GREATER SYDNEY ANALYSIS

NITHYA
IYENGAR

DATABASE DESCRIPTION

Database	Source	Description
Business.csv	Australian Bureau of Statistics	Number of businesses categorized by industry and SA2 regions, based on different turnover sizes
Stops.txt	Transport for NSW	Location of all transportation methods (buses, ferries, and trains) in a GTFS format
Population.csv	Australian Bureau of Statistics	Number of people living in SA2 regions, based on age.
PollingPlaces.csv	Australian Electoral Commission	Location of polling places for 2019 Federal Election
Catchments	NSW Department of education	Catchment regions of students wanting to primary, secondary, and future government schools.
Incomes	Australian Bureau of Statistics	Total earnings by SA2 regions.
Crimes	Sydney Suburb Reviews	Total crimes by suburbs, including avg incidents and crime rate per capita.
Recreation	The City of Sydney	List of playgrounds and fitness centers within a region.

Pre-Processing [Data Cleaning]

Data cleaning and pre-processing steps were undertaken to ensure clean and concise data is used for analysis. These steps included:

Business.csv:

- The data was spilt into two different categories: RetailTrade and HealthCare. In the process of doing this, fifteen other industries were removed as they irrelevant to the analysis. Additionally, column names which started with numbers were renamed to a letter format. All the business columns were combined into a singular “total_business” column for easier calculations and the “total_business’ column was also divided by a 1000 to represent the values per 1000 people for RetailTrade and HealthCare.

Stops.txt

- A new column “geom” was created, which stores the latitude and longitude as point data type.

Catchments

- Future:
    - All data under the “geom” column was converted into a multipolygon data type which contained the latitude and longitude. The column “add_date” was also dropped and all the column names were converted to be lowercase.
- Primary:
    - The “priority and “add_date” columns were dropped. Same as future, the “geom” column was converted into a multipolygon data type which contained the longitude and latitude. All the column names were also converted to be lowercase for consistency across the datasets. Moreover, the all the column years from kindergarten to year 6 are checked whether it had a “Y” value, it returned through and thus all columns are combined into a “primary_catchments” column. Columns with year 7 and more are dropped.

- Secondary:
    - The "geom" column was converted into a multipolygon data type, containing longitude and latitude information. It was checked if all years from year 7 to 12 had catchments, and if true, columns from year 7 to 12 were combined into a new column called "secondary_catchments". Columns with year 6 and below were dropped.

Income.csv

- Remove np values from dataset.

Pollingplaces2019.csv

- The state column was dropped as each is NSW. The FID and polling_place_name columns were also dropped as they did not provide significant information for the analysis. Columns such as division_name, polling_place_id, latitude, longitude and the_geom were kept. The “the_geom” column was renamed to “geom” and was converted to a point data type, with the longitude and latitude values.

Population.csv

- Columns which started with numbers were renamed to a letter format for easier and compatibility. A column “total_young_people” was created which contained the sum of people of ages between 0 to 19. A column for the total population per SA2 region was created, “total_people” containing the total population by adding the young people and the rest of the age groups.

SA2 Areas:

- The “geom” column was converted into a multipolygon data type, containing the longitude and latitude. Additionally, all column names were converted to lowercase. Columns “SA2_CODE21”, “SA2_NAME21”, and “geometry” were kept and every other column (SA3_CODE21, SA3_CODE21, etc) were dropped. Additionally, columns “SA2_CODE21” and “SA2_NAME21” were renamed “sa2_code” and “sa2_name” for consistency across the databases.

Crimes

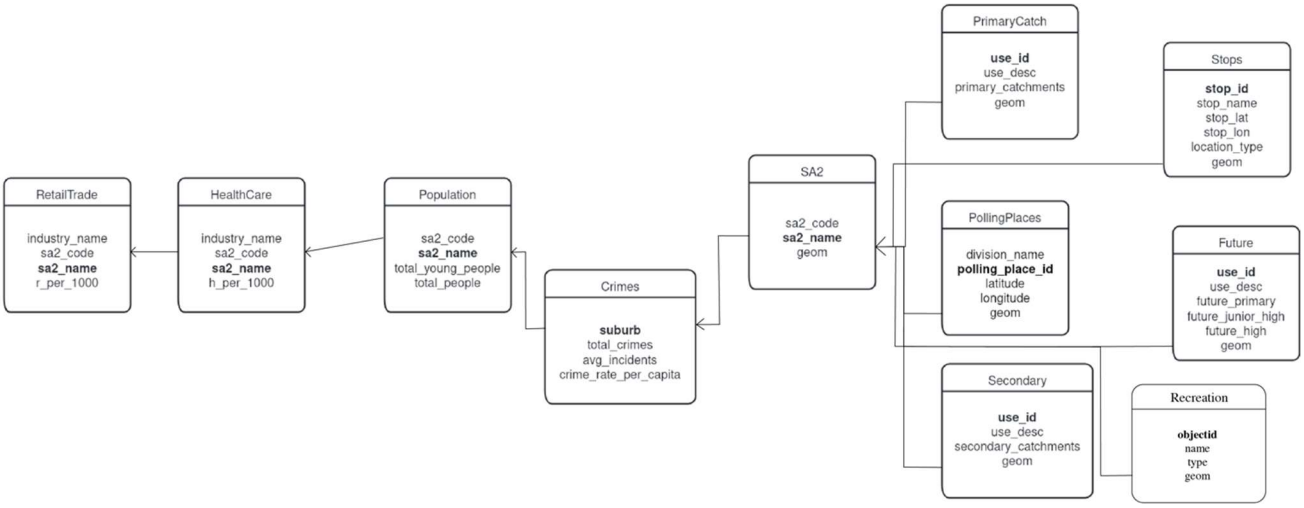
- No cleaning needed; dataset was already clean.

Recreation

- Columns ‘playground’ and ‘fitness’ were combined into a total ‘recreation’

DATABASE DESCRIPTION:

The following database diagram illustrates as to how the schema was established:



The data integration process involved numerous steps. First, relevant datasets were cleaned and extracted from relevant sources, this step was undertaken to ensure consistency and compatibility across the datasets. Then the datasets were combined into a uniform dataset where each dataset was linked to another, whether directly or through other datasets.

Datasets are joined according to their columns, for example, for the schema diagram above, RetailTrade, HealthCare, and Population were all joined based on their column “sa2_codes”, with their primary keys being “sa2_names.” This was conducted to ensure only relevant data is joined across and there aren’t any data losses. Moreover, the additional dataset Crimes was joined into Population through the “sa2_names” and “suburbs” column, with suburb also being the primary key for the Crimes dataset.

Datasets Future, PrimaryCatch, Secondary, Stops and PollingPlaces were all joined onto the SA2 dataset through the “geom” column with primary keys being use_id for Future, PrimaryCatch, Secondary, polling_place_id for PollingPlaces and stop_id for Stops. Moreover, the SA2 dataset was joined to Crimes via the columns “sa2_names” and “suburbs”, where “sa2_names” is the primary key for SA2.

Recreation was also joined into the SA2 dataset, through the common “geom” column, with the primary key for Recreation being “objectid.”

Throughout this integration process, multiple checks and tests were conducted to ensure the data stays accurate and coherent.

SCORE ANALYSIS

In this analysis, the score was computed using the Sigmoid Formula and z scores. Z-scores were calculated by each sa2_name, and its respective dataset. The sigmoid function takes in multiple factors, these include:

1. Z-score of RetailTrade:
2. Z-score of HealthCare
3. Z-score of Stops
4. Z-score of PollingPlaces
5. Z-score of Schools

Each z-score was calculated using the z-score formula, z = (score – mean)/std).
The table below illustrates the mean [represented by 0] and the standard deviation [represented by 1] for each of the given datasets.

	retail	health	primary_schools	secondary	future	stops	polls
0	0.101778	0.117486	1.011364	4.43750	0.039773	155.119318	4.309659
1	0.087904	0.102066	0.106144	1.99474	0.326613	84.487478	4.125954

For RetailTrade, the mean and standard were the number of retail business per 1000 people, from this, the z-score was calculated per sa2_name and the number of businesses within that region. It can be concluded that on average, most z-scores for sa2_name regions lie within the first standard deviation from the mean. Concluding that most regions are “well-resourced” in terms of being near a retail business.

Comparing this to healthcare, the mean and standard division were calculated in the same manner. The z-scores of healthcare exhibited a wider range and illustrated significant variation compared to the z-scores of RetailTrade. For example, the amount of health care facilities per 1000 people was slightly higher than the amount of retail business per 1000 people. As a result, it can be concluded that the regions which are “well resourced” in terms of retail business are also significantly “well resourced” in terms of health care facilities. This might be due some regions are significantly larger in terms population or land mass and thus require more health care and retail facilities compared to smaller regions.

As for Stops, the calculated z-score is a combination of all the train, bus and wharf stops within a given region. This may be a limiting factor when deciding how well resourced an area is in terms of public transport as some regions have only have bus stops and not train station or vis versa. As such, the overall z-scores for Stops lie within two standard deviations from the mean. As a result, most SA2 regions are adequately resourced in terms of access to public transport. Another reason to this may be due some regions have significantly higher buses stops and buses compared to others, and this may generate z-scores to be more scattered.

For PollingPlaces, the z-scores were relatively within the first standard deviation of the mean, meaning that all almost most of the regions were “well-resourced” when it came to the Federal Election in 2019. However, there is a limitation when comparing this dataset with the others as it is almost 4 years old. During this time and now, polling places for elections may have changed/moved and thus affecting how well resources each region is. Moreover, there also might a potential loss of data when joining tables resulting in some polling places being unaccounted for.

For future catchments, the z-score for it was the same across all the regions. This may be due to all future values being 0 in the final dataset table. As a result of this, all the z-scores lead to an inaccurate calculation and thus can be deemed invalid.

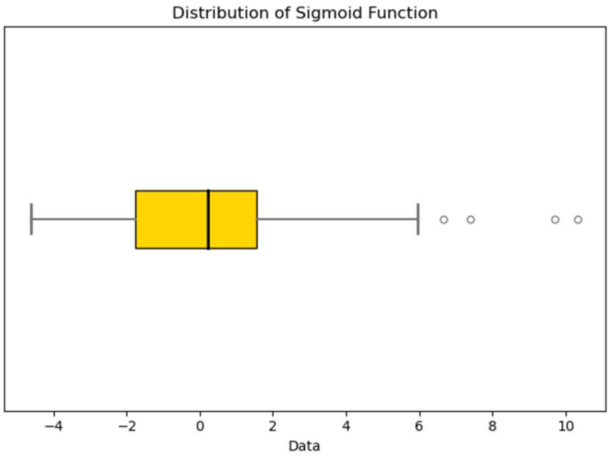
Similarly for primary catchments, the z-score was the same for all regions. This may be due to all regions being near at least one primary school and thus the catchments for all the regions for kindergarten to year 6 are 1. One limitation to this may be that the regions where there are no primary catchments are not accounted for.

For secondary catchments, there was a range of z-scores. However, there were some that were overlapping with each other. For example, the z-scores for secondary catchments for regions Gosford – SpringField, Oyster Bay - Como – Jannali, Sutherland – Kirrawee, Engadine, Loftus – Yarrawarrah and Woronora Heights was the same. One possible explanation for this may be the unequal distribution of high school catchments across regions, with certain regions having a higher number of catchments when compared to others. As a result, similar regions have the same z score. From this, it can be concluded that most regions are "well resourced" in terms of having a high school catchment near them.

	sa2_code	sa2_name	retail_zs	health_zs	primary_zs	secondary_zs	future_zs	stops_zs	polling_zs
100	115021297	Dural - Kenthurst - Wisemans Ferry	1.413146	0.357751	-0.107059	3.289902	-0.121773	6.41374	1.136789

	sa2_code	sa2_name	retail_zs	health_zs	primary_zs	secondary_zs	future_zs	stops_zs	polling_zs
80	128011605	Lilli Pilli - Port Hacking - Dolans Bay	-0.96444	-0.778769	-0.107059	-1.723282	-0.121773	-1.409905	-0.802156

The overall sigmoid rresult for these datasets was greatly wearied across the regions, with some having a sigmoid value as high as 4.2 [Pymble] and others as low as -2.4 [Zetland]. The boxplot "**Distribution of Sigmoid Function"** illustrates this. As seen from the boxplot, it can be concluded that most regions lie within the lower quartile of the plot. From this, it can be concluded that most regions are under resourced in terms of school catchment areas, retail businesses, health care facilities, transportation stops and polling places. The region that had the greatest sigmoid function (10.34) and thus the most well resources was Dural - Kenthurst - Wisemans Ferry and the region that is the most under resourced was Lilli Pilli - Port Hacking - Dolans Bay with a sigmoid value of. -4.61. One of the reasons for this may be due to Dural – Kenthurst – Wisemans Ferry being a combination for three suburbs and has having the most resources. For Lilli Pilli – Port Hacking – Dolans bay, the area can be concluded as an industrial suburb and thus would not have a lot of retail/school catchment resources.

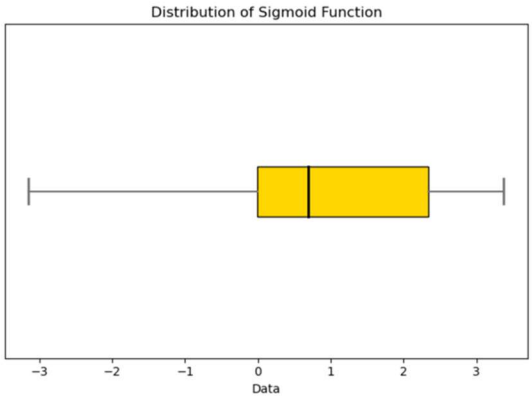


Two additional datasets that were added were crimes and recreations.

For Crimes, the z – scores lied within two standard deviations from the mean. As a result, it can be concluded that the most regions are somewhat well resourced when it comes to crimes and protection. Necessarily not all regions that have a higher z -score for crime mean that the area isn't safe or well protected. For example, those regions that have a higher crime z-score may be well-resourced in terms of safety and protection but due to the general land mass of the region and population, the crime rates might be higher when compared to a region that has the same level of safety and protection.

In terms of Recreation, the z – scores had a greater range when compared to crimes and previous datasets. One of the causing factors for this may be due to some regions have multiple parks, playground, or fitness area due to its population and land area. One of the limitations to this database was that it only covers the City of Sydney/inner Sydney. As a result, when the dataset was joined onto everything else, there was significant amount of data loss. Overall, it can be concluded that most if not all regions part of inner Sydney has some sort of fitness / playground area.

The overall sigmoid result for the given databases and the two extended ones resulted in the following boxplot. As shown, this boxplot significantly differs to the one above. One of the main reasons for this is the datasets combined all together only result in the analysis of inner Sydney. As a result, most of the sigmoid values for these regions illustrate the regions as very well/well resourced.



CORRELATION ANALYSIS

For the purposes of analyzing greater Sydney, only the datasets provided will be used for correlation analysis. The correlation between the median incomes for SA2_regions and how well resourced they can be concluded in the following heatmap. As shown in the heatmap, the median income of a SA2 region does not really play a part in how well-resourced certain areas are. For example, when comparing median income to the retail z scores, there is very littler correlation between the two. Indicating, that a higher median income does not result in the respective SA2 region to have a greater number of retail businesses. Similarly, when comparing median incomes and secondary school catchments, from the heatmap it can be deduced that there is also very little correlation between the two, almost resulting in a score of -0.2. However, this could be an indication that those living in an area with lower median income may not fall within secondary school catchments as those living in an area with a higher median income.