Final Project Progress Report

Warren Lim
DSCI 510 - Fall 2025
Predicting Hospital Readmission Rates Using Healthcare Quality and Cost Data

Project scope update
The project scope is still consistent with the original proposal. The goal is to predict 30-day hospital readmission risk using publicly available CMS datasets and the Kaggle Hospital Readmissions dataset. No major scope changes have occurred since the proposal. Current focus is on accessing and exploring the CMS Hospital Readmissions Reduction Program (HRRP) API data to validate data retrieval and understand its structure before model development.

Data sources
Data was successfully retrieved from the CMS Hospital Readmissions Reduction Program (HRRP) API endpoint. With pandas, the script src/load.py loads 18,510 rows of hospital-level readmission data directly from the API in CSV format. A simple test in src/tests.py verifies the data load and ensures the dataset is non-empty. The next steps are to integrate additional CMS hospital information and the Kaggle patient-level dataset.

Issues / difficulties
The CMS API initially returned nested JSON that needed switching to a CSV endpoint for proper parsing. Minor data-cleaning challenges are expected, including handling missing fields and merging multiple datasets with inconsistent hospital identifiers. In a way, no major blockers have happened so far.