Data Wrangling and Modeling Report

Group: Miguel Macias

Vaccination Rate Scatter Plot

Comment: A Lot of countries have only been vaccinating for 174 days when the plot was taken, that's why the range of the scatter plot is so wide. I made two scatter plots, one with all countries included and another where the range is 300 days and above to better show the distribution of vaccination rates for those countries.

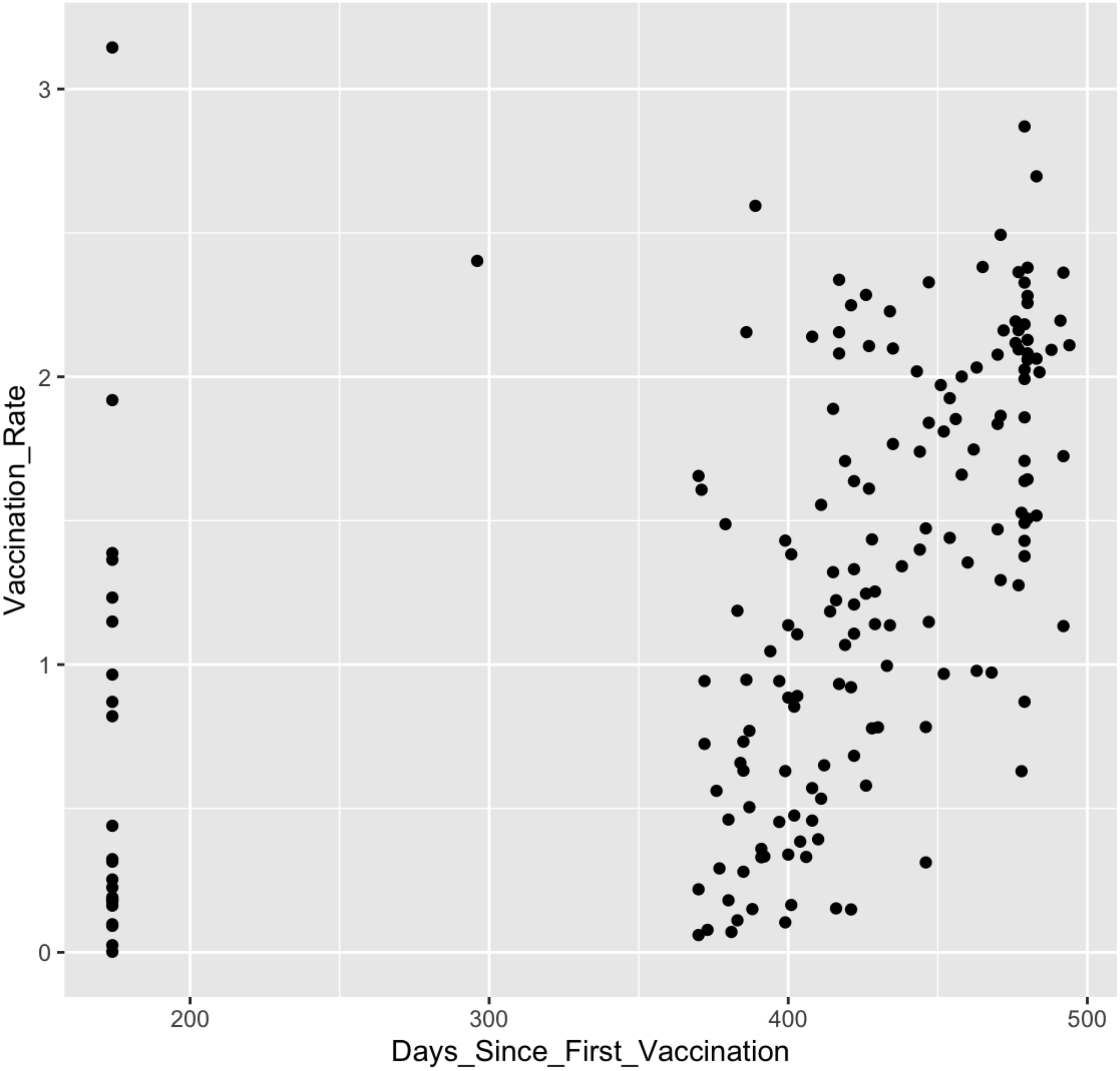


Figure 1. Scatter Plot with all vaccination countries including those who started late

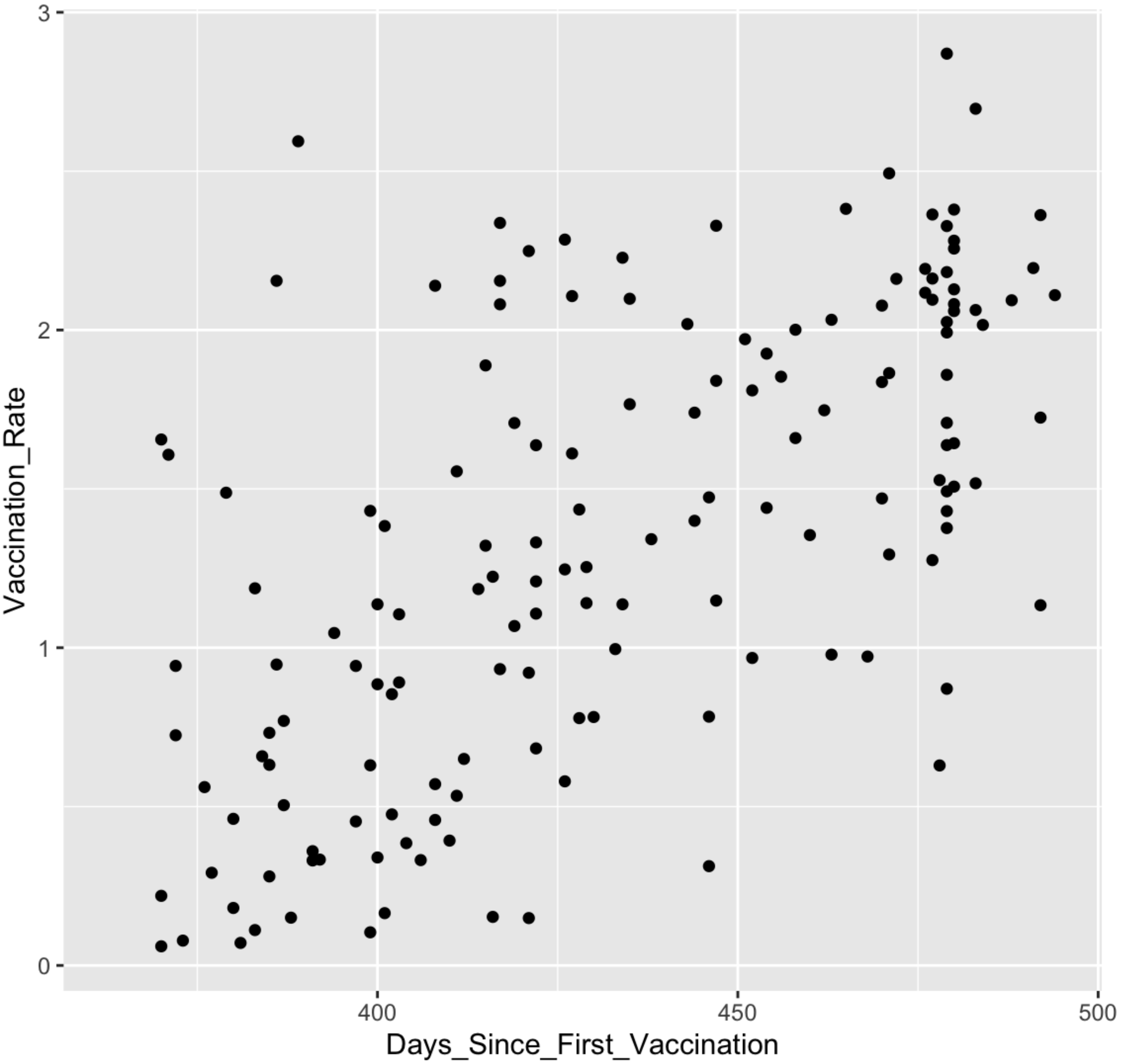


Figure 2. Scatter Plot only includes countries that have started vaccinating for 300 days and more

Modeling

Predictor Variables Chosen

1. **Days_Since_First_Vaccination** - Keeps a running total for how many days since the first vaccine was given. I chose this since I believe that for the model to be accurate it should notice how long a given country has been vaccinating its citizens, vaccination numbers

are always going to go with time as you cannot unvaccinate a person. Especially considering that every data point is a day itself.

2. **Hospital_Beds** - Grabs the latest recorded amount of hospital beds per 10,000 population. I chose this variable as a country that has more hospital beds may usually indicate that the country has a more sophisticated health sector, which may help model their capabilities of giving a vaccination.
3. **Urban_Population**- Takes note of how much of the population lives inside urban cities. I chose this predictor as having a higher denser population may help predict vaccination as it may be easier to give out doses.
4. **Mortality_Rate**- Takes the mortality rate for a given population of male and female adults between 15-60. Lower mortality rates may imply a better health sector, which in turn may help predict a country's ability to distribute vaccines.
5. **Life_Expectancy_At_Birth**- How long an average newborn would be expected to live given that current death rates stay the same. I chose this as a longer life expectancy usually means a better-established health sector in a given country. A better health sector may help indicate how well a country can distribute vaccinations.
6. **Population_80-Up** - How much of the given population is 80 years or older. I chose this predictor because since covid is more lethal to older individuals, there may be a link between vaccination rate and the elderly population.
7. **Population** - How many people live in that country. I chose this as population in a larger country may lead to harder vaccination rate due to the sheer amount of people need to be vaccinated.

Model 1

All Predictors

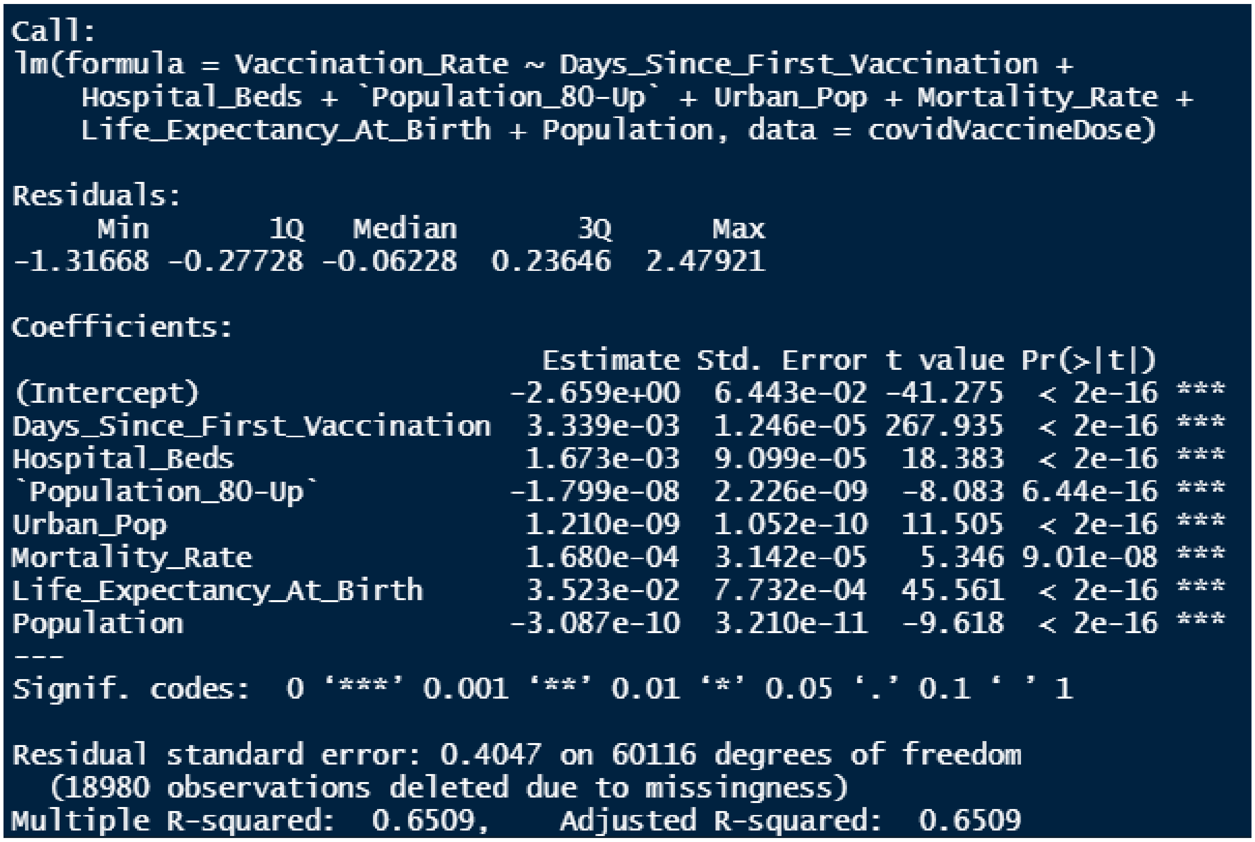
Transformations

None

Equation

$$predictedVacRate\ =\ Predictor1\ +\ Predictor2\ +\ldots+\ Predictor7$$

Summary



```
Call:
lm(formula = Vaccination_Rate ~ Days_Since_First_Vaccination +
    Hospital_Beds + `Population_80-Up` + Urban_Pop + Mortality_Rate +
    Life_Expectancy_At_Birth + Population, data = covidVaccineDose)

Residuals:
     Min       1Q   Median       3Q      Max
-1.31668 -0.27728 -0.06228  0.23646  2.47921

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -2.659e+00  6.443e-02 -41.275  < 2e-16 ***
Days_Since_First_Vaccination  3.339e-03  1.246e-05 267.935  < 2e-16 ***
Hospital_Beds                 1.673e-03  9.099e-05  18.383  < 2e-16 ***
`Population_80-Up`           -1.799e-08  2.226e-09  -8.083 6.44e-16 ***
Urban_Pop                     1.210e-09  1.052e-10  11.505  < 2e-16 ***
Mortality_Rate                1.680e-04  3.142e-05   5.346 9.01e-08 ***
Life_Expectancy_At_Birth      3.523e-02  7.732e-04  45.561  < 2e-16 ***
Population                   -3.087e-10  3.210e-11  -9.618  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4047 on 60116 degrees of freedom
  (18980 observations deleted due to missingness)
Multiple R-squared:  0.6509,    Adjusted R-squared:  0.6509
```

Figure 3 Summary statistics of model 1.

Model 2

3 Predictors: Population_Proportion_80_Up + Hospital_Beds + Days_Since_First_Vaccination

Transformations

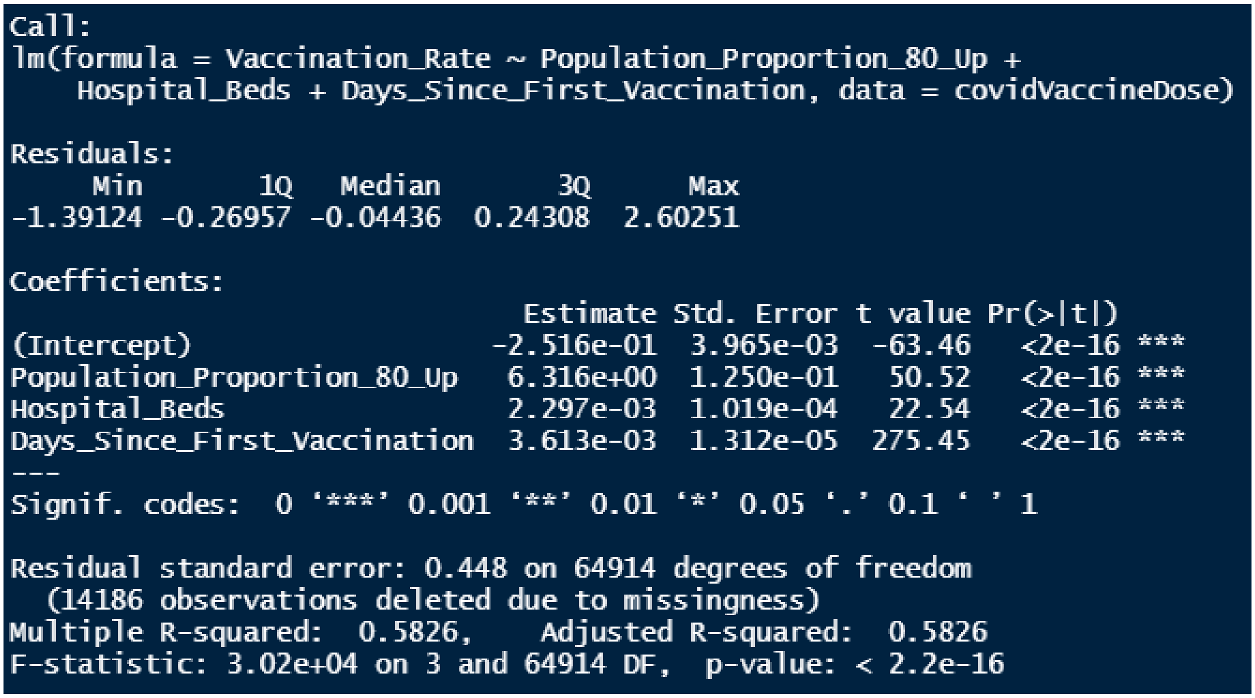
$$Population\_Proportion\_80\_Up \ = \ (Population\_80 \ - \ Up/\ Population)$$

This transformation gets the ratio of how many individuals that are over 80 in a given country.

Equation

$$predictedVacRate \ = \ Population\_Proportion\_80\_Up \ + \ Hospital\_Beds \ + \ Days\_Since\_First\_Vaccination$$

Summary



```
Call:
lm(formula = Vaccination_Rate ~ Population_Proportion_80_Up + 
    Hospital_Beds + Days_Since_First_Vaccination, data = covidVaccineDose)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.39124 -0.26957 -0.04436  0.24308  2.60251 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 -2.516e-01  3.965e-03  -63.46   <2e-16 ***
Population_Proportion_80_Up  6.316e+00  1.250e-01   50.52   <2e-16 ***
Hospital_Beds                2.297e-03  1.019e-04   22.54   <2e-16 ***
Days_Since_First_Vaccination 3.613e-03  1.312e-05  275.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.448 on 64914 degrees of freedom
  (14186 observations deleted due to missingness)
Multiple R-squared:  0.5826,	Adjusted R-squared:  0.5826 
F-statistic: 3.02e+04 on 3 and 64914 DF,  p-value: < 2.2e-16
```

Figure 4 Summary statistics of model 2.

Model 3

3 Predictors: Bed_Proportion_80_Up + Urban_Pop + Mortality_Rate

Transformations

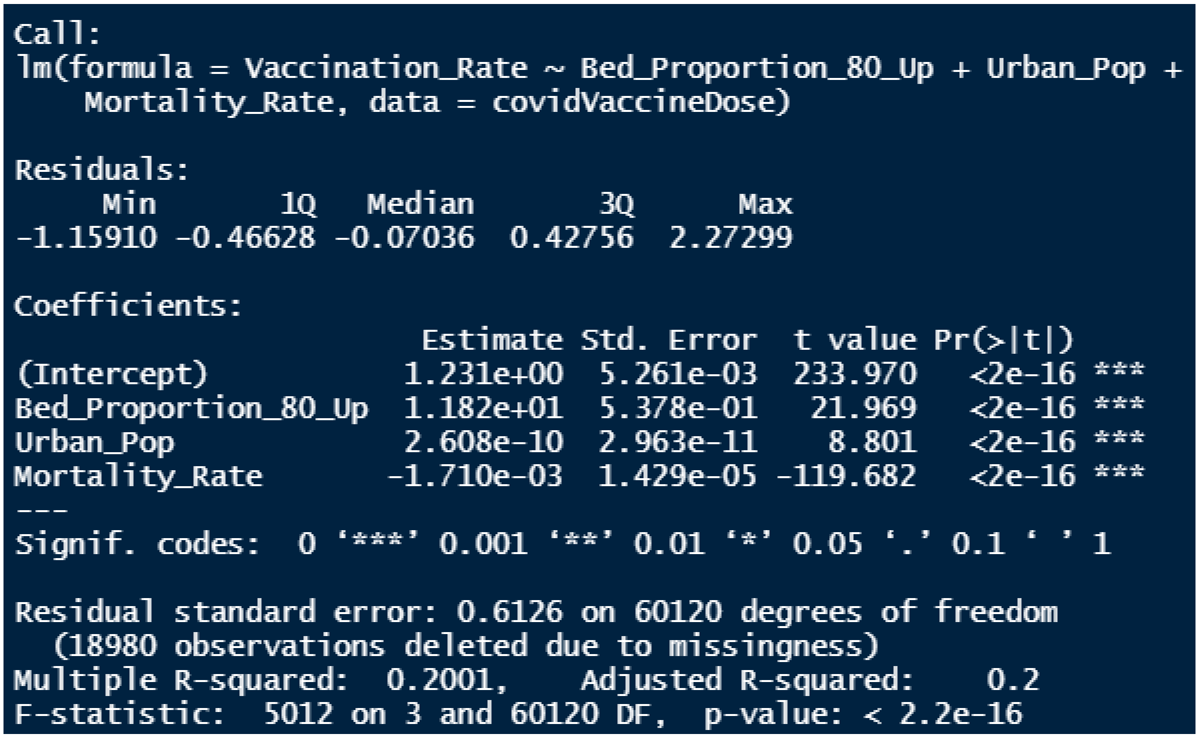
$Bed\_Proportion\_80\_Up = (Hospital\_Beds / Population\_80Up)$

Proportions of how many beds there are for every individual that is 80 years or older

Equation

$$predictedVacRate = Bed\_Proportion\_80\_Up + Urban\_Pop + Mortality\_Rate$$

Summary



```
Call:
lm(formula = Vaccination_Rate ~ Bed_Proportion_80_Up + Urban_Pop +
    Mortality_Rate, data = covidVaccineDose)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15910 -0.46628 -0.07036  0.42756  2.27299

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)           1.231e+00  5.261e-03  233.970   <2e-16 ***
Bed_Proportion_80_Up  1.182e+01  5.378e-01   21.969   <2e-16 ***
Urban_Pop             2.608e-10  2.963e-11    8.801   <2e-16 ***
Mortality_Rate       -1.710e-03  1.429e-05 -119.682   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6126 on 60120 degrees of freedom
  (18980 observations deleted due to missingness)
Multiple R-squared:  0.2001,    Adjusted R-squared:    0.2
F-statistic:  5012 on 3 and 60120 DF,  p-value: < 2.2e-16
```

Figure 5 Summary statistics of model 3.

Model 4

4 Predictors: Days_Since_First_Vaccination + Hospital_BedsSq + Population + Life_Expectancy_At_Birth

Transformations

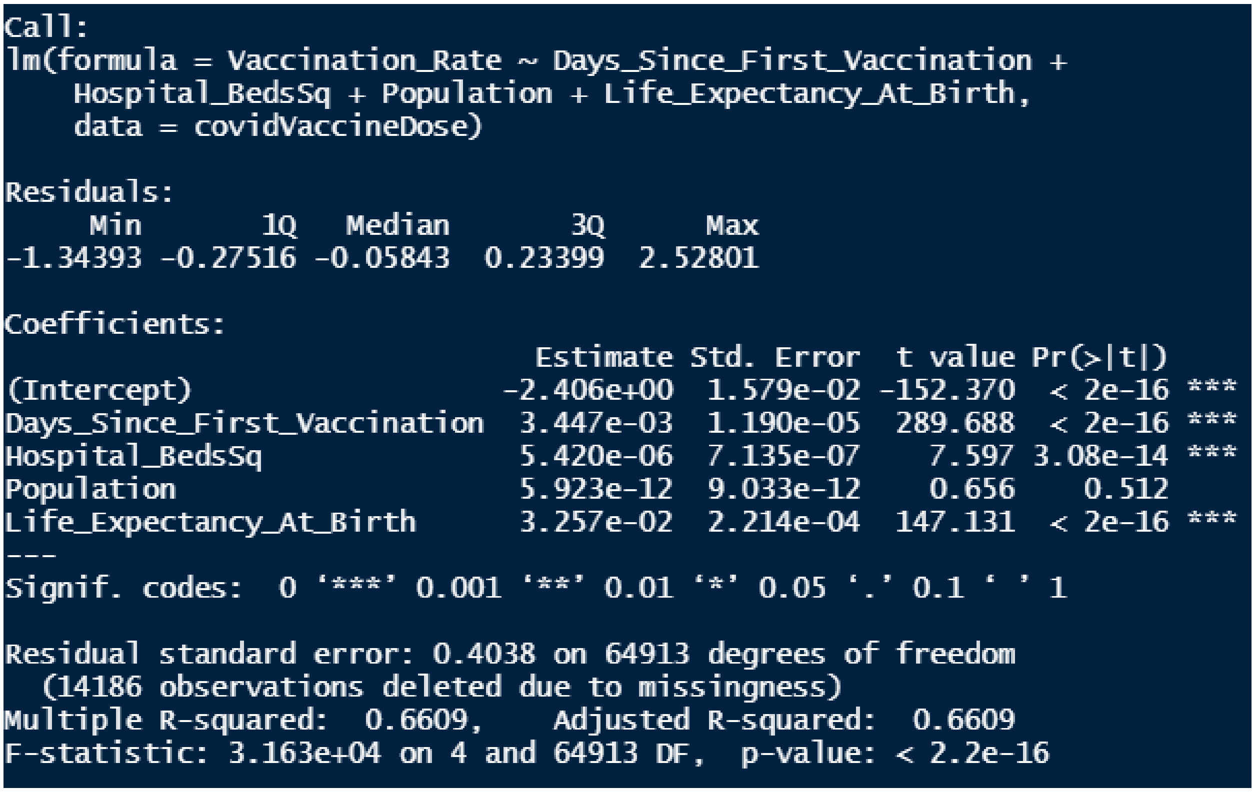
$$Hospital\_BedsSq = Hospital\_Beds^2$$

Squaring hospital beds, attemping to see if non-linearing transforming a variable will make a more accurate model

Equation

$$predictedVacRate = Days\_Since\_First\_Vaccination + Hospital\_BedsSq + Population + Life\_Expectancy\_A$$

Summary



```
Call:
lm(formula = Vaccination_Rate ~ Days_Since_First_Vaccination +
    Hospital_BedsSq + Population + Life_Expectancy_At_Birth,
    data = covidVaccineDose)

Residuals:
     Min       1Q   Median       3Q      Max
-1.34393 -0.27516 -0.05843  0.23399  2.52801

Coefficients:
                               Estimate Std. Error  t value Pr(>|t|)
(Intercept)                  -2.406e+00  1.579e-02 -152.370  < 2e-16 ***
Days_Since_First_Vaccination  3.447e-03  1.190e-05  289.688  < 2e-16 ***
Hospital_BedsSq               5.420e-06  7.135e-07    7.597 3.08e-14 ***
Population                    5.923e-12  9.033e-12    0.656    0.512
Life_Expectancy_At_Birth      3.257e-02  2.214e-04  147.131  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4038 on 64913 degrees of freedom
  (14186 observations deleted due to missingness)
Multiple R-squared:  0.6609,    Adjusted R-squared:  0.6609
F-statistic: 3.163e+04 on 4 and 64913 DF,  p-value: < 2.2e-16
```

Figure 6 Summary statistics of model 4.

Model 5

4 Predictors: Hospital_BedsSq + Bed_Proportion_80_Up + Urban_Pop + `Population_80-Up`

Transformations

$Hospital\_BedsSq \;=\; Hospital\_Beds^{2}$

Squaring hospital beds, attemping to see if non-linearing transforming a variable will make a more accurate model

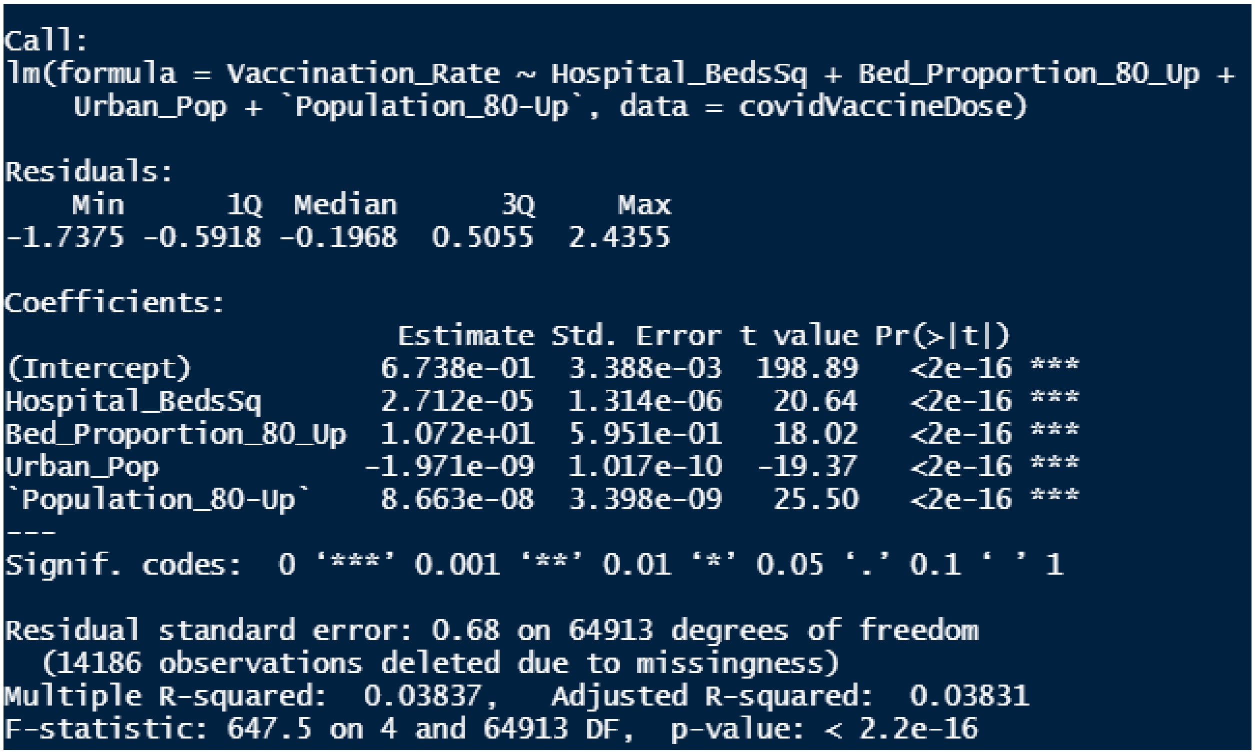
$Bed\_Proportion\_80\_Up \;=\; (Hospital\_Beds\,/\,Population\_80Up)$

Proportions of how many beds there are for every individual that is 80 years or older

Equation

$$predictedVacRate \;=\; Hospital\_BedsSq \;+\; Bed\_Proportion\_80\_Up \;+\; Urban\_Pop \;+\; `Population\_80 - Up`$$

Summary



```
Call:
lm(formula = Vaccination_Rate ~ Hospital_BedsSq + Bed_Proportion_80_Up +
    Urban_Pop + `Population_80-Up`, data = covidVaccineDose)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7375 -0.5918 -0.1968  0.5055  2.4355

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.738e-01  3.388e-03  198.89   <2e-16 ***
Hospital_BedsSq       2.712e-05  1.314e-06   20.64   <2e-16 ***
Bed_Proportion_80_Up  1.072e+01  5.951e-01   18.02   <2e-16 ***
Urban_Pop            -1.971e-09  1.017e-10  -19.37   <2e-16 ***
`Population_80-Up`    8.663e-08  3.398e-09   25.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.68 on 64913 degrees of freedom
  (14186 observations deleted due to missingness)
Multiple R-squared:  0.03837,	Adjusted R-squared:  0.03831
F-statistic: 647.5 on 4 and 64913 DF,  p-value: < 2.2e-16
```

Figure 7 Summary statistics of model 5.

Model 6

5 Predictors: Days_Since_First_Vaccination+ Hospital_BedsSq + Bed_Proportion_80_Up + Urban_Pop + `Population_80-Up`

Transformations

$Hospital\_BedsSq \ = \ Hospital\_Beds^2$

Squaring hospital beds, attempting to see if non-linearing transforming a variable will make a more accurate model

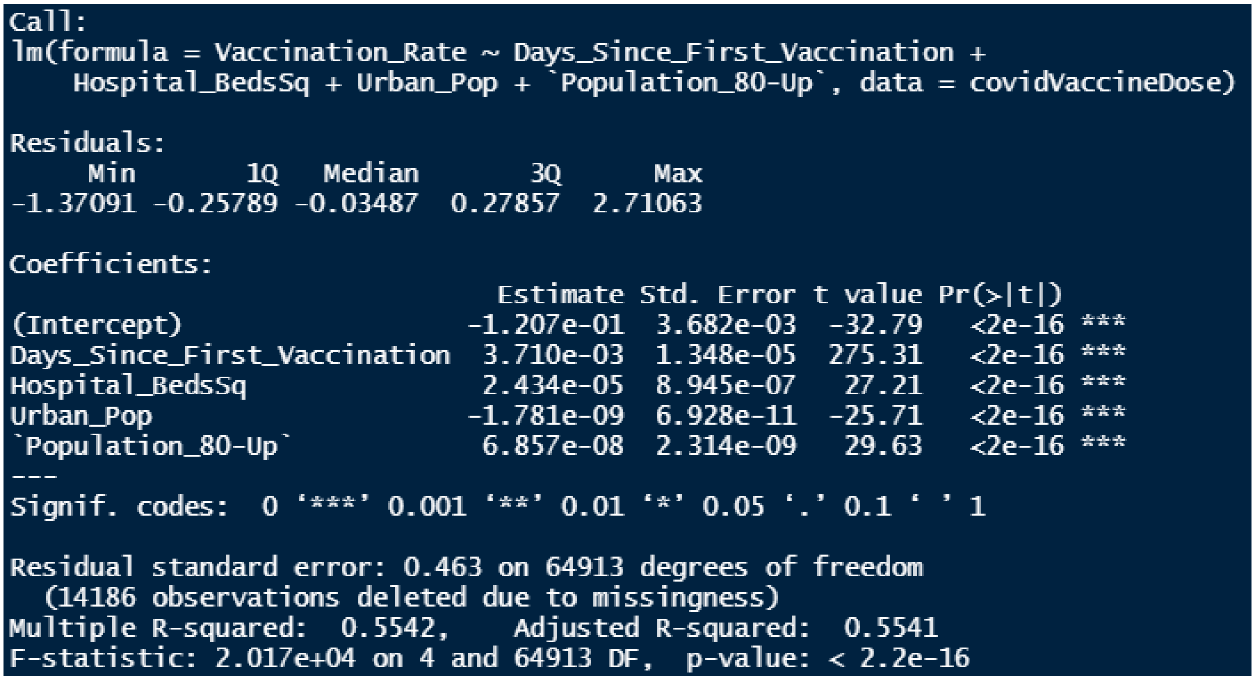
$Bed\_Proportion\_80\_Up \ = \ (Hospital\_Beds \,/\, Population\_80Up)$

Proportions of how many beds there are for every individual that is 80 years or older

Equation

$$predictedVacRate \ = \ Days\_Since\_First\_Vaccination \ + \ Hospital\_BedsSq \ + \ Bed\_Proportion\_80\_Up \ + \ Urban\_Pop \ + \ `Population\_80Up$$

Summary



```
Call:
lm(formula = Vaccination_Rate ~ Days_Since_First_Vaccination + 
    Hospital_BedsSq + Urban_Pop + `Population_80-Up`, data = covidVaccineDose)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.37091 -0.25789 -0.03487  0.27857  2.71063 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  -1.207e-01  3.682e-03  -32.79   <2e-16 ***
Days_Since_First_Vaccination  3.710e-03  1.348e-05  275.31   <2e-16 ***
Hospital_BedsSq               2.434e-05  8.945e-07   27.21   <2e-16 ***
Urban_Pop                    -1.781e-09  6.928e-11  -25.71   <2e-16 ***
`Population_80-Up`            6.857e-08  2.314e-09   29.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.463 on 64913 degrees of freedom
  (14186 observations deleted due to missingness)
Multiple R-squared:  0.5542,	Adjusted R-squared:  0.5541 
F-statistic: 2.017e+04 on 4 and 64913 DF,  p-value: < 2.2e-16
```

Figure 8 Summary statistics of model 6.

Comparison Among All Model's Adjusted R ^{2}

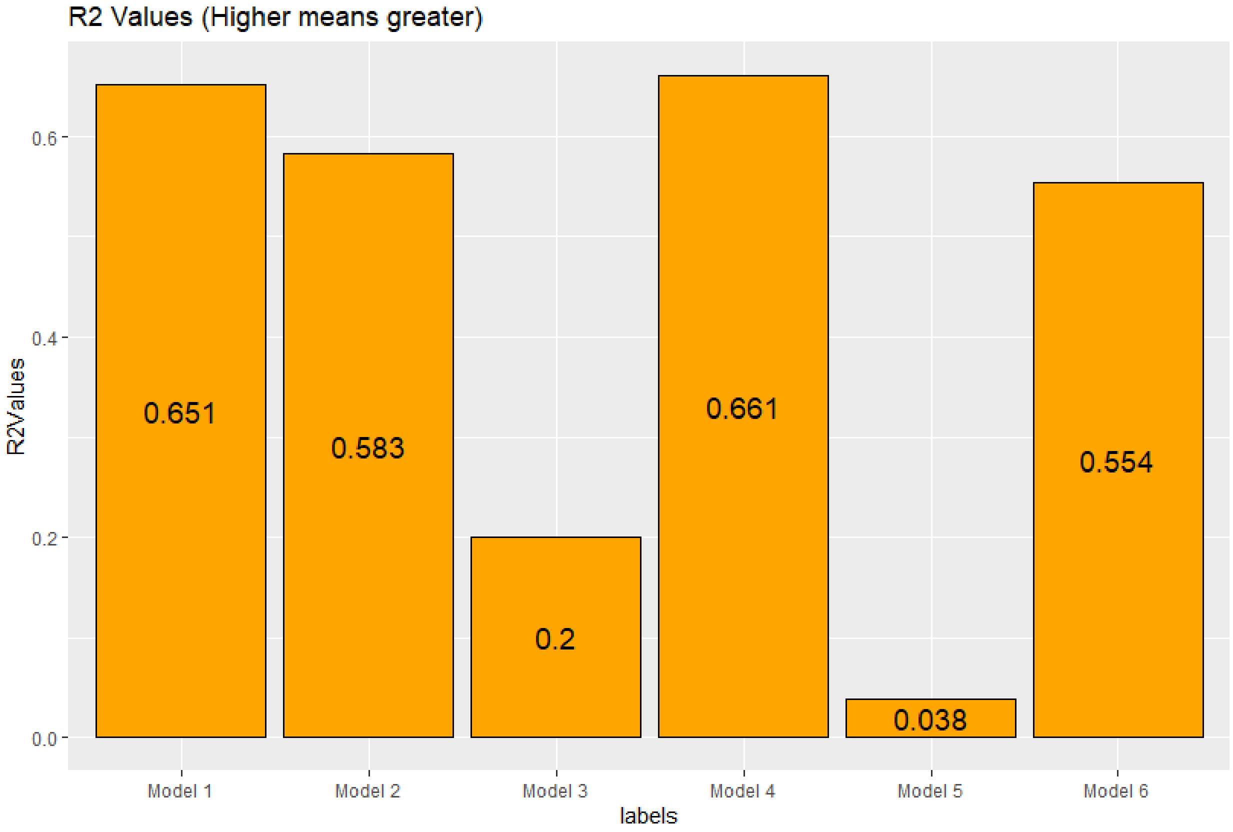


Figure 9 Boxplot of all models adjusted R ^{2}.

Conclusion

After running six different models with varying predictor sizes and combinations, we noticed that the ones that accounted for how long a country has been vaccinating were significantly more accurate than those without it. This is expected given that each entry it is trying to predict is a different day. Look at models 5 and 6. By simply adding how long a country has been vaccinating in model 6, we significantly see increases in our adjusted R^2. For that reason, we would consider knowing how long a country has been vaccinating to be the most significant factor. However, I was surprised that model 3 could land an adjusted R^2 of .2 without considering what day it was for them. As for the insignificant factors, these are much harder to spot. I would say that taking into account the proportion of the elderly population and the ratio between beds and people did yield lower results. Another significant factor to note would be the variable size. Even model 1, with one of the highest adjusted R^2, was simply all the predictors together with no transformations done to them. I wonder if this can be due to overfitting, given that we did use the most predictors here. However, our adjusted R^2 did not change much from our original R^2. I would say that our models did show us how hard it is to get an accurate model for vaccination rates. Given that our highest accurate models stayed relatively close to each other, which makes us believe that there are significantly more, or different, variables that we would need to look into for making better models.

Things to look at

Given that some countries when inserted into the data sets did not start at zero, I would advise to separate those models. As this created a giant line of data points on the left hand side which may skew our model(See Figure 1). I also noticed that all the countries that showed these characteristics were from the Africa continent, so separating the countries into their down dataset based on their respective continent may yield a better model.

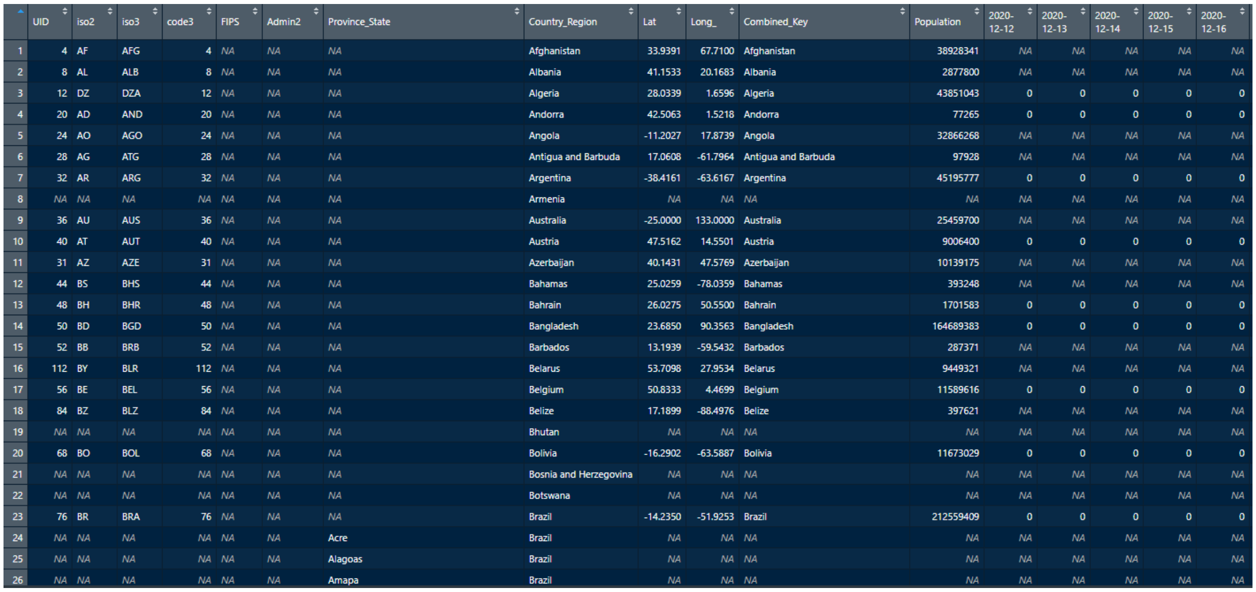
Data Wrangling Steps

#Tidying up Vaccination Doses
#Removing all information that won't be needed for modeling

```
covidVaccineDose <- covidVaccineDose %>%
select(-c(FIPS,Admin2,Lat,Long_,UID,iso2,iso3,code3, Combined_Key))
```

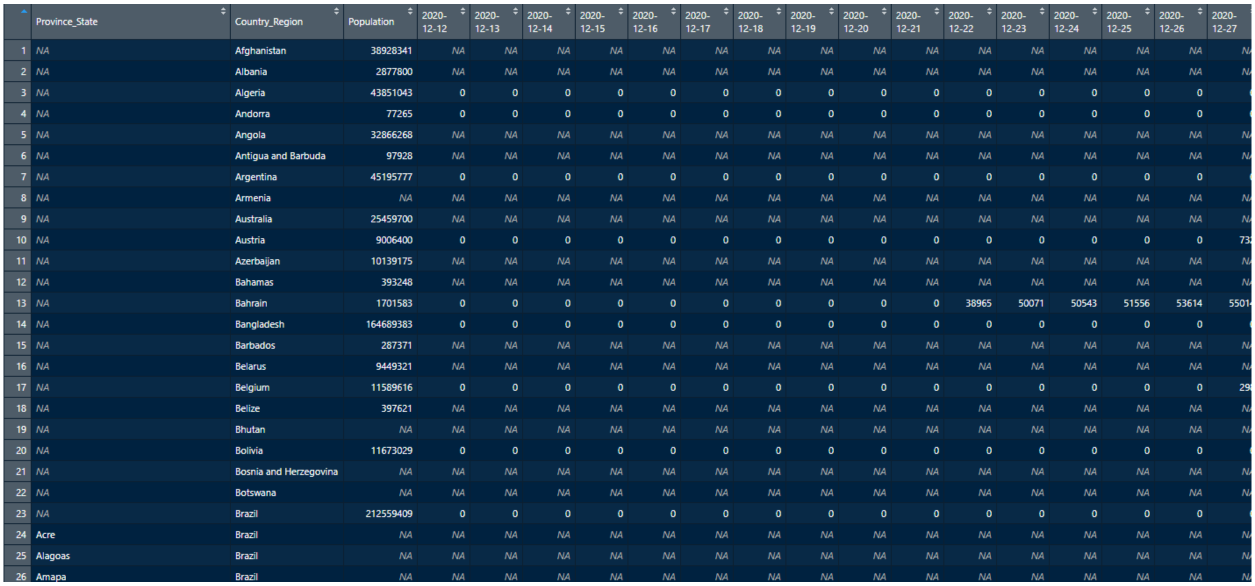
This line of r code was meant to remove columns that we did not deem to be necessary to model or would help us tidy the data.

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	NA
2	8	AL	ALB	8	NA	NA	NA	Albania	41.1533	20.1683	Albania	2877800	NA	NA	NA	NA	NA
3	12	DZ	DZA	12	NA	NA	NA	Algeria	28.0339	1.6596	Algeria	43851043	0	0	0	0	0
4	20	AD	AND	20	NA	NA	NA	Andorra	42.5063	1.5218	Andorra	77265	0	0	0	0	0
5	24	AO	AGO	24	NA	NA	NA	Angola	-11.2027	17.8739	Angola	32866268	NA	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	NA
7	32	AR	ARG	32	NA	NA	NA	Argentina	-38.4161	-63.6167	Argentina	45195777	0	0	0	0	0
8	NA	NA	NA	NA	NA	NA	NA	Armenia	NA	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	NA
10	40	AT	AUT	40	NA	NA	NA	Austria	47.5162	14.5501	Austria	9006400	0	0	0	0	0
11	31	AZ	AZE	31	NA	NA	NA	Azerbaijan	40.1431	47.5769	Azerbaijan	10139175	NA	NA	NA	NA	NA
12	44	BS	BHS	44	NA	NA	NA	Bahamas	25.0259	-78.0359	Bahamas	393248	NA	NA	NA	NA	NA
13	48	BH	BHR	48	NA	NA	NA	Bahrain	26.0275	50.5500	Bahrain	1701583	0	0	0	0	0
14	50	BD	BGD	50	NA	NA	NA	Bangladesh	23.6850	90.3563	Bangladesh	164689383	0	0	0	0	0
15	52	BB	BRB	52	NA	NA	NA	Barbados	13.1939	-59.5432	Barbados	287371	NA	NA	NA	NA	NA
16	112	BY	BLR	112	NA	NA	NA	Belarus	53.7098	27.9534	Belarus	9449321	NA	NA	NA	NA	NA
17	56	BE	BEL	56	NA	NA	NA	Belgium	50.8333	4.4699	Belgium	11589616	0	0	0	0	0
18	84	BZ	BLZ	84	NA	NA	NA	Belize	17.1899	-88.4976	Belize	397621	NA	NA	NA	NA	NA
19	NA	NA	NA	NA	NA	NA	NA	Bhutan	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	68	BO	BOL	68	NA	NA	NA	Bolivia	-16.2902	-63.5887	Bolivia	11673029	0	0	0	0	0
21	NA	NA	NA	NA	NA	NA	NA	Bosnia and Herzegovina	NA	NA	NA	NA	NA	NA	NA	NA	NA
22	NA	NA	NA	NA	NA	NA	NA	Botswana	NA	NA	NA	NA	NA	NA	NA	NA	NA
23	76	BR	BRA	76	NA	NA	NA	Brazil	-14.2350	-51.9253	Brazil	212559409	0	0	0	0	0
24	NA	NA	NA	NA	NA	NA	Acre	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA
25	NA	NA	NA	NA	NA	NA	Alagoas	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA
26	NA	NA	NA	NA	NA	NA	Amapa	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA

After:

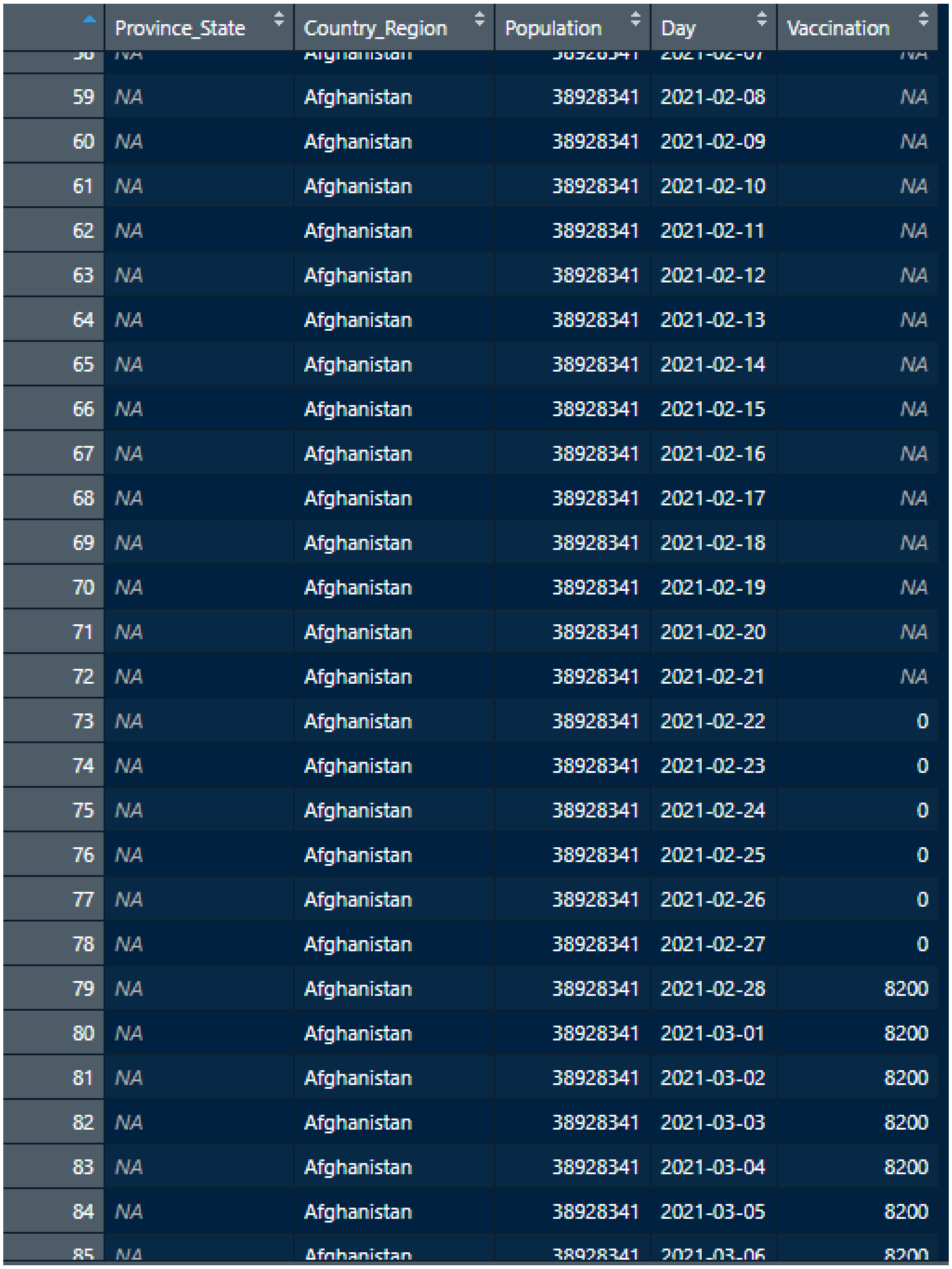


	Province_State	Country_Region	Population	2020-12-12	2020-12-13	2020-12-14	2020-12-15	2020-12-16	2020-12-17	2020-12-18	2020-12-19	2020-12-20	2020-12-21	2020-12-22	2020-12-23	2020-12-24	2020-12-25	2020-12-26	2020-12-27
1	NA	Afghanistan	38928341	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
2	NA	Albania	2877800	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
3	NA	Algeria	43851043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	NA	Andorra	77265	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	NA	Angola	32866268	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
6	NA	Antigua and Barbuda	97928	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
7	NA	Argentina	45195777	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	NA	Armenia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
9	NA	Australia	25459700	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
10	NA	Austria	9006400	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	73
11	NA	Azerbaijan	10139175	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
12	NA	Bahamas	393248	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
13	NA	Bahrain	1701583	0	0	0	0	0	0	0	0	0	0	38965	50071	50543	51556	53614	5501
14	NA	Bangladesh	164689383	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	NA	Barbados	287371	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
16	NA	Belarus	9449321	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
17	NA	Belgium	11589616	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29
18	NA	Belize	397621	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
19	NA	Bhutan	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
20	NA	Bolivia	11673029	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21	NA	Bosnia and Herzegovina	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
22	NA	Botswana	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
23	NA	Brazil	212559409	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
24	Acre	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
25	Alagoas	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
26	Amapa	Brazil	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N

#Filter where Province_State == NA as that is the overall vaccination rate for that country.
#Then make the dataframe longer by giving every day a separate row

```
covidVaccineDose <- covidVaccineDose %>% filter(is.na(Province_State)) %>%
pivot_longer(4:496, names_to = 'Day', values_to = 'Vaccination')
```

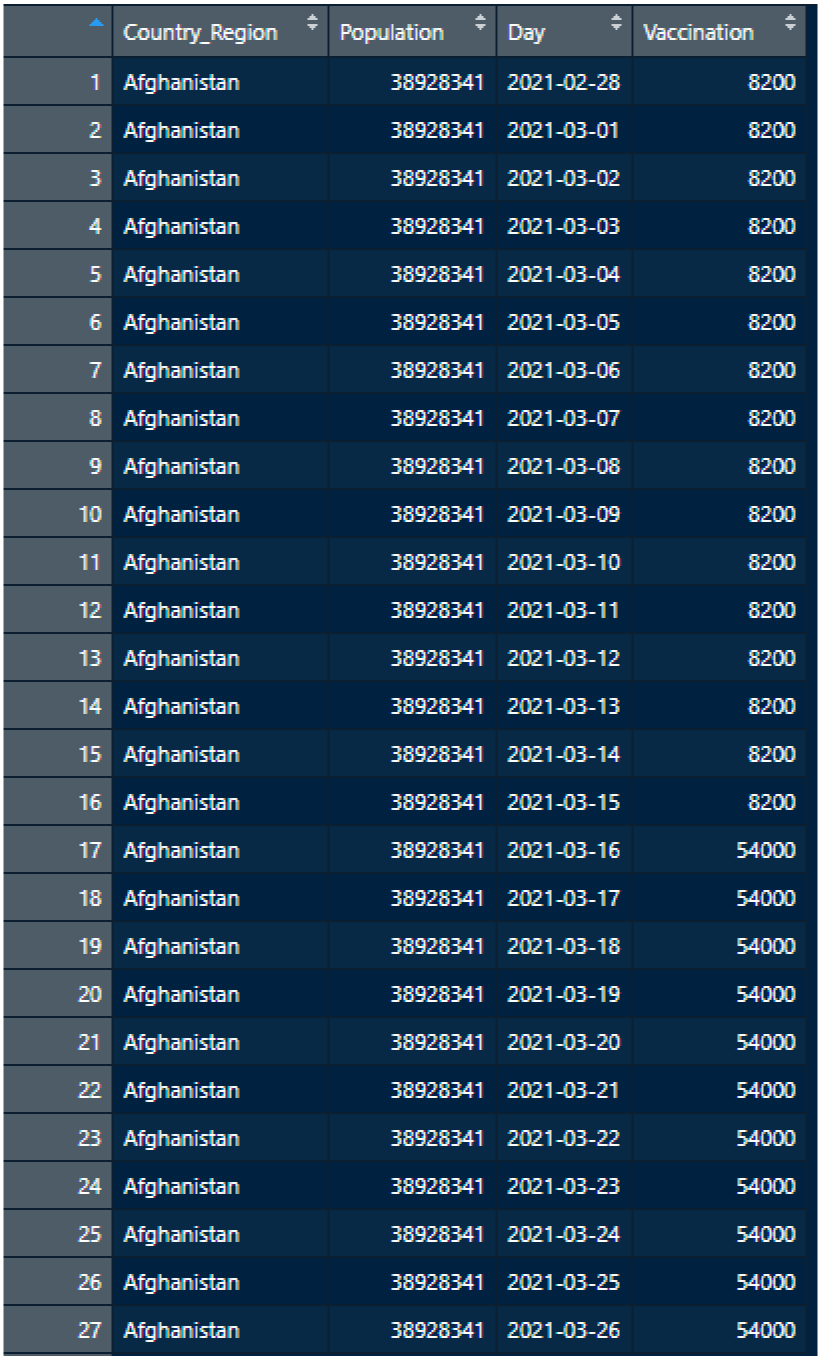
I noticed that every time province_state was NA it was going over the overall vaccination rates for that country; therefore we filtered only those that are NA, and then pivot the dataframe longer.



	Province_State	Country_Region	Population	Day	Vaccination
58	NA	Afghanistan	38928341	2021-02-07	NA
59	NA	Afghanistan	38928341	2021-02-08	NA
60	NA	Afghanistan	38928341	2021-02-09	NA
61	NA	Afghanistan	38928341	2021-02-10	NA
62	NA	Afghanistan	38928341	2021-02-11	NA
63	NA	Afghanistan	38928341	2021-02-12	NA
64	NA	Afghanistan	38928341	2021-02-13	NA
65	NA	Afghanistan	38928341	2021-02-14	NA
66	NA	Afghanistan	38928341	2021-02-15	NA
67	NA	Afghanistan	38928341	2021-02-16	NA
68	NA	Afghanistan	38928341	2021-02-17	NA
69	NA	Afghanistan	38928341	2021-02-18	NA
70	NA	Afghanistan	38928341	2021-02-19	NA
71	NA	Afghanistan	38928341	2021-02-20	NA
72	NA	Afghanistan	38928341	2021-02-21	NA
73	NA	Afghanistan	38928341	2021-02-22	0
74	NA	Afghanistan	38928341	2021-02-23	0
75	NA	Afghanistan	38928341	2021-02-24	0
76	NA	Afghanistan	38928341	2021-02-25	0
77	NA	Afghanistan	38928341	2021-02-26	0
78	NA	Afghanistan	38928341	2021-02-27	0
79	NA	Afghanistan	38928341	2021-02-28	8200
80	NA	Afghanistan	38928341	2021-03-01	8200
81	NA	Afghanistan	38928341	2021-03-02	8200
82	NA	Afghanistan	38928341	2021-03-03	8200
83	NA	Afghanistan	38928341	2021-03-04	8200
84	NA	Afghanistan	38928341	2021-03-05	8200
85	NA	Afghanistan	38928341	2021-03-06	8200

```
#Remove Province_State and drop all the NA Vaccinations or Vaccination is 0
covidVaccineDose <- covidVaccineDose %>% select(-c(Province_State)) %>%
filter(!is.na(Vaccination)) %>% filter(Vaccination != 0)
```

I then removed the province_state column and removed all vaccination dates that are NA or 0



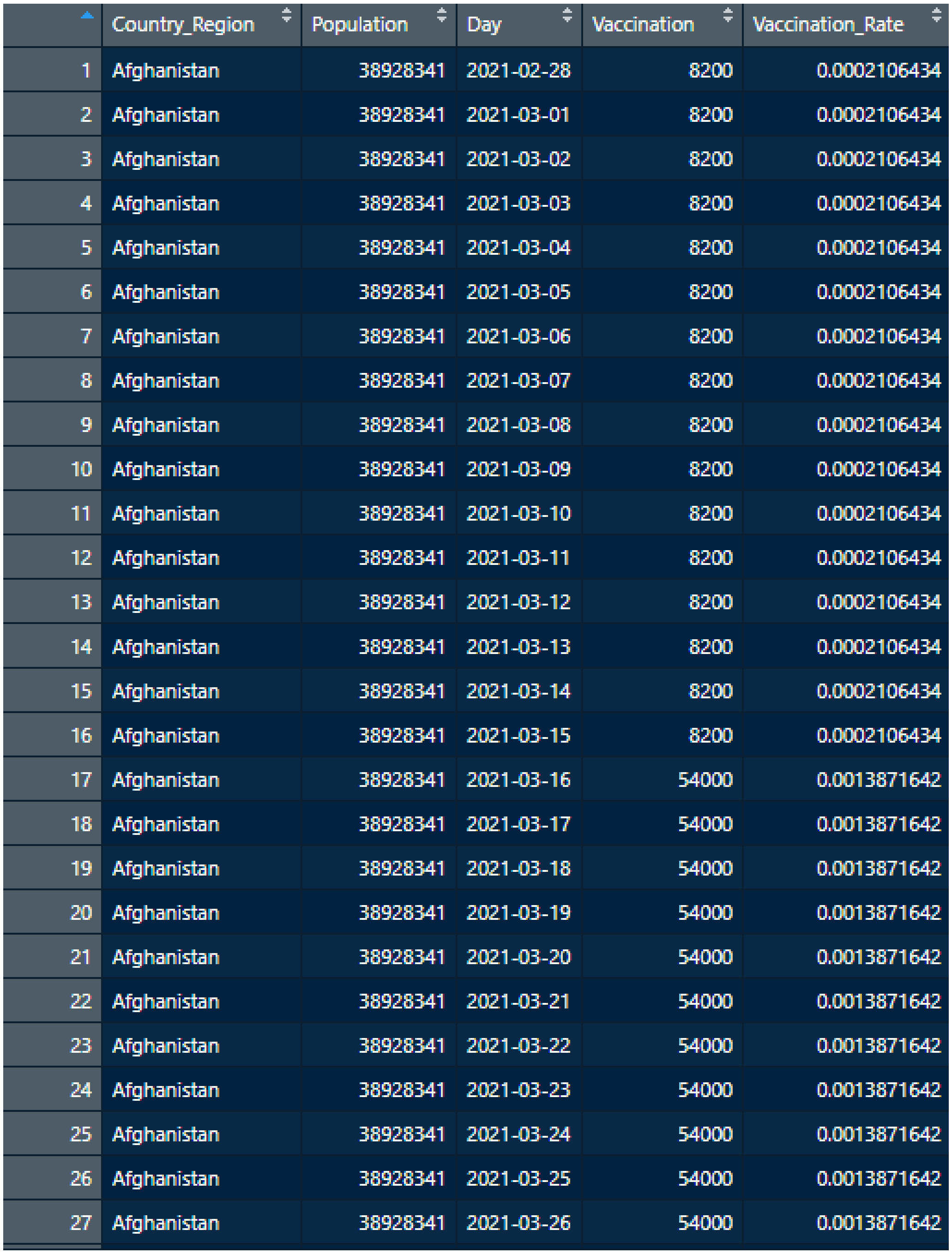
	Country_Region	Population	Day	Vaccination
1	Afghanistan	38928341	2021-02-28	8200
2	Afghanistan	38928341	2021-03-01	8200
3	Afghanistan	38928341	2021-03-02	8200
4	Afghanistan	38928341	2021-03-03	8200
5	Afghanistan	38928341	2021-03-04	8200
6	Afghanistan	38928341	2021-03-05	8200
7	Afghanistan	38928341	2021-03-06	8200
8	Afghanistan	38928341	2021-03-07	8200
9	Afghanistan	38928341	2021-03-08	8200
10	Afghanistan	38928341	2021-03-09	8200
11	Afghanistan	38928341	2021-03-10	8200
12	Afghanistan	38928341	2021-03-11	8200
13	Afghanistan	38928341	2021-03-12	8200
14	Afghanistan	38928341	2021-03-13	8200
15	Afghanistan	38928341	2021-03-14	8200
16	Afghanistan	38928341	2021-03-15	8200
17	Afghanistan	38928341	2021-03-16	54000
18	Afghanistan	38928341	2021-03-17	54000
19	Afghanistan	38928341	2021-03-18	54000
20	Afghanistan	38928341	2021-03-19	54000
21	Afghanistan	38928341	2021-03-20	54000
22	Afghanistan	38928341	2021-03-21	54000
23	Afghanistan	38928341	2021-03-22	54000
24	Afghanistan	38928341	2021-03-23	54000
25	Afghanistan	38928341	2021-03-24	54000
26	Afghanistan	38928341	2021-03-25	54000
27	Afghanistan	38928341	2021-03-26	54000

#Get vaccination rate per population, vaccination/population

```
covidVaccineDose <- covidVaccineDose %>% mutate(Vaccination_Rate = Vaccination/Population)
```

I moved this step further in the R file as I get missing population values and recalculate Vaccincation rate. This should not affect the outcome of vaccination rate however

I then added a column that gets vaccination rate on a country on a given date

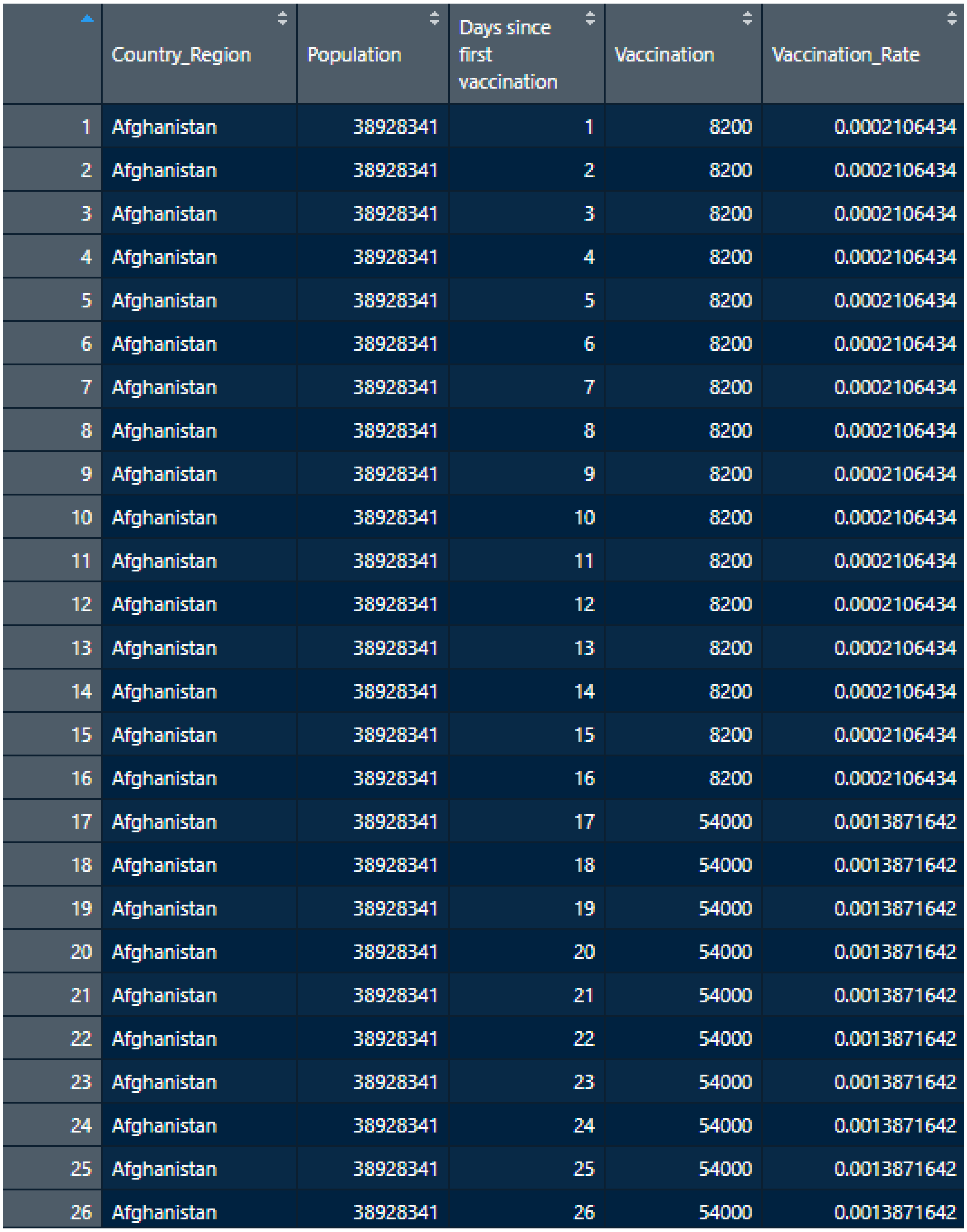


	Country_Region	Population	Day	Vaccination	Vaccination_Rate
1	Afghanistan	38928341	2021-02-28	8200	0.0002106434
2	Afghanistan	38928341	2021-03-01	8200	0.0002106434
3	Afghanistan	38928341	2021-03-02	8200	0.0002106434
4	Afghanistan	38928341	2021-03-03	8200	0.0002106434
5	Afghanistan	38928341	2021-03-04	8200	0.0002106434
6	Afghanistan	38928341	2021-03-05	8200	0.0002106434
7	Afghanistan	38928341	2021-03-06	8200	0.0002106434
8	Afghanistan	38928341	2021-03-07	8200	0.0002106434
9	Afghanistan	38928341	2021-03-08	8200	0.0002106434
10	Afghanistan	38928341	2021-03-09	8200	0.0002106434
11	Afghanistan	38928341	2021-03-10	8200	0.0002106434
12	Afghanistan	38928341	2021-03-11	8200	0.0002106434
13	Afghanistan	38928341	2021-03-12	8200	0.0002106434
14	Afghanistan	38928341	2021-03-13	8200	0.0002106434
15	Afghanistan	38928341	2021-03-14	8200	0.0002106434
16	Afghanistan	38928341	2021-03-15	8200	0.0002106434
17	Afghanistan	38928341	2021-03-16	54000	0.0013871642
18	Afghanistan	38928341	2021-03-17	54000	0.0013871642
19	Afghanistan	38928341	2021-03-18	54000	0.0013871642
20	Afghanistan	38928341	2021-03-19	54000	0.0013871642
21	Afghanistan	38928341	2021-03-20	54000	0.0013871642
22	Afghanistan	38928341	2021-03-21	54000	0.0013871642
23	Afghanistan	38928341	2021-03-22	54000	0.0013871642
24	Afghanistan	38928341	2021-03-23	54000	0.0013871642
25	Afghanistan	38928341	2021-03-24	54000	0.0013871642
26	Afghanistan	38928341	2021-03-25	54000	0.0013871642
27	Afghanistan	38928341	2021-03-26	54000	0.0013871642

#Group by Country_Region, add by one for each row

```
covidVaccineDose <- covidVaccineDose %>% group_by(Country_Region) %>% mutate(Day = row_number()) %>% rename("Days_Since_First_Vaccination" = Day)
```

Group by country, when then increment days since first vaccination by one, which resets when a new country is found in the group.



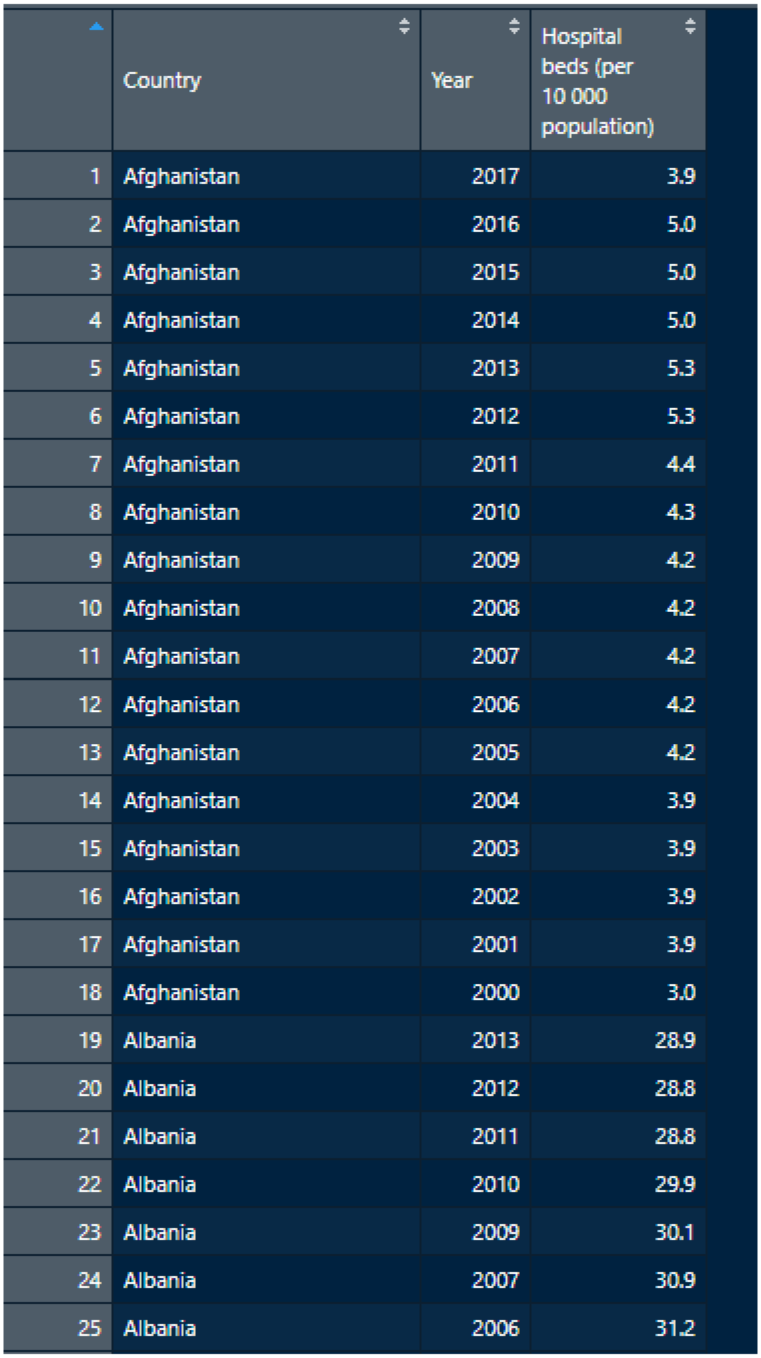
	Country_Region	Population	Days since first vaccination	Vaccination	Vaccination_Rate
1	Afghanistan	38928341	1	8200	0.0002106434
2	Afghanistan	38928341	2	8200	0.0002106434
3	Afghanistan	38928341	3	8200	0.0002106434
4	Afghanistan	38928341	4	8200	0.0002106434
5	Afghanistan	38928341	5	8200	0.0002106434
6	Afghanistan	38928341	6	8200	0.0002106434
7	Afghanistan	38928341	7	8200	0.0002106434
8	Afghanistan	38928341	8	8200	0.0002106434
9	Afghanistan	38928341	9	8200	0.0002106434
10	Afghanistan	38928341	10	8200	0.0002106434
11	Afghanistan	38928341	11	8200	0.0002106434
12	Afghanistan	38928341	12	8200	0.0002106434
13	Afghanistan	38928341	13	8200	0.0002106434
14	Afghanistan	38928341	14	8200	0.0002106434
15	Afghanistan	38928341	15	8200	0.0002106434
16	Afghanistan	38928341	16	8200	0.0002106434
17	Afghanistan	38928341	17	54000	0.0013871642
18	Afghanistan	38928341	18	54000	0.0013871642
19	Afghanistan	38928341	19	54000	0.0013871642
20	Afghanistan	38928341	20	54000	0.0013871642
21	Afghanistan	38928341	21	54000	0.0013871642
22	Afghanistan	38928341	22	54000	0.0013871642
23	Afghanistan	38928341	23	54000	0.0013871642
24	Afghanistan	38928341	24	54000	0.0013871642
25	Afghanistan	38928341	25	54000	0.0013871642
26	Afghanistan	38928341	26	54000	0.0013871642

#Now we tidy up hospital beds

```
hospitalBeds <- hospitalBeds %>% group_by(Country) %>% slice(which.max(Year))
```

Group by country then grab the row that has the max year value (The latest)

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
14	Afghanistan	2004	3.9
15	Afghanistan	2003	3.9
16	Afghanistan	2002	3.9
17	Afghanistan	2001	3.9
18	Afghanistan	2000	3.0
19	Albania	2013	28.9
20	Albania	2012	28.8
21	Albania	2011	28.8
22	Albania	2010	29.9
23	Albania	2009	30.1
24	Albania	2007	30.9
25	Albania	2006	31.2

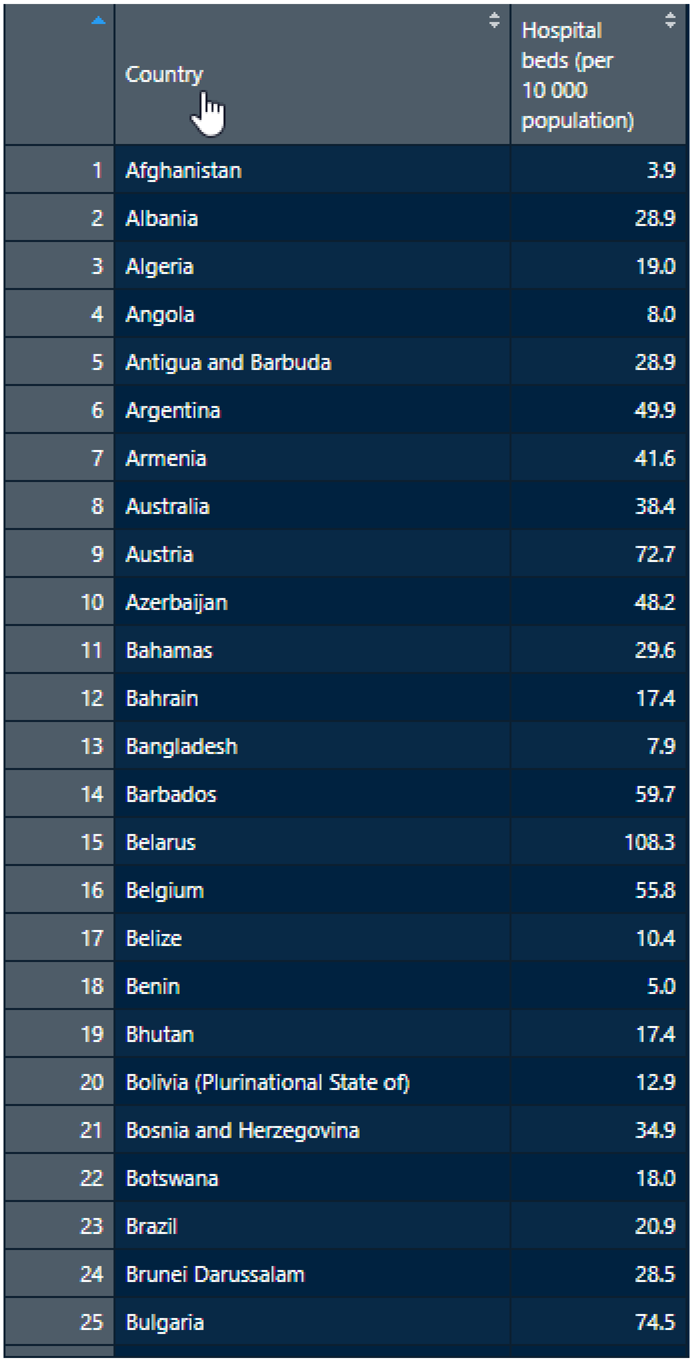
After:

	Country	Year	Hospital beds (per 10 000 population)
1	Afghanistan	2017	3.9
2	Albania	2013	28.9
3	Algeria	2015	19.0
4	Angola	2005	8.0
5	Antigua and Barbuda	2017	28.9
6	Argentina	2017	49.9
7	Armenia	2014	41.6
8	Australia	2016	38.4
9	Austria	2018	72.7
10	Azerbaijan	2014	48.2
11	Bahamas	2017	29.6
12	Bahrain	2017	17.4
13	Bangladesh	2016	7.9
14	Barbados	2017	59.7
15	Belarus	2014	108.3
16	Belgium	2019	55.8
17	Belize	2017	10.4
18	Benin	2010	5.0
19	Bhutan	2012	17.4
20	Bolivia (Plurinational State of)	2017	12.9
21	Bosnia and Herzegovina	2014	34.9
22	Botswana	2010	18.0
23	Brazil	2017	20.9
24	Brunei Darussalam	2017	28.5
25	Bulgaria	2017	74.5

#Dropping year variable from hospitalBeds as that's unneeded

```
hospitalBeds <- hospitalBeds %>% select(-c(Year))
```

I drop year from the table

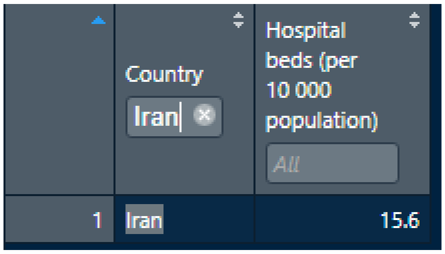


	Country	Hospital beds (per 10 000 population)
1	Afghanistan	3.9
2	Albania	28.9
3	Algeria	19.0
4	Angola	8.0
5	Antigua and Barbuda	28.9
6	Argentina	49.9
7	Armenia	41.6
8	Australia	38.4
9	Austria	72.7
10	Azerbaijan	48.2
11	Bahamas	29.6
12	Bahrain	17.4
13	Bangladesh	7.9
14	Barbados	59.7
15	Belarus	108.3
16	Belgium	55.8
17	Belize	10.4
18	Benin	5.0
19	Bhutan	17.4
20	Bolivia (Plurinational State of)	12.9
21	Bosnia and Herzegovina	34.9
22	Botswana	18.0
23	Brazil	20.9
24	Brunei Darussalam	28.5
25	Bulgaria	74.5

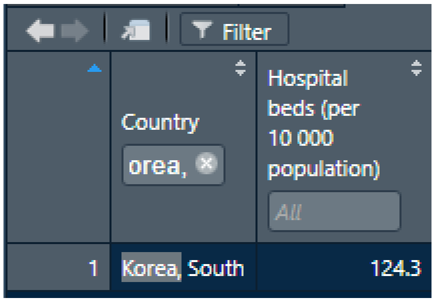
```
#Rename countries to be consistent between all tables
#All name changes are based from covidVaccineDose country names
hospitalBeds <- hospitalBeds %>% mutate(Country = replace(Country, Country == "Iran (Islamic Republic of)", "Iran"))
hospitalBeds <- hospitalBeds %>% mutate(Country = replace(Country, Country == "Republic of Korea", "Korea, South"))
hospitalBeds <- hospitalBeds %>% mutate(Country = replace(Country, Country == "United Kingdom of Great Britain and Northern Ireland", "United Kingdom"))
```

```
hospitalBeds <- hospitalBeds %>% mutate(Country = replace(Country, Country == "Venezuela
(Bolivarian Republic of)", "Venezuela"))
hospitalBeds <- hospitalBeds %>% mutate(Country = replace(Country, Country == "United
States of America", "US"))
```

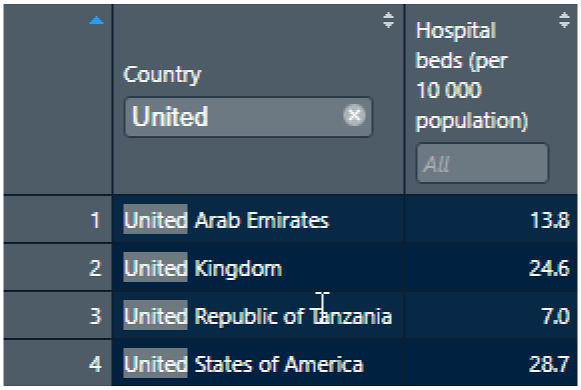
#I noticed venezuela and US was also miss spelled so I decided to fix it :D



	Country Iran	Hospital beds (per 10 000 population) All
1	Iran	15.6



	Country orea,	Hospital beds (per 10 000 population) All
1	Korea, South	124.3

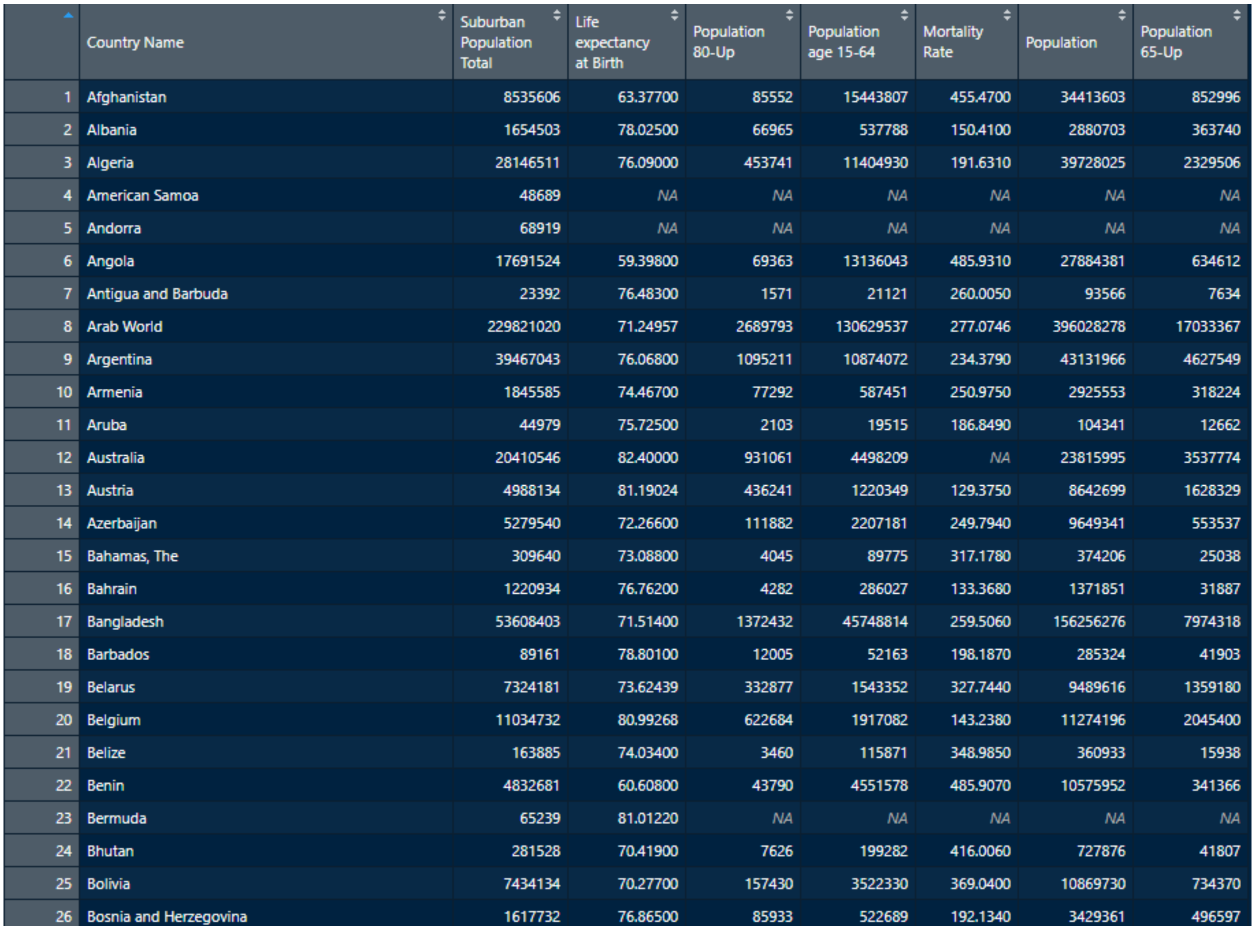


	Country United	Hospital beds (per 10 000 population) All
1	United Arab Emirates	13.8
2	United Kingdom	24.6
3	United Republic of Tanzania	7.0
4	United States of America	28.7

##Tidying for demographics
#Adding both male and female, pulled from homework 5 with some tidying up

```
demographics <- demographics %>% pivot_wider(-'Series Name', names_from = `Series Code`,
values_from = YR2015) %>%
  mutate(`Population 80-Up`=(SP.POP.80UP.FE+SP.POP.80UP.MA), `Population age
15-64`=(SP.POP.1564.MA.IN+SP.POP.1564.FE.IN), `Population age
15-64`=(SP.POP.0014.MA.IN+SP.POP.0014.FE.IN), `Mortality Rate`
=(SP.DYN.AMRT.FE+SP.DYN.AMRT.MA),
`Population`=(SP.POP.TOTL.FE.IN+SP.POP.TOTL.MA.IN), `Population 65-Up`
=(SP.POP.65UP.FE.IN+SP.POP.65UP.MA.IN)) %>%
  select(`Country Name`, SP.URB.TOTL, SP.DYN.LE00.IN, `Population 80-Up`, `Population age
15-64`, `Population age 15-64`, `Mortality Rate`,`Population`, `Population 65-Up`) %>% view()
demographics <- demographics %>% rename("Suburban Population Total" = SP.URB.TOTL)
demographics <- demographics %>% rename("Life expectancy at Birth" = SP.DYN.LE00.IN)
```

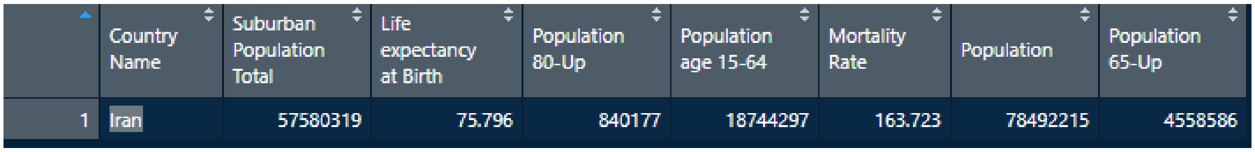
This is pulled directly from homework 5, makes the dataset wider and adding male and female population numbers together



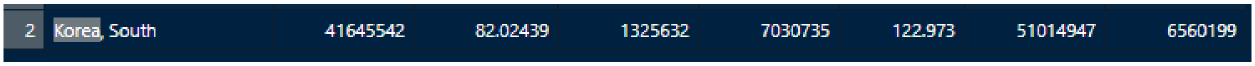
	Country Name	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population	Population 65-Up
1	Afghanistan	8535606	63.37700	85552	15443807	455.4700	34413603	852996
2	Albania	1654503	78.02500	66965	537788	150.4100	2880703	363740
3	Algeria	28146511	76.09000	453741	11404930	191.6310	39728025	2329506
4	American Samoa	48689	NA	NA	NA	NA	NA	NA
5	Andorra	68919	NA	NA	NA	NA	NA	NA
6	Angola	17691524	59.39800	69363	13136043	485.9310	27884381	634612
7	Antigua and Barbuda	23392	76.48300	1571	21121	260.0050	93566	7634
8	Arab World	229821020	71.24957	2689793	130629537	277.0746	396028278	17033367
9	Argentina	39467043	76.06800	1095211	10874072	234.3790	43131966	4627549
10	Armenia	1845585	74.46700	77292	587451	250.9750	2925553	318224
11	Aruba	44979	75.72500	2103	19515	186.8490	104341	12662
12	Australia	20410546	82.40000	931061	4498209	NA	23815995	3537774
13	Austria	4988134	81.19024	436241	1220349	129.3750	8642699	1628329
14	Azerbaijan	5279540	72.26600	111882	2207181	249.7940	9649341	553537
15	Bahamas, The	309640	73.08800	4045	89775	317.1780	374206	25038
16	Bahrain	1220934	76.76200	4282	286027	133.3680	1371851	31887
17	Bangladesh	53608403	71.51400	1372432	45748814	259.5060	156256276	7974318
18	Barbados	89161	78.80100	12005	52163	198.1870	285324	41903
19	Belarus	7324181	73.62439	332877	1543352	327.7440	9489616	1359180
20	Belgium	11034732	80.99268	622684	1917082	143.2380	11274196	2045400
21	Belize	163885	74.03400	3460	115871	348.9850	360933	15938
22	Benin	4832681	60.60800	43790	4551578	485.9070	10575952	341366
23	Bermuda	65239	81.01220	NA	NA	NA	NA	NA
24	Bhutan	281528	70.41900	7626	199282	416.0060	727876	41807
25	Bolivia	7434134	70.27700	157430	3522330	369.0400	10869730	734370
26	Bosnia and Herzegovina	1617732	76.86500	85933	522689	192.1340	3429361	496597

#Rename countries to be consistent between all tables

```
demographics <- demographics %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Iran, Islamic Rep.", "Iran"))
demographics <- demographics %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Korea, Rep.", "Korea, South"))
demographics <- demographics %>% mutate(`Country Name` = replace(`Country Name`, `Country Name` == "Venezuela, RB", "Venezuela"))
hospitalBeds <- hospitalBeds %>% mutate(Country = replace(Country, Country == "United States of America", "US"))
```



	Country Name	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population	Population 65-Up
1	Iran	57580319	75.796	840177	18744297	163.723	78492215	4558586



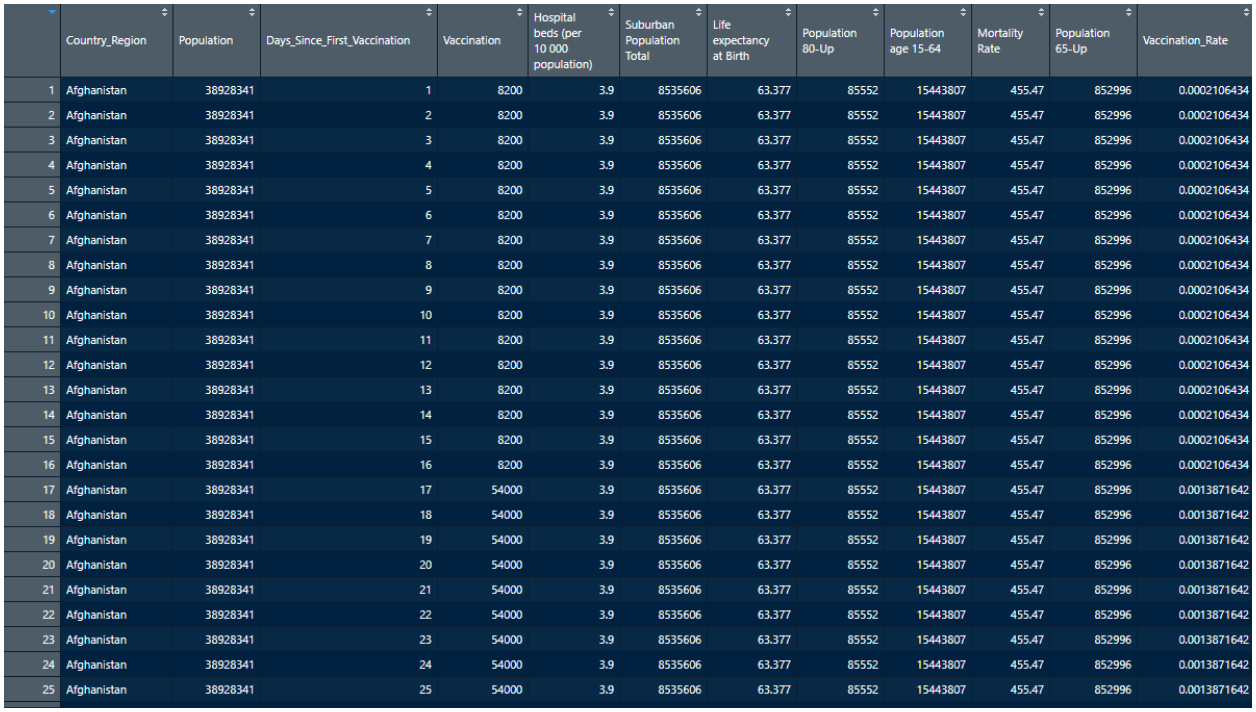
2	Korea, South	41645542	82.02439	1325632	7030735	122.973	51014947	6560199

1	Venezuela	26518336	72.584	349225	8537342	283.108	30081829	1914673

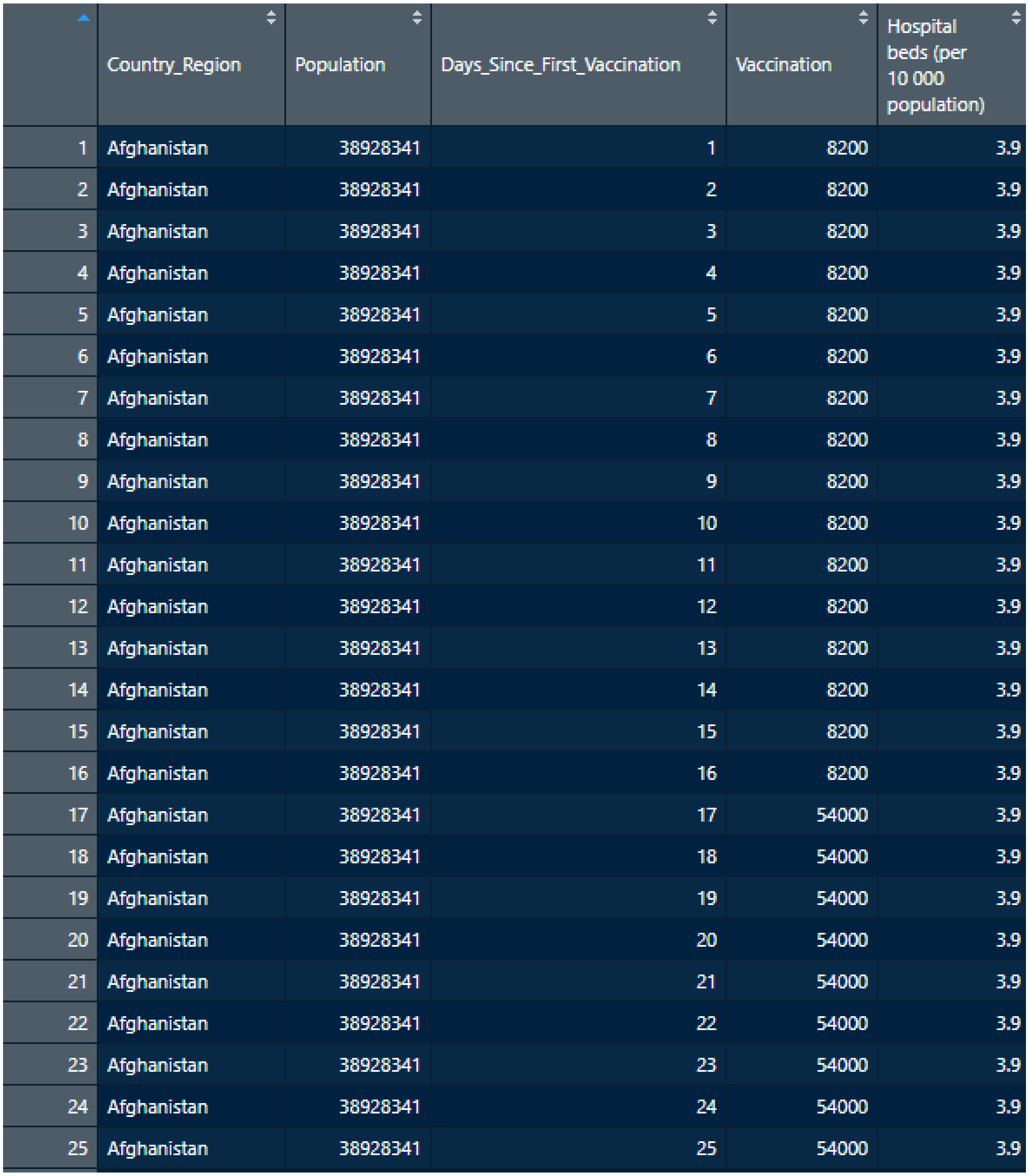
#Join the tables together

```
covidVaccineDose <- covidVaccineDose %>% left_join(hospitalBeds, covidVaccineDose, by =
c("Country_Region" = "Country"))
```

Joining the tables together using country



	Country_Region	Population	Days_Since_First_Vaccination	Vaccination	Hospital beds (per 10 000 population)	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population 65-Up	Vaccination_Rate
1	Afghanistan	38928341	1	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
2	Afghanistan	38928341	2	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
3	Afghanistan	38928341	3	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
4	Afghanistan	38928341	4	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
5	Afghanistan	38928341	5	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
6	Afghanistan	38928341	6	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
7	Afghanistan	38928341	7	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
8	Afghanistan	38928341	8	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
9	Afghanistan	38928341	9	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
10	Afghanistan	38928341	10	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
11	Afghanistan	38928341	11	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
12	Afghanistan	38928341	12	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
13	Afghanistan	38928341	13	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
14	Afghanistan	38928341	14	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
15	Afghanistan	38928341	15	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
16	Afghanistan	38928341	16	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
17	Afghanistan	38928341	17	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
18	Afghanistan	38928341	18	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
19	Afghanistan	38928341	19	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
20	Afghanistan	38928341	20	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
21	Afghanistan	38928341	21	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
22	Afghanistan	38928341	22	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
23	Afghanistan	38928341	23	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
24	Afghanistan	38928341	24	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
25	Afghanistan	38928341	25	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642

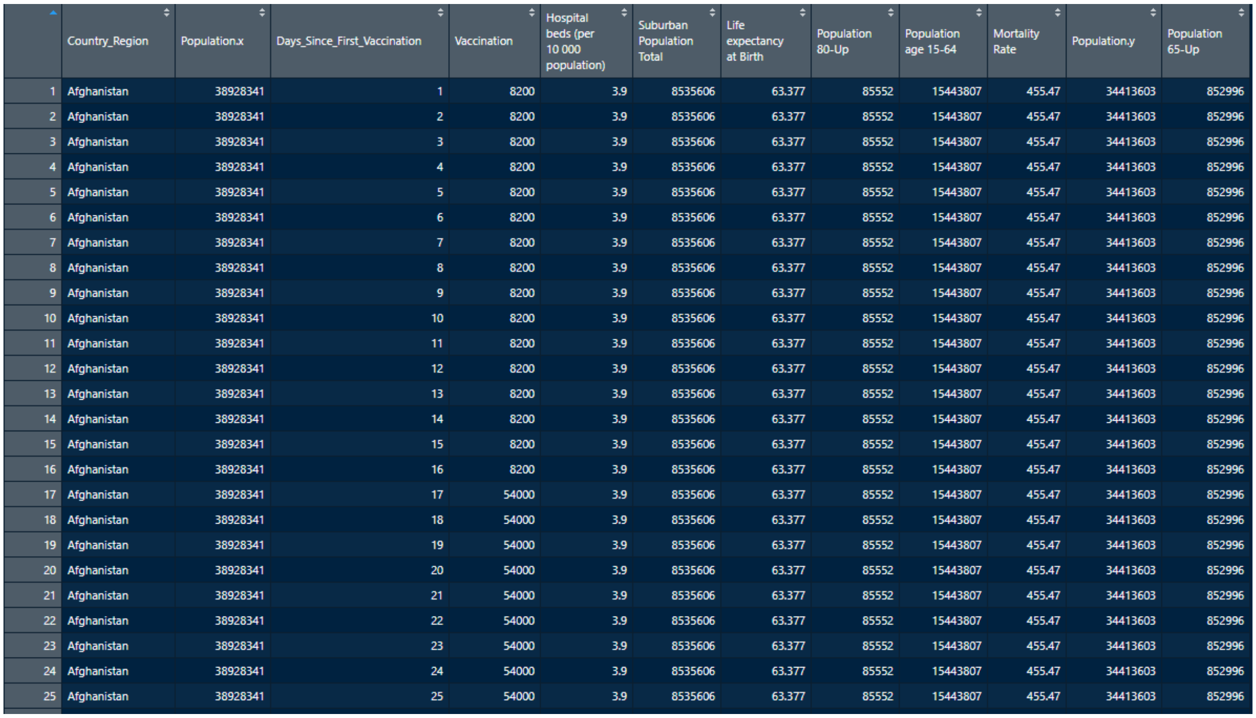


	Country_Region	Population	Days_Since_First_Vaccination	Vaccination	Hospital beds (per 10 000 population)
1	Afghanistan	38928341	1	8200	3.9
2	Afghanistan	38928341	2	8200	3.9
3	Afghanistan	38928341	3	8200	3.9
4	Afghanistan	38928341	4	8200	3.9
5	Afghanistan	38928341	5	8200	3.9
6	Afghanistan	38928341	6	8200	3.9
7	Afghanistan	38928341	7	8200	3.9
8	Afghanistan	38928341	8	8200	3.9
9	Afghanistan	38928341	9	8200	3.9
10	Afghanistan	38928341	10	8200	3.9
11	Afghanistan	38928341	11	8200	3.9
12	Afghanistan	38928341	12	8200	3.9
13	Afghanistan	38928341	13	8200	3.9
14	Afghanistan	38928341	14	8200	3.9
15	Afghanistan	38928341	15	8200	3.9
16	Afghanistan	38928341	16	8200	3.9
17	Afghanistan	38928341	17	54000	3.9
18	Afghanistan	38928341	18	54000	3.9
19	Afghanistan	38928341	19	54000	3.9
20	Afghanistan	38928341	20	54000	3.9
21	Afghanistan	38928341	21	54000	3.9
22	Afghanistan	38928341	22	54000	3.9
23	Afghanistan	38928341	23	54000	3.9
24	Afghanistan	38928341	24	54000	3.9
25	Afghanistan	38928341	25	54000	3.9

#Join demographic by country name as well and then adding population value to those that are NA and match with the demographics country name
##Got help from here
https://stackoverflow.com/questions/42027390/r-how-to-fill-in-missing-value-with-another-dataset-effeciently

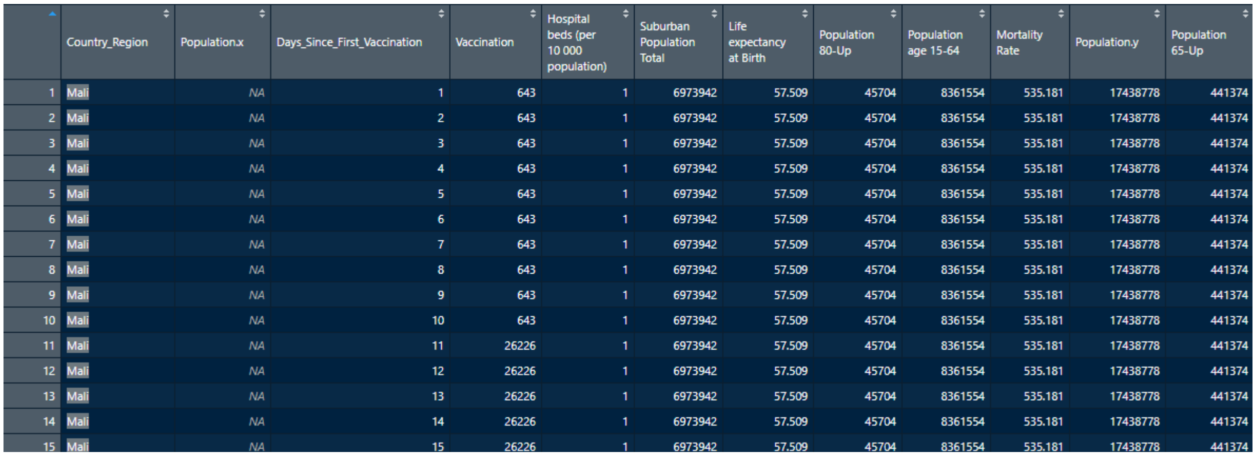
```
covidVaccineDose <- covidVaccineDose %>% left_join(demographics, covidVaccineDose, by =
c("Country_Region" = "Country Name")) %>% mutate(Population.x =
ifelse(is.na(Population.x),Population.y, Population.x))
```

This is the join



	Country_Region	Population.x	Days_Since_First_Vaccination	Vaccination	Hospital beds (per 10 000 population)	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population.y	Population 65-Up
1	Afghanistan	38928341	1	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
2	Afghanistan	38928341	2	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
3	Afghanistan	38928341	3	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
4	Afghanistan	38928341	4	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
5	Afghanistan	38928341	5	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
6	Afghanistan	38928341	6	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
7	Afghanistan	38928341	7	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
8	Afghanistan	38928341	8	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
9	Afghanistan	38928341	9	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
10	Afghanistan	38928341	10	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
11	Afghanistan	38928341	11	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
12	Afghanistan	38928341	12	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
13	Afghanistan	38928341	13	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
14	Afghanistan	38928341	14	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
15	Afghanistan	38928341	15	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
16	Afghanistan	38928341	16	8200	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
17	Afghanistan	38928341	17	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
18	Afghanistan	38928341	18	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
19	Afghanistan	38928341	19	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
20	Afghanistan	38928341	20	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
21	Afghanistan	38928341	21	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
22	Afghanistan	38928341	22	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
23	Afghanistan	38928341	23	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
24	Afghanistan	38928341	24	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996
25	Afghanistan	38928341	25	54000	3.9	8535606	63.377	85552	15443807	455.47	34413603	852996

Before Mutate:



	Country_Region	Population.x	Days_Since_First_Vaccination	Vaccination	Hospital beds (per 10 000 population)	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population.y	Population 65-Up
1	Mali	NA	1	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
2	Mali	NA	2	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
3	Mali	NA	3	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
4	Mali	NA	4	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
5	Mali	NA	5	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
6	Mali	NA	6	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
7	Mali	NA	7	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
8	Mali	NA	8	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
9	Mali	NA	9	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
10	Mali	NA	10	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374
11	Mali	NA	11	26226	1	6973942	57.509	45704	8361554	535.181	17438778	441374
12	Mali	NA	12	26226	1	6973942	57.509	45704	8361554	535.181	17438778	441374
13	Mali	NA	13	26226	1	6973942	57.509	45704	8361554	535.181	17438778	441374
14	Mali	NA	14	26226	1	6973942	57.509	45704	8361554	535.181	17438778	441374
15	Mali	NA	15	26226	1	6973942	57.509	45704	8361554	535.181	17438778	441374

After Mutate:

	Country_Region	Population.x	Days_Since_First_Vaccination	Vaccination	Hospital beds (per 10 000 population)	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population.y	Population 65-Up
1	Mali	17438778	1	643	1	6973942	57.509	45704	8361554	535.181	17438778	441374

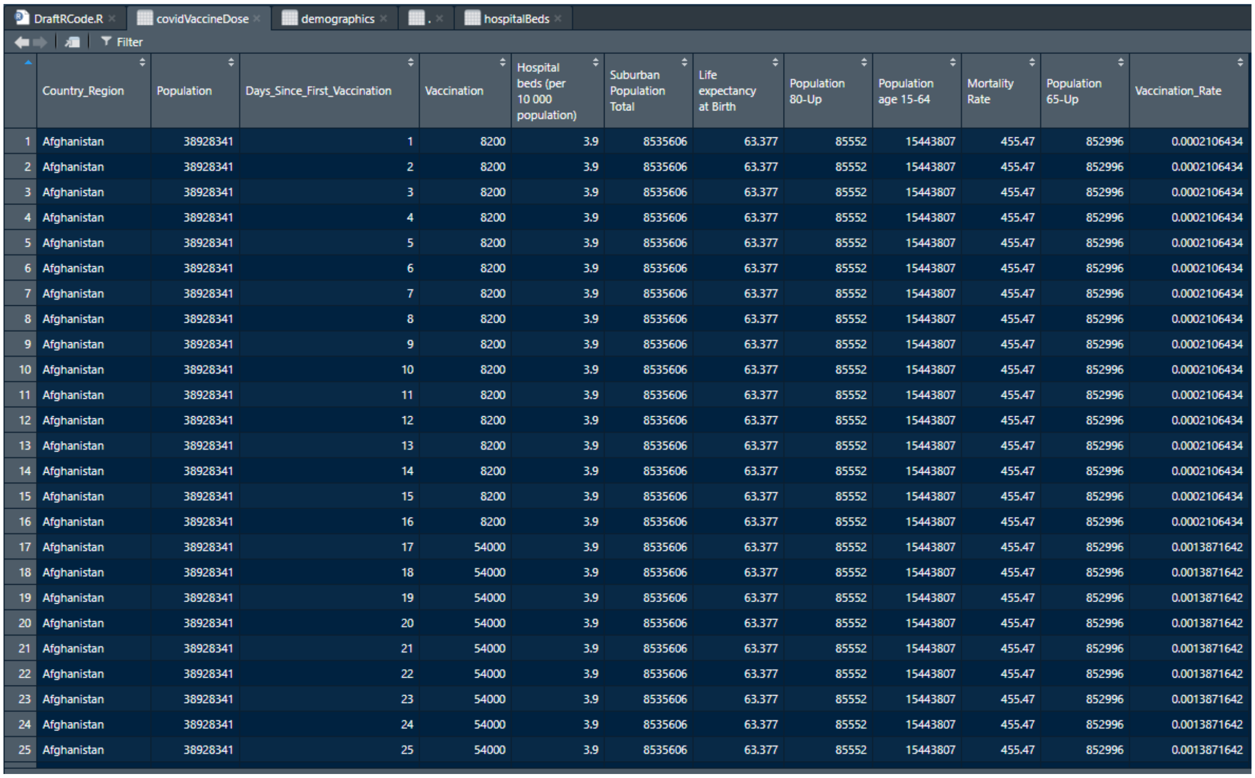
#Remove the column and rename the original column

```
covidVaccineDose <- covidVaccineDose %>% select(-c(Population.y)) %>% rename(Population = c(2))
```

#Now that we filled in missing population values, calculate vaccination rate

```
covidVaccineDose <- covidVaccineDose %>% mutate(Vaccination_Rate = Vaccination/Population)
```

Change naming again so it can look neater and dropped Population.y



	Country_Region	Population	Days_Since_First_Vaccination	Vaccination	Hospital beds (per 10 000 population)	Suburban Population Total	Life expectancy at Birth	Population 80-Up	Population age 15-64	Mortality Rate	Population 65-Up	Vaccination_Rate
1	Afghanistan	38928341	1	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
2	Afghanistan	38928341	2	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
3	Afghanistan	38928341	3	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
4	Afghanistan	38928341	4	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
5	Afghanistan	38928341	5	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
6	Afghanistan	38928341	6	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
7	Afghanistan	38928341	7	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
8	Afghanistan	38928341	8	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
9	Afghanistan	38928341	9	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
10	Afghanistan	38928341	10	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
11	Afghanistan	38928341	11	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
12	Afghanistan	38928341	12	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
13	Afghanistan	38928341	13	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
14	Afghanistan	38928341	14	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
15	Afghanistan	38928341	15	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
16	Afghanistan	38928341	16	8200	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0002106434
17	Afghanistan	38928341	17	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
18	Afghanistan	38928341	18	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
19	Afghanistan	38928341	19	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
20	Afghanistan	38928341	20	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
21	Afghanistan	38928341	21	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
22	Afghanistan	38928341	22	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
23	Afghanistan	38928341	23	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
24	Afghanistan	38928341	24	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642
25	Afghanistan	38928341	25	54000	3.9	8535606	63.377	85552	15443807	455.47	852996	0.0013871642