Reem Almijmaj

ID: 3217747723

Principles of Programming for Data Science

December 3rd, 2021

Project- HW5

Introduction and Motivation:

Nowadays, our world is facing a pandemic that is called coronavirus, COVD-19. It is a viral disease that affects many countries and kills millions of people worldwide. January 20, 2020, CDC confirms the first U.S. laboratory-confirmed case of COVID-19 in the U.S. [1] Because this pandemic is affecting the U.S. hardly, and I am currently living in the U.S. I am interested in calculating the total number of Covid-19 death cases for each state and comparing the result between the states of the U.S. to find the top 5 states with the highest death cases rates. My aim is to see which states were *hit* the *hardest* by covid-19 and which *state* outbreaks are under control. In addition, I will compute the death cases density for each state to know the states that have a high death density, so the government can provide more support to these states that are affected the most by the pandemic and help them to overcome this struggle. Furthermore, I will compute the case fatality rate in the U.S. The aim is to measure the severity among detected cases. Not only that but also comparing the death cases number for the whole U.S. before and after the invention of the covid-19 vaccine. The objective is to see if the vaccine helps in reducing the number of death cases in the U.S.

Data Resources:

Extracted data from three different datasets, two is an API and one is a web-scraping. All these resources are freely available.

- The first dataset is from this website [3]. I web scraping this website to take the name of the US states and the area size of each state. I extract the data from the html. It is an open data source because we can access the website by searching on the web. It has the name of the US states only 50 states and its area size in square miles.

```
[<tr>
<td> </td>
<td><strong>State</strong></td>
<td><strong>Square Miles (Land Area)</strong></td>
</tr>, <tr>
<td>1</td>
<td>Alaska</td>
<td>570,641</td>
</tr>, <tr>
<td>2</td>
<td>Texas</td>
<td>261,914</td>
</tr>, <tr>
<td>3</td>
<td>California</td>
<td>155,973</td>
</tr>, <tr>
<td>4</td>
<td>Montana</td>
```

A sample of the original dataset.

	StateNam	SquareM
2	AK	570,641
3	TX	261,914
4	CA	155,973
5	MT	145,556
6	NM	121,365
7	AZ	113,642
8	NV	109,806
9	CO	103,730
10	WY	97,105
11	OR	96,003
12	ID	82,751
13	UT	82,168
14	KS	81,823
15	MN	79,617

A sample after I cleaned the data

- Second dataset is an API, it is free and open source. I got it from covidtracking.com. And the Jason file [4] is open source as well. It provides information about US individual states for Covid- tracking. The API consists of columns such as date, state name, number of positive cases and death cases and others.

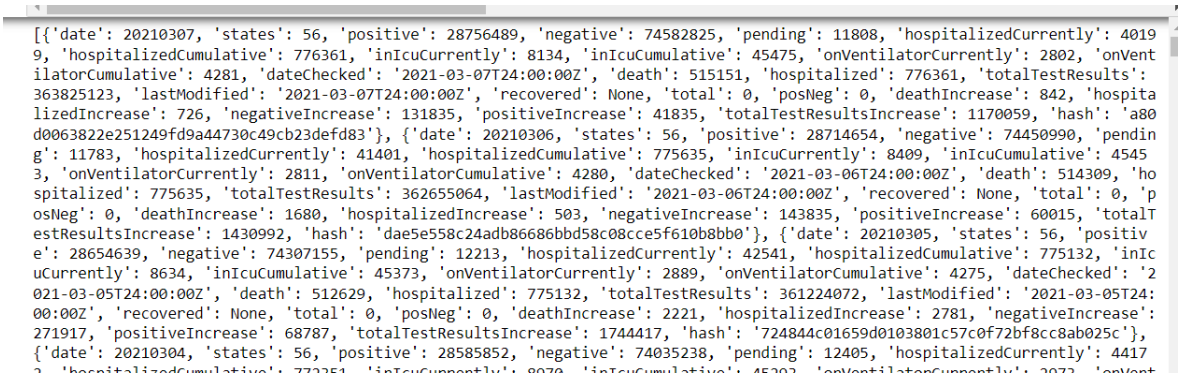

```
[{"date":20210307,"state":"AK","positive":56886,"probableCases":null,"negative":null,"pending":null,"totalTestResultsSource":"totalTestsViral","totalTestResults":1731628,"hospitalizedCurrently":33,"hospitalizedCumulative":1293,"inIcuCurrently":null,"inIcuCumulative":null,"onVentilatorCurrently":2,"onVentilatorCumulative":null,"recovered":null,"lastUpdateEt":"3/5/2021 03:59","dateModified":"2021-03-05T03:59:00Z","checkTimeEt":"03/04 22:59","death":305,"hospitalized":1293,"hospitalizedDischarged":null,"dateChecked":"2021-03-05T03:59:00Z","totalTestsViral":1731628,"positiveTestsViral":68693,"negativeTestsViral":1660758,"positiveCasesViral":null,"deathConfirmed":null,"deathProbable":null,"totalTestEncountersViral":null,"totalTestsPeopleViral":null,"totalTestsAntibody":null,"positiveTestsAntibody":null,"negativeTestsAntibody":null,"totalTestsPeopleAntibody":null,"positiveTestsPeopleAntibody":null,"negativeTestsPeopleAntibody":null,"totalTestsPeopleAntigen":null,"positiveTestsPeopleAntigen":null,"totalTestsAntigen":null,"positiveTestsAntigen":null,"fips":"02","positiveIncrease":0,"negativeIncrease":0,"total":56886,"totalTestResultsIncrease":0,"posNeg":56886,"dataQualityGrade":null,"deathIncrease":0,"hospitalizedIncrease":0,"hash":"dc4bccd4bb885349d7e94d6fed058e285d4be164","commercialScore":0,"negativeRegularScore":0,"negativeScore":0,"positiveScore":0,"score":0,"grade":""},
{"date":20210307,"state":"AL","positive":499819,"probableCases":107742,"negative":1931711,"pending":null,"totalTestResultsSource":"totalTestsPeopleViral","totalTestResults":2323788,"hospitalizedCurrently":494,"hospitalizedCumulative":45976,"inIcuCurrently":null,"inIcuCumulative":2676,"onVentilatorCurrently":null,"onVentilatorCumulative":1515,"recovered":295690,"lastUpdateEt":"3/7/2021 11:00","dateModified":"2021-03-07T11:00:00Z","checkTimeEt":"03/07 06:00","death":10148,"hospitalized":45976,"hospitalizedDischarged":null,"dateChecked":"2021-03-07T11:00:00Z","totalTestsViral":null,"positiveTestsViral":null,"negativeTestsViral":null,"positiveCasesViral":392077,"deathConfirmed":7963,"deathProbable":2185,"totalTestEncountersViral":null,"totalTestsPeopleViral":2323788,"totalTestsAntibody":null,"positiveTestsAntibody":null,"negativeTestsAntibody":null,"totalTestsPeopleAntibody":119757,"positiveTestsPeopleAntibody":null,"negativeTestsPeopleAntibody":null,"totalTestsPeopleAntigen":null,"positiveTestsPeopleAntigen":null,"totalTestsAntigen":null,"positiveTestsAntigen":null,"fips":"01","positiveIncrease":408,"negativeIncrease":2087,"total":2431530,"totalTestResultsIncrease":2347,"posNeg":2431530,"dataQualityGrade":null,"deathIncrease":-1,"hospitalizedIncrease":0,"hash":"997207b430824ea40b8eb8506c19a93e07bc972e","commercialScore":0,"negativeRegularScore":0,"negativeScore":0,"posi
```


This is a sample data.

date	state	deathIncrease
3/7/2021	AK	0
3/7/2021	AL	0
3/7/2021	AR	22
3/7/2021	AS	0
3/7/2021	AZ	5
3/7/2021	CA	258
3/7/2021	CO	3
3/7/2021	CT	0
3/7/2021	DC	0
3/7/2021	DE	9
3/7/2021	FL	66
3/7/2021	GA	1
3/7/2021	GU	0
3/7/2021	HI	1

This is a sample after I clean the data and take the related columns, I am interested in.

- Third is an API for the US covid tracking, it is free and open source. I got it from covidtracking.com and the Jason file [5] is open access as well. It provides information about US individual states for Covid- tracking. The API consists of columns such as date, state name, number of positive cases and death cases and others.



```
[{'date': 20210307, 'states': 56, 'positive': 28756489, 'negative': 74582825, 'pending': 11808, 'hospitalizedCurrently': 4019
9, 'hospitalizedCumulative': 776361, 'inIcuCurrently': 8134, 'inIcuCumulative': 45475, 'onVentilatorCurrently': 2802, 'onVent
ilatorCumulative': 4281, 'dateChecked': '2021-03-07T24:00:00Z', 'death': 515151, 'hospitalized': 776361, 'totalTestResults':
363825123, 'lastModified': '2021-03-07T24:00:00Z', 'recovered': None, 'total': 0, 'posNeg': 0, 'deathIncrease': 842, 'hospita
lizedIncrease': 726, 'negativeIncrease': 131835, 'positiveIncrease': 41835, 'totalTestResultsIncrease': 1170059, 'hash': 'a80
d0063822e251249fd9a44730c49cb23defd83'}, {'date': 20210306, 'states': 56, 'positive': 28714654, 'negative': 74450990, 'pendin
g': 11783, 'hospitalizedCurrently': 41401, 'hospitalizedCumulative': 775635, 'inIcuCurrently': 8409, 'inIcuCumulative': 4545
3, 'onVentilatorCurrently': 2811, 'onVentilatorCumulative': 4280, 'dateChecked': '2021-03-06T24:00:00Z', 'death': 514309, 'ho
spitalized': 775635, 'totalTestResults': 362655064, 'lastModified': '2021-03-06T24:00:00Z', 'recovered': None, 'total': 0, 'p
osNeg': 0, 'deathIncrease': 1680, 'hospitalizedIncrease': 503, 'negativeIncrease': 143835, 'positiveIncrease': 60015, 'totalT
estResultsIncrease': 1430992, 'hash': 'dae5e558c24adb86686bbd58c08cce5f610b8bb0'}, {'date': 20210305, 'states': 56, 'positiv
e': 28654639, 'negative': 74307155, 'pending': 12213, 'hospitalizedCurrently': 42541, 'hospitalizedCumulative': 775132, 'inIc
uCurrently': 8634, 'inIcuCumulative': 45373, 'onVentilatorCurrently': 2889, 'onVentilatorCumulative': 4275, 'dateChecked': '2
021-03-05T24:00:00Z', 'death': 512629, 'hospitalized': 775132, 'totalTestResults': 361224072, 'lastModified': '2021-03-05T24:
00:00Z', 'recovered': None, 'total': 0, 'posNeg': 0, 'deathIncrease': 2221, 'hospitalizedIncrease': 2781, 'negativeIncrease':
271917, 'positiveIncrease': 68787, 'totalTestResultsIncrease': 1744417, 'hash': '724844c01659d0103801c57c0f72bf8cc8ab025c'},
{'date': 20210304, 'states': 56, 'positive': 28585852, 'negative': 74035238, 'pending': 12405, 'hospitalizedCurrently': 4417
```


Sample of the original data.

	date	positiveIncrease	deathIncrease
0	2021-03-07	41835	842
1	2021-03-06	60015	1680
2	2021-03-05	68787	2221
3	2021-03-04	65487	1743
4	2021-03-03	66836	2449
5	2021-03-02	54248	1728
6	2021-03-01	48092	1241

Sample after I clean the data.

Analysis:

- Finding CFR [case fatality rate] in the USA

CFR = total number of deaths from covid-19 / total number of infected individuals' total number of infected individuals with covid-19 in the USA

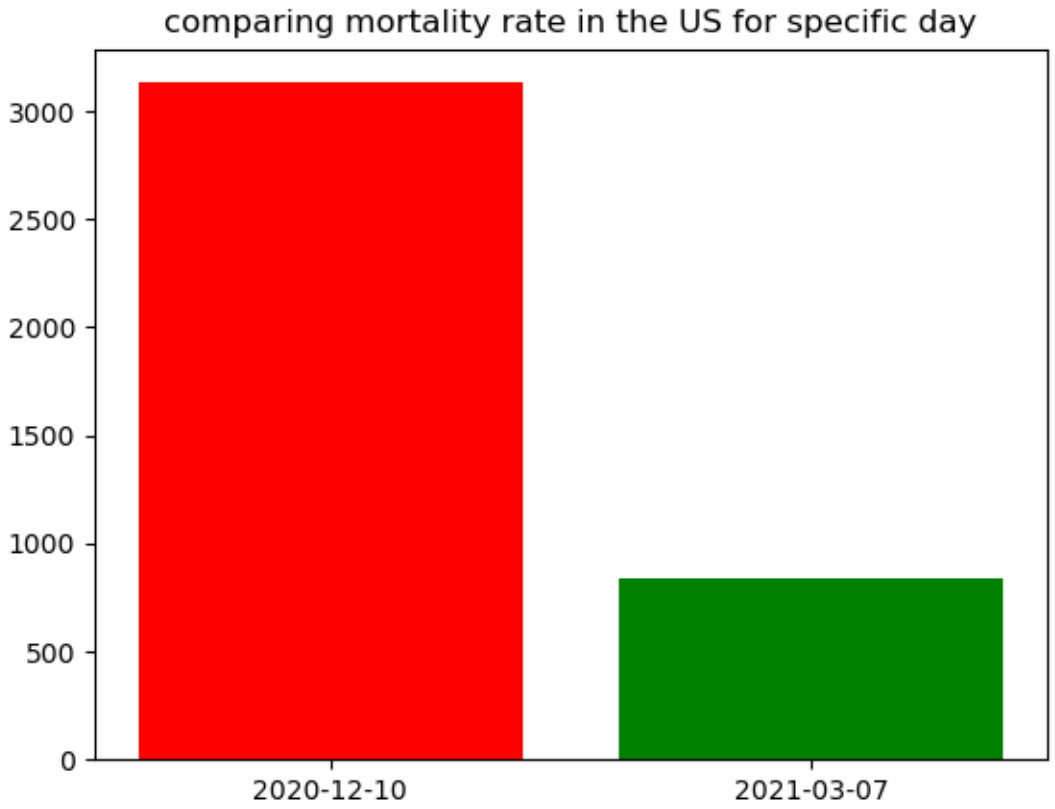
The result of calculating the case fatality rate for the COVID-19 disease in the USA is CFR * 100 which is:

```
1.7914252327535536
```

This helps to understand how to estimate the fatality rate and know if covid is really causing a lot of death.

- Compare mortality rate in the US for specific day in 2021 with a specific day in 2020 where vaccine was not invented yet.

Last day I have on my dataset is 2021-03-07 so I used it as a day where the rate of vaccine is high compare to the 2020-12-10 where the covid-19 vaccine was not used.

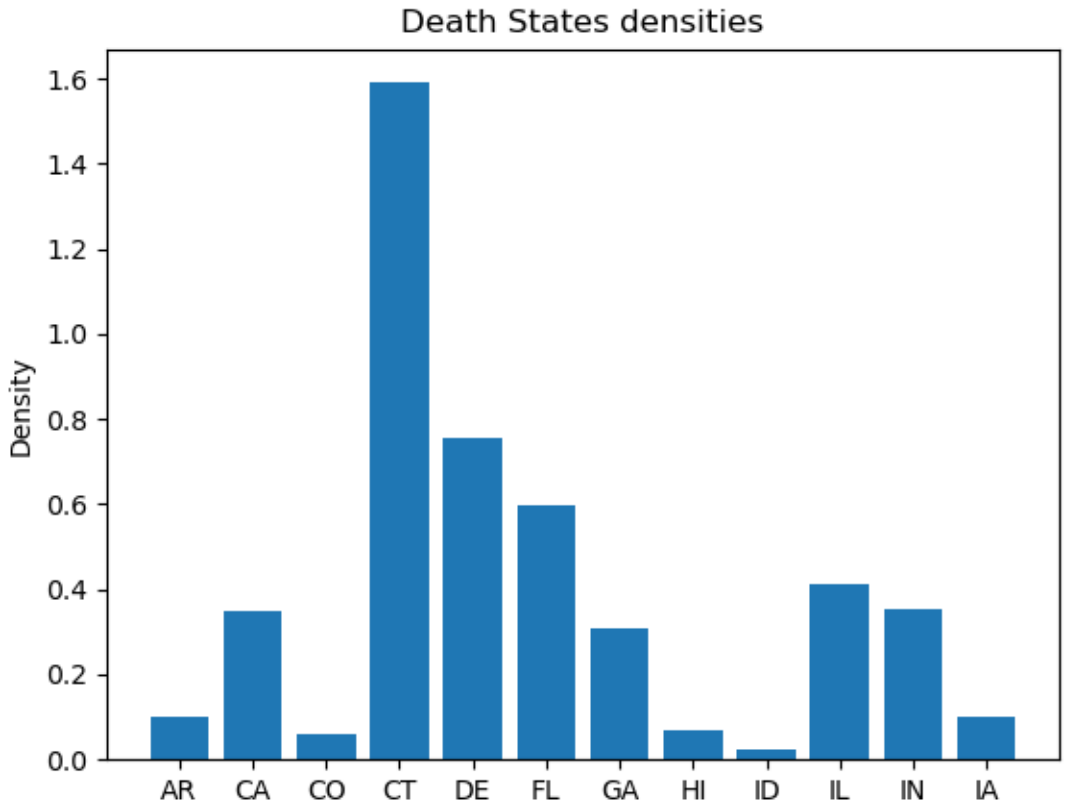


The result of comparing mortality rate in the US for a specific day in 2021 after a mass of covid-19 vaccination processes with a specific day in 2020 where the covid-19 vaccine was not invented

As you can see from the figure above that the number of cases in 2020-12-10 is dramatically high, so my conclusion is that vaccine did help to decrease the number of death cases because in 2021-03-07 less than one thousand deaths.

- Calculate the Death cases density of a state

Death cases density of a state = total death cases of a state / total area size in square miles of a state.



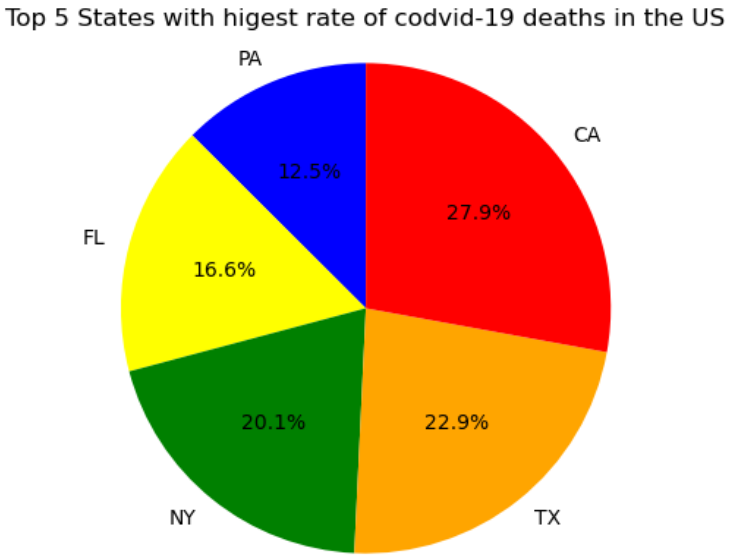
The result of calculating the death cases density of COVID-19 disease for each state in the US.

As you can see from the sample above (sample of Covid-19 death cases densities of 12 states of US). The Connecticut has 1.5 death density rate but if I compare the whole data set, NJ is the state that has the highest death density rate. I calculated this for all the states, but I randomly chose 12 states to represent. This is one way where people can understand where outbreaks are the strongest.

	DeathDesnsity
AL	0.200296
AK	0.000534
AZ	0.143758
AR	0.105559
CA	0.347009
CO	0.057949
CT	1.590093
DE	0.753964
FL	0.597552
GA	0.309156
HI	0.069594
ID	0.022767
IL	0.413973
IN	0.355088
IA	0.099472
KS	0.058871
KY	0.121288
LA	0.223752
ME	0.022939
MD	0.813811
MA	2.100281
MI	0.294735
MN	0.082570
MS	0.145117
MO	0.118523
MT	0.009508
NE	0.028734
NV	0.045872
NH	0.132010
NJ	3.177787
NM	0.031376
NY	0.826465
NC	0.236093
ND	0.021480
OH	0.434962

```
OK      0.066017
OR      0.023916
PA      0.547747
RI      2.473888
SC      0.290724
SD      0.025034
TN      0.280131
TX      0.169716
UT      0.024048
VT      0.022489
VA      0.242538
WA      0.079196
WV      0.096525
WI      0.130869
WY      0.007023
```

- Finding top 5 States with highest rate of Covid-19 deaths in the US



```
[('PA', 24349), ('FL', 32266), ('NY', 39029), ('TX', 44451), ('CA', 54124)]
```

Top 5 States with highest rate of Covid-19 deaths in the US. The aim is to see the state that hit hardly by covid-19. It might help the government to put more efforts to reduce the impact of COVID-19 on these states. Also, could help people to know which state is highly contagious so they protect themselves.

Extensibility and Maintainable:

Extensibility:

- Working with real-time data because the data I worked on within this project is static (already stored data) and not up-to-date.
- Using sophisticated statistical tools to get a meaningful insight out of the data because I used simple calculations because the more advanced techniques were used, the more accurate the result became.
- Improve the graphs, and choose better plotting libraries, also, create a better user interface that could help the user to see the result clearly and accurately.

Maintainability:

- Scraping the data could take sometime around 2 minutes to run in default mode, and the time might vary from computer to computer. If this happens, I highly suggest running the code in a static mode.
- If you did not download the right packages/libraries, the code might not work, and you might get unexpected results.
- There is no issue regarding the API keys, as I can access the data directly without issuing a key and without any limitation for daily access.

Conclusion & Future Work:

COVID-19 is an infectious disease that affect millions of people worldwide. Infections have been growing rapidly and tremendous efforts are being made to fight the disease. This report tries to overcome this challenge by finding top 5 States with highest rate of Covid-19 deaths in the US, finding CFR, case fatality rate, in the USA, compare mortality rate in the US for specific day in 2021 with a specific day in 2020 where vaccine was not invented yet and calculating the death cases density of a state. The calculations I have made can be used to understand the growth, spread of COVID-19 and which state is hit the most by Covid-19. I hope that my work provides useful knowledge to the field for researchers interested in this area. For the future work, I suggest work on real-time data for COVID-19, real time analysis. Also, using the data for prediction to draw a real insight from the data not only calculating the cases. Because if it is used for prediction the county will be more prepared. Furthermore, I chose some type of plots that

might not be perfect for visualizing the data, so I would say in the future the plotting and graphs could be enhanced.

References:

[1] https://www.cdc.gov/museum/timeline/covid19.html

[2] https://www.fda.gov/news-events/press-announcements/fda-approves-first-covid-19-vaccine

[3] https://statesymbolsusa.org/symbol-official-item/national-us/uncategorized/states-size

[4] https://api.covidtracking.com/v1/states/daily.json

[5] https://api.covidtracking.com/v1/us/daily.json