Final Report

Jonathan Dao

Mahmoud Khafagy

Richard Stanley

Written Report

4.a. Brief descriptions of the data wrangling steps

For data wrangling, the three tables, demographics, covid, and the gdp data were converted into tidy data. Then all the unnecessary data were filtered out. For the demographics table, LE00 and the URB were converted to a smaller tables and joined by country code so each country can have each population labeled individually. We added a variable days since the first day of vaccination, and a mutation for the vaccination rate based on shots given to people per one hundred thousand. Since the vaccination rate is already based on population, later in our modeling, there was no change in the adjusted r-squared value. While adding shots per hundred thousand doesn't change the model, it would be a decent metric to use to report possible predictions. At the end all the tables were merged by iso3 and country code.

4.b. Brief description of how variables were chosen for data modeling

For the data modeling, we used a table to run various linear models, checked each summary, and compared the adjusted r-squared values. Here is that table:

Please note that DSS is variable: days_since_start and VR is the vaccination rate variable.

multiple r2	adjusted r2	Linear model - pseudocode
0.5759	0.5759	DSS_URB <- lm VR~DSS+URB
0.01808	0.01802	DSS_GDP_URB <- lm VR~URB+GDP
0.5979	0.5979	DSS_GDP_URB_DST <- lm VR~DSS+URB+GDP
0.6931	0.693	BEST <- lm VR~DSS+URB+GDP+LE00
0.6931	0.693	DSS_URB_GDP_LE00_POP <- lm VR~DSS+URB+GDP+LE00+POP
0.1779	0.1778	URB_GDP_LE00_POP <- lm VR~URB+GDP+LE00+POP

The results of these models are used later in this report in a bar graph to compare these values.

4.c. Description of any variable transformations

One of the more obvious variable transformations is the vaccination Rate [vacRate] which represents shots per population and used as the dependent variable for all the models that we needed to create. Mutating the days_since_start was a bit of challenge, although in order to graph a progression based on time, it needed to be done.
We also made one variable transformation which was vaccination rate * 100,000 which equals the number of shots per one hundred thousand people. Though we didn't use ShotsPerHundredK in our reports, we kept it in because it's a decent metric.

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

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

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

When checking out the bar graph which represents whether or not the adjusted r2 values are significant, the last two models GDP+URB, and URB+GDP+DYN+POP didn't have time as a factor, which we wanted to emphasize the importance of time as an additional independent variable to the other factors that made the more successful models. Ultimately, the inclusion of all these data points didn't give us a linear model that was significant enough to conclude that even the best model should be considered to be a *good* model. Therefore, in our conclusion, we would say that the independent variables used to depict vaccination rate for any given country in any of our models are not sufficiently significant and reject the claim that they would be in any current combination of the data given.

4.g. Clarity of the report

Ultimately, we didn't find a significant combination of independent variables that gave us a significant conclusive model for vaccination rate. We will say that aside from the days since the start of shots that the most impactful independent variable was **life expectancy at birth** which only effected the models at a difference of 0.0951 or roughly 10% of the adjusted r2.

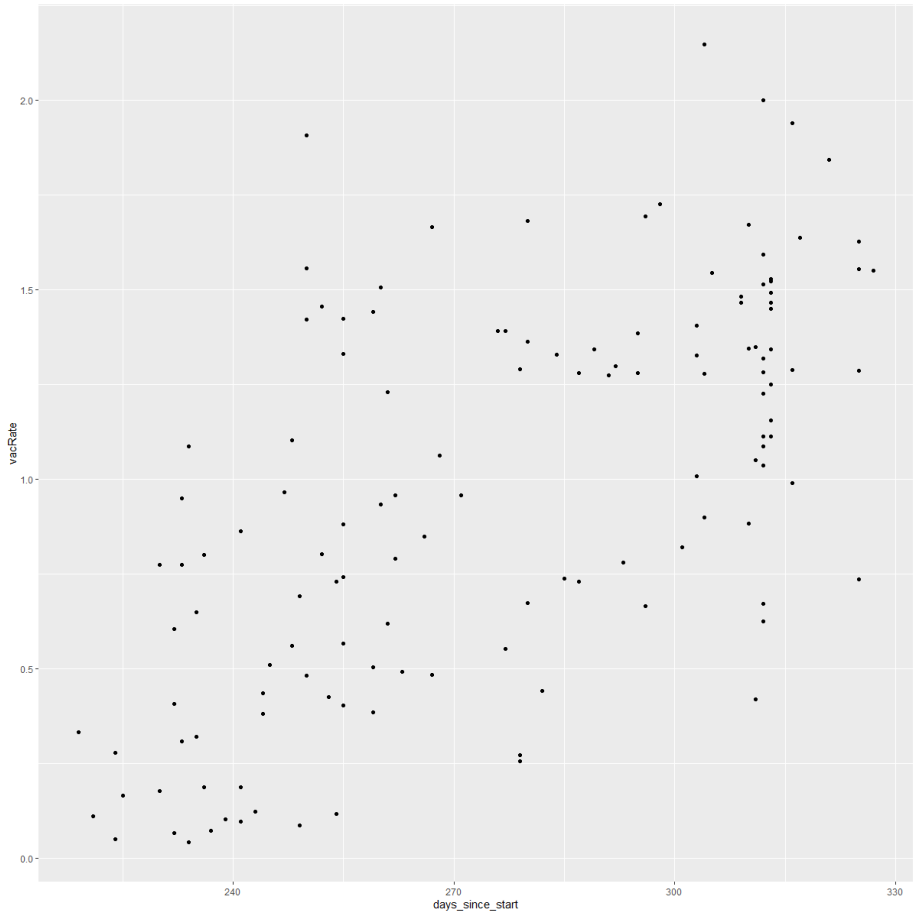

Figure 1: scatterplot of most recent vaccination

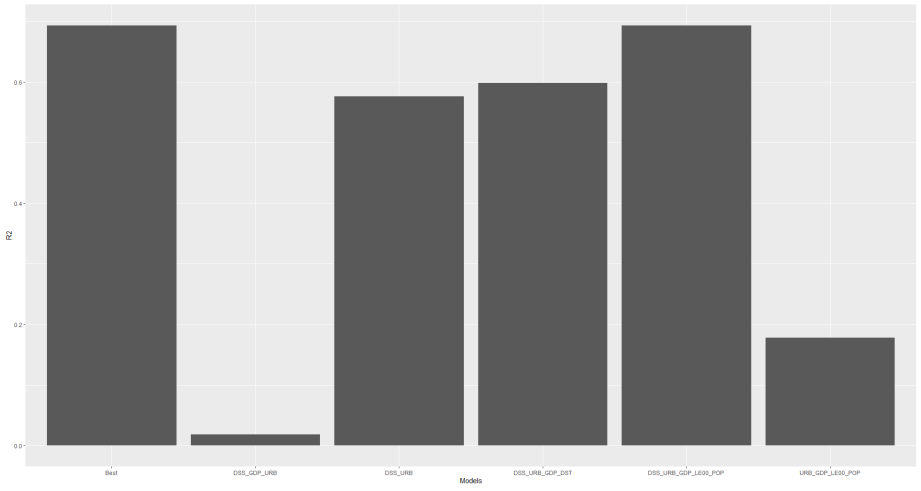

Figure 2: A plot that shows the R2 values of the different models