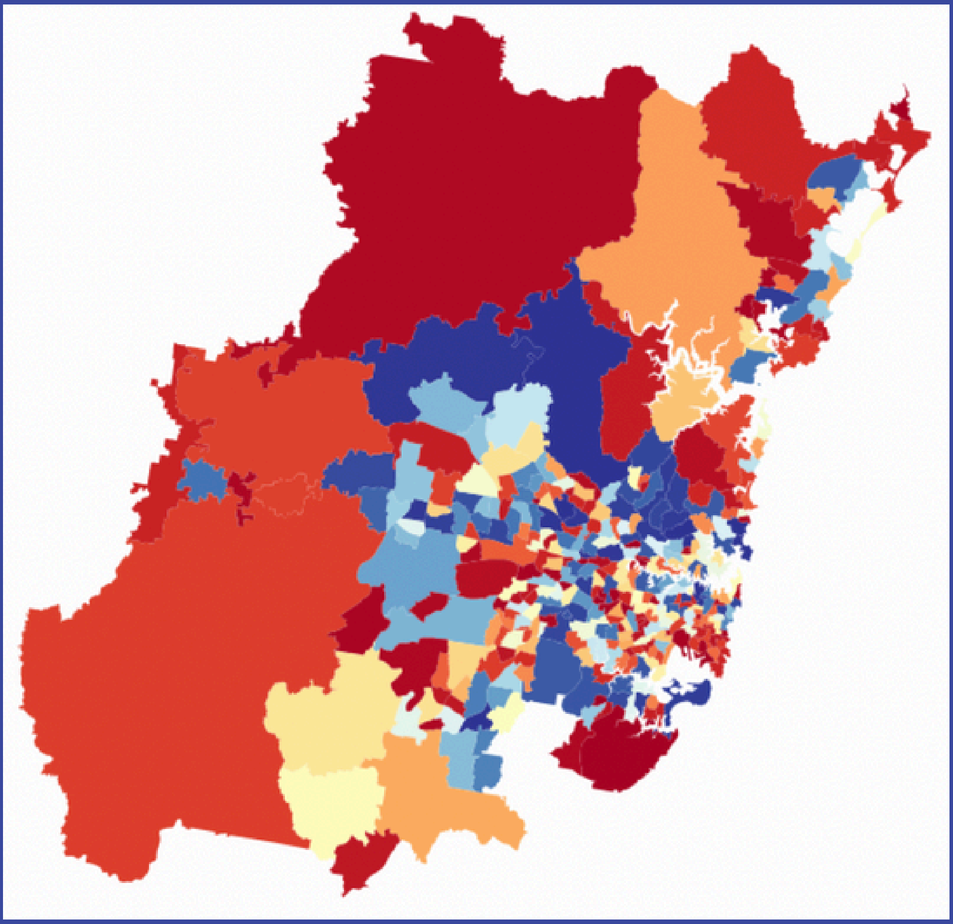
DATA2001

A Bustling City



Datasets

SA2 REGIONS

The ABS SA2 region dataset contains the geospatial boundaries of each SA2 region. Columns describing different statistical areas (SA3, etc.) were removed as SA2 is the only relevant category for this purpose. Geospatial data format was adjusted for ingestion.

INCOME

The ABS income dataset provides the average income for each SA2 region, data types were converted for ingestion.

CATCHMENT

The Department of Education catchments dataset contains geospatial data for primary and secondary school catchment areas. Primary and Secondary school data was combined into one table, and geospatial formats were updated for ingestion.

BUSINESSES

The ABS stops dataset provides a list of bus stops including their locations. Columns describing the stop (location time, wheelchair accessibility, etc.) were removed as the quantity of stops within SA2 regions is our focus. Geospatial data format was adjusted for ingestion.

ALCOHOL LICENCES

The Department of Customer Service's licenses dataset lists all premises licensed to provide alcohol. Columns describing the premise or the nature of the license were excluded. Geospatial format was adjusted for ingestion.

POPULATION

The ABS populations dataset provides the population by age group for every SA2 region. Age group data columns for people between 20-85+ were removed, as only 'young people' (ages 0-19), compared to the total population are relevant.

POLLING

The AEC polling places dataset lists all polling places with their geospatial point data. Geospatial data was reformatted for ingestion. Null values for latitude and longitude were excluded, as they represent moving polling stations, typically for nursing homes, and can move between multiple SA2 regions. Substitution of average or common values was not done to avoid inaccurate geospatial data.

STOPS

The Transport for NSW stops dataset provides a list of bus and train stops including their locations. Unnecessary columns were removed as the quantity of stops within SA2 regions is our focus. Geospatial format was adjusted for ingestion.

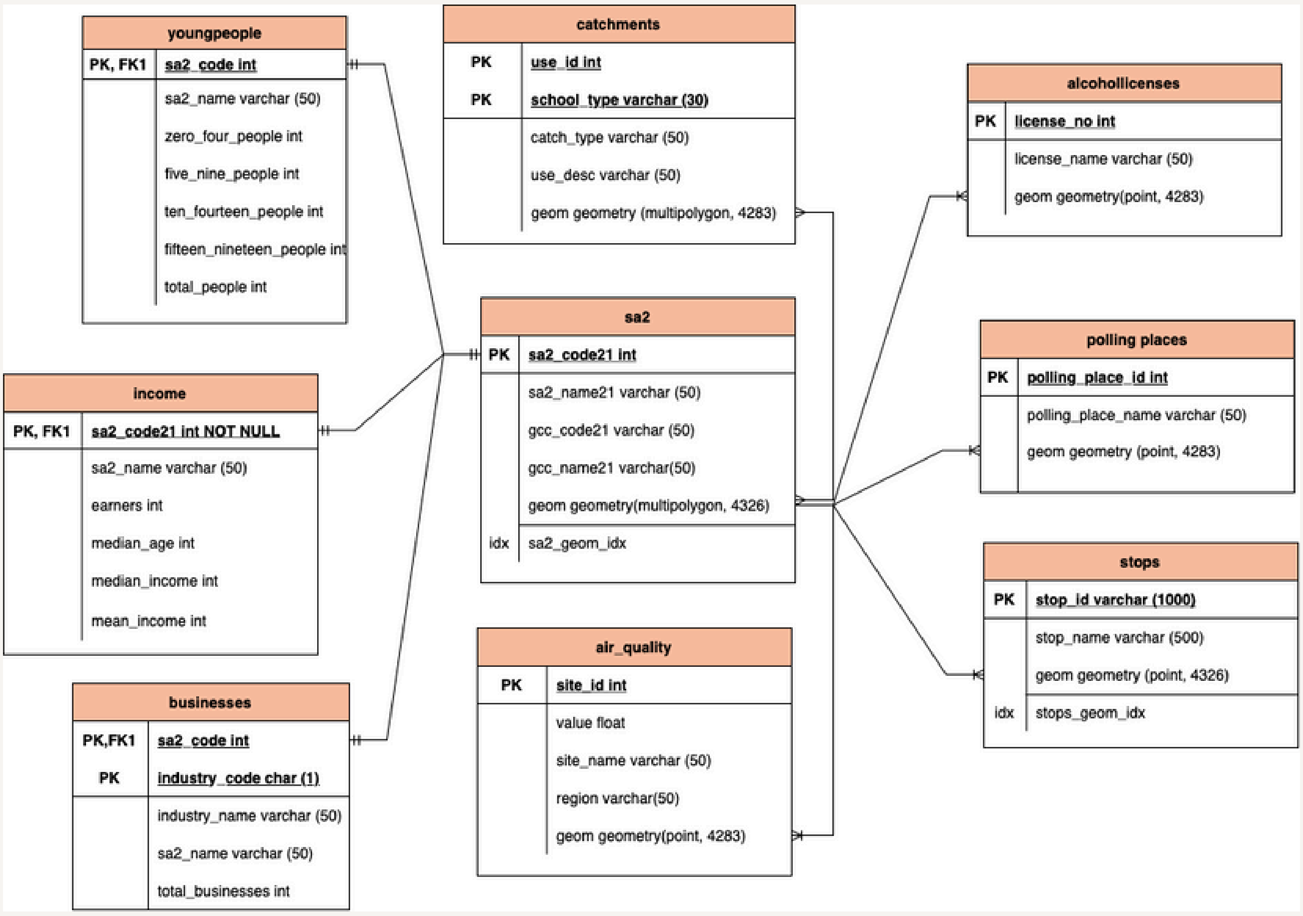
AIR QUALITY

The air quality dataset provides ozone readings for each measurement station in NSW. Station locations and ozone readings were taken from the NSW Air Quality API, converted from JSON format. Both tables were adjusted to merge together, then the merged dataset was ingested.

Database

The database was established from our initial datasets: Population, Catchments, SA2, Income, Businesses and Stops. The schema centres around the SA2 region boundary dataset, as the aim was to correlate all other data with an SA2 code. Geospatial Indexes were created on the SA2 Boundary and Stops to improve efficiency as SA2 joins spatially with 5 other datasets and Stops consists of over 110,000 geospatial points, slowing operations . Additional Datasets of Alcohol Licenses and Air Quality were later added and joined to SA2 via a spatial join.

The full database schema is detailed in Figure 0.1



youngpeople	
PK, FK1	sa2_code int
	sa2_name varchar (50)
	zero_four_people int
	five_nine_people int
	ten_fourteen_people int
	fifteen_nineteen_people int
	total_people int

catchments	
PK	use_id int
PK	school_type varchar (30)
	catch_type varchar (50)
	use_desc varchar (50)
	geom geometry (multipolygon, 4283)

alcohollicenses	
PK	license_no int
	license_name varchar (50)
	geom geometry(point, 4283)

sa2	
PK	sa2_code21 int
	sa2_name21 varchar (50)
	gcc_code21 varchar (50)
	gcc_name21 varchar(50)
	geom geometry(multipolygon, 4326)
idx	sa2_geom_idx

income	
PK, FK1	sa2_code21 int NOT NULL
	sa2_name varchar (50)
	earners int
	median_age int
	median_income int
	mean_income int

polling places	
PK	polling_place_id int
	polling_place_name varchar (50)
	geom geometry (point, 4283)

stops	
PK	stop_id varchar (1000)
	stop_name varchar (500)
	geom geometry (point, 4326)
idx	stops_geom_idx

air_quality	
PK	site_id int
	value float
	site_name varchar (50)
	region varchar(50)
	geom geometry(point, 4283)

businesses	
PK,FK1	sa2_code int
PK	industry_code char (1)
	industry_name varchar (50)
	sa2_name varchar (50)
	total_businesses int

Figure 0.1: Database Schema - ER diagram

Score Analysis

Computation

Firstly, standard scores (z-score) for the following features were calculated:

- The total number of businesses per 1,000 people in each SA2 region. The industries included were ‘Accommodation and Food Services’, ‘Retail Trade’, ‘Arts and Recreation Services’, ‘Professional, Scientific, and Technical Services’, ‘Education and Training’, and ‘Health Care and Social Assistance’. These industries were chosen as they involve direct interaction with the average resident.
- The total number of school catchments per 1,000 young people (i.e., people below the age of 20) in each SA2 region.
- The number of bus stops in each SA2 region.
- The number of polling places in each SA2 region.

Then, a sigmoid function was applied to the z-scores to generate a score between 0 and 1 for each SA2 region. The higher the score, the more bustling the region is said to be. The formula used is as follows:

$$\frac{1}{1+e^{-x}}$$

where x is the sum of the z-scores for each region.

Extension

Two further datasets were included. The first dataset contained information about alcohol licenses. As nightlife is an important measure of how bustling an area is, the number of premises licensed to provide alcohol in each region was seen as a good measure to include. The second dataset contained information about ozone levels. As higher ozone indicates more activity, the level of ozone in each region was seen as a good measure to include as well. However, a major limitation of this dataset is that it does not cover data for most regions in Sydney.

From these two datasets, the z-scores for the number of premises licensed to provide alcohol in each SA2 region and ozone levels in each SA2 region were calculated. These values were then included in the sigmoid function.

Overall, as Fig. 1.2 and Fig 1.3 show, the extension does not seem to have impacted the distribution of Bustling Scores in a significant manner.

Trends and distribution

With the extra datasets included in the Bustling Score, the SA2 regions in Greater Sydney with the highest scores are:

- Sydney (North) - Millers Point
- Sydney (South) Haymarket
- Dural - Kenthurst - Wisemans Ferry
- Chatswood (East)
- Surry Hills

Meanwhile, the SA2 regions with the lowest scores were:

- Lilli Pilli - Port Hacking - Dolans Bay
- Woronora Heights
- Putney
- Strathfield South
- Yennora Industrial

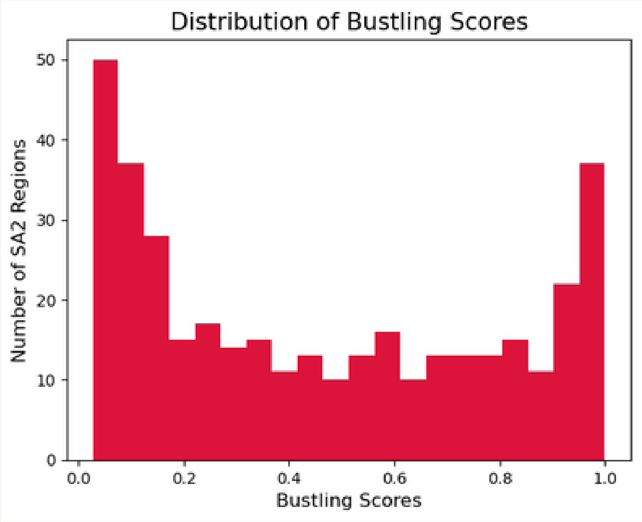


Figure 1.1: The distribution of bustling scores

Fig. 1.3 shows that there is a wide range of activity levels across the city. However, the most bustling areas tend to surround the city center. This makes sense: four of the most bustling areas are quite central, whilst all five of the least bustling areas are further out. Wisemans Ferry seems to be the only exception, but only because it has the highest number of school catchments per thousand young people out of any SA2 region.

Furthermore, Fig 1.1 shows that the distribution of activity levels across the city is bimodal and U-shaped. This means that there are many regions with very little activity and many regions with lots of activity. This largely makes sense. On the one hand, there are many residential and industrial areas, as well as parks, across the city where there is very little activity. On the other hand, there are also many areas where socioeconomic activity is very concentrated.

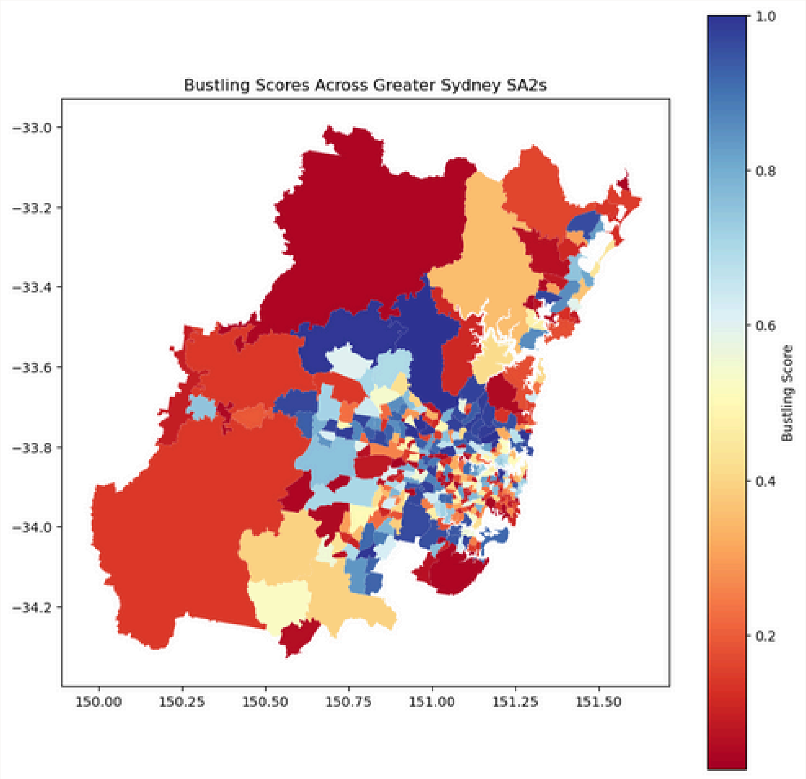


Figure 1.2: The map **before** the inclusion of the extra datasets

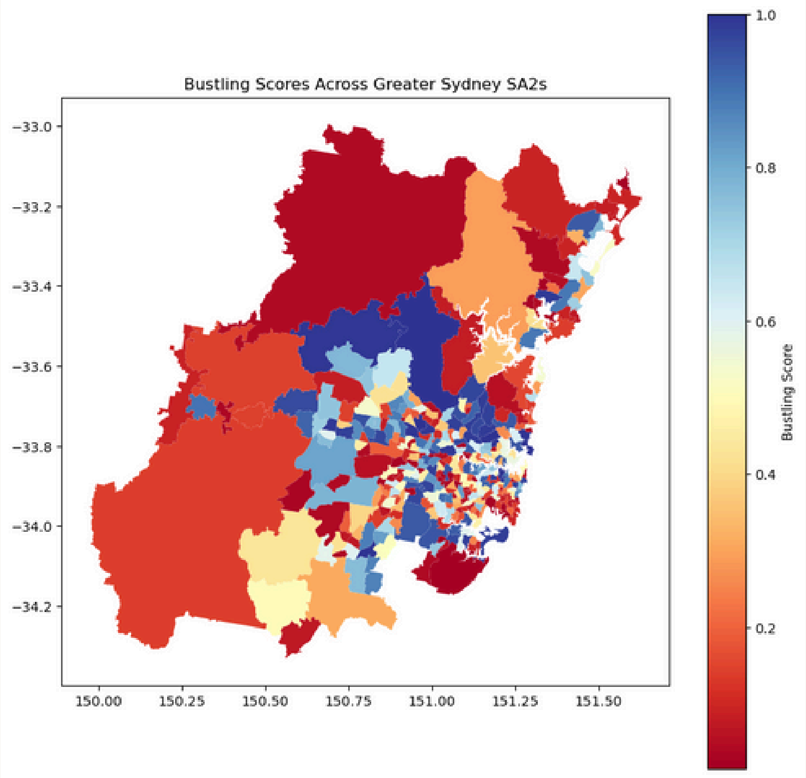


Figure 1.3: The map **after** the inclusion of the extra datasets

Correlation Analysis

Statistical Test

The evidence suggests that there is no significant correlation between the median income and Bustling scores

1. A scatterplot was produced representing the median income and the Bustling Score, then a linear regression was fitted to the data (Figure 2.1).
2. Visually there was no correlation, which was confirmed by a very weak Pearson correlation coefficient of 0.03.
3. A residual plot was produced (Figure 2.2) with no pattern shown: there is no suggestion of alternate correlation.

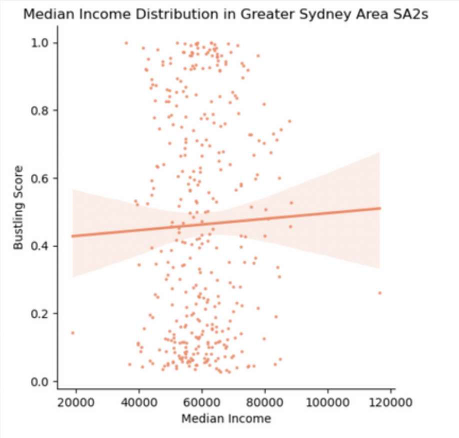


Figure 2.1: Scatterplot of relationship between Income and Bustling Score

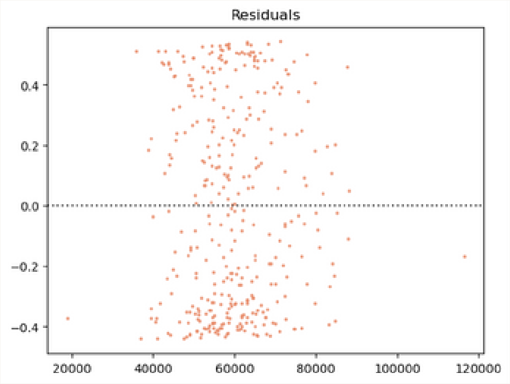


Figure 2.2: Residual plot of relation

Summary

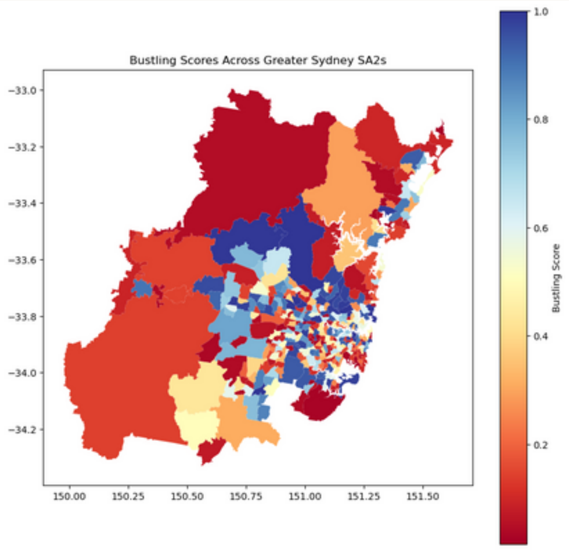


Figure 2.3: Heatmap of Bustling Scores

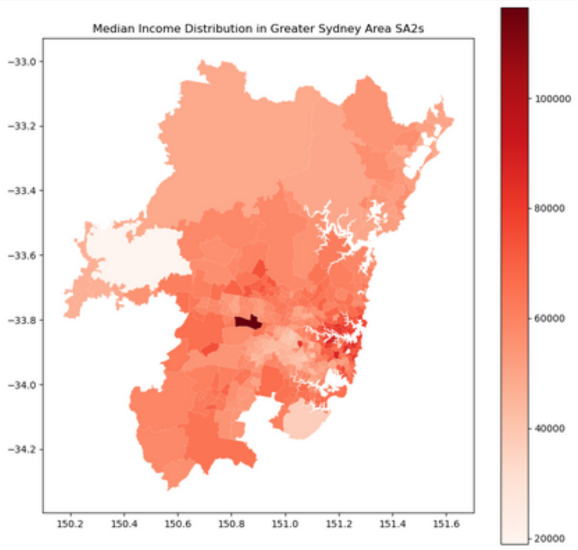


Figure 2.4: Heatmap of Median Income

The lack of correlation should not be surprising. When comparing a map view of the Bustling Score distribution (Figure 2.3) with a map view of the income distribution (Figure 2.4), higher income levels are concentrated closer to the City of Sydney, whereas activity levels are more widely distributed throughout Greater Sydney. A significant exception is the Prospect Reservoir, with a very low population of high earners presumably as the majority are specialists for the Prospect Dam.

This may be because an area can be valuable to a high-income earner regardless of its level of activity. For example, SA2 regions in Sydney's Eastern Suburbs (Rose Bay, Bellevue Hill, etc.) have a low level of activity but very high incomes. Conversely, the CBD has both a very high level of activity level and also very high levels of income. With people of high incomes choosing to live in areas with both the highest and lowest activity levels, it seems likely that both ends of the scale of activity can be a draw to an area. If this were true, a higher salary would not necessarily correlate with any particular level of activity.

The bustling metric seems to be appropriate. However, there are limitations. The metric can be impacted by the area of the SA2 region. For example, Dural - Kenthurst - Wisemans Ferry has a very high catchments score, as it covers a large area with a low population size. This inflated its bustling score when its other scores were quite low, limiting the score's effectiveness.