Greater Sydney Analysis

By,
Abtahee Islam, SID: 530405083 –

Dataset Description

A range of datasets had to be imported. This included SA2 Regions, Businesses, Stops, Schools, Population and Income. The SA2 Regions boundaries shapefile was collected from the Australian Bureau of Statistics (ABS) which provided the spatial data types to distinguish each region and allowed for analysis. The Businesses dataset was sourced from the ABS as well, which collected the data through quarterly and annual Australian Business Number (ABN) registrations recorded in the Australian Business Register (ABR). The dataset consisted of the number of businesses in each industry for each SA2 region. The data of public transport Stops were collected through the NSW Government Transport Open Data Hub and consisted of stops, geographical locations and wheelchair boarding. School Catchment zones were provided by the NSW Government's Department of Education where each school's geographical location, name and available years were provided. The Points of Interest (POI) were extracted from the NSW Points of Interest API for each SA2 region. The income and population datasets were not provided with an original source. The datasets were given for analysis for the assignment.

Database Description

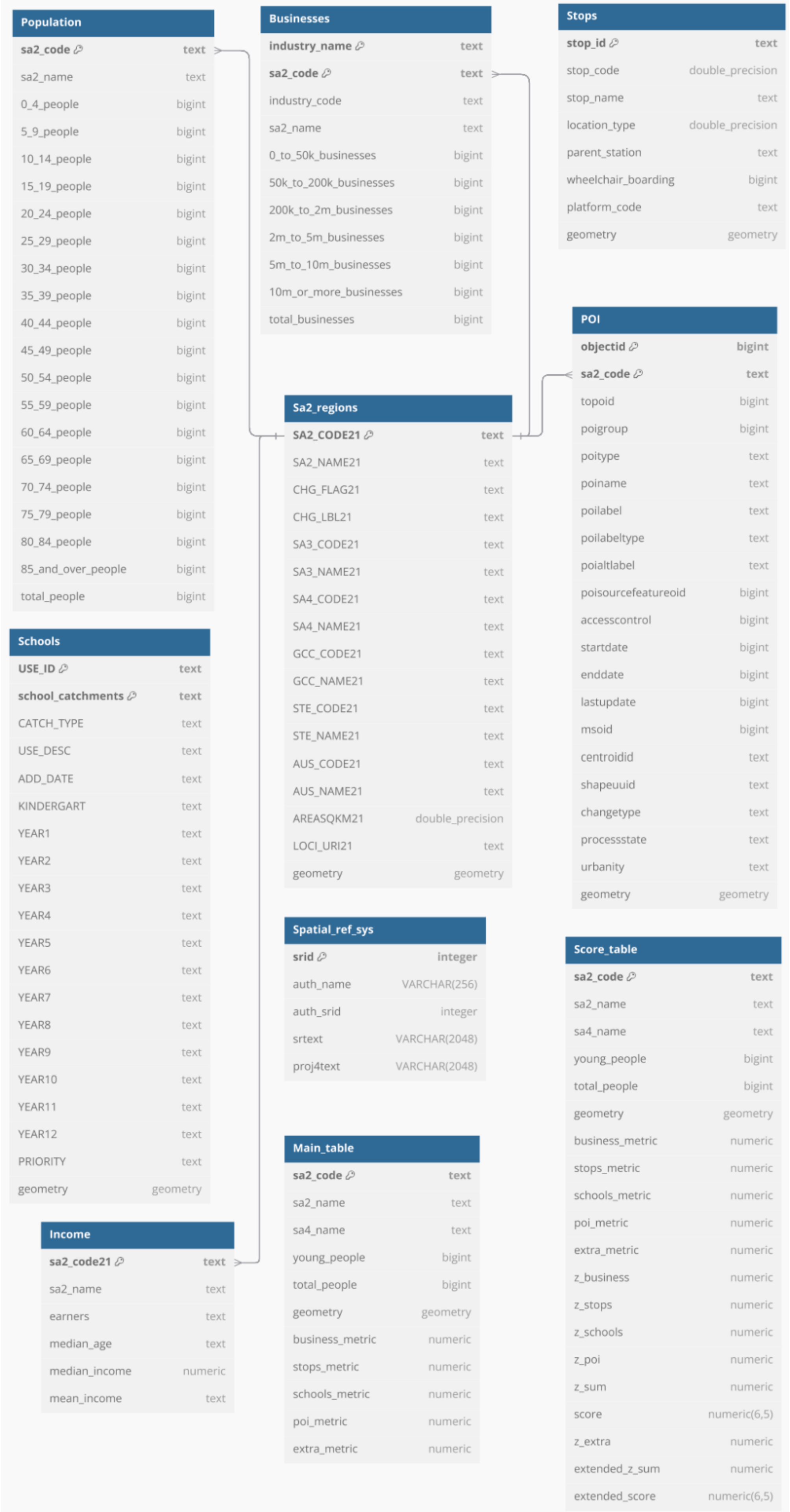
To ensure efficient data management, each table was assigned primary keys. In most tables there was a single column such as sa2_code or stop_id which was used as a primary key. For datasets where one column was not enough to differentiate rows, such as the businesses, school catchments and POI datasets, a composite primary key (e.g. industry_name and sa2_code) was implemented. Foreign keys were used to ensure relationships between tables, such as linking the population table to the main SA2 regions table.

Additionally, indexes were created to improve query performance. B-tree indexes were created for the sa2_regions dataset and the population dataset to help speed up filtering and joining when calculating metrics and scores. A spatial index was created for the SA2 regions dataset using GIST, which improves the speed of queries that check whether points fall within the boundaries (multipolygons). Overall, the indexing methods were essential as they reduced query times for large datasets.

To clean the datasets, rows were removed to maintain data consistency and prevent errors in the calculation of scores and metrics. This was prevalent in the income dataset as multiple rows consisted of records with 'np', which would affect correlation calculations, so they were removed. The school catchment datasets which consisted of primary, secondary and future tables were all combined into a single table. An extra column called "school_catchments" was added, which helped differentiate between school types. This was done to simplify querying as all data was combined in one table.

The datasets were all set to EPSG:4326, which means all spatial data follows the WGS 84 coordinate system. This was chosen as it is the standard coordinate system for GPS globally used by companies like Google.

Data Schema Diagram



Population	
sa2_code	text
sa2_name	text
0_4_people	bigint
5_9_people	bigint
10_14_people	bigint
15_19_people	bigint
20_24_people	bigint
25_29_people	bigint
30_34_people	bigint
35_39_people	bigint
40_44_people	bigint
45_49_people	bigint
50_54_people	bigint
55_59_people	bigint
60_64_people	bigint
65_69_people	bigint
70_74_people	bigint
75_79_people	bigint
80_84_people	bigint
85_and_over_people	bigint
total_people	bigint

Schools	
USE_ID	text
school_catchments	text
CATCH_TYPE	text
USE_DESC	text
ADD_DATE	text
KINDERGART	text
YEAR1	text
YEAR2	text
YEAR3	text
YEAR4	text
YEAR5	text
YEAR6	text
YEAR7	text
YEAR8	text
YEAR9	text
YEAR10	text
YEAR11	text
YEAR12	text
PRIORITY	text
geometry	geometry

Income	
sa2_code21	text
sa2_name	text
earners	text
median_age	text
median_income	numeric
mean_income	text

Businesses	
industry_name	text
sa2_code	text
industry_code	text
sa2_name	text
0_to_50k_businesses	bigint
50k_to_200k_businesses	bigint
200k_to_2m_businesses	bigint
2m_to_5m_businesses	bigint
5m_to_10m_businesses	bigint
10m_or_more_businesses	bigint
total_businesses	bigint

Sa2_regions	
SA2_CODE21	text
SA2_NAME21	text
CHG_FLAG21	text
CHG_LBL21	text
SA3_CODE21	text
SA3_NAME21	text
SA4_CODE21	text
SA4_NAME21	text
GCC_CODE21	text
GCC_NAME21	text
STE_CODE21	text
STE_NAME21	text
AUS_CODE21	text
AUS_NAME21	text
AREASQKM21	double_precision
LOCI_URI21	text
geometry	geometry

Spatial_ref_sys	
srid	integer
auth_name	VARCHAR(256)
auth_srid	integer
srtext	VARCHAR(2048)
proj4text	VARCHAR(2048)

Main_table	
sa2_code	text
sa2_name	text
sa4_name	text
young_people	bigint
total_people	bigint
geometry	geometry
business_metric	numeric
stops_metric	numeric
schools_metric	numeric
poi_metric	numeric
extra_metric	numeric

Stops	
stop_id	text
stop_code	double_precision
stop_name	text
location_type	double_precision
parent_station	text
wheelchair_boarding	bigint
platform_code	text
geometry	geometry

POI	
objectid	bigint
sa2_code	text
topoid	bigint
poigroup	bigint
poitype	text
poiname	text
poilabel	text
poilabeltype	text
poialtlabel	text
poisourcefeatureoid	bigint
accesscontrol	bigint
startdate	bigint
enddate	bigint
lastupdate	bigint
msoid	bigint
centroidid	text
shapeuuid	text
changetype	text
processstate	text
urbanity	text
geometry	geometry

Score_table	
sa2_code	text
sa2_name	text
sa4_name	text
young_people	bigint
total_people	bigint
geometry	geometry
business_metric	numeric
stops_metric	numeric
schools_metric	numeric
poi_metric	numeric
extra_metric	numeric
z_business	numeric
z_stops	numeric
z_schools	numeric
z_poi	numeric
z_sum	numeric
score	numeric(6,5)
z_extra	numeric
extended_z_sum	numeric
extended_score	numeric(6,5)

Score Analysis

Formula Computation and Rationale:

The scoring formula is used to evaluate how well resourced each SA2 region is based on four normalised z-scores. This includes the number of businesses in a chosen industry per 1000 people, number of public transport stops in the region, school catchment areas per 1000 young people and the number of points of interest (POIs). These components were chosen as they provided an accurate representation of infrastructure and services within the zone. The z-scores were calculated for each metric and the final score for each SA2 region used the sigmoid function on the sum of the four z-scores calculated previously. This returned a result between 0 and 1. The specific industries selected were 'Information Media and Telecommunications', 'Transport, Postal and Warehousing' and 'Manufacturing'. These were chosen because they are strongly associated with employment and economic activity. These were also used to see the strength of infrastructure and services within the regions. The selected POI groups were 3, 5, 8 and were chosen as they represent the important aspects of the social wellbeing and functionality within the SA2 zones.

The impact of different components on the overall score:

Each metric component in the scoring formula had a direct impact on the overall scores of each SA2 region. The businesses metric calculated the number of businesses per 1000 people in selected industries which consisted of 'Information Media and Telecommunications', 'Transport, Postal and Warehousing' and 'Manufacturing'. The standard deviation of 30.1835(4.dp) shows that there are large differences in business density through SA2s which impacted our score significantly. The Stops metric, which calculated the number of public transport stops in each SA2 zone and had a mean of 88.9394(4.dp) played a key role in determining the accessibility and how well-connected a region is. School catchment areas per 1000 young people had a very low mean and standard deviation of 0.0064(4.dp) and 0.0103(4.dp). This indicates that only a small number of SA2s have good school catchment coverage, meaning that it would heavily impact the score if an SA2 has multiple within it. The POIs introduced more variety to the score with a standard deviation of 28.6147(4.dp) as it took into account the recreational areas, area types (urban, suburban, city, etc) and infrastructure-related services. The most in favour with high scores were the SA2 regions with high amounts of recreational facilities (e.g. park, clubs, etc.) which helped increase community wellbeing. Overall, business density and school catchment areas had the strongest impacts on SA2 scores, while POIs and public transport stops had less significant impacts.

Summary of Overall Distribution:

The overall distribution of scores across all SA2 regions within the selected SA4 regions shows moderate variability, which is supported by the standard deviation of the scores being 0.2769(4.dp). The scores range from a minimum of 0.0510(4.dp) to 0.9999(4.dp). The median of the discovered scores is 0.3791(4.dp). This means that most SA2 regions in the selected zones are below the threshold of being well-resourced, meaning based on metrics and z-scores, more than half are under-resourced. The spread of scores shows a moderate variability and highlights how moderately resourced regions will experience uneven access to resources such as services and infrastructure. These statistics give an insight into the overall distribution of infrastructure and services such as business, stops, schools and POIs.

Furthermore, the three SA4 zones show disparity in how resources are distributed throughout the region. The Inner West has a consistent resource distribution, while City and Inner South and Eastern Suburbs show a high degree of spread with a few extreme outliers and mostly low resourced infrastructure and resources. This is prevalent in both the box plot (Figure 1.4) as it shows both Eastern Suburbs and City and Inner South being highly positively skewed. This is also shown in the choropleth map (Figure 1.3) as the region of Inner West appears greener compared to the other regions. Overall, regions should strive for more consistent distribution of resources across the SA2 zones.

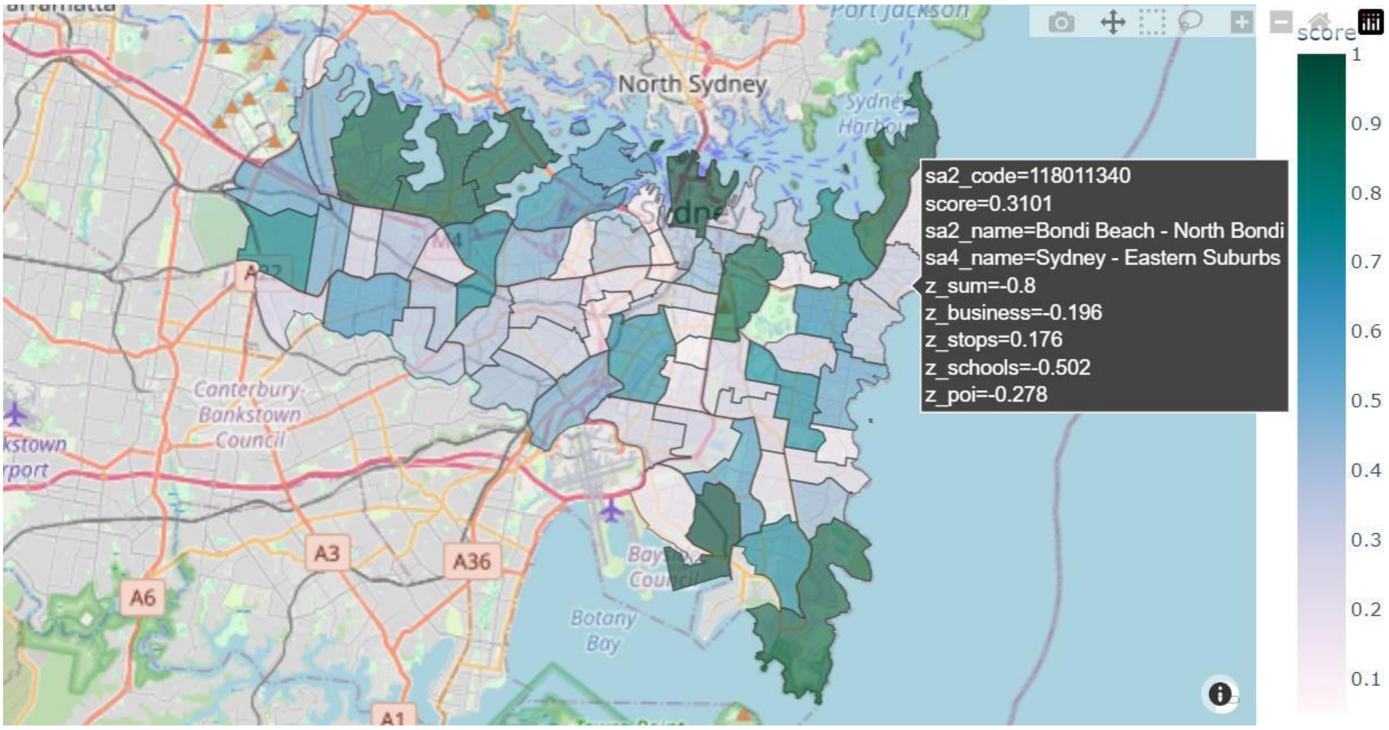

sa2_code=118011340
score=0.3101
sa2_name=Bondi Beach - North Bondi
sa4_name=Sydney - Eastern Suburbs
z_sum=-0.8
z_business=-0.196
z_stops=0.176
z_schools=-0.502
z_poi=-0.278

Figure 1.2 Interactive Map of SA2 regions showing score of each zone

Commentary on how the distribution of scores differ across SA4 zones:

Comparing the results of the statistics gathered in selected SA4 zones of City and Inner South, Eastern Suburbs and Inner West, the major imbalance in resource distribution is highlighted.

Sydney - Inner West

The Inner West SA4 region is considerably more well-resourced than the other SA4 zones, as it boasts the highest median among all SA4s with a score of 0.4999(4.dp) as well as the highest mean of 0.5128(4.dp). This implies that the average SA2 zone in the Inner West had reliable access to resources, such as infrastructure and services. A narrow IQR of 0.3108(4.dp) highlighted the consistency and showed fewer outliers. These statistics highlight a well-distributed infrastructure network and support equal access to services.

Sydney – City and Inner South

This zone displays a low median and mean score of 0.2935(4.dp) and 0.3469(4.dp) respectively which indicates that most SA2 regions are under-resourced. The zone has a Q3 of 0.4380(4.dp), which shows that more than 75 percent of SA2 regions have a score below this value, highlighting that the limitations of infrastructure and services are prevalent throughout the SA4. The IQR of 0.3312(4.dp) shows low spread. The wide range of scores from 0.0510(4.dp) to 0.9999(4.dp) shows that there are a few high-scoring SA2 outliers. This emphasises that there is uneven resource distribution, where only a small number of zones have high resources while generally most zones do not.

Sydney – Eastern Suburbs

The Eastern Suburbs have a low median score of 0.3533(4.dp) but a high IQR of 0.5029(4.dp). This suggests there is high variability in resources across the region. The high spread of scores indicates some SA2 zones in the top 25 percent, with scores of at least 0.7133(Q3), are well-resourced, but a significant number still fall far below the average score. The low mean score of 0.2993(4.dp) further supports this, as many SA2s have low scores, emphasising limited access to resources despite a moderate number of high-performing areas.

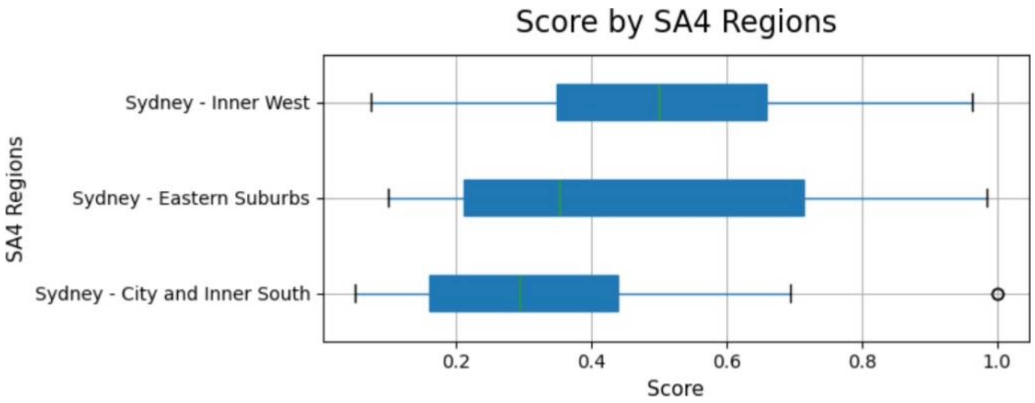


Figure 1.3 Boxplot of Scored Distribution in Selected SA4 Zones

Trends, regions or scores of interests:

When analysing statistics for SA2 zones, several outliers and patterns were found. This is most prevalent in the SA2 zones of Millers Point and Banksmeadow, which have extremely high scores of 0.9999. These scores result from very high z scores in business and school metrics. This is despite the fact that these SA2 are part of the lowest-scoring SA4 zone. This is shown in the stacked z-sum plot (Figure 1.5) as the SA2s have a large bar in both the business and school components.

Another observation is that some SA2 zones, whilst not having high scores in any individual metric, can achieve a moderately high score by maintaining moderately positive values in all metrics. For example, the SA2 zone of Five Dock – Abbotsford did not dominate in any single metric but showed positive z-scores in business, stops and POIs which resulted in the region returning a strong score, which was fifth highest in scores measured for all SA4s. This shows that an SA2 does not need to be an outlier in one z metric to get a high score, it would be better if it had a consistent and even distribution of infrastructure and services which can lead to better accessibility and liveability without relying on one specific metric.

Additionally, another pattern is the large variation within City and Inner South SA4 region, where some SA2s significantly outperform the rest, while in contrast, regions such as Zetland and Pyrmont have the lowest scores. This highlights inequality within the region, where only some have a high concentration of infrastructure and services, while most remain under-resourced. This suggests resource allocation is uneven, leading to only small pockets of high-resource zones.

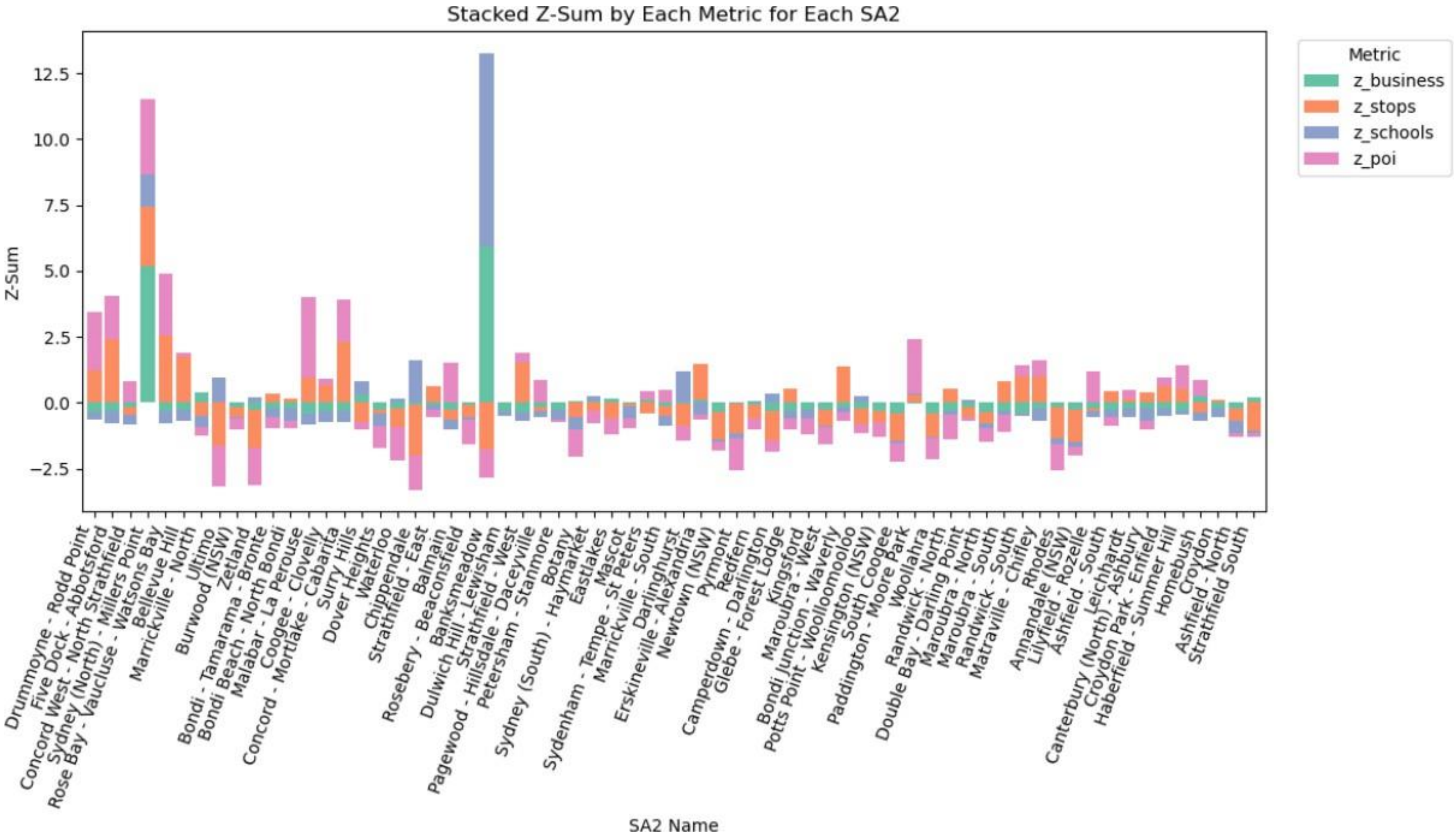


Figure 1.4 Stacked Z-sum Plot of Each SA2 Zone

Correlation Analysis

When exploring the correlation between resource availability and the median income of each SA2 region, a Pearson correlation coefficient was used, as it is an effective way to measure the relationship between two normally distributed variables. The coefficient provides a score between -1 and 1, which indicates whether the median income of a region is associated with how well-resourced it is.

Sydney - Inner West

The Inner West shows a moderate positive correlation between resource score and median income with a Pearson correlation coefficient of 0.35. This suggests that higher income SA2 regions have better access to resources. However, the correlation is not strong, indicating that while income might have some influence, it is not the sole factor. This is supported by the consistent resource distribution within the Inner West, reflected by a high mean of 0.5128, indicating that the region is well-resourced even though the coefficient is weak. This implies that other factors play a much more important role as there is an even spread of infrastructure and service availability.

Sydney – Eastern Suburbs

In the Eastern Suburbs region, the correlation coefficient of 0.27 indicates a weak positive relationship between median income and resource score. The low correlation emphasises that there is almost no link between high income levels and high resource availability. The large IQR reflects high variability and aligns with the coefficient calculated, showing that median income is not strongly associated with scores.

Sydney - City and Inner South

City and Inner South also show a weak positive correlation of 0.28, which points to a weak association between higher income areas and better resource availability. The correlation is not supported by the data as the region has many low-scoring SA2 regions and a few extreme outliers, which further reduces the strength of correlation between income levels and resource scores, which creates a misleading idea that higher income levels result in higher amounts of infrastructure and services. Once again, this further accentuates that income is not the sole factor affecting resource availability within a SA2.

The overall correlation of all SA4 regions returned a Pearson correlation coefficient of 0.28, which indicates a weak positive correlation. This highlights that there is a weak trend where higher median income regions tend to have better resources. However, the relationship remains weak, so median income by itself cannot be a reliable indicator. In Figure 2.1, the scatter-plot further supports this pattern. There is a large spread of scores across the various median income levels. Notably, the highest scores were found between $60000 and $70000, which is considered an average salary. Several low median income SA2s achieved high scores while several high median income zones had low scores. This implies that while median income has a slight influence on access to infrastructure and services, other factors are much more important in determining how many resources an SA2 region receives.

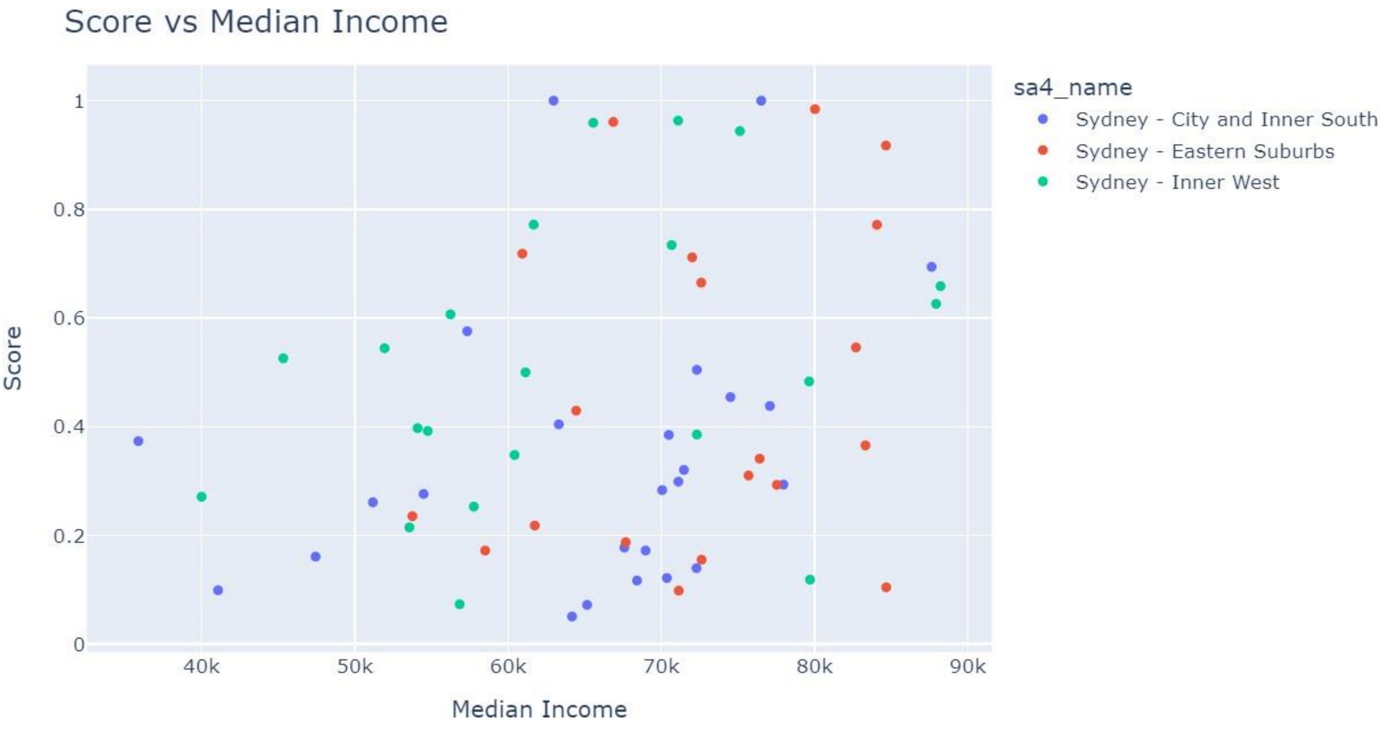


Figure 2.1 Scatter Plot of Score vs Median Income