

Data2001 report

Sydney livability suburb

Dataset Description

General method: Loading data phase is conducted by using python code and in jupyter notebook platform and we manipulate the data frame by using the Panda library.

1. Business statistics

The table has the dimension 2301 x 9.

This dataset is downloaded as a CSV file from the CANVA page which contains several business-related attributes for each suburb area including a number of businesses, accommodation and food, retail, agriculture, healthcare, administration and transportation. We create a variable holding a data frame by using the function `read_csv()` and cleaning the data by using `drop_duplicates()` and `dropna()` functions. Create a schema using CREATE TABLE function and name it "business_stats" followed by the name of the column and appropriate type.

2. Neighbourhood

This dataset is downloaded as a CSV file with a dimension of 309 x 13, from the CANVA page which contains several statistics about the neighbourhood area around NSW: name of the area, the area of the land, population, number of businesses, number of dwelling, average rent and income and the number of young people categorized in 4 ranges 0-4, 5-9, 10-14, 15-19. We create a variable holding a data frame by using the function `read_csv()` and cleaning the data by using `drop_duplicates()` and `dropna()` functions to get rid of null value because it will affect the average value when calculating z score. Create a schema using SQL's CREATE TABLE function and name it "neighbour" followed by the column's name and appropriate type.

3. School catchments

The data of school catchments include data from 3 categories: primary school, secondary school and data of schools that will open in the year 2023. This is the spatial data set that includes the information about the school name, type of the school, which education level the school provides, the date the information was added, the priority of the school and the geometry data that illustrate the region surrounding the schools. Three different data sets were inserted into the database with different names : "catchments_future", "catchments_primary" and "catchments_secondary" which have the dimension of 43x17, 1666x18 and 435x18 respectively. The catchments_future dataset provides information which year the school is open for enrollment.

4. SA2 2016 AUST data

This is a spatial data set that is used by the Australian Bureau of Statistics (ABS) to gain the geographically classified statistics. The data has the dimension of 2310x13 and include the main attributes such as region name , state name, area of the region, as well as the geometry border of the region (illustrated as polygon type). The data was originally a zip file, it is downloaded from the CANVA page and unzipped by code in jupyter notebook (Appendix I)

A schema was created with the matching name to the column's name with the appropriate type. A new copy of the dataset was created because the type of the geometry column will be changed by the function derived from tutorial 9 (Appendix II) . Changing all the elements in the column to the same type will keep the data consistency and fit with the schema initial data type. The data is then loaded into the database by the `to_sql()` function and it is named "sa2_statistical_areas".

5. Break and Enter

This dataset is compressed inside a zip file, which represents the density of break enter events in the certain region in Sydney. The dataset has the dimension of 2594 x 7 which includes the information about the density of the area and number representing the density level as well as the geometry data and the area of the region.

6. Cycle network

This dataset is a GeoJson file, downloaded from the City Sydney data hub which has a 1594 x 6 dimension. It has the information of the cycle route in the Sydney central region and it shows the route's type such as : "separated off-road sideway", "Off-road shared path", etc., and the information about the source of the information, the length and the geometry type that illustrate the route.

7. Greenhouse gas emission

This dataset is a GeoJson file, downloaded from the City Sydney data hub which has dimensions of 116 x 20 and the data contain the total greenhouse gas emission each year from 2005 to 2019 and there are 4 major causes : electricity, waste, transport and gas. In addition, the data includes information about the data area and the geometry shape of each region.

Database Description

All the columns that present the id in the table (eg : area_id, objectid, use_id) which are unique attributes are set as primary key automatically by the pgadmin and manually. The foreign key is set up to reference the area_id in business_stats table because the z score data depends mostly on the data from the table, so the region in this table is the region the z score will be calculated. The column sa2_main16 is the area_id that is important to identify the region that the research will be conducted. The cyclenetwork and greenhouse emission have the foreign key geom reference to table neighbour_business because the spatial join will be conducted to calculate the number of routes and the gas emission in the specific region. The spatial index was created using gist which is named geom_id in the sa2_statistical_areas table so that can help to join the spatial faster. The diagram showing the relation of the table is in Appendix III.

To construct the table that is able to support the calculator process, first, join the tables neighbour and sa2_statistical_areas using the area_name because the neighbour table

does not have geom type, the newly created table is stored as "sa2_neighbour". Then, create a new table "neighbour_business" which is formed by joining the sa2_neighbour and business_stats table using area_name and area_id. Then a table combining all three catchments was created by union three table future, primary and secondary school catchments dataset. Using spatial join with function ST_WITHIN to join the table break_and_enter with the neighbour_business table, because calculate the z score require the total area of crime hot spot divide to land area, so it is hard to calculate if the area is partially intersect with the region, so WITHIN function is preferred than INTERSECTS function. The column name "geom" is changed to "geometry" to avoid conflict and the newly created one is called "neighbour_break". Join the school table with the neighbour_break table using ST_INTERSECTS function because the catchment area intersects with the region because it can calculate how many catchments in the following area. Before joining, change the "geom" column in the "school" table to "geom1" to prevent the ambiguous, this table is named calculation table because it has all the data needed for Z score calculation.

Livability Score analysis

To calculate the z score, I need to calculate the measurement by the following equations.

1. school_density = school*1000/number_of_young,
2. accom_density = accommodation_and_food_services*1000/number_of_young,
3. retail_density = retail_trade*1000/number_of_young,
4. crime_density = hotspot_area/land_area,
5. health_density = health_care_and_social_assistance*1000/number_of_young

The problem requires that to calculate the density by 1000 young people, a column named number_of_young is created as the sum of the number of people in 4 columns showing the age range in the neighbour table. Divide by the number of young people and multiply by 1000 so that you can get the number of each attribute per 1000 young people. After getting all the density scores, calculate the z score (formula in appendix IV). Create a column z_sum and calculate it by the formula in Appendix V. Finally, the formula for the livability score S (Appendix 6) is

$S(z) = 1/(1+e(-z))$ (Sigmoid function) with z is the z_sum

Correlation Analysis

The correlation used is the Pearson Correlation Coefficient because it is simple to understand by representing the relation between two variables by a numeric in the range [-1,1]. When the correlation is -1 means two variables are negative correlation and vice versa.

The correlation of median household income and livability score is 0.256141 which is close to 0. Hence the median income may not have much linear relationship with the livability score of the suburb. The result is supported by the plot in appendix VII.

The correlation of average rent and livability score is 0.41092 which is closer to 0. But the score lies in the middle value of 0 and 1 so it can be told that they have a weak positive correlation. The result is supported by the plot in appendix VIII. Last, the map about the livability of the region is draw by matplotlib . It is indicate by the color, from dark orange to green. Dark orange mean the score is pretty low and the green vice versa. The map is in appendix IX.

Sydney Livability Analysis

Our client contacted our real estate company to have an inquiry about a good place to live in Sydney City. They are a typical rich Asian family with parents and 2 kids. The daughter and the son are 7 and 13 years old respectively. We gave them our research about the livability score of each suburb in Sydney city. The client does not have a high confidence in the data because it does not include the environment and accessibility aspect to the data. Our research team improved the accuracy of the dataset by adding two more dataset which is greenhouse gas emission of the suburb and the cycle route that is available in the city.

In the previous calculation table, we join both greenhouse_emission and cyclenetwork table using ST_WITHIN with geoms from both tables. Then add more columns greenhouse_density, cycle_rout_density and the z_greenhouse and z_cycle_route. We also change the crime_density calculation because each area has a different density level, so just using the area may not reflect the correct status of crime in the area. So I multiply the area with the density indicator (column ORIG_FID). The formula to calculate is below :

1. emission_density = (total_emission/category)/Shape_Area
2. cycle_route_density = num_of_route/land_area,
3. crime_density = crime_density_level*hot_spot_area/land_area

The emission is divided into 4 different categories, so I calculate the sum of all emissions in a year and divide by the number of categories to get the average emission in that year.

The new z sum calculation is :

$$\mathbf{z_sum = z_school + z_accom + z_retail - z_crime + z_health - z_emission + z_cycle_route}$$

Hence, changing the density of each measurement and dividing the number of attributes by the population of the area, so the score is more accurate and can be applied to any age range. Our client seem very happy with the new solution that our research team provided

Appendix

I. The code to unzip the file

```
# extract files, only use one time

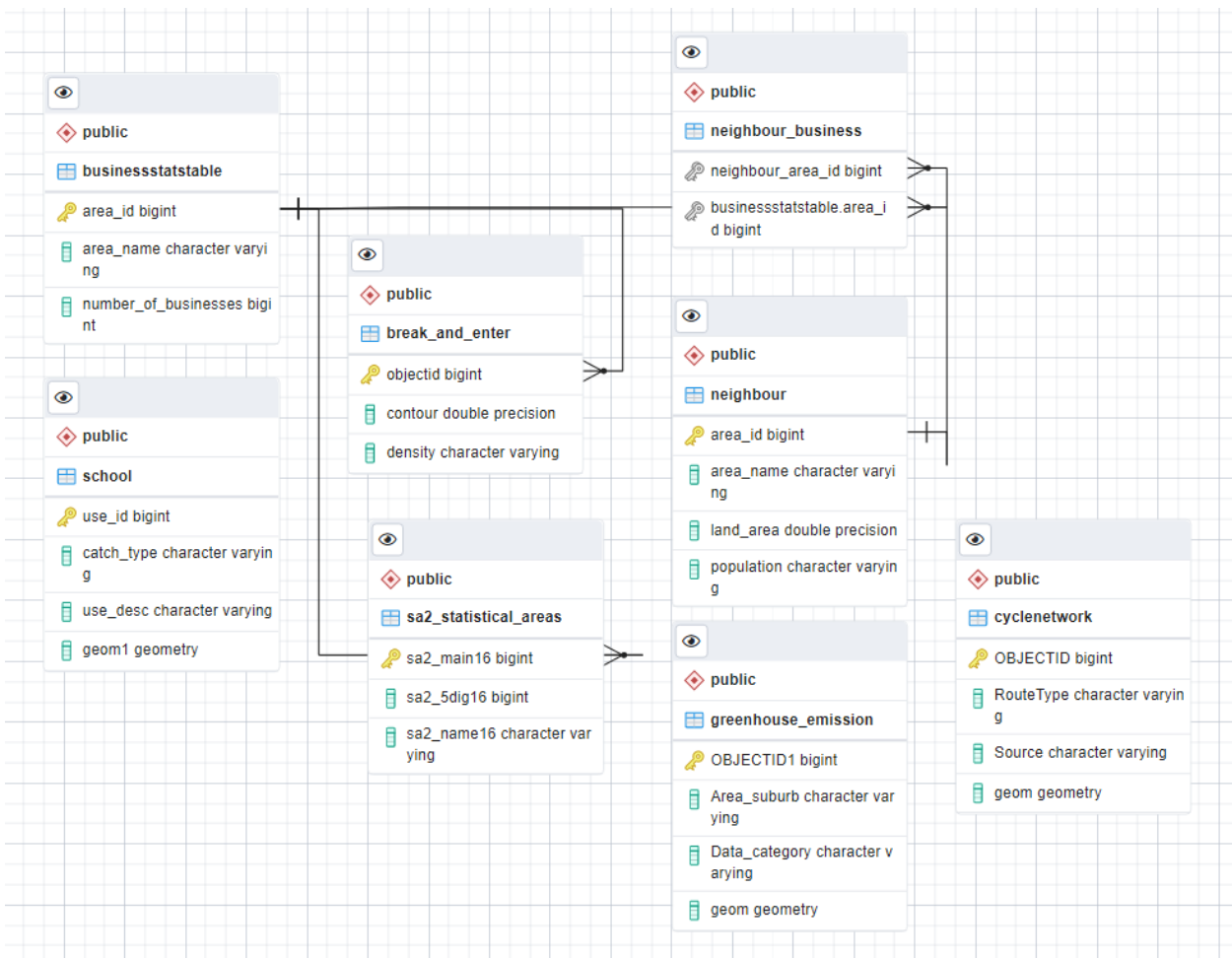
import zipfile

zip_file = zipfile.ZipFile('break_and_enter.zip')
zip_extract = zip_file.extractall()
zip_extract.close()
```

II. The code to standardise the geometry type

```
def create_wkt_element(geom,srid):
    if (geom != None):
        if (geom.geom_type == 'Polygon'):
            geom = MultiPolygon([geom])
        if(geom.geom_type == 'LineString'):
            geom = MultiLineString([geom])
    return WKTElement(geom.wkt, srid)
```

III. The Entity relation diagram of the tables



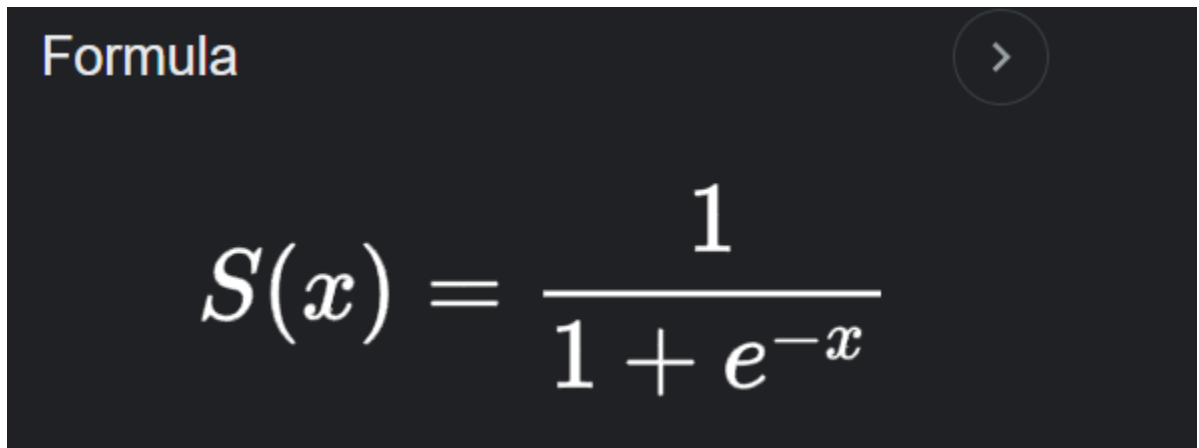
IV . Calculate the z score

$$z(\text{measure}, x) = (x - \text{avg}_{\text{measure}}) \div SD_{\text{measure}}$$

V. The formula for z sum

$$(z_{\text{school}} + z_{\text{accomm}} + z_{\text{retail}} - z_{\text{crime}} + z_{\text{health}})$$

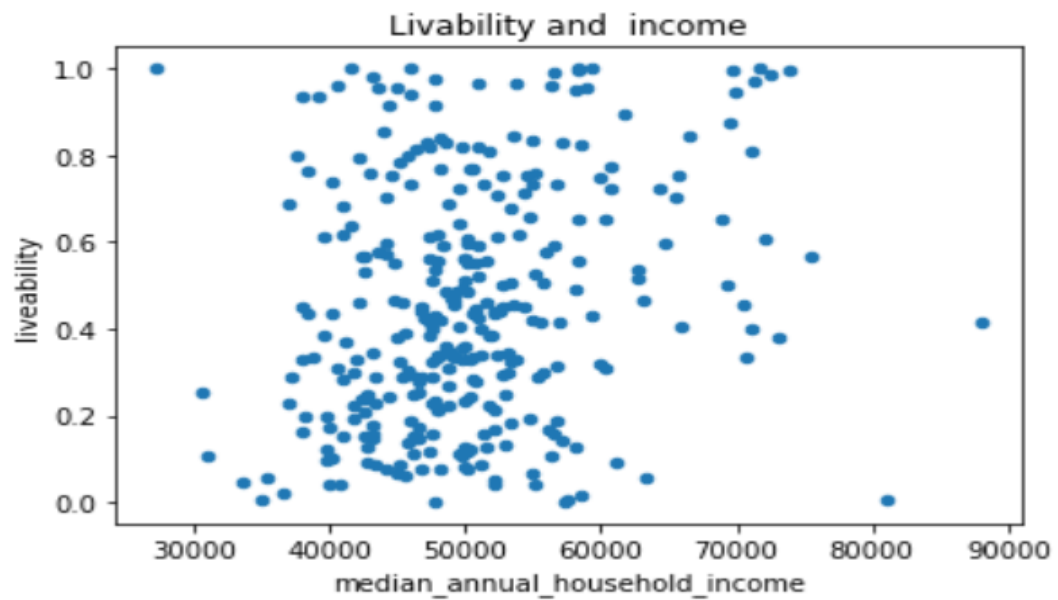
VI Sigmoid function



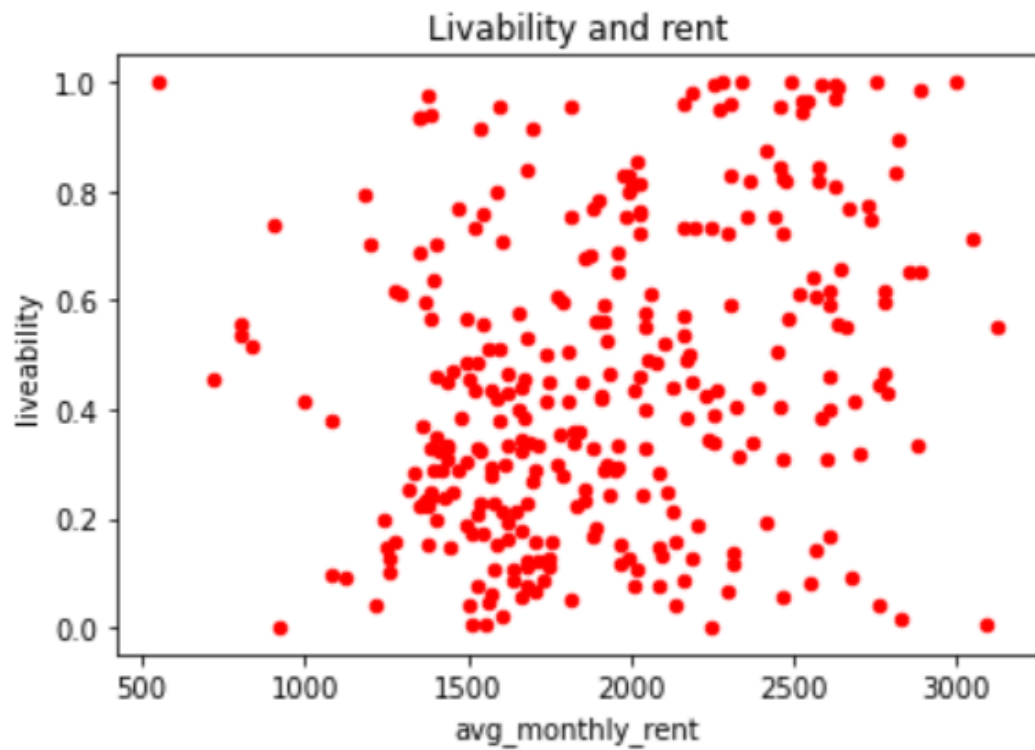
Formula

$$S(x) = \frac{1}{1 + e^{-x}}$$

VII The plot to show the correlation between median income and livability score



VIII.The plot to show the correlation between average rent and livability score



IX. The map show that the livability of Greater Sydney and Sydney region

