Project 1 - Analyzing COVID-19 Vaccination Rates

By Armanul Ambia, Iftekharul Islam

Data Wrangling

All data wrangling is performed using the Tidyverse library in R.

Population - demographics.csv

	Country Name	Country Code	Series Name	Series Code	YR2015
1	Afghanistan	AFG	Life expectancy at birth, total (years)	SP.DYN.LE00.IN	6.337700e+01
2	Afghanistan	AFG	Urban population	SP.URB.TOTL	8.535606e+06
3	Afghanistan	AFG	Population, total	SP.POP.TOTL	3.441360e+07
4	Afghanistan	AFG	Population ages 80 and above, female	SP.POP.80UP.FE	4.831900e+04
5	Afghanistan	AFG	Population ages 80 and above, male	SP.POP.80UP.MA	3.723300e+04

The demographics.csv data set consists of population data for a large list of countries. These populations are further divided into categories based on age and sex. All of these categories fall under two columns called Series Name and Series Code, where Series Code is an abbreviated version of Series Name. This redundancy makes the data set untidy and difficult to work with. To fix this, the Series Name column is removed entirely and the values of the Series Code are converted into their own individual columns. Additional columns are created as well to combine the sex-specific population columns into total population columns regardless of sex. The only columns necessary for the linear modeling are Country Code, SP.DYNLE00.IN (life expectancy at birth), SP.URB.TOTL (total urban population), SP.POP.0014.IN (number of people age 14 and under), SP.POP.1564.IN (number of people between age 15 and 64), SP.POP.65UP.IN (number of people age 65 and above), SP.POP.80UP (number of people age 80 and up), and SP.DYN.AMRT (mortality rate). The remaining columns are all removed. The following shows the first few rows of the demographics data set after data wrangling is complete.

	Country Code	SP.DYN.LE00.IN	SP.URB.TOTL	SP.POP.0014.IN	SP.POP.1564.IN	SP.POP.65UP.IN	SP.POP.80UP	SP.DYN.AMRT
1	AFG	63.37700	8535606	15443807	18116800	852996	85552	455.4700
2	ALB	78.02500	1654503	537788	1979175	363740	66965	150.4100
3	DZA	76.09000	28146511	11404930	25993589	2329506	453741	191.6310
4	ASM	*NA*	48689	*NA*	*NA*	*NA*	*NA*	*NA*
5	AND	*NA*	68919	*NA*	*NA*	*NA*	*NA*	*NA*

GDP - API_NY.GDP.MKTP.CD_DS2_en_csv_v2_3011433.csv

	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961
1	Aruba	ABW	GDP (current US$)	NY.GDP.MKTP.CD	*NA*	*NA*
2	Africa Eastern and Southern	AFE	GDP (current US$)	NY.GDP.MKTP.CD	1.929193e+10	1.970186e+10
3	Afghanistan	AFG	GDP (current US$)	NY.GDP.MKTP.CD	5.377778e+08	5.488889e+08
4	Africa Western and Central	AFW	GDP (current US$)	NY.GDP.MKTP.CD	1.040732e+10	1.113130e+10
5	Angola	AGO	GDP (current US$)	NY.GDP.MKTP.CD	*NA*	*NA*

The API_NY.GDP.MKTP.CD_DS2_en_csv_v2_3011433.csv data set consists of gross domestic product (GDP) data for a list of countries. The GDP for each country is listed from the years 1960 to 2021. For the linear model, only the most recent GDP data for a country is needed. To calculate this, multiple steps are required. Each year in the data set is designated its own column. All of these years are put into a single column. The data set is then arranged by descending order of the year while keeping all the years for a single country together. Then, the data set is grouped by the Country Code before selecting the top year, which represents the most recent year since the years are listed in descending order. Now that the most recent GDP has been found, only the Country Code and GDP columns are kept since the remaining columns are unnecessary for the linear model. The following shows the first few rows of the GDP data set after data wrangling is complete.

	Country Code	GDP
1	ABW	3.202189e+09
2	AFE	8.984741e+11
3	AFG	1.980707e+10
4	AFW	7.865850e+11
5	AGO	6.230691e+10

Vaccinations - time_series_covid19_vaccine_doses_admin_global.csv

	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	Population	2020-12-12
1	4	AF	AFG	4	*NA*	*NA*	*NA*	Afghanistan	33.9391	67.7100	Afghanistan	38928341	*NA*
2	8	AL	ALB	8	*NA*	*NA*	*NA*	Albania	41.1533	20.1683	Albania	2877800	*NA*
3	12	DZ	DZA	12	*NA*	*NA*	*NA*	Algeria	28.0339	1.6596	Algeria	43851043	0
4	20	AD	AND	20	*NA*	*NA*	*NA*	Andorra	42.5063	1.5218	Andorra	77265	0
5	24	AO	AGO	24	*NA*	*NA*	*NA*	Angola	-11.2027	17.8739	Angola	32866268	*NA*

The time_series_covid19_vaccine_doses_admin_global.csv data set consists of global data for COVID-19 vaccinations. Each country is listed in its own individual row with individual columns for every single day vaccinations occurred. This data is tidied by making a single column of dates and another column counting the number of vaccinations given that day. Because not all countries started vaccinating on the same day, many rows in the date column have values of 0. These rows are removed from the data set. Then, the vaccination rate is calculated for each day by dividing the number of shots given by the population of the country and adding these values to a new column.

For the linear model, another column counting the days since vaccinations began in a country is needed. To do this, a new data frame called startDate is created by grouping the COVID-19 vaccination data set by the Country Region and creating a column consisting of the earliest vaccination date. Then, the original vaccination data set is merged with startDate. This gives the original data set an additional column consisting of the earliest vaccination dates for every country. This date is then subtracted from the date of vaccinations for every row to create a new column counting the days since vaccinations started for each individual day of every country. This number is off by 1 so 1 is added to this value as well. Finally, the date and startDate columns are removed from the data set since they are not needed for the linear model. The following shows the first few rows of the data set after data wrangling is complete.

	iso3	Country_Region	Population	Shots	vacRate	daysSinceStart
1	AFG	Afghanistan	38928341	8200	0.0002106434	1
2	AFG	Afghanistan	38928341	8200	0.0002106434	2
3	AFG	Afghanistan	38928341	8200	0.0002106434	3
4	AFG	Afghanistan	38928341	8200	0.0002106434	4
5	AFG	Afghanistan	38928341	8200	0.0002106434	5

Final Combined Data Frame

Now that all the tables have been modified, they can be joined together. This function is called twice to first combine COVID-19 data with GDP data and then combine the COVID-19 and GDP combination with the demographics data. In both of these join calls, the iso3 and Country Code columns are identified as the same values across all tables to ensure that the tables are joined correctly despite having different column names that represent the same thing. After the join is complete, the columns are rearranged to separate the dependent variable (vacRate) from the predictor variables. The combined data set is now ready for linear modeling.

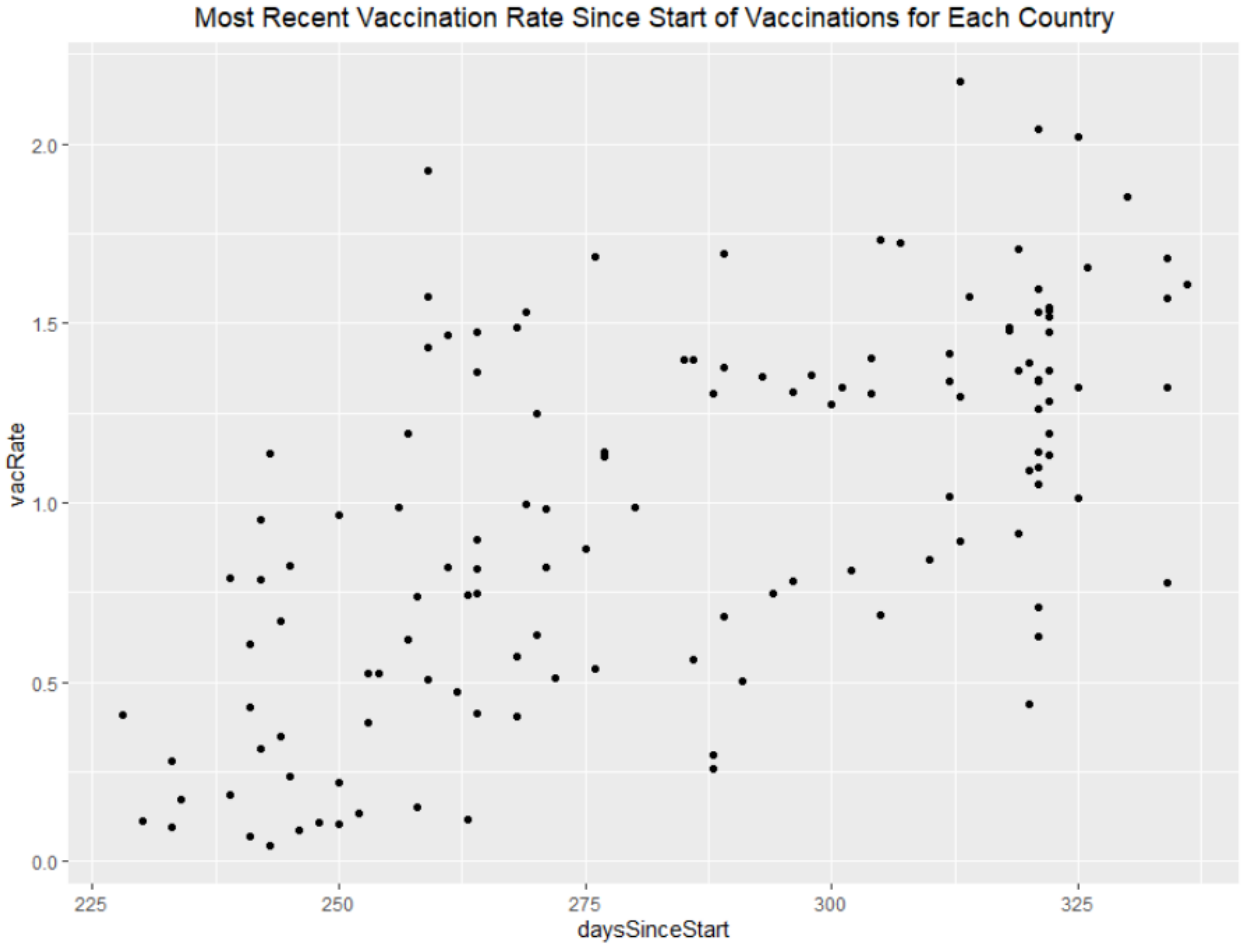
Columns 1-7:

	iso3	Country_Region	vacRate	Shots	Population	daysSinceStart	GDP
1	AFG	Afghanistan	0.0002106434	8200	38928341	1	19807067268
2	AFG	Afghanistan	0.0002106434	8200	38928341	2	19807067268
3	AFG	Afghanistan	0.0002106434	8200	38928341	3	19807067268
4	AFG	Afghanistan	0.0002106434	8200	38928341	4	19807067268
5	AFG	Afghanistan	0.0002106434	8200	38928341	5	19807067268

Columns 8-14:

SP.DYN.LE00.IN	SP.URB.TOTL	SP.POP.0014.IN	SP.POP.1564.IN	SP.POP.65UP.IN	SP.POP.80UP	SP.DYN.AMRT
63.377	8535606	15443807	18116800	852996	85552	455.47
63.377	8535606	15443807	18116800	852996	85552	455.47
63.377	8535606	15443807	18116800	852996	85552	455.47
63.377	8535606	15443807	18116800	852996	85552	455.47
63.377	8535606	15443807	18116800	852996	85552	455.47

Scatter Plot



Linear Modeling

The following displays the models used and their corresponding R-squared values:

1. vacRate ~ GDP + daysSinceStart

   This model was chosen to show how vaccination rate changed relative to the number of days since vaccination started and the GDP.

```
> model1 = lm(formula = vacRate ~ GDP + daysSinceStart, data =
combined)
> r1 <- summary(model1)$r.squared
[1] 0.5818341
```

2. vacRate ~ GDP + Life Expectancy + daysSinceStart

   This model was chosen to show how vaccination rate changed with respect to the number of days since vaccination started, the GDP, and life expectancy.

```
> model2 <- lm(formula = vacRate ~ GDP + SP.DYN.LE00.IN +
daysSinceStart, data = combined)
> r2 <- summary(model2)$r.squared
[1] 0.68798
```

3. vacRate ~ Population Up to Age 14 + daysSinceStart

   This model was chosen to show how vaccination rate changed relative to the number of days since vaccination started and the number of people of age 14 and below.

```
> model3 = lm(formula = vacRate ~ SP.POP.0014.IN + daysSinceStart,
data = combined)
> r3 <- summary(model3)$r.squared
[1] 0.5890672
```

4. vacRate ~ Population Age 15 to 64 + daysSinceStart

   This model was chosen to show how vaccination rate changed relative to the number of days since vaccination started and the number of people between age 15 and 64.

```
> model4 = lm(formula = vacRate ~ GDP + SP.POP.1564.IN +
daysSinceStart, data = combined)
```

```
> r4 <- summary(model4)$r.squared
[1] 0.5852769
```

5. vacRate ~ Population Age 65 and Up + daysSinceStart

   This model was chosen to show how vaccination rate changed relative to the number of days since vaccination started and the number of people of age 65 and above.

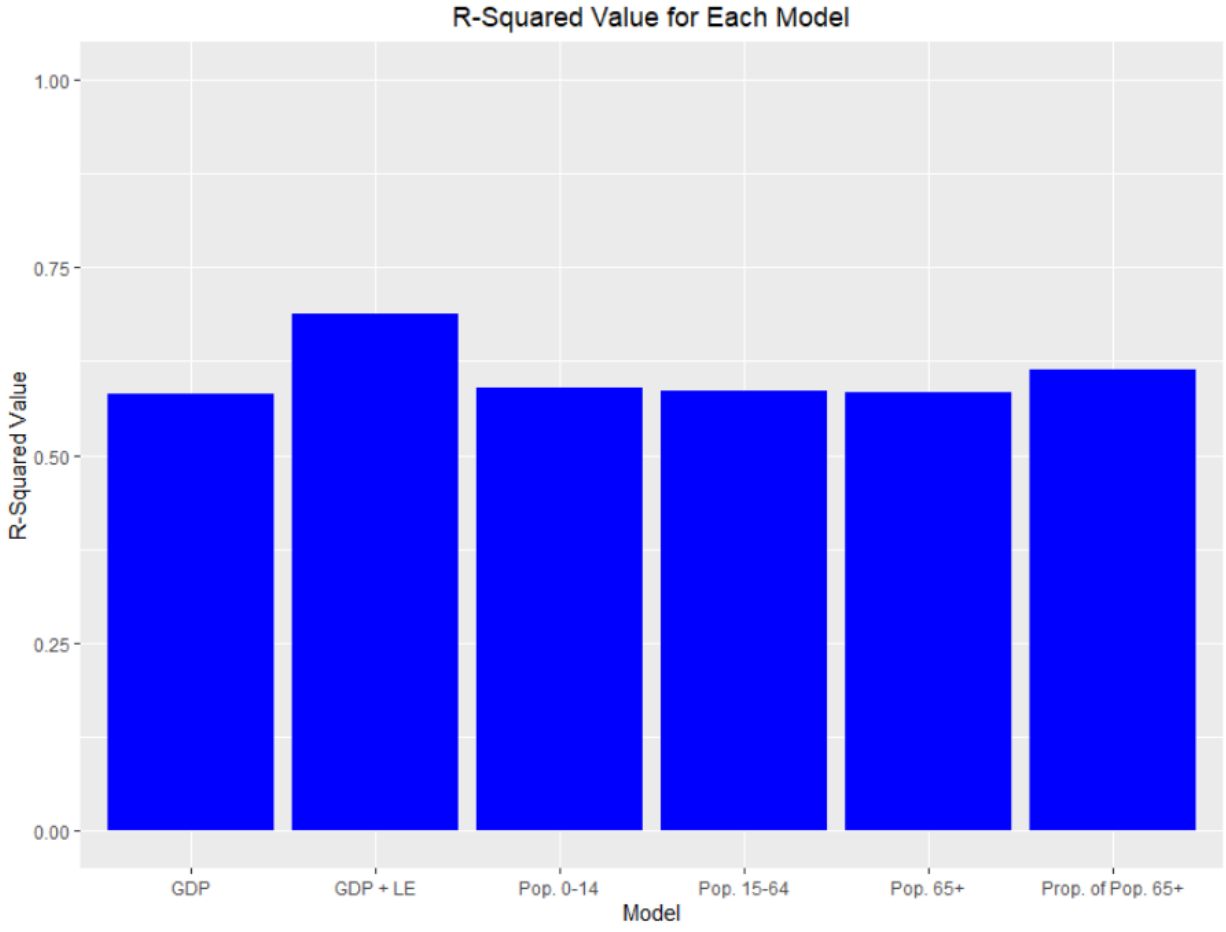
```
> model5 <- lm(formula = vacRate ~ SP.POP.65UP.IN + daysSinceStart,
data = combined)
> r5 <- summary(model5)$r.squared
[1] 0.5843081
```

6. vacRate ~ Proportion of Age 65 and Up + daysSinceStart

   This model was chosen to show how vaccination rate changed with respect to the proportion of the number of people of age 65 and above and the total population.

```
# Transform the dataset by adding a new variable, proportion65
> combined <- combined %>% mutate(proportion65 =
SP.POP.65UP.IN/Population)
> model6 = lm(formula = vacRate ~ GDP + proportion65 +
daysSinceStart, data = combined)
> r6 <- summary(model6)$r.squared
[1] 0.6130355
```

Bar Plot of R-squared Values



Conclusion

After tidying the data, we analyzed the R-Squared values of 6 different models. These models included GDP, GDP and Life Expectancy, Population from 0 to 14, Population from 15 to 64, Population 65 and up, and the proportion of Population 65 and up. Each of these was also coupled with the daysSinceStart variable in their respective linear models. The plot of these R-Squared values revealed relatively similar values ranging from .58 to .69. Our clear best predictor was GDP and Life Expectancy. This hints that wealthier countries with higher life expectancy correlate to a high vaccination rate. Additionally, we saw a slight increase in the proportion of 65 and Up population models as well. This model had an R-Squared value higher than the previous model that only accounted for the population of 65 and up. This suggests that the proportion of people 65 and up compared to the entire population is more important than the sheer number of elderly people in the population.