Project 1 - COVID Vaccination Rates

Vaccine Data Wrangling

With the vaccine data, we were only studying country-level rates so I first got rid of any rows containing provincial data as well as unusable countries that do not have any population data. I then tidied the data and put the dates and number of vaccinations into two separate columns instead of having them spread out over 467 columns. After tidying the data, I then deleted irrelevant columns. Since we only care about the days since the first vaccination of each country, I filtered out dates where 0 vaccinations took place. Finally, I calculated and added a column for the vaccination rate of each country as well as a column that tracks the days since the first vaccination.

```
# VACCINE DATA
# Filter out provinces and countries with no population
vax <- vax %>% filter(is.na(Province_State), !is.na(Population)) %>% view()
# Tidy number of vaccinations on given date
vax <- vax %>% pivot_longer(-c(1:12), names_to = "Date", values_to = "Vaccinations", values_drop_na = TRUE) %>% view()
# Delete irrelevant columns
vax <- vax[,-c(1,2,3,4,5,6,7,9,10,11)] %>% view()
# Filter out rows containing dates with 0 vaccinations
vax <- vax %>% filter(!Vaccinations == 0) %>% view()
# Calculate vaccination rate and add respective column
vax <- vax %>% select(Country_Region, Population, Vaccinations) %>% group_by(Country_Region) %>% mutate(Vaccination_Rate = Vaccinations / Population) %>% view()
# Add column that tracks days since first vaccination
vax <- vax %>% group_by(Country_Region) %>% mutate(Days_Since_First_Vaccination = 1:n()) %>% view()
```

Before:

	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	Population	2020-12-12	2020-12-13	2020-12-14	2020-12-15	2020-12-16
1	4	AF	AFG	4	NA	NA	NA	Afghanistan	33.9391	67.7100	Afghanistan	38928341	NA	NA	NA	NA	
2	8	AL	ALB	8	NA	NA	NA	Albania	41.1533	20.1683	Albania	2877800	NA	NA	NA	NA	
3	12	DZ	DZA	12	NA	NA	NA	Algeria	28.0339	1.6596	Algeria	43851043	0	0	0	0	
4	20	AD	AND	20	NA	NA	NA	Andorra	42.5063	1.5218	Andorra	77265	0	0	0	0	
5	24	AO	AGO	24	NA	NA	NA	Angola	-11.2027	17.8739	Angola	32866268	NA	NA	NA	NA	
6	28	AG	ATG	28	NA	NA	NA	Antigua and Barbuda	17.0608	-61.7964	Antigua and Barbuda	97928	NA	NA	NA	NA	
7	32	AR	ARG	32	NA	NA	NA	Argentina	-38.4161	-63.6167	Argentina	45195777	0	0	0	0	
8	NA	NA	NA	NA	NA	NA	NA	Armenia	NA	NA	NA	NA	NA	NA	NA	NA	
9	36	AU	AUS	36	NA	NA	NA	Australia	-25.0000	133.0000	Australia	25459700	NA	NA	NA	NA	
10	40	AT	AUT	40	NA	NA	NA	Austria	47.5162	14.5501	Austria	9006400	0	0	0	0	
11	31	AZ	AZE	31	NA	NA	NA	Azerbaijan	40.1431	47.5769	Azerbaijan	10139175	NA	NA	NA	NA	
12	44	BS	BHS	44	NA	NA	NA	Bahamas	25.0259	-78.0359	Bahamas	393248	NA	NA	NA	NA	
13	48	BH	BHR	48	NA	NA	NA	Bahrain	26.0275	50.5500	Bahrain	1701583	0	0	0	0	
14	50	BD	BGD	50	NA	NA	NA	Bangladesh	23.6850	90.3563	Bangladesh	164689383	0	0	0	0	

After:

	Country_Region	Population	Vaccinations	Vaccination_Rate	Days_Since_First_Vaccination
1	Afghanistan	38928341	8200	0.0002106434	1
2	Afghanistan	38928341	8200	0.0002106434	2
3	Afghanistan	38928341	8200	0.0002106434	3
4	Afghanistan	38928341	8200	0.0002106434	4
5	Afghanistan	38928341	8200	0.0002106434	5
6	Afghanistan	38928341	8200	0.0002106434	6
7	Afghanistan	38928341	8200	0.0002106434	7
8	Afghanistan	38928341	8200	0.0002106434	8
9	Afghanistan	38928341	8200	0.0002106434	9
10	Afghanistan	38928341	8200	0.0002106434	10
11	Afghanistan	38928341	8200	0.0002106434	11
12	Afghanistan	38928341	8200	0.0002106434	12
13	Afghanistan	38928341	8200	0.0002106434	13
14	Afghanistan	38928341	8200	0.0002106434	14
15	Afghanistan	38928341	8200	0.0002106434	15

Hospital Beds Data Wrangling

The hospital beds data did not need much wrangling. The bed data of the most recent year was all that was necessary in this data set. The actual year column was not necessary so I went ahead and dropped that column too.

```r
# BEDS DATA
# Most recent year appears first, keep the first bed value per country using summarize()
# Year column is not needed
beds <- beds %>% group_by(Country) %>% summarize(Beds=first(`Hospital beds (per 10 000 population)`)) %>% view()
```

Before:

	Country	Year	Hospital beds (per 10 000 population)
1	Afghanistan	2017	3.9
2	Afghanistan	2016	5.0
3	Afghanistan	2015	5.0
4	Afghanistan	2014	5.0
5	Afghanistan	2013	5.3
6	Afghanistan	2012	5.3
7	Afghanistan	2011	4.4
8	Afghanistan	2010	4.3
9	Afghanistan	2009	4.2
10	Afghanistan	2008	4.2
11	Afghanistan	2007	4.2
12	Afghanistan	2006	4.2
13	Afghanistan	2005	4.2

After:

	Country	Beds
1	Afghanistan	3.9
2	Albania	28.9
3	Algeria	19.0
4	Angola	8.0
5	Antigua and Barbuda	28.9
6	Argentina	49.9
7	Armenia	41.6
8	Australia	38.4
9	Austria	72.7
10	Azerbaijan	48.2
11	Bahamas	29.6
12	Bahrain	17.4
13	Bangladesh	7.9
14	Barbados	59.7
15	Belarus	108.3

Demographics Data Wrangling

To tidy the demographics data, I gave each series code their own column and gave them their corresponding YR2015 data. The series name and country codes were unnecessary so those columns were dropped.

```
# DEMOGRAPHICS DATA
# Tidy data
demo <- demo %>% pivot_wider(-'Series Name', names_from = 'Series Code', values_from = YR2015) %>% view()
# Add male and female data together
demo <- demo %>% mutate(SP.POP.0014.IN=SP.POP.0014.MA.IN+SP.POP.0014.FE.IN) %>% mutate(SP.POP.80UP=SP.POP.8
# Drop country code and gender specific columns, filter NAs
demo <- demo[,-c(2,6:17)] %>% filter(!is.na(SP.DYN.LE00.IN), !is.na(SP.URB.TOTL), !is.na(SP.POP.0014.IN), !
```

Before:

	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
6	Afghanistan	AFG	Population ages 15-64, male	SP.POP.1564.MA.IN	9.386355e+06
7	Afghanistan	AFG	Population ages 15-64, female	SP.POP.1564.FE.IN	8.730445e+06
8	Afghanistan	AFG	Population ages 0-14, male	SP.POP.0014.MA.IN	7.905639e+06
9	Afghanistan	AFG	Population ages 0-14, female	SP.POP.0014.FE.IN	7.538168e+06
10	Afghanistan	AFG	Mortality rate, adult, female (per 1,000 female adults)	SP.DYN.AMRT.FE	2.067460e+02
11	Afghanistan	AFG	Mortality rate, adult, male (per 1,000 male adults)	SP.DYN.AMRT.MA	2.487240e+02
12	Afghanistan	AFG	Population, female	SP.POP.TOTL.FE.IN	1.672744e+07
13	Afghanistan	AFG	Population, male	SP.POP.TOTL.MA.IN	1.768617e+07
14	Afghanistan	AFG	Population ages 65 and above, female	SP.POP.65UP.FE.IN	4.588240e+05
15	Afghanistan	AFG	Population ages 65 and above, male	SP.POP.65UP.MA.IN	3.941720e+05

After:

Filter

	Country Name	SP.DYN.LE00.IN	SP.URB.TOTL	SP.POP.TOTL	SP.POP.0014.IN	SP.POP.80UP	SP.POP.1564.IN	SP.DYN.AMRT	SP.POP.TOTL.IN	SP.POP.65UP.IN
1	Afghanistan	63.37700	8535606	34413603	15443807	85552	18116800	455.4700	34413603	852996
2	Albania	78.02500	1654503	2880703	537788	66965	1979175	150.4100	2880703	363740
3	Algeria	76.09000	28146511	39728025	11404930	453741	25993589	191.6310	39728025	2329506
4	Angola	59.39800	17691524	27884381	13136043	69363	14113726	485.9310	27884381	634612
5	Antigua and Barbuda	76.48300	23392	93566	21121	1571	64812	260.0050	93566	7634
6	Arab World	71.24957	229821020	396028278	130629537	2689793	248365376	277.0746	396028278	17033367
7	Argentina	76.06800	39467043	43131966	10874072	1095211	27630345	234.3790	43131966	4627549
8	Armenia	74.46700	1845585	2925553	587451	77292	2019878	250.9750	2925553	318224
9	Aruba	75.72500	44979	104341	19515	2103	72164	186.8490	104341	12662
10	Austria	81.19024	4988134	8642699	1220349	436241	5794021	129.3750	8642699	1628329
11	Azerbaijan	72.26600	5279540	9649341	2207181	111882	6888622	249.7940	9649341	553537
12	Bahamas, The	73.08800	309640	374206	89775	4045	259393	317.1780	374206	25038
13	Bahrain	76.76200	1220934	1371851	286027	4282	1053937	133.3680	1371851	31887
14	Bangladesh	71.51400	53608403	156256276	45748814	1372432	102533145	259.5060	156256276	7974318
15	Barbados	78.80100	89161	285324	52163	12005	191259	198.1870	285324	41903

Uniforming country names

Uniforming the country names of the hospital beds and demographics data to match the vaccine data is important so there are not multiple entries for the same country.

```r
# UNIFORMING COUNTRY NAMES TO MATCH VACCINE DATA
beds <- beds %>% mutate(Country = replace(Country, Country == "Iran (Islamic Republic of)", "Iran"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Republic of Korea", "South Korea"))
beds <- beds %>% mutate(Country = replace(Country, Country == "United Kingdom of Great Britain and Northern Ireland", "United Kingdom"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Bolivia (Plurinational State of)", "Bolivia"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Lao People's Democratic Republic", "Laos"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Venezuela (Bolivarian Republic of)", "Venezuela"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Republic of Moldova", "Moldova"))
beds <- beds %>% mutate(Country = replace(Country, Country == "United States of America", "US"))
beds <- beds %>% mutate(Country = replace(Country, Country == "Viet Nam", "Vietnam"))

demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Korea, Rep.", "South Korea"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Iran, Islamic Rep.", "Iran"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Venezuela, RB", "Venezuela"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "St. Vincent and the Grenadines", "Saint Vincent and the Grenadines"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "St. Lucia", "Saint Lucia"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Slovak Republic", "Slovakia"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Czech Republic", "Czechia"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Bahamas, The", "Bahamas"))
demo <- demo %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "United States", "US"))
```

Join/merge data sets

Using inner join, I merged the data of all three sets in respect to country.

```r
# Perform inner joins to merge tables
join <- beds %>% inner_join(vax, by=c(Country="Country_Region")) %>% inner_join(demo, by=c(Country="Country Name")) %>% view()

# Rearrange column order to match example
final <- join[,c(1,5,4,3,6,2,7,8)]
```

	Country	Vaccination_Rate	Vaccinations	Population	Days_Since_First_Vaccination	Beds	SP.DYN.LE00.IN	SP.URB.TOTL
1	Afghanistan	0.0002106434	8200	38928341	1	3.9	63.377	8535606
2	Afghanistan	0.0002106434	8200	38928341	2	3.9	63.377	8535606
3	Afghanistan	0.0002106434	8200	38928341	3	3.9	63.377	8535606
4	Afghanistan	0.0002106434	8200	38928341	4	3.9	63.377	8535606
5	Afghanistan	0.0002106434	8200	38928341	5	3.9	63.377	8535606
6	Afghanistan	0.0002106434	8200	38928341	6	3.9	63.377	8535606
7	Afghanistan	0.0002106434	8200	38928341	7	3.9	63.377	8535606
8	Afghanistan	0.0002106434	8200	38928341	8	3.9	63.377	8535606
9	Afghanistan	0.0002106434	8200	38928341	9	3.9	63.377	8535606
10	Afghanistan	0.0002106434	8200	38928341	10	3.9	63.377	8535606
11	Afghanistan	0.0002106434	8200	38928341	11	3.9	63.377	8535606
12	Afghanistan	0.0002106434	8200	38928341	12	3.9	63.377	8535606
13	Afghanistan	0.0002106434	8200	38928341	13	3.9	63.377	8535606
14	Afghanistan	0.0002106434	8200	38928341	14	3.9	63.377	8535606
15	Afghanistan	0.0002106434	8200	38928341	15	3.9	63.377	8535606

Linear modeling and plots

```r
# PLOTS AND LINEAR MODELS
# Scatterplot of only the most recent vaccination rate for every country and the number of days since first vaccination
forplot <- final %>% group_by(Country) %>% summarize(Days_Since_First_Vaccination=max(Days_Since_First_Vaccination), Vaccination_Rate=last(Vaccination_Rate))
scatter <- ggplot(data=forplot) + geom_point(mapping=aes(x=Days_Since_First_Vaccination, y=Vaccination_Rate))

m1 <- lm(data = final, Vaccination_Rate ~ Days_Since_First_Vaccination)
summary(m1) # R-squared: 0.6125

m2 <- lm(data = final, Vaccination_Rate ~ Days_Since_First_Vaccination + Beds)
summary(m2) # R-squared: 0.6341

m3 <- lm(data = final, Vaccination_Rate ~ Days_Since_First_Vaccination + SP.DYN.LE00.IN)
summary(m3) # R-squared: 0.7473

m4 <- lm(data = final, Vaccination_Rate ~ Days_Since_First_Vaccination + SP.URB.TOTL)
summary(m4) # R-squared: 0.6131

m5 <- lm(data = final, Vaccination_Rate ~ Days_Since_First_Vaccination + SP.URB.TOTL + SP.DYN.LE00.IN)
summary(m5) # R-squared: 0.7478

# Organize 5 models and corresponding R2 values into data frame
df <- data.frame(Model=c("M1", "M2", "M3", "M4", "M5"),
R2=c(summary(m1)$r.squared, summary(m2)$adj.r.squared, summary(m3)$adj.r.squared, summary(m4)$adj.r.squared, summary(m5)$adj.r.squared))

# Create bar plot comparing models and their R2 values
bar <-ggplot(data=df, aes(x=Model, y=R2)) + geom_bar(stat="identity")
bar
```

Dependent variable - Vaccination rate

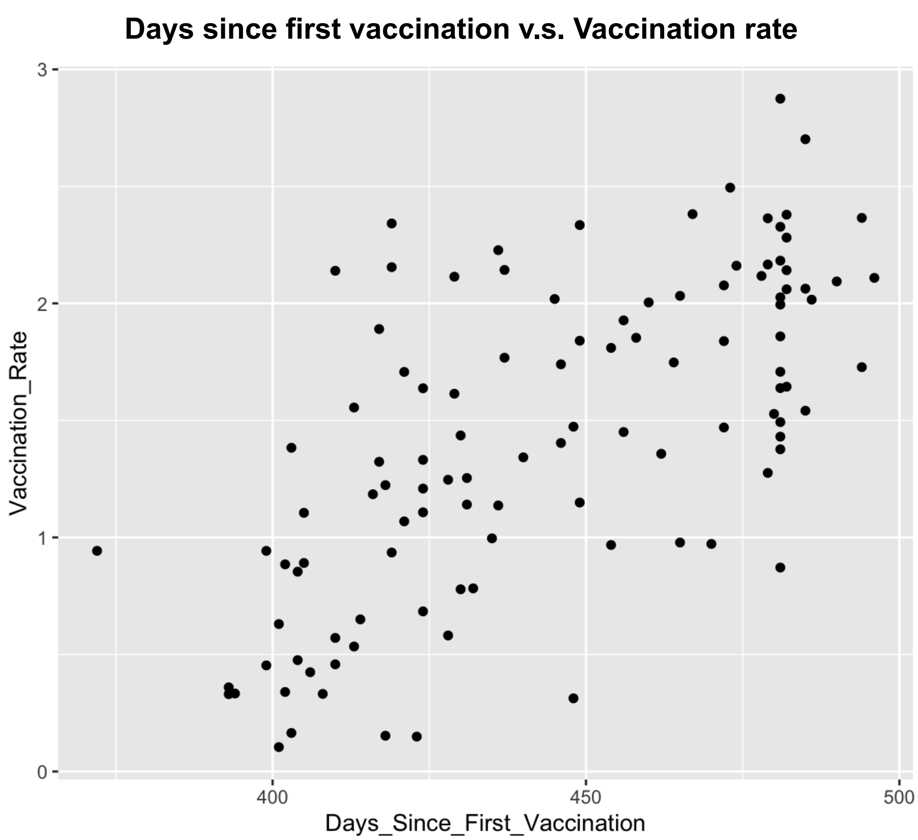
Model 1 (m1) - Days since first vaccination as predictor

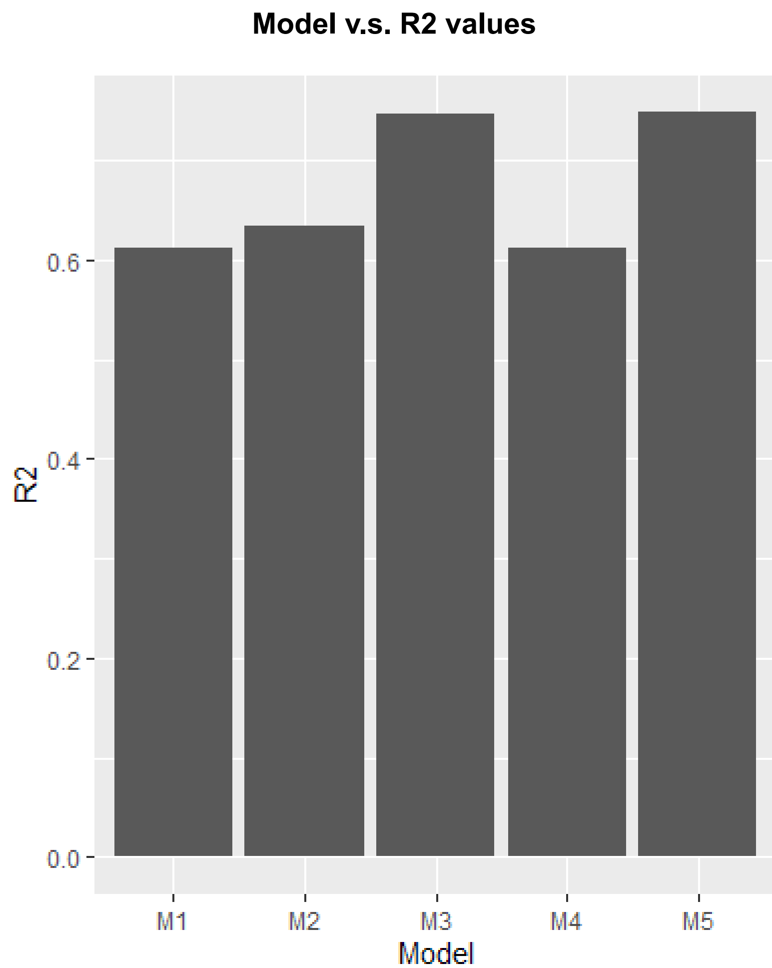
Model 2 (m2) - Days since first vaccination + hospital beds as predictor

Model 3 (m3) - Days since first vaccination + Life expectancy at birth as predictor

Model 4 (m4) - Days since first vaccination + Urban population as predictor

Model 5 (m5) - Days since first vaccination + Urban population + life expectancy at birth as predictor





Conclusion

As seen in the bar plot, it is clear the models that contain the 'life expectancy at birth' (SP.DYN.LE00.IN) predictor are the most accurate. This is implied by their R2 values which are closer to 1 than the models not containing SP.DYN.LE00.IN as a predictor. I believe this is the case because countries that have a higher life expectancy are usually more developed. Vaccines and other medical necessities are a lot more accessible in developed countries compared to underdeveloped countries. This would obviously have an effect on the life expectancy at birth which is why it is the most accurate in predicting the vaccination rate per country.