THE UNIVERSITY OF SYDNEY

School of Computer Science
Uwe Roehm

DATA2001/DATA2901: Data Science, Big Data, and Data Diversity **1.Sem./2021**

Practical Assignment: Bushfire Risk Analysis

Group Assignment (20%) **10.05.2021**

Introduction

In this practical assignment of DATA2001/DATA2901 you are asked to gather and integrate several datasets to perform a data analysis of the *bushfire risk* of different neighbourhoods in Sydney.

You find links to online documentation, data, and hints on tools and schema needed for this assignment in the 'Assignments' section in Canvas.

Disclaimer: This assignment is mainly about data integration. Note that the age and varying quality of the provided data do not allow to reliably assess the actual bushfire risk.

Data Set Description and Preparation

Your task in this assignment is to calculate a bushfire risk score with regard to bushfire protection for different neighbourhoods in Sydney. The neighbourhood 'score' is expressed as a measure of several factors which we *assume* to affect the risk of bush fires within an area — vegetation, population density, number of dwellings etc.

In order to calculate this score, you will need to integrate different data sources. As a starting point, we provide you with some census-based datasets which give you input on at least three factors: population density, dwelling and business locations. We also provide some spatial data with the vegetation and risk categories provided by the NSW Rural Fire Service. We leave it up-to you to integrate further data and to refine the suggested risk score. Some ideas would be the availability of specific emergency services, or the prevalence of waterways etc.

Based on your computed risk scores, perform then a correlation analysis against the ABS provided median income and median rent costs of each neighbourhood.

Your submission should consist of your Jupyter notebook that you used for integrating the data sets and for performing and visualising your analysis.

Milestone 1: Load and integrate the provided datasets into the university provided PostgreSQL database by the tutorials in Week 11.

Provided datasets: We provide in Canvas several CSV files with Statistical Area 2 (SA2) data from the Australian Bureau of Statistics (ABS), as well as some bush fire prone land vegetation spatial data from the NSW Rural Fire Service (keep checking Canvas/Ed for any later additions or updates at `https://edstem.org/courses/5592/discussion/462995`):

`StatisticalAreas.csv:`	`area_id, area_name, parent_area_id`
`Neighbourhoods.csv:`	`area_id, area_name, land_area, population, dwellings, businesses, median_income, av`
`BusinessStats.csv:`	`area_id, number_of_businesses, accommodation_and_food, retail_trade, agriculture_for`
`RFSNSW_BFPL.shp:`	`gid, category, shape_leng, shape_area, geom`

Task 1: Data Integration and Database Generation

Build a database using PostgreSQL that integrates data from the following sources:

1. Sydney neighbourhood dataset (based on provided CSV files with SA2-data from ABS).
2. Spatial data in the SA2 ESRI Shape data file from the ABS at `https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.001July%202016`)
3. Census data for the given neighbourhoods including population count, dwelling and businesses counts.
4. Bush Fire Prone Land in NSW; Originally from the Rural Fire service, but modified for this task - you will need to do some transforming of this data
5. **You are encouraged to extend and refine both scoring function and source data**. For full points when integrating at least one additional data set.

Milestone 1: Load and integrate the provided datasets into PostgreSQL by the tutorials in Week 11.

Task 2: Fire Risk Analysis

1. Compute the fire risk score for all given neighbourhoods according to the following formula and definitions (adjust as needed if you integrated any additional datasets):

$$\mathit{fire\_risk} = S(z(\mathit{population\_density})+z(\mathit{dwelling\_\&\_business\_density})+z(\mathit{bfpl\_density})-z(\mathit{assistive\_service\_density})$$

With S being the logistic function (sigmoid function), and z the *z-score* ("standard score") of a measure - the number of standard deviations from the mean (assuming a normal distribution):

$$z(measure, x) = \frac{x - avg_{measure}}{stddev_{measure}}$$

Measure	Definition	Risk	Data Source
$population\_density$	population divided by neighbourhood's land area	+	`Neighbourhoods.csv`
$dwelling\_density$	number of dwellings divided by neighbourhood land area	+	`Neighbourhoods.csv`
$business\_density$	number of businesses divided by neighbourhood land area	+	`BusinessStats.csv`
$bfpl\_density$	area and category of BFPL divided by neighbourhood land area	+	`RFSNSW_BFPL.shp`
$assistive\_service\_density$	number of assistive services divided by neighbourhood land area	-	`BusinessStats.csv`

2. Store the computed measures and scores of each neighbourhood in your database. **Create at least one index** which is helpful for data integration or the fire risk score computation.
3. Determine whether there is a correlation between your fire risk score and the median income and rent of a neighbourhood.

Task 3: Documentation of your Bushfire Risk Analysis

Write a document (Jupyter notebook or Word document or PDF file, no more than 5 pages plus optional Appendix) in which you document your data integration steps and the main outcomes of your fire risk data analysis, including the correlation study with the bush fire statistics. Your document should contain the following:

1. **Dataset Description**
   What are your data sources and how did you obtain and pre-process the data?
2. **Database Description**
   Into which database schema did you integrate your data (preferable shown with a diagram)? Which index(es) did you create, and why?

3. **Fire Risk Score Analysis**
   Show which formula you applied to compute the Fire Risk score per neighbourhood, and give an overview of fire risk results. This can be done either in text by highlighting some representative results, or with a graphical representation onto a map (preferred).
4. **Correlation Analysis**
   How well does your score correlate to the affluence of the neighbourhoods? Compare both the median household incomes and the rental prices of each region.

Task 4: DATA2901 Task for Advanced Class Only

1. For teams in the advance class, integration of at least one additional data set is compulsory.
2. One of the additional data sources must come from a web source such as be Web Scraping or using a Web-API, rather than just a downloadable additional CSV data set.
3. Include in the fire risk analysis some data that was inferred using a machine learning or natural language processing step. **NOTE: This may change depending on the difficulty of the task, please keep apprised with the clarification thread for details -** `https://edstem.org/courses/5592/discussion/462995`

General Coding Requirements

1. Solve this assignment with a Python Jupyter notebook in Python and SQL (Adv: also Unix).
2. Use the provided Jupyter and PostgreSQL servers from the tutorials.
3. If you use any extra libraries which are not installed in the labs, disclose in your documentation which library and what version.

Deliverables and Submission Details

There are four deliverables:

1. **source code** of the data integration and analysis tasks,
2. a brief **report/documentation** (up to 5 pages, as of content description above), and a
3. **short demo** in the labs of Weeks 12 and 13 with the whole team present.
4. Please also provide **access to your database** with the schema and the processed data.

All deliverables are due in Week 12, no later than **8pm, Friday 28 May 2021**. Late submission penalty: -20% of the awarded marks per day late. The marking rubric is in Canvas.

Please submit the source code and a soft copy of your documentation as a zip or tar file electronically in Canvas, one per each group. Name your zip archive after your $Class$ and group number X with the following name pattern: **data2001_assignment2021s1_*Class*_group*X*.zip**

Students must retain electronic copies of their submitted assignment files and databases, as the unit coordinator may request to inspect these files before marking of an assignment is completed. If these assignment files are not made available to the unit coordinator when requested, the marking of this assignment may not proceed.

All the best!

Group member participation

This is a group assignment. The mark awarded for your assignment is conditional on you being able to explain any of your answers to your tutor or the lecturers if asked.

If members of your group do not contribute sufficiently you should alert your tutor as soon as possible. The tutor has the discretion to scale the group's mark for each member as follows, based on the outcome of the group's demo in Week 12 or 13:

Level of contribution	**Proportion of final grade received**
No participation or no demo.	0%
Passive member, but full understanding of the submitted work.	50%
Minor contributor to the group's submission.	75%
Major contributor to the group's submission.	100%