THE UNIVERSITY OF SYDNEY

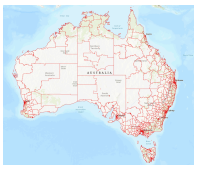
School of Computer Science
Nasim Ahmed, Jarrod Jones

DATA2x01: Data Science, Big Data and Data Variety

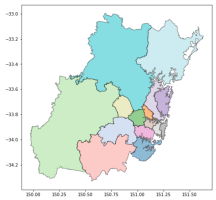
Practical Assignment: Greater Sydney Analysis

Group Assignment (20%) Due: Sunday 18th of May 2025 @ 11:59pm

Introduction



Australia is formally defined by more than 2,000 "Statistical Area Level 2" (SA2) distinct geographical regions, designed to represent communities of between 3,000-25,000 people "that interact together socially and economically". The ultimate aim of this assignment will be to focus on a subset of these regions, and develop a **metric for how "well-resourced" each SA2 region is**, through the integration of multiple publicly available datasets containing information about geographical features.



In Greater Sydney, there are 350+ SA2s, split between 15 SA4 zones (visualised on the left, e.g. "Northern Beaches", "Parramatta"). Each member in your group is to select one of these SA4 zones (i.e. a group of three people requires three distinct SA4 zones) for which scores will be generated based on the guidelines below. This will involve spatially joining the provided datasets with the SA2 boundary information, and also utilising a specific web API to extract information for an additional dataset. Results will be collated in a summative report, which should involve tables and visualisations that are succinct and informative.

Preparation

Form a **group of 3 students** (within your enrolled tutorial where possible, or with your tutor's permission otherwise).

- Initial data loading and cleaning should be completed in **Python**, then **SQL** should be used to merge datasets and produce scores. This code should be collated in a neat, concise **Jupyter notebook** file.
- This unit's Week 8 tutorial covers instructions for managing spatial data and the installation of **PostGIS** (the spatial extension of PostgreSQL) on your local database server.
- This unit's Week 7 tutorial covers instructions for web **API** querying, including examples directly relevant to this assignment.
- A shapefile of the **SA2 digital boundaries** can be accessed on the ABS website here. Use these, alongside the data sources on Canvas, to complete the tasks below.

Tasks

Task 1

Import all datasets (clean if required) into your PostgreSQL server, using a well-defined data schema. These sources include:

- SA2 Regions: Statistical Area Level 2 (SA2) digital boundaries (feel free to filter this down to the "Greater Sydney" GCC).
- Businesses: Number of businesses by industry and SA2 region, reported by turnover size ranges.
- Stops: Locations of all public transport stops (train and bus) in General Transit Feed Specification (GTFS) format.
- Schools: Geographical regions in which students must live to attend primary, secondary and future Government schools.
- Population: Estimates of the number of people living in each SA2 by age range (for "per capita" calculations).
- Income: Total earnings statistics by SA2 (for later correlation analysis).

Note: Ensure spatial datasets consider the correct SRID, which may differ within datasets (e.g. 4283 vs 4326)
Note: It is not essential to ingest all columns for each dataset, but any of potential importance should be retained.
Note: Consult the marking rubric for the full requirements (e.g. around the implementation of indexes).

Task 2

Utilise the NSW Points of Interest API to extract information relevant to each SA2 region and form our additional dataset:

i) Develop a function that returns all points of interests from the API within a specified bounding box of coordinates (this can be done by adjusting the similar function in the Week 7 tutorial which finds all points of interest *near* a specified *point*).

ii) Build a loop that cycles through each SA2 region within your selected SA4 region, waits a second before executing, then runs the function for that region's bounding box, to find all points of interest within that SA2 region.

iii) The cumulative results of this loop should then be similarly ingested into your localhost database, as the final geographic dataset to be used in score calculations. Retain any columns of interest, use the NSW Topographic Data Dictionary as needed for data definitions, and ensure your table is well-defined accordingly.

Task 3

For every SA2 region in your selected zones, compute a score for how "well-resourced" they are according to the formula provided below, where S is the sigmoid function, z is the normalised z-score, and 'young people' are defined as anyone aged 0-19. Feel free to ignore any SA2 regions with a population below 100, if they exist amongst your chosen areas. You are welcome to extend the scoring function however you deem necessary, so long as rational explanation is provided (e.g. other mathematical standardisation techniques, mitigating the impact of outliers, calculating some metrics per-capita or per-sqkm, etc).

As a small means of encouraging extensions of the basic suggested scoring function, note that the z_{business} definition encourages your own selection of specific industries, and the z_{POI} definition similarly allows for "POIgroup" values to be chosen as desired.

$$\text{Score} = S(z_{\text{business}} + z_{\text{stops}} + z_{\text{schools}} + z_{\text{POI}})$$

Metric	Definition	File	Data Source
Business	Businesses per 1000 people, in selected industries	Businesses.csv	Australian Bureau of Statistics
Stops	Number of public transport stops	Stops.txt	Transport for NSW
Schools	Catchments areas per 1000 'young people'	SchoolCatchments.zip	NSW Department of Education
POI	Number of places of interest, in selected groups	*(generated in Task 2)*	NSW Points of Interest API

Task 4

Prepare a summative report presenting thorough analysis of your results. The guidelines below are not necessarily exhaustive nor prescriptive - but would be wise to consider incorporating, as they are reflected in the marking rubric.

- **Summarise** key findings (e.g. Which regions scored highest/lowest? How did scores compare across SA4 zones? Interesting findings?).
- **Visualise** your scores in an engaging way (e.g. How were your scores distributed? What do they look like in a map-overlay visual?).
- **Scrutinise** the results (e.g. Which underlying score components played a significant role? What are the limitations of your scoring?).
- Additional requirement: Determine if any **correlation** exists between your scores and the median income of each region.

Task 5: Advanced Class Only

There are two additional components for DATA2901 students.

1. Create a new version of your score using **ranks** (r) rather than z-scores (z). As a theoretical example, rather than considering a particular SA2 to have 42 public transport stops, you would use the fact that this would rank it 14th of the regions. This will require a new standardisation technique other than the simple sigmoid z-score summation of before, so additionally consider how to convert these values into a comparable, interpretable score. Compare this new score to your previous one from Task 3 - discuss their differences, and conclude which (if any) is more reliable.

$$\text{Score}_{\text{adv}} = f(r_{\text{business}}, r_{\text{stops}}, r_{\text{schools}}, r_{\text{POI}})$$

2. Use a supervised or unsupervised **machine learning** technique to add further depth to your results. This task is intentionally broad to allow creative applications, but some examples could include:
    - A regression model to evaluate which features are statistically significant in predicting the median income of a region.
    - A decision tree classifier to predict the broader SA3 region of a particular SA2 area, given some of its features.
    - An unsupervised clustering algorithm to find similarities between SA2s that might otherwise not be considered alike.

Deliverables

All deliverables are due at the end of Week 11, no later than **11:59pm on Sunday the 18th of May**.

1. PDF Report: This should be no more than 6 pages (plus an optional appendix), in which you document your data integration steps and the main outcomes of your analysis. Your document should contain the following:
    - *Dataset Description*: What are your data sources? How did you obtain and pre-process the data?
    - *Database Description*: How was your schema established (preferably a database diagram included), and how was the data integrated? What index(es) did you create and why?
    - *Score Analysis*: Describe the formula used to compute your score for each region, and give an overview of your results (see the full description of Task 4 for further detail). This section will likely be the longest and most in-depth.
    - *Correlation Analysis*: How well does your score correlate with the median income of each SA2 region? Are these results surprising? Make any final observations about the usefulness or limitations of your scores.
    - *Additional Analysis*: A final section for DATA2901 students, based on their extra requirements.
2. Jupyter Notebook: A file containing your entire data workflow.
3. Short Demo: A brief conversation with your tutor (not a formal presentation) in the Week 11 tutorials (or Week 12, if necessary). This allows time to discuss the decisions behind your work, and is not a marked component, but is mandatory for any marks to be received.

The **marking rubric** will be available on Canvas.
Late submission penalty: -5% of the available marks per day late; minimum 0% after 5 days.
Please submit a **single zip file** containing all deliverables electronically in Canvas, one for each group.
Students must **retain electronic copies** of their submitted assignment files and databases, as the unit coordinator may request to inspect these files before marking of an assignment is completed. If these assignment files are not made available to the unit coordinator when requested, the marking of this assignment may not proceed.

Participation

As a group assignment, the mark awarded for your assignment is conditional on contribution to the group, and a baseline ability to explain the contents of your submission to your teaching team if asked. If members of your group do not contribute sufficiently, please alert your tutor as soon as possible. The tutor will have the discretion to scale the the mark received by an individual as below, based on the outcome of the group's demo.

Level of Contribution	Proportion of Final Grade Received
No participation or no demo	0%
Passive member, but full understanding of the submitted work	50%
Minor contributor to the group's submission	75%
Major contributor to the group's submission	100%

Conclusion

All the best for your assignment! Please direct any questions to our Ed discussion forum's FAQ megathread.