Christopher Scott and Alexander Kim
Team Name: We Hate Coronavirus
CIS 3715 – Dr. Slobodan Vucetic

COVID-19 Final Project – Week 4 Final Report

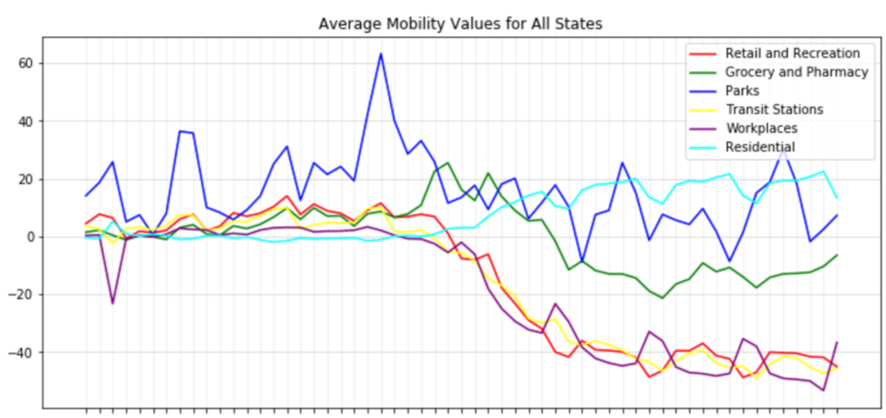
The project/task that we selected for our final COVID project was listed as number 5 amongst our options. This task required EDA to discover a relationship between distancing measures and mobility, and also required us to answer the question of whether or not there was an obvious link between these two aspects of the Coronavirus pandemic. As stated in our report from last week, we felt that it would be interesting to see how quickly mitigation/social distancing measures would take effect, and how much weight a Governor's words have on the public as a whole. The way we approached this project was to first think of questions that we wanted to have answered, then we performed EDA on the dataset in order to try and answer those questions. Since EDA in itself can be done without a clear purpose at times, we wanted to isolate a few key questions so that our EDA would be meaningful, rather than simply comparing and visualizing random bits of data without a goal. Through this project, we gained valuable skills in EDA (Exploratory Data Analysis) and experience in working with real world datasets. We were also able to extract some valuable insights into how the public reacted to these government-mandated Coronavirus mitigation orders, and how quickly they followed them as well. We felt that this project was an interesting one and allowed us to see relevant data being analyzed in a situation that directly impacts us today.

In order to perform EDA with a purpose, we came up with several questions that we wanted answers to. Some of the questions that we wanted answered include the following – "Were mitigation measures followed by a drop in mobility? If yes, was there a delay?", "Which mitigation measures had the greatest effect on mobility?", "What was the frequency of mitigation measures per state? How effectively was the government communicating with the public in each state?", and "Which state responds to mitigation orders the fastest? Which states take the least amount of time to practice mitigation measures?". The task of answering these questions required numerous hours of experimenting and playing around with different datasets. Our initial plans for the project proved to be too ambitious, and we did not find an answer to every one of our questions. However, by focusing on a few important aspects of the data we got results that were satisfying enough and allowed us to gain valuable insights.

Though this project was interesting and rewarding in the end, it came with its own set of trials and tribulations that required us to change course many times. One of the first issues we ran into was with the dataset itself. As stated in last week's report, we realized that the Apple mobility dataset that was provided to us was very limited. The dataset itself was based on a time series and only had three modes of transportation/mobility per city. In addition, the dataset itself only included select cities and was not consistent with the states/cities that we had collected data on as a class. As a resolution to this issue, we ended up using Google's mobility data. At first we were hesitant to use the data because it was presented in a .pdf format, but as we started the project, Google released a .csv version of the data which allowed us to work with it more easily. The next issue that we ran into was during the data cleaning and preprocessing phase of the project. There were two main conflicts that arose during this stage – the first being that the mitigation events related to counties were not considered in the dataset, and the second issue being that examples from the mitigation dataset were only considered if they had a valid mitigation type category. It was for this reason that the data from New Jersey was excluded from our calculations and dataset visualizations. The next issues that we ran into were restrictions due to time. It was challenging to gauge how long a particular task would take from inception to completion, and because this portion of the project was only to be done within a two-week timeframe we were limited to achievable amounts of exploration. We focused on identifying the most important facets of the data and how we could perform quality analysis of those topics.

Our project is split up into three main .ipynb files. Upon cloning or download of our project, you will notice many files within the project directory. The three main files to look for are "Data Cleaning.ipynb", "google_analysis.ipynb", and "visualizations.ipynb". This is where the majority of the work and coding is done. The csv files are data files that we used throughout the project and the text files (not including this word document) were used to write down important thoughts that we wanted to incorporate into our project.

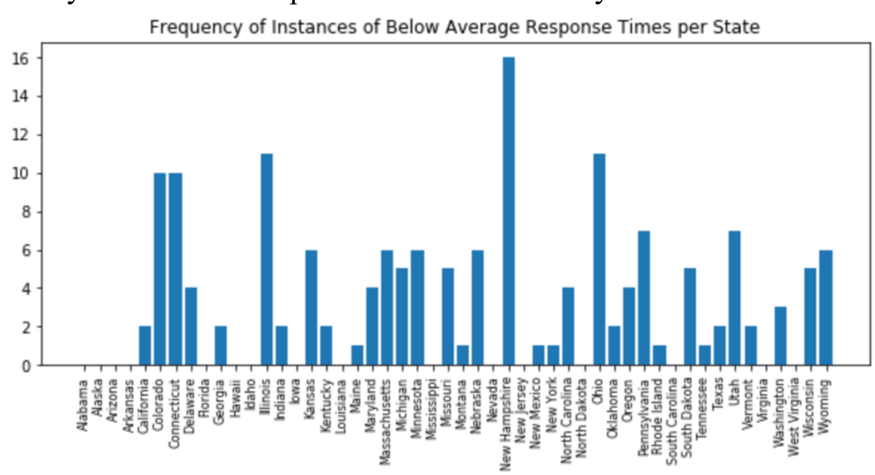
The two main files that we used to better understand the data and perform EDA were the "google_analysis.ipynb" and "visualizations.ipynb" files. They are included with this submission and report. After performing EDA and visualizations of the average mobility and mitigation events, we were able to conclude a few points that we found interesting. From the google mobility dataset, we found that it indicated large decreases in mobility for retail and recreation, workplaces, and transit stations. We also found that grocery and pharmacy mobility experienced a significant increase (> 20%) coinciding with the period of most mitigation events. This could be due to the fact that groceries and medication are considered necessary, and with the arrival of the pandemic, many people stocked up on food, supplies, and medication because they felt it was necessary to survive. This analysis was performed over the US as a whole and also on each individual state. The following figure (line chart) below illustrates our conclusions that we drew:



You may notice, however, that the "parks" category for the average mobility has a significant increase and variance in general. It's important to note the baseline mobility for parks started in February, which is still considered the Winter season. With colder weather, people are less likely to go to parks, which leads to a lower baseline and higher variance when the weather gets warmer in March and April. Another thing to note from the figure is that the residential category sees around a 20% increase in mobility after mitigation events began to be implemented. This is trend is self-explanatory. Next, we compared the average mobility and the number of mitigation events. From this we saw that there were significant drops in mobility that coincided with increases in mitigation events from state governments. The comparison of the line chart shown above and the last bar chart in the visualizations file implies this conclusion.

After performing analysis on the google data through calculated changes in mobility and delays via a “window method”, we were able to come to a few conclusions. Before mentioning our conclusions however, it is important to describe our “window” methodology that we used to perform our analysis. For the dataset, we set a “time window” of five days, where the change in mobility within a window was calculated as the difference between the greatest magnitude (min or max) mobility value in the window, and the mobility value at the mitigation event date. The delay from the mitigation event date to the mobility minimum date was also calculated. Now that the methodology has been somewhat clarified, we are able to delve into our conclusions. At first, we considered only drops in mobility. However, we realized that this was insufficient due to the fact that there were net increases in residential mobility and parks mobility, so we then turned our attention to overall changes in mobility. We also found that the average delay between mitigation events and changes in mobility was very consistent across categories within the time window. The heatmaps in the file titled “google_analysis.ipynb” illustrate the conclusions that we drew from the data. One potential drawback to this methodology, we realized, was that the concurrency of mitigation events would not allow us to conclude on the casual effect of mitigation events themselves.

The final piece of analysis that we performed was in regard to which states were the quickest/best at following mitigation measures that were announced by the governor. The way we measured this was through the use of two baselines – the average mobility change and the average delay for the given state. We then looked at each time window (time window = 5) and measured how long that specific state would take to reach the average change in mobility. Then we compared that time period to the average delay, and if the delay for that time period took longer than the average delay, then we incremented a variable that kept count of the frequency of these types of delays. The bar chart produced from this analysis can be found below.



From this figure, we can see that the three states that seem to be the worst at responding to mitigation orders are New Hampshire, Ohio, and Illinois. We wanted to delve into reasons as to why this was (literacy rates in these states, governor approval ratings, etc.) but we were restricted by time. Overall, these conclusions that we drew from this project were able to help us gain insights on this Coronavirus pandemic that we hadn’t had before, and the conclusions in this section help us feel satisfied. We have spent a total of 15 hours on this project in weeks 3 and 4.