Mobility and COVID-19

Daniel Lampert, Jack Wang, Joshua Chung, Akaash Kambath

1. An Introduction

The COVID-19 pandemic has altered life as we know it in a myriad of ways. Its virulence has changed the fabric of social life and interaction, making us rethink activities at even the most mundane level. In the United States, stay-at-home orders were put in place by each state during the end of March 2020 and early April 2020, mandating that individuals stay at home as much as possible. Such mobility restrictions were enacted by government and medical officials as an effort to combat the spread. These stay-at-home mandates decreased visits to areas such as parks, transit stations, grocery stores, and workplaces.

However, such stay-at-home mandates came at a mental, emotional, and economic toll. No one enjoys being forced to stay at home, and many small businesses suffered as a result of the mobility restrictions. With the country beginning to open back up, cases have been increasing in recent months, despite medical officials imploring that public areas be avoided as much as possible. As a result of the uptick in cases, it has been rumored that stay-at-home mandates may be placed again, much to the distress of the general public. This has led our group to ask the following questions: **how much are changes in mobility in public areas associated with the case rate of Covid-19 in the past 7 days?** Knowing the answers to such a question could provide a better understanding of how effective mobility restrictions placed on certain public places can be in preventing the spread of Covid-19.

To operationalize our problem statement, we will be examining changes in mobility via the `Residential Areas`, `Workplaces`, `Parks`, `Retail & Recreation`, `Grocery & Pharmacy`, and `Transit Stations` variables in our dataset. Our dependent variable will be the `Case Rate per 100,000 in the past 7 days`. The reason we chose this as our dependent variable as opposed to one such as the `Case Rate per 100,000`[1] is because the mobility variables are updated every 2-3 days, so we felt that this metric would be more reasonable instead of the cumulative `Case Rate per 100,000`.

2. A Model Building Process

We would like to measure the change in COVID-19 case rates per 100,000 people in the past 7 days by examining changes in mobility at public spaces such as parks, grocery stores, and workplaces. This is a descriptive modeling goal since we are interested in determining which locations are most strongly correlated with recent COVID-19 rates in different states in the United States. For each state, Google measures the change in number of visitors with their opt-in location data by comparing it to the state's baseline value. The baseline value for a given day of the week is computed as the median number of visitors for that day of the week in the 5-week period from January 3 – February 6, 2020. This is then compared with the opt-in location data collected on days between September 11 through October 23. For example, the number of visitors for each Monday in the date range September 11 to October 23 is compared to the median number of visitors for Mondays in the 5-week period from January 3 to February 6. Google then takes the change in mobility for each day in the 5-week period of Sept. 11 to Oct. 23 and gives the average change for that period. This is the value we are using for each state as a datapoint.

Since the Google mobility data was gathered from individuals who allowed Google to track their live location, it is likely that these individuals tended to be more tech savvy than the average American. This may suggest that these individuals are younger than the general population. As a result, this may have an impact on the

[1] We will refer to this as Case Rate

mobility characteristics of these individuals and the external validity of our findings should be interpreted with caution.

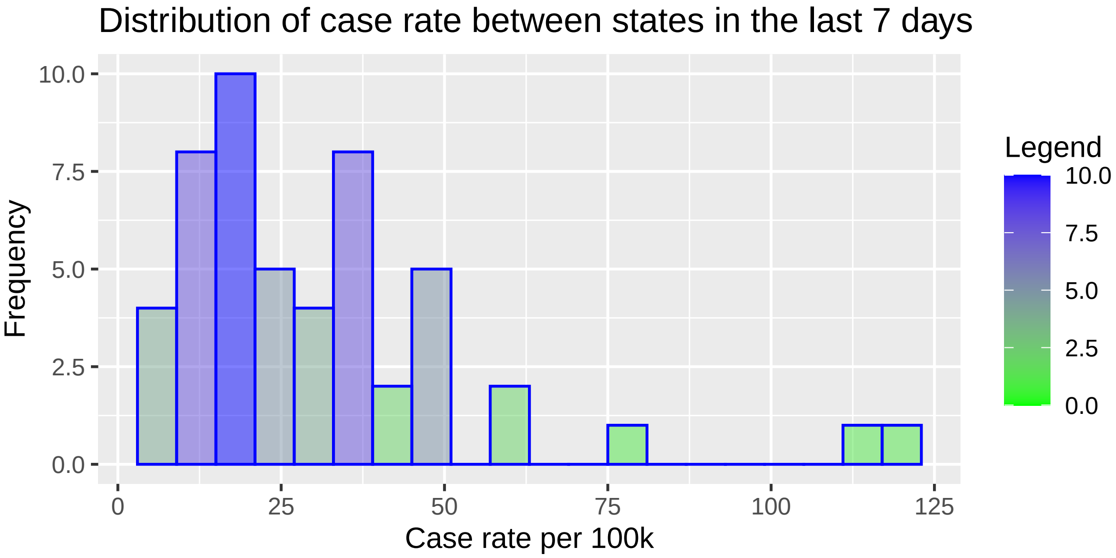
For the key variable we are looking at change in mobility in transit stations. We plan on using the change in mobility of transit stations as our key variable as we believe it will encompass the general movement of the population in each state. We want to investigate whether transit mobility is a good indicator of a state's `case rate per 100,000 in the last 7 days`, which may serve as a proxy for how well a state is responding to the pandemic.

There are no missing data points. The mobility variables all appear to be normally distributed, with slight skews for certain covariates. Specifically, we see a left skew in the distribution of park mobility, right skew in workplace mobility, and right skew in transit station mobility.

We acknowledge that this change in mobility at certain locations may be legally enforced or voluntary, which may be indicative of other factors that have relationships with case rate in the last seven days. Given this information, the coefficients and the significance values should be interpreted with caution.

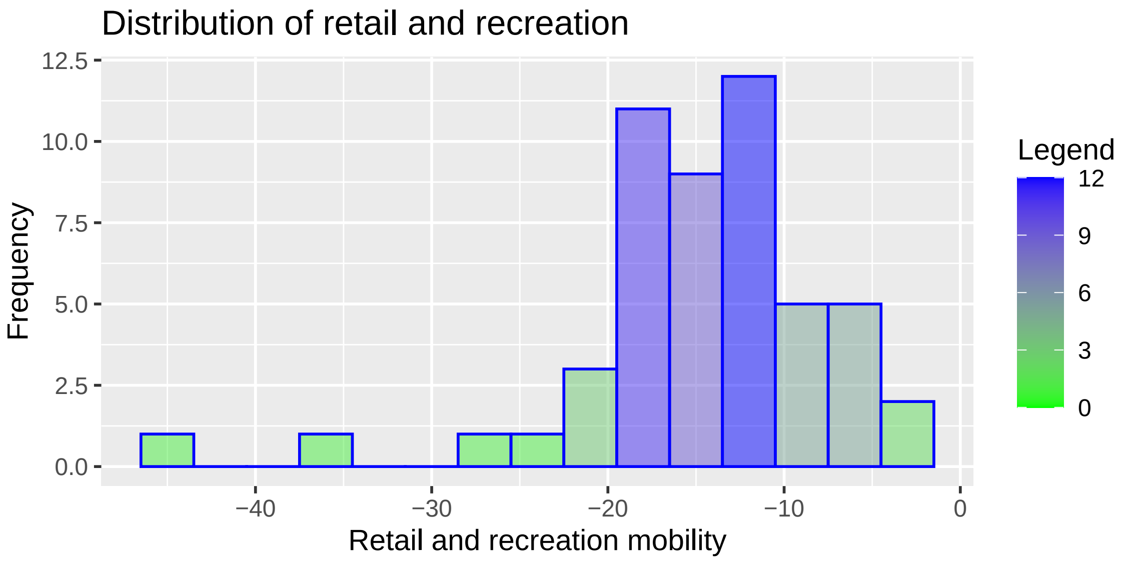
Exploratory Data Analysis

```
hist_case_rate <- ggplot(covid, aes(x = `Case Rate per 100000 in Last 7 Days`,
                                   fill = ..count..)) +
  geom_histogram(binwidth = 6, , alpha = 0.5, color = 'blue') +
  ggtitle("Distribution of case rate between states in the last 7 days") +
  scale_fill_gradient("Legend",low = "green", high = "blue") +
  xlab("Case rate per 100k") + ylab("Frequency")
hist_case_rate
```



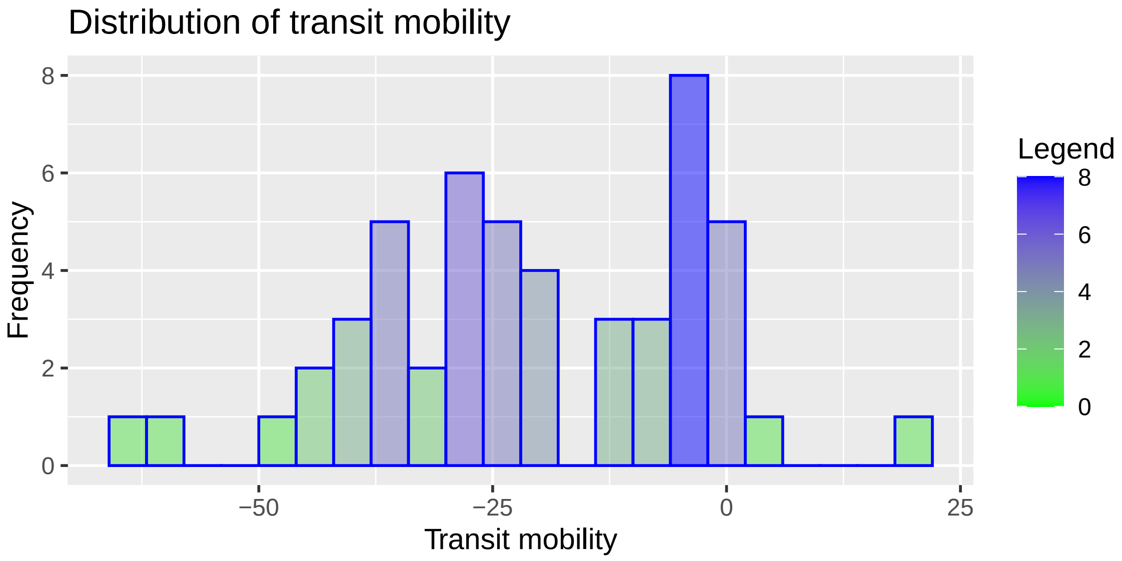
We can see a right skew in this distribution, showing that certain states have dramatically higher case rates than the majority.

```
hist_rec <- ggplot(covid, aes(x = `Retail & recreation`, fill = ..count..)) +
  geom_histogram(binwidth = 3, , alpha = 0.5, color = 'blue') +
  ggtitle("Distribution of retail and recreation") +
  scale_fill_gradient("Legend",low = "green", high = "blue") +
  xlab("Retail and recreation mobility") + ylab("Frequency")
hist_rec
```



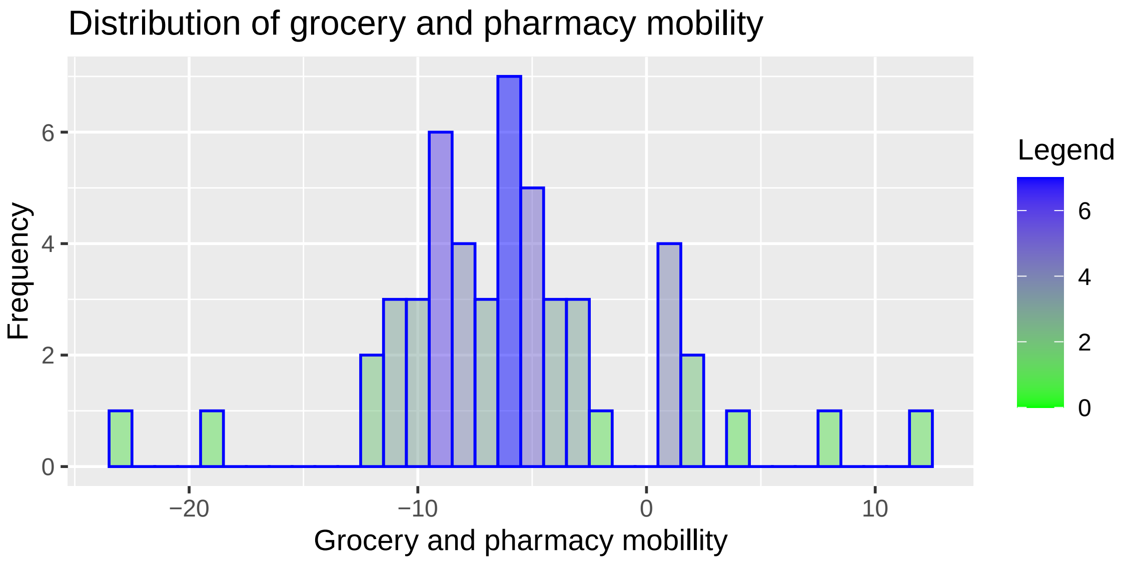
We see a left skew in this distribution, indicating that certain states closed retail and recreation stores at a much higher rate than most. Also no states had a positive change in retail and recreation mobility.

```
hist_transit <- ggplot(covid, aes(x = `Transit stations`, fill = ..count..)) +
  geom_histogram(binwidth = 4, , alpha = 0.5, color = 'blue') +
  ggtitle("Distribution of transit mobility") +
  scale_fill_gradient("Legend",low = "green", high = "blue") +
  xlab("Transit mobility") + ylab("Frequency")
hist_transit
```



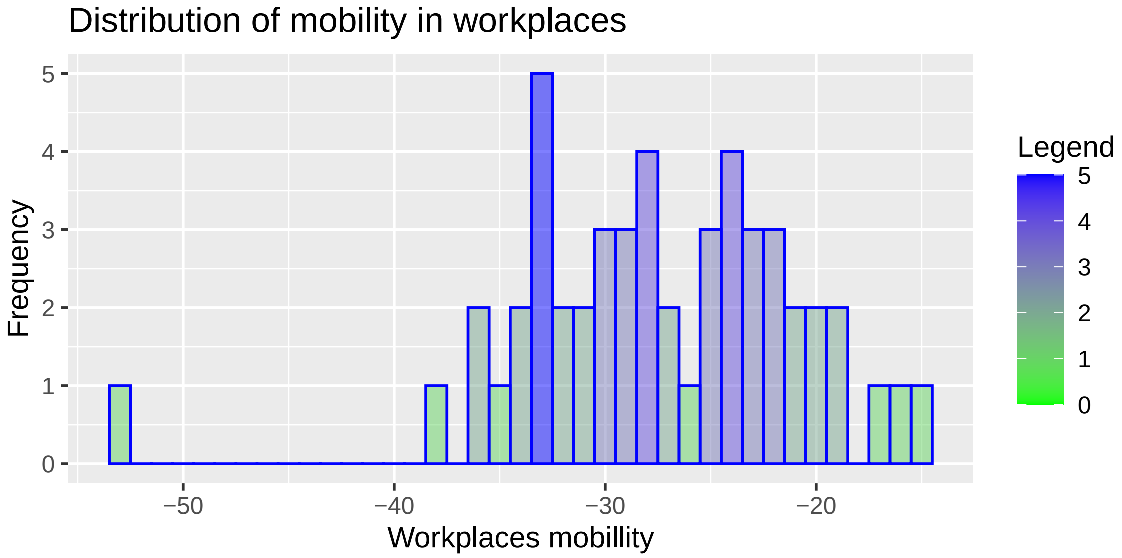
This distribution appears to be somewhat normal with no particular skew. However, there is a high density around zero change in transit mobility, which suggests that 15 states had little to no change.

```
hist_grocery_pharm <- ggplot(covid, aes(x = `Grocery & pharmacy`, fill = ..count..)) +
  geom_histogram(binwidth = 1, , alpha = 0.5, color = 'blue') +
  ggtitle("Distribution of grocery and pharmacy mobility") +
  scale_fill_gradient("Legend",low = "green", high = "blue") +
  xlab("Grocery and pharmacy mobillity") + ylab("Frequency")
hist_grocery_pharm
```



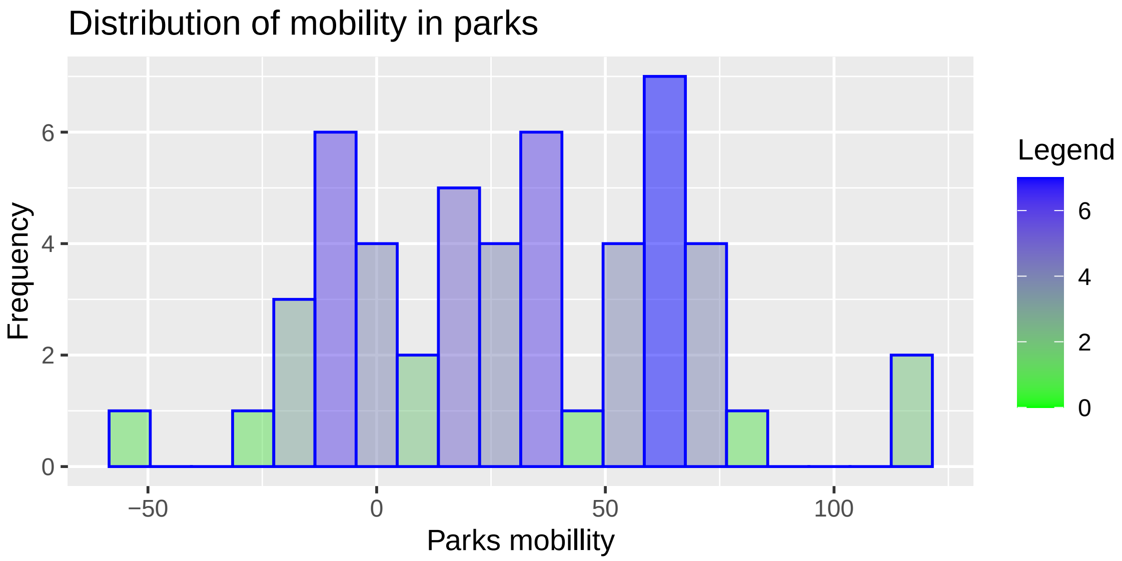
Grocery and pharmacy mobility appears to have a fairly normal distribution centered around a -7% change in mobility.

```
hist_workplaces <- ggplot(covid, aes(x = Workplaces, fill = ..count..)) +
  geom_histogram(binwidth = 1, , alpha = 0.5, color = 'blue') +
  ggtitle("Distribution of mobility in workplaces") +
  scale_fill_gradient("Legend",low = "green", high = "blue") +
  xlab("Workplaces mobillity") + ylab("Frequency")
hist_workplaces
```



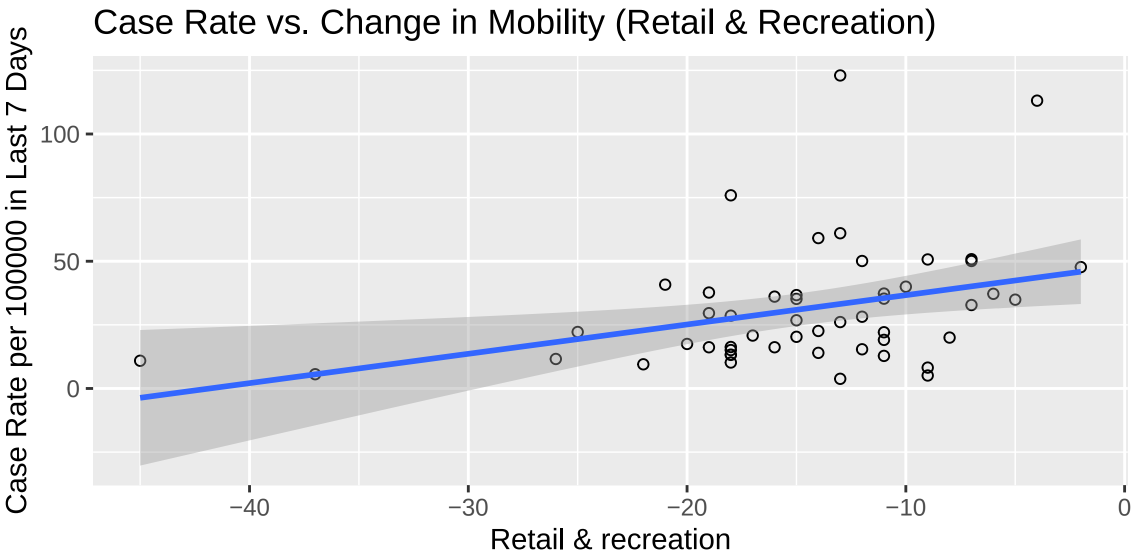
This distribution appears to be fairly normal except for a single outlier at -53% for workplace mobility. We also observe that there are no states with a positive change in workplace mobility.

```
hist_parks <- ggplot(covid, aes(x = Parks, fill = ..count..)) +
  geom_histogram(binwidth = 9, , alpha = 0.5, color = 'blue') +
  ggtitle("Distribution of mobility in parks") +
  scale_fill_gradient("Legend",low = "green", high = "blue") +
  xlab("Parks mobillity") + ylab("Frequency")
hist_parks
```



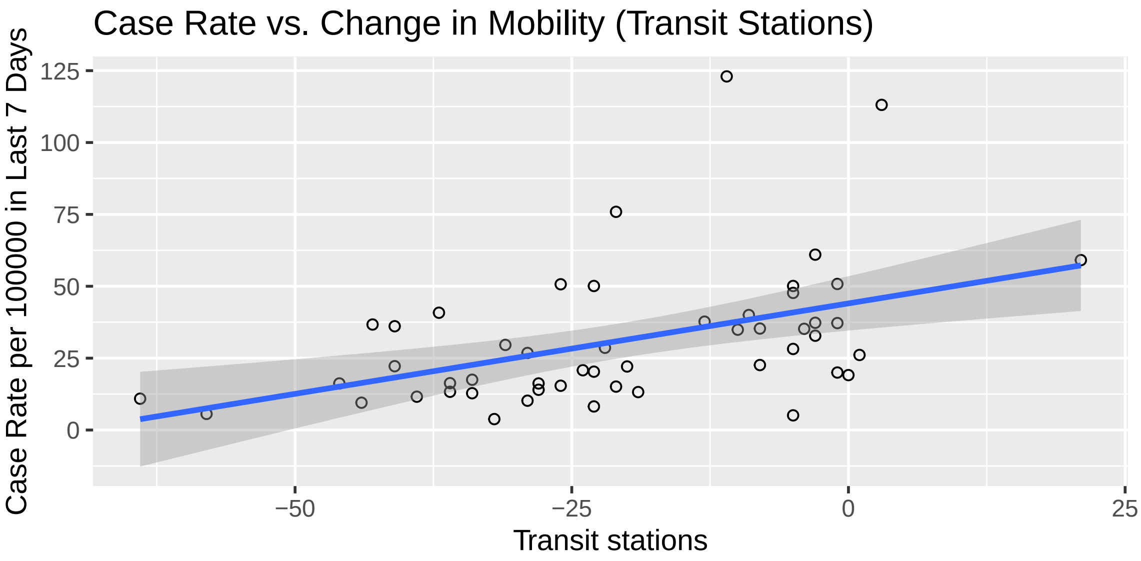
This distribution seems slightly non-normal as there are several peaks. Generally, park mobility seems to have increased in the majority of states.

```
ggplot(covid, aes(x=`Retail & recreation`, y=`Case Rate per 100000 in Last 7 Days`)) +
  ggtitle("Case Rate vs. Change in Mobility (Retail & Recreation)")+
  geom_point(shape=1) +
  geom_smooth(method=lm)
```



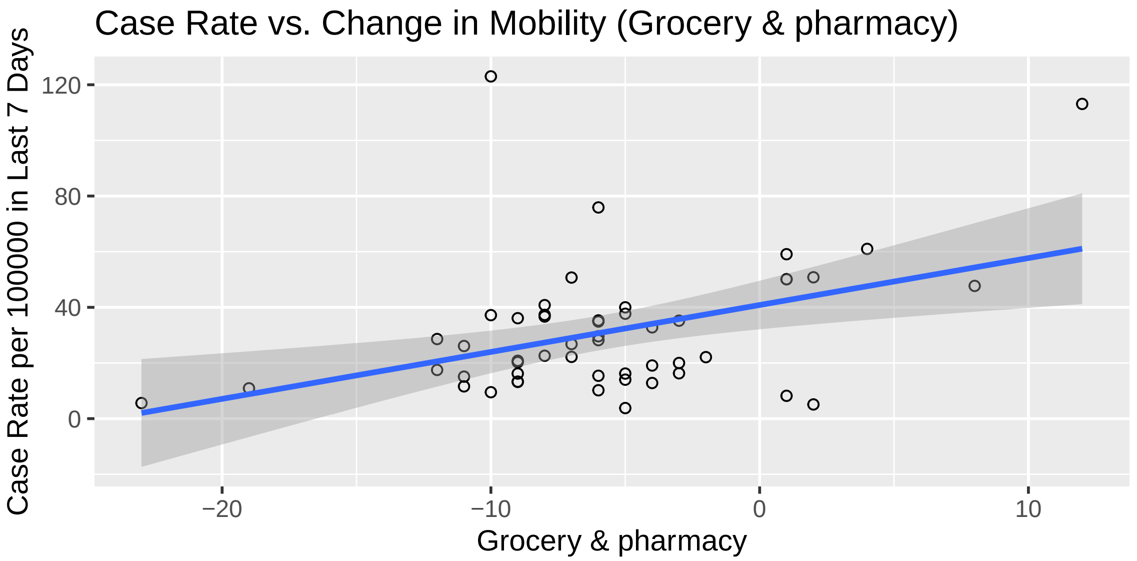
An upwards-sloping regression line indicates that increased mobility in retail and recreation is associated with an increase in case rate. It makes sense because places for retail and recreation are typically indoor, where the chance of being infected is high.

```
ggplot(covid, aes(x=`Transit stations`, y=`Case Rate per 100000 in Last 7 Days`)) +
  ggtitle("Case Rate vs. Change in Mobility (Transit Stations)")+
  geom_point(shape=1) +
  geom_smooth(method=lm)
```



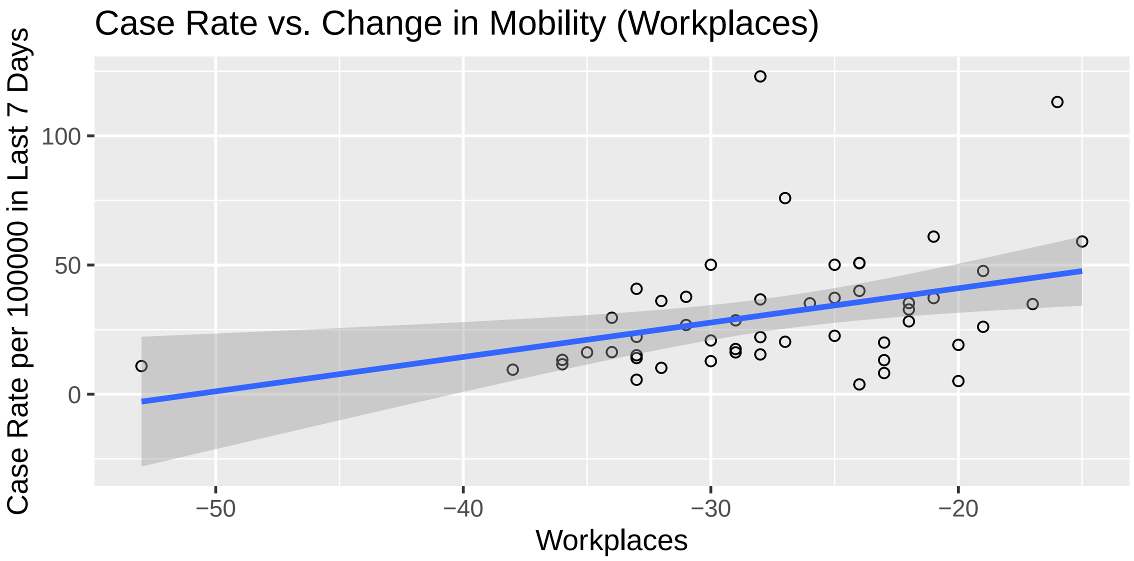
A positive regression line indicates that an increase in transit station mobility is associated with an increase in case rate. The slope is even higher than that of recreation and retail mobility, which makes sense because people moving around and not staying at home is generally associated with COVID-19 spread.

```
ggplot(covid, aes(x=`Grocery & pharmacy`, y=`Case Rate per 100000 in Last 7 Days`)) +
  ggtitle("Case Rate vs. Change in Mobility (Grocery & pharmacy)")+
  geom_point(shape=1) +
  geom_smooth(method=lm)
```



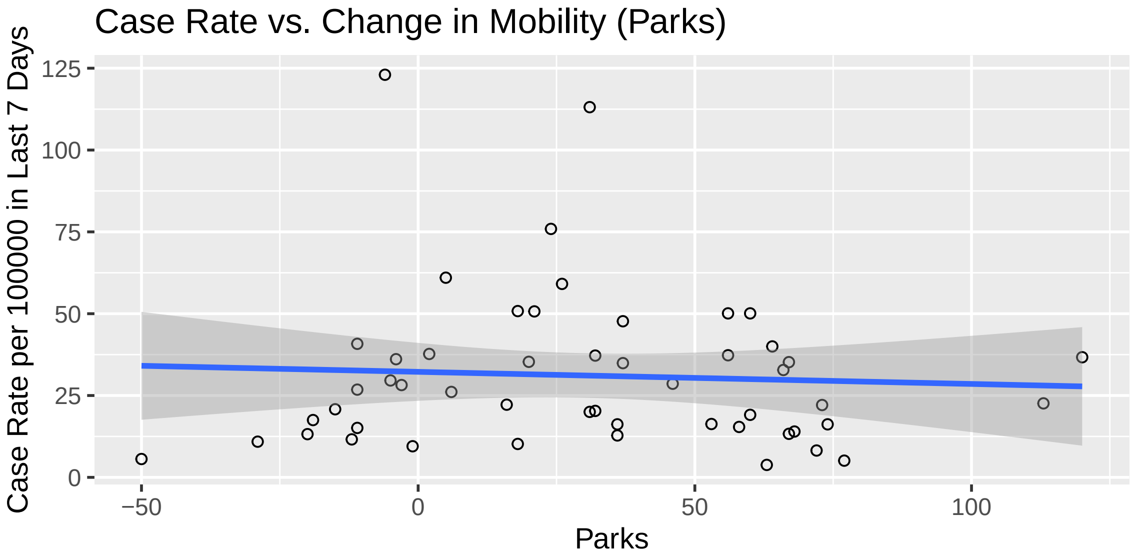
We see a slightly positive relationship with case rate and mobility in grocery and pharmacy stores. We may look at population density as a reason for this. Densely populated areas likely have a higher increase in grocery and pharmacy mobility since they tend to have more options for food that are not just grocery stores. Since many of these options shut down with the stay at home orders, grocery stores and pharmacies saw an uptick in mobility from people who previously frequented other stores.

```
ggplot(covid, aes(x=Workplaces, y=`Case Rate per 100000 in Last 7 Days`)) +
  ggtitle("Case Rate vs. Change in Mobility (Workplaces)")+
  geom_point(shape=1) +
  geom_smooth(method=lm)
```



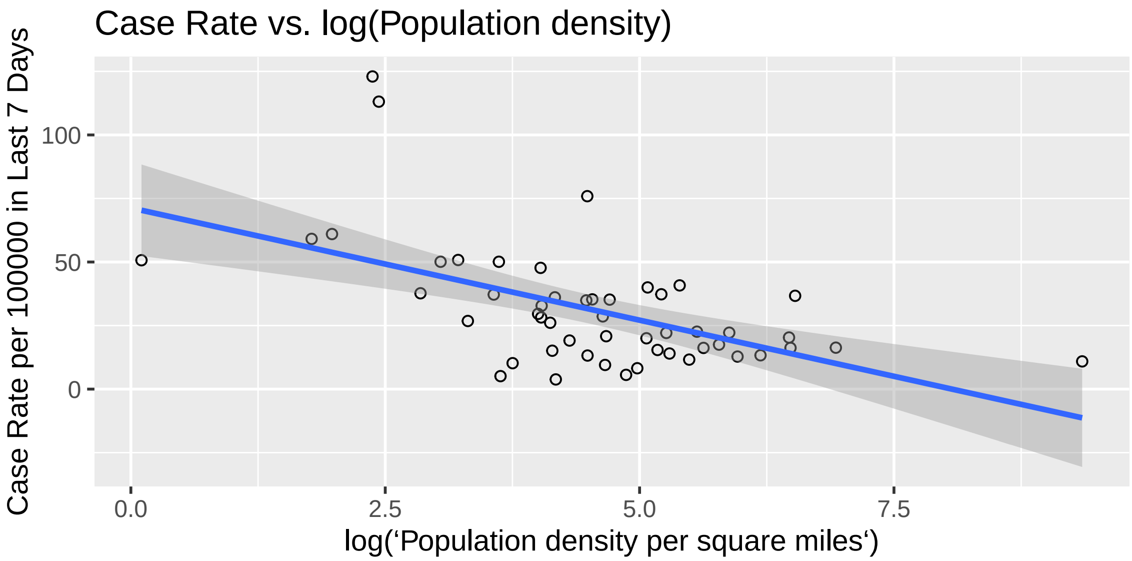
We see a positive relationship between case rate and workplace mobility. Workplace mobility has a positive relationship with case rates likely because people have to work together in close proximity.

```
ggplot(covid, aes(x=Parks, y=`Case Rate per 100000 in Last 7 Days`)) +
  ggtitle("Case Rate vs. Change in Mobility (Parks)")+
  geom_point(shape=1) +
  geom_smooth(method=lm)
```



There does not seem to be a relationship here. We assume this is due to park having large, open air spaces that are conducive to preventing the spread of the virus.

```
ggplot(covid, aes(x=log(`Population density per square miles`),
 y=`Case Rate per 100000 in Last 7 Days`)) +ggtitle("Case Rate vs. log(Population density)") +geom_poin
  geom_smooth(method=lm)
```



For model 3, we utilized a log transformation[2] on population density as opposed to the non-transformed population density as this visualization shows a linear relationship between the log of population density and case rate in the last 7 days. There seems to be a negative relationship between the log of population density and the case rate. We postulate that there may be a reverse casual relationship where states with higher population density may have populations that are more willing to obey mandates and follow recommendations that would protect others and themselves from COVID-19 as they feel that they have a higher risk of contracting the virus, leading to lower case rates.

```
covid_sub <- covid %>%
  select(Parks, Workplaces,`Grocery & pharmacy` , `Transit stations`,
         `Retail & recreation`)
covariance_matrix <- cov(covid_sub)
correlation_matrix <- cor(covid_sub)
```

```
library(xtable)
print(xtable(covariance_matrix, caption = 'Covariance Matrix of Mobility Variables'),
      type="latex",scalebox='0.9',comment=FALSE)
```

	Parks	Workplaces	Grocery & pharmacy	Transit stations	Retail & recreation
Parks	1348.20	77.13	90.81	168.71	139.19
Workplaces	77.13	45.01	23.23	98.57	40.62
Grocery & pharmacy	90.81	23.23	34.54	61.47	32.00
Transit stations	168.71	98.57	61.47	301.09	96.66
Retail & recreation	139.19	40.62	32.00	96.66	56.03

Table 1: Covariance Matrix of Mobility Variables

```
print(xtable(correlation_matrix,, caption = 'Correlation Matrix of Mobility Variables'),
      type="latex",scalebox='0.9',comment=FALSE)
```

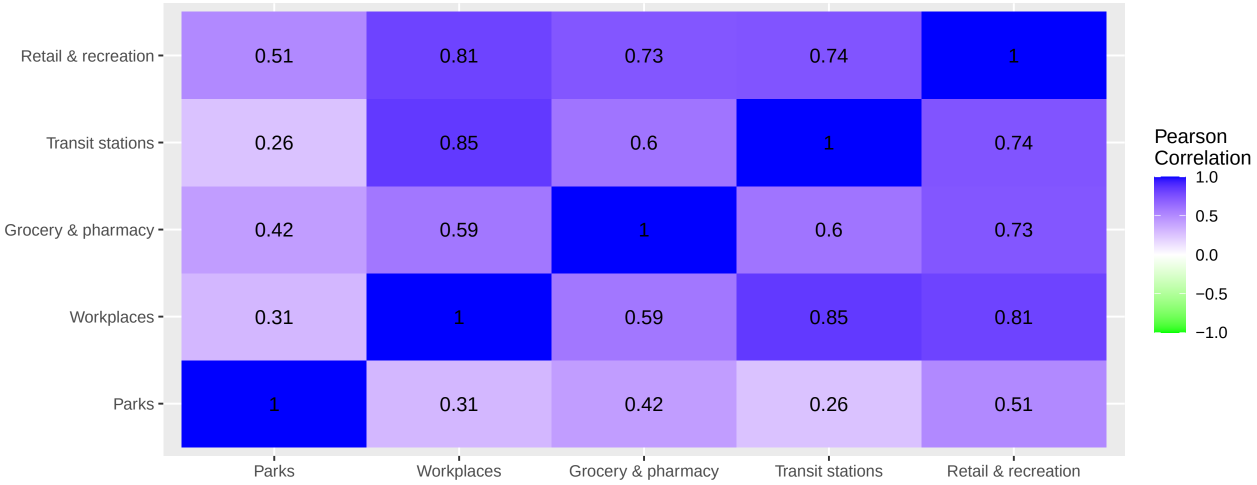
```
cor_cov <- round(cor(covid_sub),2)
covid_sub_melted <- melt(cor_cov)

covid.heatmap <- ggplot(data = covid_sub_melted, mapping = aes(x = Var1,
                                   y = Var2,fill = value)) +geom_tile() +  xlab(label = "Sample") +
 scale_fill_gradient2(low = "green", high = "blue", mid = "white",
```

[2] We utilized a natural log transformation

	Parks	Workplaces	Grocery & pharmacy	Transit stations	Retail & recreation
Parks	1.00	0.31	0.42	0.26	0.51
Workplaces	0.31	1.00	0.59	0.85	0.81
Grocery & pharmacy	0.42	0.59	1.00	0.60	0.73
Transit stations	0.26	0.85	0.60	1.00	0.74
Retail & recreation	0.51	0.81	0.73	0.74	1.00

Table 2: Correlation Matrix of Mobility Variables

```
  midpoint = 0, limit = c(-1,1), space = "Lab",
  name="Pearson\nCorrelation") + geom_text(aes(Var2, Var1, label = value),
 color = "black", size = 4) + xlab("") + ylab("")
covid.heatmap
```



Although there is no perfect correlation between the covariates, there is substantial correlation between many of them. The areas in dark blue are the areas with the highest collinearity. Since many forms of mobility are related to one another, it makes sense that there is high multi-collinearity. For example, the highest correlation is between transit stations and workplaces. This is not surprising since transit stations are often used to go to work. The same scenario occurs for several of the other covariates.

Here are the models:

```
    parks = as.matrix(covid['Parks'])
    transit = as.matrix(covid['Transit stations'])
    caseRate7 = as.matrix(covid['Case Rate per 100000 in Last 7 Days'])
    grocery = as.matrix(covid['Grocery & pharmacy'])
    workplaces = as.matrix(covid['Workplaces'])
    retail_and_recreation = as.matrix(covid['Retail & recreation'])

    model1 = lm(caseRate7 ~ transit )
    model2 = lm(caseRate7 ~ grocery+transit)
```

```
log_pop_density = log(covid$`Population density per square miles`)
model3 <-  lm(caseRate7 ~ parks + grocery + transit + retail_and_recreation
              + workplaces + log_pop_density)
car::vif(model3)
```

```
##                 parks               grocery               transit
##              1.934483              2.274814              3.884990
## retail_and_recreation            workplaces       log_pop_density
##              4.876566              4.951975              2.193816
```

As we can see, none of the covariates suffer from perfect multicollinearity since each has a VIF score less than 5.

Model Analysis

Our model 1 took on the form[3]:

$$\widehat{y} = 44.043 + 0.629 * transit$$

In model 1, we only include one key variable, the change in transit station mobility. This variable encompasses all kinds of transit, not just public transportation. This includes seaports, highway rest stations, and train stations, as well as many others. Therefore we can use this variable as a proxy to measure the general movement in a specific state. We believe that the positive coefficient for transit stations is inline with the expectation that having people travel around and also come into contact with each other would lead to higher rates of COVID-19. The coefficient's value of 0.629, which is statistically significant, suggests that a 1% increase in visits to transit stations is associated with a 0.629 increase in the case rate.

Our model 2 takes on the form:

$$\widehat{y} = 45.441 + 0.447 * transit + 0.891 * grocery$$

In model 2, we added another variable, change in grocery and pharmacy mobility, into our model. This variable does not have a significant association with increase in case rate, which makes sense since groceries and pharmacies have reduced the maximum capacity of customers and have mandated wearing masks in most areas, unlike other recreational activities. Also, people generally try to maintain a relatively large distance with each other in these areas so the chance of getting infected is low. We believe that the p-value for this coefficient may be high since grocery stores and pharmacies are a necessity for societies, and this change in mobility will not vary as greatly between states when compared to variations in case rate.

We interpret the positive coefficient of grocery and pharmacy mobility change as the following: a 1% increase in visits to grocery stores and pharmacies is associated with a 0.891 increase in the case rate.

Our model 3 took on the form:

$$\widehat{y} = 55.064 + 0.410 * transit + 1.004 * grocery - 0.079 * parks$$
$$-0.037 * retail\ and\ recreation - 0.716 * workplaces - 6.056 * log(population\ density)$$

In model 3, we included changes in park, retail and recreation, workplace mobility, and log of population density, where the coefficient for log population density is significant. The coefficient for the change in mobility for transit stations is not significant anymore. Surprisingly, the coefficient for the logarithm of population density is negative.As stated earlier, we hypothesize that states that have higher populations tend to be more willing to enact and obey COVID-19 health related mandates, which is associated with lower case rates. A 1% change in transit mobility in this model, holding all else constant will lead to a 0.41 change in case rates. Holding all else constant, a 1% change in grocery mobility will lead to 1.004 change in case rate. The interpretation for the other covariates follows the same logic.

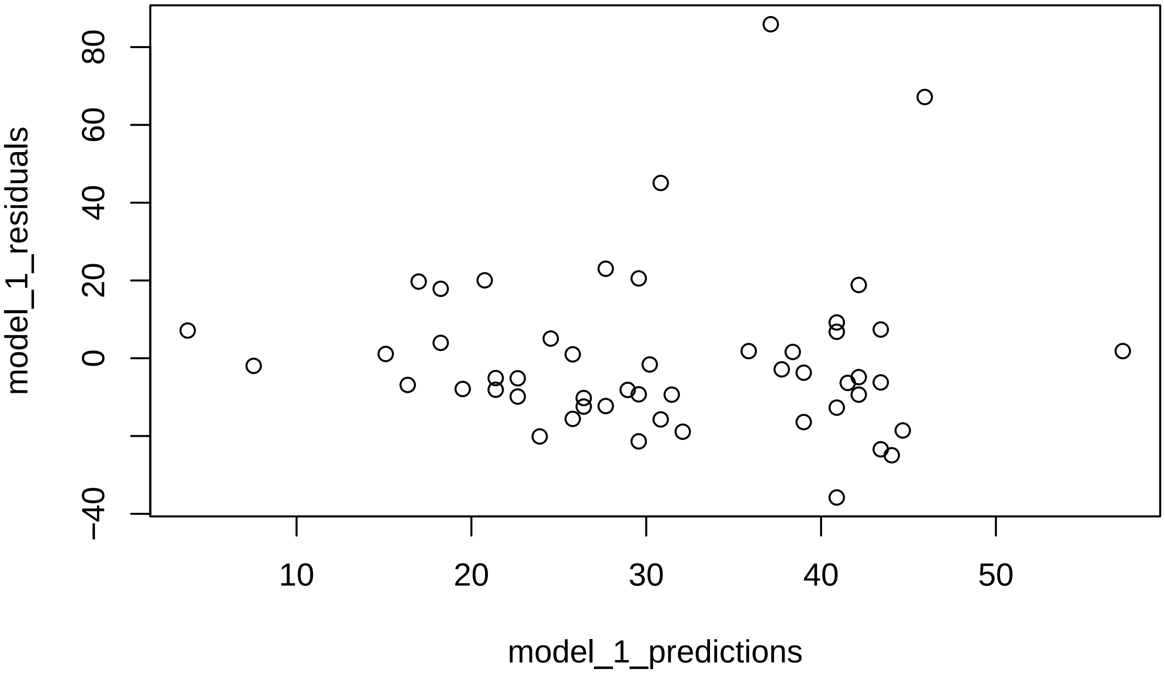
3. Limitations of the Models

1. IID Random Sampling: We acknowledge that our data is not identically distributed nor independent between each data point. States geographically near each other may share similar case rates and states may belong to different distributions. Moreover, states that have leadership from a similar political orientation likely share similar policies in regards to combating COVID-19. Our data is extremely limited as there are only 51 data points that exist.

[3] y:case rate per 100,000 in the past 7 days

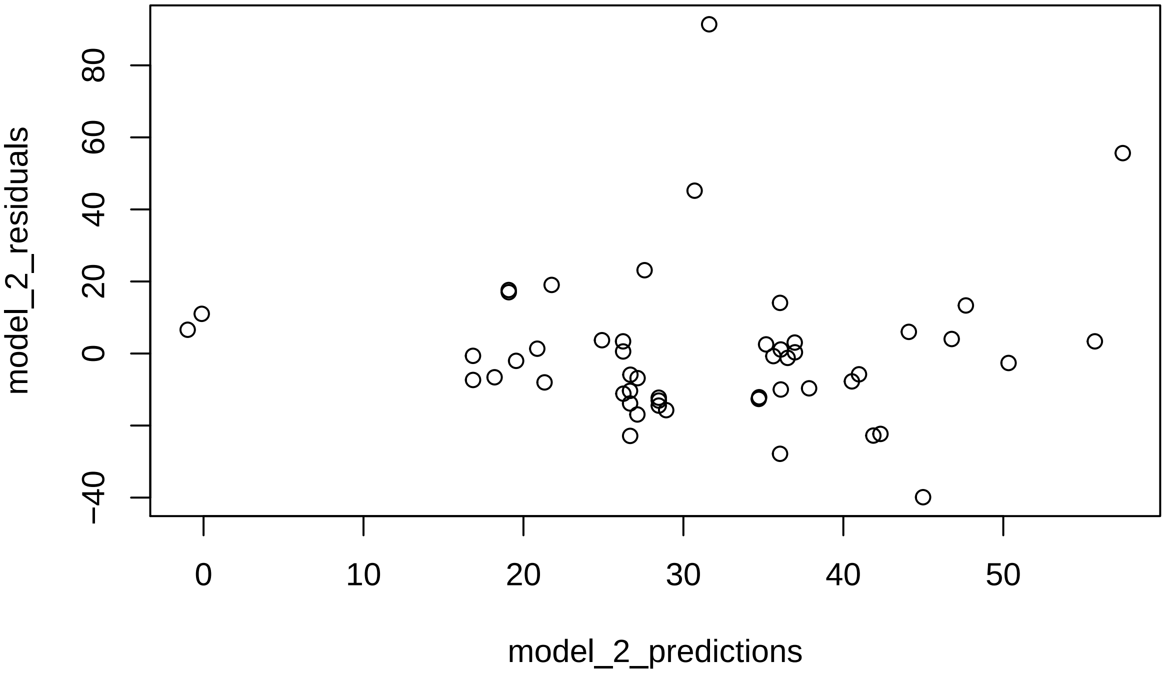
2. Linear Conditional Expectation Holds (will be addressed in R file by composing plots of residuals vs predictions for each model)

```
model_1_predictions = predict(model1)
model_1_residuals = resid(model1)
plot(model_1_predictions, model_1_residuals)
```



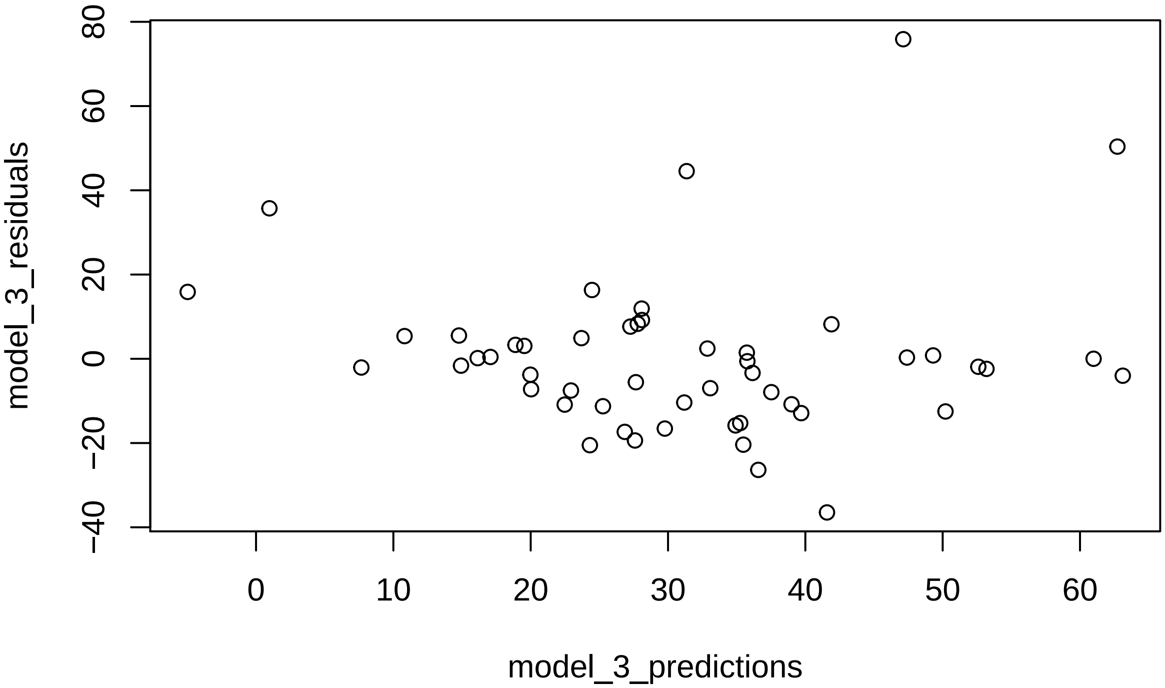
There appears to be no relationship between model 1 residuals and model 1 predictions.

```
model_2_predictions = predict(model2)
model_2_residuals = resid(model2)
plot(model_2_predictions, model_2_residuals)
```



There appears to be no relationship between model 2 residuals and model 2 predictions.

```
model_3_predictions = predict(model3)
model_3_residuals = resid(model3)
plot(model_3_predictions, model_3_residuals)
```



There appears to be no relationship between model 3 residuals and model 3 predictions.

Thus all 3 models satisfy Linear Conditional Expectation.

3. One CLM assumption of concern is the assumption of no perfect collinearity. Multicollinearity in the OLS regression models can undermine the statistical significance of the dependent variable, which in our case is `Case Rate per 100,000 in the Last 7 Days`. To discover any instances of perfect or near-perfect collinearity, a Variance Inflation Factor (VIF) test was conducted on the covariates Retail and Recreation, Parks, Workplaces, Grocery & Pharmacy, Transit Stations, and log Population Density. A score of 5 and above for any covariate indicates a strong multicollinearity for that variable. All of the covariates received a score less than 5, so this assumption was satisfied.

4. The assumption of homoskedastic errors can largely be avoided by using robust standard errors rather than classical standard errors. For this reason, we decided to use robust standard errors in the creation of our models to ensure the reliability of the coefficients and measures of significance.

As a sanity check, let's run a Breusch-Pagan test on each model. The null hypothesis of this test states that there is no evidence for heteroskedastic error variance. Note that failing to reject the null hypothesis does not mean the assumption is satisfied, though failing it does tell us that the assumption is not satisfied.

```
lmtest::bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 1.8758, df = 1, p-value = 0.1708
```

Here, we fail to reject the null hypothesis that there is no evidence for heteroskedastic error variance in model 1.

```r
lmtest::bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 1.0794, df = 2, p-value = 0.5829
```

Here, we fail to reject the null hypothesis that there is no evidence for heteroskedastic error variance in model 1.

```r
lmtest::bptest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 3.0641, df = 6, p-value = 0.8008
```

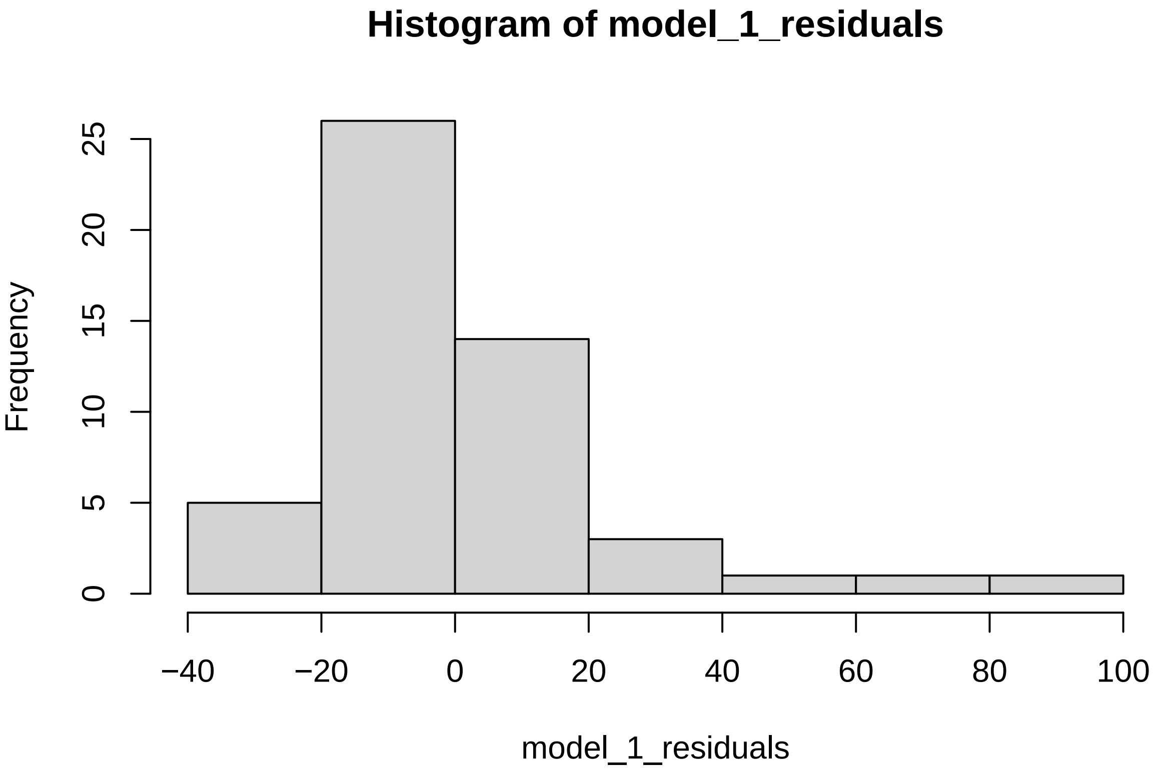
Here, we fail to reject the null hypothesis that there is no evidence for heteroskedastic error variance in model 1.

5. Normally Distributed Errors (will be addressed in R file, will make histograms of residuals for each model) To analyze this assumption, histograms of the distributions of the residuals for each model will be created and analyzed.

```r
mean(model_1_residuals)
```

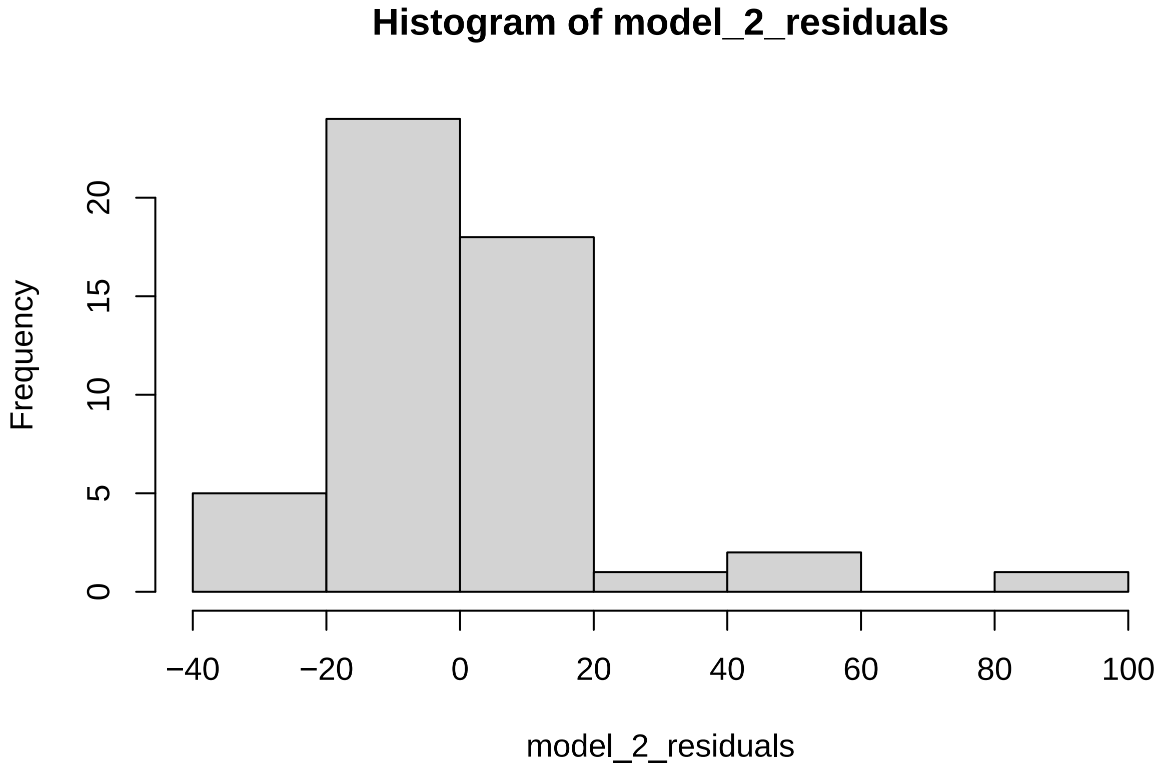
```
## [1] 0
```

```r
hist(model_1_residuals)
```



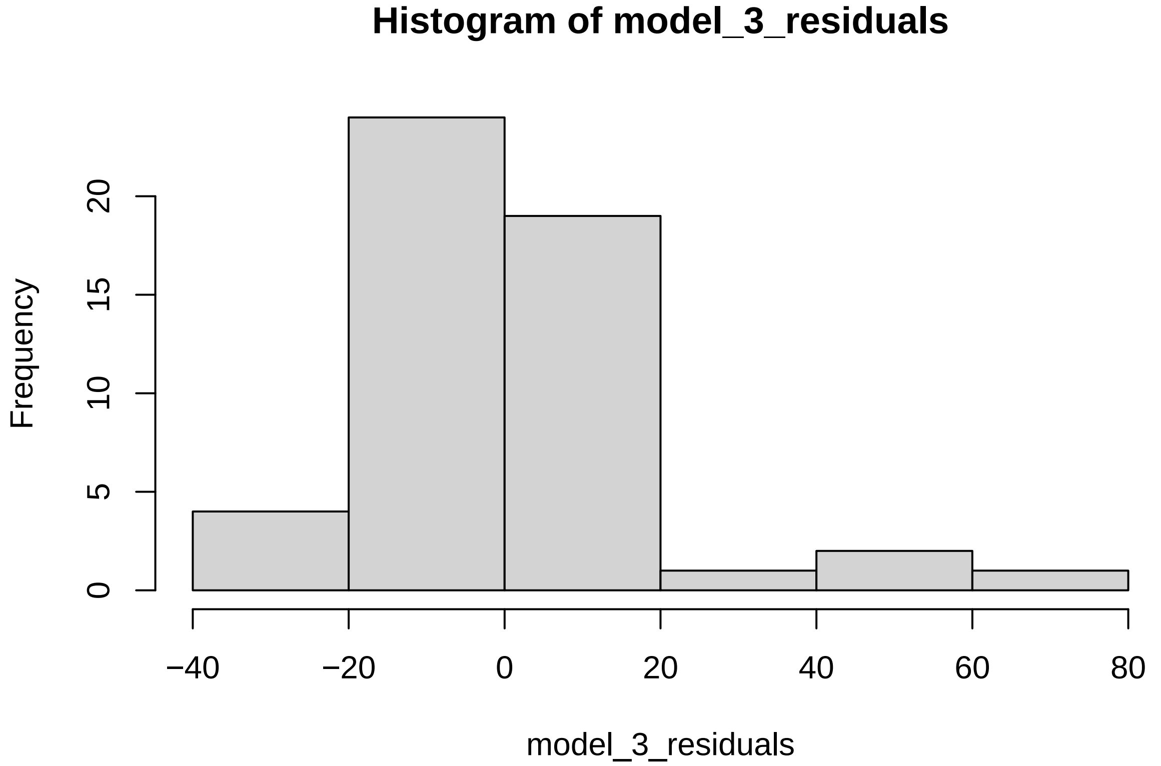
The distribution of the model 1 residuals is skewed to the right.

```
hist(model_2_residuals)
```



The distribution of the model 2 residuals is skewed to the right.

```
hist(model_3_residuals)
```



The distribution of the model 3 residuals is skewed to the right.

This assumption is violated, so hypothesis testing and confidence intervals derived from our analysis should be interpreted with caution.

4. A Regression Table

```
stargazer(model1, model2, model3, type="latex",
          title = "Regression Table for 3 Models",
          covariate.labels = c("Parks","Grocery","Transit stations","Retail and Recreation"
                               ,"Workplaces","log Population Density"),
          dep.var.labels = "Case Rate per 100000 in Last 7 Days",
          header=FALSE,ci=FALSE, no.space = TRUE, report = c("vc*"))
```

Table 3: Regression Table for 3 Models

	Dependent variable:		
	Case Rate per 100000 in Last 7 Days		
	(1)	(2)	(3)
Parks			−0.079
Grocery		0.891	1.004
Transit stations	0.629***	0.447**	0.410
Retail and Recreation			−0.037
Workplaces			−0.716
log Population Density			−6.056**
Constant	44.043***	45.441***	55.064***
Observations	51	51	51
R^2	0.208	0.238	0.373
Adjusted R^2	0.191	0.206	0.287
Residual Std. Error	21.547 (df = 49)	21.349 (df = 48)	20.230 (df = 44)
F Statistic	12.837*** (df = 1; 49)	7.496*** (df = 2; 48)	4.359*** (df = 6; 44)
Note:			*p<0.1; **p<0.05; ***p<0.01

*The analysis of our models was discussed above here

5. Discussion of Omitted Variables

1. Attitudes towards scientific evidence of COVID-19 - Proxy(Mask Usage) - Toward zero
2. Political orientation - Proxy (majority political party in state Conservative being 1) - Away from zero
3. How cautious people are about the virus - Toward zero
4. Age distribution of each state - Proxy (average age) - Toward zero
5. Health Consciousness - Toward zero

Even though we are evaluating a descriptive question, we believe that there are underlying causal relationships that lead to the omitted variable bias we see in our descriptive models. Therefore it is important to discuss omitted variables we believe may affect the variables that we are using in our models.

Omitting a variable that represents attitudes towards scientific evidence about COVID-19, or its proxy mask usage, likely biases the model towards zero. This occurs because individuals who care more about scientific evidence are less likely to conduct activities that put them at a greater risk of contracting COVID-19, such as going retail shopping. This biases the model towards zero because the higher the mask usage, the lower the case rate would be. Unlike attitudes towards science, the omission of political orientation with the proxy of political party likely biases away from zero. This is probably the case because conservative individuals are more likely to undertake activities such as retail shopping that fall into the category of movement. For this reason, this omitted variable biases the model away from zero since conservativeness has a positive impact on the models coefficients. If we consider how cautious people are about the virus as an omitted variable, we would argue that the bias is towards zero because they would likely take more precautions to prevent contracting the virus, lowering the case rate. Age distribution by state is also an omitted variable and we believe that states with younger populations would be less afraid to contract the virus, and thus have higher case rates. The inverse is also true for states with older populations. Therefore, we believe this bias is towards zero. Health consciousness similarly biases the model towards zero because individuals who are more health conscious are less likely to undertake activities that put them at greater risk of contracting COVID-19, including retail shopping and other movements. Since the effect on the coefficients is negative, this omission biases the model towards zero.

6. Conclusion

The COVID-19 pandemic has led numerous medical and government officials to clamor for a strong decrease in visits to public places. Given that mobility restrictions are not enjoyable for anyone, and with rumors that another stay-at-home order may be mandated again in the future, we hope that the three descriptive models we constructed will provide more insight on the level of association between public places and the COVID-19 case rate.

Though not significantly enough to invalidate our results, there was a fair amount of collinearity between each of the covariates as shown by our VIF scores and analysis of the CLM assumptions.

One key finding from the EDA portion of this analysis was that there was a negative association between the log population density and the case rate of COVID-19. This was a relatively surprising discovery, since our group had hypothesized that the more densely populated a state was, the more likely they were to have a higher case rate of COVID-19. However, this discovery appears to suggest that the more densely populated a state is, the more careful a state and its people are in trying to perform safer practices. An alternate interpretation may be that the more sparsely populated a state is, the more safe its individuals think they are, and as a result, the more lax they may be with regards to taking COVID-19 safety precautions.

Another key finding was that in models 1 and 2, the positive coefficient for the change in transit station mobility was statistically significant. We interpret this to mean that there is a positive association with overall change in mobility of each state with the case rate per 100,000 in the last 7 days, since we used change in mobility in transit stations as a proxy for overall mobility.

With regards to answering the research question at hand and the larger contexts surrounding it, we found that a 1% increase in a change in transit station mobility was associated with a 0.447% increase in the case rate of COVID-19. As a result, it may make sense for states to go the full mile when enforcing stay-at-home mandates and tightening restrictions on visiting transit stations.