2022

SYDNEY LIVEABILITY ANALYSIS

PREPARED BY

Dawei Yun

Zirong Wen

ASSIGNMENT GROUP

F12D-RE08 - 10

Dataset Description

Neighbourhoods.CSV

This dataset is from the Canvas Data2001 page, and it provides information about the neighbourhoods in Australia. It includes area code, area name, population, income, rent, and youth population.

BusinessSrats.CSV

This dataset is from the Canvas Data2001 page, and it provides information about the business statistics data in Australia. These include area codes, area names, and a range of statistics on the local economy.

SA2_2016_AUST.zip

This dataset is from Canvas Data2001 Page, and it provides information about Australian administrative region data. This includes the region code and the name of each region, and the geometry of the area.

break_and_enter.zip

This dataset is from Canvas Data2001 Page, and it provides information about Australian crime data, including crime id, areas affected and their geometry.

school_catchments.zip

This dataset is from Canvas Data2001 Page, and it provides information about Australian schools, Including primary school, secondary school, and future school statistics, as well as their geometric data.

Wayfinding_signage.SHP

This dataset is from the City of Sydney Datahub("Wayfinding signage", 2022), and it provides information about the landmark statistics of Sydney Inner City and their geometric Data.

Raingardens.SHP

This dataset is from the City of Sydney Datahub("Rain gardens", 2022), and it provides information about the garden statistics of Sydney Inner City. And their geometric data.

Data Preprocessing

In order to do our research to be more accurate and efficient, we need to pre-process the data. In CSV file neighbourhood.csv and BusinessStats.csv, we define a 'clean' function to replace the invalid value in these two datasets with null. After that, in the neighbourhoods.csv dataset, we created a new column called 'young' to store the total number of young people from 0-19. Finally, we use the drop function to remove the column we do not need, and upload processed data 'neighbourhoods' and' Business' to the database.
We use 'geopandas' to read shapefiles and use CRS to check their EPSG code. The next step

is to unify their coordinate system information. We unify all shapefiles to EPSG = 4283(GDA94) for future calculation. For the shapefiles having geometry types ‘polygon’ and ‘multpolygon’. We define a function called 'clean_geom', which can change geom_type from polygon to multipolygon. For the shapefiles with geometric data type ‘point’, we convert it to well-known Text format to ensure that geopandas is the same type as PostGIS expects.

In SA2_2016_AUST.zip, we remove all the null values and only keep the row 'GCC_NAME16' equal to 'Greater Sydney'. This could make the rest of our calculations more efficient. The school_catchments.zip contains three shapefile, primary school, secondary school, and future school because their composition is almost the same, so we merge them and create a new table 'school'.

For all the shapefile SA2_2016_AUST.zip, school_catchments.zip, break_and_enter.zip, wayfinding_signage.shp and raingardens.shp, we use the drop function to delete the lines we don't need and upload them to the database, and they are ‘sa2’, 'school', 'Breakk', 'signs' and ‘garden' respectively.

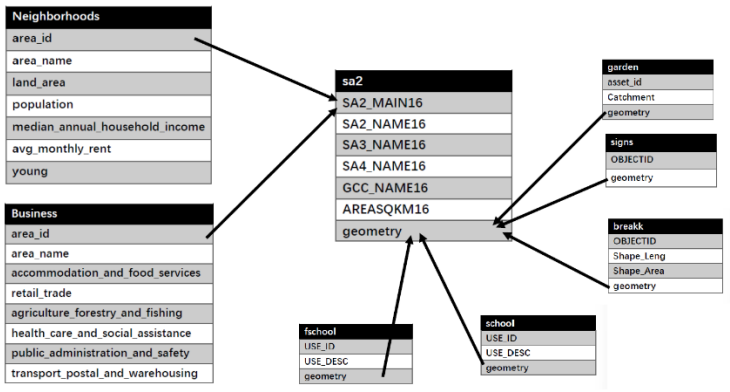
Outliers

Everything is not perfect, and so does the dataset. We found many outliers in our dataset. In businessStats.csv, there is a place called Sydney-Haymarket-the Rocks, where all the values are significantly higher than the mean.

At first, we first tried to delete the outliers with Interquartile Range, but it didn't work. Because there is a wide variation in the data from place to place, it deletes too many rows. So, we tried to delete outliers higher than four times its mean, which is also not good. It directly led to a significant deviation in the result. The highest livability score was changed to the lowest, which was unreasonable. We then tried to replace the outliers with median values, but the results were unsatisfactory.

After reflection, we believe it is unfair to delete and replace the value; If we remove the outliers, the region will not be counted in the formula, which is unjust to the area. As part of central Sydney, it makes sense that The Rocks has more business than any other place. With resignation, we chose to keep the outliers.

Database Schema



The data set Neighbourhoods and Business can join the SA2 file using the primary key 'area_id'. The rest of the data sets are shapefiles, which gives them the 'geometry' column as foreign

keys. The geometric data of the 'geometry' column can be connected through spatial Join.

Index

1. Spatial index area_idx on 'geometry' column in sa2 file
2. Normal index suburb_idx on 'SA2_MAIN16' column in sa2 file

When we use a sa2 dataset, these two indexes can speed up our data retrieval speed and reduce the grouping and sorting time, significantly reducing our query time.

Greater Sydney Livability Analysis

This section will analyse Greater Sydney's livability score using the described data set. Also, present the correlation score of livability with median income and average monthly rent.

The formula for calculating Livability score is:

$$Score = Sigmoid(z_{school} + z_{accomm} + z_{retail} - z_{crime} + z_{health})$$

Z-school

The school score represents the number of school catchment areas per 1000 'young people', where young people are 0-19 years old. We first need to calculate the number of schools in each district of Greater Sydney and then use $\frac{\text{The number of schools}}{\text{Total number of young people} \div 1000}$。

Z-accomm, retail, health

We need to calculate the number of accommodation and food services per 1000 people, retail services per 1000 people, and health services per 1000 people. Then calculate the number of the above three services in each district of Greater Sydney and then use $\frac{\text{The number of services}}{\text{Population} \div 1000}$

Z-crime

The crime was calculated as the sum of hotspot areas divided by total area, where total area represents the area where the crime occurred. We should use $\frac{\text{Sum of hotspot areas}}{Total\ area}$

Z-score and Sigmoid

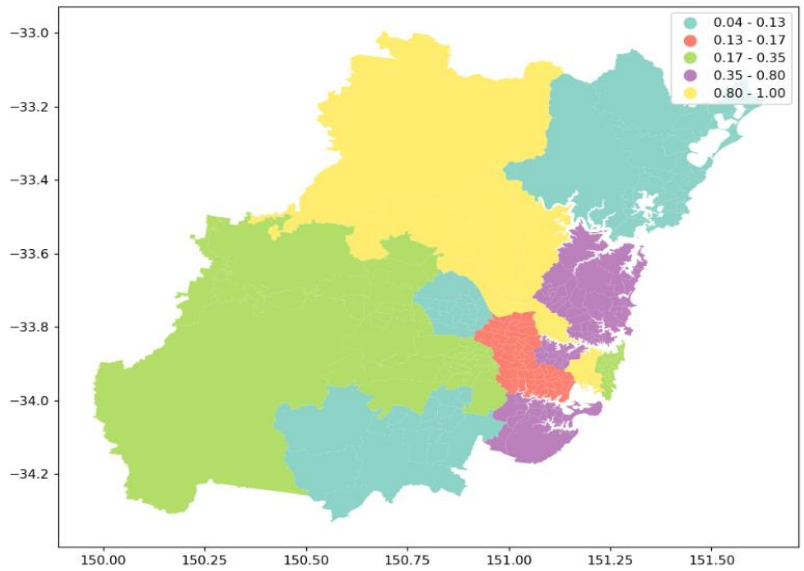
To calculate z-score, we import stats from scipy, and then use stats to calculate the Zscore. For the sigmoid, we from scipy.stats import logistic, then use logistic.cdf to calculate the sigmoid score.

Table of Greater Sydney Livability score

The results are shown in the figure below, where the mean value of greater Sydney Livability calculation is 0.4262, and the median value is 0.2649

	geater_sydney	crime_average	accommodation_food_average	retail_average	health_average	school_average	sum_z_score	sigmoid
0	Sydney - North Sydney and Hornsby	1.578850e-05	4.563279	6.213383	11.694416	0.784564	1.417399	0.804930
1	Sydney - Northern Beaches	4.901909e-06	3.923854	6.373943	6.232877	0.875774	-0.184999	0.453882
2	Sydney - Inner South West	5.623162e-05	3.709106	6.257785	5.084246	0.845291	-1.746669	0.148468
3	Sydney - Baulkham Hills and Hawkesbury	4.494037e-07	3.505514	6.732086	7.968592	1.405832	2.629669	0.932747
4	Sydney - Parramatta	9.604390e-05	3.867302	6.261248	5.253380	0.975800	-1.763453	0.146358
5	Sydney - Sutherland	5.561315e-06	3.262203	5.385505	5.584151	1.124444	-0.242301	0.439719
6	Sydney - Outer South West	2.520301e-06	2.020297	3.775816	3.612596	1.225862	-1.886881	0.131600
7	Sydney - South West	1.132173e-05	2.580386	4.867914	3.945524	1.139448	-1.443834	0.190952
8	Sydney - Eastern Suburbs	1.528548e-04	4.875726	6.380196	11.176558	0.745790	-1.020453	0.264939
9	Sydney - Inner West	1.664556e-04	5.476598	6.961237	9.656526	1.092706	0.179237	0.544690
10	Sydney - City and Inner South	1.450854e-04	11.377764	12.073835	9.956975	1.660599	8.290116	0.999749
11	Sydney - Outer West and Blue Mountains	1.687038e-06	2.536645	4.103211	3.899150	1.412347	-0.614323	0.351074
12	Central Coast	2.407538e-06	3.016596	4.180013	4.662550	0.966195	-1.887236	0.131560
13	Sydney - Blacktown	3.308118e-05	2.315066	4.054148	3.007916	1.013181	-3.196204	0.039309
14	Sydney - Ryde	2.927077e-05	4.333277	5.387039	8.656273	1.255765	1.469930	0.813047

Data Visualization



This figure summarizes the distribution of Livability scores in each district of Greater Sydney. The regions with the highest scores are Sydney-City and Inner South, Sydney-Baulkham Hills, and Hawkesbury. The areas with the lowest scores were Sydney-Blacktown and The Sydney-Outer Southwest. From this point, we can infer that the livability score is related to the area's prosperity, and the location near the CBD has a good score.

Correlation Analysis

	median_income	monthly_rent	sigmoid
median_income	1.000000	0.661919	0.434470
monthly_rent	0.661919	1.000000	0.557529
sigmoid	0.434470	0.557529	1.000000

As shown in the figure, we calculated the correlation coefficient of Livability Score, median income, and monthly rent. A correlation coefficient between 0.4 and 0.6 was interpreted as a Moderate positive association("The Correlation Coefficient (r)", 2022). Substantially, the more livable places are, the more prosperous they are, and they will have higher incomes and higher rents.

Inner Sydney Livability Analysis

Our stakeholders are the older people who want to move or immigrate to Inner Sydney, so we need to consider all the factors that affect the life of the older people. The formula we use to calculate livability will also change. The new formula is:

$$Score = Sigmoid(z_{signs} + z_{forestry} + z_{safety} - z_{crime} + z_{health} + Z_{gardens})$$

Z-signs

Signs can help older people who don't use smartphones find their way, and areas with more signs will be more accessible to older people. We need to calculate the number of signs in each region and divide it by the population. Because older people prefer quiet places, the population data has a negative impact. We should use $\frac{\text{Number of Signs}}{Population}$

Z-forestry, safety, health

The data we choose here is the number of forestries, public safety and health care. More health services will be more convenient for the elderly medical treatment. Forestry can improve air quality. Better social security offers a secure place to live. So, we figured out the number of three variables in each region and divided it by the population. We should use $\frac{\text{Number of servises}}{Population}$

Z-crime

The crime was calculated as the sum of hotspot areas divided by total area, where total area represents the area where the crime occurred. Because crime also has a negative impact on the elderly. We should use $\frac{\text{Sum of hotspot areas}}{Total\ area}$

Z-gardens

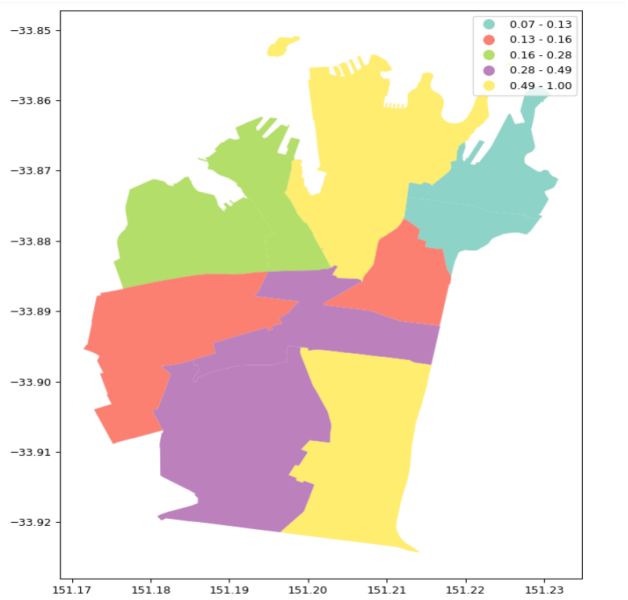
The garden can provide a comfortable place for the elderly to walk and exercise while improving the environment and increasing the green areas. So, we calculate the number of gardens in each region, divided by the population. We should use $\frac{\text{Sum of hotspot areas}}{Total\ area}$

Table of Greater Sydney Livability score

The results are shown in the figure below, where the mean value of Inner Sydney Livability calculation is 0.334, and the median value is 0.173

	Sydney_Inner_City	crime	avg_signs	avg_health	avg_forestry	avg_safety	avg_gardens	sum_z_score	sigmoid
0	Darlinghurst	0.000065	0.015708	0.022555	0.000644	0.000242	0.000483	-2.066887	0.112357
1	Erskineville - Alexandria	0.000014	0.010437	0.006284	0.000729	0.001122	0.000449	-0.262703	0.434699
2	Glebe - Forest Lodge	0.000024	0.007086	0.008383	0.000509	0.000139	0.000880	-1.610537	0.166514
3	Newtown - Camperdown - Darlington	0.000017	0.006375	0.013115	0.000401	0.000182	0.000219	-1.811594	0.140446
4	Potts Point - Woolloomooloo	0.000038	0.007014	0.006636	0.000924	0.000252	0.000798	-2.548476	0.072529
5	Pyrmont - Ultimo	0.000037	0.009716	0.006934	0.000706	0.000789	0.000872	-1.520221	0.179429
6	Redfern - Chippendale	0.000026	0.009163	0.005051	0.000117	0.000196	0.001997	-0.187498	0.453262
7	Surry Hills	0.000042	0.009522	0.011404	0.001107	0.000830	0.000498	-1.811283	0.140483
8	Sydney - Haymarket - The Rocks	0.000013	0.030700	0.039245	0.010029	0.003676	0.000193	11.252616	0.999987
9	Waterloo - Beaconsfield	0.000016	0.005827	0.004060	0.000264	0.000475	0.002215	0.566582	0.637974

Data Visualization



The chart above summarizes the distribution of livability scores in the inner Sydney area. The areas with the highest scores were Sydney-Haymarket-The Rocks and Waterloo - Beaconsfield. The place with the lowest score was Potts Point-Woolloomooloo. Regions that are livable for the elderly may not be liable for other groups of people. Woolloomooloo is the CBD area of Sydney with many shops and high-end restaurants. But these commercial areas are not appropriate for older people, so it makes sense that it gets the lowest score.

Correlation Analysis

	monthly_rent	sigmoid
monthly_rent	1.000000	0.629489
sigmoid	0.629489	1.000000

As shown in the figure, we calculated the correlation coefficient between livability score and monthly rent. A correlation coefficient between 0.6 and 0.8 is interpreted as strong positive association ("The Correlation Coefficient (r)", 2022). This means that the more livable the place, the higher the monthly rent. Because our formula includes health services, virescence, and public safety, which are attributes of new communities, older communities may not have these amenities because of urban planning. In comparison,contemporary communities are more focused on quality of life. Therefore, we recommend that older people choose a more livable area to live.

References

The Correlation Coefficient (r). (2022). Retrieved 24 May 2022, from https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module9-Correlation-Regression/PH717-Module9-Correlation-Regression4.html

Rain gardens. (2022). Retrieved 24 May 2022, from https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::rain-gardens/explore?location=-33.891734%2C151.204117%2C13.97&showTable=true

Wayfinding signage. (2022). Retrieved 24 May 2022, from https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::wayfinding-signage/explore?location=-33.889226%2C151.204472%2C13.78&showTable=true