Sydney Livability Analysis

Author: Junrou Chen, Yang Zhou, Aoni Fu

September 14, 2022

1 Introduction

With the advanced urbanisation of Greater Sydney in recent years, the livability scores of the areas have arisen increasingly more interest and play more significant roles in residents' daily life, especially for those who want to settle down and buy properties in Sydney. In this project, the aim is to help our stakeholders who are young couples with children to choose the most livable area in Sydney for living. To measure the livability, a score including several factors (i.e., school, accommodation, retail service, crime, and health service) was calculated. The score was further modified to better reflect our stakeholders' needs.

2 Data Ingestion

2.1 Dataset description and cleaning

Datasets used in this study were collected from different sources including *Australian Bureau of Statistics (ABS)*, *City of Sydney Open Data Hub*, and *domain.com.au* . Detailed dataset description is provided in a seperate file.

To begin with, all null values and duplicated values were checked and entries containing missing values were dropped for all datasets except for three school catchment zone tables. Although missing values in school-related tables were observed in the *add_date* columns, they were not dropped since these columns were not used in this study.

All column names were changed to lowercase for consistency purposes and polygon geometry information was converted to multipolygon type. In neigbourhood table, commas between numbers were removed and the geometry information in Industry_Occupation table was flatten from 3D points into 2D points.

For datasets related to school catchment zones, all three datasets (i.e., primary, secondary and future) were concatenated into one dataframe. A new column 'is_future' is created to indicate whether the zone is still in the planning stage. In the combined school table, duplicate IDs were observed, indicating some schools have both primary and secondary catchment zones. A new column to indicate the school ID instead of using the 'use_id' was created such that those schools are treated as different schools in the rest of this study.

To reduce the size of database, only entries related to Greater Sydney region were selected from SA2, neighbourhood and business tables before uploading into database.

2.2 Database schema description

Figure 1 illustrates the relationship of tables in the database system used in this study, where a key icon and arrow (pointing to the referenced table) indicates a primary and foreign key, respectively.

The columns *sa2_main16* and geometry information of several tables are highly used in this study during joining and spatial joining processes. To speed up SQL querying process, indices were created as follows:

1. sa2_geom_idx on the geom column and sa2_main16_idx on the sa2_main16 column were created in SA2 table.

2. neighbourhood_area_idx on the area_id column was created in neighbourhood table.

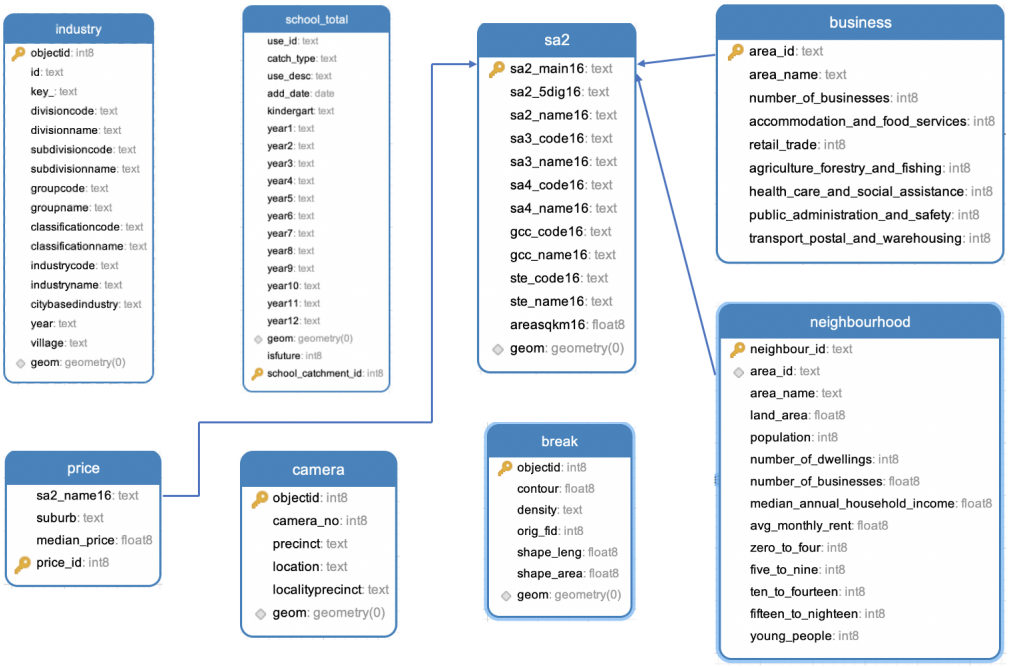

Figure 1: Schema of datasets (golden key and arrow indicates primary and foreign key)

3. business_area_idx on the area_id column was created in business table.
4. school_geom_idx on the geom column was created in school_total table.
5. industry_geom_idx on industry_ geom was created in the industry table.

No additional index on geometry information was created for camera and price tables since they only contain small amount of entries were not frequently queried.

3 Livable Score Analysis

In this study, a livable score (LS) that characterising the living quality of an SA2 area is defined as follows:

$$LS = S(Z_{school} + Z_{accomm} + Z_{retail} - Z_{crime} + Z_{health}) \quad (1)$$

where S is the Sigmoid function and the definition of each factor is described in Table 1. It is worthwhile to mention that all values calculated are further standardized by Z-score (i.e., the difference between input and mean value divided by standard deviation).

Category	Weight(%)	Measure	Definition
Education	20	School	number of schools catchment areas per 1000 young people (age<20)
Diversity	40	Accomm Retail	number of accommodation and food services per 1000 people number of retail services per 1000 people
Safety	20	Crime	sum of hotspot areas divided by total area
Health	20	Health	number of health services per 1000 people

Figure 2 illustrates the livable score of SA2 areas within Greater Sydney region. The colour bar ranging from 0 to 1 represents the degree of livability score, where yellow and purple color indicates a high and low score, respectively. An interactive map can be found in the Jupyter notebook attached separately as well.

The average score of Greater Sydney region was calculated at 0.45. It was found that *Sydney - Haymarket - The Rocks regio* exhibited the highest score of 1.0 while *Bidwill - Hebersham - Emerton* region showed the lowest score.

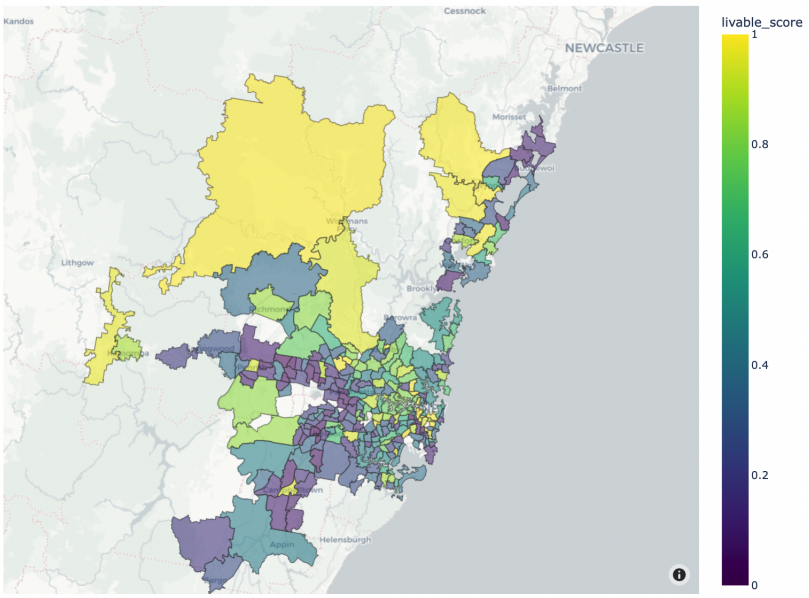

Figure 2: Livable Score in Greater Sydney

It is noticed that the color map does not cover all SA2 areas within the Greater Sydney region since SA2 areas containing incomplete information were dropped before calculation rather than being filled with 0 for consistency purposes.

According to Figure 2, it is observed that the livable scores of northern areas of Greater Sydney region are higher than those of southern areas in general. The high scores in some areas that are far away from central Sydney areas could be attributed to the low population density, which is involved in the calculation of many factors. Additionally, suburbs that are close to the central of Greater Sydney region have relatively good livability. A possible reason for this could be these areas are better developed and have more facilities than other counterparts.

3.1 Correlation Analysis in Greater Sydney

As illustrated in Figure 3(a) and (b), correlation between livable scores and another two factors (i.e., median annual household income and average monthly rent) was investigated. Their linear regression models can be expressed by the following equations:

$$median\_annual\_household\_income = 7484 * livable\_score + 46478 \tag{2}$$

$$average\_monthly\_rent = 639 * livable\_score + 1673 \tag{3}$$

3.1.1 Correlation between score and median annual household income:

Figure 3(a) indicate that people with lower income will live in an area with a lower livable score. However, such linear relationship between score and median annual household income was not strong. Although a positive relationship was observed, the coefficient was relatively small and the R^2 was calculated at only 0.079, indicating the regression model does not fit well.

The Pearson's correlation coefficient between these two factors was calculated at 0.281, which agrees well with the weak correlation observed in the linear model.

3.1.2 Correlation between score and the average monthly rent:

Based on Figure 3(b), a positive relationship between rent and livable score was observed, indicating that people living in an area with a lower livable score tend to pay lower rent. However, such linear relationship was not strong as well since the R^2 was calculated at only 0.181.

The Pearson's correlation coefficient between these two factors was calculated at 0.425, suggesting a week correlation between them as well.

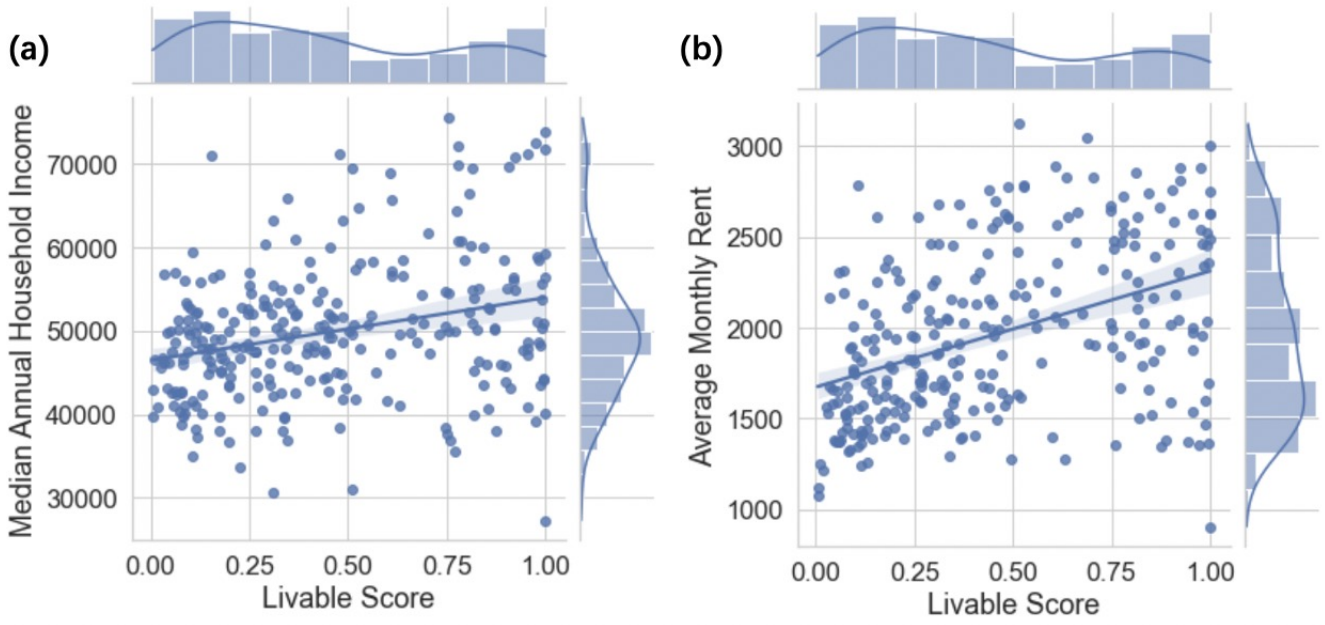

Figure 3: Livable Score in Greater Sydney

4 City of Sydney Analysis

4.1 A brief introduction of the stakeholder

The stakeholders in this study are young couples with children who want to move to Sydney and potentially buy a property near City of Sydney area. Hence, they prefer to live in an area that is safe and has plenty education resources for their children as well as good health resources. Hence, more weight should be given to these factors and they tend to less care about the diversity or prosperity of the area.

4.2 Tailor the livability based on the stakeholders' needs

According to the preference of stakeholders, a modified new livable score (NLS) that characterising the livability score of City of Sydney area is defined as follows:

$$NLS = S(4 * Z_{school} + Z_{accomm} + Z_{retail} - 2 * Z_{crime} + 3 * Z_{health} + 2 * Z_{camera} + Z_{industry} - Z_{price}) \quad (4)$$

where S is the Sigmoid function and the definition of each factor is described in Table 2 below. Similarly, all values calculated are further standardized by Z-score (i.e., the difference between input and mean value divided by standard deviation).

According to stakeholders' preferences, the weighting of Safety, Education categories were adjusted upwards from 20% to 26.7% while the weight of diversity was adjusted downwards from 40% to 20%. A new category cost was introduced since the price of properties could play a significant role in determining the livability for our stakeholders. Specifically, three additional measures were included by exploring three additional datasets as shown below:

1. Camera

   The total number of street cameras of an SA2 region divided by the area of that region were included in the calculation. By including this measure, we believe it will provide additional aspects to characterise

Category	Weight(%)	Measure	Definition
Education	26.7	School	number of schools catchment areas per 1000 young people (age¡20)
Diversity	20	Accomm	Same as the definition in Table 1
		Retail	Same as the definition in Table 1
		Industry	self-defined industrial diversity index
Safety	26.7	Crime	sum of hotspot areas divided by total area
		Camera	number of street cameras divided by region area
Health	20	Health	Same as the definition in Table 1
Cost	6.6	Price	average median price for 2-bedroom units of the most recent year

whether a region is safe since less traffic accidents and criminal activities could occur in areas with large number of street cameras.

2. Industry

   The index defined to can be expressed as follows:

$$Industry\_Diversity = \sum_{i=1}^{j} (\frac{n_i}{N_i})^2 \tag{5}$$

   where j is the total number of industry types (excluding health, food, accommodation, retail facilities) within the City of Sydney region (i.e., 19 industries), n_i is the total number of businesses belong to the i^{th} industry within a specific SA2 area, and N_i is the total number of businesses belong to the i^{th} industry within the entire City of Sydney region, respectively.

   In addition to health, retail, accommodation and retail services, it is believed that the above index can better reflect the prosperity of a SA2 area and can provide information on potential job opportunities for stakeholders.

3. Price

   This measure the cost of buying a 2-bedroom unit in a SA2 area within City of Sydney region, where the average of median price extracted from transaction history within the most recent year was extracted for calculation. This measure negatively contributed to the livable score and added an additional category to better reflect stakeholders' needs.

4.3 Visualisation of new livable score for City of Sydney region:

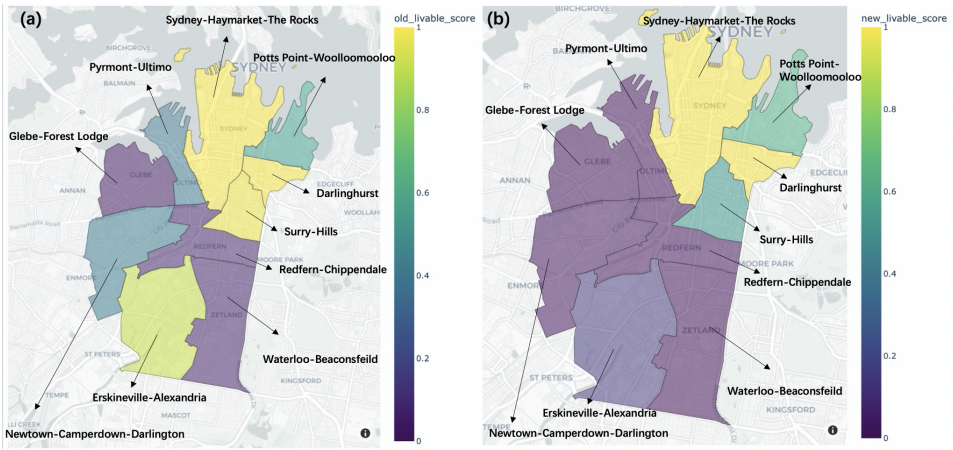

Figure 4: Compared to the original and refined livable score in the city of Sydney

In Figure 4, a compassion between livable scores calculated using old standards (left) and new standards (right) is presented. It is evident that the livable scores of area changed significantly after new standards

applied. For instance, the livability score of *Erskineville - Alexandria* decreases from an above-average level to a low value around 0.2. The employment of new standards made it easier for our stakeholders to select ideal living area since many "average" areas became less attractive. According to Figure 4(b), *Sydney-Haymarket-The Rocks* area exhibited the highest livability score of 1.0, closely followed by *Darlingurst*. These two areas were outstanding based on the new criteria since the average score of City of Sydney region was only 0.331. *Glebe - Forest Lodge* showed the lowest livability score of 0.016.

It can be indicated that some factors are highly correlated, such as the categories of education and safety, while some factors highly fluctuate.

4.4 Correlation Analysis in Greater Sydney

As illustrated in Figure 5(a) and (b), correlation between new livable scores and another two factors (i.e., median annual household income and average monthly rent) was investigated. Their linear regression models can be expressed by the following equations:

$$median\_annual\_household\_income = -9686 * livable\_score + 57586 \tag{6}$$

$$average\_monthly\_rent = 334 * livable\_score + 2306 \tag{7}$$

4.4.1 Correlation between new livable scores and median annual household income:

In Figure 5(a), a negative linear relationship was observed, indicating that people with higher income will on the contrary live in an area with a lower livable score. However, the model failed to fit the relationship well since the R^2 was calculated at only 0.135, which could be attributed to the fact that only limited data were provided. Pearson's correlation coefficient between these two factors was calculated at -0.367, indicating a week negative linear correlation as well.

4.4.2 Correlation between new livable scores and average monthly rent:

Based on Figure 5(b), there is a positive relationship between rent and livable score, suggesting that the people paying higher rent tend to live in areas with higher livability scores. Although R^2 was calculated at only 0.307, Pearson's correlation coefficient between these two factors was calculated at 0.554, which could be classified as a medium-level correlation.

5 Conclusion

In this study, a livable score considering safety, education, diversity and health was firstly calculated to establish a basic understanding on the liability of Greater Sydney region. To provide better suggestions for our stakeholders to identify ideal areas for living, the team modified the score by including more features and adjusting weights of each measures.

In our analysis, both *Sydney-Haymarket-The Rocks* and *Darlingurst* areas exhibited high scores and they are ideal areas to live for our stakeholders.

6 Appendix

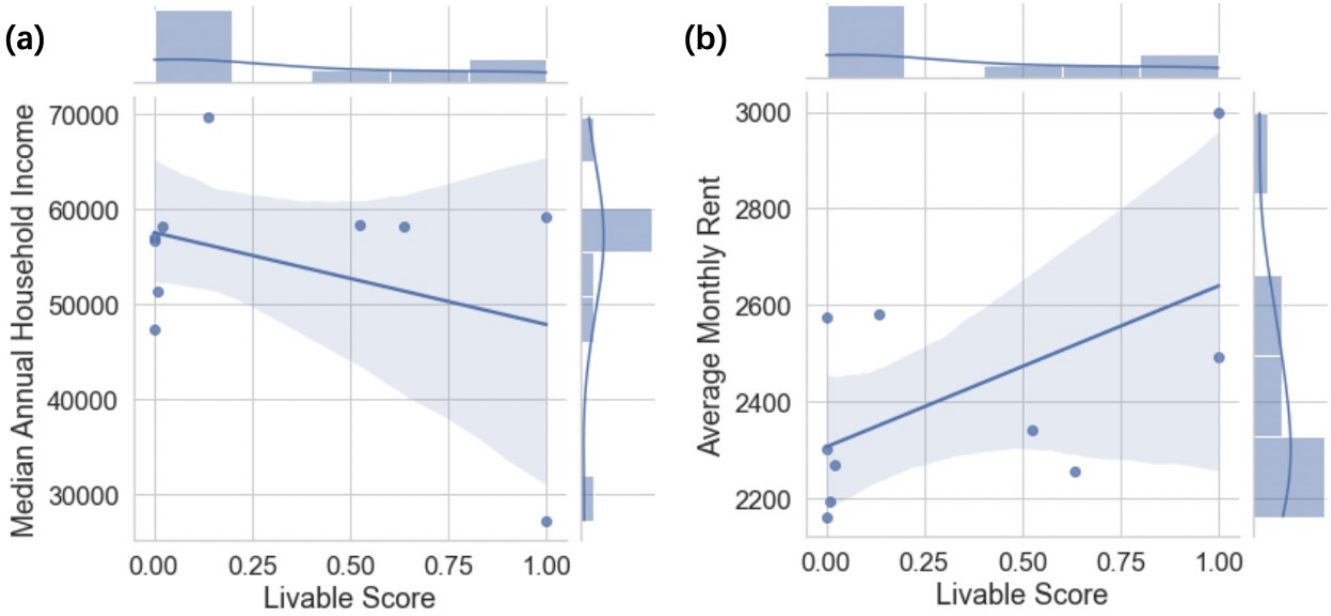

Figure 5: Correlations between livable score and median household income /avg monthly rent in city of Sydney