DATA2001 REPORT (RE08 - 13)

Team Members (SID): 500521681,500497113,500477034

1. Introducion

The purpose of this report is to investigate and analyse for Sydney residents which are the most 'livable' suburbs to buy in Sydney and which Sydney suburbs have the best environment.

We aggregated multiple datasets and calculated a 'liveability' score (Sigmoid score) for each Sydney suburb to reveal the most liveable suburbs. Subsequent analysis and investigation of the environment by adding data on "Greenhouse_gas_emissions_profile_by_suburb" and "Waste_generation_by_suburb".

2. Dataset description

2.1. Sources:

- Both Neighborhoods.csv and BusinessStats.csv are derived from Australian Bureau of Statistics (ABS) census data.
- SA2_2016_AUST.zip is derived from the Australian Bureau of Statistics (ABS) and associated parent areas.
- break_and_enter.zip is sourced from BOCSAR.
- school_catchments.zip is sourced from NSW Department of Education.
- Greenhouse_gas_profile.geojson.geojson from Utilities and Government Statistics.
- Waste_generation.geojson from City of Sydney local government area Statistics.

2.2. Data content:

• Neighborhoods.csv:

The data includes housing statistics for different parts of Australia, including information on the number of dwellings, population by age group and average monthly rent.

• BusinessStats.csv:

The dataset contains business statistics for different regions of Australia, including the number of businesses, restaurants, retail and more.

• SA2_2016_AUST.shp:

Contains a broad range of detailed statistics including social, demographic and economic statistics, separated and labelled by the concept of functional area.

- **break_and_enter.shp:**

A map containing statistics regarding the hotspots and intensity of crime in suburbs of NSW.

- **school_catchments.zip:**

This document contains datasets for 3 different types of schools, all of which are non-government schools.

- **Greenhouse_gas_profile.geojson**

This GeoJOSN data contains a profile of greenhouse gas emissions by suburb from 2005 to 2019, including emissions from electricity, natural gas, transportation, and more.

- **Waste_generation.geojson**

This GeoJOSN data contains the amount of waste generated by suburbs from 2005 to 2019.

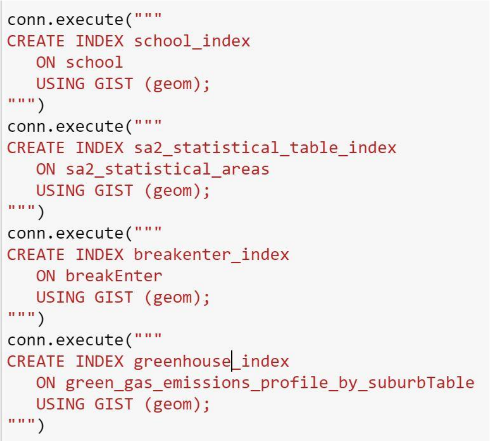
2.3. Data cleaning:

For the csv datasets Neighborhoods.csv and BusinessStats.csv we used **drop_duplicates()** to clean up the duplicates. After that, we joined the 2 csv files with another 3 shape files and changed all Nan values to zero.

For the extra dataset (2 geojson files), since we only surveyed data from 2016-2017, we also performed cleaning on these numerical data to change the Nan values to zero.

3. Database description

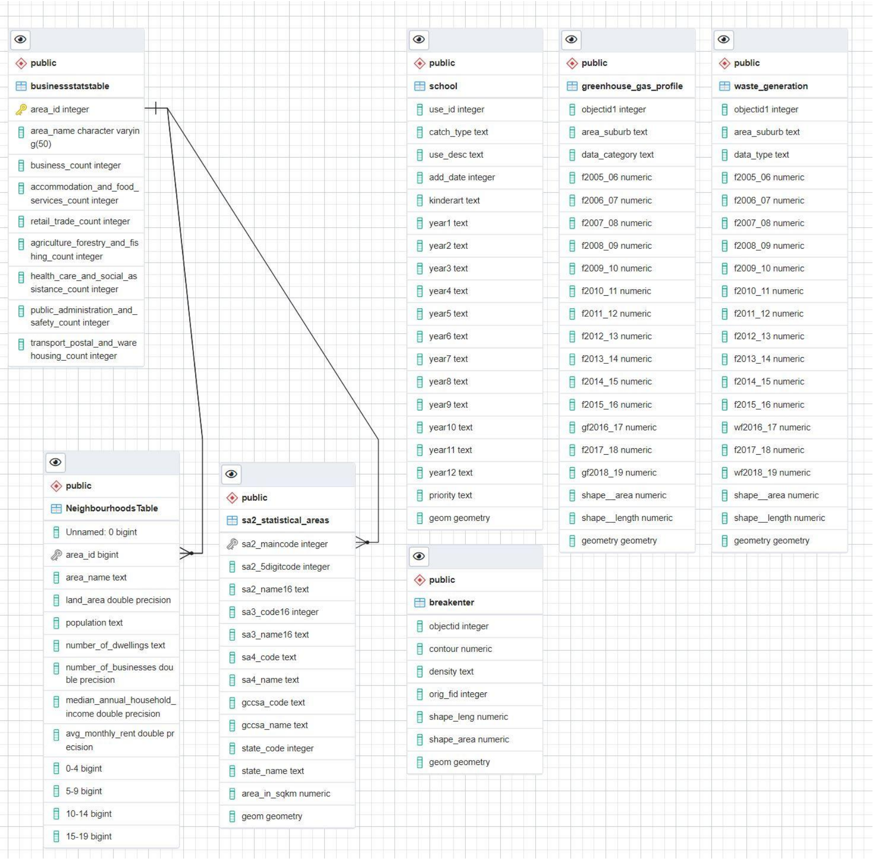
In the following section, we will demonstrate how we utilised schemas to organise our databases and how we interconnected them through well defined constraints. Additionally, we'll also show how we implemented an index system to speed up our runtime and queries.



```
conn.execute("""
CREATE INDEX school_index
   ON school
   USING GIST (geom);
""")
conn.execute("""
CREATE INDEX sa2_statistical_table_index
   ON sa2_statistical_areas
   USING GIST (geom);
""")
conn.execute("""
CREATE INDEX breakenter_index
   ON breakEnter
   USING GIST (geom);
""")
conn.execute("""
CREATE INDEX greenhouse_index
   ON green_gas_emissions_profile_by_suburbTable
   USING GIST (geom);
""")
```

We created indexes for each of the databases to speed up each of their searches/queries.

• The Database Schema Diagram



public
businessstatstable
area_id integer
area_name character varying(50)
business_count integer
accommodation_and_food_services_count integer
retail_trade_count integer
agriculture_forestry_and_fishing_count integer
health_care_and_social_assistance_count integer
public_administration_and_safety_count integer
transport_postal_and_warehousing_count integer

public
NeighbourhoodsTable
Unnamed: 0 bigint
area_id bigint
area_name text
land_area double precision
population text
number_of_dwellings text
number_of_businesses double precision
median_annual_household_income double precision
avg_monthly_rent double precision
0-4 bigint
5-9 bigint
10-14 bigint
15-19 bigint

public
sa2_statistical_areas
sa2_maincode integer
sa2_5digitcode integer
sa2_name16 text
sa3_code16 integer
sa3_name16 text
sa4_code text
sa4_name text
gccsa_code text
gccsa_name text
state_code integer
state_name text
area_in_sqkm numeric
geom geometry

public
school
use_id integer
catch_type text
use_desc text
add_date integer
kinderart text
year1 text
year2 text
year3 text
year4 text
year5 text
year6 text
year7 text
year8 text
year9 text
year10 text
year11 text
year12 text
priority text
geom geometry

public
breakenter
objectid integer
contour numeric
density text
orig_fid integer
shape_leng numeric
shape_area numeric
geom geometry

public
greenhouse_gas_profile
objectid1 integer
area_suburb text
data_category text
f2005_06 numeric
f2006_07 numeric
f2007_08 numeric
f2008_09 numeric
f2009_10 numeric
f2010_11 numeric
f2011_12 numeric
f2012_13 numeric
f2013_14 numeric
f2014_15 numeric
f2015_16 numeric
gf2016_17 numeric
f2017_18 numeric
gf2018_19 numeric
shape__area numeric
shape__length numeric
geometry geometry

public
waste_generation
objectid1 integer
area_suburb text
data_type text
f2005_06 numeric
f2006_07 numeric
f2007_08 numeric
f2008_09 numeric
f2009_10 numeric
f2010_11 numeric
f2011_12 numeric
f2012_13 numeric
f2013_14 numeric
f2014_15 numeric
f2015_16 numeric
wf2016_17 numeric
f2017_18 numeric
wf2018_19 numeric
shape__area numeric
shape__length numeric
geometry geometry

- In order to better define the relationship between our tables, we chose to add the area_id in the businessstatstables as our unique primary key, so that foreign keys could be established within NeighbourhoodsTable and sa2_statistical_areas tables, such that referential integrity throughout our main tables could be maintained.

4. Greater Sydney Score Analysis

We calculate the liveability score for all given neighbourhoods by summing up each of the z scores other than crime and then subtracting the Z score of crime according to the definition given by the assignment specification. The formula for the overall score are given as follows:

$$Score = S(z_{school} + z_{accommodation} + z_{retail} - z_{crime} + z_{health})$$

Assume the variable “young people” and “population” we are given have already divided by 1000(as required by the assignment spaces and this has been computed in preprocessing steps), we can calculate all the corresponding z scores using the method indicated by McLeod [1] in her online article:

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

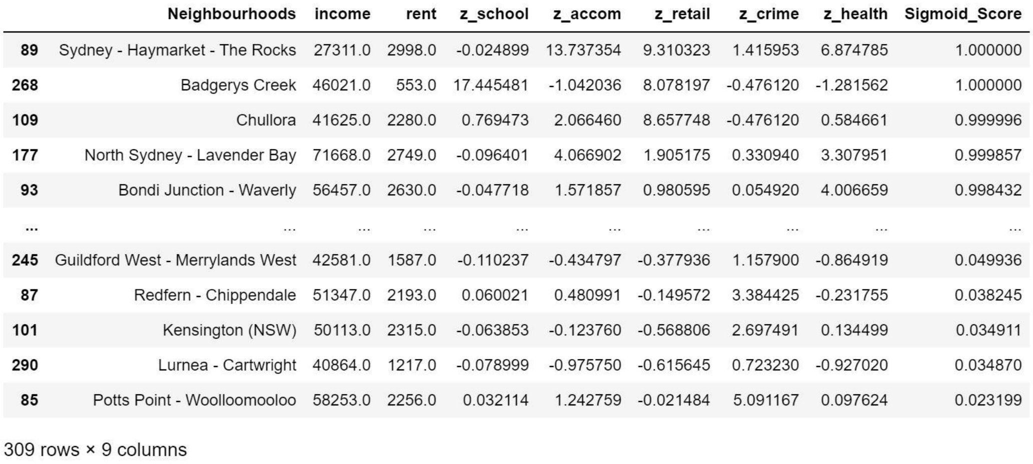
After we have computed z_{school}, $z_{accommodation}$, z_{retail}, z_{crime}, z_{health}, we can apply sigmoid’s function to combine multiple z-scores into a single z-score value as our overall liveability score. A mathematical representation of the function can be seen as follows:

$$S(x) = \frac{1}{1 + e^{-x}}$$

Where x is the combination of all z-score, which can be represented by:

$$(z_{school} + z_{accommodation} + z_{retail} - z_{crime} + z_{health})$$

By applying the method above we have generated a table that provides an overview of all the corresponding z-scores and our livability score is attached as the last column.

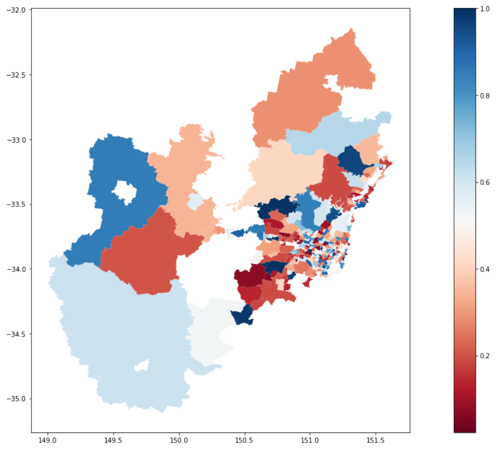


	Neighbourhoods	income	rent	z_school	z_accom	z_retail	z_crime	z_health	Sigmoid_Score
89	Sydney - Haymarket - The Rocks	27311.0	2998.0	-0.024899	13.737354	9.310323	1.415953	6.874785	1.000000
268	Badgerys Creek	46021.0	553.0	17.445481	-1.042036	8.078197	-0.476120	-1.281562	1.000000
109	Chullora	41625.0	2280.0	0.769473	2.066460	8.657748	-0.476120	0.584661	0.999996
177	North Sydney - Lavender Bay	71668.0	2749.0	-0.096401	4.066902	1.905175	0.330940	3.307951	0.999857
93	Bondi Junction - Waverly	56457.0	2630.0	-0.047718	1.571857	0.980595	0.054920	4.006659	0.998432
...	...	...	...	...	...	...	...	...	...
245	Guildford West - Merrylands West	42581.0	1587.0	-0.110237	-0.434797	-0.377936	1.157900	-0.864919	0.049936
87	Redfern - Chippendale	51347.0	2193.0	0.060021	0.480991	-0.149572	3.384425	-0.231755	0.038245
101	Kensington (NSW)	50113.0	2315.0	-0.063853	-0.123760	-0.568806	2.697491	0.134499	0.034911
290	Lurnea - Cartwright	40864.0	1217.0	-0.078999	-0.975750	-0.615645	0.723230	-0.927020	0.034870
85	Potts Point - Woolloomooloo	58253.0	2256.0	0.032114	1.242759	-0.021484	5.091167	0.097624	0.023199

309 rows × 9 columns

The result shows that the z_{retail} plays a substantial part in computing our final livability score. When we rank all the neighbourhoods by their individual livability score in decreasing order we see that the top 3 neighbourhoods all have a very high z_{retail} whereas the z_{retail} of the 5 neighbourhoods at the bottom are all negative. Meanwhile, we see that suburbs with a larger land size usually benefit from the crime score as the sum of hotspot areas often only take a relatively small portion of its land size, hence its crime score will become lower compared to the suburbs with the same size of hotspot areas but a smaller land size. It is also notable that the population density greatly affects the outcome of calculation, as 4 of the 5 z scores involve the population variable. Population, in our calculation, is mostly negatively correlated with the z-scores as calculations largely involve dividing the population density (e.g. per 1000 people). With the table shown below that is generated by tabling our dataframe, the importance of population density and z_{retail} can be argued in general.

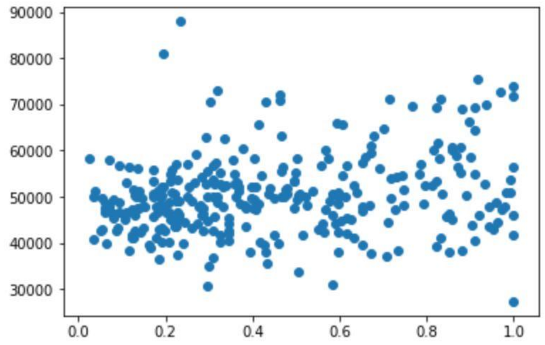
• The Graph Visualisation



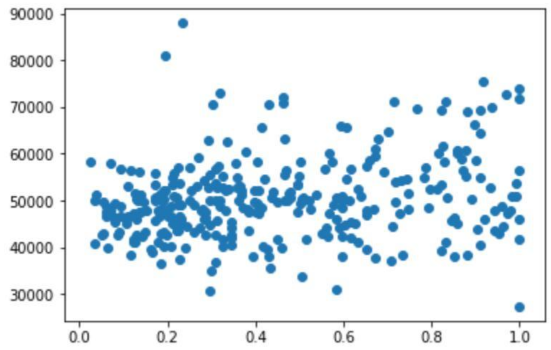
(The deep the blue, the higher the livability score, while the deep the red, the lower the livability score)

5. Correlation Analysis

cor(score,income) = 0.22



cor(score,rent) = 0.44



From the result displayed by the table, Our liveability score is positively correlated with rent and income. Meanwhile, the correlation between our livability score and rent is higher than the correlation between our livability score and income.

6. City of Sydney Analysis

Our stakeholders are the Australian residents who do not currently live in Sydney but are planning to live in the city of Sydney for a set period of time.

We are using Greenhouse_gas_emissions_profile_by_suburb.geojson and Waste_generation_by_suburb.geojson to propose a recommendation on which suburb they should be looking at in terms of renting or long-term stay based on the waste generation by suburb and Greenhouse gas emissions profile by suburb.

We calculat the z score individually for greenhouse gas emissions and waste generations using the same method we have used in section 5 which is $Z = \frac{x - \mu}{\sigma}$

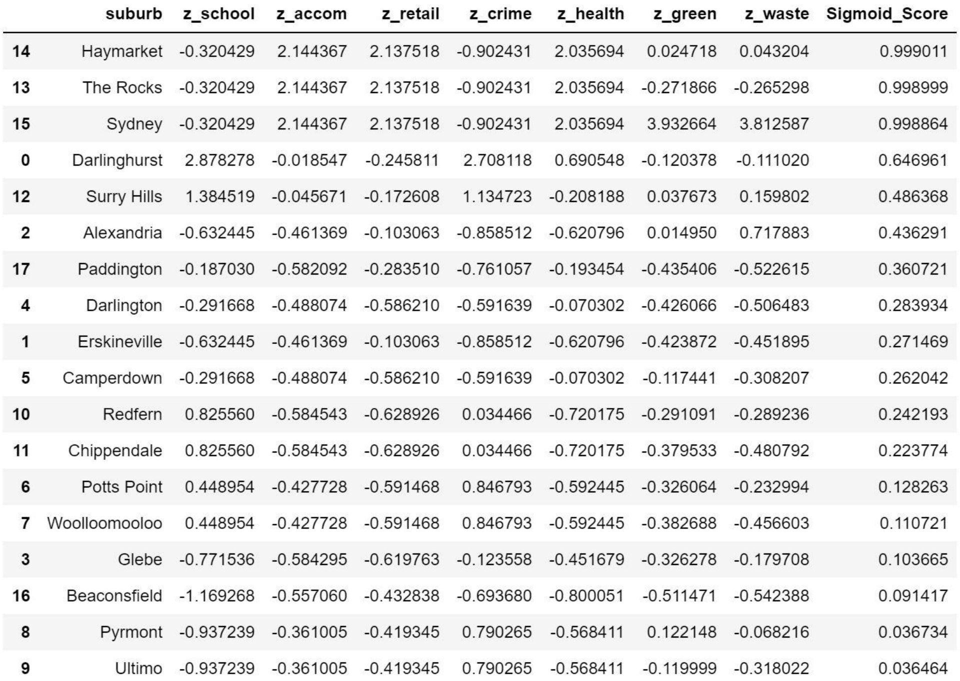
For greenhouse gas emissions, x = electricity + gas + other scope 3 + transport + water waste + waste. To avoid too much NA involved when selecting the latest data(2018-2019) as shown below, we choose to use data from 2016 to 2017 so our analysis is complete and none of the suburbs involved in the z score calculation is disadvantaged.

For Waste generations, x = C&I discposal + C&I recycling + MSW disposal + MSW recycling and we use data from 2016 to 2017 to coordinate with the z score of greenhouse gas emissions. The formula to calculate the z score will also be the same.

After we have obtained $z_{greenhouse}$,and z_{waste}, we will update our livability score formula as following to measure the livability of inner city suburbs:

$$Score = S(z_{school} + z_{accommodation} + z_{retail} - z_{crime} + z_{health} - z_{greenhouse} + z_{waste})$$

The analysis has shown that amongst all inner suburbs, Haymarket has the highest livability score after we take $z_{greenhouse}\ and\ z_{waste}$ into account whereas Ultimo has the lowest livability score.

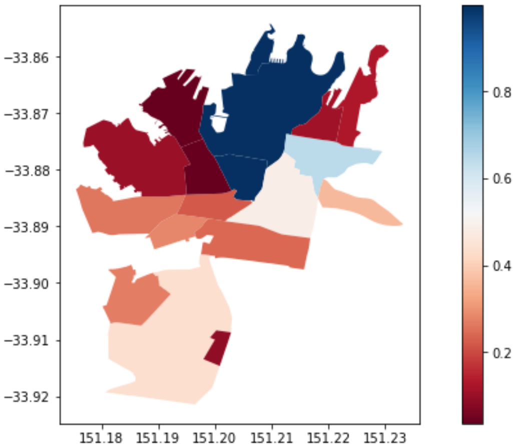


	suburb	z_school	z_accom	z_retail	z_crime	z_health	z_green	z_waste	Sigmoid_Score
14	Haymarket	-0.320429	2.144367	2.137518	-0.902431	2.035694	0.024718	0.043204	0.999011
13	The Rocks	-0.320429	2.144367	2.137518	-0.902431	2.035694	-0.271866	-0.265298	0.998999
15	Sydney	-0.320429	2.144367	2.137518	-0.902431	2.035694	3.932664	3.812587	0.998864
0	Darlinghurst	2.878278	-0.018547	-0.245811	2.708118	0.690548	-0.120378	-0.111020	0.646961
12	Surry Hills	1.384519	-0.045671	-0.172608	1.134723	-0.208188	0.037673	0.159802	0.486368
2	Alexandria	-0.632445	-0.461369	-0.103063	-0.858512	-0.620796	0.014950	0.717883	0.436291
17	Paddington	-0.187030	-0.582092	-0.283510	-0.761057	-0.193454	-0.435406	-0.522615	0.360721
4	Darlington	-0.291668	-0.488074	-0.586210	-0.591639	-0.070302	-0.426066	-0.506483	0.283934
1	Erskineville	-0.632445	-0.461369	-0.103063	-0.858512	-0.620796	-0.423872	-0.451895	0.271469
5	Camperdown	-0.291668	-0.488074	-0.586210	-0.591639	-0.070302	-0.117441	-0.308207	0.262042
10	Redfern	0.825560	-0.584543	-0.628926	0.034466	-0.720175	-0.291091	-0.289236	0.242193
11	Chippendale	0.825560	-0.584543	-0.628926	0.034466	-0.720175	-0.379533	-0.480792	0.223774
6	Potts Point	0.448954	-0.427728	-0.591468	0.846793	-0.592445	-0.326064	-0.232994	0.128263
7	Woolloomooloo	0.448954	-0.427728	-0.591468	0.846793	-0.592445	-0.382688	-0.456603	0.110721
3	Glebe	-0.771536	-0.584295	-0.619763	-0.123558	-0.451679	-0.326278	-0.179708	0.103665
16	Beaconsfield	-1.169268	-0.557060	-0.432838	-0.693680	-0.800051	-0.511471	-0.542388	0.091417
8	Pyrmont	-0.937239	-0.361005	-0.419345	0.790265	-0.568411	0.122148	-0.068216	0.036734
9	Ultimo	-0.937239	-0.361005	-0.419345	0.790265	-0.568411	-0.119999	-0.318022	0.036464

People from other places in Australia can therefore choose the most favourable inner-city suburb in Sydney for their visit according to the final sigmoid_score, e.g. Haymarket, The Rocks, Sydney all share a similar high Sigmoid score due to its convenience and massive number of accommodations. Choosing to live in those suburbs should be a prior choice over choosing to live in suburbs on the bottom of the lists(e.g. Pyrmont,Ultimo).

• The Graph Visualisation

We visualised the score using the shape file of Waste_generation and the previously calculated sigmoid score:



(same interpretation as 4)

7. References

[1]McLeod, S. A. (2019, May 17). ***Z-score: definition, calculation and interpretation*****. Simply Psychology. www.simplypsychology.org/z-score.html**

Greenhouse gas emissions profile by suburb from CITY OF SYDNEY: "https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::greenhouse-gas-emissions-profile-by-suburb-1/explore?filters=eyJGMjAwNl8wNyI6WzEwLjA0NzM1NTU2LDIxNzQ0OTguMTE3XX0%3D&location=-33.888930%2C151.203975%2C13.56&showTable=true"

Waste generation by suburb from CITY OF SYDNEY: "https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::waste-generation-by-suburb/explore?location=-33.888925%2C151.203975%2C13.56&showTable=true"