COMP3430/COMP8430 – Data Wrangling - 2018

Assignment 2 **Due 11:55pm Sunday 7 October 2018**

Worth 15% of the final grade for COMP3430/COMP8430

Draft – Last update August 16, 2018

Overview and Objectives

This assignment focuses on identifying possible data quality problems in data sets and taking necessary steps to correct them. We provide you with two data sets, similar as assignment 1, and for this second assignment you need to both identify data quality problems in these data sets and fix them. Fixing all the data quality issues with these data sets will likely take more time than you have. **This is intentional and we don't expect you to correct everything**. It also reflects the real-world where there is always more data cleaning that can be done. So **prioritise your effort** based on the time a task will take and the likely benefit to the end use of the data set, and make sure you justify these choices.

Important

- Submit **one zip archive file**, named **uNNNNNNN_*assignment_2.zip***, where uNNNNNNN is your ANU ID. For example if your ANU ID is u1234567 you should submit the file *u1234567_assignment_2.zip*. **Only use underscores** and not spaces, and **only lower-case letters** in your file name (as this will greatly help our marking efforts).
- Your zip file must contain,
  1. Your report, a **.pdf** document named **uNNNNNNN_*assignment_2_report.pdf***
  2. Your merged and cleaned data set, a **.csv** file named **uNNNNNNN_*assignment_2_ds.csv***
- 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 allowed total maximum length of your report is **three (3) A4 pages** and **1,200 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 the assignment before the submission deadline.

Submission

Submission will be done using Wattle. Click on the link Assignment 2 submission (to be made available) in week 9 to upload your report. You may submit as many draft versions of the assignment 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 assignment, therefore check carefully what you submit as the final version!**

Deadlines, Extensions and Late Submissions

The assignment is due 11:55pm, Sunday October 7, 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 assignment. We do encourage you to discuss your work, but we expect you to do the assignment work by yourself. 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 assignment.

Marking

This assignment will be marked out of 15. Note that not all assignment parts are equally difficult. **The assignment will count for 15% of your final course grade.** For this assignment there are no single right or wrong answers. Marks will be awarded based on your reasoning and the justification 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 assignment, we will re-mark the entire assignment and your mark may go up or down as a result**.

Assignment Questions

The purpose of this assignment is to fix data quality problems. Similar to assignment 1, we provide two data sets ***edu_emp.csv*** and ***medical.csv***, available for download from Wattle in week 6. By assuming **the final task to be conducted on these data sets is to examine links between an individual's education and employment history and their health you need to correct the data quality issues in these data sets and merge them into a single data set**.

We note that for these two data sets we have generated the records randomly, but used attribute values that come from real-world data sets (all of which are, or were, in the public domain). **Again, any similarity to real persons or places is entirely coincidental**.

We do not require the use of any specific tool, software package or programming language and you are free to choose whichever you feel most comfortable with. However, please note that due to the size of the data sets, manual inspection and correction of individual records will be very time consuming. **So make your decisions on data cleaning accordingly**.

For each of the three parts below you must do two things:

- Clean the actual data sets provided, either by writing code yourself or using a tool or software package.
- Describe in your report what you did and why. It is particularly important to justify your choices with respect to the end use of the data set – **examining links between an individual's education, employment and health**.

In your ZIP file you must include a single .csv file that contains your merged and cleaned data set. As part of marking your submission we will compare your submitted data set against a cleaned data set.

- **Part 1 - Merging the data sets** (7 marks): You must merge the two individual data sets into a single data set. Please describe and justify in the report how you did this and why you chose that approach. **Your new data set must include a header line, must use the original attribute names (except for any new attributes you have generated) and must include the Social Security Number (SSN) attribute**.

  Some things you may wish to consider in the merging process (and explain in your report) include:

  - How did you find corresponding records in the two different data sets?
  - If there were records that only occurred in a single data set, what did you do with them?
  - If there were duplicate records in a single data set, what did you do with them?
  - If there were any inconsistencies between the two data sets, how did you resolve them?

- **Part 2 - Missing and incorrect values** (4 marks):

  - If there were any missing values in the data sets, please describe in your report how you dealt with them and why you chose this approach.
  - If you detected any values that were not missing but were (likely) incorrect or impossible, how did you deal with them and why? If it was different to your approach to dealing with missing values, please justify this distinction.

- **Part 3 - Other data cleaning** (4 marks): Perform other data cleaning tasks that you think are important, keeping in mind the final use of the cleaned data set. Things you may wish to consider include:

  - The data quality dimensions from lecture 5 that are not covered by Part 1 and Part 2.
  - Any data reduction you think is warranted. Give particular consideration to the likely impact on the end use of the cleaned data set (**examining links between an individual's education, employment and health**).
  - Any attribute values that are in the wrong attribute.

Marking: You will receive up to 7 marks for describing and justifying a good approach to merging the two data sets and dealing with any issues that arise. You will receive up to 4 marks for your treatment of missing or otherwise problematic values in the merged data set, along with a justification of your decisions. You will receive up to 4 marks for any further cleaning you undertake, along with a justification of your approach. You also need to describe and justify why you chose to correct the things you did, particularly if there were other data quality issues you were unable to fix. We will mark you on both the actual data cleaning you have done (as evidenced by comparing your submitted data set to the clean data set) and the justification of the choices you made, as evidenced by your report.

Keep your report short and to the point, and focus your efforts on the actual data cleaning. **You do need to describe what you did, and why**. However, please do not include definitions, descriptions of how techniques work (unless you modified them somehow), a theoretical overview of data cleaning, etc.