Sydney Liveability Analysis

DATA2901

Group F10-Adv-01-5

Pratul Singh Raghava	prag9782	510604305
Amanda Walpitage	awal8482	510657987

Dataset Description

Datasets used in the analysis are listed below.

- ***Lights.geojson***
- ***BusinessStats.csv***
- ***school_catchments.shp***
- ***SA2_2016_AUST.shp***
- ***Playgrounds.geojson***
- ***Neighbourhoods.csv***
- ***break_and_enter.shp***

In addition to the provided datasets, we used two additional datasets. The Playgrounds dataset provides data on all the playgrounds and outdoor fitness stations in the city of Sydney, and the Lights dataset provides data on lights controlled by the Sydney city council. These are GeoJSON files (files format that encode geo-spatial data) sourced from the City of Sydney Open Data Hub. Prior to analysis, all the data was transformed, cleaned and reduced to facilitate data integration and to improve the accuracy of our results.

Data Transformation and Reduction

To facilitate data integration, each dataset was stored into a pandas dataframe. Shape files and Geojson files were read using *geopandas* and *json* libraries into pandas dataframes.

As our project focused on the liveability in the Greater Sydney Area, a new dataframe with data comprising only the Greater Sydney area was created from SA2_2016_AUST dataset. Furthermore the X,Y coordinates in datasets were stored to POINT geometry type in order to apply the functions on geospatial data in POSTGIS.

Data Cleaning

Each data frame was tested for their quality especially for missing values. The data quality issues in each data frame was found to be minimal. Although several missing values were found in some datasets, none of those values were related to attributes which were used in our study except for the Neighbourhoods dataset in which we had to remove a couple of rows.

Moreover the population field in the neighbourhoods table had to be converted to integer type from string type, eliminating the commas used as decimal separators.

Database Description

Database Schema

The database tables were created to include only the interested attributes. Refer to the database schema diagram.

Database Indexes

As our analysis required results to be filtered out by Greater Sydney and Inner Sydney, indexes were created on fields SA2_NAME and SA3_NAME.

Also indexes on geom columns in break_and_enter and school_catchment_combined tables were created to make the spatial joins faster in the system.

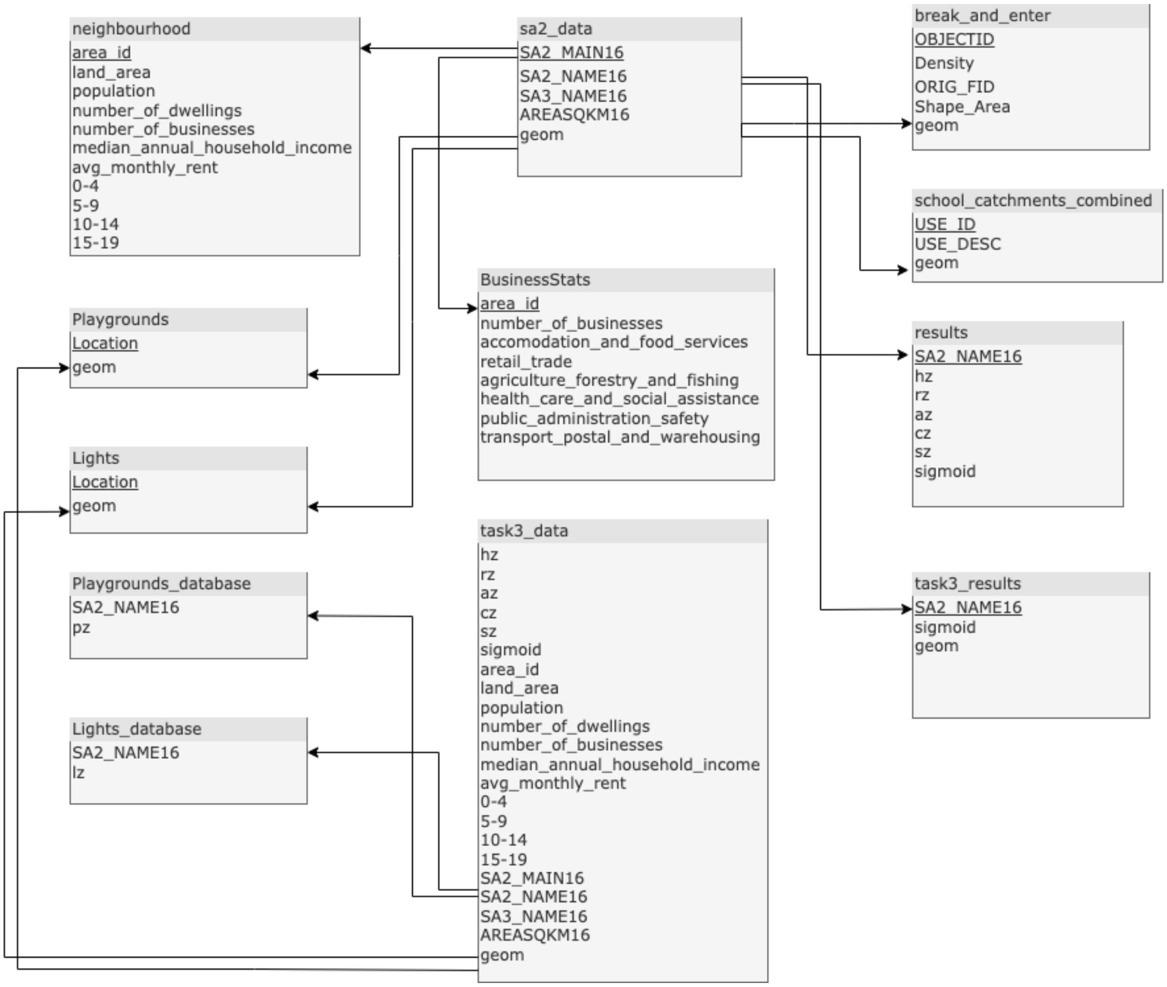


Fig 1 : The schema diagram for the database

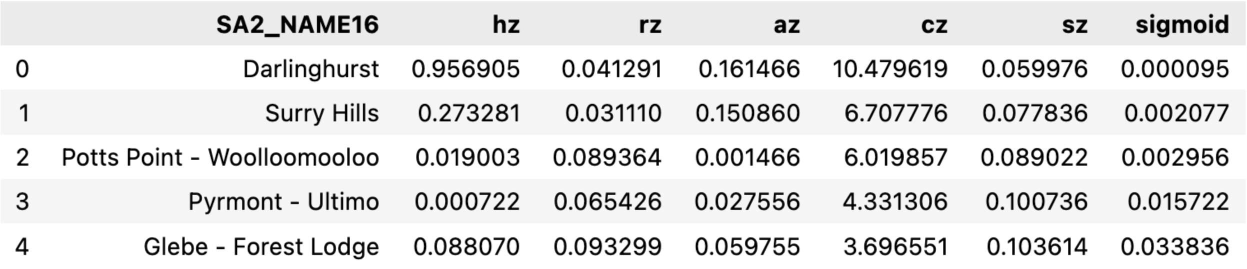
Greater Sydney Score Analysis

The formula used :

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

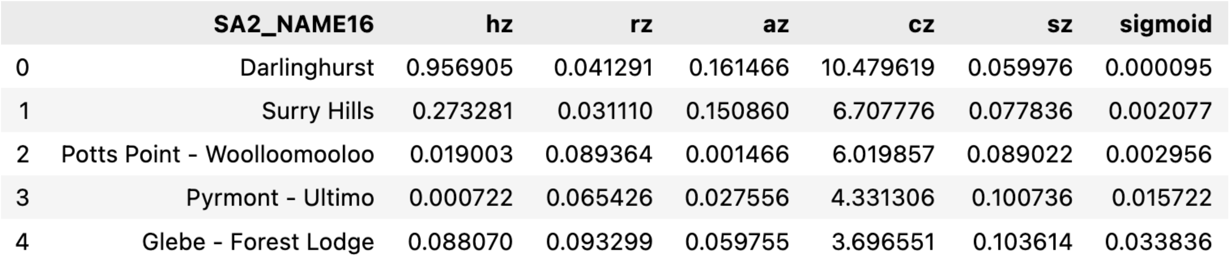
Our results discovered ***Banksmeadow*** as the most liveable suburb as it ranked first in 3/5 measures used in our liveability formula .Moreover it was discovered that liveability score of the suburbs **Darlinghurst** was the lowest due to a higher crime score.

It should be mentioned that the crime measure in our liveability formula skewed the overall liveability score significantly. This resulted in our results being quite different from official liveability indexes such as The Urban Living Index. However, it is to be noted that among the 20 parameters which the above mentioned index uses, crime data is not one of them, therefore, it can be said that it does not paint a full picture of the actual liveability score.



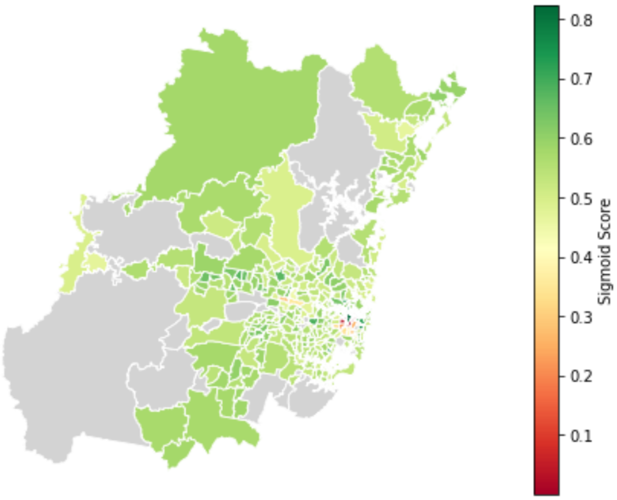
	SA2_NAME16	hz	rz	az	cz	sz	sigmoid
0	Darlinghurst	0.956905	0.041291	0.161466	10.479619	0.059976	0.000095
1	Surry Hills	0.273281	0.031110	0.150860	6.707776	0.077836	0.002077
2	Potts Point - Woolloomooloo	0.019003	0.089364	0.001466	6.019857	0.089022	0.002956
3	Pyrmont - Ultimo	0.000722	0.065426	0.027556	4.331306	0.100736	0.015722
4	Glebe - Forest Lodge	0.088070	0.093299	0.059755	3.696551	0.103614	0.033836

Table 1 - Top 5 suburbs in Greater Sydney in terms of liveability



	SA2_NAME16	hz	rz	az	cz	sz	sigmoid
0	Darlinghurst	0.956905	0.041291	0.161466	10.479619	0.059976	0.000095
1	Surry Hills	0.273281	0.031110	0.150860	6.707776	0.077836	0.002077
2	Potts Point - Woolloomooloo	0.019003	0.089364	0.001466	6.019857	0.089022	0.002956
3	Pyrmont - Ultimo	0.000722	0.065426	0.027556	4.331306	0.100736	0.015722
4	Glebe - Forest Lodge	0.088070	0.093299	0.059755	3.696551	0.103614	0.033836

Table 2 - Bottom 5 suburbs in Greater Sydney in terms of liveability



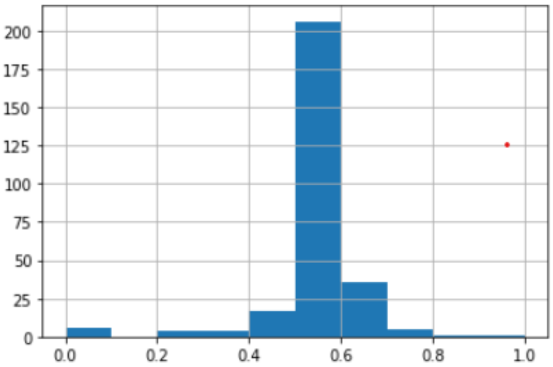
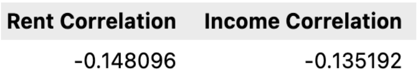


Figure 1- Visual Representation of Greater Sydney Liveability Score per Each Suburb

Figure 2 - Greater Sydney Liveability Score's Uneven Distribution

Correlation Analysis

Our score doesn't correlate at all with the *median rent* and *median income of each neighbourhood*. This was determined by calculating the correlation coefficient of our score with the above parameters. A near zero value of the coefficient calculated by both python and SQL separately, indicated no correlation at all.



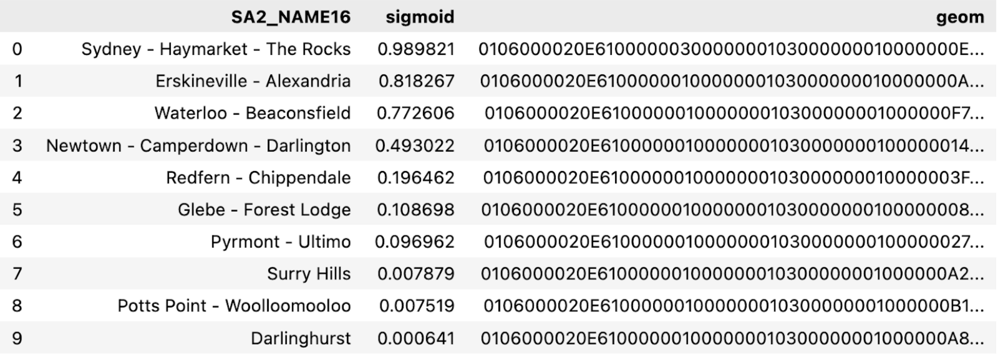
Rent Correlation	Income Correlation
-0.148096	-0.135192

City of Sydney Analysis

This part of the analysis was performed to assess the suitability of suburbs within the City of Sydney for our **Stakeholder : a couple with a young child**. Recreation opportunities, schools and safety around the neighbourhood are considered as additional requirements because of the importance they have on physical and emotional health of the child and parents.Therefore, the formula used in this task gives more preference to schools, health and safety than other parameters. The additional datasets, Lights and Playgrounds, were used in this context. The liveability score formula used in the previous task was modified by adding new measures $Zplaygrounds$ and $Zlights$.

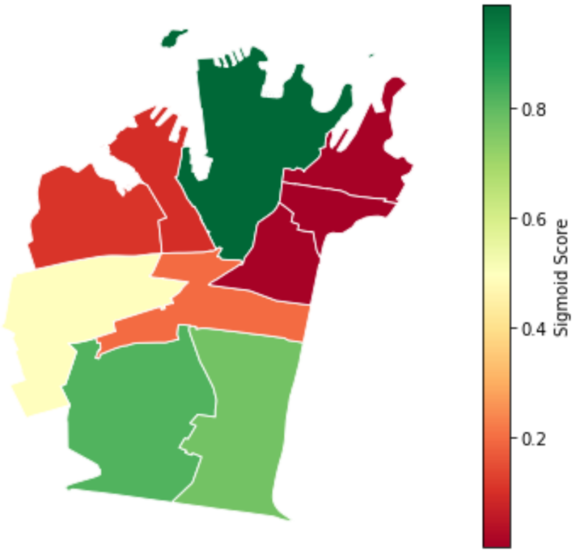
Score = Sigmoid(2*Zschool + 2*Zaccomm + 0.5*Zretail − Zcrime + 2*Zhealth + Zplaygrounds + Zlights)

According to the results, **Sydney Haymarket - The Rocks** followed by **Erkineville-Alexandria** and **Waterloo-Beaconsfield** are the top suburbs in the city that are highly desirable for our stakeholder. This would be our recommendation for them to live in.



	SA2_NAME16	sigmoid	geom
0	Sydney - Haymarket - The Rocks	0.989821	0106000020E6100000030000000103000000010000000E...
1	Erskineville - Alexandria	0.818267	0106000020E6100000010000000103000000010000000A...
2	Waterloo - Beaconsfield	0.772606	0106000020E610000001000000010300000001000000F7...
3	Newtown - Camperdown - Darlington	0.493022	0106000020E61000000100000001030000000100000014...
4	Redfern - Chippendale	0.196462	0106000020E6100000010000000103000000010000003F...
5	Glebe - Forest Lodge	0.108698	0106000020E61000000100000001030000000100000008...
6	Pyrmont - Ultimo	0.096962	0106000020E61000000100000001030000000100000027...
7	Surry Hills	0.007879	0106000020E610000001000000010300000001000000A2...
8	Potts Point - Woolloomooloo	0.007519	0106000020E610000001000000010300000001000000B1...
9	Darlinghurst	0.000641	0106000020E610000001000000010300000001000000A8...

Table 3 - Results of Inner Sydney Liveability Scores



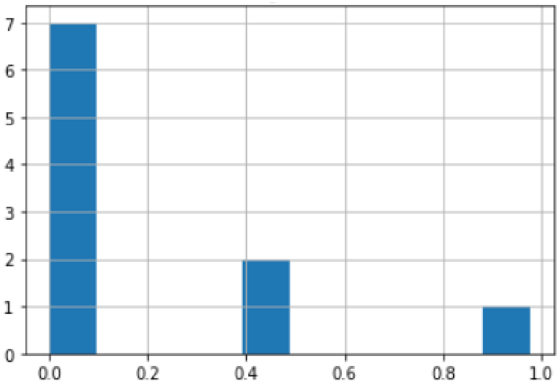


Figure 3- Visual representation of Inner Sydney liveability Score per Each Suburb

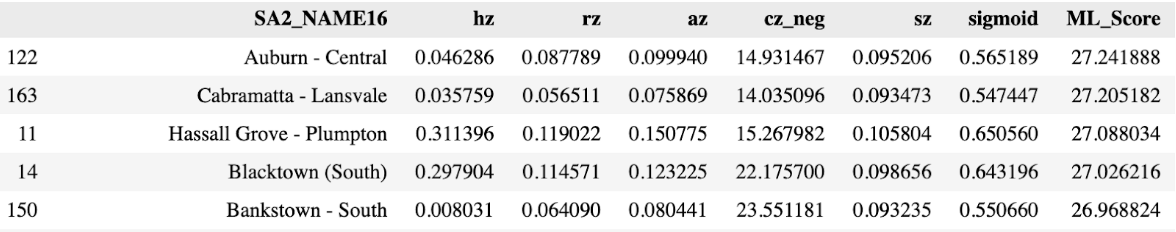
Figure 4 - Inner Sydney Liveability Score's Uneven Distribution

Greater Sydney Score Analysis using ML

Liveability of Greater Sydney was analysed again with the aid of an unsupervised Machine Language Algorithm : TSNE. We opted against using a supervised algorithm, since we would have to train it on our Task-2 results and thus, the algorithm would make the same potential mistakes made in Task-2 and would therefore be of no real use.

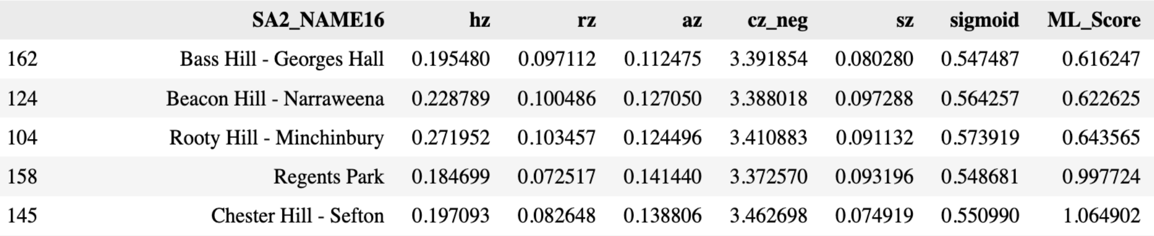
Interestingly the results from the equation used in Task-2 previously and the results from the machine learning analysis are quite different.

In our ML analysis, along with the other parameters, we used a *cz_neg* parameter instead of the *cz* parameter used in the above tasks. The *cz_neg* parameter was just a score inversely proportional to the crime score. We used this to avoid the heavy skewing of results by the crime scores and doing so actually produced results which were far closer to [The Urban Living Index](), than the earlier analysis. Therefore, the results produced by the Unsupervised Machine Learning Algorithm can be said to be much better than the formula used in Task-2 Analysis of Greater Sydney.



	SA2_NAME16	hz	rz	az	cz_neg	sz	sigmoid	ML_Score
122	Auburn - Central	0.046286	0.087789	0.099940	14.931467	0.095206	0.565189	27.241888
163	Cabramatta - Lansvale	0.035759	0.056511	0.075869	14.035096	0.093473	0.547447	27.205182
11	Hassall Grove - Plumpton	0.311396	0.119022	0.150775	15.267982	0.105804	0.650560	27.088034
14	Blacktown (South)	0.297904	0.114571	0.123225	22.175700	0.098656	0.643196	27.026216
150	Bankstown - South	0.008031	0.064090	0.080441	23.551181	0.093235	0.550660	26.968824

Table 4 - Top 5 most liveable suburbs in Greater Sydney per ML algorithm



	SA2_NAME16	hz	rz	az	cz_neg	sz	sigmoid	ML_Score
162	Bass Hill - Georges Hall	0.195480	0.097112	0.112475	3.391854	0.080280	0.547487	0.616247
124	Beacon Hill - Narraweena	0.228789	0.100486	0.127050	3.388018	0.097288	0.564257	0.622625
104	Rooty Hill - Minchinbury	0.271952	0.103457	0.124496	3.410883	0.091132	0.573919	0.643565
158	Regents Park	0.184699	0.072517	0.141440	3.372570	0.093196	0.548681	0.997724
145	Chester Hill - Sefton	0.197093	0.082648	0.138806	3.462698	0.074919	0.550990	1.064902

Table 5 - Top 5 least liveable suburbs in Greater Sydney per ML algorithm

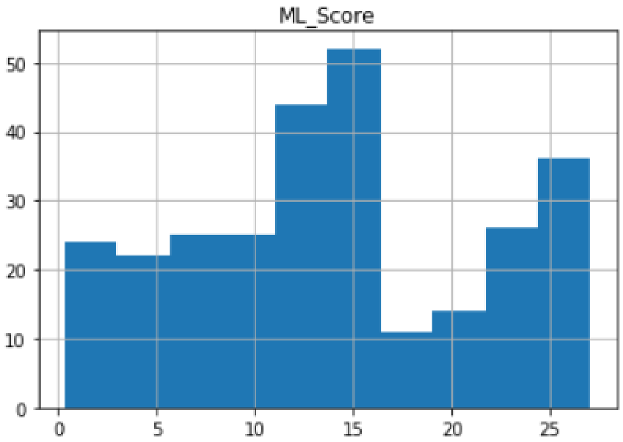


Figure 5 - Greater Sydney Liveability score distribution computed with TSNE ML algorithm

Bibliography

1. Playgrounds.City of Sydney Open Data Hub.(2020).Retrieved 18 May 2022, from https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::playgrounds/about

2. Lights. City of Sydney Open Data Hub.(2020). Retrieved 18 May 2022, from https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::lights/about

3. Urban Living Index. (2022). Retrieved 18 May 2022, from https://urbanlivingindex.com/