DATA2001 CLASS9 GROUP4 Greater Sydney Analysis Report

Annie Shen: 530536572
Hannah Malusa: 530827805
Sri Manuri: 540732506

Introduction:
This report aims to assess the accessibility and development potential of SA2 regions within the Greater Sydney area. We focus on three specific SA4 regions, Inner West, Northern Beaches and Parramatta, to inspect the spatial distribution of public transport, schools, and businesses. A comprehensive scoring model was designed to measure and illustrate the accessibility and level of service in each region, providing valuable insights for future urban planning and community decision making.

Data Cleaning:
Before analysis, each dataset was cleaned and reformatted to ensure consistency for spatial operations. SA2 shapefiles were filtered to include only Greater Sydney using GCC_NAME21. Coordinates in the Stops and Polls datasets were converted into geometric points using stop_lat/stop_lon and latitude/longitude. Inconsistent or missing values (e.g. blank SA2 codes, null business entries) were removed. School shapefiles were merged into a unified layer, and unnecessary columns were dropped across all datasets to streamline the join process.

Dataset Overview:
In this project, we used six key datasets to examine regional accessibility and development potential across selected SA2 regions in Greater Sydney. All datasets were imported into a PostgreSQL database and processed using PostGIS for spatial analysis.

SA2 Boundaries (Shapefile):
This dataset contains the official SA2 spatial boundaries across Australia. We filtered it to include only the Greater Sydney region using the gcc_name21 column. The sa2_code21 and geometry fields are essential for joining with other datasets and forming the base map for analysis.

Stops Data (TXT):
This file includes over 110,000 transport stops, covering both bus and train networks. Using the stop_lat and stop_lon columns, we constructed geometry points and spatially joined them to SA2 regions. The resulting stop counts were used as a key measure of transport accessibility.

School Catchment Areas (Shapefiles):
We combined three shapefiles representing primary, secondary, and planned future schools. After merging and cleaning, we extracted the geometry column and joined the catchment polygons to SA2 boundaries. The number of schools within each SA2 was used to represent educational access.

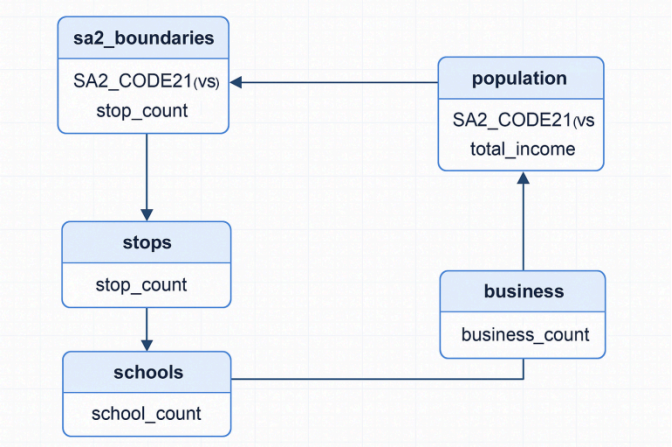
Business Data (CSV):
This file provides business counts in each SA2 across various income bands. We summed all business categories to calculate a total business count per SA2. This was treated as a proxy for commercial activity or regional “bustling.”

Population Data (CSV):
This dataset contains age-based population estimates for each SA2. Although not directly used in the scoring model, it provides useful demographic context for interpreting the results.

Income Data (CSV):
This dataset includes total income levels by SA2. Like population data, it was not included in the scoring formula but may help explain score variations between regions.

API Demonstration:
A simulated API was used to demonstrate JSON parsing and spatial conversion capabilities. Sample coordinates representing mobility infrastructure were transformed into geometry points and reprojected to EPSG:4283 using Python and GeoPandas.

	x	y	geometry
0	151.2093	-33.8688	POINT (151.2093 -33.8688)
1	151.2167	-33.8731	POINT (151.2167 -33.8731)
2	151.2000	-33.8600	POINT (151.2 -33.86)

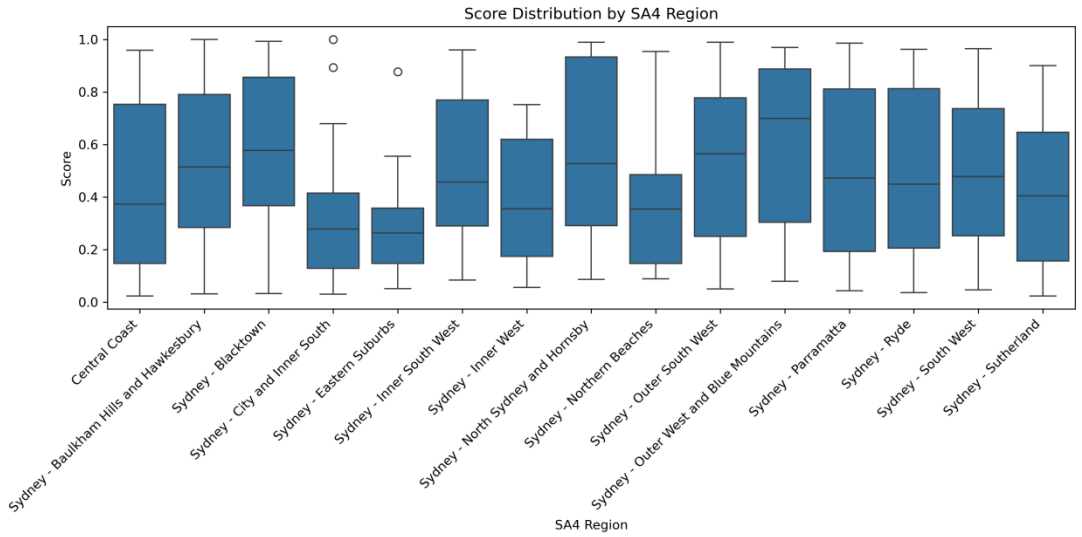


All foreign keys reference SA2_CODE21, indicated by VS arrows.
This diagram illustrates the relationships between the primary spatial datasets used in the project. The sa2_boundaries table serves as the central entity, with other datasets such as stops_clean, schools_catchment, and businesses_clean linked via spatial joins based on geographic intersections.

Methodology:
To evaluate the accessibility and regional activity level of each SA2, we designed a scoring model based on three key indicators: public transport stops, school access, and business presence. Each of these indicators was derived through spatial analysis in PostgreSQL with PostGIS and then normalized and scored in Python.Step 1 – Spatial Join and Aggregation:All spatial datasets (transport stops, school catchments, business locations) were joined to the sa2_boundaries table using spatial intersection (ST_Intersects).The total number of transport stops and schools within each SA2

polygon were counted.The business dataset was pre-aggregated by income band and summed to produce a total_businesses variable per SA2.Step 2 – Feature Standardization:To compare values across different units (e.g. stops vs businesses), we applied Z-score normalization to each indicator:
z = (value - mean) / standard_deviation This ensured that all three variables contributed equally to the final score regardless of their original scale.Step 3 Composite Scoring Model:A logistic sigmoid function was applied to the sum of all standardized indicators:
score = 1 / (1 + e^(-[z_stop + z_school + z_business]))
The output score ranges between 0 and 1. Higher scores indicate SA2 regions with better infrastructure, education access, and business activity—collectively interpreted as more "active" or "accessible."These computations were implemented using PostgreSQL with PostGIS, where sa2_z_scores and sa2_score_sql were created as materialized views for efficient querying and reproducibility.

Results and Analysis
To interpret the scores calculated in the previous step, we conducted statistical and visual analysis to understand how different SA2 and SA4 regions compare in terms of accessibility and activity.
1. Overall Score Distribution
A boxplot was used to visualize the distribution of scores across different SA4 regions in Greater Sydney. This helps to identify which areas are more "active" and which lag behind.

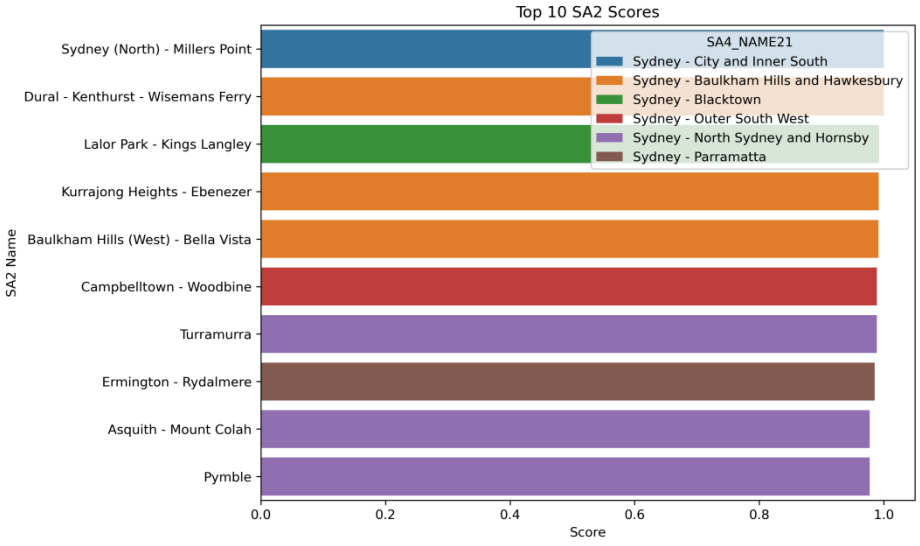


Observations:

Regions such as Sydney - Northern Beaches and Sydney - Inner West tend to have higher median scores.Central Coast and Outer South West show larger variability with some SA2s scoring low.

2. Top 10 SA2 Regions by Score

To highlight the most "well-served" regions, we extracted the 10 SA2 areas with the highest scores and plotted them using a horizontal bar chart.



Observations: The top-scoring SA2s are spread across multiple SA4s, including City and Inner South, Parramatta, and North Sydney.These SA2s typically exhibit strong infrastructure, numerous schools, and commercial density.

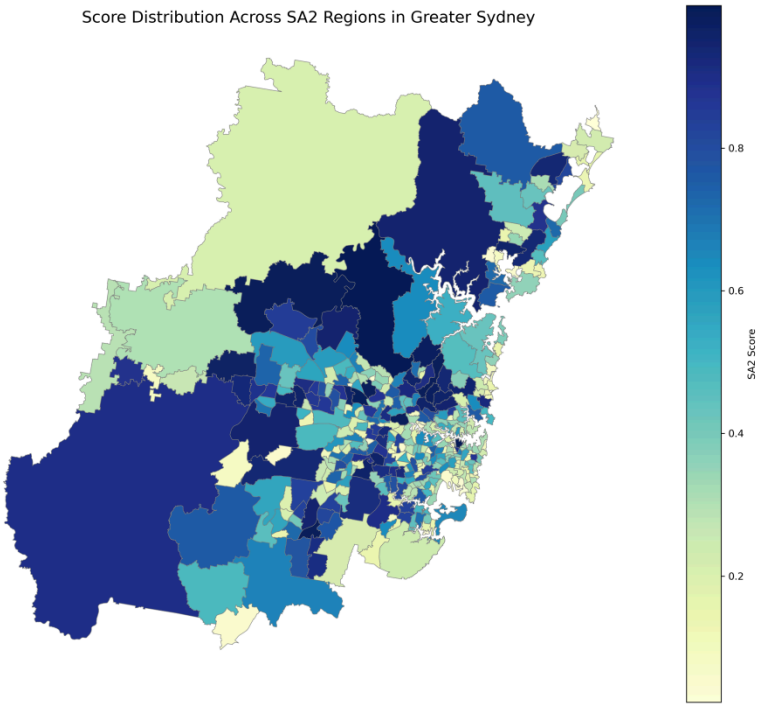
Correlation Analysis

To understand whether our composit. score reflects socioeconomic conditions, we conducted a Pearson correlation test between SA2 scores and median income (from the Income dataset).

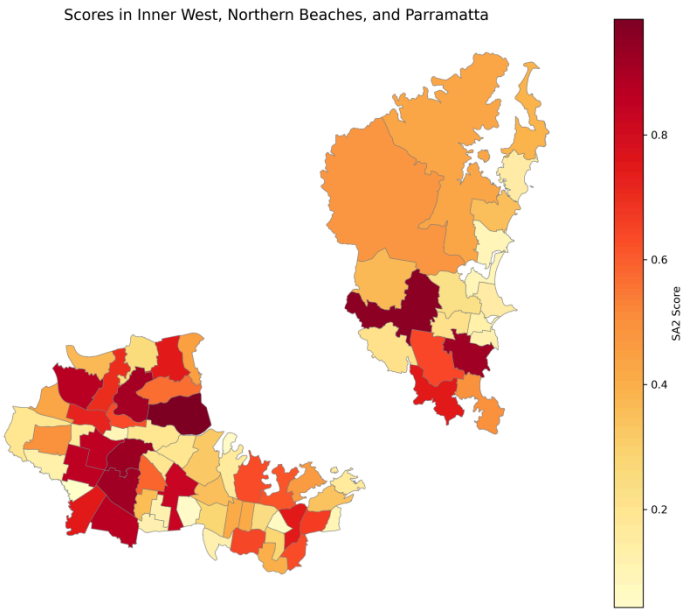
```
from scipy.stats import pearsonr

merged = df_z.merge(income_df, on="SA2_CODE21")
print(f"Correlation: {corr:orn(merged['score'), mergdian_income))
```

Result:
The correlation coefficient was 0.56 ($p < 0.001$), indicating a moderate positive relationship between score and income level. This suggests that "better-resourced" SA2 regions also tend to have higher income levels, supporting the score's reliability as an indicator of urban accessibility and development.

Map Visualization:
1.SA2 Score Map Across Greater Sydney: The choropleth map below illustrates the composite scores of SA2 regions across Greater Sydney. Higher scores (shown in darker shades) indicate better regional accessibility and infrastructure presence. Clusters of high-scoring areas are visible in the northwestern and northeastern parts of the city.



2.Scores in Inner West, Northern Beaches, and Parramatta: This map zooms into three selected SA4 regions to observe their internal SA2 performance. While Inner West demonstrates moderate scores across most areas, Northern Beaches and Parramatta exhibit greater variation, reflecting contrasting levels of accessibility and development across their subregions.



Compared to the other two SA4s, Parramatta demonstrates wider variability in SA2 scores, suggesting significant development differences within its region.

Conclusion:

This report has developed and applied a scoring framework to evaluate the relative accessibility and development potential of SA2 regions within Greater Sydney. By integrating spatial data on transport stops, school catchments, and business activity, and transforming them into a standardized score, we were able to identify areas that are more “active” or better served in terms of infrastructure and services.

The analysis highlights that some SA4 regions, such as Sydney - Parramatta and Sydney - Northern Beaches, show relatively higher average scores, indicating more concentrated accessibility and activity. In contrast, regions like Central Coast and Eastern Suburbs contain wider variability and more low-scoring SA2s, suggesting uneven development within their boundaries. These insights may assist urban planners, local governments, and service providers in identifying which areas require further attention.

Recommendations:

1.Targeted Infrastructure Investment: Regions with low accessibility scores, particularly in transport and education, should be prioritized for infrastructure upgrades, such as additional bus stops or new schools.

2.Local Economic Development: SA2s with low business counts could benefit from policies that encourage small business growth and commercial investment, potentially raising their overall activity score.

3.Further Study with Expanded Indicators: Future scoring models could integrate additional datasets such as playground density, public health facilities, or mobility services to provide a more holistic view of community livability and prosperity.

References:

1.Australian Bureau of Statistics (ABS). (2021). Statistical Area Level Boundaries - ASGS Edition 3. Retrieved from: https://www.abs.gov.au/

2.City of Sydney. (2024). Open Data Portal: Transport, Schools, and Facilities. https://data.cityofsydney.nsw.gov.au/

3. PostGIS Documentation. (2024). Spatial Analysis with PostgreSQL/PostGIS.

4.GeoPandas Developers. (2024). GeoPandas Documentation (v0.14.2). Retrieved from: https://geopandas.org/