Viral Vulnerability Analysis

F12D GROUP 8 480452061 490032284 480556587

Dataset Description

The datasets we used in this research mainly come from the Australian Bureau of Statistics (ABS), NSW Ministry of Health (MH) and Open Data Transport NSW within 3 different types: csv, shp and json.

Original data set 1--StatisticalAreas.csv from ABS. This dataset consists of three attributes: area_id, area_name and parent_area_id and 414 columns, which provide the area with a unique id for the research.

Original data set 2--Neighbourhoods.csv from ABS. This raw data set has 312 records with 8 attributes: area_id, area_name, land_area, population, number_of_dwellings, number_of_businesses, median_annual_household and avg_monthly_rent, which provide the area with detailed information.

Original data set 3--PopulationStats2016.csv from ABS. This dataset was extracted in 2016 with 517 samples and primarily consists of the population of different age groups from Sydney areas, also has four attributes which are area_id, area_name, male and female. These describe the population information of an area.

Original data set 4--HealthService.csv from ABS. This dataset is also cosponsored by Health services in NSW. It consists of 3027 samples with 12 attributes which are id, name, category, num_beds, address, suburb, state, postcode, longitude, latitude, comment and website. These provide the health service condition in an area.

Original data set 5--NSW_postcodes.csv from ABS. This dataset has 5640 samples in five attributes: id, postcode, locality, longitude and latitude. These provide specific area coverage with a unique postcode.

Original data set 6--SA2_2016_AUST.shp from ABS. The shapefile has 13 attributes which will provide a polygon to mapping area with all samples.

Original data set 7-- Journey to Work 2011.json from Opendata Transport NSW. This additional dataset has 97363 samples in three attributes: origin, destination and people, which provide the movement of people.

Original data set 8-- NSW COVID-19 tests by location and result.csv from MH. This dataset has 442624 records in 7 attributes, provides the tested people's area and result.

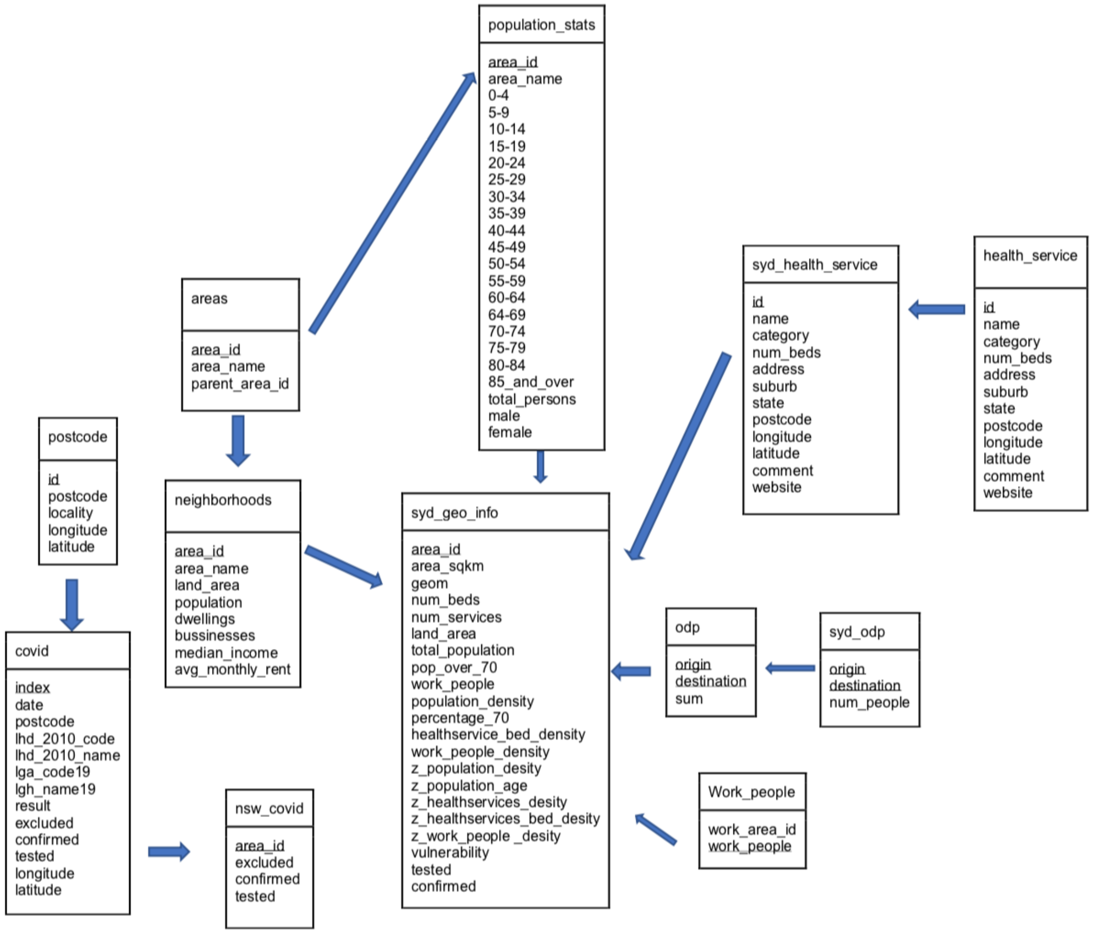
Methods of pre-processing of DATA and Some DATA Limitation

At first, we built our local host in jupyter and connected to the pgadmin, then we loaded all our data sets into the pgadmin database system. Moreover, we have joined our datasets with the same attribute(area_id) and it will be analysed more further in our next step database description. Moreover, all eight raw datasets have the same limitation, like some categories in the dataset have no recorded values, indicating that some non-response bias may have been captured in the data. Also, for example, maybe the data was administered electronically, some people may not have adequate resources in accessing the website allowing for selection bias. Beside the null records were transferred to 0 for easy evaluate

Database Description

We have loaded our raw datasets into 12 database schemas(the table as shown above). In these tables, primary keys indicated with an underline. For the provided datasets, we join it based on area_id and use shapefile to load the map. In order to analyze and show the data clearly, we set syd_geo_info as a master schema and add other details' columns in it such as the total population, health service number and the density measures.

As the additional data set we choose the journey to work dataset and transfer the destination attributes as people workplace, then group by the destination. Using the people as one of the measures to calculate the Vulnerability score. For the Covid19 test data, we use ST_MakePoint to transfer the postcode to area_id and join it to syd_geo_info schema. Tested and confirmed number was added to the master schema for correlation analysis.



Furthermore, we have created 4 indexes and a spatial index, neighbourhoods schema has the area_id index, syd_health_services schema has an index which is called iat_long_idx, covid schema has an index called covid_postcode_idx, syd_geo_info has an index which is called geo_area_id_id and a spatial index called geom_idx. As these columns are mostly used, we created indexes for them to speed up searching in the databases. Therefore, these 5 indexes are crucial for us to join our datasets thereby we can determine vulnerability and correlation for our next analysing work.

Vulnerability Score Analysis

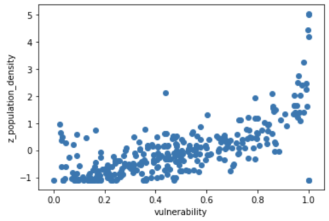
To compute the vulnerability score for all given neighbourhoods, we used the similar formula given in the assignment specification while adding one more variable (work people density) as we integrated an additional dataset. For the percentage of the population over 70, we used the value between 0 - 1. Here is our vulnerability score formula:

$$vulnerability = S(z(population\ density)+z(population\ age)+z(work\ people\ density)-z(health\ \text{service density})-\text{z(hospital bed density))}$$

With S being the sigmoid function, and z the z-score of a measure - the number of standard deviations from the mean.

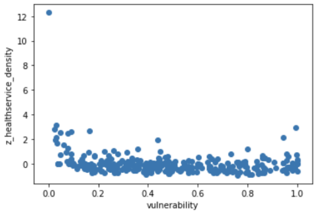
$$z(measure, x) = \frac{x - avg(measure)}{stddev(measure)}$$

To show the relations between the vulnerability score and each measure's z score, we used pearsonr to draw the correlation relationship. The examples are given below representing each part in the formula and an additional density(right) added from our own dataset. We can see that the vulnerability score is mainly affected by the density of population and the density of working people. It's clearly shows the vulnerability score is more effected by population density and work people density. The number of bed is not influence a lot may because there are lots of 0 records. Similar with number health service, null records and distribution uneven may lead to not correlation.



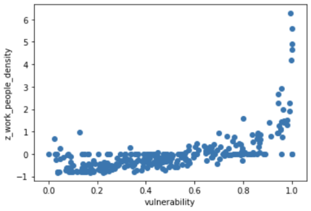
```
correlation: 0.7063375523020066
```

Correlation z_population_density with vulnerability



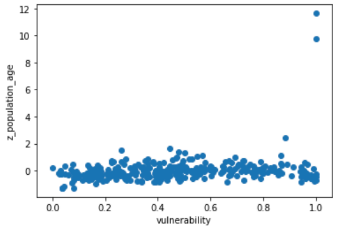
```
correlation: -0.23880753487570675
```

Correlation z_healthservice_density with vulnerability



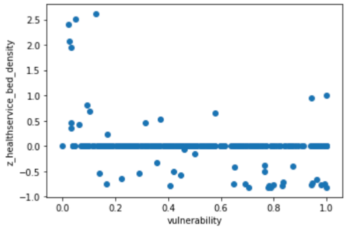
```
correlation: 0.6337753467036823
```

Correlation z_work_people_density with vulnerability



```
correlation: 0.23738308515314366
```

Correlation z_population_age with vulnerability

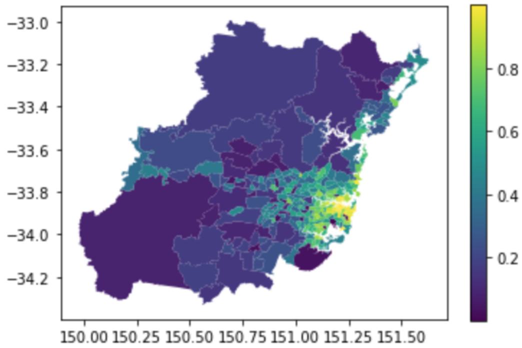


```
correlation: -0.2816192041924447
```

Correlation z_healthservice_bed_density with vulnerability

The map below is a map-overlay visualisation of vulnerability score in Sydney. From the graph, we found that in the middle-east of Sydney, the vulnerability score is high while in the rest of Sydney it is relatively low. The top-right and bottom-left of Sydney have even lower scores. The reason may be that there are more people living and working in the middle-east of Sydney which increases the z scores of population density and working people density. Although it has higher z scores of hospital service density and hospital bed density, they have less impact on the vulnerability than other z scores.

Vulnerability Score in Sydney Area

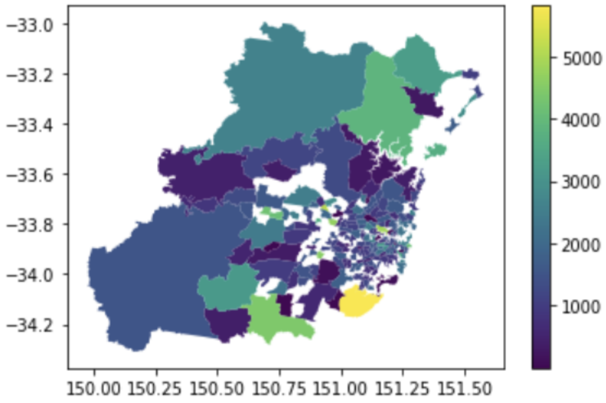


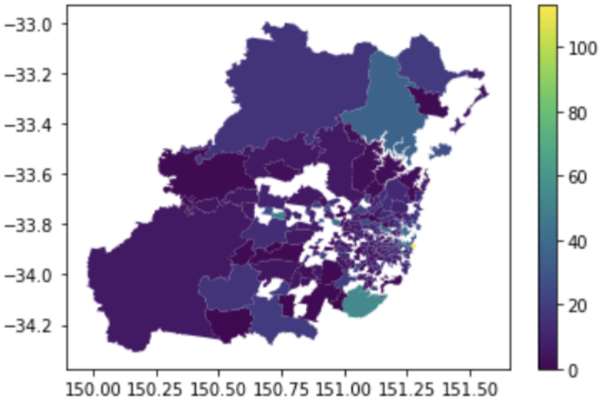
Correlation Analysis and data visualization

Visualization

Maps below show the tested cases and confirmed cases in Sydney. For the left graph, the middle-east of Sydney still shows relatively higher test cases similar to the vulnerability score. However, the yellow part shows a large number of test cases which is even more than its total population, so it's a data mistake and should be ignored. For the right graph, the overall confirmed cases have few differences. It's worth noting that there are still high confirmed cases in the middle-east of Sydney although the area is really small. Empty spaces in the both maps mean that those areas have no test cases or confirmed cases recorded.

Tested Cases in Sydney Area **Confirmed Cases in Sydney Area**





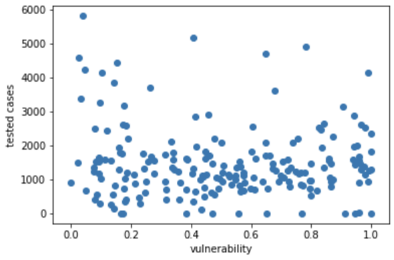
Correlation Analysis

In order to effectively make reasonable comparisons thereby investigating the relationship of Vulnerability between these attributes. We have plotted our attributes on a grid made by two categorical axes by using a scatter plot.

Based on these three scatter plots from below, we have identified that tested cases and confirmed cases have a strong correlation to each other, as in the term of statistical language it looks pretty linear. The reason might be the test cases and confirmed cases are affected by the total population. Also, as a result, this scatter plot clearly shows that as the number of tested cases increases the number of confirmed cases increases regularly too.

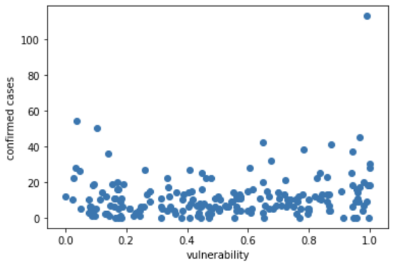
On the other hand, another two scatter plots do not have similar trends like it indicates these attributes in these two scatter plots are not very correlated. Moreover, it also indicates that test cases and confirmed cases don't have a huge impact on our vulnerability score. The reason might be the population density of Sydney is small compared to other global alpha cities and the government social distance restriction based COVID -19 situation, these two main reasons might cause vulnerability are not affected very much by test cases and confirmed cases. To sum up, the above result indicates that the vulnerability score does not depend on tested cases and confirmed cases.

Vulnerability Score and Tested Cases



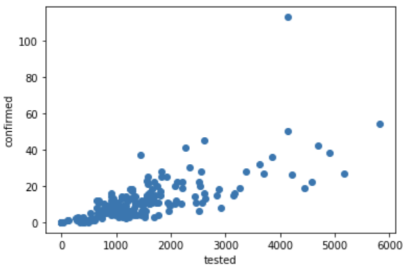
```
correlation: -0.06583281512978366
```

Vulnerability Score and Confirmed Cases



```
correlation: 0.15060919543313717
```

Tested Case and Confirmed Case



```
correlation: 0.7241141651826642
```