REPORT

Data Mining project

Dzhidzhavadze Luiza

MLDM M1

1. Problem Understanding

In this project I present some statistics and predictions on Covid-19 and its vaccinations. It's been more than one year that Covid-19 has become an integral part of our lives. This fact made me think that making a prediction model on this virus could be important and a very interesting idea. That is why the main point of my project is based on making a model which can answer the question: "When can a certain country achieve a herd immunity of Covid-19?". [1]

2. Data Understanding

For my model I used a data called "COVID-19 World Vaccination Progress" which I found on the www.kaggle.com.[2] This data is daily updated from Our World in Data GitHub repository for covid-19.

It has **15** columns and **8,451** rows of information about 154 countries.

The data contains the following information:

- **Country**- this is the country for which the vaccination information is provided;
- **Country ISO Code** - ISO code for the country;
- **Date** - date for the data entry; for some of the dates we have only the daily vaccinations, for others, only the (cumulative) total;
- **Total number of vaccinations** - this is the absolute number of total immunizations in the country;
- **Total number of people vaccinated** - a person, depending on the immunization scheme, will receive one or more (typically 2) vaccines; at a certain moment, the number of vaccination might be larger than the number of people;
- **Total number of people fully vaccinated** - this is the number of people that received the entire set of immunization according to the immunization scheme (typically 2); at a certain moment in time, there might be a certain number of people that received one vaccine and another number (smaller) of people that received all vaccines in the scheme;
- **Daily vaccinations (raw)** - for a certain data entry, the number of vaccination for that date/country;
- **Daily vaccinations** - for a certain data entry, the number of vaccination for that date/country;
- **Total vaccinations per hundred** - ratio (in percent) between vaccination number and total population up to the date in the country;
- **Total number of people vaccinated per hundred** - ratio (in percent) between population immunized and total population up to the date in the country;
- **Total number of people fully vaccinated per hundred** - ratio (in percent) between population fully immunized and total population up to the date in the country;
- **Number of vaccinations per day** - number of daily vaccination for that day and country;
- **Daily vaccinations per million** - ratio (in ppm) between vaccination number and total population for the current date in the country;
- **Vaccines used in the country** - total number of vaccines used in the country (up to date);

[1] https://www.kaggle.com/xleong3/statistics-predictions-on-covid-19-vaccinations

[2] https://www.kaggle.com/gpreda/covid-world-vaccination-progress

- **Source name** - source of the information (national authority, international organization, local organization etc.);
- **Source website** - website of the source of information.

3. Data analysis and Data Preparation

Before building the model I needed to do some data preprocessing procedures.

Firstly, I deleted all the columns which were useless for building my model ('source name' and 'source website’). After that I deleted all rows with no "total vaccinations" value (since I need these values to build my model after). For the beginning I used the data set to answer some general questions about an adoption of Covid-19 vaccines such that:

1. How many countries have started vaccinations?
2. Which country has administered the most number of vaccines?
3. What is the TOP-20 countries of total number of vaccines administered? (Figure 1)
4. What is the TOP-20 countries by total vaccinations per hundred of population? (Figure 2)

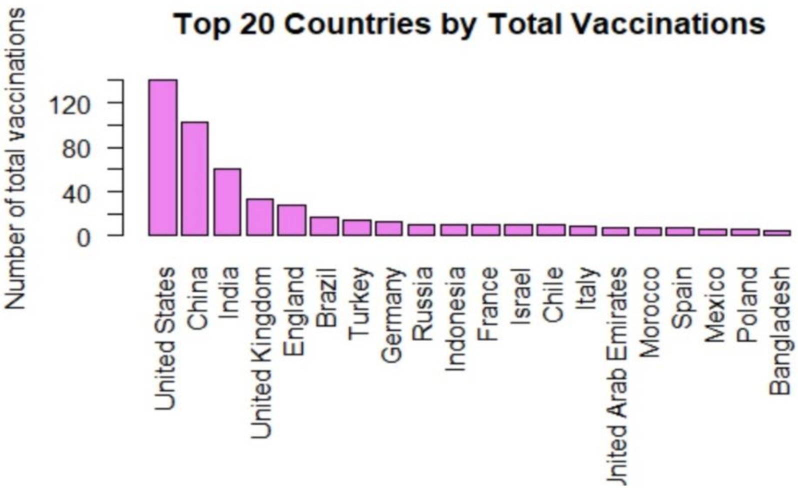

Figure 1. TOP-20 countries by Total number of vaccinations

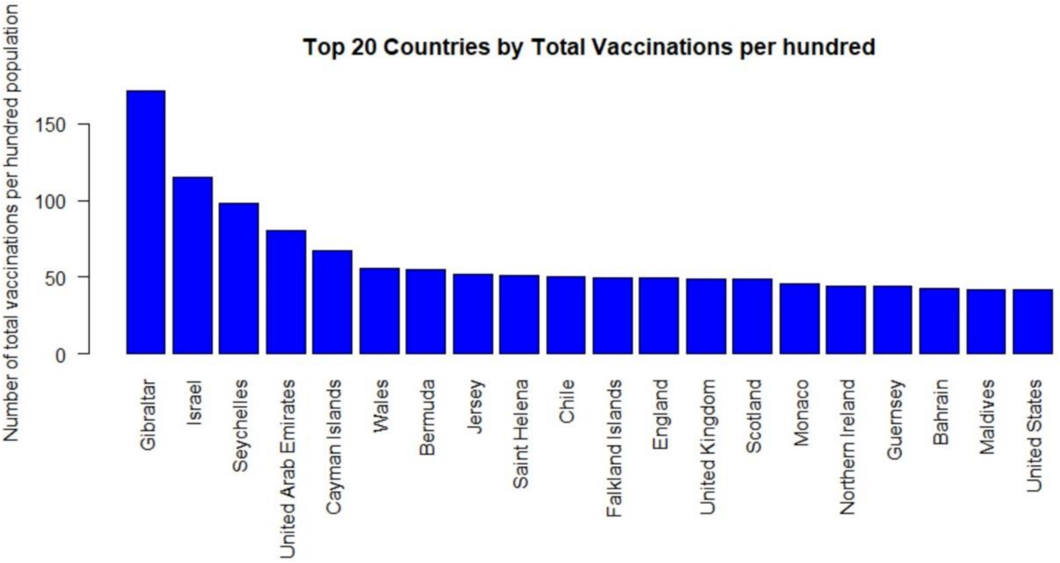

Figure 2. TOP-20 countries by Total number of vaccinations per hundred people

To run my model I decided to choose 3 countries : USA (because it has administered the most number of vaccines), France and Germany.

4. Modeling

To create my model I used a Linear Regression. Before modeling I created subsets of the main data set with the records corresponding to each country only. After that I decided to build the graphs of daily number and total number of vaccinations administered for each of the chosen country to see the correlation between them and the variable "Date" (the variable "Date" was converted to the as.POSIXct format beforehand). I also checked the correlation with the function cor(x, y). It was important to check how one variable depend on another before building the regression and choose the right predictor.

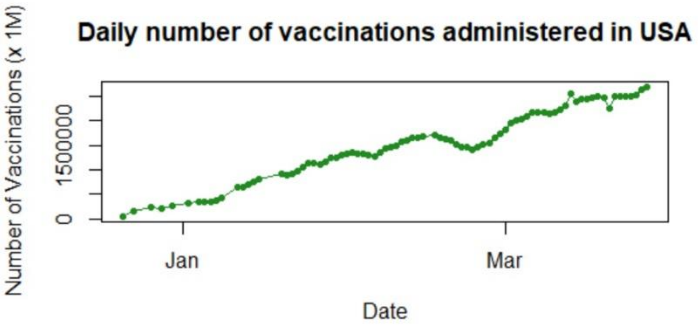

Figure 3. Daily number of vaccinations administered in USA

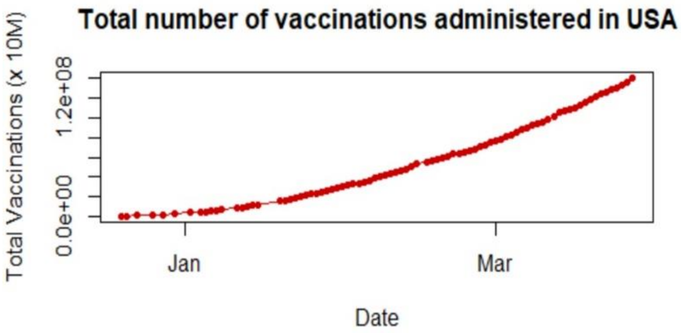

Figure 4. Total number of vaccinations administered in USA

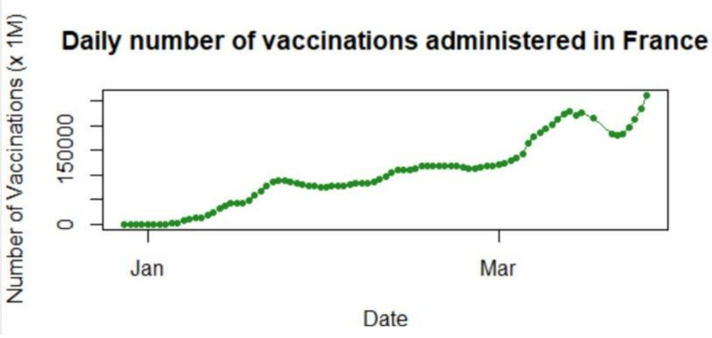

Figure 4. Daily number of vaccinations administered in France

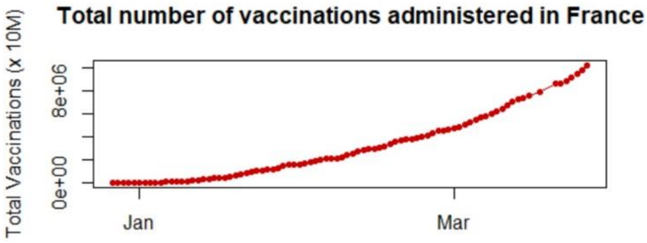

Figure 6. Total number of vaccinations administered in France

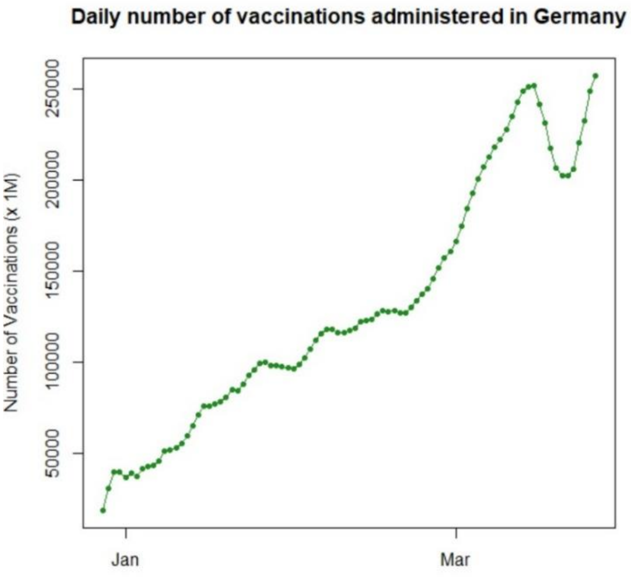

Figure 7. Daily number of vaccinations administered in Germany

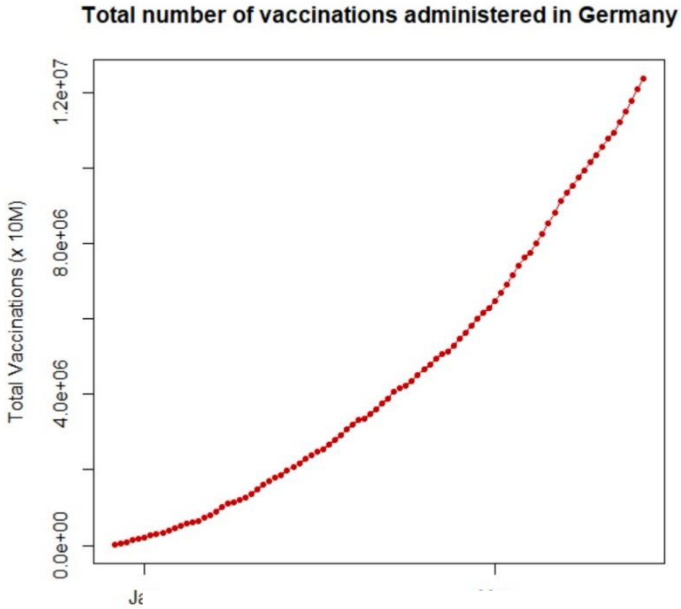

Figure 8. Total number of vaccinations administered in Germany

Germany had the most number of records in the data set, that is why I needed to use a separate window (calling the **windows()** to open it and **dev.off()** to close after) to build the graphs for it.

After the predictor was chosen (total vaccination number), I created the small data set for each country which was consisted only of the predictor and target value (in my case it's columns "date" which is target and "total_vaccinations" as a predictor).

The main problem was to present the variable "Date" as the numeric the way that I could come back to the normal date format after making a prediction. For that I took the first date in each subset as a starting point to compute the difference between it and all dates after and then I converted it to the numerical value (for this I used the function **difftime()** with units="days").

The prediction model was build using the lm function: **model <- lm(y~X, data =*country*_data_small)**, where y- variable "date", X- variable "total_vaccinations" and data is equal to the data of the corresponding country.

5. Evaluation

To evaluate the accuracy of my model I used a function **summary()** which gave me the information the R-squared scores. Since these scores were quite close to 1, I concluded that my linear regression fits well and I can use it for my predictions (Table 1). Also the p-value for each model was really close to zero which shows that I chose a good predictor.

	Linear Regression USA	**Linear Regression Germany**	**Linear Regression France**
Multiple R-squared	0.9515	0.9527	0.9282
Adjusted R-squared	0.9509	0.9522	0.9274
p-value	2.2e-16	2.2e-16	2.2e-16

Table 1. R-squared scores and p-value of the model

Plots of linear regression and its residuals (differences between observed and predicted by the model values of data) are presented on graphs below, where blue seeds on the linear regression graph represent the real behavior of the data and red line is the line built by linear regression.

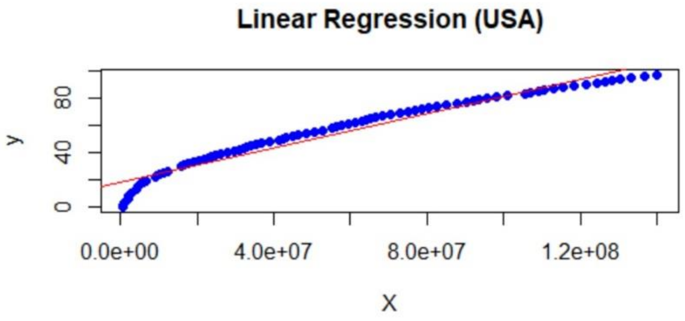

Figure 9. Linear Regression model (USA)

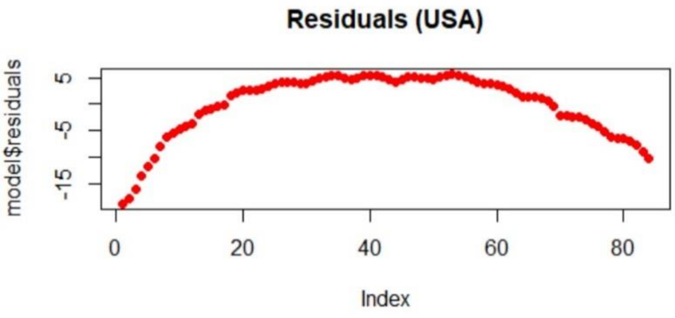

Figure 10. Residuals of the model (USA)

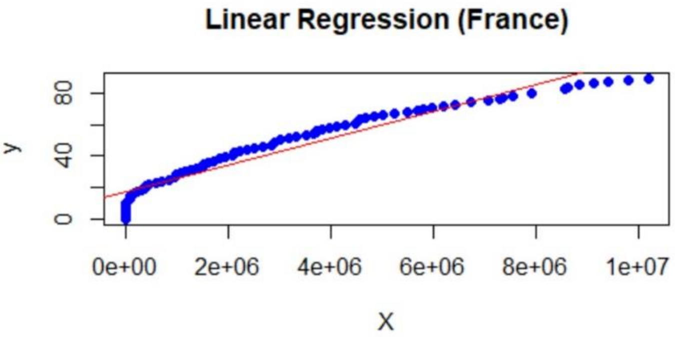

Figure 11. Linear Regression model (France)

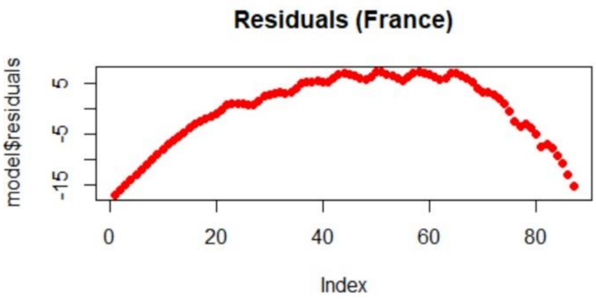

Figure 12. Residuals of the model (France)

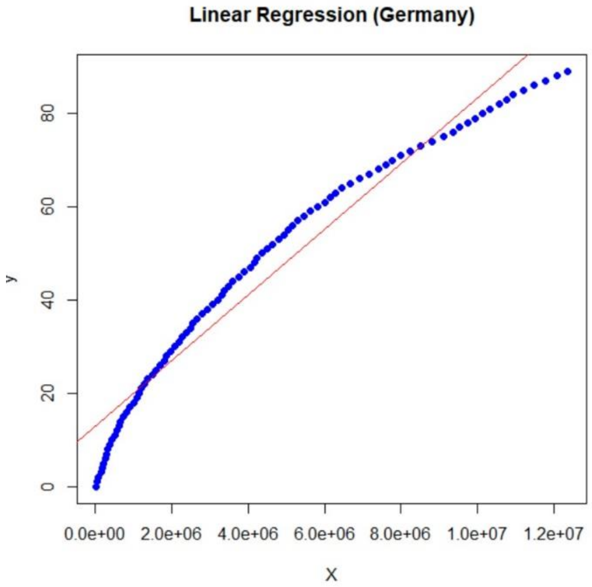

Figure 13. Linear Regression model (Germany)

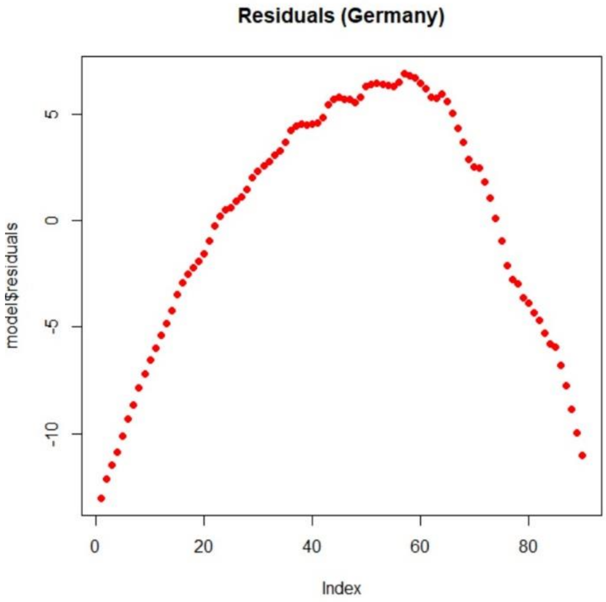

Figure 14. Residuals of the model (Germany)

6. Deployment

To predict the date when each country will achieve a herd immunity of Covid-19 I had to learn how many people of the whole country population should be vaccinated.

In the article of WHO (World Health Organization) the threshold was about 80% of the population.[3] Also, in the sources of John Hopkins Bloomberg School of Public Health was said that the vaccination rate should be around 70% of the population.[4] Since I wanted to predict the earliest date of a Covid-19 herd immunity, I took the 70% as my benchmark. Based on this knowledge, I computed the number of vaccines need for each country (considering that each person need 2 vaccines). I put the needed number of vaccines to my model and predicted the date using the function **predict()** and converting the result into the date format.

[3] https://www.who.int/news-room/q-a-detail/herd-immunity-lockdowns-and-covid-19#:~:text=The%20percentage%20of%20people%20who,among%20those%20who%20are%20vaccinated

[4] https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html

The results of my predictions are presented in the table below (Table2).

	Population of the country (x 1 M)	Number of vaccines needed (x 1 M)	The earliest date when the country will achieve a herd immunity
USA	332	464	2021-10-28
France	68	95	2023-04-01
Germany	84	117.6	2023-04-10

Table 2. Results of the predictions

As the result we can see that France and Germany have almost the same dynamic on Covid-19 vaccination and they have a chance to achieve a herd immunity at the same period of time, while USA can probably achieve it way earlier as it has administrated the biggest number of vaccines among all other countries due to today.

7. Conclusion

It was a good idea to use a Linear Regression model for the task which I chose to do as it is not hard to apply and analyze the results. The model performed really well, R-scores are close to zero and the final predictions look real. This model can also be used to make the same predictions on anu other country.