Are national parks becoming more attractive to locals during and post-pandemic?

Peter Morgan, Frances Leung, Nitin Swarup Sokhey

8/6/2021

Section 1: Introduction

With travel restrictions, border closures and shutdown of businesses across the globe throughout the pandemic, the tourism industry has taken a significant financial hit and its outlook is still uncertain. According to the United Nations World Tourism Organization (UNWTO), international tourist arrivals have fallen 72% between January 2020 and July 2021 for the Americas [1]. And for the world, the overall decline in the same period was reported to be 85% [1]. For the California Tourism Board, rebuilding tourism as a more sustainable and resilient sector in the future is a priority.

Tourism has accounted for $144.9 billion dollars in the California economy and resulted in $12.2 billion in local and state tax revenue [2]. Tourism also supports 1.2 million jobs in California [2]. There is a strong motivation to have a deeper understanding on what drives tourism especially during the pandemic where everything is so uncertain.

To get a greater understanding of what drives tourism, we will be using national park visits as our outcome variable and gain a greater understanding of what increases visit. Understanding which variables increases visits to national parks is valuable information as it opens the door to follow up studies to see if those same variables cause other sectors of tourism to increase.

This report is structured with nine additional sections. Section 2 describes the research question and the the variables of interest for our research question. Section 3 describes our data sources along with initial exploratory analysis of our data. Section 4 describes our models that we will be using. Section 5 describes the results of our model. Section 6 describes the limitations of our model using the Classical Linear Model (CLM) assumptions. Section 7 is a discussion of omitted variables. Section 8 is our concluding remarks. Section 9 cites our sources, and Section 10 is an appendix of additional charts and figures that were not able to be included in this report.

Section 2: Research Question

With statewide travel guidelines to avoid non-essential out-of-state and out-of-country travel in effect throughout the pandemic, our hypothesis is that this would heighten residents' interest in taking local trips and exploring nature in sites near them during and shortly post-reopening. The purpose of this research is to examine the effect of the pandemic on visitation to national parks. Our goal is to determine the relationship between the number of visitations to parks in the United States and variables including the number of COVID-19 cases and vaccination rate.

The main reason we hypothesize this position is because we suspect that higher COVID rates decrease travel outside a persons home state, people are more likely to visit national parks close to them where it is easier to reach and easier to adhere to social distancing guidelines.

Understanding our research question is extremely valuable for the California Tourism Board. If we find what increases national park visits, we will have a great framework for follow up studies to be performed to

understand what increases more financially lucrative areas of tourism.

We will use a causal model to address this research question and determine if higher vaccination rates and lower COVID cases cause more national park visits. The model will be described further in section 4.

Section 3: Data

This section describes each of the datasets used and any manipulation that was performed on those datasets. It will also show some exploratory data analysis. More exploratory data analysis can be found in the appendix. For the purposes of this study, we will consider all data prior to or within February 2020 to be "pre-COVID" data and any data points within or after March 2020 to be "post-COVID" data.

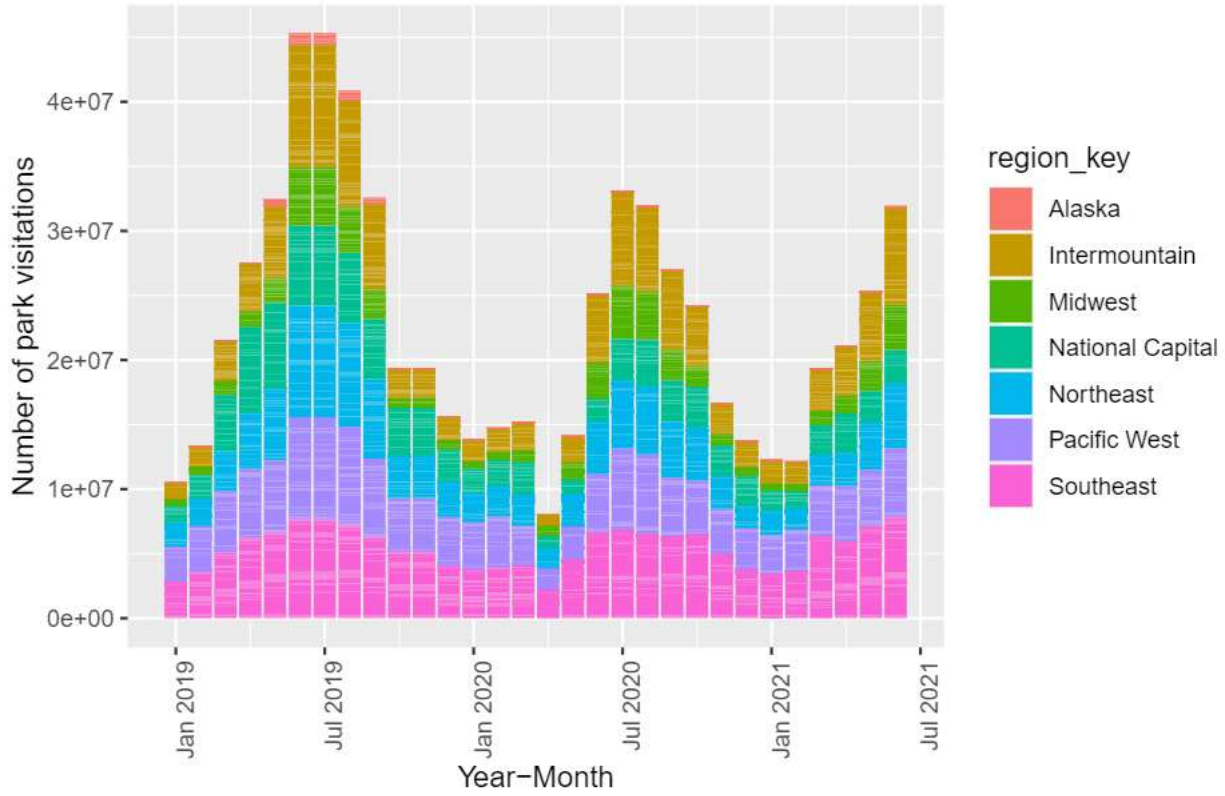
National Park Data

Our national park data is downloaded from the national park service website [4]. To get our initial dataframe, some manual manipulation was necessary. The reason for this is because the national park data was extracted from the website on a monthly basis. This means that the data needed to be downloaded from every month and combined. We collected data from January of 2019 to June of 2021.

Once all the months were combined in this dataset, each row is a unique park month combination. The additional columns of data that we collect are the number of park visits that month, the difference in visits from this month to last month, the number of visits from the same month last year, the number of year to date visits this year, the number of year to date visits last year, the number of year to date visit differential from this year to last year, and the region of the park. The national park service defines seven regions to classify national parks into. These regions are Alaska, Intermountain, Midwest, National Capital, Northeast, Pacific West, and Southeast.

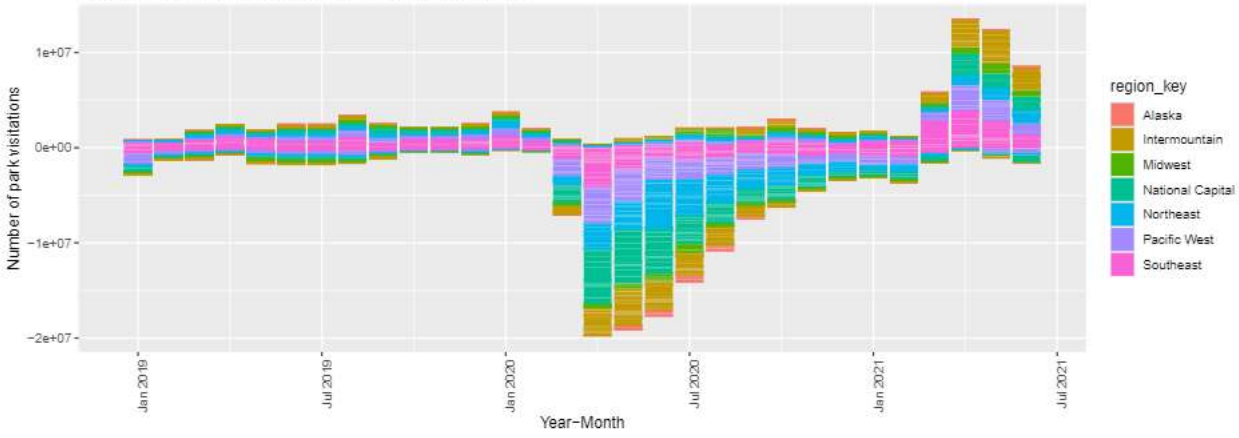
For initial exploration of this data, we are interested in examining how park visits have changed over time. Figure 1 below shows national park visits by month, filtered by the region. The first thing we observe is that it seems that there were more park visits before COVID. You can see that the more park visits in the dataset occurs in the summer of 2019. Another observation is that there appears to be seasonal variation in our data. There are more visits in the summer of COVID time than in the winter of non COVID times.

Figure 1: US Park Visitations



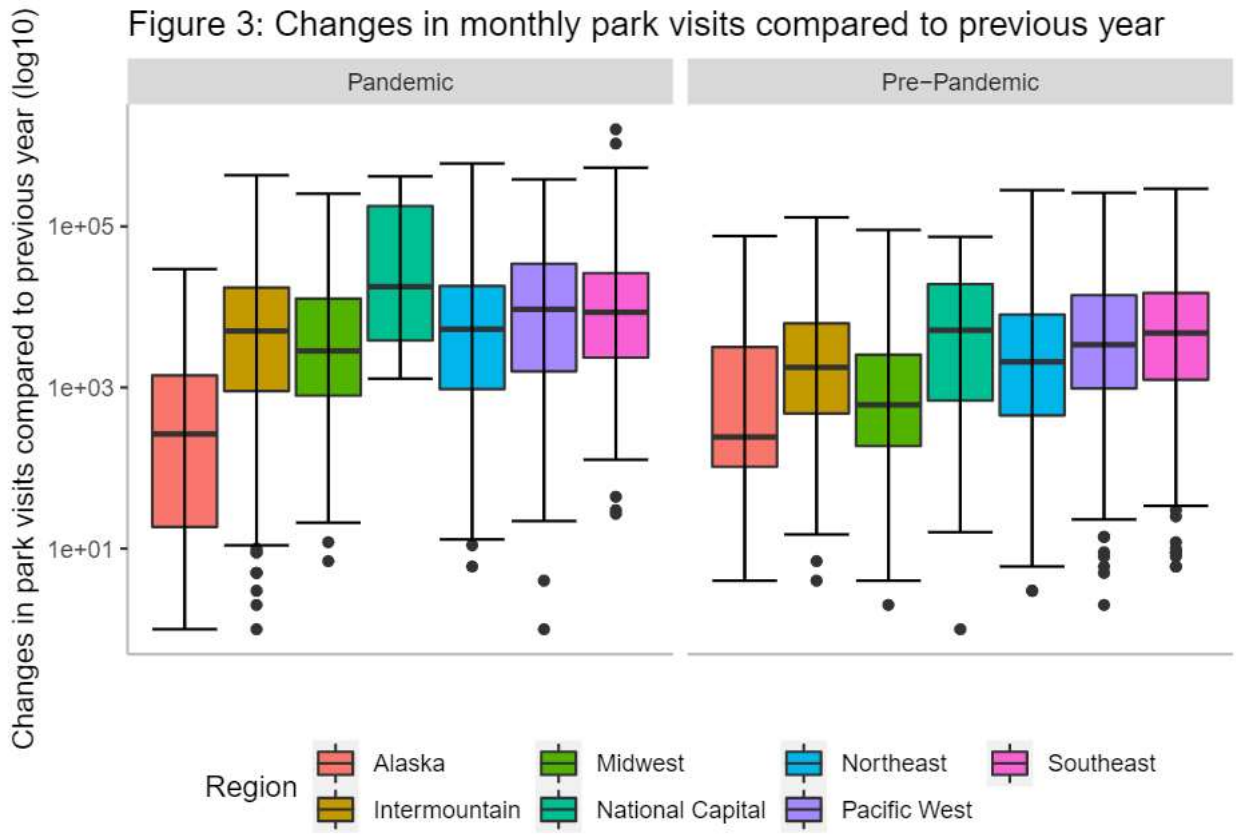
To account for seasonal effect in our exploration, we turn to figures 2 and 3. Figures 2 and 3 show us the difference in the number of visitations in parks compared to last year, also filtered by region. The chart shows that visitations in 2019 were fairly static and followed a similar trend to the data in 2018. However, once February of 2020 hits, we see a drastic decrease in park visits compared to the previous year. This trend continues up until March of 2021. March of 2021 had slightly more park visitations than March of 2020 and that trend continues up until June of 2021.

Figure 2: US Park Visitations YOY difference by month



The boxplots in Figure 3 shows the monthly changes in park visits compared to the previous year in the same month. We can see that when the two pools of pre-pandemic and post-pandemic data are compared side-by-side, there is an evident increase of park visits during pandemic for all regions. It is because of this

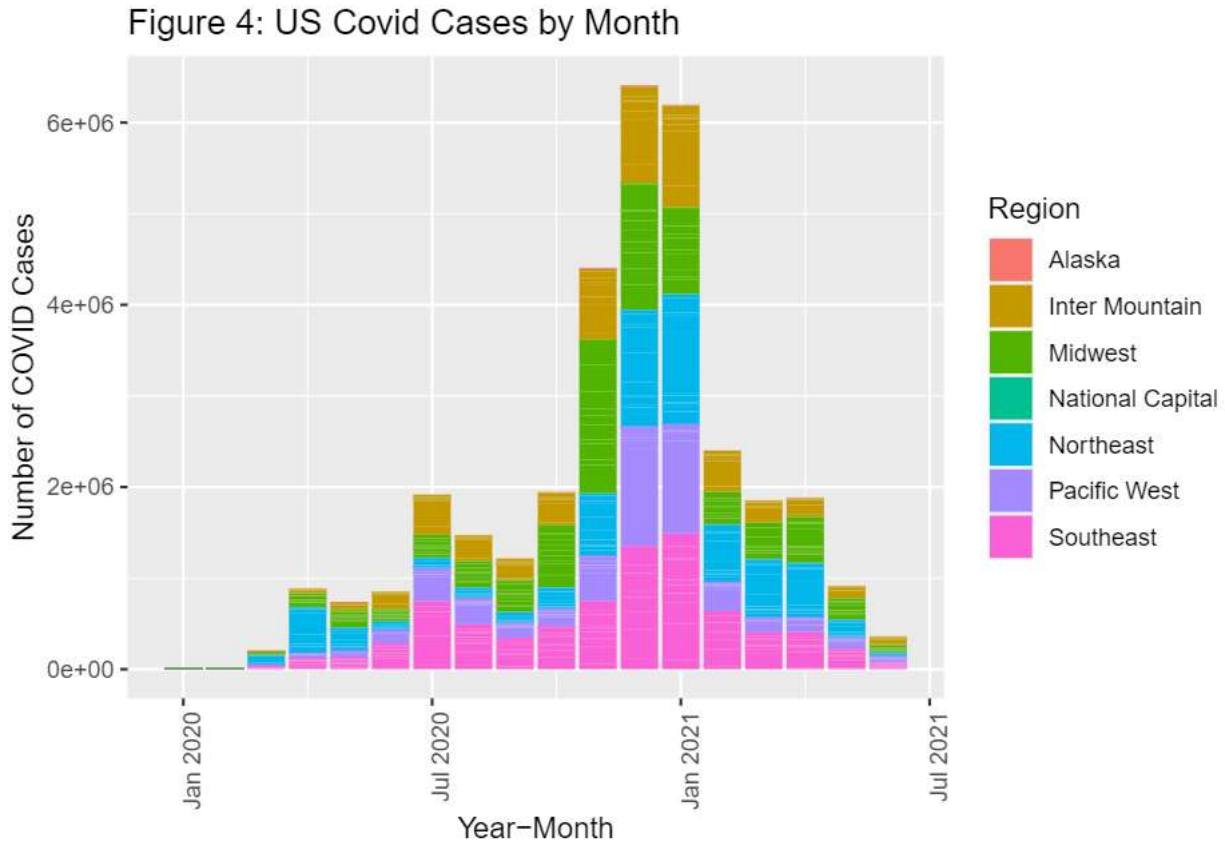
observation in particular that motivates our investigation on the effect of pandemic related variables such as COVID cases, vaccination numbers, etc. on park visitations.



COVID Case Data

Our COVID Cases dataset was downloaded from the New York Times COVID-19 dataset [5]. This file contains data where each datapoint is a unique date and state combination. The raw data has dates from January 21st 2020 to present day. The data will be cut off at the end of June because that is how far we are going for our park data and any data before January 21st 2020 will be reported as values of zero because there are no reported COVID cases in the United States before that date [5].

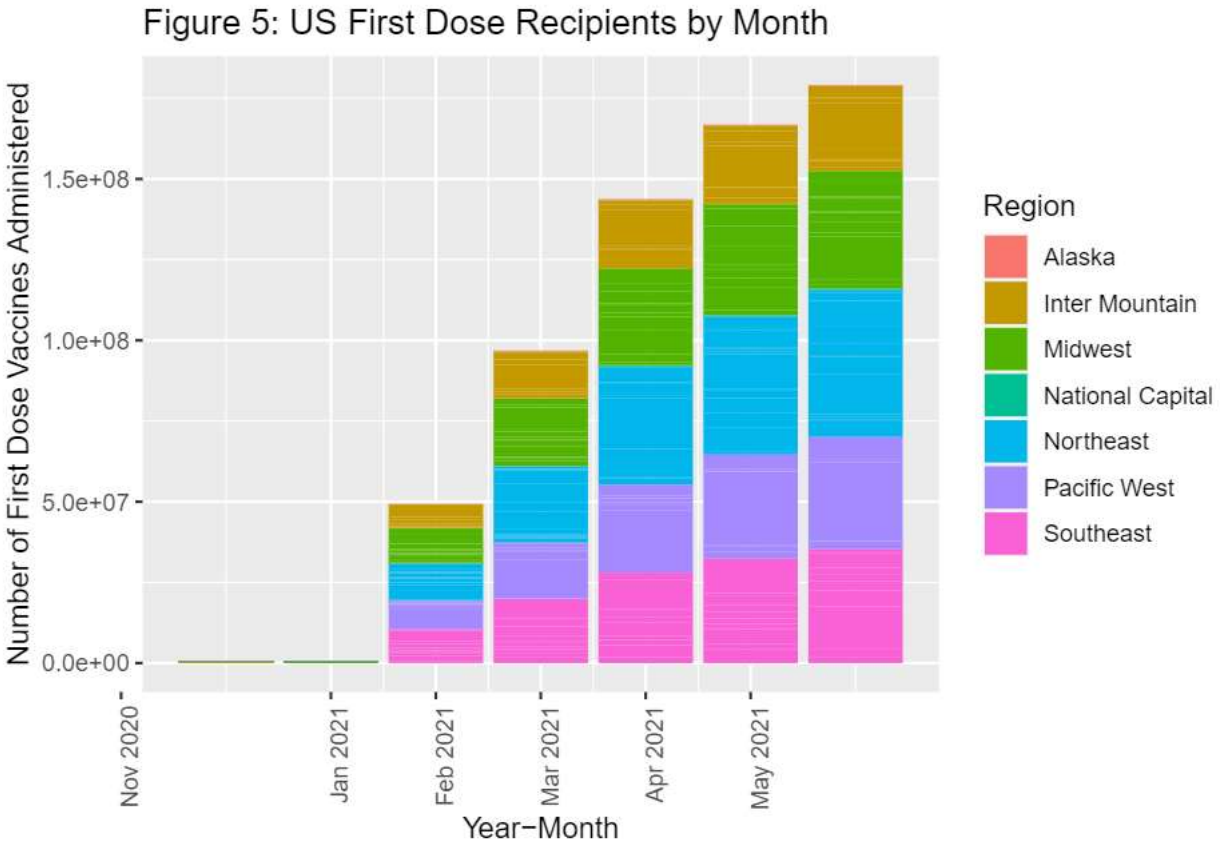
The raw data file is made up of three additional columns beyond date and state. These columns are fips which is a state identifier that will be removed, the number of COVID cases on that date, and the number of COVID deaths on that date. To get the number of COVID cases per month, the sum of cases and deaths for all days in a specified month are reported and the data is transformed to have monthly rows instead of daily rows. Figure 4 below shows the distribution of COVID cases across by month filtered by region. As you can see, the number of cases maximizes in the beginning of 2021 and end of 2020. According to our hypothesis, this should be the area with the least quantity of park visits



Vaccination Data

Our vaccination data was downloaded from the CDC website [6]. The raw data file contains 27 columns. Each row has a unique state date combination. The first step in cleaning the data is removing any unnecessary columns. This leaves us with 6 remaining columns which are date, state, percentage of fully vaccinated people, number of fully vaccinated people, number of people who have received the first does of the vaccine, and percentage of people who have received the first dose of the vaccine. The region column is appended on based on the state using a mapping table. After this, some rows are dropped as well. This dataset contains data on every single day. We are only interested in monthly data so we drop all data that is not the last day of the month. The last day of the month will contain accurate data on how many people were vaccinated by the end of the month.

We explore this data using figure 5 below. This figure shows the number of people who have received at least the first dose of the vaccine each month filtered by region. This plot shows that ever since February, there has been an increase in vaccinations administered among all regions while the trend plateaus as we move closer to present day. Our hypothesis says that higher national park visits should occur more recently as there are more vaccinations administered at this time.



Mapping Tables and Other Data Sources

We have discussed the main sources of data that will be analyzed in our project. However, there are a few more data sources and mapping tables that are used in this analysis that will be described here.

The first data source is from the 2020 U.S. census [7]. This data is used for population metrics to calculate statewide COVID case data normalized by population. We analyzed data from 2019, 2020, and 2021 in this analysis. For the purposes of this study we will assume that all the state population values are equal to the population reported in 2020.

We also have two manually created mapping tables to assist with data joining. The first mapping table maps every state to a region. This makes it easy to add a region column when necessary. The second mapping table maps every national park to a state. This data was not available for download in the national park dataset as only the region was included. This will assist with data joining as described below.

Data Joining

First, the state column is added to the national park dataset using the national park state mapping table. Next, the COVID case dataset is joined to the national park dataset by state, and month. The vaccination data is then joined by state and month. In the end, we have a dataframe where each row is a unique park, month combination.

Section 4: Statistical Model

To investigate the research question **Are national parks becoming more attractive to locals during and post-pandemic?**, we examined the effects of pandemic related input variables on park visitations by building numerous linear regression models for comparisons. The input/output variables that were considered are described in the two tables below.

Table 1 describes the model setup for 5 models (Models 1a to 5a) using pandemic input variables in the form of numbers.

Table 2 describes the model setup for the equivalents (Models 1b to 5b) using pandemic input variables in the form of rates. Model 5a is the same as Model 4a except for the use of y2 for month park visits instead of y1 for park visit changes as a comparison. The same is true for Models 5b and 4b.

Table 1: Linear regression models (With pandemic inputs as #'s)

Variables	Model 1a	Model 2a	Model 3a	Model 4a	Model 5a
y1 - Park visit changes	y1	y1	y1	y1	–
y2 - Monthly park visits	–	–	–	–	y2
x1 - # of COVID cases	x1	x1	x1	x1	x1
x2 - # of 1st dose vaccination	x2	x2	x2	x2	x2
x3 - # of COVID deaths	–	x3	x3	x3	x3
x4 - # of full vaccination	–	–	x4	x4	x4
x9 - Pandemic indicator	-	-	-	x9	x9

Table 2: Linear regression models (with pandemic inputs as rates)

Variables	Model 1b	Model 2b	Model 3b	Model 4b	Model 5b
y1 - Park visit changes	y1	y1	y1	y1	–
y2 - Monthly park visits	–	–	–	–	y2
x5 - Rate of COVID cases	x5	x5	x5	x5	x5
x6 - Rate of 1st dose vaccination	x6	x6	x6	x6	x6
x7 - Rate of COVID deaths	–	x7	x7	x7	x7
x8 - Rate of full vaccination	–	–	x8	x8	x8
x9 - Pandemic indicator	-	-	-	x9	x9

Description of variables

The letter **y** denotes output variables. The letter **x** denotes input variables. A total of 2 output variables and 9 input variables were examined. For each of COVID cases, COVID deaths, 1st dose vaccination and full vaccination, both measures in monthly counts (variables x3) and monthly rates (based on the respective state's population) were available in the sample data.

Output variable - Park visitations

The full data set assembled has 11,310 monthly park visitation data points from 385 national parks in 7 regions across the United States. Park visitations data from January 2019 to June 2021 were available at the time the research was conducted. Both the changes in monthly park visitations compared to the previous year (**denoted as y1**) and monthly park visitations (**denoted as y2**) were available.

Given the research question focuses on analyzing the **difference** in park visitation numbers between pre- and during/post-pandemic, we have chosen **y1** as the output variable for Models 1a to 4a. As a comparison or the effect of the use of rates vs. numbers, we used **y2** as the output variable for Model 5a and 5b.

Categorical variable - Pandemic indicator

Aside from the source data retrieved on park visitations, COVID cases/deaths and COVID vaccination, we also added a categorical variable to depict whether the data point is related to the pre-pandemic vs. during/post-pandemic time frame. For all monthly data points that occurred prior to March 2020, we considered that as being "Pre-pandemic" and denoted the indicator as 0. For all data points that occurred within or after March 2020, we considered that as being "Pandemic" and denoted the indicator as 1. The Pandemic Indicator variable is represented by x9 in the model. This categorization simplifies the modeling as we were not performing any time series analysis in this research. Time become either it is pandemic or not pandemic.

Selection of data points to use for modelling

Given the full sample dataset covers the entire United States, we extracted data from the top 3 out of the 7 regions for modeling purposes. This step was taken because it is very unlikely for a single model to describe patterns from the entire country. We also decided not to use a single region but a few to ensure that there is at least some randomness in the sample data and covers geographies that are not entirely adjacent to each other. The 3 regions chosen to be included in the modeling were Intermountain, Northeast and Pacific West.

These were also made cause these are the regions where there is a higher concentration of parks. Pacific West was chosen because California state alone has the highest number of parks in the country. The below tables depicts somem summary statistics on parks by regions and states.

Table 3: Park distribution by 7 US regions and top 7 states

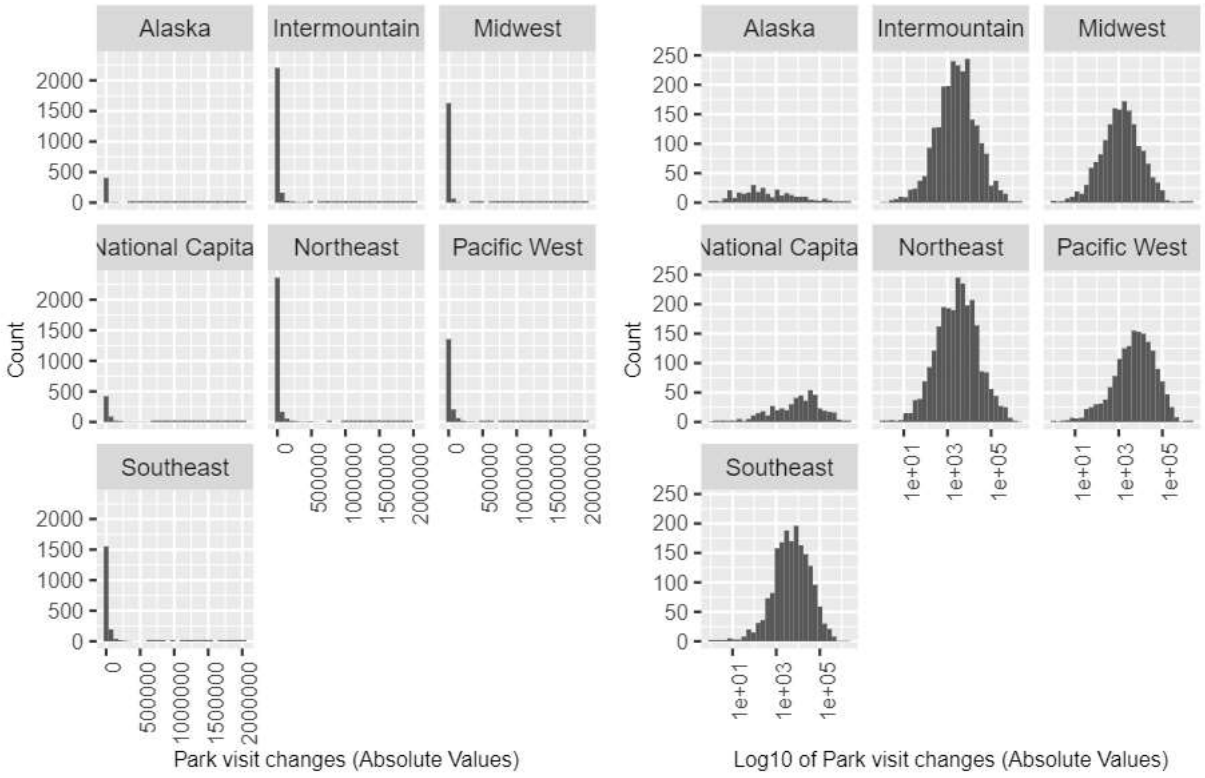
region	n	percent	state	n	percent
Total	385	-	Total	385	-
Northeast	92	23.90%	California	26	6.75%
Intermountain	82	21.30%	New York	23	5.97%
Southeast	62	16.10%	District of Columbia	21	5.45%
Midwest	57	14.81%	Arizona	19	4.94%
Pacific West	56	14.55%	New Mexico	16	4.16%
National Capital	21	5.45%	Alaska	15	3.90%
Alaska	15	3.90%	Pennsylvania	15	3.90%

Transformation on variables

A review of park visitation distribution with histograms across all 7 US regions shows that it has a very skewed distribution to the left. Because of this, we applied a logarithmic transformation on park visits related output variables. To enable logarithmic transformation on changes to park visits, we also needed to take the absolute values of the park visit changes to convert negative values into positive values. See figure 6 below for the comparison before (left panel) and after (right panel) the transformation.

Another transformation to the data that was applied was to convert any NA values that arose after joining the park visits data with the COVID and vaccination data to zero. The NA would be because there are no COVID cases or vaccination happening yet pre-pandemic.

Figure 6: Histograms of Park Visits



Linear regression model equations

The 10 models that we formulated have mathematical equations shown below. There are 2 sets with 5 models for each.

Set 1 - Linear regression models (With pandemic inputs as #'s)

Model 1a - Base model with # of COVID cases and # of 1st dose vaccination as inputs, changes in park visits as output

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 &+ \beta_1\ \text{(Number of COVID cases)}\ + \beta_2\ \text{(Number of 1st dose vaccination)}\end{aligned} \tag{1}$$

Model 2a - Model 1a plus # of COVID deaths

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 &+ \beta_1\ \text{(Number of COVID cases)}\ + \beta_2\ \text{(Number of 1st dose vaccination)} \\ &+ \beta_3\ \text{(Number of COVID deaths)}\end{aligned} \tag{2}$$

Model 3a - Model 2a plus # of full vaccination

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 &+ \beta_1\ \text{(Number of COVID cases)}\ + \beta_2\ \text{(Number of 1st dose vaccination)} \\ &+ \beta_3\ \text{(Number of COVID deaths)} \\ &+ \beta_4\ \text{(Number of full vaccination)}\end{aligned} \tag{3}$$

Model 4a - Model 3a plus categorical variable pandemic indicator

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 &+ \beta_1\ \text{(Number of COVID cases)}\ + \beta_2\ \text{(Number of 1st dose vaccination)} \\ &+ \beta_3\ \text{(Number of COVID deaths)} \\ &+ \beta_4\ \text{(Number of full vaccination)} \\ &+ \beta_9\ \text{(Pandemic indicator)}\end{aligned} \tag{4}$$

Model 5a - Same as Model 4a but output variable becomes number of park visits

$$\begin{aligned}\text{Park visits} = \beta_0 &+ \beta_1\ \text{(Number of COVID cases)}\ + \beta_2\ \text{(Number of 1st dose vaccination)} \\ &+ \beta_3\ \text{(Number of COVID deaths)} \\ &+ \beta_4\ \text{(Number of full vaccination)} \\ &+ \beta_9\ \text{(Pandemic indicator)}\end{aligned} \tag{5}$$

Set 2 - Linear regression models (With pandemic inputs as rates)

Model 1b - Base model with rate of COVID cases and rate of 1st dose vaccination as inputs, changes in park visits as output

$$\text{log10(abs(Park visit changes))} = \beta_0 + \beta_5\ \text{(Rate of COVID cases)}\ + \beta_6\ \text{(Rate of 1st dose vaccination)} \quad (6)$$

Model 2b - Model 1b plus rate of COVID deaths

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 + \beta_5\ &\text{(Rate of COVID cases)}\ + \beta_6\ \text{(Rate of 1st dose vaccination)} \\ &+ \beta_7\ \text{(Rate of COVID deaths)}\end{aligned} \quad (7)$$

Model 3b - Model 2b plus rate of full vaccination

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 + \beta_5\ &\text{(Rate of COVID cases)}\ + \beta_6\ \text{(Rate of 1st dose vaccination)} \\ &+ \beta_7\ \text{(Rate of COVID deaths)} \\ &+ \beta_8\ \text{(Rate of full vaccination)}\end{aligned} \quad (8)$$

Model 4b - Model 3b plus categorical variable pandemic indicator

$$\begin{aligned}\text{log10(abs(Park visit changes))} = \beta_0 + \beta_5\ &\text{(Rate of COVID cases)}\ + \beta_6\ \text{(Rate of 1st dose vaccination)} \\ &+ \beta_7\ \text{(Rate of COVID deaths)} \\ &+ \beta_8\ \text{(Rate of full vaccination)} \\ &+ \beta_9\ \text{(Pandemic indicator)}\end{aligned} \quad (9)$$

Model 5b - Same as Model 4b but output variable becomes number of park visits

$$\begin{aligned}\text{Park visits} = \beta_0 + \beta_1\ &\text{(Rate of COVID cases)}\ + \beta_2\ \text{(Rate of 1st dose vaccination)} \\ &+ \beta_3\ \text{(Rate of COVID deaths)} \\ &+ \beta_4\ \text{(Rate of full vaccination)} \\ &+ \beta_9\ \text{(Pandemic indicator)}\end{aligned} \quad (10)$$

Section 5: Results from Model

The results from the regression for selected models are described in the below sections. Models 1b, 3b and 5b from **Set 2 - Linear regression models (with pandemic inputs as rates)** are as follows. The models are listed as (1), (2) and (3) respectively in the table.

```
##
## ===============================================================================================
##                                               Dependent variable:
##                     ---------------------------------------------------------------------------
##                           log10(abs(park_month_diff) + 1)           log10(park_ty_monthly + 1)
##                             (1)                    (2)                         (3)
## -----------------------------------------------------------------------------------------------
## covid_cases_rate          4.058                -8.219                     38.489***
##                          (4.177)               (5.399)                     (6.923)
## one_dose_rate            -0.003**               0.008                      0.025***
##                          (0.001)               (0.007)                     (0.008)
## covid_deaths_rate                             798.754***                -2,824.811***
##                                              (247.683)                    (301.526)
## fully_vax_rate                                 -0.014                     -0.023**
##                                               (0.008)                     (0.010)
## pandemic_indicator                                                        -0.431***
##                                                                           (0.075)
## Constant                 3.376***              3.361***                    3.914***
##                          (0.029)               (0.029)                     (0.041)
## -----------------------------------------------------------------------------------------------
## Observations              2,250                 2,250                       2,250
## R2                        0.002                 0.009                       0.067
## Adjusted R2               0.001                 0.007                       0.065
## Residual Std. Error  1.121 (df = 2247)     1.118 (df = 2245)          1.337 (df = 2244)
## F Statistic        2.371* (df = 2; 2247) 4.911*** (df = 4; 2245)  32.232*** (df = 5; 2244)
## ===============================================================================================
## Note:                                                          *p<0.1; **p<0.05; ***p<0.01
```

Based on the results, the majority of the pandemic related variables are statistically and significantly different from zero. With this, we can reject the null hypothesis that the coefficients are zero. There is however some instability in terms of the signs of the coefficients that switch between positive and negative. This applies to both the one_dose_rate and covid_cases_rate.

From a residual standard error perspective, all models above have relatively low residuals. From a significance perspective, Model 5b (depicted as (3) above) appears to describe the patterns of the sample data points the best out of all. However, with an adjusted R2 of 0.065, it means that the model is only capable of describing ~ 6.5% of the data points.

Based on the fitted model, Model 5b above, our multivariate least squares regression formula becomes:

$$\begin{aligned}\text{Park visits} = 3.914 &+ 38.489 \cdot \text{ (Rate of COVID cases) } + 0.025 \cdot \text{ (Rate of 1st dose vaccination)} \\ &- 2824.811 \cdot \text{ (Rate of COVID deaths)} \\ &- 0.023 \cdot \text{ (Rate of full vaccination)} \\ &- 0.0431 \cdot \text{ (Pandemic indicator)}\end{aligned} \tag{11}$$

The interpretation of the above equation is as follows:

- When the rate of COVID cases increases by 1, there will be 38.489 times more monthly park visits, all else being equal

- When the 1st dose vaccination rate increases by 1, there will be 2.5% more park visits, all else being equal
- When the COVID death rate increases by 1, there will be a decrease in 2,824.811 times fewer monthly park visits, all else being equal
- When the full vaccination rate increase by 1, there will be a decrease of 2.3% of park visits, all else being equal
- During a pandemic, park visits decrease overall by 43.1%, all else being equal

While this model that encapsulates the most variables (all being statistically significant) predicts that the higher the number of COVID cases and the higher the 1st dose vaccination, the more park visits there would be, it suggests that COVID death rate and full vaccination rate would however would counter park visitations. It also predicts that when there is a pandemic, the park visits overall would drop. All of these predictions appear to be in alignment with our hypothesis that interests in national parks heightens at the time of a pandemic. A logical explanation for why there may be a decrease in park visits when the fully vaccination rate goes up is that people perhaps become more confident and feel less restrictive to expand the types of activities that they would like to engage outside of nature seeking outdoor activities such as visitation to parks. Furthermore, there are likely some collinearity that exists across some of the variables such as fully_vax_rate and one_dose_rate as well as COVID_cases_rate and COVID_deaths_rate, despite all of them resulted in being significant in the model.

Models 1a, 3a and 5a from **Set 1 - Linear regression models (with pandemic inputs as #'s)** are as follows. The models are listed as (1), (2) and (3) respectively in the table.

```
##
## ================================================================================================
##                                              Dependent variable:
##                     ---------------------------------------------------------------------------
##                           log10(abs(park_month_diff) + 1)        log10(park_ty_monthly + 1)
##                                     (1)                                (2)                  (3)
## ------------------------------------------------------------------------------------------------
## covid_cases                      0.00000**                          -0.00000            0.00000***
##                                  (0.00000)                          (0.00000)           (0.00000)
## one_dose_num                       0.000                             0.00000            0.00000***
##                                    (0.000)                          (0.00000)           (0.00000)
## covid_deaths                                                       0.00005***           -0.0001***
##                                                                    (0.00001)           (0.00002)
## fully_vax_num                                                      -0.00000            -0.00000***
##                                                                    (0.00000)           (0.00000)
## pandemic_indicator                                                                      -0.445***
##                                                                                          (0.063)
## Constant                         3.346***                           3.334***            3.914***
##                                   (0.026)                            (0.026)             (0.041)
## ------------------------------------------------------------------------------------------------
## Observations                       2,250                              2,250               2,250
## R2                                 0.003                              0.010               0.059
## Adjusted R2                        0.002                              0.008               0.057
## Residual Std. Error          1.121 (df = 2247)                  1.117 (df = 2245)   1.343 (df = 2244)
## F Statistic                2.836* (df = 2; 2247)              5.509*** (df = 4; 2245) 28.033*** (df = 5; 2244)
## ================================================================================================
## Note:                                                                    *p<0.1; **p<0.05; ***p<0.01
```

Section 6: Limitations of Model

Below we discuss several of the assumptions of the Classical Linear Model (CLM) adn whether our model satisfies them:

Assumption 1: Independent and Identically Distributed (IID)

Unfortunately, given the nature of the data, our model fails the first requirement of any classical model, that the data is IID. Given the broad natrue of our observations (i.e. cases per region) we acknowledge that there may be several issues with the assumption that our data was drawn independently from an identically distributed population (IID), which may diminish the ability of our final model to fully answer the research question as originally 10 posed, or may result in greater uncertainty than the model's standard errors suggest. In considering the IID assumption it is first useful to explicitly identify the population from which our sample is drawn, and about which it is supposed to make inferences.

The population is effectively all the people in the United States and the sampling in our research is just the aggregation of the population into the regions and states. Although we have regional indicators such as "Southeast, Midwest, ..." that group of states that share similar characteristics and help to draw some of the dependence between the data points into measurable form. Viewed in this way it is fairly obvious that this poses some clear challenges to the assumption of independence.

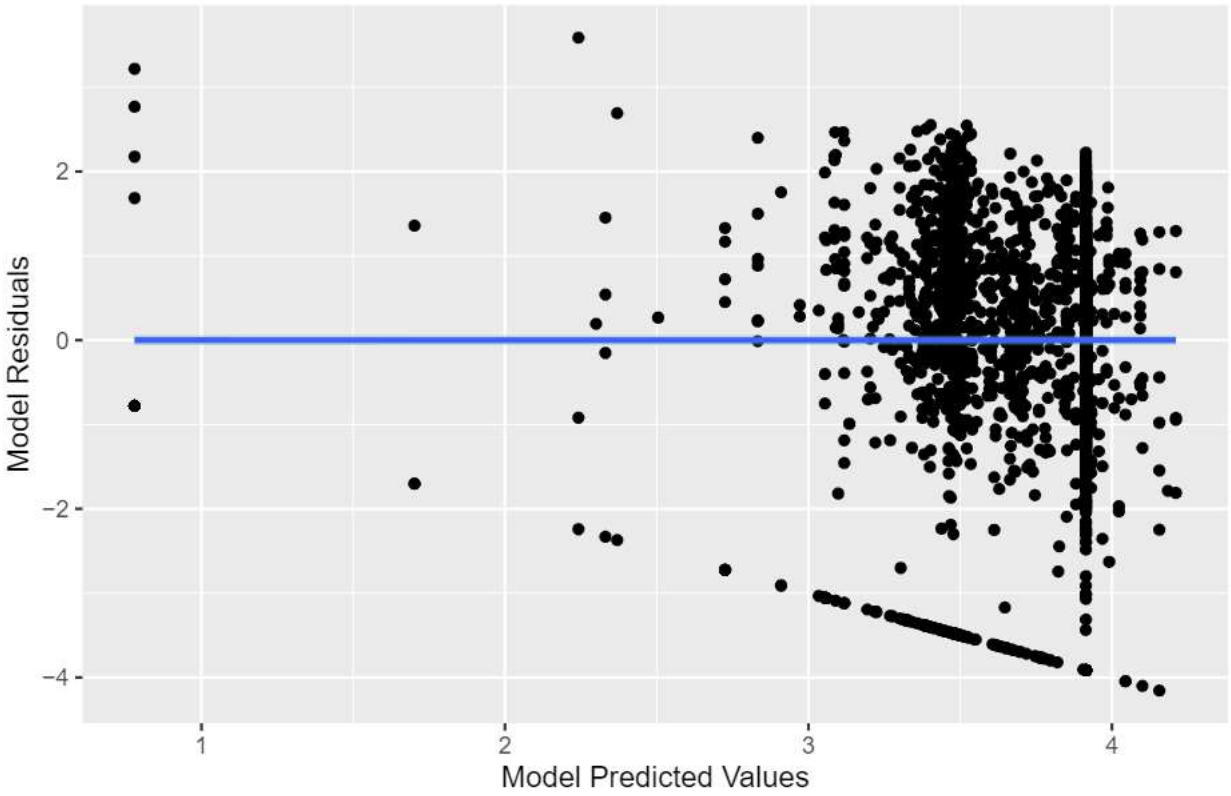
These issues with the IID assumption are one reason why we need to be cautious with how we interpret the model's results, and do not overstate how they would generalize to the population. The clustering of states is problematic, but geographical and political clustering of observations would still be an issue with more granular data, e.g. at the county level.

Assumption 2: Linear Conditionaly Expectation

The second assumption required for a classical linear model is that there is a linear conditional expectation in the exploratory variables. The best way to check for this assumption is through visually inspecting plots of the model residuals on the y-axis against each explainer on the x-axis. A plot of the fitted points against the residuals is also a useful tool, especially when there are too many variables to plot, as it captures the model to the whole. Looking at only fitted values against the residuals can hide issues in individual components though, so we prefer to look at each explainer individually if able. Having mainly 3 variables helps to do so concisely.

Our model as a whole passes the visual test for linearity at higher values, with our model consistently over predicting these points (as shown in Figure 7). Given the weight of evidence, we must conclude that this assumption is met. What this means for our models is that while it remains the best linear predictor for the data given, the coefficients are reflective of the true relationship between the variables and outcome.

Figure 7: Residuals vs. Predicted Values



Assumption 3: Homoscedasticity

This assumption refers to the variance of the residual, or error term, in a regression model being constant. The variance of the errors should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred condition is known as homoscedasticity (same scatter). If the variance changes, we refer to that as heteroscedasticity (different scatter).

The easiest way to check this assumption is to create a residuals versus fitted value plot. On this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction. The diagnostic plot can be found in the same figure as above;i.e. Figure 7.

It can be observed in the plot above that the majority of time, the variance is constant for a sizable portion of the observations. This means that the variance has been constant and thus homoscedasticity of the variance is maintained. Hence, this assumption is met.

Assumption 4: No Perfect Collinearity

Multicollinearity means that two or more regressors in a multiple regression model are strongly correlated. If the correlation between two or more regressors is perfect, that is, one regressor can be written as a linear combination of the other(s), we have perfect multicollinearity. While strong multicollinearity in general is unpleasant as it causes the variance of the OLS estimator to be large (we will discuss this in more detail later), the presence of perfect multicollinearity makes it impossible to solve for the OLS estimator, i.e., the model cannot be estimated in the first place.

In the case of strong collinearity, we calculated the paired-correlations of the variables contained in our model. Since we have only 3 major variables, the analysis was pretty easy and it turned out that none of the paired combinations of variables were having 1 as their correlation. Thus, this assumption is met.

```
## [1] "Covid Cases and Vaccinations"
## [1] 0.05801357
## [1] "Covid Cases and Deaths"
## [1] 0.6565995
## [1] "Vaccinations and Deaths"
## [1] 0.1089828
```

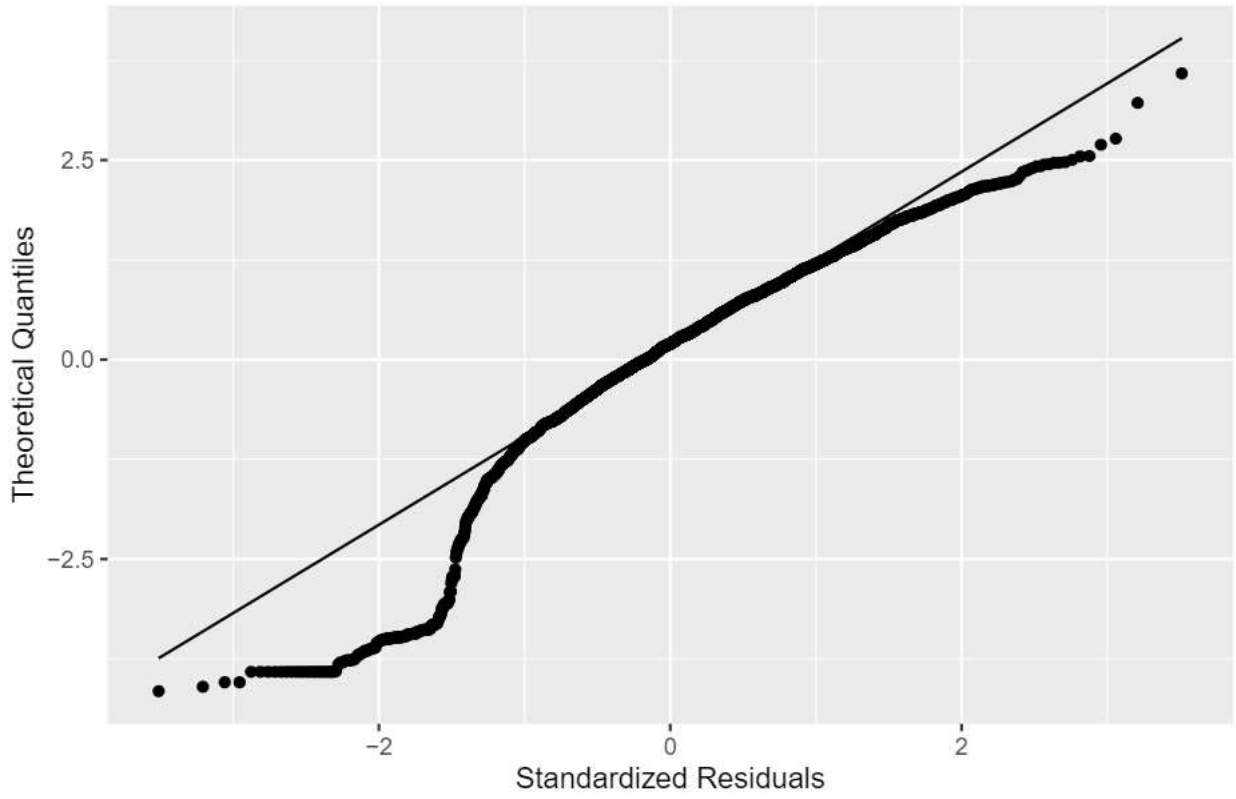
Assumption 5: Normally Distributed Errors

The normality assumption is needed to determine the shape of the sampling distribution for our estimator coefficients. By looking at the Normal QQ plot of the model's residuals, we see that the model is almost normal except that the tails are smaller than a normal curve. It of course fails the strict Shapiro-Wilk test of normality.

However, even if the errors are non-normal, a version of the CLM says that the sampling distribution for the estimator coefficients becomes normal as the number of observations increases. We have plenty of observations more than the typical rule of thumb value of 30 where the CLM becomes applicable. Given the large number of observations, we can confidently apply the CLM in this case and the spirit of the requirement is satisfied. (Spirit of requirement just wants the sampling distribution of beta to be normal. CLM says it will be.)

On inspection (Figure 8) we find no obvious evidence of deviations from normality that concerns us; there are some slight deviations from the perfect normal line at the upper and lower ends of the plot, but these are within the limits. Hence, this assumption is also met.

Figure 8: QQ plot



Section 7: Discussion of Omitted Variables

One of regression analysis' most serious problems occurs when omitted variables affect the relationship between the dependent variable and included explanatory variables. If researchers estimate without considering that the true slope is affected by other variables, then they obtain a slope estimate that is a constant, in contrast to the true slope.

The main relationship we explored in our model was between visits in parks in different regions of the US along with COVID-19 cases and vaccination rate. Park's visit is an extremely complicated topic with complex relationships with many of the variables we considered, in which directions of causality are often unclear or even disputed. For example, there is a correlation of ¥ =0.108 between vaccination of people and deaths due to COVID-19. In terms of causality we could easily argue that being vaccinated might not lead to fatal death, and likewise we could argue that the death could be avoided if the citizens are vaccinated. This is called the effect of the reverse Omitted variable. Similar bi-directional causal arguments could be made for the relationship between park visit and several of the variables we considered (e.g. COVID cases, COVID deaths, etc.) We do not suggest that a simple linear model could capture such a tangled web of causality, and hence we do not attempt to determine specific omitted variables that may causally introduce large biases to our model. However, in terms of additional variables which we didn't include that we believe may improve our model's performance in explaining the variance in COVID-19 deaths between different regions, we do suggest two particular variables that we would seek to explore further in any follow up of this study:

Seasons: There are three main seasons, Spring, Summer and Fall. Usually national parks would have more visitations in summer and spring season as compared to fall because of the snowfall which covers several roads and mountains across North America. It thus affects the ability to travel by road. However, due to the pandemic many people could not go out of their houses, thus impacting the visitations at the parks. We would seek to introduce a variable that could characterize overall political leanings of the state, e.g. the political party of its governor. Thus we would seek to introduce a variable that could give us insights about the seasons.

Unemployment: Prior to the pandemic, the unemployment rate was lower than during the pandemic. People used to spend on entertainment and social gatherings as part of the usual norm. This was impacted as a number of businesses were shut and unemployment resulted in people, focusing on day to day cost of living and reduced spending on entertainment and social gatherings, like visiting national parks for recreational activities. So we could also seek to introduce a variable which would account for the unemployment rate on a monthly time period.

Section 8: Conclusion

Overall, the economy has been severely impacted by COVID-19. Statewide travel guidelines forced residents to avoid non-essential out-of-state and out-of-country travel throughout the pandemic. We observe that changes in monthly park visits compared to previous year have increased during the pandemic. We believe this is due to reduced out-of-state travel. This infers a change in behavior where travelers are more open to visiting National Parks and outdoor recreational activities.

The top 3 visited locations in 2021 were the Great Smokey Mountains, Blue Ridge Parkway, and Golden Gate National Recreation Area where we have seen growth in visitors at the parks.

To come to this conclusion, we have reviewed the entire data across the United States over the period - 2019-Jun 2021. We see a correlations between park visits , COVID cases, deaths and vaccination numbers.

Section 9: References

[1] Tourist arrivals down 87% in January 2021 as UNWTO calls for stronger coordination to restart tourism. UNWTO. July 2021. https://www.unwto.org/news/tourist-arrivals-down-87-in-january-2021-as-unwto-calls-for-stronger-coordination-to-restart-tourism

[2] Why Travel Matters. Visit California. (n.d.). https://industry.visitcalifornia.com/partner-opportunities/programs/why-travel-matters

[3] https://www.businessinsider.com/are-national-parks-open-covid-19-coronavirus-united-states-nps-2020-5

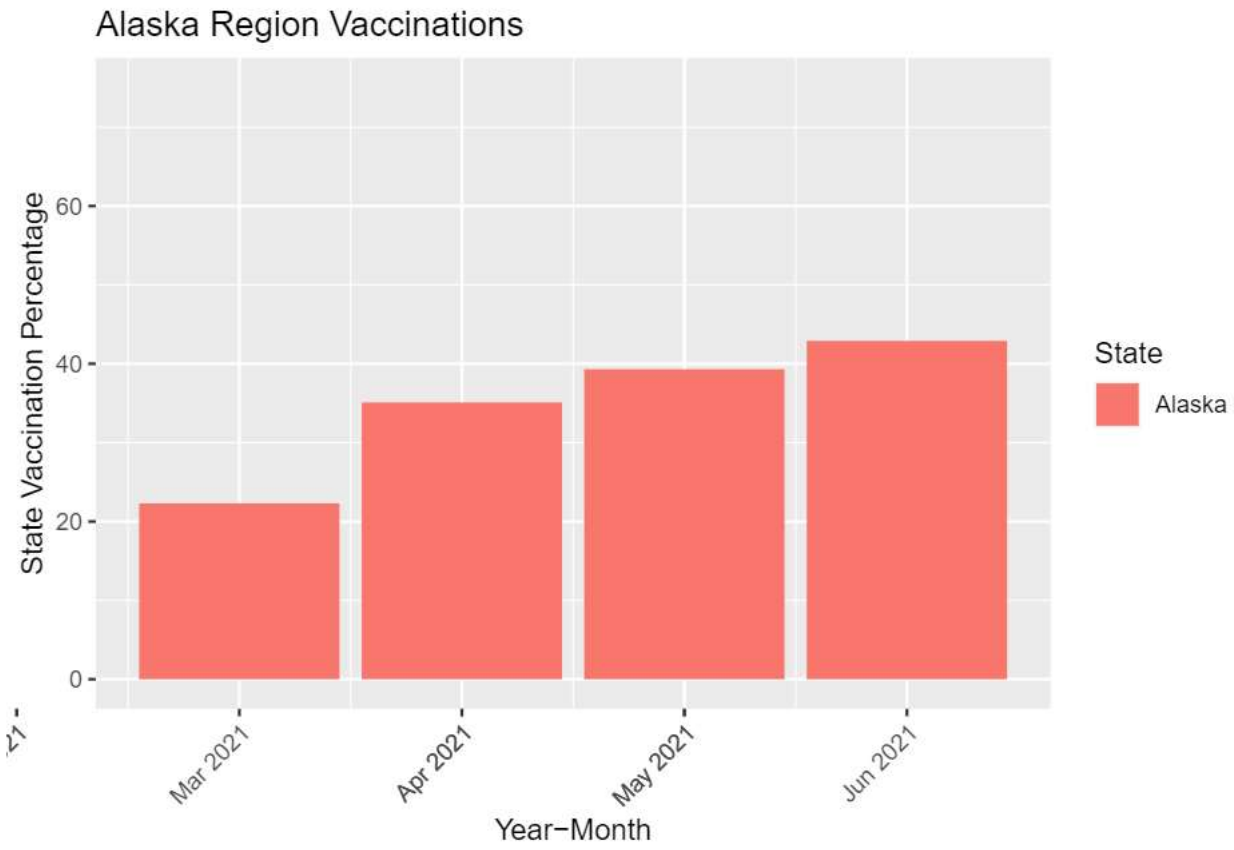
[4] https://irma.nps.gov/STATS/Reports/National

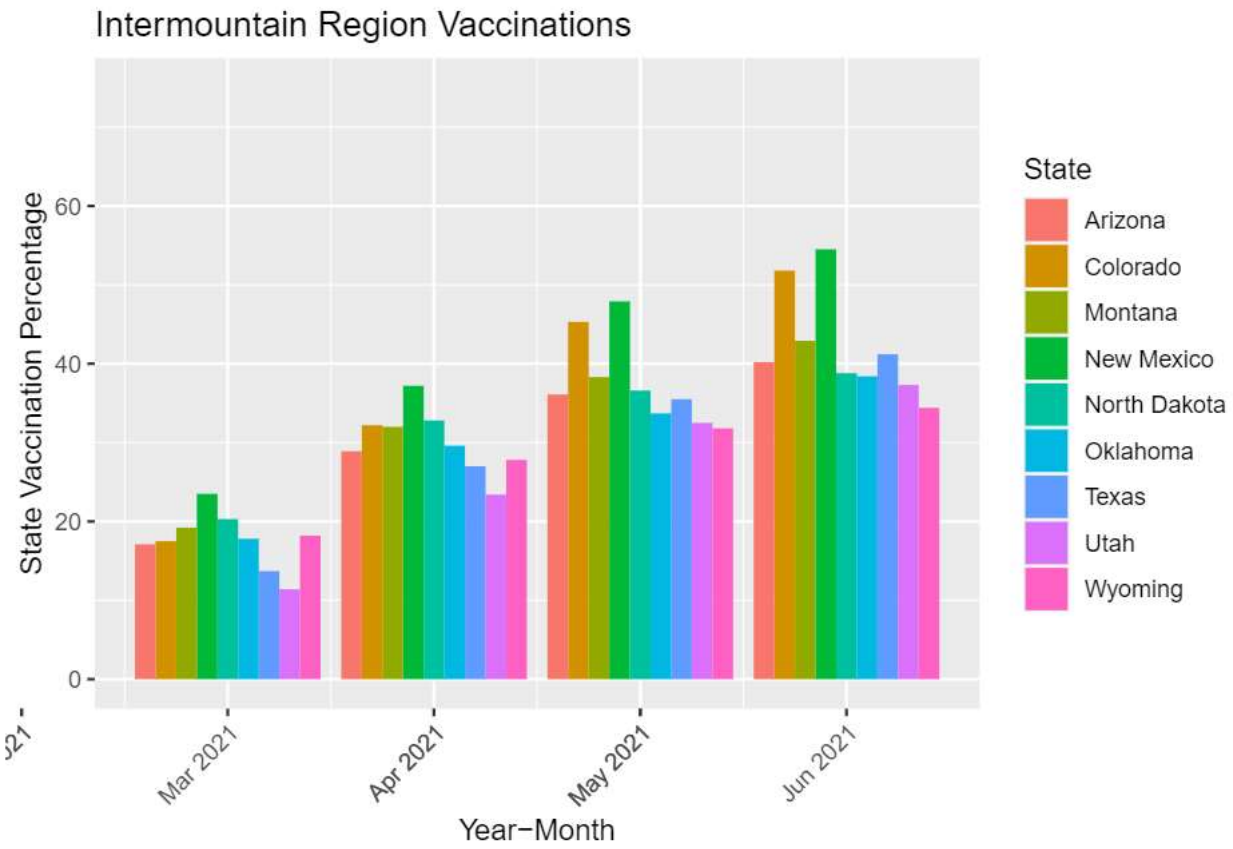
[5] https://github.com/nytimes/covid-19-data/tree/master/live

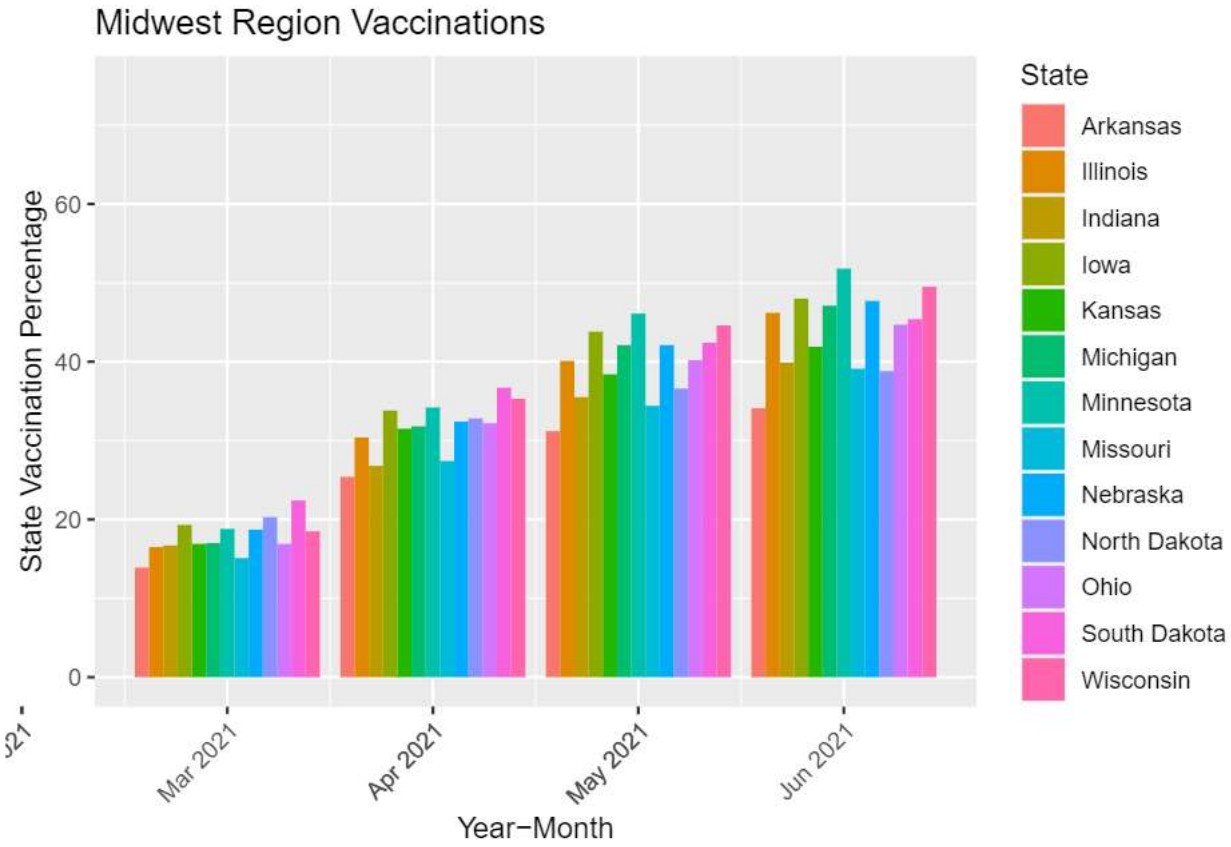
[6] https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

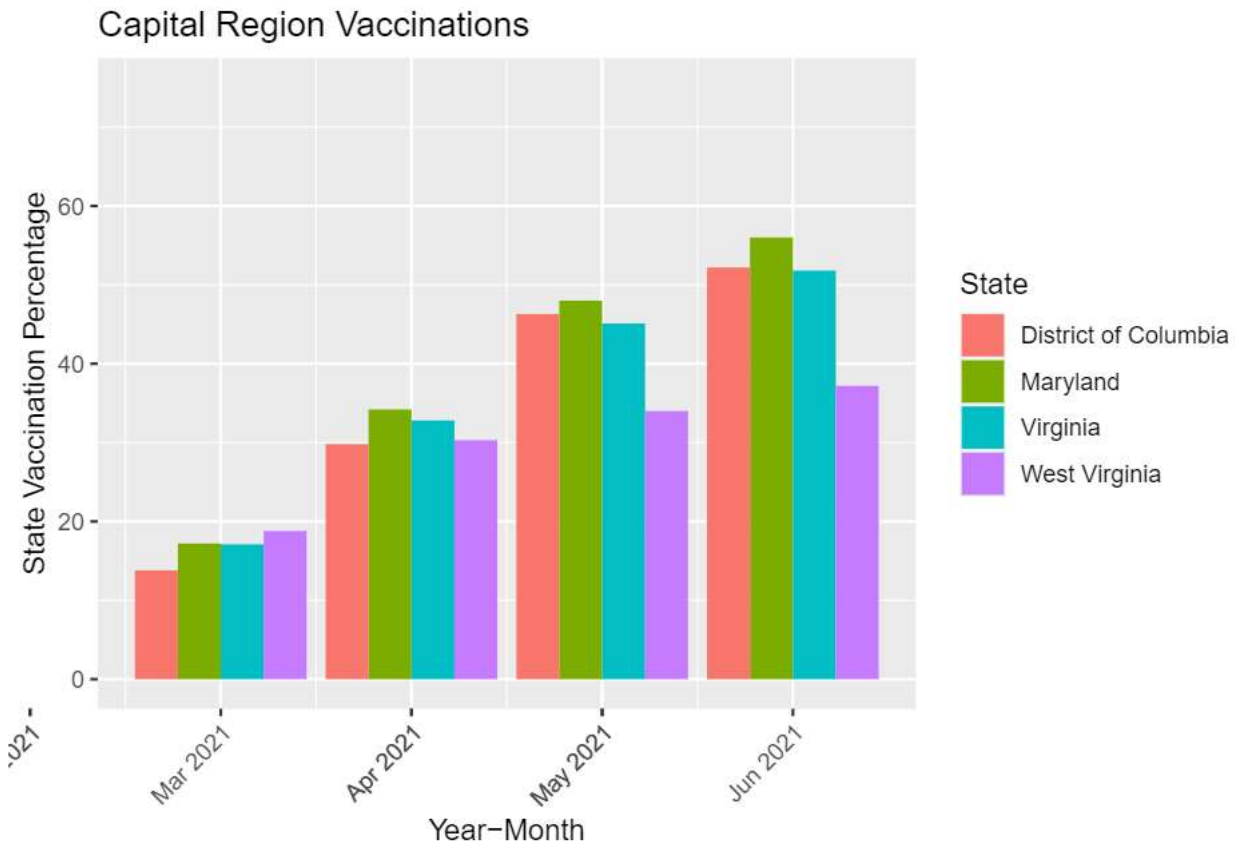
[7] National Population Totals: 2010-2020. United States Census Bureau. Annual Estimates of the Resident Population for the Nation and States. https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-totals-national.html

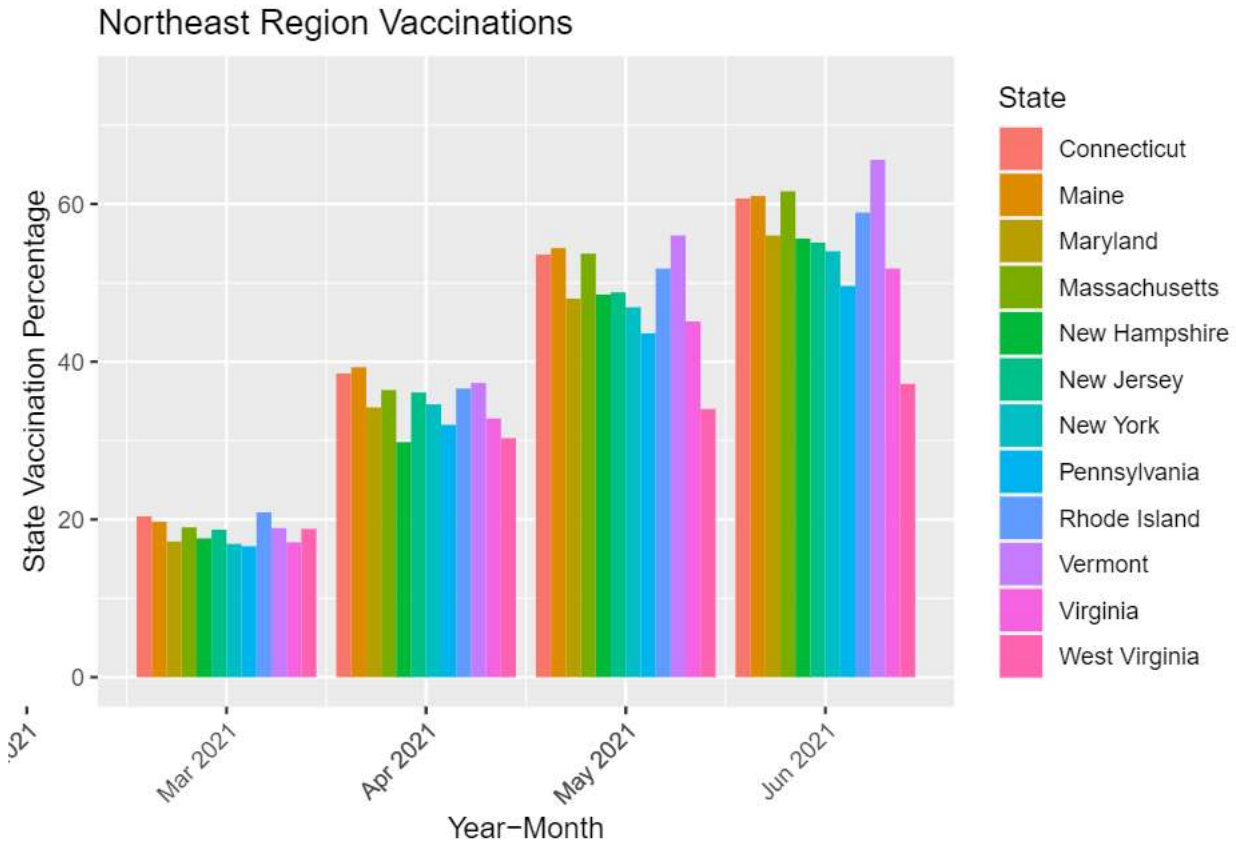
Section 10: Appendix

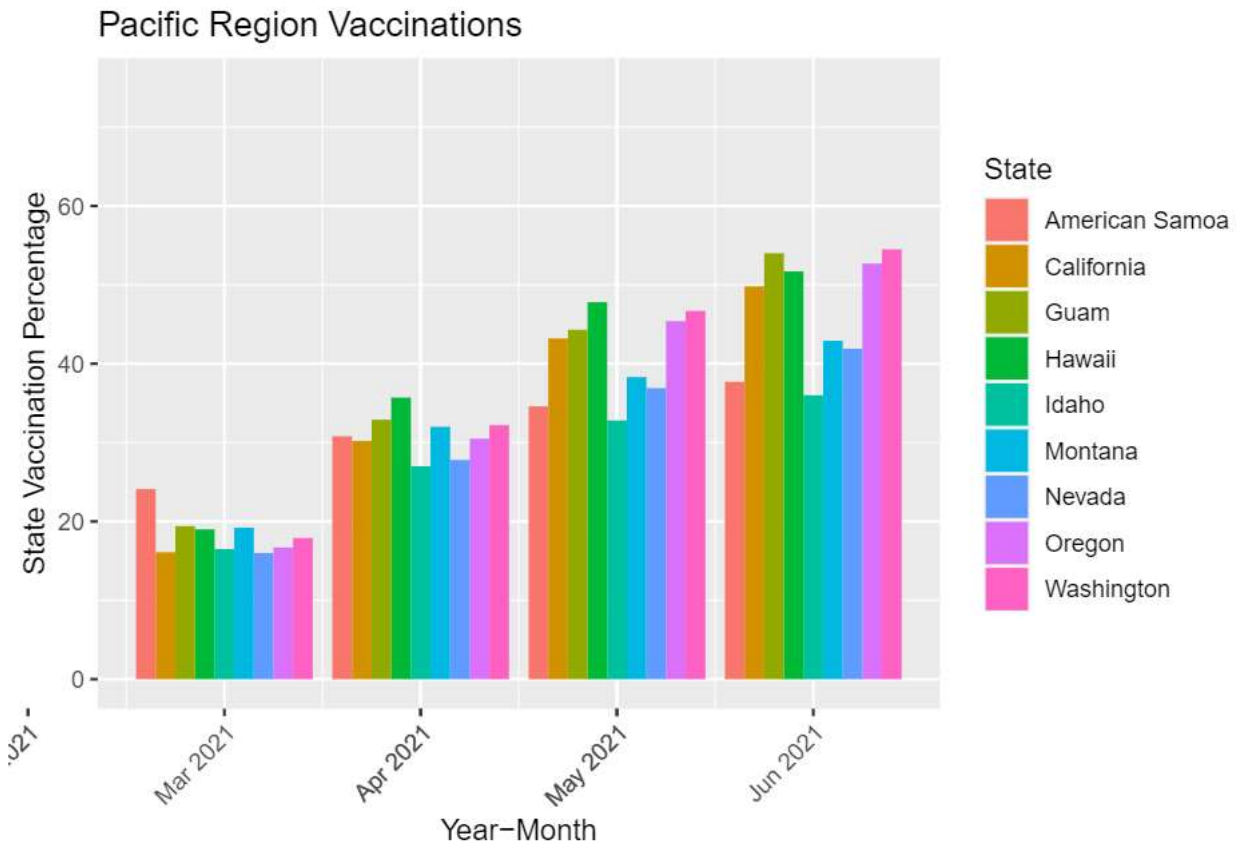


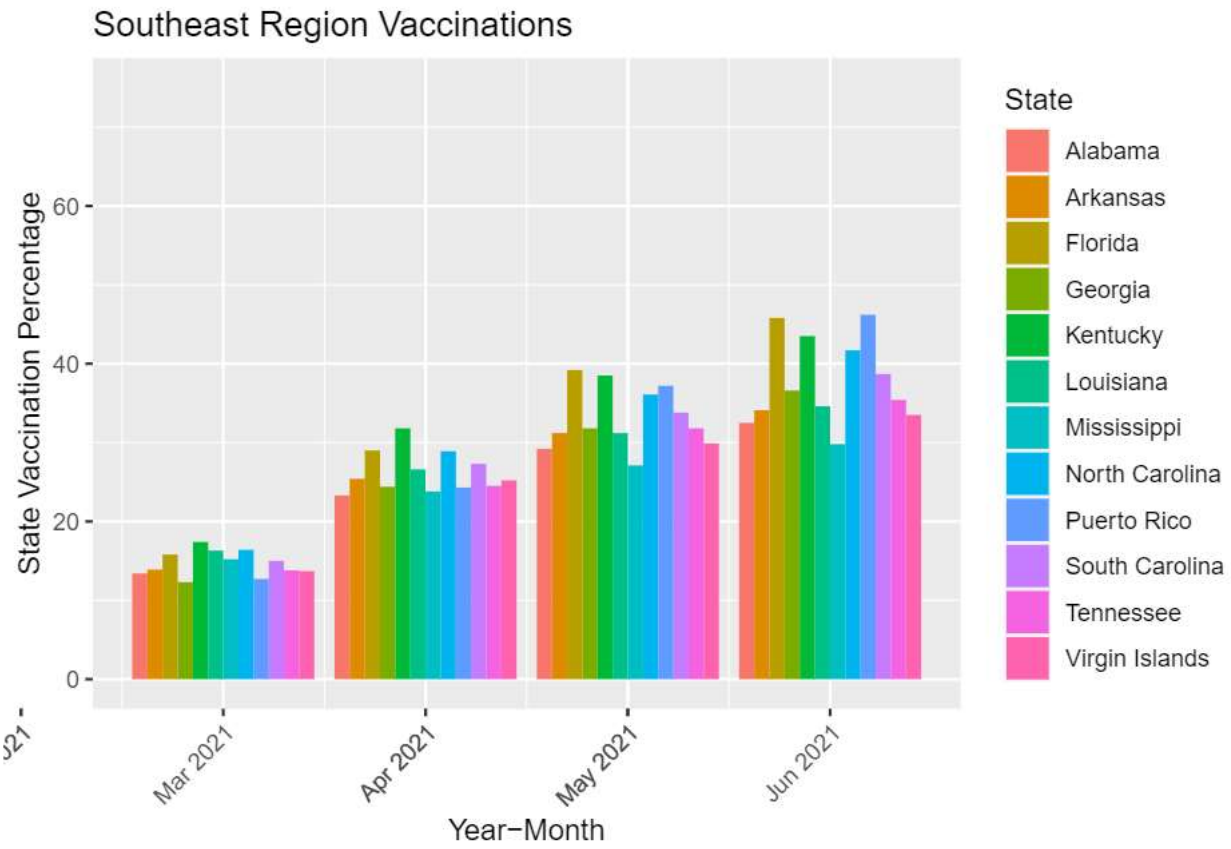

Intermountain Region Vaccinations
State Vaccination Percentage
60
40
20
0
Mar 2021
Apr 2021
May 2021
Jun 2021
Year−Month
State
Arizona
Colorado
Montana
New Mexico
North Dakota
Oklahoma
Texas
Utah
Wyoming


Midwest Region Vaccinations
State Vaccination Percentage
60
40
20
0
Mar 2021
Apr 2021
May 2021
Jun 2021
Year−Month
State
Arkansas
Illinois
Indiana
Iowa
Kansas
Michigan
Minnesota
Missouri
Nebraska
North Dakota
Ohio
South Dakota
Wisconsin


Capital Region Vaccinations
State Vaccination Percentage
60
40
20
0
Mar 2021
Apr 2021
May 2021
Jun 2021
Year−Month
State
District of Columbia
Maryland
Virginia
West Virginia


Northeast Region Vaccinations
State Vaccination Percentage
60
40
20
0
Mar 2021
Apr 2021
May 2021
Jun 2021
Year−Month
State
Connecticut
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
Virginia
West Virginia


Pacific Region Vaccinations
State Vaccination Percentage
60
40
20
0
Mar 2021
Apr 2021
May 2021
Jun 2021
Year−Month
State
American Samoa
California
Guam
Hawaii
Idaho
Montana
Nevada
Oregon
Washington



Regression results for Models 2a, 4a from Set 1 models are shown below. They are depicted by (1) and (2) respectively in the table.

```
##
## ================================================================
##                              Dependent variable:
##                    ------------------------------------------
##                      log10(abs(park_month_diff) + 1)
##                            (1)                   (2)
## ----------------------------------------------------------------
## covid_cases              -0.00000              -0.00000**
##                          (0.00000)             (0.00000)
## one_dose_num               0.000                 0.00000
##                           (0.000)               (0.00000)
## covid_deaths             0.0001***             0.00003**
##                          (0.00001)             (0.00001)
## fully_vax_num                                  -0.00000
##                                                (0.00000)
## pandemic_indicator                              0.423***
##                                                 (0.051)
## Constant                 3.334***              3.149***
##                           (0.026)               (0.034)
## ----------------------------------------------------------------
## Observations               2,250                 2,250
## R2                         0.009                 0.039
## Adjusted R2                0.008                 0.037
## Residual Std. Error   1.117 (df = 2246)     1.101 (df = 2244)
```

```
## F Statistic         6.800*** (df = 3; 2246) 18.127*** (df = 5; 2244)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Models 2b, 5b from Set 2 models are shown below. They are depicted by (1) and (2) in the table below. These are variations of models that were run outside of the core results described in the report and are placed here for reference.

```
##
## ========================================================================
##                                    Dependent variable:
##                     ---------------------------------------------------
##                            log10(abs(park_month_diff) + 1)
##                              (1)                       (2)
## ------------------------------------------------------------------------
## covid_cases_rate            -7.917                  -31.458***
##                             (5.398)                  (5.624)
## one_dose_rate              -0.003**                  -0.004
##                             (0.001)                  (0.006)
## covid_deaths_rate          856.094***                280.666
##                             (245.289)               (244.934)
## fully_vax_rate                                       -0.008
##                                                      (0.008)
## pandemic_indicator                                   0.698***
##                                                      (0.061)
## Constant                    3.363***                 3.149***
##                             (0.029)                  (0.033)
## ------------------------------------------------------------------------
## Observations                  2,250                    2,250
## R2                            0.007                    0.064
## Adjusted R2                   0.006                    0.062
## Residual Std. Error    1.118 (df = 2246)       1.086 (df = 2244)
## F Statistic         5.649*** (df = 3; 2246) 30.582*** (df = 5; 2244)
## ========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Summary result for Model 5b is as follows.

```
##
## Call:
## lm(formula = log10(park_ty_monthly + 1) ~ covid_cases_rate +
##     one_dose_rate + covid_deaths_rate + fully_vax_rate + pandemic_indicator,
##     data = df_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1562 -0.6043  0.1974  0.8878  3.5863
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.914e+00  4.121e-02  94.980  < 2e-16 ***
## covid_cases_rate    3.849e+01  6.923e+00   5.559 3.03e-08 ***
## one_dose_rate       2.498e-02  7.957e-03   3.140  0.00171 **
## covid_deaths_rate  -2.825e+03  3.015e+02  -9.368  < 2e-16 ***
## fully_vax_rate     -2.334e-02  1.012e-02  -2.307  0.02114 *
## pandemic_indicator -4.314e-01  7.481e-02  -5.767 9.18e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.337 on 2244 degrees of freedom
## Multiple R-squared:  0.06701,    Adjusted R-squared:  0.06493
## F-statistic: 32.23 on 5 and 2244 DF,  p-value: < 2.2e-16
```