1 Executive Summary

This report focuses on the livability analysis of each neighbourhood in Greater Sydney, with a documentation of the data cleaning, schema of the database, livability result and analysis. This report also provides recommendations for a young couple looking to rent a home in the City of Sydney.

2 Dataset Description

2.1 Neighbourhoods

The neighbourhood data is census data obtained from the NSW government in the form of a CSV file. When this data is stored in the PostgreSQL database, if areas contain null values in median renting and/or median income columns, they are removed from the database. By carefully looking into the areas that do not have records for renting or income, these are the areas that are either industry sites, national parks or airports. These are not livable spaces or not suitable environments for humans and, therefore, can not contribute to our livability analysis. Furthermore, prior to loading this data set, columns such as population and dwelling are expected to be numerical values; however, there were commas after the thousands digits, which stopped the transformation of those columns to numeric. As a result, during the data cleaning process, any commas appearing in those two columns have been replaced with empty string, (“”) through the Pandas package to obtain numerical values. An extra column called ‘young’ was also added to the database, calculating the sum of all teenagers and young adults for each neighbourhood.

2.2 Businesses

The business data set is obtained from the NSW government in the form of CSV. This dataset initially contains business information all around Australia. However, this report only focuses on neighbourhoods in Greater Sydney. Business information related to Sydney is obtained by making comparisons with the neighbourhood dataset, such that only rows that contain area IDs that match the area id in neighbourhood data were kept and saved to the server. Furthermore, with respect to the goal of this report, columns other than accommodation and food services, health care assistance, and retail trade were removed. We will not consider these for neither section 2 nor 3 calculations.

2.3 Break and enter

The break and enter data is obtained from BOSCAR as spatial data. This data is pre-processed by first transforming all polygons to multipolygons to ensure consistency for later queries and conducting WKT conversion with srid = 4283 when uploaded to the Postgres server.

2.4 Schools

This dataset is obtained from the school file via the NSW government. It is a concatenation of primary schools, secondary schools, and future schools. Firstly, we keep the school ID, school name, and geometry columns in the database as there is a consideration of the number of schools rather than types, dates and priority in this analysis. Secondly, we checked that there was no invalid value.

Thirdly, we removed the schools that appear in both future and existing schools. Finally, we transform the SRID in the geometry column.

2.5 SA2
This dataset is from the SA2 shapefile from the NSW government. Firstly, we only kept five columns which are SA2_MAIN16 for SA2 ID, SA2_NAME16 for SA2 name, SA3_NAME16 for City of Sydney analysis, GCC_NAME16 for Greater Sydney analysis, and geometry. Secondly, we dropped the columns whose geometry is none. Finally, we transform the srid in the geometry column.

2.6 Water fountains
This dataset is obtained from the City of Sydney Data Hub API in the form of GeoJSON. It contains the point locations of 215 drinking fountains (water bubblers) within the City of Sydney. It contains 5 columns, including the site name where the fountain is located, the suburb of the site, type of site, accessibility status and point locations by coordinates. Some accessibility statuses are unknown, but this column is unnecessary in our z-score calculation and livability analysis, so we dropped this column.

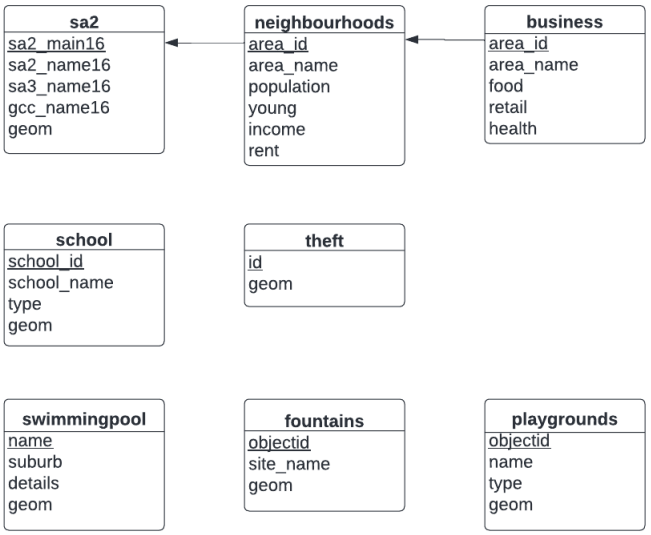
2.7 Playground
This dataset is obtained from the City of Sydney Data Hub API in the form of GeoJSON. It contains the point locations of 151 playgrounds within the City of Sydney. Each row contains an object ID, name of the playground, type of playground, and a point location by coordinates. Since all playgrounds are treated equally regardless of the type, we have discarded the “type” column.

2.8 Swimming Pool
This dataset is obtained from the City of Sydney Data Hub API in the form of GeoJSON. It contains the point locations of 5 public swimming pools within the City of Sydney. Each row contains a unique object ID, name of the pool, class of the pool, phone number, URL, pool description details, postcode, suburb, street address and a point location in coordinates. Since we only needed the object ID and name as an index and the point location for counting purposes, we have dropped all other columns.

3 Database Description

3.1 Data schema description
The database consists of eight tables. Keys underlined are primary keys, and area ids in neighbourhoods and business are one-to-one foreign keys such that the three tables (sa2, neighbourhoods, business) are linked by area id. Five tables (school, theft, swimmingpool, fountains, playgrounds) are linked with neighbourhoods by topological relationships using geometry columns.

3.2 Data schema diagram



3.3 Indexing description

3.3.1 SA2.geom

Due to the large dataset in SA2 and its frequent use, we created a spatial index in SA2.geom. SA2.geom is an intermediary between non-spatial neighbourhoods and their spatial data, thus in later queries, it is an important database for us to draw connections between neighbourhoods and its relevant spatial information (by finding matching area id in neighbourhood). Index in SA2.geom speeds up the queries of livability scores.

3.3.2 School.geom

As there is also a large dataset in school, We created a spatial index in School.geom to speed up the calculation of school numbers when matching its geometry points to SA2.

4 Greater Sydney Score Analysis

4.1 Formula

$$Score = S\ (\ Zschool + Zaccomm + Zretail - Zcrime + Zhealth)$$

Where S is the sigmoid function, z is the normal z score: $z = (x - mean) / stddev$

Where school is the number of schools per 1000 young people, accomm is the number of accommodation and food services per 1000 people, retail is the number of retail services per 100 people, crime is the sum of hotspot areas divided by total area, health is the number of health services per 100 people.

4.2 Result

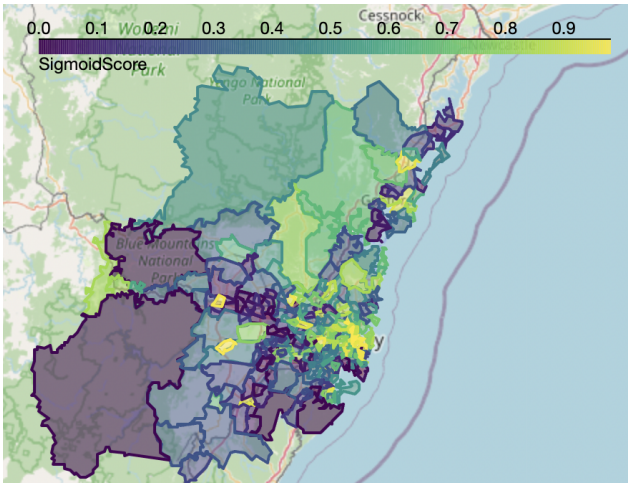
The top 10 neighbourhoods in Greater Sydney is demonstrated below:

NeighbourhoodName	SigmoidScore
Sydney - Haymarket - The Rocks	1.000000
Badgerys Creek	1.000000
Chullora	0.999997
North Sydney - Lavender Bay	0.999949
Darlinghurst	0.999834
Bondi Junction - Waverly	0.999175
St Leonards - Naremburn	0.998991
Surry Hills	0.998291
Double Bay - Bellevue Hill	0.991610
Kogarah	0.990590

The lowest score is 0, where there are 13 neighbourhoods. These are all non-residential areas such as industrial areas, military areas, parks, reservoirs, airports etc. The mean score is 0.435838. The median score is 0.391531.

4.3 Visualisation

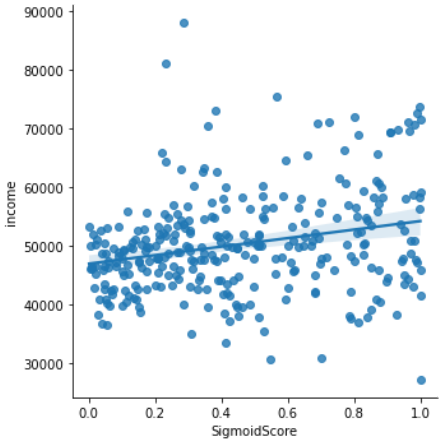
Map of Greater Sydney with livability visualisation



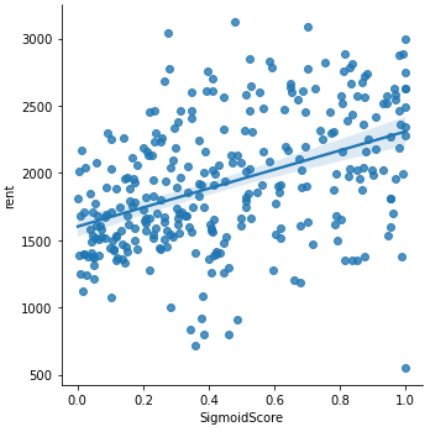
As illustrated by the visualisation above, neighbourhoods with lighter colours represent a higher livability. As a result, areas in or near the east of Sydney such Chatswood, and Ryde, as well as the north of Sydney appeared to have higher livability. (refer to jupyter notebook for interactive function)

5 Correlation Analysis

5.1 Graph



Correlation between livability and income



Correlation between livability and renting

5.2 Result

The correlation coefficient between the livability of neighbourhoods and income is 0.252. This positive result indicates that there is a proportional relationship between median income and livability. Specifically, as the median income of the areas increases, the livability is more likely to increase. However, the small value indicates a weak relationship between income and livability.

The correlation coefficient between median renting and livability is 0.434, indicating a positive and moderate relationship, where in areas with high livability it is more likely to find more expensive renting houses, yet it is not necessary due to the weak to moderate link.

6 City of Sydney Analysis

6.1 Stakeholder Introduction

The stakeholder is a young couple who also own a three-year-old puppy. This couple is working in the city of Sydney and has a shared hobby of swimming. They are demanding to rent a house in the inner Sydney areas. Furthermore, they also want areas that can provide them convenience for doing daily exercises, such as walking, jogging and swimming, and a suitable environment for them to walk the dog.

6.2 Livability Score

With consideration of the stakeholder's daily needs, the livability of the neighbourhoods in the City of Sydney takes into account the distribution of accommodations and food services, health care services,(per 1000 people) crime cases, and median renting prices of houses of each neighbourhood in the City of Sydney. As lower home renting contributes to releasing this young couple's financial pressure, this is calculated as a factor that increases liability. Furthermore, considering the stakeholder's family structure and personal lifestyle, we included the number of available swimming pools around each area to satisfy the stakeholder's hobby. Larger numbers of playgrounds and water fountains also contributed to higher livability scores in the analysis. More playgrounds provided the family with more choices to jog and walk their dog. An abundance of

water fountains provide convenience for the dog who used them as a source of water. The final livability score was calculated following the formula:

Score = S (Zfood + Zhealth + Zplayground + Zfountain + Zswimmingpool − Zcrime - Zrent)

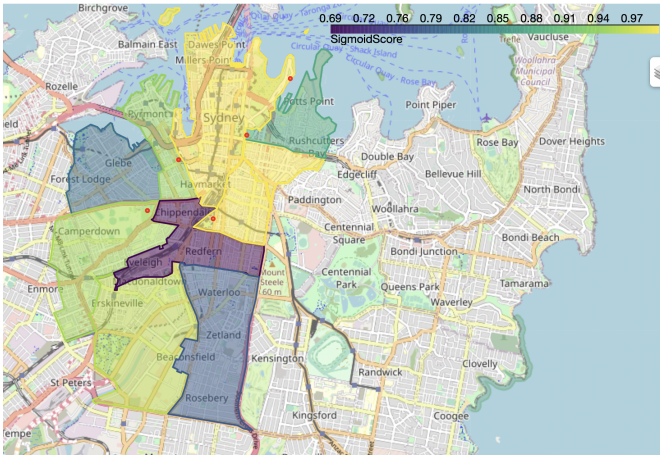
6.3 Result and Analysis

The result in descending order by their livability score is demonstrated below:

	area_name	SigmoidScore
8	Sydney - Haymarket - The Rocks	1.000000
7	Surry Hills	0.996812
3	Newtown - Camperdown - Darlington	0.977719
2	Glebe - Forest Lodge	0.975685
0	Darlinghurst	0.975169
1	Erskineville - Alexandria	0.936044
4	Potts Point - Woolloomooloo	0.898024
9	Waterloo - Beaconsfield	0.858333
5	Pyrmont - Ultimo	0.760655
6	Redfern - Chippendale	0.661457

The range of livability score is 0.34, with the highest being 1, at The Rocks and the lowest being 0.66 at Redfern.

6.4 recommendation and visualisation



As illustrated by the map above, areas with lighter, warmer colours such as yellow and light green indicate areas with higher livability scores. (refer to interactive map in jupyter notebook). As a result, the top three most recommended places for the couple to rent their house are Surry Hills, The Rocks and Darlinghurst. Furthermore, this map also highlights the locations of available swimming pools with red dots. To better satisfy the couple's hobby, Surry Hill, out of three top recommendations, is mostly recommended to be their first option for the most convenient access to local swimming pools.