Dataset Description

SA2 Regions: A shapefile of the SA2 digital boundaries. This data set is provided by the Australian Bureau of Statistics and describes Statistical Area Level 2 (SA2) digital boundaries and corresponding basic information like the name and code. In this analysis, code and GEOMETRY need to be used to join with other data.

Businesses: Number of businesses by industry and SA2 region, reported by turn over size ranges. There are multiple records for a sa2 code, and sa2_code can be joined with data set SA2.

Stops: Locations of all public transport stops (train and bus) in General Transit Feed Specification (GTFS) format. By latitude and longitude, it can be associated with SA2.

Polls: Locations (and other premises details) of polling places for the 2019 Federal election. By latitude and longitude, it can be associated with SA2.

Schools: Geographical regions in which students must live to attend primary, secondary and future Government schools. School data can be linked to SA2 through GEOMETRY. The columns of the three school data are the same, so I combined them into one for analysis.

Population: Estimates of the number of people living in each SA2 by age range (for ”per capita” calculations). There are only one record for a sa2 code, and sa2_code can be joined with data set SA2.

Income: Total earnings statistics by SA2. There are only one record for a sa2 code, and sa2_code can be joined with data set SA2.

Steal From Dwelling: This data is provided by BOCSAR describes the crime area of Steal From Dwelling.

Petrol Stations: This data is provided by AURIN and describes the petrol stations for suburb.

Data Preprocess: We use geopandas to read data set containing GEOMETRY, pandas to read other data. For the datasets Schools, SA2 Regions, Steel From Dwelling, and Petroleum Stations, filter outlier and use the function create_wkt_element convert the format for column GEOMETRY. Then importe data into the database. For the datasets Polls and Stops with latitude and longitude, use the method points_from_xy Convert latitude and longitude to Point and import it into the database. The rest of the data sets is imported into the database after filtering outlier.

Database Schema

A total of 10 tables have been established, with specific table names and fields as shown in the

following figure.

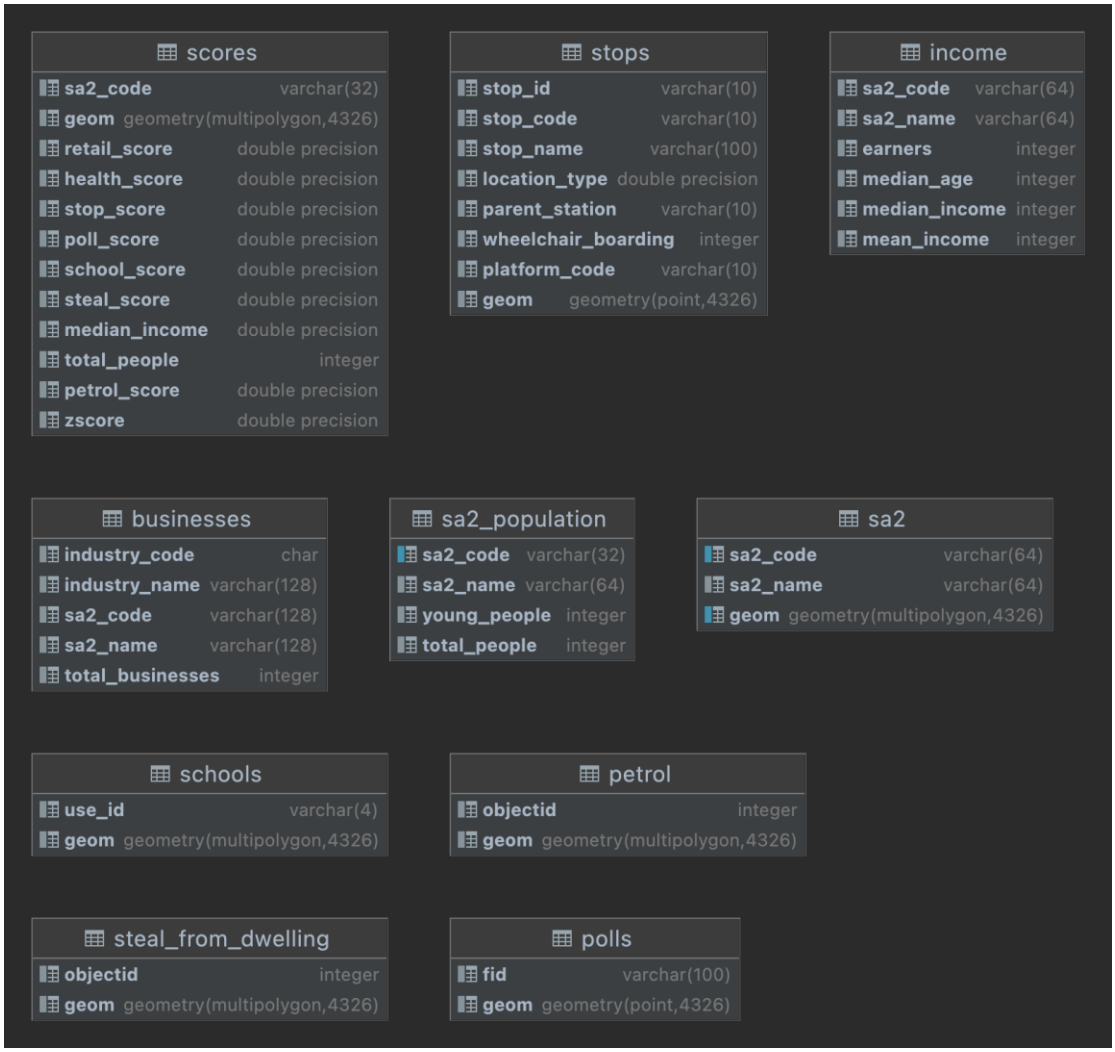


Table Indexes

stops: Established spatial index on column geom to efficiently join with SA2 region. There are 114718 pieces of data in the stops, and establishing an index significantly reduces the running time. When there is no index established, when running stops to associate the SA2 table through the "st_contains" method, the calculation time is about 7 minutes. After adding the index, the calculation time is only 20 seconds.

schools: Established spatial index on column geom to efficiently join with SA2 region.

polls: Established spatial index on column geom to efficiently join with SA2 region.

steal_from_dwelling: Established spatial index on column geom to efficiently join with SA2 region.

petrol: Established spatial index on column geom to efficiently join with SA2 region.

income: Established spatial index on column geom to efficiently join with SA2 region.

businesses: Established index on column sa2_code to efficiently join with SA2 region.

sa2_population: Established index on column sa2_code to efficiently join with SA2 region.

sa2: Established index on column sa2_code and spatial index on column geom to efficiently join with other tables.

Score Analysis

Firstly, calculate the 5 pre provided scores for the question and the 2 scores we have expanded ourselves, and save them to 7 views. The calculation logic for each score is as follows.

view_z_health

We joins the businesses table with the sa2 table using the sa2_code column, and then joins the resulting table with the sa2_population table also using the sa2_code column. This allows the query to access information about the total number of businesses and the Then selecting only the rows where the industry_name column in the businesses table equals 'Health Care and Social Assistance'. Finally, group by sa2_code to calculates the health score by dividing the sum of total businesses in the health care and social assistance industry by the total population in the region, expressed per 1000 people.

view_z_retail

The calculation method is the same as the health score, except for filtering business based on the industry name 'Retail Trade'.

view_z_school

We joins three tables: sa2, schools, and sa2_population, uses the st_intersects function to connect the sa2 table and the schools table, finding the schools in each region. Then, the GROUP BY clause groups the rows by sa2_code, uses the SUM(st_area(schools.geom)) function to calculate the sum of school areas in each region and the SUM(sa2_population.young_people) function to calculate the sum of young people in each region. The school score is calculated by dividing the former by the latter, multiplied by 1,000,000.

view_z_polls

We joins the sa2 and polls tables using a spatial comparison function and groups the rows by sa2_code. The COUNT(*) function counts the number of rows in each group, which is the number of polling stations in each SA2 region.

view_z_stops

The calculation method is the same as the poll score.

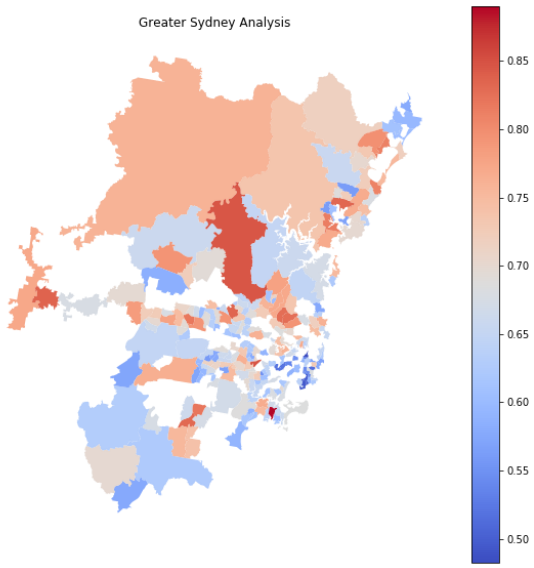
view_z_petrol

We joins the sa2 table and the petrol table using the st_contains function to find petrol stations in each region. Then group by sa2_code to calculate the petrol score for each region.

view_z_steal

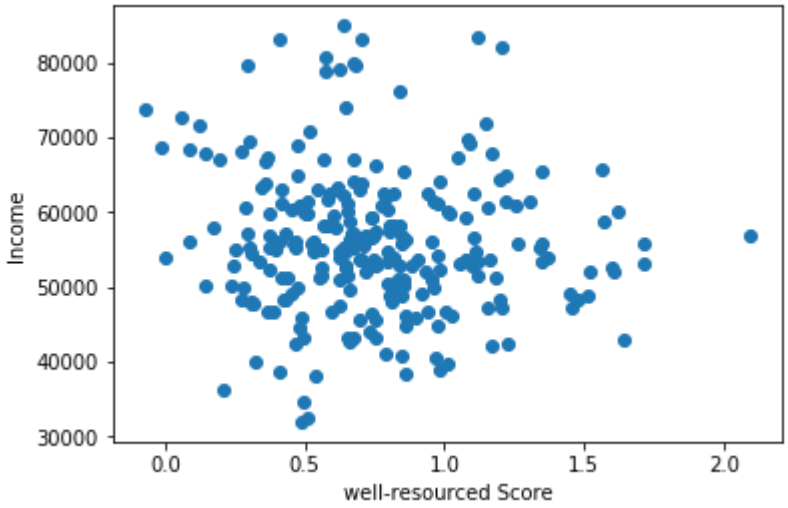
We joins the sa2 table and the steal_from_dwelling table using the st_intersects function to find theft incidents that occurred in each region. group by sa2_code tol calculate the steal score for each region, which is the sum of the area of theft incidents in each region divided by the area of the region.

After obtaining 7 scores, We using sa2_code to join all various scores and other data for each SA2 region. It joins the sa2 table with several other tables using the left join statement. Additionally, selects the median income for each SA2 region from the income table, joined using the sa2_code column. Then we use pandas merge the scores data with the population data using the sa2_code column as the join key. Filter out rows where the total population is less than 100 or the median income is less than or equal to zero. Define a max_min_scaler function that scales values of a given column to a range between 0 and 1 and fill any missing values with zero. Apply the max_min_scaler function to the columns containing the scores. Calculate the zscore for each region by adding the scaled scores for health, petrol, retail, public transport stops, polling stations, and schools, and subtracting the scaled score for theft incident, and use sigmoid to scale the score to 0-1. The final score for each region is shown in the following figure. From the graph, it can be seen that the central region has the highest score, while the southern region has higher scores than the northern region.



Correlation Analysis

Pearson correlation coefficient between zscore and median_income is -0.09, It indicates that there is no relationship between income and score, and from the scatter plot, the same effect can also be seen.


80000
70000
60000
50000
40000
30000
Income
0.0
0.5
1.0
1.5
2.0
well-resourced Score