Data Management I (Data Profiling and Quality Dimensions)
Exam Guidelines and Requirements

Prof. Dr. Ajinkya Prabhune
Academic Researcher: Ashish Chouhan

- The submission for this module has two parts; details are explained as **Goal-A** and **Goal-B** below.
- **Goal-A** -> Each student can select an unclean dataset of their choice, and it is expected that you profile and clean the dataset considering the different quality dimensions.
- The student **MUST** consider the following criteria while selecting the dataset:
    - Minimum 15 columns with at least 1 column of each datatype.
    - Minimum 20,000 records in the dataset.
- The exam format for **Goal-A** is a 20min presentation, including the question and answer session.
    - the presentation should contain the systematic steps taken by the student in profiling the data and cleaning the data considering the different data quality dimension.
    - exceeding over 20min is subject to a reduction in the grades.

  o Each student has to submit his/her presentation + code/script of the demo until **21st March 2021 23:55** in Moodle.
- **Goal-B** -> Each student has also to prepare an unclean dataset**.**
- The exam submission for **Goal-B** will be a short document, recipe/source-code if any and the unclean dataset in either CSV, XL, or a MySQL database dump.

Tips for Goal-A

a) Data Profiling

- Using Talend Data Quality Management tool, you have to explain the data profiling steps performed on the dataset.

b) Data Cleaning

- Mention all the relevant data quality dimension necessary for cleaning the dataset.
- Present in detail the steps (recipe) performed for cleaning the dataset by applying the data quality dimensions. A systematic walkthrough of the recipe is expected.
- OpenRefine, Talend Data Preparation, Tableau Data Prep, or Trifacta can be used for performing the data cleaning process.
- The datasets can be in any format, exporting them in a DB system for performing specific/advance queries is up to each individual. However, this information should be seamlessly integrated into the presentation and highlighted where necessary.
- Following are the mandatory questions that have to be answered for this goal:
    - What insights did you gain from this dataset profiling and cleaning?
    - Which all data quality dimension does your analysis fits into?
        - Explain the dimension with a user-story, the user-story can be hypothetical one created by you.
        - For example, in the classroom tutorial on the Patent dataset, the user-story was finding the different countries filing patents;
            - To answer this user-story, the dimensions conformity + consistency + accuracy was considered for the column “PublicationNumber”.

- Why did you choose a given technique and the associated steps for a task?
  - Is there an alternative technique that could have been

Goal-B

- Creating a Raw dataset refereeing to any online source like kaggle.com.
- Conditions on the Raw dataset
  - Minimum dimensions to be considered while making the dataset Raw are
    - Accuracy
    - Completeness
    - Conformity
    - Validity
    - Consistency
    - Uniqueness
    - Currency
  - You are free to include any other dimension from the Literature.
- The dataset must be made at least **25% Raw/Unclean** based on the above-stated dimensions. Make sure to mention how the 25% has been calculated after the uncleaning is performed.
- Every record/set of records that you make unclean should be documented, highlighting the quality dimension that is considered.
- The recipe/source-code should also be submitted for verification
  - You can use any or a combination of Excel, OpenRefine, Data Preparation, Tableau Data Prep, or Trifacta.