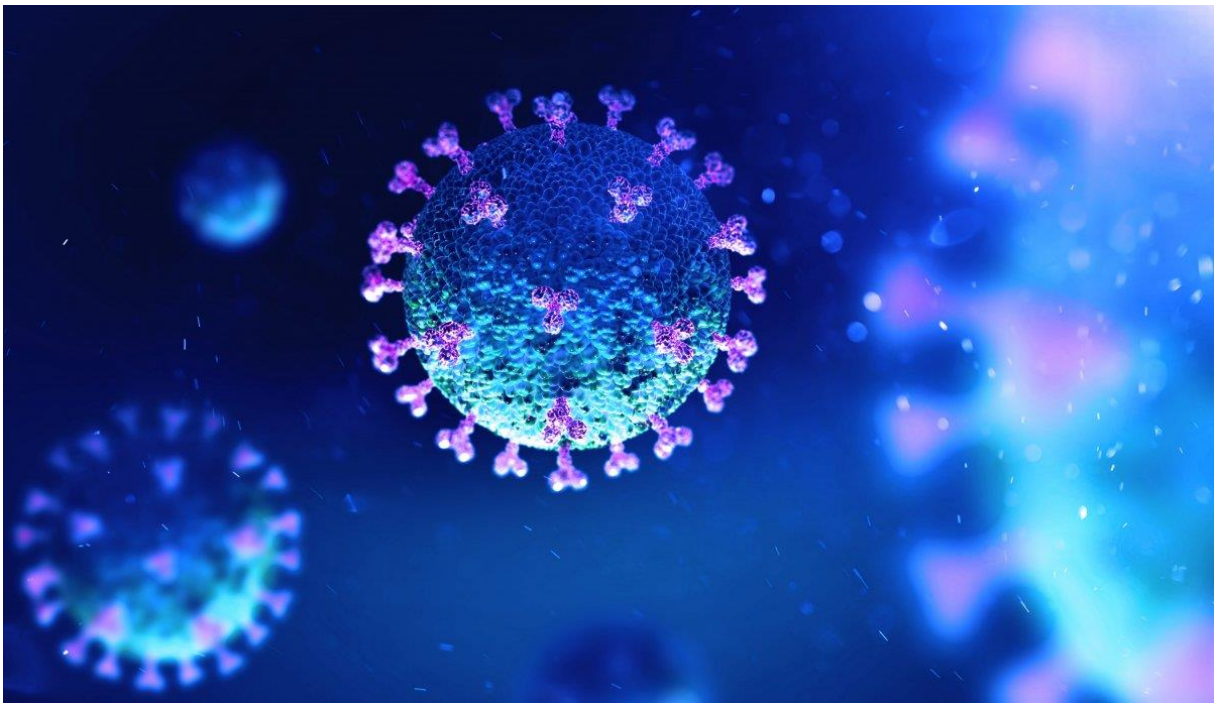
U.S. COVID-19 –
DATA ANALYSIS REPORT



Author: Balint Pataki

Engineer-Economist, Data Scientist

May 4th, 2020

1 Table of Contents

2 Study Scope

The aim of this study is to have a detailed insight of COVID-19 pandemic within the U.S. based on cases and deaths per day by state and county.

Questions to be answered:

1. Which regions are the most effected by the virus?
2. Which regions have the most deaths due to the virus?
3. Can we clearly separate and group the states based on COVID-19 effect? This can be important and during the planning of restarting the economy – we might consider to maintain more severe rules in some states than in others.
4. Can we identify states where the virus causes the most deaths per cases? What can be the root cause of this result?

3 Data

3.1 Sources

I. COVID-19 dataset on GitHub:

https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061.0	1	0
1	2020-01-22	Snohomish	Washington	53061.0	1	0
2	2020-01-23	Snohomish	Washington	53061.0	1	0
3	2020-01-24	Cook	Illinois	17031.0	1	0
4	2020-01-24	Snohomish	Washington	53061.0	1	0

Table 1: Covid-19 dataset

The dataset contains the cases and deaths related to COVID-19 per day by county/state.

II. Census data 2010 – population estimates for 2019:

https://www2.census.gov/programs-surveys/popest/tables/2010-2019/counties/totals/co-est2019-annres.xlsx

	Population Est. 2019	County	State
0	55869.0	Autauga County	Alabama
1	223234.0	Baldwin County	Alabama
2	24686.0	Barbour County	Alabama
3	22394.0	Bibb County	Alabama
4	57826.0	Blount County	Alabama

Table 2: Population dataset

III. State centroid locations (Wikipedia):

https://en.wikipedia.org/wiki/List_of_geographic_centers_of_the_United_States

	state	coordinates
1	Alabama	32°46′46″N 86°49′43″W / 32.7794°N 86.8287°W...
2	Alaska	64°04′07″N 152°16′42″W / 64.0685°N 152.2782°...
3	Arizona	34°16′28″N 111°39′37″W / 34.2744°N 111.6602°...
4	Arkansas	34°53′38″N 92°26′33″W / 34.8938°N 92.4426°W...
5	California	37°11′03″N 119°28′11″W / 37.1841°N 119.4696°...

Table 3: State centroid location dataset

IV. County boundaries (JSON file):

https://eric.clst.org/assets/wiki/uploads/Stuff/gz_2010_us_050_00_500k.json

3.2 Methods

I mainly used ***Pandas*** library during the whole project to create dataframes and to work with them. For visualization I applied ***Matplotlib*** and ***Folium*** libraries. I also used ***Scikit-learn*** to perform one of the most popular unsupervised Machine Learning techniques, K-Means Clustering.

During the study I used other libraries as well, such as *Numpy* and *BeautifulSoup*.

In order to read and work with CSV and XLSX files I applied the common Pandas functions *read_csv* and *read_excel*. In case of state centroid locations, I needed to apply ***web scraping*** techniques, while county boundaries were loaded in from ***JSON*** file.
In all the aforementioned data have been transformed into Pandas dataframe which is a powerful and useful tool to analyze data effectively in Python

4 Exploratory Data Analysis

In this section I analyzed the data from different viewpoints and I applied some advanced visualization tools like Matplotlib or Folium.

In order to perform a correct and useful analysis, some data wrangling has to be made, including the following calculations:

- sum of cases and deaths by states
- cases and deaths weighted by population where needed

First of all, I described the dataset with some basic statistical exploration.

	cases	deaths
count	109696.000000	109696.000000
mean	208.172805	9.395858
std	2410.264561	159.278032
min	0.000000	0.000000
25%	3.000000	0.000000
50%	10.000000	0.000000
75%	44.000000	1.000000
max	172364.000000	12895.000000

Table 1: Main statistical features of the dataset

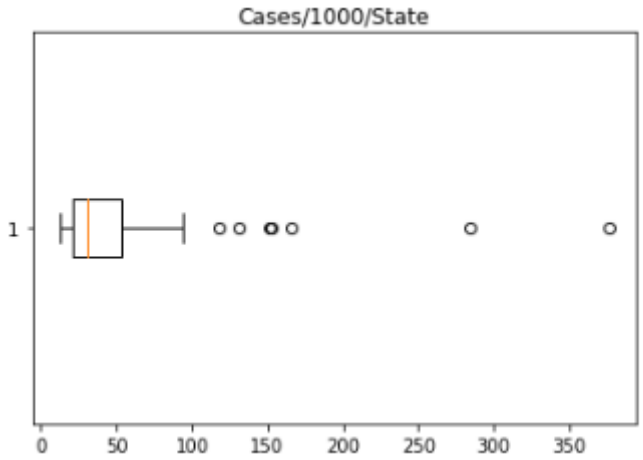


Figure 1: Box plot of the cases per 1,000 population by state

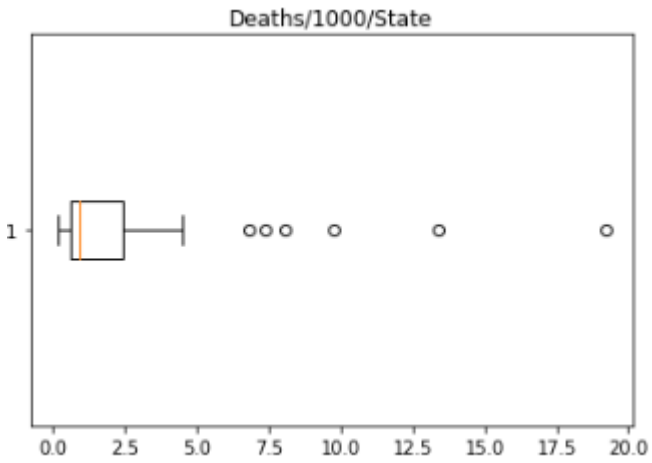


Figure 2: Box plot of the deaths per 1,000 population by state

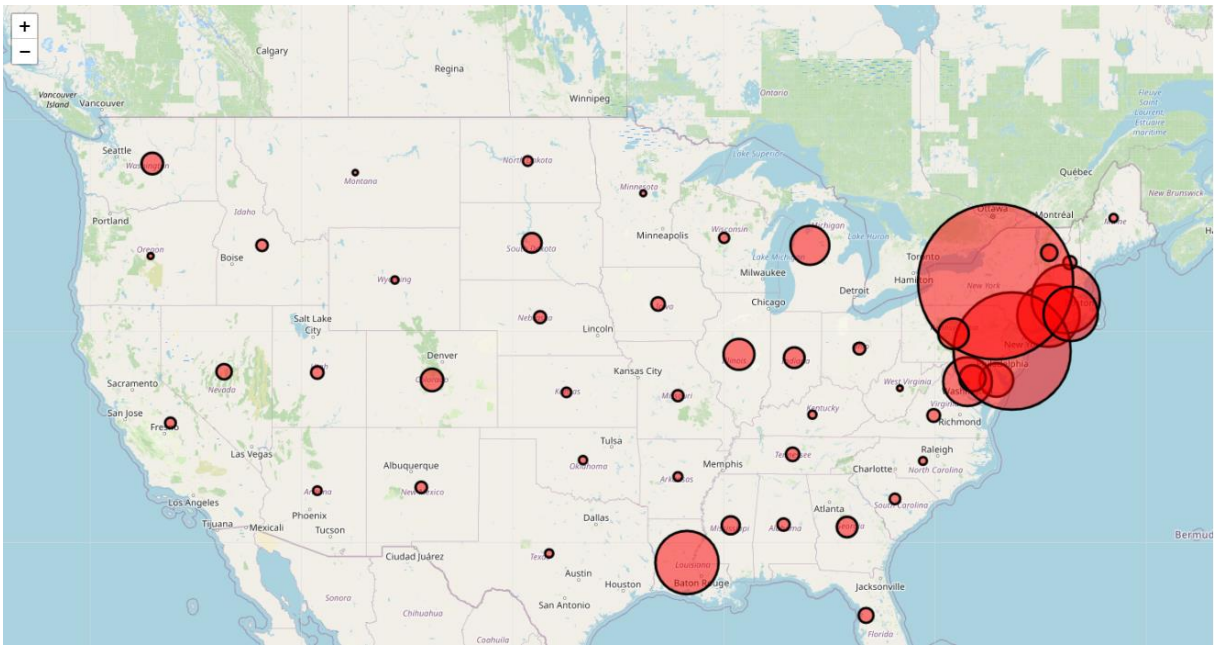


Figure 3: Cases/1,000 population by state on a map

This map shows exactly, where are the hot points, which states are the most effected by COVID-19. Please note that the bigger the circle the higher is the rate of cases per 1.000 people in the given state.

We can have a closer view on the East Coast states:

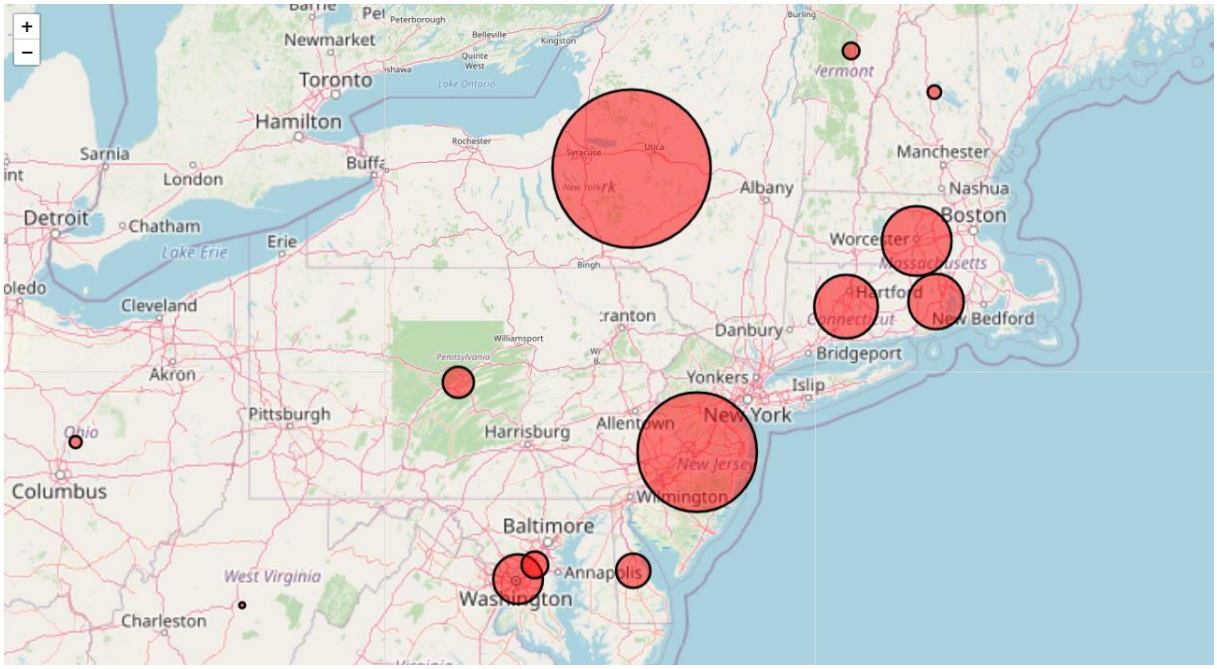


Figure 4: Cases/1,000 population by state on a map – East Coast of U.S.

To show the contrast between states, we can take a look on the below choropleth map:

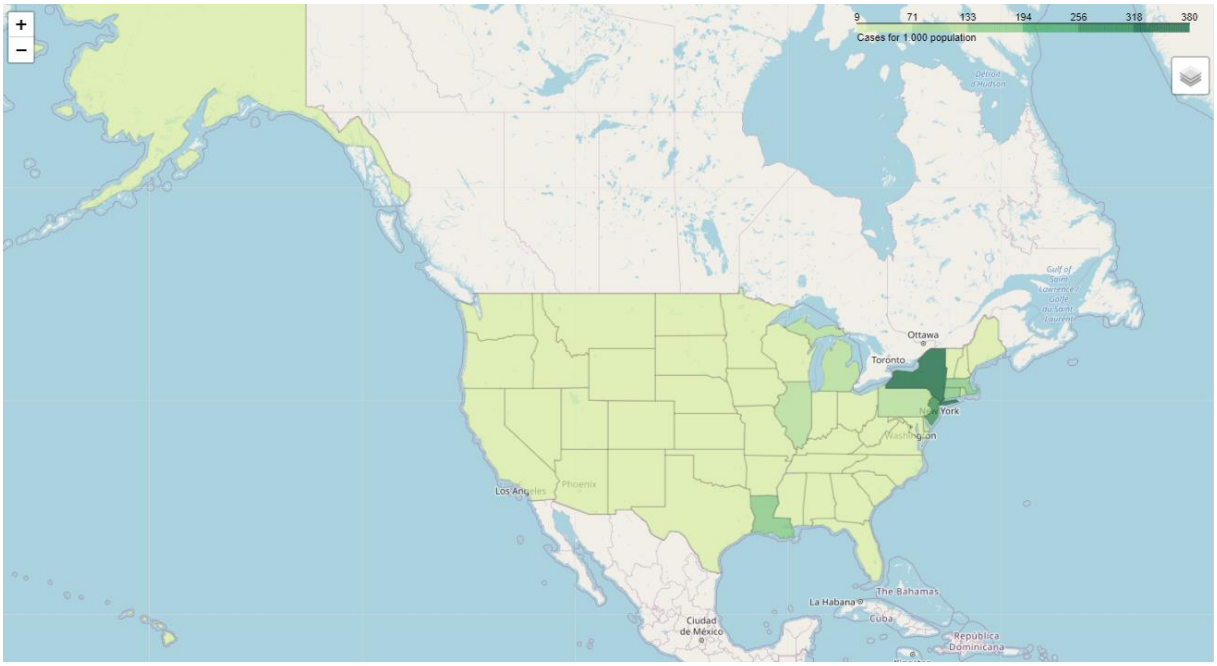


Figure 5: Choropleth map - Cases/1,000 population by state

Now we can see that there is a difference between the states regarding the cases, but we don't really know the dimension of this difference.

Thus, bar plots can be applied in order to see further details of the data.

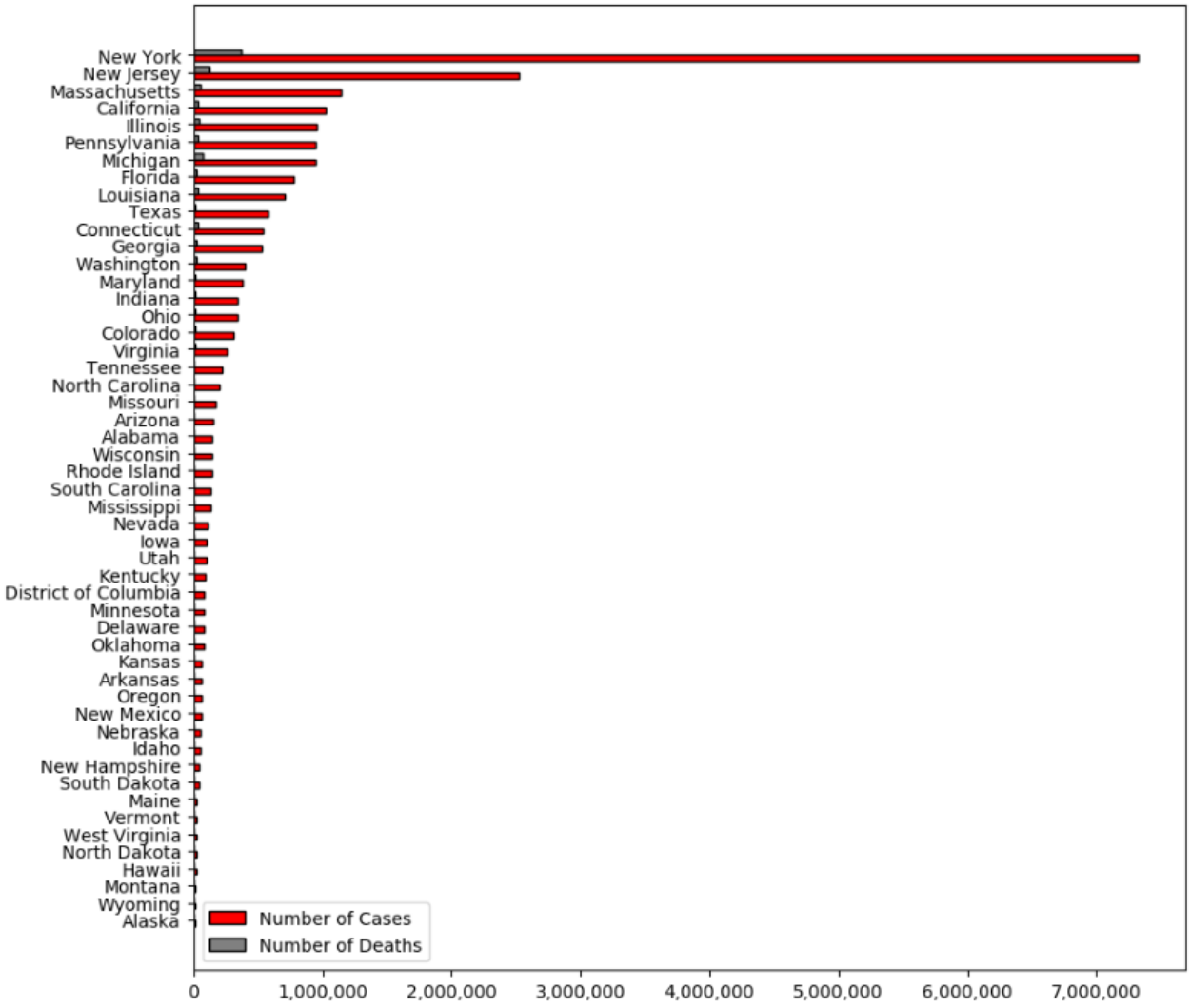


Figure 6: Total number of cases and deaths by state (bar plot)

Here we can see that in New York there are more than 7,000,000 cases but the second most effected state (New Jersey) has less than half of it.

The number of deaths shows a very similar picture:

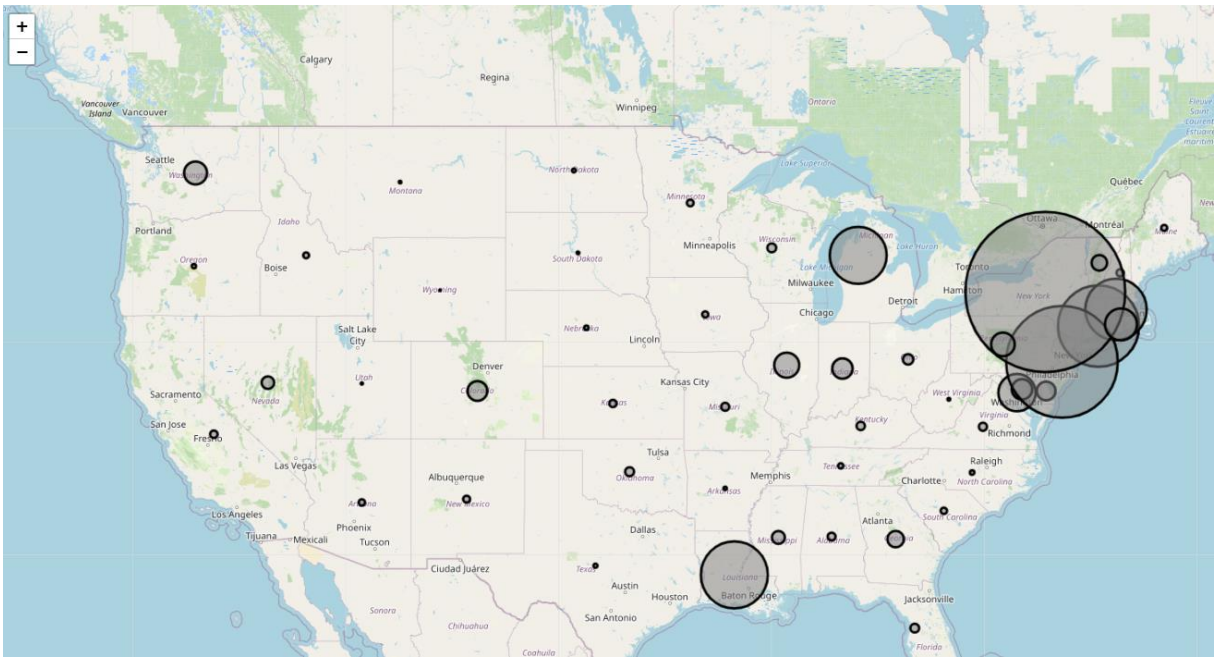


Figure 7: Deaths/1,000 population by state

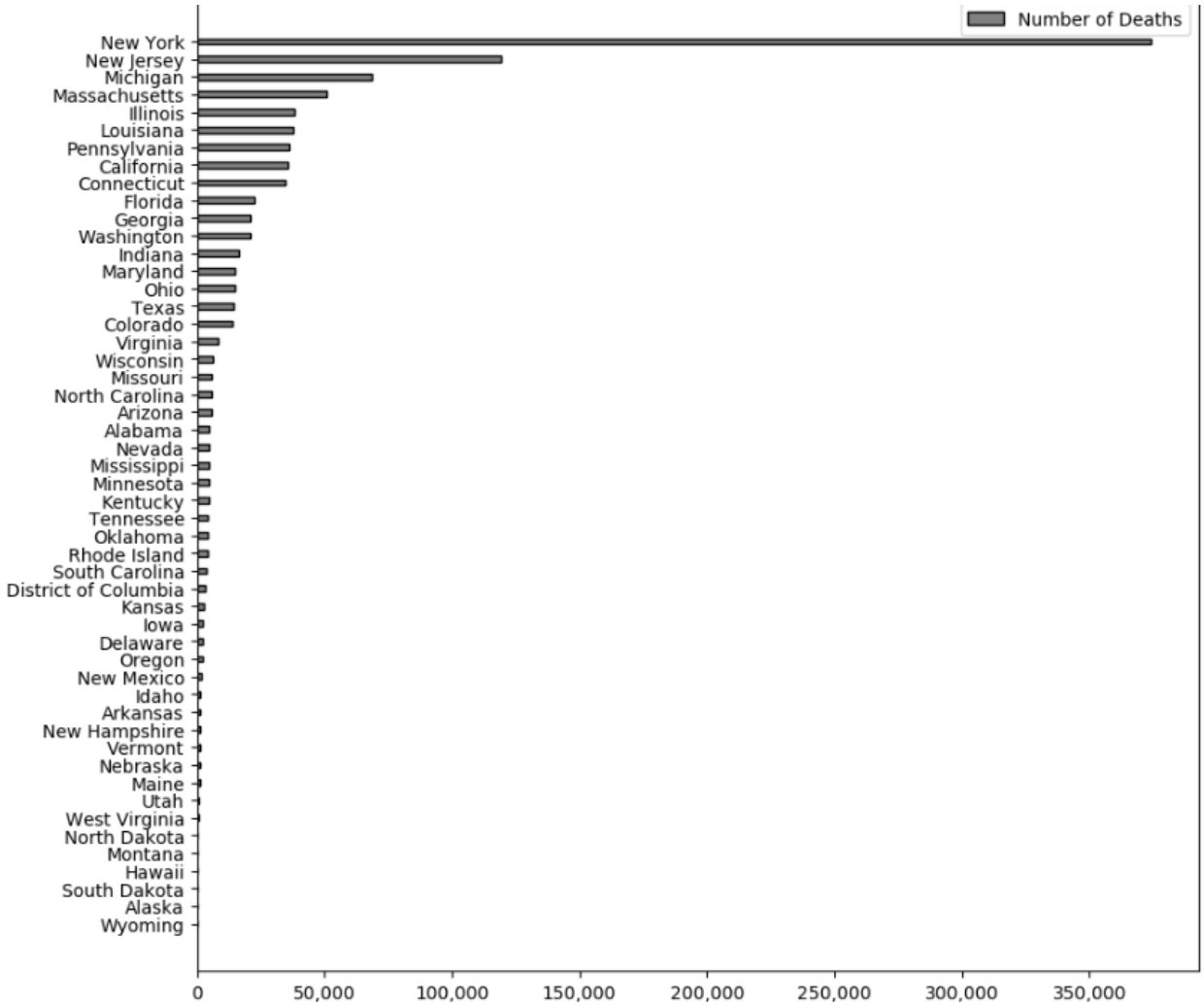


Figure 8: Total number of deaths by state (bar plot)

We can see which states are most effected by the virus and the numbers are clear. Both total numbers of cases and deaths and relative numbers per 1.000 populations shows that New York is the leader of the stats.

However, we might want to know where is the virus the deadliest – that can be calculated by dividing total number of cases by the total number of deaths. This metric can be seen below:

This is an interesting plot as it shows Michigan, Minnesota and Connecticut as leaders of the stats.

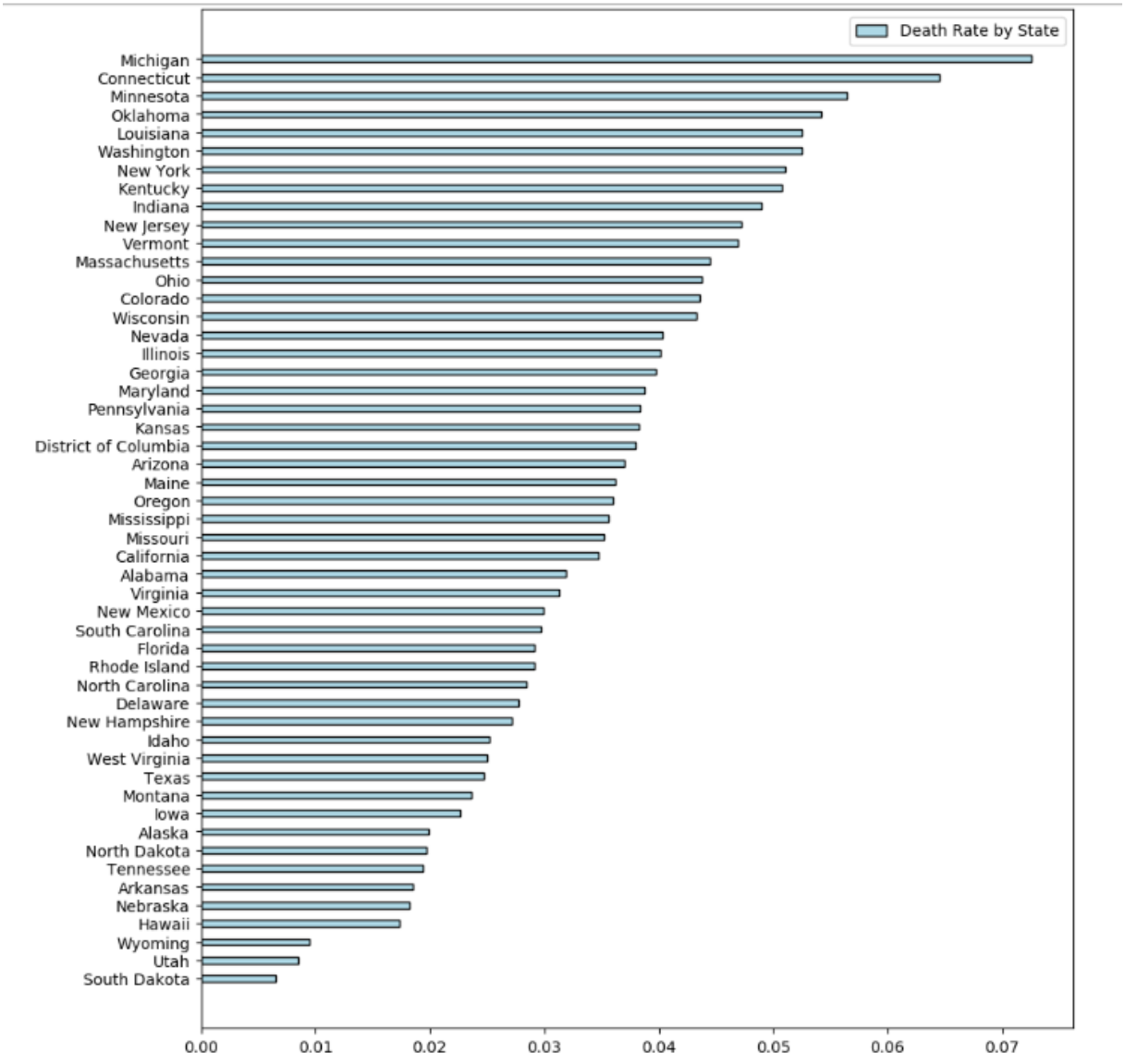


Figure 9: Death rate by state

5 Clustering

In order to divide the states into two groups I applied K-Means unsupervised machine learning algorithm.

This grouping can be useful if we should distinguish between states where the lockdown can be dissolved and the states where it should be maintained.

In order to decide how many groups should be applied during clustering, the Elbow Method has been applied:

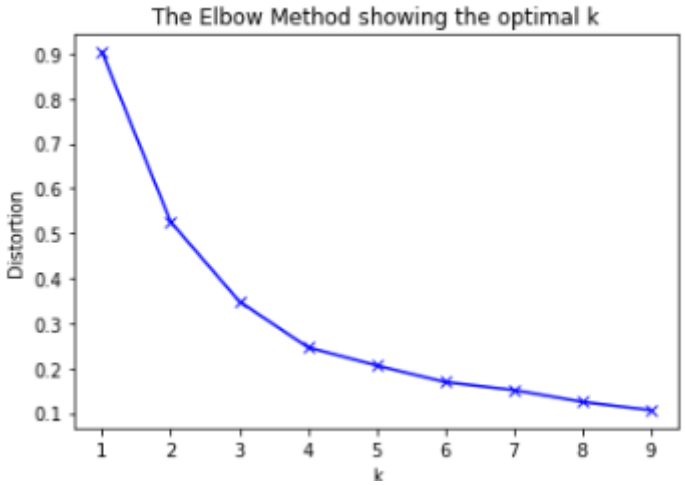


Figure 10: Elbow Method applied for K-Means algorithm

This graph can show us the optimal "k" which can be applied for K-Means algorithm – "k" indicates the clusters in which the data should be divided.

We can visualize the two clusters (clustering has been performed based on Cases/1,000 and Deaths/1,000 population):

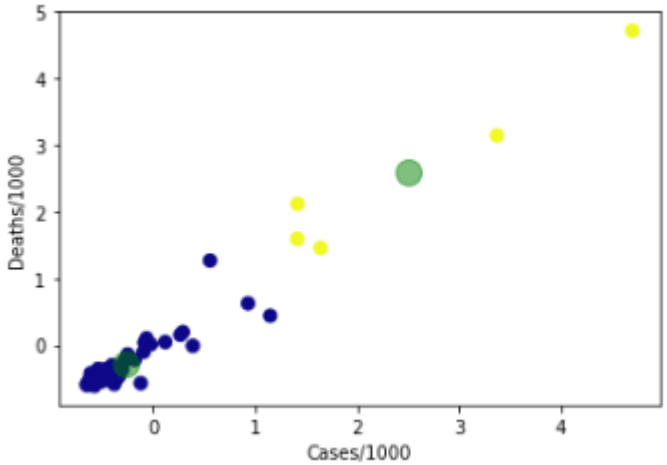


Figure 11: Visualized clusters based on cases and deaths per 1,000 people

6 Results and Discussion

Cases and deaths have been analyzed from different viewpoints: total numbers, numbers weighted by population have been measured. In the other hand a clustering has been performed, based on this result we can say that the following states should be considered as high risk regions and could be managed differently than others:

	Cluster Labels	state	cases	deaths	Population Est. 2019	Pop/1000	Cases/1000	Deaths/1000	latitude	longitude	deathRate
21	1	Massachusetts	1210886	54859	6892503.0	6892.503	175.681607	7.959228	42.2596	-71.8083	0.045305
30	1	New Jersey	2650357	127087	8882190.0	8882.190	298.390037	14.308070	40.1907	-74.6728	0.047951
32	1	New York	7637746	393240	19453561.0	19453.561	392.614288	20.214294	42.9538	-75.5268	0.051486
18	1	Louisiana	741267	39394	4648794.0	4648.794	159.453613	8.474026	31.0689	-91.9968	0.053144
6	1	Connecticut	569160	37255	3565287.0	3565.287	159.639322	10.449369	41.6219	-72.7273	0.065456

Table 5: Elements of Cluster 1 – the most effected states

Thus, the most effected states are **Massachusetts, New Jersey, New York, Louisiana and Connecticut**.

On the other part we have to take a look at **Michigan**, **Minnesota** and **Oklahoma** because of the high death/case ratios in these states.

If we decide to investigate further studies, age group distributions and population density should be considered.