Project Proposal

Assignment 2 Group 2

Our team will be conducting a project that aims to aid in understanding more about health and inclusiveness by investigating whether there is a relation between net income and COVID-19 case numbers in Victoria.

Considering how the pandemic is still ongoing, it is important for the government, healthcare sectors, Victorians and visitors to remain vigilant so as to prevent another outbreak. Hence, this project will not only assist the former two to know which suburbs are at risk and require the most COVID-19 support, it will also assist in warning Victorians and visitors which suburbs to be cautious of once the lockdown has been eased. Additionally, it will aid the government in knowing which suburbs require monetary support and to ensure that poorer suburbs are not marginalised.

To do so, it will use the datasets found in Table 1. The first two rows will be used to get information on the COVID-19 cases for each suburb. As for the net income for each suburb, we are unable to find a much more recent dataset. Hence, we have decided to use 'Income Tax by Postcode in Australian, 2018-2019' and 'Land Prices'. The former could be used to get information for 2018-2019, while the latter could be used to get information as we have assumed the higher the land prices, the more expensive a suburb is. We will be linking the datasets via postcode information, since it is the common attribute for all of these, albeit in different formats.

Table 1: Database to be used

Database Descriptions	**Links to Data**	**Data Format**
Victorian coronavirus data - Active cases by postcode (updated daily)	https://discover.data.vic.gov.au/dataset/victorian-coronavirus-data/resource/e3c72a49-6752-4158-82e6-116bea8f55c8	Excel CSV
All Victorian SARS-CoV-2 cases by government area, postcode and acquired source	https://discover.data.vic.gov.au/dataset/all-victorian-sars-cov-2-cases-by-local-government-area-postcode-and-acquired-source	Excel CSV

Income Tax by Postcode in Australia, 2018-2019 [Might not need to use]	https://data.gov.au/data/dataset/taxation-statistics-postcode-data	Excel
Land Prices [Might not need to use]	https://discover.data.vic.gov.au/dataset/victorian-property-sales-report-median-vacant-land-by-suburb/resource/f62a7c21-a7a7-4474-8df7-197f9aa6b063	Excel
Postcodes in Victoria [Might not need to use]	https://postcodes-australia.com/state-postcodes/vic https://www.worldpostalcodes.org/l1/en/au/australia/list/r1/list-of-postcodes-in-victoria	Html
Jobkeeper Payment Data	https://treasury.gov.au/coronavirus/jobkeeper/data	Excel

Additionally, it will use various data wrangling techniques to extract information from the datasets. The tables will need to be extracted from their respective excel spreadsheets or csv files. The datasets selected such as income tax use postcode numbers, but others such as land prices, active and total COVID-19 case numbers use the names of each postcode area. Hence, in order to link those datasets, we will use web crawling and scraping to match suburbs with postcodes (from sites mentioned in table above). Then the datasets will be linked by matching suburbs and postcodes. With HTML files, this will be done with regular expressions.

- If HTML files contain tables, pd.read_html()

By using these data wrangling methods, we will be able to combine the raw data more efficiently and generate tables in the csv format; descriptive statistics such as mean, median, mode and standard deviation; and graphs including scatter plots and box plots. Scatter plots will be able to show the spread of cases and income by postcode visually for users of the analysis, while box plots make analysing standard deviation and interquartile ranges easier. With this, the stakeholders will be able to find reading and understanding the information easier and see the

correlation, if there is any, much more effectively. Moreover, it will reduce time, enabling data analysts to focus on analysis instead of spending time on data wrangling. Stakeholders would make concrete and timely decisions.

- Scatter plots won't overly be useful for geospatial analysis (w/ postcodes)
    - try finding a shapefile and researching about geopandas and folium
    - at least use shapefiles and matplotlib

Some possible challenges and risks for undertaking this work include data selection bias and cost of lives. As stated previously, we are unable to find net income data for 2020-2021 and can only find data for 2019. Additionally, the data used prior to 2019 do not include the effects of COVID-19. Finally, lives will be affected if data wrangling techniques are incorrect, since the government and healthcare professionals might not be able to focus on the suburbs that require the most support. Not to mention, Victorians and visitors may be less careful when visiting the affected areas.