# Lab 2 - Regression to Study the spread of COVID-19

w203: Statistics for Data Science - Team Flu-Fighters Adi, Anders, Cynthia, Zixi

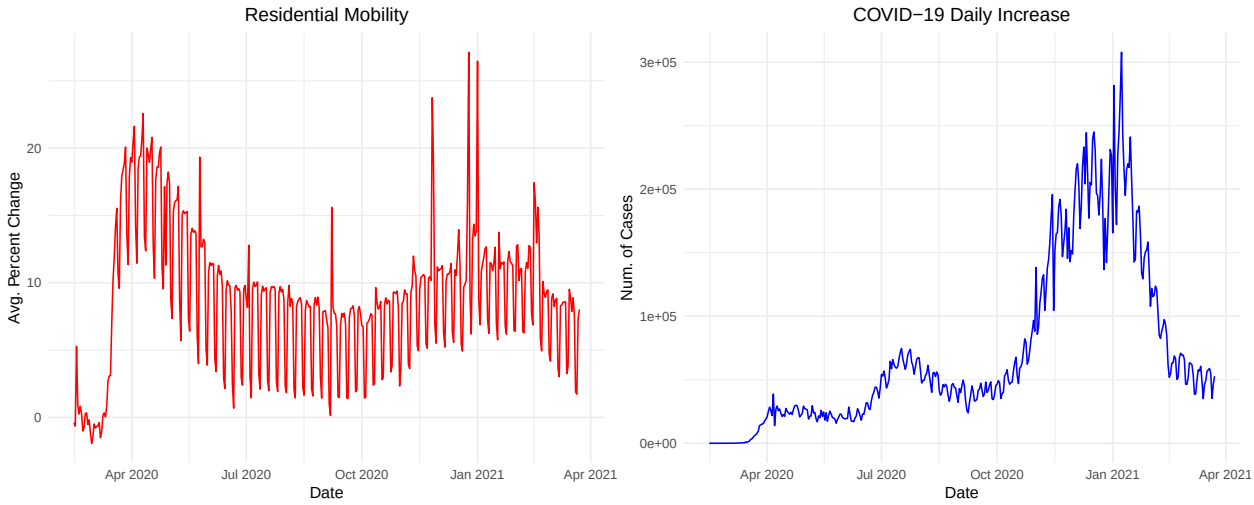
## Introduction:

In response to the COVID-19 pandemic, many countries have endeavored to control the spread of COVID-19 by limiting population mobility through enforcing face mask mandates and stay home orders, in order to reduce close contacts.

In the United States, residents changed their mobility pattern dramatically since COVID-19 pandemic started. As shown in the plot below, COVID-19 daily cases spiked during the summer and holiday season from November 2020 to January 2021. The spike in new cases was the result of multiple factors including quarantine fatigue, group activities, and social gatherings in the summer and holiday seasons. Accordingly, mobility change to residential also showed short spikes during the same periods. As the virus continues to spread through October, the population responds to COVID-19 with more precautions and their mobility trend tends to stabilize at the later period. However, during the early months, even though the number of daily increases is comparably lower than the later spikes, the mobility to residential data indicated that the average change to residential increased 10-20% compared to baseline last year. People dramatically reduced their mobility during the beginning of the pandemic even though the cases number were lower, but as the year progressed, mobility began to increase again. Therefore, we are interested in such trends and explored a few other variables to find their effects on the spread of COVID-19.

Our research question is – Is there a causal relationship between mobility and the spread of COVID-19 cases during February 2020 to June 2020 period? Furthermore, what other factors affect such relationships? We hypothesize that face mask mandates and Stay-At-Home orders, which are primary reasons for population's mobility change to residential, also have impacted the spread of COVID-19.

Our research can help build a stronger understanding of the relationship between political interventions and the trend of a pandemic. The outcomes could suggest whether implementation of political orders are effective, and also provide reference for political leaders to take effective measures preventing the spread of futher epidemics.



# Data Wrangling and Variables

## Data Sets:

In order to test our hypothesis, we obtained four data sets and selected a few interesting variables for further analysis.

1. United States COVID-19 Cases and Deaths by State A database of COVID cases that collected by Center of Disease Control
2. COVID-19 US State Policy Database A database of state policy responses to the pandemic, compiled by researchers at the Boston University School of Public Health.
3. COVID-19 Community Mobility Report A Google dataset that includes state-level measurements of individual mobility
4. The American Community Survey American Community Survey A data set collected by US Census Bureau in 2018. It contains state-level demographics and other indicators of general interest.

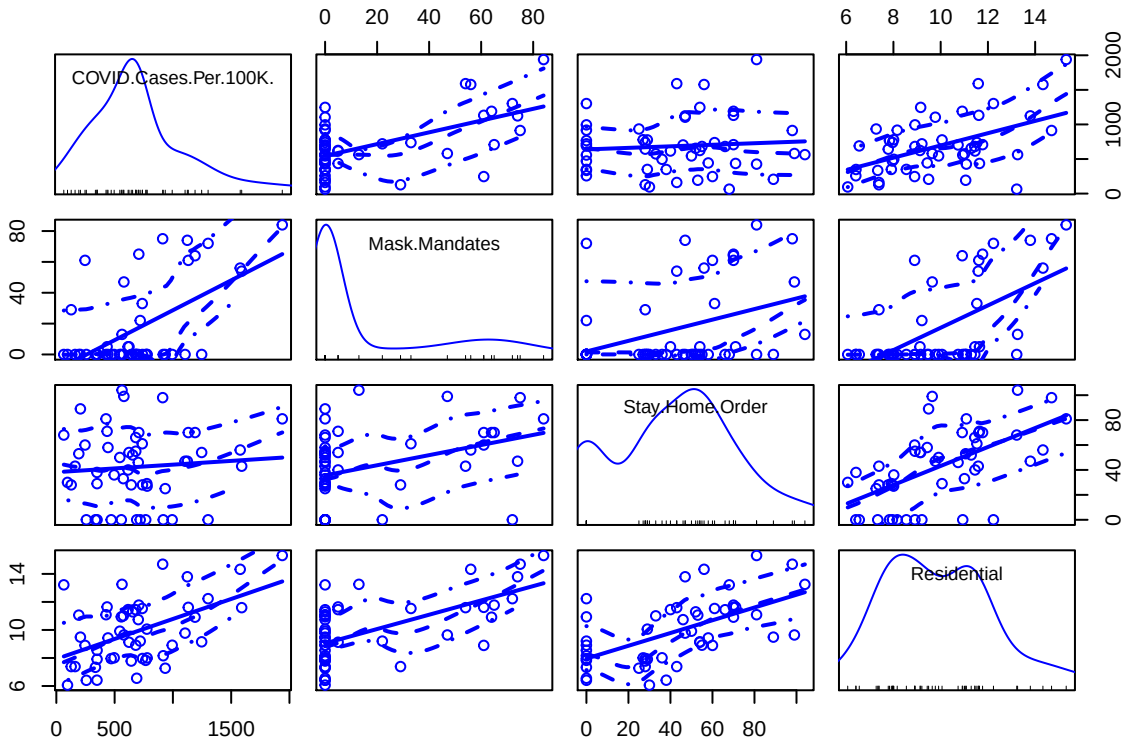
## Variables and transformation:

The variables that we interested in and transformation details are shown below. Each variable are explored at the state level within the timeframe selected (02/15/2020 - 06/30/2020). A number of demographic variables were explored and evaluated in Model 3 and 4 to further explore our hypothesis. The details of transformation would be included in the model section.

- State: Abbreviation used for all states in United States.
- COVID Cases Per 100K: Accumulated COVID-19 confirmed cases divided by 100k population by state.
- Residential Mobility: Average value of all subseted sample's mobility to residential data.
- Stay Home Order: Number of days of state-level stay home order being enforced.
- Mask Mandates: Number of days of state-level mask mandates being enforced.
- Ethnicity Populations: A percentage of a total state's population by ethnicity group
- Population Age: Percentage of the population in each age group at the state-level: 0-34, 34-65, 65+
- Family Households: Percentage of households that are families

# Exploratory Data Analysis:

First, we want to look at how each predictor variable and outcome variable correlate to each other, and their distributions. As shown below, we looked at a scatterplot matrix of the variables of interest, and it showed the correlation between each variable. A few variables are selected as we think they could help explain the spread of COVID-19, including COVID cases per 100K population, face mask mandates duration, stay home order duration, and residential mobility. Looking at the correlation of COVID cases vs each variable, all showed an upward trend. This is contradictory to our hypothesis, as we assume with a longer duration of mask mandates and stay home order, the COVID cases would decrease overtime. Also, with more mobile activity changed to residential and less contact, the COVID cases would decrease as well. This suggests that there are omitted variables and reverse causality interfering with our hypothesis and modeling accuracy, which will be discussed in depth in the model specific limitation section.




We also looked at the statistics and distribution of the four critical variables. The summary table of statistics informed us that our outcome variable COVID Cases per 100K population in each state has a minimum of 63.92, mean of 686.09, and maximum of 1938.35 within the time frame we were interested in. The distribution plot showed us a relatively normal distribution, which allows us to perform further analysis.

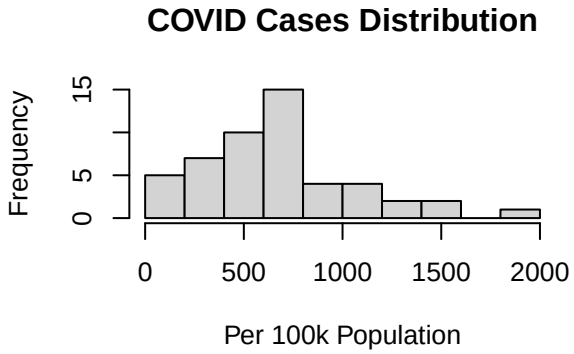
Residential Mobility - According to the summary table, it has a minimum of 6.066, mean of 9.889, and maximum of 15.314. The distribution plot showed a sparse representation of normal distribution with slight right skewness. As described in the Google Mobility Report, each raw data point represented the percentage of mobility change to specific activity compared to the baseline which was documented a similar time period in the previous year. It represents a percentage change of such activity of the sample compared to previous year, which is independent of other mobility variables. This variable helped us to see how much more time residents stayed at home at the beginning of COVID-19 pandemic.

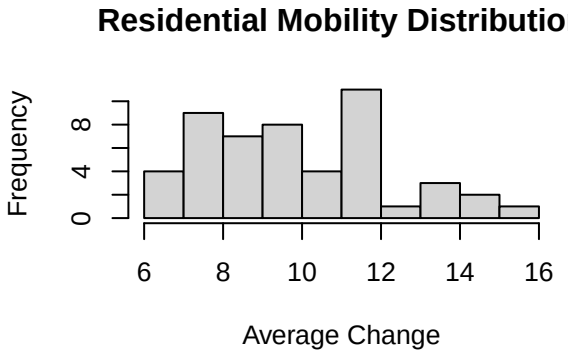
Mask mandates - Within the timeframe of interest, there were only 18 states having face mask mandate in place. At the early stage of COVID-19 pandemic, the mask was not recommended by health officials to the general population as the nation was facing a shortage of medical supply. This led to a later issued date of official state-level mask mandate. This also caused that many mask mandate issued dates were later than the time period we selected. This explained the high frequency of zero days in the distribution plot above.

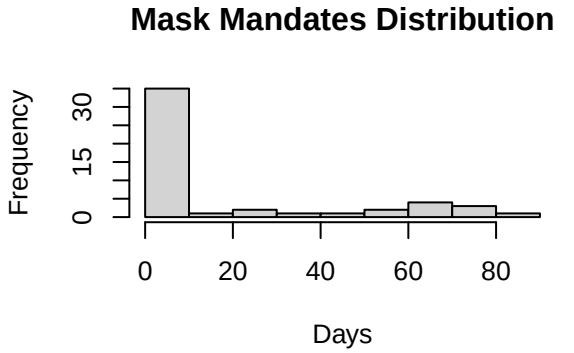
Stay Home order - There were 11 states that have zero days where they enforced a stay at home order. This led to the right skewed distribution. The raw data showed that multiple states had no start or end date of a strict state-level stay-at-home order. Even with such order, businesses tended to follow more or less strict guidelines, which made it difficult to evaluate the effectiveness of this state level order. The implementation of this order was hard to be evaluated in terms of effectiveness of social distancing the population. With this in mind, we further evaluated them in our models.

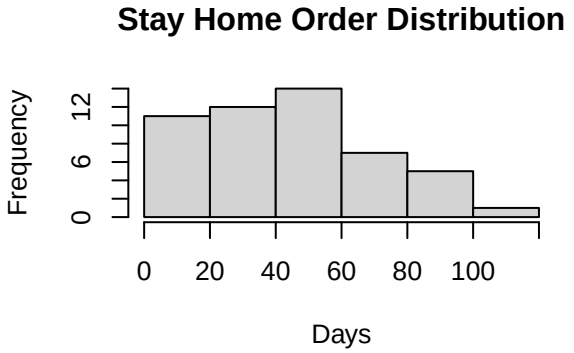
We will look at the how three predictor variables affected our outcome variable in Model 1 and 2. Many other demographic predictor variables will be evaluated and discussed in Model 3 and 4 to further test of our hypothesis.

```
##  COVID Cases Per 100K   Residential     Mask Mandates   Stay Home Order
##  Min.   :  63.92       Min.   : 6.066   Min.   : 0.00   Min.   :  0.00
##  1st Qu.: 428.59       1st Qu.: 7.974   1st Qu.: 0.00   1st Qu.: 27.25
##  Median : 648.50       Median : 9.566   Median : 0.00   Median : 44.50
##  Mean   : 686.09       Mean   : 9.889   Mean   :16.50   Mean   : 42.34
##  3rd Qu.: 878.88       3rd Qu.:11.451   3rd Qu.:27.25   3rd Qu.: 60.75
##  Max.   :1938.35       Max.   :15.314   Max.   :84.00   Max.   :104.00
```



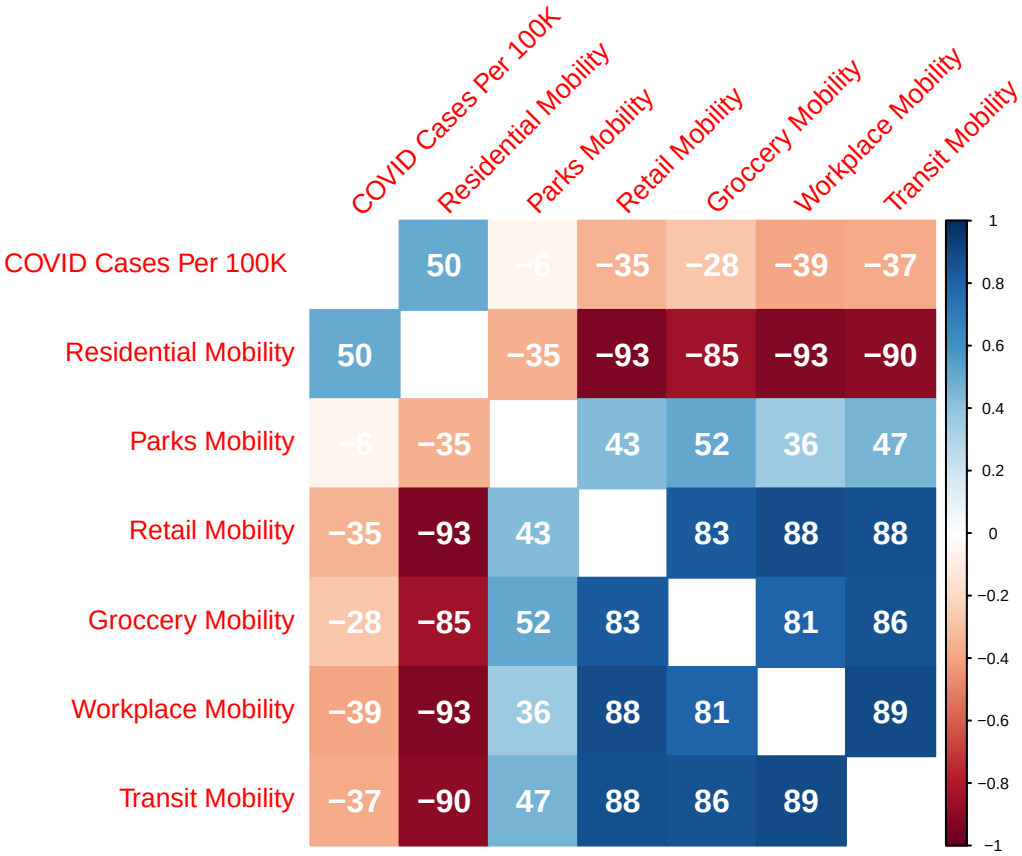




# Models

## Model 1

For our base model, we wanted to explore the relationship between cases per 100k population to mobility during the selected period (02/15/2020 to 06/30/2020). One of the key insights between the various mobility variables is that residential variables do inversely correlate with all other mobility variables. As the Google Mobility (Understand the data - Link) describes residential mobility as a residential variable shows a change in duration while the other variables measure a change in total visitors and as moist people already spend a big part of the day at places of residence, the capacity for change is not too large. Additionally, one can see residence as the default state and other locations only trigger once one has left the residence.

Other factors such as weather, availability of public transportation, availability of parks, identification of these locations appropriately would also determine the accuracy of visitor counts. Also, we could identify the reverse causality between 'residential mobility' and 'COVID Cases per 100k population' as during this period there was a significant relay of information around increases in cases counts and spread of COVID. One way to comprehend reverse causality is that significant people were concerned with the spread and would modify the travel behavior when COVID cases were on the rise and relax down once the spread started to slow down.




The next step is to identify whether 'Case per 100k population' does have normal distribution and plot the cases per 100k population per state as a variable of average residential mobility.

As you can see above the mobility and COVID cases have a positive correlation and do display patterns but also expose various outliers such as 'Hawaii', 'Louisiana' and 'Montana' which may have quite skewed plotting of mobility and COVID cases as they may represent the extreme variation.

```
##
## Regression Results - Model 1
## ================================================
##                            Dependent variable:
##                        ---------------------------
##                           COVID Cases Per 100K
##                            Model 1 - Mobility
## ------------------------------------------------
## Residential Mobility           88.62***
##                               p = 0.0002
##
## Constant                       -190.26
##                                p = 0.40
##
## ------------------------------------------------
## Observations                     50
## R2                              0.25
## Adjusted R2                     0.24
## Residual Std. Error        352.46 (df = 48)
## F Statistic              16.25*** (df = 1; 48)
## ================================================
```

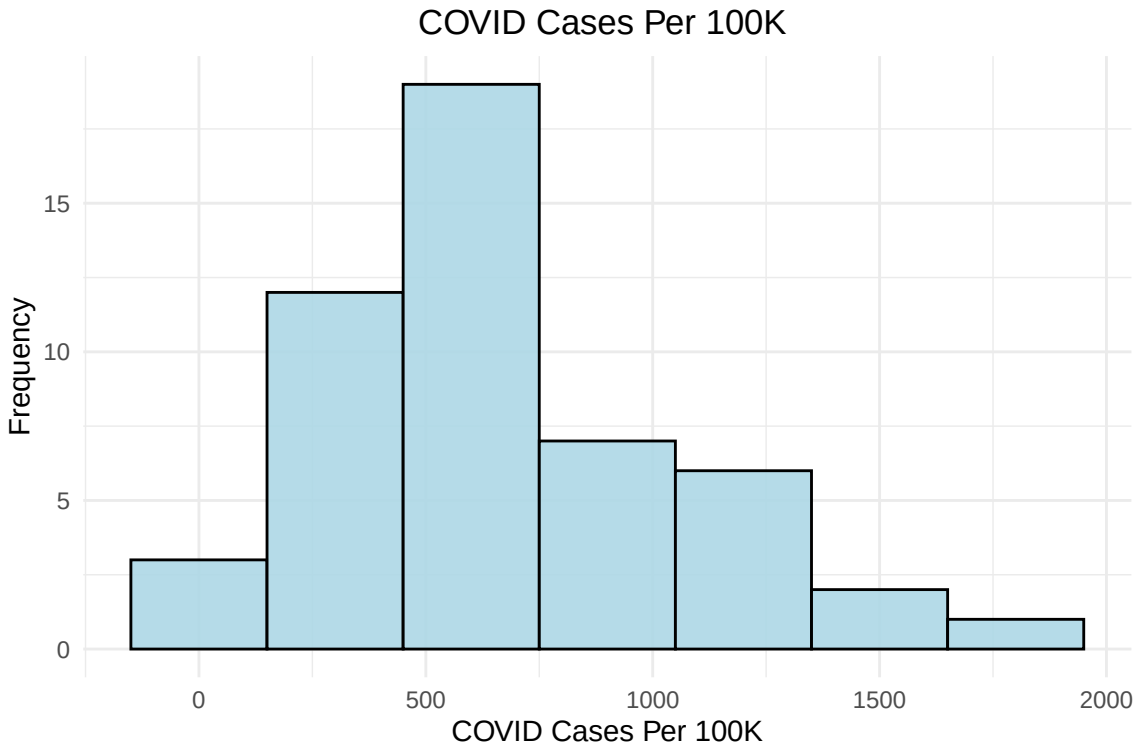



Figure 1: Model 1 - Mobility

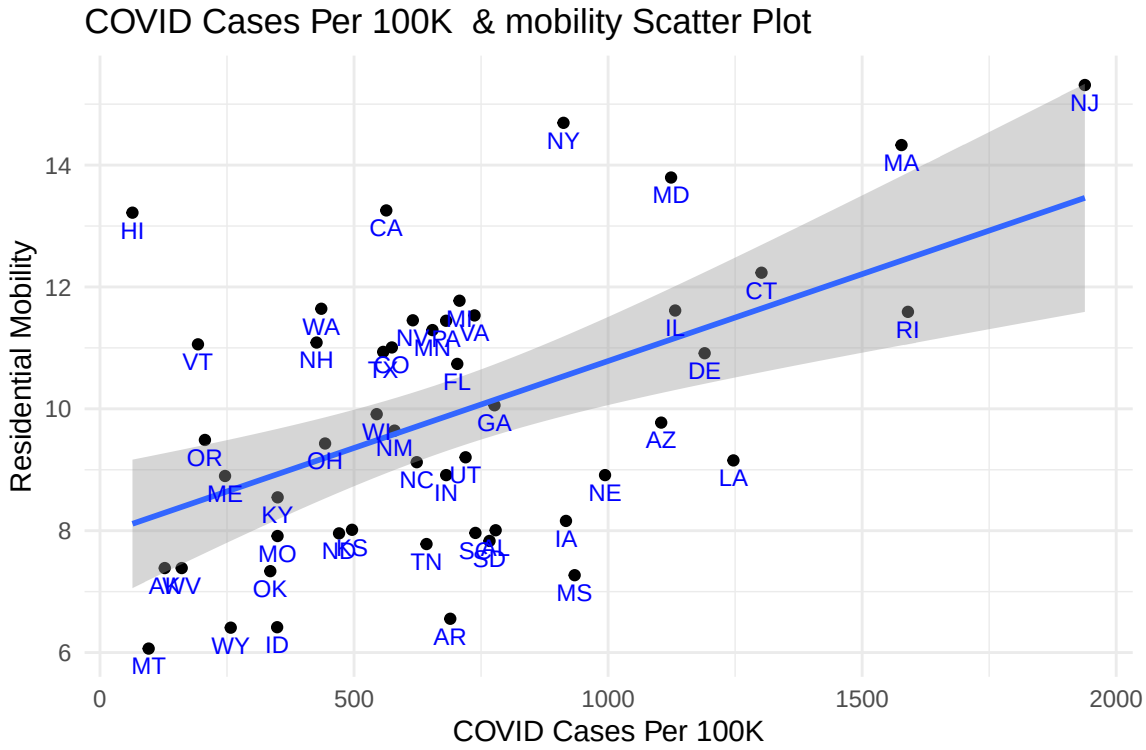


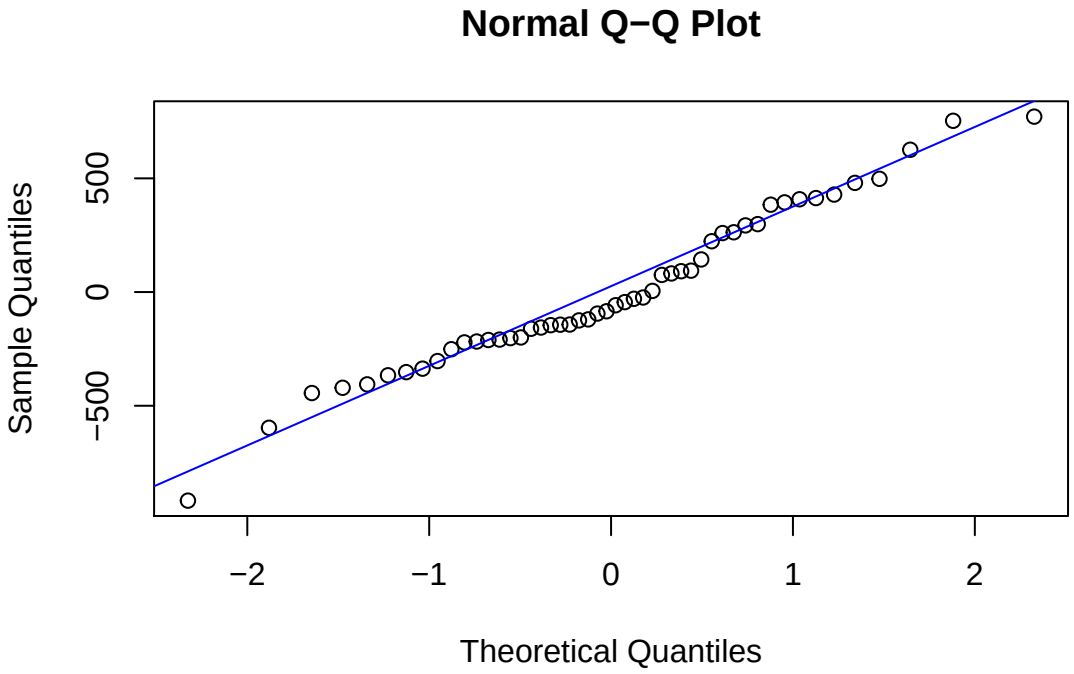

Figure 2: Model 1 - Mobility

```
## Note:                    *p<0.1; **p<0.05; ***p<0.01
```

Model one indicates that an increase in COVID cases is dependent on residential mobility. An increase in average residential mobility (people tend to stay home longer) predicted to increase in COVID cases to another 88.62 cases per 100k population with a standard deviation of 352.46. The R^2 for the model is 0.25 and a statistically significant result of less than 0.01.

Interestingly, the negative Y intercept is critical as the all the values are right top quadrant of the plot and as we are trying to summarize the variable 'COVID cases per 100k population' as an aggregate function during the period, we can cannot transform too much as the transforming aggregate functions may have multifold impact and can deviate the model significantly. Intuitively, we are expecting to have higher case counts during this period as most states were struggling with the COVID spread.

To evaluate model one further, we wanted to evaluate the residuals vs. predicted vs. actuals values to identify the areas of further explorations.




As you can notice that some actual values and predicted values do have similar trends but do not explain the behavior of certain states such as 'Hawaii' where predicted cases count is much higher than the actual case

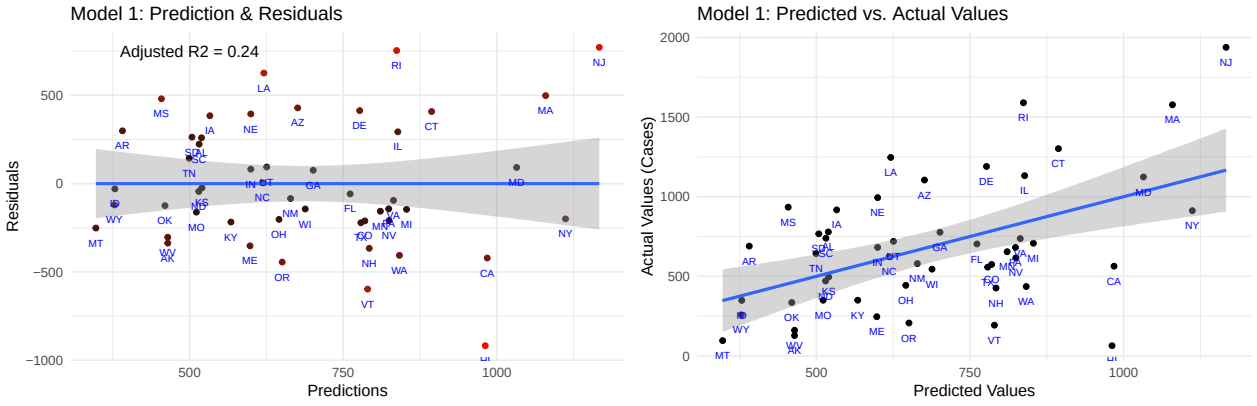



Figure 3: Evaluating Model 1

counts and 'Arkansas' where the actual values are much higher than the predicted values.

One of the interesting things about Prediction vs. residual is that as the predicted values increase the deviation of actual vs. predicted values increases. Hence, one can understand that states with higher actual case values may have other critical factors associated with the increase in COVID spread. But, given the cluster of values around 0 on 'prediction vs residual' plots suggests that the linear model can be enhanced further to include other variables that may explain the deviation but this is a decent model, to begin with. Also, based on the Normal QQ plot as well, it suggests that the model is normally distributed as most values are closely following the qq line as expected.

Couple of limitations of this base model is that grouping the mobility for an entire state over 5 months does take variations away from the data and bring the data towards the overall averages.Also, the residential mobility is difficult to reconstruct or summarize as average as the mobility will differ from each county to each day and would have seasonality of weekend, holidays and local trends.

## Model 2:

From the baseline model, we are observing that it is not able to explain most of variance due to a low R2 score. Indeed, many variables can be omitted if only a single variable is used to explain the case number per 100k. In our next step, we want to introduce more variables that are relevant to stopping the spread of COVID-19. Here we are particularly interested in how government policies are affecting the spread of COVID-19. Many states in succession started to place stay stay-at-home and, as a result, US resident mobility was significantly affected by stay-at-home order in the early of year 2020. To measure the effect of government policies, we add the variable to measure duration of stay-at-home order before June 30, 2020. This variable is also a proxy variable to measure how serious a state is treating COVID-19. If the duration is long, it means the state government is serious about COVID and decides to act early to combat its spread. Note that every state might place the order in a different way and we only consider the period when restrict order is placed. Take Texas for example. There was stay-at-home order placed but government did not restrict movement to public places. In this case, we assume the duration of stay-at-home order is zero day for Texas. This variable which affects the case number in the same direction as residential mobility (although due to reverse causality, the coefficient is positive), is omitted in our model 1 and the coefficient of residential mobility is biased away from zero because no government effort is captured to explain the slowdown of COVID case number. Another interesting government policy relevant to mobility is the mask mandate. Even though residents were encouraged to stay at home, essential activities were not forbidden, and people could still be active in public area. However, wearing masks has been proven to be an effective measure to reduce risk of infection when people must meet others in person. Therefore, including mask mandate will help to explain the cases when mobility itself is not enough to make a prediction on the spread of COVID. Here again, we take into account the effect of wearing masks by introducing the duration of mask mandate for each state in our model 2. In the end, we will include 3 variables in total to predict the case number per 100 thousand people in the population from 2020-2-15 to 2020-6-30 at state level: residential mobility, the duration of stay-at-home order, and the duration of mask mandate order.


To further understand the relationship between the variables, we plot a correlation matrix for them. Among the predictors, we do not see highly correlations, implying that we are less likely to have high multicollinearity issue in our model 2. The predictors all demonstrate some correlation with outcome variable except the duration of stay-at-home order. However, to evaluate the effect of each variable, correlation is not the best way to draw a conclusion. We will fully evaluate them by running OLS model and looking at their statistical significance, magnitude of coefficients and standard error.

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                 COVID Cases Per 100K
##                           Model 1              Model 2
##                             (1)                  (2)
## ---------------------------------------------------------------
## Residential Mobility      88.622***            77.204***
##                           p = 0.0002           p = 0.009
##
## Stay Home Order                                -4.602**
##                                                p = 0.017
##
## Mask Mandates                                  6.422***
##                                                p = 0.004
##
## Constant                   -190.261             11.521
##                           p = 0.398            p = 0.962
##
```

```
## ------------------------------------------------------------------
## Observations                    50                      50
## R2                             0.253                   0.441
## Adjusted R2                    0.237                   0.405
## Residual Std. Error    352.465 (df = 48)      311.305 (df = 46)
## F Statistic          16.246*** (df = 1; 48) 12.119*** (df = 3; 46)
## ==================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

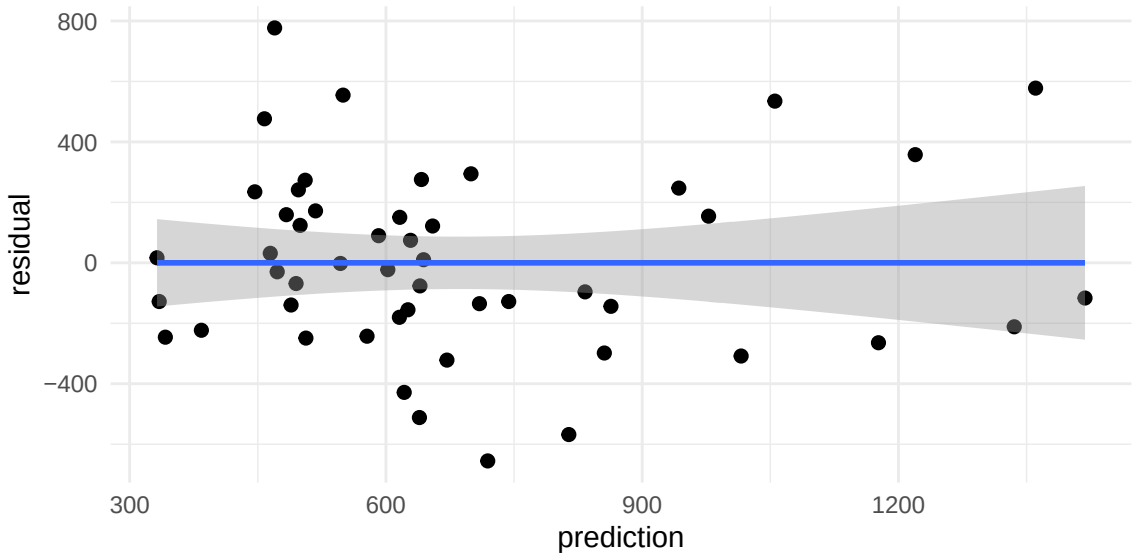
From the table above, we can see that all key variables show statistical significance but the coefficient of residential mobility has dropped to 77.204 from 88.622, indicating the additional variables have changed the bias in the coefficient for residential mobility. The coefficient for duration of stay-at-home order is -4.602. It confirms that with just one single variable itself, residential mobility has been biased away from zero by the negative sign of coefficient for duration of stay-at-home order. Practically, it indicates if a state could place stay-at-home order one day earlier, there would be about 4 less cases per 100k people in the population during the early of the 2020. Interestingly, the coefficient for duration of mask mandate is positive, which can be possibly the result of the reverse causality between duration of mask mandate and COVID cases. It is likely that the mask mandate is placed because the COVID cases already builds up its number and within the period being studied the effect of mask mandate is not well captured. As a result, the positive sign of duration of mask mandate is a limiting factor to the ability of explanation of model 2.

```
## Analysis of Variance Table
##
## Model 1: Case_Per_100k_Pop ~ mob_residential_avg + days_w_stay_home_order +
##     days_w_mask_mandates
## Model 2: Case_Per_100k_Pop ~ mob_residential_avg
##   Res.Df     RSS Df Sum of Sq     F   Pr(>F)
## 1     46 4457897
## 2     48 5963111 -2  -1505214 7.766 0.001242 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

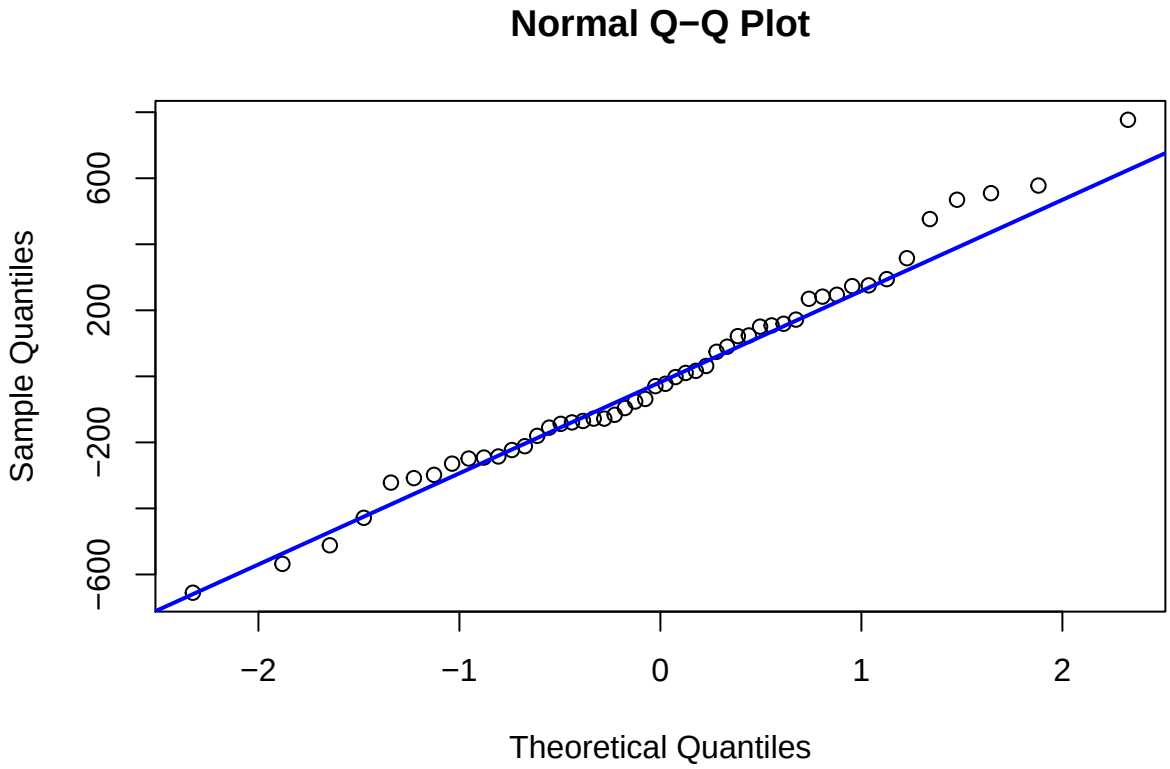
Yet we are still interested in comparing model 1 and model 2 and checking the assumptions. Running a F-test, we are able to see model 2 has indeed given more explanatory power than model 1 by reducing the sum of squared residuals. We obtain a statistical significance at 0.001 level, meaning the additional variables produce noticeable effect.

### Model 2: CLM Assumptions

Since we are using a relatively small data set of only 50 data points, it is also necessary to check the CLM assumptions to make sure the model is good shape.




First of all, we check the assumption of linear conditional expectation by plotting residual vs prediction. We see the residuals are not apparently away from zero, which means the assumption can hold for true and model indeed has captured the linear relationship between variables. Second, when running the model, we observe no variable dropped due to perfect collinearity. In the correlation matrix, we do not have highly correlated predictors either. Therefore, collinearity will not be a concern for us.




Third, we want to make sure the residuals are normally distributed by checking the normal Q-Q plot. From the plot above, the points do not form a perfect a line, meaning the residuals are not perfectly normally distributed. But it is still safe to consider the assumption of normally distributed errors can hold for true.

```
bptest(model_2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 1.3464, df = 3, p-value = 0.7182
```

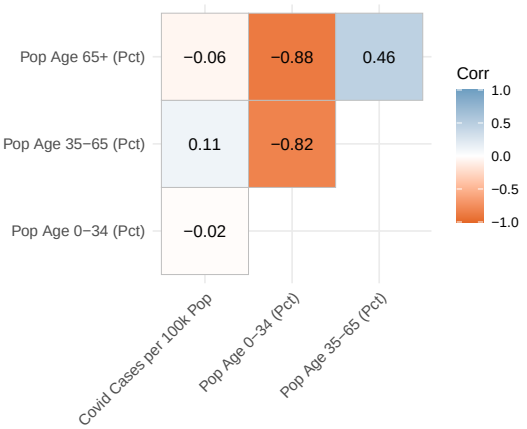
Fourth, we need to the assumption of homokedastic conditional variance by running the Breusch-Pagan test. The test gives a p-value of 0.7182, which is larger than 0.05. Therefore, we fail to reject the null hypothesis that there is no evidence for heteroskedastic error variance. Meeting this assumption has justified using standard errors in previous results.

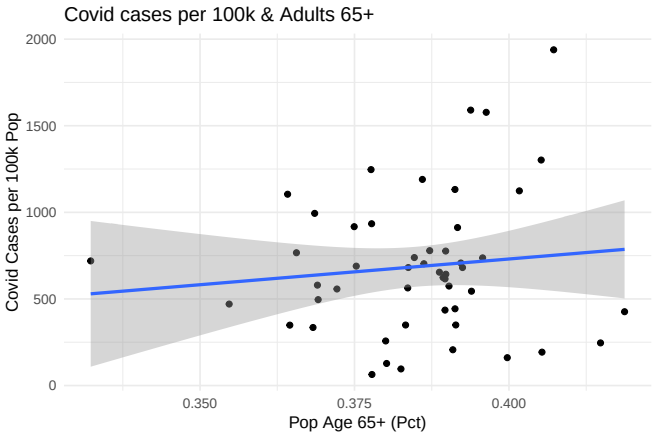
Lastly, IID assumption cannot be met. On the one hand, the data is aggregated at state level but people can travel between states, making each state not independent. On the other hand, 50 states are randomly drawn from a population of infinite number of states. In this sense, it is identically distributed. But still, IID assumption is violated.

## Model 3:

Model 2 focused on mobility and policy data to explain the increase in COVID, but there are still omitted factors that lead to variations in COVID cases. One of the most significant is how people within each state react to COVID. Demographics including age, occupation, income, ethnicity, culture, household, and health all play a role in how people respond to the virus. The goal of model 3 was to explore these demographic categories at the state level to determine if there were large enough variations between states to provide useful metrics to model on.

We began by evaluating age by grouping each state's population into three categories: 0-34, 35-65, and 65+. These categories represent the young, middle age, and elderly population, and each category have distinct differences in how they were affected by COVID. While people 65+ are at a higher risk, it may lead to a more cautious lifestyle with lower mobility and transmission compared to a younger population. Alternatively, a younger population may lead to fewer cases due to less severe symptoms that lead to less transmission. Our hypothesis was that states with a higher percentage of adults 65+ would lead to a higher number of COVID cases due to the spread of COVID in old folks' homes and other senior care facilities at the beginning of the pandemic.

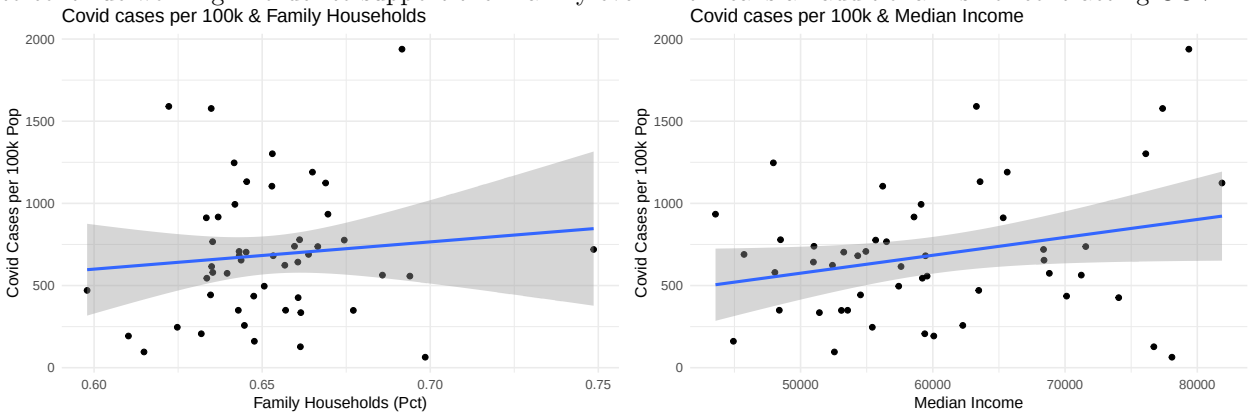



Our analysis showed that our hypothesis was incorrect since our evaluation showed there was not a strong relationship. The correlation plot shows that none of the age groups have a significant R^2's. The scatter plot graphs the percentage of each states' population against the number of COVID cases to help us understand the weak relationship. Evaluating the graph shows that the majority of states have between 35% and 40% of their population above the age of 65. The small variability in the elderly population between states and the wide range of COVID cases helps explain the weak relationship between the variables. Since we're evaluating

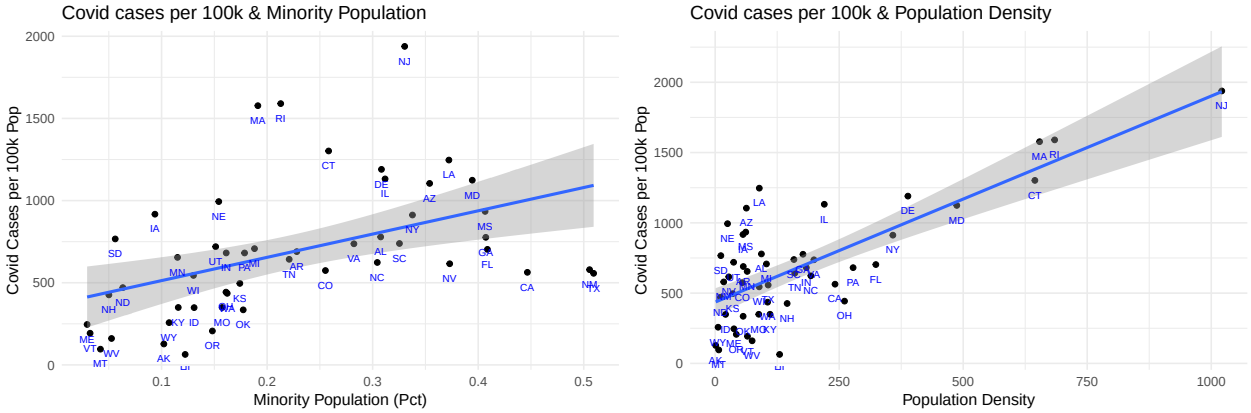
at the state level, there is not a distinctive difference between them that explains the increase in COVID cases. In order to effectively utilize age as a metrics we would need to evaluate smaller populations such as the variance between cities.

Next, we evaluated family status and income to determine if states with a larger percentage of family households or lower income households would lead to a higher number of COVID cases. Studies have shown that over half of people living with someone who has COVID ended up contracted the virus themselves. Based on this, our assumption is that state's with higher number of family households will lead to a higher number of COVID cases. In addition, a lower median household income could be an indication that people will be forced to continue working in order to support their family even if it means an additional risk of contracting COVID.

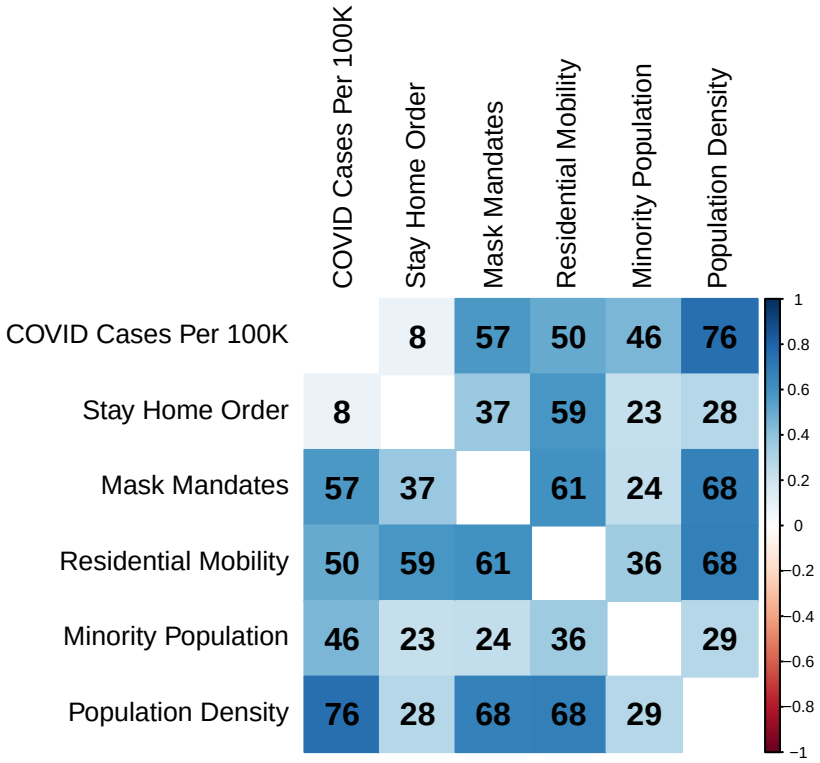



As with age categories, household demographic data between family and non-family households showed very little variability between states. The plot above shows that for most states family households represent between 62% and 68% of all households. Within states between 62% and 68%, there was a wide range of COVID cases per 100k. The small amount of variability in the percentage of family households between states makes it difficult to interpret its impact on the increase in COVID cases. Median household income shows a slightly stronger relationship with COVID but since standard of living and average income vary between states it is difficult to interpret. Evaluating the percentage of a state's population in poverty also lead to similar results. We determined that income was not an ideal metric to explain the increase in COVID cases during the beginning of the pandemic when modeling at the state-level.

The demographic metrics that proved to be the most effective in accurately predicting the number of COVID cases were ethnicity and population density. Previous studies have shown that Black, Hispanic, Native American, and other minority ethnicities were more affected by COVID than white and Asian communities. To capture this impact, we derived a variable representing the percentage of a state's population that was composed of minority communities at higher risk. The derived metric also provided us a more normal distribution with fewer outliers since certain ethnicities differ between states. In addition, population density also showed a strong correlation to Covid cases after we removed DC as an outlier because of its extremely high population density and a low number of COVID cases.



The plots above evaluates the breakdown among states for both the percentage of a state's population from an ethnicity at a higher risk of COVID and population density. Minority Population follows a more normal distribution between 0 and 50% of the population and does show a trend between the increase in population and COVID cases. While there is a lot of variability with states that have a density of 0-250 people per square mile, a clear trend is defined as population density increases. The results also show that states in the Northeastern united states including New Jersey, Rhode Island, Massachusetts, and Connecticut all have the highest population density and COVID rates per 100k population. This could be because they are all smaller states with higher density or because of their geographic proximity increasing the spread of COVID.




Our correlation results support the previous plot and show that minority populations had a .46 R^2 compared to a -.26 R^2 for white population percentage. Minority populations could be affected for a variety of reasons including cultural differences such as multigenerational households or other reasons including status of living and occupation. While ethnicity can't be used as the sole cause of an increase in COVID, there is a clear trend between the two and can be used to explain the increase in cases which was not completely explained by mobility and policy metrics.

Population density has the highest R^2 of .76 which makes it the strongest metric we have evaluated so far. While our original hypothesis was that population density would correlate strongly with minority communities, the results do not support it. In reality, population density correlates strongest with mask mandates and residential mobility. The relationship may be because higher density states and cities are more likely to implement mask mandates early in the pandemic comapred to rural areas. Even though population density has a strong correlation, its distribution is heavily skewed towards lower density states between 0-250. The high correlation to mask mandate and residential mobility could also lead to problems with collinearity when modeling. As a result, we have to be aware of these limitations when using the metric in the model.

Based on our findings, we built model 3 using the derived minority population metric: Case_Per_100k_Pop ~ mob_residential_avg + days_w_mask_mandates + days_w_stay_home_order + minority_pop_pct_at_risk

```
##
## ================================================================
##                                Dependent variable:
##                    ---------------------------------------------
##                                COVID Cases Per 100K
##                           Model 2                Model 3
##                             (1)                    (2)
## ----------------------------------------------------------------
## Residential Mobility      77.204***              59.652**
##                           p = 0.009              p = 0.032
##
## Stay Home Order            -4.602**              -4.752***
##                           p = 0.017              p = 0.009
##
## Minority Population                              976.794***
##                                                  p = 0.007
##
## Mask Mandates              6.422***               6.255***
##                           p = 0.004              p = 0.003
##
## Constant                    11.521                -22.810
##                           p = 0.962              p = 0.918
##
## ----------------------------------------------------------------
## Observations                  50                     50
## R2                          0.441                  0.528
## Adjusted R2                 0.405                  0.486
## Residual Std. Error    311.305 (df = 46)      289.445 (df = 45)
## F Statistic          12.119*** (df = 3; 46) 12.567*** (df = 4; 45)
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

After modeling, we compare the results to model 2 in the table to identify the impact of adding the minority population metric. The results show that Minority Population has a positive coefficient of 977 which confirms our hypothesis that having a higher percentage of at risk minority groups does lead to a higher number of COVID cases. For every 1% increase in population of a state's at-risk minority group, the predicted number of COVID cases per 100k increases by 9.72. Overall, the other metrics also retain their statistical significance with only Residential Mobility showing a significant change in its coefficient from 77 to 59. Like with model 2, this indicates that adding the additional metric changes the bias in the coefficient for residential mobility. This could be because the ethnicity metric is capturing another aspect of the spread of COVID that is not impacted by residential mobility. Compared to model 2, the third model captures a new characteristic of each state to better explain how COVID spreads. While ethnicity cannot be used as the only demographic reason to define the increased spread of COVID, it can help explain differences in spread between states with similar policies and mobility statistics. Additional research will have to be conducted to determine how communities with a high percentage of Black, Hispanic, and Native American residents differ and contribute to the increase of COVID cases.

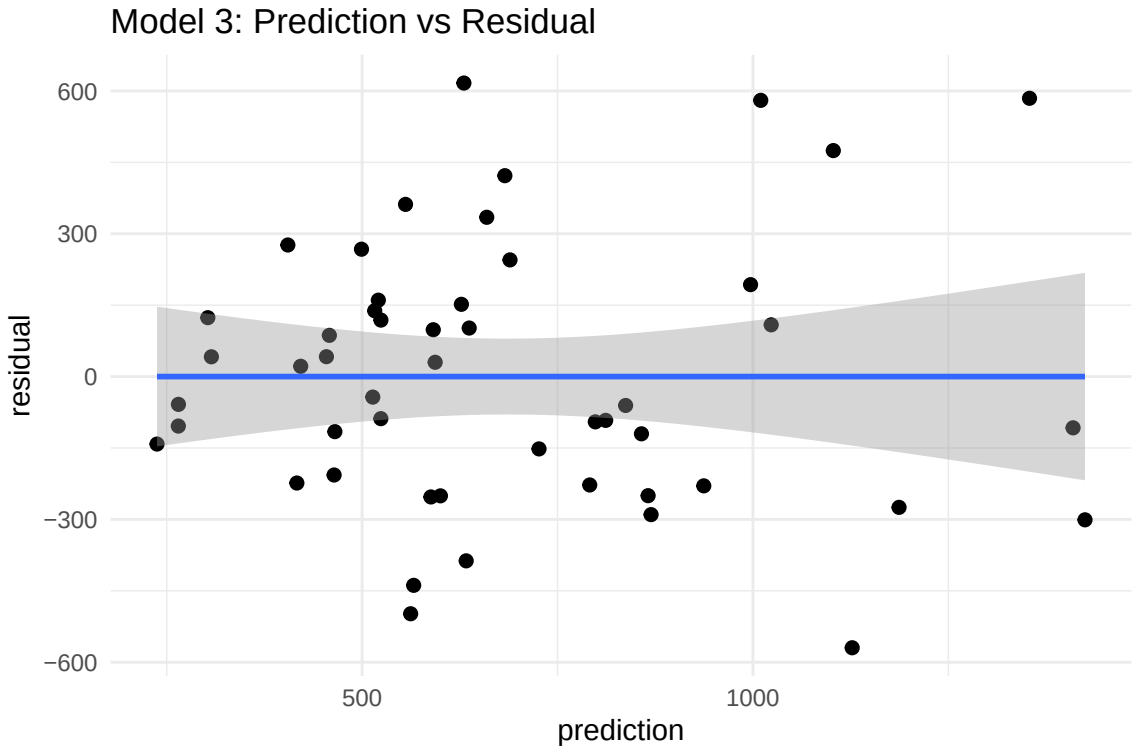
```
## Analysis of Variance Table
##
## Model 1: Case_Per_100k_Pop ~ mob_residential_avg + days_w_mask_mandates +
##     days_w_stay_home_order + minority_pop_pct_at_risk
## Model 2: Case_Per_100k_Pop ~ mob_residential_avg + days_w_stay_home_order +
##     days_w_mask_mandates
```

```
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     45 3770024
## 2     46 4457897 -1   -687873 8.2106 0.00631 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Running an F-test between models 2 and 3 confirms that including ethnicity as a metric increases explanatory power. With a p-value of .006, we are confident that the results are statistically significant and adding the additional variable provides noticeable effect. By adding new metrics we continue to reduce the sum of squares residuals and improve upon our original model.

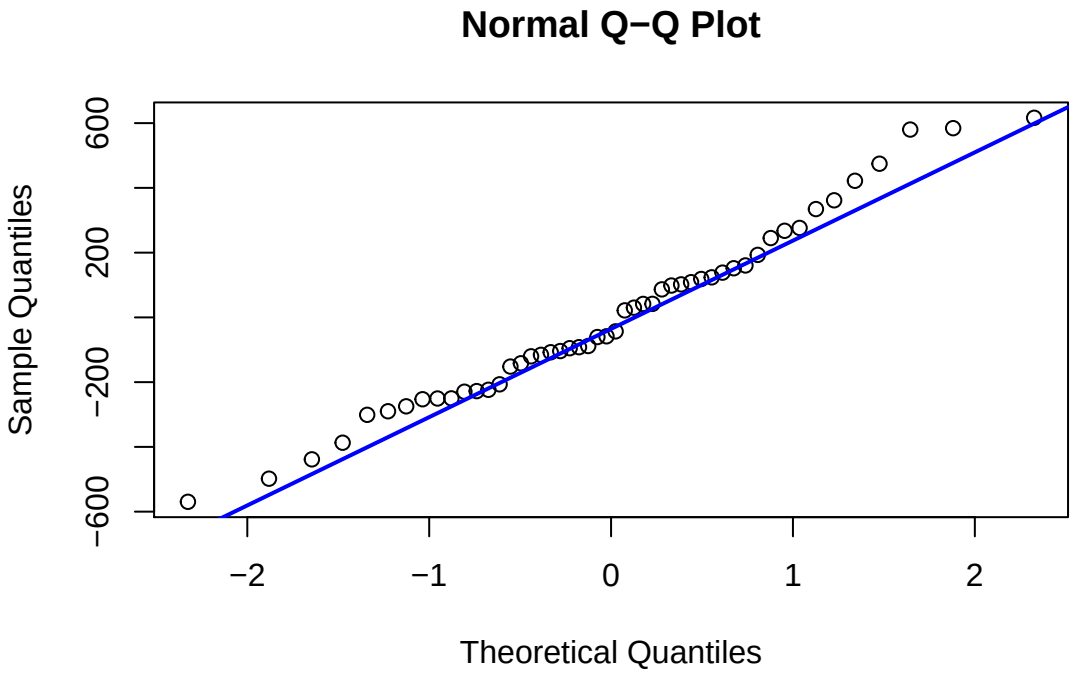
### Model 3: CLM Assumptions

As with model 2, we further assess the model to ensure the CLM assumptions are met.




First, we check the linear conditional expectation by plotting the residuals and predictions in the chart above. The results show that the residuals are not skewed away from 0. Second, our initial correlation matrix confirmed there is no perfect collinearity between the percentage of minority populations and previous metrics.

We evaluate the third condition of normally distributed residuals by plotting a normal Q-Q plot and running a Wilks-Shapiro test




```
##
##  Shapiro-Wilk normality test
##
## data:  model_3$residuals
## W = 0.97849, p-value = 0.4903
```

The plot above shows that the metrics are not perfectly normally distributed but are close enough where we can assume it holds true. Running a Shaprio-Wilk normality test also shows a high p-value of .49 which confirms we can treat the residuals as normal.

Fourth, we run a Breush-Pagen test again to confirm we meet the assumption of homoskedastic conditional variance.

```
bptest(model_3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 7.6532, df = 4, p-value = 0.1051
```

The test gives a p-value of .1051, which is larger than 0.05. Therefore, we fail to reject the null hypothesis that there is no evidence for heteroskedastic error variance. Meeting this assumption has justified using standard errors in previous results.

Finally, we still cannot meet the IID assumption. Since we are still evaluating between states that people can travel between the data is not IID. Since the states with the highest number of COVID cases are all from the Northeast there are serious concerns with it affecting our model. We tried to capture the geographic nature of each state by using one-hot encoding to create metrics representing the US regions defined by the US Census, but were unsuccessful in creating statistically significant metrics. Using only 5 regions is still too broad and further dividing the data into sub-regions complicates the data and the model.

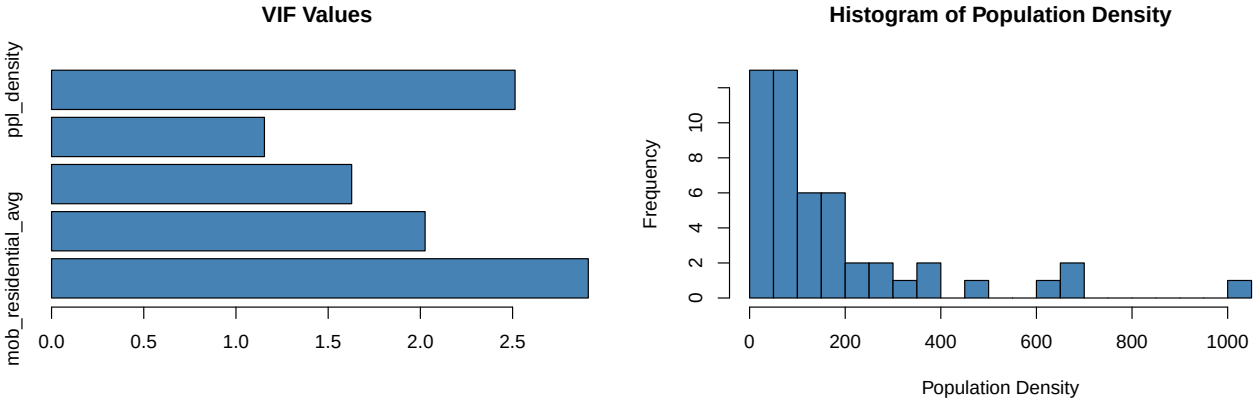
## Model 4:

After running model 3 and determining that adding minority population to the model improves its ability to predict COVID cases per 100k population, we continued to explore how demographic metrics can continue to improve accuracy. The goal of adding new data was to discover the underlying causes of different ethnicities that can lead to an increased number of COVID cases. As shown previously, after removing the outlying variable of DC since it has an extremely high population density, the metric showed a strong relationship to the number of COVID cases. Our hypothesis is that at the start of the pandemic, COVID first spread among large cities and dense populations. Using population density which represents the number of people per square mile, we can better explain how COVID spreads. Areas with high residential mobility but low population density will have different results than more populous states. Since the metric population density has high collinarity with mask mandates and residential mobility, there is a concern that the model will break CLM assumptions.

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                 COVID Cases Per 100K
##                            Model 3                Model 4
##                              (1)                    (2)
## ---------------------------------------------------------------
## Residential Mobility        59.652**               1.607
##                            p = 0.032              p = 0.951
##
## Mask Mandates               6.255***               2.416
##                            p = 0.003              p = 0.185
##
## Stay Home Order            -4.752***              -3.108**
##                            p = 0.009              p = 0.042
##
## Minority Population        976.794***             895.500***
##                            p = 0.007              p = 0.003
##
## Population Density                                 1.205***
##                                                  p = 0.00004
##
## Constant                    -22.810                357.536*
##                            p = 0.918              p = 0.080
##
## ---------------------------------------------------------------
## Observations                  50                     50
## R2                           0.528                  0.681
## Adjusted R2                  0.486                  0.645
## Residual Std. Error    289.445 (df = 45)      240.429 (df = 44)
## F Statistic          12.567*** (df = 4; 45) 18.814*** (df = 5; 44)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

The results of model 4, that includes the new metric population density, show a mix of statistical significance. Overall while the length of stay at home, percent of minority population, and population density all show statistical significance, but the model no longer shows significance for mobility and mask mandates. Our original concerns of perfect collinarity appear to be valid and reduce the significance of our residential mobility metric. Instead of explaining the underlying causes for higher rates of Covid for minority communities, population density has a stronger relationship with mobility and mask mandates. The results could be the

cause of states with denser populations implementing mask mandates sooner due to increased risk of Covid transmission. Overall population density does not help us determine the cause of increased Covid cases with states with high minority populations. We may need to explore beyond demographic into more cultural or health aspects to correctly identify the reason and give a causal explanation.




Running a Variance Inflation Factor test shows that there is potential concern for multicollinearity since Residential Mobility is approaching 3, but it is not considered severe. Another issue is the non-normal distribution of the Population Density metric. Even after removing the extreme outlier of NJ still returns a skewed distribution. Overall, model 4 reduces the impact of our key metrics and is not an improvement compared to model 3. We would need to explore new variables or transform Population Density and remove outliers in order to build a more accurate model.

## Limitations of Models

Google Mobility Data - Mobility data, to a certain extent, represents a certain population who has a Google account and allows Google to use their data for other purposes. Grouping the mobility for an entire state over 5 months does take variations away from the data and bring the overall variable towards the overall averages. During our exploration phase, we identified the certain the state did spike in visits during March 2020 as the weather started to improve and celebrate the start of spring.

Demographic Data - Evaluating demographic data at the state level captures a large population of a state even if only a small part of the population has been affected by Covid-19 during the initial months of the pandemic. Many demographic metrics evaluated did not show a significant causal effect.

Summarization of models using variables such as age, income, population density and ethnicity does take details away from the data and pushes the entire data set to similar averages. For example, the average income for a family of 4 in California 2018 is around 71k but similar average income in San Francisco 2018 was 110k while Alameda County is 54k this variation would be critical to identify the impact of income in a model.

Mask Mandate - As COVID was spreading through the country, we are all aware of the debates around executing the mask mandate and the severity of the enforcement may differ across the state. But, the current data set does provide the start and end date of the mask mandate.

Median Income - Income metrics do not account for cost of living differences between and within states. Difficult to determine lower income communities with a single metric.

Reverse Causality - We could identify the reverse causality between Mobility and COVID spread as the increase in cases do warrant folks to stay indoors and as the spread of COVID slows down people tend to travel more. But, similar to other reverse causality in the other variables.

**Omitted Variables - The following are relevant variables that were not a part of our dataset.**

- Number of COVID tests - Since testing varied by state at the start of the pandemic, COVID cases could be inflated or under reported depending on the number of tests conducted.
- Health metrics - Additional metrics such as obesity rates and other health related metrics could be more useful in describing a population compared to ethnicity or other demographic metrics.
- Tourist and Travel metrics - Travel metrics such as number of flights or hotels booked could be useful to define the increased spread of COVID especially during the first months of the pandemic.
- Housing Metrics - Additional metrics such as percentage of people living in apartments, single family homes, or public housing could better explain the relationship to COVID cases than family households.
- Elderly Care Facilities - Since Nursing homes were hardest hit at the beginning of the pandemic, having a metric to define the total number of people living in elderly care facilities could be more useful than age categories.

# Conclusion

We started with a casual question around the relationship between mobility and COVID spread. The casual nature of the question did provide us multiple avenues to explore other factors that may provide insight into the modeling COVID spread in the US.

During the exploration, we did identify these various features that would provide additional refinement and can be encapsulated within these four models that predicted COVID spread using statistically significant variables such as population density, mask mandate, state home order duration, minority population, and residential mobility. Of all the four models, model four does have the highest r squared values but we prefer model three as model four does have issues with collinearity.

But adding the variable 'minority population at risk' in model three does stabilize other variables & we are confident that the results are statistically significant and increase explanatory power. But, we can identify that adding singular ethnic groups (Asian, Hispanic, Black, or White) to model three may not be correct as being a member of a certain minority group does make an individual more likely to get COVID nor does a state with a higher population to a specific minority will put more people at risk. Ethnicity does capture some of our omitted variable characteristics such a culture, food habits, genetic predispositions, susceptibility to colder weathers, etc. hence grouping the minority groups into variable does help our model.

Overall, our model accurately describes the impact of mobility, policy, and certain demographic characteristics on the impact of COVID cases at the start of the pandemic. The next steps to continue to improve our model will be to research and implement metrics from new datasets to help handle and explain the omitted-variable bias.

##

```
## Regression Comparisons
## ================================================================================
##                                        Dependent variable:
##                     ---------------------------------------------------------
##                                        COVID Cases Per 100K
##                     Model 2-StayHome/Mask   Model 3-Ethnicity     Model 4-Minority
##                              (1)                   (2)                   (3)
## --------------------------------------------------------------------------------
## Residential Mob           77.20***              59.65**                1.61
##                           p = 0.01             p = 0.04              p = 0.96
## Stay Home                  -4.60**             -4.75***               -3.11**
##                           p = 0.02             p = 0.01              p = 0.05
## Minority Pop                                   976.79***             895.50***
##                                                p = 0.01              p = 0.003
## Population Density                                                    1.20***
##                                                                     p = 0.0001
## Mask Mandate               6.42***              6.26***                2.42
##                           p = 0.004            p = 0.003             p = 0.19
## Constant                    11.52                -22.81                357.54*
##                           p = 0.97             p = 0.92              p = 0.08
## --------------------------------------------------------------------------------
## Observations                 50                    50                    50
## R2                          0.44                  0.53                  0.68
## Adjusted R2                 0.41                  0.49                  0.65
## Residual Std. Error   311.31 (df = 46)      289.44 (df = 45)      240.43 (df = 44)
## F Statistic         12.12*** (df = 3; 46) 12.57*** (df = 4; 45) 18.81*** (df = 5; 44)
## ================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

**End of Lab 2 - Team Flu-Fighter**