DATA2001: Greater Sydney Analysis Report

Dataset Description

This assignment makes use of 10 datasets. Seven of them were given alongside the assignment and three were sourced online. Below is a short description of each dataset and a summary of the data-cleaning steps undertaken.

SA2 Regions
This dataset is obtained from the Australian Bureau of Statistics, containing the geometric boundaries, and descriptions, of Statistical Area Level 2 regions. For importing, the dataset was limited to regions in the 'Greater Sydney' area and with defined boundaries. These boundary geometries were converted to Well-Known Text for use with PostGIS, and the SA2 code and name fields were also extracted.

Businesses
This dataset is sourced from the Australian Bureau of Statistics, recording the number of businesses of different sizes, in different industries, in each SA2 region. Little cleaning was required for the data; the fields extracted described the industry (code and name), the SA2 region (code and name), and the total number of businesses in that industry, in that region.

Stops
Data describing public transport schedules, stop locations, and routes were sourced from Transport NSW's Open Data platform. This dataset was originally packaged in a GTFS format; only the subset regarding transport stops was used. Geometry POINT objects were created for each stop from their longitude and latitude. Alongside geometry, the ID and name of each stop was also extracted. Duplicates and missing values were removed from the ID field, to be used as a primary key.

Polls
Data describing polling locations in the 2019 Federal Election were sourced from the AURIN Data Catalogue. POINT objects were created to describe the position of each poll, from their longitude and latitude. The name and ID of each poll was also extracted.

Schools
This dataset was obtained from the NSW Department of Education website and contains information on the catchment areas for NSW government schools. It includes three shapefiles indicating the geographical regions in which students must reside to attend primary, secondary, and future government schools. For our purpose, we concatenated the three, and extracted the ID of each school, its type, and its catchment geometry. Duplicates of the ID-type combination were removed for use as a primary key..

Population
The dataset was obtained from the Australian Bureau of Statistics website and contains demographic data on the population distribution across various age groups of each SA2 region within the Greater Sydney area. A new column called 'young_people' was created, which aggregates the population count of individuals aged below 19. Two additional fields were extracted: SA2 code, and total population. Duplicate regions were removed, and data was filtered to ensure each SA2 code mapped to a valid SA2 code in the 'regions' dataset.

Income
The dataset was sourced from the Australian Bureau of Statistics website and contains income data for each of the SA2 regions in the Greater Sydney area. The SA2 code and the median income for each region were extracted. Missing values and duplicates were removed, and a filter was performed on the SA2 code column to ensure conformity with the 'regions' dataset.

Homelessness
The dataset was obtained from the Australian Bureau of Statistics website and contains estimated homeless populations in each SA2 region in the Greater Sydney area. The dataset was provided in Excel format and has two columns of interest: SA2 region names, and the number of homeless persons.

Health Services
Data concerning the location of ambulance stations, hospitals, and psychiatric facilities were sourced from the NSW Spatial Data Portal, provided in the Google Maps KML format. The name and position of each facility (described as a POINTZ object) were extracted.

NSW Budgets
This JSON dataset records information, particularly regarding estimated expenditure, concerning government projects enacted within the NSW region, in the 2015/16 fiscal year, sourced from the NSW Government Data website. The geometry of these projects was described in MULTIPOINT format, and was extracted alongside the estimated expenditure of each project, and the name of the responsible government agency.

Database Description

Sydney
The overall schema for the dataset, called 'sydney', contains 10 tables: SA2 region information and geometry ('regions'), businesses by industry and region ('businesses'), public transport stops ('stops'), polling booth locations in the 2019 federal election ('polls'), school catchment areas ('schools'), population estimates for each SA2 region ('population'), median income for each SA2 region ('income'), estimates for homeless populations in each region ('homelessness'), health service locations ('emergency'), and government expenditure on each region in the 2015/2016 financial year ('budget'). These tables are linked as per Figure 1 below.

Where Primary and Foreign Keys were not possible to link tables, geometry was used, according to the relationship between data and regions. For example, as hospitals exist at a POINT in the 'emergency' table, ST_Contains was used to join with 'regions'; government projects span larger areas, defined as MULTIPOINTS, and therefore ST_Contains and ST_Intersects (joined by an 'or' clause) are both used to join 'budgets' with 'regions'.

Thus, as the 'regions' geometry field is used consistently for joining, an index was created for it. The SA2 code in the 'regions' table is used to join non-spatial data, such as the 'businesses' table, and was also given an index. Additionally, as the 'stops' dataset is comparatively large, with over 60,000 valid entries, and is joined to 'regions' by its geometry field, a third index was also created for this column.

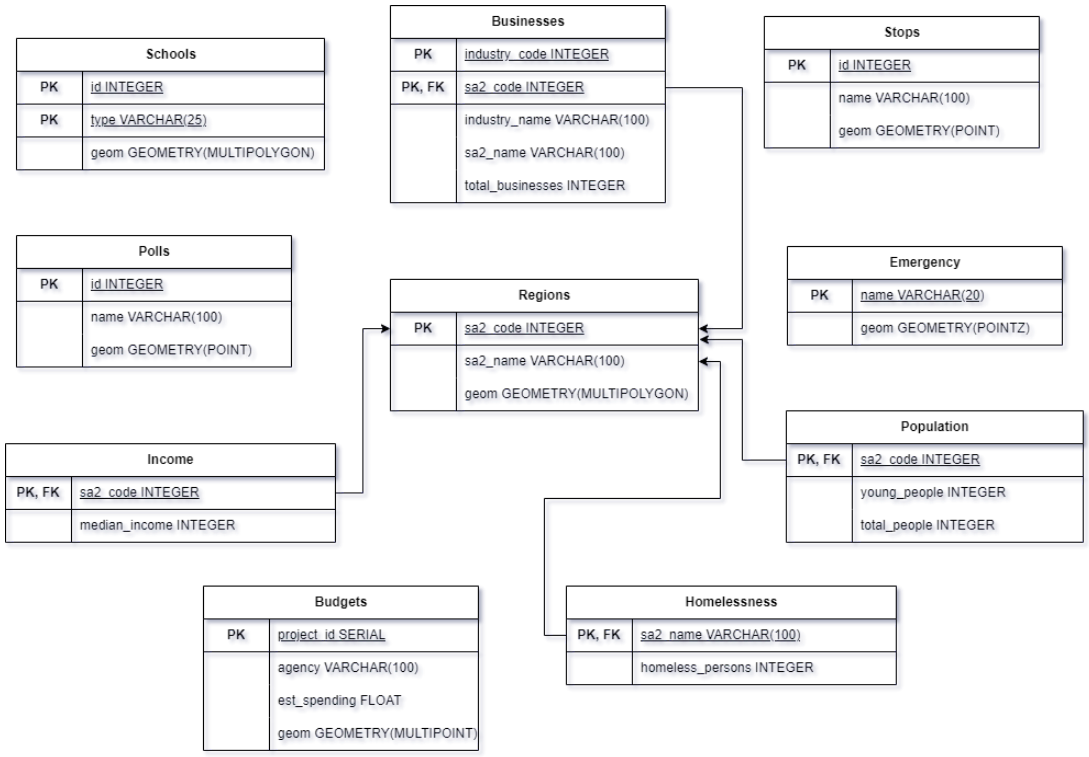

Figure 1 - Database Schema Diagram

Score Analysis

Function Construction

The score (s) of how 'well-resourced' a region is was calculated as the sum of its z-scores in each of 8 categories ($z_{category}$), calculated from the tables described above. A sigmoid function was then applied to this sum to standardise the range of scores:

$$s = \sigma(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools} - z_{homelessness} + z_{emergency} + z_{budget})$$

z_{retail} and z_{health} were calculated from the number of retail and health businesses in each region, respectively, per 1000 people. Similarly, $z_{emergency}$ and z_{budget} were calculated from the total number of health services in an area, and total amount of government expenditure, per capita. $z_{schools}$ was calculated from the number of school catchments, divided by the number of young people, as this is the most specific population the resource pertains to. The use of per-capita metrics aims to mitigate bias toward larger rural regions by controlling for population size. For example, Darlinghurst, which has a comparatively small population (10,776, while many go as high as 25,000), is able to achieve a very high score of 0.992 because it contains 6 hospitals/ambulance stations, and 10 school catchments. Other areas with much greater populations and similar resources, such as Chatswood and Auburn, score lower. This is consistent with the score being a measure of how 'well-resourced' a region is, as the same number of resources among fewer people should yield a higher score. This approach does have drawbacks, however. For instance, Banksmeadow scores 1.00 (the maximum), not because it has exceptional resources, but because its population is exceedingly small, at 507.

Similarly, z_{stops} and z_{polls} were calculated from the number of public transport stops and number of polling locations, respectively, divided by the area of the region. This metric was chosen, as opposed to per capita, because these resources are designed to be spatial; public transport stops must adequately cover a region, and polling locations must be significantly close to residents to facilitate voting. This allows smaller regions, which are closer to the CBD and generally have a smaller population, to compete with much larger rural areas. For example, the rural region 'Dural-Kenthurst-Wiseman's Ferry' has 361 public transport stops, but scores similarly to Surry Hills, which has only 36, because it is nearly 300 times the size.

$z_{homelessness}$ was calculated from the estimated number of homeless people per region, divided by the population of the region. However, because a high z-score on this metric would indicate a region is poorly resourced, this z-score was subtracted from the overall sum.

The summation of all z-scores was seen as an adequate aggregation function, because it avoids the inclusion of arbitrary weights to represent the importance of different categories. Rather, the sum represents, unambiguously, the total number of standard deviations above the mean a region possesses across various metrics. Any further abstraction here may risk obfuscating this underlying premise, and biassing the results. The sigmoid function was used to standardise the range of values of this sum, and constrains the distribution of scores to be more regular; it does not qualitatively alter the scores. Figure 2 shows the distribution of scores, and supports this notion; the masses at either end of the scale represent the long tails of the unscaled z-score sum.

Figure 2 - Distribution of Scores Across SA2 Regions

Summary and Distribution

Figure 2 below is a screenshot of the interactive map visualising SA2 regions in the Greater Sydney area, summarising their score, and the resources which produced that score; this map can be accessed through the 'sa2_score_map.html' file in the submission folder. Note that scores for areas in grey could not be calculated due to missing data, and that areas with populations fewer than 100 are not represented. See the Appendix for a snapshot of the top scoring regions.

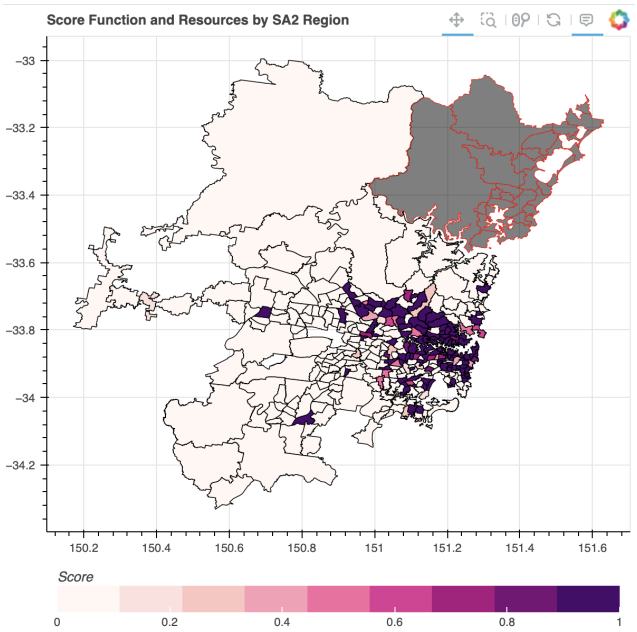

Figure 3 - Score Function and Resources by SA2 Regions Map Snapshot

Additional Dataset Discussion

The inclusion of homelessness data in our analysis serves a different purpose compared to metrics like the number of retail outlets or schools. While these metrics directly represent resources, homelessness reflects the overall resource situation of a region. A higher homeless population indicates a lack of housing, support, and other important factors that contribute to the well-being of an area. By including homelessness data, we can consider this broader context, indirectly including variables that cannot be adequately accounted for within our analysis.

The hospital and ambulance service data was included as a supplement to the z_{health} metric calculated from the 'businesses' table. While this score does reflect the medical resources of a region, it also includes non-essential businesses, such as cosmetic clinics and medical technology businesses, which while contributing to the economic resource of an area, do not contribute to its tacit resources. Explicitly including hospitals and emergency services was thought to represent the brick-and-mortar institutions which the residents of an area are most likely to access. However, the limitation of this data is that hospitals and emergency services tend to cluster; Darlinghurst, for example, has both a public hospital, private hospital, and ambulance station, on the same street, while many areas have none at all. While this could be seen as a flaw in the data, it could equally be construed as a feature: areas with clusters of these regions are extremely well-resourced to deal with medical emergencies, and should score commensurately.

The budget expenditure data was included to represent the resources flowing into a region, and may give an idea of change in resources. 'Schofields-East' for example, scores negatively on every metric, except per capita expenditure, where it is in the 98th percentile, and thus receives a mid-range score overall, 0.47. This is intended to indicate direction of resources, in that Schofields-East is likely to improve its other z-scores in future, because of this expenditure. While this could be expected to bias against rural regions, as large public works projects may be funded disproportionately in inner-city regions, where density, and therefore utility, is higher, the data does not support this. Although Miller's Point (an inner-city area) has the second highest z_{budget}, the other top spots are dominated by areas further from the city: Castle Hill, Epping, Ryde, Kellyville, and Chatswood.

Correlation Analysis

Correlation analysis was conducted to determine the strength and nature of relationships between the calculated score and the median income for each SA2 region, first broken down into each component:

Correlation between	Retail	Health	Stops	Polls	Schools	Emergency	Homelessness	Expenditure
and	Median Income							
Coefficient r =	0.145	0.484	0.090	0.30	0.042	0.197	-0.177	0.020

Table 1 - Correlations Between Score Components and Median Income

Table 1 shows a weak-to-moderate positive correlation between income and all metrics, except for homelessness. This is consistent with the intention of the measures; as each metric is supposed to reflect how "well-resourced" an area is, areas associated with higher-income populations should generally score higher. This finding also substantiates the notion that homelessness implies a lack of resources, justifying its inclusion as a negative term in the scoring formula.

The correlation between the aggregated score and the median income of each area was also investigated, with results presented in Figure 4.

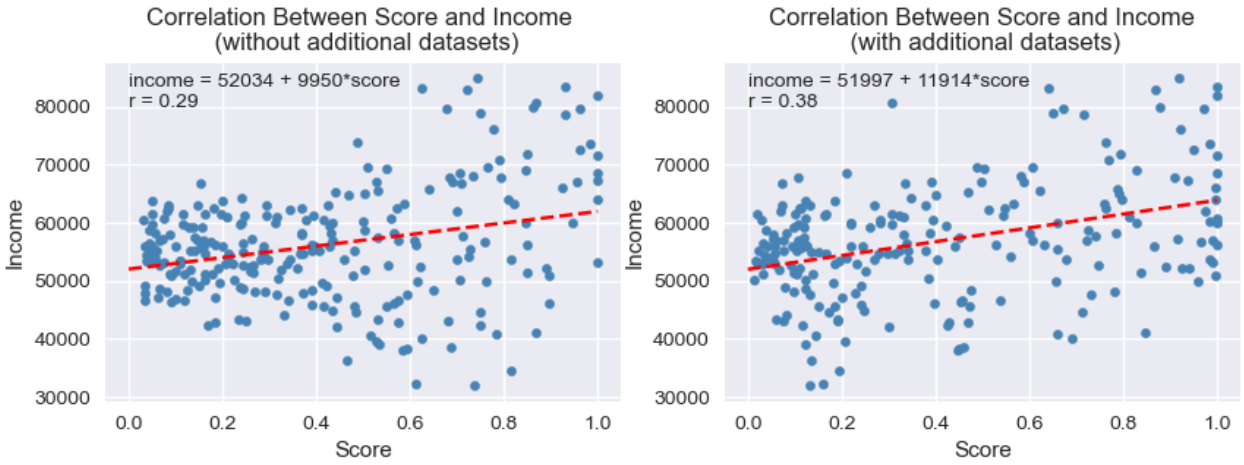

Figure 4 - Correlations Between Aggregated Score Function and Median Income

Both when the additional datasets (homelessness, emergency, and budgets) are included, and excluded, a positive correlation is observed between the aggregated score and median income. This positive correlation is consistent with the assumption that well-resourced areas are typically associated with higher-income groups, and provides evidence for the validity of the scoring function.

Appendix - Top Scoring SA2 Regions

Name	Population	Median Income (AUD)	Score
Sydney (North) - Miller’s Point	8199	–	1.00000
Banksmeadow	507	68584	1.00000
Sydney (South) - Haymarket	20346	–	1.00000
Chatswood - East	19770	–	0.99998
Castle Hill - Central	7685	56193	0.99997
Darlinghurst	10776	71676	0.99993
Hurstville - Central	12143	–	0.99983
North Sydney - Lavender Bay	12578	82003	0.99971
Camperdown - Darlington	8442	–	0.99962
Double Bay - Darling Point	10046	–	0.99958