SID 480062228
SID 510141497
SID 510524520

Dataset Description

The data used originated from various sources: businesses.csv, SA2_2021_AUST_GDA2020.shp, stops.txt, pollingplaces2019.csv, catchments_future.shp, catchments_primary.shp, catchments_secondary.shp, population.csv, and income.csv.

The initial cleanup was done on the population dataset. This involved converting column names that begin with a number into string format because postgresql only accepts column names that start with letters or an underscore. Subsequently, NULL values in the geometry columns from the sa_region, stops, polls, and schools datasets were addressed.

Furthermore, the longitude and latitude columns in the stops and polls datasets were transformed into geometry points. Finally, all polygon types in the sa_regions, stops, polls, and schools datasets were converted to multipolygon format.

Database Description

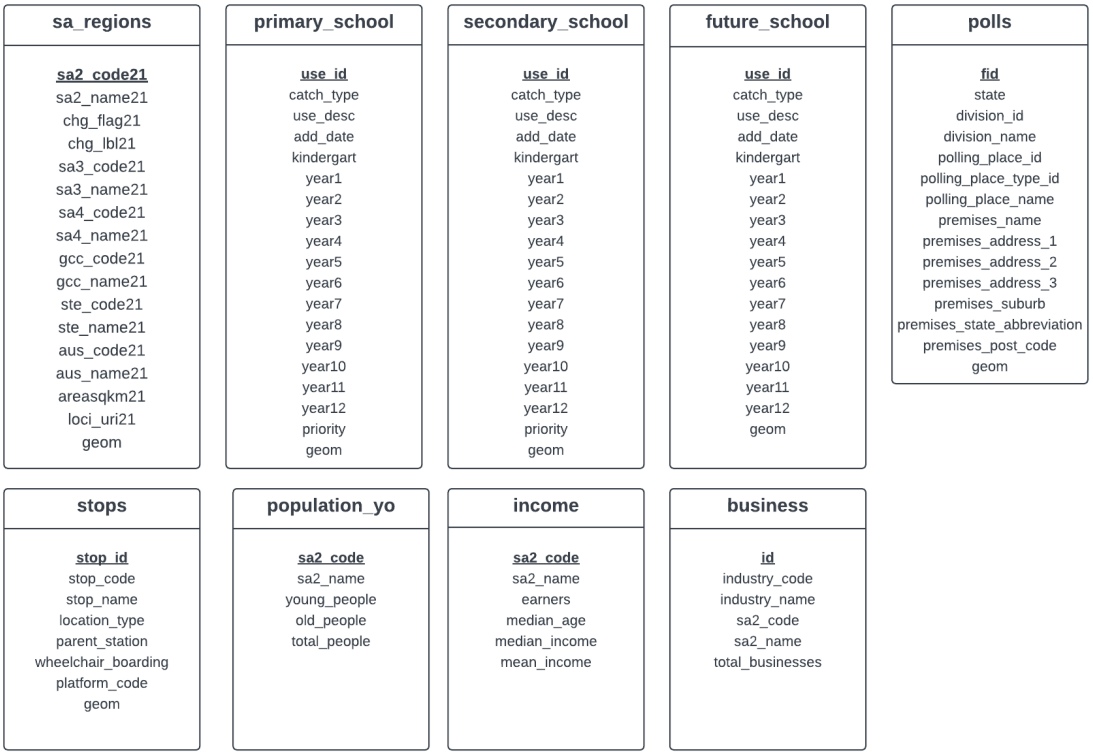

Figure 1. Database diagram

When we first cleaned the datasets, there were some rows with null value on the column we used to join. Those data were taken out as they could not be used to compare with other tables.

SID 480062228
SID 510141497
SID 510524520

sa2_code21 is initially used to combine all the original tables with sa2_code present. Those without sa2_code are joined by comparing their coordinates in the geom column to see if the region contains the respective facilities.

Index on id in the business table is created so that it will reduce the time taken to find appropriate entries as some of them have overlapping industry_code for sa2_code.

For the spatial index, the geom column in the sa_regions table is used as that is the row used to see if the given resources are present in the particular sa2 region.

Score Analysis

Our resource scoring model is based on the sigmoid function that was derived from various sources. The score generated provides a quantified measure of the living quality in separated regions around the Greater Sydney area.

Model and formula
The resource score is computed using the following formula:

- Score$=S(Z_{retail} + Z_{health} + Z_{stops} + Z_{polls} + Z_{schools})$
- Z represents the scoring function of each variable which is computed by the formula below.

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

Figure 2. Z-score Formula

To ensure the robustness and fairness of our model, we extended our dataset with additional data from various sources:

- Heat Vulnerability Index: This metric encompasses areas to be observed, indicating where populations within the Greater Metropolitan Area of Sydney are more susceptible to the detrimental impacts of urban heat as of the summer season of 2015-2016. The HVI employs exposure, sensitivity, and adaptive capacity indicators to determine an overall heat vulnerability score.
- Traffic Light Count: This dataset includes the locations of traffic lights in New South Wales. By tallying the number of traffic lights in each area, we can assess the level of resources in a given neighborhood.
- Hospital Accessibility: In adherence to privacy laws and regulations, certain private hospitals have been excluded from the total hospital count per region. This measure assists in evaluating a community's health and wellbeing stability.

Results Overview

SID 480062228
SID 510141497
SID 510524520

Since we added 3 more datasets in addition to the 5 datasets provided, our score formula changed to

- Score=S($Z_{retail} + Z_{health} + Z_{stops} + Z_{polls} + Z_{schools} + Z_{hvi} + Z_{trafficlights} + Z_{hospitals}$).

Also, the added datasets represent useful resources and hence, the z-scores are added instead of getting subtracted.

	sa2_name21	retail_zscore	health_zscore	stops_zscore	polls_zscore	schools_zscore	hvi_zscore	trafficlights_zscore	hospital_zscore	zscore_sum	normalised_sum	sigmoid
0	Acacia Gardens	-0.234085	-0.258632	-1.284334	-0.801427	-0.131607	0.727700	-0.937846	NaN	-2.920232	-0.764375	0.317697
1	Annandale (NSW)	0.150135	0.698696	-1.272737	-0.072920	-0.128648	NaN	0.024466	NaN	-0.601008	-0.146205	0.463514
2	Arncliffe - Bardwell Valley	-0.000833	-0.535537	-0.750886	0.169916	-0.130615	NaN	0.385333	NaN	-0.862622	-0.215936	0.446225
3	Artarmon	-0.332335	0.174957	-1.330720	-0.558591	-0.129788	NaN	0.265044	NaN	-1.911433	-0.495488	0.378602
4	Ashcroft - Busby - Miller	-0.467989	-0.619453	0.779877	-0.072920	-0.132458	1.513119	-1.058135	NaN	-0.057958	-0.001460	0.499635
...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3. Table of z-scores, sums and sigmoid value

After adding up the z-scores, we normalized it so that it represents more accurate and standardized information. Hence,

- Score=S(normalized sum of Z-scores).

Among the regions, there are regions without some facilities. In z-score calculation, they were taken out in the z-score sum. However, this does not reduce the correctness as the sum is calculated by adding normalized z-scores of resources and adding 0 just represents the absence of such resources in that particular region.

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^{x}}{e^{x} + 1}$$

Figure 4. Sigmoid Function Formula

After finding the normalized sum, the sigmoid function is implemented where x is the normalized sum. It is implemented for a normalized sum of every region so that we will be able to compare the result between every region.

SID 480062228
SID 510141497
SID 510524520

Version with sa2_name21, normalised_sum and sigmoid:

	sa2_name21	normalised_sum	sigmoid
0	Acacia Gardens	-0.764375	0.317697
1	Annandale (NSW)	-0.146205	0.463514
2	Arncliffe - Bardwell Valley	-0.215936	0.446225
3	Artarmon	-0.495488	0.378602
4	Ashcroft - Busby - Miller	-0.001460	0.499635
...	...	...	...
355	Wyoming	-0.445140	0.390517
356	Wyong	-0.262069	0.434855
357	Yagoona - Birrong	0.252755	0.562855
358	Yarramundi - Londonderry	-1.082000	0.253128
359	Zetland	-0.603467	0.353551

360 rows × 3 columns

Figure 5. Simplified Table

The value of sigmoid will represent how resourced the area is in relation to other regions in Greater Sydney.

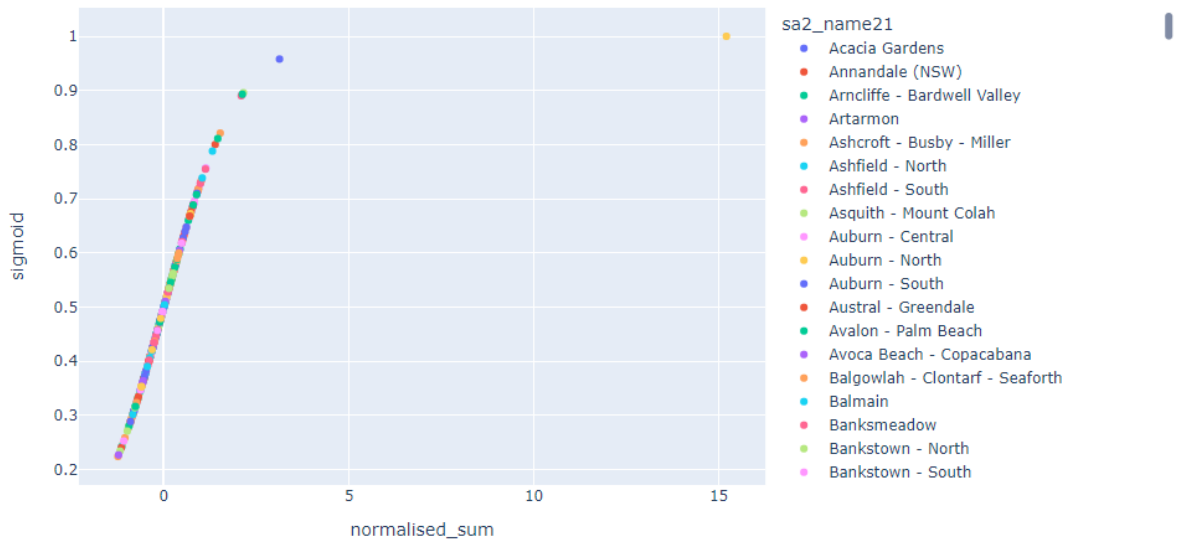

Figure 6. Sigmoid against normalised_sum graph

This graph shows the sigmoid graph created. As it can be seen, there was an anomaly at normalised_sum = 15.2 which shows that there was an area which had much more resources than other areas and it is "Sydney (North) - Millers Point". Beside that area, other areas are resourced pretty evenly with normalised_sum being in range from -2 to 4.

SID 480062228
SID 510141497
SID 510524520

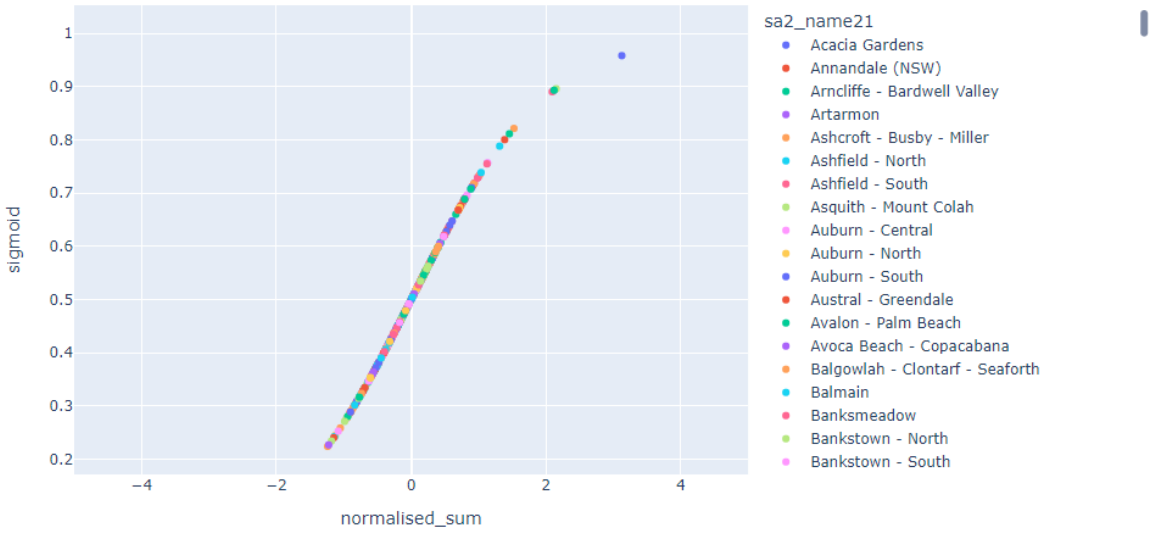

Figure 7. Cleaned Sigmoid against normalised_sum graph

Without the anomaly point at normalised_sum = 17, the graph only looking at normalised_sum between -4 and 4 looks like this.

Correlation Analysis

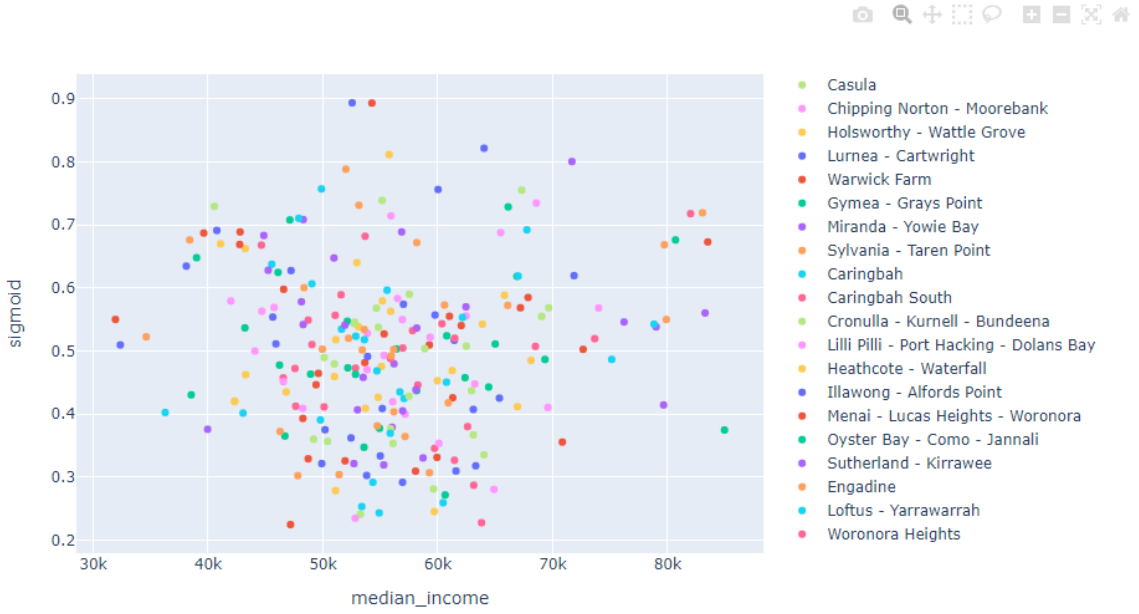

Figure 8. Scatter plot of sigmoid value against median_income

The graph generated does not depict any clear pattern correlating increased income with residing in areas rich in resources. This result is quite surprising, as it is generally assumed that those with higher earnings tend to inhabit regions with more plentiful resources. Therefore, the computed value from the sigmoid function does not serve as valid evidence to

support the theory that a rise in median income increases the likelihood of living in better-resourced areas.

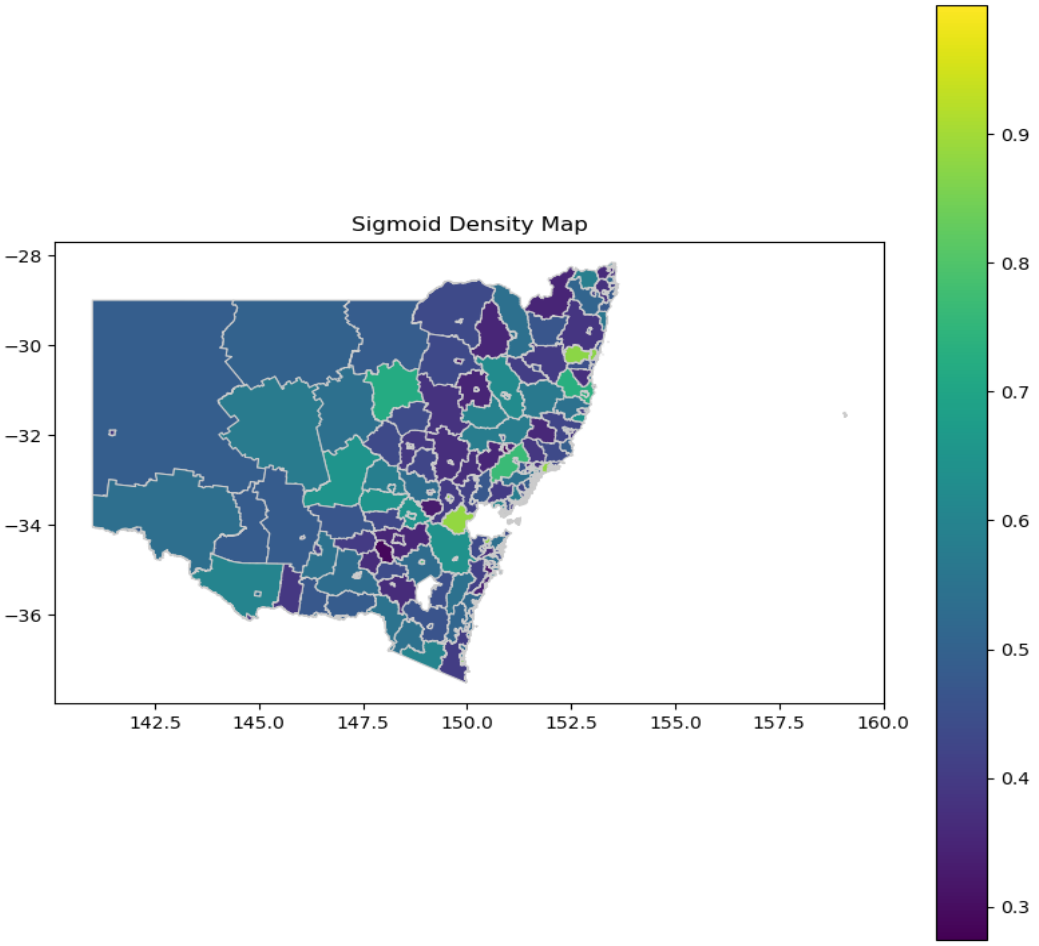

Figure 9: Map Overlay Visualization(Sigmoid)

The depicted map demonstrates the regions utilized in the computation of the sigmoid function, with darker colors signifying lower values and brighter colors indicating higher ones. This map specifically focuses on the Greater Sydney area, which is our primary area of interest. However, the map's limitations lie in its incomplete coverage of all regions, mainly due to the presence of numerous null values within the various datasets we've gathered.