DATA2001 Report: Livability Analysis

SID: 510288769
SID: 510359092

* Note: We also had another group member but he has been passive. We have a Facebook group, but it's hard for us to reach him (because he is often not online). He did not contribute to the report and the code he wrote for some of the z-score calculations was incompatible on Jupyter as it didn't use the database table, so we had to redo it. He also did not regularly turn up to meetings too (we had 9 meetings after he joined the group, only twice he attended) and even if he did, he remained mostly silent.

Dataset Description

We employed a variety of file types and, for spatial datasets, all spatial data columns were converted to the well-known text (WKT) format with the spatial reference identifier (SRID) of 4283 because it was indicated in the source document of SA2.

CSV documents

Both BusinessStats and Neighbourhoods were obtained from ABS census data and were in CSV format. Columns that contained irrelevant data for analyzing the livability scores were dropped in the cleaning processes, as this minimized the extra number of rows that had to be removed due to missing values in these irrelevant columns.

BusinessStats needed to be cleaned up again because some entries had 'Unknown' names in the 'area name' column, therefore their indexes were used to discard the rows (Appendix 1). This is to ensure that foreign keys can be used with the SA2 table.

Shapefiles

The NSW Department of Education provided the three independent school catchment datasets. Concatenating primary, secondary, and future catchment shapefiles with the 'add_date' and the different school years, which contained boolean values, resulted in a unionized dataset. We dropped the school years column as we thought the 'catch_type' column was sufficient enough to display which years the schools taught using well-known terms like 'primary' and 'secondary.'

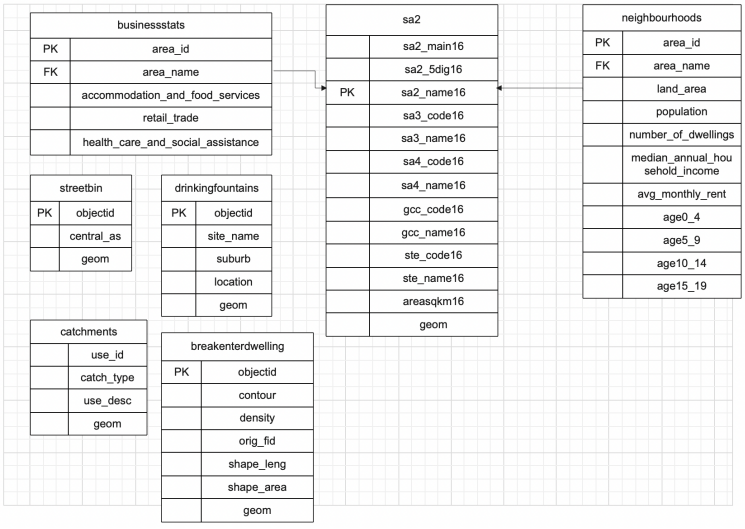
The break and enter dataset is a BOCSAR shapefile that was cleaned similarly by dropping rows with missing values.

Finally, there was the ABS-sourced SA2 dataset, in which the document stated that they used the Geodetic CRS of GDA94, which is equivalent to the SRID of 4283. We cleaned the dataset by removing any rows with missing values using the same procedure.

Independently sourced datasets

We used the 'City of Sydney Open Datahub' to find two spatial datasets that we thought were relevant to our stakeholders. The first was a geojson file containing information on all of the drinking fountains in Inner Sydney. We omitted columns with incomplete rows and deleted the 'Accessible' column, which did not contain essential data needed in computing the livability scores. The second dataset was a shapefile with information about street bins, which we cleaned similarly by deleting missing values. The links for both datasets are below:
https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::drinking-fountains-water-bubblers/explore?location=-33.903903%2C151.211923%2C13.44
https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::street-litter-bins/explore?location=-33.902833%2C151.192993%2C12.57

ERD Diagram



Database Description

We used the y22s1d2x01_zjin6432 database, in which seven tables were added to the public schema as it allowed for geographical data integration.

Index

Indexes are used to speed up the retrieval of data from a database. It is one of the best techniques to increase the efficiency of queries by swiftly assisting users in finding the data that need to be searched. Three indexes were created:

- suburb_idx which is focused on the ‘geom’ attribute from the sa2 table
- area_idx which is focused on the ‘gcc_name16’ attribute from the sa2 table
- sa3_idx which is focused on the ‘sa3_name16’ attribute from the sa2 table

The reason for a spatial index was because we used ST_CONTAINS, which is a spatially indexed function, where possible for cross-joining neighbourhoods. Both the area_idx and sa3_idx, which are based on the region name (sa3_idx) and greater capital city respectively, helped the query performance as data are filtered out based on being in the Greater Sydney region and the inner city areas utilizing both indexes regularly.

Greater Sydney Score Analysis:

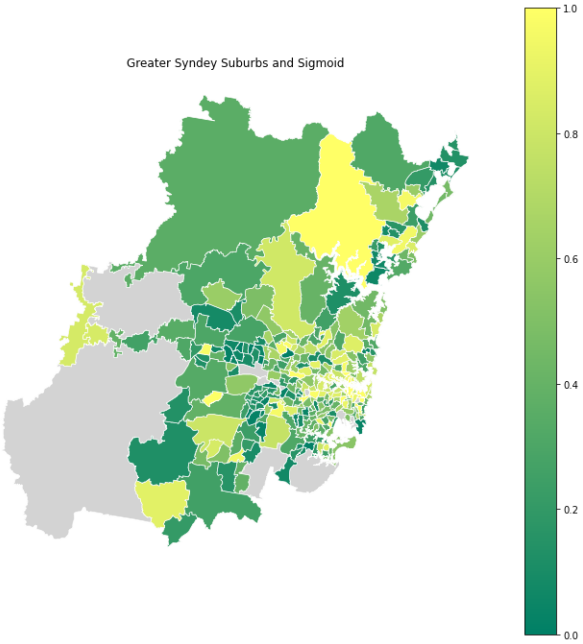
Z-score

$$\text{Score} = S\,(\text{Zschool} + \text{Zaccomm} + \text{Zretail} - \text{Zcrime} + \text{Zhealth})$$

The Z-score is a standardized score done by dividing the difference between the measured value and the mean by the standard deviation. This allows us to compare multiple datasets using the same set of criteria. To calculate the z-score, we first created a SQL query that allowed us to use PostGIS to link neighborhoods and filter out data. Because some of the data, such as 'BusinessStats' and

'Neighbourhoods,' had no 'Geom' columns to join with, we joined them by their area_names. We also divided the number of services by the population per 1000 and totaled the hotspot areas based on the task requirements. After that, the query is converted to a dataframe, and the z-score was calculated using the module from scipy.stats.

Greater Sydney Map



In the map above, the color bar illustrates that the sigmoid is close to 1 around yellow regions, while it is closer to 0 near dark green regions. We plotted grey regions to guarantee there were no voids due to missing data in several suburbs inside Greater Sydney, which may mislead the reader if they are unfamiliar with that map of Sydney. According to the degree of color difference, the closer the area is to yellow, the higher the sigmoid score, suggesting that the area is more livable; conversely, the closer the area is to green, the lower the habitability. There are some grey spots in this graph where no z-score could be generated due to missing numbers, hence it has limits. However, it provides a broad perspective of each Greater Sydney suburb region. Furthermore, without further clarification, it is difficult to see all of the small suburbs close to the CBD.

Correlation Analysis

Correlation:

A positive correlation value ranges from 0 to 1, with the value closer to 1 indicating a stronger correlation between the two variables. Furthermore, the association is smaller the closer the value gets to 0.
The following table shows our correlation coefficients:

	Livability Score
Median rent	0.429 (3 d.p)
Median income	0.329 (3.d.p)

According to our correlation calculation:

- The sigmoid scores and the median rent for each neighborhood have a moderate positive association.
- The sigmoid scores and the median income for each neighborhood have a weak positive association.

Because both correlations between each suburb are positive, it follows that if one variable increases, the other will most likely increase as well. Furthermore, this shows that high-rent and high-income neighborhoods are more likely to have better livability scores.

City of Sydney Analysis

Proposed stakeholder

We proposed our stakeholders to be families with young children eager to relocate to the inner city for the ease of living near the CBD. We thought it was appropriate because previous z-score calculations for Greater Sydney contained data relevant to young children, such as the number of schools per 1000 youths.

With our stakeholders in mind, we decided to obtain datasets related to the local environment. As a result, the street bins dataset was chosen because the presence of bins encourages people not to litter, which not only keeps the neighborhood clean and free of visual and odor pollution but also instills environmentally friendly habits in children by showing them how adults in their neighborhood live.

Furthermore, we chose drinking fountains, where higher counts increased the livability score. This is because we believe it will alleviate parents' concerns about their children becoming dehydrated while playing outside.

Refined Z score

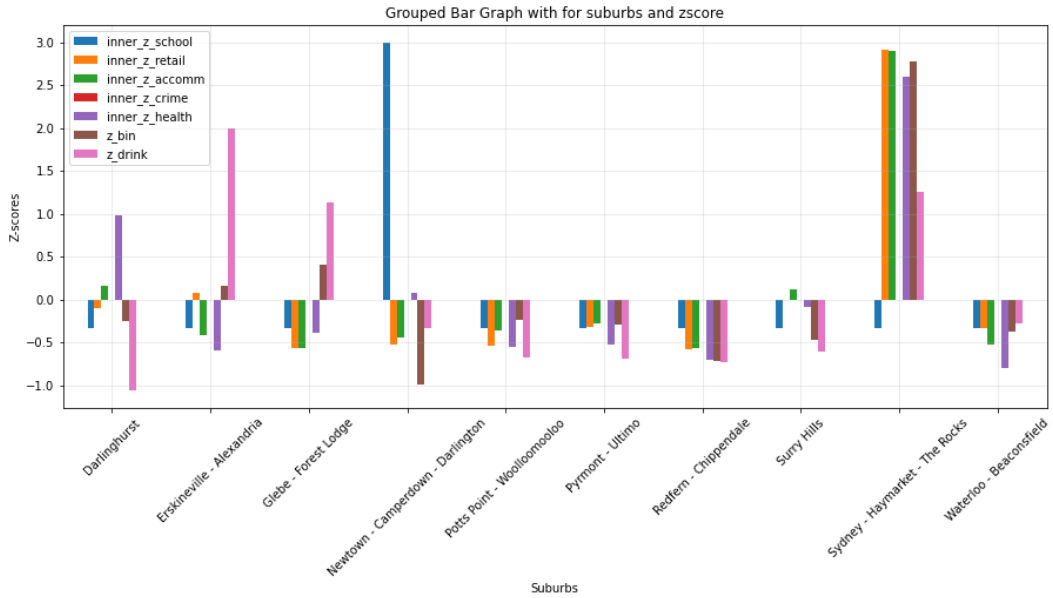
To appeal to our stakeholders, we narrowed the z-score to only include data from the inner city with our revised formula, rather than focusing on the entire greater capital city. This guaranteed that comparisons between inner-city suburbs were not influenced by external means and standard deviation.
Therefore we adjusted the score to be:

$$\text{Score} = S\,(Z_{school} + Z_{retail} + Z_{accomm} - Z_{crime} + Z_{health} + Z_{bin} + Z_{drinking})$$

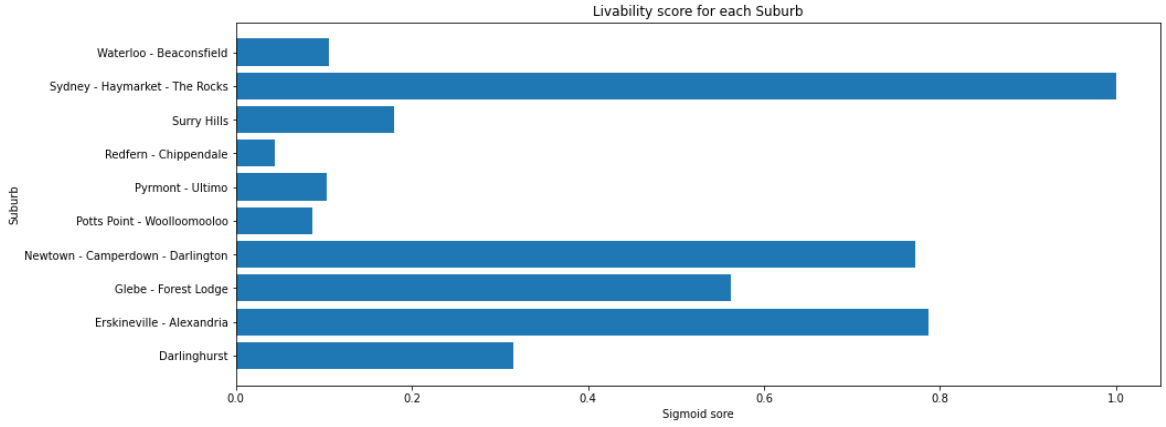
Where:

- Zbin represents the number of bins per 1000 people in the neighborhood
- Zdrinking represents the number of drinking fountains per 1000 people in the neighborhood

Charts



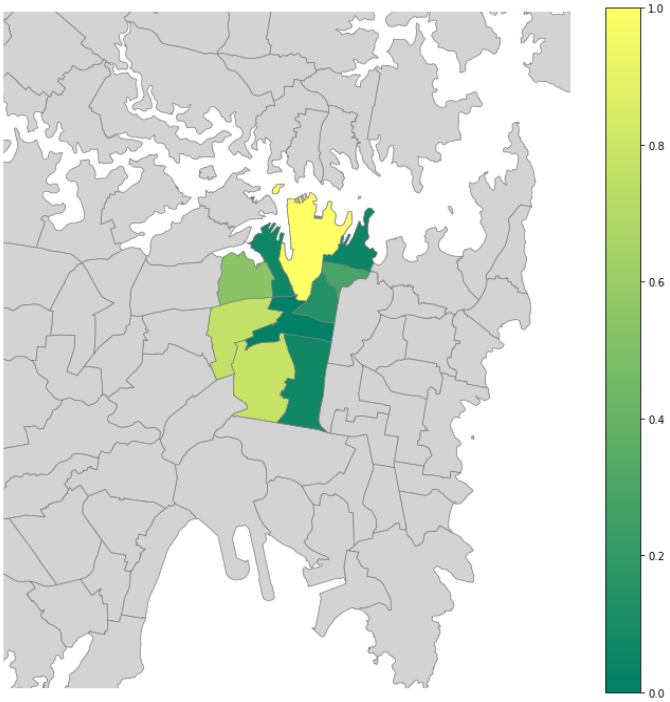
The z-scores determined by refining the search areas to the inner city of Sydney, including the z-score calculations for schools, retail, accommodation, crime, and health, are encoded in each of the colors of the bars. At first glance, the CBD region, which includes Sydney, Haymarket, and The Rocks, appears to have mostly favorable z-scores, making it one of the best choices for families. It does not, however, have a positive z-score for schools, which implies that if the family wishes to reside near their children's school, they need to look into alternative options. In contrast to the CBD, the majority of other suburbs have negative z-scores, implying that when comparing specific suburbs in the CBD, most locales provide fewer public services for residents. Furthermore, by limiting the z-score calculations to just the inner city, the chart indicates that the z-score for crime in all locations is zero, indicating that the record most likely lack some recording of crime hotspots, hence outside knowledge should be used when the family decides to relocate to central Sydney.



In the chart above, the Sigmoid score is stored by the bar lengths on the x-axis, whilst the suburb names are encoded by the individual bars in the figure above. The score for ten grouped suburbs in the inner city is shown in the bar chart. Some suburbs, such as Redfern, Chippendale, Potts Point, and Woolloomooloo

have low scores compared to other suburbs, indicating a smaller number of services for every thousand people. From the chart above, our stakeholders should be more inclined to choose Sydney, Haymarket, and The Rocks area as their place of residence. However, some may reconsider this area as there are low numbers of schools and some families with young children may prefer to live in areas with abundant educational resources.

City of Sydney Map



The color bar represents the sigmoid on the map, which depicts the inner city suburbs. The higher the sigmoid, the more yellow a region is, indicating a higher livability score. When we use our revised model, we can see that regions like Darlinghurst, Sydney, Haymarket, and The Rocks have extraordinarily high livability scores. On the other hand, Areas like Red Fern and Chippendale, have the lowest livability scores, which correlate with the lowest z-score. Furthermore, if the family chooses based on these criteria, they should avoid Pyrmont, Ultimo, Potts Point, and Woollooomo, as these areas are on the dark green scale.

Appendices

Appendix 1: Identifying extra cleaning on rows BuinessStats

```
i = BusinessStatsData[BusinessStatsData['area_name'].str.contains("Unknown")]
i
```

	area_id	area_name	accommodation_and_food_services	retail_trade	health_care_and_social_assistance
576	199999499	Currently Unknown (NSW)	238	419	258
1039	299999499	Currently Unknown (Vic.)	148	258	118
1568	399999499	Currently Unknown (Qld)	100	135	83
1741	499999499	Currently Unknown (SA)	31	35	29
1994	599999499	Currently Unknown (WA)	34	47	35
2094	699999499	Currently Unknown (Tas.)	11	5	3
2163	799999499	Currently Unknown (NT)	14	25	3
2295	899999499	Currently Unknown (ACT)	20	5	13
2300	999999499	Currently Unknown (Aus.)	7	80	39

Appendix 2: Sigmoid and Z-score calculations

	suburb	z_school	z_accomm	z_retail	z_crime	z_health	zscore	sigmoid
0	Acacia Gardens	-0.225784	-0.759114	0.160031	0.000000	-0.610528	-1.435395	0.192259
1	Arncliffe - Bardwell Valley	-0.225784	-0.072988	-0.180618	0.833253	-0.775117	-2.087759	0.110292
2	Ashcroft - Busby - Miller	-0.225784	-0.644599	-0.879738	-0.447085	-0.825852	-2.128887	0.106321
3	Ashfield	-0.225784	0.117469	-0.125762	-0.254238	-0.175108	-0.154948	0.461340
4	Asquith - Mount Colah	-0.225784	-0.626545	-0.393178	-0.700200	-0.487283	-1.032590	0.262582
...	...	...	...	...	...	...	...	...
294	Woy Woy - Blackwall	-0.225784	0.109574	-0.379859	1.266429	-0.764849	-2.527347	0.073963
295	Wyoming	-0.225784	-0.527358	-0.609285	-0.023166	-0.447649	-1.786910	0.143452
296	Wyong	-0.225784	1.058632	-0.315705	-0.508426	-0.618265	0.407304	0.600441
297	Yagoona - Birrong	-0.225784	-0.615950	-0.245182	-0.836500	-0.635966	-0.886380	0.291857
298	Yarramundi - Londonderry	-0.225784	-0.617350	-0.484660	0.000000	-0.977882	-2.305676	0.090654

299 rows × 8 columns