DATA2001 Group Assignment-Report

Shan-Shan, Liu 520065802

Yang, Nie 510239417

1. Dataset Description

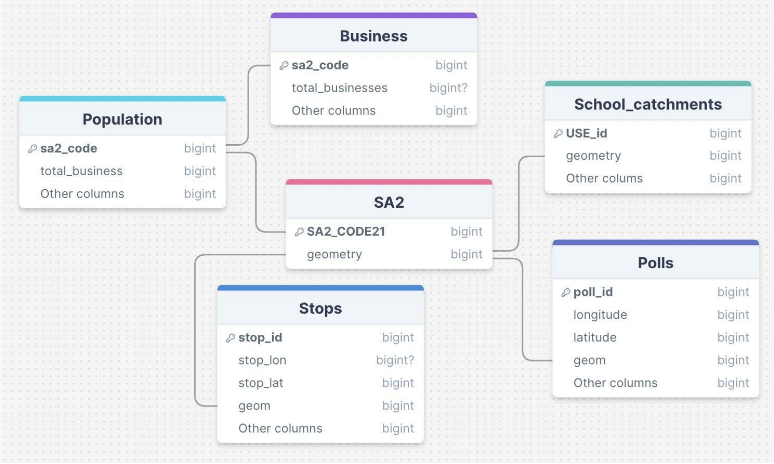
The Datasets are provided by the DATA2001 teaching team and were downloaded directly on Canvas. The "Statistical Area Level 2 (SA2) digital boundaries" file are sourced from the Australian Bureau of Statistics and local government databases, providing geographical boundaries and demographic information for Sydney's neighbourhoods, for further spatial analysis. We use the "Business" dataset to assess economic activity which is distributed across each sa2 region, and we focused on the 'Retail Trade' industry for this project. "Public Transport Stops", "Polling Location" and "School Catchments" datasets play the same role as Business dataset, to construct a comprehensive "well-resourced" score for each neighbourhood, providing information include the accessibility and infrastructure of the area, and also the civic engagement and educational resources.

Before we started to create tables and import datasets, we went through a basic data cleaning process, including dropping empty or duplicated rows. That is to ensure the data accuracy, and do not include irrelevant information.

2. Database Description

For schema establishment, we set up schema in PostgreSQL database using a combination of manual SQL commands and Python scripts leveraging libraries such as SQLAlchemy and GeoPandas. Firstly, we created tables with appropriate data types for each file, and then defined primary keys to ensure data integrity. Lastly, we established relationships through foreign keys where relevant.

The central dataset," sa2_boubaries" contains spatial definitions for SA2 regions, identified by a primary key, typically "sa2_code21" This table serves as the geographical anchor for other datasets. The "stops" and "polls" datasets are integrated through spatial queries, using their respective 'geom' columns to determine which stops and polls are located within the boundaries of the SA2 regions. This spatial relationship does not rely on direct database foreign keys but on geographic proximity and overlap determined by functions such as "ST_within()". The "business" dataset, while not explicitly confirmed to have direct foreign key references to "sa2_ boundaries" sa2_boubaries", is presumed to be related through the "sa2_code21" field, allowing businesses to be associated with specific SA2 regions for detailed economic analysis. Lastly, the "catchment_primary" and "catchment_secondary" datasets, which include spatial data for school catchment areas, are similarly integrated through spatial joins to the SA2 regions.



Business	
sa2_code	bigint
total_businesses	bigint?
Other columns	bigint

Population	
sa2_code	bigint
total_business	bigint
Other columns	bigint

School_catchments	
USE_id	bigint
geometry	bigint
Other colums	bigint

SA2	
SA2_CODE21	bigint
geometry	bigint

Polls	
poll_id	bigint
longitude	bigint
latitude	bigint
geom	bigint
Other columns	bigint

Stops	
stop_id	bigint
stop_lon	bigint?
stop_lat	bigint
geom	bigint
Other columns	bigint

3. Score Analysis

Computing the "well-resourced" score for each neighborhood in the Greater Sydney SA2 regions combines several aspects including business density, public transport accessibility, polling station availability, and school coverage, each tailored to reflect essential services and infrastructures.

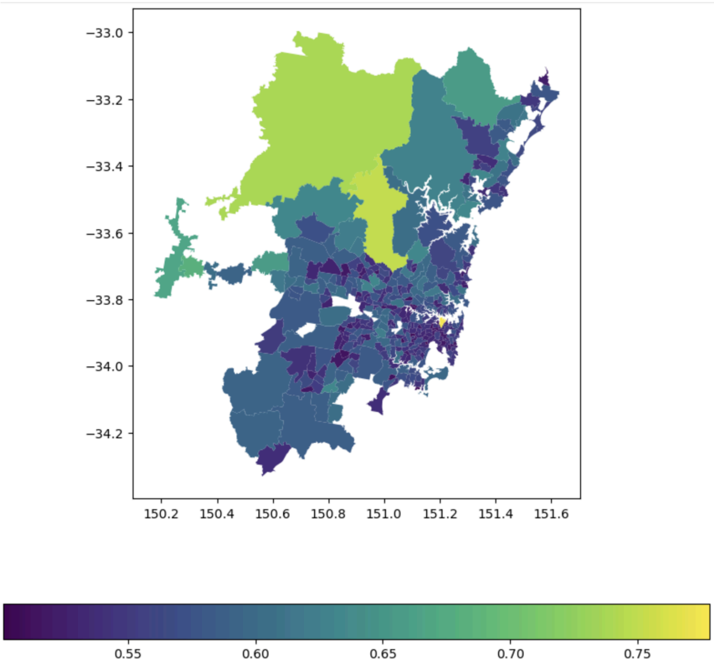
The business density score is calculated by dividing the total number of businesses by the population per 1,000, providing a measure of commercial activity relative to the size of the population, ensuring comparability across different-sized regions. This approach highlights regions potentially offering more employment opportunities and services.

Public transport accessibility is gauged through the number of transport stops within each region's boundaries, reflecting the ease of mobility for residents and emphasizing the region's connectivity.

Polling station availability is assessed similarly by counting the polling stations within each region, underlining the civic infrastructure's readiness and public service accessibility.

School coverage is determined by the ratio of the total area of school catchments to the number of school-age children per 1,000 people, showing the spatial adequacy of educational facilities relative to the demographic needs.

These scores undergo min-max normalization to ensure each score contributes equally to the final aggregate, preventing any single aspect from skewing results due to different value ranges. The final "well-resourced" score is derived by summing these normalized scores and applying a sigmoid transformation, which constrains the final score within the range of 0 to1, making it straightforward to interpret and compare, ensuring that the evaluation of resourcefulness is balanced and reflective of a comprehensive view of neighborhood amenities.



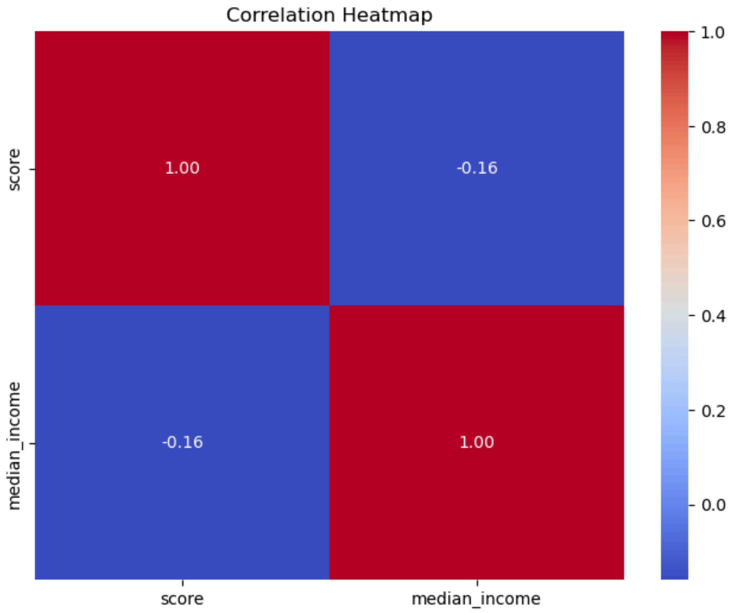
(graph1)

According to Graph1, the overall score shows the yellow and green region have the highest scores where it is located at the northern part of the map. There might be more businesses, schools , amenities, and higher population density that are contributing to the score. Most areas of medium blue (around 0.65) located in the central and eastern parts of the map have moderate scores than the northern part.

4. Correlation Analysis

The r is -0.16 which is very low, and it indicates that the variation cannot be well-explained by median income. Further statistical analysis or additional data might be required to draw more robust conclusions. These results are quite surprising that the low correction we finally got since the median income may influence the score somehow. The number of metrics is limited. We need more potential metrics to develop how bustling the districts within the city are. Moreover, there are some missing values in these datasets that may affect our final scores.

Graph2 is the correction heat map illustrating the relationship between the score and the median income. The correlation coefficient is -0.16 between these two variables, suggesting a slight negative correlation. This means that as the median income increases, the score tends to decrease slightly, although the correlation is weak. It is not strong enough to imply a definitive or substantial relationship between the score and the median income.

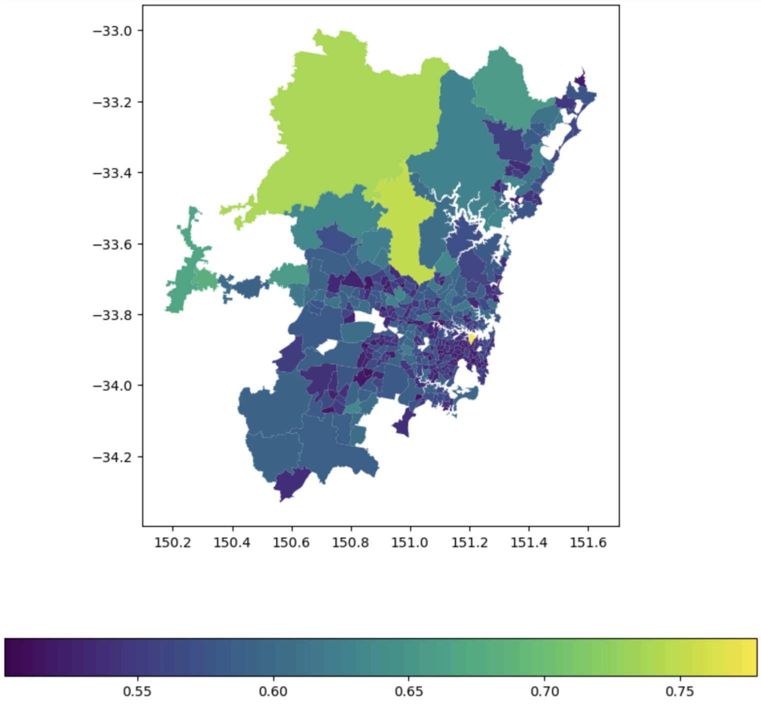


(graph2)

5. Other Supporting Visualization

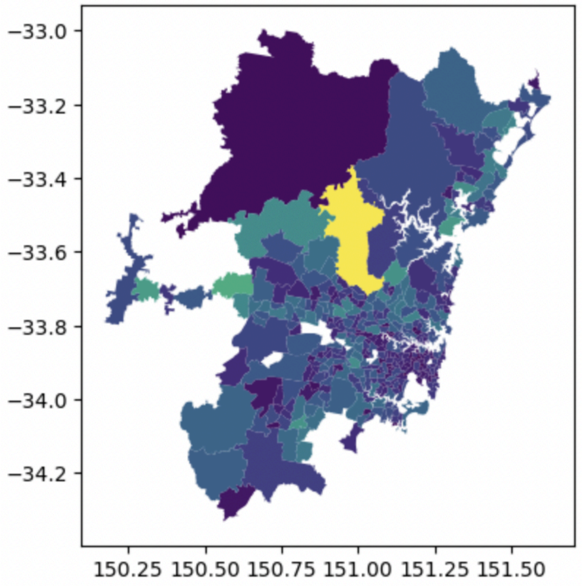
a. Score by SA2 region

Higher bustling scores, represented by shades of green to yellow, are observed in the central and northern parts of Sydney, suggesting these regions are more vibrant with higher population density, commercial activity, and social interactions. Conversely, lower bustling scores, shown in shades of blue to dark purple, are found in the southwestern and southern outskirts, indicating these areas are quieter and likely more residential with less commercial activity.



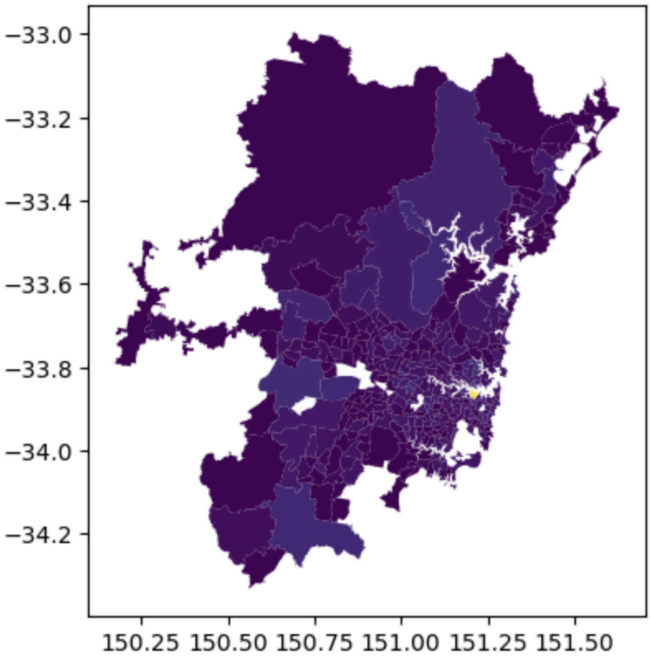
b. Stops Z-score

Areas with high transportation density, shown in yellow, are primarily located in the northern parts of Sydney, suggesting a concentration of transport hubs such as bus, train, and ferry stations in these regions. The central areas also exhibit moderate transportation density, represented by shades of green to blue, reflecting significant but less concentrated transport infrastructure. Conversely, the western and southern outskirts, depicted in dark purple, indicate lower transportation density.



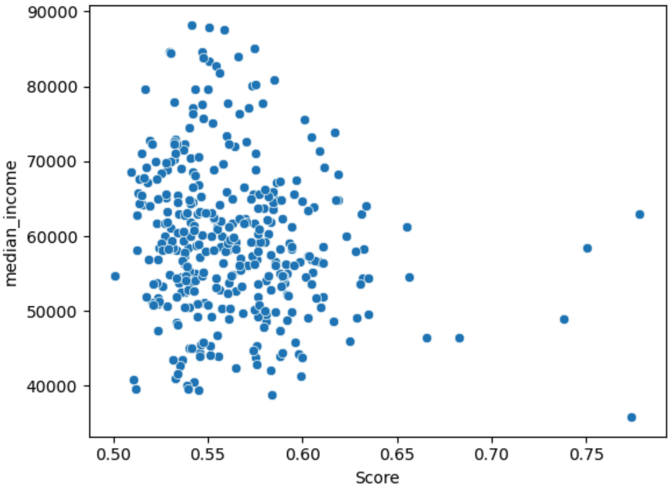
c. Business Z-score

The highest business densities, shown in yellow, are predominantly located in central Sydney, suggesting these areas are major commercial hubs with a significant concentration of businesses and economic activity. The surrounding areas with shades of light purple indicate moderate business density, reflecting regions with considerable but less concentrated business presence. The outskirts of Sydney, depicted in dark purple, indicate low business density, pointing to more residential or less commercially developed regions. This distribution highlights the centralization of business activities in Sydney's core, which is critical for urban development, economic planning, and infrastructural investment to support and potentially decentralize business growth.



d. Correlation between Score and median income

suggesting a complex and potentially weak correlation between these variables. The data points are spread out, indicating considerable variability in median income at similar levels of bustling scores. There is no clear trend or pattern suggesting a strong linear relationship; instead, the spread of incomes across the range of scores highlights the diversity in economic conditions associated with different levels of bustling activity. The cluster of points seems denser at lower scores and median incomes, possibly indicating that regions with lower bustling scores tend to have a narrower range of, generally lower, incomes. Conversely, at higher bustling scores, the variability in income increases, suggesting that these areas might encompass a wider economic spectrum.


median_income
Score