Assignment 2: COMP20008 Data Science Project

March 3, 2021

Due Dates

- Friday 19th March - Group Formation
- Wednesday 14th April - Project Proposals
- Thursday 22nd April - Peer Feedback
- Friday 21st May - Final Report

Objectives

The objectives of this assignment are

- To gain an appreciation of the variety of open datasets that are available and what data wrangling / processing methods might be applied to them.
- To gain experience in practical data wrangling with real-world datasets.
- To build e-portfolio as a data scientist in training.
- To gain experience in written communication of analysis and results on a topic related to data science.
- To practice the initial ”pitch” phase of a data science project.
- To practice giving evaluation and feedback to projects.

Hypothetical Scenario and Objective

The Victorian Department of Innovation wishes to understand how it can use data to improve the services offered to Victorians. Student teams are invited to work on a project that aligns with the theme *Using open data for understanding the liveability, inclusiveness, health and sustainability of communities in Victoria.*

You are part of a team of four for this project. Each team is asked to design and undertake a pilot investigation using open datasets, which aligns with the above theme. At the end of the project, each team is to produce a report on the project.

The purpose of the project is to facilitate and showcase demonstrations of ways in which open data can be wrangled to help understand issues which are important for communities within Victoria. These may be either in terms of liveability, health, inclusiveness, or sustainability.

Your team must propose a project with a question that aligns with the above theme and identify at least 2 open datasets that could assist in answering the above question. The proposal should also explain which wrangling techniques could be applied to these open datasets to address your chosen question, what analysis will be performed and what information might be revealed, why that might be interesting for addressing the question and who might care.

You and your team members will need to:

1. Produce a project proposal. For your proposal, you will need to:
    - Choose an objective that you would like to investigate (liveability, inclusiveness, health or sustainability).
    - Select a community granularity (all of Australia, Victoria, Melbourne, a particular region with Victoria, a collection of regions, ..)
    - Propose a question relating to your objective. Some questions may relate to multiple objectives, but this isn't a requirement.
    - Identify which audience would be interested in an answer to this question and why.
    - Identify at least 2 open datasets (or more) that can used together to help shed light on this question.
    - Propose a list of data wrangling and analysis tasks on these datasets. This includes an explanation of what techniques (data pre-processing, transformation, integration, correlation, analysis, visualisations and prediction) would be applied, what their purpose would be and what information they could reveal to address the question.
2. Perform the various tasks in your proposal.
3. Report on the data science project at the end of the project.

Examples of questions

You have considerable freedom in selecting a question. The key aspect is to explain how it connects with understanding any of the aspects of liveability, or inclusiveness, or health or sustainability of communities in Victoria. Some indicative examples that provide a flavour (n.b. these are not exhaustive):

- Is there sufficient health services availability for residents in Victoria? (health)
- Which suburbs in Melbourne need new sports facilities? (inclusiveness, health)
- Is population growth in Victoria impacting access to transport? (sustainability, liveability)
- Is there a relation between household income and access to education across Victoria? (inclusiveness)

- Is access to green spaces an important factor in the success of Melbourne's restaurants? (liveability)
- What are the key factors influencing the adoption of recycling technology in Melbourne? (sustainability)
- Is there a correlation between the amount of pollution and socio-economic level in rural regions across Victoria? (liveability, inclusion, health)

Datasets

The LMS Project page has a list of repositories that can be used as a starting point for finding datasets. Data from any of these is fine to use.

You are also welcome to use other datasets that have been made publicly available by reputable entities, or which are readily available via a registration process open to University of Melbourne staff/students.

You should not select datasets that have been illegally obtained or published (E.g. data violating copyright permissions or that has been hacked).

If in any doubt as to whether a particular dataset is ok to select, please post a question on the discussion forum or contact the subject Head Tutor for clarification.

Git repository

All of the code you develop as part of this project should be stored in a GitHub repository. As with assignment 1, you should visit the following link to create your GitHub repository: https://classroom.github.com/g/Dk6Dm5sv

Only one member of your group should create a GitHub repository, the other group members should be added to the same GitHub repository. This will ensure that all group members are able to collaborate on the same codebase.

Assessment 30 marks

Proposal (4 marks), Due Wednesday 14th April

These 4 marks are awarded as long as your team submitted your own proposal which included relevant information for the elements of a proposal. The marks are not based on the quality of your proposal, but a good proposal will set out a logical structure for your final project report (which is marked). You will receive feedback on your proposal which is designed to help focus your research and improve your final report.

Your proposal should be roughly 600 - 800 words long and should cover the following points

1. What is the research question and how is it related to the theme of *understanding the liveability, inclusiveness, health and sustainability of communities in Victoria*?

2. Why is it worth tackling (i.e. motivation) and who would care? Who are the stakeholders in this project? In what respects might it provide innovative information? (you do not want to have a question which is trivial, or for which the answer already publicly exists and can be readily found)?

3. What are the two or more open datasets you could use and why? What is their format, size and what information do they contain? Can they be linked together, and if so how?

4. What data wrangling and analysis methodologies could you use to investigate your research question? Be specific about how they could be applied on your selected datasets.

5. What could be achieved by using these data wrangling methodologies? What would be output/product of the wrangling (type of data, graph, table, statistic(s), ...)? How will this add value compared to having just the raw data?

6. What might be the challenges and risks for undertaking this work?

Peer Review (4 marks), Due Thursday 22nd April

After the week 7 submission, each individual student will be provided with the proposals from two other student groups. Your task is to evaluate those reports and provide peer-feedback to help those groups undertake their research. You will be assigned proposals to review and enter your feedback in LMS. Your feedback should cover the following points:

Content

- Does the report address points 1-6 above in a clear, logical, comprehensive and persuasive way?

- How clear, well defined and specific are the topic and question?

- How clear is the description of the proposed datasets and does it help the reader to easily get a better understanding of the data?

- Would the proposed investigation be likely to generate new information? (An example of an investigation which wouldn't, is one where the answer to the question is straightforwardly on a web page/document/article found via a Google search).

Data Wrangling & Analysis

- Are the wrangling steps appropriate for the project?

- Are the analysis steps appropriate for the project?

- To what extent can the wrangling and analysis steps produce sound and conclusive results? Any major gaps?

Written Communication

- Did the proposal communicate the purpose and outcome?
- Is the proposal written in clear, concise language?
- Does the proposal explain the pipeline clearly?

As a rough guide, you should spend no more than one hour on each review.

After this peer review process you will be provided with the peer feedback of your proposal, as well as a tutor's feedback. As each individual student is evaluating two group proposals, this will result in a number of peer reviews of your group's proposal. We strongly encourage you to consider the feedback provided when undertaking your research and drafting your final report.

Final Report (22 marks) Due Friday 21st May

Your next task is to undertake the research task described in your proposal. You may find that your wrangling and analysis methods change as you proceed with the project, based on the feedback you've received on your proposal and the results you obtain in your initial investigation. This is to be expected and does not pose a problem. You will need to write a report on your findings.

Your final report should be no more than 1500 words in length excluding figures and tables. Your report should include the following information:

1. What is the research question and how is it related to the theme of *understanding the liveability, inclusiveness, health and sustainability of communities in Victoria*?
2. What are the datasets you've used and how have you linked them together?
3. What wrangling and analysis methods have you applied? Why have you chosen these methods over other alternatives?
4. What are the key results your research has obtained?
5. Why are your results significant and valuable?
6. What are the limitations of your results and how can the project be improved for future?

Your report should make effective use of visualisations to support your argument.

Submission Instructions

Your final report must be uploaded via Canvas by the due date. All of your code files, and any other supporting files used, should be placed in a .zip archive and uploaded via Canvas by the due date. It is essential that any numerical results or visualisations used in the final report can be reproduced by running your code. You must also include a link to your GitHub repository.

Your report, code files and any other supporting documentation must also be pushed to your git repository. You must ensure that the README file within your git repository contains the names and student IDs of each member of your group.

Other

Extensions and Late Submission Penalties: If requesting an extension due to illness, please submit a medical certificate before the submission deadline to the lecturer. If there are any other exceptional circumstances, please contact the lecturer with plenty of notice. Late submissions without an approved extension will attract a penalty of 10% of the marks available for that phase per 24hr period (or part thereof) that it is late. After 144 hours your submission will no longer be marked and you will receive 0 marks for that phase.

Academic Honesty

You are expected to follow the academic honesty guidelines on the University website
`https://academichonesty.unimelb.edu.au`

A project discussion forum has also been created on the subject LMS. Please use this in the first instance if you have questions, since it will allow discussion and responses to be seen by everyone.