**INFOSYS 722: DATA MINING AND BIG DATA**

**Iteration III-OSAS**

**Meteorological effect on**

**Air quality in Beijing**



**By**

**Tianyi Yang**

**Tyan227**

# 1.Business and/or Situation understanding

### Identify the objectives of the business and/or situation

Over the past decades, China keep developing with a fastest economic growth in various fields, but at the same time, the country also paid a heavy environmental price. China is now facing a great challenge of the infamous “airpocalypse” --air pollution kills thousands of people every year.

The quality of urban air quality is closely related to the seasons and meteorological conditions[[1]](#footnote-2) like wind, rain, air pressure and temperature. Due to China's vast territory, collecting meteorological data from each region can make the process of collecting data very complicated, which will adversely affect the pre-processing of the data and the efficiency of the generation of models. On the other hand, factors like the terrain, elevation, population density, and degree of urban development in different regions of China vary notably and may interfere with the final result of the report.

Therefore, the goal of this report is to study the how major meteorological factors may affect the quality of Beijing, the capital city of China, in the absence of major changes in the emissions of pollution sources. In order to achieve this goal, we will compare the influence of different meteorological phenomenon to the air quality and find the key factor. Alternatively, it may turn out that the meteorological factors are not determinant for the air quality. The goal of this report is completed by getting either of these two results.

### 1.2 Assess the situation

## Air is one essential element to human-being. As a matter of fact, a person inhales around 14,000 liters of air every day[[2]](#footnote-3), in this case, any contaminations in the air can dramatically affect people’s health little by little. What’s worth, patients with pre-existing respiratory and heart conditions, including the young and old, are particularly vulnerable towards bad air quality.

## Air pollution is a complex phenomenon, and air pollutant concentrations are affected by many factors at specific times and locations. Most of the air pollution in many cities are caused by coal-fired smog, as well as automobile exhaust and suspended particulate matter. Their combined effect makes air pollution even worse.

In these circumstances, China has declared war against air pollution by reducing the production of steel and of coal-fired electricity, setting restrictions for vehicles on road, and making investment in sustainable energy like solar and wind power.

### 1.3 Data mining objectives

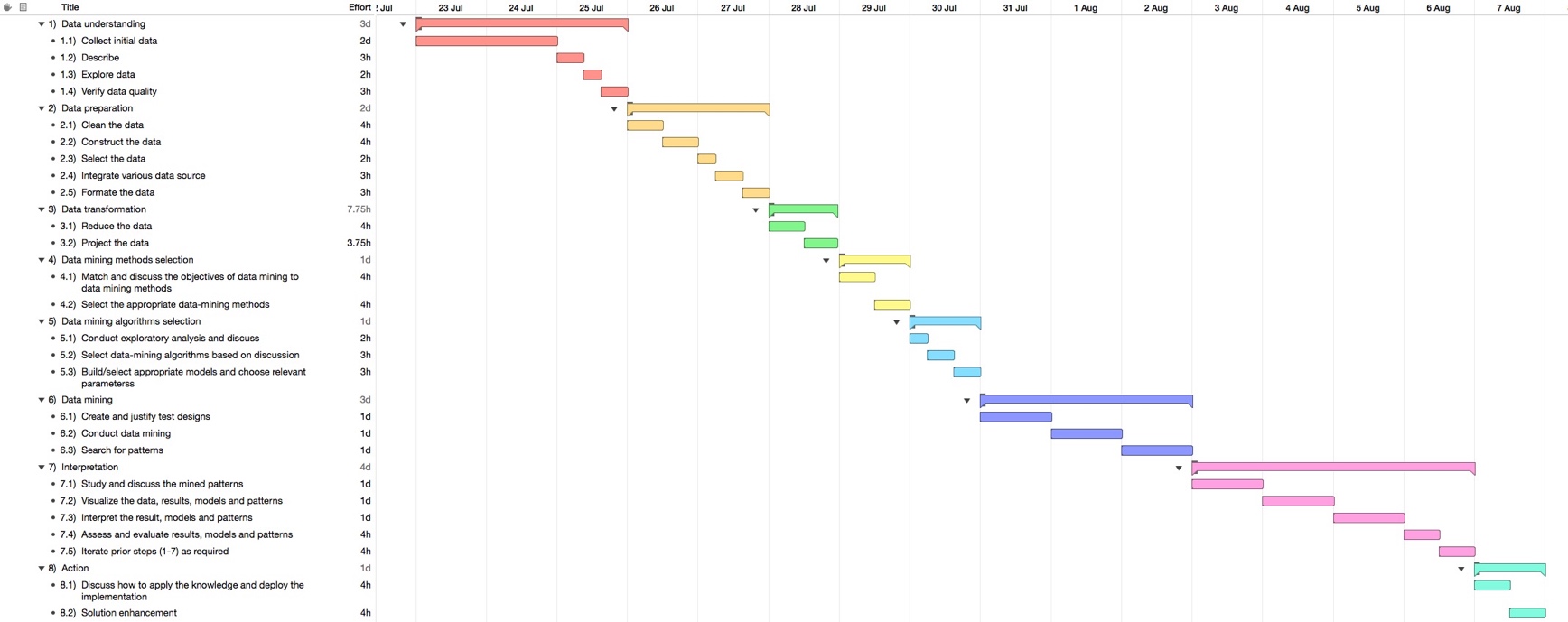
The goal of this report is to find the relationship between meteorological conditions and air quality by both prediction and classification.

An air quality index (AQI) is a number indicating how polluted the air currently is or how polluted it is forecast to become. Many meteorological factors from the data, like temperature and humidity are also numbers. In order to model continuous-valued functions, prediction is applied here for the regression analysis. If there exists regression relationship, we may use some of these factors to predict the air quality.

On the other hand, meteorological data may also contain binary values, like whether it rains or snows during a specific day. For these factors a classifier would be constructed to predict categorical labels such as “snow” or “no snow” for a meteorological data. Also, in order to realize the classification, AQI may need to be sorted in to different air pollution categories.

### 1.4 Project plan

This project of analyzing meteorological effect on air quality in Beijing may take around 2 weeks, followed by the visualized data mining roadmap:



# 2. Data understanding

2.1 Collect initial data

Both air pollutants and meteorological factors are essential for the study of their effect on air quality. However, we may find great difficulties in the process of collecting data of air pollutant like factory and automobile exhaust, because the numerical detection of these pollutants is challenging, and although it is included in the relevant reports in Beijing, most of the data is in the unit of years, which means we may not find sufficient data for the data mining process.

On the other hand, the collection of meteorological information is more accessible. Beijing Meteorological Service regularly tests and records meteorology information in Beijing, among them, some data are even recorded by hours or even minutes. The majority of the data in the meteorology daily records is actually the average value of multiple factors for each day, which may reduce the impact from potential inaccurate values.

Therefore, the main purpose of this paper is to analyze whether meteorological factors have effect on air quality and their relationships by collecting the daily air quality and meteorological records of Beijing from the year 2015 to 2017.

Source data of meteorological information is collected from National Meteorological Information Center[[3]](#footnote-4). Air quality data of Beijing is collected from Tianqihoubao[[4]](#footnote-5) which presents and records the detailed data about air quality indicators measured in Beijing published by the China National Environmental Monitoring Centre[[5]](#footnote-6).

Because the data source is gathered from different sources, their format needs to be unified and then combined together for the following data analysis. The raw dataset is named “raw-data.csv” in the folder.

2.2 Describe the data

In this paper, meteorological information is represented by the following factors:

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

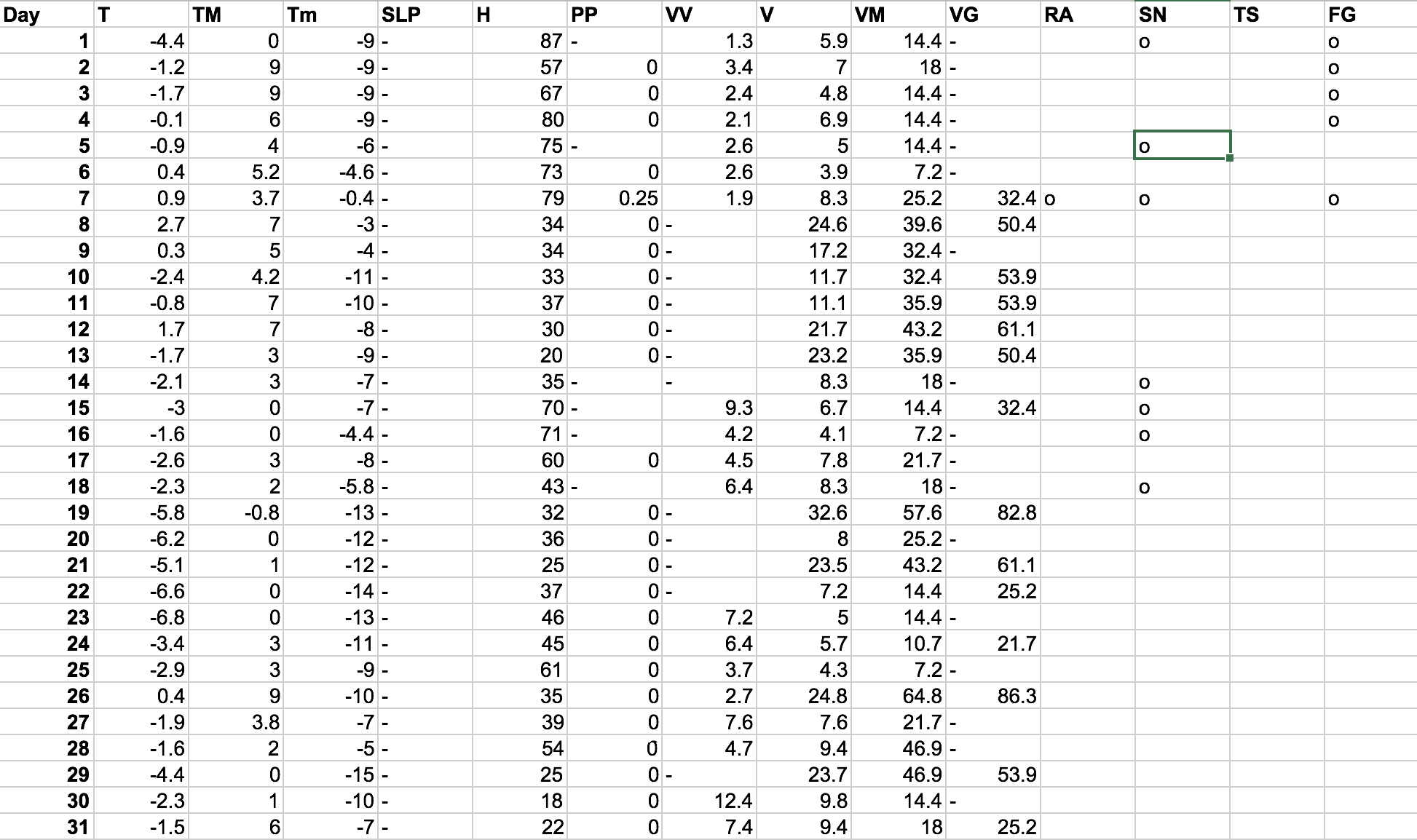
An air quality index (AQI) is a number indicating how polluted the air currently is or how polluted it is forecast to become. As the AQI increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects. Different countries have their own air quality indices, corresponding to different national air quality standards[[6]](#footnote-7).

In China, The AQI level is based on the level of six atmospheric pollutants, namely sulfur dioxide (SO2), nitrogen dioxide (NO2), suspended particulates smaller than 10 μm in aerodynamic diameter (PM10), suspended particulates smaller than 2.5 μm in aerodynamic diameter (PM2.5), carbon monoxide (CO), and ozone (O3) measured at the monitoring stations throughout each city. In this way, the AQI level plays an important role as a comprehensive indicator that reveals potential air pollution problem.

Therefore, the air quality takes factors like PM2.5, PM10, SO2, NO2, CO, O3 and AQI into consideration.

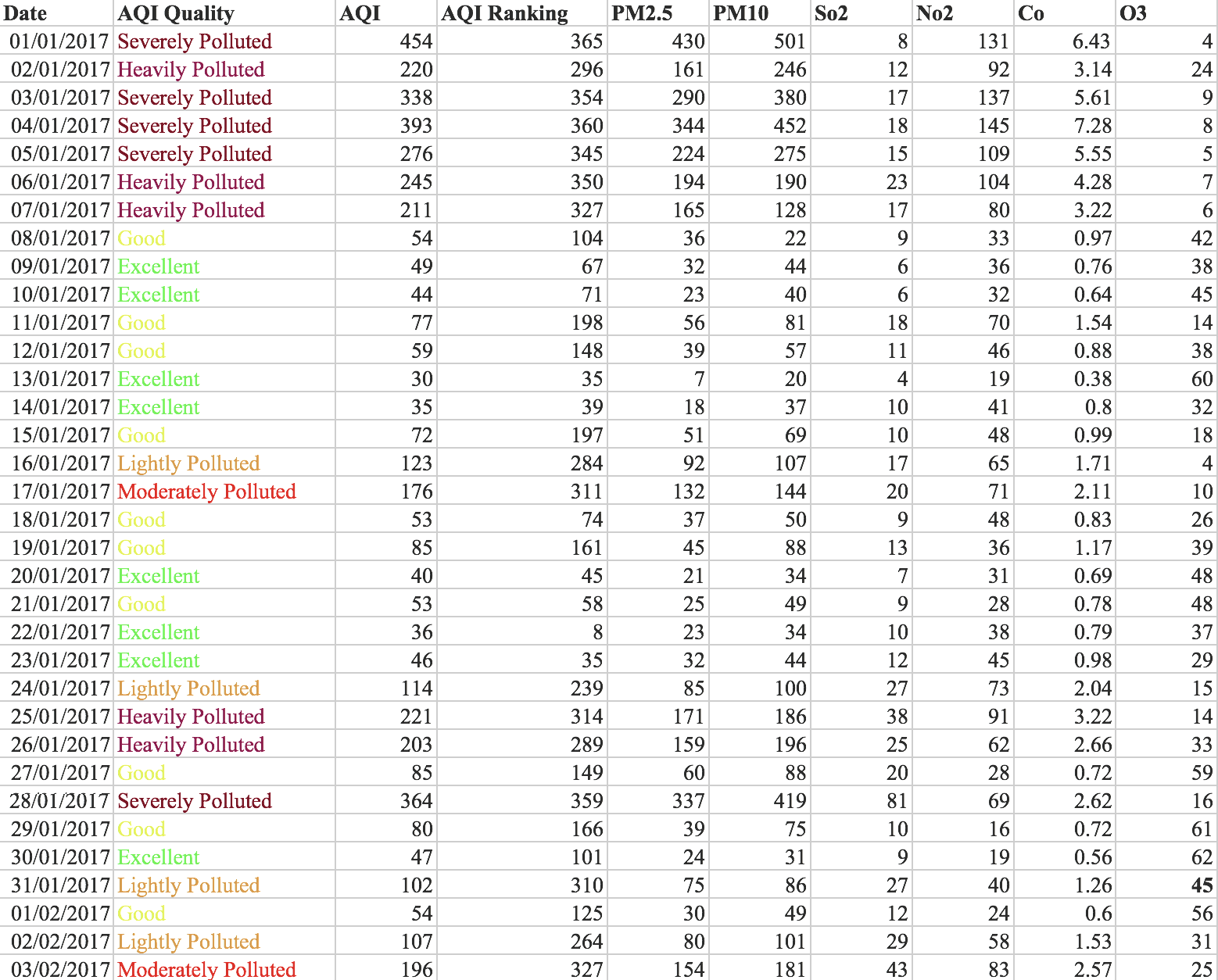
2.3 Explore the data

Original meteorological data is shown below:



We can find that there are some deficiencies in the data, especially in the SLP column. Columns like RA, SN, TS and FG are binary factors, but they are only marked “o” when the meteorological phenomenon happens. These cases should be taken into consideration in further data preparation.

And the original air quality data is as followed:

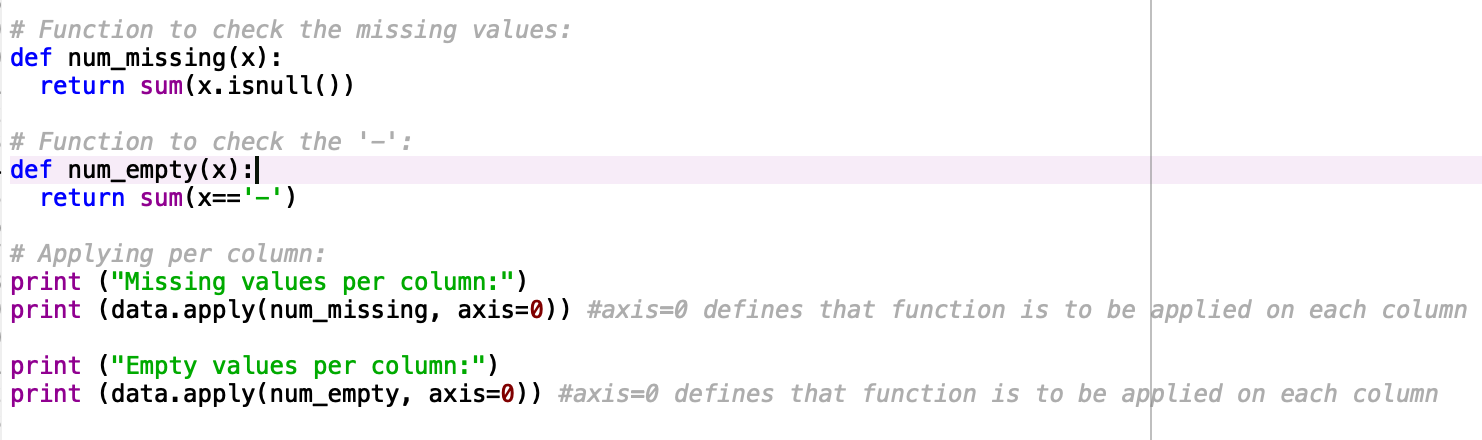


This table is much more complete comparing to the previous one, but there are still variables that is not appropriate to locate in the dataset. The column “AQI Ranking” indicates the ranking of air quality in Beijing comparing to other Chinese cities. As this paper focus on one city, this column is not very useful and may need to be removed in further data preparation.

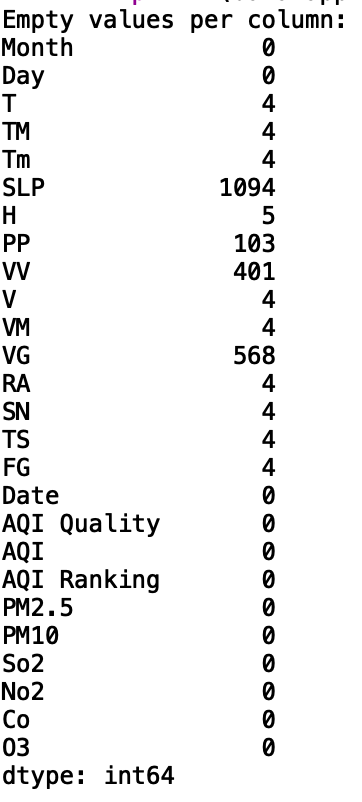
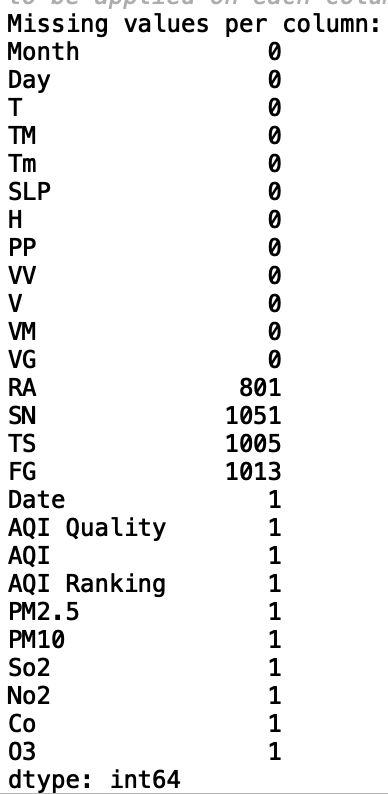
You may also have noticed the AQI quality in this table, which is actually derived from the AQI. It can be used as the ordinal representation of the AQI value and may lead to a better visualization output in the later process. As I mentioned before, this report will analyze the influence of meteorological factors on air quality by both classification and prediction method. Since AQI quality a very good nominal target for the classification, it will be remained in the dataset.

2.4 Verify the data quality

Python is a dynamic tool for knowledge development in databases. By making new functions to check the missing values and empty values (data marked by “-”), we can analyze the quality of the data:



And the outputs of the functions are as followed:

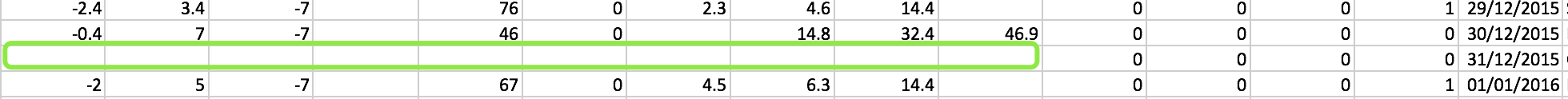


For from the result, we may find some data quality patterns:

As we mentioned before in step 2.3, columns like SLP, PP, VV and VG incomplete, but in the quality check it appears to be fine (the count is 0). That’s because in the raw data, it uses “-” to mark the null value, and it has been proved by the output above on the right hand.

On the other hand, we also find that inside binary columns RA, SN, TS and FG, there are lots of missing values. As we mentioned in step 2.3, that’s probably because the raw data only marks “o” as the value of “yes”, and all the values stands for “no” remain null.

What’s more, it’s noticed that all air quality records have 1 null value, and at the same time, most meteorological records have 4 null values. This means that there are some entire missing rows in the dataset, for example:



From this screenshot, we can see that the data on the left (meteorological data) is missing while the air quality records is still fine. All these values mentioned above need to be pre-processed during the data preparation process. (All these missing values are kept in the dataset for later inspection.)

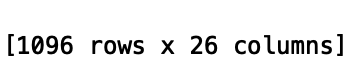
These discoveries indicate that the raw data requires some preparation processing, in order to perform the data mining later on.

3. Data preparation

Our goal of this report is to find the relationship between the meteorological factors and air quality. For the columns like TM, Tm, VM and VG, they stand for the maximum value of a corresponding factor like temperature or wind speed. They are valuable records for meteorological research, however, the average value is more representative and accurate for the study of its influence on air quality for this paper. In this case, these values may get removed for a clear result.

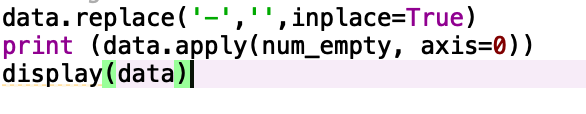
In step 2.4, we also find some problems during the inspection of the data quality, in order to improve it, we need to do some modifications. I changed all the “-” marks in the csv file into null, and for the binary columns, I changed all the “o” marks into “1” and all the empty records into “0”.

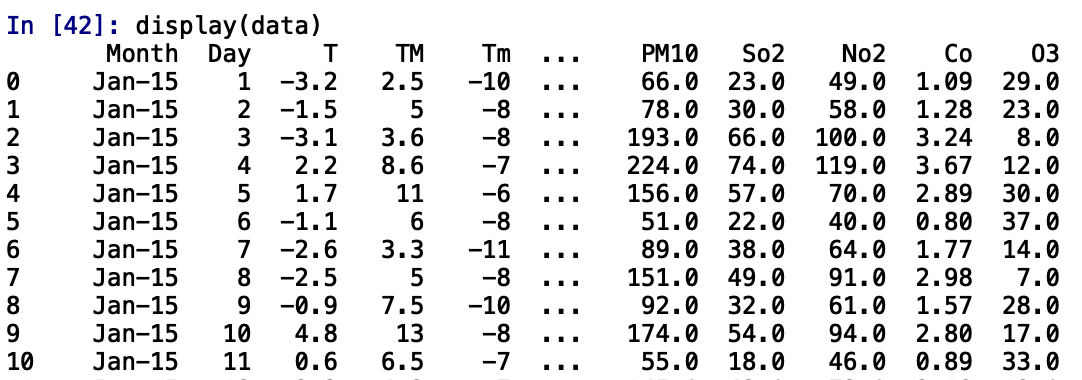
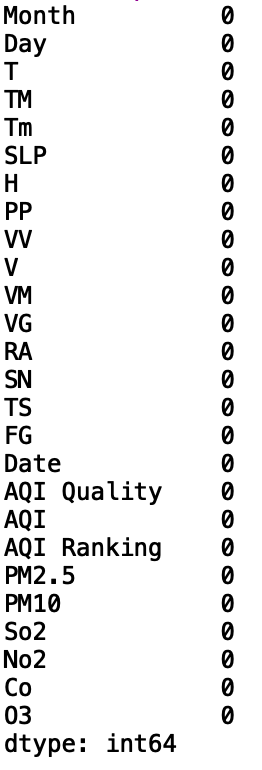
Before we prepare the data, we will record the shape of the data, for further compare:



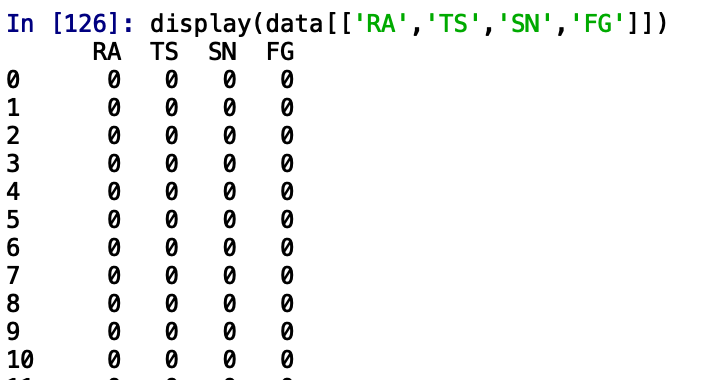
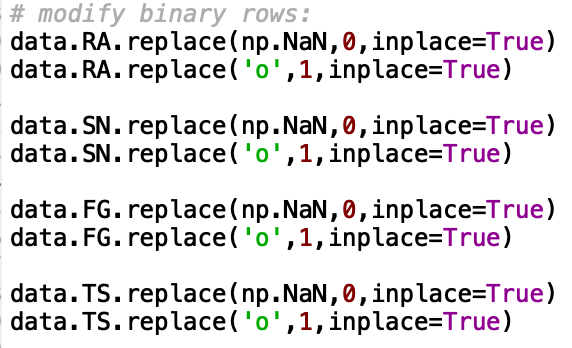
And as the output indicates, now the data is consists of 1096 rows and 26 columns.

Firstly, we change all the data marked with “-” in to null:





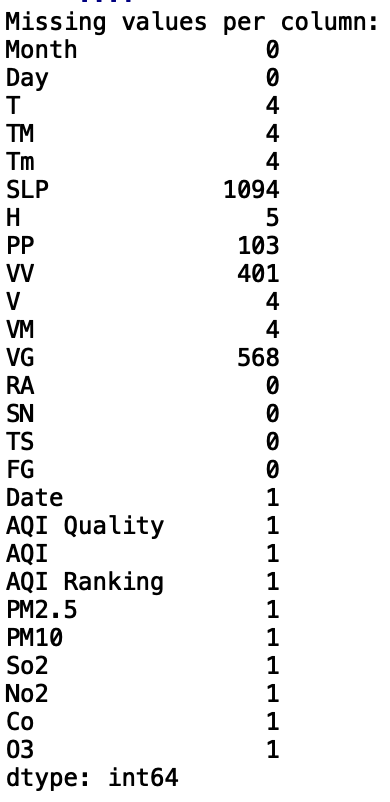
As you can see above, each “-” is now gone, and all the negative numbers remain still, as we hope. Then, we will modify the values for the binary columns:



As you can see, now the binary columns are shown correctly.

Till this step, we finished the first step of the data preparation, the changed dataset is saved as “raw-data-modified.csv” in the folder.

And then we can do the quality inspection again:



Now we can see that all the previous problems are solved. From the output of the data check, it’s noticed that fort the column of SLP, a huge number of data are missing. There are also lots of missing values in the column of VG, VV and PP. In this report, we consider 1000 as the baseline number of the records for analysis, because if there’s no enough data, the result of the report may be misleading. In this case, these columns may need to be removed later in the process due to lack of data quality.

Also, outliers and extreme values are also found in this stage, as shown above. Image on the right is the boxplot of the AQI column. Based on the result we can find that, AQI data with values more than 250 is considered as outliers. These are the AQI quality classified as “heavily polluted” and “severely polluted”. Either direct remove or replace by mean may greatly affect our classification process later on. In this case, the data remains still.

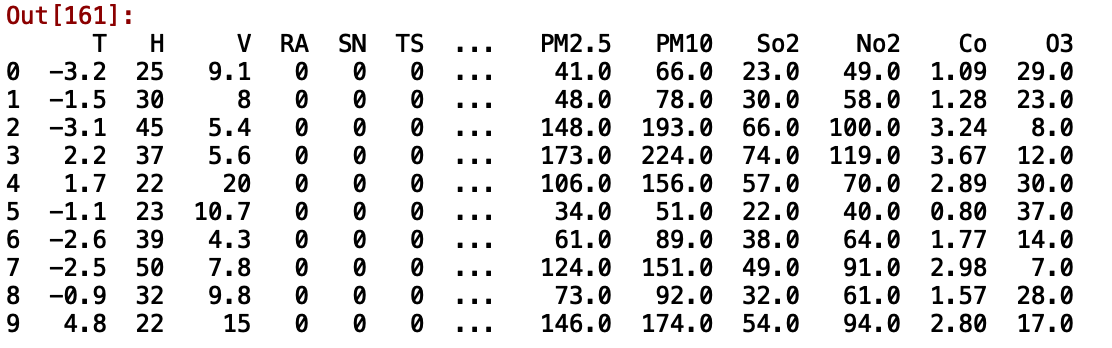
In order to generate an integrated dataset, the meteorological information and air quality indicators of the same date need to be combined together. Some modifications need to be made before the data mining, for example, the date is recorded in both data source, and we may keep only 1 of them as the id of the data.

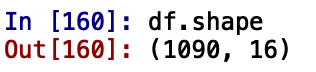
4. Data transformation

4.1 Reduce the data

As it was mentioned in the step 2.3 and 2.4, column stands for maximum values (TM, Tm, VM and VG) are removed, repetitive date columns (Month, Day), as well as data with too much null values (SLP, VG, VV, PP) or with useless value (AQI ranking) for the goal. Then, we can remove all the rows with null values.

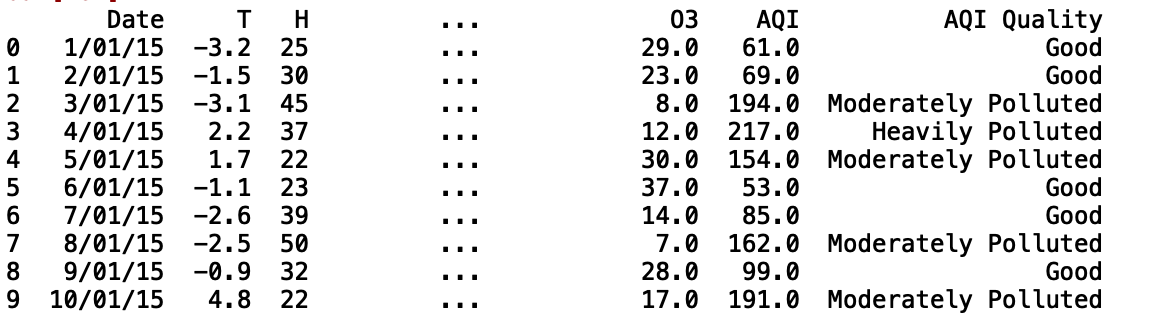
After the reduction of the data, the result of the dataset is shown below:





As you can see, now there are only 16 columns and 1090 lines of data (compares to 26 columns and 1096 lines in the original raw dataset).

The order of the data (like position of the Date and AQI) could be changed for a better understanding:



And it can also be visualized to have a preliminary understanding:



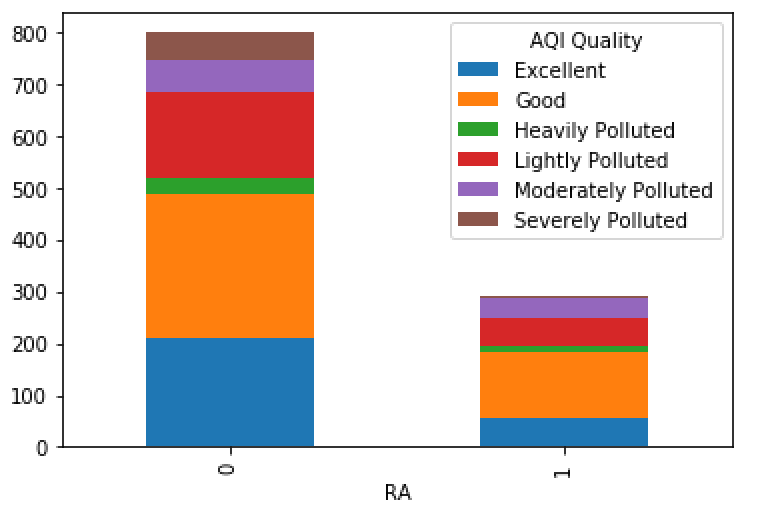
The result of this processing can also be found in the source folder, named “final-data”.

4.2 Project the data

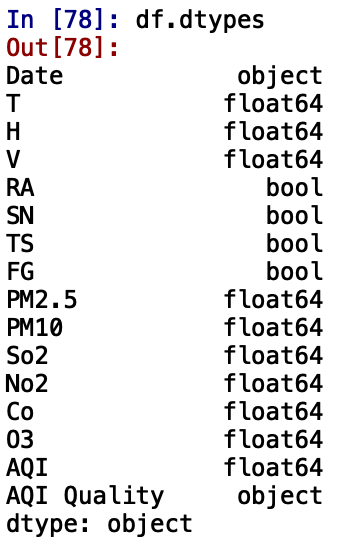
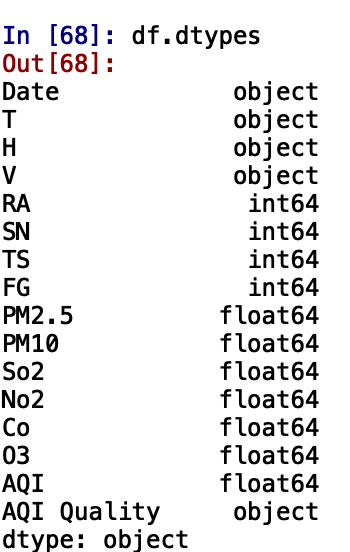
Firstly, we try to have a general overview of the data through feature selection function. However, due to the AQI is a continuous value, the distribution of the AQI is hard to get visualized. However, we can separate AQI into different groups according to the classification of the AQI as shown below:

|  |  |  |
| --- | --- | --- |
| **AQI** | **Air Pollution Level** | **Air Pollution** **Category** |
| 0–50 | Level 1 | Excellent |
| 51–100 | Level 2 | Good |
| 101–150 | Level 3 | Lightly Polluted |
| 151–200 | Level 4 | Moderately Polluted |
| 201–300 | Level 5 | Heavily Polluted |
| >300 | Level 6 | Severely Polluted |

In this way, the distribution of AQI in various level can be directly visualized by each variable, after it is divided into 6 groups, the graph appears to be more representative:



And we can modify the type of each factors using “dtypes” function:



After the modification (right), we can find that the types of the factors are now all correct, and then we can move on to select the data mining methods.

5. Data-mining method(s) selection

5.1 Match and discuss the objectives of data mining to data mining methods

The data mining methods is normally divided into two kinds, supervised learning and unsupervised learning. For this report, because we already know the output (AQI) and input (meteorological factors), supervised learning method should be the appropriate choice to make further study.

Normally, there are two kinds of supervised learning: classification and prediction. Regression and classification are similar in many ways, the major difference is their output has different range of values. For the classification, the output is normally several values, which is called labels; For the regression, the output can take any real number.

For our project, because we have two kind of outcomes—AQI (continuous value) and AQI Quality (label), we will use both methods.

5.2 Select the appropriate data-mining method(s) based on discussion

The goal of this report is to find the relationship between meteorological conditions and air quality. Many meteorological factors from the data, like temperature and humidity, also the target—AQI are continuous values. In order to model continuous-valued functions, prediction is applied here for the regression analysis. If there exists regression relationship, we may use some of these factors to predict the air quality.

At the same time, the dataset also contains binary values, like whether it rains during a specific day. For these factors a classifier would be constructed for a better prediction of categorical labels such as “snow” or “no snow” for a meteorological data.

Therefore, both prediction and classification are considered the appropriate data mining methods for this report.

6. Data-mining algorithm(s) selection

6.1 Conduct exploratory analysis and discuss

Like we mentioned in the step 5, variables in this paper includes continuous values like temperature and binary values like whether rain comes at a specific day. Targets are also divided into 2 possibilities: continuous AQI value or categorical AQI Quality level. Thus, algorithms for both classification and prediction methods need to be taken into consideration.

6.2 Select data-mining algorithms based on discussion

There are lots of algorithms we can use in the data mining, for our project, regression and classification algorithms will be used.

Binomial classification normally divides the data into two parts: negative and positive. It’s very helpful for binary label dataset. However, as we mentioned before, the outcome of the AQI Quality (multinomial) analysis is not binary, thus algorithms based on binary outcome (like Huffman) may not fit in this situation.

On the other hand, because our model for regression consists of float type data and continuous target, common regression algorithms could be used in our analysis.

Considering the applicability of the models to our data mining goal, the following models are chosen:

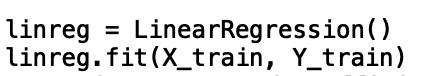
For classification, decision tree classifier, logistic regression and random forest classifier will be utilized.

For prediction, linear regression will be used.

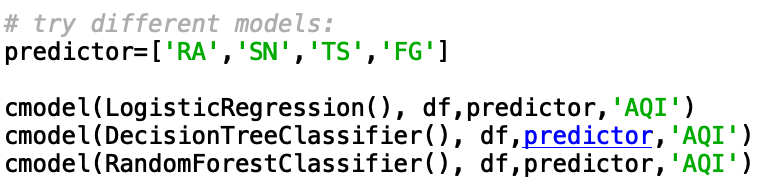
6.3 Build/Select appropriate model(s) and choose relevant parameter(s)

For this report, continuous parameters may have a better result with the continuous target, while binary parameter works better with categorical target. So, we will try to use Linear Regression to study the temperature, humidity, rainfall and wind speed, while applying decision tree, logistic regression and random forest for binary variables like rain, storm, fog and snow.

The regression model is built as followed:



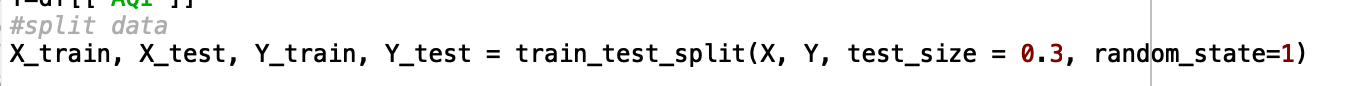
The classification models are built as followed:



7. Data Mining

7.1 Create and justify test designs

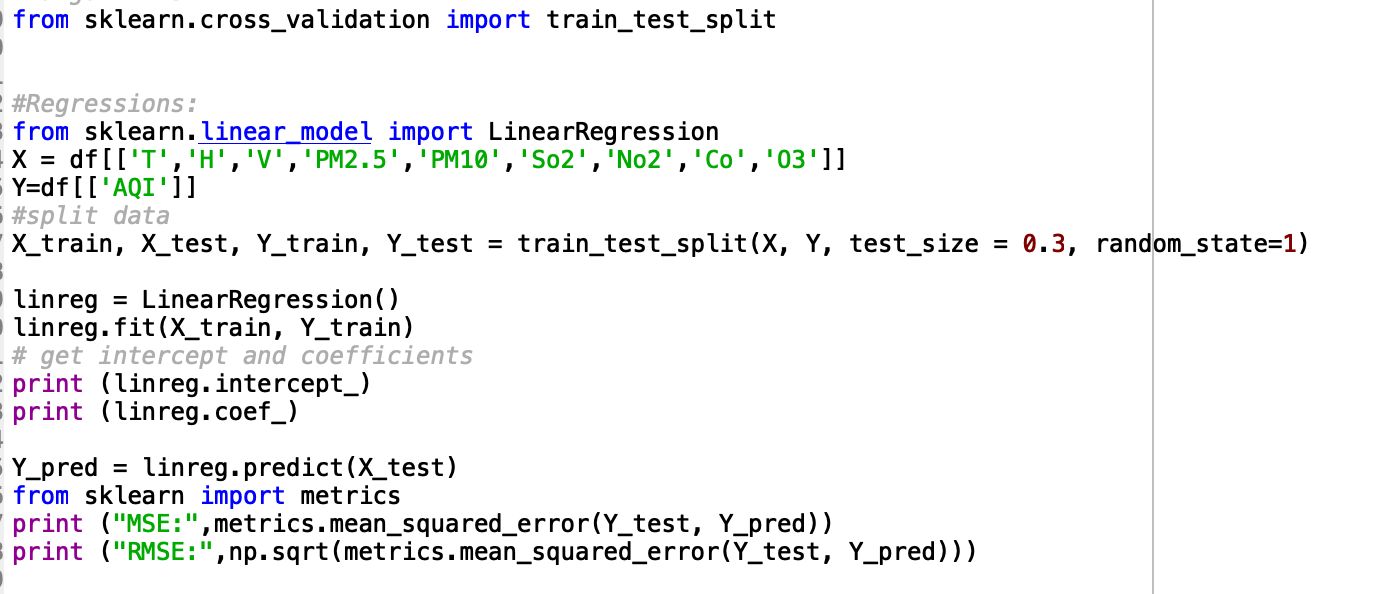
The main purpose is to conduct the data mining as it is planned in step 5-6, and check their accuracy using the analysis node. Dataset is split into two distinct datasets, where training data is used to create the model and testing data is used to test and validate the model’s capability.  Fort this step, a 70/30 training/testing split is used considering the total amount of the data is only 1090. Small testing set may lead to great deviation, for this reason, in order to guarantee the accuracy of the testing result, the size of testing partition is increased to 30%.



7.2 Conduct data mining – classify, regress, cluster, etc. (models must execute)

Data mining processing is as followed:

Linear Regression:



Classifications:

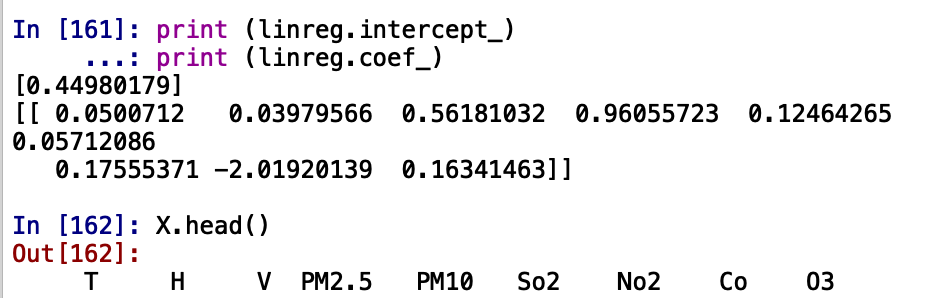


As you can see, prediction method is applied in the data mining with Linear Regression algorithm. Classification method is also applied, using decision tree, logistic regression and random forest algorithms for the analysis of the meteorological influence on the air quality.

7.3 Search for patterns

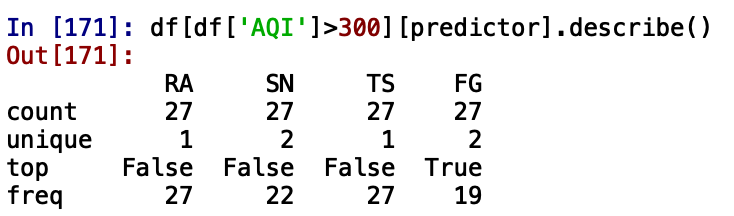
Some patterns are found in the process of the data mining.

Firstly, Co (-2.019), PM2.5 (0.961) and wind speed (0.562) are weighted mostly in the linear model:



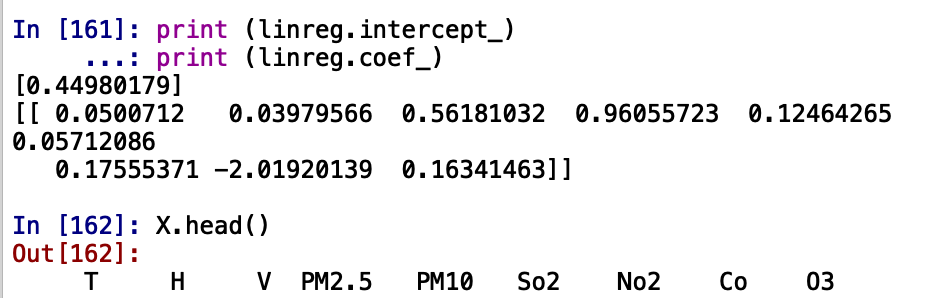
At the same time, the relationship between AQI and some meteorological factor seems to be irrelevant, like temperature (0.050) and humidity (0.040).

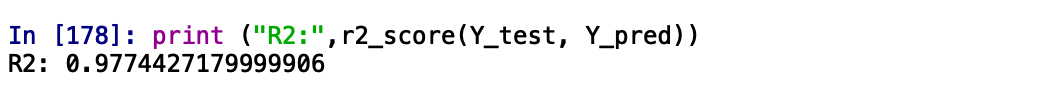
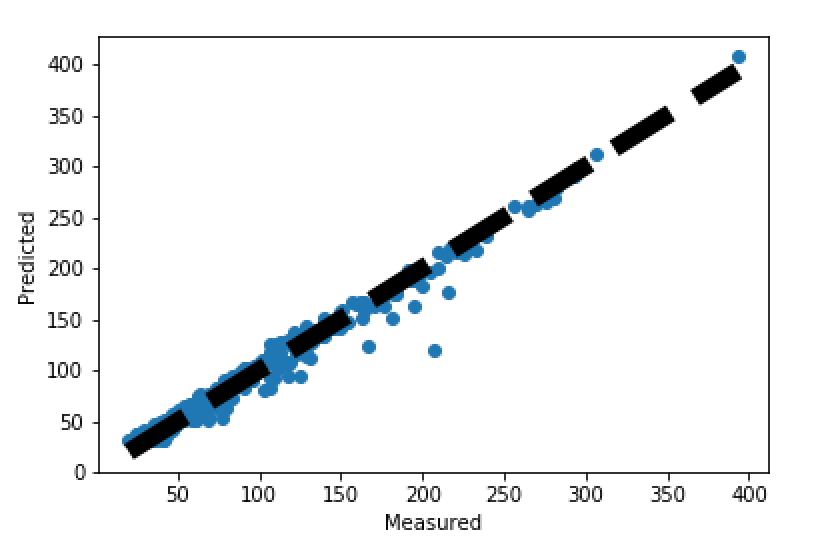
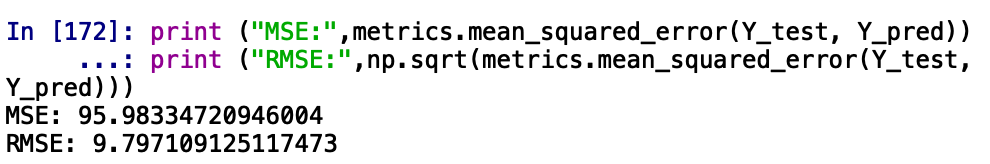
On the other hand, during the process of generating a Decision tree classifier model, we can easily find the total numbers of each meteorological binary factors for each AQI quality. When the air is severely polluted, it seems to have no rain or storm meteorological phenomenon (unique =1):



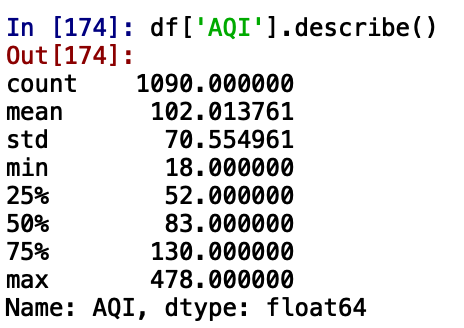
# 8. Interpretation

8.1 Study and discuss the mined patterns

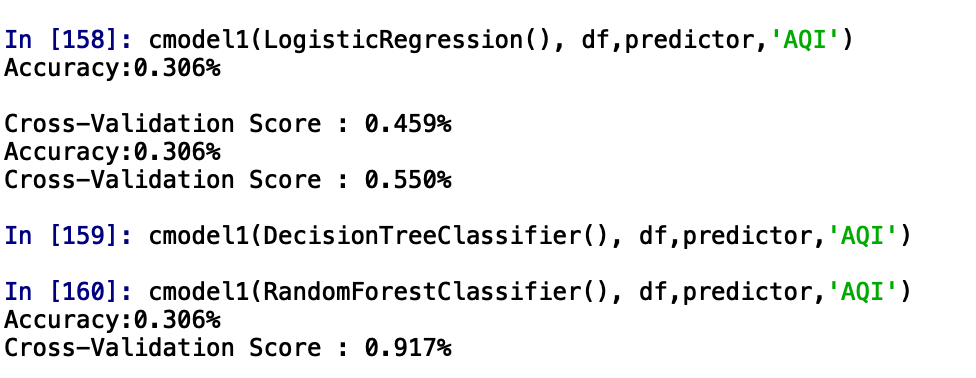




From the data mining result of the prediction part is shown above, we can see that the R2 (0.977), represents the predictive power of the model as a value between 0 and 1) of Linear Regression is very close to 1, this means that there is a linear relationship. Based on the Root Mean Squared Error (RMSE, which measures how close the predictions are to the actual outcomes; thus, a lower score is better), we can know that the prediction outcome (9.797) is also very close to the actual value, comparing with the actual "AQI" value, such as mean, min and max:



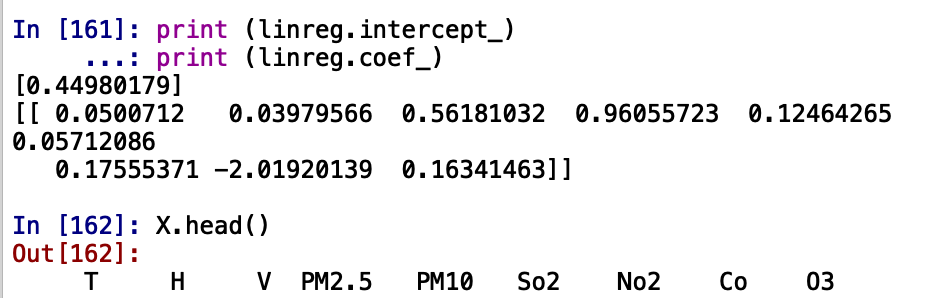
On the other hand, the results of classification output generated from binary values are not satisfactory (accuracy is only 0.306%):



This means that meteorological factors like rain, storm, fog and snow don’t have a direct influence on the AQI quality.

8.2 Visualize the data, results, models, and patterns

Firstly, we look into the linear models:



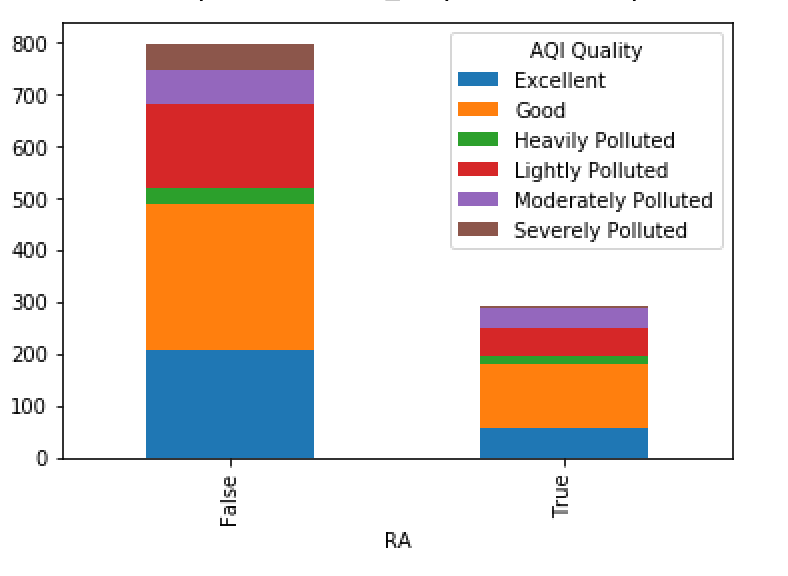
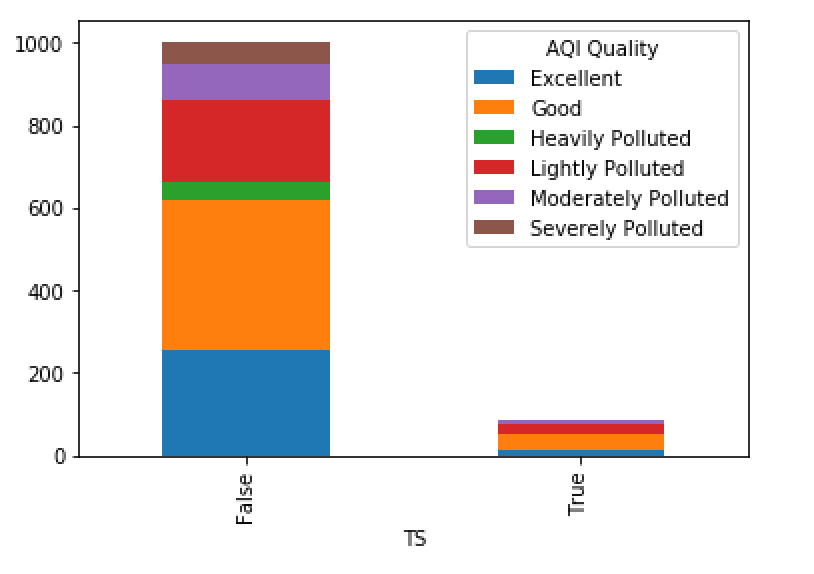
From this table we can know that CO (feature weight =-2.019) and pm2.5 (feature weight = 0.961) are the most important predictors for AQI, while the predictor importance of other pollutants is relatively low: NO2 (0.176), O3 (0.163), PM10 (0.125) and SO2 (0.057). Compares to these values, the predictor importance for meteorological factors are lower: V (0.562), H (0.040) and T (0.050). In order to have a better understanding of the meteorological impact on the air quality, we may need to do another iteration (step 8.5) without all the pollutants factors.

This can also be proved by the graphs below:

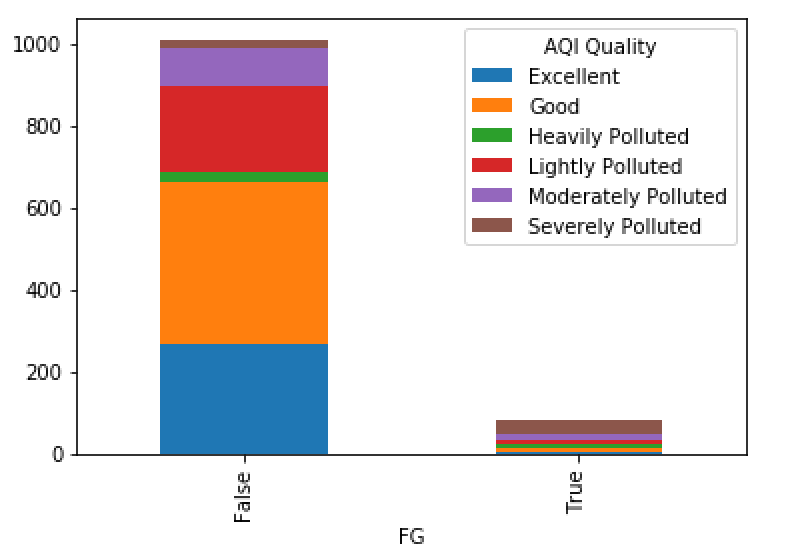




Then, we will look in to the classification. As we mentioned before, the error rate of the Decision Tree classifier is too high to find reliable patterns. However, as we mentioned in 7.3, when the air is severely polluted, it seems to have no rain or storm meteorological phenomenon.

The graphs above shows that there is no severe polluted air quality when rain or storm come, so they may have a positive effect on the air. On the contrary, more severe polluted phenomenon take place when fog comes:



Comparing these two groups, we can conclude that the portion of excellent, good, light and moderate AQI of RA and TS is higher than SN and FG. In general, rain (RA) and TS (storm) have a good influence on the AQI quality, the impact of snow (SN) remain uncertain, and Fog (FG) seems to have a bad effect on the air quality.

8.3 Interpret the results, models, and patterns

The interpretation of the patterns and results found during the data mining is as followed:

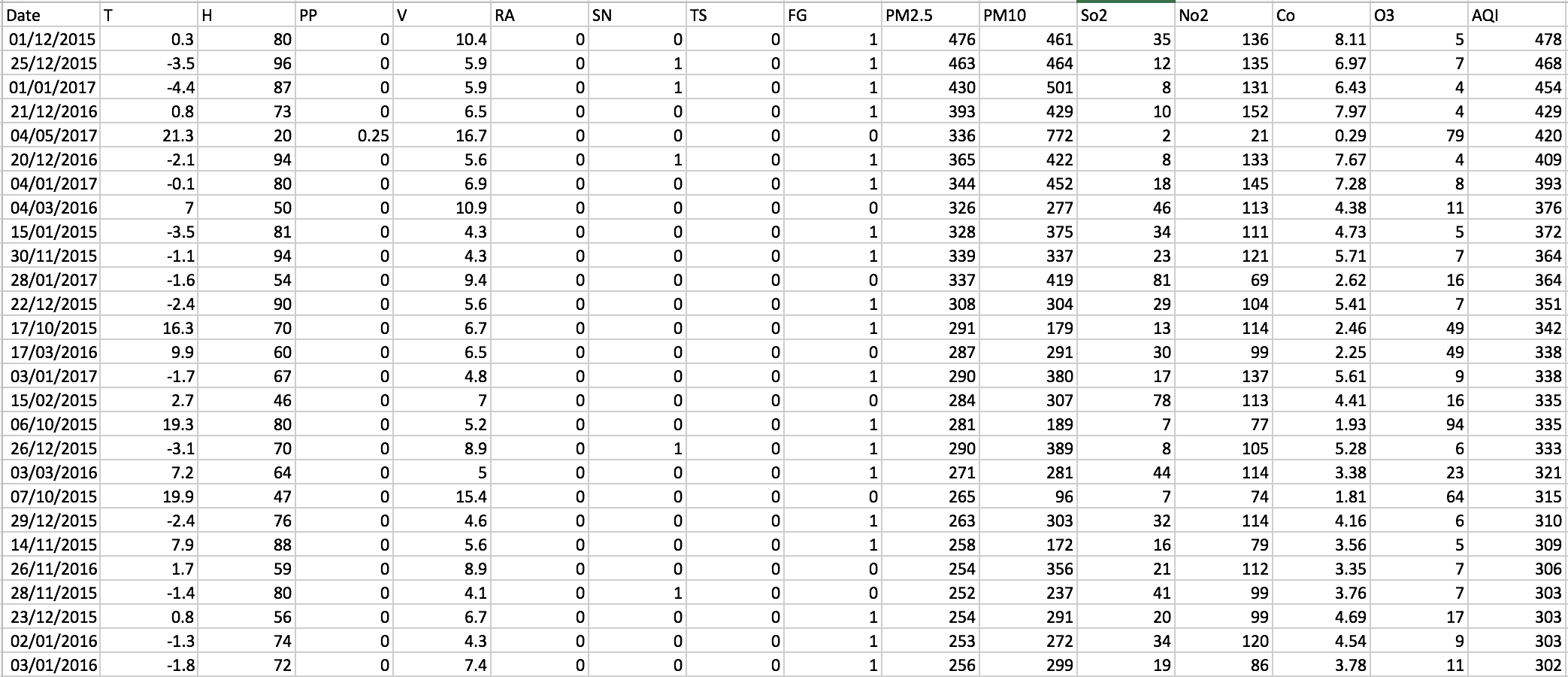
1. By using both prediction and classification methods, we find out that meteorological phenomenon are not decisive factors for the air quality.
2. Specific meteorological phenomenon may affect some circumstances (When storm comes, the air won’t be severely polluted, and when fog come, the air quality won’t be excellent and is very likely to be severely polluted). This is probably caused by the meteorological condition for the accumulation and spread of pollutant.
3. Comparing to other pollutants, PM2.5 and CO has the greatest impact on the AQI of Beijing.

8.4 Assess and evaluate results, models, and patterns

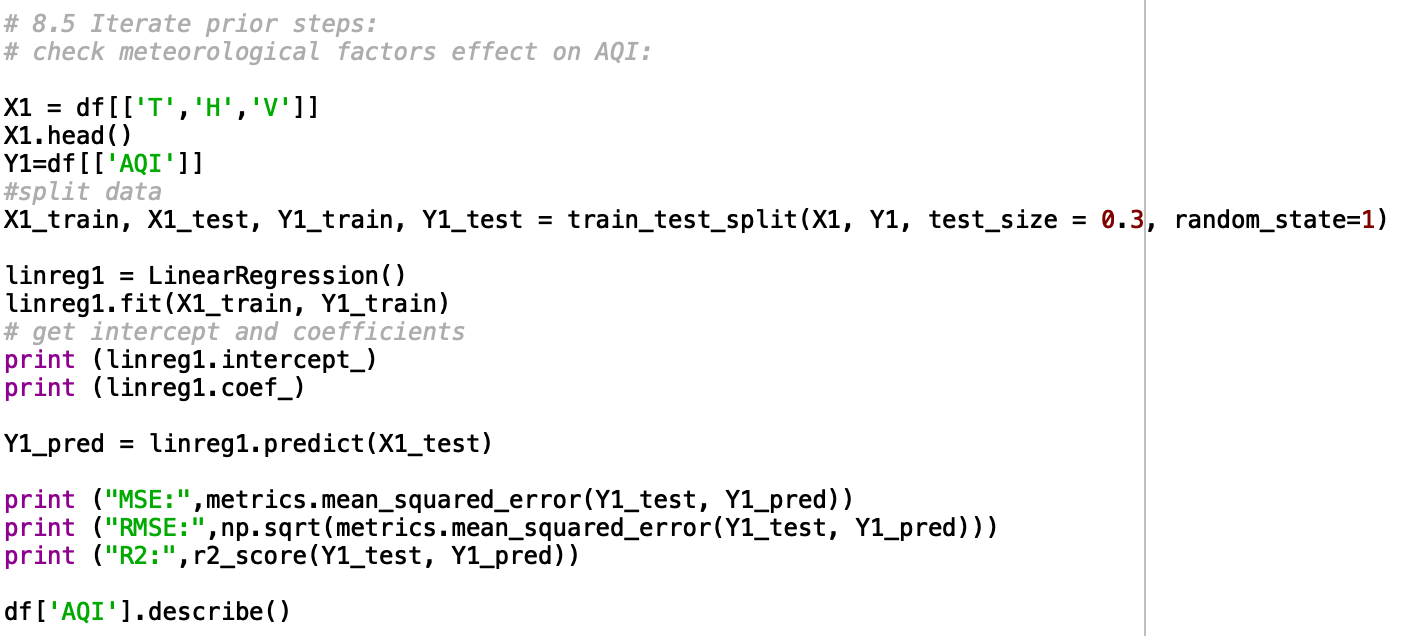
The effect of the meteorological phenomenon is less than I expected. This is due to two reasons: First is that it’s not possible to control all the variables. The meteorological indicators change every day, and the pollutes generated by different sources are hard to be balanced at one specific level at a long time. Secondly, meteorology is always the secondary factor that influence air quality. It never creates or eliminate pollutants in the air, but it does effectively promote the diffusion of the polluted air when wind comes.

8.5 Iterate prior steps (1 – 7) as required

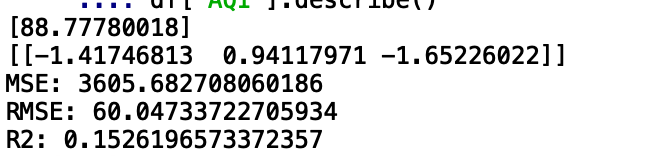
In order to prove the point mentioned in 8.2, I sorted the data with AQI greater than 300 (severe polluted). As you noticed, there is no rain or storm during these days (mostly in winter), which in this way is indeed a factor that is very important for the dispersion of pollutants.



Also, in order to check the linear relationship between meteorological factors and AQI, I make another iteration to check their relation using the same method and algorithm:



You can find these codes right behind the linear regression, with the comment says (Linear model with only meteorological factors). This time, only the meteorological factors are left for the generation of the model, and the output is as followed:



This time, the model accuracy is very low: RMSE reaches 65.766 (huge number comparing to mean and standard deviation), and the R2 is only 0.153. Comparing to the previous results of the linear models, we conclude that there’s no linear relationship between these selected meteorological factors and the AQI. This result also proves our conclusion in the step 8.3.

1. Effect of climate change on air quality, Danile J. Jacob, Darrell A. Winner [↑](#footnote-ref-2)
2. Source: European Environment Agency, 2013 [↑](#footnote-ref-3)
3. National Meteorological Information Center, [data.cma.cn](http://data.cma.cn) [↑](#footnote-ref-4)
4. Tianqihoubao, [www.tianqihoubao.com](http://www.tianqihoubao.com) [↑](#footnote-ref-5)
5. China National Environmental Monitoring Centre, www.cnemc.cn [↑](#footnote-ref-6)
6. Air quality index, Wikipedia, https://en.wikipedia.org/wiki/Air\_quality\_index#cite\_note-2 [↑](#footnote-ref-7)