Greater Sydney Analysis Report

Bu Seong, Hantha Nguyen, Jenny Yuan

The area of Greater Sydney can be defined by over 350 “Statistical Area Level 2”, or “SA2”, regions, within which small communities interact. In this report, the “liveability” of each region will be analysed, using collated datasets regarding different aspects of these regions.

1. Dataset Description

All datasets were imported into Python and cleaned with Pandas and GeoPandas by the following:

- Filtering null values where appropriate
- Renaming columns into simple lower-case labels for convention
- Dropping columns that are not of interest
- Converting spatial columns into Well-Known Test (WKT) format for consistency with SQL, by importing the geoalchemy2 library and using the helper function create_wkt_element

Cleaned datasets were loaded to PostgreSQL for analysis after defining appropriate schemas. Below are brief descriptions of each dataset:

SA2: The data source for SA2 was provided and is a shapefile containing information on and the geometries of all SA2 regions in Australia.

Businesses: The provided CSV file contains data regarding the number of businesses in different industries within each SA2 region.

Stops: The provided text file contains data regarding the locations of bus and train stops in Greater Sydney. A new column of point geometries was created using the longitude and latitude data.

Polls: The provided CSV file contains the 2019 Federal election polling locations.

Schools: The provided shapefiles contain the geographical regions in which students must live to attend primary, secondary, and future government schools.

Population: The provided CSV file contains estimations of people in each SA2 region by age range.

Income: The provided CSV file contains statistics for the earnings and earners in each SA2 region.

Task 3 Additional Datasets:

Aged Care Services: The GeoJSON file contains the locations of aged care services.

Index of Socio-Economic Advantage: The CSV file contains socio-economic scores for each SA2 region, where a score increase indicates more high-income households and skilled occupations.

Index of Education and Occupation: The CSV file contains scores for the educational and occupational level of each SA2 Region, where a score increase indicates higher educational attainment and higher employment rates.

2. Database Description

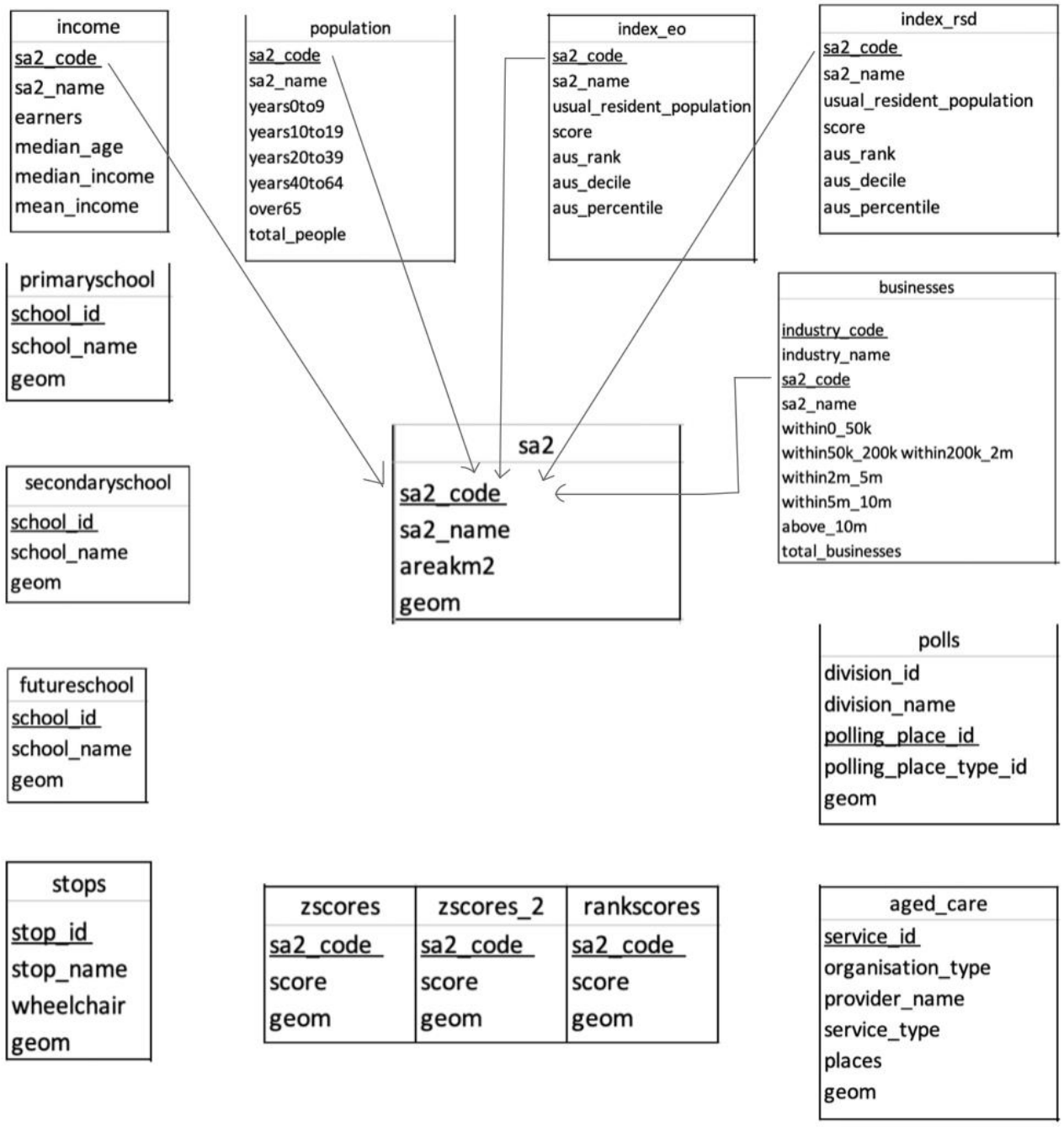


Figure 1. Database diagram.

The columns chosen to be kept upon initial cleaning are the ones listed on the database diagram (Fig. 1). Ultimately, not all were needed and were dropped during score analysis. In the database diagram, the underlined attributes are the primary keys of each table. Many of these are also foreign keys, referring to the sa2 column of the sa2 table (shown using arrows in Figure 1). The businesses table has a composite primary key, consisting of sa2 code and industry code, as this creates a unique key for the table. For other tables, foreign keys could not be established as they refer to the sa2 table using geometries, which are not unique values and hence cannot become foreign keys. These tables were joined to the sa2 table using the ST_Contains() function in SQL, which identified whether their geometries were contained within a particular SA2 region.

A spatial index named sa2_index was created on the sa2 table using the geometry, to increase the speed of queries for additional analysis in task 3 which required the use of multiple tables.

3. Score Analysis

The formula used to calculate scores for each SA2 region with at least 100 people is as follows:

Score = $S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools})$ where:

- z_{retail}: z-score of the total number of retail businesses per 1000 people
- z_{health}: z-score of the total number of health services per 1000 people
- z_{stops}: z-score of the number of public transport stops
- z_{polls}: z-score of the number of polling places (as of 2019)
- $z_{schools}$: z-score of the total number of school catchments per 1000 people aged 0-19
- z-score = $(x - \mu) / \sigma$ where μ=mean, σ=standard deviation, and x is defined above

S is the sigmoid function, $S(x) = 1/(1+e^{-x})$, taking the sum of all z-scores as an input.

	sa2_code	score	z_retail	z_health	z_stops	z_polls	z_schools
0	102011028	0.194375	-0.151934	0.441162	-0.512237	-1.028006	-0.170813
1	102011029	0.166567	-0.404105	-0.486794	0.479566	-1.028006	-0.170813
2	102011030	1.000000	0.895103	0.293690	-0.005065	1.292995	12.715386
3	102011031	0.975979	0.593465	1.184555	1.268500	0.828795	-0.170813
4	102011032	0.991259	0.274663	1.276619	2.057434	1.292995	-0.170813
...	...	...	...	...	...	...	...
242	128021536	0.728417	-0.444973	-0.609718	0.919115	1.292995	-0.170813
243	128021538	0.956928	-0.268278	0.018315	0.254156	0.828795	2.267871
244	128021607	0.638276	-0.510268	-0.591211	0.547189	1.292995	-0.170813
245	128021608	0.052605	-0.504943	-0.866499	-0.320638	-1.028006	-0.170813
246	128021609	0.010840	-0.749630	-0.630709	-1.470228	-1.492206	-0.170813

Figure 2. Overall score and z-scores for each SA2 region using original datasets.

Implications of additional datasets: The additional datasets chosen as mentioned previously were Australian aged care services, the index of relative socio-economic advantage and the index of education and occupation. As the assignment was investigating the ‘liveability’ of each SA2 region, we focused on the positive aspects of what made a region well-resourced, being in this case the accessibility of aged care services and the index/scores for the latter two datasets. Higher levels of which all indicate a more desirable place to live and hence that the region is well-resourced.

Taking these datasets into consideration, the scoring function then follows:

Score = $S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools} + z_{aged\_care} + z_{index\_rsd} + z_{index\_eo})$ where:

- z_{aged_care}: z-score of the total number of aged care services in each region
- z_{index_rsd}: z-score of the socio-economic score for each region
- z_{index_eol}: z-score of the education and occupation score for each region

The z-score and sigmoid function was calculated the same way as done in task 2.

Similarly, after incorporating the new datasets into the scoring function z_{aged_care}, z_{index_rsd}, and z_{index_eo} it returned different scores as to the scores without it in the regions (Fig. 3).

	sa2_code	score	z_retail	z_health	z_stops	z_polls	z_schools	z_aged_care	z_index_rsd	z_index_eo
0	102011028	0.258748	-0.151934	0.441162	-0.512237	-1.028006	-0.170813	-0.905047	0.762490	0.511900
1	102011029	0.138598	-0.404105	-0.486794	0.479566	-1.028006	-0.170813	-0.905047	0.599861	0.088355
2	102011030	0.999999	0.895103	0.293690	-0.005065	1.292995	12.715386	-0.905047	0.198710	-0.390920
3	102011031	0.995255	0.593465	1.184555	1.268500	0.828795	-0.170813	1.679307	0.263762	-0.301752
4	102011032	0.989986	0.274663	1.276619	2.057434	1.292995	-0.170813	0.387130	-0.278335	-0.246023
...	...	...	...	...	...	...	...	...	...	...
242	128021536	0.914918	-0.444973	-0.609718	0.919115	1.292995	-0.170813	-0.258959	0.968487	0.679089
243	128021538	0.998835	-0.268278	0.018315	0.254156	0.828795	2.267871	2.971484	0.437232	0.244398
244	128021607	0.733856	-0.510268	-0.591211	0.547189	1.292995	-0.170813	-0.258959	0.784174	-0.078834
245	128021608	0.070988	-0.504943	-0.866499	-0.320638	-1.028006	-0.170813	-0.905047	0.957645	0.266690
246	128021609	0.020290	-0.749630	-0.630709	-1.470228	-1.492206	-0.170813	-0.905047	1.174483	0.367003

Figure 3. Overall score and z-scores for each SA2 region with additional datasets.

An interactive Choropleth map was implemented for the scores, with darker red colours reflecting the more well-resourced areas, or higher scoring regions and lighter yellow colours reflecting the less well-resourced ones. No notable pattern or trend can be identified using purely the map.

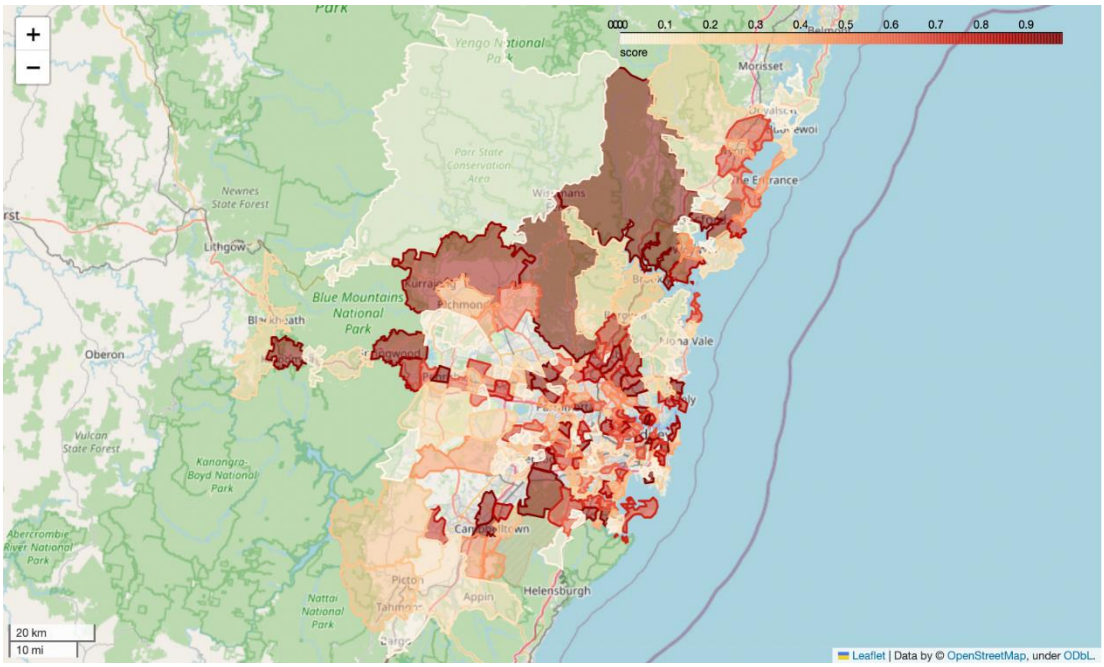


Figure 4. Interactive Choropleth map showing scores for each SA2 region (using z-scores).

Before and after the addition of new datasets, as seen in Figures 5 and 6, the distribution of the scores is more even in the middle and peak at the extremities. This is more evident after the addition of new datasets (Fig. 6). This was unexpected as we would have expected a normal distribution due to the Central Limit Theorem, which states that sample means often follow a normal distribution. This suggests that the z-score method may be ineffective at indicating the liveability of each region.

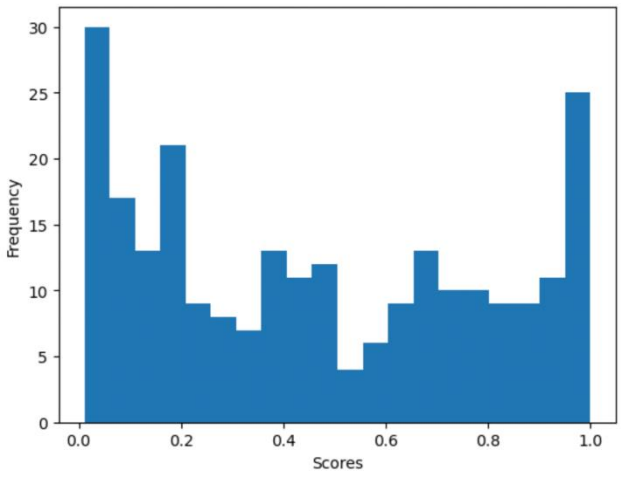


Figure 5. Original score frequency.

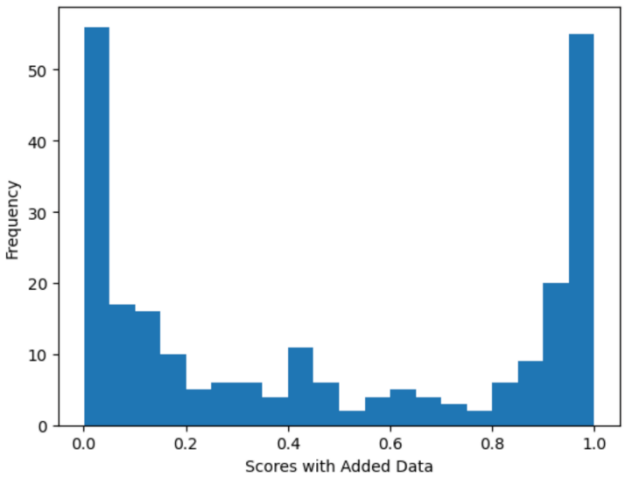


Figure 6. Score frequency after dataset addition.

The scatter plots of scores against distance from the CBD (which was defined as Surry Hills as the Central/Town Hall areas could not be included due to Nulls) show little to no relation between the two variables. Like the histograms, the data points also gathered more at extremities after the inclusion of the additional datasets, which suggests that there may be a severe divide in socio-economic state or education and occupation in Greater Sydney. Otherwise, due to the random distribution/lack of a trend, there is little to conclude from the visualisation of the data.

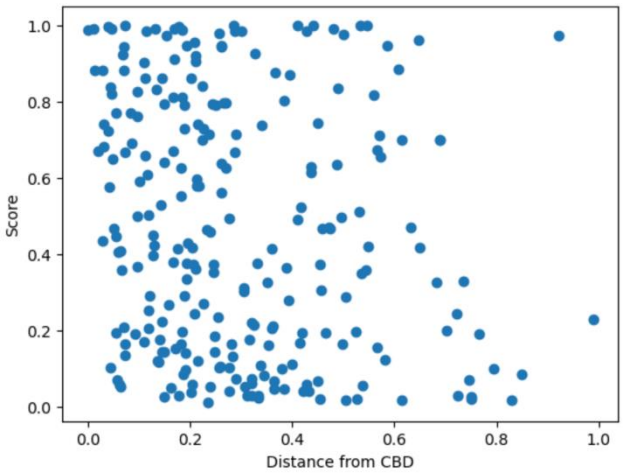


Figure 7. Original score vs. distance scatterplot.

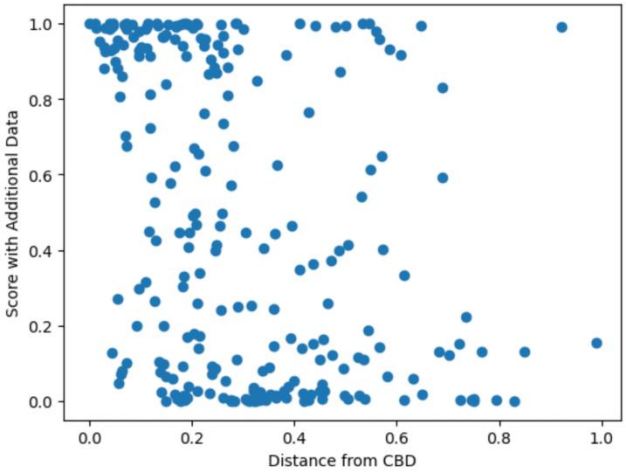


Figure 8. Score vs. distance after dataset addition.

Limitations: As seen Figures 7 and 8, it is difficult to deduce any correlation as the scatter plots do not indicate any score-distance relationship. The histograms of score frequencies in Figures 5 and 6 also show that many regions scored close to the extremities, which is unexpected as explained previously. These issues may mean that the qualities/data used in analysis do not accurately reflect how well-resourced a region is. In addition, the scoring method itself may be flawed as the scores in different aspects may not be completely independent from each other. For example, regions scoring high in occupation/education and retail are generally higher income, which may lead to fewer stops due to smaller demand in public transportation. If the details were to be further analysed, the use of different weighting in each z-score component may be useful, but it may be difficult to determine which aspects should have a heavier/lighter weighting.

4. Correlation Analysis

To evaluate how well the scores, as calculated previously, correlate with the median income of each SA2 region, the correlation coefficient was calculated. The correlation coefficient was used to reflect the statistical measure of the strength of the relationship between the two variables. The value was calculated to be 0.5495, which indicates a moderate positive relationship, and thus that the score increases with median income. The result is not surprising as areas that are well-resourced are also more desirable to live in. If the demand to live in the region is high, the aspect of property valuation would be higher, leading to a higher concentration of high-income people living within the area.

However, correlation may not imply causation, hence it is difficult to determine whether the score is dependent on the median income. Furthermore, the standard of a well-resourced neighbourhood differs for people with different incomes. As mentioned earlier, people in a lower-income bracket generally want regions with many transport stops and closer access to schools, while people in higher-income brackets would consider these aspects of less value. In addition, since the issue is social/geographical, deviations from expectation are commonplace due to variations in personal preferences. These may be possible explanations for a moderate correlation coefficient value.

5. Additional Analysis

5.1 Rank Scoring and Analysis

Below is the formula used to calculate scores for each SA2 region based on rank:

Score = mean(standardized $score_{retail}$ + standardized $score_{health}$ + standardized $score_{stops}$ + standardized $score_{polls}$ + standardized $score_{schools}$) where:

- The regions are ranked using the same criteria from which z-scores were calculated
- Rank-scores were standardised using the method:
  standardised rank score = 1 - (rank - minimum rank)/(maximum rank - minimum rank), giving a scoring scale of 0 to 1, with 1 reflecting the ‘best’ regions and 0 reflecting the ‘worst’
- The maximum ranks (in terms of integer value) are lowest and minimum ranks are highest, hence minimum rank ≡ 1 and thus: standardized score = 1 - (rank - 1)/(lowest rank - 1)
- The mean of each standardized rank-score was taken to return the overall score for each region, as each factor (retail, health, stops, polls, and schools) was taken to have equal weighting in their contribution to the well-resourced-ness of a region
- Density ranking was used here as logically, regions with the same value in a certain aspect should receive the same rank, and those following are placed into the rank below

	sa2_code	score	stdrank_retail	stdrank_health	stdrank_stops	stdrank_polls	stdrank_schools
0	102011028	0.365890	0.532520	0.768293	0.346821	0.181818	0.000000
1	102011029	0.310305	0.280488	0.418699	0.670520	0.181818	0.000000
2	102011030	0.760841	0.934959	0.735772	0.497110	0.636364	1.000000
3	102011031	0.642664	0.902439	0.898374	0.867052	0.545455	0.000000
4	102011032	0.660987	0.796748	0.906504	0.965318	0.636364	0.000000
...	...	...	...	...	...	...	...
242	128021536	0.390532	0.223577	0.300813	0.791908	0.636364	0.000000
243	128021538	0.562719	0.439024	0.650407	0.595376	0.545455	0.583333
244	128021607	0.364375	0.170732	0.321138	0.693642	0.636364	0.000000
245	128021608	0.171417	0.182927	0.093496	0.398844	0.181818	0.000000
246	128021609	0.081813	0.020325	0.280488	0.017341	0.090909	0.000000

Figure 9. Overall score and standardized rank-scores for each SA2 region using original datasets.

An interactive Choropleth map was implemented for the rank-scores (Fig. 10), with darker red colours reflecting higher scoring regions and lighter yellow colours reflecting the lower ones. In contrast to the z-score map, there is a notable clustering of high-scoring regions near the centre and to the north of the City of Sydney, while regions in the south/west and on the edges of Greater Sydney are generally lower-scoring.

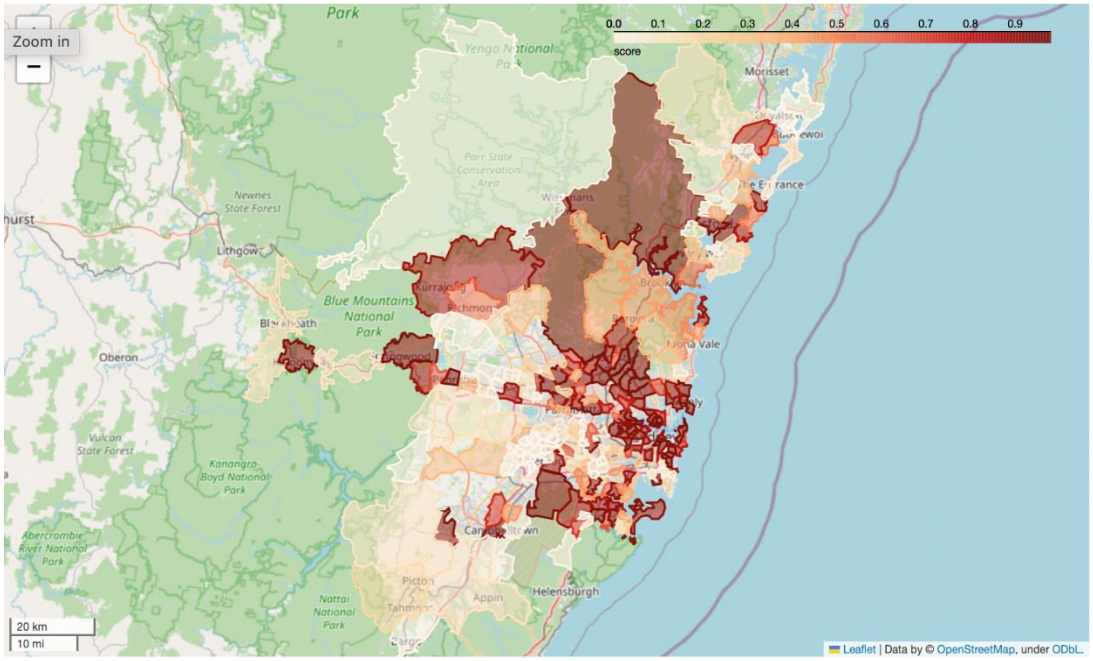
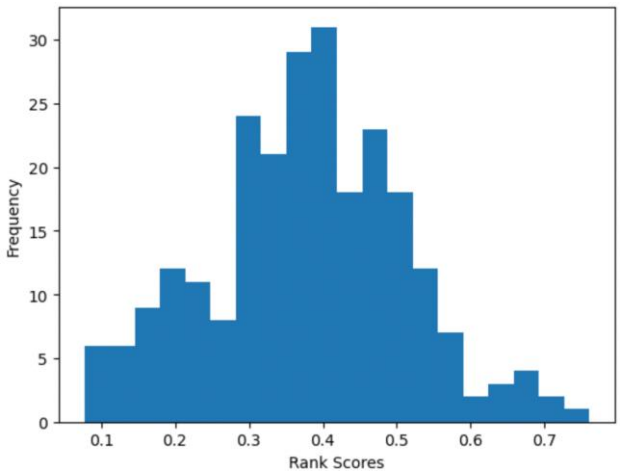


Figure 10. Interactive Choropleth map showing scores for each SA2 region (using rank).

The scores of the regions follow an approximate normal distribution (Fig. 11), which is expected due to the Central Limit Theorem as explained previously. The scatterplot of the scores (Fig. 12) follows a weak negative trend when plotted against distance from the CBD, which is also to be expected as the CBD is defined as the Central Business District, the heart of Sydney's businesses and retail, hence regions closer to the CBD generally have better/more facilities while those further away are often neglected. When compared to the lack of score-distance relation and the unexpected shape of the histogram from the z-scores, using rankings seems to be more effective in reflecting liveability.



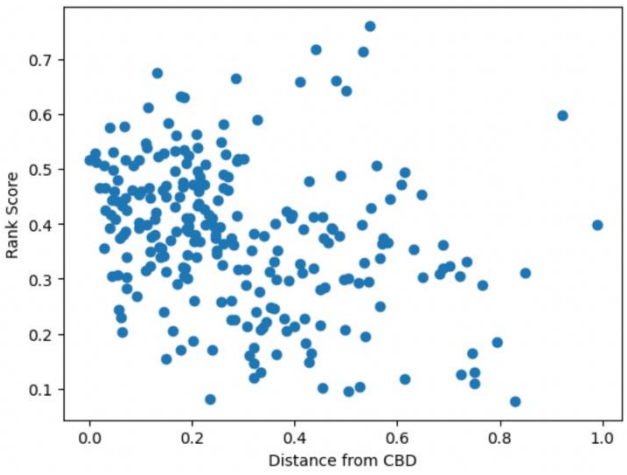


Figure 11. Score frequency histogram with rank scoring. Figure 12. Score vs. distance with rank scoring.

Limitations: Although in this particular analysis, the rank method produced more reasonable results than the z-score method, it is often more intuitive to use z-scores rather than rank. For example, the difference in number of bus stops between the first rank and the second may be vastly different than that between the second and third, though the rank difference is 1 in both cases. The unexpected

effectiveness of the rank system in comparison with the z-score method may be due to the limited number of data/regions used, hence not fully demonstrating the true usefulness of each method.

5.2 Median Income Prediction

A multiple linear regression was built to predict the median incomes of SA2 regions with population 100+. There are three explanatory variables including z-scores of the average number of stops per km^2, the total number of health services per 1000 people and the socio-economic advantage index.

	sa2_code	z_stops_km2	z_health	z_index_rsd	median_income
0	102011028	-0.374030	0.441162	0.762490	52450
1	102011029	-1.306543	-0.486794	0.599861	48724
2	102011030	-1.770296	0.293690	0.198710	46228

Figure 13. The first three observations in the dataframe.

We used train_test_plit in the Scikit-Learn library, and the model_selection module to divide the dataset into training and testing sets with a ratio of 7:3. The model was created using linear_model.LinearRegression(). It trained on the training data and predicted on the testing data.

The final model and evaluation are as below:

median_income = 56502.85 + 2229.06*z_stops_km2 + 1378.41*z_health + 6807.63*z_index_rsd

- R squared: 0.55473
- Mean squared error: 39714894.38
- Mean absolute error: 5080.12

Over 55% of the median income variance can be explained by the model features. The mean squared error and mean absolute error were recorded relatively high. Therefore, the model needs further improvements with more relevant independent variables. Though, with this model, all features are statistically significant since their p-values are smaller than the level of significance of 0.05:

- z_stops_km2: 1.856228e-04
- z_health: 1.779959e-03
- z_index_rsd: 1.376950e-37

Hence, we can conclude that the number of stops, health services and the index of socio-economic advantage can be positively correlated with the median income of an SA2 region.

REFERENCES

Australian Bureau of Statistics. (2022). *Counts of Australian Businesses, including Entries and Exits* [Data set]. https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads

Australian Bureau of Statistics. (2020). *SA2 digital boundaries* [Data set]. https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files/SA2_2021_AUST_SHP_GDA2020.zip

Australian Bureau of Statistics. (2021). *Socio-Economic Indexes for Areas (SEIFA), Australia* [Data set]. https://www.abs.gov.au/statistics/people/people-and-communities/socio-economic-indexes-areas-seifa-australia/2021#data-downloads

Australian Government Department of Health and Aged Care. (n.d.). *Australian Aged Care Services* [Data set]. https://www.nationalmap.gov.au/

Australian Electoral Commission. (2019). *AEC - Federal Election - Polling Places (Point) 2019* [Data set]. https://data.aurin.org.au/dataset/au-govt-aec-aec-federal-election-polling-places-2019-na

NSW Department of Education. (2023). *School intake zones (catchment areas) for NSW government schools* [Data set]. https://data.cese.nsw.gov.au/data/dataset/school-intake-zones-catchment-areas-for-nsw-government-schools

Transport for NSW. (2022). *Timetables Complete GTFS* [Data set]. https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs