COMP20008 Project Phases 2-4

V1.0: 26th March 2018

Due Dates

- Phase 2 (10%): 11:59pm 23rd April. Submission will be via the LMS (Turnitin will be used).
- Phase 3 (13%): 11:59pm 8th May. Submission will be via the LMS (Turnitin will be used).
- Phase 4 (10%): 5min oral presentation plus questions. Presentations will be scheduled to occur in Week 11 (14-18 May) 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 (Turnitin will be used).

Phases 2-4: 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. In Phase 2, 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, you will complete the investigation and deliver a written report outlining your methodology and findings. In Phase 4 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 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 written report or in slides of a Powerpoint presentation.

Types of questions

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.

Richard Sinnott will be giving a presentation to the class on 28 March 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).

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 co-ordinator for clarification.

Phase 2: Concept Formulation and Preliminary Investigation (10%)

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 perform a preliminary investigation to provide evidence that the project will be feasible.

Please submit a pdf file of no more than 3 pages (11pt font. Top, bottom, left and right margins of at least 2cm) addressing the following items. You should answer each of them in the order 1, 2, ..., 8:

1. Title of Project (choose this according to your chosen domain and question)
2. Domain: Select either one of, or a combination of: transport, health, business, tourism, sport, education, the environment, communities, the arts, commerce, public amenities, employment, usage of facilities, real estate, finance or urban planning.
3. What is the question you are seeking to answer? Who would be interested in an answer to this question and why? How might the information be used and who could it benefit? (1 paragraph)
4. In what respects will your answer to this question provide innovative information? (you do not want to have a question which is trivial, or for which the answer already publically exists and can be readily found). (1 paragraph)
5. Datasets: What 2 datasets will you will use? Provide a short description of the information and a link (URI) to where the datasets can be downloaded.

6. In what ways will your processing, integration, analysis and visualization add value compared to having just the raw data? (2 paragraphs)

7. A short summary of the initial investigations and work you have done in wrangling your datasets to provide confidence that the plan is feasible (i.e. describe any summaries, initial transformations, visualisations and integrations you have performed and discuss their effectiveness). (Length=2 pages)

8. An explanation of why the remainder of the project is feasible and likely to yield interesting results, in light of your initial investigations (1-2 paragraphs).

Marking for Phase 2

Marking for this phase will include consideration of the following aspects:

- How clear, well defined and specific are the topic and research question? How well does the title indicate the purpose of the investigation?

- How clear is the provided summary of proposed datasets and does it help the reader to easily get a better understanding of the data? In the report, it is recommended to include a table summarising features of the data such as rows and columns. Such a table could consider aspects such as: What are the features and what types are there (continuous/categorical/binary/...)? What is the typical range of features? What percentage of the data contains noise/outliers?

- Analysis and results: Explanation of any data pre-processing performed. Why or why not was it done? Clear communication of the logic behind wrangling methods used - why done and with what purpose?

  In Q7, where findings are discussed, clear presentation and grouping of findings. Is context provided for results being shown (why is figure shown, what can be concluded)? It is important to separate speculation (guessing something might be true) from conclusions reached based on evidence.

- Completeness (answers provided to all questions), clarity and presentation (e.g. logical flow, definition of any terminology likely to be unfamiliar to an average student in COMP20008, readability, appropriate use of section headings, bullet points, labelling of all figures and tables, adherence to formatting guidelines).

- How persuasive and plausible are the arguments for Q4 and Q8?

- For Q4 and Q6, is the value that will be added clear?

- For Q7, the depth of work should be equivalent to 6-10 hours wrangling.

Phase 3: Report (13%)

In this phase, you must complete your investigation and then write a report of 4-5 pages (11pt font. Top, bottom, left and right margins of at least 2cm). Based on the feedback you received for Phase 2, or based on further investigations, you may need to make adjustments

to the Question you proposed in Phase 2. This is fine, provided you explain and justify the adjustments in the Phase 3 report. If planning to make a major adjustment to your Question, you are advised to check with either your workshop tutor, head tutor or the subject co-ordinator.

The format and approximate length of each section in the report is listed below.

1. *Title (1 sentence)*: The title of your project.
2. *Domain (1 sentence)*: The domain of your project (one or more of transport, health, business, tourism, sport, education, the environment, communities, the arts, commerce, public amenities, employment, usage of facilities, real estate, finance or urban planning.)
3. *Question (1 paragraph)*: A statement of the question you sought to answer.
4. *Datasets (2 paragraphs)*: A description of the datasets used. Explain what information was is in each dataset and what was included in the dataset schema. Provide URIs of where the datasets can be downloaded.
5. *Pre-processing (0.75 pages):* A description of the pre-processing methods you applied to the datasets and why. A description of any issues encountered and limitations of your methods.
6. *Integration (0.5 pages):* A description of how the datasets were integrated and a description of any issues encountered and limitations of your methods.
7. *Results (2 pages):* A description of how you did the analysis/visualisation of the integrated data and any issues encountered. Presentation and discussion of the results/findings (tables/graphs and discussion) and limitations of your methods or findings.
8. *Value (0.25 page)*: An explanation of how your pre-processing, integration and analysis/visualisation added value compared to having just the raw data. How do they help answer the Question?
9. *Challenges and Reflections (0.25 page)*: What difficulties and challenges did you encounter in processing, integrating and visualising these datasets to answer your question? Describe any unsuccessful efforts or any dead ends encountered and how you addressed these. (You could include mention here about any modifications you needed to make to your Question).
10. *Question Resolution (0.25 page)*: Explain how your results help answer the question that you proposed. Who might be interested in the results and why?
11. *Code (2 paragraphs):* How much code (in Python) did you write from scratch? What are the major Python libraries you used? What other other publically available code did you use? Where you used code other than Python, explain and justify why.
12. *Bibliography*: List any references that you have you used.

Code: Separate to the report, you are asked to also submit a zipfile of the Python code that you wrote yourself for the project. Include comments documenting the functionality in your Python code files and also include comments explaining about any libraries or external code that you used. Include a README file explaining the structure of the collection of code files.

Marking Scheme for Phase 3

Marking will include consideration of the following questions. A more detailed description of marking criteria for this phase will be released by mid April.

- To what extent has your data wrangling helped answer the question that was posed? How persuasive is your answer to the question? For projects where it proved difficult to answer the question, to what extent is it clear why it proved difficult/challenging?
- To what extent has your pre-processing, integration and analysis/visualisation added value to the raw datasets?
- To what extent does your solution make use of multiple technologies, methods and data representations presented in the subject (E.g. HTML, XML, JSON, data wrangling using Python, missing value methods, data transformation methods, outlier detection methods, visualisation methods, clustering methods, linkage and integration methods, correlation and feature ranking methods)? How complex is your use of these technologies, methods and data representations?
- How clear and complete is your report? To what extent is it evident that you are aware of the limitations of your findings? Does it adhere to the formatting guidelines?
- How well have you implemented your code in Python ? The implementation should be easy enough to follow by a tutor. How complete and clear is your Python code documentation?

Phase 4: 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 (14-18 May). The detailed schedule will be released by mid April.

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 read your phase 2 or 3 submissions. 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 to the Head Tutor. If there are any other exceptional circumstances, please contact the Head Tutor 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 2 (10 marks total) will be penalised 1 mark if 4 hours late, 2 marks if 28 hours late, 3 marks if 50 hours late, etc.

Phases 2-4 are expected to together require approximately 36-43 hours work.

Academic Honesty

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

Further Information

The Project page on the subject LMS contains various resources

- A list of frequently asked questions
- Pointers to datasets
- Some example projects that were submitted by students in 2016.

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.