

Practical Assignment: Viral Vulnerability Analysis

Course: DATA2001

Due: 8pm, Friday 29 May 2020

Tutorial F10D, Group 15

Data2001\_assignment2020s1\_group15

Student Details:

Yap Hui Xuan, Rachel; 500356946

Joshua Chew, 490408180

Table of Contents

[Dataset and Database Description 2](#_Toc41664469)

[Database Schema 2](#_Toc41664470)

[StatisticalArea 2](#_Toc41664471)

[Neighbourhoods and HeathServices data 3](#_Toc41664472)

[PopulationStats2016 3](#_Toc41664473)

[NSW\_Postcodes 4](#_Toc41664474)

[covid19\_nsw\_testsites\_simulated\_capacity 4](#_Toc41664475)

[covid\_19\_tests\_by\_date\_and\_location\_and\_result 4](#_Toc41664476)

[tests\_per\_postcode 4](#_Toc41664477)

[confirmed\_per\_postcode 4](#_Toc41664478)

[Journey\_to\_work dataset(2011JTW\_Table01\_V1.0) 4](#_Toc41664479)

[Vulnerability Score - Correlation Analysis 5](#_Toc41664480)

# Dataset and Database Description

## Database Schema

A screenshot of a social media post

Description automatically generated

## StatisticalArea

It contains a total of 3 columns:

Area\_id => unique id for an area. (primary key)

Area\_name => name of an area in Sydney.

Parent\_area\_id => area\_id is a subset of an unique parent\_area\_id.

## Neighbourhoods and HeathServices data

Neighbourhoods table consist of 9 + 1 columns:

area\_id=> unique id for a particular area. (primary key)

area\_name=> name of an area in Sydney.

land\_area=> area of a land

population=> number of people with respect to area\_id

number\_of\_dwellings => number of dwelling with respect to area\_id

number\_of\_businesses=> number of businesses with respect to area\_id

median\_annual\_household\_income=> median annual household income with respect to area\_id

avg\_monthly\_rent=> average monthly rent with respect to area\_id

population\_density (newly added) => density of the population with respect to area\_id

HealthServices table consist of 12 columns:

Id=> unique id for each hospital or General Practice (GP) clinic (primary key)

Name=> name of the hospital/GP

Category=> 2 categories: GP or Hospital

Num\_beds=> number of beds in the hospital or GP

Address=> address of the hospital/GP location

Suburb=> suburb of the hospital/GP

State=> state of the hospital/GP

Postcode=> postcode of the hospital/GP

Longitude, latitude=> longitude and latitude values of the hospital/GP

Comment=> comments on their operating hours and many other information.

Website=> website of the hospital/GP

Spatial join has been attempted using these 2 databases.

1. Create a table: *SA2\_2016\_AUST* and insert the shape file data into the table.
2. Create a spatial index on *SA2\_2016\_AUST* table. A spatial index is extremely helpful for large data sets.
3. Create a view table: *neighbourhood\_SA2\_2016\_AUST* to combine *SA2\_2016\_AUST* table and *neighbourhoods* table using *area\_id = SA2\_MAIN16*:
4. Create a table: *healthservices\_geo* and insert geometry data based on *latitude* and *longitude* in *health\_service* table
5. Create a table: *neighbourhood\_sa2\_2016\_aust\_\_healthservices\_geo* to do a spatial join between *neighbourhood\_SA2\_2016\_AUST* table and *healthservices\_geo* table using *ST\_Contains(n.goem, h.location)*

## PopulationStats2016

It contains the 23 + 2 columns:

area\_id=> unique id for a particular area. (primary key)

Area\_name => name of an area in Sydney.

Age\_distribution=> number of people in the respective age range. (18 columns)

Total\_persons=> total number of people with respect to area\_id.

Females=> number of females with respect to area\_id.

Males=> number of males with respect to area\_id.

Above\_70 (newly added) => number of people having an age >= 70 with respect to area\_id.

Percent\_age\_above\_70 (newly added) => percentage of people who are age above 70 with respect to area\_id.

## NSW\_Postcodes

It contains the 5 columns:

id => unique id for an area In NSW. (primary key)

postcode => represents a NSW postal address linked to the area’s id.   
locality => represents the name of an area in NSW where the id and postcode are linked to.

Longitude, latitude => represents longitude and latitude values of the postcode id

## covid19\_nsw\_testsites\_simulated\_capacity

It contains 7 columns:

Site\_id=>unique id for a COVID-19 test site. (primary key)

centre\_name => name of the test site.

Phone\_number => phone/mobile number attached to a test\_site.

Opening\_hours => the length of time the test sites remain open.

Latitude, longitude => represents longitude and latitude values of the test\_site.

Test\_capacity => the estimated capacity of each test site.

## covid\_19\_tests\_by\_date\_and\_location\_and\_result

As the dataset was over 25MB (32,000+ rows) we removed several columns that were irrelevant to the vulnerability score and correlation analysis:

For example: 1hd\_2010\_code and 1hd\_2010\_name => correspond to SA3 regions of Sydney

It contains 2 columns for 2 separate datasets (split):

## tests\_per\_postcode

Postcode=> represents a unique Australian postal address. (primary key)

Result=> represents the number of tests conducted at their respective postal address.

## confirmed\_per\_postcode

Postcode=> represents a unique Australian postal address. (primary key)

Result=> represents the number of positive COVID-19 tests found at their respective postal addresses.

## Journey\_to\_work dataset(2011JTW\_Table01\_V1.0)

This dataset details the number of employed people in each region.

It has just 3 columns from an original 27: As all our datasets we relevant to areas in SA2 we removed all other areas for more streamlined imports (25+ MB dataset).

O\_SA2\_11=>Area\_id=> corresponds to a unique area in NSW. (primary key)

O\_SA2\_NAME11=>Area\_name=> is the area corresponds to the area\_id.

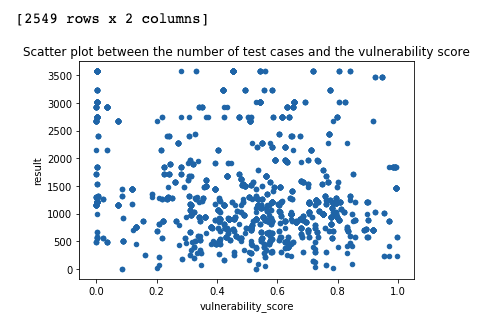
EMPLOYED\_PERSONS =>Employed\_persons=> the aggregate of all people employed in SA2. (so corresponding to Area\_id and Area\_name)

# Vulnerability Score - Correlation Analysis

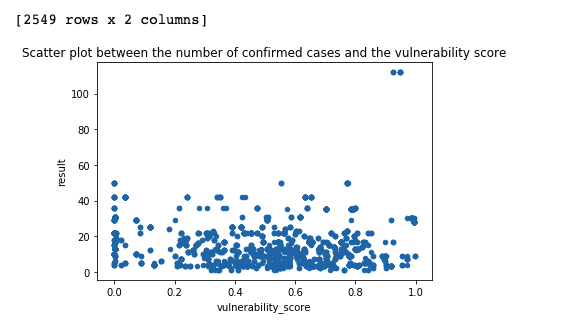
We found a weak negative correlation between both the number of testcases and the vulnerability score and the number of confirmed cases and the vulnerability score. This could be due to the following reasons:

* The datasets age may have influenced the results to an extent. The population dataset is dated to 2016 and the travel dataset (journey to work) is dated to 2011. Combined with highly recent covid\_19\_test\_by\_date\_and\_location dataset, which is dated to early May of 2020 and we can expect unreliable results.
* The significant number of null values in the HealthServices dataset may also make the results unrepresentative.
* On average we would expect risk scores to be higher than the number of confirmed cases.

Thus, there is reasonable doubt to infer that the correlation results are invalid.



r= -0.19611988707687444



r= -0.05900450728362349