# DATA2001 Project Report

Bushfire Risk Analysis
F14-28

25/5/2021

Faye He 500030507
Abby Guo 490006654

# Dataset & Database Description

Various data sets are used to analyze the bushfire protection capabilities of different communities in Sydney and calculate bushfire risk scores. They are as follows:

## StatisticalAreas.csv

This data set comes from the Australian Bureau of Statistics (ABS). This data provides information about Sydney neighbourhoods that can be obtained by "area_id". This makes data analysis easier because the unique "area_id" variable is used instead of the complex string of area names.

## Neighbourhooods.csv

This data set also comes from the Australian Bureau of Statistics (ABS) and provides unique "area_id" and related information about the area, population, number of houses, number of businesses in each neighbourhood, and rent.

## BusinessStats.csv

This data set is still from the Australian Bureau of Statistics (ABS). Through the unique "area_id", we can find the name of the designated area, the total number of companies, and the categories involved in the area. These categories are: health care/social assistance, accommodation/food service, retail, agriculture/forestry/fishing, transportation post/storage and public administration/security.

## RFSNSW BFPL.shp

This data includes the specific "category", "gid", the length of the fire risk called "shape_leng", the area of the fire risk called "shape_area", and the variable called "geom" that can be compared with the geometry number from SA2.

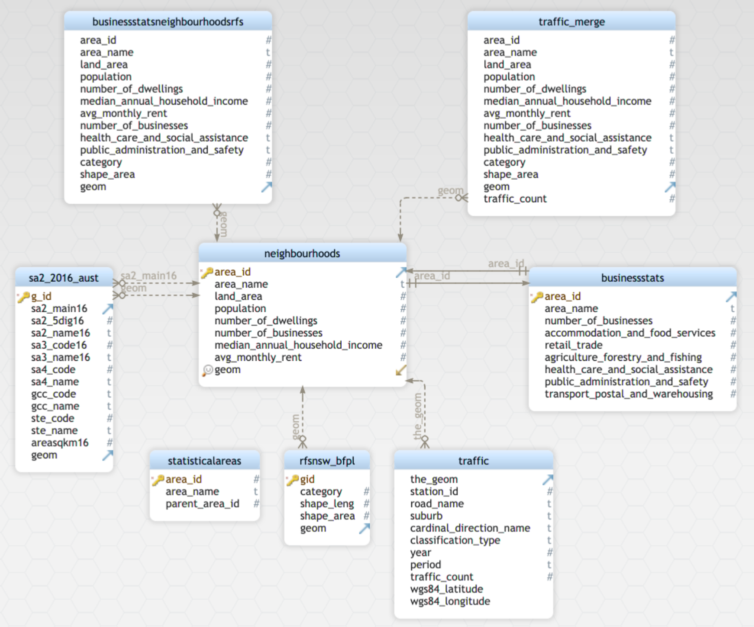
## SA2 2016 AUST.shp

This data set also comes from the Australian Bureau of Statistics (ABS).This data includes a specific "gid", the only "sa2_main16" can be matched with the "area_id" of other data, and "sa2_name16" can be matched with the "area_name" of other data. And the other eleven variables are auxiliary variables.

## Traffic.geojson

This additional data set comes from Transport of NSW, It provides the average road traffic volume of some permanent and sample roadside collection equipment stations in key locations in New South Wales in 2016. Sourced From: [2016. Traffic Volume Viewer, NSW, Australia, URL: Transport Traffic Volume Viewer]

## Database Schema



After our analysis, we found that Neighborhoods.csv and BusinessStats.csv use the same "area_id", so we set the primary key from BusinessStats.csv as a foreign key to Neighborhoods.csv. At the same time, the "sa2_main16" in the SA2 shapefile is also set as a foreign key by us and set to Neighborhoods.csv.

## Pre-processing the data

For this data analysis, we used the Python-based Jupyter Notebook. First, we defined a function called "pgconnect" to connect to the PostgreSQL server (pgAdmin4), and imported psycopg2, sqlalchemy, json and os. Before officially processing the data, we imported a series of useful packages such as pandas, geopandas, numpy, etc. Then we performed "data cleaning", cleaning up the value of "NA" or replacing it with 0. This is more conducive to our data analysis. We found that the rfsnsw shapefile does not have a corresponding "area_name", we first match the "geom" data in the SA2 shapefile with the "area_name" in Neighborhoods.csv and then change the "geom" Add the data in the rfsnsw shapefile and Neighborhoods.csv can use the "geom" data to match Neighborhoods.csv. We matched Neighborhoods.csv, BusinessStats.csv, StatisticalAreas.csv, rfsnsw shapefile and SA2 shapefile into a summary table, which is convenient for calculating the fire factor later. For the additional geojson, we matched it by the region name.

# Fire Risk Score Analysis

$mx.density \ = \ \frac{mx}{Land\ Area}$ (mx is variable)

$Z(mx.density,\ x) \ = \ \frac{x - Mean(mx.density)}{Std(mx.density)}$

**Zdensities =**

$z(population\ density) \ + \ z(dwelling\ \&\ business\ density) \ + \ z(bfpl\ density) \ - \ z(assistive\ service\ density)$

**FireRisk** = $S(Zdensities)$

## Population Density

The purpose of population density calculation is to calculate the population of the area divided by the land area of the area (mx.density). Then find the z value by finding the z formula.

## Dwelling Density

Dwelling density is calculated by dividing the total number of dwellings and businesses in neighbourhoods by the land area of neighbourhoods. Then find the z value by finding the z formula.

## Business Density

Business density is  calculated by dividing the sum of 'number_of_dwellings' and 'number_of_businesses' by the land area of the neighbourhoods. Then obtain the z value by finding the z-formula.

## Assistive_service Density

assistive_service density is calculated by dividing the sum of 'public admin/security' and ' health care/social assistance' by the land area of the neighbourhoods. Then obtain the z value by finding the z-formula.

## BFPL Density

The calculation of BFPL is divided into three steps. Firstly, we use the current shape_area to calculate the radius of the shape, and then add different types of different buffer lengths to the radius of the circle (the default is a circle), and then recalculate the current shape_area.

Secondly, we use this shape_area to divide by the land area of the neighbourhoods, and finally, we obtain the z value by finding the z formula.

## Traffic Density

The traffic density is calculated by dividing the 'traffic_count' by the 'land_area' from Neighborhoods.csv.

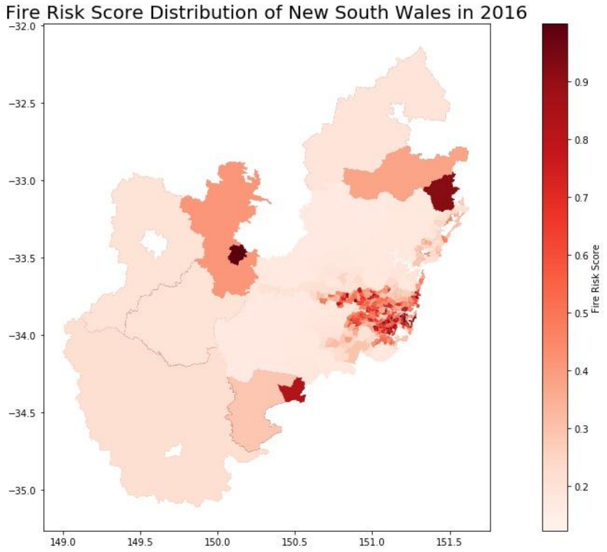
Then calculate Z(traffic_count) according to the formula.

**ZwithTraffic =**

$Zdensities \ + \ z(traffic\ density)$

**FireRiskwithTraffic =** $S(ZwithTraffic)$

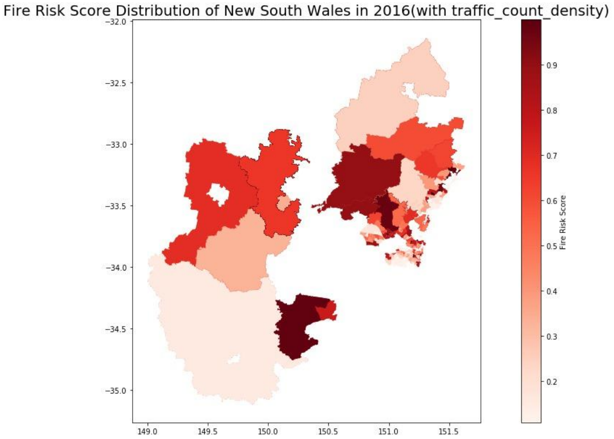
## Fire Score Distribution NSW 2016



After calculating the first five densities, use the formula of fire-risk to find the score and visualize the results. We found that the fire risk is higher in the urban areas of NSW.

Our inference is that combining the first five densities, the reason is the urban areas of NSW are densely populated and commercial areas, which are more prone to fires.

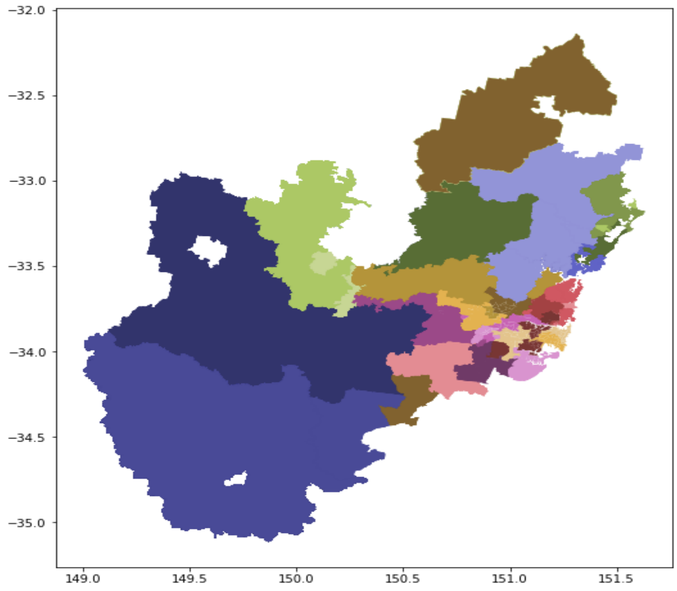
## Fire Score with Traffic NSW 2016



Combine the traffic density with the first five densities and use the formula called ZwithTraffic for finding S to get a new 'risk_score'.

Because some areas cannot be matched, there will be some vacancies on the map. However, according to the existing areas, in 2016, some areas were originally low fire risk areas, but due to the influence of traffic density, they became high fire risk areas.

### IDENTITY OVERLAP BETWEEN TWO MAPS



## Correlation Analysis

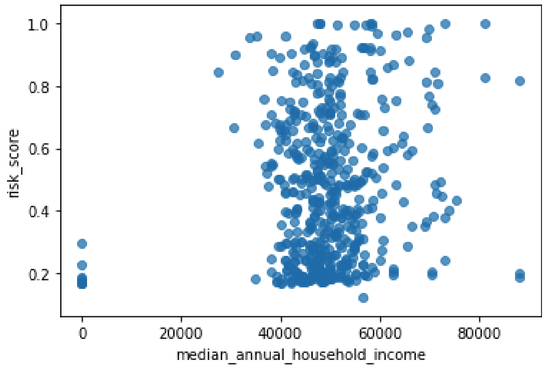
Compare the fire risk score with the average of 'avg_monthly_rent' and the median of 'median annual household income' to determine whether there is a correlation between the two variables.

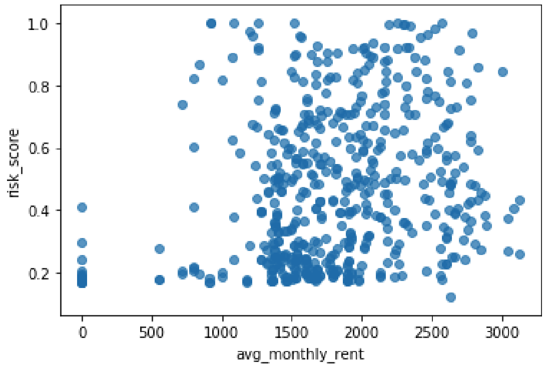
We use the Pearsonr correlation coefficient formula to calculate the correlation between these two variables.

Correlation coefficient between risk_score and median_annual_household_income: 0.189
Correlation coefficient between risk_score and avg_monthly_rent: 0.257

Our correlation coefficient is weak, but there is a slight positive linear correlation between the fire risk score and the median of 'median_annual_household_income' and the average of 'avg_monthly_rent'.

Perhaps people with higher annual incomes are more inclined to live in areas with higher living costs or areas with higher monthly rents, so the trends in income and rent are similar.





We also used the Pearsonr function to get the p-value between the two variables. Our original hypothesis is: the fire risk score is independent of the median income/monthly rent, the alternative hypothesis is not.

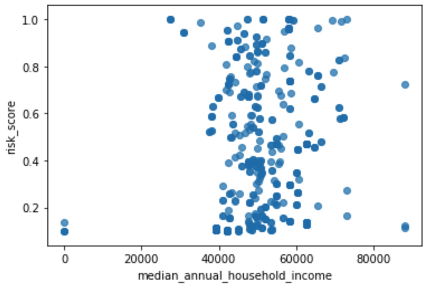
The Pearsonr correlation p-value between the fire risk score and the median income is 0.00001 less than 0.05. Therefore, we should reject H0 and prove that there is basically no correlation between them, which is consistent with our correlation coefficient results.

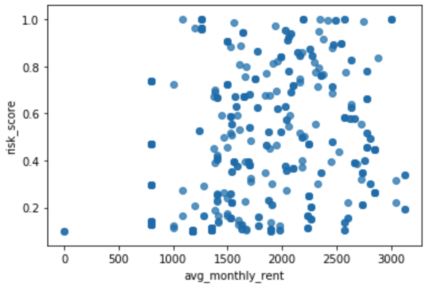
But for other reasons, this is different from what we expected. The reason may be that the level of income is not important, and other facilities in the area are well established, which results in whether the area is high-risk or low-risk.

The Pearsonr correlation p-value between fire risk score and rent is 0.00000 less than 0.05. Therefore, we still reject the original hypothesis. The fire risk score and the average monthly rent are not independent, which is also consistent with our correlation coefficient results. But for other reasons, this is different from what we expected. The reason may be that the level of rent is not important, and it cannot be said that high-income people have different fire risk scores from other areas.

Correlation coefficient between risk_score and median_annual_household_income: 0.077
Correlation coefficient between risk_score and avg_monthly_rent: 0.298

After we added the traffic density, the correlation of our correlation coefficient was still weak, but there was a slight positive linear correlation between the fire risk score and the average of 'avg_monthly_rent'.





After we added additional data, the p-value changed significantly. Our original hypothesis is: the fire risk score is independent of the median income/monthly rent, the alternative hypothesis is not.

The Pearsonr correlation p-value between the fire risk score and the median income is 0.03258 less than 0.05. Therefore, we should reject H0 and prove that there is basically no correlation between them, which is consistent with our correlation coefficient results. But for other reasons, this is different from what we expected. The reason may be that the additional data is related to traffic, so it has a certain impact.

The Pearsonr correlation p-value between fire risk score and rent is 0.00000, which is less than 0.05. Therefore, we still reject the null hypothesis. The fire risk score and the average monthly rent are not independent, which is also consistent with our correlation coefficient results. But for other reasons, this is different from our expectations. The reason may be that traffic does not have enough influence on the current fire risk score.

## Indices

We created four indexes in the database in order to speed up the query speed:
1. The sa2_geom_idx in the geom column of the sa2_2016_aust table. (Space citation)
2. rfsgeom_idx in the geom column of the rfsnsw_bfpl table. (Space citation)
3. The geom_idx on the geom column in the neighbourhoods table. (Space citation)
4. The area_name_idx in the area_name column of the neighbourhoods table.
(Normal index)

## Libraries Used

sqlalchemy, psycopg2, psycopg2.extras, json, os, __future__, pandas, geopandas, math, numpy, shapely, matplotlib, matplotlib.pyplot, scipy.stats, geoalchemy2, shapely.geometry and geopandas.tools.