w203_lab2: regression models draft

```
schema <- cols(
  state = "c",
  cases_total = "i",
  cases_last_7_days = "i",
  case_rate = "n",
  case_rate_last_7_days = "n",
  deaths_total = "i",
  deaths_last_7_days = "i",
  death_rate = "n",
  death_rate_last_7_days = "n",
  tests_total = "i",
  tests_positive = col_factor(
    levels = c("0-5%", "6-10%", "11-20%"),
    ordered = TRUE
    ),
  test_rate = "i",
  white_cases = "i",
  white_pop = "i",
  black_cases = "i",
  black_pop = "i",
  hispanic_cases = "i",
  hispanic_pop = "i",
  other_cases = "i",
  other_pop = "i",
  white_deaths = "i",
  black_deaths = "i",
  hispanic_deaths = "i",
  other_deaths = "i",
  emerg_date = col_date(format = "%d/%m/%Y"),
  beg_bus_close_date = col_date(format = "%d/%m/%Y"),
  end_bus_close_date = col_date(format = "%d/%m/%Y"),
  bus_close_days = "i",
  beg_shelter_date = col_date(format = "%d/%m/%Y"),
  end_shelter_date = col_date(format = "%d/%m/%Y"),
  shelter_days = "i",
  mask_date = col_date(format = "%d/%m/%Y"),
  mask_use = "l",
  mask_legal = "l",
  beg_maskbus_date = col_date(format = "%d/%m/%Y"),
  end_maskbus_date = col_date(format = "%d/%m/%Y"),
  maskbus_use = "l",
  gov_party = col_factor(
    levels = c("R", "D"),
    ordered = FALSE
  ),
  pop_dens = "n",
```

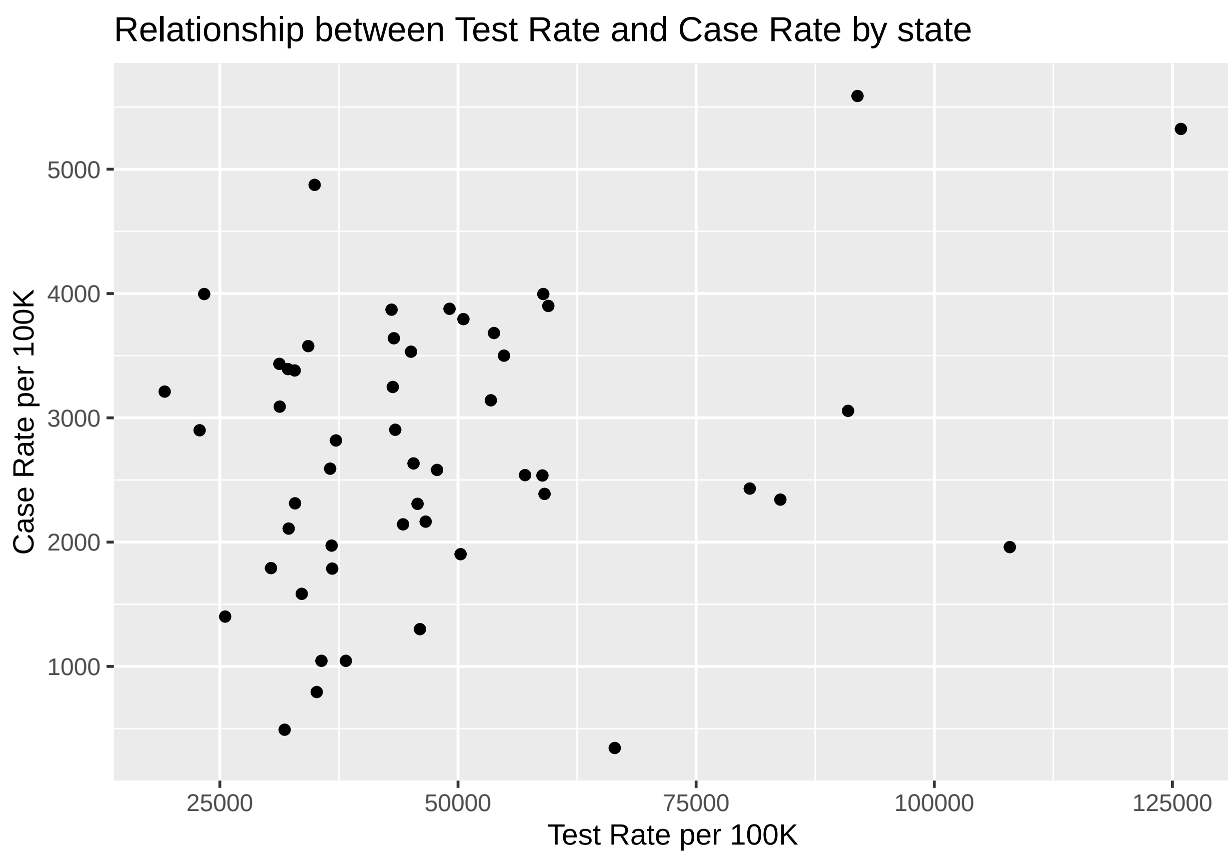
```r
  pop_total = "i",
  pre_cond_total = "i",
  serious_illness_pct = "n",
  all_cause_deaths_total = "i",
  homeless_total = "i",
  medicaid_pct = "i",
  life_expectancy = "n",
  unemployment_rate = "n",
  poverty_rate = "n",
  weekly_UI_max_amount = "i",
  household_income = "i",
  age_0_18 = "i",
  age_19_25 = "i",
  age_26_34 = "i",
  age_35_54 = "i",
  age_55_64 = "i",
  age_65 = "i",
  mob_RR = "i",
  mob_GP = "i",
  mob_PK = "i",
  mob_TS = "i",
  mob_WP = "i",
  mob_RS = "i"
)
```

```r
df <- read_delim(
  file = "clean_covid_19_LB_version.csv",
  delim = ";",
  col_names = TRUE,
  col_types = schema,
  na = ""
  )
```

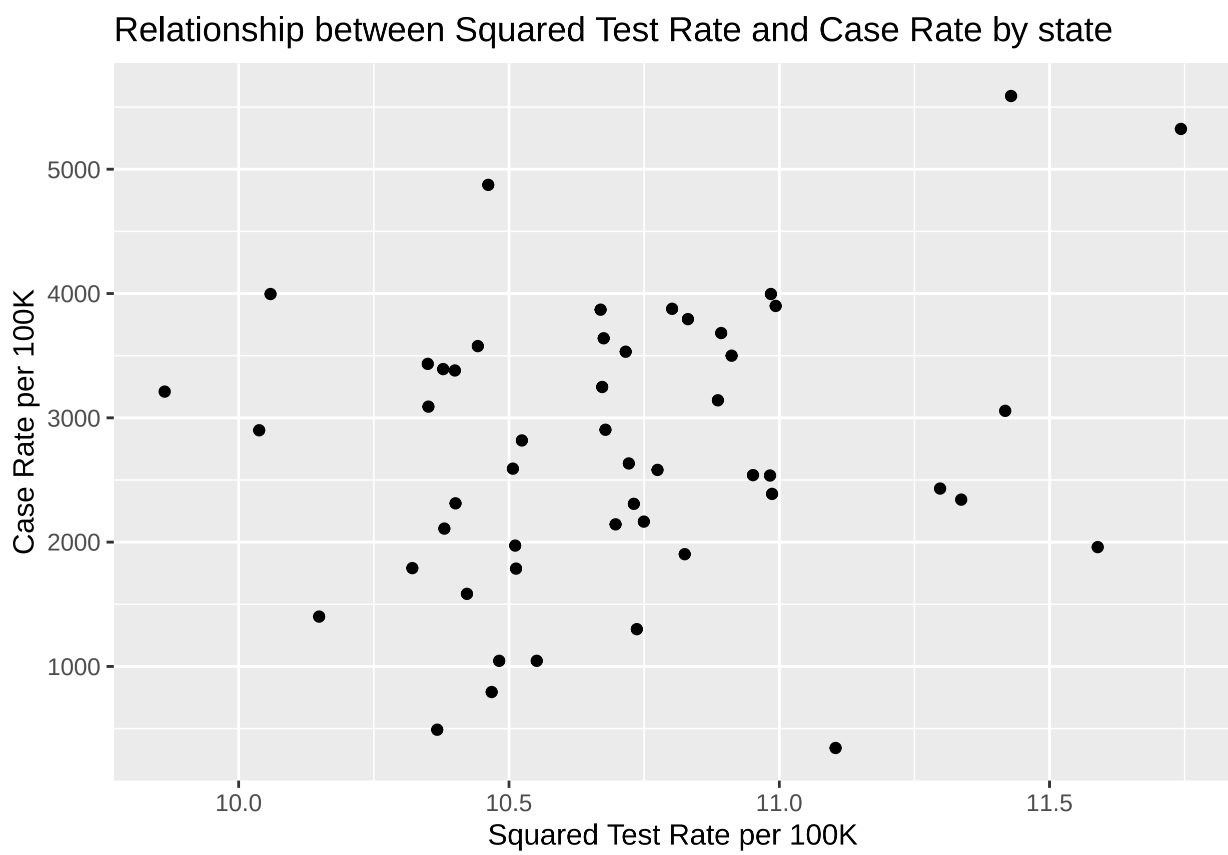
Question: Should we include test_rate (or any transformation of it) as an initial variable on our model?

Answer: Yes, we should include test_rate on our initial model version with no transformation

```r
plot1 <- 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"
  )
plot1
```




```
plot2 <- df %>%
  ggplot(aes(y = case_rate, x = log(test_rate))) +
  geom_point() +
  labs(
    title = "Relationship between Squared Test Rate and Case Rate by state",
    x = "Squared Test Rate per 100K",
    y = "Case Rate per 100K"
  )
plot2
```




```r
df$sqrt_test_rate = df$test_rate^2
```

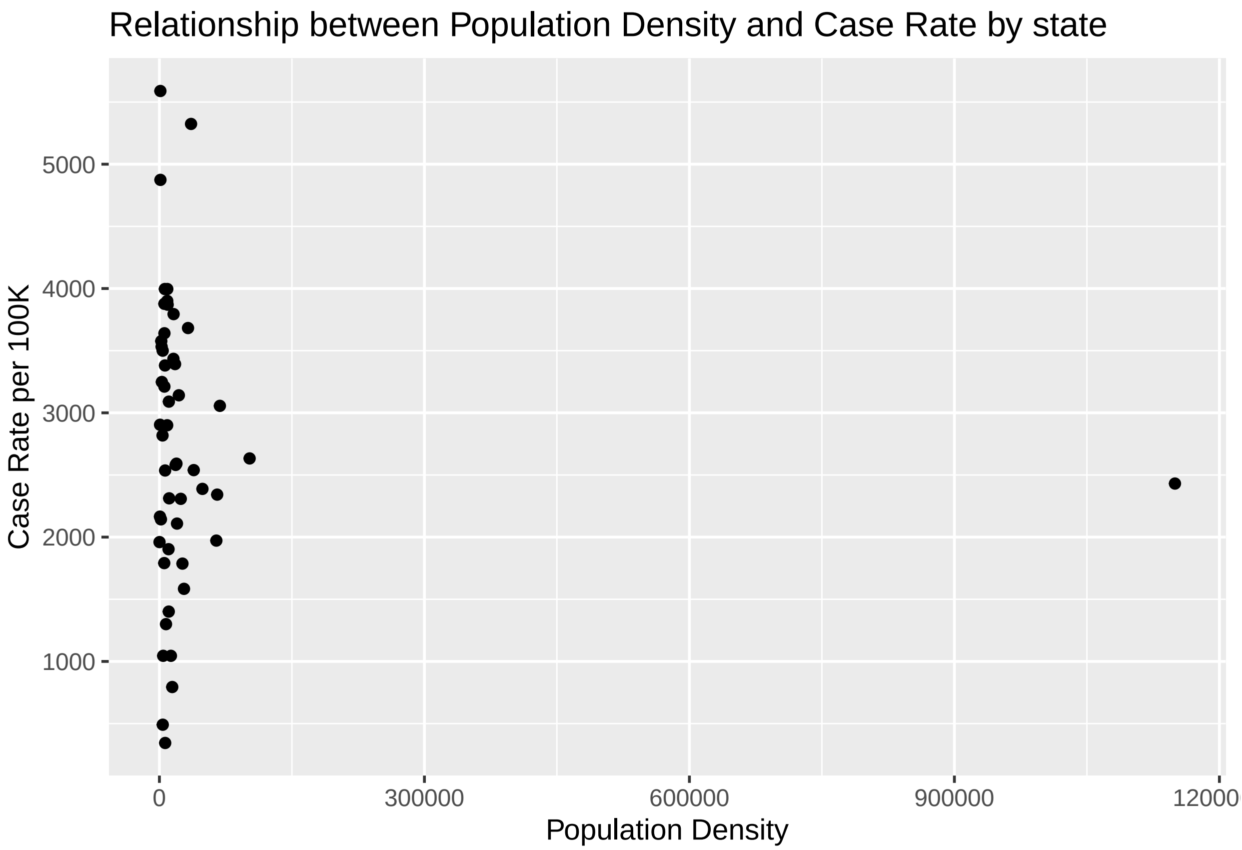
```r
mod1_1 <- lm(case_rate ~
               mask_use,
             data = df
             )

mod1_2 <- lm(case_rate ~
               mask_use +
               test_rate,
             data = df
             )

mod1_3 <- lm(case_rate ~
               mask_use +
               log(test_rate),
             data = df
             )

std_errors = list(
  sqrt(diag(vcovHC(mod1_1))),
  sqrt(diag(vcovHC(mod1_2))),
  sqrt(diag(vcovHC(mod1_3)))
  )

stargazer(mod1_1, mod1_2, mod1_3, se = std_errors, type = "text")
```

```
##
## ===================================================================================
##                                          Dependent variable:
##                     ---------------------------------------------------------------
##                                               case_rate
##                              (1)                    (2)                    (3)
## -----------------------------------------------------------------------------------
## mask_use                 -830.000**            -990.470***            -983.274***
##                          (343.609)              (324.753)              (337.720)
##
## test_rate                                        0.018*
##                                                 (0.010)
##
## log(test_rate)                                                         893.463*
##                                                                        (523.727)
##
## Constant                3,302.765***          2,530.239***            -6,155.628
##                          (293.589)              (501.044)             (5,533.903)
##
## -----------------------------------------------------------------------------------
## Observations                 51                     51                     51
## R2                         0.121                  0.236                  0.211
## Adjusted R2                0.103                  0.204                  0.178
## Residual Std. Error 1,076.417 (df = 49)   1,013.835 (df = 48)   1,030.367 (df = 48)
## F Statistic         6.738** (df = 1; 49) 7.416*** (df = 2; 48) 6.416*** (df = 2; 48)
## ===================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

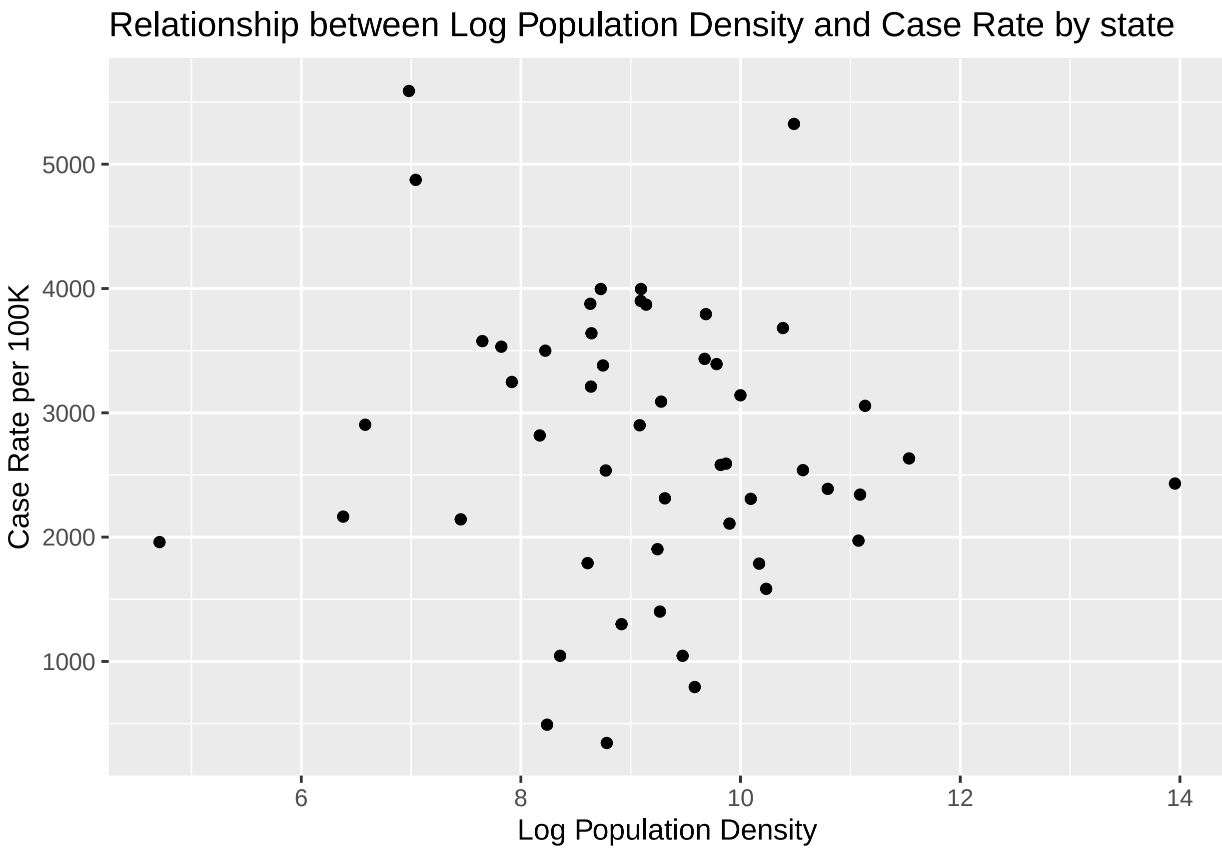
Question: Should we include pop_dens as an another initial variable on our model on top of test_rate?

Answer: No, we should not add pop_dens to our regression model

```
plot3 <- df %>%
  ggplot(aes(y = case_rate, x = pop_dens)) +
  geom_point() +
  labs(
    title = "Relationship between Population Density and Case Rate by state",
    x = "Population Density",
    y = "Case Rate per 100K"
  )
plot3
```




```r
plot4 <- df %>%
  ggplot(aes(y = case_rate, x = log(pop_dens))) +
  geom_point() +
  labs(
    title = "Relationship between Log Population Density and Case Rate by state",
    x = "Log Population Density",
    y = "Case Rate per 100K"
  )
plot4
```




```
mod2_1 <- lm (case_rate ~
                mask_use +
                test_rate,
              data = df
                )

mod2_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                pop_dens,
              data = df
                )

mod2_3 <- lm (case_rate ~
                mask_use +
                test_rate +
                log(pop_dens),
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod2_1))),
  sqrt(diag(vcovHC(mod2_2))),
  sqrt(diag(vcovHC(mod2_3)))
  )
```

```r
stargazer(mod2_1, mod2_2, mod2_3, se = std_errors, type = "text")
```

```
##
## ========================================================================================
##                                            Dependent variable:
##                       ------------------------------------------------------------------
##                                                 case_rate
##                               (1)                    (2)                    (3)
## ----------------------------------------------------------------------------------------
## mask_use                  -990.470***            -972.696***            -984.312***
##                            (324.753)              (325.461)              (370.777)
##
## test_rate                   0.018*                 0.019*                 0.018
##                            (0.010)                (0.011)                (0.012)
##
## pop_dens                                          -0.001
##                                                   (0.002)
##
## log(pop_dens)                                                            -6.833
##                                                                         (152.039)
##
## Constant                 2,530.239***           2,493.487***            2,588.169**
##                            (501.044)              (521.334)             (1,200.805)
##
## ----------------------------------------------------------------------------------------
## Observations                  51                     51                     51
## R2                          0.236                  0.242                  0.236
## Adjusted R2                 0.204                  0.194                  0.187
## Residual Std. Error  1,013.835 (df = 48)    1,020.343 (df = 47)    1,024.514 (df = 47)
## F Statistic         7.416*** (df = 2; 48) 5.011*** (df = 3; 47) 4.843*** (df = 3; 47)
## ========================================================================================
## Note:                                                       *p<0.1; **p<0.05; ***p<0.01
```

Question: Should we include any variable to control for age demographics? If yes, which variable does the better job in improving our model explanability?

Answer: Yes, we should include age_below_25

```r
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
```

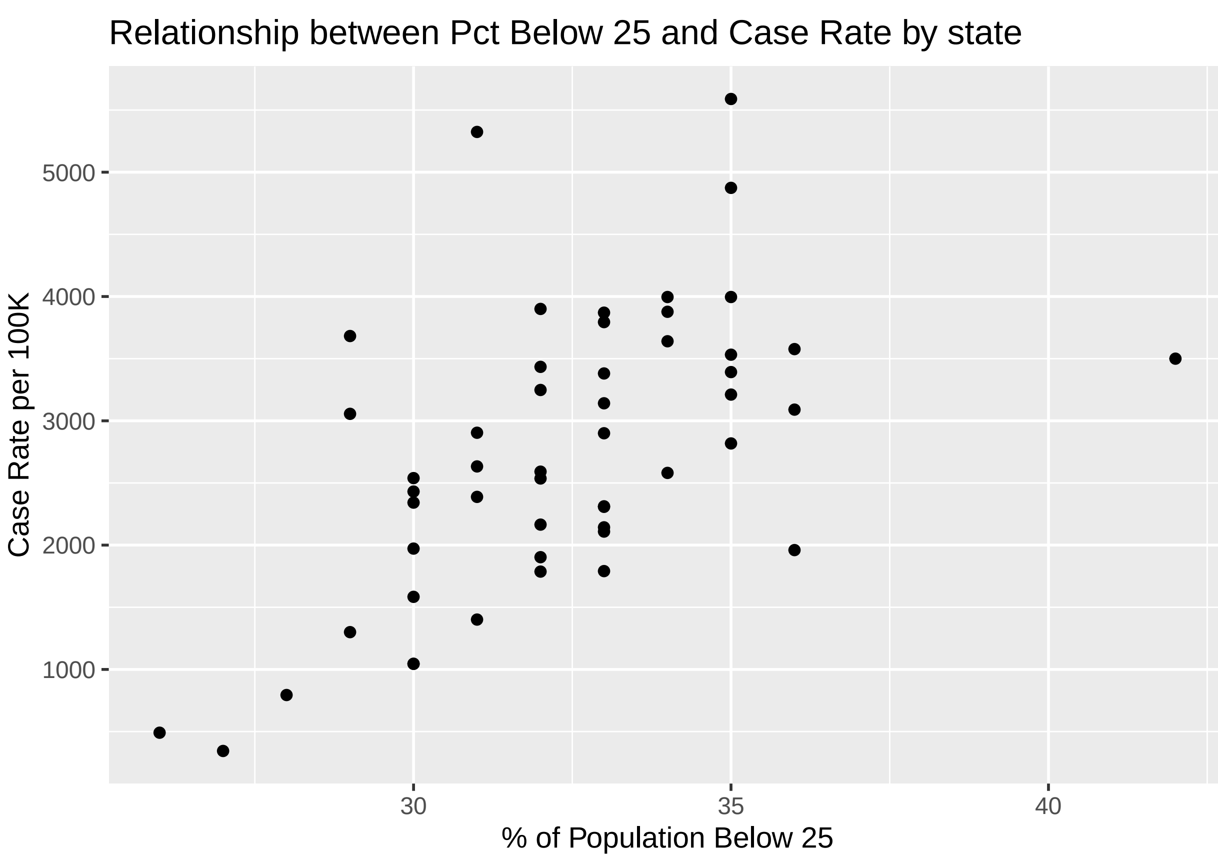
```
## age_35_54    0.05333333   -0.440000
## age_55_64    1.53019608    1.872941
## age_65       1.87294118    4.294118
```

```
df$age_below_25 = df$age_0_18 + df$age_19_25
df$age_above_55 = df$age_55_64 + df$age_65
```

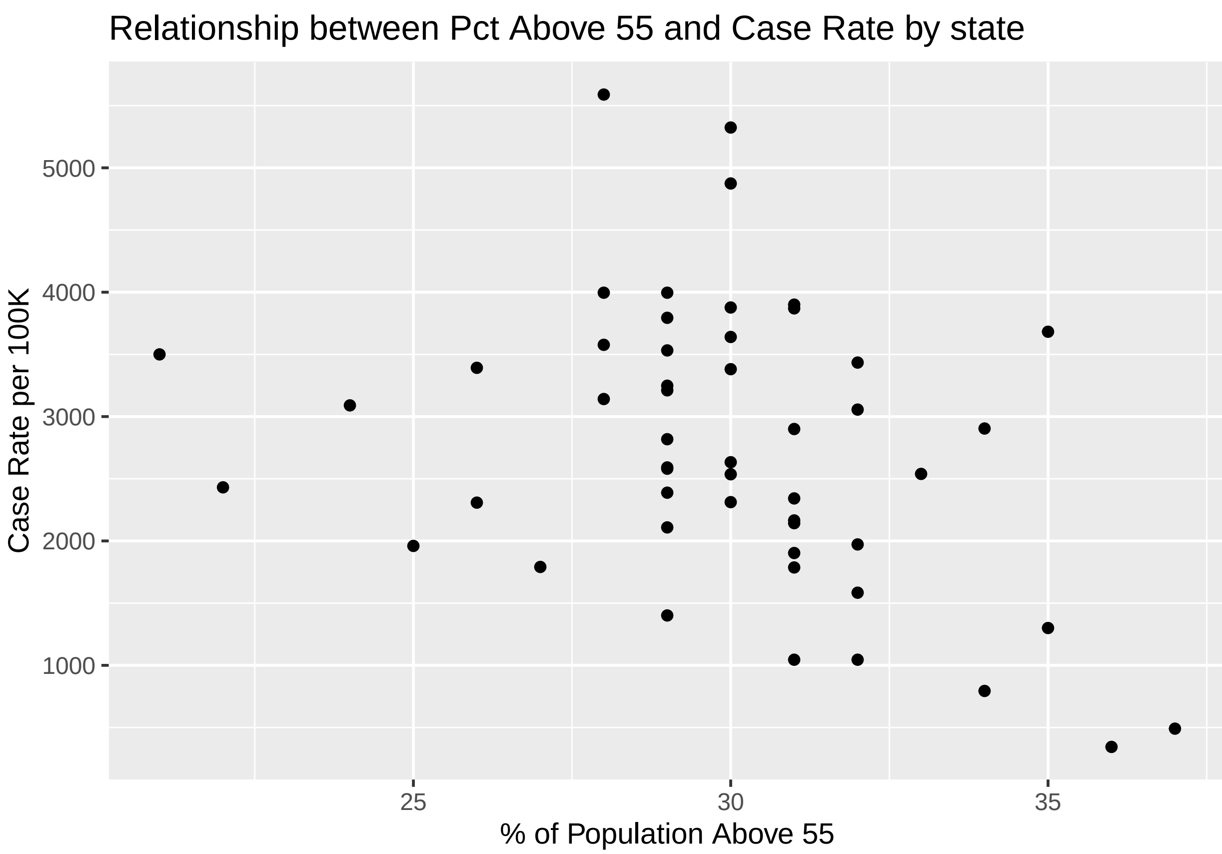
```
var(df[ ,c(4, 64:65)], na.rm=TRUE)
```

```
##                 case_rate age_below_25 age_above_55
## case_rate     1291651.530  1713.384706 -1218.636863
## age_below_25     1713.385     7.352941    -6.324706
## age_above_55    -1218.637    -6.324706     9.570196
```

```
plot5 <- 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"
  )
plot5
```



```r
plot6 <- df %>%
  ggplot(aes(y = case_rate, x = age_above_55)) +
  geom_point() +
  labs(
    title = "Relationship between Pct Above 55 and Case Rate by state",
    x = "% of Population Above 55",
    y = "Case Rate per 100K"
  )
plot6
```




```r
mod3_1 <- lm (case_rate ~
                mask_use +
                test_rate,
              data = df
                )

mod3_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25,
              data = df
                )

mod3_3 <- lm (case_rate ~
                mask_use +
                test_rate +
```

```r
                age_above_55,
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod3_1))),
  sqrt(diag(vcovHC(mod3_2))),
  sqrt(diag(vcovHC(mod3_3)))
  )

stargazer(mod3_1, mod3_2, mod3_3, se = std_errors, type = "text")
```

```
##
## ==========================================================================================
##                                          Dependent variable:
##                       --------------------------------------------------------------------
##                                               case_rate
##                              (1)                   (2)                    (3)
## ------------------------------------------------------------------------------------------
## mask_use                 -990.470***          -806.717***          -1,059.154***
##                           (324.753)            (265.274)              (286.999)
##
## test_rate                  0.018*               0.020*                  0.016
##                            (0.010)              (0.012)                (0.012)
##
## age_below_25                                  224.736***
##                                                (76.416)
##
## age_above_55                                                        -129.544**
##                                                                      (52.140)
##
## Constant                 2,530.239***         -4,942.849*            6,553.500***
##                           (501.044)           (2,790.278)            (1,555.372)
##
## ------------------------------------------------------------------------------------------
## Observations                 51                   51                     51
## R2                          0.236                0.516                  0.357
## Adjusted R2                 0.204                0.485                  0.316
## Residual Std. Error  1,013.835 (df = 48)    815.744 (df = 47)      939.646 (df = 47)
## F Statistic         7.416*** (df = 2; 48) 16.684*** (df = 3; 47) 8.715*** (df = 3; 47)
## ==========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

Question: Should we include any variable to control for socio-economic differences among states? If yes, which variable does the better job in improving our model explanability?

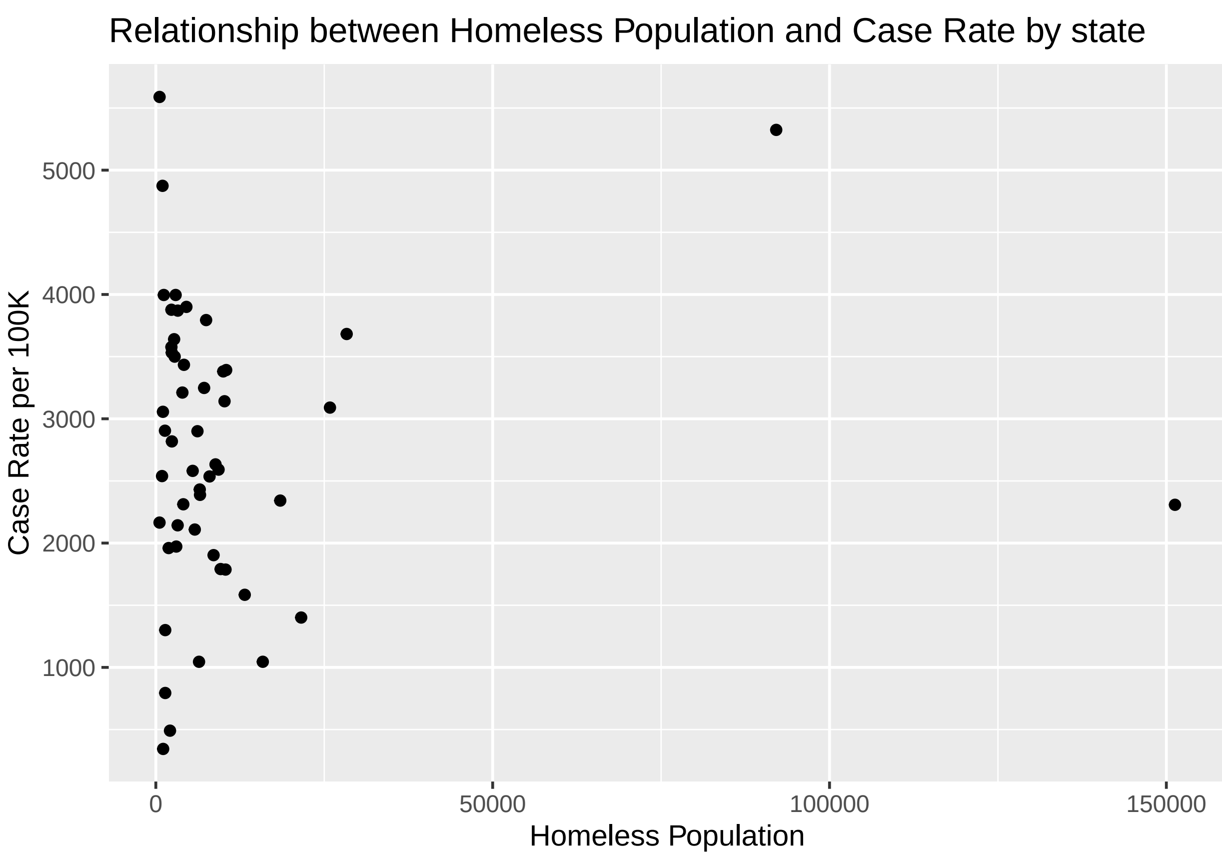
Answer: No, we should not include any variable to control for socio-economic differences. Poverty_rate could be an option, but it has high collinearity with black_pop. And at the final model black_pop does a better job than poverty_rate.

```r
var(df[ ,c(4, 44, 47, 48, 50)], na.rm = TRUE)
```

```
##                        case_rate homeless_total unemployment_rate poverty_rate
```

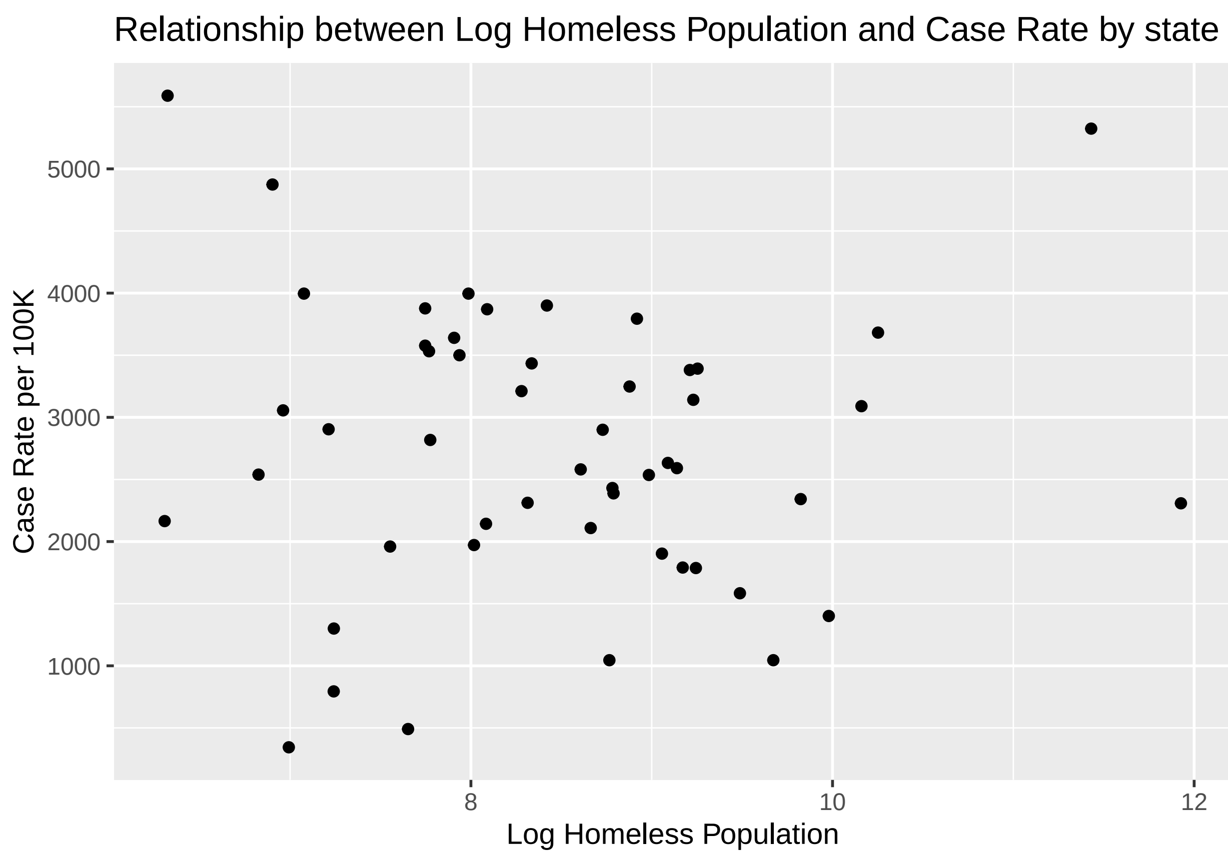
```
## case_rate           1310901.7404    2298722.238          -534.4865     8944.9616
## homeless_total      2298722.2384  593414662.410        32545.7747    -7124.5159
## unemployment_rate      -534.4865      32545.775          113.7861      184.7886
## poverty_rate           8944.9616      -7124.516          184.7886      816.6220
## household_income   -3030031.8669    3380857.778        -2281.5012  -16739.5698
##                    household_income
## case_rate              -3030031.867
## homeless_total          3380857.778
## unemployment_rate         -2281.501
## poverty_rate             -16739.570
## household_income      104263090.902
```

```
plot7 <- df %>%
  ggplot(aes(y = case_rate, x = homeless_total)) +
  geom_point() +
  labs(
    title = "Relationship between Homeless Population and Case Rate by state",
    x = "Homeless Population",
    y = "Case Rate per 100K"
  )
plot7
```

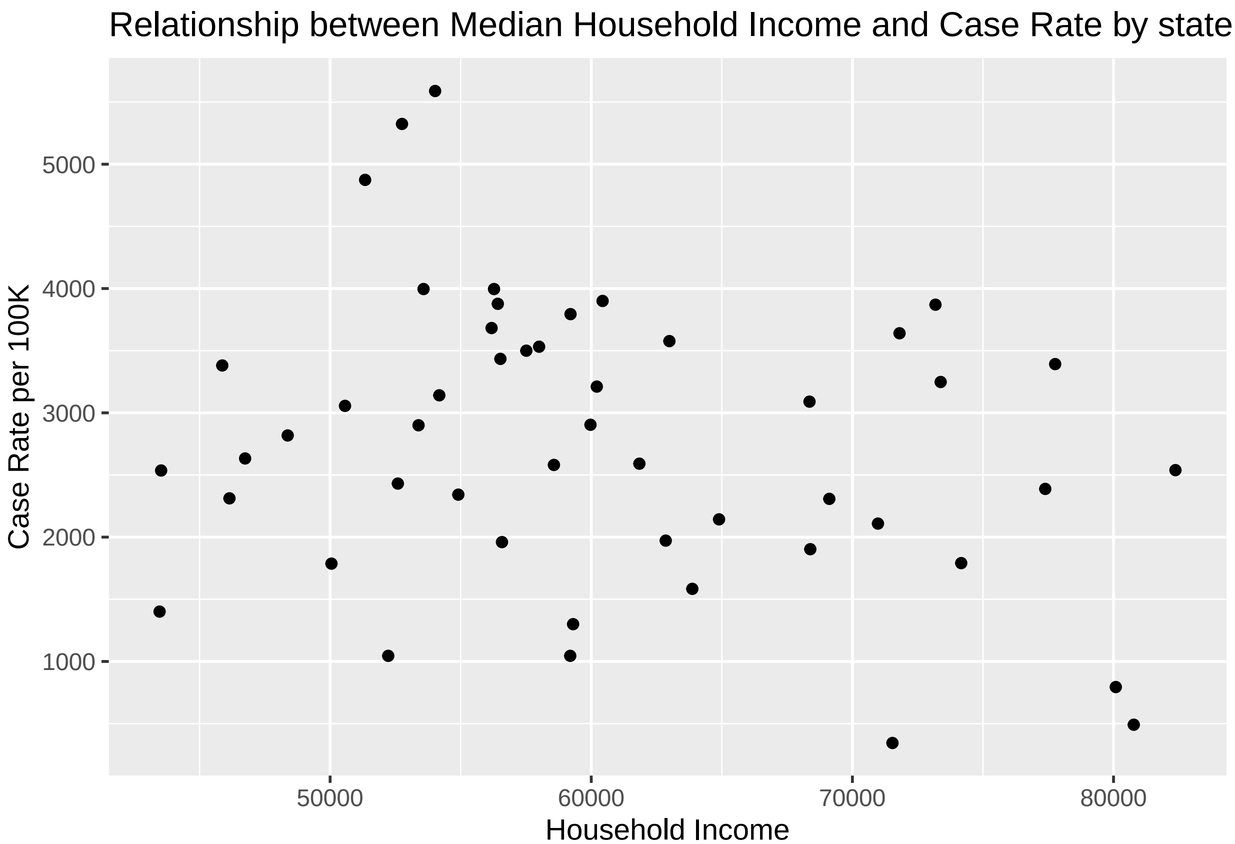



```
plot8 <- df %>%
  ggplot(aes(y = case_rate, x = log(homeless_total))) +
  geom_point() +
  labs(
```

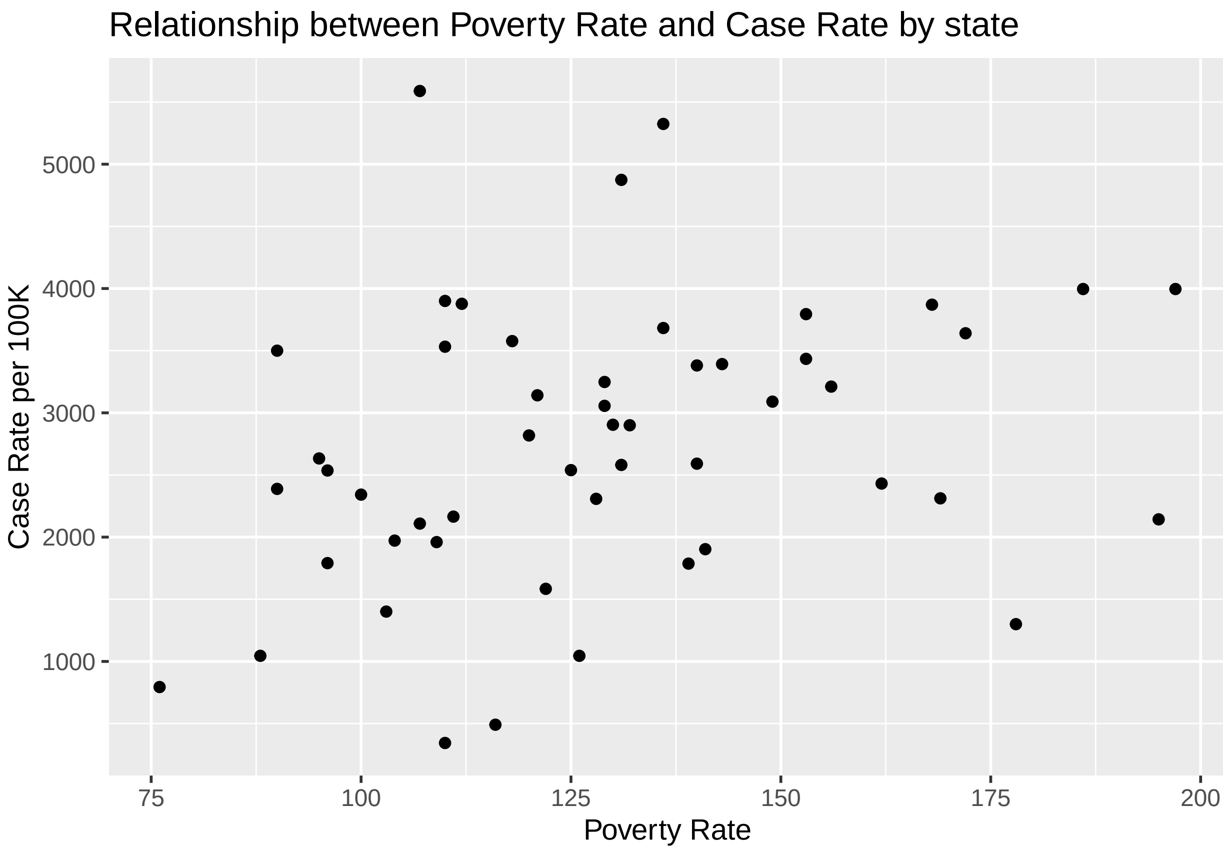
```
    title = "Relationship between Log Homeless Population and Case Rate by state",
    x = "Log Homeless Population",
    y = "Case Rate per 100K"
  )
plot8
```




```
plot9 <- df %>%
  ggplot(aes(y = case_rate, x = household_income)) +
  geom_point() +
  labs(
    title = "Relationship between Median Household Income and Case Rate by state",
    x = "Household Income",
    y = "Case Rate per 100K"
  )
plot9
```




```
plot10 <- df %>%
  ggplot(aes(y = case_rate, x = poverty_rate)) +
  geom_point() +
  labs(
    title = "Relationship between Poverty Rate and Case Rate by state",
    x = "Poverty Rate",
    y = "Case Rate per 100K"
  )
plot10
```




```
mod4_1 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25,
              data = df
                )

mod4_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                log(homeless_total),
              data = df
                )

mod4_3 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                household_income,
              data = df
                )

mod4_4 <- lm (case_rate ~
                mask_use +
                test_rate +
```

```r
                age_below_25 +
                poverty_rate,
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod4_1))),
  sqrt(diag(vcovHC(mod4_2))),
  sqrt(diag(vcovHC(mod4_3))),
  sqrt(diag(vcovHC(mod4_4)))
  )

stargazer(mod4_1, mod4_2, mod4_3, mod4_4, se = std_errors, type = "text")
```

```
##
## ==============================================================================================
##                                                         Dependent variable:
##                       ------------------------------------------------------------------------
##                                                              case_rate
##                              (1)                    (2)                    (3)                    (4
## ----------------------------------------------------------------------------------------------
## mask_use                -806.717***            -857.694***            -854.375***            -849.6
##                          (265.274)              (300.554)              (267.011)              (226.
##
## test_rate                 0.020*                 0.020                  0.019                  0.02
##                           (0.012)                (0.012)                (0.012)                (0.0
##
## age_below_25             224.736***             225.388***             216.980***             210.0
##                           (76.416)               (74.665)               (76.178)               (61.
##
## log(homeless_total)                              81.191
##                                                 (154.671)
##
## household_income                                                       -0.007
##                                                                        (0.011)
##
## poverty_rate                                                                                  11.0
##                                                                                               (4.5
##
## Constant                -4,942.849*            -5,625.272**           -4,165.459             -5,958.
##                         (2,790.278)            (2,718.923)            (2,981.378)            (2,177
##
## ----------------------------------------------------------------------------------------------
## Observations                 51                     51                     50                     5
## R2                          0.516                  0.522                  0.534                  0.5
## Adjusted R2                 0.485                  0.481                  0.493                  0.5
## Residual Std. Error   815.744 (df = 47)      818.951 (df = 46)      815.392 (df = 45)      759.470 (
## F Statistic         16.684*** (df = 3; 47) 12.573*** (df = 4; 46) 12.903*** (df = 4; 45) 16.492*** (
## ==============================================================================================
## Note:                                                                            *p<0.1; **p<0.05
```

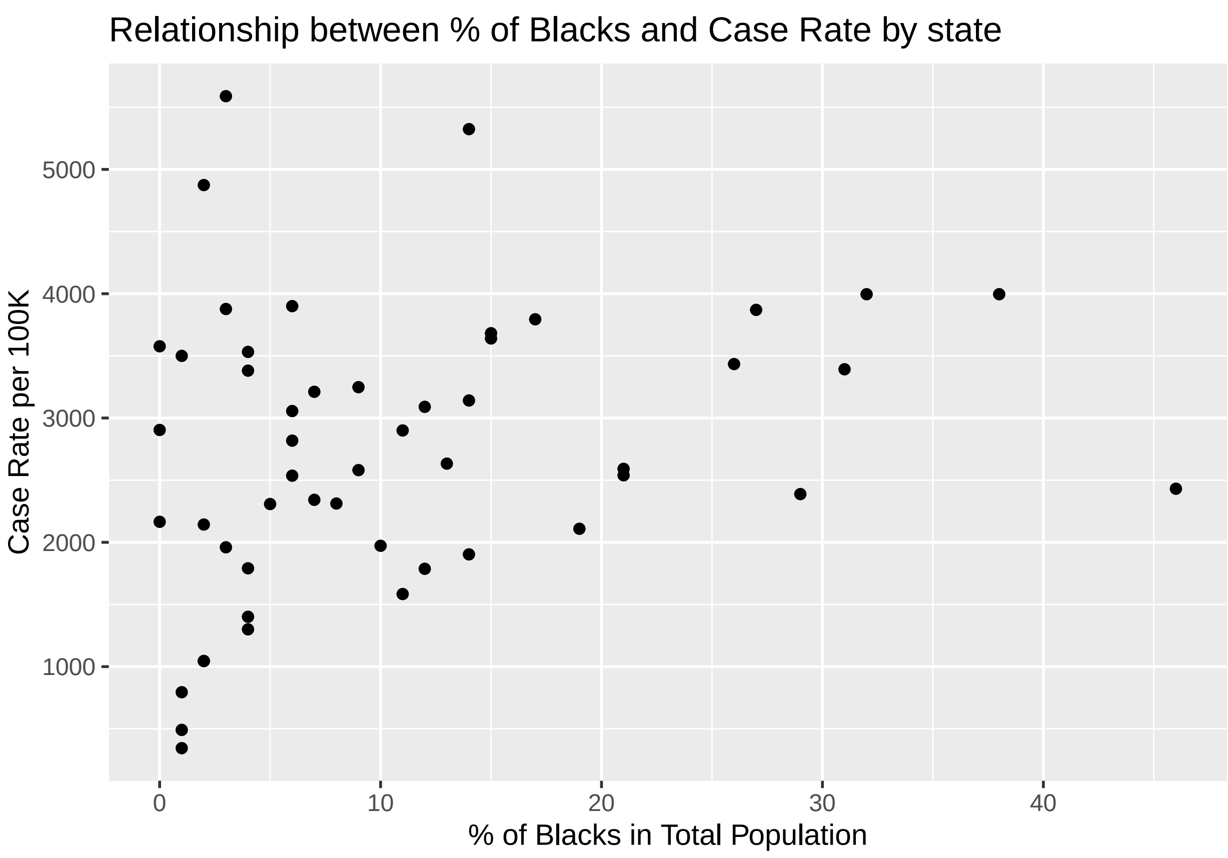
Question: Should we include any variable to control for race mix differences among states? If yes, which variable does the better job in improving our model explanability?

Answer: Yes, we should include the log(black_pop)

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

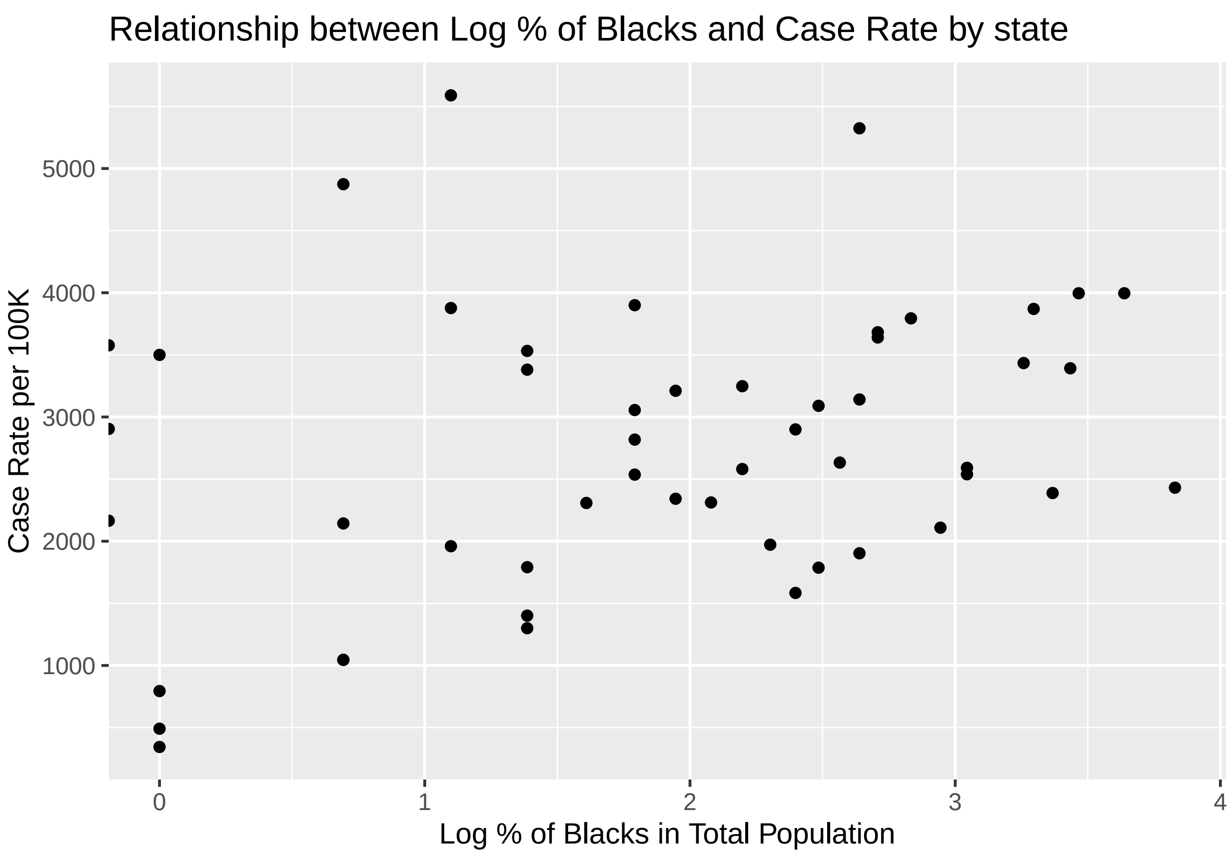
```
##                  case_rate   white_pop   black_pop hispanic_pop
## case_rate     1291651.5302 -1410.79725 3061.76588    612.26275
## white_pop       -1410.7973   292.47843  -76.75765   -116.18157
## black_pop        3061.7659   -76.75765  113.49647    -14.41765
## hispanic_pop      612.2627  -116.18157  -14.41765    108.31843
```

```r
plot11 <- df %>%
  ggplot(aes(y = case_rate, x = black_pop)) +
  geom_point() +
  labs(
    title = "Relationship between % of Blacks and Case Rate by state",
    x = "% of Blacks in Total Population",
    y = "Case Rate per 100K"
  )
plot11
```




```r
plot12 <- df %>%
  ggplot(aes(y = case_rate, x = log(black_pop))) +
  geom_point() +
  labs(
    title = "Relationship between Log % of Blacks and Case Rate by state",
```

```
    x = "Log % of Blacks in Total Population",
    y = "Case Rate per 100K"
  )
plot12
```




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

```
mod5_1 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25,
              data = df
                )

mod5_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                black_pop,
              data = df
                )

mod5_3 <- lm (case_rate ~
                mask_use +
                test_rate +
```

```
                age_below_25 +
                log(black_pop),
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod5_1))),
  sqrt(diag(vcovHC(mod5_2))),
  sqrt(diag(vcovHC(mod5_3)))
  )

stargazer(mod5_1, mod5_2, mod5_3, se = std_errors, type = "text")
```

```
##
## ==========================================================================================
##                                              Dependent variable:
##                      ---------------------------------------------------------------------
##                                                   case_rate
##                               (1)                    (2)                    (3)
## ------------------------------------------------------------------------------------------
## mask_use                 -806.717***            -881.390***          -1,050.513***
##                           (265.274)              (243.679)              (249.837)
##
## test_rate                  0.020*                 0.020*                 0.020*
##                           (0.012)                (0.012)                (0.011)
##
## age_below_25             224.736***             220.471***             220.014***
##                           (76.416)               (64.117)               (63.693)
##
## black_pop                                        29.167**
##                                                  (14.469)
##
## log(black_pop)                                                         182.231**
##                                                                        (71.793)
##
## Constant                 -4,942.849*           -5,068.198**           -4,922.534**
##                          (2,790.278)            (2,316.645)            (2,354.761)
##
## ------------------------------------------------------------------------------------------
## Observations                 51                     51                     51
## R2                          0.516                  0.589                  0.595
## Adjusted R2                 0.485                  0.554                  0.560
## Residual Std. Error   815.744 (df = 47)      759.202 (df = 46)      754.292 (df = 46)
## F Statistic         16.684*** (df = 3; 47) 16.512*** (df = 4; 46) 16.878*** (df = 4; 46)
## ==========================================================================================
## Note:                                                        *p<0.1; **p<0.05; ***p<0.01
```

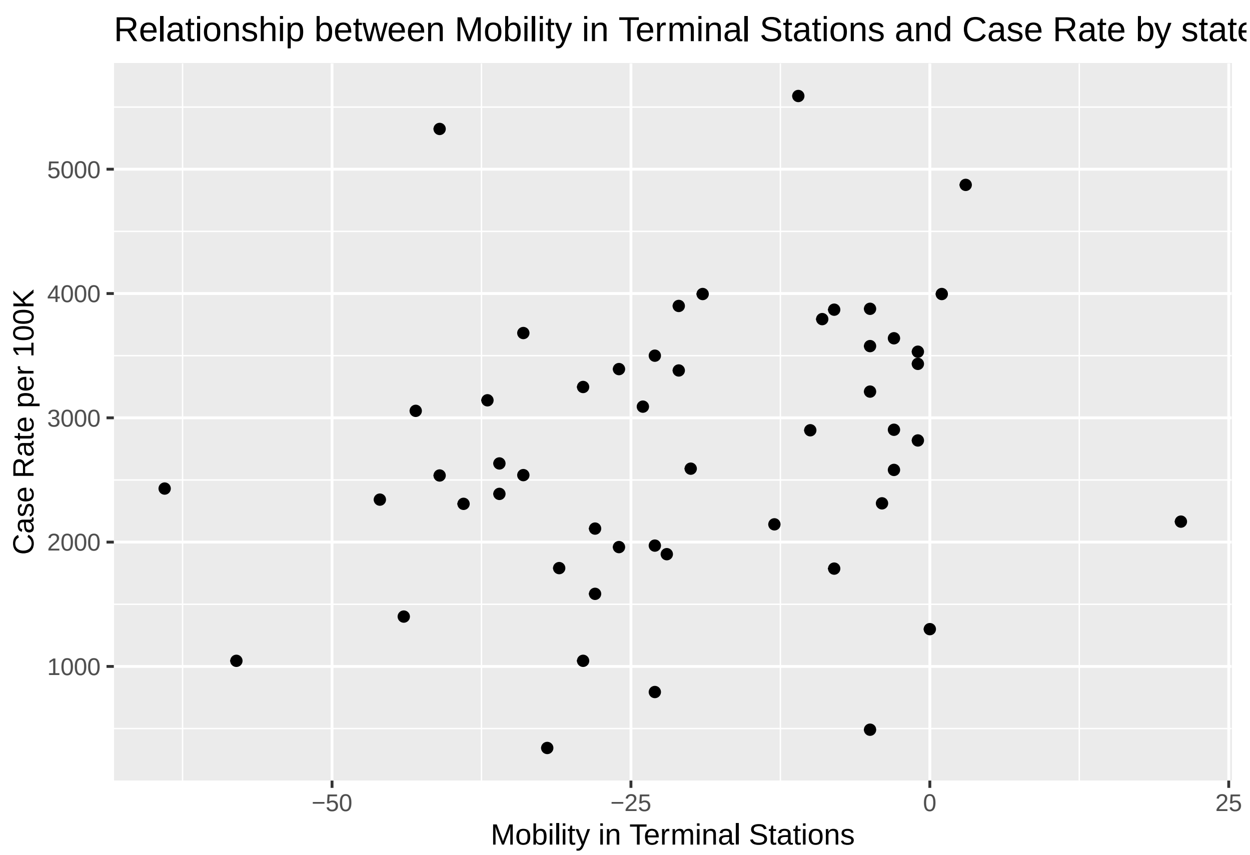
Question: Should we include any indicator from Google mobility? If yes, which variable does the better job in improving our model explanability?

Answer: Yes, we should include the mob_TS variable

```r
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
```

```r
plot13 <- df %>%
  ggplot(aes(y = case_rate, x = mob_TS)) +
  geom_point() +
  labs(
    title = "Relationship between Mobility in Terminal Stations and Case Rate by state",
    x = "Mobility in Terminal Stations",
    y = "Case Rate per 100K"
  )
plot13
```




```
mod6_1 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                log(black_pop),
              data = df
                )

mod6_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                log(black_pop) +
                mob_TS,
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod6_1))),
  sqrt(diag(vcovHC(mod6_2)))
  )

stargazer(mod6_1, mod6_2, se = std_errors, type = "text")
```

```
##
## ================================================================
```

```
##                                  Dependent variable:
##                      --------------------------------------------
##                                       case_rate
##                              (1)                    (2)
## ----------------------------------------------------------------
## mask_use                 -1,050.513***          -961.366***
##                            (249.837)              (240.032)
##
## test_rate                   0.020*                 0.024**
##                            (0.011)                (0.012)
##
## age_below_25              220.014***             191.802***
##                            (63.693)               (57.522)
##
## log(black_pop)            182.231**              221.195***
##                            (71.793)               (83.120)
##
## mob_TS                                            16.017**
##                                                   (7.479)
##
## Constant                 -4,922.534**           -3,999.773*
##                          (2,354.761)            (2,133.506)
##
## ----------------------------------------------------------------
## Observations                  51                     51
## R2                          0.595                  0.633
## Adjusted R2                 0.560                  0.593
## Residual Std. Error   754.292 (df = 46)      725.343 (df = 45)
## F Statistic         16.878*** (df = 4; 46) 15.550*** (df = 5; 45)
## ================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

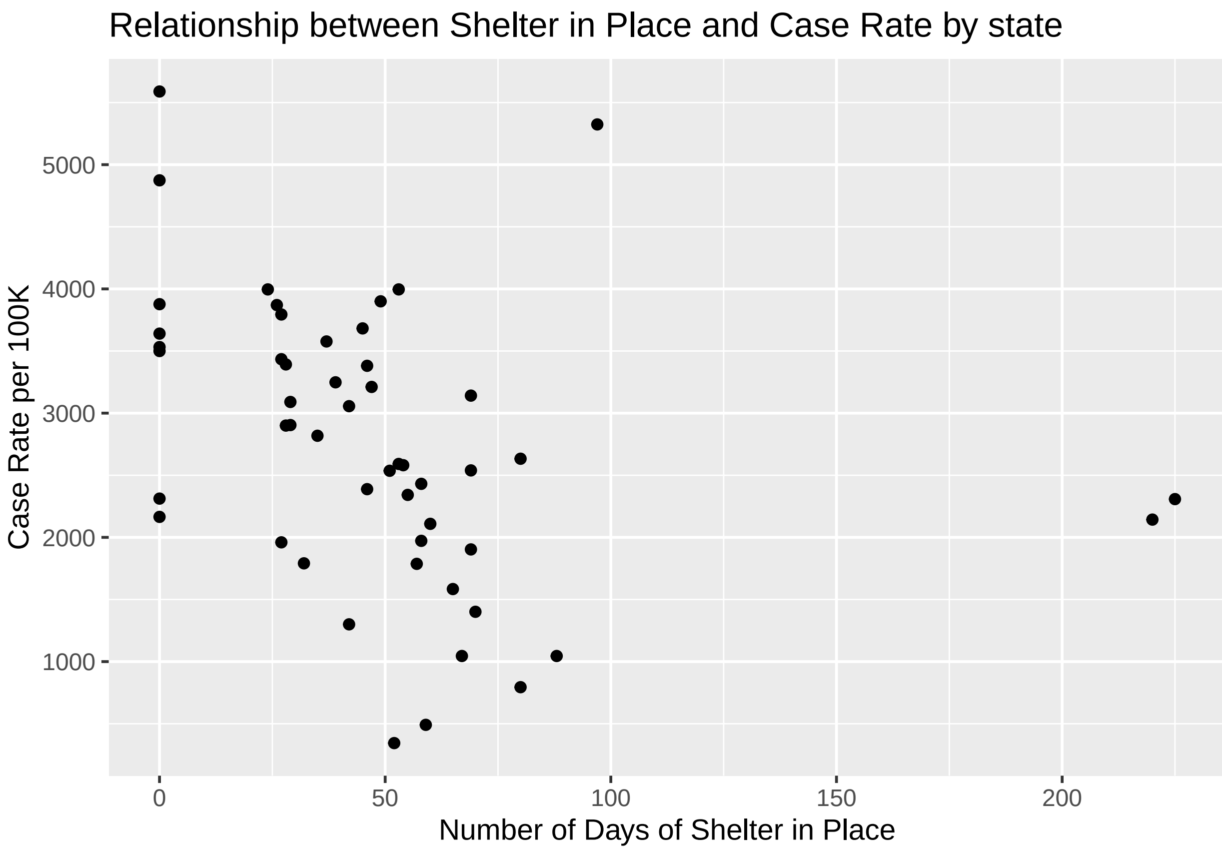
Question: Should we include any other variable related to policies adopted by states? If yes, which variable does the better job on improving our model explanability?

Answer: Yes, we should include shelter_days and bus_close_days just a matter of performing an acid test on the mask_use (see if it continues to be statistically and pratically significant)

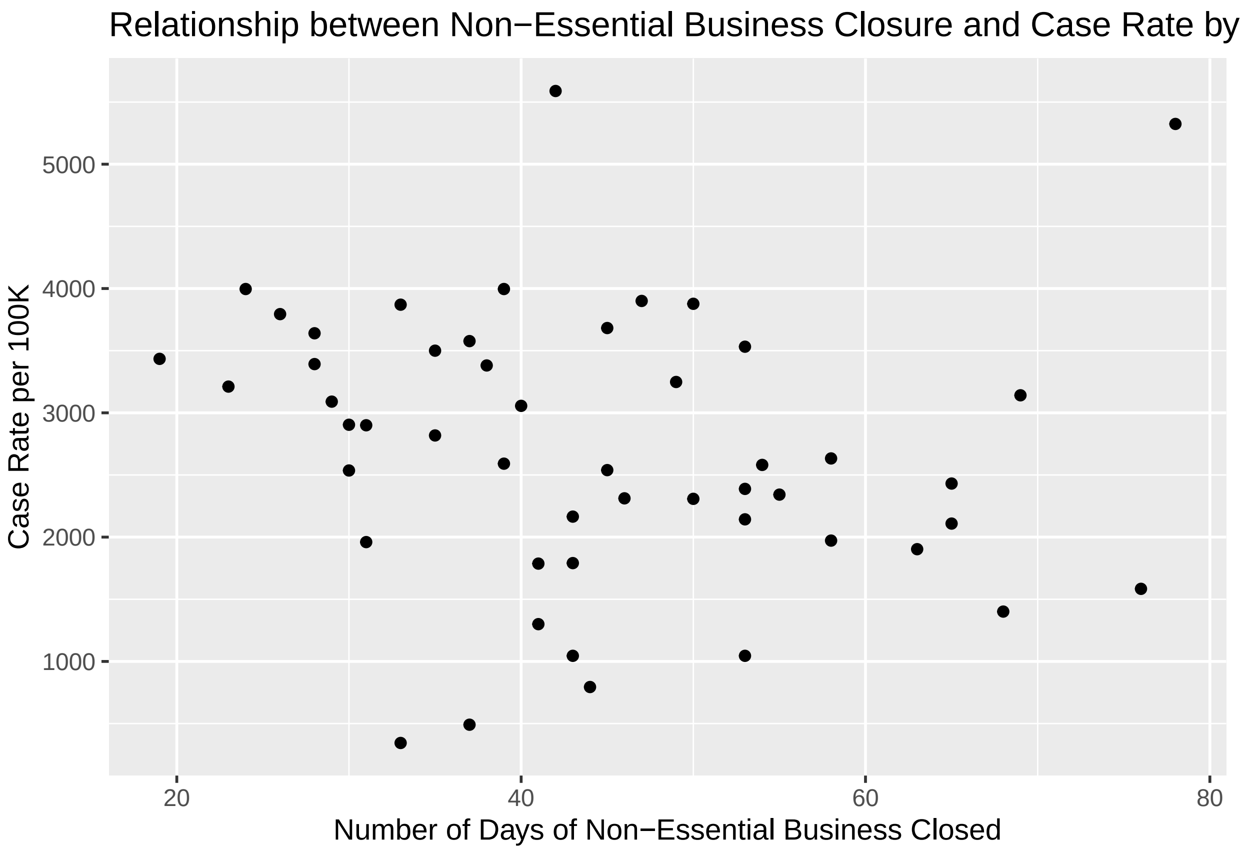
```
var(df[,c(4, 28, 31, 34, 37)], na.rm=TRUE)
```

```
##                    case_rate bus_close_days  shelter_days   mask_legal
## case_rate      1224051.20041  -2318.7163265 -14179.758367 -96.71877551
## bus_close_days   -2318.71633    194.9897959    231.138776  -0.69183673
## shelter_days    -14179.75837    231.1387755   1849.307755   4.39510204
## mask_legal         -96.71878     -0.6918367      4.395102   0.19632653
## maskbus_use        -78.98816      1.5734694      3.795102   0.01673469
##                  maskbus_use
## case_rate       -78.98816327
## bus_close_days    1.57346939
## shelter_days      3.79510204
## mask_legal        0.01673469
## maskbus_use       0.12285714
```

```r
plot14 <- df %>%
  ggplot(aes(y = case_rate, x = shelter_days)) +
  geom_point() +
  labs(
    title = "Relationship between Shelter in Place and Case Rate by state",
    x = "Number of Days of Shelter in Place",
    y = "Case Rate per 100K"
  )
plot14
```




```r
plot15 <- df %>%
  ggplot(aes(y = case_rate, x = bus_close_days)) +
  geom_point() +
  labs(
    title = "Relationship between Non-Essential Business Closure and Case Rate by state",
    x = "Number of Days of Non-Essential Business Closed",
    y = "Case Rate per 100K"
  )
plot15
```




```
mod7_1 <- lm (case_rate ~
                mask_use +
                sqrt_test_rate +
                age_below_25 +
                log(black_pop) +
                mob_TS,
              data = df
                )

mod7_2 <- lm (case_rate ~
                mask_use +
                sqrt_test_rate +
                age_below_25 +
                log(black_pop) +
                mob_TS +
                shelter_days,
              data = df
                )

mod7_3 <- lm (case_rate ~
                mask_use +
                sqrt_test_rate +
                age_below_25 +
                log(black_pop) +
                mob_TS +
                bus_close_days,
```

```r
               data = df
                 )

mod7_4 <- lm (case_rate ~
                 mask_use +
                 sqrt_test_rate +
                 age_below_25 +
                 log(black_pop) +
                 mob_TS +
                 shelter_days +
                 bus_close_days,
               data = df
                 )

std_errors = list(
  sqrt(diag(vcovHC(mod7_1))),
  sqrt(diag(vcovHC(mod7_2))),
  sqrt(diag(vcovHC(mod7_3))),
  sqrt(diag(vcovHC(mod7_4)))
  )

stargazer(mod7_1, mod7_2, mod7_3, mod7_4, se = std_errors, type = "text")
```

```
##
## ==================================================================================================
##                                                      Dependent variable:
##                      -----------------------------------------------------------------------------
##                                                           case_rate
##                               (1)                    (2)                    (3)                    (4
## --------------------------------------------------------------------------------------------------
## mask_use                  -941.458***            -923.895***            -930.044***            -914.3
##                           (254.014)              (266.207)              (294.184)              (299.
##
## sqrt_test_rate             0.00000*               0.00000*               0.00000*               0.00
##                           (0.00000)              (0.00000)              (0.00000)              (0.00
##
## age_below_25              189.009***             186.940***             188.893***             186.9
##                            (55.740)               (56.016)               (51.068)              (52.
##
## log(black_pop)            214.014***             211.922***             214.376**              212.1
##                            (80.323)               (82.095)               (87.554)              (88.
##
## mob_TS                     15.181**               14.441*                15.499**               14.7
##                            (7.521)                (8.017)                (7.885)               (8.3
##
## shelter_days                                      -0.847                                        -1.
##                                                   (2.048)                                       (1.8
##
## bus_close_days                                                           5.860                  6.5
##                                                                         (12.697)               (12.
##
## Constant                 -3,252.670*             -3,163.101            -3,524.995**            -3,456
##                          (1,937.655)            (1,942.548)            (1,689.708)            (1,735
```

```
##
## ---------------------------------------------------------------------------------------------------
## Observations                  51                        51                        50                        50
## R2                          0.630                     0.631                     0.628                     0.62
## Adjusted R2                 0.589                     0.581                     0.576                     0.50
## Residual Std. Error   728.564 (df = 45)      736.018 (df = 44)      720.555 (df = 43)      727.921 (
## F Statistic         15.334*** (df = 5; 45) 12.536*** (df = 6; 44) 12.087*** (df = 6; 43) 10.171*** (
## ===================================================================================================
## Note:                                                                                  *p<0.1; **p<0.05
```

Question: What should be our final three model versions?

Answer: model_1 ~ mask_use + test_rate model_2 ~ mask_use + test_rate + below_25 + log(black_pop) model_3 ~ mask_use + test_rate + below_25 + log(black_pop) + shelter_days + bus_close_days

model_1 is point of departure model_2 is our best model model_3 is aimed to stress the significance of our coefficient when we add another policies that compete for variability with mask_use

```r
mod8_1 <- lm (case_rate ~
                mask_use +
                test_rate,
              data = df
                )

mod8_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                log(black_pop) +
                mob_TS,
              data = df
                )

mod8_3 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                log(black_pop) +
                mob_TS +
                shelter_days +
                bus_close_days,
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod8_1))),
  sqrt(diag(vcovHC(mod8_2))),
  sqrt(diag(vcovHC(mod8_3)))
  )

stargazer(mod8_1, mod8_2, mod8_3, se = std_errors, type = "text")
```

```
##
## ==========================================================================================
```

```
##                                           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
```

What would it look like if we had added poverty rate?

```r
mod8_1 <- lm (case_rate ~
                mask_use +
                test_rate,
              data = df
                )

mod8_2 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                poverty_rate +
                log(black_pop) +
                mob_TS,
              data = df
```

```r
                )

mod8_3 <- lm (case_rate ~
                mask_use +
                test_rate +
                age_below_25 +
                poverty_rate +
                log(black_pop) +
                mob_TS +
                shelter_days +
                bus_close_days,
              data = df
                )

std_errors = list(
  sqrt(diag(vcovHC(mod8_1))),
  sqrt(diag(vcovHC(mod8_2))),
  sqrt(diag(vcovHC(mod8_3)))
  )

stargazer(mod8_1, mod8_2, mod8_3, se = std_errors, type = "text")
```

```
##
## ======================================================================================
##                                            Dependent variable:
##                        ---------------------------------------------------------------
##                                                 case_rate
##                                  (1)                    (2)                    (3)
## --------------------------------------------------------------------------------------
## mask_use                    -990.470***            -971.910***            -940.918***
##                              (324.753)              (244.600)              (276.727)
##
## test_rate                     0.018*                 0.024**                0.022*
##                               (0.010)                (0.011)                (0.012)
##
## age_below_25                                       193.802***             191.382***
##                                                     (53.306)               (49.384)
##
## poverty_rate                                          5.400                  8.045
##                                                      (4.658)                (6.156)
##
## log(black_pop)                                      181.651**              157.591*
##                                                     (84.299)               (91.597)
##
## mob_TS                                               11.207                  7.960
##                                                      (8.462)               (10.611)
##
## shelter_days                                                                -2.655
##                                                                             (3.235)
##
## bus_close_days                                                              10.489
##                                                                            (13.039)
##
```

```
## Constant              2,530.239***          -4,776.968**          -5,374.913***
##                        (501.044)            (1,942.991)           (1,812.792)
##
## ------------------------------------------------------------------------------------------
## Observations               51                    51                    50
## R2                        0.236                 0.646                 0.658
## Adjusted R2               0.204                 0.597                 0.591
## Residual Std. Error  1,013.835 (df = 48)    721.192 (df = 44)      707.701 (df = 41)
## F Statistic         7.416*** (df = 2; 48) 13.362*** (df = 6; 44) 9.844*** (df = 8; 41)
## ==========================================================================================
## Note:                                                      *p<0.1; **p<0.05; ***p<0.01
```