THE UNIVERSITY OF SYDNEY

The University of Sydney

School of Computer Science

Practical Assignment: Sydney Liveability Analysis

Author:
Harry Lockyer
Ke Lyu

Unit Coordinator &
Lecturer:
Uwe Roehm

An Assignment submitted for the UoS:

DATA2001 Data Science, Big Data and Data Variety

May 21, 2022

Contents

Dataset Description

What are your data sources and how did you obtain and pre-process the data?

In this article, we have obtained 7 datasets for *the Greater Sydney Liveability Score Analysis* and *City of Sydney Analysis*. Datasets: **Neighbourhoods, BusinessStats, Catchments, BreakAndEnter** and **SA2 Shape Data** were provided by the course. The other 2 datasets: **Chairs** [3] and **Mobility Parking** [2] were acquired from City of Sydney Open Data Hub [1].

After collect the datasets from Canvas and City of Sydney Open Data Hub. Data cleaning should be performed since we should keep data consistency and remove data we might not need for our further research. At this step, we mainly use Jupyter Notebook and Python as our data cleaning tools. For the two csv files: **BusinessStats** and **Neighbourhoods**, Python module **Pandas** was used to read the file. Furthermore, we dropped columns not needed in our analysis, converted data type and rename for readability. In terms of spatial data, module geopandas has been used for geospatial operations. We also dropped any columns not needed in the further research, and applied name and type conversion.

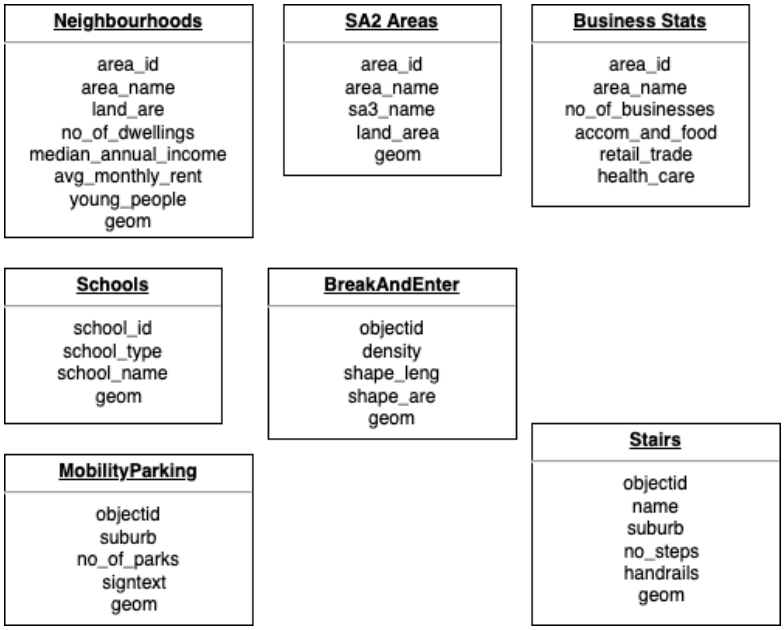
Moreover, in order to fully utilise the spatial data, we defined a WKT function to make sure all geometries are the same as those expected by PostGIS requires conversion to the **Well-Known Text (WKT)** format: MultiPolygons and POINTS. We also specify the **Spatial Reference Identifier (SRID)**: 4283 (GDA94) as per the **SA2** file. Datasets from City of Sydney DataHub uses 28356, which is still GDA94. Then we create a geometry column for each table to store suburbs and locations' geographical location, which could be used in the analysis.

Finally, we build a database using PostgreSQL that integrates data from cleaned datasets in order to starting **Greater Sydney Score Analysis** and **City of Sydney Analysis**.

Database Description

Into which database schema did you integrate your data (preferable shown with a diagram)? Which index(es) did you create, and why?

In order to place the database objects into a logical group and allow researchers quickly access enormous information, several schema was created during the process of integrating and populating the data. The following diagram demonstrate the schema has been using during the analysis.



To make spatial joins faster than usual since the datasets in this research are massively large. The analysis will have to go through every record in the database if no spatial index presents [4]. Hence, we choose to create spatial indexes to accelerate this process. At this step, The **GiST (Generalised Search Tree)** method has been implemented to break up data into "things to one side", "things which overlap", "things which are inside" [4]. We built 5 indexes based on each spatial data table's geometry column.

Greater Sydney Score Analysis

Show which formula you applied to compute the liveability score per neighbourhood, and give an overview of the results through

To analyse Greater Sydney liveability, the following formula has been applied in terms of school, accommodation, retail services, crime and health service. Accommodation, retail and health services are calculated as the number of service per 1000 people. School catchment areas z-score is the number of areas per 1000 young people which we defined as anyone from 0 to 19 years old. And crime is the sum of hotspot areas divided by total area. We assume that more school catchments, accommodation and food services, retail services and health services will positively impact on suburb's overall life quality. On the other hand, more crime in the neighbourhoods will lead to a negative affect.

$$S = \mathcal{S}(z_{school} + z_{accomm} + z_{retail} - z_{crime} + z_{health})$$

The $\mathcal{S}$ is the sigmoid funtion [5], z is the normal z score.

Measure	Definition	Risk	Data Source
school	number of schools catchment areas per 1000 'young people'	+	school_catchments.zip
accom	number of accommodation and food services per 1000 people	+	BusinessStats.csv
retail	number of retail services per 1000 people	+	BusinessStats.csv
crime	sum of hotspot areas divided by total area	−	break_and_enter.zip
health	number of health services per 1000 people	+	BusinessStats.csv

Firstly, according to the formula, we built tables including area id and their performance per 1000 young people, per 1000 people and per area for different factors. This allows us to calculate the liveability score in the Greater Sydney in the next step. Afterwards, by combining each table and apply formula to each factor, the z score has been generated. Finally, by using the sigmoid function, we could acquire the final score for every area in the Greater Sydney. In the research, the final score falls between 0 and 1. Higher score means the area is more liveable

Top 10 most liveable Areas

	area_id	area_name	sigmoid
79	117031337	Sydney - Haymarket - The Rocks	1.000000
99	119011355	Chullora	1.000000
71	117031329	Darlinghurst	0.999964
167	121041417	North Sydney - Lavender Bay	0.999945
149	121011401	St Leonards - Naremburn	0.999429
78	117031336	Surry Hills	0.999248
2	102011030	Calga - Kulnura	0.998685
83	118011341	Bondi Junction - Waverly	0.996477
182	122031432	Terrey Hills - Duffys Forest	0.993600
34	115011553	Castle Hill - Central	0.992633

The table illustrates the most 10 liveable areas in the Greater Sydney in the descending order.

It can be seen that Haymarket - The Rocks and Chullora received highest score and tied as the most liveable area. The top 5 areas' score is relatively close. However, the score drops over 0.002 between each area after Surry Hills. This leave us a question, what factors lead to this trend? We could conduct further research on this in the future.

Overall, these areas we considered are the most 10 liveable areas in the Greater Sydney based on school catchments, accommodations, retail services, crime and health services performance in the research.

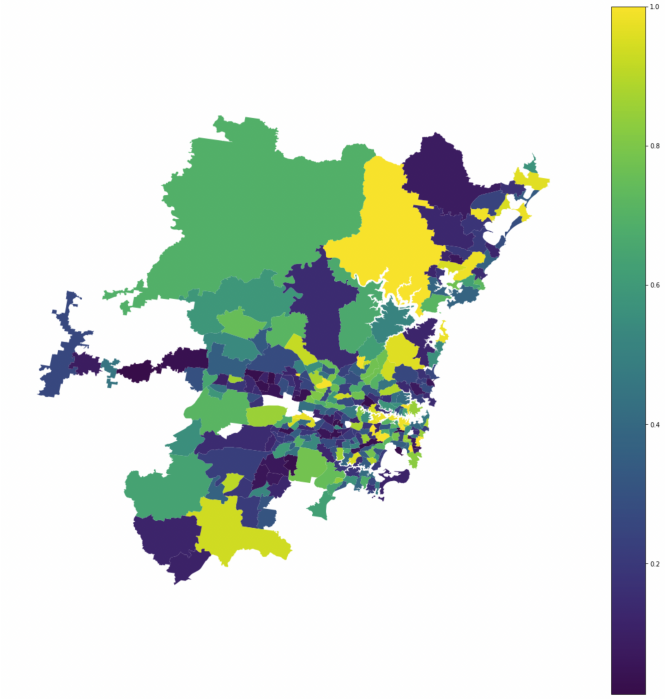
Top 10 most unlivable Areas

	area_id	area_name	sigmoid
217	124051470	St Clair	0.012168
90	118021350	Malabar - La Perouse - Chifley	0.015050
61	116031315	Hassall Grove - Plumpton	0.016713
65	117011321	Botany	0.020470
82	118011340	Bondi Beach - North Bondi	0.021451
207	124031460	Glenmore Park - Regentville	0.021748
255	127011504	Ashcroft - Busby - Miller	0.025996
62	116031316	Lethbridge Park - Tregear	0.026425
235	125031484	Guildford West - Merrylands West	0.028619
218	124051580	Colyton - Oxley Park	0.033160

In addition, we also disclosed the 10 most unlivable areas in the Greater Sydney areas in the ascending order. The table displayed St.Clair is the most unlivable area because it obtained lowest score in each factor.

Visualise the Greater Sydney

Finally, we create a map of the Greater Sydney so that the audience has a more direct experience over which area is more liveable and which area is less.

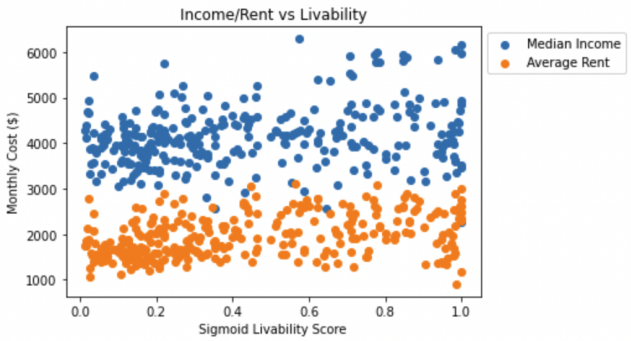


The diagram illustrates the Greater Sydney's liveability over each area. Colour yellow to black represents the final score of each area from highest(1.0) to lowest(0.0) in the map. The colour closer to yellow, the area is more liveable for

people to live. In the contrast, the colour more close to the black, the area is more unlivable for people. It can be seen that northwest of the city, areas along the Parramatta River and southwest of the city considered as liveable areas. However, interestingly, some areas with a low composite score are adjacent to areas with a higher score.

Correlation Analysis

How well does your score correlate with the median rent and median income in each neighbourhood?



The diagram displays how the median rent and median income correlate with the final livability score in the Greater Sydney. The figures are given as monthly cost.

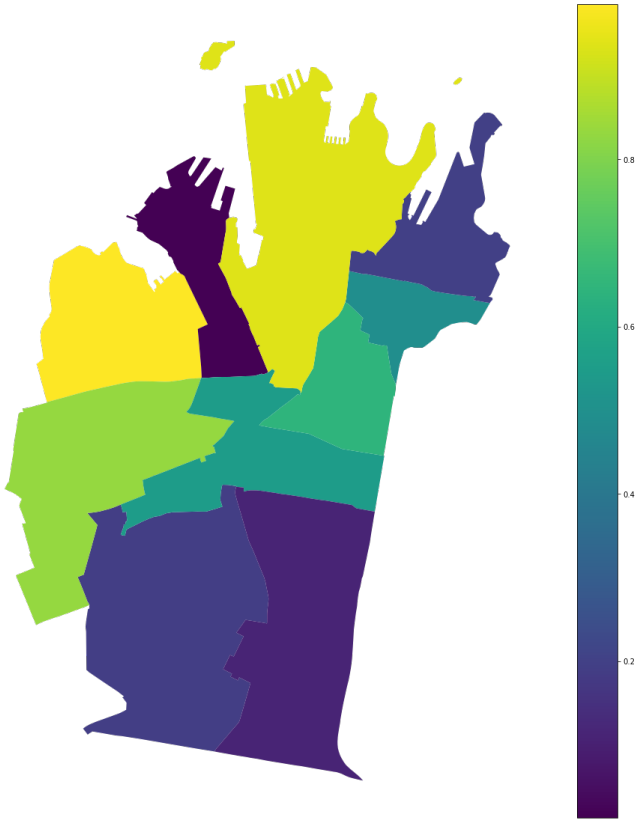
Overall, there is no significantly correlation between median income and final score as the score increases. Also the median rent shows week correlation with the final score increases. However, we did observe there could be a small correlation between low income/low rent and low liveability. Even so, we still could not see any obvious trend in the correlation analysis.

City of Sydney Analysis

Propose a stakeholder and give a brief introduction. Show how you tailored your score for their needs. Demonstrate the results on a map.

Our stakeholder is a disabled international student who is decided to move to Sydney in July. She has asked us to help her to find an area has the most accessible infrastructure in the city of Sydney since they could provide great convenience to her life. After receiving her proposal, we thought we could build on our previous

research by adding data of stairs and mobility parking. But since she was looking for properties near university of Sydney, we narrowed down our search to the city of Sydney. Therefore, in this part, we will continue previous research and place extra emphasis on stairs and mobility parking when calculating the overall score.



For our research within the City of Sydney, we re-calculated all the z-scores for each factor, giving us a better understanding of the liveability of just City of Sydney, ignoring the rest of Sydney. We also considered the importance of each of the factors to our stakeholder, adjusting each score by a factor between 0 and 2 (0 being less important and 2 being more important). For the school score, adjusted the score using a factor of 0.1. This is because the stakeholder already is at university and has little to no use for the data. For the accommodation score, we also adjusted it by a factor of 0.1. As they are looking to move to City of Sydney, they will not need to know how much accommodation is around. We left the crime score as it is, as crime is no more important to our stakeholder as it is to anyone else in the city. We adjusted the retail score by a factor of 1.0 too, as retail businesses was not unimportant to our stakeholder but did not need to be given any more importance. To our stakeholder, health services are quite important. Thus, we adjusted the health score by a factor of 1.5. We also added in a score based off the average monthly rent for the area and adjusted it by a factor of 1.2. Being a student, our stakeholder will have to consider how much they are paying for rent. With all these scores calculated and passed through the same sigmoid function, the top-rated area was "Glebe – Forest Lodge". From the data, this area would be best suited for our stakeholder. Its health score is in the middle of the pack, but it has a very low stair score and a very high parking score, meaning that it would be quite easy for our stakeholder to move around the area. It also has the lowest average monthly rent of the area, which is beneficial for our stakeholder.

References

[1] City of Sydney. City of Sydney Data hub, 2022. Last accessed 23 May 2022.

[2] City of Sydney. Mobility parking, 2022. Last accessed 23 May 2022.

[3] City of Sydney. Stairs, 2022. Last accessed 23 May 2022.

[4] PostGIS. Chapter 4. data management.

[5] Wikipedia. Sigmoid function, Apr 2022.