Final Project: End-to-End Data Cleaning Workflow

One way to think about the final project and the associated report is as follows: Imagine you have been hired by a client to perform data cleaning for a particular use case and you're now documenting what you've done: Can you explain clearly what you've done? Can you document and explain your decisions? Can you provide evidence? With this in mind, here is a possible structure / outline for your report:

- **1. Introduction and Overview.** Briefly describe your data cleaning project, and state the main aim(s) or goals. In particular, you should describe a (hypothetical or real) **use case** of the dataset, e.g., what data analysis you would like to do after cleaning the data. Your target use case should also lead you to some specific **data cleaning goals** that can achieve the desired **fitness for use** of this dataset. In addition to your main use case, also answer these questions:
  - Are there use cases for which the dataset is already "clean enough"? That is for those use cases you would not have to do any data cleaning.
  - Conversely, are there use cases for which the dataset will never be good enough? That is, while at first it would appear that your dataset can be used, upon closer inspection (see below) one would find that there is no hope to support that use case (e.g., because the information is not complete enough, of too low quality, or not of the right kind).

- **2. Initial Assessment of the dataset**. Here you should describe the **structure** (i.e., schema) and **content** of the dataset and **quality issues** that are apparent from an initial inspection. A good way to describe the structure is via an ER diagram. You can then use the entities and relationships of the ER diagram to explain the content of the concrete table(s) you have. The quality issues (problems) should be described in narrative form and examples used for illustration as needed.

- **3. Data Cleaning methods and process**
  - (a) For most projects, we encourage you to use **OpenRefine**, unless there are good reasons not to. Document how you have used OpenRefine and explain both what worked well and what problems you might have encountered when using the tool.
    - Document the results of this phase, both in narrative form and with supplemental information (e.g., which columns were cleaned and what changes were made?) Can you quantify the results of your efforts? Also provide provenance information from OpenRefine. Pay close attention to what OpenRefine includes and does not include in its **Operation History**!

    - If important information is missing in the latter, provide that information in other ways, e.g., explaining things in the narrative.
  - (b) If you find that certain steps are not well suited for OpenRefine (e.g., due to scalability or other issues), consider applying an alternative solution, e.g., using **Python**, **R**, or another tool such as **Trifacta Data Wrangler, Tableau**, etc. Document your choice and provide the corresponding similar artefacts as OpenRefine (i.e., narrative and supplemental files). If you chose a script-based alternative (e.g., Python) we encourage you to provide a Jupyter notebook as well. That way, you can combine narrative text, explanations, and outputs along with the actual code in one place. Your report should then also include a link to your notebook (output).

- **4. Data Cleaning Results**
  - (a) Develop a **Relational Database Schema** for your dataset. What logical integrity constraints (ICs) can you identify? Load the data into a **SQLite** database with your target schema. Use SQL queries to profile the dataset and to check the ICs that you have identified! Think of the queries as denials that retrieve those rows that can be used as “witnesses” of the IC violations.
  - (b) Create a **Workflow Model** of your data cleaning workflow: What are the key inputs and outputs of your overall workflow? What are the dependencies? Create a visual version of your overall workflow using the yw tool. If you’ve used OpenRefine, also use the **or2yw** tool to create a YW model from your OpenRefine recipe.
  - (c) Provide a clear way to showcase the overall changes you’ve done with the dataset. This can be tables that documents how many cells you changed in what steps of the cleaning process. Or this can be a way to show the dataset before and after cleaning (something similar to the ‘diff’ codes you’ve seen in our OpenRefine assignment).

- **5. Conclusions and Future Work**. Summarize the main outcomes of your data cleaning project -- what are the takeaways? Do you think the clients that asked you to clean the dataset will have a clear sense of what you’ve changed? Have you encountered any problems or challenges along the cleaning process? Describe them. What would be your next steps for the project if additional time was available? (e.g., for additional data cleaning or for subsequent analysis).