Data Analytics: Air Pollution in Seoul Between 2017 and 2019

Santosh Bahadur (518904) | Kazi Omar (513569) | Sandeep Giri (534325)

Assignment Group: 22

P3T Bachelor of Information and Communication Technology

University of Tasmania

Abstract – **The following research report is based on Seoul's air pollution from 2017 to 2019. The data analytics in this paper comprises mainly six levels namely: Background Information and Problem; Data Preparation; Data Mining; Evaluation and Result from analysis and Future work. The aim of data analytics is to investigate past data on air pollutants to gather insights on how the air pollutants have evolved and see if a predictive model can be created. Some of the factors investigated are the time of the day, the month of the year, and year time between 2017 and 2019, etc. Machine learning techniques have also been used to forecast the types of pollutants that will be present in the future.**

Keywords-data analysis, machine learning, pollutants

I. INTRODUCTION

Air pollution has been a major factor leading to deaths worldwide. According to WHO, seven million people die every year due to many diseases caused by air pollution. The WHO data also shows that 9 out of 10 people breathe air that exceeds the WHO guideline limits containing a high level of pollutants (Air pollution, 2021). The major pollutants in urban areas include Carbon monoxide (CO), Carbon dioxide (CO2), Nitrogen Monoxide (NO), Nitrogen Oxides (NO2), and particulate matter PM2.5, PM10. (Air Pollution Forecasting, 2018). However, particulate matters such as PM2.5 and PM10 is the area of concern as this matter cause a serious health problem.

PM10: Inhalable particles, with diameters that are generally 10 micrometers and smaller.

PM2.5: fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller (Particulate Matter (PM) Basics | US EPA, 2021).

It is very important to study the past events on air pollution and react accordingly for making future changes to reduce air pollution as air pollution has been a concern for many countries in contributing to diseases and deaths.

In this research, our team has chosen Air pollution dataset for Seoul to study the past events for making future analysis in generating a predictive model for solving a problem. Splunk has been used to prepare data and generate various models using different data mining techniques. First, the team will prepare the data following various data pre-processing and transformation techniques then the data will be explored using visualization tools, clustering, and using various data mining techniques the models will be generated following the objective.

II. UNDERSTANDING THE DATA

Air Pollution Seoul Dataset deals with the measurement of air pollution, specifically the six pollutants (SO2, NO2, CO, O3, PM10, PM2.5) in Seoul, South Korea. The data were measured for 25 districts between 2017 and 2019 in an interval of an hour. The dataset is divided into four files:

Measurement_info.csv: This dataset contains 5 attributes which are measurement date, station code, item code, average value, instrument status. This file provides information on the average value of the six pollutants which are identified by the item code. The pollutants level was measured from the 25-station in the interval of one hour from the date 1/01/2017 0:00 to 31/12/2019 23:00.

Measurement_item_info.csv: This dataset has 7 attributes which are item code, item name, unit of measurement, and pollutant level such as Good (Blue), Normal (Green), Bad (Yellow), Very bad(red). This file provides an indicator of the pollutants level such as Good (Blue), Normal (Green), Bad (Yellow), Very bad(red) if the pollutants level measured falls in the given range.

Measurement_station_info.csv: Measurement station info has 5 attributes which are station code, station name, address, latitude, and longitude. The dataset provides detailed information on the location of each station. Each station can also be identified using station code.

Measurement_summary.csv: Measurement summary provides the summary of the above three data sets. Various relations are being made between the given three datasets to produce the measurement summary. The measurement summary provides attributes described in the three datasets but a more compact form. We can find the following attributes in the measurement

summary: Measurement date, Station code, Address, Latitude, Longitude, SO2, NO2, O3, CO, PM10, PM2.5.

III. CHARACTERISTICS OF THE DATA

- The data is the measurement of six pollutants in 25 districts. (SO2, NO2, O3, CO, PM10, PM2.5)
- The measurement is done in an interval of one hour from the year 2017-2019
- Not all data was measured with the instrument in good status so, data contains faulty values.
- Item codes are used to indicate pollutants. (1 -> SO2; 3 -> NO2; 5 -> CO; 6 -> O3; 8 -> PM10; 9 -> PM2.5)
- Pollutant's level is better when the average value is lower and can be represented as Good, Normal, Bad, Very Bad.

IV. PROBLEM

Use various visualization techniques to study the data.

- Visualize the overall trend of the 6 pollutants level over the past 3 years.
- Analyze what time in the day the pollutants level rise.
- Analyze the pollutants which exceed the harmful limit for all 25 districts.
- Identify how the model can be useful and how can it be improved.
- Make descriptive analysis by looking in-depth at past events to gather insights on how things looked and see if a predictive model can be created.

Since the datasets were large with various attributes it was crucial for the team to refine the goal and hence come up with a major objective for the project. The team decided to use Exploratory Data Analysis and Data visualization techniques to get a general sense of the data using charts to visualize patterns, detect outliers, etc.

Initially, the team decided to pick a measurement info dataset as it contained major attributes that would be relevant to the analysis task.

V. DATA PREPARATION

Data preparation is an important step in creating a model with improved accuracy. The air pollution Seoul dataset is large and may contain poor quality data or missing values. Using such data in creating a predictive model can alter the accuracy of insights or could lead to incorrect insights (Why data preparation is an important part of data science? 2021). The team undertook various data preparation steps in getting the data ready for mining.

A. Incomplete (Missing) Data

Using Splunk doing a quick index search we found out that all the fields had 100% event which signifies the dataset doesn't contain any empty value. Following Splunk command was also used to verify:

```
|inputlookup measurement_info.csv | search isnull(attributes)
```

However, looking at the instrument status not all measurements were done using the instrument in good condition. So, we needed to handle noisy data.

B. Handle Noisy Data

Since the 1-hour interval measurement is provided after calibration it seems not all the measurements are accurate. Data was measured while the instrument status needed calibration, Abnormal state, Power cut off, Under Repair and some data were Abnormal. The team decided to re-fill the faulty measurement values with the value recorded in the nearest time for the pollutants level.

To achieve this, we needed to sort the data. We sorted the Station code and Item code by smallest to highest, and with the sorted data we replaced the faulty value with NULL and later Filled Down.

The following command could accomplish the task:

```
| inputlookup measurement_info.csv
| table "Measurement date","Station code","Item
code","Average value","Instrument status"
| sort "Station code"
| sort "Item code"
| eval "Average value"=if ('Instrument status'!=0, null(),
'Average value')
| filldown
```

C. Data Aggregation

Since the measurement is done in an interval of the hour, the dataset will be huge and too detailed to fit in the model. The team decided to transform the data into the interval of the day by taking the average of the measurement recorded in the 24-hour interval. Also, to visualize the data in a year or monthly form the dataset was transformed accordingly.

The following command was used to transform the hour interval dataset to a day:

```
| eval _time=strptime('Measurement date', "%Y-%m-%d")
| timechart span=1day avg("Average value") by "Item code"
```

D. Data Generalization

The address was converted to City by removing the Street Address and only keeping the city name, this way it becomes easier to visualize data from a higher abstraction view.

E. Data Reduction

Since Instrument status was used as a reference to clean data for faulty values, we no longer need it. Instrument Status was excluded from the dataset.

F. Feature Creation

The date was converted to Unix timestamp as the Measurement date in the dataset was not an epoch time but rather a String.

VI. Exploratory Data Analysis

Explore the overall trend of the six pollutants from the year 2017 - 2019

The team decided to investigate the history of air pollutants for all districts to examine how the level of each pollutant has evolved over the years.

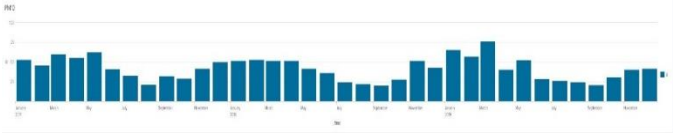

Figure 1: PM10 Evolution Between 2017 and 2019

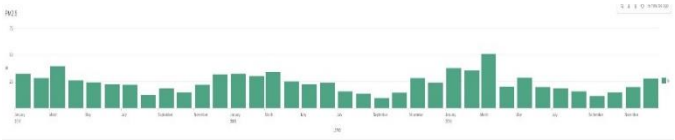

Figure 2: PM 2.5 Evolution Between 2017 and 2019

Particulate Matters PM10 and PM2.5 show a similar pattern over the years. The pollutants level seems to be rising in the month of November to March and slowly going down and the lowest in the month of July-September.

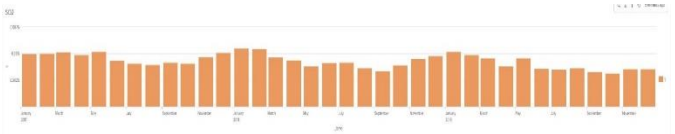

Figure 3: SO2 Evolution Between 2017 and 2019

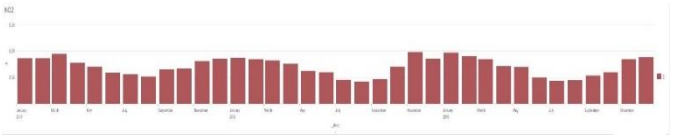

Figure 4: NO2 Evolution Between 2017 and 2019

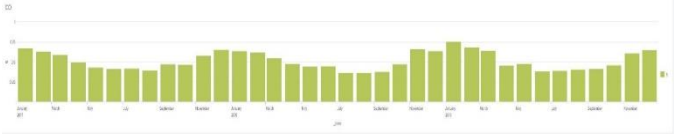

Figure 5: CO Evolution Between 2017 and 2019

The overall trend of SO2, NO2 and CO looks similar, there seems to be rise in the pollutants level mostly in the month of November to March and slowly going down in the rest of the month.

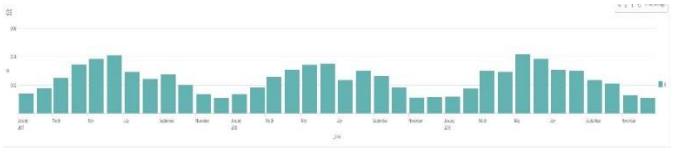

Figure 6: O3 Evolution Between 2017 and 2019

O3 seems to have a different trend than the rest pollutants level, there seems to be a rise in March to June and the trend is going down the rest of the month.

Overall, all pollutants seem to be good or below average range. However, PM2.5 seems to have a spike in some months so, the team decided to dig deep into PM2.5 and visualize the trend for all districts.

Visualizing PM2.5 for all districts we can see that all the station follows similar trend. Among all the station, station 119 has the highest level in March. Further, the team decided to visualize the PM2.5 for Station 119 in more detail.

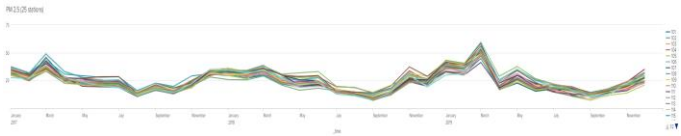

Figure 7: Visualization of PM 2.5 for all station

Figure 8: Visualization of PM 2.5 for station 119 in hourly basis

As we can see from the visualization above, the air pollution in Seoul seems to be pretty good and most of the time under the average limit. PM2.5 and PM10 seem to be at average or a little bit high towards bad in some of the time.

VII. Clustering

The purpose of cluster analysis techniques is to locate different groups of similar elements. One of the most prominent strategies for discovering the underlying structure is grouping, which has the advantage of not requiring human supervision (Represa, N.S., et al. 2019).

We used clustering to further explore the data to understand the structure. For this project, we wanted to explore the correlation between pollutants by clustering and looking for the pollutants that are close in values to each other. We used K-mean clustering to achieve this. K-means is a prototype-based, simple partitional clustering algorithm that attempts to find K non-

overlapping clusters. These clusters are represented by their centroids (a cluster centroid is typically the mean of the points in that cluster) (Wu, J. 2012). For the algorithm, we chose the value of k to be 3, we used Standard Scaler to standardize all the data fields by scaling their mean and standard deviation to 0 and 1 so, that the data with different scales are standardized by rescaling to have a standard deviation of one or both and centering about the mean (Preprocessing your data using MLTK Assistants - Splunk Documentation, 2021). We also used PCA to reduce the dimension of the dataset. Looking at the result, there seems to be a co-relation between PM10 and PM2.5 and NO2 and SO2.

VIII. PREDICTIVE DATA MINING

The dataset we have comes under the Time-series category so, we decided to perform Time-Series Data analysis using the Univariate Time-Series Forecasting method using an algorithm like Kalman, Arima for the identification and the prediction of the values of pollution level for the future timeframe. Using the different algorithm, we will generate models and later compare in the evaluation section. The team decided to create a predictive model to predict future pollutants levels of PM2.5 for Seoul's 25 districts. In this method, the forecasting problem will contain only two variables in which one is time and the other will be the field to forecast, in our case PM2.5 pollutant.

IX. TIME SERIES FORECASTING

A. Kalman Filter

Kalman filter is an algorithm that considers a subset of the features. It takes time series as input and performs smoothing and denoising so, the smoothed series can be predicted. There are several methods used for prediction which are Local Level used to predict local levels of time series without any trend, Local level trend to predict the trend only, Seasonal Local Level to predict only seasonal component and combination of LLT and LLP to account for both trend and seasonality. For our model, we generated a time chart for 1 month using the search below:

```
| inputlookup Measurement_info.csv
| where 'Item code' = 9 | eval "Average value"=if('Instrument status'!=0, "", 'Average value')

| filldown

| eval _time=strptime('Measurement date', "%Y-%m-%d")

| timechart span=1mon avg("Average value") by "Item code"
```

We have selected Algorithm as Kalman Filter, Field to Forecast as PM2.5, Method to the seasonal local level, future timespan 24 months, confidence interval as 95.

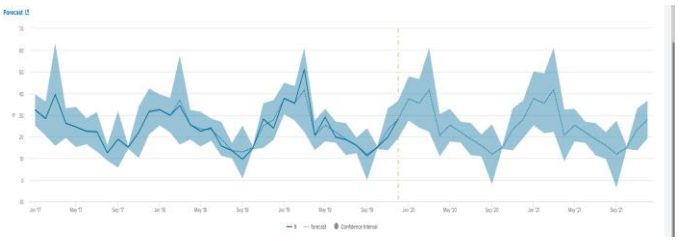

Figure 9: PM2.5 Prediction Using Kalman Filter algorithm

As we can see in the line chart, behind the dotted line is the data that the algorithm used to learn and the prediction for the events can be seen after the dotted lines.

Figure 10: PM2.5 R2 Statistic & RMSE Accuracy using Kalman filter

We can also use R2 Statistic and Root Mean Squared Error to measure the accuracy of the model.

R2 Statistics measures the model fit for the forecasted algorithm which means the closer the value is to 1 the better is the model. The Root Mean Squared Error is the result of the absolute measure of fit, which is the smaller the number the better the model. The model generated has an R2 of 0.93 which is close to 1 so, the model seems to be good and similarly, the root means the squared value is relatively less i.e., the model has fewer residuals.

B. Arima

Arima is a linear regression model that uses its own past values which is its own lags and the lagged forecast errors, so that the equation can be used to forecast future values (ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+, 2021). We need to select p, d, and q for this model. We have selected p as 6 which means we will use 6 previous of our time series in the autoregressive portion for the calculation. D is the difference between the data point and the point that follows it so it makes the time series stationary, in our case we will use d as 1. Q is used to forecast values using x previous prediction error and we will use q as 0.

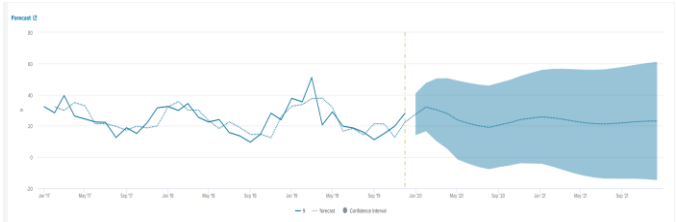

Figure 11: PM2.5 Prediction Chart Using Arima Algorithm

As we can see in the line chart, behind the dotted line is the data that the algorithm used to learn and the prediction for the events can be seen after the dotted lines.

Figure 12: R2 Statistics and RSME Accuracy Using Arima

For the Arima model, the R2 is closer to 0 which indicates the model didn't fit properly even though the Root mean squared error is relatively low.

X. EVALUATION

For the evaluation we used two models, model generated using Arima and Kalman-Filter. We used various elements to find the model with good accuracy and high performance. As we can see from the comparison of different metrics below for the two model, model generated using Kalman Filter tends to have higher accuracy as the R2 Statistics is closer to 1 and Root Mean Squared Error is also low.

Models	R2 Statistics	Root Mean Squared Error
Kalman Filter	0.93	2.20
Arima	0.38	6.96

Table 1: Accuracy Results of Kalman Filter and Arima

XI. CONCLUSION

Particulate matters have numerous health impacts, and many countries are suffering damages due to an increase in PM levels. Exposures to the particulate matter have been associated with early age mortality, increase the chance of heart and lungs diseases and respiratory symptoms. It is important to analyze past data on air pollution to make a future prediction so, the country can take appropriate steps before it can create a bigger impact. During our investigation, we analyzed trends of various pollutants levels for districts in Seoul and dig deeper to understand the PM2.5 level and how it is evolving. We also visualized the PM2.5 for the district which had the highest measurement recorded. With the model generated, in the future, we can further optimize the model accuracy and use it to predict PM2.5. The finished product will display the predicted PM2.5 level of the district in Seoul.

XII. REFERENCES:

1. Who.int. 2021. *Air pollution*. [online] Available at: <https://www.who.int/health-topics/air-pollution#tab=tab_1> [Accessed 18 September 2021].
2. ProjectPro. 2021. *Why data preparation is an important part of data science?*. [online] Available at: <https://www.projectpro.io/article/why-data-preparation-is-an-important-part-of-data-science/242#mcetoc_1fafnlp3v5> [Accessed 21 September 2021].
3. 2018. *A Deep Learning Approach for Air Pollution Forecasting in South Korea Using Encoder-Decoder Networks & LSTM.* [online] Available at: <https://www.researchgate.net/publication/324716980_A_Deep_Learning_Approach_for_Air_Pollution_Forecasting_in_South_Korea_Using_Encoder-Decoder_Networks_LSTM> [Accessed 4 October 2021].
4. US EPA. 2021. *Particulate Matter (PM) Basics | US EPA*. [online] Available at: <https://www.epa.gov/pm-pollution/particulate-matter-pm-basics#effects> [Accessed 4 October 2021].
5. Wu, J., 2012. *Advances in K-means Clustering*. [online] Google Books. Available at: <https://books.google.com.au/books?hl=en&lr=&id=pI2_F8SqWcQC&oi=fnd&pg=PR10&dq=clustering+data+mining&ots=chaN2N9mkj&sig=CVJ-YNSWGgLO_mmNQ5qyJrIpsIE&redir_esc=y#v=onepage&q=clustering%20data%20mining&f=false> [Accessed 7 October 2021].
6. Represa, N.S., Fernández-Sarría, A., Porta, A. and Palomar-Vázquez, J. (2019). Data Mining Paradigm in the Study of Air Quality. Environmental Processes, 7(1), pp.1–21.
7. Docs.splunk.com. 2021. *Preprocessing your data using MLTK Assistants - Splunk Documentation*. [online] Available at: <https://docs.splunk.com/Documentation/MLApp/5.3.0/User/Preprocessing> [Accessed 7 October 2021].
8. Machine Learning Plus. 2021. *ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+*. [online] Available at: <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/> [Accessed 7 October 2021].

XIII. Team Contribution

Throughout the project phase, the team worked together. The combined effort of the team from selecting the dataset to making predictive analysis made the project possible from start to finish. In the beginning, we had 1 or 2 sessions every week and then slowly increasing it to 3-4 sessions a week to work on the project. During our meet, we would spend some time discussing the topic and worked with Splunk to do data analysis. At the same time, we would write the draft report of the progress made each day. For the Background information and Data Preparation, we worked for 4 weeks, Data analysis and data mining for 2 weeks and rest of the time was spent in evaluation, analysis, and further improvement of the report.

Kazi Faruque (513569):
I took the leading role with all the work with Splunk as we discussed in our first team meeting about the role. Splunk was very new to me as I have worked with other software for data analysis like weka, rapid miner etc. I have studied both academic and online resources to resolve the problem we got in every phase of the research. As we divided the part into weeks interval, I have tried to follow the specific goal for every week that were assigned. I have worked efficiently with the documentation part such as, background information, data preparation, data mining steps. I have helped the team with resolving issues related clustering and algorithm problems. In the data preparation, I have helped my team with pre-processing which helped us to use the dataset in machine learning. In data mining steps, I have helped my team with creating modelling and finalizing results to put in the report. In documentation have helped my team to present the document and proofreading etc. Overall, it was a great experience working with my team.

Sandeep (534325):
I took the leading role in documenting the work the team discussed during the meeting. I worked with the team member closely in all the parts, researching any confusion that arrived by going back to the unit or asking questions to tutor and lecturer during the consultation or by email so, the team doesn't lag. I helped with the Problem background, Data Preparation, making both descriptive and predictive analysis using descriptive techniques like clustering, visualization through charts, and predictive techniques like forecasting using Splunk. In the end, I helped the team to evaluate the best model by comparing the evaluation metrics and discussed the future work. After the draft was documented, I helped the team member to finalize the document, proofread and make reports error-free.

Santosh (518904):
I helped in the report to finalize the structure and referencing. I mapped out how we can understand and distinguish the different pollutant levels, stations, and time frames with all team members. Helped in making decision on how we can sort the missing data sets and with what command we can use in SPLUNK to handle those missing data. Helped in clustering and understood how it can be beneficial in this report. Helped Sandeep with the descriptive and predictive analysis. Motivated the team at peak moments when everyone was getting overwhelmed with the study load as it was reaching the end of semester. Finally, I also helped my team to proofread the report to make sure no mistakes occurred.