Lab 2 Draft Report: COVID-19 Case Rate vs Population and 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.") %>%
  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)
# 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 case rate related to the population density and distribution of demographics of a state? How does the effect of population make-up on case rate compare with the effect of policy decisions made in each state?

For this report, the investigation will be centered around the effect of population metrics such as density and demographics and policy on the COVID case rates across the United States (US), which is grouped by the 50 states plus the District of Columbia (D.C.). This research question aims to measure the 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 cases. In this sense, the aim is to measure how dependent the COVID case rate is on that of which cannot be controlled (i.e., population and demographics) vs that of which can be controlled to a certain degree (i.e., implementation of policies to attempt to combat case rate). With this information, one could suggest whether or not the proliferation of COVID can be controlled, and to the level at which these policies help with this control.

This question can be broken up into two distinctive areas of investigation. The first and primary area of investigation involves how the COVID case rate is affected by population metrics. For this, the key variables are COVID case rate per 100,000 residents and the population density per square mile. This will provide an initial understanding in the relationship between the spread of the virus and how densely populated a state is. Following this, the analysis considers variables that address the make-up of these populations, specifically focusing on the distribution of demographics such as race/ethnicity and age. In order to account for the varying population sizes across the states, the variables are operationalized by specifically looking at variables that contain a rate, in order to standardize across the samples. In this sense, the analysis quantifies the spread of COVID by leveraging the COVID case rate per 100,000 rate, and for population demographics it investigates the percentage distribution of the race/ethnicity and age categories (e.g., white percentage of total population).

The secondary area of investigation will be to also include variables that measure policy implementation within states, and to compare the effect of these as opposed to the primary area of investigation that is population metrics. Specifically, there is a focus on policies that address the mandates of wearing masks within the state. Through this, the aim is to analyze whether the implementation of mask related policies will have an effect on case rate, as well as the degree of this effect relative to population metrics. To operationalize these policies and simplifying the measurements, the models will only consider whether or not these policies were implemented through indicator variables (1 = implemented, 0 = not implemented).

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 (with the addition of the District of Columbia) is treated as a unique data point, the sample contains 51 data points. Although this size meets the general rule of an adequate sample size, 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 factors such as geographic location and employment opportunities.

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 much of the information is not captured given the methods of operationalization. Among others, the loss of date-related information within policy implementation variables 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 policy implementation has some effect on case rate, this difference in length of time could play a large 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 mainly on mask mandates. 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). This follows for the population-related variables, as they capture race/ethnicity and poverty, but not other characteristics such as proportions of residents with pre-existing conditions. 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 or

demographic measures were simply not readily available in the dataset.

2. A Model Building Process

The primary variable of interest will be the total covid case rate per 100,000 residents. The covid case rate was judged to give a better understanding than the death rate because of the potential for other, unrecorded variables that could be correlated with a person dying from covid versus just becoming infected. For example, the availability and quality of medical care in a state may impact its ability to keep covid infected patients alive, although this is not represented 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 infection rate was chosen so that these other variables which may correlate with the death rate would not have to be considered. Finally, the covid infection rate was also chosen on a per capita basis so that potential total population differences between states would already be accounted for in the variable.

The modeling goal for the covid case rate will be one of description. A modelf of causation may be created out of this model if potential causal relationships are examined and incorporated into the model's structure. This may later be expanded to predict the covid case rate in the past seven days in order for a state to use it to potentially judge what effect a modeled policy may have on its current situation.

There are two broad groups of covariates which will be examined in building the model. The first is demographic information on the state, such as population density, the unemployment rate, or rate of people living in poverty. The second general group is that of policy decisions taken by the state, such as mask mandates, legal enforcements of policies, or the political party of the state's governor (which is assumed to be related to the policies they may support). Problematic covariates would include any of the other direct measures of covid severity in the state, such as total infection rate not on a per capita basis or death rates. It can be assumed that these variables would be colinear with the primary variable of interest since more covid cases per capita would imply more total cases not adjusted for population or be strongly correlated with more deaths per capita, etc.

Model Variables

The three models aim to investigate the relationship between the COVID rate, state demographic information, and mask mandates. The list of variables included in the current dataset that pertain to these concepts are listed below.

State Characteristics:

- Governor
- Population density per square miles
- Population 2018
- Nonelderly Adults Who Have A Pre-Existing Condition
- Percent at risk for serious illness due to COVID
- All-cause deaths 2018
- Number Homeless (2019)
- Medicaid Expenditures as a Percent of Total State Expenditures by Fund
- Life Expectancy at Birth (years)
- Percent Unemployed (2018)
- Percent living under the federal poverty line (2018)
- Weekly UI maximum amount with extra stimulus (through July 31, 2020) (dollars)
- Median Annual Household Income
- Closed other non-essential businesses
- Stay at home/ shelter in place
- Mandate face mask use by all individuals in public spaces
- No legal enforcement of face mask mandate

Demographics:

- Children 0-18 Percentage
- Adults 19-25 Percentage
- Adults 26-34 Percentage
- Adults 35-54 Percentage
- Adults 55-64 Percentage
- 65+ Percentage
- white Percentage of Total Population,
- Black Percentage of Total Population,
- Hispanic Percentage of Total Population,
- Other Percentage of Total Population

Currently, Case rate per 100,000 is chosen as a dependent variable to represent the pandemic status in each state. Variables such as population density per square miles, unemployed percentage(2018)/percentage living under the federal poverty line(2018), white/black percentage of total population, percentage of seniors(65+), and mask mandate policies are chosen as candidate covariates to explore potential relationships between pandemic status demographic characteristics, economic status, racial composition, and state policy.

Data Exploration

Through data exploration, most data investigated is clean as numerical type except the black percentage of total population. This information of some states is reported vaguely as less than 0.01, which is replaced as 0 for further work. This could be updated with more precise data with the help of external resources.

Several transformations have been made for variables as population density per square miles and white percentage of total population. Population density is extremely skewed due to the high population density of the District of Columbia. A logarithm transformation of this variable is of more uniform distribution. In addition, the square of the white percentage of the total population is of a distribution closer to the normal distribution. It also represents the chance that the interaction happens between two white people in practice.

In addition, correlation values of white percentage and black percentage, unemployed percentage and percentage living under the federal poverty line shows these two groups of variables are correlated respectively.

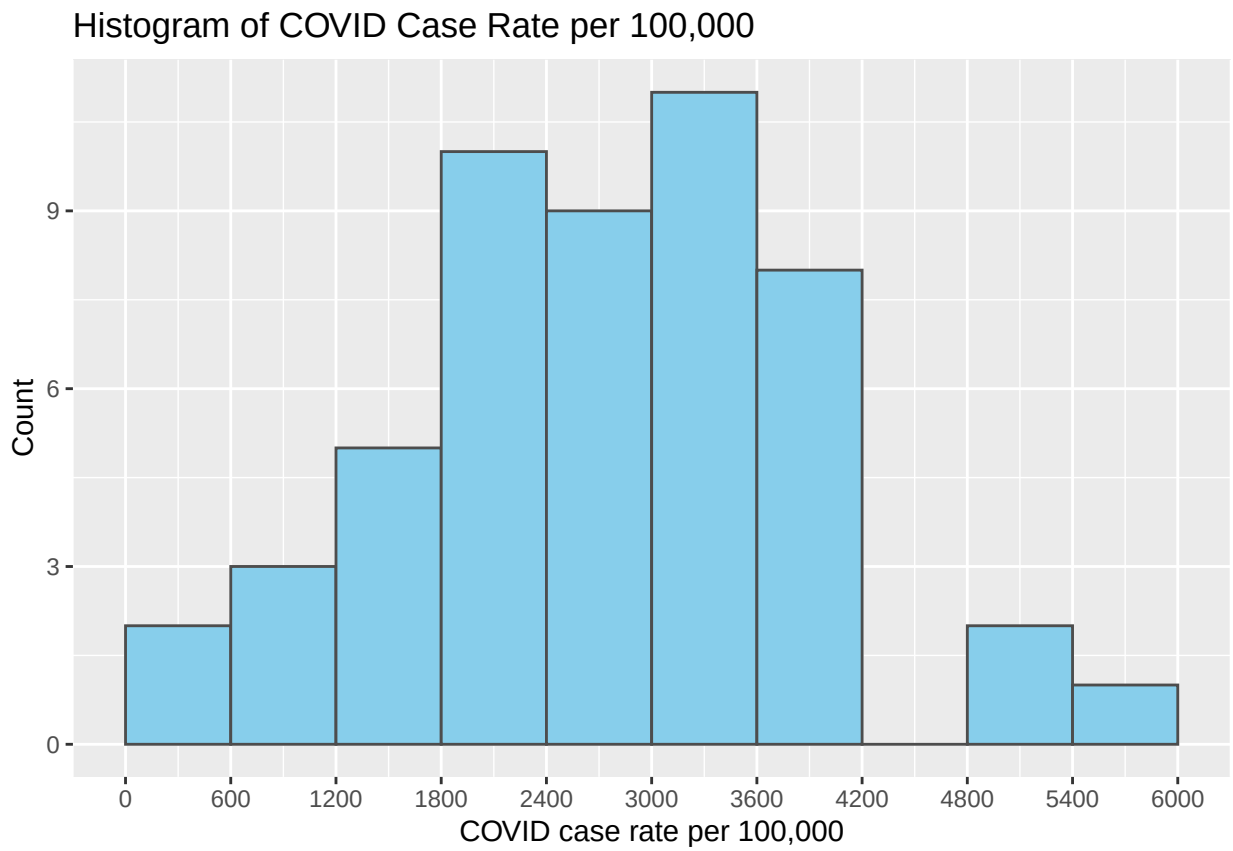
Model 1

For the first model, the relationship between the COVID case rate per 100,000 and the population density per square mile of states was analyzed. First, the distribution of COVID case rate per 100,000 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 Case Rate per 100,000",
       x = "COVID case rate per 100,000", y = 'Count') +
  scale_x_continuous(breaks=seq(0, 6000, 600))
```

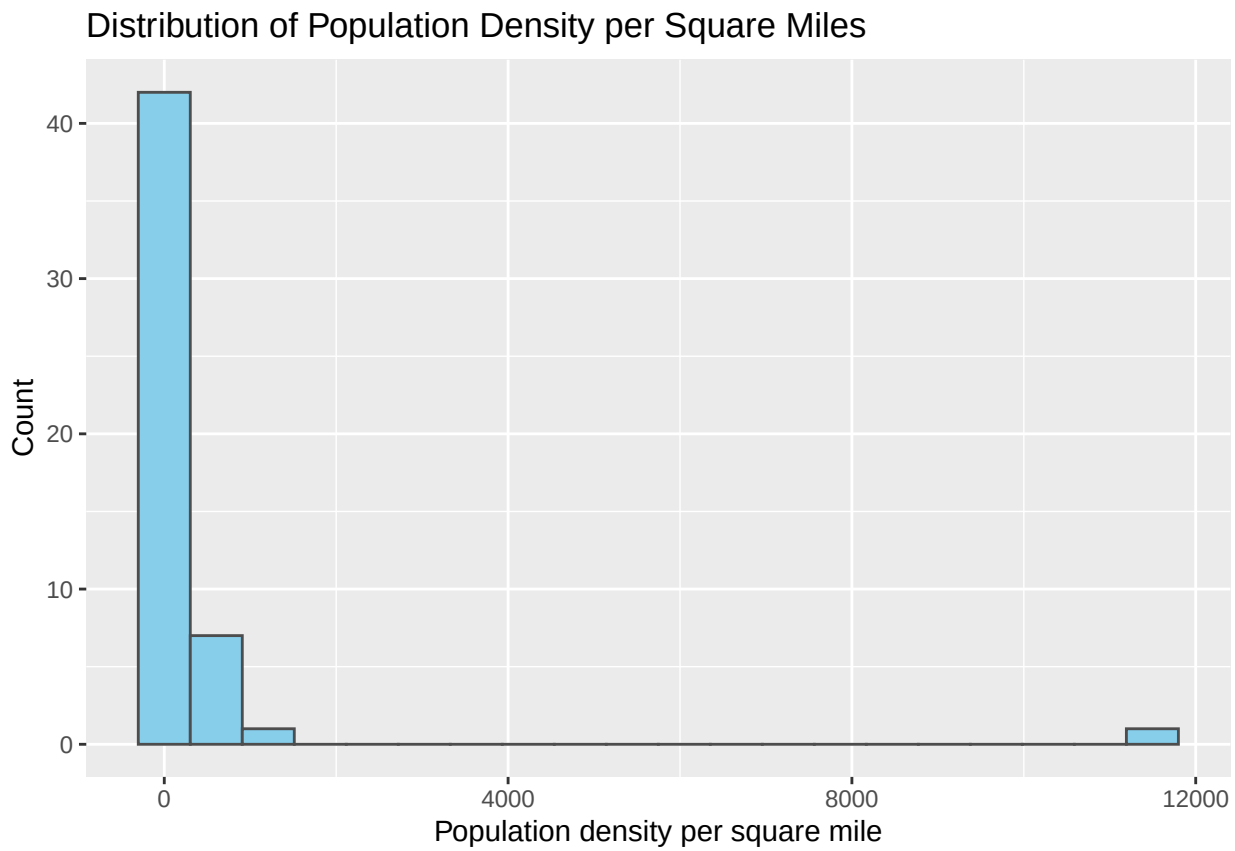


As can be seen above, the distribution is fairly normal, and given that it has already been standardized as a rate across all states, there is no need to perform any transformations on this variable. Thus, the case rate per 100,000 variable can be leveraged as is as the dependent variable for all three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

Next, the distribution of population density per square mile variable is examined.

```
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 the District of Columbia (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)
```

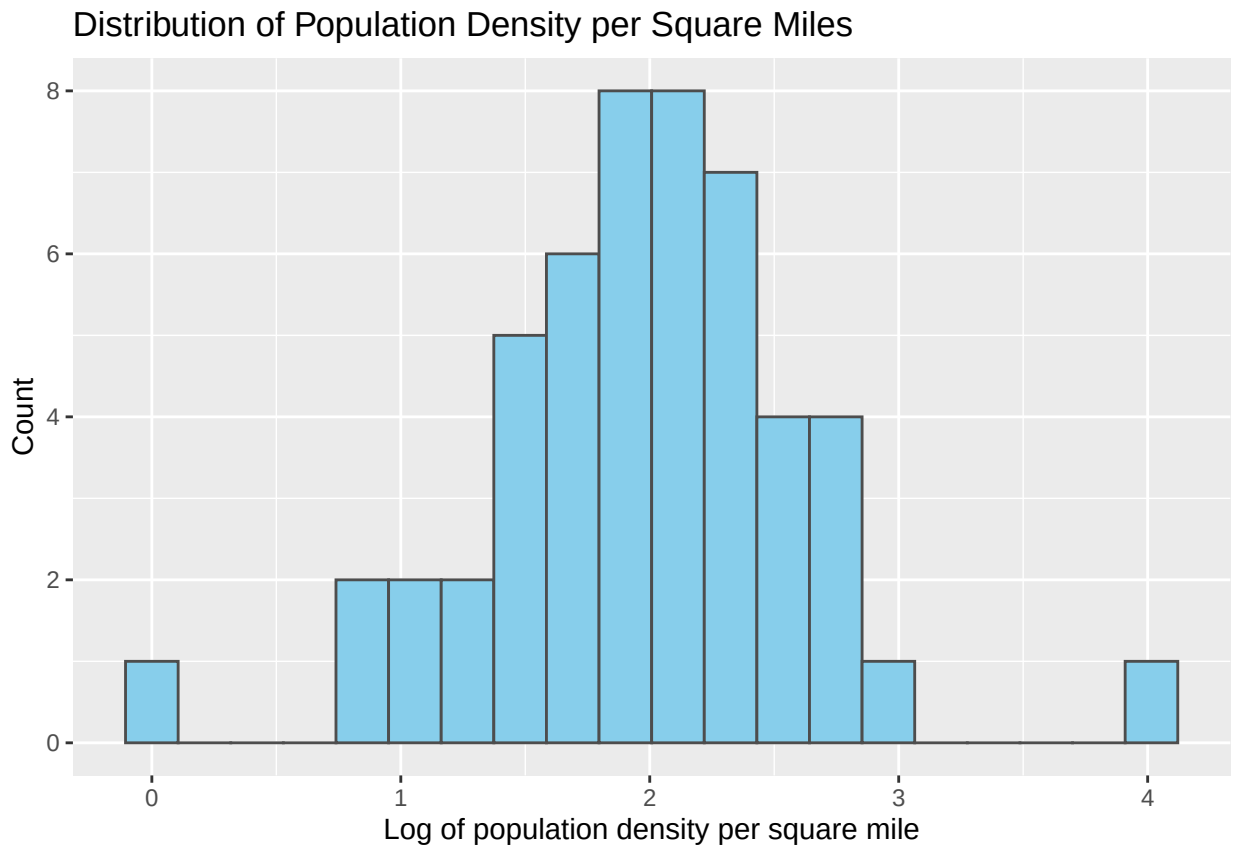
```
## [1] "District of Columbia = 11496.81"
```

```
# 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',
```

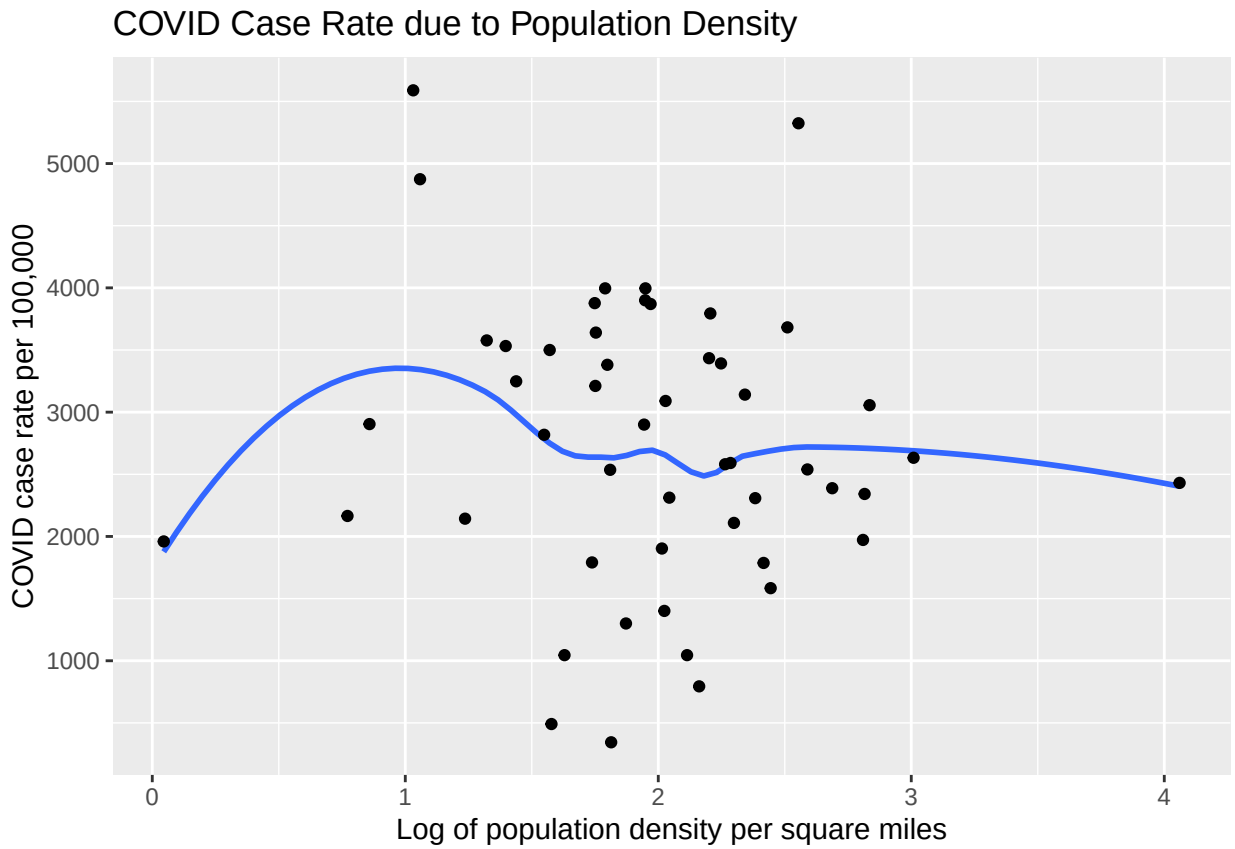
```
    x = 'Log of population density per square mile', y = 'Count')

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'
  )
```



From the above plot, it can be posited that there is no discernible relationship between the variables given the non-linear relationship. 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.

$$Case.Rate.Per.100000 = \beta_0 + \beta_1 log(Population.Density.Per.Square.Miles)$$

```
model1 <- lm(case_rate_100k ~ log10(population_density) , data = df)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                 3103.38     560.18  5.5400 1.187e-06 ***
## log10(population_density)  -179.22     244.27 -0.7337    0.4666
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the results of the regression, two things are identified. Firstly, the result of the coefficient is nowhere near significant, and therefore we fail to reject the null hypothesis that there is no correlation between case rate and population density, and there is no detectable effect of the independent variable population density. However, the coefficient of population density appears to be negative, which would concur with the assumption that the higher the population density, the lower the COVID case rate in a state.

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.

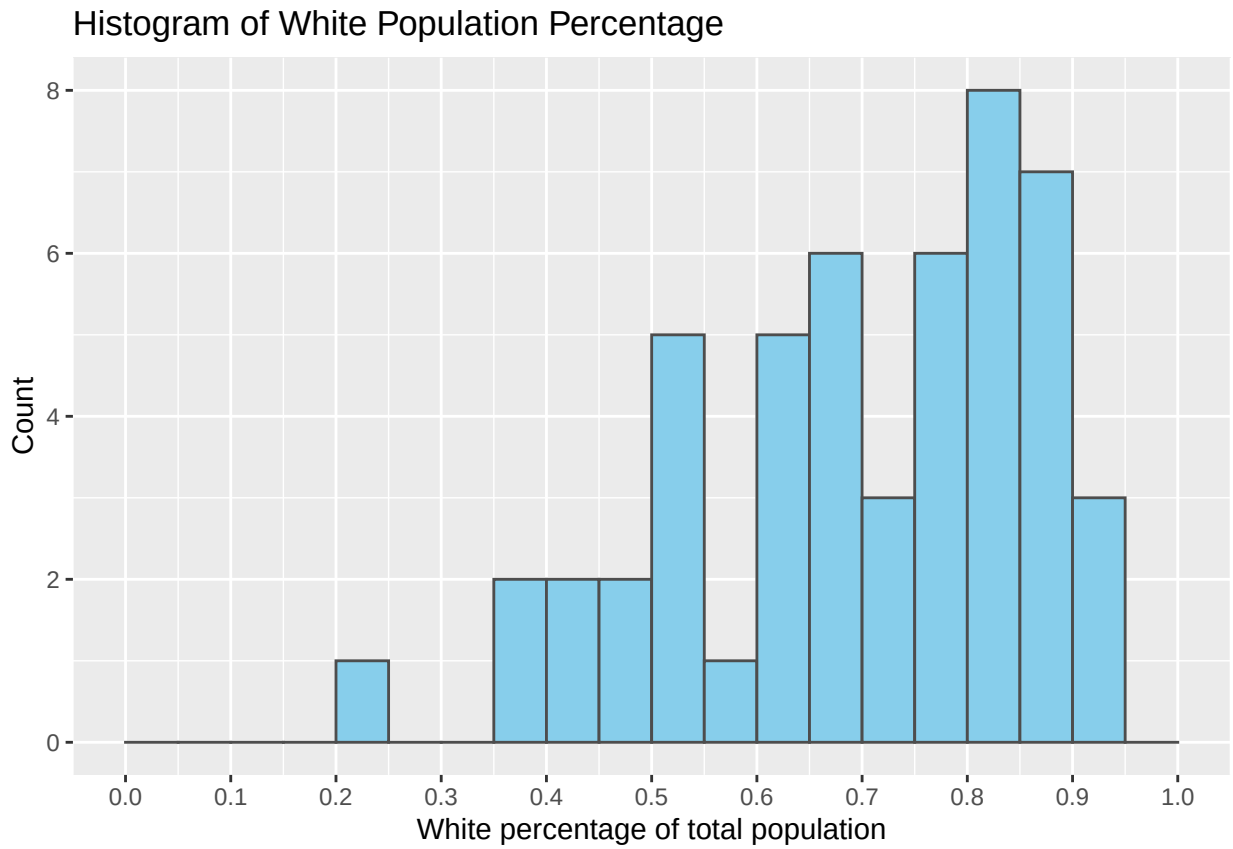
```
black_num <- as.numeric(df$black_population_pct)
black_num[is.na(black_num)]<- 0
cor(df$white_population_pct, black_num)
```

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

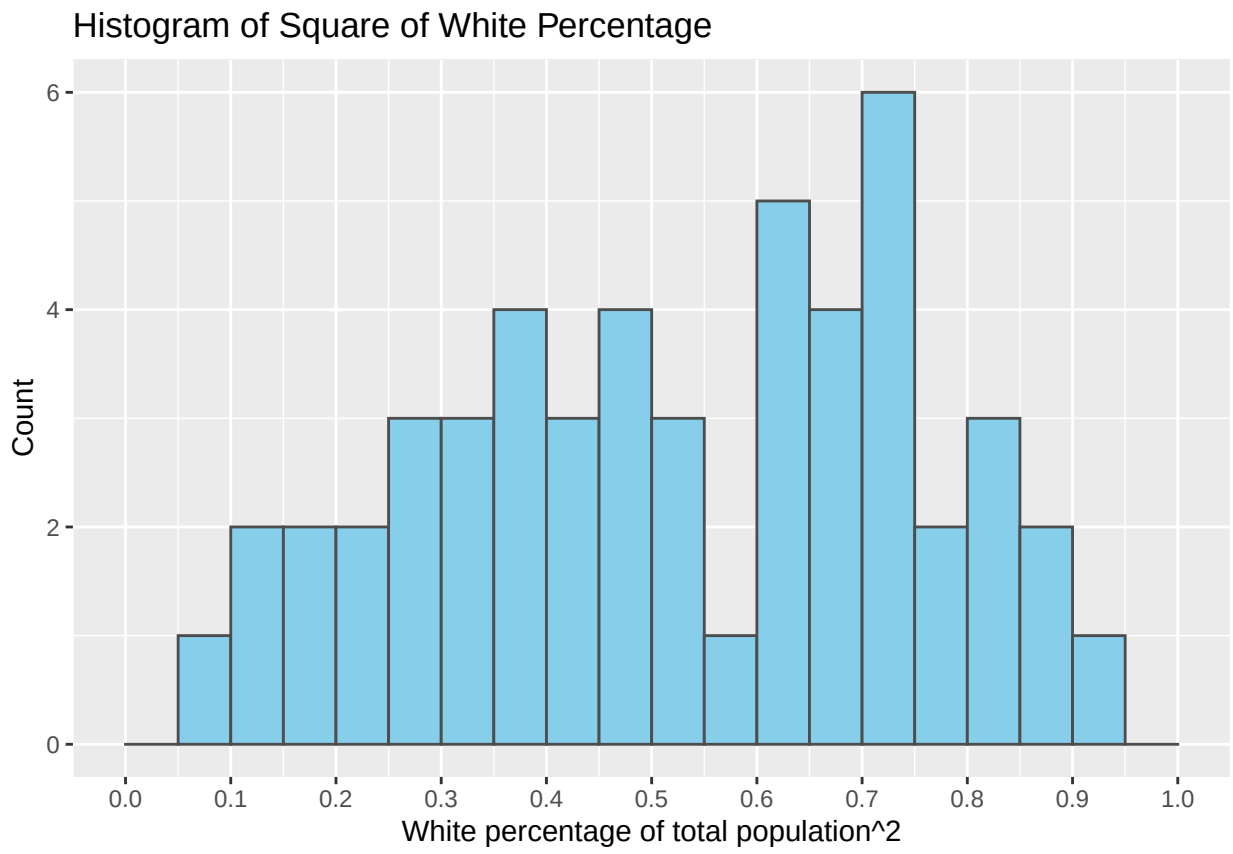
```
summary(df$white_population_pct)
```

```
##    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.

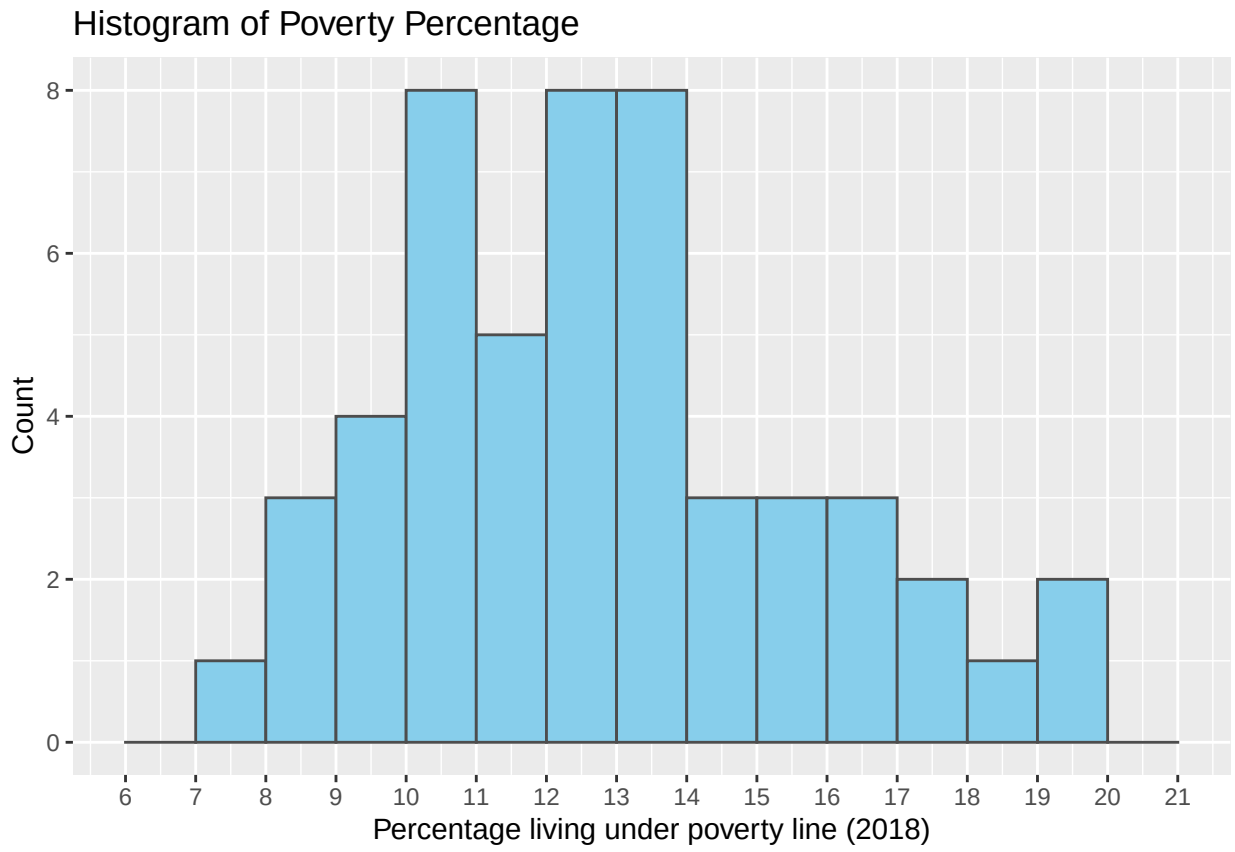
```
summary(df$poverty_pct)
```

```
##    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.

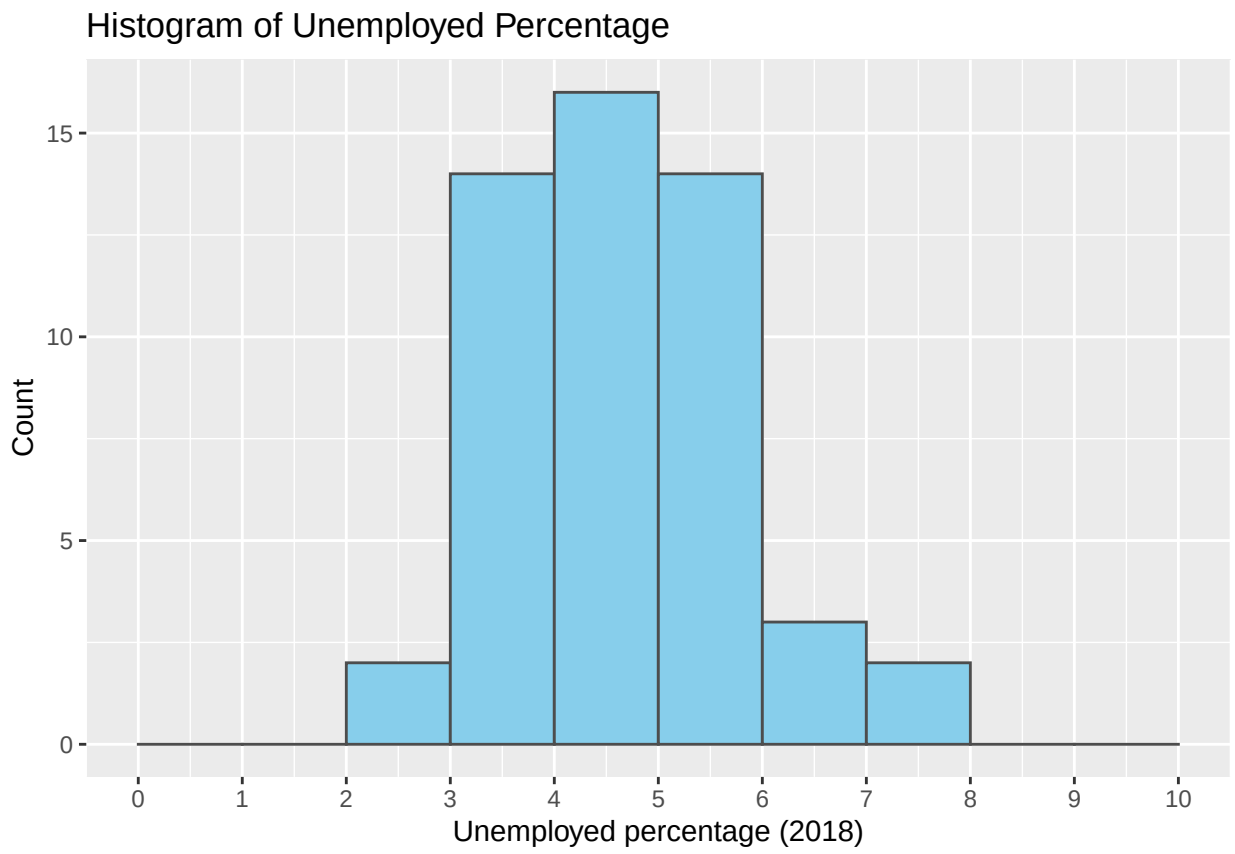
```
summary(df$unemployed_pct)
```

```
##    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))
```

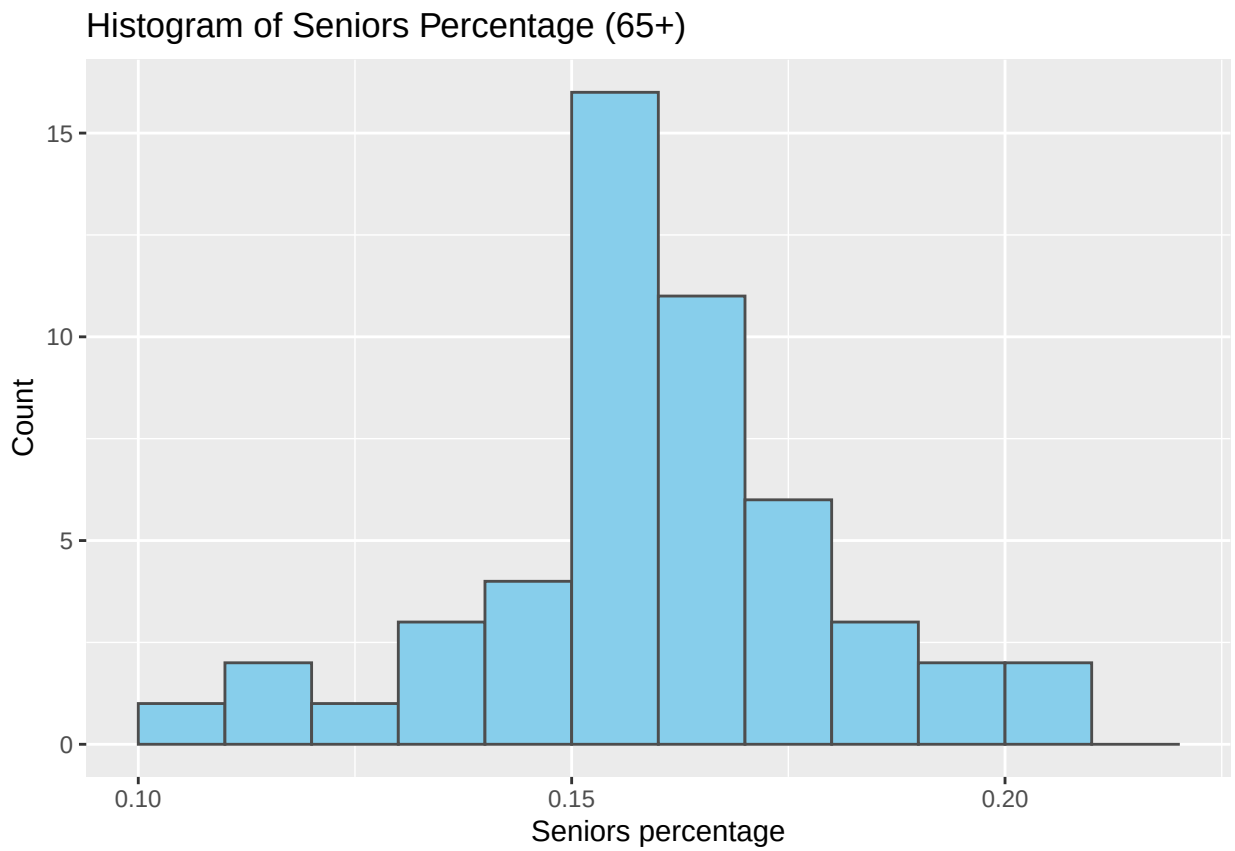


2-d) Seniors Percentage (Age 65+) The seniors percentage (age 65+) is of a nearly normal distribution.

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

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

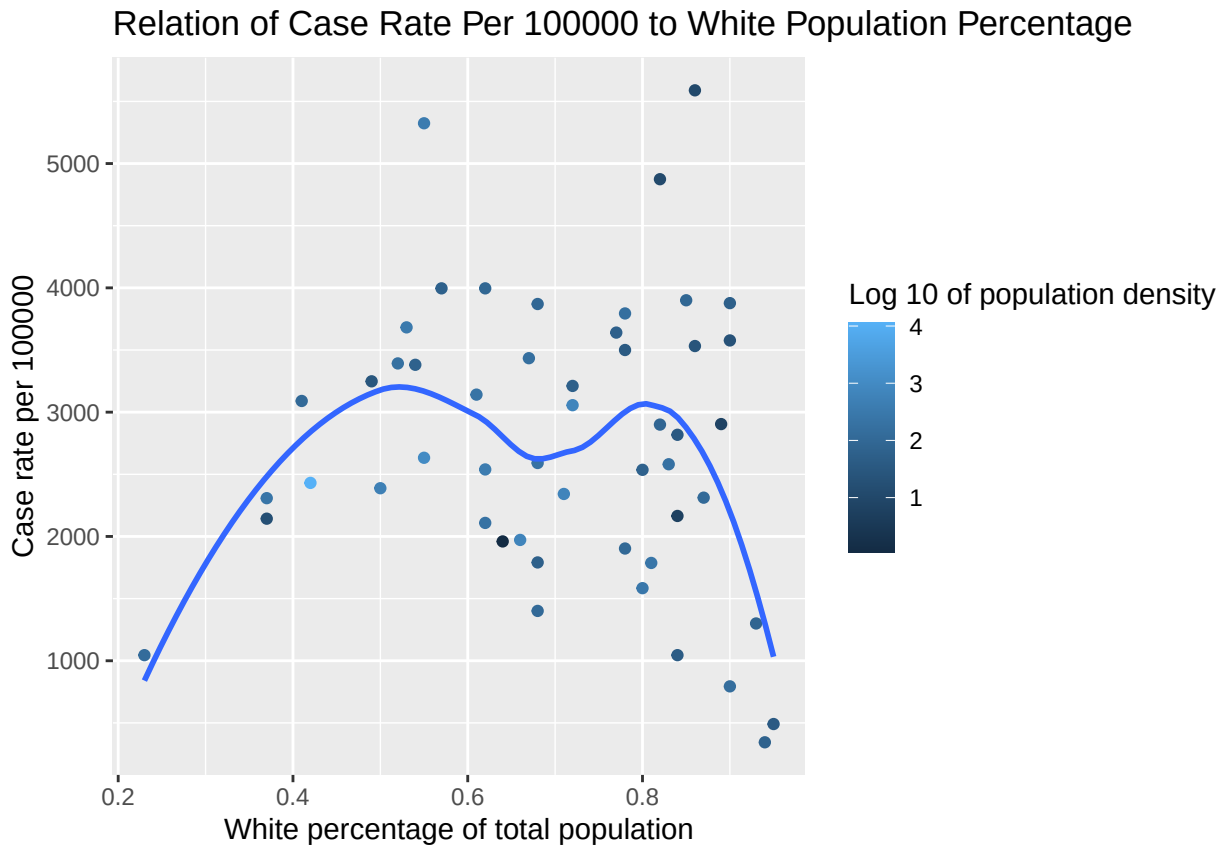
```
ggplot(data = df,
  mapping = aes(x= senior_pct))+
  geom_histogram(fill = 'skyblue', color = 'grey30', breaks = seq(0.10,0.22,0.01)) +
  labs(title =  "Histogram of Seniors Percentage (65+)", x = "Seniors percentage",
       y = 'Count')+
  scale_x_continuous(breaks=seq(0.1, 0.25, 0.05))
```



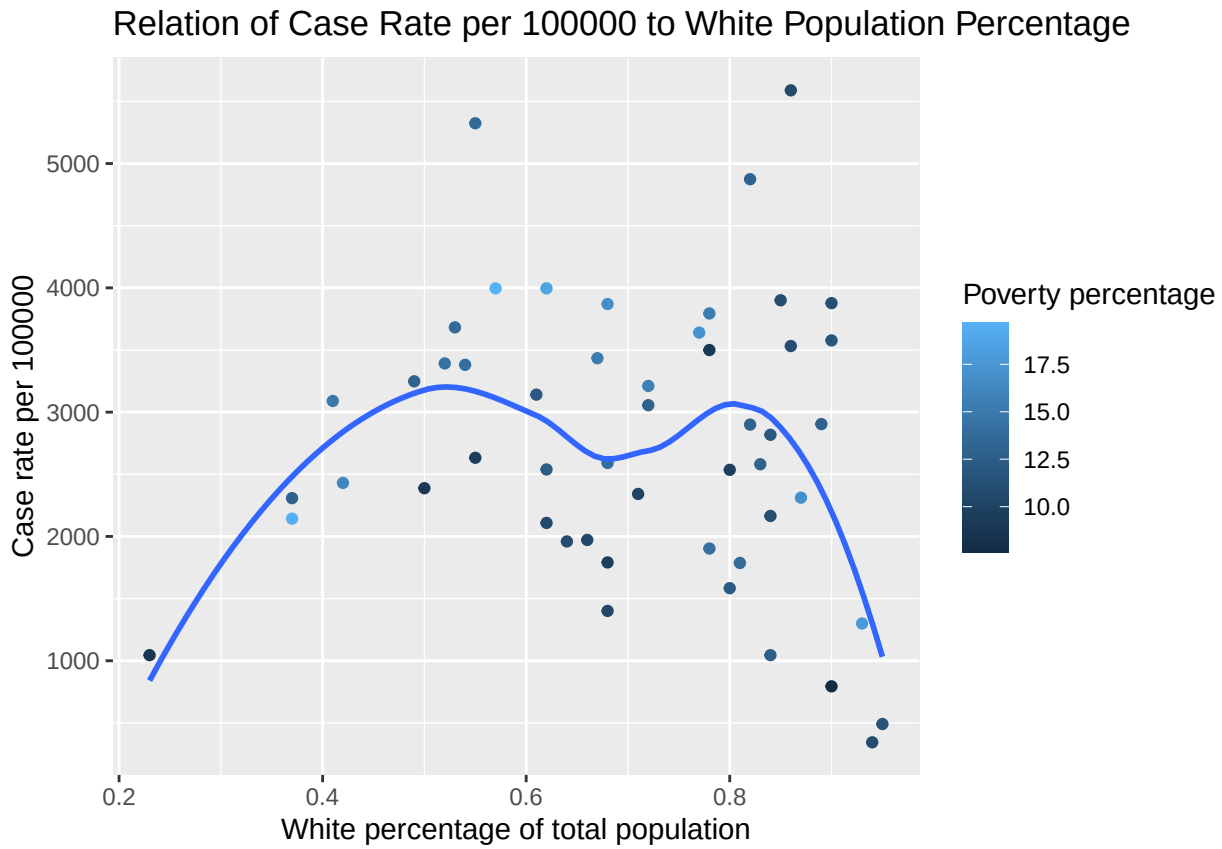
Model 2 Plots

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

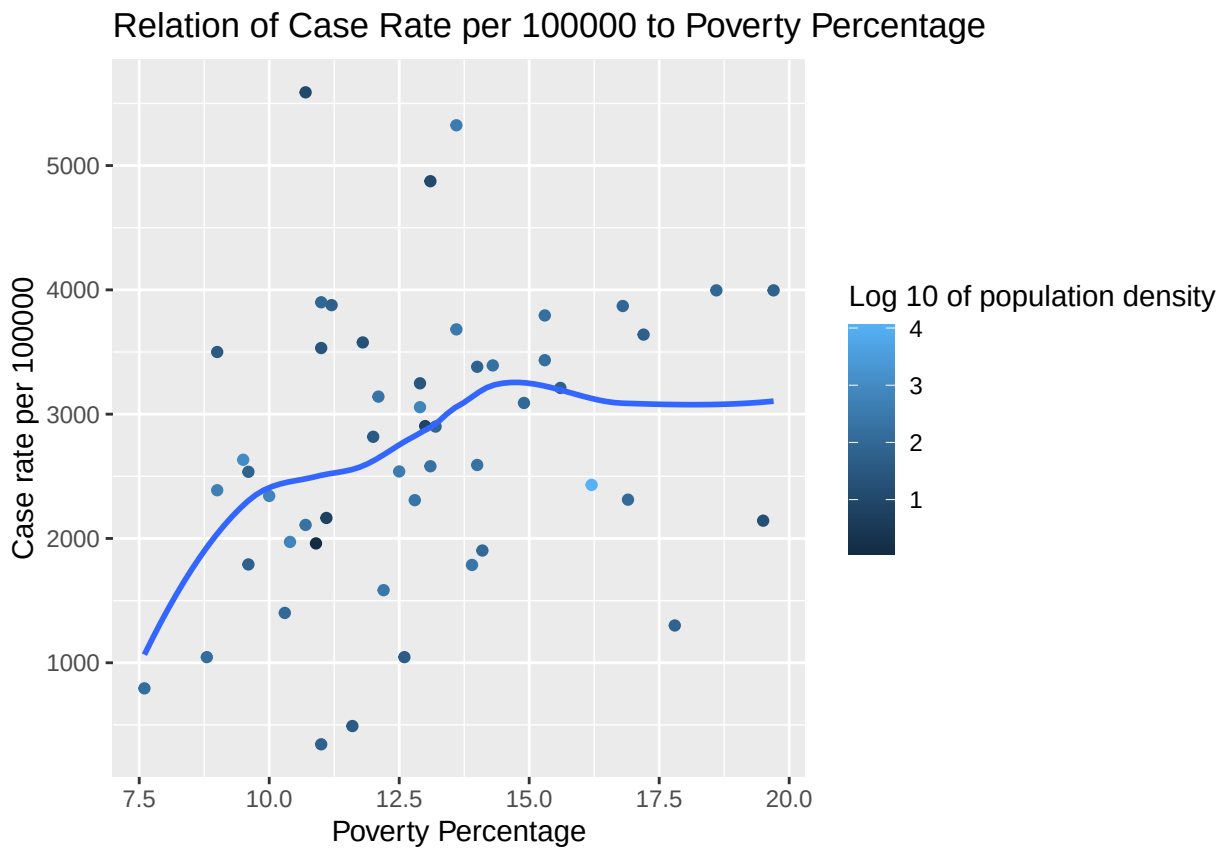
```
df %>%
  ggplot(aes(white_population_pct, case_rate_100k, color = log10(population_density))) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate Per 100000 to White Population Percentage',
    x = 'White percentage of total population',
    y = 'Case rate per 100000',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



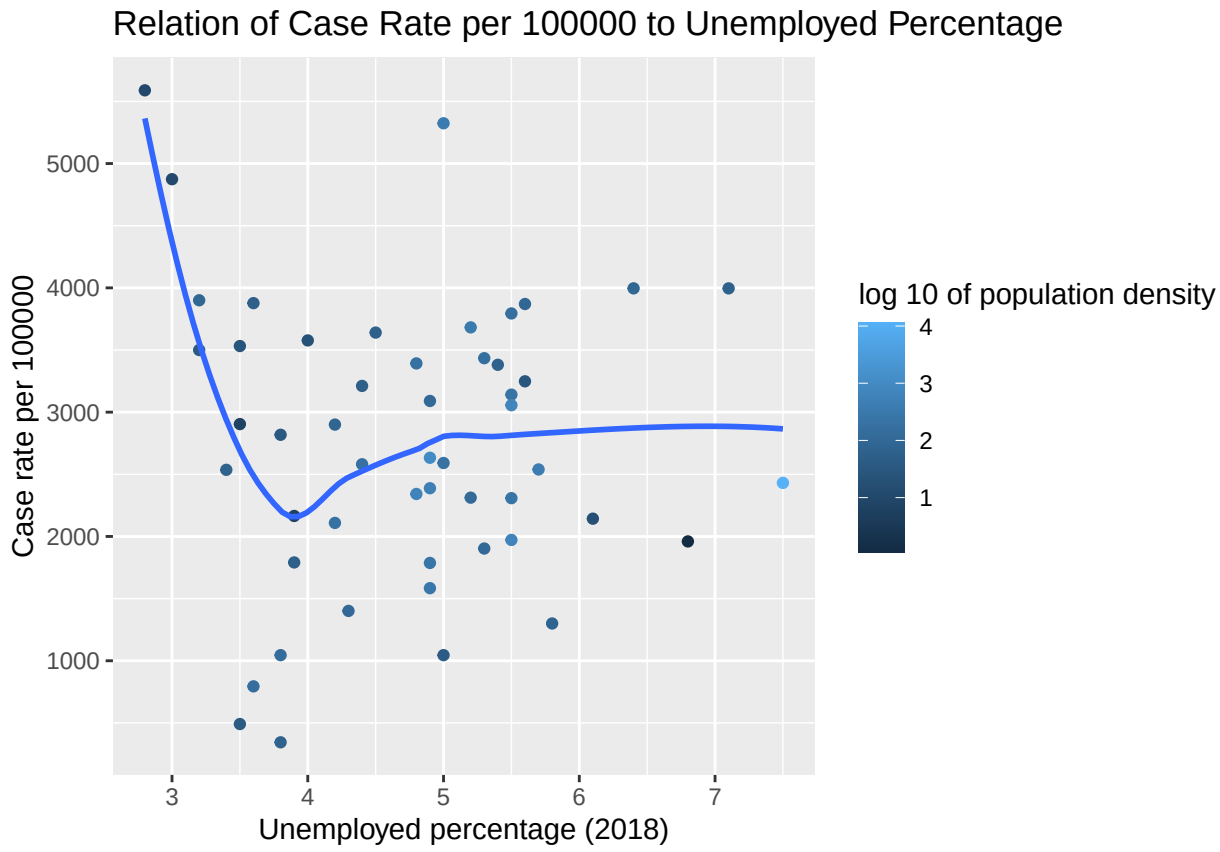

```
df %>%
  ggplot(aes(x = white_population_pct, y = case_rate_100k, color = poverty_pct)) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate per 100000 to White Population Percentage',
    x = 'White percentage of total population',
    y = 'Case rate per 100000',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



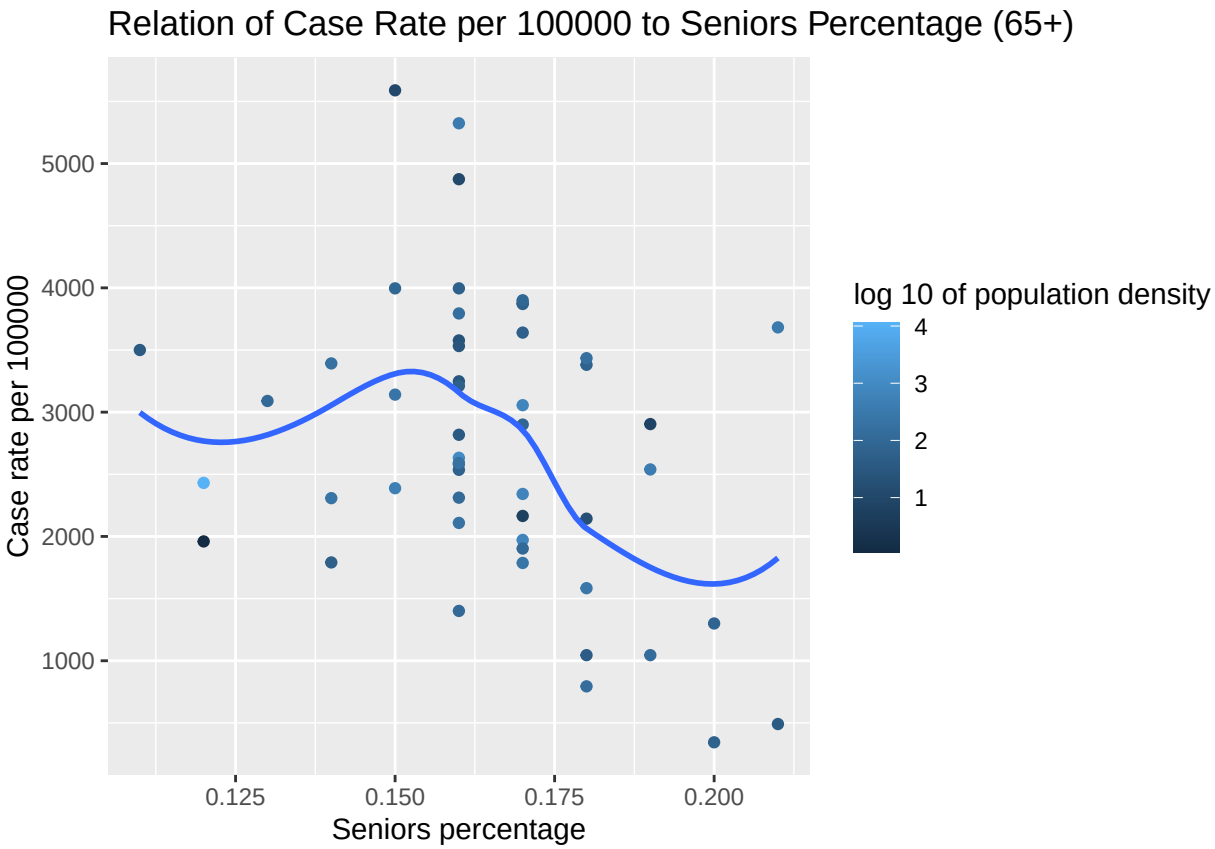

```
df %>%
  ggplot(aes(x = poverty_pct , y = case_rate_100k, color = log10(population_density))) +
  geom_point() +
  geom_smooth(se=FALSE)+
  labs(
    title = 'Relation of Case Rate per 100000 to Poverty Percentage',
    x = 'Poverty Percentage',
    y = 'Case rate per 100000',
    color = 'Log 10 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.However, no coefficients regarding these variables in model 2 are statistically significant.

Model 2-a

$$Case.Rate.Per.100000 = \beta_0 + \beta_1 log(Population.Density.Per.Square.Miles) + \beta_2(White.Percentage.of.Total.Population)$$

```
model2a <- lm(case_rate_100k ~ log10(population_density)+ white_population_pct, data = df)
coeftest(model2a, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3838.47    1329.29  2.8876 0.005806 **
## log10(population_density)  -257.19     291.12 -0.8834 0.381406
## white_population_pct       -829.71    1357.62 -0.6111 0.543984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$Case.Rate.Per.100000 = \beta_0 + \beta_1 log(Population.Density.Per.Square.Miles) + \beta_2 (White.Percentage.of.Total.Population)^2$$

```
model2a <- lm(case_rate_100k ~ log10(population_density)+ I((white_population_pct)^2), data = df)
coeftest(model2a, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                   3858.81    1013.37  3.8079 0.0003979 ***
## log10(population_density)     -305.61     313.53 -0.9747 0.3345845
## I((white_population_pct)^2)   -974.21     999.73 -0.9745 0.3347088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2-b

$$Case.Rate.Per.100000 = \beta_0 + \beta_1 log(Population.Density.Per.Square.Miles) + \beta_3 (Percentage.Under.Poverty.Line.2018)$$

```
model2b <- lm(case_rate_100k ~ log10(population_density) + poverty_pct, data = df)
coeftest(model2b, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1675.148    988.000  1.6955  0.09646 .
## log10(population_density) -189.854    245.881 -0.7721  0.44382
## poverty_pct                112.241     57.684  1.9458  0.05755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2-c

$$Case.Rate.Per.100000 = \beta_0 + \beta_1 log(Population.Density.Per.Square.Miles) + \beta_4 (Percentage.Unemployed.2018.)$$

```
model2c <- lm(case_rate_100k ~ log10(population_density) + unemployed_pct, data = df)
coeftest(model2c, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3097.6820   991.5362  3.1241 0.003022 **
## log10(population_density) -180.1410   311.9827 -0.5774 0.566363
## unemployed_pct               1.5824   216.5590  0.0073 0.994200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2-d

$$Case.Rate.Per.100000 = \beta_0 + \beta_1 log(Population.Density.Per.Square.Miles) + \beta_4(Seniors.Percentage)$$

```
model2d <- lm(case_rate_100k ~log10(population_density)+ senior_pct, data = df)
coeftest(model2d, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  5988.17    2062.51  2.9033 0.005564 **
## log10(population_density)    -184.80     336.59 -0.5490 0.585526
## senior_pct                 -17447.88   10047.30 -1.7366 0.088877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

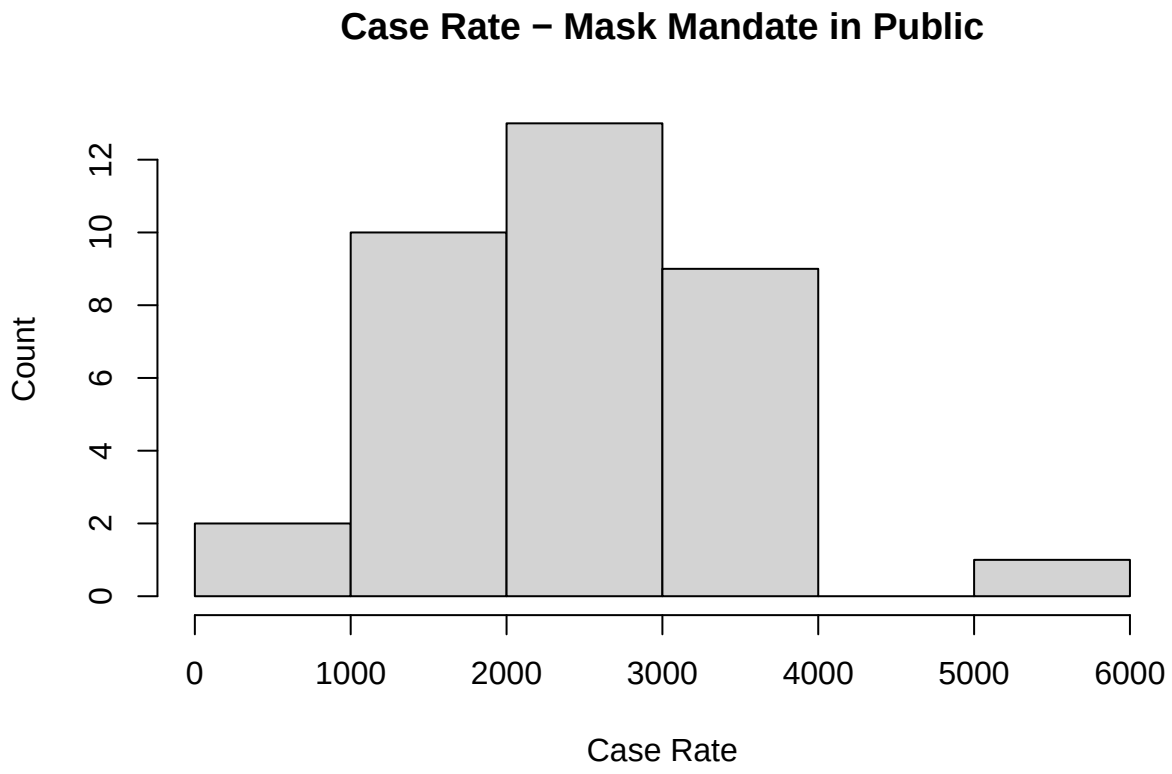
It is worth to note that, the mask_public is in date format, which has value of zero or a actual date when the policy is enforced. According to the documentation, zero represents "the absence of an order or directive", which can be interpreted as the policy is not enforced by the state explicitly. For a leaner 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 hard to measure the time effect as different states that enforced the public mask policy may have different threshold.

Firstly, "mask_public_bool" column is created from "mask_public", with 1 represents that the state has a pulbic mask policy, and 0 means there's no explicit public mask policy from the state.

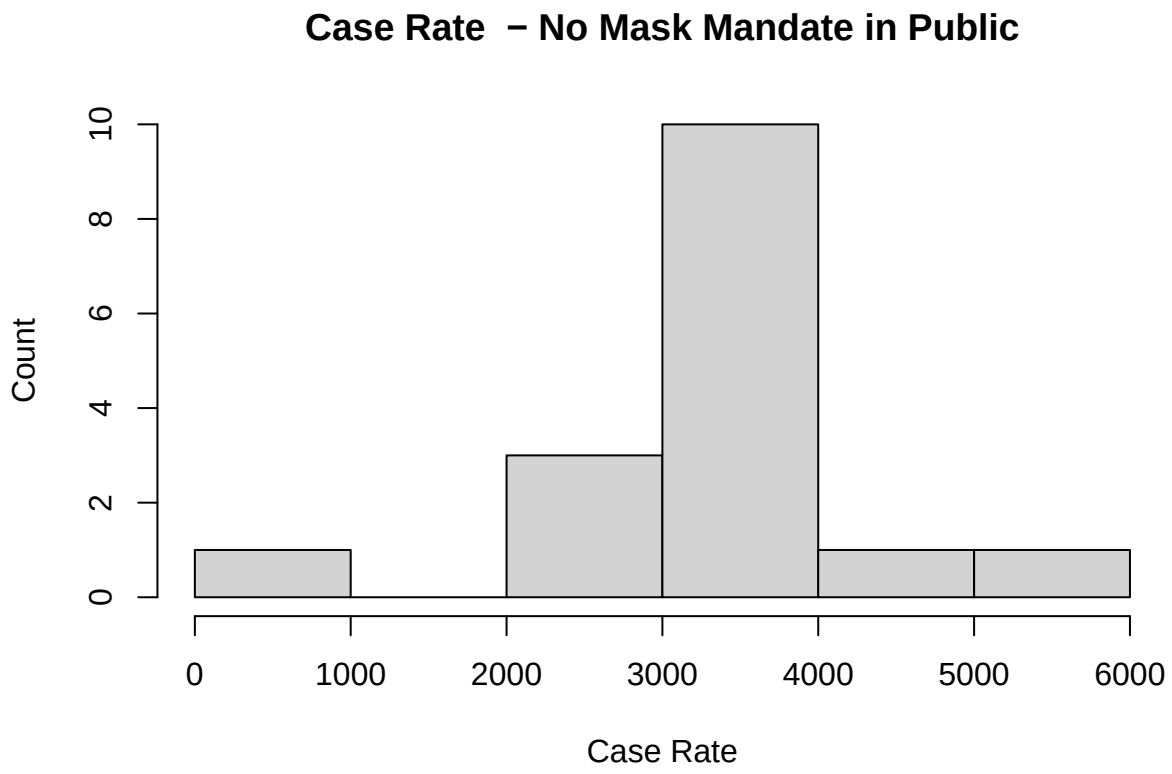
```
df <- df %>%
  mutate(
    mask_public_bool = case_when(
      mask_public == 0 ~ 0,
      !(mask_public == 0) ~ 1
      )
  )
```

Histograms below that shows the distribution of states that has public mask policy or not. Both distributions exhibits certain degree of normality.

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

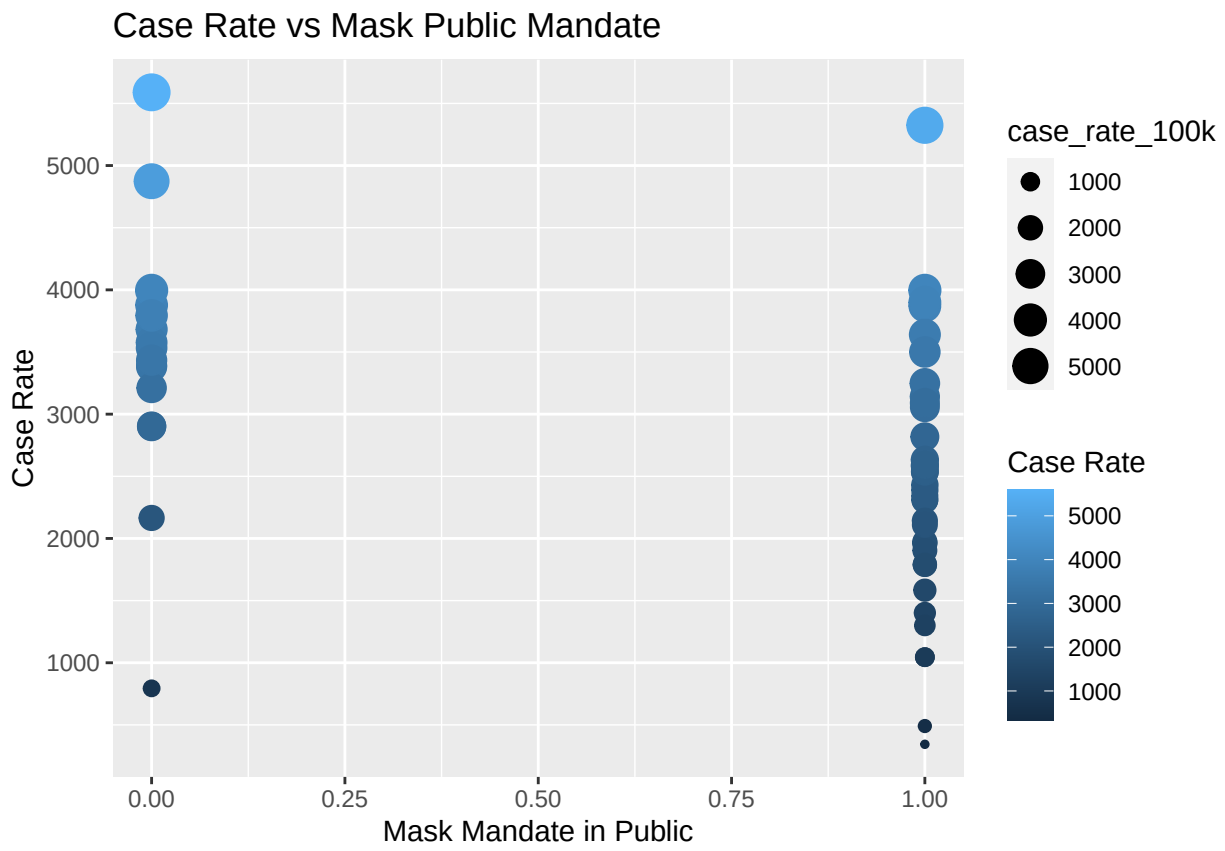



```
hist(df$case_rate_100k[df$mask_public_bool == 0], xlab='Case Rate', ylab='Count',
     main='Case Rate  - No Mask Mandate in Public')
```



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'
  )
```



Model 3-a 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.

```
model3a <- lm(case_rate_100k ~ log10(population_density)
              + poverty_pct + mask_public_bool, data = df)
coeftest(model3a, vcov = vcovHC)
```

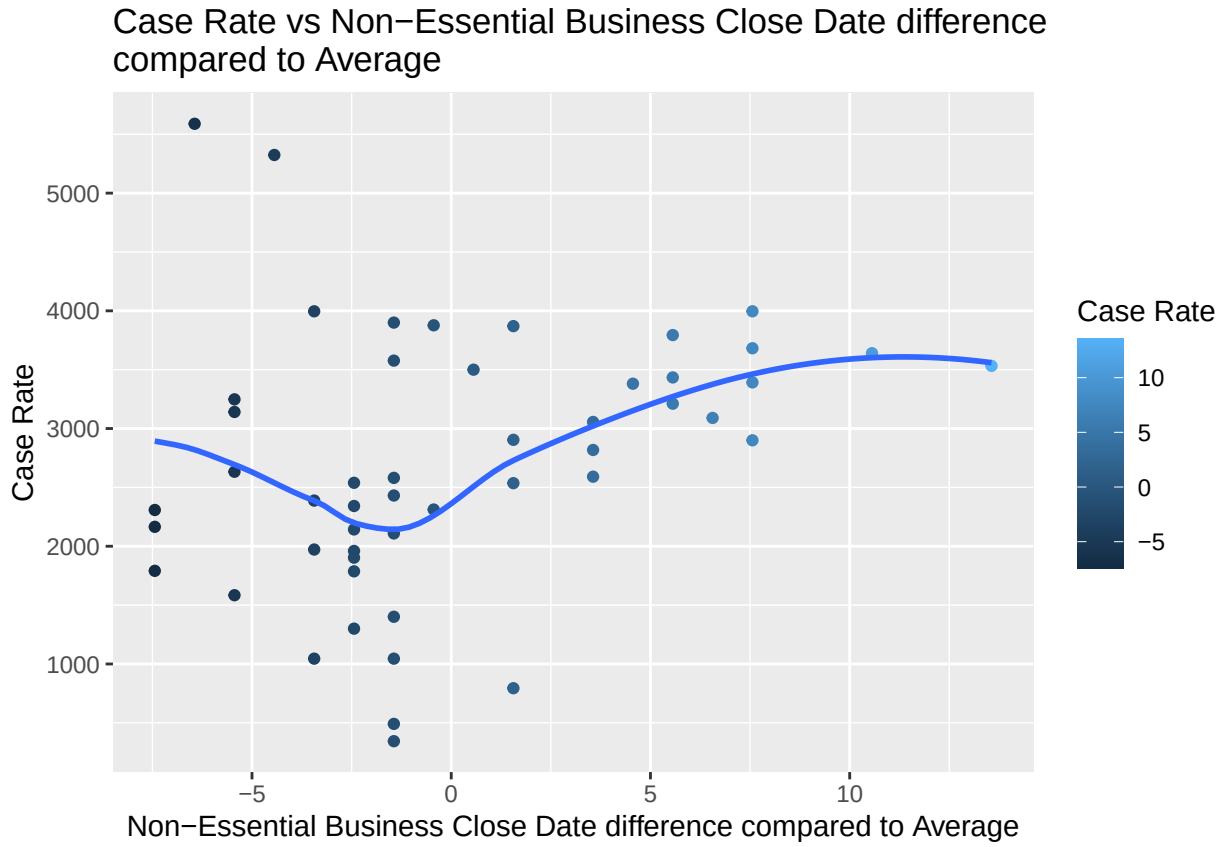
```
##
## t test of coefficients:
##
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2027.409   1003.465  2.0204 0.049060 *
## log10(population_density)    31.377    206.451  0.1520 0.879852
## poverty_pct                 104.191     59.583  1.7487 0.086875 .
## mask_public_bool           -998.485    334.011 -2.9894 0.004437 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 3-b & 3-c Considering that there could be multiple policy variables that are relevant to predicting the case number. We have also examined the "Closed.other.non.essential.businesses" variable (renamed as "business_closed") which represents the date when the state issued closure of non-essential businesses. As "business_closed" is a date variable, a transformation was performed to convert the date into a variable "business_close_diff" which represents the days difference between the business close date of the state compared to the average date of all the states.

```
df$business_close_diff =as.numeric(df$business_closed-mean(df$business_closed, na.rm=TRUE))
```

From the plot of business close date difference against case rate, the graph shows that there is some degree of linearity between the two variables.

```
df %>%
  ggplot(aes(x = business_close_diff, y = case_rate_100k, color = business_close_diff)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(
    title = 'Case Rate vs Non-Essential Business Close Date difference \ncompared to Average',
    x = 'Non-Essential Business Close Date difference compared to Average',
    y = 'Case Rate',
    color = 'Case Rate'
  )
```



By running the model with "business_close_diff" against case rate variable, the coefficient shows that the variable is significant.

```
model3b <- lm(case_rate_100k ~ business_close_diff, data = df)
coeftest(model3b, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2706.940    153.925 17.5861  < 2e-16 ***
## business_close_diff   63.176     26.696  2.3665  0.02204 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, the regression with both policy variables "mask_pubic_bool" and "business_close_diff" shows that neither the the variable is significant.

```
model3c <- lm(case_rate_100k ~ business_close_diff + mask_public_bool, data = df)
coeftest(model3c, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         3249.007    371.367  8.7488 1.991e-11 ***
## business_close_diff   36.060     37.072  0.9727   0.33568
## mask_public_bool    -774.381    437.943 -1.7682   0.08351 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The findings above suggests that there is colinearity between the two variables "mask_pubic_bool" and "business_close_diff". Therefore, we decided to use "mask_public_bool" variable which has higher significance to represent our policy variable. As a result, model 3-a is selected as the final model for model 3.

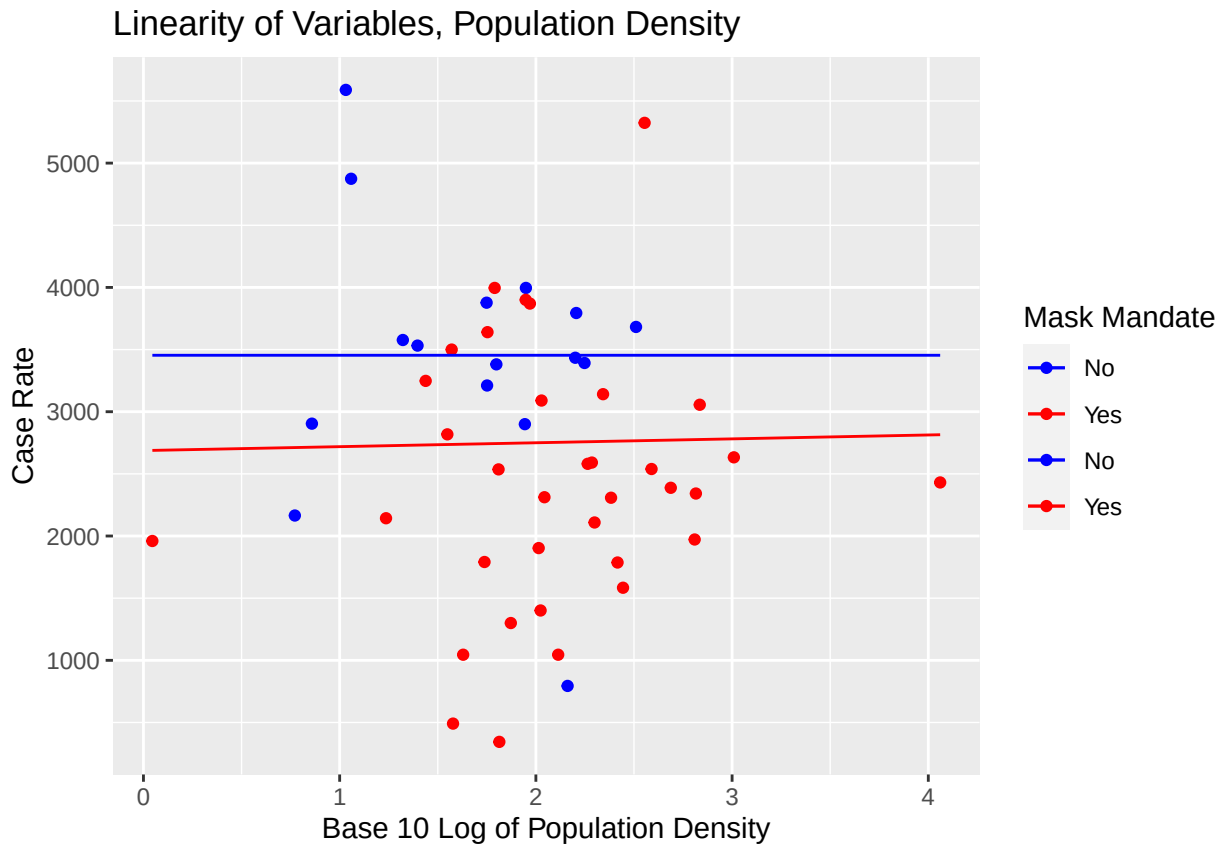
3. Limitations of the Models

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.

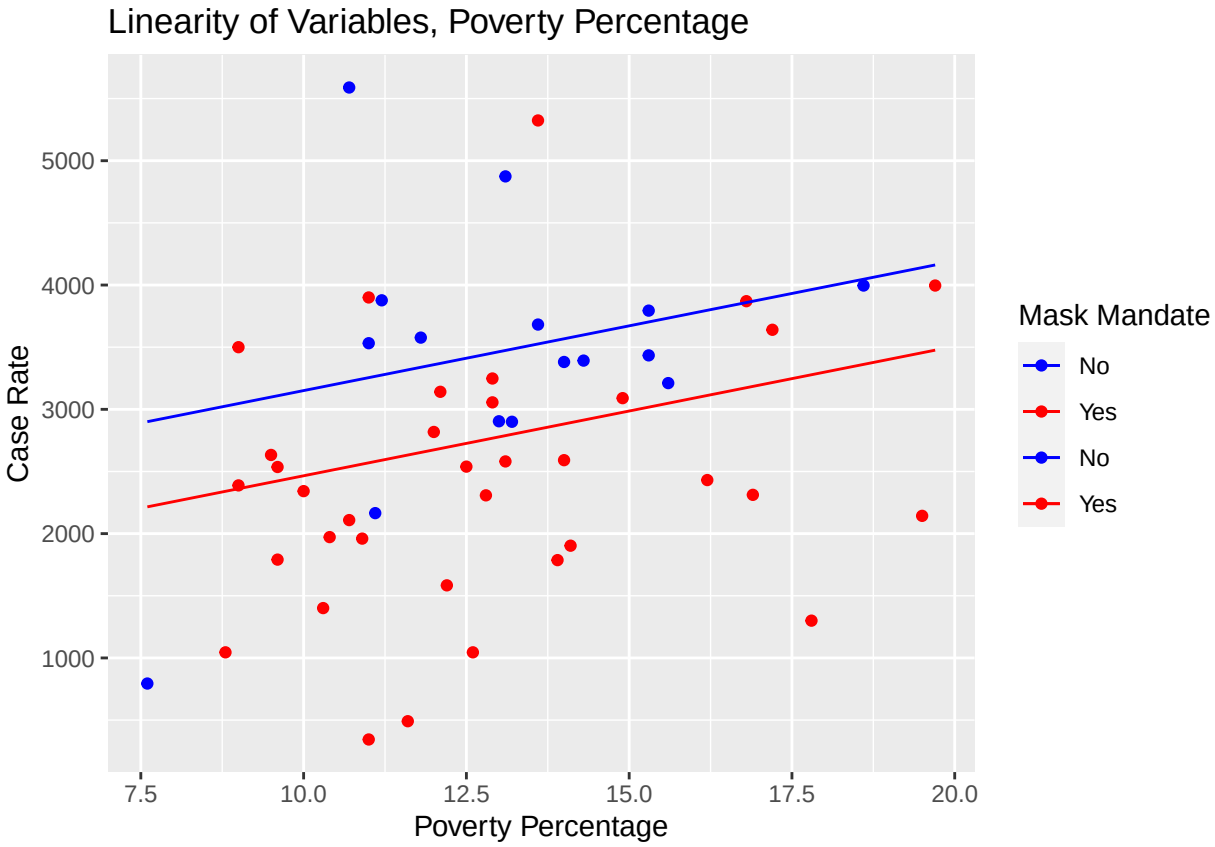
```
avg_pop <- log10(mean(df$population_density))
avg_pov <- mean(df$poverty_pct)
avg_mask <- mean(df$mask_public_bool)
y_pop <- model3a$coefficients[1] + model3a$coefficients[2]*log10(df$population_density) + model3a$coeffi
y_pov <- model3a$coefficients[1] + model3a$coefficients[2]*avg_pop + model3a$coefficients[3]*df$poverty_
y_pop_nomask <- model3a$coefficients[1] + model3a$coefficients[2]*avg_pop + model3a$coefficients[3]*avg_
y_pov_nomask <- model3a$coefficients[1] + model3a$coefficients[2]*avg_pop + model3a$coefficients[3]*df$p
```

```r
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'
  )
```




```r
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.

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.

Finally, 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. Conceptually, the variables which may have an important colinearity in the final model are variables such as different percentage make ups of race and ethnicity, or related public policy. These variables are appropriately evaluated and excluded throughout the model building process.

4. Regression Table

Although there were several versions of models created for model 2 and 3, through the analysis of the appropriate variables respective to each model a single model version was selected to be included in the regression table. This was determined based on factors of significance when investigating the effect relationship of these variables on the outcome variable, which is detailed in section 2.

For the regression table below, model 1, model 2-b, and model 3-a were selected as the most significant models to include.

```
se.model1 = coeftest(model1, vcov = vcovHC)[ , "Std. Error"]
se.model2 = coeftest(model2b, vcov = vcovHC)[ , "Std. Error"]
se.model3 = coeftest(model3a, vcov = vcovHC)[ , "Std. Error"]

stargazer(model1, model2b, model3a, type = "text", omit.stat = "f",
          se = list(se.model3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 1: The relationship between COVID case rate and population density")
```

```
##
## Table 1: The relationship between COVID case rate and population density
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                             case_rate_100k
##                                  (1)              (2)              (3)
## --------------------------------------------------------------------------------
## log10(population_density)      -179.220         -189.854          31.377
##                                (206.451)        (241.530)        (233.878)
##
## poverty_pct                                     112.241*         104.191*
##                                                 (55.125)         (50.872)
##
## mask_public_bool                                                 -998.485**
##                                                                  (323.790)
##
## Constant                      3,103.375**       1,675.148        2,027.409*
##                               (1,003.465)       (862.256)        (802.855)
##
## --------------------------------------------------------------------------------
## Observations                      51                51               51
## R2                               0.010             0.089            0.242
## Adjusted R2                      -0.010            0.051            0.194
## Residual Std. Error       1,142.031 (df = 49) 1,107.048 (df = 48) 1,020.297 (df = 47)
## ================================================================================
## 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. 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.