# Dataset Description

Data source: The shapefile of SA2 digital boundaries, and four metrics used to finish Task 2 are downloaded from the links in the spec sheet.

Extra 1: This provides the number of spatial information of walking count sites in several SA2 regions.

<https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::walking-count-sites/explore>

GeoJSON format which contains spatial data is chosen, satisfying two criteria at once.

Extra 2: [This provides the number of dwelling units and value of buildings approved in SA2 regions.](https://www.abs.gov.au/statistics/industry/building-and-construction/building-approvals-australia/latest-release)

<https://www.abs.gov.au/statistics/industry/building-and-construction/building-approvals-australia/latest-release>

Note that this XLSX file could not be directly used, so we removed some unwanted rows.

Obtain data: Saving the files in the same folder, and then imported them into python using corresponding modules, step by step in our Jupyter file.

Preprocessing SA2 regions and Extra 1 data:

* GeoPandas to convert “polygon” into “multi polygon” for geo-spatial data as the latter is more compatible.
* SA2\_code should be converted into integer type.

Preprocessing four metrics:

* Some columns in population/business data start with numbers, which is not allowed in SQL. So we have to rename those columns. We also rename other columns for visibility.
* Converting datetime data in school data.

Common preprocessing approaches:

* Converting geometry data into WKT so that it can be loaded to SQL.
* Converting the column names into lowercase in PostgreSQL, which is a good convention.
* Removing columns that will not be used to make the table cleaner.
* Removing some invalid, empty data.

# Database Description

Attributes in **bold** are primary keys. Arrows signify foreign keys.

Index created: We created indexes for … for which table …

Reason: The primary reason is to make data retrieval faster. Without indexing, the database system would have to scan the entire collection of rows, which might take minutes and therefore inefficient when faced with massive queries. Indexing can also enforce constraints such as uniqueness (E.g., we do not want same SA2 code to be applied to different rows), improving data integrity and consistency.

# Score Analysis

Original: Score = S(zbusiness + zstops + zpolls + zschools)

Extra 1: Score with counter = S(zbusiness + zstops + zpolls + zschools) + 0.01 \* count

The walking counters are mainly in CBD regions where the areas are small so there are few bus stops so we modify the score accordingly to compensate for the loss.

Extra 2: Score with buildings = S()

The

# Correlation Analysis

Extra 1:

The result is

The limitation of this is that the dataset only provides a small number of SA2 regions with counting objects, limiting our evaluation for all regions.

Extra 2: The scatter plot and the Pearson correlation coefficient -0.040075960483588614, signify an **extremely weak negative relationship**. The reason might be that SA2 regions are of different sizes, if the polls, schools, stops are distributed unequally then the median income of each SA2 will have nothing to do with the bustling score.

The result is a bit surprising as one may expect regions with higher median income to have a higher bustling score. One explanation for this is that people with higher income might bustle somewhere else (

The usefulness is that it does provide some measurements on how bustling each SA2 region is, compared to ambiguous plain language.

One limitation of our score is that there might be time delay in the bustling as approved buildings take time to finish construction and contribute to the bustling level of the region. Another might be investment in building might not

<https://docs.google.com/document/d/1ngUkQpjBVBXy7e50qf9TAfYfv7Y6PN-y0zHAXAQSuOE/edit?usp=sharing>

Todo

Finish extra 1

Extra 2 analysis

Overview of results (mean, median, outliers)

Visualization of score