DATA2001 REPORT

Hardik Chojar (530377104), Dylan Uno Syahfaril (530675262), Wilson Peter Yulianto (530762481)

1. Dataset Description

1.1. Data Sources

The dataset used for this assignment are obtained through various resources, including publicly accessible websites, API and files provided on the assignment specifications, these datasets came from multiple raw formats including csv, txt, and shp. All data was accessed between april and may of this year. The data provided by the Australian Bureau of Statistics are released under creative common attribution license which are open government license, meanwhile the data provided by NSW government are under open data license.

Dataset	Source	Access Method	Format	Licence Info	Usage
SA2/SA4 Boundary Files	Australian Bureau of Statistics	Downloaded from ABS website	Shapefile	Open Government Licence / CC BY 4.0	Spatial joins (POI, stops), SA2 code/name reference
NSW Points of Interest (POI) API	NSW Government SIX Maps POI API	API (no key required)	JSON	NSW Open Data Licence	Counting POIs in SA2 for zPOI calculation
Businesses.csv	Provided in Canvas	Provided (CSV)	CSV	Open Government Licence	Business per 1000 people -> z business calculation
Public Transport Stops (GTFS stops.txt)	Provided in Canvas	Provided in dataset	CSV (GTFS format)	NSW Open Data Licence	Counting stops in SA2 -> z stops calculation
School Catchments (SchoolCatchments.zip)	Provided in Canvas	Provided (shapefile)	Shapefile	NSW Open Data Licence	Calculating catchment per young people -> z schools calculation
Population Estimates by SA2 (Population.csv)	Provided in Canvas	Provided (CSV)	CSV	Open Government Licence	Business per 1000 people, filter SA2s with population >= 100
Census Median Income (Income.csv)	Provided in Canvas	Provided (CSV)	CSV	Open Government Licence	Extended analysis / reporting (correlation, descriptive stats)

1.2. Acquisition and Preprocessing

Each of the dataset above are processed from various raw format where SA2 were loaded into PostGIS environment to enable spatial operations, from the SA2 geometry bounding boxes were calculated using spatial functions like ST_extent to prepare API based extraction, the NSW POI API were extracted using HTTP request and bounding box parameter, the JSON data was parsed to extract relevant attributes. CSV files are read with pandas library with standard cleaning operations such as renaming columns, filtering and type casting. All of this cleaned and structured data was then inserted to SQL tables with defined data type.

Our data processing pipeline implemented several filtering, imputation, and coordinate reprojection steps to ensure data relevance, quality, and spatial consistency. Geographically, analysis was focused by selecting SA2 regions exclusively from our chosen SA4s (Sydney - Blacktown, Sydney - Parramatta, Sydney - Ryde) within the "Greater Sydney" GCC. All spatial data, including SA2 boundaries (converted from GDA2020 to EPSG:4326), school catchments (standardized to EPSG:4326), and transport stops (raw latitude/longitude converted to EPSG:4326 point geometries), were standardized to the WGS84 (EPSG:4326) coordinate reference system before database ingestion. Points of Interest (POI) for these SA2s were retrieved via API using their EPSG:4326 bounding boxes. POI selection for scoring focused on community-facing amenities such as libraries, parks, healthcare facilities (General Hospital, Community Medical Centre), educational institutions (Primary/High School, Child Care Centre), transport hubs (Transport Interchange), and cultural venues (Art Gallery, Museum), while excluding industrial, purely commercial, or highly niche sites. For data integrity, duplicate entries were systematically removed. SA2s with a total population below 100 were

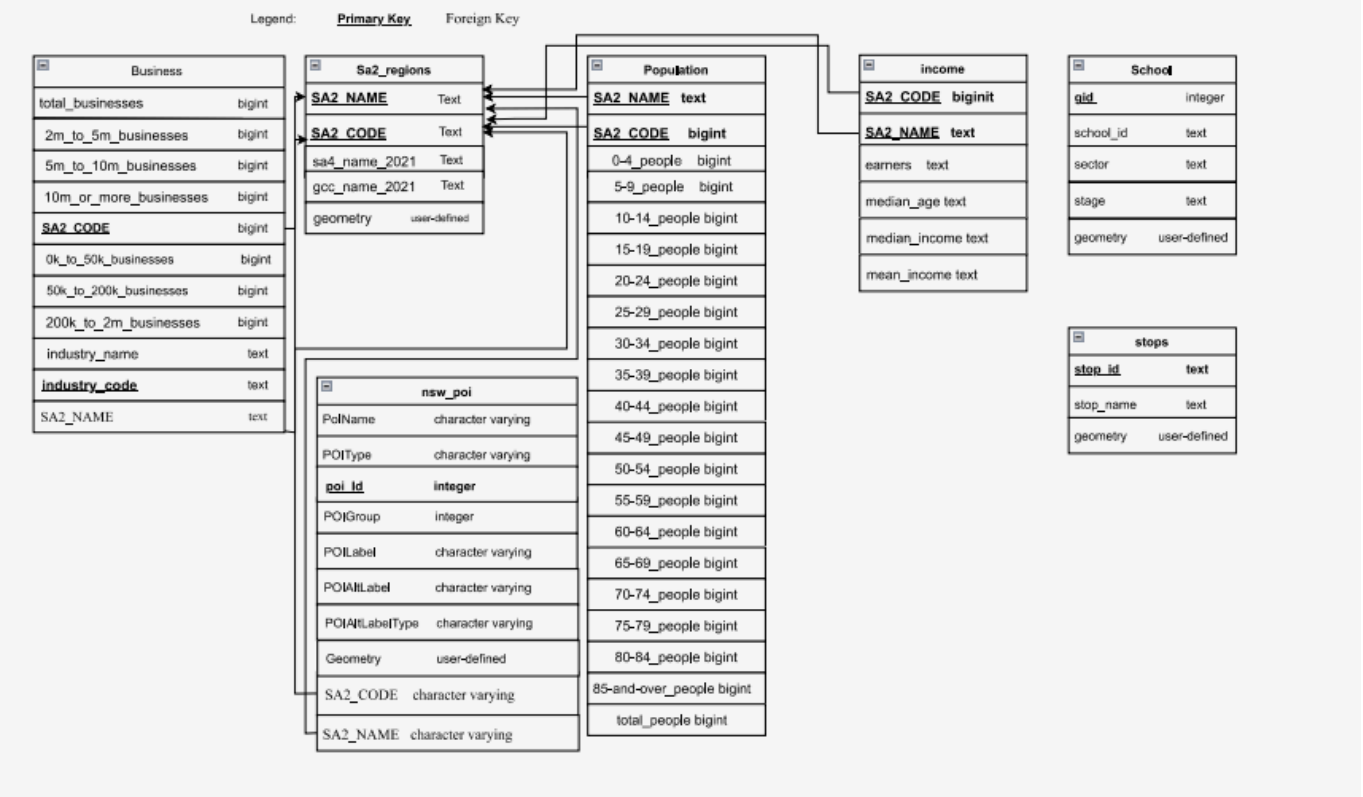
filtered out from relevant calculations (e.g., business and school metrics per capita) to maintain statistical stability. Missing raw counts for metrics (businesses, stops, schools, POIs) were explicitly imputed as zero using COALESCE directly within our Z-score calculation logic. Finally, for the correlation analysis with median income, rows with missing income data were dropped.

2. Database Description

2.1. Description

To store and integrate the datasets required for scoring SA2 regions, we designed a normalized relational schema implemented using PostgreSQL with the PostGIS extension for spatial operations. The schema diagram is included below, showing all core tables with their primary keys and foreign key relationships.

2.2. Schema Diagram



2.3. Schema Overview

Our schema the following tables:

- **sa2_regions**: Contains SA2 geometries and names. The sa2_code and sa2_name field serves as the primary keys and is referenced by other tables. sa2_name provides a human-readable identifier.
- **businesses:** Stores industry-wise business counts by SA2. Joined to sa2_regions using sa2_code.
- **stops**: Holds all GTFS-format public transport stops. Includes geometry data for spatial joins with SA2 boundaries.
- **school**: Represents school catchment zones. Spatially intersected with SA2 boundaries to compute coverage per 1000 youth.
- **population**: Age-segmented population data per SA2, used for per-capita calculations. Includes total population and youth counts.
- **income**: Holds median income values per SA2 region for later correlation analysis.
- **points_of_interest (POI)**: Populated via a custom API query loop across SA2 bounding boxes. Contains a poi_id primary key and its SA2_CODE foreign key references sa2_regions. Includes POI type/group for filtering and a geometry column for spatial joins.

2.4. Data Integration and Spatial Joins

All geometry fields across datasets were standardized to SRID 4326 (WGS84) to ensure spatial consistency. We performed spatial joins using PostGIS functions to integrate datasets at the SA2 level. Key spatial operations included:

- Mapping GTFS stop locations (points) to their containing SA2 regions using ST_Contains().
- Overlaying school catchment polygons with SA2 boundaries using ST_Intersects() (and NOT ST_Touches() to ensure genuine overlap) to estimate coverage.
- Assigning each NSW Point of Interest (POI) (points) to its containing SA2 region using ST_Within(). These joins were fundamental for aggregating data and calculating the components of our well-resourced score at the SA2 level.

2.5. Indexes

To optimize query performance, particularly for spatial operations and joins, the following indexes were created:

- Spatial indexes (GIST): Implemented on all geometry columns (geom) in the public.sa2_regions, public.stops, public.schools (referring to school catchments), and public.nsw_poi tables.
- B-tree indexes: Created on foreign key fields such as sa2_code in the public.businesses, public.income, and public.population tables, and on other frequently queried attribute columns like industry_code in public.businesses. These indexes significantly accelerated data retrieval for spatial joins and the score calculation queries.

3. Score Analysis

3.1. Scoring Formula and Rationale

Our "well-resourced" score is calculated by summing standardized z-scores from four components: business density, public transport stops, school availability, and points of interest (POIs). Then, we apply a sigmoid function to this sum. Z-score normalization was selected to standardize each component, removing bias due to differing scales. For example, large numbers of bus stops compared to relatively fewer schools. The summation of z-scores assumes equal importance of all four components, reflecting our belief that each contributes significantly to community resources. Applying a sigmoid function ensures our final scores are intuitively interpretable, bounded within a range of 0 (poorly resourced) to 1 (very well-resourced), and resistant to distortion by extreme values. This method yields a balanced, robust measure of resource availability, directly comparable across all SA2 regions.

3.2. Implication of Extensions

Integrating genuinely distinct datasets, such as detailed healthcare service levels (e.g., GP wait times, specialist availability by SA2, which go beyond our current POI counts per capita) or specific environmental quality indicators (e.g., air pollution levels, accessible green space quality ratings), would offer a more nuanced "well-resourced" assessment. While our existing Z-score and sigmoid framework can accommodate new standardized metrics, the primary implication is managing increased model complexity. This necessitates careful evaluation to ensure each novel dataset adds unique value rather than redundancy, and requires thoughtful recalibration of metric weightings to maintain the score's coherence and ensure it evolves into a more comprehensive, not just more convoluted, measure of regional resourcefulness.

3.3. Distribution Summary

The overall "well-resourced" scores across all 67 analyzed SA2 regions exhibit a mean of 0.475 and a median of 0.445, suggesting a slightly right-skewed distribution (as seen in Figure 2, Appendix). The standard deviation of 0.300 and an Interquartile Range (IQR) of 0.478 indicate considerable variability in resource levels across these areas. When examining the selected SA4 regions (Table 1):

- Sydney - Blacktown shows the lowest average resource level (Mean: 0.339, Median: 0.234), coupled with a relatively high standard deviation (0.314) and IQR (0.445), pointing to significant internal disparities.
- Sydney - Parramatta presents a more moderate and consistent resource profile (Mean: 0.529, Median: 0.507) with the lowest spread (Std Dev: 0.232, IQR: 0.322), indicating more evenly distributed resources.
- Sydney - Ryde has the highest average resource level (Mean: 0.588, Median: 0.524) but also the largest spread (Std Dev: 0.349, IQR: 0.631), suggesting it contains both some of the most and least resourced SA2s among the three SA4s.

	SA4 Region	Mean Score	Median Score	Standard Deviation	IQR
0	Overall	0.475424	0.444820	0.299508	0.478354
1	Sydney - Blacktown	0.339371	0.234107	0.314169	0.445494
2	Sydney - Parramatta	0.529292	0.507460	0.231663	0.322356
3	Sydney - Ryde	0.587680	0.523660	0.348750	0.631074

Table 1: Summary Statistics of Well-Resourced Scores by SA4 Region

3.4. Cross-Zone Comparison

A comparative boxplot of score distributions across the selected SA4 zones (Figure 3, Appendix) visually highlights these inter-SA4 differences and allows for direct comparison of their median, interquartile range (IQR), and overall range. For instance, Sydney - Parramatta exhibits a relatively compact distribution around a higher median, while Sydney - Blacktown shows a lower median and wider spread, visually confirming the statistical variations noted in Table 1. Sydney - Ryde, while having the highest median, also displays the widest overall range and several outliers, underscoring significant inequality in resource distribution within this SA4.

3.5. Key Pattern and Outlier

Analysis of the well-resourced score distribution highlights several keys. Blacktown has the lowest well-resourced score with low resource and high inequality among areas, meaning most areas in Blacktown are generally under-resourced. Parramatta represents a more balanced distribution with a high mean and median, indicating generally well-resourced areas. Ryde averages as the best-resourced area, but inequality is high, with some areas extremely well-resourced and the majority not. Interestingly, the majority of Ryde areas are less well resourced than its median suggests. Figure 4 (Appendix) pinpoints the specific SA2s with the highest and lowest 'well-resourced' scores, illustrating these extremes of resource availability and revealing that highly resourced and under-resourced areas can exist even within the same broader SA4 zone.

4. Correlation Analysis

4.1. Correlation Testing

To investigate the relationship between the computed “well-resourced” scores and the median household income for each SA2 region within our selected SA4 areas, we conducted a Pearson correlation analysis. Pearson correlation was chosen as it is suitable for measuring the linear association between two continuous numeric variables, which both our well-resourced score and median income are. The Pearson correlation coefficient (r) measures the strength and direction of this linear relationship, with significance determined by the corresponding p-value. The results are:

1. Overall Analysis (n = 67): Pearson correlation coefficient (r) = -0.196, p-value = 0.112.
2. Sydney - Blacktown (n = 23): Pearson correlation coefficient (r) = -0.623, p-value = 0.002.
3. Sydney - Parramatta (n = 31): Pearson correlation coefficient (r) = 0.192, p-value = 0.301.
4. Sydney - Ryde (n = 13): Pearson correlation coefficient (r) = -0.083, p-value = 0.787. Additionally, the regression analysis (Figure 5, Appendix) visually confirms these findings.

The scatter plot with an overlaid regression line (Figure 5, Appendix) visually illustrates the relationship between these two variables, showing each SA2 as a point and the line of best fit to help discern the direction and strength of any linear trend.

4.2. Interpretation

The overall analysis across 67 SA2 regions revealed no statistically significant linear relationship between "well-resourced" scores and median incomes ($r = -0.196$, $p = 0.112$). However, for Sydney - Blacktown specifically, the correlation was negative and statistically significant ($r = -0.623$, $p = 0.002$). This suggests a moderate inverse relationship, highlighting that areas with higher resource availability scores in Blacktown tend to have lower median incomes, reflecting unique socioeconomic characteristics in this region. Conversely, results for Sydney - Parramatta ($r = 0.192$, $p = 0.301$) and Sydney - Ryde ($r = -0.083$, $p = 0.787$) indicated no significant linear correlations. The regression plot (Figure 5, Appendix) further supports these findings, depicting a slight overall negative trend and particularly highlighting the dispersion of data points for Sydney - Blacktown, thereby reinforcing the complexity and variability of socioeconomic dynamics within Greater Sydney.

5. Key Findings

The Z-score distributions for individual resource components (Figure 1, Appendix) reveal that the median Z-score for businesses is -0.12, for transit stops -0.08, for schools -0.28, and for Points of Interest -0.06, suggesting that, on average, most SA2s score slightly below the mean for these individual metrics before aggregation. This boxplot (Figure 1) shows the central tendency, spread, and outliers, helping to understand the typical contribution and variability of each component.

The composite "well-resourced" score distribution, visualized in Figure 2 (Appendix), peaks at lower and higher ends, with fewer regions in the mid-range, suggesting some polarisation in resource levels. Regionally (Table 1 and Figure 3, Appendix), Sydney - Ryde (mean 0.588) scores highest on average but with high inequality (std dev 0.349), Sydney - Parramatta (mean 0.529) is more consistently resourced (std dev 0.232), and Sydney - Blacktown (mean 0.339) is the least resourced on average, also with high internal variation (std dev 0.314). Analysis of extreme SA2s (Figure 4, Appendix) reveals that both highly and poorly resourced areas exist across different SA4s, underscoring intra-SA4 disparities.

No statistically significant overall correlation was found between well-resourced scores and median income ($r = -0.196$). However, a significant moderate negative correlation ($r = -0.623$, $p = 0.002$) exists in Sydney - Blacktown. The regression plot (Figure 5, Appendix) visually supports the weak overall trend.

6. Data Visualisations

6.1. Static Choropleth Map

To visualize the spatial distribution of the "well-resourced" scores across the selected SA2 regions, a static choropleth map was generated (Figure 6, Appendix); note that a few SA2 areas may appear without color as they were excluded from scoring due to having fewer than 100 young residents.

6.2. Interactive Map

For interactive exploration, please refer to the map in Section 6.2 of the submitted Jupyter Notebook.

7. Scrutiny of Results

7.1. Component Impact Analysis

To assess each component's influence on the "well-resourced" composite score, we performed three analyses suitable for understanding different facets of impact: direct linear influence (Pearson correlation for strength

of linear relationship with the total score), rank stability (Spearman's rank correlation in a leave-one-out approach to see how rankings change), and distributional changes (comparing mean, median, std dev of scores when a component is removed).

All components exhibited statistically significant positive Pearson correlations with the composite score ($p < 0.05$). Points of Interest (POIs) showed the strongest linear influence ($r = 0.715$), followed by transit stops ($r = 0.504$), businesses ($r = 0.401$), and school catchments ($r = 0.291$).

The leave-one-out Spearman rank correlation analysis revealed that removing businesses had the least effect on ranking stability ($\rho = 0.889$), while excluding POIs had the greatest impact on rank order ($\rho = 0.803$). Removing stops ($\rho = 0.832$) and schools ($\rho = 0.846$) had intermediate effects on ranking. This underscores the key role of POIs and, to a lesser extent, stops and schools in determining the relative ranking of SA2s.

Distributional statistics (mean/median/std dev of the composite score, originally 0.475/0.445/0.300) showed notable shifts:

- Removing POIs: Mean changed to 0.477, median to 0.416, std dev to 0.254 (largest reduction in spread and median).
- Removing Schools: Mean changed to 0.498, median to 0.521, std dev to 0.312 (median increased most).
- Removing Businesses: Mean changed to 0.491, median to 0.489, std dev to 0.293.
- Removing Stops: Mean changed to 0.473, median to 0.403, std dev to 0.287. This suggests POIs significantly contribute to score variability and higher scores, while schools tend to moderate scores. Removing businesses or transit stops caused smaller distributional shifts.

In summary, POIs are the most impactful component, driving both linear correlation and score variability. School catchments and stops also significantly affect ranking stability and score distribution, while businesses exert the least influence on these aspects among the four components.

7.2. Limitations and Biases

Our "well-resourced" score offers a snapshot of amenity access but is shaped by several key methodological choices and data characteristics. The definition of a "resource" itself is subjective, as our selection of specific business industries (e.g., retail, healthcare, excluding heavy industry) and POI types (e.g., parks, libraries, excluding quarries) prioritizes everyday community-facing services; different criteria would yield different results. Furthermore, by equally weighting the z-scores of business, transport, school, and POI metrics, we assume equal importance for each, which may not reflect diverse community needs. The z-score normalization, while standardizing metrics, can be sensitive to outliers, potentially amplifying the impact of exceptionally high or low raw counts in certain SA2s, and the subsequent sigmoid transformation compresses differences at the extremes of the 0-1 scale.

Data-wise, the score relies on counts (e.g., number of businesses/stops, school catchment overlaps) and does not inherently capture the quality, capacity, or actual utilization of these resources. For instance, our school metric, which counts catchment overlaps per 1,000 young people (excluding SA2s with under 100 young residents to maintain statistical stability), indicates choice but not necessarily school quality or availability of places. Similarly, spatial analysis counting resources within SA2 boundaries doesn't account for "edge effects" where residents might easily access amenities in adjacent SA2s. Future enhancements could explore weighted metrics, robust scaling techniques, and methods to incorporate resource quality or cross-boundary accessibility to provide an even more nuanced understanding.

Appendix

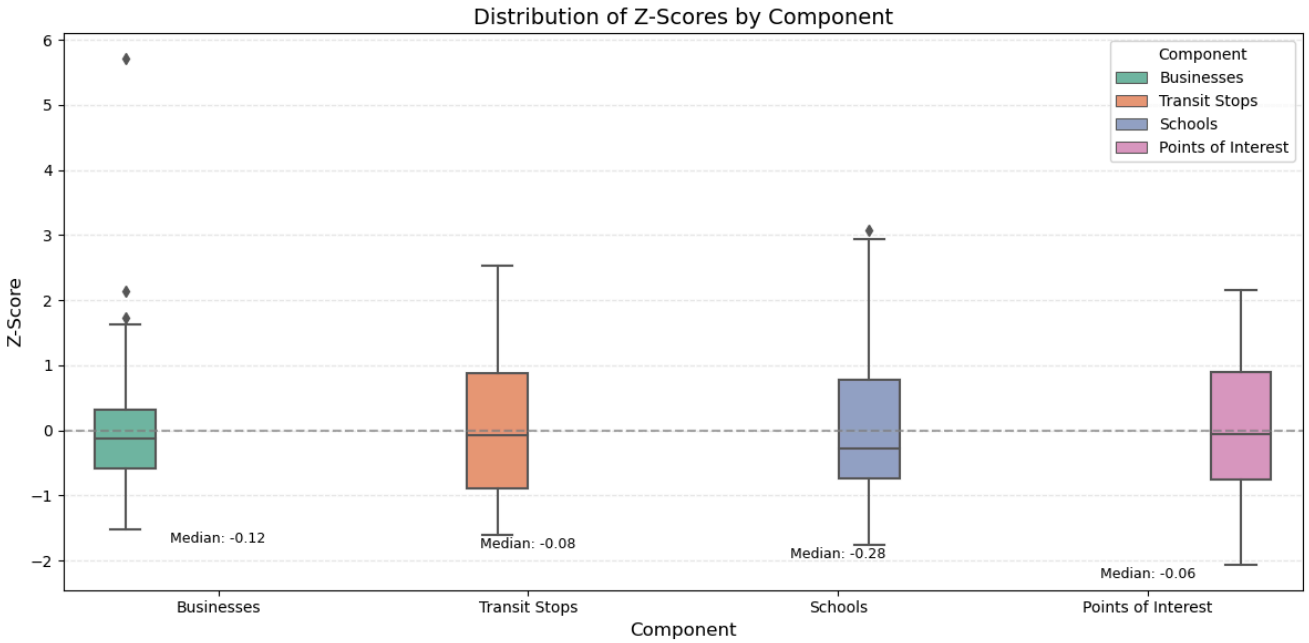


Figure 1. Z-Score Distribution Boxplot

Figure 2. Well-Resourced Score Distribution Barplot

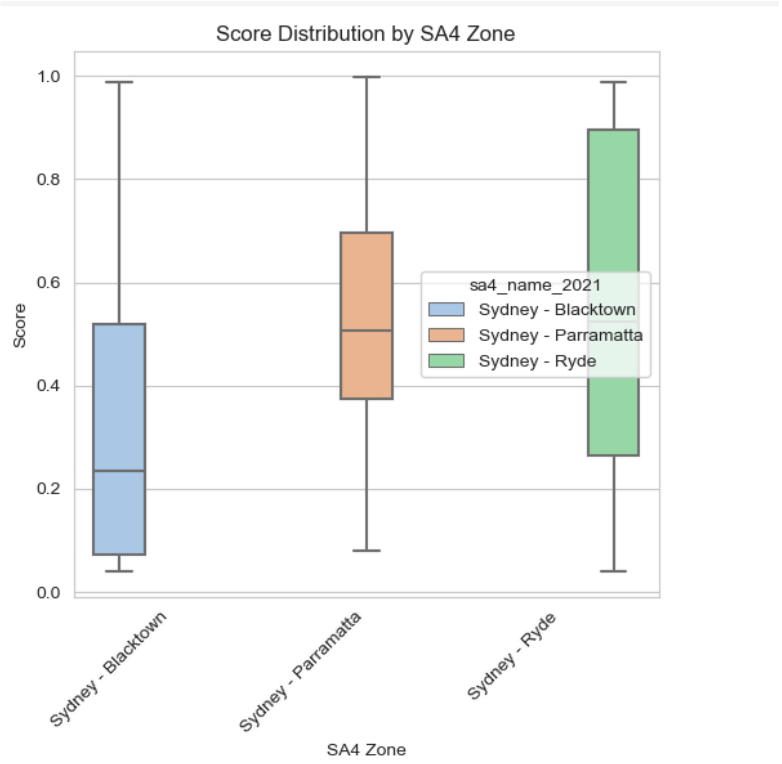


Figure 3.Well-Resourced Score Distribution by SA4 Regions Boxplot

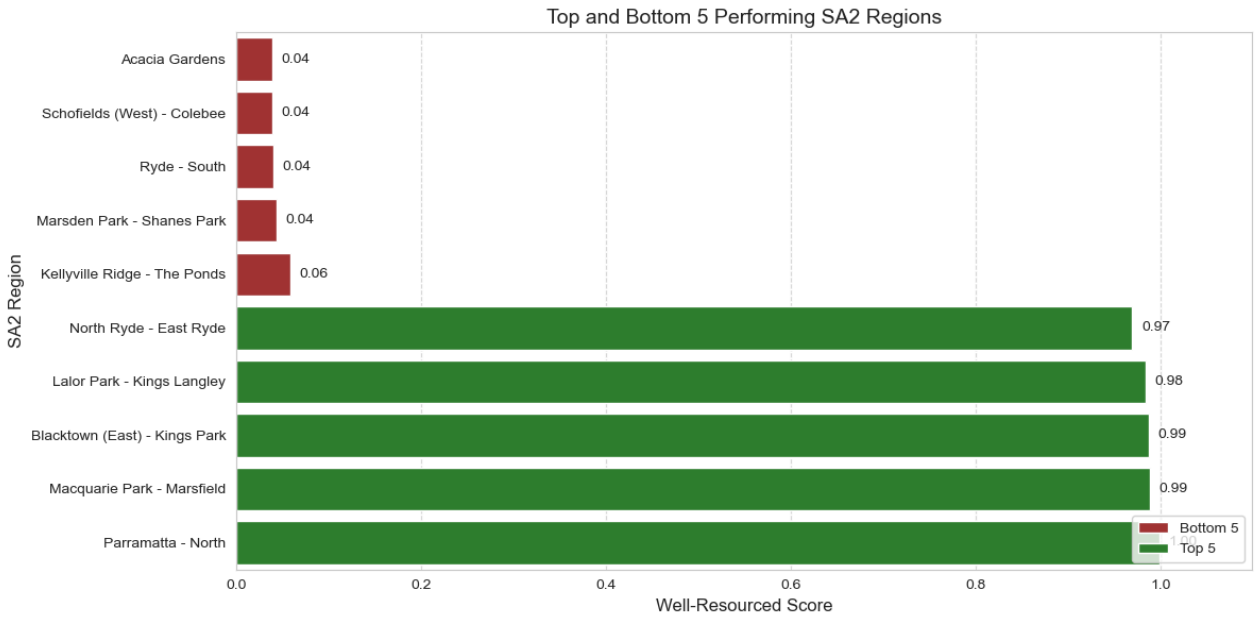


Figure 4.Top and Bottom 5 Performing SA2 Regions by Well-Resourced Score

Figure 5. Well Resourced Score vs Median Income Regression Plot

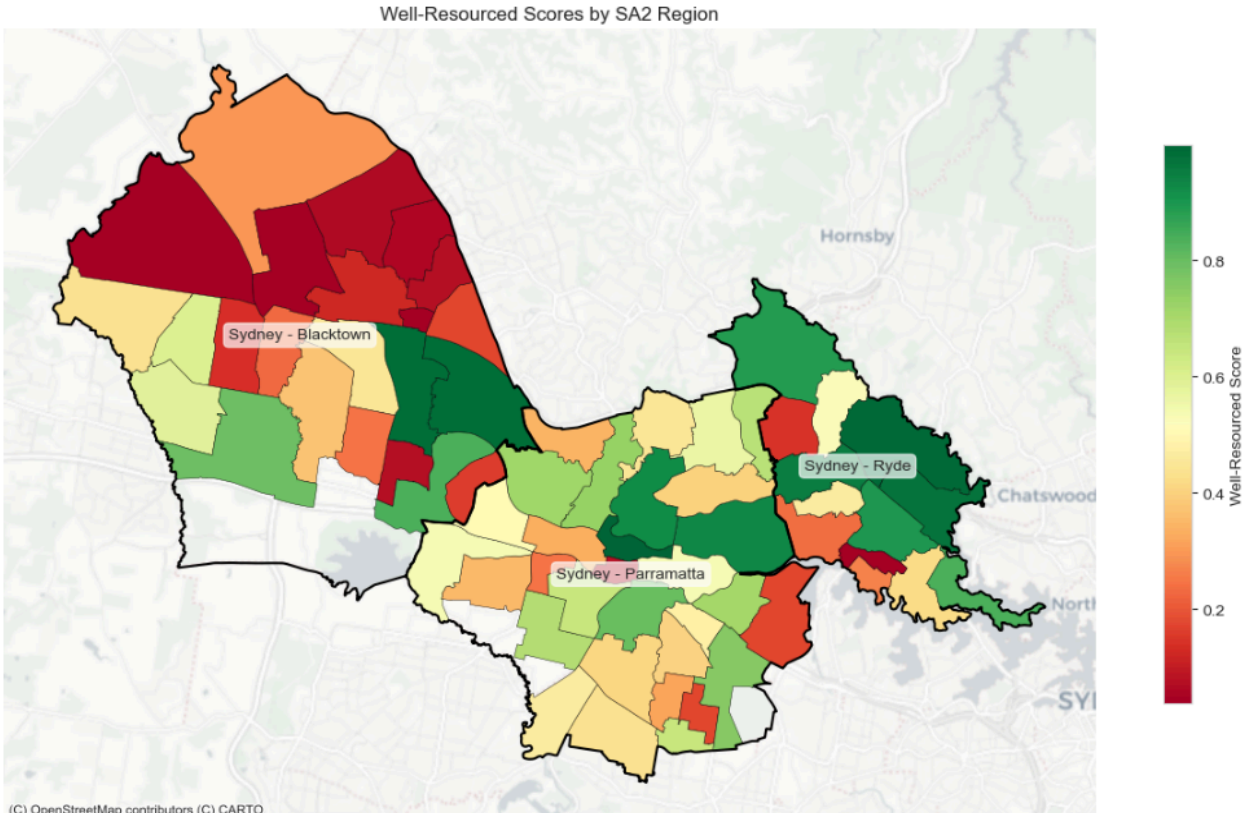


Figure 6. Geographic Distribution of Well-Resourced Scores across Selected SA2 Regions