Unifying Data Science 2023 Project Proposal

Suzanna Thompson, Echo Chen, Pragya Raghuvanshi, Ahmed Ibrahim

TOTAL POINTS

1.9 / 2

QUESTION 1

1 Exploratory Questions 0.9 / 1

+ 0.85 pts A reasonable start, but the data is not very well presented and/or thoughtfully interpreted.

✓ **+ 0.9 pts** *Good!*

+ 0 pts Always cite your data sources!

+ 0.95 pts Great! salient questions investigated well

+ 0.8 pts The data is not very well presented and/or thoughtfully interpreted.

+ 0.65 pts You didn't answer any of your exploratory questions, and it's not clear how they relate to your problem

I don't know that you actually answered your second exploratory question. I'm giving you the full benefit of the doubt on your third one only because you showed me plots the shows you weren't just trying to figure out which states were not party to the first wave of stay at home orders, but also what the distribution of implementations were for the first wave.

1 I don't think that sentence says what you think it does...

2 won't counts almost by construction tell you which states are the biggest?

3 You made plots of who did what when! You should have included those.

4 This only just barely counts as "answering a exploratory question"—what are the relative frequencies?

QUESTION 2

2 Backwards Design 1 / 1

+ 0.9 pts A promising project! Some facets still to be worked out, but you're off to a good start.

- 0.05 pts In terms of solving the problem you state motivates your project, the question you're seeking to answer doesn't quite connect.

✓ **+ 0 pts** *Analyzing the effects of anything on the spread of covid is a little tricky, and I will caution you that we have had several teams attempt to study the spread of covid in the past and those projects have generally ended up feeling a little unsatisfying to the teams.*

First, when studying covid, what we are often most interested in is the effect of an intervention on _infection rates_. But the data that's available—positive covid test counts—is only a loose proxy for infections, both in the sense that many people who get sick don't get tested (in the likelihood

that somebody tries to get a test may be related to other factors like their political inclinations), and in the sense that it can be hard to know when exactly somebody got infected who tests positive (one can test positive as little as two days after infection, or as many as 10 days). That means that you are trying to measure the effect of interventions on an outcome that you are only able to measure with significant noise.

Second, covid often follows exponential growth curves, which makes modeling with a simple linear regression a little simplistic.

And third, often times interventions that have a simple name ("stay at home orders" or "mask requirements") have very different implementations and compliance rates in different places.

None of that means that it's something you aren't allowed to study, I just want to give you a heads up that there are a lot of challenges surrounding this topic.

✓ **+ 1 pts** *Really nicely developed!*

+ 0.95 pts Great!

+ 0 pts By far the easiest way to get us census data at a county level or for cities is by using https://www.nhgis.org/ —it has everything the senses bureau has, but with a much better interface and with results pre-aggregated to different geographic levels

+ 0.85 pts This is a reasonable start, but there are some issues to address.

+ 0.95 pts Giving it a solid score since this is a subtle thing, but something we need to discuss more/refine.

5 But not proofread for typos apparently... ;P

6 You may find it more fruitful to look at the growth rate in infections, at least during the first wave. But that's the type of functional form exploration you will definitely need to undertake to ensure your modeling they data correctly.

7 Same point.

8 □□□

9 You could potentially do the analysis at the county level. Granularity is basically always good.

10 □□□

11 Is there enough overlap in demographic/geographic characteristics to allow for matching these treatment states with these control states?

12 If you're able to match on county level, you'll end up with a lot more overlap

13 Cool! I really like the idea of using mobility data

Estimating the Impact of Stay-at-Home Orders during the COVID-19 Pandemic

1 Topic

1.1 Background

COVID-19 was declared a global pandemic on 11, March 2021. In the subsequent weeks and months, states in the United States began to implement measures in order to slow the pandemic or to "flatten the curve." Such measures include stay at home orders, school closures, masking mandates, social distancing requirements, and a required increased sanitization routine. This followed worldwide measures, both health and regulatory to slow the pandemic.

In the beginning of the pandemic, Italy received a lot of media attention as they were a community to be hit with COVID-19 early. In their paper, "The impact of first and second wave of the COVID-19 pandemic in society: comparative analysis to support control measures to cope with negative effects of future infectious diseases", Coccia writes about Italy's response to the pandemic in the first and second waves. They cite that Italy's response to their first and second waves had an impact on the overall health. They write, "The results here suggest that the impact of COVID-19 on the health of people depends on manifold environmental, climate, social and economic factors and policy responses of governments."

1.2 The Problem

The problem that we are seeking to explore is about the impact of early intervention, specifically the intervention of a stay-at-home order or a lockdown order. Stay-at-home orders were one of the most important noon pharmaceutical intervention (NPI) issued by the government during the pandemic as a community mitigation strategy to reduce the movement and thereby the spread of the pandemic. The problem that we're seeking to investigate is how does a stay-at-home order, which is considered to be an extreme mitigation measure, enacted in the first wave of a pandemic effect the subsequent outcomes after catching the viru

1.3 Domain and Application

The domain that this project focuses on is Healthcare and Public Health. It is important that public health decisions are evidence-based and statistically sound. In our project, we hope to provide a scientific and statistically sound approach to evaluating how lockdown and stay-at-home orders affected the first-wave outcomes of COVID-19. The theoretical application of this research is to inform future early government interventions for future pandemics. Ideally, our research will show if the lockdowns were effective or not, which could inform public health officials on how to operate if another pandemic occurs.

2 Project Question

What was the causal effect of state government-imposed lockdowns on the spread of COVID-19 (measuring death rate and infection rate) during the first wave, controlling for other factors such as population density and compliance with spread slowing guidelines?

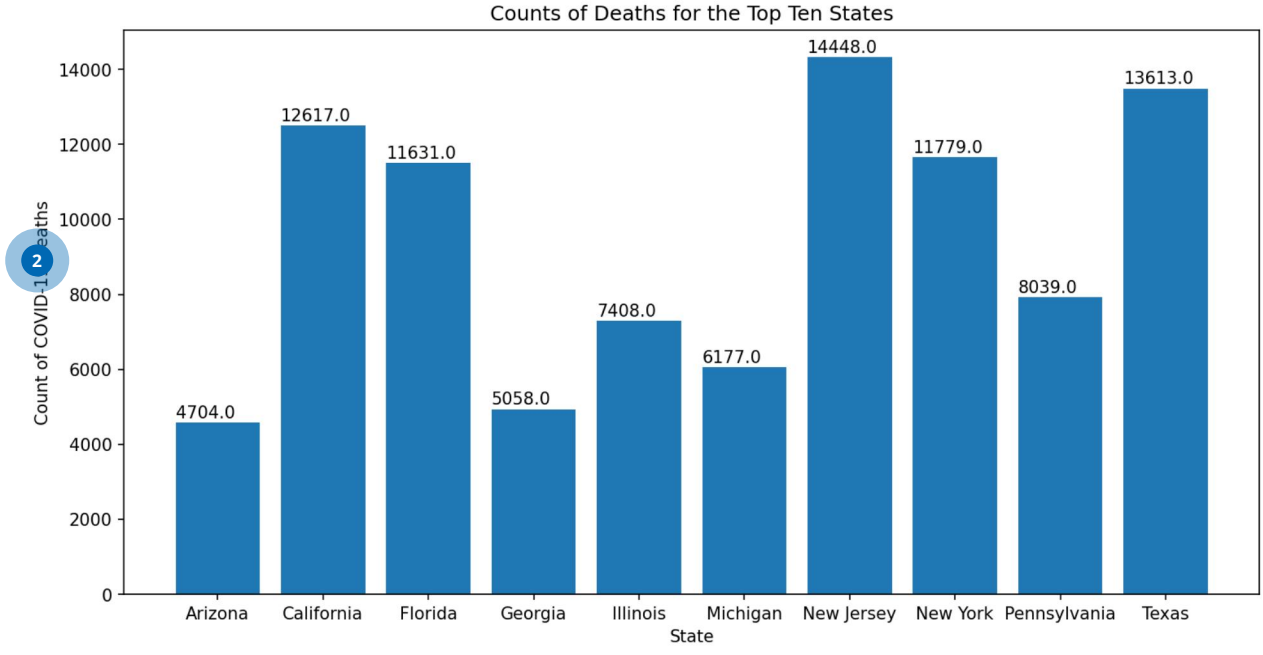
3 How Will Answering This Question Help Address Your Problem?

Answering the exploratory questions will inform our understanding of the causal effect of state government-imposed lockdowns on the spread of COVID-19 during the first wave. Each question contributes essential context and preliminary data to analyze the effectiveness of stay-at-home orders and other policies implemented in response to the pandemic.

- What states had the highest count of deaths within the first wave of COVID-19?

  Identifying the states with the highest death and case growth rates will inform our analysis of the effectiveness of various policy measures and lockdown strategies implemented by each state. Comparing the outcomes in states with different levels of severity will help us understand the factors that contributed to these high rates, such as population density, demographics, and healthcare infrastructure. This information will enable us to control for these factors when assessing the causal effect of lockdowns on death and infection rates.

  The first wave of COVID-19 is defined to be from March 11, 2020 until August 20, 2020. Using weekly COVID-19 death data, we found New Jersey, Texas, California, New York, and Florida to have the five highest number of deaths, with 14,326, 13,491, 12,495, 11,657, and 11,509 deaths respectively. (1) Below is a graph showing the ten states with the highest number of deaths.



- What types of policies were implemented in response to COVID-19 during the first wave?

  Understanding the different types of policies implemented in response to COVID-19 will inform our evaluation of various approaches taken by states to manage the pandemic. This information will help us identify which policies were more successful in controlling the spread of the virus and inform our analysis of the causal effect of stay-at-home orders. Moreover, we can reveal potential correlations between specific policies and their impacts on death and infection rates, further refining our understanding of lockdown measures' effectiveness.

  The most common policies that were enacted by states were stay-at-home orders, mask mandates, social distancing requirements, closure of 'non-essential' businesses, where the definition of 'non-essential' varied from state to states, travel restrictions, school closures, and contacting tracing. (2)

- How many states did not impose a stay-at-home order within the first two weeks of the defined first wave of COVID-19?

  Determining the number of states that did not impose a stay-at-home order during the initial weeks of the pandemic will provide a basis for comparison when evaluating the causal effect of lockdown measures. By comparing the outcomes in states with and without stay-at-home orders, we can better understand the role of these policies in curbing the spread of COVID-19. Furthermore, this information can shed light on alternative strategies that were used in the absence of stay-at-home orders, which might offer additional insights into effective pandemic management.

  We found that seven states did not impose a stay-at-home order in the first wave of COVID-19. (3). They are Arkansas, Iowa, Nebraska, North Dakota, South Dakota, Utah, Wyoming.

4 Ideal Experiment

Since I am God have the ability to observe parallel universes and control all variables to create an ideal experimental setting

4.1 Experimental Setup:
Create two identical parallel universes, Universe A and Universe B, with the same states and populations, all experiencing the first wave of the COVID-19 pandemic. Ensure that all other factors, such as population density, demographics, healthcare infrastructure, and compliance with spread-slowing guidelines, are identical between the two universes.

4.2 Treatment Variable:
In Universe A, all states will implement the treatment variable: state government-imposed lockdowns. These lockdowns will include strict stay-at-home orders, school closures, masking mandates, social distancing requirements, and increased sanitization routines.

In Universe B, none of the states will implement government-imposed lockdowns. However, they will receive guidelines on preventing the spread of the virus, but without any strict enforcement of stay-at-home orders or other restrictive measures.

4.3 Outcome of Interest:
The primary outcomes of interest would be:
Infection rate: The number of new COVID-19 cases per capita in each state during the first wave of the pandemic in both universes.
Death rate: The number of COVID-19-related deaths per capita in each state during the first wave of the pandemic in both universes.

4.4 Comparison and Analysis:
By comparing the outcomes (infection rate and death rate) between the states in Universe A (lockdown) and Universe B (no lockdown), we can accurately determine the causal effect of state government-imposed lockdowns on the spread of COVID-19 during the first wave.

Since all other factors are identical between the two universes and no other policy interventions are allowed to occur simultaneously, the results will provide a clear understanding of the effectiveness of stay-at-home orders and other restrictive measures in curbing the spread of COVID-19. This valuable information will inform future pandemic management strategies and guide policymakers and public health officials in making well-informed decisions.

5 Pick a Study Context

To measure the outcome variables (infection rate and death rate) during the first wave of the COVID-19 pandemic, we will need data on confirmed cases and deaths on a state-by-state basis in the United States. This data should include daily or cumulative cases and death counts over time.

Potential data sources: National health departments, reputable news organizations, or research institutions often maintain and publish such datasets. In the U.S., on websites of organizations like the Centers for Disease Control and Prevention (CDC), COVID Tracking Project, or Johns Hopkins University.

To include variation in the treatment variable (state government-imposed lockdowns), we will need data on the specific policies and measures enacted by each state during the first wave of the pandemic. This data should contain information on the start and end dates of stay-at-home orders, masking mandates, social distancing requirements, and other relevant policies.

Potential data sources: Government websites, think tanks, or research institutions that maintain policy databases or compile state-level policy information. Examples include the National Governors

Association, the National Conference of State Legislatures (NCSL), or the COVID-19 US State Policy Database (CUSP).

Additionally, To find similar states for comparison, we can use "matching", this method involves identifying pairs of states with similar characteristics, such as population density, demographics, and healthcare infrastructure, but with different policy implementations (lockdown vs. no lockdown)., we can prepare some date on these factors.

6 Project Design

The question of interest in this project is how effective were the lockdown and stay at home orders, issued during the COVID-19 pandemic, in curbing the spread of pandemic measured by growth of infection and death rate. For this purpose, we define our hypothesis as:

- H0: Stay at home orders do not significantly impact the growth of the infection and death rate (COVID-19 spread).
- H1: Stay at home orders do significantly impact the infection and death rate (COVID-19 spread).

To answer the question we will use the following method design:

1. **Pre-post analysis**: Pre-post analysis is a research design that compares the outcome before and after the implementation of an intervention to estimate its effect. In the context of our objective, pre-post analysis will be used to estimate the impact of stay-at-home or lockdown orders on the growth rates of infections and deaths. This involves defining the intervention, the pre-intervention and post-intervention periods, and the outcome variables. The treatment effect, that is the stay at home orders, can be estimated by comparing the observed outcomes in the post-intervention period with the outcomes in the pre-intervention period. Statistical tests can be used to determine the significance of the treatment effect. However, pre- post analysis is a simpler and straightforward approach to estimating causal effects and it's important to note there may be multiple other factors that could affect the outcomes, and it may be challenging to establish causality between the intervention and the observed outcomes. Therefore, this necessitates the use of more comparative analysis of the effects of policy change on the base.

2. **Difference in difference analysis**: The difference-in-differences (DiD) method is used to determine the effectiveness of interventions or treatments by comparing changes in outcomes over time between a treatment group and a control group. In our scenario, we will compare the treatment effect, that is stay at home orders, on the treatment group by comparing the, to a counterfactual group that will mimic the treatment group had there been no policy change. In other words, we can say that for difference in difference to be effective the counterfactual states should follow a similar trend to the treatment state under consideration, prior to the policy change. The idea is to create a parallel universe for the base state where the policy change has not taken place, and compare it with the reality where the policy has been put in place. This will help

us better gauge the effectiveness of policy in a conclusive manner as compared to the earlier approach of pre-post analysis. Therefore, we will perform a difference in difference to gauge the effectiveness of stay at home orders by comparing our treatment states, that is, states that issued a stay at home orders, to our control states, that is, states that did not issue stay at home orders. The effectiveness will be determined by measuring the outcome of the COVID-19 spread, that is the growth in the infection and the death rate.

3. Our analysis will involve estimating the following four parameters:
    a. $\bar{y}\, T = 1,\ Post$, which is the average outcome of treatment group, post treatment
    b. $\bar{y}\, T = 0,\ Post$, which is the average outcome of control group, post treatment
    c. $\bar{y}\, T = 1,\ Pre$, which is the average outcome of treatment group, pre treatment
    d. $\bar{y}\, T = 0,\ Pre$, which is the average outcome of control group, pre treatment

4. The pre- post analysis will involve estimating the following:

$$\bar{y}\, T = 1,\ Post - \bar{y}\, T = 1,\ Pre$$

5. The difference in difference estimator $\hat{\delta}$ is defined as follows:

$$\hat{\delta} = (\bar{y}\, T = 1,\ Post - \bar{y}\, T = 1,\ Pre) - (\bar{y}\, T = 0,\ Post - \bar{y}\, T = 0,\ Pre)$$

Assumptions in difference in difference analysis:

1. Treatment/intervention and control groups have Parallel Trends in outcome pre intervention
2. Intervention unrelated to outcome at baseline (allocation of intervention was not determined by outcome)
3. SUTVA

7 Model Results

The methodology will involve performing a regression analysis with the outcome variable as the dependent variable and the treatment status, time period, and their interaction as independent variables. The interaction term captures the difference in the change in the outcome variable between the treatment and control groups over time.

$Y = \beta0 + \beta1*[Time] + \beta2*[Intervention] + \beta3*[Time*Intervention] + \beta4*[Covariates] + \varepsilon$

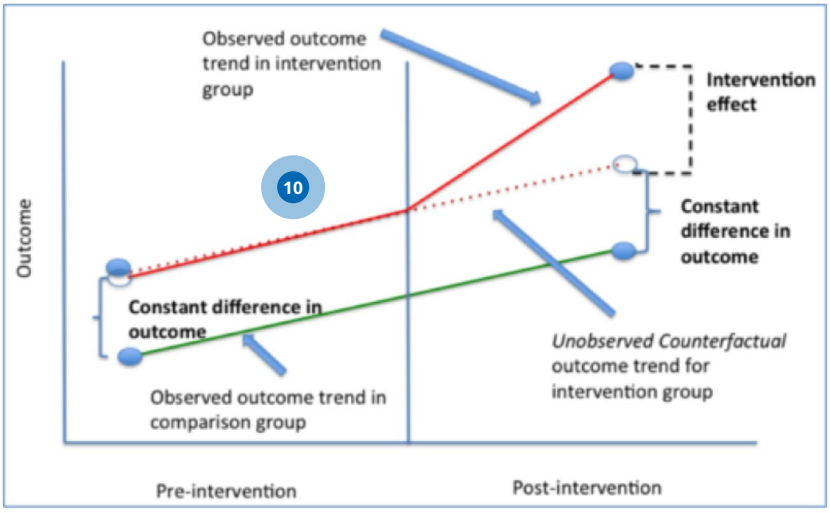
The results of a difference-in-differences (DiD) analysis can be depicted in two ways:

1. The results are typically presented in a table that shows the estimated treatment effect and its statistical significance.

Coefficient	**Interpretation**
β0	Baseline Average
β1	Time trend in control group
β2	Difference between the two groups due to the treatment
β3	Difference in changes in groups over time.

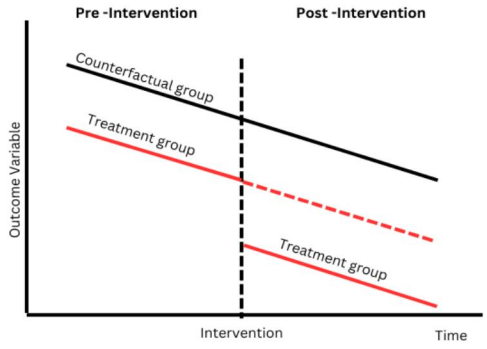
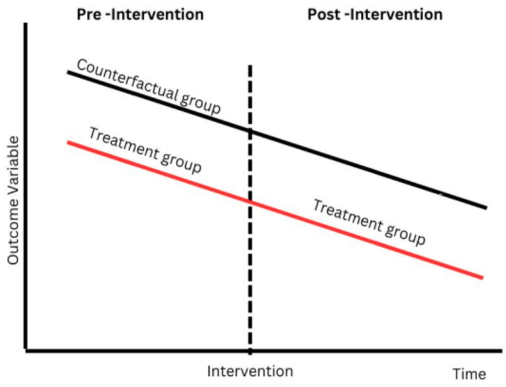
In our analysis the interaction term will capture the difference in the change in the outcome variable, that is death rate and infection rate, between the treatment and control groups after the treatment period.

2. The results can also be visualized through a graphical representation to show outcome trends before and after treatment intervention, that is, stay at home orders. The visualization helps better to gauge if the treatment and counterfactual group were following a parallel trend or not. The parallel trend assumption is one of the most critical assumptions to ensure internal validity of the DiD model. It requires that in the absence of treatment, the difference between the 'treatment' and 'control' group is constant over time. Although there is no statistical test for this assumption, visual inspection is useful when you have observations over many time points. Hence a graphical representation is the best way to inspect if there is a violation of this assumption.



Difference-in-Difference estimation, graphical explanation

Results obtained from Difference in Difference analysis:

If Hypothesis is True	If Hypothesis is False
Reject H0 and retain H1	Reject H1 and retain H0
The coefficient of our interaction term as obtained in the regression analysis is statistically significant.	The coefficient of our interaction term as obtained in the regression analysis is statistically insignificant.
	

8 Final Variables Required: Ibrahim

Our treatment variable will be the states where stay at home order policy intervention took place as of a fixed date (i.e. march 27, 2020), and our control variable are the states where stay at home policy intervention did not take place. The following are our treatment and control variables. In order to study the effectiveness of COVID-19 stay at home order, we will be performing pre-post and difference in difference analysis where states and the counterfactuals will be chosen based on the presence or absence of policy interventions, demographics such as population density, and other measures. On the next page we list the treatment and control states.

Please find below the list of treatment and control states:

Treatment State	Control States
Minnesota	Arkansas
Montana	Iowa
Nevada	Nebraska
North Carolina	North Dakota
Rhode Island	Oklahoma
Alabama	South Dakota
Arizona	Wyoming
Florida	
Georgia	
Kansas	
Maine	
Maryland	
Mississippi	
Missouri	
New Hampshire	
South Carolina	
Tennessee	
Texas	
Utah	
Virginia	

All of the COVID-19 data listed in Section 9 are at the county level. Therefore, for our comparison, our unit of observations will be at county level as well as state level. For instance, population density will be calculated both at county and state level using the population dataset and the land m[illegible]s dataset. On the other hand, the mobility report dataset will provide insights into what changed in response to those stay at home policies, such as movement trends over time by geography, across different categories of places like retail recreation and workplaces. The data is provided at the county level and provides a percent difference from the baseline, which was calculated before the first wave of COVID-19.

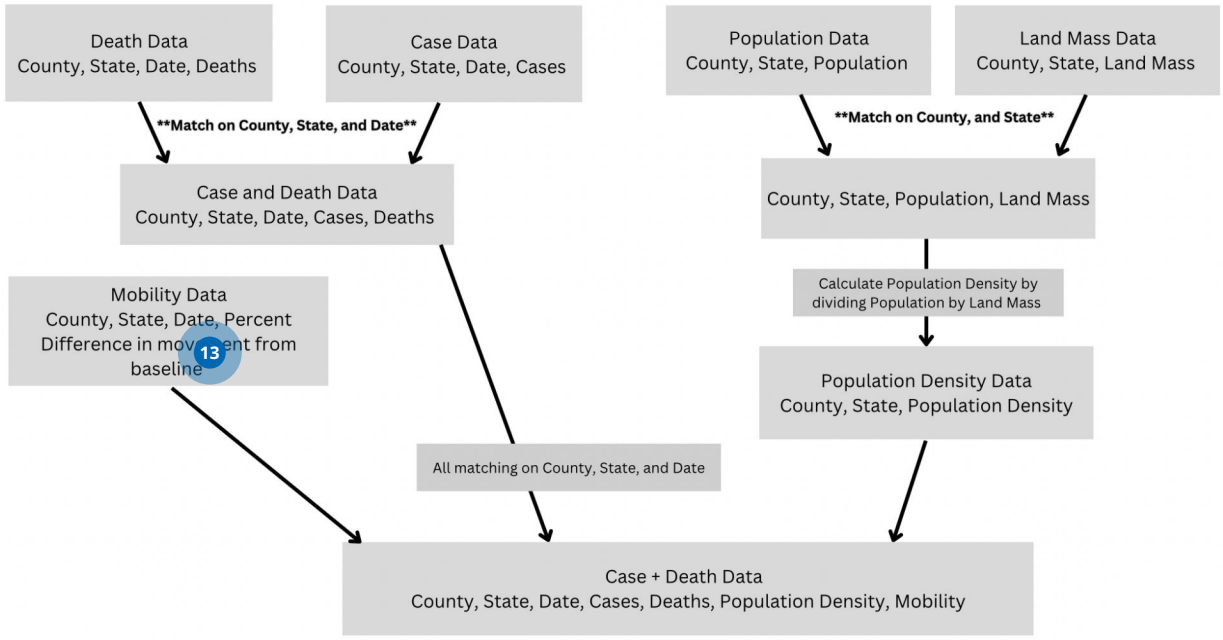
Our outcome variables here are the difference in confirmed cases day over day and difference in death rate day over day, also referred to as the COVID-19 positive rate or infections rate, and the difference in death rate day over day. We will be collecting positive case rate based on a 7 day period lagged data as we would expect that it takes about 7 days to fully identify and diagnose a COVID-19 case. In other words, we will be collecting a positivity rate about 7 days after the stay at home order policy is implemented.

Similarly, for death rates, we will be using a 14-15 day period as the expected since when patients die from the day they catch COVID-19. The data is provided at county level and will be aggregated on state level for comparative analysis. The difference in death rate and positivity rate will then be calculated by taking a change between the consecutive days/weeks.

9 Data Sources

Death: https://usafacts.org/visualizations/coronavirus-COVID-19-spread-map
Case: https://usafacts.org/visualizations/coronavirus-COVID-19-spread-map
Mobility: https://www.google.com/COVID-1919/mobility/data_documentation.html?hl=en
Land mass: https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html
Population: https://usafacts.org/visualizations/coronavirus-COVID-19-spread-map

Below is how we plan to merge the data.



Links for Citations

1. https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab
2. https://www.ncsl.org/health/state-action-on-coronavirus-COVID-19
3. https://ballotpedia.org/States_that_did_not_issue_stay-at-home_orders_in_response_to_the_coronavirus_(COVID-19)_pandemic,_2020
4. https://COVID-1919-interventions.com/#tab-6961
5. https://www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html