Project - COVID-19 Vaccination Progress

16 April, 2021

Introduction

With the aim of immunizing everyone with a vaccine, COVID-19 vaccination process has now started and progressed in many countries all around the world. There are some vaccines which are much more used than other vaccines in different countries. Some countries are very advanced in the process of vaccinating major percentage of its country population. The economic state is also different in every country due to this global pandemic. In this project, we are using the COVID-19 vaccination data of the world and web-scraping techniques to find answers to some of our key analysis questions related to COVID-19 vaccination progress around the world -

(1) Which vaccine is mostly used in countries around the world?
(2) Which countries have the highest daily average vaccinations?
(3) Where are more people vaccinated per day?
(4) What is the proportion of fully vaccinated people from entire population of a country?
(5) Is the association between country GDP and total vaccinations statistically significant?

The purpose of this project is to learn the overall progress of vaccinations in different countries all around the world by analyzing these questions above.

Data and Methods

Primary data source

The primary data source of our project is the COVID-19 Vaccination data from Kaggle. Link - https://www.kaggle.com/gpreda/covid-world-vaccination-progress?select=country_vaccinations.csv

This dataset (country vaccinations) contains information about Country, Country ISO Code, Date, Total number of vaccinations, Total number of people vaccinated, Total number of people fully vaccinated, Daily vaccinations (raw), Daily vaccinations, Total vaccinations per hundred, Total number of people vaccinated per hundred, Total number of people fully vaccinated per hundred, Number of vaccinations per day, Daily vaccinations per million, Vaccines used in the country, Source name, Source website.

Secondary data sources

The secondary sources of data for our project contains information about country population and GDP. Since both these data are absent in our primary data source, we collected these data using web-scraping techniques from Worldometer website.

(1) Country Population - https://www.worldometers.info/world-population/population-by-country
(2) Country GDP - https://www.worldometers.info/gdp/gdp-by-country/

Methods

For our first three analysis questions, we used the primary data source, data visualization principles and summarized our results in bar charts to answer these questions as accurately as possible.

The last two analysis questions are answered using web-scraping techniques which required fair amount of data cleaning after scraping these data from our secondary data sources. Linear regression is also used to answer the last analysis question.

Data Analysis Results

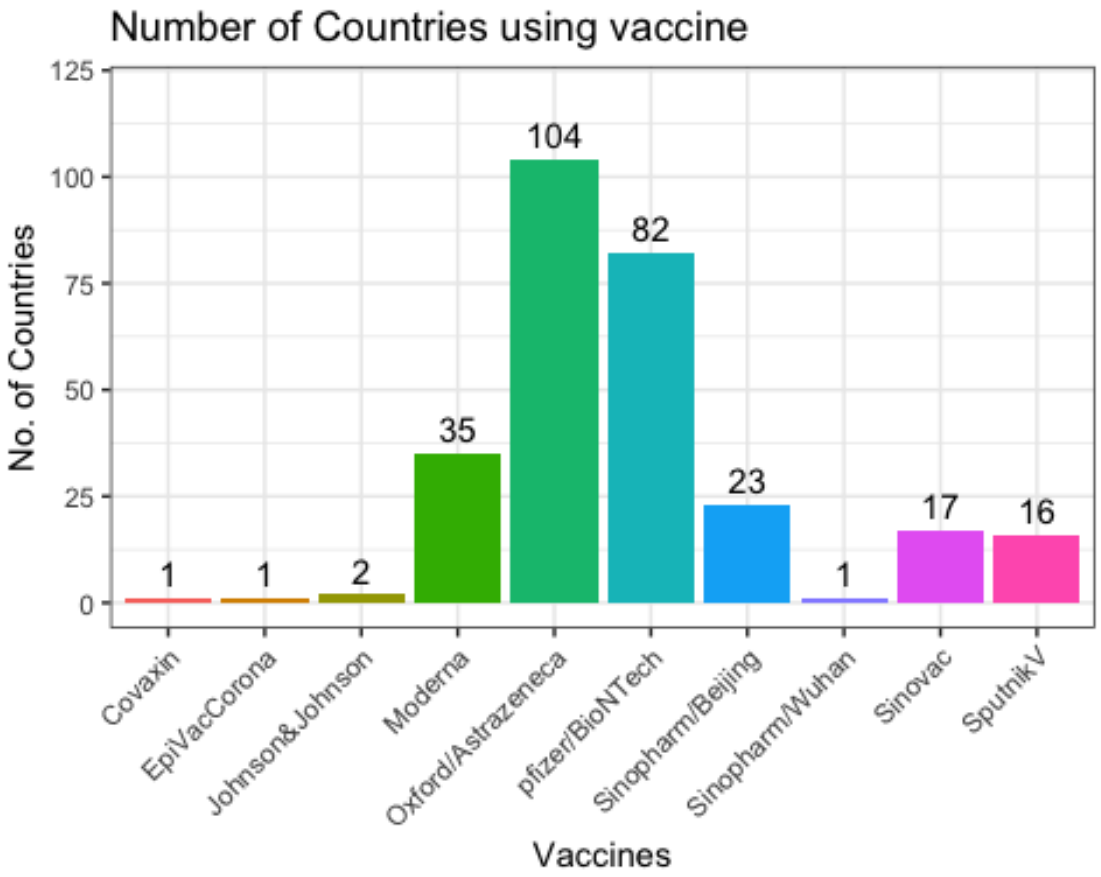
In this section, we will analyze each of our analysis question and present the results of our data analysis separately. We will be using data visualization, web-scraping techniques and linear regression depending on what analysis question we are answering.

Question 1 - Which vaccine is mostly used in countries around the world?

For answering this question, firstly we need to find how many vaccines are in there in this dataset. After exploring the dataset, we found that these are the 10 unique vaccines which are being used by countries all over the world.

```
##  [1] "Oxford/AstraZeneca" "Pfizer/BioNTech"    "SputnikV"
##  [4] "Moderna"            "Sinovac"            "Sinopharm/Beijing"
##  [7] "Covaxin"            "EpiVacCorona"       "Johnson&Johnson"
## [10] "Sinopharm/Wuhan"
```

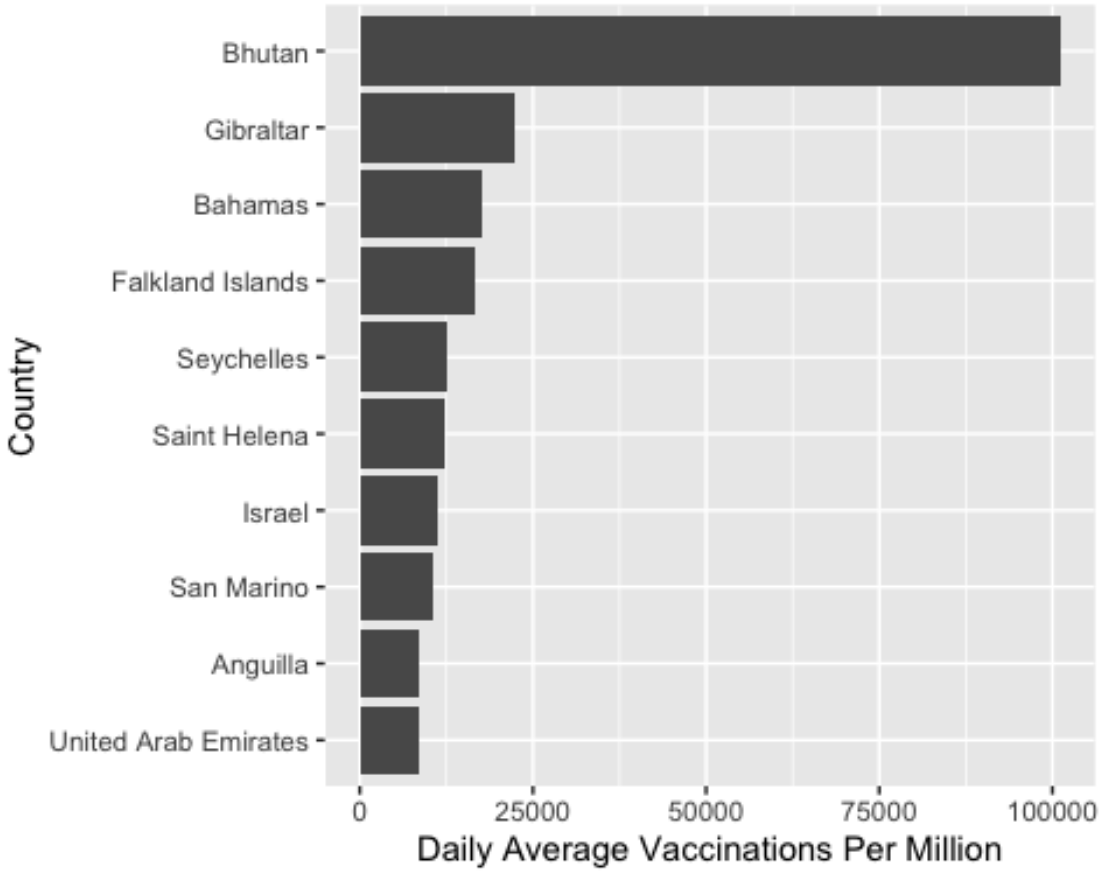
Now, we find the number of countries using the above vaccines and visualize our results in a bar plot.



From the above analysis and data visualization, we can see that the most used vaccine around the world is Oxford/Astrazeneca followed by pfizer/BioNTech. A total of 99 countries using Oxford/Astrazenca vaccine and a total of 81 countries are using the pfizer/BioNTech vaccine.

Question 2 - Which countries have the highest daily average vaccination rate?

For this analysis question, we need to find the mean of `daily_vaccinations_per_million` column grouped by country and find the top 10 countries with highest rate of daily average vaccination per million.



As we can see above, Bhutan has the highest rate of daily average vaccinations per million among the list of top 10 countries.

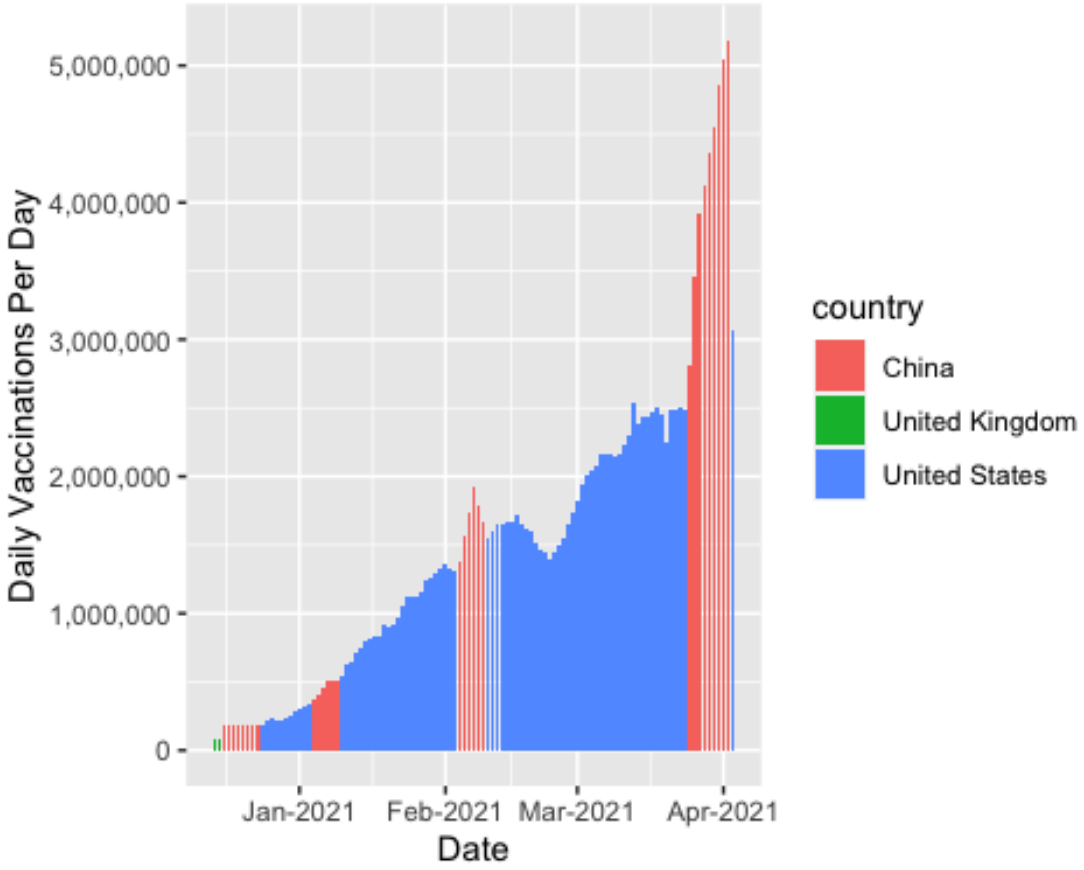
Question 3 - Where are more people vaccinated per day?

For answering this analysis question, firstly we need to explore our actual dataset and find how many dates are there for the `daily_vaccination` column. Here, we found that there are 108 distinct dates with valid data(daily_vaccinations). This means we should have 108 rows(dates) in our final dataframe. We can use this dataframe to find out the countries with the highest number of people vaccinated per day.

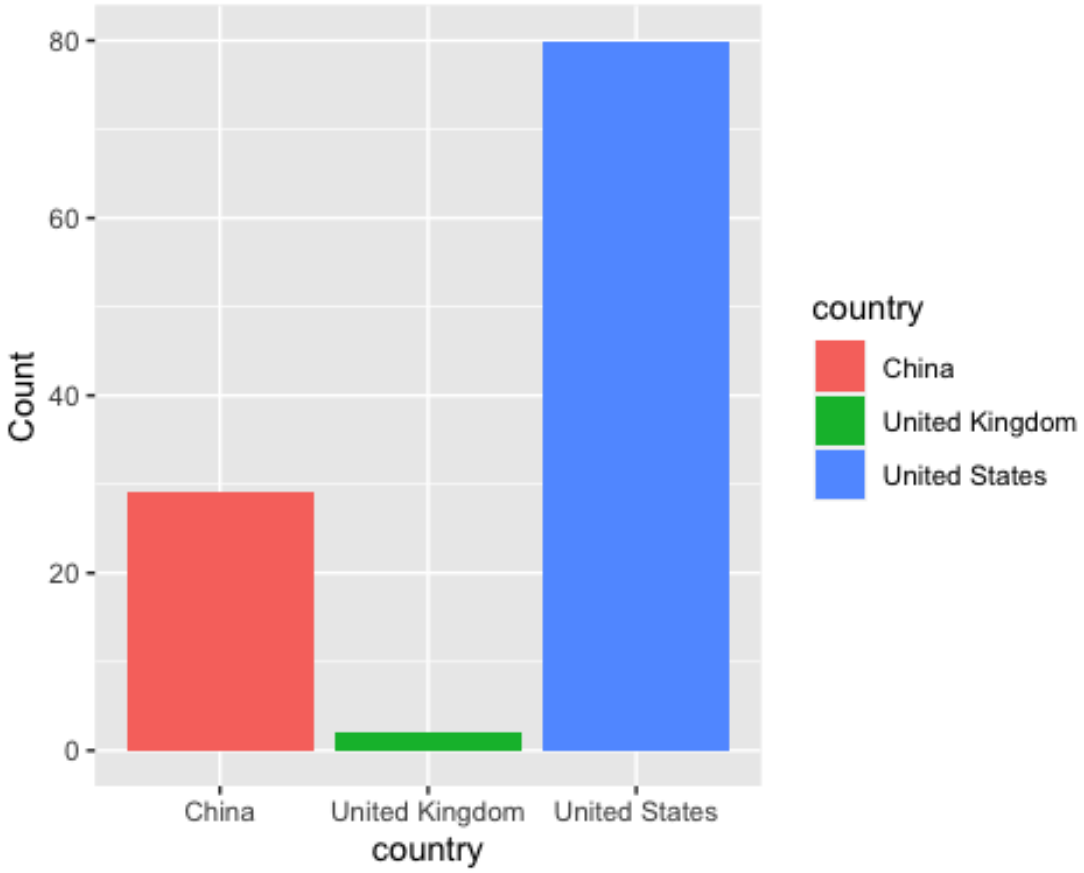
```
## [1] 111

## Rows: 111
## Columns: 3
## $ date                <chr> "2020-12-14", "2020-12-15", "2020-12-16",
"2020-12-…
## $ country             <chr> "United Kingdom", "United Kingdom", "China",
"China…
## $ daily_vaccinations <dbl> 84117, 84117, 187500, 187500, 187500, 187500,
18750…
```

As we can see above, it shows the name of countries with the highest number of people being vaccinated per day from our chosen dataset.

Now, we can visualize the above data in a bar-chart to show the countires with the highest number of vaccinated people per day.




```
## # A tibble: 3 × 2
## # Groups:   country [3]
##   country            n
##   <chr>          <int>
## 1 China             29
## 2 United Kingdom     2
## 3 United States     80
```



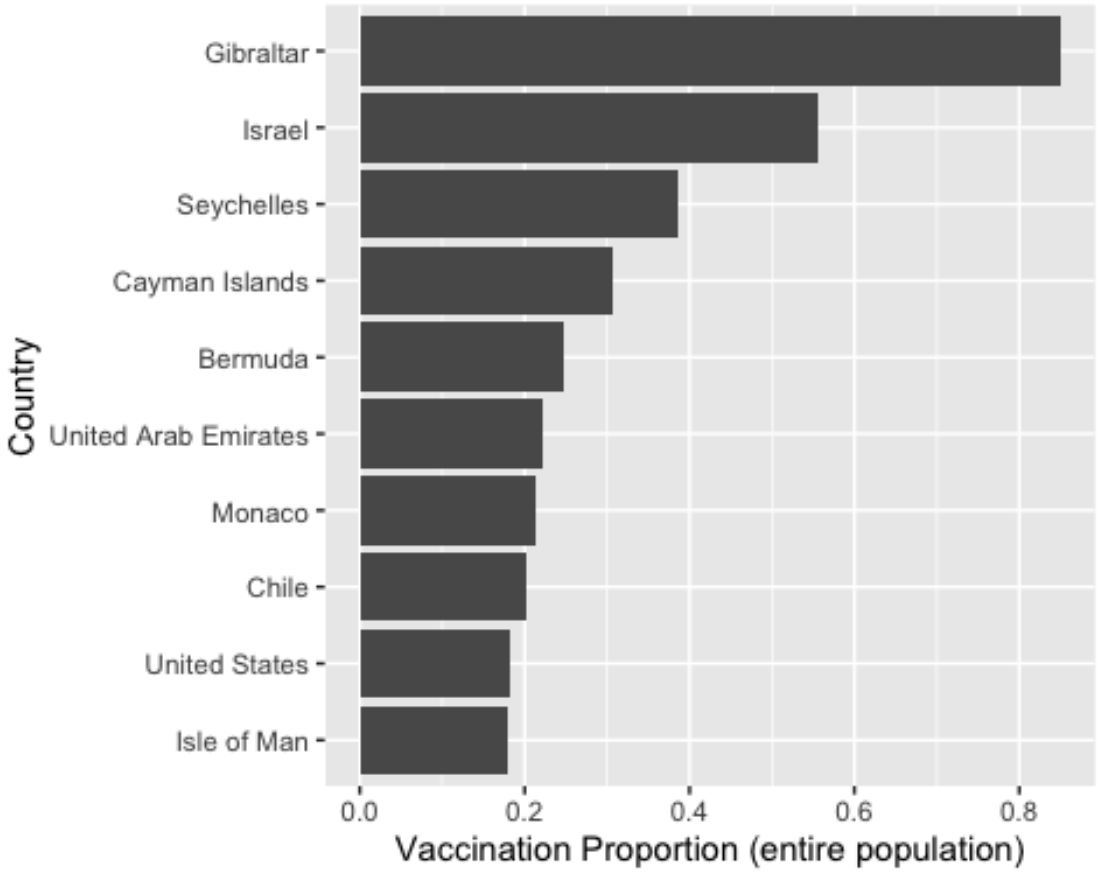
As we can see from the chart above, only 3 countries have the highest number of vaccinated people on 108 different dates. China has appeared 27 times which means that China has the highest number of vaccinated people on 27 different dates. We can also say that United Kingdom has appeared only 2 times which is the lowest. On the other hand, United States has the highest number of vaccinated people on 79 different dates which is the highest among the countries in our chosen data-set.

Question 4 - What is the proportion of fully vaccinated people from entire population of a country?

In this analysis question, we will be using web-scraping to get country population data from worldometer website as described above in data and methods section.

Firstly, we get the data of fully vaccinated people from our primary data source. Then, we use web-scraping to get data of country and it's population from worldometer website. Since we don't need any other info after web-scraping this table, we clean up our data to only keep information about country and population. Since population column has commas and character datatype, we remove the commas from the column and convert it to a numeric datatype.

Secondly, we inner-join our data from primary source and secondary source by country. Then, we can visualize top 10 countries with highest proportion of fully vaccinated people out of their entire population in the following way.



As we can see above from this data visualization, we have a list of top 10 countries of highest proportion of fully vaccinated people out of their entire population. Gibraltar is leading on top by having highest proportion of fully vaccinated people out of its entire population whereas United States is at the bottom of the list.

Question 5 - Is the association between country GDP and total vaccinations statistically significant?

In this analysis question, we will be using web-scraping to get country GDP data from worldometer website as described above in data and methods section.

Firstly, we get the data of total vaccinations of each country from our primary data source. Then, we use web-scraping to get data of country and it's GDP from worldometer website. Since we don't need any other info after web-scraping this table, we clean up our data to only keep information about country and GDP. Since GDP column has commas and $ signs, we remove them from the column and make it to a numeric datatype from character datatype for further analysis later.

Secondly, we inner-join our data from primary source and secondary source by country. Since the values of GDP and total vaccinations are very large, we also need to transform the variables GDP and total vaccination using a log transformation.

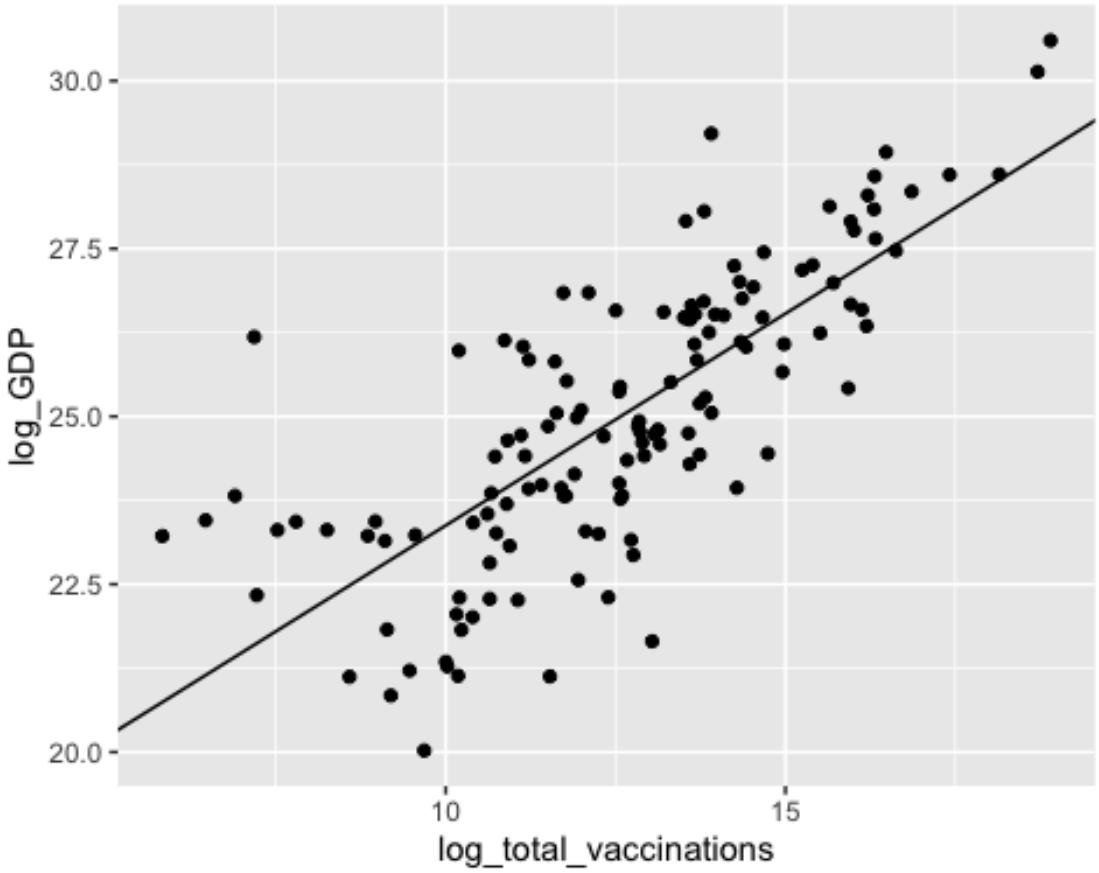
Now, we can fit a linear model and find the confidence interval to determine if the association between log_GDP and log_total_vaccinations statiscally significant.

```
##
## Call:
## lm(formula = log_GDP ~ log_total_vaccinations, data =
vaccination_summary2_log)
##
## Coefficients:
##            (Intercept)  log_total_vaccinations
##                17.0640                  0.6312

##                             2.5 %     97.5 %
## (Intercept)            15.884726 18.2432248
## log_total_vaccinations  0.539404  0.7230014
```

To determine whether this difference is statistically significant, we can look at 95% confidence intervals for the regression coefficients.

Because 0, which corresponds to no difference in average value, is not in the confidence interval for the regression coefficient log_total_vaccinations, we conclude that the association between log_GDP and log_total_vaccinations is statistically significant.

The following scatterplot shows the relationship between log_total_vaccinations and log_GDP and the linear line provides a good description of the dataset.



Conclusion

In summary, this data analysis project gave us the chance to explore and understand the overall progress of COVID-19 vaccinations all around the world. We learned about the most used vaccines in different countries. Additionally, we learned about the vaccination progress of different countries in different aspects. Lastly, we investigated to find association between country GDP and their total vaccinations count using linear regression model. In short, we are able to achieve the real purpose of our project which we originally intended in the beginning through extensive data analysis of all of our analysis questions.

Project Tweet (280 characters)

The most used vaccine around the world is Oxford/Astrazeneca currently. Bhutan has the highest rate of daily average vaccinations per million. People are getting vaccinated more daily in USA. The association between country GDP and total vaccinations is statistically significant.

Appendix

Analysis Question 1 Code

```
library(tidyverse)
covid<-read.csv("country_vaccinations.csv")

data<-covid%>%
  separate(col="vaccines",into=c("vaccine1","vaccine2","vaccine3"),sep=",")

data1<-unique(data$vaccine1)
data2<-unique(data$vaccine2)
data3<-unique(data$vaccine3)
data4<-c(data1,data2,data3)

#data4
data4<-data4%>%str_replace_all(" ","")
data5<-unique(data4)
data5<-data5[!is.na(data5)]
data5

##  [1] "Oxford/AstraZeneca" "Pfizer/BioNTech"    "SputnikV"
##  [4] "Moderna"            "Sinovac"            "Sinopharm/Beijing"
##  [7] "Covaxin"            "EpiVacCorona"       "Johnson&Johnson"
## [10] "Sinopharm/Wuhan"

library(tidyverse)
library(janitor)
library(stringr)

#Oxford/AstraZeneca
data1<-data%>%
  filter(vaccine1=="Oxford/AstraZeneca"|vaccine2==" Oxford/AstraZeneca"
         |vaccine3==" Oxford/AstraZeneca")
data1<-unique(data1$country)
data1<-as.data.frame(data1)

#Pfizer/BioNTech
data2<-data%>%
  filter(vaccine1=="Pfizer/BioNTech"|vaccine2==" Pfizer/BioNTech"
         |vaccine3==" Pfizer/BioNTech")
data2<-unique(data2$country)
data2<-as.data.frame(data2)
names(data2)[names(data2) == "data2"] <- "Pfizer/BioNTech"

#Sputnik
sputnik<-data%>%
  filter(vaccine1=="Sputnik V"|vaccine2==" Sputnik V"|vaccine3==" Sputnik V")
sputnik<-unique(sputnik$country)
sputnik<-as.data.frame(sputnik)
```

```r
#Moderna
moderna<-data%>%
  filter(vaccine1=="Moderna"|vaccine2==" Moderna"|vaccine3==" Moderna")
moderna<-unique(moderna$country)
moderna<-as.data.frame(moderna)

#Sinovac
sinovac<-data%>%
  filter(vaccine1=="Sinovac"|vaccine2==" Sinovac"|vaccine3==" Sinovac")
sinovac<-unique(sinovac$country)
sinovac<-as.data.frame(sinovac)

#Sinobei
sinoBei<-data%>%
  filter(vaccine1=="Sinopharm/Beijing"|vaccine2==" Sinopharm/Beijing"
         |vaccine3==" Sinopharm/Beijing")
sinoBei<-unique(sinoBei$country)
sinoBei<-as.data.frame(sinoBei)

#Covaxin
covaxin<-data%>%
  filter(vaccine1=="Covaxin"|vaccine2==" Covaxin"|vaccine3==" Covaxin")
covaxin<-unique(covaxin$country)
covaxin<-as.data.frame(covaxin)

#EpivacCorona
epi<-data%>%
  filter(vaccine1=="EpiVacCorona"|vaccine2==" EpiVacCorona"
         |vaccine3==" EpiVacCorona")
epi<-unique(epi$country)
epi<-as.data.frame(epi)

#Johnson&Johnson
john<-data%>%
  filter(vaccine1=="Johnson&Johnson"|vaccine2==" Johnson&Johnson"
         |vaccine3==" Johnson&Johnson")
john<-unique(john$country)
john<-as.data.frame(john)

#Sinopharm/Wuhan
sinoWuhan<-data%>%
  filter(vaccine1=="Sinopharm/Wuhan"|vaccine2=="Sinopharm/Wuhan"
         |vaccine3=="Sinopharm/Wuhan")
sinoWuhan<-unique(sinoWuhan$country)
sinoWuhan<-as.data.frame(sinoWuhan)

#Vaccine data
vaccines<-c("Oxford/Astrazeneca","pfizer/BioNTech","SputnikV",
```

```
            "Moderna","Sinovac","Sinopharm/Beijing","Covaxin",
            "EpiVacCorona","Johnson&Johnson","Sinopharm/Wuhan")
num_countries<-c(nrow(data1),nrow(data2),nrow(sputnik),nrow(moderna),
                 nrow(sinovac),nrow(sinoBei),nrow(covaxin),nrow(epi),
                 nrow(john),nrow(sinoWuhan))
df<-data.frame(vaccines,num_countries)

#Data visualization
country_vaccine <- ggplot(mapping=aes(x=vaccines, y=num_countries,
                                      fill = vaccines))+
geom_col() +
labs(x = "Vaccines", y = "No. of Countries",
     title  = "Number of Countries using vaccine")+
geom_text(aes(label =num_countries ), vjust=-0.5)+
theme_bw()+
theme(axis.text.x = element_text(angle = 45, hjust = 1),
      legend.position = "None") +
expand_limits(y = 120)
```

Analysis Question 2 Code

```
vac_rate<-covid%>%
  group_by(country)%>%
  summarise(mean=mean(daily_vaccinations_per_million,na.rm=TRUE))

top_ten<-vac_rate%>%
  top_n(n=10,wt=mean)

daily_avg_vaccinations <-top_ten%>%
  ggplot(aes(x = reorder(country, mean), y = mean)) +
  geom_bar(stat="identity") +
  coord_flip() +
  labs(x = "Country", y="Daily Average Vaccinations")
```

Analysis Question 3 Code

```
#Find how many dates are there in actual dataset for the daily_vaccination
column
library(tidyverse)

data <- read.csv("country_vaccinations.csv")

total_dates_with_data <- data %>%
  group_by(date) %>%
  filter(!is.na(daily_vaccinations)) %>%
  distinct(date) %>%
  nrow()

total_dates_with_data

## [1] 111
```

```r
#Find the countries with highest number of people vaccinated per day.
total_vaccinations <- data %>%
  group_by(date) %>%
  filter(!is.na(daily_vaccinations)) %>%
  filter(daily_vaccinations == max(daily_vaccinations)) %>%
  summarise(country, daily_vaccinations) %>%
  distinct()

glimpse(total_vaccinations)
```

```
## Rows: 111
## Columns: 3
## $ date               <chr> "2020-12-14", "2020-12-15", "2020-12-16",
"2020-12-…
## $ country            <chr> "United Kingdom", "United Kingdom", "China",
"China…
## $ daily_vaccinations <dbl> 84117, 84117, 187500, 187500, 187500, 187500,
18750…
```

```r
library(dslabs)
library(scales)

#bar-chart showing the countires with the highest number of
#vaccinated people per day.
total_vaccinations$date <- as.Date(total_vaccinations$date)

plot<- ggplot(total_vaccinations,
       aes( date, daily_vaccinations, fill = country)) +
  geom_bar(stat="identity") +
  scale_x_date(labels = date_format("%b-%Y"))+
  scale_y_continuous(labels = comma) +
  labs(x = "Date", y="Daily Vaccinations Per Day")

#How many times does a country appear on the above plot
#showing highest number of vaccinated people per day?
max_vaccinated_per_day<- total_vaccinations %>%
  group_by(country) %>%
  count(country)

max_vaccinated_per_day
```

```
## # A tibble: 3 × 2
## # Groups:   country [3]
##   country            n
##   <chr>          <int>
## 1 China             29
## 2 United Kingdom     2
## 3 United States     80
```

```r
plot2 <- max_vaccinated_per_day %>%
  ggplot(aes(country, n, fill = country)) +
```

```
  geom_col()+
  labs(y = "Count")
```

Analysis Question 4 Code

```
library(rvest)
library(tidyverse)
library(purrr)
library(tidytext)

#Get totatl vaccinations done in each country from vaccine dataset
#Dataset source: https://www.kaggle.com/gpreda/covid-world-vaccination-
progress
vaccine_dataset <- read_csv("country_vaccinations.csv")

total_vaccinations <- vaccine_dataset %>%
  group_by(country) %>%
  filter(!is.na(people_fully_vaccinated)) %>%
  mutate(people_fully_vaccinated =
           max(people_fully_vaccinated, na.rm = TRUE))%>%
  summarise(country, people_fully_vaccinated) %>%
  distinct()

#Web scraping to collect population of each country in 2020
url_population <-
  "https://www.worldometers.info/world-population/population-by-country"

resource <- read_html(url_population)

country_populations <- resource %>%
  html_node("table") %>%
  html_table() %>%
  rename(
    country = `Country (or dependency)`,
    population = `Population (2020)`
  ) %>%
  select(country, population)

#Clean data - remove commas from population column,
#then convert datatype from chr to numeric
country_populations$population <-
  as.numeric(gsub(",","",country_populations$population))

#Inner-Join total vaccinations data and country populations by country
#proportion of fully vaccinated people out of
#entire population - data visualization of top 10 countries
vaccination_summary1 <- inner_join(total_vaccinations,
                         country_populations,
                       by = "country") %>%
  summarise(vaccination_proportion = people_fully_vaccinated/population) %>%
  top_n(n = 10) %>%
  ggplot(aes(x = reorder(country, vaccination_proportion),
             y = vaccination_proportion)) +
```

```
  geom_bar(stat="identity") +
  coord_flip() +
  xlab("Country")
```

Analysis Question 5 Code

```
#Get total vaccinations in countries
vaccinations <- vaccine_dataset %>%
  group_by(country) %>%
  filter(!is.na(total_vaccinations)) %>%
  mutate(total_vaccinations = max(total_vaccinations, na.rm = TRUE)) %>%
  summarise(country, total_vaccinations) %>%
  distinct()

#Web scraping to collect GDP of each country
url_gdp <- "https://www.worldometers.info/gdp/gdp-by-country/"

resource_gdp <- read_html(url_gdp)

country_gdp <- resource_gdp %>%
  html_node("table") %>%
  html_table() %>%
  rename(
    country = `Country`,
    GDP = `GDP (nominal, 2017)`
  ) %>%
  select(country, GDP)

#data cleaning - remove comma and $ from GDP
country_gdp$GDP <- gsub(",","",country_gdp$GDP)
country_gdp$GDP <- as.numeric(gsub("\\$", "", country_gdp$GDP))

#Inner-Join vaccinations data and country gdp by country
#Data analysis to find if there is any corelation
#between gdp and total vaccinations
vaccination_summary2 <- inner_join(vaccinations,
                                   country_gdp,
                                   by = "country")

vaccination_summary2_log <- mutate(vaccination_summary2,
                                    log_GDP = log(GDP),
                                    log_total_vaccinations =
                                      log(total_vaccinations))

#Fit a linear model and find confidence interval
fit <- lm(log_GDP ~ log_total_vaccinations, data = vaccination_summary2_log)

fit
```

```
##
## Call:
## lm(formula = log_GDP ~ log_total_vaccinations, data =
vaccination_summary2_log)
##
## Coefficients:
##            (Intercept)  log_total_vaccinations
##                17.0640                  0.6312

confint(fit)

##                             2.5 %     97.5 %
## (Intercept)            15.884726 18.2432248
## log_total_vaccinations  0.539404  0.7230014

#Scatterplot for showing relationship between log_total_vaccinations and
log_GDP
scatter_plot <- ggplot(vaccination_summary2_log, aes(x =
log_total_vaccinations,
                       y = log_GDP)) +
  geom_point() +
  geom_abline(intercept = coef(fit)[1],
              slope = coef(fit)[2])
```