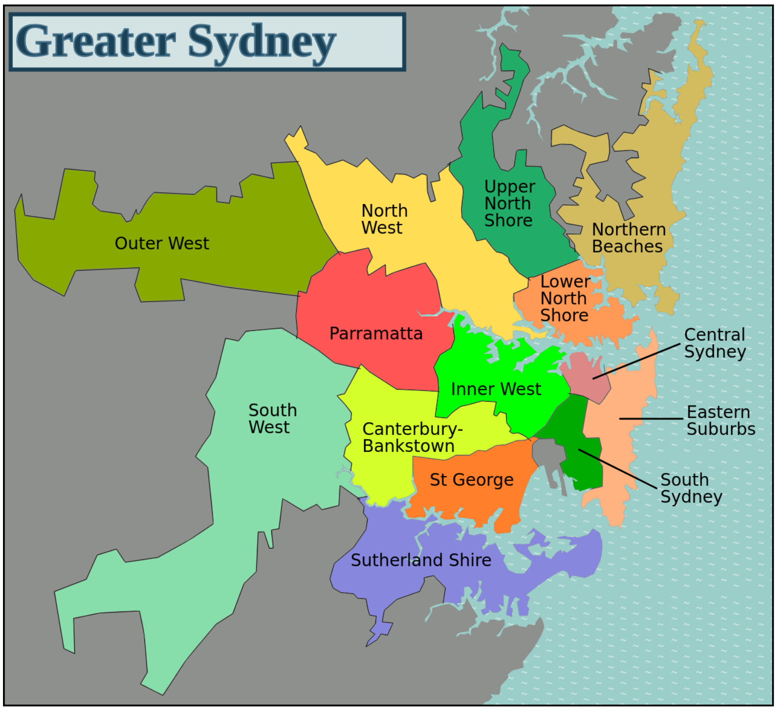
DATA2901 Assignment 1

Greater Sydney Analysis: Group Project Report



Preran Apinakoppa, Luke Brutto, Linh Ngo

May 2023

1 Introduction

As part of the group assignment, members were assigned to collecting and handling geo-spatial data from various government data-hubs from the GOV. of NSW and wide range of publicly available datasets. This assignment assesses and reviews their skills in handling data using SQL and python programming paradigms.

2 Data

2.1 Dataset Description

There are in total 10 datasets utilized for this project, with the first 7 datasets given from the specifications and the last 3 additional datasets incorporated from online sources. The description of each dataset is as follows:

1. **SA2 Regions:** Geographcal datasets about Statistical Area Level 2 (SA2) digital boundaries, which is sourced from Australian Bureau of Statistics.
2. **Businesses:** Number of businesses by industry and SA2 region, reported by turnover size ranges, which is sourced from Australian Bureau of Statistics.
3. **Stops:** Locations of all public transport stops (train and bus) in General Transit Feed Specification (GTFS) format.
4. **Polls:** Locations (and other premises details) of polling places for the 2019 Federal election, which is sourced from Australian Electoral Commission.
5. **Schools:** Geographical regions in which students must live to attend primary, secondary and future Government schools, which is sourced from NSW Department of Education.
6. **Population:** Estimates of the number of people living in each SA2 by age range (for ”per capita” calculations), which is sourced from the specification packages.
7. **Income:** Total earnings statistics by SA2. This data set is sourced from the specification packages.
8. **Clubs and hotels:** Locations of all clubs and hotels in NSW, sourced from the NSW Department of Customer Service. Last updated March 2, 2021.
9. **Public toilets:** Locations of public and private-public toilet facilities across Australia, sourced from the Australian Department of Health. Last updated April 30, 2023.
10. **School locations and enrolment:** Master dataset containing comprehensive data for all public schools in NSW including local electorate information and school enrolment numbers in JSON (JavaScript Object Notation) format, sourced from the NSW Department of Education. Data as of May 14, 2023.

2.2 Data Preprocessing

For datasets above, the following data preprocessing steps are applied:

- Drop unnecessary columns: To keep the database concise, only relevant columns for later analysis of each dataset are kept.
- Convert all the columns name of dataframes to lowercase: This step supports the creation of tables when using PostgreSQL
- Convert to geometric: Convert data to more useful spatial datatypes such as points found as floating point numbers to a paired tuple containing the latitude and longitude of point location. This makes it easier for SQL implementation and object draws for creating visual maps.
- Filter rows: As to clean the data of unnecessary data and to reduce time and allocated system resources toward processing and handling data as well to remove garbage entries as a preemptive measure of error handling.
- Transform dataframe from wide format to long format: This step is applied to the ”Population” dataset. Instead of having the number of people of all age groups stored as columns for each SA2 region, each row in the dataset is now corresponded to number of people in specific age group in a specific SA2 region. This step helps reduce the amount of columns required to store the population for each SA2 region and provides more convenient in analysis.

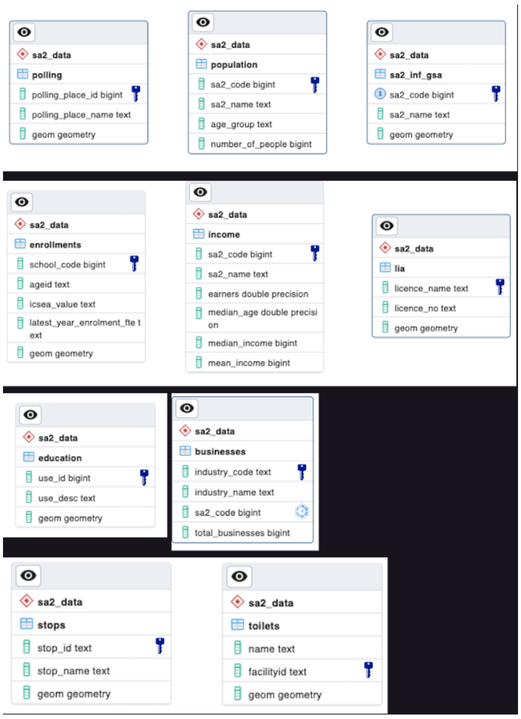


Figure 1: Schema diagram

3 Database Description

The schema "sa2_data" is created in PostgreSQL to store all the datasets mentioned above. For each dataset, primary key and applicable foreign key has been enforced to ensure data integrity and establish relationships between the tables in the database. For example, "sa2_code" columns in all tables are identified as foreign key that references to the "sa2_code" in the main SA2 regions dataset. Furthermore, spatial indexes are also created on geometric columns to enhance the performance when querying spatial data.

4 Results Analysis

4.1 Original Scoring Calculation

Well resourced refers to the extent which SA2 areas are facilitated with newer and usable infrastructure and sustainable living standards of the given population evident through census data. To model this, Z-scores were used as a measure of variance or how many standard-deviations away from the mean the data point sits. The sum of the Z-scores were passed to the sigmoid function as shown below. This transformation process comparative analysis through a much smaller scale from 0 to 1.

$$\text{Score} = \quad S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools})$$

Or more specifically:

$$S(x) = \frac{1}{1+e^x}$$

Figure 1 details the results of the initial data as a histogram.

4.2 Formula Extension Rationale

Aside from the normalised variables used to calculate the first iteration of the sigmoid score, we extended the score to include 3 additional datasets: Public Toilets, Enrollments and Clubs and Hotels. These data sets were chosen to grant the score a wider and more accurate view of 'well-resourced' SA2s. The metrics below were normalised and included the score calculation:

- **Public toilets per person:** A higher value implies that a given SA2 is able to provide additional resources and investment (such as water, building materials and toilet paper) for a safe, convenient public amenity.

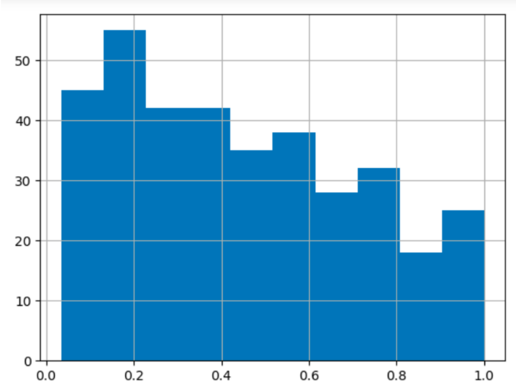


Figure 2: Initial Histogram of SA2 Scores.

- **Clubs/Hotels per person:** A higher value implies that a given SA2's residents have the disposable income in to spend on food and alcohol, and in addition represents the region's ability to supply these goods.
- **Educational Advantage Per Young Person:** Despite the number of school catchment areas already being included, this is an important metric to determine educational potential, an important factor regarding a region's income[1]. The advantage is measured using socioeconomic factors selected by the Australian Curriculum, Assessment and Reporting Authority. To account for multiple schools within areas, we weight each school according to its student population:

$$\# \, Full \, Time \, Enrollments \cdot School \, ISCEA \, Value \, / \, \# \, Young \, People$$

Hence, the formula calculation used was given by:

$$\text{Score} = S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools} + z_{clubs} + z_{toilets} + z_{EduAdv})$$

4.3 Implications of Extending the Dataset

However, extending the base formula came with a host of challenges. The first issue was that the data sets are not updated to the same day. Hence, there is the risk that the conclusions we draw do not reflect the current state of the Greater Sydney Region.

The second (and main) issue with extending the base formula was that, as all metrics are normalised, we had to ensure that they fit a normal curve. An investigation (summarised in appendix) into the distribution of these additional metrics revealed that the curves were only normal when the logarithm (we elected for base 10) was taken. However, this solution presented further issues - any region with 0 toilets or hotels/clubs or no schools could not have the logarithm performed. We found that 3.89% of SA2s had no public toilets, 15.3% had no clubs/hotels and 6.67% had no applicable schools.

The solution we came to was providing a penalty in place of the 0s. This penalty was equal to the lowest z-score calculated for that metric, as the region was at this level or more extreme.

4.4 Overall Score Distribution and Correlation Analysis

The overall distribution of the scores was observed using a map-overlay visualisation, histogram and summary table.

In addition to the shown data, the correlation between the extended sigmoid scores and the median income of a given SA2 was **0.28**. A correlation analysis was also attempted using a form of median income that was normalised and put into a sigmoid function, although this only returned a value of **0.26**. Both values highlight that there is a weak correlation between how 'well resourced' the nation is (according to our definition) and their median income total. This is an improvement from the previous scoring method (**0.22**). This seems to imply that at least one of the scores added (clubs, educational advantage and toilets) denote a higher income. This will be determined in the following section (additional analysis).

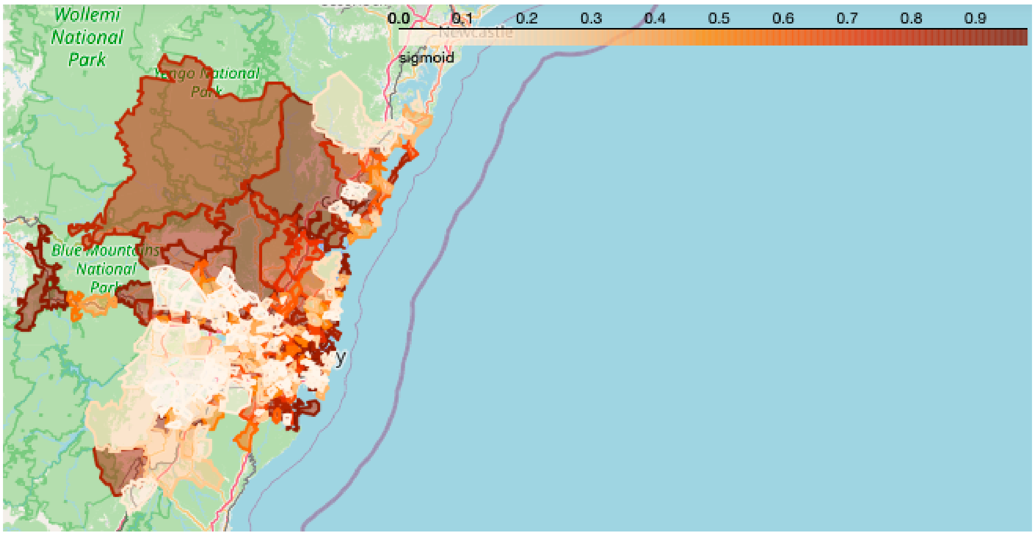


Figure 3: Map-overlay Visualisation of SA2 Scores (interactive version in notebook).

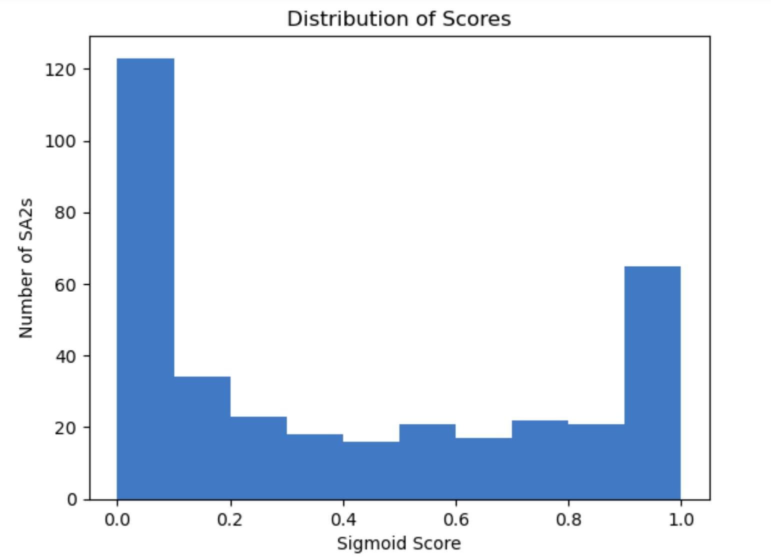


Figure 4: Histogram of SA2 Scores.

Table 1: Breakdown of SA2 scores

Score Bin	Percentage of SA2s in Bin (3.s.f)
(0.0,0.1]	34.2
(0.1,0.2]	9.44
(0.2,0.4]	11.4
(0.4,0.6]	10.3
(0.6,0.8]	10.8
(0.8,0.9]	5.83
(0.9,0.1]	18.1

However, the implications of this are improvement quite jarring. The noticeable increase in correlation implies that Figure 4 and Table 1 more accurately represent the state of resource equality in the Greater Sydney area than Figure 2, with over a third of SA2s in the lowest possible bin and less than 1 fifth in the highest. The histogram's shape also tells us that, in agreeance with the binned percentages, less than 50% of SA2s are within the (0.1,0.9] score range.

In addition, the map-overlay visualisation also tells us that the darker, higher-scoring regions of the map are concentrated mostly along the coast and further North, whilst the lower-scoring areas are mostly concentrated to the South and inland. As these areas are mostly in the same area, we may conclude that there are factors affecting the availability of resource. This is logical as high transportation costs and long distances from trading hubs (such as coastal areas with sea access) plague these regions. In addition, less resources also imply that these regions will have inadequate infrastructure, and as such have lower levels of education.

However, despite the startling results, the weak correlation does not allow us to be confident in our results, despite how they might provide insight into an Australian inequality issue.

4.5 Particular Trends, Regions and Scores of Interest

Despite the overall correlation being weak, we area able to supplement this with a short case study of the top and bottom 3 scoring SA2s from Table 2.

Table 2: Highest 3 and Lowest 3 Scoring SA2s

SA2 Name	Score
Sydney (North) - Millers Point	1
Banksmeadow	1
Parramatta - North	0.999939
Box Hill - Nelson	0.000016
South Wentworthville	0.000008
Castle Hill - West	0.000008

Millers Point and Banksmeadow are both connected to sea and contain a variety of stores and cultural spaces such as sculptures and theatre. Similarly, North Parramatta is near the CBD, and as such has a similar level of business activity. Given the concentration of people regularly visiting these areas, it is logical that their health, transport and toilet facilities would experience investment - bolstering their score.

On the other hand, the lower scoring SA2s are somewhat confusing. In the case of Box Hill (Nelson) and Castle Hill, it appears true that they are relatively remote compared to the higher-scoring areas. However, South Wentworthville is only a short drive away from the high-scoring Parramatta. Furthermore, new housing projects being put in place by the Hills Shire Council[2], imply that resources are more easily able to access the area than the score claims.

Hence, we can conclude from the score and correlation analysis that, whilst the scoring formula allows insights into Sydney's inequality that appear to be somewhat correct, the short case study and weak nature of the correlation prevent the investigation from being firm evidence to this end.

5 Additional Analysis

5.1 Rank Scoring System

The ranking system in place of the z-score system, takes in the aggregate data whereas the score-system leaves out any outliers/noise. Outliers may have a large affect on how skewed the data is and could allow for investigation on why they do not conform to the natural trend of the dataset.

As opposed to the sigmoid summation function, the individual metrics of each SA2 area will be ranked before having a function applied to potentially provide a more accurate representation of the score distribution. The function selected is the 'rank product', which is often used in biology. It is calculated by taking the geometric mean of all ranks, in this case 8 values were considered:

$$\text{Rank Score} = \left(r_{retail} \cdot r_{health} \cdot r_{stops} \cdot r_{polls} \cdot r_{schools} \cdot r_{clubs} \cdot r_{toilets} \cdot r_{EduAdv}\right)^{1/8}$$

As seen in Figure 5, the data has been summarised into a histogram. The Rank Product Score histogram resembles that of a normal distribution termed "bell-shaped" and portrays that most SA2 regions are of similar progress (at a score of 167). This means they share the similar level of development in the aforementioned areas of infrastructure, education levels and/or school retention rates, and other co-factors such as health and available readily accessible public goods.

The accuracy of the rank score in representing how 'well resourced' an SA2 is can also be determined through a correlation test with median income. Note that in this test, the negative of the rank score was used, as a lower rank means a higher level of development. A result of 0.30848 shows that, whilst still a weak correlation, it is a more accurate predictor than the sigmoid summation scoring system.

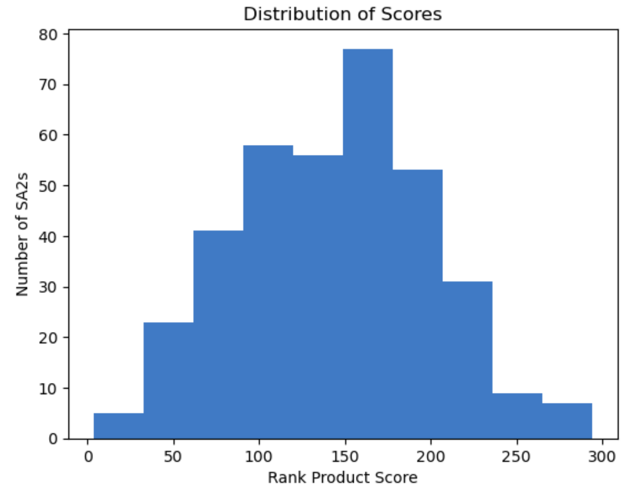


Figure 5: Histogram of the Rank Product Score (RPS)

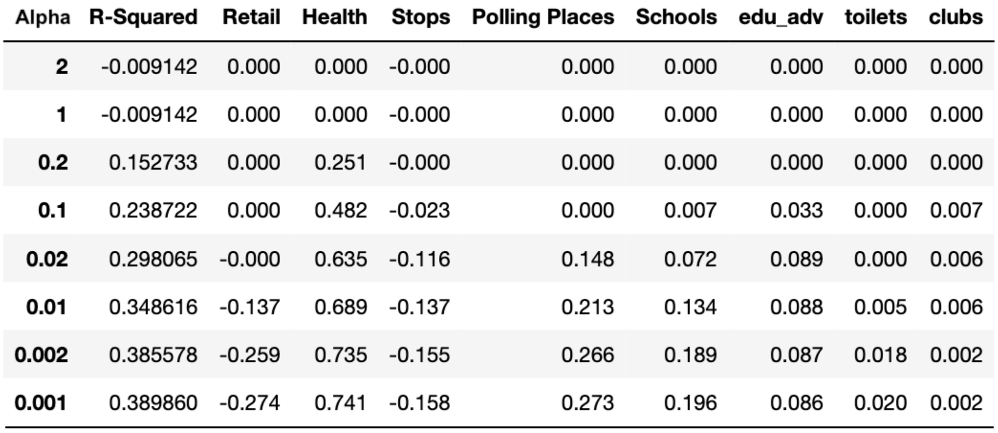
5.2 Machine Learning Analysis - LASSO

Despite the correlation between the median income and sigmoid score being quite low **(0.28)**, it inspired us to question whether considering individual metrics may offer a more accurate model.

The Least Absolute Shrinkage and Selection Operator (LASSO) regression method is a linear modelling technique favouring aggressively shrinking the model down to the most important predictor variables. Its 'L1 regulariser' penalises a given model equal to the absolute magnitude of its coefficients. Unlike other closely related techniques (e.g. ridge regression), it has the capability to set a coefficient to absolute 0, effectively removing variables it deems 'unnecessary'. LASSO's harsh regularisation hence allows us to determine which variables have the greatest impact on median income levels, and potentially create a model which is easier to deploy given the reduced variable count.

The investigation was conducted with a 90-10 split of the data into training and testing respectively. In addition, the median income was normalised to allow for more effective feature selection. With 247 observations in total, the sample size should be sufficient to create the model.

Table 3: Results of the Machine Learning Investigation



Alpha	R-Squared	Retail	Health	Stops	Polling Places	Schools	edu_adv	toilets	clubs
2	-0.009142	0.000	0.000	-0.000	0.000	0.000	0.000	0.000	0.000
1	-0.009142	0.000	0.000	-0.000	0.000	0.000	0.000	0.000	0.000
0.2	0.152733	0.000	0.251	-0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.238722	0.000	0.482	-0.023	0.000	0.007	0.033	0.000	0.007
0.02	0.298065	-0.000	0.635	-0.116	0.148	0.072	0.089	0.000	0.006
0.01	0.348616	-0.137	0.689	-0.137	0.213	0.134	0.088	0.005	0.006
0.002	0.385578	-0.259	0.735	-0.155	0.266	0.189	0.087	0.018	0.002
0.001	0.389860	-0.274	0.741	-0.158	0.273	0.196	0.086	0.020	0.002

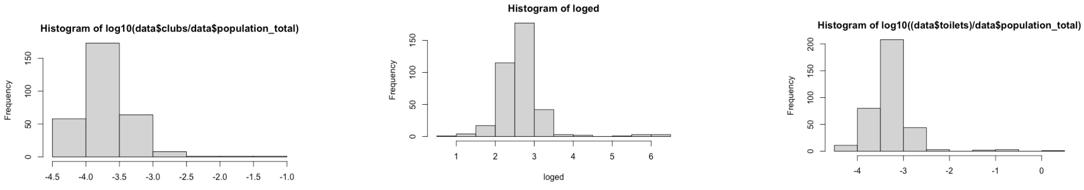
As per Table 3, as **Alpha** (hyper parameter denoting penalty magnitude) increases, the value of all coefficients moves towards zero. When alpha reaches 0.02, Retail and Toilets are the first coefficients to be reduced to zero. This suggests that they are the least useful in the model to predict the median income of a given SA2. At 0.1, the Polling Places coefficient is reduced to 0, then at 0.2 all except for health are reduced to 0. Past this point, all coefficients were reduced to 0. Hence, the LASSO regression suggests that the 'Health' metric is the most valuable out of those investigated to predict income. This makes sense as one's health is one of the most important determining factors in the human wellbeing (even being a core component in the Human Development Index (HDI) ranking of a nation) and as such individuals with higher income might be more willing to invest in their health. Furthermore, at alpha = 0.1, educational advantage and stops are also coefficient with a high value relative to others remaining (clubs and schools) bar health. Educational advantage being important is logical as parents that have achieved higher levels of education will have access to a wider variety of jobs, as "Higher educational attainment leads to higher total incomes, more diverse sources of income and reduced reliance on the aged pension" [1]. The number of stops having a negative correlation also makes sense, as residents of lower income SA2s are less likely to own personal vehicles due to their cost. As such, they are more reliant on the public transport stops the government provides, increasing their prominence. Assuming the opposite is true in richer areas, the negative correlation is justified.

However, whilst some variables are more important than others in predicting income levels, they all play a part in increasing the accuracy of the model. This is shown as the R-Squared value (measuring goodness of fit) decreases as the regularisation becomes harsher, denoting a poorer predictive model. As such, we may conclude that, whilst health, stops and educational advantage are the most important variables in determining the median income level of an economy out of the 8 factors studied, they would be largely ineffective in predicting income levels on their own. Even all together, the R-Squared value barely reaches 0.4, showcasing a weak correlation, the same result as the original sigmoid correlation.

Conclusively, whilst Stops, Health and Educational Advantage are all important indicators in predicting the median income level of a given SA2, all metrics act to increase the model's effectiveness. However, even with all predictors, the model would still be largely ineffective given its low R-Squared value.

Appendices

1 Histograms From RStudio



(a) Clubs (b) Educational Advantage (c) Toilets

Figure 6: Histograms of Logged Metrics. R ignores log(0), and a penalty was applied to these measurements. RStudio was used.

References

[1] Australian Government Department of Education. Income. https://www.education.gov.au/integrated-data-research/benefits-educational-attainment/income, 2022.

[2] ArchitectureAU Editorial. High-density office and housing precincts proposed for north-west sydney. https://architectureau.com/articles/high-density-office-and-housing-precincts-proposed-for-north-west-sydney/, 2023.