# Impact of Demographic Characteristics of a Country on COVID Vaccine Distribution and Manufacturer Prevalence.

By: Simranjit Kaur, Joanna Kim, Brandon Lee, and Laila Voss.

### Author contributions

Simranjit Kaur created several scatterplots of the share of vaccinations against socioeconomic variables, explored the linear regression of share of vaccinations against socioeconomic variables, wrote part of the abstract, and wrote part of the discussion.

Joanna Kim created several scatterplots of the share of vaccinations against socioeconomic variables, explored the linear regression of share of vaccinations against socioeconomic variables, wrote each figure's explanation, wrote about the scatterplot's results, wrote part of the abstract, and wrote part of the discussion

Brandon Lee tidied the data, merged the initial datasets, interpreted each principal component, created the scatterplots for the PC combinations and wrote the accompanying analysis, worked on the multiple linear regression on the dataset used in the PCA analysis, and wrote part of the discussion

Laila Voss prepared the data for principal component analysis and developed the substance for this section, including creating the correlation table, heatmap, and the PC loading plots. She was also responsible for the writing accompanying these portions and the multiple linear regression, and made contributions to the synthesis of the final report.

### Abstract

The speed of vaccine invention, manufacturing and distribution have undoubtedly been crucial in allowing us to see a "light at the end of the tunnel" in regard to the COVID-19 pandemic. However, many countries vary wildly in the number of vaccine doses they have been able to secure and adminster, which has contributed to widespread inequality and raised questions about the role of privilege, wealth and development of countries in the health of their citizens. Our analysis was primarily related to the demographic and socioeconomic characteristics of a country, with two main aims: to identify the relationship between these traits and the number of vaccinations distributed per hundred people, and to identify the relationship between these traits and the distribution of vaccine manufacturers in a country. In regard to the first aim, we found that the level of development of a country (as proxied through birth rate, life expectancy and GDP per capita) is the most important factor in the success of vaccine supply and distribution. In regard to the second aim, we found that covariates such as total reserves, female labor force, and expenses were the most important factors that determined which vaccine manufacturer was most utilized in a country.

---

# Introduction

## Background

The pandemic has been going on for over a year and vaccines have started to distribute recently. There are gaps among countries with the rate of vaccination as well as even the type of vaccines. The most popular vaccines globally are Pfizer BioNTech and AstraZeneca. However, there are many more, such as the Moderna, Johnson & Johnson, and Sinopharm. Seychelles, U.A.E. and Israel are the countries with the highest doses administered per 100 people. But what factors play into those numbers?

Apart from the location of a country, there are other factors that may play a part in which vaccine each country uses. Some of the variables we will look into are the country's GDP, total reserves, and female labor force.

The data that we are looking at is a merged dataset of countries' vaccination data and countries' economic, agricultural, and societal data. We combined these two datasets to see if we can notice any correlations or patterns between a country's vaccination characteristics and the country's demographic characteristics.

## Aims

We aimed to understand how a country's demographic statistics, including life expectancy and birth rate, impact the total number of vaccinations per hundred people. Since these factors are also correlated to a country's status as developed or developing, we were interested in discerning a "profile" for countries that have been more successful in vaccinating their population. To do this, we used principal component analysis to understand drivers of variation in the data and also undertook a multiple linear regression to identify key predictors of the number of vaccines distributed. In the end, we found that the level of development of a country (as proxied through birth rate, life expectancy and GDP per capita) is the most important factor in the success of vaccine supply and distribution.

Moreover, we aimed to explore the relationships between a country's economic indicators, such as GDP per capita, and the manufacturer's distributions on the vaccines they are administering. We examined which societal characteristics, such as the prevalence of gender inequality, correlated with the different vaccine manufacturers. To understand these different relationships, we utilized regression and visual explorations. In the end, our regression model failed to provide us with any valuable information, so we focused on analyzing the trends through scatter plots. We found that covariates such as total reserves, female labor force participation, and expense seemed to have a clear relationship with the different vaccine manufacturers. Other variables, such as GDP per capita displayed relationships that were a bit more difficult to discern, but ultimately aligned with our findings from the previous plots.

---

# Materials and methods

## Datasets

---

## 1. Data description

The data analyzed in this report is a collection of various vaccine counts sampled daily in different countries globally. Our data set contains covariates such as diabetes, varying population, imports, GDP per capita, expenses, total reserves, and labor force. These data are publicly available:

> Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. Nat Hum Behav (2021).
>
> World Bank. World Development Indicators, The World Bank Group, https://data.worldbank.org/indicator. 20 May 2021.

### Sampling and Measurement Information

The raw data on vaccination doses administered was collected through a combination of manual and automated means. Data was collected from government websites, health ministries, dedicated dashboards and social media accounts of national authorities manually. Data from official sources in a machine-readable format were collected using Python scripts. Data that was not machine-readable was aggregated by third-party sources also using automated methods. The World Bank uses the Development Data Group to collect data through various ways, the main one being through a globally coordinated program.

### Data Structure

For this study, the observational units are countries and the relevant population and sampling frame from the vaccine data is all people who received a dose of the various vaccines along with the countries. In terms of scope, we found that the vaccine data accounts for every country, and, therefore, we believe it is representative of the whole population. It still might be limited, however. This limitation may be due to constraining our analysis to more recent dates, which might lead to some information loss.

> **Table 1**: Variable descriptions and units for each variable in the dataset.

| Name | Variable description | Type | Units of measurement |
|---|---|---|---|
| location | Name of the country (or region within a country) | Character | NA |
| vaccine | Name of the vaccine being adminsitered | Character | NA |
| total_vaccinations | Total number of doses administered. This is counted as a single dose, and may not equal the total number of people vaccinated, depending on the specific dose regime (e.g. people receive multiple doses). again. | Numeric | Ones |
| birth rate | The number of live births occurring during the year, per 1,000 population estimated at midyear. | Numeric | Thousands |
| life expectancy | The number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. | Numeric | Years |
| hospital beds | Inpatient beds available in public, private, general, and specialized hospitals and rehabilitation centers. | Numeric | Thousands |
| population % (15-64) | Total population between the ages 15 to 64 as a percentage of the total population | Character | Percentage of the total population |
| population % (65+) | Population ages 65 and above as a percentage of the total population | Numeric | Percentage of the total population |
| population, total | Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. | Numerica | Billions |
| urban population % | People living in urban areas as defined by national statistical offices. | Numeric | Percentage of the total population |
| labor force female | Labor force participation rate is the proportion of the population ages 15 and older that is economically active | Numeric | Percentage of total work force |
| imports | Imports of goods and services represent the value of all goods and other market services received from the rest of the world. | Numeric | Percentage of GDP |
| gdp per capita | GDP per capita is gross domestic product divided by midyear population. | Numeric | US Dollars in Thousands |
| expense | Expense is cash payments for operating activities of the government in providing goods and services | Numeric | Percentage of GDP |
| total reserves | Total reserves comprise holdings of monetary gold, special drawing rights, reserves of IMF members held by the IMF, and holdings of foreign exchange under the control of monetary authorities. | Numeric | NA |

The first few rows of the data are shown in Table 2.

> **Table 2**: Example rows of vaccine data.

| | location | total_vaccinations | people_vaccinated | daily_vaccinations | total_vaccinations_per_hundred | people_vaccinated_per_hundred | daily_vaccinations_per_millio |
|---|---|---|---|---|---|---|---|
| 0 | Afghanistan | 504502 | 448878 | 13921 | 1.3 | 1.15 | 35 |
| 1 | Albania | 622507 | 440921 | 12160 | 21.63 | 15.32 | 422 |
| 2 | Antigua and Barbuda | 31262 | 31262 | 63 | 31.92 | 31.92 | 64 |

## Methods

First, to examine relationships between the number of vaccines distributed per hundred people and characteristics of a country, principal component analysis was performed on demographic data, including variables such as life expectancy and birth rate, in order to identify drivers of variation. We looked at proportional and cumulative variance by number of components, created loading plots and examined different combinations of the principal components. We also performed a multiple linear regression on the number of vaccines distributed per hundred people with the demographic information as the dependent variables. Second, to examine the relationships between the variety of vaccine manufacturers and characteristics of a country, a linear regression model was used to see which variables could be statistically significant for the distribution of the vaccine. Scatterplots were then made to visualize the distribution of vaccines based on a socioeconomic factor, with points color coded for each vaccine manufacturer.

---

# Results

## Relationship between demographic variables and number of vaccines distributed

First, we examined the correlation between our demographic characteristics and the total number of vaccinations per hundred people; we chose this vaccine measure because it was available for every country in the set and is scaled by population since we divide by hundreds of people. We drop our other vaccine variables so that we do not identify vacuous relationships.

### Correlation between total vaccinations per hundred and demographic variables

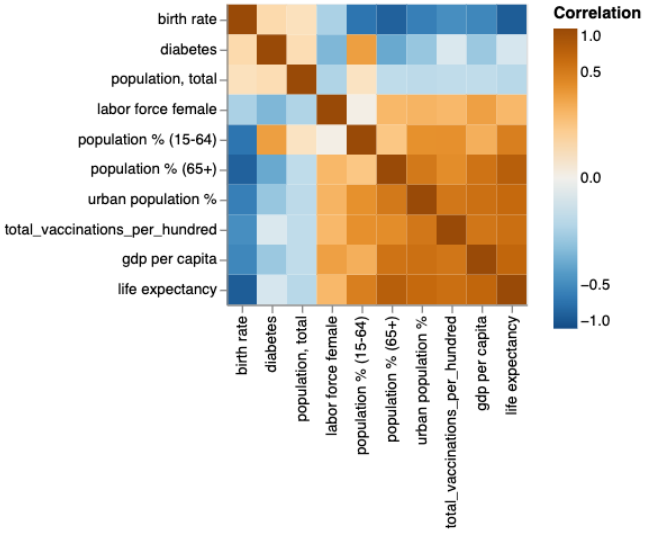
> **Table 3**: A table showing the correlation between total vaccinations per hundred and demographic variables.

| Variable/Indicator | Correlation [0,1] |
|---|---|
| Birth rate | -0.481145 |
| Population, total | -0.072408 |
| Diabetes | -0.023062 |
| Labor force female | 0.184706 |
| Population % (15-64) | 0.375170 |
| Population % (65+) | 0.392794 |
| Urban population % | 0.535046 |
| GDP per capita | 0.542155 |
| Life expectancy | 0.597680 |

As a result, from examining these entries, it can be seen that the number of total vaccinations per hundred people is:

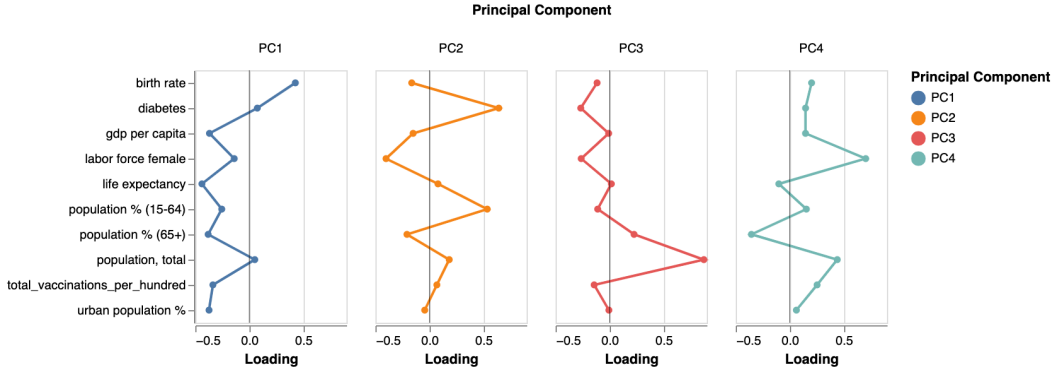
- strongly *negatively* correlated with birth rate and total population, meaning it *tends to vary in opposition* with these variables;
- strongly *positively* correlated with life expectancy, GDP per capita and the percentage of population that is urban, meaning it *tends to vary together* with these variables.

> **Figure 1**: A heatmap showing the correlations between the different variables.




After normalizing our data and carrying out PCA, we examined the proportion of variance explained by each component as well as the cumulative proportion. We found a sharp drop off in variance explained after the first component. In order to select a number of components, we plotted the variances to determine the fewest number of principal components that capture a considerable portion of variation and covariation. With four components (i.e. PC1, PC2, PC3 and PC4), we are able to cumulatively explain about 78% of the data so we will use these moving forward. Plotting the principal components allows us to examine the upweights and downweights to derive interpretations for each.

> **Figure 2**: Principal component loading plot.




### Interpretation of Principal Components

The first principal component seems to be a representation of a country's level of development due to birth rate being up-weighted while GDP, life expectancy, population percentage (ages 65+), and urban population percentage are fairly equally down-weighted.

The second principal component seems to be a representation of the general level of health in a country since diabetes and population percentage of people aged 15 to 64 are significantly up-weighted.

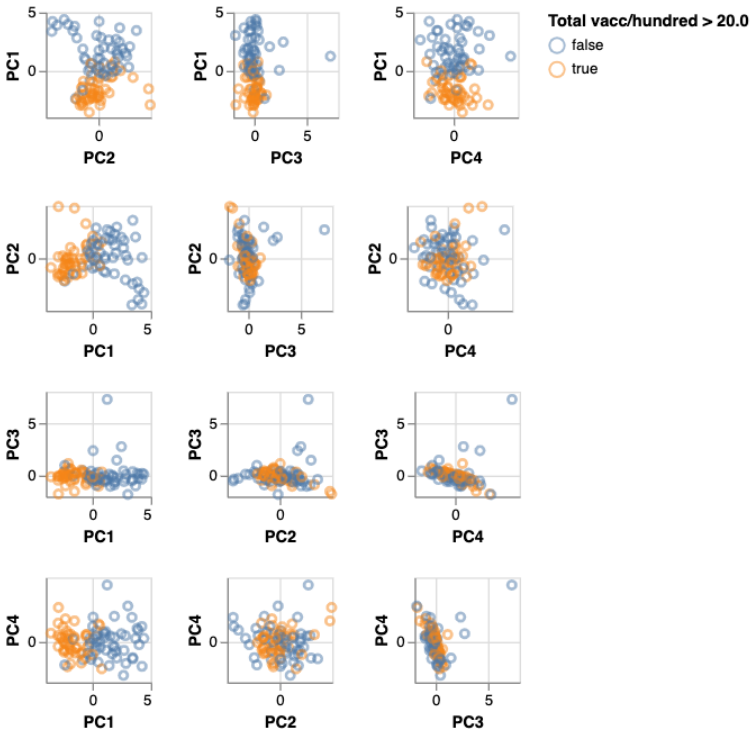
The third principal component seems to be a measure of a country's population size due to the total population variable being so heavily up-weighted compared to the other variables.

The fourth and last principal component seems to be a measure of a country's level of gender equality in the workforce due to the labor force female variable being so highly up-weighted.

### Interpretation of Principal Component Combinations

From the scatterplots below, we can see that PC1 seems to be the only principal component that significantly influences the total number of vaccinations per hundred people, since PC1 is the the only principal component that had a clear divergence on whether or not total number of vaccinations per hundred people is greater than 20 or less than 20 (which is around the median of 18.635). The other principal components do not seem to indicate a clear pattern where a higher or lower PC value correlates with more or less total vaccinations per hundred people due to there being no evidence of a clear divergence similar to what we observed in the PC1 scatterplot in any other plots. As such, it seems that the variables that are related to a country's level of development (birth rate, GDP per capita, life expectancy, population percentage (ages 65+), and urban population percentage) are the variables that have the largest influence on a country's ability to successfully distribute the vaccine to a larger portion of the population.

**Figure 3**: Scatterplot comparisons between each combination of principal components




### Regression

To verify these results, we run a multiple linear regression and examine both the estimates and standard errors for each variable; in the table below, we also include the number of standard errors from 0 to allow for interpretation of signifiance. Birth rate, life expectation and population are all significant and have relatively high coefficient estimates, which indicates that the level of development as related to these variables is crucial in predicting the number of vaccines distributed.

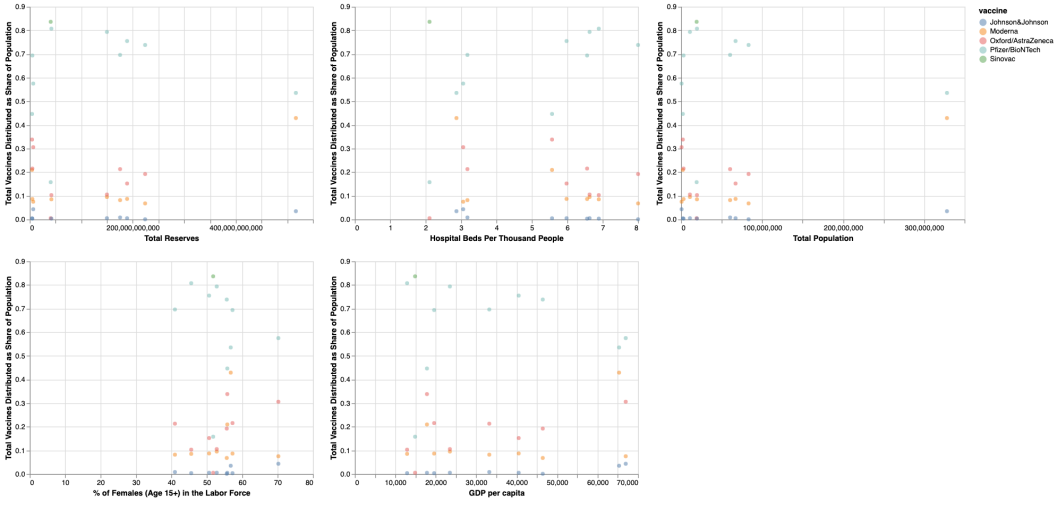
**Table 4**: A summary of the regression model showing the significance of each covariate.

| Variable | Coefficient Estimate | Standard Errors | # SE from 0 |
|---|---|---|---|
| intercept | -527.381 | 170.304 | 5.37183e-232 |
| birth rate | 4.37517 | 1.8254 | 43.5265 |
| life expectancy | 2.08511 | 0.720343 | 11.169 |
| diabetes | -0.0907378 | 0.624472 | 1.46245 |
| population % (15-64) | 4.22373 | 1.64909 | 41.4094 |
| population % (65+) | 3.08519 | 1.56958 | 13.9347 |
| population, total | -3.04847e-10 | 2.25303e-09 | 4.43848e+08 |
| urban population % | 0.178782 | 0.133722 | 8.94216 |
| labor force female | 0.0293287 | 0.177705 | 5.79479 |
| gdp per capita | 0.000194029 | 0.000127703 | 7832.17 |

### Relationship between demographic variables and vaccine manufacturers

To find trends in each vaccine manufacturer based on socioeconomic variables, the percent of total distributed vaccines were plotted against the socioeconomic variables. Then, each point was color coded based on vaccine manufacturer. That way, percent of total distribution of each vaccine based on the socioeconomic variable could be easily seen.

**Figure 4**: Scatterplots of the various covariates against total vaccine distributed as a share of the population, to show the trends in each vaccine manufacturer based on socioeconomic variables.




#### Scatterplot of total reserves against vaccines distributed as a share of the population

As total reserves of a country increases, the Pfizer/BioNTech distribution seems to increase increases, as Oxford/AstraZeneca decreases.

#### Scatterplot of hospital beds per hundred people against vaccines distributed as a share of the population

As hospital beds of a country increases, Pfizer/BioNTech distribution increases, as Oxford/AstraZeneca decreases. Although an increased number of hospital beds does not necessarily indicate wealth, it did follow the same trend as the plot of total distributed vaccines percentage against total reserves.

#### Scatterplot of population against vaccines distributed as a share of the population

A larger population does not signify greater wealth; for example the United States has a population of 328.2 million people with a GDP per capita of around $65,000 while India has a population of 1.366 billion with a GDP per capita of around \$2,000. Still, this does exhibit the same relationship as the two previous figures.

#### Scatterplot of percent of females in the labor force against vaccines distributed as a share of the population

As the percentage of females over the age of 15 in the labor force of a country increases, Pfizer/BioNTech distribution decreases, as Oxford/AstraZeneca increases. This shows the opposite relationship of the first two figures since as the variable increases, Pfizer/BioNTech's distribution percentage decreases and Oxford/AstraZeneca increases.

#### Scatterplot of GDP per capita against vaccines distributed as a share of the population

As the percentage of females over the age of 15 in the labor force of a country increases, Pfizer/BioNTech distribution decreases, as Oxford/AstraZeneca increases. This shows the opposite relationship of the first two figures since as the variable increases, Pfizer/BioNTech's distribution percentage decreases and Oxford/AstraZeneca increases.

---

## Discussion

From the principal component analysis we conducted earlier, we found that variables that are related to a country's level of development such as birth rate and life expectancy had the most significant influence on how successful a country's vaccination distribution is (based on the number of vaccinations per hundred people), and we further supported this finding by running a multiple linear regression and concluding that birth rate, life expectancy and population are all significant in predicting the number of vaccines distributed due to having relatively high coefficient estimates. After these findings, we would have liked to explore the main causes for why countries with lower levels of development (in terms of birth rate, life expectancy, and GDP per capita) seem to be more capable of supplying and distributing vaccines to a higher proportion of their population, since, intuitively, countries with higher levels of development would have more resources and better healthcare systems to better enable successful distributions of vaccinations to the populace.

Upon building our regression model to further explore relationships between (socio)economic factors and vaccine distributions for the various manufacturers, we found that the p-values for all of our covariates were greater than 0.05, making them insignificant. While useful, p-values are not the whole story since we can end up with an insignificant p-value but a reliable estimate of the trend. Moreover, we found that the coefficient estimates were very small and not reliable (before and after normalizing). Because the regression model gave us insignificant/inconclusive results, we decided to focus our attention on visual exploration. From the six scatterplots kept, some of the relationships' meanings were unclear. More specifically, although the hospital beds plot showed the same relationship as the total reserves plot, it does not mean that the same conclusion can be drawn because hospital bed numbers are dependent on the pandemic impact on the country, not wealth. The total population and female labor force plots showed the same relationship as the total reserves plot. However, what an increased population and female labor force means for a country's socioeconomic status is unclear.