Greater Sydney Regional Standard of Living Index

Nicholas Lukman
Ethan Yong

Dataset Description

Businesses.cvs
This data comes from the dataset ‘Counts of Australian Businesses, including Entries and Exits’ by the Australian Bureau of Statistics. It provides data on the counts of active businesses across Australia by sa2 code, split across different industries. We use this data for our ‘Retail’ and ‘Health’ metrics to measure retail businesses and health services per 1000 people respectively. As the dataset was well formed without any null values and already formatted correctly for our use, we did not clean it further.

Stops.txt
This data comes from the dataset “Timetables Complete GTFS” from Transport for NSW. This dataset contains data on all public transport stops across NSW, each with longitude and latitude coordinates. We use this data for our ‘Stops’ metric to measure how many public transport stops are located within the sa2 code. We transformed the data by converting the longitude and latitude values into a shapely point value and then transformed it to Well Known Text(WKT) format with the SRID value 7844 to represent the Geocentric Datum of Australia 2020.

PollingPlaces2019.csv
This data comes from the dataset “AEC - Federal Election - Polling Places (Point) 2019” from the Australian Electoral Commission. This dataset provides data on all the polling places across Australia. We use this data for our ‘Polls’ metric to measure how many federal election polling locations are contained in each sa2 code. This dataset also contains longitude and latitude values for each entry which we similarly transformed into WKT format like with Stops.

SchoolCatchments.zip
This data comes from the dataset “Individual catchment areas for NSW government schools” from the NSW Department of Education. It contains shape data on school catchments within NSW. We use this data for our ‘primary’ and ‘secondary metrics to measure the density of school catchments per 1000 young people within each sa2 code. We preprocess these datasets by also converting their geometry columns into WKT format.

Existing_Bicycle_Network.shp
This data comes from the dataset “Infrastructure Cycleway Data” from Transport NSW. It contains shape data on every cycling path within NSW as linestring values. We use this data for our ‘bikes’ metric which measures the length of cycling paths within each sa2 code. We preprocess the dataset by transforming the linestring values into multilinestring values for ease of storage within the database.

Electric Vehicle Charging Stations NSW - 20211207.csv
This data comes from the dataset “Electric Vehicle Charging Stations NSW” from Transport NSW. It contains shape data on the locations of each electric vehicle charging station within

NSW. We use this data for our ‘ev’ metric to measure how electric vehicle chargers are in each sa2 code. We preprocess the data by converting the geom column to WKT format.

Database Description

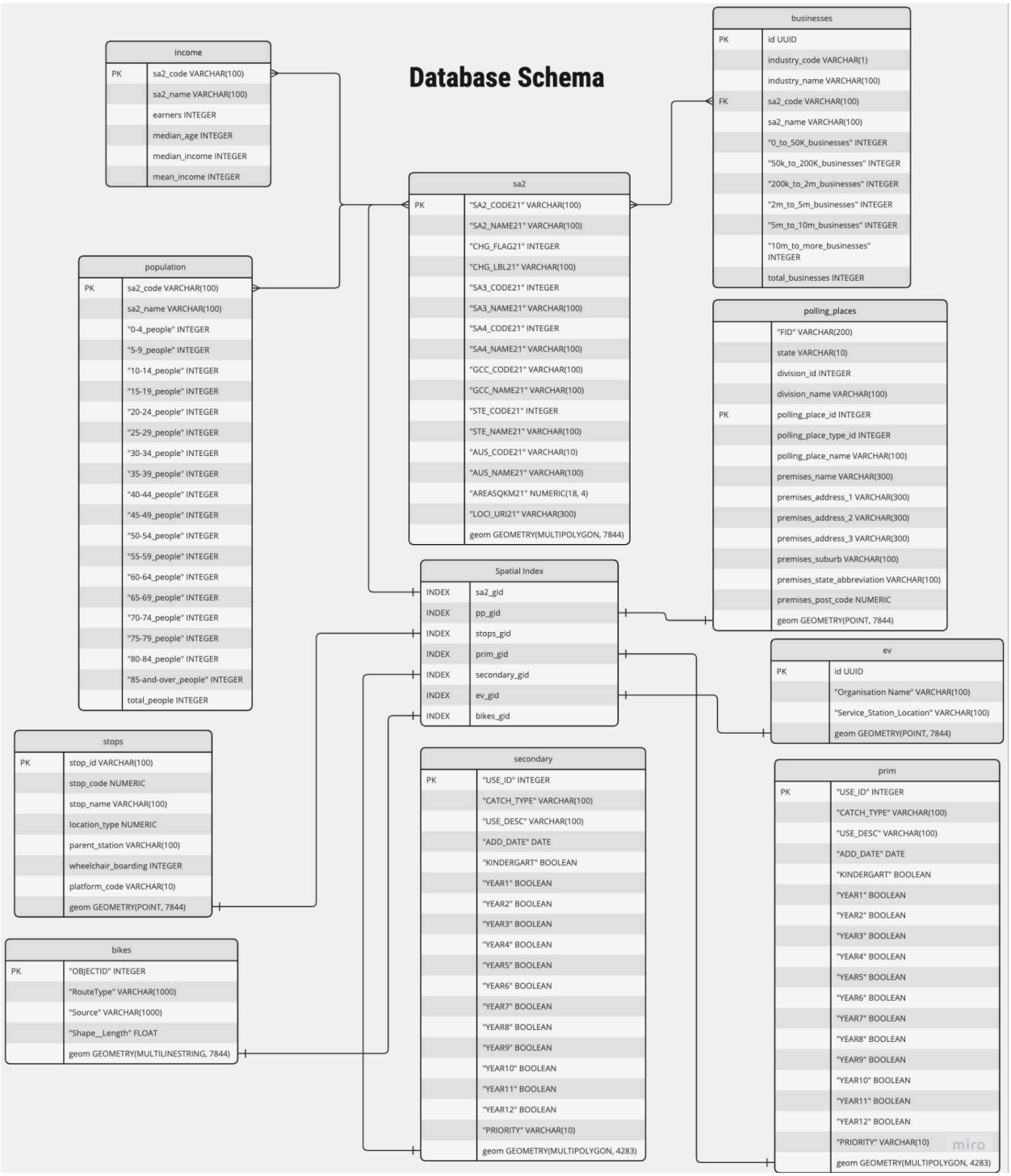

Figure 1. Database schema diagram

We uploaded the datasets into our local SQL server with the data schema shown above (Figure 1). For most of the tables, sa2_code was set as the primary key. For tables without sa2_codes, the tables were joined using the spatial operations, ST_Intersects and ST_Contains. To speed up the process of our spatial joins we created spatial indexes on the tables with geometry attributes. Several transformations (ST_Transform) were performed for spatial joins as some datasets did not have matching SRIDs, some were GDA2020 (SRID: 7844) and some were GDA1994 (SRID: 4283).

Score Analysis

Our “well resourced” score consists of several components: retail businesses per 1000 people, health businesses per 1000 people, public transport stops, federal polling locations and school catchment areas per 1000 people aged 19 years and under. For each component, z scores were calculated, summed together and inserted into a sigmoid function to produce a “well resourced” score. Scores were calculated for each SA2 region.

$$"well\ resourced"\ score\ = S(z_{health} + z_{retail} + z_{stops} + z_{polls} + z_{schools})$$

To compute the score function for each region, we queried our local SQL database server to obtain the necessary data needed to compute z scores for each component.

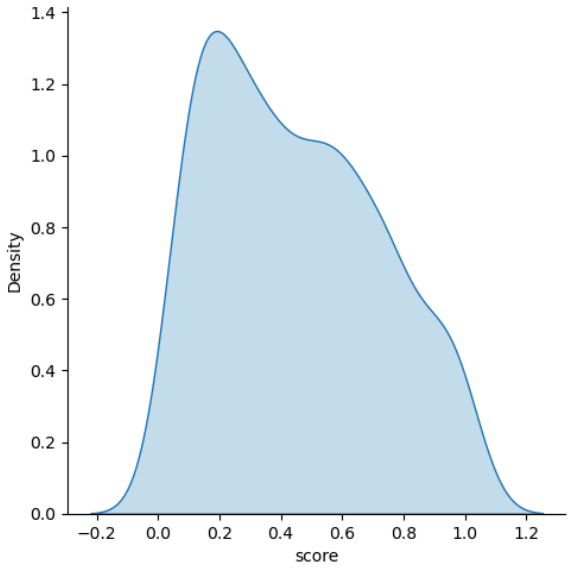

Figure 2. Kernel Distribution Density of scores

From plotting the density of scores across the sa2 scores we can observe a mostly normal distribution with a slight positive skew which seems to suggest a relatively balanced resource allocation across the Greater Sydney region (Figure 2).

From our heat map of the “well resourced” scores over Sydney, we observe a large variance of scores within the inner metropolitan regions of Sydney. Further out however, large clusters of both very high and very low scores appear. The northern regions appear to have more frequent high scoring regions, whilst the southern regions tend to have lower scoring regions indicating some disparity between the two areas (Figure 3).

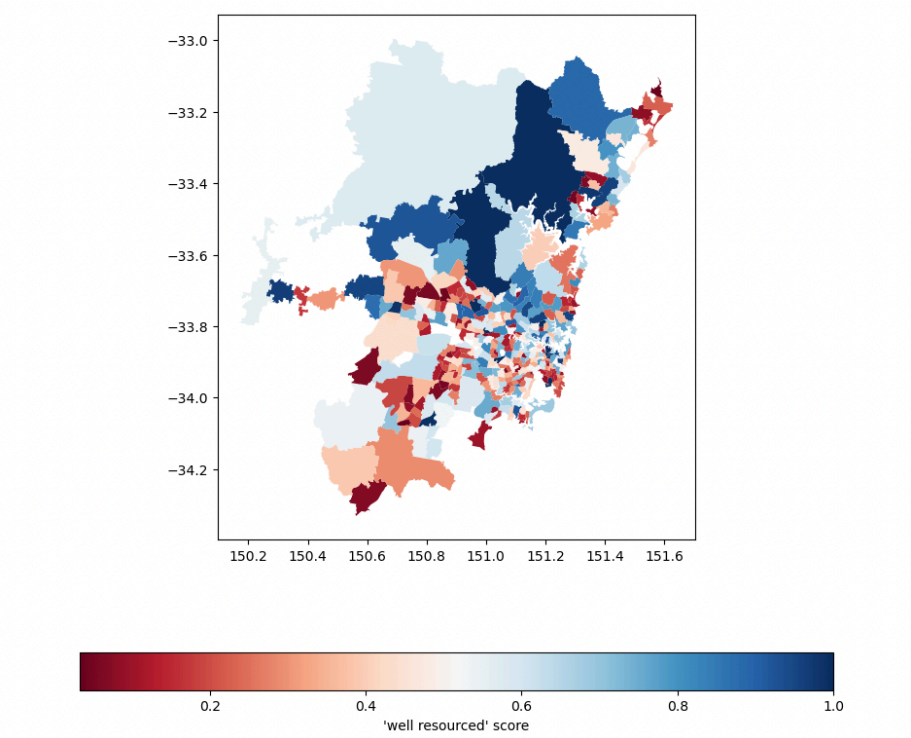

Figure 3. Map of greater sydney coloured by “well resourced” score

With the addition of the two extra variables, electric vehicle charging stations and bike lanes, into our scoring function we observe very little differences in both the distribution of score densities as well as in the spatial distributions. This would suggest the new variables don’t have much variance across the different sa2 codes.

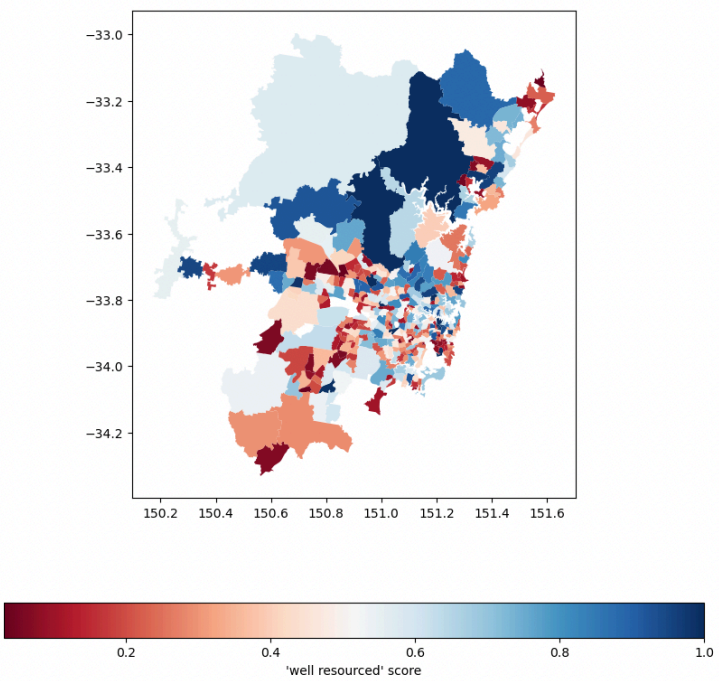

Figure 4. Map of greater sydney coloured by "well resourced" score with extension $"well\ resourced"\ score\ = S(z_{health} + z_{retail} + z_{stops} + z_{polls} + z_{schools} + z_{bike\ lanes} + z_{electric\ vehicle\ stations})$

The extended scoring function above does pose a few key limitations which might inhibit the scoring function's validity.
The bike path data only contains data of bike paths within the City of Sydney LGA which pertains to only a small segment of the Greater Sydney region. This would introduce an implicit bias in the scoring function to better favor sa2 codes within the inner Sydney area.

Due to the relative recency of electric vehicle technology within the Greater Sydney region, there are only 57 charging stations in the region. This would significantly bias the few sa2 codes that have one.

Correlation Analysis

Figure 5. Scatterplot of median income plotted against "well resourced" score

There appears to be a positive correlation between score and median income as represented by the upward sloping fitted line (Figure 5). For every one unit increase in score, median income increases by $6684.81. This implies that wealthier areas tend to have better public facilities and more retail stores. However, the fitted lines only has an r^2 value of 0.034056 which means that only 3.41% of the variation in median income can be explained by our "well resourced" scores, indicating that our "well resourced" score may not be a very good determinant of median income.

Figure 6. Scatterplot of median income plotted against "well resourced" score

With the addition of two extra variables (bike lanes and electric vehicle stations) into our scoring system the correlation between median income and “well resourced” score actually decreased, with an r^2 value of 0.0311 (Figure 6), 0.3% less than before. Following the new scoring system, our linear regression predicts that for every unit increase in “well resourced” score, median income increases by $6276.14, over $400 less than previously with the original scoring system. However, the changes to the correlation was not too significant as very minimal data could be obtained for our additional variables, mostly only encompassing the city of Sydney whereas our scoring system applies throughout Sydney's Greater Sydney area.