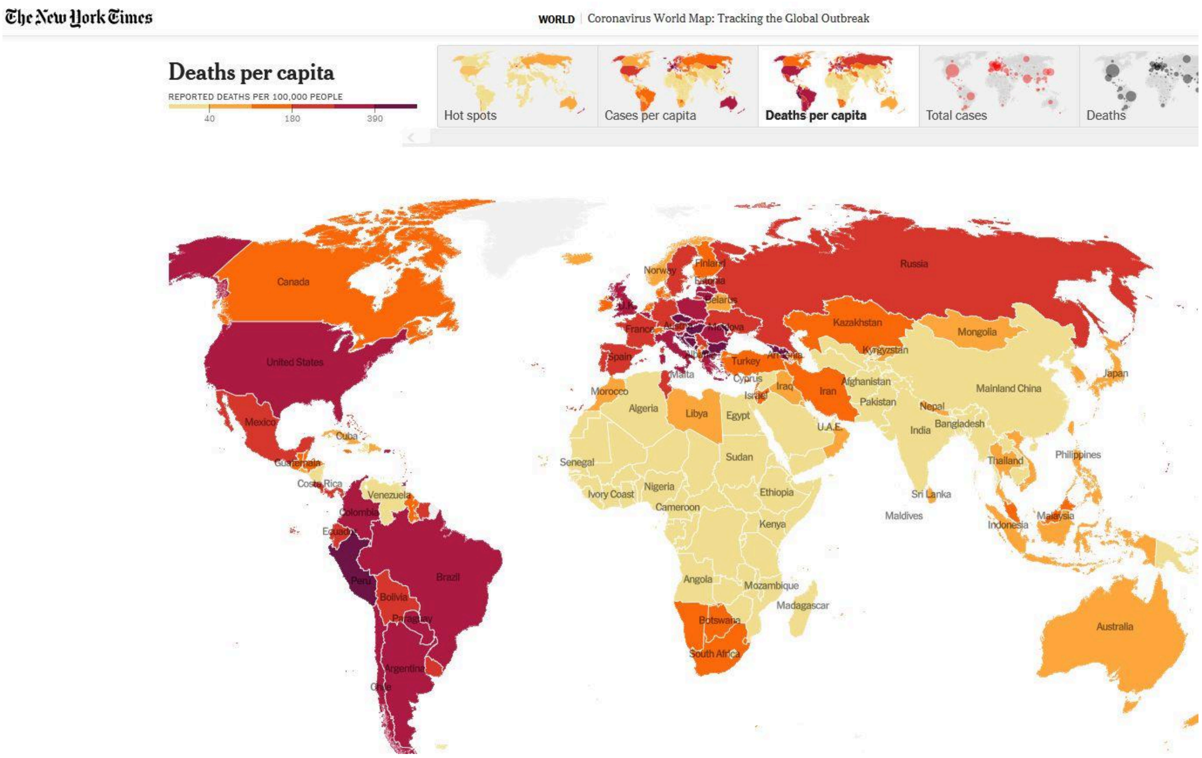
# Project Requirements




Though the public health emergency from the COVID-19 pandemic has ended, one question has not been adequately answered - why did countries have widely different rates of death from COVID? Only a small proportion of COVID cases lead to death, typically after a few weeks. However, the daily number of deaths in a country does not depend only on the number of cases. It also varies with several other factors including the extent of vaccination, the level of development (e.g., countries with more hospitals should see fewer deaths), age demographics (countries with a larger proportion of older people should see more deaths), pre-existing differences in medical conditions such as diabetes. The goal of this project is to use linear modeling to **predict two weeks ahead** the number of daily COVID deaths in different countries using a range of factors.

## Datasets

1. Data on COVID-19 from *Our World in Data*
   Their complete dataset contains a lot of information including the number of deaths, cases, vaccinations, hospitalizations, and several other country-specific pieces of information relevant to understanding the effects of COVID. Note that you can read the "raw" CSV file from a URL directly, like so:
   ```
   read_csv("https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv")
   ```
2. Population estimates from The World Bank's DataBank

Use the above web page to

a. Select all countries
b. Select a few Series (variables) that you think will be relevant to predicting death from COVID. For example, populations in certain age groups, mortality rates, and expected lifetime. At the very least, select *Population ages 80 and above, female* and *Population ages 80 and above, male*.
c. Select time: Only 2023
d. Download your selection as a CSV file (you get a .zip file, which contains the .csv file; delete the last few lines of the csv file which has the license information)

## Approach

There are three steps in this project:

1) **Data wrangling** to get all the data into **one table** that can be used for linear modeling
   a) read the two data files using `read_csv()`
   b) Keep only country-level data by removing all rows where the country_code is not exactly 3 letters (these represent larger regions like continents). Hint: `nchar(string)` returns the number of characters.
   c) Remove countries whose total population is less than 1 million.
   d) Add a new column `new_deaths_smoothed_2wk` that has the same values as `new_deaths_smoothed` *but two weeks ahead* (will be used for linear modeling as described later). R has a `Date` type that enables calculations with dates like `mutate(date= date - 14)` and `filter(date >= as.Date("2023-01-01"))`.
   e) tidy tables, as needed. (Hint: only the population data is not tidy.)
   f) Merge the tables (Hint: `join` using the 3-letter ISO code)

At the end of these steps, the data should be in one table, ready for linear regression (only a small sample of the data is shown below):

dependent variable

Predictor variables (many more)

| iso_code | location | date | new_deaths_smoothed_2wk | new_cases | new_cases_smoothed | total_vaccinations | SP.DYN.LE00.IN | SP.URB.TOTL | SP.POP.TOTL | SP.POP.80UP.FE |
|---|---|---|---|---|---|---|---|---|---|---|
| AFG | Afghanistan | 2022-12-18 | 0.571 | 72 | 63.000 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-19 | 0.429 | 51 | 62.000 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-20 | 0.429 | 73 | 58.286 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-21 | 0.571 | 39 | 58.000 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-22 | 0.571 | 65 | 58.857 | 12448870 | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-23 | 0.429 | 25 | 51.429 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-24 | 0.143 | 23 | 49.714 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-25 | 0.286 | 60 | 48.000 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-26 | 0.429 | 90 | 53.571 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-27 | 0.714 | 26 | 46.857 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-28 | 0.571 | 32 | 45.857 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-29 | 0.571 | 23 | 39.857 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-30 | 0.714 | 18 | 38.857 | NA | 63.37700 | 8535606 | 34413603 | 48319 |
| AFG | Afghanistan | 2022-12-31 | 0.714 | 43 | 41.714 | NA | 78.69024 | 261953748 | 320742673 | 7400961 |
| USA | United States | 2022-01-01 | 1977.714 | 584647 | 354503.286 | 521579175 | 78.69024 | 261953748 | 320742673 | 7400961 |
| USA | United States | 2022-01-02 | 2054.286 | 471965 | 387434.000 | 522079475 | 78.69024 | 261953748 | 320742673 | 7400961 |
| USA | United States | 2022-01-03 | 2127.286 | 302967 | 414167.286 | 523331209 | 78.69024 | 261953748 | 320742673 | 7400961 |
| USA | United States | 2022-01-04 | 2152.143 | 390858 | 439820.714 | 524757059 | 78.69024 | 261953748 | 320742673 | 7400961 |
| USA | United States | 2022-01-05 | 2079.429 | 902391 | 502377.286 | 526221050 | 78.69024 | 261953748 | 320742673 | 7400961 |

2) **Linear modeling**
   The goal is to predict `new_deaths_smoothed` *two weeks in the future*. Hint: this is the dependent variable.
   a) Make a list of all predictor variables that are available. The challenge is to identify which combination of these predictors will give the best predictive model.
   b) Generate some (at least 3) transformed variables. E.g., these could combine variables (e.g., `cardiovasc_deaths= cardiovasc_death_rate*population`).
   c) Split your dataset into train and test subsets: only data from 2022 should be used for building/training the linear models in `lm()`. (Data from 2023 will be used for evaluation as described later). Note: ***each day*** is one data point.
   d) Run linear regression with **at least 5 different combinations of predictor variables.** Hint: each model will look like:
   `new_deaths_smoothed_2wk~new_cases_smoothed+gdp_per_capita+diabetes_prevalence+icu_patients+SP.URB.TOTL`

3) **Evaluating the linear models**
   You should evaluate each of your linear models by predicting the number of daily deaths in each day in January-June 2023 (the test data) and comparing it with the actual number of deaths on those days. Specifically:
   a) For each of your models, calculate the Root Mean Squared Error (RMSE) over all days in January-June 2023 and all countries. Hint: use `rmse()` in library(modelr).
   b) For only your best model, calculate the Root Mean Squared Error for **every country**. Hint: use `group_by()` and `summarise(rmse(model= my_best_model, data=cur_data()))`. `cur_data()` gives the data in each group.

## Group work

You may work in groups of 1-3. Include all group member names in the PDF reports.

## Submission in two stages[1]:

### Stage 1 (Group formation and data wrangling) Due: Friday, April 19.

To submit:
1. A draft report that describes the (partially completed) data wrangling steps [PDF]
2. A listing of your R code in one file [.R file]

The draft report should include the names of everyone who will work on the project. This same group of students will continue to work together and submit the final project. You can continue to work on the data wrangling steps also until the final due date.

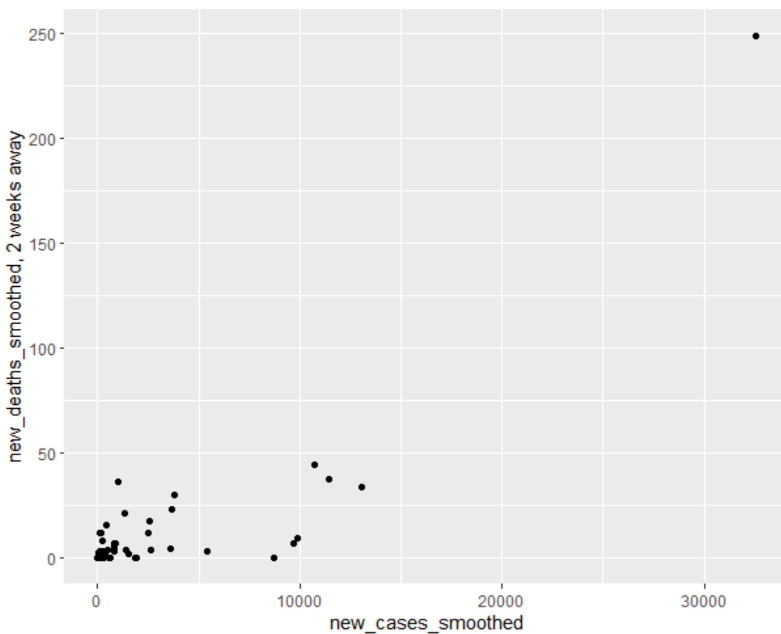
---

[1] This two-stage submission is to encourage you to work on the project early. Most of the grade will be based on the final submission, but you should have formed groups and completed downloading the data and started the data wrangling in two weeks.

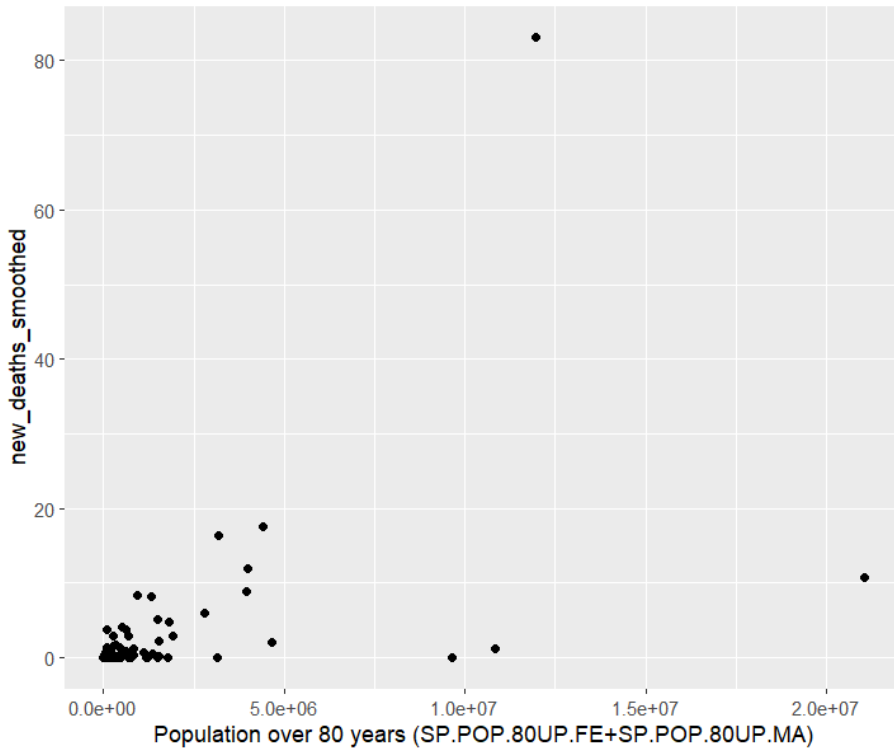
## Stage 2 (Final submission) Due: ~~Friday, May 3~~ Sunday, May 5

To submit:

1. A short report describing your work. Specifically, your report should include:
   - brief description of only the important data wrangling steps,
   - List of the variables ("series") that you selected from the Population estimates webpage,
   - a scatterplot of only the most recent new deaths per day two weeks ahead (new_deaths_smoothed_2wk) in the test dataset (i.e., 2023-06-30) and the corresponding new cases per day (new_cases_smoothed) for every country (i.e., one point per country), like so:




   - a scatterplot of only the most recent new deaths (new_deaths_smoothed) in the test dataset (i.e., 2023-06-30) and the total (female+male) population over 80 for every country (i.e., one point per country), like so:




- descriptions of variable transforms,
- list of the different combinations of predictor variables in your models,
- brief reasons for *why* you chose these predictor variables (e.g., your prior knowledge, or a plot showed a trend),
- a table listing the R2 and RMSE of **all** your models
- a table showing the RMSE of only your best model for the 20 most populous countries arranged in decreasing order of population, like so:
- a conclusion that describes in words the implication of your most accurate model.

2. A listing of your R code in one file [.R file]

## Project checklist/grading rubric

1. Draft submission (approximately 10% of total grade)
    a. Data wrangling is at least partially complete
    b. Brief report of completed steps
    c. Group member names are included in the report
    d. R code for completed data wrangling
    e. Submission on time
2. Data wrangling (final)
    a. Code to load and wrangle OWID data
    b. Code to load and wrangle demographics data
    c. Code to join datasets to one table
3. Modeling:

    a. Tried at least 5 different combinations of variables for modeling
    b. Included at least 3 variable transformations
    c. Code that correctly implements the above
4. Evaluation:
    a. Generate the R2 and RMSE of all models
    b. Identified the best model and calculated its RMSE for all countries
    c. Code that correctly implements the above
    d. Note: having a high R2/low RMSE is *not* important for grading
5. Written report (final)
    a. Brief descriptions of the data wrangling steps
    b. Brief description of how variables were chosen for data modeling
    c. Descriptions of variable transformations
    d. Scatterplot of only the most recently available new_deaths_smoothed_2wk and new_cases_smoothed for every country
    e. Scatterplot of only the most recent new deaths per day and the urban population
    f. A table that shows the R2 and RMSE of the different models
    g. A table that shows the RMSE of the best model for 20 most populous countries
    h. A conclusion – what does your modeling say about death rates (e.g., what are the significant factors and what are not)
    i. Clarity of the report (e.g., appropriate section headings)
6. Code
    a. Readability: use of `tidyverse,` no unnecessary use of complex functions.
    b. Code has adequate comments
    c. Note: include only the final code, i.e., do not submit just the RStudio history