DATA2001 Assignment Report

Jacob Whiteford (520351554), Carlos Gonzalez (530239167)

Dataset Description

SA2 Regions (**SA2_2021_AUSTR_GDA2020.shp**) - Contained information on statistical area digital boundaries (SA2, SA3, SA4). For the purpose of our analysis, only SA2 data was considered, and only for the Greater Sydney Region (done by filtering variable GCC_NAME21 for Greater Sydney). Spatial data, in which we changed geom type to multipolygon.

Businesses.csv - Contains information on the number of business in different SA2 regions with counts based on the value of the business (e.g. 0_to_50k_businesses). Data provided by the Australian Bureau of Statistics. This dataset was separated into Health and Retail businesses for analysis.

Population.csv - Contained information on population statistics for total people and for every age group living in each SA2 area. SA2 code and names included in the dataset. Filtered for analysis; only total population statistics and age groups between 0-19 were necessary for this project.

SchoolCatchments.zip (**catchments_future.shp**) - Contained information on future catchment areas (school intake zones) for NSW Government schools. Data provided by the NSW Department of Education. Spatial data with shapefile, in which we changed geom type to multipolygon. Contains information on catchment type, school names and date added to specific catchment area.

SchoolCatchments.zip (**catchments_primary.shp**) - Contained information on catchment areas (school intake zones) for NSW Government Secondary Schools. Data provided by the NSW Department of Education. Spatial data with shapefile, in which we changed geom type to multipolygon. Contains information on catchment type, school names and date added to specific catchment area.

SchoolCatchments.zip (**catchments_secondary.shp**) - Contained information on catchment areas (school intake zones) for NSW Government Primary Schools. Data provided by the NSW Department of Education. Spatial data with shapefile, in which we changed geom type to multipolygon. Contains information on catchment type, school names and date added to specific catchment area.

Crime.zip (**crime/AlcoholRelatedAssault_JanToDec2021.shp"**) - Shape file (spatial data) which contained information on Alcohol related assault cases in NSW. Data was provided by NSW Bureau of Crime Statistics and Research (BOCSAR). Geom type was changed to multipolygon. 4 main columns; ‘OBJECTID’, ‘Contour Density’, ‘ORIG_FID’, ‘geom’.

Fire.php.rss - Is a GeoRSS format file sourced from New South Wales Fire and Rescue[1]. Formatted as an XML file, the dataset contained all Fire and Rescue NSW Fire Station locations

[1] https://www.fire.nsw.gov.au/feeds/stations-georss.php

in latitude-longitude format. The coordinates were then converted into point type geometry for analysis. The dataset contained 5 main columns: ‘pubDate’, ‘title’, ‘description’, ‘link’, ‘geo:lat’, ‘geo:long’.

PollingPlaces2019.csv - Contained data on the AEC - 2019 Federal Election Polling Places. Data provided by the Australian Electoral Commission. The file contained coordinates for polling locations along with some geometry data that needed WTK transformations.

All data was integrated using the Pandas Data frame .to_sql() function, after defining appropriate schemas for each table/Pandas Data Frame.

Database Description

After cleaning the data, and dropping attributes that would not be used for analysis, the database consisted of 11 tables. The database diagram is shown below with each attribute’s corresponding data type, table’s primary and foreign key where necessary. The data was integrated using the Pandas Data frame .to_sql() function, after defining appropriate schemas for each table/Pandas Data Frame.

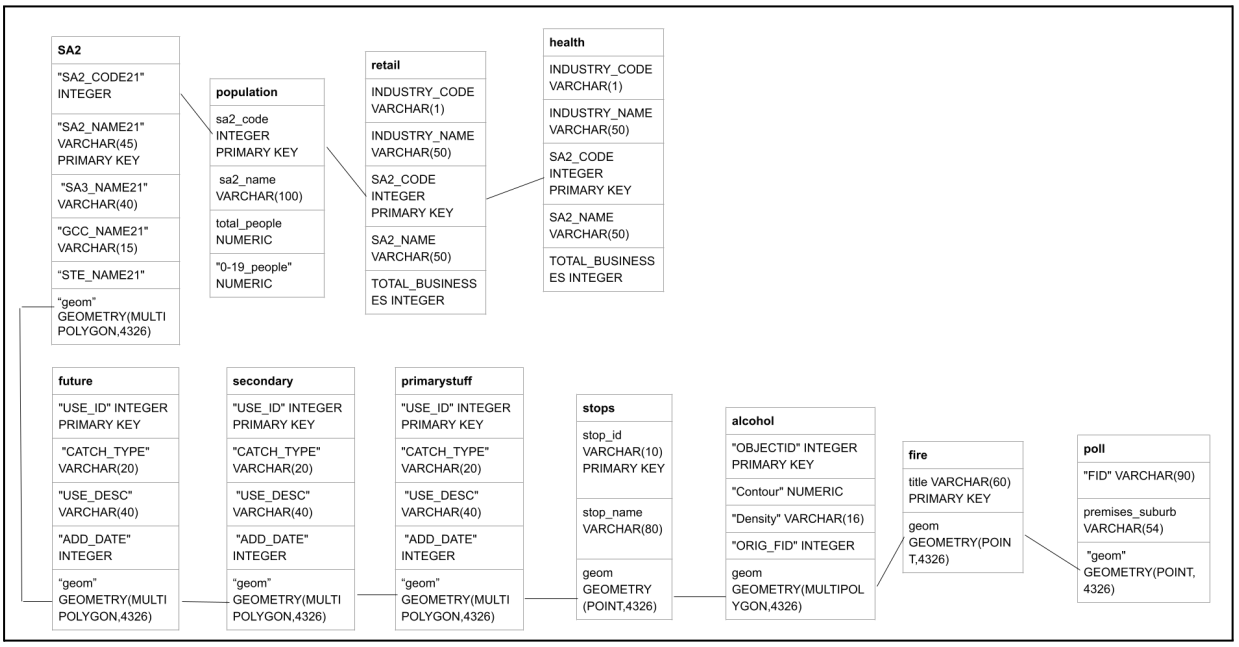

Figure 1: Database Diagram

In terms of table indexes, we used GiST (Generalized Search Tree) which is supported out of the box by PostgreSQL. Particularly, we used this index on the ‘geom’ attributed for the stops, primarystuff, future, and secondary tables. The raw stops.txt file had more than 100,00 observations, with testing we determined that GiST greatly improved queries with this table. We then decided to also utilize indexes for the other tables aforementioned, as they also had a considerable number of observations.

Results Analysis

This project aimed to calculate how “well-resourced” each SA2 region is based on the metrics detailed in Table 1, see appendix.

$$\textit{Score 1} = S(z_{retail} + z_{health} + z_{retail} + z_{poll} + z_{schools})$$

The score function is based on the sum of the normalized z-scores for each of the metrics below. In the formulae, S, represents the sigmoid.

The sigmoid function, is monotonically increasing, and maps an input *x* to an output ∈ [0,1], thus a higher input *x* will lead to a score closer to 1, which would be associated with a good neighbourhood. In this first scoring function, all the metrics included can be considered as positive attributes in a neighbourhood (e.g. more schools are better for the community), therefore all the z-scores contribute to a greater input *x,* and thus a better score, unless the metric is below the state average. The second iteration of the scoring function extended on this by adding two new metrics, Alcohol-Related Assaults and Fire Stations.

$$\textit{Score 2} = S(z_{retail} + z_{health} + z_{retail} + z_{poll} + z_{schools} + z_{fire\ stations} - z_{alcohol})$$

In this new scoring function, since a higher number of alcohol-related assaults in a neighbourhood is a negative attribute, instead of adding the z-score for alcohol, it is subtracted from the combined sum of z-scores such that the output of the sigmoid function (score) decreases, reflecting that a neighbourhood is not as “well-resourced”. The inverse should happen given below average number of assaults, as the score should improve. In terms of the number of fire stations, it was incorporated as usual to the scoring function, by adding the normalized z-score.

The score distributions for both scoring functions are presented below:

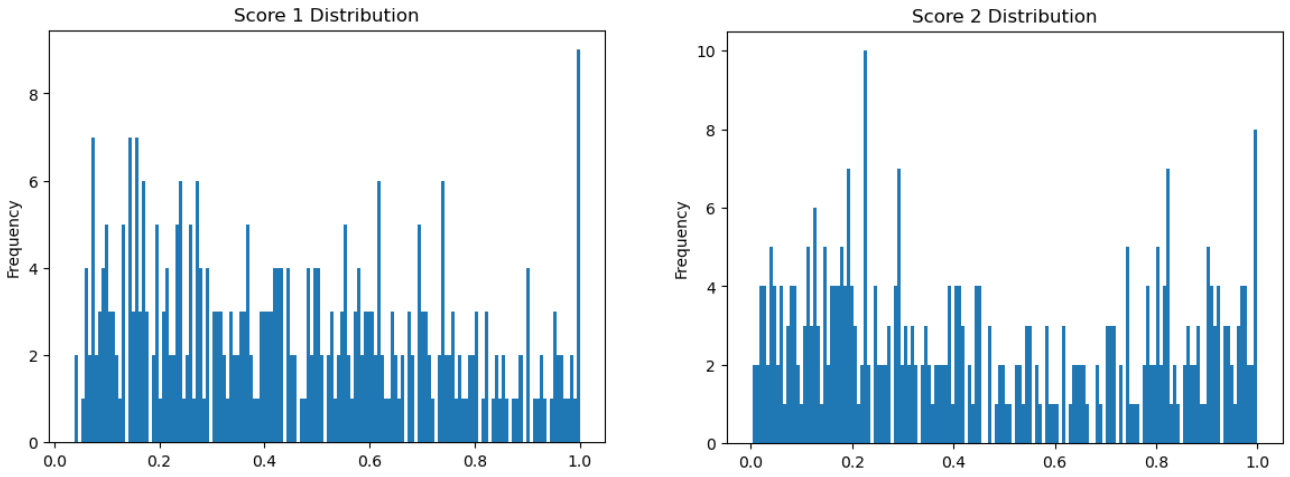

Figure 1: Score Distributions

Overall, for the first model we see that 154 of the SA2 regions in the model have a score >0.5 meaning that their aggregate metrics are above average. On the other hand, 206 regions had a score <0.5. Some regions of interest include Miller’s Point and Banksmeadow, who had a particularly high z-score sum of 46.87 and 20.45 respectively. This led to their scores being ~1. While Miller’s Point is one of Sydney’s most renowned and wealthy districts was unsurprising, Banksmeadow having a such a high Z-Score Sum/Score was surprising as an industrial area

east of the Sydney Airport. Upon closer inspection, we noticed that with a 0-19 population of 57, and 6 schools, the number of School Catchments Per 1000 Young People was an incredibly high 105.26, and thus 16.37 out of its Z-score sum of 20 can be attributed to only the schools, leading to a very high score.

On the other hand, with the second iteration, displaced 2 SA2 Regions to a score <0.5. The top 5 SA2 Regions are Sydney (North) - Millers Point (1.00), Banksmeadow (1.00), Blackheath - Megalong Valley (0.99), Sydney (South) - Haymarket(0.99), Dural - Kenthurst - Wisemans Ferry(0.99). Except for the 5th region which replaced Chatswood East (0.99) in the first iteration, despite expected differences in the Z-Score Sum, which can be partially attributed to the direction of the Alcohol metric, the top regions remained the same.

The most noticeable difference between the two models was how the bottom-ranked regions changed. With the first scoring function the bottom regions were Jordan Springs - Llandilo (0.0621), Edmondson Park (0.059), Kariong (0.0510), Summerland Point - Gwandalan (0.0393), Spring Farm (0.0379). With the second score the bottom 5 cities are now Spring Farm (0.016), Colyton - Oxley Park (0.010) Regents Park (0.009) Canterbury - South (0.007), Summerland Point - Gwandalan (0.002). Overall, it appears that these bottom regions are small suburbs with very small populations and located quite significantly far from the city centre, they are expected to have less public transportation stops, school, and businesses in the area because of this.

It also appears that some regions with a very large extension and a low population, particularly the SA2 regions around the Blue Mountains regions have received high scores. Both scoring functions did not involve considerations for region size, as these regions are partially determined by number of residents and their interactions. However, similar to the Banksmeadow case, SA2 areas such as “Dural - Kenthurst - Wisemans Ferry” have a total of 35 schools, in an area of ~355 km2, while “South Coogee” has 5 schools. Hence, a metric such as schools which is mostly determined by location, to improve accessibility to as many people as possible, might not be ideal by itself.

Correlation Analysis

Statistical analysis was undertaken to test the original score and additional score against the median income in available SA2 regions. Two correlation test types were conducted in which we calculated the correlation coefficient between the computed scores and median income, as well as the utilisation of linear regression analysis to further test for the relationship between the two variables.

The original score returned a correlation coefficient of 0.19680628607428152, suggesting a weakly positive linear relationship between the Score and Median income variables. This means that as the computed sigmoid score for the original function increases, there is a slight tendency for median income to increase congruently, albeit weakly.

Regression analysis using a linear regression model based off of the original score and median income provided us with a coefficient of 5.42010408e-06, an intercept of 0.1963449363371459, and a p-value of 6.89482458e-84. The coefficient suggests that a one-unit increase in median income is associated with a 5.42010408e-06 unit increase in the Score. The intercept represents the expected value for the score if the median income is 0. Interpreting the model, the positive coefficient suggests that even though the effect of median income on the Score is small, there still exists a positive linear relationship between the 2 variables. In addition to this, the p-value

indicates that there is a significant relationship between median income and the score, as it is well below the significance level of 0.05.

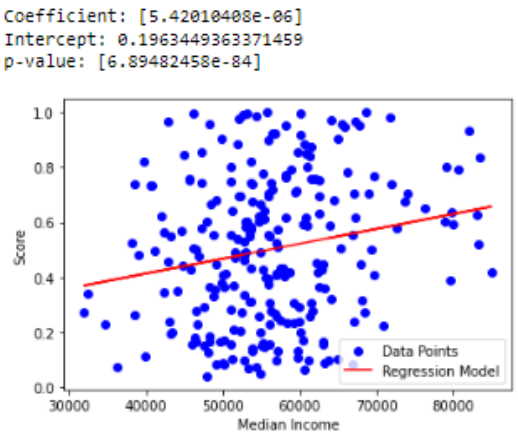

Figure 2: Linear regression model: Original Score with Median Income

The updated score with the inclusion of the datasets based on fire station and alcohol related assaults returned a correlation coefficient of 0.23254239831215667, suggesting a slightly stronger positive correlation with median income when compared to the previous score. Again, this means that as the sigmoid score for the updated function increases, there is a slight tendency for median income to increase.

Regression analysis using a linear regression model based off of the updated score and median income provided us with a coefficient of 7.57745042e-06, an intercept of 0.05768569021161707, and a p-value of 2.24008481e-69. The coefficient indicates that a one-unit increase in median income is associated with a 7.57745042e-06 unit increase in the Score, suggesting that the relationship between median income and the updated score is slightly stronger than our previous model. The intercept is also significantly smaller, implying that the expected score of median income is 0 is lower when compared to the previous model. In addition to this, the p-value indicates that there is still a significant relationship between median income and the updated score, as it is well below the significance level of 0.05.

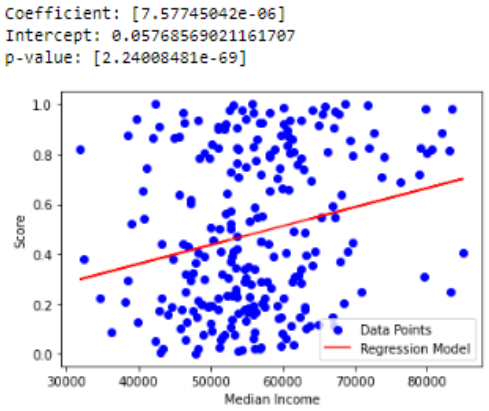

Figure 3: Linear regression model: Updated Score with Median Income

With reference to the results from the computed correlation coefficients and regression analysis, it is clear that median income has a stronger relationship with our updated score. The updated model provides a more accurate estimation of the impact of median income on the score, as it has a lower intercept yet higher coefficient.

Data Visualisations

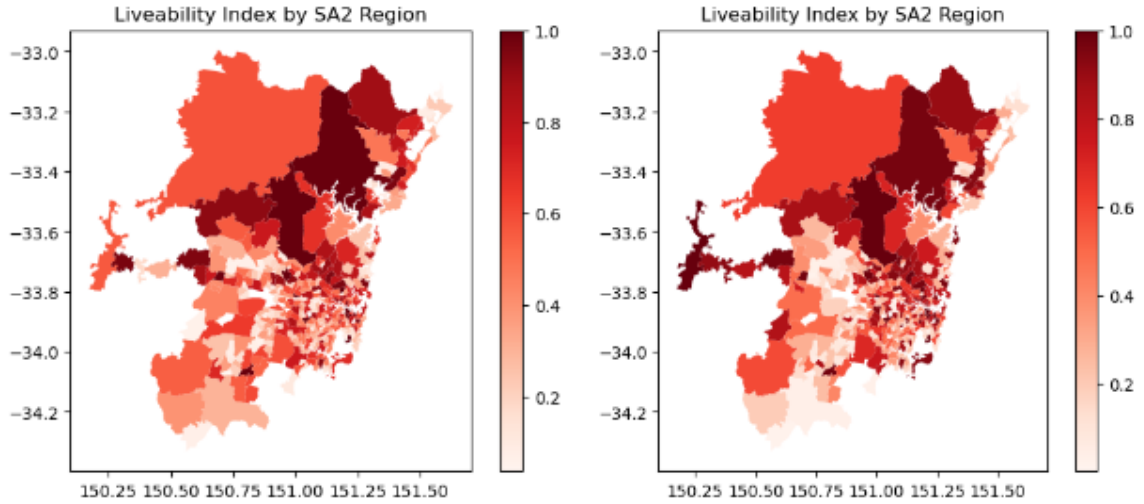

Figure 4: Heat Map of Liveability Index by SA2 Region. Original Score (left), Updated Score (right)

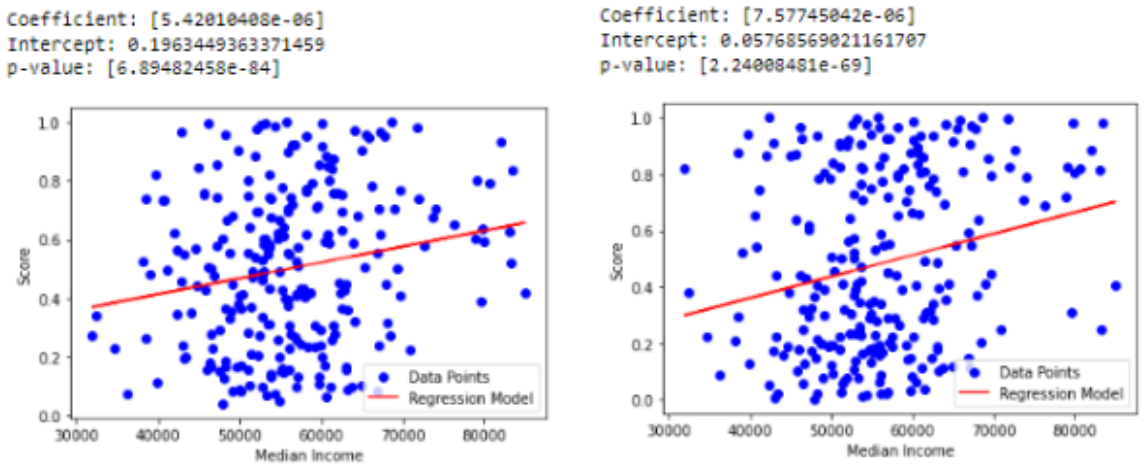

Figure 5: Linear Regression Models (Score vs Median Income). Original Score (left), updated score (right)

Appendix

Metric	Definition	Table/Dataset
Alcohol Related Assaults	Alcohol Related Assaults per 1000 residents	alcohol
Fire Stations	Number of Fire Stations per 1000 residents	fire
Health	Health-related Businesses per 1000 residents	health
Retail	Retail Businesses per 1000 residents	retail
Stops	Number of Public transportation Stops	stops
Poll	Number of Federal election polling locations	poll
Schools	Number of Schools per 1000 young residents (0-19 years old)	futurestuff, primary, secondary

Table 1: Score Metrics

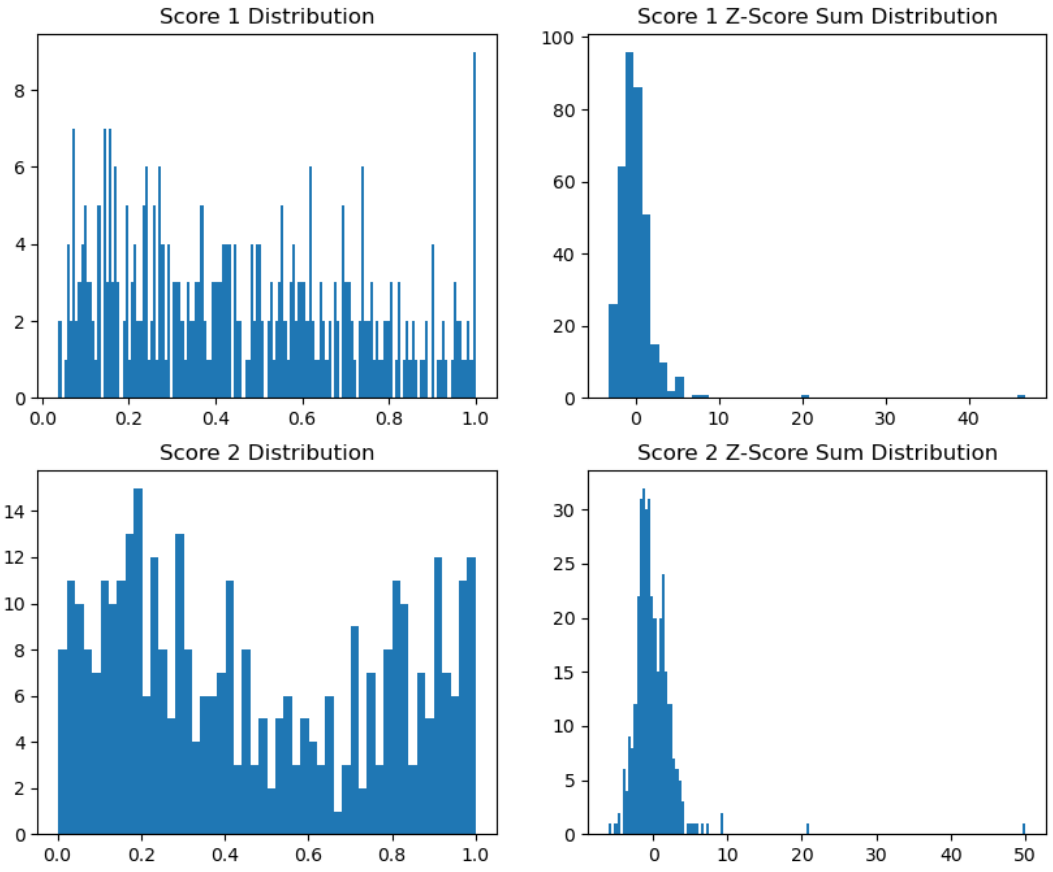

Figure 3: Distribution Graphs