Dataset Description

- **SA2 regions:**

This is a shapefile containing the digital boundaries of SA2 regions located within Australia. This file was obtained on the Australian Bureau of Statistics website. Columns that were deemed unnecessary for querying were dropped, and the SA2 regions were filtered to only include those located in Greater Sydney. Spatial data types were converted to WKT format, using SRID 4326.

- **Businesses:**

This is a csv file that contains how many businesses by industry (e.g health services) are in each SA2 region, reported by turnover size ranges. This file was obtained on the Australian Bureau of Statistics website. Unneeded columns for aggregation were dropped, and SA2 regions include those in Greater Sydney. Spatial data types were converted to WKT format, using SRID 4326.

- **Stops:**

This is a txt file that contains the location of public transport stops (train and bus) by General Transit Feed Specification format. This was located on the Transport NSW Gov Data website. Spatial data types were converted to WKT format, using SRID 4326.

- **Polls:**

This is a csv file that contains the locations of polling places for the 2019 federal election. This dataset comes from the Australian Electoral Commission. and SA2 regions include those in Greater Sydney. Spatial data types were converted to WKT format, using SRID 4326.

- **Schools:**

These contain shape files which denote the regions in NSW that students must live in to attend primary, secondary and future Government schools. and SA2 regions include those in Greater Sydney. This was obtained on the NSW department of education website. Spatial data types were converted to WKT format, using SRID 4326.

- **Population:**

This is a csv file that contains estimates of the number of people that live in a SA2 region by age range. This was obtained on the Australian Bureau of Statistics website. SA2 regions were filtered to only include those in Greater Sydney SA2 regions. Data was also filtered to only include individuals of age range 0 – 19 years old

- **Income:**

This is a csv file consisting of total earning statistics in the years 2019-2021, sorted by SA2 code in regions of Greater Sydney. This was located on the Australian Bureau of Statistics website. Columns were filtered to not include values that are invalid for aggregation.

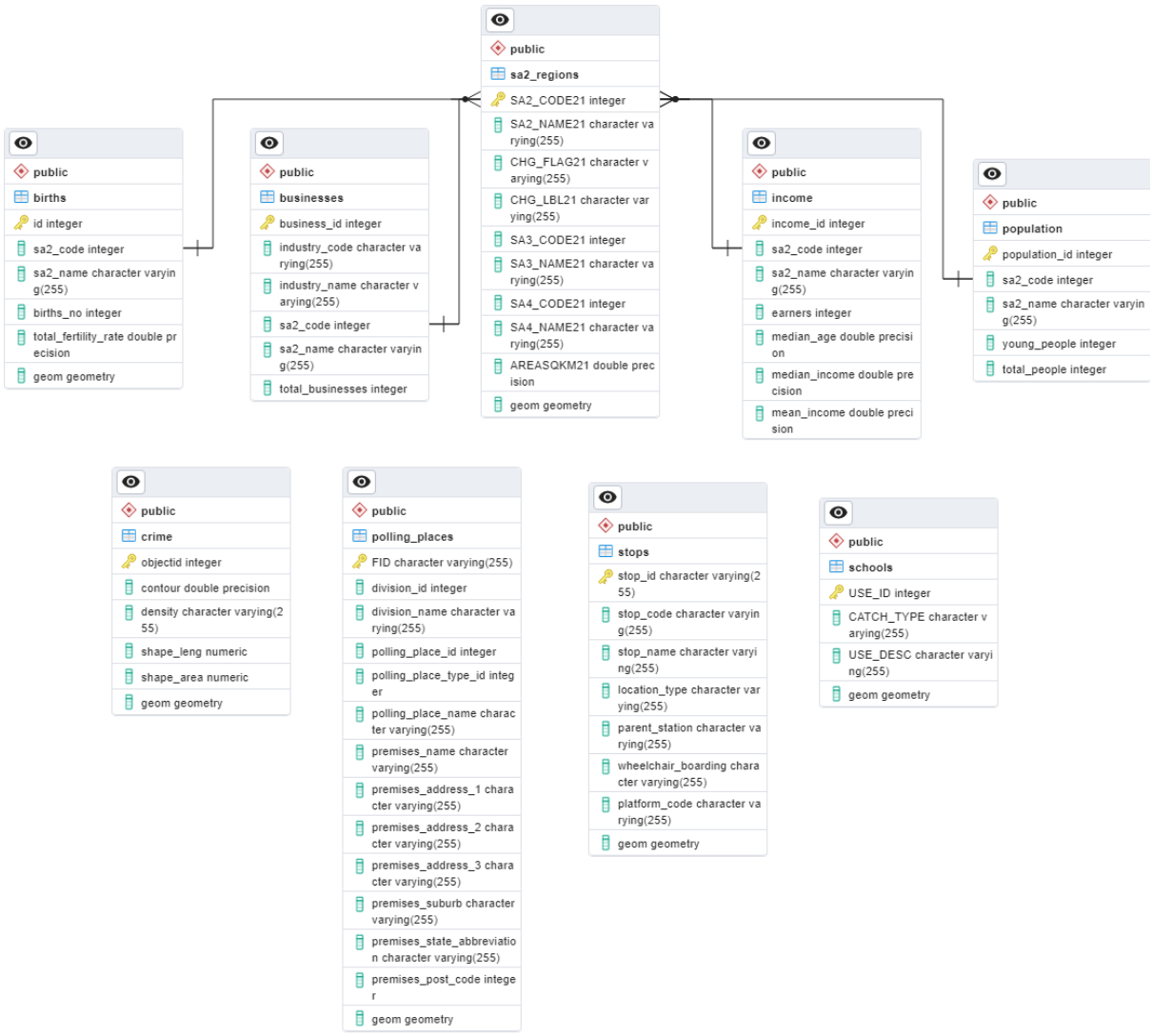
- **Births**

This is a json file containing statistics about birth rates, fertility rates in Australia from the years 2010-2020 by SA2 region. This dataset comes from the Australian Bureau of Statistics. The dataset was filtered so only regions in Greater Sydney were included, and unneeded columns for aggregation were dropped. Birth rates and fertility rates were also filtered to not contain invalid types for aggregation. Spatial data types were converted to WKT format, using SRID 4326.

- **Crime**

This is spatial data that contains crime hotspots of the type ‘stolen from vehicle’, obtained from the NSW Bureau of Crime Website, sorted by number of crime spots within a particular area. Unneeded columns for aggregation were dropped, and spatial data types were converted to WKT format, using SRID 4326.

Database Description



This schema was established by executing the SQL script ('schema.sql') which contains the necessary CREATE TABLE statements for each dataset along with their respective columns and data types. The script also includes the creation of primary keys and foreign key references where applicable.

The data integration into the database is done using the 'to_sql' method from SQLAlchemy. Each dataset is inserted into its respective table that was created in the previous SQL script. For tables that contain geometry data, the 'dtype' parameter is specified with the 'Geometry' type and the appropriate SRID (4326) to ensure the geometry data is stored correctly in the database.

The indexes created improves query performance by allowing faster data retrieval based on the indexed columns. Non-spatial indexes optimise standard column-based filtering and joining, while spatial indexes optimise spatial operations on geometry data for efficient spatial queries.

Score Analysis

Task 2

1. Health Dataset:
    - Formula: total_businesses / (total_people / 1000)
    - This formula calculates the number of health businesses per 1000 people in each region.
    - The total number of health businesses in a region is divided by the total number of people in that region, multiplied by 1000.
2. Polling Places Dataset:
    - Formula: num_polls
    - The score for each region is simply the count of polling places in that region.
3. Retail Dataset:
    - Formula: total_businesses / (total_people / 1000)
    - This formula calculates the number of retail businesses per 1000 people in each region.
    - The total number of retail businesses in a region is divided by the total number of people in that region, multiplied by 1000.
4. Schools Dataset:
    - Formula: (COUNT(s."USE_ID") / pop.young_people) * 1000
    - This formula calculates the number of schools per 1000 young people in each region.
    - The count of schools in a region is divided by the count of young people (population aged below a certain threshold) in that region, multiplied by 1000.
5. Stops Dataset:
    - Formula: COUNT(s.stop_id)
    - The score for each region is simply the count of stops in that region.

Task 3

6. Births Dataset:
    - Formula: births_no / (total_people / 1000)
    - This formula calculates the number of births per 1000 people in each region.
7. Crimes Dataset:
    - Formula: COUNT(crime.objectid) / (pop.total_people / 1000)
    - This formula calculates the number of crimes per 1000 people in each region.

The scripts calculate z-scores for different factors such as the number of stops, number of schools, number of retail businesses, number of polling places, number of health businesses, number of births and number of crime spots in each region. Z-score represents how many standard deviations a data point is from the mean.

These z-scores are then normalized by subtracting the total mean and dividing by the standard deviation, resulting in normalized z-scores. Normalization helps to standardize the data and bring all the factors to a comparable scale.

Finally, a sigmoid function is applied to the normalized z-scores to obtain the final score for each region. The sigmoid function maps the normalized z-scores to a range between 0 and 1, providing a smoothed and bounded score. The sigmoid function used in the script is the logistic sigmoid function: sigmoid(x) = 1 / (1 + exp(-x)).

Task 2 Results

sa2_code	sa2_name	normalised_zscore	sigmoid
117031644	Sydney (North) - Millers Point	15.7334	1.0000
115021297	Dural - Kenthurst - Wisemans Ferry	2.6728	0.9354
117031645	Sydney (South) - Haymarket	2.4932	0.9237
121011684	Chatswood - East	2.0714	0.8881
115011291	Baulkham Hills (West) - Bella Vista	1.9056	0.8705
123021437	Campbelltown - Woodbine	1.7525	0.8523
125041717	Parramatta - North	1.6857	0.8437
117011320	Banksmeadow	1.6761	0.8424
102011030	Calga - Kulnura	1.6684	0.8414
124031464	Penrith	1.4788	0.8144

Based on the results table, the top-ranked region is "Sydney (North) - Millers Point" with a very high normalised z-score of 15.7334 and a sigmoid score of 1.0000, indicating it is an exceptional region in terms of the computed factors. The subsequent regions in the table, such as "Dural - Kenthurst - Wisemans Ferry" and "Sydney (South) - Haymarket," also have high scores, demonstrating their favourable attributes based on the included datasets.

Task 3 Results

sa2_code	sa2_name	normalised_zscore	sigmoid
117031644	Sydney (North) - Millers Point	13.1094	1.0000
119011355	Chullora	2.7808	0.9416
123021437	Campbelltown - Woodbine	2.1499	0.8957
124031464	Penrith	1.7653	0.8539
122031429	Freshwater - Brookvale	1.6535	0.8394
115011291	Baulkham Hills (West) - Bella Vista	1.6163	0.8343
116011303	Blacktown (East) - Kings Park	1.5861	0.8301
117031645	Sydney (South) - Haymarket	1.5411	0.8236
125031483	Guildford - South Granville	1.5375	0.8231
102011032	Gosford - Springfield	1.3614	0.7960

In this updated results table, the region "Sydney (North) - Millers Point" remains at the top with a high normalised z-score of 13.1094 and a sigmoid score of 1.0000. This region continues to exhibit exceptional characteristics based on the birth and crime datasets. Other regions that appear in the table, such as "Chullora," "Campbelltown - Woodbine," and "Penrith," also demonstrate notable scores, suggesting favourable attributes in terms of births and relatively low crime rate per 1000 people.

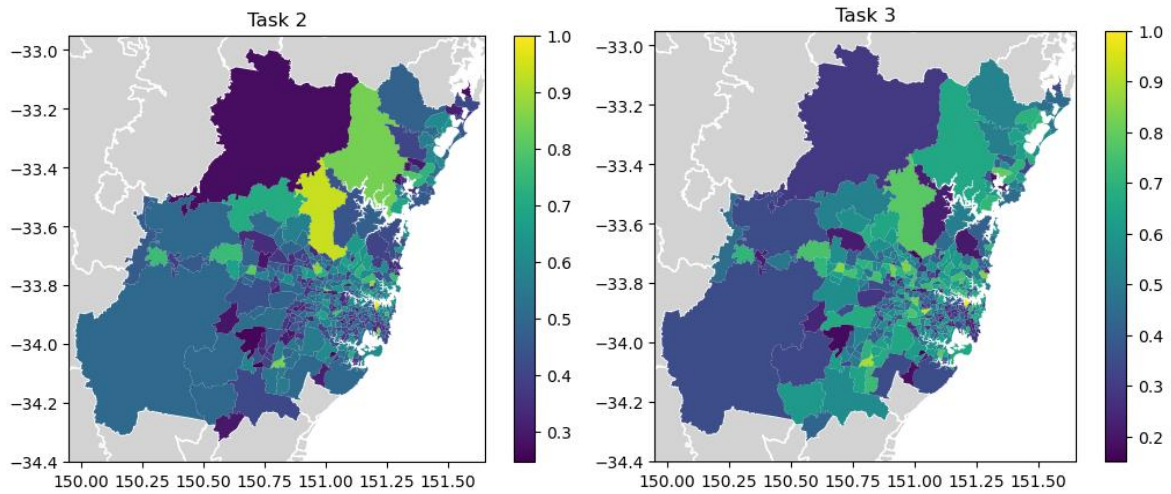
Analysis between Task 2 and Task 3 result tables

In the first results table (including multiple datasets), the top region is "Sydney (North) - Millers Point" with a normalised z-score of 15.7334 and a sigmoid score of 1.0000. This region retains its position as the highest-scoring region even after considering only the birth and crime datasets in the second table, with a slightly lower normalised z-score of 13.1094 and the same sigmoid score of 1.0000. This suggests that "Sydney (North) - Millers Point" consistently stands out across different datasets and criteria.

However, there are some differences in the rankings of other regions. For example, in the first table, "Dural - Kenthurst - Wisemans Ferry" and "Chatswood - East" were ranked second and fourth, respectively. In the second table, these regions have shifted in ranking, with "Chullora" appearing as the second-highest scoring region and "Freshwater - Brookvale" ranked fifth.

Additionally, some regions that were not present in the first table, such as "Chullora" and "Guildford - South Granville," have emerged as significant regions in terms of births and crimes in the second table. On the other hand, regions like "Banksmeadow" and "Calga - Kulnura," which appeared in the first table, are not included in the top 10 regions of the second table.

Map Overlay Visualisation to show scores



In task 2, the scores appear to be randomly scattered with no clear directional spread. This suggests that the scores from task 2 do not exhibit any particular pattern or trend across the regions.

On the other hand, in task 3, the scores are higher in regions that are more centralised. This indicates that the scores in task 3 exhibit a clear spatial pattern, where the central regions tend to have higher scores compared to the peripheral regions.

These contrasting patterns between task 2 and task 3 highlight the differences in the factors or variables being analysed or measured in each task. It suggests that the criteria used to calculate the scores in task 2 are unrelated to the spatial location or centrality of the regions, while task 3 specifically captures and reflects the relationship between centrality and the assigned scores.

Correlation Analysis

We used the Pearson's correlation coefficient to represent the strength and direction of the linear relationship between median income and the overall score. This number was calculated by summing the product of the differences between median income and the mean income, and sigmoid values and the mean sigmoid, divided by the product of the counts and the standard deviations of median income and sigmoid values. This calculation was based on the scores from task 3.

This value came out to be -0.0259 which indicated a very weak and almost negligible negative linear relationship between the median income and sigmoid scores per region. From the sigmoid scores determined from task 3's model, we can say that there is no meaningful linear association between median income and the overall sigmoid score calculated based on the factors considered in the script. Other factors not included in this analysis may have a more significant impact on the overall score. Furthermore, no statistical tests of significance were conducted and therefore it cannot be concluded that the negative correlation between the two variables happened by chance or is statistically significant. Thus, we cannot say for sure that median income has no relationship with the sigmoid score of a region due to a variety of factors not yet analysed.

References

AURIN. (n.d.). Births in Australia 2010 – 2020 by SA2 (json). Retrieved from https://data.aurin.org.au/dataset/au-govt-abs-abs-births-sa2-2010-2020-sa2-2016

AURIN. (n.d.). Federal Election Polling Places 2019 (na). Retrieved from https://data.aurin.org.au/dataset/au-govt-aec-aec-federal-election-polling-places-2019-na

Australian Bureau of Statistics. (n.d.). Counts of Australian Businesses, including Entries and Exits. Retrieved from https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads

BOCSAR. (n.d.). Spatial datasets. Retrieved from https://www.bocsar.nsw.gov.au/Pages/bocsar_datasets/Spatial.aspx

Centre for Education Statistics and Evaluation. (n.d.). School intake zones (catchment areas) for NSW government schools. Retrieved from https://data.cese.nsw.gov.au/data/dataset/school-intake-zones-catchment-areas-for-nsw-government-schools

Transport for NSW. (n.d.). Timetables Complete GTFS. Retrieved from https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs