W203 Lab 2 Report: How Does Moblity Affect COVID-19 Cases

Jeff Adams, Brittany Dougall, Li Jin, Jerico Johns

Introduction

It has been over a year since the initial outbreak of COVID-19 in the US. During this year, many states have adopted various policies to slow down the spread of the pandemic. For example, in New York City, Governor Andrew Cuomo shut indoor dining on March 16, 2020. Many companies also started to operate in telecommuting mode. As a result, people's mobility patterns are very different from the pre-COVID period. This presents an opportunity for studying the impact of these mobility changes on the spread of COVID-19.

We think this is an interesting research question now because some states are starting to gradually revert some policies. For example, as of 3/19/21, indoor dining can increase to 50% capacity in New York City (NYC Business 2021). The research result of the effects of change of mobility on new COVID-19 cases would help people and states make decisions on whether to increase their retail and recreation activity.

We propose exploring a causal relationship between mobility, public mask mandate, and demographics within a state and COVID-19 cases. Formally, our primary research question is:

How much does a percentage change in mobility impact the percentage change in new COVID-19 cases within a given state, given a 2 week lag period?

We choose to use 2 weeks (end of the first time period to the beginning of the second time period) between mobility & mask mandate and new COVID-19 cases based on the current CDC guidelines (Centers for Disease Control and Prevention 2020b) suggesting an individual is likely to incubate the disease $2 - 14$ days after initial symptoms appear (which could be up to 10 days after initial infection).Thus 2 weeks after (i.e. days 21 to 28) mobility/mask mandate best covers this period of 12 to 24 days with some buffer for discovery of the infection and test turn around times.

Model Building Process

Data Sources

New Resources

- New York Times COVID-19 Cases Data (The New York Times 2021): A series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time

Class Resources

- COVID-19 US State Policy Database (Raifman, J., et al. 2020): A database of state policy responses to the pandemic, compiled by researchers at the Boston University School of Public Health
- COVID-19 Community Mobility Report (Google 2021): A Google data set that includes state-level measurements of individual mobility
- The American Community Survey (US Census Bureau 2019): A product from the US Census Bureau that contains state-level demographics and other indicators of general interest

Choosing the variables

In this section we will define the dependent variable and independent variables, as well as our reasons to choose them.

Dependent variable

percent_change_in_week_over_week_covid_cases (July 19th - 25th) Percentage change in new COVID cases, comparing the number of new cases from Sunday 07/19/20 - Saturday 07/25/20 vs. Sunday 07/12/20 - Saturday 07/18/20.

The time period of the dependent variable is 2 weeks lagged from the mobility variable. We chose the 2 weeks lagging period based on the CDC suggestion that an individual is likely to incubate the disease $2 - 14$ days after initial symptoms appear (which could be up to 10 days after initial infection). We chose to use percentage change in COVID-19 cases rather than case counts in order to study the impact of mobility when states have very different population sizes, since total case count is related to a state's population size.

We chose to use this specific time period for the following reasons:

- This time period captures the return towards baseline mobility after the initial lock down dip (quarantine fatigue) which is an interesting time period to study (Figure 2).
- This is a week in which 31 states did not have public face mask mandates vs 20 states which did, and therefore is interesting time period to study the effects of public face mask mandates on slowing down the spread of COVID-19 (policy variation and closes to a balanced count between groups).
- Is a counter to the seasonal nature of the disease, i.e. weather plays less of a role in transmission during this summer time period which helps our causal question (Lipsitch, Marc, DPhil 2021).
- It's also early enough in the pandemic that we'd expect a much lower proportion of the population to have contracted the disease already and develop antibodies, which gives us a clearer picture of our causal question since antibodies / disease resistance could be a potential confounding variable.

Our method of choosing percent_change and the lag between change in COVID-19 cases and change in mobility is similar to the 'differenced panels' approach on async video 11.14. By taking the percent change in new COVID-19 cases, we can reduce some impacts of the confounding variables.

Independent variables

percent_change_in_week_over_week_mobility_retail_and_recreation (Sat,Sun,M...F) Seven variables, each a percentage point change in median mobility (compared to baseline) for Retail & Recreation for the week of Sunday 06/28/20 - Saturday 07/04/20 vs Sunday 06/21/20 - Saturday 06/27/20 for each day of the week. Sunday is only compared to Sunday, etc.

Since we have no indication of how to weight these values based on what the real baseline number is (not provided in Google's dataset), we can't simply average all 7 days. For example, a -10% change of mobility on Monday is very different from a -10% change of mobility on Saturday since the baseline number of people for Saturday is much higher. Therefore we decide to keep these variables separate.

public_mask_mandate_flag (In place as of July 1st) Flag for whether or not a public face mask mandate was in place for the state as of July $1st$ (to capture the majority of the week of July $4th$, our primary week of interest for change in mobility)

percent_female (2019) American Community Survey (2019) estimate of state's percentage of female residents.

This study (HBS Working Knowledge 2020) suggests that "women are much more likely than men to view COVID-19 as a severe health problem. They are also more willing to wear face masks and follow other public health recommendations to prevent the spread of the virus countries."

Due to this research, we decided to implement percent female as a control variable when studying the impact of mobility upon each state's percent change in COVID-19 cases.

percent_white (2019) American Community Survey (2019) estimate of state's percentage of white residents.

This article (Mayo Clinic 2020) suggests that "Research increasingly shows that racial and ethnic minorities are disproportionately affected by COVID-19 in the United States."

Therefore we decide to include this variable as a control variable.

percent_65_and_older (2019) American Community Survey (2019) estimate of state's percentage of residents 65 or older (66 or older at time of analysis).

CDC suggests that "Older adults are At greater risk of requiring hospitalization or dying if diagnosed with COVID-19" (Centers for Disease Control and Prevention 2020a).

Therefore we think that states with a high percentage of older adult population could have lower mobility due to the fear of getting COVID-19 and therefore, lead to lower COVID-19 cases.We decide to include this variable as a control variable.

percent_24_and_younger (2019) American Community Survey (2019) estimate of state's percentage of residents 24 or younger (25 or younger at time of analysis).

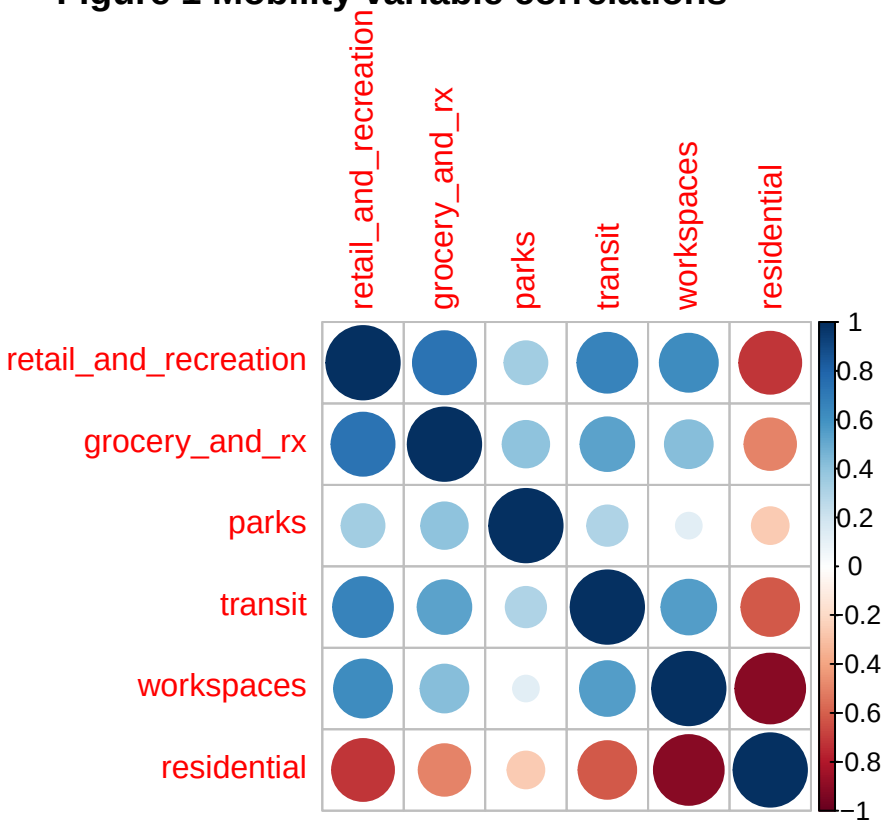
According to the news, some college students have shown reckless behavior with respect to social distancing (Inside Higher Ed 2020).

We include this variable to study if the difference in percentage of younger residents has an effect on COVID-19 cases.

Choosing the mobility variable

In this section, we will show the process of choosing the mobility variable. The two charts show (1) the correlation plot between different variables and (2) the time series plot show the relationship between different mobility variables by time.

Figure 1 Mobility variable correlations



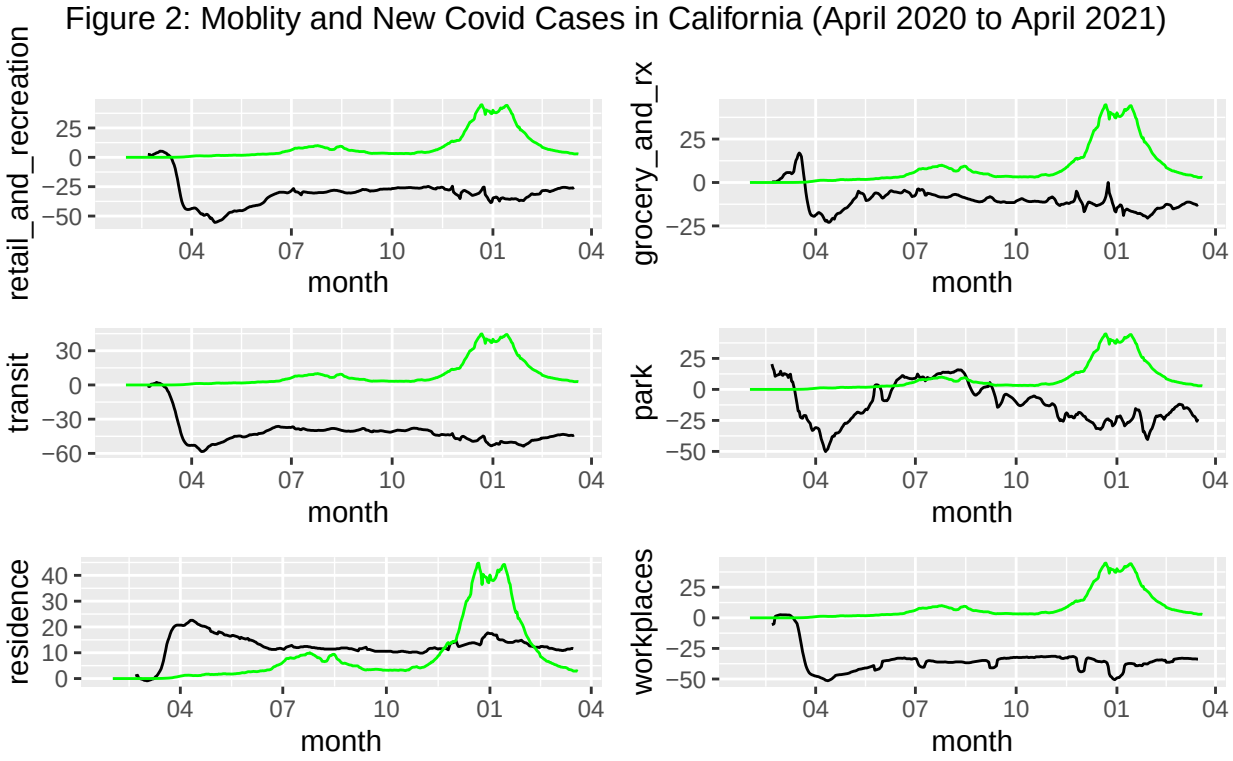


Figure 1, 2 show that retail, transit and grocery have positive correlation of > 0.70 (figure 1) and they tend to move in the same direction (figure 2). Retail and residential have a negative correlation of -0.75 and they tend to move approximately in the opposite direction.

Since retail mobility is positively correlated with transit and grocery and negatively correlated with residence mobility, we decide to only include one of them as our independent mobility variable. We decide to use retail mobility because we think its practical meaning is the most interesting among all the mobility variables, given the political focus on restaurants and storefronts. We decided not to use workplaces and park mobility variable because (1) parks are outdoor spaces and easier to maintain 6 foot social distances, and therefore, we don't expect changes in park mobility to cause change in new cases of COVID-19 and (2) for workplaces, generally speaking people have less control of whether to go to workplaces or not, and therefore, this variable does not have sufficient practical impact.

Another observation is that there is no clear correlation between any of the mobility variables and new COVID-19 cases. Still, we decide to include mobility as an independent variable to study its combined effects with other independent variables (public mask mandate and demographic variables).

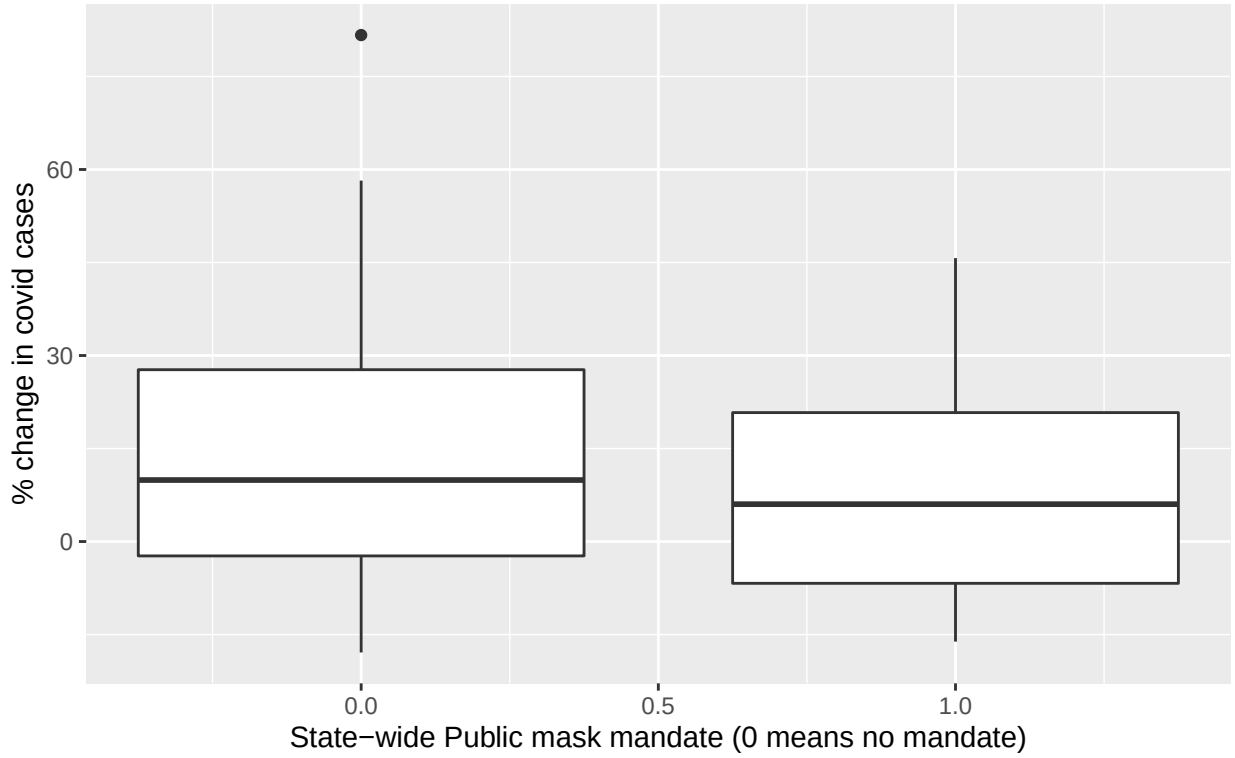
Examine the control variables

In this section, we will examine the relationship between control variables (public_mask_mandate_flag, percent_female, percent_white, percent_65_and_older, percent_24_and_younger) and the dependent variable (percent_change_in_week_over_week_covid_cases) for the study time period (July 19th - 25th).

Public mask mandate Looking at the box plot for percent change in COVID-19 cases group by public mask mandate variable for the 51 states (Figure 3), we see that there is a slight higher mean in percent change in COVID-19 cases in states without public mask mandate than with one. We decide to include this control variable and further examine its coefficient and p-value.

Note that the outlier in Figure 3 is Hawaii. Fewer tests were conducted over the weekend amidst a hurricane threat (Honolulu Civil Beat 2020). We decided to include this in the regression because there was also a cluster of cases linked to failure to comply with face mask mandates and social distancing at two bars (Honolulu Civil Beat 2020) and we don't know how many cases were impacted by the reporting delay.

Figure 3:
change in covid cases with/without public mask mandate



Demographics variables Here is a combined scatter plot for the four demographics variables we are interested in (percent_female, percent_white, percent_65_and_older, percent_24_and_younger). The scatter plots don't show a clear pattern of the linear relationship just by these control variables themselves. Nonetheless, we decide to include these variables in the model specification to study further.

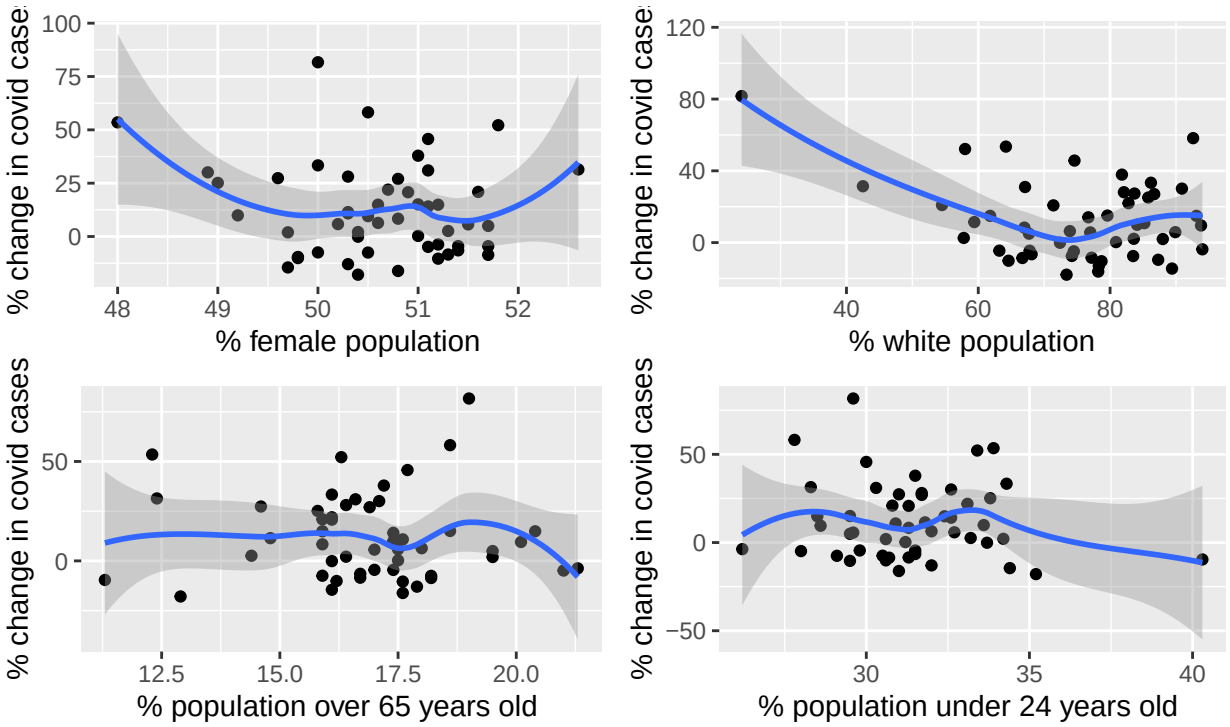


Figure 4:
Relationship between demographics control variables and % change in covid cases

Modeling Goal

Our modeling goal is explanatory. Primarily, we are interested in the causal relationship between mobility change in retail and recreation of the week of 06/28/2020 and the change of new COVID-19 cases 2 weeks later. Secondarily, we are interested in the causal question described above, with the control of public mask mandate and demographic variables as control variables.

We are only studying the causal relationship within the chosen time period and not planning to extend any conclusion beyond the original time period.

Causal Theory

Increase in retail and recreation mobility causes increase in new covid cases From CDC (Centers for Disease Control and Prevention 2020b): "The virus that causes COVID-19 most commonly spreads from person to person by respiratory droplets during close physical contact (within 6 feet). The virus can sometimes spread from person to person by small droplets or virus particles that linger in the air for minutes to hours."

Based on this, our causal theory is that an increase in retail and recreation mobility would result in an increase in the number of people that each person is in contact with, and therefore, increase the chances of each individual infected by COVID-19, and finally, result in an increase in new COVID-19 cases, given a 2 week lag period for incubation and infection.

Public Facemask Mandate causes decrease in new covid cases COVID-19 is airborne and spread by respiratory droplets which enter through the nose and mouth. Wearing a mask will help prevent the spread of these droplets into the air. It is possible for someone to spread the virus even if they do not have symptoms (Olmsted Medical Center 2021).

Based on this, our causal theory is that having a public face mask mandate causes a decrease in new COVID-19 cases, so long as mobility is constant. In other words, public mask mandates moderate the impact mobility has on new COVID-19 cases.

Model Specifications

Model One percent_change_in_week_over_week_covid_cases ~ percent_change_in_week_over_week_mobility_retail_and_recreation (Sunday, Monday, .. , Saturday)

Model Two percent_change_in_week_over_week_covid_cases ~ percent_change_in_week_over_week_mobility_retail_and_recreation (Sunday, Monday, .. , Saturday) + public_mask_mandate_flag

Model Three percent_change_in_week_over_week_covid_cases ~ percent_change_in_week_over_week_mobility_retail_and_recreation (Sunday, Monday, .. , Saturday) + public_mask_mandate_flag + percent_white + percent_female + percent_under_24 +percent_over_65

Regression Table

We chose to use classical standard errors in our regression model and coefficient tests because in our case, they were larger than robust standard errors (more conservative). As will be discussed in our assessment of the classical linear model assumptions, we have not found evidence of heteroscedastic variance of errors. We chose to use a general linear F-test for nested models to determine if there is a statistically significant improvement in our F-statistic between reduced and full models. This gives us a statistically testable comparison of our models based on the error sum of squares (SSE/RSS) for each. We prefer the comparison of the SSE statistic to R-squared because it is in the same unit as the dependent variable, whereas R-squared is a percentage term, and thus SSE is more easily interpretable in both absolute and relative terms. For assessing the significance of the coefficients of our predictors and for conducting our F-test, we chose to use an α level of 0.05 due to the popularity of this threshold value and since we do not have a preference, given our limited background in the study of COVID-19 case analysis.

```
##
## Regression Results
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                      Percent Change in New Cases
##                                (1)                (2)                (3)
## --------------------------------------------------------------------------------
## Sunday (Retail)                -1.268             -0.445              0.604
##                               (1.709)            (1.746)            (1.576)
##
## Monday (Retail)                1.629              1.676              0.331
##                               (1.615)            (1.583)            (1.519)
##
## Tuesday (Retail)               -0.158             -0.511             -2.105
##                               (1.791)            (1.767)            (1.649)
##
## Wednesday (Retail)             1.371              1.291              -0.378
##                               (1.495)            (1.466)            (1.399)
##
## Thursday (Retail)              2.606              2.805*             3.837**
##                               (1.684)            (1.654)            (1.502)
```

```
##
## Friday (Retail)                -0.215             -0.267               0.572
##                               (1.074)            (1.053)             (0.978)
##
## Saturday (Retail)               0.819             1.390*              -0.403
##                               (0.665)            (0.736)             (0.830)
##
## Public Mask Mandate Flag                          -11.812           -14.086**
##                                                   (7.071)             (6.672)
##
## Percent Female                                                       -8.972**
##                                                                      (3.959)
##
## Percent White                                                       -0.972***
##                                                                      (0.286)
##
## Percent Under 24                                                     -3.740*
##                                                                      (2.100)
##
## Percent Over 65                                                      -2.119
##                                                                      (2.163)
##
## Constant                      -12.313             2.470            656.068***
##                               (15.949)          (17.959)           (239.735)
##
## ------------------------------------------------------------------------------------
## Observations                    51                 51                  51
## R2                             0.221              0.270               0.487
## Adjusted R2                    0.094              0.130               0.325
## Residual Std. Error       20.445 (df = 43)   20.032 (df = 42)    17.656 (df = 38)
## F Statistic              1.744 (df = 7; 43) 1.938* (df = 8; 42) 3.002*** (df = 12; 38)
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Comparison of Model 2 and Model 1:

```
## Analysis of Variance Table
##
## Model 1: p1_p2_percent_change_in_cases ~ Sunday_retail + Monday_retail +
##     Tuesday_retail + Wednesday_retail + Thursday_retail + Friday_retail +
##     Saturday_retail + public_mask_mandate_flag
## Model 2: p1_p2_percent_change_in_cases ~ Sunday_retail + Monday_retail +
##     Tuesday_retail + Wednesday_retail + Thursday_retail + Friday_retail +
##     Saturday_retail
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     42 16854
## 2     43 17974 -1   -1119.8 2.7905 0.1023
```

Comparison of Model 3 and Model 2:

```
## Analysis of Variance Table
##
## Model 1: p1_p2_percent_change_in_cases ~ Sunday_retail + Monday_retail +
##     Tuesday_retail + Wednesday_retail + Thursday_retail + Friday_retail +
##     Saturday_retail + public_mask_mandate_flag + percent_female +
```

```
##     percent_white + percent_under_24 + percent_over_65
## Model 2: p1_p2_percent_change_in_cases ~ Sunday_retail + Monday_retail +
##     Tuesday_retail + Wednesday_retail + Thursday_retail + Friday_retail +
##     Saturday_retail + public_mask_mandate_flag
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     38 11846
## 2     42 16854 -4   -5008.2 4.0164 0.008184 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of our regression analysis in the results table using the stargazer library (Hlavac, Marek 2018) indicate that, generally, changes in mobility (as proxied by recreation and retail) for the weeks of June 28, 2020 to July 4th, 2020 do not explain changes in new COVID-19 cases two weeks later (unless the effect of mobility has been masked by an omitted variable). Model 1 in the regression table above demonstrates that there is not a statistically significant weekday mobility change that explains changes in statewide COVID-19 cases. As a whole, we fail to reject the null hypothesis that Model 1 has a significantly higher F-statistic than an intercept only model ($p > 0.05$). We also observe that adding a control for a public mask mandate flag in Model 2 does not significantly improve our model compared to Model 1 (we fail to reject the null hypothesis that Model 2 has a statistically different F-statistic than Model 1 since conducting an F-test to compare models 1 and 2 results in a p-value greater than our α of 0.05). In Model 2, the public mask mandate variable is not significant at an α level of 0.05.

Interestingly, we see significant improvements in our explanatory model as a whole as measured by a reduction in the sum of squared estimates of errors (also known as RSS) and certain explanatory variables when adding controls for state-level demographic features in Model 3. We reject our null hypothesis that Model 2 and Model 3 have the same F-statistic (since the resulting p-value from the F-test of models 2 and 3 is less than our α of 0.05), and conclude that Model 3 is a statistically significant improvement upon Model 2 and an intercept only model (i.e. a regression with no independent variables).

When controlling for a state's gender balance (percent female), racial composition (percent white), and age (percent under 24 and percent over 65), we observe that the change in mobility for Thursday (July $2nd$ compared to June $25th$) is statistically significant at an α level of 0.05. However, since only one weekday's mobility change is statistically significant, we do not deem this finding a practically significant finding regarding mobility more generally (for this week long period). Six of seven mobility variables are still not statistically significant. Thus, we conclude that changes in mobility do not explain changes in new COVID-19 cases for this time-period and this model specifically. Mobility may still explain changes in COVID-19 cases in different time periods or after accounting for the impact of all omitted variables (see Omitted Variables section).

In our Model 3, we observe that the public mask mandate flag is statistically significant ($p = 0.04$) at an α level of 0.05, when controlling for mobility and the aforementioned demographic variables. In our observation period, state's with a public mask mandate flag in place as of July $1st$, 2020 (i.e. the majority of the week that includes the July $4th$ holiday) saw 14% fewer new COVID-19 cases week-over-week, two weeks after the mobility time period than states with no public mask mandate flag in place, holding mobility and demographics constant. This is both statistically and practically significant. This indicates that the oft-debated mask policies instituted during this time period may be effective at reducing new COVID-19 cases (holding other variables constant), if this observation is not an artifact of omitted variable bias.

We also observe that percent female ($p = 0.03$) and percent white ($p = 0.002$) are statistically significant at an α level of 0.05 in our Model 3. A 1 percentage point (ppt) increase in a state's proportion of female residents was associated with 9% fewer new COVID-19 cases week-over-week, when holding mobility, public mask mandate flag, and other demographic variables constant. This is both statistically and practically significant. Although a state can not simply adjust demographic composition, it is practically significant in helping understand how COVID-19 may impact different genders and/or how different gender groups may adhere to COVID-19 health guidelines. A 1 ppt increase in a state's proportion of residents who identify as white was associated with 1% fewer new COVID-19 cases week-over-week, when holding mobility, public

mask mandate flag, and other demographic variables constant. This coefficient is statistically significant and practically significant, if this observation is not an artifact of omitted variable bias. Even though this is our smallest coefficient estimate, the implications of COVID-19 trending differently for different racial groups points to a need for further research in this critical area.

Limitations of the Model

I.I.D. The COVID-19 and mobility observations that our sample is pulled from do not pass the independent assumption requirement of the classic linear model. One contributing factor to the independence violation is state proximity. State observations are not independent due to their proximity to one another. As people travel they can spread the virus across state borders. This can cause the mobility in one state to have an impact on COVID-19 cases in a bordering state. States' COVID-19 policies may be influenced by trends in nearby states. This causes each of our samples to not have true independence. Since we are unable to collect new COVID-19 and state policy observations to sample from, we acknowledge that our findings will be informative of the clusters rather than the population and proceed with our study in spite of this Classical Linear Model (CLM) violation. Regarding the identical distribution requirement of the CLM, we fail to reject the null hypothesis that our sample is taken from identically distributed observations. Our study is taken across a short period of time that is identical for every state, during which the same federal policies are in place. This means that each of our samples are being pulled from the exact same population. In our study, the population is the United states from 6/28/21 to 7/24/21.

No perfect collinearity In order to test the collinearity between the predictors in our models, we chose to assess the Variance Inflation Factor (VIF) scores of our respective models. The square root of the VIF score compares the standard error for each predictor to the standard error for that predictor if it was uncorrelated with any other predictor in the model and therefore allows us to determine collinearity between predictors. To calculate VIF scores for our respective models, we used the function summ from the jtools package in R and set the vifs argument to true. The threshold for VIF scores suggestive of problematic multicollinearity differs between statistical texts, with the threshold occurring at a VIF score of 5 in some sources and a 10 in others. For our tests, we chose to compare the model scores to a VIF score of 5 to be conservative. In our interpretation, a VIF score of 5 or greater suggests moderate to high correlation between model predictors. If found, this correlation can be addressed through eliminating redundant features or through dimension reduction such as PCA analysis.

The VIF scores for the first model below correspond to those of our base model with only the changes in retail mobility for each day of the week paired observation. All VIF scores are below 2.7, suggesting although there is some correlation between predictors, the degree of collinearity is not problematic.

Observations	51
Dependent variable	p1_p2_percent_change_in_cases
Type	OLS linear regression

F(7,43)	1.74
R^2	0.22
Adj. R^2	0.09

	Est.	S.E.	t val.	p	VIF
(Intercept)	-12.31	15.95	-0.77	0.44	NA
Sunday_retail	-1.27	1.71	-0.74	0.46	2.01
Monday_retail	1.63	1.61	1.01	0.32	2.15
Tuesday_retail	-0.16	1.79	-0.09	0.93	2.26
Wednesday_retail	1.37	1.50	0.92	0.36	2.64
Thursday_retail	2.61	1.68	1.55	0.13	2.63
Friday_retail	-0.22	1.07	-0.20	0.84	1.62
Saturday_retail	0.82	0.66	1.23	0.22	1.29

Standard errors: OLS

The 2nd test result shows the VIF scores when the predictor for an in-place public mask mandate is added to the model. The VIF scores for the predictors in this model experience small increases but scores continue to remain below 5, suggesting that the addition of the mask mandate flag has not added problematic collinearity.

Observations	51
Dependent variable	p1_p2_percent_change_in_cases
Type	OLS linear regression

F(8,42)	1.94
R^2	0.27
Adj. R^2	0.13

	Est.	S.E.	t val.	p	VIF
(Intercept)	2.47	17.96	0.14	0.89	NA
Sunday_retail	-0.44	1.75	-0.25	0.80	2.18
Monday_retail	1.68	1.58	1.06	0.30	2.15
Tuesday_retail	-0.51	1.77	-0.29	0.77	2.29
Wednesday_retail	1.29	1.47	0.88	0.38	2.64
Thursday_retail	2.80	1.65	1.70	0.10	2.65
Friday_retail	-0.27	1.05	-0.25	0.80	1.63
Saturday_retail	1.39	0.74	1.89	0.07	1.65
public_mask_mandate_flag	-11.81	7.07	-1.67	0.10	1.51

Standard errors: OLS

When additional demographic predictors were introduced, as seen in the below result, VIF scores increased, with the highest scores now over 3. Although scores are still below our threshold of problematic collinearity, the larger scores suggest a greater degree of correlation than in our second model.

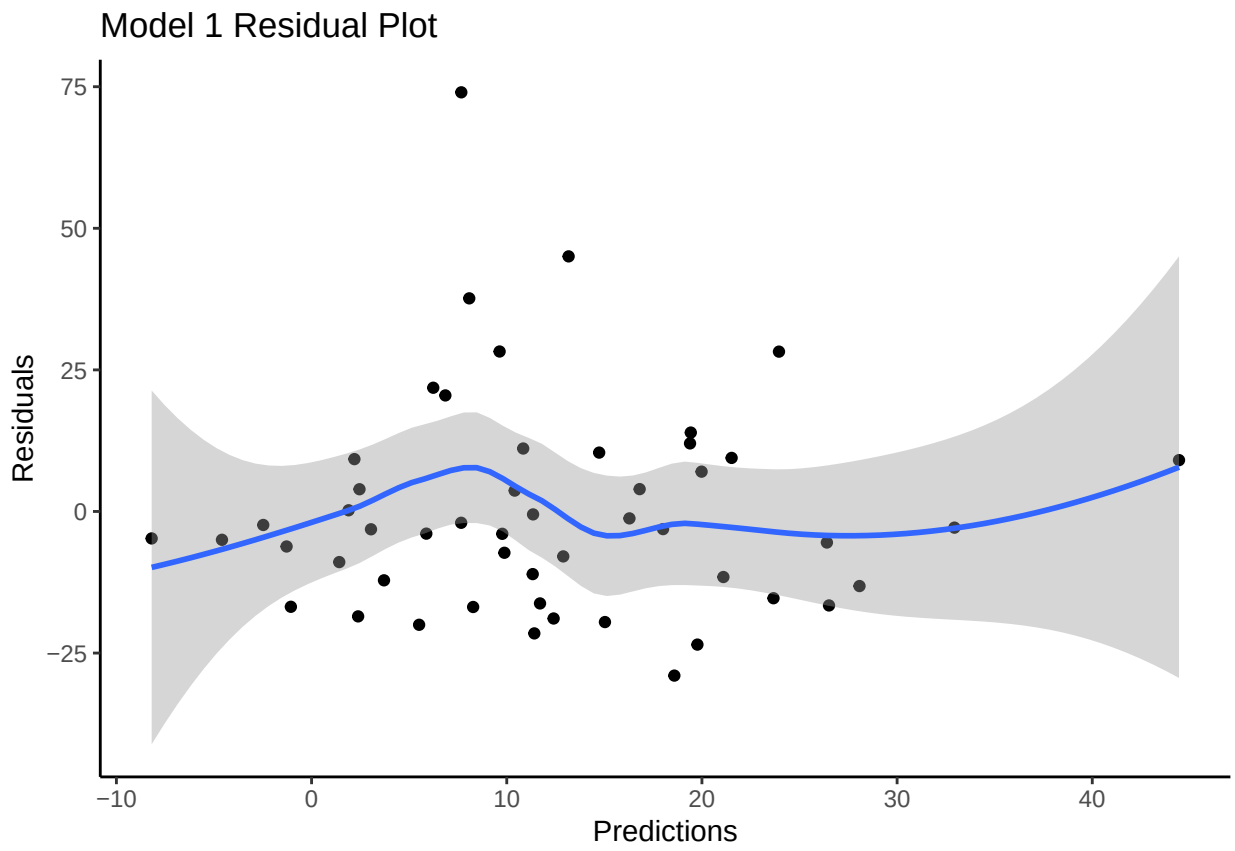
Observations	51
Dependent variable	p1_p2_percent_change_in_cases
Type	OLS linear regression

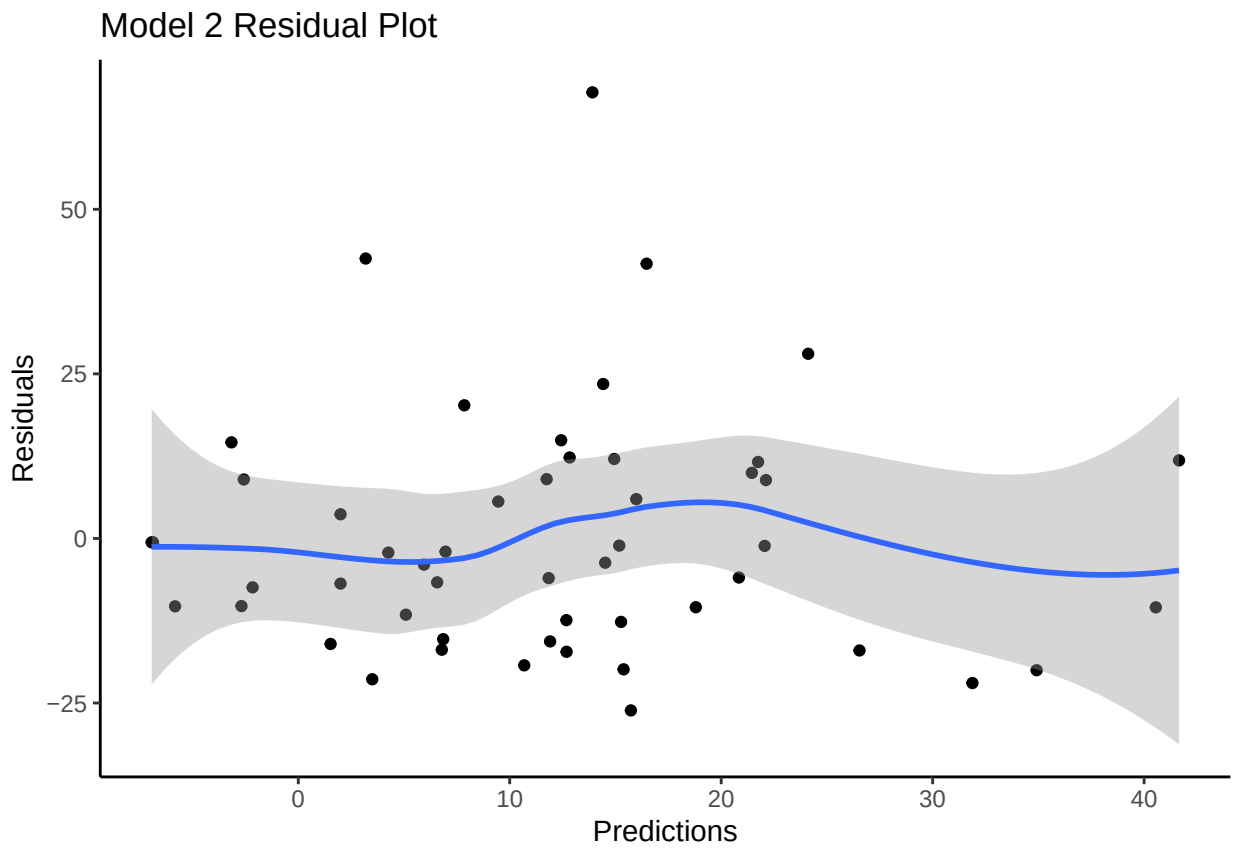
F(12,38)	3.00
R^2	0.49
Adj. R^2	0.32

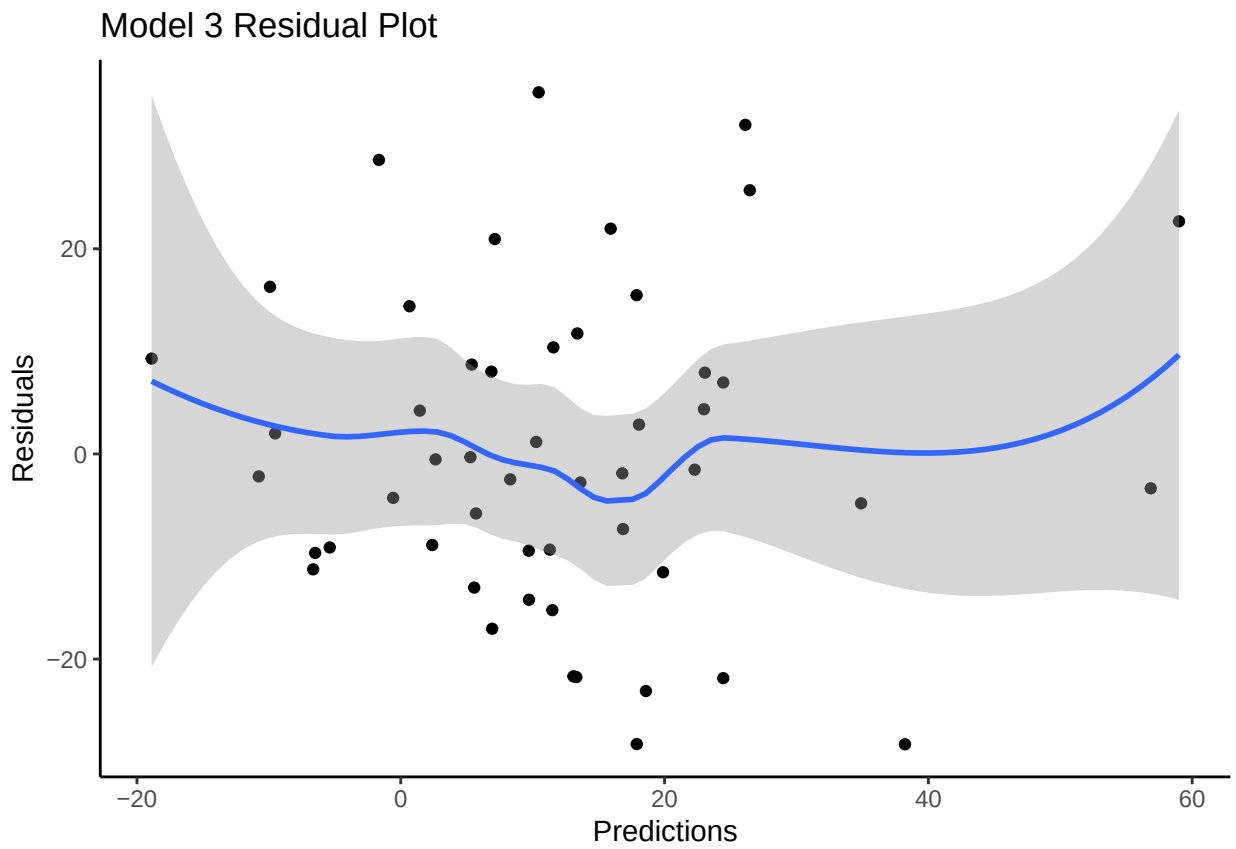
	Est.	S.E.	t val.	p	VIF
(Intercept)	656.07	239.74	2.74	0.01	NA
Sunday_retail	0.60	1.58	0.38	0.70	2.29
Monday_retail	0.33	1.52	0.22	0.83	2.55
Tuesday_retail	-2.10	1.65	-1.28	0.21	2.57
Wednesday_retail	-0.38	1.40	-0.27	0.79	3.10
Thursday_retail	3.84	1.50	2.55	0.01	2.81
Friday_retail	0.57	0.98	0.58	0.56	1.81
Saturday_retail	-0.40	0.83	-0.49	0.63	2.70
public_mask_mandate_flag	-14.09	6.67	-2.11	0.04	1.74
percent_female	-8.97	3.96	-2.27	0.03	1.80
percent_white	-0.97	0.29	-3.40	0.00	2.41
percent_under_24	-3.74	2.10	-1.78	0.08	3.78
percent_over_65	-2.12	2.16	-0.98	0.33	3.10

Standard errors: OLS

Linear conditional mean Since our 2*nd* and 3*rd* models contain multiple predictors, we proceeded to test this CLM assumption in a higher dimensional space - i.e. plotting predictions vs. residuals, as shown below. For the first and second model, the plots show a relatively flat linear trend line around zero and do not appear to have a non-linear pattern - the average residuals change little with the predicted values. The 3rd model has a greater bend in the linear trend line, but still does not show a non-linear pattern between residual and predicted values.


Model 1 Residual Plot
Residuals
75
50
25
0
−25
−10
0
10
20
30
40
Predictions


Model 2 Residual Plot
Residuals
50
25
0
−25
0
10
20
30
40
Predictions



Since the plots of residuals versus predicted values do not demonstrate a clear non-linear relationship, we have not found evidence that the addition of additional predictors in the second and third models result in a clear violation of the CLM assumption of linear conditional expectation.

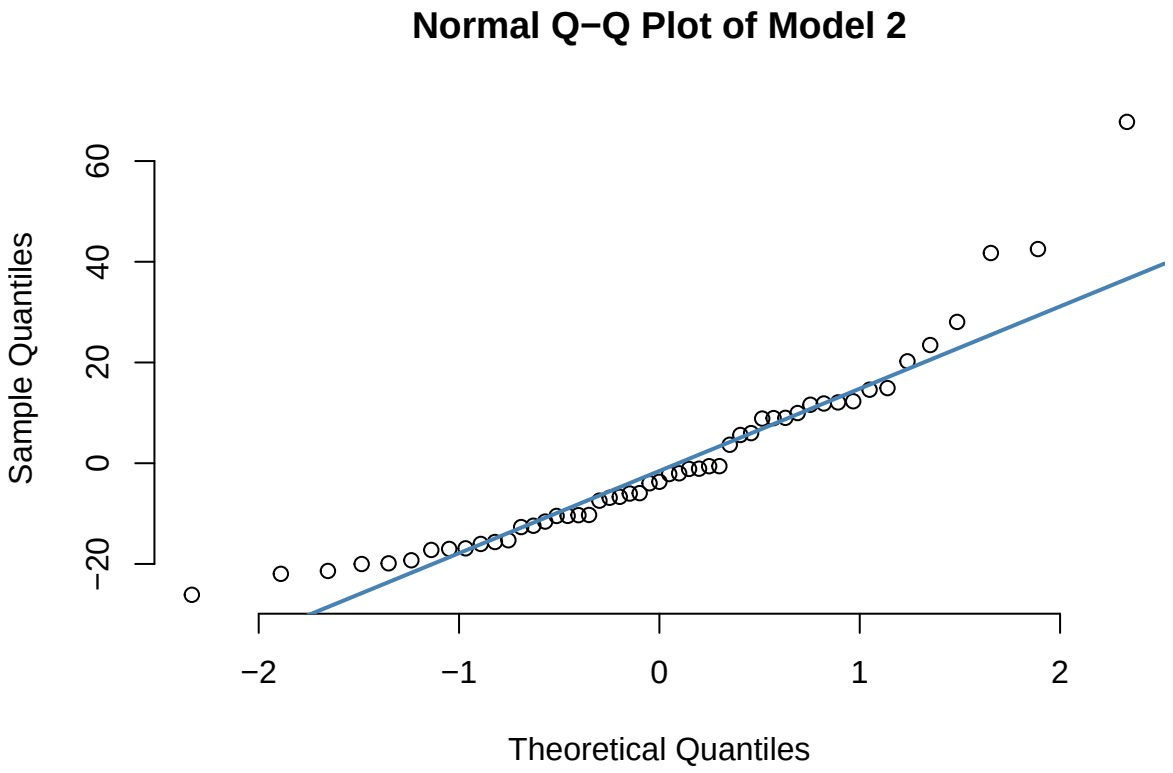
Homoscedasticity The CLM assumption of homoscedasticity can also be assessed visually, through the plots of the residuals versus the predicted values used to assess the assumption of linear conditional expectation. In the plots of residuals versus predicted values for model 1 and 2, the distance of the residuals from the fitted line increases toward the center and is lower on the ends where there are fewer data points. Since we failed to reject the null hypothesis of a linear conditional expectation for our models, we decided to test these models for homoscedasticity using the Breusch-Pagan test. For our α, we chose 0.05 due to the popularity of this threshold value and since we do not have a preference, given our limited background in the study of COVID-19 case analysis. Breusch-Pagan assumes that there is no evidence for heteroscedastic error variance. For all 3 models, we fail to reject the null hypothesis of homoscedasticity, as demonstrated by the below p-value (all p-values are greater than our threshold of 0.05). For this reason and since classical standard errors were larger than robust standard errors for our models, we chose to use classical standard errors to be conservative.

```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 2.1001, df = 7, p-value = 0.9541
```

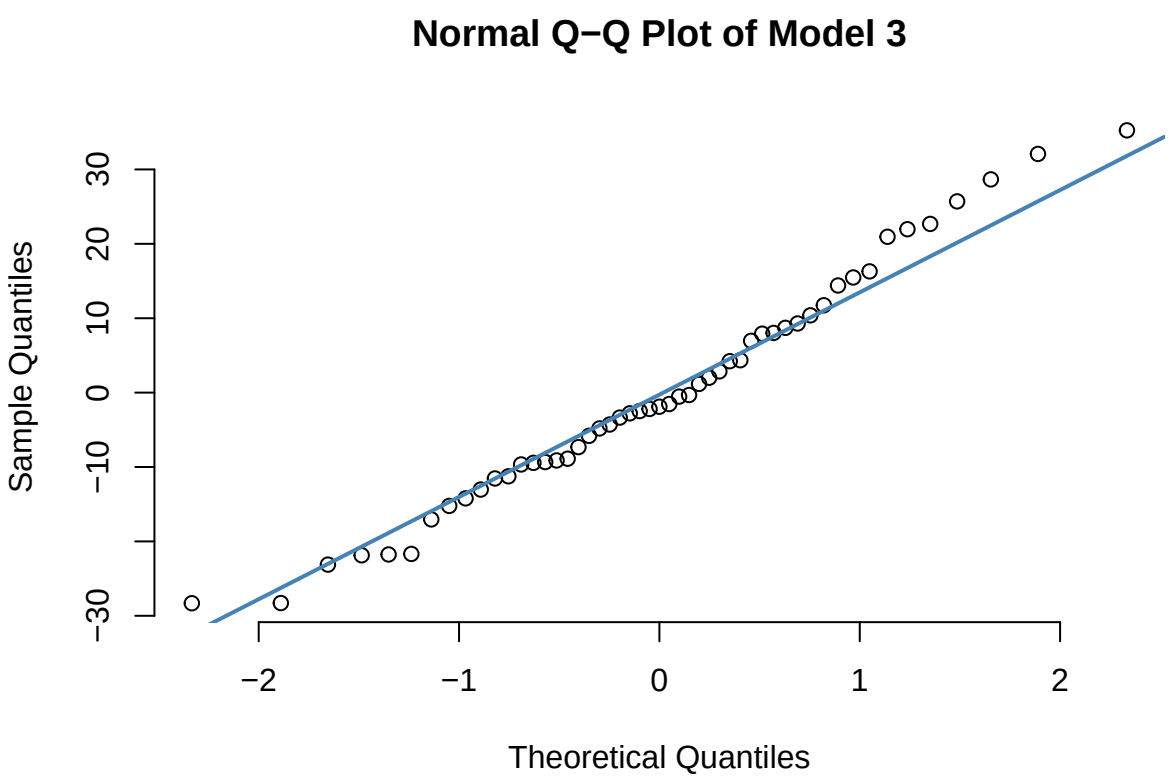
```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 4.0628, df = 8, p-value = 0.8514
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 18.216, df = 12, p-value = 0.1093
```

Normally distributed errors To test the assumption of normally distributed errors, we plotted the residuals from models 2 and 3 against the normal distribution using QQ-plot. On the residual plot of model 2 the data points are curved up and away from the regression line at both ends, indicating a right skew to our observations. This indicates that our uncertainty estimates are biased.



Conversely, the residuals of model 3 with the demographic predictors are more normally distributed, according to the below Q-Q plot (the tails are closer to the qq-line).



We confirmed these visual observations using the Shapiro Wilks test to check the normality of our residuals - in Shapiro Wilks test, the null hypothesis is that the residuals are normally distributed. For our α, we again chose to use 0.05 due to the popularity of this threshold value and since we do not have a preference, given our limited background in the study of COVID-19 case analysis. As seen below, the test result for model 2 aligns with our visual observation - we reject the null hypothesis that the residuals are normally distributed since our p-value is smaller than our α of 0.05. This means that for our second model, our uncertainty estimates are biased. For the third model, our p-value is larger than our α, meaning that we fail to reject the null hypothesis that the residuals are normally distributed - we have not found evidence that our uncertainty estimates are biased.

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_2)
## W = 0.8973, p-value = 0.0003422
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_3)
## W = 0.98082, p-value = 0.5743
```

Discussion of Omitted Variables

In this section we will highlight the five omitted variables that we believe are most important for our model. If possible, we will determine the direction of the omitted variable bias for the independent variables in our model that have statistical significance. We chose to look at only the statistically significant variables because

the sign of the coefficient cannot be trusted if there is no statistical significance. The statistically significant variables from our models are: 'Thursday Retail and Recreation Mobility', 'Public Mask Mandate Flag', 'Percent Female', and 'Percent White'. Below you can find a summary of our results in Table 1.

Table 1:

	Covid Cases	Mobility (+)		Mask (-)		% Female (-)		% White (-)	
	Corelation	Corelation	Bias	Corelation	Bias	Corelation	Bias	Corelation	Bias
Location Tracking	+	+	Away from Zero	NA	NA	NA	NA	NA	NA
Mask Mandate Compliance	-	-	Away from Zero	+	Away from Zero	+	Away from Zero	-	Towards Zero
Testing	+	-	Towards Zero	+	Towards Zero	NA	NA	NA	NA
Population Density	+	-	Towards Zero	+	Towards Zero	NA	NA	NA	NA
Asymptomatic	-	+	Towards Zero	-	Towards Zero	NA	NA	NA	NA

Proportion of population without location tracking enabled

The changes in mobility in our observations are obtained from smart phone devices where the user opted in to location tracking (Newton, Casey 2020). This definition suggests the omission of one key variable in our model predicting COVID-19 cases by changes in mobility - the proportion of the population without a phone with location tracking enabled. This variable is not captured in the mobility, state policy, or demographic features directly. Additionally, none of the available variables can adequately proxy for this omitted variable.

We expect that this omitted variable has a positive correlation with our dependent variable, percent increase in COVID-19 cases. Given the high cost associated with smart phone devices, we believe that mobility among lower income groups may be underrepresented. Members of low income groups may have higher representation in essential work such as grocery stores, in which they cannot work from home and in which they have greater risk of COVID-19 exposure than the general population. Similarly, those who possess a smart phone but intentionally opt out of location sharing may also be at higher risk for COVID-19 exposure, but due to engaging in higher risk behaviors.

Thursday Retail and Recreation Mobility

We expect that individuals who do not have location tracking enabled on a smart phone will have higher mobility near the holiday. This may be due to greater staffing needs for essential workers including in low-income essential work and due to higher risk behaviors such as traveling out of the area or congregating with non-household members for those who have intentionally opted out of location sharing. For this reason, we expect a positive relationship between this omitted variable and the predictor Thursday Retail and Recreation Mobility.

Since the expected correlation between the omitted variable and Public Mask Mandate is positive and the expected correlation between the omitted variable and percent change in COVID-19 cases is also positive, the direction of the bias is positive. Since the coefficient for Thursday Retail and Recreation Mobility is positive, the omitted variable is making the coefficient more positive than it would otherwise be - the bias for this coefficient is away from 0 and the coefficient is overestimated. If this is the case, the percentage increase in COVID-19 cases due to Thursday Retail and Recreation Mobility may be an artifact of omitted variable bias - since the coefficient for Thursday Retail and Recreation Mobility is overestimated, it may actually be zero.

However, since we are measuring percent changes in COVID-19 cases in each state over a short window of time, we believe that this omitted variable proportion of the population without location tracking enabled on a smart phone will experience little change from our starting time period to our final time period, so it will be consistent within each state.

Public Mask Mandate

We do not expect a relationship between this omitted variable and Public Mask Mandate, in part because of limited knowledge of the demographic of the group whose mobility is not captured. If, for example, the group primarily consisted of low income essential workers who experience greater potential COVID-19 exposure than the general population, this may factor into state policies to implement a mandatory mask mandate in the interest of protecting these workers. However, the demographics of those who intentionally opt out of

location sharing are not known - the demographics of these individuals may be at a higher or lower risk for COVID-19 than the general population and without the knowledge of who these individuals are and their risk behavior around COVID-19, it seems unlikely that they would factor into state mask mandate policy.

Percent Female and Percent White

Studies have shown high rates of smart phone ownership across gender and race (Pew Research Center 2021). The demographic composition of the group not captured by location tracking may vary, with potentially large differences in demographic composition for essential work by industry. Furthermore, we have found little research on who opts out of location sharing. For these reasons, we do not expect this omitted variable to have a relationship with Percent Female and Percent White.

Compliance with public mask mandate

Whether or not the public complies with state mandates to wear a mask is another omitted variable that cannot reasonably be proxied with other predictors in the state policies, mobility, or COVID-19 case features. This predictor cannot be directly measured. Additionally, there are degrees of compliance (e.g. wearing a mask but not covering one's nose or wearing a mask in some settings but not others) that hinder assessing the impact of this variable. Since we cannot introduce other predictors to proxy for this variable, we will focus on the expected bias from its omission.

We expect increased compliance with public state mask mandates to have a negative relationship with percent change in COVID-19 cases, due to the airborne nature of COVID-19 transmission and reduced opportunity for disease spread with proper mask wearing.

Thursday Retail and Recreation Mobility

Similarly, we expect a negative relationship with the difference in changes in mobility from baseline. We hypothesize that if residents of an area comply with the mask wearing mandate that they also comply with other public health protocols designed to reduce transmission such as limiting travel.

Since the expected correlation between the omitted variable and mobility is negative and the expected correlation between the omitted variable and percent change in COVID-19 cases is also negative, the direction of the bias is positive. Since the coefficient for mobility is positive, the omitted variable is making the coefficient more positive than it would otherwise be - the bias for this coefficient is away from 0 and the coefficient is overestimated. If this is the case, the percentage increase in COVID-19 cases due to Thursday Retail and Recreation Mobility may be an artifact of omitted variable bias - since the coefficient for this mobility predictor is overestimated, it may actually be zero.

However, since we are measuring percent changes in COVID-19 cases in each state over a short window of time, we believe that this omitted variable compliance with public mask mandate will experience little change from our starting time period to our final time period, so it will be consistent within each state.

Public Mask Mandate

We expect a positive relationship between the predictors Public Mask Mandate and the omitted variable mask compliance. In areas where a mask mandate is in place, we expect that the majority of individuals will choose to wear a mask while fewer individuals will choose to wear a mask if not explicitly required to do so.

Since the expected correlation between the omitted variable and Public Mask Mandate is positive and the expected correlation between the omitted variable and percent change in COVID-19 cases is negative, the direction of the bias is negative. Since the coefficient for public mask mandate is negative, the omitted variable is making the coefficient more negative than it would otherwise be - the bias for this coefficient is away from 0 and the coefficient is overestimated. If this is the case, the percentage decrease in COVID-19 cases due to Public Mask Mandate may be an artifact of omitted variable bias - since the coefficient for this mobility predictor is overestimated, it may actually be zero.

However, since we are measuring percent changes in COVID-19 cases in each state over a short window of time, we believe that this omitted variable compliance with public mask mandate will experience little change from our starting time period to our final time period, so it will be consistent within each state.

Percent Female

Several studies have found gender to be correlated with mask compliance. Females were found to be more likely to comply with public mask mandates (Haischer, Michael H., et al 2020), which may be due to differences in gender perception of face masks (Howard, Matt C 2021). Therefore, we expect that the omitted variable compliance with a public mask mandate will have a positive relationship with Percent Female.

Since the expected correlation between the omitted variable and Percent Female is positive and the expected correlation between the omitted variable and percent change in COVID-19 cases is negative, the direction of the bias is negative. Since the coefficient for Percent Female is negative, the omitted variable is making the coefficient more negative than it would otherwise be - the bias for this coefficient is away from 0 and the coefficient is overestimated. If this is the case, the percentage decrease in COVID-19 cases due to having a higher female population proportion may be an artifact of omitted variable bias - since the coefficient for the percent female predictor is overestimated, it may actually be zero.

However, since we are measuring percent changes in COVID-19 cases in each state over a short window of time, we believe that this omitted variable compliance with a public mask mandate will experience little change from our starting time period to our final time period, so it will be consistent within each state.

Percent White

Some studies have found that members of minority groups are more likely to comply with mask mandates than whites, which may be related to higher rates of COVID-19 infection and deaths among these groups (Hearne, Brittany N., and Michael D. Niño 2021),(The Day 2021). For this reason, we expect the omitted variable to have a negative relationship with the predictor Percent White.

Since the expected correlation between the omitted variable and Percent White is negative and the expected correlation between the omitted variable and percent change in COVID-19 cases is also negative, the direction of the bias is positive. Since the coefficient for Percent White is negative, the omitted variable is making the coefficient less negative than it would otherwise be - the bias for this coefficient is toward 0 and the coefficient is underestimated. If this is the case, the omitted variable compliance with public mask mandate does not impact the trustworthiness of the Percent White coefficient.

Ease and availability of testing

One of the more difficult metrics to track is the ease of testing within a state. In the early months of the pandemic some states were better than others at testing individuals. Because of such low availability of tests in many state and local governments, there were guidelines that you could not get tested unless you were showing symptoms. This variable is not captured in the mobility, state policy, or demographic features directly. Additionally, none of the available variables can adequately proxy for this omitted variable.

We believe this omitted variable has a positive correlation with the number of COVID-19 cases reported. The easier it is for an individual to get tested the more likely they are to go get a test. If it is more difficult, then many individuals may have COVID-19 but never get tested and their case will never get counted in the observations.

Thursday Retail and Recreation Mobility

We expect ease of testing to have a negative correlation with Thursday Retail and Recreation Mobility. We hypothesize that a state that is taking the pandemic more seriously is more likely to have better availability of testing. As well they will have residents that will be taking it more seriously and not being as mobile. Therefore, the higher the ease of testing the lower the mobility. Since the coefficient for Thursday Retail and Recreation Mobility is positive, this omitted variable would make the coefficient less positive than it would otherwise be. So we estimate that this omitted variable biases Thursday Retail and Recreation Mobility toward zero and the coefficient is underestimated. Therefore, the omitted variable ease of testing should not we have statistical significance for this mobility coefficient. This omitted variable could have a decent impact on our model because we are measuring changes over time. This is right when testing started to ramp up across the country, so it could be different from our starting time period to our final time period.

Public Mask Mandate

We expect ease of testing to have a positive correlation with public mask mandates. We hypothesize that states that are taking the virus more seriously will have a mask mandate in place and will have a higher ease of testing. Because the coefficient for mask mandate is negative, the effect of the bias from this omitted variable on Public Mask Mandate would be toward zero (underestimated). We believe this omitted variable could have a decent impact on our model because we are measuring changes over time. This is right when testing started to ramp up across the country, so it could be different from our starting time period to our final time period. We need to consider this when looking at the significance of the Mask Mandate Variable.

Percent Female and Percent White

We do not expect this omitted variable (ease of testing) to have a relationship with Percent Female and Percent White. No studies can be found that show a correlation between ease of testing and Percent Female and Percent White.

Population Density

The density of the population in the state is another omitted variable from our observations. This is something that varies greatly from state to state. States like New York have a denser population than states like Wyoming. We believe this will cause cases to vary from state to state. They could have similar mobility numbers but because of the density of the population the virus could spread much faster. This variable is not captured in the mobility, state policy, or demographic features directly. Additionally, none of the available variables can adequately proxy for this omitted variable. We expect this omitted variable to have a positive correlation with percent change in COVID-19 cases. The denser the population the easier this airborne virus can spread and the more COVID-19 cases they will have.

Thursday Retail and Recreation Mobility

Due to the higher risk of contracting the virus in dense populations strict lock-down guidelines were generally implemented and followed. Therefore, we hypothesize that population density is negatively correlated with mobility. If this is true, then the direction of omitted variable bias is towards zero and the coefficient is underestimated. So, we should be able to trust that we have statistical significance for this mobility coefficient after accounting for the impact of the omitted variable population density. As well we believe that this omitted variable would have minimal impact on our model, because we are measuring changes over time. There is little to no change in a state's population density from our starting time period to our final time period, so it will be consistent within each state.

Public Mask Mandate

Similar to the logic for mobility we believe that states with denser populations will be more likely to implement a mask mandate. This would mean that we have a positive correlation between population density and public mask mandates. Because the coefficient for mask mandate is negative, we estimate that this omitted variable has a bias towards zero for Public Mask Mandate (the coefficient is underestimated). In addition, we believe that this omitted variable would have minimal impact on our model, because we are measuring changes over time. There is little to no change in a state's population density from our starting time period to our final time period, so it will be consistent within each state.

Percent Female and Percent White

We do not expect this omitted variable (population density) to have a relationship with Percent Female and Percent White. No studies can be found that show a correlation between population density and Percent Female and Percent White for each state.

Asymptomatic

A portion of people that contract COVID-19 display no symptoms from the virus. Unfortunately, these people can still spread the virus to others. The percentage of asymptomatic people varies depending on many different factors like age and ethnicity, but no clear trends between them have been established. This makes

it difficult to track and quantify how many people are asymptomatic in each state, but this variable will have implications for our model because many people only get tested if they are showing symptoms, so a varying number of COVID-19 cases will not be captured in our observations.This variable is not captured in the mobility, state policy, or demographic features directly. Additionally, none of the available variables can adequately proxy for this omitted variable.

We expect the omitted variable proportion of COVID-19 infected individuals who are asymptomatic to have a negative correlation with the number of COVID-19 cases reported. We hypothesize that the more asymptomatic people in a state the less likely they are to go and get tested. If this is true, then many people who have COVID-19 will not have reported COVID-19 cases.

Thursday Retail and Recreation Mobility

We believe the more asymptomatic individuals the higher the mobility. If an individual believes they are healthy they are more likely to go out like normal. This would mean a positive correlation between proportion of COVID-19 infected individuals who are asymptomatic and mobility. If true, then this omitted variable biases the coefficient for Thursday Retail and Recreation Mobility towards zero (the coefficient is underestimated). So, we can trust the statistical significance of our coefficient for Thursday Retail and Recreation Mobility after accounting for the omitted variable proportion of COVID-19 infected individuals who are asymptomatic. This omitted variable would have minimal impact on our model, because we are measuring changes over time. There is little to no change in the number of asymptomatic individuals from our starting time period to our final time period, so it will be consistent within each state.

Public Mask Mandate

Similar to the logic for mobility we believe that states with more asymptomatic individuals will be less likely to implement a mask mandate. This would mean that we have a negative correlation between the number of asymptomatic individuals and public mask mandates. Because the coefficient for mask mandate is negative, we estimate that this omitted variable has a bias towards zero for Public Mask Mandate (the coefficient is underestimated) . We believe that this omitted variable will have minimal impact on our model, because we are measuring changes over time. There is little to no change in a state's number of asymptomatic individuals from our starting time period to our final time period, so it will be consistent within each state.

Percent Female and Percent White

We do not expect this omitted variable (population density) to have a relationship with Percent Female and Percent White. No studies can be found that show a correlation between population density and Percent Female and Percent White for each state. As well it varies from state to state.

Conclusion

Early public health guidance in the COVID-19 pandemic centered on limiting one's mobility and mask-wearing as the most effective strategies available for inhibiting disease transmission. Shutdowns of non-essential businesses were implemented across the nation and shelter-in-place mandates were enacted. For these reasons, we sought to investigate the impact of changes in mobility upon percent changes in COVID-19 cases in a causal model. We chose to use changes in retail and recreation mobility as our predictor given the political focus upon restaurants and storefronts and due to correlation between this measure of mobility and other mobility metrics in the Google mobility report which could introduce problematic collinearity if multiple mobility measures were used.

Contrary to public health guidance, we failed to find evidence in our $2nd$ main model that positive increases in mobility (as proxied by retail and recreation mobility) were associated with a percent increase in cases of COVID-19. This finding held true even when controlling for the presence of a public mask mandate. In our $3rd$ model, controlling for demographic variables observed to impact COVID-19 disease transmission (Percent Female, Percent White, Percent Under 24, and Percent Over 65), we also failed to find evidence that changes in retail and recreation mobility contributed to a percent change in COVID-19 cases for most metrics. Although the change for Thursday retail and recreation mobility demonstrated a statistically significant result in this full model (as measured by a p-value less than 0.05), this finding lacks practical significance. The

day does not coincide with the July 4*th* holiday and we have not found evidence that suggests widespread holiday closures on Thursday July 2*nd* which could result in increased mobility and disease spread. However, mobility may still explain changes in COVID-19 cases in different time periods or if the impact of changes in mobility has been suppressed by omitted variables such as those discussed in this report.

Conversely, we found that the implementation of a public mask mandate did support public health guidance on mask-wearing. When adding in additional demographic controls, the presence of a public mask mandate was observed to result in a 14% reduction in percent change in COVID-19 cases, holding all else constant. Similarly, Percent Female and Percent White were also observed to result in a reduction in the percent change in cases - approximately 9% and 1% respectively (holding all else constant). Although these findings may be artifacts of omitted variable bias, they do support other studies on the relationship between mask-wearing, demographic controls, and the spread of COVID-19. Therefore, further research to understand the relationship between gender, race, and disease transmission may prove valuable.

Sources

Centers for Disease Control and Prevention. 2020a. "COVID-19 and Your Health." Online Article. www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html.

———. 2020b. "Healthcare Workers." Online Article. www.cdc.gov/coronavirus/2019-ncov/hcp/faq.html#:%7E:text=Based%20on%20existing%20literature%2C,2%E2%80%9314%20days.

Google. 2021. "COVID-19 Mobility Reports." Data Set. www.google.com/covid19/mobility.

Haischer, Michael H., et al. 2020. "Who Is Wearing a Mask? Gender-, Age-, and Location-Related Differences During the Covid-19 Pandemic." *PLOS ONE* 15 (10): 186–93. www.ncbi.nlm.nih.gov/pmc/articles/PMC7561164.

HBS Working Knowledge. 2020. "The Covid Gender Gap: Why Fewer Women Are Dying." Online Article. hbswk.hbs.edu/item/the-covid-gender-gap-why-fewer-women-are-dying#:%7E:text=Research%20shows%20that%20men%20are,reluctance%20to%20seek%20health%20care.

Hearne, Brittany N., and Michael D. Niño. 2021. "Understanding How Race, Ethnicity, and Gender Shape Mask-Wearing Adherence During the Covid-19 Pandemic: Evidence from the Covid Impact Survey." *Journal of Racial and Ethnic Disparities*. pubmed.ncbi.nlm.nih.gov/33469866.

Hlavac, Marek. 2018. "Stargazer: Well-Formatted Regression and Summary Statistics Tables. R Package Version 5.2.2." Computer Program. https://CRAN.R-project.org/package=stargazer.

Honolulu Civil Beat. 2020. "VIRUS Tracker — July 27: 28 New Covid-19 Cases; Bar Clusters Emerge." Online Article. https://www.civilbeat.org/2020/07/coronavirus-daily-news-hawaii/.

Howard, Matt C. 2021. "Gender, Face Mask Perceptions, and Face Mask Wearing: Are Men Being Dangerous During the Covid-19 Pandemic?" *Personality and Individual Differences* 170: 110417. www.ncbi.nlm.nih.gov/pmc/articles/PMC7543707/#:~:text=Recent%20popular%20press%20authors%20have,during%20the%20COVID%2D19%20pandemic.&text=Therefore%2C%20although%20gender%20does%20not,relate%20to%20face%20mask%20perceptions.

Inside Higher Ed. 2020. "Blame Game." Online Article. www.insidehighered.com/news/2020/08/24/college-covid-strategies-dont-adequately-address-typical-student-behavior.

Lipsitch, Marc, DPhil. 2021. "Seasonality of Sars-Cov-2: Will Covid-19 Go Away on Its Own in Warmer Weather?" Online Article. https://ccdd.hsph.harvard.edu/will-covid-19-go-away-on-its-own-in-warmer-weather/.

Mayo Clinic. 2020. "Coronavirus Infection by Race: What's Behind the Health Disparities?" Online Article. www.mayoclinic.org/diseases-conditions/coronavirus/expert-answers/coronavirus-infection-by-race/faq-20488802.

Newton, Casey. 2020. "Google Uses Location Data to Show Which Places Are Complying with Stay-at-Home Orders — and Which Aren't." Online Article. www.theverge.com/2020/4/3/21206318/google-location-data-mobility-reports-covid-19-privacy.

NYC Business. 2021. "NYC Restaurant Reopening Guide." Online Article. www1.nyc.gov/nycbusiness/article/nyc-restaurant-reopening-guide.

Olmsted Medical Center. 2021. "Wearing a Mask." Online Article. www.olmmed.org/covid-19-information/how-to-wear-a-mask.

Pew Research Center. 2021. "Mobile Fact Sheet." Online Article. www.pewresearch.org/internet/fact-sheet/mobile.

Raifman, J., et al. 2020. "COVID-19 Us State Policy Database (Cusp)." Google Docs. github.com/nytimes/covid-19-data.

The Day. 2021. "Your Politics and Race Predict Your Likelihood of Wearing a Mask." Online Article. www.theday.com/article/20200515/OP03/200519635.

The New York Times. 2021. "Coronavirus (Covid-19) Data in the United States." Git Repository. github.com/nytimes/covid-19-data.

US Census Bureau. 2019. "ACS Demographic and Housing Estimates." Data Set. data.census.gov/cedsci/table?q=ACS&g=0100000US.04000.001&tid=ACSDP1Y2019.DP05&moe=false&hidePreview=true.