2022-5-18

Sydney livability analysis

DATA2001 Group Report

Group: F14D-RE10-3

Presented by:

510015316, 510496102, 510407902

Dataset Description

Data sources

Our group used seven datasets for this assignment, five of which were provided and an additional two were extra datasets. Both the Neighbourhoods.csv and the BusinessStats.csv in the provided datasets are from the ABS Census Data. SA2_2016_AUST.zip was obtained from the Australian Bureau of Statistics (ABS) and the relevant parent area. break_and_enter.zip is data on theft 'hot spots' in NSW as identified by the BOCSAR. Finally, school_catchments.zip covers all school information from kindergarten to high school, provided by the NSW Department of Education.

New datasets

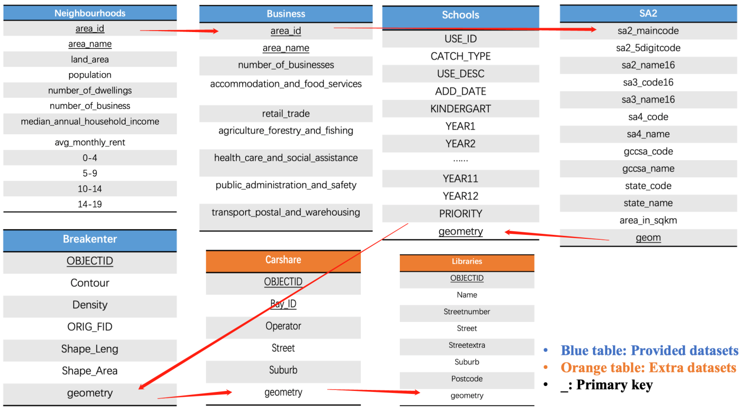
The additional datasets used by our group were all derived from the City of Sydney Open Data Hub. The first one is a dataset called Libraries, which contains information about libraries across Sydney where you can find any books you want to borrow, as well as public Wi-Fi, free printers, and other information. The second dataset is called car share bay. All of Sydney's car sharing-enabled stations are available here, providing city dwellers with a convenient, affordable, and sustainable transport option.

Pre-processing the data

When using data for analysis, we consolidate and cleanse the data. For analyzing data, we use PostgreSQL in our Jupyter Notebook. At the beginning of the analysis of the data, the 'psycopg2' package was imported and a connection was made with the PostgreSQL server at the University of Sydney with its own ID number. The first five data sets are provided by the university and can be downloaded directly from the Canvas; the latter two sets of additional datasets are selected according to the needs of the assumed stakeholder and downloaded as GeoJson file format. It is relatively easy to clean these seven sets of data. What we do is to find out NA/NULL value and some data that inconsistent. The solution is to delete the entire row with this data. In addition, we also renamed some column headings to facilitate our analysis.

Database Description

The seven-database schema are showing below as graph:



For the Neighborhoods dataset and the BusinessStats dataset, our group created an index for each of them. To gain better data integration, the indexes are set to divide different area from the column ‘area_name’. The indexes are named as ‘name_area’ and ‘b_name_area’. This naming method directly adopts the reverse column name to make it easier to remember and use. When this index is established, we can quickly retrieve data from a large database, making access to the RDBMS faster. Without a doubt, indexing a table or view is one of the best ways to improve query and application performance.

Sydney livability analysis

For international students from overseas, the livability score and correlation are accurately and intuitively provided to them to choose a suitable residence.

Livability score

the livability score is generated via the following formula:

livability score = z(school) + z(accom) + z(retail) - z(crime) + z(health)

$$z = \frac{x - \mu}{\sigma}$$

Where:

Z=number of standard deviations a raw score (x-score)

X=an interval/ratio variable

μ=the mean of x

σ=the standard deviation of x

School

The column school is the number of schools catchment areas for each thousand people aged 0-19.

Accom

The column accom is the number of accommodation and food services per 1000 people.

Retail

The column retail is the number of retail services for each thousand people.

Crime

The column crime is the sum of hotspot areas divided by total area.

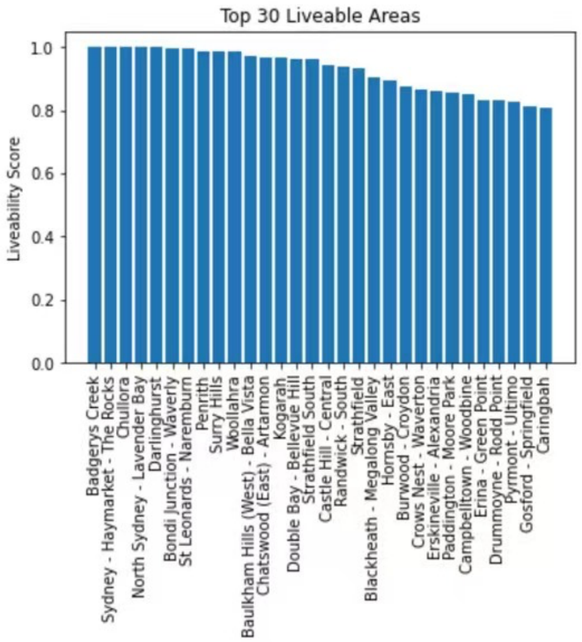
Health

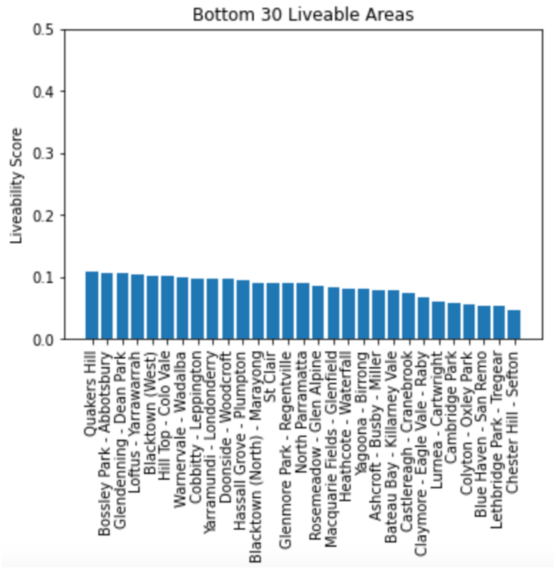
The column health is the number of health services per 1000 people.

Result of livability score

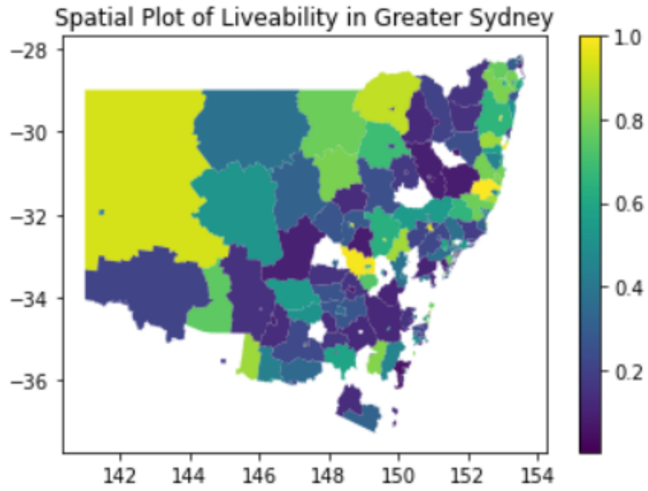
For Sydney's suitability score, the average score was 3.860495e-01, the standard divination was 3.898288e-01, the maximum score was 1.000000e+00, and the minimum score was 8.051704e-07.

Areas with suitability





The first map is the top 30 highly habitable areas such as Double Bay and Bondi. Most of their z are above 0.8, which shows that there are many school malls in this area and the environmental safety factor is high. The second map is Sydney's top 30 low-livability areas. Their z-score average is around 0.5, with the highest score being only 0.1 (Quakers Hill) and the lowest score being less than 0.3 (Chester Hill-Sefton). This is a very unsuitable environment for human habitation. This may be accompanied by inconvenient transportation, low safety factor, remote location, and sparse population.



This map shows the livability score for each area of Sydney. Where the x and y axes correspond to the SRID index, which corresponds to a spatial reference system based on a specific ellipsoid used for flat-earth mapping. In the map, we can see that different areas display different colors. The color index can be compared by the color bar on the right, which shows the livability score. The lighter and closer to yellow the color is, the better the conditions in this area and suitable for everyone to live in, the darker the color and closer to purple, the lower the score. If we put a virtual coordinate system in the middle of the map, it can be concluded that there are more rough blocks in the south and east of Sydney, and it is best not to choose to live. However, the north and west-north locations on the map are colored between green and yellow, which means it's a good place to live and possibly with a higher happiness index.

Correlation analysis

To determine whether there is a correlation between these two variables, we compared suitability scores with median income and median rent. Median house income and average monthly rent are from Neighbourhoods.csv. The correlation between these two variables is calculated using Pearson’s correlation coefficient formula.

The correlation coefficient of median house income to Livability Score is 0.2502081816756671.

The correlation coefficient of average monthly rent and Livability Score is 0.48936557355016713.

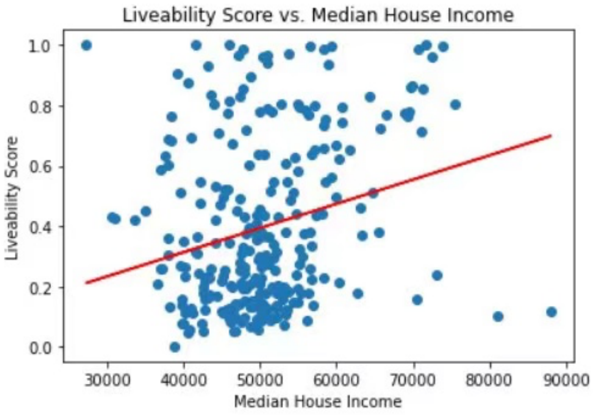
These correlation coefficients are not very high, but it is indeed a positive correlation, showing a slight positive trend on these two factors. The occurrence of this situation may be due to the following factors:

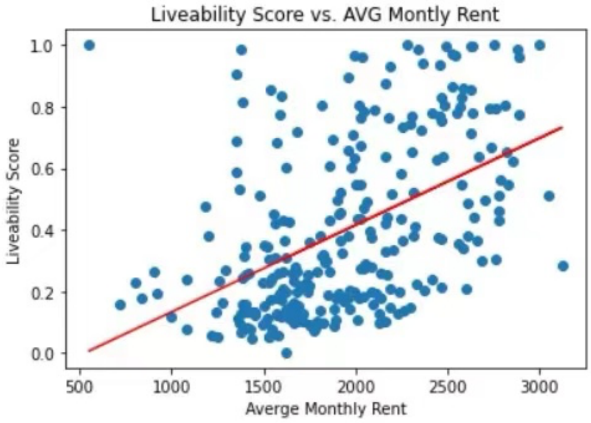
1）areas with good community environment, have higher safety factor and high cost of living, rather than low-cost housing and remote areas. Therefore, both income and rent are positively correlated with the livability score, and the trends of the two correlation coefficients are similar.

2）In addition, not everyone has high incomes and sufficient funds, and those with low incomes will have to choose houses far away from the city center. This leads to the reason why the two correlation coefficients are not very high even if they are positively correlated.

3）Variables cannot all be considered perfectly in determining livability. And some external datasets may affect the results of correlation.

The correlation coefficient





The graph above shows a linear relationship between the two factors and the livability score, but there are many outbounds. Except for higher and lower exits, the correlation was positively correlated. Due to the large number of outliers, this also explains why the two variables are associated with a lower correlation of habitability.

City of Sydney Analysis

The stakeholders

For our hypothetical client, our setting is a group of college students who are about to graduate or have graduated and want to stay in the local to find a job. They are all international students from overseas who do not want to return to their home countries for the time being but want to stay in Sydney to continue their job search. So, for them, they want to live in a relatively prosperous area, close to many office areas or business districts, so that they can have more job opportunities. At the same time, there should be a lot of public transportation facilities, such as subway, buses, or car rental centers. It is also necessary to choose a place with a very low crime rate, because as a foreigner, you are very concerned about your personal safety, and you will feel more at ease if you choose a community with better public security. Coupled with the supporting living facilities, the ability to easily buy food or daily necessities is also within the scope of consideration. For our additional data sets, one is the dataset of the library, and the other is the car share bay. The reason for us to choose datasets is because, firstly the library can provide a comfortable working place, and it is more perfect if it comes with free Wi-Fi or a free printer; at the same time, because some office areas are not near the subway station, or the two places are a little far apart. For those who can drive, if they can't afford to buy a car, they can rent one for their own use or even use them for future commuting. This is a very economical and cost-effective solution.

New formula

We collected data from two datasets (libraries.geojson, carshares.jeojson) and calculated the z-scores to get a new livability score.

livability score = z(school) + z(accom) + z(retail) - z(crime) + z(health) + z (population density) + z (business density).

Each index is weighted and divination differently in the livability calculation due to re-updating customer needs. As foreigners are most concerned about safety, it is extremely harmful to livability if a crime occurs in the area. As a job seeker, the population density in an area needs to be moderate to deal with enough people. At the same time, the business density in the area must also be high to better increase the chances of job hunting. As a single person, there is no family and children around, so they don't care about the number of schools in the area. Also, for a better quality of life, there should be an absolute number of retail stores and food stores around the community to ensure supplies of daily necessities. Therefore, based on a variety of considerations, the calculation results of livability have also changed.

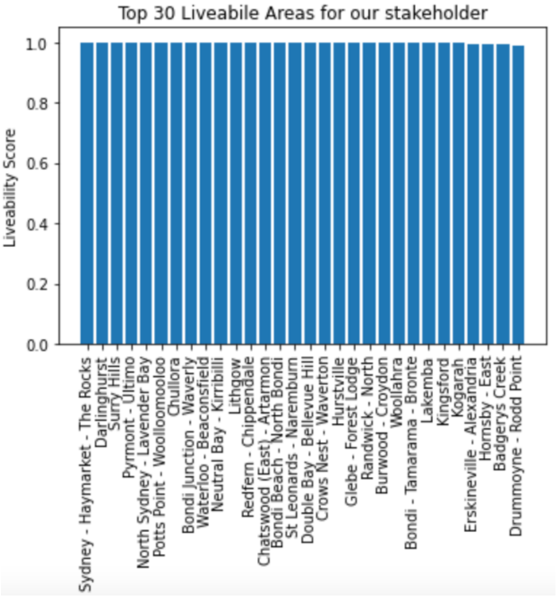
Population density

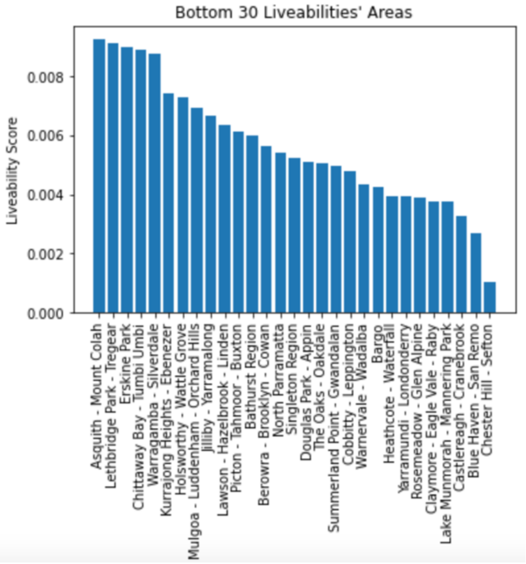
Population density is the population divided by the land area. Calculate z-score on population density by python and standard score equation.

Business density

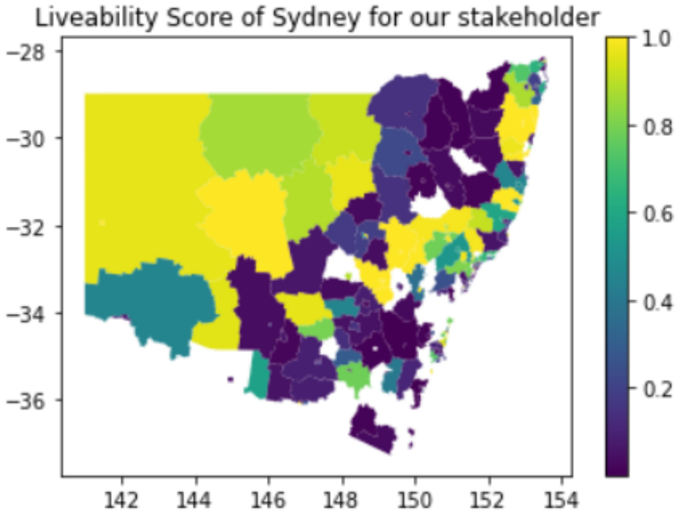
Business density is the number of dwellings plus the number of businesses and then divided by the floor space. Calculate z-score on business density by python and standard score equation.

Result of curve





According to the average suitability score 3.860495e-01, there are some places in Sydney that are suitable for stakeholders, but not all areas are suitable for living. As shown in the image above, stakeholders will select any area on the first map instead of the area on the second map such as Redfern and Bondi, among others. This means that these areas may have a high safety factor, dense population, easy transportation, and a large number of malls and parks nearby. In addition, we analyzed the correlation between public administration plus safety, accommodation food services to livability scores, which were 0.25466845559282736 and 0.38867409170036743, respectively. Both correlation coefficients were positive, even though they were not highly correlated with the livability score. At the same time, these coefficients and curve provide a lot of useful information and assistance for stakeholders.



This map has been re-adjusted to consider the needs of our stakeholders. Because of the needs of our customer service, they need to live in a place with a prosperous economy that can provide more employment opportunities, and at the same time pay great attention to community safety. So, from our results, it is highly recommended that they choose the northern area of Sydney or the area close to the central area. The colors of these areas are all above green, and there are two areas that are very bright yellow, which represent the most satisfying needs of our stakeholders.