Lab 2 Final Report: COVID-19 Case Rate vs Population Demographics and Mask Policy

Aidan Jackson, Frank Liu, Sam Temlock, Haoyu Zhang

Initial reassignment of common data used across models:

```
df <- read.csv("covid-19.csv", header = TRUE)
df<-df%>%
  rename(case_rate_100k = 'Case.Rate.per.100000',
         death_rate_100k = 'Death.Rate.per.100000',
         population_density = 'Population.density.per.square.miles',
         white_pct = 'White...of.Cases',
         black_pct = 'Black...of.Cases',
         hispanic_pct = 'Hispanic...of.Cases',
         other_pct = 'Other...of.Cases',
         state_emergency = 'State.of.emergency',
         business_closed = 'Closed.other.non.essential.businesses',
         business_reopen = 'Began.to.reopen.businesses.statewide',
         mask_public='Mandate.face.mask.use.by.all.individuals.in.public.spaces',
         mask_legal='No.legal.enforcement.of.face.mask.mandate',
         black_population_pct ="Black...of.Total.Population",
         white_population_pct = "White...of.Total.Population",
         poverty_pct = "Percent.living.under.the.federal.poverty.line..2018.",
         unemployed_pct = "Percent.Unemployed..2018.",
         senior_pct = "X65.",
         population_2018 = "Population.2018", ## extra variable in test
         non_elderly_pre_existing = "Nonelderly.Adults.Who.Have.A.Pre.Existing.Condition",
         homeless = "Number.Homeless..2019.") %>%
  select(State, case_rate_100k, death_rate_100k,population_density, white_pct, black_pct, hispanic_pct,
         other_pct, state_emergency, business_closed, business_reopen, mask_public, mask_legal,
         black_population_pct, white_population_pct, poverty_pct, unemployed_pct, senior_pct,
         population_2018, non_elderly_pre_existing, homeless)
# Assign correct data types
cols.num <- c("white_pct","black_pct","hispanic_pct","other_pct")
df[cols.num] <- sapply(df[cols.num],as.numeric)
df$state_emergency=as.Date(df$state_emergency, format = "%m/%d/%Y")
df$business_closed=as.Date(df$business_closed, format = "%m/%d/%Y")
df$business_reopen=as.Date(df$business_reopen, format = "%m/%d/%Y")
head(df)
```

1. An Introduction

Research Question: How is the COVID-19 case rate related to the distribution of population demographics within a state, and how does the effect of these demographics compare with the effect of mask-related policy decisions made by that state?

For this report, the investigation will be centered around the effect of population demographics and policy on the COVID-19 case rates across the United States (US), which is grouped by the 50 states plus the District of Columbia (D.C.). Within the report, this collection will be referred to as the "states", and each member a "state", inclusive of D.C..

The research question aims to measure the COVID-19 case rate per 100,000 residents within each state based on the make-up of the population, as well as the policy decisions that were or were not put into place in order to combat the rise of COVID-19 cases. In this sense, the aim is to examine how dependent the COVID-19 case rate is on that of which cannot be controlled (i.e., population demographics), as well as that of which can be controlled to a certain degree (i.e., implementation of policies to attempt to combat case rate) by each state. With this information, one could suggest whether or not the proliferation of COVID-19 within a state seems to be related to either or both controllable and uncontrollable variables, and to the level at which the effect of these variables differ help. The modeling goal of this research question will be one of description, and will be broken up into three phases of investigation

The first and primary phase of investigation involves how the COVID-19 case rate is affected by key population demographics. For this, the key variables are COVID-19 case rate per 100,000 residents and the percentage population distribution of seniors (defined as 65 years old or older) within a state. The case rate per 100,000 was selected as the key output/dependent variable as it provides a standardized measure of the spread of COVID-19 across the states, taking into account the absolute population of the states. The senior percentage was selected as the key input/independent variable given that guidance had been released by the Centers for Disease Control and Prevention (CDC) that seniors belong to the age category for those at higher risk of COVID-19, and thus are a well studied and documented group. This is likely due to the both the fact that seniors in this age group are more likely or had greater access to be tested given their categorization and the fact that they would have a greater likelihood of exhibiting detectable COVID-19 related symptoms that would prompt them to get tested. An additional rationale for measuring the senior percentage is that they may also be more likely to contract COVID-19 at lower viral loads, making them more susceptible to the virus. Conversely, this may result in seniors being more wary of the threat of COVID-19, and taking additional precautions to prevent infection relative to other age categories, such as limiting social interactions and taking more preventative measures in terms of hygiene. These key variables will provide an initial understanding of the relationship between the spread of the virus and population demographics.

Following this, in the second phase of investigation the analysis considers other variables of state-level population that are considered factors of contractability, namely the rate of poverty (defined as the percent of individuals living under the federal poverty line in 2018) and the rate of unemployment (defined as the percent unemployment in 2018). Those living below the poverty line may have less access/are unable to afford to preventative controls such as sanitation products and masks that help prevent the spread of the virus, while those who are designated as unemployed may not have the flexibility of sheltering at home as well as access to the aforementioned preventative controls. As mentioned, these variables operationalize population metrics that may lead to greater case rates, and are already standardized as rates to account for varying absolute populations across states.

The tertiary and final phase of investigation includes a variable that measures the implementation of policy as a response to COVID-19, and to understand the added effect of this variable in conjunction with the first phase of investigation that is state population metrics. Specifically, there is a focus on a policy that mandates the wearing of masks within the state. Through this, the aim is to analyze whether the implementation of a mask-related policy will have an effect on case rate, as well as the degree of this effect relative to population metrics. To operationalize this policy and simplify the measurements, the models only consider whether or not this policy was implemented through transformed indicator variables (1 = implemented, 0 = not implemented). As a result, factors of when and for how long the policy was implemented will be lost. Although this loss of information may fail to capture the impact of length of time of a policy on case rate (the policy may take time to be adopted/show meaningful efficacy), as there is no time-series data for case rate included in the data set, it was adjudged to be incongruent with the analysis.

Additionally, the final phase of investigation continues to expand on the examination of state-level population demographics via the population density variable (defined as the population density in square miles, population/square miles). A key factor to the spread of COVID-19 is the idea of social distancing, where

the virus is considered to be more likely to spread when people are close in proximity. Thus, population density is measured as a conduit to indicate the level of proximity within each state and will operationalize the concept of social distancing.

Assumptions

Prior to the analysis, it is important to identify a set of assumptions that have been made throughout the report and to assess the appropriateness of the data. Although there may be other considerations against the appropriateness of the data, the following highlights three particular arguments that must be taken into account when interpreting the results of the analysis.

Firstly, it is important to note that given each state is treated as a unique data point, the sample contains 51 data points. Although this size meets the general rule that an adequate sample size is 30 data points or more, as the analysis begins to factor in the wide range of potential population distributions and demographics within the states, it is clear that there is large variation within the sample. For example, the population density can have large variability depending on the population concentration among a few large areas as well as the level of uninhabited or sparsely inhabited land within a state, while the demographics of population can depend on a wide range of things such as regional factors and employment opportunities. This point is further discussed in the IID assumption addressed in section 3 of this report.

Secondly, note that there are many internal and external aspects of the selected variables that have not been included within the models. Just addressing the internal information that is lost, it can be seen that some of the information is not captured given the methods of operationalization discussed previously. Among others, the loss of information on the date of implementation within the mask policy variable strips out any contextual knowledge regarding the length of policy implementation, and precludes the identification of how long a specific policy was implemented. Assuming that this has some effect on case rate, this difference in length of time may play a part in the effect. There are also other external aspects that cannot be included, such as the level of population density broken out among different population demographics.

Finally, with regard to the policy variables, this report focuses on a form of mask mandate. Given this, it does not capture the effect of other policies that may or may not have a greater effect on the case rate (e.g., implementation of stay-at-home orders, closure of non-essential businesses, etc.). In justification, these choices were largely made due to there either being an appropriate amount of samples in each category (e.g., most states have implemented basic policies such as the closure of non-essential businesses and the implementation of stay-at-home orders, and thus these were not included given the lack of samples for the states that have not implemented them), or that data on other policies were simply not readily available in the dataset. The same argument can be applied in the selection of population demographics.

2. A Model Building Process

The primary variable of interest will be the total COVID-19 case rate per 100,000 residents. The COVID-19 case rate was judged to be able to provide a better understanding than the related death rate because of the potential for other, unrecorded variables that could be correlated with a person dying from COVID-19 versus just becoming infected. For example, the availability and quality of medical care in a state may impact its ability to keep COVID-19 infected patients alive, and these variables are not included in the data set. The health status of the residents in one state compared to another state may also affect the death rate, but this is also not included. Instead, the COVID-19 case rate was chosen so that these other variables which may correlate with the death rate would not have to be considered.

The covariates that will be examined in building the models fit into two broad categories. The first category is demographic information on the state, namely percentage population distribution of seniors, the unemployment rate, the rate of people living in poverty, and the population density. The second general group is that of policy decisions taken by the state, specifically a mask mandate. Problematic covariates would include any of the other direct measures of COVID-19 severity in the state, namely total infection rate not on a per capita basis and COVID-19 death rates. It can be assumed that these variables would be collinear

with the primary variable of interest since COVID-19 case rate is simply a linear transformation of total cases not adjusted for population, and death rate is derived from case rate.

As mentioned in the introduction, the modeling goal for the COVID-19 case rate will be one of description. A model of causation may follow this research if potential causal relationships are examined and incorporated into the model's structure. This may later be expanded to predict the COVID-19 case rate in the past seven days in order for a state to use it to potentially judge what effect a mask policy could have.

Model Variables

The three models aim to investigate the relationship between the COVID-19 case rate, state population demographics, and a mask mandates. COVID-19 case rate per 100,000 is chosen as the key dependent variable to represent the proliferation of the COVID-19 pandemic in each state. The population percentage of seniors is chosen as the key independent variable as a proxy for susceptibility by innate state population make-up. The variables of poverty percentage, unemployment percentage, mask mandate policy, and population density per square miles are chosen as candidate covariates to further explore the aforementioned potential relationship.

In order to allow for more simplistic and direct interpretations when analyzing the effect of the input variables on the output, it was determined that all percentage variables (senior, poverty, and unemployment) would be transformed into rates per 100,000 to align with the case rate variable. It will also serve to ease the burden on the readers of this report as like-for-like scaled relationships can be drawn between the variables. Given that this transformation is constant for all values (scalar multiplication of 100,000), it will have no effect on the distributions of the variables.

```
# Percentage variables are re-scaled to rates per 100k

df <-df %>%
  mutate(
    senior_per_100k = 100000*senior_pct,
    poverty_per_100k = 100000*poverty_pct,
    unemployed_per_100k = 100000*unemployed_pct,
  )
```

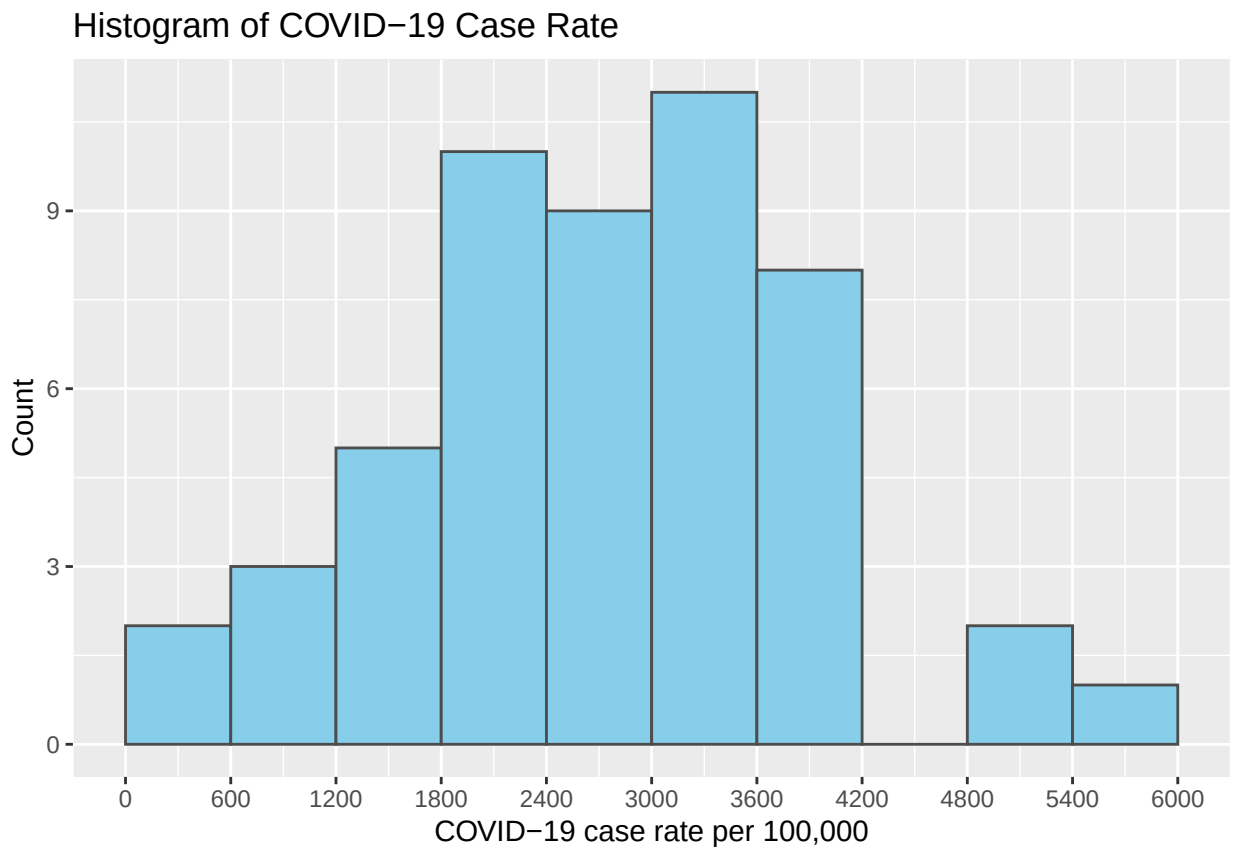
Model 1

For the first model, the relationship between the COVID-19 case rate per 100,000 and the population percentage distribution of seniors was analyzed. First, the distribution of the case rate dependent variable is examined.

```
summary(df$case_rate_100k)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     344    2040    2633    2749    3516    5589
```

```
ggplot(data = df,
  mapping = aes(x= case_rate_100k)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0,6000,600)) +
  labs(title =  "Histogram of COVID-19 Case Rate",
       x = "COVID-19 case rate per 100,000", y = 'Count') +
  scale_x_continuous(breaks=seq(0, 6000, 600))
```



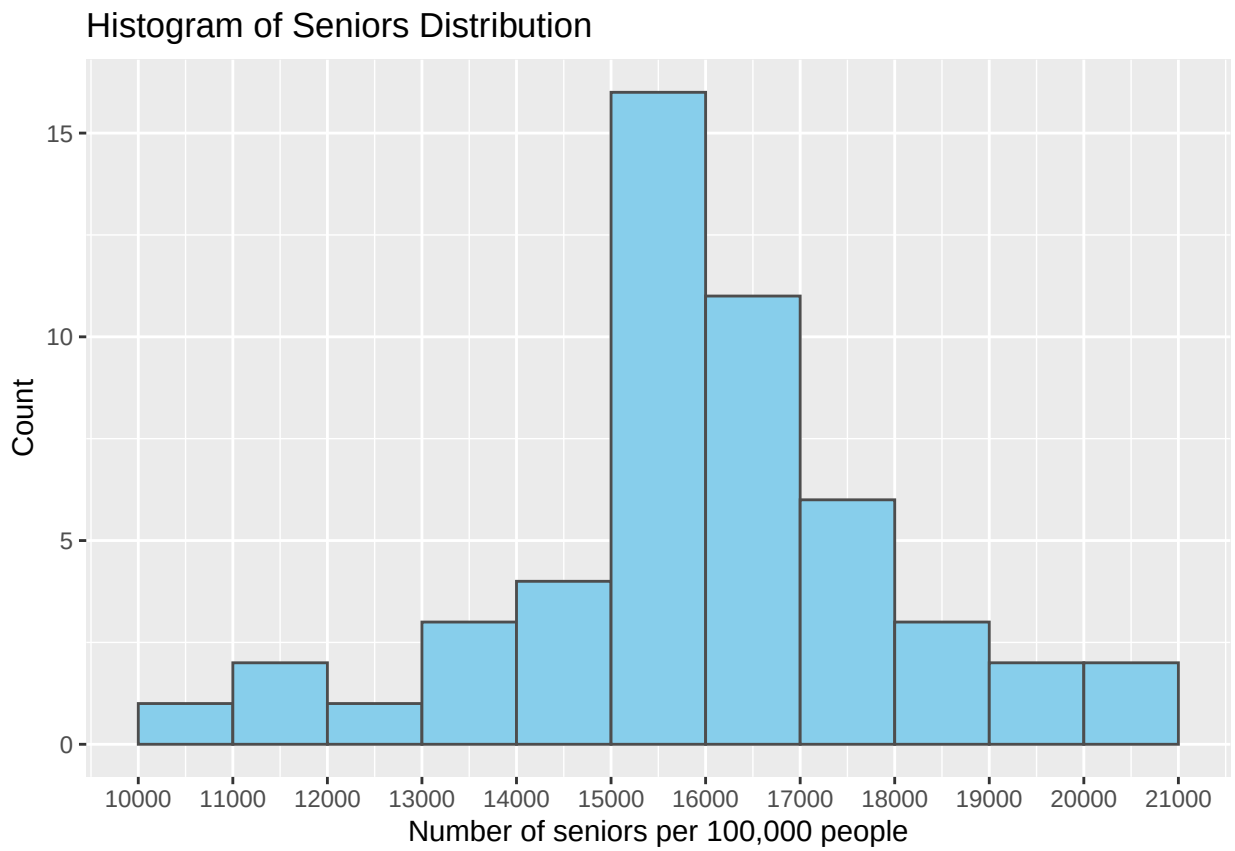
As can be seen above, the distribution is fairly normal, and given that it has already been standardized as a rate across all states, there is no need to perform any transformations on this variable. Thus, the case rate per 100,000 variable can be leveraged as is as the dependent variable for all three models.

Next, the distribution of the senior percentage transformed into the rate per 100,000 (as discussed in the Model Variables section) is examined.

```
summary(df$senior_per_100k)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11000   16000   16000   16471   17500   21000
```

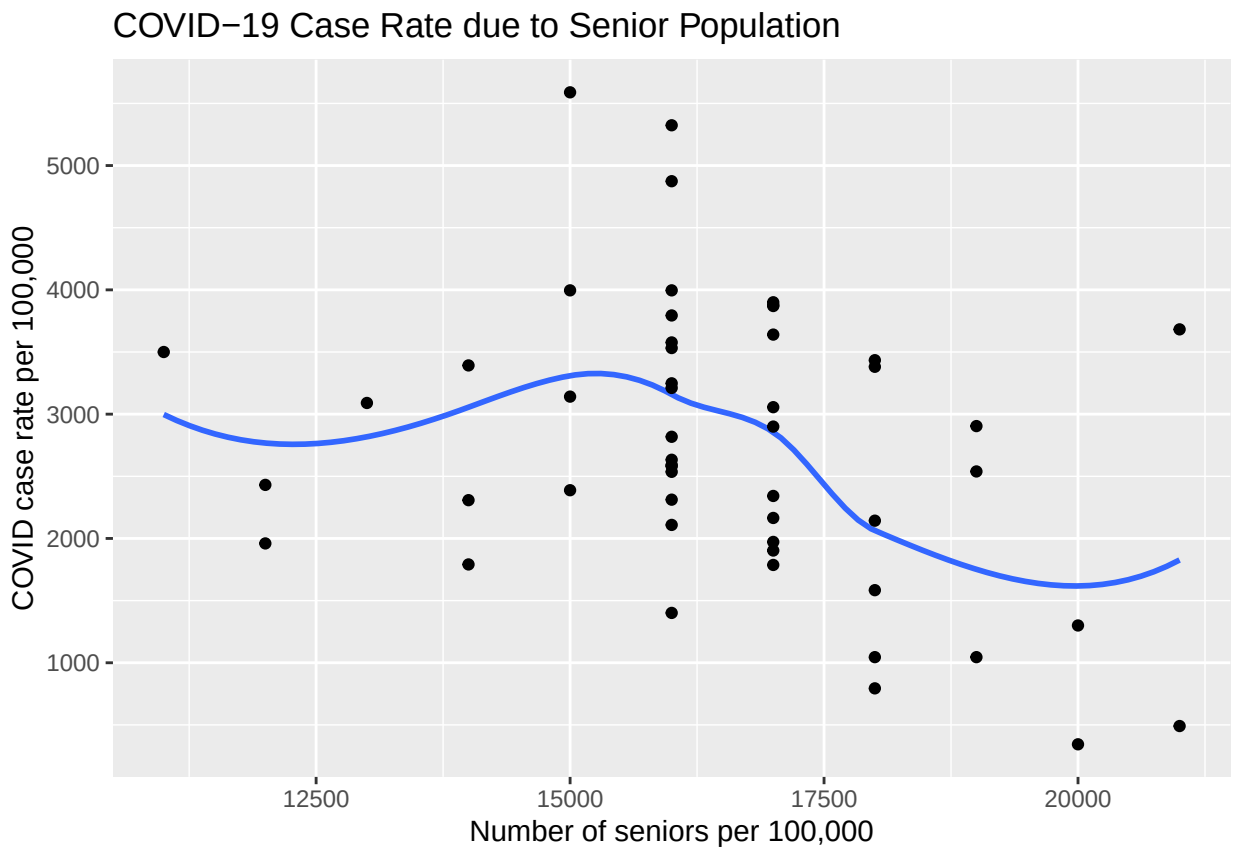
```
ggplot(data = df,
  mapping = aes(x= senior_per_100k)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(10000,21000,1000)) +
  labs(title = 'Histogram of Seniors Distribution',
  x = 'Number of seniors per 100,000 people', y = 'Count') +
  scale_x_continuous(breaks=seq(10000,22000,1000))
```



As can be seen from the distribution above, again there is a fairly normal distribution in the population of seniors per 100,000 across the state. Additionally, there do not seem to be any outliers within the distribution, given that the range of rates fall between 11,000 and 21,000. Therefore, there are no transformations that need to be made in order to restructure the distribution, and this variable can be used across all three models.

With the appropriate variable distributions investigated, a plot is created to analyze the relationship between them.

```
df %>%
  ggplot(aes(senior_per_100k, case_rate_100k)) +
  geom_smooth(se = FALSE) +
  geom_point() +
  labs(
    title = 'COVID-19 Case Rate due to Senior Population',
    x = 'Number of seniors per 100,000',
    y = 'COVID case rate per 100,000'
  )
```



From the above plot, there is noticeable variation in the smoothed blue line. However, it can be posited that there is a inverse relationship between the variables given the roughly linear relationship in the overall downward trend in the line. In order to test this, the following equation is used to create a regression model to determine the true relationship between the case rate and population density variables.

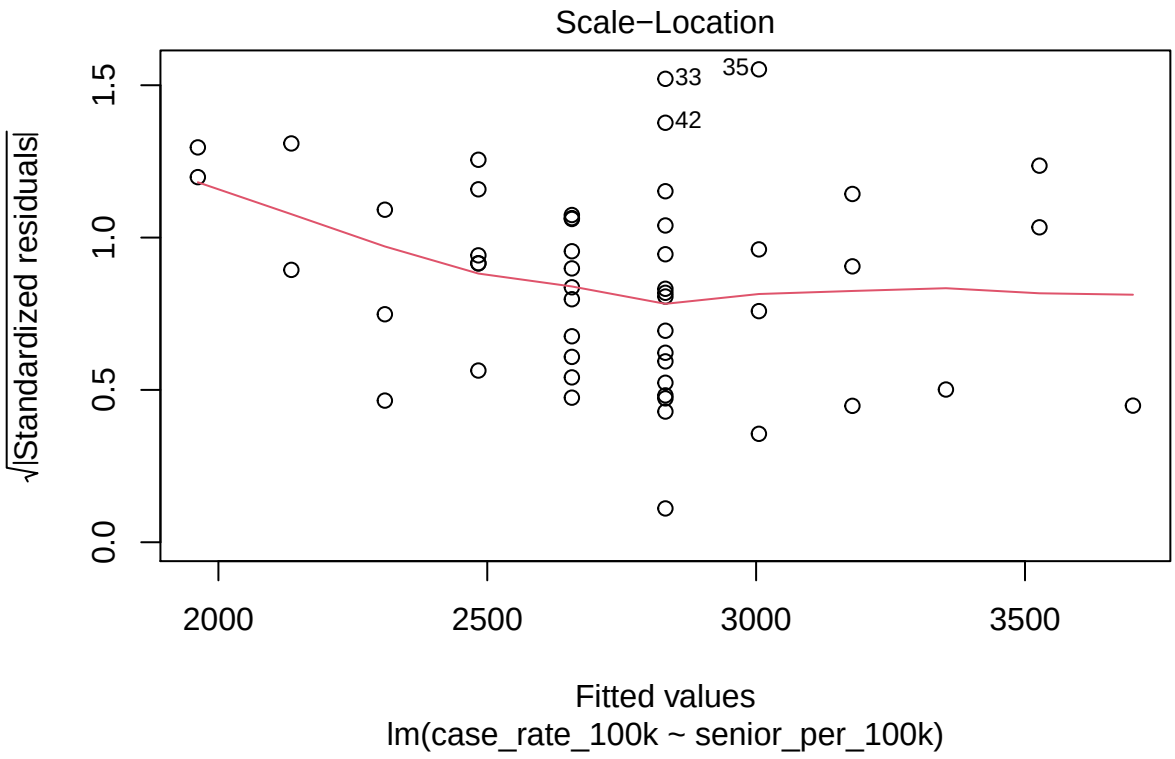
$$case\_rate\_100k = \beta_0 + \beta_1 senior\_per\_100k$$

```
# Build the regression model for Model 1
model1 <- lm(case_rate_100k ~ senior_per_100k , data = df)
```

To improve the precision of the t-test of coefficients, classical standard errors can be used over robust standard errors to test the regression. To determine the applicability of using classical standard errors, in addition to the three Classical Linear Model (CLM) assumptions necessary for the robust standard errors, the CLM assumption for homoskedastic conditional errors should be evaluated. If this additional assumption is met, the robust standard errors can be replaced by the classical standard errors. The assumption of homoskedasticity is met if the data does not have large variance among the residuals.

To evaluate this assumption, the square root of the residuals (to remove any negative values) is plotted against the fitted values of the model. Additionally, a Breusch-Pagan test is run to check the level of heteroskedasticity.

```
# Check the homoskedasticity of errors in model 1
plot(model1, which=3)
```




```
# Run the Breusch-Pagan test for model 1
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 0.46099, df = 1, p-value = 0.4972
```

From the above plot, the variance in residuals can be seen to be roughly constant given the relative flatness of the plotted red line, thus meeting the assumption of homoskedastic conditional errors. This is supported by the p-value of 0.497 from the Breusch-Pagan test, meaning it fails to reject the null hypothesis that the conditional errors are not heteroskedastic. With this additional assumption met, the regression model is run using a t-test with classical standard errors.

```
coeftest(model1)
```

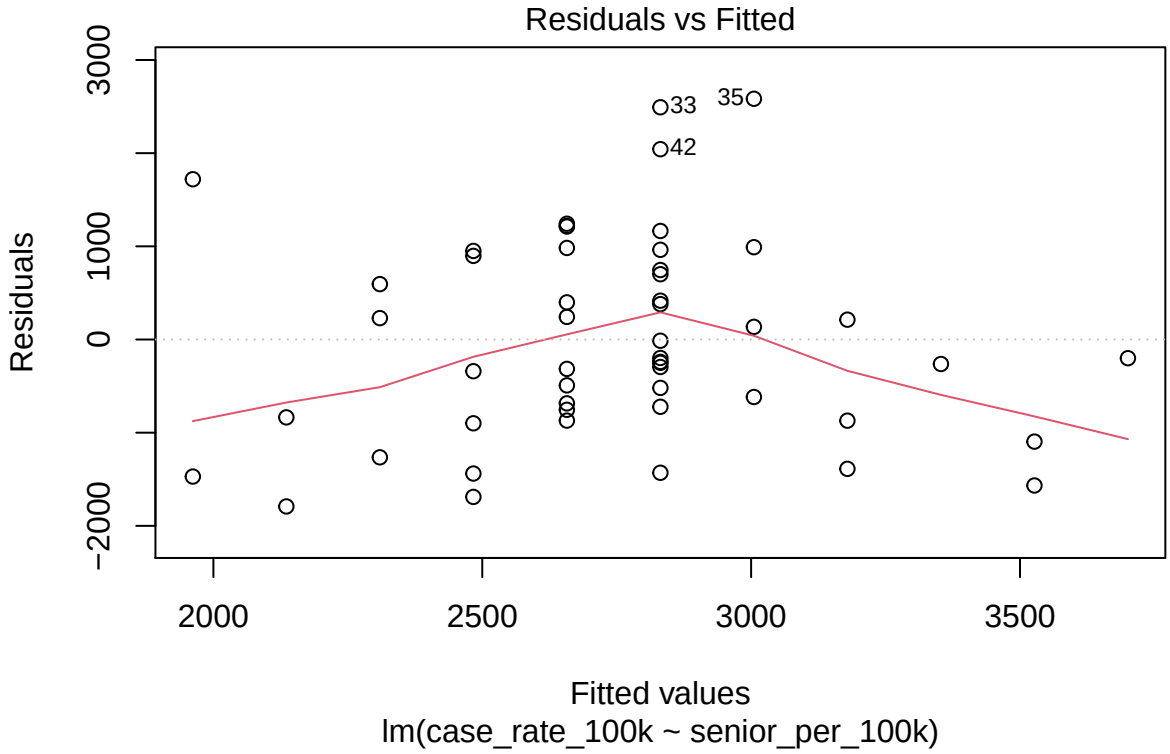
```
##
## t test of coefficients:
##
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     5613.666667 1233.338521  4.5516 3.531e-05 ***
## senior_per_100k   -0.173900    0.074307 -2.3403   0.02338 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen from the above output, the variable for senior percentage is shown to be significant within the model specification with a p-value of 0.023. Furthermore, The coefficient of -0.174 for the variable is negative, supporting the earlier evaluation that there is an inverse relationship between the senior percentage and the case rate. In practical significance, this can be interpreted as for every additional senior per 100,000 in the population, there is a decrease of 0.174 COVID-19 cases per 100,000. More intuitively, for roughly every 6 additional seniors per 100,000, there is a decrease of 1 case per 100,000.

Next, the other CLM assumptions are revisited to determine the whether the coefficients are unbiased and consistent. However, the assumption for IID is not detailed here given that it is evaluated more generally for all three models within section 3.

The second CLM assumption to be evaluated here is that of a linear conditional expectation relationship within the data, so that the model is accurately estimating the relationship between the variables. To test for this, the residuals of model 1 are plotted against its fitted values.

```
# Check for linearty in model 1
plot(model1, which=1)
```

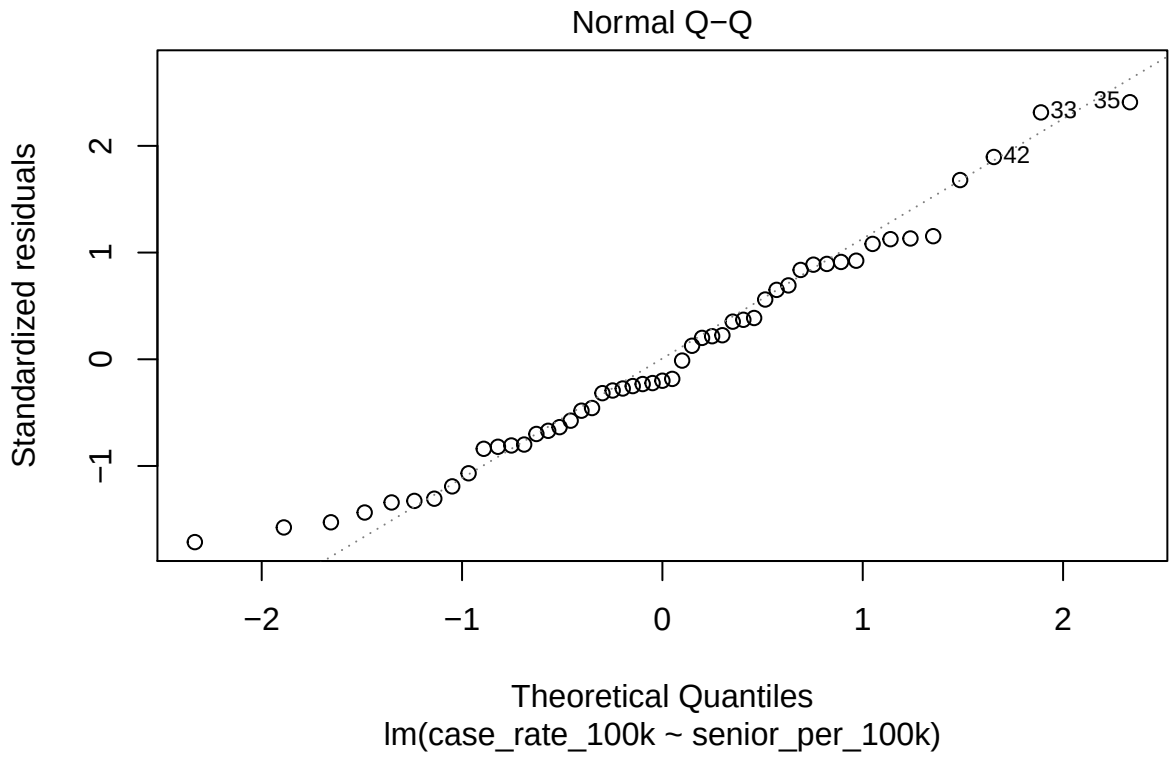


When examining the plot above for linearity in the conditional expectation, there is a noticeable slightly quadratic shape to the relationship of the conditional expectation. This may suggest that the relationship between the variables may not be as linear as the model assumes. However, as discussed previously in the scatter plot of the two variables, the relationship can be approximated as roughly linear. It is also difficult to explain the complexity behind the COVID-19 case rate with just a single variable that does not fully capture the broader population demographics. Additionally, there is hope that given the addition of other input variables within models 2 and 3, the relationship within the data will be better explained and result in more linearity.

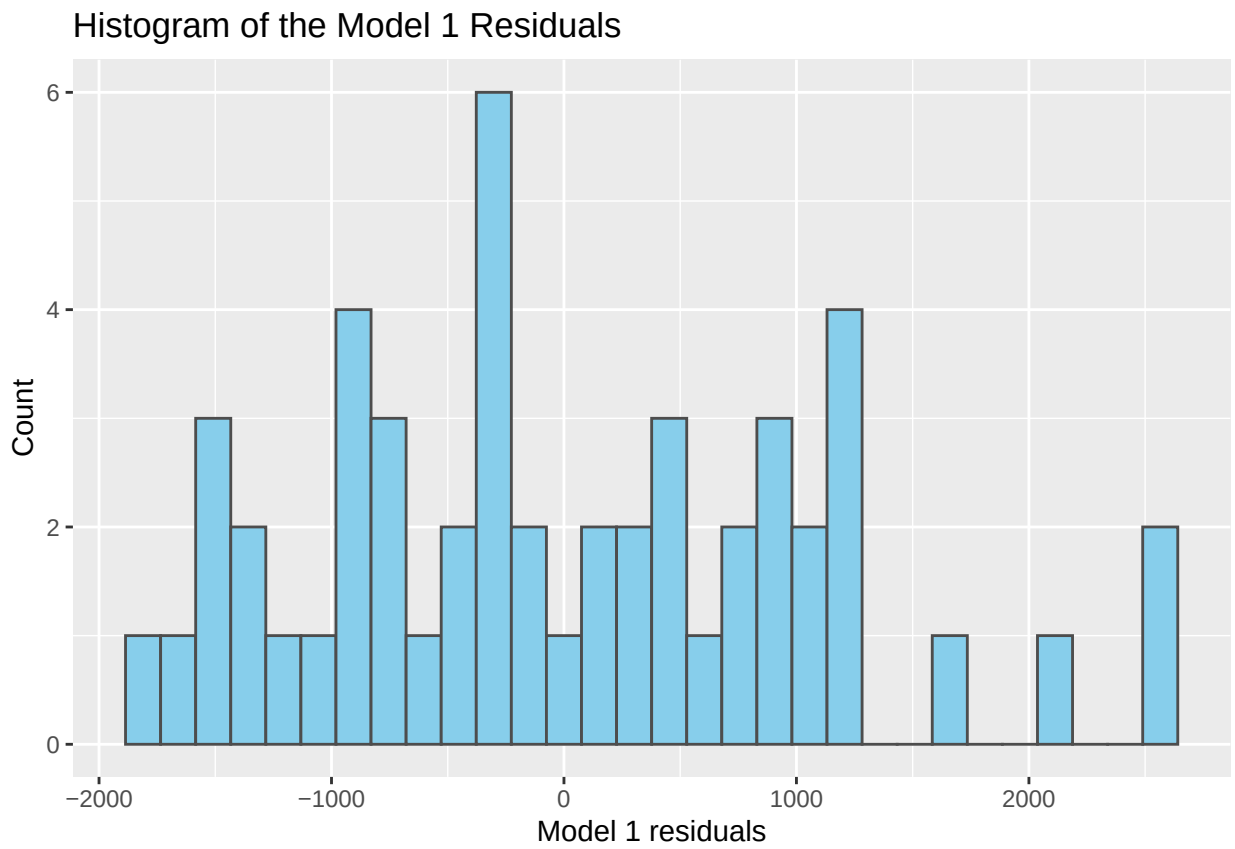
The third CLM assumption examined for model 1 is the assumption of normality of errors. This is a necessary

assumption to ascertain that the errors used in the regression are drawn from a normal distribution, so that they can be accepted when used to calculate the significance levels. Both a QQ plot and histogram of the distribution of the residuals is plotted to observe the normality.

```
# Create a QQ plot for the model 1 residuals
plot(model1, which=2)
```



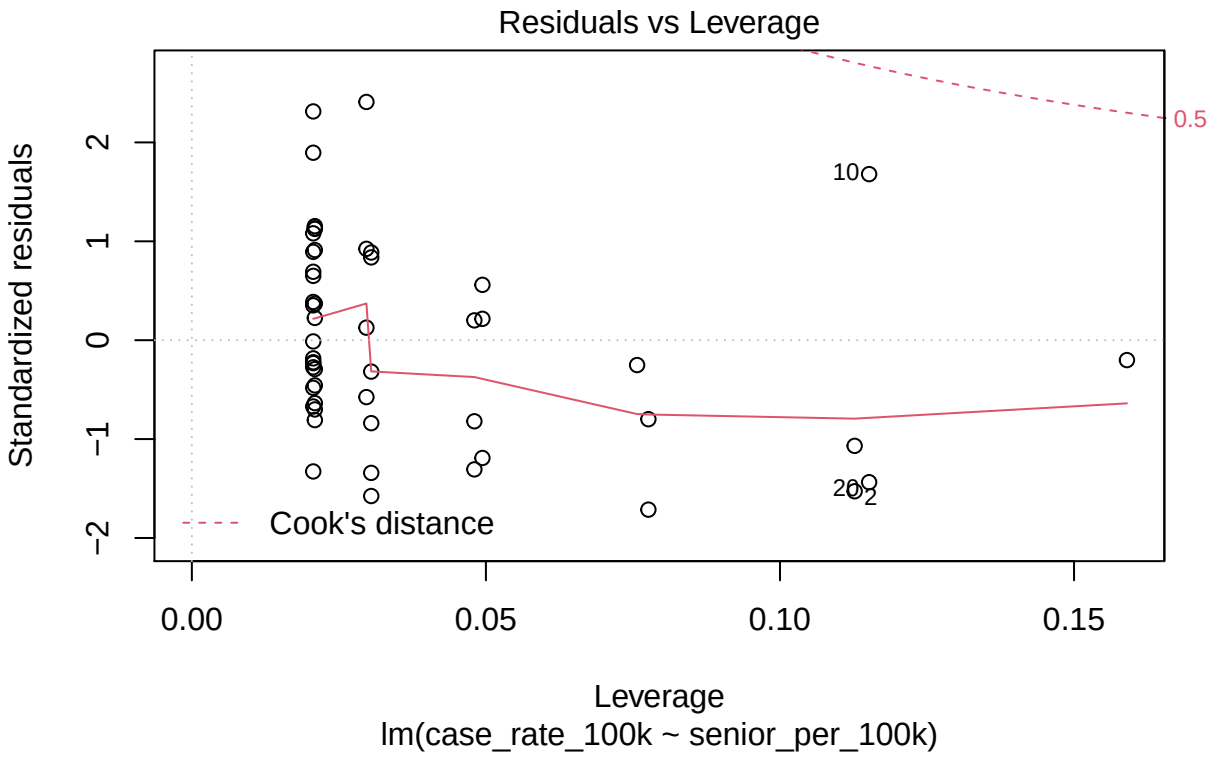

```
# Create a histogram of the distribution of the model 1 residuals
df %>%
  ggplot(aes(x = resid(model1))) +
  stat_bin() +
  geom_histogram(fill = 'skyblue', color = 'grey30', bins=30) +
  labs(title =  "Histogram of the Model 1 Residuals",
       x = "Model 1 residuals", y = 'Count')
```



From the QQ plot, normality in the residuals can be detected given the proximity of the points to the normal line. Additionally, although the shape is a little harder to observe, the histogram of the residuals also gives a fairly normal distribution.

Finally the Cook's distance is examined by a plot of the residuals vs the model 1 leverage to look for any outliers in the values. From the plot below, no obvious outliers are detected.

```
# Investigate outliers using Cook's distance
plot(model1, which=5)
```



It is also noted that there is no test conducted for collinearity in model 1 given there is only one input variable.

Model 2

For model 2, state demographic information regarding racial composition(white percentage of total population /Black Percentage of Total Population), economic status(unemployed percentage/poverty percentage), and population seniors percentage was investigated to explore its potential connection to the COVID pandemic status.

2-a) White Percentage of Total Population/ Black Percentage of Total Population Variable of black percentage is stored as string variables. Moreover, three entries of black percentage of total population is "<0.01", which is to be dropped or replaced by values determined by extra resources for further analysis.

Here, the value of black percentage is replaced by 0 when it is "<0.01". Afterwards, the correlation of white percentage and black percentage is -0.42, which shows these two variables are correlated. Considering that the value of other race groups is much smaller, it is proper to only include white percentage in the regression model to explore the relation of care rate and racial composition.

The distribution of the white % is skewed, not an ideal normal distribution. While the distribution of the square of white % is more close to a normal distribution. In practice, this square could reflect the chance that the interaction happens between two white people.

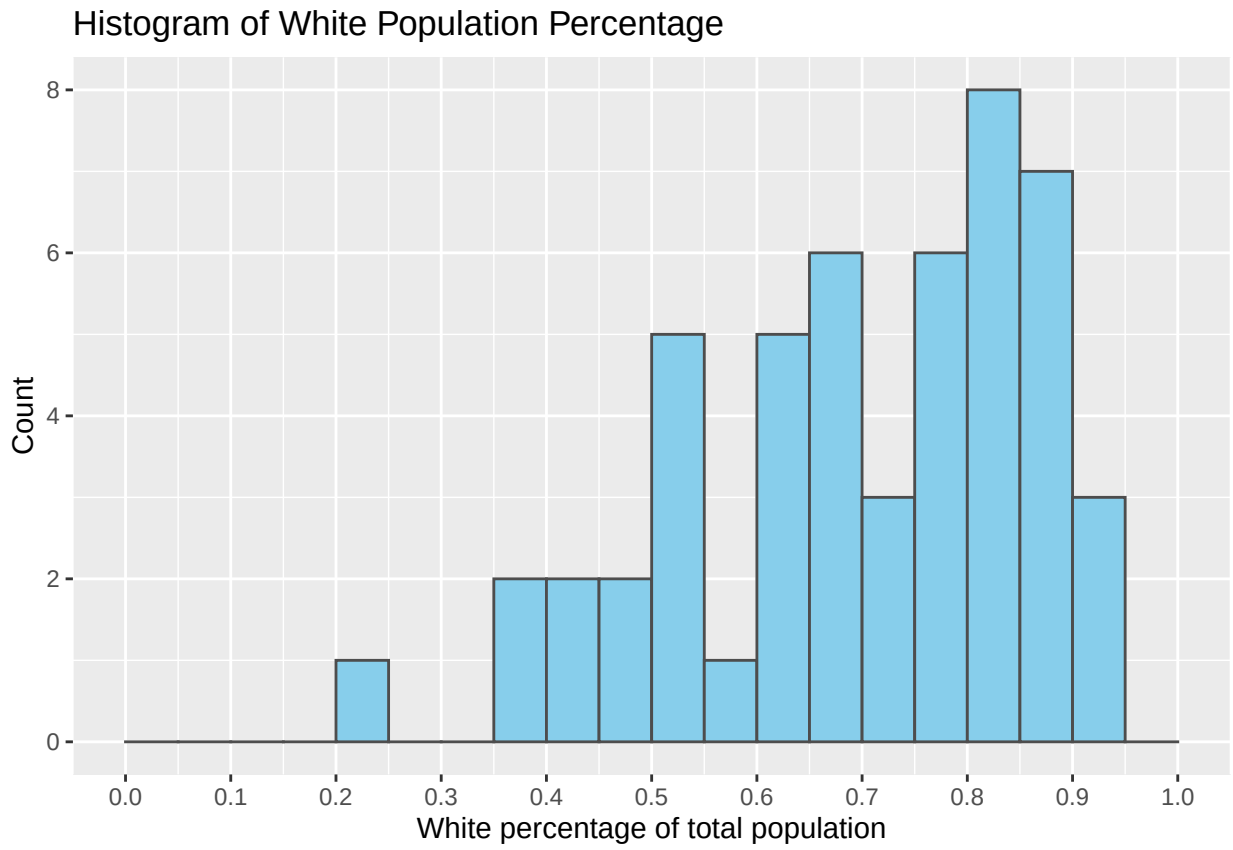
```
black_num <- as.numeric(df$black_population_pct)
black_num[is.na(black_num)]<- 0
cor(df$white_population_pct, black_num)
```

```
## [1] -0.4212923
```

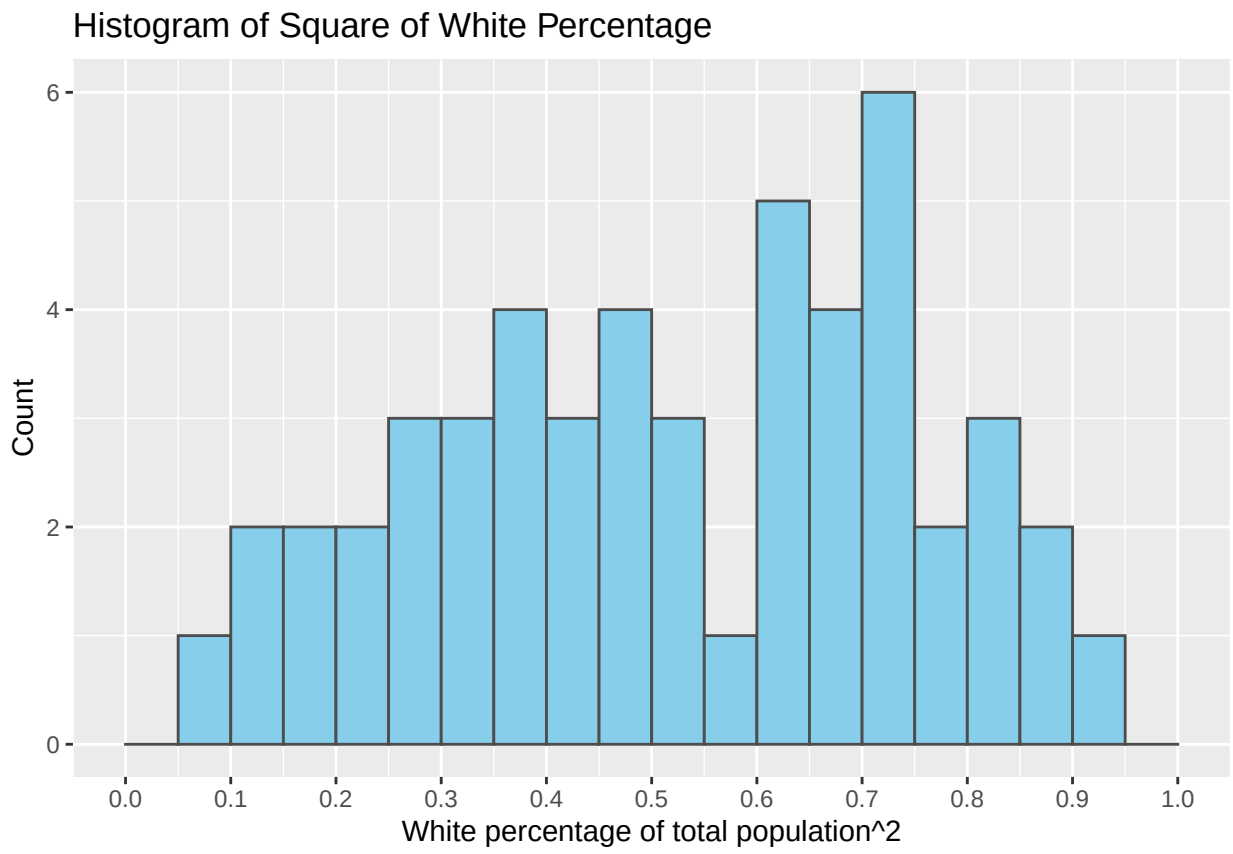
```
summary(df$white_population_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2300  0.5900  0.7200  0.7004  0.8400  0.9500
```

```
ggplot(data = df,
  mapping = aes(x= white_population_pct))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0,1,0.05)) +
  labs(title =  "Histogram of White Population Percentage",
       x = "White percentage of total population", y = 'Count')+
  scale_x_continuous(breaks=seq(0, 1, 0.1))
```




```
ggplot(data = df,
  mapping = aes(x= (white_population_pct)^2))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0,1,0.05)) +
  labs(title =  "Histogram of Square of White Percentage",
       x = "White percentage of total population^2", y = 'Count')+
  scale_x_continuous(breaks=seq(0, 1, 0.1))
```



2-b) Percentage living under the federal poverty line (2018) Correlation of pecentage living under poverty line and white percentage of total population shows that they are not significantly related. The distribution of the poverty percentage is not heavily skewed.

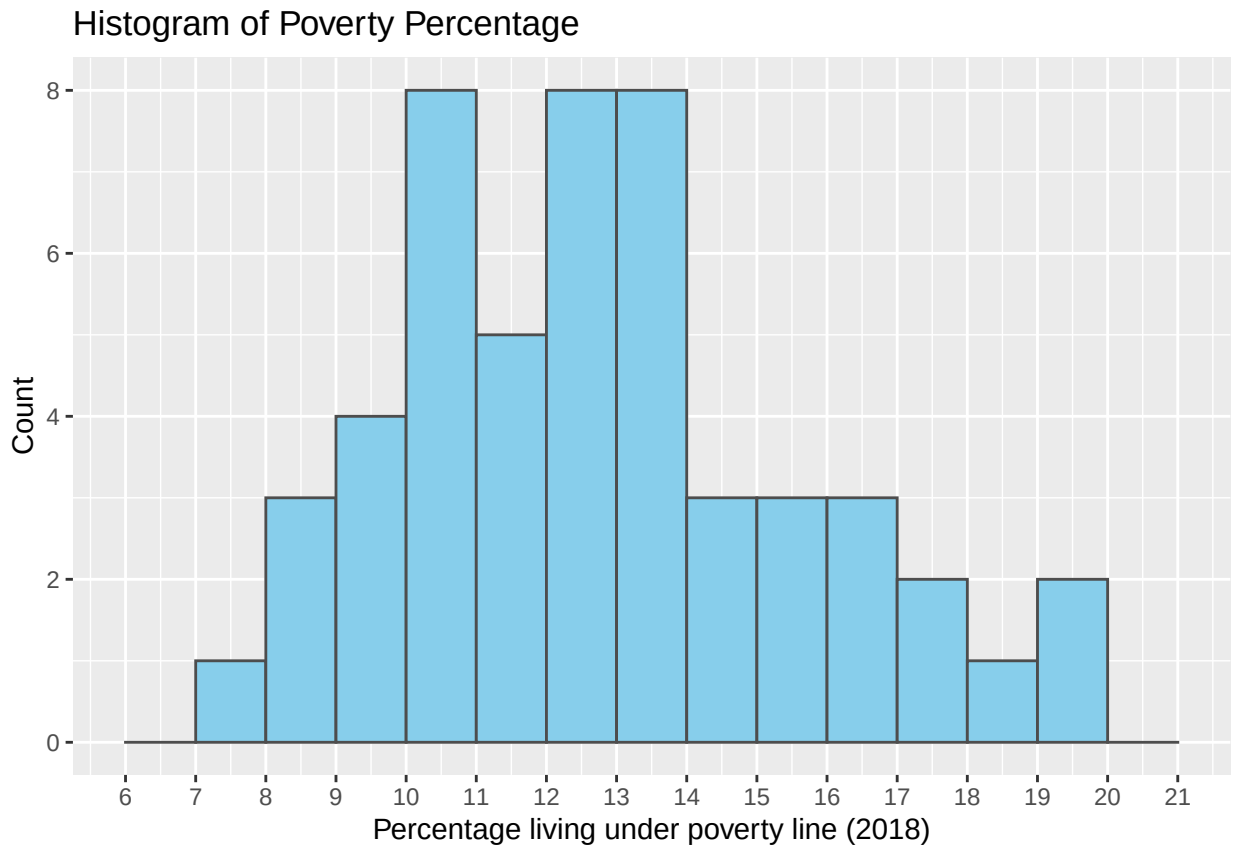
```
summary(df$poverty_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.60   10.95   12.80   12.91   14.20   19.70
```

```
cor(df$poverty_pct, df$white_population_pct)
```

```
## [1] -0.1571033
```

```
ggplot(data = df,
  mapping = aes(x= poverty_pct))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(6,21,1)) +
  labs(title =  "Histogram of Poverty Percentage",
       x = "Percentage living under poverty line (2018)", y = 'Count')+
  scale_x_continuous(breaks=seq(6, 21, 1))
```



2-c) Unemployed Percentage (2018) Although, the distribution is of high peak in the middle of the range. Overall, it is not highly skewed or heavily tailed. In addition, the unemployed rate is correlated to the variable to the poverty rate, which is in an agreement with intuitive expectation.

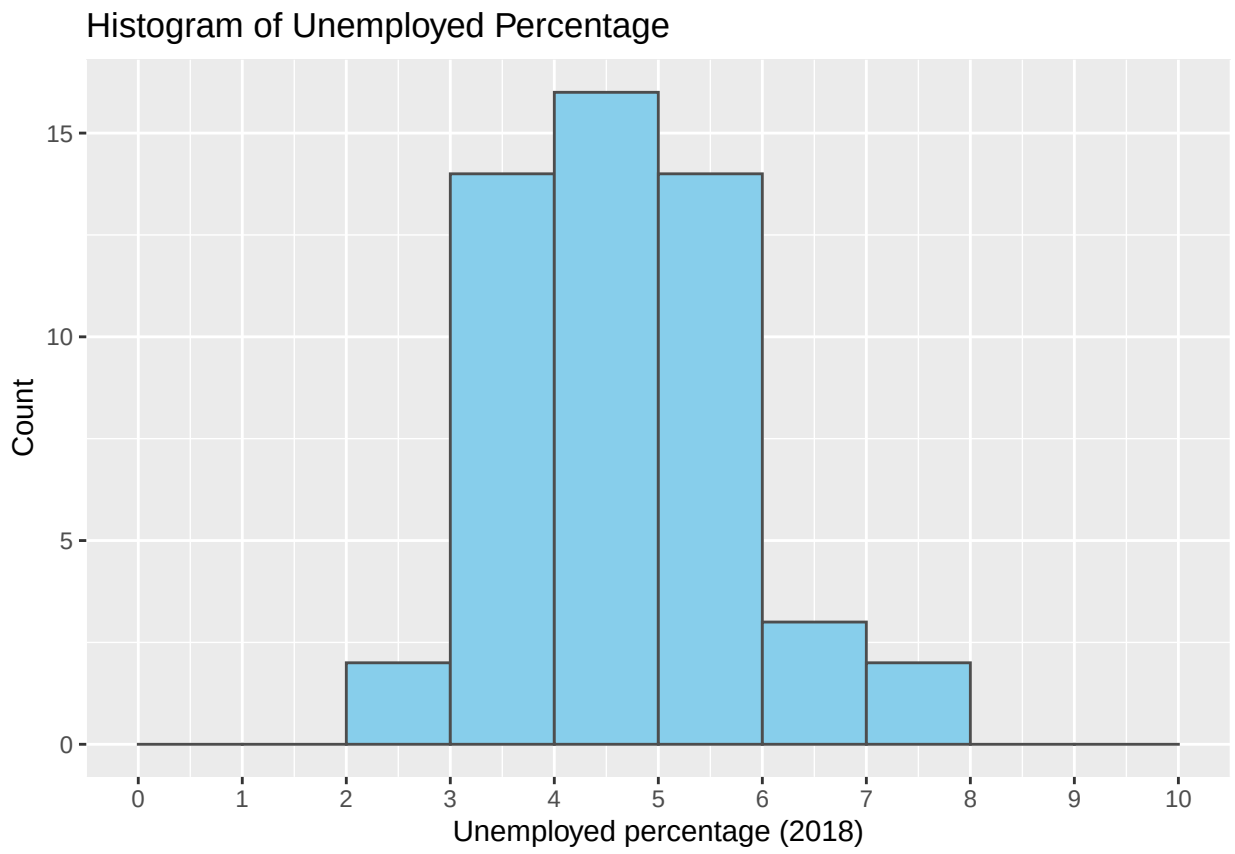
```
summary(df$unemployed_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.800   3.850   4.900   4.747   5.500   7.500
```

```
cor(df$unemployed_pct, df$poverty_pct)
```

```
## [1] 0.6101304
```

```
ggplot(data = df,
  mapping = aes(x= unemployed_pct))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0,10,1)) +
  labs(title =  "Histogram of Unemployed Percentage",
       x = "Unemployed percentage (2018)", y = 'Count')+
  scale_x_continuous(breaks=seq(0, 10, 1))
```

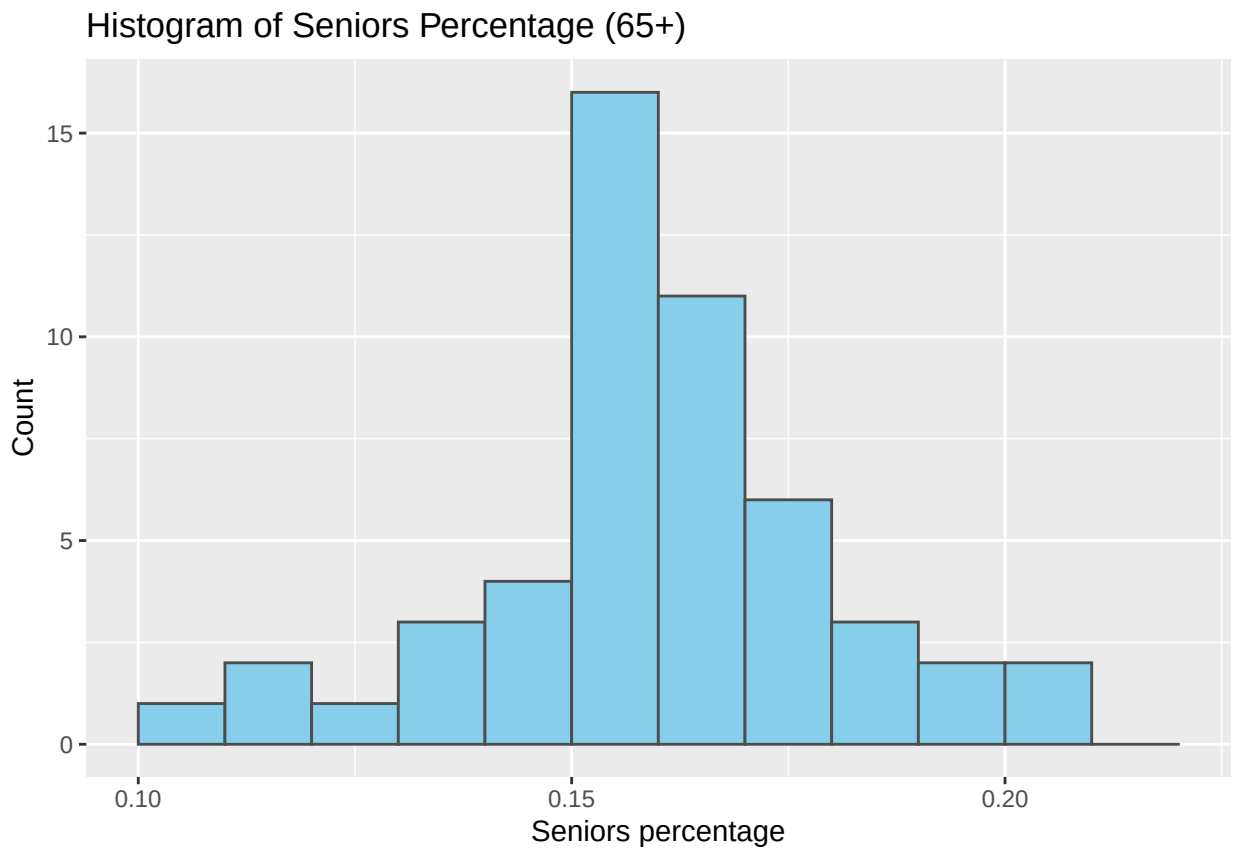


2-d) Seniors Percentage (Age 65+) The seniors percentage (age 65+) is of a nearly normal distribution.

```
summary(df$senior_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1100  0.1600  0.1600  0.1647  0.1750  0.2100
```

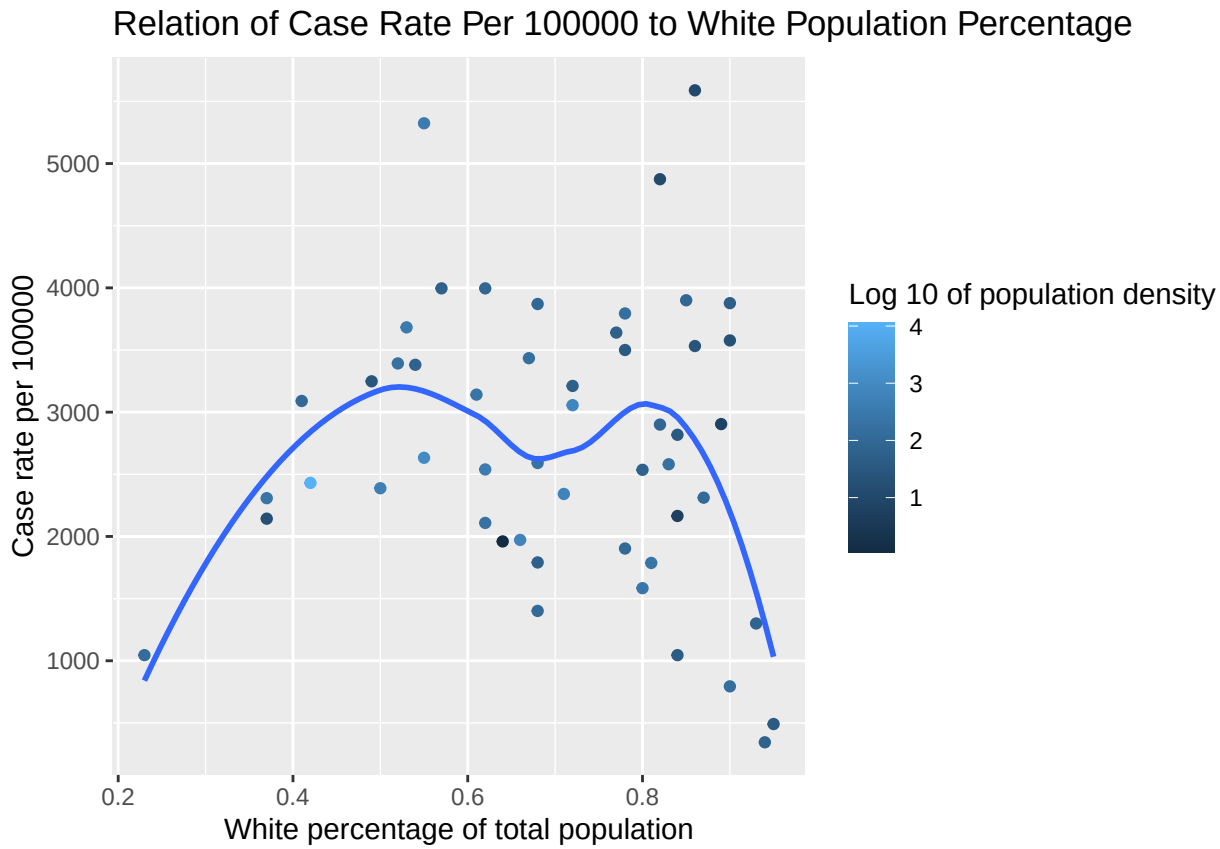
```
ggplot(data = df,
  mapping = aes(x= senior_pct))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0.10,0.22,0.01)) +
  labs(title =  "Histogram of Seniors Percentage (65+)", x = "Seniors percentage",
       y = 'Count')+
  scale_x_continuous(breaks=seq(0.1, 0.25, 0.05))
```



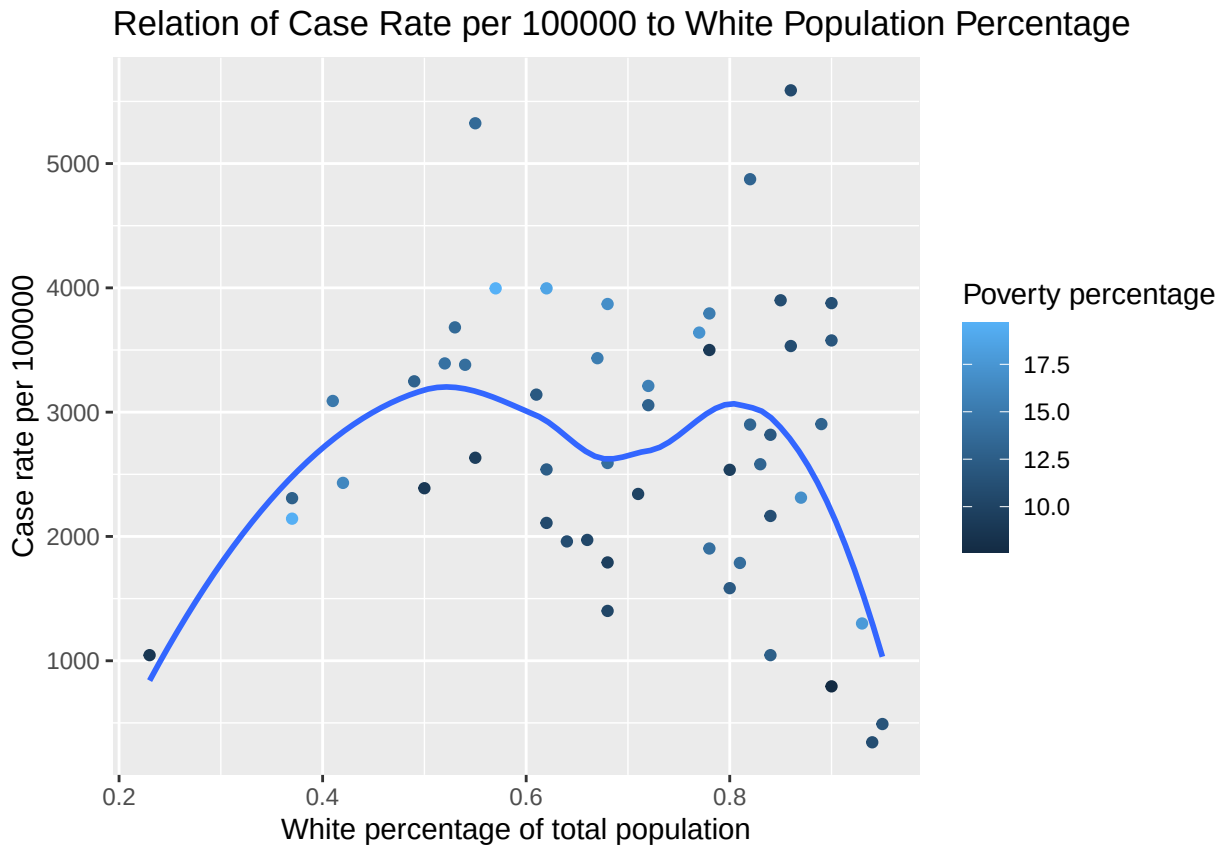
Model 2 Plots

No obvious trend could be easily observed or concluded from the plots regarding the dependent variables and candidate covariates.

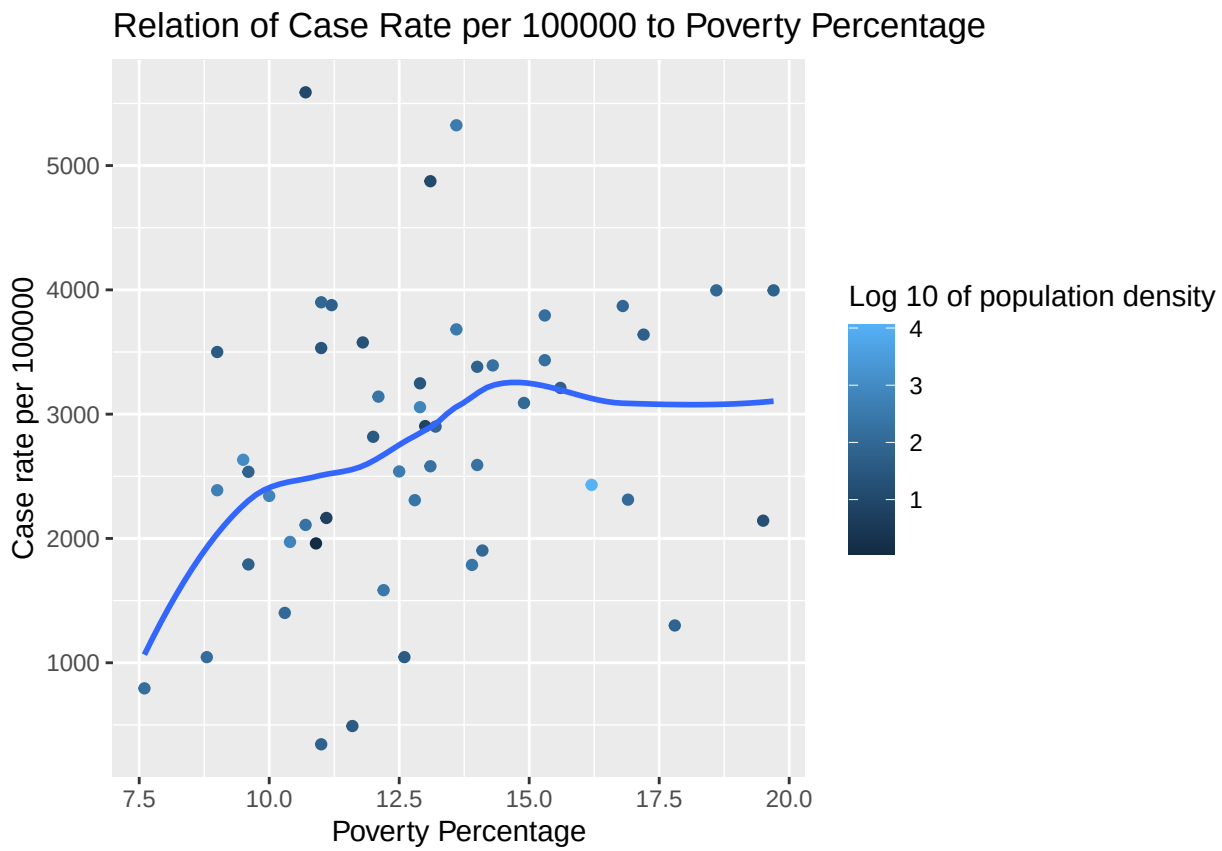
```
df %>%
  ggplot(aes(white_population_pct, case_rate_100k, color = log10(population_density))) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate Per 100000 to White Population Percentage',
    x = 'White percentage of total population',
    y = 'Case rate per 100000',
    color = 'Log 10 of population density'
  )
```



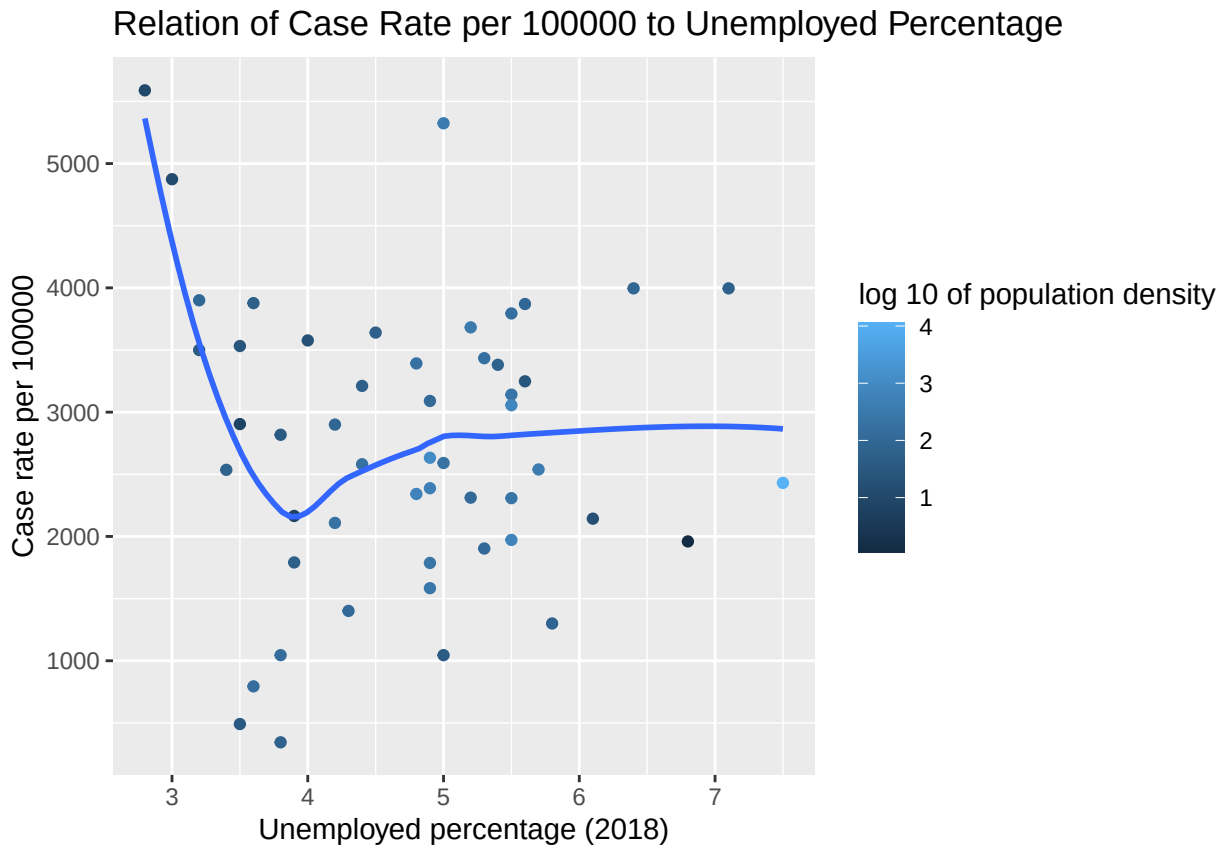

```
df %>%
  ggplot(aes(x = white_population_pct, y = case_rate_100k, color = poverty_pct)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate per 100000 to White Population Percentage',
    x = 'White percentage of total population',
    y = 'Case rate per 100000',
    color = 'Poverty percentage'
  )
```



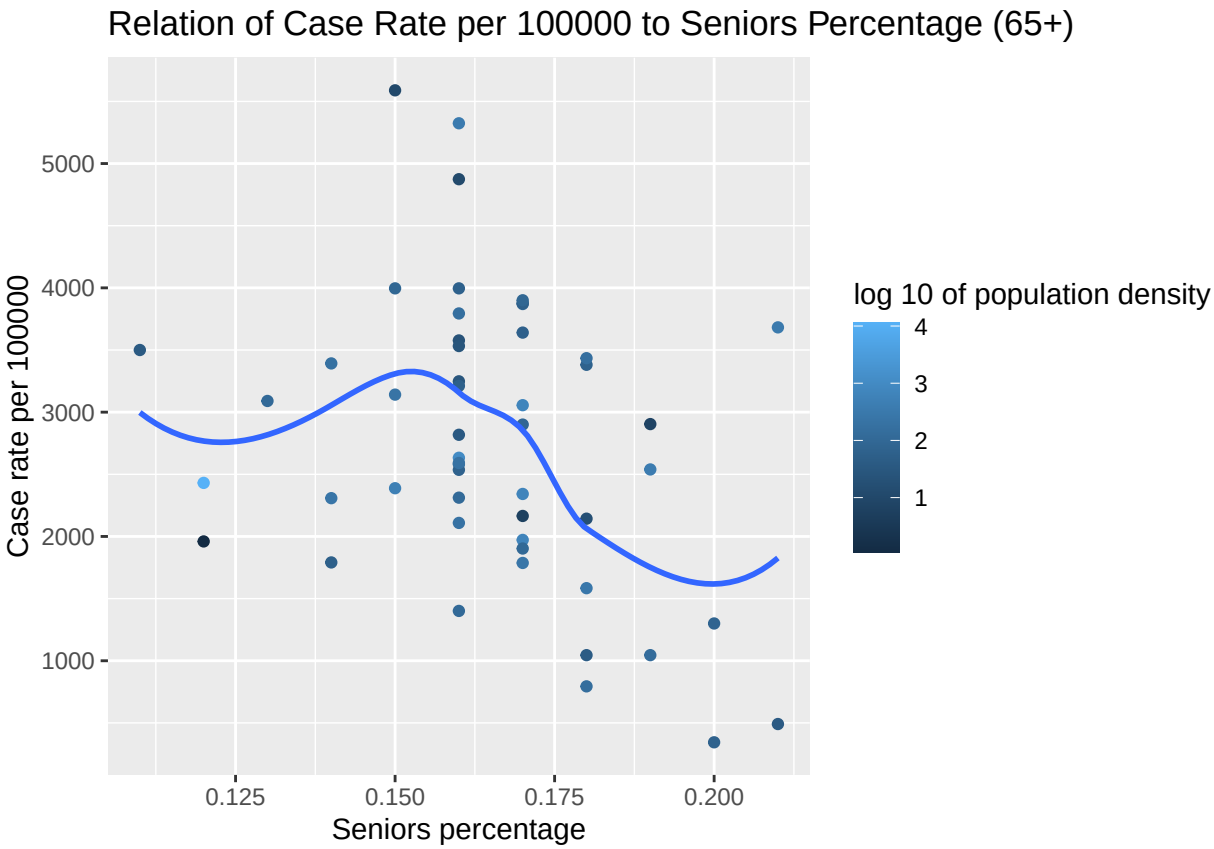

```
df %>%
  ggplot(aes(x = poverty_pct , y = case_rate_100k, color = log10(population_density))) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate per 100000 to Poverty Percentage',
    x = 'Poverty Percentage',
    y = 'Case rate per 100000',
    color = 'Log 10 of population density'
  )
```




```
df %>%
  ggplot(aes(unemployed_pct, case_rate_100k, color = log10(population_density))) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate per 100000 to Unemployed Percentage',
    x = 'Unemployed percentage (2018)',
    y = 'Case rate per 100000',
    color = 'log 10 of population density'
  )
```




```
df %>%
  ggplot(aes(senior_pct, case_rate_100k, color = log10(population_density))) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate per 100000 to Seniors Percentage (65+)',
    x = 'Seniors percentage',
    y = 'Case rate per 100000',
    color = 'log 10 of population density'
  )
```



Model 2 Regression

The regression models regarding the demographic variables discussed previously are listed as below.

$$Case.Rate.Per.100K = \beta_0 + \beta_1 Senior.Rate.Per.100K + \beta_2 Poverty.Rate.Per.100K + \beta_3 Unemployed.Per.100K$$

```
model2 <- lm(case_rate_100k ~ senior_per_100k + poverty_per_100k + unemployed_per_100k, data = df)
coeftest(model2)
```

```
##
## t test of coefficients:
##
##                         Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)           5.8693e+03  1.3629e+03  4.3067 8.371e-05 ***
## senior_per_100k      -2.3097e-01  6.8570e-02 -3.3684 0.0015170 **
## poverty_per_100k      2.3175e-03  6.2345e-04  3.7172 0.0005355 ***
## unemployed_per_100k -4.8619e-03  1.6846e-03 -2.8860 0.0058754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(need to update) From the results of the regression, none of the coefficients of investigated variables in model 2 are significant (p-value < 0.05). Among these regression coefficient tests, the p-value regarding the poverty percentage variable is the lowest (0.05755). In addition, its positive estimated coefficient is consistent

with the statement that low-income communities are at high pandemic risk claimed by mainstream media these days.

Model 3

For Model 3, policy related variables will be introduced for the analytics. Policy variable "Mandate face mask use by all individuals in public spaces" (renamed as mask_public) is selected as control variable to predict the case rate. Given it has been proved that mask is effective in preventing the transmission of disease, it is expected that the mask policy mandate in public would have an effect in reducing the case rate.

There are other variables related to mask such as "No legal enforcement of face mask mandate" and "Mandate face mask use by employees in public-facing businesses", we believe these variables have a lower impact than the policy that is enforced to public. Also, it is very likely that if a state enforce public mask policy, they will by default enforce mask policy for employees in public-facing business. As a result, we believe mask policy for all individuals in public space can be used to effectively represents the policy impact.

It is worth to note that, the mask_public is in date format, which has value of zero or a actual date when the policy is enforced. According to the documentation, zero represents "the absence of an order or directive", which can be interpreted as the policy is not enforced by the state explicitly. For linear regression, the date values in mask_public column are transformed as value 1, so that we can distinguish whether states has public mask policy or not. It is also noted that by transforming the variable, some important information will be lost because the actual date (early or late) to enforce the policy can also have an impact to the case rate. However, it is difficult to measure the time effect as different states that enforced the public mask policy may depend on related situations or different thresholds.

Firstly, "mask_public_bool" column is created from "mask_public", with 1 represents that the state has a pubic mask policy, and 0 means there's no explicit public mask policy from the state. The summary shows that there are 35 states that enforced public mask mandate, and 16 states that had no public mask policy. The ratio looks reasonable for the analysis considering sample size of 51 as there is decent number of samples for each state policy type.

```
df <- df %>%
  mutate(
    mask_public_bool = case_when(
      mask_public == 0 ~ 0,
      !(mask_public == 0) ~ 1
      )
  )

cat('Count for states that enforced mask mandate in public: ',
    length(df$mask_public_bool[df$mask_public_bool==1]))
```

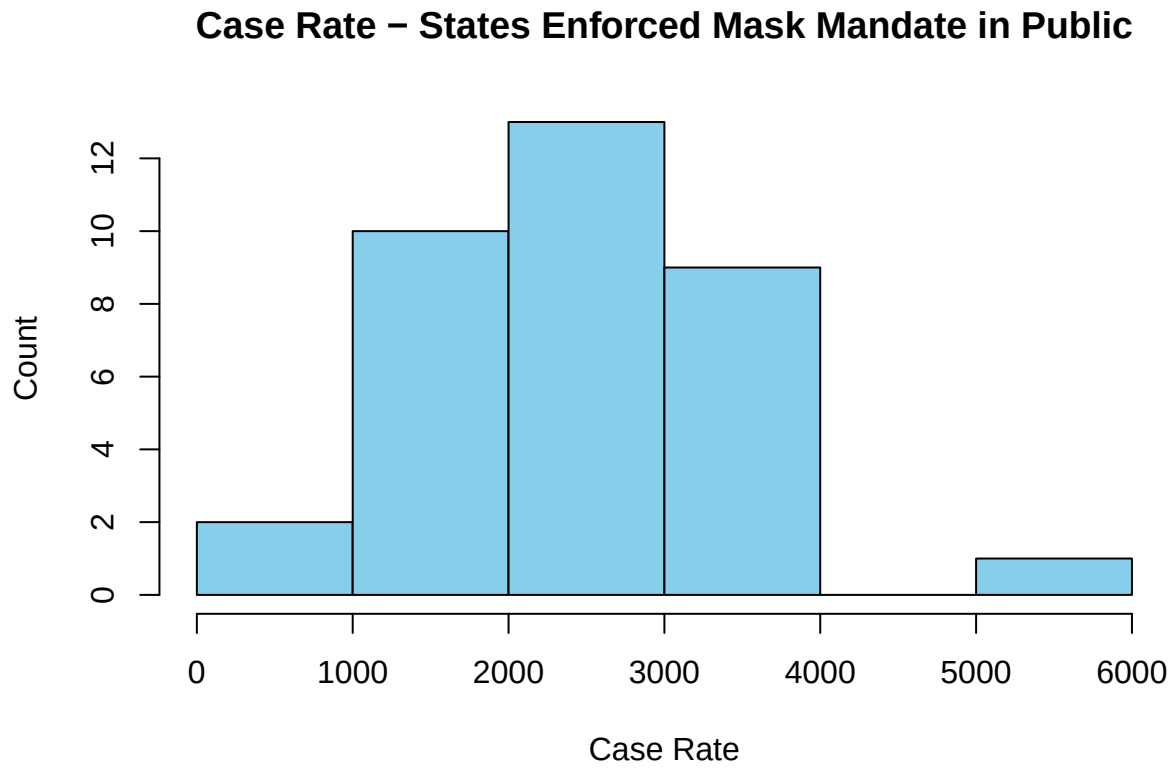
```
## Count for states that enforced mask mandate in public:  35
```

```
cat('Count for states did not enforced mask mandate in public: ',
    length(df$mask_public_bool[df$mask_public_bool==0]))
```

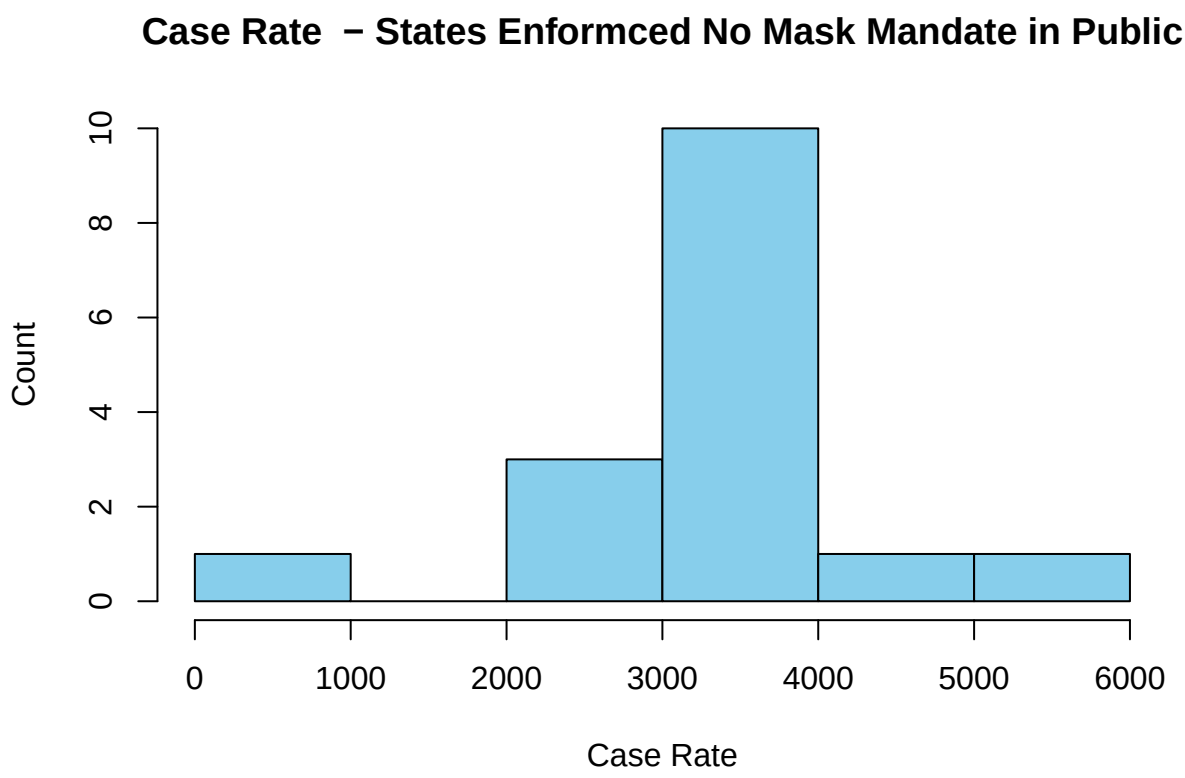
```
## Count for states did not enforced mask mandate in public:  16
```

Histograms below that shows the distribution of states that has public mask policy or not. Both distributions are not very normal with relatively small skews toward both tails.

```
hist(df$case_rate_100k[df$mask_public_bool == 1], xlab='Case Rate', ylab='Count',
     col='skyblue', main='Case Rate - States Enforced Mask Mandate in Public')
```

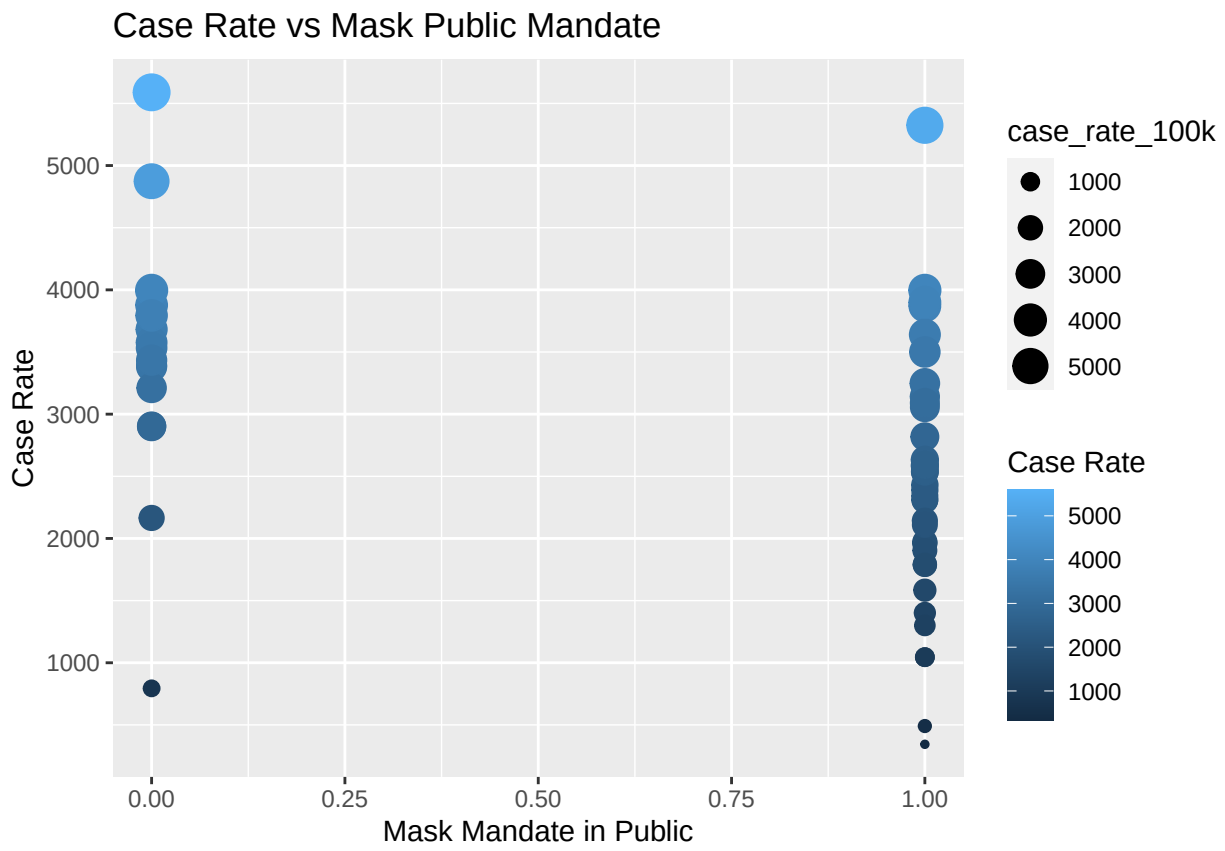



```
hist(df$case_rate_100k[df$mask_public_bool == 0], xlab='Case Rate', ylab='Count',
     col='skyblue', main='Case Rate  - States Enformced No Mask Mandate in Public')
```



A plot that shows the relationship between different policies vs case rate.

```
df %>%
  ggplot(aes(x = mask_public_bool, y = case_rate_100k, color = case_rate_100k)) +
  geom_point(aes(size=case_rate_100k)) +
  labs(
    title = 'Case Rate vs Mask Public Mandate',
    x = 'Mask Mandate in Public',
    y = 'Case Rate',
    color = 'Case Rate'
  )
```

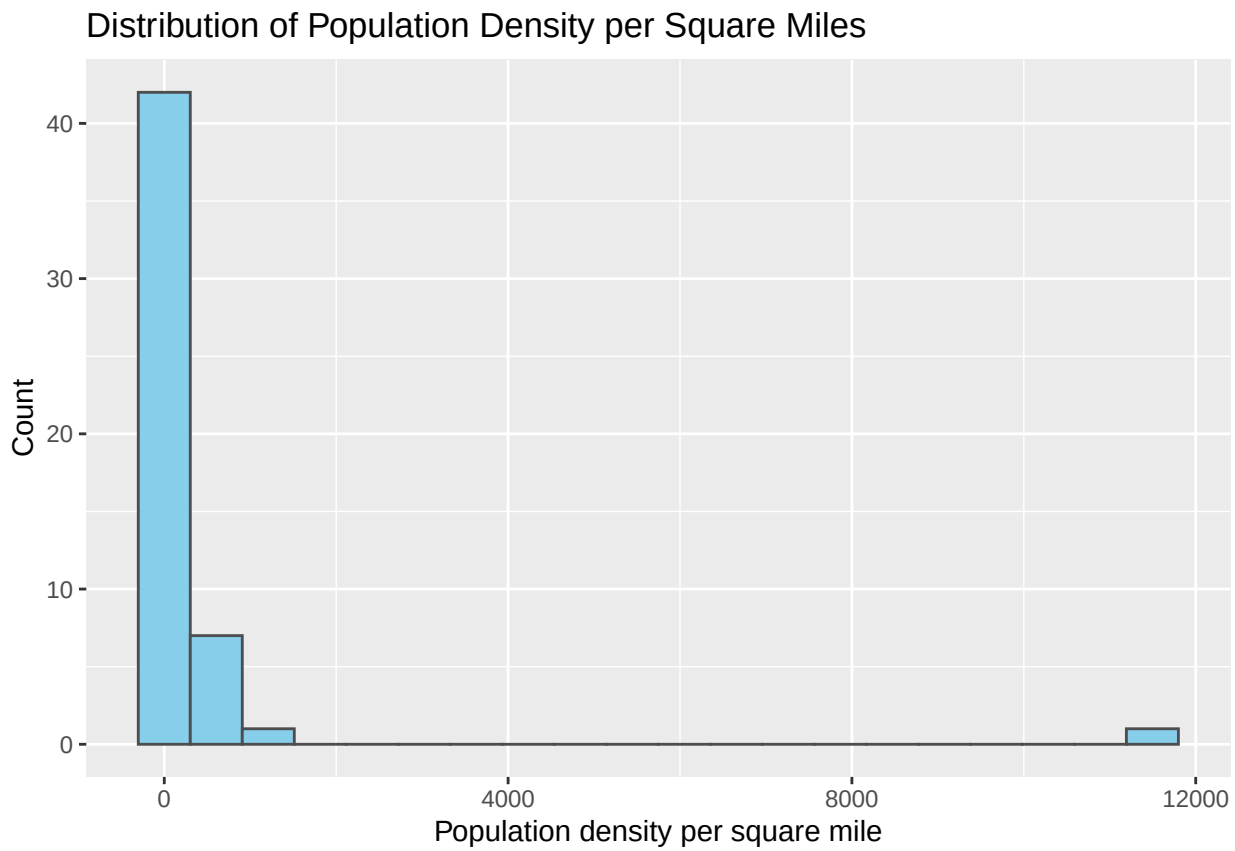


@Frank, this was my previous analysis of the POPULATION DENSITY variable, feel free to use any of this

Next, the distribution of the population density variable is examined.

```
# Plot the distribution in a histogram
histogram_of_pdensity <- df %>%
  ggplot(aes(x = population_density)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', bins = 20) +
  labs(
    title = 'Distribution of Population Density per Square Miles',
    x = 'Population density per square mile', y = 'Count')

histogram_of_pdensity
```



As can be seen from the histogram, although most of the population density is concentrated in the 0 to 1500 range, there are some grouping of outliers that are very far from this concentration. When an analysis is performed, it can be seen that there is only one data sample that is the outlier, which is D.C. with a value of 11,496. This is given due to the fact that D.C. is a district that solely consists of a large city, as mentioned as a possibility in the introduction. Given that this causes the data to be skewed, the logarithm is taken to scale the variable. Alternatively, the data point could have been droppped from the sample, but given the already small sample size, it was determined that a better approach would be to keep it within the sample. Once this transformation was performed, it is shown that there is a relatively normal distribution of population densities (see figure below).

```
# Find the outlier data points
outliers <- subset(df, population_density > 4000)
paste(outliers$State, '=', outliers$population_density)
```

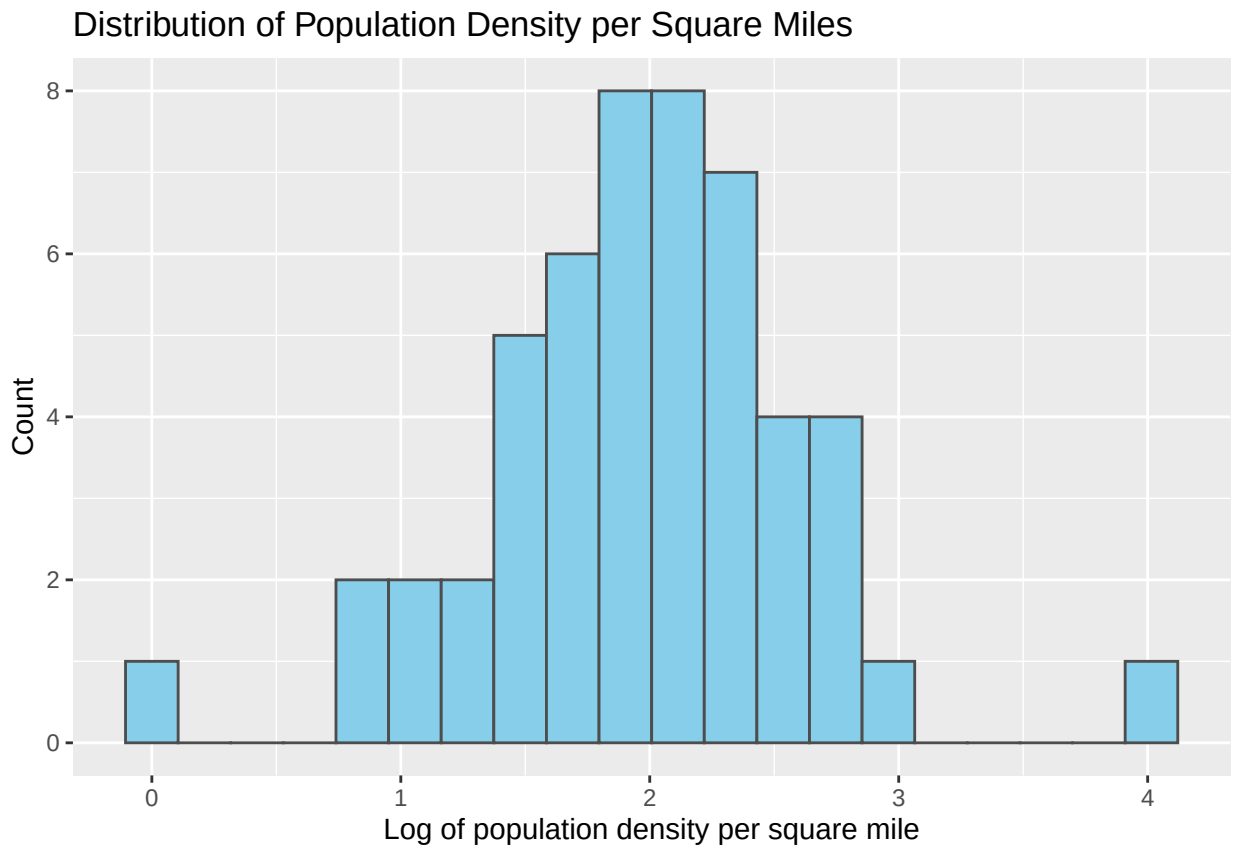
```
## [1] "District of Columbia = 11496.81"
```

```
# Transform the variable by taking the logarithm and assign it to a new variable
df <- df %>%
  mutate(l_population_density = log10(population_density))

# Plot the new distribution in a histogram
histogram_of_pdensity <- df %>%
  ggplot(aes(x = l_population_density)) +
  geom_histogram(fill = 'skyblue', color = 'grey30', bins = 20) +
  labs(
    title = 'Distribution of Population Density per Square Miles',
```

```
    x = 'Log of population density per square mile', y = 'Count')

histogram_of_pdensity
```



With the appropriate variables transformed, a plot is created to show the relationship between them.

```
df %>%
  ggplot(aes(l_population_density, case_rate_100k)) +
  geom_smooth(se = FALSE) +
  geom_point() +
  labs(
    title = 'COVID Case Rate due to Population Density',
    x = 'Log of population density per square miles',
    y = 'COVID case rate per 100,000'
  )
```

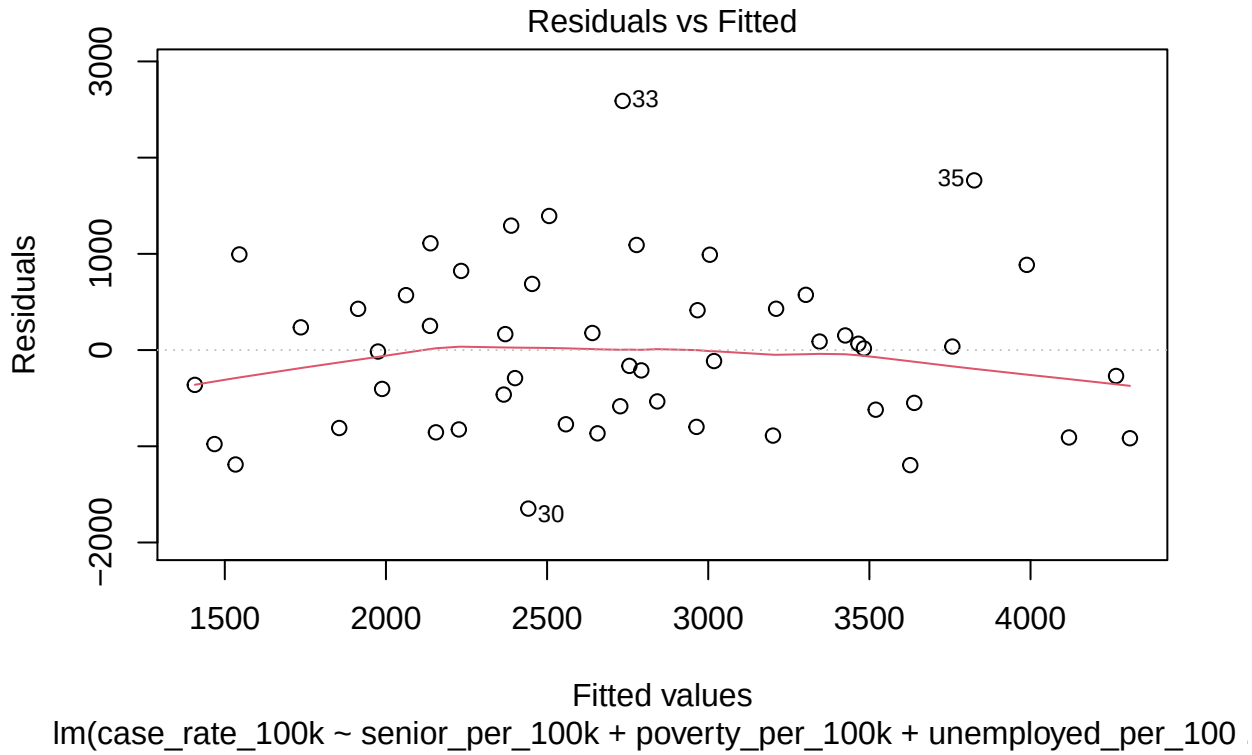
Model 3 Regression Next, the linear model is created from model 2 with the mask policy variables added for the regression. The regression coefficient shows that Mask_public_bool variable is highly significant. The result suggests that public mask mandate policy has postive effect in bringing the case number down.

```
model3 <- lm(case_rate_100k ~ senior_per_100k
             + poverty_per_100k
             + unemployed_per_100k
            + mask_public_bool
            + log(population_density), data = df)
coeftest(model3, vcov = vcovHC)
```

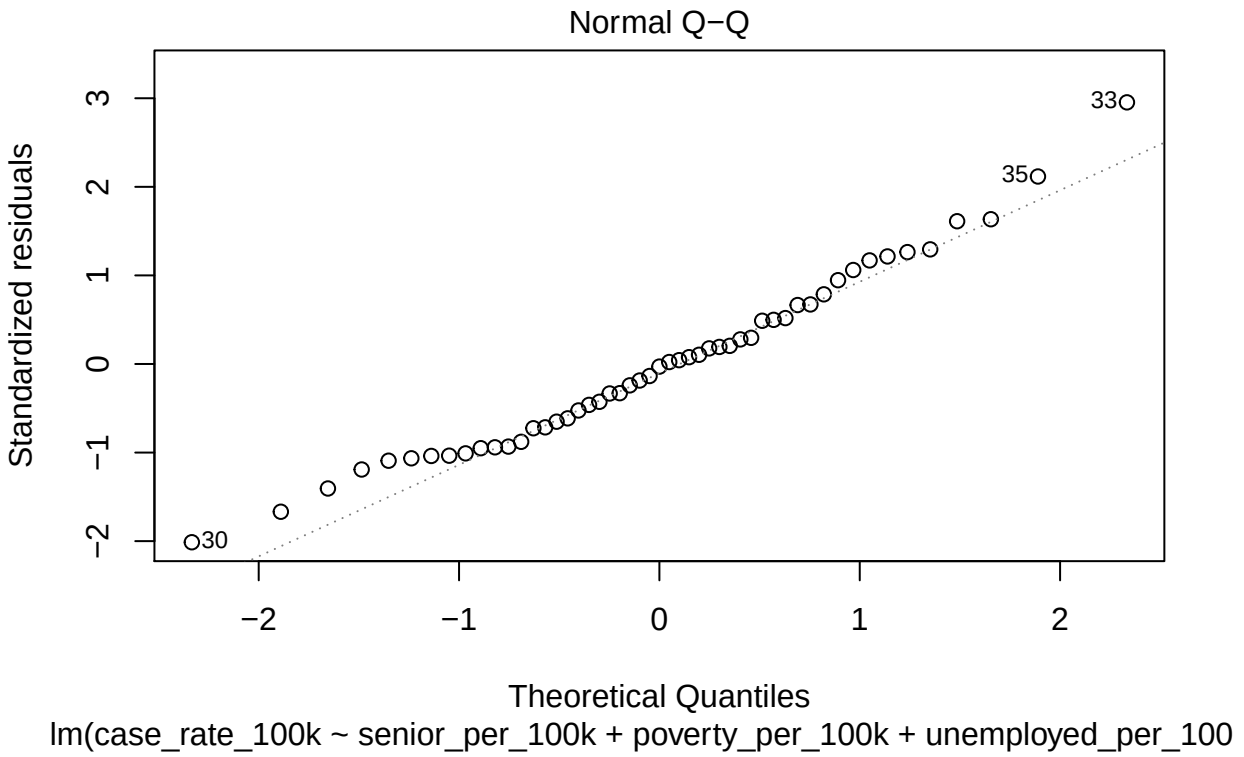
```
##
## t test of coefficients:
##
##                            Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)              6.1462e+03  1.4059e+03  4.3718 7.198e-05 ***
## senior_per_100k         -2.4185e-01  7.7714e-02 -3.1121  0.003223 **
## poverty_per_100k         1.9671e-03  6.0669e-04  3.2423  0.002236 **
## unemployed_per_100k     -3.6426e-03  1.6836e-03 -2.1636  0.035839 *
## mask_public_bool        -9.4564e+02  3.2972e+02 -2.8680  0.006267 **
## log(population_density)  9.3457e+01  1.0535e+02  0.8871  0.379717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CLM assumptions check for Model 3:

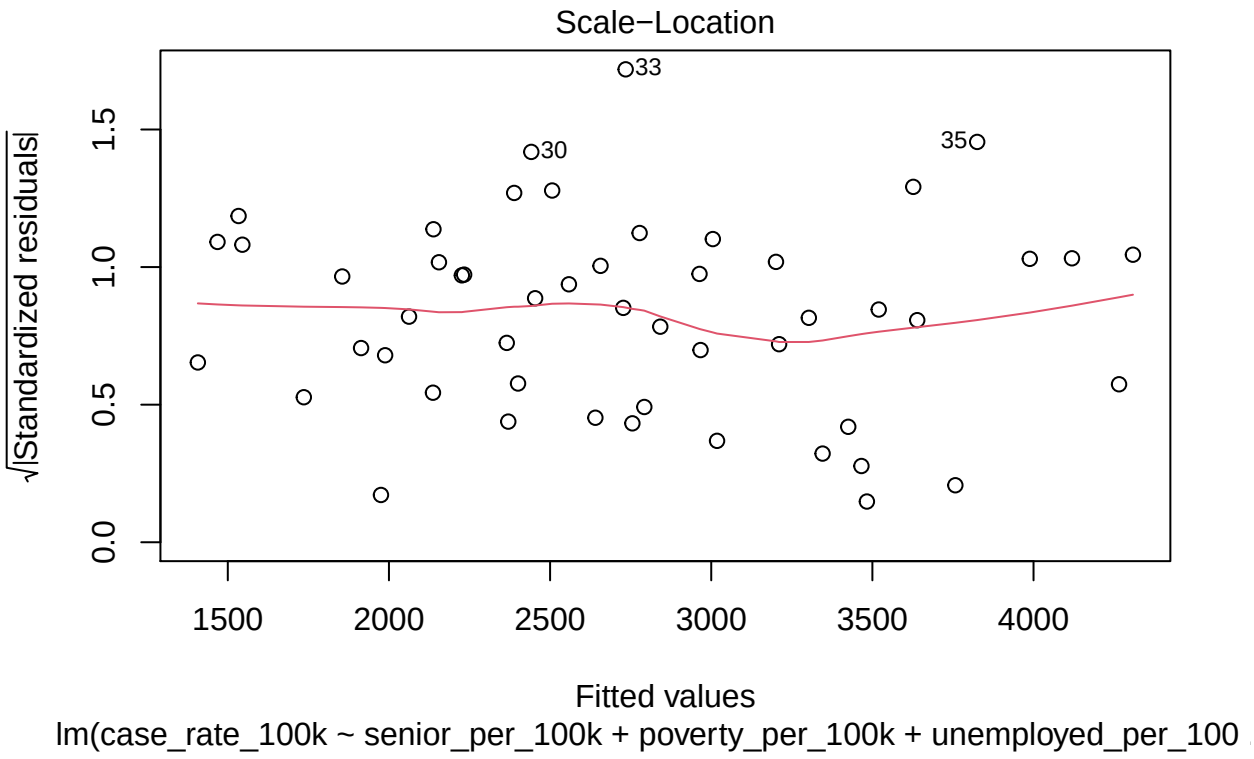
```r
#Zero Conditional Mean
plot(model3, which=1)
```



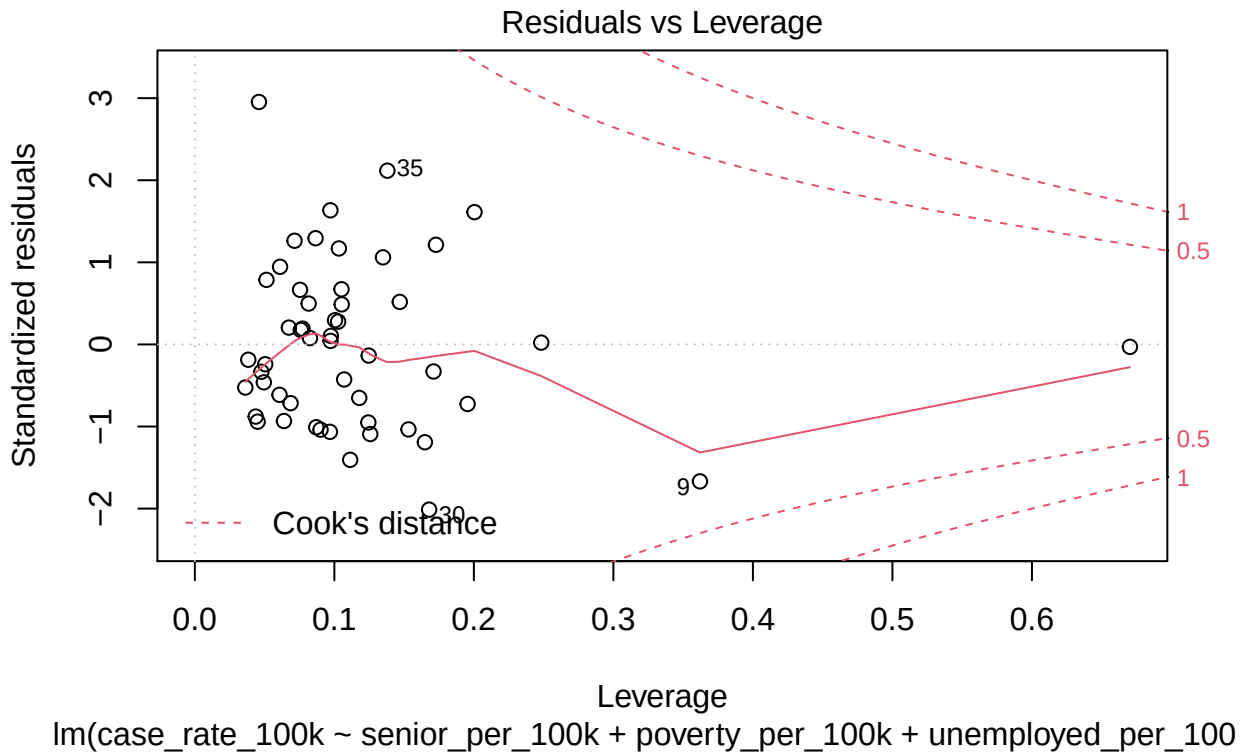

```r
#Normality of errors
plot(model3, which=2)
```




```
#Homoskedasticity
plot(model3, which=3)
```




```
#Cook distance
plot(model3, which=5)
```




```
#Collinearity
vif(model3)
```

```
##         senior_per_100k        poverty_per_100k     unemployed_per_100k
##                1.079612                1.882855                2.268679
##        mask_public_bool log(population_density)
##                1.208219                1.282422
```

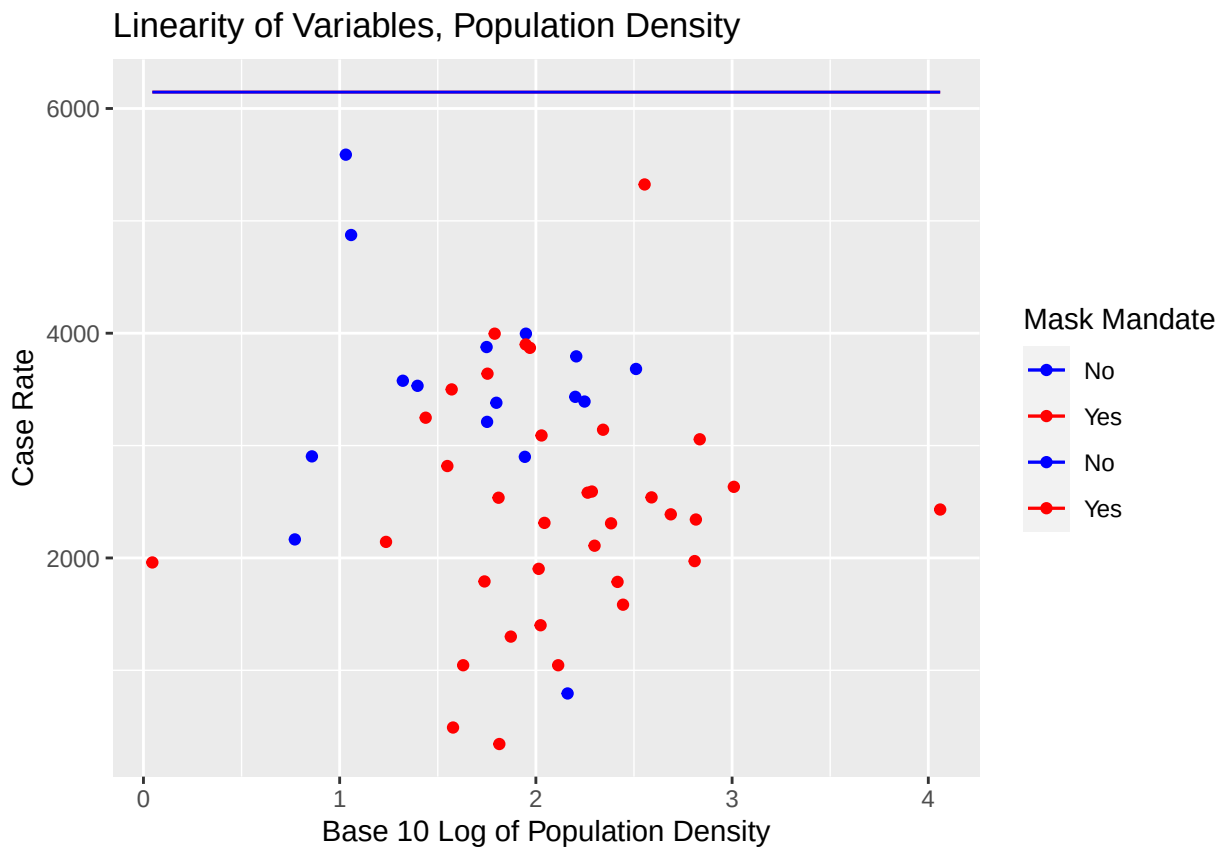
3. Limitations of the Models

Firstly, the most general requirement for the linear model is that data points be independent and identically distributed (IID). While for the most part the events of one state will primarily affect that single state, travel by individuals spreading the virus across state lines is an inevitable occurence. This may cause the spread of a virus within one state to be influenced by those around it, or generally by its position in the country. In addition, all states are involved in the same market economy system of the country to varying degrees, even though there are economic performance divisions across the states. Likewise, certain regions of the country may also have similar racial compositions due to a shared history across general geographic regions. While this is acknowledged to occur, the effect is likely minor compared to the policy data collected state by state which would affect the spread of the virus. Therefore it is relatively safe for this assumption to be met.

Secondly, the process being modeled must be described by a linear conditional expectation function of the variables which are included, in order for the linear regression to be valid. This can be assessed after the fact by visualizing the model over the range of each variable alongside the data, or also by visualizing the residuals over the range of each variable. For the purpose of this assumption and other model or data related assumptions, only the Model 3-a will be used as it is the most inclusive.

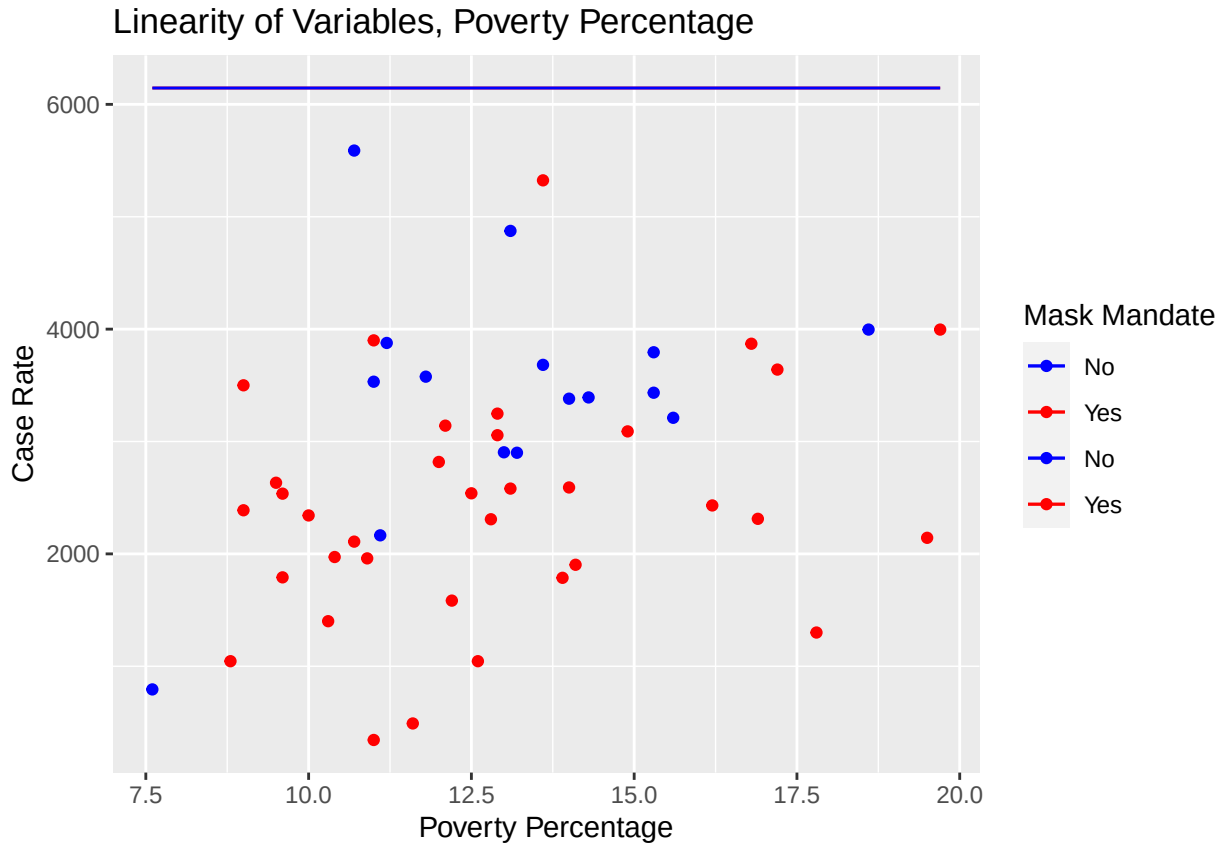
```
avg_pop <- log10(mean(df$population_density))
avg_pov <- mean(df$poverty_pct)
avg_mask <- mean(df$mask_public_bool)
y_pop <- model3$coefficients[1] + model3$coefficients[2]*log10(df$population_density) + model3$coefficie
y_pov <- model3$coefficients[1] + model3$coefficients[2]*avg_pop + model3$coefficients[3]*df$poverty_pc
y_pop_nomask <- model3$coefficients[1] + model3$coefficients[2]*avg_pop + model3$coefficients[3]*avg_po
y_pov_nomask <- model3$coefficients[1] + model3$coefficients[2]*avg_pop + model3$coefficients[3]*df$pov

df %>%
  ggplot() +
  geom_point(aes(x = log10(population_density), y = case_rate_100k,
                 group = factor(mask_public_bool), color = factor(mask_public_bool))) +
  geom_line(aes(x = log10(population_density), y = y_pop, color = "red")) +
  geom_line(aes(x = log10(population_density), y = y_pop_nomask, color = "blue")) +
  scale_color_manual(labels = c("No", "Yes", "No", "Yes"),
                     values = c("blue", "red", "blue", "red")) +
  labs(
    title = 'Linearity of Variables, Population Density',
    x = 'Base 10 Log of Population Density',
    y = 'Case Rate',
    color = 'Mask Mandate'
  )
```




```
df %>%
  ggplot() +
```

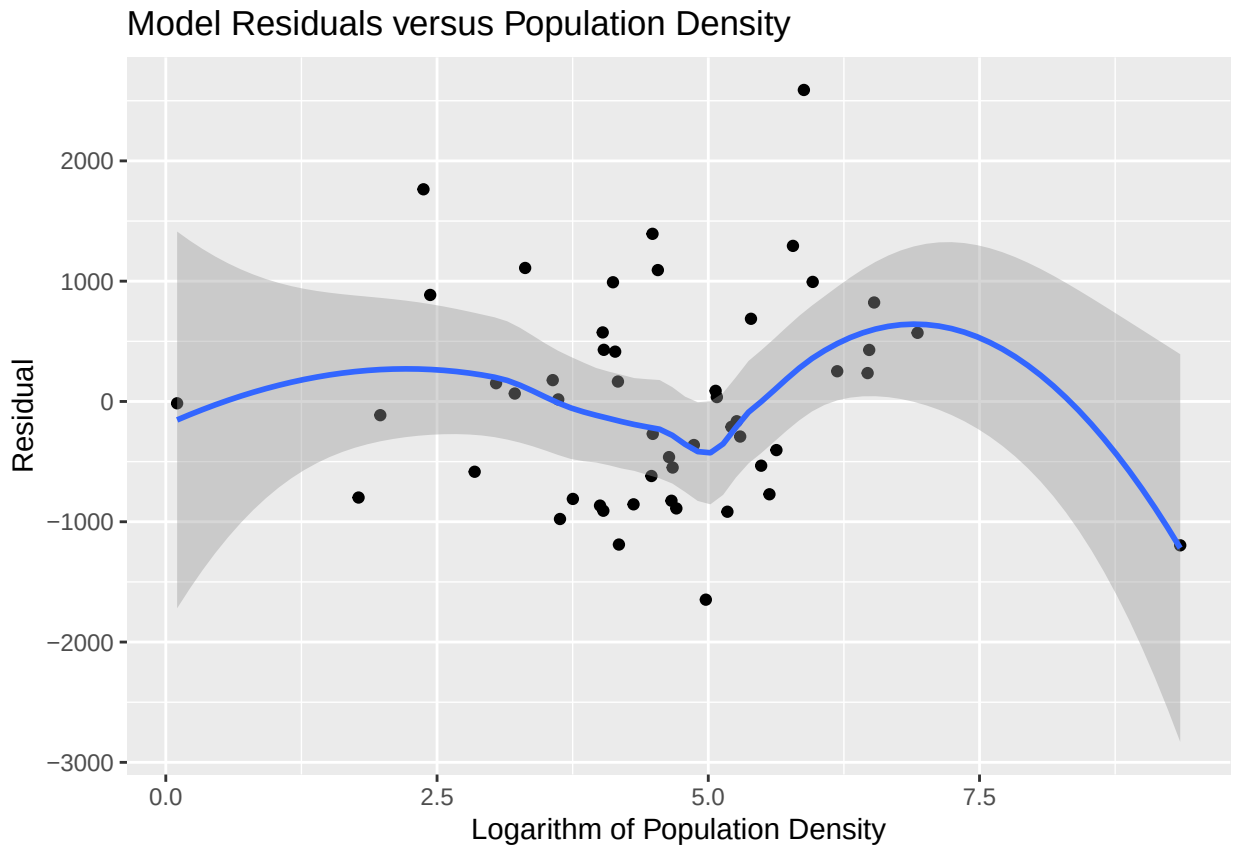
```
  geom_point(aes(x = poverty_pct, y = case_rate_100k, group = factor(mask_public_bool),
                 color = factor(mask_public_bool))) +
  geom_line(aes(x = poverty_pct, y = y_pov, color = "red")) +
  geom_line(aes(x = poverty_pct, y = y_pov_nomask, color = "blue")) +
  scale_color_manual(labels = c("No", "Yes", "No", "Yes"),
                     values = c("blue", "red", "blue", "red")) +
  labs(
    title = 'Linearity of Variables, Poverty Percentage',
    x = 'Poverty Percentage',
    y = 'Case Rate',
    color = 'Mask Mandate'
  )
```



Linearity for the three regressed variables is demonstrated above. As can be seen when mask policy and population density are varied, the data generally follows a linear trend. This is similar when the poverty percentage is varied, although at higher levels of poverty the presence of a mask mandate made the case rate unexpectedly low. It appears that possible a negative quadratice function may fit the data better, although it appears close enough to linear that this assumption may hold.

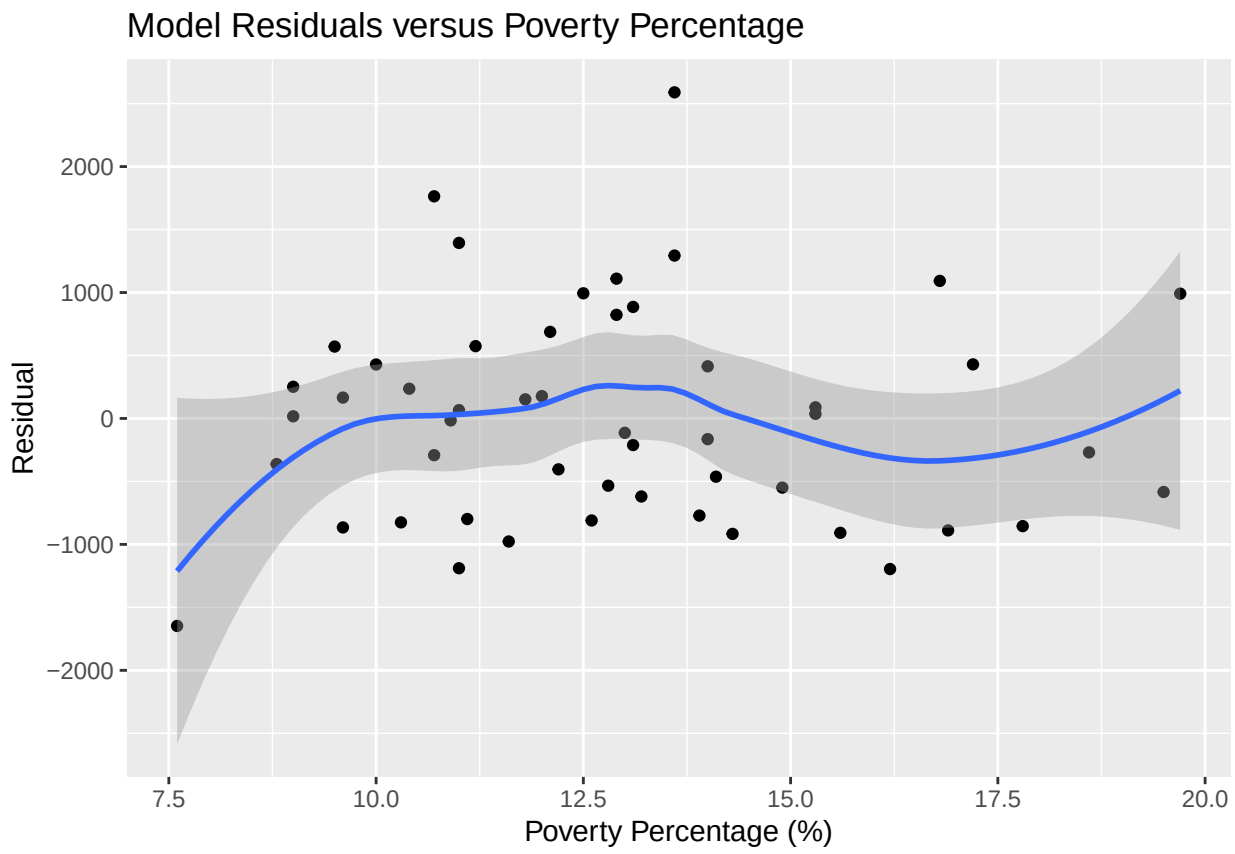
```
df %>% {
   ggplot(mapping = aes(x = log(df$population_density), y = model3$residuals)) +
      geom_point() + stat_smooth(se = TRUE) +
      ggtitle("Model Residuals versus Population Density") +
      xlab("Logarithm of Population Density") +
      ylab("Residual")
}
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
df %>% {
   ggplot(mapping = aes(x = df$poverty_pct, y = model3$residuals)) +
      geom_point() + stat_smooth(se = TRUE) +
      ggtitle("Model Residuals versus Poverty Percentage") +
      xlab("Poverty Percentage (%)") +
      ylab("Residual")
}
```

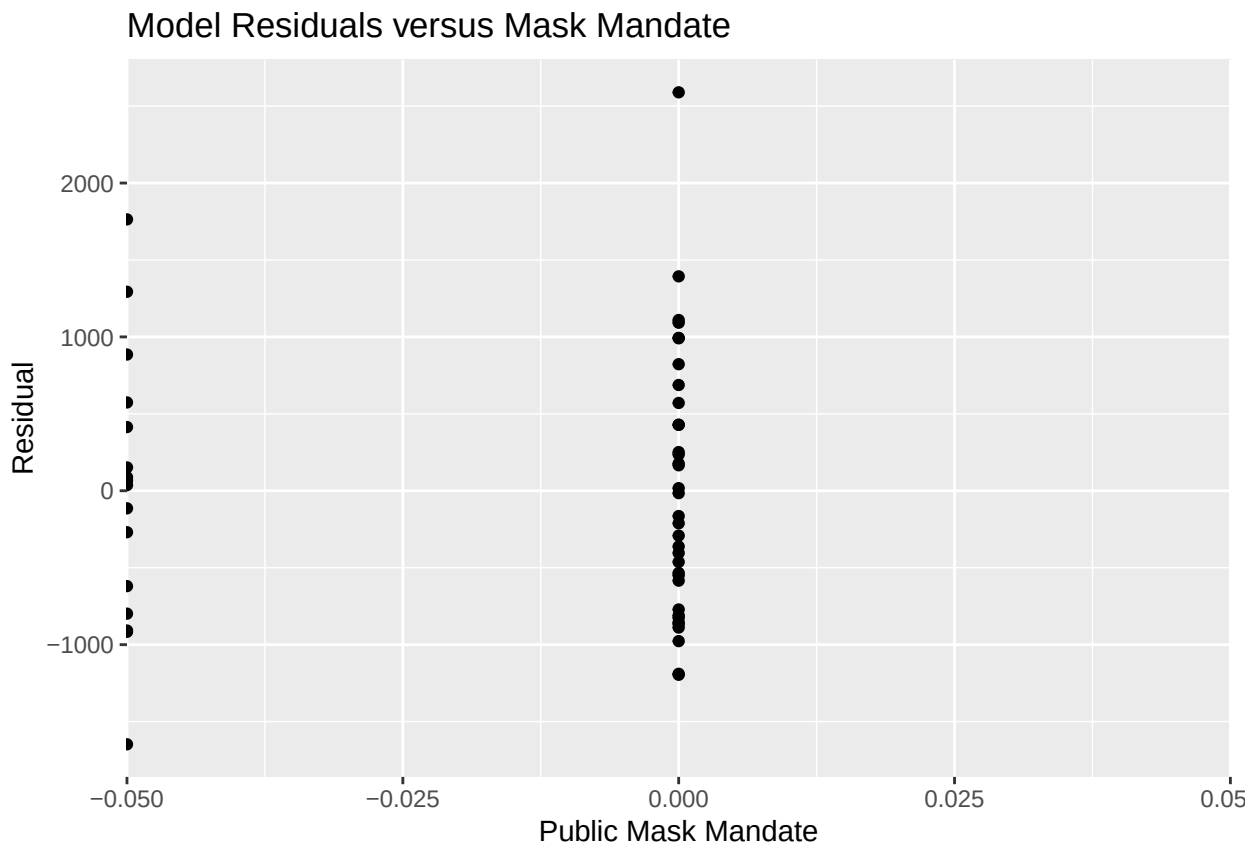
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
df %>% {
   ggplot(mapping = aes(x = log(df$mask_public_bool), y = model3$residuals)) +
      geom_point() + stat_smooth(se = TRUE) +
      ggtitle("Model Residuals versus Mask Mandate") +
      xlab("Public Mask Mandate") +
      ylab("Residual")
}
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## Warning: Removed 16 rows containing non-finite values (stat_smooth).
```



Thirdly, related to the relationship of the residuals to the data, homoscedasticity can be assessed by visualizing the size of residuals across the range of each variable. Alternatively, it can also be done via the Breusch-Pagan test to more objectively determine whether homoscedasticity is not present in the data. Using the first method with the plots above, the residuals to each of the four regressions appear to be evenly spread throughout the range of the x axis. Therefore, this assumption can also be reasonably assumed to be met. The normal distribution of these errors also appears to be well met by the linear models, with most data points close to their respective line causing most residuals to be centered around zero.

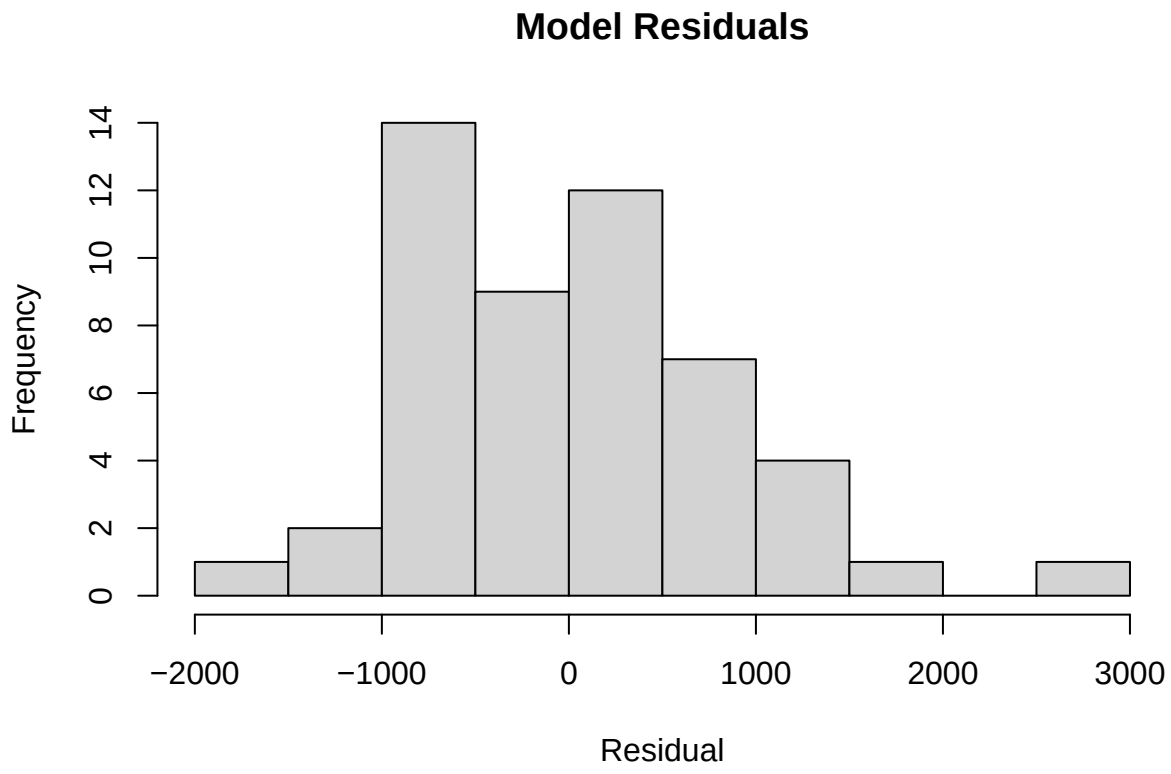
Perfect colinearity is automatically detected by the R regression algorithms, which drop potential variables if they have complete colinearity with another variable. Near perfect colinearity is more problematic to detect, but a common hint to its existence is large standard errors on colinear features. In Model 3-a, all of the standard errors for model coefficients are at least less than half of the value of the coefficient itself. Therefore, near-perfect colinearity can be assumed to not be taking place between these variables.

```
correlation_one <- cor(log(df$population_density), df$poverty_pct)
vif_one <- 1/(1-(correlation_one^2))
correlation_two <- cor(log(df$population_density), df$mask_public_bool)
vif_two <- 1/(1-(correlation_two^2))
correlation_three <- cor(df$poverty_pct, df$mask_public_bool)
vif_three <- 1/(1-(correlation_three^2))
```

A more methodical approach for evaluating if there is too high of colinearity between features is through the calculation of the Variance Inflation Factor (VIF). For the three variables in Model 3-A, the VIF values between the three combinations of the three variables are 1.0004677, 1.1029272, and 1.0017876. Generally, as long as the VIF is below four there can be assumed not enough colinearity to violate the assumption as is the case in this model.

Finally, the linear model is also assumed to generate residual error that is normally distributed away from the predicted values. A simple and straightforward way to assess this is through a histogram of the residuals of the model.

```
hist(model3$residuals, main = "Model Residuals", xlab = "Residual")
```



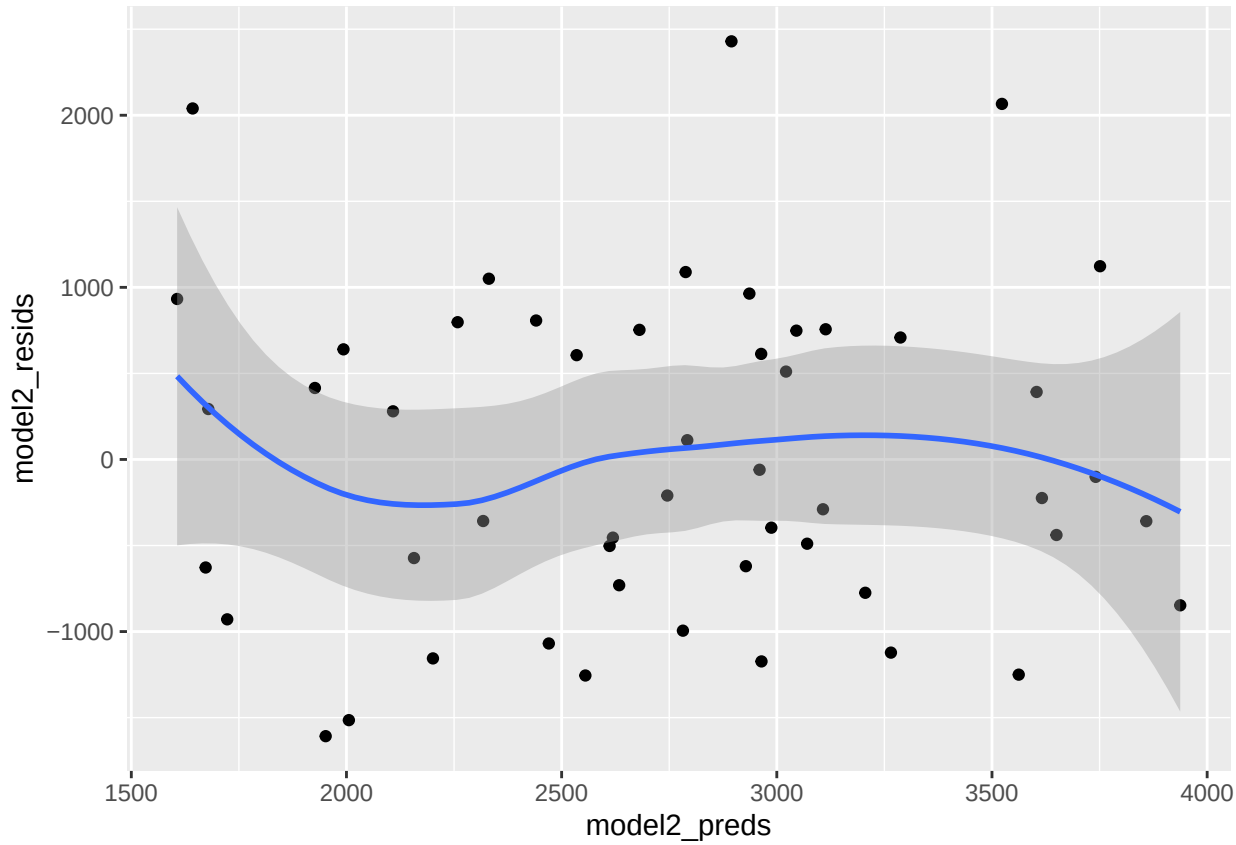
As seen above, this assumption is generally well met.

Model 2 limits

i. IID Sampling (Discussed in the common section)

ii.Linear Conditional Expectation Assessing non-linear in higher-dimensional space: look at the predicted vs. residuals of the model.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



####iii.No Perfect Collinearity First, check if any variables were dropped by R. No coefficient is missing, which mean there is no perfect collinearity.

```
model2$coefficients
```

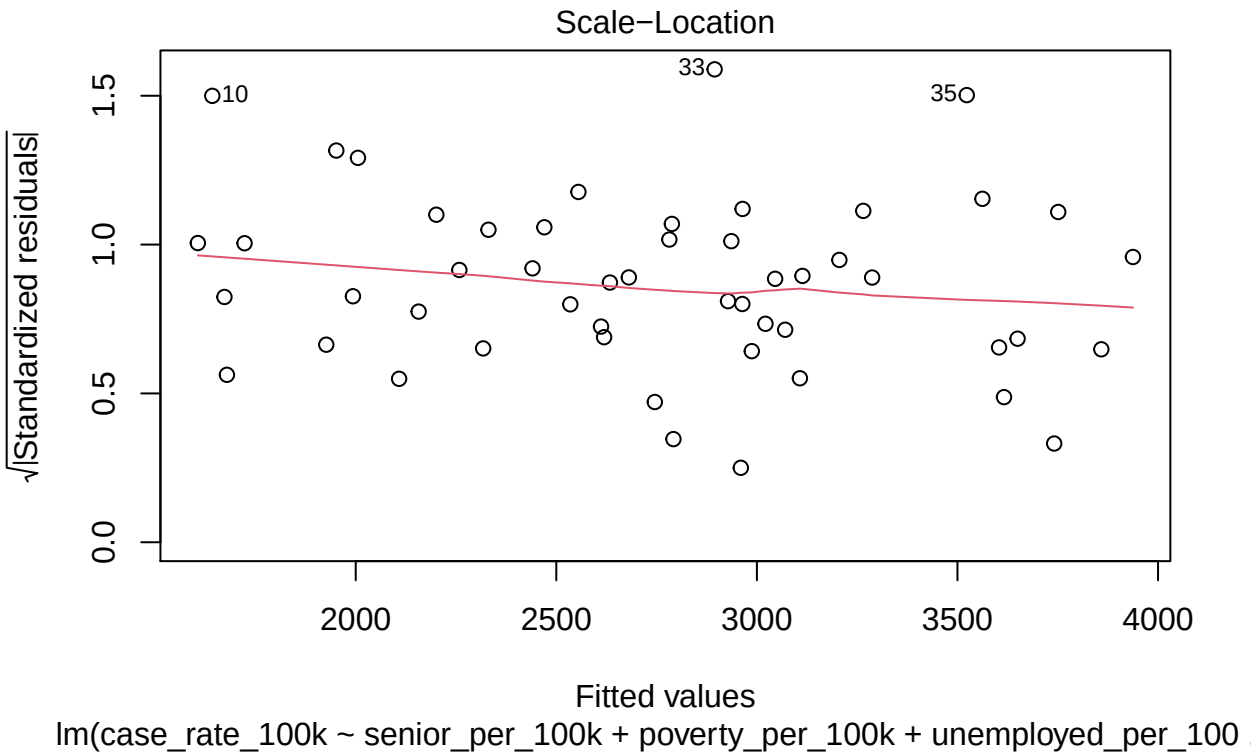
```
##          (Intercept)     senior_per_100k    poverty_per_100k unemployed_per_100k
##         5.869330e+03       -2.309704e-01        2.317495e-03       -4.861906e-03
```

In addition, all the variance inflation factors are less than 4, which doesn't indicate the exisitence of collinearity.

```
vif(model2)
```

```
##     senior_per_100k    poverty_per_100k unemployed_per_100k
##            1.065676            1.655600            1.692290
```

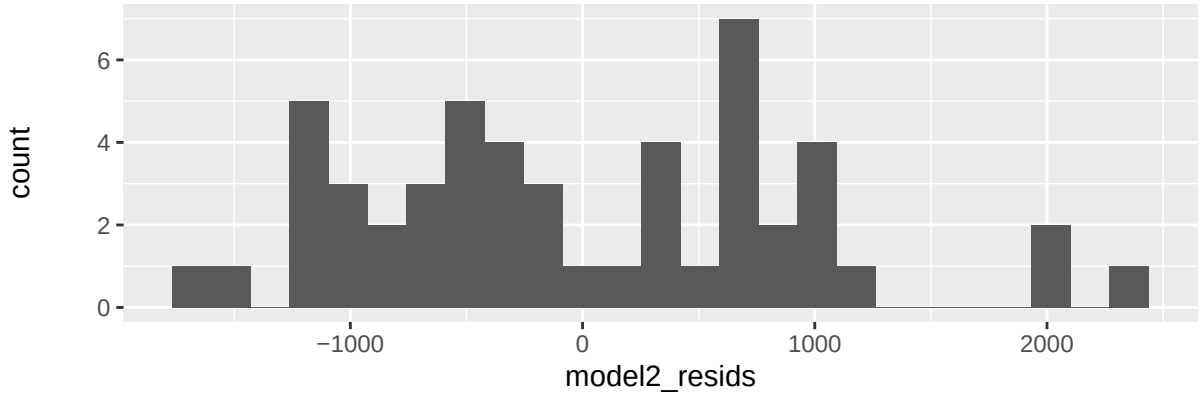
####iv.Homoskedastic Errors 10 is Florida, 33 is New York, 35 is North Dakota

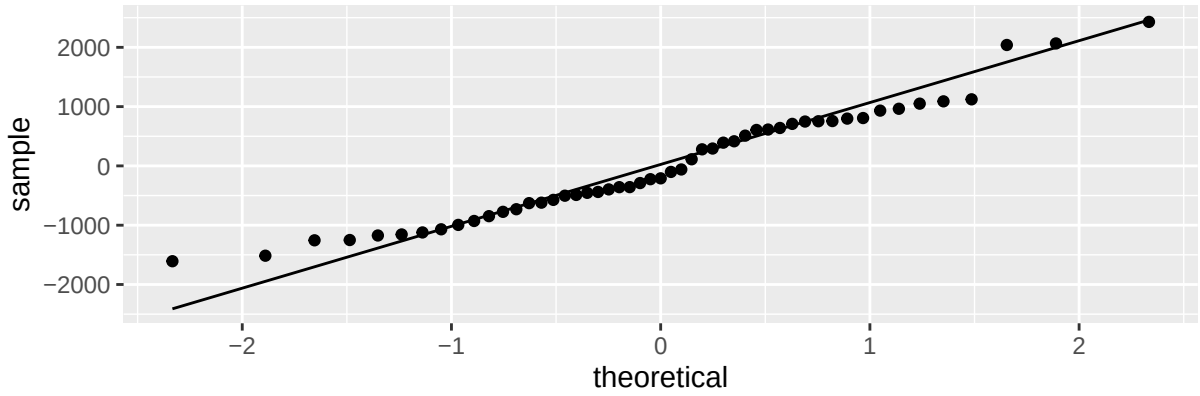


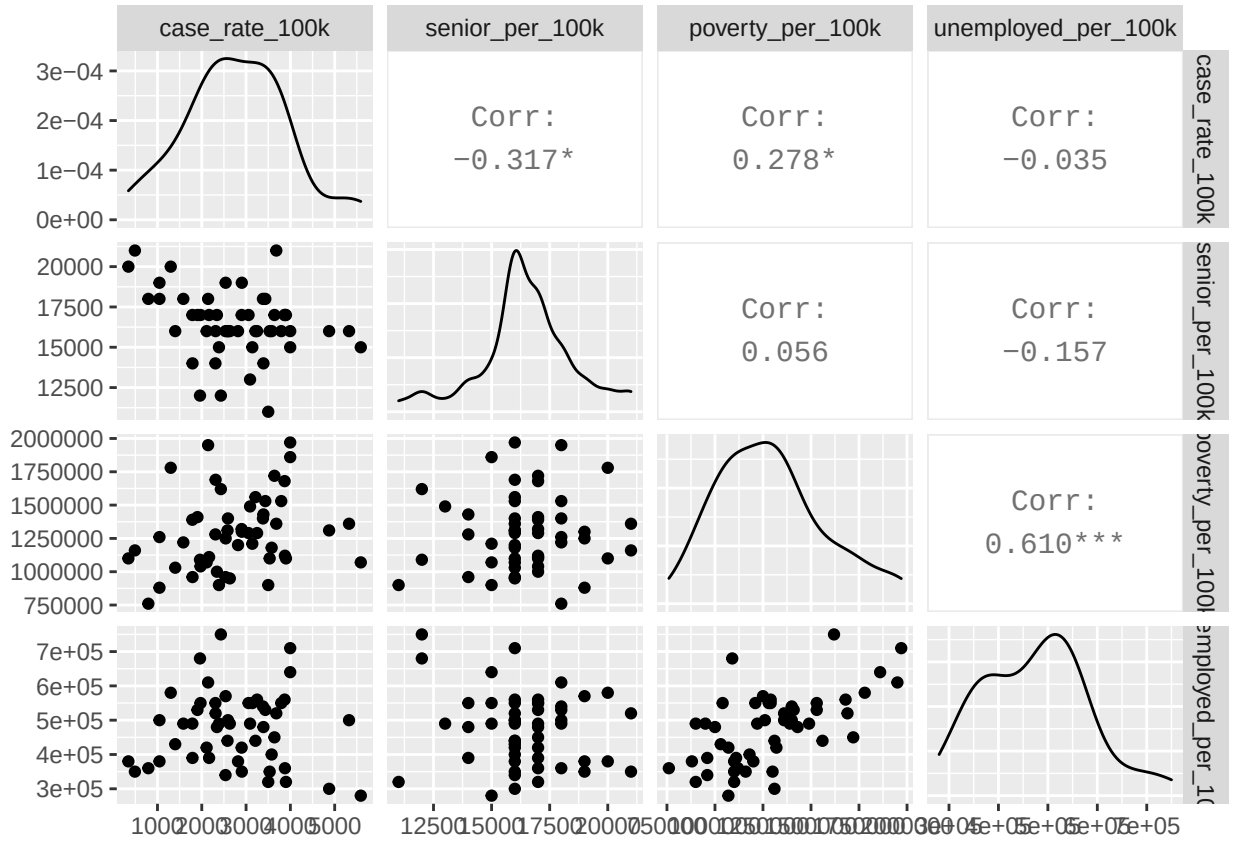

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 4.5045, df = 3, p-value = 0.2119
```

####v.Normally Distributed Errors Check the normality of error distribution based on the histogram of residuals and the qqplot.

```
## geom_bar: na.rm = FALSE, orientation = NA
## stat_bin: binwidth = NULL, bins = NULL, na.rm = FALSE, orientation = NA, pad = FALSE
## position_stack
```

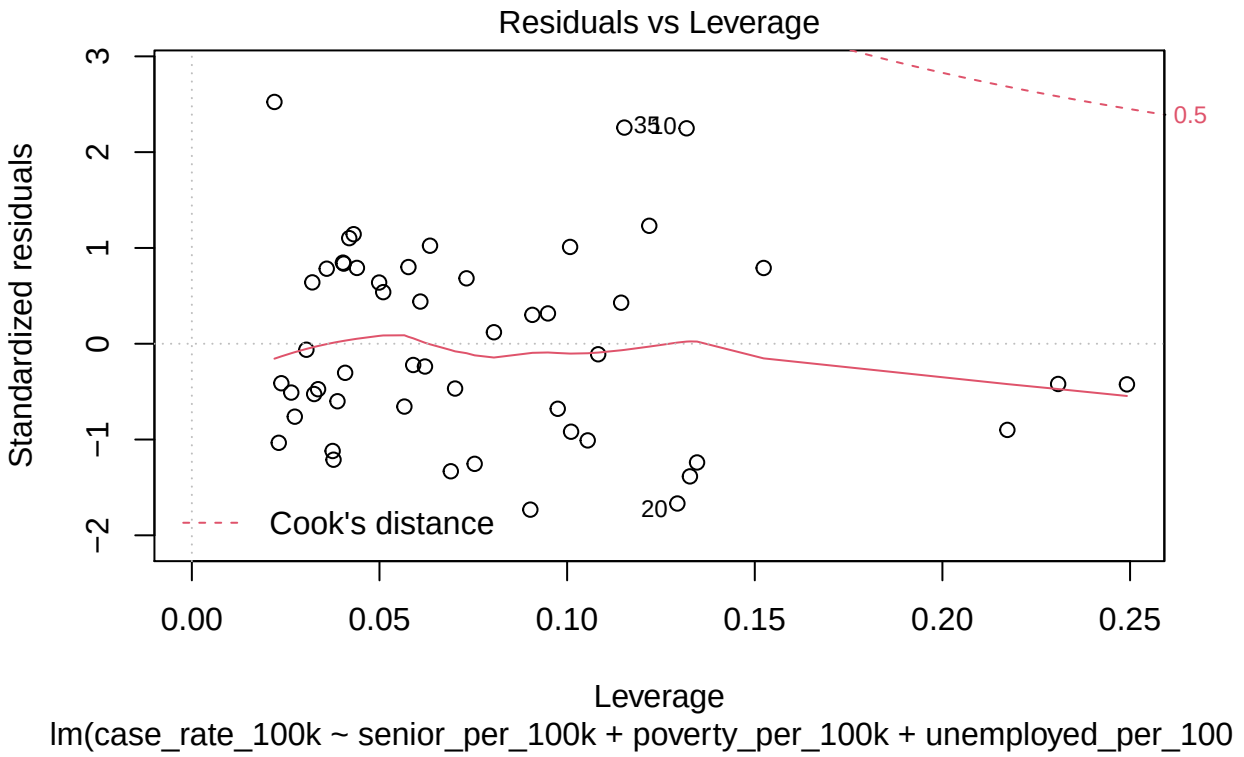

count
6
4
2
0
−1000
0
1000
2000
model2_resids


sample
2000
1000
0
−1000
−2000
−2
−1
0
1
2
theoretical



GGPairs

Cook Distance 10 Florida, 20 Maine, 35 North Dakota



4. Regression Table

With the analysis for the selected variables, the three corresponding models, and the respective diagnostics of the CLM assumptions complete, the models can then be included for comparison in a regression table. Model 1, model 2, and model 3 are included in the regression table below,.

```
se.model1 = coeftest(model1)[ , "Std. Error"]
se.model2 = coeftest(model2)[ , "Std. Error"]
se.model3 = coeftest(model3)[ , "Std. Error"]

stargazer(model1, model2, model3, type = "text",
          se = list(se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 1: The effect of population demographics and mask policy on COVID-19 case rate
```

```
##
## Table 1: The effect of population demographics and mask policy on COVID-19 case rate
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                              case_rate_100k
##                                  (1)                 (2)                 (3)
## --------------------------------------------------------------------------------
## senior_per_100k               -0.174**            -0.231**            -0.242***
```

```
##                                (0.064)             (0.069)              (0.064)
##
## poverty_per_100k                                  0.002***             0.002**
##                                                    (0.001)              (0.001)
##
## unemployed_per_100k                               -0.005**             -0.004*
##                                                    (0.002)              (0.002)
##
## mask_public_bool                                                      -945.638**
##                                                                        (297.729)
##
## log(population_density)                                                93.457
##                                                                        (96.285)
##
## Constant                     5,613.667***         5,869.330***         6,146.213***
##                               (1,273.196)          (1,362.851)          (1,273.196)
##
## ---------------------------------------------------------------------------------------
## Observations                      51                   51                   51
## R2                               0.101                0.311                0.439
## Adjusted R2                      0.082                0.267                0.376
## Residual Std. Error     1,088.808 (df = 49)   973.292 (df = 47)     897.547 (df = 45)
## F Statistic             5.477* (df = 1; 49) 7.059*** (df = 3; 47) 7.034*** (df = 5; 45)
## =======================================================================================
## Note:                                                  *p<0.05; **p<0.01; ***p<0.001
```

When observing the table above, three elements stand out.

Firstly, there is a difference in the calculated standard error between the coefficient test run on the linear models and the regression table. Taking a look at the standard error for the poverty percentage coefficient, the value in the table returns a significant result with a standard error of 55.125, as opposed to the insignificant result of a standard error of 57.684. This follows for model 3, with the table showing a significant result with standard error of 50.872, as opposed to an insignificant result of a standard error of 59.583.

Secondly, following the first observation, when designating the poverty percentage variable as significant, the regression table suggests that both covariate variables (including the mask mandate variable) are indeed significant with the COVID case rate. For the poverty percentage effect, the coefficient highlights that there is a positive correlation between the poverty percentage and case rate (as poverty percentage increases, so does case rate). The practical significance of this shows that for every percentage increase of population below the federal poverty line, there is an increase of roughly 112 cases per 100k for model 2, and roughly 104 cases per 100k for model 3. As for the mask mandate variable, the coeeficient highlights a negative correlation between it and the case rate (if the mask mandate is implemented, the case rate decreases). The practical significance is such that if the mask mandate has been implemented, the case rate per 100k will decrease by roughly 1000 cases.

Thirdly, we notice that the R2 value increases by a large value for every additional variable that is included within the model. Initially, we start with a very low R2 value, suggesting that almost no variance is explained in model 1 and that the model is a poor fit and predictor for the data. However, although still low, there is close to a 9 times increase in the amount of variance that is explained in model 2. Finally, in model 3, the R2 value increases by roughly 2.75 times the previous value, and a decent amount of the variance can be seen to be explained by the model. The adjusted R2 that accounts for the increase in number of variable terms in the model also correspondingly increases as the covariates are added. It also holds a negative value for model 1, signifying that there is no variance that can be explained by the model, further highlighting the poor fit.

As a note, it should be stressed that the models and variables included in this regression table may change

when the final report is created to provide a more descriptive analysis of the original research question.

5. Omitted Variables

Omitted variables from the model may affect both the internal specifications of included variables as well as the general applicability of the model to explain the output variable. For this analysis, some of these variables are within the dataset but not used while others are hypothetical.

5-1: Urban vs Rural Population Ratio

One of the variables that was hypothesized as being potentially influential is a ratio of a state's urban to rural population. Currently, the dataset contains population information as averaged across each state as a whole. While population density is potentially useful, it was found to not be significant in the models. This may be due to states which have large urban centers, where population density woukd be high, losing their impact due to also containing large areas of land that is sparsely populated. One example of this would be New York, which contains several of the most densely populated counties in the country, but also large rural areas.

An urban to rural ratio would be explanatory of the population density variable as a whole, where a higher urban to rural ratio would result in a higher statewide population density generally. With a greater fraction of urban residents, people would also be less likely to be able to social distance as effectively, increasing the per capita case rate. Therefore, the coefficient for population density in the original model would be larger and more positive than a model where this ratio is included, resulting in a bias that is positive and away from zero.

5-2: Median Income

Knowledge of the median income for a state would give an indication of the relative job status for a working person in that state. This would be important for the case rate to understand about how many people work in low paying jobs that are more frequently in close contact with others (such as the service industry, gig industry, etc.). Higher median income may imply that more workers are able to function remotely with greater social distancing. A higher median income would then be negatively related to the covid case rate. It would also be explanatory in a way that makes the poverty percentage variable less impactful, where either of these variables may be considered as proxies for one another. Because inclusion of the median income variable would led to a lower magnitude poverty percentage variable, and that the coefficient of the poverty percentage variable is positive, this would cause a positive omitted variable bias away from zero.

5-3: Gender Ratio

The ratio of gender between women and men may impact the covid case rate in that men are more suscepitble to viral infections and generally have lower life expectancies from lifestyle differences. Greater susceptibility may relate to the case rate in that men may more often become infected when exposed to COVID under similar circumstances as a women. Their shorter life expentancy may also impact the case rate in that more COVID cases may result in greater or more serious symptoms, with less occurances of an infection that is asymptomatic or goes undetected. In either of these cases, a higher fraction of men among the population would lead to an overall higher covid case rate.

This variable could also be partially explained by the senior percentage variable, because of the increase of women in gender ratios generally as people get older. Therefore, introducing the gender ratio variable into the model would lower the value of the coefficient to the senior percentage variable to be less negative. This would be a negative omitted variable bias away towards zero.

However, the magnitude of this change is likely to not be large due to other causal mechanisms between seniors and covid case rate. While men may more frequently present COVID symptoms, older people in general are also more susceptible to the virus due to their physical condition and would also present more symptoms than non seniors. When comparing the two, generally seniors are considered as more at-risk of COVID complications than men across all ages, and so the senior percentage would still have a greater impact on the overall covid case rate than the gender ratio.

5-4: Mask Use Among Population

Currently, the model incorporates state policy around masks with a variable indicating whether or not a state mandated a mask policy for the general public. While this is useful for describing an action a state government can make and its affect on the case rate, it doesn't full capture whether how much the population actively wears a mask to drive down transmission. Instead, a variable which would indicate the fraction of the population actively wearing a mask would better capture the mechanism of how the virus is transmitted among people.

A variable capturing the mask use among the population would likely reduce the explanatory power of the state mask mandate variable to be less negative. This would cause the bias of the omitted variable to be negative and towards from zero.

5-5: GDP Percentage of Tourism

The economic standing of a state not only affects the individual households of its population, but also includes how it interacts with other states and how its peoples' lives might have been economically changed by the pandemic. One way that a state may increase its number of covid cases is both from tourists or others visiting during the pandemic and bringing disease, but also if its own population must work in close contact with others serving them in the tourism industry. In addition, if certain key industries to a state, such as tourism, layed off many workers, they would likely have to take lower paying close contact jobs such as in the service or gig economy in order to continue having a livelihood. This would further lead to a higher covid case rate

These effects may be included with a variable that demonstrates how much of a state's GDP is from tourism. It would be related to other economic indicator variables, such as the percent of the population living in poverty. While poverty is a broader phenomena, the number of people in poverty in the pandemic could be potentially higher due to negative growth of the tourism industry. Therefore, it is expected that states with a greater focus on tourism may also have more people in poverty in the pandemic. This would take some of the explanatory power of the poverty variable away, resulting in its model coefficient becoming less positive. This would result in a positive omitted variable bias away from zero.

6. Conclusion

In conclusion, although there is no evidence to show that although there is a direct correlation between the COVID case rate and population density as originally hypothesized in our research question, it can be seen from further development of the models that there is some correlation in other variables examined, namely the mask mandate variable. As can be seen from model 3, the regression model gives a mask mandate variable coefficient with p-value of 0.003058, which is highly significant. The negative coefficient also suggests that the original conjecture that the implementation of mask policies within a state would decrease the COVID case rate was indeed accurate.

Conversely, the hypotheses that population and demographics would have an effect on the COVID case rate was unable to be accepted. Although upon investigating the poverty percentage variable in model 2 there was a detection of notable adjacency to a significant result (and an actual significant result when using the `stargazer` package to compute the regression table) with a p-value of 0.05755, it was not enough to be truly

significant and reject the null hypothesis. Similarly, the relationship of race/ethnicity as well as age on case rate was also insignificant with no apparent correlation.

Given these results in conjunction with the original aim of the research question, it can be suggested that variables that can be controlled (in this case, mask policy) has a greater effect on stymieing the spread of COVID-19 than the variables that may be inherent to a state, such as population density and age/ethnicity demographics. This is a promising insight, as it provides opportunity for states to implement and apply policies to intervene and control the case rate without being worried about its varying effects depending on population characteristics. However, it should once again be stressed that the final report may include and analyze alternative variables and apply them to different models based on the findings from this draft report, which as a result may also lead to more concrete or slightly varied insights.