THE UNIVERSITY OF SYDNEY

DATA2001 Group Assignment

F14D-RE10-2

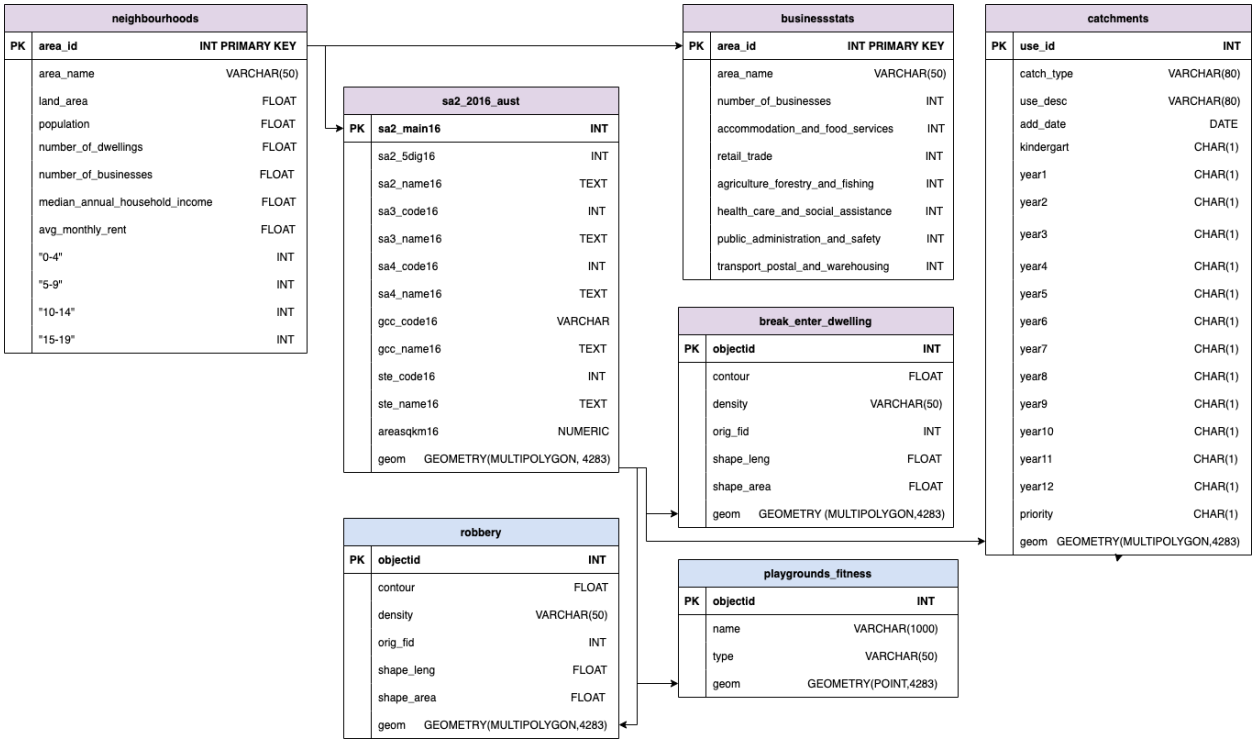
490018222, 510615460

1. Dataset Description

Dataset	Data Sources and Description	Pre-Processing
Neighborhoods Dataset	The neighborhoods dataset is a provided dataset that was acquired from the University of Sydney's Canvas site. Accordingly, this dataset consists of the number of dwellings, and businesses, median annual household income, the average monthly rent, and the mean number of people (in regards to age) in a certain area. This dataset is a CSV file that was derived from the ABS census data.	This dataset was pre-processed using Pandas by replacing null values with 0 and dropping duplicate values (if any). The number_of_dwelling and population columns were also converted into a float in addition to replacing existing commas with an empty string.
Business Stats Dataset	The Business Statistics dataset is a provided dataset that was obtained through the University of Sydney's Canvas site. This dataset consists of the number of operating businesses in an area and its related industry. This dataset is a CSV file taken from the ABS census data.	This dataset was cleaned by replacing null values with 0 and dropping duplicate values (if any) using Pandas.
Statistical Area 2 Dataset (SA2)	The SA2 dataset is another provided dataset that was taken from the University of Sydney Canvas site. This dataset is an SHP file that was taken from ABS census data.	Before loading this dataset to the database, this dataset was cleaned by dropping null values and duplicates (if any) using Pandas.
Break and Enter Dataset	The Break and Enter dataset is also a provided dataset that was downloaded from the University of Sydney's Canvas site. This dataset is an SHP file that was acquired from NSW BOSCAR's open data catalog and is one that represents the crime rates involving the breaking and entering into dwellings or personal residences.	After loading this dataset to Pandas, the column names were converted into lowercase letters, null values were dropped and duplicates were removed as well (if any).

School Catchments Dataset	The School Catchments datasets are datasets taken from the University of Sydney's Canvas site. These datasets are SHP files derived from the NSW Department of Education website. There are 3 catchments datasets, all of which visualize the school intake zones or catchment areas for differing types of schools.	This dataset was first pre-processed by appending the 3 different shape files, dropping duplicate values, dropping null values, and creating a new geom column which is used to convert it to WKT format followed by dropping the old column.
Robbery Dataset	The Robbery dataset is also an SHP file that was acquired from NSW BOSCAR's open data catalog [1]. This dataset showcases the crime rates that involve robbery in Sydney.	To pre-process this dataset, the column names were converted into lowercase letters, null values were dropped and duplicates were removed as well (if any).
Street Safety Camera Dataset	The Street Safety Camera dataset is a dataset that was acquired from the City of Sydney Data Hub [2]. This dataset was downloaded in the form of a CSV. This dataset shows the installation of street safety cameras in the areas which are deemed unsafe (in Sydney).	No pre-processing was done before reading this dataset to the database using Pandas.
Playground and Fitness Station Dataset	The Playground and Fitness Station dataset is a dataset that was acquired from the City of Sydney Data Hub [2]. This dataset was downloaded in the form of a GeoJSON. This dataset pinpoints the locations of playgrounds and outdoor fitness stations in Sydney.	The dataset was also cleaned by removing both null and duplicate values and column names were changed into lowercase letters using Pandas.

2. Database Description



The above diagram visualized the datasets that were integrated and correspondingly, the following spatial indices were created to assist and speed up the spatial joins:

Index	Index Description
sa2_idx	Created for the geom column of the joint neighborhood, business stats, and shape dataset
catch_idx	Created to access the geom column of the school catchments dataset
break_idx	Created for the geom column of the break and enter dataset

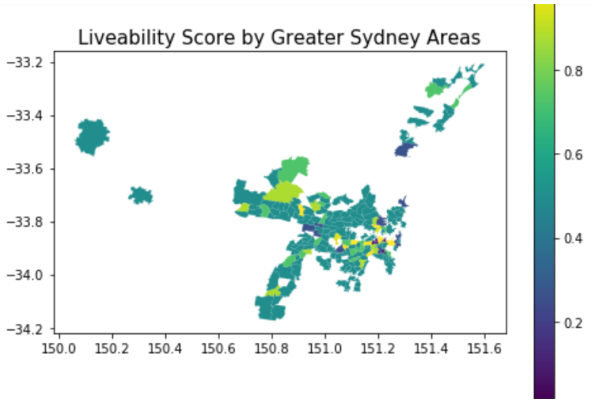
3. Greater Sydney Score Analysis

$$Liveability\ Score = S(z(school) + z(accomm) + z(retail) - z(crime) + z(health))$$

In accordance with 'liveability' scores of each of the given neighborhoods, the above formula was used - where S is the sigmoid function and z is the normal z score. In calculating the liveability scores, we take

into account five factors. The first factor would be school catchment or school intake areas per 1000 'young people' or more specifically, people of the ages 0-19 (school measure) - this factor plays a role in dictating the liveability of an area as those studying seek to live in areas that are nearby schools as it would be convenient. Even so, some may even consider school zones safer. The second factor accounted for is the number of accommodations and food services available per 1000 people (accom measure) and the third factor measures the number of retail services available per 1000 people (retail measure) - the two said factors can be used to consider the liveability of an area to judge how convenient the area is as most people would want easy access to a variety of businesses. The fourth factor is the sum of break-and-enter-related crime hotspot areas per total area (crime measure) - this factor is used to 'assess' the safety of an area. The last factor measures the number of health services per 1000 (health measure). Moreover, all the z-scores of each measure and the 'liveability' score (sigmoid) were computed using NumPy.

Note: *In regards to the crime measure, we chose to only calculate the 'high-density' areas because we believe that by definition, crime hotspot areas are "a place of significant activity, danger, or violence" [3]*



Accordingly, a graphical representation was created to visualize the 'liveability' scores or in other words, the results of our calculation. As shown, we made use of a heat map to graphically represent our findings. The legend shows the liveability scale - yellow being the most liveable and purple being the least liveable (the higher the score the more liveable an area is). For clarity, the below table showcases some of our main findings - the top five most liveable areas in Greater Sydney with Sydney - Haymarket - The Rocks as the area with the highest liveability score.

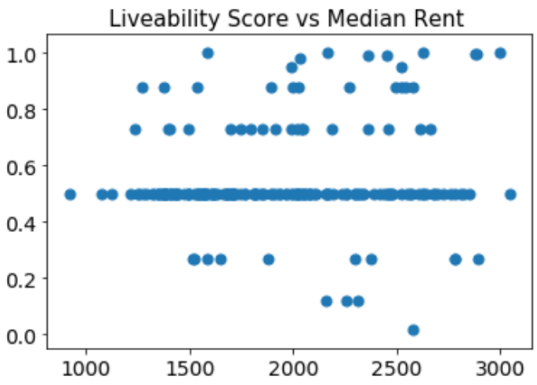
Note: *There are empty points because we made use of the JOIN function.*

	area_id	area_name	liveability_score
51	117031337	Sydney - Haymarket - The Rocks	1.000000
23	116011303	Blacktown (East) - Kings Park	0.999983
112	121011401	St Leonards - Naremburn	0.999877
101	120031391	Burwood - Croydon	0.999877
55	118011343	Double Bay - Bellevue Hill	0.999089

4. Correlation Analysis

The liveability scores were compared to the median rent of each neighborhood and the median income of each neighborhood using the Pearson Correlation Coefficient to identify whether there's a correlation between each of the said variables. Our findings are displayed in the below table and graphs.

	Liveability Score vs Median Rent	**Liveability Score vs Median Income**
Correlation Coefficient	$\rho = 0.113$	$\rho = 0.103$
Correlation	Very weak positive correlation	Very weak positive correlation



The table above shows that the correlation coefficient of the liveability score and the median rent of each neighborhood is $\rho = 0.113$. Since the correlation coefficient is less than 0.3 and is very close to 0 (no correlation), we can conclude that there is barely a correlation between the liveability score and the median rent of a neighborhood. This can also be seen on the graph where most values are predominantly lined up horizontally with some points showing a slight upward trend.

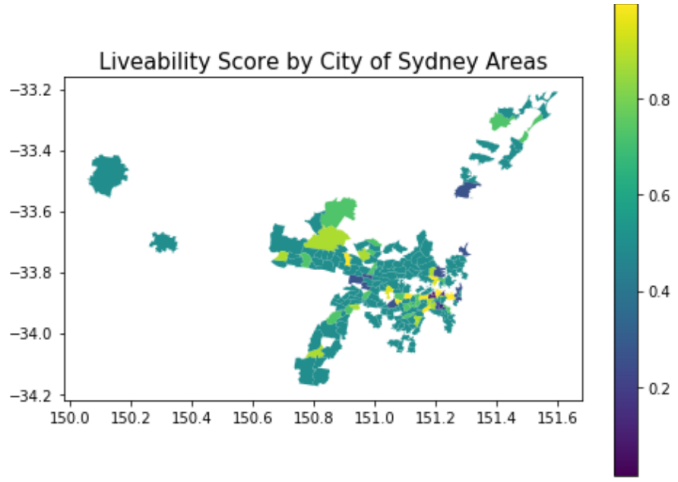
As shown in the second graph, we can see that there is no visible (upward or downward) trend. Furthermore, the table also suggests that there is an extremely weak positive correlation between the liveability score and the median income of each neighborhood as the correlation coefficient is $\rho = 0.103$ which is also less than 0.3 and is closer to 0.

5. City of Sydney Analysis

Our proposed stakeholders would be ex-pat families that are considering relocating with kids. More specifically, families with children aged 0-9. The parents of these families are those that thrive in a city setting but also would want their kids to enjoy a broad range of outdoor activities. Additionally, these families are also aware that Sydney is a family-friendly city with an extensive array of educational institutions. However, in search of a place to stay, these families do not know which area would best suit their needs.

$$Liveability\ Score = S(z(school) + z(playground) - z(robbery) - z(crime))$$

To conduct a more tailored analysis, we made use of the two extra datasets (Refer to Part 1) to calculate our 'refined' liveability score using the above formula. As mentioned previously, we believe that these families would be interested in having their kids enjoy the outdoors, thus, one added measure would fall under the category of playgrounds which by definition, is the playground density. Aside from the playground measure, we added a robbery measure using one of our extra datasets so our stakeholders would be more aware of the crime rates in these areas. Additionally, we made use of the same dataset in computing the school measure, however, modified it to catchment density. Similarly, the same break-and-enter dataset and the same calculations were done for our crime measure.



With that being said, the said calculations were represented in a heatmap as shown below. The legend shows the liveability scale - yellow being the most liveable and purple being the least liveable (the higher the score the more liveable an area is).

Appendix and Reference List

[1]

Community Relations Division and N. D. of Justice, "Open Data," *www.bocsar.nsw.gov.au*, 2020. https://www.bocsar.nsw.gov.au/Pages/bocsar_datasets/Datasets.aspx

[2]

"Street safety cameras," *data.cityofsydney.nsw.gov.au*, 2017. https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::street-safety-cameras/explore?location=-33.871009%2C151.208460%2C15.00&showTable=true.

[3]

Oxford, "hotspot noun - Definition, pictures, pronunciation and usage notes | Oxford Advanced Learner's Dictionary at OxfordLearnersDictionaries.com," *Oxfordlearnersdictionaries.com*, 2022. https://www.oxfordlearnersdictionaries.com/definition/english/hotspot?q=hotspots.