DATA 2001 Assignment Data Report

Datasets used and sources:

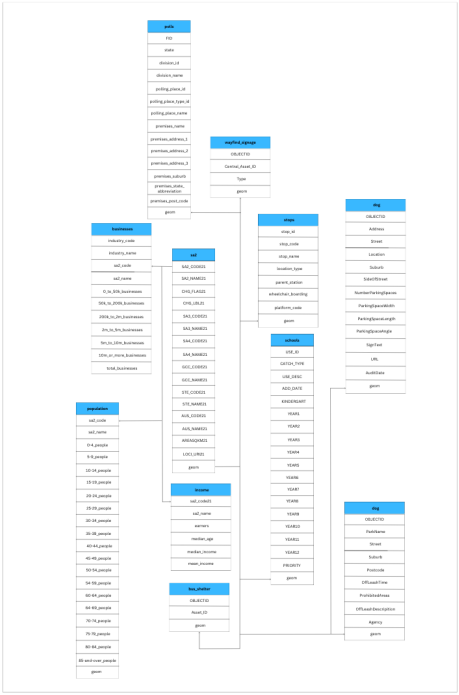
1. SA2 Regions - SA2, Statistical Area Level 2, is a dataset of Australia's functional areas and includes data about the area codes and names in Australia, along with the longitude and latitude position of these areas. We filtered this dataset by only using areas in "Greater Sydney" and turning the longitude and latitude position of the areas into a geometric data that can be read by our query. Source: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files/SA2_2021_AUST_SHP_GDA2020.zip File Type: zip, shp
2. Businesses - A dataset of business by industry in the SA2 regions that includes the number of businesses in that region for each turnover size range. We filtered this dataset by removing any unknown values in the dataset. Source: https://canvas.sydney.edu.au/courses/56224/files/36508328/download?download_frd=1 File Type: CSV
3. Stops - A dataset of all public transport stops in General Transit Feed Specification (GTFS) format, as well as the longitude and latitude values of each stop. We filtered the data by making a column that combines the longitude and latitude values into one geometric data that can be read by our query and removing the longitude and latitude columns as they are no longer needed. Source: https://canvas.sydney.edu.au/courses/56224/files/36508335/download?download_frd=1 File Type: TXT
4. Polls - A dataset of polling places for the 2019 Federal election as well as the longitude and latitude values of these polls. Same as what we did with the stops dataset, we filtered the data by making a new column that has the geometric data of the polls and removed the unnecessary longitude and latitude columns. Source: https://canvas.sydney.edu.au/courses/56224/files/36508332/download?download_frd=1 File Type: CSV
5. Schools - A dataset of geographical regions in which students must live to attend primary, secondary and future Government schools. We filtered the data by turning the longitude and latitude position of the areas of all primary, secondary, and future government schools into a geometric data that can be read by our query. Then we made another dataset called 'Schools' where we combined all the primary, secondary and future datasets into one. Source: https://canvas.sydney.edu.au/courses/56224/files/36508333/download?download_frd=1 File Type: zip, shp
6. Population - A dataset of the estimated number of people living in each SA2 by age range. We filtered the data by removing any unknown values. Source: https://canvas.sydney.edu.au/courses/56224/files/36508333/download?download_frd=1 File Type: CSV
7. Income - A dataset of the total earnings for each SA2 area with the mean and median income as well as the total number of earners and their median age. We filtered the data by removing any unknown values. Source: https://canvas.sydney.edu.au/courses/56224/files/36508339/download?download_frd=1 File Type: CSV
8. Wayfind Signage - A dataset that shows all the geographic locations of wayfinding signages in Sydney as well as the type of signs they are (braille, pylon, finger). We filtered this dataset by turning the longitude and latitude position of the areas into a geometric data that can be read by our query. Source: https://data.cityofsydney.nsw.gov.au/datasets/9b0155b887e1410d97c10030d4838209_0/explore?location=-33.889594%2C151.204472%2C13.14 File Type: GeoJSON
9. Bus Shelters - A dataset that shows all geographic locations of bus shelters in sydney. We filtered this dataset by making a column that combines the longitude and latitude values into one geometric data that can be read by our query and removing the longitude and latitude columns as they are no longer needed. Source: https://data.cityofsydney.nsw.gov.au/datasets/3f216ec15ad0498091b335afee5b6537_0/explore File Type: CSV
10. Dog Off-leash Park - A dataset of all off-leash dog parks in sydney as well as the open times of each park and their longitude and latitude position. We filtered this dataset by turning the longitude and latitude position of the areas into a geometric data that can be read by our query. Source:

https://data.cityofsydney.nsw.gov.au/datasets/20b8bb752d7943d98b9704ed66d539fa_0/explore?location=-33.890266%2C151.205909%2C13.25 File Type: GeoJSON

11. Mobility parking - A dataset of all mobility parking spaces in Sydney along with the longitude and latitude position of the parking spaces. We filtered this dataset by turning the longitude and latitude position of the areas into a geometric data that can be read by our query and removing the old geometric data. Source: https://data.cityofsydney.nsw.gov.au/datasets/2671fdce3f354b3e8f02236ca12271cb_1/explore File type: GeoJSON

Database created and used:

Schema used - “Public”
Tables used and Database Diagram -



This diagram shows all the tables used and their attributes. The lines represent how we connected the tables together when doing the query using JOIN functions.

Indexes - Two indexes were used in this schema:

1. school_geom_index - An index used for the table schools and its attribute geom
2. sa2_geom_index - An index used for the table sa2 and its attribute geom

These indexes were used to speed up queries to save time.

Score analysis:
We used a standard formula to calculate the z score which is to take the observation value minus the mean and then divide by the standard deviation.

a. Businesses
   Firstly, we calculated the number of businesses per 1000 people. This is done by taking an average number of businesses in each region divided by the total number of the population and then multiplying by 1000. After that, we start to filter all the regions with less than 100 people to increase the reliability of the data. It is then followed by the standard z score calculation. In this case, we take the value of businesses per 1000 people of each region minus its mean and then divide it by the standard deviation.
b. Stops

We start by counting the number of public transport stops in each region. This is done by joining the population table with the stops table. It is then grouped by the region name and counted for the number of entries (equal to number of stops). After that, we start to filter all the regions with less than 100 people to increase the reliability of the data. It is then followed by the standard z score calculation. In this case, we take the value of stop counts of each region minus its average and then divided by the corresponding standard deviation.

c. Polls

The first thing we did was to determine the number of polling places in each region. This was done by joining the table population, sa2, and polls, then group by each region and count the number of entries (number of polling places). After that, we start to filter all the regions with less than 100 people to increase the reliability of the data. It is then followed by the standard z score calculation. In this case, the number of polling places count is used and the corresponding value of each region is subtracted by the mean and then divided by the standard deviation.

d. Schools

This is one of the most complicated parts in our code. There are three small parts of the code:

- Young people count for each region:
  In this part, we looked into the population table and consider that there are 4 groups which to be considered young people: "0-4_people", "5-9_people", "10-14_people", and "15-19_people". We calculate the total number across the 4 columns and put it as youngpeople_num. In addition, we only take the region with more than 100 people
- School count for each region:
  For this part, we need to use the table young people which we just obtained above. We then join it with the sa2 and left join with schools to check which region does each school belong to. After that, we group by each region and then count the number of entries (number of schools).
- Number of schools per 1000 people for each region and final output table:
  The number of schools per 1000 people in each region is calculated by taking the school count divided by the number of young people for each region and then multiplied by 1000 for a more interpretable number. Lastly, the z score is calculated by taking number of schools per 1000 people for each region minus its mean and then divide by the standard deviation

e. Overall initial score

Firstly, we executed a long sql query which contains all of the tables above, along with a final output table for the score. The overall formula we have is:

$$\frac{1}{1+e^{-(z_{businesses}+z_{stops}+z_{polls}+z_{schools})}}$$

f. Mobility Parking

We first join the population with the sa2 and then join the mob_park. This is to get the coordinates of all the parking slots available and then check which region they are in. We then filter to only choose the regions with more than 100 people to increase the reliability of the data. It is then followed by the standard z score calculation. In this case, the number of parking slots available for each region will be subtracted by the mean and then divided by the standard deviation.

g. Bus Shelter

First, we combine the population data with SA2 regions and then merge it with the bus shelter data. This allows us to obtain the coordinates of all available bus shelters and determine which region they belong to. We then filter the data to select only those regions with more than 100 people, which enhances the reliability of our analysis. Next, we calculate the standard z-score for each region based on the number of bus shelters. This involves subtracting the mean number of bus shelters from the actual number for each region and then dividing by the standard deviation.

h. Wayfind Signage

Initially, we merge population data with SA2 regions and then combine it with information about wayfinding signages. This enables us to identify the coordinates of all available wayfinding signages and determine the regions they belong to. Afterward, we filter the data to include only regions with more than 100 people, thus improving the reliability of our analysis. Following this, we calculate the standard z-score for each region based on the number of wayfinding signages. This involves subtracting the mean number of signages from the actual count for each region and then dividing by the standard deviation.

i. Dog Off-leash Park

Initially, we merge population data with SA2 regions and then combine it with information about Dog Off-leash Parks. This enables us to identify the coordinates of all available Dog Off-leash Park slots and determine the regions they belong to. Afterward, we filter the data to include only regions with more than 100 people, thus improving the reliability of our analysis. Following this, we calculate the standard z-score for each region based on the number of Dog Off-leash Park slots. This involves subtracting the mean number of slots from the actual count for each region and then dividing by the standard deviation.

j. Overall extended score

Firstly, we executed a long sql query which contains all of the tables above, along with a final output table for the score. The overall formula we have is:

$$\frac{1}{1+e^{-(z_{business}+z_{polls}+z_{schools}+z_{stops}+z_{mob\_park}+z_{signage\_query}+z_{bus\_shelter}+z_{DogPark\_query})}}$$

k. Comparison and results review

	sa2_code	score
count	3.600000e+02	360.000000
mean	1.197813e+08	0.402983
std	6.551019e+06	0.291358
min	1.020110e+08	0.029508
25%	1.170316e+08	0.154820
50%	1.205215e+08	0.312936
75%	1.250116e+08	0.602388
max	1.280216e+08	1.000000

F1.1 New score analysis

	sa2_code	score
count	3.600000e+02	360.000000
mean	1.197813e+08	0.452829
std	6.551019e+06	0.269911
min	1.020110e+08	0.047316
25%	1.170316e+08	0.211000
50%	1.205215e+08	0.397600
75%	1.250116e+08	0.640490
max	1.280216e+08	1.000000

F1.2 Original score analysis

Overall, after adding four more dataset to analyse, we can see that most of the numerical values have a decreasing trend. To be specific, the min, mean, 25%, 50%, and 75% all decrease in the new score table. In addition to this, we can see that the standard deviation is actually showing an opposite signal. This means that our classifier formula has a better separation between the regions as the new scores are more spread out compared to the old one.

sa2_name	score
Sydney (North) - Millers Point	1.000000
Erskineville - Alexandria	1.000000
Calga - Kulnura	1.000000
Glebe - Forest Lodge	0.999999
Potts Point - Woolloomooloo	0.999999

F1.3 New top 5 highest

sa2_name	score
Sydney (North) - Millers Point	1.000000
Calga - Kulnura	1.000000
Dural - Kenthurst - Wisemans Ferry	0.999386
Macquarie Fields	0.998152
Ermington - Rydalmere	0.998115

F1.4 Old top 5 highest

The new and old top 5 highest show quite a lot of differences although they have the same top 1 as 3 out of the 5 from both do not match.

sa2_name	score
Wolli Creek	0.029508
Castle Hill - West	0.046810
Woronora Heights	0.047601
Spring Farm	0.048562
Acacia Gardens	0.049444

F1.5 New top 5 lowest

sa2_name	score
Wolli Creek	0.047316
Castle Hill - West	0.074261
Woronora Heights	0.075480
Spring Farm	0.076957
Chippendale	0.077870

F1.6 Old top 5 lowest

The two top 5 lowest however show a similar thing with 4 out of 5 members match but the scores are way less off from each other from our new method.

Correlation analysis:

```
import seaborn as sns
plt.figure(figsize=(10, 6))
sns.regplot(x='median_income', y='score', data=df, scatter_kws={'alpha'
plt.title('Regression Plot of Score vs Median Income')
plt.xlabel('Median Income')
plt.ylabel('Score')
plt.grid(True)
plt.show()
```

A general positive association is seen when comparing our computed scores to statistics on median income for each SA2 region: scores typically climb in parallel with rises in median income.

It's not a very strong or linear relationship, though. Results at similar income levels vary significantly. Although the median income influences the scores, it does not determine or fully explain them. Through additional data collection and analysis, the evaluations would probably benefit from including a wider variety of criteria such as commercial activity levels, transportation patterns, population density, and accessibility to facilities and infrastructure.