Analyzing factors impacting COVID-19 vaccination rates through data mining.

Dongseok Cho, Mitchell Driedger, Sera Han, Noman Khan, Mohammed Elmorsy, Mohamad El-Hajj
Department of Computer Science
MacEwan University
Edmonton, Canada
{chod5, driedgerm3, hans28, khann24}@mymacewan.ca
{elmorsym, elhajjm}@macewan.ca

***Abstract*—Since the approval of the COVID-19 vaccine in late 2020, people have been getting vaccinated around the globe. While many governing bodies have starting mandated COVID-19 vaccinations, the vaccination rate differs from one country to another. This analysis used COVID-19 vaccination data from Our World in Data and country indicators from the World Bank to determine what factors are associated with vaccination rates in a country. Unsupervised learning such as k-means clustering was implemented to cluster groups of countries with similar indicators. Within these clusters, the Spearman and Pearson correlation method was applied to display indicators that formed these groups. The final step of this analysis was comparing the correlations found from K-mean clustering to the COVID-19 and social indicator data set to see if the factors affect the vaccination rate. These methods found that GDP per capita and governance factors played a significant role in determining vaccination rates. Countries with a high GDP per capita generally had higher percentage of vaccination rates, with the opposite for low GDP per capita. Some governance factors were observed to correlate highly with vaccination rates in European countries. Countries with a high trust in government and low violence environments displayed this. Population, education, and access to information indicators were assumed to positively relate to the vaccination rate, however this analysis concluded this was not the case.**

***Keywords*—COVID-19, Data Mining, Clustering, Correlation, Vaccination Rate, Governance, GDP, Health, Education, Access to Information, Literacy Rate**

I. Introduction

Coronavirus disease (COVID-19) is a high infection disease known to affect the respiratory and cardiovascular systems. The SARS-CoV-2 virus causes the disease. According to the World Health Organization (WHO) [1], most people infected with the virus will experience mild respiratory symptoms and do not require any medical attention. However, individuals with underlying medical conditions and those older have an increased risk of serious illness. WHO recommends staying informed about the disease, practising social distancing, wearing a proper mask and staying sanitized are all methods to slow the transmission of the disease [1]. COVID-19 began in Wuhan, China, in December of 2019, where a pneumonia outbreak raised international concern [2]. Before long, the virus had spread to most countries and was starting to have a massive global impact. At the end of 2020, the world has dramatically impacted people's livelihood, public health, the workforce, etc. Countries had varying responses to the pandemic, including physical interaction lock downs, travel restrictions, and masking laws. Almost half of the world's population was at risk of losing their livelihood [3].

Throughout 2020, different pharmaceutical companies had been developing a vaccine for COVID-19. By December, the first vaccinations were approved for public use, and individuals were starting to get vaccinated [4]. Several of these companies recognized as top vaccination developers include Pfizer, Moderna, AstraZeneca, and Johnson & Johnson. The development of the COVID-19 vaccine was done extremely quickly, considering that entirely developing a vaccine often takes many years

[4]. However, the distribution of the vaccine has not seen the efficiency that development had. Multiple factors have impacted the differences between different countries' vaccination rates. When a country first began to vaccinate was likely a product of how quickly they could acquire vaccination doses in the masses. For example, lower-income countries in Africa do not typically have the means to manufacture vaccinations locally, and depending on getting doses imported slows the advancement of vaccination rates [4]. Another impacting factor on vaccination rates is the hesitancy to vaccinate. A study done in late 2020 found that is a large variance of COVID-19 vaccine hesitancy among different countries [5].

With many variables impacting the rate of people becoming vaccinated, it draws the question: are there hidden underlying factors associated with the variation of vaccination rates among different countries? What are the indicators of a country predisposed to vaccinate its population efficiently, and is there a deeper meaning to be found? The goal of this study was to take a data-driven approach to answer these questions. By analyzing the relationships between vaccination statistics and socioeconomic factors between different countries, common indicators of high or low vaccination rates could be discovered and analyzed. RStudio was used to clean, combine, and analyze the data in this study [6].

II. Related Literature

The academic environment surrounding COVID-19 is constantly developing as the global effects of the virus are observed.

A novel coronavirus outbreak of global health concern is one of the earliest publications regarding the COVID-19 pandemic. This literature describes the early onset of the virus, beginning in Wuhan, China, in late 2019. It details the clinical features of the first patients that were confirmed to be infected with COVID-19, and out of the 41 patients in the first cohort of testing, 22 developed dyspnoea, 13 had to be admitted to an intensive care unit, and six died. While these findings were stated to be treated with caution, such a high mortality rate was very alarming. The study serves as a warning, with the ending statement calling for all efforts to understand the disease and act against it [2].

How COVID-19 vaccine supply chains emerged during a pandemic is a comprehensive look at the formation of COVID-19 vaccine supply chains concerning different countries and companies. It recalls how the vaccine supply chains under Pfizer, Moderna, Novavax and other companies were developed within 2020 and 2021. The vaccine distribution rolled out differently in different world areas; for example, the United States and the European union had higher rates of mRNA vaccines such as Pfizer and Moderna being administered, while China strictly administered domestic vaccines. The study concludes by encouraging researchers to determine if there was a quicker way to manufacture these vaccines in order to prepare for the next pandemic [4].

A Vaccination Simulator for COVID-19: Effective and Sterilizing Immunization Cases is an epidemic simulator that simulates the impact of different vaccination strategies. This study was published in December of 2021 and is an encompassing simulator that yields valuable findings. According to the model, a sterilizing-age-based vaccination scenario concludes with the most positive results. In this scenario, vaccinations are based on age. Other results showed that across all scenarios, having higher vaccination rates decreases mortality rates [7].

COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates summarizes COVID-19 vaccine hesitancy displayed the findings of 31 published studies. A large variety of acceptance rates were found in this study. Countries with the highest vaccine acceptance rates include Ecuador, Malaysia and Indonesia, with over 90% acceptance, while Kuwait, Jordan and Italy showed high hesitancy rates. Some surveys were among healthcare workers only, with the acceptance range from 78.1% to 27.7%. Lower acceptance rates were observed throughout the Middle East, Russia, Africa, and some European countries. This study was completed in January 2021, and it concludes by encouraging governments and other bodies to build vaccine trust with the general public in order to prevent vaccine hesitancy from hindering rates of COVID-19 vaccination [5].

III. Data Collection and Cleaning

The data analyzed in this study include COVID-19 vaccination data and a broad range of country indicator statistics. COVID-19 data used in this study was obtained from Our World In Data [8]. This expansive data set contains a range of COVID-related statistics for most countries, including cases, deaths, vaccinations, hospitalizations, tests, and a handful of relevant country indicators. Multiple data sets were collected from the World Bank [9] and compiled into one central data set.

A. COVID-19 Data

This data set is structured as having one row per day, per country. These are presented chronologically within each country and by order of country ISO code. The starting day per country is variable when said country begins to submit COVID-19 related statistics. The ending day per country is the extraction date, February 1, 2022. The raw data includes the country's ISO code, continent, country name, date, and the previously mentioned columns. Due to the focus of this study being on the vaccination rates, the majority of the non-vaccine related data was discarded. This remaining vaccination data included the total number of vaccinations, people vaccinated, fully vaccinated, and a total number of boosters (count and per hundred people for each). Also remaining was new vaccinations (count, smoothed, smoothed per million) and new people vaccinated (smoothed, per hundred); however, these were discarded. Many countries in this set did not provide updated statistics every day. The dates that did not receive updated vaccination data were left with blank cell values. To remove these N/A values, the last filled row per each country was filled in for subsequent rows within a column until another filled cell was encountered or a new country began.

A new set was created called COVID Metadata to summarise this vaccination data. This consisted of a country for each row, along with various statistics generated from the larger vaccination set. The number of vaccinations, total vaccinations, and boosters per hundred were recorded at the beginning of the January, April, July, and October months of 2021 and January 2022. Also listed is the first vaccination date, as determined from the vaccination data set.

COVID Metadata Columns
ISO Code
Continent
Date of First Vaccination
People Vaccinated Per Hundred 2021-01-01
People Fully Vaccinated Per Hundred 2021-01-01
Total Boosters Per Hundred 2021-01-01
People Vaccinated Per Hundred 2022-04-01
People Fully Vaccinated Per Hundred 2021-04-01
Total Boosters Per Hundred 2021-04-01
People Vaccinated Per Hundred 2021-07-01
People Fully Vaccinated Per Hundred 2021-07-01
Total Boosters Per Hundred 2021-07-01
People Vaccinated Per Hundred 2021-10-01
People Fully Vaccinated Per Hundred 2021-10-01
Total Boosters Per Hundred 2021-10-01
People Vaccinated Per Hundred 2022-01-01
People Fully Vaccinated Per Hundred 2022-01-01
Total Boosters Per Hundred 2022-01-01

TABLE I
Columns in COVID Metadata data set

B. World Bank Data

Three data sets were collected from the World Bank's data catalogue: world development indicators [10], environment, social and governance data [11], and health nutrition and population statistics [12]. The world development indicators provide a basis to compare countries regarding their general quality of life. The environment, social, and governance data set and the health nutrition and population data set both contain data that has likely impact on the vaccination rate, so they were included as well.

Only data from 2010 to 2020 was included for each of these data sets. The data was organized to reflect the following format: one row for each country and indicator values for each year as columns. This was done to align the data structure with the COVID Metadata data set previously created. A hindrance to the effectiveness of this data is a large number of countries and indicators that have incomplete data. Specific countries do not supply data for many indicators, which proves an issue for analyzing data among these countries. To mitigate against this, certain indicators and countries are ignored during the analysis. This was a case-by-case basis to include as much data as possible without introducing misleading data to the analysis algorithms. These resulting sets were combined with the COVID Metadata data set, using the country

ISO code. This resulted in a complete representation of each country, including important vaccination statistics and collected indicators from the previous decade.

IV. Methodology

To determine indicators associated with vaccination rates, many tools were utilized within RStudio. First, indicators with high correlations across all data were identified using linear regression (IV-C) and Pearson/Spearman correlations (IV-D). Next, k-means clustering was applied to specific or all countries based on the indicators in the previous step (IV-B). Lastly, indicators with high vaccination correlation were identified within these clusters using Pearson and Spearman correlations (IV-D). Throughout this process, data was manipulated, filtered, and visualized using the R language libraries in RStudio (IV-A).

A. *RStudio*

RStudio was used to manipulate, filter and visualize data, as well as apply intelligent tools such as clustering. RStudio is an open source software that is used for data science [6]. It provides an interactive development environment for the data science language R. R contains many functionalities to analyze data. For example, filtering methods in R were used to derive countries with specific indicator values. The library ggplot2 [13] was used to visualize data in different ways, including histograms, scatter plots, correlation plots, and cluster plots.

B. *Clustering*

Dealing with many indicators from 208 countries was a time-consuming process. To answer the initial questions, methods were required to group countries and analyze their significance while no predicted pattern existed. Therefore, clustering was utilized, a widely-used unsupervised learning method. The k-means variation of clustering was applied in this study, due to k-means being simple and efficient. An example of this method is displayed in Figure 1 This method is proved to be a very effective way to produce quality clustering results [14]. A k-means clustering algorithm groups similar points together in the form of clusters. This number of groups is represented by K [15]. Hence, choosing the proper K value is very important, as it affects the clustering performance. The Elbow method was implemented, a commonly used K value configuration tool that distinguish the optimal K value while calculating the within-cluster sum of the square [16]. After plotting an Elbow method figure, the K value can be determined at the location of the elbow point on the graph. Once the K value is decided, the clustering process can be completed.

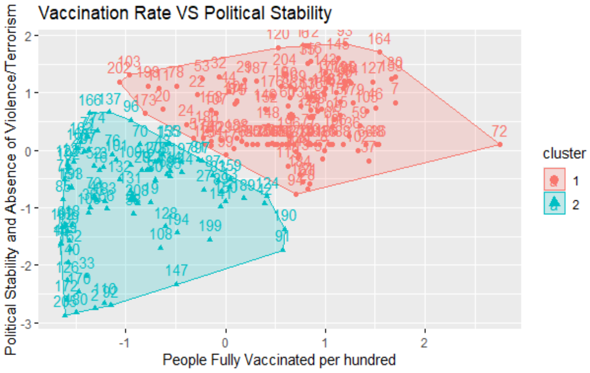

Fig. 1. Clustering example of political stability against vaccination rate.

C. *Linear Regression*

To analyze the association between indicators and vaccination rate, linear regression tables were constructed. Linear regression examines how significant variables are at predicting other variables [17]. This study used simple linear regression, which has one dependant variable and one independent variable. A line of best fit is plotted on a linear regression graph, displaying the level of positive or negative correlation between these variables. This can be observed in Figure 2.

D. *Correlational Methods*

Pearson[18] and Spearman[19] correlation coefficients were used to investigate the correlation between indicators, dependant on the condition of the data. These statistical methods are widely used to measure the association between two variables. Moreover, these methods demonstrate the degree of association between variables, represented in a range from +1 to -1. The two ends are perfect associations, with +1 being a positive association

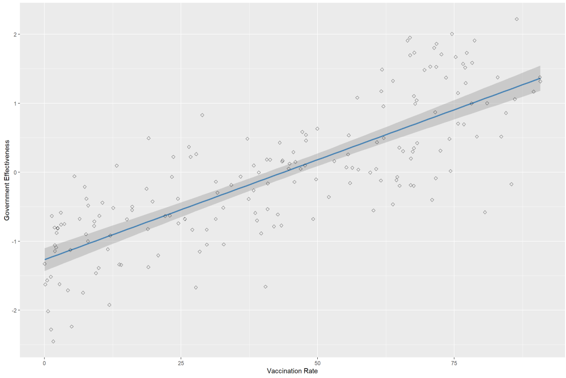

Fig. 2. Linear regression graph example of government effectiveness and vaccination rate.

and -1 as a negative association. Since the data analyzed contains mainly continuous variables, the Pearson method was used to examine the linear relationship between two variables. The Spearman coefficient was applied when the variables observed were of the ordinal type. An example of results from Spearman correlation is demonstrated in Table II.

Indicators	Correlation(%)
GDP per capita (current US$)	42.73
People using at least basic drinking water services (% of population)	41.85
Urban population (% of total population)	45.69
Government Effectiveness: Estimate	49.18
Rule of Law: Estimate	40.04

TABLE II
EXAMPLE OF SPEARMAN CORRELATIONS DERIVED FROM FIGURE 1.

V. ANALYSIS

Prior to this data analysis, assumptions were made as to which indicators would be heavily correlated with vaccination rates to investigate these indicator categories. These include GDP per capita, population, governance indicators (government effectiveness, rate of terrorism), health indicators, education levels, and access to information indicators. These assumptions were based on discussion and reason and comprised the focus of the analysis.

Figure 3 is a histogram displaying all countries' fully vaccinated percentages as of January 2022. As depicted in the figure, many countries have very little of their population vaccinated, with over 30 countries between 0% and 10%. The 10%-60% range is more evenly distributed, and there are over 6 countries in the 60%-80% range. Only a few countries can boast a 90%+ vaccination rate.

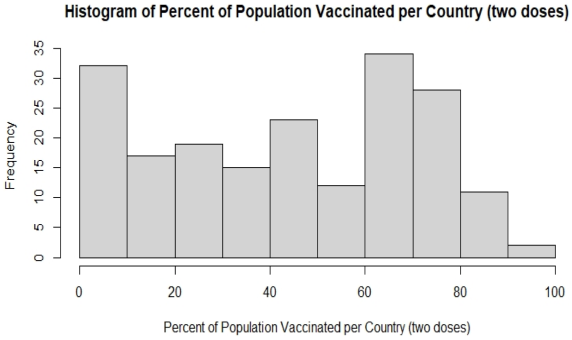

Fig. 3. Percentages of country's fully vaccinated populations as of January 1, 2022

Figure 4 illustrates the opposite sides of the vaccination spectrum by highlighting countries between 0%-10% and 75%-100% vaccinated. The majority of countries under 10% are located in Africa, with the higher end spread globally.

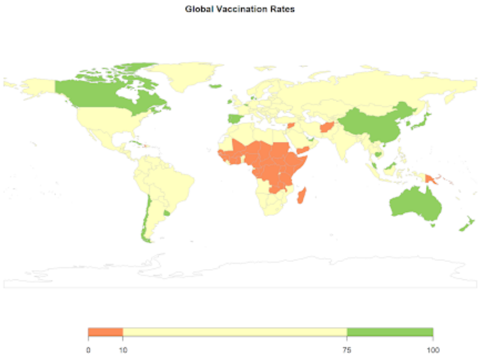

Fig. 4. Map illustration of fully vaccinated rates per country

An initial comparison between select indicators and vaccination rates was executed to provide an idea of which of the assumptions made were correct. Linear regression plots were constructed with GDP per capita, population, health expenditure, internet access, tertiary enrollment, and government effectiveness indicators as one variable, and vaccination percentages as the other. These indicators were from the most recent year with sufficient data, 2019, for all categories except health expenditure and tertiary enrollment (2018). The percentage of fully vaccinated (two doses) individuals in a country's population on January 1, 2022, was selected to represent the vaccination rate. These findings are displayed in Figure 5.

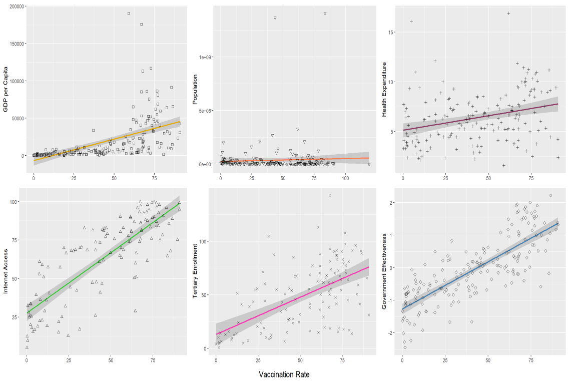

Fig. 5. Linear regression tables of six initial indicators and vaccinations per hundred in January 2022 (top to bottom, left to right): GDP per capita, population, health expenditure, internet access, tertiary enrolment, government effectiveness.

GDP per capita shows a high variance in GDP per capita values under the higher vaccination rates, while the population does not correlate with all countries. The rest of the indicators display a broad distribution of plots, with internet access and government effectiveness showing more positive lines of best fit. An important note is that countries not supplying data for selected indicators were omitted from these results, and therefore a deeper analysis is necessary.

While these indicators make initial impressions of the six categories of assumptions, they are only individual indicators across the entire country range. A further analysis was performed on these and related indicators.

A. GDP per Capita

The GDP per capita indicator displayed a considerate separation between high and low vaccination and GDP per capita values in Figure 5. The lower end of vaccination rates is populated almost entirely with low GDP per capita countries, except the Bahamas, having a very low vaccination rate compared to GDP per capita peers displayed in Figure 6.

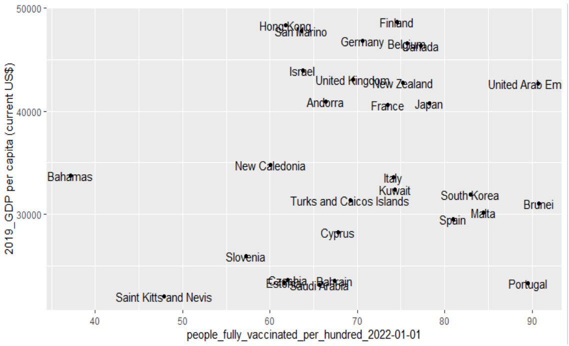

Fig. 6. Scatter plot of 2019 GDP per capita and fully vaccinated percentage in January 2022, within 20,000$ USD and 50,000$ USD.

Within the highest GDP per capita countries, a range of vaccination values is observed, with little association between the two variables. As a whole, GDP per capita is positively correlated with vaccination population values. However, within top and bottom subgroups, little correlation is observed from Figure 5. The top and bottom 30 GDP per capita countries were split into separate bins to investigate these subgroups further. 30 was chosen since it is a small enough group to isolate the GDP per capita correlation within this group and still observe a large enough number of countries to find commonalities within.

The top 30 GDP per capita countries was analyzed using k-means clustering with all other indicators, displayed in Figure 7. Within the clusters formed, cluster 3 (blue) displayed very high correlations with multiple indicators. These are listed in Table III. High positive correlations are found within rates of child immunization to other diseases, as well as negative correlations with the age dependency ratio (young) and fertility rate. Positive correlations with immunization rates is a logical conclusion, as countries that are predisposed to vaccinating are most likely willing to vaccinate against COVID-19. Having a higher age-dependency ratio is more associated with less developed countries, also coming to a logical conclusion. However, the fertility rate having a negative correlation was not expected.

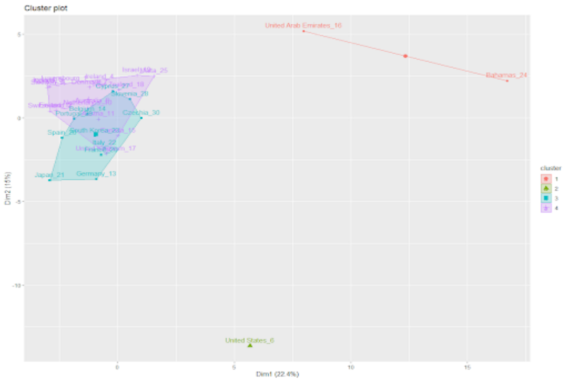

Fig. 7. Cluster of top 30 GDP per capita countries with vaccination percentage as of January 2022 and all other indicators.

Indicators	Correlation(%)
2018 Age Dependency Ratio, Young	-77.60
2018 Fertility Rate, Total (births per woman)	-70.40
2018 Immunization, DPT (% of children ages 12-23 months)	93.10
2018 Immunization, Hib3 (% of children ages 12-23 months)	91.40
2018 Immunization, measles (% of children ages 12-23 months)	87.30
2018 Immunization, Pol3 (% of one-year-old children)	93.60

TABLE III
CORRELATION VALUES WITHIN CLUSTER 3 OF THE TOP 30 GDP PER CAPITA COUNTRIES CLUSTERING.

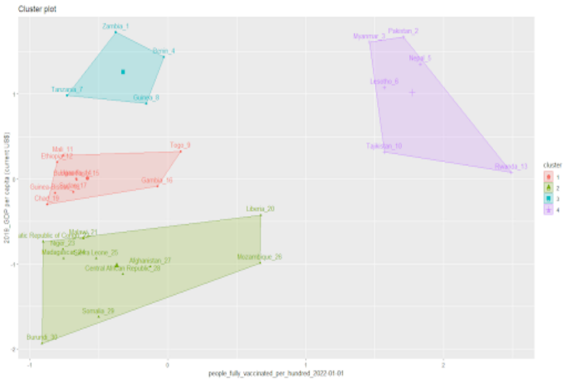

Fig. 8. Cluster of bottom 30 GDP per capita countries with vaccination percentage as of January 2022 and GDP per capita value of 2019.

Indicators	Correlation(%)
2019 Population Growth (annual %)	77.10
2019 Rural Population Growth (annual %)	71.40
2019 Urban Population Growth (annual %)	88.60
2019 Government Effectiveness	71.40
2019 Population Density (people per sq. km of land area)	71.40
2019 Proportion of Seats Held by Women in National Parliaments	82.90

TABLE IV
CORRELATION VALUES WITHIN CLUSTER 4 OF THE BOTTOM 30 GDP PER CAPITA COUNTRIES CLUSTERING.

Clustering the bottom 30 GDP per capita resulted clusters displayed in Figure 8. From these clusters, cluster 4 (purple) was far removed from the other clusters, with the notable correlation values recorded in Table IV. These are all comprised of positive correlations with population growth and governance indicators. As found in section V-C, governance values having a positive correlation with vaccination is a common finding. However, the presence of the % of women in national parliaments indicator is more notable.

Two countries with considerate similarities and differences are the Bahamas and Rwanda. These countries have very similar fully vaccinated percentages as of January 2022, with Rwanda at 41% and the Bahamas at 37%. However, when comparing to each country's GDP per capita peers, Rwanda is the highest vaccinated while the Bahamas is one of the lowest. One of the most notable indicators displaying different values is the proportion of seats held by women in politics, with Rwanda at 61% and the Bahamas at 23%. As previously mentioned, this indicator was one of the highest indicators correlated with vaccination rate within the bottom 30 GDP per capita countries. This governance indicator is further analyzed in section V-C.

B. Population

To determine the effect of the total population on vaccination rates, countries with high and low population counts were analyzed. 2019 population was used as the population indicator, due to it being the most recent population indicator with sufficient data. Keeping consistent with other indicators, the vaccination indicator chosen was the fully vaccinated percentage of population on January 1st 2022, was used. The top and bottom 30 countries sorted by population were selected, with the same reasoning as described in GDP per capita (V-A).

The top 30 countries by population were clustered with population and vaccination rate. This is displayed in Figure 9. Clusters 1 (red) and 2 (green) are mostly comprised of industrialized middle/high power nations (local and international). Cluster 1 includes the United States, Russia, and Mexico. Cluster 2 displayed higher vaccination rates and were located in Asia or Europe, except for Brazil. Some country's in this cluster were Japan, Germany and Turkey Cluster 3 (green) contained countries with lower vaccination rates, being mainly African and South Asian countries such as Ethiopia, Bangladesh, Myanmar and Kenya.

A correlation analysis on cluster 1 was performed, displayed in Table V. Water and sanitation services displayed positive correlations with vaccination rates, as did safe water and the age dependency ratio (old). Negative correlations were found within hand washing stations and life expectancy at birth for females.

Indicators	Correlation(%)
Safe Water	91.90
Handwash Station	-82.90
Age dependency Ratio, Old	72.20
Life expectancy at birth. female(years)	72.20
Number of people who are undernourished	-76.60
People using safely managed drinking water services (% of population)	91.90
People using safely managed drinking water services urban (% of urban population)	73.70
People using safely managed sanitation services (% of population)	76.70
People using safely managed sanitation services urban (% of urban population)	70.20

TABLE V
CORRELATION FOR INDICATORS IN CLUSTER TWO

Cluster 4 (purple) consists of China and India, two nations with drastically higher populations and different vaccination rates. The populations for China and India are listed as 1,407,745,000 and 1,366,417,756, respectively. China displays higher vaccinations than India, with China sitting at 83.6% and India with only 43.57%. Both of these are Asian countries that wield considerable power regionally and internationally. The largest difference between these countries is the form of government and economy properties. China has a rigid authoritarian social structure, while India is has a service-based economy with distinct social class and caste-based system. Governance indicators for China and India are displayed in the Table VI. These values follow standard normal distribution with a range from -2.5 to 2.5. As observed in these indicator statistics, China's government effectiveness, rule of law, and political stability are all rated higher than India, while India has much greater voice and accountability. Both of these countries perform poorly in regards to Control of Corruption.

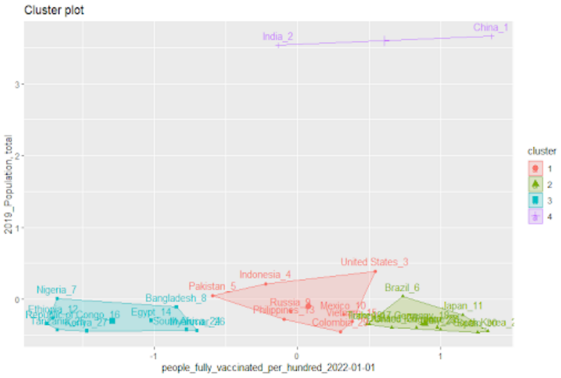

Fig. 9. Clusters for the top 30 populations with vaccination percentage as of January 2022 and total population for 2019

Indicators	China	India
Rule of Law: Estimate	-0.275	-0.030
Government Effectiveness: Estimate	0.517	0.171
Control of Corruption: Estimate	-0.295	-0.254
Political Stability Absence of Violence/Terrorism: Estimate	-0.257	-0.770
Voice and Accountability: Estimate	-1.628	0.269

TABLE VI
GOVERNANCE INDICATOR VALUES FOR CHINA AND INDIA

Clustering the bottom 30 populated countries displayed less populated countries having higher vaccination rates than more populated countries on average, as shown in figure 10. Cluster 3 (blue) contains some small, high GDP per capita nations/territories, typically falling under the umbrella of a larger nation or organization. It also contains island nations, mostly based in North America and Europe. These nations island share the trait of containing tourism-based countries. Cluster 3 accounted for 17 of the bottom 30 countries.

Cluster 1 (red) displayed very low vaccination rates. These countries have lesser economies in common, commonly heavily involved with tourism, agriculture, or fishing. Countries such as Dominica, Saint Lucia and Vanuatu fall within this range. Cluster 3 (green) contained higher populations within the bottom 30 countries, with Iceland and the Bahamas having varying vaccination rates in comparison to the other countries in this cluster. Gibraltar was clustered by itself due to it's very high vaccination rate. Alike counties found in cluster 3, this is a small territory of a larger country, reporting very high GDP per capita values.

The Bahamas and Iceland have the highest populations among the bottom 30 populated countries, with Iceland having a much higher vaccination percentage than the Bahamas. Iceland had a vaccination rate of 76.99% while Bahamas vaccination rate was 37.13%. Their populations were 389486 for Bahamas's and 360563 for Iceland. A considerate difference between these countries is the international support these countries have. The Bahamas is a member of the Caribbean Community, while Iceland is part of the Nordic Council, the Council of Europe, and the European Free Trade Association. A further analysis on the impact of these associations is needed to confirm the impact of these associations with vaccination rate.

Analyzing the top and bottom 30 populated countries did not support population having a large correlation with vaccination rate. The separation of vaccination rates displayed in these clusters can be attributed to differences in GDP per capita and other economic indicators, strongly supported in the bottom 30 countries. Analyzing India and China displayed strong differences with governance factors, with China's authoritarian government yielding higher vaccination rates than India.

C. Governance

Analyzing the bottom 30 GDP per capita countries displayed the women's seats in national parliaments indicator having over 80% correlation with vaccination rate in one cluster (V-A). In the Bahamas and Rwanda, this indicator was a notable difference where vaccination rate was the same. Further analysis was done to determine if this relation between women's seats in the national parliament and the vaccination rate was consistent across all countries.

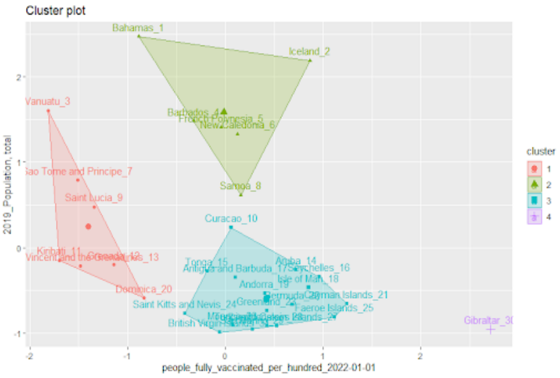

Fig. 10. Clusters for the bottom 30 populations with vaccination percentage as of January 2022 and total population for 2019

The top 10 countries with women's seats were selected at the beginning of this analysis process. These countries were clustered bases on this indicator and the fully vaccinated percentage of population on January 2022, and correlations within the largest cluster formed are recorded in Table VII. This displayed several indicators with a negative correlation, including incidence of tuberculosis (per 100,000 people), people using at least basic drinking water services (percent of the population), control of corruption, political stability and absence of violence/terrorism, regulatory quality, rule of law, and voice and accountability.

Unlike the Bahamas and Rwanda, comparing correlations results with the entire data set did not show any relations between vaccination rates and women's seats in parliaments. This same procedure was executed with an increased size, being the top 30 countries in order to determine if sample size was impacting this result. These results are listed in Table VIII. Here, women's seats in parliaments is not found to have a high correlation within these 30 countries, rather with other governance factors.

One of the clusters formed in the women's seats in politics indicators was comprised of mostly Eu-

Indicators	Correlation(%)
Incidence of tuberculosis (per 100,000 people)	-80.00
People using at least basic drinking water services (% of population)	-60.00
Control of Corruption: Estimate	-80.00
Political Stability and Absence of Violence/Terrorism: Estimate	-80.00
Regulatory Quality: Estimate	-80.00
Rule of Law: Estimate	-80.00
Voice and Accountability: Estimate	-80.00

TABLE VII
Correlation of Top 10 Women seats in parliaments with vaccination rate.

Indicators	Correlation(%)
Age dependency ratio (% of working-age population)	60.00
Age dependency ratio, old	62.42
Political Stability and Absence of Violence/Terrorism: Estimate	66.06
Voice and Accountability: Estimate	64.85

TABLE VIII
Correlation of top 30 women seats in parliaments with vaccination rate

Indicators	Correlation(%)
Age dependency ratio (% of working-age population)	48.92
Age dependency ratio, old	33.00
Age dependency ratio, young	-3.60
Political Stability and Absence of Violence/Terrorism: Estimate	64.53
Voice and Accountability: Estimate	77.04

TABLE IX
European correlation values with vaccination rate

ropean countries. These countries showed a positive relation with governance indicators. These governance factors were analyzed to determine their correlations with vaccination rates. Correlation tables were made entirely with European countries' governance indicators an vaccination rates, resulting in positive correlations with age dependency ratio indicators (percent of working-age population, old, young), political stability and absence of violence/terrorism, and voice and accountability. Among these indicators, the highest correlation was voice and accountability with 77%, meaning the correlation was strongly positive with European countries that have a government determined by their citizen. These results are recorded in Table IX.

The effect of the age dependency ratio (percent of working-age population) was nest observed within the 48-58% vaccination range. Here, the data showed that the higher the voice and accountability and political stability, the higher the vaccination rate. Outside of this range, some countries displayed results that were not related to the correlation table. This was most likely due to countries not updating the vaccination rate often, or not having enough governance data. Based on these results, a combination of higher reported governance indicators lead to higher vaccination rates. Analyzing the governance factors of European countries displayed that if people trust their government or have some control over their government, higher vaccination rate are reported.

D. Health

While analyzing the women seats in national parliaments with vaccination rate in section V-C, incidence of tuberculosis (per 100,000 people) showed a high negative correlation of 80%. Tuberculosis(TB) is a lung disease caused by germs that are spread from person to person through the air. TB usually affects the lungs, but it can also affect other parts of the body, such as the brain, the kidneys, or the spine. A person with TB can die if they do not get treatment.[20] Since COVID-19 is known to attack lung and make people hard to bread, tuberculosis expected to have a relation with COVID-19 vaccination rate. Therefore, further research was done on TB, by making visuals, table, and comparing with entire data set. However TB turned out to have no to very weak relation with COVID-19 vaccination rate.

One of the clusters formed in section V-A displayed high positive correlations with infants being vaccinated against different diseases. In order to find the relationship with these and the entire data set, further analysis needs to be done.

E. Education

The 2018 tertiary enrolment ratio and literacy rate indicators were used to prove the initial assumption that higher education level and adult literature rate are associated with higher vaccination rates. These two variables were implemented since they represent a higher education and intelligence level within a country. For the data validity, the year 2018 was chosen as it provides the most recent and largely filled data. The tertiary enrolment rate expresses the percentage of total enrollment in post-secondary institutions regardless of age, and the literacy rate counts the percentage of people ages 15 and above. Also, this study aimed to define the relationship between educational factors and vaccination rate; therefore, all countries having values had been added for analysis. The approach was to calculate the correlation coefficient between vaccination rate and two these two variables. Table X shows these correlations.

Vaccination rate VS	Correlation(%)
Tertiary Enrolment	54.54
Literacy Rate	67.22

TABLE X
PEARSON CORRELATION COEFFICIENTS OF VACCINATION RATE AND EDUCATIONAL FACTORS

As displayed in Table X, these educational factors are positively correlated with the vaccination rate. With the tertiary enrolment lesser than the literacy rate. However, it is not safe to say there is a strong correlation between intelligence indicators and vaccination rate, as filtering out the countries not recording data for tertiary enrolment rate gives only 111 countries, and the literacy rate gives only 78 countries. It can be considered just partial information as it is only 53 percent and 37 percent out of entire countries. Therefore, educational factors are only found to be partially positively correlated with vaccination rates.

F. Access to Information

The 2019 mobile subscription rate and individuals using the Internet indicators were picked from the data set to determine if access to information factors strongly affect the vaccination rate. All countries were included in this analysis as variables to observe the trend of correlation among the entire data set. The individuals using the Internet indicator refers to the percentage of the population using the Internet, while the mobile subscription rate is the actual population subscribed to mobile devices. This value has been divided by the population for each country to obtain the rate of mobile users as a percentage.

Vaccination rate VS	Correlation(%)
Mobile Subscription Rate	49.06
Internet Use Rate	81.89

TABLE XI
PEARSON CORRELATION COEFFICIENTS OF VACCINATION RATE AND ACCESS TO INFORMATION FACTORS

The correlation coefficient in table XI displays how these two factors relate to vaccination. The degree of correlation between mobile subscription and vaccination rates is moderate and not as strong as to convince their relationship. However, the positive correlation between internet use rate and vaccination is much stronger of a relationship. Figure 11 illustrates the trend lines for the relationship between those two indicators and the vaccination rate. The linear regression line for the mobile subscription rate is insufficient. On the other hand, the line in the internet using rate is a clear and positive relationship. Although one of the indicators to determine the access to information gives a high correlation with the vaccination rate, the other indicator does not provide enough evidence to conclude that a higher rate of people's access to information leads to a higher vaccination rate. Thus, access to information indicators is not wholly correlated with vaccination rate.

VI. CONCLUSIONS

As COVID-19 vaccinations are being completed around the globe in early 2022, it is interesting to notice different country indicators are correlated with these vaccination percentages of populations. Through the analytical processes of k-means clustering Pearson/Spearman correlation tables, it is clear that the vaccination rates are dependant on some of these indicators. Prior to the analysis, it was assumed that GDP per capita, population, governance, health, education, and access to information indicators would be strong predictors of vaccination rates within a country. However, not all of these indicators displayed these relationships.

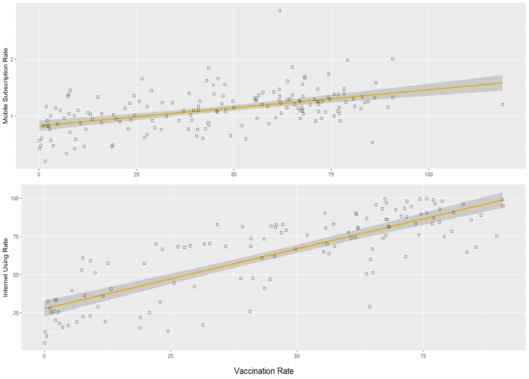

Fig. 11. Linear regression tables of two indicators and vaccinations per hundred in January 2022 (top to bottom): Mobile Subscription Rate,Individuals use the Internet Rate

GDP per Capita and governance were two indicator categories that resulted in being strong predictors of vaccination rates, with GDP per Capita being the most reliable and reoccurring factor within this analysis. The most vaccinated countries were observed to be top GDP per countries, with the opposite effect in the low GDP per capita countries. Governance factors were observed to have an impact on vaccination rates on multiple occasions through this analysis, however investigating these indicators within European countries displays the most positive results. Within Europe, governments that provide safe environment for their citizens and allow them to maintain control over their own governments resulted in higher vaccination rates.

Unlike prior assumptions, population, health, education, and access to information indicators were not found to highly correlate with vaccination rate. Population displayed very little correlations within the top and bottom 30 populated countries, where other economic and governance indicators were found to be correlated with these rates. Health indicators were found to be correlated within select occurrences throughout the analysis, however not enough was done with these indicators to ensure their validity as being associated with vaccination rates. Education and access to information did not have correlations strong and frequent enough to support their association with vaccination rate across the entire set.

Based on the findings in this study, the rate of which a country was vaccinated to COVID-19 was due to the wealth and government of a country. This is supported from related literature, as wealthier countries were able to acquire vaccinations readily and distribute more efficiently than countries with a lesser wealth. Also, a people's trust in their government and government structure directly impacts how willing individuals are to being vaccinated to this new disease. The ideal national environment for a population to become vaccinated is a wealthy one with a modernized government.

References

[1] W. H. Organization. "Coronavirus disease (covid-19)." (), [Online]. Available: https://www.who.int/health-topics/coronavirus. (accessed: 03.15.2022).

[2] G. et. al., "A novel coronavirus outbreak of global health concern," *The Lancet*, vol. 395, pp. 470–473, 10223 2020. DOI: https://doi.org/10.1016/S0140-6736(20)30185-9.

[3] F. International Labour Organization, F. Agriculture Organization, and W. H. O. Agriculture Organization. "Impact of covid-19 on people's livelihoods, their health and our food systems." (2020), [Online]. Available: https://www.who.int/news/item/13-10-2020-impact-of-covid-19-on-people's-livelihoods-their-health-and-our-food-systems. (accessed: 03.15.2022).

[4] T. J. B. Chad P. Bown, "How covid-19 vaccine supply chains emerged in the midst of a pandemic," *The World Economy*, vol. 45, pp. 468–522, 2 2022. DOI: https://doi.org/10.1111/twec.13183.

[5] M. Sallam, "Covid-19 vaccine hesitancy worldwide: A concise systematic review of vaccine acceptance rates," *Vaccines*, vol. 9, pp. 468–522, 2 2021. DOI: https://doi.org/10.1111/twec.13183.

[6] RStudio Team, *Rstudio: Integrated development environment for r*, RStudio, PBC., Boston, MA, 2021. [Online]. Available: http://www.rstudio.com/.

[7] K. et. al., "A vaccination simulator for covid-19: Effective and sterilizing immunization cases," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 4317–4327,

12 2021. DOI: https://doi.org/10.1109/JBHI.2021.3114180.

[8] A. et. al., "Coronavirus pandemic (covid-19)," *Our World in Data*, 2020, https://ourworldindata.org/coronavirus.

[9] W. Bank. "Data catalog." (), [Online]. Available: https://datacatalog.worldbank.org/home. (accessed: 03.15.2022).

[10] ——, "World development indicators." (), [Online]. Available: https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators. (accessed: 02.01.2022).

[11] ——, "Environment, social and governance data." (), [Online]. Available: https://datacatalog.worldbank.org/search/dataset/0037651/Environment--Social-and-Governance-Data. (accessed: 02.01.2022).

[12] ——, "Health nutrition and population statistics." (), [Online]. Available: https://datacatalog.worldbank.org/search/dataset/0037652/Health-Nutrition-and-Population-Statistics. (accessed: 02.01.2022).

[13] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN: 978-3-319-24277-4. [Online]. Available: https://ggplot2.tidyverse.org.

[14] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, pp. 63–67, 2010. DOI: 10.1109/IITSI.2010.74..

[15] A. K. Pandey. "A simple explanation of k-means clustering." (2020), [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/. (accessed: 02.25.2022).

[16] R. L. Thorndike. "Who belongs in the family?" (1953).

[17] S. Solutions. "What is linear regression." (2013), [Online]. Available: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/. (accessed: 04.20.2022).

[18] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[19] J. H. Zar, "Spearman rank correlation," *Encyclopedia of Biostatistics*, vol. 7, 2005.

[20] C. for Disease Control and Prevention. "Tuberculosis: General information." (), [Online]. Available: https://www.cdc.gov/tb/publications/factsheets/general/tb.htm#:~:text=Tuberculosis%5C%20(TB)%5C%20is%5C%20a%5C%20disease,they%5C%20do%5C%20not%5C%20get%5C%20treatment. (accessed: 04.19.2022).