Lab 2: Vaccination Rates & New COVID-19 Cases

Kayla W., Dera C., Pavan E., Greg T.

Contents


```
# load csv files created in Lab2_Data_Prep.Rmd
vacc <- read.csv(file = "Final Datasets/vacc.csv")
vacc_covid_final_03_31 <- read.csv(file = "Final Datasets/vacc_covid_final_03_31.csv")
```

Introduction

COVID-19 vaccines are currently being administered worldwide, although vaccine efficacy was tested in Randomized Clinical Trials (RCT), little is known how if vaccinations are currently lowering COVID-19 case counts. We decided to focus our efforts on understanding how vaccination rates impact new case counts in each state, during this initial roll out of vaccination efforts.

The main research question we focused on is: "Are vaccinations causing a decrease in new COVID-19 cases?"

This model is meant to explain in simple terms, how much vaccination impacts a population's rate of COVID-19 infections. We know, from the RCT's for each approved (for emergency use) vaccine, that vaccination reduces severe COVID cases with high success, and reduces COVID cases across the board with trial efficacy rates ranging from 70-98%. At the same time, there is still a big component of vaccine effectiveness that relies on high percentages of the population being inoculated. Currently, all states are well below herd immunity levels (typically reached when population vaccination rates are greater than 70%, with some variation based on virus features).

In addition to vaccination rates being relatively low, but steadily growing, we currently are seeing the easing of restrictions and preventative measures around transmitting COVID-19 in many states, without having vaccination numbers be at a protective level. Given this phenomena, we anticipate that once we iterate our model to include additional behavioral components such as a state's COVID-19 mandates, community movement, state demographics, etc., that the model will better explain the causal relationship (assuming one exists in the data).

Our base model will look directly at the relationship between vaccination rates and new cases of COVID-19.

Data and Preparation

Our research utilizes several data sources and required a decent amount of preparation work. For details on our data preparation please see the supplemental RMD files titled Covid_Movement_DataPrep and Lab2_Data_Prep.

These RMD files and additional information on our analysis can be found at: https://github.com/ghtully/sp21_w203_Lab2_Team1.git

Data Sources Data sources used for this study are outlined and described below.

CDC Data United States COVID-19 Cases and Deaths by State over Time by the United States Centers for Disease Control and Prevention (CDC). This data is from the CDC reported aggregate counts of COVID-19 cases and death numbers daily based on the most recent numbers reported by states, territories, and other jurisdictions.

This data is reported at the daily level, however the CDC notes that there are many inconsistencies with the reporting cadence by different states, that the recommendation and the common metric used is the weekly count of new cases. This can be seen when exploring the daily data and we see that some states report zeros for several days and only report larger numbers on two days a week. Or, when there are some days with negative case counts which are designed to be corrections from previous days. For this reason, we went with the recommended aggregation of data to the weekly counts.

Our World in Data (OWID) & Vaccination Rates Data on United States COVID-19 (coronavirus) vaccinations by Our World in Data (OWID). The dataset relies on data updated daily by the United States Centers for Disease Control and Prevention. The dataset provides key vaccination distribution data for each state such as the total number of doses administered, total number of people who received at least one vaccine dose, share of vaccination doses administered among those recorded as shipped in CDC's Vaccine Tracking System etc.

KFF Data Policy and Actions COVID-19 Data and Policy Actions – Policy Actions by Kaiser Family Foundation (KFF). This dataset contains state level actions and policies put in place by the governors to slow the spread of the virus in the state. The dataset includes mandates ordered by a state's executive branch like imposing mandatory stay at home orders, closing or limiting capacity at non-essential businesses, restaurants, and bars, limiting large gatherings and requiring face masks. This data set is helpful to understand the different measures put in place state by state that can affect the spread of the virus and help understand the number of cases in each state.

COVID-19 US State Policy Database compiled by researchers at the Boston University School of Public Health. This dataset tracks the dates when each US state implemented new social safety net, economic, and physical distancing policies in response to the COVID-19 pandemic, combined with data on existing health and social policies and information on state characteristics. The dates in this compiled database were used to validate the policy actions dataset by KFF. We utilized the state demographics characteristics that could potentially influence the spread of the virus such as population, population density, unemployment, etc.

Community Mobility Report COVID-19 Community Mobility Report by Google is a dataset that provides a metric of movement compared to a baseline for the given community. The baseline period is a five week period from January- midFebruary 2020, prior to the pandemic. Data is reported in percent change from this baseline for given counties, states, and countries. It is also parsed out into different environments, adding some nuance to knowing what kind of spaces people are moving more or less in (e.g. workplace, retail, recreation, residential). While this data set has nuanced county level data by day, we aggregated these metrics to state and month level for each category. Furthermore, since these numbers are a relative value compared within each state, we transformed these metrics to a categorical variable highlighting which communities are moving Less, Average Amounts, or High Amounts (see EDA discussion below).

Exploratory Data Analysis and Transformations of Main Dependent and Independent Variables

Time series consideration: As our research question seeks to explore the causal relationship between % population vaccinated and COVID-19 cases and since the data we have available is in the form of a time

series, we need to consider from the outset, how to ensure that the data we use for the model is independent and identically distributed (IID) so that it can satisfy the core assumption of the Central Limit Theorem and be used for hypothesis testing.

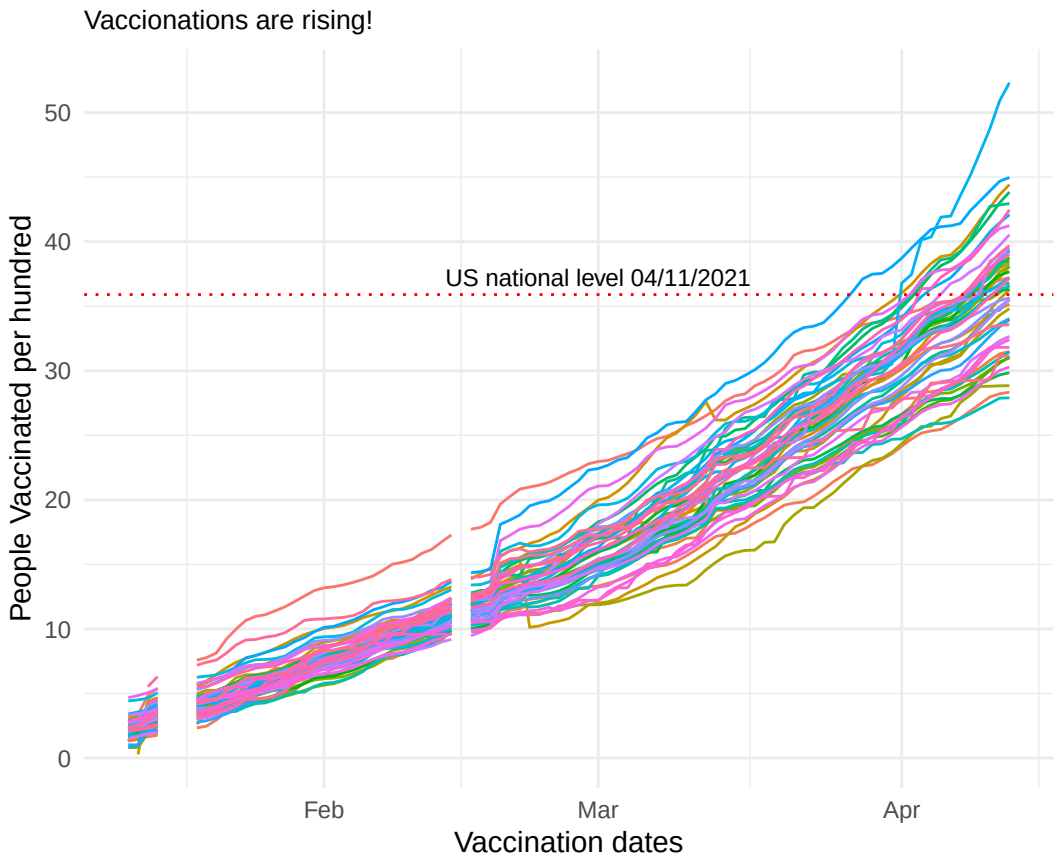
In looking at the % population vaccinated data, it is clear that the data is not independent as for any given time point, the data depends on the previous time point (e.g., today CA is 25% and yesterday it was 23%, these two data points are dependent). We considered multiple approaches to make the datapoint independent and decided that the safest way to ensure independence is to choose only one date and take the % pop vaccinated for each state at that date. Thus we have 51 (including DC) data points. Although this reduces our sample size from ~6,000] to just 51 it removes the risk of not being IID due to time dependency. Later in our model limitations discussion, we review other issues the data may have with regard to IID.

Given that we will only choose one date and that we are asking if there is a causal relationship, we aim to choose a data as current as possible so that the vaccine is potentially prevalent enough to have an effect on a population's COVID-19 incidence rate.

EDA of Vaccination data: We selected the vaccination dataset described above because it provides total people vaccinated per 100 people on a daily basis. It also provides data on people fully vaccinated per 100. We decided to use just people vaccinated (one or two shots) as studies indicate that efficacy can be as high as 85% after just one shot.

First we looked at how % vaccinated increases (it can only increase) over time starting on Jan 12, 2021 (when the dataset starts tracking % vaccinations) up to April 11, 2021. The plot below shows that % vaccination are rising and by 4/11, one state has passed 50% while most range 27% - 45% which aligns with US national total of ~36% of people with at least one dose as reported by the CDC (NPR).

```
# Time series of % pop vaccinated over time (up to yesterday)
plot_vacc_0 <- vacc %>%
  ggplot(aes(x = as.Date(date), y = people_vaccinated_per_hundred)) +
  geom_line(aes(color = abb))  +
  theme(legend.position = 'none', plot.margin = margin(0.1, 2.6, 0.1, 0.1, "cm")) +
  labs(title = "Vaccionations are rising!",
       x = "Vaccination dates",
       y = "People Vaccinated per hundred") +
  theme(plot.title = element_text(size=10), legend.position = "none")

plot_vacc_0 + geom_hline(yintercept=35.9, linetype='dotted', col = 'red') +
  annotate("text",
           x = as.Date("2021-03-01"),
           y = 35.9,
           label = "US national level 04/11/2021",
           vjust = -0.5,
           size=3)
```

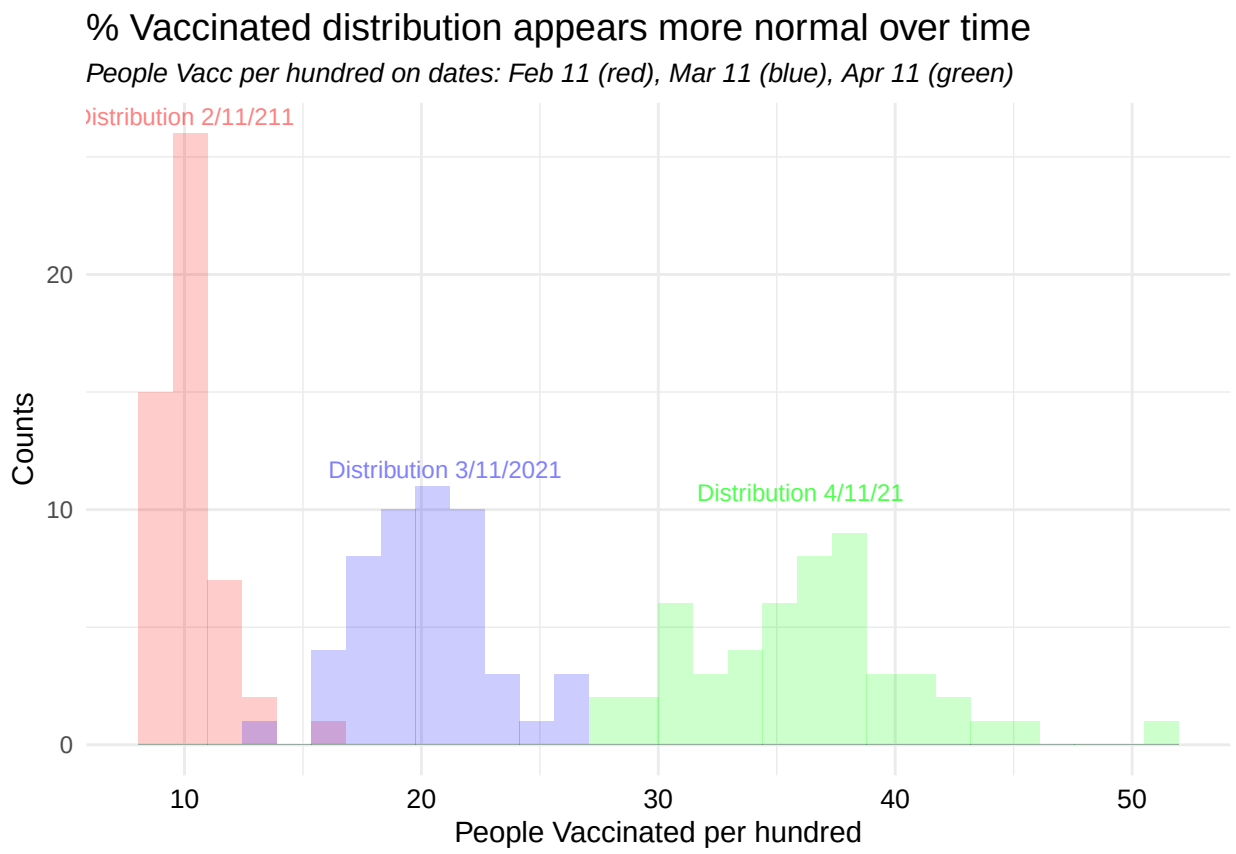


As seen in the plot of % Vaccination over time, there are a few days with no data in January and February. We were unable to identify the specific reason for missing data though we suspect it is due to changes in reporting frequency and content by the CDC during the massive rollout of the vaccination program. As seen later in the report, these missing dates will not impact our analysis.

We next aimed to see if any transformations of the data are necessary for linear regression modeling. Below is a plot of the distribution of the % of people vaccinated across the states and the distribution appears to be more normal over time.

```
# Check distribution of % Population Vaccinated over Feb, March, April
plot_vacc_distribution_over_time <- vacc %>%
  ggplot(aes(x=people_vaccinated_per_hundred)) +
  geom_histogram(data=subset(vacc, date == "2021-02-11"),fill = "red", alpha = 0.2) +
  geom_histogram(data=subset(vacc, date == "2021-03-11"),fill = "blue", alpha = 0.2) +
  geom_histogram(data=subset(vacc, date == "2021-04-11"),fill = "green", alpha = 0.2) +
  labs(title = "% Vaccinated distribution appears more normal over time",
  subtitle = "People Vacc per hundred on dates: Feb 11 (red), Mar 11 (blue), Apr 11 (green)",
  x = "People Vaccinated per hundred",
  y = "Counts") +
  theme(plot.title = element_text(size=14, color="black")) +
  theme(plot.subtitle=element_text(size=10, face="italic", color="black")) +
  theme(axis.text.x=element_text(size=10, color="black"))
plot_vacc_distribution_over_time + annotate("text",
                                            x = 10,
                                            y = 26,
                                            label = "Distribution 2/11/211",
                                            vjust = -0.5,
```

```
                                        size=3,
                                        color = "red",
                                        alpha = 0.5) +
  annotate("text", x = 21,
           y = 11,
           label = "Distribution 3/11/2021",
           vjust = -0.5,
           size=3,
           color = "blue",
           alpha = 0.5) +
  annotate("text",
           x = 36,
           y = 10,
           label = "Distribution 4/11/21",
           vjust = -0.5,
           size=3,
           color = "green",
           alpha = 0.7)
```



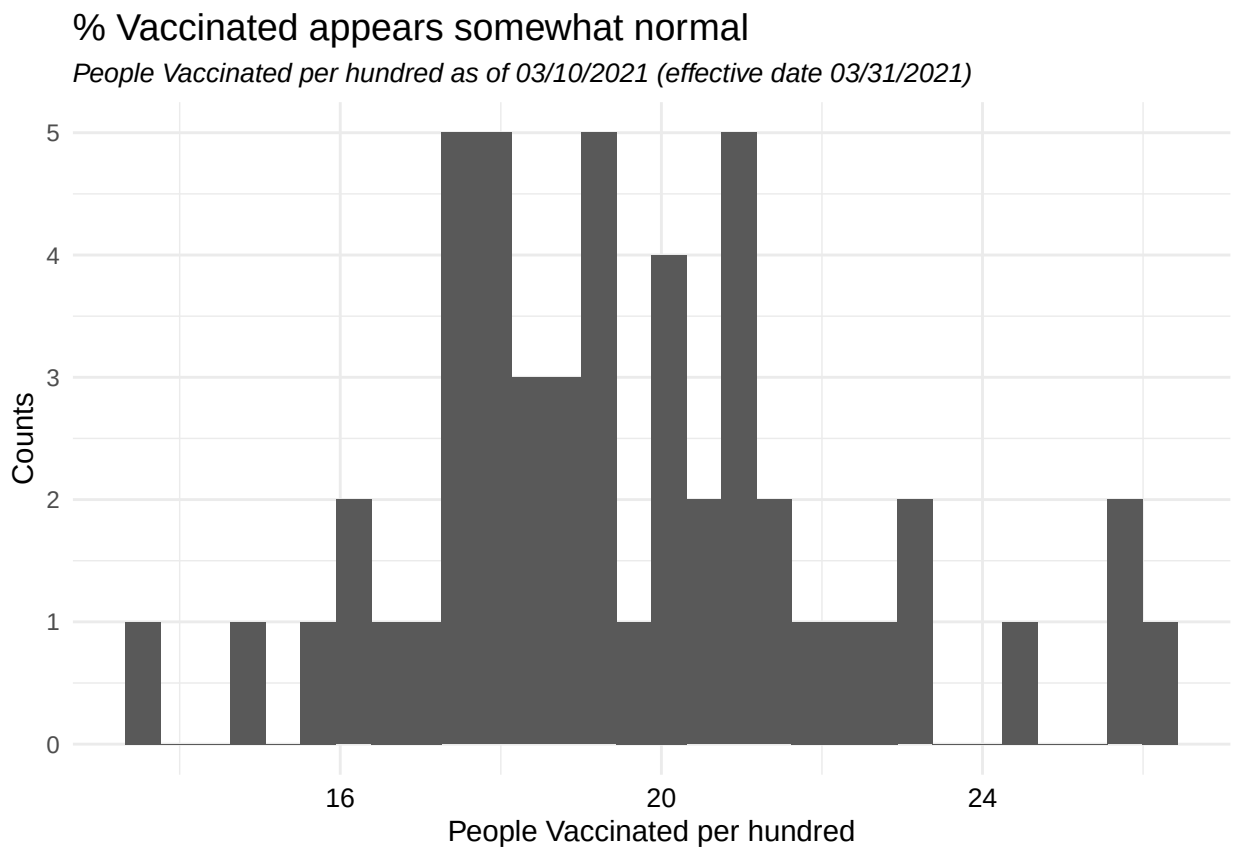
A log transformation of the data did not make the distribution appear more normal (see supplemenal RMD file: Lab2_Data_Prep for more details). Therefore, as discussed above, we will aim to choose the data as current as possible as % vaccinations appears to be distributed somewhat normal.

According to the CDC; "It typically takes two weeks after vaccination for the body to build protection (immunity) against the virus that causes COVID-19."In order to properly gauge any effect vaccination has on new covid cases, we applied a two week buffer from the vaccination date and label this date the "effective date." Covid cases over a defined time period will be evaluated with % vaccination over the same time on

the "effective dates." This means that the latest % vaccination data we can use is two weeks before the beginning of April, 2021 (time of this Lab). In addition, for additional data that we will evaluate in our model such as state COVID-19 policy mandates and community movement data the latest date is March 31, 2021. Therefore, in order to use these dataset, our set our effective date to March 31, 2021.

The distribution of % vaccinated among states is also somewhat normal.

```
# Distribution of % Vaccinated on effective date March 31, 2021
plot_vacc_effective_date_03_31 <- vacc_covid_final_03_31 %>%
  filter(effective_date == "3/31/2021") %>%
  ggplot(aes(x=people_vaccinated_per_hundred)) +
  geom_histogram() +
  labs(title = "% Vaccinated appears somewhat normal",
  subtitle = "People Vaccinated per hundred as of 03/10/2021 (effective date 03/31/2021)",
  x = "People Vaccinated per hundred",
  y = "Counts") +
  theme(plot.title = element_text(size=14, color="black")) +
  theme(plot.subtitle=element_text(size=10, face="italic", color="black")) +
  theme(axis.text.x=element_text(size=10, color="black"))

plot_vacc_effective_date_03_31
```



EDA of COVID-19 case data To gauge the impact of % vaccination in a population on COVID cases, we have to measure new COVID-19 cases and not total case count. Since we are looking at the state level and state populations differ significantly, we need to transform the number of new cases to a "per unit population metric." In its daily reporting of COVID-19 cases, the CDC uses "Cumulative Cases per 100k in Last 7 Days" as a weekly "incidence" rate of COVID-19. Adding seven days removes any risk of daily reporting inconsistencies (e.g., one state could report all cases over a weekend on a Sunday or the day

after a National holiday). Therefore, we decided the daily new case count provided in the CDC data to a Cumulative Cases per 100k in Last 7 Days as well so for any ":effective date" in the dataset, we have the sum of new cases from that date and 6 days prior, divided by the state population and multiplied by 100,000.

As mentioned above, we aim to choose the latest possible date for our analysis, March 31 is where other datasets around movement and policy go up to to so that is what we ultimately chose as the "effective date". Below is a description of the Cumulative Cases per 100k in Last 7 Days, the minimum is ~37, the maximum ~402 and nthe mean ~135.

```
vacc_covid_final_03_31 %>%
  dplyr::select(cumuluative_new_case_7_per100000) %>% summary()
```

```
##  cumuluative_new_case_7_per100000
##  Min.   : 36.92
##  1st Qu.: 75.21
##  Median :124.49
##  Mean   :135.12
##  3rd Qu.:166.29
##  Max.   :402.14
```

We sanity checked that the transformations agreed with the CDC table on their website that shows daily "Cumulative Cases per 100k in Last 7 Days" and they were the equal (see supplemental RMD file: Lab2_Data_Prep).

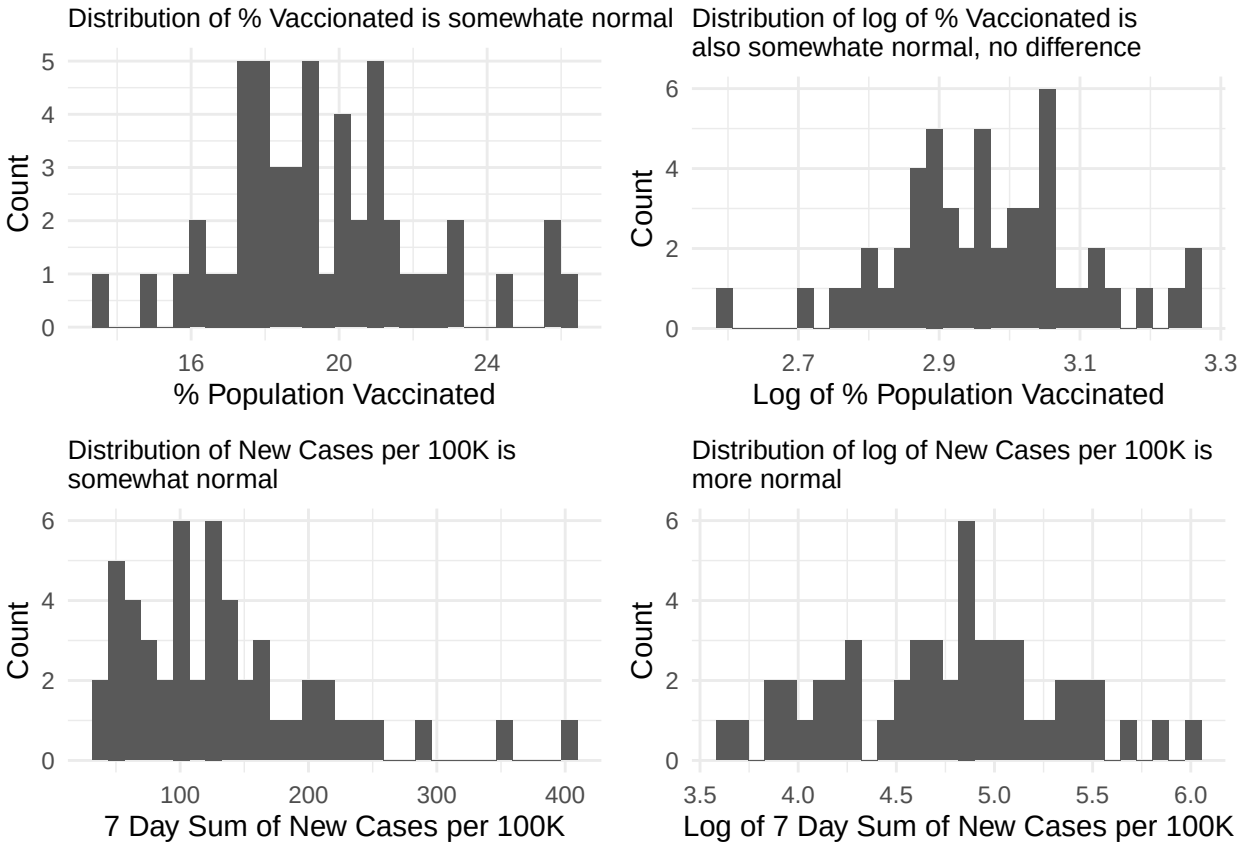
We did observe two rows in the entire CDC covid dataset (not around March 31) with negative values for new cases and decided to leave them as is because they may be corrections from the previous day and the values are small (-9 and -1).

We reviewed the distribution of Cumulative Cases per 100k in Last 7 Days and it appears skewed and with a long tail (see plot below), we also tried a Log transformation and it appears more normal in our dataset. We did this as well for the variable of Percent Population Vaccinated.

Check transformations in distribution of variables: % vaccinated and New cases

```
plot_Vacc <- vacc_covid_final_03_31 %>%
  ggplot(aes(x=people_vaccinated_per_hundred)) +
  geom_histogram() +
  labs(title = "Distribution of % Vaccionated is somewhate normal",
       x = "% Population Vaccinated",
       y = "Count") +
  theme(plot.title = element_text(size=10))

plot_Vacc_Log <- vacc_covid_final_03_31 %>%
  ggplot(aes(x=log(people_vaccinated_per_hundred))) +
  geom_histogram() +
  labs(
  title = "Distribution of log of % Vaccionated is \nalso somewhate normal, no difference",
  x = "Log of % Population Vaccinated",
  y = "Count") +
  theme(plot.title = element_text(size=10))

plot_New_Cases <- vacc_covid_final_03_31 %>%
  ggplot(aes(x=cumuluative_new_case_7_per100000)) +
  geom_histogram() +
  labs(title = "Distribution of New Cases per 100K is \nsomewhat normal",
```

```
       x = "7 Day Sum of New Cases per 100K",
       y = "Count") +
  theme(plot.title = element_text(size=10))

plot_New_Cases_log <- vacc_covid_final_03_31 %>%
  ggplot(aes(x=log(cumuluative_new_case_7_per100000))) +
  geom_histogram() +
  labs(title = "Distribution of log of New Cases per 100K is \nmore normal",
       x = "Log of 7 Day Sum of New Cases per 100K",
       y = "Count") +
  theme(plot.title = element_text(size=10))

grid.arrange(plot_Vacc, plot_Vacc_Log,plot_New_Cases,plot_New_Cases_log, nrow = 2, ncol = 2)
```

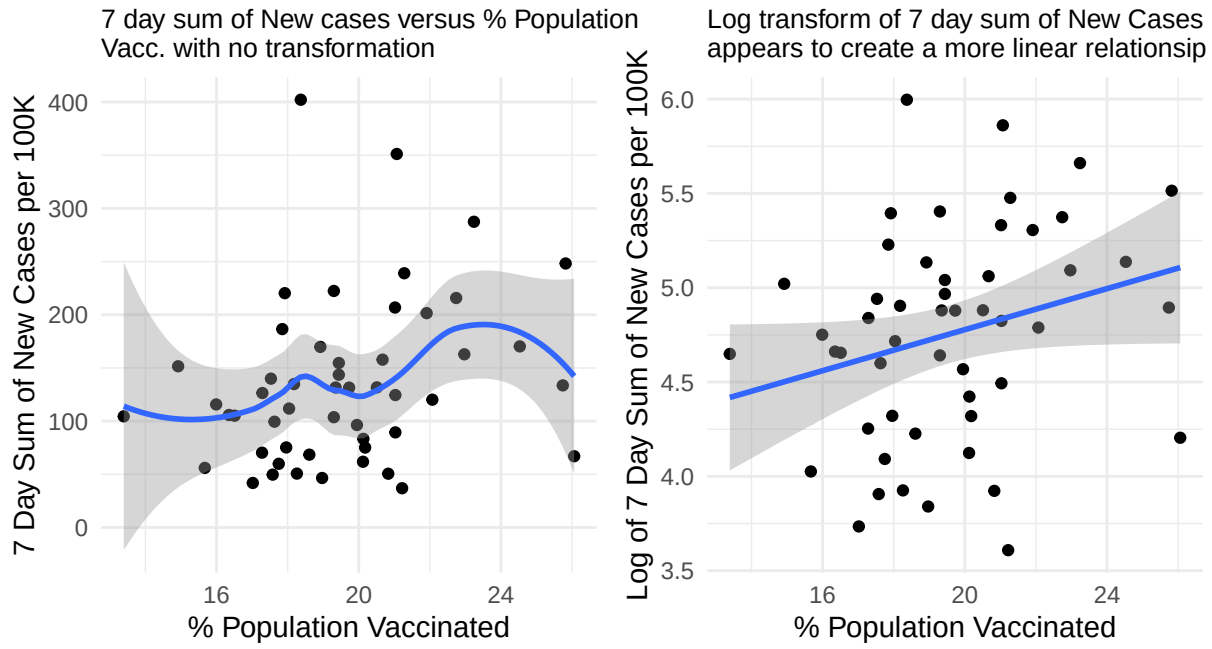


We decided to apply the log transformation to Cumulative Cases per 100k in Last 7 Days for all models. Since it is the predicted variable, interpretation of any statistically significant causal relationship would be fairy straightforward as a unit change in the variable is associated with a 100 * (Beta coefficient) percent change in Cumulative Cases per 100k in Last 7 Days.

Before entering the Model Building process using linear regression, we want to confirm that there appears to be a linear relationship between the two variables with a scatter plot. As expected, the Log Transform of Cumulative Cases per 100k in Last 7 Days with % Population Vaccinated appears to have a linear relationship. It is interesting to see that the relationship trends positively as seen in the fitted linear model. This suggests that a higher % of Population vaccinated may actually increase COVID-19 case count.

```
# Check scatter plot of both variables
```

```r
plot_3 <- vacc_covid_final_03_31 %>%
  ggplot(aes(x =  people_vaccinated_per_hundred, y = cumuluative_new_case_7_per100000)) +
  geom_point()  +
  labs(title = "7 day sum of New cases versus % Population \nVacc. with no transformation",
       x = "% Population Vaccinated",
       y = "7 Day Sum of New Cases per 100K") +
  theme(plot.title = element_text(size=10), aspect.ratio=1) +
  geom_smooth()

plot_4 <- vacc_covid_final_03_31 %>%
  ggplot(aes(x =  people_vaccinated_per_hundred,
             y = log(cumuluative_new_case_7_per100000))) +
  geom_point() +
  labs(title = "Log transform of 7 day sum of New Cases \nappears to create a more linear relationsip",
       x = "% Population Vaccinated",
       y = "Log of 7 Day Sum of New Cases per 100K") +
  theme(plot.title = element_text(size=10), aspect.ratio=1) +
  geom_smooth(method = lm)

plot_3 | plot_4
```



Model Building

Base Model For our base model we kept our variables simple and directly targeting our question. We used the percentage of the population vaccinated as our predictor variables and the count of new cases as our outcome. Our count of new cases of Covid were calculated as a weekly count with a three week delay to

account for vaccination efficacy (REF). Based on our exploratory data analysis, we used a log transformation of our count of new cases, and kept the % population vaccinated untransformed.

Our base model is:

$$f(\text{Log(Count New Cases)}) = \beta_0 + \beta_1(\text{Percent Population Vaccinated})$$

Running this model we get the following results:

```
#Build base model
model_one <- lm(log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred,
                data = vacc_covid_final_03_31)

model_one
```

```
##
## Call:
## lm(formula = log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred,
##     data = vacc_covid_final_03_31)
##
## Coefficients:
##                   (Intercept)  people_vaccinated_per_hundred
##                       3.69019                        0.05437
```

```
coeftest(model_one, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                   3.690186   0.556829  6.6271 2.509e-08 ***
## people_vaccinated_per_hundred 0.054373   0.028701  1.8945   0.06407 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see from this model that People Vaccinated per 100, or % people vaccinated, is significant at a p-value of .1. This has a coefficient of .05437, giving us a model of:

$$f(\text{Log(Count New Cases)}) = 3.69 + .054(\text{Percent Population Vaccinated})$$

This model suggests that with each percentage increase in population vaccinated we see a slight *increase* in COVID-19 cases. This is certainly a counter intuitive result, as we know from testing that vaccines reduce infections.

It is at this point that we acknowledged and explored the idea that there are many other variables that are critical to increasing COVID-19 case counts. Largely, we zeroed in on the idea of "risky behavior". "Risky behavior", in this context, is defined as behaviors that may increase spread of the virus and don't adhere to precautionary procedures. Some examples may be gathering in large groups, not wearing masks, not adhereing to social distancing.

For our second model our goal was to pull in some of these key metrics to see if our model improves.

Improved Model After running our initial model, we stopped to think critically about what other variables were at our disposal that may help explain some variation and drivers for case count. We found a counterintuitive result in our base model, which suggested that an increase in vaccination rates in a community may be driving an increase in new Covid cases. We know from vaccination trials that vaccination reduces Covid cases so this is a surprising find.

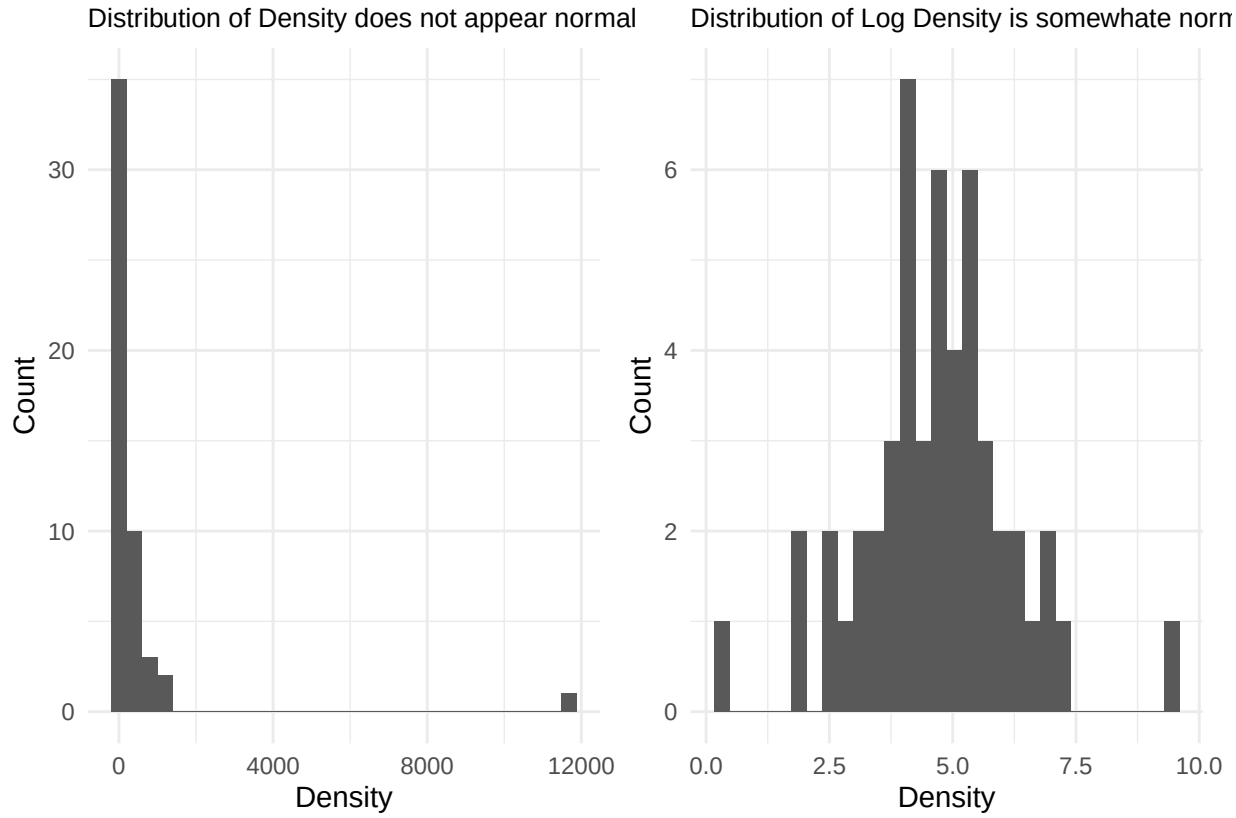
When thinking of viral transfer, we noted that there are many behavioral components that would also drive new covid cases. Our goal for the Improved Model was to identify the best metrics at our disposal to capture behaviors that may drive new case count. With this we identified both community movement data and statewide policy information. When thinking through these metrics, we decided that the raw movement data for communities would be a better proxy for "risky behavior" as it is a direct measurement of what people are actually doing. While the policy data initially seems to follow a similar logic, we realized we may not be able to account for how closely statewide policies are being followed and/or enforced.

After looking at our movement metrics we identified the Retail & Recreation, Workplaces, and Transit movement data as the most appropriate for measuring "risky behavior" in terms of congregating and spending time not social distancing. All of these variables are providing a relative measurement of movement over a specific month (e.g., 3./31) and comparing the amount of movement to a baseline pre-COVID. As the percentage change is relative to the states themself, the numbers are not a comparison between states. It is possible that some states may have a negative percent change from baseline but still be actually moving more than a state with positive percent change (assuming they were an extremely high moving state before). Given this uncertainty in our understanding of the data and how numbers compare across states we decided to take out some granularity by categorizing and grouping a state into one of three groups (Low - bottom quartile, Medium two middle quartiles, High top quartile) based on the average change over the three moths, Jan to March, 202. By grouping the data like this, we are less concerned about that actual number and instead, trying to capture a signal to see which states people are moving or less.

A second category of metrics we decided to add State population density. Given that Covid is an airborne transmitted disease, the density of people is likely a driving factor in new cases. Of course in every state there are highly dense metropolitan areas and low density rural areas so this metric does not capture this detail however, we believe it can be an important causal factor that we can use to reduce error in the model.

We transformed Density with a Log transformation to make the data appear more normal. See histograms below comparing non-transformed to Log transformed.

```
plot_Dens <- vacc_covid_final_03_31 %>%
  ggplot(aes(x=Density)) +
  geom_histogram() +
  labs(title = "Distribution of Density does not appear normal",
       x = "Density",
       y = "Count") +
  theme(plot.title = element_text(size=10))

plot_Dens_log <- vacc_covid_final_03_31 %>%
  ggplot(aes(x=log(Density))) +
  geom_histogram() +
  labs(title = "Distribution of Log Density is somewhate normal",
       x = "Density",
       y = "Count") +
  theme(plot.title = element_text(size=10))

plot_Dens | plot_Dens_log
```




```
#Add in dummy variables for the Policy information so its usable in the model:
vacc_covid_final_03_31$Bussiness_Flag <- ifelse(vacc_covid_final_03_31$Non.Essential.Business.Closures =

vacc_covid_final_03_31$Bar_Flag <- ifelse(vacc_covid_final_03_31$Bar.Closures. == "Closed", 0, ifelse(va
vacc_covid_final_03_31$Mask_Flag <- ifelse(vacc_covid_final_03_31$Statewide.Face.Mask.Requirement == "Y

vacc_covid_final_03_31$Workplace_flag <- ifelse(vacc_covid_final_03_31$workplace_mean == "HIGH", 2, ifel

vacc_covid_final_03_31$Retail_flag <- ifelse(vacc_covid_final_03_31$retail_mean == "HIGH", 2, ifelse(va

vacc_covid_final_03_31$Parks_flag <- ifelse(vacc_covid_final_03_31$parks_mean == "HIGH", 2, ifelse(vacc_

vacc_covid_final_03_31$Grocery_flag <- ifelse(vacc_covid_final_03_31$grocery_mean == "HIGH", 2, ifelse(

vacc_covid_final_03_31$Transit_flag <- ifelse(vacc_covid_final_03_31$transit_mean == "HIGH", 2, ifelse(

vacc_covid_final_03_31$Res_flag <- ifelse(vacc_covid_final_03_31$residential_mean == "HIGH", 2, ifelse(
#Create a dummy variable for positive and negative movement values:

vacc_covid_final_03_31$Workplace_pos <- ifelse(vacc_covid_final_03_31$workplaces_percent_change_from_bas

vacc_covid_final_03_31$Retail_pos <- ifelse(vacc_covid_final_03_31$retail_and_recreation_percent_change_

vacc_covid_final_03_31$Parks_pos <- ifelse(vacc_covid_final_03_31$parks_percent_change_from_baseline_202

vacc_covid_final_03_31$Grocery_pos <- ifelse(vacc_covid_final_03_31$grocery_and_pharmacy_percent_change_
```

```
vacc_covid_final_03_31$Transit_pos <- ifelse(vacc_covid_final_03_31$transit_stations_percent_change_fro

vacc_covid_final_03_31$Res_pos <- ifelse(vacc_covid_final_03_31$residential_percent_change_from_baselin

# Check all columns are created correctly
# head(vacc_covid_final_03_31)
```

Since we decided to focus on movement and not policies, our second model is the following:

$f(\text{Log(Count New Cases)}) = \beta_0 + \beta_1(\text{Percent Population Vaccinated}) + \beta_2(\log(\text{Density}) + \beta_3(\text{Workplace Movement}) + \beta_4(\text{Retail}$

The results of this model are:

```
model_two <- lm(log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred + log(Density) +
Workplace_flag +
Retail_flag +
Transit_flag,
data = vacc_covid_final_03_31)
model_two
```

```
##
## Call:
## lm(formula = log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred +
##     log(Density) + Workplace_flag + Retail_flag + Transit_flag,
##     data = vacc_covid_final_03_31)
##
## Coefficients:
##                   (Intercept)  people_vaccinated_per_hundred
##                       2.66556                        0.06803
##                   log(Density)                 Workplace_flag
##                       0.16546                        0.39167
##                    Retail_flag                   Transit_flag
##                      -0.16808                       -0.23609
```

```
coeftest(model_two, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.665557   0.635621  4.1936 0.000127 ***
## people_vaccinated_per_hundred  0.068029   0.020325  3.3471 0.001657 **
## log(Density)                   0.165464   0.050027  3.3075 0.001857 **
## Workplace_flag                 0.391674   0.157921  2.4802 0.016939 *
## Retail_flag                   -0.168079   0.200767 -0.8372 0.406913
## Transit_flag                  -0.236087   0.142132 -1.6610 0.103658
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we see that while our main variable of % People vaccinated is still positive, it is now significant to p= .01, as is log(Density), and Workplace Movement. Retail and Transit movement are negative coefficients which, again is a bit unintuitive as increasing movement suggests lower COVID-19 cases, at the same time these variables are not significant in themselves.

Comparing our Base Model and the Improved Model with an F-Test we can see if there is a statistically significant improvement:

```
anova(model_one, model_two)
```

```
## Analysis of Variance Table
##
## Model 1: log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred
## Model 2: log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred +
##     log(Density) + Workplace_flag + Retail_flag + Transit_flag
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     49 14.6606
## 2     45  9.3814  4    5.2791 6.3306 0.0003952 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-test for Model two is statistically significant in this dataset, therefore we can reject the null that Model performs just as well in this dataset and Model Two appears to be a better performing model in this dataset.

Checking the CLM Assumptions for the Improved Model Following are the five assumption of the CLM.

1. IID Sampling
2. Linear Conditional Expectation
3. No Perfect Collinearity
4. Homoskedastic Errors
5. Normally Distributed Errors

`IID Sampling:`

To assess whether our data meets IID assumptions we need to walk through each data source used. The critical component for IID is that all data points are independent and identically distributed, and approach normal as we increase our sample size.

COVID-19 case counts - this data does look identically distributed and roughly normal. At the same time, its possible these are not fully independent. People move between states and we know that viral spread occurs from travelers. While some "edge effects" like this can be small, at this point it is difficult for us to assess how impactful this may be. We did not look at travel data, nor did we look at case counts for states that belong to non-residents. Many states count cases for their residents even if that resident is outside of the state when diagnosed, which obviously presents a problem with the independence of this metric as that individual is infected in a state they do not reside.

Vaccine distribution - Vaccine distribution in many cases is limited to the amount of vaccine available to the state. We did not evaluate if all states use all vaccines provided, although recent news reports suggest some states are not maxing out their vaccinations. It is currently unclear if this is driven by individuals not opting in, the states not offering the vaccines, and/or difficulty for residents to register and/or travel to vaccination sites. Regardless, there is some maximum vaccination amount that is limited by the amount of vaccine available. So, the data used for this metric is not fully independent from each other. However, we can see from our data that vaccination rates do have a normal distribution for our sample. Once limitations on availability are removed, then this metric will be IID.

Movement data- Movement data is independent between the states and has a normal distribution. Movement data is not tied to any individuals and does not matter if the people moving are residents of the state or visitors. It is a measure of people in the community interacting and leaving their homes, and is there fore an independent data point for each state. This data is IID.

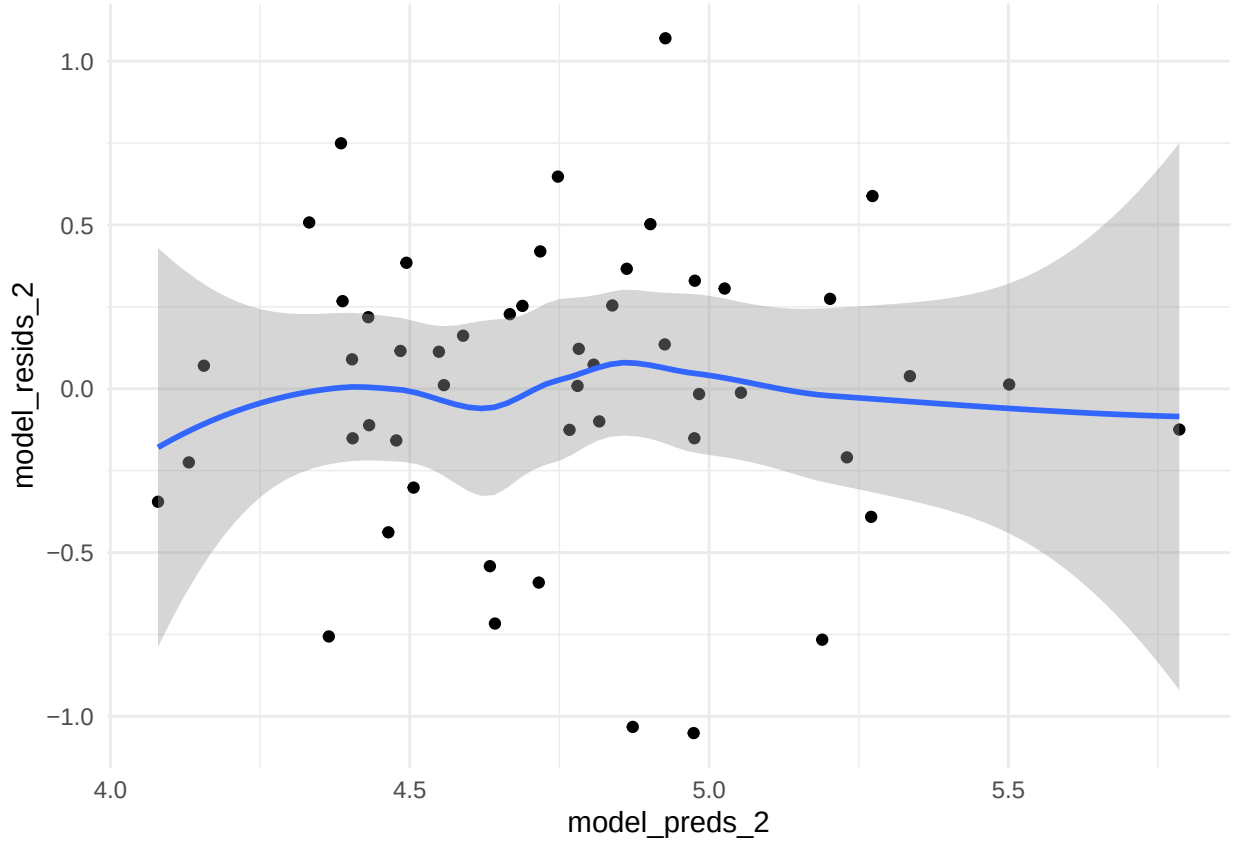
Policy data- Similarly, policy data meets the IID requirements. This data is independent from each other, as each state implements its own policies, restrictions, and consequences. Most of these varialbes are binary, or categorical with 3-4 groups.

Density- Density data is both normal (with a log transformation) and independently drawn from state demographics. This data meets the IID requirements.

`Linear Conditional Expectation:`

To check whether there is a linear conditional expectation, we looked at the predicted values vs. residuals of our improved model (model_two). It appears that the residuals are fairly evenly distributed above and below the expectation line. The line is mostly linear between predicted values of 4.2 & 5.2 with a slight dip between 4.5 and ~4.75. Given that most of our samples are concentrated between 4.5 & 5.0 We assume that Linear conditional expectation is not an issue and the slight deviation from linearity may be not be significant and would have been caused due to some noise in the data.

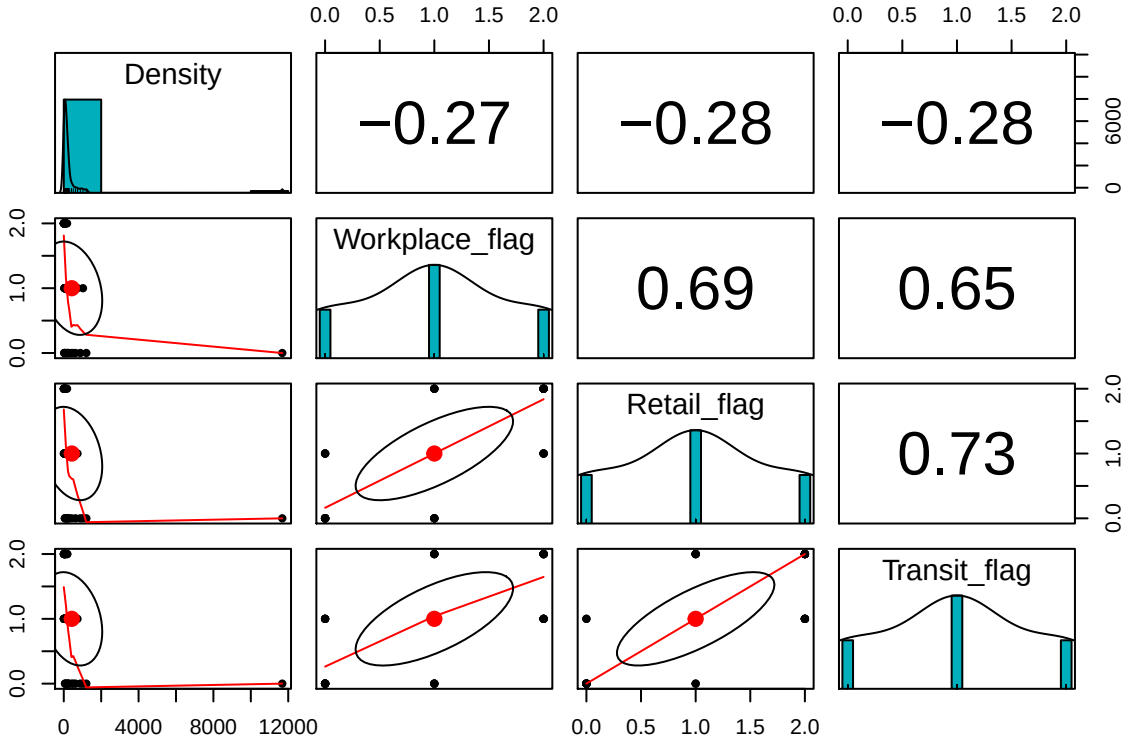
```
vacc_covid_final_03_31['model_preds_2'] <- predict(model_two)
vacc_covid_final_03_31['model_resids_2'] <- resid(model_two)

vacc_covid_final_03_31 %>%
  ggplot(aes(model_preds_2, model_resids_2)) +
  geom_point() +
  stat_smooth()
```



`No Perfect Collinearity:`

To check for Perfect Collinearity among our model variables, we looked at Pearson correlation coefficients and bivariate scatter plots between each of the variables. The correlation coefficients seem to suggest that there is a positive correlation between the flag variables as in the Improved Model (model_2) case as well. Each of the flag variables have been independently derived from the actual dataset and so, we don't anticipate there is any perfect collinearity. We will inspect the coefficients and also the output of VIF to further understand the collinearity between the variables.

```
model_2_vars <- vacc_covid_final_03_31[c('Density','Workplace_flag','Retail_flag',
                                         'Transit_flag')]

pairs.panels(model_2_vars,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```



We also checked if R has dropped any columns when fitting the model. We see that all of our variables have coefficients that are not equal to zero, suggesting that there is no perfect collinearity.

```
model_two$coefficients
```

```
##                  (Intercept) people_vaccinated_per_hundred
##                   2.66555658                     0.06802922
##                log(Density)                Workplace_flag
##                   0.16546386                     0.39167409
##                 Retail_flag                  Transit_flag
##                  -0.16807929                    -0.23608723
```

Another test we conducted to check for perfect collinearity was to use a VIF command. We see here that while they are all greater than one, there doesn't appear to be any high VIF numbers suggesting we don't have significant collinearity between our variables.

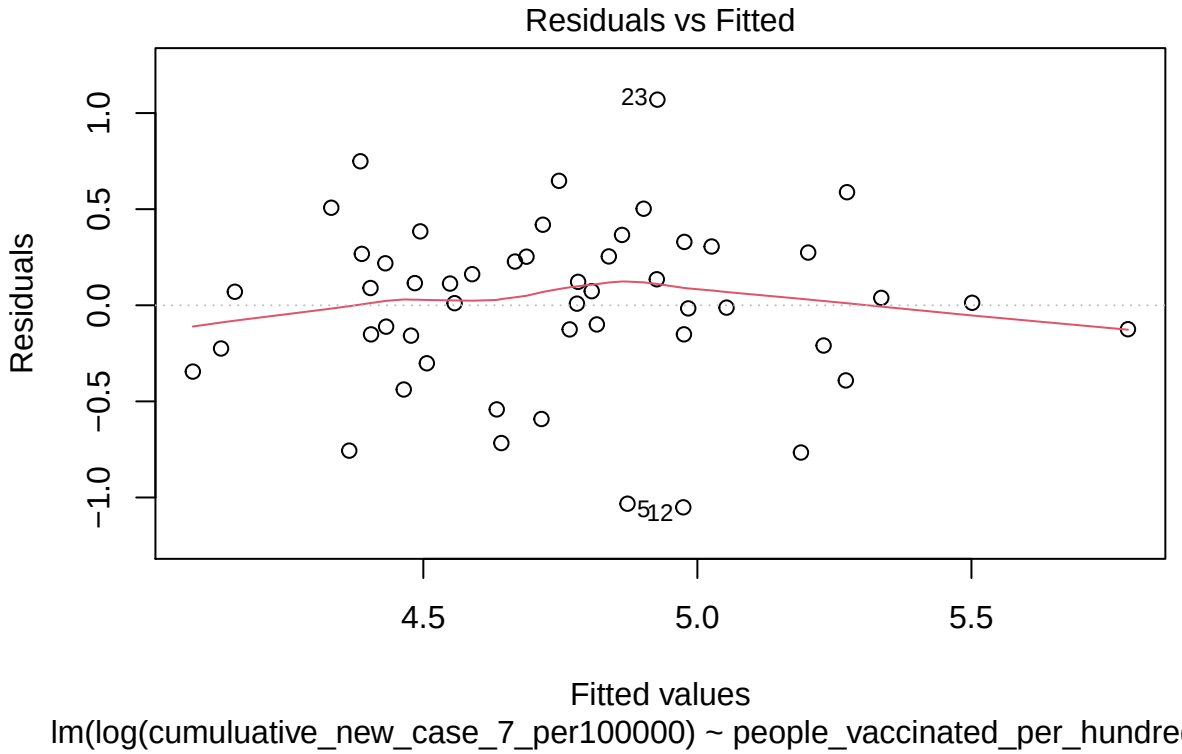
```
vif(model_two)
```

```
## people_vaccinated_per_hundred                log(Density)
##                       1.242642                    2.009528
##                Workplace_flag                 Retail_flag
##                       2.262572                    2.777849
##                   Transit_flag
##                       2.561925
```

Homoskedastic Errors:

To assess whether the distribution of the errors is homoskedastic, we examined the residuals versus fitted plot. From the plot it looks like the points above and below our expectation line are evenly spaced out with a bit more dense and "heavy" above the line but not significant. There are few residuals on the left and on the extreme right of the plot that are causing some unevenness but not too significant. This looks nice and balanced for the most part. We will perform Breush Pagan test to further evaluate if heteroskedasticity is a concern.

```
plot(model_two, which=1)
```



To further test for homoskedastic errors we also ran the Breush Pagan test:

```
bptest(model_two)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_two
## BP = 5.4459, df = 5, p-value = 0.3639
```

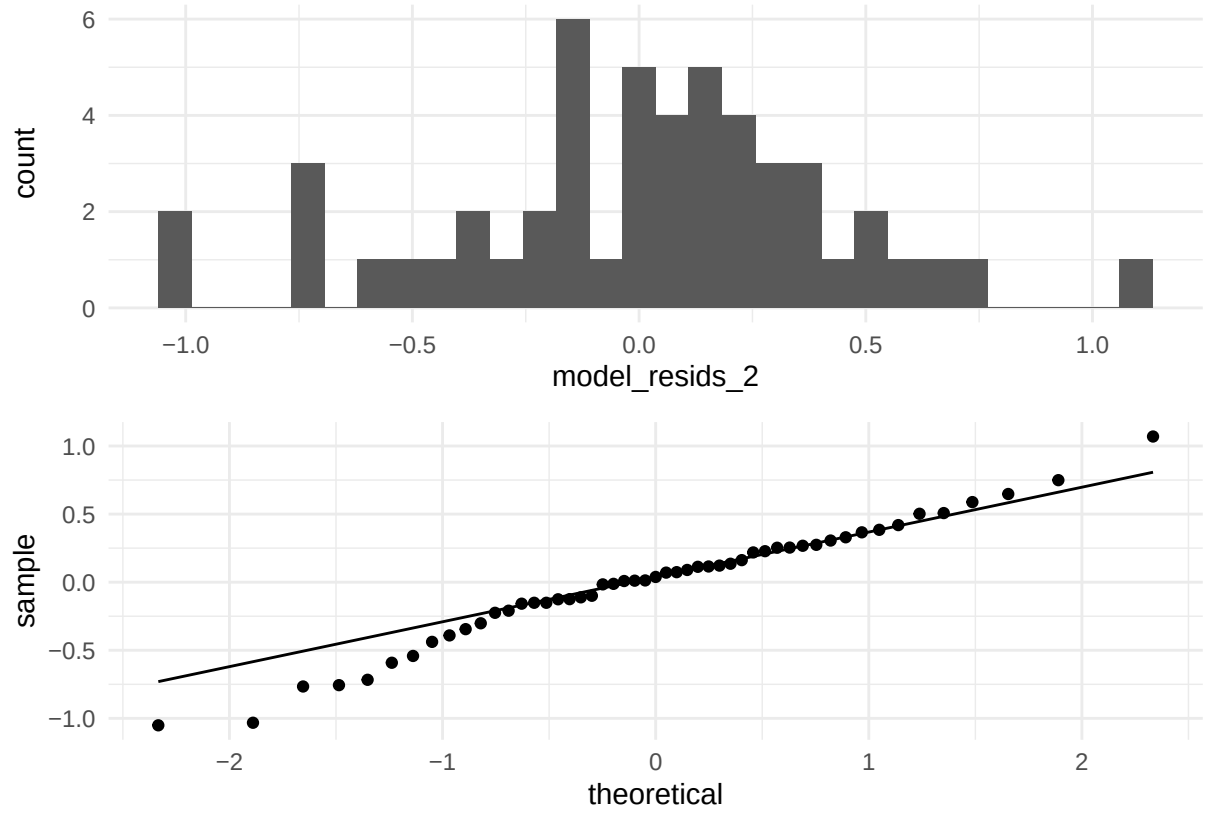
This test fails to reject the null hypothesis, which means we do not have sufficient evidence that heteroskedacity

is present in the model. However, this is still a potential source of error, and as a result we will be using the robust standard errors to combat this as much as possible.

`Normally Distributed Errors:`

To check for Normally distributed errors, we looked at the The histogram of residuals which shows that the distribution is mostly normal with some deviation from normality, specifically there appears to be long tails on the left and the right.

```
plot_one <- vacc_covid_final_03_31 %>%
  ggplot(aes(x = model_resids_2)) +
  geom_histogram()

plot_two <- vacc_covid_final_03_31 %>%
  ggplot(aes(sample = model_resids_2)) +
  stat_qq() + stat_qq_line()

plot_one / plot_two
```



We also looked at the qq plot which confirms the same with the distribution almost being close to the normal line but with some deviation towards bottom left and at the extreme right but not significant. Therefore we assume that normality of errors is satisfied and not a concern. This is something we will want to be aware of as we analyze our model.

Overall, our model fits most of the CLM assumptions, while identifying a few key components that we should be aware of when interpreting our model and applying it elsewhere.

Throwing the Kitchen Sink at It Our third and final model we included the other policy variables we initially omitted. We had initially removed these due to the inability to identify if these policies are being adhered too (see model limitations below). Here we will add them in and see if our model shows improvement.

The "Kitchen Sink" Model is as follows:

$$f(\text{Log(Count New Cases)}) = \beta_0 + \beta_1(\text{Percent Population Vaccinated}) + \beta_2(\text{log(Density)} + \beta_3(\text{Workplace Movement}) + \beta_4(\text{Retail}$$

The results of this model are as follows:

```
model_three <- lm(log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred + log(Density) +
model_three
```

```
##
## Call:
## lm(formula = log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred +
##     log(Density) + Workplace_flag + Retail_flag + Transit_flag +
##     Bussiness_Flag + Bar_Flag + Mask_Flag + Parks_flag + Grocery_flag,
##     data = vacc_covid_final_03_31)
##
## Coefficients:
##                   (Intercept)  people_vaccinated_per_hundred
##                       2.55628                        0.07501
##                  log(Density)                Workplace_flag
##                       0.14708                        0.28308
##                   Retail_flag                   Transit_flag
##                      -0.21302                       -0.34382
##                Bussiness_Flag                      Bar_Flag
##                       0.34765                       -0.23279
##                     Mask_Flag                    Parks_flag
##                      -0.02219                        0.08872
##                  Grocery_flag
##                       0.09085
```

```
coeftest(model_three, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.556280   0.904969  2.8247 0.007346 **
## people_vaccinated_per_hundred  0.075011   0.028445  2.6371 0.011851 *
## log(Density)                   0.147084   0.058204  2.5270 0.015559 *
## Workplace_flag                 0.283076   0.179545  1.5766 0.122758
## Retail_flag                   -0.213018   0.221253 -0.9628 0.341443
## Transit_flag                  -0.343816   0.174081 -1.9750 0.055194 .
## Bussiness_Flag                 0.347654   0.290626  1.1962 0.238652
## Bar_Flag                      -0.232790   0.217567 -1.0700 0.291048
## Mask_Flag                     -0.022192   0.284057 -0.0781 0.938118
## Parks_flag                     0.088718   0.130827  0.6781 0.501594
## Grocery_flag                   0.090847   0.178501  0.5089 0.613585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we compare the Improved Model to the Kitchen Sink model we see the following:

```
anova(model_two, model_three)
```

```
## Analysis of Variance Table
##
## Model 1: log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred +
##     log(Density) + Workplace_flag + Retail_flag + Transit_flag
## Model 2: log(cumuluative_new_case_7_per100000) ~ people_vaccinated_per_hundred +
##     log(Density) + Workplace_flag + Retail_flag + Transit_flag +
##     Bussiness_Flag + Bar_Flag + Mask_Flag + Parks_flag + Grocery_flag
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     45 9.3814
## 2     40 8.2599  5    1.1215 1.0863 0.3828
```

The F-test for the Kitchen Sink Model is not statistically significant in this dataset, therefore we cannot reject the null that the Improved Model is the same as the Kitchen Sink model. In other words, we do not get more explanatory power from adding the additional variables.

Regression Table and Model Comparison

To display our models in a regression table, we will use the stargazer function from the stargazer package.

First, we need to create vectors of robust standard errors for each model and then pass the standard errors into stargazer through the se argument.

% of population vaccinated is statistically significant in Model two and Model three however the coefficient slightly increases from Model One to Two to Three which implies that the variable is showing more effect on increasing new Covid cases. The adjusted R2 is highest in Model two which is also suggests it is has the most variance

```
se.model_one = coeftest(model_one, vcov = vcovHC)[ , "Std. Error"]
se.model_two = coeftest(model_two, vcov = vcovHC)[ , "Std. Error"]
se.model_three = coeftest(model_three, vcov = vcovHC)[ , "Std. Error"]
stargazer(model_one, model_two, model_three, type = "text", omit.stat = "f",
          se = list(se.model_three),
          star.cutoffs = c(0.05, 0.01, 0.001), title = "Table 1: The relationship between new covid cas
```

```
##
## Table 1: The relationship between new covid cases and % of population vaccinated
## ================================================================================
##                                          Dependent variable:
##                                 ---------------------------------------------
##                                     log(cumuluative_new_case_7_per100000)
##                                       (1)             (2)             (3)
## --------------------------------------------------------------------------------
## people_vaccinated_per_hundred      0.054          0.068*          0.075*
##                                   (0.028)         (0.027)         (0.029)
##
## log(Density)                                      0.165**         0.147*
##                                                   (0.060)         (0.061)
##
## Workplace_flag                                    0.392**          0.283
##                                                   (0.135)         (0.149)
##
## Retail_flag                                       -0.168          -0.213
##                                                   (0.149)         (0.174)
##
```

```
## Transit_flag                                -0.236          -0.344
##                                             (0.143)         (0.170)
##
## Bussiness_Flag                                               0.348
##                                                             (0.216)
##
## Bar_Flag                                                    -0.233
##                                                             (0.157)
##
## Mask_Flag                                                   -0.022
##                                                             (0.225)
##
## Parks_flag                                                   0.089
##                                                             (0.109)
##
## Grocery_flag                                                 0.091
##                                                             (0.157)
##
## Constant                          3.690***        2.666***        2.556**
##                                    (0.905)         (0.754)         (0.828)
##
## ------------------------------------------------------------------------------
## Observations                         51              51              51
## R2                                  0.069           0.404           0.475
## Adjusted R2                         0.050           0.338           0.344
## Residual Std. Error           0.547 (df = 49) 0.457 (df = 45) 0.454 (df = 40)
## ==============================================================================
## Note:                                          *p<0.05; **p<0.01; ***p<0.001
```

When comparing the values from our regression table we see a nice summary of what we explored in detail above. A note here, is that for every variable in our second model, we see an improved robust standard error from Model 1 and Model 3. This adds further support that our Improved Model (model 2) is the most successful for explaining the causal relationship between vaccination rates and new COVID- 19 Cases.

Practical Significance & Conclusion

As we've discussed, above, in all three models, People Vaccinated per hundred has a positive coefficient. In the case of our best Model (Improved Model, or Model Two), where the the predicted variable has a Log transformation, **one more person out a hundred** with a vaccine is associated with a **6.8% percent increase in Cumulative Cases per 100k in Last 7 Days**, given all else equal.

$f(\text{Log(Count New Cases)}) = \beta_0 + \beta_1(\text{Percent Population Vaccinated}) + \beta_2(\log(\text{Density}) + \beta_3(\text{Workplace Movement}) + \beta_4(\text{Retail}$

Taking a step back here, these results suggest that the answer to our initial research question; "*Do more vaccinations cause a decrease in new COVID-19 cases?*" would be"NO". Rather, we found that there is a statistically significant causal relationship between % of people vaccinated and new COVID-19 cases but in the **opposite direction** than we had expected.

This result is non intuitive and it goes against dogma because, as more people get vaccinated, more people contract COVID-19. This is a surprising result; however, it is not a novel finding and there appears to be precendence for this type of phenomenon.

In 1975, economist Sam Peltzman published a study titled "*The Effects of Automobile Safety Regulation*", (Journal of Political Economy) where he demonstrated that improvements in automobile safety (e.g, seat

belts) did not decrease the automobile death rate. He hypothesized that this was because drivers felt safer and therefore engaged in riskier behavior, leading to more car crashes and the same rate of deaths even though the cars were in fact safer.

This theory has been studied extensively and has apparently far reaching effects. It is called The Peltzman Effect (see Decision Lab website for a more background) and could help explain what we are observing with our modeling. Following the logic behind The Peltzman Effect, people (whether they are vaccinated or not) are aware that more people are vaccinated against COVID-19 and therefore, are engaging in riskier behavior (e.g., not wearing masks, socializing in groups, entering close quarters, etc.) and due to the highly contagious nature of COVID-19, more people are actually getting infected.

Of course, society has experience in using vaccines to severely diminish or even eradicate diseases. We expect that this result, if valid, will eventually weaken and likey may go in the opposite direction (i.e., statistically significant result with a negative coefficient). To establish if this is what we are seeing in the case of COVID-19 we would need to do more work to identify and test the relative levels of "risky behaviors" between the states.

This model contains many limitations (see Appendix: Model Limitations), such as sample size, scale of study, and several key ommitted variables that deserve consideration. Further work is suggested before stating any final findings. This future work should be focused on pulling in omitted variables or sufficient proxies when possible, and on analyzing the notion that some states are now engaging in "Risky Behavior" and if this is a causal relationship with the COVID-19 case increase. Based on our model, this superficially appears to be the case. However, further work should be pursued.

Appendix: Model Limitations

Sample size: We have conducted the study to observe the effect of vaccination within United States, We're only dealing with 51 data points corresponding to 51 states which is fairly small. Although we have a small sample, We have conducted tests to validate CLM assumptions to ensure that no small sample biases have been introduced into our study.

Scale: States may be too "big and clunky" for the model as variables such as population density can differ widely in a state and so can case counts. Data on vaccinations, cases in the United States is available at the county level and likely movement data can be found as well.

Omitted Variables: We identified multiple omitted variables that were not available in our datasets. As discussed below, most of these variables pull the true coefficient away from zero which means that the effect may be in fact larger than we see in our model and with our data.

Face Mask Adherence: We have omitted the variable Face Mask adherence as it was unavailable in the datasets we had access to. Given our understanding of the data, We believe Face Mask adherence has an inverse effect on the outcome variable No. of new cases. This is because as more people adhere to the Face Mask policy, the less the disease spread could be, resulting in a decrease in the number of new cases. Therefore, the sign of the beta coefficient corresponding to this variable will be 'negative'.

Although Face Mask Adherence and Vaccination Rates may not have a direct causal relationship, We anticipate that more adherence to the policy could result in less disease spread which could result in less overall vaccination rates. And so, when regressing Face Mask Adherence on the Vaccination rate variable, the sign of the delta coefficient corresponding to Vaccination rate will be 'negative' as well. Therefore, the Omitted variable bias will be a product of two 'negative' numbers resulting in a positive quantity.

From our base model, we have observed beta1_tilde to be positive. And so, the true beta coefficient has to be a smaller number compared to beta1_tilde as beta1_tilde (Positive) = beta1_true + bias (Positive) => The bias is pulling the beta coefficient away from zero in the positive direction. In other words the direction of bias is "***away from zero***."

Prevalence of Covid Variants per state: Prevalence of Covid variants could indicate that the virus may be mutating and vaccination may not be effective which could have a consequence on overall cases. We have omitted this variable as it was unavailable in the datasets we had access to. Given our understanding of the data, We believe Prevalence of Covid Variants has a positive effect on the outcome variable No. of new cases i.e. the more variants there exist, the more the cases could be.

Given that a single vaccine may not be effective in treating all the variants, We anticipate that these variants could result in increased vaccination rates. And so, when regressing Face Mask Adherence on the Vaccination rate variable, the sign of the delta coefficient corresponding to Vaccination rate will be 'positive' as well. Therefore, the Omitted variable bias will be a product of two 'positive' numbers resulting in a positive quantity.

From our base model, we have observed beta1_tilde to be positive. And so, the true beta coefficient has to be a smaller number compared to beta1_tilde as beta1_tilde (Positive) = beta1_true + bias (Positive) => The bias is pulling the beta coefficient away from zero in the positive direction. In other words the direction of bias is "***away from zero***."

Age: We have omitted the variable age as it was unavailable in the datasets we had access to. Given our understanding of the data, We believe age has a positive effect on the outcome variable No. of new cases. This is because older people could have a weaker immune system and could be at a higher risk compared to young people. Therefore, the sign of the beta coefficient corresponding to this variable will be 'positive'.

We anticipate that the older the population is the more the percentage of people vaccinated could be. And so, when we regress Age on Vaccination rate variable, the sign of the delta coefficient corresponding to Vaccination rate will be 'positive' as well. Therefore, the Omitted variable bias will be a product of two 'positive' numbers resulting in a positive quantity.

From our base model, we have observed beta1_tilde to be positive. And so, the true beta coefficient has to be a smaller number compared to beta1_tilde as beta1_tilde (Positive) = beta1_true + bias (Positive) => The bias is pulling the beta coefficient away from zero in the positive direction. In other words the direction of bias is "***away from zero***."

Returning to school: We omitted the data on how students are returning to school. We believe that more students returning to school has a positive effect on the outcome of the target variable Number of new cases. School settings promote close interaction between students which will increase the spread of the disease.

The sign of the beta coefficient corresponding to this variable will be 'positive'. We anticipate that students returning to school will be positively related to an overall increase in vaccination rates. And so, when comparing returning to school with the Vaccination rate variable, the sign of the delta coefficient corresponding to Vaccination rate will be 'positive' as well.

Therefore, the Omitted variable bias will be a product of two 'positive' numbers resulting in a positive quantity. The bias is pulling the beta coefficient away from zero in the positive direction. In other words the direction of bias is "***away from zero***."

Social gatherings: The data showing how individuals are interacting with one another in various social gatherings such as restaurants, parties, gyms, was not included in our model because the dataset was not available to us. The data describing how much social interaction is taking place at different locations would have a significant effect on our target variable.

Given our understanding of the data and our model, We suggest that increased social gatherings has a positive effect on the outcome variable No. of new cases. This is because as more people gather in different locations we expect more interactions between individuals. This increased interaction allows for more of the disease spread to take place hence increasing the number of cases.

The sign of the beta coefficient corresponding to this variable will be 'positive'. We anticipate that increased social gatherings will occur as a result of overall increase in vaccination rates. And so, when comparing social

gatherings with the Vaccination rate variable, the sign of the delta coefficient corresponding to Vaccination rate will be 'positive' as well.

Therefore, the Omitted variable bias will be a product of two 'positive' numbers resulting in a positive quantity. The bias is pulling the beta coefficient away from zero in the positive direction. In other words the direction of bias is "***away from zero***."