STAT 390 Weekly Progress Report 1

September 25-29

Cindy Ha, Willie Xie, Erica Zhang

Progress/Accomplishments

- Set up Github Repo
- Found dataset from Kaggle that met requirements
  - 67 variables
  - 341,408 observations from Jan 2020 to Sep 2023
- Added dataset to raw data folder and a dictionary for variable definitions
- Completed quick skim of dataset to check usability
  - 1 date, 4 categorical, 62 numeric variables
  - Found missingness issues that will need to be addressed (15 variables > 80% missing)

Plan/Next Steps

- Work on proposal to present on **Thursday, Oct 5th**
  - Develop a prediction question we want to answer regarding Covid-19
    - Determine target variable
    - Decide between the appropriate regression or classification models
    - Complete any necessary data pre-processing steps (e.g. mismatched data types, missingness, etc.)
- Set up timeline to complete project by deadline
- Obtain any necessary sources/references to help answer our question
- After getting approval to move forward, conduct a thorough EDA
  - Split data into training and testing sets

We are optimistic about the current progress of the project and plan to meet regularly.