Report

**Dataset description 1: Data sources and preprocessing steps**

**SA2 data:**

Source: SA2\_2021\_AUST\_SHP\_GDA2020/SA2\_2021\_AUST\_GDA2020.shp

Preprocessing: Filter out the data where GCC\_NAME21 is Greater Sydney and convert the geometric data to WKT format.

Business data:

**Businesses**

Source: Businesses.csv

Preprocessing: Remove missing values and convert column names to lowercase.

Demographic data:

**Population**

Source: Population.csv

Preprocessing: Remove missing values and convert column names to lowercase.

Revenue data:

**Income**

Source: Income.csv

Preprocessing: Process "np" values and convert column names to lowercase.

Bus stop data:

**Stops**

Source: Stops.txt

Preprocessing: Create point geometry data and convert to WKT format.

Polling location data:

**Polling**

Source: PollingPlaces2019.csv

Preprocessing: remove missing latitude and longitude data, create point geometry data and convert to WKT format.

School district data:

Source: catchments/catchments\_primary.shp, catchments/catchments\_future.shp, catchments/catchments\_secondary.shp

Preprocessing: remove missing geometry data, convert polygons to polygons, and convert to WKT format.

**Bike Lockers and Sheds Data**

Source: bike\_lockers\_sheds.csv

Description: Locations of bike lockers and sheds.

Pre-processing: Converted coordinates to WKT format.

**Petrol Stations**

Source: Petrol\_Stations.geojson

Description: Locations of petrol stations.

Pre-processing: Converted geometries to WKT format.

**Traffic Lights**

Source: traffic\_lights\_location.csv

Description: Locations of traffic lights.

Pre-processing: Converted coordinates to WKT format.

**Dataset description 2: Database schema establishment and data integration**

**Database schema**: Create each table (sa2, businesses, population, income, stops, poll, catchments\_primary, catchments\_future, catchments\_secondary, school\_catchments) through SQL commands, and define the type of each field.

**Data integration**: Use the to\_sql function to insert the preprocessed data frame into the corresponding table.

**Index creation**:

**Sa2**:

**idx\_sa2\_code21**: SA2\_CODE21 is the main identifying field of the sa2 table and is often used in queries for joins or filtering. Creating this index can speed up these operations.

**idx\_gcc\_code21**: The GCC\_CODE21 field is used for geographic or region related queries and filtering. Indexes help improve query speed.

**idx\_ste\_code21**: The STE\_CODE21 field is used for state-level queries and analysis, and the index can optimize these operations.

**Businesses**:

**idx\_businesses\_sa2\_code**: sa2\_code is the key field for connecting business data to the sa2 table, and the index can speed up the connection operation.

**idx\_industry\_code**: industry\_code is used for query and filtering of industry categories, and the index improves query efficiency.

**Population**:

**idx\_population\_sa2\_code**: sa2\_code is a key field for connecting population data to the sa2 table. The index can speed up connection and query operations.

**Income:**

**idx\_income\_sa2\_code21**: sa2\_code21 is the key field for connecting income data to the sa2 table, and the index can speed up connection and query operations.

**Stops:**

**idx\_stops\_stop\_id:** stop\_id is a unique identifier of a bus stop and is often used for query and connection operations. Indexing can improve efficiency.

**idx\_stops\_location\_type**: The location\_type field is used to filter different types of sites, and the index helps speed up queries.

**Polling:**

**idx\_poll\_division\_id:** division\_id is used to query the constituency to which the voting location belongs. The index can speed up the query.

**idx\_poll\_state**: The state field is used for state-level query and filtering, and the index can improve query efficiency.

**Catchments**

**idx\_catchments\_primary\_use\_id,idx\_catchments\_future\_use\_id** and **idx\_catchments\_secondary\_use\_id:** USE\_ID is the main identifier of school district data and is often used for connection and query operations. Indexing can improve efficiency.

**idx\_school\_catchments\_use\_id**: USE\_ID is the main identifier of school district data and is often used for connection and query operations. Indexing can improve efficiency.

**idx\_school\_catchments\_catch\_type**: CATCH\_TYPE is used to filter different types of school districts, and the index helps speed up queries.

**Bike\_lockers\_sheds**

create index idx\_bike\_lockers\_sheds\_geometry

**Petrol\_stations**

create index idx\_petrol\_stations\_geometry

**Traffic\_lights\_location**

create index idx\_traffic\_lights\_location\_geometry

These indexes are created to optimize the performance of common query and join operations. By creating indexes on join fields and fields that are frequently used for filtering, we can significantly increase query speed and reduce the response time of database operations. This ensures that the system can perform tasks more efficiently when dealing with large data sets and complex queries.

**Score Analysis**

**Business per 1000 people Z-score:**

Where

We did not choose a specific industry because we believe that the higher the number of enterprises in each region relative to the population, that is, the higher the z-score indicates more active business activities, economic prosperity, and more convenient life.

**Stops number region Z-score:**

The higher the Z-Score, it indicates that the area has more bus stations, which usually means that the public transportation system is more developed, residents travel more conveniently, the area is closely connected inside and outside, and the community is more prosperous.

**Polls number region Z-score:**

The higher the Z-Score, the greater the number of polling places in the area, which usually means higher community participation, active political participation of residents, complete community service facilities, and a higher degree of prosperity in the area.

**School catchments per 1000 people Z-score:**

**Where**

The higher the Z-Score, it indicates that the area has richer educational resources relative to the population, which usually means that the area is suitable for families to live in, has a higher education level, has a superior community environment, and is more prosperous.

**Bicycle sheds quantity Z-score:**

The higher the Z-Score, the better the bicycle parking facilities in the area are, which usually means that the area pays attention to environmental protection and healthy lifestyles, has good transportation convenience, and has a relatively prosperous community.

**Petrol stations quantity Z-score:**

The higher the Z-Score, the greater the number of gas stations in the area, which usually means that the area is frequently used by vehicles, has large traffic flow, active regional economic activities, and a relatively prosperous community.

**Traffic lights quantity Z-score**:

The higher the Z-Score, the greater the number of traffic lights in the area, which usually means that the area has a complex transportation network, large traffic flow, high degree of urbanization, and a relatively prosperous community.

**Score of business, stops, polls, schools:**

The score takes into account commercial activities, public transportation, community services and educational resources. The higher the Z-score, the better the facilities in all aspects of the community and the higher degree of prosperity.

**Score of business, stops, polls, schools, bicycle, petrol, traffic:**

The Z-Score combines enterprises, bus stations, polling places, school service areas, bicycle parking lots, gas stations and traffic lights. The higher the Z-score, the better the performance in all aspects and the higher the overall prosperity of the community.

**Correlation Analysis:**

According to the median income and z-score shown in the table, we can get from it that when the z-score is higher, the median income is higher.

The usefulness of Z-Score

Comprehensive assessment of regional development:

Multidimensional evaluation: The Z-Score combines the quality of a variety of facilities and services, such as businesses, public transportation, polling places, school service areas, bicycle parking, petrol stations and traffic lights. Through this comprehensive score, we can conduct a multi-dimensional evaluation of the overall development level of a region.

Regional comparison: Z-Score provides a standardized scoring system for different regions, allowing us to easily compare facilities and service levels in various regions. This is very helpful for governments and urban planners when formulating regional development policies.

Identify areas of strength and weakness:

Discover hotspot areas: By analyzing Z-Score, we can identify hotspot areas with higher facilities and service quality, thereby guiding more resources and policy tilt.

Discover weak areas: At the same time, Z-Score can also help us find weak areas with low quality facilities and services, so that governments and community organizations can make targeted improvements.

Data-driven decisions:

Support urban planning: Z-Score provides data-based decision support to help urban planners optimize resource allocation and improve the coverage and quality of public facilities and services.

Promote balanced development: By analyzing the Z-Score of different areas, we can promote balanced development within the city and avoid excessive concentration of resources in a few areas.

Limitations of Z-Score

Some Economic factors are not considered:

Housing prices and cost of living: Similarly, housing prices and cost of living are also important factors affecting residents’ quality of life, but these are not considered in the Z-Score.

Unable to reflect social and cultural factors:

Social capital: Such as the social capital of the community, the social network of residents, etc. These soft factors also have an important impact on the quality of life in the region, but they are difficult to quantify through Z-Score.

Cultural facilities:

Such as museums, theaters, cultural centers, etc. These facilities have an important impact on the cultural life of the region but are not reflected in the current Z-Score calculation.

Data update frequency:

Timeliness:

The accuracy of Z-Score depends on the timeliness of data. If outdated data is used, the scoring results may not reflect current reality.

Dynamic changes:

The development of cities and regions is dynamic, and the quality of facilities and services is also constantly changing, requiring regular data updates and Z-Score recalculations.

Data quality and coverage:

Data accuracy:

The calculation of Z-Score relies on a variety of data sources. The accuracy and completeness of the data have an important impact on the final scoring result. If data is inaccurate or incomplete, scoring results may be biased.

Data coverage:

Some data may be missing or insufficiently covered in some areas, which will affect the scoring results in these areas and lead to distortion in comparisons.