## **Air Pollution in Seoul**

Chiu Fan Hui

## 1 Problem Statement and Background

#### 1.1 Problem Statement

Recently, the air quality is getting worse and worse. With many serious air pollutions, people have become more aware of the importance of air protection. Among them, PM2.5 is one of the important factors that cause haze. Only when the content of PM2.5 in the air drops, can the haze situation be alleviated. Therefore, the dataset used in this study is the 2017-2019 air quality data of Seoul, South Korea. The main purpose of this study is to analyze and predict PM2.5 in district 105, so as to determine the changes in PM2.5 in the next 200 days.

### 1.2 Background

The initial data contains 647511 data including 11 attributes such as address, geographic location and pollutants (Figure 1). The pollutants contained in the Dataset are CO, NO2, O3, PM10, PM2.5 and SO2, and the monitoring interval is one hour. Therefore, it is necessary to perform data clean before analyzing this set of data to remove redundant data, thereby simplifying data and reducing operational complexity.

- 647/31 recent parties 0010/2021 (E20-62000) - No Scient Sampling *									W 300 Y		A Town Mode *
Europa Potantin Statistics (647,516) Visualization											
20 For Fags * / Format Process *									1 - 1 2	1.4.5.4	T 8 Not?
603/633 T	7 (01.7	SHIM F. Y	Degrate F. P.	Memorana Ada 1	0.	NOT 2 /	011.7	PROD 1 /	PNOST	1071	States and 3 of
rm. Imag-ro Phin-gil. Nagro-go. Socil. Penality of Karea		17,1720164	121.0000075	2917-61-61 09:00		8.007	6.002	19.0	67.0	0.094	191
In Joyce Harall, Joyceps, Soul, Maddin of Name		17.5728164	TET BETTER	2017/01/01 31/00		1.00	8.007	77.0	10.4	9.000	101
th. Jong-ro Stgergil, Jongroups. Sovel, Manuality of Marea		27.1720164	121.0000075	2017-01-01 02:00		8.890	8.002	79.0	10.0	0.094	191
In June of Maryll, Jespens, Small, Namelia of Name		37,5730164	121,0000075	2017/01/01 01:00		8.896	9.00	19.6	18.6	9.000	191
to, Jacques Stgargil, Jorgenya, Secol, Republic of Karea		17,1728164	121,0000075	2917-01-01 04:00		8.691	8,002	10.0	61.0	0.002	101
15. Jugues Margill, Juguergo, Socal, Mandilas of Norwa		37.3128194	323,8696675	Different Month		1.00	9.00	19.8	11.6	0.000	181
16. Jang-re Stgargil, Jorgen-ye. Satel, Republic of Korea		27.1720364	121,0000075	2017/01/01 00:00		1.00	8.002	20.0	10.0	0.002	191
15. Jegres Dgryll, Jegreys, Seal, Sankist of Nova	1.4	37,5720164	121,000075	Section of the		1.86	8.80	75.0	18.6	9.00	101
15, Jang-co Higargil, Jangeorga, Secol, Republic of Sarea		27,1720164	127,0000073	2017-01-01 00:00		9.00	8.000	12.0	10.0	0.004	ter
15. Jagres Egyrgil, Segroups, Seoil, Robblit of Bores		ST. STORISA	323.8656675	207-01-01-01-01		9.80	9.460	19.0	12.6	9.481	787
16. Degree Stgrigtl, Degreegs, Selet, Rockins of Karea	1.0	17.1720164	127.0000075	2017-01-01 19:00		1.90	8.004	79.0	10.0	9.004	107
15, Jacy-on Expensio, Jacymonys, Secul. Republic of Sorme	1.0	37,4730104	525,969875	2017-01-01 11/00		1.00	1,401	10.0	76.0	9.484	181
is, largers Higa-gil, Jospin-ga, Saial, Aspablic of Saras.	1.0	17,1720104	127,6000075	2017-01-01.12100		2.454(000000000000000	A. 001	10.0	79.0	0.004	101
15, Jang-on Elga-gil, Jangue-pa, Secal, Femalist of Seres	1.0	ST-ETTWISE	525,869675	2017-01-01 12-00		4.81	4.801030333333	96.6	76.6	0.00	101
ts, larg-re Higa-gil, Joseph-ga, Secal, Republic of Karea	4.1	17.1700164	127.6650675	2017-01-01 14:00		8.805	A. 000	91.0	79.0	0.000	101
15, Septem Signings, Septemps, Senal, Republic of Sorne	6.0	17,1779164	121,9656675	2017-01-01 15:00		1.00	1.00	87.8	10.0	0.000	101
18, larg-re Mga-gil, Jargu-ga, Senal, Republic of Karea	6.3	17.1700164	127,0050075	2017-01-01.16:00		8.863	0.004	87.0	65.6	9.000	101
16, Jang-re Elga-gil, Sergro-pa, Senal, Resette of Sores	5.0	37,1739104	121,849825	annered than		1.80	8.862	95.0	75.6	0.005	101
tt, lag-m liga-gil, langu-gi, letel, heaklis of kares	6.0	17.1709164		2017-01-01 16:00		1.80	8.960	91.9	11.1	0.000	101
16, Jang-to Higa-gil, Jorgen-go, benal, Resultin of Norwa	1.0	17,1709104	121.865625	2013-01-01 (0.00)		9.861	8.80	55.4	76.0	9.003	191

Figure 1. Raw data set

### 2 Methods

## 2.1 Data Preparation

The following part uses clustering, prediction and forecasting for data analysis and data mining.

Clustering is the method of separating a population or set of data points into multiple groups so that data points in the identical group are more similar than data points in other groups. To put it another way, the aim is to identify groups with related features and assign them to clusters. There are two algorithms used for clustering in this report, which are K-Means and DBSCAN.

The clearly marked groups in the data can be found by the K-means algorithm, which is very effective for identifying unknown groups in complex data and judging the types of groups existing in the data.

DBSCAN is a data clustering approach that uses density-based spatial grouping of applications with noise. DBSCAN combines together points that are close to each other based on a distance measurement (typically Euclidean distance) and a minimal number of points based on a set of points. The spots that are in low-density regions are likewise marked as outliers.

Prediction involves estimating the outcome of unknown data.

Forecasting is a sub-discipline of prediction in which it can predict its future based on time series data. A algorithm called Kalman filter is used for forecasting in this report.

There are two algorithms used for clustering in this report, which are Liner Regression and Random Forest Regressor.

Linear Regression fits a line to the observed data to describe the connection between variables. A straight line is used in linear regression models, whereas a curved line is used in logistic and nonlinear regression models. Regression can be used to estimate how a dependent variable will vary as the independent variable(s) change.

Random forest allows multiple decision trees to be integrated into the final decision, which can be used for classification tasks or regression tasks. Among them, classification tasks include discrete output or predictive classification, while regression tasks include predicting continuous output such as time and price.

### 2.1.1 Data Cleaning:

The data cleaning operation needs to be completed on the basis of the initial data firstly, so that the amount of system operation can be reduced in the subsequent data analysis and prediction process, and the opera-tion efficiency can be improved. Filtering out invalid data is the first part of data cleaning. This step directly removes data that has no effect on pollutant values such as address and geographic location. Then I ex-pressed all the data of the same pollutant every day in the form of aver-age. The data that have completed data cleaning process (Figure 2).

√ 1,095 results (01/01/1970 11.00,00.000 to 08/10/20.	21 22:16:32:000)						Job = II II • F Smort Mode =
Row Date Preview Id							
_time 1	Station_code 1	PM2.5 1	PMIO 0	00 1	502 0	NO2 0	03 0
2017-01-01	105	49.5410000000004	31.2316666666667	6.7633333333333332	0.00525000000000000	8.0385000000000000	8.00752500000000000024
2817-91-92	105	64.27100440000087	126.333333333333333	8.7410080080080087	0.00562500000000000015	0.03270033333333334	0.0202003333333334
2817-81-83	105	40.100000000000004	30.2316666666667	0.6022222222222222	0.0062588088000000000	0.82322166666666665	8.01641908808808666673
2817-81-84	105	32.79066644666064	60.458333333333333	0.70333333333333	0.0040000000000000000000000000000000000	0.840000000000000015	8.0131660680666660
2017-01-05	105	17.547686666668888	27.3333333333333333	8.5310080066068088	0.00529786666666689	8.0257160060666668	8.821041668666666677
2817-81-85	105	13.66068666666666	20.583333333333333	8.5680080088008008	0.0947500000000000025	0.8268333333333333344	0.0187003333333334
2817-81-87	105	25.54100044000008	48.75	0.74100800660088	0.0050000000000000000000000000000000000	0.037503333333333333	0.000033333333333334
2817-81-88	105	18.7883333333333322	25.25	0.512499999999998	9.004833333333333333346	0.821500000000000016	8.0271666666666683
2017-01-09	105	49.0000000000000000	68.333333333333333	0.47861133333333333	0.005875800000000000	0.8192083333333333333	8.027500000000000018
2817-01-10	105	15	12.175	0.528833333333334	0.0055000000000000000000000000000000000	8.0192003333333334	0.8254100800800000

Figure 2. Data set after filtering

### 2.1.2 Outliers:

There are 14 outliers detected in the dataset (Figure 3). However, the outliers are not removed. It is because the outliers are legitimate observations and they do not affect the result of the following analysis.

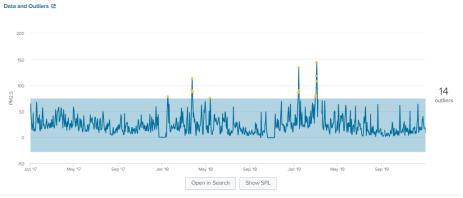


Figure 3. Outliers in dataset

### 2.1.3 Missing data and null value:

There is no missing data and null value found in the dataset.

### 2.2 Data analysis and data mining

The following part uses clustering, prediction and forecasting for data analysis and data mining.

### 2.2.1 Clustering using K-means:

The cluster silhouette\_score of clustering using K-Means algorithm is 0.4696 and it is not far from 1 (Figure 4).

silhouette\_score \$

# 0.4696494565076221

Figure 4. silhouette\_score of clustering using K-Means algorithm

### 2.2.2 Clustering using DBSCAN:

The cluster silhouette\_score of clustering using DBSCAN algorithm is 0.859 and it is close to 1 (Figure 5).

Figure 5. silhouette\_score of clustering using DBSCAN algorithm

## 2.2.2 Prediction using Liner Regression:

The predicted line of PM2.5 and the actual line of PM2.5 is very different (Figure 6). The predicted PM2.5 and the actual PM2.5 in scatter chart is also very different (Figure 7). The  $R^2$  of prediction using Liner Regression is 0.4695 and it is far from 1 (Figure 8).

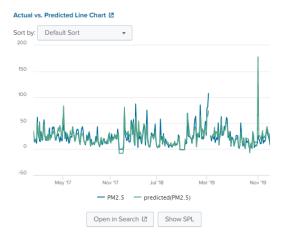


Figure 6. Line chart of prediction using Liner Regression

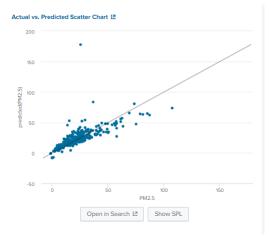


Figure 7. Scatter chart of prediction using Liner Regression

R<sup>2</sup> Statistic ☑

0.4695

Figure 8. R<sup>2</sup> of prediction using Liner Regression

## 2.2.2 Prediction using Random Forest Regressor:

The predicted line of PM2.5 and the actual line of PM2.5 is similar (Figure 9). The predicted PM2.5 and the actual PM2.5 in scatter chart is different (Figure 10). The  $\mathbb{R}^2$  of prediction using Random Forest Regressor is 0.8139 and it is close to 1 (Figure 11).

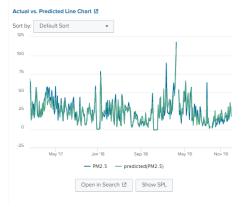


Figure 9. Line chart of prediction using Random Forest Regressor

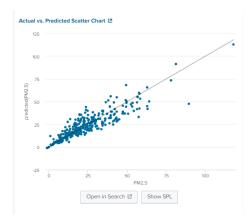


Figure 10. Scatter chart of prediction using Random Forest Regressor

R2 Statistic 12

0.8139

Figure 11. R<sup>2</sup> of prediction using Random Forest Regressor

### 2.2.2 Forecasting using Kalman filter:

The R<sup>2</sup> of forecasting using Kalman filter is 0.4598 and it is far from 1 (Figure 12).

R<sup>2</sup> Statistic 🛂

0.4598

Figure 12. R<sup>2</sup> of forecasting using Kalman filter

## 3 Evaluation

For clustering, the silhouette\_score of DBSCAN algorithm is 0.3894 higher than the silhouette\_score of K-Means algorithm. Better clustering configuration is represented by the silhouette\_score that closer to 1. Therefore, DBSCAN algorithm is recommended for clustering.

For prediction, the  $R^2$  of Random Forest Regressor algorithm is 0.3444 higher than the  $R^2$  of Liner Regression algorithm.  $R^2$  that closer to 1 represent better model. Therefore, Random Forest Regressor algorithm is recommended for prediction.

For forecasting, Kalman filter algorithm is not recommended for forecasting. It is because the  $\mathbb{R}^2$  of the algorithm is far from 1 which means it is not a suitable model.

### 4 Tools

Splunk was used in this research, and it is mainly used for big data search and monitoring. The information stored by Splunk can be correlated and indexed by it, and the data can be searched and reports can be generated and data can be visualized.

## 5 Lessons Learned

In this research experience, I have a deeper understanding and mastery of the data analysis knowledge in the course middle school, which includes the use conditions and methods of different algorithms and how to choose a more suitable algorithm to complete the research. Through continuous practice and testing, I have also increased my understanding of how to use Splunk. In addition, in the process of completing the research this time, my time planning skills and learning ability have also been enhanced.

### 6 Conclusion and future work

To conclude, DBSCAN algorithm is recommended for clustering and Random Forest Regressor algorithm is recommended for prediction.

# 7 Weekly member activities

The whole project is completed by myself, I participated in the entire process of data processing and gained a deep-er understanding and mastery of each step.