Group 16 - GreenTech
Ran Jiang (ranj3) Zhanna Sakayeva (zhannas3) Mengjia Zeng (mengjia6)

This project will present a civil air quality analysis of Australia based on the hourly air quality data from a performance monitoring station. We will mainly focus on the variation of AQI (Air Quality Index) along with the time changes, and the prediction of impending air quality, also will display the findings in a clear and concise manner. A clean atmosphere is one of the necessary conditions for human survival and plays an important role in the survival of human beings and living things. At the same time, exposure to bad-quality air is a major cause of mortality and other diseases like respiratory infections, allergic diseases, asthma, etc. Therefore, we are interested in doing the research and analysis in favor of the improvement of air quality and the control of the emission of harmful gases.

According to rapid urbanization, many developing countries suffer from serious air pollution problems (Yi, 2018). The increasing pollutants may cause long-term health problems and short-term health effects (Kang, 2018). However, most researchers focused on figuring out a prediction model with higher accuracy (Li, 2016; Kang, 2018; Yi, 2018), without emphasis on a basic trend analysis for the annual change or weekly change of air quality and discovering the correlation between specific chemicals and PM2.5. Therefore, one challenge for our project is to involve a trend analysis of changes of chemicals, as well as PM2.5 to see whether air quality may be different between weekdays and weekends and check further about its annual trend. Second, many studies related to air quality used the dataset from developing countries, especially in China, while our project will use a dataset from Australia, a developed country. Last, for the air quality problem, researchers used different techniques, such as deep learning techniques, to improve the prediction accuracy, while our group also plans to try one more advanced model to see whether we can improve the fitting accuracy as compared with linear regression.

To approach the aforementioned problem we will be using hourly air quality data from a performance monitoring station in Australia (dataset here).
In order to address the problem we will divide the project into several parts that involve transforming and preparing the data at the first stage. It will then be followed by an exploratory data analysis stage which will help in identifying any trends between chemicals and PM2.5, as well as along the time change. This idea should allow us to determine possible chemicals that may be used in the modeling stage to predict PM2.5. To approach this particular problem, linear regression might be used to model the local factors of air quality.

The idea before constructing the model and drawing any insights from it is to divide the dataset into train and test sets. Furthermore by fitting the model to the training set, we can evaluate it based on different metrics that can be used to measure model performance. Moreover, we can utilize the test dataset to make predictions about PM2.5. By this we can test the model on the data that was “not seen” by it previously, and check how it could possibly behave with external data. The above mentioned idea of approaching the problem strongly aligns with the course focus. It consists of the tools and methods that will be discussed in the course. Based on the course objectives we hope to apply visualization techniques, statistical algorithms and libraries like pandas, scikit-learn to fit our regression model.

Bibliography:

1. Kang, G. K., Gao, J. Z., Chiao, S., Lu, S., & Xie, G. (2018). Air quality prediction: Big data and machine learning approaches. *Int. J. Environ. Sci. Dev*, *9*(1), 8-16.
2. Li, X., Peng, L., Hu, Y., Shao, J., & Chi, T. (2016). Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, *23*(22), 22408-22417.
3. Yi, X., Zhang, J., Wang, Z., Li, T., & Zheng, Y. (2018, July). Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 965-973).

Contributions:

Ran Jiang (33%): Discussed about the topics, read previous papers to get familiar with the background, and wrote the introduction part of the project.

Zhanna Sakayeva (33%): Found an appropriate dataset, wrote about the idea to address the problem and discussed the methods which will be used in our project.

Mengjia Zeng (33%): Discussed about the topics, searched for some papers, wrote challenges and the origin for our project.