Analysis of PM2.5 Pollution

ITERATION 2 ISAS

Name: Ziteng Li

Student ID: 733922565

Table of Content

1. Business or Situation understanding.

1.1. Identify the objectives of the business/situation

1.1.1. Background

Air pollution may be the significant driver of NCDs(non-communicable diseases), including lung cancer, heart attacks, strokes, and accelerating climate change. Air pollution could make a global death increase of 7,000,000 per year. Global must try to prevent diseases and reduce air pollution.

1.1.2. Associate with sustainable problems

The target of 17 Sustainable Development Goals is to stop extreme poverty and create a healthy and sustainable world by 2030. There are five goals associated with the severe air pollution problems, including Goal 3 (Good Health and Well-being), Goal 7 (Affordable and Clean Energy), Goal 11 (Sustainable Cities and Communities), Goal 13 (Climate Action), and Goal 17 (Partnerships For The Goals).

1.1.3. Defining business objectives

To deal with the high concentration of PM2.5 in the air is harmful to human beings, we should understand what PM2.5 is. “PM (Particulate matter) is a standard proxy indicator for air pollution. It affects more people than any other pollutant. PM’s significant components are sulfate, nitrates, ammonia, sodium chloride, black carbon, mineral dust, and water” (Team Airveda, 2017). It consists of a complex mixture of solid and liquid particles of organic and inorganic substances suspended in the air. “While particles with a diameter of 10 microns or less, can penetrate and lodge deep inside the lungs, the even more health-damaging particles are those with a diameter of 2.5 microns or less” (WHO, 2018).

- Concentration of PM2.5
- Reduce PM2.5
- Predict PM2.5
- Other air pollution

1.1.4. Business Success Criteria

Finding the main reasons for high PM2.5 concentration, use methods and models to predict the relative inferiority of PM2.5 concentration, find out solutions to reduce PM2.5 pollution by human activities.

1.2. Assess the situation

1.2.1. General situation

While PM2.5 impacts everyone, people with breathing and heart problems, children, and the elderly are most sensitive. Due to particulate matter’s omnipresence, ambient particulate matter has proven to be more potent than alcohol and diabetes.

Exposure to PM2.5 has multiple short terms and long-term health impacts. The short term includes irritation in the eyes, nose, and throat, coughing, sneezing, and shortness of breath. Prolonged exposure to PM2.5 can cause permanent respiratory problems such as asthma, chronic bronchitis, and heart disease.

In low- and middle- income countries, exposure to pollutants in and around homes from the household combustion of polluting fuels on open fires or traditional stoves for cooking, heating and lighting further increases the risk for air pollution-related diseases, including acute lower respiratory infections, cardiovascular disease, chronic obstructive pulmonary disease, and lung cancer.

1.2.2. Area evaluation

There are severe risks to health not only from exposure to PM but also from exposure to ozone (O3), nitrogen dioxide (NO2), carbon monoxide (CO) and sulfur dioxide (SO2). As with PM, concentrations are often highest, mainly in low- and middle-income countries. "Ozone is a significant factor in asthma morbidity and mortality. At the same time, nitrogen dioxide, and sulfur dioxide also can play a role in asthma, bronchial symptoms, lung inflammation, and reduced lung function" (WHO, 2018).

There are currently more than 30,000 available air quality monitoring stations in the world, out of which more than 12,000 are published on the World Air Quality Index project. In this dataset, the data came from 11 different monitoring stations in Beijing, mainly focus on the air quality, including different air types mentioned above.

1.2.2. Requirements

The data and project results are security and legal restrictions. For example, the dataset, Kaggle, is a data modeling and data analysis competition platform. There were lots of reliable dada set on the Kaggle website and the UCI-Machine Learning Repository website. In this project, the data set was chosen from the Kaggle website. It is a useful dataset for evaluating and predicting the concentration of PM2.5.

1.2.4. Assumptions

In this project, the dataset was published in 2018. It is not hard to find the dataset used by many reports or articles during the research. There is no need for data quality assumptions. The project result would be analyzed by the SPSS Modeler and would report in later steps.

1.2.5. Constraints

This project could only predict the concentration of PM2.5 by the collected weather index. There are many uncertainties such as vehicle restrictions, factory closures, increase and decrease in human carbon travel, and another situation that cannot be recorded are not considered when predicting the concentration of PM2.5.

1.3. Determine data mining objectives

1.3.1. Data Mining Goals

The focus of this project is to study the changes in the concentration of PM2.5, conduct data mining on the data to predict the occurrence of PM2.5 in the later period effectively, and the government researches the exhaust emissions industry and automobile exhaust that lead to an increase in PM2.5. Control and help individuals prevent air pollution from harming the human body and reduce the occurrence of diseases.

1.3.2. Data Mining Success Criteria

Finding the main reasons of high PM2.5 concentration base on the dataset. Use methods and models in SPSS Modeler to predict the relative inferiority of PM2.5 concentration. Find out solutions to reduce PM2.5 pollution by human activities.

The analysis results can be used as a reference to improve air quality and achieve the ultimate goal of reducing PM2.5 concentration by reducing the emission of these related compounds in the air.

1.4. Produce a project plan

The overview plan for the study of this project is as shown in the table below.

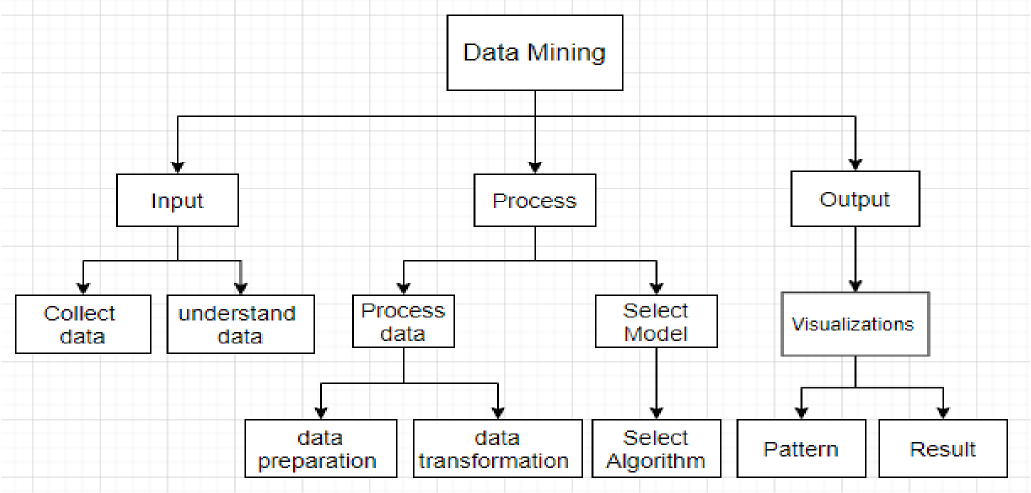
Table 1. project plan overview

Phase	Time	Resources
Business/Situation understanding	Two days (July 31st - Aug.1st)	All analysts (Finding dataset, accessing the situation)
Data understanding	Five days (Aug.2nd - Aug.6th)	All analysts (Research)
Data preparation	One day (Aug.7th)	All analysts, SPSS Modeler, Research
Data Transformation	Two days (Aug.8th- Aug.9th)	SPSS Modeler, Research
Data Mining Method Selection	3days (Aug.10th- Aug.13th)	SPSS Modeler, Research
Data Mining Algorithm Selection	Four days (Aug.14th - Aug.18th)	SPSS Modeler, Research
Data Mining	Four days (Aug.19th- Aug.23rd)	SPSS Modeler, Research
Interpretation	Four days (Aug.24th – Aug.28th)	SPSS Modeler, Research

In this project, the primary technique is the SPSS Modeler. It can be used to analyze and process the data. It also contains various models that allow users to enter data sets and perform multiple operations on the data by displaying test results and diagrams.

1.4.1. WBS and Gantt chart

The process of this project is based on the Gantt chart, as shown below.



2. Data Understanding.

2.1. Collect initial data

Asian regions are the typical areas that suffered from PM2.5 pollution, especially in Beijing, China. In order to find the analytical and usable data for PM2.5 pollution, the data set of this project has been found from Kaggle (https://www.kaggle.com/sid321axn/beijing-multisite-airquality-data-set), and UCI-Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data#). However, there are 12 different regions have been tested in this data set, research has been done for finding a moderate area (Changping in Beijing) in this data set to make the results of this project is significant and usable.

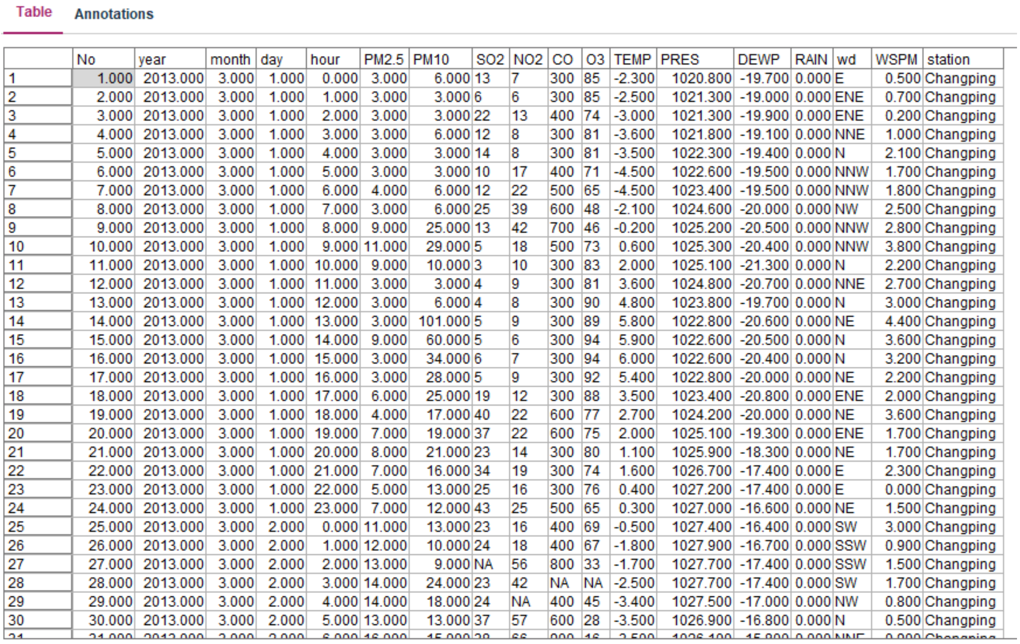
2.2. Describe the data

For the whole data set, there are 12 Beijing regions under-tested, respectively.

The data set was recorded every hour while testing the air quality in the Beijing Changping district, which was in total 35065 hours' air-quality report.

There are 18 columns, 35065 rows in the data set, containing two data types: numeric and string.

Figure 1. Examples of data in the dataset – Evidence from SPSS Modeler



	No	year	month	day	hour	PM2.5	PM10	SO2	NO2	CO	O3	TEMP	PRES	DEWP	RAIN	wd	WSPM	station
1	1.000	2013.000	3.000	1.000	0.000	3.000	6.000	13	7	300	85	-2.300	1020.800	-19.700	0.000	E	0.500	Changping
2	2.000	2013.000	3.000	1.000	1.000	3.000	3.000	6	6	300	85	-2.500	1021.300	-19.000	0.000	ENE	0.700	Changping
3	3.000	2013.000	3.000	1.000	2.000	3.000	3.000	22	13	400	74	-3.000	1021.300	-19.900	0.000	ENE	0.200	Changping
4	4.000	2013.000	3.000	1.000	3.000	3.000	6.000	12	8	300	81	-3.600	1021.800	-19.100	0.000	NNE	1.000	Changping
5	5.000	2013.000	3.000	1.000	4.000	3.000	3.000	14	8	300	81	-3.500	1022.300	-19.400	0.000	N	2.100	Changping
6	6.000	2013.000	3.000	1.000	5.000	3.000	3.000	10	17	400	71	-4.500	1022.600	-19.500	0.000	NNW	1.700	Changping
7	7.000	2013.000	3.000	1.000	6.000	4.000	6.000	12	22	500	65	-4.500	1023.400	-19.500	0.000	NNW	1.800	Changping
8	8.000	2013.000	3.000	1.000	7.000	3.000	6.000	25	39	600	48	-2.100	1024.600	-20.000	0.000	NW	2.500	Changping
9	9.000	2013.000	3.000	1.000	8.000	9.000	25.000	13	42	700	46	-0.200	1025.200	-20.500	0.000	NNW	2.800	Changping
10	10.000	2013.000	3.000	1.000	9.000	11.000	29.000	5	18	500	73	0.600	1025.300	-20.400	0.000	NNW	3.800	Changping
11	11.000	2013.000	3.000	1.000	10.000	9.000	10.000	3	10	300	83	2.000	1025.100	-21.300	0.000	N	2.200	Changping
12	12.000	2013.000	3.000	1.000	11.000	3.000	3.000	4	9	300	81	3.600	1024.800	-20.700	0.000	NNE	2.700	Changping
13	13.000	2013.000	3.000	1.000	12.000	3.000	6.000	4	8	300	90	4.800	1023.800	-19.700	0.000	N	3.000	Changping
14	14.000	2013.000	3.000	1.000	13.000	3.000	101.000	5	9	300	89	5.800	1022.800	-20.600	0.000	NE	4.400	Changping
15	15.000	2013.000	3.000	1.000	14.000	9.000	60.000	5	6	300	94	5.900	1022.600	-20.500	0.000	N	3.600	Changping
16	16.000	2013.000	3.000	1.000	15.000	3.000	34.000	6	7	300	94	6.000	1022.600	-20.400	0.000	N	3.200	Changping
17	17.000	2013.000	3.000	1.000	16.000	3.000	28.000	5	9	300	92	5.400	1022.800	-20.000	0.000	NE	2.200	Changping
18	18.000	2013.000	3.000	1.000	17.000	6.000	25.000	19	12	300	88	3.500	1023.400	-20.800	0.000	ENE	2.000	Changping
19	19.000	2013.000	3.000	1.000	18.000	4.000	17.000	40	22	600	77	2.700	1024.200	-20.000	0.000	NE	3.600	Changping
20	20.000	2013.000	3.000	1.000	19.000	7.000	19.000	37	22	600	75	2.000	1025.100	-19.300	0.000	ENE	1.700	Changping
21	21.000	2013.000	3.000	1.000	20.000	8.000	21.000	23	14	300	80	1.100	1025.900	-18.300	0.000	NE	1.700	Changping
22	22.000	2013.000	3.000	1.000	21.000	7.000	16.000	34	19	300	74	1.600	1026.700	-17.400	0.000	E	2.300	Changping
23	23.000	2013.000	3.000	1.000	22.000	5.000	13.000	25	16	300	76	0.400	1027.200	-17.400	0.000	E	0.000	Changping
24	24.000	2013.000	3.000	1.000	23.000	7.000	12.000	43	25	500	65	0.300	1027.000	-16.600	0.000	NE	1.500	Changping
25	25.000	2013.000	3.000	2.000	0.000	11.000	13.000	23	16	400	69	-0.500	1027.400	-16.400	0.000	SW	3.000	Changping
26	26.000	2013.000	3.000	2.000	1.000	12.000	10.000	24	18	400	67	-1.800	1027.900	-16.700	0.000	SSW	0.900	Changping
27	27.000	2013.000	3.000	2.000	2.000	13.000	9.000	NA	56	800	33	-1.700	1027.700	-17.400	0.000	SSW	1.500	Changping
28	28.000	2013.000	3.000	2.000	3.000	14.000	24.000	23	42	NA	NA	-2.500	1027.700	-17.400	0.000	SW	1.700	Changping
29	29.000	2013.000	3.000	2.000	4.000	14.000	18.000	24	NA	400	45	-3.400	1027.500	-17.000	0.000	NW	0.800	Changping
30	30.000	2013.000	3.000	2.000	5.000	13.000	13.000	37	57	600	28	-3.500	1026.900	-16.800	0.000	N	0.500	Changping

Table 2. Data details – Evidence from the dataset

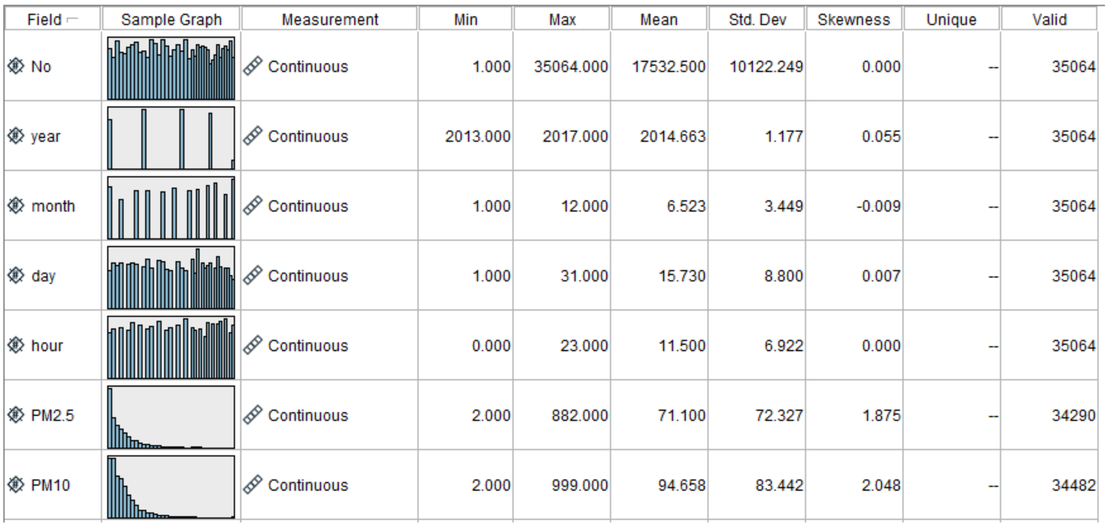
Data Name	Unit	Data Type	Description
NO		Numeric	Rows of data
Year		Numeric	The observation time (year)
Month		Numeric	The observation time (month)
Day		Numeric	The observation time (day)

Hour		Numeric	The observation time (hour)
PM2.5	ug/m^3	Numeric	PM2.5 (Particulate matter) concentration
PM10	ug/m^3	Numeric	PM10 (Particulate matter) concentration
SO2	ug/m^3	Numeric	SO2 (Sulfur dioxide) concentration
NO2	ug/m^3	Numeric	NO2 (Nitrogen dioxide) concentration
CO	ug/m^3	Numeric	CO (Carbon monoxide) concentration
O3	ug/m^3	Numeric	O3 (Ozone) concentration
TEMP	degree Celsius	Numeric	Hourly temperature in particular area
PRES	hPa	Numeric	Pressure
DEWP	Celsius	Numeric	Dew point temperature
RAIN	mm	Numeric	Precipitation
WSPM	m/s	Numeric	Wind speed
wd	N S E W	String	Wind direction
station		String	Air-quality monitoring site (Changping)

2.3. Explore the data

This data set includes hourly air pollutants data from 12 nationally controlled air-quality monitoring sites. The air-quality data are from the Beijing Municipal Environmental Monitoring Center. The meteorological data in each air-quality area are matched with the nearest weather station from the China Meteorological Administration. The period is from March 1st, 2013 to February 28th, 2017.

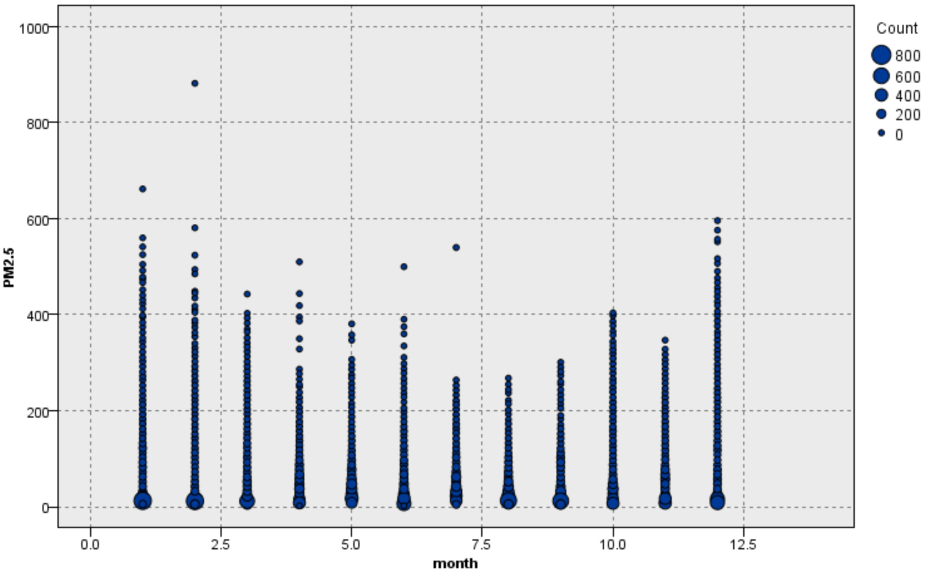
Figure 2. some fields of the data in SPSS Modeler



Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
No		Continuous	1.000	35064.000	17532.500	10122.249	0.000	--	35064
year		Continuous	2013.000	2017.000	2014.663	1.177	0.055	--	35064
month		Continuous	1.000	12.000	6.523	3.449	-0.009	--	35064
day		Continuous	1.000	31.000	15.730	8.800	0.007	--	35064
hour		Continuous	0.000	23.000	11.500	6.922	0.000	--	35064
PM2.5		Continuous	2.000	882.000	71.100	72.327	1.875	--	34290
PM10		Continuous	2.000	999.000	94.658	83.442	2.048	--	34482

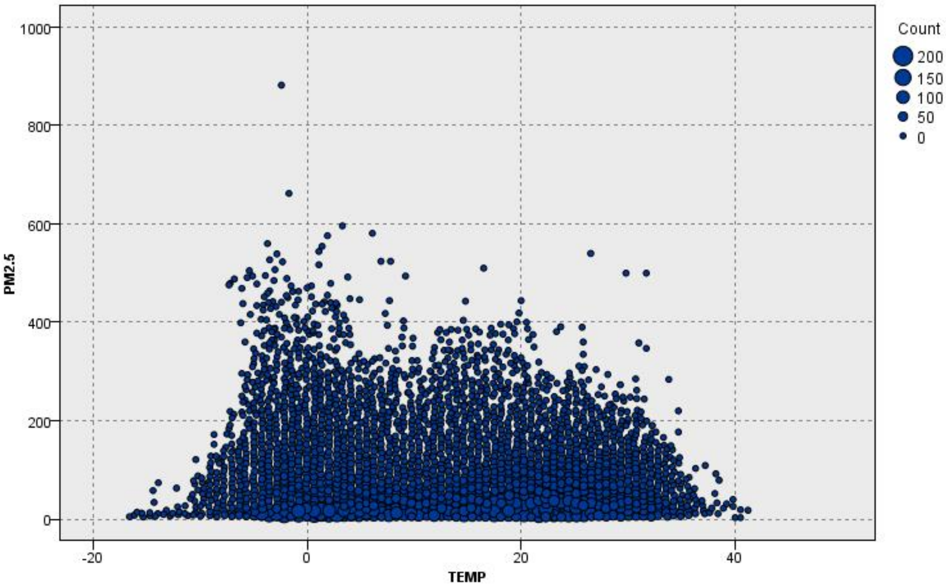
The charts below show the distribution of PM2.5 concentration against month, *temperature, pressure, dew point temperature, precipitation, and wind speed.*

Figure 3. PM2.5 concentration against hour – Evidence from SPSS Modeler



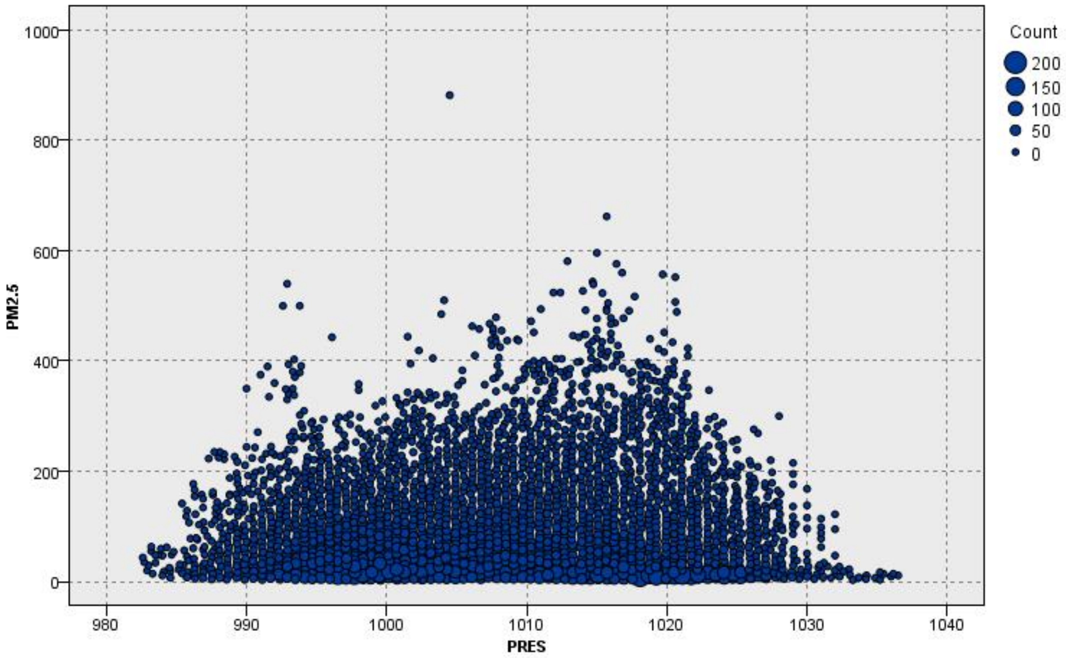
The PM2.5 concentration (>800) appears in the winter (December, January, February) and spring (March, April, May) in Beijing.

Figure 4. PM2.5 concentration against temperature – Evidence from SPSS Modeler



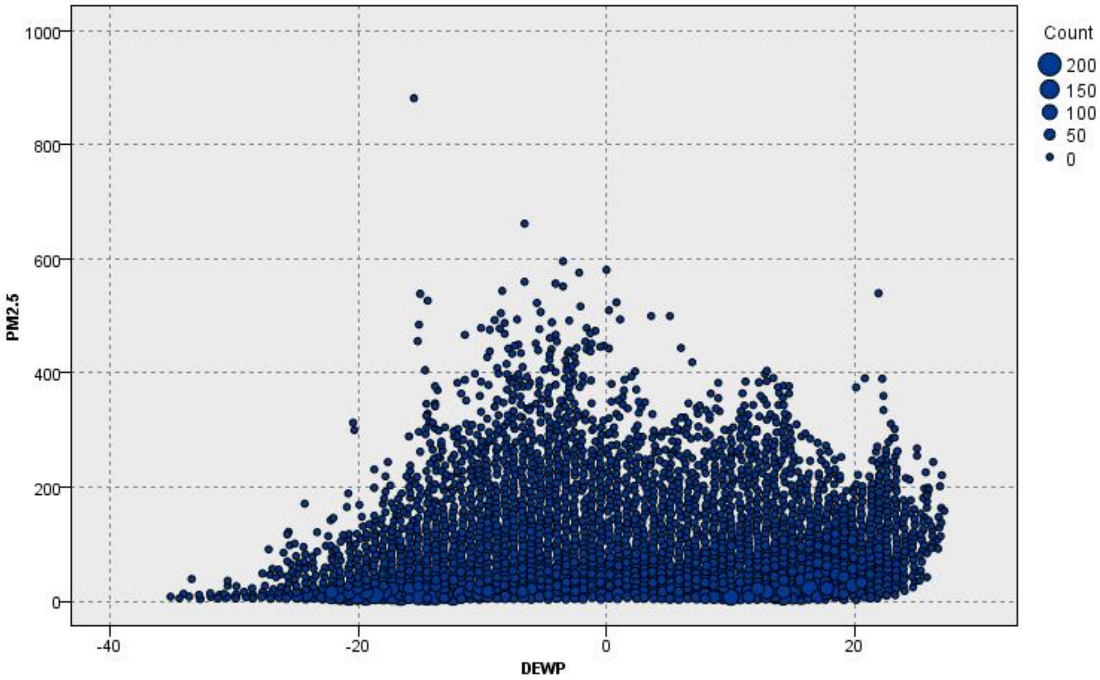
The PM2.5 concentration is affected by temperature. In this chart, the temperature between -5 to 10 degrees has a greater concentration of PM2.5(around 200) than the temperature between 20 with 40 degrees.

Figure 5. PM2.5 concentration against pressure – Evidence from SPSS Modeler



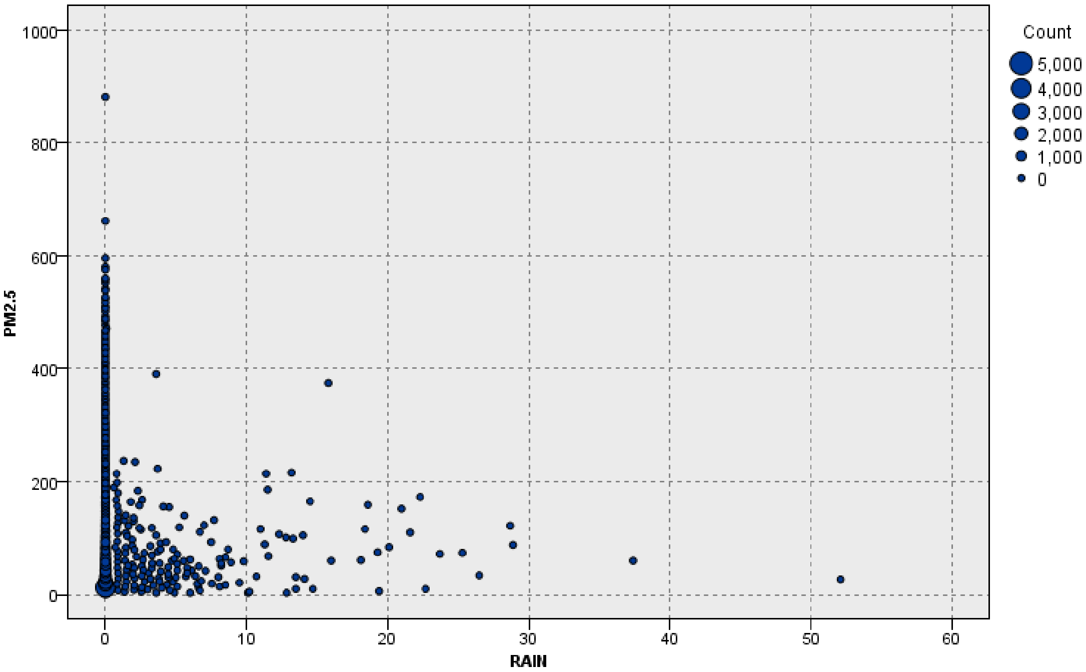
The PM2.5 concentration is not affected by the pressure. According to normal atmosphere pressure (around 1000 hPa), the concentration is slightly centralized near 200 ug/m^3.

Figure 6. PM2.5 concentration against dew point temperature – Evidence from SPSS Modeler



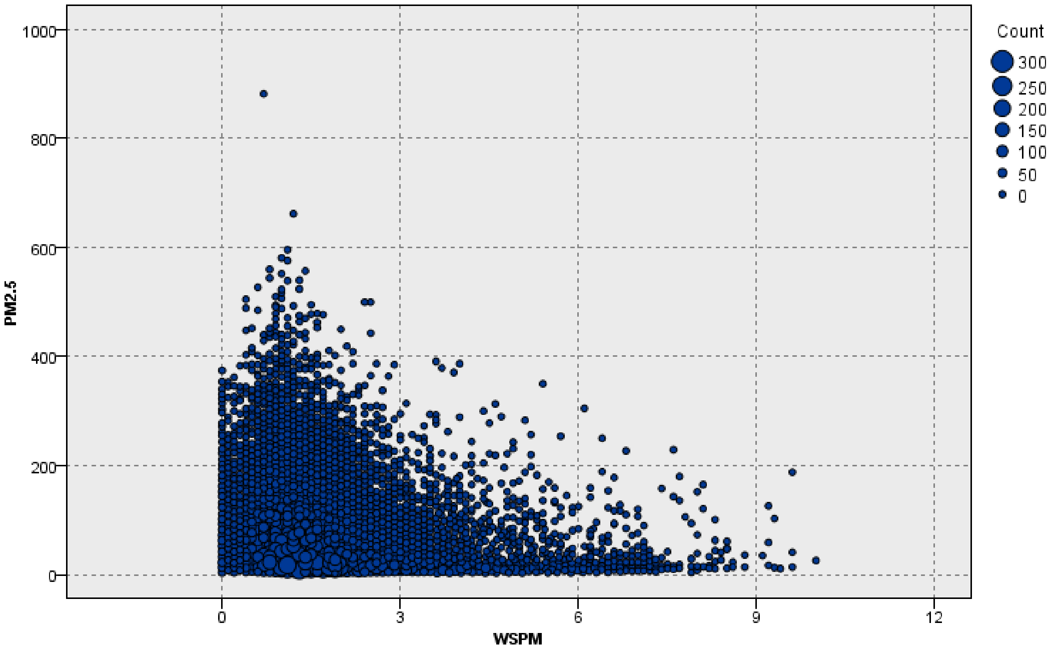
The PM2.5 concentration increased when the dew point temperature grows.

Figure 7. PM2.5 concentration against precipitation – Evidence from SPSS Modeler



The concentration of PM2.5 in 5000-1000 ug/m^3 when 0 precipitation is more generous than other precipitation levels.

Figure 8. PM2.5 concentration against Wind speed – Evidence from SPSS Modeler



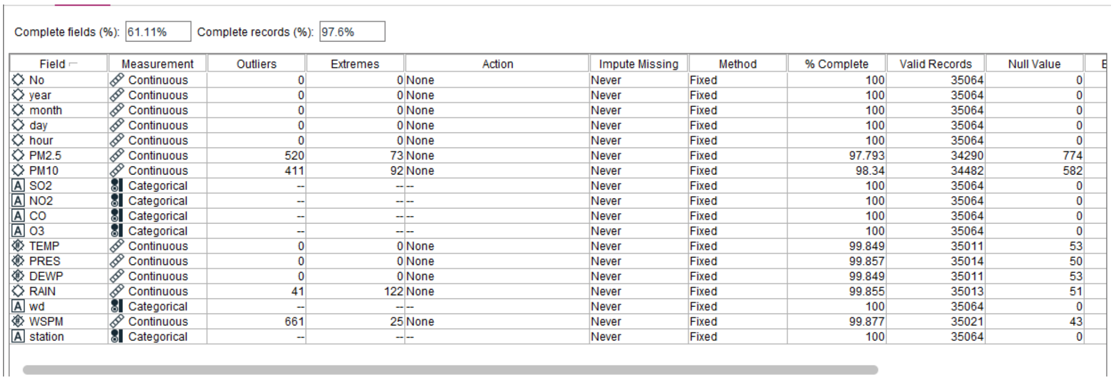
In the SPSS modeler, Plots show values of a *Y* field against values of an *X* field. Often, these fields correspond to a dependent variable and an independent variable, respectively.

It is clear that the PM2.5 concentration is affected significantly by the precipitation and wind speed; the concentration would decrease while the weather is rainy or windy. To demonstrate the concentration of PM2.5 affected by different variables by using the plot chart. Meanwhile, the PM2.5 concentration is higher than average while under the lower temperature (-20 to 10).

The PM2.5 concentration in winter (December, January, February) and spring (March, April, May) is much higher than the other two seasons. Because most weather in summer and autumn in Beijing is rainy and windy, the concentration of PM2.5 decreased clearly in these two seasons.

2.4. Verify the data quality

Figure 9. the quality of data before analysis – Evidence from SPSS Modeler



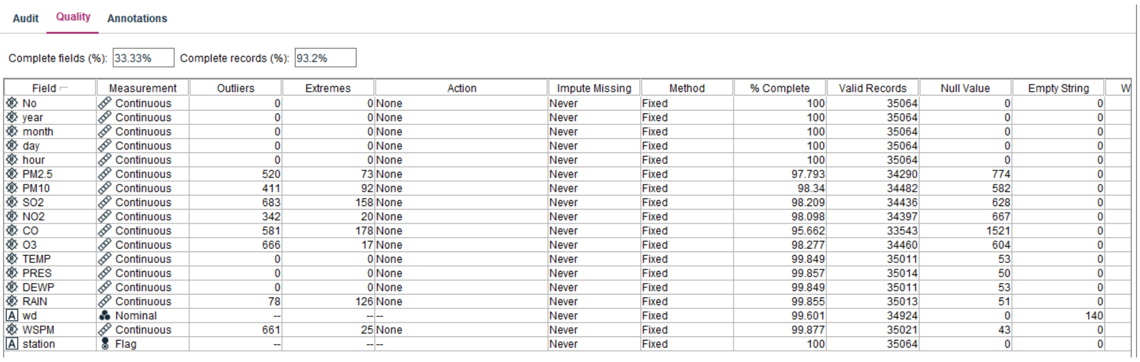
Complete fields (%): 61.11% Complete records (%): 97.6%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	E
No	Continuous	0	0	None	Never	Fixed	100	35064	0	
year	Continuous	0	0	None	Never	Fixed	100	35064	0	
month	Continuous	0	0	None	Never	Fixed	100	35064	0	
day	Continuous	0	0	None	Never	Fixed	100	35064	0	
hour	Continuous	0	0	None	Never	Fixed	100	35064	0	
PM2.5	Continuous	520	73	None	Never	Fixed	97.793	34290	774	
PM10	Continuous	411	92	None	Never	Fixed	98.34	34482	582	
SO2	Categorical	--	--	--	Never	Fixed	100	35064	0	
NO2	Categorical	--	--	--	Never	Fixed	100	35064	0	
CO	Categorical	--	--	--	Never	Fixed	100	35064	0	
O3	Categorical	--	--	--	Never	Fixed	100	35064	0	
TEMP	Continuous	0	0	None	Never	Fixed	99.849	35011	53	
PRES	Continuous	0	0	None	Never	Fixed	99.857	35014	50	
DEWP	Continuous	0	0	None	Never	Fixed	99.849	35011	53	
RAIN	Continuous	41	122	None	Never	Fixed	99.855	35013	51	
wd	Categorical	--	--	--	Never	Fixed	100	35064	0	
WSPM	Continuous	661	25	None	Never	Fixed	99.877	35021	43	
station	Categorical	--	--	--	Never	Fixed	100	35064	0	

The measurement type of some fields may still have some problems.

The measurements of SO2, NO2, CO, O3 may contain errors and may need to change to continuous.

Figure 10. after fixed data – Evidence from SPSS Modeler



Complete fields (%): 33.33% Complete records (%): 93.2%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	W
No	Continuous	0	0	None	Never	Fixed	100	35064	0	0	
year	Continuous	0	0	None	Never	Fixed	100	35064	0	0	
month	Continuous	0	0	None	Never	Fixed	100	35064	0	0	
day	Continuous	0	0	None	Never	Fixed	100	35064	0	0	
hour	Continuous	0	0	None	Never	Fixed	100	35064	0	0	
PM2.5	Continuous	520	73	None	Never	Fixed	97.793	34290	774	0	
PM10	Continuous	411	92	None	Never	Fixed	98.34	34482	582	0	
SO2	Continuous	683	158	None	Never	Fixed	98.209	34436	628	0	
NO2	Continuous	342	20	None	Never	Fixed	98.098	34397	667	0	
CO	Continuous	581	178	None	Never	Fixed	95.662	33543	1521	0	
O3	Continuous	666	17	None	Never	Fixed	98.277	34460	604	0	
TEMP	Continuous	0	0	None	Never	Fixed	99.849	35011	53	0	
PRES	Continuous	0	0	None	Never	Fixed	99.857	35014	50	0	
DEWP	Continuous	0	0	None	Never	Fixed	99.849	35011	53	0	
RAIN	Continuous	78	126	None	Never	Fixed	99.855	35013	51	0	
wd	Nominal	--	--	--	Never	Fixed	99.601	34924	0	140	
WSPM	Continuous	661	25	None	Never	Fixed	99.877	35021	43	0	
station	Flag	--	--	--	Never	Fixed	100	35064	0	0	

The chart above shows the correct fields. After changing the ‘NA’ into empty value, numbers of outliers and extremes value came up, these data need to be deleted or modified in future steps.

Some of the fields contain the Outliers and Extremes, these data should be analyzed and process in the later steps. The data in this particular situation may depend on the day’s weather, so the data may not be removable.

Missing values and null values need to be modified or processed. All the missing value in this data set is lower than 5%.

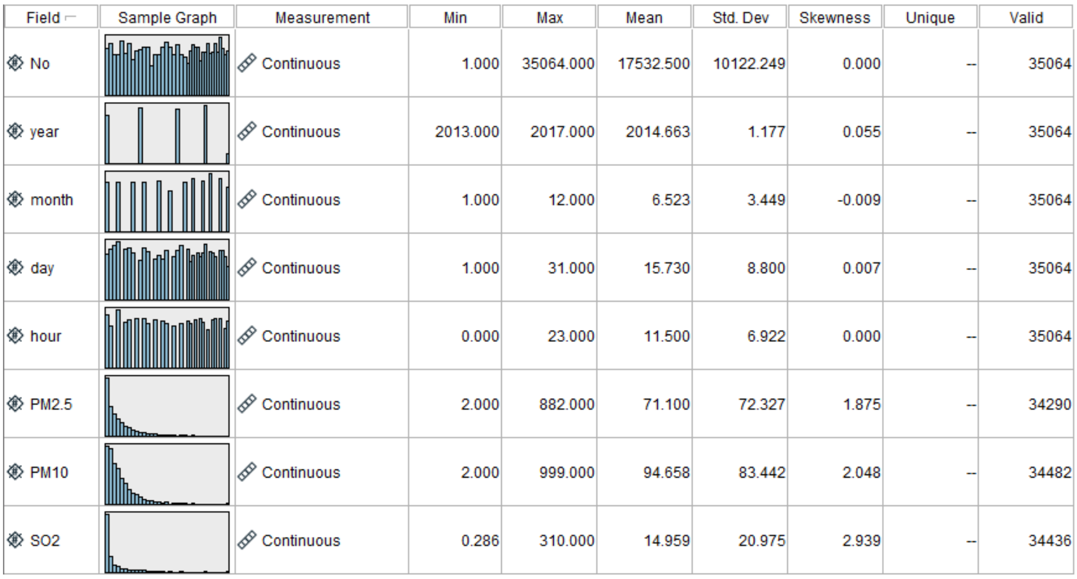
3. Data preparation

3.1. Select the data

By observing complete data and the distribution of missing data, I decided to select all the data in the dataset. The data set got an entire record of the weather conditions for every hour from March 1st, 2013 to February 28th, 2017. The missing data of each attribute is irregularly distributed every hour during this period.

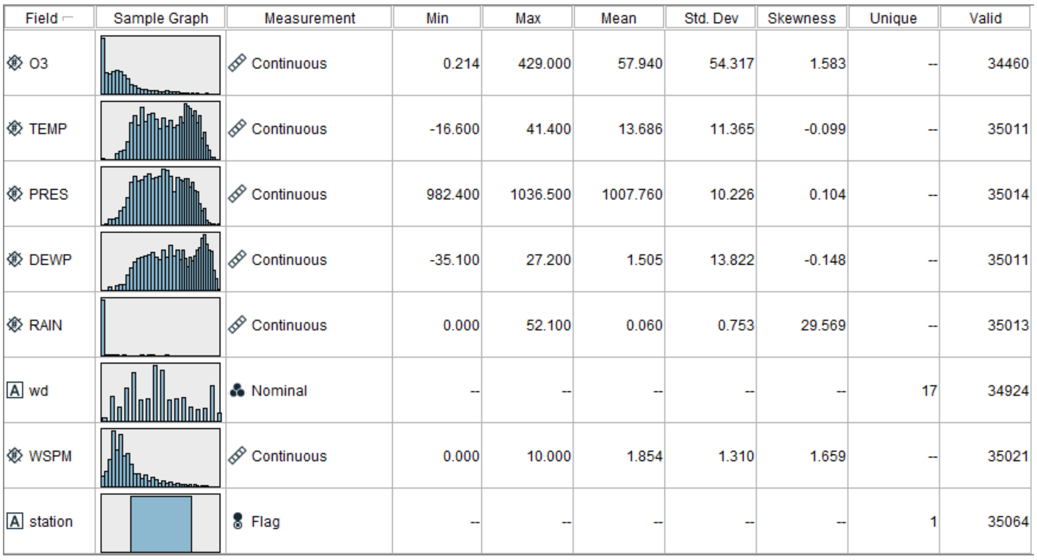
Therefore, all the data in the data set should be considered.

Figure 11. the selected data(part one) – Evidence from SPSS Modeler



Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
No		Continuous	1.000	35064.000	17532.500	10122.249	0.000	--	35064
year		Continuous	2013.000	2017.000	2014.663	1.177	0.055	--	35064
month		Continuous	1.000	12.000	6.523	3.449	-0.009	--	35064
day		Continuous	1.000	31.000	15.730	8.800	0.007	--	35064
hour		Continuous	0.000	23.000	11.500	6.922	0.000	--	35064
PM2.5		Continuous	2.000	882.000	71.100	72.327	1.875	--	34290
PM10		Continuous	2.000	999.000	94.658	83.442	2.048	--	34482
SO2		Continuous	0.286	310.000	14.959	20.975	2.939	--	34436

Figure 12. the selected data(part two) – Evidence from SPSS Modeler



Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
O3		Continuous	0.214	429.000	57.940	54.317	1.583	--	34460
TEMP		Continuous	-16.600	41.400	13.686	11.365	-0.099	--	35011
PRES		Continuous	982.400	1036.500	1007.760	10.226	0.104	--	35014
DEWP		Continuous	-35.100	27.200	1.505	13.822	-0.148	--	35011
RAIN		Continuous	0.000	52.100	0.060	0.753	29.569	--	35013
wd		Nominal	--	--	--	--	--	17	34924
WSPM		Continuous	0.000	10.000	1.854	1.310	1.659	--	35021
station		Flag	--	--	--	--	--	1	35064

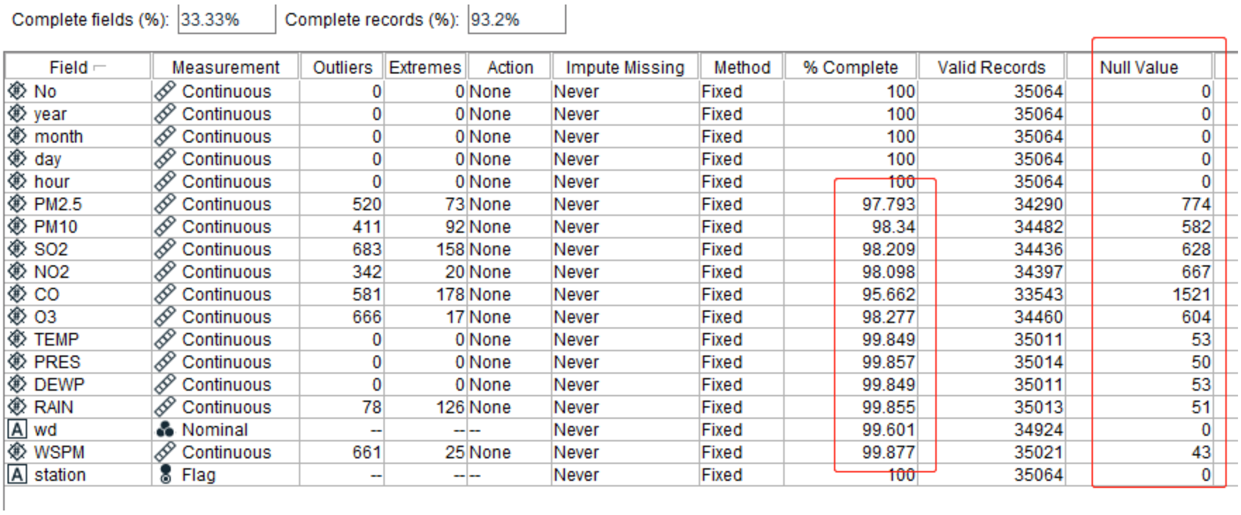
There are 18 fields in this data set. 'No' has marked as the record id. 'wd' is the wind direction, was recorded in abbreviation. 'Station' has been marked as a Flag value, because there is only one monitor station(Changping) in this data set. Other fields are all numeric values.

3.2. Clean the data

There are blank values, outliers, and extreme values in the data, affecting the analysis results. As mentioned in 2.4, verify the data, the missing value should be replaced.

This data set's total record is 35064, and there is the null or empty value show in the chart below. In terms of replacing or handling the missing value, the 'Missing Values SuperNode' has been used.

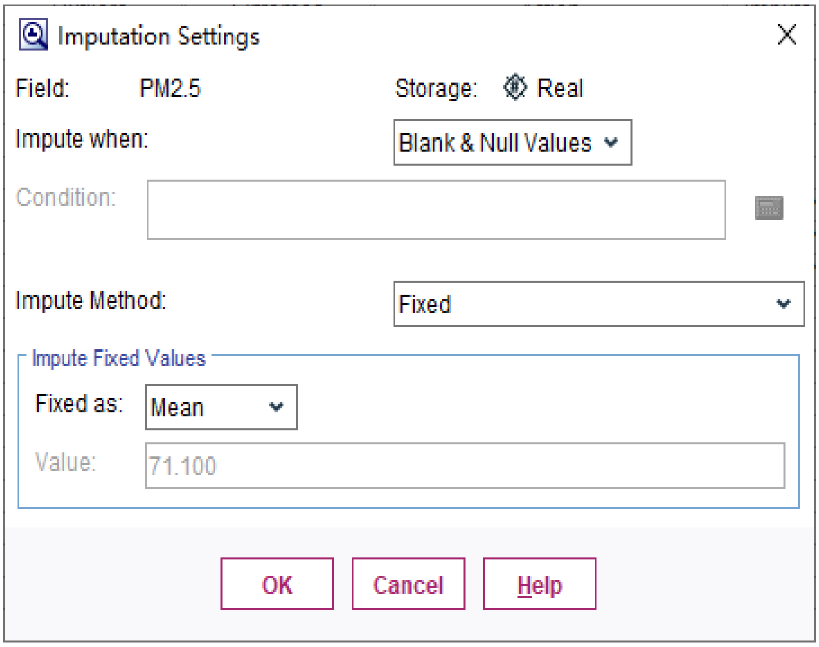
Figure 13. data quality before handle missing value – Evidence from SPSS Modeler



Complete fields (%): 33.33% Complete records (%): 93.2%

Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value
No	Continuous	0	0	None	Never	Fixed	100	35064	0
year	Continuous	0	0	None	Never	Fixed	100	35064	0
month	Continuous	0	0	None	Never	Fixed	100	35064	0
day	Continuous	0	0	None	Never	Fixed	100	35064	0
hour	Continuous	0	0	None	Never	Fixed	100	35064	0
PM2.5	Continuous	520	73	None	Never	Fixed	97.793	34290	774
PM10	Continuous	411	92	None	Never	Fixed	98.34	34482	582
SO2	Continuous	683	158	None	Never	Fixed	98.209	34436	628
NO2	Continuous	342	20	None	Never	Fixed	98.098	34397	667
CO	Continuous	581	178	None	Never	Fixed	95.662	33543	1521
O3	Continuous	666	17	None	Never	Fixed	98.277	34460	604
TEMP	Continuous	0	0	None	Never	Fixed	99.849	35011	53
PRES	Continuous	0	0	None	Never	Fixed	99.857	35014	50
DEWP	Continuous	0	0	None	Never	Fixed	99.849	35011	53
RAIN	Continuous	78	126	None	Never	Fixed	99.855	35013	51
wd	Nominal	--	--	--	Never	Fixed	99.601	34924	0
WSPM	Continuous	661	25	None	Never	Fixed	99.877	35021	43
station	Flag	--	--	--	Never	Fixed	100	35064	0

Figure 14. handle missing value, fixed as mean



Imputation Settings

Field: PM2.5

Storage: Real

Impute when: Blank & Null Values

Condition:

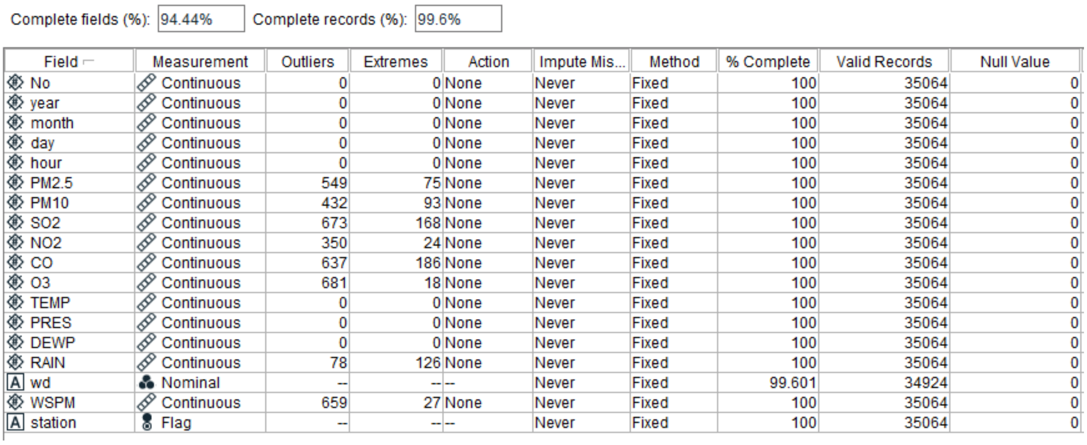
Impute Method: Fixed

Impute Fixed Values

Fixed as: Mean

Value: 71.100

Figure 15. data quality after handle missing value – Evidence from SPSS Modeler



Complete fields (%): 94.44% Complete records (%): 99.6%

Field	Measurement	Outliers	Extremes	Action	Impute Mis...	Method	% Complete	Valid Records	Null Value
No	Continuous	0	0	None	Never	Fixed	100	35064	0
year	Continuous	0	0	None	Never	Fixed	100	35064	0
month	Continuous	0	0	None	Never	Fixed	100	35064	0
day	Continuous	0	0	None	Never	Fixed	100	35064	0
hour	Continuous	0	0	None	Never	Fixed	100	35064	0
PM2.5	Continuous	549	75	None	Never	Fixed	100	35064	0
PM10	Continuous	432	93	None	Never	Fixed	100	35064	0
SO2	Continuous	673	168	None	Never	Fixed	100	35064	0
NO2	Continuous	350	24	None	Never	Fixed	100	35064	0
CO	Continuous	637	186	None	Never	Fixed	100	35064	0
O3	Continuous	681	18	None	Never	Fixed	100	35064	0
TEMP	Continuous	0	0	None	Never	Fixed	100	35064	0
PRES	Continuous	0	0	None	Never	Fixed	100	35064	0
DEWP	Continuous	0	0	None	Never	Fixed	100	35064	0
RAIN	Continuous	78	126	None	Never	Fixed	100	35064	0
wd	Nominal	--	--	--	Never	Fixed	99.601	34924	0
WSPM	Continuous	659	27	None	Never	Fixed	100	35064	0
station	Flag	--	--	--	Never	Fixed	100	35064	0

The above chart shows the data quality after replacing the missing value.

However, the wd data means the wind direction, which is the string value, and the wind direction did not affect the testing concentration of PM2.5, PM10, SO2, etc. So I leave the blank or null value in the wd field.

3.3. Construct the data

In this project, the primary purpose is to predict the concentration of PM2.5, and the data set already contains all the required indicators; as a result, it is not necessary to add more datasets or tables.

3.4. Integrate various data sources

In this project, the dataset contains all valid data. It is an unnecessary need to collect data and does not need to integrate any tables or attributes.

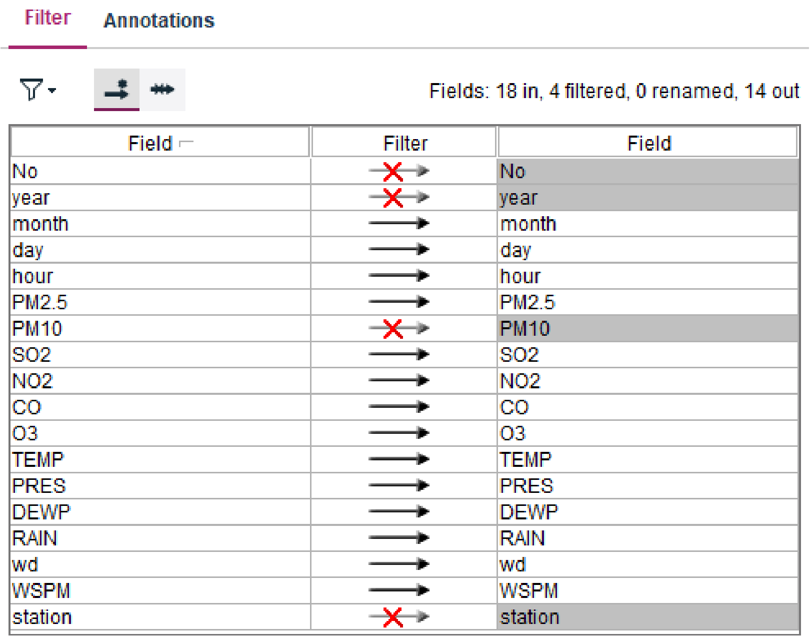
3.5. Format the data as required

In this dataset, reformatting is not required. The data set comes from a professional statistical team, and All data were obtained from the Meteorological Monitoring Center. The format of data is suitable for this project to predict the PM2.5 concentration.

4. Data transformation

4.1. Reduce the data

Figure 16. deletion of the fields – Evidence from SPSS Modeler



Filter Annotations

Fields: 18 in, 4 filtered, 0 renamed, 14 out

Field	Filter	Field
No	✗→	No
year	✗→	year
month	→	month
day	→	day
hour	→	hour
PM2.5	→	PM2.5
PM10	✗→	PM10
SO2	→	SO2
NO2	→	NO2
CO	→	CO
O3	→	O3
TEMP	→	TEMP
PRES	→	PRES
DEWP	→	DEWP
RAIN	→	RAIN
wd	→	wd
WSPM	→	WSPM
station	✗→	station

After cleaned the data, replaced the missing value, some of the useless attributes should be reduced. Additionally, the data may influence accuracy by making the model overfit the data set.

Reasons for deletion

- ‘No’, ‘year’ and ‘station’ are the useless fields in predicting the PM2.5.
- ‘No’ is the record number, there are 35064 rows of data.
- ‘year’ is the year from 2013 to 2017, which did not affect the PM2.5 concentration, only for recording the time.
- ‘station’ is the region that had been monitored(Changping, Beijing), there is only one station in this dataset.
- ‘PM10’ is the concentration of PM10, which is also a particulate matter, and it may affect the role of other air attributes.

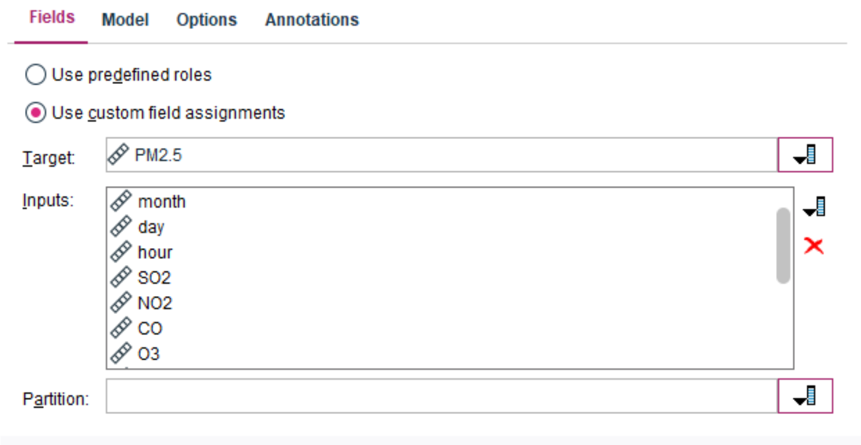
4.2. Project the data

In this step, run the dataset with the “feature selection” the modeling from the SPSS Modeler, analyze the ‘PM2.5’ field as the target, all other fields were set to the input.

The diagram below shows 11 essential fields and two unimportant fields(‘Rain’ and ‘day’) in predicting the PM2.5. One screened field is the ‘PRES’(pressure) because it is a coefficient of variation below the threshold.

Then use filter to select all the essential fields for the future steps.

Figure 17. set PM2.5 as the target, other fields are input– Evidence from SPSS Modeler



○ Use predefined roles

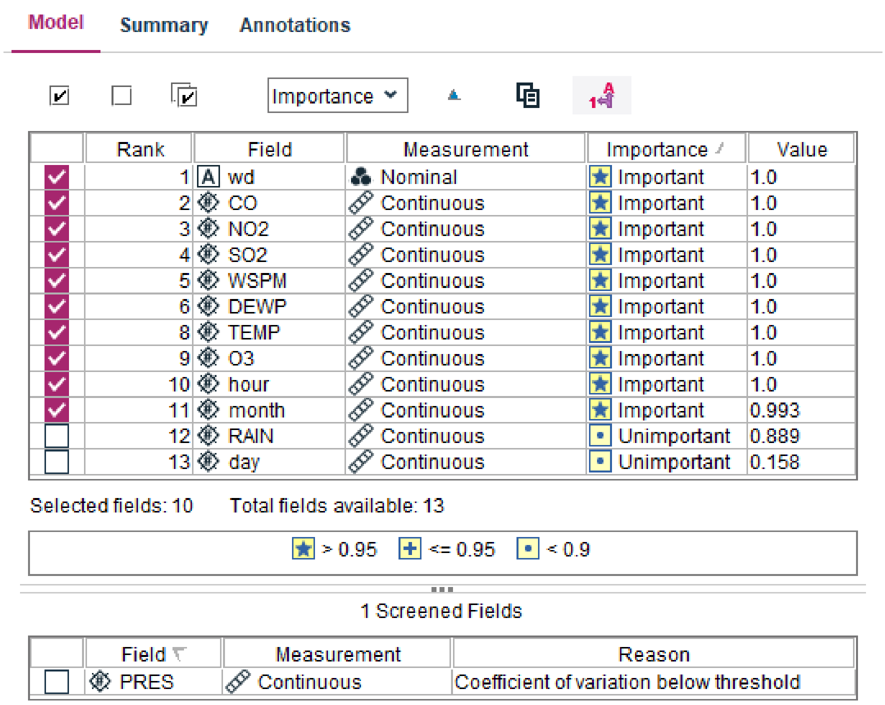
● Use custom field assignments

Target: PM2.5

Inputs:
month
day
hour
SO2
NO2
CO
O3

Partition:

Figure 18. result after feature selection – Evidence from SPSS Modeler



	Rank	Field	Measurement	Importance	Value
☑	1	wd	Nominal	★ Important	1.0
☑	2	CO	Continuous	★ Important	1.0
☑	3	NO2	Continuous	★ Important	1.0
☑	4	SO2	Continuous	★ Important	1.0
☑	5	WSPM	Continuous	★ Important	1.0
☑	6	DEWP	Continuous	★ Important	1.0
☑	8	TEMP	Continuous	★ Important	1.0
☑	9	O3	Continuous	★ Important	1.0
☑	10	hour	Continuous	★ Important	1.0
☑	11	month	Continuous	★ Important	0.993
☐	12	RAIN	Continuous	• Unimportant	0.889
☐	13	day	Continuous	• Unimportant	0.158

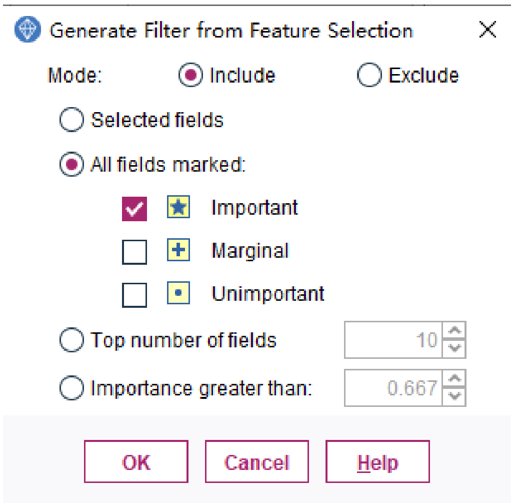
Selected fields: 10   Total fields available: 13

★ > 0.95 + <= 0.95 • < 0.9

1 Screened Fields

	Field	Measurement	Reason
☐	PRES	Continuous	Coefficient of variation below threshold

Figure 19. select the important fields – Evidence from SPSS Modeler



Generate Filter from Feature Selection

Mode: ● Include ○ Exclude

○ Selected fields

● All fields marked:

☑ ★ Important

☐ + Marginal

☐ • Unimportant

○ Top number of fields 10

○ Importance greater than: 0.667

5. Data-mining methods selection

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

5.1.1. Supervised learning and unsupervised learning

The training data are labeled, which is supervised learning. Training data without a label is unsupervised learning (Sathya & Abraham, 2013).

The supervised learning uses the algorithm or model to understand the relationship of the data input and output, including classification and regression. The unsupervised learning is that only have input data and no corresponding output, for example, clustering.

5.1.2. Discuss the method with the Data mining objective

Base on the evidence above supervised learning is suitable for this project. Meanwhile, learning the existing function from a given dataset is useful to predict new data results.

In this project, the data mining objective is to predict the concentration of PM2.5 by the given fields, including the effect of weather and other pollute air. The target is to predict the PM2.5 concentration and consider all the other attributes.

The classification method is used to divide a group of data objects in the database into different classes according to the classification model and find out the similar characteristics of the data objects (Data for Data Mining, n.d.).

The regression analysis method reflects the temporal features of attributes in the dataset. At the same time, the regression method generates a function that maps data elements to a real value prediction variable and discovers the dependence between variables or attributes (Data for Data Mining, n.d.).

5.2. Select the appropriate data-mining methods based on discussion

The clustering method is not suitable for the project, both input data and output data are required in this project. In this case, unsupervised learning method is not allowed to be used.

The classification method is also not suitable for this objective. The classification algorithms estimate the collection from inputs data to the categorical output variables. On the other hand, the data type of classification output might be a category. However, this project's prediction data is numeric, which is continuous data, but not discrete.

Regarding the data-mining objectives, there are both input variables and output variables; thus, it is necessary to choose the supervised learning methods. Because the predicted target is PM2.5, it is the continuous value, and the data mining method should be decided as the regression method.

6. Data-mining algorithms selection

6.1. Conduct exploratory analysis and discuss

In terms of selecting algorithms, there are three main data-mining algorithms for solving regression problems, including “Linear and Polynomial Regression, Neural Networks, and Regression Trees and Random Forests” (Seif, 2018).

To choose a suitable algorithm for the project, it is necessary to find out the definition and function of the algorithm and discover the advantages and disadvantages.

- Linear and Polynomial Regression

  To begin with the simple case, the single variable linear regression is one of the techniques to function the relationship between the independent input variable and the dependent output variable, which would use a linear model. On the other hand, the multivariable linear regression is a model for multiple input independent attributes (feature variables) and an output dependent variable. This linear model would combine various input variables (Seif, 2018). Another general case is called polynomial regression; this model becomes a non-linear combination of the independent variables and would require knowledge of how data relates to the output.

  Advantages:

  Linear regression is useful for small and straightforward size dataset, fast to model.

  Linear regression is easy to understand the valuable business decisions.

  Limitations:

  Because the polynomial regression model must contain the data structure and the feature variables’ relationship, it is hard for non-linear data to design.

  Linear and Polynomial Regression do not have enough benefits during analysis of complex data.

  For data mining objective:

  The most appropriate algorithm for the project should be the multivariable linear regression. We set the PM2.5 concentration as the target, which is the output, and combine all other attributes as the input, which is the independent variables.

- Neural Networks

  An interrelated group of nodes is called neurons; it consists of a neural network. For each neuron, the independent input variables would model as a multi-variable linear combination, and all the feature variables may multiply a value called weights. A non-linearity gives the neural network the ability to model complex non-linear relationships, as is applied to the linear combination of neurons. A neural network can have multiple layers where one layer’s output would pass other results in the same way. There is no generally non-linearity applied to the output. “Neural Networks are trained using Stochastic Gradient Descent (SGD) and the backpropagation algorithm” (Seif, 2018).

Advantages:

Neural networks could effectively model complex non-linear relationships because it contains many non-linearities' parameters.

We generally do not need to worry about learning the data structure and the feature variables' relationships in neural networks.

Limitations:

The neural network models are not easy to understand and translate.

Other machine learning algorithms commonly exceed neural networks in "small data" cases.

They demand an amount of data to accomplish high performance.

For data mining objective:

As the neural networks could effectively model complex non-linear relationships and achieve high performance, the dataset's quantity could train the neural network. Therefore, set the concentration of PM2.5 as the target, the neural network in SPSS Modeler could be a suitable algorithm for the project.

- Regression Trees and Random Forests

Regarding the decision tree's base case, it is an intuitional model where one traverses down the tree branches, then selects the next chapter to keep going down based on the decision at the node. The induction of tree is input training instances, decide the best attributes to split on. All training instances should be categorized after splitting the dataset and recurring on the resulting split datasets. While building the tree, the goal is to break into the attributes that create the purest child nodes possible, keeping to a minimum the number of splits that would need to be made to classify all instances in our dataset. The purity relates to a previously unseen example for it to be appropriately organized.

Decision trees are a part of random forests, and the input elements are run through several decision trees. For classifications, a voting system is used to establish the final class. For regression, the output value of all trees is averaged (Seif, 2018).

Advantages:

Regression trees could easily handle complex non-linear relationships. They could often accomplish high performance, better than polynomial regression, and the same functions with neural networks.

Very easy to understand and translate. The decision boundaries are realistic and easy to understand.

Limitations:

The training decision trees can be capable of major overfitting because of its nature. The over complex and unnecessary structure may occur in a completed decision tree model.

It would require more memory and time to reach higher performance while using more extraordinary random forest.

For data mining objective:

Because the relationships between the target(PM 2.5) and other attributes are uncertain, the tree regression model may successfully find the connection. The 35064 instances should be able to prevent overfitting.

- XGBoost

According to the XGBoost algorithm steps, first initialized to a constant, GB is based on the first derivative RI, XGBoost is based on the first derivative GI, and the second derivative HI, iteratively generate a base learner, then add to update the learner. The primary learner in XGBoost could be either CART (GBtree) or linear classifier (GBlinear) (Regression Trees, 2013).

Advantages:

XGBoost model could enumerate several candidates, thus finding the best split point from the candidates based on the formula (Random Forest, GBDT, and XGBoost, n.d.).

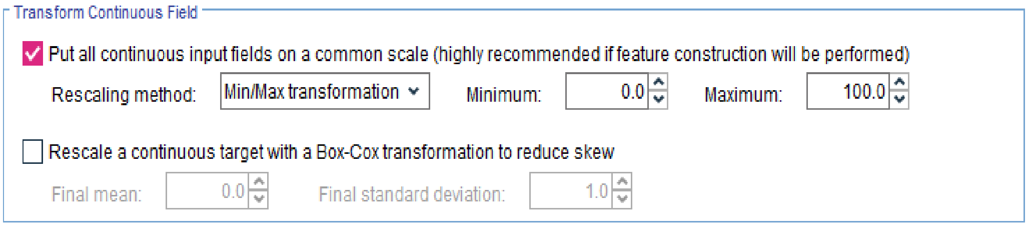
For data mining objective:

In this project, the dataset contains many data, and the XGBoost model can segment and manage the data efficiently because it could find the best split point.

All the algorithms above show the advantages and disadvantages. Because we want to find out the prediction of PM2.5, I would select the Random Forest algorithm, XGBoost Tree algorithm, and the Neural Net algorithm for this project to study the final results.

6.2. Select data-mining algorithms based on discussion

Figure 20. "Auto Data Prep" setting – Evidence from SPSS Modeler



Transform Continuous Field

☑ Put all continuous input fields on a common scale (highly recommended if feature construction will be performed)

Rescaling method: Min/Max transformation   Minimum: 0.0   Maximum: 100.0

☐ Rescale a continuous target with a Box-Cox transformation to reduce skew

Final mean: 0.0   Final standard deviation: 1.0

Because the large values would affect more on prediction results than the smaller amounts. As a result, without generate the significant value after the "feature selection" node in the SPSS Modeler, I choose to add another node "Auto Data Prep" and set the rescaling method Min/Max transformation and set the range from 0 to 100, as the screenshot shown above.

After understanding the algorithms associated with the regression, I run the "Auto Numeric" node to get the algorithms automatically from the SPSS Modeler; the models have shown below. Comparing the correlations, all three models are higher than 0.850, which means there are significances in this model and associated well with the target.

Figure 21. Two important nodes for selecting algorithms– Evidence from SPSS Modeler

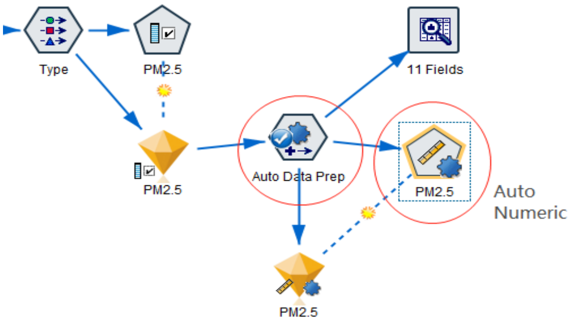
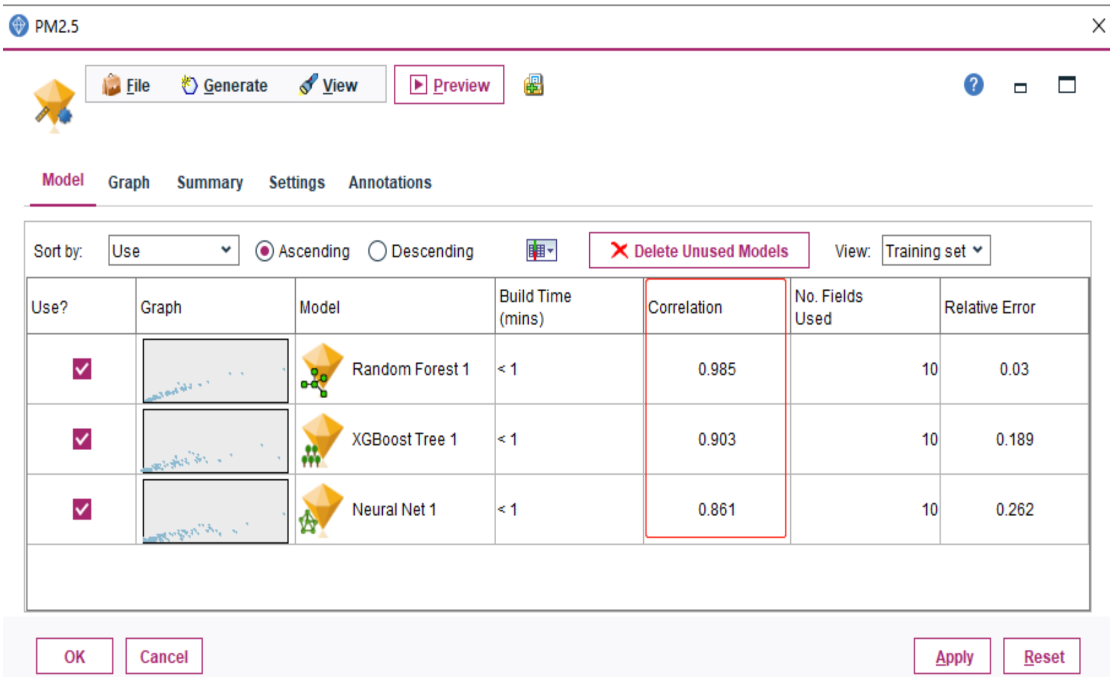


Figure 22. The model chooses after running the "Auto Numeric" node – Evidence from SPSS Modeler



Sort by: Use ● Ascending ○ Descending View: Training set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		Random Forest 1	< 1	0.985	10	0.03
☑		XGBoost Tree 1	< 1	0.903	10	0.189
☑		Neural Net 1	< 1	0.861	10	0.262

Base on the 6.1 discussion and the node's generated results, the project will analyze the "Random Forest," "XGBoost Tree," and "Neural Net" model for the prediction of PM2.5.

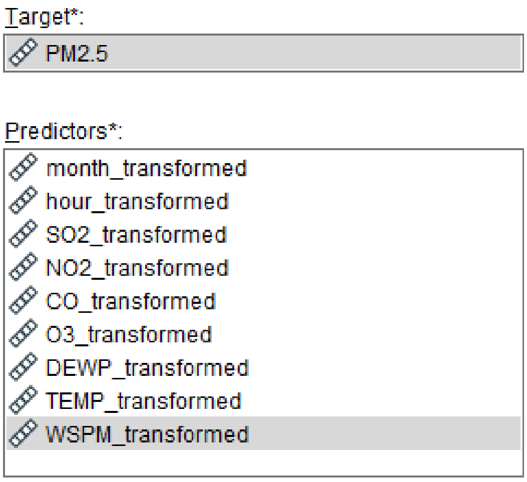
6.3. Build/Select appropriate models and choose relevant parameters

After use SPSS Modeler(figure 21) to select the algorithms, I would recommend an analysis of how I build the appropriate models. The order of models is based on the correlation levels.

Model 1: Random Forests

The target(PM2.5) and predictors(all other fields after transformed, but without the nominal fields 'wd') are still the same, as the screenshot shows below. Figure 24 and 25 describe the settings of the random forest model, and the system automatically sets both the basics and advanced settings.

Figure 23. Target and predictors in the Random Forest model– Evidence from SPSS Modeler



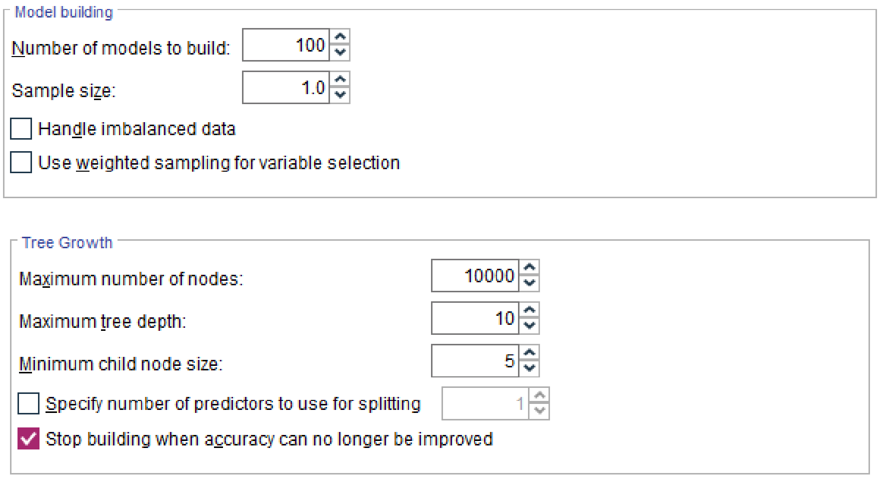
Target*:

PM2.5

Predictors*:

month_transformed
hour_transformed
SO2_transformed
NO2_transformed
CO_transformed
O3_transformed
DEWP_transformed
TEMP_transformed
WSPM_transformed

Figure 24. Basics settings of Random Forest Model – Evidence from SPSS Modeler



Model building

Number of models to build:	100
Sample size:	1.0

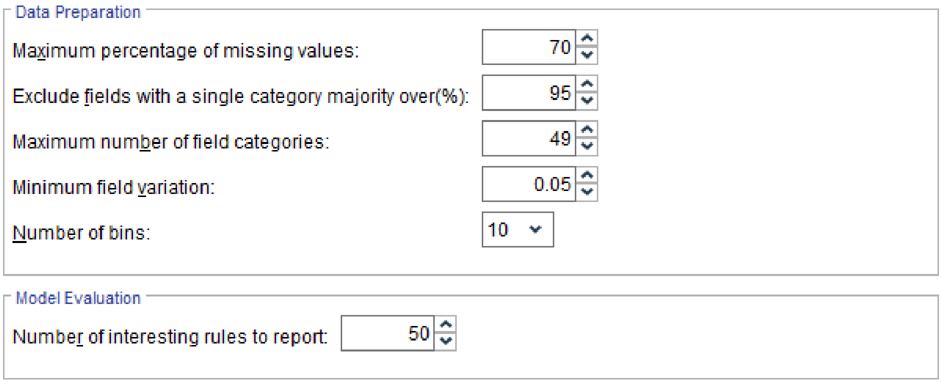
□ Handle imbalanced data
□ Use weighted sampling for variable selection

Tree Growth

Maximum number of nodes:	10000
Maximum tree depth:	10
Minimum child node size:	5
□ Specify number of predictors to use for splitting	1

☑ Stop building when accuracy can no longer be improved

Figure 25. Advanced settings of Random Forest Model – Evidence from SPSS Modeler



Data Preparation

Maximum percentage of missing values:	70
Exclude fields with a single category majority over(%):	95
Maximum number of field categories:	49
Minimum field variation:	0.05
Number of bins:	10

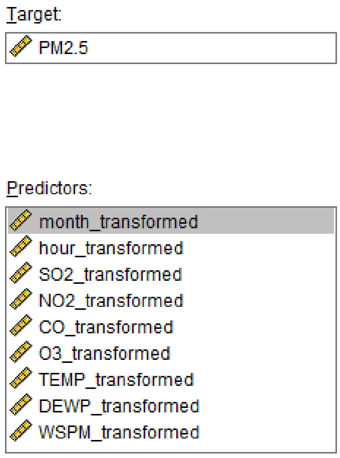
Model Evaluation

Number of interesting rules to report:	50

Model 2: XGBoost Tree

The target and predictors are still the same fields(without the nominal fields ‘wd’).

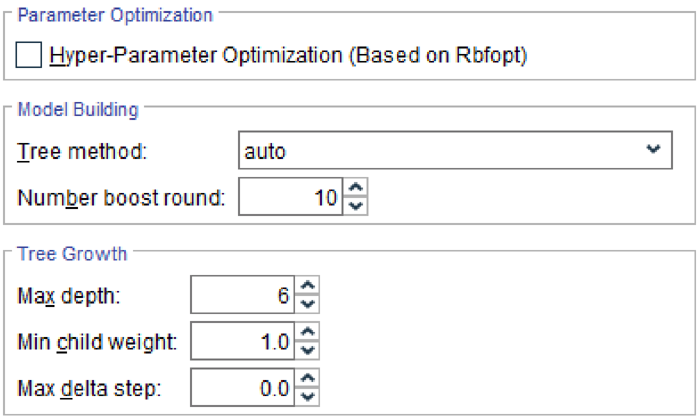
Figure 26. Target and predictor of XGBoost Tree model – Evidence from SPSS Modeler



Target:
PM2.5

Predictors:
month_transformed
hour_transformed
SO2_transformed
NO2_transformed
CO_transformed
O3_transformed
TEMP_transformed
DEWP_transformed
WSPM_transformed

Figure 27. basics setting of XGBoost Tree model – Evidence from SPSS Modeler



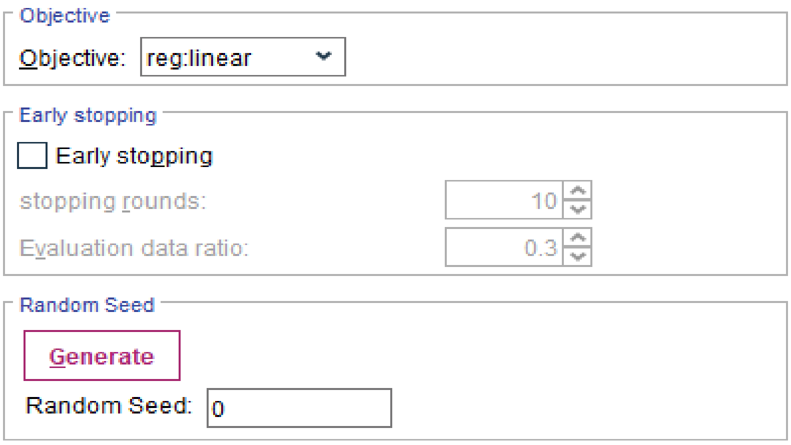
Parameter Optimization
□ Hyper-Parameter Optimization (Based on Rbfopt)

Model Building
Tree method: auto
Number boost round: 10

Tree Growth
Max depth: 6
Min child weight: 1.0
Max delta step: 0.0

As the XGBoost tree model is scalable and flexible gradient tree, by setting the tree method as auto and the number boost round as 10. Tree growth depth cannot greater than 6; the minimize child weight are set to 1.0; the maximum delta step is 0.0 (shown in figure 27).

Figure 28. learning task setting of XGBoost Tree Model – Evidence from SPSS Modeler



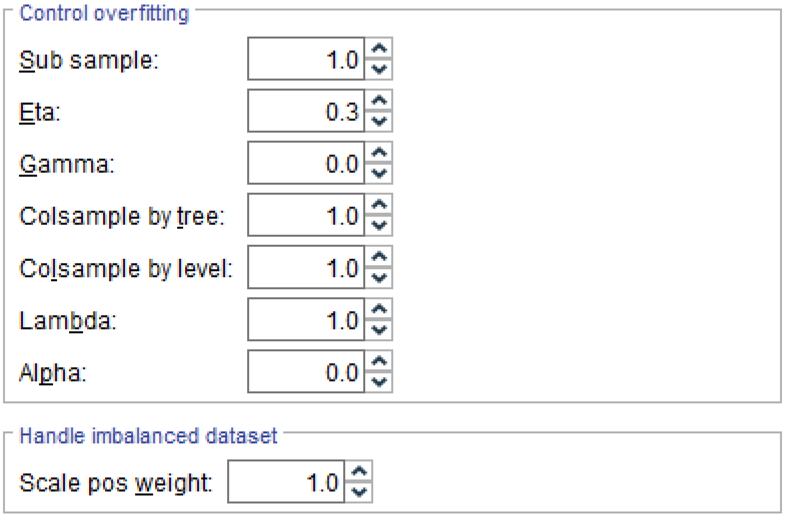
Objective
Objective: reg:linear

Early stopping
□ Early stopping
stopping rounds: 10
Evaluation data ratio: 0.3

Random Seed

Random Seed: 0

Figure 29. The advanced setting of XGBoost Tree Model – Evidence from SPSS Modeler



Control overfitting

Sub sample: 1.0

Eta: 0.3

Gamma: 0.0

Colsample by tree: 1.0

Colsample by level: 1.0

Lambda: 1.0

Alpha: 0.0

Handle imbalanced dataset

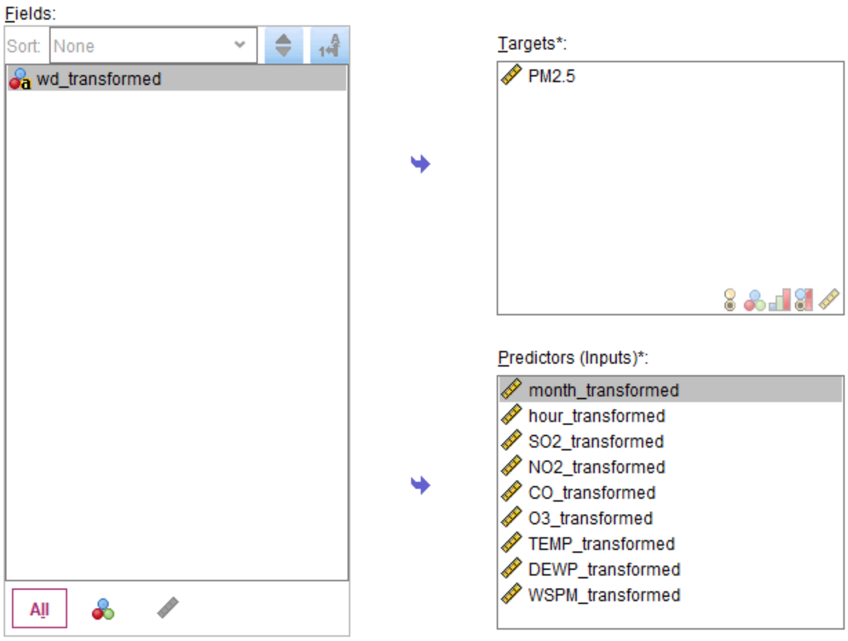
Scale pos weight: 1.0

As the settings in figure 28, the objective has been set as linear regression, the scenes in figure 29 could control overfitting and avoid imbalanced dataset.

Model 3: Neural networks

The Neural networks model, the target, and the predictors(inputs) are still the same with other models(as shown in figure 30), without the wind directions. The basics settings have been shown in figure 31, where we are building a new model, and the main objective is to create a standard model. Figure 32 shows the advanced setting of neural networks and set the appropriate overfit prevention set to 30%.

Figure 30. Target and predictor of Neural Network model – Evidence from SPSS Modeler



Fields:

Sort: None

wd_transformed

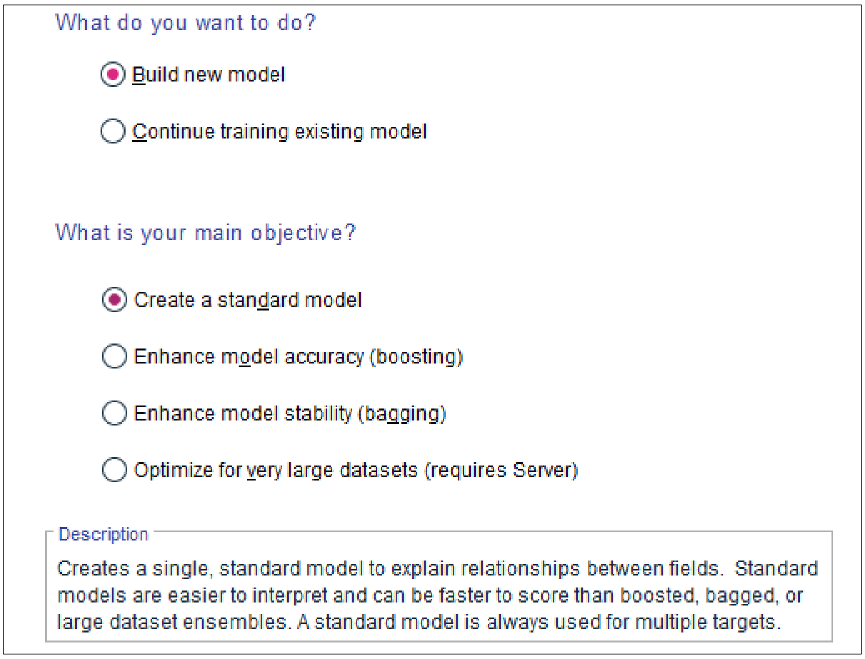
Targets*:

PM2.5

Predictors (Inputs)*:

month_transformed
hour_transformed
SO2_transformed
NO2_transformed
CO_transformed
O3_transformed
TEMP_transformed
DEWP_transformed
WSPM_transformed

Figure 31. Basics setting of Neural Network model – Evidence from SPSS Modeler



What do you want to do?

● Build new model

○ Continue training existing model

What is your main objective?

● Create a standard model

○ Enhance model accuracy (boosting)

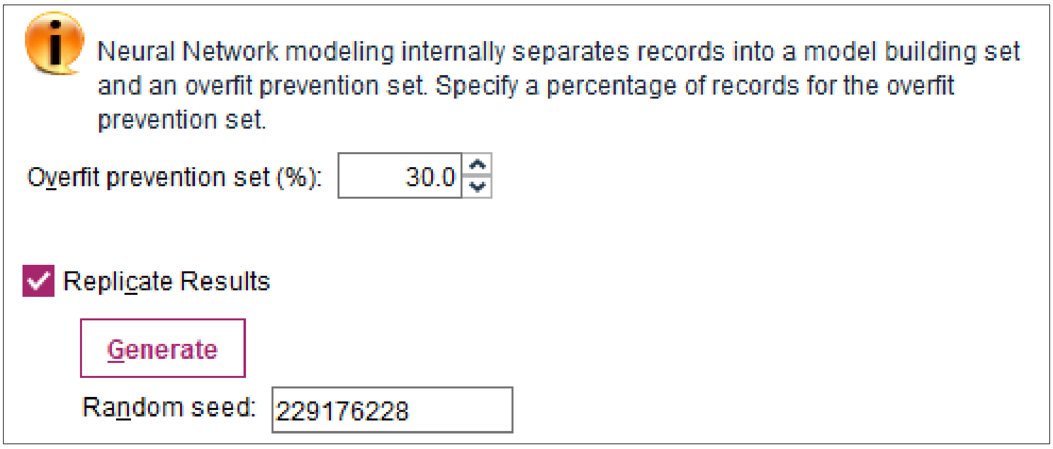
○ Enhance model stability (bagging)

○ Optimize for very large datasets (requires Server)

Description

Creates a single, standard model to explain relationships between fields. Standard models are easier to interpret and can be faster to score than boosted, bagged, or large dataset ensembles. A standard model is always used for multiple targets.

Figure 32. The advanced setting of Neural Network model – Evidence from SPSS Modeler



Neural Network modeling internally separates records into a model building set and an overfit prevention set. Specify a percentage of records for the overfit prevention set.

Overfit prevention set (%): 30.0

☑ Replicate Results

Random seed: 229176228

7. Data Mining

7.1. Create and justify test designs

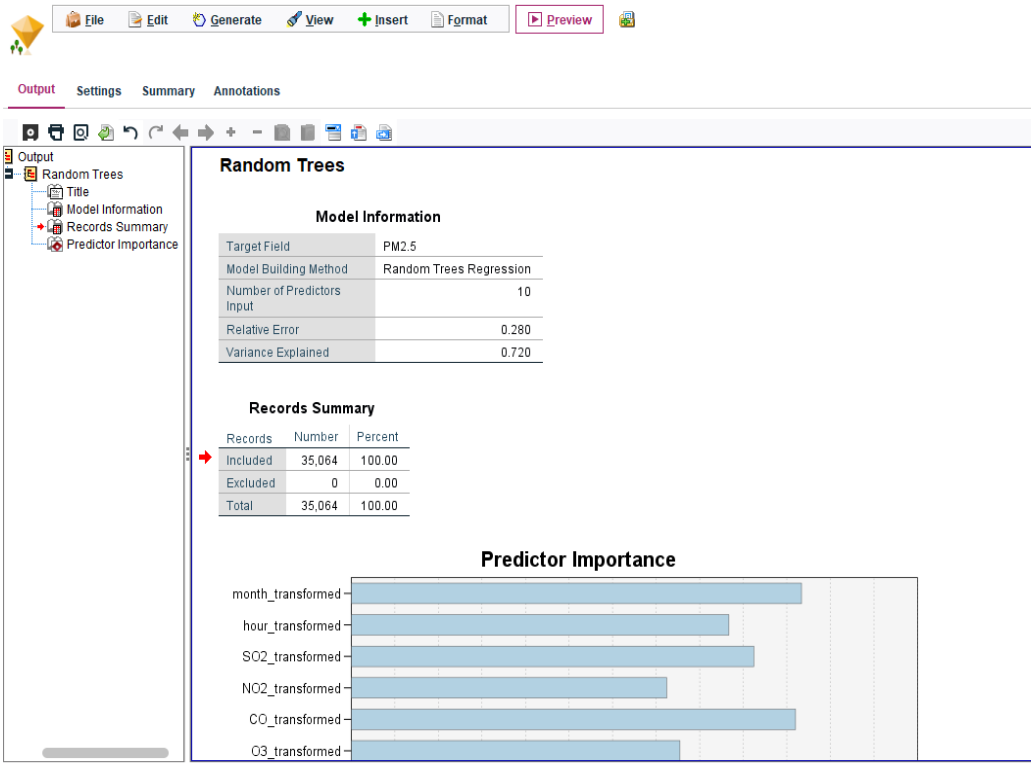
The SPSS modeler could generate their test model. The neural networks model has been set as 70/30 separate models to prevent the overfit. For the
XGBoost tree model's and Random Forest model's settings, there are also some settings for controlling overfitting and handing an imbalanced dataset.

7.2. Conduct data mining

After completing the preparations, the three models run successfully, and the results would show up, respectively.

Model 1: Random Forests

Figure 33. The output of the Random Forest model – Evidence from SPSS Modeler



Random Trees

Model Information

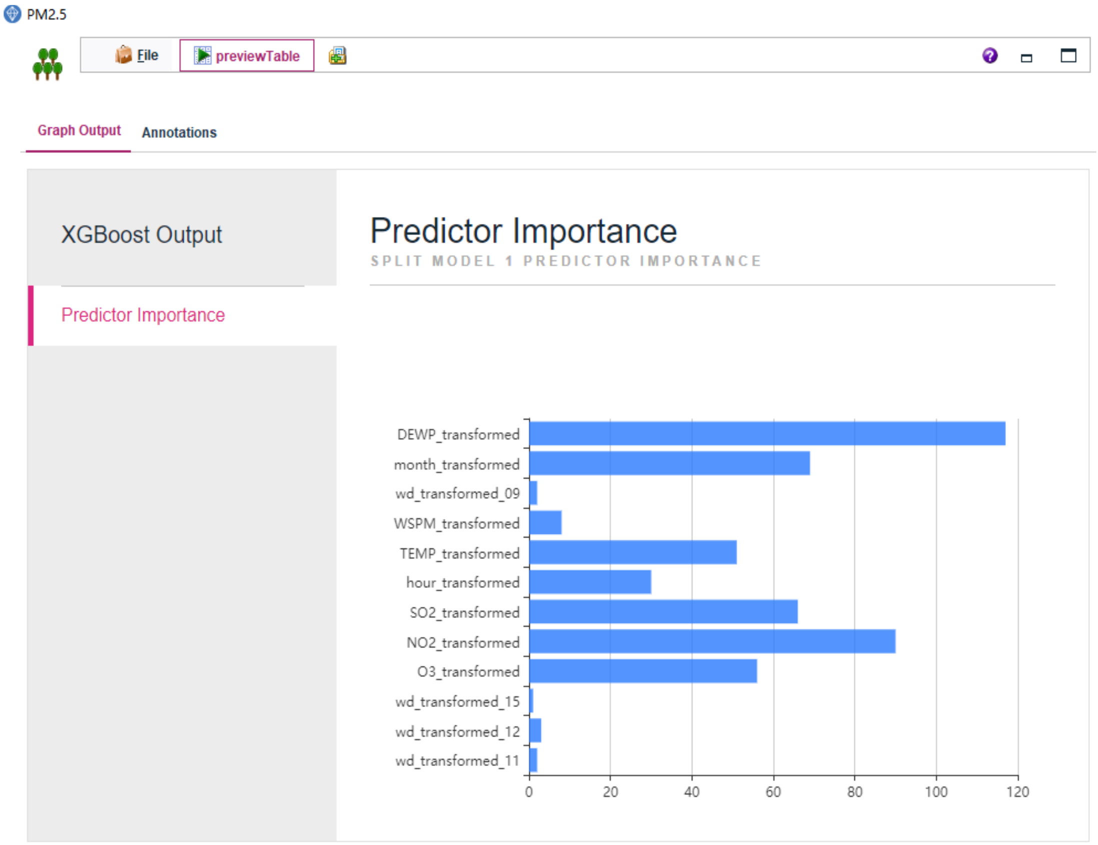
Target Field	PM2.5
Model Building Method	Random Trees Regression
Number of Predictors Input	10
Relative Error	0.280
Variance Explained	0.720

Records Summary

Records	Number	Percent
Included	35,064	100.00
Excluded	0	0.00
Total	35,064	100.00

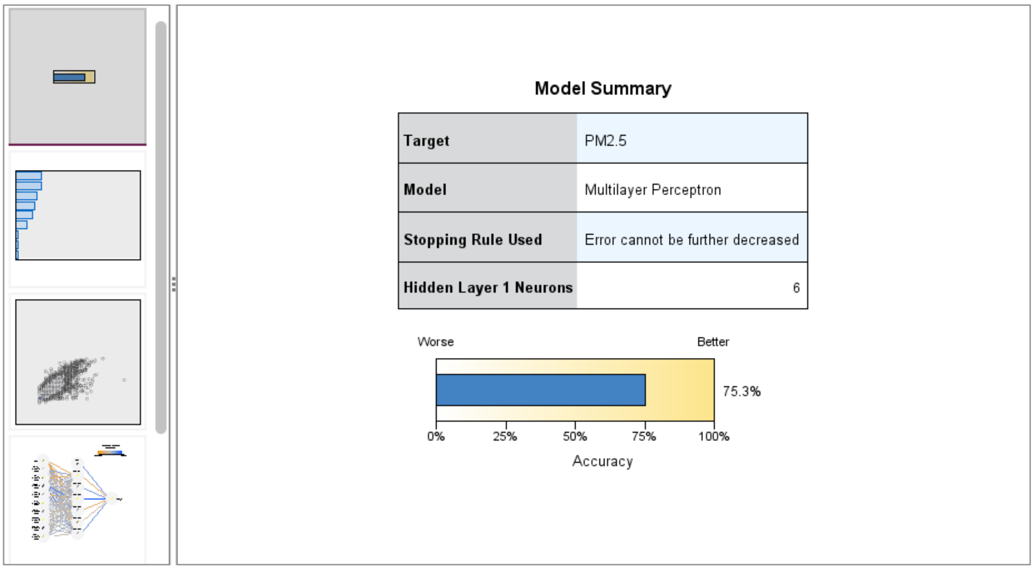
Model 2: XGBoost Tree

Figure 34. The output of XGBoost Tree model – Evidence from SPSS Modeler



Model 3: Neural networks

Figure 35. The output of the Neural networks model – Evidence from SPSS Modeler



Model Summary

Target	PM2.5
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	6

In the Random Forest model, the variance explained shows 0.720, and the relative error is 0.280.

XGBoost Tree model provided each predictors' importance to the target.

In the Neural networks model, the accuracy shows 73.8%.

The results from the three models are satisfying and straightforward for further analysis.

7.3. Search for patterns

Data mining mode – mined pattern

SPSS Modeler provides many patterns, and they are beneficial for beginners. The objectives and conditions of data mining patterns are apparent. It may reduce the time for completing the project.

Using the "feature selection" node to automatically select the essential and relevant predictors, as mentioned in step 4, provides the models' valuable and high correlation attributes.

In step 6, access the "Auto Data Prep" node, use the rescaling method to transform predictors' data by limiting the minimum(0.0) and the maximum(100.0).

Simultaneously, use the "Auto Numeric" node to automatically find out the most effective algorithms: the Random Forest model, the XGBoost tree model, and the Neural network model for the regression problems.

The output and results of the three models could be found in the following graphs.

*Figure 36. Predictor importance of **Random Forest model** – Evidence from SPSS Modeler*

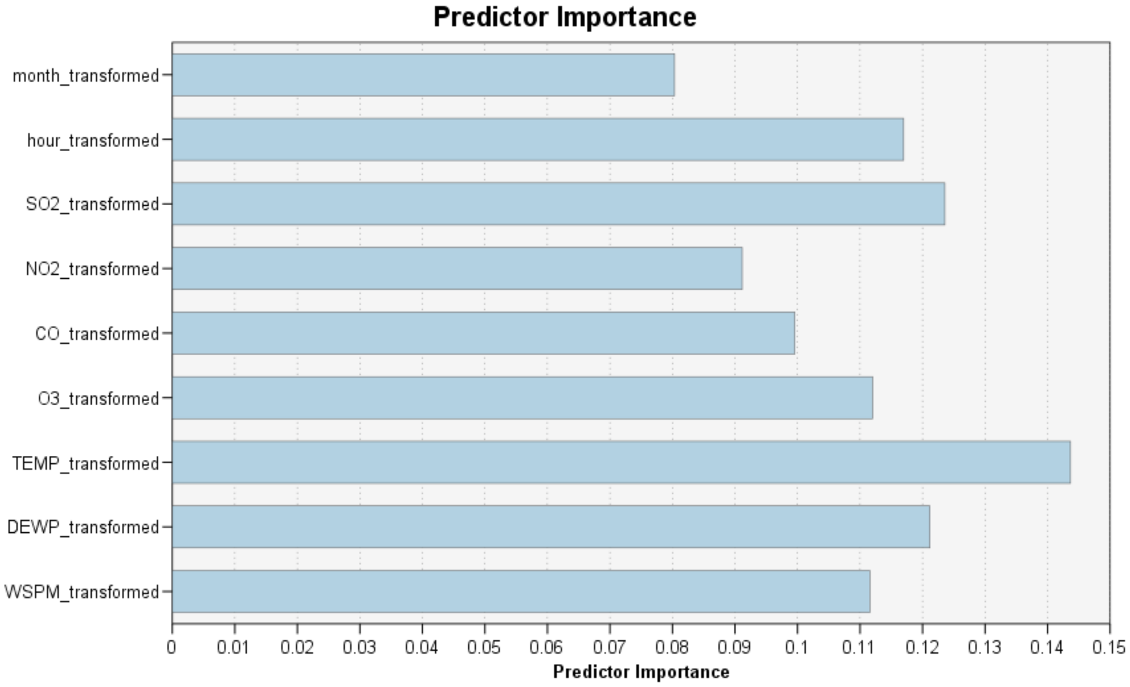
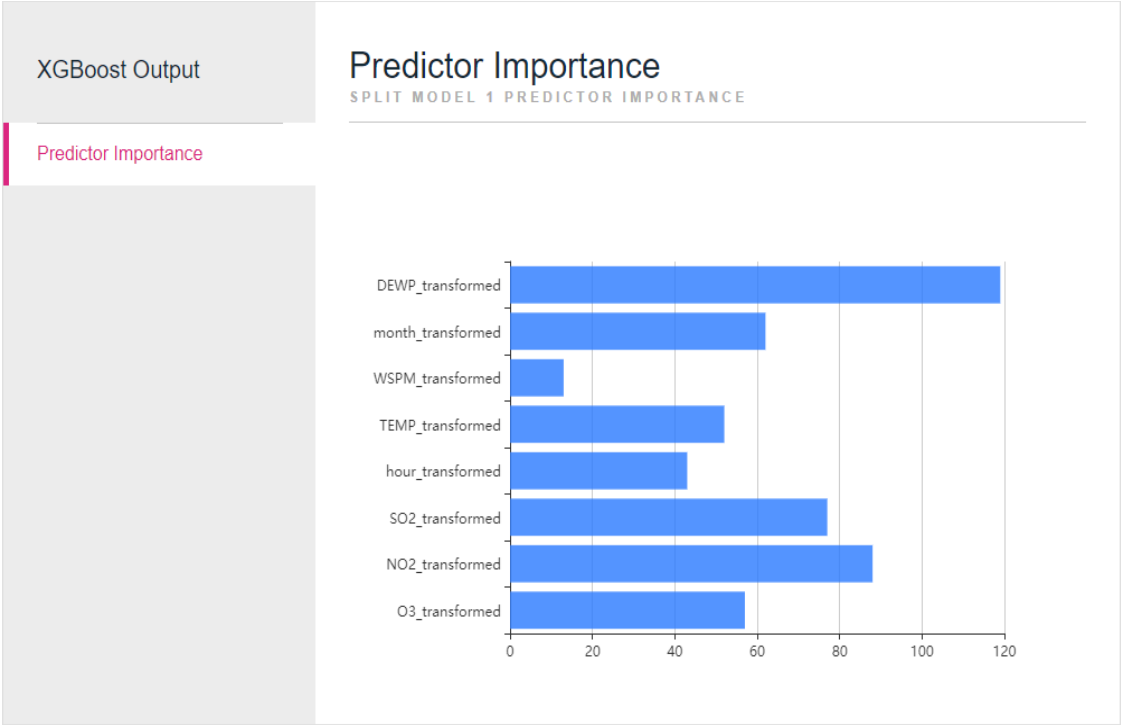


*Figure 37. Predictor importance of **XGBoost Tree model** – Evidence from SPSS Modeler*

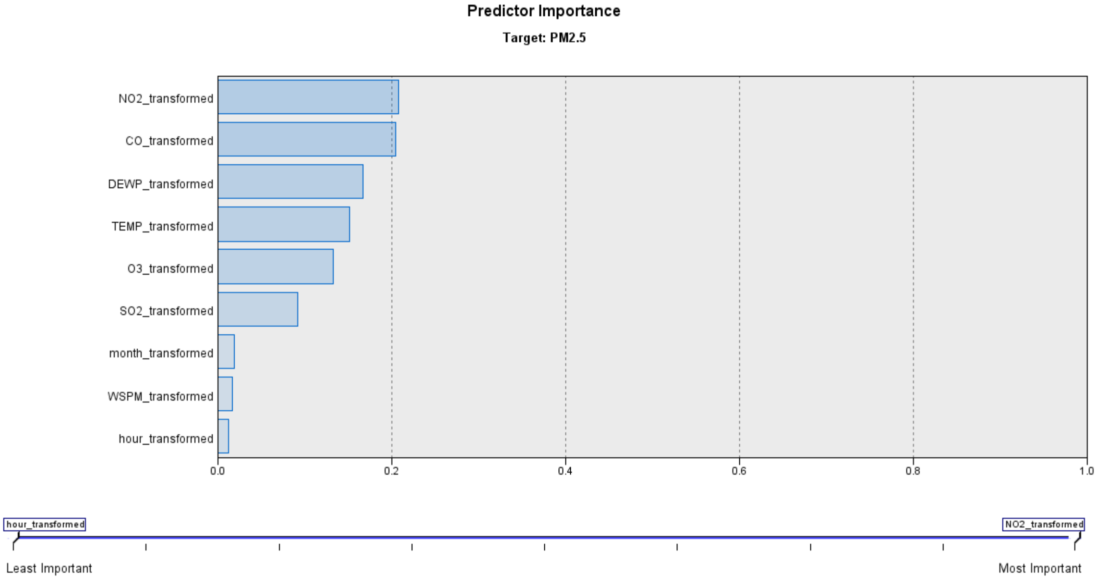


The essential predictor based on Random Forest's output is the temperature ('TEMP'), with the number of importance is 0.143 approximately (Figure 36).

Additionally, based on the output of the XGBoost Tree model, the most significant predictor is the Dew point temperature ('DEWP'); the second is nitrogen dioxide('NO2') (Figure 37).

Meanwhile, based on the output of the Neural Network model, the most important predictor is the concentration of nitrogen dioxide('NO2'); the second significant predictor is the Dew point temperature ('DEWP') (Figure 38).

Figure 38. *Predictor importance of* ***Neural Networks model*** *– Evidence from SPSS Modeler*



8. Interpretation

8.1. Study and discuss the mined patterns

The missing data’s replacement helped construct the complete relationship between the independent variables and the dependent variable. Before the data preparation, there are a few portions of space or NA value data for some fields. Replacing the missing data helps us to delete the noise and make every instance useful.

It is necessary to generate important and valid attributes before running the models. ‘ Feature selection’ node helped us to filter unnecessary or less critical features. The model could focus on the importance of the attributes that the ‘feature selection’ remained the significant relative predictors.

While selecting the model and algorithms, each model’s correlation is more significant than 0.85(see figure 22), after the analysis by the ‘Auto Numeric’ node. These high correlations help us to decide the algorithms and keep moving on the study.

Based on the discussions in step five(Data-mining methods selection) and step six(Data-mining algorithms selection), this project’s main pattern is **regression**. By using the models in SPSS Modeler, some significant predictors have shown in the result, including the temperature (‘TEMP’), shows in figure 36; Dew point temperature (‘DEWP’), shows in figure 37; nitrogen dioxide(‘NO2’), shows in figure 38.

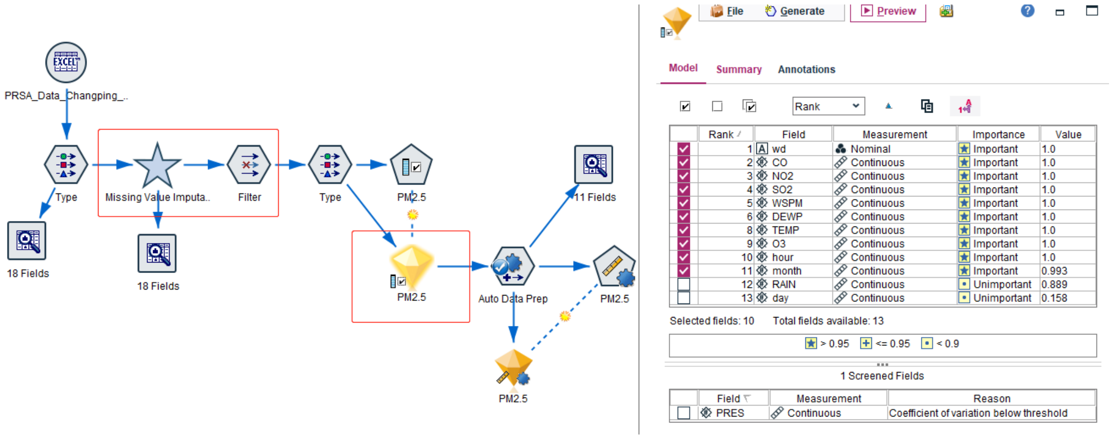
Overall, the predictor importance rate in each model shows significant results. Dew point temperature and nitrogen dioxide could be the primary input of predicting the concentration of PM2.5.

8.2. Visualize the data, results, models, and patterns

8.2.1. Data visualization:

All the data has been modified after step three(Data preparation) and step four(Data transformation). The data using for algorithms and models are all been transformed and selected by “Missing Value Imputation,” “Filter” node, “Feature selection” node. Figure 39 shows the execution of processing the data. Figure 40 shows the fields after processed.

Figure 39. Processing the data – Evidence from SPSS Modeler steam DMAS



	Rank	Field	Measurement	Importance	Value
☑	1	wd	Nominal	Important	1.0
☑	2	CO	Continuous	Important	1.0
☑	3	NO2	Continuous	Important	1.0
☑	4	SO2	Continuous	Important	1.0
☑	5	WSPM	Continuous	Important	1.0
☑	6	DEWP	Continuous	Important	1.0
☑	8	TEMP	Continuous	Important	1.0
☑	9	O3	Continuous	Important	1.0
☑	10	hour	Continuous	Important	1.0
☑	11	month	Continuous	Important	0.993
□	12	RAIN	Continuous	Unimportant	0.889
□	13	day	Continuous	Unimportant	0.158

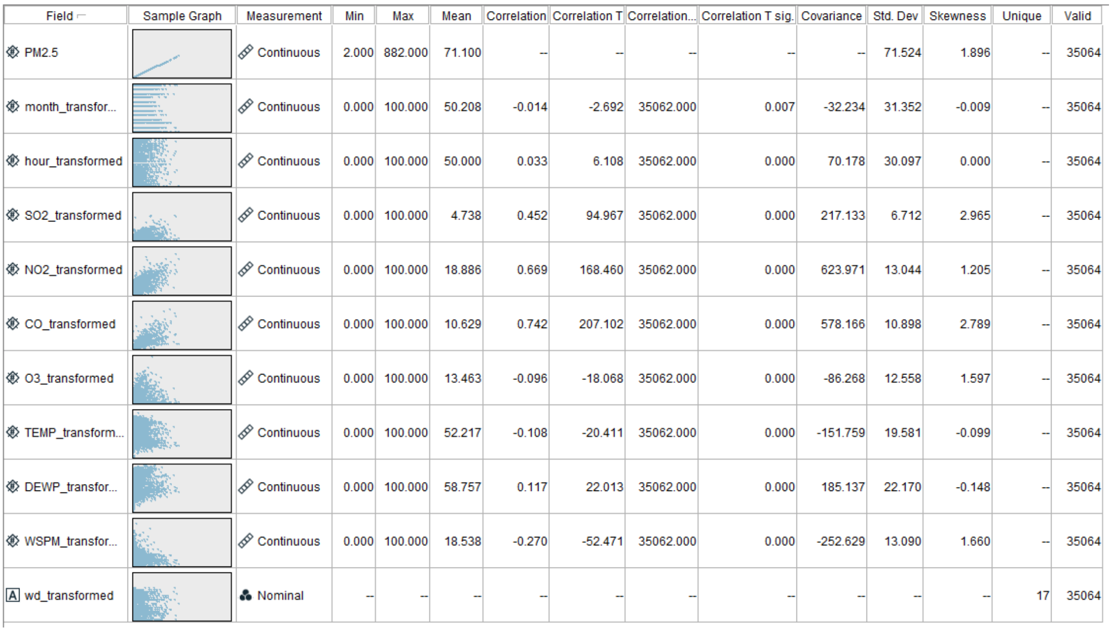
Selected fields: 10 Total fields available: 13

★ > 0.95 + <= 0.95 • < 0.9

1 Screened Fields

	Field	Measurement	Reason
□	PRES	Continuous	Coefficient of variation below threshold

Figure 40. Visualization of the data(the fields using in the models) – Evidence from SPSS Modeler



Field	Sample Graph	Measurement	Min	Max	Mean	Correlation	Correlation T	Correlation...	Correlation T sig.	Covariance	Std. Dev	Skewness	Unique	Valid
PM2.5		Continuous	2.000	882.000	71.100	--	--	--	--	--	71.524	1.896	--	35064
month_transfor...		Continuous	0.000	100.000	50.208	-0.014	-2.692	35062.000	0.007	-32.234	31.352	-0.009	--	35064
hour_transformed		Continuous	0.000	100.000	50.000	0.033	6.108	35062.000	0.000	70.178	30.097	0.000	--	35064
SO2_transformed		Continuous	0.000	100.000	4.738	0.452	94.967	35062.000	0.000	217.133	6.712	2.965	--	35064
NO2_transformed		Continuous	0.000	100.000	18.886	0.669	168.460	35062.000	0.000	623.971	13.044	1.205	--	35064
CO_transformed		Continuous	0.000	100.000	10.629	0.742	207.102	35062.000	0.000	578.166	10.898	2.789	--	35064
O3_transformed		Continuous	0.000	100.000	13.463	-0.096	-18.068	35062.000	0.000	-86.268	12.558	1.597	--	35064
TEMP_transform...		Continuous	0.000	100.000	52.217	-0.108	-20.411	35062.000	0.000	-151.759	19.581	-0.099	--	35064
DEWP_transfor...		Continuous	0.000	100.000	58.757	0.117	22.013	35062.000	0.000	185.137	22.170	-0.148	--	35064
WSPM_transfor...		Continuous	0.000	100.000	18.538	-0.270	-52.471	35062.000	0.000	-252.629	13.090	1.660	--	35064
wd_transformed		Nominal	--	--	--	--	--	--	--	--	--	--	17	35064

8.2.2. Model visualization:

Figure 41 shows the model visualization of the models using in this project. The following graphs (figure 42, 43, 44, 45) show the model selection and the full plot chart of every model.

Figure 41. Model visualizations – Evidence from SPSS Modeler

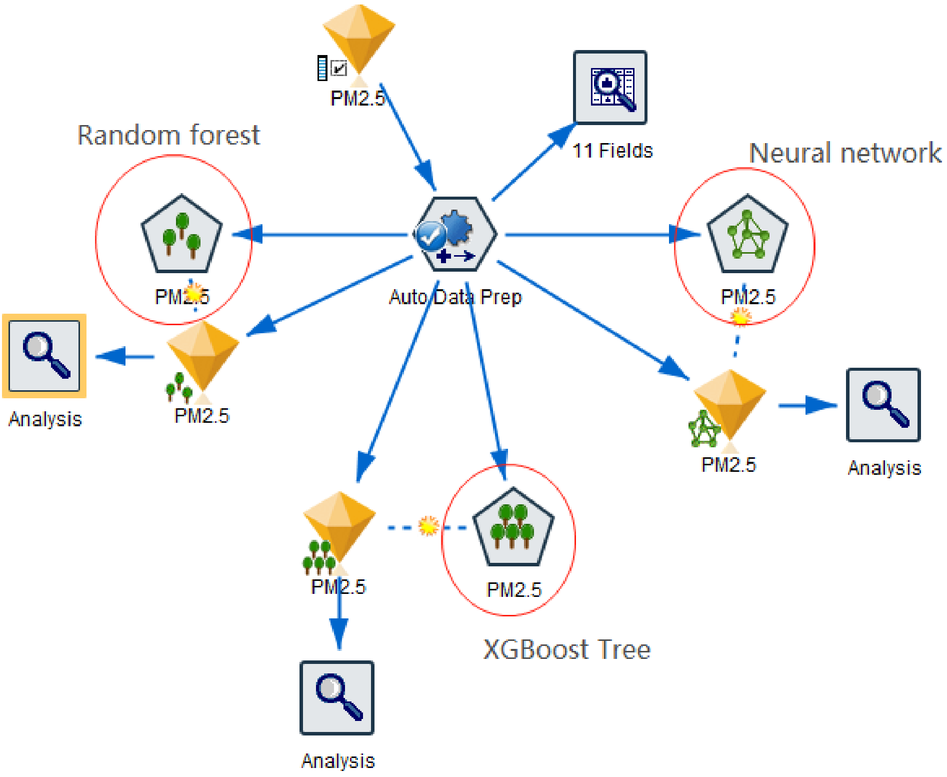
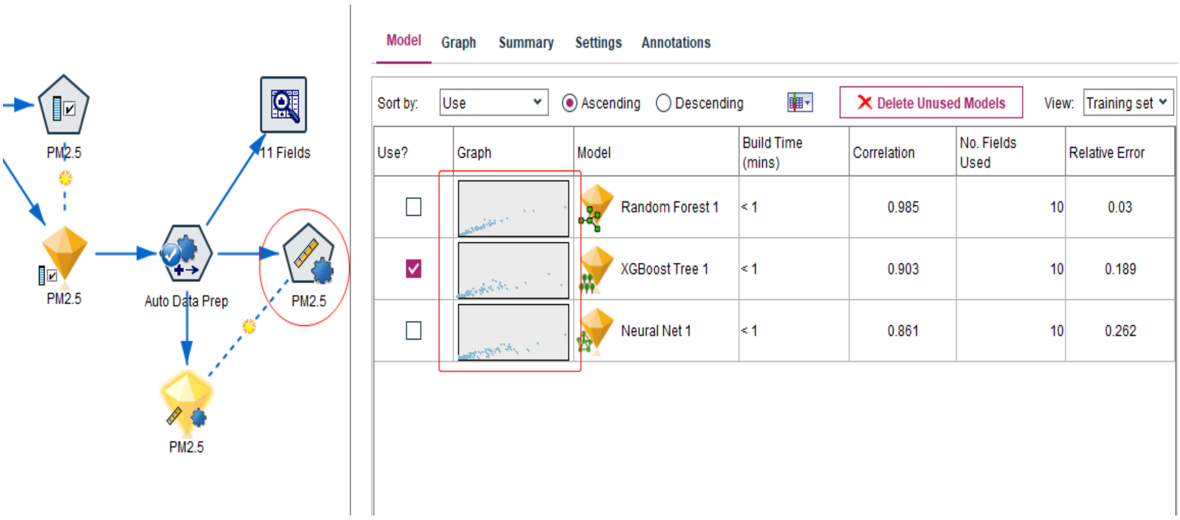


Figure 42. The output from 'Auto Numeric' node – Evidence from SPSS Modeler



Sort by: Use ● Ascending ○ Descending View: Training set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
□		Random Forest 1	< 1	0.985	10	0.03
☑		XGBoost Tree 1	< 1	0.903	10	0.189
□		Neural Net 1	< 1	0.861	10	0.262

Figure 43. Visualization of the full plot of PM2.5 against RL-PM2.5(after generated by Random Forest) – Evidence from SPSS Modeler

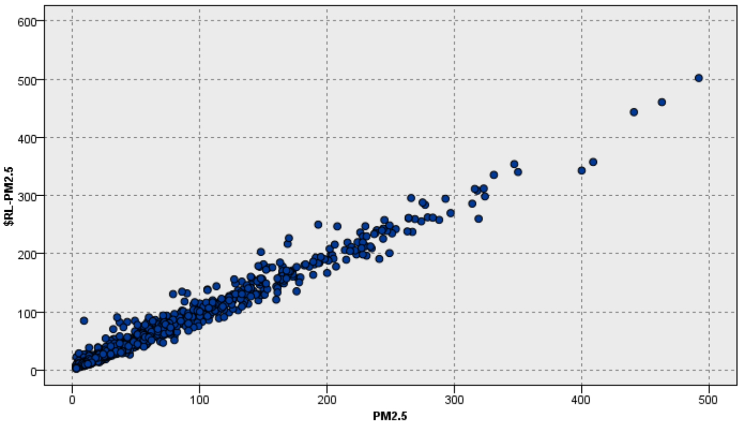


Figure 44. Visualization of the full plot of PM2.5 against XGT-PM2.5(after generated by XGBoost Tree) – Evidence from SPSS Modeler

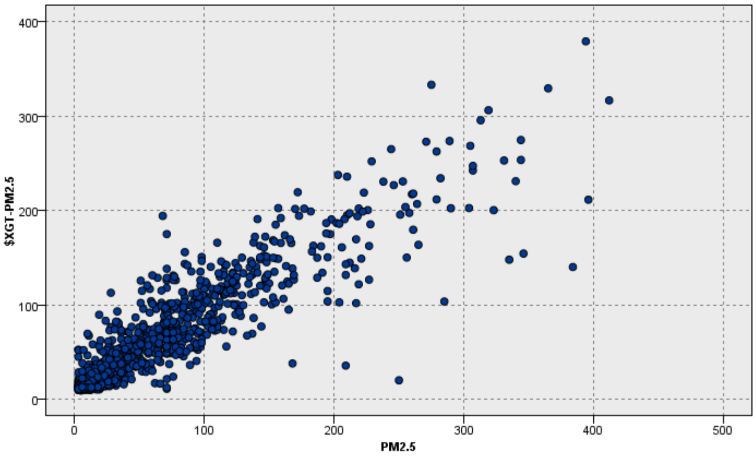
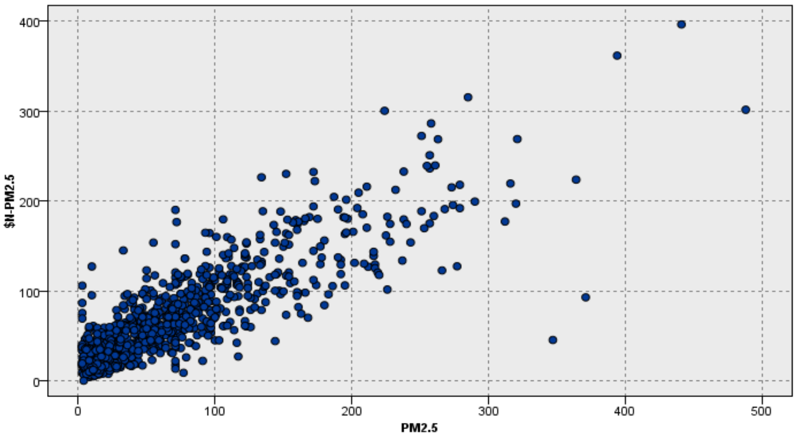


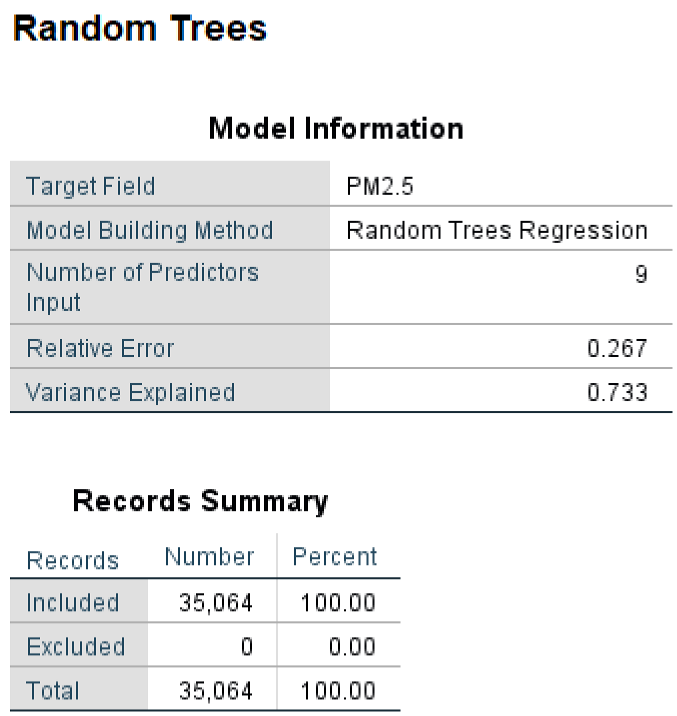
Figure 45. Visualization of the full plot of PM2.5 against N-PM2.5(after generated by Neural Network) – Evidence from SPSS Modeler



8.2.3. Results and pattern visualization:

In figure 46 - 48, the details show the output of the Random Tree model. The work from the XGBoost Tree model has shown in figure 49, 50. The output from the Neural Network model has shown in figure 51 - 56.

*Figure 46. Results from **Random Tree** model – Evidence from SPSS Modeler*



Random Trees

Model Information

Target Field	PM2.5
Model Building Method	Random Trees Regression
Number of Predictors Input	9
Relative Error	0.267
Variance Explained	0.733

Records Summary

Records	Number	Percent
Included	35,064	100.00
Excluded	0	0.00
Total	35,064	100.00

*Figure 47. Results from **Random Tree** model – Evidence from SPSS Modeler*

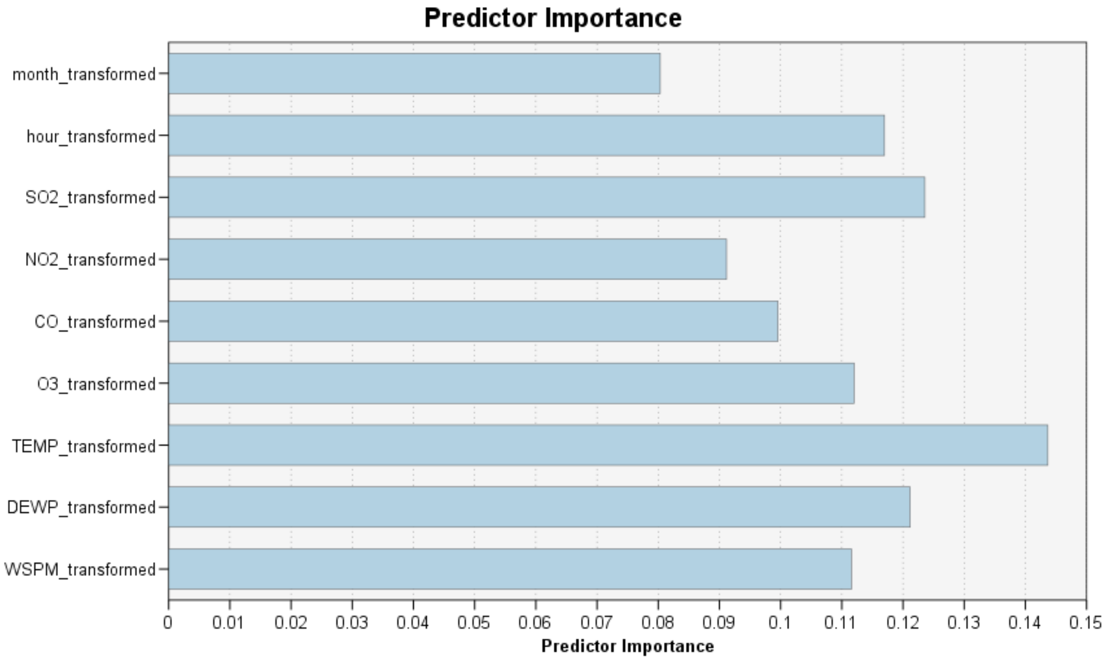
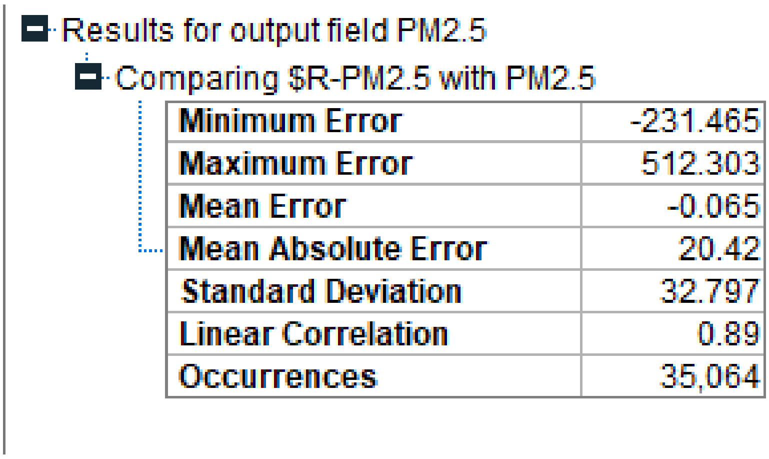


Figure 48. Analysis of result for output of ***Random Forest*** *– Evidence from SPSS Modeler*



Results for output field PM2.5

Comparing $R-PM2.5 with PM2.5

Minimum Error	-231.465
Maximum Error	512.303
Mean Error	-0.065
Mean Absolute Error	20.42
Standard Deviation	32.797
Linear Correlation	0.89
Occurrences	35,064

Figure 49. Result of ***XGBoost Tree*** *model – Evidence from SPSS Modeler*

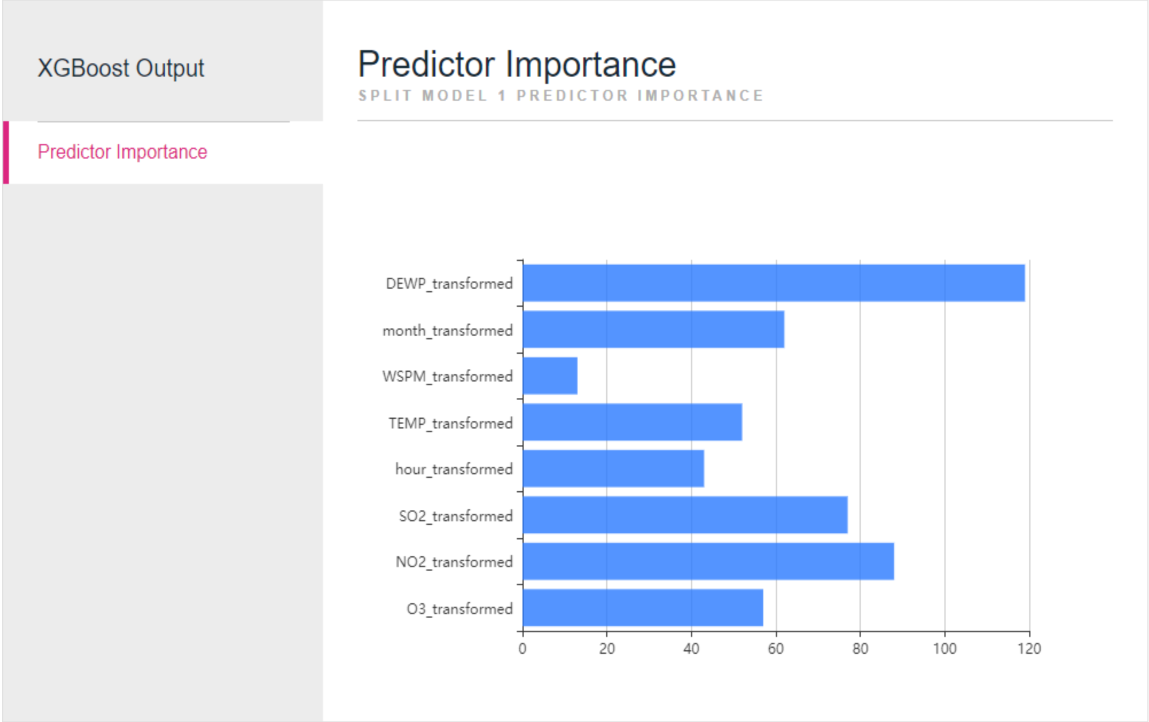
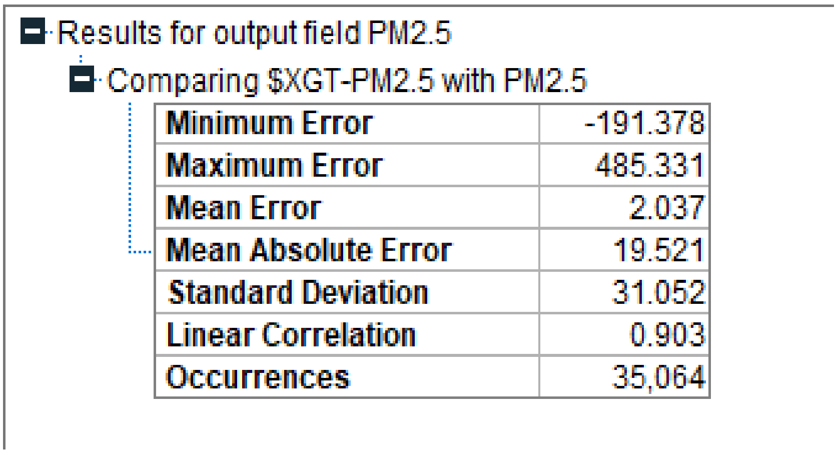


Figure 50. Analysis of result for output of ***XGBoost Tree*** *– Evidence from SPSS Modeler*

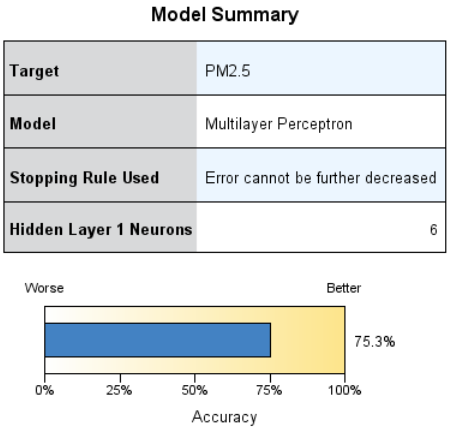


Results for output field PM2.5

Comparing $XGT-PM2.5 with PM2.5

Minimum Error	-191.378
Maximum Error	485.331
Mean Error	2.037
Mean Absolute Error	19.521
Standard Deviation	31.052
Linear Correlation	0.903
Occurrences	35,064

Figure 51. Result of ***Neural Network*** *model – Evidence from SPSS Modeler*



Model Summary

Target	PM2.5
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	6

Figure 52. Result of ***Neural Network*** *model – Evidence from SPSS Modeler*

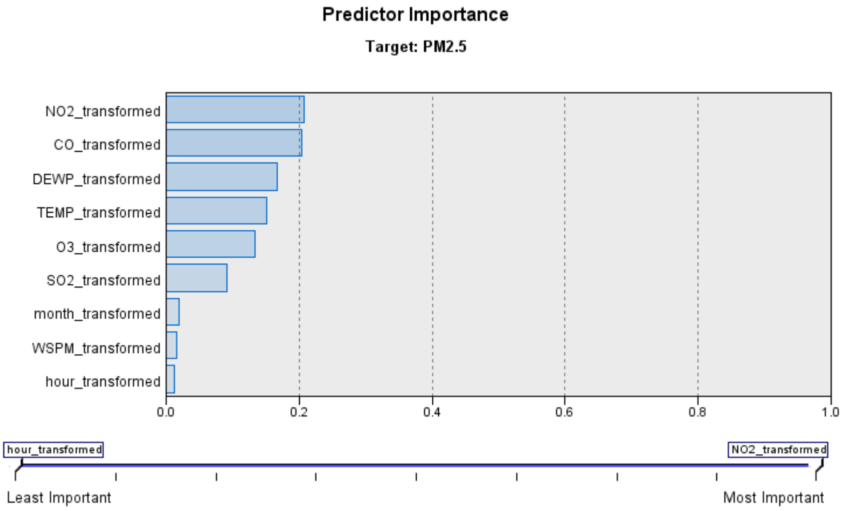


Figure 53. Result of ***Neural Network*** *model – Evidence from SPSS Modeler*

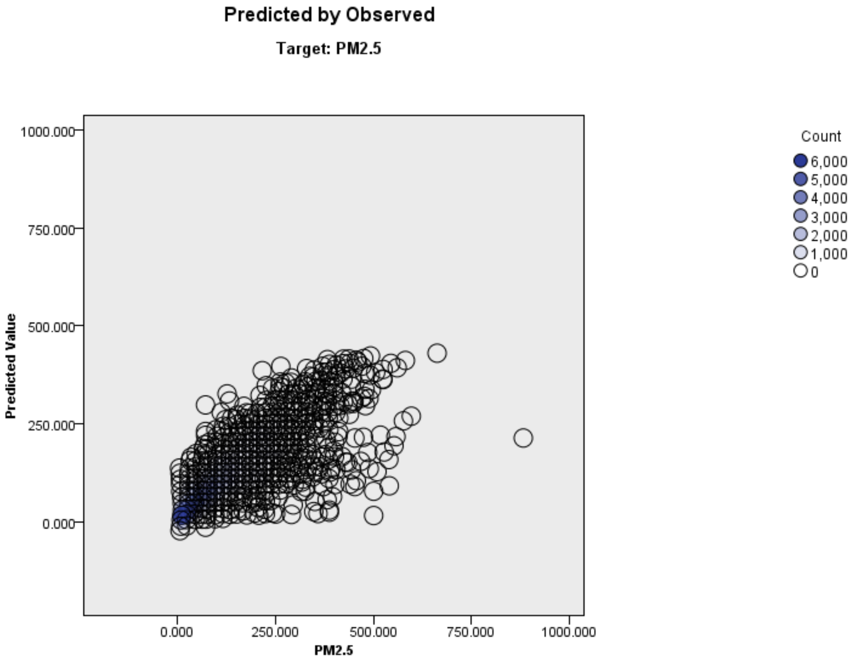


Figure 54. Result(effects) of ***Neural Network*** *model – Evidence from SPSS Modeler*

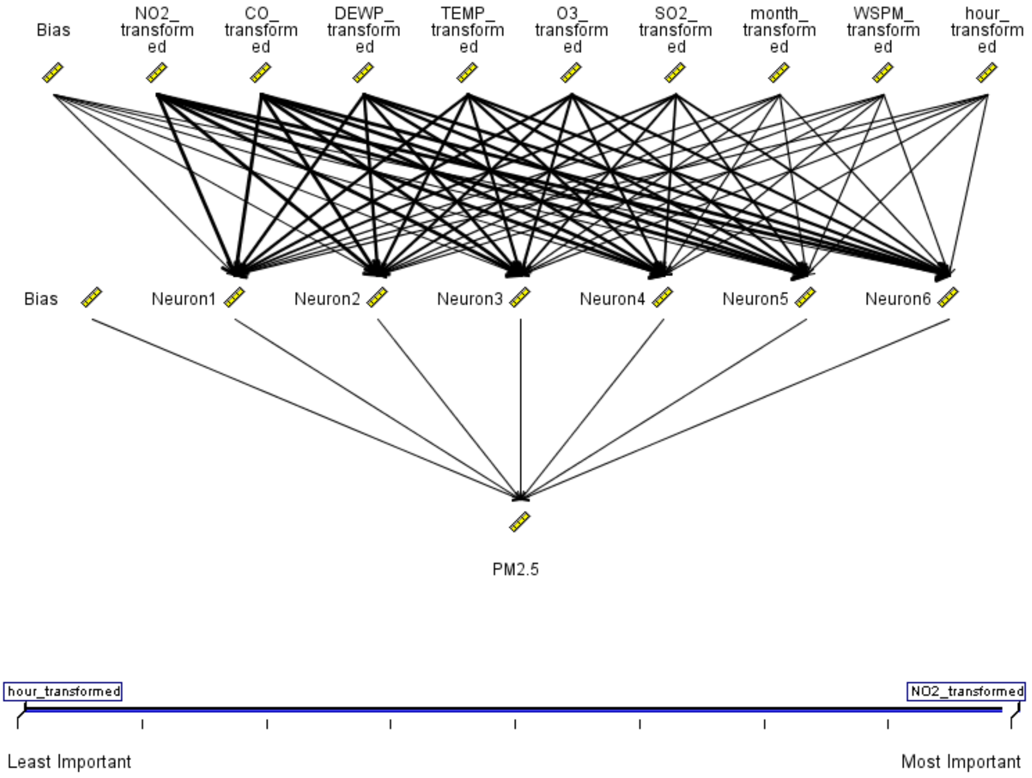


Figure 55. Result(coefficients) of the ***Neural Network*** *model – Evidence from SPSS Modeler*

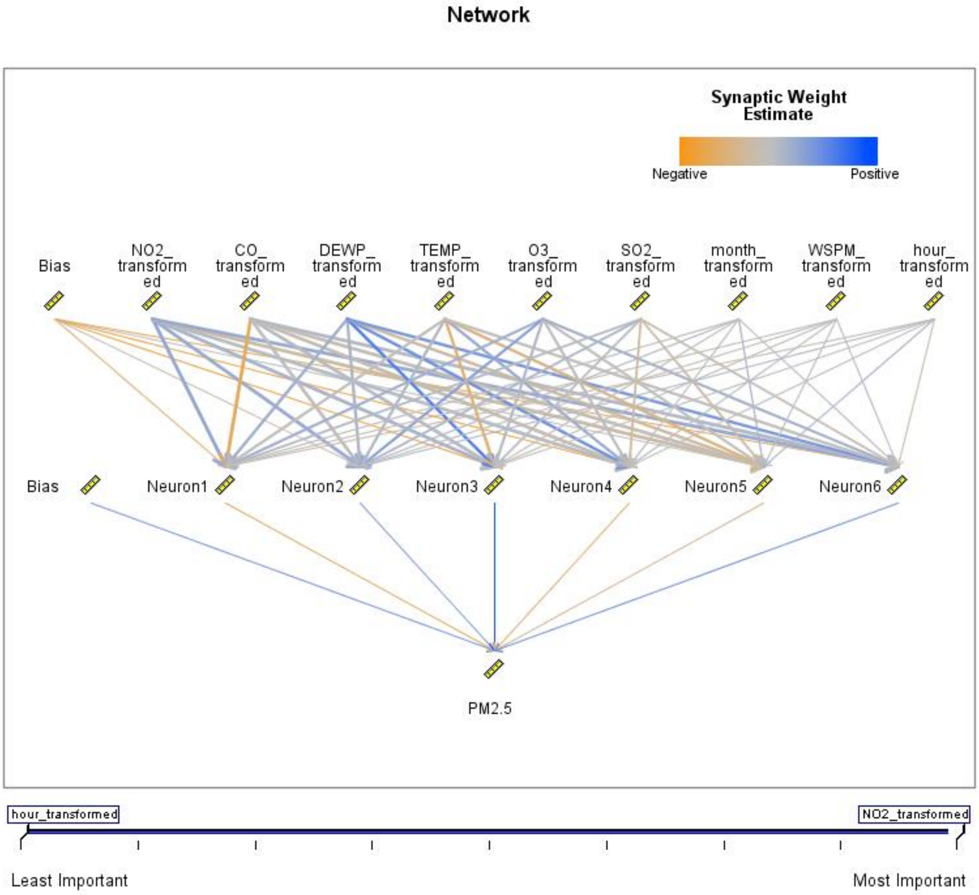
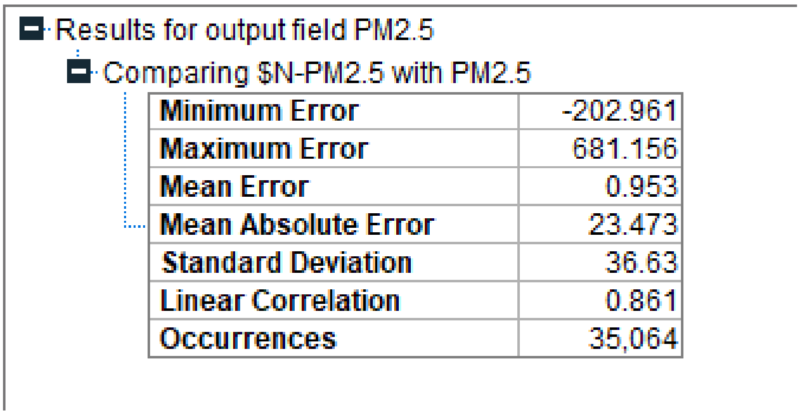


Figure 56. Analysis of result for output of ***Neural Network*** *– Evidence from SPSS Modeler*



Results for output field PM2.5

Comparing $N-PM2.5 with PM2.5

Minimum Error	-202.961
Maximum Error	681.156
Mean Error	0.953
Mean Absolute Error	23.473
Standard Deviation	36.63
Linear Correlation	0.861
Occurrences	35,064

8.3. Interpret the results, models, and patterns

8.3.1. Results and models Interpretation

- Random Tree model:

  In figure 46 - 48, the details show the output of Random Tree model.

  The correlation of the Random Forest model (with the target is PM2.5) is 0.985, which is the highest correlation of the three models.

  In this model, the most important predictor is 'TEMP' (temperature), with the figure is 0.143; the second significant element is SO2, with the figure is 0.123; the 'DEWP' (Dew point temperature) is also on the second level of substantial predictor, with the model is 0.121 approximately. The effect of other chemical elements also plays a role, the value of NO2, CO, O3 is 0.092, 0.10, 0.113, respectively. The time variables such as month and hours, show a big difference, month predictor becomes the least important attribute, with the figure is 0.08; however, the value of hours predictor is estimated at 0.17.

  Figure 48 shows the result from the 'analysis' node for this model offers the Mean Error is -0.052, and Standard Deviation is 32.281, Linear correlation 0.893, which is satisfied. The Occurrences of the dataset are 35064, with no data missed.

- XGBoost Tree model:
  The output from the XGBoost Tree model has shown in figure 49, 50.
  The correlation of the XGBoost Tree model (with the target is PM2.5) is 0.903, which is suitable for this dataset.
  In this model, the most important predictor is 'DEWP' (Dew point temperature), with the figure is 119 (120 in total); the second most significant predictor is 'NO2' (nitrogen dioxide), with the figure is 88; the lowest figure is 13, which is the 'WSPM' (Wind speed), most unimportant predictor in this model. The influence of chemical elements on predictions is also important, the value of SO2 and O3 is 77, 57.
  Additionally, the time influencing makes some changes, the month predictor becomes the fourth in this model, with 62 predictor importance, also the hour predictor increases its' level a bit against the Random Forest model, with the figure 43.

In figure 50, the result from the 'analysis' node for the XGBoost Forest model shows the Mean Error is 2.038, and Standard Deviation is 31.267, Linear correlation 0.901, which is the highest compared to the other two models. The Occurrences of the dataset is 35064, no data missed.

- Neural Network model:

  The output from the Neural Network model has shown in figure 51 - 56.

  The correlation of the Neural Network model (with the target is PM2.5) is 0.861, which is suitable for this dataset. This model's accuracy is 75.3%, which is a satisfying result (shown in figure 51). As regards to the predictor importance of this model (see figure 52), the highest-ranking is the NO2, with the percentage is 21%, is the most important predictor; the second highest-rated is CO, with the rate is 20%; the third and fourth is the DEWP and TEMP (temperature features), with the percentage 17% and 15% respectively; O3 (13%) and SO2 (9%) are the rest of chemical elements. It is worth noting that the time features do not show enough importance in this model, with only 2% and 1% for month and hour, respectively.

  In figure 53, the plot chart shows the predicted by observed in this model. There are few outliers shown in this chart; however, based on the large dataset we have (35064 data), this chart's normality is satisfied.

  In figure 54, the network effect chart shows the different importance between each predictor, and It has the same result with the predictor importance shown in figure 51, the NO2 is the most important predictor in this model. Meanwhile, the network coefficients chart shows the synaptic weight estimate, and the DEWP has three light blue lines connected with the target neurons; however, the most important predictor is NO2.

  In figure 56, the result from the 'analysis' node for the XGBoost Forest model shows the Mean Error is 0.196, relatively small value for the error, and Standard Deviation is 35.523, Linear correlation 0.868, which is acceptable. The Occurrences of the dataset is 35064, no data missed.

8.3.2. patterns Interpretation

1. The missing data's replacement helped construct the complete relationship between the independent variables and the dependent variable. Before the data preparation, there are a few portions of empty value or NA value data for some fields. Replacing the missing data helps us to delete the noise and make every instance useful.

2. It is necessary to generate essential and valid attributes before running the models. ' Feature selection' node helped us to filter unnecessary or less critical features. The model could focus on the importance of the attributes that the 'feature selection' remained the significant relative predictors.

3. While selecting the model and algorithms, each model's correlation is more significant than 0.85(see figure 22), after the analysis by the 'Auto Numeric' node. These high correlations help us to decide the algorithms and keep moving on the study.

8.4. Assess and evaluate results, models, and patterns

8.4.1. Assess and evaluate the results and models:

Each model's unit is different; however, the most important predictors are around the DEWP, and NO2, which is clearly showing the results. However, the impact of different models has demonstrated that the DEWP (dew point temperature) is an important predictor with PM2.5.
In terms of the data mining success criteria, we have found that the DEWP temperature is one of the most significant feature associates with the concentration of PM2.5, the chemical elements in this dataset (SO2, O3, NO2, CO) also became the most critical predictor in some of the models, especially the NO2.

The time variables, month, and hour also play a role in the models' results.

In conclusion, in my opinion, the PM2.5 concentration is associated with the DEWP temperature, as the lower DEWP is, the higher concentration of PM2.5 is. Similarly to the month and hour features, the concentration of PM2.5 is much higher in winter and spring (from December to May) than the other two seasons. Simultaneously, the temperature at night is lower than the temperature in the daytime; consequently, the concentration of PM2.5 would higher in the night.

8.4.2. Assess and evaluate the patterns

Using the "feature selection" node to select the essential and relevant predictors automatically, as mentioned in step 4, provides the valuable and high correlation attributes for models.

In step 6, access the "Auto Data Prep" node, use the rescaling method to transform predictors' data by limiting the minimum(0.0) and the maximum(100.0).

At the same time, use the "Auto Numeric" node to find out the most significant algorithms automatically, which are the Random Forest model, XGBoost tree model, and Neural network model for the regression problems.

Based on the discussions in step five(Data-mining methods selection) and step six(Data-mining algorithms selection), this project's main pattern is **regression**. By using the models in SPSS Modeler, some important predictors have shown in the result, including the temperature ('TEMP'), shows in figure 36; Dew point temperature ('DEWP'), shows in figure 37; nitrogen dioxide('NO2'), shows in figure 38.

Overall, the predictor importance rate in each model shows significant results. Dew point temperature and nitrogen dioxide could be the main input of predicting the concentration of PM2.5.

8.5. Iterate prior steps (1 – 7) as required

8.5.1. Iterate business/situation understanding

The PM2.5 weather situation in Changping(Beijing) would not change significantly for the time being, and the project prediction target would not change. Additionally, train the data set, build, and generate the model and predict the PM2.5 concentration according to the other relevant weather index.

8.5.2. data understanding

The data set is still the weather situation in Changping(Beijing). The data quality can be observed through the specific analysis and understanding of the data by processing the database.

8.5.3. Data preparation

Processing the missing values and finding the desired data column against the target could ensure data prediction quality.

8.5.4. Data transformation

According to the target, filter the unwanted rows, then convert the used columns to the agreed range to ensure fairness.

8.5.5. Data-mining method selection

Because the variables in the project are both input and output, supervised learning is required. The regression models are appropriate for our project because the prediction target is continuous.

8.5.6. Data-mining algorithms selection

SPSS modeler provides several reliable algorithms. After investigation, all these algorithms show their advantages. It is necessary to compare the results of each algorithm to verify our prediction.

8.5.7. Data-mining

Data mining results are as follows: for the random forest model, CO is the most critical factor affecting PM2.5; For the XGBoost Tree model, the temperature is the most crucial factor affecting PM2.5.

These steps have not changed compared to before; this project has been iterated many times and repeated various measures to ensure that the model is significant.

References

Data for Data Mining. (n.d.). Principles of Data Mining, 11–21. Doi: 10.1007/978-1-84628-766-4_1.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37-54.

IISD. (2018). *WHO Global Conference Recommends Reducing Deaths from Air Pollution by Two-Thirds by 2030.* Retrieved from https://sdg.iisd.org/news/who-global-conference-recommends-reducing-deaths-from-air-pollution-by-two-thirds-by-2030/

IQAIR. (2020). *2019 WORLD AIR QUALITY REPORT*. Retrieved from https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf

Random Forest, GBDT and XGBoost. (n.d.). Retrieved from https://blog.csdn.net/yingfengfeixiang/article/details/80210145?utm_medium=distribute.pc_relevant.none-task-blog-BlogCommendFromBaidu-3

Regression Trees. (2013). *Regression Methods for Medical Research*, 204-235. doi:10.1002/9781118721957.ch9

Team Airveda. (2017). *What Is PM2.5 and Why Is It Important?* Retrieved from https://www.airveda.com/blog/what-is-pm2-5-and-why-is-it-important

World Health Organization. (2018). *Ambient (outdoor) air pollution.* Retrieved from https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

World-wide Air Quality Monitoring Data Coverage. (n.d.). Retrieved from https://aqicn.org/sources/

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z. & Chen, S.X. (2017). Cautionary Tales on Air-Quality Improvement in Beijing. *Proceedings of the Royal Society A, 473*(2205), 20170457.

Sathya, R., & Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence, 2*(2). doi:10.14569/ijarai.2013.020206

Seif, G. (2018, March 5). Selecting the best Machine Learning algorithm for your regression problem. Retrieved from https://towardsdatascience.com/selecting-the-best-machine-learning-algorithm-for-your-regression-problem-20c330bad4ef

"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.).

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."