COMP20008 Elements of Data Processing

Semester 1, 2021

Group 147 Fernando Teodoro, Archie Embleton-Mew, Andy Yan

I. Introduction

In the given assignment, the students were set the task of coming up with a proposal regarding wrangling open data following the theme of "issues which are important for communities within Victoria". This was to be tackled in terms of: liveability, health, inclusiveness, and/or sustainability. Group 147 decided to focus the scope on the question of whether there is *correlation between the lack of travel due to Covid-19 and Air Quality in Victoria*. This tackles both health and liveability, in the sense that air quality can be considered a health risk factor if it were to be significantly contaminated with pollutants; and general community health and quality of life contribute to overall liveability of a specific area.

The outbreak of the Covid-19 pandemic provided the world with a unique opportunity to see the effects of significantly decreased human activity on certain aspects of the environment. It is well known that our modes of transport have damaging side effects, in the form of, for example, Carbon Monoxide emission. Given increased restrictions on movement and travel, one could assume there would be significantly less emissions produced by combustion engines. If a correlation were to be found between these two factors, it could provide incentive to work on, for example, electric engines. Given this kind of development, there is potential for this topic to lead itself to sustainability.

II. Datasets

The three main sources of information for the project are the *EPA Air Watch 2020*, *Apple Mobility Trends Report*, and *Victorian Covid Cases* datasets. Aside from the components of air quality measured in the dataset, two important columns present are 'location' and 'datetime'. As for the Mobility Trends Report, once pre-processed, the columns of note were 'date', and 'transportation type'. Finally, for the Covid Cases dataset, the relevant components were 'diagnosis date' and 'postcode'. In order to relate the AirWatch and Covid Cases datasets, the Australian Postcodes dataset by Matthew Proctor. This was done because

Air Watch had their location column as the name of the suburb of the site, whereas the Covid Cases dataset had its location column by postcode.

There were two components in the project: 1) Victoria as a general region; and 2) particular suburbs within Victoria. Because the Mobility Report only had data for the entire region, this could not be analysed in terms of specific areas such as Geelong.

List of Datasets directly related to Air Quality and Movement:

EPA Air Watch | https://www.epa.vic.gov.au/EPAAirWatch Format: xlsx Information: hourly measures of pollutants (such as CO, NO2, SO2) in the air.
Apple Mobility Trends Report | https://covid19.apple.com/mobility Format: csv Information: measures the difference in travel in relation to a baseline, starting 13 January 2020
Victorian Covid Cases by LGA | https://www.dhhs.vic.gov.au/victorian-coronavirus-covid-19-data Format: csv Information: daily records of every Covid-19 case in Victorian LGAs

List of Datasets used to link the aforementioned Datasets:

Australian Postcodes | https://www.matthewproctor.com/australian_postcodes Format: csv Information: data regarding suburbs and corresponding postcodes

III. Data Wrangling and Analysis

Converting the Datasets to Matching Formats

The EPA Air Watch dataset, originally, was in the xlsx filetype. Additionally, it had multiple sheets corresponding to the location of the testing site. Due to the time required to process an XLSX file, the decision was taken to convert each of the sheets into a CSV file. This was done manually through saving each of them from Microsoft Word. The Apple Mobility Report dataset had a very different format in regard to how it recorded the dates and transportation type. 'transportation type' was a column, whereas each date was a column in itself. To standardise the formats, this had to be transposed and outputted into another CSV file.

Cleaning the Datasets

The datasets initially were reduced to the columns as stated in *Part II: Datasets* and outputted to a Victorian CSV or Regional CSV File. Please note that here, 'Regional' could be any region found within the datasets.

Groupby

For the Air Watch dataset, every one of the Regional CSV files from the were combined under conditions corresponding to suburb and date. This would put all the data into one Dataframe. Once this was done, a mean groupby by date was used on this combined Dataframe to change the data hourly data to daily data. For the Covid dataset, a groupby was used to sum up all the covid cases by date.

Combining

Inner and Outer merges were used on a csv file with the 366 days in 2020 to join the processed datasets, to ensure each datapoint had a place in the merged dataset. Also, for the Mobility dataset, an average column was made, which is the mean of 'driving' and 'transit' columns.

Statistical Analysis

For this specific analysis, Footscray was used because among the locations of the Air Quality testing sites, it had relatively rich data.

Prior to statistical analysis, graphs of each column from the merged dataset were generated. This was done to serve as visual aid in choosing which analysis method to use. Most of the graphs produced, visually, looked very scattered. There was not a necessarily clear trend for Victoria and Footscray. This would rule out the use of Pearson's Correlation and Linear Regression. We decided to use Mutual Information (MI) with a 3-bin equal spread as a means of analysis because of this. It could be argued that there was a slightly noticeable trend, in for example, Carbon Monoxide in Victoria. However, this wasn't linear and wasn't necessarily fully scattered, and was thus outside of the scope of what we learned so far in the course. That being said, it was only outside the scope for analysis outside of MI. Furthering the decision to use it.

To perform MI, two components to be correlated were isolated into a single Pandas dataframe. Before this however, outliers were detected and removed through calculating IQR, Upper, and Lower Whiskers. Once that was done, a 'df.dropna()' function was performed on the two-column dataframe to remove inconsistencies between both columns, in order to perform Mutual Information. The output was then printed onto a text file.

IV. Results

Table 1: Movement vs Air Quality Components in Victoria

AIR QUALITY COMPONENTS	CORRELATION COEFFICIENT
BPM 2.5	0.0262
PM10	0.0215
O3	0.0204
NO2	0.0499
CO	0.0520
SO2	0.0293

Table 2: Covid vs Air Quality Components in Victoria and Footscray

AIR QUALITY COMPONENTS	CORRELATION COEFFICIENT	
	Victoria	Footscray
BPM 2.5	0.01640	0.0704
PM10	0.00899	0.0653
O3	0.00484	0.0444
NO2	0.02810	0.0450
CO	0.04490	0.0529
SO2	0.00767	N/A

Figure 1: Average Movement across Victoria in 2020

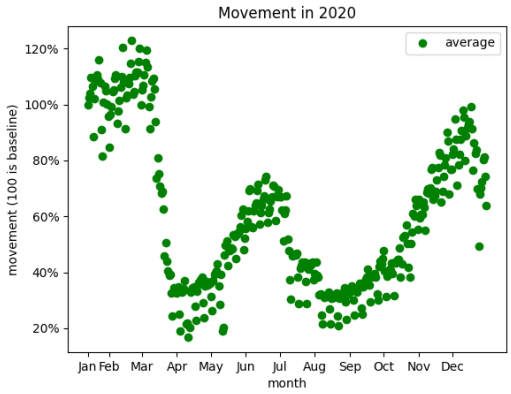


Figure 2: Average Carbon Monoxide across Victoria in 2020

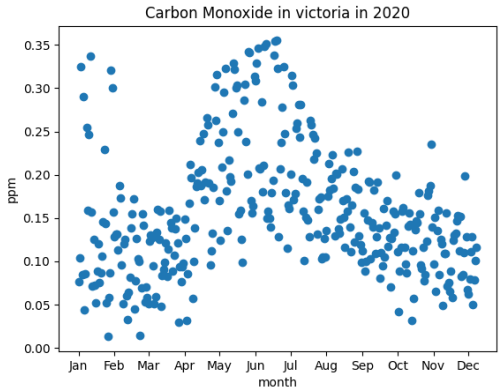


Figure 3: Confirmed Covid Cases across Victoria in 2020

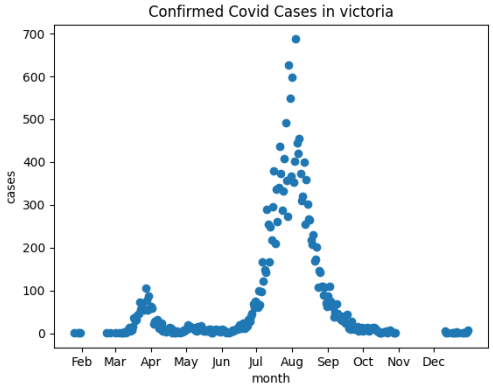


Figure 4: Average Sulphur Dioxide across Victoria in 2020

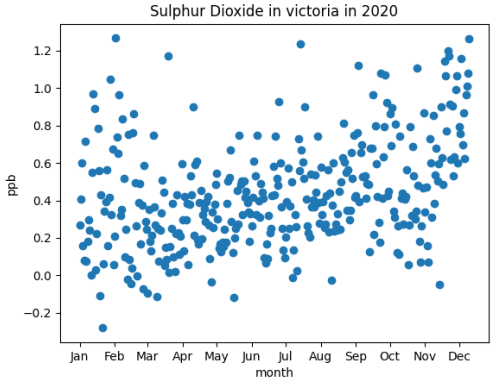


Figure 5: Covid Cases across Footscray in 2020

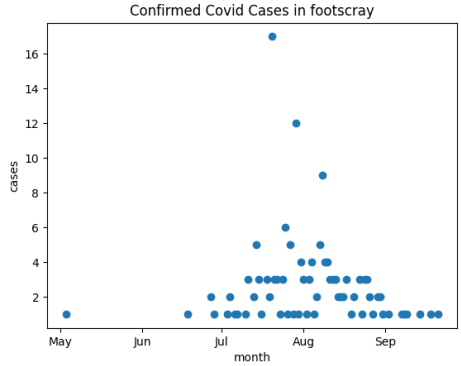
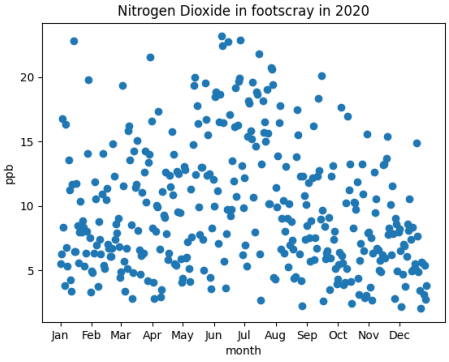


Figure 6: Nitrogen Dioxide across Footscray in 2020



Values on the tables were rounded to three significant figures. There are more components to air quality than what is listed here but can be seen in the appendices.

It is noticeable that NO2 and CO generally have the largest correlations. BPM and PM in Footscray are also notably larger than the average Victorian levels.

V. Significance

One of the bases of the project was under the assumption that, with the decrease in travel due to Covid and lockdowns, that there would be a corresponding decrease in common emissions. From the graphs generated and results found from our analysis, it could be inferred that Covid-19 and the following Movement overall did not have a significant correlative relationship with Air Quality. It could be argued that because people had to stay in their houses, something else (i.e. energy production) had to increase, and thus produced more pollutants potentially to the level of cars pre-Covid-19. The results could provide insight to the role that transportation plays with regard to air quality, and thus may be of benefit to the government, auto manufacturers, etc.

VI. Limitations

One of goals of the team was to be able to analyse specific parts of Victoria. While, in theory and in the code written this is possible, a large number of open datasets only have data on Victoria as a whole. Additionally, there were discrepancies in the data in the sense that, for example, different testing locations in the EPA Air Quality dataset did not all measure the same pollutants. In the future, perhaps more specific and complete datasets could be found.

Due to the relatively small correlations found in the study, alone, finding significance in the data would be limited. It could be useful to examine the data alongside datasets with different conditions (i.e. energy or weather). Moving forward, these could be used to help identify each component's role in environmental air quality.

Regarding the data in the study, in accordance with EPA Victoria[1], the pollutant levels were under the 'good' ranges across the board. Because of this, identifying bins for the

[1] https://www.epa.vic.gov.au/for-community/monitoring-your-environment/about-epa-airwatch/calculate-air-quality-categories

Mutual Information analysis was difficult. We then arbitrarily chose to use a 3-equal-length bin in an outside sourced function to calculate MI. Moving ahead, more research could be done to identify a better bin for more robust analyses.

Appendix

Figure 7: Average PM<=2.5 across Victoria in 2020

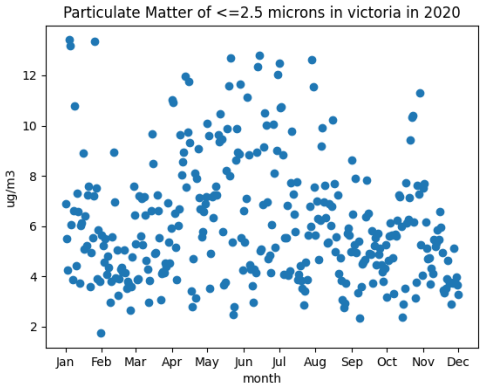


Figure 8: Average PM<=10 across Victoria in 2020

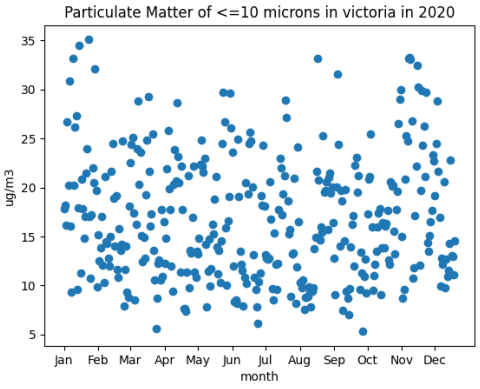


Figure 9: Average Ozone across Victoria in 2020

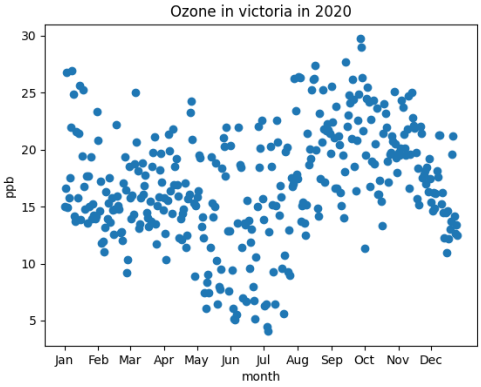


Figure 10: Average Nitrogen Dioxide across Victoria in 2020

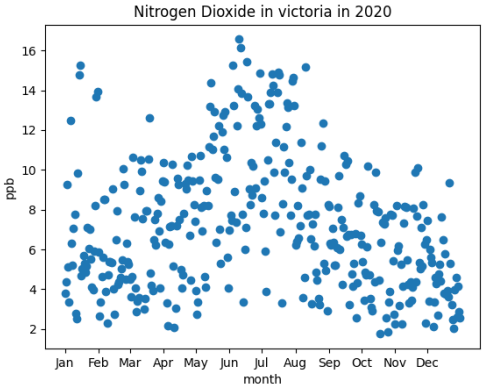


Figure 11: PM<=2.5 across Footscray in 2020

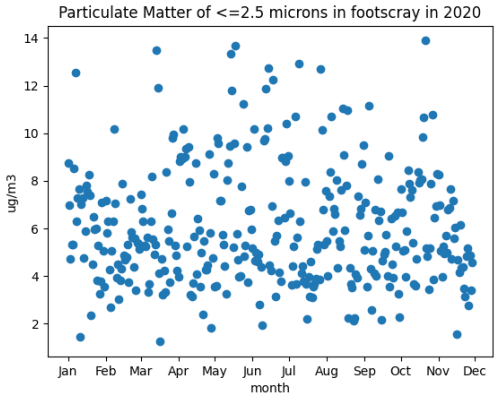


Figure 12: PM<=10 across Footscray in 2020

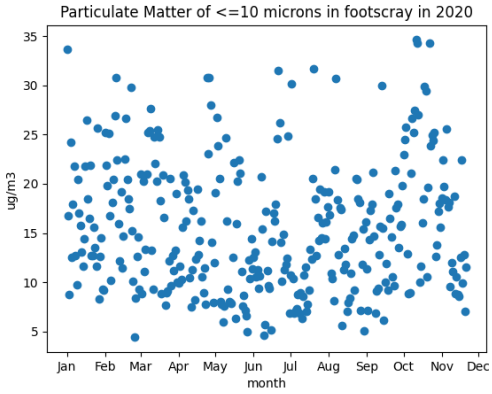


Figure 12: Ozone across Footscray in 2020

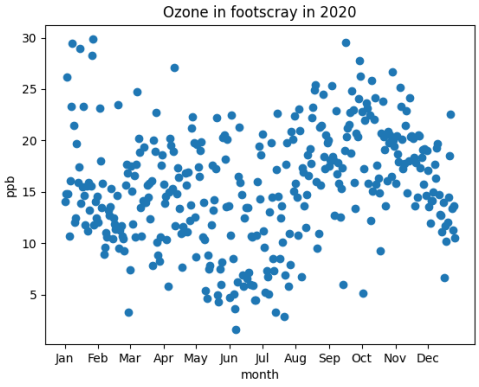


Figure 13: Carbon Monoxide across Footscray in 2020

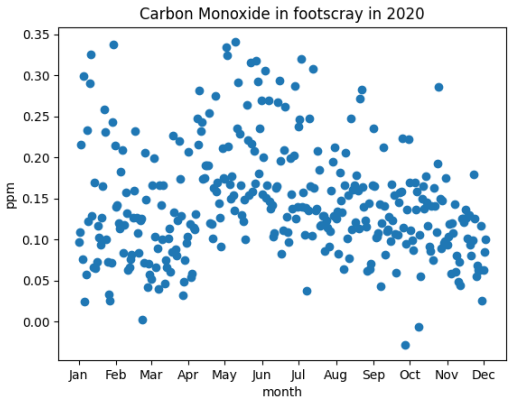


Figure 14: Covid across Footscray in 2020