School of Computer Science
Uwe Roehm

DATA2001/DATA2901: Data Science, Big Data, and Data Diversity **1.Sem./2020**

Practical Assignment: Viral Vulnerability Analysis

Group Assignment (20%) **06.05.2020**

Introduction

In this practical assignment of DATA2001/DATA2901 you are asked to gather and integrate several datasets to perform a data analysis of the *viral vulnerability* 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 COVID19 risk.

Data Set Description and Preparation

Your task in this assignment is to calculate a vulnerability score with regard to infectious diseases for different neighbourhoods in Sydney. The neighbourhood 'vulnerability' is expressed as a measure of several factors which we *assume* to affect the spread of a virus within a community — population density, age distribution, pre-existing health conditions, and access to healthcare services.

In order to calculate this score, you will need to integrate different data sources. As a starting point, we provide you with a few census-based datasets which give you input on at least three factors: population density, age distribution, and locations of health services (hospitals and GPs). We leave it up-to you to integrate further data and to refine the suggested vulnerability score. Some ideas would be percentage of population with pre-existing health conditions such as asthma or diabetes, presence of meeting hotspots such as large shopping centres or sports venues, intensity of international travel (either by locals there or by tourists in an area), or public transport usage.

Based on your computed vulnerability scores, perform then a correlation analysis with the official COVID-19 data per neighbourhood as provided by NSW Health (also provided, resp. linked).

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 postgres 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 health service location data from Sydney (keep checking Canvas for any later additions or updates):

`StatisticalAreas.csv:`	`area_id, area_name, parent_area_id`
`Neighbourhoods.csv:`	`area_id, area_name, land_area, population, dwellings, businesses, median_income, av`
`PopulationStats2016.csv:`	`area_id, area_name,` *`age_distribution`*`, total_persons, females, males`
`HealthServices.csv:`	`id, name, category, num_beds, address, ..., longitude, latitude, comment`
`NSW_Postcodes.csv`	`id, postcode, locality, longitude, latitude`
`COVID-19 Statistics`	recent daily data can be accessed from data.gov.au
	e.g.: `https://data.gov.au/dataset/ds-nsw-5424aa3b-550d-4637-ae50-7f458ce327f4`

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. Census data for the given neighbourhoods including population count and age distributions.
3. Health services in NSW; **Todo:** spatial join with neighbourhoods.
4. **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: Viral Vulnerability Analysis

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

$$vulnerability = S(z(population\_density) + z(population\_age) - z(healthservice\_density) - z(hospitalbed\_density))$$

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	+	`nNeighbourhoods.csv`
$population\_age$	percentage of a neighbourhood's population age 70+	+	`PopulationStats2016.csv`
$healthservice\_density$	number of health services per suburb per 1000 people	−	`HealthServices.csv`
$hospitalbed\_density$	number of hospital beds per suburb per 1000 people	−	`HealthServices.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 vulnerability score computation.
3. Determine whether there is a correlation between your viral vulnerability score and the number of COVID-19 tests or COVID-19 cases (positive tests) per neighbourhood.

Task 3: Documentation of your Viral Vulnerability 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 vulnerability data analysis, including the correlation study with the COVID-19 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. **Vulnerability Score Analysis**
   Show which formula you applied to compute the vulnerability score per neighbourhood, and give an overview of vulnerability 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 number of COVID-19 cases in the given suburbs? Is there any correlation with the number of COVID-19 tests in the neighbourhoods?

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 vulnerability analysis some data that was inferred using a machine learning or natural language processing step. For example, you could retrieve and count named entities from the scrapped content of a website about international visitors or travel infrastructure in different neighbourhoods in Sydney, or you could try to train a neighbourhood classifier.

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 Week 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 13, no later than **8pm, Friday 29 May 2020** (extended). 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 group number X with the following name pattern: **data2001_assignment2020s1-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 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%