With the COVID-19 pandemic sweeping the entire world in 2020, I set out to study the spread of the virus by examining factors of countries that led to the spread of COVID-19 being contained within certain countries or spreading by one certain date, 11/27/2020. The way in which I measured how much spread each country had was by: Total Cases per one million people in the country, total deaths per one million, and death rate. These were my target variables. I used data from Our World in Data COVID numbers, which provided several country factors, as well as mobility data from Google Mobility Reports, which determined on average how much the citizens of the country changed their behaviors such as going to the retail store, going to work, etc. during the pandemic.

First off, I began the capstone by reading in the OWID (Our World in Data) csv file followed by the Google Mobility csv file. I then wanted to find the date of the first case in each country since it varied and could change the perception of the data since certain countries may have started their outbreaks later but handled it worse since the first case. I made a new dataframe for the first case and days since the first case and merged them making a new column for them on the main OWID data frame. I filtered the data for the date of 11/27/20 and then merged the data with the mobility data as well to get the mean mobility change as well as the peak for each category for each country to have two different ways to represent the mobility data rather than just one. I then cleaned up the dataset and also merged the data with another dataset from OWID which had the tests per million data which I thought would be useful for my models.

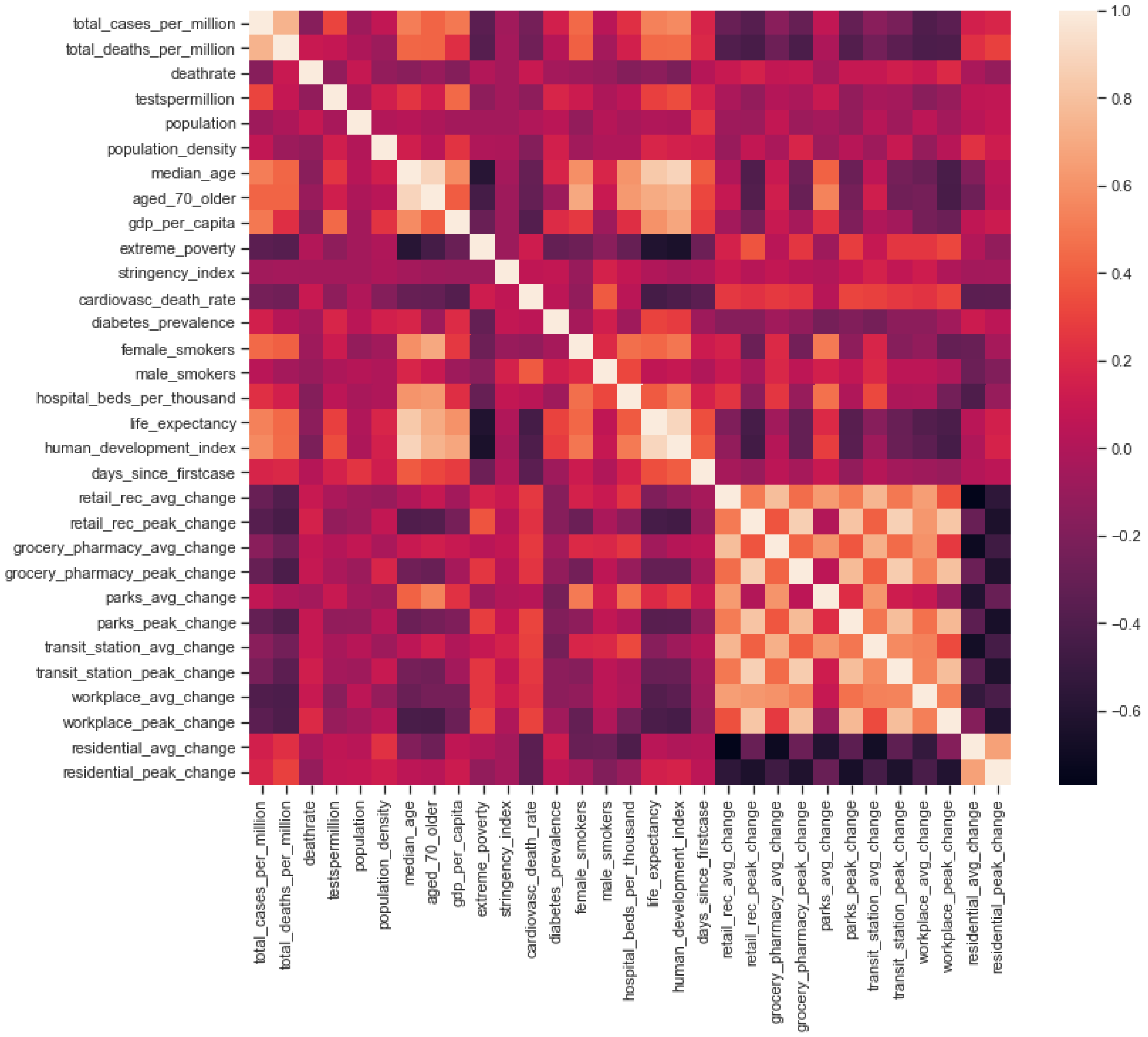
Next came, filling in the null values. I first had to set my index to the column so I could see where the missing values were rather than numeric values. I used the sum of null values by column as well as the missingno graphic in order to visualize missing values until I cleared them all. First, I deleted all countries with more than half of the columns missing and deleted the handwashing column because it had an excessive amount of null values. To fill in the null values, I looked at alternative sources and filled the null values to the best of my ability so that they would as accurately fit the data as possible and would be better than simply using the mean. For the rest of the null values I simply filled them in using the mean.

I then started my visualization process. First, I made a heatmap and used the corr() function to visualize and look at the correlations. The heatmap is shown below and this along with the numbers shown in the corr() function show a few major observations:

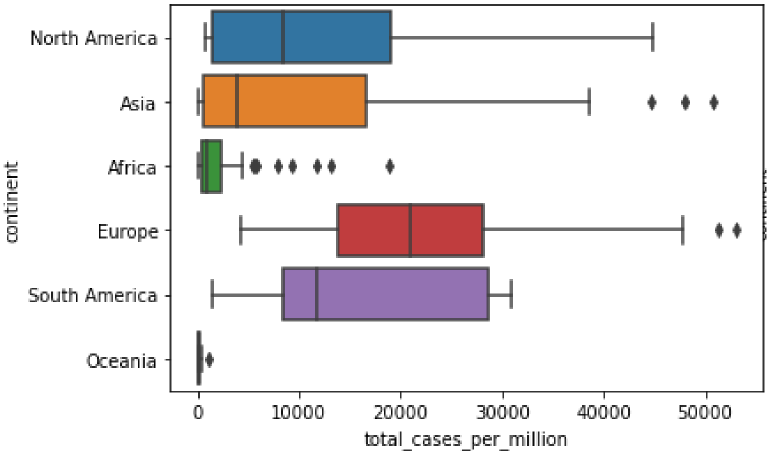
Total cases per million strongly positively correlated with median age, gdp_per_capita, female smokers, human development index, and life expectancy, and strongly negatively correlated with extreme poverty and every type of mobility change especially workplace change and retail/recreation change.

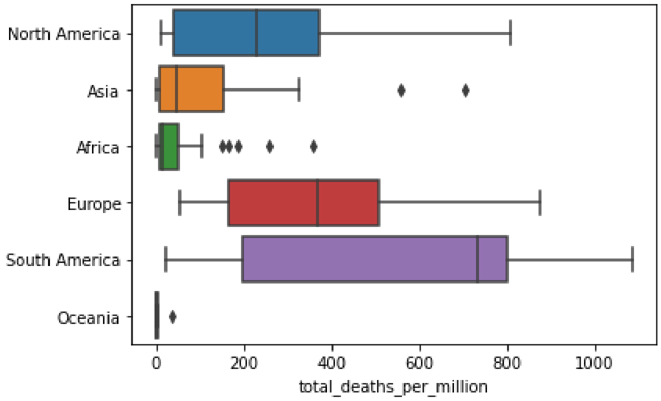
Total deaths per million strongly positively correlated with median age, female smokers, human development index, and life expectancy, and strongly negatively correlated with extreme poverty and every type of mobility change especially workplace change and retail/recreation change.

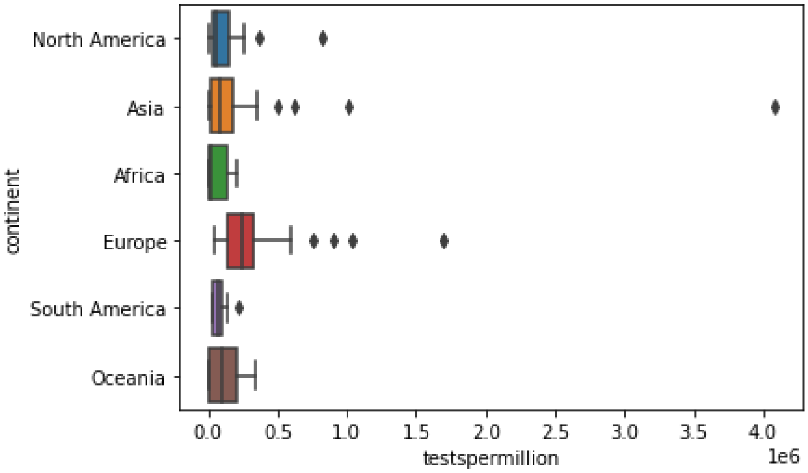
Death Rate is not strongly correlated with any variable, which is already not a great sign for being able to use this model to predict the death rate.



I then had to reset the index to numeric values and remove the non numeric columns so that I could run scatterplots and other graphs. I called this new temporary dataframe with only numeric values “numeric”. I plotted scatterplots of target variables along with all the predictor variables using the pairplot method. I also made a distplot for every column to gauge the overall distribution. Some of the variables had close to a normal distribution but some others started with several values of near zero and progressively became less amount of high volumes of that variable such as cases/deaths by country since several countries had none altogether. I followed this with box plots as another way to show the distribution along with histograms and boxplots of the target variables by continent.

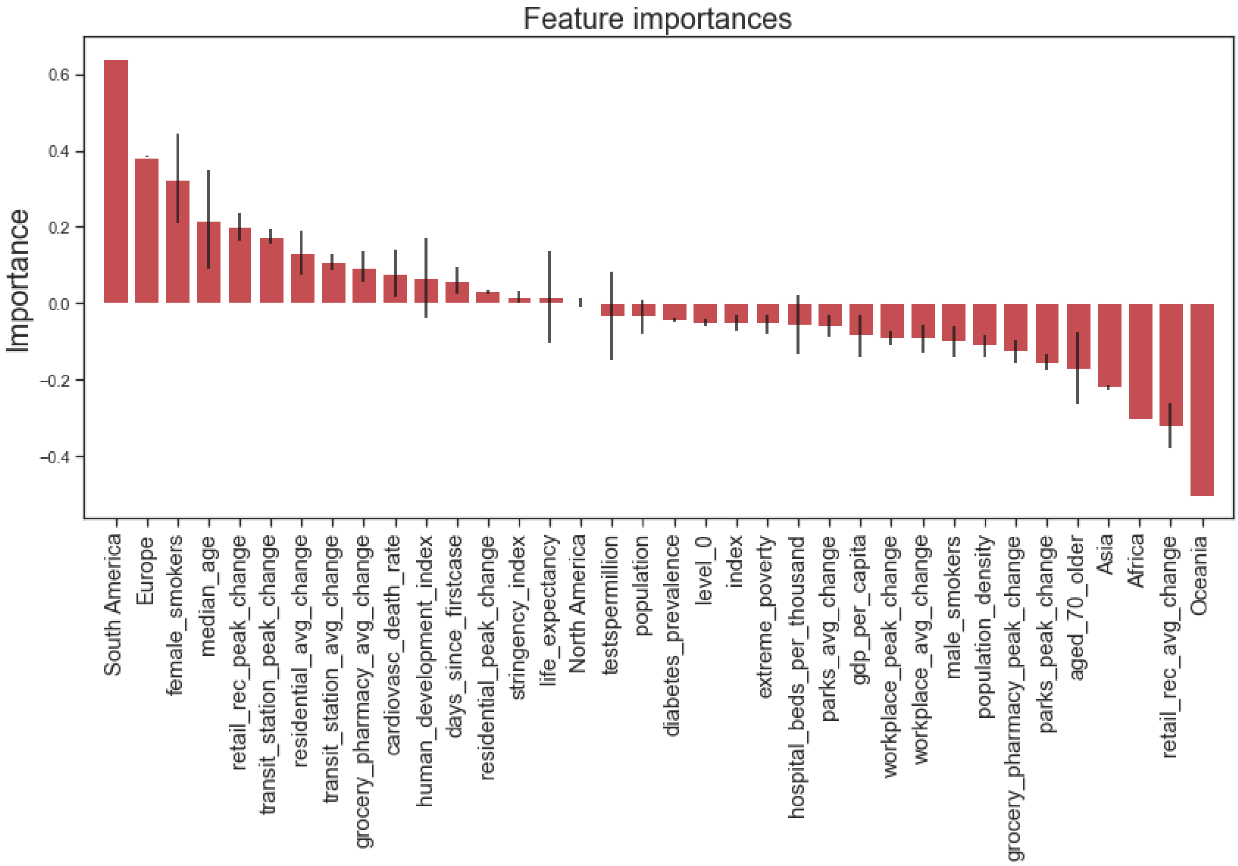




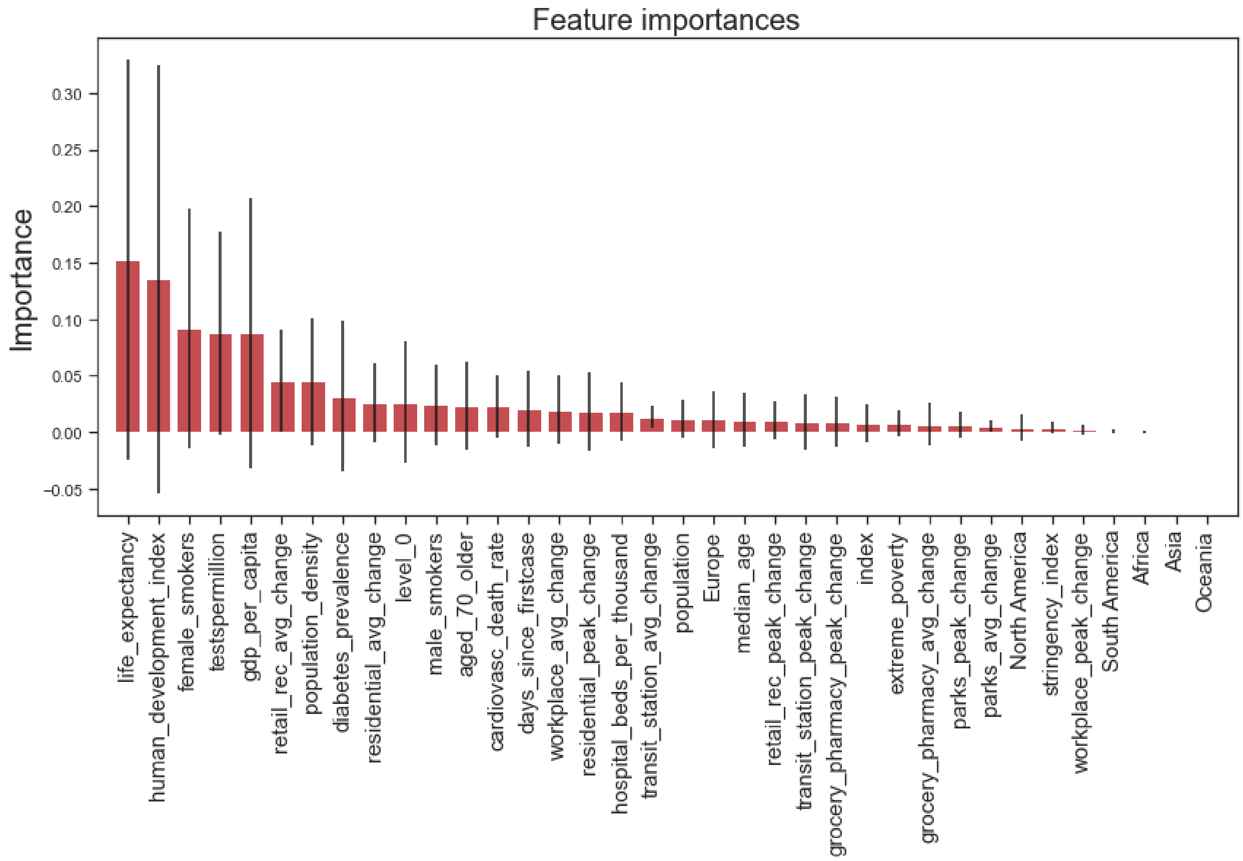


Through some of the boxplots we see here strikingly that Europe and South America seem to have the biggest spread of COVID-19 in terms of both cases and deaths. South America leading with deaths per million and Europe leading with cases per million. Europe leads testing decisively but still is relatively high in cases and deaths, which is surprising. Another observation I would make here is that generally based on total numbers it seems like Asia would have the most cases but this is actually because Asia's population is so large- 59.51% of the world, so relatively they are not doing terribly. Africa seems to be doing the best at managing the virus.

I then started my modeling process by training and preprocessing the data. I first made dummy variables to be able to classify my classification variables. I set the value X as every predictive variable, Y equal to cases per million, Y2 equal to deaths per million and Y3 equal to death rate. I ran a linear regression, random forest, and gradient booster for each Y value and returned a r-squared value which determines the score of the model based on how much of it the model can predict as well as a scatter plot with the predicted y values on the x axis and the y-test values on the y axis. For the random forest and gradient booster methods, I used GridSearchCV which is a way to optimize the hyperparameters for the model. For each target variable, I chose the best performing model of the three. What was found was that the best model for predicting Y or cases per million was the Random Forest model, the best model for Y2 or deaths per million was the Random Forest was the Linear Regression. The death rate had no model with an r-squared higher than 0.0296, thus as we predicted earlier by looking at the spread of the death rate, this data will not be sufficient to predict the death rate of the disease. I was able to include a list of feature importances for the best Y and Y2 models. The importances for the Linear Regression model for Y2 or total deaths per million was:



The feature importances for the Random Forest for cases per million was as followed:



Some things that stand out are that for the linear regression of deaths per million, the continents mattered significantly. Female smokers,median age, retail/recreation change, transit change, residential change were positively correlated with deaths per million. It is interesting that female smokers in particular were such highly correlated with this and not male smokers. One reason for this could be potentially that females smoke far less than males so the countries where females do tend to smoke more have a higher chance of causing COVID tests, while male smokers are more prevalent in each country as shown by the distribution. Median age being positively correlated with deaths from COVID also makes a lot of sense given countries with older citizens would intuitively have more deaths from this virus which affects older people more.

For cases per million, the most important factors were life expectancy, human development index, female smokers, tests per million, and gdp per capita. Tests per million makes sense because the more tests there are the more reported cases there will be. However it is very interesting that human development, life expectancy, and gdp per capita are all important but negatively correlated with cases. It seems counterintuitive and is a curious case.

Overall this capstone provided us with helpful information through graphs and models about which factors in countries led to handling the Coronavirus pandemic well and which did not. I was able to predict how a certain country did through November of 2020 in this pandemic through the data I used.

Future Questions:

1. Which other data could we include to improve the death rate data to potentially find variables to be able to model the death rate in a country? Or is death rate simply consistent throughout the world?
2. Why exactly are COVID cases and deaths positively correlated with human development index, GDP per capita, and life expectancy? In other words, why is it that developing countries seemed to be doing worse with this virus? This seems counterintuitive since one might expect it to be the opposite case.
3. How can we use this data to help contain the next pandemic better, or potentially stopping it from becoming one at all?