Project

Context

The following two datasets contain US hospital data.

- hospital_1.csv
- hospital_2.csv

After an interview of the data owner, we obtain the following information about the two datasets

- Many hospitals have similar names across different cities (Saint Lukes, Saint Mary, Community Hospital)
- In urban areas, hospitals can occupy several city blocks so addresses can be ambiguous
- Hospitals tend to have many clinics and other associated and related facilities nearby
- Hospitals also get acquired and name changes are common

To do list

- Carry out some profiling analyses on the two datasets
- Carry out some pre-processing technique that you consider necessary
- Match the two datasets according to different approaches
- Evaluate, compare, and analyze the matching performance of chosen approaches

Deliverables

- python code with comments (Jupyter Notebook or any other format)
- a short report (maximum 2 pages in MS Word) of lesson learned from this project