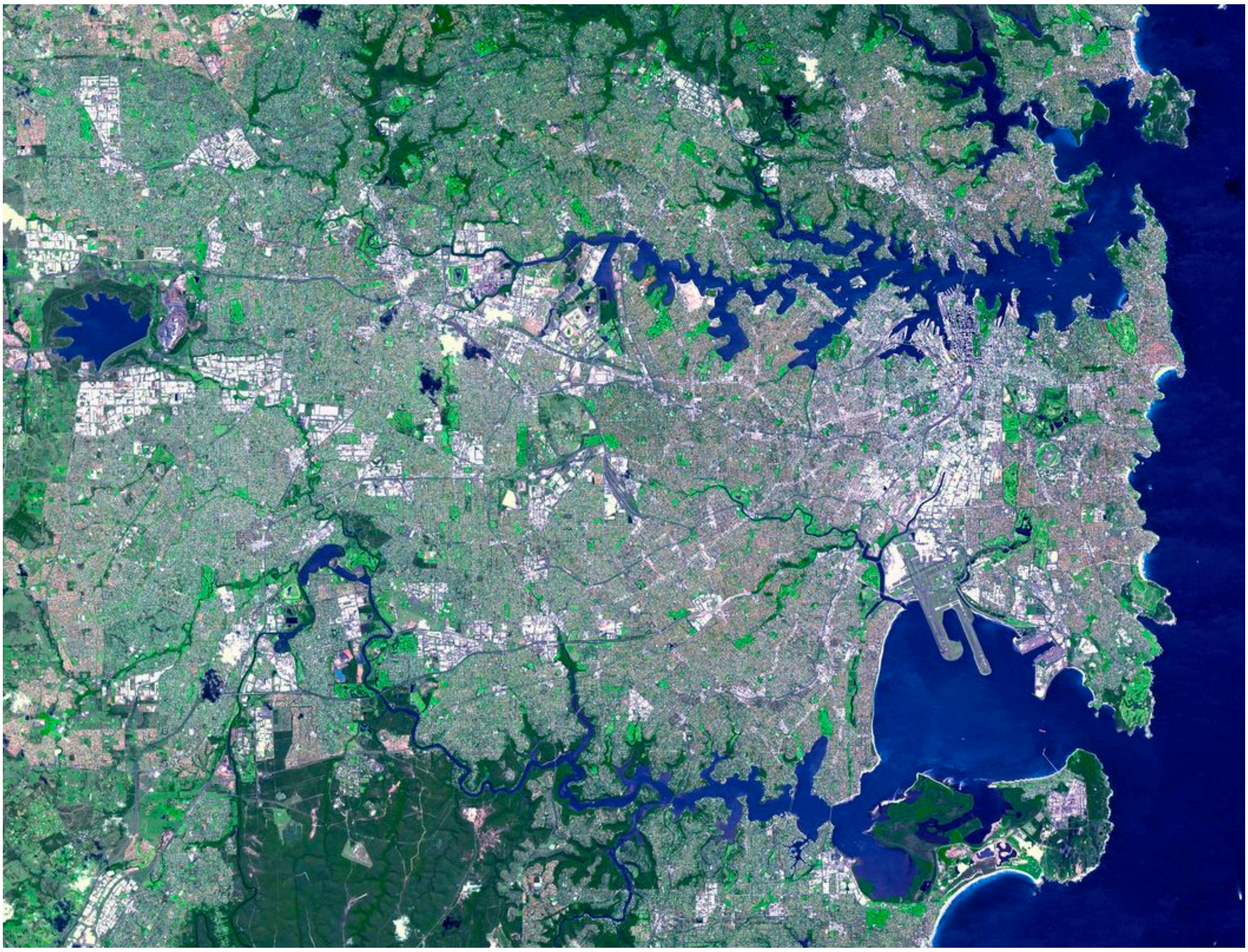


ASSIGNMENT

DATA2001 : Data Science, Data Variety, and Data Variety

Lab20 - Group 02

Members :
Ayush Singh(520584277),
Devang Jindia(530206893),
Raghav Dogra (520214754)

Introduction

This report outlines the group assignment on the Greater Sydney Area. We have integrated various datasets into a PostgreSQL server, along with the PostGIS extension for spatial data queries. Our datasets are combinations of publicly available records and data provided to us at the beginning of this assignment. We have used several python libraries along with SqL to read, clean, find co-relations, and plot the data frames.

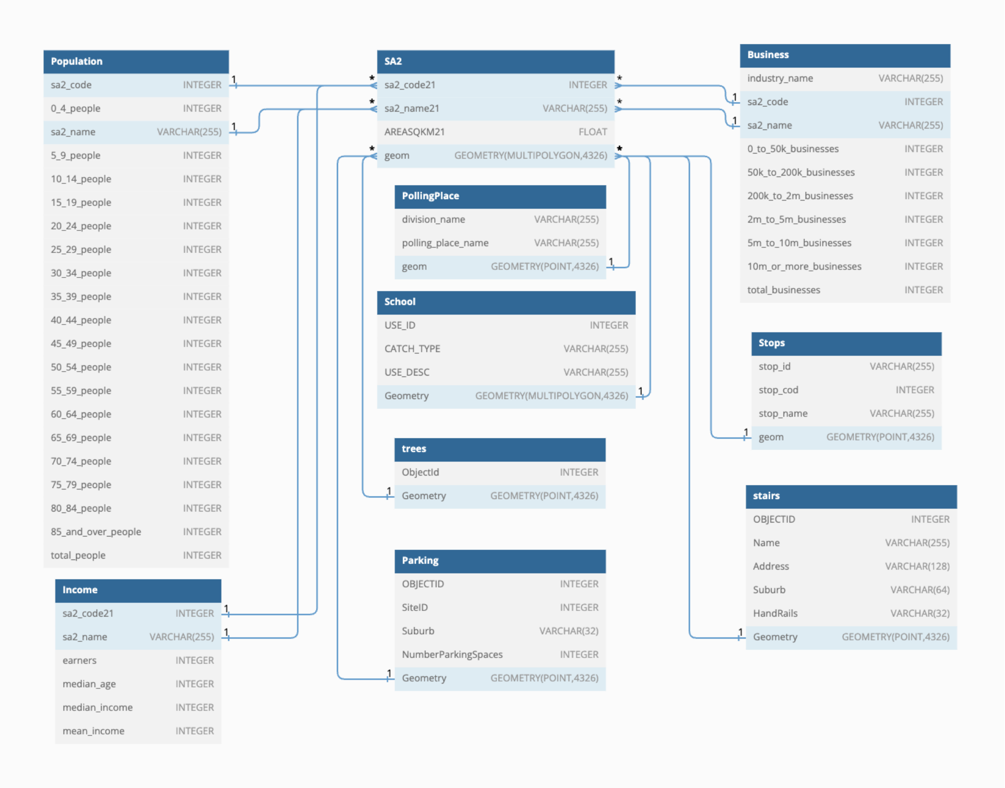
Dataset Description

Our datasets are named SA2 Digital Boundaries, Businesses, Public Transportation Stops, Polling Places, School Catchments, Population estimates, Income Statics, Mobility Parking, Trees, and Stairs. Brief information about their sources and preprocessing are as follows.

SA2 Digital Boundaries : This data frame was downloaded from the ABS website. The data frame originally consists data of whole Australia, hence we filtered out the data set to only represent the Greater Sydney Area in our pre-processing. The data-frame had several columns to begin with, but we decided to only keep the *SA2_Code21*, *SA2_Name21*, *AreaSqKm*, and *Geom* for our analysis, since other columns had irrelevant data. Business : This data is sourced from Australian Bureau of Statistics. In our pre-processing phase, we dropped all the duplicated and null values and the '*industry_code*' column since it was redundant information given '*industry_name*' existed.' Public Transportation Stops : The stops data frame is sourced from Transportation for NSW. In pre-processing we handled data type conversion along with the creation of a '*geom*' column form the '*latitude*' and '*longitude*' columns. We also dropped '*location_type*', '*parent_station*', '*wheelchair_boarding*', '*platform_code'* since they were irrelevant to our analysis. Polling Places : This data frame was sourced form the Australian Electoral Commission. In the preprocessing we dropped all the columns except '*division_name*' , '*polling_place_name*' and '*geom*'. School Catchments : This data frame was secured from NSW Department of Education. We decide to not use the CatchmentFuture data frame since it had no impact on the current time and our data analysis. In prepousseing we converted the geospatial data to suitable forms and merged the other two data frames. Population : The population data frame was provided to us with the assignment scaffold. In preprocessing we dropped the rows with zero '*total_people*' since those rows were not adding value to our analysis. Income : This data Fram was also provided to us with the assignment scaffold. In preprocessing, we did type correction and dropped all the duplicate and NaN values. Mobility Parking : This data frame is from the *CityOfSydney* website. In it's preprocessing we have created the WKT Element for the '*geom*' column and dropped rows with irrelevant values. Trees : This data frame is from *CityOfSydney*. In it's preprocessing we have created the '*geom*' column from '*latitudes*' and '*longitudes*' column and dropped all the irrelevant columns. Stairs : This data frame is also from *CityOfSydney* and it's preprocessing includes creating the WKTElement along with dropping columns not relevant to our analysis.

Data Schema

The database schema is primarily based on the data frames being joint on SA2 Code, SA2 Name, along with the Geospatial Data Available from the SA2 Digital Boundaries data frame. We have indexed on SA2 Digital Boundaries data frame using the ‘geom’ column for all the geospatial data-frame. For the data frames with non-geospatial data, we have used the ‘SA2_Code’ or ‘SA2_Name’ for the creation of indexes. We have primarily used SA2 column since they are the basis our analysis structure and the digital boundaries of SA2 cover the whole Greater Sydney Area and correlate to all the other data frames.



Population	
sa2_code	INTEGER
0_4_people	INTEGER
sa2_name	VARCHAR(255)
5_9_people	INTEGER
10_14_people	INTEGER
15_19_people	INTEGER
20_24_people	INTEGER
25_29_people	INTEGER
30_34_people	INTEGER
35_39_people	INTEGER
40_44_people	INTEGER
45_49_people	INTEGER
50_54_people	INTEGER
55_59_people	INTEGER
60_64_people	INTEGER
65_69_people	INTEGER
70_74_people	INTEGER
75_79_people	INTEGER
80_84_people	INTEGER
85_and_over_people	INTEGER
total_people	INTEGER

Income	
sa2_code21	INTEGER
sa2_name	VARCHAR(255)
earners	INTEGER
median_age	INTEGER
median_income	INTEGER
mean_income	INTEGER

SA2	
sa2_code21	INTEGER
sa2_name21	VARCHAR(255)
AREASQKM21	FLOAT
geom	GEOMETRY(MULTIPOLYGON,4326)

PollingPlace	
division_name	VARCHAR(255)
polling_place_name	VARCHAR(255)
geom	GEOMETRY(POINT,4326)

School	
USE_ID	INTEGER
CATCH_TYPE	VARCHAR(255)
USE_DESC	VARCHAR(255)
Geometry	GEOMETRY(MULTIPOLYGON,4326)

trees	
ObjectId	INTEGER
Geometry	GEOMETRY(POINT,4326)

Parking	
OBJECTID	INTEGER
SiteID	INTEGER
Suburb	VARCHAR(32)
NumberParkingSpaces	INTEGER
Geometry	GEOMETRY(POINT,4326)

Business	
industry_name	VARCHAR(255)
sa2_code	INTEGER
sa2_name	VARCHAR(255)
0_to_50k_businesses	INTEGER
50k_to_200k_businesses	INTEGER
200k_to_2m_businesses	INTEGER
2m_to_5m_businesses	INTEGER
5m_to_10m_businesses	INTEGER
10m_or_more_businesses	INTEGER
total_businesses	INTEGER

Stops	
stop_id	VARCHAR(255)
stop_cod	INTEGER
stop_name	VARCHAR(255)
geom	GEOMETRY(POINT,4326)

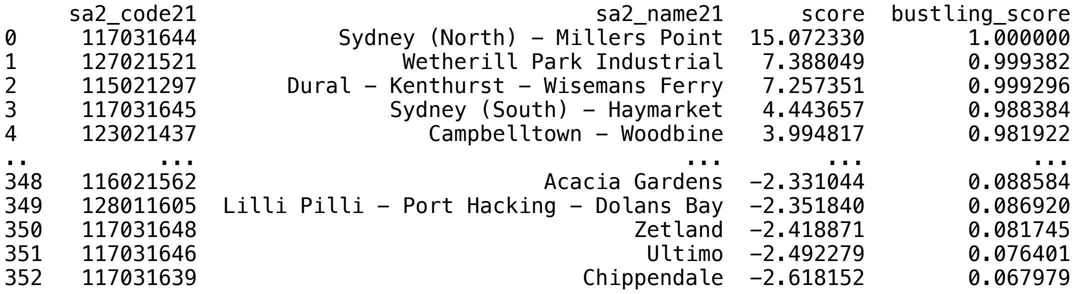
stairs	
OBJECTID	INTEGER
Name	VARCHAR(255)
Address	VARCHAR(128)
Suburb	VARCHAR(64)
HandRails	VARCHAR(32)
Geometry	GEOMETRY(POINT,4326)

Score Analysis

For the z-Score of the Business data frame we chose to group the data by Electricity, Gas, Water and Waste Services industry, since the their presence supports excess demand of consumers of these resources, and more the resources are consumed, more bustling the area becomes. We have joined *Business* and *Population*, on the ‘*sa2_code*’ column to get information about each region. Finally we filter the data to include only regions related to the 'Electricity, Gas, Water and Waste Services' industry. For the z-score of the stops in each SA2 area we subtracted the average count of stops across all SA2 areas from the count of stops in each SA2 area, and then divided by the standard deviation of the count of stops across all SA2 areas. For the polling places we calculated the difference between the count of polling places in a specific region and the average count of polling places across all regions, and then scales this difference by the standard deviation of the count of polling places across all regions. Finally for schools in each region based on population under age 19, the z-score is determined by subtracting the mean from the average schools per 1000 under-19 population for that area and dividing the result by the standard deviation. To find our Sigmoid function, which is defined as :

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

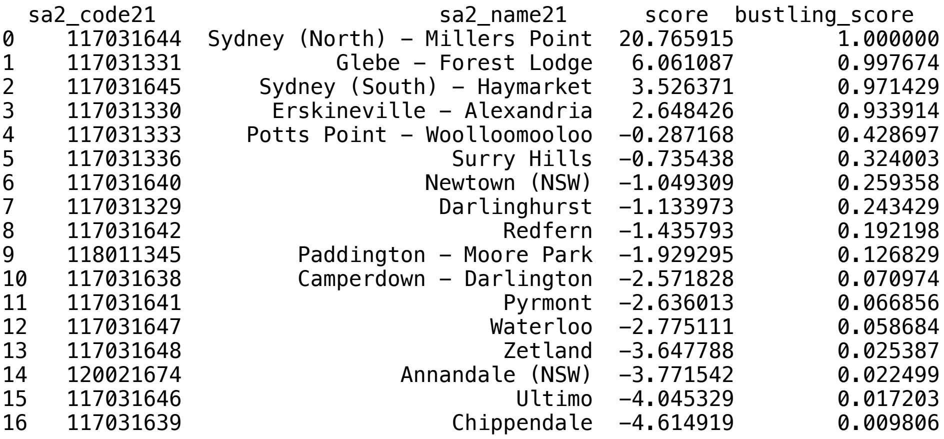
x is changed to the combined z-score calculated for each SA2 area based on the selected variables. This changes the linear z-score values into a bounded range between 0 and 1, representing the bustling score of each SA2 area, which is in-terms indicating the bustling score of each SA2 area. The table is as follows :



```
     sa2_code21                          sa2_name21      score  bustling_score
0     117031644          Sydney (North) - Millers Point  15.072330        1.000000
1     127021521            Wetherill Park Industrial   7.388049        0.999382
2     115021297      Dural - Kenthurst - Wisemans Ferry   7.257351        0.999296
3     117031645            Sydney (South) - Haymarket   4.443657        0.988384
4     123021437              Campbelltown - Woodbine   3.994817        0.981922
..          ...                                 ...        ...             ...
348   116021562                      Acacia Gardens  -2.331044        0.088584
349   128011605  Lilli Pilli - Port Hacking - Dolans Bay  -2.351840        0.086920
350   117031648                             Zetland  -2.418871        0.081745
351   117031646                              Ultimo  -2.492279        0.076401
352   117031639                         Chippendale  -2.618152        0.067979
```

This data frame show that areas with the highest bustling scores are "Sydney (North) - Millers Point", "Wetherill Park Industrial", and "Dural - Kenthurst - Wisemans Ferry". These areas have high combined bustling-score, indicating a significant presence of businesses, public transportation stops, polling places, and schools relative to their population under 19. The lowest score is of “Chippendale” suggesting comparatively much lower bustling area.

After this we added the z-scores of our own imported data frame, i.e trees, parking spaces, and stairs. The z-score is calculated separately for each SA2 area based on the count of trees, parking spaces, and stairs in each area respectively. After merging our data sets, we get the following data frame :



```
   sa2_code21                      sa2_name21      score  bustling_score
0    117031644  Sydney (North) - Millers Point  20.765915        1.000000
1    117031331           Glebe - Forest Lodge   6.061087        0.997674
2    117031645      Sydney (South) - Haymarket   3.526371        0.971429
3    117031330       Erskineville - Alexandria   2.648426        0.933914
4    117031333     Potts Point - Woolloomooloo  -0.287168        0.428697
5    117031336                    Surry Hills  -0.735438        0.324003
6    117031640                  Newtown (NSW)  -1.049309        0.259358
7    117031329                   Darlinghurst  -1.133973        0.243429
8    117031642                        Redfern  -1.435793        0.192198
9    118011345         Paddington - Moore Park  -1.929295        0.126829
10   117031638         Camperdown - Darlington  -2.571828        0.070974
11   117031641                        Pyrmont  -2.636013        0.066856
12   117031647                       Waterloo  -2.775111        0.058684
13   117031648                        Zetland  -3.647788        0.025387
14   120021674                 Annandale (NSW)  -3.771542        0.022499
15   117031646                         Ultimo  -4.045329        0.017203
16   117031639                    Chippendale  -4.614919        0.009806
```

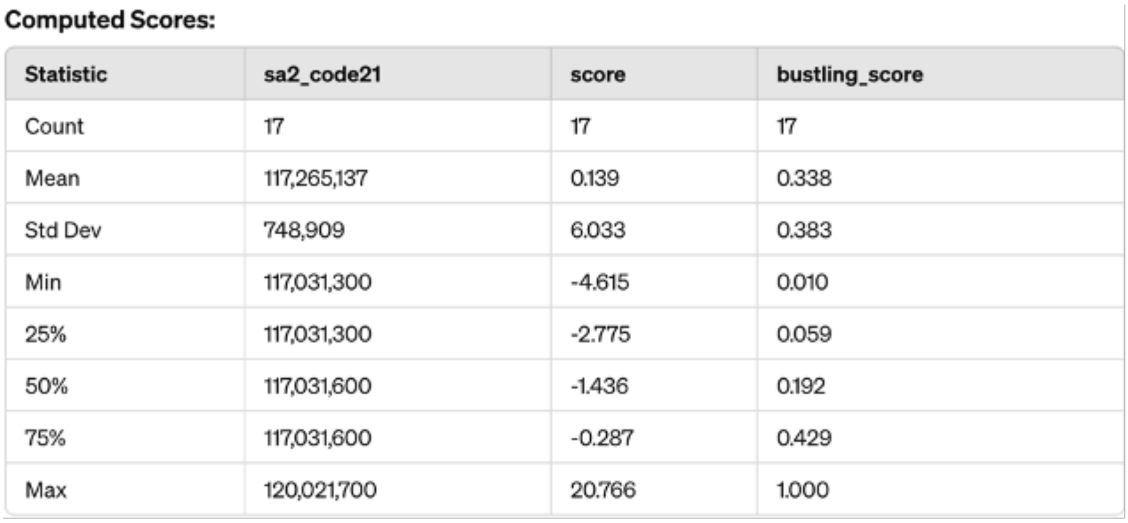
We can see that "Sydney (North) - Millers Point" is still at the top and "Chippendale" is still at the bottom but the other arrangements have changed, implying that trees, parking spaces, and stairs have high impact one the bustling of an area. Overall, "Sydney (North) - Millers Point" can be considered as one of the most bustling places in the Greater Sydney Area.

Correlation Analysis

The correlation coefficient of -0.0101 indicates an almost negligible negative relationship between the bustling score and median income. This suggests that higher bustling scores do not correspond to higher or lower median incomes in a meaningful way, which is surprising since income often relates to high social activities. Therefore, factors contributing to the bustling score appear to be independent of the median income levels in these regions.

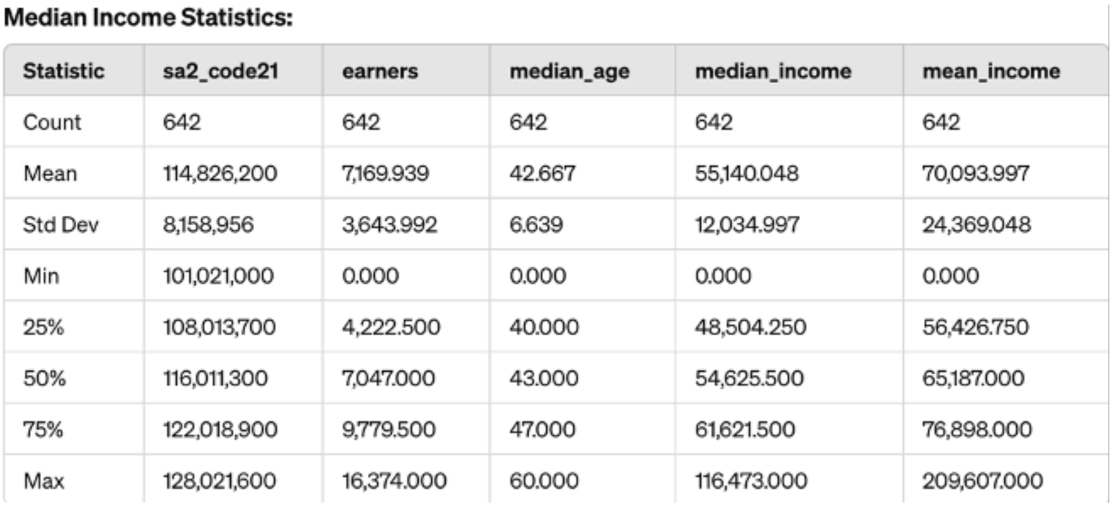
The data used in the analysis provide additional context:

- The '*bustling_score*' ranged from 0.009806 to 1.000000 with a mean of 0.338173, indicating a wide variance in the level of bustling activity across different regions.
- Median income statistics ranged broadly upto $116,473, with a mean income of approximately $55,140.



Computed Scores:

Statistic	sa2_code21	score	bustling_score
Count	17	17	17
Mean	117,265,137	0.139	0.338
Std Dev	748,909	6.033	0.383
Min	117,031,300	-4.615	0.010
25%	117,031,300	-2.775	0.059
50%	117,031,600	-1.436	0.192
75%	117,031,600	-0.287	0.429
Max	120,021,700	20.766	1.000

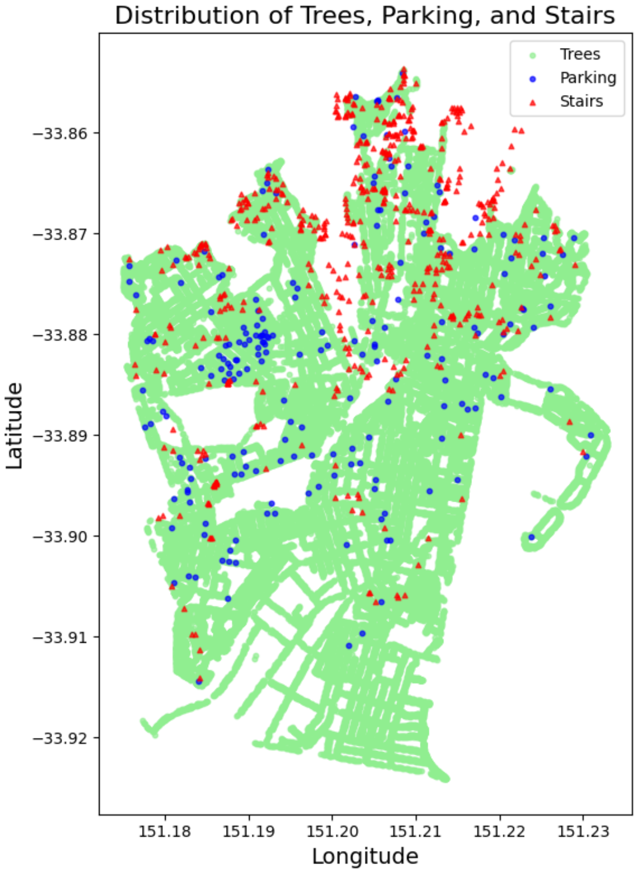


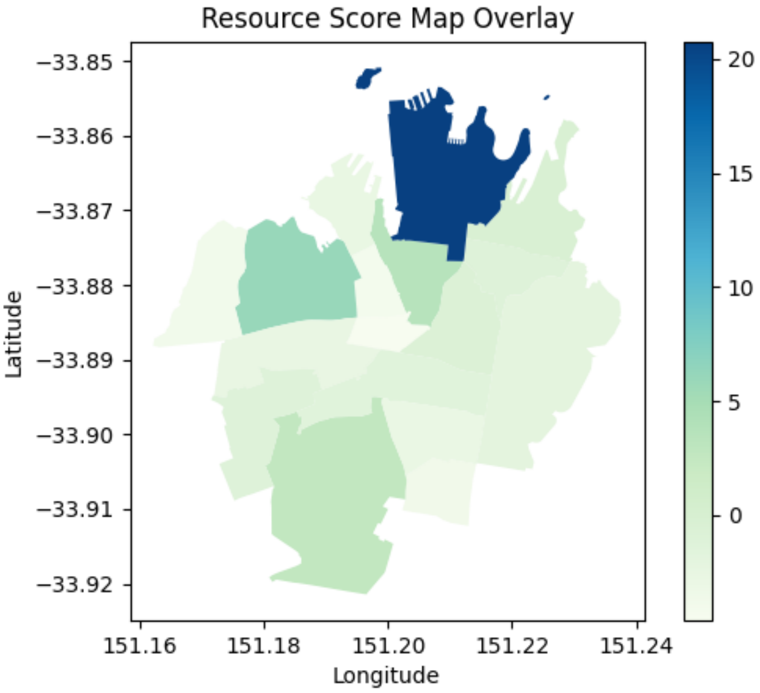
Median Income Statistics:

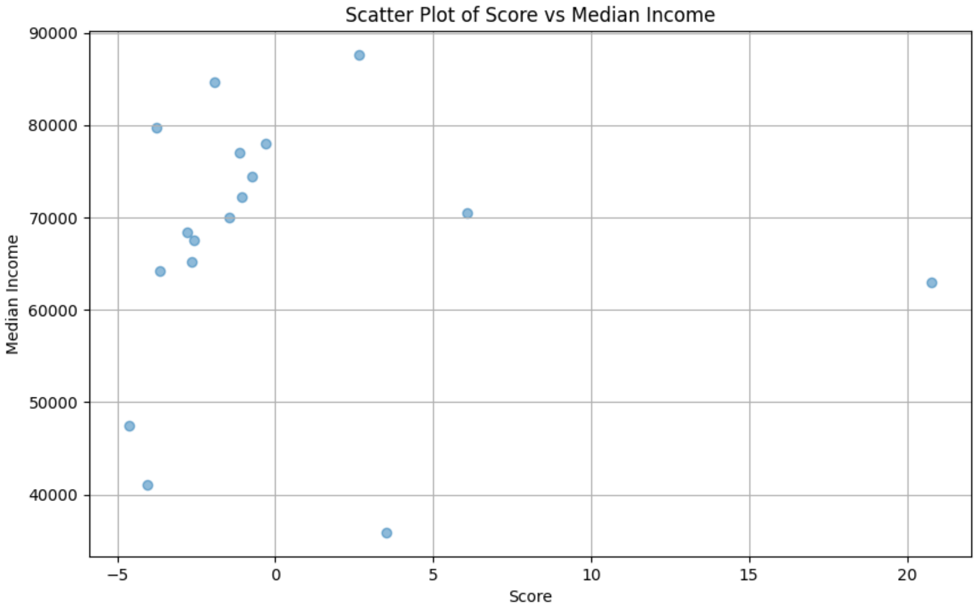
Statistic	sa2_code21	earners	median_age	median_income	mean_income
Count	642	642	642	642	642
Mean	114,826,200	7,169.939	42.667	55,140.048	70,093.997
Std Dev	8,158,956	3,643.992	6.639	12,034.997	24,369.048
Min	101,021,000	0.000	0.000	0.000	0.000
25%	108,013,700	4,222.500	40.000	48,504.250	56,426.750
50%	116,011,300	7,047.000	43.000	54,625.500	65,187.000
75%	122,018,900	9,779.500	47.000	61,621.500	76,898.000
Max	128,021,600	16,374.000	60.000	116,473.000	209,607.000

Data Visualisation

The following plots represent the overlayed plots chosen by our team members and their geospatial positioning over the Greater Sydney Area. The next graph shows the bustling score over different regions in Sydney and finally the last one is the scatter plot representing the correlation between the Score and median income.


Distribution of Trees, Parking, and Stairs
Trees
Parking
Stairs
−33.86
−33.87
−33.88
−33.89
−33.90
−33.91
−33.92
Latitude
151.18
151.19
151.20
151.21
151.22
151.23
Longitude


Resource Score Map Overlay
−33.85
−33.86
−33.87
−33.88
−33.89
−33.90
−33.91
−33.92
Latitude
151.16
151.18
151.20
151.22
151.24
Longitude
20
15
10
5
0


Scatter Plot of Score vs Median Income
Median Income
Score