Course Project on Exploratory Data Analysis for Machine Learning

Subject: Analysis of data on COVID

Brief description of the data set and a summary of its attributes

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, for every continent and for the whole world. All the data is taken from official sources and updated regularly (most of data is updated daily).

- This data set contains:
- Confirmed cases and deaths (total confirmed cases of COVID-19, new confirmed cases of COVID-19, total deaths attributed to COVID-19, new deaths attributed to COVID-19, etc.)
- Hospitalizations and intensive care unit (ICU) admissions (number of COVID-19 patients in ICUs, number of COVID-19 patients in hospital, etc.)
- Testing for COVID-19 (total tests for COVID-19, new tests for COVID-19, etc.)
- Vaccinations against COVID-19 (total number of COVID-19 vaccination doses administered, total number of people who received at least one vaccine dose, total number of people who received all doses prescribed by the initial vaccination protocol, total number of COVID-19 vaccination booster doses administered (doses administered beyond the number prescribed by the vaccination protocol), etc.)
- Other variables (population, etc.)

Initial plan for data exploration

My goal for data exploration in this project is to analyze dynamics of the new cases of COVID in the world, progress of vaccination against COVID (vaccination with at least one dose of vaccine, full vaccination, and buster vaccination). Also, data will be prepared for making machine learning model (regression) to model progress of vaccination.

The initial plan for data exploration includes:

- Data Collection. This step includes downloading data set from Internet and its examination.
- Data Cleaning. This includes dealing with NaNs, outliers, etc.

- Analysis. This includes drawing diagrams of dynamics of new cases of COVID and progress of vaccination.
- Feature Engineering. In this stage, data will be prepared for machine learning model.

Actions taken for data cleaning and feature engineering

Data cleaning. The main problem with data in this data set is that there are many NaNs. I used three strategies to deal with them:

1) removing columns that contain only NaNs;
2) replacing NaNs with zeros;
3) in order not to deal with many zeros in the beginning, I created several new datasets: 1) the dataset for analysis and visualization of cases of COVID - since the beginning of pandemic; 2) the dataset for analysis and visualization of the number of people who got vaccinated (by, at least, one injection of vaccine) - since the beginning of vaccination; 3) the dataset for analysis and visualization of the number of people who got fully vaccinated - since the beginning of full vaccination; 4) the dataset for analysis and visualization of the number of total boosters - since the beginning of revaccination.

Feature engineering. In this stage, I prepare data for developing models of the progress of vaccination and full vaccination in the world.

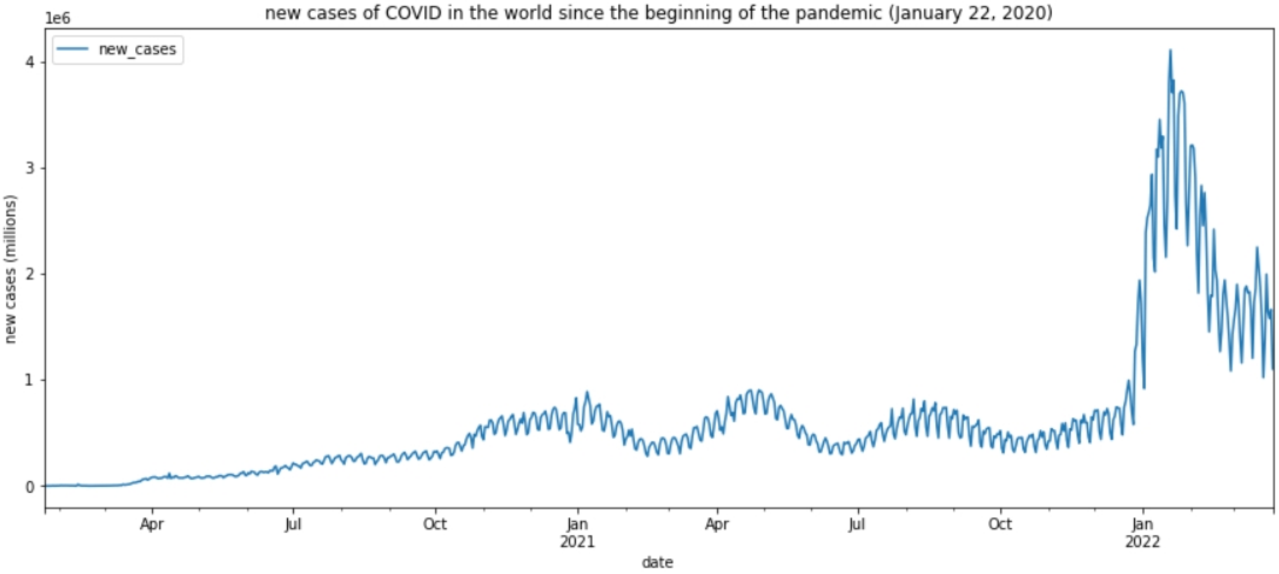
1. For machine training, dates need to be transformed from object type into float64 type (date_delta). Other data in the dataset are already in float64 type and do not need transformation.
2. Also, I create new datasets for machine learning that contain only people vaccinated / people fully vaccinated, dates, and date_delta.

Key Findings and Insights, which synthesizes the results of Exploratory Data Analysis

The most important findings as results of EDA were drawing diagrams that show dynamics of new cases of COVID and progress of vaccination, and also correlation between them which will be described in the next points.

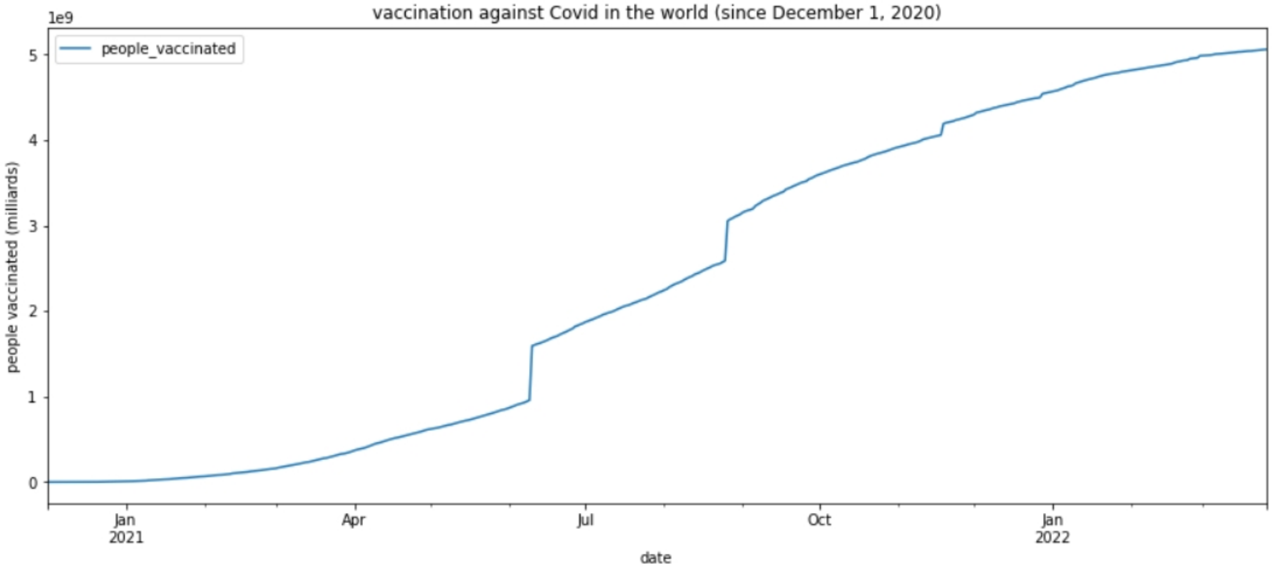
1. The diagram of dynamics of new cases of COVID in the world since the beginning of the pandemic (that is, since January 22, 2020) is shown in the next page.

It shows that in January-February 2022 the number of new cases of COVID was the highest during the whole time of pandemic. The number of new cases is now decreasing but it is still very high and much higher than the number of new cases of COVID in 2020 and 2021. It also shows that the number of cases of COVID in January-October 2020 was quite small, but since October 2020, the number of cases of COVID became much larger with periods of increasing and decreasing. Between October 2020 and October 2021, there were three waves of COVID with approximately similar number of new cases. But the fourth wave, which began in January 2022, the number of new cases increased approximately in 4 times.

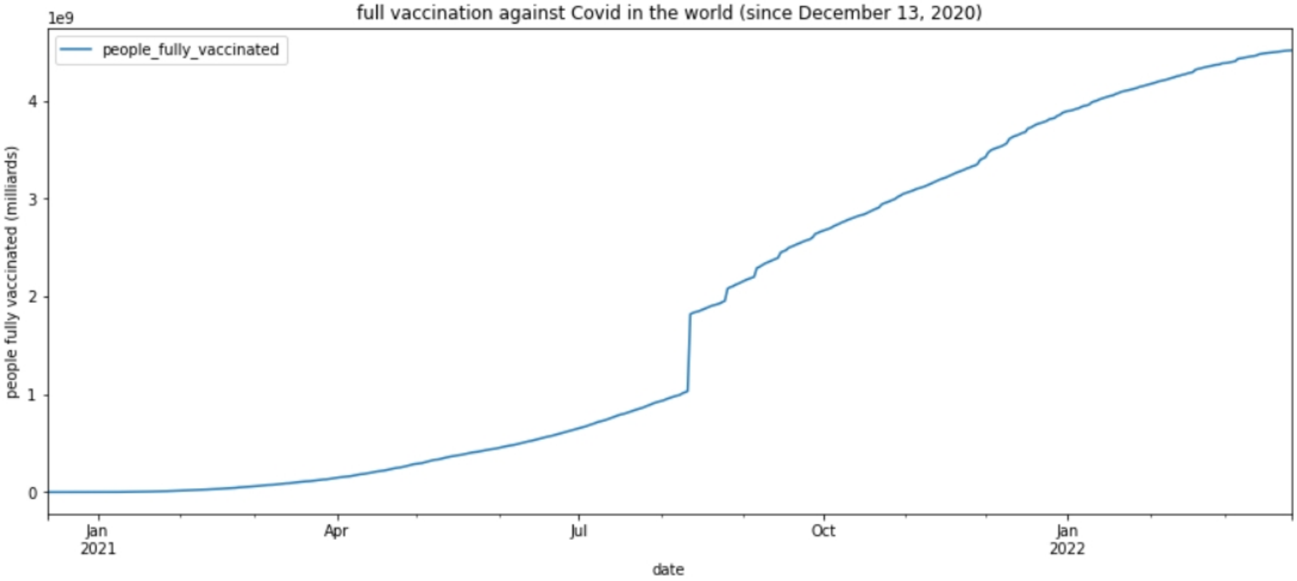


2. The diagrams of vaccination (with at least one dose of vaccine), full vaccination and booster vaccination in the world look like shown in the next page.

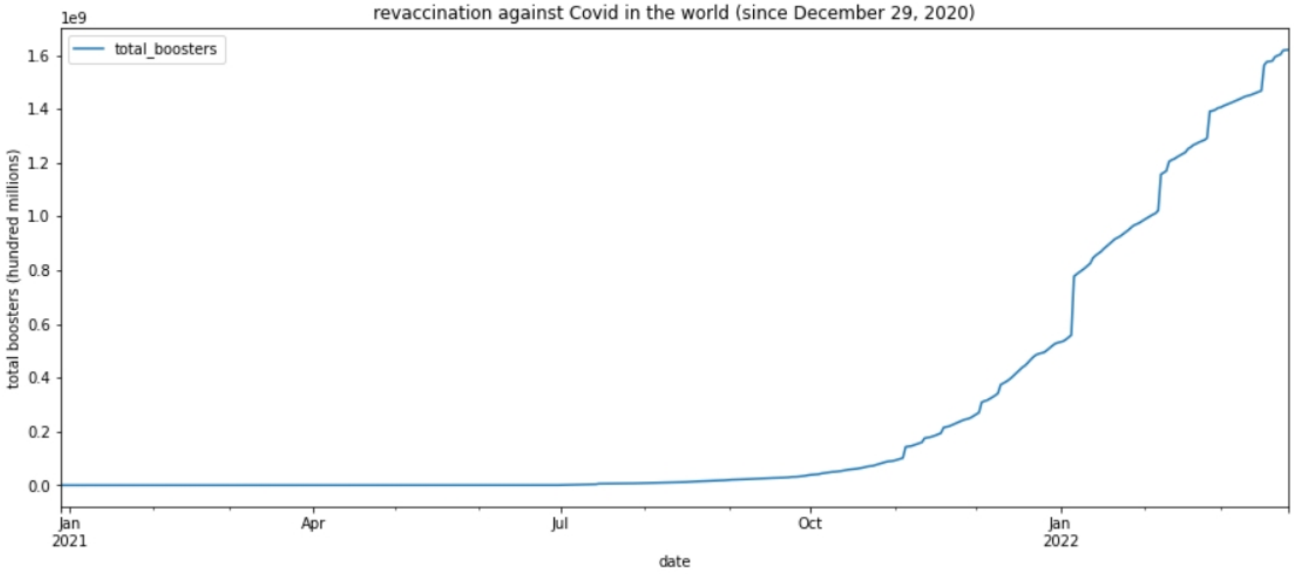
1) The diagram of the number of people in the world who were vaccinated with, at least, one does of vaccine:



2) The of the number of people in the world who were fully vaccinated:



3) The of the number of total booster vaccinations in the world:



For these diagrams, I used data beginning with the first date for which there is data in the data set for the number of people who were vaccinated, fully vaccinated, and for booster vaccinations.

The numbers of vaccinated and fully vaccinated people show similar dynamics. However, the number of booster vaccinations was growing very slowly in January-October 2021 and since November 2021 it began to grow much faster.

Formulating at least 3 hypothesis about this data

Hypothesis 1. There is a correlation between vaccination and new cases of COVID.

Null: there is no correlation between vaccination and new cases of COVID.

Alternative: there is a correlation between them.

Hypothesis 2. The more people are vaccinated, the less new cases of COVID.

Null: growth of the number of vaccinated people does not affect the number of new cases of COVID.

Alternative: growth of the number of vaccinated people decreases the number of new cases of COVID.

Hypothesis 3. There is a correlation between the number of fully vaccinated people and booster vaccinations.

Null: the number of fully vaccinated people and booster vaccinations are not correlated.

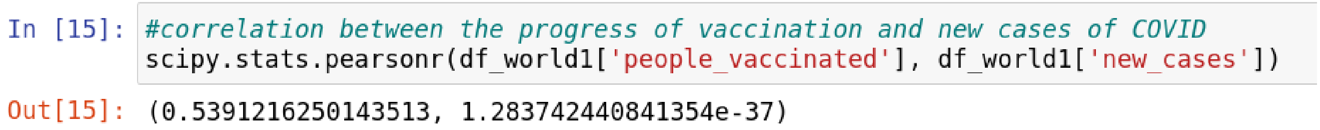
Alternative: there is a correlation between them.

Conducting a formal significance test for one of the hypotheses and discuss the results

Test for hypothesis 1 (about correlation between vaccination and new cases of COVID).

In order to test this hypothesis, I calculated Pearson correlation coefficients and drew plots.

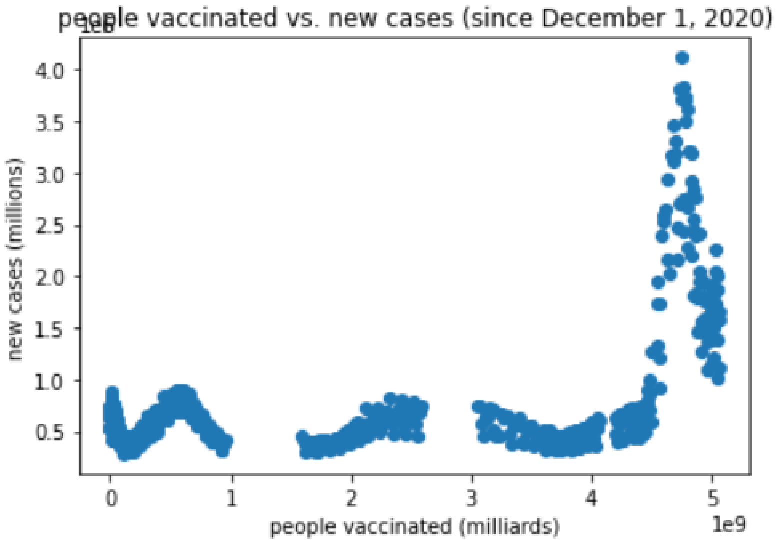
1) For the number of people who were vaccinated with at least one dose of vaccine:



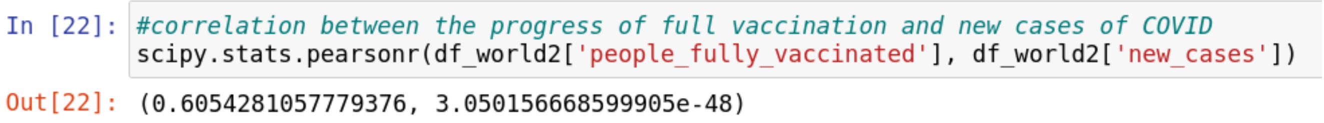
```
In [15]: #correlation between the progress of vaccination and new cases of COVID
         scipy.stats.pearsonr(df_world1['people_vaccinated'], df_world1['new_cases'])

Out[15]: (0.5391216250143513, 1.283742440841354e-37)
```

The first value here is Pearson correlation coefficient, the second one is p-value. This result indicates that there is correlation between the number of people vaccinated with at least one dose and the number of new cases. In addition, the value of Pearson correlation coefficient is positive, which indicates that the growth of vaccinated people does not decrease the number of new cases of COVID.

The plot of the number of people vaccinated vs. new cases of COVID shows the same things (see the next page).



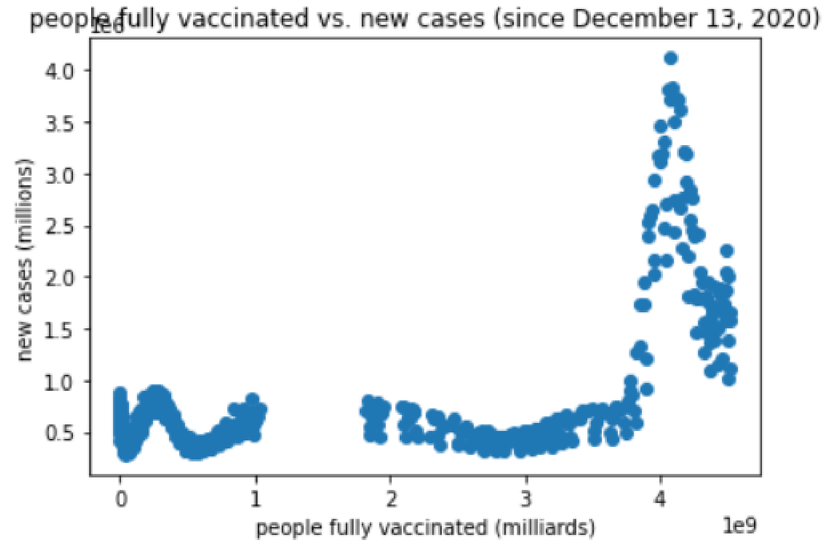
2) For the number of people who were fully vaccinated:



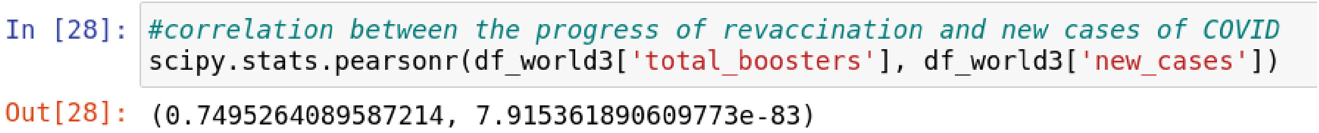
```
In [22]: #correlation between the progress of full vaccination and new cases of COVID
         scipy.stats.pearsonr(df_world2['people_fully_vaccinated'], df_world2['new_cases'])

Out[22]: (0.6054281057779376, 3.050156668599905e-48)
```

This result is similar to the previous one. It indicates that there is correlation between the number of people who were fully vaccinated and the number of new cases. The value of Pearson correlation coefficient is positive, which indicates that the growth of fully vaccinated people does not decrease the number of new cases of COVID.

The plot of the number of people fully vaccinated vs. new cases of COVID shows the same things (see the next page).



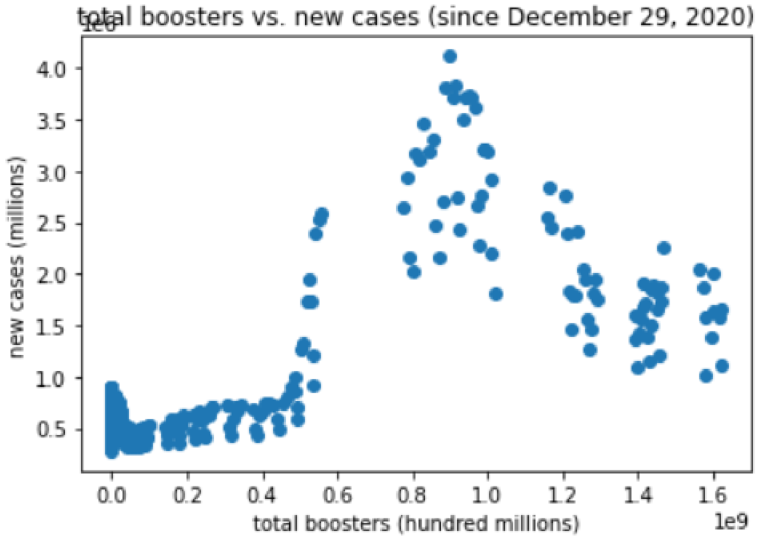
3) For the number of booster vaccinations:



```
In [28]: #correlation between the progress of revaccination and new cases of COVID
         scipy.stats.pearsonr(df_world3['total_boosters'], df_world3['new_cases'])

Out[28]: (0.7495264089587214, 7.915361890609773e-83)
```

This result is similar to the previous two. It indicates that there is correlation between the number of total booster vaccinations and the number of new cases of COVID. The value of Pearson correlation coefficient is positive and therefore indicates that the growth of people who received booster vaccinations does not decrease the number of new cases of COVID.

The plot of the number of total boosters vs. new cases of COVID shows the same things (see the next page).



Therefore, there is correlation between the number of new cases of COVID and the number of people vaccinated with at least one dose of vaccine, fully vaccinated, and vaccinated with booster vaccines. That is, there is correlation between the number of new cases of COVID and progress of vaccination. However, the growth of the number of vaccinated people does not reduce the number of cases of COVID.

Suggestions for next steps in analyzing this data

The next step should be developing machine learning models for the progress of vaccination (vaccination with one dose of vaccine, full vaccination, and with booster vaccines). Also, it could be good to try to develop a model of dynamics of new cases of COVID, but it does not seem to be an easy task now.

Also, since the pandemic of COVID is not over yet, dynamics of new cases will be changed over time. And since vaccination against COVID also continues, the numbers of vaccinated people will keep increasing. The data set used in this project is being updated every day. Therefore, it would be good to repeat analyzing this data in future.

Summary of the quality of this data set and a request for additional data if needed

The data in this data set are taken from official sources and therefore are reliable. Since all the data is from official sources, it can be assumed that it is unbiased, unprejudiced, and impartial. The data is relevant and applicable for the tasks of this project. The data is accessible, it can be easily and quickly retrieved from OWID website. The data is up-to-date, it is updated daily.

However, diagrams show some problems with data for numbers of people vaccinated with at least one vaccine, fully vaccinated people, and total booster vaccinations. All of them are smooth most of the time, but also have some “surges” when the number of vaccinated people in one day is much higher that the number in the previous day. This does not happen because of missing data. No values are missing there. But this data set has this problem. I did not find other datasets that would have different values. But it might be good to try to find values for each country from other sources and then calculate the values for the world, which might be different. However, of course, there is no guarantee that it will correct this problem.

In addition, this dataset contains only the numbers of total booster vaccines but not the number of people who had booster vaccinations. Since there are people who got more than one booster vaccine, these numbers are not equivalent. It is not very convenient to compare the number of people who got vaccinated or fully vaccinated with the number of total booster vaccines. Therefore, it would be good to find the data for the number of people who got booster vaccines.