Data Analysis of Air Quality in Australia

Group 16 - GreenTech

Ran Jiang
ranj3

Zhanna Sakayeva
zhannas3

Mengjia Zeng
mengjia6

***Abstract*—Because of rapid urbanization, many developing countries have suffered from serious air pollution problems [5]. Governments and citizens continued to express increasing concerns on air quality, because it may cause long-term health problems and short-term health effects [2]. With these concerns in mind, air quality evaluation and prediction has become an important research area. Some previous studies have reported the effectiveness of some complex models, like different deep learning architecture. In this study, we used a less-used dataset collected in Australia, a developed country, to see whether the influence of pollutants on air quality would be different than that in the developed country. Furthermore, one limitation for previous studies was not including some trend analysis of chemicals and our project also included exploratory analysis of trend analysis of changes of chemicals, as well as PM2.5 to see whether air quality may be different among months and check further about its annual trend. Last, our project tried some easier models instead of deep learning architectures in order to balance between the accuracy and processing cost among the process of predicting PM2.5 in the future.**

***Index Terms*—air quality, regression, exploratory analysis**

I. Introduction

Today, air condition has becoming an overwhelming health-related problem among the world. Multiple researches have been conducted to help prevent and predict the air conditions worldwide. In our project, our target is to build a model which ensures both interpretability and accuracy for predicting the AQI of PM2.5 using the chemicals in the present hour.

To achieve this target, this model should not be too complex while it should still have enough capability for predicting air conditions. Currently, we have conducted correlation analysis between all the air pollutants' Air Quality Indexes (AQI) including AQI_CO, AQI_NO2, AQI_O3_1hr, AQI_O3_4hr, AQI_PM10 and AQI_PM2.5 for better understanding the relationship between different chemicals. Additionally, visualization techniques like heatmap have been applied to the AQIs for us to better interpret their correlations. After careful data preprocessing and analyzation, we start building two models which predicts the AQI_PM2.5 in the next hour with OLS regression and evaluated these two model's performance with RMSE. In our current setting so far, we have achieve similar results from both of the OLS regression models. We hope to evaluate more models with our current analzation and understanding of the chemical levels.

II. Related Work

A. Published Work that relates to Our Project

As mentioned above, air quality evaluation and prediction has become an important research area and encouraged a lot of studies.

One study has reviewed various machine-learning techniques, like Artificial Neural Network, Random Forest Model and Decision Tree Model, used in the field of prediction of air quality. Those deep-learning architectures were constructed based on different kinds of datasets, which are mostly located in China, a developing country, with different accuracy varying from 0.55 to 0.89 [2]. A similar study also tried to develop a deep-learning technique, which was named as the stacked autoencoder model that incorporates autoencoders as building blocks to construct a deep network. By the experiments conducted at different stations in China, the air quality, as indicated by the concentration of PM2.5, were separated into different ranks and the overall prediction for ranks could be as round 0.82 [4]. Another similar study proposed a deep neural network (DNN)-based approach that consisted of a spatial transformation component and a deep distributed fusion network to predict air quality. Within this study, the accuracy could reach 0.81 in predicting the air quality in the next hour [5]. Moreover, Shwet (2022) Proposed a model using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF) to predict the Air Quality Index (AQI) and Air Pollutant Concentration (NOx) levels. And for the model comparison, seven well-established machine learning models have been taken, which are aimed to compare with the proposed model to find the balance between accuracy and suitability. By using MAE, MSE, RMSE, and R2 scores to compare prediction models' performance, the proposed hybrid RFERF model is the best-suited and highly accurate. The Linear Regression and the Pearson correlation had been used by Wenrong for analyse the air quality in Shanghai, as well as considering the multicollinearity between the six factors, in order to reduce the influence of multicollinearity on the model [1].

B. How Our Approach will be Similar or Different from Others

Based on the previous research, most of the study focused on developing a prediction model to predict PM2.5, and not paying super attention in the trend analysis part. There may

be a specific trend of chemicals, as well as PM2.5 annually and air quality may even be different between weekdays and weekends, which was what this study would cover. Also, the generally used datasets are from stations based in China, a developing country, while our study will try a different dataset that is rarely used, as collected in Australia, a developed country, to check the trends and develop possible prediction models. Even though we will try to use AQI of PM2.5 as the response variable, as most of the studies did, we will try to fit linear regression models and also try to divide different concentration levels of PM2.5 into different ranks in order to develop some further classification models.

III. Data

In our project, we will be using the air quality data from a performance monitoring station. It was obtained from ACT Government Open Data Portal Australia Government. Our dataset Civic Air Quality Station mainly focuses on the air pollution data from Civic station which is based on Air Quality Monitoring Data Air Quality Monitoring Data — Open Data Portal. The website provides publicly available datasets that contain independent scientific measurements of various pollutants obtained from several active monitoring sites of Australia. The portal provides accessible and shareable data from the Australian Capital Territory and allows us to download datasets in csv or json forms Civic Air Quality Station — Open Data Portal. So, our dataset was available to download as a CSV file and was accessible through URL as well.The time period chosen was from January 1, 2011 and it is updated every hour. At this moment, the dataset contains 104152 records and 18 features. Examples could be seen in Fig.1.

Further we did some analysis on our dataset and performed data wrangling, which was followed by a data modelling stage. As we discussed, before constructing the model and drawing any insights from it we divided our dataset into train and test sets. We implemented our model and trained it with a training dataset and validated it on a testing dataset. The dataset was splitted in such a way that 0.75 was training data and the rest 0.25 was testing data. It was done so that the model could be first trained and then could be tested on the test set such that the error in prediction could be checked and proper results could be used. The variable to be predicted was AQI of PM2.5.

IV. Exploratory Data Analysis

As a first step to get a good understanding of some data preprocessing steps, we provided the numerical summary of the data below.

	count	mean	std	min	25%	50%	75%	max
NO2	25472.0	0.007554	0.016522	0.0	0.001	0.002	0.006	0.271
O3_1hr	98018.0	0.015364	0.009888	0.0	0.008	0.016	0.022	0.105
O3_4hr	102208.0	0.015711	0.009411	0.0	0.009	0.016	0.022	0.098
CO	26686.0	0.259621	0.223538	0.0	0.110	0.200	0.350	1.960
PM10	75444.0	12.928034	29.798101	-0.9	6.200	9.400	13.700	1107.400
PM2.5	63471.0	7.589173	26.999084	-1.7	3.400	5.100	7.400	1010.600
AQI_CO	26686.0	2.883609	2.501523	0.0	1.000	2.000	4.000	22.000
AQI_NO2	25897.0	6.468548	5.177982	0.0	3.000	5.000	9.000	38.000
AQI_O3_1hr	97965.0	15.364834	9.889201	0.0	8.000	16.000	22.000	105.000
AQI_O3_4hr	102208.0	19.548118	11.768187	0.0	11.000	20.000	27.000	123.000
AQI_PM10	75444.0	25.852951	59.597605	-2.0	12.000	19.000	27.000	2215.000
AQI_PM2.5	63471.0	30.353059	107.997517	-7.0	13.000	20.000	29.000	4043.000
AQI_Site	104179.0	32.302230	85.634297	0.0	19.000	26.000	33.000	4043.000

Fig. 2: Numerical summary.

We may observe that the mean value for PM2.5 is 7.589173 and it ranges from -1.7 to 1010.600. This may indicate the existence of outliers. In order to observe that and to get initial impressions about the data, the time series visualisation of the dataset was plotted. It can be seen that we have very high values(≈ 1000) of PM2.5 for the year 2020. This might indicate the greater the level of air pollution and health concern. Generally, the value that is below 50 might indicate good air quality and over 300 represents hazardous air quality.

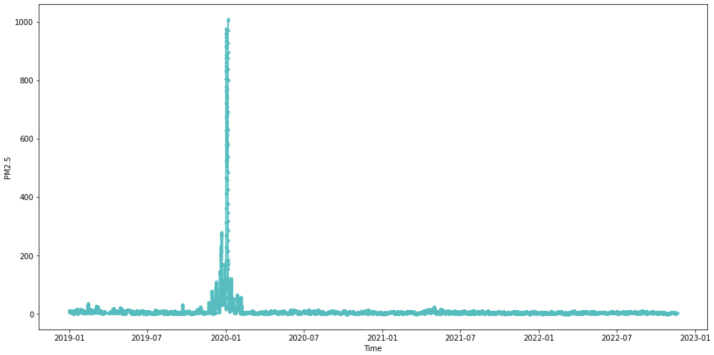

Fig. 3: Time series visualization from year 2019.

This could also be observed from the box plot in Fig.4, which represents PM2.5 by month in 2020.

	Name	GPS	DateTime	NO2	O3_1hr	O3_4hr	CO	PM10	PM2.5	AQI_CO	AQI_NO2	AQI_O3_1hr	AQI_O3_4hr	AQI_PM10	AQI_PM2.5	AQI_Site	Date	Time
0	Civic	(-35.285307, 149.131579)	10/11/2022 11:00:00 AM	NaN	0.024	0.020	0.0	7.00	3.62	0.0	0.0	24.0	25.0	14.0	14.0	25.0	10 November 2022	11:00:00
1	Civic	(-35.285307, 149.131579)	10/11/2022 12:00:00 PM	NaN	0.026	0.023	0.0	6.65	3.50	0.0	0.0	26.0	29.0	13.0	14.0	29.0	10 November 2022	12:00:00
2	Civic	(-35.285307, 149.131579)	10/11/2022 01:00:00 PM	NaN	0.027	0.025	0.0	6.58	3.55	0.0	0.0	27.0	31.0	13.0	14.0	31.0	10 November 2022	13:00:00
3	Civic	(-35.285307, 149.131579)	10/11/2022 02:00:00 PM	NaN	0.029	0.026	0.0	6.60	3.52	0.0	0.0	29.0	33.0	13.0	14.0	33.0	10 November 2022	14:00:00
4	Civic	(-35.285307, 149.131579)	10/11/2022 03:00:00 PM	NaN	0.029	0.028	0.0	6.38	3.44	0.0	0.0	29.0	35.0	12.0	13.0	35.0	10 November 2022	15:00:00

Fig. 1: First five observations from the whole dataset.

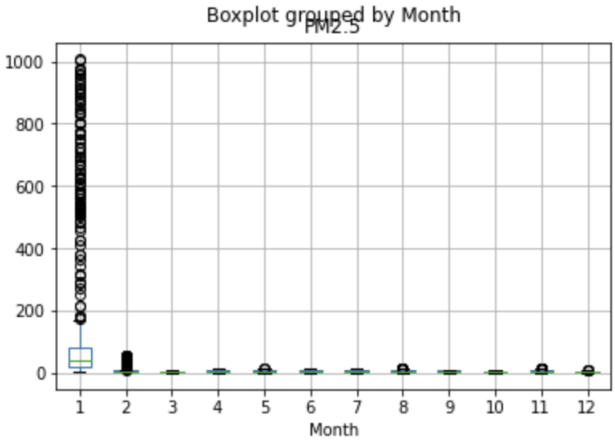

Fig. 4: PM2.5 by month in year of 2020.

Further, in order to construct models we will need to study the relationship between different features of our dataset. It can be done by constructing correlation matrix (heatmap) (Fig 5.).

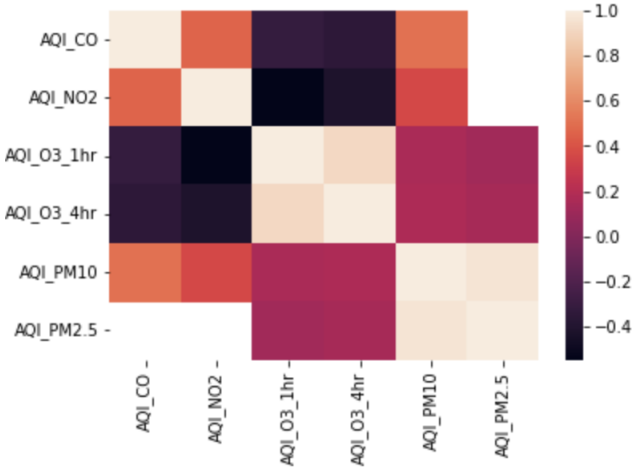

Fig. 5: Correlation matrix for the dataset.

It can be observed that there is some positive relationship between the PM2.5 and AQI_O3_1hr, AQI_O3_4hr, AQI_PM10 variables.

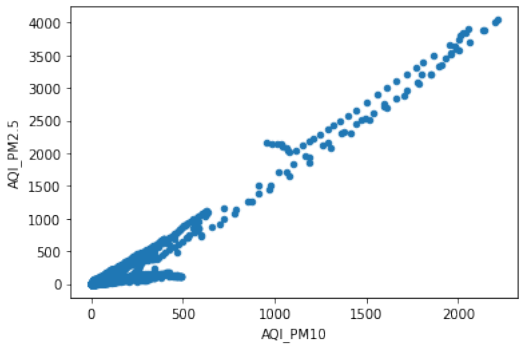

Fig. 6: The relationship between AQI_PM2.5 and AQI_PM10

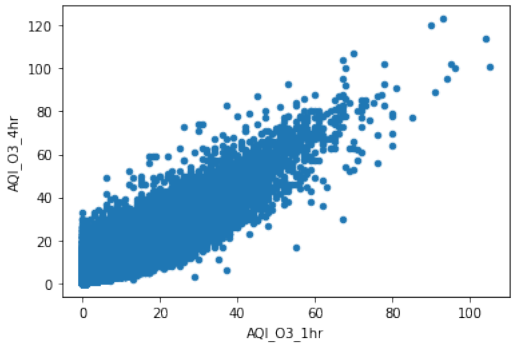

Fig. 7: The relationship between AQI_O3_1hr and AQI_O3_4hr

By observing the above correlation coefficients, a scatter plot of partial variables is made to more intuitively observe the relationship between the variables.

From Figure 6, we can see the obvious linear relationship between AQI_PM2.5 and AQI_PM10. From Figure 7, we can see the linear relationship between AQI_O3_1hr and AQI_O3_4hr.

V. Preliminary Technical Details and Results

This project first did some data wrangling and data cleaning. First, this study used AQI of chemicals, AQI_CO, AQI_NO2, AQI_O3_1hr, AQI_O3_4hr and AQI_PM10 as the predictor variables and AQI of PM2.5 as the response variable. After checking the summary statistics of the non-null PM2.5, this study used all the observations with non-null AQI_PM2.5.

Second, this study wanted to develop a prediction model using the chemicals in the present hour to predict the AQI level of PM2.5 in the next hour. Therefore, the AQI_PM2.5 in the next hour is considered as the response variable, while chemicals at the present hour are considered as the predictors. One more point to mention was that the number of non-null CO and non-null NO2 were too small, which may not be useful when adding into the model.

Therefore, after choosing observations based on the criterion above, this study then divided all the observations into train and test dataset with a division as 0.75 to 0.25. And this study first started with fitting multiple regression with three predictor variables, AQI_O3_1hr, AQI_O3_4hr and AQI_PM10. And the p-values from this multiple regression showed that all the three variables are considered as significant, which is denoted as Model1. Also, since AQI_O3_1hr and AQI_O3_4hr are both related to the concentration level of O3, therefore, this study further tried a smaller model with only AQI_O3_4hr and AQI_PM10 as the predictor variables, which is denoted as Model2. This study kept AQI_O3_4hr as the indicator of the concentration level of O3, since the p-value of AQI_O3_4hr is much smaller, indicating its significance.

As shown by the table below, the two models have similar RMSE. Since Model2 realized more with our simplicity concern, this study used Model2 to analyze further. As indicated by the coefficients of the multiple regression, with a positive coefficient of AQI_PM10 to be 1.6438, PM10 exerts a positive influence on the concentration level of PM2.5. However, as indicated by the negative coefficient of AQI_O3_4hr, O3 reversely exerts a negative influence on the concentration level of PM2.5.

Model	Model1	Model2
RMSE	26.3732	26.3785

TABLE I: RMSE with Two Multiple Linear Regression

Take the observation on 10/22/2018 23:00:00 as an example, the predicted concentration from Model2 was 23.587, while the actual concentration is 26.0, which is pretty close to each other.

For the future modelling, this project aimed to categorise different concentration levels into several ranks, like Excellent, Good, Lightly Polluted, Moderately Polluted and Severely Polluted due to the concern that different warning levels may help people better with the warning instead of a numeric concentration of PM2.5.

VI. Appendix

Our work timeline and distributions so far are summarized below as the Fig. 8 Gantt chart below.

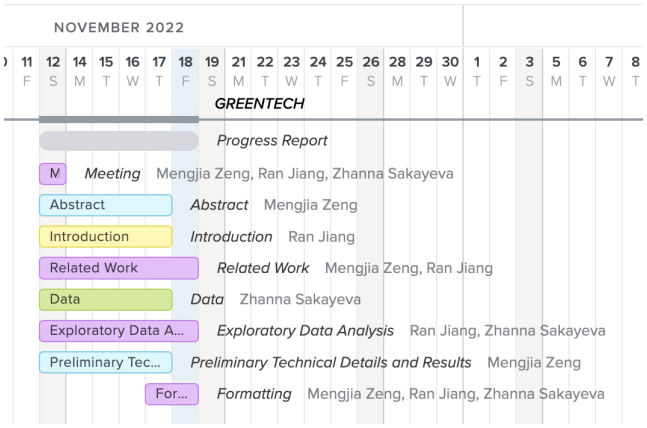

Fig. 8: Timeline of Work

VII. Contributions

Ran Jiang (33%): Wrote the introduction, related work and some part of the explratory data analysis.

Mengjia Zeng (33%): Wrote the abstract, related work and preliminary results. Wrote codes for data wrangling and data modeling.

Zhanna Sakayeva (33%): Wrote the paragraphs about data and exploratory data analysis, wrote some codes for data visualization.

VIII. Code

The code can be accessed through this link Code Link.

References

[1] Jiang, Wenrong. "The data analysis of Shanghai air quality index based on linear regression analysis." Journal of Physics: Conference Series. Vol. 1813. No. 1. IOP Publishing, 2021.

[2] Kang, Gaganjot Kaur, et al. "Air quality prediction: Big data and machine learning approaches." Int. J. Environ. Sci. Dev 9.1 (2018): 8-16.

[3] Ketu, Shwet. "Spatial Air Quality Index and Air Pollutant Concentration prediction using Linear Regression based Recursive Feature Elimination with Random Forest Regression (RFERF): a case study in India." Natural Hazards 114.2 (2022): 2109-2138.

[4] Li, Xiang, et al. "Deep learning architecture for air quality predictions." Environmental Science and Pollution Research 23.22 (2016): 22408-22417.

[5] Yi, Xiuwen, et al. "Deep distributed fusion network for air quality prediction." Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. 2018.