DATA2001 Bush Fire Risk Analysis Assignment

F10-50

ZengCheng 490006698
Meian Bao 490440320

Dataset Description

This project contains seven datasets and many of them come from different sources. Some of the datasets come from the Australian Bureau of Statistics(ABS) such as Neighbourhoods.csv, SA2_2016_AUST.shp, StatisticalAreas.csv, and BusinessStats.csv. The BFPL we use is the Small version. Moreover, there are two extra datasets, Fire_NPWSFireHistory_13052021.shp from SEED, The Central Resource for Sharing and Enabling Environmental Data in NSW, and SUA_2016_AUST.csv from ABS.

- Neighbourhoods.csv:
  It is a census data on neighbourhoods (SA2-level areas) in Greater Sydney from ABS. This dataset includes 8 columns which are area id, area name, land area, population, number of dwellings, number of business, median annual household income, and average monthly rent. It will be indispensable data in the later project as some datasets need to join with this dataset.
- StatisticalAreas.csv:
  This dataset is an area identifier and parent area identifiers, including three columns, area id, area name, and parent area id.
- BusinessStats.csv:
  This dataset contains business information about the area, Include nine columns, area id and name, the number of businesses/services(such as health care and social assistance, public administration and safety, transport postal and warehousing)/ accommodation and food/ retail trade in each area.
- RFSNSW_BFPL.shp:
  This data includes the geographic location information of each region, including GEOM, SHAPE_AREA, SHAPE_LENG and CATEGORY
- SA2_2016_AUST.shp:
  This data is a SA2 boundary data, which is a shapefile.
- Fire_NPWSFireHistory_13052021.shp:
  The data is used to evaluate the combustion status and environmental impact, as well as to estimate the severity of the fire in the event of a fire accident.
- SUA_2016_AUST.csv:
  The data is used to evaluate the combustion status and environmental impact, as well as to estimate the severity of the fire in the event of a fire accident.

Data clean

StatisticalAreas.csv SUA_2016_Aust.csv, Fire_NPWSFireHistory_1305221.shp, Sa2_2016_ust.shp and businessstats.csv all use dropna(). For the neighbourhoods.csv, RFSNSW_BFPL_small.shp was processed by fillna (' 0 '). This is to ensure the integrity of the neighbourhoods. For the two SHP files, we also cleaned up the names (from uppercase to lowercase) and converted the polygon to Multi Plogons

Database Description

- **Database Schema:**

  Our ER diagram has the seven tables above plus one spatial_ref_sys. (See Appendix)
- **Primary key:**
  For tables businessstats, statisticalareas, and neighbourhoods, 'area_id' is selected as the primary key. 'area_id' represents the unique id of a certain area. For tables rfsnsw_bfpl and fire_history, 'gid' is selected as the primary key. For table sua_2016_aust, 'SUA_CODE_2016' is selected as the primary key. The primary key can be used to differentiate a column or a group of columns for each row since the primary key is unique and it cannot be a duplicate. Therefore, it ensures the same value will not appear twice in a table.
- **Foreign key:**
  The foreign key is a column that shows the relationship between two tables. It maintains data integrity and acts as a cross-reference between two tables. The column 'area_id' in table neighbourhoods linked the table sa2_2016_aust with 'sa2_main16' as 'area_id' is the primary key of table neighbourhoods and 'sa2_main16' is the foreign key of table sa2_2016_aust. Both of them represent the unique area id of a certain area, by using the foreign key, these two tables can join together.
- **Index:**
  There are many indexes created in the project. For example, 'Area_Name_Index' for column 'area_name'in table neighbourhoods, "BFPL_Gemo_Index" for column 'geom' in table rfsnsw_bfpl, 'update_risk' for column 'land_area' in table neighbourhoods. The purpose of all the indexes is to speed up the next query
- **Join tables:**
  In task2, we need to do some density calculation through different tables, to make the coding process more convenient, all the columns we will use need to join in the same table. We join tables neighbourhoods, businessstats and rfsnsw_bfpl together by using 'area_id' column and 'geom' column. The 'area_id' column in table neighbourhoods is used to join with the 'area_id' column in table businessstats as they all represent the same specific id of an area, The 'geom' column in table neighborhoods is used to join with the 'geom' column in table rfsnsw_bfpl, where the 'geom' column in table neighbourhoods is adding from table sa2_2016_aust.

Fire Risk Score Analysis

Final score formula:
new_fire_risk = S(z(population density)+z(dwelling & business density)+z(bfpl density)+z(agroforestry density)−z(assistive service density))

Note: bfpl_ density = (shape_area * category_coefficient)/ land area. Coefficient of Category1 is 100, Coefiicient of Category2 is 30, Coefficient of Category3 is 40.

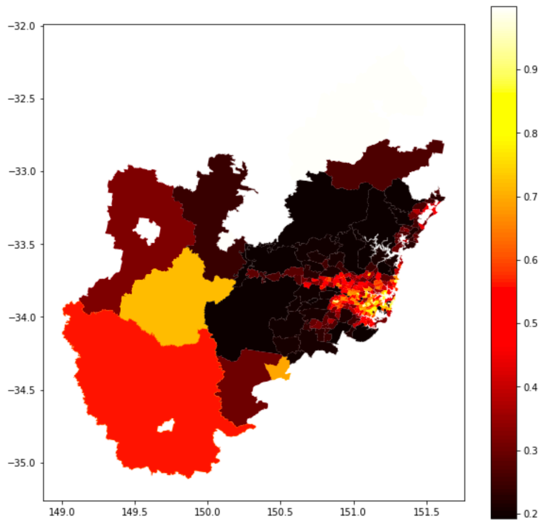
Reason:

I added an agroforestry density. It means that number of agriculture forestry and fishing divided by neighbourhood land area. Risk is +. Data Source is BusinessStats.csv

Because I think if there were more agriculture and forestry, the risk of fire would be higher. Because if there are more forest farms, then flammability is likely to rise. There is a higher risk of fire if fire protection is not done properly

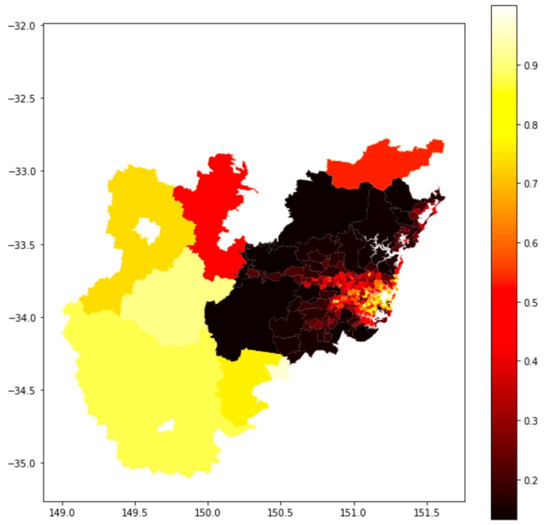
The map:

According to the formula given (**original formula**), we get the map, look like this:



The X and Y axes represent geographic location data from the GEOM variable in Neighbourhoods, which is POLYGON. Black to white represents the level of fire risk. White represents the highest fire risk score and black represents the lowest fire risk grade. This level was set based on the Fire Risk Score we had previously obtained. As we can see from the graph, the risk is higher at the edge and in the middle of the map.

According to the **new formula**, we get the map, look like this:



And obviously, the images we get from the new formula show that the surrounding cities are more dangerous than the previous picture showed. The most likely reason is that the surrounding area has a higher forest coverage, which leads to a higher fire factor.

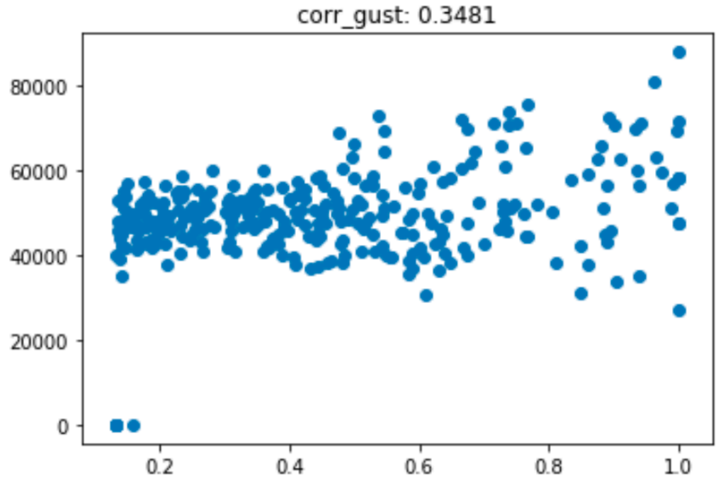
The fire risk score in the middle of the map (city) is also relatively high. According to the data analysis, we believe that the reason is that the city has a high population density and many buildings (including dwelling and business). Although the city's corresponding assistance will be better, but still can not fundamentally prevent the occurrence of fire. So this leads to a very high fire risk score in the center of the city.

Correlation Analysis

Median income :

Firstly, We calculated the correlation coefficient between **fire risk score** and average income, which is 0.1771.

Then, We calculated the correlation coefficient between **new fire risk score** and average income, which is 0.3481. The following is a scatter plot of the two:



In this graph, the X axis represents fire risk score and the Y axis represents average income. We can see that the relationship is not so obvious.

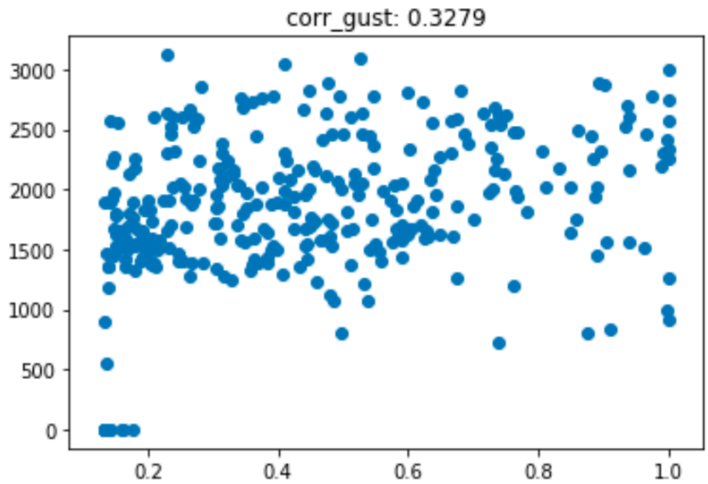
And we can see that this coefficient is a little bit higher than it used to be according to the new formula. There is still no clear trend, but we can assume that people with higher incomes are more likely to live in urban centers, where the risk of fire is higher. However, since this change is not very obvious, we can not get the conclusion that income is positively correlated with the fire risk score.

We get a common reason for the explanation through discussion, because the fire risk is related to the population density and dwelling/business density of the area, and the income of people within a city must be uneven. People in the same building may have different incomes. So this also leads to the result that the fire risk score is not related to the average income.

Average rent (new formula)

Firstly, We calculated the correlation coefficient between **fire risk score** and average rent, which is 0.2592.

Then, We calculated the correlation coefficient between **new fire risk score** and average rent, which is 0.3279. The following is a scatter plot of the two:



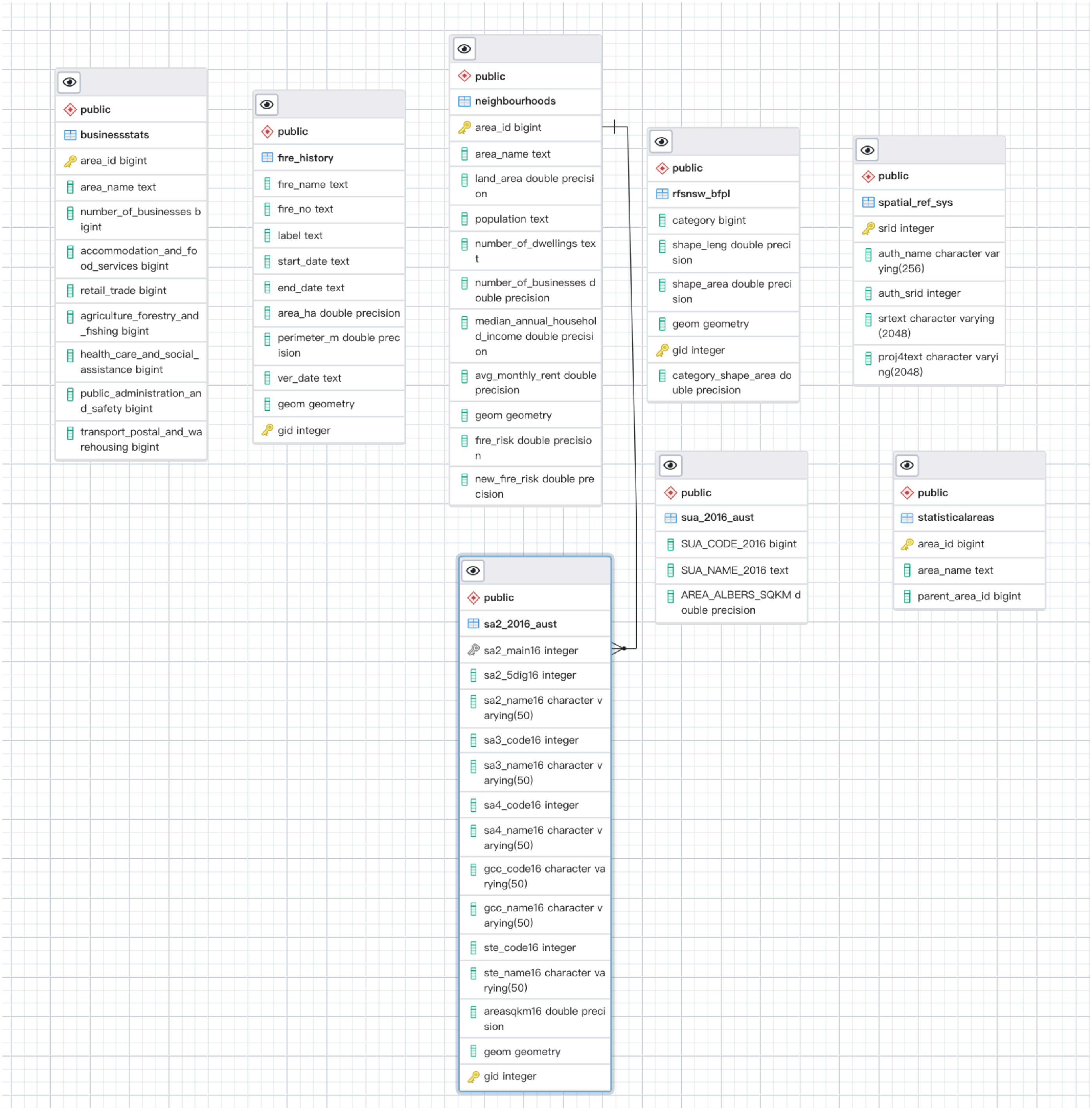
In this graph, the X axis represents fire risk score and the Y axis represents average rent. We can see that the relationship is not so obvious.
We can see that there is still no clear relationship between new fire risk score and average rent. But the correlation is higher than before. We speculate that this is due to the addition of agroforestry density in the new formula. We agree with the previous point, because the fire coefficient is directly related to dwelling/business density and population density. The rent level does not directly affect the housing density. Residential areas with high rent also have many high-rises and wealthy people, while areas with low rent also have many ordinary people and ordinary houses. Therefore, we believe that rent has no relationship with fire risk, and there is no obvious trend.

The relevant code:

```
neighbours = pd.read_sql_query('''SELECT * FROM neighbourhoods''',conn)
rent_list = [] #income_list = []
related_risk = []
for i in neighbours['avg_monthly_rent']: # neighbours['median_annual_household_income']
    rent_list.append(i)
for x in neighbours['new_fire_risk']:
    related_risk.append(x)
corr_gust = round(neighbours['new_fire_risk'].corr(neighbours['avg_monthly_rent']),4)
print(corr_gust)
```

We created two lists to store all the average rent and fire risk scores. Using the corr method in Pandas, we calculated the correlation coefficient between the two, preserving only four decimal places.
The same code is also used to calculate the correlation coefficient between average income and fire risk score, only changing the variable from 'avg_monthly_rent' to 'median_annual_household_income'.

Appendix


public
businessstats
area_id bigint
area_name text
number_of_businesses bigint
accommodation_and_food_services bigint
retail_trade bigint
agriculture_forestry_and_fishing bigint
health_care_and_social_assistance bigint
public_administration_and_safety bigint
transport_postal_and_warehousing bigint
public
fire_history
fire_name text
fire_no text
label text
start_date text
end_date text
area_ha double precision
perimeter_m double precision
ver_date text
geom geometry
gid integer
public
neighbourhoods
area_id bigint
area_name text
land_area double precision
population text
number_of_dwellings text
number_of_businesses double precision
median_annual_household_income double precision
avg_monthly_rent double precision
geom geometry
fire_risk double precision
new_fire_risk double precision
public
rfsnsw_bfpl
category bigint
shape_leng double precision
shape_area double precision
geom geometry
gid integer
category_shape_area double precision
public
spatial_ref_sys
srid integer
auth_name character varying(256)
auth_srid integer
srtext character varying (2048)
proj4text character varying(2048)
public
sua_2016_aust
SUA_CODE_2016 bigint
SUA_NAME_2016 text
AREA_ALBERS_SQKM double precision
public
statisticalareas
area_id bigint
area_name text
parent_area_id bigint
public
sa2_2016_aust
sa2_main16 integer
sa2_5dig16 integer
sa2_name16 character varying(50)
sa3_code16 integer
sa3_name16 character varying(50)
sa4_code16 integer
sa4_name16 character varying(50)
gcc_code16 character varying(50)
gcc_name16 character varying(50)
ste_code16 integer
ste_name16 character varying(50)
areasqkm16 double precision
geom geometry
gid integer