Modelling Chapter - Lucas

Causal Diagram

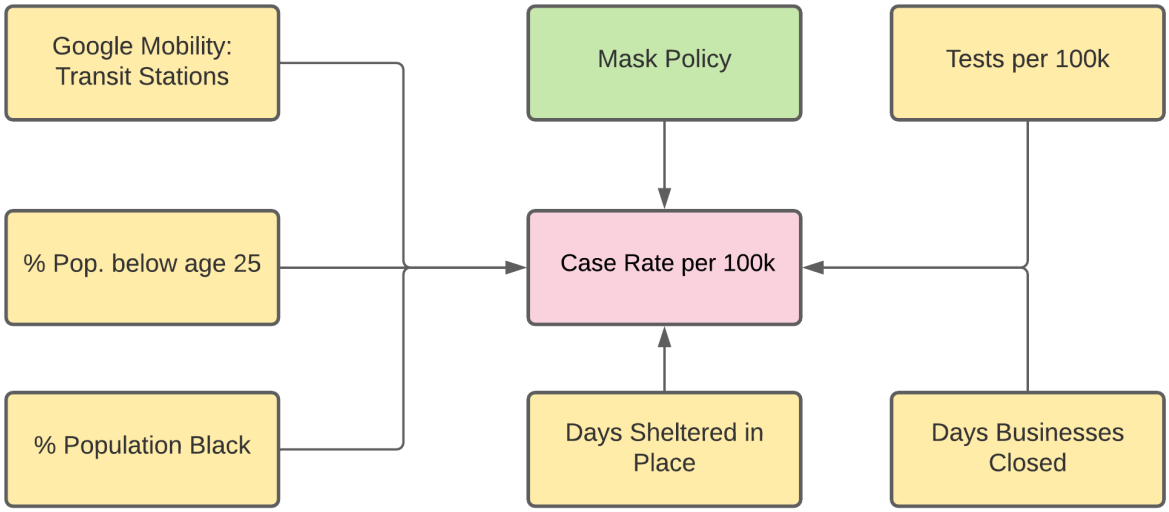


Figure 1: Causal Diagram

Data Summary and Cleaning

The data used in the model is taken from the provided covid_19 dataset. The provided dataset is up-to-date as of October 30th, 2020. The final model uses eight variables that are either created from existing data or supplemented from online resources. Below are the adjustments made to variables that were either created or supplemented.

Mask Use

This binary/logical variable was created by assigning a 1 if the state had a mask mandate and 0 if it did not (based on the mask mandate date column). Some of the data from the mask mandate was missing and was manually research and populated.

Percent Age Below 25

This column was created by combining the 0-18 and 18-25 age groups. No other adjustments were made.

Days in Shelter-in-Place

The number of days each state was under the Shelter-in-Place mandate. This data was missing some data and supplemented by researching and populating the missing data. The column was creating by subtracting the end and start dates.

Days Businesses Closed

The number of days each state closed non-essential businesses. Similar to the days in SIP, missing data was populating through research of the state's specific mandates, then calculated by subtracting the end and start dates.

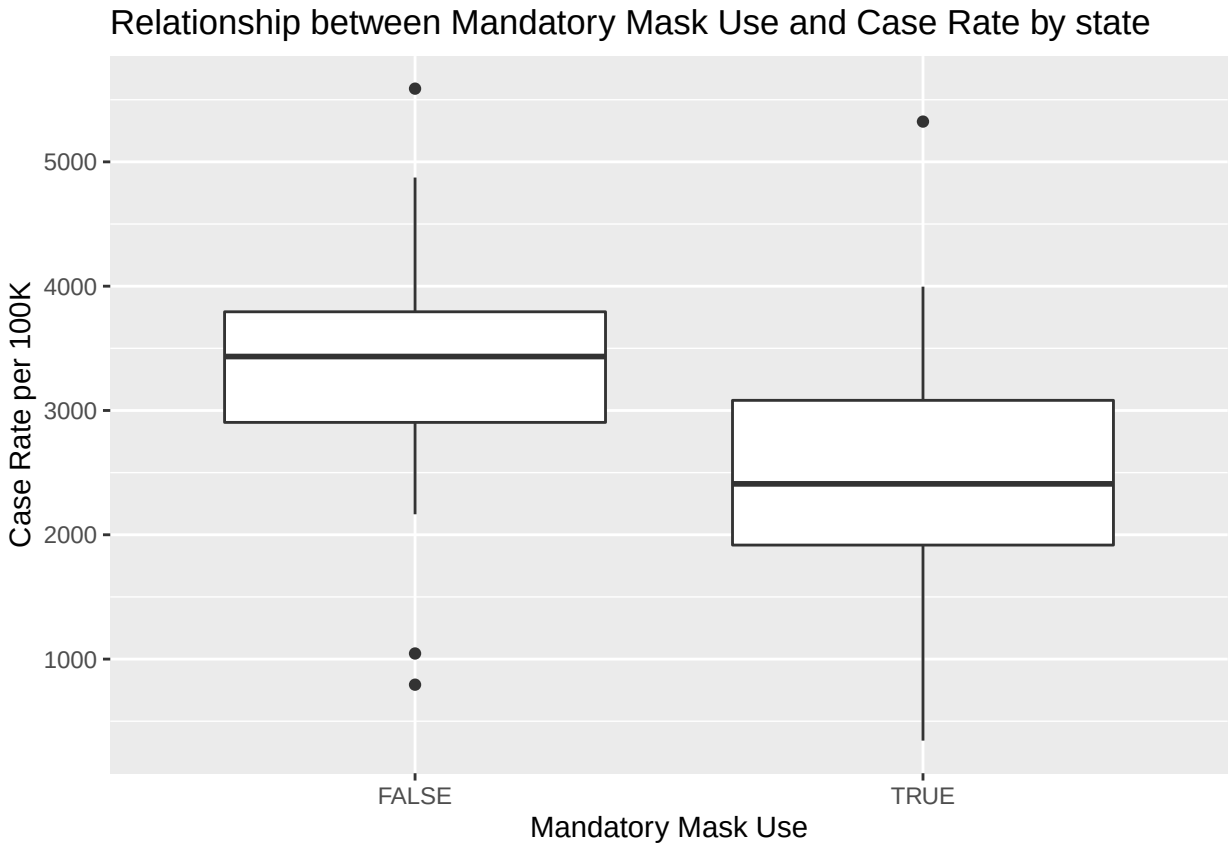
Percentage of Pouplation: Black

Observations that were marked as "< 0.01" were rounded up to 0.01 for a log transformation to be applied.

Exploratory Data Analysis

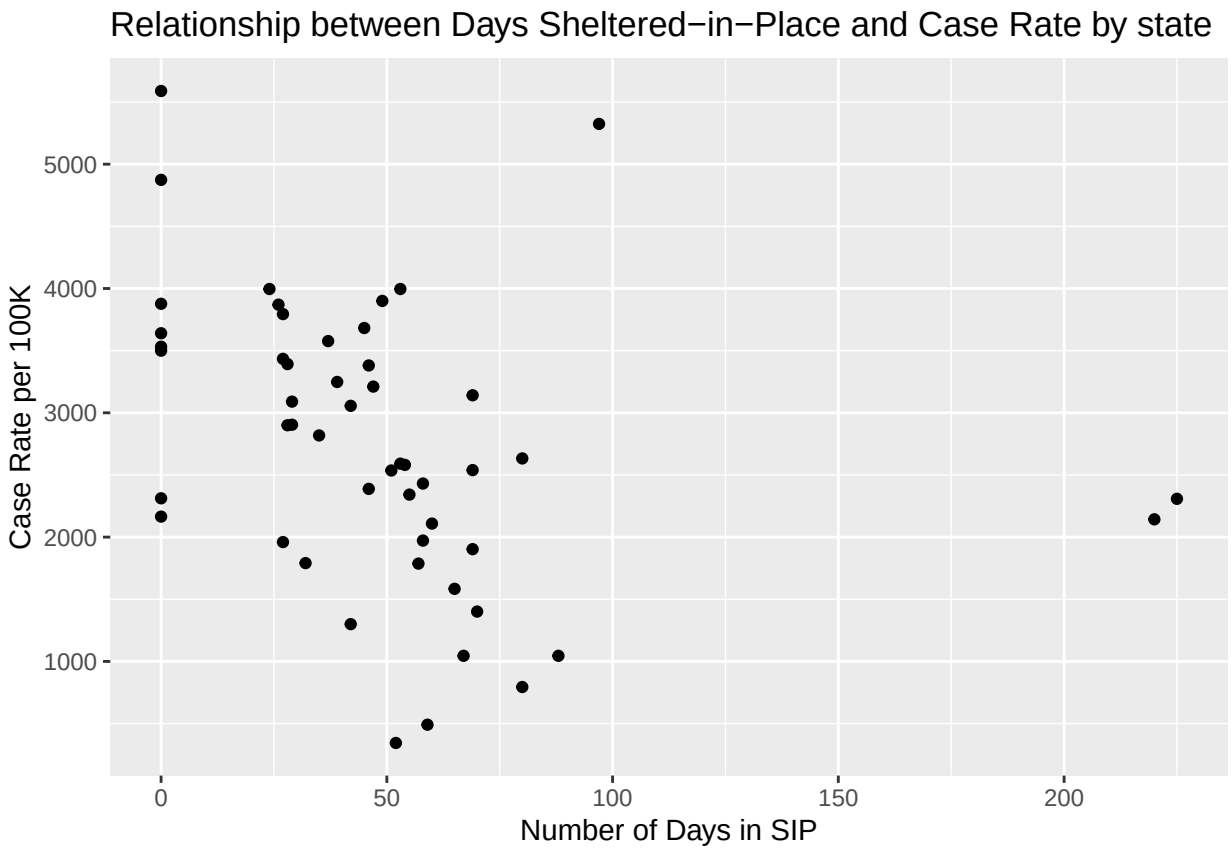
Talk about it being a policy and mandatory Based on our causal model, mask use should lead to a decrease in the number of covid cases. The boxplot below suggests that this initial assumption at least holds to some degree.

```
df %>%
  ggplot(aes(y = case_rate, x = mask_use)) +
  geom_boxplot() +
  labs(
    title = "Relationship between Mandatory Mask Use and Case Rate by state",
    x = "Mandatory Mask Use",
    y = "Case Rate per 100K"
  )
```

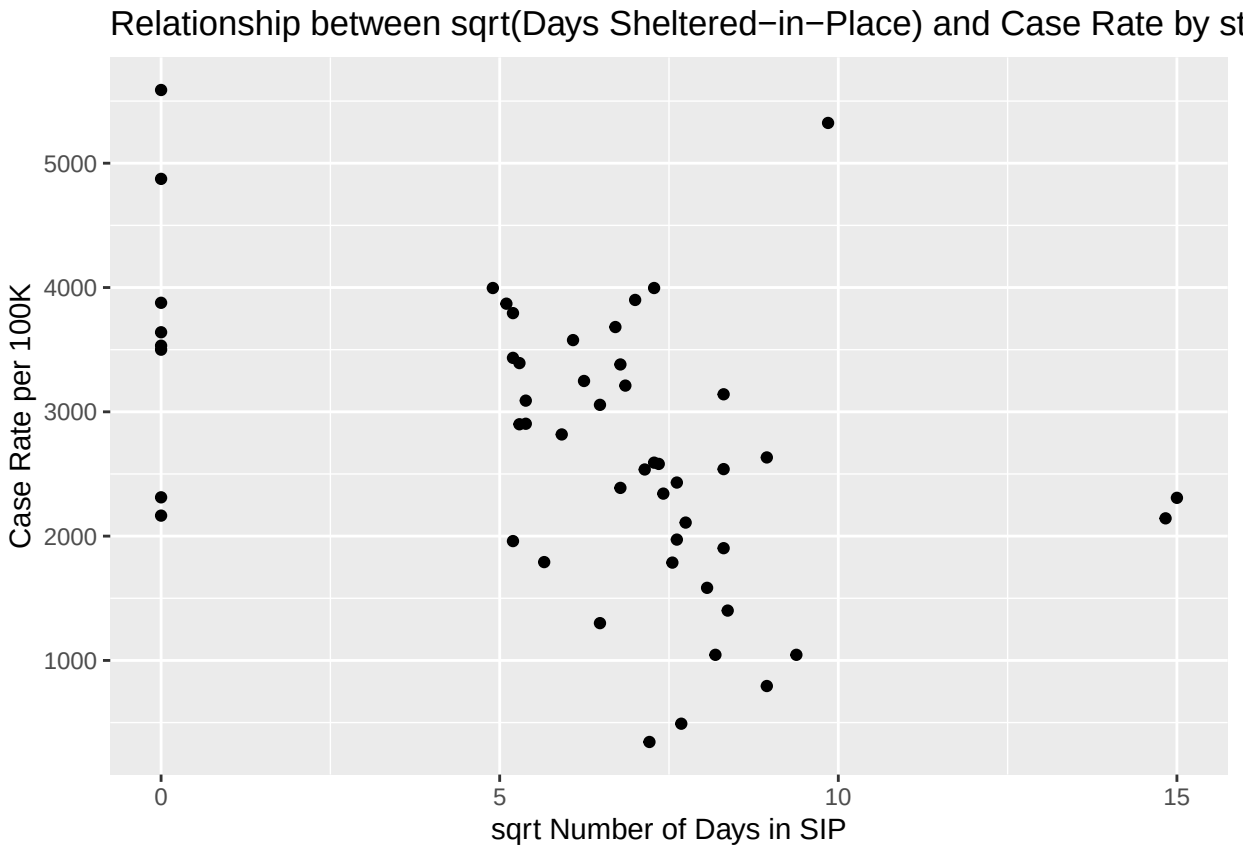


COVID-19 is a virus that spreads from person to person and the first stop gaps implemented were the closing of non-essential businesses and a Shelter-in-Place (SIP) mandate. A square-root transformation of the data better interprets the data as the correlation increases. *Diminishing returns and more linear relationship.*

```
df %>%
  ggplot(aes(y = case_rate, x = shelter_days)) +
  geom_point() +
  labs(
    title = "Relationship between Days Sheltered-in-Place and Case Rate by state",
    x = "Number of Days in SIP",
    y = "Case Rate per 100K"
  )
```




```
df %>%
  ggplot(aes(y = case_rate, x = sqrt(shelter_days))) +
  geom_point() +
  labs(
    title = "Relationship between sqrt(Days Sheltered-in-Place) and Case Rate by state",
    x = "sqrt Number of Days in SIP",
    y = "Case Rate per 100K"
  )
```




```
sip <- lm(case_rate ~ shelter_days, data = df)
sqrt_sip <- lm(case_rate ~ sqrt(shelter_days), data = df)
summary(sip)$r.squared
```

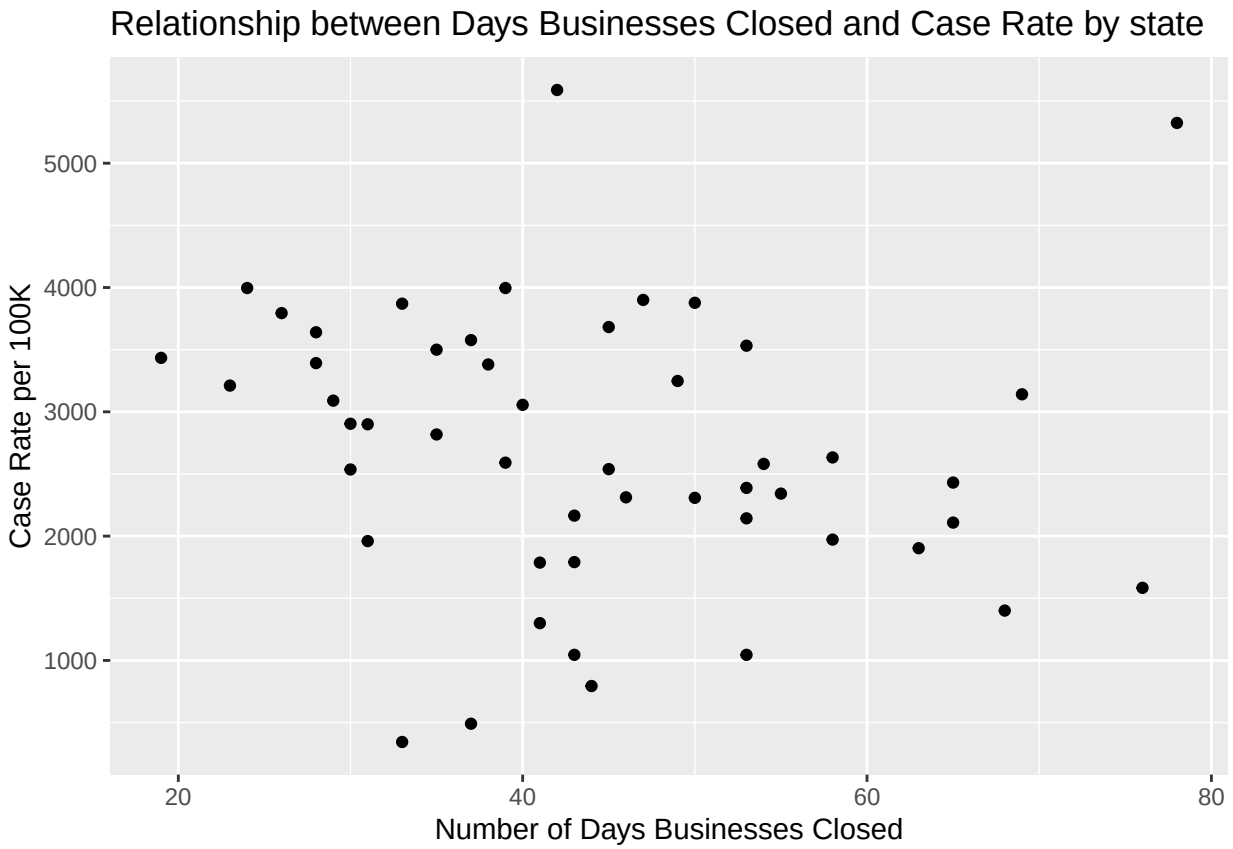
```
## [1] 0.1068833
```

```
summary(sqrt_sip)$r.squared
```

```
## [1] 0.1770349
```

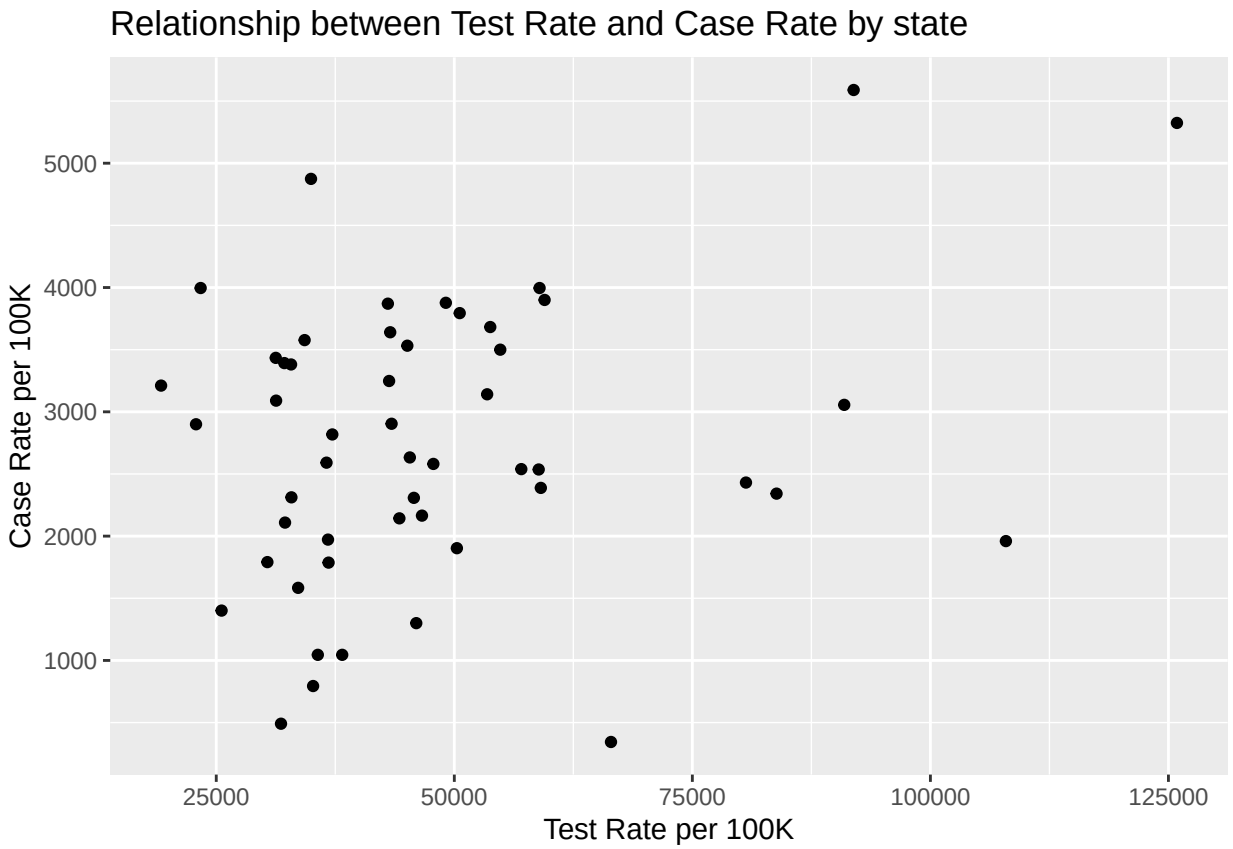
Similarly, the number of days a non-essential business was closed would greatly reduce the number of human-human interactions. A log transformation of the number of businesses days closed also serves to increase the correlation slightly.

```
df %>%
  ggplot(aes(y = case_rate, x = bus_close_days)) +
  geom_point() +
  labs(
    title = "Relationship between Days Businesses Closed and Case Rate by state",
    x = "Number of Days Businesses Closed",
    y = "Case Rate per 100K"
  )
```



The number of positive cases a state has is strictly dependent on the number of tests administered. From the scatter plot below, we do see a positive relationship between the number of tests administered and the number of cases. Log and square-root transformations do not lead to a higher correlation therefore the variable is left as is.

```
df %>%
  ggplot(aes(y = case_rate, x = test_rate)) +
  geom_point() +
  labs(
    title = "Relationship between Test Rate and Case Rate by state",
    x = "Test Rate per 100K",
    y = "Case Rate per 100K"
  )
```



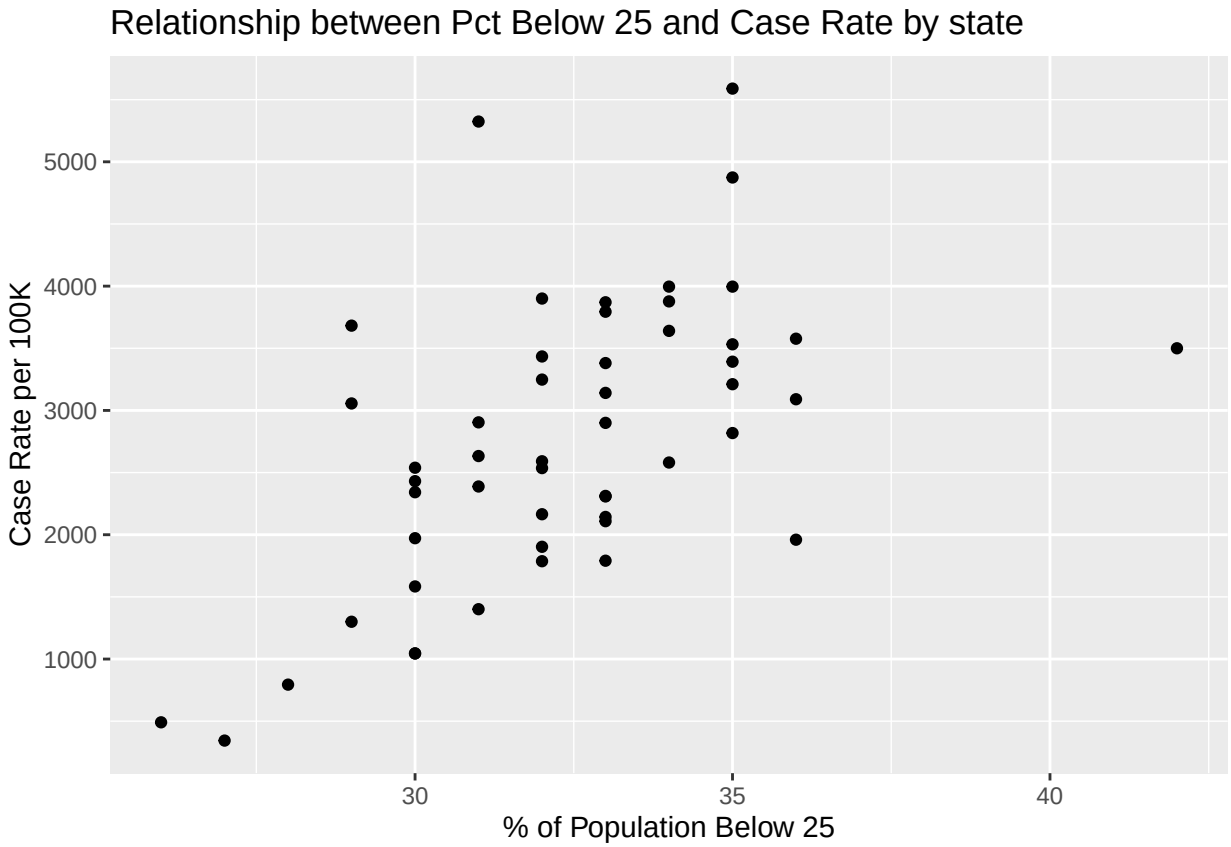
Next, we look at the correlation between case rate and age. From the covariance matrix below, age groups 0-18 and 65+ are not only the most extreme of the age groups, but covariances as well.

```
var(df[ ,c(4, 51:56)], na.rm=TRUE)
```

```
##               case_rate     age_0_18   age_19_25  age_26_34     age_35_54
## case_rate 1291651.53020 1200.3352941 513.0494118 54.7219608 -335.82666667
## age_0_18     1200.33529    5.0329412   0.8741176 -0.1870588   -0.76000000
## age_19_25     513.04941    0.8741176   0.5717647  0.4541176   -0.22000000
## age_26_34      54.72196   -0.1870588   0.4541176  2.2596078    0.31333333
## age_35_54    -335.82667   -0.7600000  -0.2200000  0.3133333    0.98666667
## age_55_64    -471.88980   -1.8047059  -0.5505882 -1.0980392    0.05333333
## age_65       -746.74706   -2.9505882  -1.0188235 -1.6705882   -0.44000000
##                age_55_64      age_65
## case_rate -471.88980392 -746.747059
## age_0_18    -1.80470588   -2.950588
## age_19_25   -0.55058824   -1.018824
## age_26_34   -1.09803922   -1.670588
## age_35_54    0.05333333   -0.440000
## age_55_64    1.53019608    1.872941
## age_65       1.87294118    4.294118
```

```
df$age_below_25 = df$age_0_18 + df$age_19_25
df %>%
  ggplot(aes(y = case_rate, x = age_below_25)) +
  geom_point() +
```

```
  labs(
    title = "Relationship between Pct Below 25 and Case Rate by state",
    x = "% of Population Below 25",
    y = "Case Rate per 100K"
  )
```



We also examine the case rate vs ethnicity. The black population is much more correlated with case rate than any of the other ethnic groups. Being white is also negatively correlated with the case rate.

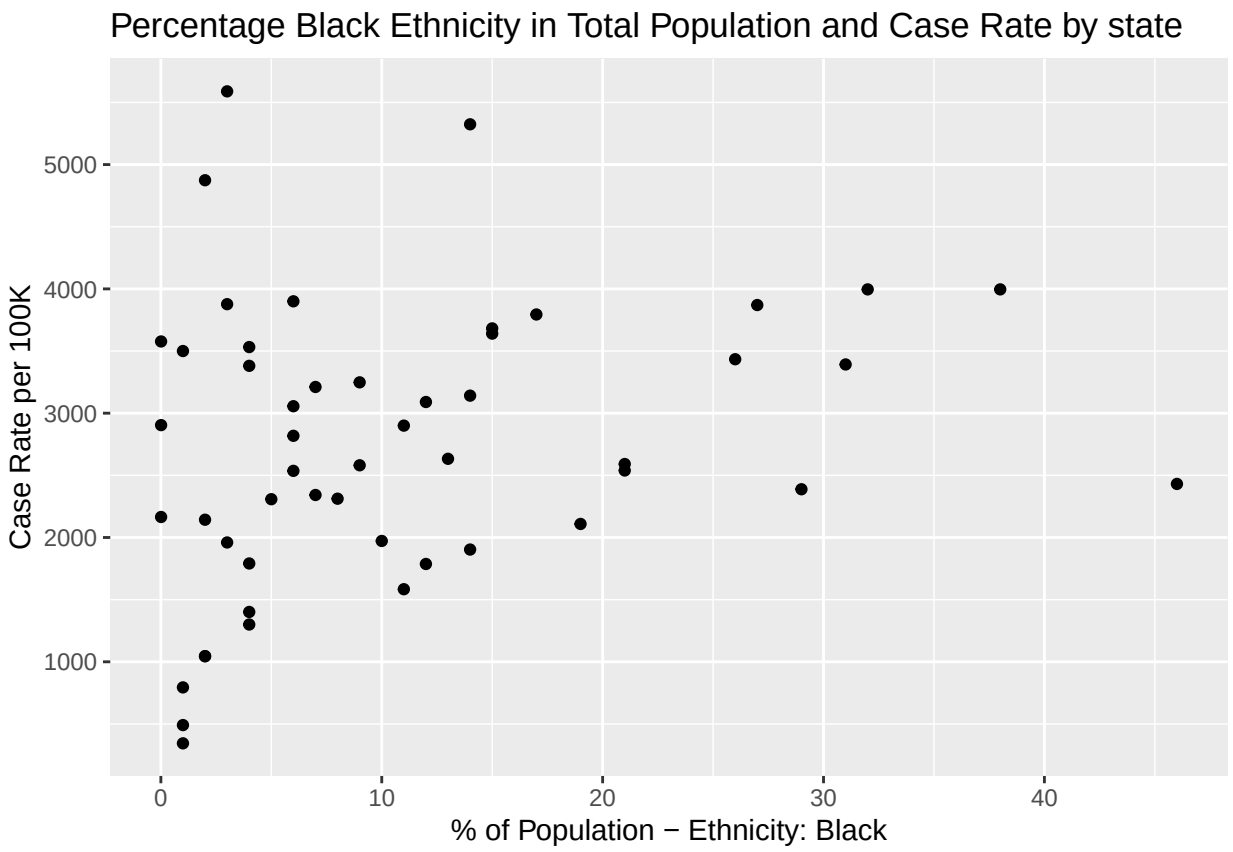
```
var(df[ ,c(4,14,16,18,20)], na.rm=TRUE)
```

```
##                case_rate   white_pop   black_pop hispanic_pop   other_pop
## case_rate    1291651.5302 -1410.79725 3061.765882   612.262745 -781.869412
## white_pop      -1410.7973   292.47843  -76.757647  -116.181569  -17.231765
## black_pop       3061.7659   -76.75765  113.496471   -14.417647   -3.542353
## hispanic_pop     612.2627  -116.18157  -14.417647   108.318431   -4.711765
## other_pop       -781.8694   -17.23176   -3.542353    -4.711765   13.931765
```

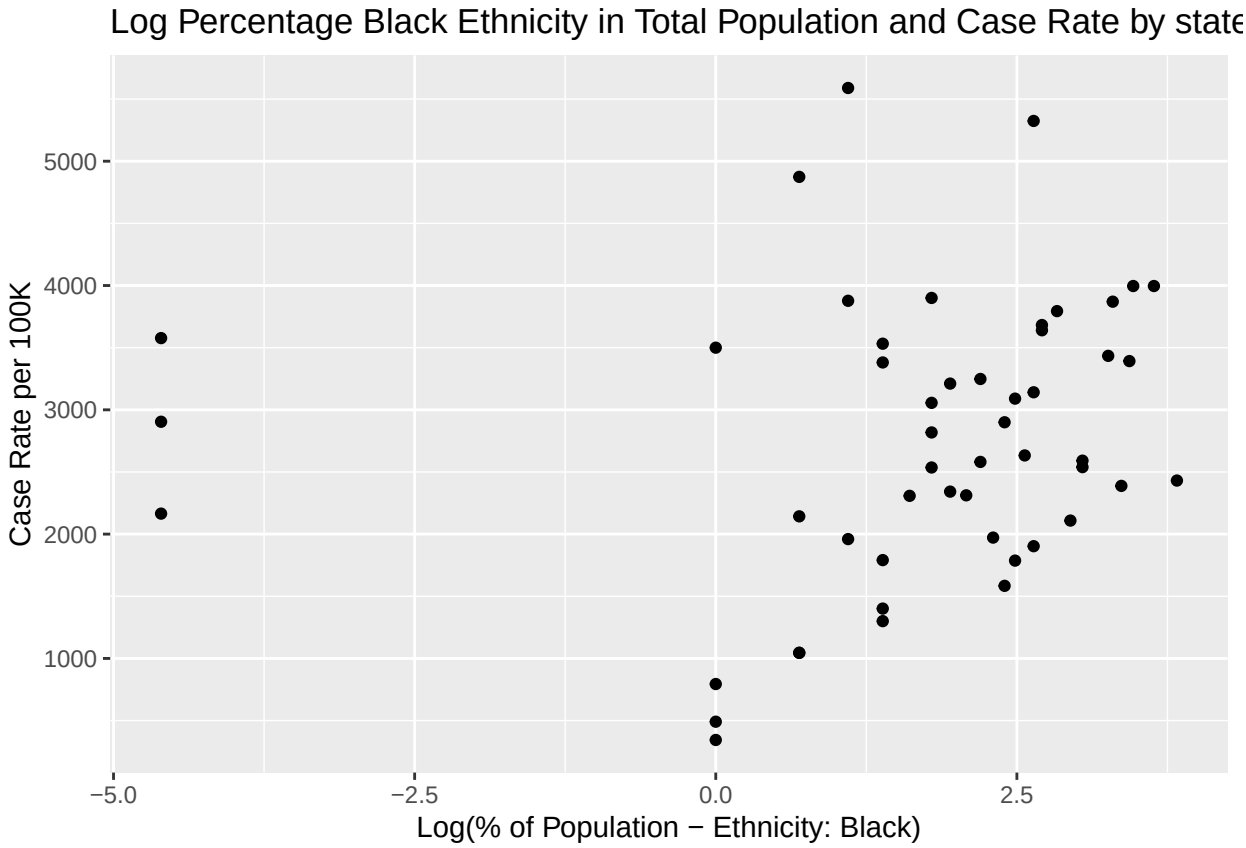
We more closely examine the correlation between black population percentage, the most highly correlated ethnicity variable, and case rate. From the graph below, we see that there is a slightly concaving arc as the proportion of black increases. ~~We apply a log transformation to see if the correlation increases and see that it does.~~

```
df$black_pop[df$black_pop == 0] = 0.01
df %>%
```

```
  ggplot(aes(y = case_rate, x = black_pop)) +
  geom_point() +
  labs(
    title = "Percentage Black Ethnicity in Total Population and Case Rate by state",
    x = "% of Population - Ethnicity: Black",
    y = "Case Rate per 100K"
  )
```




```
df %>%
  ggplot(aes(y = case_rate, x = log(black_pop))) +
  geom_point() +
  labs(
    title = "Log Percentage Black Ethnicity in Total Population and Case Rate by state",
    x = "Log(% of Population - Ethnicity: Black)",
    y = "Case Rate per 100K"
  )
```




```
blk_pop <- lm(case_rate ~ black_pop, data = df)
log_blk_pop <- lm(case_rate ~ log(black_pop), data = df)
summary(blk_pop)$r.squared
```

```
## [1] 0.06395714
```

```
summary(log_blk_pop)$r.squared
```

```
## [1] 0.02772813
```

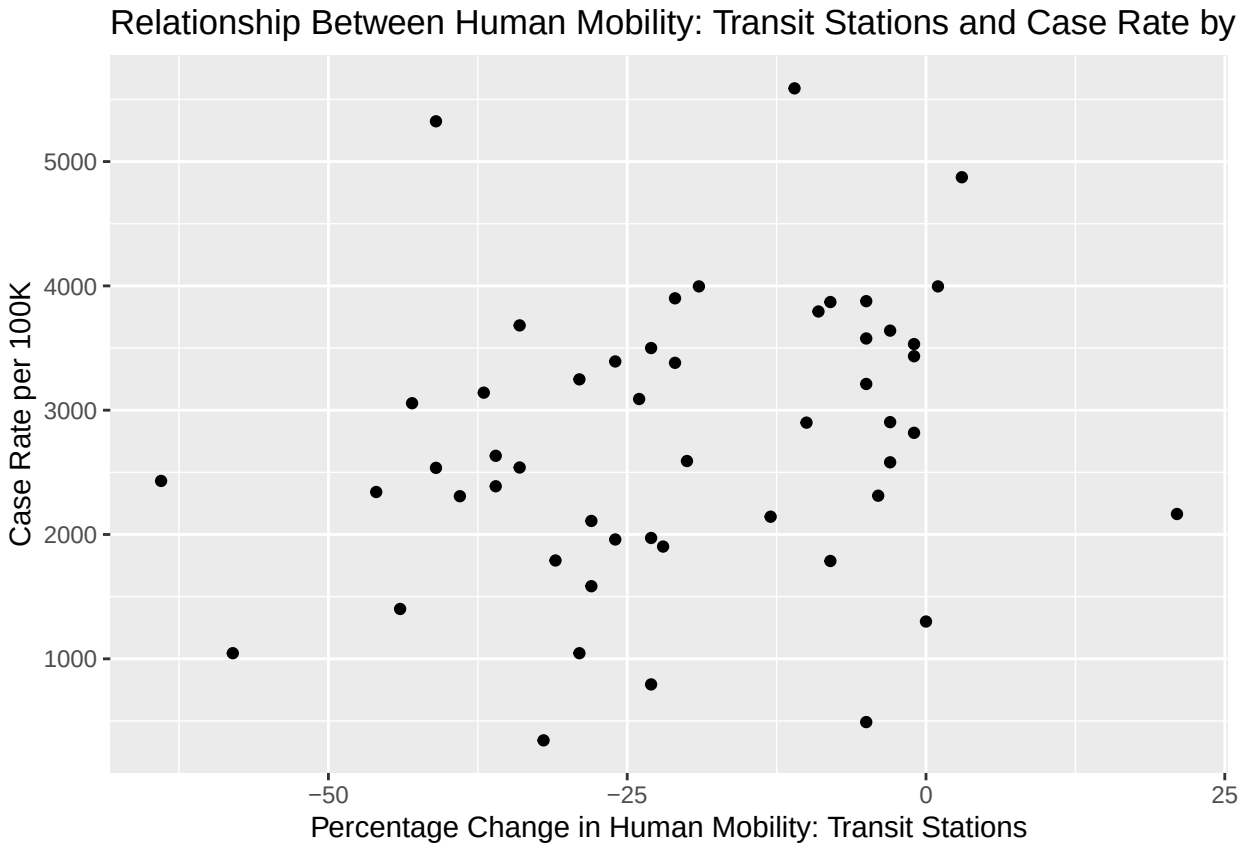
*Add data from Google Human Mobility on time period of data*Included in the dataset is the Google Human Mobility metric regarding the change in activity of popular spots (retail, groceries, parks, etc.,). We find that the highest correlation of the mobility variables lies with the change in Transit Stations (mob_TS). Oddly enough, Wyoming is one of three states with a positive change in transit station use. Wyoming also has by far the largest change at 21% with South Dakota next at 3%.

```
var(df[ ,c(4, 57:62)], na.rm=TRUE)
```

```
##              case_rate    mob_R&R    mob_G&P       mob_P     mob_TS     mob_WP
## case_rate 1291651.5302 1618.18235 995.136471 -9142.82078 4669.04157 1177.39843
## mob_R&R      1618.1824   56.02824  31.997647   139.19059   96.65882   40.62118
## mob_G&P       995.1365   31.99765  34.543529    90.81412   61.47176   23.22824
## mob_P       -9142.8208  139.19059  90.814118  1348.20314  168.71373   77.12627
## mob_TS       4669.0416   96.65882  61.471765   168.71373  301.09255   98.56745
```

```
## mob_WP       1177.3984   40.62118  23.228235    77.12627   98.56745   45.01255
## mob_RES      -582.8475  -16.50941  -8.345882   -45.63020  -39.23961  -15.12039
##                mob_RES
## case_rate -582.847451
## mob_R&R    -16.509412
## mob_G&P     -8.345882
## mob_P      -45.630196
## mob_TS     -39.239608
## mob_WP     -15.120392
## mob_RES      8.043137
```

```
df %>%
  ggplot(aes(y = case_rate, x = mob_TS)) +
  geom_point() +
  labs(
    title = "Relationship Between Human Mobility: Transit Stations and Case Rate by state",
    x = "Percentage Change in Human Mobility: Transit Stations",
    y = "Case Rate per 100K"
  )
```



Model Building

Model 1

Model description

Our first regression model has **Covid-19 Case Rate per 100,000 habitants** as our outcome variable and two covariates: our variable of interest (**Mandatory Mask Use**) and **Test Rate per 100,000 habitants**.

We have included **Test Rate** because we've seen before there is variability in the test rates among US states which could potentially affect the integrity of our outcome variable - e.g. states presenting lower case rate not because actual infection rate by covid is lower, but because test availability is lower. In order to mitigate such shortcomings we decided to have **Test Rate** as an independent variable since the very first version of our regression model.

Our Model 1 has the format:

case_rate = β_0 + β_1 * mask_use + β_2 * test_rate

Model summary

```
model_1 <- lm(case_rate ~ mask_use + test_rate, data = df)
std_errors = sqrt(diag(vcovHC(model_1)))
stargazer(model_1, se = std_errors, type = "text", title = "Model 1 Summary")
```

```
##
## Model 1 Summary
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               case_rate
## ------------------------------------------------
## mask_use                     -990.470
##
##
## test_rate                      0.018
##
##
## Constant                   2,530.239***
##                              (501.044)
##
## ------------------------------------------------
## Observations                    51
## R2                             0.236
## Adjusted R2                    0.204
## Residual Std. Error     1,013.835 (df = 48)
## F Statistic            7.416*** (df = 2; 48)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Overall model significance (F-test)

Under a significance level of 0.05, we can reject the null hypothesis ($H_0 : \beta_1 = \beta_2 = 0$) in favor of our fuller model ($H_a : \beta_1 \neq 0$ and $\beta_2 \neq 0$), which now includes the covariates **mask_use** and **test_rate**. The F-Statistic = 7.416, and the p-value < 0.01. Our Model 1 has an adjusted R-squared of 0.204.

Coefficient significance (t-test)

Under a significance level of 0.05, we can accept the alternative hypotheses $H_{a1} : \beta_1 \neq 0$ and $H_{a2} : \beta_2 \neq 0$, which means both of our covariates do explain part of the variability observed in the **case_rate**.

Our estimate for β_1 (the coefficient of our variable of interest) is $\tilde{\beta}_1 = -990.5$, with a standard error of 307.0 and a p-value of 0.002.

```
coeftest(model_1, vcovHC = vcovHC(model_1, type = "HC3"))
```

```
##
## t test of coefficients:
##
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  2530.239162  378.070268  6.6925 2.178e-08 ***
## mask_useTRUE -990.469625  307.005273 -3.2262  0.002261 **
## test_rate       0.018295    0.006801  2.6900  0.009800 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical significance

According to Model 1, states that have adopted mandatory mask use would expect to have -990.5 covid cases per 100,000 habitants. Given that the median of covid case rate among US states is of 2,633 per 100,000 habitants, with 1st Quartile = 2,040 and 3rd quartile = 3,516, the coefficient estimate has practical significance, with an effect size corresponding to a reduction of 37.6% of the median of the covid case rate among states.

Model 2

Model description

Our second regression model has **Covid-19 Case Rate per 100,000 habitants** as our outcome variable and five covariates: our variable of interest (**Mandatory Mask Use**), **Test Rate per 100,000 habitants**, **Percentage of Population Below 25 Years Old**, **Log of Percentage of Black Ethnicity in Total Population**, and **Human Mobility Change in Transit Stations**.

Model 2 represents our best understanding of the relationship among the variables, striking a balance between accuracy and parsimony. Our variable of interest continues to be **Mandatory Mask Use** and our measurement goal continues to be to assess the significance and practical impact of **Mandatory Mask Use** in the **Case Rate**. The other mediating variables we added to the model work like *controls* in order to allow us to better capture the significance and practical relevance of the **Mandatory Mask Use** in the decrease of **Case Rate**.

Our Model 2 has the format:

case_rate = β_0 + β_1 * mask_use + β_2 * test_rate + β_3 * age_below_25 + β_4 * log(black_pop) + β_5 * mob_TS

Model summary

```
model_2 <- lm(case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop) + mob_TS, data = df)
std_errors = list(
  sqrt(diag(vcovHC(model_1))),
  sqrt(diag(vcovHC(model_2)))
)
stargazer(model_1, model_2, se = std_errors, type = "text", title = "Model 2 Summary")
```

```
##
## Model 2 Summary
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                       case_rate
##                              (1)                   (2)
## ---------------------------------------------------------------
## mask_use                  -990.470***           -961.366***
##                            (324.753)             (240.032)
##
## test_rate                   0.018*                0.024**
##                             (0.010)               (0.012)
##
## age_below_25                                     191.802***
##                                                   (57.522)
##
## log(black_pop)                                   221.195***
##                                                   (83.120)
##
## mob_TS                                            16.017**
##                                                   (7.479)
##
## Constant                2,530.239***           -3,999.773*
##                           (501.044)            (2,133.506)
##
## ---------------------------------------------------------------
## Observations                 51                    51
## R2                          0.236                 0.633
## Adjusted R2                 0.204                 0.593
## Residual Std. Error  1,013.835 (df = 48)    725.343 (df = 45)
## F Statistic         7.416*** (df = 2; 48) 15.550*** (df = 5; 45)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Overall model significance (F-test)

Under a significance level of 0.05, we can reject the null hypothesis (**Model 1**) in favor of our fuller **Model 2**, which now includes the covariates **mask_use**, **test_rate**, **age_below_25**, **log(black_pop)**, and **mob_TS**. The F-Statistic = 16.258, and the p-value < 0.01. Our Model 2 has an adjusted R-squared of 0.593. Our residual standard error decreased from 1,013.8 to 725.3. **Model 2** is the most robus model we will build in the scope of this study.

```r
anova(model_1, model_2, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: case_rate ~ mask_use + test_rate
## Model 2: case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop) +
##     mob_TS
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1     48 49337332
## 2     45 23675518  3  25661814 16.258 2.675e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coefficient significance (t-test)

Under a significance level of 0.05, we can accept the alternative hypotheses: $H_{a1} : \beta_1 \neq 0$, $H_{a2} : \beta_2 \neq 0$, $H_{a3} : \beta_3 \neq 0$, and $H_{a4} : \beta_4 \neq 0$, which means 4 out of 5 of our covariates do explain part of the variability observed in the **case_rate**.

Our estimate for β_1 (the coefficient of our variable of interest) is $\tilde{\beta}_1 = -961.4$, with a standard error of 240.0 and a p-value of 0.0002. It continues to be statistically significant and with an estimated value that changed little from **Model 1** (-990.5) to **Model 2** (-961.4).

```r
coeftest(model_2, vcovHC = vcovHC(model_2, type = "HC3"))
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)    -3.9998e+03  1.3709e+03 -2.9176 0.0054877 **
## mask_useTRUE   -9.6137e+02  2.3881e+02 -4.0257 0.0002153 ***
## test_rate       2.4053e-02  5.2235e-03  4.6048 3.386e-05 ***
## age_below_25    1.9180e+02  4.0512e+01  4.7344 2.215e-05 ***
## log(black_pop)  2.2119e+02  6.1183e+01  3.6153 0.0007542 ***
## mob_TS          1.6017e+01  7.3530e+00  2.1783 0.0346619 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical significance

According to Model 2, states that have adopted mandatory mask use would expect to have -961.4 covid cases per 100,000 habitants. Given that the median of covid case rate among US states is of 2,633 per 100,000 habitants, with 1st Quartile = 2,040 and 3rd quartile = 3,516, the coefficient estimate has practical significance, with an effect size corresponding to a reduction of 36.5% of the median of the covid case rate among states.

Model 3

Model description

Our third regression model has **Covid-19 Case Rate per 100,000 habitants** as our outcome variable and seven covariates: our variable of interest (**Mandatory Mask Use**), **Test Rate per 100,000 habitants**,

Percentage of Population Below 25 Years Old, **Log of Percentage of Black Ethnicity in Total Population**, **Human Mobility Change in Transit Stations**, **Number of Days of Shelter in Place**, and **Number of Days of Non-Essential Businesses Closure**.

Model 3 includes the previous covariates, and other new 2 covariates, erring on the side of inclusion. A key purpose of **Model 3** is to demonstrate the robustness of the results of our measurement goal ($\tilde{\beta}_1$). New variables on **Model 3** represent other commom policies US states have adopted to combat the virus spread. They have some collinearity with mask use as it would be expected, since typically a state enact not a single, but a set of policies against covid-19.

Despite the fact that **Model 3** loses some explanatory power due to the inclusion of the new variables, the result we would like to highlight is that our coeficient of interest ($\tilde{\beta}_1$) continued to be both statistically significant, and with an estimated value that has practical significance in terms of informing public policies in the combat to the virus.

Model 3 main role was to work as an *acid test* of the relevance of mandatory mask use in reducing the case rate of covid 19 among US states. And, in this sense, it helped us to confirm the robustness of our study results.

Our Model 3 has the format:

case_rate = β_0 + β_1 * mask_use + β_2 * test_rate + β_3 * age_below_25 + β_4 * log(black_pop) + β_5 * mob_TS + β_6 * shelter_days + β_7 * bus_close_days

Model summary

```
model_3 <- lm(case_rate ~ mask_use + test_rate + age_below_25 + log(black_pop) + mob_TS + shelter_days +
std_errors = list(
  sqrt(diag(vcovHC(model_1))),
  sqrt(diag(vcovHC(model_2))),
  sqrt(diag(vcovHC(model_3)))
)
stargazer(model_1, model_2, model_3, se = std_errors, type = "text", title = "Model 3 Summary")
```

```
##
## Model 3 Summary
## ================================================================================================
##                                                Dependent variable:
##                                 ---------------------------------------------------------------
##                                                       case_rate
##                                       (1)                (2)                      (3)
## ------------------------------------------------------------------------------------------------
## mask_use                          -990.470***        -961.366***              -940.666***
##                                    (324.753)          (240.032)                (284.877)
##
## test_rate                           0.018*             0.024**                  0.023*
##                                    (0.010)            (0.012)                  (0.012)
##
## age_below_25                                          191.802***               190.437***
##                                                       (57.522)                 (54.186)
##
## log(black_pop)                                        221.195***               219.955**
##                                                       (83.120)                 (91.170)
##
```

```
## mob_TS                                          16.017**              15.920*
##                                                 (7.479)               (8.452)
##
## shelter_days                                                           -0.768
##                                                                       (1.764)
##
## bus_close_days                                                         6.819
##                                                                      (12.360)
##
## Constant                  2,530.239***          -3,999.773*          -4,228.707**
##                             (501.044)           (2,133.506)          (1,982.136)
##
## ------------------------------------------------------------------------------------
## Observations                   51                    51                    50
## R2                            0.236                 0.633                 0.633
## Adjusted R2                   0.204                 0.593                 0.572
## Residual Std. Error  1,013.835 (df = 48)    725.343 (df = 45)     723.819 (df = 42)
## F Statistic         7.416*** (df = 2; 48) 15.550*** (df = 5; 45) 10.355*** (df = 7; 42)
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

Overall model significance (F-test)

Under a significance level of 0.05, we can not reject the null hypothesis (**Model 2**) in favor of our fuller **Model 3**, which now includes the covariates **shelter_days** and **bus_close_days**. Our Residual Std. Error almost remained the same, even with the inclusion of two new variables. This demonstrates that there is collinearity between our new variables and the old ones. The inclusion of the new variables did not help to increase the explained variability of the outcome variable. The adjusted R-squared of **Model 3** is of 0.572.

On the other hand, as it was asserted above, our focus of interest on **Model 3** is not on the overall roustness of the model (for that sake we have **Model 2**), but more on performing an *acid test* around the statistical and practical significance of our coeficient of interest ($\tilde{\beta}_1$).

Coefficient significance (t-test)

Under a significance level of 0.05, we can accept the alternative hypotheses: $H_{a1} : \beta_1 \neq 0$, $H_{a3} : \beta_3 \neq 0$, and $H_{a4} : \beta_4 \neq 0$, which means only 3 out of 7 of our covariates do explain part of the variability observed in the **case_rate**.

Our estimate for β_1 (the coefficient of our variable of interest) is $\tilde{\beta}_1 = -940.7$, with a standard error of 284.9 and a p-value of 0.0006. It continues to be statistically significant and with an estimated value that changed little from **Model 2** (-961.4) to **Model 3** (-940.7).

```r
coeftest(model_3, vcovHC = vcovHC(model_3, type = "HC3"))
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)    -4.2287e+03  1.4726e+03 -2.8716 0.0063744 **
## mask_useTRUE   -9.4067e+02  2.5302e+02 -3.7177 0.0005890 ***
## test_rate       2.3496e-02  5.3170e-03  4.4190 6.848e-05 ***
## age_below_25    1.9044e+02  4.1432e+01  4.5963 3.903e-05 ***
```

```
## log(black_pop)  2.1995e+02  6.1515e+01  3.5756 0.0008954 ***
## mob_TS          1.5920e+01  7.9817e+00  1.9945 0.0526119 .
## shelter_days   -7.6843e-01  2.7989e+00 -0.2745 0.7850081
## bus_close_days  6.8192e+00  9.0098e+00  0.7569 0.4533532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical significance

According to Model 3, states that have adopted mandatory mask use would expect to have -940.7 covid cases per 100,000 habitants. Given that the median of covid case rate among US states is of 2,633 per 100,000 habitants, with 1st Quartile = 2,040 and 3rd quartile = 3,516, the coefficient estimate has practical significance, with an effect size corresponding to a reduction of 35.7% of the median of the covid case rate among states.