Data C102 Final Project

Jillian Criscuolo, Eliza van Hamel Platerink, Noah Prozan, Zoe Parcells

May 10th 2021

1 Data Overview

1.1 2021 US Region Mobility Report

The 2020-21 US Region Mobility Report dataset we used was provided by Google collected through Google Maps. The data shows how traffic en route to specific, categorized places changed compared to prior to the pandemic, with a baseline pre-pandemic day being represented by the median value of the 5-week period Jan 3 – Feb 6, 2020. It is significant to note that these dates were an extraordinary time for the US economy, meaning increased retail and workplace movement, and may not be entirely representative of all mobility before the pandemic (i.e, the baseline may look different if a five week period from June-July 2019 was used). Each row represents movement trends compared to the determined "baseline" across different categories of places for a sub-region of the United States on a single day.

We decided to merge this dataset on date, sub-region 1, sub-region 2 with the US Counties dataset that was provided by the New York Times. By incorporating this dataset, we were able to access the daily number of new cases and deaths, the seven-day rolling average, and the seven-day rolling average per 100,000 residents for each state. The sample of data is drawn from New York Times journalists using county, state, and national government sources to compile these average cases and deaths. Each row represents a different county with totals of how many cases, deaths, and averages of cases and deaths per 100,000 residents were recorded there each day.

Moreover, the data from the 2020-21 US Region Mobility Report is drawn from a sample of the population of drivers on the road. Google Map users are aware that their location data was being used as only sets of data from users who have turned on the Location History setting were included. This setting is off by default, meaning that consent is insured but also that a large sector of the population (non-users and users with their Location History setting off) were not accounted for. This creates a relevant concern for convenience sampling as well as selection bias as only those with smartphones that can host Google Maps are represented, systematically excluding those without access to technology. This is important to note when looking at the results of this report as essential workers in the height of the pandemic primarily came from low-income communities who may have not have had access to Google Maps and therefore are not accounted for in the data.

In terms of features/columns that would have been helpful, had they been available in this dataset, are vaccine information and percent of the population that was tested. How many vaccines per state have been administered as well as how many people were getting tested for COVID in each state would have been helpful as these are two potentially confounding variables. This information would have provided additional insight into what factors contributed to how mobility was affected during the pandemic.

2 Exploratory Data Analysis (EDA)

We start with cleaning our data!

After some investigation, we quickly realized that we were not going to need all of the columns, or that they were not going to be exceptionally helpful. We dropped these columns accordingly. We also created a dataset where we dropped the NAN values for the Retail column, as well as the covid_avg and deaths_avg columns. We did this in order to ensure a successful Causal Inference later in the notebook. The removal of the NAN values was reasonable, especially for the covid_avg column and for the deaths_avg column, because in the early stages of the pandemic or even pre-pandemic, states and counties were not reporting this data. In some cases this data was not even existent, as COVID-19 was non-existent in these places. We also realized that there were places in the data set where we needed to remove NAN values, specifically in the place where there was no sub_region_1, or state label. We did this in order to ensure that we were accurately analyzing the relationship between different states, as our question entails. Removing these values allows for this discussion and deep dive into the difference in mask mandates/COVID restrictions to ensue.

The decision to remove the columns will not influence our analysis later down the line because they were not related to the questions we were trying to answer. In terms of the NAN values, these were not influential to the analysis because the NAN values mainly occurred in the earliest stages of the pandemic, or even the pre-pandemic stages. This means that the NAN values were before anything was really understood about restrictions or even COVID-19 in general.

We also split the data into different subgroups by state in order to allow for further discussion about the restrictions. The information about high vs low COVID restrictions was taken from the AARP website about the statewide mask mandates in each state. There are 25 states who either never had a statewide mask mandate, or have removed the mask mandate. There are 25 states with high restrictions, who still have a state-wide mask mandate in place.

We finally created a smaller data set that consists of two states with two relatively high COVID restrictions (California and New York), and two states with relatively low COVID restrictions (Alabama and Texas). We did this for mainly visual purposes.

2.1 Heatmaps

We first wanted to see the general trends in percent change from baseline by state. For each listed category (retail/recreation, parks, etc.) we grouped the available data by state, ultimately plotting the mean percent change from baseline for the years 2020 and 2021, respectively.

2.2 Transit

Based on the two heatmaps, it is clear that states that have a there is a significant difference in transit between states. We wanted to analyze the difference in two states that had particularly high COVID restrictions (CA and NY) and two states that had typically low COVID restrictions (AL and TX).

Texas removed the mask mandate on March 10 (68 days after January 1) Alabama removed the mask mandate on April 9 (98 days after January 1)

From the two graphs above, we can see a clear grouping between the states with high COVID restrictions and their transit percent change, and the states with low COVID restrictions and their transit percent change. We hope to do some investigation into why this happens!

2.3 Retail

Similarly to transit, we decided to explore retail and recreation further after seeing significant variance between different states.

From the two graphs above, we can see a clear grouping between the states with high COVID restrictions and their retail and recreation percent change, and the states with low COVID restrictions and their retail and recreation percent change. We hope to do some investigation into why this happens!

3 Research Questions

3.1 Question 1

The first question we seek to address is whether or not having high vs low COVID restrictions affect retail and recreation mobility? This question can be used to form decisions about COVID restrictions and whether or not restrictions truly affect mobility within a county. We have chosen to use the method of causal inference for this question to formally prove correlation equaling causation. This method allows us to determine the true effect of an independent phenomenon, in this case high vs low COVID restrictions, on the change observed in retail and recreation mobility.

3.2 Question 2

The second question we are looking at is can we predict retail and recreation mobility from states with low vs high COVID restriction as well as can we classify states as low or high COVID restriction states based on mobility? This question could help shape decisions for business owners and determine the feasibility of reopening as each tier of reopening is shaped by different levels of capacity. We have chosen to use Generalized Linear Models (GLMs) for this prediction because it allows us to generalize regression models while finding the posterior distribution over the coefficients for models that may be nonlinear after choosing

a likelihood distribution and link function. Therefore, predictions with GLMs are a good fit as they allow us to fit our model more precisely. We also wanted to predict whether a state had low or high COVID restrictions based on retail and recreation mobility data. This question could inform decision making about what level of restrictions are needed in a region based on mobility and death rate. We chose to use the random forest method of Bootstrap Aggregating (BAGGING) as each tree in the model is trained on a bootstrap/resampling of the data, capturing a different aspect of the data from each model and making it less likely to overfit.

4 Technique 1

We addressed our first research question by use of Causal Inference.

4.1 Methods

The variable which corresponds to treatment is the level of restriction in each state. We split the data into the high vs low restriction states (25 states with high restrictions, treatment variable marked FALSE, and 25 states with low COVID restrictions, treatment variable marked TRUE. We differentiated the relationship between high vs low restrictions using aforementioned AARP link[4]), and are investigating the relationship between the mask mandate, and the non mask mandate in the four different states. The variables which correspond to the outcome are the change in retail and recreation.

A variable that we believe is a confounder is the average number of COVID cases and COVID deaths. We believe that these are confounders because if, per capita, COVID is better or worse in some states, it will likely influence their restrictions and their abilities. The unconfoundedness assumption holds because the sheer quantity of data we have removes bias from the variables.

To adjust for confounders, we will use inverse propensity weighting. The very nature of IPW score to estimate helps us with our confounders!

4.2 Results

Since this estimate of the treatment effect is positive, it indicates that having low COVID restrictions and no mask mandates improves the recreation and retail percent change from average compared to those with high COVID restrictions. This means that states that have low COVID restrictions have a higher percentage of people visiting stores and frequenting recreational activities than those that have high COVID restrictions. The increase in the naive estimator indicates that our confounders are actually confounders and have an impact on the outcome.

To adjust for uncertainty, we did a bootstrap analysis of the IPW. Demonstrated above, we can clearly see that our estimate for the average treatment effect is fairly consistent throughout all samples. Our sample lies within our confidence interval as well!

4.3 Discussion

In order to exhibit Simpson's paradox, there needs to be a trend when the data is taken from a fine grain, and as the data gets coarser and coarser, we see a different trend. This analysis would be hard to do, given that our data is so finely separated (states, counties, AND dates!). If we had access to a nationwide, or even a statewide sample, as opposed to such small sample sizes, we would be able to do this analysis. Taking an average of the whole sample would also not work in this analysis because we need the whole group numbers, not an average..

The limitations of this method are that we have to have a finite number of samples. We also have to have the unconfoundedness assumption hold.

A confounder that we cannot account for, given the lack of data on the subject, is the vaccine availability. If we had this information, we might be able to answer more questions, and have more information in regards to why the retail and recreation numbers are higher, or why these specific states have low restrictions. Maybe the amount of people who are vaccinated also impacts the number of people who frequent the shops!

We are fairly confident that there is a causal relationship between our chosen treatment and outcome. In the early stages of the pandemic especially, when the COVID restrictions were high among almost all of the states, the percent change from these baselines was fairly similar, as evident by the very beginning of the graphs of 2020-21 in section 2.2 and 2.3. However, as the states took their own decision with how lenient they would be with the restrictions (closure of shops, mask mandates, closure of public parks, or, on the other hand, no closure/mask mandate), the difference in states becomes more and more obvious.

The states that have high COVID restrictions tend to form a conglomerate that follows a similar trend of lower percent changes, while those with low COVID restrictions have a percent change that is closer to the baseline.

5 Technique 2

We addressed our second research question by use of a Generalized Linear Model.

5.1 Methods

We are trying to predict the percent change from baseline of retail and recreation after the release of the mask mandates. We are using 1 feature, which is the percent change from baseline of retail and recreation starting from 2021. We use 2021 as a start date since the vaccine rollout became more prominent at the start of the new year. We believe 2020 data is not as significant for prediction of mobility for 2021on a day-day basis.

We will be using a Gaussian GLM. We are using a Gaussian GLM because it more closely aligns with our prior, in the sense that the percent change from baseline is increasing as a function of the date. As we move further into 2021, we expect more people will go out to stores as there are more vaccinations. Moreover, we are dealing with percentages, which are non-discrete values, and this aligns with a Gaussian GLM. The assumptions we are making is that we are unlikely to see values more than 3 away from the mean. That means that we're implicitly assuming that the vast majority of y y-values we see will be within 3 of the mean (i.e., the prediction X).

The nonparametric method we will be using is random forest. We chose to use the random forest method of Bootstrap Aggregating (BAGGING) as each tree in the model is trained on a bootstrap/resampling of the data, capturing a different aspect of the data from each model and making it less likely to overfit by using death and mobility rates to classify whether a state was considering a low or high restriction state.

5.2 Results

For states with eased masked mandates, we found that they started the year about -19% $\pm$ 2.5% below the baseline average. The day by day increase from the New Year's Day starting point is 0.25% $\pm$ 0.04%, delineated [0.21%, 0.29%]. Thus, variability of about a 0.08% increase each day in eased masked mandate states based on the standard errors, providing us with a quantitative statement of uncertainty. For states with continued masked mandates, we found that they started the year about 27.5% +- 2% below the baseline average. The day by day increase from the New Year's Day starting point is 0.27% $\pm$ 0.03%, delineated [0.24%, 30%]. Thus there is variability of about a 0.06% increase each day in eased masked mandate states based on the standard errors, providing us with a quantitative statement of uncertainty. Our GLM's findings show us that the initial change of baseline was significantly lower in states with masked mandates. However, the rate of increase with respect to the baseline is actually slightly higher in masked mandates relative to no masked mandate states. Both of the GLM models have a positive slope coefficient, which aligns with our assumption that people are more likely to spend time in retail and recreation as the number of people with the vaccine increases.[1]

5.3 Discussion

In terms of the GLM models, the states with masked mandates GLM performed slightly better due to a better log-likelihood score. The no mandate state model had a log-likelihood of -430, and the masked mandate state model had a log-likelihood of -390. Both models could be improved significantly moving forward. We could implement more robust bootstrapping techniques to improve our data, along with additional features. We could get more data from retail stores such as sales volume, traffic congestion, or any other data that significantly impacts retail and recreation mobility. Moreover, vaccine percentage could be helpful as it was a guiding thesis in our research but not explicitly included in our regression model. The random forest model provided a higher accuracy score than the k-Nearest Neighbors method as it searched for the best feature in a random subset of the resampled data whereas kNN is more sensitive to outliers in the data. The accuracy was around 60% means more than half of the time this model accurately classifies whether a state is labeled as a low or high restriction state. The plots show that states with increased retail and recreation and higher deaths are primarily classified as low restrictions regions and states ¿0% for retail and recreation labels as high restrictions regions with one outlier classified as no mask with over 20 deaths per 100,000 residents. However, the confidence we have in applying this to

future datasets is difficult to accurately determine as this model remains less interpretable than the GLM model with its random nature.

6 Conclusion

Our key findings are that having low COVID restrictions, little to no statewide mask mandates, and more open policy on COVID restrictions helps encourage mobility to return to the normal baseline it was before the pandemic. The rate of mobility increase with respect to the baseline did appear to be slightly higher in high restriction states relative to low restriction states, but both of the GLM models had a positive slope coefficient, which aligns with our assumption that people are more likely to spend time in retail and recreation COVID cases decrease in the US. Therefore, people are more inclined to visit stores, transit stations, parks, and work if the restrictions are lower. This feels very obvious, however it is important to note that these restrictions are not in place for naught. Many of the times, these states are also states with fairly high rates of COVID, and low vaccine numbers. For example, even though Georgia has released a mask mandate, "More than 6 million doses of vaccines have been given in Georgia, but the state ranks 44th in doses administered per capita to people 18 and older. Georgia still ranks among the 10 worst states per capita for newly reported deaths and among the 10 worst states per capita for hospitalized COVID-19 patients".[1] This only demonstrates a growing divide between the states who have high COVID restrictions, and typically have high vaccine rates subsequently low COVID numbers, vs states with low COVID restrictions, and low vaccine rates. Another key point to consider is that mobility may be decreased for states with higher restrictions due to greater job loss and economic instability in these regions.

Based on our results, one intervention might be to allow the states with high vaccination records and low COVID numbers to decrease restrictions more so as to allow them to increase mobility and boost the economy again. It would be ideal to decrease low risk restrictions such as outdoor dining. While our findings may reveal that retail and recreation mobility is greater in states with lower restrictions, it is important to use this data to consider the cost benefit of what truly matters. These results would force policymakers to address if returning to full retail and recreation mobility is more important than lowering the death rate, bringing ethics into the conversation of economic gain.

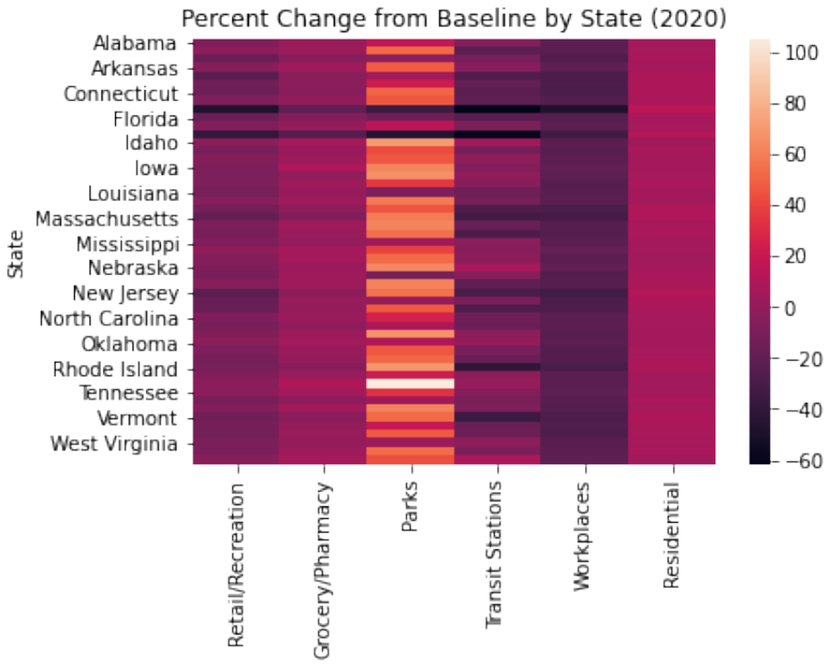
We merged the Google Mobility US data set with the NYT dataset provided by course staff on Piazza. We did so in order to account for different confounding variables for our causal analysis. The consequences were that we lost some of the rows due to the fact that they may not have been in both data sets (for example, on certain days, certain counties may not have reported their COVID numbers or may not have reported their mobility change. This would cause issues in the joining of the two tables.) However, given the expansive breadth of data we had, this felt like a small price to pay in order to ensure a stronger causal analysis. In comparison we did not do a merging of data sets on the GLM analysis. We only used the US mobility dataset from 2021.

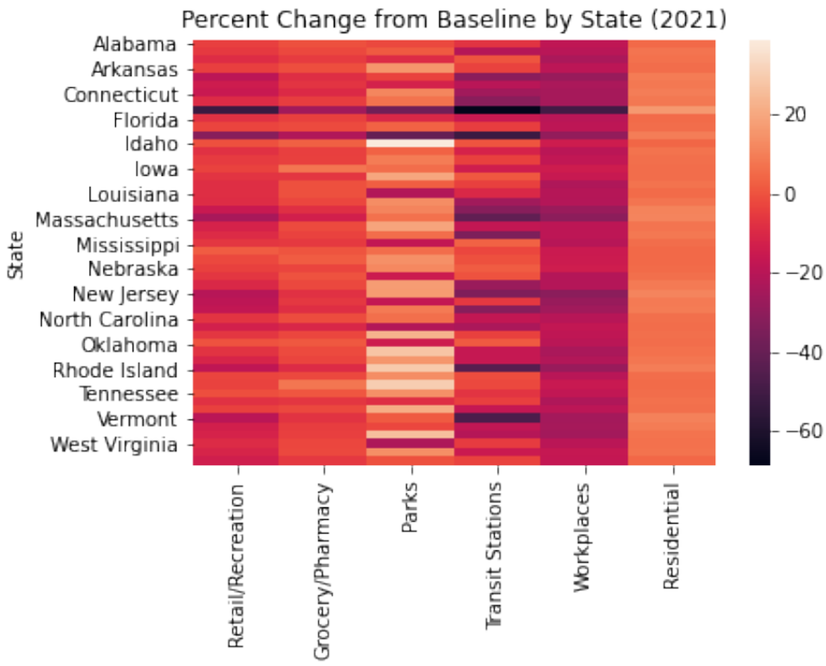
A few limitations we encountered was the lack of vaccine data per state or per county per day. This information would have been increasingly helpful in our causal analysis of the high-vs-low COVID restriction on retail mobility. We believe that this feature may have also been a confounding variable that we were unable to account for in our analysis. As for the limitations with GLMs had we had access to more features such as, hospitalizations, unemployment, and aforementioned vaccine data, we would have been able to better fit the model and improve our predictions by taking into account parameters that contribute to retail and recreation mobility.

A few future studies that could build on our work are understanding the mobility for certain regions post pandemic. For these research questions we simply looked at how deaths and mask mandates relate to retail and recreation mobility as of this spring, but a development would be to see how the mobility will continue to change with states' mask mandates evolving, vaccine rollout, and companies transitioning to permanent remote work. With the mass vaccination of US residents, many states will begin to remove mask mandates and lessen other restrictions. However, COVID has made remote work significantly more prevalent. Many companies, such as Twitter, Uber, Google, have given their employees the option to work from home permanently due to the efficiency and effectiveness of remote work.[2] Only time will reveal the long lasting effect of remote work, but a future study on how retail and recreational mobility has changed with these additional factors of evolving restrictions, vaccines, and long term remote work would be an exciting area to explore.

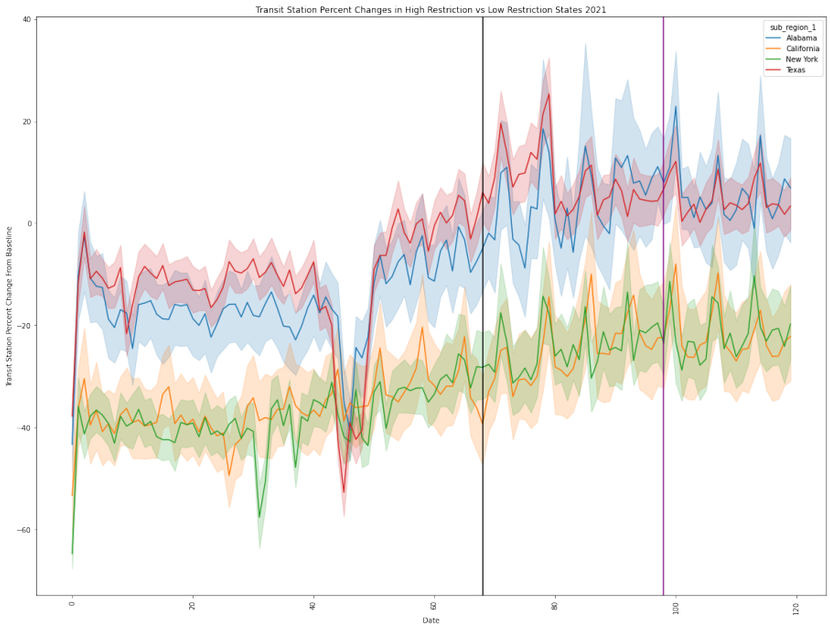
7 Appendix

7.1 Changes from Baseline by State

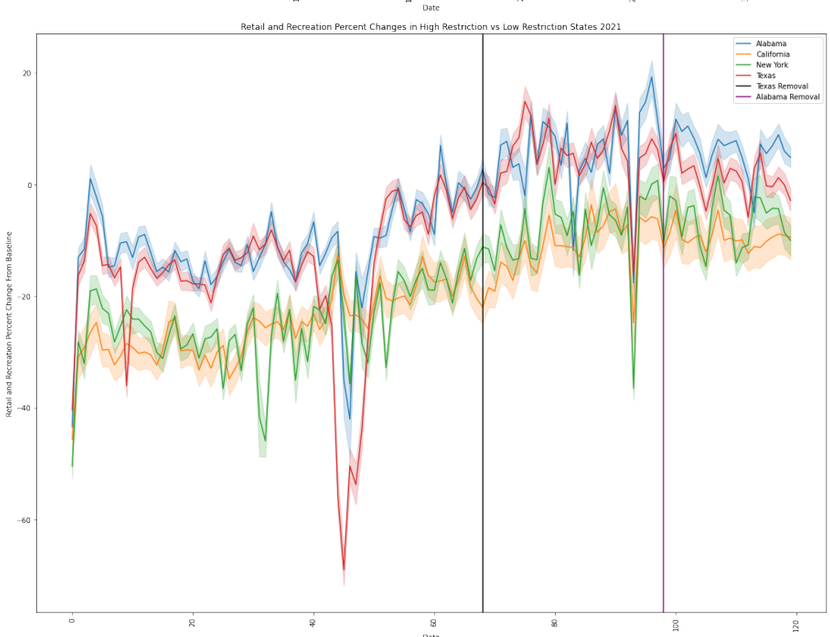


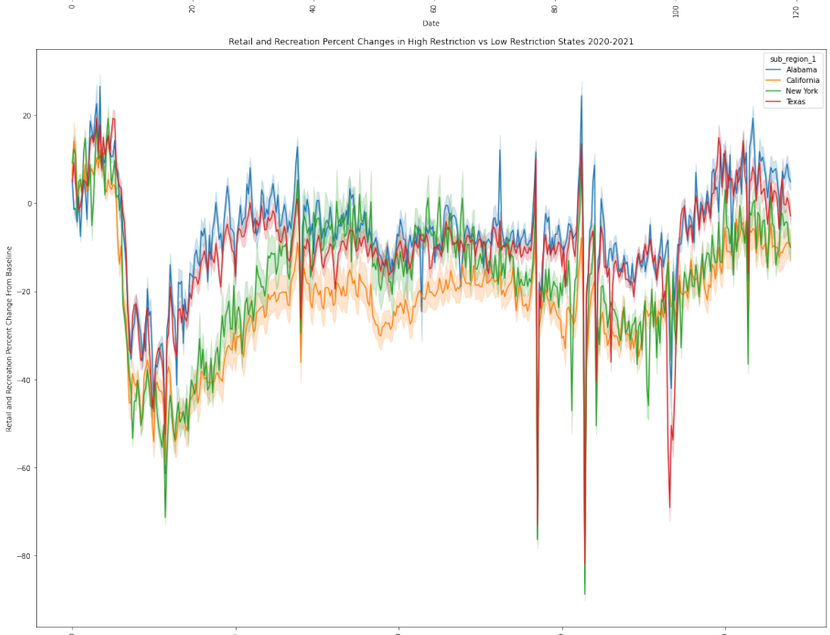


7.2 Transit Percent Changes

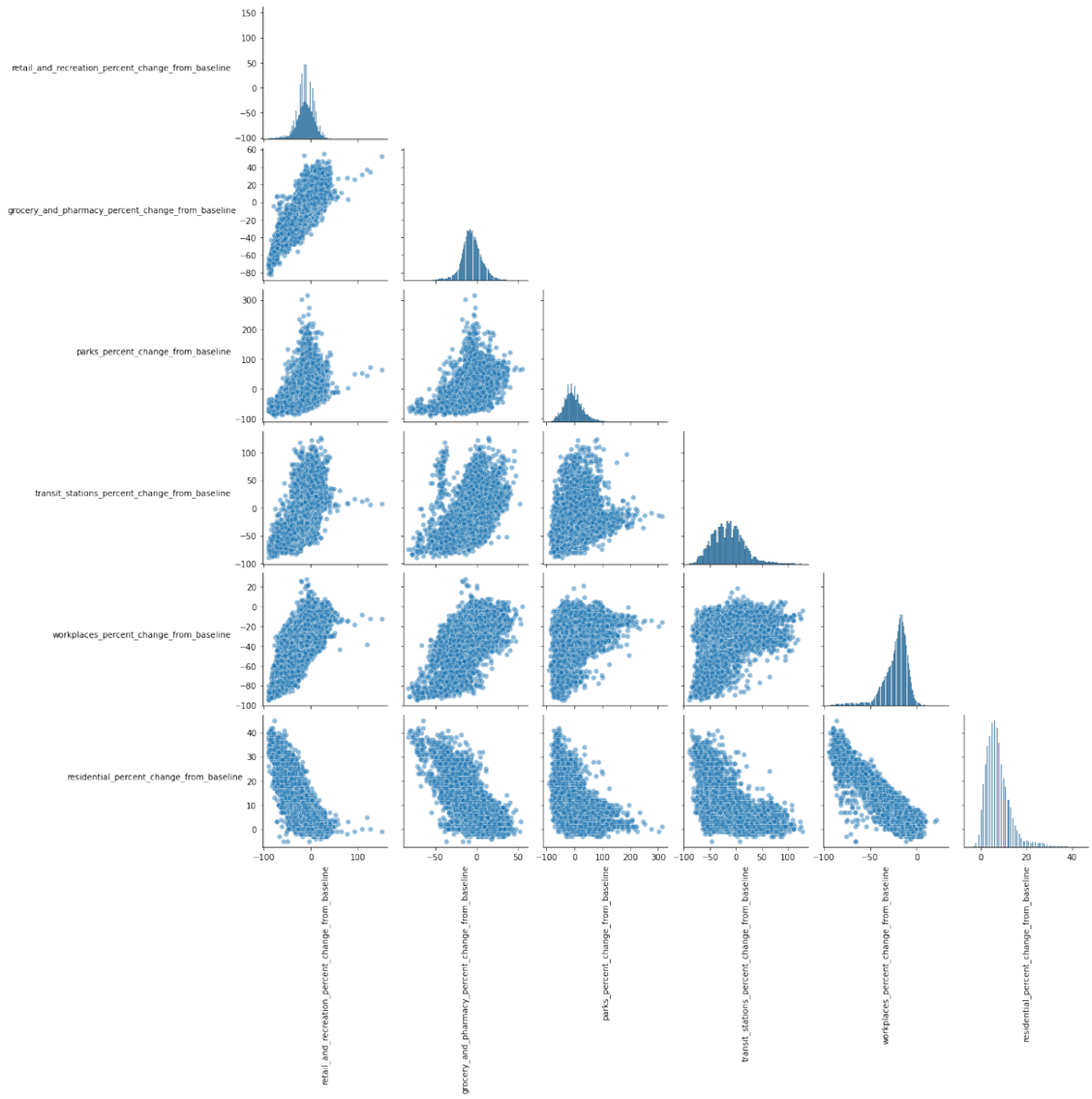


7.3 Retail and Recreation Percent Changes

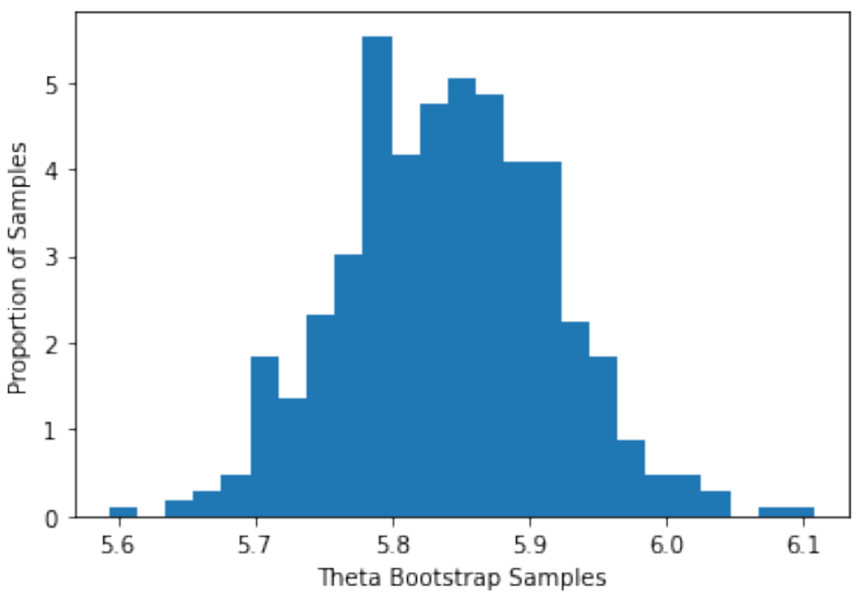




7.4 Causal Inference



7.5 Bootstrap



7.6 Generalized Linear Model

Figure 1: No Mask Mandate States

Generalized Linear Model Regression Results

Dep. Variable:	retail_and_recreation_percent_change_from_baseline	No. Observations:	122
Model:	GLM	Df Residuals:	120
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	52.248
Method:	IRLS	Log-Likelihood:	-413.42
Date:	Sun, 09 May 2021	Deviance:	6269.7
Time:	14:02:58	Pearson chi2:	6.27e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-19.4316	1.301	-14.938	0.000	-21.981	-16.882
ddate	0.2503	0.019	13.469	0.000	0.214	0.287

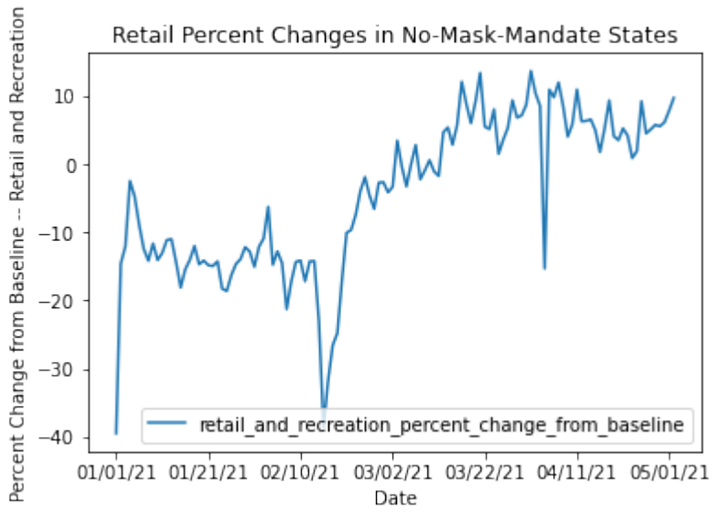
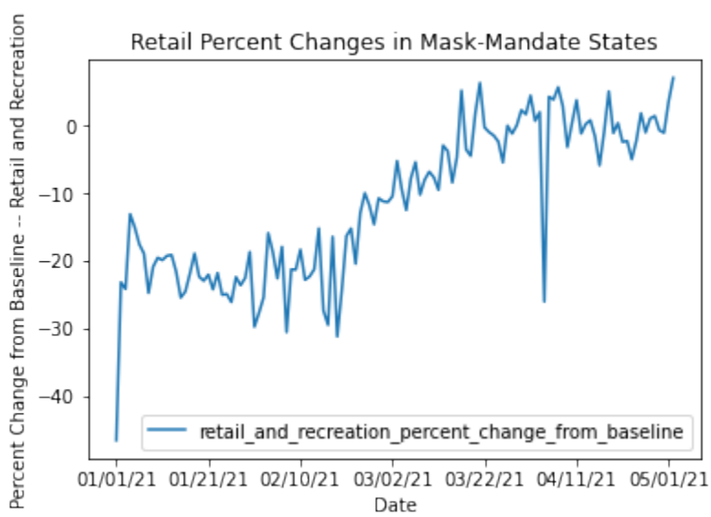


Figure 2: Mask Mandate States

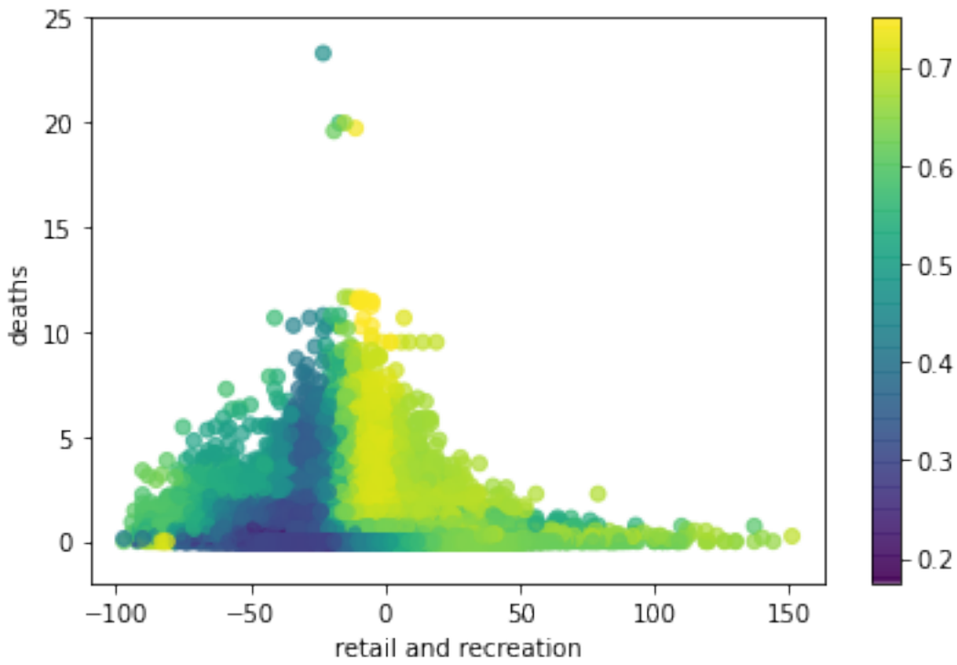
Generalized Linear Model Regression Results

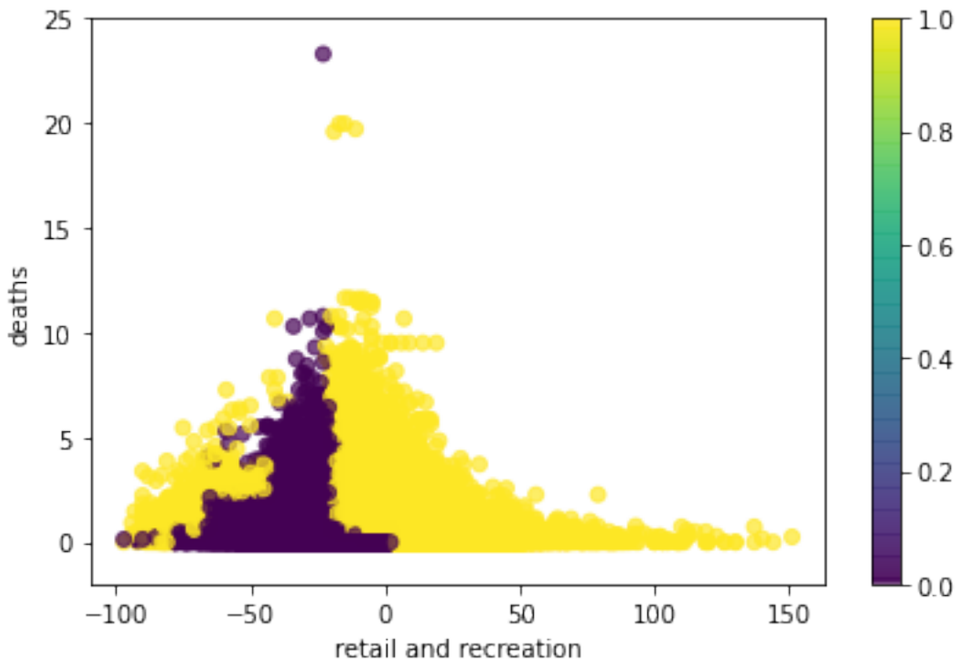
Dep. Variable:	retail_and_recreation_percent_change_from_baseline	No. Observations:	122
Model:	GLM	Df Residuals:	120
Model Family:	Gaussian	Df Model:	1
Link Function:	identity	Scale:	35.650
Method:	IRLS	Log-Likelihood:	-390.10
Date:	Sun, 09 May 2021	Deviance:	4278.0
Time:	14:02:58	Pearson chi2:	4.28e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-27.6640	1.075	-25.745	0.000	-29.770	-25.558
ddate	0.2685	0.015	17.489	0.000	0.238	0.299



7.7 Nonparametric





7.8 References

1. "Georgia's Kemp Lifts Many COVID-19 Rules on Businesses." U.S. News and World Report, U.S. News World Report, www.usnews.com/news/best-states/georgia/articles/2021-04-30/georgias-kemp-lifts-many-covid-19-rules-on-businesses.
2. Joey Hadden, Lara Casado. "17 Major Companies That Have Announced Employees Can Work Remotely Long Term." Entrepreneur, Entrepreneur, 17 Aug. 2020, entm.ag/ggh1.
3. Graphics, WSJ. "Tracking Covid-19 Vaccine Distribution." The Wall Street Journal, Dow Jones Company, 8 Apr. 2021, www.wsj.com/articles/tracking-covid-19-vaccine-distribution-11611355360?mod=theme$_c oronavirus - ribbon.d$
4. Markowitz, Andy. "Does Your State Have a Mask Mandate Due to Coronavirus?" AARP, www.aarp.org/health/healthy-living/info-2020/states-mask-mandates-coronavirus.html.