MALAYSIA DATA INNOVATION TALENT x DOSM DATATHON 2023

"Breathing in Malaysia: An In-Depth Analysis of Pollutant Origins, Contributors & Impacts in Malaysia"

by

Team Lionhearts

Ang Xin[1], Chan Hui Wen[1], Ewen Lim Yuan Khai[1], Ng Jun Fei[1]

[1]School of Mathematical Sciences, Faculty of Engineering, University of Nottingham Malaysia, 43500, Semenyih, Malaysia

Under the Guidance of

Dr Gloria Ai Hui Teng

School of Mathematical Sciences

University of Nottingham Malaysia

Date of Submission

3rd September 2023

1. Introduction

Air pollution is the contamination of the outdoor or indoor environment with any chemical, physical or biological agent that alters the quality of the atmosphere. Although the Industrial Revolution has advanced our technology and society, it has also introduced a huge amount of pollutants into the air that pose a risk to human health. The primary atmospheric pollutants are carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen oxides (NO_x), ozone (O_3) and particulate matter (PM). According to WHO's latest data, 99% of the global population breathe air that exceeds the WHO guideline limits, with the low- and middle-income exposed to the worst [1].

The current trend of industrialisation and urbanisation has reached dangerous proportions worldwide. Anthropogenic air pollution causes the death of around 7 million people each year, making it one of the top contributors to human death [2, 3] and the world's largest environmental risk factor for disease and premature death, which has not made significant progress since 2015 [4]. These pollutants have harmful effects on both human health and the environment in the short- and long-term. The effects of air pollution on human health is far reaching but it primarily increases the risk of developing respiratory and heart diseases, with children and the elderly being more susceptible [5]. Furthermore, because many of the pollutants are greenhouse gases, climate change and air pollution are closely related. As a result, the increase in air pollution causes global warming, resulting in altered weather patterns, sea level rise and more frequent and severe extreme weather events.

Air pollution has reached a critical point of posing a significant threat to public health, the environment and the overall quality of life. Addressing this problem requires strategies to mitigate emissions, improve air quality monitoring and promote sustainable practices in all sectors.

By examining the relationships between anthropogenic activities and atmospheric quality with the current data and trends, this report aims to highlight the major contributors to air pollution in Malaysia and suggest potential solutions to this global concern.

2. Literature Review

2.1 Measures to Reducing Air Pollution

The study in [6] focuses on the impact of social distancing measures implemented in Seoul, South Korea during the COVID-19 pandemic on air quality, particularly on particulate matter ($PM_{2.5}$) and other pollutants. The authors analysed short-term air quality data collected during the 30 days before and after social distancing, comparing it to historical patterns in the past five years.

The analysis results show that the concentration of CO and NO_2 pollutants decreased by more than 15% during the period of social distancing. The diurnal variation of $PM_{2.5}$ concentrations showed lower levels during social distancing, particularly during high human activity hours suggesting reduced human activity on air quality was more pronounced during periods of normally high activity. Reduction in $PM_{2.5}$ concentration was observed across most monitoring stations in Seoul during social distancing, especially in densely populated areas and industrial complexes, indicating the influence of local emissions.

The study suggests that a measure as weak as social distancing, compared to a lockdown, can lead to significant improvements in air quality, emphasising the importance of human activity in the role of air pollution. The paper acknowledges the need for further research to investigate factors such as meteorological conditions and to quantitatively separate the contributions of local emissions and long-range transport to air quality changes. However, there are certain aspects and limitations that the paper does not fully address. The paper primarily focuses on the measurement of air quality and changes in pollutant concentration in 30 days before and after social distancing measures, which is a relatively short time frame. While the paper does a detailed analysis of Seoul's air quality, it does not compare its results with other cities or regions that have implemented similar measures of social distancing. A comparative analysis could provide better insight into the relationship between human behaviour changes and air quality improvements.

2.2 Factors Affecting Air Quality

The study in [7] aims to investigate the complex relationship between urbanisation, land use changes and air quality trends in Kuala Lumpur, Malaysia. The authors analysed data collected on ambient air quality, population, land use and traffic control which were obtained from the Department of Environment Malaysia and the Ministry of Works Malaysia. The objectives are to understand the temporal trends of air quality levels, urban growth and the relationship between these variables.

The analysis revealed a significant and strong relationship between the number of unhealthy/hazardous air quality days and various urban parameters such as shopping and office spaces, industrial units and traffic volume. Contrary to the belief that transboundary pollution is the primary contributor, this study suggests that urban land uses are a significant factor in Kuala Lumpur's air quality deterioration. The paper also highlights the increase in unhealthy/hazardous air quality days despite improvements in good air quality days, highlighting the importance of addressing air pollution issues.

The study underscores the role of urbanisation, land use changes and associated activities in shaping air quality trends. The findings indicate that addressing pollution from various sources within the city itself, particularly traffic-related emissions and commercial activities is crucial for improving air quality in Kuala Lumpur while emphasising the need for sustainable urban planning to mitigate air quality degradation. Despite that, the paper relies on historical data for air quality, population, land use and traffic volume and does not discuss the potential limitations of using older data and how changes over time might have influenced the results. The paper also does not compare its findings with similar studies conducted in other cities with similar air quality issues. Such comparisons could provide a broader perspective on the unique factors influencing Kuala Lumpur's air quality.

The study in [8] highlights the significant role of livestock housing as a source of particulate matter (PM) emission and the associated environmental and health concerns.

The study found that livestock housing is a notable source of PM emissions, with high concentrations of PM adversely affecting the air quality of both indoors and outdoors. The inhaled particles from livestock housing can pose health risks to both humans and animals, resulting in various respiratory diseases.

In conclusion, the reviewed literature emphasises the need for a better understanding of the sources, characteristics and impacts of PM emissions from livestock housing. Addressing these issues is crucial to mitigating the environmental, health, and welfare risks associated with particulate matter in the context of modern livestock production systems. The paper, however, primarily focuses on the pig and poultry houses when discussing PM concentrations and factors influencing them. It could broaden its scope by including other livestock types and geographical regions, allowing for a more comprehensive understanding of PM emissions across different contexts. The paper does not delve into the specific data sources used in the reviewed studies or the methodologies employed. Providing more information about the quality and limitations of the data sources and methodologies would improve the paper's credibility.

3. Methodology

The average gas composition by state and year is used as the primary data and is analysed against possible determinants which are the number of livestock, type of energy supply, percentage of forested area and daily average traffic. The overall average gas composition of Malaysia is compared against GDP by year. The goal of the analysis is to explore the relationships between the average gas composition of air pollutants and the determinants proposed by the findings of the literature review and common consensus.

The data is mainly sourced from the data catalogue of the Department of Statistics Malaysia (DOSM) [9] with other reputable sources outlined by in Table 1:

Table 1*: Data Source and Extracted Variables*

Data Source	Features
DOSM	Monthly concentration of pollutants, Number of Livestock, Temperature
OpenDOSM [10]	Gross Domestic Product (GDP)
Ministry of Transport Malaysia	Average Daily Traffic
Environment Statistics	Energy production

The data collected was cleaned and processed in Microsoft Excel, Microsoft Power Query and Python. It was analysed and visualised in Microsoft PowerBI. All missing data was replaced with the value of 0 and data for $PM_{2.5}$ is only available from 2018 onwards as $PM_{2.5}$ only started being monitored during 2017. The analysis was carried out on data from 2017 to 2021. During the data modelling process, association tables for year and state were used to mitigate and manage many-to-many relationships within the dataset, transforming complex relationships into more manageable one-to-many or many-to-one associations. This ensured the integrity and efficiency of our data model.

An abnormality analysis was performed using the Spectral Residual with Convolutional Neural Network (SR-CNN) algorithm developed by Microsoft [11]. The algorithm borrows the Spectral Residual algorithm used in computer vision for saliency detection which identifies areas in an image that stand out visually [12]. The analysis was performed with a 90% sensitivity on the average gas composition of air pollutants to identify any data points that deviate from the pattern between 2017 to 2021 so further investigation can be done on them to understand the possible causes of these anomalies.

A prediction for the average gas composition of air pollutants from 2022 to 2030 was performed using the AAA version (additive error, additive trend and additive seasonality) of the Exponential Triple Smoothing (ETS) [13]. The algorithm is a more noise-resistant version of the Holt-Winters' seasonal algorithm which is an exponential smoothing method that incorporates linear trend and seasonal components into its prediction [14]. The prediction is based on data from 2017 to 2021. The model uses a 12-month seasonality that allows for the algorithm to assume a recurring pattern for each year. The projected average gas composition in the atmosphere has a confidence interval of 95% where the numeric value of the interval increases as time gets further and the possible variation increases. The prediction was performed to determine the condition of air quality in Malaysia in the future so estimates for future air quality milestones can be made.

4. Findings

4.1 Trend of Pollutants Among States

Air pollutant concentrations in Malaysian states exhibited distinct trends over the years. In 2019, all pollutants, except for NO_2, reached their peak levels across the nation but experienced a significant decline in 2020. This decline was primarily attributed to the COVID-19 pandemic, which necessitated a nationwide lockdown, leading to reduced road traffic, industrial production, and human activities. However, in 2021, pollutant concentrations rebounded, although they did not reach the same levels as in 2019, as lockdown measures were progressively relaxed.

Among the states, Wilayah Persekutuan (WP) consistently had the highest CO concentration in 2017, 2020, and 2021, while Pulau Pinang and Selangor recorded the highest CO levels in 2018 and 2019, respectively. In contrast, NO_2 concentration showed a consistent decline in most states over the years, with WP generally having the highest levels, except in 2021 when Selangor took the lead. All states experienced a notable increase in O_3 concentration from 2017 to 2018, with Melaka and Perlis consistently recording the highest levels among the states from 2018 to 2021. $PM_{2.5}$ concentration steadily rose from 2018 to 2019 in all states before reaching a minimum in 2020. After 2020, there was little change in $PM_{2.5}$ concentration across the states. Melaka consistently had the highest sulphur dioxide SO_2 concentration from 2017 onward, with Perlis having the highest in 2017. SO_2 concentrations did not show a significant decrease in 2020 for many of the states, indicating relatively stable levels during the pandemic-induced lockdown.

4.2 Key Factor Correlation Analysis

4.2.1 Average Daily Traffic

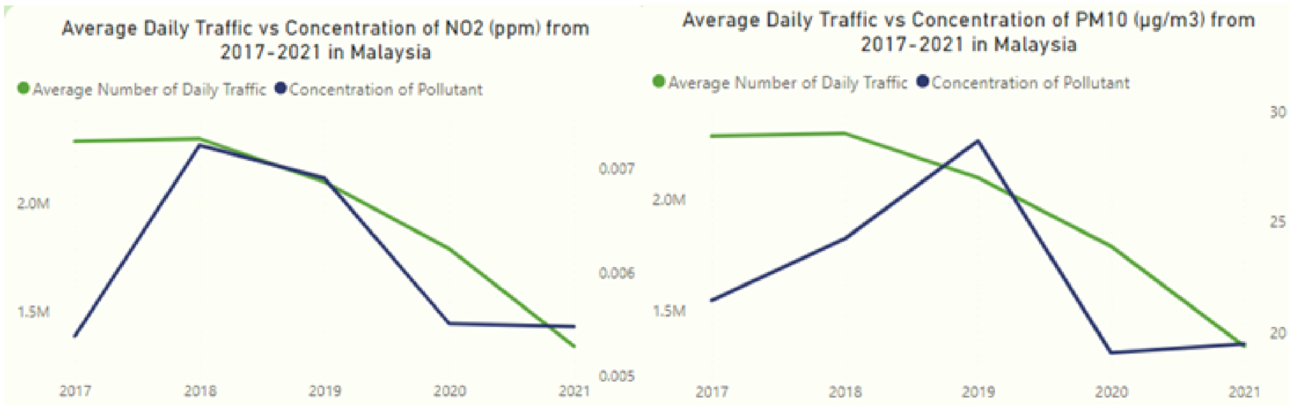

Figure 1: *Average Daily Traffic vs NO_2 and PM_{10} concentration*

From Figure 1, it is observed that average daily traffic steadily decreased across all states from 2017 to 2021. As expected, O_3 shows no correlation with average daily traffic since transportation emissions do not contribute to O_3 concentration. There is a strong correlation between NO_2 and average daily traffic, especially in WP, which had the highest NO_2 concentration. Overall, CO concentration in Malaysia exhibits a slight correlation with average daily traffic. $PM_{2.5}$ concentration follows a similar pattern to CO regarding its correlation with average daily traffic in Malaysia. PM_{10} shows a strong correlation with average daily traffic, particularly pronounced in WP. SO_2 exhibits no correlation with average daily traffic. Interestingly, SO_2 levels appear to increase in certain states, such as Pahang and Kelantan, in 2020.

4.2.2 Temperature

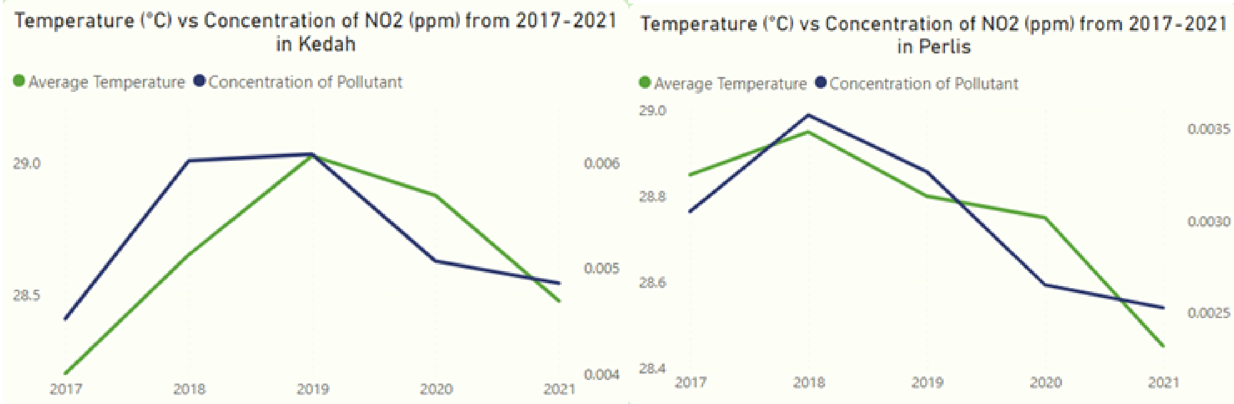

Figure 2*: Temperature vs NO_2 concentration in Kedah and Perlis*

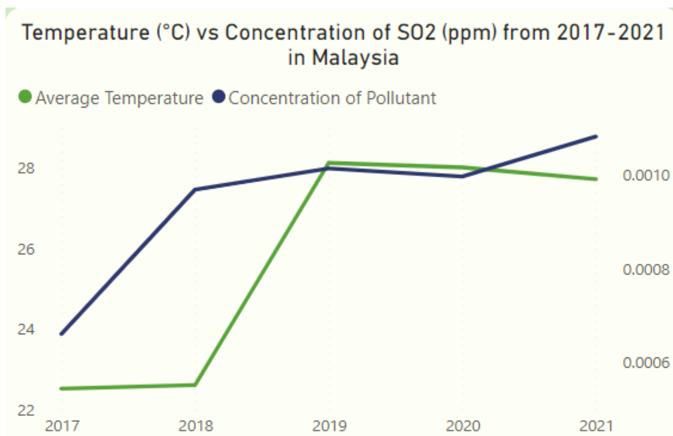

Figure 3*: Temperature vs SO_2 concentration*

As seen in the left side of Figure 3, the overall temperature trend increases gradually except for a sharp increase during 2019. O_3 shows a mild correlation to temperature although in Kedah, Melaka, Perak and Perlis it shows a strong correlation. CO shows a stronger correlation to temperature than O_3. NO_2 overall does not show much of a correlation with temperature but in Figure 2 Perlis and Kedah it had a strong correlation. SO_2 shows a mild correlation and a strong correlation in Terengganu and Negeri Sembilan. There was a common upward trend for PM_{10}, $PM_{2.5}$ and the temperature from 2017 to 2019 The sudden drop in PM_{10} and mild drop in $PM_{2.5}$ in 2020 showed no common pattern with temperature.

4.2.3 Energy Production

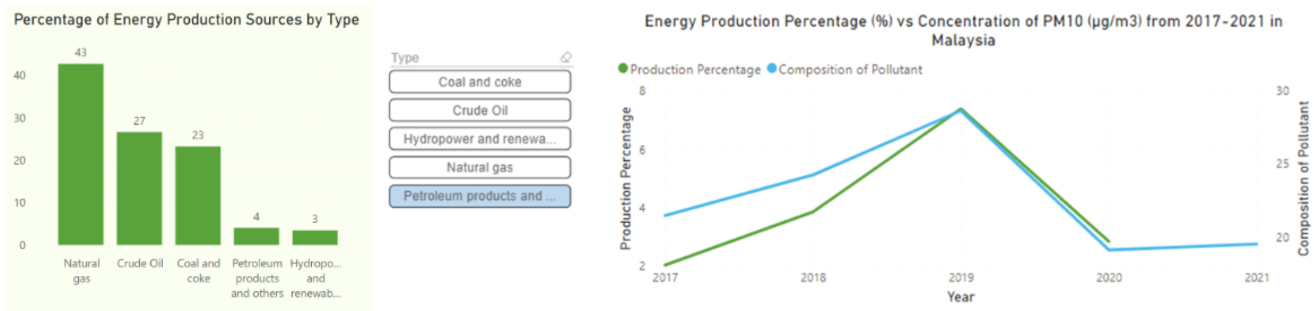

Figure 4: *Energy Production vs PM_{10}*

As seen in the left bar plot Figure 3, natural gas has been the highest energy source from 2017 to 2021 making up around 42 - 44% of the energy produced, while crude oil usage for energy production has decreased by 5% in the same period. Coal, coke, hydrocarbon and other renewables have had an overall increase in energy production. Petroleum products and others are observed to have a correlation to all pollutant concentrations but have the strongest correlation to CO and PM_{10}. Although only contributes on average 4% to the energy production of Malaysia, fluctuations in its production percentage have a significant impact on all pollutant concentrations, especially CO and PM_{10}. Other forms of energy production have little to no observable relation to the pollutant concentrations.

4.2.4 GDP and Number of Livestock

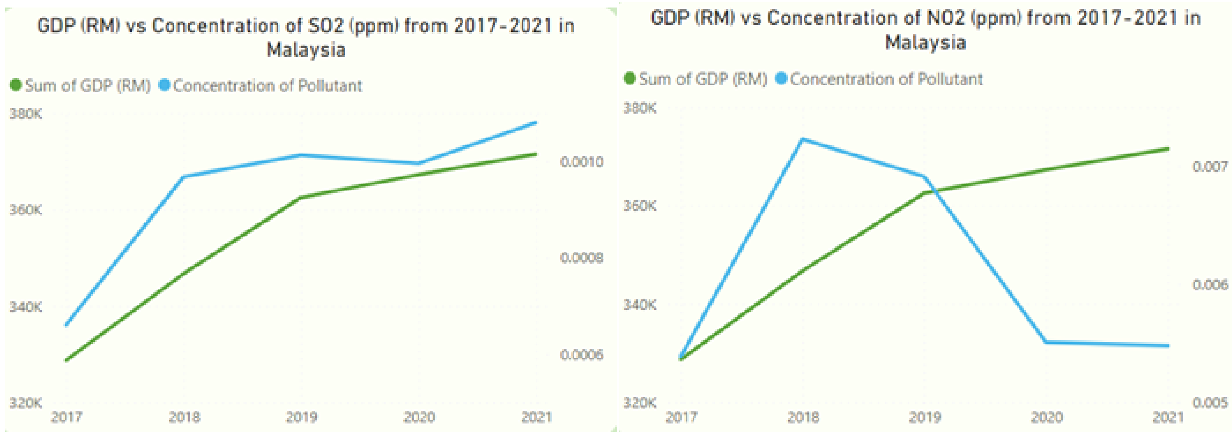

Figure 5*: GDP vs SO_2 and NO_2 concentration*

The GDP of the country increased throughout the years of 2017 to 2021. All pollutants except NO_2 increased with GDP from 2017 to 2019. All pollutants except SO_2 experienced a sudden drop during 2020 which caused deviation from the pattern. SO_2 has a strong correlation to GDP. NO_2 does not show any correlation with GDP. This is illustrated in Figure 4.

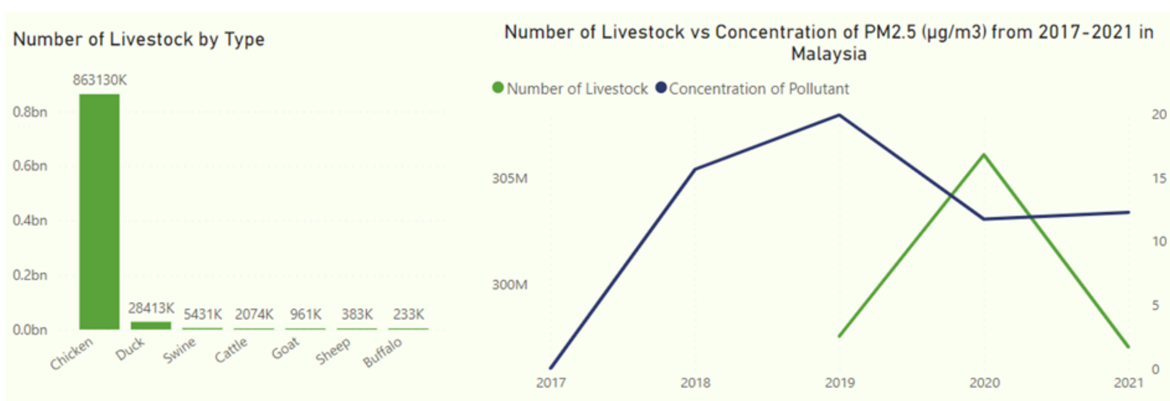

Figure 6*: Number of Livestock vs $PM_{2.5}$ concentration*

Figure 5 shows that chicken has the highest number of livestock in Malaysia, around 23 times more than all the other livestock combined. In 2020, there was an unusual spike in the amount of livestock production. None of the gases show any correlation to the number of livestock.

4.3 Anomaly detection

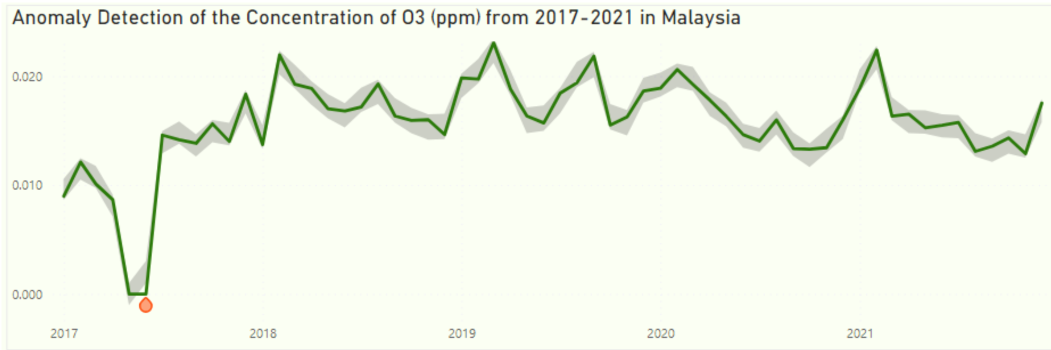

Figure 7*: Anomaly Detection for O_3*

In Figure 6 it can be seen that the abnormality analysis identifies a common point of abnormality across every type of gas in the second quarter of 2017 which is always at 0. This is due to the unavailability of data during the first and second quarter of 2017 which has been imputed as 0.

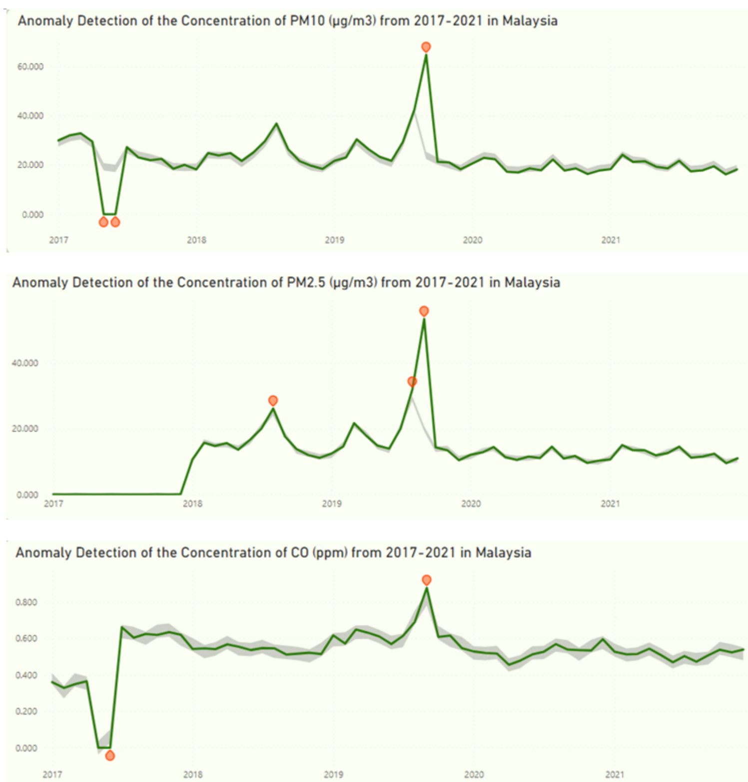

Figure 8*: Anomaly Detection of* PM_{10}, $PM_{2.5}$ *and CO*

Figure 7 shows an abnormal spike in pollutant composition for pollutants PM_{10}, $PM_{2.5}$ and CO throughout Malaysia can be seen during the 3rd quarter of 2019. This may be due to the uncontrolled forest burning in Jakarta, Indonesia [15], resulting in the 3rd quarter of 2019 being a season of haze in Malaysia and other Southeast Asia countries.

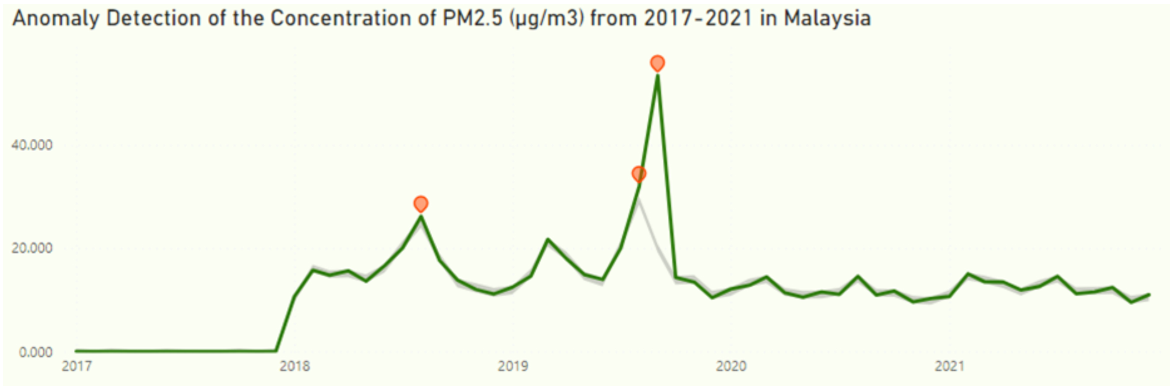

Figure 9*: Anomaly Detection for* $PM_{2.5}$

Another abnormality of $PM_{2.5}$ can be seen in Figure 8 in the third quarter of 2018. The cause of the anomaly is unknown.

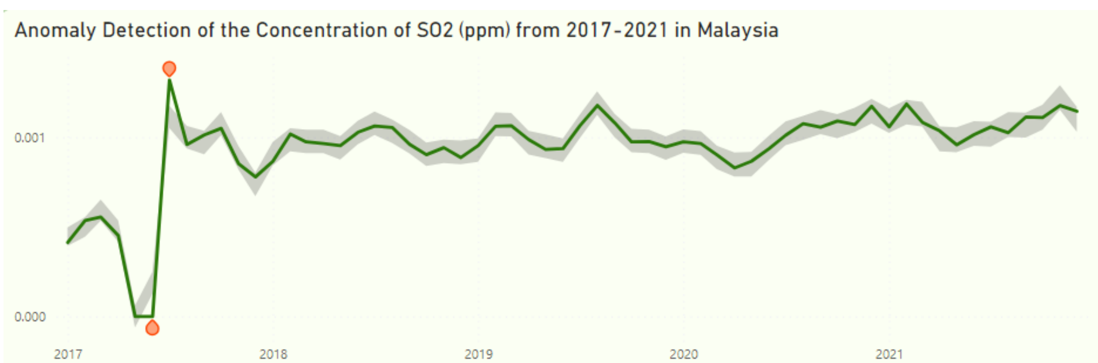

Figure 10*: Anomaly Detection for* SO_2

SO_2 experienced an anomalously high spike in the 3rd quarter of 2017 as shown in Figure 9. The cause of the anomaly is unknown.

4.4 Prediction

The prediction for the pollutant composition from 2022 to 2030 shows that all pollutant concentration follows a trend throughout the year, indicating that air pollution may be seasonal in Malaysia. A common trend of a drop in emissions can also be seen during the 2nd quarter of each year. This could be due to the monsoon season in Malaysia which brings heavy rainfall throughout the country, reducing pollutant concentration in the atmosphere.

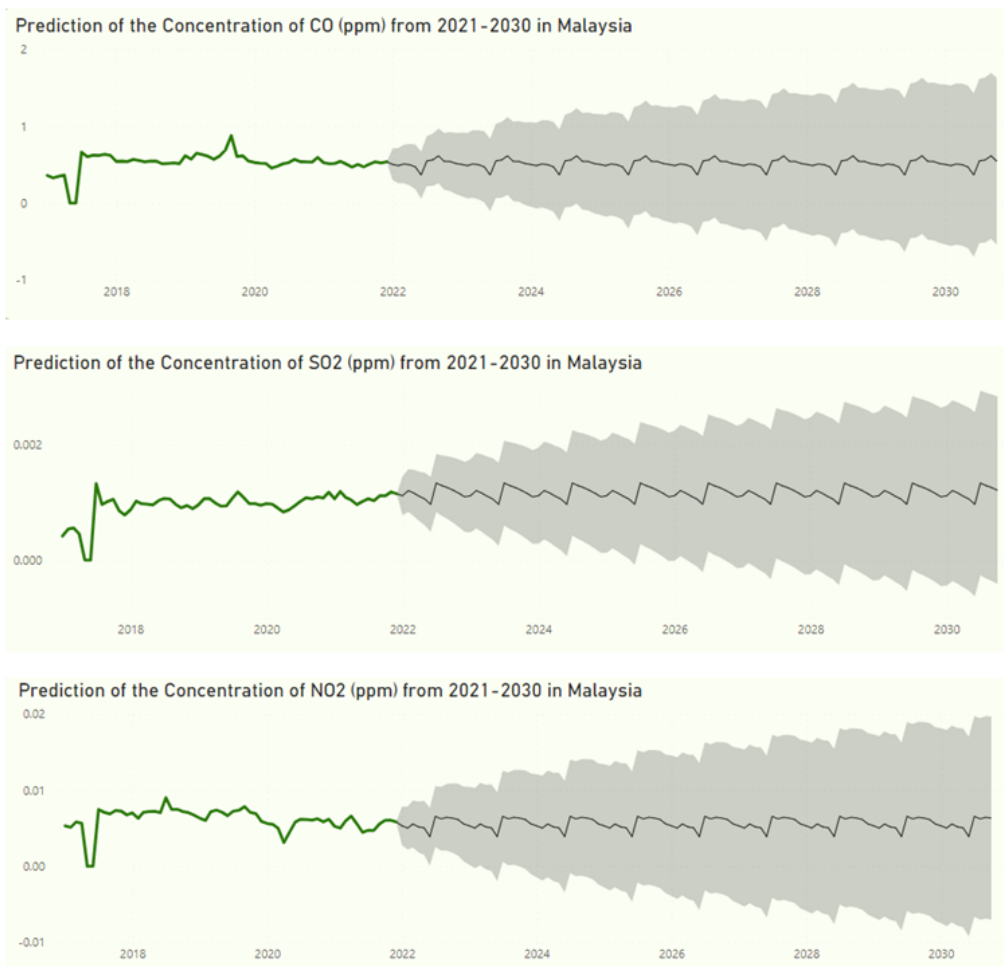

Figure 11*: Predictive model of CO, SO_2 and NO_2*

CO, SO_2 and NO_2 predictions all show a similar trend of a minimum predicted pollutant composition during the 2nd quarter of the year followed by a maximum in the 3rd quarter of the year.

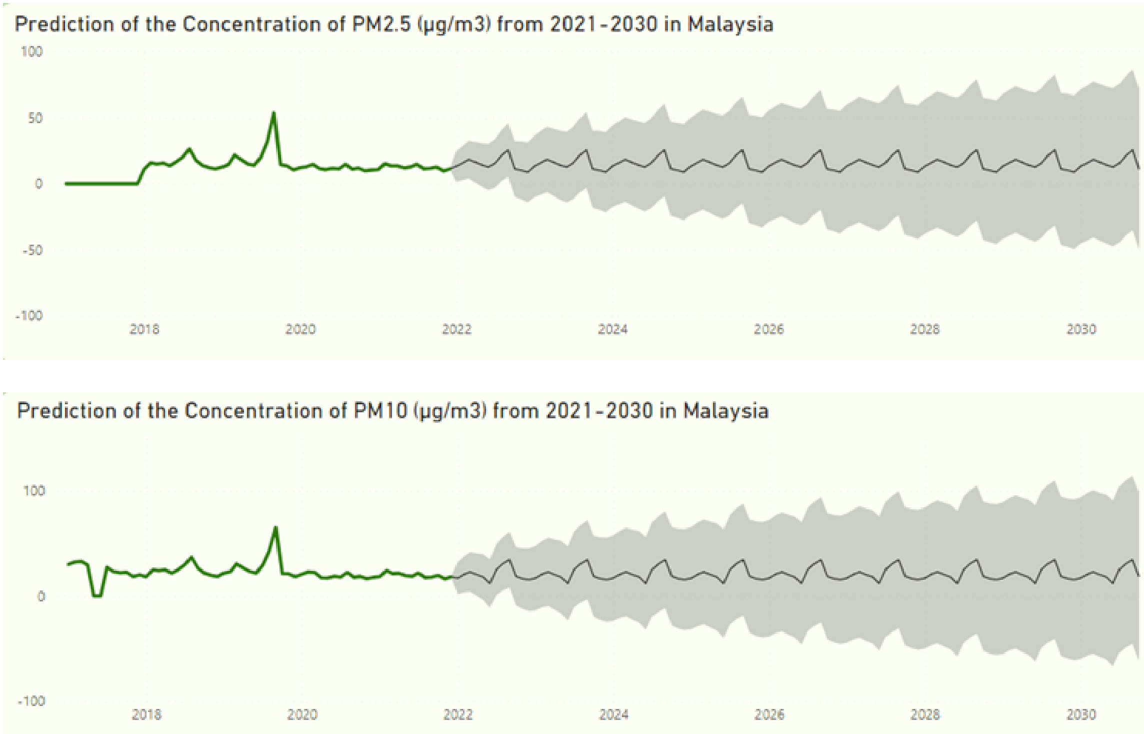

Figure 12*: Predictive Model of* $PM_{2.5}$ *and* PM_{10}

$PM_{2.5}$ and PM_{10} have similar oscillating patterns between quarters. The 1st quarter of the year is a local maximum, the second quarter is a local minimum and so on. $PM_{2.5}$ has steeper and sharper changes compared to PM_{10}.

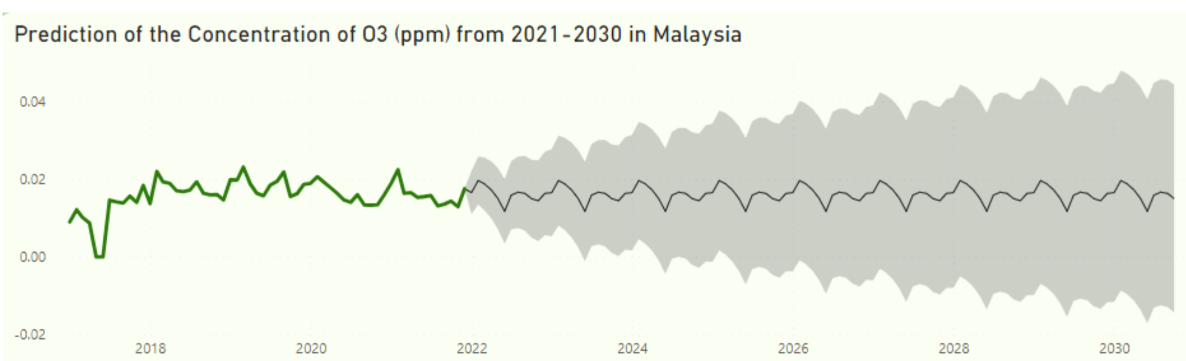

Figure 13*: Predictive Model of* O_3

O_3 has its pattern where the 1st quarter has the maximum concentration and the 2nd quarter has the minimum composition. The concentration of O_3 steadily rises with a local minimum from the 4th quarter to the 1st quarter of the following year.

5. Output

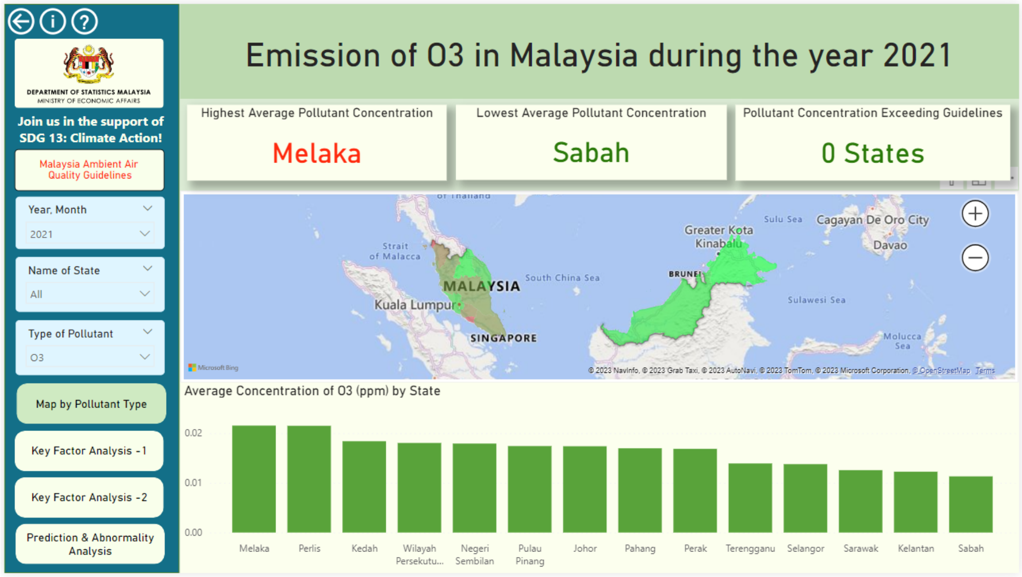

Figure 14*: Front Page of the dashboard (Map by Pollutant Type)*

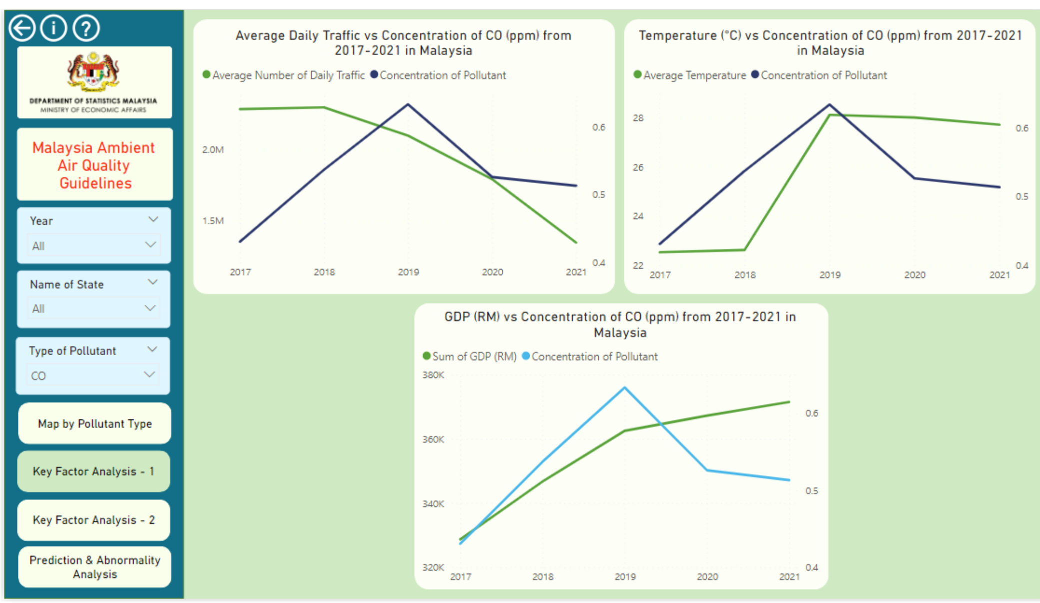

Figure 15*: Second page of the dashboard (Key Factor Analysis 1)*

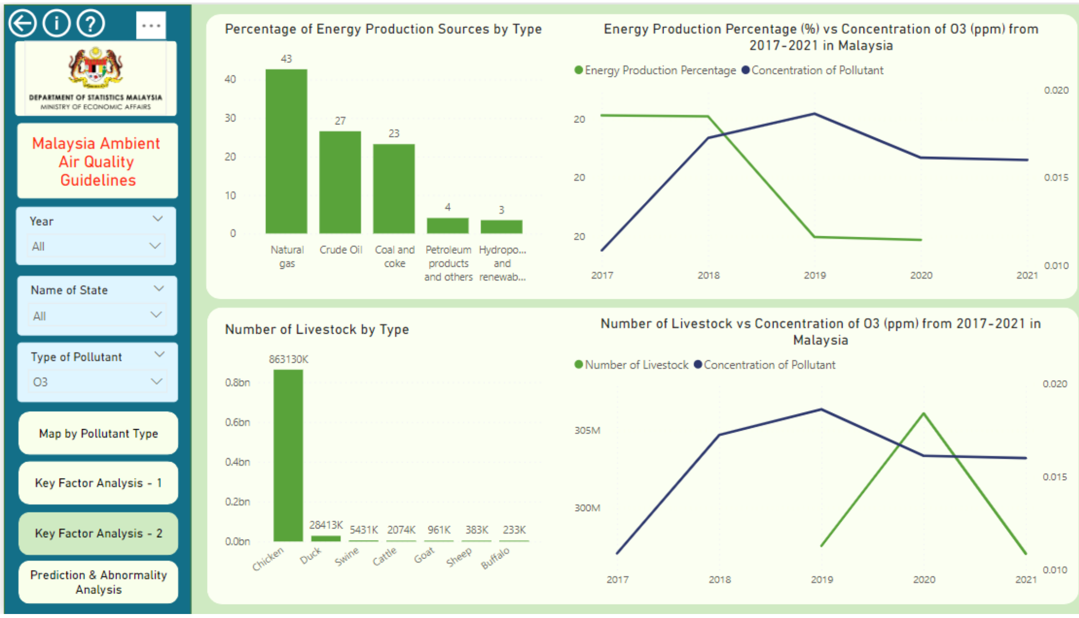

Figure 16: *Third page of the dashboard (Key Factor Analysis 2)*

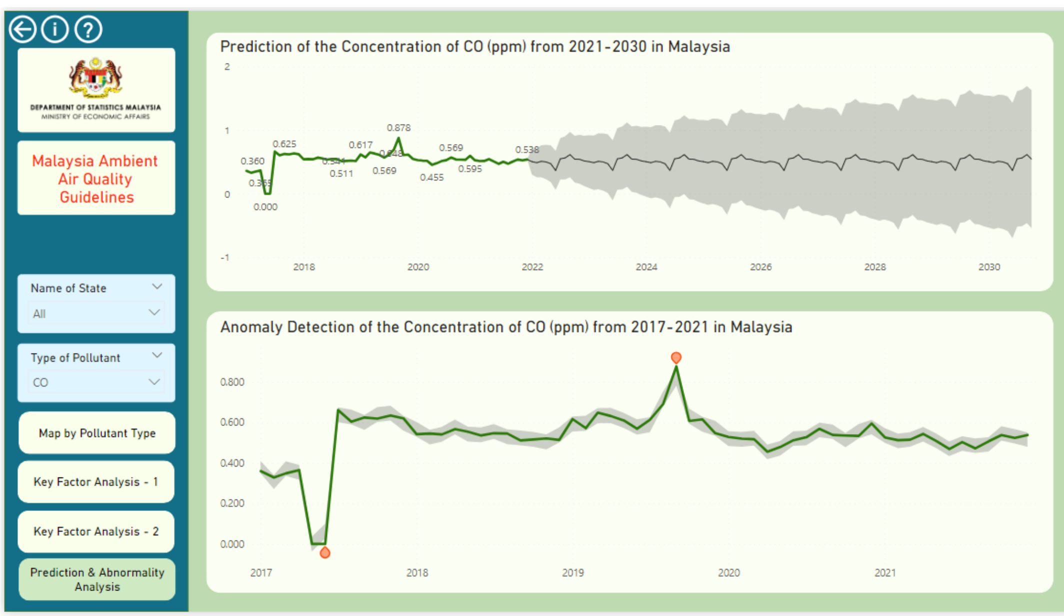

Figure 17: *Fourth page of the dashboard (Prediction & Abnormality Analysis)*

As shown in Figures 14, 15, 16 and 17, the produced dashboard displays clear purpose & objectives with great visual clarity. All graphs are displayed in a neat and concise manner with minimal clutter and the dashboard allows users to apply filters to customise the displays of the data. Dynamic titles and interactive elements such as filters, and dynamic hover-over tooltips, are also included to enhance users' experience by allowing them to understand what they are looking at and help users to gain additional insights. Finally, the dashboard contains data analysis using machine learning algorithms for users to analyse and gain additional insights

6. Conclusion

The analysis shows that the trend that is seen in [6] Seoul, South Korea during the COVID-19 pandemic can also be found in Malaysia. The graphs show a drop in every air pollutant except O_3 during 2020 when lockdowns and social distancing measures were implemented which could be due to the reduction in daily traffic and energy production that can be seen at the same time.

As seen in Figure 8, the transboundary haze has a prominent effect on Malaysian air quality that appears to have a bigger effect than the study [7] makes it seem. The analysis that has been performed in this paper appears to contradict the findings of the study [8]. No correlation between the number of livestock produced could be found with any of the air pollutants from 2019 to 2021, however, this could be due to the lack of data but the lack of clear methodology in the study could also be the cause.

The prediction model identifies that every air pollutant has a 12-month cycle where the second quarter of the year is seen to have a reduction of air pollutants possibly due to the monsoon season.

Due to the limited amount of data, this paper is unable to produce a comprehensive analysis of the historical data on air pollution in Malaysia. The air quality in Malaysia should continue to be monitored to accumulate data that can be used to identify trends and form correlations with possible factors. With the determining factors of air pollution identified, action can be taken to minimise the damaging effects.

7. Appendix

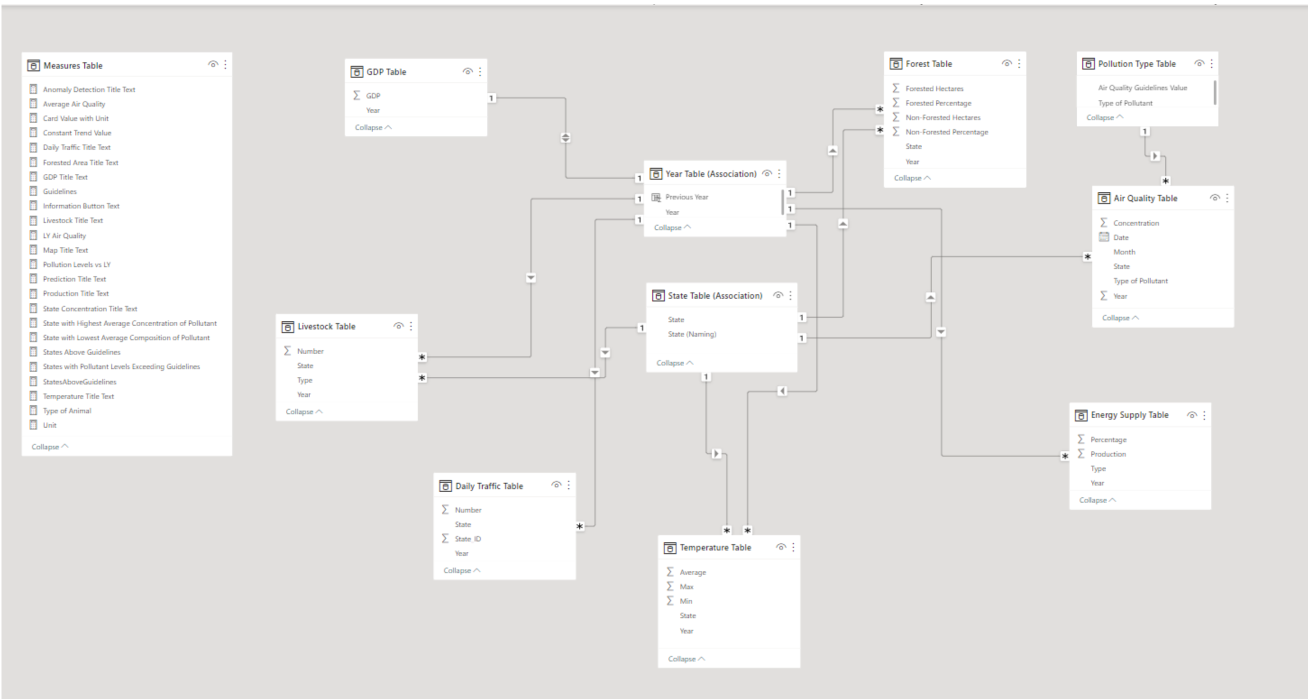

Figure 18*: Data Model for Dashboard*

```
import pandas as pd

#Convert the data to a pandas DataFrame
df = pd.DataFrame(data, columns=["State", "Year", "Jan", "Feb", "March", "April", "May", "June", "July", "August", "Sept", "Oct", "No

#Melt the DataFrame to reshape the months and their values into two columns
melted_df = pd.melt(df, id_vars=["State", "Year", "Type of Gas"], var_name="Month", value_name="Value")

#Filter out rows with "—" values
melted_df = melted_df[melted_df["Value"] != "—"]

#Display the reshaped DataFrame
print(melted_df)
```

Figure 19*: Python Code*

8. References

[1] Air pollution. (n.d.). World Health Organization (WHO). Retrieved September 2, 2023, from https://www.who.int/health-topics/air-pollution

[2] Lelieveld, J., Pozzer, A., Pöschl, U., Fnais, M., Haines, A., & Münzel, T. (2020). Retrieved September 2, 2023, from Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective. Cardiovasc Res, 116(11), 1910-1917.

[3] Lelieveld, J., Klingmüller, K., Pozzer, A., Burnett, R. T., Haines, A., & V., R. (2019). Retrieved September 2, 2023, from Effects of fossil fuel and total anthropogenic emission removal on public health and climate. Proceedings of the National Academy of Sciences of the United States of America, 116(15), 7192-7197.

[4] Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J., Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B., Lichtveld, M., Yan, C. (2022). Retrieved September 2, 2023, from Pollution and health: a progress update. *The Lancet. Planetary health*, *6*(6), e535–e547.

[5] Health consequences of air pollution on populations. (2019). *World Health Organization (WHO)*. Retrieved September 2, 2023, from https://www.who.int/news/item/15-11-2019-what-are-health-consequences-of-air-pollution-on-populations

[6] Han, B.-S., Park, K., Kwak, K.-H., Park, S.-B., Jin, H.-G., Moon, S., Kim, J.-W., et al. (2020). Air Quality Change in Seoul, South Korea under COVID-19 Social Distancing: Focusing on $PM_{2.5}$. *International Journal of Environmental Research and Public Health*, *17*(17), 6208. MDPI AG. Retrieved September 2, 2023, from https://www.mdpi.com/1660-4601/17/17/6208

[7] Ling, O. H. L., Ting, K. H., Shahruddin, A., Kadaruddin A., and Yaakob, M.J. (2010). Urban Growth and Air Quality in Kuala Lumpur City, *Malaysia. Environment Asia*, 3(2). Retrieved September 2, 2023, https://dlc.dlib.indiana.edu/dlc/bitstream/handle/10535/6482/urban%20growth.pdf?sequence=1&isAllowed=y

[8] Cambra-López, M., Aarnink, A., Zhao, Y., Calvet, S., & Torres, A. P. (2010). Airborne particulate matter from livestock production systems: A review of an air pollution problem.

Environmental Pollution, *158*(1), 1–17. Retrieved September 2, 2023, https://sci-hub.mksa.top/10.1016/j.envpol.2009.07.011

[9] Dosm. (n.d.). *Department of Statistics Malaysia*. Retrieved September 2, 2023, https://newss.statistics.gov.my/newss-portalx/ep/epProductFreeDownloadSearch.seam?cid=18527&pid=18519

[10] Department of Statistics Malaysia. (n.d.). *Data Catalogue | OpenDOSM*. OpenDOSM. Retrieved September 2, 2023, from https://open.dosm.gov.my/data-catalogue

[11] Ren, H., Xu, B., Wang, Y., Chen, Y., Huang, C., Kou, X., Xing, T., Yang, M., Tong, J., & Zhang, Q. (2019). Time-Series Anomaly Detection Service at Microsoft. *Time-Series Anomaly Detection Service at Microsoft*. Retrieved September 2, 2023, from https://doi.org/10.1145/3292500.3330680

[12] *Overview of SR-CNN algorithm in Azure Anomaly Detector*. (n.d.). TECHCOMMUNITY.MICROSOFT.COM. Retrieved September 2, 2023, from https://techcommunity.microsoft.com/t5/ai-customer-engineering-team/overview-of-sr-cnn-algorithm-in-azure-anomaly-detector/ba-p/982798

[13] Team, P. B. (2018, December 10). Describing the forecasting models in Power View. *Blog de Microsoft Power BI | Microsoft Power BI*. Retrieved September 2, 2023, from https://powerbi.microsoft.com/es-mx/blog/describing-the-forecasting-models-in-power-view/#which

[14] *ETS models - statsmodels 0.15.0 (+59)*. (n.d.). Retrieved September 2, 2023, from https://www.statsmodels.org/dev/examples/notebooks/generated/ets.html#Holt-Winters%27-seasonal-method

[15] Regencia, T. (2019, September 10). Haze blankets Kuala Lumpur, Singapore as fires rage in Indonesia. *News | Al Jazeera*. Retrieved September 2, 2023, from https://www.aljazeera.com/news/2019/9/10/haze-blankets-kuala-lumpur-singapore-as-fires-rage-in-indonesia