# DATA 2001 – ASSIGMENT 01

***DOCUMENTATION OF VIRAL VULNERABILITY ANALYSIS*** ***29-05-2020***

## INTRODUCTION

The following report is a documentation of the data integration steps and the main outcomes of the vulnerability data analysis, including the correlation study with the COVID-19 statistics. The vulnerability score in this report is expressed as a measure of several factors which is assumed to affect the spread of COVID-19 within a community.
The process in the calculation of the viral vulnerability analysis could be separated into 4 sections. A brief overview of the process is listed below.

### SECTION 1 : DATASETS

For the score to be calculated data of the factors believed to affect the virus spread is needed. Datasets are sources, pre-processed then loaded onto a database for further processing.

### SECTION 2 : DATABASE

All the loaded datasets are loaded onto PostgreSQL. A database is formed pooling together data to be used in the vulnerability score analysis calculations.

### SECTION 3 : VULNERABILITY SCORE ANALYSIS

The vulnerability score is calculated using an equation and math operations.

### SECTION 4 : CORRELATION ANALYSIS

A correlation analysis is performed to analysis the correlation between score results and COVID-19 cases and tests.

## SECTION 1 : DATASETS

For the vulnerability score to be calculated. Different data sources and datasets are needed. Datasets provides input on factors assumed to affect the spread of COVID-19 within a community. These datasets are first loaded onto a database.

### 1.1 DATA SOURCES

Table 1 : Information of datasets used

| File name | Table name | File type | Data source |
|---|---|---|---|
| StatisticalAreas | statisticalareas | CSV file | Provided |
| Neighbourhood | neighbourhoods | CSV file | Provided |
| populationstats2016 | populationstats2016 | CSV file | Provided |
| HealthServices | healthservices | CSV file | Provided |
| NSW_Postcodes | nswpostcodes | CSV file | Provided |
| covid-19-tests-by-date-and-location-and-result | tests_by_location | CSV file | Australian Bureau of Statistics : https://data.nsw.gov.au/data/dataset/209cc6ff-9f28-47c8-b440-5fb52ac0b23f/resource/a675ed8c-2baf-4685-8183-48c743257f3f/download/covid-19-tests-by-date-and-postcode-local-health-district-and-local-government-area.csv |

| | | | |
|---|---|---|---|
| covid-19-cases-by-date-and-location-and-result | cases_by_location | CSV file | Australian Bureau of Statistics : https://data.nsw.gov.au/data/dataset/aefcde60-3b0c-4bc0-9af1-6fe652944ec2/resource/21304414-1ff1-4243-a5d2-f52778048b29/download/covid-19-cases-by-notification-date-and-postcode-local-health-district-and-local-government-area.csv |
| SA2_2016_AUST | sa2_loaction | shp file | Australian Bureau of Statistics : https://www.abs.gov.au/AUSSTATS/subscriber.nsf/log?openagent&1270055001_sa2_2016_aust_csv.zip&1270.0.55.001&Data%20Cubes&9F6E4EB4E23B269FCA257FED0013A4F8&0&July%202016&12.07.2016&Latest |
| Web scraped data | commute | JSON file | Australian Bureau of Statistics : http://soit-app-pro-4.ucc.usyd.edu.au:3000/api/v1/json |

Most of the data above will be used in the calculation of the viral vulnerability score. The datasets ‘tests_by_location’ and ‘cases_by_location’ are used later on in the correlation analysis.

## 1.2 PRE-PROCESSING OF PROVIDED DATASETS

5 of the 9 total datasets were provided. The CSV files of each of the dataset were loaded using pandas. Table names were then assigned for referencing and the data was loaded onto SQL for further processing.

2 of the 5 provided datasets were cleaned to enhance server performance and statistical accuracy.

The datasets ‘healthservices’ and ‘tests_by_location’ were cleaned using python before it is loaded into the SQL servers.

In the ‘healthservices’ table two columns ‘comment’ and ‘website’ were dropped as the two columns are of no use to the vulnerability analysis and takes up a considerable amount of memory space.

In the ‘tests_by_location’ table values ‘0’ and ‘9999’ were set to null.

## 1.3 POST-PROCESSING OF DATASETS

An additional ‘geom’ column is added to the dataset ‘healthservices’ using the longitude and latitude data columns in the dataset. This column is created to preform the spatial join between ‘healthservices’ and ‘neighbouhoods’ later on.

## 1.4 EXTERNAL DATASETS

4 of the 9 datasets were not provided and was sourced from the Australia Bureau of Statistics website. The links to the datasets could be found in table 1 above. The Australia Bureau of Statistics or ABS is an independent statistical agency of the Government of Australia. With the purpose of the bureau being to "inform Australia's important decisions by partnering and innovating to deliver relevant, trusted, objective data, statistics and insights". With the ABS being an independent bureau ran by the government of Australia. The data collected from it could be considered to be reliable and accurate.

Both the ‘cases_by_loaction’ and the ‘SA2_2016_AUST’ datasets were first downloaded from the websites and the uploaded onto the servers as a CSV and a shp file respectively.

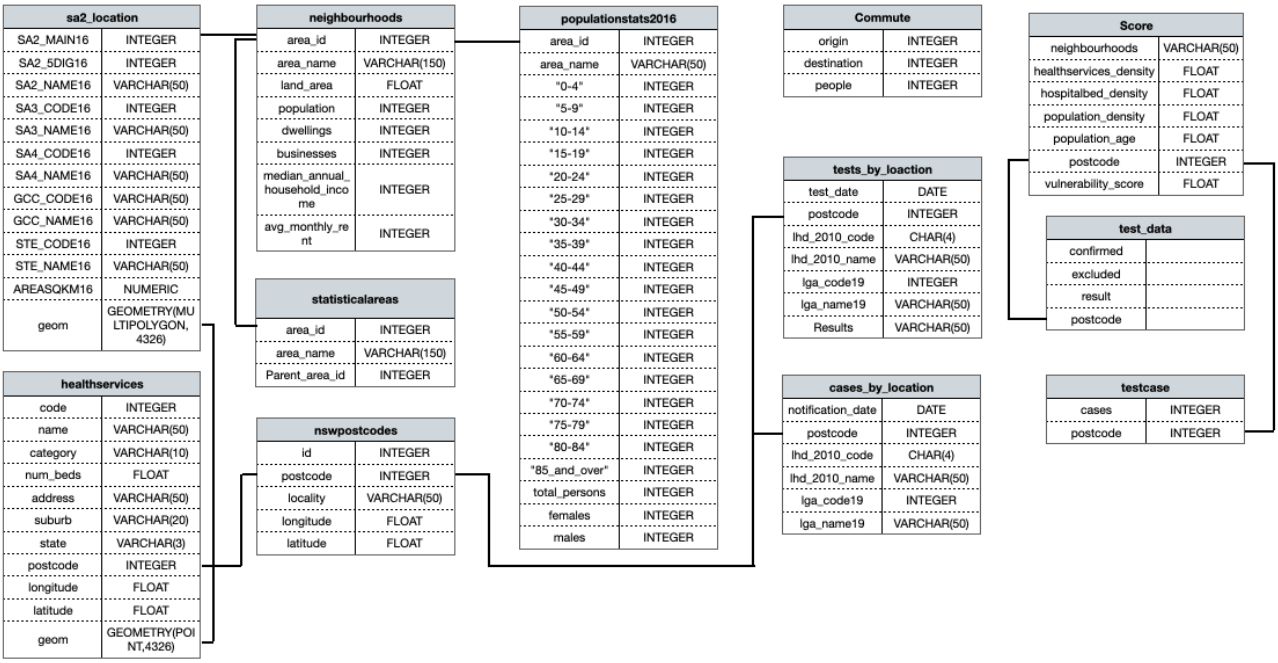
The ‘commute’ dataset, however, was uploaded through web-scraping. Instead of downloading the JSON file. The dataset was loaded using the JSON and pandas library. The dataset was then loaded onto the SQL servers. Webscraping data from the web instead of downloading the datafile and then uploading it keeps the data up to date and strengthens results of future analysis.

## SECTION 2 : DATABASE

With all 8 datasets loaded. A database is created to pool all the data to be used in the viral vulnerability score calculations.

### 2.1 DATABASE SCHEMA DIAGRAM

Figure 1 : Database schema diagram (enlarged diagram attached as appendix a)




As shown in figure 1 above. The various datasets are joined together to form a Database. Datasets are joined using common attributes and is displayed in the above diagram.

A Spatial join is used to join the ‘healthservices’ and ‘neighbourhoods’ datasets. This join could be broken down into two steps. First ‘helthservices’ is joined to the ‘sa_location’ through ‘geom’. Then ‘sa_location’ is joined to ‘neighbourhoods’ through ‘area_id’. The Spatial Reference Identifier (SRID) is 4326.

The datasets ‘tests_by_location’ and ‘cases_by_location’ of COVID-19 is joined to the database through postcodes. The detail of the join is illustrated in the figure above. These datasets will later be used in correlation analysis.

### 2.3 INDEXES CREATED

A spatial index is created from the geometry data. The spatial index is created to allow efficient access to the spatial object. The index is used to minimise processing time. Using the index means that a searching for a feature in the database would not require scanning through every record in the database.

## SECTION 3 : VULNERABILITY SCORE ANALYSIS

The vulnerability score is calculated for all given neighbourhoods using the equation listed in Section 3.1. The theory behind this analysis is the find the z-score of each factor. Calculating the number of standard deviations from the mean each measurement is for each of the neighbourhoods. It finds the relative level of risk each factor poses to each neighbourhood. With beneficial factors being decreasing the risk and vice versa. The sigmoid function is then used to predict the probability and gives a score between 0 to 1 with 0 suggesting a low probability and 1 suggesting the opposite. In other words the vulnerability score of that neighbourhood to COVID-19.

### 3.1 VULNERABILITY SCORE ANAYSIS FORMULA

$$vulnerability\ score = S\left[z(population\_density) + z(population\_age) - z\left(healthservice_{density}\right) - z\left(hospitalbed_{density}\right)\right]$$

Table 2 : Factors in vulnerability score

| Measure | Definition | Risk |
|---|---|---|
| population_density | Population divided by neighbourhood's land area | + |
| population_age | Percentage of a neighbourhood's population above age of 70 | + |
| healthservice_density | Number of health services per suburb per 1000 people | - |
| hospitalbed_density | Number of hospital beds per suburb per 1000 people | - |

The factors were calculated from the data within the database through math operators and functions.

### 3.2 OVERVIEW OF VULNERABILITY RESULTS

Table 3 : Table of caculated factors, postcode and vulnerability score (first 4 rows)

| | neighbourhoods | healthservice_density | hospitalbed_density | population_density | population_age | postcode | score |
|---|---|---|---|---|---|---|---|
| 0 | Arncliffe - Bardwell Valley | 0.008 | 0 | 49.8157 | 0.0654271 | 2205 | 0.349534 |
| 1 | Ashcroft - Busby - Miller | 0.006 | 0 | 32.8098 | 0.0836573 | 2168 | 0.422468 |
| 2 | Ashfield | 0.016 | 0 | 71.1853 | 0.100955 | 2131 | 0.387014 |
| 3 | Asquith - Mount Colah | 0.004 | 0 | 5.50872 | 0.0837259 | 2077 | 0.292128 |

Table 3 shows the results after each factor is calculated. The results of the factors are then added or subtracted together depending on the risk it represents. Applying the sigmoid function of the sum gives the vulnerability score of each neighbourhood. The final vulnerability score is shown in the score column on the far right.

As shown in table 3 above. The hospital bed density is 0 for all 4 rows. This is expected as not all neighbourhoods will have a hospital. All this means is that risk was not reduced from this factor

Health service density could be seen to reduce the risk in neighbourhood Ashfield. Despite Ashfield having a relatively high population density. It also bolsters a high health services density which balances out the risk.

The score is the highest at the second row out of the 4 rows. This could be seen as the result of having a relatively low health service density with an average to high population density and population age.

The vulnerability score above takes into account some major factors believed to be correlated with COVID-19. However, there is a lot more than what meets the eye when it comes to factors that affect the spread of COVID-19. A correlation analysis is conducted to determine the effectiveness of the vulnerability score above.

A map overlay is generated over neighbourhoods and the vulnerability score to visualise the data set.

## SECTION 4 : CORRELATION ANALYSIS

A correlation analysis between the score and COVID 19 cases and test rate for each neighbourhood could provide a valuable insight to the significance of the score in terms of predicting the vulnerability of a neighbourhood to COVID-19. The score is compared against actual data collected on COVID-19. Scatterplots would be used to provide a visualisation of the correlation between the two variables. A correlation coefficient would also be calculated through the use of the pandas function '.corr'. Map overlays are generated for tests on map, and cases on map to provide a visualisation of the COVID-19 data in a graphical setting.

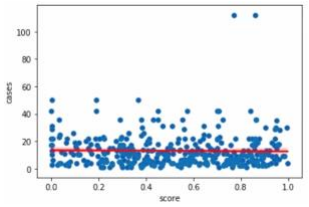
### 4.1 CORRELATION ANALYSIS BETWEEN VULNERABILITY SCORE AND COVID-19 CASES

As shown in figure 2 on the right. There is no evident correlation between COVID-19 cases and the vulnerability score.

$$\text{correlation coefficient} = -0.01886398311316438$$

A correlation coefficient ranges form -1 to 1. A coefficient extremely close to zero suggests no correlation between the vulnerability score and the number of COVID-19 cases in each of the neighbourhoods.

Figure 2 : Scatter plot of vulnerability score against COVID-19 cases for each neighbourhood



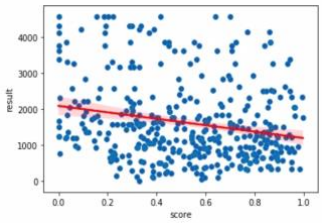

### 4.2 CORRELATION ANALYSIS BETWEEN VULNERABILITY SCORE AND COVID-19 TESTS

As shown in figure 3 on the right. This is also no evident correlation between COVID-19 testing rate and the vulnerability score.

$$\text{correlation coefficient} = -0.2175550433359228$$

Again a correlation coefficient extremely close to zero suggests no correlation between the vulnerability score and the number of COVID-19 cases in each of the neighbourhoods.

Figure 3 : Scatter plot of vulnerability score against rate of COVID-19 tests for each




*Additional correlation analysis was conducted between the score and excluded cases (tested and negative) and the score and confirmed cases (tested and positive). The coefficient score is -0.21905165387465017 and 0.11305636349486105 respectively.

## CONCLUSION

To conclude, the above report covered the dataset used, databased created, the vulnerability score, and correlation analysis. It also contains a brief walkthrough into the steps of data integration and processing. The vulnerability score result in the report shows no correlation with both the testing rate and the number of cases. The score could not be used as an approximation of the spread of COVID-19 within a community.

The score not being valid could be the result of a multitude of reasons. One of the major reasons could be because the factors taken into consideration for the vulnerability score is far too few in the grand scheme of the factors that affect the spread of COVID-19. Some other major factors may include temperature differences between neighbourhoods affecting virus activity, the level of use and or supply of surgical masks by the populations, obedience to social distancing or other health regulations by the populations, etc. In other words, increasing the number of factors and data points included in the calculation of the score will indefinitely increase the correlation between the score and actual COVID-19 statistics and increase the significance of this score in predicting the spread of COVID-19. Out of date datasets in the population stats from 2016 may also affect the result. The data may have drastically changed between 2016 and the outbreak of COVID-19 at the end of 2019. Having the most up to date data would also definitely improve the results. However, limitations such as computer processing power, reliable and available datasets, etc all limit effectiveness of the viral vulnerability score.

## REFERENCES

A list of references of sources use :


Abs.gov.au. 2020. [online] Available at: <https://www.abs.gov.au/AUSSTATS/subscriber.nsf/log?openagent&1270055001_sa2_2016_aust_csv.zip&1270.0.55.001&Data%20Cubes&9F6E4EB4E23B269FCA257FED0013A4F8&0&July%202016&12.07.2016&Latest> [Accessed 29 May 2020].

Data.nsw.gov.au. 2020. [online] Available at: <https://data.nsw.gov.au/data/dataset/aefcde60-3b0c-4bc0-9af1-6fe652944ec2/resource/21304414-1ff1-4243-a5d2-f52778048b29/download/covid-19-cases-by-notification-date-and-postcode-local-health-district-and-local-government-area.csv> [Accessed 22 May 2020].

Geographic Information Systems Stack Exchange. n.d. *Find Row With A Polygon Containing A Specific Lat/Long*. [online] Available at: <https://gis.stackexchange.com/questions/198218/find-row-with-a-polygon-containing-a-specific-lat-long> [Accessed 25 May 2020].

Postgis. n.d. *ST_Makepoint*. [online] Available at: <https://postgis.net/docs/ST_MakePoint.html> [Accessed 29 May 2020].

Postgresql. 2020. *Postgresql: Documentation: 8.0: Mathematical Functions And Operators*. [online] Available at: <https://www.postgresql.org/docs/8.0/functions-math.html> [Accessed 13 May 2020].

Postgresql. 2020. *Postgresql: Documentation: 9.1: Window Functions*. [online] Available at: <https://www.postgresql.org/docs/9.1/tutorial-window.html> [Accessed 29 May 2020].

Postgresql. 2020. *Postgresql: Documentation: 9.5: Geometric Functions And Operators*. [online] Available at: <https://www.postgresql.org/docs/9.5/functions-geometry.html> [Accessed 22 May 2020].

Postgresqltutorial. 2020. *Postgresql ADD COLUMN: Add One Or More Columns To A Table*. [online] Available at: <https://www.postgresqltutorial.com/postgresql-add-column/> [Accessed 24 May 2020].

Stack Overflow. n.d. *Postgres Error: More Than One Row Returned By A Subquery Used As An Expression*. [online] Available at: <https://stackoverflow.com/questions/21048955/postgres-error-more-than-one-row-returned-by-a-subquery-used-as-an-expression> [Accessed 21 May 2020].

# APPENDIX

## APPENDIX A : DATABASE SCHEMA DIAGRAM