International COVID-19 Vaccination Rates and Country Characteristics

Shu Ding, Marlena Jacobs, Mary Keonoupheth, Arnaud Laprais

Math 167PS, Spring 2021

Introduction

In December 2019, Coronavirus disease 2019 (COVID-19) was first identified in Wuhan, China. In the twelve months that followed, astonishing scientific innovation led to the development and approval of multiple safe and highly effective vaccines. Global vaccination will play a key role in ending this devastating global pandemic. COVID-19 vaccination rates vary widely between countries. We conducted a data wrangling project to explore this. We are interested in how vaccination rates correlate with a country's health care spending, age demographics, COVID-19 mortality rates, COVID-19 case rates, COVID-19 case fatality rates, and democracy score.

Data Sources

We used data from two sources. The World Health Organization (WHO) is the public health agency of the United Nations. WHO, in collaboration with member states, maintains an annually updated database on global health expenditure, using sources such as health accounts studies and government expenditure records. Our World in Data (OWID) is a collaborative project between researchers at the University of Oxford and the Global Change Data Lab, a non-profit based in the United Kingdom that conducts research centering on the environment and global living conditions. OWID generally sources their data from specialized institutes like Peace Research Institute Oslo, research articles, and international institutions/statistics agencies like the World Bank and the UN.

Creating the COVID-19 DataFrame

We extracted datasets from our sources, and used Python to clean, organize, label, and bring together the data. The following describes our process.

COVID-19 Vaccination, Cases, and Mortality Data

Our source for COVID-19 data was OWID. The OWID team collects data from official reports on COVID-19 vaccinations, cases, and deaths and has had a daily update of the data since December of 2020. We extracted the data from the OWID website as a comma-separated values (csv) file on April 30, 2021. Because of missing vaccination data for some countries for the most recent dates, we chose April 24, 2021 as our date of interest. The set contains data on the share of the total population for each country that has received all doses of a vaccine against COVID-19. The vaccination-related variables of interest are: continent, location, date, total vaccinations, people vaccinated, people fully vaccinated, total vaccination doses per hundred, people vaccinated per hundred, and people fully vaccinated per hundred. We also created a column for vaccination rates by calculating the proportion of people vaccinated to the country's total population as a way to reasonably compare vaccine distribution among the countries.

Other COVID-19 variables of interest were: cumulative COVID-19 cases per capita, new COVID-19 cases per capita, total cumulative COVID-19 deaths per capita, total cumulative cases per million, and total cumulative deaths per million. Also, a new column of COVID-19 Case Fatality Rate (CFR) was created. The CFR is a straightforward measurement showing the mortality risk of COVID-19. We calculated it by finding the ratio between the confirmed total deaths per capita and the confirmed total cases per capita, then multiplying by 100.

We used the Pandas Python module to read in the csv file and create a dataframe. The raw dataset for COVID-19 vaccination, cases, and mortality rates included some rows of data that are not considered countries, so we removed these particular rows. For example, ‘Africa’ and ‘World’ were intermixed with the variable ‘countries.’ To remove these inconsistent rows of data we created a list of indices that included the string ‘OWID’ in its iso code, as these indicated non-country rows of data. We subsetted the data set by using these indices to exclude non-country rows of data along with the most recent data for each country. We then extracted the variables that we were interested in analyzing and combined these variables into a new dataframe.

Age Demographic Data

Our dataset for age demographics comes from OWID as a csv file, and contains information on raw population counts in several age brackets from 1950-2020 by Country. Specifically, the data consists of 6 columns, one denoting a ‘Country’ and 5 more giving raw population values for distinct age brackets.

We read the data into Python to create a Pandas dataframe. We first subsetted the data by selecting only the most recent year of 2020 - thus eliminating all rows pertaining to the 1950-2019 timeframe. Then, by inspecting the newly subsetted data we discovered the existence of several non-country observations such as “World”, or “Central Asia”. Since we are interested in a country-by-country comparison we wanted to remove these troublesome imposter countries. With this in mind we constructed a list of countries from Google, and used this as a “master list” with which to cross-reference and clean the data. This list needed revision, but since the number of countries was relatively small, the revisions were easily accomplished through trial and error. For example, some legitimate countries slipped through the first filtering stage as a result of errors in the master list (ie: “VieT Nam” instead of “Vietnam”). Then we cleaned the data with the following process: any “country” that did not appear in the master list was singled out, and in this process we quickly and easily located the imposter countries and removed the related rows from our dataframe.Once this task was completed, we renamed the verbose and illegible columns with more appropriate tags - marking the last step in the cleaning of the age demographic data.

Although the age demographic data was technically ready for merge, since we are interested in country comparisons we needed to convert the raw population values in the rows and columns to a measure that would more easily facilitate comparison. With this goal in mind we created a new column consisting of the total population for each country, and performed element-wise division for the appropriate columns in each row to obtain population as a percentage of total population. From this process we added 5 new columns to the dataframe, one for each age bracket column.

Finally, we created a new column ”%Pop < age 25” by summing the ”<5”, ”5-14”, and ”15-24” columns.

The final age demographic data consisted of 13 columns: country, raw population for specific age brackets, population as a percentage of total population for specific age brackets, and lastly, total population for each country.

Health Expenditure Data

We extracted our health expenditure data from the WHO Global Health Expenditure Database website. Our indicators of interest were: immunization expenditure, preventive care expenditure, and total health expenditure. For each of these, we were interested in total health expenditure across all funding sources (domestic government, domestic private, and external funding). The Data Explorer tool on WHO's website allows extraction of data for one unit of expenditure at a time. Since we were interested in three units of expenditure (as a percent of GDP, in US dollars per capita, and, for sub-categories, as a percent of total health expenditure), we extracted three separate data files. Each extracted Excel file was organized into multiple rows for each country, with one row for each country's data on a particular indicator. The columns in the data files were the years, for as recent as 2018.

We used the pandas module in Python to create pandas dataframes from the extracted data files. We treated the first row of the extracted data as a header, and omitted the second empty row from our data. After deciding that older data would be less relevant to 2021, we selected only the 2018 column, the most current year available. We next reshaped each data set: reformatting the data so that each country had just one row, with the indicators as columns. We renamed the columns to specify the units of expenditure. For example, for the expenditure as percent of GDP dataframe, the column named ”Immunization Programmes” was renamed ”Immunization expenditure (% of GDP)”. We noted that some values were entered as zeros when they likely were in fact missing, so we updated our data to treat these values as missing. Next, we merged the three health expenditure dataframes into one. The merged dataframe had one row per country, with the indicators in their different units of expenditure as the columns.

To merge the health expenditure dataframe with data from OWID, the country names needed to match. For twelve countries in the WHO health expenditure data, the country name used by WHO differed from the name used by OWID. For example, WHO used the name ”United States of America,” while OWID used the name ”United States.” We used Google as needed to confirm country name matches. We created a new variable in the health expenditure dataframe, ‘location’, with country names that align with OWID data.

Political Data

Our dataset for political data comes from OWID in the form of a csv file. It contains the columns including location, year and political regime. The “political regime” column shows the ”democracy score” which captures the regime authority spectrum on a 21-point scale ranging from 10 (full autocracy) to 10 (full democracy). The dataset records the democracy scores from the year of 1816 to the year of 2015 for 166 countries. The variable we are interested in is the democracy score for the most recent recorded year (2015) for each country.

We used the pandas module in Python to create a pandas dataframe. We filtered out the rows of the dataframe for all years except 2015, which is the most recent recorded year in the dataset. Then we extracted the columns of location and political regime. The final political data contains 2 columns which are the location(country) and the democracy score from the year 2015.

Merging the Data

Once the COVID data, the health expenditure data, the age demographics data, and the political data were wrangled into shape, our next task was to merge the four dataframes into one. We merged on country name using the pandas merge method. Since vaccination rates are our primary interest for this project, our final dataframe only includes countries that had a non-missing value for our vaccination rate variable. The dataframe has a row for each of the 77 countries, with 40 columns for our included variables.

Summary Statistics

Vaccination Rates Summary Statistics

Vaccination rate is the proportion of people vaccinated in a country compared to its total population. The distribution of vaccination rates is right skewed. There are more countries with a vaccination rate below 25%

and fewer countries with a vaccination rate above 25%.

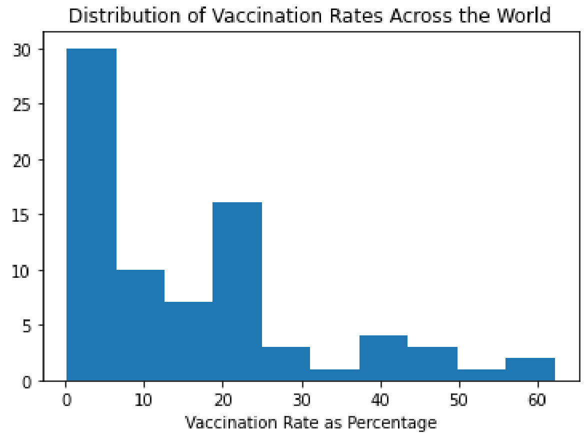
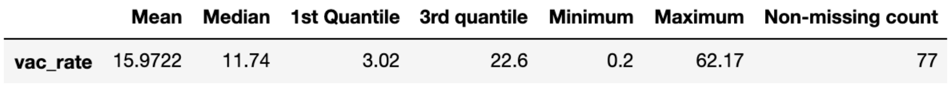


Figure 1: Distribution of Vaccination Rates

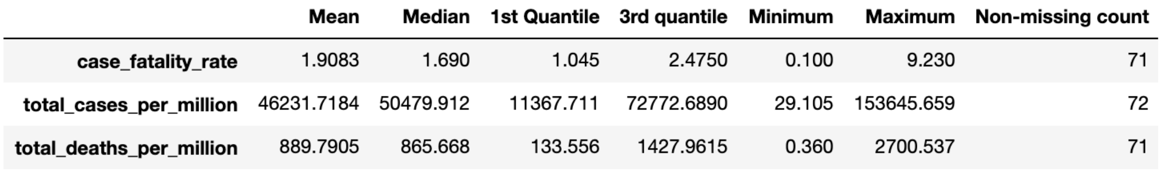
Given the COVID-19 vaccination rate summary statistics, we find that Bhutan has the highest recorded vaccination rate among the countries we analyzed with approximately 62.17% of its population vaccinated as of April 24, 2021. On the other hand, Vietnam has the lowest recorded vaccination rate with only 0.2% of its population vaccinated.



	Mean	Median	1st Quantile	3rd quantile	Minimum	Maximum	Non-missing count
vac_rate	15.9722	11.74	3.02	22.6	0.2	62.17	77

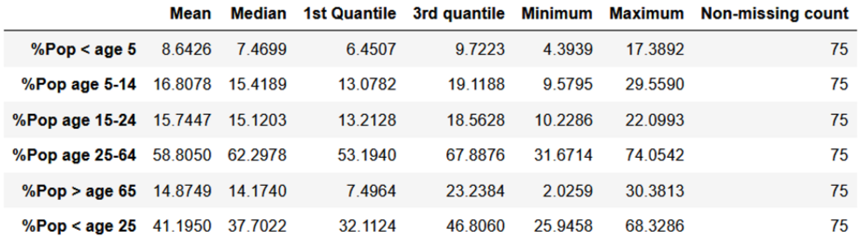
COVID-19 Cases, Mortality, and Case Fatality Rate Data Summary Statistics

The summary statistics for case fatality rate, total cases per million, and total deaths per million reveals that Vietnam has the lowest recorded cases and deaths per million with approximately 29 cases and 0.36 deaths per million, respectively. Montenegro has the highest recorded total cases per million with approximately 153645 cases per million. Czechia has the highest recorded total deaths per million with approximately 2700 deaths per million. Mexico has the highest recorded case fatality rate with a rate of 9.23% whereas Bhutan has the lowest recorded case fatality rate with a rate of 0.1%.



	Mean	Median	1st Quantile	3rd quantile	Minimum	Maximum	Non-missing count
case_fatality_rate	1.9083	1.690	1.045	2.4750	0.100	9.230	71
total_cases_per_million	46231.7184	50479.912	11367.711	72772.6890	29.105	153645.659	72
total_deaths_per_million	889.7905	865.668	133.556	1427.9615	0.360	2700.537	71

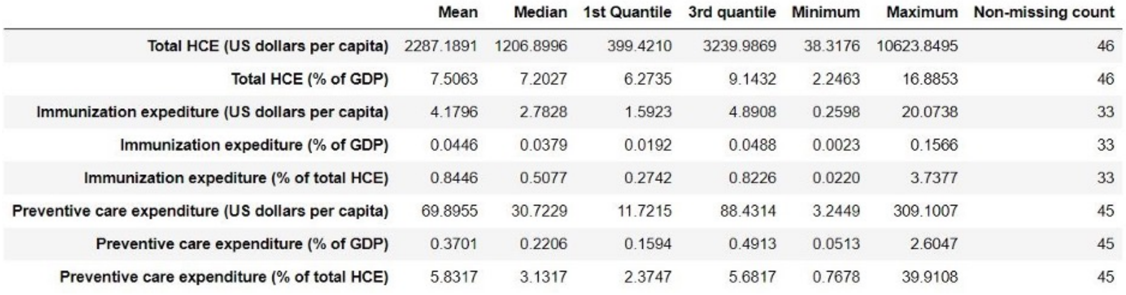
Age Demographics Data Summary Statistics



	Mean	Median	1st Quantile	3rd quantile	Minimum	Maximum	Non-missing count
%Pop < age 5	8.6426	7.4699	6.4507	9.7223	4.3939	17.3892	75
%Pop age 5-14	16.8078	15.4189	13.0782	19.1188	9.5795	29.5590	75
%Pop age 15-24	15.7447	15.1203	13.2128	18.5628	10.2286	22.0993	75
%Pop age 25-64	58.8050	62.2978	53.1940	67.8876	31.6714	74.0542	75
%Pop > age 65	14.8749	14.1740	7.4964	23.2384	2.0259	30.3813	75
%Pop < age 25	41.1950	37.7022	32.1124	46.8060	25.9458	68.3286	75

From the above table we note that on average the 25-64 age bracket is the largest age group, with the other brackets falling in far lower ranges. However, note that the final three rows each have similar ranges of around 30 percentage units, which suggests a non-trivial amount of variability in the data.

Health Expenditure Data Summary Statistics

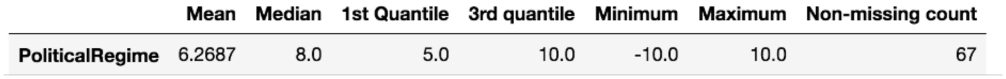


	Mean	Median	1st Quantile	3rd quantile	Minimum	Maximum	Non-missing count
Total HCE (US dollars per capita)	2287.1891	1206.8996	399.4210	3239.9869	38.3176	10623.8495	46
Total HCE (% of GDP)	7.5063	7.2027	6.2735	9.1432	2.2463	16.8853	46
Immunization expediture (US dollars per capita)	4.1796	2.7828	1.5923	4.8908	0.2598	20.0738	33
Immunization expediture (% of GDP)	0.0446	0.0379	0.0192	0.0488	0.0023	0.1566	33
Immunization expediture (% of total HCE)	0.8446	0.5077	0.2742	0.8226	0.0220	3.7377	33
Preventive care expenditure (US dollars per capita)	69.8955	30.7229	11.7215	88.4314	3.2449	309.1007	45
Preventive care expenditure (% of GDP)	0.3701	0.2206	0.1594	0.4913	0.0513	2.6047	45
Preventive care expenditure (% of total HCE)	5.8317	3.1317	2.3747	5.6817	0.7678	39.9108	45

It must be noted that health expenditure data is missing for many of the countries included in our COVID vaccination dataframe, limiting our exploration of the relationship between vaccination rates and health expenditure. The distributions of the health expenditure variables are all highly right-skewed, with several high-spending countries driving the mean higher than the median. The one exception to this is total healthcare expenditure as a percent of GDP, which has a more symmetric distribution.

Political Data Summary Statistics

The histogram on the left shows the distribution of "Democracy Score" of 77 recorded countries. The one on the right shows the distribution of "Democracy Score" of each county by 6 different continents. The "Democracy Score" goes from -10 (full autocracy) to 10 (full democracy).



	Mean	Median	1st Quantile	3rd quantile	Minimum	Maximum	Non-missing count
PoliticalRegime	6.2687	8.0	5.0	10.0	-10.0	10.0	67

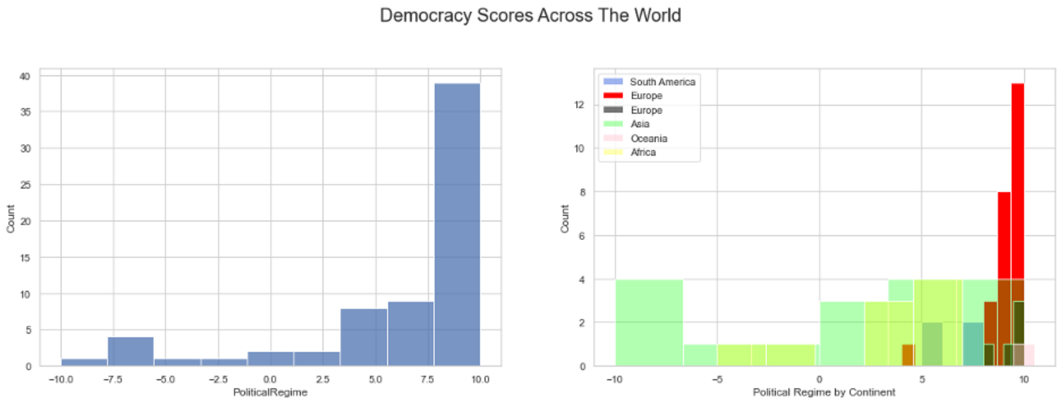


Figure 2: Distribution of Democracy Scores

The majority of the countries in our dataframe of the countries in our dataframe are democracies (around 65% countries scored 6 or higher), which includes almost all countries from North America, South America, Europe and Oceania. In contrast, 8 Asian and African countries have democracy scores below 0, which denote relatively authoritarian regimes.

Vaccination Rates Compared to Country Characteristics

Vaccination Rate Compared to Cases, Mortality, and CFR Data

We looked for the relationship between the vaccination rate and 3 different variables which are Total Cases Per Million, Total Deaths Per Million, and Case Fatality Rate.

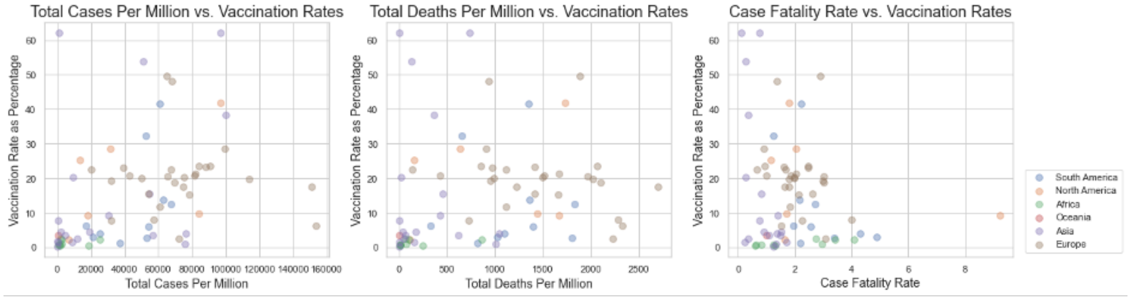


Figure 3: Vaccination Rate vs. Cases, Mortality, and CFR

The plot of "Vaccination Rates vs. Total Cases Per Million" revealed a moderate positive relationship between the 2 variables. Generally, countries with larger total case rates tend to have higher vaccination rates.

There is no obvious linear relationship between the 2 variables in the plot of "Vaccination Rates vs. Total Deaths Per Million". While it is shown in the graph that countries with higher vaccination rates (above 30) have a lower total death rate (below 2000 COVID-19 deaths per million).

We also wanted to see if there is a relationship between vaccination rates and case fatality rates. There appears to be a weak relationship between the two variables. There is a slight trend showing that countries with higher vaccination rates tend to have lower case fatality rates, but there are also many countries with lower vaccination rates and low case fatality rates. Countries in Asia have a wide range of vaccination rates, but all Asian countries tend to maintain low case fatality rates. Mexico (categorized as 'North America' in the graph) is shown in the far right portion of the graph and has the highest recorded case fatality rate at approximately 9%.

Vaccination Rate and Population Age Demographics

In this section we examine the relationship between vaccination rates among countries with certain population demographics. We are interested in exploring the behavior of countries with specific age demographics as well as whether or not there are differences in vaccination trends between larger and smaller countries.

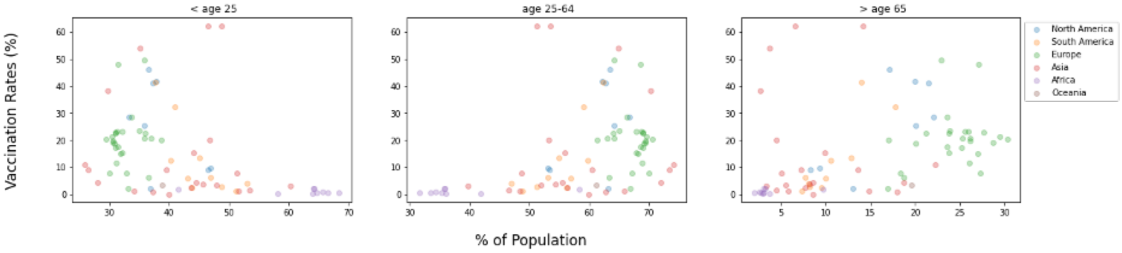


Figure 4: Vaccination Rate vs. Population Age Demographics

In the above plot we see a non-trivial relationship between the age demographics of a country and vaccination rates. From the first two (starting from the left) plots we see some support for the idea that countries with a younger population have lower vaccination rates. And the third plot ("> age 65") follows a similar pattern if Asian countries are ignored. In the following plot we see a possible explanation for the above trend.

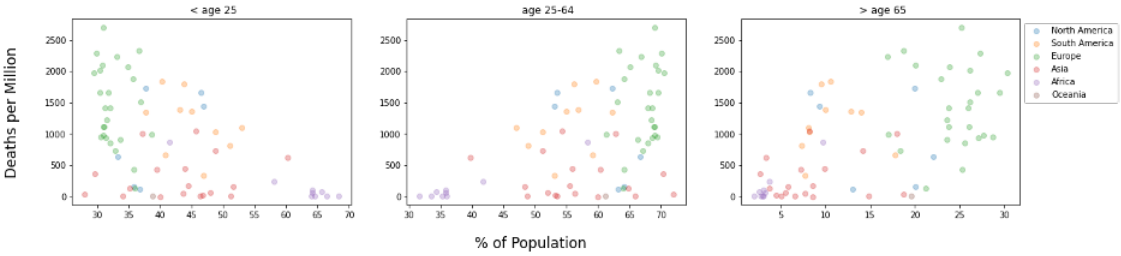


Figure 5: Deaths Per Million vs. Population Age Demographics

In the above plots, a younger population is associated with a lower death toll. Perhaps countries with younger populations are less concerned by COVID-19 because they see a far smaller death toll when compared to older countries, and as a result, those younger countries do not see vaccination as an immediate priority. On the other hand it is also possible that these younger countries are subject to factors (such as the vaccine not being approved for children) that strongly limit vaccine access.

In the following plot, we revisit Figure 3 and scale country data points by total population (larger circles indicate larger population). Here we see a far more distinct relationship between cases and vaccination rates in a country, where the deviations from the trend can be explained by population size.

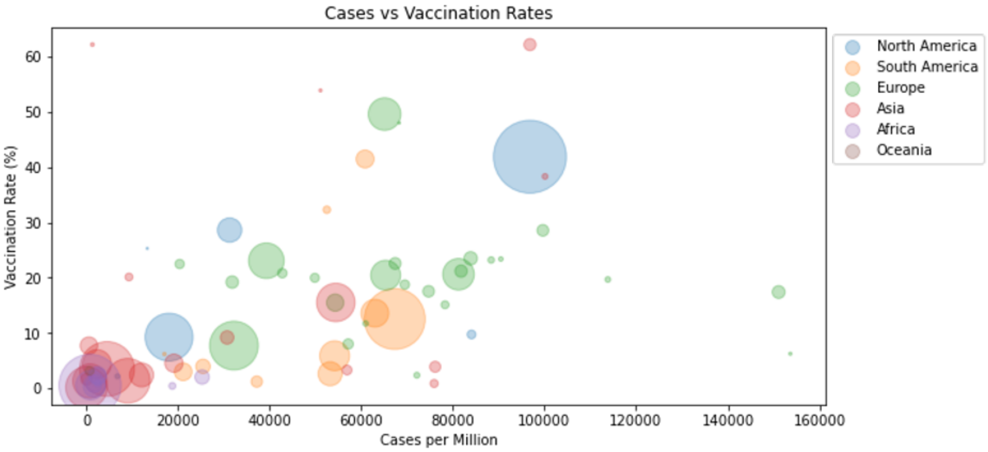


Figure 6: Cases vs. Vaccination Rate

Simply put: countries with larger populations tend to vaccinate somewhat proportionately to the amount of cases they've experienced. Many countries with smaller populations also follow this trend, and the countries that fall far outside of this trend are generally countries with smaller populations.

Vaccination Rate and Health Expenditure

We looked at health expenditure using three indicators: spending on immunization programs, on preventive care, and on all healthcare. Our data is from 2018, the most recent year available from WHO. Therefore the data will not reflect any significant increase or decrease in spending in the last few years, and will certainly not reflect influxes in spending related to COVID-19 treatment or vaccine administration. However, our data can serve as a measure of how countries have invested in healthcare prior to the start of the pandemic.

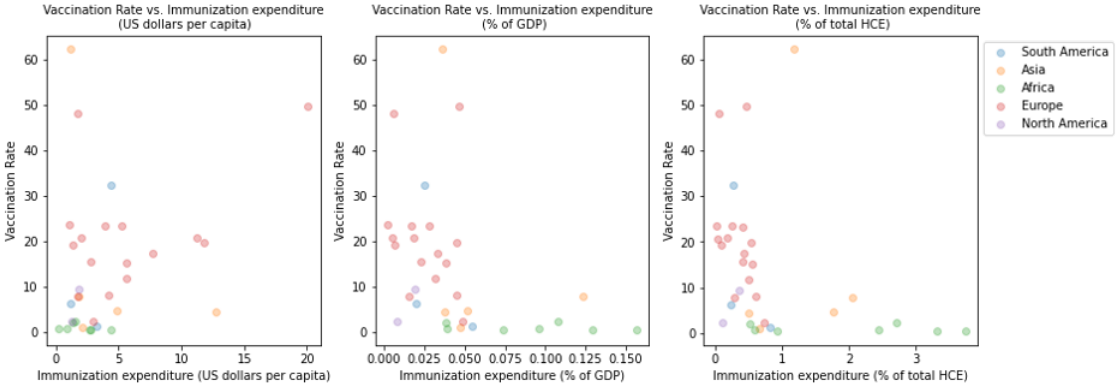


Figure 7: Vaccination Rate vs. Immunization Expenditure

With the immunization expenditure measures, it is notable that the countries with the highest COVID-19 vaccination rates (greater than 30%) tended to spend the least on immunization programs in 2018. This is surprising; we had anticipated that higher vaccination rates would be associated with higher health spending. It is also notable that the African countries, represented by green points in the scatter plots, had uniformly low vaccination rates, regardless of immunization spending in 2018. High vaccination rates may be negatively correlated with spending on immunization programs. It may be that immunization programs in 2018 were mainly aimed at immunizing children, so this investment would not result in preparation for COVID-19 vaccination of adults in 2021.

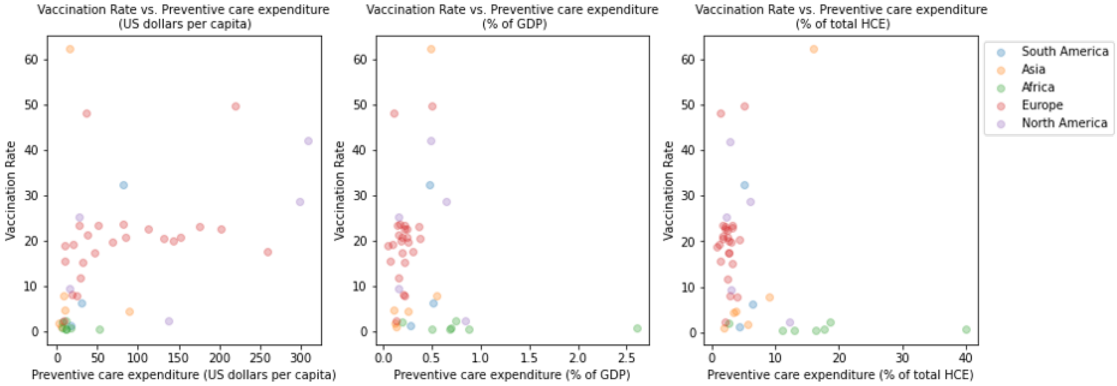


Figure 8: Vaccination Rate vs. Preventive Care Expenditure

We also looked at health expenditure in the preventive care category. With preventive care expenditure in US dollars per capita (see the plot on the left), we observed a trend more like what we were expecting: higher vaccination rates appear to be correlated with higher expenditure. We do not see this with preventive care spending as a percent of GDP, or as a percent of total health care expenditure. The actual amount spent on preventive care in 2018, as opposed to the relative amount spent, appears to be more correlated with COVID-19 vaccination rates.

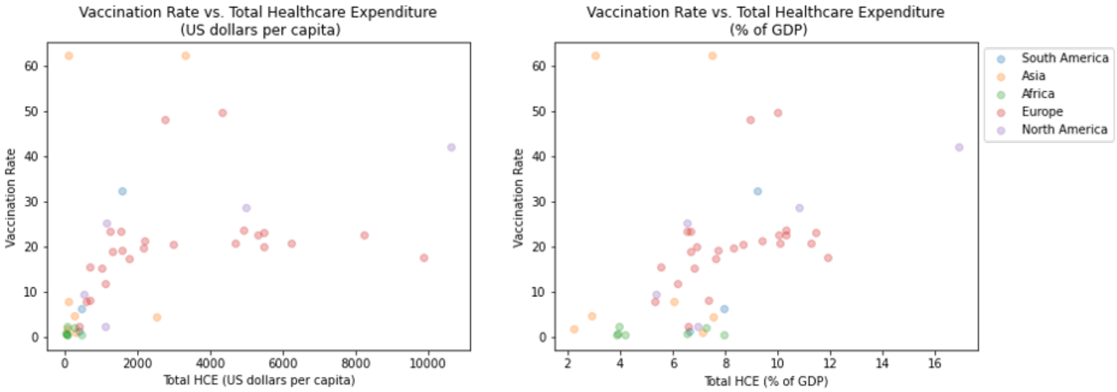


Figure 9: Vaccination Rate vs. Total Healthcare Expenditure

In our total healthcare expenditure data, we can see an association between higher COVID-19 vaccination rates and higher 2018 health spending. Interestingly, this trend does not seem to hold for the few countries with vaccination rates over 30%; there seems to be no discernible correlation for this handful of countries. Also of note is the trend in European countries, depicted with pink points in the scatterplots. These countries for the most part have vaccination rates around 20%, insensitive to the total health care expenditure. Overall, however, there appears to be a positive correlation between total health spending and COVID-19 vaccination rates. Higher health spending in 2018 may reflect a greater investment in and prioritization of healthcare, which might be predictive of higher COVID-19 vaccination rates in 2021.

Vaccination Rate and Democracy Score

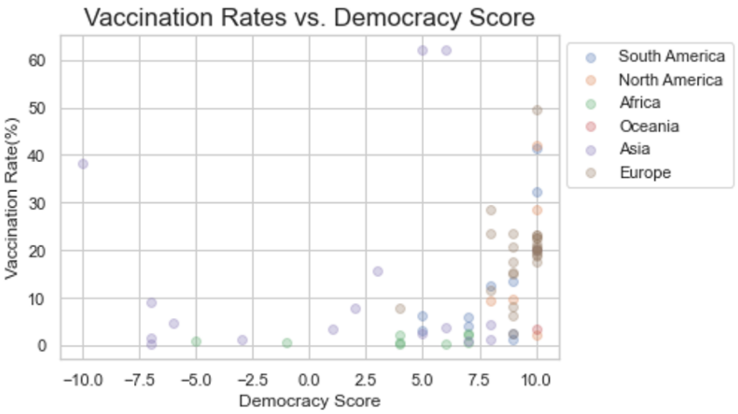


Figure 10: Vaccination Rate vs. Democracy Score

There is no obvious linear relationship between the 2 variables in the plot of ”Vaccination Rates vs. Democracy Score”. We can see from the plot that almost all the highly vaccinated countries are countries with high democracy scores.

Conclusion

We brought together data from multiple sources to answer our question: how are the characteristics of countries associated with their COVID-19 vaccination rates? We found that countries with larger total case rates tend to have higher vaccination rates, and countries with higher vaccination rates have an overall lower total death rate. Although the case fatality rate (CFR) is not clearly related to the vaccination rate, the countries with higher vaccination rates have a relatively lower CFR. Looking at countries' age demographics, we found that vaccination rates and deaths tend to be far lower among countries with younger population demographics. We also notice a clear trend among vaccination rates and cases among countries when accounting for population size. We found countries with higher total health expenditure in 2018 have higher COVID-19 vaccination rates. We also examined political trends, and found that a higher democracy score does not indicate a higher vaccination rate. On the flip side, almost all the highly vaccinated countries have high democracy scores, among which mostly are European countries.

We see that, as of April 24, 2021, countries with higher COVID-19 case rates, lower COVID-19 death rates and case fatality rates, older populations, and higher total health expenditure tended to be making the best progress in vaccinating their populations. Our data does not capture what may be the most crucial factor: access to COVID-19 vaccines. Vaccinating the world will be a global effort, and observing these trends will be important in targeting countries that need aid to improve their COVID-19 vaccination rates. This data is changing daily; our analysis could be run again with updated COVID-19 data, as well as a more targeted analysis on income/infrastructure inequalities, to continue to monitor these trends.

Data Sources

COVID-19 Data: Our World in Data
https://ourworldindata.org/explorers/coronavirus-data-explorer
https://ourworldindata.org/covid-vaccinations

Age Demographics Data: Our World in Data
https://ourworldindata.org/age-structure

Health Expenditure Data: World Health Organization
https://apps.who.int/nha/database

Political Data: Our World in Data
https://ourworldindata.org/grapher/political-regime-updated2016