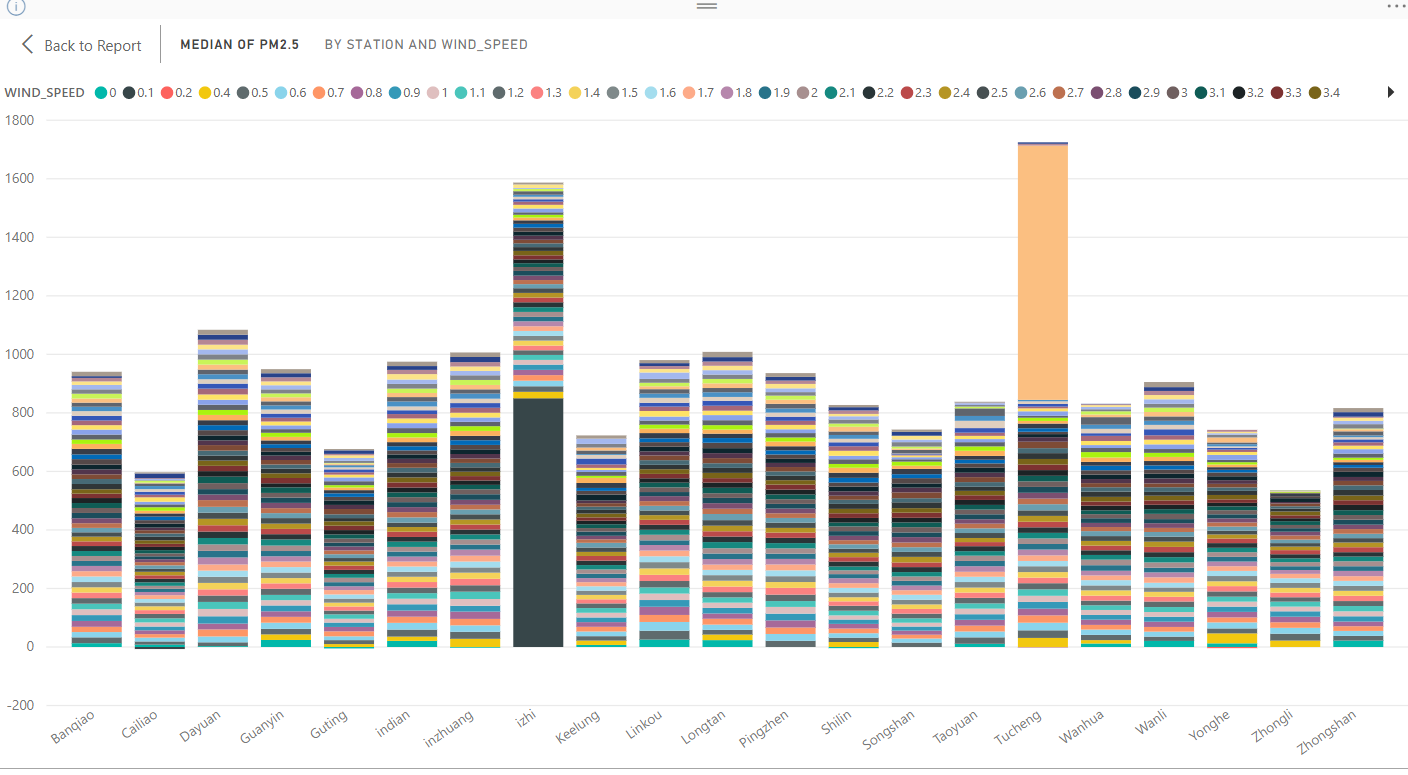


Wind-Speed – PM2.5



Based on the graph, roughly we can say that the higher speed of wind usually has a lower value of PM.

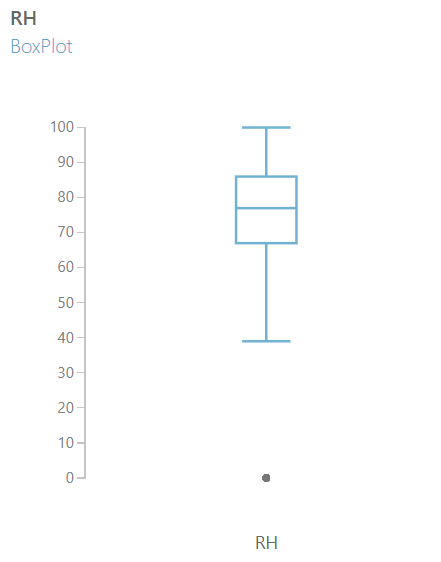
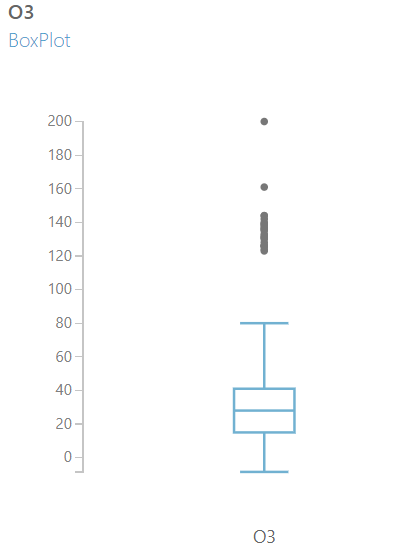
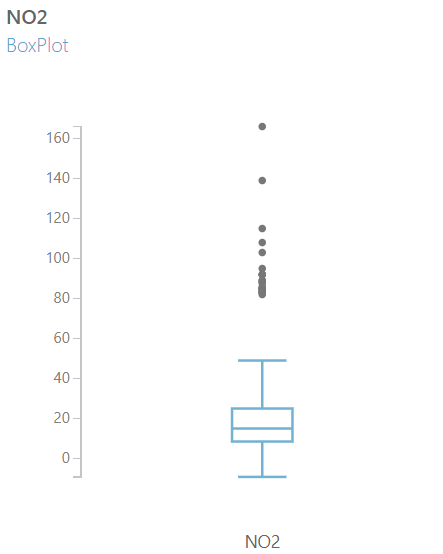
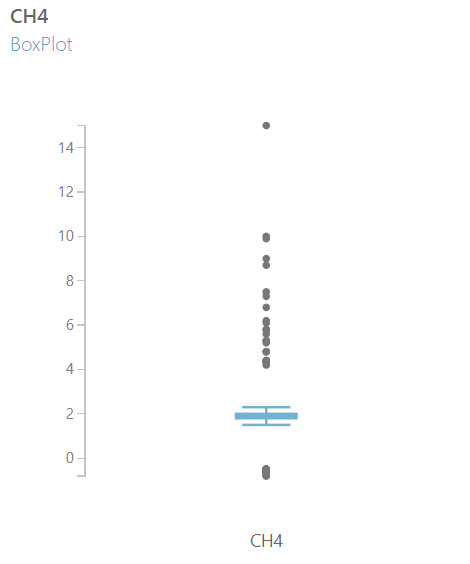
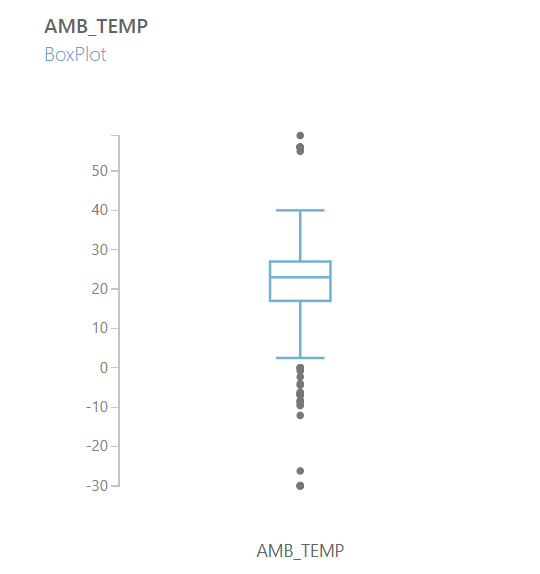
I also try to discover the overall situation of all stations.



This graph illustrates the pm2.5 values in different station with different wind speed, there are some values seems abnormal.

* 1. **Data Quality Verification:**

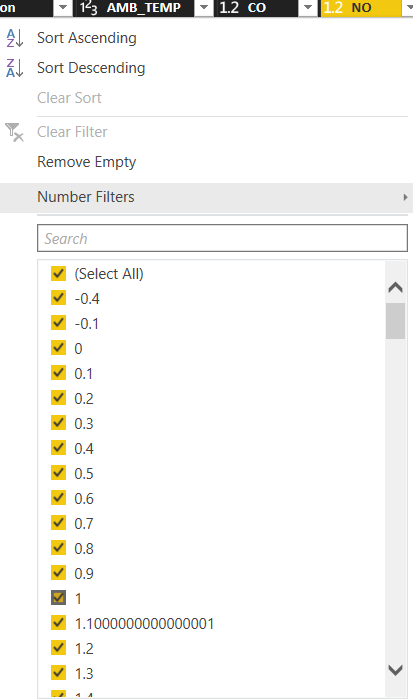
The dataset with a huge amount of records is always not perfect. There are a lot of unexpected data errors within the data set, in our air quality dataset file, there are some missing values (include missing values, ‘NR’ values), and in some attributes, the values are mixed with multiple types of data, and some records seem are not correct. All of those phenomena are not good for our data mining operation. For a better data mining result, the data set needs to be ‘cleaned.’ Within this dataset, there are some attributes have missing values, and in order to further discover the dataset, we can use the boxplot graph to find the outliers.



Boxplot of some attributes

From the chart, we can see there are some of the values are abnormal, but most of the quality of statistics are acceptable.

Also, from the Power BI we can see there are some null values, missing values and outliers and extremes.

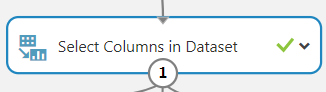


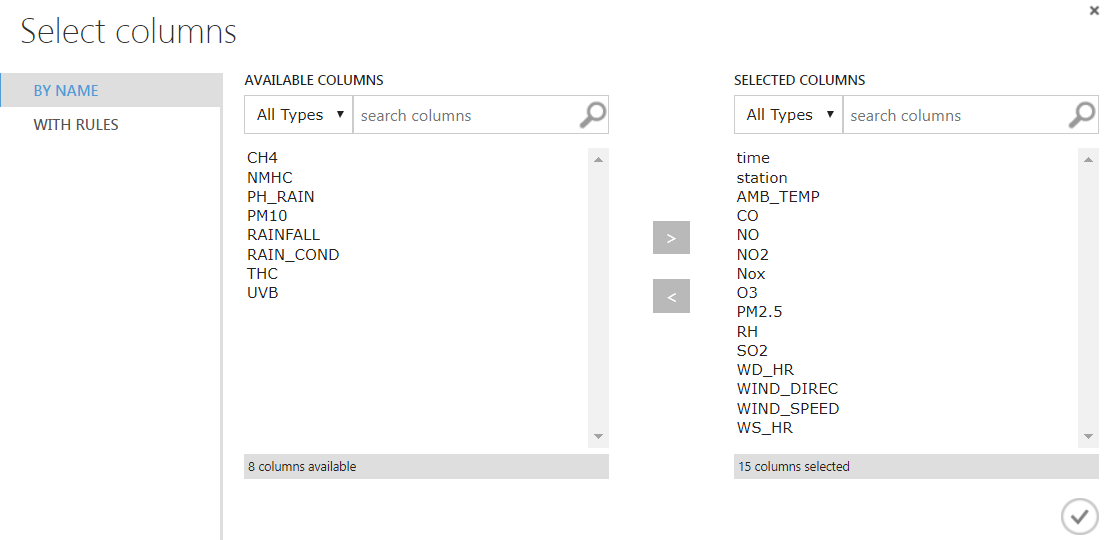
We can see there are some negative values which are not supposed to appear.

So, we need to deal with the outlier and extreme values, and the incorrect negative values.

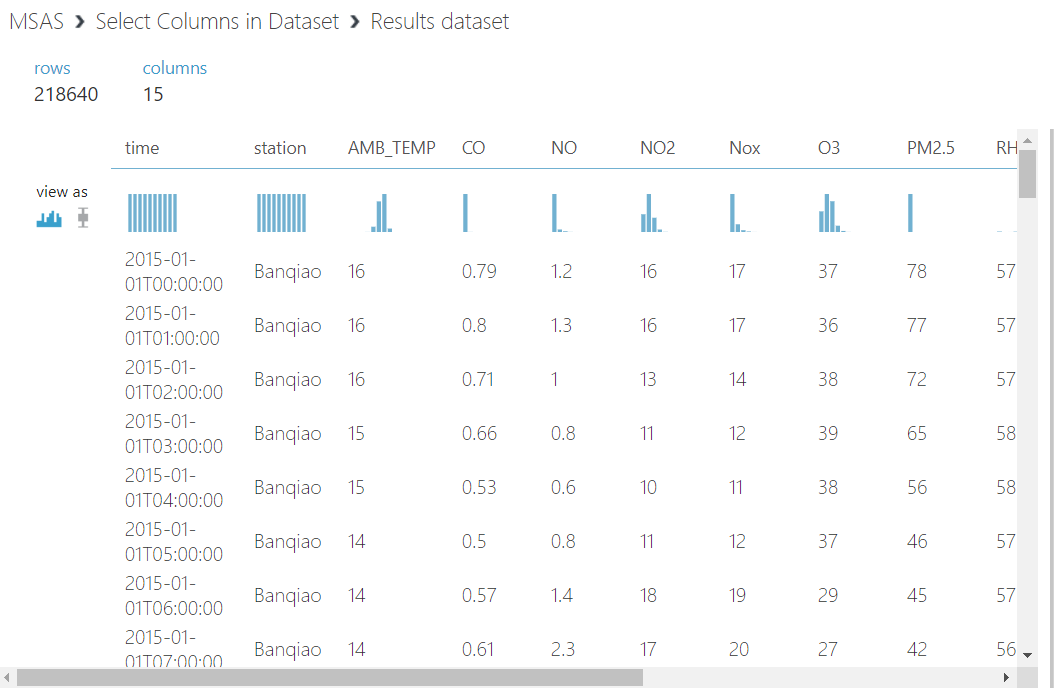
1. **Data Preparation**
   1. **Data Selection**

To select the dataset, we can use the ‘Select Columns in Dataset’ module in the Azure studio.





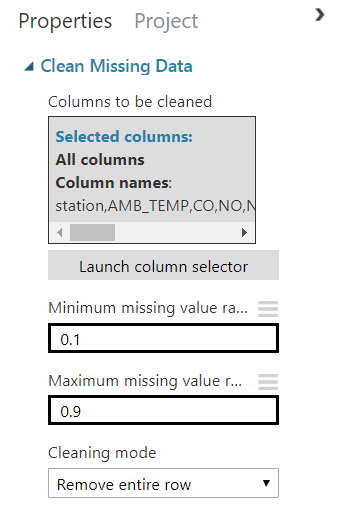
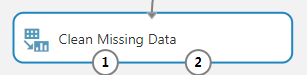
From this table, we know that some of the attributes have too many null or missing values. The PH\_Rain, Rainfall, and Rain\_Cond had too much ‘NR’ values which mean these three attributes can be considered as useless columns. Hence, we will not take them into our data-mining process. In addition, the attribute Pm10 has a strong correlation with Pm2.5, but it doesn’t have too much value, so I will not take it into consider too. Meanwhile, the other attributes also have some missing values, but in an acceptable range.

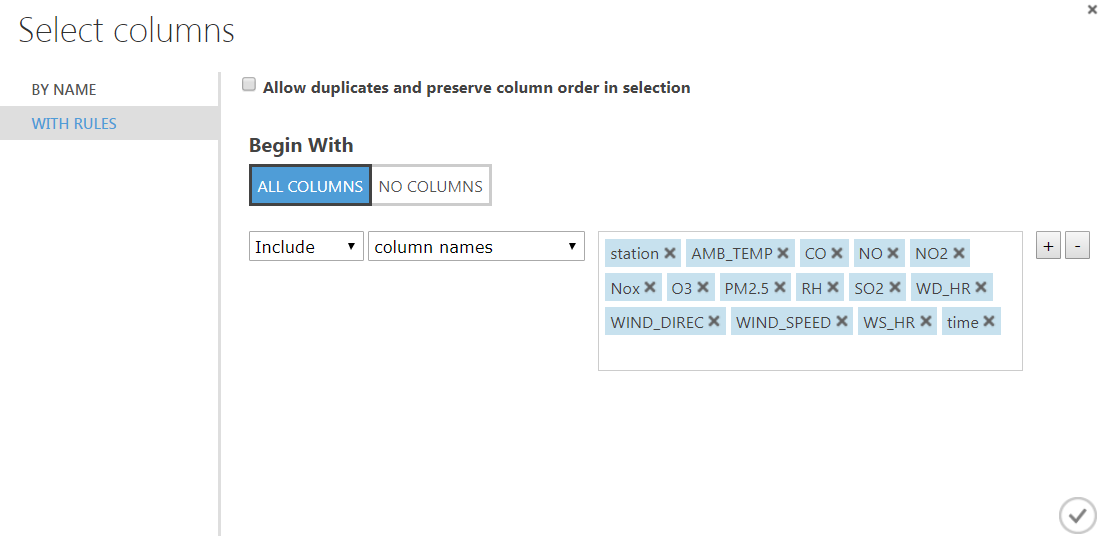


Result after ‘Select’

* 1. **Data Clean**

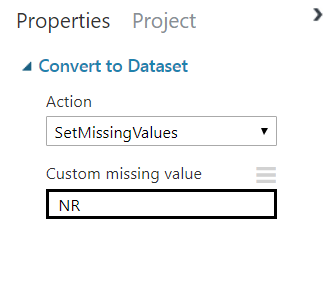
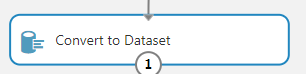
After the data selection phase, now we are going to clean the dataset. From the visualization result, we can notice that most of the attributes’ values are valid. We can use the ‘Clean missing data’ module to clean the dataset very conveniently.



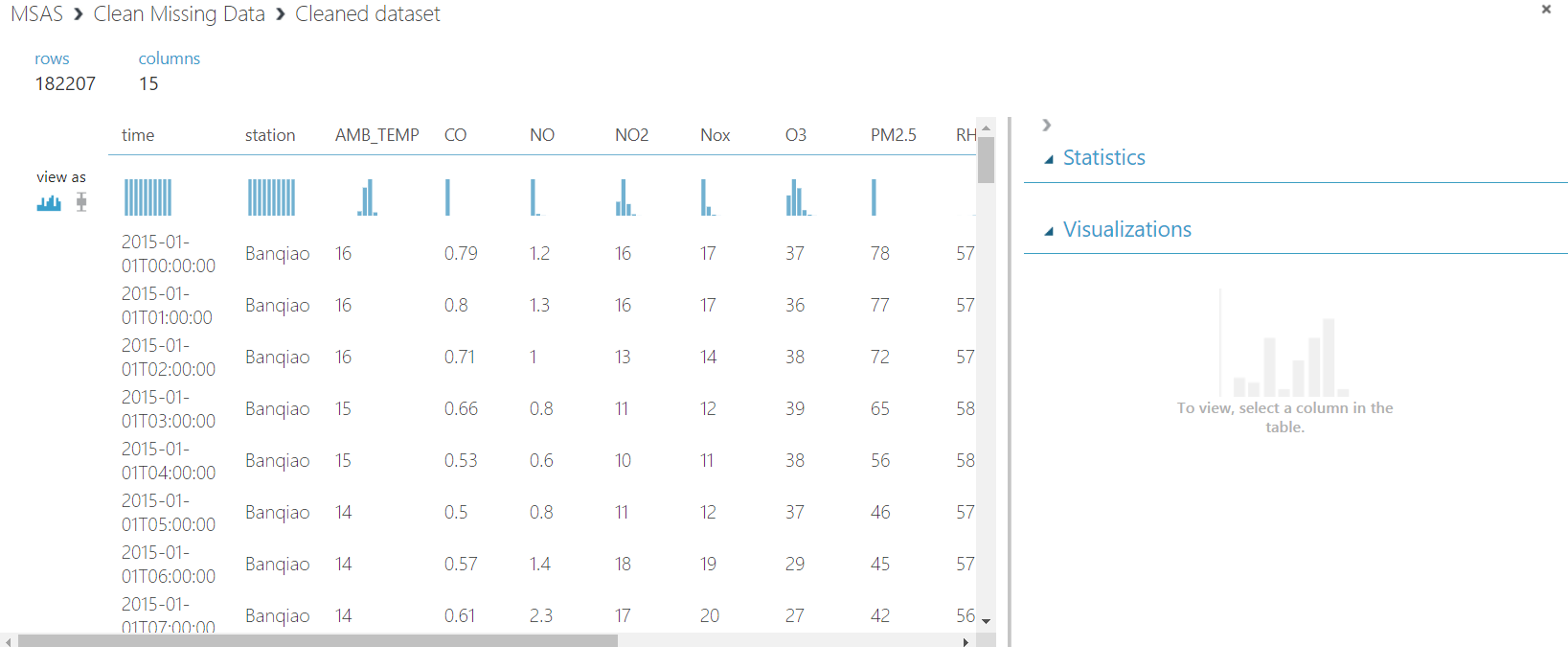


Select the columns that need to be cleaned

The ‘Clean Missing Value’ function can clean the null values, however, in my dataset, there are some values are not null, instead, they are ‘NR’ values, but they mean nothing, so I will set the ‘NR’ values as null and clean them too.



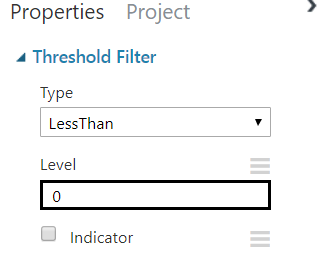
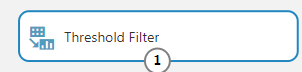
Setting NR as Null values



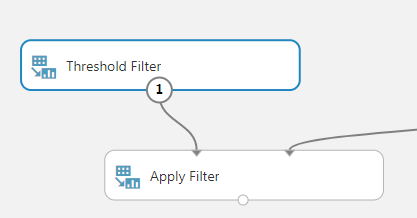
Results After Initial Clean of Null Values

Click the ‘Visualize’ button and we can get the picture above.

We can notice that there are some negative values that should not be presented, so we decide to discard them. We can also use the ‘Filter’ module in the Azure studio.

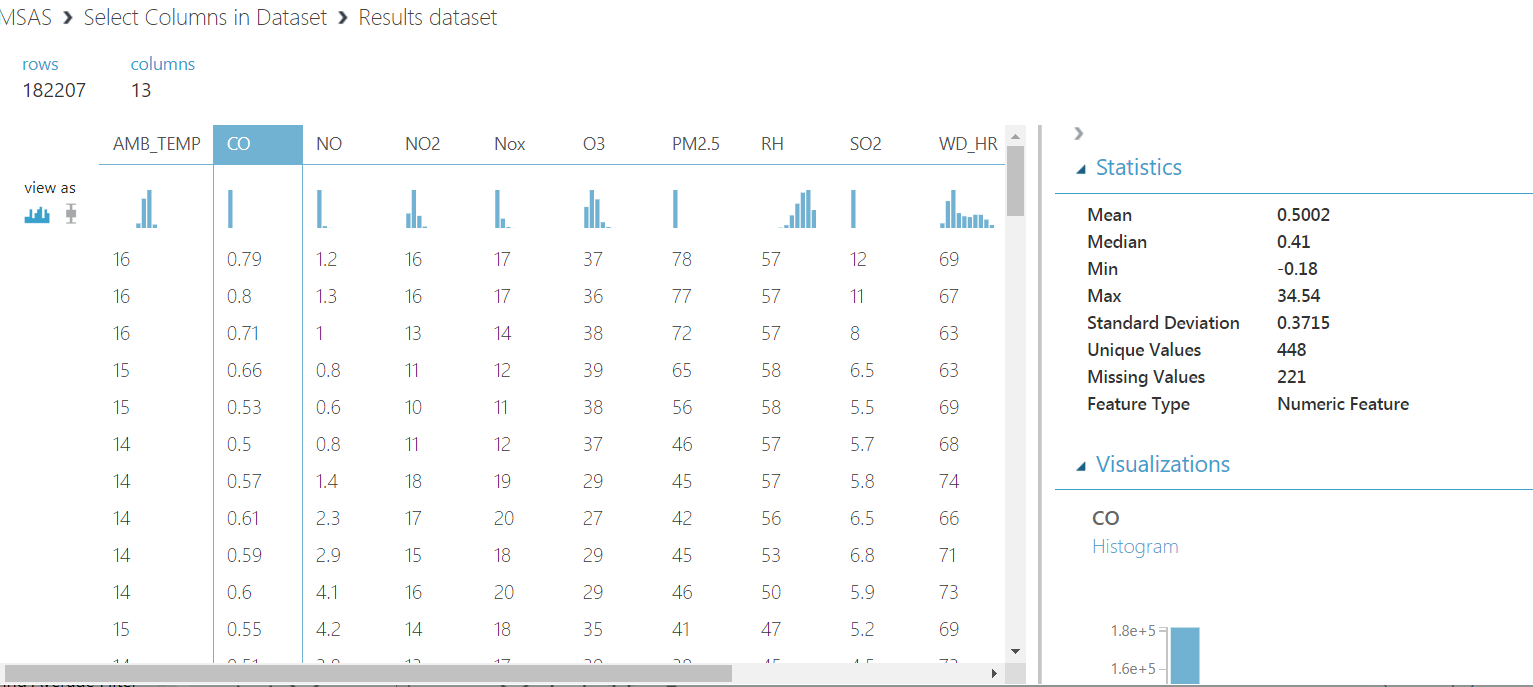


Discard all the negative values



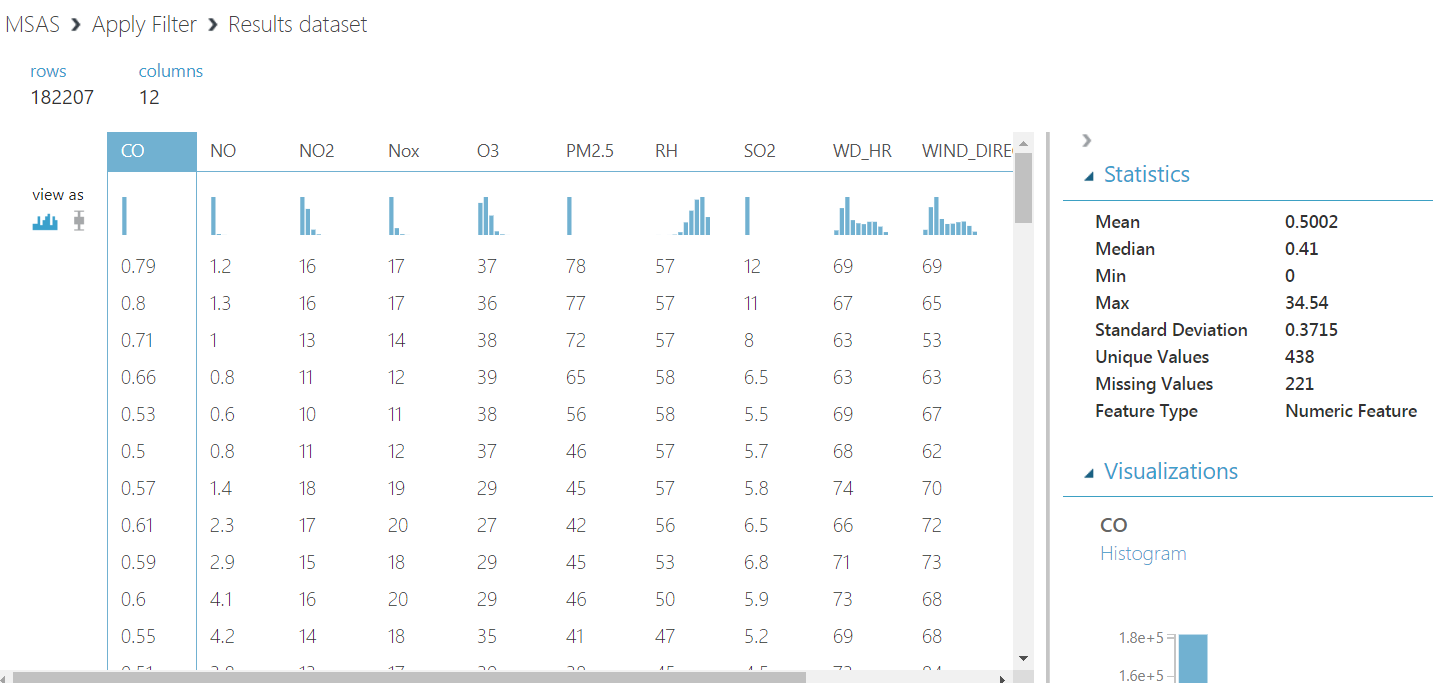
Apply the Filter

Before I run the negative cleaning process, the dataset value is like the screenshot below. In the ‘Statistics,’ the ‘min’ indicates whether the dataset has negative values.

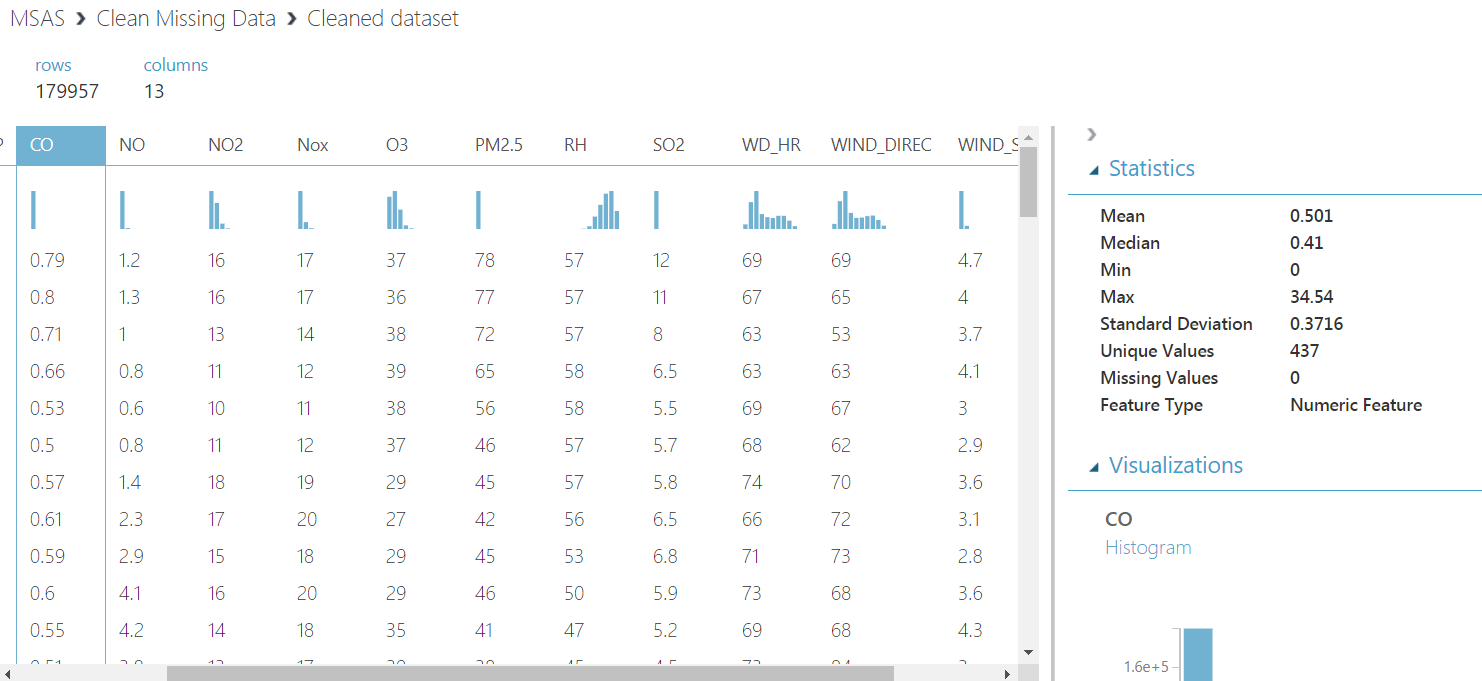


Negative Values Exist

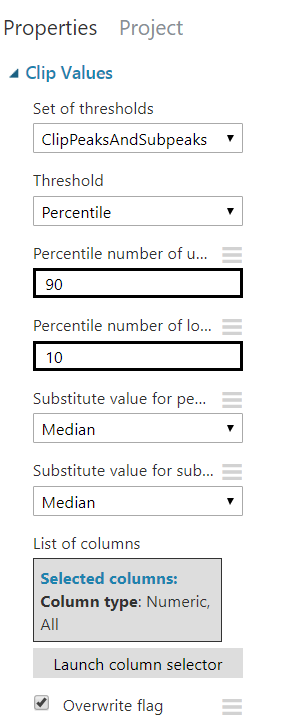
After I run this negative value cleaning process, there is no negative value exists.



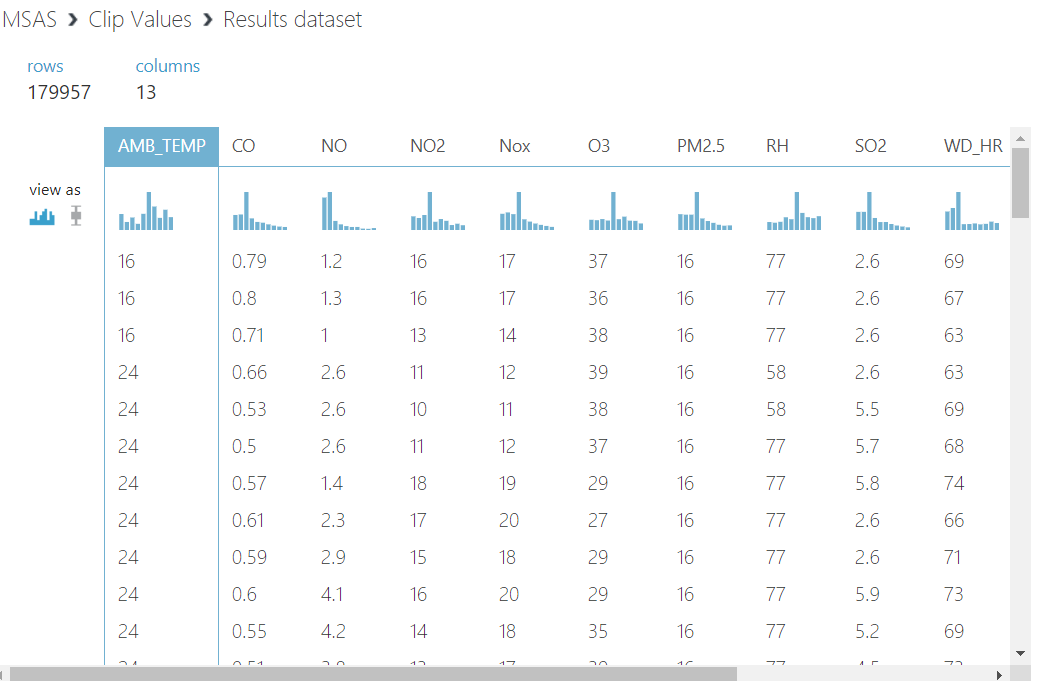
Until now, the dataset was reduced about 40k records. The missing value still exists because the threshold filter will replace all the negative values to ‘Level,’ which I typed ‘0.’ I will clean them by using the ‘Clean missing values’ again.



Next, we will deal with the extreme values. We can use the ‘Clip values’ in the ‘Scale and Reduce’ module.



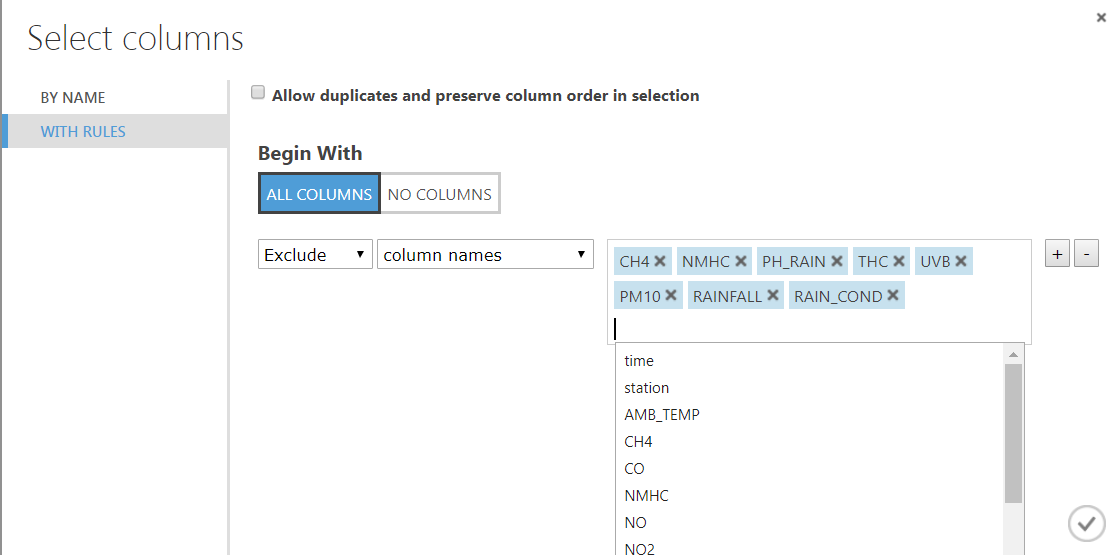
The thresholds are set with ‘Clip peaks and subpeaks’ because I want to discard the highest and lowest values to decrease the influence of extreme values, and I set the substitute method as using median values to replace them.



Result after Clip

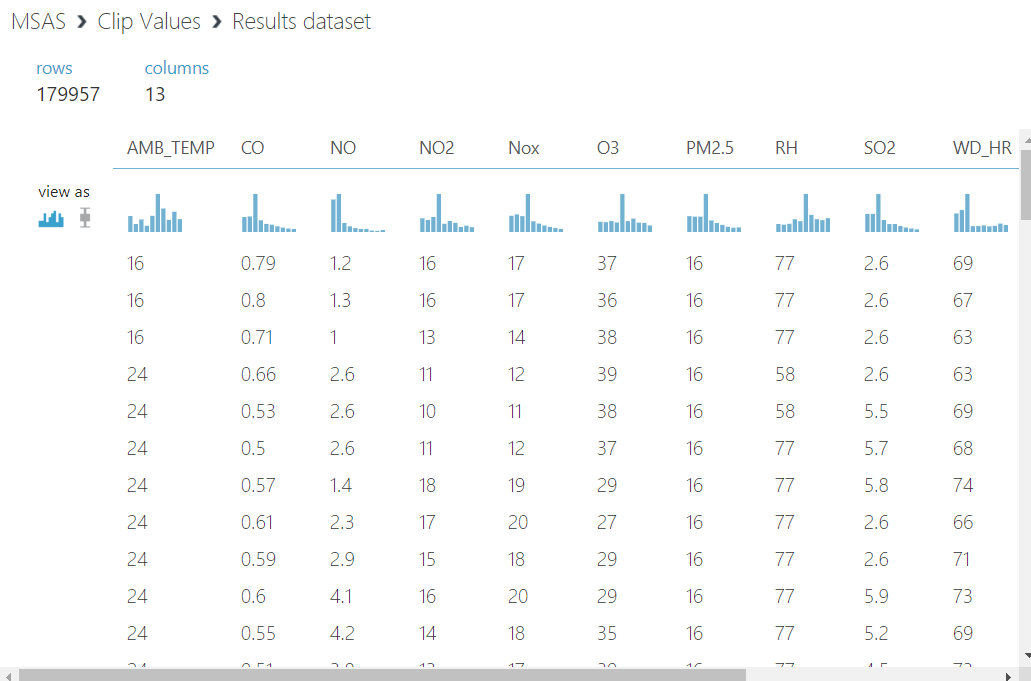
* 1. **Data Construction**

To construct the data, first, I observed that some of the attributes in the initial dataset having too many missing values, as consequence, they may seriously influence our data mining process, so I will use the ‘Select columns in dataset’ to discard them. As I only have one dataset, so there is no need to ‘merge’ different datasets.



Select columns

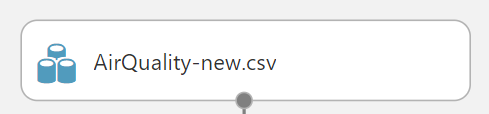
Below is the number of columns and rows I got after clean.



179,957 rows and 13 columns

* 1. **Integrate Data Sources**

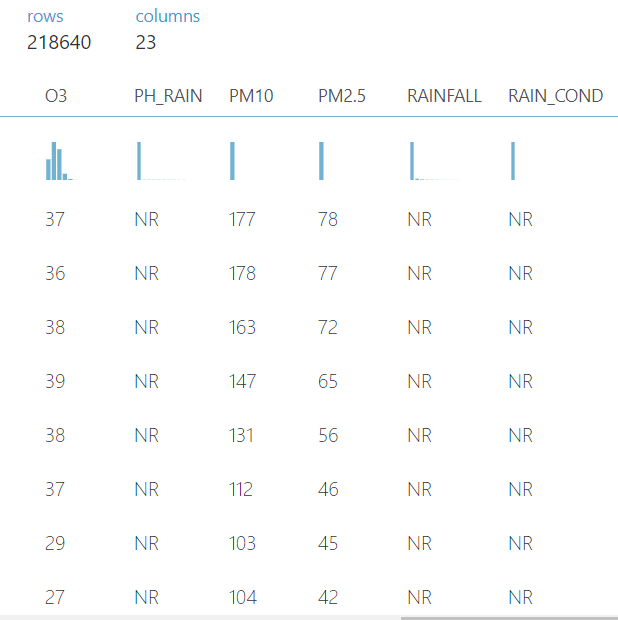
As we only have one dataset that is stored in one single CSV file, so we do not need to use the ‘Merge’ node which integrates several files together. The source file is imported at the beginning.



Source File Imported

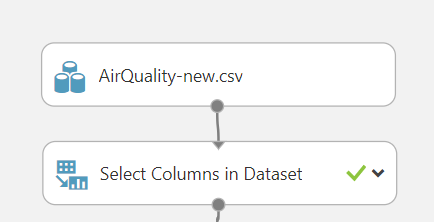
1. **Data Transformation**
   1. **Reduce the Data**

By the observation of the dataset contents, we found that some attributes have too many blank values (‘NR’) or Null values, and compared to the entire dataset these attributes cannot be easily ‘processed’ with the assurance of data quality, also, the Pm10 has too strong correlation with the target Pm2.5, so I decided to drop it too. To make sure the accuracy of further data mining process, we decide to discard these attributes – reducing the dataset.



‘NR’ Values Example

By using the ‘Select columns in dataset’ I can easily remove some attributes.

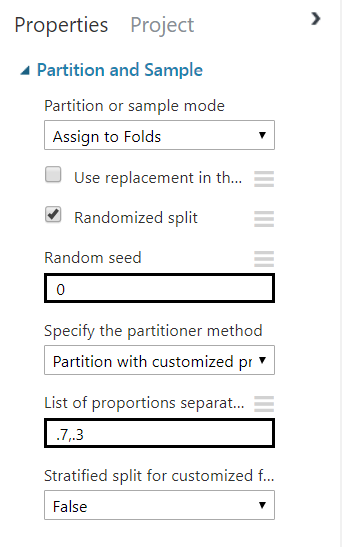
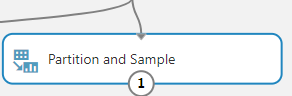


Reduce Columns

In this case, the ‘NR’ value is the blank value which we want to discard. To some extent, we can reduce some noisy data through the data reducing phase.

* 1. **Project the Data**

In this phase, I will do the projection of the data. With “Partition and Sample” module, I can split the data into separate subsets for the training, testing and validation, with percentage of 0.7, 0.2 and 0.1 respectively. By using this module, we can avoid the appearing of the ‘overfit’ phenomenon, as we test and validate the dataset, so the dataset will not be constrained in an over-specified environment.

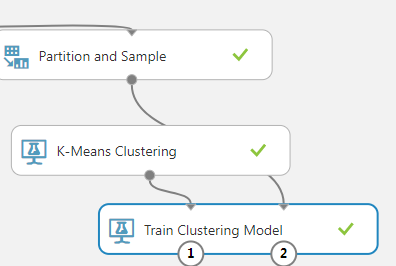


Partition and sample

1. **Data-mining Method Selection**
   1. **Match and discuss the objectives of data mining to data mining methods**

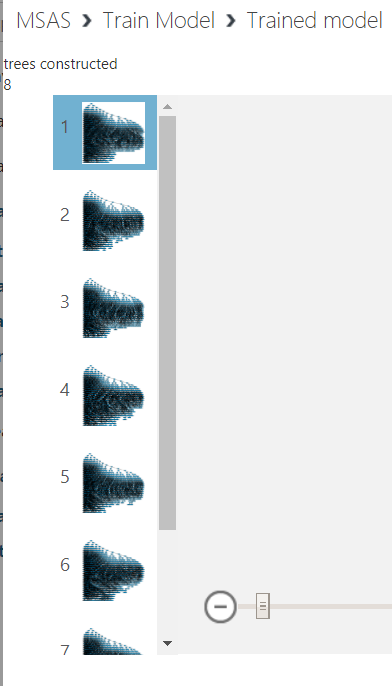
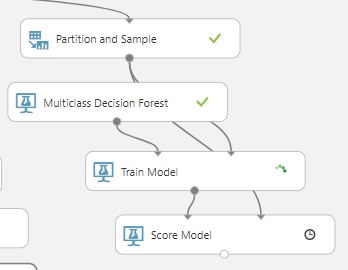
In this case, our goal is discovering the relationship between the various kinds of chemicals in the air and the PM values. In addition, we want to know whether we can predict the PM value if we have the other pollutants’ values, they are our objectives. In general, there are three types of methods – classification, clustering, and regression. The classification is the task of approximating a mapping function from input variables to discrete output variables, its output is usually named label, some classical classification algorithms such as Naïve Bayes, decision tree, and support vector machine. The regression is the task of approximating a mapping from input variables to a continuous output variable, in other words, numeric values. The classical regression algorithms include linear regression, regression tree, and logistic regression. The clustering method is unsupervised algorithm, if we don’t know our target yet, we can use the clustering, such as K-mean algorithm, Hierarchical clustering, and Gaussian mixture model.

Here I tried the clustering model, by using the K-Means algorithm, but the result is not good, which means the clustering is not suitable for my dataset.



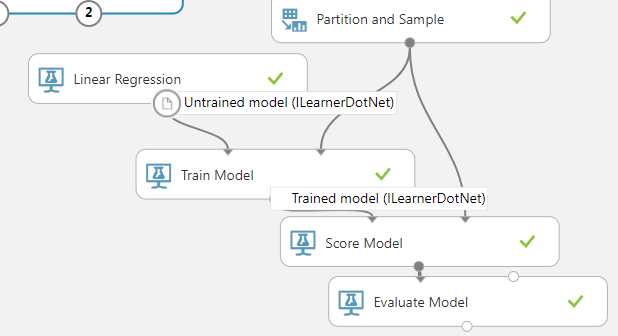
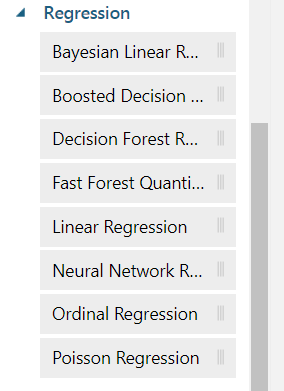
K-Means Clustering

Also, I tried one classification algorithm, which names multiclass decision forest. The classification method is usually used to ‘labeled’ a thing, such as predicting whether it will rain tomorrow, the result ‘yes it will rain’ and ‘no it will not rain’ are the labels.



Multiclass Decision Forest Classification

In my dataset, both the target attributes and the other attributes are numeric values, so it is better for this case to use the regression method to do the data mining work.

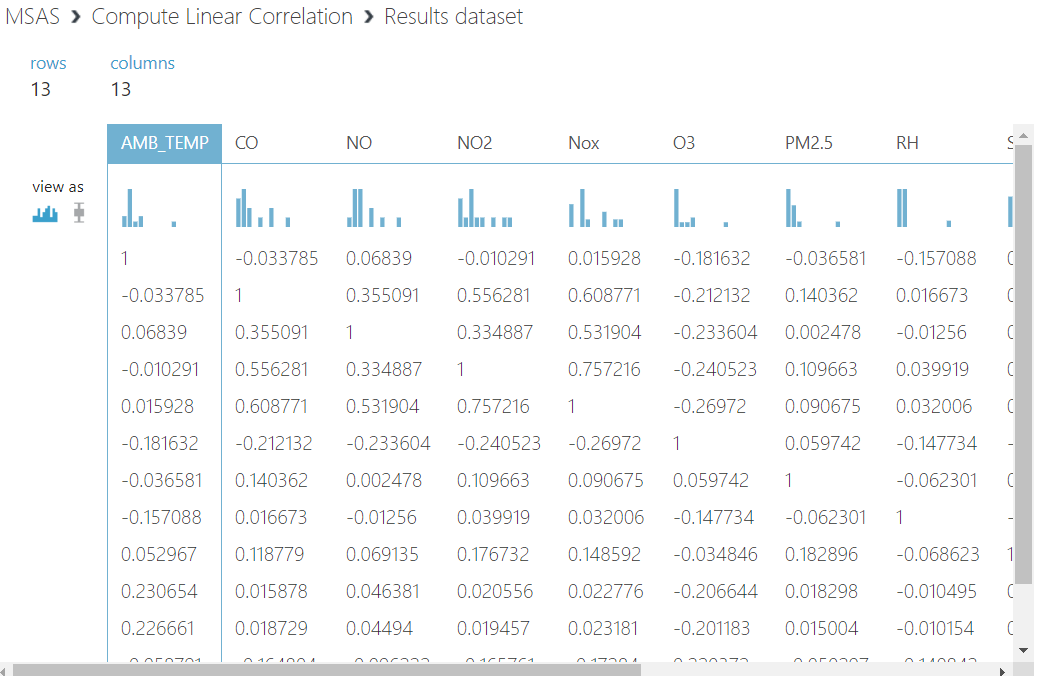
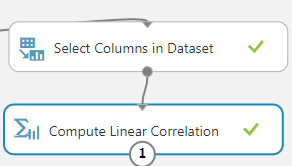


* 1. **Select the appropriate data-mining method(s) based on discussion**

Within the Azure machine learning studio, we can find the regression algorithms under the ‘Initialize Model’ tag. There are Bayesian linear regression, boosted decision tree regression, linear regression and so on. As the discussion above, the clustering and classification are both not very suitable for this dataset, so here I choose the regression algorithms, from the simple linear regression, to the neural network, I will try different regression algorithms and find a better one based on their results.

1. **Data-mining Algorithm(s) Selection**
   1. **Conduct exploratory analysis and select algorithms**

To begin with, here I can use a module called ‘Compute Linear Correlation’ to initially analyze the dataset.



Correlation

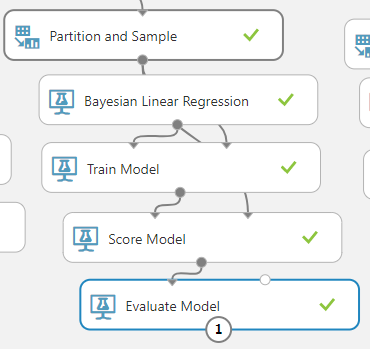
From the screenshot above, we can notice that the attributes have correlations to a certain degree, so we can use the regression algorithms, however, because the classification and regression have some overlapping aspects, so I would like to try some classification algorithms later.

* 1. **Select Algorithms**

First, under the ‘Regression’ tag we have 8 different regression algorithms in total. Based on the discussion before, I’ve already tried the linear regression, then I will try some others, such as Bayesian linear regression, decision forest regression, and neural network regression which will be compared to the two-class neural network classification algorithm.

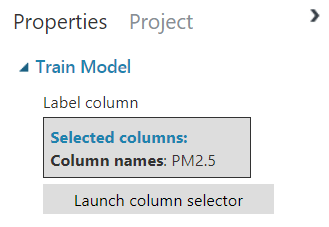
* 1. **Build/Select appropriate model(s) and choose parameters**

Now we will start running our models. To begin with, we need to connect all the necessary modules.



Bayesian Linear Regression

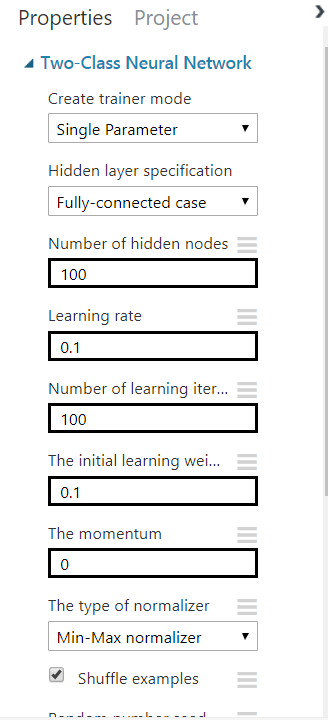
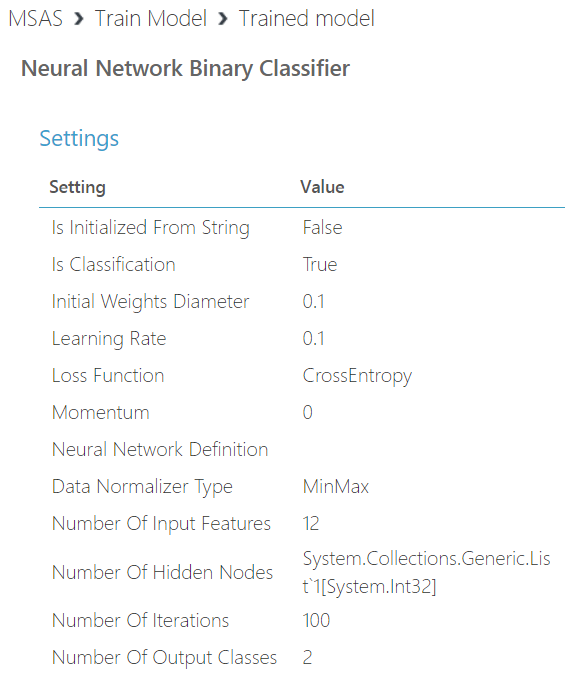
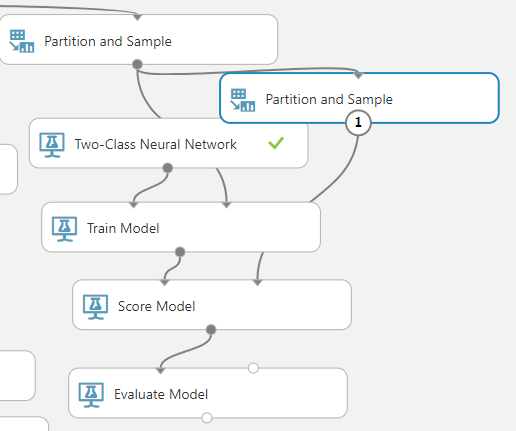
After finish all the preparation work, then we click ‘run.’



Train Model Target

First, we set the PM 2.5 as our target, and then we connect the ‘score model’ and the ‘Evaluate model.’ As we can see, a model stream has been built, and we can apply this to other algorithms.

Now I will try a classification algorithm – Two-class neural network.



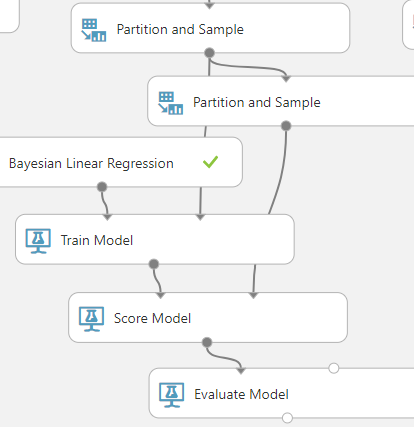
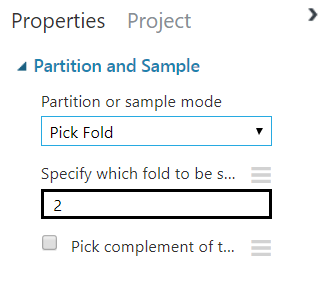
Two-class Neural Network Settings (Classification)

1. **Data Mining**
   1. **Create and justify test designs**

The ‘Partition and sample’ module is used to separate the dataset to different subsets, training, testing and evaluation. By using this, we can avoid the ‘overfit’ phenomenon at a certain degree, that would make the entire design has a more solid result. Before each algorithm is executed, I assign a ‘Partition and sample’ module to them



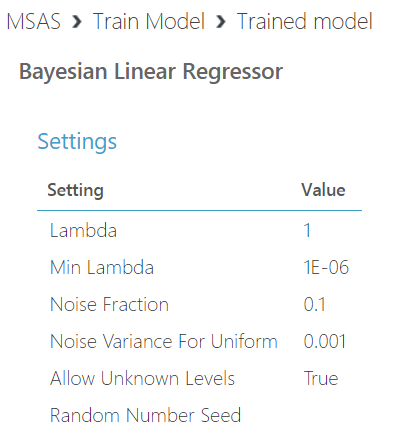
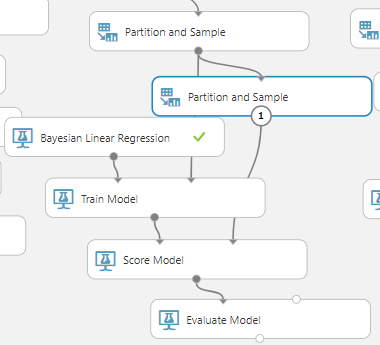
By using the ‘Picking fold’ function, I can easily assign the specified fold to the correct module.



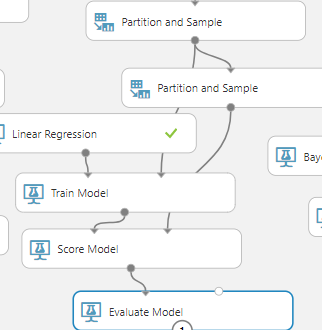
Partition Module

* 1. **Conduct data-mining – Classify, Regression, Clustering**

First, I will run the regression, start with the Bayesian linear regression algorithm.

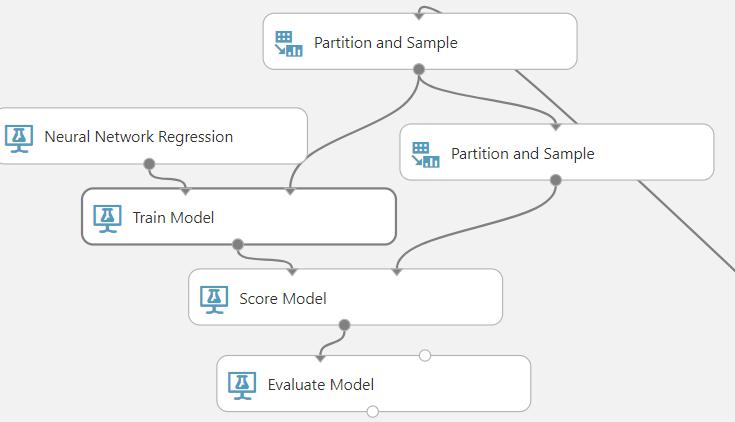


Bayesian Linear Regression

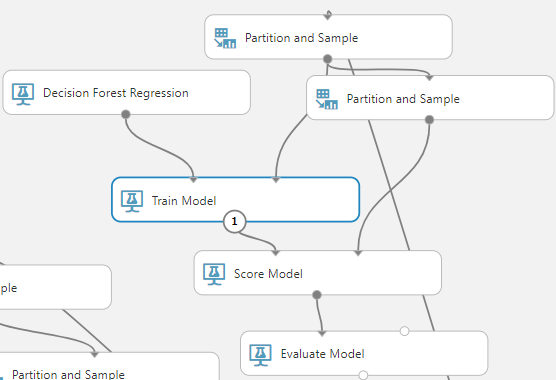


Linear Regression

The other regression algorithms conduct like the stream above.

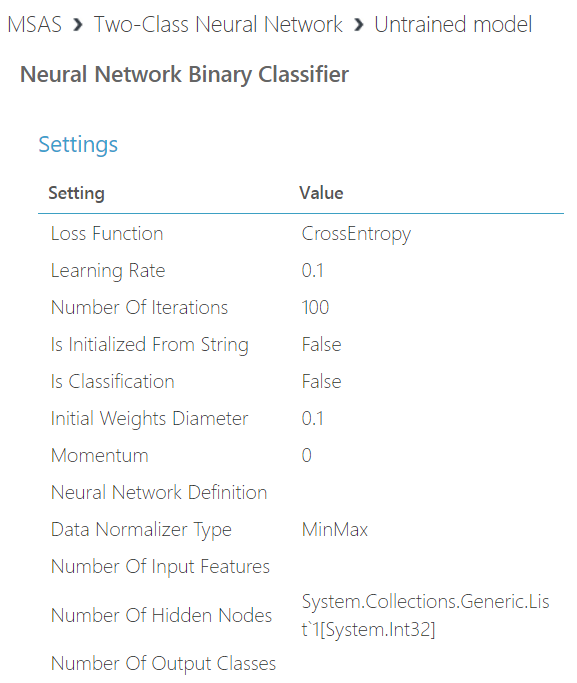
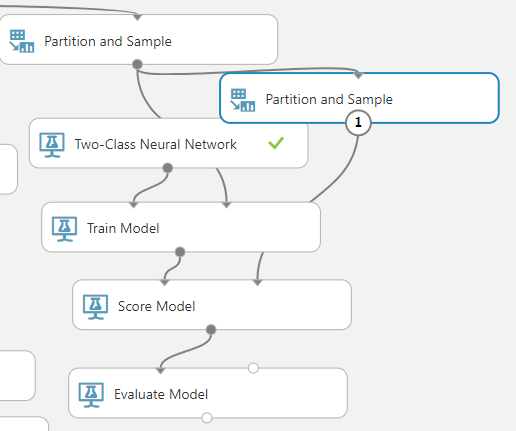


Neural Network Regression



Decision Forest Regression

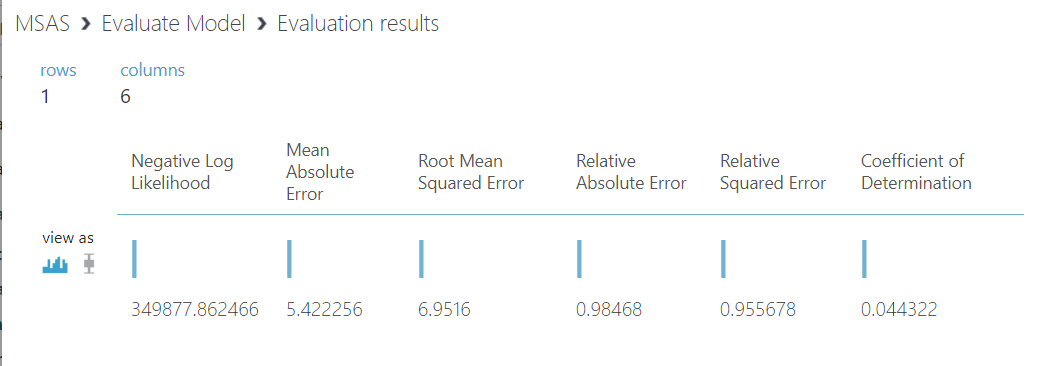
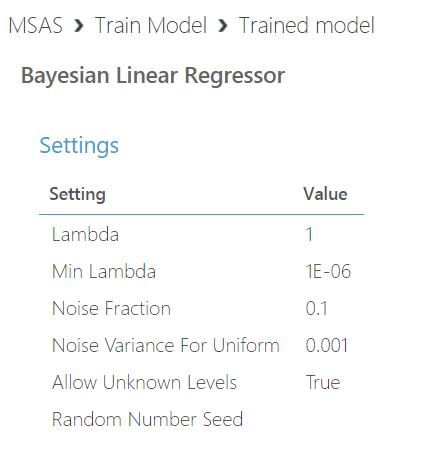
Next, I will construct the classification model – two-class neural network.



Two-class Neural Network

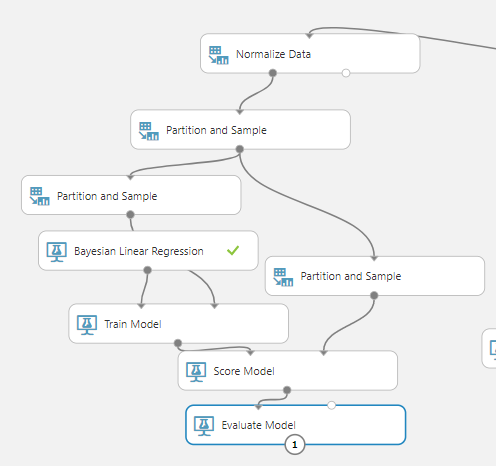
* 1. **Search for patterns**

First, I will run the Bayesian linear regression model.

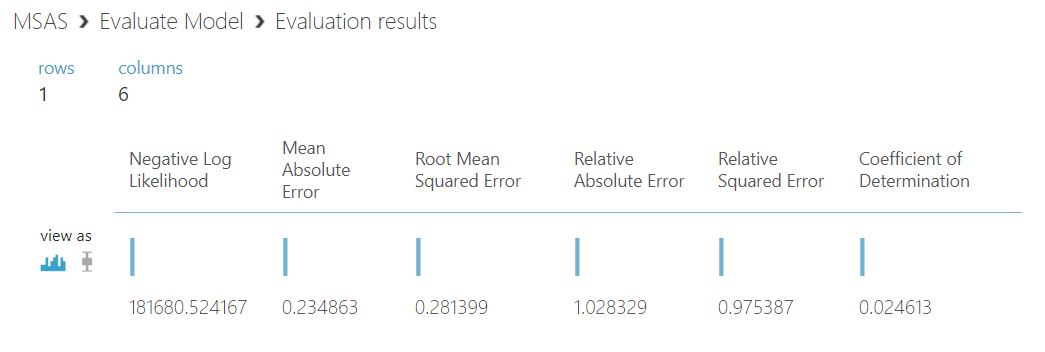


Bayesian Linear Regression

Here we can notice, the result is not very good, so can I improve this if I try to first normalize the data? So, I built another Bayesian linear model.



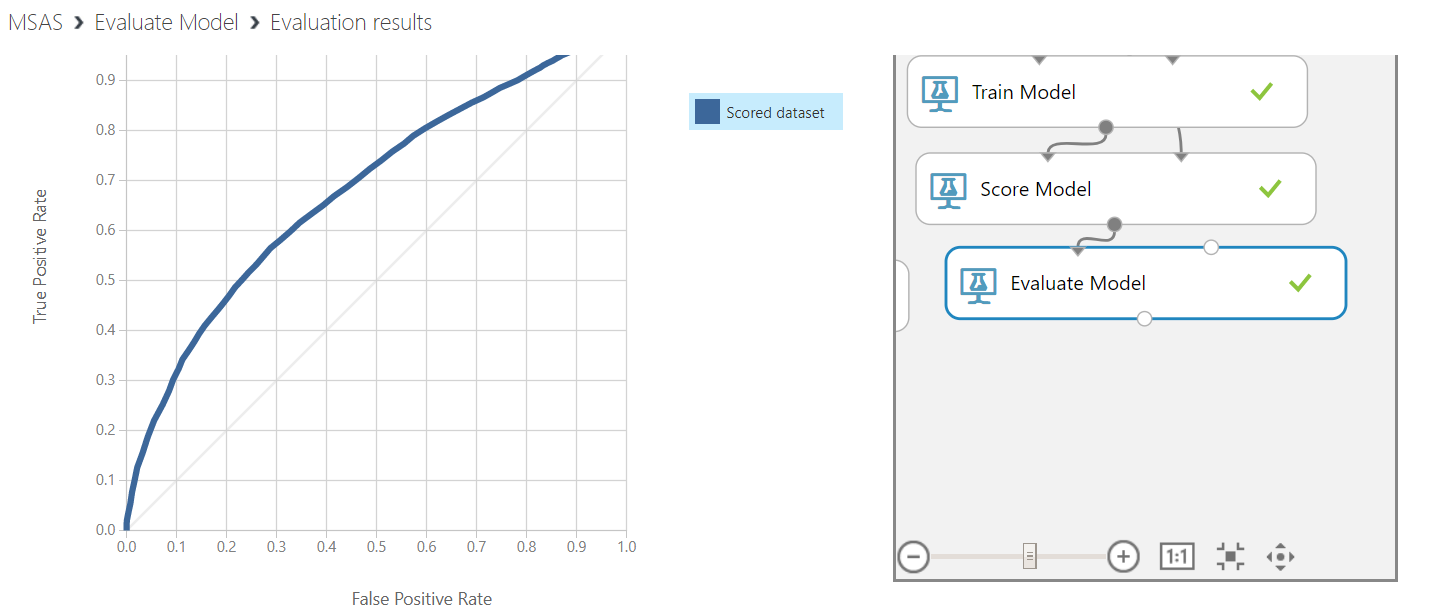
Normalized Bayesian Linear Regression

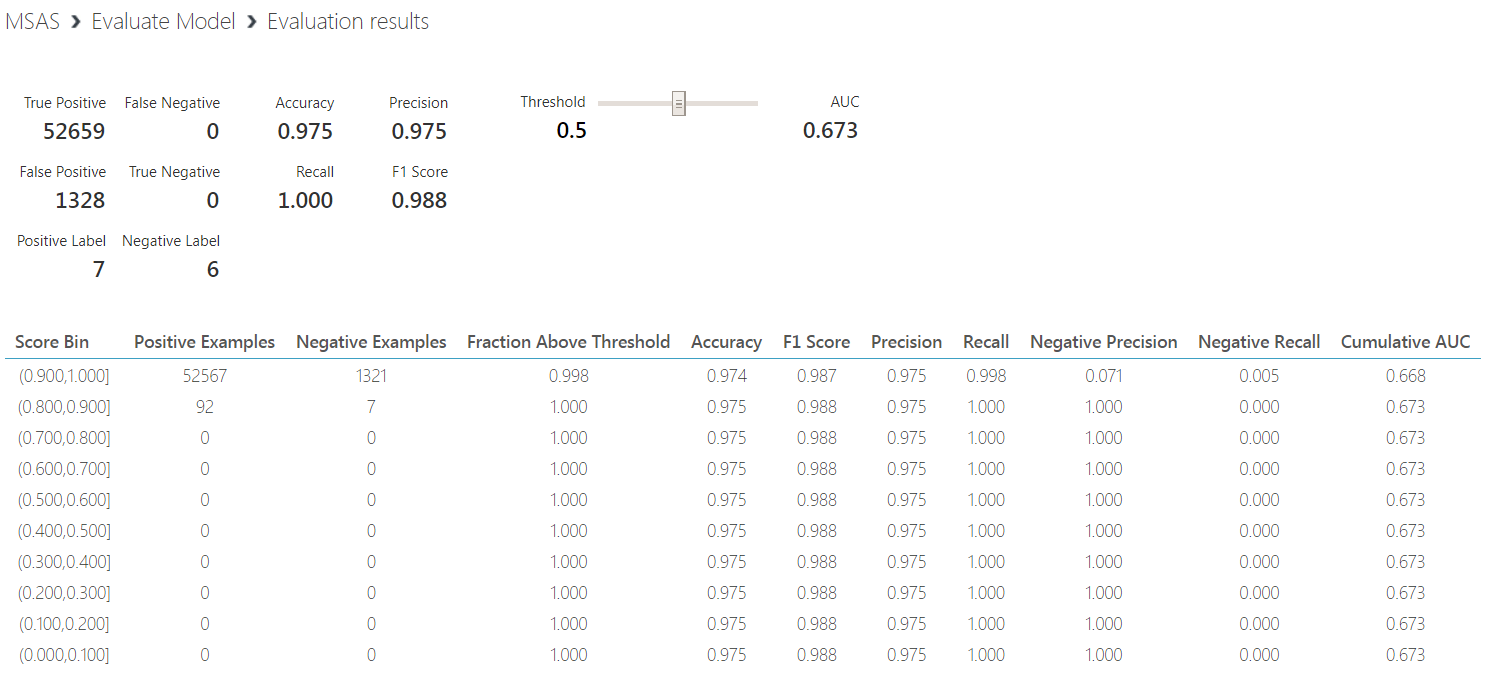


Normalized Bayesian Result

Based on the pic above, the result is improved on a certain degree. (The decrease of mean absolute error)

The next result is the classification, two-class neural network.

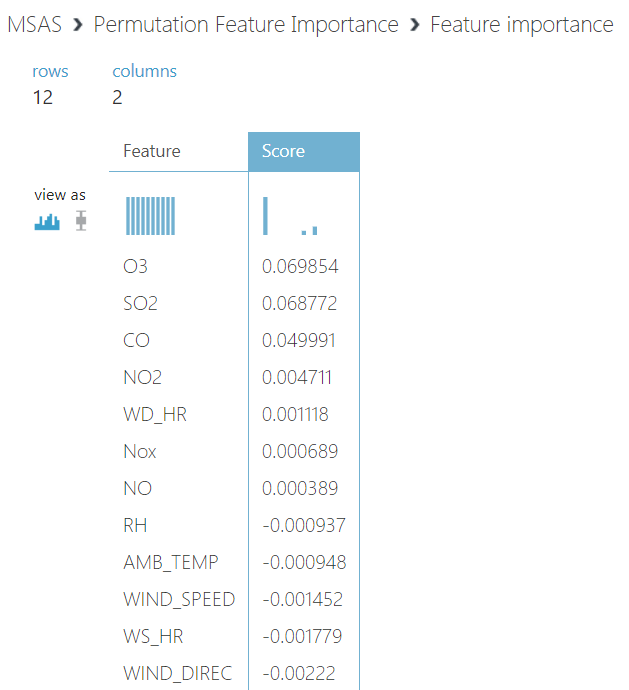
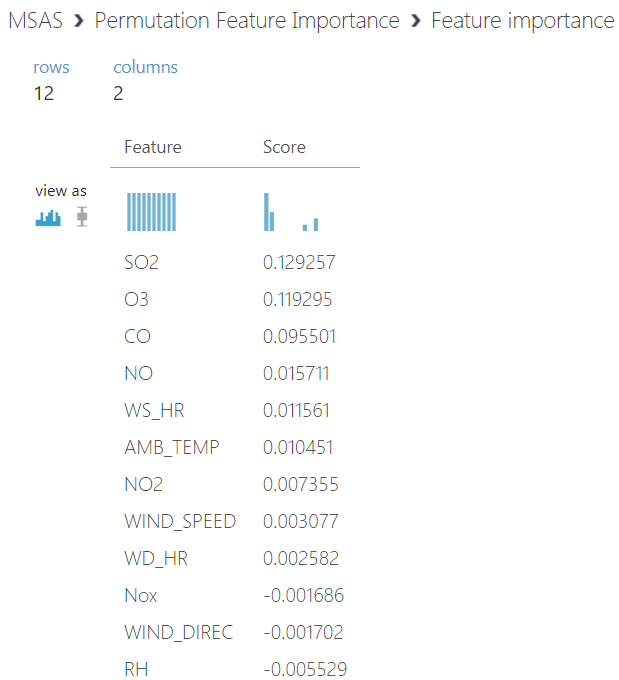




Two-Class Neural Network Results

1. **Interpretation**
   1. **Study and discuss the mined patterns**

After running all the modules, here we got a lot of results. Now I will connect a ‘Permutation feature importance’ module to discover the different importance of variables.



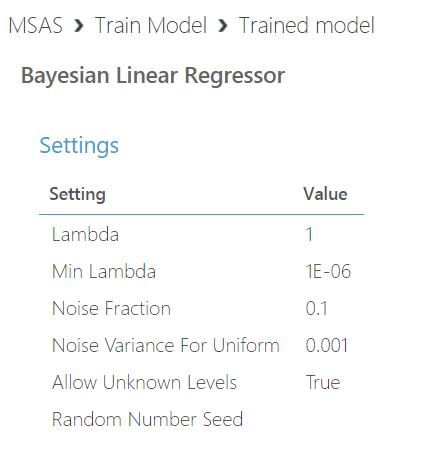
Feature Importance

The left one is the Bayesian linear regression, we can see the SO2 is the most important predictor in this case, while the right one is the normalized Bayesian linear regression, the O3 is the relatively most important predictor. The predictor importance indicates us that it is true if there are more pollutants like SO2, O3, NOx and so on in the air, the PM value is expected with a high value.

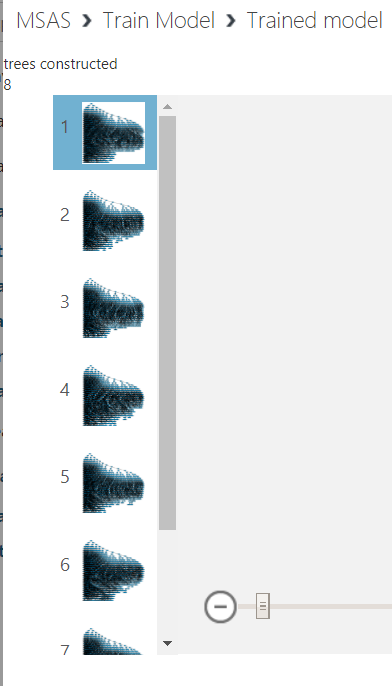
In my opinion, this result is understandable as it is suitable to our common knowledge. Also, if all the pollutants indicators are aggregated together, their result might have a stronger relationship with the PM value. But according to this, we can have a view about which kind of pollutant contributes the most to the air pollution.

* 1. **Visualize the data, results, models, and patterns**

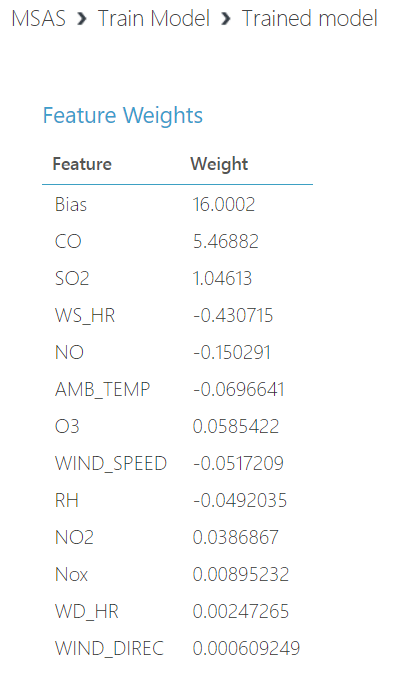
As the model building process has been showed above, so here I will skip the screenshots of model.



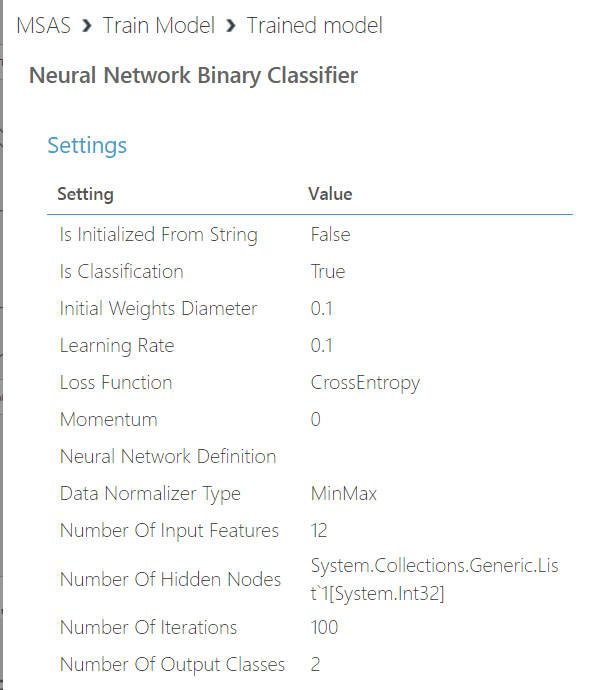
Bayesian Linear Regression



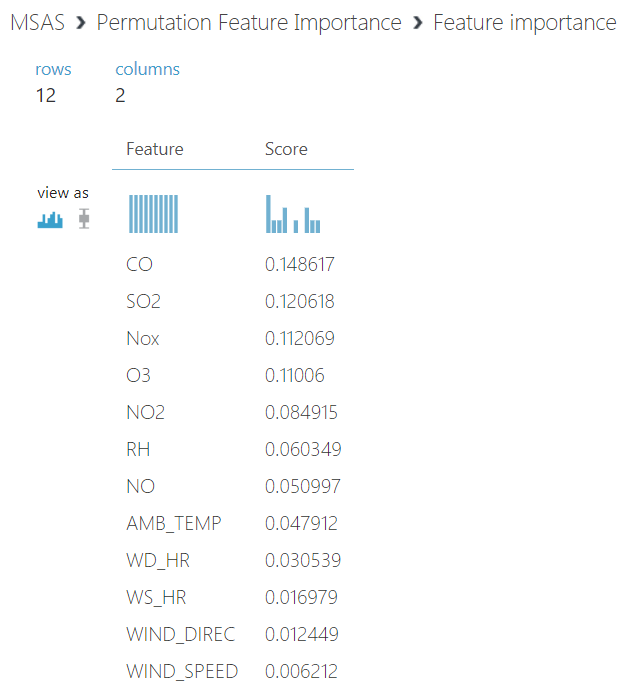
Multiclass Decision Forest



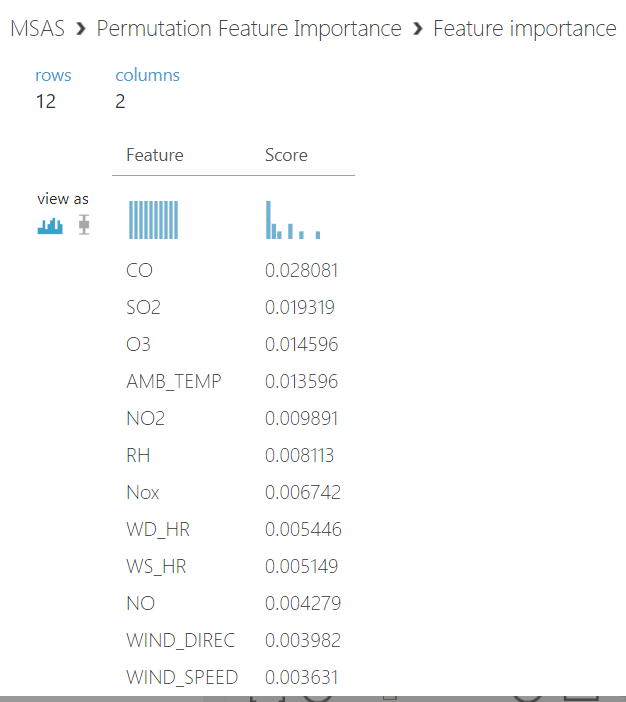
Linear Regression



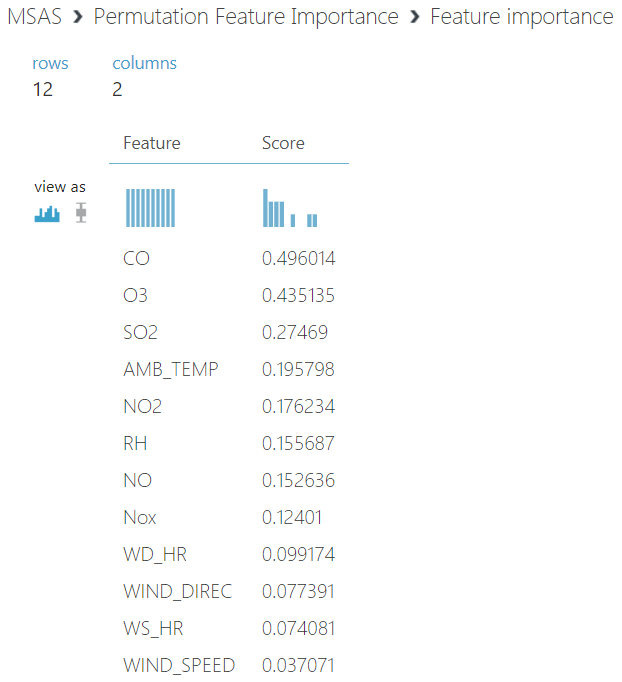
Two-class Neural Network



Neural Network Regression Feature



Multiclass Decision Forest Feature



Decision Forest Regression Feature

From the pictures above we can see that in the decision forest regression, neural network regression, and multiclass decision forest, the CO is the most important predictor, which is different with the linear regression and Bayesian linear regression.

* 1. **Interpret the results, models, and patterns**

In this phase, we will interpret these models. From the different models we can infer that, first, if there are more chemical pollutants in the air, especially the CO, SO2, O3, the PM2.5 value will increase to a high level; second, if the wind speed is relatively high, the PM2.5 value will decrease.

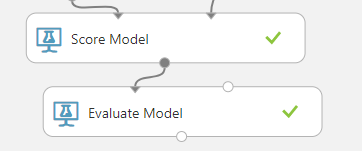
For the linear regression and Bayesian linear regression algorithm, the most important predictors are SO2 and O3, with the importance number 0.129257 and 0.069854 respectively. However, for the other algorithms, the CO becomes the most important predictor. (0.148617 for neural network regression, 0.028081 for multiclass decision forest, and 0.496014 for decision forest regression)

It is reasonable enough for us to understand, and also, it is useful for our environmental control department. For example, we can see that CO has a serious impact on the PM2.5, so the environment department now has a clue – figure out the where does the CO come from, and how it be produced and exhausted into the air. Such model indicates the things that influence our air quality. Also, we can utilize the different ‘contribution’ of air pollution from the various kinds of chemical pollutants to make up a specialized plan, which will solve the air problem based on the importance of predictors one by one.

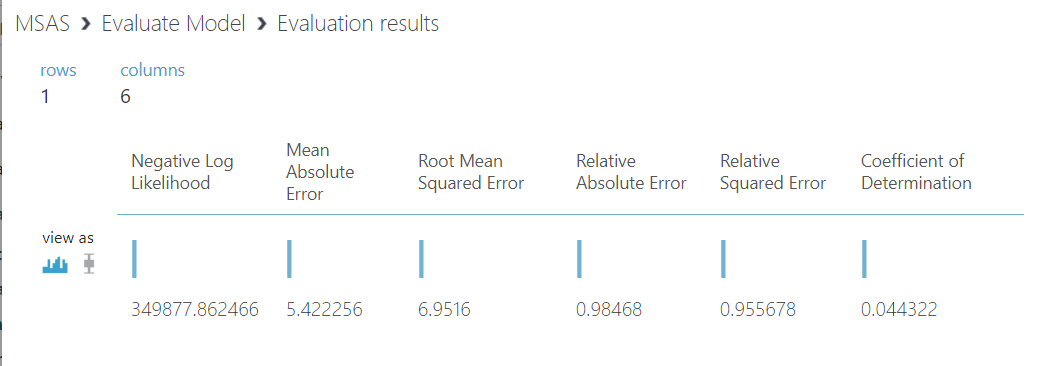
In this case, we need to care more about the CO, SO2, and O3, as they are the predictors contribute the most.

* 1. **Assess and evaluate results, models, and patterns**

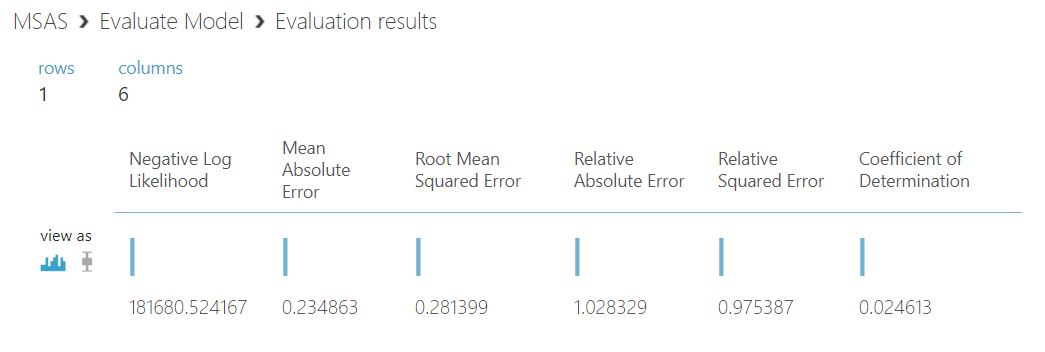
In order to assess our results, we can use the ‘Evaluate model’ module within the Azure studio.



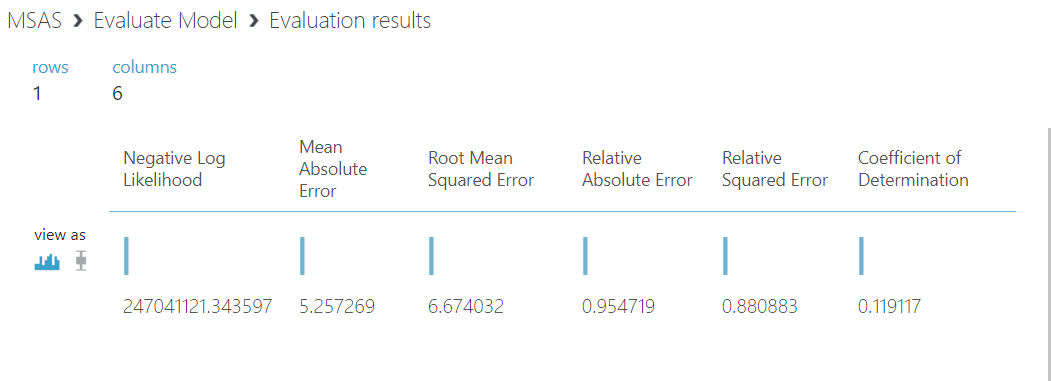
Evaluate Model



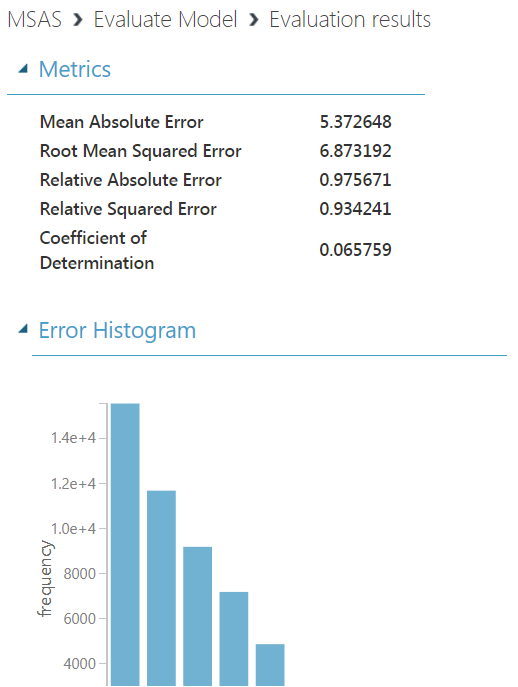
Bayesian Linear Regression Evaluation



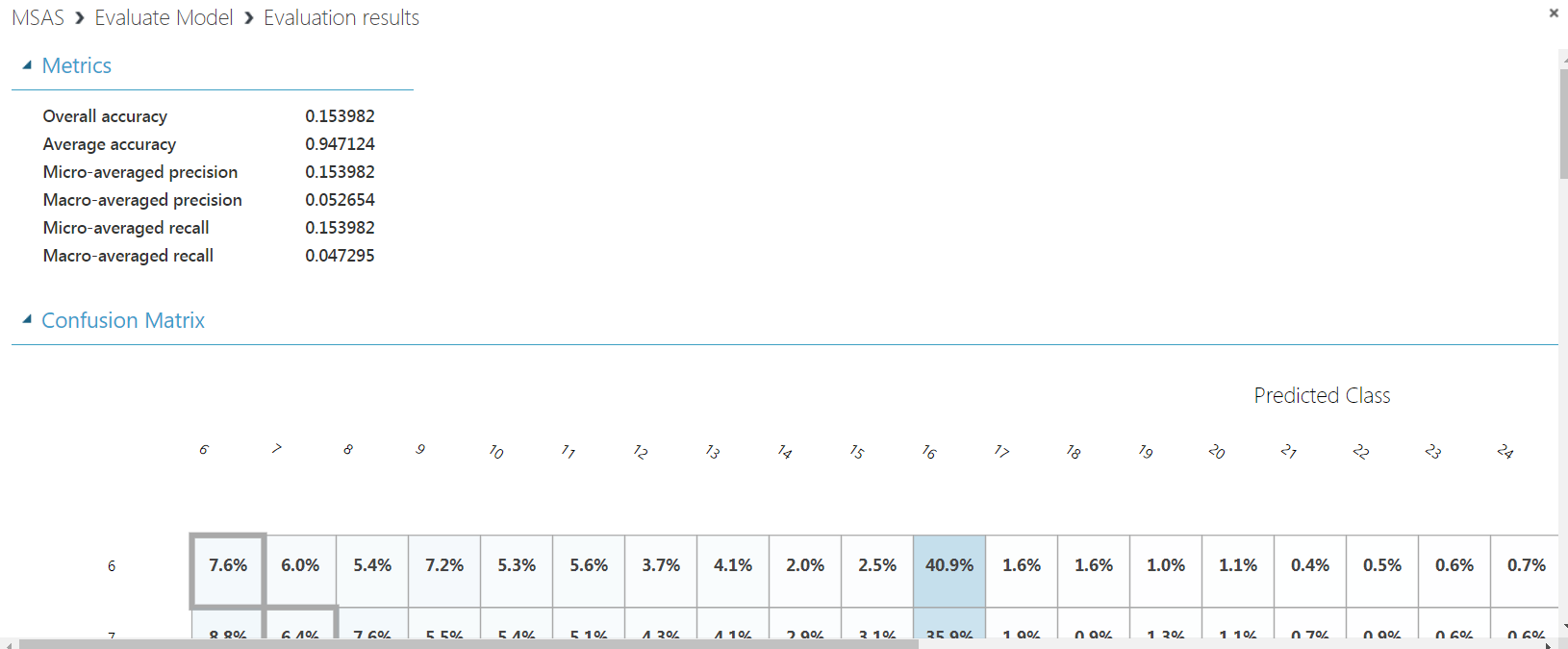
Bayesian Linear Regression Evaluation (Normalized)



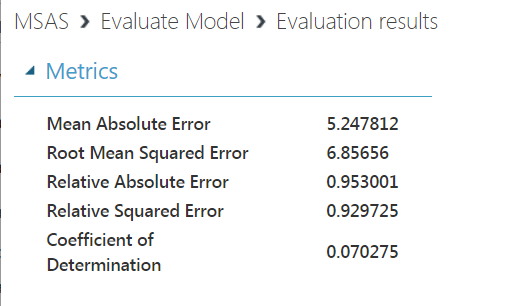
Decision Forest Regression Evaluation



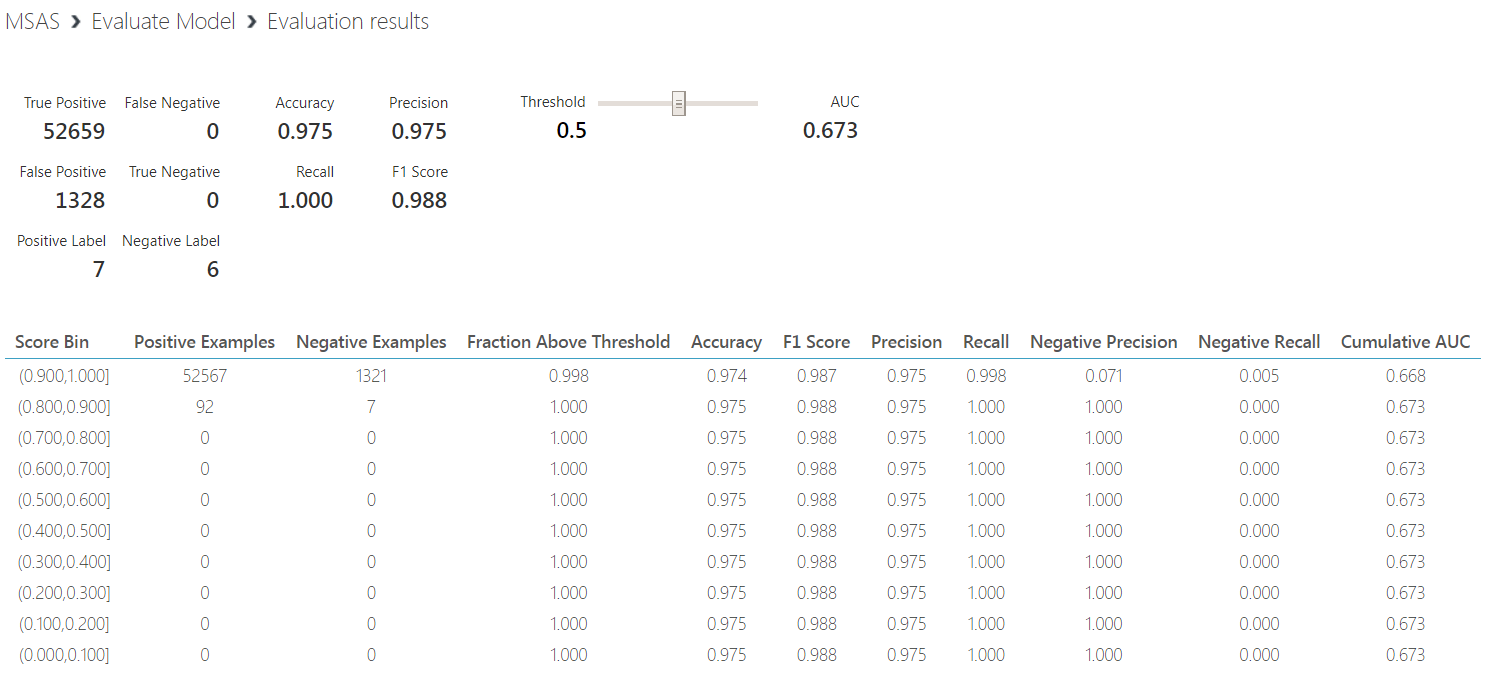
Linear Regression Evaluation



Multiclass Decision Forest Evaluation



Neural Network Regression Evaluation



Two-Class Neural Network Evaluation

From these evaluation results, we can see that the normalized Bayesian linear regression has the lowest mean absolute error, while the two-class neural network has a high accuracy with 97.5% which is pretty high. The result multiclass decision forest is not that optimistic; however, the rest of the models are not very bad.

By analyzing these models, our environment department may have a better understanding of the air quality and have a clearer view of what are the main pollutants in the air, also, how the external factors such as wind speed, humidity and so on, influence the PM2.5 concentration, and our physical health. In addition, by tracking the source of these pollutants, the environmental department may come up some solutions to deal with the ‘pollutant-producers.’ The data mining process reveals the relationship between air quality and chemicals, and it helps people to determine the priority when they want to deal with the pollutants.

* 1. **Iterate prior steps as required**

The data mining process is iterative, as we need to discover and try different algorithms to find the better result. In this case, for example, the first time I run the Bayesian linear regression, the mean absolute error is a little high, so I thought if I try to do something with the data, does the result will be better? So, I iteratively build the Bayesian linear regression model again, the difference is this time, the dataset was normalized by the ‘Normalize’ module within the Azure studio. Also, in the data cleaning step, it is very often that we need to execute the cleaning operation iteratively, and then we can get the dataset that we want.

In general, the data mining job requires us to do plenty of iterative job, and then we can have a robust and accurate model.