Bushfire Risk Analysis Report

INTRODUCTION

Pre-assigned and our own datasets detailing the basic information of regions in NSW and their respective geometry locations, this report aims to compute the bush risk scores for all the neighborhoods in the given data by measuring the density of potential leading factors, mainly focusing on the Greater Sydney whilst evaluating how the bushfire risk effects on the affluence of the neighborhoods.

DATASET DESCRIPTION

This project includes many datasets and various data formats.

“Neighborhoods.csv”

This census data source is collected from Australian Bureau of Statistics (ABS) agency, which is the pivot of this project. The data provides the information of neighborhoods in Greater Sydney including population, the number of dwellings, income, rental, etc. Each neighbourhood has corresponding “area_id”. The concrete id code is the pivot for bridging other pre-assigned datasets, and helps further data processing and analysis.

“StatisticalAreas.csv”

The dataset sources from the Australian ABS as well and lists the geographically hierarchical regions with their corresponding "area_id" and “parent_id”.

“BusinessStats.csv”

This data source is also provided by ABS. It covers the number of overall business and counts of different business categories for per regions in NSW including many assistive services.

“RFSNSW.shp”

The shapefile dataset is from the NSW Rural Fire Service that provides the locations of which categories of vegetation in NSW as well as the area and length of vegetations.

“SA2_2016_AUST.shp”

The shapefile, also from ABS provides the geographic information of NSW regions from various hierarchies. This dataset is as a reference to obtain the physical locations of each neighborhood in the "neighbourhoods.csv"

“NSW_Wetlands_2016.shp”

The shapefile sources from data.gov.au, containing various information of wetland and non-wetland in NSW, ranging from names to geographic locations.

Pre-process Data

1. Data Wrangling

It is necessary of performing data cleaning to ensure more precise data analysis in the further steps. After loading the datasets, check the null and any duplication, and then drop them if suitable. Transform the column names to lower case to ensure consistency. Then check the types of each column in datasets. In "neighborhood.csv", population, number of dwelling, and business objects, all of which are cast to “int” by deleting the commas in the number. It is meaningful to explore wetland impacting on the bushfire in

neighborhoods and thus we filter out the non-wetland data in the "wetland" file.

2. Data integration

First, to obtain geographic locations per neighborhood, insert neighborhood dataset into boundary dataset by merging "sa_2016_aust" shapefile with “neighborhood” one, and then select needed columns. In this case, the new dataset, "neigh_bound" is Geodataframe which is helpful for further spatial joining. A similar process occurs in the following. Sort out the vegetation information in terms of "area_id" in the neighborhood dataset by spatial joining "neigh_bound" and "RFSNSW" shapefile so to load the vegetation information into "neigh_bound". Similarly, spatially join of wetland and vegetation as well as neigh_bound dataset. These integrated datasets are prepared for the following visualization and scope into relationships of different variables.

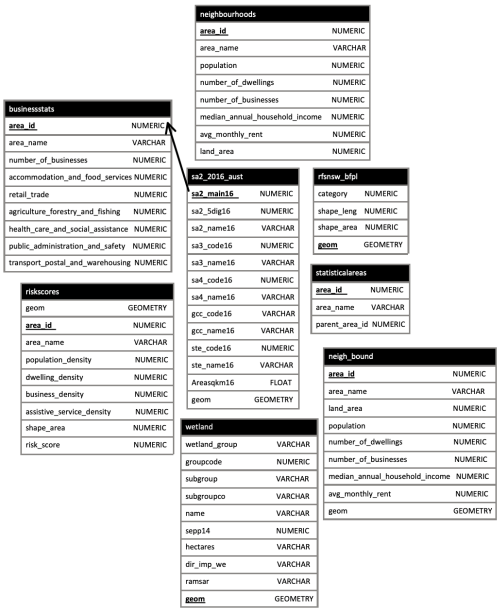
3. Set up schema

To store and manage datasets effectively, it is necessary to connect to the University of Sydney's PostgreSQL server and to execute SQL query stated in Python language in Jupyter Notebook, the required library, "psycopg2" need to be loaded.

Then create schema and establish tables for each original datasets as well as merged or joined ones. The table includes column names as well as variable types. Since some datasets include the geometry type of columns, it is necessary to project the geometry by the same spatial reference identifier to the same system. Then insert datasets into respective tables and identity primary key (Pk) and foreign key (FK).

DATABASE DESCRIPTION

Schema



The tables, besides "rfsnsw_bfpl" and "wetland", provide the related information with "area_id" as the primary key. Since the two other datasets contain the geographic location, a Primary key "geom" is created. Although neighborhoods data is the pivot among all datasets, as "neighbourhoods" dataset contains the least number of area_ids, we could not create foreign keys outside of neighborhoods in any other schemas when we have to retain all valid original data of these datasets. Therefore, we only refer "area_id" in sa2 dataset to counterpart in business one.

Index

1. Index *Neigh_id* (normal index) for the *area_name* column in *neighbourhoods* table.
2. Index *Busin_areaId* (normal index) for *area_id* column in businessstats table.
3. Index *Rfsnsw_geom* (spatial index) for *geom* in *rfsnsw_bfpl* table.
4. Index *Wetland_geom* (spatial index) for *geom* in wetland table.

5. Index *Sa2_id* (normal index) for *sa2_main16* in *sa2_2016_aust* table.
6. Index *StatArea_name* (normal index) for *area_name* in *statisticalareas* table

FIRE RISK SCORE ANALYSIS

Calculation Explanation

We calculate the fire risk score for per neighbourhood by following formula:

$$
\begin{aligned}
fire\ risk\ score \\
&= S(z(population_{density}) \\
&+ z(dwelling_{density} \\
&+ z(business_{density}) \\
&+ z(bfpl_{density}) \\
&- z(assistive\ service_{density}) \\
&- wetland_{denstiy})
\end{aligned}
$$

S represents logistical function(sigmoid function)
Z represents z-score(standard score) of a measure

1. Population density

Using neighbourhood dataset, the population density is calculated by dividing the population per region by responding land area (the value from neighbourhood dataset, same at all following calculations).

2. Dwelling density

Using neighbourhood dataset, the dwelling density is calculated by dividing dwelling per region by responding land area.

3. Business density

In terms of assignment instruction, we need to use the number of business information in the business dataset, however, there is no land area information in that. We first merge the neighborhood and business datasets to load the data into the neighborhood. Therefore, the business density is calculated by dividing the population per region by responding land area.

4. Assistive service density

Using the integrated business and neighbourhood dataset before, the assistive density is calculated by dividing the number of "health_care_and_social_assistance" and "public_administration_and_safety" of per region by responding land area.

5. BFPL_density

Adjust the land area of every plot of vegetation by establishing a risk parameter in terms of different categories. In the scenario, enlarge the land area of categories 1 and 3 by 1.01 and 1.001 times respectively and remain the category 2 the same, because from the original website, the potential risk of three categories are: category1 > category3 > category2. Sum up the overall area of vegetation per neighborhood and then divide it by the corresponding neighborhood area to get the density.

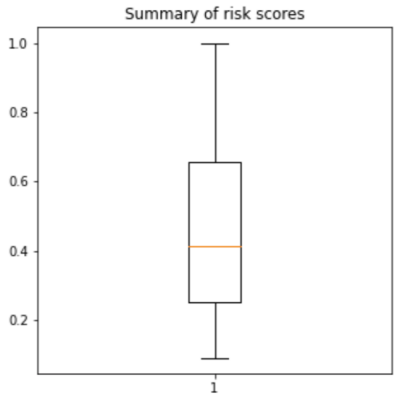
6. Wetland density

Map the wetland geographic location into the neighborhood to calculate the total of the overlapped area per region to get the value of all wetland areas in each neighborhood. Then the wetland density is calculated by dividing the area of all wetlands per neighborhood by responding land area square. We do not use z-score to standardize wetland because the wetland area from our dataset is too small.

7. Z score/logistical function

All the z scores and the logistical function calculated by the hardcode import from “scipy.stat” library by Python.

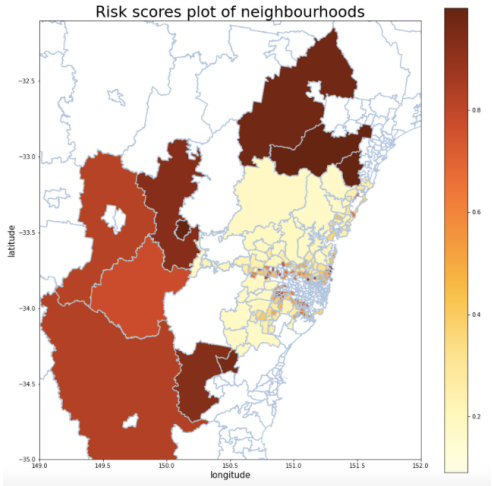
Result Description



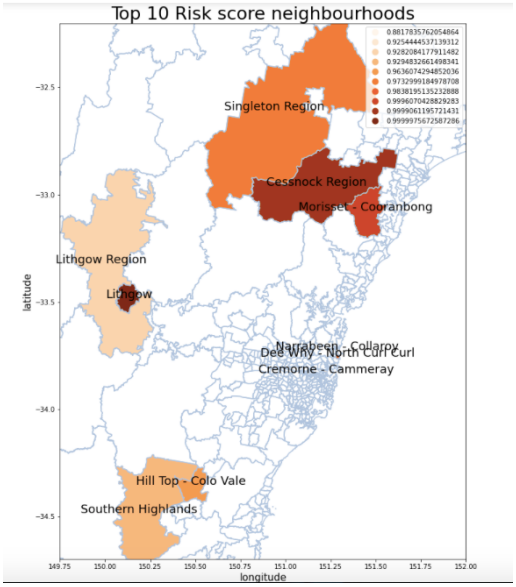
[plot 1]

From the boxplot of bushfire risk score, the majority of scores cluster around 0.3 to 0.6. It means that most regions in Greater Sydney are at moderate or low risk.

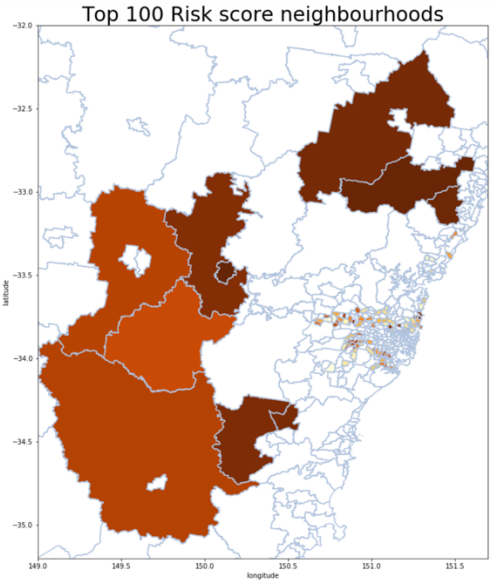
From the following graphs, it is clear outlining the spread of risk score for neighbourhoods across Greater Sydney.



[Map 1]



[Map 2]



[Map 3]

The graph (Map1) clearly outlines the spread of risk scores for neighborhoods across Greater Sydney. Generally, the inner and particularly North part are facing higher potential bushfire risks while the East part and Sydney city area less incur bushfires. The rural areas are more prone to have bushfires compared to urban regions. In the scope of the Sydney and its adjacent regions, there are a bunch of regions with high risks scattering around Sydney.

When zooming into the top 10 areas of high bushfire risk scores, the graph indicates that the regions near the city

are also at high risks. From map 2 and 3, they show that there are massive regions around Sydney with high risks, which indicates that most of bushfire risky areas concentrate closely surrounding Sydney.

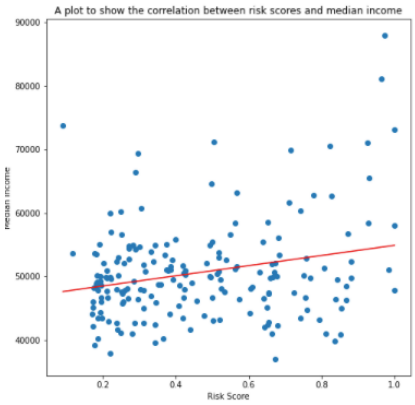
Thus, from the graphs, we could figure out that the regions scored high risks spread around the Sydney and inner East and North part of Great Sydney.

CORRELATION ANALYSIS

Comparing the risk score with median annual income and monthly rent respectively could discover the potential impact of bushfire on the affluent for the neighbourhoods. From the Pearson correlation coefficient,

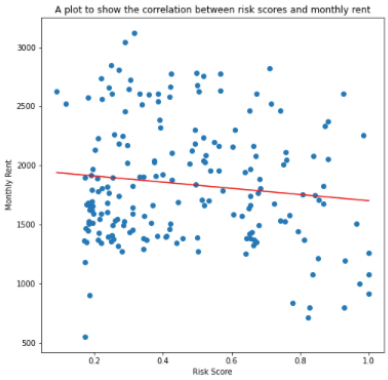
risk score and median annual income is `0.2470157534792021`

significance p-value is `0.00059079136534668 71`



risk score and monthly rent is `-0.12160328625518221`

significance p-value is `0.09465596093662912`



According to the correlation coefficients and p-values, there is slight positive tren d between risk score and annua income, which is significant according to P-value (>0.05), while the other one is insignifica nt(P-value>0.05).

Therefore, we could conclude that the bushfire risk and income have a positive trend but the association is weak. The association might due to the influences from many aspects including population, businesses situation and so on, and it may be a further question to our future research.

Although the correlation coefficient of risk score and rent is insignificant from a statistical perspective, there still shows a negative trend but the possible association between them is weak as well.