Project 1 Requirements

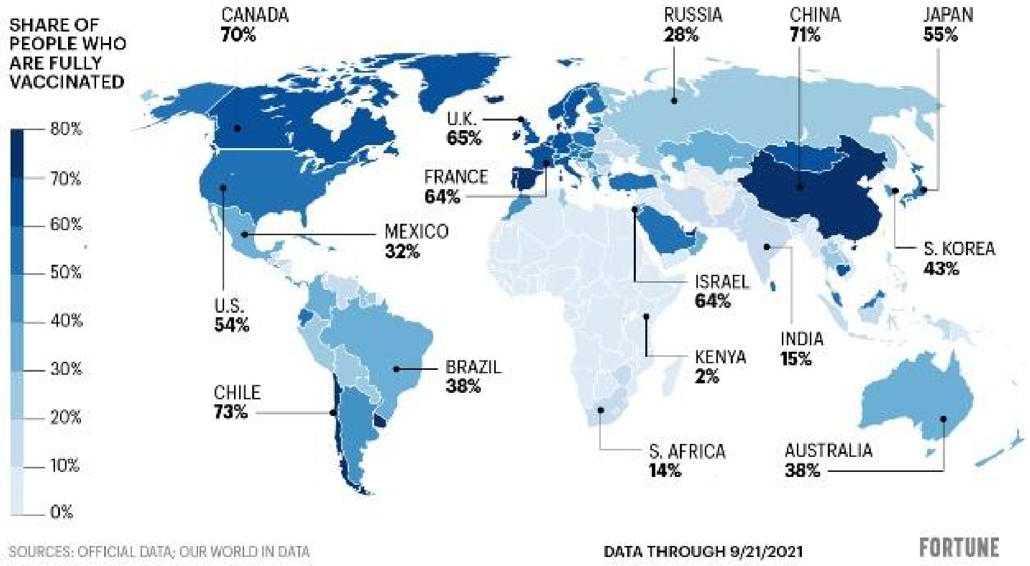


Image from: https://fortune.com/2021/09/22/covid-vaccine-rate-world-us-latest-update-coronavirus-vaccines/

COVID vaccination rates vary a great deal between countries. There are several reasons, including access to healthcare and demographics (countries prioritize recipients by age). The goal of this project is to use linear regression to model the vaccination rates in different countries in terms of their number of hospital beds per capita (a measure of access to healthcare) and demographics.

Datasets

1. time_series_covid19_vaccine_doses_admin_global.csv[1]
   - This file contains the number of vaccine doses given every day in different countries. Note that you can read the "raw" CSV file from a URL directly, like so:
     ```
     read_csv("https://raw.githubusercontent.com/govex/COVID-19/master/data_tables/vaccine_data/global_data/time_series_covid19_vaccine_doses_admin_global.csv")
     ```
2. Hospital beds data
   - Download the Hospital bed density Data by country from the World Health Organization (WHO):
     - http://apps.who.int/gho/data/view.main.HS07v
     - Click the "*Download filtered data as: CSV table*" link near the top-right.
3. demographics.csv[2]

[1] More information: International vaccine data
[2] Original dataset: https://databank.worldbank.org/source/population-estimates-and-projections/Type/TABLE/preview/on#

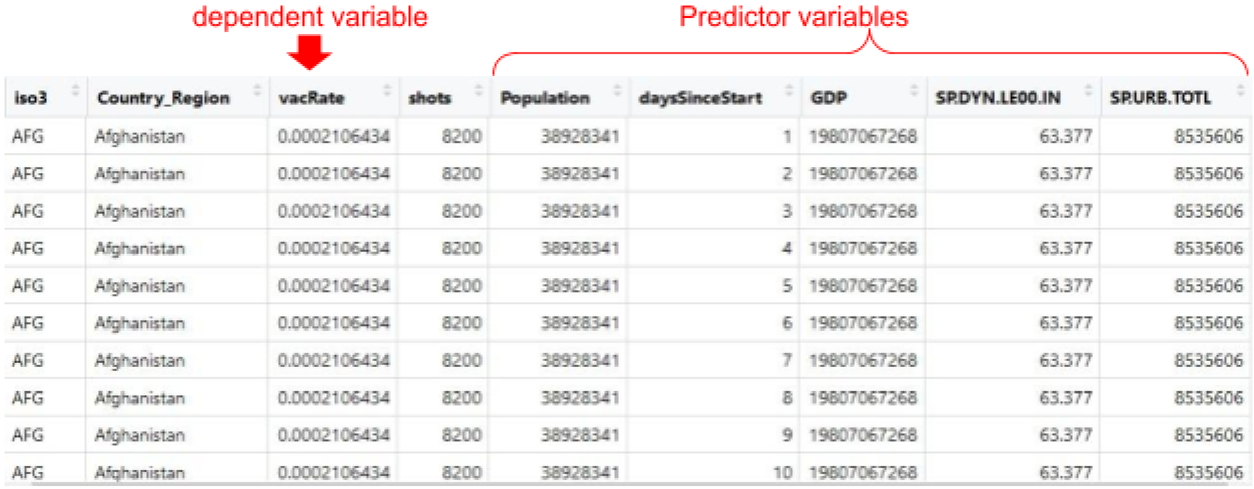
- ○ This gives the proportion of a country's population in different age groups and some other demographic data such as mortality rates and expected lifetime. [File in Datasets module on Canvas. Same dataset used in Homework #5.]

Hints

There are two major steps in this project:

1) **Data preparation/wrangling** to get all the data into **one table** that can be used for linear modeling
    a) reading the data files using read_csv()
    b) Removing unneeded rows (e.g., countries like Brazil and India report Province_State-level data that is not needed as we are studying only country-level rates) and columns.
    c) tidying tables, as needed. For example, the vaccinations data is not tidy.
    d) Calculate the vaccination rate: vaccinations/population
    e) Since the most important factor affecting vaccination rate is the number of days since vaccination began (vaccination rate always increases), calculate a variable that is: number of days since first non-zero vaccination number. This variable will be important for modeling.
    f) Discard data that is not needed. For example, only the number of hospital beds from the most recent year is necessary.
    g) You can ignore sex-related differences in demographics in this project, so add the male/female population numbers together (already done in HW #5).
    h) Merge all tables (Hint: Join using the country name)

At the end of these steps, the data should be in one table, in a format ready for linear regression:



iso3	Country_Region	vacRate	shots	Population	daysSinceStart	GDP	SP.DYN.LE00.IN	SP.URB.TOTL
AFG	Afghanistan	0.0002106434	8200	38928341	1	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	2	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	3	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	4	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	5	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	6	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	7	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	8	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	9	19807067268	63.377	8535606
AFG	Afghanistan	0.0002106434	8200	38928341	10	19807067268	63.377	8535606

2) **Linear modeling the Covid vaccination rate**
    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. You should also try

transforming some of the variables (e.g., transforming population counts to proportion of total population). Run linear regression with **at least 5 different combinations of predictor variables.**
Note: ***each day*** becomes one data point, i.e., the vaccination rate is calculated for each day for each country. The number of vaccinations should *not* be used as an independent variable as this is essentially what you are predicting.

Country names:
The country names across datasets do not all match (e.g., "Iran" and "Iran, Islamic Rep."). Such mismatches are a common problem in data science and this should be fixed before joining tables. However, since the larger project goal is to learn a model of vaccination rates in every country/every day, missing a few countries will not greatly reduce the data to build the model. So it is sufficient to fix only these countries' names which have been severely affected by the pandemic as we don't want to exclude their data, for example:

1. "Iran (Islamic Republic of)", "Iran", "Iran, Islamic Rep."
2. "South Korea", "Republic of Korea", "Korea, Rep."
3. "United Kingdom", "United Kingdom of Great Britain and Northern Ireland"

The names can be fixed by code like this:

```
mydata <- mydata %>% mutate(Country = replace(Country, Country == "Republic of Korea", "South Korea"))
```

Group work

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

Submission in two stages:

Stage 1 (Group formation and data wrangling):

Due: Friday, April 8. Please 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]

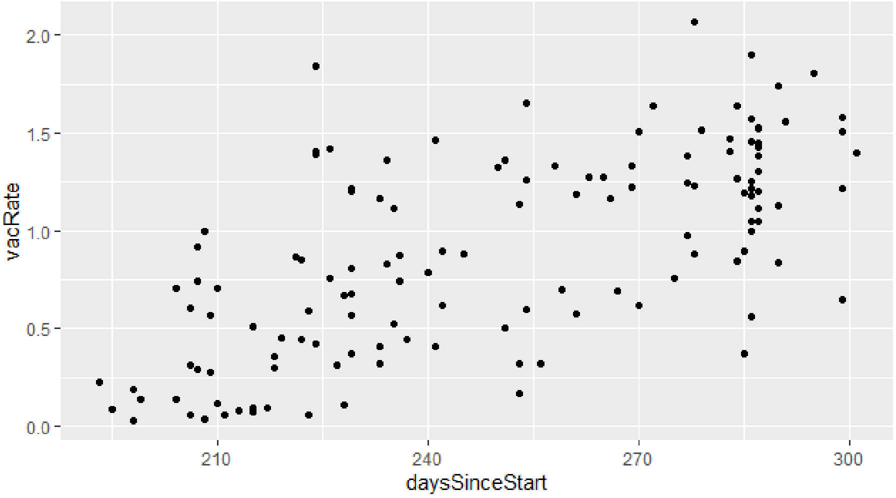
The draft report should include the names of everyone who will work on the project. This same group of students will continue to work together and submit the final project. You can continue to work on the data wrangling steps also until the final due date.

Stage 2 (Final submission):

Due: Friday, April 22. Please submit:

1. Write a short report describing your data wrangling steps and the different combinations of predictor variables you tried, and any variable transforms. [A PDF file]

- The report should include the following plots:
  i. a scatterplot of only the most recent vaccination rate for every country and the number of days since first vaccination, like:



  ii. a summary bar graph with the R2 values on the y-axis and a corresponding model name on the x-axis (include all the different models you tried).
- There should be 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 vaccinations data
   b. Code to load and wrangle hospital beds data
   c. Code to load and wrangle demographics data
   d. Code to join datasets to one table
3. Modeling:
   a. Tried different combinations of variables for modeling; tried some variable transformations
   b. Code that correctly implements the model
4. Written report (final)
   a. Brief descriptions of the data wrangling steps

  b. Brief description of how variables were chosen for data modeling
  c. Description of any variable transformations
  d. A scatterplot of most recent vaccination rates for different countries
  e. A plot that shows the R2 values of the different models
  f. A conclusion – what does your modeling say about vaccination rates (e.g., what are the significant factors and what are not)
  g. Clarity of the report (e.g., appropriate section headings)

5. Code
  a. Style: readability, use of `tidyverse`, 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