

Individual Project: Extra Credit Assignment

Instructions My Assignment

Submitted on November 30, 2018

10/10 points (100%)

QUESTION 1

10 POINTS

This extra credit exercise is an opportunity to explore a data set, assess the data representation, and apply data curation concepts from the course.

As extra credit, it is not a required assignment and doing it (or not) cannot have a negative impact on your grade. Doing this exercise can raise your grade, boosting it by one letter-grade level at most. In short, it can be a way to overcome lower than expected scores on assignments. If you have a 90 or above, it might not be helpful.

This exercise will be graded using the 10 point rubric below. We encourage you to review the rubric and ask the instruction team questions.

We will add the extra credit points manually after Coursera has calculated your course grade. You will be notified about how many extra credit points you received.

Due Date: December 2

You are being given a data file, EC_A.xls.

Give a short (~400 to 700 words) natural language account of the data file. You

can supplement your account with an illustrative diagram using Chen's notation or UML notation, but note that the main purpose of this assignment is to get you to engage with the data in the file in a thoughtful manner.

When writing your account, be sure to address each of the following questions.

1. Speculate as to the content of the data set. What is it describing?
2. Say at least two things about how the data is organized. (E.g., Is it in first normal form? Is there anything significant regarding how it is visually organized? Is it significant that the first two (or three) rows appear different than the subsequent rows? Is the text formatting significant? Etc.)
3. How many different entities and/or attributes are apparent in the data? (You can illustrate this with a diagram if you wish.)
4. Describe the dataset's shortcomings with respect to data curation objectives.
5. Given that one of the sources of the data is the following webpage <http://www.geocities.jp/kawabemasatake/index.html>, name at least one assumption that was made when the data file EC_A was created. (There are at least three assumptions that are relatively easy to pick out.) Why are the assumptions significant? What do they say about the quality of the data?

EC_A.xlsx

Submission: Please submit a pdf file of your narrative account.

Grading Rubric:

Prompt	Points
Evidence of understanding data representation, schemas and independence (Item 1 & 2)	2
Evidence of in-depth examination of data set (Item 2 & 3)	2
Discussion of curation activities, needs, and decisions (Item 4 & 5)	4
Overall quality analysis and completeness (Item 1, 2, 3, 4 & 5)	2

Analysis of dataset

🔗 Analysis of dataset

In depth analysis of dataset

RUBRIC

Evidence of understanding data representation, schemas and independence (Item 1 & 2)

Evidence of in-depth examination of data set (Item 2 & 3)

Discussion of curation activities, needs, and decisions (Item 4 & 5)

Overall quality analysis and completeness (Item 1, 2, 3, 4 & 5)

