**Weekly Progress Report (Oct 30 – Nov 3): Tilly, Victoria, Lalla-Aicha**

**Progress/Accomplishments:**

* Tilly redid all data preprocessing steps after realizing that the data needed to be aggregated by date to be able to predict on the region level. She first split the data into 2 regions of interest (New England and South). She then grouped the data by date and found the mean for all numerical features, mode for all categorical features, and sum of the target variable (covid\_19\_confirmed\_cases). This meant the data no longer had any duplicate dates and were on region-level instead of county-level. Tilly then proceeded with feature engineering and other preprocessing steps, such as scaling and one-hot encoding.
* Lalla redid the EDA for our newly preprocessed data. First, she checked the correlations between our feature engineered columns and our confirmed cases one. Then, she conducted a test to find any outliers in our test data for New England and the South. Due to our new method of pre-processing and the aggregation, there was only one outlier for New England, and no outliers for the South. However, univariate EDA on confirmed covid cases revealed vastly different distributions of cases for each region. For New England we saw the median to be around 2300, while for the South it was around 1600, with the highest number of cases going up to 27500. Lalla also conducted a multivariate analysis for each region.
* Victoria conducted further data preprocessing after the data was re-aggregated at the county level to be more suitable for ARIMA analysis and conducted exploratory data analysis of the confirmed cases and death cases in New England and the South to have a general idea of the trend and pattern. Then, she visualized the correlation between different features and eliminated highly correlated. For the time series analysis on COVID-19, she built ARIMA models by 1) examining autocorrelation and conducting differencing to make the series stationary, 2) Performing Augmented Dickey-Fuller tests to determine the order of differencing, 3) confirming that second-order differencing was suitable for the New England region's data.

**Next Steps:**

* Tilly, Lalla, and Victoria will go ahead and build 3 models (ARIMA, Auto-ARIMA, and XgBoost) using the newly processed data
* Tilly, Lalla, and Victoria will meet up next week to discuss their finished models and improve them using insights gained from looking at each other’s models
* Tilly, Lalla, and Victoria will then continue with the other 3 models and meet up again later next week to check on those