COMP20008 Project Phase 3

V1.0: 10th September 2018

Due Dates

- Phase 3-A (5%): 11:59am 5th October. Submission will be via the LMS.
- Phase 3-B (10%): 5min oral presentation plus questions. Presentations will be scheduled to occur in Week 11 (8-Oct 2018 - 14-Oct 2018) during the workshop you are officially enrolled in. You will be asked to also submit the slides via the LMS within 30min of delivering the presentation.

Phases 3: Hypothetical Scenario and Objective

The Victorian Minister for Data Science wishes to understand more about how open data can be used to benefit Victorians.

At a high level, they would like to see demonstrations of how open data can be wrangled to gain insight into issues affecting Victoria, for a broad range of areas such as transport, health, business, education, tourism, the environment, communities, the arts, commerce, public amenities, employment, sport, usage of facilities, real estate, finance or urban planning.

You are a data science consultant who is hoping to convince the Minister about the benefits of open data. You will formulate a question in a chosen domain that is relevant to Victoria, and complete a brief pilot investigation using open datasets. In Phase 3-A, you will complete the investigation and deliver a written code outlining your methodology and findings. In Phase 3-B you will make a brief oral presentation about what you have done.

The aim of the project is to provide experience in processing, analysing and visualising real world datasets.

This project will be done individually. You will need to

- Choose a domain (education or sport or transport or the environment or health, etc)
- Propose a question for your domain that relates to Victoria and for which an answer would be likely to interest politicians or policy makers in the Victorian Government.
- Identify 2 open datasets (or more) that can be linked together to help shed light on this question.
- Using Python, process these datasets, integrate them and provide analysis and visualisations which help answer the question you have posed.

You are not expected to develop an interactive tool for browsing your chosen datasets. Rather, the results of your investigation can be reported as tables or graphs or static visualisations suitable for inclusion in a Jupyter notebook (Python code) or in slides of a Powerpoint presentation.

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.

Prof. Richard Sinnott will be giving a presentation to the class on 14 September about the AURIN platform, which contains an abundance of open datasets relevant to Victoria and which you are welcome to choose to use.

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

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

The selected dataset (or datasets) should have reasonable amount of data (i.e. hundreds of rows). It should contain a number of dimensions (categorical values) and several measurements (numerical values).

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

Phase 3-A: Concept Formulation and Investigation (5%)

Your task for this phase is to describe your domain, the question to be investigated and the 2 datasets that will be used to answer the question. You must also complete your investigation and deliver Python code (ipynb file) which has the results of any work you have done in wrangling your datasets. This includes data pre-processing, transformation, integration, correlation, analysis, visualisations and prediction. Include comments documenting the functionality in your Python code and also include comments explaining about any libraries or external code that you used. The implementation should be easy enough to follow by a tutor.

You should start your ipynb file with the following two items:

1. Title of Project (choose this according to your chosen domain and question)
2. What is the question you are seeking to answer?

You should also end your ipynb file with one paragraph (conclusion, up to 100 words): Explain how your results help answer the question that you proposed. What difficulties and challenges did you encounter in processing, integrating and visualising these datasets to answer your question?

Phase 3-B: Oral Presentation (10%)

In your workshop you will make a short (5 minute) presentation on what you have found. You will be expected to develop slides in powerpoint or pdf. Presentations will occur during the workshop you are registered in and held during Week 11 (8-14 October). The detailed schedule has been released on LMS.

Presentations should be 5 minutes in length plus 2 minutes time for questions.

The presentation should be a set of slides (between 5 and 10 in number) produced using Powerpoint or similar software and should include some visual material. Have your presentation ready to display on the computer using a PDF file. Do not include videos or animations or interact with external websites (instead could use screenshots or figures in your presentation).

In the presentation, you should cover the following points.

1. What is the research question?
2. Why is it worth tackling (i.e. motivation)?
3. What are the datasets you used and why?
4. What data wrangling methodologies have you used to investigate your research question?
5. What did you find? Why is it interesting? What have you learnt?
6. What have been the challenges and what (if anything) would you have done differently?

Assessment

In preparing your presentation, do not assume that your audience has seen your phase 3-A code submission. The criteria for assessment are:

- Did the presenter communicate the purpose and outcomes of the talk early enough?
- Did the presenter communicate the structure of the talk?
- Did the presenter include sufficient information in the presentation and address the 6 items above in their slides?
- Did the presenter show progression and connection between the parts of the presentation?
- Did the presenter use visual resources well?
- Did the presenter use language suited to the assessor?
- Did the presenter use voice well and make eye contact with the assessor ?
- Did the presenter allow sufficient time for the presentation?
- Did the presenter allow sufficient time for questions?

- Was the presenter familiar with the topic?
- Did the presenter handle questions well?
- Were the presentation slides clear ? (We will also look at your slides. Please submit your slides PDF file to the LMS within 30 minutes of your presentation completing.)

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.

E.g. A late submission for phase 3-A (5 marks total) will be penalised 0.5 mark if 4 hours late, 1 marks if 28 hours late, 1.5 marks if 50 hours late, etc.

Phase 3 is expected to require approximately 18-24 hours work.

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.