Lab 13-Group 1: Abigail Kreutz & Paige (Siyi) Wu

Tutor: Tianying Sheng

DATA2001

14 May 2024

'Greater Sydney' Data Analysis Report

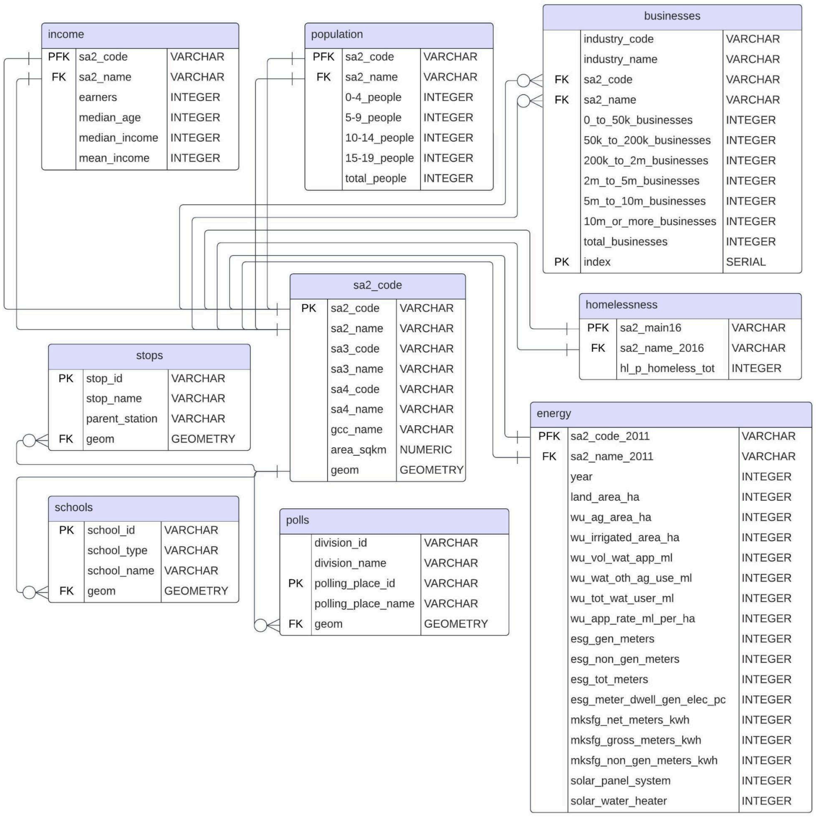
Dataset & Database Description

The datasets for the initial data loading and cleaning of Task 1 were taken from Canvas as provided by the course. These included 'SA2_2021_AUST_GDA2020,' a shapefile of the Statistical Area Level 2 (SA2) digital boundaries; 'Businesses.csv,' the number of businesses by industry and SA2 region reported by turnover size ranges; 'Stops.txt,' the location of all public transportation stop (bus and train) in General Transit Feed Specification (GTFS) format; 'PollingPlaces2019.csv,' the locations (and other premises details) of polling places for the 2019 Federal election; 'Catchments.zip,' a file containing the datasets of geographical regions in which students must live to attend primary, secondary and future Government schools; 'Population.csv,' the estimates of the number of people living in each SA2 by age range; and 'Income.csv,' the total earnings, age and income statistics by SA2 region.

Each of the .csv and .text datasets were imported into the jupyter notebook file using the pandas package in Python. We used the geopandas package of Python to load in the shapefile of the SA2 digital boundaries, as well as generate a 'schools' geodataset by combining the imported 'Catchment.zip' files of future, primary and secondary school catchment areas. Both geometries for the SA2 regions and school catchment areas are represented as MULTIPOLYGONS. For more efficient spatial querying capabilities in subsequent tasks, we created new 'geom' columns for the 'stops' and 'polls' tables. These columns stored POINT geometry types (derived from latitude and longitude coordinates), replacing the original latitude/longitude attributes, which were subsequently dropped from the 'stops' and 'polls' tables.

For Task 3, two GeoJson datasets were sourced externally, ‘Homelessness NSW.json’ and ‘Energy NSW.json.' Both datasets were originally published by the Australian Bureau of Statistics and downloaded from the Australian Urban Research Infrastructure Network (AURIN). Both datasets have been narrowed down exclusively to NSW data, as the original files contained information from the entirety of Australia were excessively large. ‘Homelessness NSW.json’ provides estimates of homelessness prevalence in 2016 by SA2 regions (ABS, 2016). The original dataset of 'Energy NSW' contained data from 2011 focusing on energy and environmental metrics within SA2 regions (ABS, 2011). However, when importing data from 'Energy NSW.json' into SQL, we opted to exclude several columns related to 'protected areas' to concentrate on analyzing the energy consumption patterns of each SA2 region rather than the environmental factors.

We established a connection to our database using the helper functions created in the Week 8 Spatial Data Tutorial. For the ingestion of our data, we defined schemas for each of the tables in our database using SQL. We created individual tables for each of the datasets, titled 'sa2_regions,' 'businesses,' 'stops,' 'polls,' 'schools,' 'population,' and 'income,' respectively, before populating them with the imported and cleaned datasets. See schema diagram of database below.



	income	
PFK	sa2_code	VARCHAR
FK	sa2_name	VARCHAR
	earners	INTEGER
	median_age	INTEGER
	median_income	INTEGER
	mean_income	INTEGER

	population	
PFK	sa2_code	VARCHAR
FK	sa2_name	VARCHAR
	0-4_people	INTEGER
	5-9_people	INTEGER
	10-14_people	INTEGER
	15-19_people	INTEGER
	total_people	INTEGER

	businesses	
	industry_code	VARCHAR
	industry_name	VARCHAR
FK	sa2_code	VARCHAR
FK	sa2_name	VARCHAR
	0_to_50k_businesses	INTEGER
	50k_to_200k_businesses	INTEGER
	200k_to_2m_businesses	INTEGER
	2m_to_5m_businesses	INTEGER
	5m_to_10m_businesses	INTEGER
	10m_or_more_businesses	INTEGER
	total_businesses	INTEGER
PK	index	SERIAL

	sa2_code	
PK	sa2_code	VARCHAR
	sa2_name	VARCHAR
	sa3_code	VARCHAR
	sa3_name	VARCHAR
	sa4_code	VARCHAR
	sa4_name	VARCHAR
	gcc_name	VARCHAR
	area_sqkm	NUMERIC
	geom	GEOMETRY

	homelessness	
PFK	sa2_main16	VARCHAR
FK	sa2_name_2016	VARCHAR
	hl_p_homeless_tot	INTEGER

	stops	
PK	stop_id	VARCHAR
	stop_name	VARCHAR
	parent_station	VARCHAR
FK	geom	GEOMETRY

	energy	
PFK	sa2_code_2011	VARCHAR
FK	sa2_name_2011	VARCHAR
	year	INTEGER
	land_area_ha	INTEGER
	wu_ag_area_ha	INTEGER
	wu_irrigated_area_ha	INTEGER
	wu_vol_wat_app_ml	INTEGER
	wu_wat_oth_ag_use_ml	INTEGER
	wu_tot_wat_user_ml	INTEGER
	wu_app_rate_ml_per_ha	INTEGER
	esg_gen_meters	INTEGER
	esg_non_gen_meters	INTEGER
	esg_tot_meters	INTEGER
	esg_meter_dwell_gen_elec_pc	INTEGER
	mksfg_net_meters_kwh	INTEGER
	mksfg_gross_meters_kwh	INTEGER
	mksfg_non_gen_meters_kwh	INTEGER
	solar_panel_system	INTEGER
	solar_water_heater	INTEGER

	schools	
PK	school_id	VARCHAR
	school_type	VARCHAR
	school_name	VARCHAR
FK	geom	GEOMETRY

	polls	
	division_id	VARCHAR
	division_name	VARCHAR
PK	polling_place_id	VARCHAR
	polling_place_name	VARCHAR
FK	geom	GEOMETRY

Some of the columns in 'sa2_regions,' 'stops,' 'polls,' and 'schools,' datasets were dropped before populating the tables in our database, as they were not necessary for future computations or analysis. Naming conventions were normalized across all fields of the 'sa2_regions' dataset, converting all column headers to lowercase and removing suffix '21'. We converted all column headers to lowercase for the 'schools' dataset, as well as renamed certain attributes for better clarity of what the values represented. 93 rows with duplicate values for the renamed attributes 'school_id' and 'school_name' were dropped from the dataset. We created a unique serial primary key called 'index' for the 'businesses' table. All other datasets involved were assigned primary keys for rows with unique values and foreign keys where applicable ('sa2_code' and 'geom').

Results Analysis

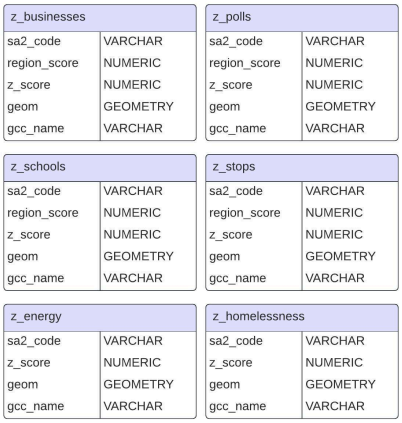
For Task 2, we wanted to calculate a score for how "bustling" each individual SA2 region (neighbourhood) where the population was at least 100. We created a z-score table for each of the datasets, titled 'z_businesses,' 'z_stops,' 'z_polls,' 'z_schools,' 'z_polls,' 'z_homelessness,' and 'z_energy,' respectively. Although the group assignment specification mentioned the description of the businesses z-score was intentionally left broad to encourage extensions of the scoring function, we decided to not select a cross-section of specific industries for the z-score calculation, as we believe all industries play an important part in the economy and thus should be included in calculating how "bustling" a SA2 region is, no matter how big or small the industry and its turnover size is.

For our energy z-score, we applied our calculation function to the values in column 'esg_tot_meters' (Energy Supply & Generation Total Meters) as the attribute encompasses a wide variety of metrics related to energy supply and generation, such as electricity consumption from the grid, electricity generation from renewable sources (e.g., solar panels), gas consumption, and water consumption. We chose this column because it focuses on residential usage, which is more related to our "bustling" score (compared to other usages, e.g., farming), and because it included a broad range of measures related to energy, which we believe is better for calculating a more holistic score than just focusing on one specific energy attributes. For our energy z-score, we filtered our dataset to only include values from the year 2011, as rows from the other years contained mostly null values. Due to geographic constraints affecting data collection, a limitation affecting both our energy and homelessness tables was missing SA2 regions. This resulted in certain regions' rows being populated with all null values upon merging the datasets, so the SA2 regions not originally included in these tables were given an energy and homelessness z-score of zero.

Each of our z-scores was calculated for every SA2 region within the Greater Sydney region with a population of at least 100. We first calculated region scores for each metric, dividing the total number of either businesses, public transport stops, federal election polling places, or school catchment areas by the total population of the SA2 region. For the businesses and school catchment z-scores, the total population of each SA2 region was multiplied by 1000 people or 1000 'young people' (ages zero to 19), respectively. The z-scores for each metric were then calculated by subtracting the average (mean) region score from the original region score, which we then divided by the standard deviation of that region's metric.

After calculating z-scores for the various metrics across SA2 regions in Greater Sydney, we synthesized these into a comprehensive "bustling" score. To achieve this, we applied the sigmoid function to the sum of all z-scores, except for the z-score of homelessness, where we subtracted its value from the sum, due to its inverse relationship with resource abundance. This sigmoid function transformed the aggregated z-scores into a bounded range between 0 and 1, representing the holistic level of resource abundance or "bustling" activity within each SA2 region. The resulting scores were then populated into a table named well_resourced_score.

The distribution of scores has a weak, positive (right) skew, with the most frequently occurring scores between 0.1 and 0.4. The tail of our distribution was greatly affected by outliers, as suggested by our standard deviation of 0.246, which indicates that there is a lot of variance in the observed data around the mean, 0.429. See schema diagram below for the tables used for score analysis and visual distribution of scores.



z_businesses	
sa2_code	VARCHAR
region_score	NUMERIC
z_score	NUMERIC
geom	GEOMETRY
gcc_name	VARCHAR

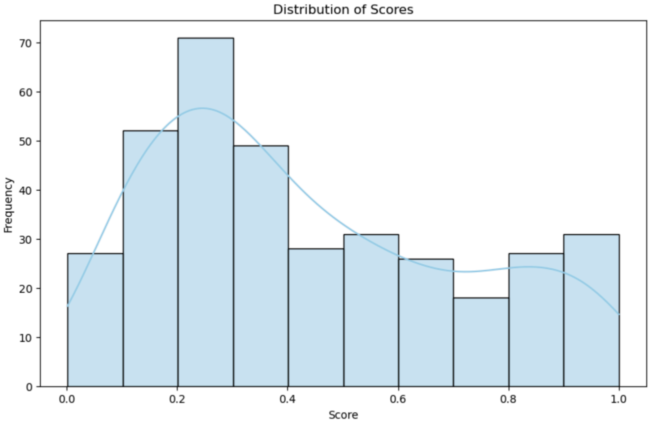
z_polls	
sa2_code	VARCHAR
region_score	NUMERIC
z_score	NUMERIC
geom	GEOMETRY
gcc_name	VARCHAR

z_schools	
sa2_code	VARCHAR
region_score	NUMERIC
z_score	NUMERIC
geom	GEOMETRY
gcc_name	VARCHAR

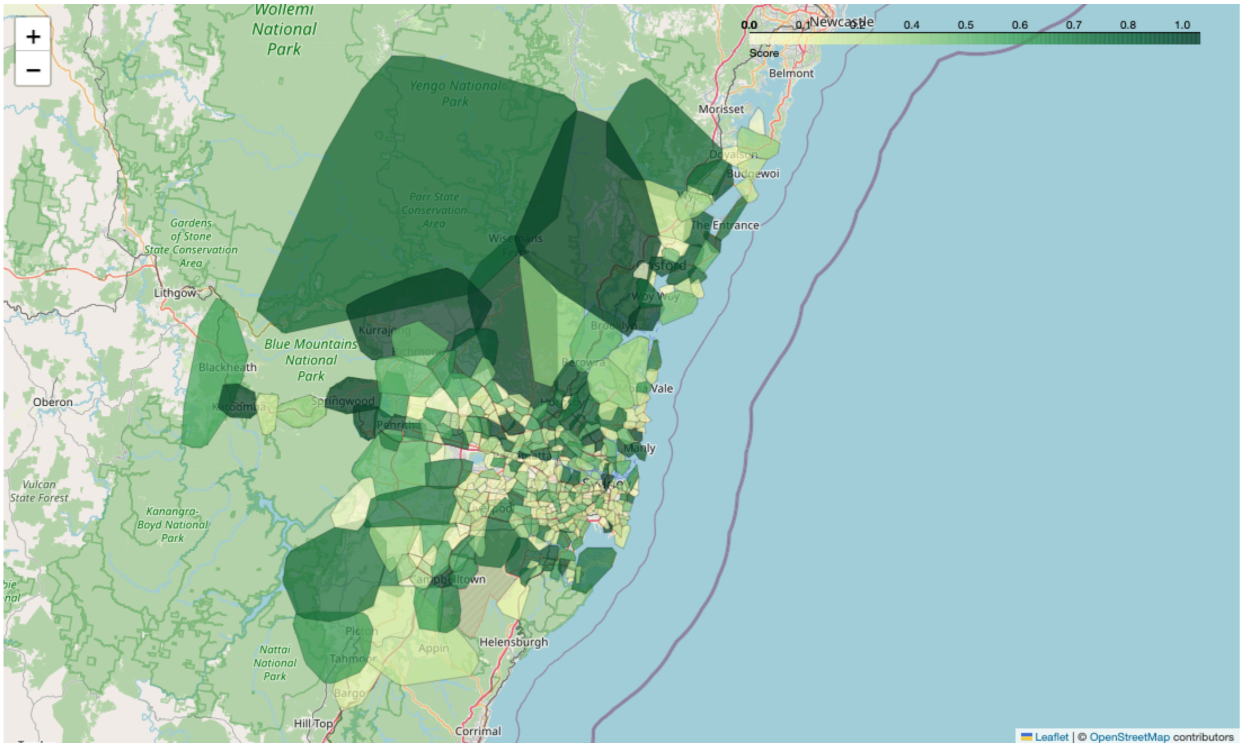
z_stops	
sa2_code	VARCHAR
region_score	NUMERIC
z_score	NUMERIC
geom	GEOMETRY
gcc_name	VARCHAR

z_energy	
sa2_code	VARCHAR
z_score	NUMERIC
geom	GEOMETRY
gcc_name	VARCHAR

z_homelessness	
sa2_code	VARCHAR
z_score	NUMERIC
geom	GEOMETRY
gcc_name	VARCHAR



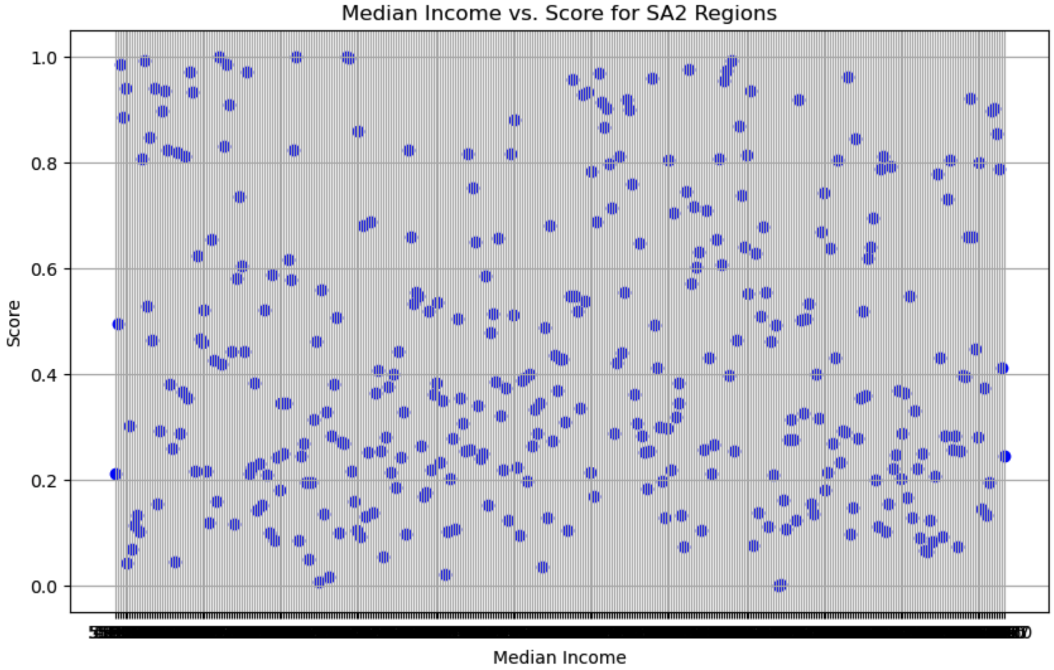
As expected, the Sydney CBD area had the highest businesses z-scores, greatly influencing its overall "well-resourced" score. However, this outlier affected all other 'businesses' z-scores, which limited and skewed every SA2 region to score extremely low. It was interesting to find that 'schools' z-score was higher for the outer regions of 'Greater Sydney'. We suspect that this has to do because there are less young people located in these regions, as the score for schools is calculated by number of schools/number of young people, so a smaller denominator leads to a higher score. The relevance of the 'schools' score is thus limited in reflecting how "bustling" a region is. For our first additional dataset looking at 'homelessness' across SA2 regions, we found that Inner Sydney regions scored higher due to homeless people living where there is more infrastructure, and thus rural regions scored lower for this metric. It is also important to note that for our second additional dataset looking at residential 'energy' usage, significant absences in data limited the consistency of the metrics used to calculate the overall “bustling” score for different regions. See below for a map of the overall "bustling" score.



Note. See Appendix for all map-overlay visualizations for each individual metric and details of particular trends, regions, or scores of interest, as well as limitations to the analysis.

Correlation Analysis

To assess the relationship between the computed scores and the median income of each SA2 region in Greater Sydney, a Pearson correlation test was conducted using the SciPy library on a join query of the income table and the well_resourced_score table from the SQL database. The Pearson correlation coefficient (r) was calculated to be 0.06, indicating a weak positive correlation between the two variables. This suggests that regions with higher computed scores tend to have slightly higher median incomes. However, the correlation was found to be not statistically significant, as indicated by the p-value of 0.285, which is greater than the commonly used threshold of 0.05. Therefore, while there is a trend of association between the computed scores and median income, this relationship is not robust enough to be considered significant within the context of this analysis. This is evident in the scatter plot below which exhibits no obvious patterns.



References

Australian Bureau of Statistics. *SA2 Energy & Environment - National Regional Profile 2010-2014*. (2011). Data.aurin.org.au. Retrieved May 14, 2024, from https://data.aurin.org.au/dataset/au-govt-abs-sa2-nrp-environment-2010-2014-sa2

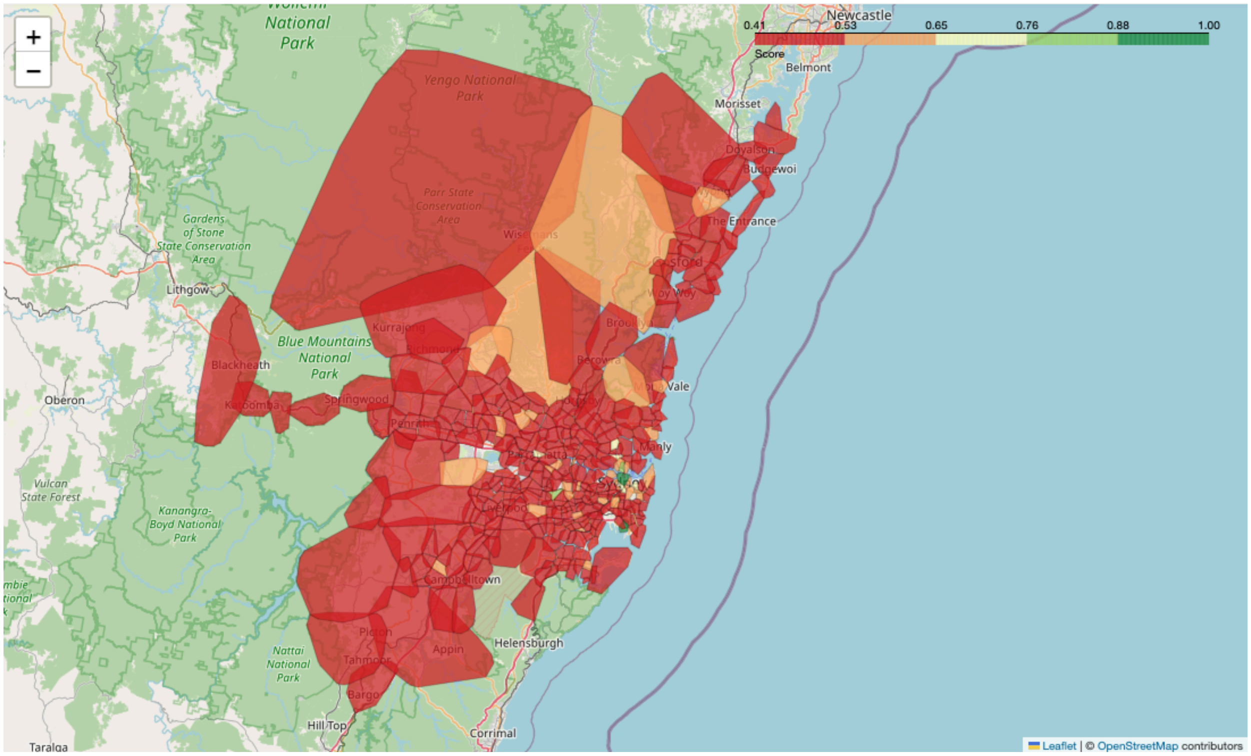
Australian Bureau of Statistics. *SA2 Estimating Homelessness 2016*. (2016). Data.aurin.org.au. Retrieved May 14, 2024, from https://data.aurin.org.au/dataset/au-govt-abs-sa2-estimating-homelessness-2016-sa2-2016

Appendix

Map Visualisations of Individual Metrics

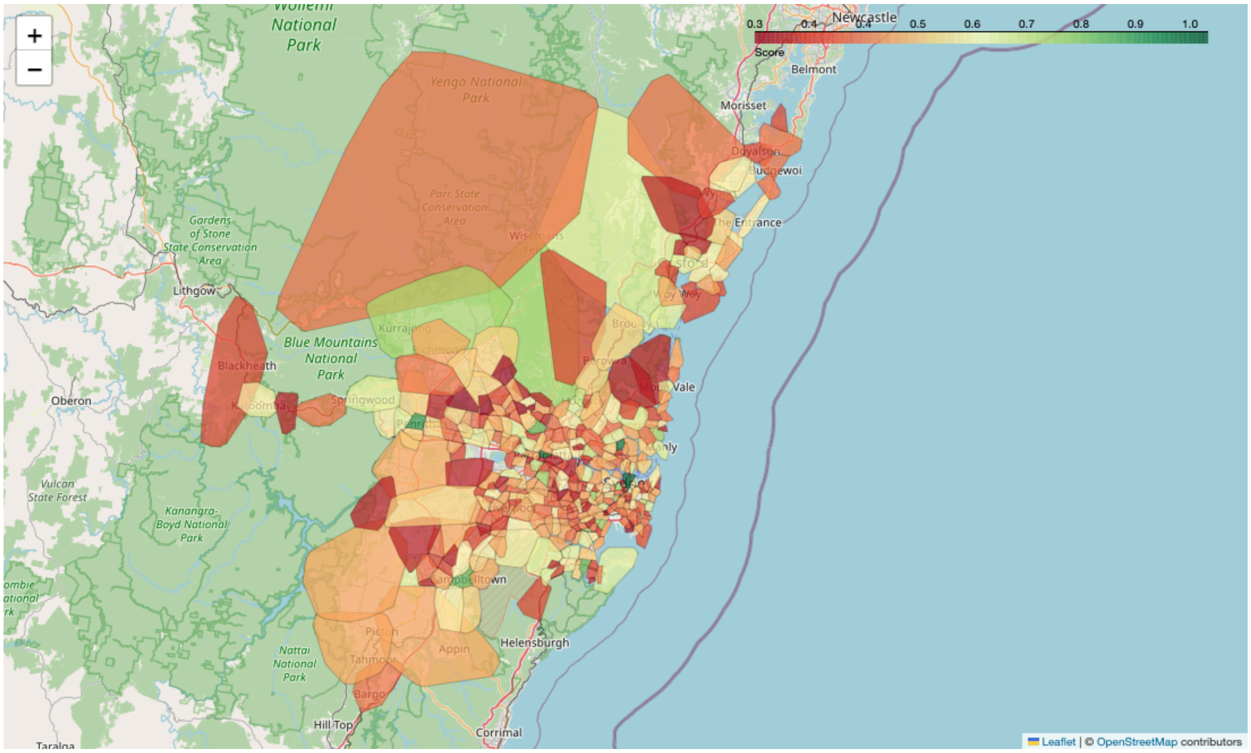
All following maps-overlay visualizations for each of the individual metrics are based on S(*metric* z-score). The sigmoid function is applied to each individual data point of the metric.

Businesses Score

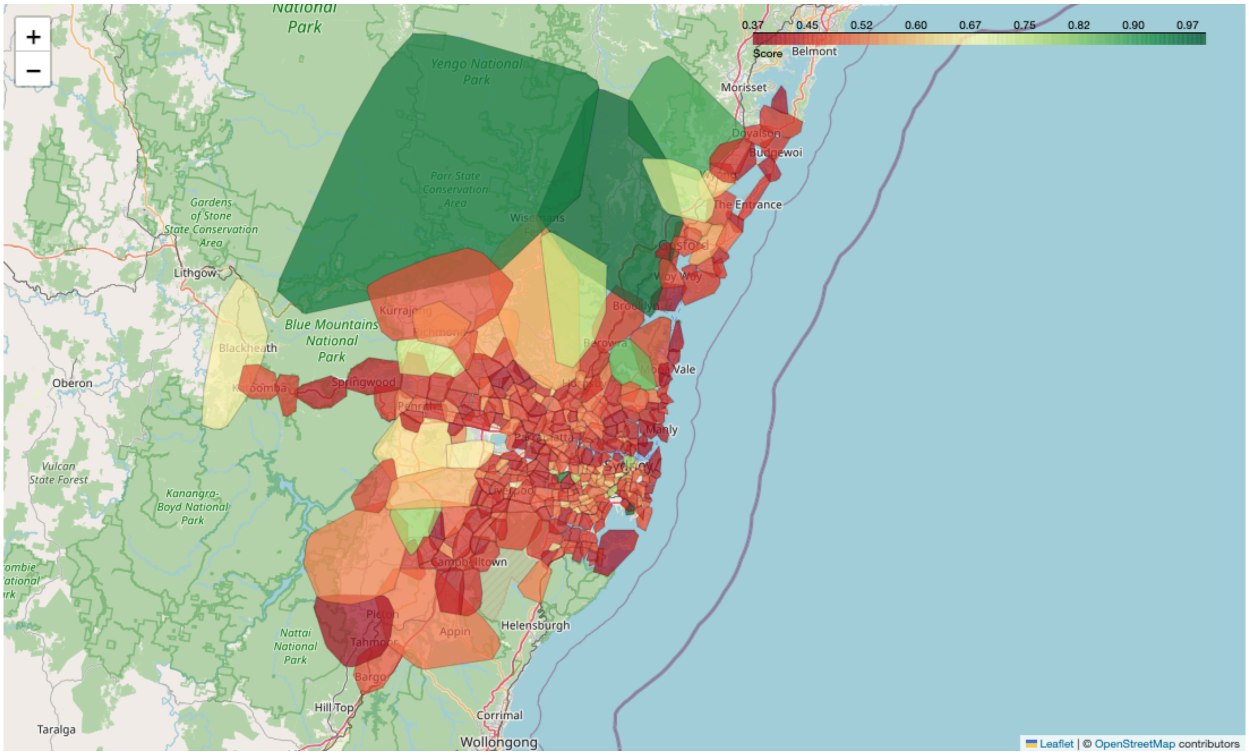


Note. Sydney CBD area scored highest, making it an outlier that skews businesses z-scores for all other regions.

Polls Score

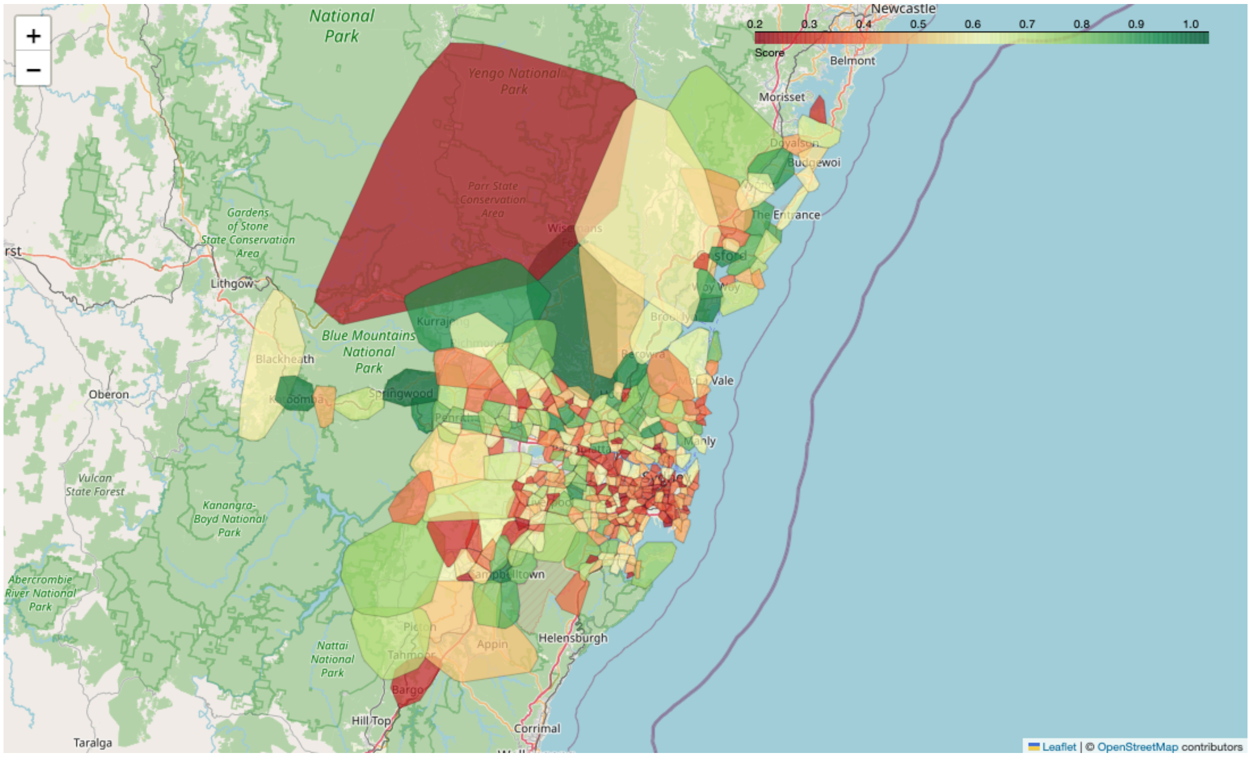


Schools Score

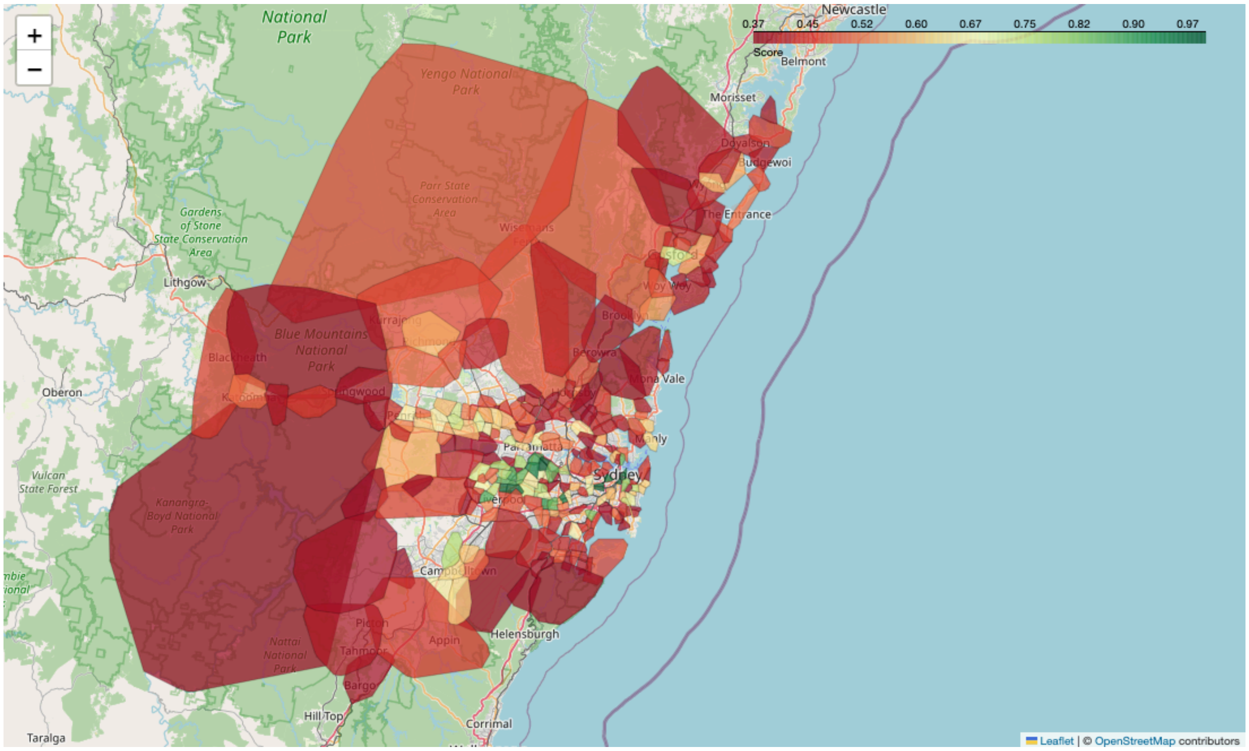


Note. Higher scores for outer regions of 'Greater Sydney' does not necessarily mean more schools are in these areas.

Stops Score

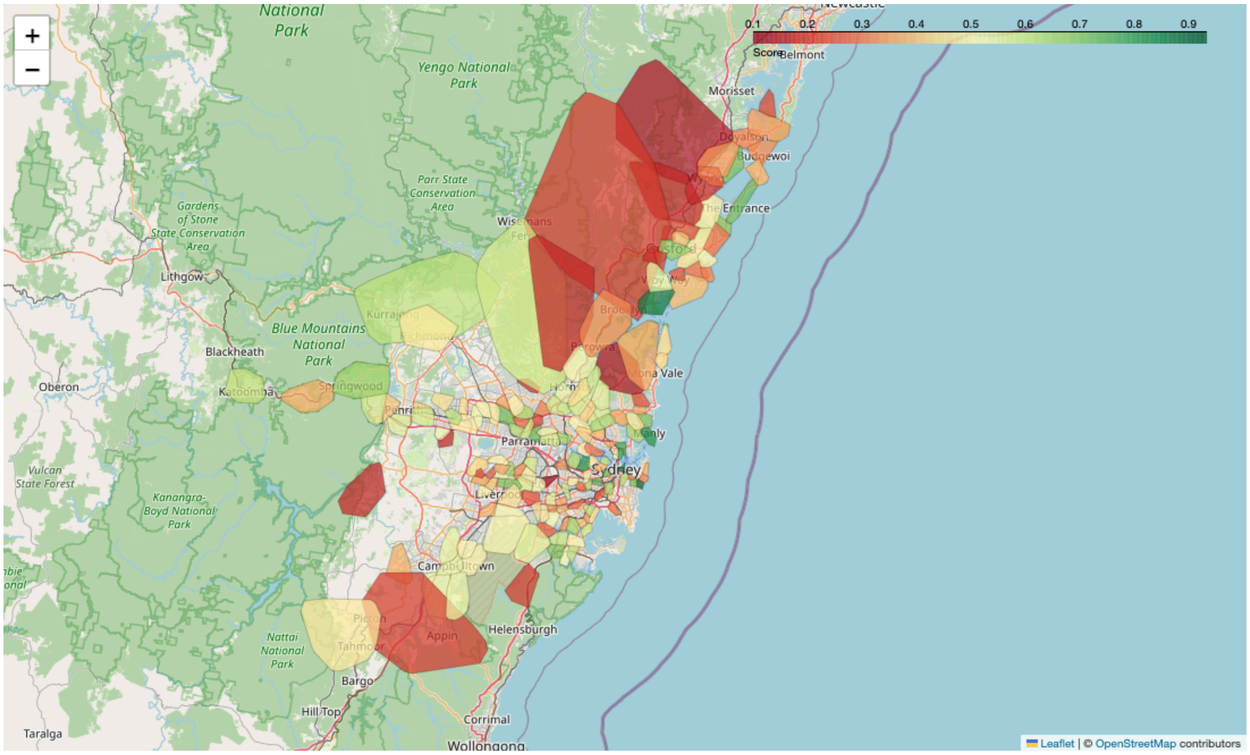


Homelessness Score



Note. Lower scores are better for this index and means less homelessness in the SA2 region.

Energy Score



Note. Significant number of absent data points from data set resulted in many regions not shaded.