THE UNIVERSITY OF SYDNEY

DATA2001
ASSIGNMENT ONE

Sydney Liveability Analysis

GROUP MEMBERS:

YUSHANG CHEN

ZHEHAN WANG

Introduction

Everything has changed a lot since the covid-19 pandemic hit. Sydney was one of the most livable cities around the world. However, a study shows that the livability index of Sydney has dropped from the third to eleven (White, 2021). We will have our algorithm for livability and give suggestions to people who have these requirements. A score is determined by using the sigmoid function. This report will investigate the most livable suburb in Greater Sydney, so a big variety of groups could benefit from the result, including married couples, elder people, a family with kids, and so on. Specifically, we will advise international college students to buy or rent a house in City of Sydney.

Datasets

7 individual datasets are used in this study. 5 datasets are used to calculate each area's livability score, including *neighborhoods*, *businessstats*, *school catchments, break_and_enter*, *sa2*. The others (*playgrounds, car_share*) are used in calculating a special score for international college students.

Neighborhoods
Neighborhoods is a dataset that describes the information about the neighbors, land area as well as the number of dwellings. It consists of 322 rows and 13 columns. This dataset is derived from *ABS census data*. We define young people in the greater Sydney score analysis as humans aged from 0 – 19 years which is the sum of three columns in this dataset.

BusinessStats
BusinessStats is a dataset that is derived from ABS Census data. It contains 2301 rows and 9 columns. It includes the number of infrastructures like hospitals and restaurants in an area. This dataset can help us understand whether the area we live in has enough infrastructure to improve the convenience of life and whether there are too many enterprises and industries interfering with life.

Break_and_enter
This dataset is primarily sourced from BOCSAR. It contains 2594 rows and 7 columns. This dataset mainly describes the crime density in a certain area using spatial data and records the area. Crime density is the number of crime incidents within a certain area. When the crime incidents are clustered, the crime density is defined as high. It is one of the important issues for all people when we decide to have our own property. It will be used frequently in the below calculation.

SA2
This dataset is derived from the Australian Bureau of Statistics. It contains 2310 rows and 13 columns. SA2 records the geographical location, area, name, and some

geographical information of the whole of Australia. We will focus on SA2 in this report. The column named SA2_MAIN16 is an identity column that can join with the previously introduced dataset - neighborhoods and BusinessStats.

School_Catchments

School_Catchments is resourced from the NSW Department of Education. It consists of three shapefiles, including Geographical coordinates for primary school, secondary and schools offered in the future. Primary school has 1666 rows and 19 columns. Secondary school has 435 rows and 19 columns. Future school has 44 rows and 18 columns.

Car_Share

Car share is a geojson file resourced from City of Sydney Data Hub. It includes 794rows and 6 columns. The dataset describes the specific location and the bay operator. Also, the suburb of the location is recorded. We used this data as a variable in the analysis for international college students.

Playgrounds

Playgrounds is also derived from City of Sydney Data Hub. It includes 151 rows and 4 columns. This dataset records the location of playgrounds and outdoor fitness stations in the local area. It includes two attributes: Name and Type. Type includes 'Playground' and 'fitness station'. This dataset is only used for international college students' analysis.

Data Cleaning

Firstly, we removed irrelevant columns from our datasets because these columns are not necessary to our interests. Then, if a table has a geometry column and this column has polygon type data, we change it to multipolygon type to ensure consistency. Next, for sa2 table, we select the data in Greater Sydney Because we only need the information of this location in this report. We combine the three school tables into one table named school. Additionally, for rows with duplicate names in school table, we keep one row and delete the rest because we want to get the number of schools that not duplicated.

Database Description

Every dataset we described before is structured data. We create tables for each dataset and store them in SQL database. The primary key for *neighborhoods* and *businessstaus* would be their *area_id. sa2 has sa2_main16* and *neighbourhoods* has *area_id* as a foreign key for referring to *area_id* in *businessstats*. Besides, *school_catchments, break_and_enter*, *school*, *car_share*, and *playgrounds* have geometry data to spatial join with *sa2's* spatial data. The schema is depicted as below:

neighbourhoods

PK &FK	__area_id__
	area_name
	population
	median_annual_household_income
	avg_monthly_rent
	young_people
Index	*n_pop_index*

break_and_enter

PK	**objectid**
	density
	shape_area
	geom

School

PK	**use_id**
	catch_type
	use_desc
	geom

businessstats

PK	__area_id__
	area_name
	accommodation_and_food_services
	retail_trade
	health_care_and_social_assistance
	public_administration_and_safety

sa2

PK/FK	__sa2_main16__
	sa2_5dig16
	sa2_name16
	sa3_code16
	sa3_name16
	sa4_code16
	sa4_name16
	gcc_code16
	gcc_name16
	ste_code16
	ste_name16
	areasqkm16
	geom
index	*sa2_index*

car_share

PK	**objectid**
	bay_id
	operator
	street
	suburb
	geom

playground

| PK	**objectid**
	name
	Type
	geom

Index

We created two indexes to be more efficient. One of them is spatial data – geometry data in *sa2*. One of them is created on the population in *neighbourhoods* dataset because it is used most frequently throughout the project.

Greater Sydney Score Analysis

We will use a "score" to measure the livability of an area in Greater Sydney. The score is determined by five aspects, including school catchment areas, accommodation and food services, retail services, crime density, and health services. Before we get into the real analysis, we find all the areas with a population of young people less than 500 and set the z score of school to be zero. If the population is smaller than 500, we set all the z scores to be zero except for z score for the school catchments. If any cell is a null value, we set all the scores to zero.

We define school as number of schools catchment areas per 1000 young people. Crime is the sum of hotspot areas divided by total area, and others are corresponding services per 1000 people. After calculating the statistics for these attributes. We calculated the statistical z score for each attribute, then we add up all of the z scores and minus the z score for crime. The reason is that only crime density will negatively affect the livability of people. Finally, we use the sigmoid function on the whole livability score. The sigmoid function is a mathematical function that has an S-shaped curve and it could map all values between 0 and 1 (Saeed, 2021). A score close to 1 indicates good livability and a score close to 0 indicates bad livability (Saeed, 2021).

As we can see from the table below (see *table 1*), The Rocks in Haymarket was ranked top one in our ranking system. The z score for accommodation, retail, and health are very outstanding. The crime density is worse than the average. The second shortage is the number of school catchments is also relatively lower. All in all, the livability of The Rocks is extremely high which means it could be one of the best locations to live in Greater Sydney. At the same time, the average monthly rent in The Rocks is very expensive, at 2998 dollars. Marsden Park is an area that has one of the best resources for school catchments, however, all the other attributes are very normal. It is a good choice for a family with children. In addition, the average monthly rent is much cheaper compare to The Rocks.
Chullora, Lavender Bay, and Narembum are the areas that also have high livability scores.

figure 1: score map for Greater Sydney
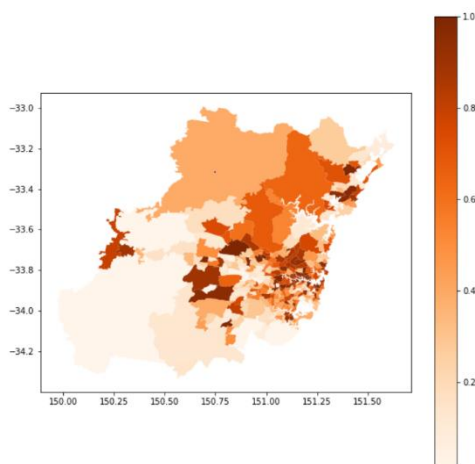
area_name	median_annual_household_income	avg_monthly_rent	population	young_people	Zaccom	Zretail	Zhealth	Zschool	Zcrime	Sydney liveability score
Sydney - Haymarket - The Rocks	27311.0	2998.0	31010.0	2617	13.600087	10.314945	6.781622	-0.482947	1.162882	1.000000
Chullora	41625.0	2280.0	1225.0	246	2.078130	9.596891	0.603946	-1.193241	-0.497786	0.999991
Riverstone - Marsden Park	49900.0	1891.0	21668.0	5214	-0.346283	-0.015235	-0.682550	10.452736	-0.345189	0.999942
North Sydney - Lavender Bay	71668.0	2749.0	11819.0	1576	4.053044	2.166775	3.278553	-0.564192	-0.214477	0.999894
St Leonards - Naremburn	73803.0	2626.0	11400.0	1972	1.694527	1.278446	3.408412	1.571768	0.001849	0.999648

table 1: statistics for top 5 livability (Greater Sydney)

A score map for Greater Sydney is shown above (see *figure 1*). As we mentioned, we ignore areas whose population is smaller than 500 and we record their score as zero. We can deduce that the population on the west side of the map is quite small because there are plenty of white indicating the population is smaller than 500. The high score areas are in the eastern of Greater Sydney and most of them are close to the river. For example, the area with the highest score is almost surrounded by a river.

Correlation Analysis

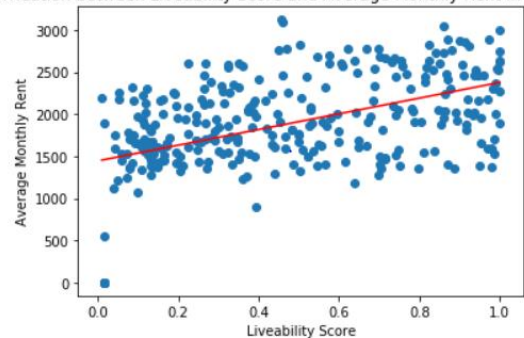*figure 2: correlation between livability and median income*

figure 3: correlation between livability and monthly rent

Figure 2 shows the correlation between livability score and median income in Greater Sydney. We can observe that the median annual household income is increasing as the livability score raises because the regression line has an increasing shape. The correlation coefficient is 0.33, so we would conclude that they are related moderately. Figure 3 indicates the correlation between livability score and average monthly rent in Greater Sydney. We can see the slope of the regression line is much sharper than the regression line in figure 3. However, average monthly rent is also correlated to the livability positively. The coefficient is 0.49 and we would consider it to be a strong positive correlation.

The limitation here is we have a record that has zero annual household income and zero average monthly rent. The correlation coefficient will get smaller if we remove this data.

City of Sydney Analysis

In the past two years, a lot of international college students are locked up in their own countries and therefore do not have a chance to take the face-to-face course on campus in Sydney. We have more international college students back on campus. Choosing a residence in a new country is probably the hardest problem for them. In this report, we will look into areas in the city of Sydney and give some recommendations for international college students who came to Australia for the first time.

An international college student, he/she probably interested in traveling around Sydney when they have spare time. Then, if they have a driving license, renting a car would be one of the best choices for a couple of friends to have a short trip. This could be the most convenient and affordable option. Besides, there are lots of international students who are facing psychological challenges. For example, they may get depressed because of the massive studying load, or they may feel lonely because of cultural differences. regular exercise could efficiently prevent or even improve their psychological issues; so it is good to have playgrounds and fitness stations near their residence. Therefore, we added the z score for the number of car share bay per 1000 population and the z score for the number of playgrounds per 1000 population to our sigmoid function.
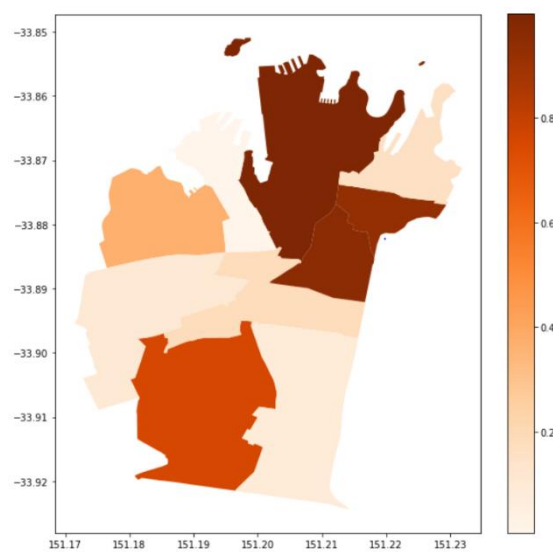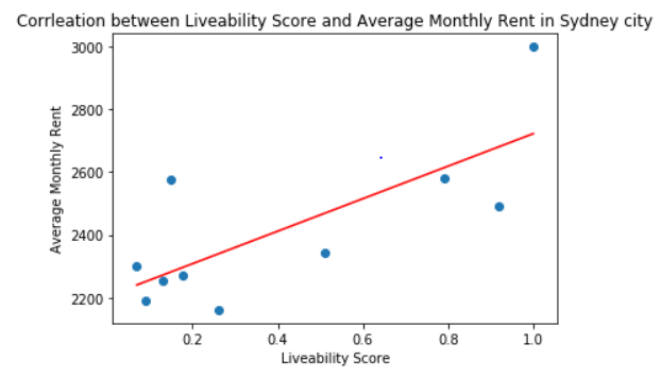

figure 4: livability score for City of Sydney

figure 5: correlation between livability and monthly rent

Figure 4 is the livability score map for the City of Sydney. The livability score is greater when the color is darker. The Rocks are still the area with the highest livability score. The food services and retail services are very rich there. However, it also has the most expensive median rent (in figure 5). Especially, students need to consider a high quality-price ratio. An international college student could also consider Surry Hills which is the fourth-best area. It has more public playgrounds than The Rocks. Also, it is much more convenient to rent a car there. The disadvantage is the crime score is high meaning there are more crimes that happened in the past several years.

Conclusion

In conclusion, The Rocks are the best area to live for most people and international college students. It is a place with rich food resources, retail resources, and school catchments. However, the average monthly rent is also the highest in the City of Sydney. Everyone has their own concern when renting a residence, and our report could be implemented with more attributes.

References

1. White, D. (2021, June 9). *Sydney Slides down global liveability rankings*. The Sydney Morning Herald. Retrieved May 25, 2022, from https://www.smh.com.au/national/nsw/sydney-slides-down-global-livability-rankings-20210609-p57zdv.html

2. Saeed, M. (2021, August 17). *A gentle introduction to sigmoid function*. Machine Learning Mastery. Retrieved May 25, 2022, from https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/