Project Phase-I Report Outline and Checklist

We strongly encourage you to follow the Phase-I report outline below, as it aligns well with the checklist and grading rubric.

The title / header of your Phase-I report should list (i) the **Team-ID** of your group[1] and (ii) for each group member the **name** and **Illinois email address**.

Report Outline / Checklist

The **project instructions** (separate document!) describe what exactly you should do as part of Phase-I of the project. The following outline should be followed for your Phase-I report:

- ☐ **1. Dataset Chosen** (**5 points**)
  a. Name the dataset you will be using for your project.
- ☐ **2. Description of Dataset**
  a. Here you will provide an ER diagram, an ontology, or a detailed database schema (**10 points**), and
  b. a **narrative** description of the dataset covering structure and content (**15 points**)
- ☐ **3. Use Cases**
  a. "Zero cleaning" use case U0: data cleaning is *not necessary* (**5 points**)
  b. "Main" use case U1: data cleaning is *necessary* and *sufficient* (**20 points**)
  c. "Never enough" use case U2 : data cleaning is *not sufficient* (**5 points**)
- ☐ **4. Data Quality Problems**
  a. List obvious data quality problems with evidence (examples and/or screenshots) (**20 points**)
  b. Explain why / how data cleaning is necessary to support the main use case U1 (**10 points**)
- ☐ **5. Initial Plan for Phase-II** (**10 points**)
  a. Below is a possible plan, listing typical data cleaning workflow steps. In your *Plan for Phase-II*, fill in additional details for the project **steps** as needed. In particular, include **who of your team members will be responsible for which steps**, and list the **timeline** that you are setting yourselves!
    - S1: Review (and update if necessary) your use case description and dataset description
    - S2: Profile *D* to identify DQ problems: How do you plan to do it? What tools are you going to use?
    - S3: Perform DC "proper": How are you going to do it? What tools do you plan to use? Who does what?
    - S4: Data quality checking: is *D'* really "cleaner" than *D*?
      - Develop test examples / demos
    - S5: Document and quantify change (e.g. columns and cells changed, IC violations detected: before vs after, etc.)

[1] In addition you can give yourself a (cute) team name for ease of identification ;-)