Weekly Progress Report: Oct 9 - Oct 13

Contributors: Ryan Choe, Rakin Hussain, Sid Taneja

Progress/Accomplishments:

- Exploratory Data Analysis performed by all three members
- Rakin identified missingness proportions for all the columns/variables in the dataset. By setting a threshold to < 0.2, our group tried to see which variables have less than 20% of their values NaN or missing
- Sid and Ryan created visualizations for a handful of specific countries, looking at how new_deaths, total_deaths, new_cases, total_cases changed for these countries over time. Evaluation/Analysis of visualization shape have sparked new ideas for next steps in project (ie. how to account for uninformed/empty data from the beginning of the time series).
- Created functions to look at each country in the dataset in a modular fashion

Problems/Challenges:

- Trouble with creating functions for any country. Sorted out during Thursday's class and a bit after.
- Needed to understand the pushing/pulling process to the main branch in GitHub so all three of us could access each other's code via the updated main branch.
- Still need to decide what to do about the missing values in our variables of interest (new_cases, new_deaths, total_cases, total_deaths). Juggling between imputing and other options.

Plan/Next Steps:

- Prioritize steps/process for imputing on missing data
- Divide data into training dataset (2020-2021 data) and testing dataset (2022 January - March data). Need to discuss with TA and Professor about how to do this for time series (briefly mentioned in work class Thursday).
- Begin planning to construct tree-based models after deciding which variables are used and how to deal with their missing values.
    - Additionally, if we use a decision tree we cannot involve future time in the training data. Will need to confirm/check this with Professor next week

Overall, the team has started to delve deeper into some Exploratory Data Analysis. The group needs to, ideally, conclude by the end of next week how to proceed with missingness in all the variables and which variables will be used in predictive modeling.