COMP3430 / COMP8430 – Data Wrangling – 2024

Assignment 1 **Due 11:55 pm on Friday 30 August 2024**

Worth 10% of the final grade for COMP3430 / COMP8430

Last update July 29, 2024

Overview and Objectives

This assignment covers the topics of data quality, data exploration, and data 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 task refers to the required readings from week 1 of the course while others ask you about practical aspects of data exploration.

Important

- The answers to this assignment have to be submitted online in Wattle, see the link **Assignment 1 Submission** under the Assignments tab.
- Follow instructions given for maximum text length in free format answers. If your answer is too long it will attract a penalty (for details see the individual questions below and the corresponding answer submission forms in Wattle).
- You can edit your answers many times and they will be saved by Wattle.
- Make sure you submit the **final version** of your assignment answers **before the submission deadline**.
- Note that **Wattle does not allow us to access any earlier edited versions of your answers, so check very carefully what you submit as the final version!**

IMPORTANT: You can only submit your assignment once! Make sure you do not forget to submit your assignment!

Penalties

Each textual question has a maximum word limit. If you write more than these provided limits we will have to apply an over-word-limit penalty. For details of limits see the individual questions below and the corresponding pages in the assignment submission in Wattle.

Deadlines, Extensions and Late Submissions

The assignment is due 11:55 pm on Friday 30 August 2024.

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 CECC 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 four (4) tasks as described below which can be worth different numbers of marks. Make sure you answer all aspects of each task.

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

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 carefully read the Types of Academic Misconduct** (https://www.anu.edu.au/students/academic-skills/academic-integrity/best-practice-principles/types-of-academic-misconduct) and the **Academic Integrity Rule 2021** (https://services.anu.edu.au/learning-teaching/academic-integrity/academic-integrity-rule-2021-for-learners).

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 submitted assignment answers. 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, and it will contribute to 10% of your final course mark.

Note that not all tasks and questions are equally difficult. For some of the tasks there is no single right or wrong answer. Marks will be awarded based on your reasoning and the justification of your decisions and explanations, as well as clarity and correctness of writing.

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**.

Data Set Generation for this Assignment and for Assignment 2

For this assignment and the upcoming Assignment 2 each of you will work on an individual data set that will be based on a **master data set** we will provide, and a **data generation program** we will also provide.

Note that we have generated the master data set based on real data (such as lookup tables of names, addresses, etc.), and we have then corrupted and modified certain aspects of that data set. We have intentionally tried to include the types of relationships, features, errors, and other data quality issues that you might find in real data sets. **Any similarity to real persons or places is entirely coincidental**.

Download the master data set named **dw_assignment_master.csv.gz**, and the data generation program named **generate-student-dataset.py** (to be made available under the Assignment tab on Wattle). Copy both these files into one folder / directory, and run the code using Python 3 in the following way:

```
python3 generate-student-dataset.py your_ANU_ID dw_assignment_master.csv.gz
```

The program will generate an output data set named **data_wrangling_medical_2024_your_ANU_ID.csv**, and print some output which contains the following important lines (for the example ANU ID u1234567):

```
$ python3 generate-student-dataset.py u1234567 dw_assignment_master.csv.gz

Your student data set for the data wrangling 2024 assignments has been generated and written into file:

    data_wrangling_medical_2024_u1234567.csv

  Your ANU ID check code is:            d76225bc
  Your student data set check code is:  a2523232f8e2

  *** Check this pair of numbers is in the list provided on Wattle, if not contact the course convenor.
```

Important

- **Write down your two check codes because you must provide them with the assignment submission.** This will allow us to validate that you have generated and used the correct data set.
- **Check that the pair of check codes you get** (like in the example above d76225bc and a2523232f8e2) **is in the list of check codes we will provide on Wattle** (under the Assignments tab Assignment 1 section). This will allow you to check that you have generated the correct data set.
- **You must use your individual generated data set for task 4 of this assignment (and the tasks on data cleaning in the upcoming Assignment 2)**.

Assignment Tasks

- **Task 1 (2 marks)**:

  (1) The Rahm and Do paper *Data Cleaning: Problems and Current Approaches*; IEEE Data Engineering Bulletin, 2000, was published over twenty years ago. This makes it relatively old by the standards of academic research in computer science.

(a) Can you think of two (2) new data quality problems that have arisen since then? (maximum 150 words)

(b) Are the issues and problems raised in this paper still relevant today? Justify why you think so or why not. (maximum 200 words)

(2) The Australian census conducted every year facilitates important decision and policy making by the government. Assume that you are employed by the Australian Bureau of Statistics as a data wrangler to clean the latest census data set and integrate it with previous years' census data to support decision making, for example, about Aboriginal and Torres strait islanders' access to health care. Identify and describe three (3) data wrangling aspects you will have to consider when dealing with such data sets. (maximum 250 words)

- **Task 2 (1 mark)**: Following is a list **L** of age values (in years) of a group of people:

$$\mathbf{L} = [\mathbf{25}, \mathbf{11}, \mathbf{40}, \mathbf{17}, \mathbf{17}, \mathbf{41}, \mathbf{21}, \mathbf{31}, \mathbf{46}, \mathbf{26}, \mathbf{86}, \mathbf{74}, \mathbf{100}, \mathbf{28}, \mathbf{15}, \mathbf{97}]$$

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 u1204067 then split it into: 12, 04, 06, 7 and add these numbers to **L**, so the final list becomes: $\mathbf{L} = [25, 11, 40, 17, 17, 41, 21, 31, 46, 26, 86, 74, 100, 28, 15, 97, \mathbf{12}, \mathbf{4}, \mathbf{6}, \mathbf{7}]$.

Now calculate and enter into the corresponding answer fields on Wattle (Round your answers to two decimal places, eg: 27.36):

1. the **mean and standard deviation of the final list L**,
2. the **median and median absolute deviation of the final list L**, and
3. the **mode of the final list L**.
4. Based on the mode, median, and mean values, what can you say about this distribution (negative/ positive skewed, normal distribution, etc.)? Explain in one or two sentences

- **Task 3 (2 marks)**: Apply binning as covered in the lectures to the numbers in the list **L** as generated in the previous task (i.e. **L** including the number segments based on your ANU ID appended).

Calculate and enter into the corresponding answer fields on Wattle the results when binning **L** using (Round your answers to two decimal places, eg: 27.36):

1. **equal depth** with two bins and **smoothed by bin median**,
2. **equal width** with three bins and **smoothed by bin mean**,
3. **equal width** with four bins and **smoothed by bin boundaries**, and
4. **equal depth** with four bins and **smoothed by bin boundaries**.

Clearly show the bins you generated when you enter your answers into Wattle answer fields by showing one bin per line, for example (assume we have binned [1,2,3,4,5,6,7,8,9] into three bins with smoothing by bin medians):

```
Bin 1:  [2, 2, 2]
Bin 2:  [5, 5, 5]
Bin 3:  [8, 8, 8]
```

- **Task 4 (5 marks)**:

For the last task of this assignment **you must use the data set you generated as per instructions above**. We ask you to explore this data set using tools of your choice (Rattle, R, Python, Pandas, etc.) and answer the specific questions about this data set given below.

Make sure to follow the instructions on the individual Wattle answer fields with regard to rounding, the number of digits to provide after the decimal point, etc.

1. Provide the missingness patterns of values (as we discussed in the labs) for the three attributes: **postcode**, **phone**, and **email**. You should provide the 0-1 missing value pattern table we discussed in the labs for the above three attributes.
2. Calculate the correlation between the attributes (a) **BMI** and **age_at_consultation**, and (b) **state** and valid **marital_status**. In your answers you need to provide the numerical correlation value (Round your answers to two decimal places, eg: 27.36), the name of the correlation method you used, and a brief (one sentence) explanation why you used that specific correlation function for each pair of attributes. (Note: the correlation statistic is not the same as p-value)
3. For the following attributes, calculate numerical values for the following data quality dimensions:
   (a) Completeness for **middle_name** and **email** (consider these attributes individually).
   (b) Validity for **weight** and **email** (with a valid email containing the @ symbol. Only consider non empty email values for the calculation).
   (c) Uniqueness for **first_name**.
   (d) Consistency between **age_at_consultation** and **birth_date** (for valid age values).
   Clearly describe how you calculated each of your results (Round your answers to two decimal places, eg: 27.36%).

4. Calculate the distributions of the first digits (Benford's law) for the attributes (a) **cholesterol_level**, (b) **blood_pressure** and (c) **medicare_number** (Round your answers to one decimal place, eg: 40.2%). Describe for each in one or two sentences if it does follow Benford's law or not, and why you think it does or does not follow this law.
5. Assume you constructed a data cube representing certain clinical information contained in your generated data set, where the three dimensions indicate the locations (states in Australia), disease type (infectious diseases, deficiency diseases, hereditary diseases, and physiological diseases), and the consultation time (year).

   Briefly describe two data warehousing operations you can apply on this data cube, clearly specifying on which dimension the operation is applied, and an example for the result you may obtain. (maximum 250 words)

You will receive up-to one mark for correctly answering each of these questions, where both correct numerical values as well as correct and clearly written justifications of your answers will be considered.

Other Aspects

For all textual answers in this assignment, English writing mistakes and typographical errors will attract small penalties. **Do not upload any code into the answer boxes in Wattle.**