**Weekly Progress Report (Oct 16 – 20): Tilly, Victoria, Lalla-Aicha**

**Progress/Accomplishments:**

* Tilly reinvestigated the data imputation process done last week and figured out that there was a data leakage problem. To determine whether some features were worth imputing, she checked the EDA done by Lalla and Victoria. She then test-train split the data before using SimpleImputer and KNNImputer to impute missing values for only the features she deemed were important.
* Lalla started exploring possible tree models. She first chose a sample region to work with since it would be quicker to compute and then disregarded any classification variables (social distancing grades, etc.). From there, she was able to select a sample of features to put into a random forest model and began to tune it. Right away, she noticed that this model would not be a good fit for the data as the predicted varies greatly from the actual data points.
* Victoria undertook the time series analysis to predict the confirmed cases. She initially adopted a naive approach focusing only on the average monthly confirmed cases and tried different sets of ARIMA parameters. However, she recognized the importance of the most recent data, as opposed to the mean. Thus, she refined her approach and placed a greater emphasis on the confirmed case. She used auto ARIMA and redeveloped the time series model using a lag of one week, aiming to predict the subsequent day's confirmed cases based on the past week's data.

**Next Steps:**

* Prepare for the EDA presentation next week
* Start building basic models to predict the death rate by COVID-19
  + Tilly: LSTM model
  + Lalla: Tree-based model
  + Victoria: ARIMA model