DATA2001 Practical Assignment: Viral Vulnerability Analysis

Dataset Description

Data Sources

Datasets were provided by the University of Sydney, obtained from the Australian Bureau of Statistics (ABS) as well as the central source of Australian open government data: Data.gov.au. Datasets from the ABS were primarily census-based datasets. The additional JSON dataset was obtained via a webservice API.

Dataset	Rationale
NSW_Postcodes.csv	Postcodes in NSW with providing geographical information such as locality.
StatisticalAreas.csv	Regions of NSW divided into areas and subdivisions.
Neighbourhoods.csv	Census data of NSW neighbourhoods providing demographical information such as population size, land area and number of dwellings in neighbourhood.
PopulationStats2016.csv	Census data providing demographical information based on area. Includes total population, males, females and senior citizen population.
HealthServices.csv	Health service providers in NSW providing information such as hospital bed capacity and location of these providers.
SA2_2016_AUST.shp	Provides geometries to be used for PostGIS relating to SA2 areas and their boundaries for spatial joins
JSON "Journey to Work 2011"	Number of people commuting from an origin SA2 ID to a destination SA2 ID for work purposes.

Pre-processing

Firstly, after connecting to the PostgreSQL database from Python and verifying that no existing tables existed, we created schemas with appropriate domain types for each table corresponding to the relevant datasets (Section 2).

Data Cleaning

The datasets needed to be cleaned, removing rows that contained irrelevant or invalid values.

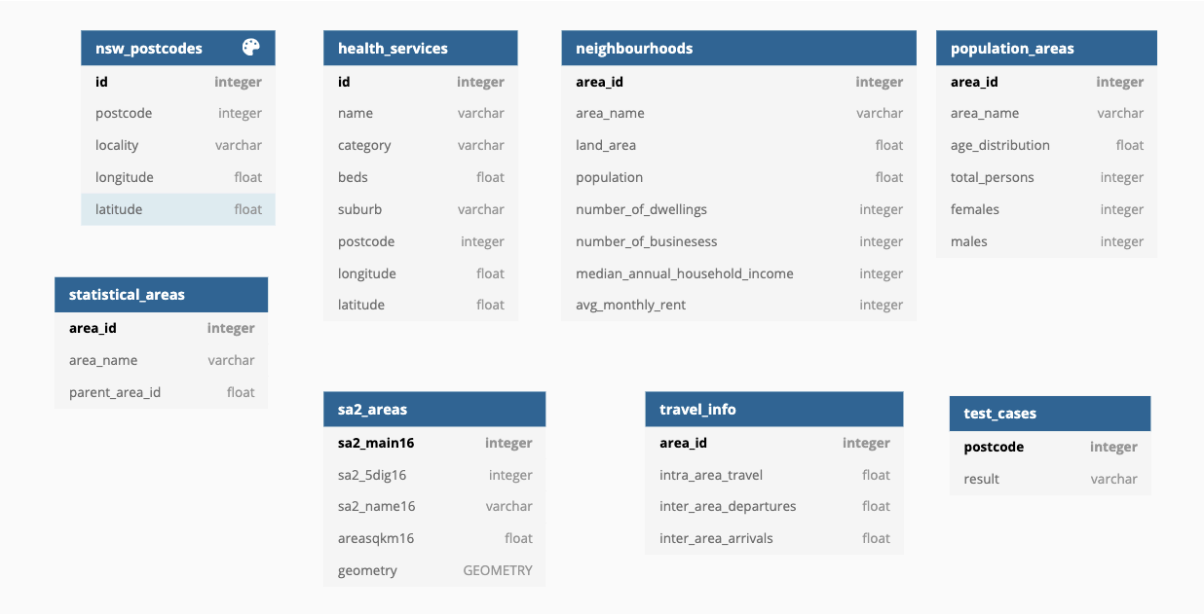
Dataset	Cleaning Process
NSW_Postcodes.csv	Remove the row from the table if longitude or latitude values are 0
PopulationStats2016.csv	Remove row if total persons are 0. Concatenate values for persons aged 70 and over into one column called "over_seventy".
HealthServices.csv	Remove irrelevant columns such as website and comment
SA2_2016_AUST.shp	Filtering for rows relevant to NSW and dropping rows where geometry values are empty/invalid
COVID-19 Tests from data.gov.au	Filtering for rows that did not contain valid postcode and results that were in not a recognisable form. Removal of irrelevant columns; leaving postcodes and results
JSON Travel Information (Provided additional dataset)	Organised data into 3 bins; inter area departures, arrivals and intra area travel.

Data Loading

Once the datasets were cleaned, they were added into PostgreSQL by connecting to the database and loading it into the corresponding schemas. Finally, connections to the database were closed. Manual inspection through PgAdmin4 was used to ensure data loading and cleaning was executed successfully.

Database Description

Schemas



Indexes

We created four indexes to improve the database performance, to retrieve specific columns used frequently in our query.

Index Name	Rationale
sa2_areas_idx	Spatial index to access sa2_areas geometries used.
lat_long_idx	Index to access latitude and longitude coordinates together of neighbourhoods in Sydney with relevant postcodes.
lat_long_2_idx	Index to access latitude and longitude coordinates together of health services located in NSW.
result_idx	Index to access COVID-19 test case results; positive or negative.

Vulnerability Score Analysis

Vulnerability Score Formula

$$vulnerability = S(z(population\_density) + z(population\_age) - z(healthservice\_density) - z(hospitalbed\_density) + z(intratravel\_area) + z(netarea\_travel))$$

Measure	Definition	Risk	Data Source
population_density	Population divided by neighbourhood's land area	+	Neighbourhoods.csv
population_age	Percentage of a neighbourhood's population aged 70 and over	+	PopulationStats2016.csv
healthservice_density	Number of health services per suburb per 1000 people	–	HealthServices.csv
hospitalbed_density	Number of hospital beds per suburb per 1000 people	–	HealthServices.csv
intra_area_travel	Number of people travelling within a given area where the origin suburb is same as	+	JSON Travel Information

	destination as proportion of total population in a given neighbourhood		
inter_area_travel	Net number of people travelling in minus travelling out of a given area proportion of total population in a given neighbourhood	+	JSON Travel Information

Formula Description

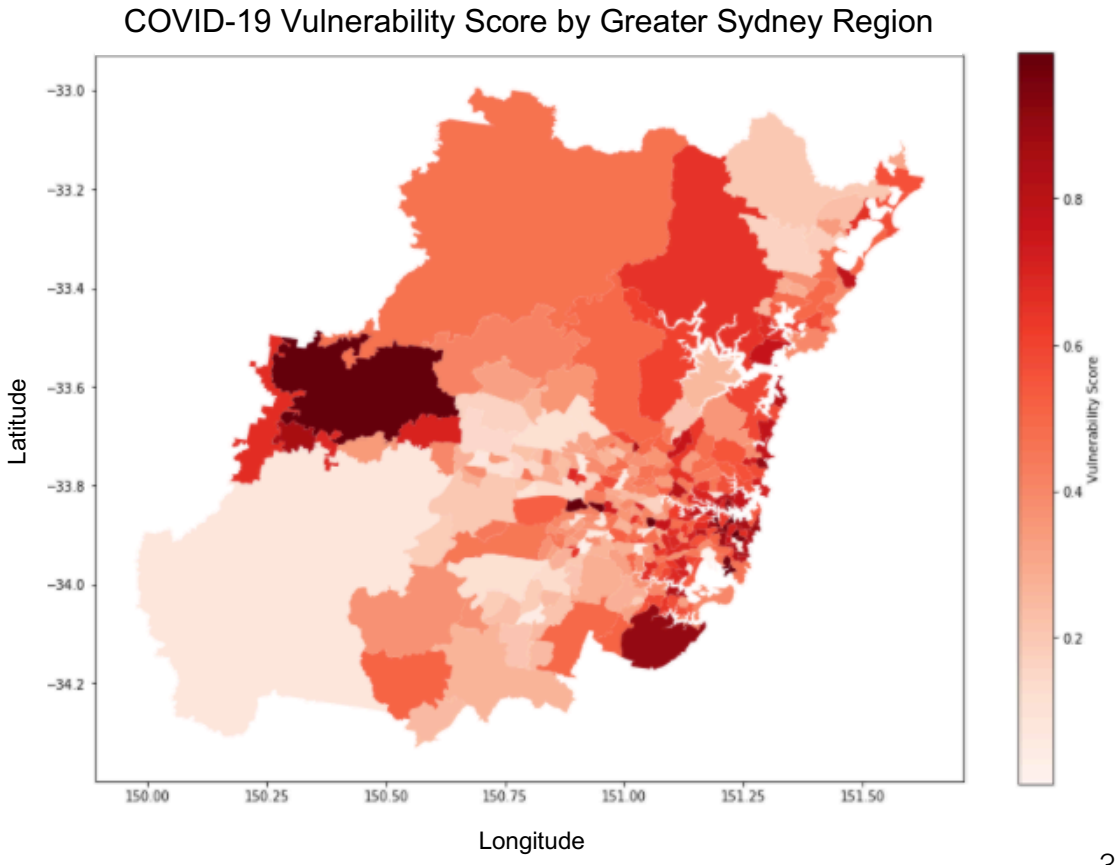
Increased population density in a neighbourhood can be associated with greater viral vulnerability due to greater transmission rates between individuals. Neighbourhoods with a higher proportion of senior citizens (aged 70 and over), are at greater risk due to reduced immunity of these elderly people. As access to health services and hospital beds within a neighbourhood increase, this reduces the vulnerability of viral infections through earlier diagnosis and greater capacity for treatment. Increased rates of intra and inter area travel increases the vulnerability of neighbourhoods due to higher possibility of community transmission from non-locally acquired cases of the virus. As such, strict travel restrictions in Australia have been implemented in conjunction with "social distancing" guidelines to minimise the spread of COVID-19 supporting our decision to include these factors in our vulnerability equation.

Vulnerability Results: Five Most and Five Least Vulnerable Areas

	sa2_name16	vulnerability_score
70	Port Botany Industrial	1.000000
88	Centennial Park	0.999970
229	Rookwood Cemetery	0.999965
287	Wetherill Park Industrial	0.999733
206	Blue Mountains - North	0.996170

	sa2_name16	vulnerability_score
71	Sydney Airport	0.000001
130	Kogarah	0.008946
291	Liverpool	0.024085
112	Bankstown - South	0.027549
102	Randwick - South	0.028969

Graphical Representation – Heatmap



Vulnerability Analysis

In the calculation of the vulnerability score, one factor taken into account was the inter-area travel within a given region. In regions with high inter-area travel, such as industrial, agricultural or tourist centres within NSW, vulnerability scores would be higher as increased population travel is related to greater COVID-19 risk. This is confirmed graphically, with the heatmap indicating both Blue Mountains – North, a tourist hotspot, and Port Botany Industrial area have high vulnerability scores.

Similarly, another factor that affects vulnerability score calculation is the number hospital beds within a given region. Particularly, industrial and remote regions within NSW often do not possess the similar health service infrastructure as more populated suburban areas. Often, workers in these regions work in relatively close proximity and in conjunction with the lack of medical infrastructure, this significantly increases the vulnerability score.

Note: athough these factors *should* theoretically affect the vulnerability score of neighbourhoods, these may not often correlate to actual case numbers.

Correlation Analysis

The coefficient of correlation (ρ) measures the relative strength between two variables, in this case vulnerability score vs. the number of positive COVID-19 test cases as a proportion of total tests conducted in a given area. A value closer to +1 indicates a stronger positive linear relationship, whereas a value closer to -1 indicates a stronger negative linear relationship.

$$\rho = \frac{cov(X, Y)}{\sigma_X, \sigma_Y}$$

Correlation Analysis Summary

Variables	Vulnerability Score vs Conducted Tests	Vulnerability Score vs Confirmed Cases	Vulnerability Score vs Confirmed Cases as a proportion of Conducted Tests
$\boldsymbol{\rho}$	-0.11149	0.00031	0.28256
Equation of regression line	$y = -7692.3x + 12122.2$	$y = 0.2x + 65.3$	$y = 0.0x + 0.0$
Correlation	*Very weak negative correlation*	*No correlation*	*Weak positive correlation*

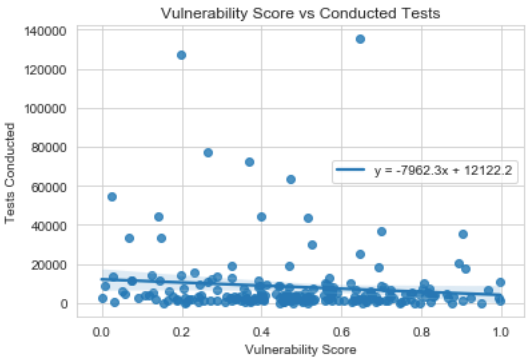
Interpretation

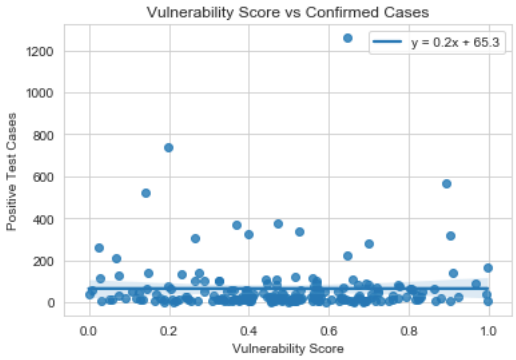
Correlation analysis reveals there is a *weak negative* correlation between vulnerability score and number of tests conducted ($\rho = -0.11149$). This suggests there is a weak negative linear relationship between vulnerability score and number of tests conducted in a neighbourhood. One example for this is due to duplicate postcode numbers associated with more than one suburb. For example, many neighbourhoods on the NSW Central Coast were associated with the postcode corresponding to Calga-Kulnura, making it an outlier. Non-distinct postcodes affected the accuracy of this correlation substantially.

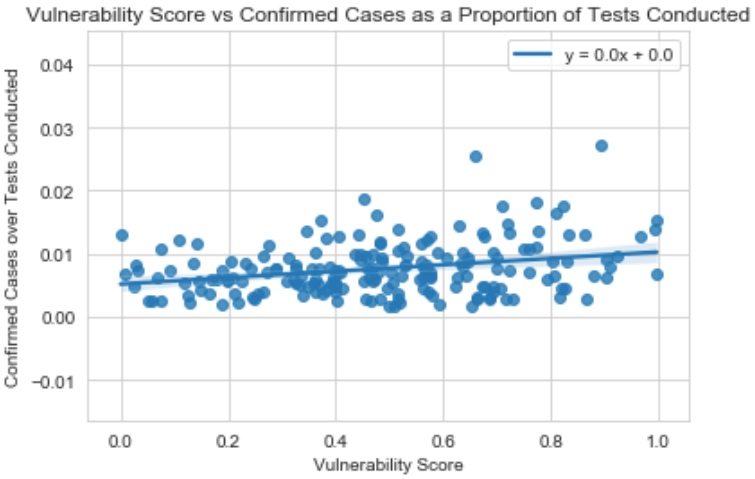
For meaningful intents and purposes, there is *no correlation* between vulnerability and number of confirmed cases of COVID-19 within a given area. ($\rho = 0.00031$). This reveals there is no linear relationship between vulnerability and confirmed cases. These results initially appear counterintuitive as the vulnerability score is a measure of several factors which we assume to affect the spread of a virus within a community.

We conducted an additional correlation analysis between vulnerability score and the number of confirmed cases as a proportion of tests conducted (in a given neighbourhood). The results of this correlation indicate a weak *positive* linear relationship ($\rho = 0.28256$).

We believe this correlation measure is more appropriate to gain an understanding of the viral vulnerability of different neighbourhoods in Sydney. By comparing vulnerability scores to the number of confirmed cases out of tests conducted, we can achieve a more standardised measure of vulnerability across neighbourhoods. For example, a rural town in remote NSW may conduct significantly less tests than a highly populated suburb in Sydney due to smaller population. Similarly, the number of confirmed cases varies with neighbourhood population size and intensity of testing. Therefore, determining the correlation between vulnerability scores and proportion of confirmed cases over tests done produces a more accurate linear relationship, assuming such relationship exists.







Limitations

Although correlation analysis can be used to identify a relationship between vulnerability score and test cases or vulnerability score and confirmed cases, it does not prove that one variable causes a change in another – causation. Similarly, we cannot determine whether increased testing rates or confirmed cases increased or decreased the viral vulnerability of a neighbourhoods. Additionally, external factors that were not considered in this viral vulnerability assessment may have influenced the outcome of correlation analysis. For example, in the context of COVID-19 the number of international arrivals into Sydney and whereabouts of these passengers would be a significant external factor to consider. Therefore, we must consider these limitations of correlation analysis when determining conclusions about the viral vulnerability of neighbourhoods.