Description: For this project, the problem I chose to address is understanding the influences that affect COVID19, particularly total_deaths_per_million. By influences, I mean predictor variables that contribute to the severity of total_deaths_per_million per each country in the world, given three following categories: age variables, health, and quality of living/infrastructure. COVID19 is a modern pandemic of unprecedented size, with many countries still in critical states today. Given the widespread nature of the virus, understanding its effects through data is crucial. We live in an era where society is data-driven, and the collection, visualization, and interpretation of data is critical in solving complex global issues. By using data (e.g., predictor and response variables), the global community can better understand which countries are suffering the most and why, thus allowing for amped up efforts in certain sectors, necessary distribution of international resources, etc. It is interesting how our society is able to maintain such consistent tracking of COVID19 data, with all 195 countries displaying some level of statistical reporting. After conducting my analysis, I found that every category of variables I defined (i.e., age, health, quality of living/infrastructure) contained atleast one variable that was statistically significant in predicting total_deaths_per_million, though the strength of correlation coefficients were not as high as I was expecting, though this is valid given the multifaceted nature of the virus, with so many variables (some yet to be documented) that affect the outcome of total_deaths_per_million.

Main Questions: 1. Are age variables (i.e., aged_70_older and median_age) statistically significant predictors of total_deaths_per_million? (using May 08 2021 data for all countries) || 2. Are health variables (i.e., cardiovasc_death_rate, diabetes_prevalence, female_smokers, male_smokers) statistically significant predictors of total_deaths_per_million? (using May 08 2021 data for all countries) || 3. Are quality of life/infrastructure variables (gdp_per_capita + handwashing_facilities + hospital_beds_per_thousand + human_development_index) statistically significant predictors of total_deaths_per_million? (using May 08 2021 data for all countries)

Methods: To conduct this multiple linear regression analysis I used a COVID19 data set that tracked multiple variables for every country for every day of the virus's activity, boiled it down to the variables mentioned above, filtered down to rows from May 08 2021, and eliminated rows which contained empty cells to describe any of the necessary variables (individually for age, health, and quality of life/infrastructure categories). I used R with the 'lm' capability and plotting capability.

Data Source and Decription:
Link to Data: https://ourworldindata.org/coronavirus-source-data

Our World in Data provides a comprehensive Coronovirus Source Data sheet, which is updated daily and includes data on confirmed cases, deaths, testing, etc. Each of the 195 countries have entries for each day. A major shortcoming of the data is that a good amount of information (i.e., specific variables only have few countries reporting numbers on specific days). Benefit-wise, however, the data is rich with critical information, can be easily formatted for an analysis, and offers a substantial sample.

Variables I am using: median_age. aged_70_older, cardiovasc_death_rate, diabetes_prevalence, female_smokers, male_smokers, testing gdp_per_capita, handwashing_facilities, hospital_beds_per_thousand, human_development_index

Link to Video Presentation: https://youtu.be/6dEPrgfPgvg

Summary of Findings: Each of the three sets of analyses indicated that some societal variable. has an influence in affecting the total_deaths_per_million from COVID19. For the 'age' analysis, testing median_age and aged_70_older, the results are as follows:
- Presence of collinearity between independant variables
- Regression equation: y(total death per million) = -289.75 + 78.11 (aged_70_older) + 13.82 (median_age)
- aged_70_older is statistically significant as a predictor, w/ a p-value < 0.000647
- An adjusted R^2 of 0.3972 makes the regression model moderately reliable
- A p-value of < 2.2e-16 for the F-statistic means some variable(s) in the model is influential in predicting outcome

For the 'health' analysis, testing cardiovasc_death_rate, diabetes_prevalence, female_smokers, and male_smokers, the results are as follows:
- female_smokers is the most impactful coefficient in the equation, and its p-value < 2e-16 → statistically significant in predicting outcome
- y(total_deaths_per_million) = 209.43761 + -0.06476 (cardiovasc_death_rate) + 3.92145 (diabetes_prevalence) + 49.99089 (female_smokers) + -3.32595 (male_smokers)
- Adjusted R^2 of 0.4822 → fairly reliable
- P-value for F-Statistic: < 2.2e-16 → indicating atleast one variable is influential in predicting outcome

For the 'quality of living and infrastructure' analysis, testing gdp_per_capita + handwashing_facilities + hospital_beds_per_thousand + human_development_index, the results are as follows:
- gdp_per_capita and human_development_index are statistically significant in predicting the outcome, with significance codes of 0.1 and 0 respectively.
- P-value for F-Statistic: < 9.145e-06 → indicating atleast one variable is influential in predicting outcome
- Adjusted R^2 of 0.2977 → not definitively reliable

Upon using an interaction model for 'quality of living and infrastructure', the results showed that:
- Interaction exist between human_development_index and gdp_per_capita, as well as handwashing_facilities and hospital_bed_per_thousand as predictors (along with some tri-variable significant p-values)
- The regression model is fairly reliable with an adjusted R^2 of 0.4561, and an F-statistic p-value of 3.041e-06

These analyses address the question on whether variables of 'age', 'health', and 'quality of living and infrastructure' are statistically significant in predicting total_deaths_per_million of all countries. The results are overall fairly reliableUsing statistical modelling is essential in understanding what parts of the globe are being hit the worse due to COVID19, and why - data-drive understanding will help in the government and international community's response to the situation

Resources: Our World in Data - Coronavirus Pandemic (COVID-19) by Oxford Martin School
https://ourworldindata.org/coronavirus