As a reminder about your overall grade: The **final project** consists of two parts (Phase-1 and Phase-2) which are weighted ⅓ and ⅔, respectively. Both parts together count 25% of your overall grade (the best 5 out of 6 assignments count the other 75% of your grade).

Handling of Phase-1 and Phase-2 Reports:

The final project report ("Phase-2 report") **includes your Phase-1 report …**

- (A) **either** "as is" (i.e., a verbatim copy-pasted version of your Phase-1 report)
- (B) **or** it includes your **improved** Phase-1 report (based on feedback you received)

In other words, if you are happy with your Phase-1 report and feedback, then you can simply choose option (A) and say so at the beginning of your final report. On the other hand, if you want to improve your Phase-1 report (based on TA feedback or new insights/work for Phase-1), then you should include a brief "change report" to help regrade your improved Phase-1 work.

Therefore, please **include the following information at the top of your final report**:

1. **Points received for initial Phase-1 submission** : ____ (out of 100)
2. **Our team chooses the following Phase-1 option:**
   - [ ] **(A) No change** to Phase-1 report
   - [ ] **(B) Improved** (or extended) Phase-1 report

If you choose Option (A), we will **not** regrade your Phase-1 report and just take your original result. If you choose Option (B), we **will regrade** your Phase-1 report. Independent of whether you choose (A) or (B), please **include your Phase-1 answers** in the final report, to **make the final report (PDF-file)**.

If you have chosen (B), then please **highlight** the new/changed parts (e.g., using **boldface**) and provide a simple change report (e.g., after the title page or table-of-contents), e.g., like this:

Phase 1 Change Report

Phase 1 Remarks (TA feedback):	**Student Comments / Summary:**
Use cases: *[-2] Missing a reason why data cleaning is "never (good) enough" for U2.*	*This missing reason has been addressed in Section 1.3 (see page 5). We provided more reasoning for the use case U2.*
Data Quality: …	…

Final Project Report: Phase-1/Part-1 [100 points]

The final project report (Phase-1 and Phase-2) needs to be written in **narrative form** (i.e., as a *report*, not just as a list of bullet points!) Please refer to the project instructions for further details.

When grading your final report, we will look for the following elements:

1. [5 Pts] **IDENTIFY** the dataset *D* (or datasets) that you are working with. You have to make sure that *D* has no access/sharing limitations, i.e., please state the **origin** of your dataset (e.g., the website / URL where you got it from) and confirm (if you "bring your own" dataset") that you have permission to share it.
    - Refer to your dataset by a suitable **name** and provide a **reference** to it (similar to a bibliographic reference). Include a **URL** (or other means to get access if necessary).
2. [25 Pts] **DESCRIBE** the dataset, i.e., its **structure** (schema) and **content**. You can make use of the following formalisms to describe the **structure**:
    - **conceptual model** (e.g., an ER diagram or an ontology) or a **database schema** (e.g., CREATE TABLE statements in SQL)
    - Use a narrative to describe the **content**. In most cases, this would include a brief description of what each column/attribute represents (e.g., don't just say "column 5 is a date column" - instead elaborate what date is this: e.g., the date-of-birth of a person, the transaction date of a purchase, etc.)
        - If you're not sure what all the columns mean, state this explicitly in your report and justify why this isn't a problem for your use case.
        - If your dataset has a spatial and/or temporal extent (e.g., loan data from the US, time range: from 2000-2010), then include this metadata in your dataset description.
    - Note: describe the dataset in its **raw form**, i.e., before any conversions, transformations, etc. For example, if you load a dataset into Excel for an initial inspection, Excel might automatically guess a data type for certain columns and convert the column. This data conversion already forms part of your data cleaning workflow. In case you do that, make sure you document this in the parts of your report that describe the overall workflow.

3. [25 Pts] List all **DATA QUALITY PROBLEMS** that you found. You can focus on those data quality problems that are essential for your target use case (U1), but you can also list other data quality problems that you have encountered. **For each column/attribute that has a DQ problem** …
    - [15 Pts] Explain what the problems are (use narrative form)
    - [10 Pts] Include evidence of the problems (e.g., through copy-pasting examples or screenshots)

4. [35 Pts] Describe a **TARGET USE CASE**, and two **minor** use cases:
   a. [25 Pts] The **main/target use case U1** is the one that motivates and justifies your data cleaning project!
      i. [20 Pts] In narrative form explain your target application or question/query that you would like to support and for which you need to clean the data. For U1, data cleaning should be **necessary and sufficient** (see instructions and fireside chat recordings for details).
      ii. [5 Pts] List which of the columns (and DQ problems) from Part (3a) are needed to implement U1.
   b. [5 Pts] A **minor use case U0** should be described that could be realized with “zero data cleaning”, i.e., the dataset *D* is “good enough as it is”. This description can be simpler than the description of U1, i.e., a brief description of a use case where no data cleaning is needed (data cleaning is **not necessary**). Provide a brief explanation why for U0 no cleaning is needed.
   c. [5 Pts] A **minor use case U2** should be described briefly that cannot be realized with any amount of data cleaning, (i.e., data cleaning is **not sufficient**), even though the dataset might look useful for U2 at first sight. Again, provide a brief explanation why U2 is “hopeless”.

5. [10 Pts] Include a short description of your **initial plan** for cleaning the dataset. For example, you may follow the steps S1-S5 in the *Instruction* document as an initial plan of the entire project and adjust it to match your data and user case accordingly:
   ○ S1: Description and summary of dataset *D* and matching use case U1.
   ○ S2: Profiling of D to identify DQ problems that need to be addressed to support U1.
   ○ S3: Data cleaning *process* and *tools* (and reasons). Here you should describe which tools you are planning to use, e.g., OpenRefine; Python; etc.
   ○ S4: Check that the new dataset is improved and problems are fixed to support the use case; e.g.: How do we expect the columns to change? Which ICV checks can be used to detect problems (or their absence)?
   ○ S5: Document dataset changes (e.g., through a summary table)
   ○ Assignments of tasks [who does what] (if applicable)

Final Project Report: Phase-2/Part-2 [100 points]

Remember that the final report has to be written in **narrative** form. For Phase-2 of your final report, we are looking for the following information, now that you have actually performed the data cleaning:

1. [40 Pts] **Data cleaning performed (with OpenRefine, Python, and/or other tools)**
   a. [20 Pts] **Identify and describe** all data cleaning steps you have performed.
   b. [20 Pts] For each data cleaning step you have performed, **explain its rationale**. Was it required to support the use case U1? If not, explain why those steps were still useful.
2. [20 Pts] **Document data quality changes**
   a. [10 Pts] Quantify the results of your efforts (e.g., by providing a summary table of changes: Which columns changed? How many cells (per column) have changed? Etc.
   b. [10 Pts] Demonstrate that data quality has been improved, e.g., by devising ICV queries and showing the difference between "before" and "after" (cleaning).
3. [20 Pts] **Create a workflow model**
   a. [10 Pts] A visual representation of your **overall (or "outer") workflow W_o**, e.g., using a suitable tool such as YesWorkflow. At a minimum, you should identify key inputs, outputs, and steps of the workflow, along with dependencies between these. Key phases and steps of your data cleaning project may include, e.g., data profiling, data loading, data cleaning, IC violation checks, etc. Explain the design of W_o and why you've chosen the tools that you have in your overall workflow.
   b. [10 Pts] A detailed (possibly visual) representation of your "**inner**" **data cleaning workflow W_i** (e.g., if you've used OpenRefine, you can use the OR2YW tool).
4. [10 Pts] **Conclusions & Summary.** Please provide a concise summary and conclusions of your project, including lessons learned. If you haven't done so earlier in the report, you should also summarize the contributions of each team member here (for teams with >= 2 members).
5. [10 Pts] **Submitting the report and supplementary materials**
   Submit a **single ZIP file** with the **final project report** (including both Part-1 and Part-2) and the following **supplementary materials:**
   a. Name of the report should include team number, e.g.: `team1001-final-report`**`.pdf`**
   b. Supplementary materials:
      i. **Conceptual model / database schema**: Provide these (e.g., your ER-diagram and/or your SQL schema) as separate files
      ii. If you used OpenRefine for data cleaning:
      Operation History: A copy of the OpenRefine operation history (copy-paste it into a json file named OpenRefineHistory.json)

If you used **other tools,** e.g., Trifacta Wrangler or Python programs: include those scripts or programs and any other auxiliary files required.

iii. **Queries**: A copy of the queries written in SQL or Datalog that you used to profile the dataset and/or check integrity constraints (copy-paste them into a text file named Queries.txt or Queries.sql).

iv. **Outer and Inner Workflow Models:** For both workflow models, provide the necessary files. For example, if you've used YesWorkflow for defining the model, then please include the following: the YW annotations files (OverallWorkflow.yw), and the YW-generated Graphviz/DOT file (OverallWorkflow.gv). If you've used another diagramming tool, then include your workflow diagrams as separate files as well.

v. **Raw and Cleaned Datasets**: Please do **not** include the datasets in the ZIP file! Rather, upload the raw and cleaned datasets to a Box folder and share the link in a plain text file (DataLinks.txt). Make sure the Box folder link is open to the public (or at least to Illinois.edu-registered users)!

6. [up to 5 Pts] **Bonus/Extra Credit.** Occasionally, we may give extra credit for particularly informative elements, e.g., use of visualizations, data analysis/data mining, etc.