**Title of report**

**BDAS (Steps 1-8)**

**Investigation of Beijing air pollution (PM25) using Machine Learning methodology**

**Commissioned by and prepared for:**

[**Data Mining and Big Data**](https://canvas.auckland.ac.nz/courses/46903) **course**

**Professor:**

**David Sundaram**

**Student name:**

**Mehdi Tadayoni**

**Submission Date:**

**October 2020**

Contents

[1 Business and/or Situation understanding 8](#_Toc54234895)

[1.1 Identify the objectives of the business/situation 8](#_Toc54234896)

[ Business Success Criteria 13](#_Toc54234897)

[1.2 Assessing the Situation 14](#_Toc54234898)

[ Resource Inventory 14](#_Toc54234899)

[ Requirements, Assumptions, and Constraints 14](#_Toc54234900)

[ Risks and Contingencies 15](#_Toc54234901)

[ Terminology 15](#_Toc54234902)

[ Cost/Benefit Analysis 15](#_Toc54234903)

[1.3 Determine data-mining objectives 15](#_Toc54234904)

[ Data Mining Success Criteria 16](#_Toc54234905)

[1.4 Project Plan 17](#_Toc54234906)

[2 Data Understanding 17](#_Toc54234907)

[2.1 Collecting Initial Data 17](#_Toc54234908)

[ Additional data 19](#_Toc54234909)

[2.2 Describing Data 20](#_Toc54234910)

[2.3 Exploring Data 22](#_Toc54234911)

[2.4 Verifying Data Quality 30](#_Toc54234912)

[3 Data preparation 31](#_Toc54234913)

[3.1 Select the data 32](#_Toc54234914)

[3.2 Clean the data 36](#_Toc54234915)

[3.3 Construct the data 40](#_Toc54234916)

[3.4 Integrate various data sources 41](#_Toc54234917)

[3.5 Format the data as required 44](#_Toc54234918)

[4 Data transformation 45](#_Toc54234919)

[4.1 Reduce the data 45](#_Toc54234920)

[4.2 Project the data 54](#_Toc54234921)

[5 Data-mining method(s) selection 54](#_Toc54234922)

[5.1 Match and discuss the objectives of data mining (1.1) to data mining methods 54](#_Toc54234923)

[5.2 Select the appropriate data-mining method(s) based on discussion 57](#_Toc54234924)

[6 Data-mining algorithm(s) selection 57](#_Toc54234925)

[6.1 Conduct exploratory analysis and discuss 57](#_Toc54234926)

[6.2 Select data-mining algorithms based on discussion 60](#_Toc54234927)

[6.3 Build/Select appropriate model(s) and choose relevant parameter(s) 62](#_Toc54234928)

[ Generalized linear regression 64](#_Toc54234929)

[ Multilayer perceptron classifier 64](#_Toc54234930)

[7 Data Mining 65](#_Toc54234931)

[7.1 Create and justify test designs 65](#_Toc54234932)

[7.2 Conduct data mining 67](#_Toc54234933)

[ Generalized linear regression 71](#_Toc54234934)

[7.3 Search for patterns 71](#_Toc54234935)

[8 Interpretation 76](#_Toc54234936)

[8.1 Study and discuss the mined patterns 76](#_Toc54234937)

[8.2 Visualize the data, results, models, and patterns 78](#_Toc54234938)

[8.3 Interpret the results, models, and patterns 80](#_Toc54234939)

[8.4 Assess and evaluate results, models, and patterns 81](#_Toc54234940)

[8.5 Iterate prior steps (1 – 7) as required 82](#_Toc54234941)

[9 RefErence: 83](#_Toc54234942)

Figures

[Fig. 1: total death per 100000 due to air pollution among countries from 1990 to 2017 in 50 countries that mortality 9](#_Toc54234943)

[Fig. 2: death per 100000 population due to air pollution among some countries from 1990 to 2017 in three portions 10](#_Toc54234944)

[Fig. 3: summation of death per 100000 due to air pollution among countries from 2000 to 2017 in three portions 10](#_Toc54234945)

[Fig. 4: distribution of air pollution total death from 1990 to 2017 in china 11](#_Toc54234946)

[*Fig. 5: Distribution of National Air Quality Monitoring Stations in Beijing(Wei Chen,2015)* 19](#_Toc54234947)

[Fig. 6 Relationship between PM25 and temperature data from 2013 to 2017 in all station 23](#_Toc54234948)

[Fig. 7 Relationship between PM25 and month data from 2013 to 2017 23](#_Toc54234949)

[Fig. 8 Relationship between PM25 and NO2 data from 2013 to 2017 in all station 24](#_Toc54234950)

[Fig. 9 Relationship between PM25 and wind direction data from 2013 to 2017 in all station 25](#_Toc54234951)

[Fig. 10: Relationship between PM25 and CO data from 2013 to 2017 in all station 25](#_Toc54234952)

[Fig. 11: Relationship between PM25 and O3 data from 2013 to 2017 in all station 26](#_Toc54234953)

[Fig. 12: Relationship between PM25 and PM10 data from 2013 to 2017 in all station 26](#_Toc54234954)

[Fig. 13: Relationship between PM25 and PRES data from 2013 to 2017 in all station 27](#_Toc54234955)

[Fig. 14: Relationship between PM25 and WSPM data from 2013 to 2017 in all station 27](#_Toc54234956)

[Fig. 15: distribution of PM25 vs PM10 28](#_Toc54234957)

[Fig. 16: distribution of O3 vs CO 29](#_Toc54234958)

[Fig. 17: distribution of SO2 vs NO2 29](#_Toc54234959)

[Fig. 18: Data import in all stations 31](#_Toc54234960)

[Fig. 19: Data description in all station , includes: Min, Max, Mean, std,…. 31](#_Toc54234961)

[Fig. 20: Data quality in stations after removing the missing data in all rows 31](#_Toc54234962)

[Fig. 21 appended data vs PM25 in all station 33](#_Toc54234963)

[Fig. 22: appended data vs NO2 in all station 34](#_Toc54234964)

[Fig. 23 appended data vs SO2 in all station 35](#_Toc54234965)

[Fig. 24: the first model for PM25 prediction and relevant inputs ( air pollutant data). Target data is in red region 35](#_Toc54234966)

[Fig. 25: the second model for PM25 prediction and relevant inputs (weather data). Target data is in red region 36](#_Toc54234967)

[*Fig. 26: Outliers in the scatter plot ;PM10 and PM25* 37](#_Toc54234968)

[*Fig. 27:removing the Outliers in the scatter plot ;PM10 and PM25* 37](#_Toc54234969)

[*Fig. 28: Outliers in the scatter plot ;CO and PM25* 38](#_Toc54234970)

[Fig. 29: distribution of CO 38](#_Toc54234971)

[*Fig. 30:removing the Outliers in the scatter plot ;CO and PM25* 39](#_Toc54234972)

[*Fig. 31: Outliers in the scatter plot ;NO2 and PM25* 39](#_Toc54234973)

[Fig. 32: distribution of PM25 40](#_Toc54234974)

[Fig. 33 Bar chart of the average of NO2 in all station 42](#_Toc54234975)

[Fig. 34 Bar chart of the average of PM25 in all station 43](#_Toc54234976)

[Fig. 35 Bar chart of the average of SO2 in all station 44](#_Toc54234977)

[Fig. 36 Feature selection to choose the best input; PM25 , NO2 and SO2 46](#_Toc54234978)

[Fig. 37 Feature selection to choose the best input; O3, PRES and TEMP 46](#_Toc54234979)

[Fig. 38Correlation between PM25 and PM10 47](#_Toc54234980)

[Fig. 39 Correlation between PM25 and CO 47](#_Toc54234981)

[Fig. 40: Correlation between PM25 and SO2 48](#_Toc54234982)

[Fig. 41 Correlation between PM25 and NO2 48](#_Toc54234983)

[Fig. 42 Correlation between PM25 and O3 49](#_Toc54234984)

[Fig. 43 Correlation between PM25 and pressure data 49](#_Toc54234985)

[Fig. 44 PM10 and PM25 distribution in 12 station 50](#_Toc54234986)

[Fig. 45 CO and PM25 distribution in 12 station 50](#_Toc54234987)

[Fig. 46 SO2 and PM25 distribution in 12 station 51](#_Toc54234988)

[Fig. 47 SO2 and PM25 distribution in 12 station 51](#_Toc54234989)

[Fig. 48 Aggregation of data average in selected station 52](#_Toc54234990)

[Fig. 49 Boxplot of PM25 distribution in 9 station 52](#_Toc54234991)

[Fig. 50 Boxplot of NO2 distribution in 9 station 53](#_Toc54234992)

[Fig. 51 linear representation of the data 58](#_Toc54234993)

[Fig. 52 SVM approaches (James, G. , 2015) 59](#_Toc54234994)

[Fig. 53 Correlation between PM25 and PM10 as a another pollutants 60](#_Toc54234995)

[Fig. 54 Correlation between PM25 and NO2 as a another pollutants 61](#_Toc54234996)

[*Fig. 55 Correlation between PM25 and TEMP as a climate data* 61](#_Toc54234997)

[*Fig. 56 Correlation between PM25 and DEWP as climate data* 62](#_Toc54234998)

[Fig. 57Residual plot between PM25 and PM10 63](#_Toc54234999)

[*Fig. 58:correlation between CO and PM25* 63](#_Toc54235000)

[*Fig. 59: correlation between NO2 and PM25* 64](#_Toc54235001)

[Fig. 60: Train and test data; blue one is model-1 as input data, red one is model-2 and green one is target 66](#_Toc54235002)

[Fig. 61 : Description of train and test set in first model 66](#_Toc54235003)

[Fig. 62 : Description of train and test set in first model 67](#_Toc54235004)

[Fig. 66 : Transformed data in first model 69](#_Toc54235005)

[Fig. 67 : Transformed data in second model 69](#_Toc54235006)

[Fig. 68 : Description of train and test set in first model 70](#_Toc54235007)

[Fig. 69 : Description of train and test set in first model 70](#_Toc54235008)

[Fig. 70: data pattern 1 72](#_Toc54235009)

[Fig. 71: Data pattern 2 73](#_Toc54235010)

[Fig. 72: data pattern 3 73](#_Toc54235011)

[Fig. 73: data pattern 4 75](#_Toc54235012)

[Fig. 74: data pattern 5 76](#_Toc54235013)

[Fig. 75 Multi Linear regression algorithms which have the best correlation with PM25 in model 1 and pattern-1 76](#_Toc54235014)

[Fig. 76 Three best algorithms which have the best correlation with PM25 in model 1 and pattern 2 77](#_Toc54235015)

[Fig. 77 Multi Linear regression algorithms which have the best correlation with PM25 in pattern-3 77](#_Toc54235016)

[Fig. 78 GLR algorithms which have the best correlation with PM25 in pattern-4 77](#_Toc54235017)

[Fig. 79 GLR algorithms which have the best correlation with PM25 in pattern-5 77](#_Toc54235018)

[Fig. 80 Correlation between test data and prediction data by multi-regression model in pattern 1 78](#_Toc54235019)

[Fig. 81 Correlation between test data and prediction data by multi-regression model in pattern 2 79](#_Toc54235020)

[Fig. 82 Correlation between test data and prediction data by simple linear regression model in pattern 3 79](#_Toc54235021)

[Fig. 83 Correlation between test data and prediction data by GLR method in pattern 4 80](#_Toc54235022)

[Fig. 84 Correlation between test data and prediction data by GLR method in pattern 5 80](#_Toc54235023)

Tables

[Tab. 1: a data description 15](#_Toc49413991)

[Tab. 2: Project plan for data mining project 17](#_Toc49413992)

[Tab. 3: data sample from Aotizhongxin station 18](#_Toc49413993)

[Tab. 4: Brief description of data properties 21](#_Toc49413994)

[Tab. 5: Appended 12 file in a unique dataset 21](#_Toc49413995)

[Tab. 6 Number of Missing data in all station 30](#_Toc49413996)

**Summary**

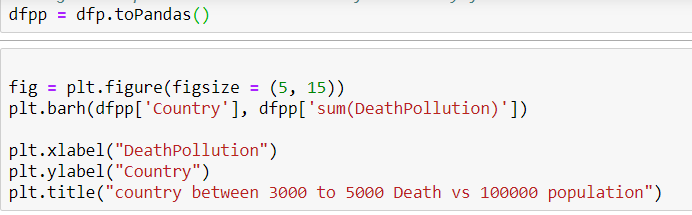
The main target of this study is Data mining and using different kind of machine learning (supervised method) to predict the most important air pollution matter which is called PM25 (particulate matter with diameter less than 2.5 \_mµ). In this study, Beijing City is divided into 6 12 stations with same data. Each air quality monitor stations location, as well as the date information (month, day, hour) are included in the feature spaces. Major air pollutant (PM25, PM10, CO, NO2, O3) measurements and climate information (Temperature, Pressure, Wind speed, DEWP) are also recorded they were used as an input. In this project after data cleaning and choosing the best data mining methodology some Regression algorithms such as Multi Linear and Generalized Linear Regression approaches by using the different PySpark Packages were used. In this study the target was using simple method to get high correlation between inputs and target. But the important aspect of this study is searching to find relationship between only one parameter (PM10) as an input and air pollutants such as PM25. However, climate data such as Temperature, pressure, wind speed were used as an independent data in pattern 2 and 4 but this study shows the regression models are not effective to predict pollutant material by only climate data. The first pattern includes combination of climate data and air pollutants worked very well and the correlation was about 0.84. However, the amazing achievement was in pattern-3, using only one input (PM10) was only imported to Simple Linear model and correlation was 0.82 approximately and based on meeting the criteria (r2>=0.7), we can generalize this approach to many station to other cities in China.

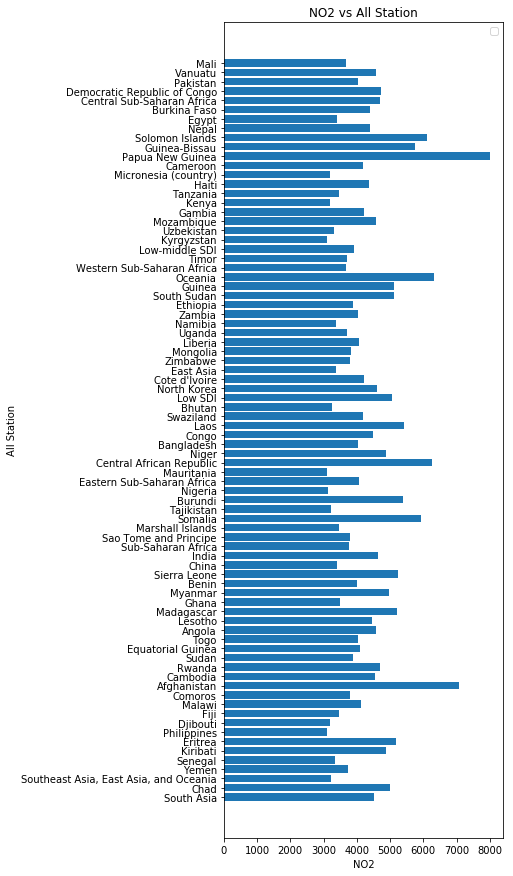
# Business and/or Situation understanding

## Identify the objectives of the business/situation

These days most of the developing countries in different kinds of regions are suffering from air pollution. Various elements like fossil fuels consumption and industrial plant emissions play significant roles in air pollution in different countries. Based on WHO statistics, Air pollution kills about seven million people worldwide each year. WHO information indicates 9 out of 10 people breathe air containing high levels of [pollutants](https://www.who.int/airpollution/ambient/pollutants/en/). Meanwhile, air pollution leads to a major threat to people's health and weather quality.  Based on WHO data, air pollution causes about 4.2 million deaths per year due to stroke, heart disease, lung cancer, and chronic respiratory diseases (World Health Organization, 2020).

Also, WHO shows air pollution influence developed and developing countries, low- and middle-income countries experience the highest-burden, with the greatest toll in the South-East Asia regions. **Fig. 1** shows the summation of mortality per 100,000 populations due to air pollution from 1990 to 2017 in different countries. Also, the distribution of mortality in some countries is clear in Fig. 2. Based on Fig. 3, the polluted countries are divided into 3 sections. The blue part reflects the most polluted countries and the red one indicates the developing countries with a high ranking in air pollution. In this category, China as a country with the most population in the world, high GDP, and high in air pollutants will be discussed in this research (Samet, J.M,2000). Below shows the PySpark code to generate the total death due to pollution number vs countries:





**Fig. 1:** total death per 100000 due to air pollution among countries from 1990 to 2017 in 50 countries that mortality

[<https://www.kaggle.com/akshat0giri/death-due-to-air-pollution-19902017>].

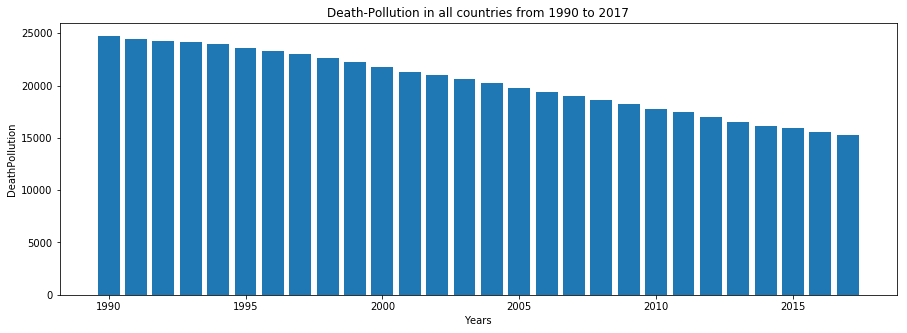


Fig. 2: death per 100000 population due to air pollution among some countries from 1990 to 2017 in three portions

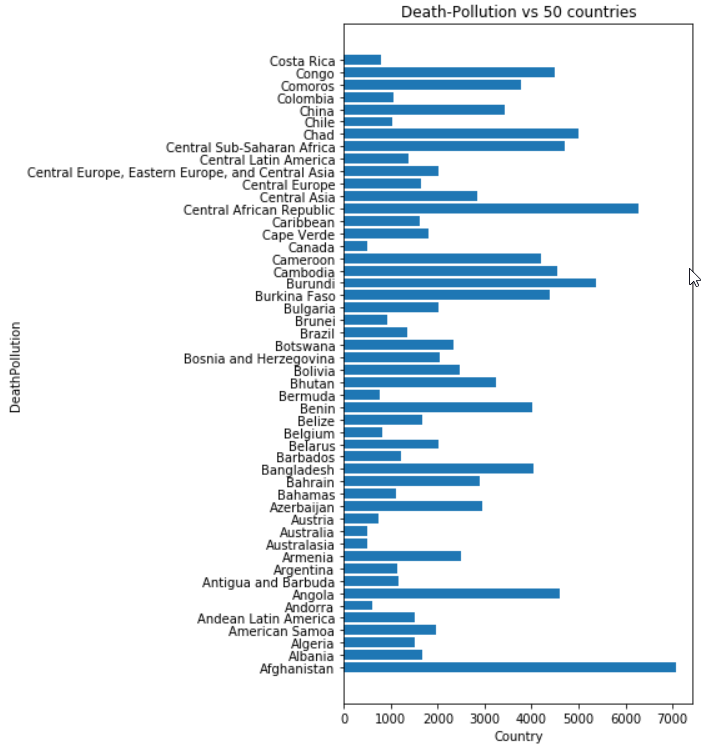


Fig. 3: summation of death per 100000 due to air pollution among countries from 2000 to 2017 in three portions

It is estimated that about 90% of the world’s population lives in countries where the air pollution problem. The most important pollution sources are cars, power generation, the heating systems in houses, and industrial activities and also, there is a close relationship between air pollution and weather quality globally. In mentioned countries, the government or Environmental organizations usually consider the regulation of pollutants in the air as a vital issue. Air Pollution topic has been one of the most important environmental consideration in China. In 2017, 75 per 100000 people have been died due to a lack of the national standard for air quality (Fig. 4). Among the cities, Beijing with about 20 million people encounters a rising population on one hand and economic development on the other hand. In all industry, coal and fossil fuel are the majority type of energy which are used in this city. Air pollution in Beijing depends on two major elements; weather conditions and human activity. Local weather parameters such as Wind speed, Humidity, Temperature, Pressure in Beijing have a great impact on air pollution.

Outdoor air pollutants in Beijing mostly include ozone (O3), particle matter (PM25 and PM10), sulfur dioxide (SO2), carbon monoxide (CO), nitrogen oxides (ozone). Increased mortality rates have been determined in association with rising air pollutants including sulfur dioxide, PM25 , and PM10 concentrations (Du, X., Kong, 2010).

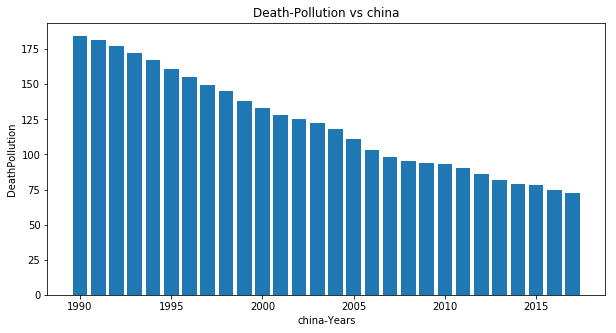
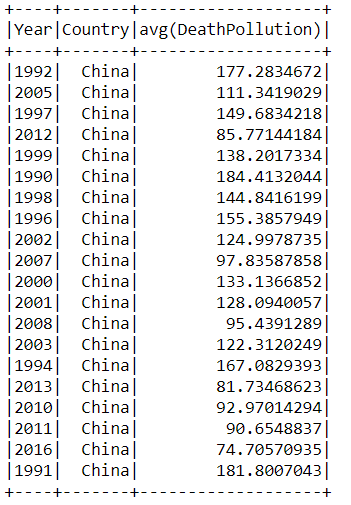


Fig. 4: distribution of air pollution total death from 1990 to 2017 in china



**Hence, this research will be conducted with the following objectives as below:**

* Based on above fining most of countries are involved yearly death because of air pollution. This issue is very important for china due to high population who are concentrated in Beijing. Above table and figure 4 shows the average of death in china because of air pollution. However, any investigation to get information about air pollutant material and research about their relationship and factors which have impact on them are very important.
* The fine particles (PM25 to 10) are associated with negative impacts on the respiratory morbidity of people and these elements are able to deposit into the lung gas-exchange region, so any investigation or prediction of this element in the air can help to prevent more death among people in Beijing (Du, X., Kong, 2010). Also, SO2, Co, O3, and No2 are important gases in the air which are regarded with respiratory problems.
* Local weather parameters such as Wind speed, Humidity, Temperature, Pressure, are very vital in calculating the air pollutant concentrations (Chan, C.K, 2008 and Chen, W.Y. & Xu, R.N,2010). Based on the different research, there is a direct relationship between the weather elements such as temperature, wind speed, pressure, and humidity with the concentration of gases and particle matter (PM25) (Holloway, 2008). For example, High humidity usually happens with high concentrations of PM25, CO, and SO2 but with low concentrations of other air gases like NO2 (Elminir,2005).
* On the other hand, Increasing or decreasing wind speed has a different impact on air pollution measurement. For instance, when the wind speed was low, the pollutants (PM25) were found at the highest concentrations. However, by increasing wind speeds we have a high concentration of dust and other particles (Elminir,2005). Therefore, these important meteorological attributes have a great impact on air pollution and they can be considered for the prediction of air pollutant concentrations.
* In the previous study, statistical computation has been used frequently for air pollution estimation for a couple of days based on some limited weather elements (Zheng, Y, 2013). Currently, some studies on statistical modeling have been focused on some simple regression models to predict PM25 or very limited particles. PM25 is a very important particle that has not been investigated with comprehensive input data.
* Besides statistical models, machine learning methodology has been enhanced recently with a huge amount of application in different aspects of life and industries (Yuan, Z, 2017). Fortunately, different new methods in Machine learning and deep learning have been invented which allow us for data mining of air pollution and Meteorological parameters comprehensively
* In this research, the main objective is data mining of massive air pollution particles (PM25), other pollutant data and weather data (Temperature, Pressure, Humidity, Wind speed) as a first step and evaluating all Machine Learning methods to predict the pollutant effectively in 12 regions (or stations).
* A considerable difference between this study and the previous ones is that this research will try to focus on utilizing all ML models to enhance the generalization performance of a complex model from big historical data from 2013 to 2017 in 12 regions in Beijing (Kurt, A,2010).

### Business Success Criteria

In this research, all important variables such as meteorological parameters and air pollution data in 2 different and separate models will be used as inputs for the prediction of PM25. Evaluation methods such as SSE, SSR, or SST will be used for the accuracy of different models. However, in Machine Learning methodology the R2 about 0.70 and more could be reliable.

## Assessing the Situation

For the prediction of the pollutant material, it needs some Meteorological parameters such as Temperature, Pressure, Humidity, and wind speed. All data comes from different regions in Beijing in CSV files and any station includes 18\*35000 rows of data. However, in this research, pollution factors (PM25) will be predicted and weather conditions and other pollutants in separate ML models are inputs in different kinds of machine learning approaches (regression, classification, clustering, deep learning…). In this research, different risks must be considered. Firstly, we face with big data as input. In this case, by the concept of data mining, we can refine the data and remove unrelated data (p-value can be a good method). Secondly, utilizing different ML approaches needs considerable time. This problem needs to make an effective plan. The last risk is using wind direction as an input to the model. By some research on different papers and also criteria such as R2, we can check the relationship with other parameters.

### Resource Inventory

For running different ML approaches and software such as, Python, R, SQL, Weka, it needs a normal laptop with enough memory. There is no certain limitation for any hardware specification. The air pollutant and meteorological data came from the UCI repository[<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>]. They are in CSV format in 12 categories in different regions in the Beijing area. They are free and stored in 2019 in the UCI repository website and show air pollution data from 1990 to 2017.

### Requirements, Assumptions, and Constraints

There are no special requirements or any constrain to access the data. All of them are free and downloadable from the website. This data set contains air pollution data in 12 nationally-controlled air-quality monitoring sites in China-Beijing. The air-quality data such as temperature, pressure…are from the Beijing Municipal Environmental Monitoring Center legally. The meteorological data in each air-quality site are matched with the nearest weather station from the China Meteorological Administration [<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>].

### Risks and Contingencies

Data comes from a valid source and there is a guarantee for data quality. Also, the first screening of data shows relative validity. However, if there is any missing or Null data, we can monitor and estimate them in the data mining phase.

### Terminology

There is some expression of data as Tab. 1:

Tab. 1: a data description

|  |  |
| --- | --- |
| Specification | Description |
| station code | **Twelve stations in Beijing** |
| SO2 | **sulfur dioxide** |
| NO2 | Nitrogen Dioxide |
| PM10 | **particle matter** |
| PM25 | **particle matter** |
| O3 | **ozone** |
| Co | **carbon monoxide** |

### Cost/Benefit Analysis

There is no cost in this project. Only it needs time which will be based on the project plan. One of the benefits of this project is the optimization of various Machine learning methods to predict air pollution elements by available and free meteorological data.

## Determine data-mining objectives

Up to now the business objective is obvious and it needed to turn it into a data mining concept. Relationship of rising the air pollutants to increasing of rate of mortality is the fundamental objective of the business but the estimation and measuring this kind of pollutants (in Beijing) are very critical information to help the business to protect the health of the people in society.

Based on understanding of air pollution problem in China (Beijing) we can summaries the data mining goals as below:

* The dataset which was collected includes 12 station in Beijing with two category of air pollution and climate data with suitable history from 2013 to 2017.
* It supposed there is close relationship between all air pollutants together. For example, it seems by increasing the NO2, probably the PM25 will increase. So one of the data mining goal in this study is fining the relationship between all materials; PM25, SO2, CO, O3, and NO2 together.
* Another important goal is finding relationship between air pollutant materials with climate data. It supposed to by decreasing the temperature, the PM25 will increase.
* By fining the correlation and relationship between all input data, using the historical data about previous information to generate a model by applying the machine learning is the final goal of data mining.
* The effective plan is dividing all data into 2 stages; training, testing which will be done by Python (PySpark). the air pollution matters such as PM25 will be the target with climate data with other gas data in *2 different machine learning models are the inputs*. In the first model, climate data such as temperature, pressure, wind speed are used as an input data and in the second model, more data such as climate and other air pollution matter such as Co2, No2, So2, will be used as a second input data. Finally, the best model will be introduced.
* In the future, by having the correct model of air pollution prediction, we are able to forecast the quantity of martials easily.

### Data Mining Success Criteria

Different evaluation methods will be used for Machine learning assessment. RMSE is the most popular evaluation metric used in the ML regression process. It says that errors are unbiased and follow a normal distribution and R-Squared or R2 is another reliable method for prediction evaluation. Generally, in the training and testing sections, because of its importance in the ML workflow, the R2 could be at least 0.70 to 0.80. However, the other models such as F-Measure or Confusion Matrix will be considered in ML (classification approach) workflow.

## Project Plan

Tab. 2: Project plan for data mining project

|  |  |  |
| --- | --- | --- |
| Phase | Time | Description |
| Business Understanding | **1 week** |  |
| -Objectives | 2 days | Business objectives, success, and criteria |
| -Situation Assessment | 2 days | Inventory of requirement, assumptions, constraints |
| -Data mining Goals | 2 days | Determination of Data mining goals and success |
| -Project Plan | 1 day | Making a project plan |
| Data Understanding | **1 week** |  |
| -Initial Data Collection | 2 days | Proving data collection report |
| -Data description | 2 days | Proving data description report |
| -Data exploring | 2 days | Proving data exploring report |
| -Data quality verification | 1 day | Proving data verification report |
| Data Preparation | **3 weeks** |  |
| -Data selection | 5 days | Inclusion and Exclusion of data |
| -Data cleaning | 4 days | Proving data cleaning report |
| -Data construction | 4 days | Attributes derive |
| -Data integration | 5 day | Data merging |
| -Data formatting | 3 days | Data reformatting |
| Modeling | **2 weeks** |  |
| -Model selection | 4 days | Different ML models selection |
| -test designing | 4 days | Test designing |
| -Model building | 3 days | Parameters models setting |
| -Model assessment | 3 day | Revised parameters settings |
| Evaluation | **1 week** |  |
| -Result Evaluation | 2 days | Assessment of data mining results |
| -Process Review | 2 days | Review of process |
| -possible action decision | 3 days | Providing the list of possible action |
| Deployment | **2 week** |  |
| -Plan deployment | 4 days |  |
| -Plan monitoring | 4 days |  |
| -Produce Final Report | 3 days |  |
| -Project review | 3 days |  |

# Data Understanding

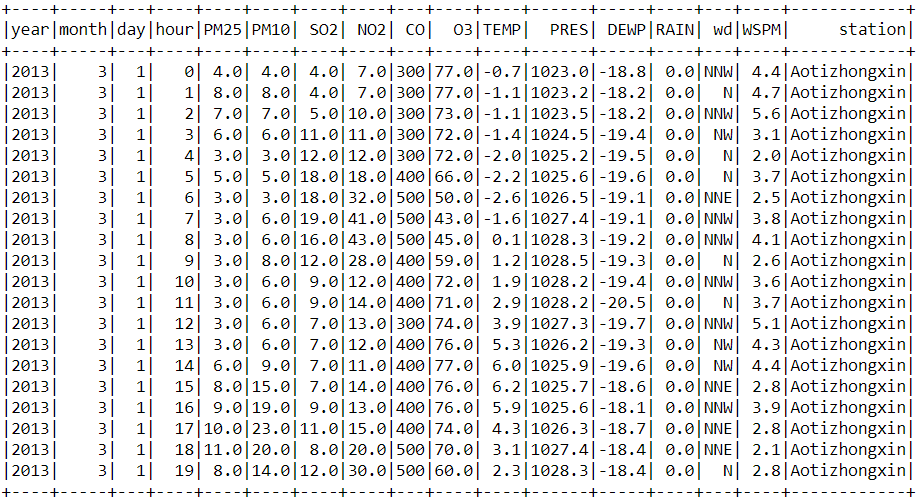
## Collecting Initial Data

Daily and hourly API from 2013 to 2017 at the data center of the Ministry of Environment Protection of China was gathered. Data from a total of 12 stations (include: Gucheng, Nongzhanguang, Guanyuan, Tiantan, Dongsi, Wanshouxigong, Aoti, Wanliu, Changping, Huairou, Shunyi, Dingling in Beijing area) was publicized on this website and can be downloaded freely. Each dataset was divided into three sections; the first section includes Year, Month, Day, and Hour from 2013 to 2017. The second one contains air pollution elements; PM10, PM25 , SO2, NO2, CO, and O3. The last portion includes climate parameters such as Temperature, Pressure, Humidity, Wind speed measurement, and the name of the station. Among 12 distribution of the monitoring stations in Beijing, eight one is located in urban stations (red zone in Fig. 5), three of them are situated in the suburb of Beijing (purple one in Fig. 5) and one background station is located at Dingling, far away from human activity (green one in Fig. 5).

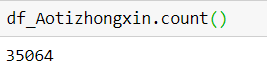
This information comes from air-quality monitoring sites in the Beijing Municipal Environmental Monitoring Center legally. Also, the meteorological data was taken from the nearest weather station from the China Meteorological Administration

[<https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data>].

Tab. 3: data sample from Aotizhongxin station









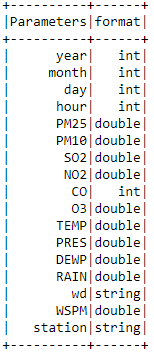
*Fig. 5: Distribution of National Air Quality Monitoring Stations in Beijing(Wei Chen,2015)*

### Additional data

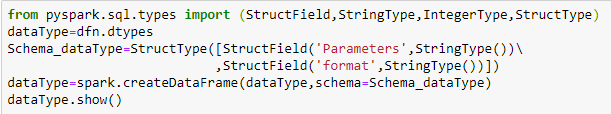
Air Pollution is one of the important reasons for the killing of innocent people in different countries. So for in the previous chapter, this data was used to open the eyes and give you a great insight into the importance of air pollution's effect on lives. However, this kind of data will not be used in our Machine learning model. This data includes some column as below:

* Entity: It is the name of the country.
* Code: It is the Code of the country.
* Year: it shows the period 1990 to 2017
* Air pollution (total) (deaths per 100,000): total death
* Indoor air pollution (deaths per 100,000): death due to indoor air pollution.
* Outdoor particulate matter (deaths per 100,000): death due to outdoor pollution.
* Outdoor ozone pollution (deaths per 100,000) : Death due to ozone pollution [<https://www.kaggle.com/akshat0giri/death-due-to-air-pollution-19902017>].

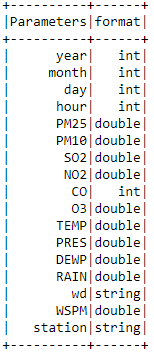
## Describing Data

Tab. 4 explains the data type and size in all stations. It shows specification, data type, data unit, and description. All 12 stations data were loaded in Python software and they were appended together as a unique dataset. Based on that the number of row data about 420000 (

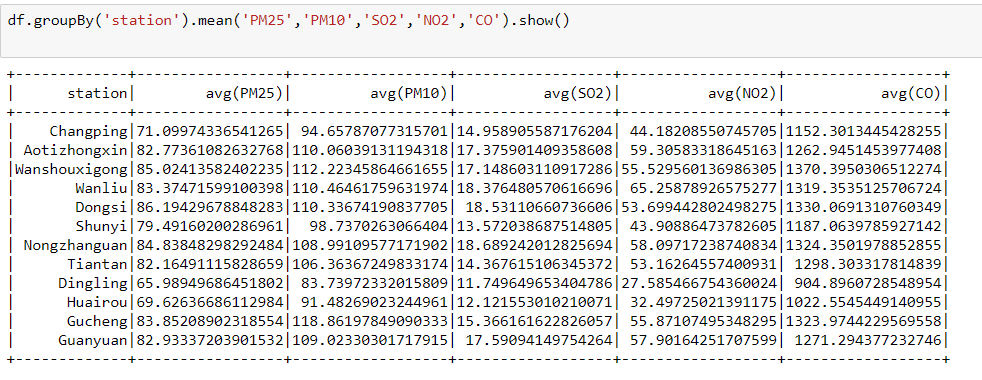
Tab. 5). All air pollution data (So2, No2, PM25 , Co, O3) have a negative impact on lives and they could be as a target in the model. They are in the same format in unit and type. Other climate data such as Temperature, Pressure, Wind Speed are the input to the model. The time is divided into different categories from year to an hour in 12 situations in Beijing (Tab. 4). Below shows the Python code to show the data format including string, integer, and double. However, Table-5 indicates average of air pollution pollutants in all stations together.

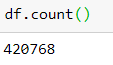


Tab. 4: Brief description of data properties



Tab. 5: Appended 12 file in a unique dataset

****

****

## Exploring Data

Based on business and datamining goals, exploration the data with charts, graphs or tables can assist us to address the goals which were mentioned in business understanding stage. The most important issue in this stage is understanding those air pollution data and their relationship to each other beneficially. Sometimes by declining of some parameters such as temperature, the air pollutants such as PM25 will be increased, Fig. 6 indicates that the PM25 is maximum in temperature -10 to 0. Another important exploration was the relationship between increasing of PM25 and time. Fig. 7 shows that PM25 usually rise in end and beginning of the year. It should say that increasing of air pollution in this time could be due to increasing of travelling of people in Beijing. Meanwhile by increasing the other air pollutant such as NO2, the PM25 will increased intensively. So the exploration of data pattern shows different relationship between PM25 and the other data.Another important of data exploration is investigating on air pollutants pattern from 2013 to 2017.

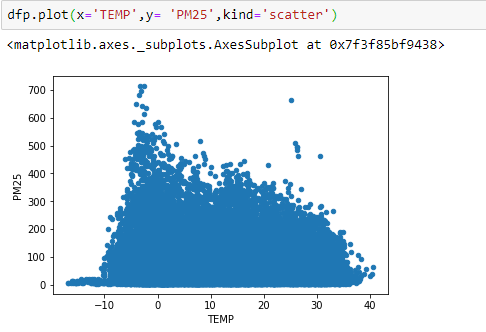


Fig. 6 Relationship between PM25 and temperature data from 2013 to 2017 in all station

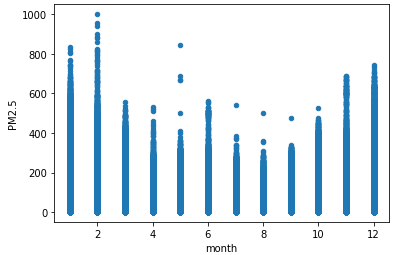


Fig. 7 Relationship between PM25 and month data from 2013 to 2017

Investigation to find the best correlation between parameters are very important in this section. Fig. 10 to Fig. 14 indicate the relationship between air pollution materials by PM25 from 2013 to 2017. Based on them, PM25 have a strong correlation with PM10, NO2 and CO. but the impact of SO2 and O3 is not the same as others and there is no clear and obvious pattern with PM25. In most of them, the amount of pollutants is increasing gradually from 2013 to 2016.

Also, the distribution plot between PM25 and all climate parameters (Fig. 13 and Fig. 14) will be seen. Also there is no strong correlation between PM25 and other pressure and temperature data. It shows the correlation between air pollutant material with climate data is weak and probably for modelling it needs to use air pollution data as an input. These kinds of graphs help us to find relevant or irrelevant climate parameters effectively.

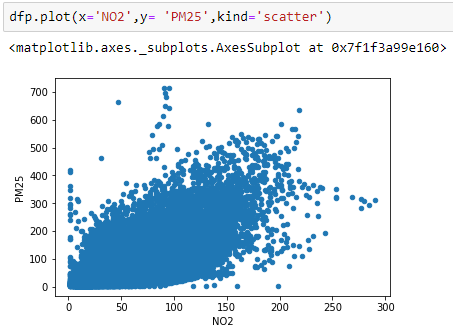


Fig. 8 Relationship between PM25 and NO2 data from 2013 to 2017 in all station

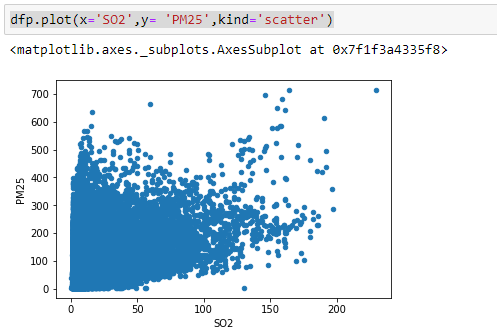


Fig. 9 Relationship between PM25 and wind direction data from 2013 to 2017 in all station

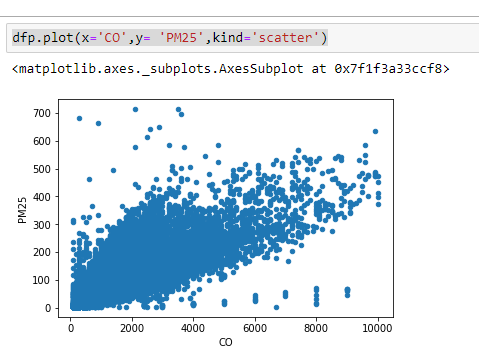


Fig. 10: Relationship between PM25 and CO data from 2013 to 2017 in all station

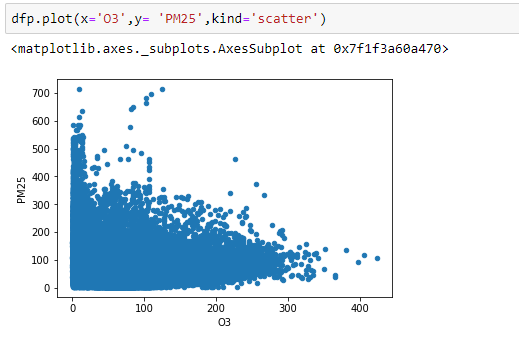
:

Fig. 11: Relationship between PM25 and O3 data from 2013 to 2017 in all station

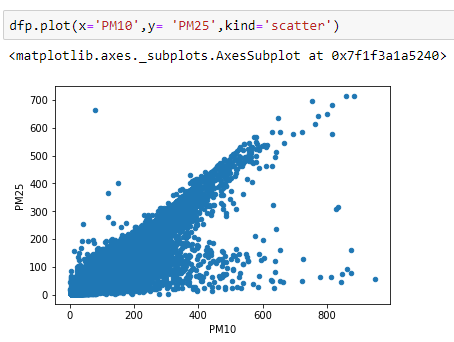


Fig. 12: Relationship between PM25 and PM10 data from 2013 to 2017 in all station

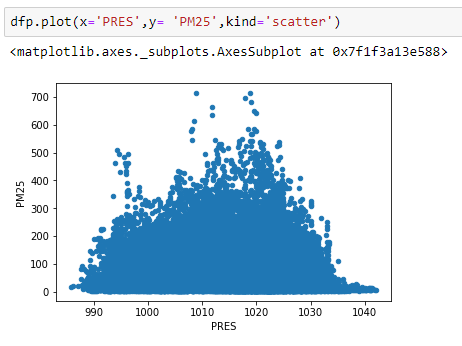


Fig. 13: Relationship between PM25 and PRES data from 2013 to 2017 in all station

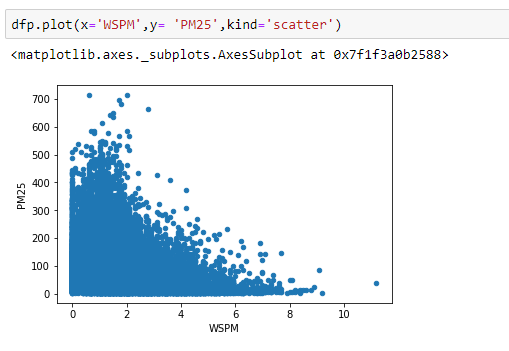


Fig. 14: Relationship between PM25 and WSPM data from 2013 to 2017 in all station

Investigation on data distribution is another important issue that we have to care about that. Actually based on above fining there is close relationship between two parameters PM10 and an input and PM25 as an output or Target. Fig. 15 to

Fig. 17 indicate on distribution of some critical parameters as three histograms. Distribution of PM25 and PM10 (Fig. 15) is very similar in sharing and frequency between 1 to 100. But the other ones have a little fluctuation to each other. Based on the

Fig. 17 most of the SO2 data are concentrated at the beginning of graph between 1 to 50 but NO2 distributed from 1 to 250 and it shows their relationship is not very clear and need more investigation.

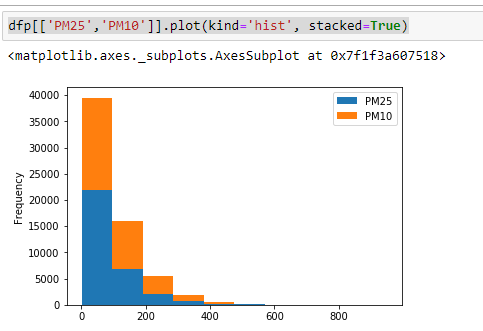


Fig. 15: distribution of PM25 vs PM10

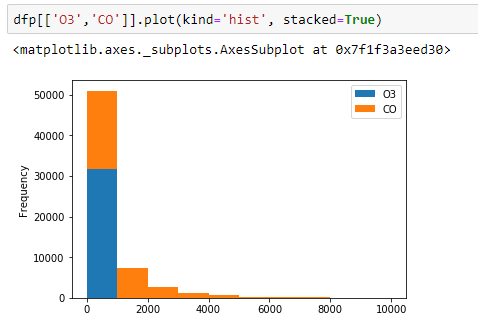


Fig. 16: distribution of O3 vs CO

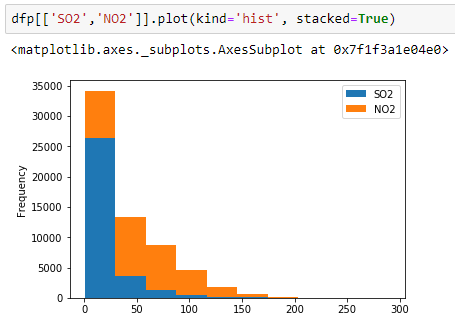


Fig. 17: distribution of SO2 vs NO2

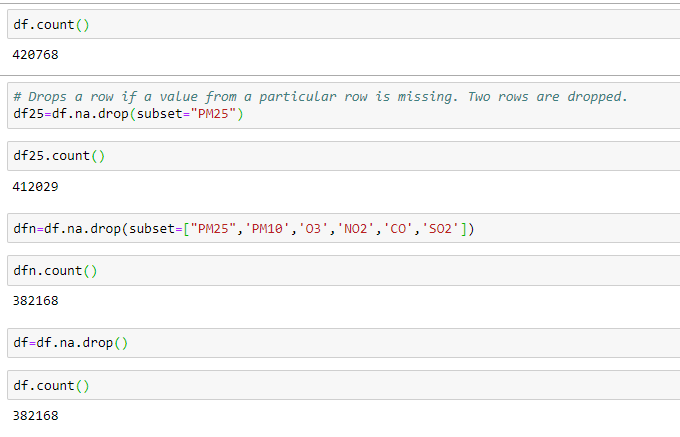
## Verifying Data Quality

All 12 station data was loaded to Python (Fig. 19) shows about 420000 rows. Also Fig. 19 indicates the data audit in all 12 station from 2013 to 2017, more than 95 percent of data are valid and they can be used for modeling steps. In this phase, there are a few null values among air pollution data (Tab. 6).

The appended file includes more than 415000 valid row data for all 12 stations in Beijing.

Missing or null data are very vital and before any simulation, they must be investigated precisely. Among all data about 2 to 4 percent of data are missing. Tab. 6 shows the missing and complete data in all stations for all input and target data. Co and O3 are data that have the most null data (about 4 percent of all data).

By more investigation on null value, PM25 as an Target must not include the Null or NaN values. By dropping Null value in PM25 and other parameters, the size of data is about 412029 (df25). Also, by dropping all null in all input data the size is about 382168 (dfn), so it seems to better the dfn model because we lost only few data by dropping all Null values.



Tab. 6 Number of Missing data in all station

On the other hand Fig. 18, and Fig. 19 show the all station mean values for all parameters after removing the Null values. Fortunately, the size of all data are the same about 382168 number.

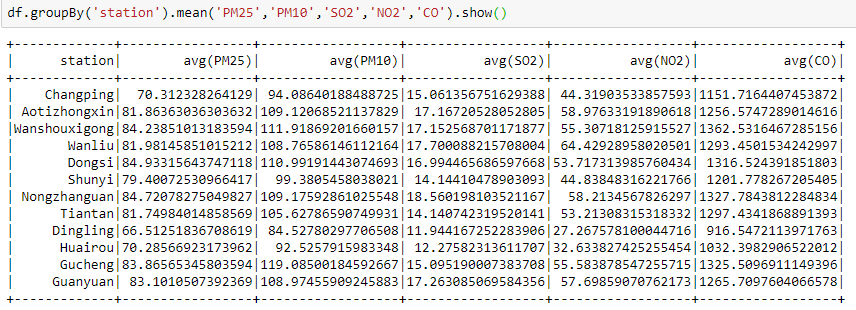


Fig. 18: Data import in all stations

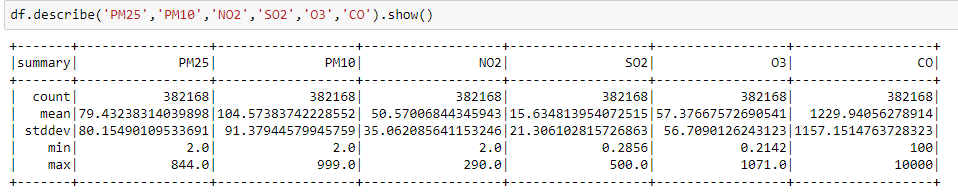
****

Fig. 19: Data description in all station , includes: Min, Max, Mean, std,….

Fig. 20: Data quality in stations after removing the missing data in all rows

# Data preparation

Data preparation is one of the most important and time-consuming aspects of data mining. In fact, it is estimated that data preparation usually takes most of a project's time and effort. However, spending suitable attempt on earlier business understanding and data understanding phases can lessen this overhead, but it still needs to expend a good amount of effort preparing and packaging the data.

## Select the data

In this section the relevant data can be selected correctly and based on that there are two different approaches to choose the best data as below:

* **Selecting items (rows)**

As the data properties were discussed in previous chapters, this dataset includes 12 stations in Beijing area and this stations measure and collect the air pollutants concentration and climate parameters from 2013 to 2014. Each station includes more than 35000 rows data they were imported to Python and appended as a unique dataset. About 415000 rows after dropping the missing data were selected to make a model (Fig. 21 to Fig. 23).

* **Selecting attributes or characteristics (columns)**

Two categories of data were selected for modelling which include air pollution data; PM2, PM10, CO2, NO2, CO, C and climate data; Temperature, Pressure, Wind direction and speed, DEWP.

For more investigation on the efficiency of input data, two kinds of models with different input data will be applied in the prediction phase. In the first model, only climate data such as temperature, pressure, humidity, and wind speed (or wind direction) are inputs (Fig. 24). Meanwhile, in the second model all data will be applied as an input (Fig. 25). Day, year and Rain data were removed from selected data.

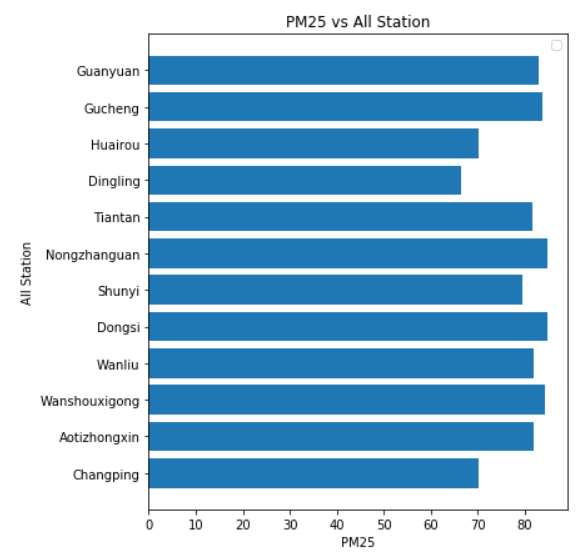


Fig. 21 appended data vs PM25 in all station

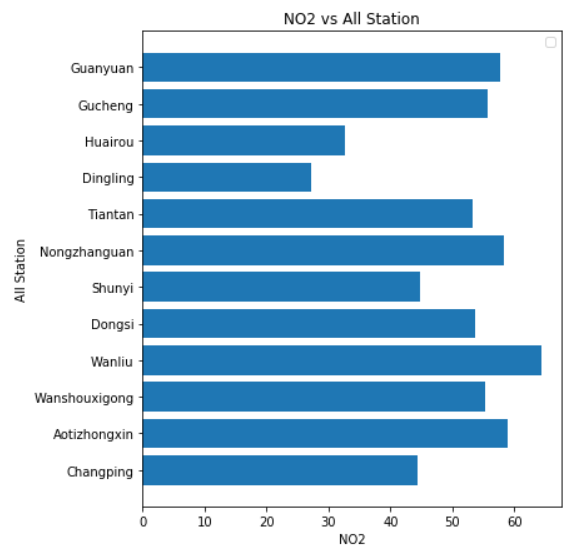
****

Fig. 22: appended data vs NO2 in all station

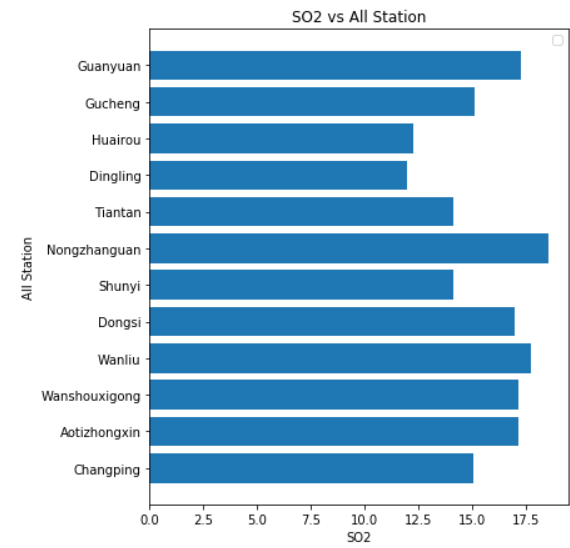


Fig. 23 appended data vs SO2 in all station

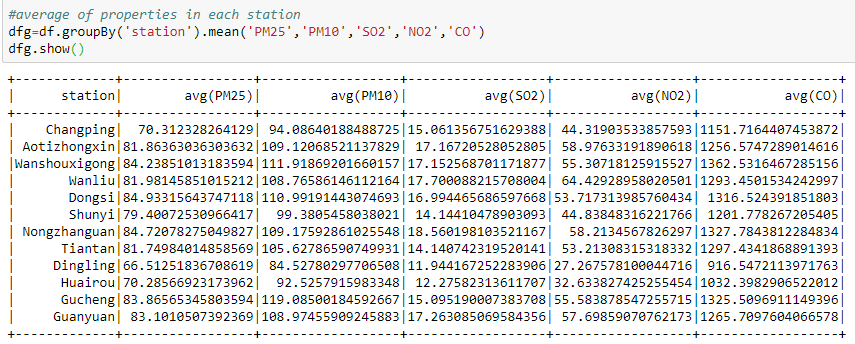


Fig. 24: the first model for PM25 prediction and relevant inputs ( air pollutant data). Target data is in red region

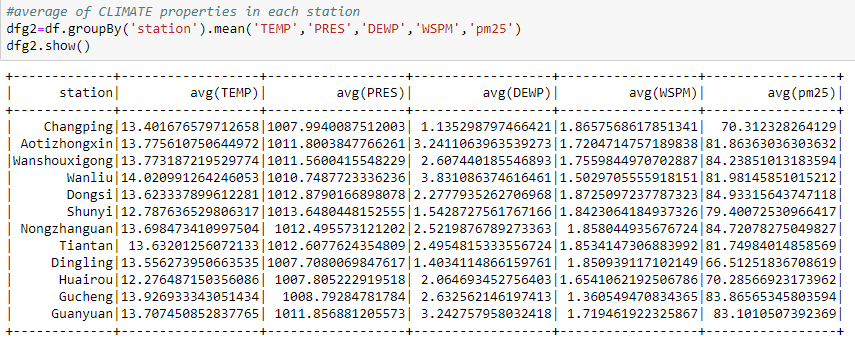
****

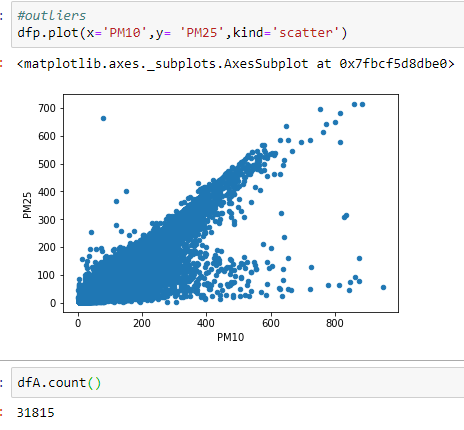
Fig. 25: the second model for PM25 prediction and relevant inputs (weather data). Target data is in red region

## Clean the data

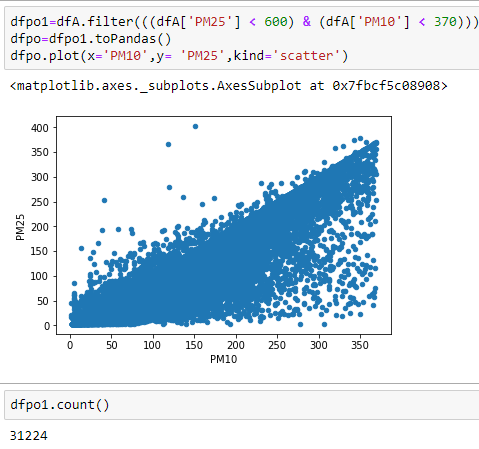
Cleaning the data involves taking a precise and serious investigation on the missing or null data in the above dataset. As *Fig. 26*, *Fig. 28*, and *Fig. 31* show outliers or extreme values that they are clear among all data in PM25, PM10 , CO and NO2. However, they could be less than 5 percent among all data. Most Outliers, Extreme and Null values belong to PM10, PM25, and CO and NO2. In Action column, the Outliers and Extreme values can be optimized. Coerce strategy was chosen for this issue, because of huge among of correct data we can change the outliers to precise data. On the other hand, two input data; TEMP and DEWP exclude any outliers’ data.

*Fig. 27* show the effect of removing the outliers from PM10 and PM25 as a using filter function and all data more than 600 and 400 for PM25 and PM10 respectively was removed. Also by using right threshold for NO2 and CO all outliers which were determined by red line, were omitted in *Fig. 30*.

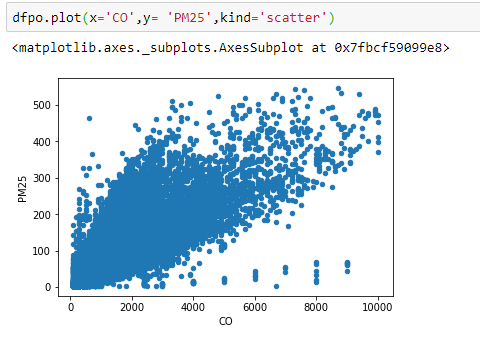
However, based on programming in Python max and min quantile were chosen for mentioned parameters. **Error! Reference source not found.** to **Error! Reference source not found.** reflect the impact of correct data on the cross plots. However, the other data remains as before and all of them will be used for prediction models.



*Fig. 26: Outliers in the scatter plot ;PM10 and PM25*



*Fig. 27:removing the Outliers in the scatter plot ;PM10 and PM25*



*Fig. 28: Outliers in the scatter plot ;CO and PM25*

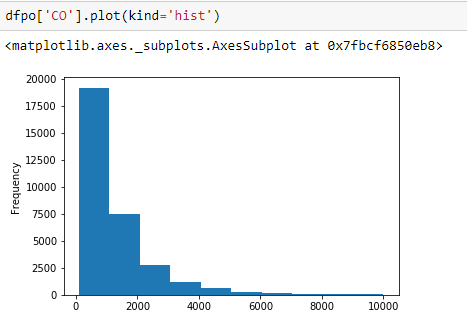
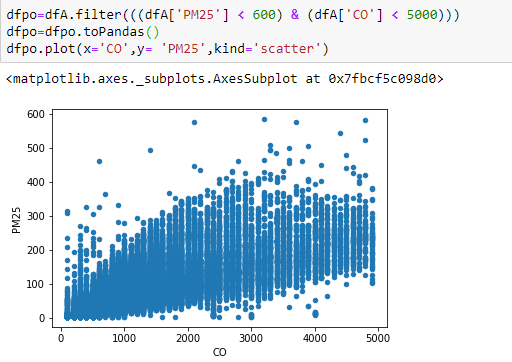
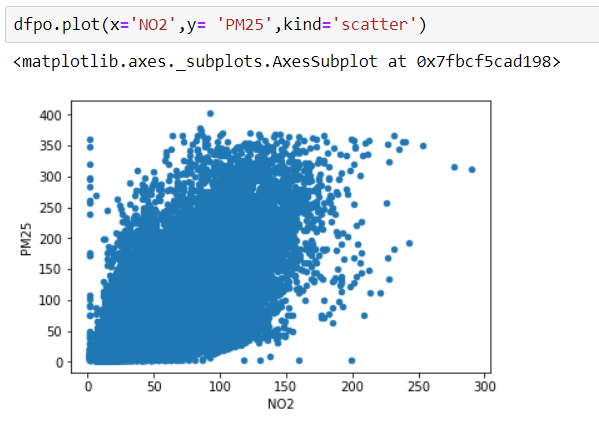


Fig. 29: distribution of CO

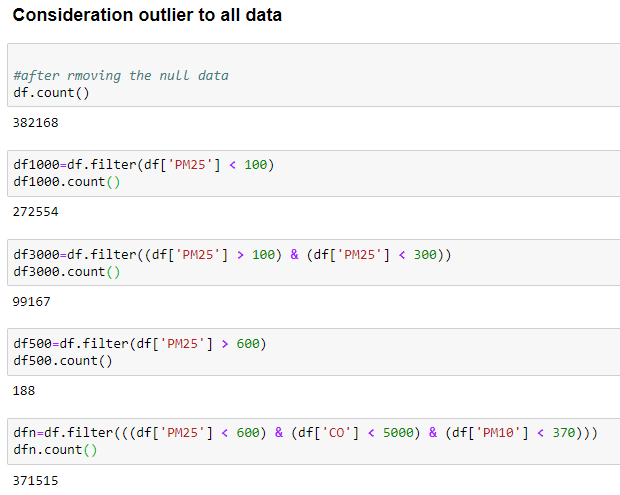


*Fig. 30:removing the Outliers in the scatter plot ;CO and PM25*

**

*Fig. 31: Outliers in the scatter plot ;NO2 and PM25*

By using the threshold for outliers in NO2, CO, PM25 and PM10, all data was filtered and based on below code programming in PySpark ‘dfn’ is the final dataset which is excluded from Null data and Outliers.



## Construct the data

In this stage, we mention the analysis results about the different PM25 concentration values for each station in Beijing. Fig. 49 shows PM25 distribution in all station and displays the duration of three PM25 concentration ranges PM25 <100, 100<PM25 <300, 600<PM25 µg/m3. However, based on the PM25 distributions in Fig. 32, 50 percent of data is less than 100 µg/m3 and this level could be good indicator of pollution boundary in the Beijing. By Drive application in PySpark, the new data as a flag was generated which is called pollution –index. However, this data will not be used as an input to model. Because it was calculated from target PM25.

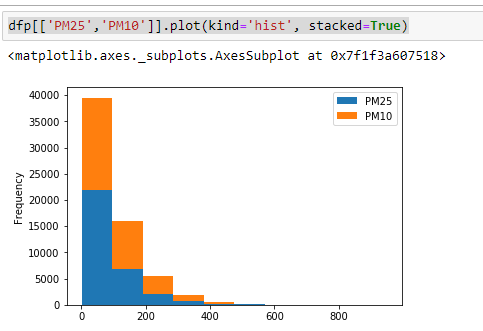
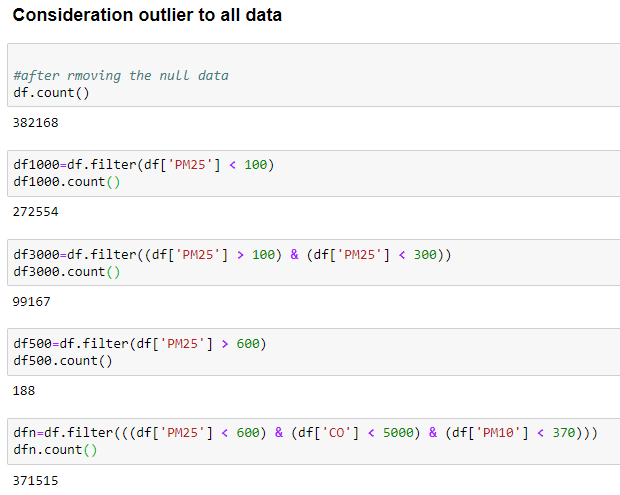
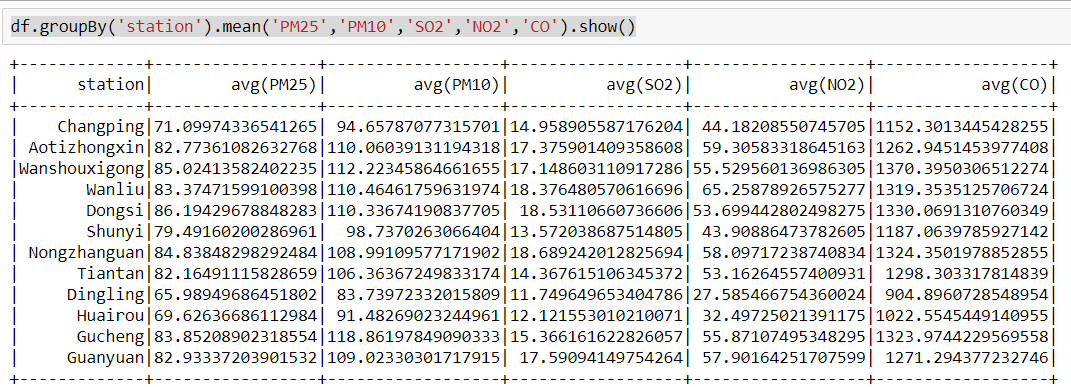


Fig. 32: distribution of PM25



## Integrate various data sources

As the data description was described in previous section, in this study, 12 station data were imported to model and because of high data size, all Null and outliers were deleted. Below code in PySpark explains the data of any station and import them as a unique file with applying all thresholds. However, working on unique dataset for missing value investigation is very effective. Based on aggregation function all 12 station in Beijing are integrated to each other. Average values for important factors such as NO2, PM25 and SO2 in all station in figure Fig. 35 to Fig. 36.



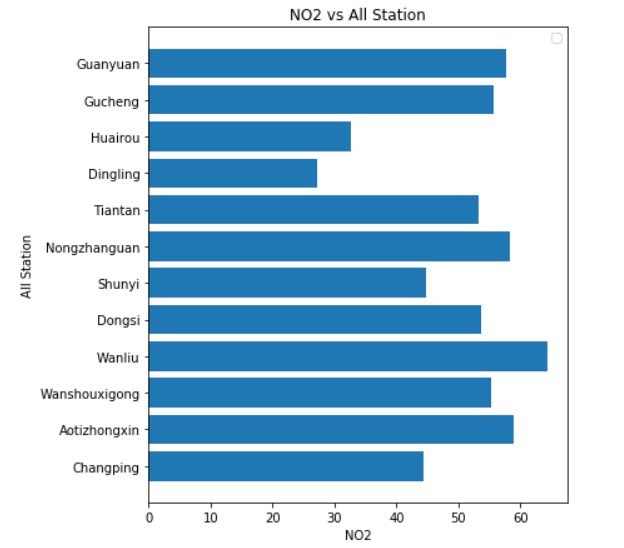


Fig. 33 Bar chart of the average of NO2 in all station

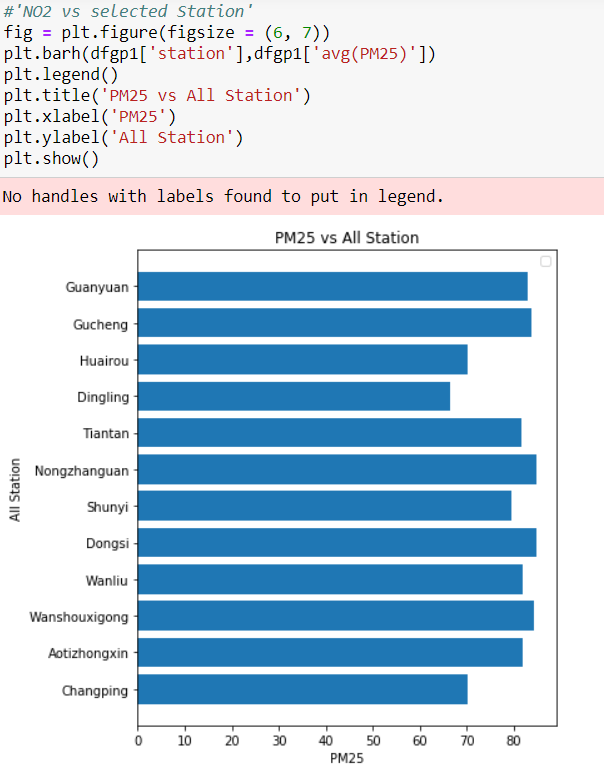


Fig. 34 Bar chart of the average of PM25 in all station

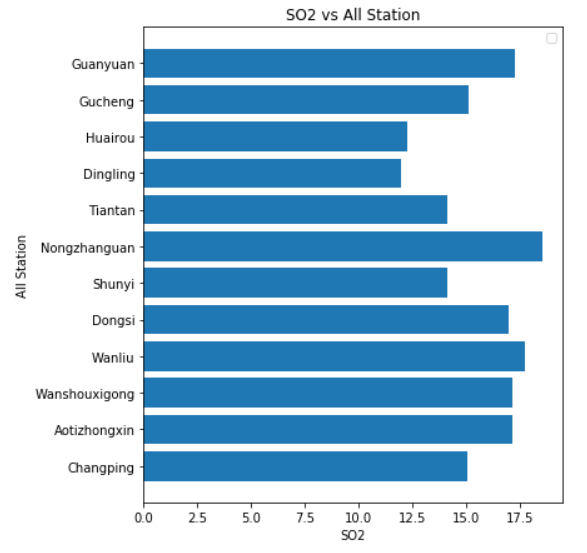
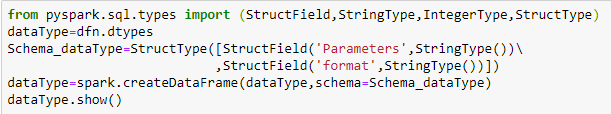


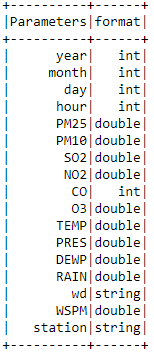
Fig. 35 Bar chart of the average of SO2 in all station

## Format the data as required

The last part of data preparation is data formatting and in this stage the input and target must be checked whether certain techniques require a particular format or not. For example, all pollution data are numeric (integer and float data) and double but the pollution index is flag type data. However, some other data such as station are text or categorical data which will not to import to the model.



Tab. 7: Brief description of data properties



# Data transformation

## Reduce the data

In this section we must reduce some data which is not very important in modelling. in data mining workflow, we usually may involve huge amount of data that can potentially be used as and inputs. However, we have to spend plenty of time to find relationship between all variable step by step. For limitation of choices, the Feature Selection algorithm can be utilized to recognize the fields that are most important for a given analysis. Here for predicting the PM25 as a vital parameter in air pollution issue we must discover which inputs have a great impact on PM25. the below coefficient shows there is good correlation between PM25 and other data. Among them PM10 and CO have a proper correlation with PM25 rather than the parameters (Fig. 36 and Fig. 37, because of the memory the feature selection was run separately in two steps). Fig. 38 to Fig. 43 indicates the relationship between PM25 and other variables. However, the climate parameters such as Pressure (PRES) have a low correlation with PM25, but because of its important role to estimate the PM25, it will be used as an input to simulation.

Specs Score

3 CO 2.177728e+08

0 PM10 3.061565e+07

2 NO2 4.324122e+06

1 SO2 2.787069e+06

4 O3 4.883934e+05

5 TEMP 1.273665e+05

6 PRES 1.559138e+03

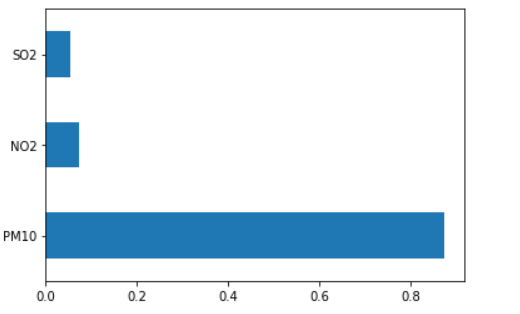


Fig. 36 Feature selection to choose the best input; PM25 , NO2 and SO2

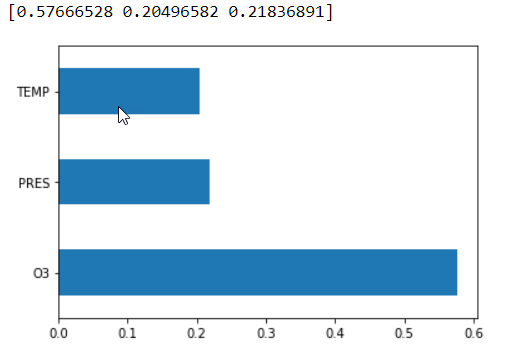


Fig. 37 Feature selection to choose the best input; O3, PRES and TEMP

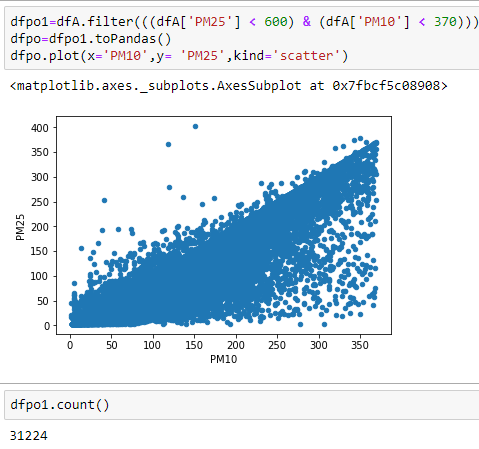


Fig. 38Correlation between PM25 and PM10

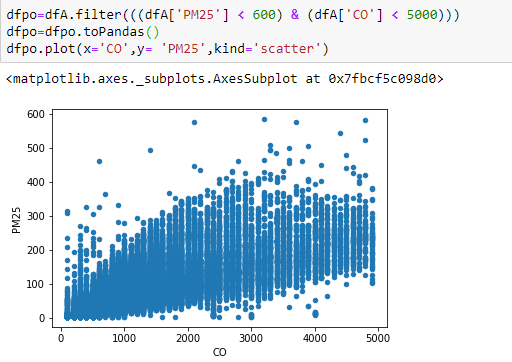


Fig. 39 Correlation between PM25 and CO

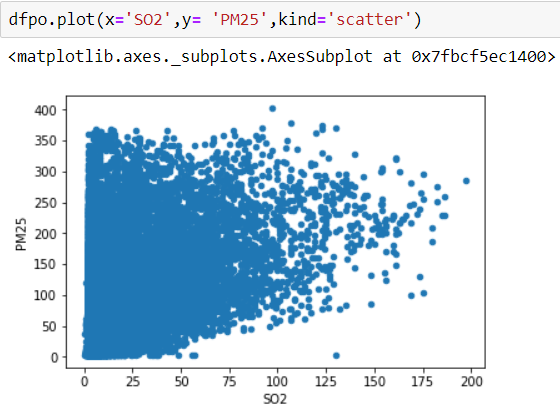


Fig. 40: Correlation between PM25 and SO2

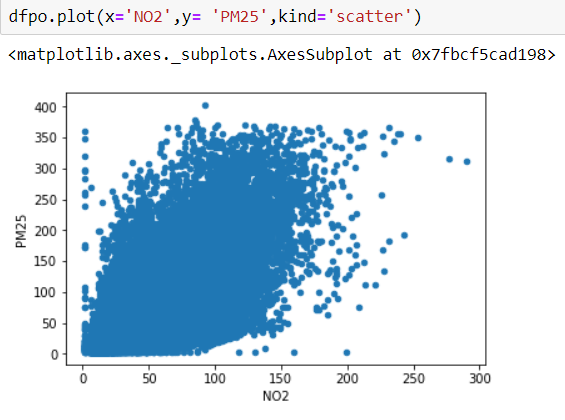


Fig. 41 Correlation between PM25 and NO2

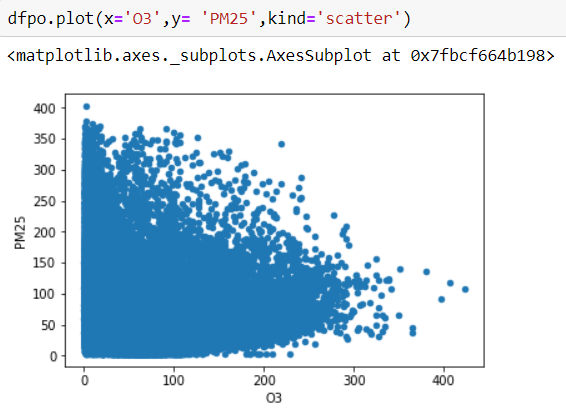


Fig. 42 Correlation between PM25 and O3

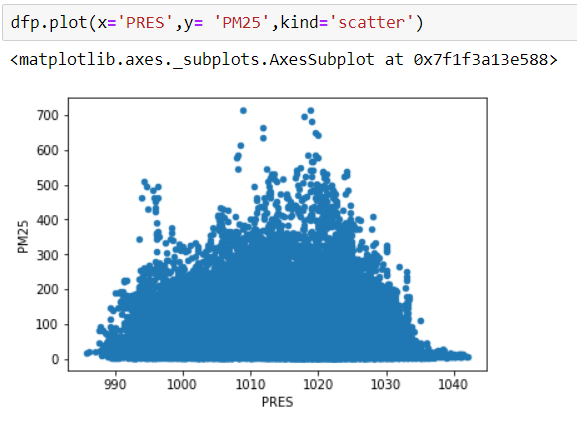


Fig. 43 Correlation between PM25 and pressure data

Based on station location and relationship of for all parameters in each stations, it needs to ignore some of the the station that their values are not inline to each other. Fig. 44 to Fig. 47 show distribution of average of PM10, PM25, SO2, CO, O3, NO2 in all station. Based on that only the values of 3 stations are not in line with others and finally the data in these station; Changding, Dinglig and Huairou were skipped from final data set. Fig. 48 indicates the aggregation of parameters in selected station and also Fig. 49 and Fig. 50 explain the NO2 and SO2 value in selected station. It indicates that now the data are consistent more.

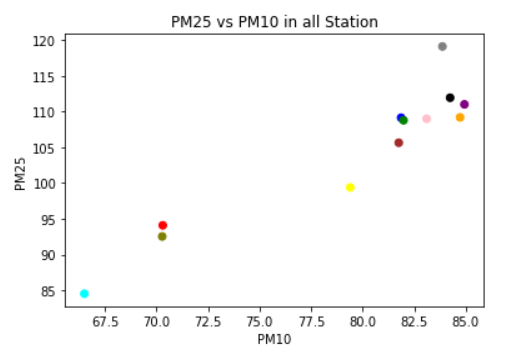


Fig. 44 PM10 and PM25 distribution in 12 station

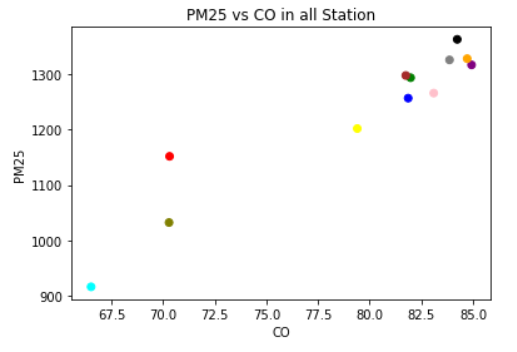


Fig. 45 CO and PM25 distribution in 12 station

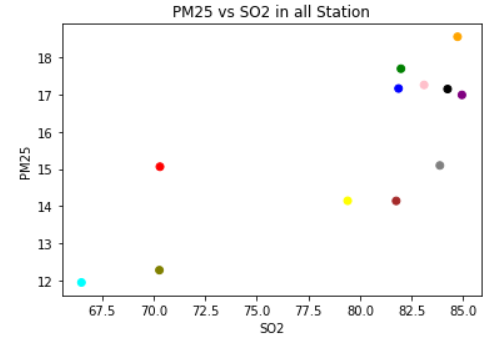


Fig. 46 SO2 and PM25 distribution in 12 station

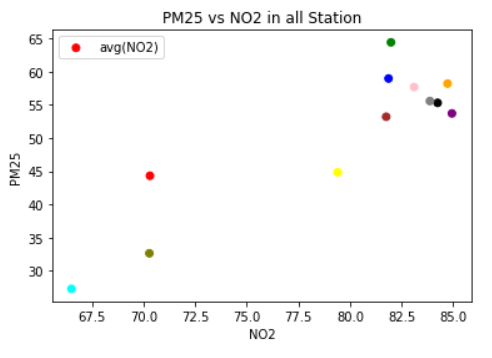
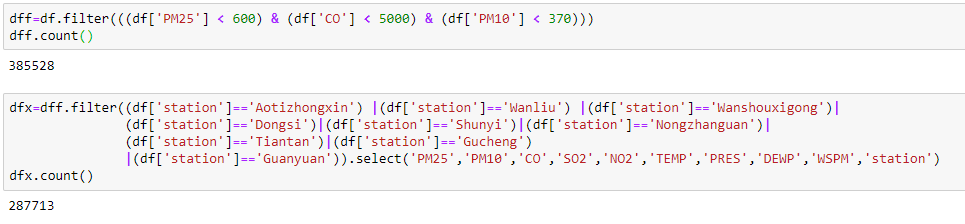


Fig. 47 SO2 and PM25 distribution in 12 station



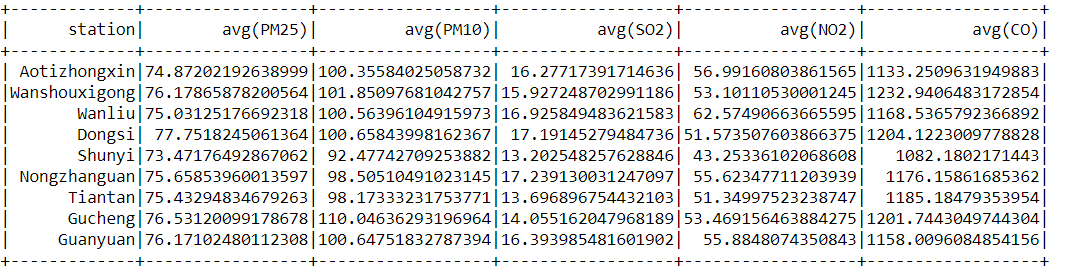


Fig. 48 Aggregation of data average in selected station

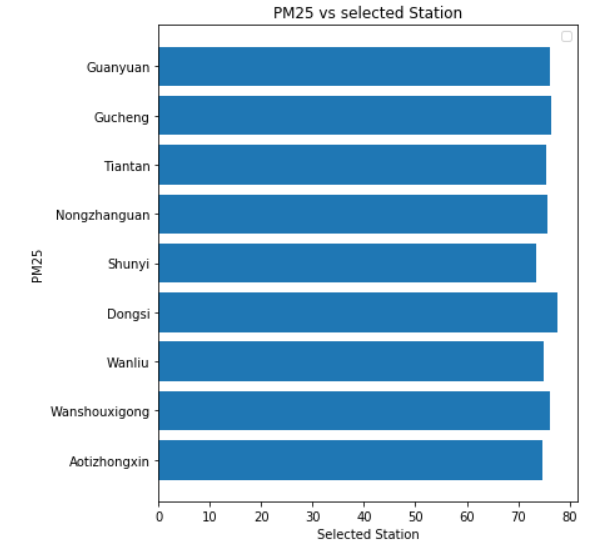


Fig. 49 Boxplot of PM25 distribution in 9 station

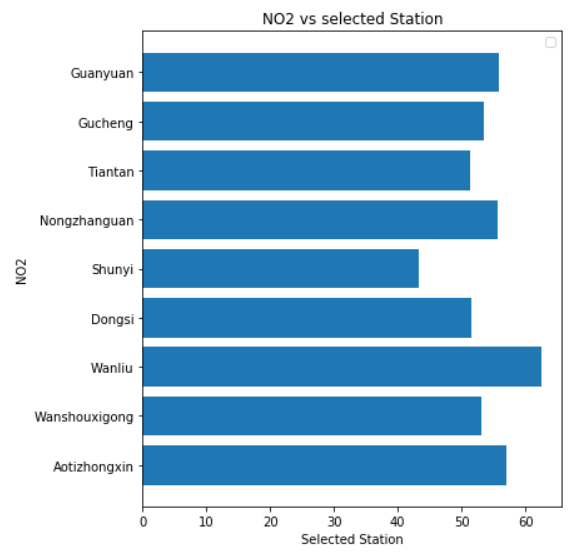
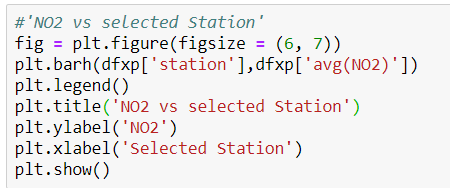
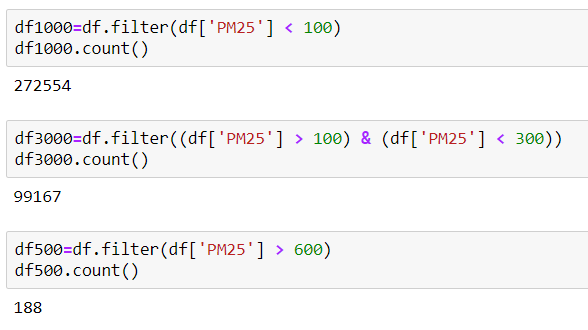


Fig. 50 Boxplot of NO2 distribution in 9 station



## Project the data

Pollution index is one of the new data which was generated as a new feature to classify the pollution as a different category. The objective is to have index with flag type from low, medium, high pollution. Based on the below category we are able to generate new flag data and consider it as a new value as a target except than PM25. most of pollution data distribution is low to medium.



# Data-mining method(s) selection

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

Air Pollution has been always considered as one of the most important problem about people heath in Beijing. Major air pollution matters, especially Fine particles (the matter with diameter less than 2.5 \_µm) which is called PM25 are robustly associated with adverse health effects, including cardiac and respiratory morbidity and finally mortality. Research from an independent group even showed that air pollution causes 1.6 million deaths in 2016 (Rohde, 2015). So measuring, monitoring and prediction of this features like PM25 by using data mining approach can help to reduce the mortality rate.

Two different data will be used for model the target. The first group is combination of climate data with other related air pollutants and second group is only climate data such as Pressure, Humidity, Temperature….

ML teaches the computers to do what comes naturally to humans: learn from all experience. ML algorithms utilize computational approach to learn all knowledge from data without relying on a predetermined equation as a model. The algorithms adaptively enhance their performance as the number of samples available for learning rising up. Supervised and unsupervised are two kinds of Machine Learning methods who are applied a lot. Supervised learning entails learning a mapping between a different relevant input data and a target and applying this mapping to predict target (PM25 ) in future. Unsupervised ML focus on dataset that the target is not available and the goal is looking for the relationship between input data (Dayan, P. ,1999).

There are different types of the methodology to categorize the data mining technique but base of PYTHON system, Data mining modeling methods can be divided into these categories ( Knowledge Center):

* **Supervised**
* **Association**
* **Segmentation**

***Supervised approach:***

In this method the objective is predicting the outputs by using one or more input data. This technique includes wide range of approaches such as *decision trees* as a one of the important classification methodology (C&R Tree, QUEST, CHAID and C5.0 algorithms), *regression* approach which includes variety of methods algorithm (linear, logistic, generalized linear, and Cox regression algorithms), *Artificial Neural Networks* (ANN), *support vector machines (SVM, LSVM)*.

Supervised models is useful method to predict a target which is clear and the objective is prediction of that data. It this project, the main goal is predicting the air pollutants such as PM25 by using two different types of the data.

Among of the supervised method, the ***classification and Regression*** (C&R method in PYTHON) Tree node generates a decision tree that allows to predict or classify target. The method uses recursive partitioning to divide the training dataset into some similar segments.

A generalized linear mixed model (GLMM) is another supervised method which is used a lot for finding the logical relationship between input data and output. Generalized linear mixed models in PYTHON include different of regression models such as simple linear regression to complex multiple models for non-normal longitudinal data.

Neural networks work similar of the way of human`s nervous system works. The basic units are neurons, which are typically organized into input, middle and output layers.  It operates by modelling of the huge amount of interconnected processing units that resemble abstract versions of neurons. The ANN workflow are arranged in various layers; **input layer** includes the input data, **hidden layers** which cover different number of neurons, and **output layer** contain the target or outcome neuron. ANN approach has ability to predict more than one targets., with a unit or units representing the target field(s). The units are connected with varying connection strengths (or weights). Input data are presented to the first layer, and values are propagated from each neuron to every neuron in the next layer. Eventually, a result is delivered from the output layer.

**Association approach**

*Association models* tries to explore patterns in input dataset which are associated with one or more other entities. This method builds rule sets that explain these relationships. Apriori and Carma can be the algorithm of this model. One other type of association model is a sequence detection model, which finds sequential patterns in time-structured data.

Association method are applied mostly for predicting more than one targets—this method is very useful to predict more than one air pollutants such as CO2, SO2, Co and other air pollution elements. Association models associate a particular conclusion (such as the decision to buy something) with a set of conditions.

**Segmentation approach**

*Segmentation models* classifies the data into different clusters, of records that have similar patterns of input fields. This method could be one of the unsupervised method which focus only on input data similarities. Kohonen networks, K-Means clustering, two-step clustering and anomaly detection are some well-known example of this methodology.

Segmentation or clustering approach are useful when the target or outcome is not clear. Clustering models concentrate on recognizing segments of similar input data and labeling them based on the segments to which they belong. Accuracy of the input data is very important because the result of this method is not right or wrong outcome.

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

In this study, based on input nature different Data mining / Machine Learning Techniques such as Multiple Regression, Decision Tree, and Support Vector Machine can be used to find the best method to generalize the target. However, based on good relationship between PM25 and some input we can use some kind of approaches. At this stage supervised method can help to predict the PM25 by using input data. One set of input data include both weather data (Temperature, pressure, win speed…) and air pollutants elements and the other input set is only climate data. So it seems for prediction of PM25 by using second input data we need to test different algorithms in supervised method. But based on previous section description the Regression and Artificial Neural Network could be the one of the optimized approach.

# Data-mining algorithm(s) selection

## Conduct exploratory analysis and discuss

With a huge number of data in all stations (420,000 rows data), we are now interested in finding out the trends and predicting them. The data extraction techniques help in converting the raw data into useful knowledge. An algorithm in machine learning for massive dataset is a kind of calculations that makes a model from row data. For making a proper model, one algorithm first analyzes the data which was provided, looking for specific types of patterns or trends. One appropriate algorithm utilizes the consequences of this analysis over many iterations to find the best coefficient for creating the mining model. The coefficients are then applied to all data set to extract actionable model and detailed statistics.

However, selecting the optimized algorithm to utilize for some analytical task can be a critical. by using various algorithms to perform the same business task, each algorithm generates a diverse result, and some algorithms can generate more than one type of result. For example, Decision Trees algorithm is not only for prediction, but also as a way to reduce the number of columns in a dataset, because the decision tree can identify columns that do not influence the final model.

**Linear Regression algorithm** as a variation of Decision Trees algorithm that can estimate a linear relationship between an input and target variable, and then use that relationship for prediction. The relationship is a linear equation which represents a series of data. For example, the line in the following diagram is the best possible linear representation of the data.

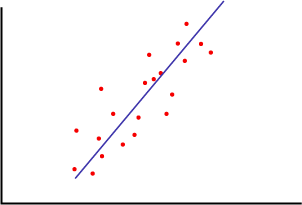


Fig. 51 linear representation of the data

Each data point in the diagram has an error associated with its distance from the regression line. The coefficients a and b in the regression equation adjust the angle and location of the regression line. the regression equation would be adjusting a and b until the sum of the errors that are associated with all the points reaches its minimum.

multiple and Polynomial regression are another regression method can be used for predicting the parameters. However, linear regression is a useful and famous method for modeling a response to a change in some underlying factor.

**Decision tree** as a one of the supervised method (classification technique) is one of the more famous method among other Regression methodology. The algorithm is a method that is called as recursive binary splitting. Beginning from the top, all observations are in the same region. Then a set of questions are asked at each node for dividing of different observations into sub regions based on the answer of those questions. The algorithm stops when the threshold stopping criterion is met.

Support vector classifier divides a test observation depending on which of a hyperplane it lies. It let some observations to be on the wrong side of the margin or hyperplane. It is the solution to the following optimization problem. An example of Support Vector Classifier is as a below figure.

When **Support Vector Machine** emerged to [ML](https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/) methodology, SVM was known as support vector networks as a one of the supervised learning models that accompanies with associated learning algorithms which then investigates data that are utilized for the analysis of regression and classification. This algorithm is a representation of the examples as points in space, that are further mapped so that the examples of the separate categories are then classified by a obvious blank data that is should to be as wide range as possible.

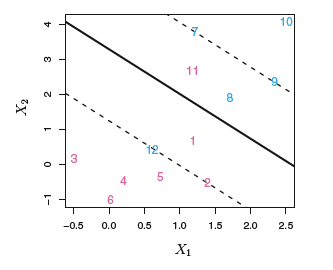


Fig. 52 SVM approaches (James, G. , 2015)

**Neural Network algorithm** is one of the well-known and flexible neural network architecture for ML. The algorithm operates by testing each input data against each possible state of the predictable attribute, and estimating probabilities for each combination based on the training data. All data after data cleaning and dealing with missing data must be divided to 2 or 3 categories such as Training set, Testing set and validation set data. One of the strangest point of this algorithm is application in both classification or regression tasks and to predict a target based on some input. A neural network can also be used for association analysis. Even also there is possibility to predict more than one target by using ANN (Microsoft Neural Network Algorithm, 2018).

## Select data-mining algorithms based on discussion

Based on the objective of this project PM25 as a dangerous air pollutant must be predicted by using combination of climate and pollutants data. Based on Fig. 53 and

Fig. 54, there is high and low correlation between PM25 and PM10 and O3. On the other side correlation between PM25 and climate parameters are complicated and it needs to use algorithm to find this relationship between target and input effectively (Fig. 55 and Fig. 56). Different regression algorithms are effective approach for reduce the uncertainty and predict the PM25 as a target. However, by using the Evaluation methodology, the best algorithm can be recognized well. Also, PM25 as continuous data which change from 0 to 1000 so using some algorithm such as logistic regression is not recommended for this kind of data.

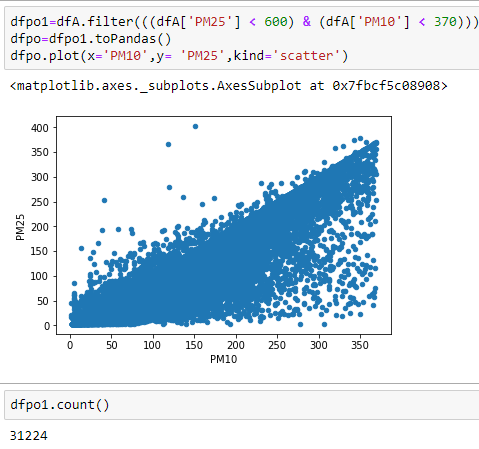


Fig. 53 Correlation between PM25 and PM10 as a another pollutants

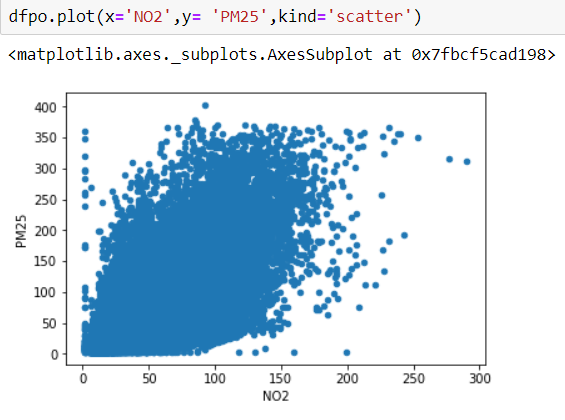
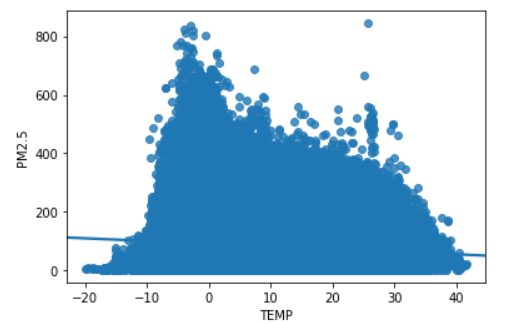
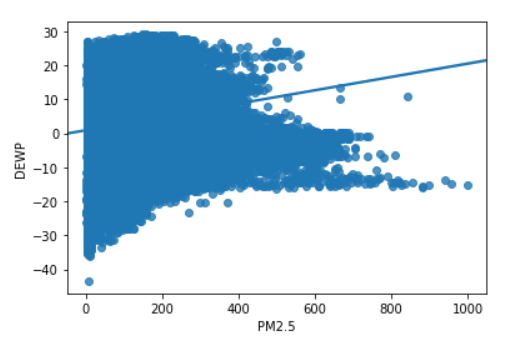


Fig. 54 Correlation between PM25 and NO2 as a another pollutants



*Fig. 55 Correlation between PM25 and TEMP as a climate data*



*Fig. 56 Correlation between PM25 and DEWP as climate data*

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

* **Linear Regression mechanism:**

This section, we use PySpark to implement simple linear regression. Then, we split our data into training and test sets, create a model using training set, evaluate your model using test set, and finally use model to predict unknown value. Simple Linear Regression is a method to help us understand the relationship between two variables:

* The independent variable (X)
* The dependent variable (that we want to predict) (Y)

The result of Linear Regression is a linear function that predicts the target variable as a function of the input variable. Based on very good correlation between PM25 and PM10 we are able to use this method to predict PM25 effectively. Fig. 57 shows good correlation between two variable and we can use simple linear model based of this data.

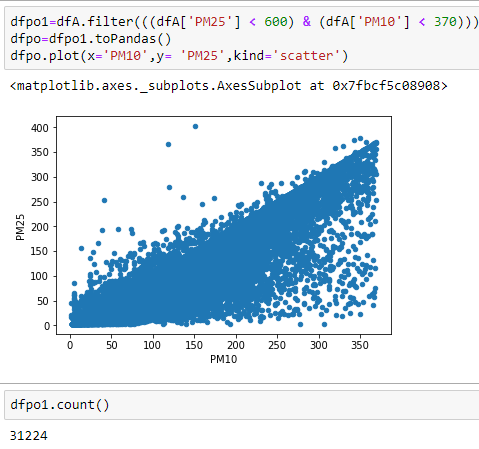
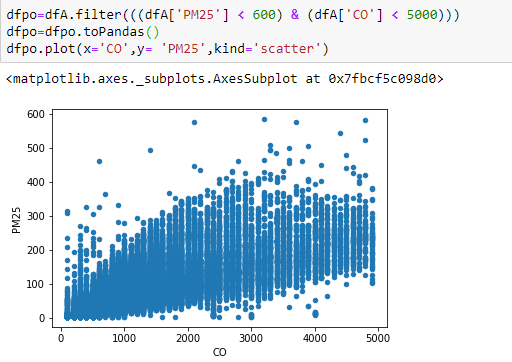


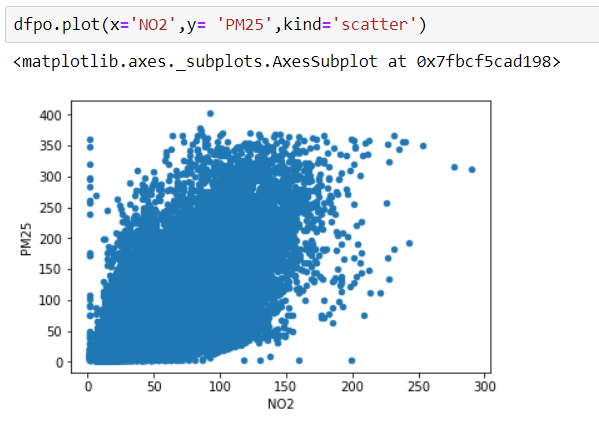
Fig. 57Residual plot between PM25 and PM10

* **Multiple Linear Regression**

If we want to use more variables in our model (such as NO2, CO, 03, TEMP, PRESS…) to predict PM25, we can use Multiple Linear Regression. Multiple Linear Regression is very similar to Simple Linear Regression, but this method is used to explain the relationship between one continuous response (dependent) variable and two or more predictor (independent) variables. We will explain the structure by using several predictor variables, but these results can generalize to any integer. Based on the Fig. 58 and Fig. 59 there is a suitable correlation between PM25, CO AND NO2. So we can run new model as a multi linear regression approach to obtain the result



*Fig. 58:correlation between CO and PM25*

**

*Fig. 59: correlation between NO2 and PM25*

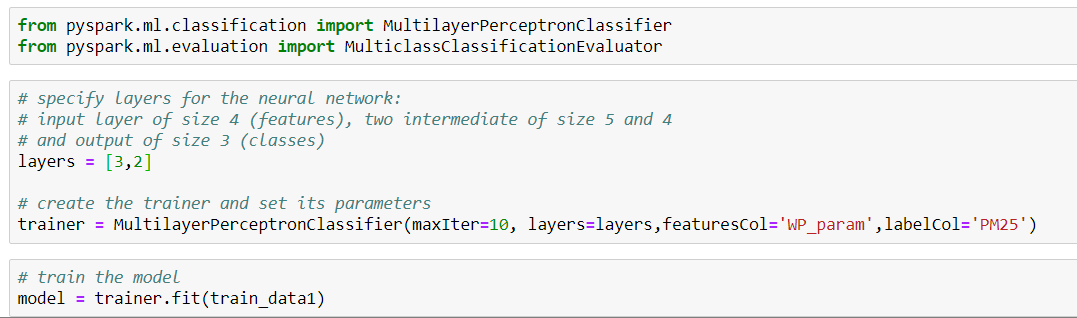
## Generalized linear regression

Spark’s GeneralizedLinearRegression interface let for adjustable characterization of GLMs which able to apply for different prediction issues contain linear regression, Poisson regression, logistic regression, and others. Currently in spark.ml, only a subset of the exponential family distributions are supported and they are listed [below](https://spark.apache.org/docs/2.1.1/ml-classification-regression.html#available-families).

GLMs require exponential family distributions that can be written in their “canonical” or “natural” form, aka [natural exponential family distributions](https://en.wikipedia.org/wiki/Natural_exponential_family). Spark’s generalized linear regression interface provides summary statistics for diagnosing the fit of GLM models, including residuals, p-values. However different parameters combination of PM10, NO2, SO2 and CO were used in different model of GLM in this study.

## Multilayer perceptron classifier

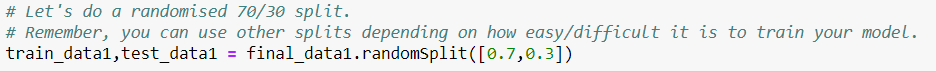
Multilayer perceptron classifier (MLPC) is a classifier based on the [feedforward artificial neural network](https://en.wikipedia.org/wiki/Feedforward_neural_network). MLPC include of multiple layers of nodes. Each layer is fully connected to the next layer in the network. Nodes in the input layer represent the input data. All other nodes map inputs to outputs by a linear combination of the inputs with the node’s weights and bias b and applying an activation function. Below code program was used in this study. All air pollutant material were used as an first layer or input layer.



# Data Mining

## Create and justify test designs

Dataset includes different row data which is classified into one target, the all climate and air pollution data was imported to the first model (**Error! Reference source not found.**) and also the climate data was imported to second model . After running the model probably some irrelevant input data will be skipped to final model. However, in both models, all data was divided to Test, Train data set. Based on them 383590 rows, 77000 rows and 306600 rows are dedicated to test, and train approaches respectively. There is no missing data in all of them and Extreme data mostly was dropped. Meanwhile most of outliers from air pollution data was removed. In addition, Fig. 61 and Fig. 62 indicate the train and test set spec in model 1 and 2 respectively. Model-1 includes all available data but model 2 includes only climate data. Also data sharing between test and train set data must be logical. Data share in training set must be more than test set data. Because in this step all relationship must be built. But based on more investigation, divided all data into 70, 30 or 80,20 could be recommended.



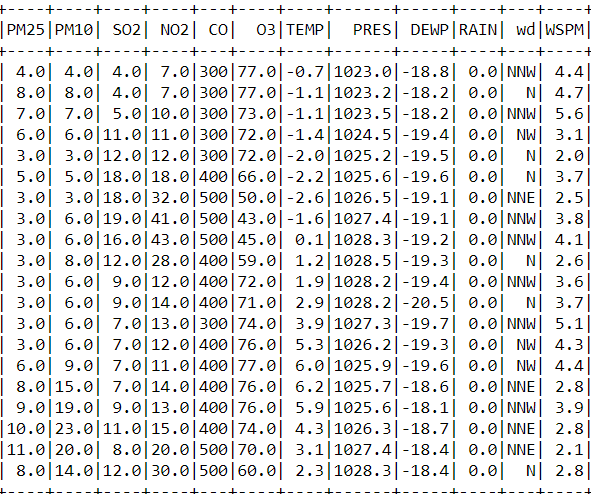


Fig. 60: Train and test data; blue one is model-1 as input data, red one is model-2 and green one is target

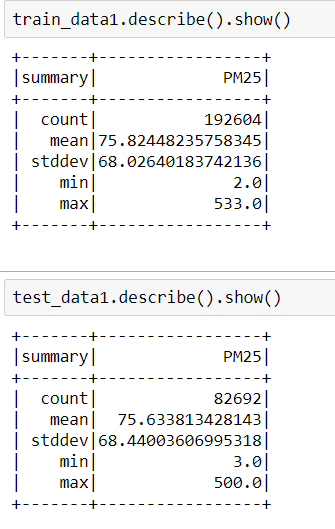


Fig. 61 : Description of train and test set in first model

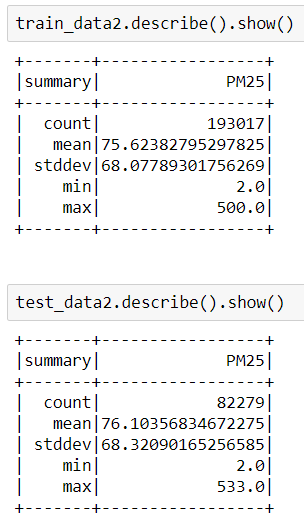


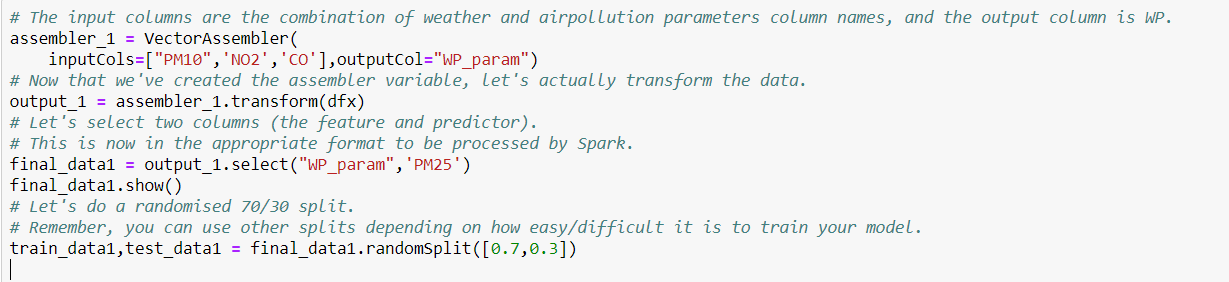
Fig. 62 : Description of train and test set in first model

## Conduct data mining

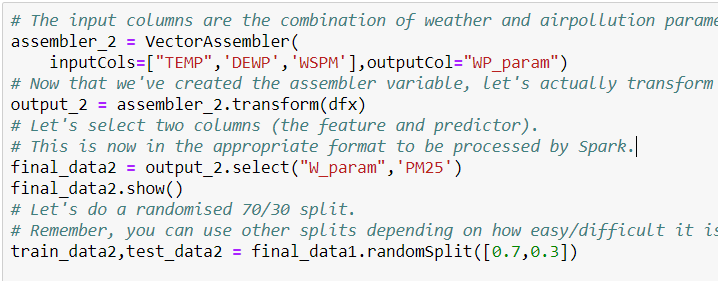
* **Multi Linear Regression**

Based on data mining methods and algorithms which were described in previous chapter, Supervised method was selected. Multi-regression approach was chosen for PM25 predictions. In the first model by using combination of all input, regression algorithms have a very good correlation and it is about 0.82 (**Error! Reference source not found.**). In the second model that I used only climate data, the correlation shows 0.2. in PySpark the data must be assembled and transformed as a unique data we call it features. But here based on below coding 2 kind of data was used as an input the first category include PM10, CO and NO2 which is called WP\_param and second model contain only weather parameters which is called W\_Param. Based on the result only the first model had a suitable result. The transformed data in both models are shown in Fig. 63 and Fig. 64. Also the description of test and train set data in both models are available in Fig. 65 and Fig. 66.

***First model***



***SECOND MODEL***



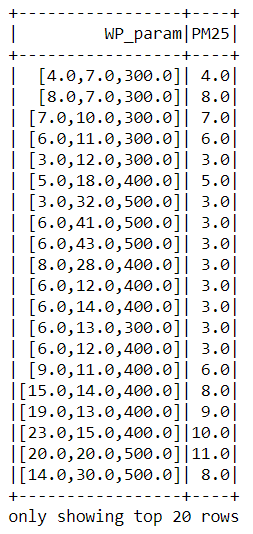


Fig. 63 : Transformed data in first model

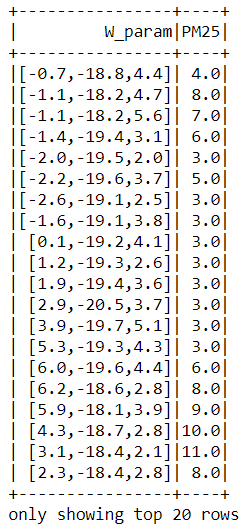


Fig. 64 : Transformed data in second model

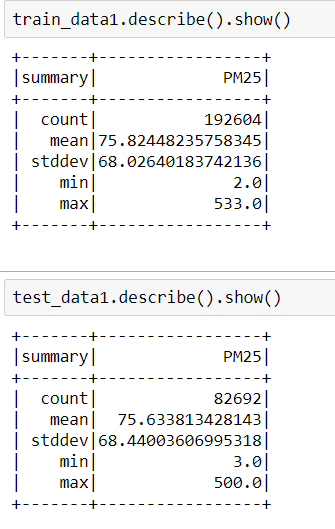


Fig. 65 : Description of train and test set in first model

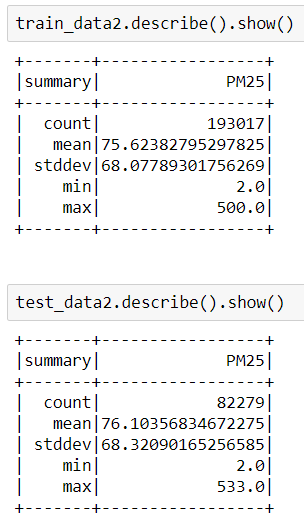
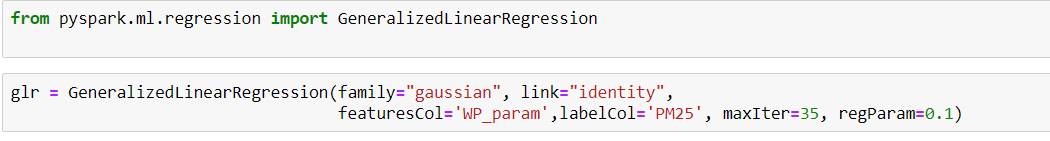


Fig. 66 : Description of train and test set in first model

## Generalized linear regression

In this approach both data models were tested and the code program was used as below.



## Search for patterns

**Pattern\_1**: The first pattern contain all data; PM10, SO2, NO2, CO, O3, TEMP, PRES and DEWP as a pollutant and climate data Also Fig. 67 as a table explains relationship between target and input clearly. Multi Linear regression model works well. There is high correlation between depended and variable data. Green dolor show target, blue and red one indicate pollution and climate input data respectively.

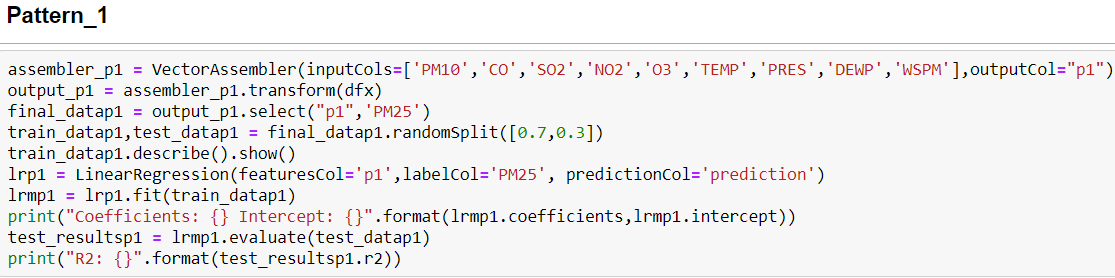




Fig. 67: data pattern 1

**Pattern\_2:** this pattern includes only 4 climate elements; TEMP,DEWP,WSPM and PRESS (Fig. 68). Multi linear regression algorithms which were used. In this pattern only 4 kind of data was used and its relationship with target is clear.

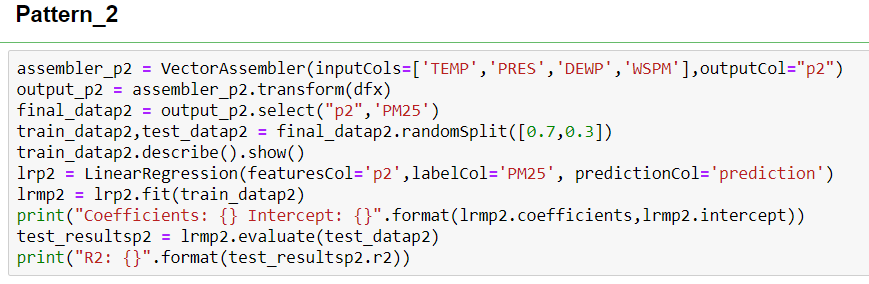


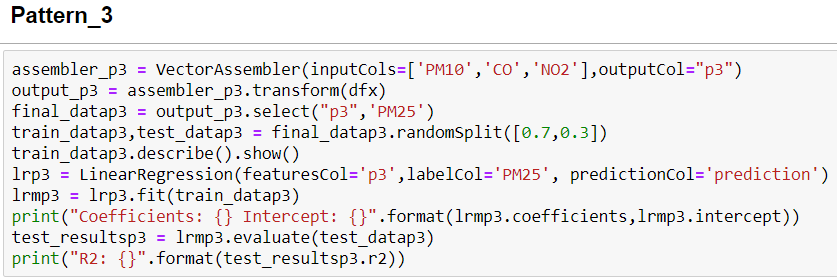


Fig. 68: Data pattern 2

**Pattern\_3:** this pattern includes PM10, CO, and NO2 as in input (Fig. 69). simple regression regression gives us proper correlation among the algorithms which were used in this pattern. In this pattern the relationship with target is clear.



Fig. 69: data pattern 3



**Pattern 4**: this pattern used the data from pattern 1. It includes all air pollution and weather data. But Generalized Linear regression was used for prediction of PM25. different family were tested and Gaussian was selected. Also the relationship of the input data with target is clear.

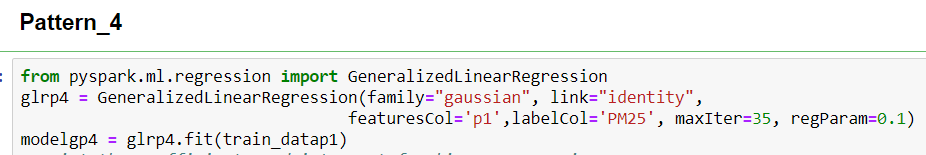




Fig. 70: data pattern 4

**Pattern\_5**: this pattern used the data from pattern 3. It includes all air pollution data. But Generalized Linear regression was used for prediction of PM25. different family were tested and Gaussian was selected. Also the relationship of the input data with target is clear.

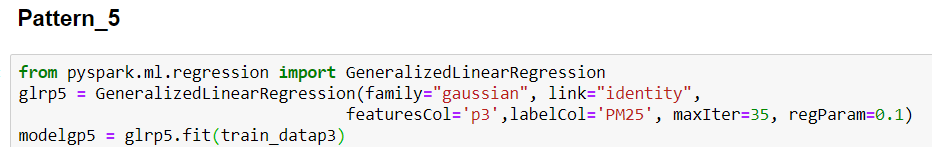




Fig. 71: data pattern 5

# Interpretation

## Study and discuss the mined patterns

In previous section, by using different input 5 pattern have been built. In all approach, Python packages was chosen and different regression algorithm was used.

**Pattern-1**: in this pattern as a complete pattern, all air pollution and climate parameters was used and correlation in all algorithm are very good. Multi Linear Regression method shows the correlation about 0.86 between original and predicted PM25. The error in (training, testing and validation) are very small and ignorable. This approach is the best among all patterns.

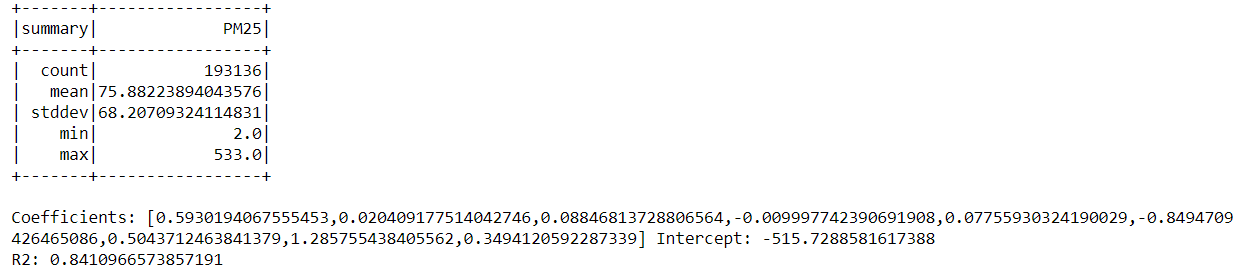


Fig. 72 Multi Linear regression algorithms which have the best correlation with PM25 in model 1 and pattern-1

**Pattern-2**: in this pattern, only climate data was imported to Multi linear regression model, however based on the evaluation correlation in all algorithm is not good. Multi Linear Regression method shows the correlation about 0.18 between original and predicted PM25 . The error in (training, testing and validation) is considerable. This approach is not recommended.

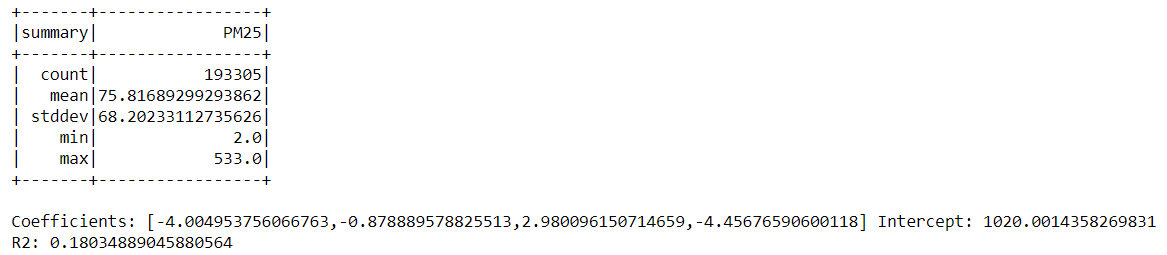


Fig. 73 Three best algorithms which have the best correlation with PM25 in model 1 and pattern 2

**Pattern-3**: in this pattern as a subsidiary of pattern-1, include only air pollution materials as an important input data and correlation in all algorithm are very good. Multi Linear Regression method shows the correlation about 0.82 between original and predicted PM25

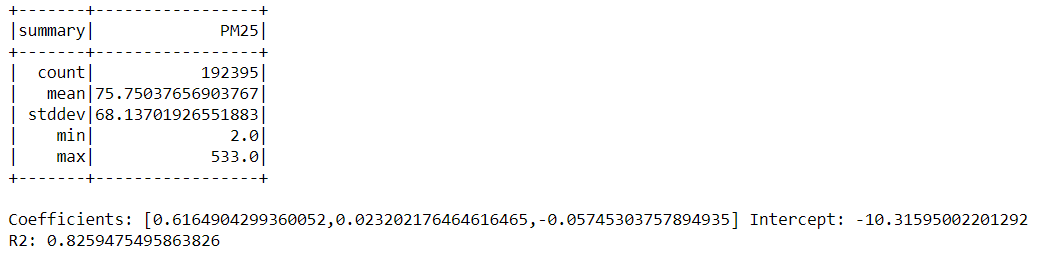


Fig. 74 Multi Linear regression algorithms which have the best correlation with PM25 in pattern-3

**Pattern-4**: in this pattern as a subsidiary of pattern-1, include air pollution and weather data and correlation in all algorithm are very good. GLR method shows the correlation about 0.6 between original and predicted PM25.

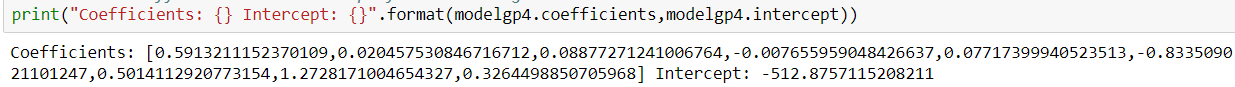


Fig. 75 GLR algorithms which have the best correlation with PM25 in pattern-4

**Pattern-5**: in this pattern includes only pollution data and correlation in all algorithm is suitable. GLR method shows good correlation about 0.62 between original and predicted PM25.

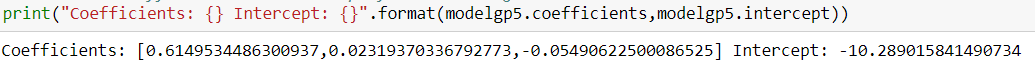


Fig. 76 GLR algorithms which have the best correlation with PM25 in pattern-5

## Visualize the data, results, models, and patterns

In previous chapter all relevant data for each pattern was introduced and based on different kind of data was used to increase the correlation between input and target. Pattern-1, 3 and 5 by using the various air pollution and data show very suitable correlation by the graph. Multi Linear Regression algorithm and GLR as a supervised method was the best approach to simulate PM25 (Fig. 77, Fig. 79 and Fig. 80). However pattern 4 and 5 by using only climate data show moderate to weak correlation between input data and target (Fig. 78and Fig. 81).

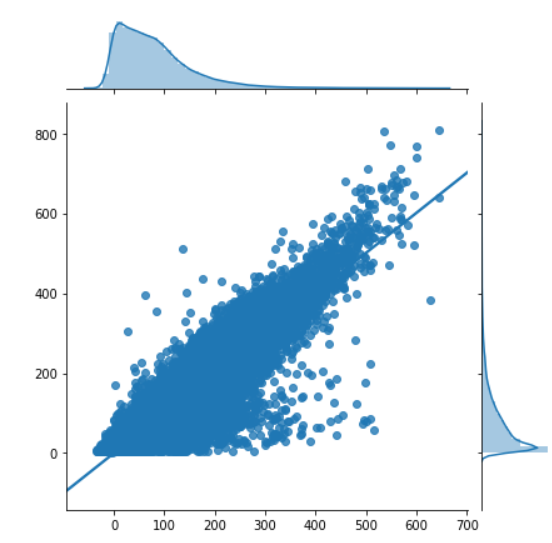


Fig. 77 Correlation between test data and prediction data by multi-regression model in pattern 1

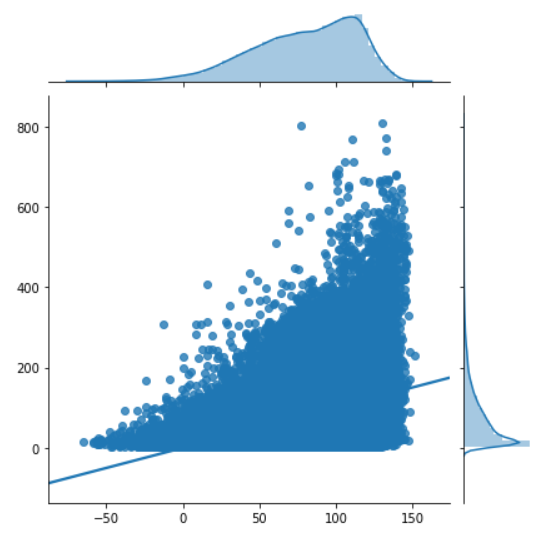


Fig. 78 Correlation between test data and prediction data by multi-regression model in pattern 2

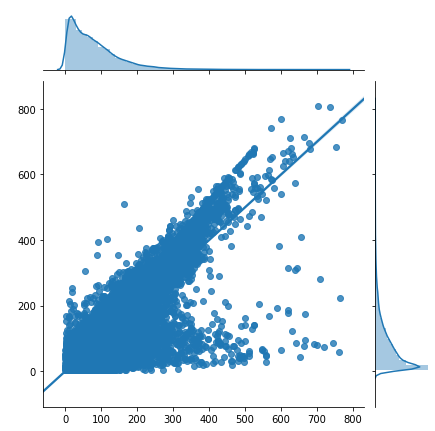


Fig. 79 Correlation between test data and prediction data by simple linear regression model in pattern 3

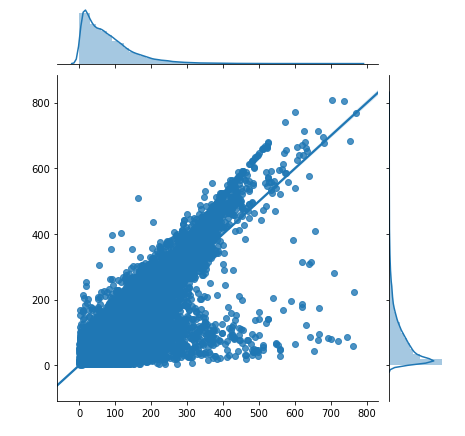


Fig. 80 Correlation between test data and prediction data by GLR method in pattern 4

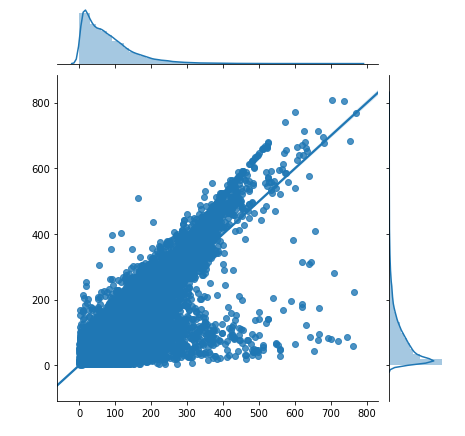


Fig. 81 Correlation between test data and prediction data by GLR method in pattern 5

## Interpret the results, models, and patterns

Pattern-1, 3 and 5 are very important to model the target. In the first one all data was used as an input. the output by using Multi Linear regression algorithm are shown as a Fig. 77, Fig. 79 and Fig. 80. On the other hand, Fig. 78 and Fig. 81 display the correlation between original and predicted PM25 by only climate data. The first pattern has the best correlation but it used different kind of data. But the second pattern, the input is only normal climate data such as temp, pressure, wind speed and prediction of air pollution parameters such as PM25 by correlation about 0.2 is not recommended. At the end only pattern 1, 3 and 4 are suitable with high correlation.

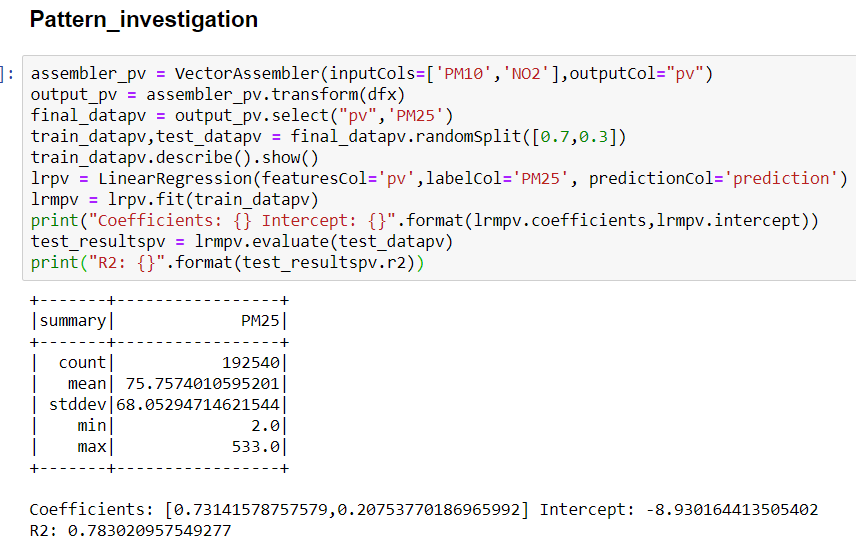
## Assess and evaluate results, models, and patterns

The tables below categorize pattern 1 and 5 together and based on that Multi Linear Regression by using all air pollution and climate data works better than the others (pattern 1 and model 1). However, based on the result in previous chapter, PM25 could be predicted only by climate data with correlation about 0.2. this achievement was very useful and beneficial for all organization to predict the air pollution by some usual weather and air pollution parameters which can be measured easily.

|  |  |  |  |
| --- | --- | --- | --- |
| Patterns | method | data | correlation |
| Pattern\_1 | Multi Linear Regression | Pollution & climate data | 0.84 |
| Pattern\_2 | Multi Linear Regression | climate data | 0.18 |
| Pattern\_3 | Multi Linear Regression | pollution | 0.82 |
| Pattern\_4 | Generalized Linear regression | Pollution & climate data | 0.6 |
| Pattern\_5 | Generalized Linear regression | pollution | 0.62 |

## Iterate prior steps (1 – 7) as required

In the pattern-1 and 3 used Multi Linear regression. Determining the correct input is very important to reduce the error. Different combination of input data was tested. However, the result shows by using PM10 in every combination the accuracy will be increased so the best combination of input data is including PM10. This investigation showed us the best final and suitable input data is combination of climate and air pollutants together.



# RefErence:

Chan, C.K. & Yao, X. (2008). *Air pollution in megacities in China.* *42*, 1–42.

Chen, W.Y. & Xu, R.N. (2010). *Clean coal technology development in China.* Energy Policy, *38*, 2123–2130.

Dayan, P. (1999). Unsupervised Learning. The MIT Encyclopedia of the Cognitive Science.

Du, X., Kong & Q., Ge & W., Zhang, S. & Fu, L. (2010). *Characterization of personal exposure concentration of particles for adults and children exposed to high ambient concentrations in Beijing, China*. Journal of Environmental Sciences, 22(11), 1757-1764. DOI:10.1016/s1001-0742(09)60316-8

Elminir, H.K. (2005). *Dependence of urban air pollutants on meteorology*. Sci. Total Environ, 350, 225–237.

Holloway, T. & Spak, S.N. & Barker, D. & Bretl, M. & Moberg, C. & Hayhoe, K. & Van Dorn, J. & Wuebbles, D. (2008). *Change in ozone air pollution over Chicago associated with global climate change.* J. Geophys. Res. Atmos. 113,DOI:10.1029/2007JD009775.

Knowledge Center *.com/support/knowledgecenter/en/SS3RA7\_sub/modeler\_mainhelp\_client\_ddita/clementine/understanding\_modeltypes.html*

James, G. (2015). An introduction to statistical learning: With applications in R.(pp. 337-372) New York: Springer.

Kurt, A. & Oktay, A.B.(2010). *Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks*. Expert Syst. 37, 7986–7992.

Li, L.J. & Wang, Y. & Zhang, Q. & Li, J.X. & Yang, X.G. & Jin, J. (2008). *Wheat straw burning and its associated impacts on Beijing air quality*. Sci. China Ser*. 51*, 403–414.

Microsoft Neural Network Algorithm, (2018). <https://docs.microsoft.com/en-us/analysis-services/data-mining/microsoft-neural-network-algorithm?view=asallproducts-allversions>

Rohde, R. A., Muller, R. A. (2015). Air Pollution in China: Mapping of Concentrations and Sources. Plos One, 10(8). doi:10.1371/journal.pone.0135749

Samet, J.M.& Zeger, S.L. & Dominici, F. & Curriero, F. & Coursac, I. & Dockery, D.W. & Schwartz, J. & Zanobetti, A.(2000). *The national morbidity, mortality, and air pollution*

*study. Part II: Morbidity and mortality from air pollution in the United States*. Res. Rep. Health Eff. Inst. 94, 5–79.

Sun, Y.L. & Zhuang, G.S. & Tang, A.H. & Wang, Y. & An, Z.S. (2006). *Chemical characteristics of PM25 and PM10 in haze-fog episodes in Beijing*. *Environ. Sci. Technol.* *40*, 3148–3155.

Wang, Q.Q. & Shao, M. & Liu, Y. & William, K. & Paul, G. & Li, X.H. & Liu, Y.A. & Lu, S.H. (2007) *Impact of biomass burning on urban air quality estimated by organic tracers: Guangzhou and Beijing as cases.* Atmos. Environ, 41, 8380–8390.

Wei Chen. (2015). Air Quality of Beijing and Impacts of the New Ambient Air Quality Standard. 6, 1243-1258. DOI:10.3390/atmos6081243

World Health Organization, (2020). Air Pollution. Retrieved from <https://www.who.int/health-topics/air-pollution#tab=tab_2>

Yuan, Z. & Zhou, X. & Yang, T. & Tamerius, J. & Mantilla, R. (2017). *Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study*. In Proceedings of the 6th International Workshop on Urban Computing, Halifax, NS, Canada.

Zheng, Y. & Liu, F. & Hsieh, H.-P. (2013). *When urban air quality inference meets big data*. In Proceedings of the19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA,

"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. (See: [https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.htmlLinks to an external site.](https://www.auckland.ac.nz/en/students/forms-policies-and-guidelines/student-policies-and-guidelines/academic-integrity-copyright.html)).

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."