Introduction

Australia consists of over 2000 distinct geographical regions known as “Statistical Area Level 2”(SA2), which are specifically defined to represent communities ranging from 3000 to 2500 individuals who interact socially and economically. For this particular report, we will focus on the 350+ SA2s located within the Greater Sydney area. This report will include a computation of the scores that reflect the level of available resources for each region along with the data description, and step by step integration, correlation analysis, and visualisation.

Dataset and preprocessing description

The general preprocessing method followed for all datasets is as follows:

- Load data into dataframe using pandas or geopandas when geospatial data is concerned
- Remove columns not necessary for the purposes of this report
- Rename and change the order of columns
- Remove rows with ‘na’, ‘NaN’, ‘0’, or NULL values
- Convert column names to lower case to simplify use in sql queries
- For data frames with spatial data:
    - SRIDs were converted to a consistent type (4326)
    - Longitude and latitude columns were used to create Point data (where applicable)

This method was followed for all datasets unless otherwise specified.

Provided Datasets

Businesses.csv

The business.csv dataset is a dataset related to industries and businesses in various regions sourced from The Australian Bureau of Statistics. After removing unnecessary columns the businesses data frame contained variables industry_name, sa2_code, sa2_names, and total_businesses. During preprocessing we filtered the data frame so that it only contained data relating to the industries ‘Retail Trade’ and ‘Health Care and Social Assistance’.

Stops.txt

This spatial dataset contains information about public transportation stops and was sourced from Transport for NSW. After removing unnecessary columns the stops data frame contained variables columns stop_id, stop_name, and the_geom.

PollingPlaces2019.csv

This spatial dataset gives information about polling places in NSW as sourced from The Australian Electoral Commission. After removing unnecessary columns the polling data frame contained variables polling_place_id, polling_place_name, and the_geom. During preprocessing we removed all rows containing NA values.

Catchments.zip

The catchments.zip contains several datasets containing school catchment information, as sourced from The NSW Department of Education. For the purposes of this report we only made use of the catchments_primary.shp and catchments_secondary.shp files. We concatenated the two data frames containing these two files to form the school dataframe. After removing unnecessary columns the school data frame contained variables use_id, use_desc, and geometry.

Population.csv

This dataset provides information about the total people in each sa2 region and gives a breakdown of the age groups in these regions as sourced from The Australian Bureau of Statistics. During preprocessing we created a new variable called young_people which contains the sum of all data from the 0 to 19 year old range. After removing unnecessary columns the population data frame contained variables sa2_code, sa2_name, young_people, and total_people.

Income.csv

The income dataset provides information about the number of earners, median_age, median_income, and mean_income for each sa2 neighbourhood as sourced from The Australian Bureau of Statistics (table 1). After removing unnecessary columns the income data frame contained variables sa2_code, sa2_name, and median_income. During preprocessing we removed all rows with np values.

SA2_2021_AUST_GDA2020.shp

The SA2 dataset contains the SA2 digital boundaries of neighbourhoods in Australia as sourced from The Australian Bureau of Statistics. After removing unnecessary columns the sa2 data frame contained variables sa2_code21, sa2_name21, and geometry. During preprocessing we filtered the data frame so that it only contained data relating to neighbourhoods in Greater Sydney.

Additional Datasets:

Toiletmap.csv

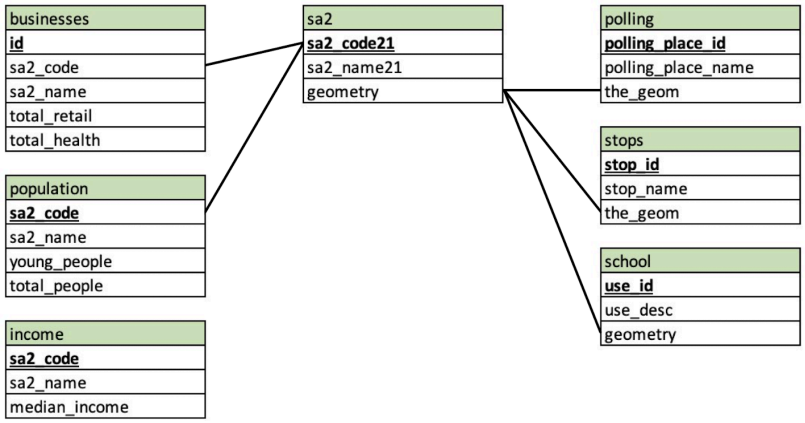
The public toilet dataset contains the location and facilities of public toilets within Australia as sourced from The Australian Government Department of Health. After removing unnecessary columns the toilet data frame contained variables facilityid and the_geom. During preprocessing we filtered the data frame so that it only contained data relating to NSW.

Rent.csv

The rent dataset contains rent information across Australia detailing dwelling type and weekly rent, as sourced from The Australian Bureau of Statistics. After removing unnecessary column the rent data frame contained variables sa2_code, sa2_name, obs_value, and rent. During preprocessing the data was filtered such that only data pertaining to SA2 regions within NSW remained, we then used the upper and lower bound of rent ranges to compute average weekly rent. The column containing region information used to form the sa2_code and sa2_name columns by splitting the data in the ASGS_2016: Region column.

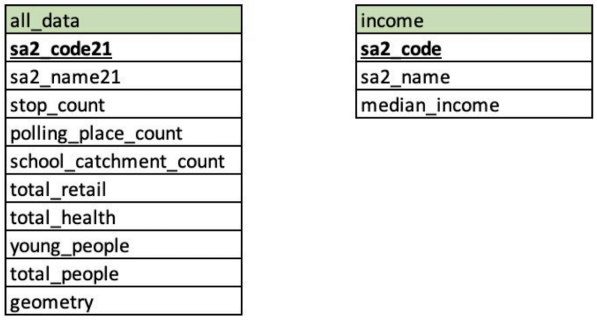
Database Description

Once SQL was used to create tables for each data frame, these tables were populated with the data from their respective dataframes. Primary keys were created for every table and the businesses table was altered such that instead of having one column for industry name, which contains either ‘Retail Trade’ or ‘Health Care and Social Assistance” , it now has two columns ‘total_retail’ and ‘total_health’ which store the total businesses for their industry. Below is a database schema diagram before the tables were merged:



Two views, sa2_spatial and sa2_nonspatial, were created and populated using joins on the existing tables. These two views are the merged versions of all the above tables, with spatial and non-spatial data divided accordingly. We chose to form these two views before creating a final table containing all data in order to simplify the joins used.

The table containing all data relevant for calculating scores was created by using joins on the previously mentioned view. Below is a database schema diagram after all data for the scores has been merged (the tables from the diagram above are still in the database but are not included in this diagram to reduce redundancy):



Once the additional datasets were ingested into the DBMS the database schema diagram could be extended as below (the tables from the previous diagrams remain in the database but are not included to reduce redundancy):

toilet
facilityid
the_geom

rent
sa2_code
sa2_name
obs_value
rent

Note: Primary keys are underlined and written in bold text.

Score Analysis

Score Computation

The well-resourced score of each SA2 region in the Greater Sydney area was computed according to the following formula:

$$Score = S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools})$$

where S is the sigmoid function and z is the normalised z-score. The score was only calculated for SA2 regions with a population of at least 100 people. These z scores are the z scores of the number of retail businesses, health services, public transport stops, federal election polling places, and school catchment areas per region.

The score was computed by first calculating the mean and standard deviation of each of the aforementioned criteria, this information was then used to calculate the z score according to the following formula:

$$Z = (x - \mu)/\sigma$$

where x is the observed value, μ is the mean of the sample, and σ is the standard deviation of the sample. In the definitions as given in the assignment requirements we were asked to determine the z score for retail, health, and school per 1000 people (or young people for school) however upon computing the score this lead to means and standard deviations of zero which made computing the score impossible, for this reason we chose not to incorporate this per capita criteria.

Once the z-scores of every region were calculated we normalised these z scores using min-max scaling. We first calculated the minimum and maximum z score for each variable and then used these values to normalise the z scores according to the following formula:

$$Z_{normalised} = (Z_{not\ normalised} - Z_{min})/(Z_{max} - Z_{min})$$

We then finally computed the well-resourced scores according to the following formula:

$$S = 1/(1 + e^{-(z_{retail}+z_{health}+z_{stops}+z_{polls}+z_{schools})})$$

where the normalised z-scores were used.

Note: Once score was extended to include the additional datasets the same procedure was followed.

Score analysis for original score computation

The key results of the well-resourced score computation to 3 decimal places are as follows:

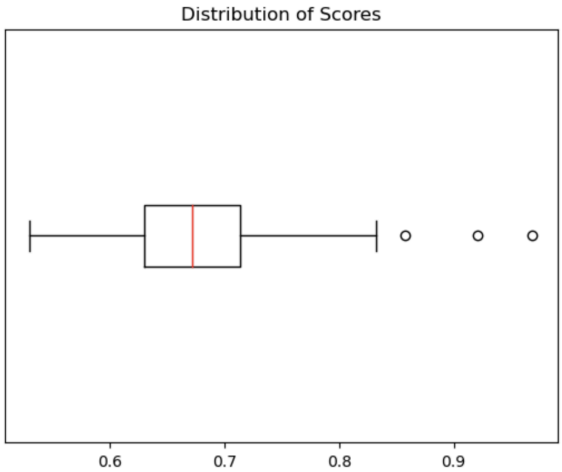
Minimum	0.531
Maximum	0.969
Mean	0.675
Median	0.672
Standard Deviation	0.064

The highest and lowest 3 scores to 3 decimal places and their respective regions are as follows:

Highest	
SA2 region	Score
Sydney (North) - Millers Point	0.969
Dural - Kenthurst - Wisemans Ferry	0.921
Baulkham Hills (West) - Bella Vista	0.857

Lowest	
SA2 region	Score
Lilli Pilli - Port Hacking - Dolans Bay	0.531
Summerland Point - Gwandalan	0.531
Banksmeadow	0.539

The scores are relatively normally distributed but do have several outliers, as can be seen in the following box plot:



These outliers are higher than the upper boundary of the box plot. This could indicate that it would be wise to remove these outliers and determine the scores and visualisations once more, this would likely lead to a more positive result of the score computation.

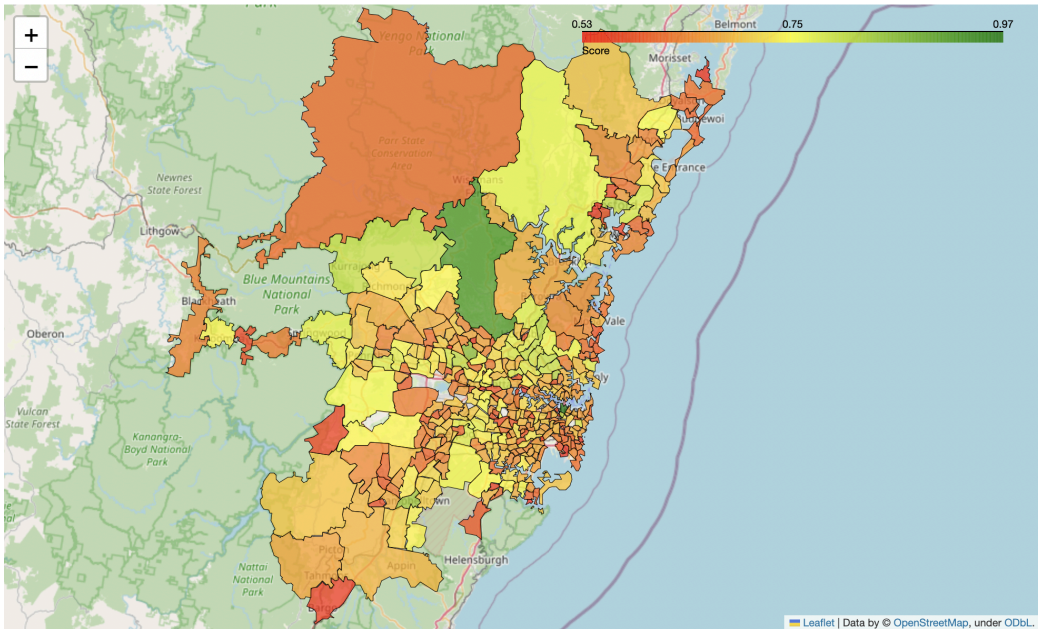
Extended score

Before the extended score could be computed to incorporate the rent and public toilet datasets it was necessary to create a view for each dataset which, by using joins with the all_data table, contains the number of public toilets per region, and the average rent per region. The extended scores were then calculated in the same way as the original score except that the normalised z score for rent was deducted instead of added as higher rent would make a region less desirable to individuals wanting to live in the region and therefore its score should be lower.

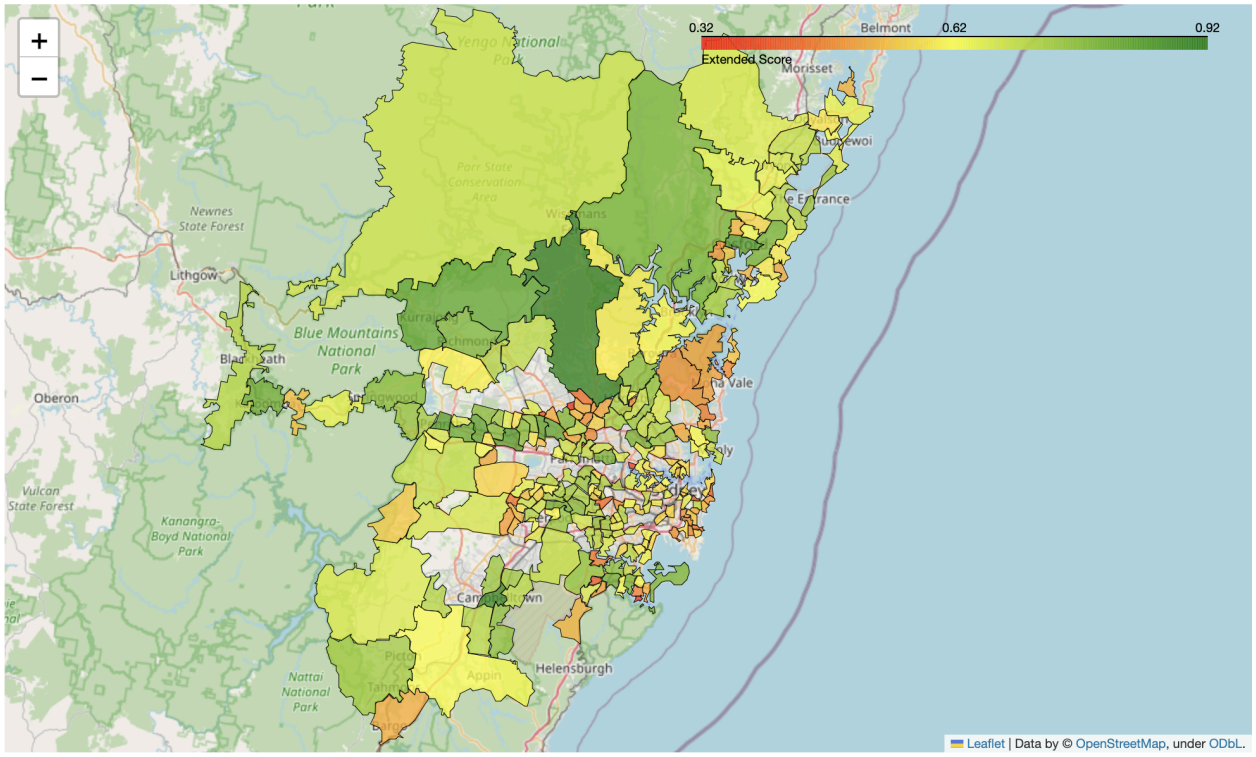
Extending the score to the additional datasets can be useful as the score now offers a more complex analysis of how well-resourced a region is, however extending the score also has a negative impact on the usefulness of our analysis. This negative impact is due to the fact that the additional datasets have missing data for certain regions and therefore the score is a less complete picture of the regions within the Greater Sydney area. Another limitation to extending the score is the way in which the normalised z score of rent is incorporated into the score, the decision made to deduct this z score in the sigmoid function was made considering the data from the perspective of an individual living in a given region, however if the perspective of a property owner was to be taken the score would have to be managed differently.

Visualisation

An interactive map was generated using the well-resourced scores as calculated above. Below is a static image of the map, the interactive map can be found here.



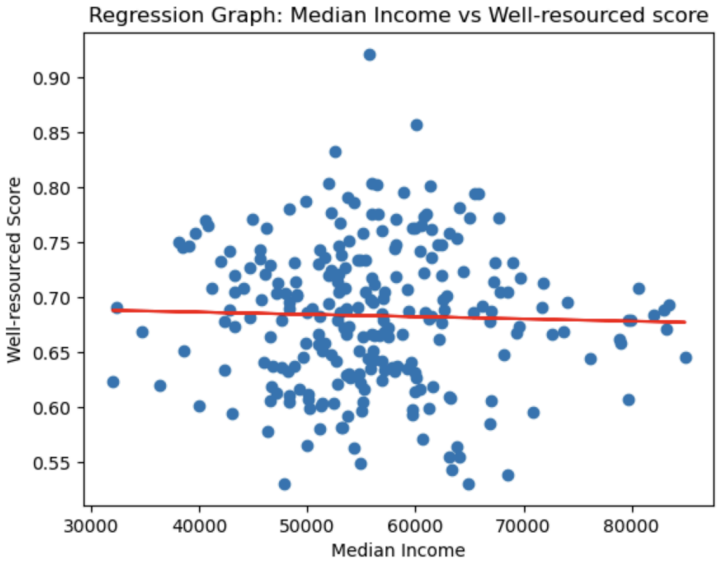
An interactive map was generated using the extended well-resourced scores as calculated above. Below is a static image of the map, the interactive map can be found here.



As you can see there are regions for which the additional datasets had no data, therefore no score was calculated for these regions. It is also important to note that the values assigned to each end of the colour spectrum in the extended score map differ from those in the score map, this can cause the data to be misconstrued if the difference of scale is not taken into consideration.

Correlation Analysis

To determine the correlation between the well-resourced score and the median income of a region we created a table income_score using SQL. We then used pandas to make a data frame for this table and used this data frame to determine any correlation. Below is a graph fitted with a regression line showing the correlation between income and score. Correlation analysis between the well-resourced scores and median income revealed a weak negative correlation of -0.03045277822359784, suggesting that the well-resourced scores are not strongly influenced by the median income of an area.



Usefulness, Limitations and Conclusion

One of the biggest limitations of the computed scores is the fact that the datasets used were not all current or from the same time period. For example the rent data was from 2016 whereas the polling data was from 2019. This likely leads to some inaccuracy in the scores. The scores are also not fully representative of how well-resourced an area is, an area may have many health services within it but the value of these health services lies in their quality. This aspect of the data is not considered in the computation of the well-resourced scores.

Despite the limitations of the scores, these scores could be useful to individuals or enterprises deciding on an area to locate themselves in. These scores could also assist government officials in determining which regions to focus their efforts on in order to improve the Greater Sydney area as a whole.

In conclusion, the analysis of Greater Sydney using the provided datasets and additional datasets has provided valuable insights into the well-resourced scores of different SA2 regions. The scores, although limited in certain aspects, can be useful for individuals, enterprises, and government officials in their decision-making processes.

Overall, this analysis provides a starting point for further exploration and decision-making, acknowledging the strengths and limitations of the scores and the datasets used. Future work could involve incorporating more current and comprehensive datasets, considering qualitative factors of resources, and refining the scoring methodology for a more accurate assessment of well-resourced areas within Greater Sydney.