Project Report

Names: Akshat Javiya, Valentyna Shyyan, Tauheed Janjua

Data Wrangling:

There are several important data-wrangling steps that need to be completed before any modeling can begin. The first step is to load the data from the csv into two different data frames. The next step is to filter out the data that does not have a three-character iso code and a population of less than a million. After this, the next step is to populate the new deaths in the advance column two weeks in advance. Then, it is necessary to tidy the world population data, remove the Series Name column, and pivot wider to ensure each observation is in one row. Lastly, the two data frames will be joined based on the three-letter country code.

Linear Modeling:

After getting Data Wrangling, the data is good for making predictions and creating linear models. The strategies that we used in picking variables were simple. We opted for variables that are likely to cause death in COVID patients. There was a catch because some variables were NA, so they were available only for developed countries. For transformed variables, we choose new_deaths_smoothed, population_density, SP.POP.80UP.FE, SP.POP.80UP.MA, life_expectancy, and human_development. New_deaths_smoothed was important because we were going to predict that column two weeks ahead. Population_density helps to detect if the population density is higher or not; if the density is higher, then the chances of more cases and more deaths increase. SP.POP.80UP.FE and SP.POP.80UP.MA are important because, in these columns, people have higher chances of getting COVID and dying due to health issues or due to their age. Life expectancy and human_development are also important because these factors give a better understanding of a country's health system and the availability of health systems to people. For linear regression, we relied upon other variables like gdp_per_capita, diabetes_prevalence, total_deaths, excess_mortality, median_age, handwashing_facilities, total_vaccinations, population, cardiovasc_death_rate, male_smokers, female_smokers, extreme_poverty, and new_vaccinations_smoothed_per_million. Gdp_per_capita, handwashing_facilities, excess_mortality, and extreme_poverty are part of predicting a country's position in affordability, stability, and caution in preventing COVID-19. Diabestes_prevalence, cardiovasc_death_rate, male_smoker, and female_smokers are inter-connected to each other because people who fall under these columns have a high chance of dying from COVID due to their condition. Since COVID-19 is a respiratory disease, people with heart or lung problems have a high chance of dying. new_vaccinations_smoothed_per_million, total_vaccinations, population, total_deaths, and median_age; these columns help if the COVID-19 affected median_age and population; other columns help us indicate whether the COVID-19 vaccination helped the population in slowing down covid.

Evaluation & Conclusion:

It is very difficult to predict something as unbeknownst as deaths in the future because there could always be so many factors, like hunger, economy, war, etc., that can change an entire country's death rate. However, based on factors that we do know, like life expectancy, deaths in the past, disease rates, poverty rates, and more, we can get a very accurate idea of what the future could look like. After generating the r-squared and root mean squared, we get this

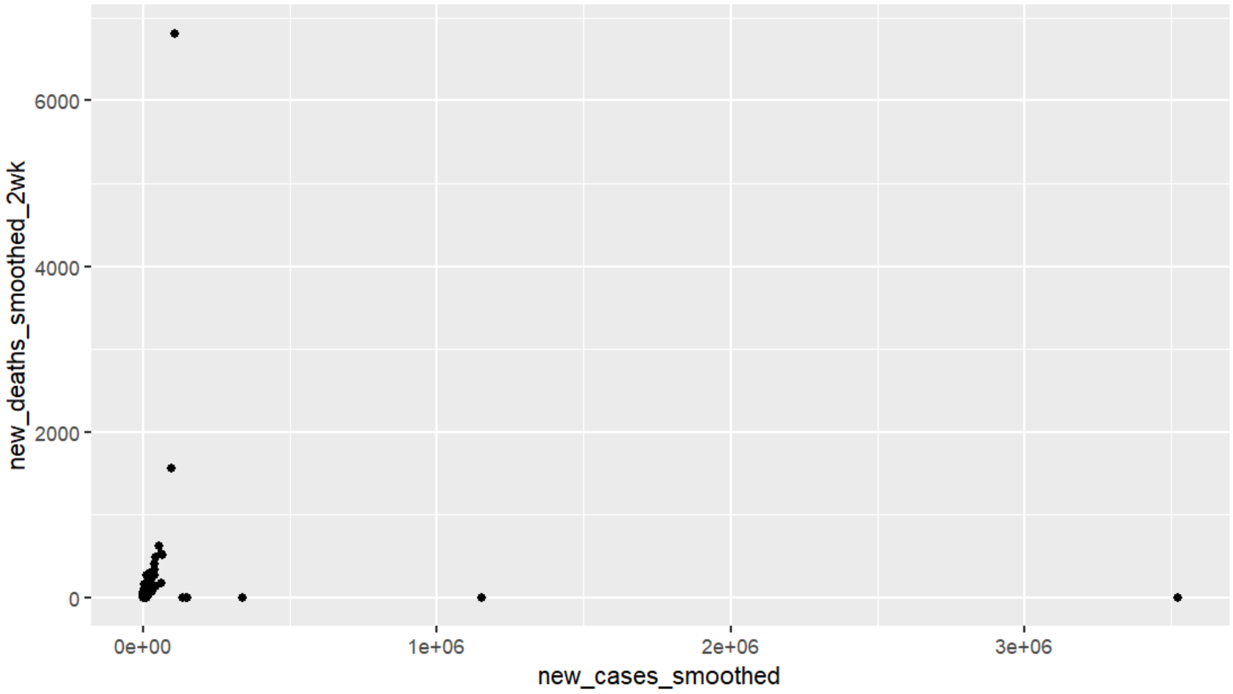
Model	R2
Model 1	0.1892
Model 2	0.8682
Model 3	0.5865
Model 4	0.04861
Model 5	0.02603

Model	RMSE
Model 1	149.4905
Model 2	168.5743
Model 3	69.10188
Model 4	169.1687
Model 5	30.23559

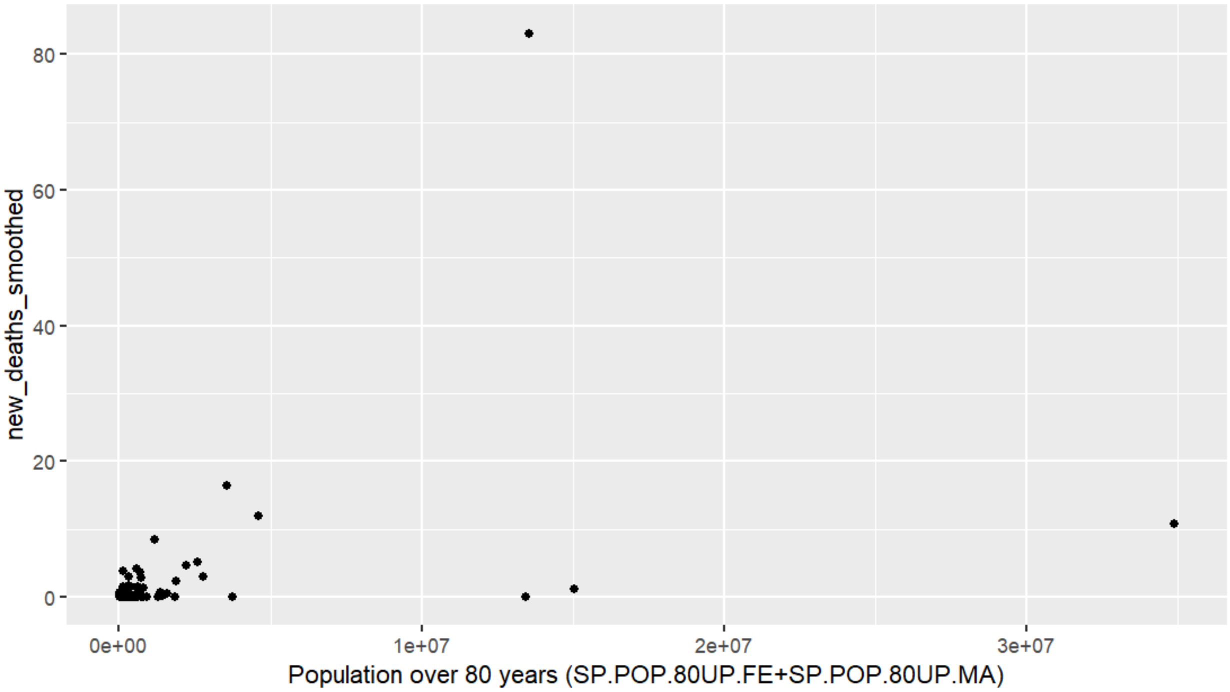
Countries	RMSE
China	1902.917523
India	7.989310
United States	88.476485
Indonesia	6.748229
Brazil	39.527279
Mexico	7.616738
Japan	406.561711
Philippines	15.315313
Germany	24.194039
France	17.893073
United Kingdom	17.208719

Italy	13.520740
Colombia	7.260634
Spain	7.650032
Canada	13.086806
Peru	17.101437
Australia	62.549438
Chile	7.312262
Sweden	10.923394
New Zealand	9.230213

The scatterplot that was generated based on the most recently available new_deaths_smoothed_2wk and new_cases_smoothed for each country is as follows.



The scatterplot of only the most recent new deaths per day and the urban population is as follows


new_deaths_smoothed
80
60
40
20
0
0e+00
1e+07
2e+07
3e+07
Population over 80 years (SP.POP.80UP.FE+SP.POP.80UP.MA)