DATA2001 Practical Assignment: Bushfire Risk Analysis

Group: F16 - 33

Student Name	Unikey
Lavender Li	shli5792
Yuxuan Xiao	yxia8543

Dataset description

StatisticalAreas Dataset: The “StatisticalAreas” dataset is a comma-separated value (CSV) file obtained from USYD canvas which has 3 columns. It introduces the area identifier and parent area identifiers.
Neighbourhoods Dataset: The “Neighbourhoods” dataset is a comma-separated value (CSV) file obtained from USYD canvas which has 8 columns. It introduces census data on neighbourhoods (SA2-level areas) in Greater Sydney, such as “area name”, “population”, “dwellings” and so on.
BusinessStats Dataset: The “Neighbourhoods” dataset is a comma-separated value (CSV) file obtained from USYD canvas which has 8 columns. It introduces the number of different types of businesses in Greater Sydney, such as the “accommodation and food” number and so on.
RFSNSW_BFPL Dataset: The “RFSNSW_BFPL” dataset is a ESRI Shapefile (shp) file obtained from USYD canvas which has 5 columns. It introduces three types of vegetation in different locations on the map with shape and geometry data.
SA2_2016_AUST Dataset: The “SA2_2016_AUST” dataset is a ESRI Shapefile (shp) file obtained from USYD canvas which has 14 columns. It is Australian Statistical Geography Standard (ASGS) Edition 2016.
Sydney Special 1:250 000 GIS: This dataset is shapefile (shp). The data is part of a series of maps covering the entire Australia, including specific data such as elevation, habitation, transport, and hydrology. In this assignment, we only use part of data which introduces “lakes” in Greater Sydney. (link: http://pid.geoscience.gov.au/dataset/ga/64665)

Data import

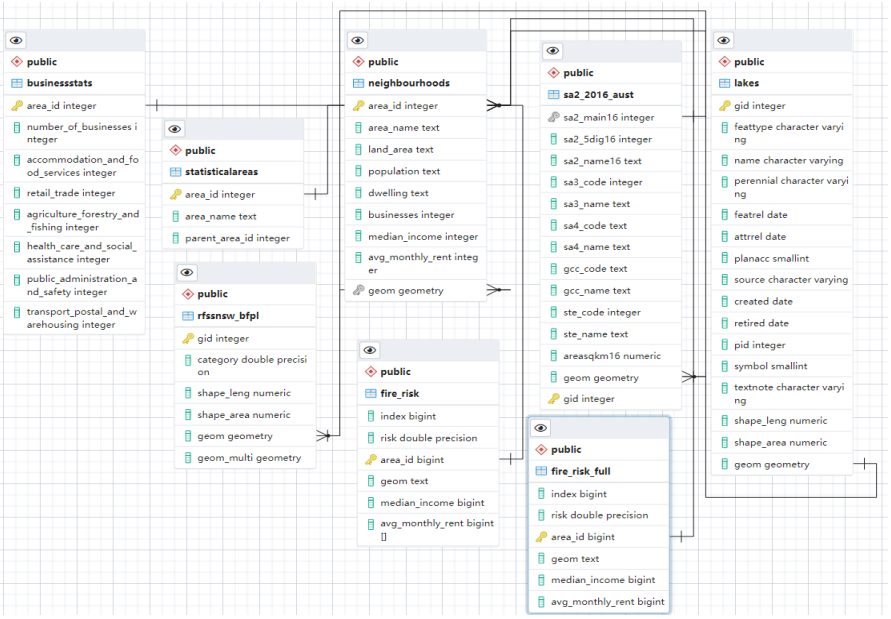
There are six files needed to use in this assignment including three CSV files and three SHP files. In this assignment, we chose to use both PYTHON and POSTGIS to import all the dataset and change the correct table name in the database. Using POSTGIS software imported “RFSNSW_BFPL” and “Lakes" two Shapefile. And all the other files are used by PYTHON to import. All the table names in the database have been set lowercase version.

Data cleaning and preprocessing

1. **dealing with lack data situation**
   After importing all the datasets into the database, we observed that many rows in the dataset contain NULL values or lack some values in different fields. Because we need to analyze the whole 322 neighbourhoods in Greater Sydney and avoid the incomplete map situation, we chose to use the “pandas.DataFrame.fillna” method to fill these values with 0.
2. **dealing with different data types situation**
   After importing all the datasets into the database, we observed that there are many attribute types that are not suitable for analyzing processing. Thus, we changed all the geom types into “GEOMETRY(MULTIPOLYGON, 4283)”. Also changed the “population”,”dwelling” types into int in the “neighbourhoods” table which firstly changed the form with a comma to a comma free form and then transfered data types.
3. **adding some columns in some tables**
   There are many tables in our dataset which need to link among them, especially links to the “neighbourhoods” table. Also we need to draw the graph in this assignment, thus adding the “geom” column into the “neighbourhoods” table is necessary.

Database description

Schema



Primary key

The Primary Key constraint uniquely identifies each record in the database table. Thus, before setting the primary key, we used the “pandas.DataFrame.drop_duplicates” method to ensure that the column values are unique. In our database tables, we set “area_id” for tables “statististicalareas”, “neighbourhoods” and “businessstats” which because each neighbourhoods’ ID is different. And we set “gid” for tables “rfsnsw_bfpl”, “sa2 2016 aust” and “lakes" which because “gid” is just the number unique and starting from 0. Because the shapefiles using POSTGIS software imported set the primary key automatically, we don’t need to add them by manual operation.

Foreign key

By setting the foreign key the “area_id_fk”,let column “sa2_main16” in the table "sa2_2016_aust" refers to the table "neighbourhoods" column “area_id”. Because “sa2_main16” is not the primary key in its table, thus, in this way, the data of the two tables can be linked, and it is convenient to calculate the data of other columns in the two tables. When inserting or updating data entries, the referential integrity of the data between the two tables can be guaranteed.

Index

A database index is an identifier that is attached to a table field to increase query speed. Through the use of indexes, in the process of querying, can greatly speed up data retrieval, and the optimization hider is used to improve the performance of the system. Some of our datasets have lots of rows, thus, using indexes is necessary.
There are two types index using in this assignment:

1. **normal index**
   Indicates a normal index, and can be used in most cases. Thus in this assignment, we set the normal index in all the values needed to be used in fire-risk calculation processing which means “area_id”, “population”, “dwellings” in “neighbourhoods” table;

"number_of_businesses" and "health_care_and_social_assistance" in "businessstats" table; "shape_area" and "category" in "rfsnsw_bfpl" table.

2. **spatial index**
   A spatial index is an index on a field of a spatial data type.Thus in this assignment, all need to be used "geom" data can be set the spatial index which in "sa2_2016_aust", "neighbourhoods" and "rfsnsw_bfpl" tables.

Fire risk analysis

Risk score formula

original fire risk:

fire risk = S(z(population density)+z(dwelling & business density)+z(bfpl density)−z(assistive service density)

after adding our own dataset fire risk:

fire risk = S(z(population density)+z(dwelling & business density)+z(bfpl density)−z(assistive service density)-z(lake density)

Compared with the original formula, we have added a new influencing factor through an additional data set: the density of lakes. In the vicinity of the lake, the humidity of the air is relatively high. Compared with dry and water-deficient places, it is not easy to catch fire, and after a fire, it is easy to take water from the lake and cover the fire. The existence of lake water can reduce the risk of forest fires, so the standard score of lake density is always subtracted from the formula.

Fire risk calculation processing

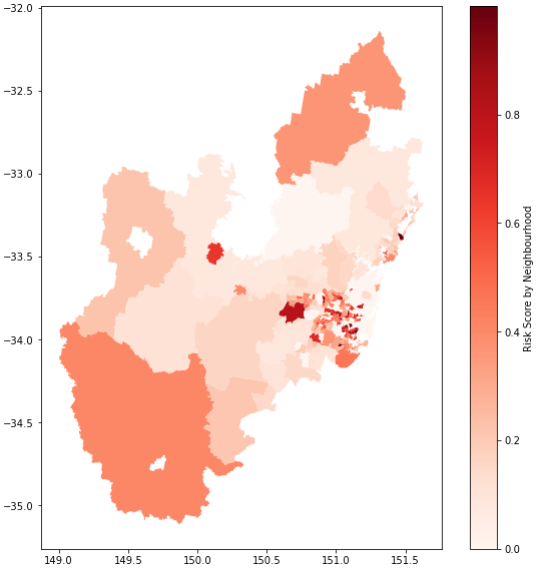
1. **create methods**
   The fire risk result calculation will use logistic function (sigmoid function), and z-score ("standard score"). These two functions will be used more than once in our assignment. Thus, we created two methods called "z_score" and "sigmoid" used to realized. And the final goal is to calculate fire risk, we also created a "fire_risk" method to do that.
2. **join**
   During calculation of density values, we need to use "join" operations. And there are two types of "join" using in this assignment:
   a. **join**
      Join "neighbourhoods" with "sa2_2016_aust" table to calculate population_density and dwellings_density, join "neighbourhoods" table with "businessstats" and table "sa2_2016_aust" to calculate business_density and assistive_service_density, and join "neighbourhoods" table with "lakes" and table "sa2_2016_aust" to calculate lake_density. These columns are specifically chosen to join on each other to ensure referencing to the same area, and then compute geometry related queries.
   b. **spatial join**
      Spatial join refers to the matching of rows in a joining element to rows in a target element based on the relative spatial position of the element. In this assignment, when we link some tables using "geom" values, we need to use spatial join.
3. **create new table "fire_risk" and "fire_risk_full"**
   After calling the "fire_risk" method in our code, we got all the results about each neighbourhood's fire risk. Aiming to make it easier to draw graphs and calculate correlation,

we created new tables to store the fire risk data and add three columns “geom”, “median_income” and “avg_monthly_rent”.

4. **weighted average to calculate “bfpl_density”**
   According to the dataset description, there are three categories for different types of areas. And because we need to calculate fire risk, thus, areas that are not prone to fire weight are low and prone to fire weight high.

Graph visualization

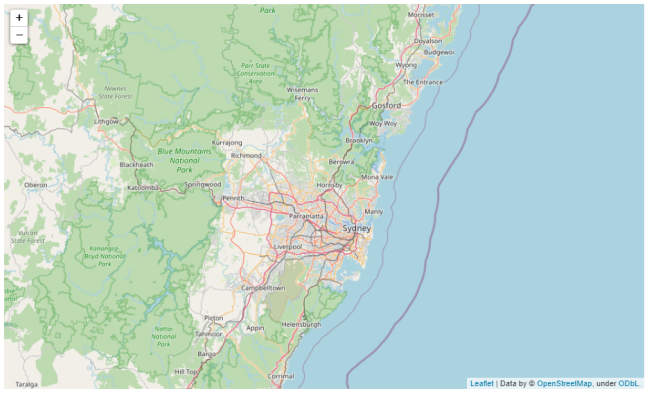
The x-axis and y-axis of the image correspond to longitude and latitude respectively. The label represents Risk Score by Neighbourhood, the darker the color, the greater the risk.
Bushfire Risk is mainly affected by six factors, population density, dwelling density, business density, bf pl density, assistive service density, and lake density. In the picture, the north and southeast of Sydney and the Sydney City are reddish.



- The risk in southeastern Sydney and northern Sydney is relatively high compared to most areas, mainly due to the density of bf pl.The bush vegetation here is very dense and flammable, and there are few assistive services. Although other new books have little impact, the assistive service density and bf pl density directly affect the risk.

- The risk of the location of Sydney is generally higher. Although the vegetation cover is not much, the residential population is relatively large, the activities are frequent, and there are many commercial shops in many areas, and the bustling city center. Population density, dwelling density and commercial density have a much higher impact on risk than assistive service density.

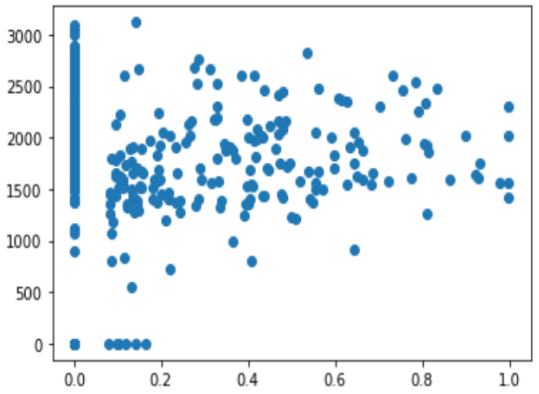
Aiming to show clearly about the Greater Sydney each neighbourhood's, we created a zoomable map about sydney which is an interactive graph.

- 

Correlation analysis

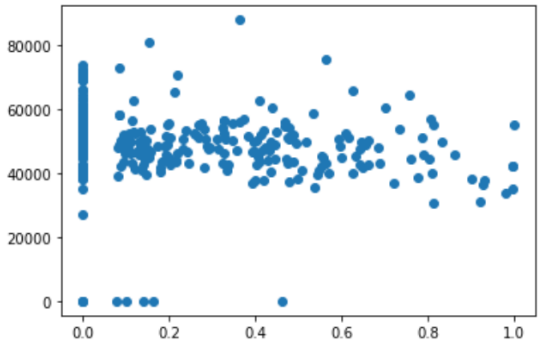
Scatter plot

- Compare with the median household incomes



The x-axis represents fire risk, and the y-axis represents household income.
In the figure, when the fire risk is 0-0.4, the points are concentrated in the range of household income 4000-6000, and a small part is 6000-8000. When the income is 0, there are four or five points scattered below 0.2 risk. When the fire risk is between 0.6 and 1.0, the distribution of points gradually becomes sparse, and the upper limit of the median household income becomes lower and lower. It shows that the higher the household incomes, the lower the probability of bush fire.

- Compare with the rental prices



The x-axis represents fire risk, and the y-axis represents rental prices.
Between the risk value of 0 and 0.2, the points are scattered among all the rent values. Between the risk value of 0.2 to 1.0, the distribution range of the points is gradually shrinking, from 500-2000 to 1500-2500. The rent lower limit of the point distribution is gradually increasing, and the higher limit is decreasing.It shows that where the rent is low, the fire risk is relatively low. In communities with high rents, the fire risk is not the highest.

Correlation

The correlations calculated by the two methods are negative, between -0.1 and -0.3, close to 0. The fire risk score and the affluence of the neighbourhoods have a weaker negative correlation. The richer the neighbourhoods, the lower the fire risk. According to the analysis of the above two scatter plots, household incomes and rental prices are integrated together, and the risk of wealthy neighbourhoods is lower. Since pearson's correlation is a correlation analysis between continuous variables or equidistant measures, it is not applicable to our calculation of fire risk, so the correlation is closer to 0. In addition, the p-value calculated by Spearman is less than 0.001, which is extremely statistically different, and the probability of the conclusion is extremely low.

```
from scipy.stats import pearsonr
corr, p_value = pearsonr(data1, data2)
print('Pearsons correlation: %.3f' % corr)
print(p_value)

corr1, p_value = pearsonr(data1, data3)
print('Pearsons correlation: %.3f' % corr1)
print(p_value)
```

```
Pearsons correlation: -0.169
0.0023544529413270544
Pearsons correlation: -0.062
0.26711308791185095
```

```
from scipy.stats import spearmanr
corr, p_value = spearmanr(data1, data2)
print('Spearmans correlation: %.3f' % corr)
print(p_value)

corr1, p_value = spearmanr(data1, data3)
print('Spearmans correlation: %.3f' % corr1)
print(p_value)
```

```
Spearmans correlation: -0.278
4.0554913668676157e-07
Spearmans correlation: -0.210
0.00015118764242690868
```

	index	risk	area_id	median_income	avg_monthly_rent
index	1.000000	0.063177	1.000000	-0.021639	0.033796
risk	0.063177	1.000000	0.063177	-0.201948	-0.136019
area_id	1.000000	0.063177	1.000000	-0.021639	0.033796
median_income	-0.021639	-0.201948	-0.021639	1.000000	0.450129
avg_monthly_rent	0.033796	-0.136019	0.033796	0.450129	1.000000