COMP3430 / COMP8430 – Data Wrangling – 2024

Assignment 2 **Due 11:55 pm on Friday 27 September 2024**

Worth 15% of the final grade for COMP3430 / COMP8430

Last update September 19, 2024

Overview and Objectives

This assignment covers the topics of record linkage and data cleaning, with a focus on identifying possible data quality problems in data sets and taking necessary steps to correct them. Similar to assignment 1, we ask you to generate a second data set, and to then both identify data quality problems in these data sets and to also 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 almost always more data cleaning that can be done. So **prioritise your effort** based on the tasks we ask you to do, and the likely benefits to the end use of the data set. Make sure you justify and describe these choices.

Important

- The answers to this assignment have to be submitted online in Wattle, see the link **Assignment 2 Submission** under the Assignment 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

Textual questions have maximum line and maximum word limits. 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 27 September 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 15, and it will contribute 15% 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.

IMPORTANT: We do not accept any type of code, screenshots, or external links as answers. Please do not waste the space given to you to provide answers by writing external links or code in that space. We will not mark such answers and you will lose marks if the correct answer is not inside the text fields.

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

As with Assignment 1, for this assignment each of you will again work on individual data sets that will be based on a second **master data set** we will provide, and a new **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 second master data set from Wattle (under the Assignment tab Assignment 2 section) named **dw_assignment_master2.csv.gz**, and the new data generation program named **generate-student-dataset2.py**. Copy both these files into one folder / directory, and run the code using Python 3 in the following way:

```
python3 generate-student-dataset2.py your_ANU_ID dw_assignment_master2.csv.gz
```

The program will generate an output data set named **data_wrangling_education_2024_your_ANU_ID.csv**, and print some output which contains the following important lines (for the example ANU ID u1234567):

```
$ python3 generate-student-dataset2.py u1234567 dw_assignment_master2.csv.gz

Your student data set for the data wrangling 2024 assignment 2 has been generated and written into file:

    data_wrangling_education_2024_u1234567.csv

  Your ANU ID check code is:            d76225bc
  Your student data set check code is:  3dd9e57a0f08

  *** 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 3dd9e57a0f08) **is in the list of check codes we will provide on Wattle** (under the Assignment tab Assignment 2 section). This will allow you to check that you have generated the correct data set.
- Note that the check codes are different for the data set you generated for assignment 1 and this new data set.
- **You must use your individual generated data set, data_wrangling_education_2024_your_ANU_ID.csv (generated based on your ANU ID), for tasks 2 to 4 of this assignment, together with the data set you generated for Assignment 1, data_wrangling_medical_2024_your_ANU_ID.csv**.

Assignment Tasks

- **Task 1 (3 marks)**: Generate two strings of numbers based on your ANU ID (excluding the first character 'u'):
  - s_1 is the string '123' concatenated with the **first four digits of your ANU ID**.
  - s_2 is the string '123' concatenated with the **last four digits of your ANU ID. Include any leading zeros**.

  For example, if your ANU ID is u9800765 then s_1 = '123<u>9800</u>' and s_2 = '123<u>0765</u>'

  Now **manually** calculate the following similarities between these two strings and **include in your assignment both your workings** (equations or edit matrices) as well as **the final results** for:

  1. The Dice coefficient similarity based on unigrams ($q = 1$).
  2. The Jaccard similarity based on bigrams ($q = 2$).
  3. The bag distance similarity
  4. The Levenshtein edit distance between the two strings, assuming a cost of 2 for substitutions, cost of 1 for inserts, and cost of 1 for deletes.
  5. In a couple of sentences, explain the relationship between the bag distance and the edit distance.

  Round the final numerical results to two decimal places (eg: 0.01 or 42.42). If you do not include your workings you will not receive any marks. Python or R code are not acceptable workings. For sub-task 4 you must show the full edit matrix as your workings (as discussed in the lectures).

  Your will receive 0.5 marks for correct calculations and results for sub-tasks 1 to 3, 1 mark for the correct calculation and result for sub-task 4, and 0.5 marks for subtask 5.

For the following tasks, you need to use your individual education data set generated as per the instructions above together with your individual medical data set generated for Assignment 1.

By assuming **the final task to be conducted (by a data analyst after data wrangling has been completed) on these data sets is to examine links between an individual's education, their employment history, and their health**, **you need to merge the two data sets into one single data set and correct data quality issues in this merged data set. Note that we do not expect you to do the final analysis, but only the data wrangling tasks which will prepare the merged data set for the final analysis**.

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**.

Once you cleaned and merged the two data sets, you need to upload a single **.csv** file that contains your final merged and cleaned data set, named **data_wrangling_merged_2024_your_ANU_ID.csv** in Wattle. For example, for an ANU ID u1234567 the final data set should be named as **data_wrangling_merged_2024_u1234567.csv**.

As part of marking your submission we will compare your submitted data set against a cleaned data set that we have generated.

- **Task 2 – Merging the data sets (4 marks)**: You must merge your two individual data sets into a single data set using the Social Security Number (SSN) attribute that is in common to both data sets. **Your new data set must include a header line with all the original attribute names and must include the SSN attribute – named 'ssn' as in the two data sets to be merged. In addition, you may include new attributes you have generated, if any**.

  You will need to address the following four aspects of this merging step in the assignment answer field in Wattle:

  1. How many unique SSNs occurred in common in both data sets? How many occurred only in the medical data set, and how many occurred only in the education data set?
  2. If there were records that only occurred in a single data set, describe what you did with them, and explain / justify why.
  3. If there were duplicate records in an input data set (with the same SSN), describe what you did with them, and explain / justify why.
     Note: Please only explore and obtain a count of the duplicates in each data set in this step, and describe how you plan to handle them in the very end. Do not perform the deduplication until you have answered all other questions. Apply deduplication at the very end, before you save the final merged dataset.
  4. If there were any inconsistencies between records in the two data sets with the same SSN, what attributes had inconsistencies, and how many from each attribute? How did you deal with these inconsistencies, by either resolving them or processing them otherwise? Describe and justify the approaches you took.
     Note: When counting inconsistencies, where one SSN from one dataset has more than one match in the other, please count inconsistencies only once, and if at least one of the matching record pairs is consistent for a given SSN for an attribute, please consider that the attribute value for that SSN is consistent.

  Write a maximum of 400 words for the above task.

- **Task 3 – Missing and incorrect values (4 marks)**:

  1. Following the missing patterns table we discussed in the labs and used in Assignment 1, find the combination of three attributes with the highest number of missing values (i.e. the three-attribute combinations with the largest numbers of records with missing values) in (a) education data set, and (b) your merged data set. Provide the attribute names and the corresponding number of records with missing values in each of these data sets.

     Note: Please consider the original education data set when calculating missing values in question (a). In question (b), please first merge records with common SSN across the two data sets, and ignore the records where the SSNs appear in only one of the two data sets. Furthermore, for common attributes, if the attribute value is missing in only one data set for records merged across the two data sets, please replace the missing value with the non-missing value, before doing missing value calculations. However, as highlighted in Task 2 (3) above, please do not do deduplication before conducting missing value calculation. See the example below.
     DB 1 – [[SSN: 123, first_name: Tom], [SSN: 123, first_name: Tommy], [SSN: 456, first_name:'']]
     DB 2 – [[SSN: 123, first_name: ''], [SSN: 456, first_name:''], [SSN: 789, first_name: Mary]]
     Merged data set to consider for counting missing values – [[SSN: 123, first_name: Tom], [SSN: 123, first_name: Tommy], [SSN: 456, first_name:'']]
     Therefore, the merged dataset has only one missing value in first name (note that we have not applied deduplication yet). Also notice that we have dropped the record with SSN 789 since it does not appear in DB 1.
  2. For the two attributes with the highest number of missing values (individually) in your merged data set, either:
     – consider if you can impute these missing values. If so describe the approach you have taken to impute missing values, and justify why you have taken this approach; or
     – if you decided you cannot impute missing values in an attribute then describe and justify why you have not done any imputation.
  3. Describe what incorrect or impossible values you found in attributes in your merged data set, and provide how many such incorrect or impossible values are there for each attribute. Also describe why you believe these values are impossible or incorrect.
  4. Describe how you dealt with the incorrect or impossible values identified in your merged data set (for example correcting them in some way or another).

  Write a maximum of 450 words for the above task.

- **Task 4 – 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, to **examine links between an individual's education, employment, and their health**.

  Things you may wish to consider include (but are not limited to):

  1. Do any of the attributes in your merged data set have data quality issues with regard to the data quality dimensions we discussed in Lecture 5 that are not covered by Tasks 2 and 3.
  2. Are any data reduction or transformation tasks required?
  3. Are there values that seem to be in the wrong attribute(s)? If so how can you correct them?
  4. Any other problems you detected that you think are worth correcting.

  Write a maximum of 400 words for the above task.

  You should consider minimum of four extra data exploration and data cleaning tasks, and for each describe what you have done, justify why you have done it (within the context of the final use of the cleaned data set described above), and any numerical results relevant to the task that you think are important.

Marking: For each question in Tasks 2 to 4 you will receive up to one mark for appropriately describing and justifying your approach.

For task 2, you will receive up to 4 marks for describing and justifying your approach to merging the two data sets and dealing with any issues that arise.

For task 3, you will receive up to 4 marks for your treatment of missing or otherwise problematic values in the merged data set, along with appropriate justifications of your decisions.

For task 4, you will receive up to 4 marks for any further data exploration and data cleaning you undertake, along with appropriate justifications of your approaches.

We will mark you on both the actual data cleaning you have done (as evidenced by comparing your submitted merged data set to our clean data set) and the justification of the choices you made, as evidenced by your answers provided in Wattle.

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.**