Model 2 & 3 clean_script

Jun Qian, Lucas Schroyer, Ryan Mitchell, Oliver Chang

4/8/2021

Introduction here

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 team's primary research question was "How does mobility impact the spread of COVID-19?" To begin to answer that question, our research team decided to conduct an exploratory observational analysis using OLS regression to measure the complex relationships that exist between changes in a states' population mobility and state-level COVID-19 case counts per 100K people in the 365 days following each state's declaration of a state of emergency (SOE). As part of this endeavor, we also explore whether other variables, such as state age demographics, population density, and state level policies on mask mandates, stay at home orders, quarantine restrictions, enhanced unemployment benefits, and business closures might also have an impact on COVID 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 mobility and COVID case counts. However, we ultimately decided against this because of the high likelihood of reverse causality between these variables (i.e. a change in mobility causes a change in 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 state level differences in temperature and humidity, behavioral differences in terms of mask compliance, COVID testing availability, the percentage and absolute numbers of people using mass transit, and the percentage and absolute numbers of people who are able to work from home, to name a few). These omitted variables will be discussed in the model limitations section of this report. We decided to focus on COVID cases as our dependent variable, as opposed to COVID deaths, because we inferred that COVID deaths counts are directly dependent on the number of COVID cases and other causal inputs, such as genetic predisposition, underlying health conditions/comorbidities, hospital utilization rates, the availability of various treatment options (ventilators, experimental drugs, etc.). Given COVID cases are a direct input to COVID deaths and much of the data needed to analyze these supplemental variables is not easily available or likely does not exist, a study on COVID death counts may suffer from substantial omitted variable bias. This would raise questions about the interpretation of our model in terms of both statistical and practical significance. Our motivation for normalizing the dependent variable per 100K people 365 days after each state's declaration of a state of emergency was as follows:

Absolute population counts vary substantially from state to state, and population being the most important factor in absolute COVID case counts is not an interesting finding. One year was a natural ending point given that we recently passed the one year mark for each state's declaration of a state of emergency. Additionally, we believe that the mere act of the state declaring an emergency may have contributed to shifts in the behaviors of the residents of each state, so we wanted to index our analysis against that moment in time. We decided to regress on population mobility as our key variable of interest because: Mobility fits well into

a causal regression framework for COVID cases, because physical proximity is a requirement for disease transmission, and proximity is a function of mobility (and population density, which we will also explore in our modeling efforts). The 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.

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

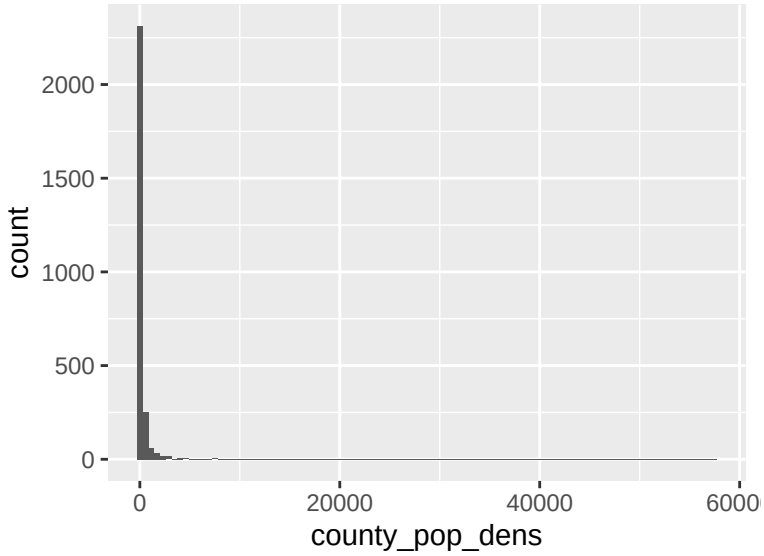
First, we read in the NYT Covid database from an excel workbook (see the data folder in our repository for all of the raw data files used in this analysis).

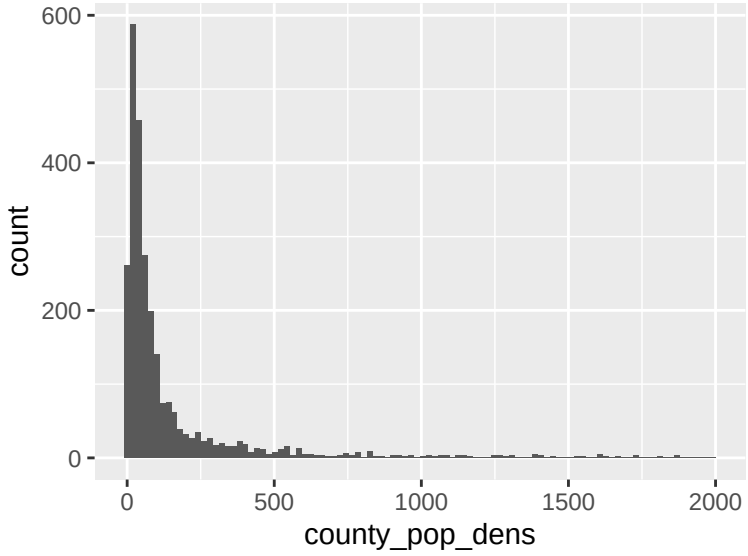
Then, we read in the policy data from an excel workbook downloaded from the US State Policy Database.

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

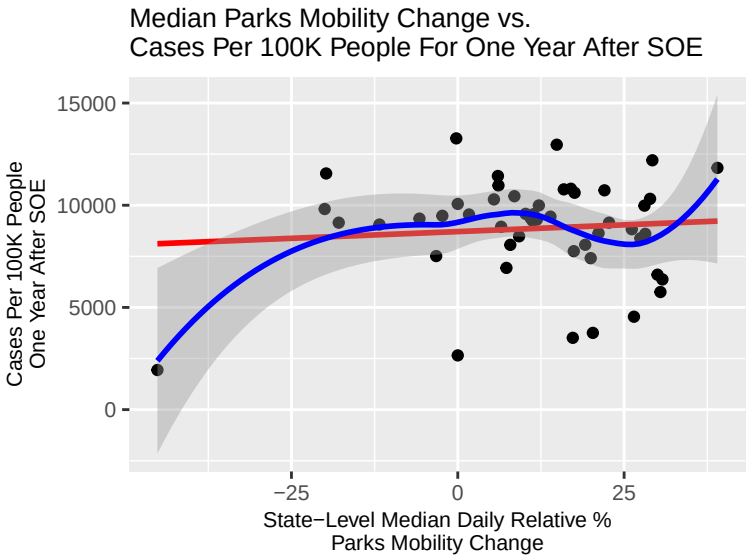
For our analysis, we aggregated the year-plus worth of COVID-19 data into a single metric per state for a total data frame size of 50 observations. We decided to aggregate one-years worth of observations after a

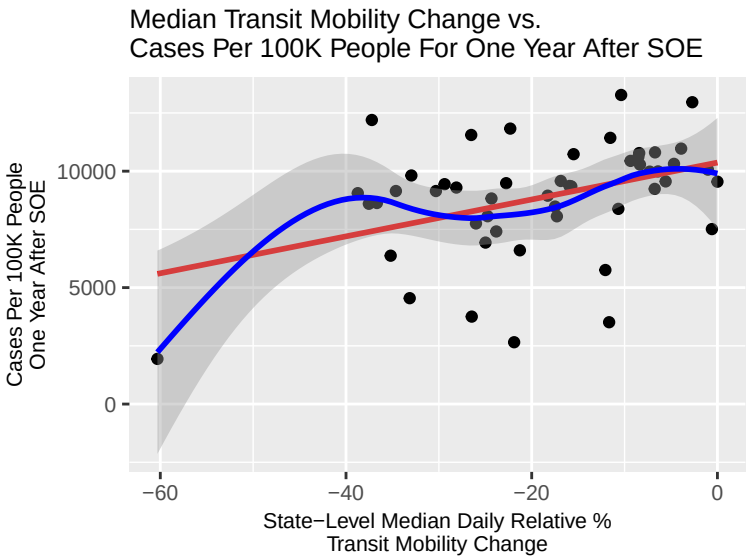
state declared a state of emergency, which we are using as a proxy to indicate the "start date" of a state's serious attempts to curb the spread of the virus. We ultimately decided upon one year's worth of observations to aggregate rather than a shorter time period as we wanted to capture a significant 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.

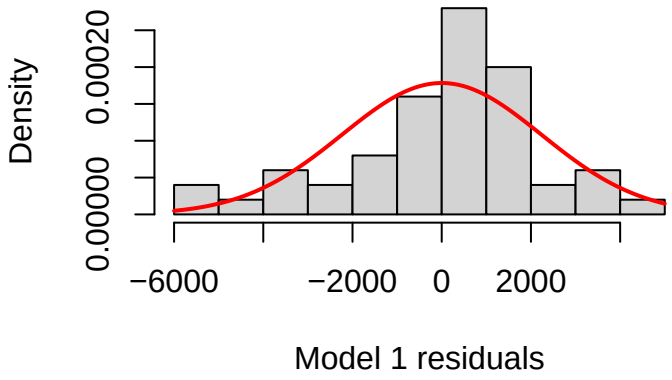
4) Given that the distribution of relative transit mobility change had a left skew, our team decided to use the median value for each state within the 365 day window as a better measure of central tenancy.

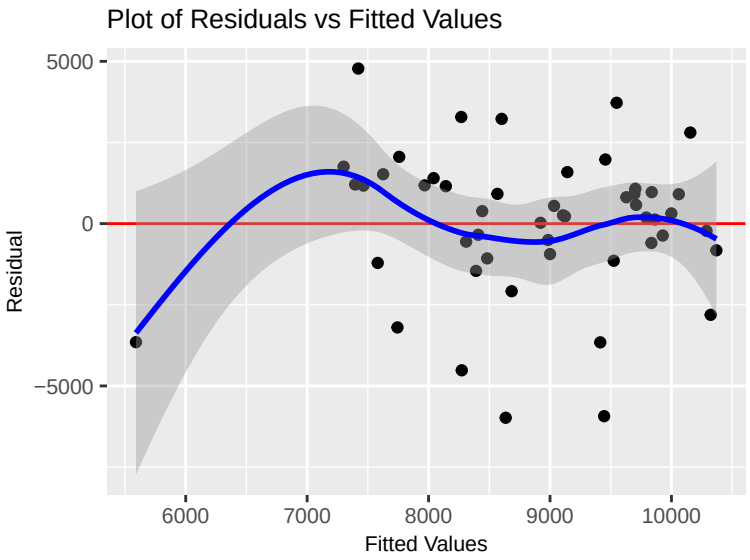
Base Model

```
##
## -------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
```

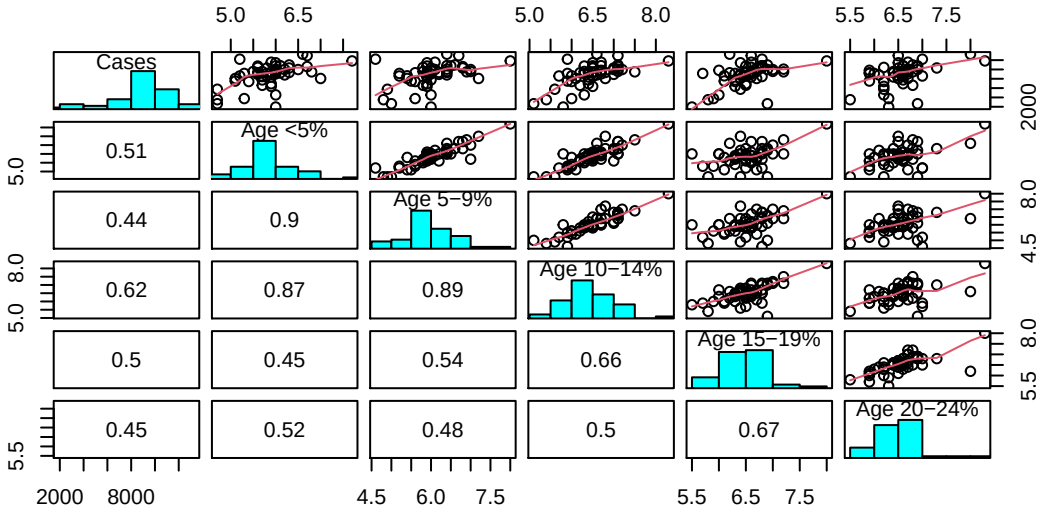
Model 1 Discussion Here:

Second Model.

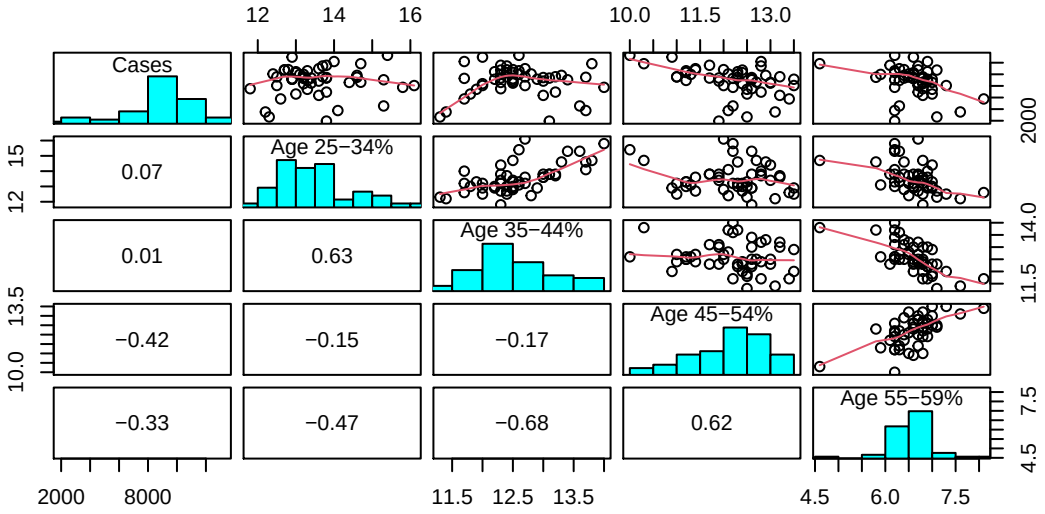
As we thought about an approach to building our Model 2 specifications, the general approach that our group decided to take was to incrementally test the addition of new features to our Model 1 specification in descending order of expected importance, according to our general understanding of how viruses spread. At each iteration, we began by examining the relationship (using scatterplots) between a new demographic or policy-related feature and our outcome of interest: COVID cases per 100K people 365 days after each state declared a state of emergency. If we observed a relationship, we ran a combination of t-Tests and/or ANOVA F-Tests to determine whether or not to add the feature to our Model 1 specification. We proceeded in this 'greedy' algorithmic fashion by adding variables to our Model 1 specification until we could no longer justify further additions based on results from ANOVA F-Tests. The first incremental feature we tested was derived from observational data from the census bureau on population age distributions by state.

One of the contributing factors to the spread of COVID-19 is asymptomatic spread. Younger people have been shown to have milder symptoms and therefore it stands to reason that they may be less likely to get tested, and ultimately end up spreading the disease at a greater rate than older age groups. To make matters worse, younger people tend to interact with more people (source?) as a result of being in school (switching between classrooms, touching desks and other surfaces where other students have been, congregating in large cafeterias, etc.), which increases the probability of viral transmission. Hence, for our second model, we were interested in testing our general hypothesis that age demographics may play a role in the spread of COVID-19. We began by doing some basic exploratory data analysis (EDA) with respect to age distributions and cumulative COVID-19 case counts per 100K, which is our target variable for Lab 2.

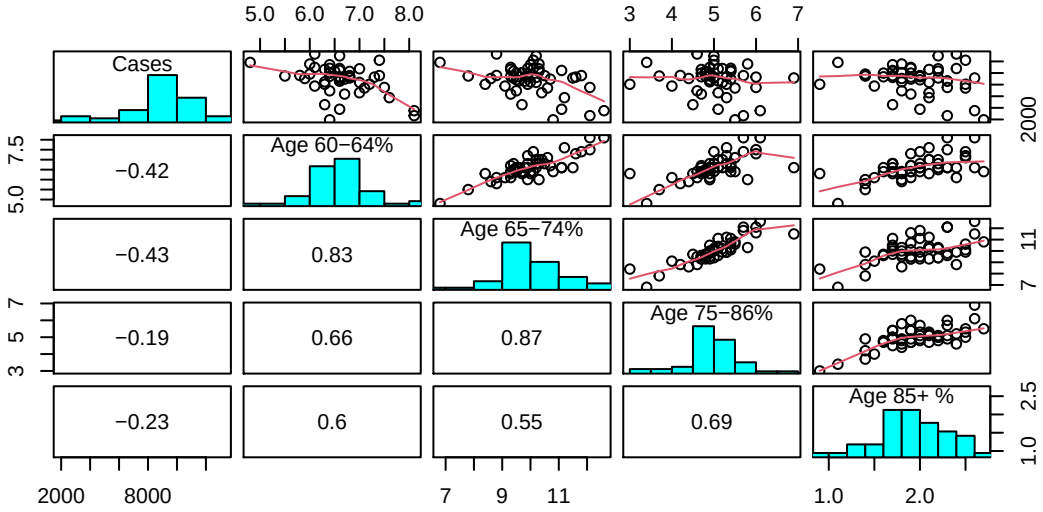
Case vs Age Bins Distribution and Scatter Plots



Case vs Age (Contin.) Bins Distribution and Scatter Plots



Case vs Age (Contin.) Bins Distribution and Scatter Plots




```
#Significant
anova(model_a, model_1_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 47 184378849
## 2 48 245004785 -1 -60625936 15.454 0.0002765 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
#Insignificant
anova(model_b, model_a, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 46 170183297
## 2 47 184378849 -1 -14195553 3.837 0.05621 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```r
#Significant F-test
anova(model_c, model_a, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 45 132785225
## 2 47 184378849 -2 -51593624 8.7424 0.00062 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```r
#Insignificant relative to model_c
anova(model_d, model_c, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 44 132369352
## 2 45 132785225 -1 -415873 0.1382 0.7118
```

```r
anova(model_e, model_d, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years +
Percent.Sex.And.Age.Total.Population.20.To.24.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
```

```
## 1 43 124070515
## 2 44 132369352 -1 -8298836 2.8762 0.09713 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

```
anova(model_e, model_c, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years +
## Percent.Sex.And.Age.Total.Population.15.To.19.Years +
Percent.Sex.And.Age.Total.Population.20.To.24.Years
## Model 2: cases_per_100k_at_365d ~ median_transit_change
+ Percent.Sex.And.Age.Total.Population.Under.5.Years +
## Percent.Sex.And.Age.Total.Population.5.To.9.Years +
Percent.Sex.And.Age.Total.Population.10.To.14.Years
## Res.Df RSS Df Sum of Sq F Pr(>F)
## 1 43 124070515
## 2 45 132785225 -2 -8714710 1.5102 0.2324
```

When performing EDA on the categorical age group distributions and their relationship to our target variable, we noticed a strong positive correlation to the target among age groups 0-24. Age groups in the 24+ range did not appear to follow a consistent pattern with respect to correlation with our target variable and were therefore not a key focus area for our analysis.

Next, we performed a series of ANOVA F-tests on a nested set of linear regression models with parameter estimates for this set of age groups (Under 5, 5-9, 10-14, 15-19, and 20-24) as well as our main variable of interest (median transit mobility change). Our motivation here was to understand whether the regression residuals were measurably different from one another between the different (nested) model specifications. For context, the null hypothesis for an F-test is that fitting additional coefficients for a longer model does not measurably reduce the residuals relative to a nested model with fewer parameters.

Our first ANOVA F-test compared our Model 1 with a new model that had an additional parameter estimate for the percentage of the population under 5 years old. With a p-value of 0.0003, we rejected the null hypothesis and used this new model (model_a) as the baseline model for subsequent comparisons with additional parameter estimates for different age categories.

We failed to reject the null hypothesis when adding an estimator for ages 5-9 (model_b), and succeeded in rejecting the null hypothesis when adding an estimator for ages 10-14 (model_c), with a p-value of 0.001. Adding additional estimators for 15-19 and 20-24 failed to reject the null hypothesis that these models (model_d and model_e) were measurably better at reducing residuals than model_c. These results suggested creating two new features for the percentage of the population ages 0-9 and 10-24 and adding them to our base Model 1 to create the first specification for our Model 2.

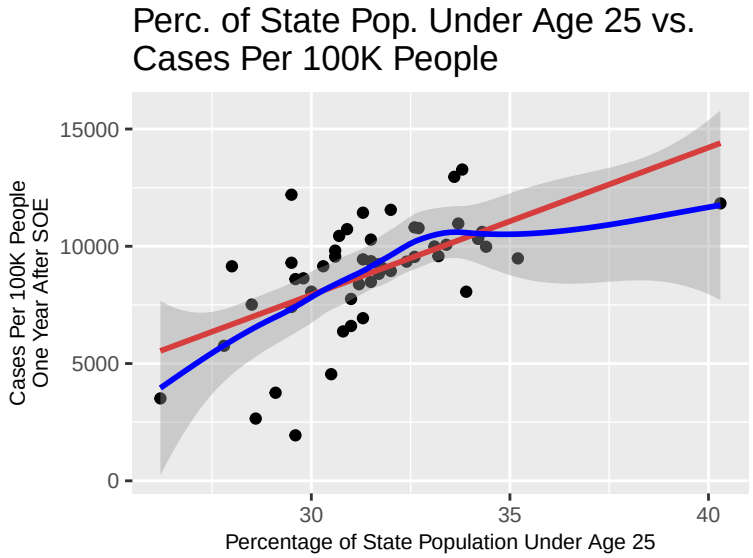
```
final_df <- final_df %>%
  mutate(pop_pct_age_0_9 = Percent.Sex.And.Age.Total.Population.Under.5.Years +
           Percent.Sex.And.Age.Total.Population.5.To.9.Years,
                    pop_pct_age_10_24 =
           Percent.Sex.And.Age.Total.Population.10.To.14.Years +
           Percent.Sex.And.Age.Total.Population.15.To.19.Years +
```

```
          Percent.Sex.And.Age.Total.Population.20.To.24.Years)

cor(final_df$pop_pct_age_0_9, final_df$pop_pct_age_10_24)
```

```
## [1] 0.7701834
```

```
final_df <- final_df %>%
  mutate(pop_pct_age_0_24 = Percent.Sex.And.Age.Total.Population.Under.5.Years +
           Percent.Sex.And.Age.Total.Population.5.To.9.Years +
           Percent.Sex.And.Age.Total.Population.10.To.14.Years +
           Percent.Sex.And.Age.Total.Population.15.To.19.Years +
           Percent.Sex.And.Age.Total.Population.20.To.24.Years)
```

However, after measuring a 0.77 correlation between these two features - and with the goal of increased model parsimony - we decided to group them together to prevent the standard errors for their respective coefficients from increasing substantially. Let's take a look and assess whether this new variable for percentage of the population < 24 years old visually satisfies the conditional linearity expectation with respect to our target.



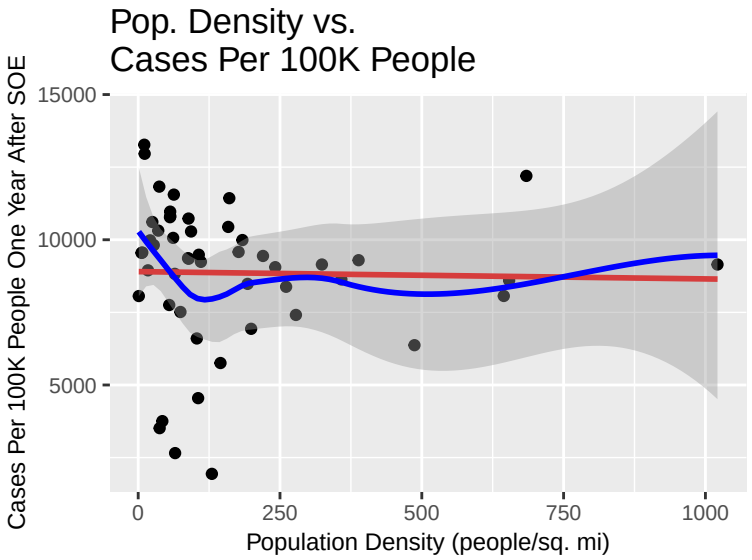
The conditional linear expectation between our population percentage aged 0-24 and our target appears to be met. In addition to evaluating the linear relationships between the demographic age variables and our target, we also evaluated the log, square root, and square transformations to the age distribution data but did not find them to aid in reducing the frequency or magnitude of outlier data points.

Indeed, including a feature for the percentage of the population aged < 24 years old appears to have improved our model's performance. An ANOVA F-test returned a p-value of 0.0001, suggesting that we reject the null hypothesis that the (interim) Model 2's residuals were not measurably different from the residuals of Model 1.

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24
## Model 2: cases_per_100k_at_365d ~ median_transit_change
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1     47 176654365
## 2     48 245004785 -1 -68350419 18.185 9.595e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After accounting for changes in mobility and demographic age differences between states, the next variable we wanted to explore as part of our descriptive model for COVID case counts was population density. According to the World Health Organization, "COVID-19 virus is primarily transmitted between people through respiratory droplets and contact routes." (https://www.who.int/news-room/commentaries/detail/modes-of-transmission-of-virus-causing-covid-19-implications-for-ipc-precaution-recommendations). In other words, COVID spreads primarily through physical interactions between infected and uninfected hosts, regardless of whether the actual mechanism of transmission is airborne or surface based contact. Hence, it stands to reason that more densely populated areas would see greater rates of infections, because the frequency of these physical interactions will increase with population density. This was the motivation for our group exploring whether a relationship existed between state population density and our outcome variable of interest.



To our surprise, however, there was no clear relationship between population density and our outcome variable. A t-Test on the parameter estimate for population density's relationship with our target variable failed to reject the null hypothesis that the coefficient value was not measurably different from zero.

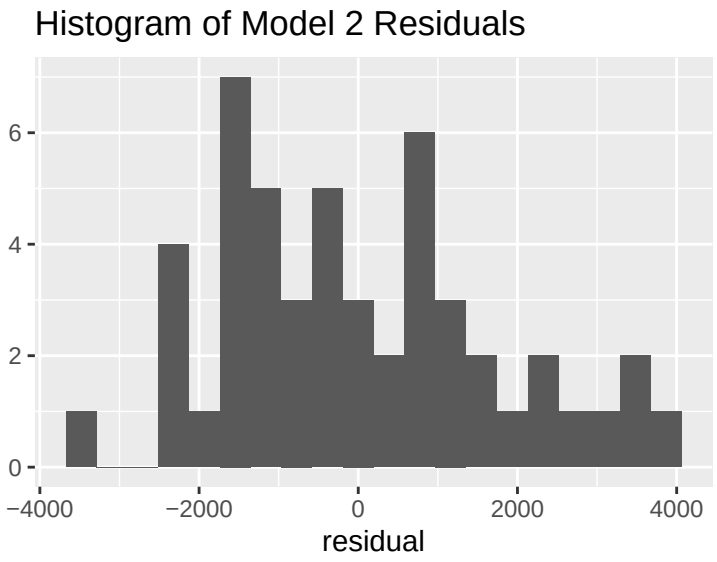
```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ population_density, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6930.5  -799.2   487.9  1416.9  4372.2
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8902.5248   454.6106  19.583   <2e-16 ***
## population_density   -0.2508     1.6991  -0.148    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

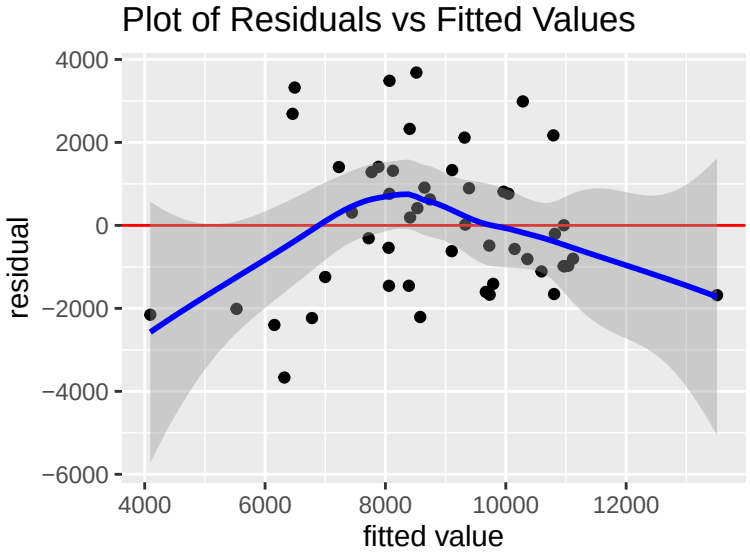
```
##
## Residual standard error: 2477 on 48 degrees of freedom
## Multiple R-squared:  0.0004535,  Adjusted R-squared:  -0.02037
## F-statistic: 0.02178 on 1 and 48 DF,  p-value: 0.8833
```

In spite of this test, our group decided to move forward and try including the population density feature in Model 2, as we believed it to be a conceptually meaningful variable in describing the population prevalence of COVID in each state one year after each state declared a state of emergency. Therefore, we proceeded with an ANOVA F-Test to test whether the incremental population density feature measurably improved model performance (via reduction of residuals) relative to our current model with features for median transit mobility change and percentage of the population < 24 years old. This ANOVA F-test returned a p-value of 0.002, enough to reject the null hypothesis that the model residuals were not measurably different from one another.

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

##
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change +
##     pop_pct_age_0_24 + population_density, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3666.7 -1367.9  -257.5  1191.5  3684.0
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -10160.586   3917.035  -2.594  0.01268 *
## median_transit_change     82.223     23.514   3.497  0.00106 **
## pop_pct_age_0_24         628.592    119.313   5.268 3.55e-06 ***
## population_density         4.665      1.446   3.225  0.00232 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1770 on 46 degrees of freedom
## Multiple R-squared:  0.5109, Adjusted R-squared:  0.479
## F-statistic: 16.02 on 3 and 46 DF,  p-value: 2.875e-07

##
## -------ANOVA F-Test Significance Test Relative To Model 1-------

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
## Model 2: cases_per_100k_at_365d ~ median_transit_change
##   Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
## 1     46 144077125
## 2     48 245004785 -2 -100927659 16.112 4.974e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## -------ANOVA F-Test Significance Test Relative To Interim Model 2-------

## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24
##   Res.Df       RSS Df Sum of Sq      F  Pr(>F)
## 1     46 144077125
## 2     47 176654365 -1 -32577240 10.401 0.00232 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

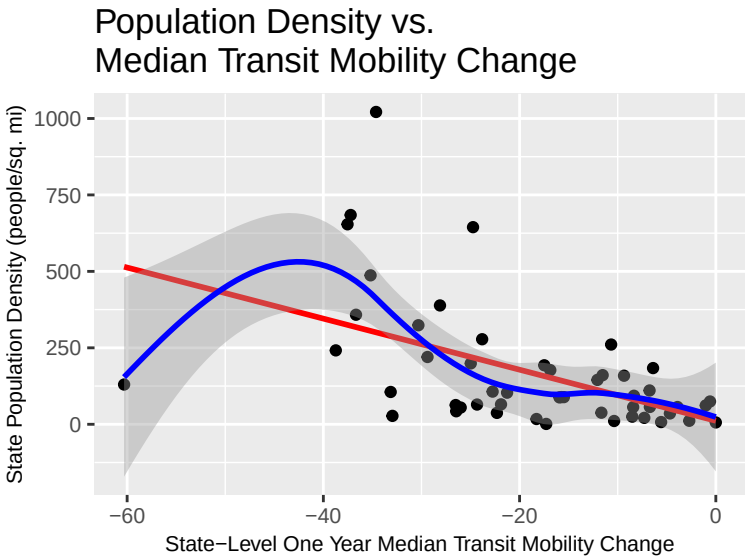


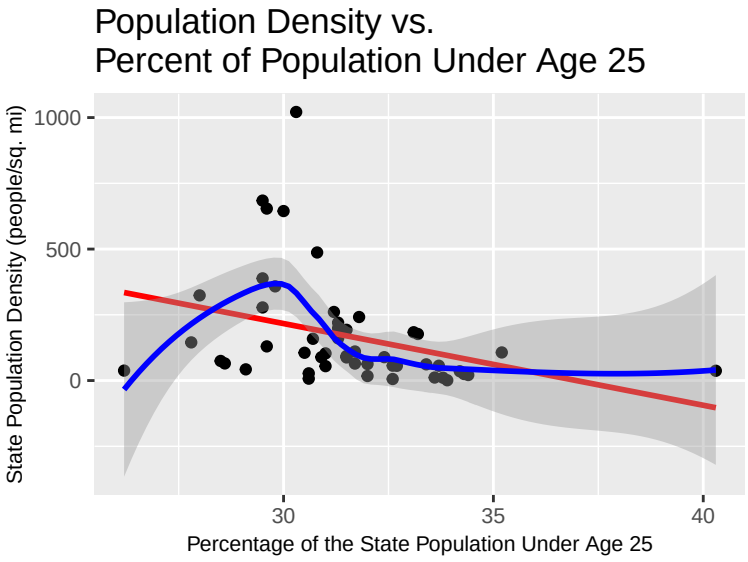



```
##
## -------Homoskedasticity Test-------
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2_final
## BP = 8.3386, df = 3, p-value = 0.03951

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

##
##  Shapiro-Wilk normality test
##
## data:  model_2_final$residuals
## W = 0.97275, p-value = 0.2984
```

Because the population density feature only becomes statistically significant at the p = 0.05 level in our model when we include features for median change in transit mobility and the percentage of the population under age 25, we say that the population density has a conditional relationship with our outcome of interest. Both of these two co-variates (median transit mobility change and percentage of the population under age 25) are negatively correlated with population density and positively correlated with our outcome variable.





However, we are only interested in the unique variation of population density with respect to our outcome variable. When the OLS regression algorithm calculates the parameter estimate for population density, it starts by regressing population density on the other model co-variate (input) features. The residuals from that regression represent the portion of population density that is *not* colinear with median transit mobility change and percentage of the population under age 25. Then OLS regresses our target values on those residuals to derive an estimate for the population density parameter. The model summary tells us that if we hold the percentage of the population under 25 and the median transit mobility change for a state constant (we do not allow them to co-vary), that there exists a positive correlation between population density and our outcome variable at a statistically significant level ($p = 0.002$) using classical standard errors.

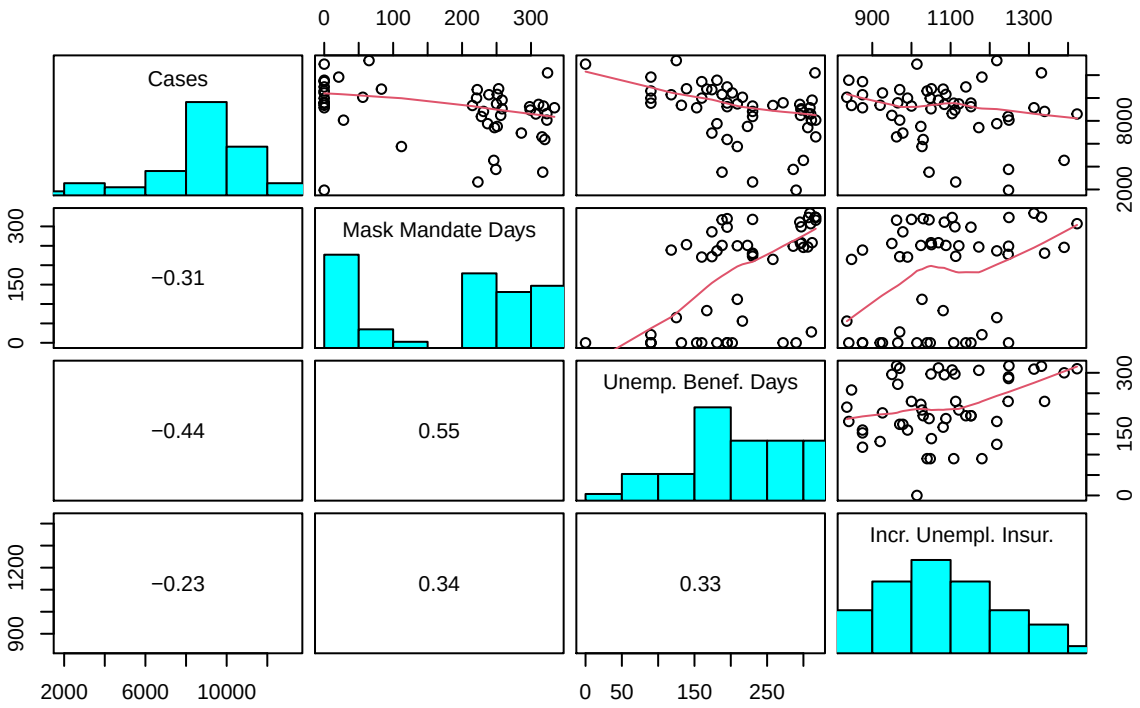
At this stage, we wanted to explore whether any state policy changes aimed at reducing the spread of COVID-19 added incremental descriptive power beyond our current Model 2 specification. In particular, we wanted to examine whether the timing of mask mandates, length and amount of increased government assistance via enhanced unemployment benefits, business closures, stay at home mandates, and travel quarantine restrictions had measurable effects on our outcome of interest after accounting for the features already in our Model 2 specification (which included state-level features for median change in transit mobility, percentage of the population under 25 years of age, and population density). All the policy-related features were recorded as dates, except for the increased unemployment insurance amount, which was recorded as an integer.

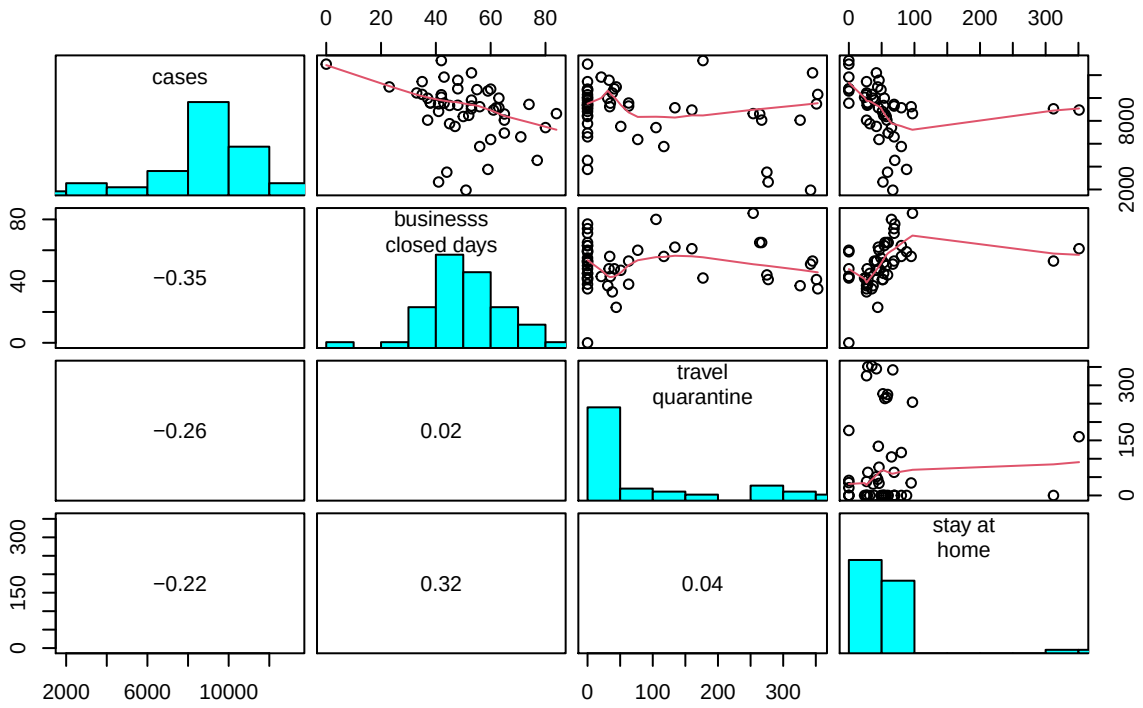
To align with our target variable of COVID-19 Cases per 100K one year after state of emergency declaration, we encoded the date related policy variables as the total number of days each policy was in place for after the state of emergency was announced for each state (up to 365 days). For transparency:

1. If a policy had no beginning and end dates, the total days were assigned as zero.
2. If a policy had beginning but not end dates, the total days were calculated by the date of state emergency declared + 364 days - the date of the beginning of the policy.
3. If a policy had both beginning and end dates, the total days were calculated by the difference in days of the two dates.

Once the policy-related features of our interest were transformed into days in force, we conducted EDA on the policy-features and COVID cases per 100k population using scatter plot and correlation matrices.

Policy Days vs Cases Per Capita Distribution and Scatter Plots





From the scatterplots, we see varying degrees of linearity between the policy features and our outcome of interest. We tested each of these features iteratively in the same manner as before, using significance from ANOVA F-tests as the benchmark to decide whether or not to include incremental policy features as part of our final Model 2. Ultimately, none of these policy variables (length of mask mandates, length and amount of increased government assistance via enhanced unemployment benefits, length business closures, length of stay at home mandates, and length of travel quarantine restrictions) returned a p-value that would allow us to reject the null hypothesis that the model residuals had not measurably improved.

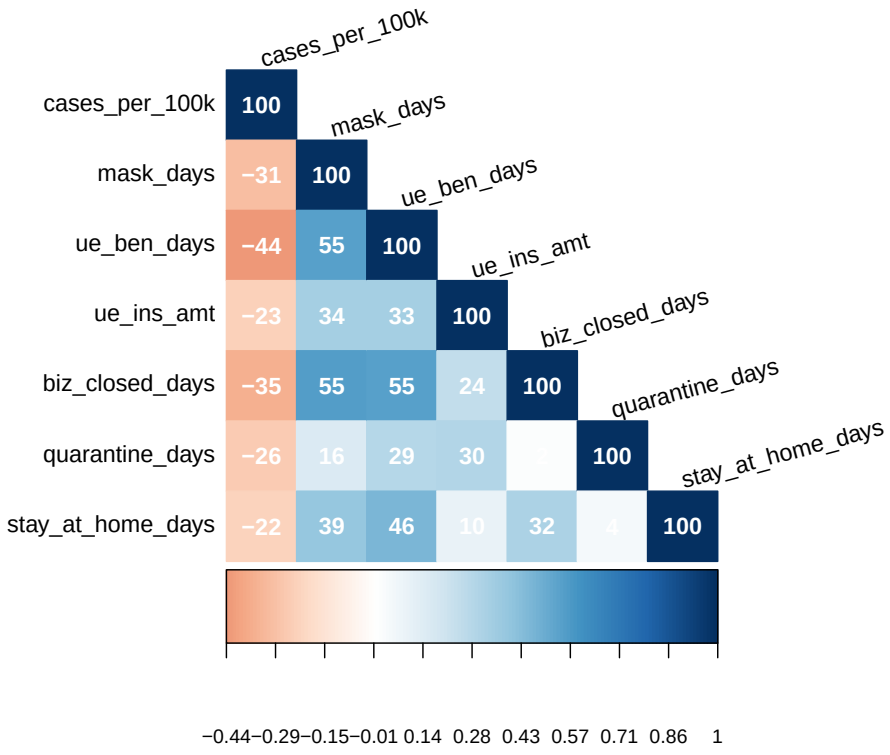
Still, it is worth noting that all of the aforementioned policy features, which were declared to mitigate COVID spread, demonstrated a negative correlation (from approximately -0.5 to -0.2) with our outcome variable of COVID cases per 100k people. The fact that these policy features failed to reject the null hypothesis in the ANOVA F-Test relative to the features already in our Model 2 was not entirely unexpected. Conceptually, several of the features share a significant amount information with median transit mobility change. One could make the argument that business closures, quarantine mandates, and stay at home mandates are all captured, to some extent, in the transit mobility change.

Ryan fleshing out writing here

Unemployment insurance extensions and benefit increases

We were surprised, however, that mask mandates

Correlation Matrix: State Policies vs. Cases



Jun's policy tests (all failed) for Model 2 here (to be reviewed...):

```
#Testing mask mandate days: Failed F-Test
model_2_final_incremental_test1 <- lm(cases_per_100k_at_365d ~
                                        median_transit_change +
                                        pop_pct_age_0_24 + population_density +
                                        mask_mandate_days, data = final_df)
anova(model_2_final_incremental_test1, model_2_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + mask_mandate_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 138865334
## 2     46 144077125 -1  -5211792 1.6889 0.2004
```

```
#Testing extended unemployment benefit days: Failed F-Test
model_2_final_incremental_test2 <- lm(cases_per_100k_at_365d ~
                                        median_transit_change +
                                        pop_pct_age_0_24 + population_density +
                                        unemployment_benefits_days,
                                      data = final_df)
anova(model_2_final_incremental_test2, model_2_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + unemployment_benefits_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 137886393
## 2     46 144077125 -1  -6190732 2.0204 0.1621
```

```
#Testing increased unemployment insurance amount: Failed F-Test
model_2_final_incremental_test3 <- lm(cases_per_100k_at_365d ~
                                        median_transit_change +
                                        pop_pct_age_0_24 + population_density + increased_weekly_unemplo
anova(model_2_final_incremental_test3, model_2_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + increased_weekly_unemployment_insurance_amt_thru_jul31
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 135803753
## 2     46 144077125 -1  -8273373 2.7415 0.1047
```

```r
#Testing mask mandate days: Failed F-Test
model_2_final_incremental_test4 <- lm(cases_per_100k_at_365d ~
                                        median_transit_change +
                                        pop_pct_age_0_24 + population_density +
                                        business_closed_days_round1,
                                      data = final_df)
anova(model_2_final_incremental_test4, model_2_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + business_closed_days_round1
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 140581017
## 2     46 144077125 -1  -3496108 1.1191 0.2958
```

```r
#Testing quarantine mandate days: Failed F-Test
model_2_final_incremental_test5 <- lm(cases_per_100k_at_365d ~
                                        median_transit_change +
                                        pop_pct_age_0_24 + population_density + travel_quarantine_manda
anova(model_2_final_incremental_test5, model_2_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + travel_quarantine_mandate_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 140733564
## 2     46 144077125 -1  -3343562 1.0691 0.3067
```

```r
#Testing stay at home mandate days: Failed F-Test
model_2_final_incremental_test6 <- lm(cases_per_100k_at_365d ~
                                        median_transit_change +
                                        pop_pct_age_0_24 + population_density +
                                        stay_at_home_days, data = final_df)
anova(model_2_final_incremental_test6, model_2_final, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + stay_at_home_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     45 144017384
## 2     46 144077125 -1    -59741 0.0187 0.8919
```

```
#model_base=cases_per_100k_at_365d ~ median_transit_change +
#pop_pct_age_0_24 + population_density
#Model_3=cases_per_100k_at_365d ~ median_transit_change +
#pop_pct_age_0_24 + population_density + mask_mandate_days+
#unemployment_benefits_days
#+increased_weekly_unemployment_insurance_amt_thru_jul31+
#business_close_open_days+travel_quarantine_mandate_days+stay_at_home_days

#lm.II=lm(Model_3,data=final_df)
#summary(lm.II)
#resid(lm.II)
#anova(lm.II, lm.I, test = "F")
#shapiro.test(lm.II$residuals)
#bptest(lm.II)
```

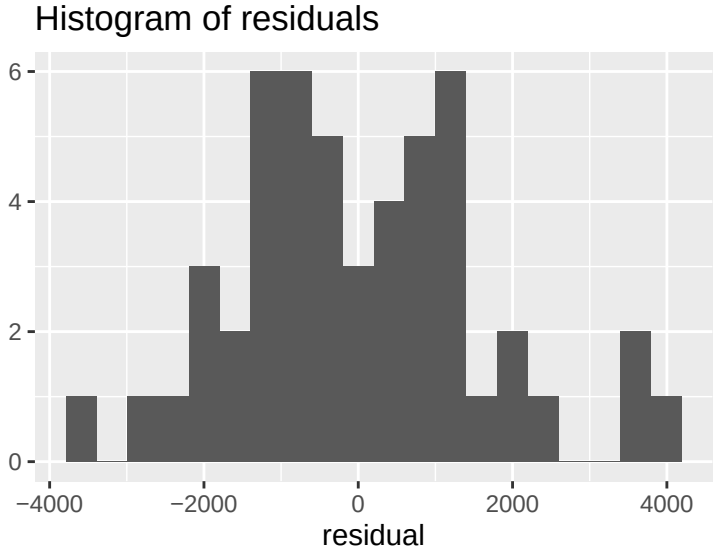
Third Model

Model 3:

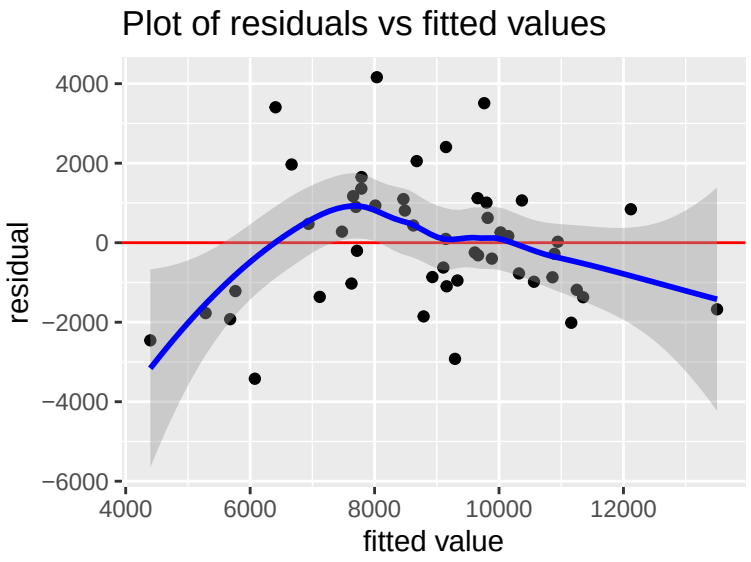
```
##
## Call:
## lm(formula = cases_per_100k_at_365d ~ median_transit_change +
##     pop_pct_age_0_24 + population_density + mask_mandate_days +
##     unemployment_benefits_days + increased_weekly_unemployment_insurance_amt_thru_jul31 +
##     business_closed_days_round1 + travel_quarantine_mandate_days +
##     stay_at_home_days, data = final_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3421.8 -1079.3   -90.2   990.0  4162.1
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                     -4556.676   4720.588
## median_transit_change                              54.348     28.765
## pop_pct_age_0_24                                  569.302    129.214
## population_density                                  5.790      1.575
## mask_mandate_days                                  -1.188      2.766
## unemployment_benefits_days                         -3.873      4.991
## increased_weekly_unemployment_insurance_amt_thru_jul31   -2.339      2.031
## business_closed_days_round1                       -16.598     23.816
## travel_quarantine_mandate_days                     -1.779      2.379
## stay_at_home_days                                   1.841      4.816
##                                                 t value Pr(>|t|)
## (Intercept)                                      -0.965 0.340206
## median_transit_change                             1.889 0.066103 .
## pop_pct_age_0_24                                  4.406 7.68e-05 ***
## population_density                                3.677 0.000694 ***
## mask_mandate_days                                -0.429 0.669918
## unemployment_benefits_days                       -0.776 0.442269
## increased_weekly_unemployment_insurance_amt_thru_jul31  -1.151 0.256369
## business_closed_days_round1                      -0.697 0.489895
```

```
## travel_quarantine_mandate_days                            -0.748 0.458904
## stay_at_home_days                                          0.382 0.704251
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1781 on 40 degrees of freedom
## Multiple R-squared:  0.5695, Adjusted R-squared:  0.4727
## F-statistic:  5.88 on 9 and 40 DF,  p-value: 3.369e-05
```

```
## Analysis of Variance Table
##
## Model 1: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density + mask_mandate_days + unemployment_benefits_days +
##     increased_weekly_unemployment_insurance_amt_thru_jul31 +
##     business_closed_days_round1 + travel_quarantine_mandate_days +
##     stay_at_home_days
## Model 2: cases_per_100k_at_365d ~ median_transit_change + pop_pct_age_0_24 +
##     population_density
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     40 126808331
## 2     46 144077125 -6 -17268794 0.9079 0.4991
```




```
## `geom_smooth()` using formula 'y ~ x'
```



Regression Table

```
robust_se_1 <- coeftest(model_1_final,
                        vcovHC(model_1_final, type = 'HC3'))[ , "Std. Error"]

robust_se_2 <- coeftest(model_2_final,
                        vcovHC(model_2_final, type = 'HC3'))[ , "Std. Error"]

robust_se_3 <- coeftest(model_3_final,
                        vcovHC(model_3_final, type = 'HC3'))[ , "Std. Error"]
```

stargazer output

Table 1: OLS models for COVID-19 Spread

	Dependent variable:		
	cases_per_100k_at_365d		
	(1)	(2)	(3)
median_transit_change	79.250** (32.943)	82.223*** (31.481)	54.348 (45.057)
pop_pct_age_0_24		628.592*** (154.756)	569.302*** (160.445)
population_density		4.665** (2.213)	5.790** (2.789)
mask_mandate_days			-1.188 (3.116)
unemployment_benefits_days			-3.873 (5.252)
increased_weekly_unemployment_insurance_amt_thru_jul31			-2.339 (2.110)
business_closed_days_round1			-16.598 (23.841)
travel_quarantine_mandate_days			-1.779 (3.121)
stay_at_home_days			1.841 (5.388)
Constant	10,370.170*** (610.184)	-10,160.590** (5,070.544)	-4,556.676 (5,891.585)
Observations	50	50	50
R2	0.168	0.511	0.570
Adjusted R2	0.151	0.479	0.473
Residual Std. Error	2,259.262 (df = 48)	1,769.777 (df = 46)	1,780.508 (df = 40)
F Statistic	9.714*** (df = 1; 48)	16.018*** (df = 3; 46)	5.880*** (df = 9; 40)
Note:			*p<0.1; **p<0.05; ***p<0.01

Plots, Figures, and Tables

Do the plots, figures and tables that the team has chosen to include successfully move forward the argument that they are making? Has the team chosen the most effective method (a table or a chart) to display their evidence? Is that table or chart the most communicative it could be? Is every plot, figure, and table that is included in the report referenced in the narrative argument?

Assessment of the CLM.

Has the team presented a sober assessment of the CLM assumptions that might be problematic for their model? Have they presented their analysis about the consequences of these problems (including random sampling) for the models they estimate? Did they use visual tools or statistical tests, as appropriate? Did they respond appropriately to any violations?

An Omitted Variables Discussion.

Did the report miss any important sources of omitted variable bias? Are the estimated directions of bias correct? Was their explanation clear? Is the discussion connected to whether the key effects are real or whether they may be solely an artifact of omitted variable bias?

Conclusion.

Does the conclusion address the research question? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?

Are there any other errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?

General Notes:

"In principle the SE reflects the degree of uncertainty or the lack of information for getting a 'good' (that is reliable) estimate of a parameter. Therefore if you keep everything else the same (eg the same variation in the response, the same number of observations) but you increase the number of separate parameters to be estimated there will be less information per parameter to get the estimate, and hence larger standard error. Precisely what happens will depend on the the degree of variation in the additional X variable that is included and how colinear it is with already included variables."

Known IID violations: Geo-spatial dependence (states near each other are not independent…physical proximity) Policy coordination dependence (states near each-other have coordinated policies (like NY/NJ quarantine policies, etc))

Other limitations: Mobility data is based on Google-Maps cell phone users. Not everyone has access to a smart phone or uses Google Maps and allows their location to be traced, so this data may not be representative of the population. Additionally, we do not have absolute numbers, only relative change data.