1. DATASET DESCRIPTION

The 9 datasets used were sourced from different websites as listed in the table:

Datasets	Source
Businesses.csv	Australian Bureau of Statistics (https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads)
Stops.txt	Transport for NSW (https://opendata.transport.nsw.gov.au/dataset/timetables-complete-gtfs)
PollingPlaces2019.csv	Australian Electoral Commission (https://data.aurin.org.au/dataset/au-govt-aec-aec-federal-election-polling-places-2019-na)
Catchments.zip	NSW Department of Education (https://data.cese.nsw.gov.au/data/dataset/school-intake-zones-catchment-areas-for-nsw-government-schools)
Population.csv	Australian Bureau of Statistics (https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads)
Income.csv	Australian Bureau of Statistics (https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads)
SA2_Regions.zip	Australian Bureau of Statistics (https://www.abs.gov.au/statistics/economy/business-indicators/counts-australian-businesses-including-entries-and-exits/latest-release#data-downloads)
Internet.json	Aurin.org.au (https://data.aurin.org.au/dataset/tua-phidu-sa2-internetaccessathome-sa2)
Homelessness.csv	Aurin.org.au (https://data.aurin.org.au/dataset/au-govt-abs-sa2-estimating-homelessness-2016-sa2-2016)

After being directly downloaded from Canvas and above websites, all datasets were loaded into our Jupyter notebook and cleaned via Python code. For all datasets, we checked and removed all NaN and duplicated values (if exists). Besides, each dataset required some additional cleaning steps as below:

For the file containing the SA2 Regions zip file, we first removed any rows that were not part of the ‘Greater Sydney’ region and dropped the redundant columns. Then, any rows that did not have any geometry were removed before the spatial data in the “geom” column was converted to WTK.

T**he Businesses.csv file** had data about the total number of businesses in each area based on industry. We filtered the data to ensure that it only contained data for Retail Trade and Health Care and Social Assistance since those are the only two industries we focus on within the scope of this assignment. We then removed duplicated rows and irrelevant columns and ended up with just necessary columns 'industry_code', 'industry_name', 'sa2_code', 'sa2_name' and 'total_businesses' columns.

The Stops.txt file had the coordinates of bus stops in each area. We removed the redundant columns such as 'stop_code', 'location_type', 'parent_station', 'wheelchair_boarding' and 'platform_code' since they were irrelevant in our study, Since this was a spatial dataset we had the geometry datasite converted to WTK so that the geometry could be read and used with postGIS.

The 'PollingPlaces2019.csv' file had the number of polling places in each area. We essentially did the same for the Stops.txt file and removed redundant columns and converted the geometry datasite to WTK.

The Population.csv file contained data regarding the total number of people in each area and grouped them based on age groups. We removed rows where 'total_people' was less than 100, as said in the specifications. We also created a new column ('young_people') which contains the aggregated total number of people between the ages 0-19 which will be later used to calculate the z-score later on.

The Income.csv file contained data regarding the mean and median income in each area, as well as the median age and total number of earners. We cleaned the dataset by transforming the age and income columns to numeric values for easy calculations later. We removed any rows that contained 'np' values and removed any rows that didn't match the sa2_code from the population dataset so that the data we will analyse will be complete.

For the Catchments_primary.shp, catchments_secondary.shp, catchments_future.shp files, as they all share the same attributes, we combined them into one dataset for easier analysis later on. Again we removed all unnecessary columns and converted the geometry datasite to WTK.

The homeless.json file is one of the two files we sourced that came in the form of JSON. We cleaned it the same way we did other datasets (removing NaN values and dropping duplicate values). This file contained data regarding the number of people who have faced homelessness from 2011-2016. This dataset needed no complex cleaning except the basic cleaning that was applied to every dataset.

The Internet.csv file is the other dataset we got that contains data about the total percentage of private dwellings that have Internet access. It had explicit information about the type of internet people were subscribed to (broadband/ dial-ups etc) but we removed the columns and only kept the one with the total percentage. This dataset needed no complex cleaning except the basic cleaning that was applied to every dataset. One limitation we encountered for this dataset was that it only took in consideration of dwellings so it isn't an accurate representation of internet access in each area as it did not include office buildings etc.

Note: As the whole analysis and report is about The Greater Sydney, we decided to use the spatial reference identifier, or SRID = "4283" for all spatial data, which is the Geodetic coordinate system for Australia. According to the source, the SRID of the **'PollingPlaces2019.csv'** file is also "4283".

2. DATABASE DESCRIPTION

Figure 2.1 shows the completed database schema with all columns renamed correctly. Figure 2.1 shows the completed database schema with all columns renamed correctly. In the table diagrams, an underline and bold text denotes a primary key or unique key and an underline and italic text denotes a foreign key that is referenced by the table. The arrows represent the foreign key that is referenced to the primary key being references. For example, sa2_code is a foreign key in *Businesses* that is being referenced in the primary key sa2_code in *SA2_Region*.

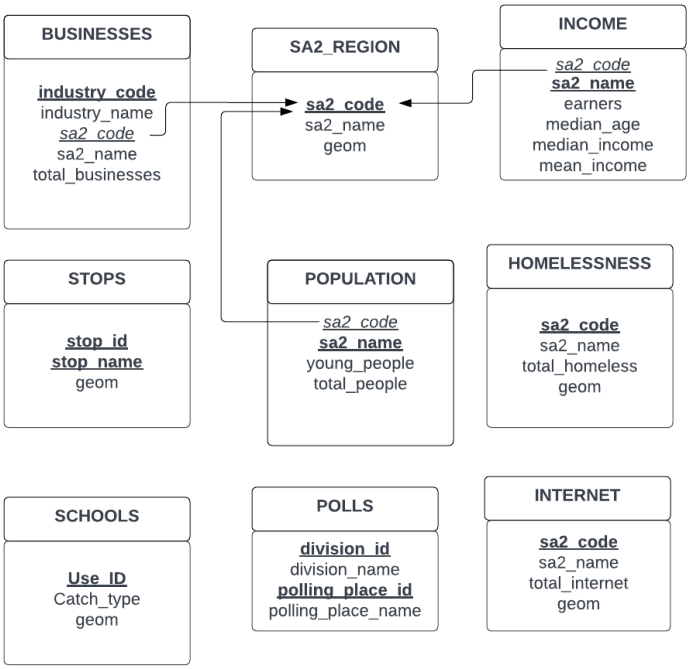

Figure 2.1

After that some indexes were created to enable the SQL servers to efficiently and effectively locate rows in a dataframe and immediately utilise its key values. This eliminates the need to traverse external code or navigate through multiple columns/rows in other dataframes. We chose to base our indexes on variables used frequently, such as the total number of businesses (total_businesses) and the total population (total_people) variables in the schema.

We also used a spatial index on the geom attribute of the stops table since it was a variable we used to join with other variables in other tables. This improves the performance of join operations between the tables so the database engine can efficiently match and retrieve the related rows, resulting in faster query execution.

3. RESULTS ANALYSIS

The resourcefulness per area was computed through a series of SQL queries and Python codes. We first standardised the values for each individual metric (retail, health, stops, polls, and schools) to z-scores for each SA2 region. For stops and polls metrics, we simply calculate the z-score of the sum of that variable in each region (total transportation stops, total polling places). However, regarding health and retail metrics, we used the ratio of total retail businesses and health services per 1000

people to minimise outliers and inflated values. For the schools metric, we used the ratio of total number of school catchments per 1000 young people to provide a more accurate analysis. Then, the "resourcefulness" score of each region is calculated by putting the sum of their previously computed z-scores into the Sigmoid function.

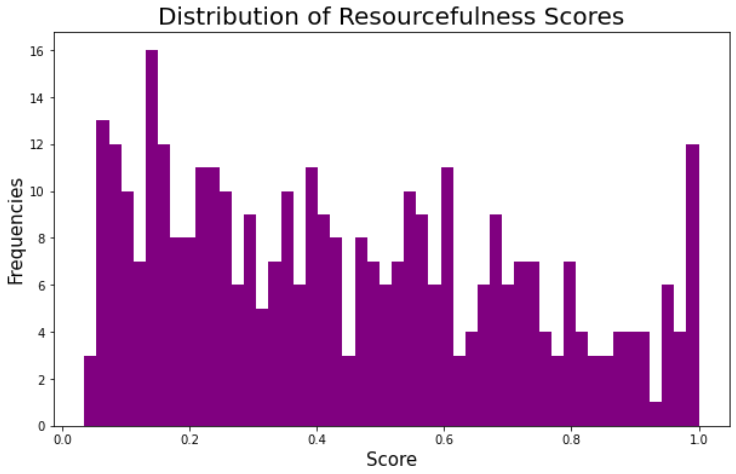

Figure 3.1

Overall, the score distribution can be considered to be relatively right skewed (Figure 3.1). However, there is a noticeably high number of regions with the score of 1. This might be due to the use of ratios when computing the figures, which could result in inflated values, especially in areas with smaller populations. As a result, the mean and standard deviation increase, impacting all standardised z-scores due to the presence of these outliers.

Later, two new datasets were added to extend the score calculation. We used datasets we thought would be relevant to the topic such as the number of people who have experienced homelessness in the span of 4 years and the number of dwellings that have internet access in each area. The z-scores of each respective dataset was then computed and added (deducted the z score of the homeless attribute) to the Sigmoid function for a new score.

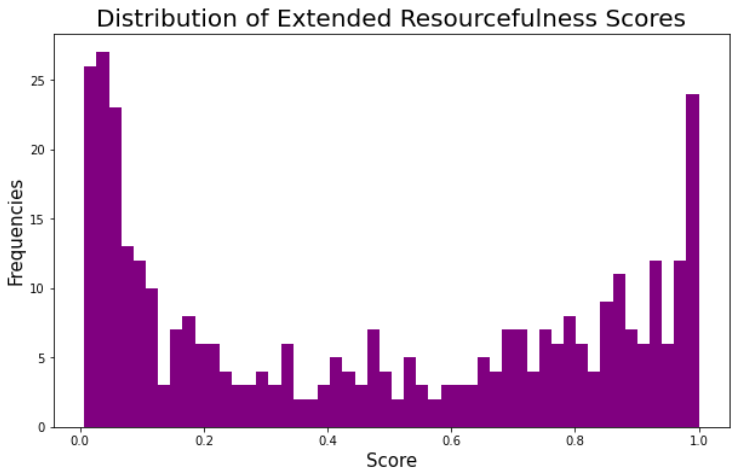

Figure 3.2

As can be seen from Figure 3.2, adding the two new dataset completely changes the score distribution into a U-quadratic distribution. This could be due to extreme values or outliers in the

homelessness and internet accessibility factors that significantly impact the distribution. These outliers can distort the shape of the distribution and lead to a U-quadratic pattern. It's also possible that threshold effects are at play. This happens when certain levels of homelessness or internet accessibility may have a more pronounced effect on the overall z-scores, thus leading to a drastic change in shape of the distribution.

Some noticeable regions with highest/lowest z-scores can be summarised as below:

Z-Score Metric	Highest Region	Lowest Region
Retail	Sydney (North) - Millers Point	Narara
Health	Sydney (North) - Millers Point	Warragamba - Silverdale
Stops	Chippendale Wolli Creek	Rural - Kenthurst - Wisemans Ferry
Schools	Umina - Booker Bay - Patonga	Banksmeadow
Overall	**Sydney (North) - Millers Point**	**Auburn - Central**

With highest z-scores in Retail and Health Metric, it is understandable that Millers Points ends up being the most resourceful region. Auburn - Central, meanwhile, appears to be the least resourceful region based on our metric and calculation.

Using the final new score, the "resourcefulness" score can be visualised on the geographic map of Greater Sydney:

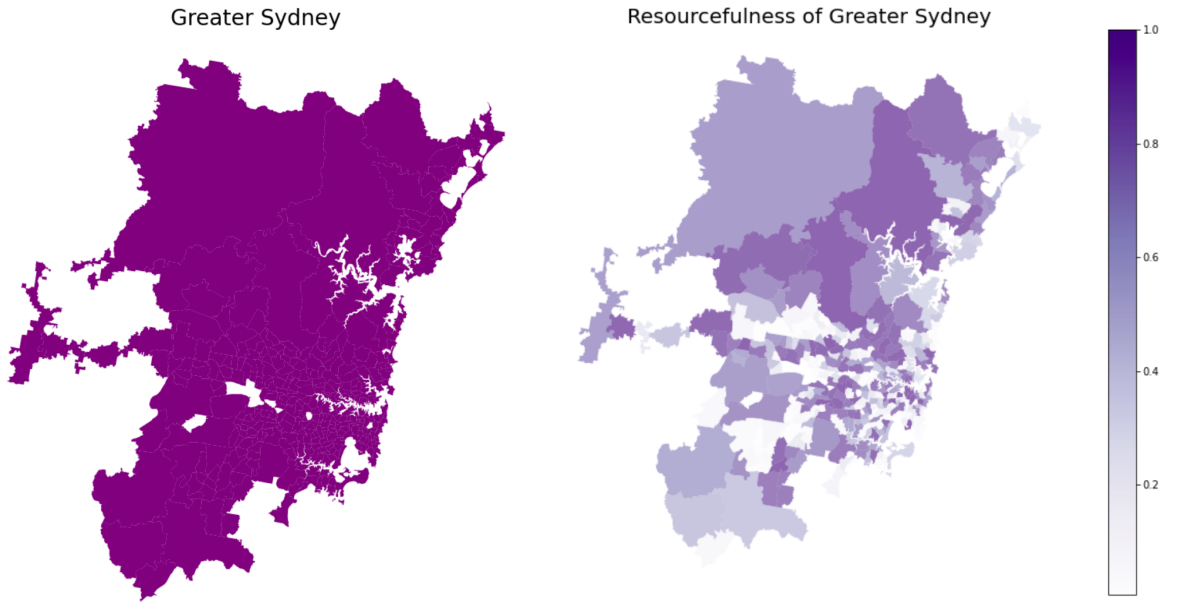

Figure 3.3

The conclusion that our analysis produced was that neighbourhoods in the coastal areas and residential suburbs had higher scores compared to other neighbourhoods, meaning they are generally more well-resourced than other neighbourhoods. This could be due to the fact that these areas are residential and tourist areas where public transport, schools and businesses would be more accessible.

This became one of the limitations we encountered while computing the final score as it did not take into account the size of the neighbourhood. Due to them being mainly a residential/ tourist suburb, they will ultimately have a higher score. Another point to note is that areas that may have high scores for other factors may also have higher homelessness (such as Darlinghurst and Surry Hills). This could be due to the fact that Darlinghurst and Surry Hills are closer to the CBD where despite being more well-resourced, rental prices in the area may be less affordable, leading to a higher rate of homelessness.

4. CORRELATION ANALYSIS

In order to evaluate the correlation of the computed score and median income, the Pearson correlation coefficient between them is firstly calculated which is equal 0.08209167. The value being significantly small (or can be considered approximately zero) indicates that there is a very weak (or no) correlation relationship between the resourcefulness of a region and its median income.

Plotting the income against the score does support the aforementioned comment:

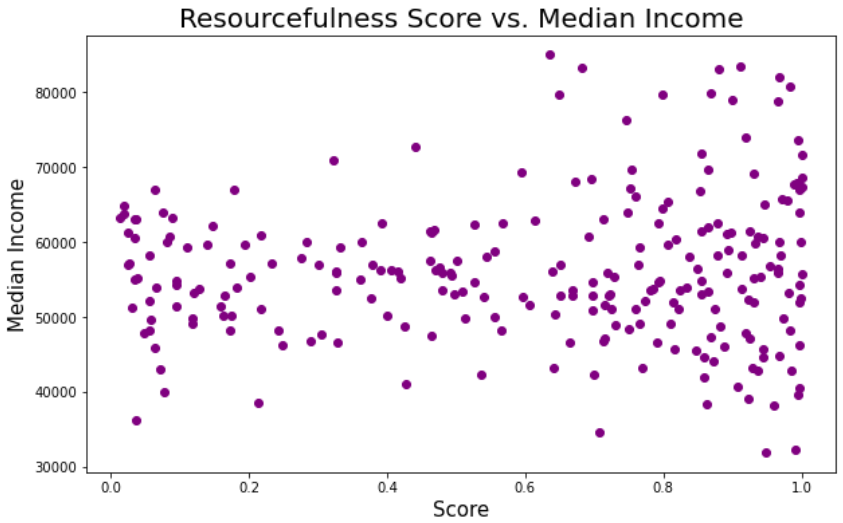

Figure 4.1

Intuitively, the more resourceful a region is, the bigger economic value it has and the higher income its residents tend to receive. However, according to our analysis, although the highest median income regions usually have “resourcefulness” scores higher than 0.6, the data is overall randomly distributed without any notable patterns. This could be due to the fact that people may not always live in the same area they choose to work in. Our computed score is also calculated based on a limited number of factors, which could be why the result was incomprehensive.

In conclusion, within the scope of this analysis, the resourcefulness is not correlated with the median income of that region. This conveys that the resourcefulness of a region cannot be evaluated and concluded based on its median income, and vice versa.