Aaron Tsui
31346171

Analyzing COVID-19 data to make decisions for the Spring 2021 semester

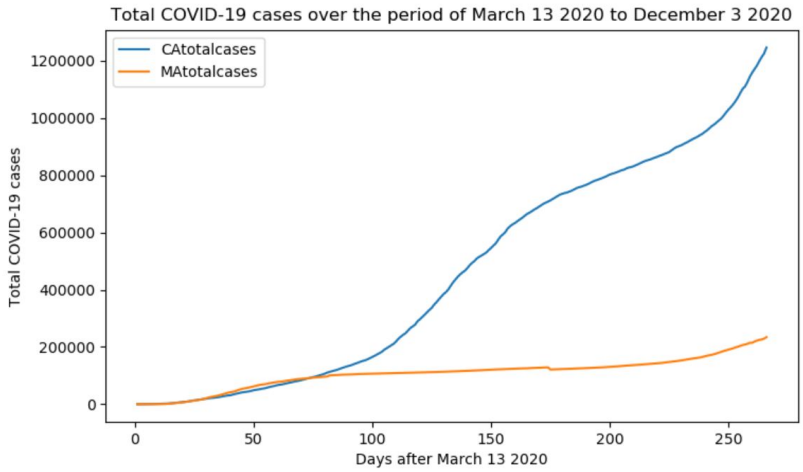
I obtained this dataset from https://covid19-projections.com/. I changed the dataset multiple ways to make it more accurate to current estimates. I eliminated most of the dates where total cases for one state were not recorded and also extrapolated the elimination of those dates to the other state's data. This gave me a consistent dataset from March 13th, 2020, to December 3rd, 2020.

Team Members: Aaron Tsui

Main Question:
Should I go back to UMass for the Spring? Is the rate of COVID-19 cases in Massachusetts increasing more quickly than it is here at home in California?

Subquestion: Is the number of total cases accelerating faster in Massachusetts than it is in California?

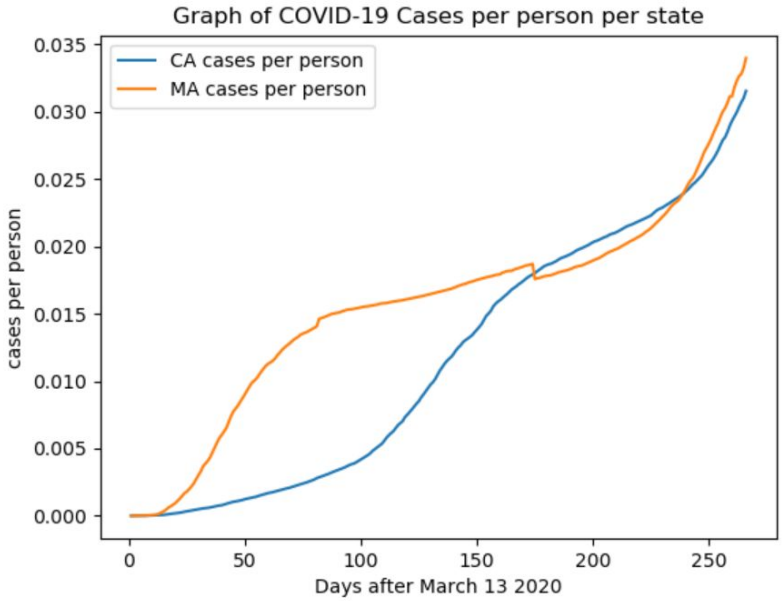
```
df.plot(x='date',y=['CAtotalcases', 'MAtotalcases'])
# tried to run with df['date'] and df['CAtotalcases'], didn't work. debugged for 3 hours, this worked
plt.ylabel('Total COVID-19 cases')
plt.xlabel('Days after March 13 2020')
plt.title('Total COVID-19 cases over the period of March 13 2020 to December 3 2020')
plt.show()
```



- After making this graph I realized that this is misleading because California has a far higher population than Massachusetts, making the graph faulty. So I added two columns to my CSV data: one that divided California Total Cases with California's population, and another that divided Massachusetts Total Cases with Massachusetts' population, and reran my code with changed y axis values.

```
df.plot(x='date',y=['CA cases per person', 'MA cases per person'])
```

```
plt.ylabel('cases per person')
plt.xlabel('Days after March 13 2020')
plt.title('Graph of COVID-19 Cases per person per state')
plt.show()
```

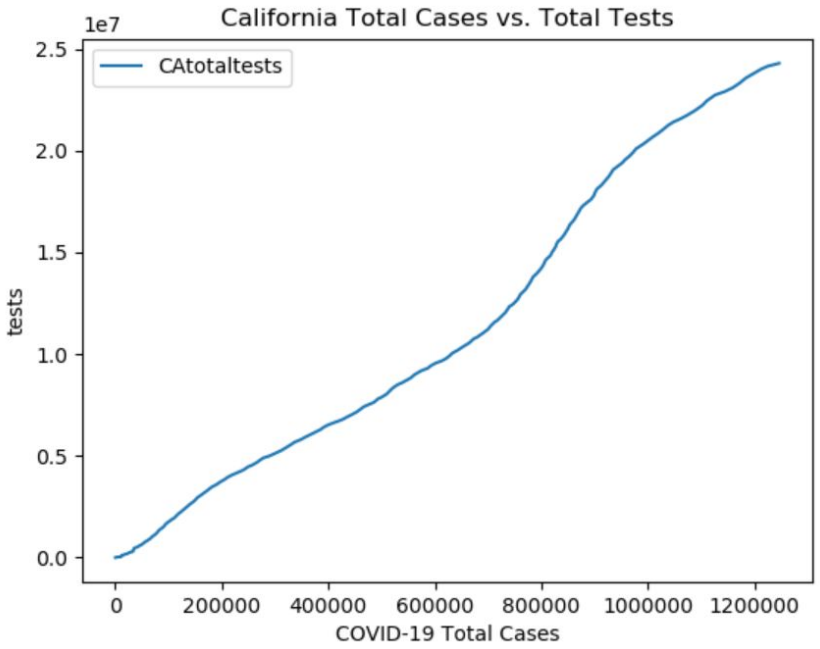


- This graph gives a much better picture of how COVID-19 cases have been increasing since March 13th, 2020 relative to each states' estimates.
- Massachusetts' cases per person increased sharply around May, and decreased in acceleration as it neared July. California's cases per person started off at a slow pace, but instead increased sharply in acceleration around the end of June. California's cases per person shortly increased past Massachusetts' numbers at the beginning of September, but dipped back down below at the start of November.
- It can be inferred from our graph that people traveling in and out of California for the Memorial Day holiday increased the speed at which COVID-19 cases were discovered, while many people in Massachusetts stayed home for Memorial Day.

Subquestion: Is the number of total COVID tests accelerating as fast as total COVID cases? Is it lower? Or is it higher?

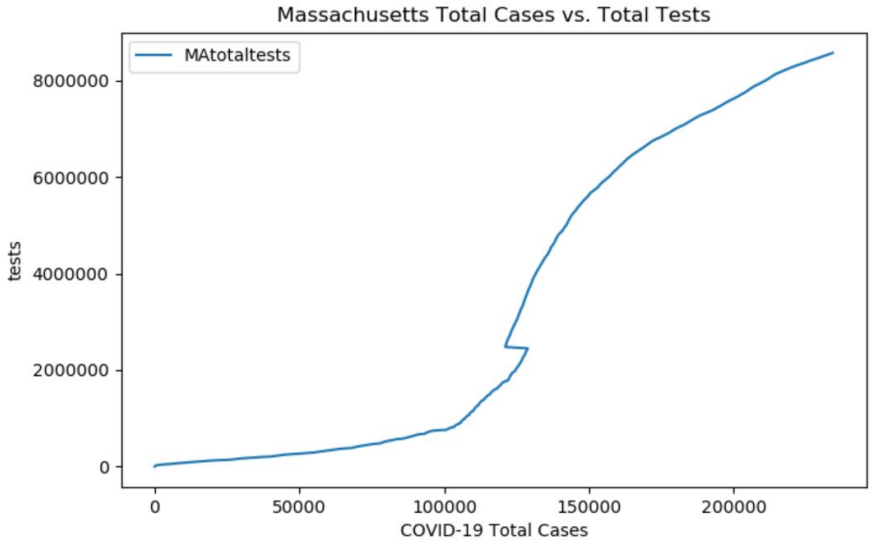
Why this question: This was fascinating me considering these past few months we've heard many things from politicians and political figures saying tests have been massively produced and distributed. This analysis was intended to show if we are going to 'beat COVID-19' anytime soon or do the distribution of COVID-19 tests need to be ramped up even further than it is now?

```
df.plot(x='CAtotalcases',y=['CAtotaltests'])
plt.ylabel('tests')
plt.xlabel('COVID-19 Total Cases')
plt.title('California Total Cases vs. Total Tests')
plt.show()
```



- What we can see on this graph is that Total COVID-19 Tests increased at a faster speed at around 70,000 Total COVID-19 cases, and that it has since reached a slower speed after we passed the 97,000 cases mark.
- We can infer that tests need to be produced and distributed at a faster pace than it currently is.

```
df.plot(x='MAtotalcases',y=['MAtotaltests'])
plt.ylabel('tests')
plt.xlabel('COVID-19 Total Cases')
plt.title('Massachusetts Total Cases vs. Total Tests')
plt.show()
```



- In Massachusetts, Total COVID-19 Tests largely started off low and only started increasing at a faster rate after 100,000 Total Cases. The “blip” in the data where tests are around 2.5 million

was most likely due to a large number of deaths from COVID-19, lowering the number of cases and increasing the number of deaths (total deaths and total cases are separated into two groups of data in this dataset).
- We can infer that tests need to be produced and distributed at a faster pace than it currently is.

- Overall, the slope of the Total Cases vs. Total Tests line for California is higher than it is for Massachusetts.

Conclusions:
My conclusion is that I should stay here at home in California rather than return to UMass for the Spring 2021 semester.

Originally, my plan was to **not** return for the spring semester due to costs, since it costs me less to attend UMass via online lectures at home than stay in UMass Housing doing the same thing. My results from this analysis of COVID-19 data only strengthens my decision to not return for the spring.

Variables that were not accounted for in the data include travel restrictions put in place by governing bodies as well as geographic location. It would be disingenuous to not recognize the physical distance between San Francisco, California and Reno, Nevada is far larger than the distance between Amherst and New York City.

Challenges:
I could not get my plotting to work for the first 6 hours I worked on this project; in the first 3 hours I was brainstorming ideas and questions to ask. I debugged and sought help through stackoverflow multiple times, a few solutions didn't work but one finally did. Setting delim_whitespace as True was one solution I found but only ended up causing even more problems. I also had some challenge in trying to come up with a question to explore and a question I could answer with the skills I've learned in this class. A question that I wasn't able to answer was "What is the relationship between Total Deaths by COVID-19 and Predicted Deaths to COVID-19?" Future exploration into COVID-19 could include that question.

How to run my code:
What I imported:
```
from matplotlib import pyplot as plt
import pandas as pd
```

Reading in the CSV data:
```
df = pd.read_csv(r"C:\Users\aaron\Desktop\CICS397AFP\CAMACOVID.csv", header=0)
```

The main parts of my code are after each question I posed to this data and before every graph.

My preprocessed dataset, CAMACOVID.csv, can be found here:
https://drive.google.com/file/d/1mwZ1auFkjyxnOsY3RrTrXZl1hjRN2FFu/view