Final Report

Assignment 2 Group 2

GitHub Repository Link

https://github.com/COMP20008/assignment-2-comp20008-assignment-2-group-2

Research Question

Is there a relation between income and COVID-19 case numbers of each postcode area in Victoria?

The relationship between COVID-19 cases and income may affect livability in Victoria. Where there is an increase in either case numbers or low income, it has been shown that residents feel increased anxiety about potential outbreaks of disease as well as health problems caused by living conditions. The relationship between income and COVID-19 cases also affects inclusiveness, health, and sustainability of communities. A study by Cash et al. has shown that 90% of reported COVID-19 deaths are in the world's richest countries, not including the numbers of China, Iran and Brazil[1]. It cites many factors, including poverty, the approach to the pandemic and the locations of care homes for the elderly.

COVID-19 is an international health crisis that has been causing worldwide lockdown, quarantine, and some restrictions. The coronavirus outbreak has rapidly spread across the world, posing enormous health and economic challenges for the human population. New strains have led almost every nation struggling to curb transmission

[1] Cash, R., & Patel, V. (2020). Has COVID-19 subverted global health?. *The Lancet*, *395*(10238), 1687-1688.

even with testing and treating patients while quarantining potential individuals with the virus through contact tracing or restricting large gatherings[2].

The goal of this report is to investigate whether there is a correlation between income and COVID-19 cases in Victoria. We aim to understand health and inclusion by examining how the average wealth of individuals in each postcode area might be related with the spread of COVID-19. With the ongoing pandemic, it is important for the government, healthcare sectors and Victorians to remain vigilant so as to prevent another outbreak. This project aims to highlight whether more actions should be taken to improve the health and sustainability of lower income postcodes. It also serves to show which suburbs should remain cautious once lockdowns are eased or lifted and to aid the nation by knowing what areas require more financial support than others rather than marginalising those who need it most.

Datasets

Datasets found in Table 1 are used in our analysis. The first dataset is used to obtain information on COVID-19 cases for each postcode. As for a measure of income for each suburb, we have decided to use the JobKeeper application numbers for a year, where the higher application numbers signify a lower income in the postcode area. We will be linking the datasets via postal code as it is the common attribute. As such, we will be using the html link to access the names of each postcode area for identification.

The shapefile is used to visualise our data further into our analysis to give us an understanding of the actual locations of the spread of the virus and that of the application numbers.

[2] Chakraborty, I., & Maity, P. (2020). COVID-19 outbreak: Migration, effects on society, global environment and prevention. *Science of the Total Environment*, *728*, 138882.

Table 1: Databases to be used

Database Descriptions	**Data Format**	**Links to Data**
Victorian coronavirus data - All cases by postcode (file from 11 September - although updated daily)	Excel CSV	https://discover.data.vic.gov.au/dataset/victorian-coronavirus-data/resource/e3c72a49-6752-4158-82e6-116bea8f55c8
Postcodes and respective names in Victoria	Html	https://www.worldpostalcodes.org/l1/en/au/australia/list/r1/list-of-postcodes-in-victoria
Jobkeeper Application Count by Postcode (April 2020 - March 2021)	Excel	https://treasury.gov.au/coronavirus/jobkeeper/data
Shapefile for Victoria Region	Shapefile	https://s3-ap-southeast-2.amazonaws.com/cl-isd-prd-datashare-s3-delivery/Order_LZMU0O.zip from https://datashare.maps.vic.gov.au/search?md=1553f19f-3b03-5e40-924e-6355eb9a3f89

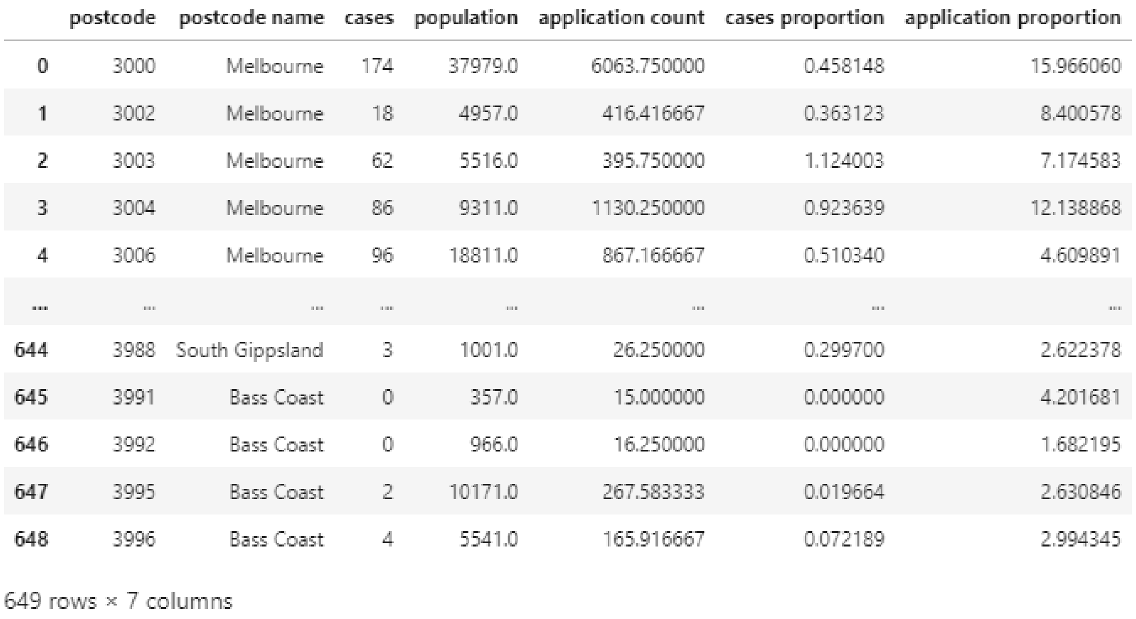
Wrangling and Analysis Methods

Data wrangling is the process of transforming and mapping data from one "raw" form into a more understandable one with the intention to make it more appropriate for analysis.

In order to link these datasets, we used web crawling and scraping techniques on the site listed above to match suburb names with postcodes. We filtered through the html file and obtained this information that is presented in a table on the website. We accessed the excel and csv files by downloading them and using the module *openpyxl* to read each sheet. As the JobKeeper file separated data by month, we collated all 12 months and took an average of application counts across the year. We also extracted the population number and total number of cases for each postcode from the Victorian cases excel csv.

By using these data wrangling methods, we were able to combine the raw data more efficiently and generate a DataFrame that contained the postcode, postcode area name, case numbers, population, and application count. We then added columns that divided the application count and case numbers over the population of the area, producing the columns: cases proportion and application proportion (Table.1). DataFrames allow us to use data we have processed more efficiently during our analysis process as compared to lists or Series.

Table.1 DataFrame created for further analysis



	postcode	postcode name	cases	population	application count	cases proportion	application proportion
0	3000	Melbourne	174	37979.0	6063.750000	0.458148	15.966060
1	3002	Melbourne	18	4957.0	416.416667	0.363123	8.400578
2	3003	Melbourne	62	5516.0	395.750000	1.124003	7.174583
3	3004	Melbourne	86	9311.0	1130.250000	0.923639	12.138868
4	3006	Melbourne	96	18811.0	867.166667	0.510340	4.609891
...	...	...	...	...	...	...	...
644	3988	South Gippsland	3	1001.0	26.250000	0.299700	2.622378
645	3991	Bass Coast	0	357.0	15.000000	0.000000	4.201681
646	3992	Bass Coast	0	966.0	16.250000	0.000000	1.682195
647	3995	Bass Coast	2	10171.0	267.583333	0.019664	2.630846
648	3996	Bass Coast	4	5541.0	165.916667	0.072189	2.994345

649 rows × 7 columns

We printed a correlation matrix and descriptive statistics for our DataFrame to understand the relationship of each column with the others. This is a good preliminary overview to support future analysis and the identification of outliers.

Scatter plots are a great way to show the spread of cases and application count by postcode for users visually. Although it does not show us the geospatial distribution of cases and applications, we use the scatter plot and a corresponding regression line to identify if there is any relationship between case proportion and application proportion. A regression analysis summary is also printed to understand the visual data better.

We also used geopandas to produce a visual representation of the case proportion and application proportion across the entire state of Victoria through the aforementioned shapefile in Table 1. This gives us a more realistic overview of the spread of both income and cases in the state as the scatter plots do not show geospatial data.

These graphs and data have assisted us in concluding whether there is a relationship between income and case numbers in each postal area across Victoria.

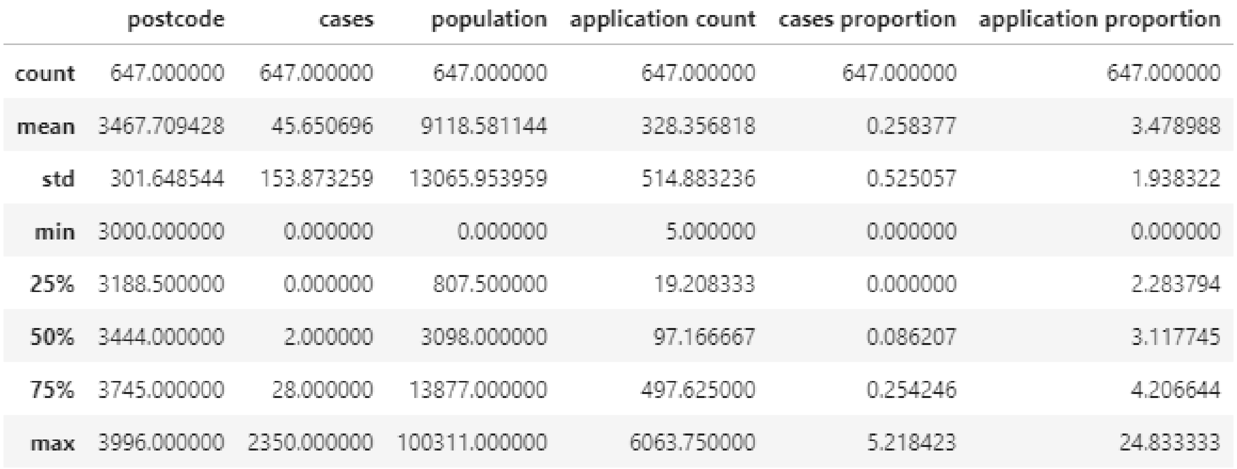
Key Results of Research

Preliminary Analysis of Data

While producing graphs and plots, we noticed a significant outlier through a scatter plot shown in Fig.1. One of the largest outliers being postcode 3026 with an application proportion of 680.56 and case proportion of 156.94 which may have stemmed from an error in the data that shows 72 as a population but has 94 cases of COVID-19.

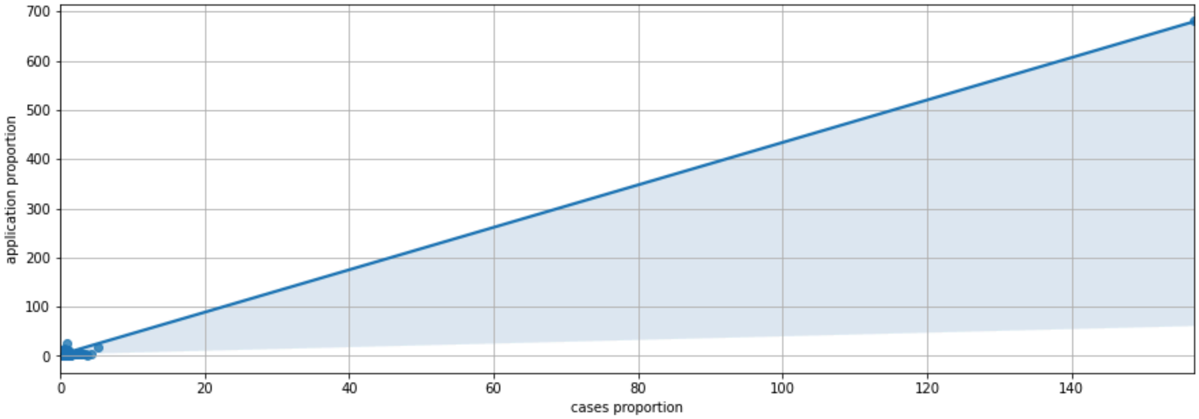
The descriptive statistics produced without this outlier in Table.2 shows a mean of 45 cases for every 9119 people in each suburb, which in turn shows a mean application proportion of 3.47. As this data heavily skewed our graphs, we decided to remove this outlier to prevent further distortion in our analysis.

Table 2 Descriptive statistics of data after data wrangling



	postcode	cases	population	application count	cases proportion	application proportion
count	647.000000	647.000000	647.000000	647.000000	647.000000	647.000000
mean	3467.709428	45.650696	9118.581144	328.356818	0.258377	3.478988
std	301.648544	153.873259	13065.953959	514.883236	0.525057	1.938322
min	3000.000000	0.000000	0.000000	5.000000	0.000000	0.000000
25%	3188.500000	0.000000	807.500000	19.208333	0.000000	2.283794
50%	3444.000000	2.000000	3098.000000	97.166667	0.086207	3.117745
75%	3745.000000	28.000000	13877.000000	497.625000	0.254246	4.206644
max	3996.000000	2350.000000	100311.000000	6063.750000	5.218423	24.833333

Fig.1 Scatter plot of application proportion and cases proportion by postal code



As seen in our scatter plot in Fig.2, it is unclear if there is a relationship between application proportion and cases proportion in Victoria. We took a look at our regression summary to aid our understanding in Table 3. The adjusted R-squared of 0.367 shows that 36.7% of COVID-19 case proportions data can be explained by application proportions and there is a moderate positive relationship between income and case proportions.

Our correlation analysis in Table 4 also supports this moderate positive relationship with a correlation coefficient of 0.252 between application and case proportions.

Fig.2 Scatter plot of application proportion and cases proportion by postal names

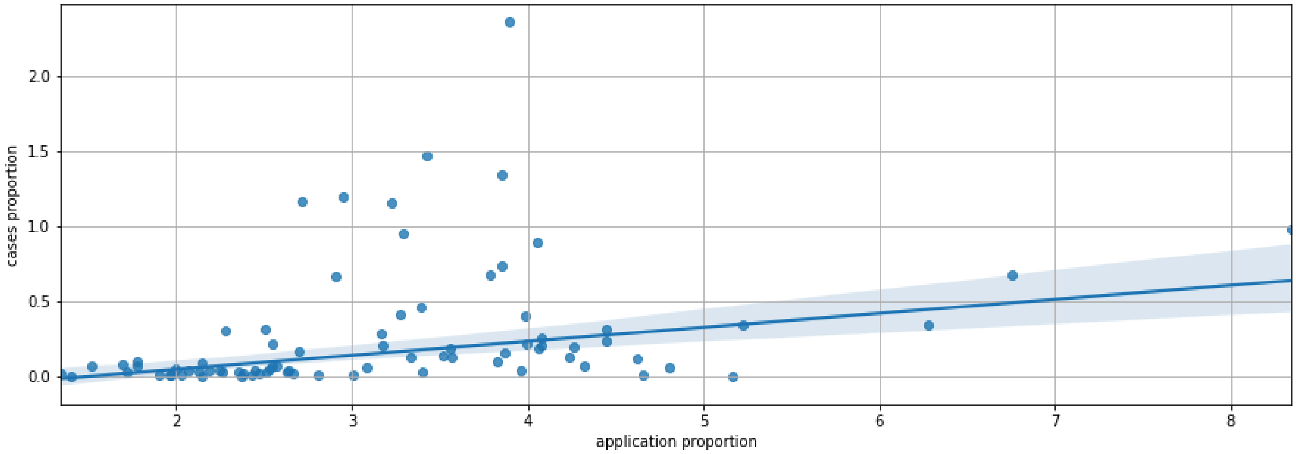
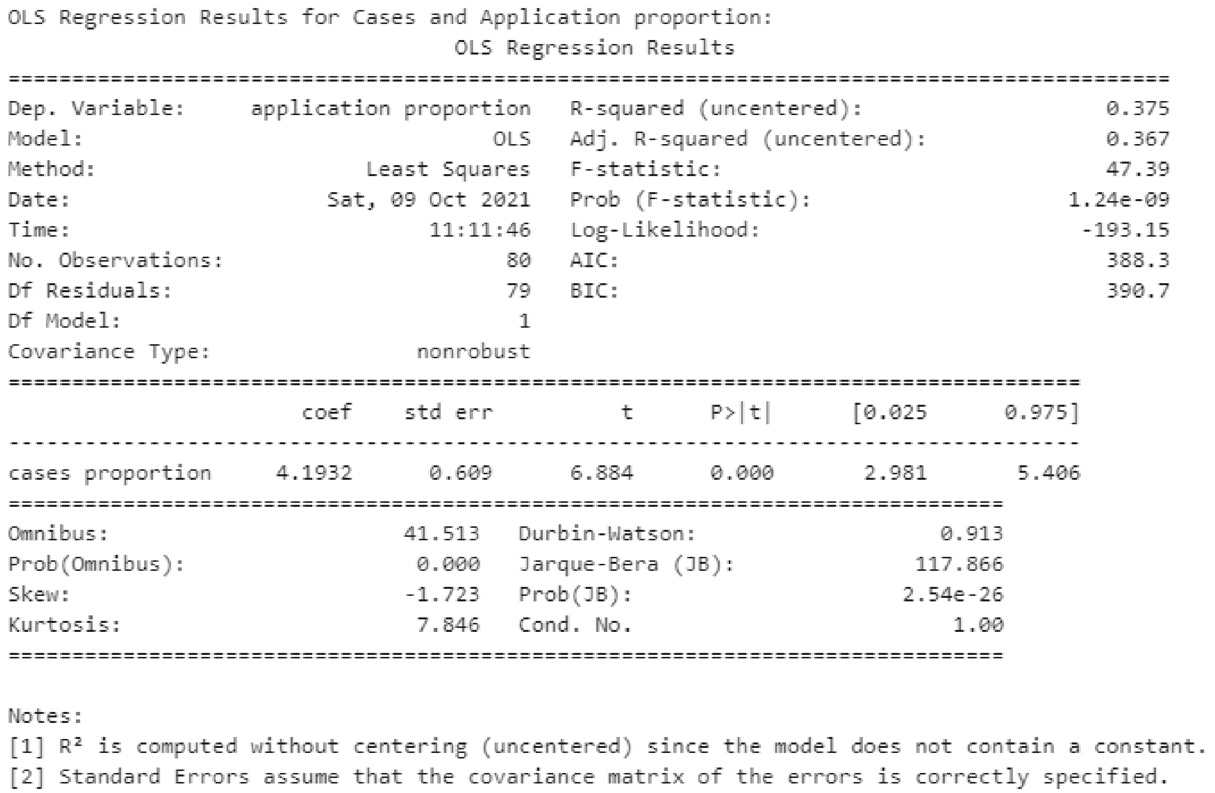


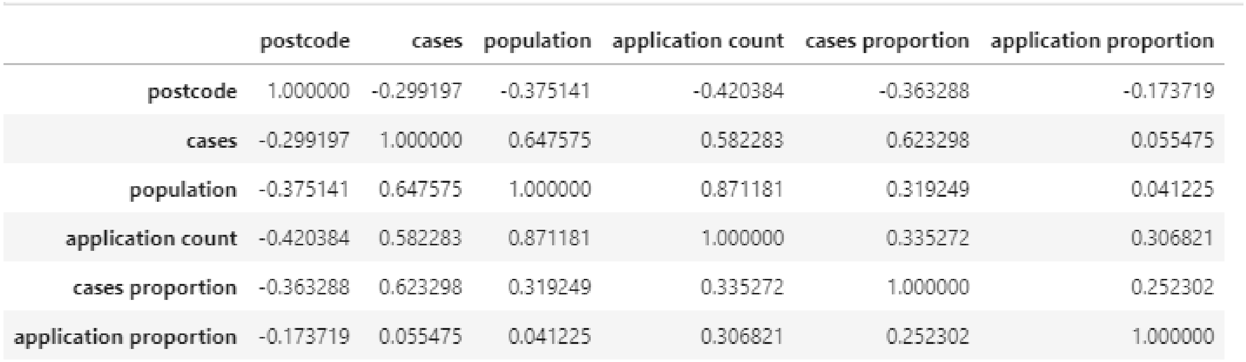
Table 3 OLS Regression Results



```
OLS Regression Results for Cases and Application proportion:
                                   OLS Regression Results                                   
===========================================================================================
Dep. Variable:     application proportion   R-squared (uncentered):                   0.375
Model:                                OLS   Adj. R-squared (uncentered):              0.367
Method:                     Least Squares   F-statistic:                              47.39
Date:                    Sat, 09 Oct 2021   Prob (F-statistic):                    1.24e-09
Time:                            11:11:46   Log-Likelihood:                         -193.15
No. Observations:                      80   AIC:                                      388.3
Df Residuals:                          79   BIC:                                      390.7
Df Model:                               1                                                  
Covariance Type:                nonrobust                                                  
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
cases proportion     4.1932      0.609      6.884      0.000       2.981       5.406
==============================================================================
Omnibus:                       41.513   Durbin-Watson:                   0.913
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              117.866
Skew:                          -1.723   Prob(JB):                     2.54e-26
Kurtosis:                       7.846   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Table 4 Correlation between columns in DataFrame



	postcode	cases	population	application count	cases proportion	application proportion
postcode	1.000000	-0.299197	-0.375141	-0.420384	-0.363288	-0.173719
cases	-0.299197	1.000000	0.647575	0.582283	0.623298	0.055475
population	-0.375141	0.647575	1.000000	0.871181	0.319249	0.041225
application count	-0.420384	0.582283	0.871181	1.000000	0.335272	0.306821
cases proportion	-0.363288	0.623298	0.319249	0.335272	1.000000	0.252302
application proportion	-0.173719	0.055475	0.041225	0.306821	0.252302	1.000000

Geospatial Analysis of Data

We also looked at the geospatial data created via GeoPandas. Based on the two spatial graphs in Fig.3 and Fig.4, COVID-19 cases are prevalent around the central part of Victoria (especially around Melbourne), whereas application proportions are spread evenly across the state. However, some exceptionally high application proportions can be observed from the central part of Victoria, which might be related to the inner parts of Victoria being the ground zero of the COVID-19 virus.

Fig.3 Geospatial data of Case proportions

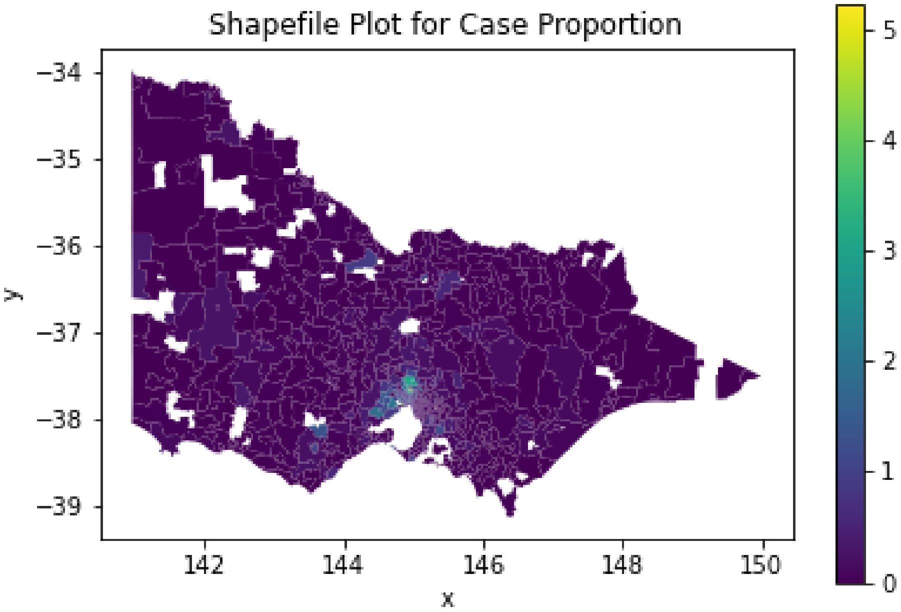
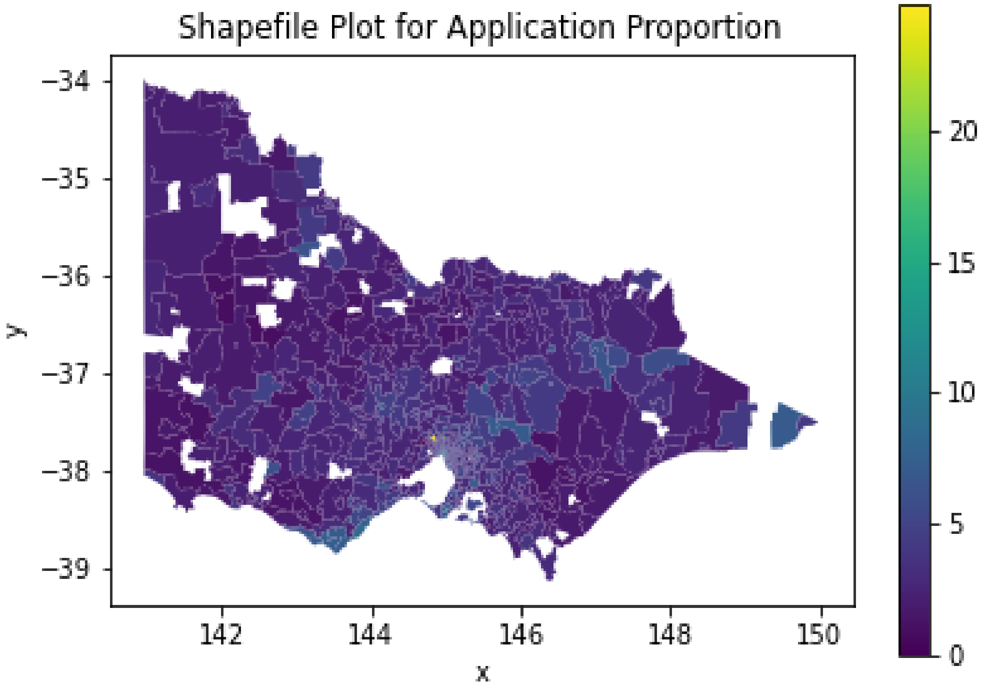


Fig.4 Geospatial data of Application proportions



Significance and value of results

Summary of Analysis

Taking into account the results of the scatterplot and descriptive statistics, we can see a moderately positive relationship between the cases proportion and application proportion. This would signify that as the proportion of applications for JobKeeper increases in a postcode area, the proportion of cases also increases.

This conclusion is somewhat supported by the geospatial data we produced, where there are higher application proportions for areas with more cases. However, we also note that there are areas not as significantly affected by the spread of COVID-19 but still see higher proportions of applications in the postcode area.

We can conclude that there is a somewhat positive relationship between income and case numbers in postcode areas.

Significance

Through this research, we have found that although there is a significant relationship between income and cases of COVID-19, it is not a very strong one. This would mean that the lower income areas of Victoria are not neglected and that Victoria shows inclusiveness and sustainability of its communities during this fight against COVID-19.

Authorities and Victorians alike should continue to ensure that this level of inclusiveness is maintained as the pandemic is likely to continue to have a significant impact in our lives for the foreseeable future.

Limitations and Improvements

One limitation of this project is the lack of net income data from 2020-2021. Although it is reasonable to assume areas with lower income would have higher percentages of applications, it is unverifiable without the actual data. While we were able to improvise by making use of Jobkeeper application counts, it would make the results more significant if we had used relevant net income data.

Another limitation is that the spread of COVID-19 is likely due to many more factors than analysed in our report. Some potential reasons would be the spread of households in a given postcode or with the newly introduced vaccination programme, the percentage of vaccinated people in each postcode area. Hence, our report that focuses on the relationship between income and COVID-19 spread may be skewed by other factors.

We should also note that the scatter plot and regression used do not denote a causal relationship between both variables. It only signifies a relationship, and the spread of COVID-19 cannot be said to depend on income levels.

One way to improve the project for the future would be to use updated and relevant datasets. We could also make use of GeoPandas View or Folium for a more interactive visualisation.

Another key improvement would be to include more potential factors and their correlation with each other to ensure a more robust analysis on the spread of COVID-19 and livability in Victoria.