DATA2001: Data Science, Big Data and Data Variety

Practical Assignment Rubric: Greater Sydney Analysis

SID: 510286673 & 520617829

Section 1: Dataset Description

The aim of this task is to calculate how 'well resourced' each sa2 region in Sydney is. This is achieved by calculating the Z score for a number of Resources (different files) and applying the Sigmoid function to the sum of these Z scores for each region. The higher the Z score the greater the number of attributes a region is for a particular resource (e.g retail businesses) leading to a higher Sigmoid function value and thus eventuating in a 'well resourced' SA2 region in Greater Sydney.

The Datasets we used in for this task may be organized as per their use as follows:

Provided Resources datasets to transform to Z scores

Buisnessess.csv (cleaned into two separate datasets for use)

1) Retail Businesses
2) Health Businesses

The Businesses.csv dataset was pre-cleaned and obtained via Canvas. This data gives a breakdown of businesses by industry, turnover, and geographic region. The dataset contains this information for all industries and SA2 regions. The filtering process for the retail and health subsections of this dataset was straightforward and concise as it was done using the Pandas library in Python. It mainly comprised dropping unnecessary columns and filtering out null values that hindered the overall analysis. The steps for obtaining the data for the "Retail" and "Health" metrics as required in the assignment specification were:

1) Retail: After initial data cleaning, the data was further filtered to include only rows where the 'industry_name' column had the value "Retail Trade". The retail metric is an integral part in judging the quality of life and the satisfaction of the people living in it because it shows how well connected a place is and might even show how well the people in the region are employed.
2) Health: After initial data cleaning, the data was further filtered to include only rows where the 'industry_name' column had the value "Health Care and Social Assistance". The health metric is an integral part in judging the quality of life of the people living in it because it shows how easily people can access healthcare and if basic needs are being met.

Stops.txt

This data can be found in the open data section of the transport NSW website, it gives the coordinates of all the public transport stops in NSW. After obtaining the file via the canvas page for data2002 it was read into jupyter notebook via the pandas extension of python, where it was read in as a pandas dataframe. Since it was a text file with no geometric attribute the longitude and latitude columns were then combined and converted to a well known text format, the Spatial Reference Identifier (SRID) - in this case 4326, to represent the WGS84 world geodetic coordinate system. In other words a set of point coordinates were created for each row in the data set. This can then be converted into a geometric object when the data is imported in postgreSQL.

Before this crucial step the data was cleaned; columns not necessary for Z score calculations were dropped, if NA values existed in necessary columns the row was dropped, and all duplicate rows were dropped.

PollingPlaces2019.csv

This file may be found on the aurin website, and is a dataset provided by the AEC (Australian Electoral Commision), it gives the coordinates of all the polling stations in NSW. Sourced to us through canvas this csv file was read into jupyter as a pandas dataframe and the longitude a latitude was converted into a spatial object before the cleaned geospatial file was imported into postgreSQL. The data was cleaned by removing duplicate rows, checking for null values, removing null values and dropping rows unnecessary for calculating the Z score.

Catchments.zip, (future, primary and secondary files were merged into a single all inclusive file)

This dataset is provided by the NSW government on their website under the Department of Education. Obtained via the canvas page for data2002, this shape file needed to be unzipped and read into jupyter as three separate geo dataframes using the geopandas package (future, primary and secondary files). Each of these files contains the schools and the coordinates of the multipolygon that bounds their catchment area.
The three geodata frames were cleaned as follows. Columns irrelevant to the Z Score calculation, and duplicate rows were then dropped. Then, the geometric object was turned into consistent multigons and the coordinates were converted into the well known universal form.
Finally, the geodata frames were merged and then cleaned so that NA values were checked for duplicate rows were removed before this clean, concise, merged geo dataframe was imported into postgreSQL.

Datasets found by group members to be used for Z score calculations

Homeownership.csv

This dataset was found on the data.aurin.org.au, it is a csv file detailing how many homeowners there are per sa2 region, the geometric multipolygon that bounds each region, and also higher geographic boundaries including Greater Capital region. This is important as it allows for filtering to just the Greater Sydney region in later queries. This dataset has been cleaned such that the unnecessary columns were dropped, null values were dropped, and columns were renamed. The data was then imported to postgreSQL ready for Z Score calculations.

datasource-UNSW_CFRC-UNSW_CFRC_GEONODE_geonode_RentalHouseholds_2016.json

1) This section was done by SID 520617829 this data is a json file obtained via data.aurin.org.au . This dataset provides statistics about renters (number and z-score) for different geographic regions in Australia, identified by SA2 codes and names. The geometry column allows mapping the regions spatially and the file type is different from the datasets that were provided.
2) The major cleaning was done using geopandas and the geometry column was converted to the Well-Known Text (WKT) format, as an intermediate step using the geoalchemy2 library, to convert from the types in GeoPandas to the WKT format in PostGIS with the Spatial Reference Identifier (SRID) - in this case being 4326. For data cleaning the unnecessary columns were dropped, null values in the geometry column were removed and the z score of total renters by sa2 region was calculated.
3) The reason this dataset was chosen is because the number of renters in an sa2 region signifies how affordable housing is in a specific region and that is a major indicator of the quality of life and satisfaction.

Datasets used to organize, index, calculate and join Datasets

- Population.csv

This data can be found on the ABS (Australian Bureau of Statistics Website) but was retrieved from the data2002 canvas page. It was read into jupyter as a dataframe and then split and additional dataframe, a subset of the population dataframe was created: ‘Youngpeople’. This file contains just the population of those aged 0-19 per sa2 region whilst the population data frame includes the entire population of each sa2 region. Both datasets were cleaned by removing duplicate rows, dropping unnecessary columns, and checking for null values.

- SA2regions.zip

SA2 region data sourced from the ABS (Australian Bureau of Statistics Website) in the form of a shape file contains the SA2 codes and geometric boundaries for SA2 regions across australia. This data was read into Jupyter as a geo dataframe. This was then cleaned by checking for NA values and dropping any, removing duplicates, and converting all spatial objects to multipolygons for consistency. Finally the columns unnecessary for the Z Score calculation were dropped and some columns were renamed for similarity. Finally the data was imported into postgreSQL

- Income

The Income data was sourced from the ABS (Australian Bureau of Statistics Website) it was a csv file in tabular form. To calculate correlation all unnecessary columns were dropped using pandas that was all the cleaning required after

which the Scores of all metrics that were calculated were merged with the median income column using the column containing sa2 region as a key.

Section 2: Database Description

The Schema was established by defining the db_schema as the new schema 'Sydney Regions', and then executing a CREATE SCHEMA… statement in SQL and setting the searchpath to this schema. Then, we sequentially cleaned each file and once the desired table had been created we added this table to the schema using a CREATE TABLE statement. (this part is no longer included in the code in Jupyter as we had a problem with postGIS in pgadmin meaning we could only import spatial data under the public schema…):

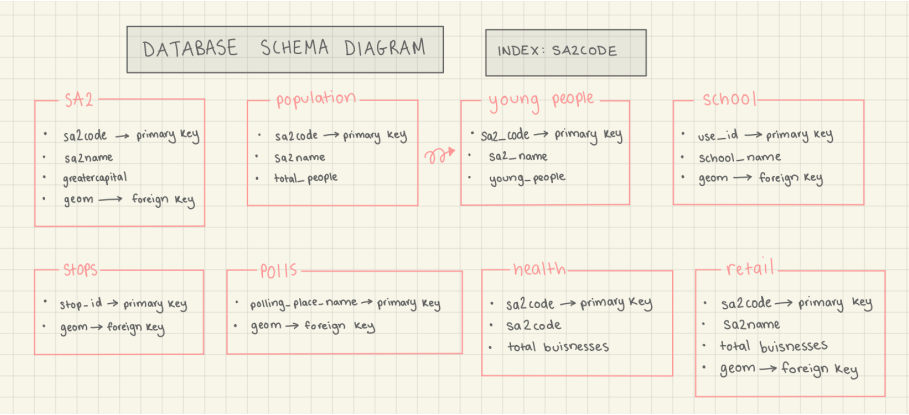

Fig 1.0: Database Schema for provided datasets

Section 3: Results Analysis

We used the Sigmoid Function to score each SA2 region in Greater Sydney: 1 / (1 + EXP(-x)) where x is the sum of all Z score values for a given SA2 region. These Z scores were calculated for each of the 5 datasets in the following ways:

Polls:

First, the data must be prepared to create the Z score. The total polling places per region SA2 is found through using the COUNT(*) function when the polls and SA2 region data is joined using the ST_CONTAINS() function and their common 'geom' keys which are of the same coordinate system. Thus all polling stations (points) can be grouped into their corresponding SA2 region (multipolygon). A condition clause is also included to filter out any regions not in greater sydney.. The group by statement groups the data by SA2 region and the ST_AREA() function is used to find the total area per sq km for each SA2 region. Finally, a new column is added, 'polling places per sqkm' which uses the formula below to calculate how many polling places there are per sq km of each sa2 region.

Now, a Z score can be calculated for each SA2 region using the following formula:

Z Score = (Polling Places per sq km - Mean) / Standard Deviation;

*Mean = mean of all polling places per sq km**

*Standard Deviation = standard deviation of all polling places per sq km**

**Found using numpy as this simplified the SQL query dramatically.*

Stops

First, the data must be prepared to create the Z score. The number of stops per region SA2 is found through using the COUNT(*) function when the stops and SA2 region data are joined using the ST_CONTAINS() function and their common 'geom' keys which are of the same coordinate system. Thus all public transport stops (points) can be grouped into their corresponding SA2 region (multipolygon). A condition clause is also included to filter out any regions not in greater sydney. The group by statement groups the data by SA2 region and the ST_AREA() function is used to find the

total area per sq km for each SA2 region. Finally, a new column is added, ‘stops per sqkm’ which uses the formula below to calculate how many stops there are per sq km of each sa2 region.
Now, a Z score can be calculated for each SA2 region using the following formula:

Z Score = (Stopsper sq km - Mean) / Standard Deviation;

Health

First, the data must be prepared to create the Z score. The number of health businesses per region SA2 is found through using the COUNT(*) function when the health and SA2 region data are joined on their common primary key ‘sa2code’. A condition clause is also included to filter out any regions not in greater Sydney and any non health businesses. The group by statement groups the data by SA2 region. This forms a new table ready to be then merged with the population dataset.
The SA2 and health data is joined with the population data on the primary key ‘sa2code’ and a column calculating the number of health businesses there are per 1000 people in the population for each SA2 region is added. A condition is also added using HAVING so that the total_people is greater than 0 since this will be our denominator when calculating the businesses per 1000 people and a denominator cannot be 0.
Now, a Z score can be calculated for each SA2 region using the following formula:

Z Score = (Health Businesses per 1000 people - Mean) / Standard Deviation;

Retail

First, the data must be prepared to create the Z score. The number of retail businesses per region SA2 is found through using the COUNT(*) function when the retail and SA2 region data are joined on their common primary key ‘sa2code’. A condition clause is also included to filter out any regions not in greater Sydney and any non retail businesses. The group by statement groups the data by SA2 region. This forms a new table ready to be then merged with the population dataset.
The SA2 and Retail data is joined with the population data on the primary key ‘sa2code’ and a column calculating the number of retail businesses there are per 1000 people in the population for each SA2 region is added. A condition is also added using HAVING so that the total_people is greater than 0 since this will be our denominator when calculating the businesses per 1000 people and a denominator cannot be 0.
Now, a Z score can be calculated for each SA2 region using the following formula:

Z Score = (Retail Businesses per 1000 people - Mean) / Standard Deviation;

Schools

First, the data must be prepared to create the Z score. The number of school catchment areas that intersect part of or all of an SA2 is found through using the COUNT(*) function when the schools and SA2 region data is joined using the ST_INTERSECT() function and their common ‘geom’ keys which are of the same coordinate system. Thus all school catchment areas (multipolygon) can be grouped into the corresponding SA2 region they intersect (multipolygon). A condition clause is also included to filter out any regions not in greater sydney. The group by statement groups the data by SA2 region. This forms a new table ready to be then merged with the population dataset.
The SA2 and schools data is joined with the young people data on the primary key ‘sa2code’ and a column calculating the number of school catchment areas there are per 1000 young people in the population for each SA2 region is added. A condition is also added using HAVING so that the total_people is greater than 0 since this will be our denominator when calculating the businesses per 1000 people and a denominator cannot be 0.
Now, a Z score can be calculated for each SA2 region using the following formula:

Z Score = (Schools per 1000 young people - Mean) / Standard Deviation;

For our own datasets the following methods were used:

Homeowners

The owners per 1000 people was calculated by joining the population with the newly introduced homewners dataset using the primary key ‘sa2code’. The Z score was then calculated in the same way as for the number or retail business or health facilities above;

Z Score = (Homeowners per 1000 young people - Mean) / Standard Deviation;

Renters (“rent”)

For calculating the Z-scores of the number of renters per sa2 region first the mean and standard deviation of the 'Renters' column were calculated using the mean() and std() functions, respectively. The mean is stored in the variable renters_mean, and the standard deviation is stored in the variable renters_std then the the z-score for each value in the **'Renters'** column was calculated using the formula **(value - mean) / standard deviation**. This formula standardizes each value by subtracting the mean and dividing by the standard deviation and then the resulting z-scores were stored in a new column called **'Renters_zscore'**

Merging Together and Finding Sigmoid function value:

Both new datasets and the original merged z score table were joined using the primary key ‘sa2 code’ and the Sigmoid function was applied using the exact same method described above for this newly merged table.

Overview of results

We created an overlay map (figure 1) to create a visual summary of the results. In summary, SA2 regions in the inner city have higher Sigmoid function values (closer to one; closer to yellow) indicating they are better resourced with the factors we used to calculate this value: health facilities, retail venues, public transport, polling stations, school catchment.

This map aligns with my knowledge of Sydney regions. The inner city, although densely populated, has very small SA2 areas and tends to be 'richer', thus resources such as those listed are more accessible giving these areas a higher Sigmoid function value and more yellow color on the map. Alternatively as Sydney spreads to the north east and west the area size of SA2 regions increase meaning there would be lower Z scores for polling places and stops per sqkm. Moreover these suburbs are not as well off so not as much government focus is placed on providing these regions with facilities.

HOWEVER, there are some obvious outliers to this explanation, those high sigmoid scoring functions in Sydney's North West. This is because these regions are mostly bushland and national parks and therefore have much lower population, thus the number of health facilities, retail venues, and school catchment areas score much higher as these are measured per 1000 people (or young people) in the population.

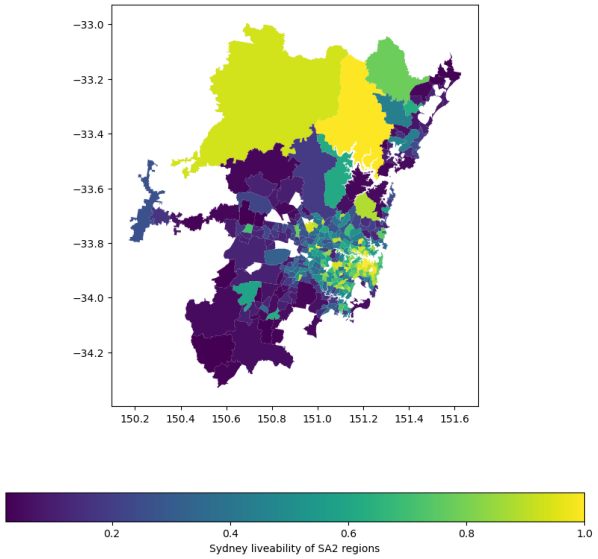

Fig 2.0: Overlap Map showing distribution of Sigmoid Function values based on Z scores calculated for the 5 provided datasets.

Section 4: Correlation Analysis

There is a positive correlation between median income of each SA2 region and score but it also has many outliers and the points are spread out as seen in the figure, the line of best fit shows that with higher 'Scores' the median income of a specific region increases (it was added for seeing the correlation more clearly using Sklearn). The column 'median_income' from the Income dataset and the 'Score' column after it was calculated was merged into one csv file using sa2_regions as a key. The data in this plot was rescaled because there was a huge difference in the numerical value between the median income and score, so simply plotting those two columns would make the correlation unclear. Some reasons due to which the correlation between the above mentioned attributes is positive could be:

1) The scores were calculated using Z Scores for different 'Metrics' according to the following formula: $\text{Score} = S(z_{retail} + z_{health} + z_{stops} + z_{polls} + z_{schools})$.With better healthcare and access to different retail facilities along with better transportation, and academic facilities an individual has better chances of obtaining and retaining a job that pays well.
2) With access to polling facilities more individuals would feel like they have a say in how their area is governed making them feel important and heard which helps a person become more productive in general.
3) Overall the scores seem to be a fairly accurate measure of how good the quality of life is in a specific region which is why there might be a positive correlation between the two attributes.

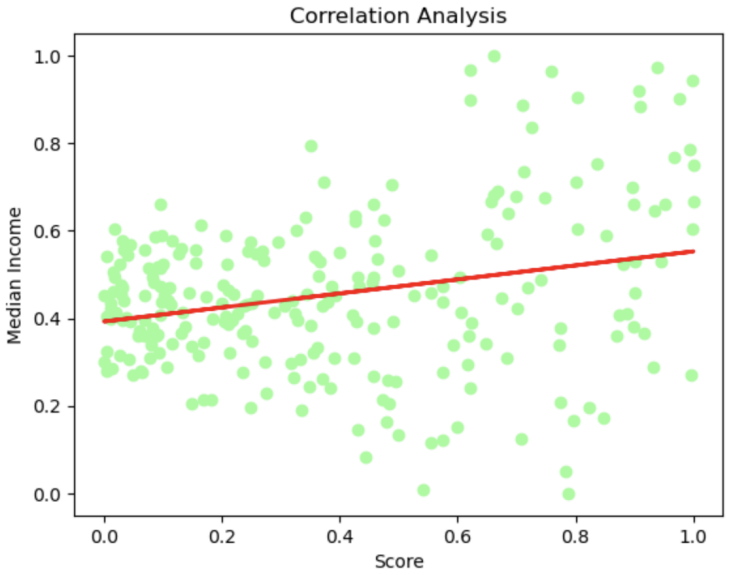

Fig 3.0: Correlation between Score of different metrics and Median Income

References

1) *Welcome - AURIN - Data Catalogue. (n.d.). Data.aurin.org.au. Retrieved May 24, 2023, from https://data.aurin.org.au*
2) *DATA2001. (n.d.). Week9_Tutorial_Solutions [Review of Week9_Tutorial_Solutions]. The University of Sydney.*
3) *Connecting postgresql with sqlalchemy. (n.d.). Stack Overflow. Retrieved May 24, 2023, from https://stackoverflow.com/questions/9353822/connecting-postgr*
4) *matplotlib.pyplot.scatter() in Python. (2020, March 26). GeeksforGeeks. https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/*
5) *NSW Department of Education | education.nsw.gov.au. (2019, August 6). Nsw.gov.au. https://education.nsw.gov.au*
6) *TfNSW Open Data Hub and Developer Portal. (n.d.). Opendata.transport.nsw.gov.au. https://opendata.transport.nsw.gov.au*

7) *Australian Bureau of Statistics. (2022). Australian Bureau of Statistics, Australian Government. Abs.gov.au. https://www.abs.gov.au*