STATISTIC 198 FINAL PROJECT SUMMER 2020: Investigating the Association Between COVID-19 Response and Climate Change Sentiment in the U.S.

Anabella Buckvar, Chandler Naylon, William Reynolds, Natalie Wong

Due: Wednesday, July 29, 2020 at 11:59p

###Introduction and Data In recent months, the world has been bombarded with a public health crisis from the COVID-19 pandemic. Although the reality and severity of the public health issue has been proven by scientists around the world, many Americans are still skeptical of whether or not COVID-19 is "real" and if it is truly something to be taken seriously. Similarly, climate change is another major crisis that our world is facing – and like COVID-19, many are still doubtful of the reality of global warming despite it being proven and confirmed by scientists, time after time. Given the similar scientific backing behind both crises and the surprisingly widespread amounts of disbelief about both issues, especially among United States residents, our project group was curious to see if the two types of "denial" are associated in any way. Thus, our research question asks, is a state's societal response to COVID-19 a strong indicator of its overall climate change sentiment?

The general research question that we strove to answer is whether there is a relationship between COVID-19 denial and climate change denial. We felt that many people who do not believe in climate change or fail to recognize its severity have a distrust in science, and we thus also wondered if they would doubt the scientific validity of COVID-19. Specifically, is a state's societal response to COVID-19 a strong indicator of its overall acceptance of the issue of human-caused climate change? When referring to "how seriously" state residents are taking either the climate change or COVID-19 crisis, we are referring to the extent of residents' responses to the COVID-19 pandemic and state regulations as well as their opinions about the presence and severity of climate change. Our data was obtained from numerous sources. We used one single data set for our climate change data, but combined multiple data frames to garner information about COVID-19 related actions and behaviors for each state.

The data we used to gain insight on climate change sentiment is from the Yale Project on Climate Control Communication, which estimated beliefs, risk perceptions, and policy preferences at both state and local levels for US citizens after surveying over 24,000 US citizens during the spring of 2019. The survey asked 29 questions regarding whether or not climate change is real, to what extent it is happening, how it is affecting humans, and how the government and other stakeholders should respond to it. The dataset has over 61 variables. However, for the sake of our project, we only included the 6 variables that related to the survey questions that we found to be most relevant for our research question. The variables we used from this dataset included human, consensus, worried, futuregen, personal, and harmUS. These variables represent the percentage of residents surveyed in the Yale Climate Change survey who think global warming is caused by human activity (human), who believe that most scientists think global warming is happening (consensus), are worried about global warming (worried), think that global warming will harm future generations (futuregen), think that global warming will harm them personally (personal), and who think that global warming will harm people in the United States (harmUS).

We utilized 3 different raw datasets to gain information about COVID-19 actions and behaviors in each US state.

One of the sources of COVID-19 data that we utilized in our project was from the Kaiser Family Foundation's (KFF) dataset on State Data and Policy Actions to Address Coronavirus. This set provided insight into a

given state's social distancing actions and gave us an idea of "how seriously" a state was taking COVID-19. For each state, it gave information on the status of reopening, details on a state's stay at home order, whether or not there's a mandatory quarantine policy for travelers, the state of nonessential businesses closures, and many more. Ultimately, these actions are quantified in order to give each state a social distancing score. There are 39 variables in this dataset, but we only used 2 of the most pertinent ones for our research purposes. The 2 variables that we used were bar_closures and status_of_reopening, which denoted if a state had closed, newly closed, reopened, or implemented new restrictions in bars in their state. and status_of_reopening, which gave information about whether a state was proceeding with reopening, pausing new reopenings, implementing new restrictions, or completely reopened. The information used in this data set was derived largely from state government websites and sources, and this dataset consolidated them all into one place. Because the data is taken from state government sources, we are largely confident that it is reliable. Furthermore, this data is being updated regularly, so we are confident that it is relevant to the present time period.

We also utilized Google's COVID-19 Community Mobility Reports. This source offers daily insight into how mobility in community spaces has changed during COVID-19. It measures how many people went to a given location and how much time they spent there, and compared those levels to a baseline day before COVID-19, between January 3rd and February 6, 2020. Areas addressed include retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential spaces. The data shows how movement and interaction patterns in such community spaces have changed.

This dataset provided information on community mobility for many different countries, but we only used the data given for the United States in our project. This dataset included 6 variables, all of which we used in our model. These included Retail_and_recreation_percent_change_from_baseline, grocery_and_pharmacy_percent_change_from_baseline, Parks_percent_change_from_baseline, Transit_stations_percent_change_from_baseline, Workplaces_percent_change_from_baseline, Residential_percent_change_from_baseline, which measured the percentage amount of change in mobility in the retail and recreational locations, grocery stores and pharmacies, public parks, public transit stations, workplaces, and residential areas respectively before and during COVID-19.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Linking to GEOS 3.5.1, GDAL 2.2.2, PROJ 4.9.2
```

###Methodology After formulating our research question, we outlined the procedure that we would use to address it. Essentially, we procured relevant information from each COVID-19 related data set and merged it into a single data frame, and then modeled these COVID-19 variables into a multiple linear regression that would output a climate score created by a function that combined the various climate change variables we used. This climate score was a response variable used to predict outcomes later in our research. These steps are outlined below.

First, we did a good deal of data wrangling to procure the necessary and relevant information from our

datasets. We first modified the Google COVID-19 Community Mobility Report to only show data for the United States, since it contained data for countries around the world. Then, we selected data points that corresponded to entries from March 15 to July 23rd, which was a time frame that we felt corresponded to the United States time frame of the COVID-19 crisis. Next, we took the mean value of each predictor per state within that time frame. We then merged this dataset with the KFF State Social Distancing Actions dataset and the Yale Public Opinion Estimates on Climate Change dataset, leaving us with a consolidated csv to model with. This was all done in Excel using the VLOOKUP function, then uploaded to RStudio. These variables would serve as the input predictors in the linear regression model.

We eliminated some of the variables that we found to be less pertinent, including the variables related to whether a state implemented stay at home orders, mandatory quarantine for travelers, nonessential business closures, bans on large gatherings, limits on restaurant capacity, closing bars, requiring face coverings, and whether a state postponed primary elections. The names of the variables in the data set were stay_at_home_order, mandatory_quarantine_for_travelers, non_essential_business_closures, large_gatherings_ban, restaurant_limits, bar_closures, face_covering_requirement, and primary_election_postponement. We eliminated these specific variables because we reasoned that many of them would be collinear with other ones in the overall dataset, and would thus be unnecessary and possibly negatively affect our model. We also eliminated the variable emergency_declaration that denoted whether states declared emergency or not, since all states had answered "yes."

Next, we created a function that would create a "climate score," denoted by the variable name score.y, for each US state. We identified several questions from the Yale Public Opinion Estimates on Climate Change dataset, using this function was created by taking the average of the percent of people that responded affirmatively to questions about climate change responses to climate change questions. This score would quantify how seriously each state's residents view the issue of climate change to be, based on the percentage of residents surveyed in the Yale Climate Change survey who think global warming is caused by human activity (human), who believe that most scientists think global warming is happening (consensus), are worried about global warming (worried), think that global warming will harm future generations (futuregen), think that global warming will harm them personally (personal), and who think that global warming will harm people in the United States (harmUS). After creating the climate score for each state, we used the mutate function in R to add each state's climate score variable, titled score.y, into our dataset. This "climate change score" would serve as the output of the linear regression model.

We then created a multiple linear regression model in R. This regression took the COVID-19 mobility and behavioral predictors as the inputs, and output the climate regression score.

Next, we set seed using the randomly generated value of 100 so that all of our data and figures would be reproducible. We then created a training set with 80% of the data, as well as a test set containing 20% of the data. This ultimately allowed us to fit a linear model that predicted the climate score based on the COVID-19 predictors.

```
##     State              score
##  Length:51          Min.   :41.55
##  Class :character   1st Qu.:49.16
##  Mode  :character   Median :53.44
##                     Mean   :53.34
##                     3rd Qu.:56.72
##                     Max.   :66.53
```

###Our Results In order to answer our research question about if a state's societal response to COVID-19 was associated with residents' opinions on climate change, we utilized a multiple linear regression to test the relationship between a state's climate change score and numerous predictors that measure its societal response to COVID-19.

To determine the appropriateness of this model, we verified all assumptions for the linear regression model. We created a residual plot that showed that both the assumptions for equal variances and linearity were met, since the residuals were fairly symmetric about the x-axis and showed no clear pattern. A histogram of

residuals was also created and demonstrated that the Normality assumption was also met, since the histogram was fairly Normally distributed. Although the histogram was slightly left-skewed, overall, we deemed that it was Normally distributed enough to meet the Normality assumption.

Initially, we created a linear regression model that included 8 COVID-19 related predictors, including grocery_and_pharmacy_percent_change_from_baseline, workplaces_percent_change_from_baseline, status_of_reopening, and residential_percent_change_from_baseline, bar_closures. However, we found that while the overall F test was significant at the alpha level of 0.05, none of the individual predictors were significant. Although collinearity doesn't affect the prediction accuracy of our test, we were curious to see its effect on the statistical inference. We suspected that this lack of significance might be due to collinearity.

We then calculated the variable inflation factor (VIF) to measure the amount of collinearity in the set of multiple regression variables used in our model. We determined that any predictor with a GVIF score of over 5 should be removed from the dataset, since such variables might have too much collinearity and could negatively affect the model. After running a test to measure the variable inflation factor for each of the variables, we found that numerous variables were indeed collinear and had high VIF values, with some VIF values even being over 10.

The variables we removed included Grocery_and_pharmacy_percent_change_from_baseline, Workplaces_percent_change_from_baseline, Status_of_reopening, and Residential_percent_change_from_baseline. We removed Grocery_and_pharmacy_percent_change_from_baseline because it was insignificant with a p-value of 0.0691 in the final model and had a negligible estimated slope of -0.048, while also being collinear with a VIF value of 6.804. The status_of_reopening was also removed from the model because the individual p-values for all three dummy variables were all well above the alpha = 0.05 level. Therefore, we have insufficient evidence to conclude that the expected slopes corresponding to "Paused", "Proceeding with Reopening", and "Reopened" are different than the baseline of "New Restrictions Imposed," while holding all other variables constant. We thus decided against including status_of_reopening as a predictor in our final model. Removing this predictor made sense to us through a public health lens: states recovering (i.e. New York) and states reopening with rising cases (Florida and Texas) both made those decisions with very different health situations—mostly due to the time they were affected.

Furthermore, we also removed Residential_percent_change_from_baseline because it was found to be highly collinear with retail_and_recreation_percent_change_from_baseline. When residential_percent_change_from_baseline was included in the model, it had a very high VIF value of 10.881 and an individual p-value that was insignificant (0.485). Once it was removed from the model, the individual p-value corresponding to retail_and_recreation_percent_change_from_baseline was 0.012, and thus significant at the alpha = 0.05 level and had a much lower VIF value of 5.340. This proved that the residential_percent_change_from_baseline variable was causing a great deal of collinearity with the retail_and_recreation_percent_change_from_baseline, which is why we ultimately decided to leave it out of the final model.

We then ran the model again after removing these highly collinear variables. We also removed some of the predictors that we ruled insignificant after performing sensitivity analysis multiple times. We performed sensitivity analysis by running the same model without the predictors that had high VIF values, and continued to repeat the process until we had a model with a reasonable R^2 value of 0.845, a root mean squared error (RMSE) of 2.727, and predictors that were less collinear and had lower VIF values.

We deemed this to be a comprehensive and effective multiple linear regression model to measure the association between climate change denial and COVID-19 denial.

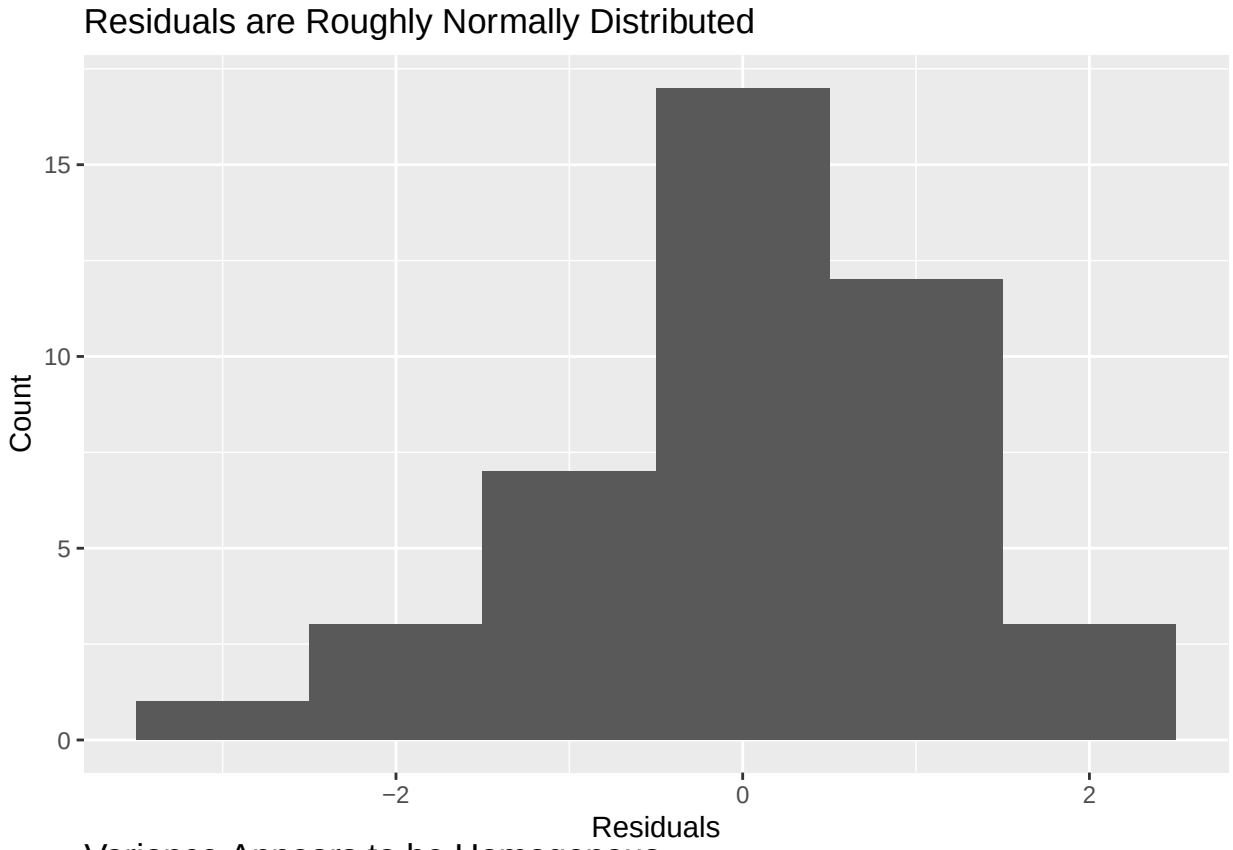
```
## # A tibble: 7 x 5
##   term                                  estimate std.error statistic  p.value
##   <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                             43.9      2.09       21.0   8.54e-22
## 2 transit_stations_percent_change_from_ba~ -0.165    0.0429     -3.86  4.57e- 4
## 3 bar_closuresNew Service Limits           0.947    2.00        0.474 6.39e- 1
## 4 bar_closuresNewly Closed                -2.99     1.36       -2.20  3.44e- 2
```

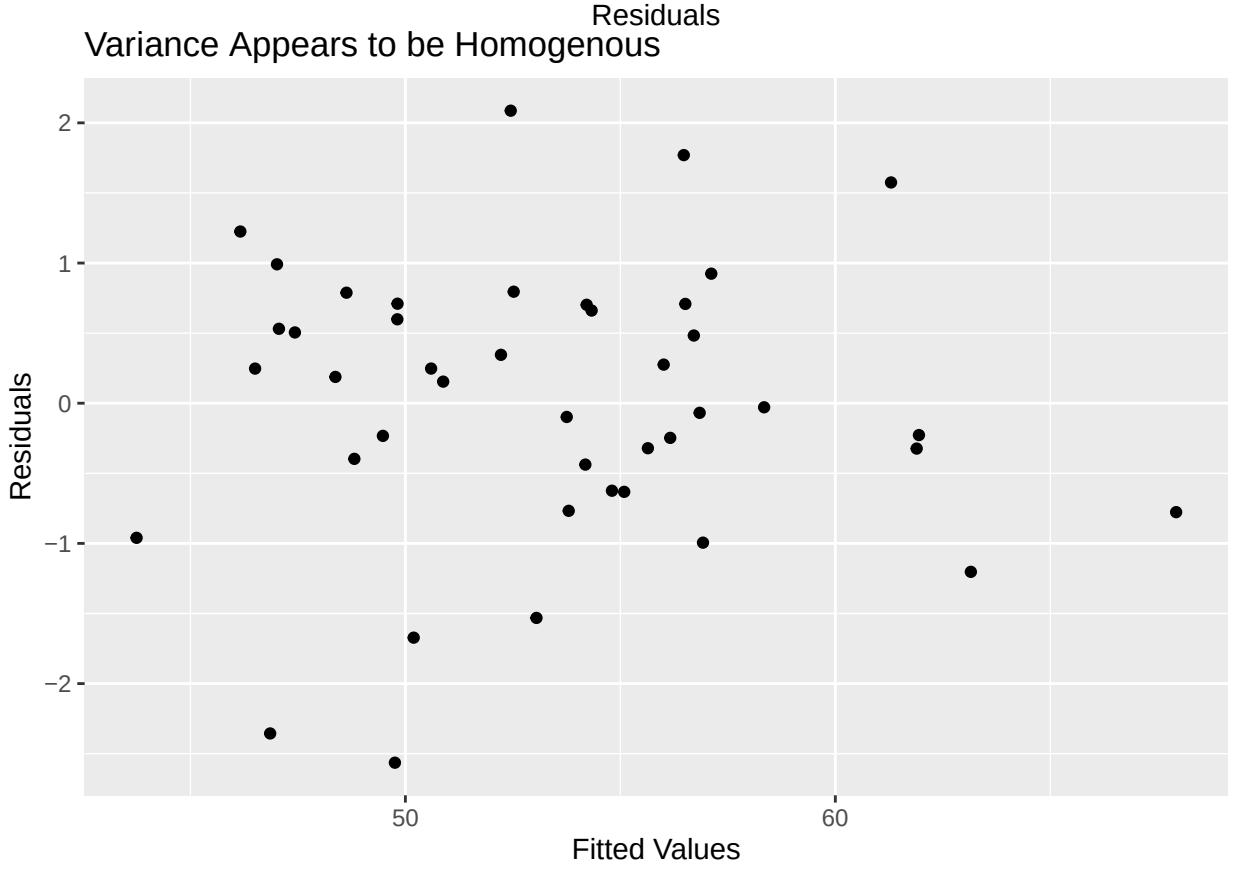
```
## 5 bar_closuresReopened                       -1.83     1.04       -1.76  8.76e- 2
## 6 retail_and_recreation_percent_change_fr~  -0.235    0.0892     -2.63  1.24e- 2
## 7 parks_percent_change_from_baseline         0.0136   0.00978     1.39  1.73e- 1
```

```
##
## Call:
## lm(formula = score.y ~ transit_stations_percent_change_from_baseline +
##     bar_closures + retail_and_recreation_percent_change_from_baseline +
##     parks_percent_change_from_baseline, data = train.data, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0369 -1.1403  0.3581  1.4495  4.4333
##
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                     43.928161   2.091018  21.008
## transit_stations_percent_change_from_baseline   -0.165444   0.042895  -3.857
## bar_closuresNew Service Limits                   0.947292   1.999916   0.474
## bar_closuresNewly Closed                        -2.985548   1.357571  -2.199
## bar_closuresReopened                            -1.825907   1.039849  -1.756
## retail_and_recreation_percent_change_from_baseline -0.234965   0.089215  -2.634
## parks_percent_change_from_baseline               0.013579   0.009779   1.389
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## transit_stations_percent_change_from_baseline   0.000457 ***
## bar_closuresNew Service Limits                  0.638598
## bar_closuresNewly Closed                        0.034373 *
## bar_closuresReopened                            0.087605 .
## retail_and_recreation_percent_change_from_baseline 0.012370 *
## parks_percent_change_from_baseline              0.173490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.404 on 36 degrees of freedom
## Multiple R-squared:  0.8453, Adjusted R-squared:  0.8195
## F-statistic: 32.78 on 6 and 36 DF,  p-value: 3.57e-13
```

```
## [1] 2.526815
```

```
##                                                    GVIF Df GVIF^(1/(2*Df))
## transit_stations_percent_change_from_baseline   5.103388  1        2.259068
## bar_closures                                    1.500989  3        1.070031
## retail_and_recreation_percent_change_from_baseline 5.340161  1        2.310879
## parks_percent_change_from_baseline              1.786698  1        1.336674
```





###Discussion After conducting our project procedures, we saw that our model was indeed effective in

conducting such analysis, as we had a final R^2 value of 0.845, which we found to be fairly high and indicated that there is indeed a relationship between a state's COVID-19 societal response and their beliefs in climate change. The model also yielded a p-value of 3.57e-13, which showed that at least one of the predictors was significant, or not equal to 0. This is affirmed by the fact that we had 3 significant predictors in our model, which included transit_stations_percent_change_from_baseline (p-value=0.000457), bar_closuresNewlyClosed (p-value=0.0343), and retail_and_recreation_percent_change_from_baseline (p-value=0.0123). These 3 predictors also showed acceptable VIF levels for collinearity, including 5.103, 1.500, and 5.340 respectively.

Although we are ultimately satisfied with the efficacy and output of our model, we also recognize that there is also room for improvement! Some of the possible weaknesses of our model are described and explained below.

In terms of the research question itself, two of the COVID-19 behaviors that we measured (reopening status and bar closures) were determined and decided upon by state governments rather than state residents themselves. Thus, they reflect the state government's views on the pandemic, but not necessarily those of the individual residents themselves. Since the climate change sentiment was based on the beliefs of individual residents, there could be slight discrepancy between what the government decided in terms of COVID-19 actions and the resident's individual beliefs of climate change. Although we don't believe that the effect of this would be obstructive or constraining, it is something to take into account while analyzing and discussing results.

Furthermore, the Yale Climate Change data that we utilized to measure sentiment and beliefs about climate change was published last spring. Although this is still fairly recent, it would have been more ideal if we could find data that was more accurate and comprehensive. Since this dataset gauged opinions on climate change before COVID-19, it does not take into account any possible effects that the pandemic may have had on people's climate change opinions. However, this could be seen as a positive aspect, since this could reduce potential confounding between the two perceptions.

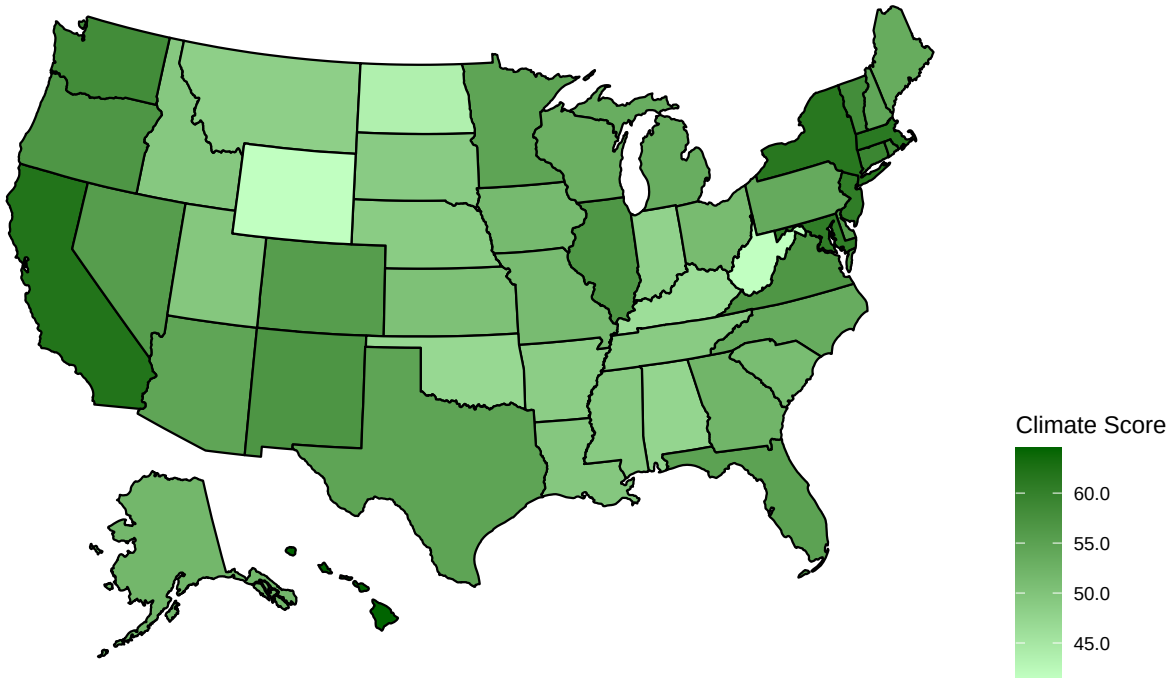
Moreover, the Google COVID-19 Community Mobility Reports has multiple potentially confounding factors that could negatively affect our overall study. Vacations and public holidays can help you understand what your community looks like when people don't go to places of work. -> might affect workplace and residential percentages. In many regions, you can see this clearly in the Parks category—visitors to parks are heavily influenced by the weather. Remember that these mobility reports show relative changes, and not absolute visitors or duration. For example, if few people normally visit places of work on a Sunday, you wouldn't expect to see large changes to Sunday visitors as your community responds to COVID-19.

Another issue that we encountered was collinearity. Numerous variables that we used in our model were likely to be highly correlated. Our model ultimately consisted of predictors with acceptable VIF values, but to achieve this we had to remove those with excessively high values (above ~6).

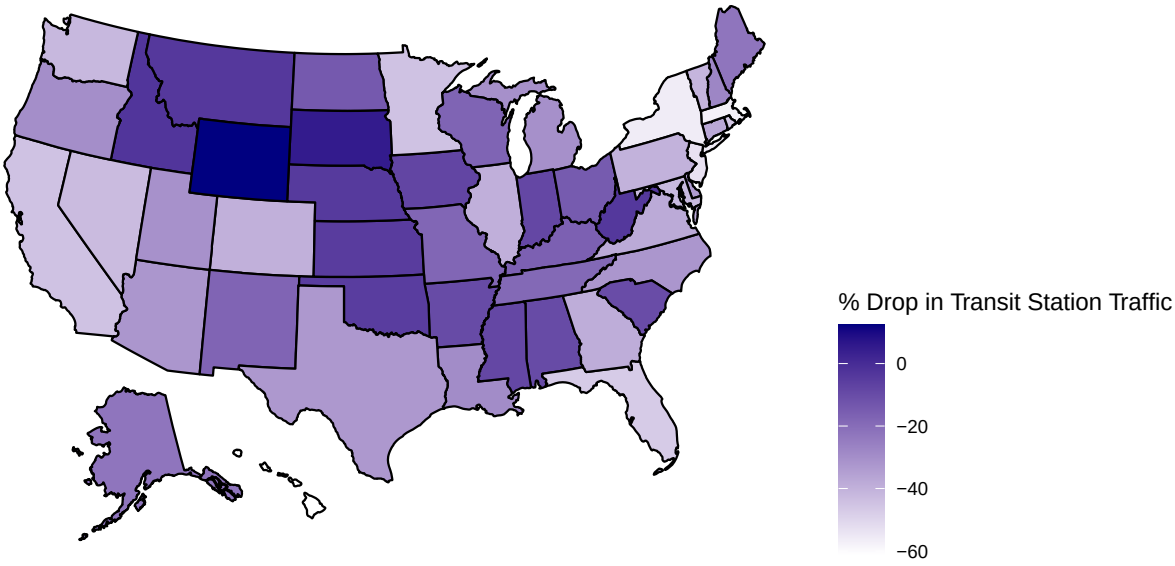
If we were to or able to start the project over, we would ideally use climate change data from 2020 rather than 2019, so that the climate change sentiment data lines up chronologically with the COVID-19 data. However, such data is currently unavailable to the public (at least from our research).

Furthermore, if we were to continue working on this project, it would be interesting to investigate the effects of more predictors. For example, we were curious about the correlation between states who mandated mask wearing and how those states' residents perceived climate change, but there was not enough data to make meaningful inference about such a relationship. If we were to continue this project further into the year when there would be more available data about such predictors, we would most definitely include those into our model to investigate their effects on climate change perception.

Hawaii, California, New York, and Massachusetts Have the Highest Climate Scores in the U.S.



Hawaii, Massachusetts, New York, and New Jersey Have the Largest % Percent Drop in Transit



###Works Cited "COVID-19 Community Mobility Report." COVID-19 Community Mobility Report, https://www.google.com/covid19/mobility?hl=en. Accessed 30 July 2020.

Jul 29, Published:, and 2020. "State Data and Policy Actions to Address Coronavirus." KFF, 29 July 2020, https://www.kff.org/coronavirus-covid-19/issue-brief/state-data-and-policy-actions-to-address-coronavirus/.

Multicollinearity Essentials and Vif in r - Articles - Sthda. http://www.sthda.com/english/articles/39-regression-model-diagnostics/160-multicollinearity-essentials-and-vif-in-r/. Accessed 30 July 2020.

"Yale Climate Opinion Maps 2019." Yale Program on Climate Change Communication, https://climatecommunication.yale.edu/visualizations-data/ycom-us/. Accessed 30 July 2020.