CS513: Theory & Practice of Data Cleaning
Final Project Instructions (Summer 2024)

Bertram Ludäscher
ludaesch@illinois.edu
University of Illinois, Urbana-Champaign

The goal of the group project is to conduct an end-to-end data cleaning project, using the various tools and techniques that we have covered throughout the course. In addition to the main tools that we used in class (i.e., RegEx, OpenRefine, Datalog (new: Logica), SQL, and Python), you are welcome to use other tools. For example, you may want to use any of the research prototypes mentioned (e.g., YesWorkflow or the OpenRefine companion tools such as or2yw), or commercial tools (e.g., Tableau). Or maybe you find a new way to use an LLM (e.g., ChatGPT). In all these cases your report will have to include sufficient documentation about how you used these tools.

Project Phases. The project is organized into two phases, each with its own set of deliverables. See Coursera for the timeline and details.

Project Phase-I

During this phase your team needs to . . .

1. **Describe the provided dataset** D.[1] You can provide a **conceptual model** (ER diagram) that depicts the entity and relationship types, or an **ontology** that illustrates all classes and their relationships. Or you can provide a **database schema** that illustrates and explains the structure and contents of the dataset. You should also add a short **narrative**, i.e., one or more paragraphs in English to describe the origin of the data and any relevant metadata (e.g., a **temporal** or **spatial extent**). If you had a dataset about farmers markets, e.g., it could be described with a relational schema (e.g., `CREATE TABLE` statements); the narrative would then explain what the different columns (attributes) mean. Other metadata may describe, e.g., the spatial extent of the data (only Illinois markets? All of the Midwest? Or the US?), and the temporal extent (for which period is the data correct?), etc. Ditto for D.
2. **Develop three use cases.** By *use case* we mean a written scenario describing a hypothetical data analysis. (If helpful, you can think of these as a queries or questions asked of the dataset, see **Additional Information** below).
   (a) **Target (main) use case**: U_1 for D such that data cleaning is *necessary* and *sufficient* to support the data analysis use case. Thus, after performing data cleaning, your cleaned data D' is *fit-for-purpose* (i.e., for U_1).
   (b) **"Zero data cleaning" use case**: U_0 should be a use case that requires "zero data cleaning", i.e., D is "good enough as it is".

[1] NYPL data. The README for this is in the parent folder.

(c) **"Never enough" use case**: U_2 is a use case for which the given dataset D is "never (good) enough", i.e., no amount of data cleaning or wrangling will make D suitable for U_2 (even though at first sight one might think so).

(d) **Note**: The purpose of the corner cases U_0 (data cleaning is **not necessary**) and U_2 (data cleaning is **not sufficient**) is to reinforce the concept that **data cleaning should be done with a purpose in mind**, i.e., a use case such as your main use case U_1, where data cleaning really makes a difference.

3. **List obvious data quality problems** (i.e., which are easy to spot during Phase-I). In order for your dataset D and main use case U_1 to match, data cleaning must be necessary and sufficient to implement U_1. You need to **support this claim** by **documenting** data quality problems that your inspection of D has revealed and that need to be addressed before U_1 can be tackeled.

4. **Devise an initial plan** that outlines how you intend to clean the dataset in Phase-II. A typical plan for the overall project will include the following **steps**: S_1: description of dataset D and matching use case U_1; S_2: profiling of D to identify the quality problems P that need to be addressed to support U_1; S_3: performing the data cleaning process using one or more tools to address the problems P (here you should describe which tools you are planning to use, e.g., OpenRefine; Python; etc.) S_4: checking that your new dataset D' is an improved version of D, e.g., by documenting that certain problems P are now absent and that U_1 is now supported; S_5: documenting the types and amount of changes that have been executed on D to obtain D'.

   You should also include a tentative **assignment of tasks** to team members (who does what)!

Additional Information

Regarding (2): How do you specify data analysis **use cases**? Generally speaking, you can simply explain the use case in a short paragraph. You might also want to be more specific and phrase use cases as **questions**: *What is it that we want to know from (or about) the data?*

In particular, a use case may be a set of database **queries** $Q_1, \ldots, Q_n$ against the dataset D (e.g., how many farmers markets offer bakery goods in addition to vegetables and fruits?) On the other hand, use cases may also be more general, e.g., you could state that you'd like to develop a web application that serves a particular purpose.

The advantage of specifying a use case U as one or more queries Q_U is that you can be very precise about when data cleaning is necessary and sufficient for U: if running Q_U on the original ("dirty") data D would result in an answer $A = Q_U(D)$ that is incorrect and/or misleading, then data cleaning is **necessary**. Conversely, data cleaning is **sufficient** if the answer $A = Q_U(D')$ on the cleaned dataset D' is correct (and not misleading).

In (3) above, how do you document data quality problems? One simple way is to include (copy-pasted) snippets of "dirty data" in your Phase-I report (you can also use screenshots for illustration) and then explain what the problem is in narrative form.

How do you describe your **plan** in (4)? A short list of your planned steps $S_1, \ldots, S_5$ will do during Phase-I. In Phase-II, you should also include a **workflow diagram** for the actual data cleaning steps that you performed (e.g., with YesWorkflow or any other diagramming tool). Of course your Phase-I **plan** and your **actual** Phase-II **workflow** might be different.

Project Phase-II

During this phase you will **execute** the plans you've come up with in Phase-I (possibly adjusting course, based on what you find when actually working with the data . . .)

What to Submit

Phase-I:

- **A single PDF file with your Phase-I report** (in narrative form) with all elements from the list above.

Phase-II:

- **A single PDF file with your Phase-II report**. This report should include:
  - A description of the **actual data cleaning workflow** W that was performed, and a **comparison** with the original Phase-I plan: e.g., were you able to execute the steps as planned, and if not, what did you have to change and why?
  - A **narrative** that ties all steps together and explains the motivation (use case U_1), the rationale for the design of the overall workflow W and the tools used.
  - **Documentation** that data quality was improved, e.g., through running "before queries" $Q_U(D)$ and "after queries" $Q_U(D')$ on D (original) and D' (cleaned), respectively.
  - A summary of the **data changes** ΔD resulting from the overall workflow W: $D \rightsquigarrow D'$.
  - A summary of **findings**, **problems encountered**, and **lessons learned**, including **possible next steps** (e.g., how would you implement the main use case U_1).
- **Supplementary Materials**. In addition to the project report, you need to provide the following supplementary materials (as a single ZIP file):
  1. **Workflow Model**: For the **overall** workflow model W, when using YesWorkflow, provide the text file that has the YW annotations (e.g., `Workflow.yw`), and the generated Graphviz (dot) file (e.g., `Workflow.gv`). For other diagramming tools, provide a source file (e.g., PPTX, . . .) and a PDF file (`Workflow.pdf`).
  2. **OpenRefine Operation History**: If you used OpenRefine, then include a copy of the *operation history* (copy-paste it into a `json` file named `OpenRefineHistory.json`).
     - If you also want to visualize the OpenRefine history, you can use the `or2yw` tool.[2]
  3. **Other History**: If you are using an **alternative tool** (instead of, or in addition to OpenRefine), please provide an analogous file (`OtherToolHistory.json`) and other provenance information **if available for that tool**: e.g., include Python (or R) scripts, Jupyter notebook files, etc.
  4. **Queries**: A copy of the queries written in SQL or Datalog to profile the dataset and check the integrity constraints (copy-paste them into a **text file** named `queries.txt`).
  5. **Original ("dirty") and Cleaned Datasets**: Please **do not** provide the datasets in the ZIP file. Rather, upload the raw and cleaned datasets in a Box folder and **share the link** in a plain text file (`DataLinks.txt`).

[2] `https://github.com/idaks/OR2YWTool`