Dataset Description

SA2 Regions: The digital boundaries of SA2 serve as a geographical framework shapefile for location-specific analysis, in which a comprehensive and detailed perspective on the various regions within Greater Sydney is offered.

Businesses: CSV regarding the number of business locations categorised by industry and region, classified based on turnover size.

Stops: Contains locations of all public transport (train and bus) stops, as in the General Transit Feed Specification (GTFS) format. This facilitates a thorough analysis of transportation networks with insights of public transportation services' accessibility and availability across different regions.

Polls: Contains details and information for the 2019 Federal election in terms of the locations of polling places. Valuable insights can be gained regarding the polling places' accessibility and distribution.

Schools: Contains geographical regions outlining specific primary, secondary, and future Government schools catchment areas. This can be useful to facilitate the education sector's informed decision-making with the educational institutions' distribution and accessibility.

Population: Provides estimations of the population residing in each SA2 region, categorised by age range, based on "per capita" calculations and demographic analysis. This provides insights of the population dynamics within different SA2 regions.

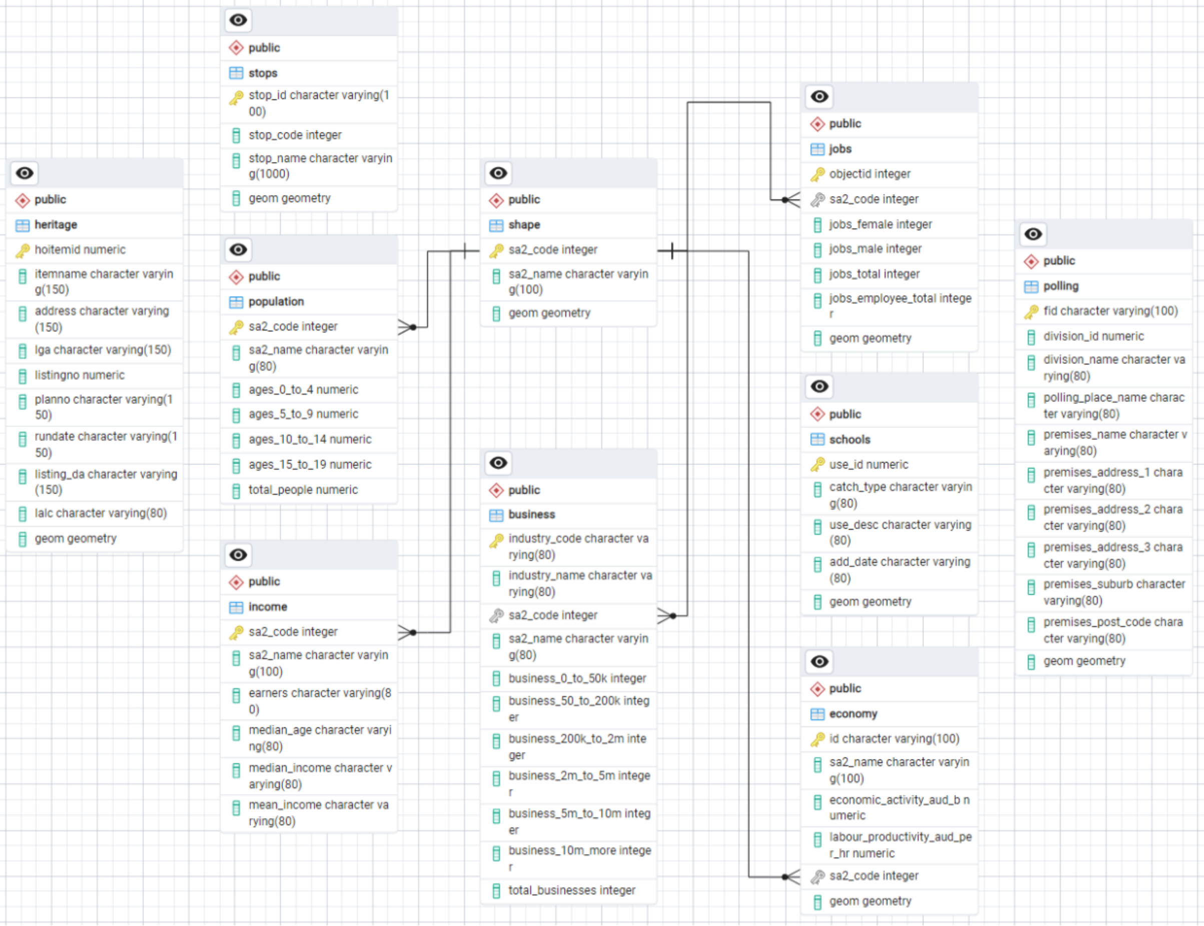
Income: Provides statistical insights on total earnings within each SA2 region, which is further utilised for correlation analysis with other socio-economic factors.

Employment [13]: ABS Education and employment themed data items, published 2022 by Statistical Areas Level 2 (SA2). Sourced from data from both Australian Bureau of statistics (ABS) and non-ABS, derived from the 7 November 2022 release of data by region. Includes a geojson file which presents employment data grouped within data by sa2 region region. Included data which gives insights of jobs in Australia and labour force status of each region.

Heritage [14]: State Heritage Register - Centroids, 2020. Published on SEED, The Central Resource for Sharing and Enabling Environmental Data in NSW on behalf of the Heritage Council of NSW. Contains the geographical locations of State Heritage Register listings in an ESRI shapefile (.shp) format. This provides the location and listing dates of heritage buildings or land that must not be demolished, damaged or moved.

Economy [15]: GI - Mapping Australia's Economy (SA2) 2014, generated by the Grattan Institute in relation to Mapping Australia's Economy with original data from ASGS 2011 SA2 Digital Boundaries. A Json file which provides economic activity analysis within regions of NSW, in which data is compiled at the Statistical Area Level 2. Aimed to showcase economic activity clustering in urban cores and inner regions.

Database Description



public
stops
stop_id character varying(100)
stop_code integer
stop_name character varying(1000)
geom geometry
public
heritage
hoitemid numeric
itemname character varying(150)
address character varying (150)
lga character varying(150)
listingno numeric
planno character varying(150)
rundate character varying(150)
listing_da character varying (150)
laic character varying(80)
geom geometry
public
population
sa2_code integer
sa2_name character varying(80)
ages_0_to_4 numeric
ages_5_to_9 numeric
ages_10_to_14 numeric
ages_15_to_19 numeric
total_people numeric
public
income
sa2_code integer
sa2_name character varying(100)
earners character varying(80)
median_age character varying(80)
median_income character varying(80)
mean_income character varying(80)
public
shape
sa2_code integer
sa2_name character varying(100)
geom geometry
public
business
industry_code character varying(80)
industry_name character varying(80)
sa2_code integer
sa2_name character varying(80)
business_0_to_50k integer
business_50_to_200k integer
business_200k_to_2m integer
business_2m_to_5m integer
business_5m_to_10m integer
business_10m_more integer
total_businesses integer
public
jobs
objectid integer
sa2_code integer
jobs_female integer
jobs_male integer
jobs_total integer
jobs_employee_total integer
geom geometry
public
schools
use_id numeric
catch_type character varying(80)
use_desc character varying (80)
add_date character varying (80)
geom geometry
public
economy
id character varying(100)
sa2_name character varying(100)
economic_activity_aud_b numeric
labour_productivity_aud_per_hr numeric
sa2_code integer
geom geometry
public
polling
fid character varying(100)
division_id numeric
division_name character varying(80)
polling_place_name character varying(80)
premises_name character varying(80)
premises_address_1 character varying(80)
premises_address_2 character varying(80)
premises_address_3 character varying(80)
premises_suburb character varying(80)
premises_post_code character varying(80)
geom geometry

Figure 1.1 *- representing our database schema diagram generated from pgAdmin*

The overall schema was designed according to the entities for storing information. These include, shape, stops, income, population, schools, polling, business, heritage, jobs, and economy, in which they are represented as tables in the database. Primary keys (the yellow key in the diagram) are defined according to the unique identifiers in each table, while foreign keys (the greyish double keys in diagram) are utilised as reference to specific attributes. With data integration into tables from our source files, we focus on the “sa2_code” attribute to establish relationships between tables. Moreover, two spatial indexes are created on the “geom” column for both “shape” and “stops” tables. This is due to these fields being accessed frequently when performing queries, in which indexes are able to speed them up.

Results Analysis

S score formulation.

The score for each region was calculated using the following formula,

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

where the z-score refers to the standard score that quantifies the number of standard deviations a data point is away from the mean. The Z-score is essential within this analysis as it normalises the variables by transforming them into a common scale. This normalisation is crucial, especially when all variables have different units, scales, and weightings, ensuring comparability across datasets.

The following z scores were calculated; z score for the number of retail trade businesses per 1000 people in each sa2 region code, z score for the number of public transport stops per sa2 code, z score for the number of polling places per sa2 code and the z score for the number of schools per 1000 young people (aged 0 to 19) per sa2 code. The Z score was calculated through SQL using the equation $Z = \frac{x - \mu}{\sigma}$, which was computed as $(df - avg(df))$ / $stddev\_pop(df)$.

The sum of all these z scores from each category were put into the Sigmoid function which produces a 'bustling score' for each sa2 region. The Sigmoid function, $S(x) = \frac{1}{1+e^{x}}$ produces a number in between 0 and 1 where values closer to 1 indicate a more bustling region and serves to normalise and make an interpretative finalised score. This function was computed through the CREATE FUNCTION SQL command.

This formula was extended to incorporate three additional datasets, an employment geojson, a heritage shapefile and an economy json file. The implications of extending the base formula with these additional datasets include improved model accuracy as it incorporates a more diverse range of factors that might influence how busy a region is. By analysing more datasets, a deeper understanding and greater insights can be gained into how well-resourced each individual neighbourhood is. However, by integrating multiple datasets, several issues may arise such as data inconsistencies, quality issues and incompatibility. In this investigation, all datasets were thoroughly cleaned and pre-processed to maintain the integrity.

Results

A map overlay visualisation was created to illustrate the bustling score per SA2 region using varying shades of orange to represent how well-resourced each neighbourhood is, as illustrated in Figure 1.2. Through this visualisation, it becomes apparent that central Greater Sydney exhibits a significantly higher bustling score, with values closer to 1, while regions further west and toward the outskirts of Sydney tend to have scores closer to 0. Colloquially, central Sydney is often perceived as more metropolitan compared to its outskirts. This is also shown in figure 1.6 (in appendix) which illustrates the bustling score in an interactive graph.

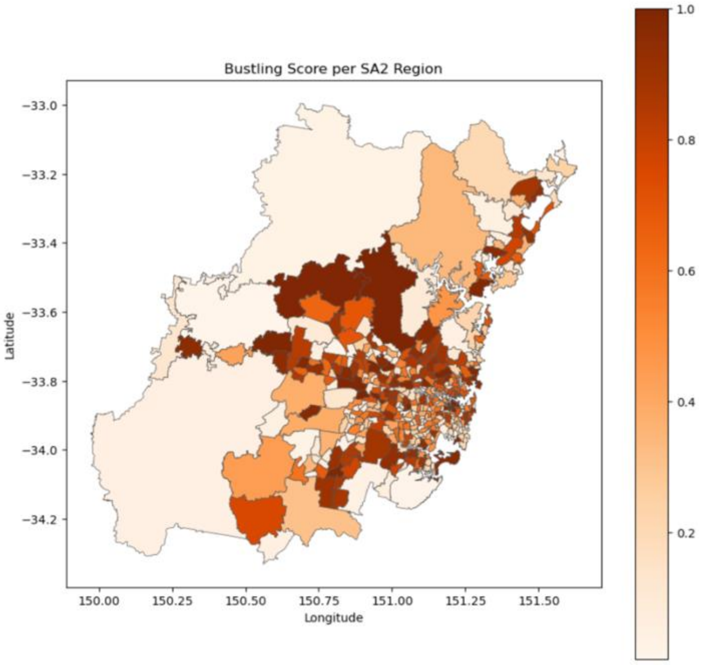
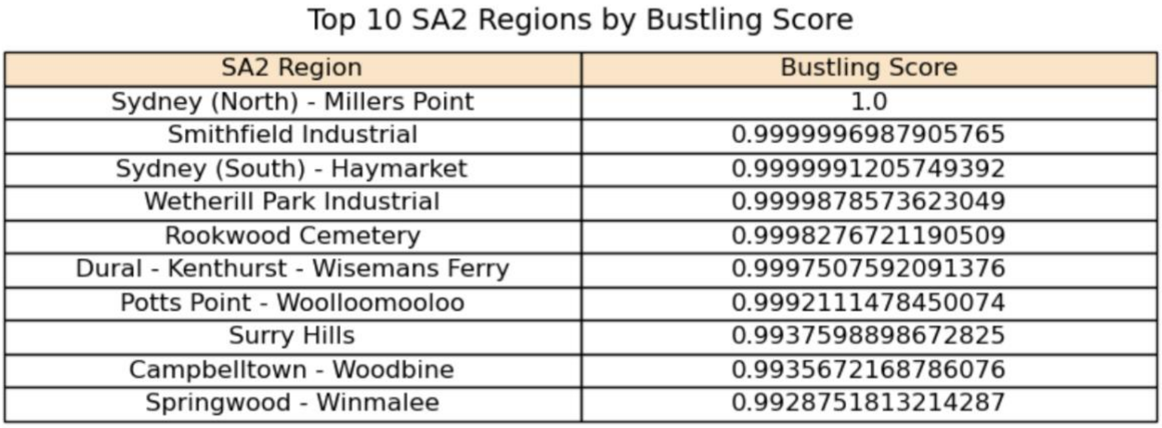


***Figure 1.2** – bustling score mapped by SA2 Region*

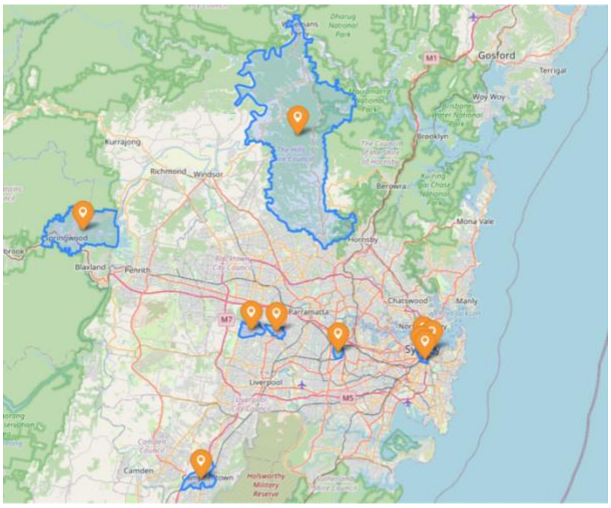


Top 10 SA2 Regions by Bustling Score

SA2 Region	Bustling Score
Sydney (North) - Millers Point	1.0
Smithfield Industrial	0.9999996987905765
Sydney (South) - Haymarket	0.9999991205749392
Wetherill Park Industrial	0.9999878573623049
Rookwood Cemetery	0.9998276721190509
Dural - Kenthurst - Wisemans Ferry	0.9997507592091376
Potts Point - Woolloomooloo	0.9992111478450074
Surry Hills	0.9937598898672825
Campbelltown - Woodbine	0.9935672168786076
Springwood - Winmalee	0.9928751813214287

Figure 1.3 *- Top 10 SA2 Regions sorted by Bustling Score*

As illustrated by figure 1.3, Millers Point within North Sydney is identified as the most bustling region. The table shows how industrial areas such as Smithfield Industrial and Wetherill Park Industrial also scored remarkably high using the bustling score metric. However, Rookwood Cemetery surprisingly received an unusually high bustling score, highlighting a potential flaw in relying solely on Z-scores for normalising a diverse range of variables. This highlights the potential loss of intuitive interpretation when using Z-scores in such contexts. To explain this anomaly, it is plausible that the high density resulting from the specialisation of this SA2 region as a cemetery could have significantly inflated the bustling score for Rookwood Cemetery. Further investigation and testing are necessary to confirm this hypothesis and to identify any other contributing factors.



Another notable trend within the top 10 regions is the prominence of bustling regions within the Sydney Central Business District (CBD). As illustrated by Figure 1.4, a clear cluster of four SA2 Regions within the top 10 can be seen condensed within the CBD: Sydney (North) - Millers Point, Sydney (South) - Haymarket, Potts Point - Woolloomooloo, and Surry Hills. Considering the CBD is the centre of commerce and businesses, these results are to be expected based on the metrics of our S score calculation.

Figure 1.4 *– Top 10 most bustling SA2 regions mapped.*
Accessed: data2001-assignment.zip

Correlation Analysis

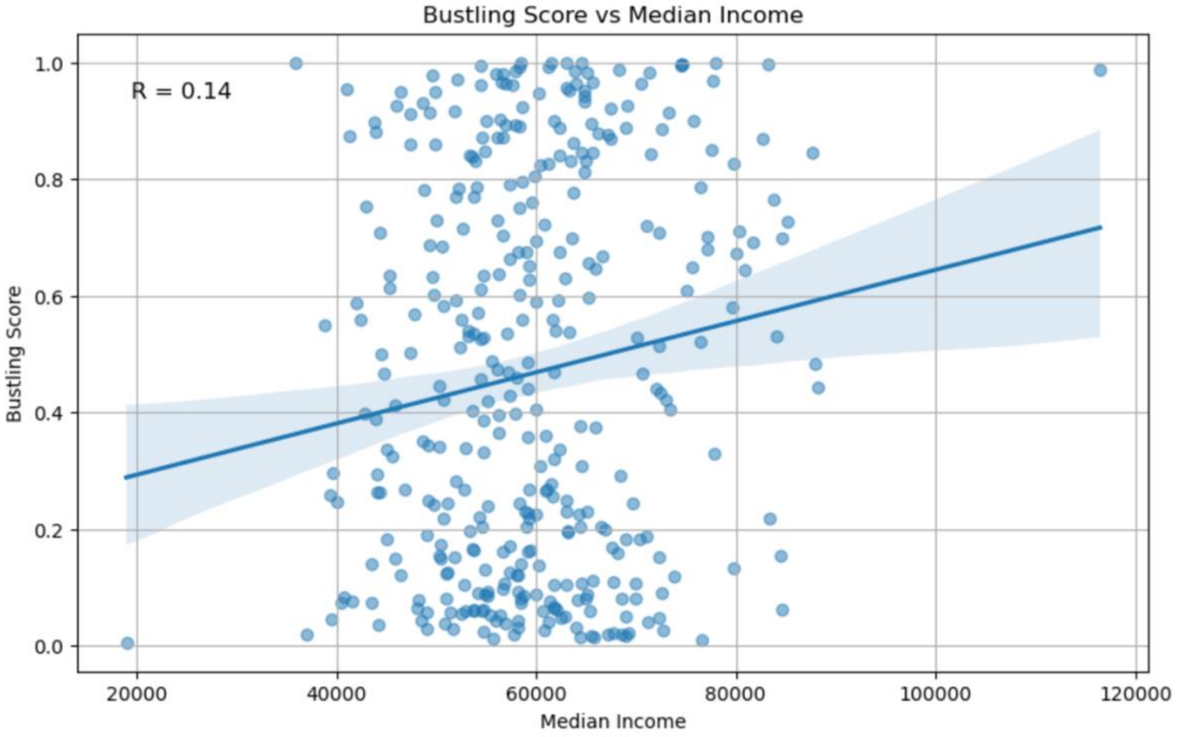


Figure 1.5 *- Scatter Plot graph of bustling score vs median income for different SA2 regions*

With our “bustling” metric revolving around the prominence of businesses, jobs, and economic prosperity, it was hypothesised that there would be a positive correlation between bustling score and median income. However, figure 1.4, shows a weak correlation present between bustling score and median income, shown by the R value of 0.14. Many of the data points are clustered around a median income of $60,000 AUD across all bustling scores from 0 to 1. There is a relatively even spread across a median income of 40,000 to 60,000 and a sparser spread from 60,000 to 80,000. This correlation indicates that there are other factors that influence how well-resourced and busy a neighbourhood is, hence highlighting a limitation of using the calculated bustling scores. Further research should be conducted to develop a better understanding of the relationship between economic factors and the bustling score.

There are two outliers present at both extremes, one at a bustling score of 0 and median income of 20,000 and another outlier at a bustling score of 1 and median income around 120,000. There are no other outliers, which suggests that there may be some correlation found if the data points were extended to include more data from both lower and higher median incomes. This would likely provide greater insights into the relationship between bustling score and median income across SA2 regions.

Appendix

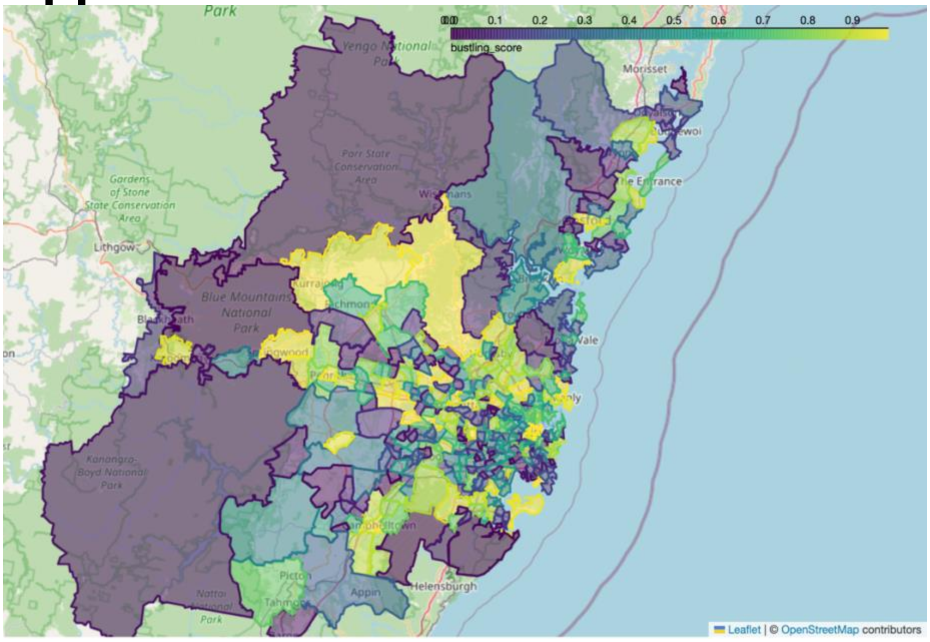


Figure 1.6 *- Interactive map of each SA2 region colour coded by bustling score. Accessed: data2001-assignment.zip*