Data Analysis

(Group name: Lab-27-group 3)

Guohan Fu 520034369

Ryan He 500178108

Introduction: Greater Sydney is Australia's largest urban area, encompassing Sydney as w ell as the surrounding region. It is densely populated and economically prosperous. Investigating and analyzing this area will provide a better understanding of the problems and future development of the area. Greater Sydney contains more than 350 SA2 areas, which are geographical areas of Sydney that represent social and economic neighborhoods and typically contain between 3,000 and 25,000 people. This assignment used datasets in a variety of formats to facilitate the processing of GIS data, statistical data, and other data sources used to develop a comprehensive view of each SA2. To facilitate the assessment of the level of resources available in each region and the calculation of the corresponding scores.

Dataset Description

SA2.zip: This dataset is from the Australian Bureau of Statistics (ABS), download on Canvas. It provides information on Statistical Area Level 2 (SA2) digital boundaries. It includes SA3 and SA4, local government (GCC) and geographic location and region, which can be used for data statistics and analysis.

Business.csv: This dataset is from ABS, download on Canvas. Which contains the number of businesses in industry and in the SA2 region, as well as a range of reports by turnover size.

Stops.txt: This dataset is from ABS, download on Canvas. Contains all public transport stops (trains and buses) in General Transport Feed Specification (GTFS) format.

Income.csv: This dataset is from ABS, download on Canvas. Which is a dataset of total revenue based on SA2 statistics for later correlation analyses.

Population.csv: This dataset is from ABS, download on Canvas. Indicates the estimated number of people living in each SA2 area, by age. "Per capita" is used in the calculations.

Pollingplaces2019.csv: This dataset is from ABS, download on Canvas. Contains the location of polling stations for the 2019 federal election, as well as details of other venues.

Catchments.zip: This dataset is from ABS, download on Canvas. Areas where students must live to attend primary, secondary and future government schools.

Dog-off-leash-parks.geojson: This dataset is from “City of Sydney”. Download on: https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::dog-off-leash-parks/about

This dataset is spatial data and provides a list of off-leash dog parks in the city of Sydney, including park names, streets, etc.

Mobility-praking.geojson: This dataset is from "City of Sydney". Download on:

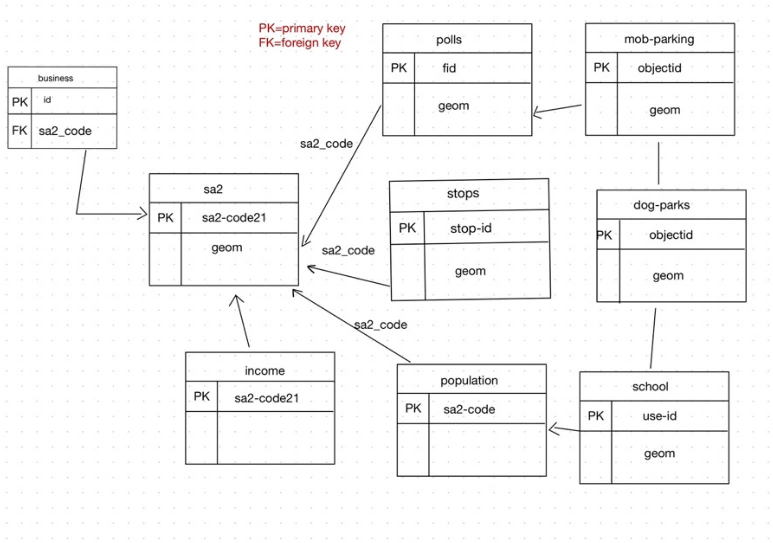
https://data.cityofsydney.nsw.gov.au/datasets/cityofsydney::mobility-parking-3/about This dataset is also spatial data, It provides information on where you can

park for free in Sydney for a certain period of time (if you have a Mobile Parking Scheme permit).

Pre-processing the data:

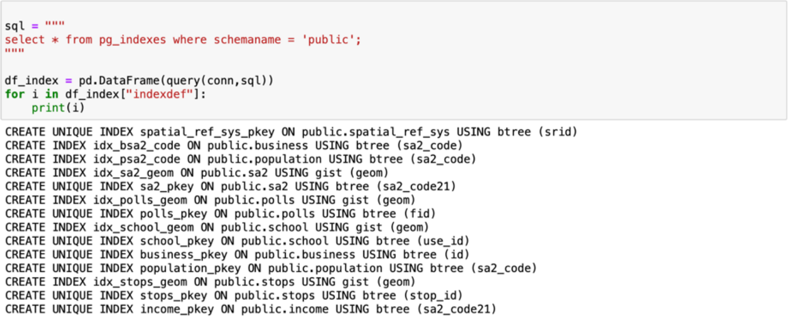
To be able to use the data more efficiently, a series of processes should be carried out before the data is analyzed. Therefore, our group believes that it is very important to pre-process the data. The database management system's query language, PostgreSQL, will be used in the Jupyter Notebook, so it's important to keep links between PostgreSQL and the python. First, we need to import the "psycopg2" package, define the "query" and "pgconnect" functions and connect to the school's postgreSQL server. The purpose is to allow python to execute SQL statements in python code. After completing the above steps, you will find that there are some missing values in each data, so we need to perform data cleaning. The purpose is to ensure the consistency of the evaluation data. Our group tried to use "dropna ()" to remove the missing values, but since there was too much data to remove, especially in the stops and polls datasets, which would have greatly affected the validity of the dataset, we gave up on using "dropna()" in the end. Then we began creating tables for each database, as well as primary key and foreign keys and variable types for each database.

Database Description:



As you can see, three of the tables are connected to each other by "sa2-code". In the sa2 table, 'sa2-code21' is the primary key, while in other tables as a foreign key is called "sa2-code", but it is essentially the same, which is why the tables can be interconnected.

Indies: To make it easier to query, we added some indexes to the database：



```
sql = """
select * from pg_indexes where schemaname = 'public';
"""

df_index = pd.DataFrame(query(conn,sql))
for i in df_index["indexdef"]:
    print(i)
```

```
CREATE UNIQUE INDEX spatial_ref_sys_pkey ON public.spatial_ref_sys USING btree (srid)
CREATE INDEX idx_bsa2_code ON public.business USING btree (sa2_code)
CREATE INDEX idx_psa2_code ON public.population USING btree (sa2_code)
CREATE INDEX idx_sa2_geom ON public.sa2 USING gist (geom)
CREATE UNIQUE INDEX sa2_pkey ON public.sa2 USING btree (sa2_code21)
CREATE INDEX idx_polls_geom ON public.polls USING gist (geom)
CREATE UNIQUE INDEX polls_pkey ON public.polls USING btree (fid)
CREATE INDEX idx_school_geom ON public.school USING gist (geom)
CREATE UNIQUE INDEX school_pkey ON public.school USING btree (use_id)
CREATE UNIQUE INDEX business_pkey ON public.business USING btree (id)
CREATE UNIQUE INDEX population_pkey ON public.population USING btree (sa2_code)
CREATE INDEX idx_stops_geom ON public.stops USING gist (geom)
CREATE UNIQUE INDEX stops_pkey ON public.stops USING btree (stop_id)
CREATE UNIQUE INDEX income_pkey ON public.income USING btree (sa2_code21)
```

We first detected the indexes that already existed, then created new ones as needed, and then detected them again to make sure all the indexes we needed existed. Obviously indexes can speed up our queries and help us improve efficiency. But at the same time, they also have disadvantages, index needs to take up a lot of storage space, in each import and update data, index also need to be updated again.

Score analysis:

The computation of score will come in task2 and task3, we will first compute the z-score of each table and then add them together to finally compute the overall score and the sigmoid function. also, we will compute the correlation. Formula：

Task2 score = S (*z*business + *z*stops + *z*polls + *z*school)

Task3 score = S (*z*business + *z*stops + *z*polls + *z*school + *z*dog + *z*moblility)

The formula for sigmoid function is $\sigma(x)=\frac{1}{1+e^{-x}}$ ， $z=\frac{x-\mu}{\sigma}$

Calculations: The *z*score for Sydney (North)-Millers Point is higher than other areas, the score is 1.0. Which means that Sydney (North)-Millers Point is more prosperous than other areas. A quick check of the relevant information also reveals that Sydney (North)-Millers Point is the second largest concentration of office buildings in NSW, with many large companies choosing to set up offices in the area. Sydney (North)-Millers Point has a full range of shopping centers and modern business centers. It also has many dining establishments. People living here don't need to go to the city center of Sydney, they can enjoy good needs if they are close to home. On the other hand, we believe that having good healthcare facilities and easy access to public transport are also important reasons for Sydney (North)-Millers Point's top ranking. Sydney (North)-Millers Point has a large train station including Chatswood, which is a very short drive from the city center. This makes it easy for residents living here to get to the city center. Wolli Creek is the lowest scoring area in the entire sa2 region, with a score of only 0.065203. By checking the relevant information, it can be found that

Wolli Creek is geographically far away from the city center, as it is in the southernmost part of Sydney, which is known as a suburb. Also, by looking at the map I realized that Wolli Creek is next to the Sydney Airport, which often makes a lot of noise, which means that the residents' lives are also affected (This was true even after task3 introduced a new dataset, the score of Sydney (North)-Millers Point is still 1.0; the score of Wolli Creek is 0.047396).

Code analysis: Business: retail connect business and population to sa2 by left join (a2_code21 = b.sa2_code) and ('s.sa2_code21 = p.sa2_code'), firstly, use the "CAST" function to convert the data type: "b.total_businesses AS NUMERIC) * 1000 / p.total_people". At the same time we need to qualify some conditions, b.industry_name = 'Retail Trade' and "p.total_people >= 100". Next step, we need use two functions AVG () and STDDEV (). Then the *z*_score was calculated by the formula in SQL.

Stops: firstly need connect stops to sa2 by left join on "ST_Contains(sa2.geom, stops.geom)", Next step, we also need use two functions "AVG()"and "STDDEV()". We need to treat '(number_of_stops - average) / standard_deviation' as z-score. Then the *z*_score was calculated by the formula in SQL.

Polls: connect polls to sa2 by left join ON ST_Contains(sa2.geom, polls.geom), the aim was to find the number of polling stations in each region. Next step, we also need use two fuctions AVG () and STDDEV (). Then the *z*_score was calculated by the formula in SQL.

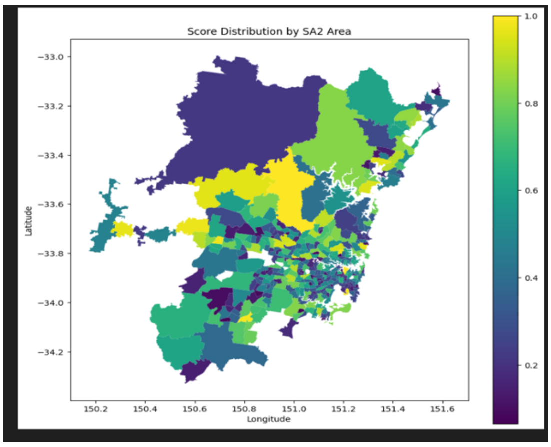
School: firstly, connect school to sa2, the aim to find the corresponding the sa2-name 21, left join school 'ON ST_Intersects(sa2.geom, school.geom)'; left join population p "ON sa2.sa2_code21 = p.sa2_code".select "p.people_0_4 + p.people_5_9 + p.people_10_14 + p.people_15_19" as young_people, use SUM() fuction as area. Then we also need use two fuctions AVG (area * 1000 / young_people) AS average and STDDEV (area * 1000 / young_people) AS standard_deviation. Then the *z*_score was calculated by the formula in SQL.

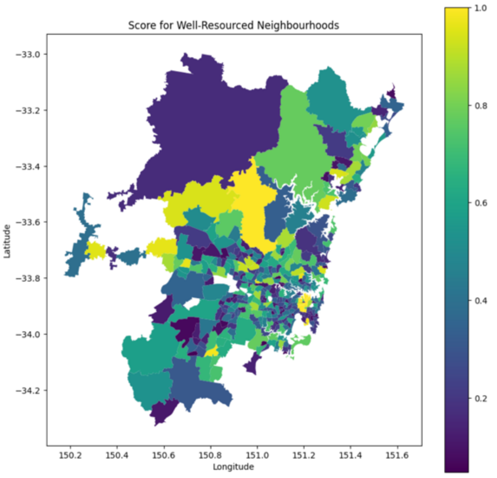
Task3: In task3, there are two new datasets in "geojson" format.

dog_park_num：connect dog_park to sa2, use LEFT JOIN "dog_parks d ON

ST_Contains(sa2.geom, d.geom)". next step is SELECT AVG (number_of_dog_parks) AS average; STDDEV (number_of_dog_parks) AS standard_deviation. In this case, we also need use two fuctions AVG () and STDDEV (). Then the *z*_score was calculated by the formula in SQL.

mob_parking: connect mob_parking to sa2, use LEFT JOIN mob_parking m ON ST_Contains(sa2.geom, m.geom), then we use two fuctions; 'AVG(number_of_mob_parking) AS average, STDDEV(number_of_mob_parking) AS standard_deviation'. Then the *z*_score was calculated by the formula in SQL.





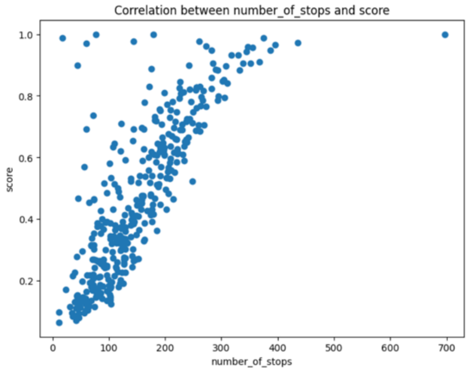
(map for task2) (map for task3)

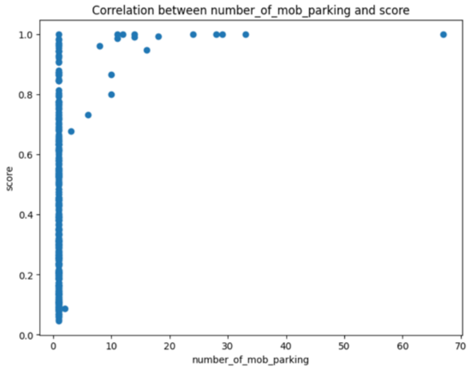
As can be seen from the indications on the far right of each graph, lighter yellow indicates a higher zscore; darker purple indicates a lower zscore. In task 2, Sydney (North)-Millers Point is the highest area, with a score of 1.0, while the lowest area is Wolli Creek, with a score of 0.065203; in task 3, Sydney (North)-Millers Point is still the highest area, with a score of 1.0, while Wolli Creek as the lowest area with a score of 0.047396.

Correlation analysis:

To investigate whether there was a correlation between the two variables in this task, our group therefore compared the number of stops in the Greater Sydney area with the mobility parking in the newly introduced dataset in task3.

The correlation coefficient of Task2: 0.7984070963935572
The correlation coefficient of Task3: 0.33856588327207154





The results show that they do correlate as the correlation coefficient is a positive value. After discussion, we believe that the reason may be that Sydney (North)-Millers Point is a very economically privileged area with a high standard of living, and therefore people living here will have a higher need for mobile parking than for bus stops, which is why there are fewer bus stops in the Sydney (North)-Millers Point are fewer in number. So, we think that when people are better off they will have private cars instead of public transport. Most Wolli Creek's residents are students and workers, which is why they have a greater need for transit.

We think that there may be limitations in the results. The new dataset imported into task3 is small compared to the other datasets, which could also lead to extreme values that could affect the results, and we applied data cleaning to remove missing values prior to the calculation, but this is also likely to affect the results.

Conclusion：

The result of this assignment is that Sydney (North)-Millers Point is a great place to live in the Greater Sydney area, by a combination of business, public transport, polling stations and schools. It has a thriving economy and is a good place to live. Wolli Creek was the lowest scoring area of the assignment, being in the southernmost part of Sydney, far from the city center while being right next to the airport, and therefore comes with noise issues and so on. Of course, there are undeniable limitations to this assignment as real data is never perfect. For example, the area of land in Sydney is very large, so some of the data may not be accurate. Also, the data itself is incomplete, containing NULL values and empty columns, The presence of a large number of missing values in the database may affect the accuracy of the results. So, our group believes that the data and results do not total match the reality of Greater Sydney.

Reference list:

City of sydney data hub. City of Sydney Data hub. (n.d.). https://data.cityofsydney.nsw.gov.au/

Map of North Sydney, NSW 2060. Map of North Sydney, NSW 2060 | Whereis®. (n.d.). https://www.whereis.com/nsw/north-sydney-2060

Map of Wolli Creek, NSW 2205. Map of Wolli Creek, NSW 2205 | Whereis®. (n.d.). https://www.whereis.com/nsw/wolli-creek-2205

Wikimedia Foundation. (2024b, April 22). *Wolli Creek, New South Wales*. Wikipedia. https://en.wikipedia.org/wiki/Wolli_Creek,_New_South_Wales

Wikimedia Foundation. (2024, March 4). *North Sydney, New South Wales*. Wikipedia. https://en.wikipedia.org/wiki/North_Sydney,_New_South_Wales