# COMP3430 / COMP8430 – Data Wrangling – 2018

## Assignment 1 **Due 11:55pm on Sunday 2 September 2018**

## Worth 10% of the final grade for COMP3430 / COMP8430

Last update August 3, 2018

---

### Overview and Objectives

This assignment covers the topics of data quality and data exploration and profiling as presented in the first few weeks of the course. It also includes questions about what *data wrangling* is, why it is important, and how it fits into the broader field of *data analytics*. One question refers to the required readings from week 1 of the course while others ask you about practical aspects of data exploration and visualisation.

### Important

- Submit **one PDF file only**, named **uNNNNNNN_*assignment_1.pdf***, where uNNNNNNN is your ANU ID. For example if your ANU ID is u1234567 you should submit the file *u1234567_assignment_1.pdf*. **Only use underscores** and not spaces, and **only lower-case letters** in your file name (as this will greatly help our marking efforts).
- 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** 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 1 submission (to be made available) in week 6 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 version of your assignment so check carefully what you submit as the final version!**

### Deadlines, Extensions and Late Submissions

**The assignment is due 11:55 pm on Sunday 2 September 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.

### Assignment Structure

The assignment consists of five (5) questions which can be worth different numbers of marks. Make sure you answer all aspects of each question.

If you have any questions on the assignment 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 assignment questions.**

### 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 10**. Note that not all questions and parts are equally difficult. **The assignment will count for 10% of your final course grade.** For some of the questions there is no single right or wrong answer. 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

- **Question 1 (2 Marks)**: According to the paper (from week 1) by Rahm and Do (*Data Cleaning: Problems and Current Approaches*), data cleaning generally deals with detecting errors and inconsistencies from data to improve the quality of data. However, as mentioned in this paper, there are many issues and problems related to data cleaning. (1) Do you think the problems and issues raised in this paper (in 2000) are still relevant today? Why or why not? (2) Can you think of new data quality problems that have arisen in the current data cleaning context other than those from this paper? **Answer these two questions by writing around 1/2 page in total.**

- **Question 2 (1 Mark)**: Following is a list $L$ of data collected for an attribute that contains the age values (in years) of a group of people: $\mathbf{L} = \{\mathbf{5}, \mathbf{20}, \mathbf{32}, \mathbf{15}, \mathbf{42}, \mathbf{55}, \mathbf{92}, \mathbf{84}, \mathbf{56}, \mathbf{42}, \mathbf{13}, \mathbf{74}\}$.

  First split your ANU ID (excluding the first character 'u') into four number segments (three pairs and a single number) and then add these four number segments to $L$. For example if your ANU ID is u1234567 then split it into: 12, 34, 56, 7 and add these numbers to $L$, so the final list $L$ becomes: $L = \{5, 20, \ldots, 13, 74, \mathbf{12}, \mathbf{34}, \mathbf{56}, \mathbf{7}\}$.

  Now calculate and **include in your report both your workings** (equations used and partially calculated results) as well as **the final results** for (1) the **median of** $L$, (2) the **mode of** $L$, and (3) the **five number summary** of $L$, as discussed in the lectures. **If you do not include your working you will not receive any marks.**

- **Question 3 (2 Mark)**: Apply binning as covered in the lectures to the list of numbers in $L$ as generated in the previous question (i.e. $L$ with the numbers based on your ANU ID appended).

  Write into your report the results when binning $L$ using (1) **equal width** with three bins and **smoothing with bin medians**, (2) **equal depth** with two bins and **smoothing by bin boundaries**, (3) **equal width** with four bins and **smoothing by bin boundaries**, and (4) **equal depth** with four bins and **smoothing by bin means**. No workings need to be included but **the generated bins need to be shown in a clear way**.

For the last two questions of this assignment we provide you with two data sets and ask you to explore them and assess their data quality. We have generated the data sets ***edu_emp.csv*** and ***medical.csv*** by sampling from real-world data sets that are in the public domain. The data sets are available for download from Wattle in week 3.

We have intentionally tried to include the types of relationships and features that you might find in real data sets like this. We note that for these two data sets we have generated records randomly, but used attribute values that come from real-world data sets. **Any similarity to real persons or places is entirely coincidental**.

- **Question 4 (3 Marks)**: **Using Rattle**, provide different types of data exploration (distribution and/or correlation) using **suitable types of plots** for any of the data sets provided. Describe in one to two paragraphs each (1) why you selected a certain type of plot, and (2) what can be learned about the data from these plots. You may use a single attribute or multiple attributes to generate these plots.

  For this question **COMP3430 students** must provide **two (2)** plots and corresponding descriptions, while **COMP8430 students** must provide **three (3)** plots and corresponding descriptions.

- **Question 5 (2 Marks)**: Consider the data quality dimensions **Completeness**, **Consistency**, and **Accuracy** as described in Lecture 5 in week 2. For this question **COMP3430 students** must select any **three (3)** attributes or attribute pairs, while **COMP8430 students** must select any **six (6)** attributes or attribute pairs from any of the two data sets. Think about what attributes or attribute pairs you can select for which quality measure.

  Calculate the percentages of completeness, consistency, and accuracy of these selected attributes / attribute pairs, and briefly describe your answers in one to two sentences. Show your answers in a table structured as shown below. You must show your workings and calculations to get full marks.

| | Attribute 1 | Attribute 2 | Attribute ... |
|---|---|---|---|
| Completeness | | | |
| Consistency | | | |
| Accuracy | | | |