Name: Thuc Duong
Project Title: The analysis of mitigation measures and its effects on death/cases/hospitalization rates.
Introduction:
+ I worked on the COVID-19 group project. I selected task 1, which is called "EDA, analysis, and visualization/dashboards of social distancing measures we collected for the states during weeks 1 and 2."
+ The motivation behind this project is to explore whether and how do mitigation measures affect death/cases/hospitalization rates.
Approach:
+ The main idea behind this project is to gather any news or general information online about COVID-19 mitigation measures and record it. This step was done in conjunction with many of my fellow students. We created a big data sheet that have the COVID-19 mitigation measures timeline for each US state and for some counties.
+ For first week task, I was involved in finding data for the states of California, New Mexico, South Carolina and Maine. The total hours that I spent on the first week was **7 hours**.
+ For the second week task, I was involved in merging my datasets with others that had the same state as me. The total hours that I spent was **4 hours**.
+ Then, I performed analysis on this dataset in conjunction with some of the other datasets that have statistics on deaths/cases/hospitalization rates to explore if there are any correlations between them.
+ Before any analysis was done, I performed quality checking and cleaned the data as I saw fit.
+ The pre-processing, data quality checking and cleaning of this dataset took me about **3 hours** because there were a lot of small errors in terms of duplicate values, empty values, misspelled mitigation types, misspelled state names, and misclassification of mitigation types.
+ I first checked for empty data cells on important columns such as states, date, and mitigation types.
+ Then I checked for any duplicates in those columns as well.
+ Then I checked for unique values in the mitigation types column and found that there were a lot of misspelled values that turned out to be duplicates of others.
+ I also found that there were some rows in which there were multiple classifications for mitigation types, so I had to dig through the raw description and the source to determine exactly one classification.
+ There were also rows in which there were literally questions being put down as mitigation type, so I had to check through the raw description and the source as well.
+ After I finished with pre-processing and cleaning the mitigation type data, I looked for additional datasets on the Internet for county-level and state-level data on deaths/cases/hospitalization rate or any other COVID-19 healthcare statistics-related details with breakdowns on the state and county-level.
+ This step took me about **1.5 hours** since there were a lot of different datasets with different kind and level of granularity out there.
+ Next, EDA was performed on the mitigation type dataset. The general EDA took about **0.5 hour**.
+ First, I performed analysis on the county-level by finding out the top 10 most mentioned counties in the mitigation measures data that we have collected during week 1 and 2. The reason I did this was because the data for any measure that involved any county was relatively small so there was no point of doing analysis on counties that were mentioned very little like counties that were mentioned one or two times.
+ The kind of analysis that I did on the county-level were the breakdowns of the mitigation type, how diverse were each of these counties are in terms of their mitigation measures, how frequent were these mitigation types (are they lumped in or spread out), the relationship between each county's population and how frequent the mitigation types were, and how does the frequency of these mitigation responses affected death/new cases/hospitalization trend.
+ The county-level analysis took about **6 hours**. Coming up with these analysis questions and continuously evaluating their merits took about 2 hours but the actual analysis/visualization took about 4 hours.
+ I then did the same kind of analysis on the state-level as well.
+ The state-level analysis took about **3 hours** to do.

+ After doing the state-level analysis, I proceeded to do analysis on the correlation between the total number of mitigation measures and death/hospitalization/positives/negative/testing rates on the state-level.
+ I also did some analysis on the correlation between the number of new public service mitigation measures and death/hospitalization/positive/negative/testing rates as well.
+ This analysis step took about an additional **5 hours**.
+ After finishing this type of analysis, I proceeded to clean up the Python Notebook to make it look more presentable and writing a final report.

Results:

+ The original dataset that my classmates and I gathered during week I and II was not clear of errors.
+ So, I had to spend about 3 hours on checking the unique values in each of the columns to see if there were any misspelled data, duplicate data or out of the ordinary data.

```
[ ]  central_data['Mitigation type'].value_counts()
```

As you can observe the figure to the right, there are misspelled errors for the "Mitigation type" column such as "Other", "work resistriction".

```
state of emergency                145
new public service                 60
wear masks                         48
work restriction                   42
Other                              30
public services reduction          19
events canceled                     7
schools closing                     6
work resistriction                  3
stay at home policy                 3
new public services\t               3
non-essential closings              3
```

There are other errors such as duplicates "stay at home policy", which means the same thing as "stay at home".

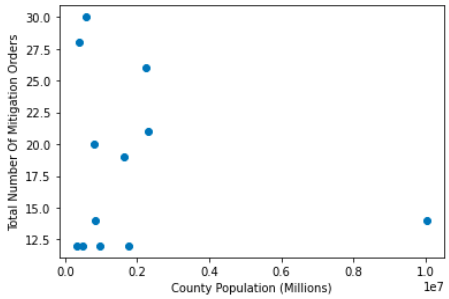
+ Before cleaning the "Mitigation type" column, there were 57 unique values, but the original intent was to have only 11 unique values for classification.
+ There were instances of values that seemed to have some sort of logical errors such as data in which the raw description might mention one type of mitigations, but the actual classification is of another.
+ There were also data that are not even about mitigating. Most of these data were about relieving the financial burdens that were observed after other mitigation effects were put in place.

```
other                          733
new public services            383
non-essential closing          322
movement restriction           248
school closing                 245
stay at home                   202
events cancelled               198
public service reduction       172
state of emergency             150
new public service              60
wear masks                      49
work restriction                49
Name: Mitigation type, dtype: int64
```

As you can see in the figure to the left, this is the unique values of the "Mitigation type" column after we cleaned it. The total number of unique values are now 11.

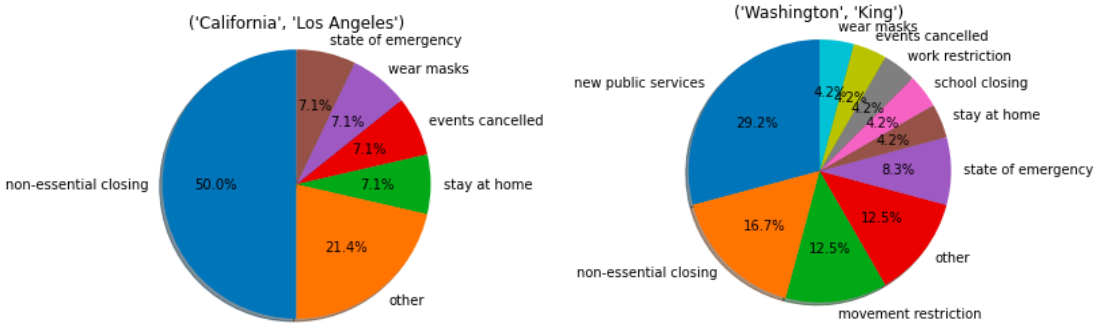
Beside the "Mitigation type" column, we also did some data quality checking and cleaning for the "Date" and "State" columns.

+ After I finished pre-processing the mitigation type data, I proceeded to find the top 10 most mentioned counties so I can do county-level analysis.
+ The first kind of analysis that I did was exploring the relationship between a county's population and the number of mitigation types it had. The assumption going to this was that if a county has a large population then we might expect to observe a higher number of mitigation orders.
+ The correlation coefficient between the county population and the number of mitigation types is about -0.16 so I would say that there is a weak relationship between them.

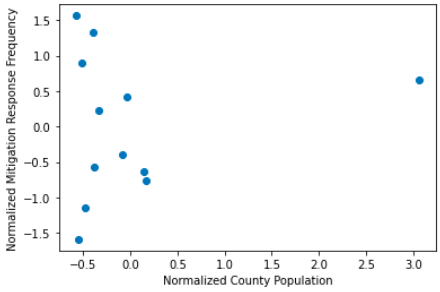


The figure to the left shows the relationship between county population and the number of mitigations. As you can see, the relationship is relatively weak. There are towns that have about half a million but with 30 mitigations while there are those with over a million that have about 12 or 13 mitigation types. So, there might not be a relationship between a county's population and the number of mitigations that it has.

+ I have also noticed that this might also be how the data was recorded too. Maybe multiple mitigation measures that were recorded separately for a county could be recorded as one.
+ Afterwards, I tried to explore the diversity of mitigation types in each of those 10 counties. The composition of the different mitigation types is different from county to county.

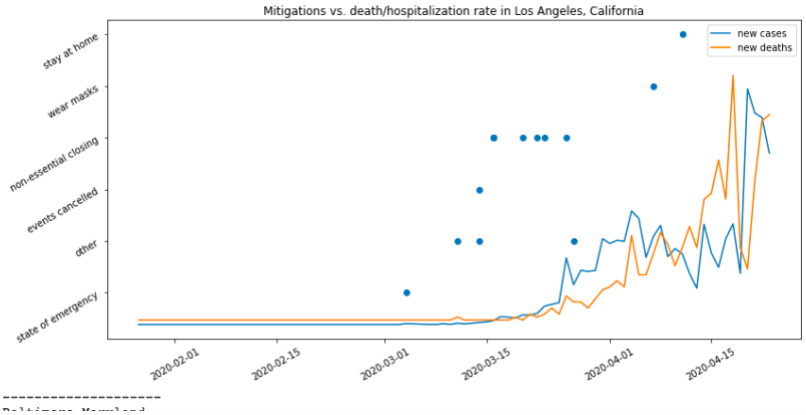


+ From observing the pie charts for all 10 counties, I would say that one county's most popular mitigation type might not be the others and these counties all have very different composition of mitigation types.
+ The most common type of mitigation responses among these 10 counties are public service reduction, non-essential closing, new public services and others. The pie charts will be included in the Appendix.
+ Next, I explored the relationship between the frequency of these orders (measured by averaging the number of orders over the date range of the dataset) and county populations. I assumed that the more people a county have, the more incentive for them to continuously putting out orders in such a short amount of time to mobilize the public and prevent the spread of COVID-19.
+ I found the correlation coefficient between the county population and the frequency of mitigation responses. It was 0.1375. So, I would say it is a relatively weak positive relationship.



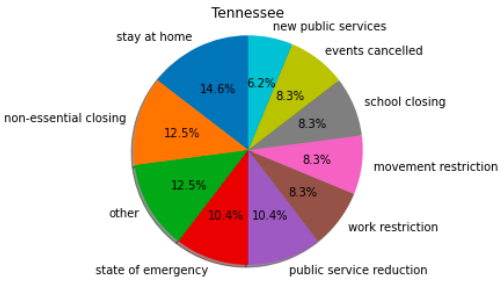
The figure to the right shows the scatter plot of normalized county population and the normalized mitigation responses frequency. The correlation coefficient is weak so there might not be a significant relationship here.

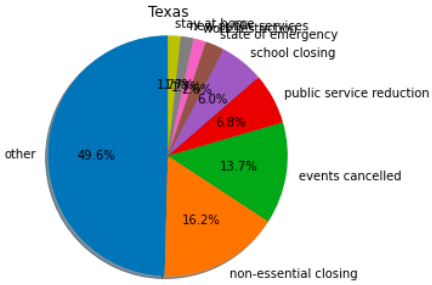
+ Then, I did some analysis and visualization on how the frequency of mitigation responses after hospitalization/death rates.
+ It looks like for the top 10 most mentioned counties; they tend to either occur before or around when the first cases/deaths reported or spread out throughout the months of March and April.
+ I also see that these mitigation measures also take place when the death rate is stabilized so we can see that these measures do have an effect on the death rate. In terms of the number of cases, I think if we can get more data on new public services that are related to testing would be great.



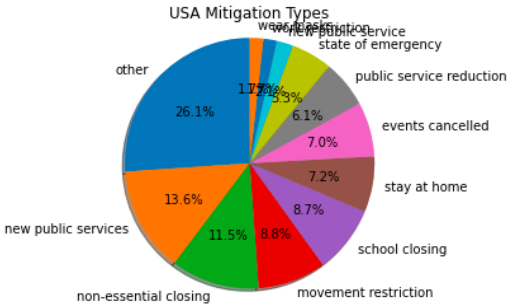
To the right, you can see that for the Los Angeles county, the mitigation responses take place before the peak in terms of new death and then it declined afterwards. All of the charts will be in the Appendix.

+ Afterwards, I followed the same analysis for state-level.
+ In terms of diversity in the different mitigation responses, every state is different from each other. Some states such as Tennessee seems to have the same proportion of every mitigation response while some states such as Texas.



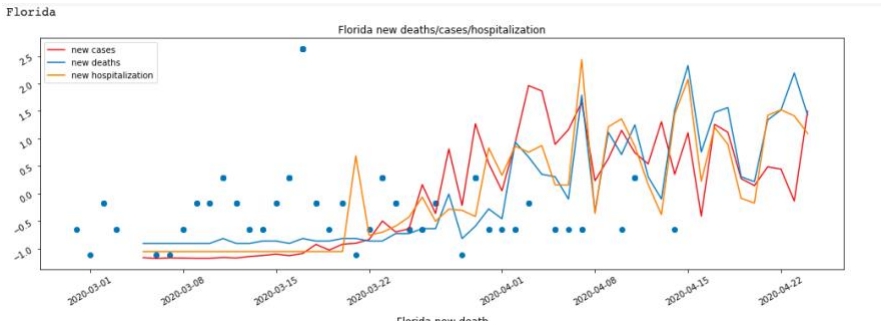


+ So not every state is the same. I also created a pie chart for the entire United States as well.

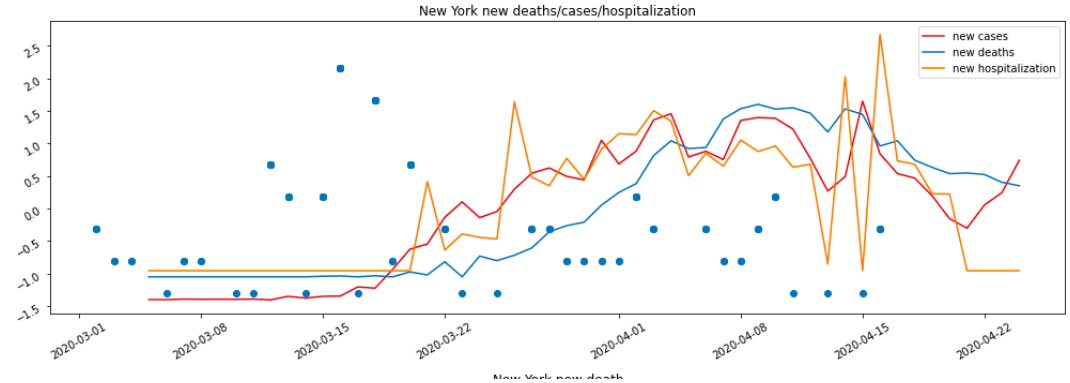


As you can see in the figure to the left, most of the mitigation measures were classified as other, new public services and non-essential closing. So, it looks like closing and active measures like new public services were rolled out to prevent coronavirus.

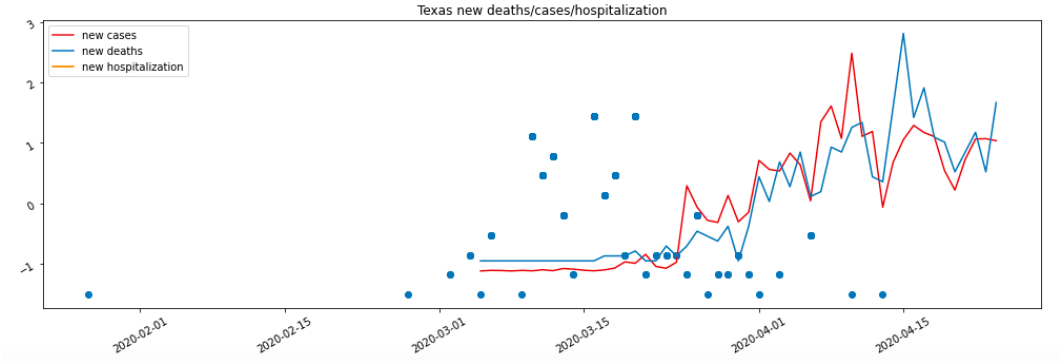
+ I wanted to explore whether mitigation responses actually help to flatten the curve by plotting the daily death/hospitalization/cases rate for the top 3 most populous states with dots as mitigation responses.



New York



Texas



+ As you can observe, the mitigations are spread out during the beginning stage up to mid-April.
+ The death/cases/hospitalization rate does not necessarily go down until late April. This does show that mitigation responses do have an effect on helping to flatten the curve.
+ Then, I explored the relationship between the number of new public services and the increase in testing/positives/negatives.
+ The correlation coefficient between a state's number of new public services and a state's total tests is 0.15924. So, there is a weak positive relationship.
+ The correlation coefficient between a state's number of new public services and a state's number of positive cases is 0.01583.
+ The correlation coefficient between a state's number of new public services and a state's number of negative cases is 0.1907. I expected this correlation coefficient to be much weaker than expected.
+ The correlation coefficient between a state's number of new public services and a state's number of deaths is 0.0226.
+ Finally, I explored the relationship between the number of mitigation responses and the increase in testing/positives/negatives.
+ The correlation coefficient between a state's number of mitigation responses and a state's total tests is 0.1954.
+ The correlation coefficient between a state's number of mitigation responses and a state's number of positive cases is -0.0373.
+ The correlation coefficient between a state's number of mitigation responses and a state's number of negative cases is 0.25547. This is a relatively modest correlation coefficient.
+ The correlation coefficient between a state's number of mitigation responses and a state's number of deaths is -0.1029. We do see a weak but negative relationship, so we do see that more mitigation responses do lower the death rate.

Conclusion:

+ Not all counties and states have similar mitigation responses. They all are very diverse.
+ There is visualization evidence that mitigation responses do in fact help in bending the curve whether that is related to death/hospitalization/new cases rate.
+ The rates in death/hospitalization/new cases do not start to flatten out or stabilize until about a month or two after the first mitigation response occurred.
+ There is a weak correlation between county population and how many mitigation responses occurred, how frequent they were.
+ There is weak to moderate correlation between the number of mitigation responses and the total test results, positive cases and negative cases.
+ There were some challenges relating to the cleaning of the mitigation dataset, but I overcame those.

Acknowledgement:

+ My classmates that I worked with during week I and II to come up with the mitigation datasets.
+ https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ (describes confirmed cases per county)
+ https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ (describes population per county)
+ https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ (describes confirmed deaths per county)
+ https://covidtracking.com/api (Under US Historical Data) Data is cumulative.
+ https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv (NYT county level data about deaths and populations)