THE UNIVERSITY OF SYDNEY

School of Computer Science
Uwe Roehm

DATA2001/DATA2901: Data Science, Big Data, and Data Diversity **1.Sem./2019**

Practical Assignment: Neighbourhood Cyclability Analysis

Group Assignment (20%) **18.04.2019**

Introduction

In this practical assignment of DATA2001/DATA2901 you are asked to gather and integrate several datasets to perform a data analysis of the cyclability 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.

Data Set Description and Preparation

Your task in this assignment is to calculate a cyclability score for different neighbourhoods in Sydney similar to the 'WalkScore' service (`https://www.walkscore.com/AU-NSW/Sydney`). According to Wikitionary, "cyclability is the extend to which [an area] is cyclable". The neighbourhood cyclability is expressed as a measure of the "five D's" — density, diversity, design, destination accessibility and distance to cycling infrastructure or transport. Obviously, a higher cyclability has benefits for health, the environmental, and probably also adds to property values (which we want to check).

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 'Ds': density, destinations and presence of bike-sharing stations (the latter two only to some extend). We leave it up-to you to integrate further data and to refine the suggested cyclability score. Some idea would be availability of cafes, restaurants or shopping centres, structure of the road network, presence of cycling infrastructure, connectivity with public transport or amount of public parks and trees.

Based on your computed cyclability scores, perform a correlation analysis with the median income and the average monthly rent per neighbourhood according to the provided Census data from the Australian Bureau of Statistics (ABS) (also provided).

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

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 (synthetic) bike-sharing data from Sydney (keep checking Canvas for any later additions such as bounding boxes of neighbourhoods):

`StatisticalAreas.csv:`	`area_id, area_name, parent_area_id`
`Neighbourhoods.csv:`	`area_id, area_name, land_area, population, number_of_dwellings, number_of_business`
`CensusStats.csv:`	`area_id, median_household_income, avg_monthly_rent`
`BusinessStats.csv:`	`area_id, num_businesses, retail_trade, accommodation_and_food, health_care, ...`
`BikeSharingPods.csv:`	`station_id, name, num_bikes, num_scooters, latitude, longitude, description`

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 number of dwellings.
3. Cycling options in terms of bike sharing pods; **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.

Task 2: Cyclability Analysis

1. Compute the cyclability score for all given neighbourhoods according to the following formula and definitions (adjust as needed if you integrated any additional datasets):
   $cyclability = z(population\_density) + z(dwelling\_density) + z(service\_balance) + z(bikepod\_density)$

   With z being 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**	**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`
$service\_balance$	*balance* of selected business types in neighbourhood	`BusinessStats.csv`
$bikepod\_density$	number of bike-sharing pods per suburb divided by area	`BikeSharingPods.csv`

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

Task 3: Documentation of your Cyclability Analysis

Write a document (Jupyter notebook or Word document or PDF file, no more than 4 pages plus optional Appendix) in which you document your data integration steps and the main outcomes of your cyclability data analysis, including the correlation study with the household income. 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. **Cyclability Analysis**
   Show which formula you applied to compute the cyclability score per neighbourhood, and give an overview of cyclability results. This can be done either in text by highlighting some representative results, or with a graphical representation onto a map.
4. **Correlation Analysis**
   How well does your score correlate to the median household income in the given suburbs? Is there any correlation with the average weekly rent 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 cyclability-score calculation 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 planned cycling infrastructure in different neighbourhoods in Sydney.

General Coding Requirements

1. You should solve this assignment with a Python Jupyter notebook in Python and SQL.
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 4 pages, as of content description above), and a
3. **short demo** in the labs of Week 12 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 **6pm, Friday 24 May 2019**. Late submission penalty: -20% of the awarded marks per day late. See also the published marking rubric 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 UniKey: *abcd1234.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:

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%

References:

”Neighbourhood Walkability in NYC” URL: `https://beh.columbia.edu/neighborhood-walkability`
”Walkability Research” URL: `https://www.walkscore.com/professional/walkability-research.php`