COMP8430 – Data Wrangling - 2018

Data wrangling project **Due 11:55pm Sunday 21 October 2018**
Worth 30% of the final grade for COMP8430

Last update September 21, 2018

Overview and Objectives

This project is an opportunity to put everything you have learned about data wrangling into practice on a problem and data sets of your choice. You have to perform a similar set of tasks as in assignment 2. However the scope is much more open, and you are free to focus on aspects that you think are important or are of particular interest to you.

Important

- Submit **one zip archive file**, named **uNNNNNNN*_data_wrangling_project.zip***, where uNNNNNNN is your ANU ID. For example if your ANU ID is u1234567 you should submit the file *u1234567_data_wrangling_project.zip*. **Only use underscores** and not spaces, and **only lower-case letters** in your file name (as this will greatly help our marking efforts).
- The zip file must contain,
  1. Your report, a **.pdf** document named **uNNNNNNN*_data_wrangling_project_report.pdf***
  2. You must also submit your original (unmodified) and final data sets. Please combine them into a single **.zip** archive, named **uNNNNNNN*_data_wrangling_data_sets.zip***. Within this .zip archive you may name the files as you like, as long as it is clear which are the original data sets and which is the final combined data set.
- Make sure that **your student ID is included on the first page** of your submitted report.
- Do **NOT** include your name anywhere in your submission. All marking will be done anonymously.
- The maximum length of your report is **five (5) A4 pages** and **2000 words**. This **does** include any figures, tables, references and appendices. Include the total word-count of your report on the first page of your report.
- **Word documents or any other formats besides PDF are not accepted** for the report and will not be marked.
- **Hand-written submissions are not accepted** and will not be marked.
- Make sure you submit a **final** version of your project before the submission deadline.

Data Set and Problem Specifications

For this project **you get to choose the problem and data sets** you want to work with, but there are **a few requirements you must follow**:

- You must use at least two data sets. You can use more if you wish.
- You need to integrate the data sets so there must be some connection between them, such a common identifier, or attributes to perform record linkage, and so on.
- The largest data set must contain at least 1000 and at most 100,000 records. If it is larger, please use a subset of records, for example only use the records for a particular state, or a particular time period, etc.
- All your data sets must be publicly available.
- **You are not allowed to use the same data sets or problems that you have used for another ANU course**. If you feel that the overlap with previous work is only minor then please contact the course convener who may choose to make an exception. **Make sure you do this before commencing work on the project**.

Some other things to keep in mind:

- You are not required to perform record linkage if there are other means of integrating the data sets. However, should your problem require record linkage then please undertake it.
- You **are** allowed to re-use any code that you wrote in the data wrangling labs, and you may also make use of the record linkage program from the data wrangling labs.
- **You must choose and describe an intended end-use analysis of the data** and then perform your work for this project with regards to this end use. Make sure your end-use is appropriate for the data sets you have chosen.

Please give careful consideration to your choice of data sets and end-use, and pick something that **allows you to demonstrate your full capabilities**. A solution to a trivial problem will not likely get you many marks.

Submission

Submission will be done using Wattle. Click on the link COMP8430 data wrangling project submission (to be made available) in week 11 to upload your report. You may submit as many draft versions of your project as you wish. However, **you must make sure you submit a final version before the submission deadline**. We will mark the **final** version present at the due date. Note that **Wattle does not allow us to access earlier submitted versions of your project, therefore check carefully what you submit as the final version!**

Deadlines, Extensions and Late Submissions

The data wrangling project is due 11:55pm, Sunday 21 October, 2018.
Students will only be granted an extension on the submission deadline in extenuating circumstances, as defined by ANU policy (http://www.anu.edu.au/students/program-administration/assessments-exams/deferred-examinations). **If you think you have grounds for an extension, you must notify the course convener as soon as possible and provide written evidence in support of your case** (such as a medical certificate). The course convener will then decide whether to grant an extension and inform you as soon as practical.

In accordance with the CECS and ANU late submission policy, **no late submissions will be accepted**, except where an extension has been approved by the course convener.

Plagiarism

No group work is permitted for this project. We do encourage you to discuss your work, but we expect you to do the project work by yourself.

If you have any questions on the project please post them on Wattle – **however do not post any partial solutions, program codes, URLs, etc. or any hints on how to solve any of the questions.**

If you are unsure about what constitutes plagiarism, **make sure you read through the ANU Academic Honesty Policy** (http://academichonesty.anu.edu.au/).

If you do include ideas or material from other sources, then you clearly have to make attribution by providing a reference to the material or source in your report. We do not require a specific referencing format, as long as you are consistent and your references allow us to find the source, should we need to while we are marking your project.

Marking

This project will be marked out of 30. Note that not all project parts are equally difficult. **The project will count for 30% of your final course grade.** For many of the questions there is no right or wrong answer. Marks will be awarded based on your reasoning, and the justifications of your decisions and explanations.

We will endeavour to release your marks and feedback within **two teaching weeks** after the submission deadline. If you feel we have made an error in marking, you have **two weeks** following the release of marks to raise any issues with the course convener, after which time your mark will be considered final. **If you request that we re-mark your project, we will re-mark the entire project and your mark may go up or down as a result**.

Project Questions

You must pick two or more data sets (following the specifications described above), **along with a hypothetical end-use for the data**. Based on this end-use, you should explore, assess, integrate, and clean the data sets. You project must address the following questions:

1. **Data and problem description and overall strategy** (2 marks):
   - You must describe a hypothetical end-use of your data sets, since this should inform all the decisions you make for the rest of your data wrangling project.
   - You must also describe the overall strategy you have taken for your data wrangling project with reference to this end-use of the data.
2. **Data description and data exploration** (4 marks):
   - You must describe where each data set comes from, what its purpose is (if known) who collected it (if known) and when it was collected (if known).
   - You must perform data exploration and describe the general characteristics of each source data set (such as number of records, numer of attributes and their types, and so on).
   - Also consider, describe and include useful statistics about the source data sets, such as outliers and unexpected values, and any other features that are relevant to your chosen end-use.
3. **Data quality assessment** (8 marks):
   - Undertake an assessment of the data quality for each of the source data sets you are using. You should consider the data quality dimensions introduced in the lectures, as well as anything else that you think is relevant.
   - Pay particular attention to data quality issues that could impact on the end-use of the data.

4. **Data integration or record linkage** (8 marks):

   - Describe the data integration tasks that you undertook to combine / merge your source data sets, and how you dealt with any problems that occurred. Make sure you describe in detail how and why you joined or linked the source data sets in the way you did it.
   - Describe what you did with any records that couldn't be integrated, and how you dealt with inconsistencies.

5. **Data preparation and cleaning** (8 marks):

   - Describe any data cleaning and other data preparation you performed, such as data reduction, data transformation, and so on. Make sure you justify your actions with respect to the end-use of the data.

Marking Criteria

For questions 2 to 5 we will assess you on the following criteria:

- **Choice of measurements and techniques**.
  How appropriate are your chosen techniques and measurements for the data sets and problem? How well do you describe and justify these choices with respect to the end-use of the data, your overall strategy, and possible alternatives?
- **Demonstration of technical proficiency**.
  How well do you apply the chosen techniques and what level of technical proficiency does this demonstrate? Do you account for any potential difficulties with the techniques and how effectively do you deal with them if they occur?
- **Analysis and explanation of the outcome**.
  How well do you explain your results and outcomes and their significance to the end-use of the data? How complete is your analysis in terms of potential ramifications or issues that could impact the end-use? Could your specified end-use analysis be performed with confidence in the validity of any conclusions reached?
- **Communication and presentation of information**.
  How effective is your report at communicating what you did and why? Are the most important points given appropriate priority? How appropriate are your choice of visualisations (figures and tables) and how well do they communicate important information?