IS 733 Data Mining (Fall 2021)
Homework 1
Data Wrangling with EdNet Data
Due before class on October 4th, 2021

Warnings: This homework takes time and effort, please plan accordingly

You are just hired as a data analyst (*Note1)* in the newly formed analytics department of RiiiD, which is a leading AI startup company specializing in providing learning resources and adaptive practices to English learners in South Korea. On the first day of your job, you are invited to attend a meeting with the business operation team in which you are briefed on the company's platform and then you are handed over a dataset ('EdNet') that was collected from this platform over the last two years from over 700K users . This dataset logged detailed user activities while they were interacting with systems. Intrigued by the sheer amount of data collected, your manager is interested in how the analytics department can help to support the company's missions to reimagine the learner's experience using AI/ML/data analytics techniques. You are asked to spend some time to look into the data and prepare a brief to your manager. Specifically, your manager is looking for answers to the following questions.

1. *(25 points) Who are the users?* To answer this question, you will need to compile a user profile table (Table 1) with information about users including
    a. Overall practice volume and performance (e.g. # of questions answered, % of questions answered correctly)
    b. Learning activity (e.g. # lectures watched, # explanation read)
    c. Add three additional metrics you would like to compute to describe users

    Create a few plots to illustrate the information in the tables. Feel free to choose the type of plots you think is appropriate.

2. *(25 points) What are the questions/items?* To answer this question, you will need to compile a question profile table (Table 2) with information including
    a. Question ID
    b. Question Type
    c. Number of times being practiced
    d. Number of times answered correctly

    Create a few plots to illustrate the information in the tables. Feel free to choose the type of plots you think is appropriate.

3. *(10 points)* Design a modified metric of "accuracy" to fairly describe users' ability by taking into account the difficulty level as derived from Table 2. Describe the procedure to compute the metrics. Be sure to be specific so that interns can use your pseudo code to implement the metrics without much trouble.

BONUS (10 points), implement the proposed metrics and plot a histogram of the metrics across all users (or subset of users of your choices).

4. (30 points) Pick a user with a reasonable amount of activity (you will define the "reasonableness" and specify the selection criteria) and create a dashboard that consists of a series of plots to tell a story of this user's activity patterns. For inspiration, you may look at the user dashboard for fitness tracker such as Fitbit.

5. (10 points) Propose two tasks for your interns to work on. The first task is of unsupervised/descriptive type and a second one is of supervised/predictive task. Please provide clear specification of the tasks so that your interns can start to work right away. You should try to propose tasks that are not attempted in the existing work with this dataset.
   BONUS (10 points): propose a reinforcement learning task

Deliverable:

- A google slide deck summarizing the above findings in the format of plots or tables (those are not table 1 or table 2, but small tables you decide to use to present information) or other contents as requested by your manager (such as those pertaining to question 3,4 or 5). Please label clearly on the slide which question you are answering. There are no lower/upper limits of the number of slides. Always keep the message concise and effective and keep your audience in mind. In this case, it is your manager. Please change the slides permission to editable, we will make comments on your slides. Two summary tables in csv format, named, table1.csv and table2.csv for user table (question 1) and item table (question2) respectively
- Please deposit the above items (a slide deck and two csv tables) into your own google folder you are asked to create (where you put your strategy plan), under this folder, create a folder named homework1. Please deposit the files there. *Please don't deposit your finalized product until after the submission deadline.*
- Please create a github account and upload codes and *optionally other files necessary to create your portfolio. You may create a github page to present your project, but this is optional for this homework assignment.*
- For submission on blackboard
    - Please upload a pdf version of your slide as your evidence of submission.
    - Please submit a link to your github account

Dataset and Background Readings:

Please download the dataset from the following link. For this homework, you may only use KT4 (uncompressed size 6.4GB). But you may need to download other small lookup tables(e.g. those in the contents folder) for the purpose of this assignment. ***Note: it may take a few hours to download/unzip the data files, please make sure you plan ahead***.
https://github.com/riiid/ednet

Please refer to this paper for details of dataset (mainly Section 1 and 2, up to page 7)

https://arxiv.org/abs/1912.03072

Software and Tools
You are free to use any software tools you feel comfortable with, which include but are not limited to Python, R, WEKA or Tableau.

Support Available
You will have the opportunity to ask specific questions to “Operation Manager” Ms. Anna who worked on this dataset during the summer. She will come to class on Sept 20 and 27 to answer your questions. You may also ask her questions via Piazza forums under HW1 tab. This forum will also be monitored by myself and TA.

Note 1:
There is an real active job ads for data analyst with RiiiD lab
https://jobs.lever.co/riiidlabs/3e447480-daf6-4301-8c20-c7da61d964cd