**Worldwide COVID-19 Vaccination Race**

**Team 4**

Assima Imataliyeva

Adam Yousif

Justin Nguyen

Tay Phan

Ari Weiss

**BSAN 460**

**6/2/2021**

**Table of Contents**

**Executive Summary** 2

**Introduction** 3

**Literature/Industry Review** 4

**Results** 6

1. Data Source & Elements 6
2. Descriptive Analytics 7
3. Regression Tree with Standardization 8
4. Multiple Linear Regression 9
5. Vaccine Classification Adoption 10

**Discussion** 11

1. Business Insights 11
2. Recommendation 12

**Appendix** 13

1. Reference 13
2. Other Visuals and Graphs 13
3. R-Code 14

**Executive Summary**

The emergence of the Covid-19 virus has affected millions of people globally with devastating consequences that changed our lives significantly. The impact of the pandemic is felt differently depending on our social identities and our lifestyles, and it has led to substantial changes in our daily lives. Its rapid spread and its effects have resulted in the use of masks, social distancing, travel restrictions, economic crisis, and loss of millions of people’s lives worldwide. Poorer countries that are already dealing with the existential humanitarian crises have been particularly exposed to the effects of Covid-19. Over the past year, research laboratories and organizations all over the world have been developing and testing vaccines. The prior research and knowledge about the structure of coronaviruses causing SARS and MERS led to the faster development of vaccines. Scientists came up with various alternatives that are being implemented worldwide, and as of April 2021, 16 vaccines have been authorized, and our report will discuss the most common vaccines like Pfizer-BioNTech, Moderna, Sputnik V, and others. Today, all countries are getting their populations vaccinated as fast as possible to get back to safe and normal life with all the pre-covid opportunities and benefits of social life. Different countries are on track to getting fully vaccinated and have achieved different results due to many factors affecting the vaccination progress. In this project, our focus would be directed towards analyzing significant data variables and external factors that influence the current Covid-19 vaccination rate worldwide.

The data we analyzed was collected from more than 100 countries representing various demographic, economic, and health situations to be compared with vaccinations in each country or region. Our approach was to thoroughly identify and combine certain variables to build a practical and sensible set of information to provide a better understanding of the current progress. On top of that, secondary research with updated developments in vaccinations have been conducted to support out data analysis findings that answer the issue true questions regarding the demographic components, factors affecting countries’ selection of vaccines and investments in their early development and supply, and pricing. The methods that were used to compare our results include descriptive analytics, predictive analytics, and regression models that will describe our key findings and interpret them for our client to better understand and predict further vaccination progress in different countries.

**Introduction**

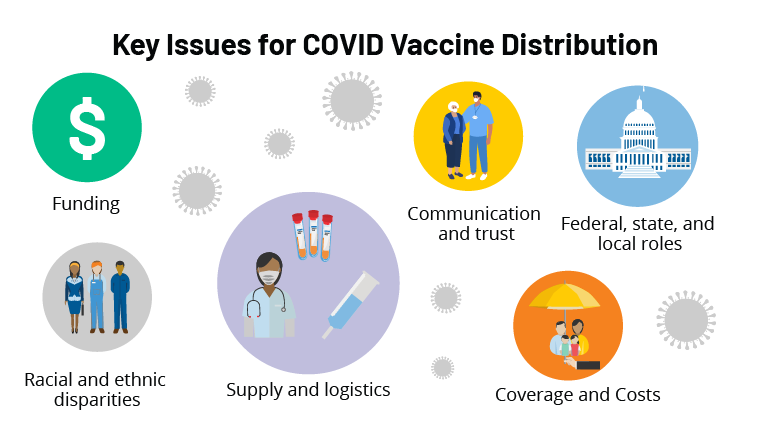
The Covid-19 vaccination process has been one of the most discussed topics for months as more and more people are getting vaccinated globally. As of today, approximately 1.9 billion doses have been administered worldwide, which is equal to 25 doses per 100 people. Among fully vaccinated countries (residents received all required doses) are United Arab Emirates, Israel, Aruba, Bahrain, and Malta. There is already a significant gap between vaccination programs in different countries, and there are still many regions waiting for the first dose. Such a gap is caused by certain discrepancies among countries that are explained by both internal and external factors. Our goal is to provide our client a detailed analysis interpreting our findings of the factors that affect current Covid-19 vaccination accomplishments to better understand and predict which features are significant and help boost the vaccination rate worldwide, which kind of vaccines are used and why, and how countries in the same region could expect different progress. This report will compare different countries and regions according to the given dataset and our findings from analyzing them to interpret them into business and public health insights as well as recommendations to our client. Provided below is the issue tree where we focus on the main question of how accessible the Covid-19 vaccine is and the factors that affect the topic. Included in the issue tree is also the analysis overview we performed to understand the main study question.

Diagram

Description automatically generated

**Literature/Industry Review**

There are various factors that serve as barriers to vaccinate the world population faster than the current speed of the process. Although the Covid-19 virus does not discriminate by gender, age, race, or moral beliefs, politics and international relations seem to be making the most impact towards such segregation that slows down the vaccination progress. The tension between certain countries caused by events in the past like sanctions, export bans, and wars are now affecting the political situation among those countries, which could determine the supply and support between them. Efficacy of vaccines and medical agency approval also play a significant role and they are identified as one of the components that determine which vaccine countries implement. For example, Russia’s Sputnik V vaccine was met with criticism for its early approval for distribution to the Russian population even before the trials ended. Some European and Western countries are still skeptical towards the vaccine, even though over a billion doses of the vaccine have been ordered worldwide. The existing skepticism along with political reasons of the West against Russia has slowed down the distribution to many countries including those in the EU because of medical agency disapproval caused by mistrust in Sputnik V efficacy. Next, vaccine prices are considered a huge influence on the vaccine choice of countries as some vaccines like Sputnik V are cheaper than most of its Western alternatives like Pfizer or Moderna, and in fact, is easier and cheaper to store and transport them as they are not required to be kept in freezers. Accessibility is another considerable element for governments to deal with. It is known that poorer nations struggle to access the most popular shots like Pfizer and Moderna, and some countries purchase a combination of different vaccines as they become available. On top of the four mentioned aspects that determine the vaccine that countries purchase, there are other key issues for Covid-19 vaccine distribution is shown below.



Distributing a COVID-19 Vaccine Across the U.S. - A Look at Key Issues

https://www.kff.org/report-section/distributing-a-covid-19-vaccine-across-the-u-s-a-look-at-key-issues-issue-brief/

Another question that was involved in our analysis was whether countries invested in the vaccine they are using today. Most of our findings fall along with the predictable outcomes of rich versus poor. Rich and powerful countries like the United States and the United Kingdom are both achieving great results in getting their populations vaccinated because they both could afford to invest in the early development of vaccines which let them stand at the front of the line for priority distribution and delivery. There are resourceful countries that did not follow the same route and some countries in the EU and Canada are currently a little further behind in the vaccination race. The reason is that they decided to invest in vaccines from European countries, afraid that the United States would issue export bans under former president Trump. However, currently, these factories are struggling to supply countries with vaccines and have been threatening export bans themselves. Even though Canada had the ability to order five times the supply it needs to vaccinate Canadians, for which they have been criticized, it did not position them for priority delivery.

Furthermore, the accessibility of the Covid-19 vaccine was an assessment of how well countries can develop and manufacture and distribute the vaccine. The Covid-19 vaccine was developed and tested after several months of the world being in shambles. The vaccine was rapidly developed and mass-produced due to the funding provided by large countries as the United States to aid testing and development and the ability to evolve preexisting vaccines similar to MERS that helped companies conduct research and manufacture the vaccines we see today.

The United States has the most extensive health expenditure per capita at around $10,966 as of 2019, which has led to the success of vaccines being readily available and administered to the current 135 million citizens that are fully vaccinated. Alongside the United States, countries such as India and China have also led the vaccination race by providing funding and manufacturing plants for these vaccines. India is currently 43.3 million that are fully vaccinated, and China with 430 million fully vaccinated. Yet, what has led these large GDP countries to their success is the availability of funds, and effective vaccination plans to lower the curve on positive Covid-19 cases.

However, what many countries around the globe from outputting the same number of vaccines with results of lowering the Covid-19 cases is each dose's price and efficacy rate. Each Pfizer dose administered is around $19.50 in the United States and 14.76 in European Union, with a 98% efficacy rate against Covid-19. Similarly, with Pfizer, the competitor Moderna is $15 in the United States and $18 in the European Union, with efficacy rates around 95%. Though, countries don't have the accessibility to billions of dollars to be spent on healthcare across the globe. This is why vaccines have been made to fit their price tag, such as AstraZeneca at $2.15 per dose. Alongside these cheaper costs for vaccines comes efficacy rates around 60%. Ultimately, there is a clear distinction between the correlation of large GDPs and the vaccines Pfizer and Moderna due to the ability to refrigerate and pay for each dose administered.

Comparably, the urban and rural populations have a strong correlation to GDP per capita. This is possible due to urban areas having a higher cost of living and accessibility to more jobs. However, rural residents have begun to lead in the vaccination race against urban residents. This could for reasons such as different age groups and the want and need to get vaccinated. Currently, 39% of all rural residents are vaccinated, and 16% have stated they will get it as soon as possible. Yet, urban populations have increased their want for vaccines that 35% of the urban population plans to get the vaccine as soon as available, on top of the current 31% fully vaccinated.

**Results**

1. **Data Source & Elements**

The dataset that had been using for this project was called *covid19.csv*. Data was collected from many countries around the world which allowed users to obtain most up to date worldwide vaccination information from the novel Corona Virus Disease started in 2019 in Wuhan City, Hubei Province of China. In summary, there are 114 countries in this dataset. The total 21 variables of data were broken into two main categories including features of different countries worldwide (gender ratio, surface area, region, health expenditure, etc.) and their Covid-19 vaccination accomplishments (people vaccinated, total vaccination, vaccines, etc.).

Rather than grouping the countries by region, GDP brackets were created through data cleaning process to ensure the accuracy for results and recommendations. There are five different brackets based on GDP per capita including High GDP per Cap for any country that has GDP per capita greater than $40,000; Upper Middle GDP per Cap for any country that has GDP per capita between $30,000 and $40,000; Middle GDP per Cap for any country that has GDP per capita between $20,000 and $30,000, Lower Middle GDP per Cap for any country that has GDP per capita between $10,000 and $20,000, and Low GDP per Cap for any country that has GDP per capita of less than $10,000. In addition to the data cleaning process, vaccines used in the country was filtered and categorized into seven binary variables indicated seven types of the vaccine instead of twenty-three vaccine combinations as given. Two quantitative variables including people vaccinated per hundred and total vaccination per hundred were used as predicted outcomes.

Vaccination campaigns are extremely complicated issues as several factors will affect a country’s efforts in different ways. In the dataset, certain variables were found to have indirect or unexpected effects that warranted further investigation on how they influenced the economy or healthcare initiatives. Notable examples include CO2 emissions, individual access to the internet, urban population, and percentage of young or older population. CO2 emission was found to both indicate strong economic productivity but also correlated to lower technological development. This factor made modeling more intricate as countries could be strong economically but may not possess strong infrastructure making it more difficult to predict their vaccination progress. The other variables followed traditional economic theory, indicating how developed a country was. Understanding these factors helped identified countries into more accurate groups for more effective insights for the client.

1. **Descriptive Analytics**

Prior research showed that countries’ vaccination efforts was a complex issue that was determined both by governmental factors as well as social-economic factors. This made each country a unique challenge due to the significant variance between their qualities and metrics. There are currently seven working vaccines approved for emergency uses in various countries internationally, adding further complexity to predicting the vaccination efforts. A chi-square test was utilized to examine the five GDP per cap brackets resulting in (X-squared = 50.757, df = 4, p-value = 2.509e-10), determining that there were significant differences between the countries based on their qualities. When running an ANOVA analysis between people vaccinated versus GDP per cap + CO2 Emissions, no significant differences were found between the factors, hinting that vaccination progress was closely tied with economic health but when analyzing people vaccinated per hundred, significant deviance was found between across all the brackets. This is visualized in figure VII.A as countries in the Middle and Lower Middle GDP per Cap brackets were able to outperform High and Upper GDP per Cap brackets on average. Healthcare systems also did not correlate with vaccination efforts; for further insight, refer to Figure VII.C in the appendix. These findings point to other factors that influence how equitable and efficient is a country’s vaccination effort.

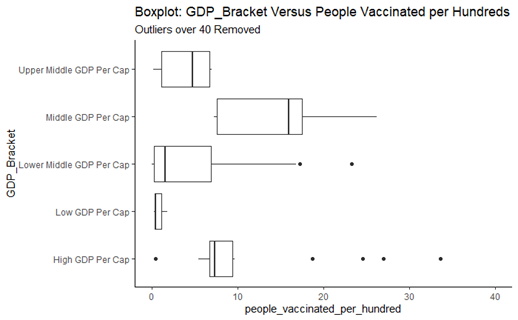


Figure VII.A

It was also important to examine how countries decided which vaccine should they adopt. It is crucial for the client to understand this due to the various direct and indirect costs each vaccine has. For this analysis, it is important to understand that J&J and Covaxin were only recently released when this dataset was created. This makes any analysis of these vaccines premature and warrants future analysis once production and adoption ramp up. Looking at Figure VII.B, Western vaccines such as Pfizer-Biotech, Moderna, and AstraZeneca-Oxford have seen rapid adoption versus Sinovac and Sputnik even though they have been released much sooner. This ties with the market research as countries prefer to trust reputable and proven pharmaceutical companies such as Pfizer and AstraZeneca over relatively untested companies such as Sinovac and Moderna when they can afford it. Sinovac and Sputnik see heavier adoption with less developed countries or countries with strong political ties with Russia or China.

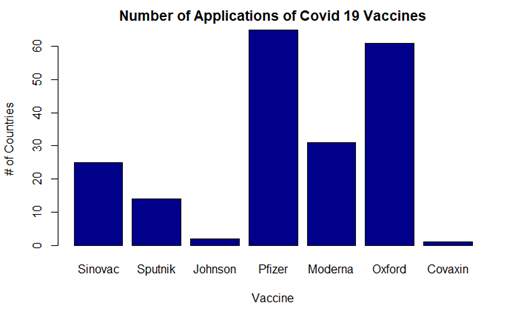


Figure VII.B

1. **Regression Tree with Standardization**

Regression Trees were conducted because the outcome of interest is numerical. The data was split multiple times to determine which country features played an important role and hold a significant impact on the overall dataset. Two separate tree-based models were built: one for people vaccinated per hundred (Figure VIII.A) and the other for total vaccination per hundred (Figure VIII.B).

Diagram, schematic

Description automatically generated  Diagram, schematic

Description automatically generated

*Figure VIII.A*

*Figure VIII.B*

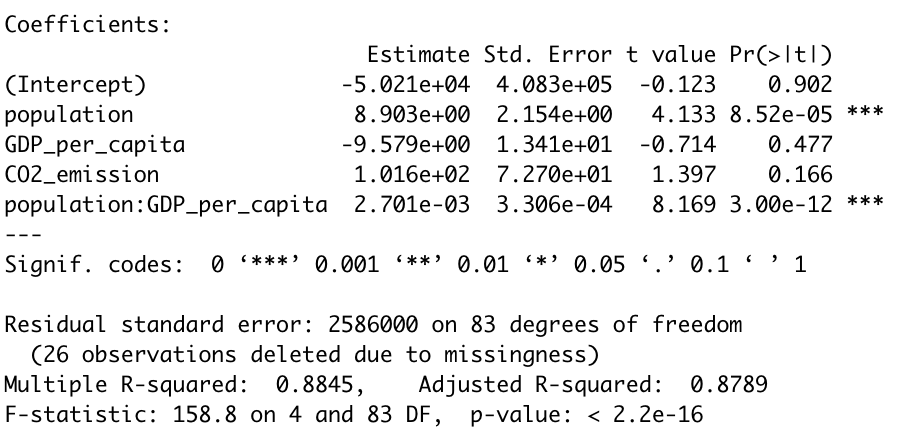
**People Vaccinated Per Hundred:**

**Total Vaccination Per Hundred:**

Both of the Regression Trees with standardization predicted the vaccination progress based on the percentage of the urban population, GDP per capita, gender ratio, and health expenditure. From these trees, several key findings were concluded. For instance, countries with a higher percentage of urban population (greater than 94.15%) are more likely to achieve higher vaccine accomplishment. As it has been showcased in the figure VIII.A, a country with an urban population of more than 94.15% should have 26.75% of people vaccinated. On the other hand, if countries with urban population less than 94.15%, then GDP per capita becomes the second indicator as well as population age 60 and above is the last indicator for people vaccinated per hundred regression tree. According to the total vaccination per hundred regression tree as figure VIII.B, if a country with a total percentage of urban population less than 94.15% but greater than 54.25% and its GDP per capita greater than $9051, it results in a higher percentage for their total vaccination as of 24.94%.

1. **Multiple Linear Regression**

Multivariate Linear Regression was performed to predict the number of people vaccinated. The variable that is people vaccinated is focused on the progress of a country’s vaccination effort. The most successful model with interaction found was people vaccinated per country is best predicted by their population multiplied by GDP per capita plus their CO2 emissions. The interaction effect means that two or more variables combined have a significantly larger effect on a feature as compared to the sum of the individual variables alone. Population and GDP per capita were found to have a strong correlation as the research showed stronger economies able to afford more vaccines for their populace. Additionally, CO2 emissions show the complexity of an economy, tying with the research on how developed countries have an edge with their vaccination progress. The results of running the regression are shown below:

The model has a very high R squared at 88.5% and adjusted R squared at 87.9%. Moreover, the p-value is well below 0.05 showing that the variables are significant indicators of people vaccinated. The coefficients of the two variables are positive which means that as population, GDP per capita, and CO2 emissions increase so do people vaccinated. This model is useful to provide insight on how vaccination efforts would progress in various countries around the globe as they start their full-scale vaccination efforts. Furthermore, this model works effectively with underdeveloped or developing countries.

1. **Vaccine Classification Adoption**

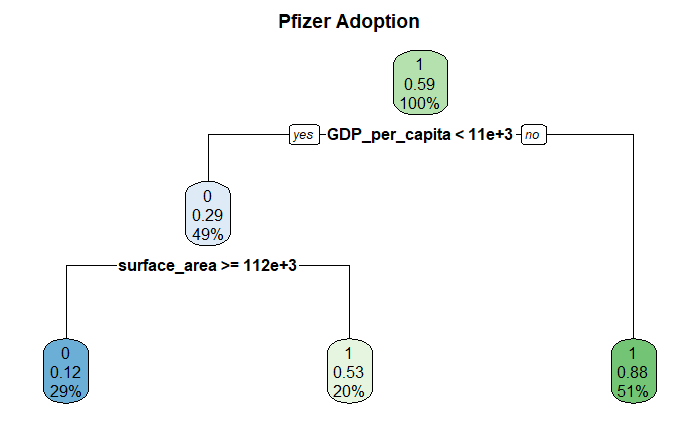


Figure X.A

The seven different approved vaccines have different direct and indirect costs for adoption. It is important for the client to understand how a country’s factors match with the vaccine. Running a classification tree for each vaccine, treating them as binary variables, allowed for analysis of which factors were determinants of the country’s adoption. It is important to note that the outcome variables from the other models were removed to prevent multi-collinearity with the findings. For the main paper, only Pfizer – Biotech (Figure X.A), Moderna (Figure X.B), and AstraZeneca – Oxford (Figure X.C) will be discussed; the other vaccines can be referred to in the appendix under Figure X.D, Figure X.E, Figure X.F, and Figure X.G. Looking at Pfizer – Biotech and Moderna first as they have the highest efficacy and higher cost, the model predicted that only denser, developed countries could afford to adopt these vaccines. This was further exacerbated as it was found through the research that priority came to countries that invested in the vaccines and had the infrastructure such as deep freeze freezers to handle storing the doses. The model noted that these two vaccines were preferred over the other five, illustrated by the less complex decisions needed to conclude if they will adopt the vaccine. The AstraZeneca – Oxford classification tree and the Sinovac Tree see more scrutiny as countries start to consider their unique circumstances and economic levels to consider if these vaccines are a good fit. The most notable is in the AstraZeneca tree, the country ability to adopt of Moderna was the largest separator for countries. This has been explained in the research from the efficacy difference of 94% to 74%. Sinovac was predicted to be the inferior vaccine, having a near reverse of the decision criteria of Pfizer, favoring countries with weaker economies.

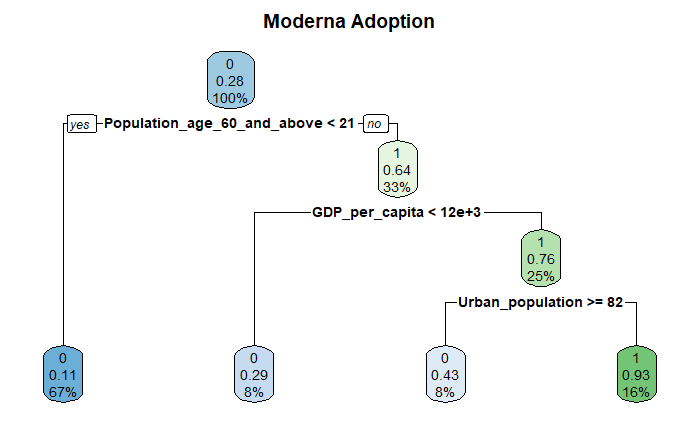


Figure X.B

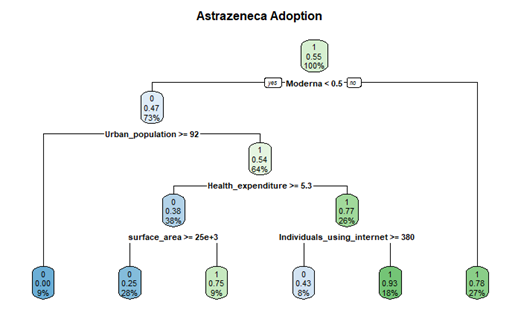


Figure X.C

At this time, the other vaccines except Sinovac did not see wide enough adoption to run effective trees. Tying with this, AstraZeneca – Oxford and J&J’s vaccines but saw health issues that might influence public perception negatively and slow down its adoption rates. These tests should be run again in the future as production of Sinovac, J&J, and Covaxin ramps up. It is also important to note that healthcare systems did not play any significant effect in the decision-making process of vaccine adoption, but research suggested that economic and political ties did influence adoption rates of some vaccines over others. Utilizing these models will help the client determine which vaccine is a good fit for their target countries.

**Discussion**

1. **Business Insights**

The models designed for this project will allow the client to predict and understand how vaccination progress and equity will be depending on a country’s features. With this, a model was also designed to determine which vaccines would be a good fit for the country. The models determined that countries with denser urban sprawls and older populations should be prioritized as the research and model found these countries to be the most vulnerable to be adversely affected by prolonged pandemics. Factors like CO2 emissions and population age became major determinants for vaccination efforts as they help predict the level of healthcare equity and capacity a country has. The model showed Co2 emission in particular to be a wild card factor, balancing between strong capacity but poor healthcare equity as emissions rose in a country. GDP per cap and larger older population was confirmed to follow the traditional economic understanding of growth and strength.

An interesting result with the model was that larger sized countries were found to favor Sinovac and AstraZeneca due to their accessibility while small countries tended to focus on Pfizer and Moderna. Pfizer and Moderna were found to be only good fits for the developed economy despite their high efficacy rates. Single-dose vaccines are imperative towards global inoculation efforts due to their low infrastructural costs and lower commitment needed from poorer populations. As the single-dose vaccines were only recently released and approved at the time of this data set creation, it is imperative that the client commit more research on a time series basis to gain more understanding on vaccination progress over time and how these new single-dose vaccines will perform. This is also crucial due to health controversies with J&J, Sinovac, and AstraZeneca which may have long term negative effects on adoption rates.

1. **Recommendation**

Although the pandemic nears its ends for the developed countries like the U.S, Canada, Germany, and China, the majority of countries are predicted to not be fully vaccinated until 2023. This makes it crucial to understand and predict how vaccination efforts will progress in lesser developed countries and which vaccines would be a good fit for their infrastructure capacities. The majority of countries do not have intricate economies or healthcare systems making it possible to predict the minimum vaccination progress in a country. The client is advised for these countries to focus on implementing the current single-dose vaccines that have been approved by the W.H.O due to the low infrastructural and logistical requirement to inoculate the populace. AstraZeneca – Oxford, J&J, Sinovac, and Covaxin are currently the leading options while also being the most accessible for poorer countries to acquire and implement. These vaccines except for Sinovac have less political commitments, allowing for wider implementation and accessibility by countries. The client should then focus its finite resource on countries with larger urban sprawls to limit the mass spread and maximize inoculation progress and equity.

**Appendix**

1. **Reference**

Hegarty, Stephanie. “Covid Vaccine Tracker: How's My Country and the Rest of the World Doing?” *BBC News*, BBC, 12 Feb. 2021, www.bbc.com/news/world-56025355.

Steenhuysen, Julie, et al. “Special Report: Countries, Companies Risk Billions in Race for Coronavirus Vaccine.” *Reuters*, Thomson Reuters, 25 Apr. 2020, www.reuters.com/article/us-health-coronavirus-vaccine-specialrep/special-report-countries-companies-risk-billions-in-race-for-coronavirus-vaccine-idUSKCN2270U1.

Litvinova, Daria. “Sputnik V Vaccine Was Met with Skepticism; Now Russia Struggles to Make Enough.” *The Times of Israel*, 9 Mar. 2021, [www.timesofisrael.com/demand-for-russian-sputnik-v-virus-vaccine-is-spiking-but-supply-snags/](http://www.timesofisrael.com/demand-for-russian-sputnik-v-virus-vaccine-is-spiking-but-supply-snags/).

Centers for Disease Control and Prevention. (2021, February 4). *Demographic Characteristics of Persons Vaccinated During the First Month of the COVID-19 Vaccination Program - United States, December 14, 2020–January 14, 2021*. Centers for Disease Control and Prevention. <https://www.cdc.gov/mmwr/volumes/70/wr/mm7005e1.htm>.

Hua, S. (2021, April 10). *China's Covid-19 Vaccination Campaign Gets Off to Slow Start*. The Wall Street Journal. <https://www.wsj.com/articles/china-knocks-on-peoples-doors-to-speed-up-domestic-covid-19-vaccination-11618069774>.

Nambi Ndugga Follow @nambinjn on Twitter, O. P. (2021, May 26). *Latest Data on COVID-19 Vaccinations Race/Ethnicity*. KFF. <https://www.kff.org/coronavirus-covid-19/issue-brief/latest-data-on-covid-19-vaccinations-race-ethnicity/>.

Ashley Kirzinger Follow @AshleyKirzinger on Twitter, C. M. (2021, April 17). *Vaccine Hesitancy in Rural America*. KFF. <https://www.kff.org/coronavirus-covid-19/poll-finding/vaccine-hesitancy-in-rural-america/>.

1. **Other Visuals and Graphs**

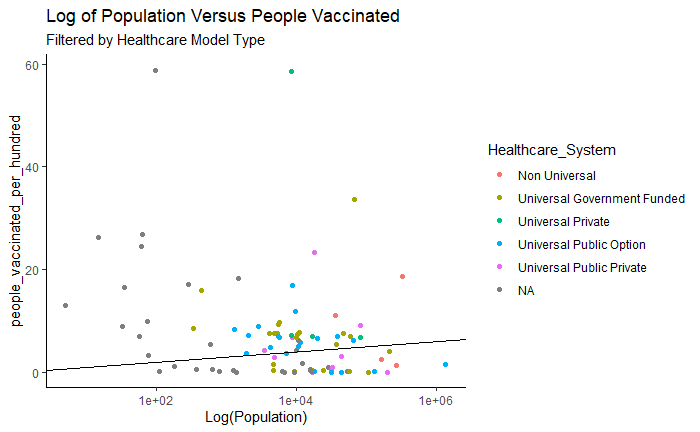
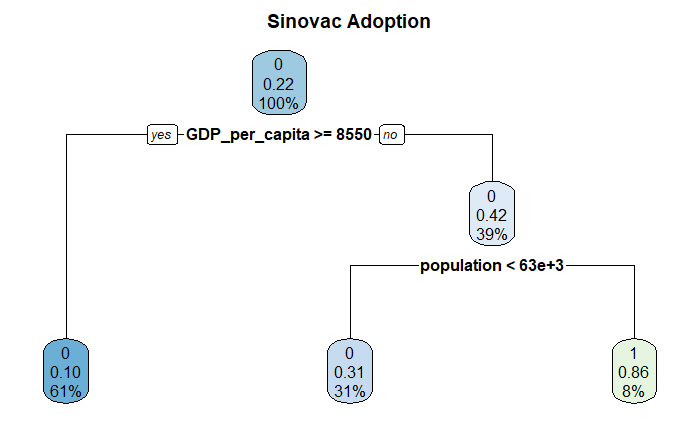
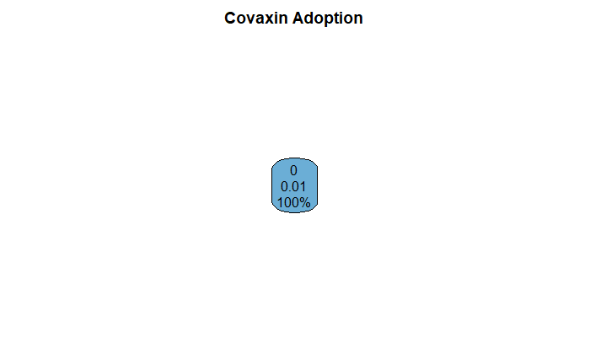


Figure VII.C

Graphical user interface, application

Description automatically generated

Figure X.D Figure X.E

Graphical user interface, application

Description automatically generated

Figure X.F Figure X.G

1. **R-Code**

#Data Cleaning

covid\_db\_raw <- as.data.frame(read.csv("covid19.csv"))

healthcare\_db\_raw <- as.data.frame(read\_excel("Country Healthcare Performance.xlsx"))

covid\_db\_vax <- covid\_db\_raw %>%

mutate(Sinovac = case\_when(str\_detect(vaccines, "Sinovac") ~ 1,

str\_detect(vaccines, "Sinopharm") ~ 1),

Sputnik = case\_when(str\_detect(vaccines, "Sputnik") ~ 1),

Johnson = case\_when(str\_detect(vaccines, "Johnson&Johnson")~ 1),

Pfizer = case\_when(str\_detect(vaccines, "Pfizer") ~ 1),

Moderna = case\_when(str\_detect(vaccines, "Moderna") ~ 1),

AstraZeneca = case\_when(str\_detect(vaccines, "Oxford") ~ 1),

Covaxin = case\_when(str\_detect(vaccines, "Covaxin") ~ 1)) %>%

dplyr::select(country, Sinovac, Sputnik, Johnson, Pfizer, Moderna, AstraZeneca, Covaxin) %>%

replace(is.na(.), 0) %>%

mutate(Total\_Vax\_Type = Sinovac + Sputnik + Johnson + Pfizer + Moderna + AstraZeneca + Covaxin)

covid\_db\_full\_clean <- covid\_db\_raw %>%

left\_join(covid\_db\_vax, by = "country") %>%

mutate(GDP\_Bracket = ifelse(GDP\_per\_capita >= 40000, "High GDP Per Cap",

ifelse(GDP\_per\_capita >= 30000, "Upper Middle GDP Per Cap",

ifelse(GDP\_per\_capita >= 20000, "Middle GDP Per Cap",

ifelse(GDP\_per\_capita >= 10000, "Lower Middle GDP Per Cap", "Low GDP Per Cap"))))) %>%

left\_join(healthcare\_db\_raw, by = c( "country" = "Country"))

#Descriptive Analytics

covid\_db\_full\_clean %>%

filter(country %in% c("United States of America", "United Kingdom", "China", "Russia", "Japan", "South Africa", "Israel", "Germany", "India", "Brazil", "Grenada")) %>%

group\_by(country) %>%

summarise(Avg\_Pop = mean(population, na.rm = T),

Avg\_GDP\_Per\_Cap = mean(GDP\_per\_capita ,na.rm = T),

AVG\_Health\_Expenduture = mean(Health\_expenditure, na.rm = T),

AVG\_Total\_Vaccination = mean(total\_vaccinations),

AVG\_people\_Vaccinated = mean(people\_vaccinated, na.rm = T),

AVG\_Vaccine\_Utilization = mean(Total\_Vax\_Type))

chisq.test(table(covid\_db\_full\_clean$GDP\_Bracket))

chisq.test(table(covid\_db\_full\_clean$GDP\_Bracket), table(covid\_db\_full\_clean$Healthcare\_System))

t.test(GDP\_per\_capita ~ Pfizer, data = covid\_db\_full\_clean)

anova(aov(people\_vaccinated\_per\_hundred ~ GDP\_growth\_rate \* CO2\_emission, data = covid\_db\_full\_clean))

#Multiple Linear Regression Model to predict number of people\_vaccinated

summary(lm(people\_vaccinated ~ surface\_area+population+Health\_expenditure +

CO2\_emission,data=covid19))

#Mean Imputation of Multiple Columns that have NAs

covid19$GDP\_per\_capita[is.na(covid19$GDP\_per\_capita)] <- mean(covid19$GDP\_per\_capita, na.rm = TRUE)

for(i in 1:ncol(covid\_db\_full\_clean)) {

covid\_db\_full\_clean[ , i][is.na(covid\_db\_full\_clean[ , i])] <- mean(covid\_db\_full\_clean[ , i], na.rm = TRUE)}

head(covid\_db\_full\_clean)

#Slipt data

selec <- sample(1:nrow(covid19), .6\*nrow(covid19))

Training <- covid19[selec, ]

Testing <- covid19[-selec, ]

#People\_vaccinated\_per\_hundred Regression Tree

a <- rpart(people\_vaccinated\_per\_hundred~Population\_age\_60\_and\_above

+Urban\_population+GDP\_per\_capita, cp=0, data=Training)

print(a)

plot(a, uniform=T)

text(a, use.n=T,cex=1.1, all=T, minlength=9, xpd=T)

#Total\_vaccination\_per\_hundred Regression Tree

a1 <- rpart(total\_vaccinations\_per\_hundred~gender\_ratio + Urban\_population + Health\_expenditure + GDP\_per\_capita, cp=0, data=Training)

print(a1)

plot(a1, uniform=T)

text(a1, use.n=T,cex=1.1, all=T, minlength=9, xpd=T)

#People\_Vaccinated\_per\_hundred Interaction Linear Regression

best\_model <- summary(lm(people\_vaccinated ~ population\*Population\_age\_60\_and\_above\*GDP\_per\_capita + CO2\_emission, data = covid\_db\_train))

#Best result so far Adjusted R of .93

#Shows strong correlation between how much of the population is over 60

#Best Model

covid\_db\_full\_clean %>%

mutate(Predicted\_People\_Vaccinated = predict(best\_model, newdata = covid\_db\_full\_clean, type = "response"),

Difference = Predicted\_People\_Vaccinated - people\_vaccinated) %>%

filter(country %in% c("China", "Mexico", "Thailand")) %>%

dplyr::select(country,

population,

GDP\_per\_capita,

Predicted\_People\_Vaccinated)

#Visualization

covid\_db\_full\_clean %>%

group\_by(region) %>%

summarise(mean(population),

mean(surface\_area),

mean(GDP\_per\_capita),

mean(Total\_Vax\_Type),

mean(total\_vaccinations\_per\_hundred))

covid\_db\_full\_clean %>%

ggplot(aes(x = population, y = total\_vaccinations\_per\_hundred, col = region)) +

geom\_point() +

scale\_x\_continuous(trans = 'log10') +

scale\_y\_continuous(trans = 'log10')+

geom\_abline()

#Should consolidate the Regions

#Histogram Stopgap Method

x <- covid\_db\_full\_clean %>%

summarise(Sinovac = sum(Sinovac),

Sputnik = sum(Sputnik),

Johnson = sum(Johnson),

Pfizer = sum(Pfizer),

Moderna = sum(Moderna),

Oxford = sum(AstraZeneca),

Covaxin = sum(Covaxin)) %>%

pivot\_longer(everything())

barplot(x$value,

xlab = "Vaccine",

ylab = "# of Countries",

main = "Number of Applications of Covid 19 Vaccines",

names.arg = x$name,

col = "dark blue")

covid\_db\_full\_clean %>% #Useless

ggplot(aes(x = population)) +

geom\_histogram(bins = 30) +

scale\_x\_log10()

covid\_db\_full\_clean %>% #Useless

ggplot(aes(x = GDP\_per\_capita)) +

geom\_histogram(bins = 30)

covid\_db\_full\_clean %>% #Need to determine What the heck is employment Industry

ggplot(aes(x = employment\_industry, y = total\_vaccinations, col = GDP\_Bracket)) +

geom\_point() +

scale\_y\_log10()

covid\_db\_full\_clean %>%

filter(!is.na(GDP\_Bracket) & !is.na(people\_vaccinated\_per\_hundred)) %>%

ggplot(aes(x = GDP\_Bracket, y = people\_vaccinated\_per\_hundred)) +

geom\_boxplot() +

scale\_y\_continuous(limits = c(0, 40)) +

coord\_flip() +

ggtitle("Boxplot: GDP\_Bracket Versus People Vaccinated per Hundreds", "Outliers over 40 Removed") +

theme\_classic()

covid\_db\_full\_clean %>%

group\_by(GDP\_Bracket) %>%

summarise(max(people\_vaccinated\_per\_hundred, na.rm = T),

mean(Total\_Vax\_Type))

covid\_db\_full\_clean %>%

filter(!is.na(GDP\_Bracket)) %>%

ggplot(aes(x = population, y = people\_vaccinated\_per\_hundred, col = GDP\_Bracket)) +

geom\_point() +

scale\_x\_log10() +

geom\_abline() +

xlab("Log(Population)") +

ggtitle("Log of Population Versus People Vaccinated", "Countries with no GDP per Capita removed") +

theme\_classic()

covid\_db\_full\_clean %>%

filter(!is.na(GDP\_Bracket)) %>%

ggplot(aes(x = population, y = people\_vaccinated\_per\_hundred, col = Healthcare\_System)) +

geom\_point() +

scale\_x\_log10() +

geom\_abline() +

xlab("Log(Population)") +

ggtitle("Log of Population Versus People Vaccinated", "Filtered by Healthcare Model Type") +

theme\_classic()

#Histogram

covid\_db\_full\_clean %>%

filter(!is.na(people\_vaccinated\_per\_hundred)) %>%

ggplot(aes(x = people\_vaccinated\_per\_hundred)) +

geom\_histogram(binwidth = 5, colour="black", fill="white")

covid\_db\_full\_clean %>%

filter(!is.na(people\_vaccinated)) %>%

ggplot(aes(x = people\_vaccinated)) +

geom\_histogram(colour="black",

fill="white")

#Histogram with layers

covid\_db\_full\_clean %>%

filter(!is.na(people\_vaccinated\_per\_hundred)) %>%

ggplot(aes(x = people\_vaccinated\_per\_hundred, fill = GDP\_Bracket)) +

geom\_histogram(binwidth = 15,

position = "dodge",

alpha=.5)

covid\_db\_full\_clean %>%

filter(!is.na(people\_vaccinated)) %>%

ggplot(aes(x = people\_vaccinated, fill = GDP\_Bracket)) +

geom\_histogram( position = "identity",alpha=.5)

# Vaccine Prediction

vaccine\_predictor <- function(vaccine\_type, model) {

Predicted\_Decision <- predict(model, test\_data, type = 'class')

table(vaccine\_type, Predicted\_Decision)

}

sample <- sample.split(test\_random$people\_vaccinated\_per\_hundred, SplitRatio = .75)

train\_data <- subset(test\_random, sample == T)

test\_data <- subset(test\_random, sample == F)

Pfizer\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = Pfizer ~ .,

method = "class")

rpart.plot(Pfizer\_Tree, main = "Pfizer Adoption")

vaccine\_predictor(test\_data$Pfizer, Pfizer\_Tree)

Moderna\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = Moderna ~ .,

method = "class")

rpart.plot(Moderna\_Tree, main = "Moderna Adoption")

vaccine\_predictor(test\_data$Moderna, Moderna\_Tree)

Sinovac\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = Sinovac ~ .,

method = "class")

rpart.plot(Sinovac\_Tree, main = "Sinovac Adoption")

AstraZeneca\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = AstraZeneca ~ .,

method = "class")

rpart.plot(AstraZeneca\_Tree, main = "Astrazeneca Adoption")

Johnson\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = Johnson ~ .,

method = "class")

rpart.plot(Johnson\_Tree, main = "J&J Adoption")

Sputnik\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = Sputnik ~ .,

method = "class")

rpart.plot(Sputnik\_Tree, main = "Sputnik Adoption")

predict(AstraZeneca\_Tree, test\_data, type = 'class')

table(test\_data$AstraZeneca, predict(AstraZeneca\_Tree, test\_data, type = 'class'))

Covaxin\_Tree <- train\_data %>%

dplyr::select(-total\_vaccinations, -people\_vaccinated, -total\_vaccinations\_per\_hundred, -people\_vaccinated\_per\_hundred, -Total\_Vax\_Type) %>%

rpart(formula = Covaxin ~ .,

method = "class")

rpart.plot(Covaxin\_Tree, main = "Covaxin Adoption")