BUSH FIRE ANALYSIS

DATA2001 – ASSIGNMENT 1 - 2021

Rahul Talla 500551998

Patrick Chang

Word Count – 1902

This document entails the entirety of the Bushfire Risk Analysis carried out for the greater Western Sydney regions as required for DATA2001 Assignment 1. The report includes data integration, wrangling, analysis and intermediate modelling of correlation analysis.

CONTENTS

OVERVIEW

The following documentation entails the findings of the carried-out bushfire risk analysis. With the eminence in climate change and its observed temperature variation, the following analysis aims to achieve an intermediate approximation on the relative fire risk scores within the Greater Western Sydney regions of NSW, Australia.

DATA SOURCING & CONTEXT

Datasets gathered include Australian SA2 areas of 2016 entails Statistical Areas Level 2 (SA2) which are medium-sized general-purpose areas built up from whole Statistical Areas Level 1. Their purpose is to represent a community that interacts together socially and economically (ABS, 2016). They vary from with population sizes from 3,000 to 25,000.

RFSNSW Bush Fire Prone Land areas dataset consists of categorised shape data of geometric format consisting of the NSW region and its corresponding risk category and vegetation status alongside quantitative measurements of land regions (RFS ,2015).

Neighbourhoods' dataset consists of the regions within NSW and their corresponding population, property and economic census data. The areas in the dataset are the only areas within Greater Western Sydney that are considered for the fire risk analysis.

Businessstats dataset is more comprehensive account of the regions within NSW and the corresponding commercial and economic census data. The dataset consists of explicit business types/activities and their operating size. Statisticalareas dataset consist of supporting SA2 and SA3 areas with corresponding area codes of Neighbourhoods and Businessstats datasets to examine eccentricities in data models

(2 ADDITIONALS) Live-JSON Weather forecast data from Yahoo weather using Rapid API which provides aggerate weather for Greater Western Sydney Region for next 10 days. Also, Wetlands' shape dataset from Department of the Environment and Water Resources consists of data on "…areas of marsh, fen, peatland or water, whether natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish or salt, including areas of marine water the depth of which at low tide does not exceed six meters…" (Department of the Environment and Energy, 2018).

DATA INTEGRATION & WRANGLING

Data integration was carried out in stages where data was first examined for expected datatypes and context of each dataset. This process enabled the following data wrangling stages.

Data Redundancy - Redundancy checks were carried out on all datasets before the integration this involved checking for duplicates by area_id, followed by row-based checks. With findings reported to notebook output cells. These were replicable on all instances of the notebook runs and data entry to the data base with use of Primary and Foreign keys where necessary. Statisticalareas dataset was found to have duplicate rows which were filtered by keeping the first appearing row.

Unwanted/Useless Data – Datasets were primarily referenced through spatial joins and area_ids where required. Hence, NULL, NA and other useless data values were first set to 0 before uploading onto PostgreSQL. In terms of usability these 0 values containing rows were then filtered based on the context of the column. Null and NA were observed in the Neighbourhoods dataset these were first set to 0 then if the column is quantitative, they were replaced by the average of all its column values. Otherwise, the row would be dropped. Intermediate ranges and median scores of quantitative columns were examined during dataset integration to identify outliers and discrepancies with context of data. Database Schema design

Assumptions, Limitations and Creative Choices - Data within all datasets was filtered such that only area_ids corresponding to the primary table (neighbourhoods) was present. Which consequently resulted in all 3 CSV and SA2 shape files consisting of 322 rows. This narrows the range of data being investigated upon to those given in the neighbourhood's dataset. After the filtration original copies of the tables are removed from PostgreSQL to prevent confusion however, they are kept locally for rollbacks if necessary.

Database Schema design – The following is the database schema diagram, note that assumptions and limitations exist for several relations with computation-based tables as these originated from the primary tables their respective relations have been assumed to exceed those of the original. Spatially joined tables have been shown to share the similar column values (e.g., rfs_sa2_join and bfpl_density).

Index Creation – Three indexes (Two spatial indexes) were created specifically to efficiently run queries and reduce time taken on each query. These were for the 'goem' columns in sa2_2016_aust and rfsnsw_bfpl datasets under the names 'sa2x' and 'rfsnsw_bfplx' respectively. Another for the Wetlands dataset under the name 'wid'.

Link to Database Schema in Appendix - (Database Schema)

FIRE RISK SCORE CALCULATION

The fire risk score was calculated by computing all densities using the assignment specifications. However, to calculate the bfpl_density a different approach was carried out. As the resulting dataset after joining point-based rfsnsw_bfpl geom values with the multi-polygon goem values in sa2_2016_aust dataset. It was observed that within a single multi-polygon/boundary in the SA2 regions existed multiple corresponding points with various shape areas. Those with the same area_id and area_name where then grouped by taking the aggregate sum of their shape area. Such that within a region the total shape area for all categories were found, an example is below:

Step 1 – Spatial Join		
area_id	category	shape_area
1234	1	12
1234	1	15
1234	2	17
1234	3	18
1234	3	2

Step 2 – Aggregate Sum Area		
area_id	category	shape_area_sum
1234	1	27
1234	2	17
1234	3	20

Step 3 – Category/Total Sum	
area_id	shape_area_sum/ category
1234	27
1234	8.5
1234	6.66...

Step 4 – BFPL Density		
area_id	SUM(shape_area_sum/ category)	bfpl_density
1234	42.16666666	42.16666666/land_area

Now after total aggregate area has been calculated by category. Each shape_area_sum is divided by its category for each corresponding area_id. From domain knowledge it is given that the category 1 for a region is considered highest risk. Therefore, dividing by 1 would keep the area value unchanged however the others are reduced accordingly. Since category 3 is medium and category 2 is considered lowest risk, we divide the area by swapping the category such that areas of category 2 are divided by 3 and areas of 3 are divided by 2. This ensures the validity of the computation.

When the aggregate sum of the shape_area/category is taken it can now be divided by the corresponding land_area found in neighbourhoods to compute reasonable bfpl_density. Stage 1 and 3 recorded on two different datasets (rfs_sa2_join and bfpl_density) for analysis when necessary. Therefore, the final fire risk formula computes to:

$$bfpl_{density} = \sum(\frac{\sum(area\ per\ category)}{category})/land\_area$$

$$water_{density} = water\_area/land\_area$$

$$Fire\ Score = S((z(population_{density}) + z(dwelling_{density} + business_{density}) + z(bfpl_{density}) - z(assistiveServices_{density}) - z(water_{density}\ )\ /AVG(forecast_{10days})))$$

Where S is the sigmoid function and z is the zscore for the respective column.

The water density is found by joining the sa2_2016_aust and neighbourhoods to determine areas that are touching have water lands. This water land area is then divided by the total land area to produce water_density. Average of forecast was taken by the transformed categorical water predictions in where Sunny is 1 (highest risk) and heavy rain (8) lowest risk. Although this poses limitations and bounds on the average due to the restricted 10-day period being considered.

FIRE RISK ANALYSIS OVERVIEW

With the adjusted fire score formula, fire scores are dependent on density variables as their value directly influences the exponential or depreciated fire score. The introduced variables water density and average of forecast data (10 days) either reduce or increase resulting fire score. Water density is subtracted as it is assumed to be risk-preventing value as water bodies prevent the rapid propagation of fire-spreading (Understanding Bush Fires, 2016). With motive to improve accuracy of fire-risk score we calculate the average forecast score that resides within the bound (10 – 90). Where score 10 would result from Sunny weather for 10 days etc. This average is divided by total density computation to estimate an adjusted fire risk score for the following 10 days. (Summary and Visualisations)

The linked graphs and statistics conclusively indicate the constricting of fire-score before additional datasets. Indicating a direct influence on the accuracy of the fire-risk scores. Also, form inference it is evident that strong clustering within 0.7-0.9 is unlikely since Western Sydney is surrounded by water lands and water basin regions. The adjusted firescore presents these findings accurately as clustering of the homogenous data within the 0.4 to 0.6 risk range is considerably meaningful for the context of the data and analysis. Although median and maximum haven’t drastically improved the standard deviation and quartile ranges have been narrowed to reflect accuracy in the lower quartiles.

CORRELATION ANALYSIS

Raw Fire-Score Analysis

Pearson coefficient hypothesis:

- H_0: p=0, mean there is no correlation (null hypothesis)
- H_1: p≠0, mean there is correlation (alternative hypothesis)

Fire-score vs Median Income – Using Pearson test, the correlation result between "neighbourhood's median income" and the "fire risk" is 0.14 (weak positive correlation), the p-value (0.012) less than 0.05 as our alpha, suggest that we should reject the null hypothesis, which mean there is correlation between "neighbourhood's median income" and the "fire risk". We try to fit "fire risk" in the x-axis or as the predictor and "neighbourhood's median income" in the y-axis in a linear model ($y = 4.77 \times 10^4 x + 5461.1$). From the equation we know that an increase of fire risk score by 1 will increase the neighbourhood's median income by 4.77×10^4. We also observe the RMSE which is the average difference between the observed value with the expected outcome is 8486.6 which is high and R-square which is the goodness of fit between the data and the model is 0.020 which is low. The scatterplot shows that there is no clear pattern and the steepness of the regression line is relatively flat.

Fire-score vs Average Rent - Using Pearson test, the correlation between "neighbourhood's average rent " and "fire risk score" is 0.04 (weak positive correlation), which much lower than the correlation that we got with the "neighbourhood's median income" with "fire risk score", the p-value (0.43) more than 0.05 as our alpha, suggest that we should retain the null hypothesis, which mean there is no correlation between "neighbourhood's average rent" and "fire risk score". We put "fire risk score" in the x-axis and the "neighbourhood's average rent" in the y-axis of a linear model $y = 95.3x + 1877$, we observe that increase of fire risk score by 1 will increase the "neighbourhood's average rent" by 95.3, the model RMSE is 481.3 which is high and the model R-square is 0.002 which is low. Scatterplot show that there is no clear pattern and the steepness of the regression line

Adjusted Fire-Score Analysis

Adjusted Fire-score vs Median Income - We make a new "fire risk score" formulation with our extra dataset. Subtracting computation of the exponential power of the old "fire risk score" with "zscore water density" and divide it with "average forecast in 10 days" give us a new fire risk score named "fire risk score extra". We repeat the step of our analysis by finding the correlation using Pearson coefficient, between "fire risk score extra" with the "neighbourhood's median income" is 0.166 which indicate positive weak correlation furthermore using the p-value provided is 0.002 less than 0.05 as our alpha, we reject the null hypothesis and conclude there is correlation between "neighbourhood's median income" and the "fire risk extra". We fit in the data with linear model ($y = 3.8 \times 10^4 x + 2.4 \times 10^4$) which mean if the "fire risk score extra" increase by 1 the "Neighbourhood's average rent" will increase by 3.8×10^4 and observe the RMSE is 8443.2 which indicate large different between observed and expected data and R-square is 0.030 which indicate weak goodness of fit to the model. Scatterplot doesn't show clear pattern and the regression line is relatively flat.

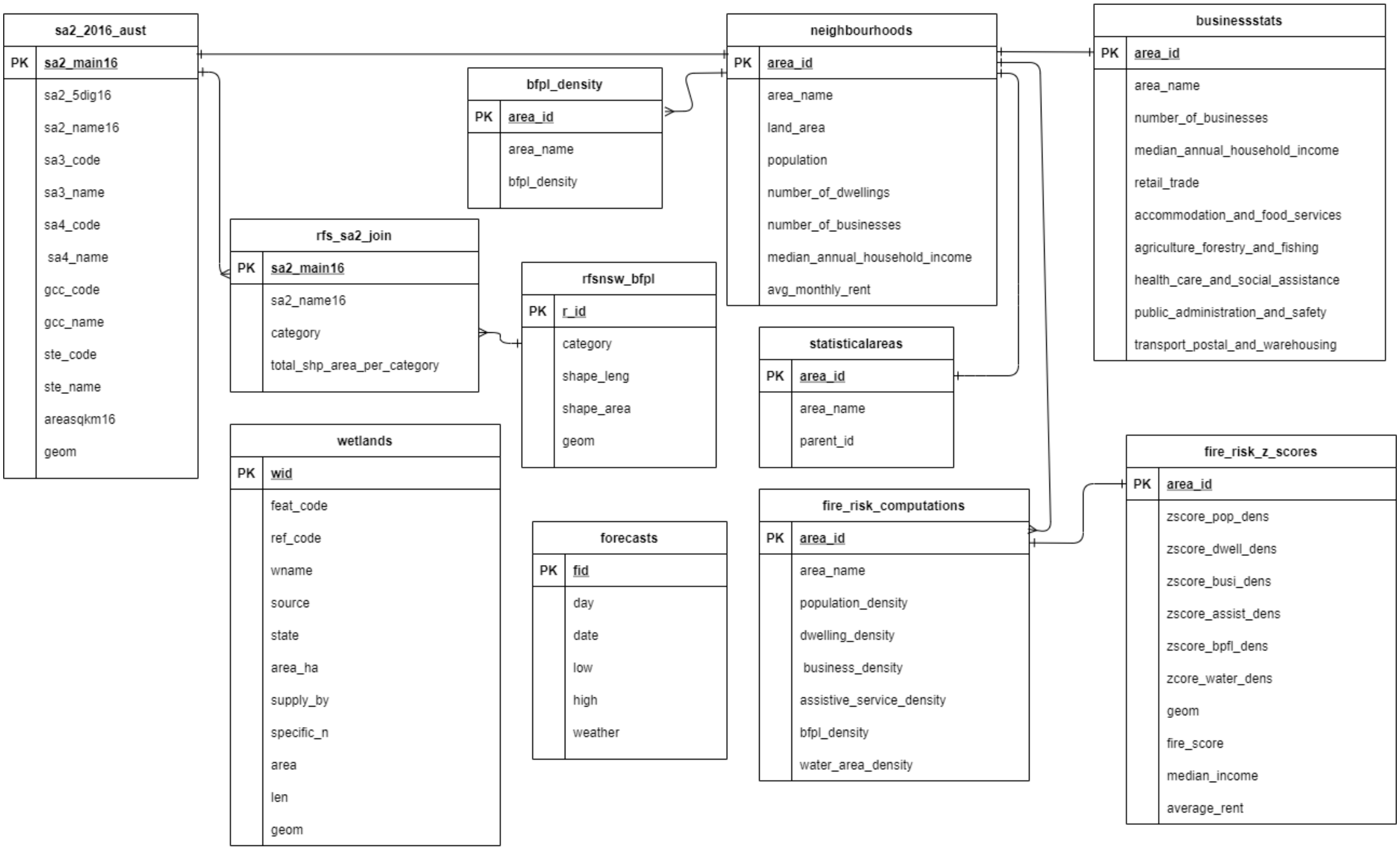
Adjusted Fire-score vs Average Rent - The correlation result between “neighbourhood’s average rent” and “fire risk score extra” is 0.031 which lower than the correlation that we got with the “neighbourhood’s median income” with “fire risk score extra” furthermore using the p-value provided is 0.57 more than 0.05 as our alpha, we reject the null hypothesis and conclude there is no correlation between “neighbourhood’s average rent” and the “fire risk extra”. We put “fire risk score extra” in the x-axis and the “neighbourhood’s average rent” in the y-axis of a linear model ($y = 178.7x + 1832.4$), we observe that increase of fire risk score by 1 will increase the “neighbourhood’s average rent” by 178.7, the model RMSE is 481.6 which indicate large different between observed and expected data and the model R-square is 0.001 which indicates weak goodness of fit to the model.

CONCLUSION

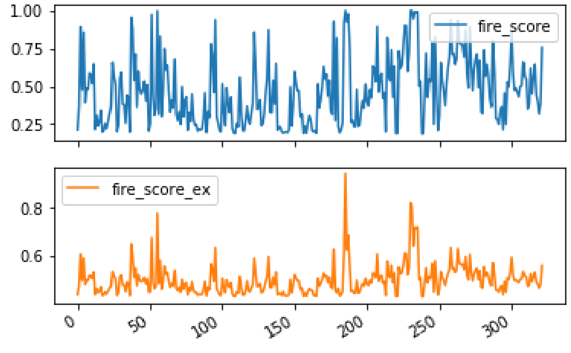
In summary, the fire score analysis for the Greater Western Sydney regions provides an insight into the correlations that exist between computed fire-scores, median income and average rent. The analysis considers the influence of environmental components that may deem computed fire-scores inaccurate. With spatial relations and weather forecast data integrated. Adjusted and raw fire-scores show that there exists a weak-positive correlation with median income while average rent shows no correlation as the p-value is greater than 0.5 threshold hence alternative hypothesis is considered.

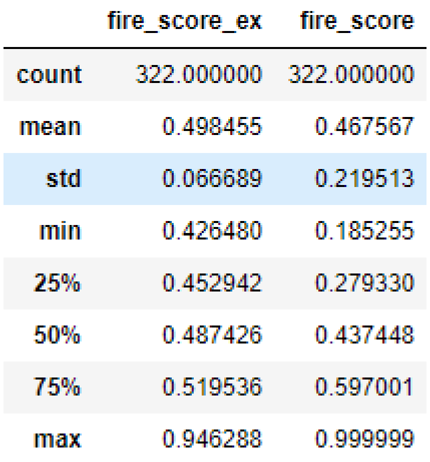
APPENDIX

Database Schema

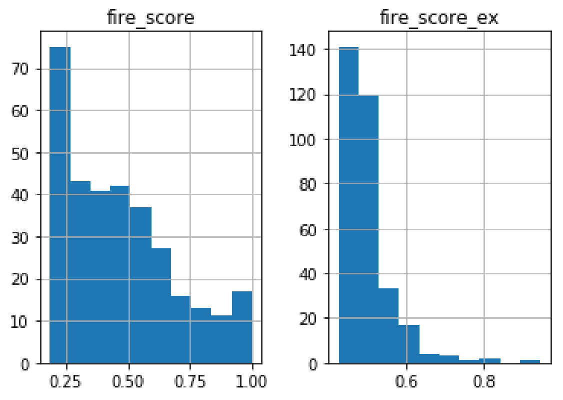


Fire Risk Analysis

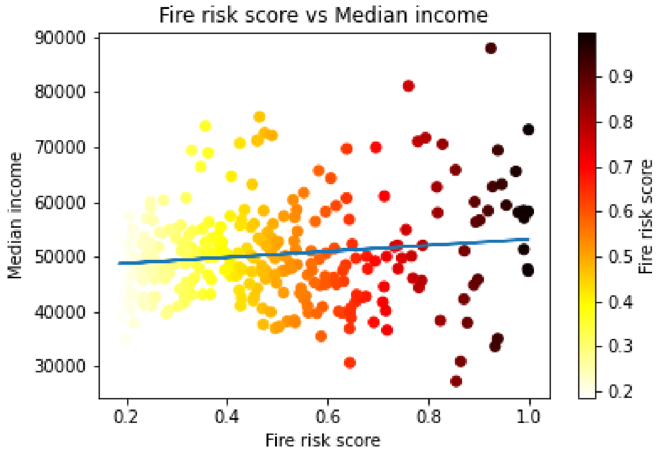


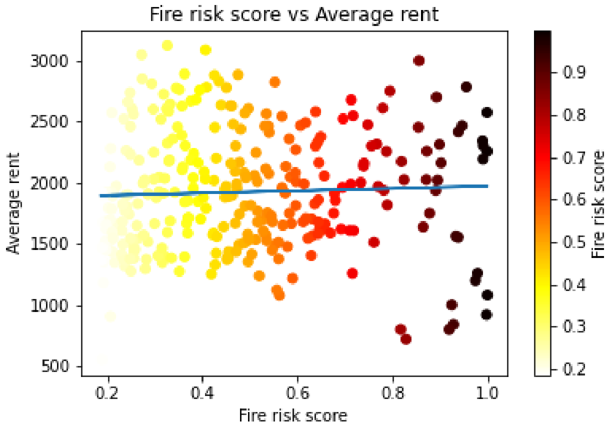


	fire_score_ex	fire_score
count	322.000000	322.000000
mean	0.498455	0.467567
std	0.066689	0.219513
min	0.426480	0.185255
25%	0.452942	0.279330
50%	0.487426	0.437448
75%	0.519536	0.597001
max	0.946288	0.999999

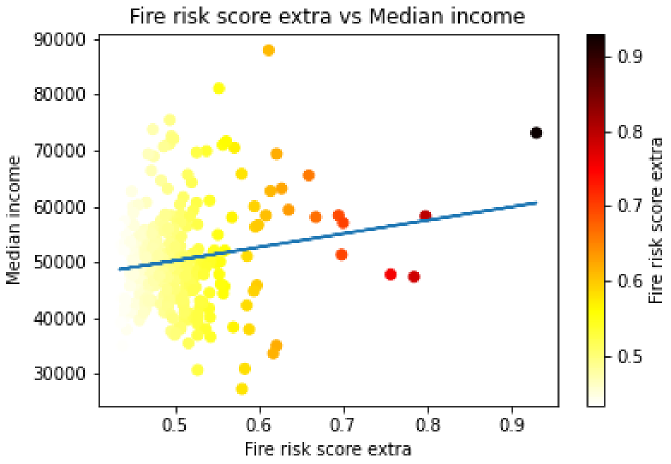


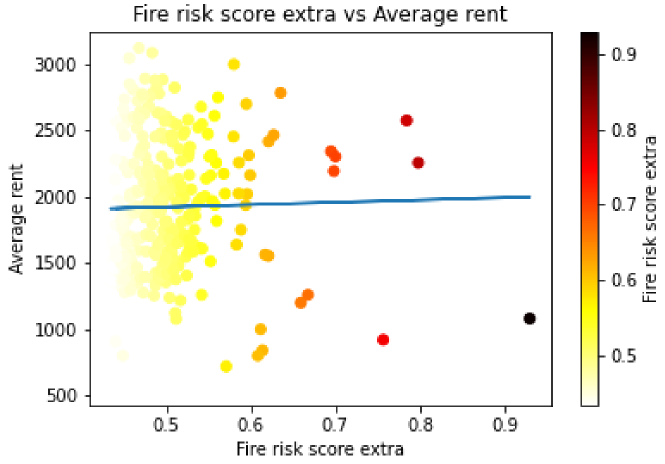
Fire-score vs Median Income and Average Rent





Adjusted Fire-score vs Median Income and Average Rent





SOURCES

ABS Australian Statistical Geography Standard (ASGS). (2016). SA2 Areas Summary

https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/1270.0.55.001~July%202016~Main%20Features~Statistical%20Area%20Level%202%20(SA2)~10014

RFS BFPL (2015). Guide for Bushfire Prone Land Mapping

https://www.rfs.nsw.gov.au/plan-and-prepare/building-in-a-bush-fire-area/planning-for-bush-fire-protection/bush-fire-prone-land

ADDITIONAL DATASET 1 - Department of the Environment and Energy (2018) Directory of Important Wetlands in Australia (DIWA) Spatial Database

https://data.gov.au/data/dataset/6636846e-e330-4110-afbb-7b89491fe567

Related article on implementation of Technology hub for poverty suffering regions of Africa

ADDITIONAL DATASET 2 - RAPID API (Live JSON) – Yahoo Weather Forecast API

https://yahoo-weather5.p.rapidapi.com/weather

Understanding Bush Fires (2016) – Science Org Bush Fires - https://www.science.org.au/curious/bushfires