Covid-19 Exploration Data Analysis

Ali Bakhshesh

Preprocessing

The dataset we used In this project includes lots of *Null* values in some columns so, at the first we specified a threshold to decide which columns should be omitted.

We decided to remove columns with more than 60000 *Null* values.

Some other preprocessing methods have been done after EDA phase such as deleting outliers.

Data Exploration

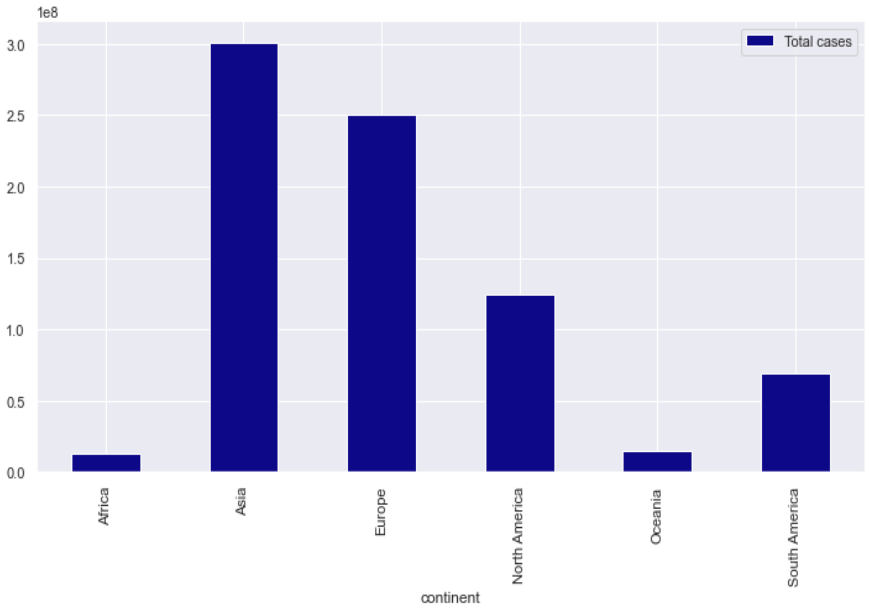
After doing preprocessing following columns left:

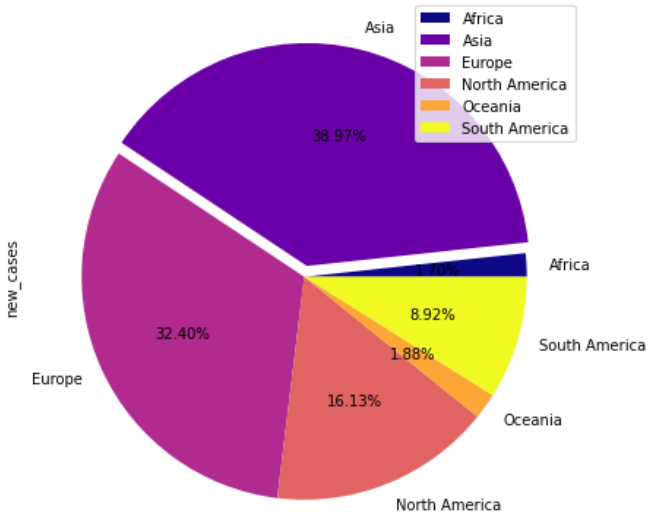
1. Iso code
2. Continent
3. Location
4. Date
5. Total cases
6. New cases
7. New cases smoothed
8. Total deaths
9. New deaths
10. New deaths smoothed
11. Total cases per million
12. New cases per million
13. New cases smoothed per million
14. Total deaths per million
15. New deaths per million
16. New deaths smoothed per million
17. Reproduction rate
18. Total tests
19. New tests
20. Total tests per thousand
21. New tests per thousand

22. New tests smoothed
23. New tests smoothed per thousand
24. Positive rate
25. Tests per case
26. Tests units
27. Total vaccinations
28. People vaccinated
29. People fully vaccinated
30. New vaccinations
31. New vaccinations smoothed
32. Total vaccinations per hundred
33. People vaccinated per hundred
34. People fully vaccinated per hundred
35. New vaccinations smoothed per million
36. New people vaccinated smoothed
37. New people vaccinated smoothed per hundred
38. Stringency index
39. Population density
40. Median age
41. Aged 65 older
42. Aged 70 older
43. Gdp per capita
44. Extreme poverty
45. Cardiovasc death rate
46. Diabetes prevalence
47. Female smokers
48. Male smokers
49. Handwashing facilities
50. Hospital beds per thousand
51. Life expectancy
52. Human development index
53. Population

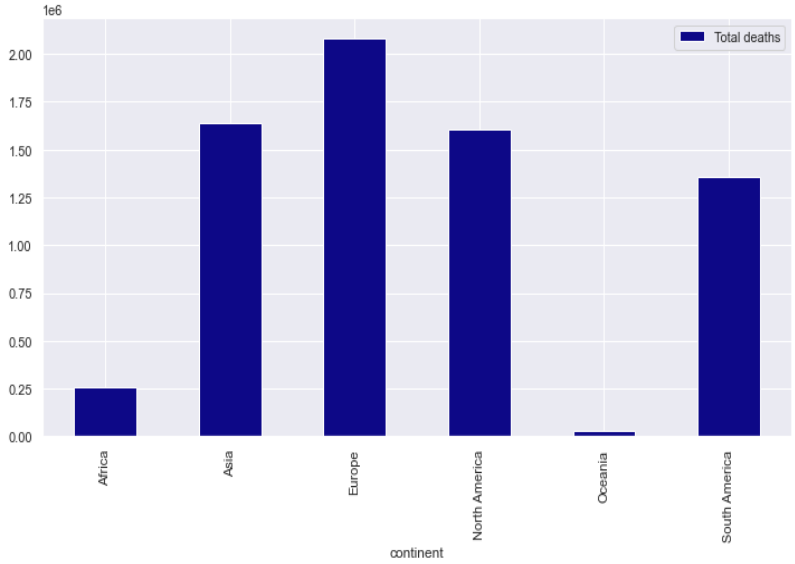
In this project we mostly tried to analyze and compare the epidemy of covid-19 in each continent and we also want to recognize the effect of some factors on the spreading of this catastrophic virus.

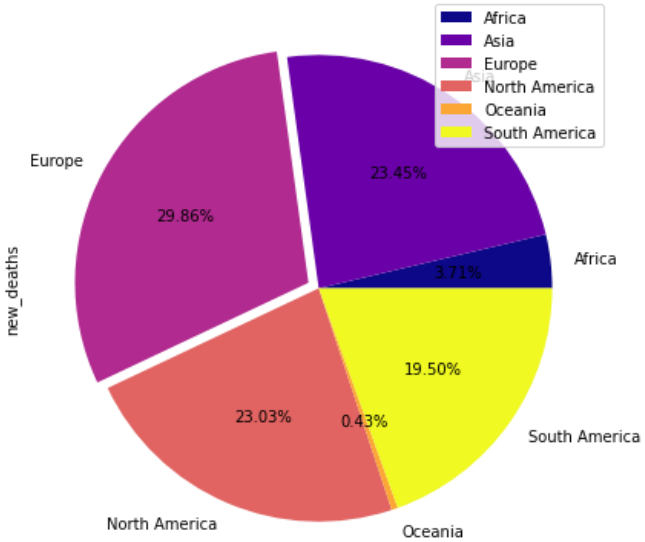
We will start with comparing Total cases in each continent:





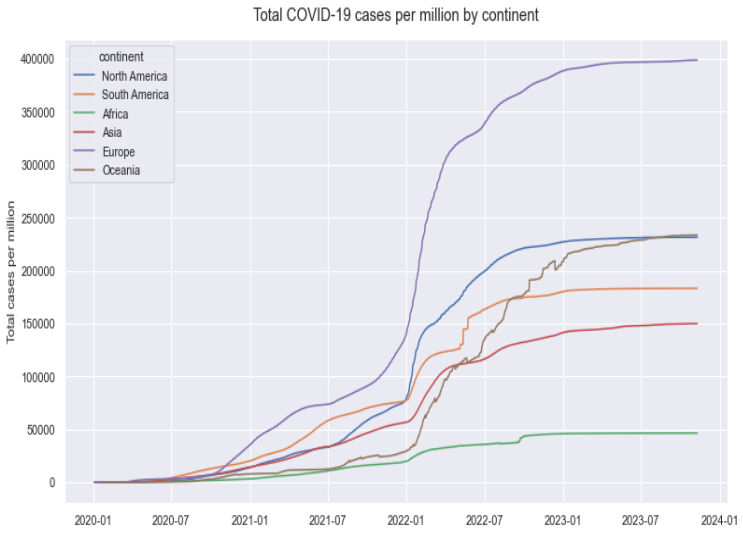
As it is clear there is a significant difference between the number of patients in Asia and Europe and other continents so it is expected that the number of death in these two continents be more than others so we will draw the plot of total deaths in each continent:

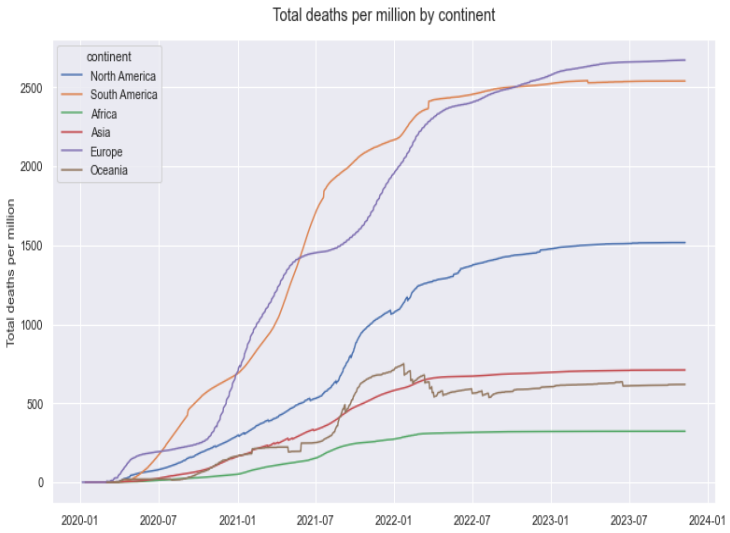




The consequences are almost as they were expected but there is a little difference, Aisa has the most total cases, but Europe has the most total deaths. Generally, we can say that continents with more patients have had more deaths.

Following plots shows the ascending trend of the speared and death of covid-19 per million in each continent.

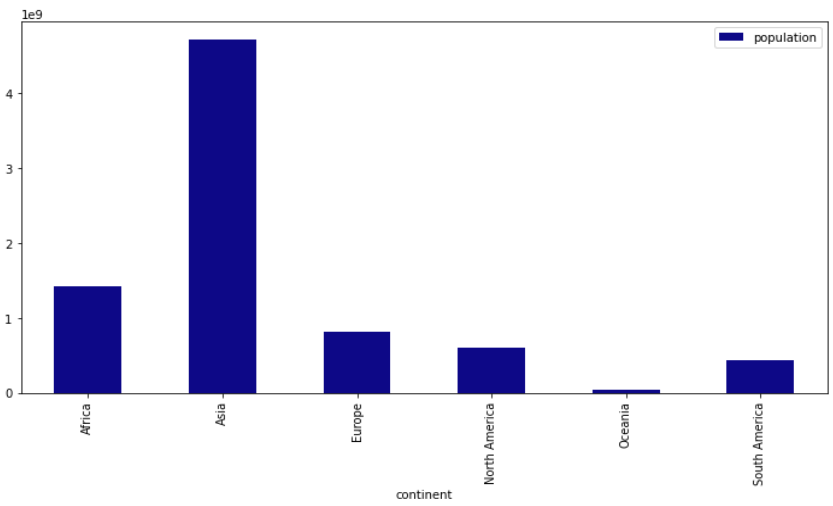


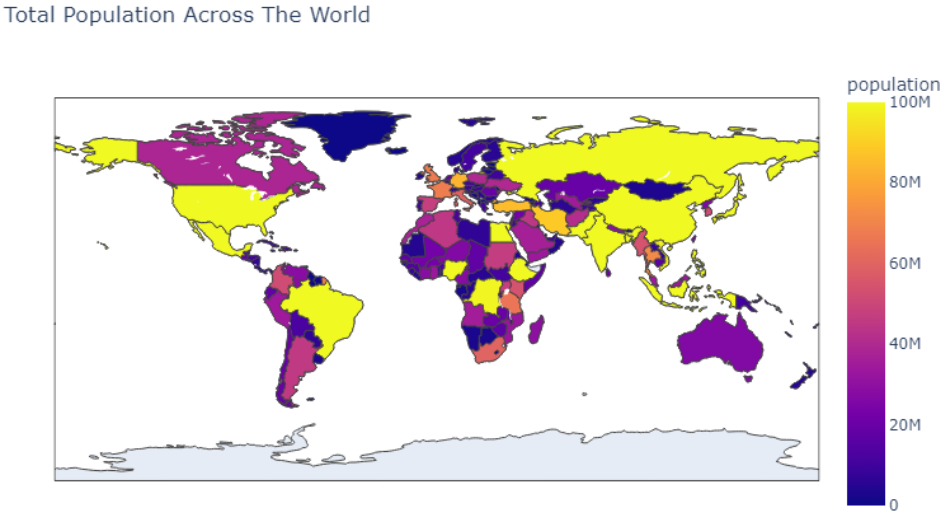


In the following we will examine the effect of several factors on the epidemy and death of covid-19.

Population:

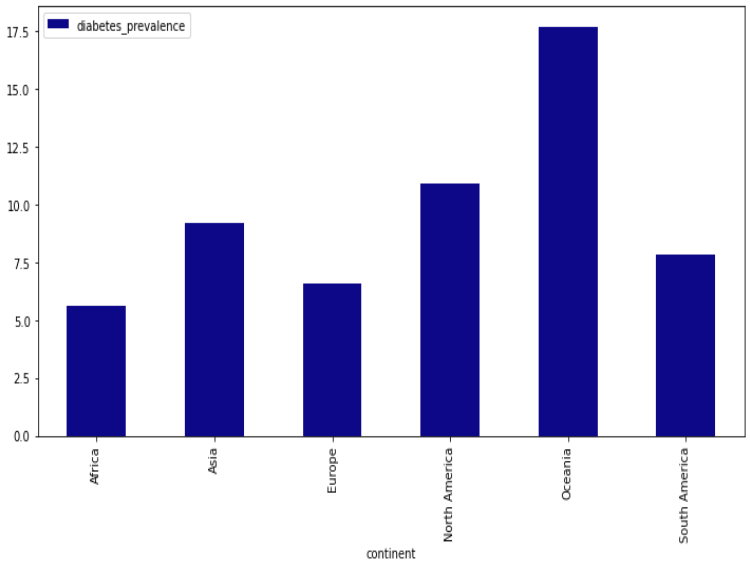
We expect that covid-19 will have a wider prevalence in more populous continents. In the blow you can see the bar plot of each continent population:

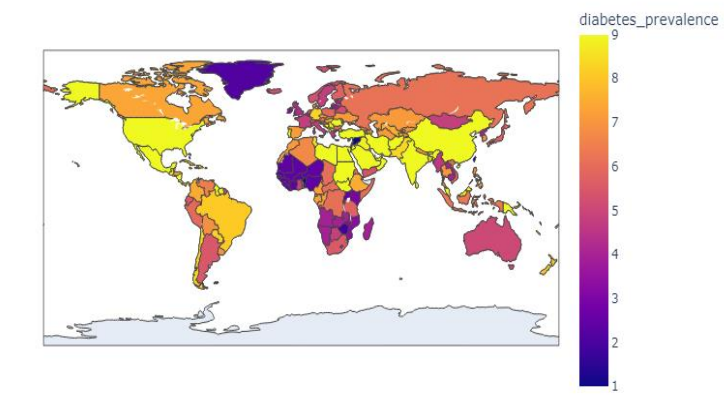




The result of this examination is weird. Asia with a significant difference in population size has the most total cases compared to other continents but about other continents we see some thing else. Also, in the population size of other continents there is no significant difference but it is clear that Africa with the largest population after Asia has the lowest patients. It could be because of the number of tests done in each country but unfortunately, we do not have enough information about this factor.

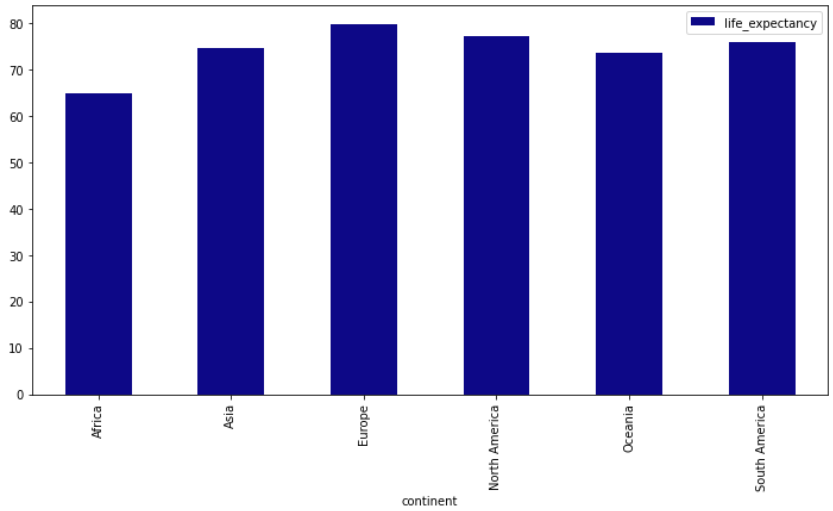
Diabetes prevalence:



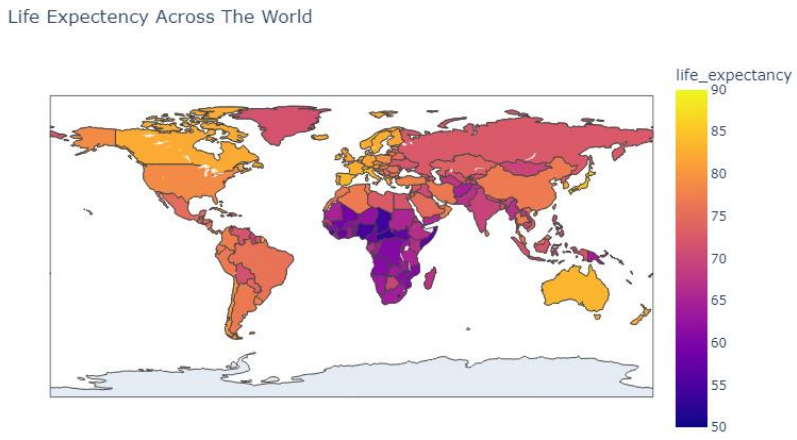


By looking at these plots we cannot see any relation between the spread of covid-19 and diabetes.

Life expectancy:

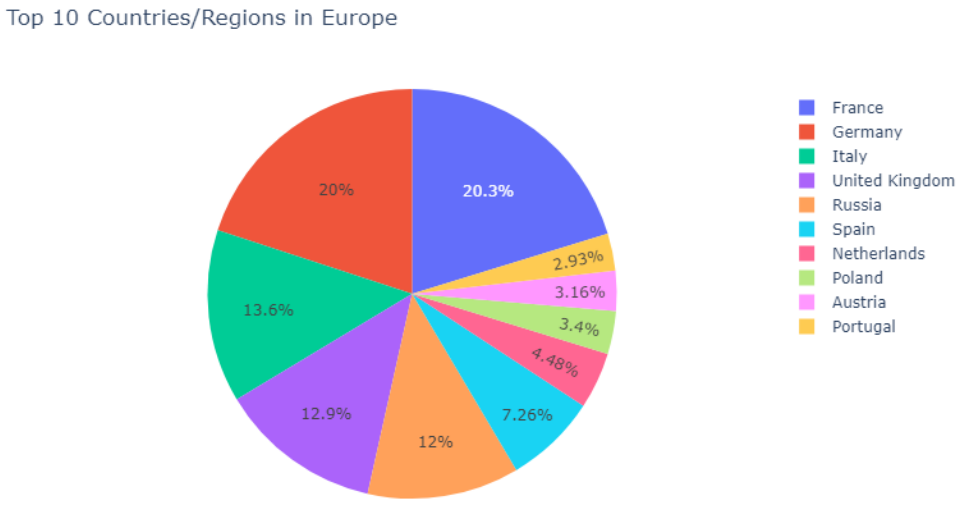


This plot shows the average amount of life expectancy in each continent. Like other factors examined up to now we do not see any relation between life expectancy and total patients, but it sounds like we've discovered something important. Continents with more life expectancy have more total deaths. So, it means that covid-19 is more dangerous for older people than young people. The following map shows the amount of life expectancy of each country.

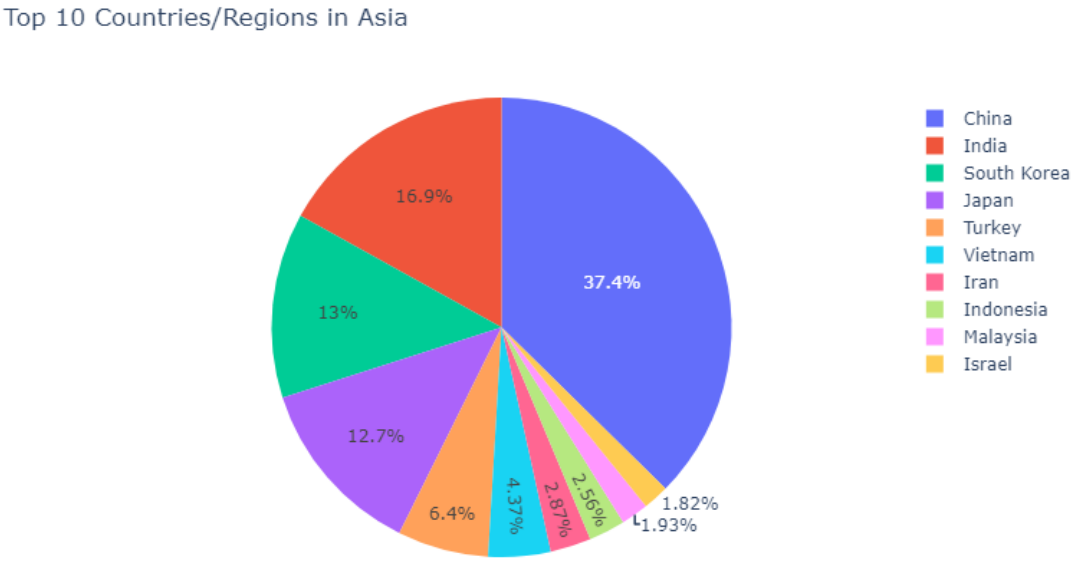


At the end of this part, we will analyze the circumstances of countries in continents with more total cases (*Europe* and *Asia*).

Europe:



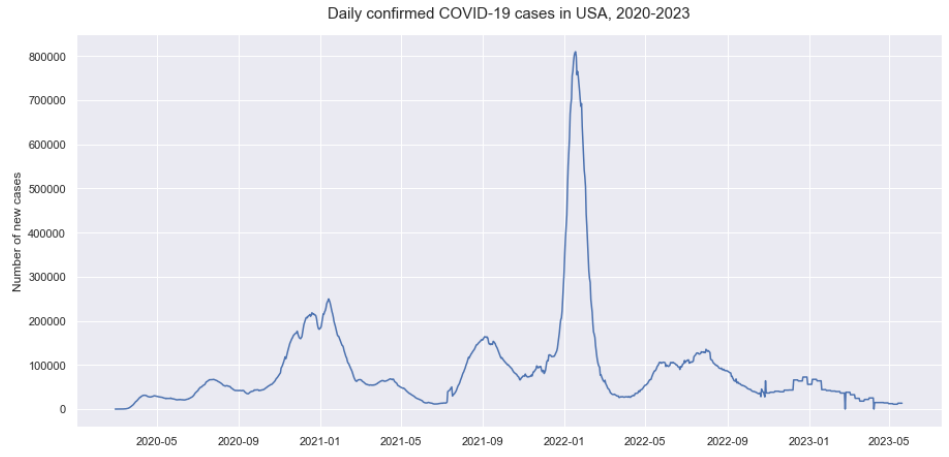
Asia:

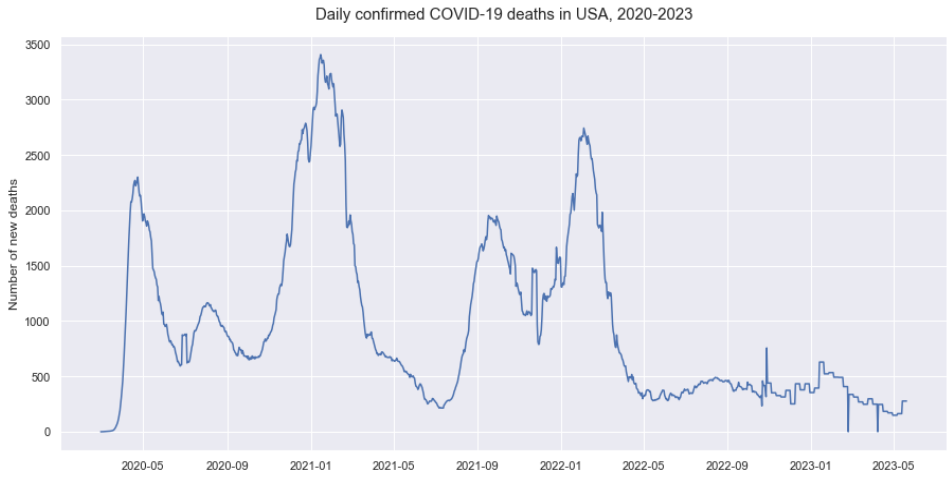


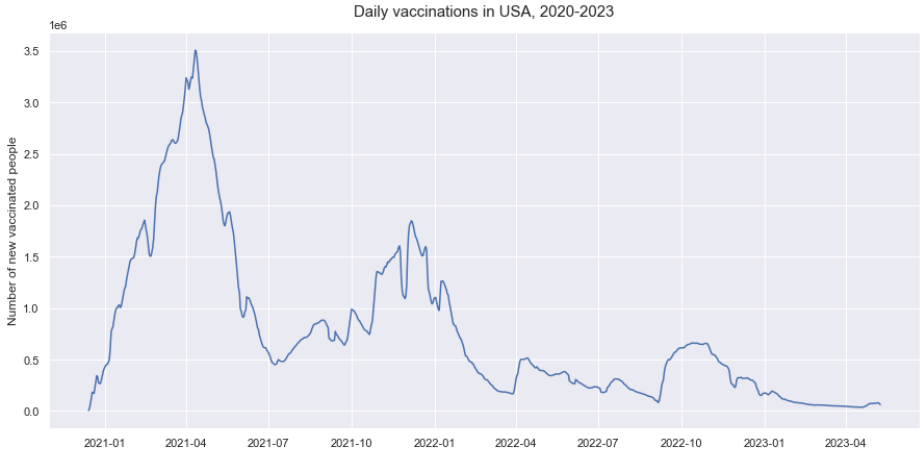
In the first part we examined the effect of population on the number of patients in each country and we couldn’t speak about it with certainty but now by looking at Asia and Europe plots it is so obvious that more populous countries have more total patients.

As the last part we are supposed to study the trend of covid-19 cases and deaths in the United States from beginning of the epidemy to 2023.

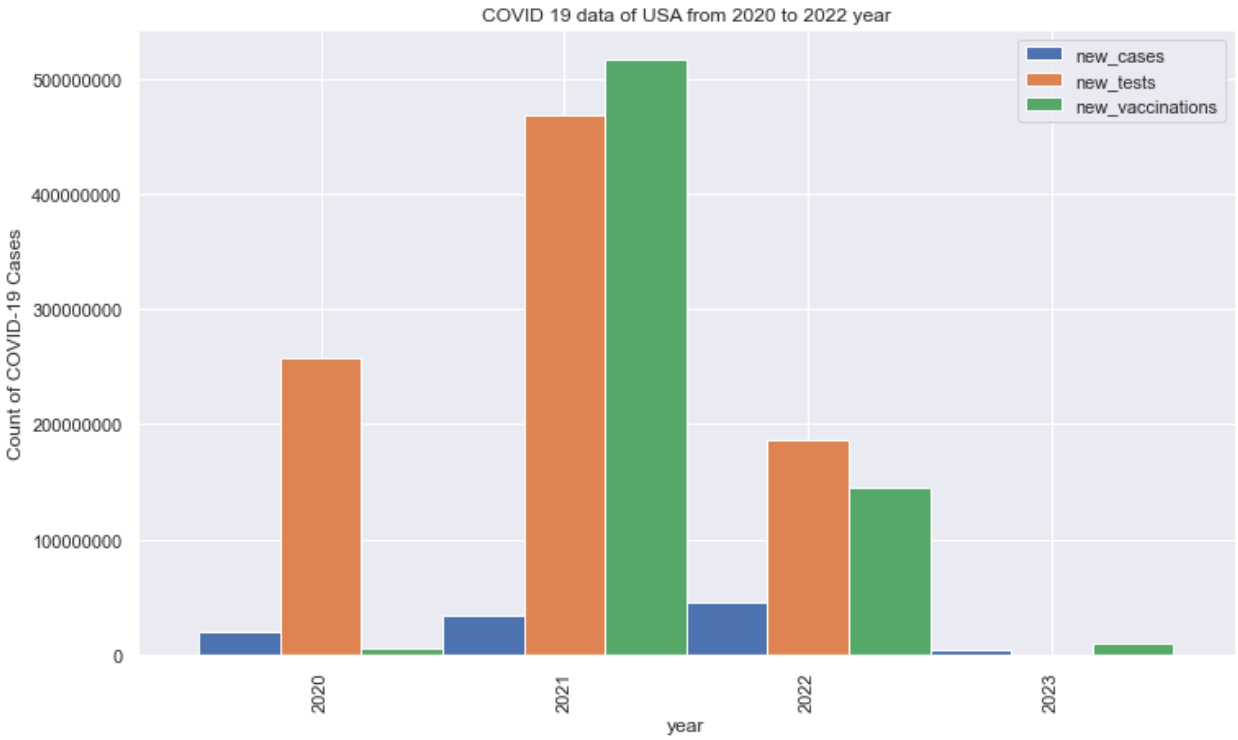
Following plots shows the number of daily cases, daily deaths and daily vaccinations in the United States:







The impact of vaccination is completely obvious. As we can see although second pick of the disease is more severe, the number of deaths is less so this demonstrates the effect of the vaccinations. This effect is clearer in the following plot:



Result:

In this project we were looking for affecting factors on prevalence of covid-19 and its mortality. We examined following factors:

- Population
- Life expectancy
- Diabetes prevalence
- Vaccination

The consequences are as follows:

About population we can say with certainty that more populous countries have more patients and more mortality either.

After studying the trend of the disease in the United States we discovered the undeniable effect of vaccination on decreasing the mortality rate. Of course, we should mention that being vaccinated does not ensure that people won’t get infected by the virus.

About two other factors we cannot comment on with certainty. About life expectancy a little effect is observable and countries with upper rate of life expectancy has upper rate of mortality, and it is maybe because these countries have elder people more than other countries.