# The Effects of Social Distancing on COVID-19

Sangseok Lee, Varun Prasad, Nathan Warren

IDS 690-04: Unifying Data Science
April 28, 2020

## Project Motivation

On December 31st, 2019, Chinese health officials confirmed the spreading of a novel acute respiratory disease. A week later, on January 7th, Chinese scientists confirmed that this disease was associated with a novel coronavirus (2019-nCov), which is now widely referred to as COVID-19.[1] Due to its worldwide spread, on March 11th, the World Health Organization (WHO) classified COVID-19 as a global pandemic.[2] COVID-19 is caused by a respiratory virus that is primarily transmitted by aerosol droplets. Close contact between individuals at home, in schools and workplaces, on transport, and at community gatherings is ideal for viral transmission.[2] As a result the WHO has recommended that people follow 'social distancing' protocols, intended to reduce close interactions between individuals, thereby slowing the spread of the virus. The focus is not necessarily on eradicating the virus but more on preventing the healthcare system from becoming overwhelmed, commonly known as "flattening the curve."[3] As the cases in the United States began to rapidly increase in March and April of 2020, many states closed schools, shut down public events, and implemented social distancing methods such as stay-at-home orders and shelters-in-place.[4] In this report, we analyze the effect of social distancing on the number of new COVID-19 cases per day and the mortality rate.

## Research Design Motivation

To thoroughly assess the effectiveness of social distancing on reducing new infections and mortality, we used a regression analysis. We hypothesize that had social distancing not been practiced, the number of new cases per day would have continued to increase exponentially (up to a certain point). With the practice of social distancing, flattening of the curve is present. It is important to note that there is a time lag associated with both the virus and with social distancing. When someone is exposed, they typically do not show symptoms for between 5 and 14 days, and only after they are diagnosed are they counted as a case. Social distancing as well will only reduce the number of cases projected to occur in a two-week period. Therefore, we

shifted the graphs of our social distancing metric by two weeks in order to account for this time lag between social distancing and number of cases.

# Data

**Census Data**

The population data was extracted from the United States Census Bureau, which had data for all US counties from 2010 to 2019.[5] In order to generate populations for 2020 for proper use with 2020 COVID data, we found the average linear increase in population from 2010 to 2019 and added this value to the population in 2019. The county names were also cleaned to remove extra characters and fully capitalized to make merging with other datasets easier. This dataset was then merged with the FIPS dataset by county and state to create matching FIPS codes for each county. County names were double-checked and adjusted to ensure proper matching with their respective FIPS code in the FIPS dataset.

**FIPS Data**

The county-level FIPS dataset was downloaded from the National Resources Conservation Service (NRCS).[6] This was used for merging our predicted 2020 county-level population to the New York Times COVID-19 dataset. Counties that were present in the Census data but not in the FIPS data were researched and added with the proper FIPS code after merging.

**COVID-19 Statistics**

COVID-19 infection and death counts were downloaded from the New York Times' Github repository for COVID-19 infection.[7] Records are only shown when a county acquired an infection, with the earliest known U.S. case being in Washington state on January 21, 2020. To properly merge this dataset with the SafeGraph dataset, we had to add in zero rows for dates that did not have infections for all states. The data also only looks at where patients are treated as opposed to where they live. Non U.S. States and Alaska were removed due to oddities in COVID-19 cases, FIPS, and population estimates for areas. This dataset also contained FIPS codes for county and state, which allowed us to merge this data with our merged projected 2020 population-FIPS dataset.

**SafeGraph - Social Distancing Metrics**

The data from SafeGraph was used to quantify social distancing. Specifically, it uses the GPS-location data from anonymized mobile devices.[8] The SafeGraph data was downloaded from a secure SafeGraph AWS bucket. The dates for this data ranged from 1/21/2020 to 4/16/2020. FIPS and date uniquely identified each row of data. Our variable of interest in this dataset was 'median_home_dwell' which was the total non-contiguous minutes at home. We also used

‘device_count’ from this dataset to help weight our ‘median_home_dwell’ when aggregating from block to county-level. To do this we multiplied ‘device_count’ by ‘median_home_dwell’ for each row and then grouped by 5-digit FIPS and date, to achieve county-level data. The summed product of ‘device_count’ and ‘median_home_dwell’ was then divided by the sum of device_count to achieve a weighted mean home dwell time for the county. This process was done once more as we grouped by 2-digit FIPS and date to achieve state-level data. While this process may not give us the true mean, it is the closest we can get with the present data.

**Data Merging**

Datasets were cleaned to ensure proper 1:1 merging. Census data was first merged with the FIPS data on county and state names. Discrepancies, such as spelling errors or new counties (ex: Broomfield County in Colorado), were identified and corrected. The New York Times COVID-19 dataset was then merged to this FIPS-POP dataset on FIPS to generate a dataset we called PF-NYT. PF-NYT was then merged with the SafeGraph data on FIPS at county-level to generate our final two datasets, which were county and state-level.

# Summary Statistics

We present several key summary statistics of the final merged dataset. These statistics include items such as the number of missing or zero entries and the states with the highest number of infections.

- Zero values for cases column: 2046
- Non-zero values for cases column: 2304

**Table 1**: State Summary Statistics as of 4/16/20

| | **Most** | **Average** | **Least** |
|---|---|---|---|
| **Infections** | New York: 99138 | Maryland: 10784 | Wyoming: 295 |
| **Deaths** | New York: 3560 | Texas: 428 | Wyoming: 2 |
| **Infections (per 10k population)** | New Jersey: 83.61 | Arizona: 14.89 | Minnesota: 3.39 |
| **Deaths (per 10k population)** | New Jersey: 3.96 | Connecticut: 0.576 | Wyoming: 0.35 |
| **Dwell Time (Average)** | California: 767 mins | Iowa: 652 mins | Vermont: 459 mins |

These statistics give us the most extreme states to look in for total cases and deaths and total cases and deaths normalized by population. What was interesting was that Minnesota had the least amount of infections when normalized by population but surprisingly Wyoming had the lowest death per capita rate. This is likely due to the fact that Wyoming only has had 2 deaths as of 4/16/20.

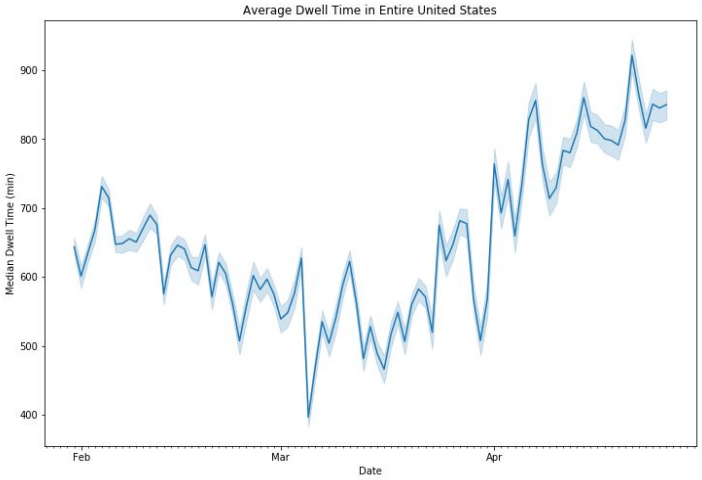


**Figure 1**:United States Average Social Distancing

To understand how the U.S. social distancing was practiced, we first created a plot that showed all state dwell time. This proved to be quite messy so here we display the social distancing average for the United States with confidence intervals. We found that all states had a very similar trend in regards to social distancing.

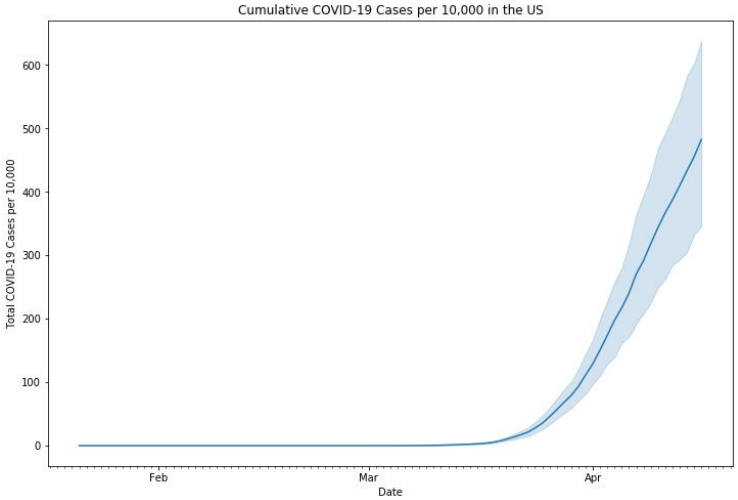


**Figure 2**: United States Cumulative COVID-19 Cases per 10,000 capita

By plotting U.S. case count we were able to see the somewhat exponential growth in COVID-19 cases in the United states.

# Analysis and Interpretation

We decided to use regression tables to analyze the impact of social distancing, as a continuous variable, on the number of new cases in a given state. Our outcome variable was new cases of COVID-19, as we wanted to predict how the effect of social distancing decreased the amount of new cases. We implemented a 2 week lag time for our social distancing metric "home dwell", or dwell as we have called it here. The purpose of having a 2 week lag time is that an increase in social distancing will only have an effect on the spread of COVID-19 around 1-2 weeks after it is employed. This is due to the fact that those infected with COVID-19 may not show symptoms for up to two weeks and additional time is required for the person to be registered as having COVID.Cases was used as a predictor variable as well. The reason behind this is that the higher the amount of cases in a given state, the more likely it is to spread to others.

**Table 2**: New Jersey Regression Table

| **New Jersey** | Coef | Std Error | t | P>\|t\| | [0.025] | [0.975] |
|---|---|---|---|---|---|---|
| Intercept | -2115.7 | 391.123 | -5.409 | 0 | -2893.710 | -1337.8 |
| Cases | 0.0447 | 0.005 | 9.218 | 0 | 0.035 | 0.054 |
| Dwell (2 week lag) | 3.3425 | 0.548 | 6.101 | 0 | 2.253 | 4.432 |

In New Jersey, which has the highest COVID infection rate per capita, we actually saw a positive and significant Dwell coefficient. The correct interpretation, in this context, of this positive value is that social distancing actually increases infection rate (new cases), although we know this not to be the case. Since our metric for social distancing was present before infections began to occur, it makes sense why a practice such as social distancing increasing may appear to look like it is contributing to new cases. However, this is a matter of correlation and not causation. We were hoping that by including cases as a predictor that this effect would not occur. While it is present here, we can use this regression table and compare it to other states. One way this might be able to be solved is to continue collecting data and thinking about other variables that may contribute to infection spread. Other situations, such as lack of tests earlier on may have also led to the positive dwell value seen in the regression table above.

**Table 3**: New York Regression Table

| **New York** | Coef | Std Error | t | P>\|t\| | [0.025] | [0.975] |
|---|---|---|---|---|---|---|
| Intercept | 2452.1728 | 686.2 | 3.574 | 0.001 | 1087.341 | 3817.005 |
| Cases | 0.0618 | 0.004 | 14.055 | 0.000 | 0.053 | 0.071 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Dwell (2 week lag) | -3.3168 | 1.067 | -3.108 | 0.003 | -5.439 | -1.194 |

New York has the most COVID-19 cases. New York actually shows a negative coefficient value for dwell which is significant. For each unit increase in dwell, our social distancing metric, we see that new cases, our outcome variable, decreases by 3.31 units (people). An additional person sick in New York increases the prediction for new cases by 0.0618.

**Table 4**: Wyoming Regression Table

| **Wyoming** | Coef | Std Error | t | P>\|t\| | [0.025] | [0.975] |
|---|---|---|---|---|---|---|
| Intercept | 4.6784 | 2.93 | 1.592 | 0.115 | -1.165 | 10.522 |
| Cases | 0.0446 | 0.008 | 5.894 | 0.000 | 0.030 | 0.060 |
| Dwell (2 week lag) | -0.0063 | 0.005 | -1.261 | 0.211 | -0.016 | 0.004 |

Wyoming had the least amount of infections. Here we see that an additional COVID-19 case led to an increase in our outcome variable, new cases, by 0.0446. On the other hand a unit increase in dwell predicts a 0.0063 unit (person) decrease in new COVID-19 cases, indicating that social distancing does help decrease the infection rate. However, the p-value here is not significant, meaning that there is a 21.1% chance that the value we see here for dwell is simply due to variable chance. We also see that the confidence intervals span over 0, indicating that dwell may have no impact on new COVID-19 cases. Although for the same reasons stated in the New Jersey regression table analysis, we do not necessarily believe this to be true.

**Table 5**: Minnesota Regression Table

| **Minnesota** | Coef | Std Error | t | P>\|t\| | [0.025] | [0.975] |
|---|---|---|---|---|---|---|
| Intercept | 16.257 | 8.271 | 1.864 | 0.066 | -1.093 | 33.597 |
| Cases | 0.0637 | 0.004 | 15.771 | 0.000 | 0.056 | 0.072 |
| Dwell (2 week lag) | -0.0190 | 0.014 | -1.385 | 0.170 | -0.046 | 0.008 |

Minnesota had the highest amount of infections per capita. For the Minnesota regression, we see similar results to what was observed with Wyoming. The p-value for dwell is not significant and the confidence interval for dwell spans over 0. Cases on the other hand, is indeed positive. For every additional case in Minnesota, the model predicts an increase in new cases by 0.0637.

# Discussion

Overall, the regression tables show that social distancing may lead to a decrease in new infections. Despite these results, there are several limitations and additional areas to explore. One is the limitation of the date range of the data. We only explored data until the middle of April, but social distancing measures were in place throughout the month and only now are starting to be slowly relaxed. As a result, we may have seen more of social distancing's effects if we extended our date range to present day. Additional analysis will encompass this full data range to more effectively quantify the effects of social distancing. There is also a limitation in the data itself regarding the number of cases. The United States did not initially test at a high capacity, so there are likely many people with COVID-19 who were not diagnosed for a long time. This is also true given how many people infected with the virus are asymptomatic. Therefore, the effects of social distancing might not accurately be seen in both statistical models or visually because many of the people infected were getting tested much later, leading to high increase in new cases.

Another limitation was that we were unable to effectively perform a diff-in-diff analysis to truly see the effect of social distancing across different states. From our exploration, we found that almost all states had a similar trend in social distancing as measured by median dwell time. Therefore, it was difficult to truly have a control and treatment group to compare how strong the effects of social distancing are. Given that this was a national emergency and a global pandemic, almost all state governments were strict on closing businesses and enforcing social distancing, so going against these guidelines was extremely unlikely to happen.

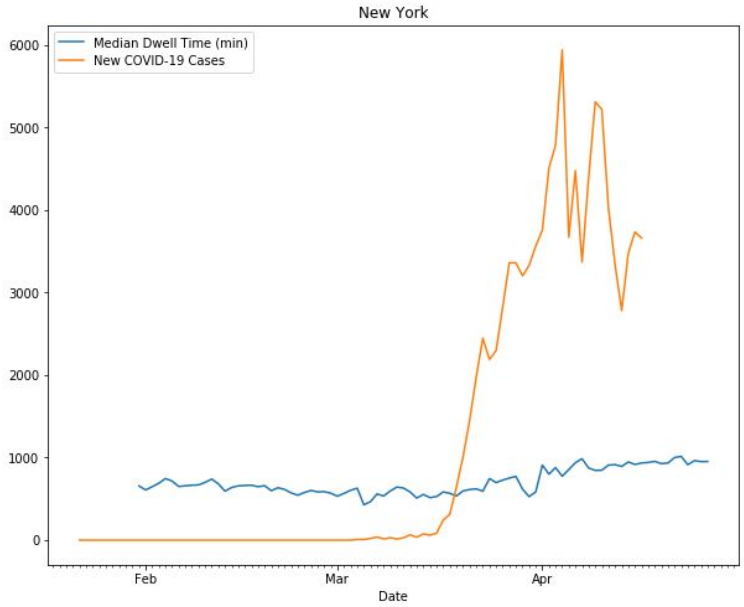

**Figure 3:** New York Dwell Time (2 week lag) and New Cases

Due to the variability of our results we wanted to show the plot above. In orange is 'daily new cases' which is the measure we use to determine how fast the infection rate is increasing or decreasing. For many states, NY included, we can see a correlation between social distancing increases and decreases related to new COVID-19 cases. Although not very scientific, an eye-test does indeed show that social distancing does indeed have a negative effect on new cases. Our 2-week lagged scale may be too high or too low and many other variables such as population density, demographics, closest hospital, etc., may have played an important role in predicting new cases (infection rate).

Overall this simple analysis did show the effects of social distancing on reducing the exponential growth of new COVID-19 cases. Still, more extensive analysis using more data and creating more complex models can better represent the data and provide more insight into the effects of social distancing on reducing the spread of COVID-19. In the future, we hope to incorporate more variables into our models to make them more robust and to better assess how effectively social distancing reduced the exponential growth of COVID-19.

# References


1. CDC. (2020, February 11). *Coronavirus Disease 2019 (COVID-19)*. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/index.html
2. *Coronavirus: COVID-19 Is Now Officially A Pandemic, WHO Says: Goats and Soda: NPR*. (n.d.). Retrieved April 28, 2020, from https://www.npr.org/sections/goatsandsoda/2020/03/11/814474930/coronavirus-covid-19-is-now-officially-a-pandemic-who-says
3. *Flattening the Curve for COVID-19: What Does It Mean and How Can You Help?* (n.d.). Retrieved April 28, 2020, from https://healthblog.uofmhealth.org/wellness-prevention/flattening-curve-for-covid-19-what-does-it-mean-and-how-can-you-help
4. *Stopping COVID-19 with New Social Distancing Dataset*. (n.d.). Retrieved April 28, 2020, from https://www.safegraph.com/blog/stopping-covid-19-with-new-social-distancing-dataset
5. Bureau, U. C. (n.d.). *County Population Totals: 2010-2019*. The United States Census Bureau. Retrieved April 28, 2020, from https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html
6. *County FIPS Codes | NRCS*. (n.d.). Retrieved April 28, 2020, from https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/home/?cid=nrcs143_013697
7. *Nytimes/covid-19-data*. (2020). The New York Times. https://github.com/nytimes/covid-19-data (Original work published 2020)

8. *Social Distancing Metrics*. (n.d.). SafeGraph. Retrieved April 28, 2020, from https://docs.safegraph.com/docs/social-distancing-metrics