Lab 02: COVID-19 Case Rate vs Age Distribution

Mrinal Chawla, YuCheng Liu & Roy Katende

12/09/2020

Introduction

In the spring of 2020, the COVID-19 pandemic washed over the USA. Spread by an airborne virus, the disease caused acute respiratory issues, leading to hospitalization and death in the most severe cases. Initially, scientists believed that elderly individuals were more at risk of contracting COVID-19. However, as the pandemic stretched over months, there were cases of people of all ages suffering from the disease. Therefore, it would be useful to describe how age distribution within each state of the USA affects the case rate of COVID-19. A descriptive model can help detail which age groups in a population were the most affected for each state. This information can then be used to plan targeted relief programs during the pandemic, and recovery programs after a vaccine is developed. Knowing how the population was affected is crucial to providing care.

In order to build this descriptive model, data from the COVID-19 US State Policy database (Raifman J, Nocka K, Jones D, Bor J, Lipson S, Jay J, and Chan P.) is used. The model outputs the expected number of cases per 100,000 individuals, using the age distribution of each state's residents. The reason that case rate per 100,000 is used instead of the total case count is that the total population in each state is different. For a huge state like California, the total case count might be even bigger than the total population of a small state. In this case, the model will be significantly skewed by the outliers of the huge and tiny states. Thankfully, case rate per 100,000 is already normalized based on population, so for each state the baseline will be the same. The age distribution data is provided in the categories of 0-18, 19-25, 26-34, 35-54, 55-64, and 65+ years of age, represented as a percentage of the state population. In order to supplement the model, the population density, and number of adults with pre-existing conditions is also considered, as these are also likely to have a factor in describing the number of cases of COVID-19 in a state. Finally, some state policies are also included, such as mask usage, business closure, and testing rate, as these policies varied from state to state, and may have contributed to fewer cases in some states. By including these in the model, the effect of the age distribution can be isolated from other factors.

However, this model is limited by the granularity of the data. The age distribution is only available in the categories described above, and thus does not provide descriptive power for finer tuned age differences. Furthermore, when considering adults with pre-existing conditions, age is not considered. So if a state has more elderly with conditions, rather than middle-aged individuals with conditions, the model will not capture that difference. Regardless of these limitations, the description of age group to case rate is still valuable to describe how people in the USA were affected by the COVID-19 pandemic, as individuals in the given age groups are likely to be biologically similar. This model can be a useful descriptive tool for each state to have a baseline on how their population was affected by the COVID-19 pandemic.

Model Building

Exploratory Data Analysis

```
#### AUTHORS: MRINAL CHAWLA, YUSUF LIU, ROY KATENDE
# install.packages('fec16')
library(tidyverse)
library(magrittr)
library(ggplot2)
library(patchwork)
library(sandwich)
library(lmtest)
# library(fec16)
library(lmtest)
library(readxl)
library(stargazer)
#setwd("~/lab02_week14")
covid <- read_excel("covid-19-v5.xlsx")
#summary(covid$‘Total Cases‘)
#summary(covid$‘Case Rate per 100000‘)
#summary(covid$‘State‘)
#summary(covid$‘Closed Business Days‘)
#summary(covid$‘Tests per 100K‘)
#summary(covid$‘Death Rate per 100000‘)
#summary(covid$‘Adults 35-54‘)
#summary(covid$‘Adults 55-64‘)
#summary(covid$‘Children 0-18‘)
#summary(covid$‘Adults 19-25‘)
#summary(covid$‘Adults 26-34‘)
#summary(covid$‘65+‘)
#summary(covid$‘Number Homeless (2019)‘)
#summary(covid$‘Total Deaths‘)
#summary(covid$‘Case Rate per 100000 in Last 7 Days‘)
#summary(covid$‘End stay at home/shelter in place‘)
#summary(covid$‘Shelter in Place Days‘)
#summary(covid$‘Population density per square miles‘)
#summary(covid$‘Nonelderly Adults Who Have A Pre-Existing Condition‘)
#summary(covid$‘Population 2018‘)
#summary(covid$‘Mandate face mask use by all individuals in public spaces‘)
#summary(covid$‘Death Rate per 100000‘)
covid$‘Pop. density / sq mile‘ <- covid$‘Population density per square miles‘
covid$‘Nonelderly Adults w/ Condition‘ <- covid$‘Nonelderly Adults Who Have A Pre-Existing Condition‘
covid$‘Mandate face mask‘ <- covid$‘Mandate face mask use by all individuals in public spaces‘
cov_sum <- covid
```

Variable Transformations

```
closed_business_days = covid$‘Began to reopen businesses statewide‘ -
   covid$‘Closed other non-essential businesses‘
```

As a part of the data transformation and cleansing process, the "Closed Business Days" variable was created. It is the difference in days between the date variables "Began to reopen businesses statewide" and "Closed other non-essential businesses". This variable is used in the latter models below, as a measure of the state's policy regarding businesses. It is analyzed to determine if length of closure had a significant effect on case rate.

```
#covid[covid$‘End stay at home/shelter in place‘ == "0"] <- "09/28/20"
shelter_in_place_days = covid$‘End stay at home/shelter in place‘ -
   covid$‘Stay at home/ shelter in place‘
```

Similarily, the "Shelter in Place Days" variable was created. It is the difference in days between the date variables "End stay at home/shelter in place" and "Stay at home/ shelter in place". It should be noted that about 10% of the States did not have one or both of these dates available, where "Stay at home/ shelter in place" was not available, the data collection date "09/28/20" was used as a proxy. The reason for getting the difference between the start and end of stay at home was to be able to measure the impact of the number of cases under the various lockdown periods for states. Numerous states did not issue lockdowns, which explains why some of the dates were not available.

Variable Exploration

The histogram below (Figure 1) is of the output variable Case Rate per 100,000. This is the variable that is being described. The distribution of the data is approximately normal.

```
# EDA of Variables used in the Basic Model

my_plot_hook <- function(x, options)
  paste("\n", knitr::hook_plot_tex(x, options), "\n")
knitr::knit_hooks$set(plot = my_plot_hook)

hist(covid$‘Case Rate per 100000‘, main="Case Rate per 100,000", xlab="Case Rate per 100000",
     col = "dark green")
```

The histograms below are of the age distribution and are normally distributed except for the age group between 26 and 34.

```
hist(covid$‘Children 0-18‘, main="Children 0-18 Age Group", xlab="Children 0-18", col = "orange")
```

```
hist(covid$‘Adults 19-25‘, main="Adults 19-25 Age Group", xlab="Adults 19-25", col = "orange")
```

```
hist(covid$‘Adults 26-34‘, main="Adults 26-34 Age Group", xlab="Adults 26-34", col = "orange")
```

```
hist(covid$‘Adults 35-54‘, main="Adults 35-54 Age Group", xlab="Adults 35-54", col = "orange")
```

```
hist(covid$‘Adults 55-64‘, main="Adults 55-64 Age Group", xlab="Adults 55-64", col = "orange")
```

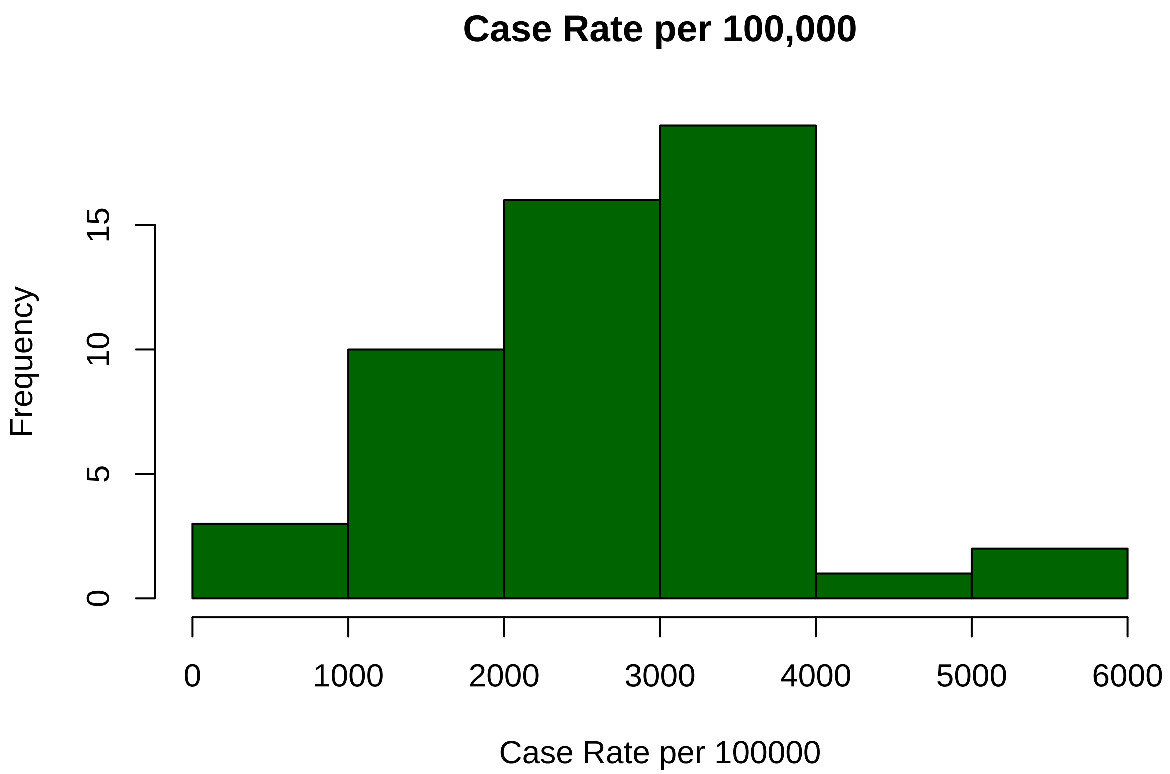

Figure 1: Distribution of Case Rate across States

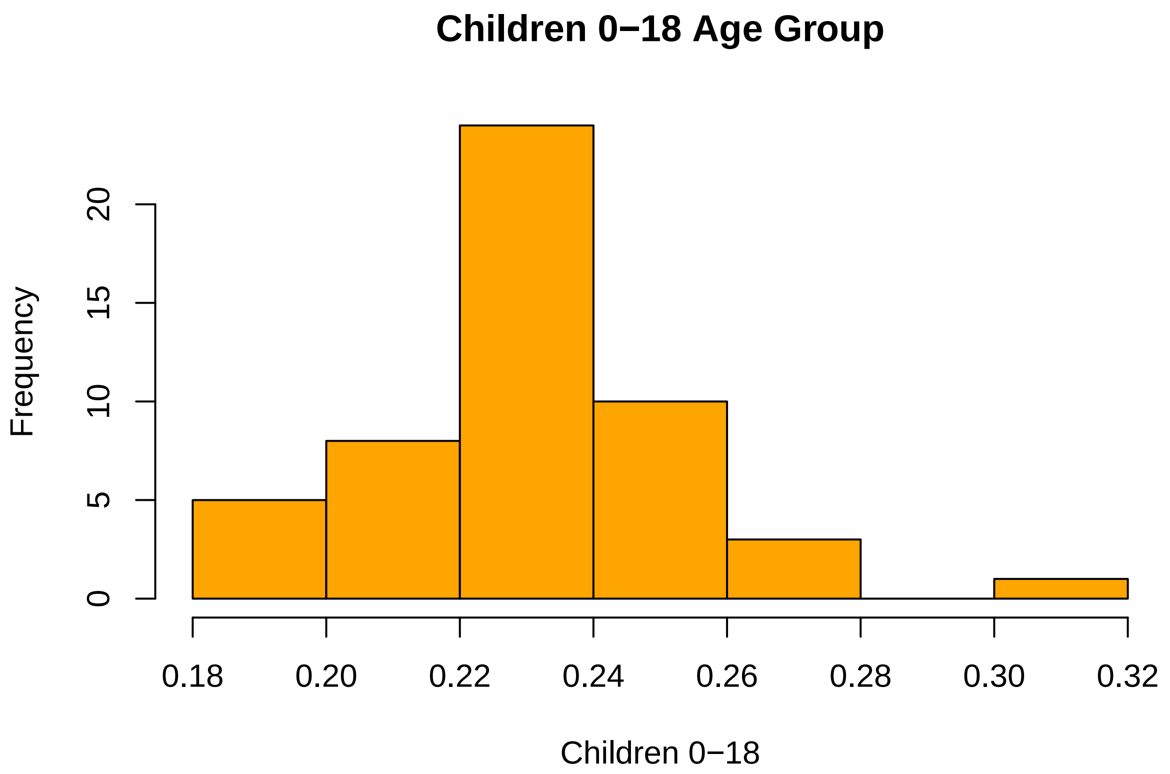

Figure 2: Distribution of Age Groups across States

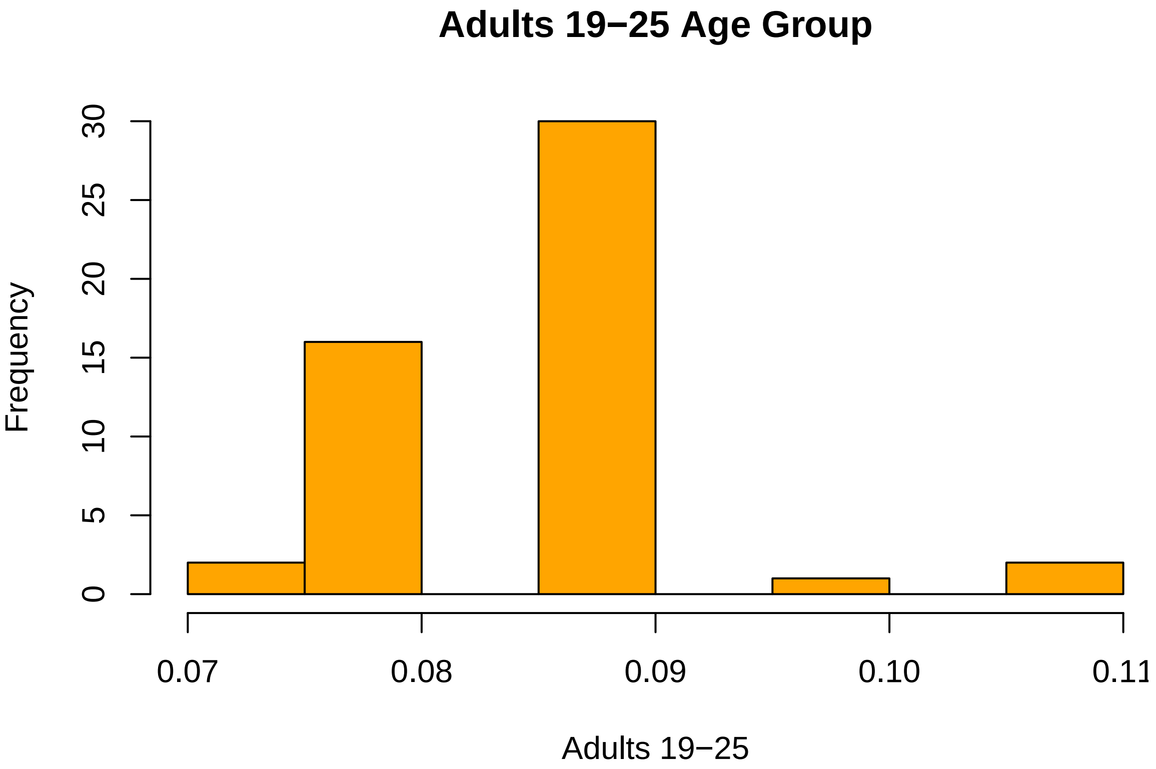

Figure 3: Distribution of Age Groups across States

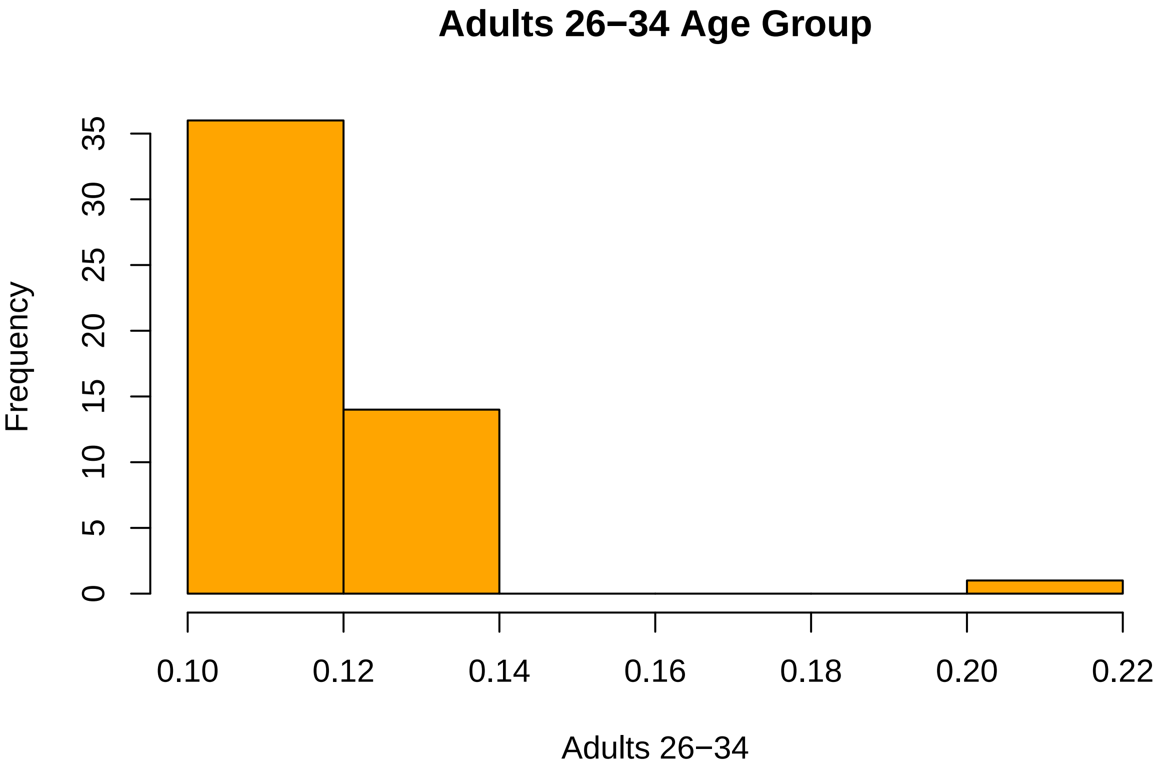

Figure 4: Distribution of Age Groups across States

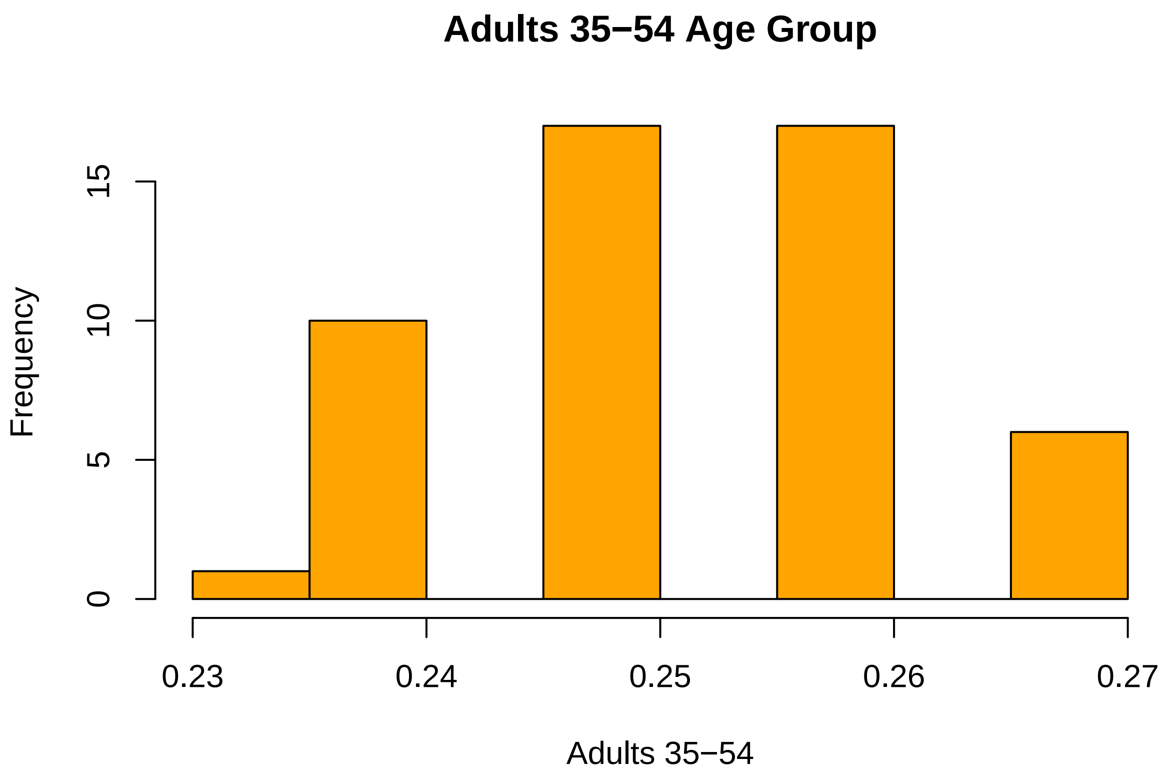

Figure 5: Distribution of Age Groups across States

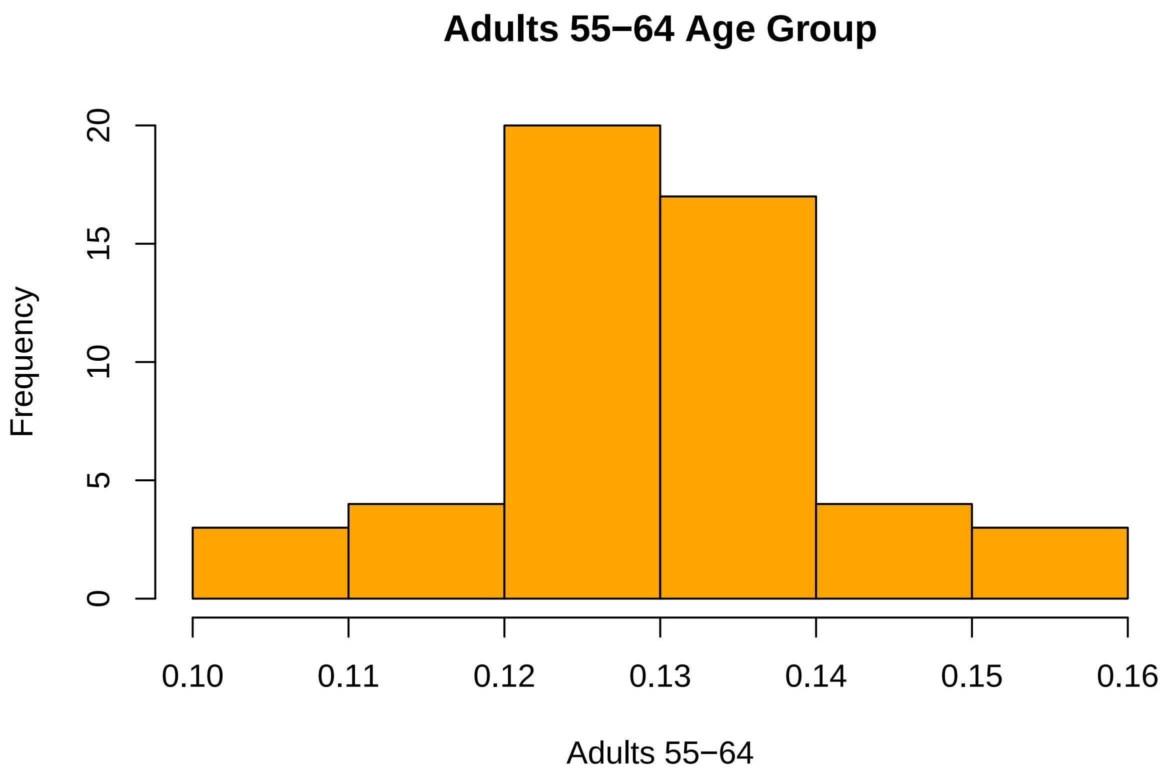

Figure 6: Distribution of Age Groups across States

```
hist(covid$‘65+‘, main="65+ Age Group", xlab="65+", col = "orange")
```

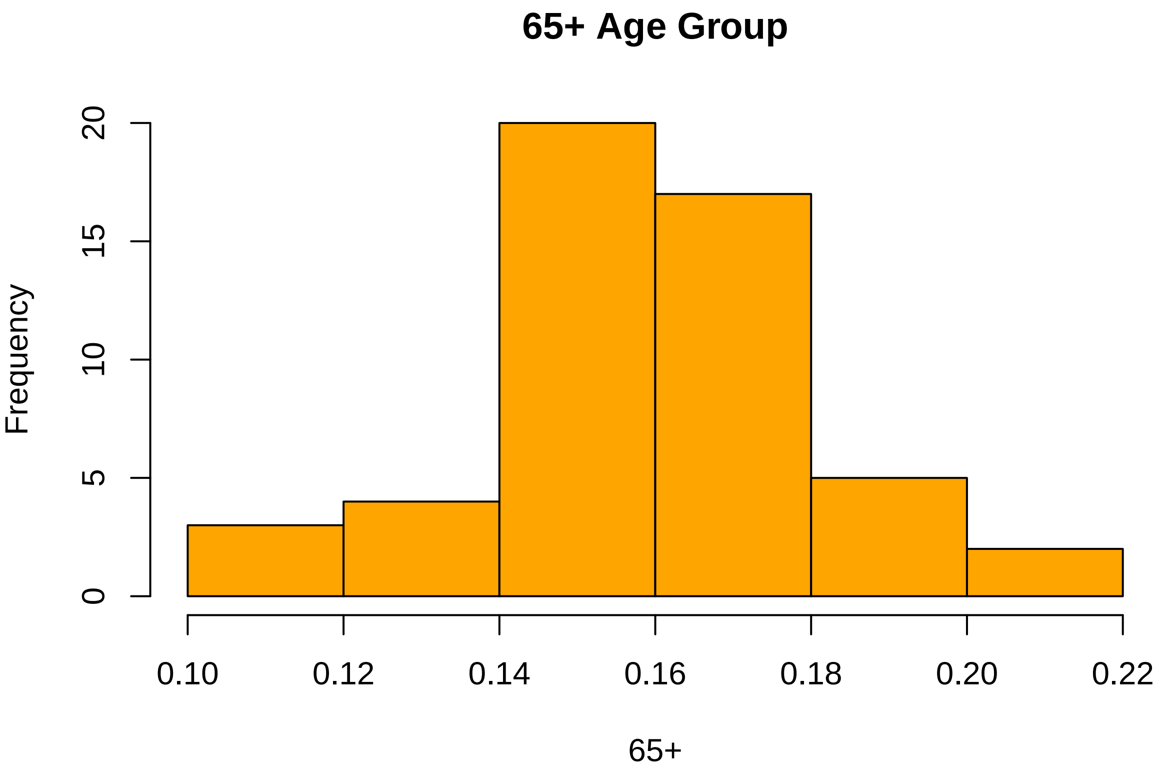

Figure 7: Distribution of Age Groups across States

Figures 2 through 3 show a normal distribution of data. Figure 4 shows that the percentage of young adults is normally low but there is an outlier in this set. The District of Columbia has a higher population of 26 to 34 year olds than other states. Since this is valid data, further transformation is not done at this point. The implications of the outliers are discussed in the model evaluation below. The other age distributions in Figures 5 through 7 show relatively normal distributions across the states. Thus no further transformation is done for the age disitribution variables.

In order to understand the additional data used for the Advanced Model, several histograms were examined. The population density per square mile (Figure 8) showed that there are outliers due to the density of New York and D.C. The rest of the country had a lower population density. As this is valid data, it is still included in the analysis. A log transformation is done (Figure 9) to handle the large range of values. This improves the normality of the distribution, and will be the variable used in the analysis.

In addition to other variables, the Nonelderly Adults with Pre-Existing Conditions was also examined (Figure 10). It was fairly normal, as was the transformed variable of Closed Business Days (Figure 11). These variables are utilized in the Advanced and Detailed models below.

```
# EDA of Variables used in the Advanced Model
hist(covid$‘Pop. density / sq mile‘,
     main="Population density per square miles",
     xlab="Population density per square miles",
     col = "red")
```

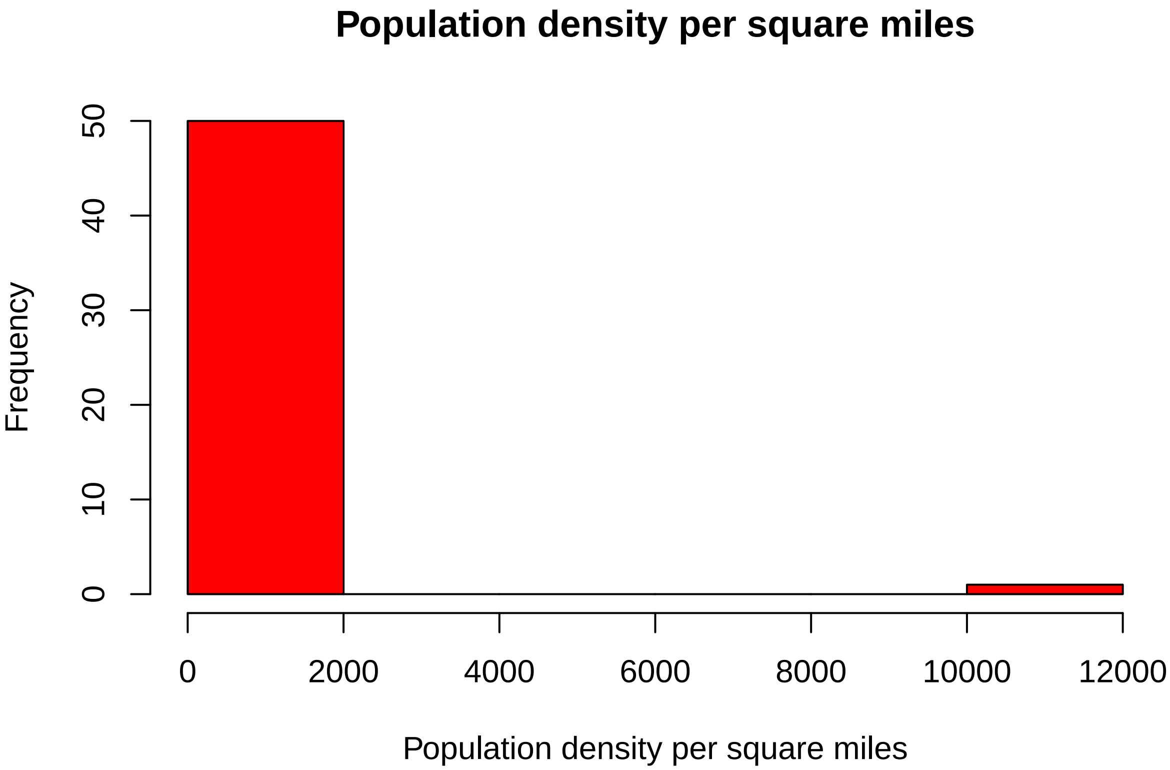

Figure 8: Distribution of Pop. Density across States

```
# EDA of Variables used in the Advanced Model
hist(log(covid$‘Pop. density / sq mile‘),
     main="Population density per square miles",
     xlab="log(Population density per square miles)",
     col = "red")
```

```
hist(log(covid$‘Nonelderly Adults w/ Condition‘),
     main="Nonelderly Adults Who Have A Pre-Existing Condition",
     xlab="Nonelderly Adults With Pre-Existing Condition",
     col = "tan")
```

```
hist(covid$‘Closed Business Days‘, main="Closed Business Days", xlab="Closed Business Days", col = "gol
```

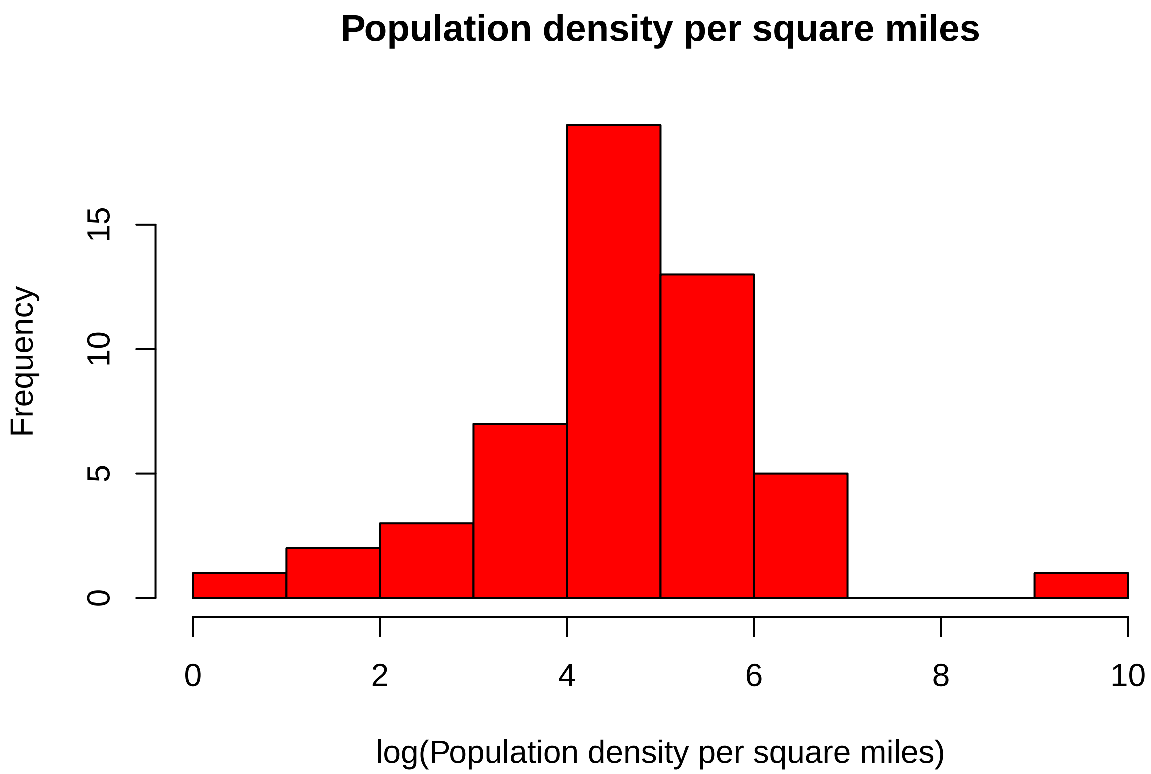

Figure 9: Distribution of log(Pop. Density) across States

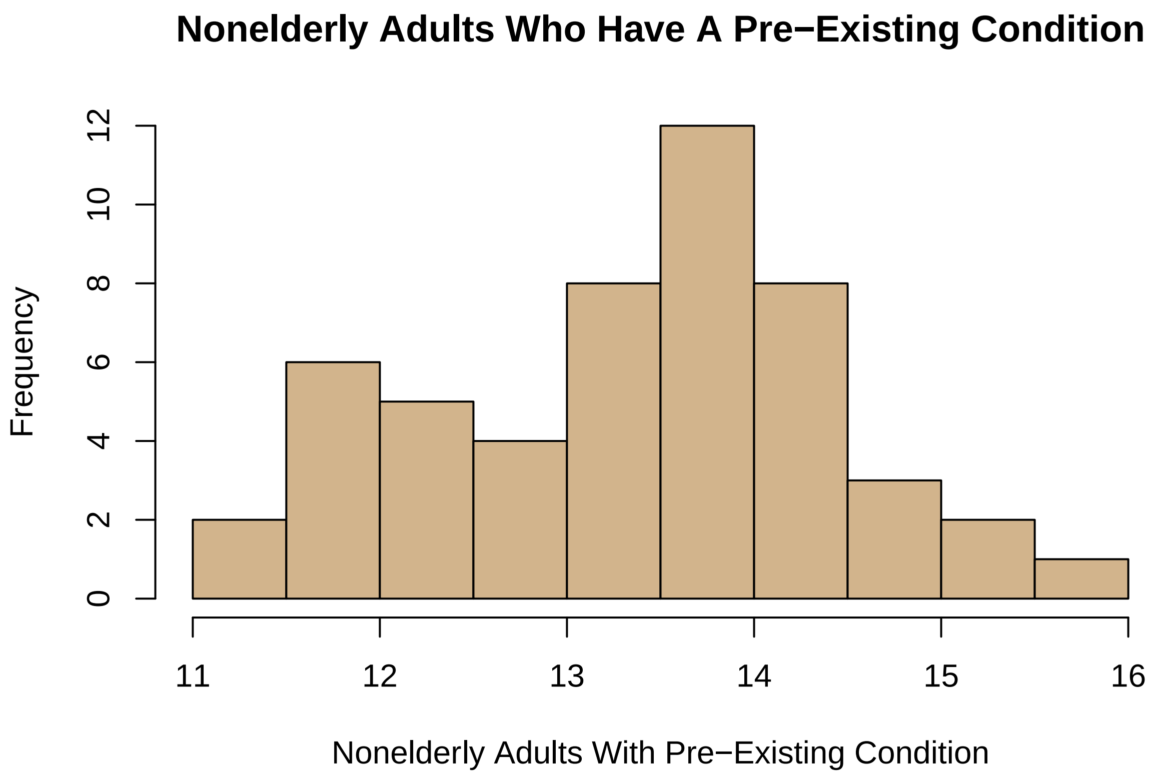

Figure 10: Distribution of Adults with Condition across States

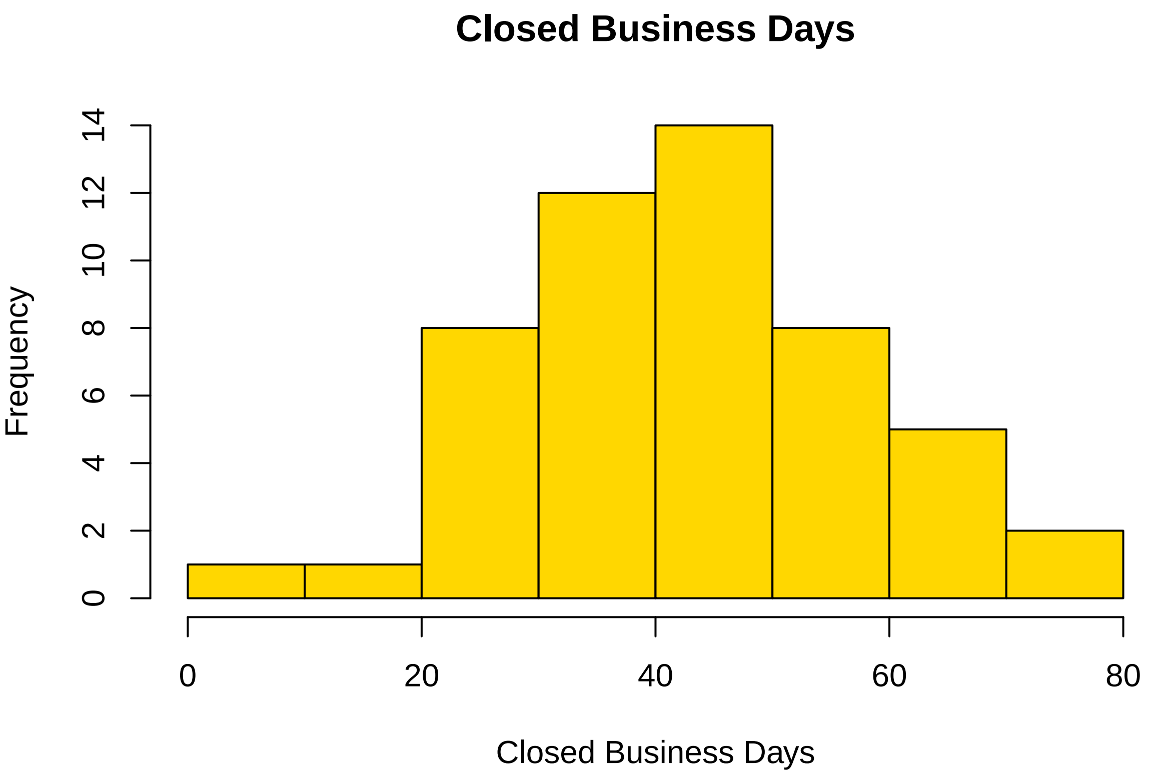

Figure 11: Distribution of Business Closure Length across States

The following variables are used only for the Detailed Model. The histogram in Figure 12 is the distribution of the number of tests per 100K people. Based on the graph, for 90% of the states, tests per 100K falls within 20000 to 60000. No further transformation is required.

```
# EDA of Variables used in the Detailed Model
hist(covid$‘Tests per 100K‘, main="Tests per 100K", xlab="Tests per 100K", col = "azure2")
```

The histogram in Figure 13 is on a binary data ”Mandate face mask use by all individuals in public spaces", which indicates either Yes or No. Therefore, there are two bars at 0 and 1 Based on the graph, 40% of the states had no mandate mask, and the other 60% had mandate mask in public spaces. This variable will serve as an indicator variable for mask mandate in the Detailed Model below.

```
covid$‘Mandate face mask‘ = ifelse(covid$‘Mandate face mask‘ > 0, 1, 0)
hist(covid$‘Mandate face mask‘, main="Mandate face mask use by all individuals in public spaces",
     xlab="Mandate face mask use by all individuals", breaks = 2, col = "cornsilk3")
```

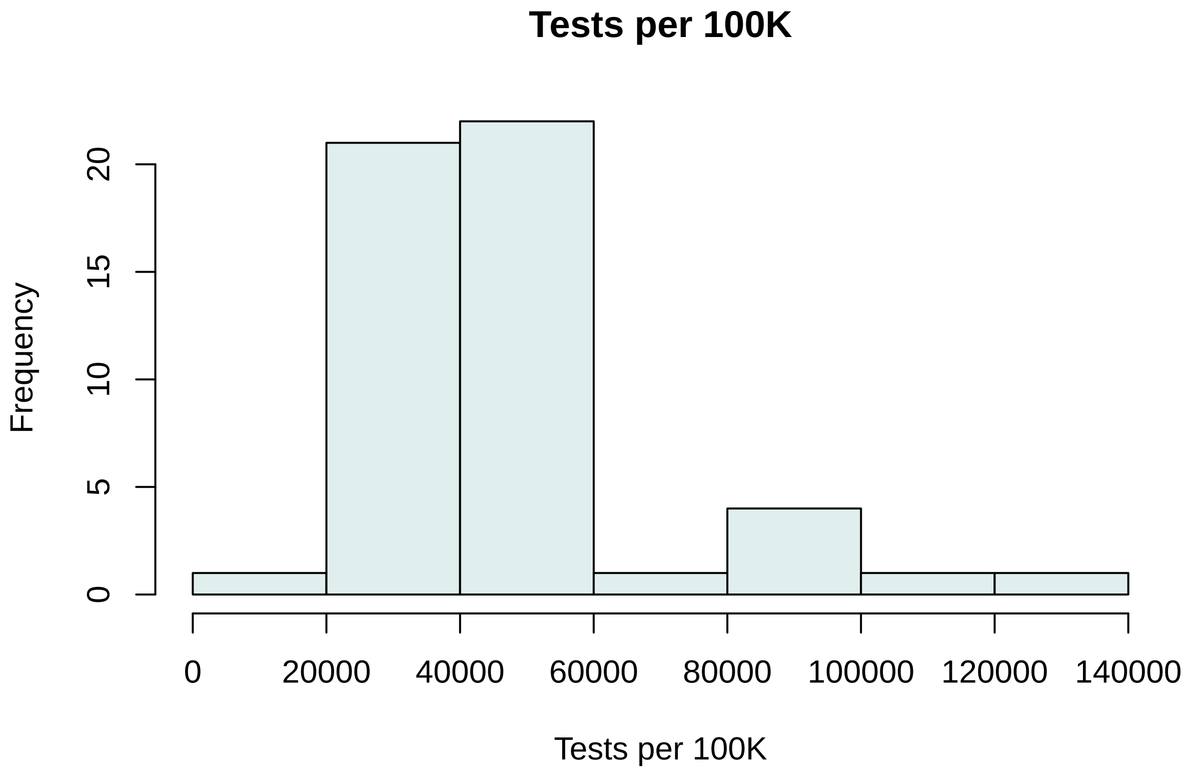

Figure 12: Distribution of Testing Rate across States

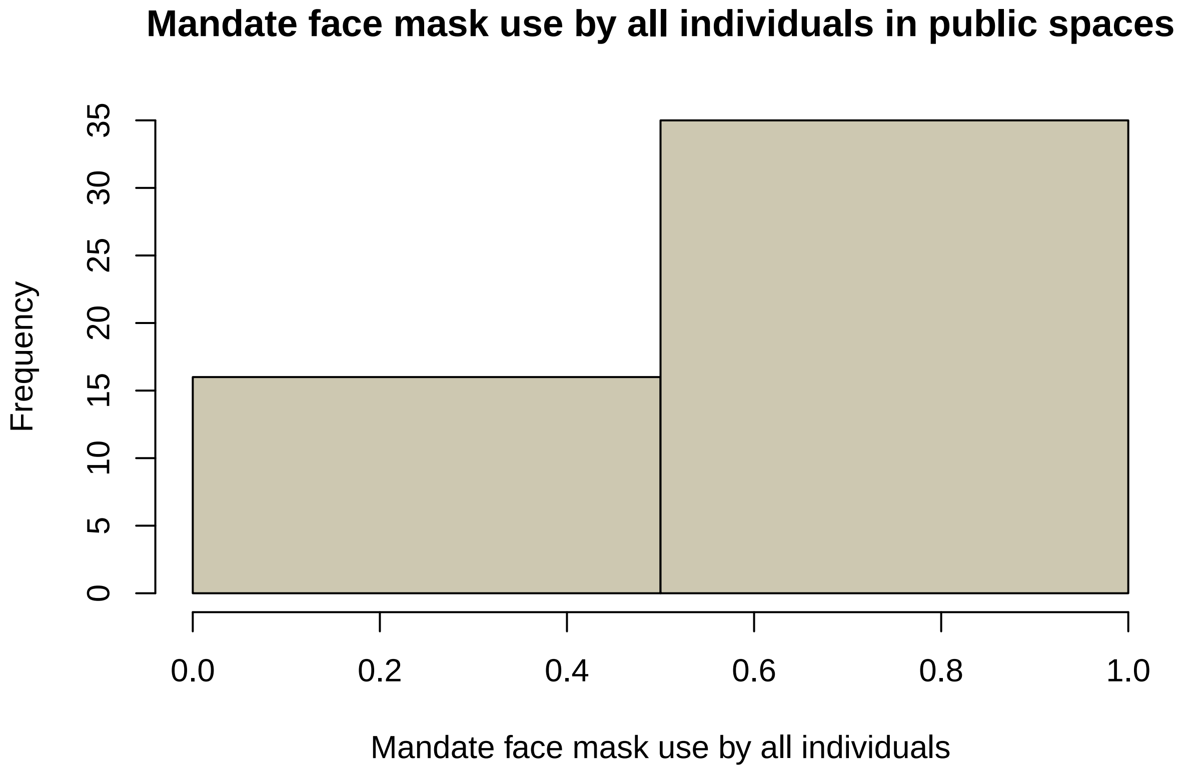

Figure 13: Distribution of Face Mask Mandate across States

The Basic Model:

The Basic Model comprises of the output variable “Case Rate per 100,000” and the input variables that are the age groups of the population from 0 years of age to over 65 years. The summary of the model is included below.

```
#### MODEL ONE - BASIC MODEL
fit01 = lm(‘Case Rate per 100000‘~‘Children 0-18‘
             +‘Adults 19-25‘
             +‘Adults 26-34‘
             +‘Adults 35-54‘
             +‘Adults 55-64‘
             +‘65+‘
             , data=cov_sum)

# The Summary of the Model Characteristics
summary(fit01)
```

```
##
## Call:
## lm(formula = ‘Case Rate per 100000‘ ~ ‘Children 0-18‘ + ‘Adults 19-25‘ +
##     ‘Adults 26-34‘ + ‘Adults 35-54‘ + ‘Adults 55-64‘ + ‘65+‘,
##     data = cov_sum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1699.3  -650.1   124.6   588.8  2589.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -31121      20631  -1.508 0.138581
## ‘Children 0-18‘    33583      21588   1.556 0.126954
## ‘Adults 19-25‘    114863      29735   3.863 0.000364 ***
## ‘Adults 26-34‘     10520      23501   0.448 0.656601
## ‘Adults 35-54‘     27674      26008   1.064 0.293105
## ‘Adults 55-64‘     16920      27390   0.618 0.539944
## ‘65+‘              32487      21726   1.495 0.141984
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 897.5 on 44 degrees of freedom
## Multiple R-squared:  0.4512, Adjusted R-squared:  0.3764
## F-statistic: 6.029 on 6 and 44 DF,  p-value: 0.0001155
```

The F-test in the summary shows that this model does hold significant descriptive power. However, there is only one significant variable, the percentage of Adults in the 19-25 age range. Based on this analysis, other age distributions could be dropped. However, they will be retained in this report, for consistency with the later models.

```
###scatter plot

plot(covid$`Adults 19-25`, covid$`Case Rate per 100000`, main="Scatterplot Of Case Rate per 100000 vs Ac
   xlab="Adults 19-25", ylab="Case Rate per 100000", pch=19)
abline(lm(covid$`Case Rate per 100000` ~ covid$`Adults 19-25`, data = covid), col = "blue")
```

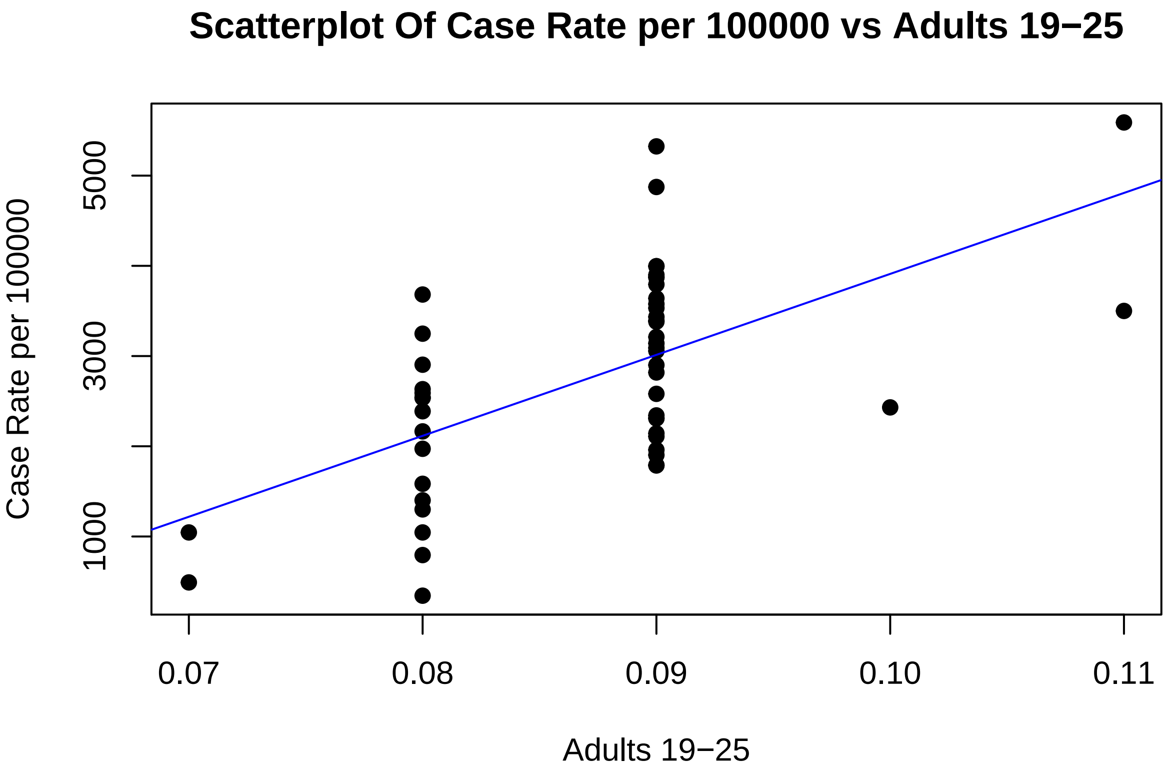

Figure 14: Scatterplot showing Case Rate per 100,000 as predicted by Young Adults Between 19 and 25

The input variable Adults 19-25 which is significant in the model, shows an uphill pattern from left to right (Figure 14), indicating a positive relationship between the percentage of population in the age group and the case rate in the state.

The practical significance of this model shows that the young adults in this age group need to be observed. They fall within the college age bracket or recently graduated. The kind of employment they usally have access to involves contact with a large number people of all ages. These jobs could be in indistries such as retail, restaurant, etc where it is easy to come in contact with infected customers and they in turn transmit the disease to family and friends - unbeknownst to them. It could also be that since they are healthy for the most part - they may not follow CDC guidelines like wearing masks and avoiding large gatherings. The model suggests that governments should be aware of individuals in this age range and their effect on the case rate of COVID-19.

The Advanced Model

The Advanced Model comprises of the output variable "Case Rate per 100,000", the input variables from the Basic Model and the log of Population density per square miles, Nonelderly Adults Who Have A Pre-Existing Condition and the transformed variable Closed Business Days. The summary of the model is included.

```
## MODEL TWO

fit02 = lm(‘Case Rate per 100000‘~ ‘Children 0-18‘
             +‘Adults 19-25‘
             +‘Adults 26-34‘
             +‘Adults 35-54‘
             +‘Adults 55-64‘
             +‘65+‘
             +log(‘Pop. density / sq mile‘)
             +‘Nonelderly Adults w/ Condition‘
             +‘Closed Business Days‘
             , data=cov_sum)

# The Summary of the Model Characteristics
summary(fit02)
```

```
##
## Call:
## lm(formula = ‘Case Rate per 100000‘ ~ ‘Children 0-18‘ + ‘Adults 19-25‘ +
##     ‘Adults 26-34‘ + ‘Adults 35-54‘ + ‘Adults 55-64‘ + ‘65+‘ +
##     log(‘Pop. density / sq mile‘) + ‘Nonelderly Adults w/ Condition‘ +
##     ‘Closed Business Days‘, data = cov_sum)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1577.06  -645.36    92.25   502.20  2467.35
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -3.482e+04  2.122e+04  -1.641   0.1085
## ‘Children 0-18‘                   4.281e+04  2.392e+04   1.790   0.0809 .
## ‘Adults 19-25‘                    1.081e+05  3.078e+04   3.511   0.0011 **
## ‘Adults 26-34‘                    2.066e+04  2.473e+04   0.836   0.4082
## ‘Adults 35-54‘                    1.908e+04  2.734e+04   0.698   0.4891
## ‘Adults 55-64‘                    3.505e+04  3.128e+04   1.120   0.2690
## ‘65+‘                             3.378e+04  2.205e+04   1.532   0.1331
## log(‘Pop. density / sq mile‘)     1.297e+02  1.363e+02   0.951   0.3470
## ‘Nonelderly Adults w/ Condition‘  1.361e-04  1.240e-04   1.098   0.2788
## ‘Closed Business Days‘           -8.086e+00  1.000e+01  -0.808   0.4236
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 899 on 41 degrees of freedom
## Multiple R-squared:  0.4869, Adjusted R-squared:  0.3743
## F-statistic: 4.323 on 9 and 41 DF,  p-value: 0.0005297
```

This model also has descriptive power, as shown by the results of the F-test in the summary. The two significant variables appear to be Children 0-18 and Adults 19-25. This model suggests that the percentage of population in the two lowest age categories are the best indicators for case rate in a state. Similar to the Basic Model, there could be many reasons for this. Practically, this indicates that states need to be aware of the effect of youth in general on the spread of COVID-19.

A chart showing the relationship of age group to case rate would be similar to Figure 14, so it is not included here.

The Detailed Model

The Detailed Model comprises of the output variable “Case Rate per 100,000”, the input variables from the Advanced Model, “Tests per 100K” and “Mandate face mask use by all individuals in public spaces”. The summary of the model is included.

```
#### MODEL THREE
fit03 = lm(‘Case Rate per 100000‘~
             ‘Children 0-18‘
             +‘Adults 19-25‘
             +‘Adults 26-34‘
             +‘Adults 35-54‘
             +‘Adults 55-64‘
             +‘65+‘
             +log(‘Pop. density / sq mile‘)
             +‘Nonelderly Adults w/ Condition‘
             +‘Closed Business Days‘
             +‘Tests per 100K‘
             +‘Mandate face mask‘
             , data=cov_sum)
#fit03

# The Summary of the Model Characteristics
summary(fit03)
```

```
##
## Call:
## lm(formula = ‘Case Rate per 100000‘ ~ ‘Children 0-18‘ + ‘Adults 19-25‘ +
##     ‘Adults 26-34‘ + ‘Adults 35-54‘ + ‘Adults 55-64‘ + ‘65+‘ +
##     log(‘Pop. density / sq mile‘) + ‘Nonelderly Adults w/ Condition‘ +
##     ‘Closed Business Days‘ + ‘Tests per 100K‘ + ‘Mandate face mask‘,
##     data = cov_sum)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1404.52  -423.83    58.21   364.72  1321.00
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -4.760e+04  1.737e+04  -2.741 0.009189 **
## ‘Children 0-18‘                   6.462e+04  1.993e+04   3.242 0.002433 **
## ‘Adults 19-25‘                    8.615e+04  2.574e+04   3.347 0.001816 **
## ‘Adults 26-34‘                    2.514e+04  2.004e+04   1.254 0.217157
## ‘Adults 35-54‘                    3.937e+04  2.264e+04   1.739 0.089924 .
## ‘Adults 55-64‘                    3.579e+04  2.559e+04   1.398 0.169923
## ‘65+‘                             4.964e+04  1.831e+04   2.711 0.009916 **
## log(‘Pop. density / sq mile‘)     2.317e+02  1.133e+02   2.045 0.047636 *
## ‘Nonelderly Adults w/ Condition‘  1.394e-04  1.004e-04   1.389 0.172850
## ‘Closed Business Days‘           -7.448e+00  8.639e+00  -0.862 0.393877
## ‘Tests per 100K‘                  2.450e-02  6.292e-03   3.893 0.000376 ***
## ‘Mandate face mask‘              -1.815e-02  5.852e-03  -3.101 0.003576 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 726.8 on 39 degrees of freedom
## Multiple R-squared:  0.681,  Adjusted R-squared:  0.591
## F-statistic: 7.569 on 11 and 39 DF,  p-value: 9.118e-07
```

This model has descriptive power as well, based on the results of the F-test. In this model, the additional variables have significant descriptive power. As they describe the additional variance, some of the original age distributions also become significant, based on the results of the t-tests. The significance level of the younger age distributions has increased. The significance of Tests per 100K and face mask mandates is also relevant to the model.

Practically, this model confirms that younger age distributions have the most significance in describing the number of cases of COVID-19 in states. These are the age distributions that governments should be aware about, when assessing how COVID-19 affected their population. Furthermore, the model controls for the number of tests in each state. The t-test confirms the significance of this variable. Therefore, for a state to have accurate case numbers, testing is required. Likewise, mask mandates have also shown to have a significant effect on COVID-19 case rate.

```
###scatter plot

plot(covid$‘Children 0-18‘, covid$‘Case Rate per 100000‘,
     main="Scatterplot Of Case Rate per 100000 vs Children 0-18",
   xlab="Children 0-18", ylab="Case Rate per 100000", pch=19)
abline(lm(covid$‘Case Rate per 100000‘ ~ covid$‘Children 0-18‘, data = covid), col = "blue")
```

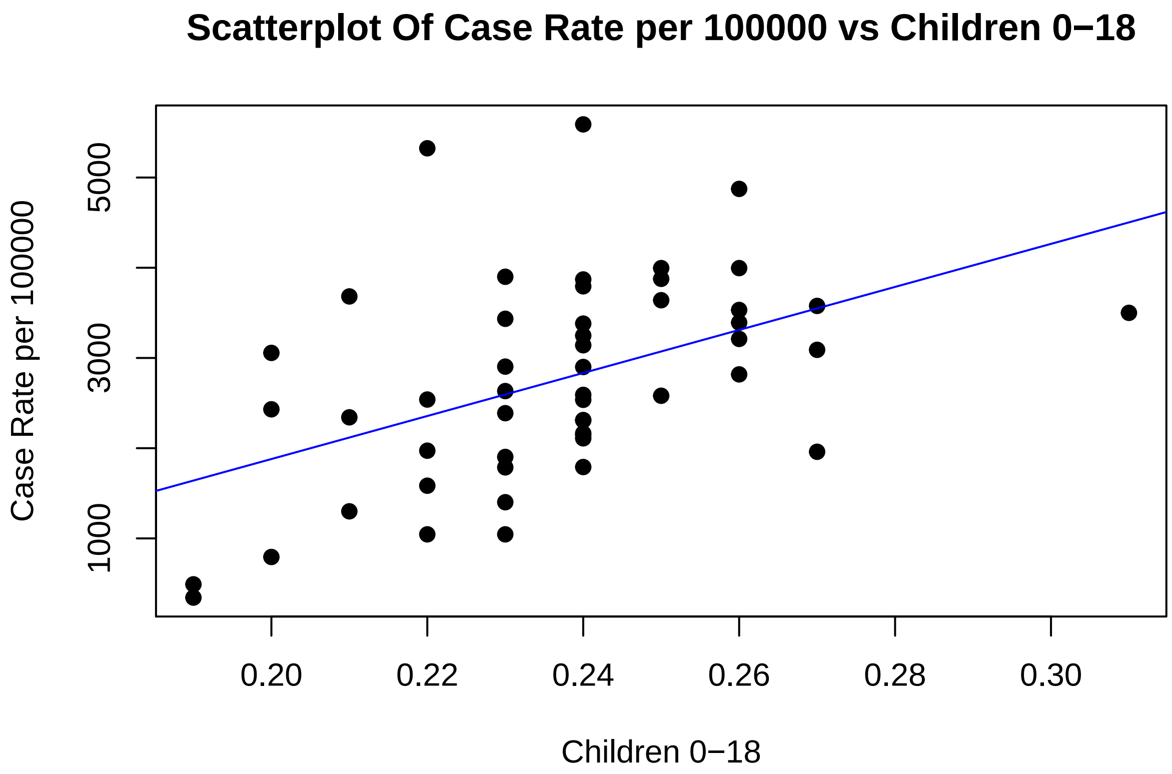

Figure 15: Scatterplot showing Case Rate per 100,000 as predicted by Significant Vaiables in the Detailed Model

```
plot(covid$‘Adults 35-54‘, covid$‘Case Rate per 100000‘, main="Scatterplot Of Case Rate per 100000 vs Ad
   xlab="Adults 35-54", ylab="Case Rate per 100000", pch=19)
abline(lm(covid$‘Case Rate per 100000‘ ~ covid$‘Adults 35-54‘, data = covid), col = "blue")
```

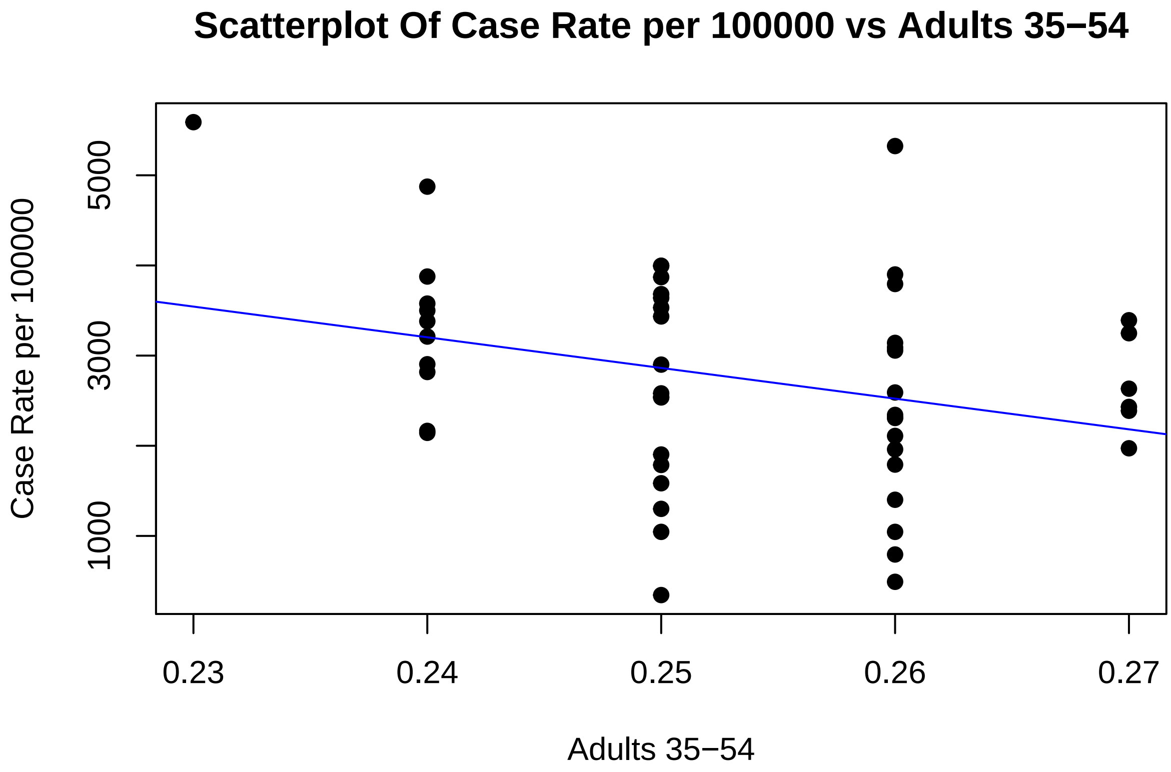

Figure 16: Scatterplot showing Case Rate per 100,000 as predicted by Significant Vaiables in the Detailed Model

```
plot(covid$‘65+‘, covid$‘Case Rate per 100000‘, main="Scatterplot Of Case Rate per 100000 vs 65+",
   xlab="65+", ylab="Case Rate per 100000", pch=19)
abline(lm(covid$‘Case Rate per 100000‘ ~ covid$‘65+‘, data = covid), col = "blue")
```

```
plot(covid$‘Tests per 100K‘, covid$‘Case Rate per 100000‘, main="Scatterplot Of Case Rate per 100000 vs
   xlab="Tests per 100K", ylab="Case Rate per 100000", pch=19)
abline(lm(covid$‘Case Rate per 100000‘ ~ covid$‘Tests per 100K‘, data = covid), col = "blue")
```

```
plot(covid$‘Mandate face mask‘, covid$‘Case Rate per 100000‘, main="Scatterplot Of Case Rate per 100000
   xlab="Mandate face mask", ylab="Case Rate per 100000", pch=19)
abline(lm(covid$‘Case Rate per 100000‘ ~ covid$‘Mandate face mask‘, data = covid), col = "blue")
```

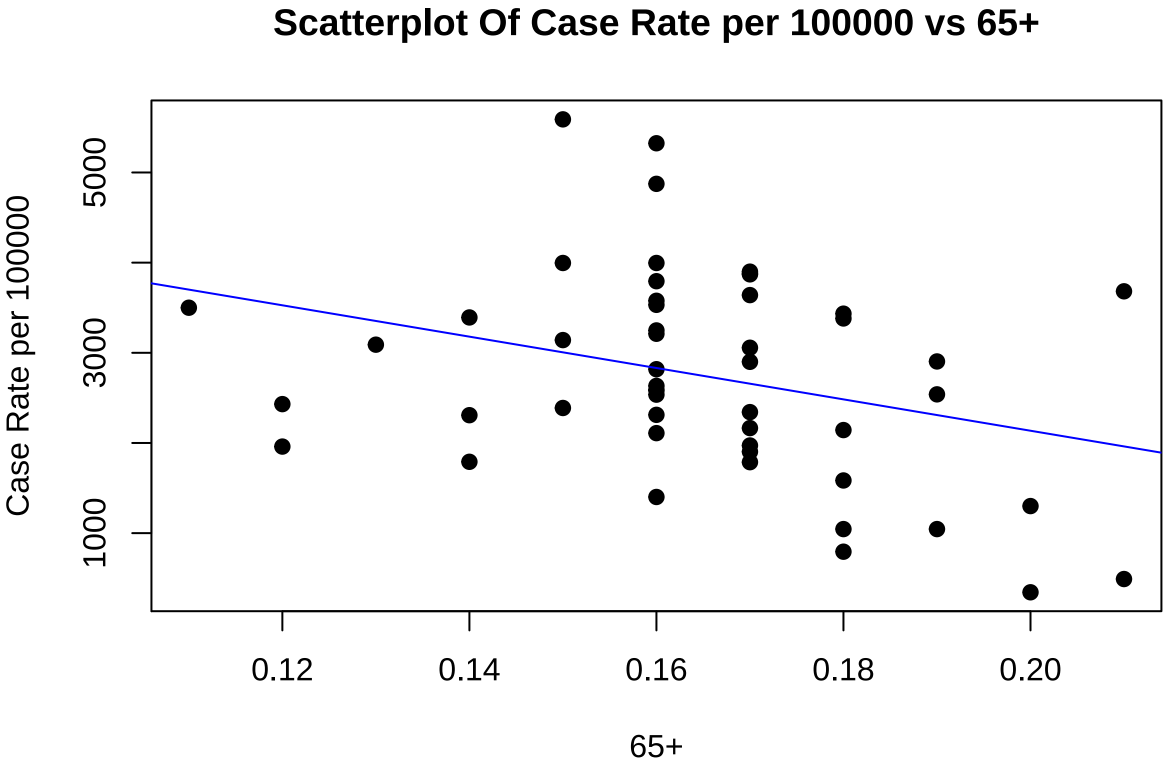

Figure 17: Scatterplot showing Case Rate per 100,000 as predicted by Significant Vaiables in the Detailed Model

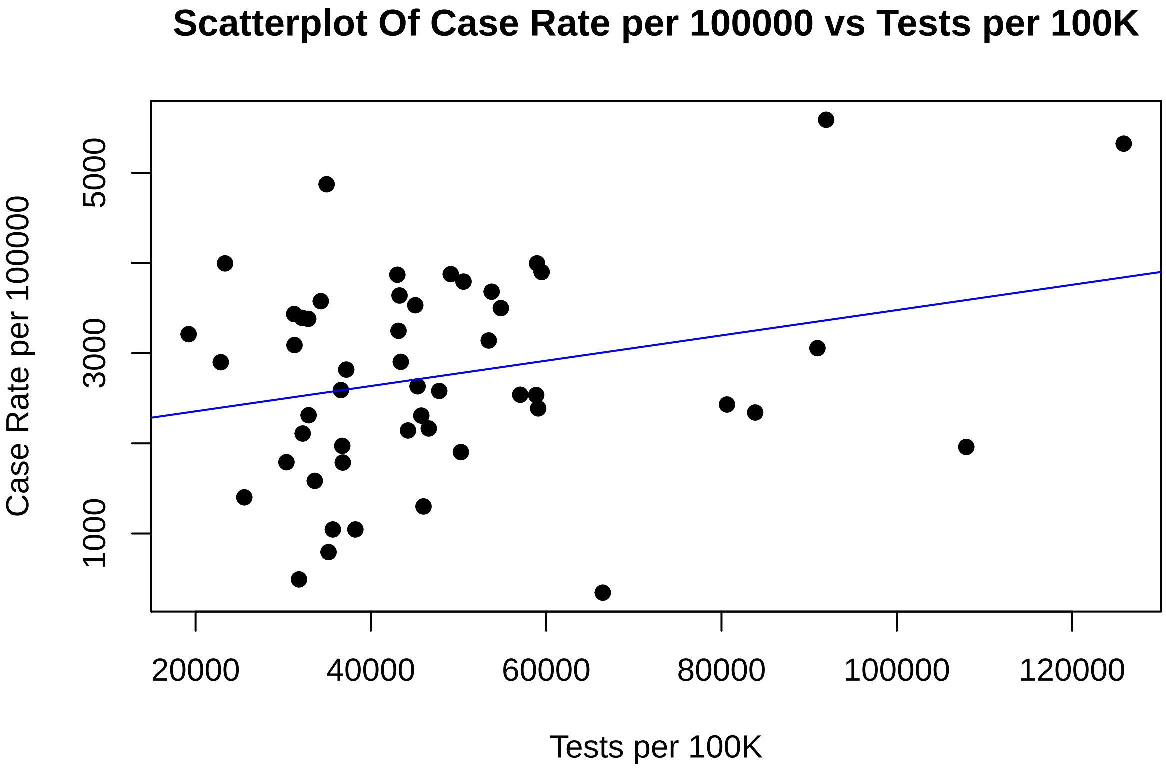

Figure 18: Scatterplot showing Case Rate per 100,000 as predicted by Significant Vaiables in the Detailed Model

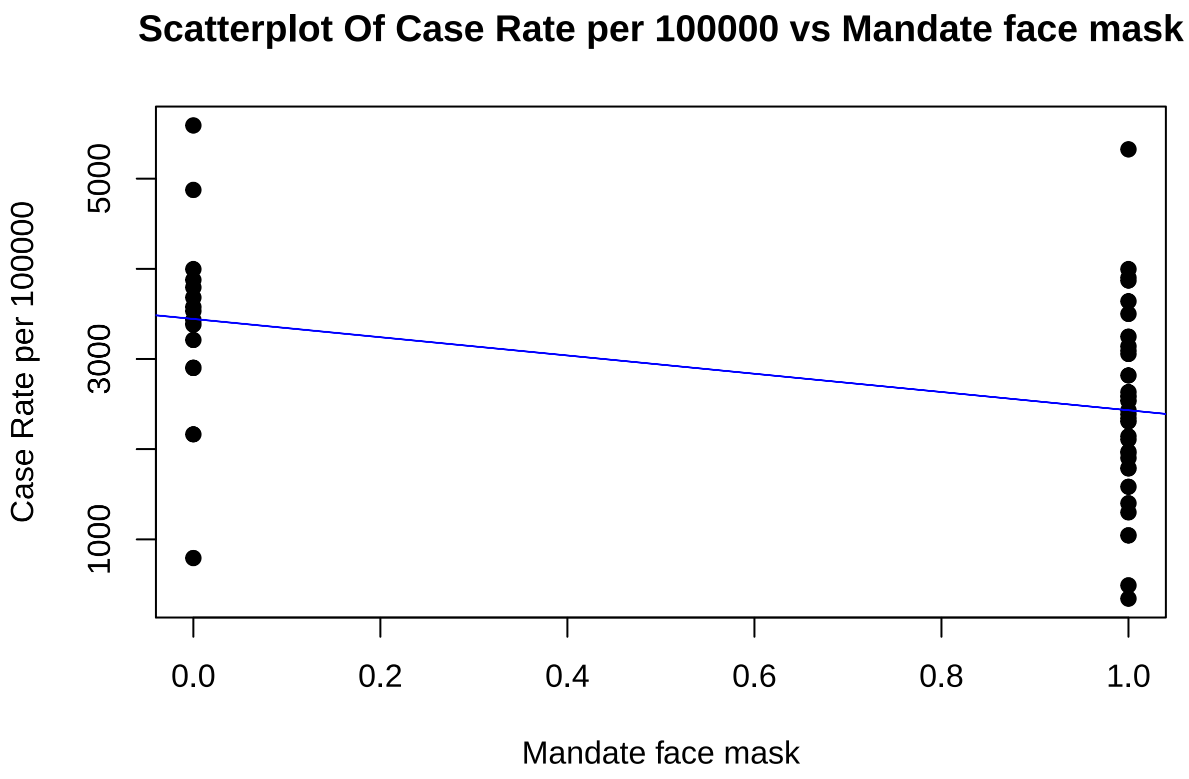

Figure 19: Scatterplot showing Case Rate per 100,000 as predicted by Significant Vaiables in the Detailed Model

All the scatterplots for the variables that are significant in the Detailed Model show that there is either a positive or negative relationship with the Case Rate per 100K. It can be seen that for younger groups(Figure 15), case rate increases as the population percentage increases. The opposite is true for older age groups (Figure 16, 17). Figure 18 shows that as testing increases, the case rate increases as well. This is required to control for the testing rate in each state. Finally, Figure 19 shows that mask usage decreases the number of cases on average. These plots help visualize the practical significance discussed above.

Model Comparison

```r
se.fit03 = coeftest(fit03, vcov = vcovHC)[ , "Std. Error"]

stargazer(fit01, fit02, fit03, type = "text", omit.stat = "f",
          se = list(se.fit03),
          star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 1: The Comparision Between All 3 Models",
          font.size = "small",
          single.row = FALSE,
          column.sep.width = "1pt")
```

```
##
## Table 1: The Comparision Between All 3 Models
## ======================================================================================
##                                                     Dependent variable:
```

```
##                                  -----------------------------------------------------
##                                                 ‘Case Rate per 100000‘
##                                         (1)                (2)               (3)
## --------------------------------------------------------------------------------------
## ‘Children 0-18‘                     33,582.830        42,813.390       64,615.970**
##                                    (25,595.020)      (23,924.760)      (19,930.970)
##
## ‘Adults 19-25‘                     114,862.800***     108,094.900**     86,145.500**
##                                    (34,850.460)      (30,784.750)      (25,735.100)
##
## ‘Adults 26-34‘                      10,520.400        20,664.510        25,142.340
##                                    (23,870.270)      (24,729.950)      (20,043.080)
##
## ‘Adults 35-54‘                      27,674.480        19,080.770        39,368.320
##                                    (33,983.360)      (27,335.790)      (22,638.490)
##
## ‘Adults 55-64‘                      16,919.620        35,051.020        35,786.830
##                                    (28,068.450)      (31,282.030)      (25,592.970)
##
## ‘65+‘                               32,486.730        33,784.790       49,636.170**
##                                    (22,453.600)      (22,047.990)      (18,307.360)
##
## log(‘Pop. density / sq mile‘)                           129.689           231.695*
##                                                       (136.306)         (113.294)
##
## ‘Nonelderly Adults w/ Condition‘                        0.0001            0.0001
##                                                       (0.0001)          (0.0001)
##
## ‘Closed Business Days‘                                 -8.086             -7.448
##                                                       (10.005)           (8.639)
##
## ‘Tests per 100K‘                                                        0.024***
##                                                                          (0.006)
##
## ‘Mandate face mask‘                                                     -0.018**
##                                                                          (0.006)
##
## Constant                            -31,121.350       -34,818.360      -47,604.250**
##                                    (22,239.670)      (21,218.580)      (17,365.860)
##
## --------------------------------------------------------------------------------------
## Observations                             51                51                51
## R2                                      0.451             0.487             0.681
## Adjusted R2                             0.376             0.374             0.591
## Residual Std. Error              897.509 (df = 44) 899.000 (df = 41) 726.815 (df = 39)
## ======================================================================================
## Note:                                                    *p<0.05; **p<0.01; ***p<0.001
```

The comparison will be conducted referencing the table above. In model 1, which only contains the age distribution, it can be seen that only one age distribution has statistical significance. While the model does appear to have descriptive power, with an R^2 of 0.451, each coefficient individually does not appear to be significant. This can be attributed to other factors affecting the COVID-19 case count. Age distribution is not the only contributing factor, since social distancing, mask usage, and other factors can also contribute. It can be seen that the model improves to $R^2 = 0.681$ as more of these factors are added, along with the significance of the age distributions. As the new features take some of the descriptive power, the coefficients for age distribution become more significant. It can be seen in model 3 that the percentage of the population older than 65, under 18 and between ages 19-25 are the most significant variables in the model. Early research has shown that the elderly are more at risk for contracting COVID-19. Furthermore, young adults are more likely to continue interacting socially and to continue working. The model would support those hypotheses, as these age distributions are the biggest contributing factors to the case rate in a state. Other than the age distribution, testing rate is also shown to be significant. Since cases, the data the model is describing, cannot be determined without tests, it stands to reason that the testing rate would also contribute to the case rate. For these reasons, the third model is the one that should be used by states for descriptive purposes. While there appears to be enough descriptive power in the other two models, the coefficients do not have enough statistical significance to be able to describe how different populations could be affected by COVID-19. It is not until the other control variables are added that the coefficients begin to become statistically significant. This makes model 3 the most practical and useful for states in the USA to utilize for descriptive purposes.

4. Model Limitations

These models are meant to imitate real life conditions. The real limitation of the model is that not all the variables can be included, because there are an infinite number of them. Thus, feature selection is done to pick the most important variables, which is a 'simplification' of the real problem. For example, in the first model, only age group distribution is considered as the only variable related to the problem. Then, in the second and third model, population variables and policy variables are added in. As expected, the more variables added, the more accurate the model became, because it is closer to the real problem. Based on the comparison done above, Model 3 provided the most information. The CLM assumptions are evaluated against this model.

Assumption 1 IID Sampling: The data being used is generated by an IID process, because it comes from a literal IID sample data of each state. Each state's situation is independent of others, and the data in each state is a count up of independent cases for different variables. When the data was plotted out, no clusters are found. Thus, the data is IID.

Assumption 2 Linear Conditional Expectation: The regression model is linear in the coefficients and the error term and is generated by the lm() function in R. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance. The assumption is then confirmed by plotting out the fitted regression with the real data points in Figure 20. The plotted line is a straight line with a clear slope, no matter how the data is distributed.

```r
plot(fit03, which = 1)
```

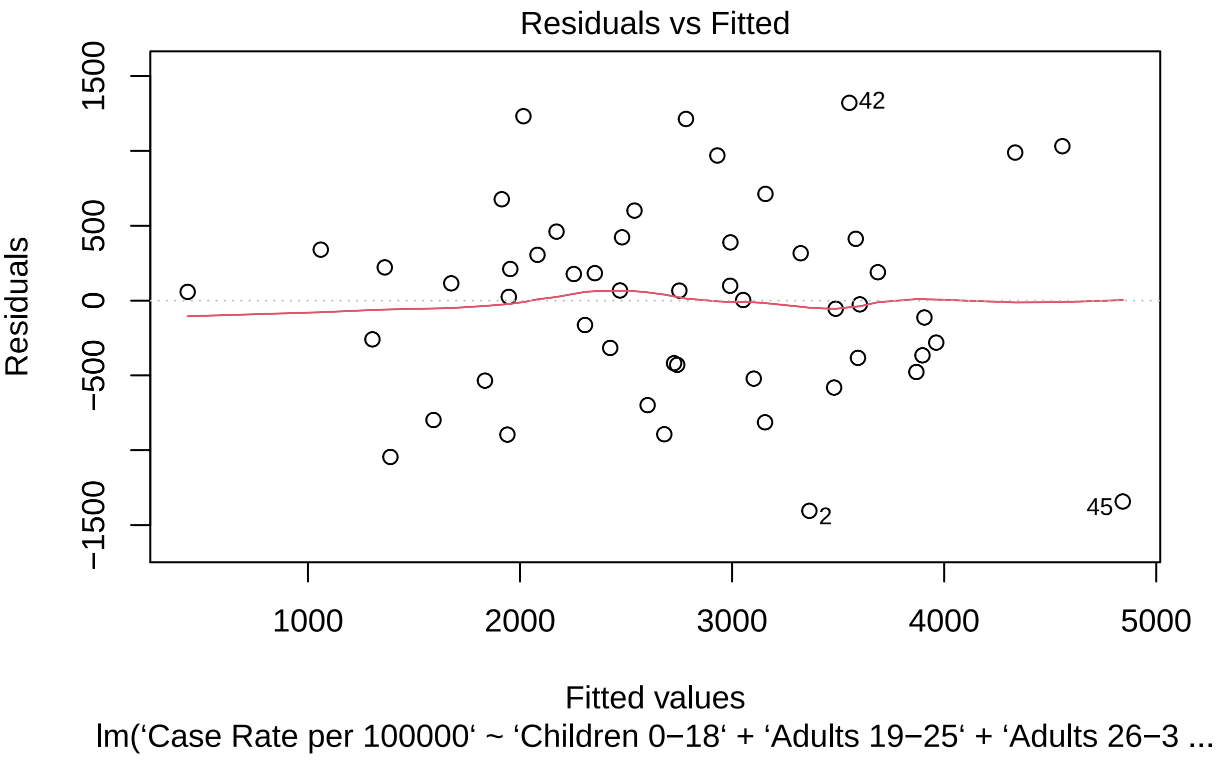

Figure 20: Check for Linear Conditional Expectation

Assumption 3 No Perfect Collinearity: For all the features that are used in the model, there was no collinearity, because no features were dropped. Also there were not any large standard errors between any of the features, as seen in the VIF matrix below.

```
# Checking for collinearity
car :: vif(fit03)
```

```
##                 ‘Children 0-18‘                   ‘Adults 19-25‘
##                       18.923456                         3.584189
##                  ‘Adults 26-34‘                   ‘Adults 35-54‘
##                        8.591796                         4.786158
##                  ‘Adults 55-64‘                           ‘65+‘
##                        9.486599                        13.622199
##     log(‘Pop. density / sq mile‘) ‘Nonelderly Adults w/ Condition‘
##                        2.707626                         1.324766
##          ‘Closed Business Days‘                 ‘Tests per 100K‘
##                        1.621821                         1.730875
##             ‘Mandate face mask‘
##                        1.421815
```

Assumption 4 Homoskedastic Conditional Variance: The error term has a constant variance (no heteroscedasticity) The first check is done with an “Eye test”, in Figure 21. On this graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction.The spread of the residuals increases as the fitted value increases. This is not the case for this model. There are some outliers, namely Alaska (2), South Dakota (42), and Utah (45). This is a limitation of the model, as these states appear to follow a different variance pattern. However, overall, the model appers to satisfy homoskedasticity.

```
plot(fit03, which = 3)
```

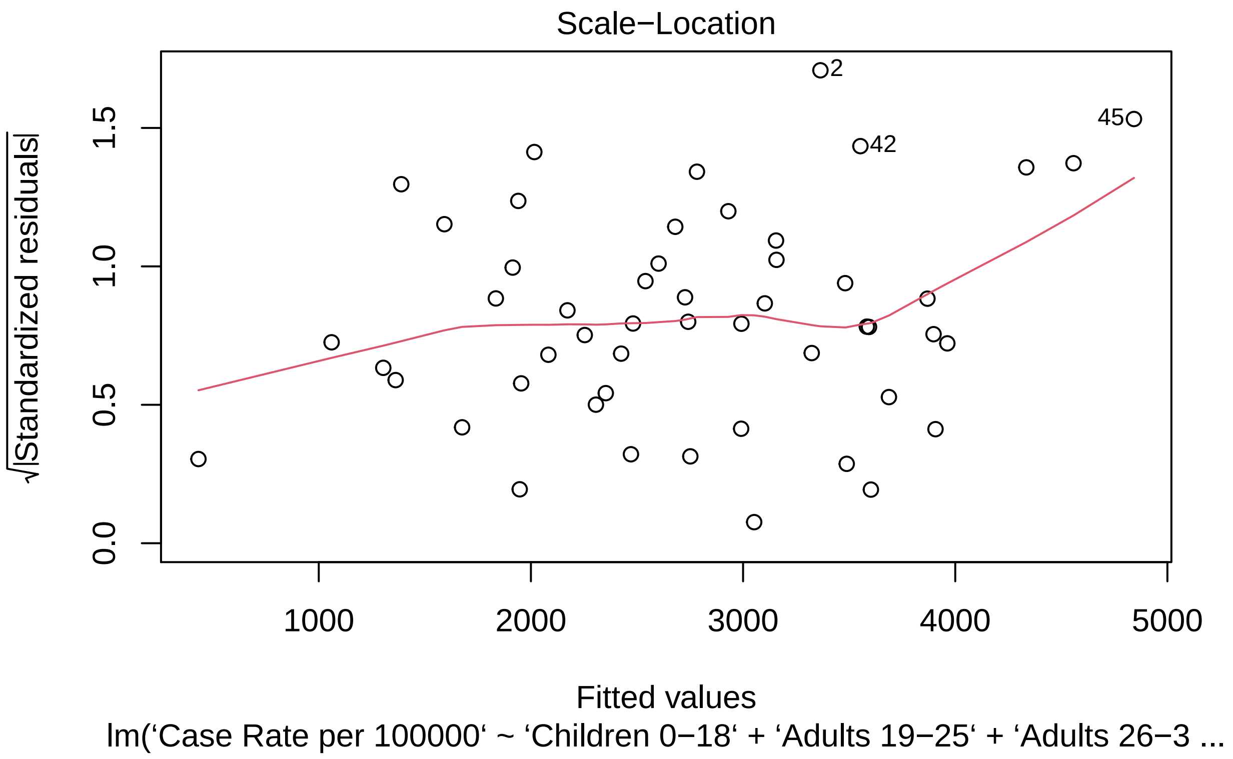

Figure 21: Check for Homoskedasticity

A Breusch-pagan test was also performed. Since the P-value is below 0.05, the null hypothesis can be rejected. This indicates that the model does not violate the assumption of homoskedasticity:

```
lmtest::bptest(fit03)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit03
## BP = 20.311, df = 11, p-value = 0.04126
```

Assumption 5 Normally Distributed Errors:

```
plot(fit03, which=2)
```

The error terms are normally distributed as the residuals follow a normal distribution by plotting them out in Figure 22.

Also, The error term has a population mean of zero. Looking at the Residuals vs Fitted graph above (Figure 20) it can be seen that there is not a straight line at 0. The line has a downward slope on the left and right hand side of the graph. This a violation of the zero conditional mean assumption.

It can be confirmed that one observation of the error term will not predict the next observation by assessing the graph of residuals in the order that the data were collected. There is a randomness in the plot (Figure 22).

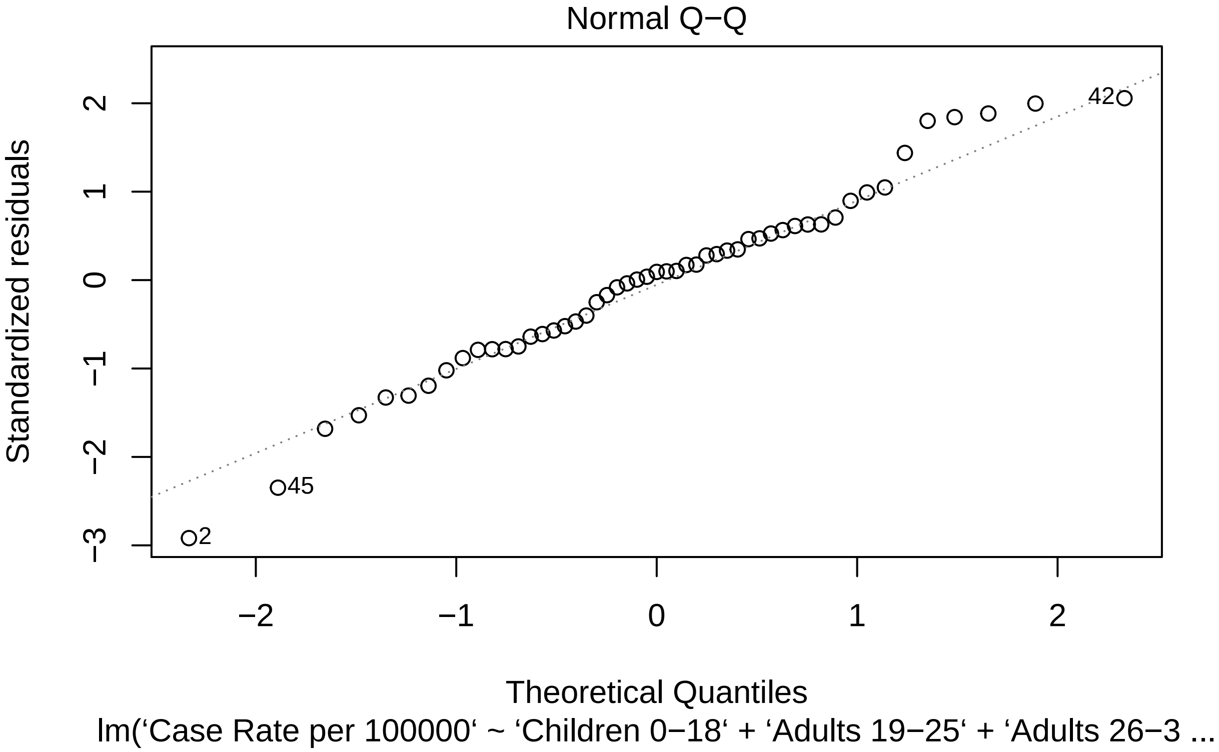

Figure 22: Check for Normally Distributed Errors

Omitted Variables:

For “shelter in place”, since the hypothesis is related to the age group, the shelter in place variable is consistent for any age group. As a result, when adding this variable into the model, the t-value was not statistically significant. It was removed in the final model, and this omitted variable had no bias on in the model at all.

Another policy that was potentially included was the variables relating to the business total close time in each state. The developed features that were used were: total business closed time using ‘Began to reopen businesses statewide’ - ‘Closed other non-essential businesses’. There was also different closed time for restaurants, bars, and other essential business. One of the assumptions was that people with younger age groups will correlate with the close time of bars and restaurants. But, after adding in those variables, there was not any improvement, and also to avoid violating variable independency, they were removed.

5. Conclusion

This model can be used as a baseline for the expected case rate of COVID-19, given the age distribution of the population. With different states having different age distributions, the description is tailored to each state’s specific needs. Furthermore, using the more detailed model, states can have an idea of how state policies helped their population. This model can serve as an important tool for pandemic mitigation and recovery, and can highlight which state populations are at risk. By knowing how the population was affected, governements can make plans and policies to help their populations recover. While the granularity is limited, the model can still provide insight into the spread of COVID-19 for an American state, and allow a state to take the appropriate actions.