Richard Gresham

CPSC 375-01

November 12, 2022

Final Report

4.a. Brief descriptions of data wrangling steps.

For data wrangling, the three tables listed, known as demographics, covid vaccine doses, and Hospital beds data, were converted into tidy data. As these data were being tidied, I also removed any unnecessary data including rows/ columns. For the demographics table for instance, I first started by removing Series name and country code, and then I pivoted wider the values for series code. Then I added male and female population together and summarized those values by country name. Then I added a new variable in the Vaccine dataset called vacrate that is the shot count divided by the population. For the vaccine data I then created days since start by first pivoting longer my data using date and shots, towards which I then grouped by country name and mutated a new column called days since start, by counting the number of days where shots > 0 for country. For the bed data set I took the most recent year bed values filtered it, and then selected only beds and Countries. Then I cleaned up some of the countries names and grouped each table by the country name to give me my completed table I will used for linear regression.

4 b.)

For the data modeling, my variables were chosen based off my final table's Predictor variables and my dependent variable. I used a table to run various linear models, and checked the summary for each of them, comparing one another based off their adjusted r-squared values. The table is listed below:

Multiple r2	Adjusted r2	Linear model – equations.
.5372	.5371	vacrate ~ days_since_start
.01324	.01321	formula = vacrate ~ SP.URB.TOTL/beds
.5394	.5394	vacrate ~ days_since_start + SP.URB.TOTL
.003049	.003035	vacrate ~ SP.URB.TOTL
.7078	.7078	vacrate ~ days_since_start + SP.DYN.LE00.IN

The results of this table of models will be used in a bar graph to compare these values for the best one.

4.c. Description of any variable transformations

The first variable transformations that occurred is the vaccination Rate, which is dubbed in my table as vacrate which represents shots per population and is used as our dependent variable within our linear regression models. Another variable transformation that was completed was the transformation and creation of days since start. This was done by using

the date column provided by the vactime data set, to create the days since start. In addition, while no other variable was created specifically in my table, I did create some temporary variable transformations, while I was creating the linear models, with one example being urban population per hospital bed, and adding days_since_start with SP.DYN.LE00.IN.

4.d. A scatterplot of most recent vaccination rates for different countries.

Listed below is a scatterplot that takes the most recent vaccination rates for different countries. This is done by grouping the table by countries, and only taking each country's highest day since they began handing out vaccines, effectively giving the most recent vaccination rate for each country.

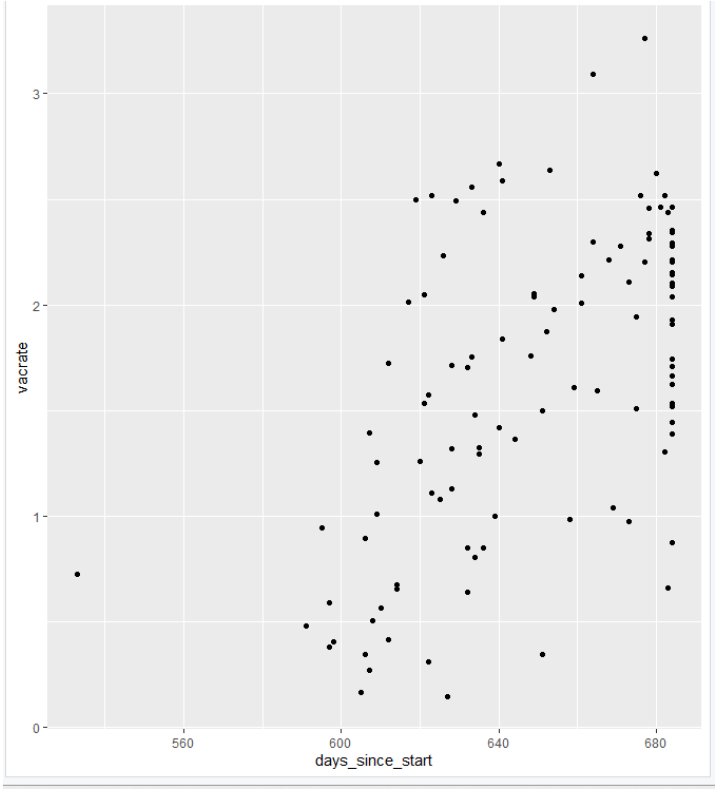

Figure 1. Scatterplot of most recent vaccination

4.e. A plot that shows the R2 values of different models.

Listed below is a bar graph comparing the different values of my 5 linear models.

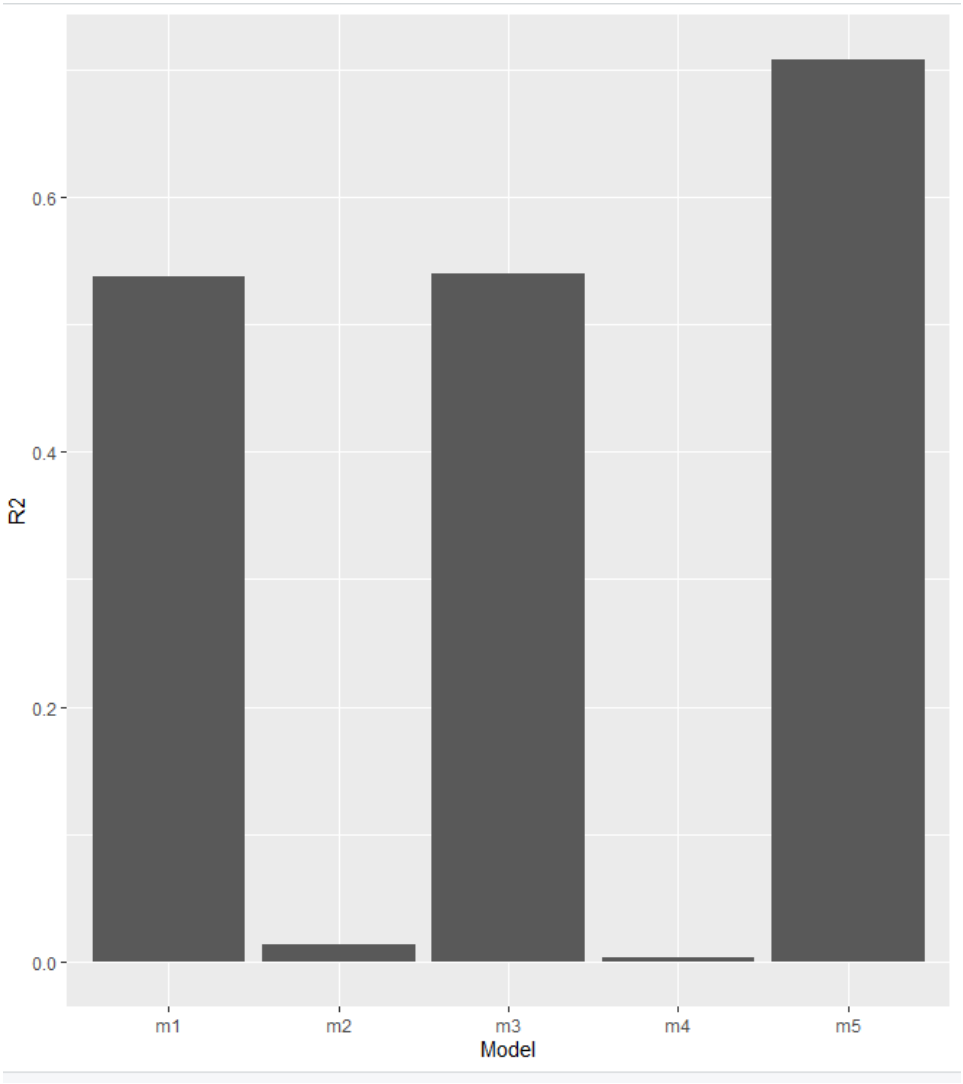

Figure 2. Bar Graph of linear models R2 value.

4.f. A conclusion – What does your modeling say about vaccination rates (e.g., what are the significant factors and what are not).

When comparing and looking at our bar graph in figure 2, which represents whether the R2 values are significant, we can see some insight on what the significant factors of our data in relation to vaccine rates are. For instance, if we look at our model 2 and model 4, we can see that in terms of R2, our independent variables are not as relevant. This independent variable is SP.URB.TOTL, which tells us that urban populations do not hold a significant factor in our dataset. Whereas in our model 5 we could see some sort of significant factor in play when we look at our independent variable's days_since_start and SP.DYN.LE00.IN, since our variables hold a high R2 value of .7078. The relation however, to m1, m3, and m5 shows us that the common independent variable days_since_start holds a more significant factor than SP.DYN.LE00.IN, since it influences a much higher R2 values. However, all of the example models seen in figure 2 is not enough to ascertain that even the best model m5, should be considered an entirely good model. Thus, in conclusion I would say that given the models present in figure 2, that the independent variables used to depict vaccination rate for any given country are only somewhat significant if they possess the independent variable days_since_start.

4.g. Clarity of the report.

In the end, I couldn't find a significant combination of independent variables present within my table that could give me a significant conclusive model for the vaccination rate. Aside from the days since a country started giving vaccine shots, there weren't many significant or impactful independent variables apart from that in our models, that could give us a truly conclusive model. Many independent variables such as SP.URB.TOTL only gave a very

small percentage of accuracy, like seen in m4 which gave a measly .003035 of our R2 value.