Introduction here

Introduction. Is the introduction clear? Is the research question specific and well defined? Does the introduction motivate a specific concept to be measured and explain how it will be operationalized. Does it do a good job of preparing the reader to understand the model specifications?

As of the writing of the document, the COVID-19 coronavirus (COVID-19) has been spreading throughout the United States for nearly 14-15 months, with the initial cases identified as having entered the country in January 2020. This report uses data from the United States Census Bureau (including state level demographics and county level population and population density data), the New York Times for COVID-19 case counts, a Google dataset on state-level mobility data, and related COVID-19 policy data from the US State Policy Database. All data used in the project was pulled in on April 10th, 2021.

Our research team has decided to create a descriptive model using OLS regression to measure the complex relationships that exist between state-level COVID-19 case counts (case counts) and changes in a states' population mobility, policy that was enacted during the pandemic, and __________. Specifically, our research question is the following: "What features from our mobility and policy databases, if any, are highly associated with changes in COVID-19 state-level case counts." This research question was initially motivated as an attempt to understand if the preventive measures that states have enacted in response to the pandemic were associated with statistically significant changes in case counts. Initially, we had hoped to find a causal relationship between our features, like policies targeted at reducing the virus and change in population mobility, however we ultimately decided against this because of the high potential reverse causality between key features like mobility and case count (i.e. a change in mobility causes a case counts which in turn will cause a change in mobility). Additionally, given how complex the nature of pandemics are, there was a strong possibility for omitted variable bias (such as ______).

Feature Selection for OLS Regression

For this study, we aggregated the different data sources across their individual study horizons (usually more than a year) to generate a small sample of 50 observations. Ultimately, we generated three linear models using OLS regressions of increasing complexity to explore the relations between our data's features and the case counts. The summary table below and the corresponding descriptions discuss the features we analyzed for our OLS regression descriptive models to understand their association with the dependent variable (i.e. case count per 100,000 people) included:

Feature	Present in which model	Source
Mobility (% change)	1, 2, 3	Google Dataset
Population Density (People/sq. mi.)	1, 2, 3	US Census Data
Percent of Population Living in a High Population Density County (%)	1, 2, 3	US Census Data
Percent of Population Under the Age of 24 (%)	2, 3	US Policy Database
Mask Mandate Days	2, 3	US Policy Database
Unemployment Benefit Days	3	US Policy Database
Increased Weekly Unemployment Insurance Amount Through July 31	3	US Policy Database
Business Close To Open Days	3	US Policy Database
Travel Quarantine Mandate Days	3	US Policy Database
Stay at Home Days	3	US Policy Database

`Case Count Per 100,000 people`: We chose the case count per 100,000 people (refered to as case counts in the rest of this research paper) as our outcome variable. We decided to normalize the case counts around the population as we were concerned that a OLS regression absolute case count would not result in any meaningful association discoveries. For example, the state of California has had some of the highest case counts despite it having some of the most aggressive policies to curb the spread of COVID-19, but this likely is due to the states massive population compared to other states with less aggressive COVID-19 policies.

`Mobility (% Change)`: We decided to regress on population mobility because physical proximity is a requirement for disease transmission and because mobility data captures the actual effects of multiple correlated policies intended to reduce COVID transmission (such as stay-at-home orders, quarantine requirements after traveling or possible exposure to a person who tested positive, business and school closures, etc.). This is effectively a method of dimensionality reduction that contributes to model parsimony. This dataset tracks the changes in mobility for the following sectors: (1) Grocery and Pharmacy, (2) Parks, (3) Residential, (4) Retail and Recreation, (5) Transit Stations, and (6) Workplaces. Ultimately, we decided to use the percent change in transit mobility in our models for two primary reasons (described in detail later). First is that most of these other features are highly correlated with transit. Second is that we couldn't determine a method for aggregating these features into a statewide change in mobility data as the raw data was captured in percent changes and not absolute changes. We theorized that changes in transit station mobility would be the most associated with changes in case counts from this list as transit stations are often identified as major spreaders of disease.

`Population Density (People/sq. mi.)`: Normalized the state's population by its area (given in square miles).

`Percent of Population Living in a High Population Density County (%)`: While normalizing a state's case count by the population is useful, we decided to add an additional feature illustrating what percentage of the population lives in a high density county, which we defined as the top 50 counties in the state when ranked by county population per square mile. An illustrative example of why we added this feature can be seen in the state of New York, which was one of the earliest states to be plagued with the virus. According to the New York Times COVID-19 tracking database, New York City alone has had 882K of New York's 1.94M (45%) total cases since the beginning of the pandemic (as of April 10th, 2021).

`Percent of Population Under the Age of 24 (%)`: The media and research studies alike have suggested that young adults and children are less likely to be impacted by the virus or may be asymptomatic compared with older adults. As a result, this young age group is more likely to continue spreading COVID-19. We hope to uncover a relationship between a state's percentage of the population with young people and a corresponding change in case counts.

`Mask Mandate`: the number of days that a state had a mask mandate.

Other model 3 policy features: We selected the following policy-related features from the COVID-19 US State Policy Database (www.tinyurl.com/statepolicies) that we thought related to COVID cases including, the date the state of emergency declared, length of mask mandate, length of stay-at-home orders, length of business closure, length of travel quarantine mandate, length of increased unemployment benefits, and increased unemployment insurance amount. All the policy-related features were recorded as dates, except for the increased unemployment insurance amount that was recorded as integers. Ultimately, what we were interested in was the total days a given policy was active for during the one year period after a state of emergency was announced for each state.

Initial Data Loading and Cleaning

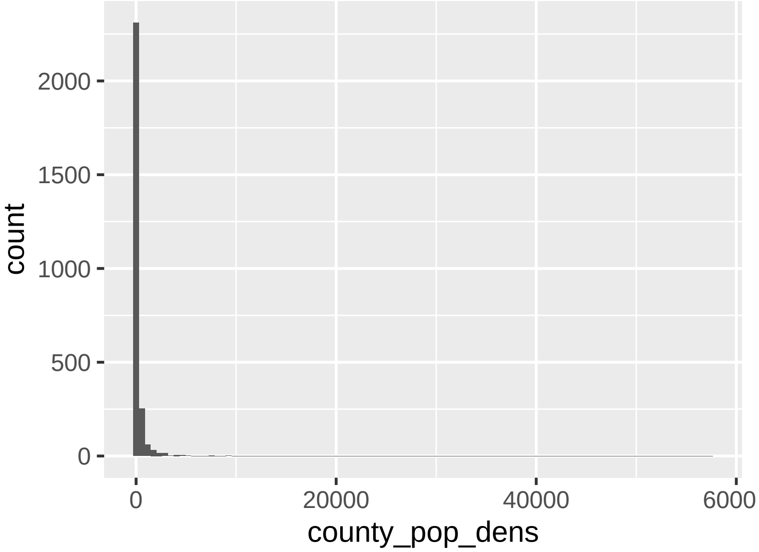
First, we read in the NYT Covid database from an excel workbook (see the data folder in our repository for all of the raw data files used in this analysis).

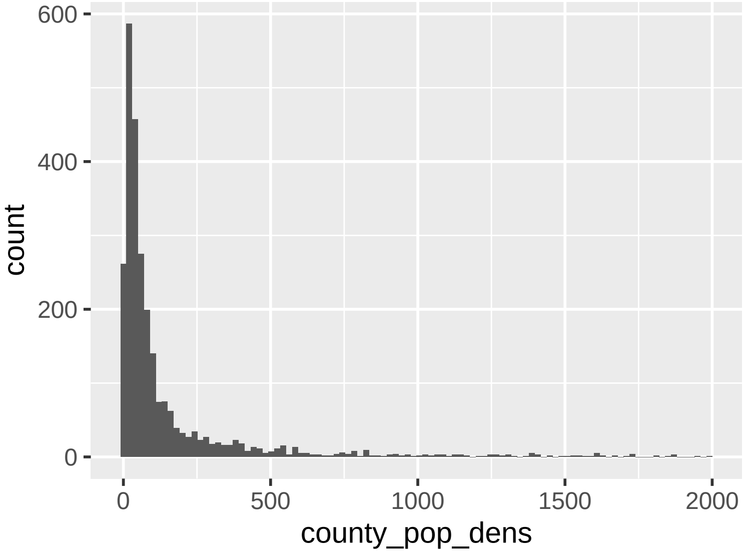
Then, we read in the policy data from an excel workbook downloaded from the US State Policy Database.

Then, we read in the Google dataset (from an excel workbook) on state-level changes in mobility data.

Next, we read in three spreadsheets available for download from the U.S. Census Bureau to obtain county-level demographics, including population and area.

Now that we have read in all of the raw datasets we will use in this study, we joined the mobility data with the county-level population and area data.





From the histograms above, we observe that the population densities are heavily skewed towards the left of the distribution. We then applied a filter to show what the distribution would resemble if we filter out the top 50 counties (which we define as the high population density counties). We then determine what percentage of a state's population lives in a high population county, with the majority of states having zero percent (i.e. the top 50 counties are found in only 19 states).

With the county level population observations and mobility observations, we then calculated the weighted average change in mobility at the state level.

Next, we join the state-level mobility observations with the added population and percentage of population in a high density area features to the NYT Covid database by state and date.

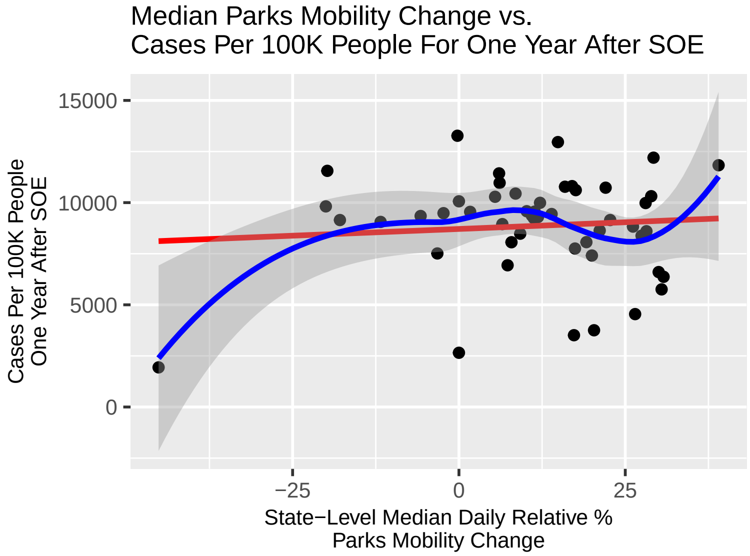
Exploratory Data Analysis

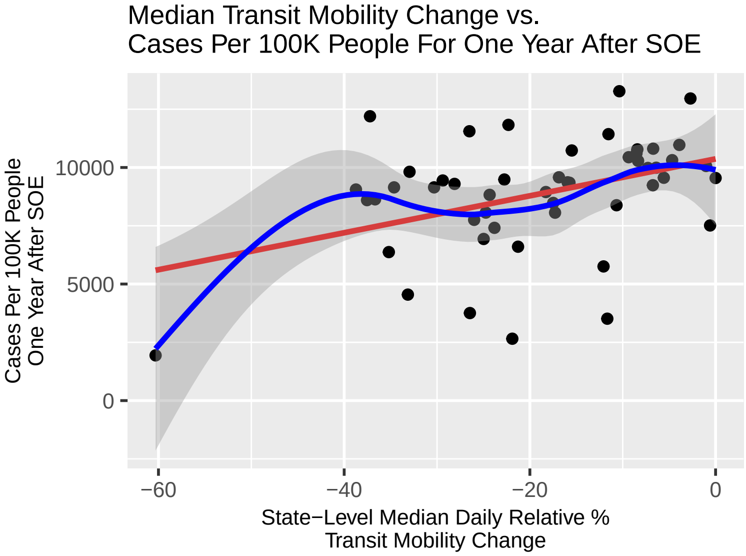
For our analysis, we aggregated the year-plus worth of COVID-19 data into a single metric per state for a total data frame size of 50 observations. We decided to aggregate one-years worth of observations after a state declared a state of emergency, which we are using as a proxy to indicate the “start date” of a state’s serious attempts to curb the spread of the virus. We ultimately decided upon one year’s worth of observations to aggregate rather than a shorter time period as we wanted to capture a significant amount of observations for which the temporal impacts would be lessoned. For example, we initially aggregated the first 90 days worth of observations after the state of emergency, however this approach actually resulted in an inverse linear relationship (i.e. negative β coefficient in our base model) between change in mobility and change in case counts when we expected a positive relationship. We concluded that 90 days was an insufficient time horizon for most states to determine the appropriate relationship between mobility and case count as the immediate time period following a state of emergency might still see an increase in case counts for a few weeks given the time-dependence of case counts with the previous case counts.

Next, we inspected the correlation between the different measures of mobility changes, we chose to use the state-level median change in transit mobility in the 365 days after each state declared an emergency as our main variable of interest for model 1.

```
## Correlation Matrix: Median Mobility Changes by Category

##             Transit Retail Grocery Parks Workplace Residential
## Transit        1.00   0.91    0.84  0.20      0.84       -0.89
## Retail         0.91   1.00    0.87  0.28      0.83       -0.86
## Grocery        0.84   0.87    1.00  0.46      0.68       -0.73
## Parks          0.20   0.28    0.46  1.00     -0.09        0.04
## Workplace      0.84   0.83    0.68 -0.09      1.00       -0.95
## Residential   -0.89  -0.86   -0.73  0.04     -0.95        1.00
```





After inspection of the correlation matrix between the different measures of mobility changes, we chose to use the state-level median change in transit mobility in the 365 days after each state declared an emergency as our main variable of interest for Model 1. Our motivation for choosing transit over the other mobility features was as follows:

1) Transit is most closely aligned with our understanding of how viruses spread, particularly from one locality or population center to another.
2) The mobility changes in Transit, Retail, Grocery, and Workplace are highly positively correlated with each other (>0.8 in each pair), and highly negatively correlated with Residential mobility changes (absolute value of >0.7 or above). Highly correlated features should be avoided in descriptive and explanatory linear regression modeling as they tend to increase the standard error estimates on the model parameter estimates for the correlated features.

The only mobility metric that was not strongly correlated with the others was Parks. Since the median change in Park mobility was not strongly correlated with any other mobility changes, we examined its relationship with our target variable but did not observe a positive or negative linear relationship. A t-Test on our calculated simple regression coefficient failed to reject the null hypothesis that there was no evidence that the coefficient for change in median Parks mobility was measurably different than zero.

3) We did not want to take an aggregation of the set of highly correlated features (Transit, Retail, Grocery, and Workplace) because the raw data provided are relative numbers and we do not have access to the underlying absolute mobility data, so we would be unable to derive a correct weighted average of these features.
4) Given that the distribution of relative transit mobility change had a left skew, our team decided to use the median value for each state within the 365 day window as a better measure of central tenancy.

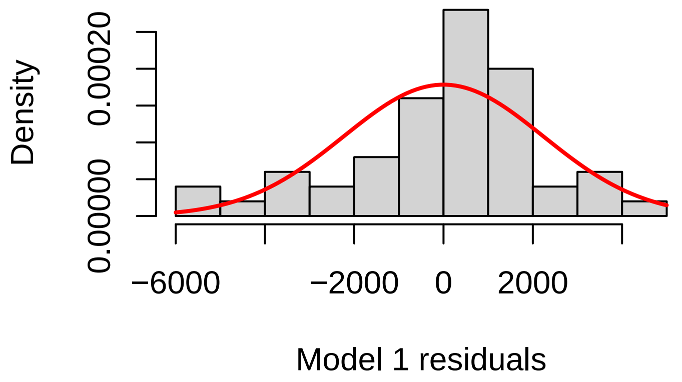
Base Model

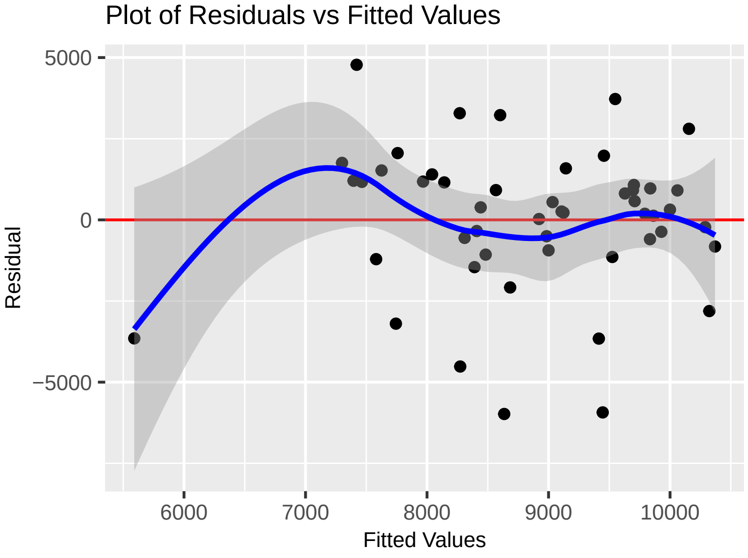
```
##
## -------Model results-------
```

```
##
```

```
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change,
##     data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5981.0  -909.3   286.3  1177.2  4778.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10370.16     580.47  17.865  < 2e-16 ***
## median_transit_change    79.25      25.43   3.117  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2259 on 48 degrees of freedom
## Multiple R-squared:  0.1683, Adjusted R-squared:  0.151
## F-statistic: 9.714 on 1 and 48 DF,  p-value: 0.003086
```

Histogram of Model 1 Residuals vs Normal





```
##
```

```
## -------Homoskedasticity Test-------
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_1_final
## BP = 1.5589, df = 1, p-value = 0.2118
```

```
##
## ------Normality of Residuals Test-------
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_1_final$residuals
## W = 0.94719, p-value = 0.02617
```

Model 1 Discussion Here:

```
save.image(file = "pwd/model_1_workspace.RData")
```