Group: Gregory Miller

Data wrangling Step description:

For the covid-19 dataset the first thing that I did was remove rows that had a value other than NA for the Province state column, since we do not care about province data, only country. Next, I pivoted the data longer since the data was not tidy, and it will make creating total vaccinations and vaccination rate possible. I also removed un-needed columns from the data such as longitude, as well as removed rows that had zero shots administered. For the covid19 dataset I also added a column that has the number of days since the first non-zero number of shots given. Next, I worked on the bed dataset where for each country in the dataset I only want the most recent years hospital bed per capita and removed un-needed columns from the data. For the last dataset demographics, the first thing that was done was pivot wider all columns other than 'series name' so that we can mutate the dataset so we can use it later. I removed columns that were not needed which is everything other than the 'Country Name', 'SP.DYN.LE00.IN', and 'SP.URB.TOTL'. The last step in cleaning the individual datasets is changing country names so they match when the tables are joined. Since all the tables have the right country name, they can be joined by that country name, and we get the final dataset for the project that can now be used for modeling. See .R file for code.

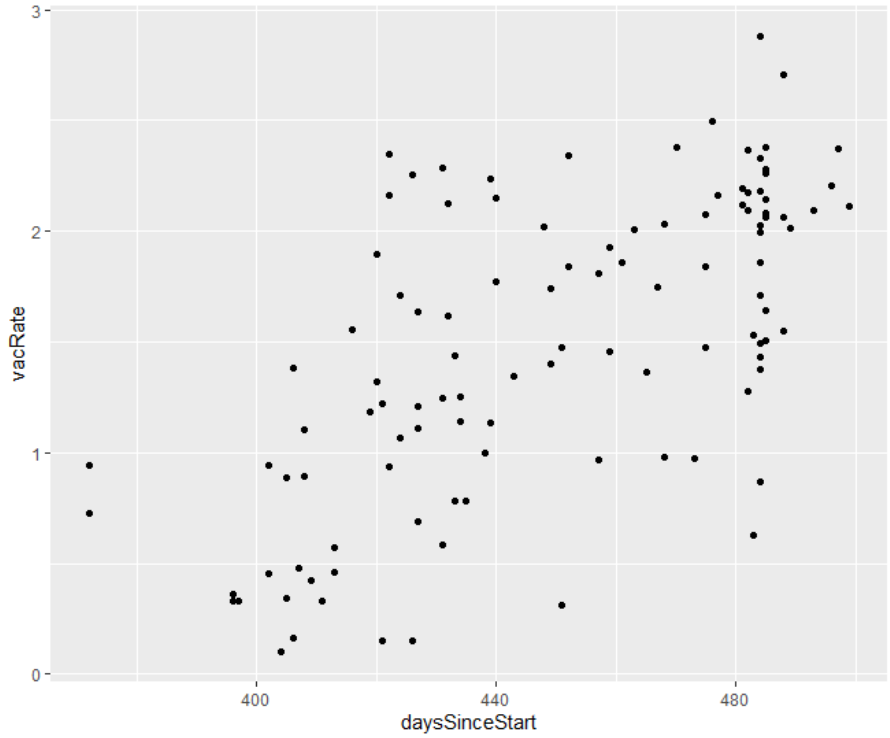
Modeling:

The first step for modeling was creating another variable that will take the total population and divide it by the total urban population. This gets us a new variable that is the percent of the total population that lives in urban cities, which will be used in the models. The first model I created was using the urban population percentage (the variable we just created) as well as beds per capita. I wanted to see how important the new variable as well as one other would be, which is not much. The next model I created was using all predictive variables that are available and this is the highest adjusted

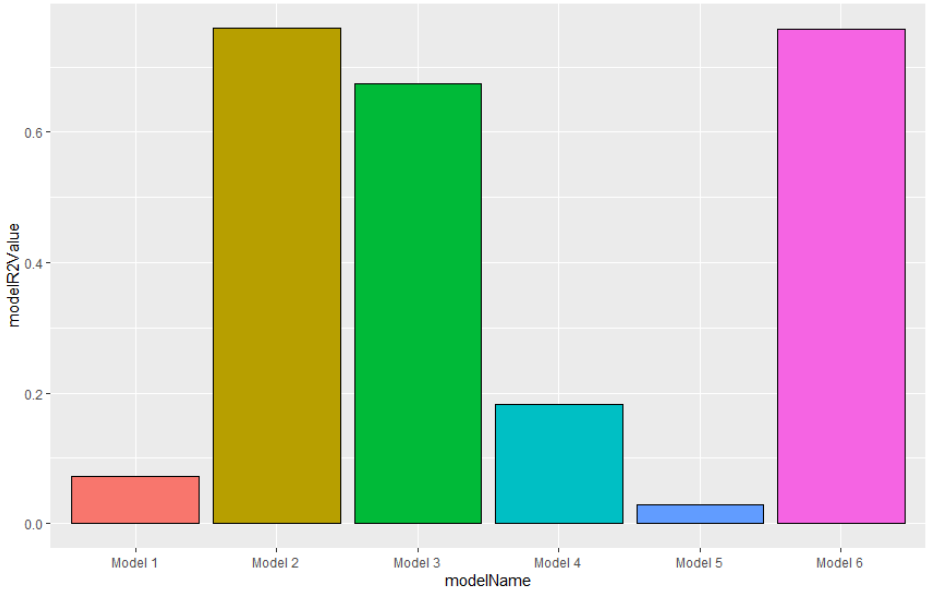
R2 value. Next, I wanted to find out which of those variables are really giving the R2 value so I created a model that has percent urban population, beds per capita, and days since start of vaccinations, that way I can compare to the first model I created. The resulting adjusted R2 value was much higher than model 1 which indicates that days since start of vaccination was one of the most important (I would find out it was the most important). The next model I tried using populations as well as beds per capita and life expectancy which did not lead to a great R2 value. I also tried modeling with population, total urban population, and beds per capita which had a very low value as well. Now that I understood which values gave be high values, I tried creating a sixth model where I do not use every predictive variable but still try and get almost as high as the model with every predictive variable. This model used the percent urban population, beds per capita, days since start of vaccinations, and life expectancy. This last model (model 6) had a R2 value of .7581 and the model with every variable (model 2) has a value of .7589, so the model having two less predictive variables has a R2 .0008 less than that of model 2. See .R file for code.

Graphs:

A scatterplot showing the most recent vaccination rate for every country (y-axis) with the number of days since start (x-axis). See .R file for code.



A bar graph that shows the R2 value on the y-axis and each model on the x-axis. See .R file for code.



Conclusion:

The most accurate model (model 2) was the one using all the predictive variables, however, most of those variables do not add very much to the total R2 value, which we can see when we look at model 6. From the most accurate models we see that days since start of vaccinations is the most important predictive variable by far. However, each predictive variable does have some weight when it comes to R2 value since model 2 is still higher than model 6. This means that all the factors are important just some are far more important than others when determining the vaccination rate of the country. This makes sense, since rich countries had access to vaccines much sooner because they funded the projects, and thus got the vaccines sooner and would be able to vaccinate their population earlier leading to higher vaccinated percentage of the total population.