Dataset Description

Data source: The shapefile of SA2 digital boundaries, and four metrics used to finish Task 2 are downloaded from the links in the spec sheet.

Extra 1: This provides the number of spatial information of walking count sites in several SA2 regions.

https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::walking-count-sites/explore

GeoJSON format which contains spatial data is chosen, satisfying two criteria at once.

Extra 2: This provides the number of dwelling units and value of buildings approved in SA2 regions.

https://www.abs.gov.au/statistics/industry/building-and-construction/building-approvals-australia/latest-release

XLSX file may not be able to be directly used, so we removed some unwanted rows. (process not in the jupyter notebook)

Obtain data: Saving the files in the same folder, and then imported them into python using corresponding modules, step by step in our Jupyter file.

Preprocessing SA2 regions and Extra 1 data:

- GeoPandas to convert “polygon” into “multi polygon” for geo-spatial data as the latter is more compatible.
- SA2_code should be converted into integer type.

Preprocessing four metrics:

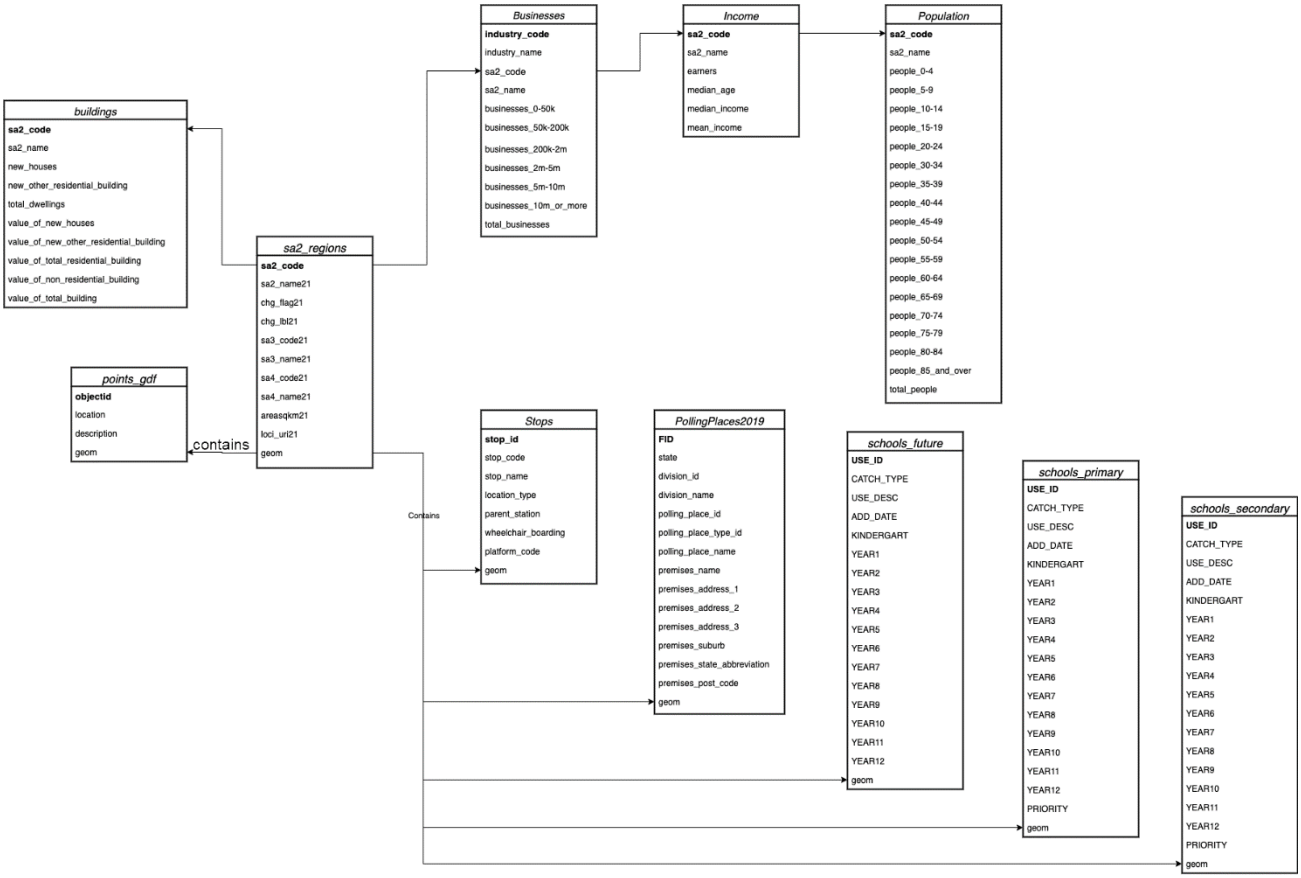
- Some columns in population/business data start with numbers, which is not allowed in SQL. So we have to rename those columns. We also rename other columns for visibility.
- Converting datetime data in school data.

Common preprocessing approaches:

- Converting geometry data into WKT so that it can be loaded to SQL.
- Converting the column names into lowercase in PostgreSQL, which is a good convention.
- Removing columns that will not be used to make the table cleaner.
- Removing some invalid, empty data.

Database Description

Attributes in **bold** are primary keys. Arrows signify foreign keys.



- sa2_regions Table:
  - Index idx_sa2_code21 on sa2_code21
  - Index idx_sa2_regions_geom on geom (Spatial)
- stops Table:
  - Index idx_stops_geom on geom (Spatial)
  - Index idx_stop_id on stop_id
- polling Table:
  - Index idx_polling_geom on the_geom (Spatial)
- schools_primary Table:
  - Index idx_schools_primary_geom on geom (SPATIAL)
  - Index idx_s_use_id on use_id
- businesses Table:
  - Index idx_sa2_code on sa2_code
  - Index idx_industry_name on industry_name

Reason for creating indexes: The primary reason is to make data retrieval faster. Without indexing, the database system would have to scan the entire collection of rows, which might take minutes and therefore inefficient when faced with massive queries. Indexing can also enforce constraints, therefore improving data integrity and consistency.

Score Analysis

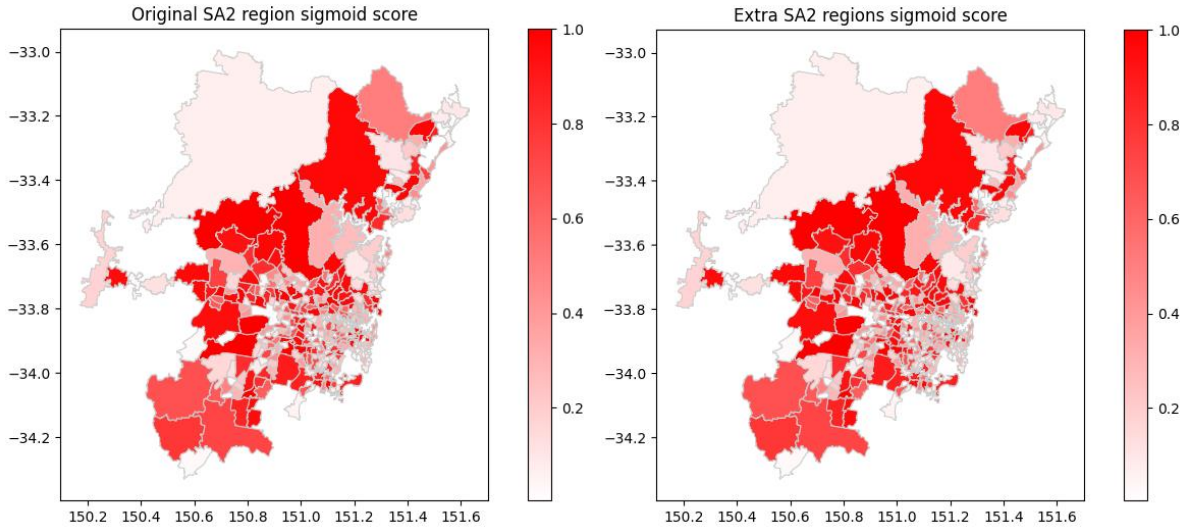
Original formula: Score = $S(z_{business} + z_{stops} + z_{polls} + z_{schools})$

Extra formula: Score with counter = $S(z_{business} + z_{stops} + z_{polls} + z_{schools} + z_{value\ of\ total\ building} + z_{number\ of\ objects})$

Where S() signifies Sigmoid function, more precisely, for Extra formula,

$$\frac{1}{1+exp\left[-\left(z_{business}+z_{stops}+z_{polls}+z_{schools}+z_{value\ of\ total\ building}+z_{number\ of\ objects}\right)\right]}$$, $\exp(x) = e^x$

The resulting map for two formulas:



From skimming the two maps, the results of two formulas are very close to each other.

1 “Bustling” means “full of activities”. To generalize the bustling score as one value from 0 to 1 with many factors, we use the sigmoid function, and standardizing the factors (getting their normalized Z-score). In the original formula, we have 4 factors: retail business, bus stops, election polls, schools. All these 4 factors are good measurements for the activity level of a region. We selected “retail trade” to represent business activity, as this is more related to everyday life. In the extra formula, we have 6 factors, adding the count objects and building values.

2 Walking count objects are placed in regions with high pedestrian activities, and regions with more count objects can be deemed to have higher pedestrian activities. This can be used in some small regions to compensate for the loss of bus stops due to their sizes. The value of buildings is the sum of the worth of buildings in a region, assessing how people value the activities in the region.

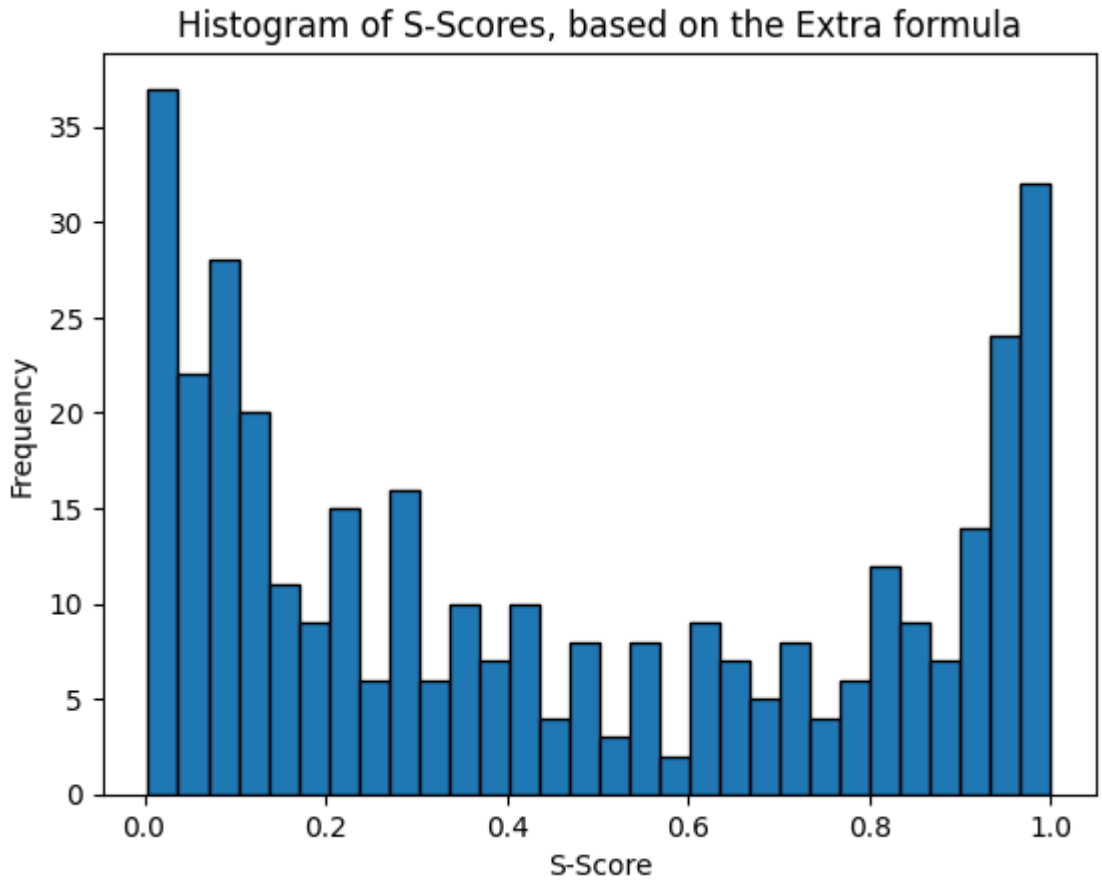
3 There are 359 regions, with 2 maximums being 1.000 and the minimums are just approximately 0.004. The distribution is something similar to “anti-normal distribution”: most regions have either an extremely high score or an extremely low score, and fewer regions

have something in between. (See Histogram in 5) Overall, regions with high and low scores are everywhere, but regions with medium scores are mostly on the coast.

4 Sometimes between every two highly bustling regions, there will be some not-so-bustling regions like intervals of cushion between them(Especially around the city center). Also, areas near Parramatta River generally have a higher bustling score.

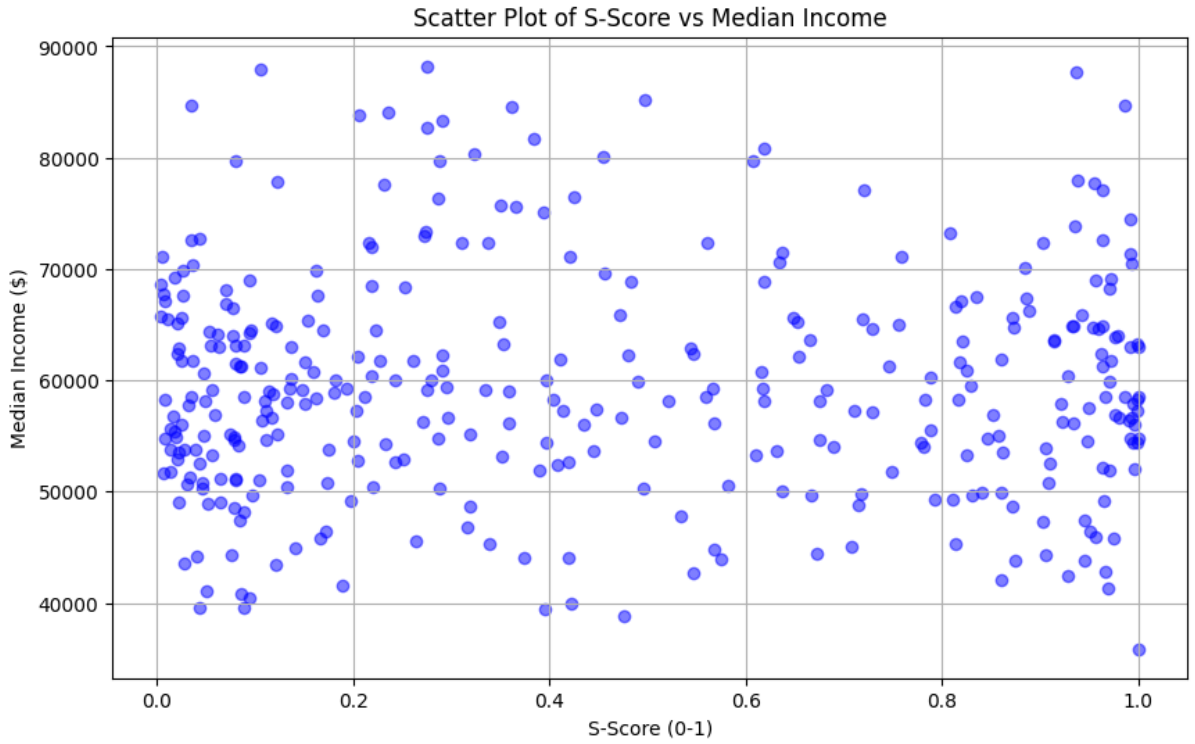
5 The table (some parts omitted) and the histogram for the data.

	sa2_code	sa2_name21	s_score	median_income	geom
0	117031644	Sydney (North) - Millers Point	1.000000	62966	MULTIPOLYGON (((151.22538 -33.85525, 151.22525...
1	117031645	Sydney (South) - Haymarket	1.000000	35875	MULTIPOLYGON (((151.19852 -33.87579, 151.19847...
2	115021297	Dural - Kenthurst - Wisemans Ferry	0.999999	58477	MULTIPOLYGON (((150.93444 -33.65020, 150.93451...
3	127011505	Austral - Greendale	0.999493	54758	MULTIPOLYGON (((150.71009 -33.91665, 150.71012...
4	115031300	Kurrajong Heights - Ebenezer	0.998970	58312	MULTIPOLYGON (((150.61982 -33.52875, 150.61986...
...	...	...	...	...	...
354	126021590	Putney	0.007201	67729	MULTIPOLYGON (((151.09996 -33.82540, 151.09996...
355	102021052	Summerland Point - Gwandalan	0.006541	51708	MULTIPOLYGON (((151.55618 -33.13680, 151.55601...
356	118021654	South Coogee	0.005212	71138	MULTIPOLYGON (((151.25113 -33.93111, 151.25109...
357	116021562	Acacia Gardens	0.004357	65756	MULTIPOLYGON (((150.91593 -33.72971, 150.91661...
358	128011605	Lilli Pilli - Port Hacking - Dolans Bay	0.004254	68606	MULTIPOLYGON (((151.12555 -34.06243, 151.12553...



Correlation Analysis

To determine whether there is a correlation between our S-score and median income, we can produce a scatter plot based on the previous table.



It is not so clear for the existence of a linear relationship to be seen. We conduct a Pearson correlation analysis:

Pearson correlation coefficient -0.040075960483588614, signify an **extremely weak, negative relationship**, as it is very close to 0. It implies that as the bustling score increases, the median income may slightly decrease, but the effect is minimal and not statistically significant. One interesting finding is, this might be a result of urban planning, where certain areas are designed to be highly active, and others are designed to be residential/less active, regardless of the median income of people there.

The result is a bit surprising as one may expect regions with higher median income to have a higher bustling score. One reason might be that SA2 regions are of different sizes, if the businesses, polls, schools, stops are more likely to be in large areas then these large areas will have a higher bustling score regardless of their possibly low median income. Another reason might be that people with higher income might bustle somewhere else.

The usefulness of our score is that it provides some measurements, based on actual data (instead of guessing), on how bustling each SA2 region is, in the format of a number from 0 to 1, which is more precise compared to ambiguous plain language.

One limitation of building value is that there might be time delay in the bustling as approved buildings take time (months or even years) to finish construction and contribute to the bustling level of the region, and investment in building might not fully reflect actual economic activities in that area. Another limitation is that we only have the number of walking count objects, not the actual pedestrian number which can more precisely signify the activity level.