Hemant Sathian

Professor Imielinski

Data 101 01:198:142 Section 09

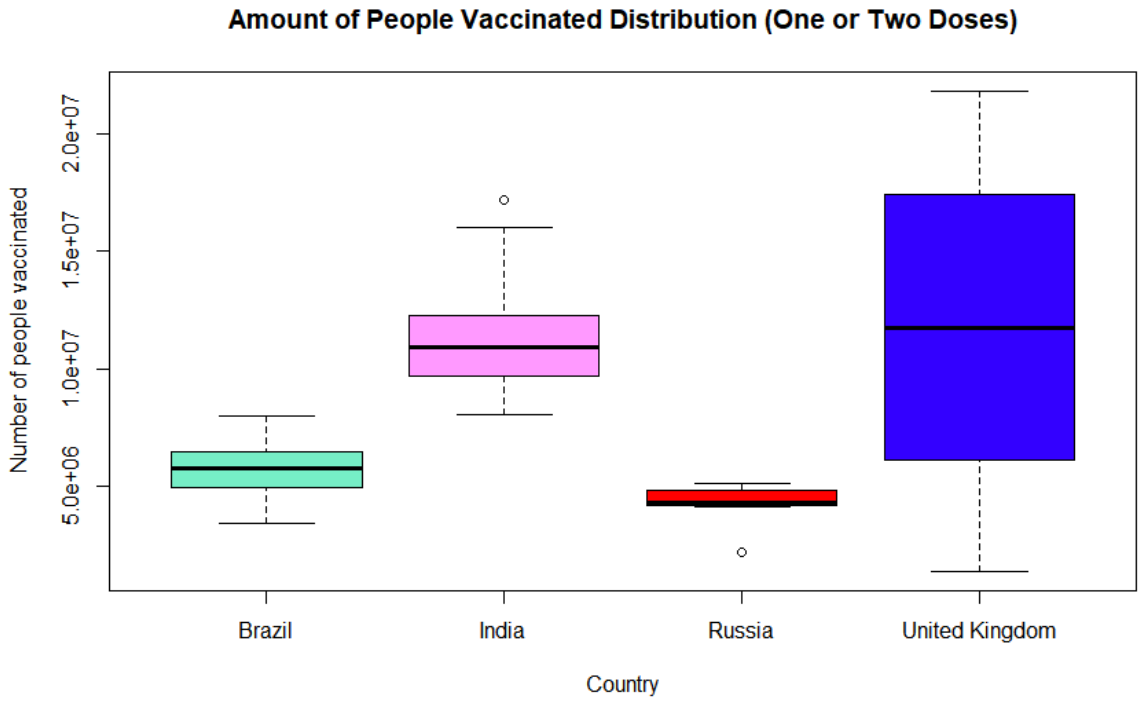
17 March 2021

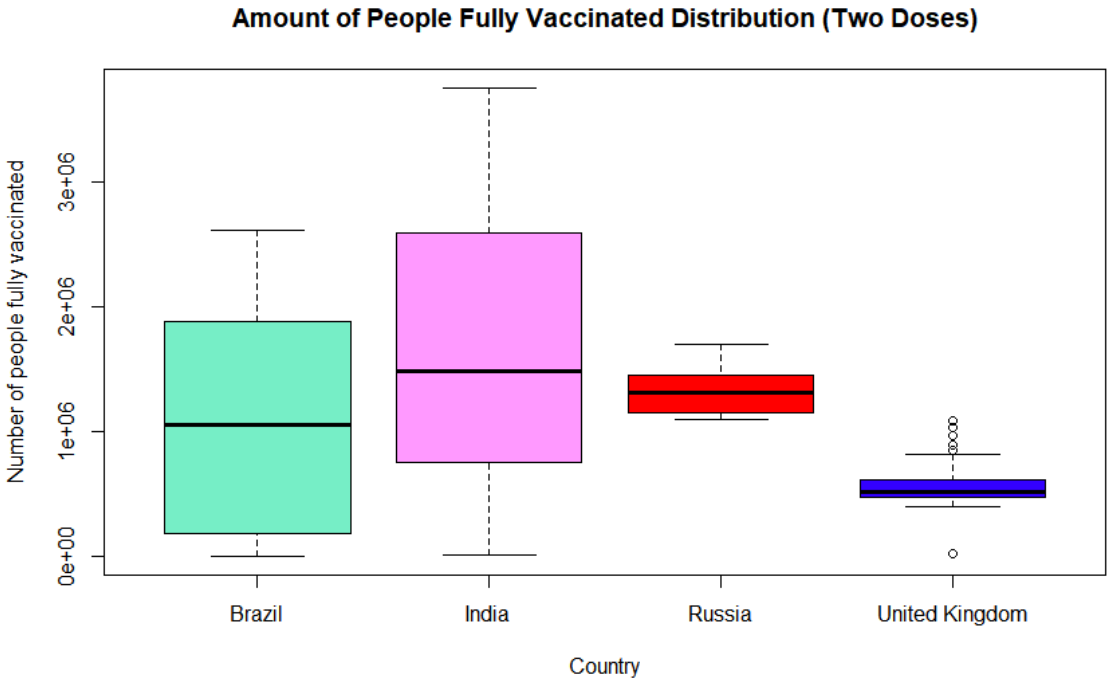
# Exploring COVID-19 World Vaccination Progress Using Data Science

Hello everyone,

My name is Hemant Sathian. I'm currently a junior studying Information Technology and Statistics at Rutgers University. During my freshman and sophomore years at Rutgers, I attended in-person lectures and lived on campus. Unfortunately, the covid-19 pandemic has pushed my junior year experience to an online format. As cases decline worldwide due to quarantine measures and the onset of vaccines, a return to normalcy may soon arrive. I was fortunate enough to have received my first dose of the covid-19 vaccine on February 19th and second dose on March 15th. As I type this blog post, I'm still feeling side-effects from the Pfizer vaccine's second dose. Having completed my vaccination, I was curious about the progress of covid-19 vaccinations worldwide. Despite the relative infancy of the covid-19 vaccination effort, I was able to find a relatively clean dataset on Kaggle.com by a reputable data scientist named Gabriel Preda.

This dataset consists of 5,321 rows of data, each accounting for the vaccination effort of different countries worldwide. Updated daily, additional data is added to the dataset from the "Our World in Data" Github repository for covid-19. This dataset can be considered trustable since each row of data in the dataset contains a source name and source website column. Each row of data also specifies information such as country name, total number of people vaccinated, total number of people "fully-vaccinated" (having received two doses), and the types of vaccines used. It is important to note that some of the data columns had missing values, represented by "NA". In order to properly represent and analyze the data, I had to use the **na.omit()** function to ignore columns of data with these values. Having observed the dataset in my RStudio, I made the decision to focus my research on countries with the highest covid-19 cases. Using information provided by the World Health Organization (WHO), I narrowed my research to these four countries: the United Kingdom, Brazil, India, and Russia. From here, I would examine the distribution between the 4 countries in terms of people who had been vaccinated (received either one or two doses of the vaccine) and people who had been fully vaccinated (received two doses of the vaccine).




Observing the first boxplot, I noticed the United Kingdom had the highest mean amount of people vaccinated (having received one or two doses) in a given data entry. When viewing the second boxplot, which describes the amount of people who had been fully vaccinated (having

received two doses), India had the largest mean amount. The difference in these boxplots was fascinating to me and could lead one to assume the United Kingdom had been the most successful in vaccinating their citizens. It could also lead one to assume India had been the most successful in fully vaccinating their citizens. To truly confirm these patterns, I would test for significance using two one-tailed tests.

My first one-tailed test consisted of comparing the mean amounts of people vaccinated (one or two doses) between the United Kingdom and India. My **null hypothesis** considered no difference between the mean amounts of people vaccinated between the United Kingdom and India. My **alternative hypothesis** considered the United Kingdom having vaccinated a higher mean amount of people than India. In my analysis, I discovered the mean amount of people vaccinated in the United Kingdom was 11,647,772 people while in India the mean amount was 11,340,410 people. Not a significant difference, but I would need to analyze further using a permutation test. The result of my permutation test (which returns a p-value for a one-sided test) was a ***p*-value of 0.32058**. This was conducted using 100,000 permutations to ensure accuracy. Since my observed *p*-value of 0.32058 is significantly higher than the significance level, which is a *p*-value of 0.05, I failed to reject my null hypothesis.

My second one-tailed test consisted of comparing the mean amounts of people fully vaccinated (two doses) between the United Kingdom and India. My **null hypothesis** was that there is no difference between the mean amounts of people fully vaccinated between the United Kingdom and India. My **alternative hypothesis** was that India has fully vaccinated a higher mean amount of people than the United Kingdom. In my analysis, I discovered the mean amount of people fully vaccinated in the United Kingdom was 561,068 people, while in India, the mean amount was 1,679,384 people. Despite there being a stark difference between the means, I would still need to analyze further using a permutation test. The result of my permutation test was a ***p*-value of 0.04378**. This was conducted using 100,000 permutations to ensure accuracy. Since my observed *p*-value of 0.04378 is lower than the significance level, which is a *p*-value of 0.05, I was able to reject my null hypothesis.

To summarize, I was not able to determine evidence that suggests the United Kingdom has vaccinated (one or two doses) a larger mean amount of people than India. Still, I was able to determine evidence suggesting that India has fully vaccinated (two doses) a more significant mean amount of people than the United Kingdom. From an alternate perspective, one might assume a developing nation such as India would not surpass the United Kingdom in fully vaccinating a larger mean amount of people, but the data analysis proves otherwise. It is through data science that these hidden narratives can be revealed.

*Caution: I had omitted data columns with missing values in my research (as stated prior) in order to properly represent and analyze the data. The omissions I made in the data set could certainly have affected my results. I had also not applied Bonferroni correction, which accounts for the simultaneous testing of the dataset for more than one hypothesis. As there are 125 unique countries in the dataset, 7,750 different hypotheses could be performed. The resulting *p*-value

cutoff to account for these hypotheses would be 0.00000645, much lower than the standard significance level of 0.05, which I had based my results on. A *p*-value of 0.00000645 would lead both of my hypothesis tests to fail to reject the null. Furthermore, the information shown in this blog may not portray an accurate representation of countries' vaccination efforts in the next following months as additional entries are made to the dataset.

Resource: https://www.kaggle.com/gpreda/covid-world-vaccination-progress

Code:

```
modr_russia

modr_united_kingdom <- subset(modr, country == "United Kingdom")

subdf <- subset(modr, country == "Brazil" | country == "India" | country == "Russia" | country == "United Kingdom")

subdf

colors <- c("aquamarine2", "#FF99FF", "#FF0000", "#3300FF", "#99FF00")


#Boxplot for people vaccinated (one or two doses)

boxplot(subdf$people_vaccinated~subdf$country,xlab = 'Country',ylab = 'Number of people vaccinated', main = "Amount of People Vaccinated Distribution (One or Two Doses)",col=colors,border="black")

#Boxplot for people fully vaccinated (two doses)

boxplot(subdf$people_fully_vaccinated~subdf$country,xlab = 'Country',ylab = 'Number of people fully vaccinated', main = "Amount of People Fully Vaccinated Distribution (Two Doses)",col=colors,border="black")


#First one-tailed test comparing means of people vaccinated (one or two doses)


#Sub-setting both samples

modr_india

modr_united_kingdom

#Amount of people vaccinated in India

India_people_vaccinated <- modr_india$people_vaccinated
```

```
#Amount of people vaccinated in the United Kingdom

UK_people_vaccinated <- modr_united_kingdom$people_vaccinated

#Means of two samples

mean.India_people_vaccinated <- mean(India_people_vaccinated)

mean.India_people_vaccinated

mean.UK_people_vaccinated <- mean(UK_people_vaccinated)

mean.UK_people_vaccinated

#Permutation Function Test

PermutationTestSecond::Permutation(modr, "country", "people_vaccinated", 100000
,"India","United Kingdom")


#Second one-tailed test comparing means of people fully vaccinated (two doses)


#Amount of people fully vaccinated in India

India_people_fully_vaccinated <- modr_india$people_fully_vaccinated

#Amount of people fully vaccinated in the United Kingdom

UK_people_fully_vaccinated <- modr_united_kingdom$people_fully_vaccinated

#Means of two samples

mean.India_people_fully_vaccinated <- mean(India_people_fully_vaccinated)

mean.India_people_fully_vaccinated

mean.UK_people_fully_vaccinated <- mean(UK_people_fully_vaccinated)

mean.UK_people_fully_vaccinated

#Permutation Function Test

PermutationTestSecond::Permutation(modr, "country", "people_fully_vaccinated", 100000
,"United Kingdom","India")
```