DATA2001 Assignment Report

1 | Dataset Descriptions

In this assignment, 7 datasets were used:

Shape data: SA2 level data

*Description:*Describes a digital boundary of areas that interact together socially and economically.
Source: Australian Bureau of Statistics: Australian Statistical Geography Standard (ASGS) Edition 3. The publication allows us to view and analyse boundaries in desktop Geospatial Information systems
Preprocessing: 2 separate dataframes were created from the SA2 level data. For both, the geometry column was converted to WGS84 coordinate system and any row with geospatial data was removed. For the greater sydney data, SA2 was filtered to only contain information from 'Greater Sydney'. Similarly, in the City of Sydney dataframe, only 'Sydney = City and Inner South' data was included.

Business statistics per SA2-area

Description: Contains the number of business in a business sector for each SA2 area.
Source: Australian Bureau of Statistics
Preprocessing: Virtually inspected for missing data. If missing data was found, placeholder 'none' was put in place.

Neighbourhoods (SA2-level areas) in greater sydney

Description: Contains information about each neighbourhood in the SA2 area. The metadata includes: area_id,area_name,land_area,population,number_of_dwellings,number_of_businesses,median_annual_household_income,avg_monthly_rent,0-4,5-9,10-14,15-19
Source: : Australian Bureau of Statistics
Preprocessing: Visually inspected for missing values. Rows with null values were dropped.

School Catchment data

Description: Contains information about school catchment zones in regional and greater NSW. School catchment zones are geographic areas from which students who live in those zones are guaranteed a place in a zone specific school. The data comes in three types; secondary, primary and future - representing the type of education offered. The secondary and primary dataset contain information about how many students can attend in each year group (Kindergarten through to year 12), while the future dataset indicates the year the school will be opened. Additionally, the datasets contain geometry information in the form of Polygons (describing the geographic area of the zone) and the name of the school pertaining to each zone (called the school description in the dataset).
Source: NSW Education Hub
Preprocessing: We are interested in obtaining the geometric and school description data for each area location. We note that each Neighborhood does not have an identical geometric correlation to the school catchment data as many neighbourhoods may intersect with a single catchment zone. Firstly, we convert the geometry information into a geometry object using the shapely library such that data can be interpreted in the WGS84 coordinate data. Some text processing is applied to convert the column names to lower case to adhere to the postgis server format. Column data that is not required is safely deleted from the catchment dataset including: "priority" and "add_data" columns. Finally, the future dataset contains different information in the year columns (kindergarten to year 12) than the other datasets. The primary and secondary datasets contain a boolean "Y" or "N" for each year group column, while the future dataset contains a numeric value, indicating the date the school will open. As part of the preprocessing step, the year group columns in the future dataset are conditionally converted to "Y/N". If the value in any of these year group columns is greater than 0 (indicating that, in future, the school will have education available for this year group), the value is set to "Y", else to "N". This brings the data in the future dataset inline with the other school catchment data sets so they may be processed at once.

Break and Enter data

Description: The break and enter dataset contains information on the density of crime within a specific geographic area. The crime data is categorised as “high”, “medium” and “low” density. The dataset also contains geometric data indicating the area covered, the total area of this area and the circumference of the geometry. As with the school catchment dataset, the geometry of each row does not directly correlate with one particular neighbourhood, and these must be correlated based upon the geometry.

Source: NSW Bureau of Crime Statistics and Research

Preprocessing: We are interested in identifying how much crime occurs in each of the neighbourhoods in our dataset. Hence, we want to geometrically associate our breakin data with one or more neighbourhoods. The data is processed similarly to the school catchment dataset by converting the geometry information into a geometry object so we can interpret it using the WGS84 coordinate system. We drop columns ‘orig_fid’ and ‘contour’ as they are not useful in identifying the crime level in each neighbourhood. Finally, we convert the crime categories from strings to numbers, representing the amount of crime; “high” is given a value of 3, “medium” a value of 2 and “low” a value of 1. This helps calculating the z score for this dataset later on.

Mobility parking

Description: Provides spatial data in the form of points, indicating where mobility parking spots are located within the City of Sydney local area.

Source: City of Sydney Data Hub

Preprocessing: The spatial data was converted to a WGS84 coordinate system to be compatible with the SA2 data and other data used for this assignment.

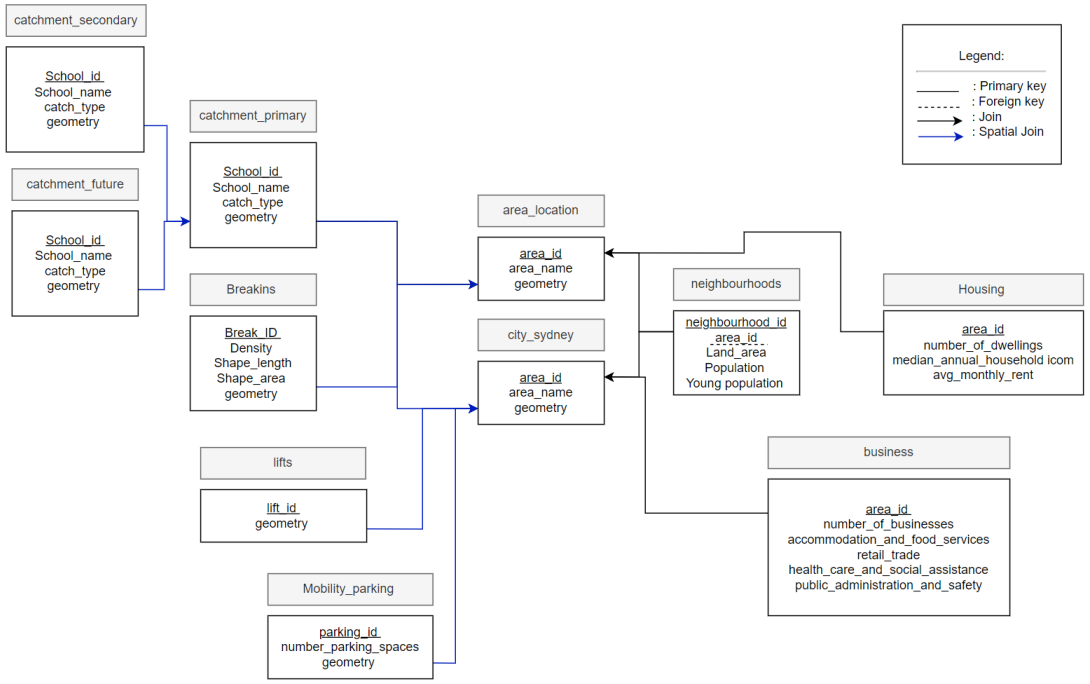
Lifts

Description: Provides spatial data in the form of points, indicating where lifts are located throughout the City of Sydney area.

Source: City of Sydney Data Hub

Preprocessing: The spatial data was converted to a WGS84 coordinate system to be compatible with the SA2 data and other data used for this assignment.

2 | Database Description



2.1 Indices

9 indices were created in total. We created indices on attributes used the most in JOINS and created similar indices on attributes used specifically for spatial joins. Indices were created on:

- 'area_id' from area_location as this was used to calculate each z-scores so it was used frequently
- 'area_id' from neighbours as this was used to calculate the z-scores too so it was used also used frequently
- Indices on the 'geometry' (which contains spatial data) attribute from: catchement_secondary, catchement_primary, catchment_future, lifts, mobililty_parking, area_location, city_sydney

3 | Greater Sydney Score Analysis

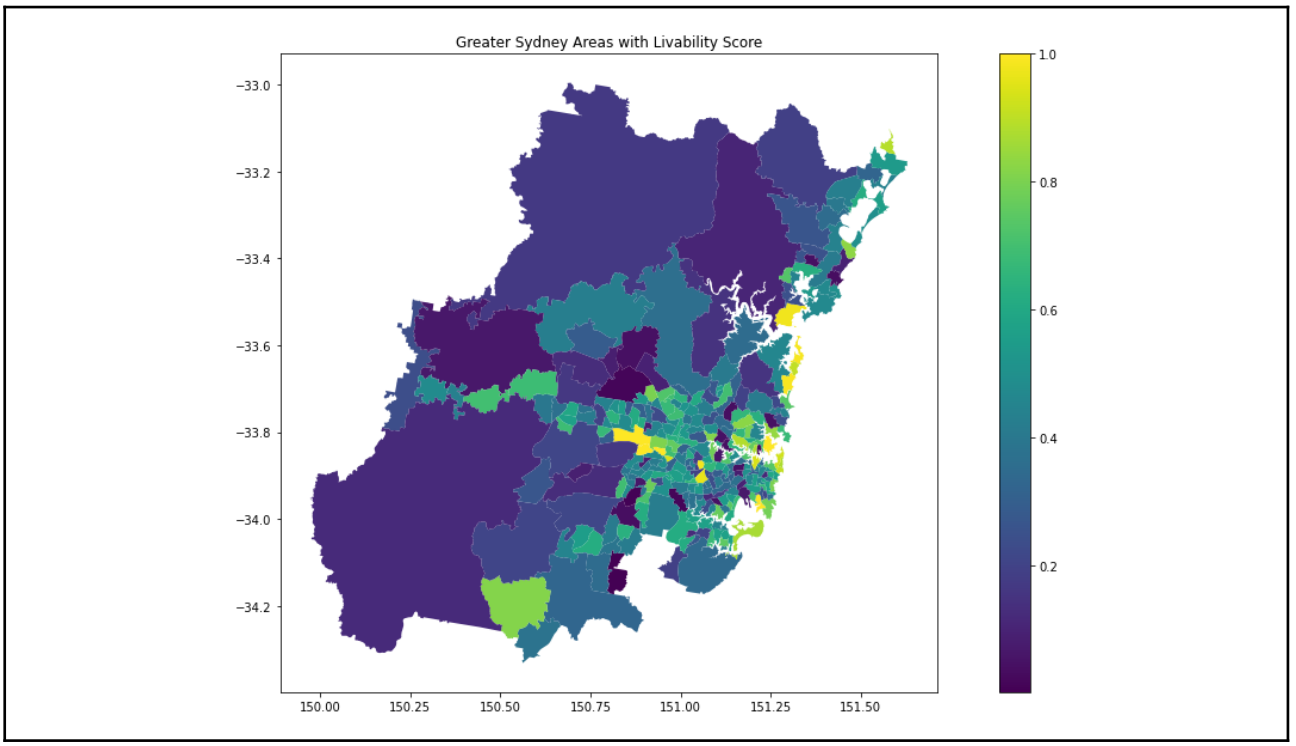
The formula used to compute our livability score was the sigmoid function of the Z score of each attribute.

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

The Z scores were calculated as followed:

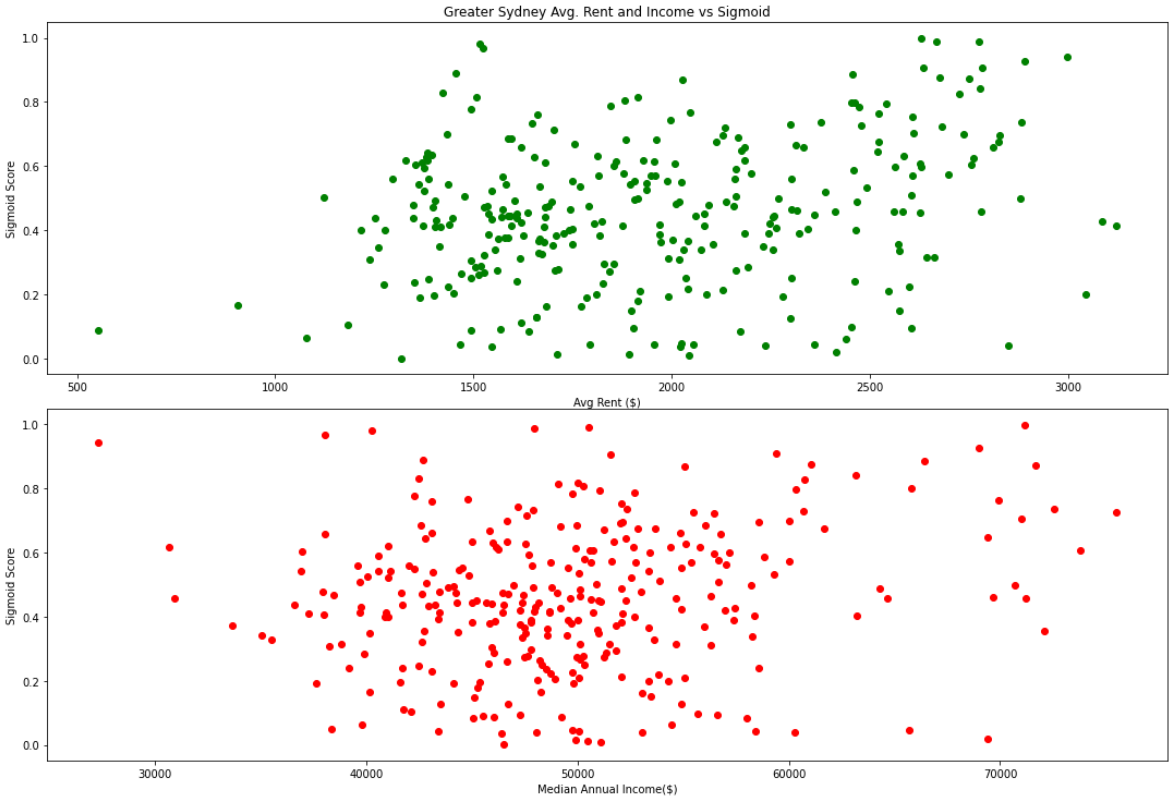
Measure	Definition
school	Number of schools catchment areas per 1000 'young people'
accom	Number of accommodation and food services per 1000 people
retail	Number of Retail services per 1000 people
health	Number of health services per 1000 people
crime	Sum of hotspot areas divided by total area

4.2 Results and Visualisation



Geometry of Greater Sydney where the colour map represents the normalised Z score. Qualitative analysis shows that there is a trend towards a higher Z score (ie. better living conditions the closer the metropolitan Sydney). However, this is by no means totally representative as we observe areas with low livability scores close to the central CBD as well.

4 | Correlation Analysis



Plots comparing the average rent and median annual income to the final livability score for each of the neighbourhoods in Greater Sydney. The plots visualise the low correlation between these attributes (as discussed in Section 4). There is a slight upwards linear trend for average rent, but overall the data appears somewhat uniformly spread.

The correlations between the median neighbourhood income and the overall livability score of the neighbourhood as well as the median rent and livability score for each neighbourhood was calculated using the pandas corr() function. It was found that the correlation between neighbourhood average income and livability was approximately 0.171 and the correlation between median rent and livability was higher at approximately 0.23. These scores were lower than what was predicted but could have been potentially influenced by missing data.

5 | City of Sydney Analysis

5.1 Stakeholder

Mark is a single 50-year-old man who had suffered a traumatic spinal cord injury falling down a flight of stairs after a night out on the town. As a result Mark is paralysed from the torso down and needs to use a wheelchair. Mark lives with his husband who is able to drive him around and park in mobility parking spots. His husband enjoys light shopping so he would still like retail shops nearby. Both would like to live in an safe area with good accessibility infrastructure and not secluded (i.e. have other houses nearby too).

5.2 Livability score modifications

The formula used to compute our livability score was taking the sigmoid function of the Z score of each attribute, similar to what was done when calculating the livability score for areas in Greater Sydney. However the 'school' attribute was removed as the stakeholder doesn't care about what schools are around. Inplace we added the additional datasets 'Mobility Parking' and 'Lifts'. This will be helpful to assess the accessibility of the areas whereby areas with more lifts and mobility parking will be more friendly to our stakeholder, who uses a wheelchair. The final livability score was calculated with the following formula:

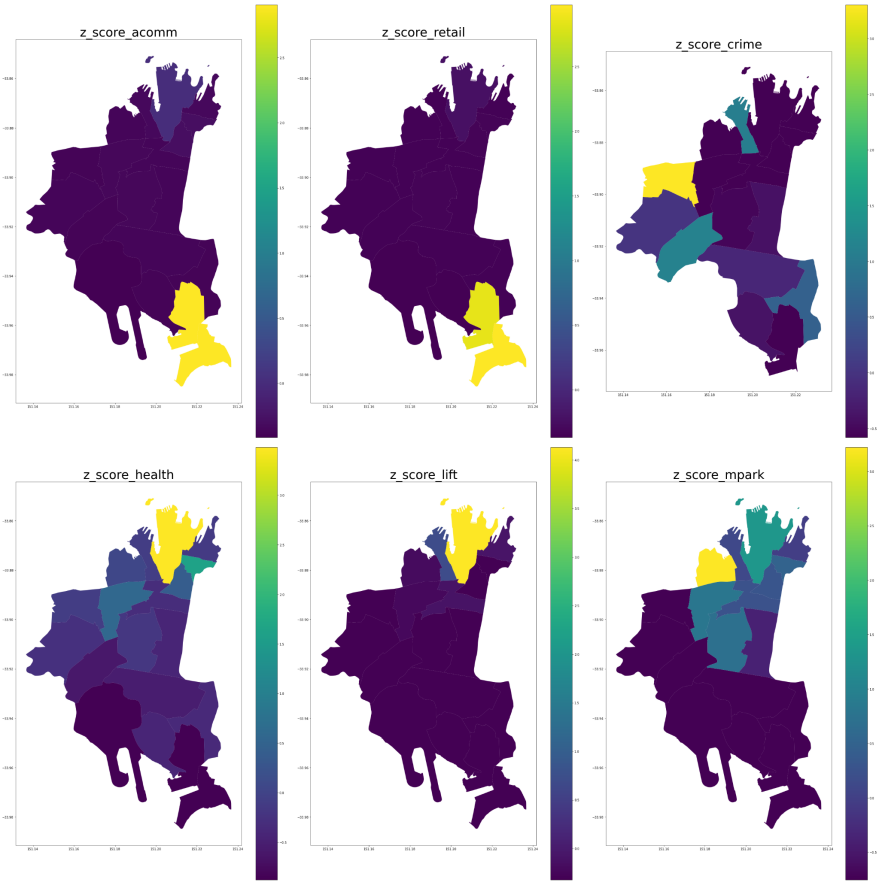
$$Score = S(z_{accomm} + z_{retail} - z_{crime} + z_{health} + z_{lifts} + z_{mobility\ parking})$$

In addition to the previously defined Z-scores we have:

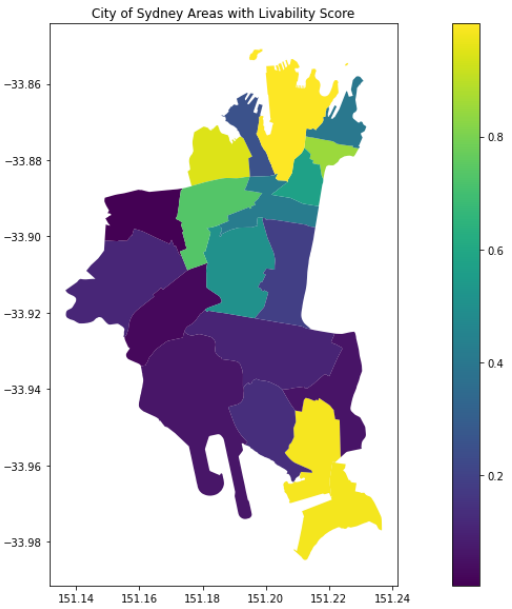
Measure	Definition
lifts	Number of lifts per 1000 people
mobility parking	Number of mobility parking spots per 1000 people

5.3 Resutls and visualisation

(next page)



Normalised Z scores for City of Sydney for each of the categories used to calculate the final livability score. A general trend shows that the northern and south-eastern sections of the city have better access to health, accommodation and general infrastructure such as lift access and mobility parking. The 3rd plot shows a general trend that crime is more frequent in the west of Sydney. These general trends are reflected in the final livability score (below) as the northern and south-easten parts of the city have the best livability scores.



Geometry of Greater Sydney where the colour map represents the normalised livability score.