THE UNIVERSITY OF SYDNEY

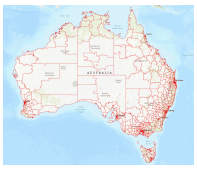
School of Computer Science
Thilina Halloluwa, Jarrod Jones

DATA2x01: Data Science, Big Data and Data Variety

Practical Assignment: Greater Sydney Analysis

Group Assignment (20%) Due: Tuesday 14th of May 2024 @ 11:59pm

Introduction



Australia is formally defined by more than 2000 "Statistical Area Level 2" (SA2) distinct geographical regions, designed to represent communities of between 3000-25000 people "that interact together socially and economically". In this assignment, we'll focus on the 350+ SA2s within the Greater Sydney area, and you will be tasked with spatially integrating several datasets of various formats to calculate a score for each region.

The picture on the left, provided by State Records NSW, is set in the 1950s, and entitled "Bustling Sydney" - an interesting way to describe our city. In many respects, it is quite "bustling" indeed, but the argument could easily be made that the appeal of Sydney is that it doesn't at all feel like a big city, given its close proximity to natural beauty (beaches, national parks, etc), overall low population density, and relatively small CBD area. Your task in this assignment is to develop a **"bustling" metric for each SA2 region of Greater Sydney**, in an attempt to quantify just how busy the districts within our city are.

Preparation

Form a **group of 2-3 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.
- 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.
- Polls: Locations (and other premises details) 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: 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)

Task 2

Compute a score for how "well-resourced" each individual neighbourhood is according to the formula provided on the next page, 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 only calculate scores for SA2 regions with a population of at least 100, and 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 is intentionally broad - select a cross-section of specific industries within the provided dataset (e.g. "Retail Trade") that you believe will be the best reflection of how "bustling" the area is (describe your rationale in the report) and use this to calculate the component.

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

Metric	Definition	File	Data Source
Business	Selected industry businesses per 1000 people	Businesses.csv	Australian Bureau of Statistics
Stops	Number of public transport stops	Stops.txt	Transport for NSW
Polls	Federal election polling locations (as of 2019)	PollingPlaces2019.csv	Australian Electoral Commission
Schools	School catchments areas per 1000 'young people'	SchoolCatchments.zip	NSW Department of Education

Task 3

Extend the score by sourcing **one additional dataset for each group member**, and then incorporating all new datasets into your scoring function. For full marks, at least one dataset should be of spatial data, and at least one should be of a type not used so far in this assignment (e.g. JSON, XML, or collated via web scraping). Almost any subject matter is permissible, so long as it can be justified as relevant to the calculation of our "bustling" metric (e.g. public facilities, other census statistics, local wildlife, etc).

For either version of your scoring function (or both!), the following subtasks should also be achieved:

- **Visualise** your score in an engaging way, and summarise key results in a table (ideally including a useful map-overlay visualisation, or an interactive graph).
- Include **in-depth analysis** into your results. Note interesting findings, discuss their limitations, and summarise key conclusions.
- Determine if there is any **correlation** between your score and the median income of each region.
- Ensure at least one useful **index** (ideally spatial) has been used for your calculations.

Task 4: 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 2 - discuss their differences, and conclude which (if any) is more reliable.

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

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 in Week 12, no later than **11:00pm on Tuesday the 14th 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 (including how it was extended with extra datasets), and give an overview of your results. This section will likely be the longest and most detailed.
    - *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 12 tutorials (or Week 13, 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.