FACULTY OF
APPLIED SCIENCES

Investigation of Covid-19 vaccination rates

Econometrics course

by

Nazar Dobrovolskyy, Volodymyr Savchuk

APPS@UCU, 2021

Dedication

We dedicate this report to the Almighty God without whom we can do nothing. We further dedicate it to our parents and guardians for their unceasing and selfless support throughout our stay in this university.

1. Introduction

In this paper, we aimed to investigate the Covid-19 vaccination rates in the world. Also, to find the relationship between demographic indicators, some macroeconomics indicators such as GDP(Gross domestic product), GNI(gross national income), and the Covid-19 vaccinations rates.

The reason for that research is to understand why some countries like Israel, the United Kingdom, the United States of America, Chile are leaders in vaccination progress. In contrast, Ukraine, Uzbekistan, Thailand, Egypt has less than three percent of citizens vaccinated.

Also, we have seen that many developed countries, which are vaccinated at a high rate, enjoy the popularization of vaccination through famous and influential people. We were interested in this fact, and we decided to check whether celebrity vaccination affects the overall pace of vaccination in the country?

2. Data description

2.1 Investigation of dependency between vaccination rates ,macroeconomics and demographics factors.

To study such a dependency we need data about the demographic indicators of all countries(source), macroeconomics indicators(source) of all countries and data about the vaccination rates in all countries.

To gather vaccination about the current vaccination rates, namely doses of vaccine already used we worked with this dataset (source). It contains a lot of information about current vaccination info in the world, but we needed only one column:
- *Daily_vaccinations_(raw)* - for a certain data entry, the number of vaccination for that date/country;

We decided to discover whether the severity of Covid in the country influance vaccination rates. So, we also collect data about deaths from Covid-19, total Covid-19 cases and current country's population. (source).

Then the data was merged to create a complete data frame to work with. The new dataframe has such columns that we use:

Vaccine.Quantity - the number of vaccine used in the country;
gdp - GDP of the country;
Population - population of the country;
Area - area of the country;
Cases - Covid-19 total cases in the country;
Deaths - Covid-19 total deaths in the country;
Recovered - total number of people recovered from Covid-19.

2.2 Vaccination, British and US celebrities data

To study the influence of celebrities, we chose Britain and the United States because they show the rapid pace of vaccination at present. Besides, these countries have a large number of influential celebrities. Because many countries promote vaccination, spread it through vaccination of celebrities. That is why our goal is to test whether celebrity vaccination affects the growth rate.

In order to check this, first of all, we need data on vaccination in countries. We decided to use this dataset (source). It contains various information about vaccination around the world with daily information. Nevertheless, for research, we need only two columns:

- *date* - date of data entry

- *daily_vaccinations_per_million* - ratio (in ppm) between vaccination number and total population for the current date in the country;

(in fact, there is no difference whether to take for per one million or total, it does not affect the result of the study, but to facilitate the work - let us take one million)

In addition to the data from this dataset, we had to manually collect two datasets about the date of publication of information about the vaccination of celebrities, the publication itself. One is dedicated to famous Americans who have been vaccinated, and the other to the British. Each of them contains two columns:

- *name* - the name of the vaccinated celebrity

- *date* - the date of publication of the celebrity vaccination.

3. Methodology explanation

3.1 Investigation of dependency between vaccination rates ,macroeconomics and demographics factors.

Multiple Linear Regression Model

We use a linear regression model to analyze the dependency and how strong the relationship between vaccination rates, macroeconomics, and demographics factors. Before running the model, we need to define which features from our dataset correlate with vaccination rates and could be used in the model.

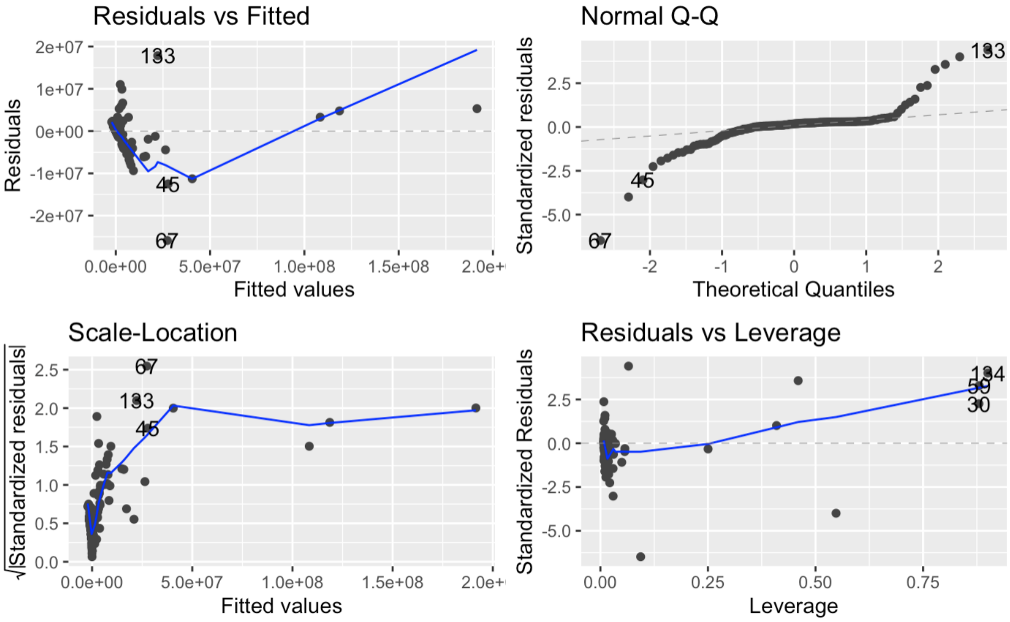
We use the correlation matrix analysis because it is very useful to study dependences or associations between variables. Running function $rquery.cormat()$ we founded feature Vaccine.Quantity correlates the most with features: gdp, Recovered, Cases, Deaths, Area, Population. So the formula for the model is:

$$Vaccine.Quantity = \beta_0 + \beta_1 * gdp + \beta_2 * Population + \beta_3 * Area + \beta_4 * Cases + \beta_5 * Deaths + \beta_6 * Recovered + u$$

Then we performed a test for multicollinearity and discovered that feature Cases and Recovered should be dropped because they are very collinear with feature Deaths.

We are ready now to run a multiple linear regression model. First, we should check the R-squared, which in our case is 0.93. So 93% the variance of the dependent variables being studied is explained by the variance of the independent variable.

Second, we check our linear regression assumptions of the data, such as Linearity of data, Normality of residuals, Homogenuity of residuals variance, Independence of residuals error terms. We do that with the help of regression diagnostic plots using function $autoplot()$ from $ggfortify$ library.



From the first plot: "Residuals vs. Fitted" we see that the relationship in the data is non-linear because otherwise, the blue line would be horizontal at zero.

From the second plot: "Scale-Location" we see that residuals are not spread equally along the range of predicators because the blue is not a horizontal with equally spread points. It can be seen that the variability (variances) of the residual points decreases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or heteroscedasticity).

From the third plot: "Normal Q-Q" we see that the normality assumption does not hold, because the normal probability plot of residuals should approximately follow a straight line(in our case is not truth).

The fourth plot: "Residuals vs Leverage" is helpful to determine the outliers. The plot highlights the top 3 most extreme points$(134, 59, 30)$. However only, one of them exceeds three standard deviations, so only that one should be dropped.

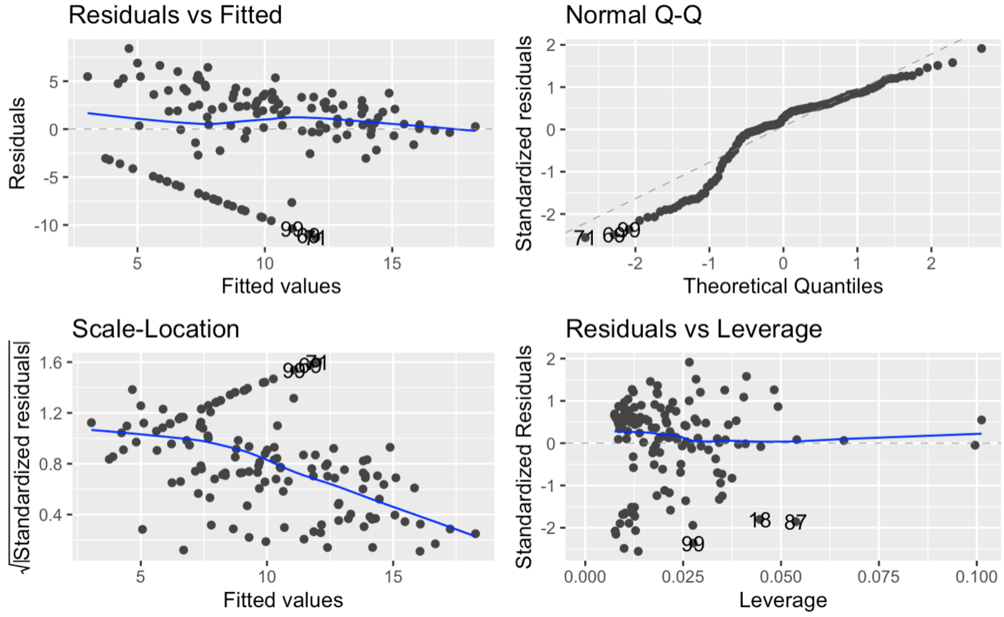
The next was to drop the outliers and to run Log-Log Linear Regression Model as a solution to heteroscedasticity and non-linear relationship in the data.

Multiple Log-Log Linear Regression

After running the model, we discovered that log(population) and log(area) are already not significant features, so the formula for our final model is:

$$log(Vaccine.Quantity) = \beta_0 + \beta_1 * log(gdp) + \beta_2 * log(population) + u$$

Let's check our linear regression assumptions of the data:



In general, things become better after performing log transformations, the data has a linear relationship, no significant outliers, but heteroscedasticity still exists.

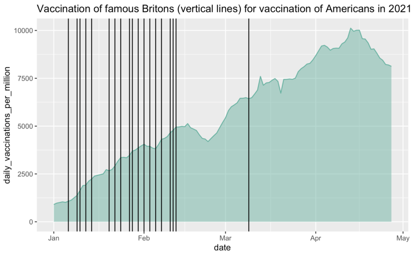
3.2 Influence of celebrities on the rate of vaccination

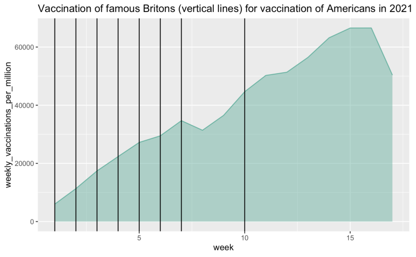
After collecting the necessary data, creating helpful data sets for selected countries and processing all the data to the desired form (Britain, USA), we wanted to see what the vaccination rate looks like on the chart for both countries. To do this, we broke the information about vaccination by day and week.

If we look at the plots below, we will see the vaccination in Britain with vertical lines that describe the time of vaccination of famous Americans and vice versa.

Why so?

We cannot investigate the influence of British celebrities in Britain due to the lack of information on what the results would be if they were not vaccinated. That is why we decided to choose the United States and Britain, which now show some of the best results in vaccination. We will compare how the vaccination of famous Britons affects the vaccination of US residents and vice versa.

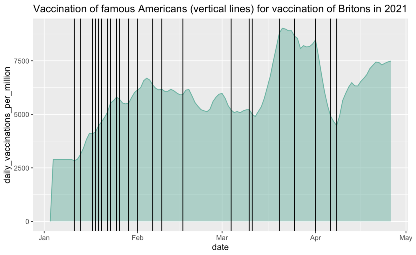


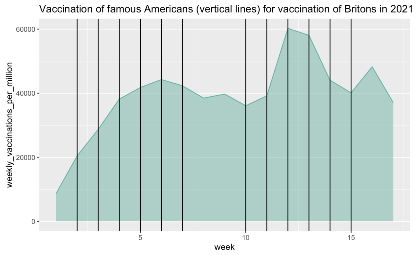


Above are data on vaccinations in the United States and vaccinations of famous people from Britain.

On the y-axis is the number of vaccinations per day per million and on the x-axis, or daily vaccination data, or weekly data. Black vertical lines are the time when celebrities were vaccinated.

Below are data on vaccinations in the UK and vaccinations of famous people from the United States.





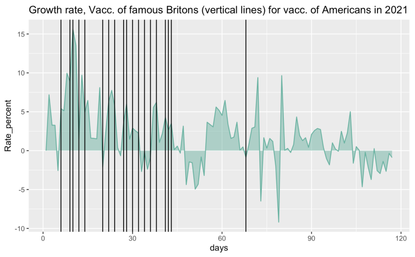
We did this according to the usual information, but now let us look at the percentage change by days and weeks to estimate the percentage dependence better, not on the number but the change in the increase or decrease in vaccinations.

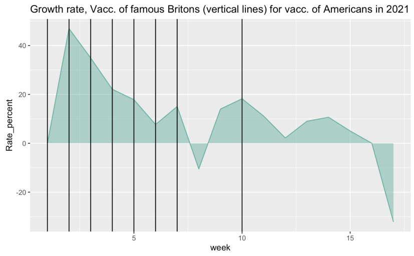
To do this, let us calculate the growth rate of vaccination in both countries. We will calculate according to the formula:

$$(\mathit{now} \;-\; \mathit{past}) \;/\; \mathit{now} \;,$$

where
now - number of vaccinations in the present time,
past - number of vaccinations in the past
- and apply this to all data in our time series.

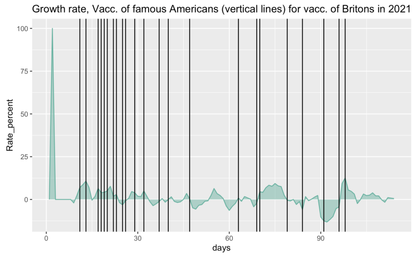
Below we will show the obtained plots.

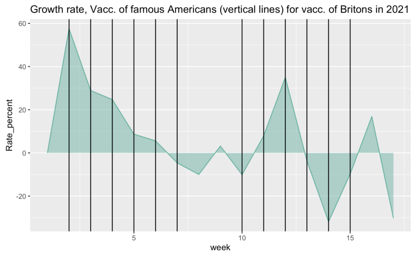




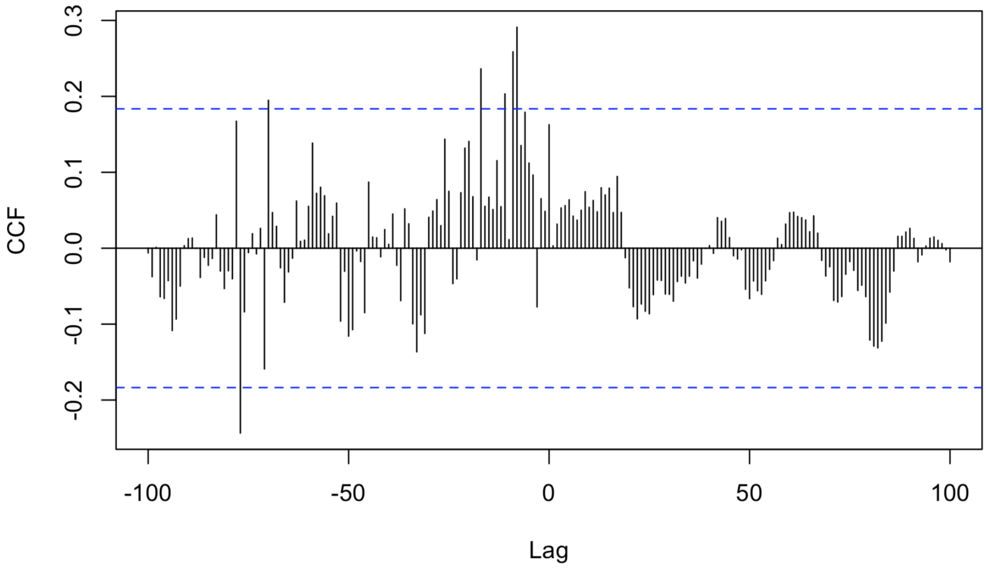
Above are data on vaccinations in the United States and vaccinations of famous people from Britain due to the growth rate.

Below are data on vaccinations in the UK and vaccinations of famous people from the United States due to the growth rate.





Now let us see if there is any relationship between these growth rates in the UK and the US. So we just combined by outer join the data from both tables of their rate percents.



Moreover, as we can see, when the lag fluctuates around -8, then the value becomes significant in a positive correlation. Nevertheless, it is still minimal (< 0.3). Thus, it can be assumed that there is a minimal positive correlation at a lag of about -8 between Britain and the United States.

Now, to calculate whether there is a correlation between celebrity vaccinations and vaccinations. First, move our day counter when the celebrity was vaccinated to the is_famous column by one. After all, if a celebrity was vaccinated at a time t, then we at least need to know whether there is a correlation at time $t + 1$ and so on.

We see that the correlation is present from the beginning of the celebrity vaccination announcement until the 33rd day from that moment.

Now let us repeat the same test, only for American celebrities and the British growth rate of vaccination.

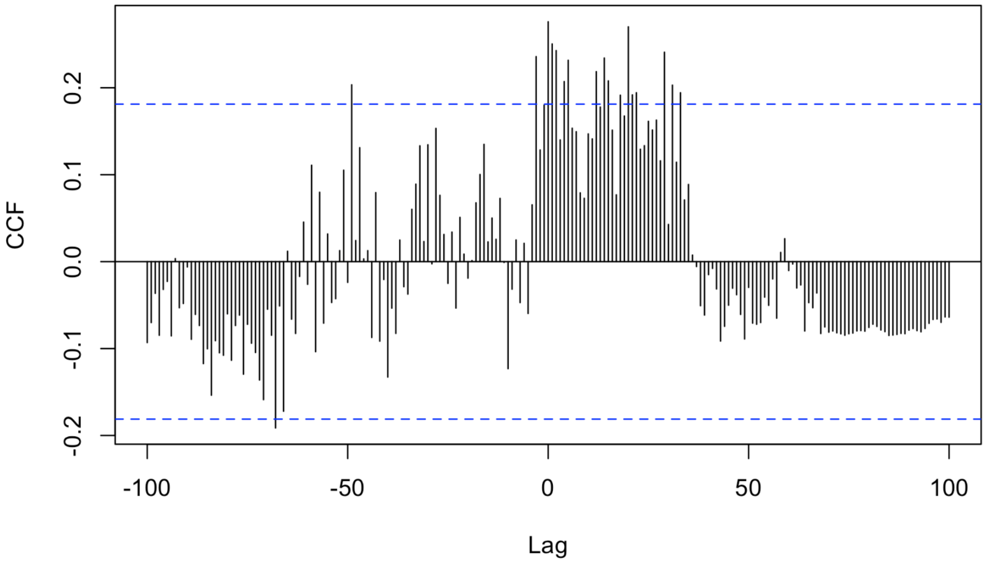


Figure 1: Cros-Correlation between UK celebrities and vaccination rates in the US

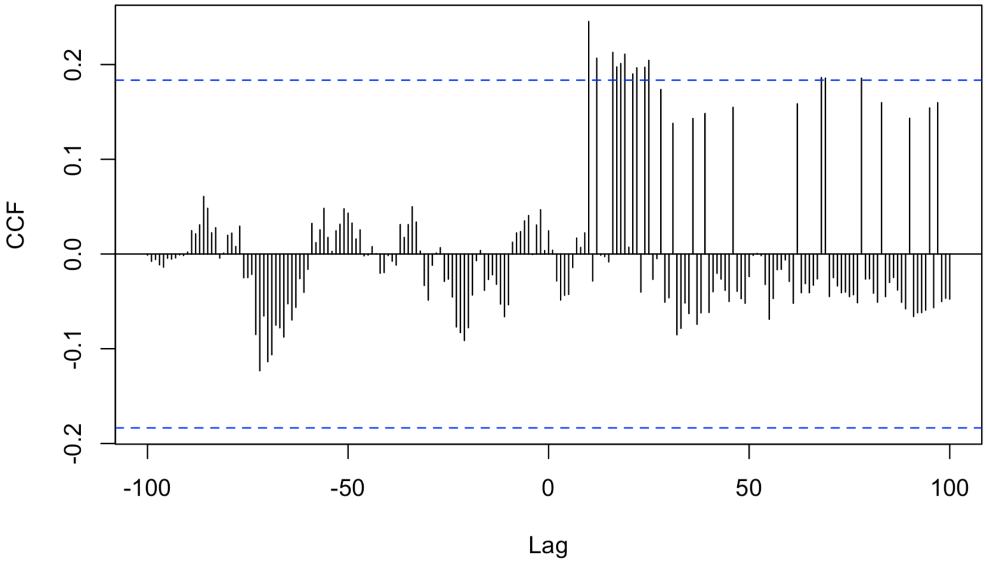
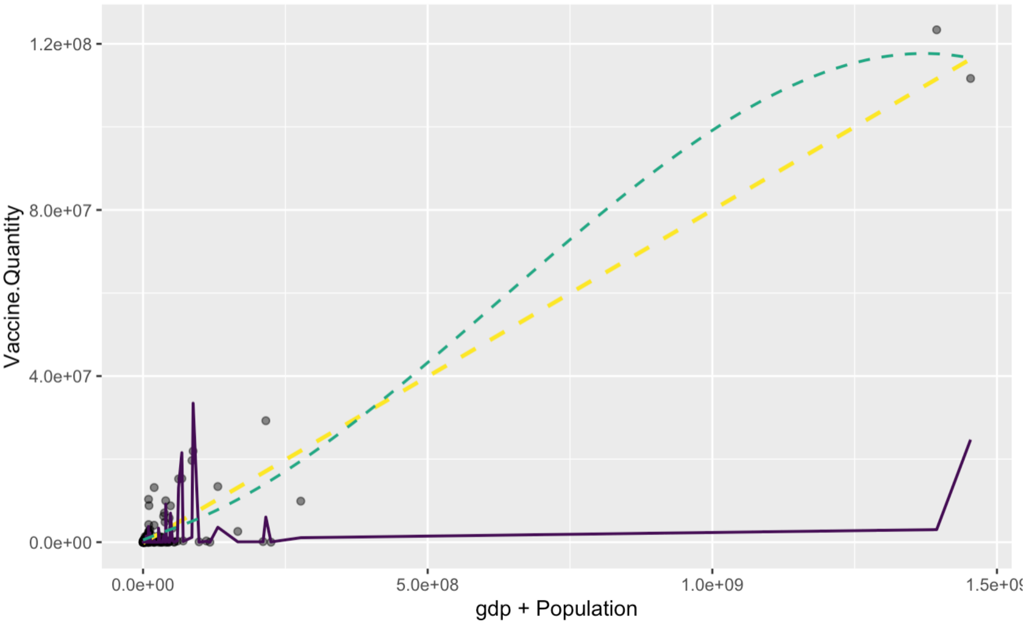


Figure 2: Cros-Correlation between US celebrities and vaccination rates in the UK

4. Results

4.1 Investigation of dependency between vaccination rates ,macroeconomics and demographics factors.

From our Log-Log Multiple Linear Regression model we get:



Plotting the results in the original scales(the regression line - purple)

$$\beta_1(log(GDP)) : 2.1167$$
$$\beta_2(log(Population)) : -0.6243$$

Both features log(GDP) and log(Population) have significant p-values(less than 0.05), so we found a significant relationship between GDP of the country, the population of the country, and vaccination rates in that country.

Specifically, we found 0.62% decrease in Vaccine.Quantity for every 1% of increase in Population. Also, we found 2.11% increase in Vaccine.Quantity for every 1% of increase in GDP.

According to our p-value of the model, which is 2.608e-13, and multiple r-squared of the model, which is 0.3553, our data is explained only up to 35%, however, our p-value shows that the significant relationship was indicated.

4.2 Influence of celebrities on the rate of vaccination

In *Figure* 1, we can see that again; this positive correlation is so minimal ($0.2 < x < 0.3$) that it barely exceeds the significance line. Nevertheless, this result is evident because the correlation between Britain and the United States is not apparent. Perhaps, if we choose the best countries with a similar vaccination rate, they will show a more precise result. And yes, the growth rate of vaccination of famous Britons has a vague positive correlation.

In Figure 2, we see that there is a correlation on day 10, up to 25. As in the previous case, there is a fuzzy positive correlation. The significance level exceeds the value by $68 - 69$ days, and it is slightly on 78 day. As we can see, the results are approximately similar; there is no clear correlation. Therefore, the conclusion remains the same. And, yes, the growth rate of vaccination of famous Britons has a vague positive correlation.

5. Conclusions and next steps

As for the impact on the vaccination of famous people, there is a correlation with a slight positive relationship. Perhaps this is due to an imperfect correlation between Britain and the United States, or it actually remains significant only with a vague dependence.

To sum up, the linear regression model is not a good estimation for Covid-19 vaccination rates. Moreover, vaccination rates could not be explained by macroeconomics indicators, demographic indicators, or Covid-19 statistics about deaths and covid cases. The main factor here is the quality of government, its ability to conduct negotiations. If we had those indicators, we would probably discover a better model.

Next steps

As for the next steps, we would like to continue research on the influence of famous people on the vaccination rate. To do this, we can go through all the countries and find two that are best correlated with vaccination rate. And see if the positive correlation increases.