# Greater Sydney Analysis

## Abstract

The study explores the liveability of SA2 regions within the Greater Sydney Area through an extensive examination of resource distribution. Leveraging a myriad of credible datasets, the study utilises the sigmoid function of the sum of z score of each metric to reflect the extent of how “well-resourced” an region is. The rationale for this report is to provide a tool for relevant stakeholders, such as urban planners, policy makers, local businesses, and community residents, to guide insights into urban planning, resource allocation, and social-economic policies.

## Dataset Description

### Description

| **Data** | **Description** |
|---|---|
| SA2 Regions: | Statistical Area Level 2 (SA2) digital boundaries. Each SA2 region represents a community in Australia that interacts together economically and socially. |
| Businesses | Records of businesses in Australia by industry and SA2 region, followed by their turnover size ranges. |
| Stops | Locations of all public transport stops (train and bus) in General Transit Feed Specification (GTFS) format |
| Polls | Locations of polling places for the 2019 Federal election. |
| Schools | Geographical regions in which students must live to attend primary, secondary and future Government schools. |
| Population | Estimation of the number of people living in each SA2 region by age range (for ”per capita” calculations). |
| Income | Total earnings statistics by SA2. |
| Safety Camera | Locations of street safety cameras in specific areas of the city of Sydney which have a higher crime rate than other areas. |
| Rain Garden | Locations of installed rain gardens in Sydney to treat stormwater, protect local waterways and green inner-city streets. |

| | |
|---|---|
| Playground | Location of playgrounds and outdoor fitness stations in central Sydney. |

## Sources and Pre-process

**SA2 Region:**
Sourced from Australian Bureau of Statistics. (2021-2026).

**Business**:
Sourced from the Australian Bureau of Statistics, (2018-2022). No changes have been made besides renaming columns to make references shorter such as from industry_code to code.

**Stops**:
Sourced from Transport for NSW. (2016-2022). Unneeded information such as location type, wheelchair_boarding and platform_code are excluded and columns are renamed to shorten the titles such as from ‘stop_id’ to ‘id’. Rows containing any null values are all excluded. Additionally, column ‘geom’, is added to contain points made with columns lat and lng.

**Polls**:
Sourced from Australian Electoral Commission, (2023). Unnecessary data such as ‘state’, ‘premises_address_1’, ‘premises_address_2’, ‘premises_address_3’ and ‘the_geom’ are dropped. Renaming is applied to shorten column names. A new column, ‘geom’, is added containing points made from columns lat and lng.

**Schools**:
Sourced from NSW Department of Education. (2017-2023). All data columns from school_primary, school_secondary and school_future are renamed to uncapitalized letters.

**Population**:
Sourced from Australian Bureau of Statistics. (2021). Columns are renamed to more suited titles and an additional column of ‘young_people’ is added to record the total number of people aged 0 to 19 by summing up the population in columns ‘0-4_people’, ‘5-9_people’, ‘10-14_people’ and ‘15-19_people’.

**Income**:
Sourced from Australian Bureau of Statistics. (2015-16 to 2019-20). Columns containing ‘np’ values are replaced with null values and dropped.

**Security Camera**:
Sourced from City of Sydney Data Hub. (2017-2023). Renamed columns to uncapitalized titles and all null values are excluded.

**Rain Gardens**:
Sourced from City of Sydney Data Hub, (2020-2023). Columns are renamed to uncapitalized titles and rows containing null values are dropped.

**Playgrounds**:

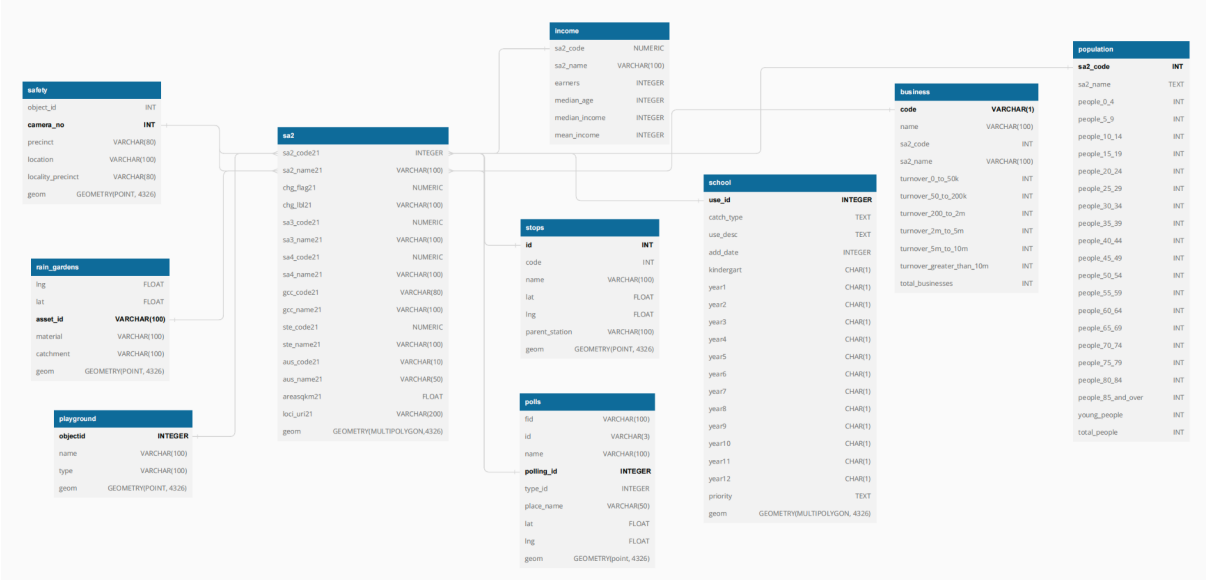
Sourced from City of Sydney Data Hub, (2020-2023). The column, ‘geometry’, is dropped after the addition of a new column, ‘geom’, to the table as this column is made using the data from ‘geometry’.

# Database Description:

The database used contains seven datasets provided by schools and three data sets sought. In the provided data set, Table sa2 contains data for ‘SA2_2021_AUSTGDA2020.zip’ with the primary key ‘sa2_code21’. The business table stores data from the ‘Business.csv’ file with the primary key ‘code’. The primary key of the stop table is the ‘stop_id’ attribute, and the data is stored in the ‘stop.txt’ file. The voting results are listed in ‘Pollsplaces2019.csv’, with ‘polling_id’ as the primary key. The table school contains information for ‘Catchments.zip’, which combines primary school with secondary school and future school in one table, where ‘use_id’ is the primary key. The population table contains information from ‘Population.csv’, with sa2_code as the primary key. The income statement saves information in ‘Income.csv’ format with the primary key ‘sa2_code’.

In the new dataset, the table safety contains the ‘Street_safety_cameras.geojson’ data with a primary key of camera_no. The Table playground contains the data related to ‘Playground.zip’. The primary key is ‘objectid’. The primary key of the table rain gardens stores the information ‘Rain_gardens.csv’ for ‘asset_id’.

After creating the database, we created two indexes on table sa2 and school to speed up the search or sorting the records in the table. The index is created based on the spatial data variable ‘gemo’ in the table, and the neighbourhood datasets are merged and sorted to form a spatial connection.




# Score Analysis

$$Score = S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools})$$

The score is calculated by using a sigmoid function on the sum of the z-score of retail business, health business, stops, polls, and schools. The z-score formula : (x - μ) / σ.

$Z_{retail}$

- Number of retail related businesses per 1000 people in certain SA2 regions.
  - Formula:
    $Total\ Retail\ Businesses\ in\ SA2\ region\ /\ Total\ Population\ of\ SA2\ Region\ *\ 1000$

$Z_{health}$

- Number of health related businesses per 1000 in certain SA2 regions.
  - Formula:
    $Total\ Health\ Businesses\ in\ SA2\ region\ /\ Total\ Population\ of\ SA2\ Region\ *\ 1000$

$Z_{stops}$

- Number of public transport stops in certain SA2 regions.
  - Determine if the geometry of the SA2 region contains or intersects with the point of the stop using ST_Contains and ST_Intersects.

$Z_{polls}$

- Number of polls as of 2019 in certain SA2 regions.
  - Determine if the geometry of the SA2 region contains or intersects with the point of the polls using ST_Contains and ST_Intersects.

$Z_{schools}$

- Number of school catchment areas per 1000 people in certain SA2 regions.
  - Check if the school area is within a certain SA2 region using ST_Contains and ST_Intersects.
  - Future_schools is excluded as it has yet to be built so there is no population

After getting the z-score, the sigmoid function is used to calculate the final score on how 'well-resourced' each region in Greater Sydney is. The closer the score is to 1, the more well-resourced that region is.

| | sa2_code | sa2_name | score |
|---|---|---|---|
| 87 | 117031644 | Sydney (North) - Millers Point | 1.000000 |
| 2 | 102011030 | Calga - Kulnura | 0.999955 |
| 88 | 117031645 | Sydney (South) - Haymarket | 0.999750 |
| 40 | 115021297 | Dural - Kenthurst - Wisemans Ferry | 0.999685 |
| 76 | 117031329 | Darlinghurst | 0.996942 |

According to the calculation, the region with the highest score of 1 is North Sydney, Millers Point. This shows that North Sydney is overall the most 'well-resourced' region in Greater Sydney.

Additional data will then be added to the formula which includes security cameras, rain gardens and playgrounds.

$Z_{camera}$

- Determine the number of security cameras in a certain region using ST_Contains.
- The higher the number of cameras, indicates that the crime rate of the region is higher than others, thus, a negative impact. It will be subtracted from the score.

$z_{playgrounds}$

- Determine the number of playgrounds in a certain region using ST_Contains.
- Playgrounds are public facilities built as a form of outdoor recreation area for children to encourage their fitness and social skills. It is a positive impact that will be added to the score.

$z_{rain\ gardens}$

- Determine the number of rain gardens in a certain region.
- X and Y values are provided and to determine if the region contains the point of the rain garden.
- Rain gardens prevent pollution to waterways and are helpful to the environment. It has a positive impact and is added to the score.

| | sa2_code | sa2_name | new_score |
|---|---|---|---|
| 1 | 117031336 | Surry Hills | 0.995611 |
| 4 | 117031331 | Glebe - Forest Lodge | 0.989295 |
| 0 | 117031329 | Darlinghurst | 0.975468 |
| 2 | 117031639 | Chippendale | 0.909237 |
| 3 | 117031333 | Potts Point - Woolloomooloo | 0.567996 |

After adding the additional data into the score, the results differ from the old score.

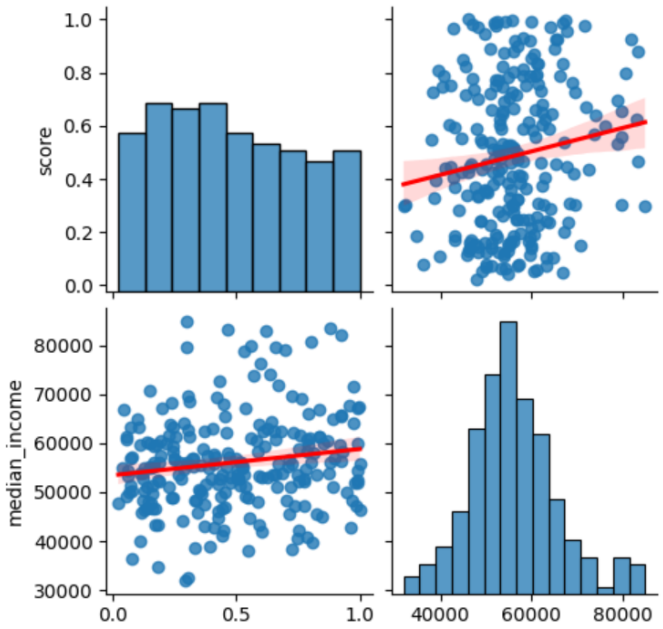
Currently, Surry hills is the most well-resourced region, with a score of 0.99 followed by Glebe - Forest Lodge with a score of 0.98.

(Note: A map visualisation of the old and new score can be obtained at page 6).

# Correlation Analysis:

**Income**

- The median income will be used to compare with the score for each region as mean income is more susceptible to outliers which may result in inaccuracies.




Both scatter plots appear to have a positive correlation but the correlation coefficient between scores and median income is *0.15*, therefore, there is a weak positive relationship between the variables.

As scores increase, median income of each region also increases. However, since the relationship between the two are weak, it also means that as scores change there is little to no effect on the median income.

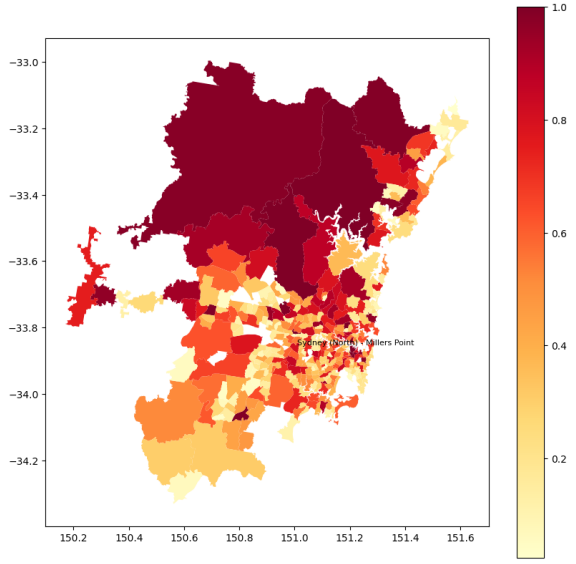
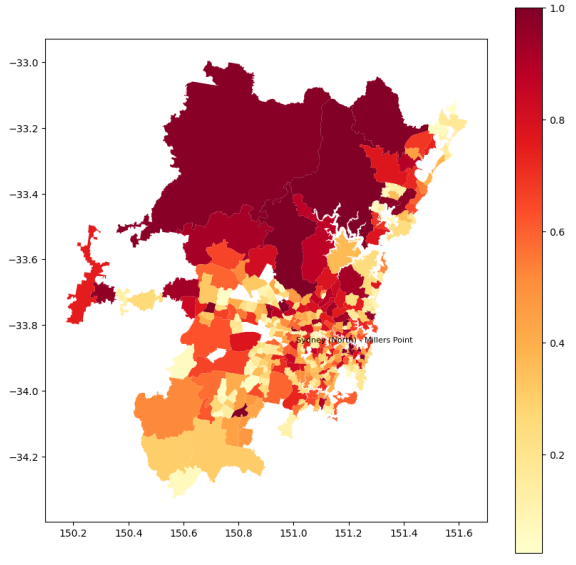
Furthermore, the data points for scores has a vertical trend while the

trend for median income is more horizontal signifying that there is little to no association between income and score. In conclusion, there is no relationship between income and scores.

## Limitation

The result gained from the updated score may be inaccurate as the additional data only records areas within the local government area of the City of Sydney, Central Sydney, and not the entirety of Greater Sydney. Therefore, a clear conclusion can not be made to determine how 'well-resourced' each region in Greater Sydney is. For clearer analysis, more data from all regions in Greater Sydney is required.

# Map Visualization

| | |
|---|---|
| **Old Score**<br>  | The map shows the scores of different regions in which the deeper the shade, the higher the score. A majority of regions with higher scores are located at the top while the lower scores are at the bottom.<br>Miller's Point, the region with the highest score, is a small region located somewhere in between the two.<br><br>The larger areas on top may be the Wollemi national park, Yengo national park and Blue Mountain national park.<br><br>Miller's point may be more well-resourced considering it's a part of central sydney, near to the harbour and has easy access to transportation. |
| **New Score**<br>  | The map has not changed much from the old score. Any changes made to the score are not visible on the map.<br><br>The cause may be that the additional data only involves regions within central sydney and thus does not change the shades of other regions. |

# Reference

Australian Bureau of Statistics. (2018-2022). *Counts of Australian Businesses, including Entries and Exits*. ABS. https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release

Australian Bureau of Statistics. (2015-16 to 2019-20). *Personal Income in Australia.* ABS. https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/personal-income-australia/latest-release

Australian Bureau of Statistics. (2021). *Regional population by age and sex*. ABS. https://www.abs.gov.au/statistics/people/population/regional-population-age-and-sex/latest-release

Australian Bureau of Statistics. (2021-2026). *Statistical Area Level 2*. ABS. https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/main-structure-and-greater-capital-city-statistical-areas/statistical-area-level-2

Australian Electoral Commission. (2023). *Federal Election - Polling Places (Point) 2019.* AURIN. https://data.aurin.org.au/dataset/au-govt-aec-aec-federal-election-polling-places-2019-na

City of Sydney Data Hub. (2017-2023). *Street safety cameras*. City of Sydney. https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::street-safety-cameras/about

City of Sydney Data Hub. (2020-2023). *Rain gardens*. City of Sydney. https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::rain-gardens/about

City of Sydney Data Hub. (2020-2023). *Playgrounds*. City of Sydney. https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::playgrounds/about

NSW Department of Education. (2017-2023). *School intake zones (catchment areas) for NSW government schools.* https://data.cese.nsw.gov.au/data/dataset/school-intake-zones-catchment-areas-for-nsw-government-schools

Transport for NSW. (2016-2022). *Public Transport - Timetables Complete GTFS.* Open data. https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs

ABS. (2021). 2021 Greater Sydney, Census All persons QuickStats | Australian Bureau of Statistics. Abs.gov.au. https://abs.gov.au/census/find-census-data/quickstats/2021/1GSYD