Data 2001 - Greater Sydney Analysis Report

Arnav Chaddha - 530288682

Aaryan Bansal - 530390550

What is SA2?

- In this report, we mention the SA2. The Statistical Areas Level 2 (SA2) regions are medium-sized general-purpose areas built up from the whole of Statistical Areas Level 1. Their Population

→Description of datasets:

As part of our data analysis, we incorporated a variety of datasets sourced from different areas, which were both independently sourced[Opal Taps, ABS - Regional Population] and provided in the Canvas tab for the assignment itself[Businesses, Stops, Polls, Income, Population, Schools]. They are listed as follows:

- SA2 Regions:
  - This spatial dataset included different regions of Australia as per their Statistical Area level 2 boundaries, which we further filtered down to the Greater Sydney area. We further pre-processed the data by dropping all placeholder values with 'NaN'
- Businesses:
  - This CSV file dataset included and described the total number of businesses as per their SA2 region. It also describes their industry and holds the value of said businesses, which range from a 0 to 10 million AUD valuation. This was downloaded off of Canvas but originally sourced from The Australian Bureau of Statistics. The placeholder values in this dataset ('',' None ', ' null', etc) were filtered out and replaced by NaN.

- Opal Taps Dataset:

    - As part of task 3, we sourced a spatial dataset that held transactional data from different Opal transportation systems, detailing tap events (tap-on and tap-off events for different public transport systems like buses, trains, and light rails) which were separated in timed intervals of 15 mins. The dataset also held different geometry objects of type “POINT”, whose X and Y values hold A specific point as latitude and longitude in the world. This dataset was originally sourced by UNSW’s City Data database [1].

- Stops:
    - This text file spatial dataset held different bus stops around the GCC area. Furthermore, it held the locations of those bus stops as geometry objects and other information such as wheelchair accessibility. This dataset was originally sourced from Transport for NSW’s OpenData database.

- Polls:
    - The polls dataset was sourced from the Australian Electoral Commission which spoke about the different polling booths in 2019 around the GCC area. This dataset held the polling place name, and the division that it was in, along with a geometric ‘POINT’ object that held the exact latitude and longitude of the place.

- Income:
    - This dataset held the different SA2 regions along with the number of earners in that region, their median age along with their median and mean income.

- Population:
    - The population dataset held the SA2 region name, along with their total population, Furthermore, it held the total number of ‘youth population’ per region which is defined as anyone aged 0-19.

- Schools:
    - This dataset was originally sourced from the NSW Department of Education. It contains three shapefiles which include spatial data for primary, secondary, and future catchements. School catchments are designated geographical areas around schools from which they primarily accept students. The data includes information about the years of schooling offered and their geographical locations. For our calculations, we have only considered primary and secondary catchments, as we believe that future catchment plans do not contribute to the current ***bustling nature*** of a region.

- ABS-Regional Population:
    - We sourced this GeoJSON dataset from Research Data Australia[2].

      This dataset consists of the different SA2 regions and the different ways their population was affected by number such as the number of births, deaths, migrations, overseas arrivals, etc. Furthermore, this region also includes a geometric MultiPolygon Object which was used to correlate SA2 regions with the original SA2 shapefile dataset.

→Database Description:

This well-defined schema consists of several interconnected tables representing different aspects of the data. Each table is designed to store specific data, with relationships defined through foreign keys. A spatial index called `idx_sa2geom` was created on the geom and name column of the SA2 table for efficient query searching, since these were used the most in our Project.

PFA the image of the schema below and A PDF version is available in the zip file for clarity.

→Result analysis:

The Jupyter Notebook provides a comprehensive examination of the bustling metric of various Greater Sydney regions. The analysis begins with a clear rationale for the formulation of the bustling score. We start by calculating the z-score $(z = (x - \mu) / \sigma)$ value for attributes from 4 data sets.

- Selected industry businesses per 1000 people
- Number of public transport stops
- Federal election polling locations (as of 2019)
- School catchment areas per 1000 'young people'

We calculated these values for each of the regions and finally calculated a sigmoid value

Using $\sigma(x) = \frac{1}{1 + e^{-x}}$

Between 0(least bustling) and 1(most bustling) for each region. In this formula, we enter the sum of all z-scores as x.

This suggests that the more of each attribute there are in each of the locations, the more Bustling it will be.

After this, we implemented the extension by involving two more attributes in the calculation Of our sigmoid value. We imported two more datasets and selected the following attributes:-

Location:

- Number of Opal Tap ons and Tap offs
- Net Population change

We find the z-scores using the same formula as earlier and find a refined sigmoid value now From 6 attributes instead of 4.

to further quantify and note trends with our datasets, we created a scatterplot of ***'mean income versus z score of population change'***.

Interestingly, we noted that a majority of the population change occurs around the **75K AUD** range, with their z-scores being the highest. We eventually deduced that this income bracket efficiently represents the middle class which encompasses a majority of the population, hence the greater number of High Z-score values. The code for the same is available in the Jupyter Notebook.

Another important bias we noticed was that the data contradicts itself in certain areas. For example, the bustling metric favors a high number of catchments per area which contradicts the fact that the majority of schools are located near suburban areas which are far away from Central Sydney, a place which in itself is bustling as per the number of stops and businesses.

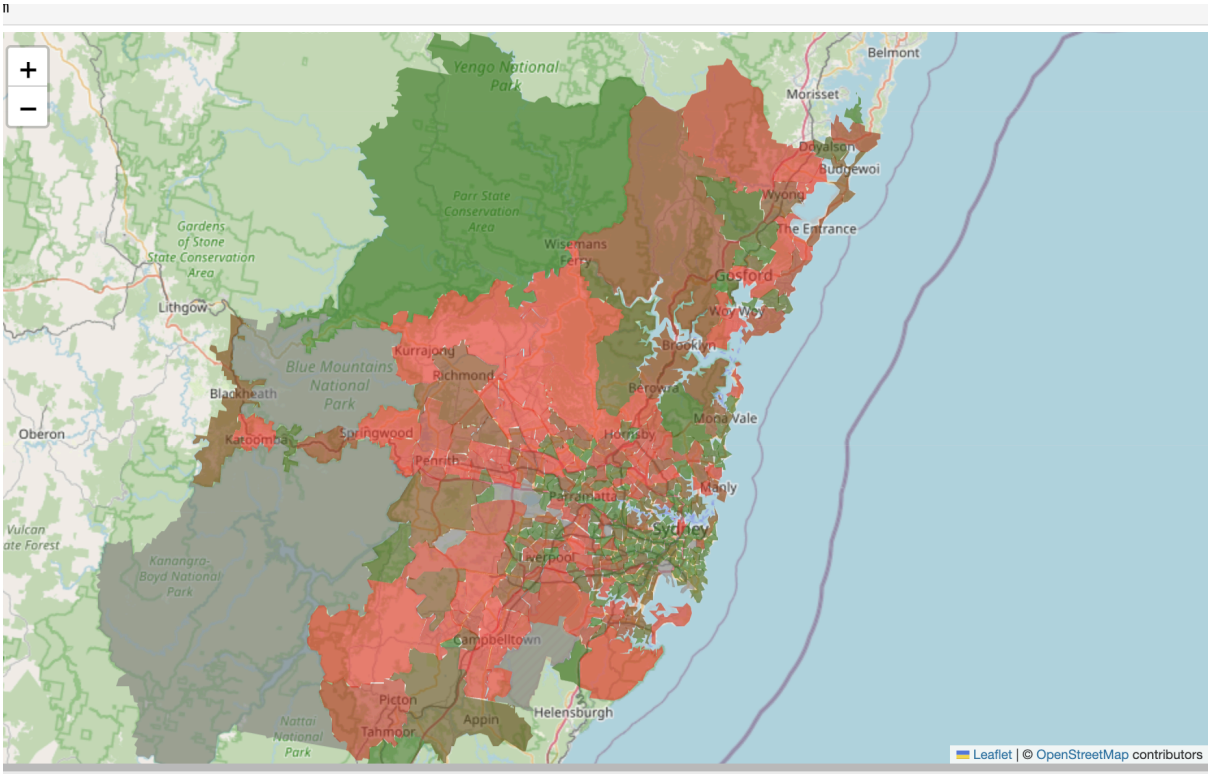
→ Correlation analysis:

We use Pearson's correlation coefficient to find the relation between a bustling score of a region and its median income. We calculated this coefficient in our Jupyter Notebook. It turned out to be ***-0.227.***

The correlation shows a slight negative correlation between the median income of a region and its bustling score. This is consistent with real-world data as the more expensive suburbs and locations tend to be more quiet and peaceful. And as these houses are more expensive, they tend to house people with higher median income.

→ Data Visualisation:

To efficiently visualize the outcome of our dataset, we created an interactive map that showed the bustling score for each Greater Sydney region on a scale of Red(Not bustling) to green (Very bustling). The grey values signify a null bustling score, which was due to insufficient data. We have attached an image here, moreover, The interactive version can be found in the .html file in the zip folder and also in the Jupyter Notebook.



Bibliography:

[1] City Futures Research Centre, "Opal Data - trips by location with SA4," City Data, [Online]. Available: https://citydata.be.unsw.edu.au/dataset/opaldatansw_byloc_withsa4. [Accessed: May 11, 2024].

[2] Australian Bureau of Statistics, "ABS - Regional Population (SA2) 2001-2021," Australian Urban Research Infrastructure Network (AURIN), [Online]. Available: https://researchdata.edu.au/abs-regional-population-2001-2021/2747979. [Accessed: May 11, 2024].