Advanced Data Science Capstone Project

Vaccination against COVID vs. New Cases of COVID

Author: Aleksandr Migunov

Part 1
Presentation for stakeholders

The Data Set

- The data set used in this project is Data on COVID-19 (coronavirus) by Our World in Data (https://github.com/owid/covid-19-data/tree/master/public/data) which contains data for every country and for the whole world taken from official sources.
- All the data is taken from the official resources and updated daily.

The Data Set

- This dataset contains: data on new cases of COVID; total cases; new vaccinations; total vaccinations; numbers of people vaccinated with, at least, one dose of vaccine; numbers of people fully vaccinated; total boosters, and a number of other information.

The Use Case

- The use case for this project is analysis of the progress of vaccination against COVID in the world (vaccination with, at least, one dose of vaccine and full vaccination) and analysis of the dynamics of new cases of COVID.
- This project has two main goals.

The Use Case

- The first goal is to analyze how the progress of vaccination affects the number of new cases of COVID, whether it is possible to stop the pandemic of COVID through vaccination or not, and if so, then, when it may happen.
- The second goal is to make a prognosis when the whole population of the world will be vaccinated (with one dose of vaccine and fully vaccinated) if the speed of the vaccination remains the same as now.

The Solution of the Use Case

- First, I analyzed the dynamics of the new cases of COVID in the world, the progress of vaccination (vaccination with one dose of vaccine, full vaccination, and with booster vaccine), and calculated Pearson correlation coefficient between them.
- Pearson correlation coefficient is a measure of linear correlation between two sets of data.

The Solution to the Use Case

- Pearson correlation coefficient always has a value between −1 and 1.
- If it is around 0, then, there is no correlation between the random variables.
- If it has a high positive value (close to 1), then, both random variables are correlated and increase of decrease together.

The Solution to the Use Case

- If Pearson correlation coefficient is has a negative value and close to -1, it means that when one variable increases, then, the other one decreases, and vice versa.
- In this part of the project, I am checking the hypothesis that the progress of vaccination decreases the number of new cases of COVID.

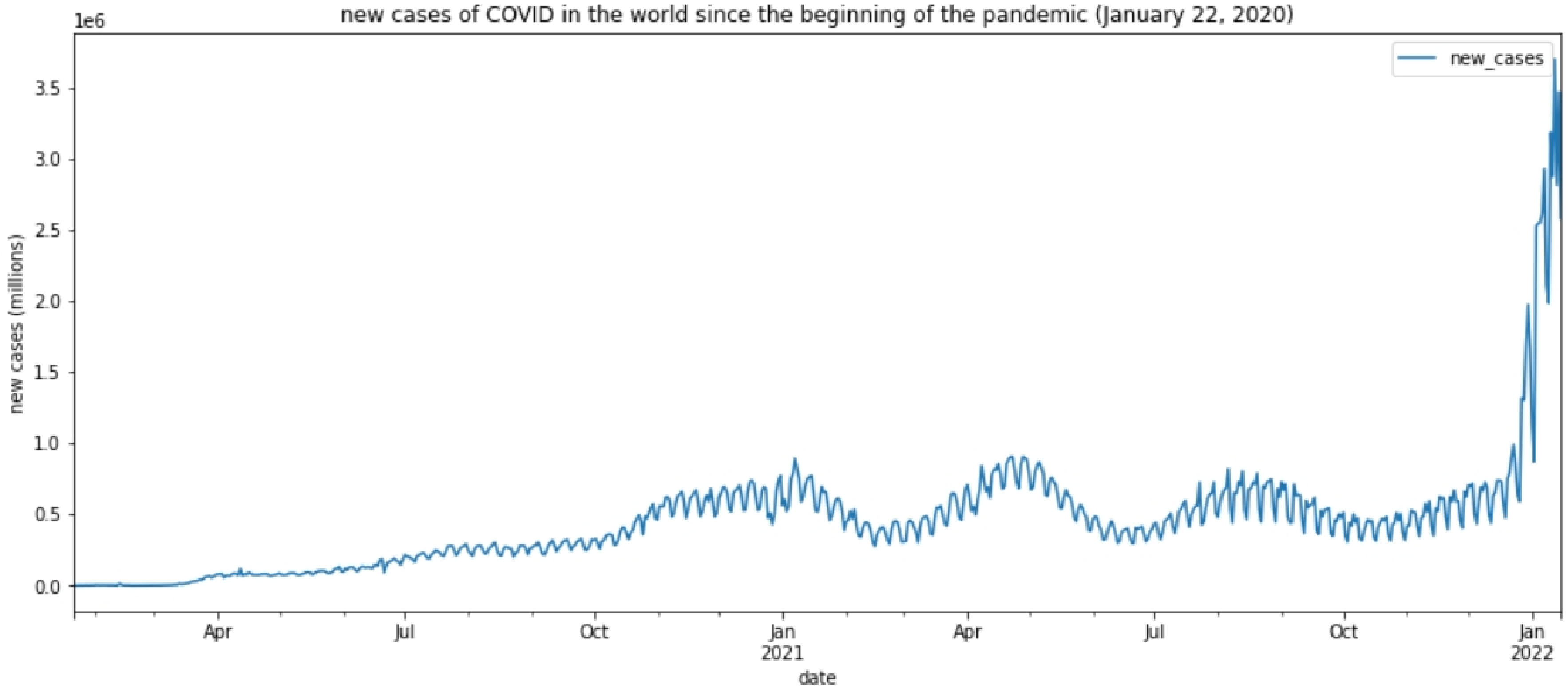
The Solution to the Use Case

- If this hypothesis is true, then, Pearson correlation coefficient must be negative. If it is close to 0 or positive, then, this hypothesis is false, which means that the progress of vaccination does not decrease the number of new cases of COVID.

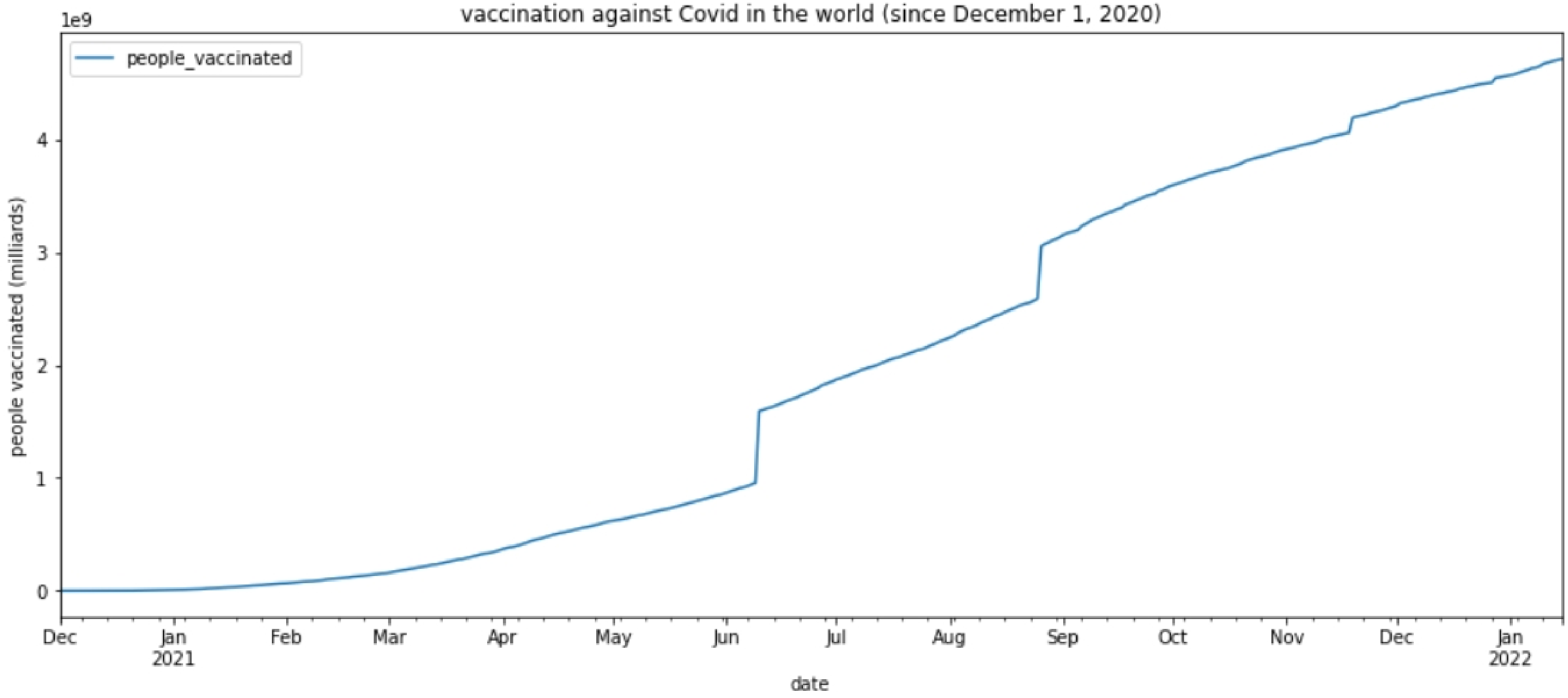
The Solution to the Use Case

- Besides the calculation of Pearson correlation coefficient, I also drew diagrams which show dynamics of new cases of COVID, progress of vaccination, and correlation between them.
- I got the following results.

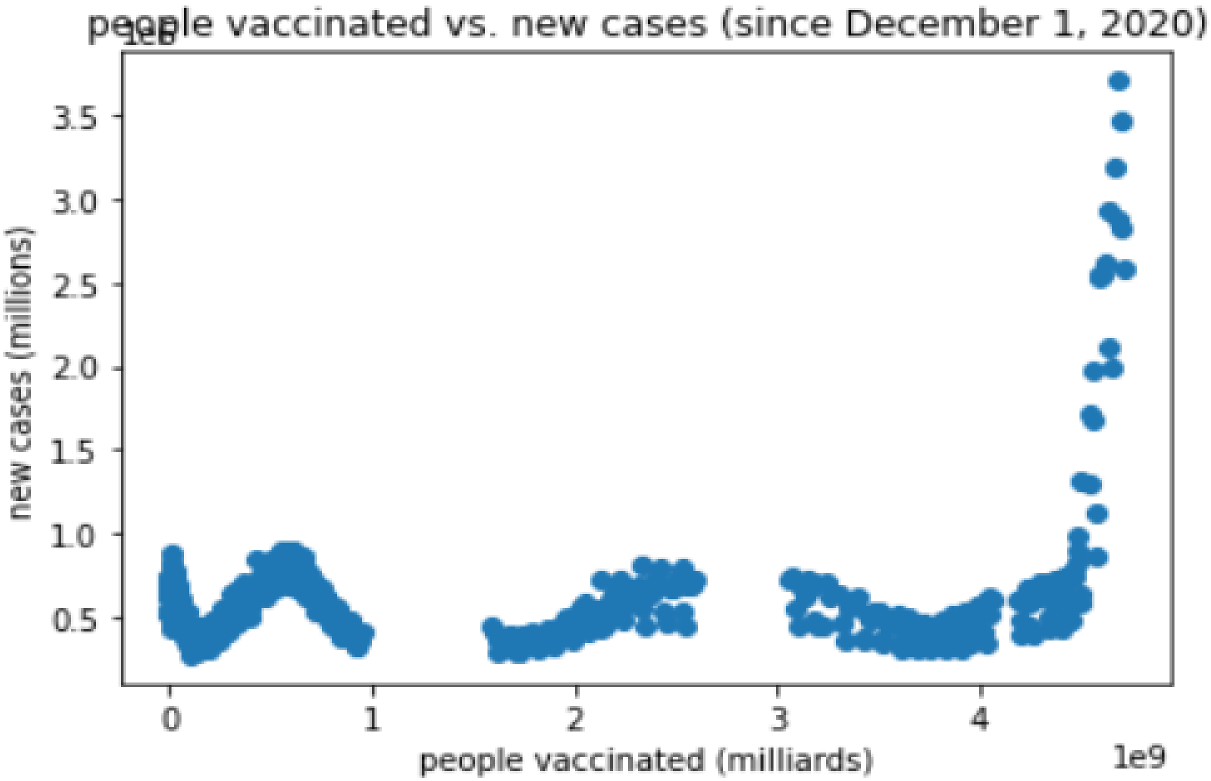
New Cases of COVID in the World



Vaccination in the World



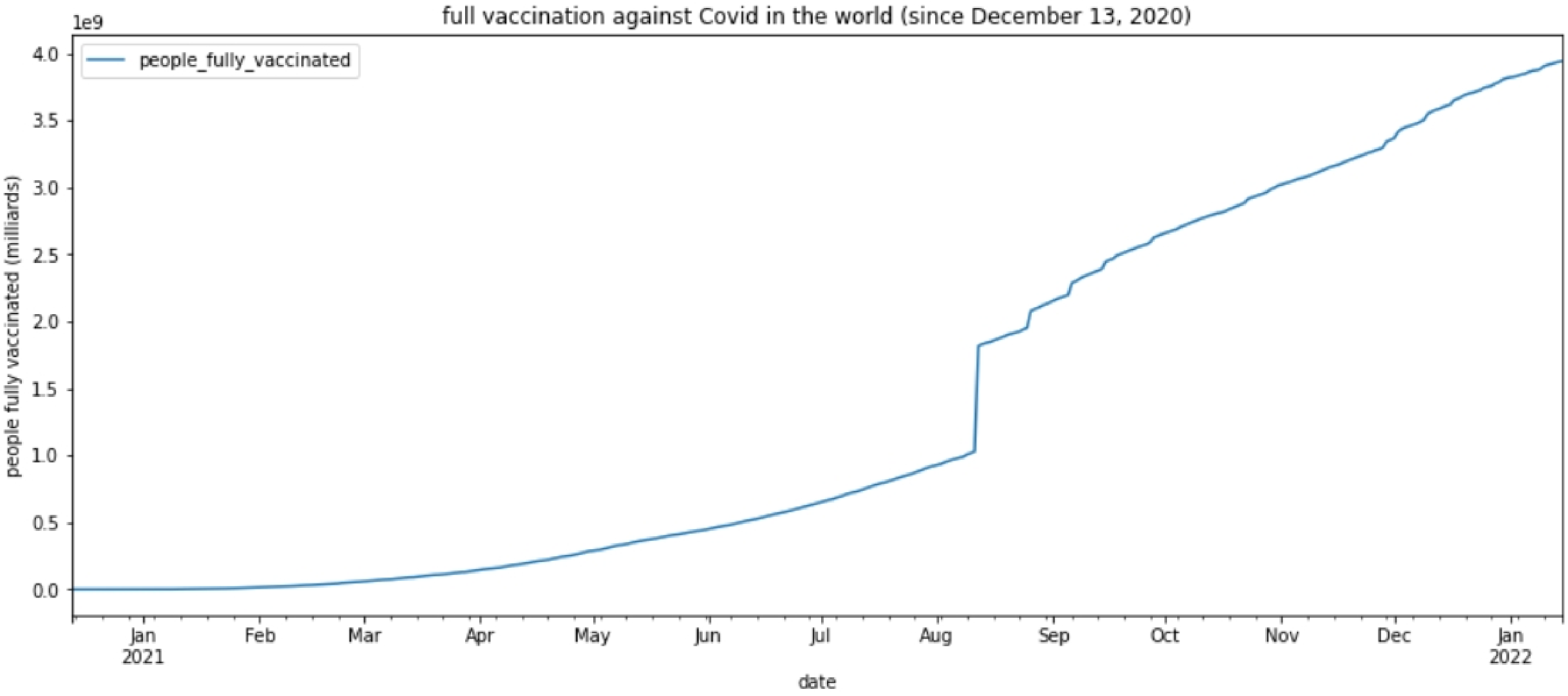
People vaccinated vs. new cases



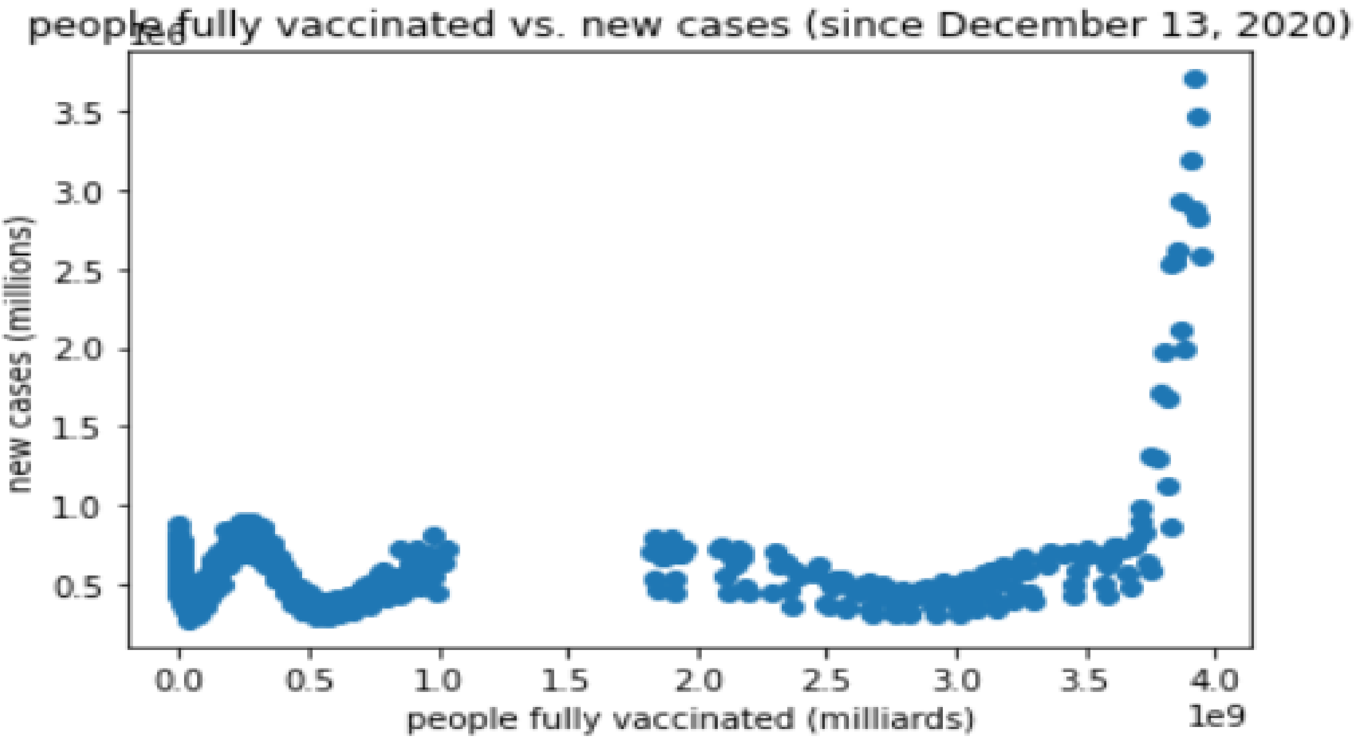
Pearson correlation coefficient

- Pearson correlation coefficient between the new cases of COVID and the number of people vaccinated with, at least one dose of vaccine is equal 0.2895.
- This means that the hypothesis that the progress of vaccination with, at least, one dose of vaccine in the world reduces the number of new cases, is false.

Full vaccination in the World



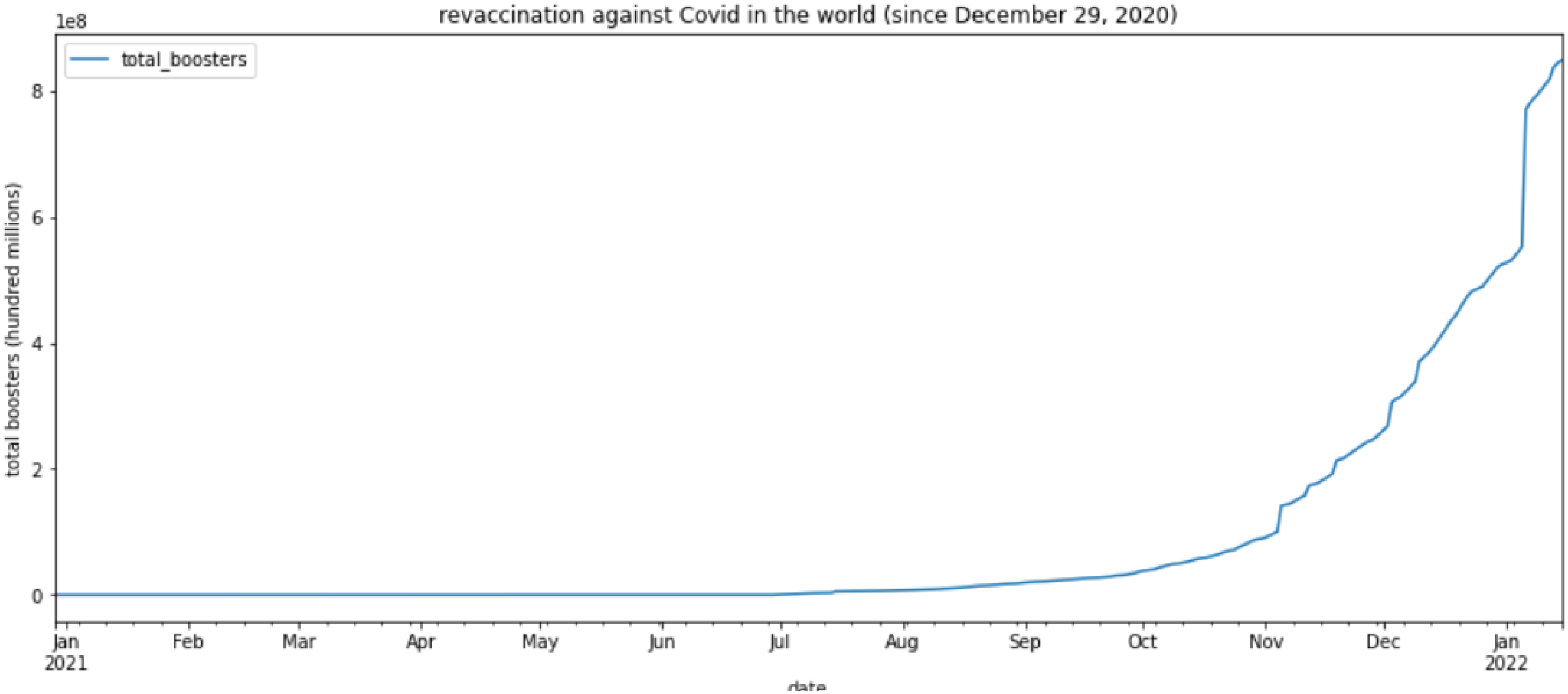
People fully vaccinated vs. new cases



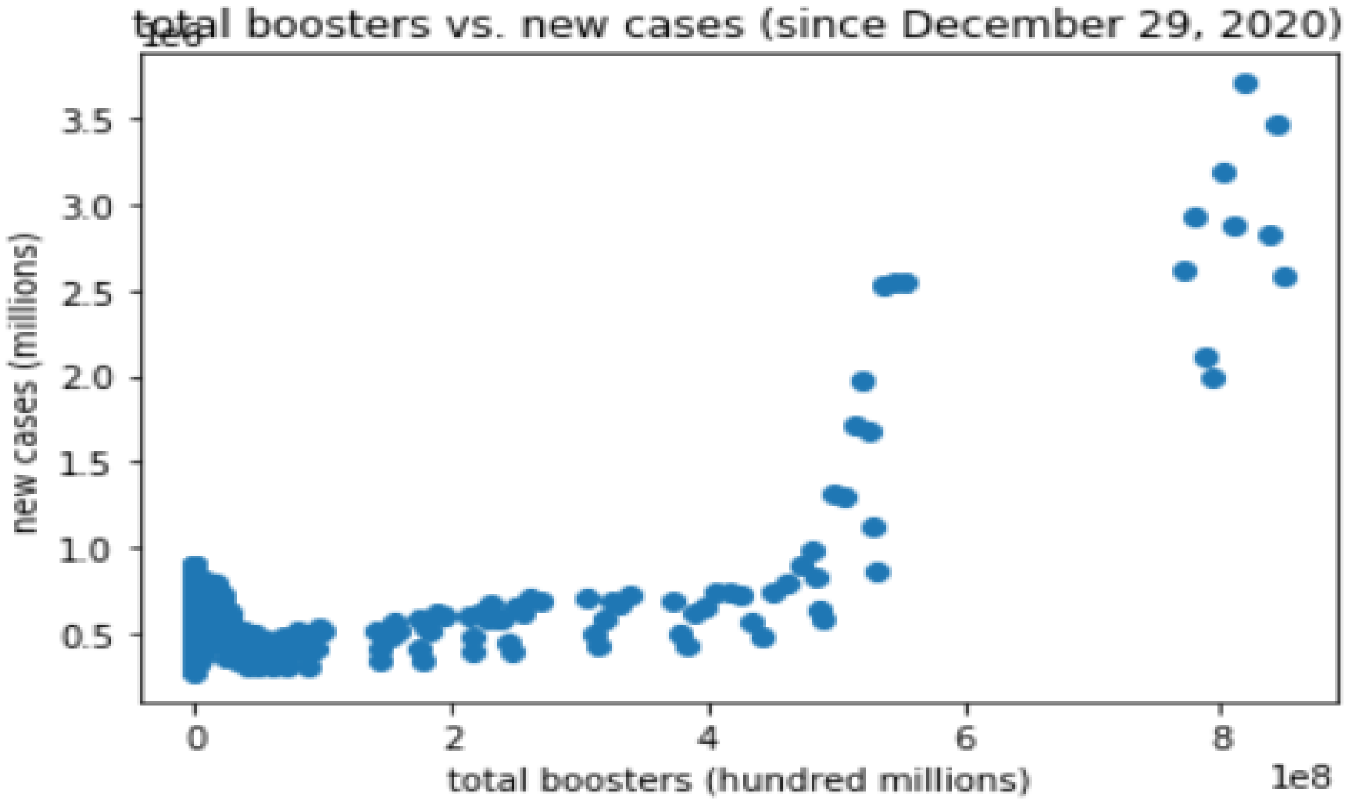
Pearson correlation coefficient

- Pearson correlation coefficient between the new cases of COVID and the number of fully vaccinated people is equal 0.3557.
- This means that the hypothesis that the progress of full vaccination in the world reduces the number of new cases, is also false.

Vaccination with booster vaccines



Booster vaccines vs. new cases



Pearson correlation coefficient

- Pearson correlation coefficient between the new cases of COVID and the total boosters is equal 0.7544.
- This means that the hypothesis that the progress of vaccination with booster vaccines in the world reduces the number of new cases, is also false.

Dynamics in individual countries

- I also analyzed dynamics of new cases, vaccination, full vaccination, and booster vaccination in several countries.
- It is similar to the dynamics in the world. Everywhere the number of new cases of COVID is either growing or very high, and the number of vaccinated people is also growing.

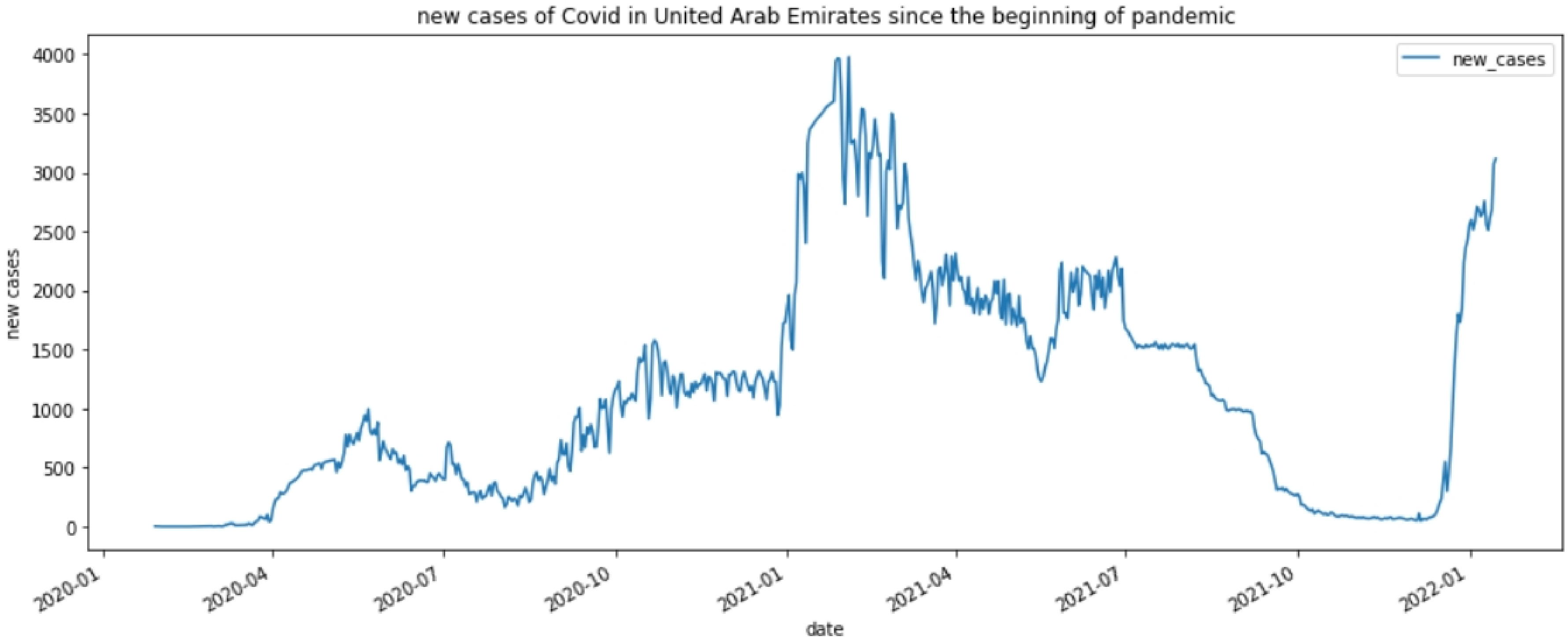
Individual countries

- Pearson coefficient coefficients between new cases of COVID and vaccination, full vaccination, and booster vaccination are either positive or close to 0.

Example of UAE

- As an example, let’s look at the United Arab Emirates.
- I chose this country because it has the highest number of vaccinated people in the world.
- In UAE, 99% of the total population are vaccinated with, at least, one dose of vaccine, and 92% of the total population are fully vaccinated.

New cases of COVID in UAE



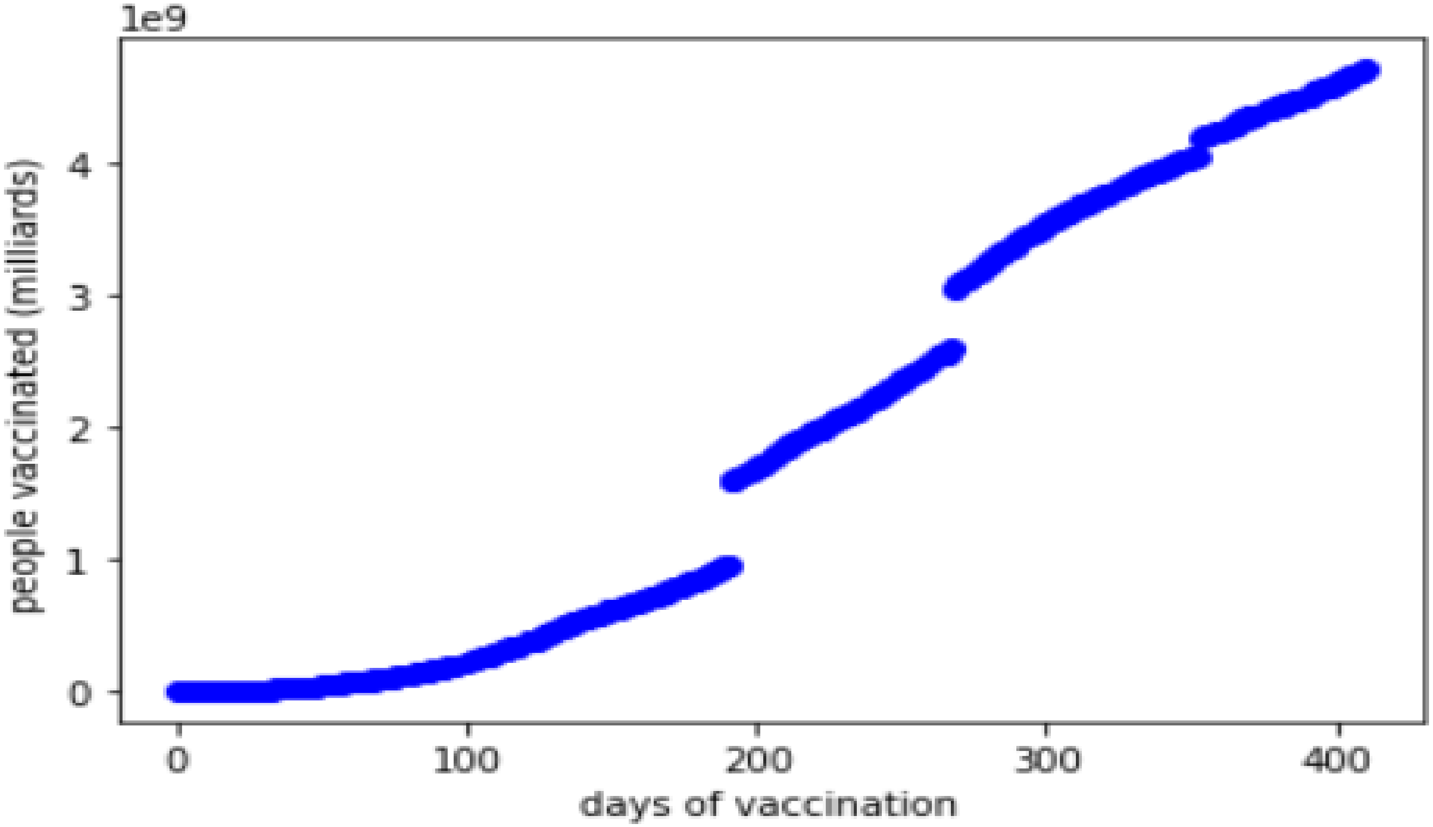
Example of UAE

- From the diagram, we see that even in the country when practically everyone is vaccinated, the number of new cases of COVID is still growing.
- This means that the progress of vaccination does not decrease the number of new cases of COVID and the pandemic of COVID cannot be stopped with vaccination.

Prognosis of vaccination

- In the next part of the project, I developed machine learning models of the progress of vaccination (with, at least, one dose of vaccine) and full vaccination.
- Then, I used these models in order to make prognosis when all the people in the world will be vaccinated if the current speed of vaccination remains the same as now.

Dynamics of vaccination



ML models of vaccination

- For modeling this function, I used linear regression model and then polynomial regression model with different number of degrees.
- I divided the dataset into train and test datasets.
- I used the train dataset for training the models.
- Then, I used the test datasets for evaluating performance of the models.

Evaluation

- For evaluation of the model performance, I used R squared.
- R squared ranges between 0 and 1.
- The higher is R squared, the better is model performance.
- The highest possible value of R squared is 1, and therefore the best models have R squared very close to 1.

ML models of vaccination

- I began with linear regression model. I defined it, trained it, and evaluated its performance. Then, I did the same with polynomial regression models, beginning with squared polynomial regression model.

Evaluation

- The linear regression model had R squared = 0.9637.
- The squared polynomial regression model had R squared = 0.9778.
- The cubic polynomial regression model had R squared = 0.9947.
- The 4-degree polynomial regression model had R squared = 0.9950.

Evaluation

- The 5-degree polynomial regression model had R squared = 0.9955.
- Although model performance increases with the increase of the degree of polynomials, polynomials with high degrees give a very bad approximation of the function for times that are higher than used in the training and test datasets, and therefore cannot be used for prognosis.

Evaluation

- They perform well within the datasets, but for higher values they either very rapidly grow or very rapidly drop.
- In the first case, they give too optimistic prognoses – that the vaccination of the whole population of the earth will be completed in a few days, which is obviously not realistic.

Evaluation

- In the second case, they give too pessimistic prognoses – that the progress of vaccination will be stopped very quickly and therefore only very new few people will be vaccinated.
- Both scenarios are incorrect, and therefore I did not use polynomial regression models with polynomials of degrees = 6 and higher.

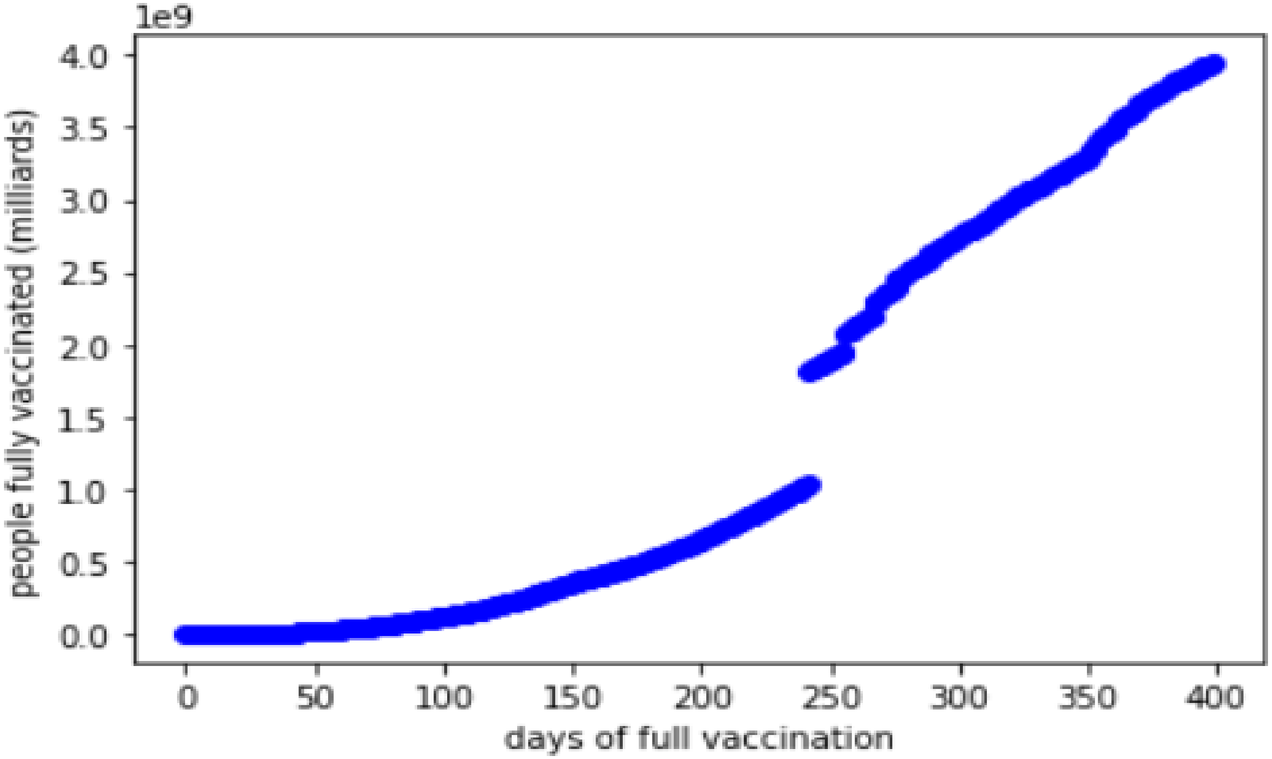
Prognosis

- According to the linear model vaccination will be completed in 100 days.
- According to the quadratic model vaccination will be completed in 107 days.
- According to the 5-degree model vaccination will be completed in 95 days.
- Other two models (the cubic and the 4-degree) do not give usable results.

Prognosis

- Therefore, three models of vaccination (with, at least, one dose of vaccine) give similar results.
- If the speed of the vaccination remains the same, then, the whole population of the world will be vaccinated (with, at least, one dose of vaccine) in 95-107 days.

Dynamics of full vaccination



ML models of full vaccination

- I did the same steps for the models of full vaccination.
- I first developed linear regression model and then polynomial regression models with polynomials of different degrees.
- I trained them, evaluated them, and used them to make prognoses.

Evaluation

- The linear regression model had R squared = 0.8961.
- The quadratic polynomial regression model had R squared = 0.9680.
- The cubic polynomial regression model had R squared = 0.9798.
- The 4-degree polynomial regression model had R squared = 0.9880.

Evaluation

- The 5-degree polynomial regression model had R squared = 0.9885.
- So, in general, performance of models of full vaccination was worse than performance of models of vaccination (with, at least, one dose of vaccine).

Prognosis

- According to the linear model full vaccination will be completed in 234 days.
- According to the quadratic model full vaccination will be completed in 123 days.
- According to the 5-degree model full vaccination will be completed in 457 days.
- Other two models (the cubic and the 4-degree) do not give usable results.

Prognosis

- So, the three models of full vaccination give very different results.
- Although 5-degree polynomial model has the best performance, its result (457 days) seems to be quite pessimistic, considering that all the models of vaccination (with, at least, one dose of vaccine) do not give the result of more than 107 days.

Prognosis

- I tried other kinds of models for full vaccination, but I could not find a satisfactory model for it.
- However, since we already have a satisfactory prognosis for vaccination (with, at least, one dose of vaccine), we can use it.

Prognosis

- Since most vaccines require two doses (and some require only one dose) and the interval between the two doses usually should be 14, 21, or 28 days, it can be concluded that the full vaccination should be completed about 28 days after the vaccination with one dose of vaccine.

Prognosis

- The result for prognosis of vaccination with, at least one dose of vaccine was that if the speed of the vaccination remains the same, then, the whole population of the world will be vaccinated (with, at least, one dose of vaccine) in 95-107 days.
- Therefore, we can conclude that if the speed of the vaccination remains the same, then, the whole population of the world will be fully vaccinated in 135 days or sooner.

Part 2
Presentation for peers

Architectural decision

- I used the Lightweight IBM Cloud Garage Method for Data Science.
- It has following seven steps:

Architectural decision

- Initial Data Exploration
- Extract, Transform, Load
- Feature Creation
- Model Definition
- Model Training
- Model Evaluation
- Deploy Model

Architectural decisions

- I wrote the details of the architectural decisions in Architectural Decisions Document

Data quality assessment

- Accessibility: data can be easily and quickly retrieved from OWID website.
- Appropriate amount of data: the volume of data is appropriate for the tasks of this project. The dataset available at OWID website contains all the necessary data.
- Believability: all the data is from official sources and therefore is reliable.

Data quality assessment

- Completeness: data which is necessary for the tasks of this project is complete.
- Objectivity: since all the data is from official sources, it can be assumed that it is unbiased, unprejudiced, and impartial.
- Relevancy: the data is applicable for the tasks of this project.
- Timeliness: the data is up-to-date, it is updated daily.

Data pre-processing

- Data cleaning: there were some missing data (NaNs) in the dataset. I dealt with them in different ways: I removed the columns that contained only NaNs; if the necessary data was missing until a certain date, I removed the rows that contained NaNs; in other cases, I replaced NaNs with zeros.

Data pre-processing

- Data integration: since I used only one dataset, there was no need in this procedure.
- Data reduction: since the dataset is not very large (around 40 MB), there was no need in data reduction.

Feature engineering

- Since the dates in the dataset were in object type which cannot be used directly for machine learning, I transformed them into date_delta in float64 type.
- I tried data normalization, but it did not improve performance of the models and made calculation of prognoses more complicated. So, I decided to use data without normalization.

Model algorithm

- Since the machine models need to predict the number of people and since the data that I am using are labeled, I need to use regression. To begin with, I chose linear regression models and also polynomial regression models with different degrees polynomials (quadratic, cubic, etc.). Other kinds of non-linear regression (exponential, logarithmic, sigmoid, etc.) would not work for these models because the diagrams of growing the number of people vaccinated and people fully vaccinated with time, show that other functions (except linear and polynomial) will not be fitting.

Model algorithm

- Since my models are very simple – the number of people vaccinated and fully vaccinated depends only on time, and since the datasets are quite small (containing data only for about 400 days), there is no need to use deep learning models. So, I used only machine learning models in this project.
- I also tried Lasso and Ridge, but they did not give any improvement of performance of the models.

Model performance indicators

- I used R square since it is quite convenient indicator for regression models which gives scores in range between 0 and 1, where 1 is the best possible performance.