DATA2001 Report

520495548 - 510588360- 520301700
CC13 - Group 8

I. Dataset Description:

The data sources provided are detailed as below:

SA2 Regions: This dataset is obtained from the Australian Bureau of Statistics. We read and processed this data using GeoPandas, dropping all unnecessary columns before loading it into our database. This dataset includes all Statistical Area Level 2 (SA2) digital boundaries, which has been filtered down to just Greater Sydney. This was done by filtering out rows from the shapefile where the *GCC_NAME21* column was not Greater Sydney. The geometric data was converted to multipolygon values.

Businesses: This dataset is obtained from the Australian Bureau of Statistics detailing the number of businesses by industry and SA2 region, reported by turnover size ranges. The dataset did not require cleaning and the CSV file was loaded into the SQL server.

Stops: This geometric dataset is obtained through Transport for NSW. This includes the locations of all public transport stops (train and bus) in General Transit Feed Specification (GTFS) format. Stops was provided as a .txt file, which was converted to a CSV file in Python. Before loading into our database, the latitude and longitude columns, *stop_lat* and *stop_lon* respectively, were combined into one column called *geom*. Coordinate values were converted to PostGIS friendly geom point values.

Polls: This dataset is obtained from the Australian Electoral Commision featuring locations (and other premises details) of polling places for the 2019 Federal election. The CSV file contained coordinate values, which had the *longitude* and *latitude* columns that had been combined into one column *geom* and converted to geom point values.

Schools: This dataset is downloaded from the NSW Department of Education. These sets include geographical regions in which students must live to attend primary, secondary and future Government schools. This dataset stored one shapefile each for primary, secondary, and future school catchments. The geometric data was converted into multipolygon values.

Population: This data estimates the number of people living in each SA2 by age range. The CSV file did not require cleaning and we have loaded it into the database directly.

Income: This dataset features total earnings statistics by SA2 (for later correlation analysis). We cleaned the CSV file by filtering out rows where attributes were given ‘np’ values.

<u>The data sources found by group members include:</u>

Emergency Services: This dataset is obtained from NSW Spatial Collaboration Portal. This dataset features spatial data representing emergency services across NSW including NSW Rural Fire Stations, Fire and Rescue NSW, NSW Police Stations and SES facilities. At first, this data set has more than 20 columns which are not related to our goal. Therefore, this dataset is cleaned to include only necessary information such as the features *generalname* which identifies the name of each emergency service and coordinate point from the *geom* column which locates the station on map before loading into the database. The *geom* column is achieved after we extract the longitude and latitude from a 3D point and convert it into a 2D point.

Source:https://portal.spatial.nsw.gov.au/portal/home/item.html?id=db511dc3fde34759997088c5cdbcf29b

Transport for NSW SHP on all the train stations in NSW: This dataset details features such as *station_id*, *station_name* and many more with the respective geometry point. For processing the data, I converted the geometry to WKT format and dropped all columns not needed, leaving only *station_id*, *station_name* and the new *geom*.

Source:https://roads-waterways.transport.nsw.gov.au/about/corporate-publications/statistics/traffic-volumes/aadt-map/index.html#/?z=6&st=2

Greater Sydney Urban Development Program: This csv dataset monitors the number of completed residential housing projects in Greater Sydney to the latest financial year. Each project entry contains a DateID, FinancialYear, LGA (Local Government Area), City, Greenfield growth data (if the area was previously undeveloped), DwellingType, NetCompletion (number of completed projects), and coordinate values. The *lga*, *city*, *netcompletion*, and the *coordinate* columns were kept.The coordinate data was turned into geometric points, converted into WKT format, and loaded into PostGIS. A sum of finished projects was maintained for each sa2 region, which was scaled by the respective area (per square km). The latter was used in the z_score calculation.

Source: https://datasets.seed.nsw.gov.au/dataset/greater-sydney-urban-development-program

II. Database Description:

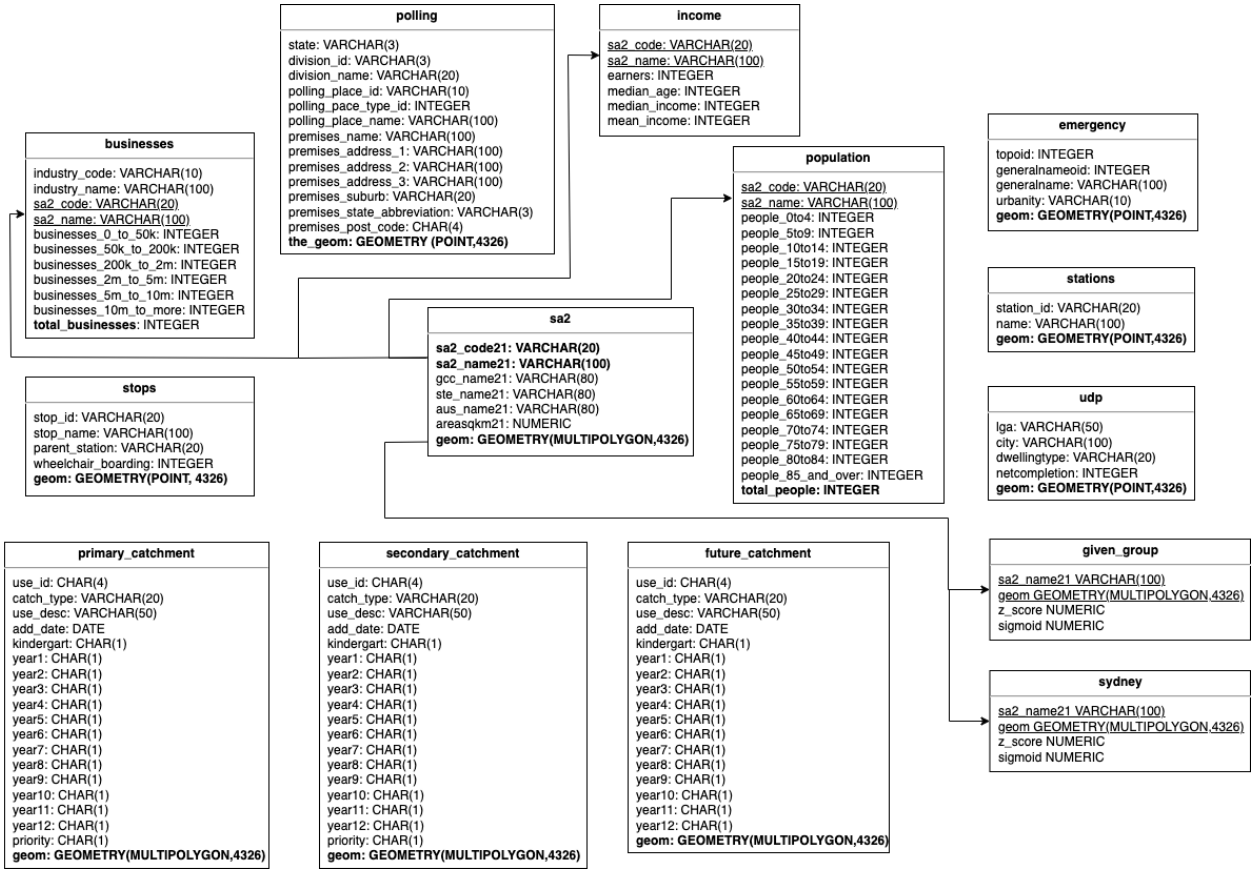
We have chosen the geom column from the sa2 table as a spatial index to speed up the queries that require a spatial join. This is used among all queries to locate whether facilities belong to that SA2 region.

The schema was established according to the diagram below. All raw files were processed to group values by SA2 regions (sa2_code21 and sa2_name21), which served as primary keys and are foreign keys in almost all tables. These unique indexes allowed aggregate table data to be joined with the respective SA2 region through another primary key of the sa2 table, geom, a column that includes geometry multipolygon values.

The datasets provided include SA2_2021_AUST_GDA2020.shp, Businesses.csv, Stops.txt, PollingPlaces2019.csv, Catchments.zip, Population.csv, and Income.csv. The stops file was converted to csv before processing. The catchments zip file contained shapefiles for primary, secondary, and future catchments areas.

The additional datasets include spatial data: **Emergency Services** (including *Fire and Rescue NSW_EPSG4326.json, NSW Police_EPSG4326.json, NSW SES Headquarters_EPSG4326.json, and NSW Rural Fire Service_EPSG4326.json*), **Station.shp**, and **dpe_netcompletion_202209.csv**.

Moreover, we have also created two more tables given_group (without additional datasets from task 30) and sydney(with additional datasets from task 3) to assist in the process of visualise data including z-score and sigmoid score to identify well-resourced area.



III. Score Analysis

Formula Computations:

z-scores for each region were calculated with the general formula:

$$Z = \frac{x-\mu}{\sigma} = \frac{score - mean}{standard\ deviation}$$

Calculations accounted for the area and population of a region. Data was filtered to leave out populations under 100.

z-scores was calculated for the number of retail businesses per 1000 people, health services per 1000 people, public transport bus stops per square kilometre, federal election polling locations (as of 2019) per square kilometre and school catchments areas per 1000 'young people' (sum of primary, secondary and future catchments) in each SA2 region.

The following formula was used to calculate the final sigmoid values:

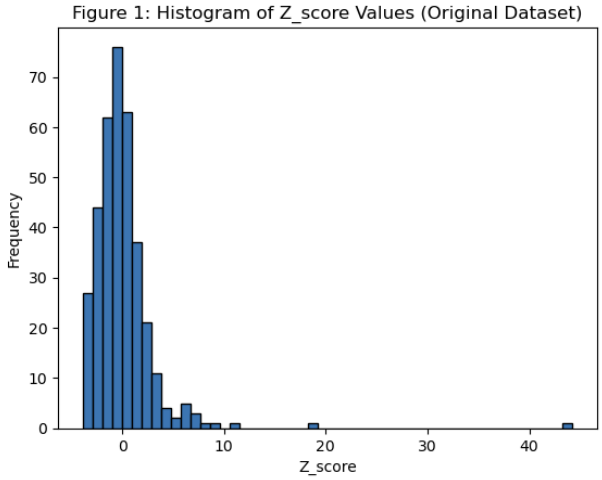
$$Score = S(Z_{retail} + Z_{health} + Z_{stops} + Z_{polls} + Z_{schools})$$

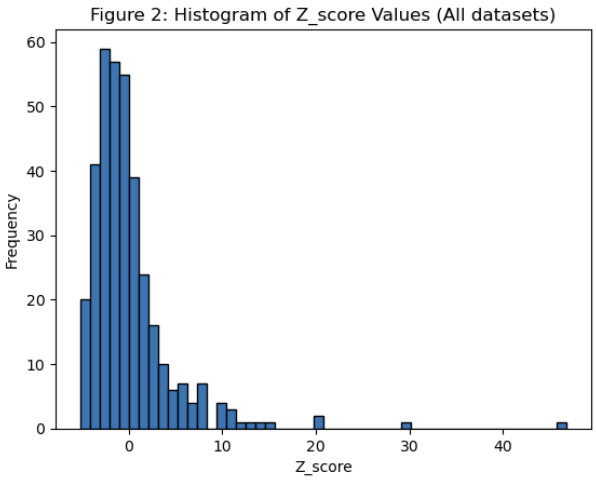
Implications of the extension of datasets and overall distributions:

We extended the base formula with the additional datasets by adding the Z scores for the number of emergency services available per square kilometre, stations per square kilometre, and the number of completed urban development projects per square kilometre. The modified base formula calculates a new sigmoid value as follows:

$$Score = S(Z_{retail} + Z_{health} + Z_{stops} + Z_{polls} + Z_{schools} + Z_{stations} + Z_{emergency} + Z_{growth})$$

Summary of the overall distribution

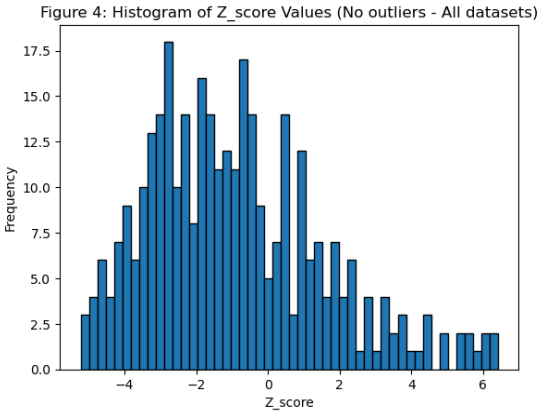




The overall distribution of the total z-scores from the given data sets is depicted in Figure 1. The distribution of the final dataset, which includes the additional datasets, is shown in Figure 2. Both histograms illustrate a higher frequency of lower z-score values, resulting in a right skewed distribution.

The additional datasets have slightly flattened the distribution, lowering the peak frequency by approximately 15. The range of scores were not noticeably affected. It seems that the standardised impact of the number of emergency services, stations, and completed urban projects follow a flatter distribution.

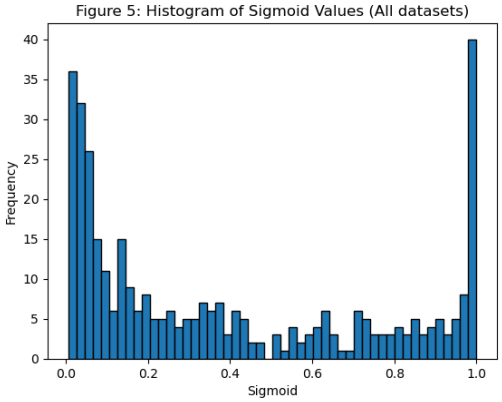
Figure 4 illustrates the distribution after removing outliers – z scores that were outside the interquartile range. We note that some extreme z scores may have been relatively larger sums of non-outlier component z scores, which could be a result of the well resourced nature of the sa2 region, which has a high z-score for all component z scores. It is not reasonable to remove them, as they comprise a significant proportion of well resourced areas. Moving forward with the analysis, we maintain all values from the combined dataset.



The total z score values from the combined dataset were transformed into sigmoid values with the following formula:

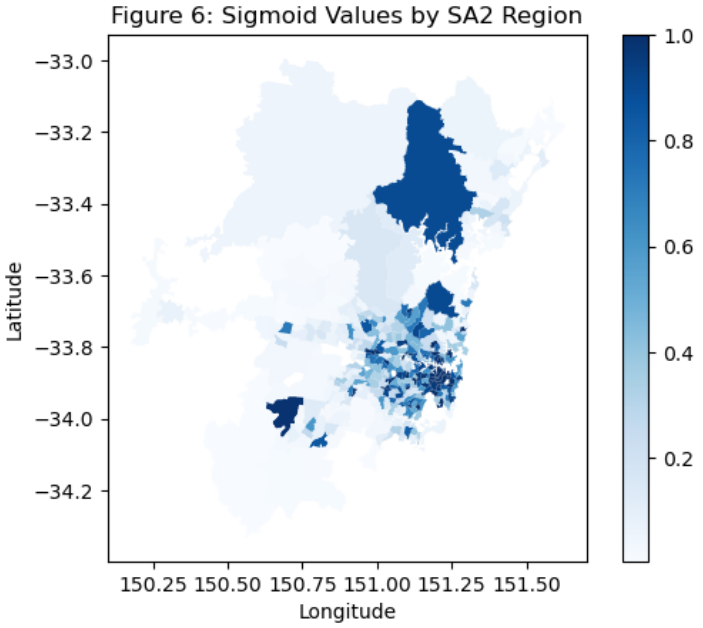
$$S(x) = \frac{1}{1+e^{-x}}$$

These results are illustrated in Figure 5. The right skew of the total z score values in Figure 2 impact the results in Figure 5, as there is a higher density of lower sigmoid values. The extreme results occur as a result of maintaining the outlier values.



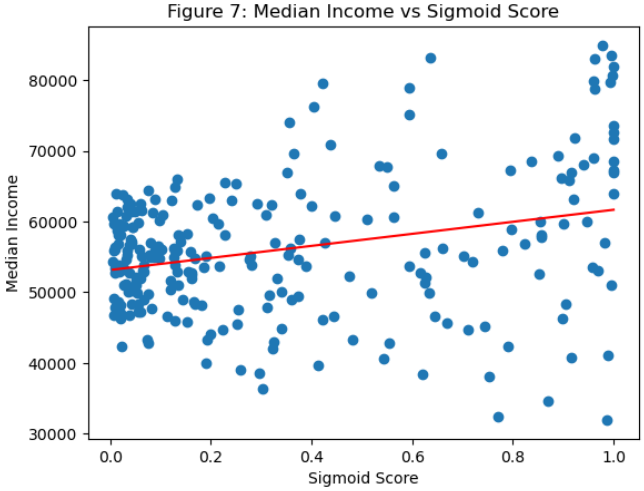
We generated a heatmap in Figure 6 which shows that well resourced areas are typically clustered in certain regions. The best resourced areas are often found around the city. This makes sense, as it is a popular area. The level of social and business activities in that area would likely require more resources and services. We note that most poorly resourced areas have sigmoid values under 0.4, which often correspond to the more rural areas within the Greater Sydney area. On the other hand, regions with higher sigmoid usually locate in Sydney CBD, North Sydney suburbs, Eastern suburbs, and Inner South and South Eastern suburbs

	sa2_name21	sigmoid
0	Sydney (North) - Millers Point	1.000000
1	Sydney (South) - Haymarket	1.000000
2	Darlinghurst	0.999974
3	Banksmeadow	0.999876
4	Parramatta - North	0.999793
...	...	...
355	Cranebrook - Castlereagh	0.025751
356	Blue Haven - San Remo	0.024833
357	Lake Munmorah - Mannering Park	0.023958
358	Bargo	0.021199
359	Warragamba - Silverdale	0.019597



IV. Correlation Analysis

Figure 7: Median Income vs Sigmoid Score



The correlation value between the median income and sigmoid value of a sa2 region is 0.292718 which is a weak positive correlation value. The weak correlation can also be observed in the figure 6 scatterplot with the line of best fit. The results were surprising, as we expected higher income regions would more strongly correlate to being well resourced.

This final observation infers that an sa2 region with a higher median income will not necessarily have a greater sigmoid value, which corresponds to being more well resourced. Ascertaining correlation against median income instead of mean income reduced the sensitivity to outliers, which is common in income data. Unfortunately, the correlation analysis is limited by its ability to describe linear relationships, which are not clearly visible in our scatterplot.

The values are dense for sigmoid values of less than 0.4, with around median income of 55000, then for sigmoid values greater than 0.4 it fans out slightly in a non-linear manner which the correlation value is unable to account for.

Some low median income areas are better resourced, likely due to government funding/support in developing areas. Higher median income areas are typically characterised by higher spending on resources, which fuels demand for more resources. We also expect people of higher income to move to better resourced areas.

Lastly, we observed that areas with lower median income tend to be poorly resourced, whereas areas with higher median income tend to be more well resourced. However, this is not a strong correlation and no clear conclusions can be made about the relationship between median income and the calculated sigmoid values.