STATS – 35500

Analysis of COVID – 19 across different States in the USA

Shivam Bairoliya

COVID – 19 is a contagious disease caused by the virus SARS-CoV-2. The virus is currently responsible for about 1.5 million deaths globally. It has affected 191 countries across the world. In the United States, the virus is responsible for nearly 270,000 deaths with the number expected to grow at a higher rate in the coming months. Through this project, I try to analyze some statistics across the various states in the country and try to answer a few questions.

Dataset source:

https://www.kaggle.com/nightranger77/covid19-state-data

Description of the Dataset:

The dataset contains the cumulative number of Coronavirus Tests, Infection, and Deaths from March through November 1st.

The numbers are divided state-wise including the District of Columbia

The variables in context are:-

- State
- Tested
- Infected
- Deaths
- Population Density
- Smoking Rate
- Pollution
- Temperature
- Case Fatality Ratio (Calculated Manually)
- Average Positivity Rate (Calculated Manually)
- ICU Beds per 10000 of the population (Calculated Manually)
- Physicians per 10000 of the population (Calculated Manually)
- Governors (Added Manually) – categorical Republican or Democrat

Description of Variables:

State:

It is the names of all the States in the United States of America including the District of Columbia listed in alphabetical order

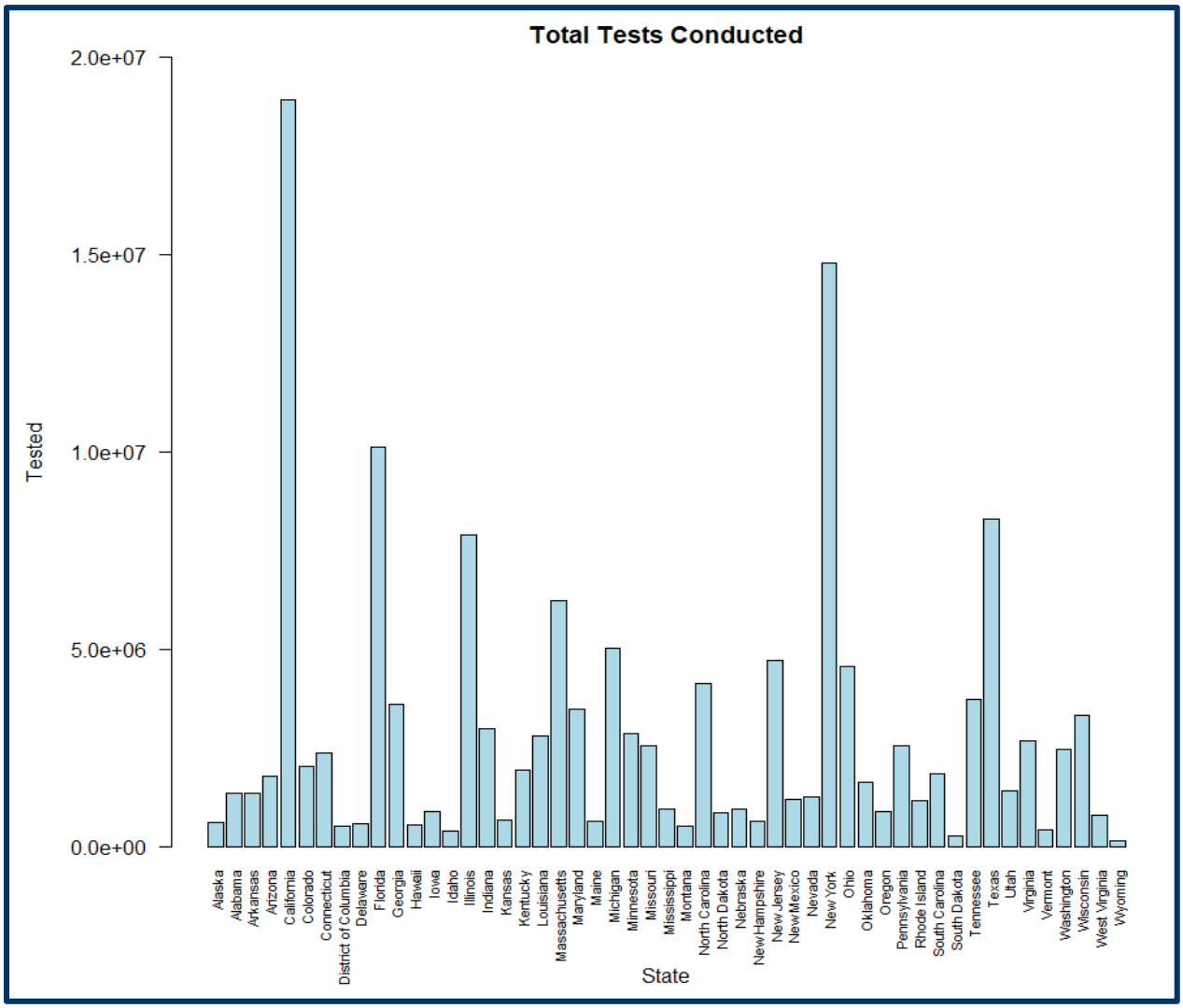
Tested:

The total number of people tested for COVID - 19 in the State until 1st November

Mean: 2904946

Standard Deviation: 3590449

Here is a bar plot



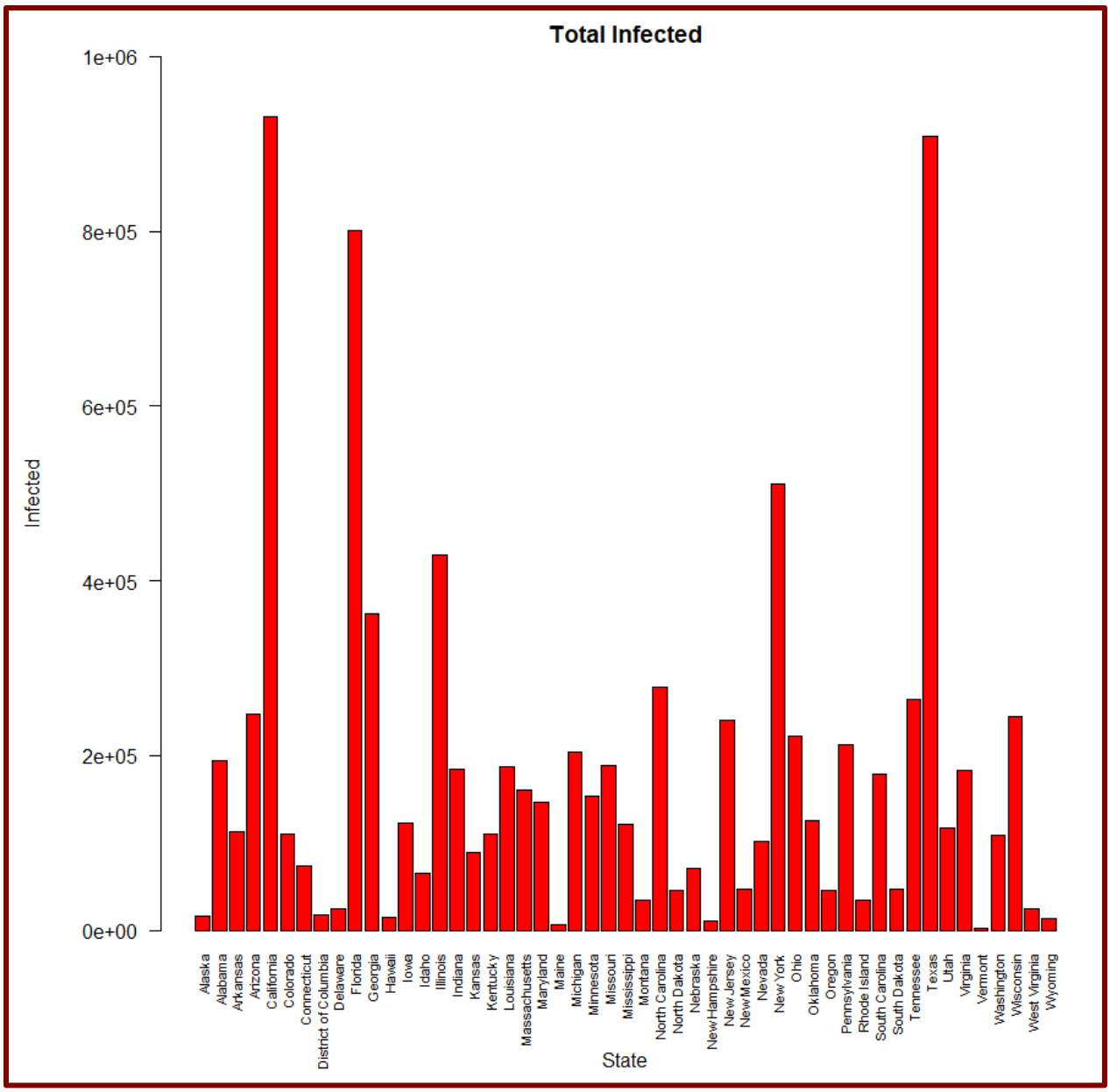
Infected:

The total number of people who tested positive for COVID - 19 in the State until 1st November

Mean: 179626.7

Standard Deviation: 208077.9

Here is a bar plot



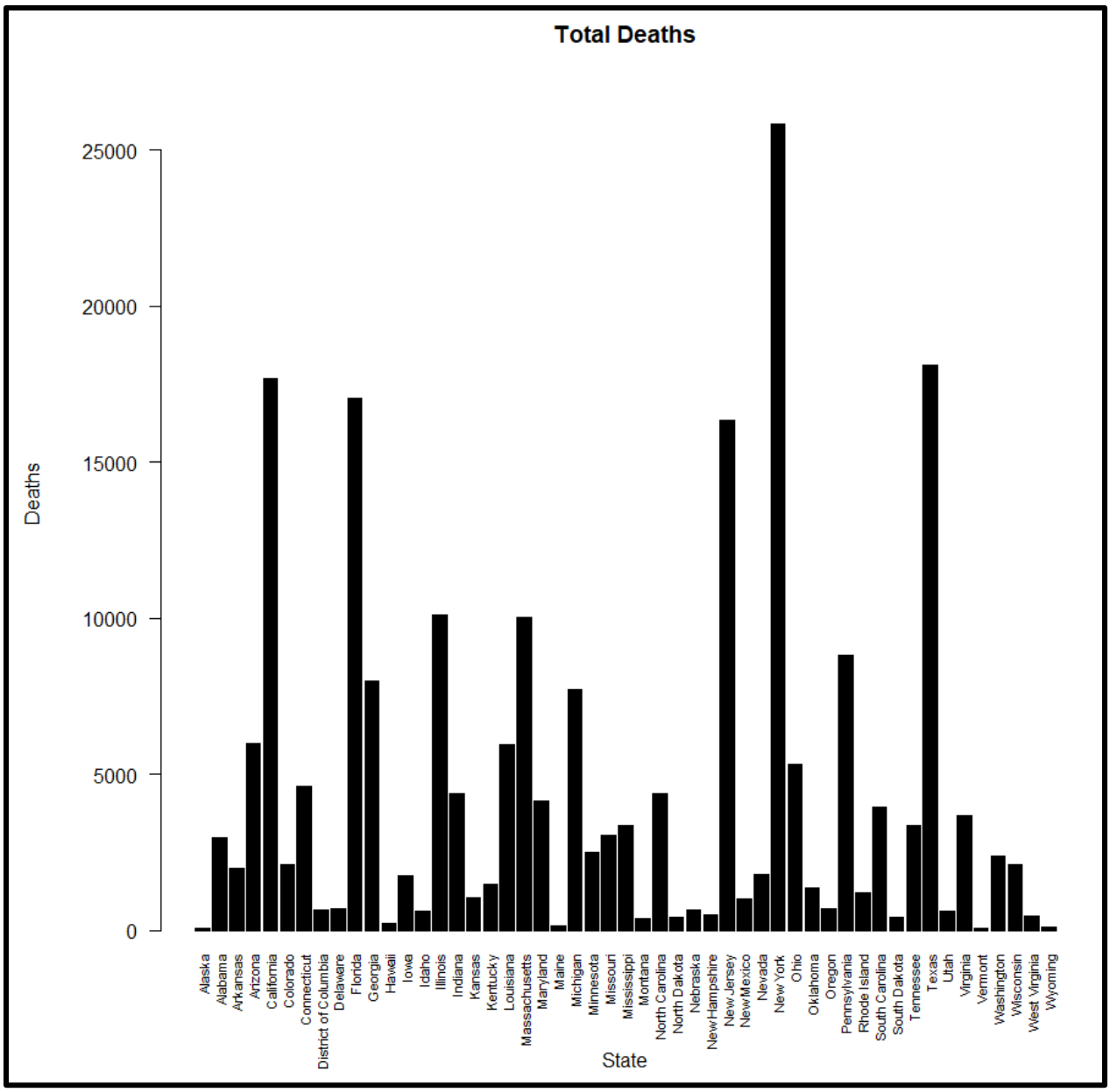
Deaths:

The total number of people who died from COVID - 19 in the State until 1st November

Mean: 4357.745

Standard Deviation: 5637.548

Here is a bar plot



Population Density:

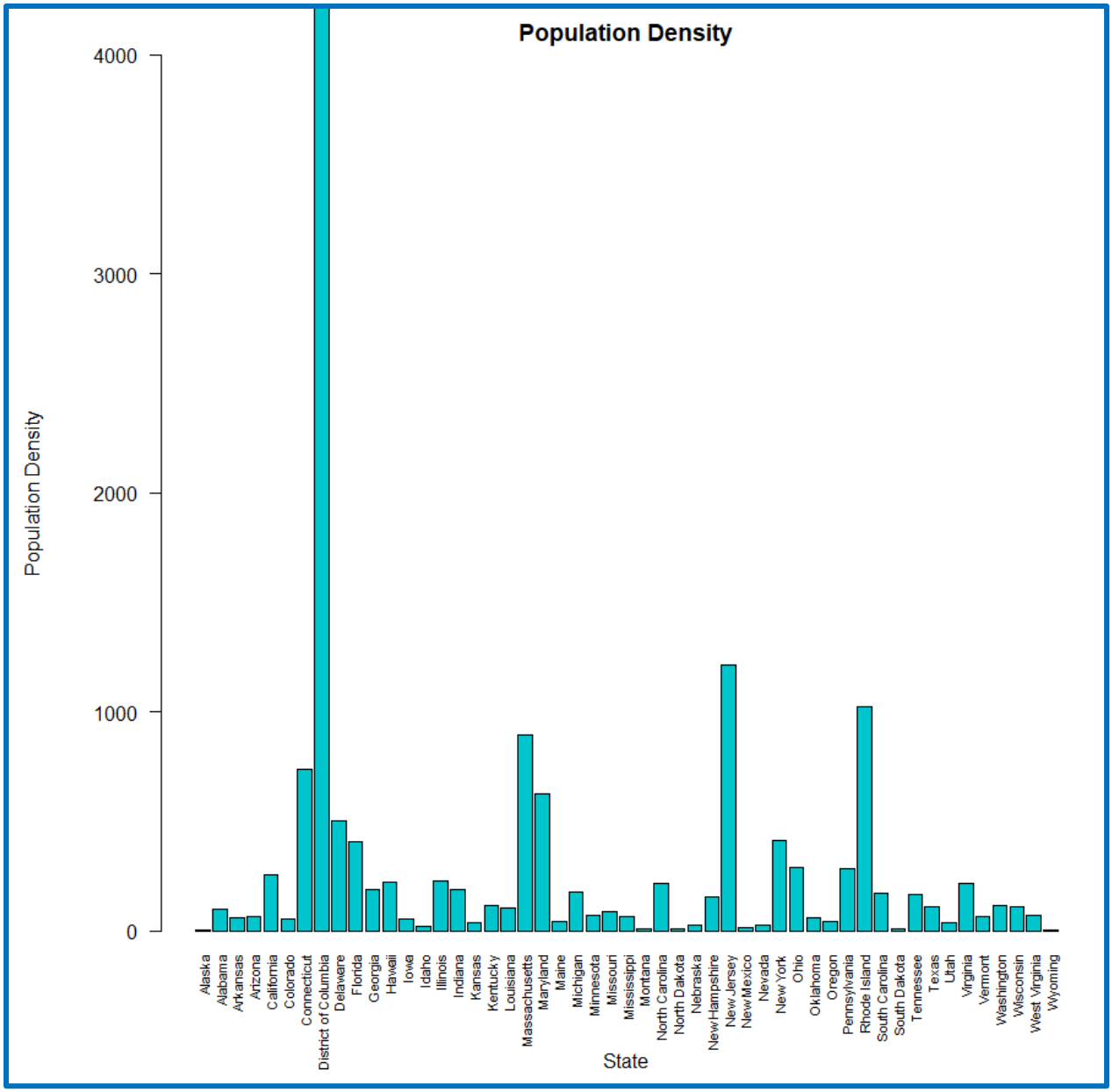
The Population Density of every State in people per mile^2

The population density of the District of Columbia is about 11814.54 people/mile^2

Mean: 431.5605

Standard Deviation: 1647.226

Here is a bar plot



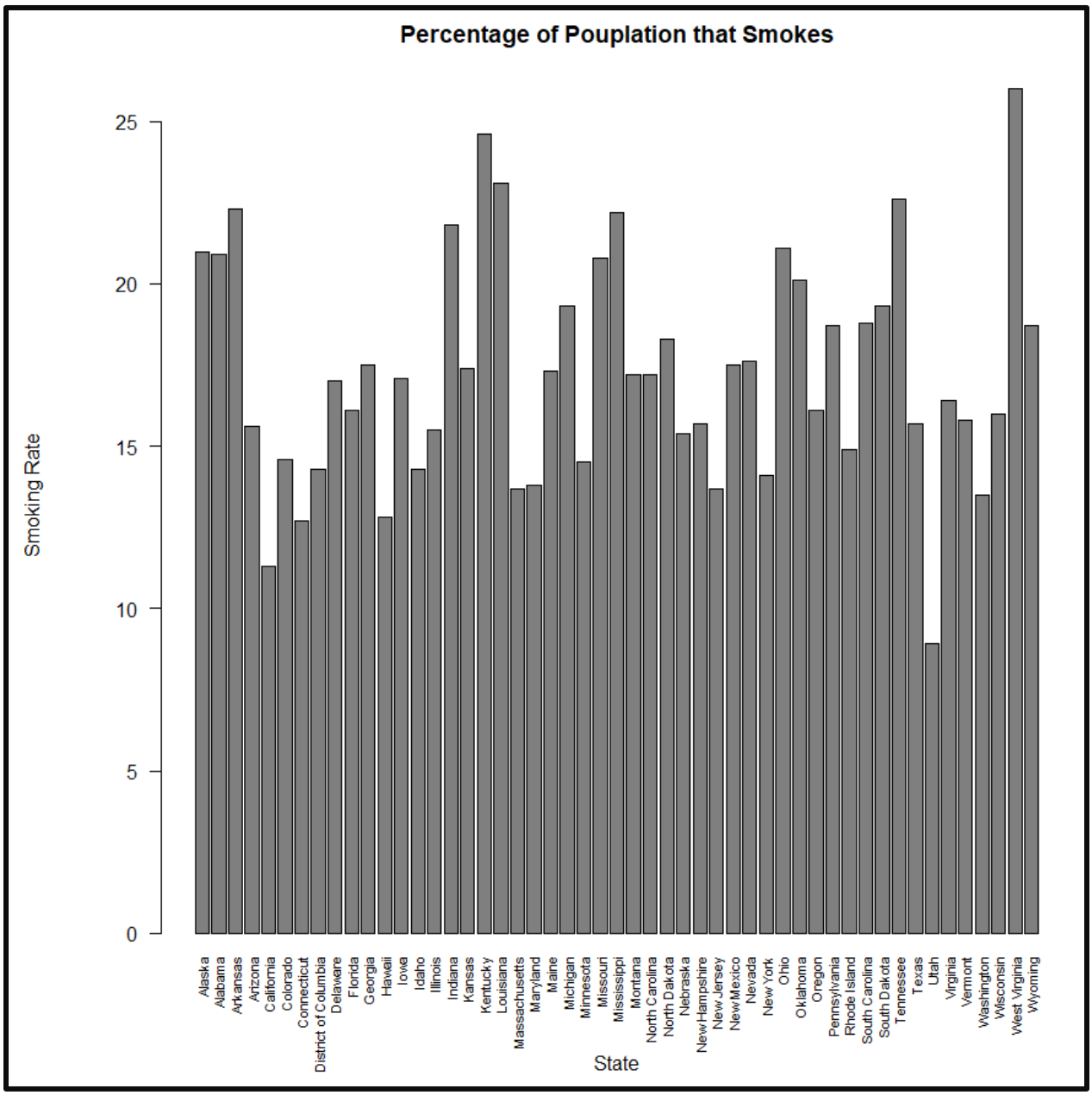
Smoking Rate:

The percentage of people in the population that smoke

Mean: 17.27059

Standard Deviation: 3.489429

Here is a bar plot



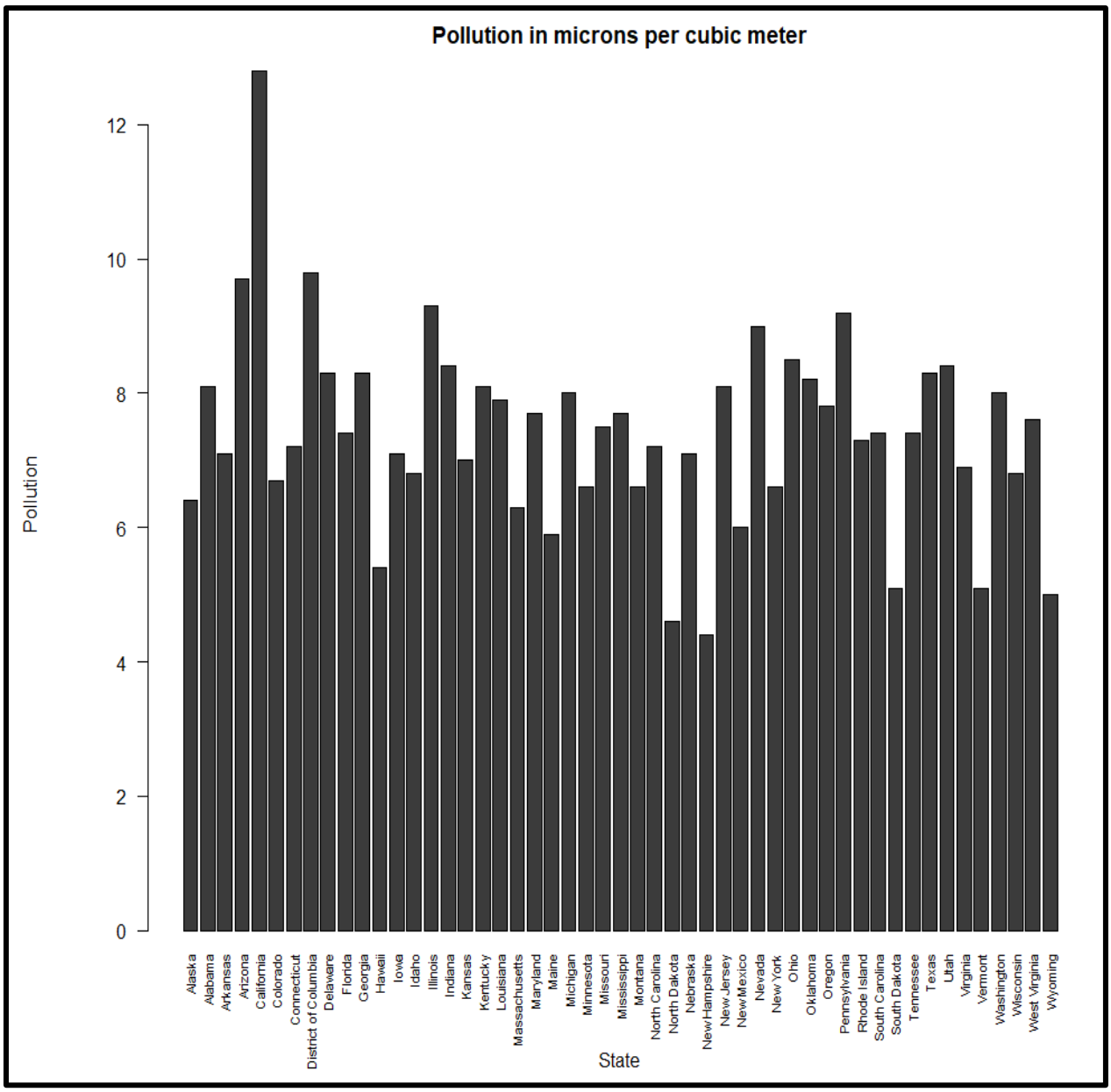
Pollution:

The average exposure of the general public to particulate matter of 2.5 microns or less (PM2.5). It is measured in micrograms per cubic meter and is a 3-year estimate from 2017 through 2019.

Mean: 7.413725

Standard Deviation: 1.457535

Here is a bar plot



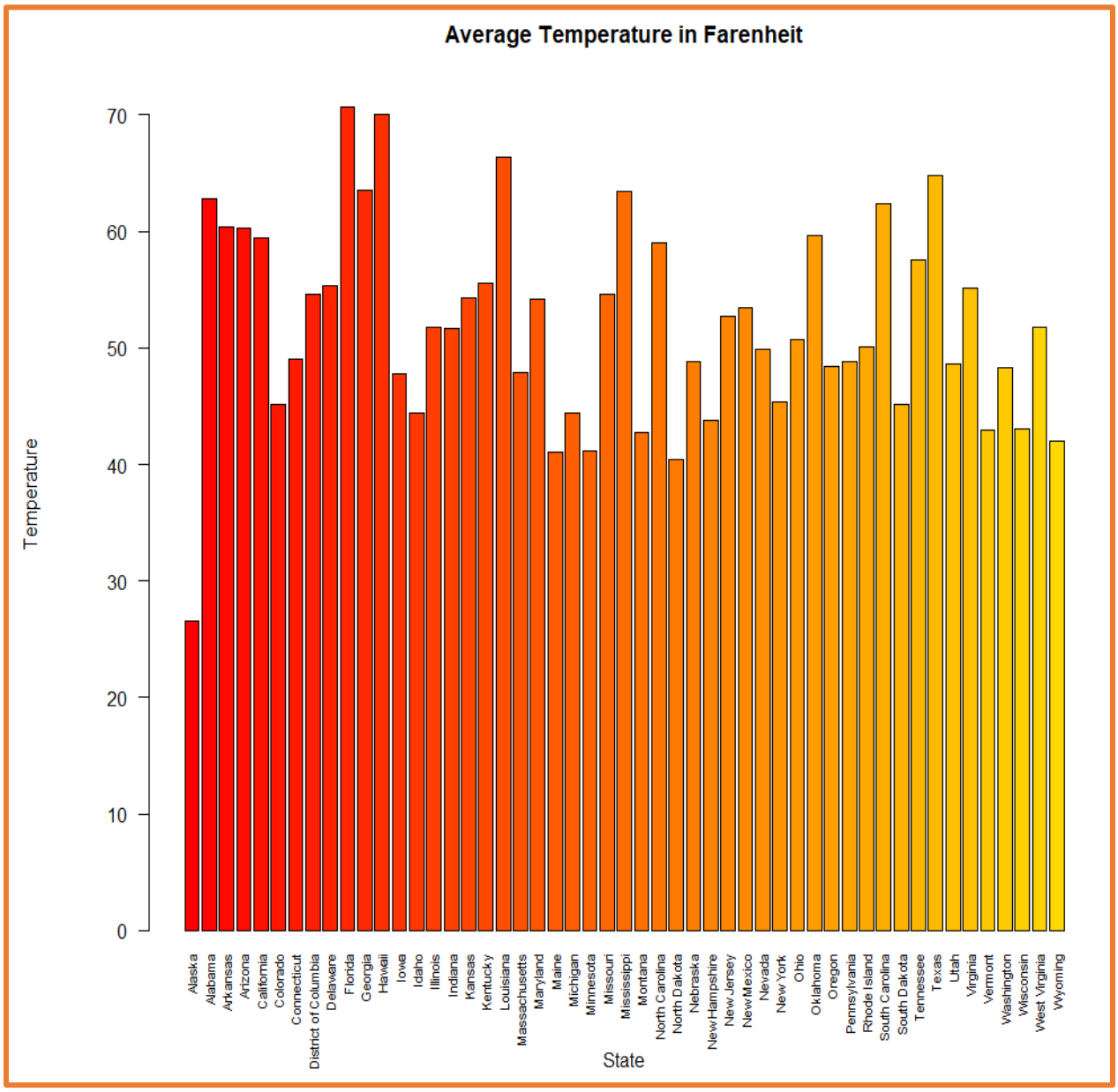
Temperature:

The average temperature of the States is measured through a calendar year. The temperature for the District of Columbia is measured by dividing the average of Maryland and Virginia. The temperatures are measured in Fahrenheit

Mean: 51.99902

Standard Deviation: 8.627992

Here is a bar plot



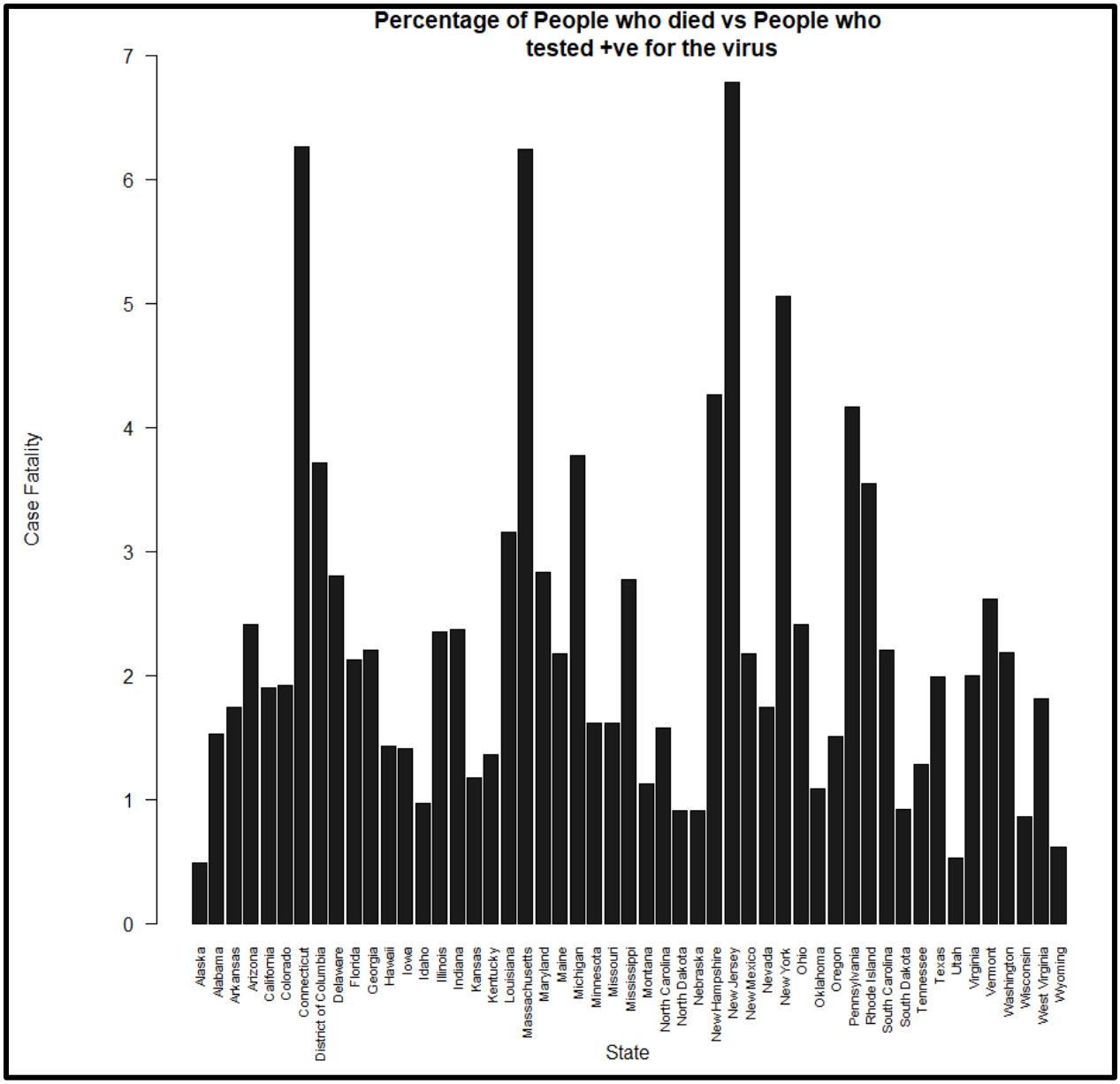
Case Fatality Ratio:

The ratio of people who dies from COVID – 19 to the people who tested positive for the virus up to November 1st expressed as a percentage

Mean: 2.285843

Standard Deviation: 1.444225

Here is a bar plot



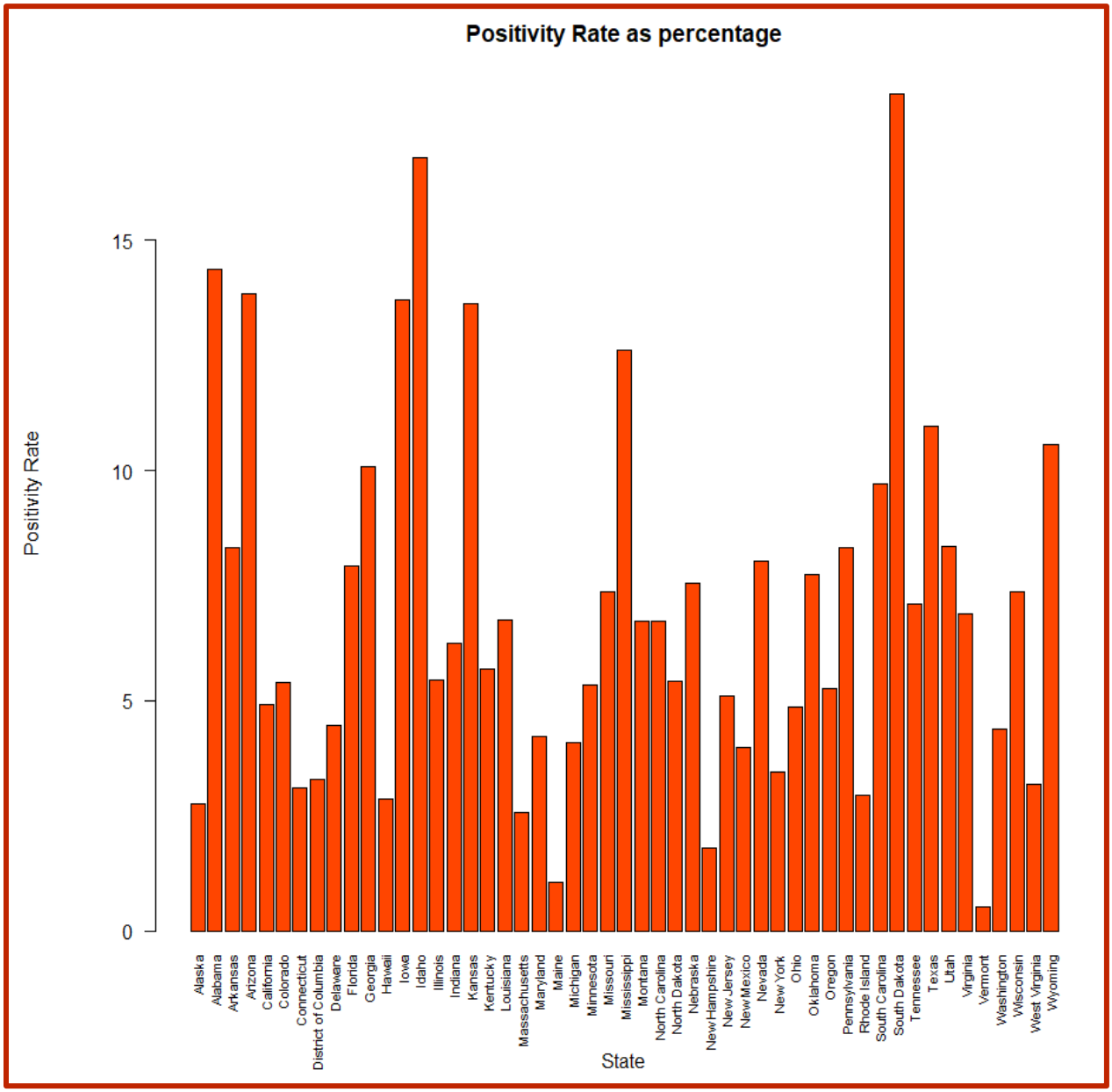
Average Positivity Rate:

The ratio of the cumulative number of people who tested positive vs the people who got tested expressed as a percentage

Mean: 6.903176

Standard Deviation: 3.981981

Here is a bar plot



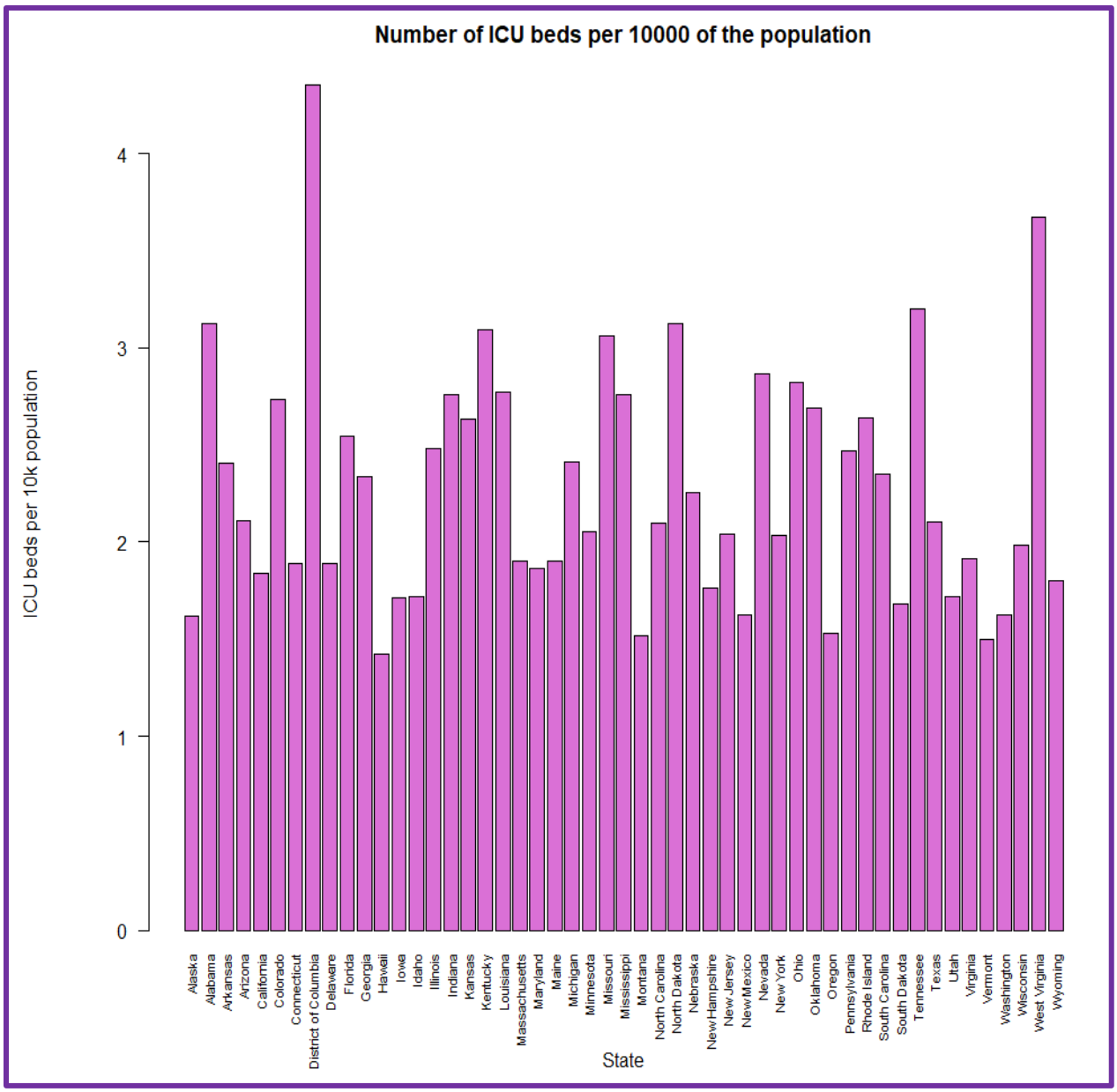
ICU Beds per 10000 of the population:

The number of ICU beds in the state per 10000 people in the population of the state

Mean: 2.283118

Standard Deviation: 0.6162801

Here is a bar plot



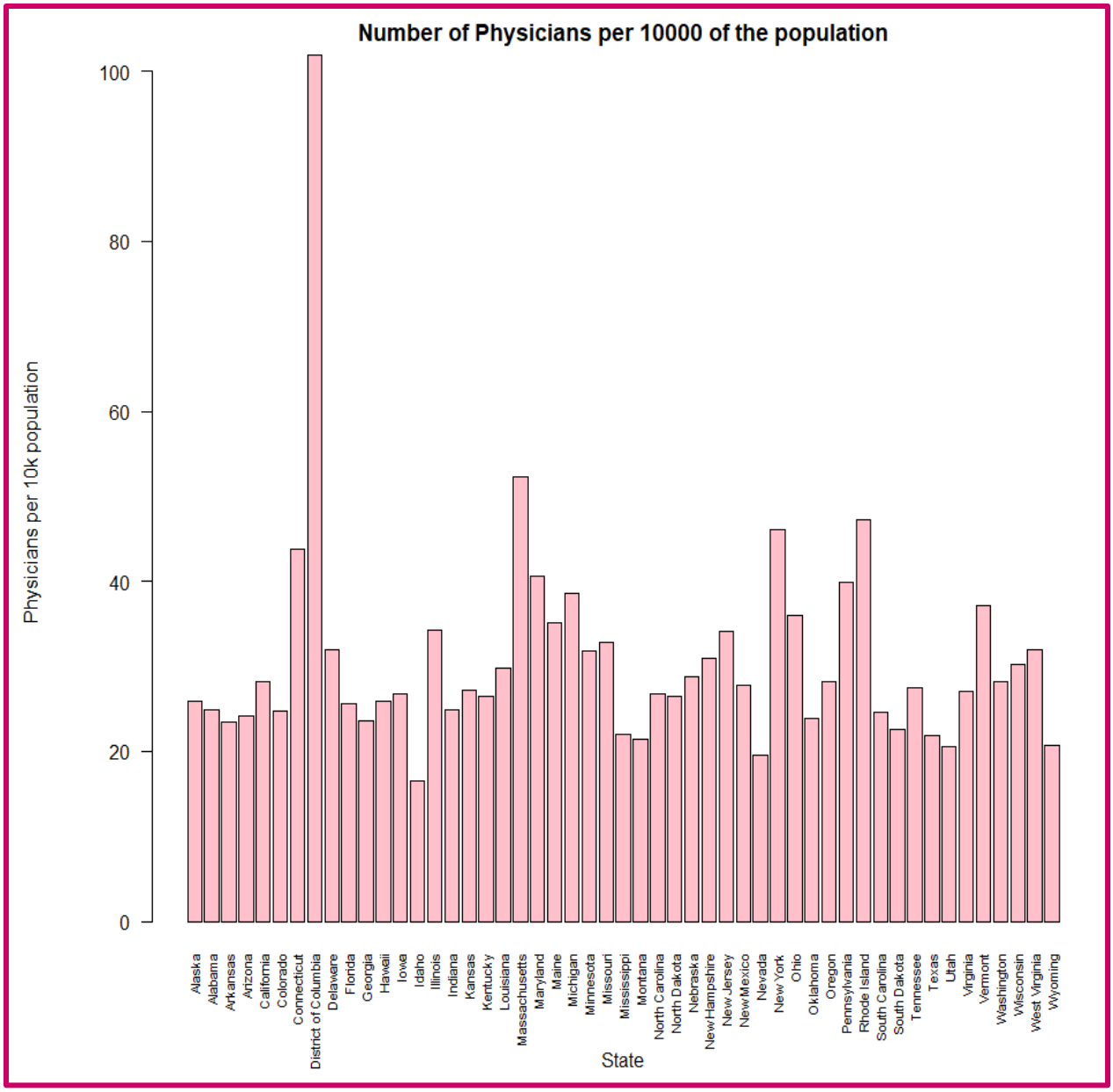
Physicians per 10000 of the population:

The number of ICU beds in the state per 10000 people in the population of the state

Mean: 30.85761

Standard Deviation: 12.61255

Here is a bar plot



Governors:

This is a categorical variable that has two letters either “D” for Democrat or “R” for Republican. It represents the political affiliation of the State’s Governor.

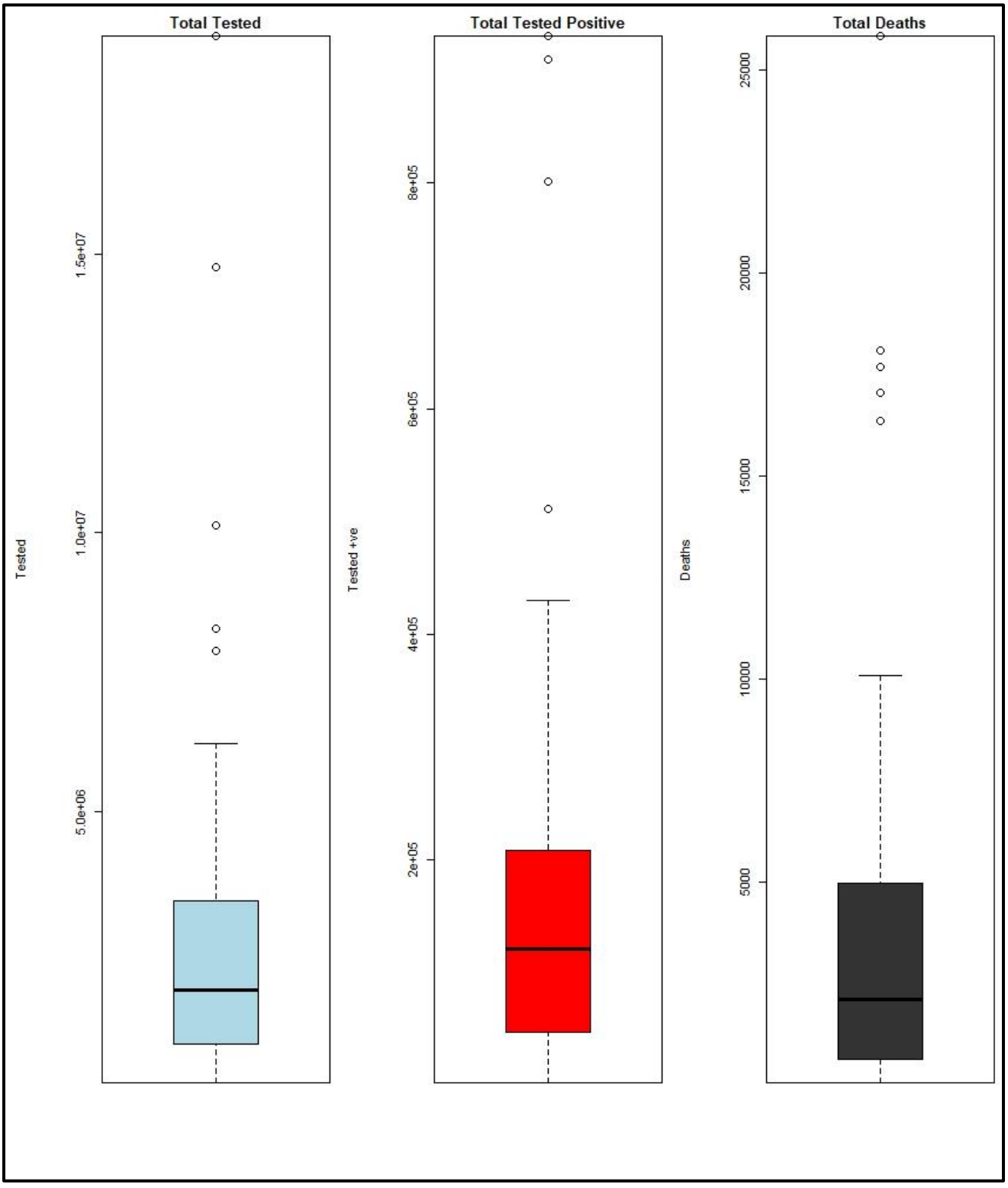
For the District of Columbia, the Mayor is considered who is a Democrat.

Count of Democrat ruled States: 25

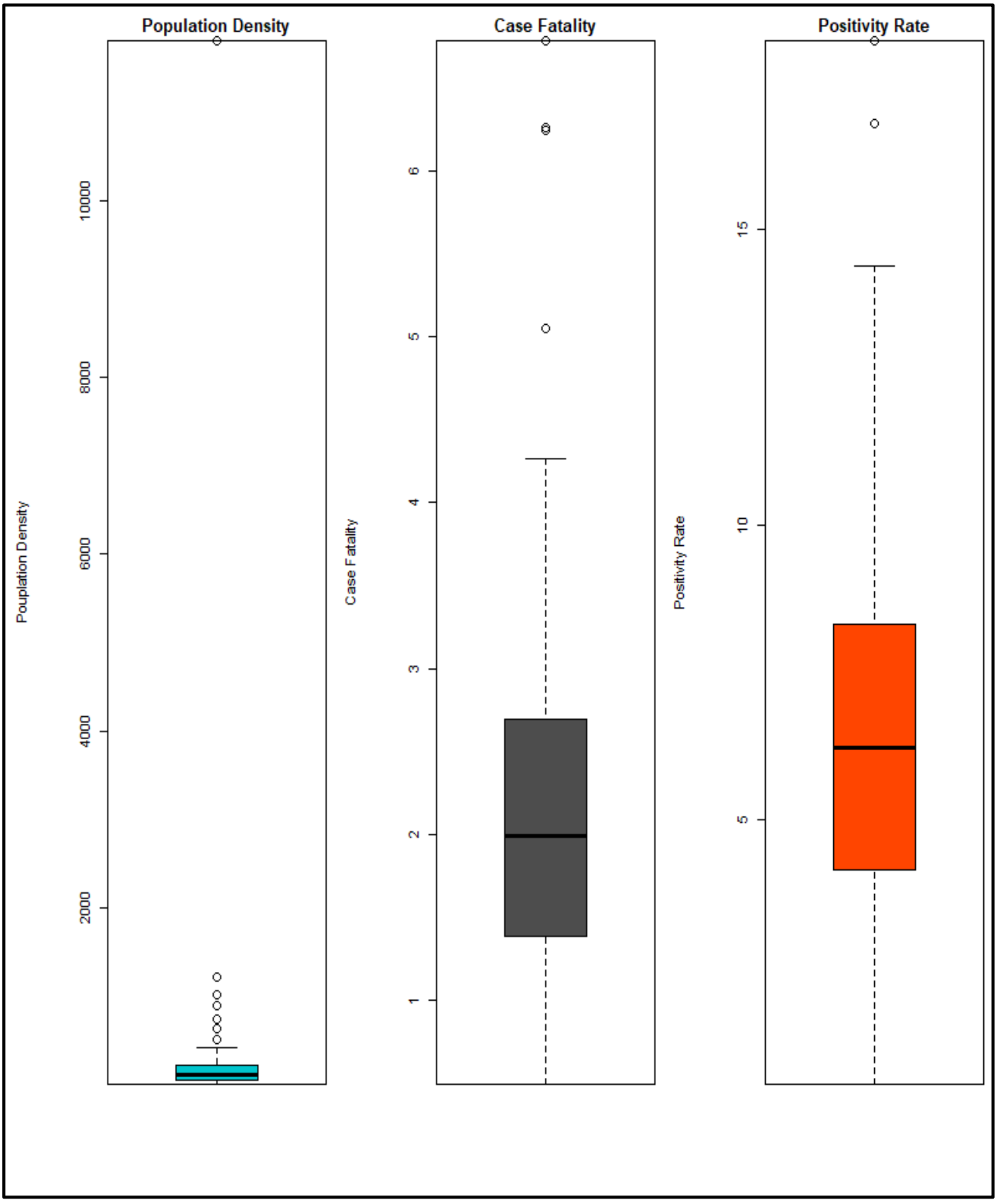
Count of Republican ruled States: 26

source: https://en.wikipedia.org/wiki/List_of_United_States_state_legislatures

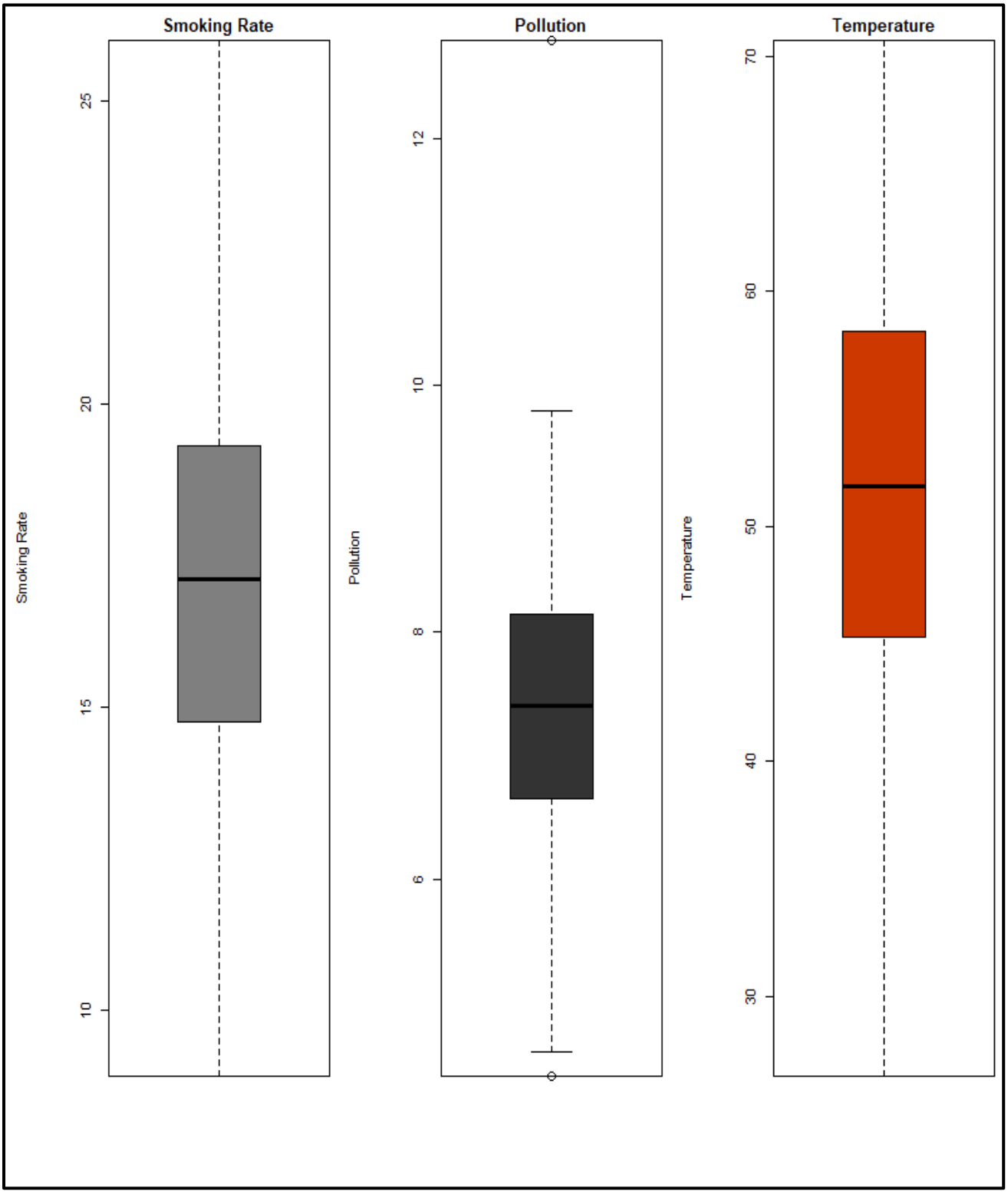
Here is a boxplot of Tested, Infected, and Deaths



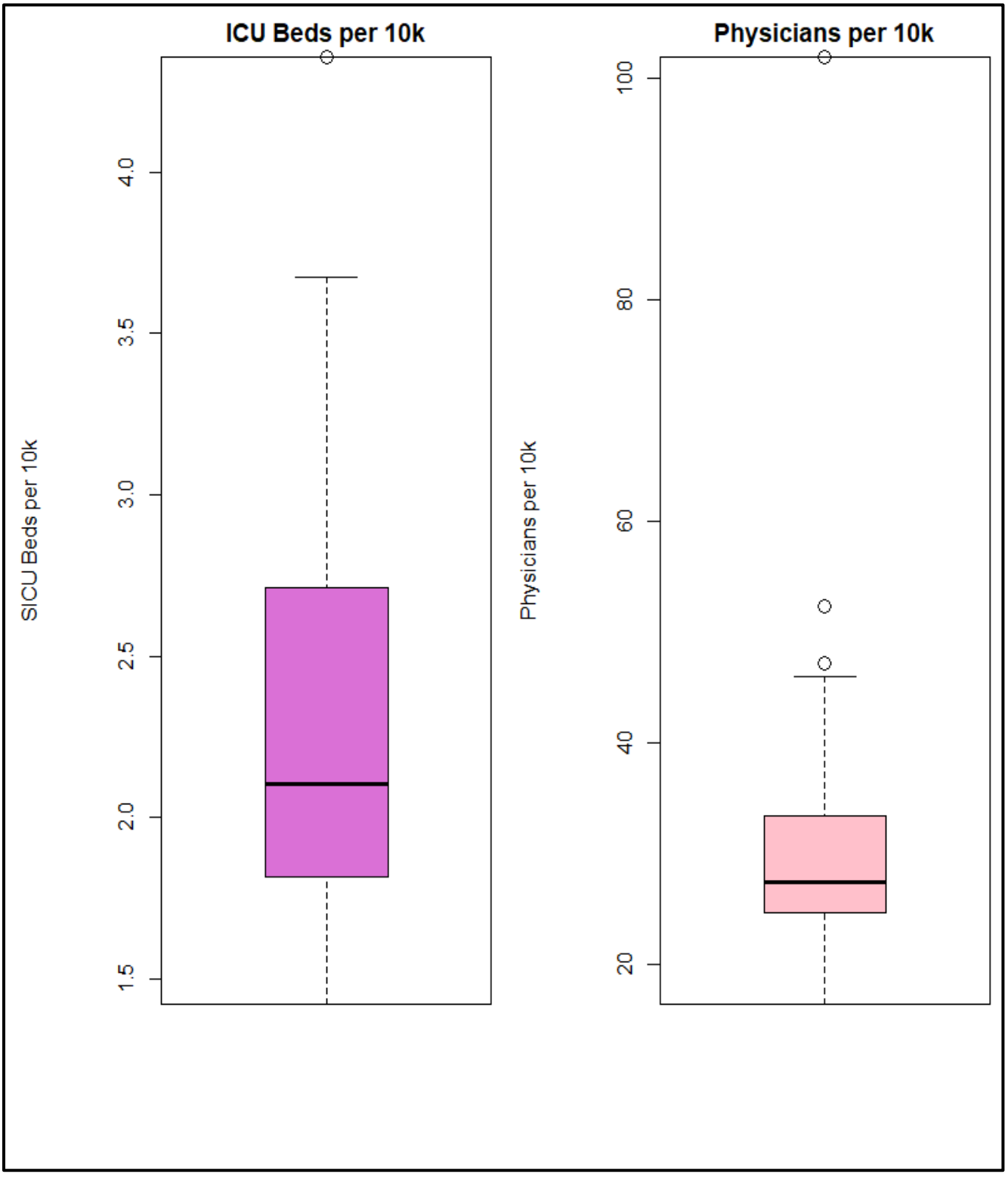
Here is a boxplot of Tested, Infected, and Deaths



Here is a boxplot of Smoking Rate, Pollution, Temperature



Here is a boxplot of ICU Beds per 10k and Physicians per 10k



Research Questions:

Question 1:

Is there a significant difference between the mean of Positivity Rate of the Republican Controlled States and the Democratic Controlled States?

Why this question?

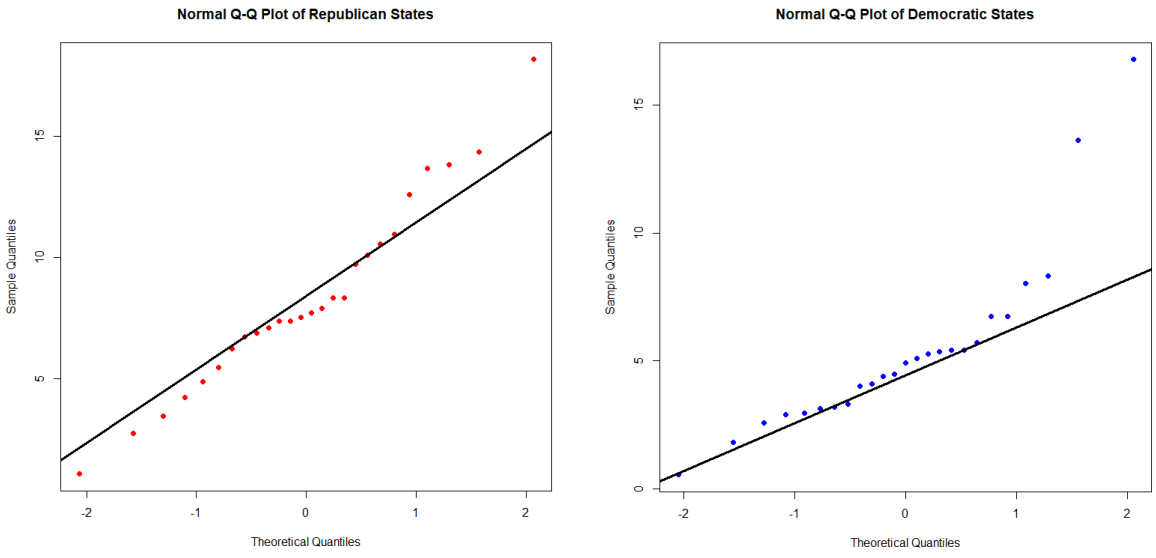
Positivity Rate is the best statistic that represents the spread of the disease of a given area. With this question, I try to figure out that whether the democratic ruled states have a lower transmission rate or the republican ruled states have a lower transmission rate. The steps taken by the Governors such as school closures, lockdowns, and restrictions on large gatherings have had some impact to control the spread of the virus. There are many factors responsible for the spread, but this can give us an idea that which of the two sets of states has a lower transmission rate.

Method Used:

A two-sample independent unpaired t-test is used to test the difference. This method is appropriate because it checks for the difference in the two means and reports it. It is easy to interpret and since the states are different the independent test is used. We will use a tow sided interval to check the difference

The validity of Assumptions:

It needs to be checked that both our samples are not much different than the normal distribution. A qqplot is used to check the validity of the assumptions.



There appears to be some skewness in both the samples but it is not significant and is acceptable

The test and Results:

Null Hypothesis(H_0): There is no difference in means of the two sets of states and the Positivity Rate means are the same. ($\mu_1 - \mu_2 = 0$)
Alternate Hypothesis(H_1): There is a difference in means of the two sets of states and the Positivity Rate means are not the same. ($\mu_1 - \mu_2 \neq 0$)

```
> t.test(x = Republican.Pos.Rate, y = Democratic.Pos.Rate,
+        alternative = "two.sided", paired = FALSE)

        Welch Two Sample t-test

data:  Republican.Pos.Rate and Democratic.Pos.Rate
t = 2.8585, df = 48.711, p-value = 0.006246
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8834677 5.0680615
sample estimates:
mean of x mean of y
 8.361885  5.386120
```

$\alpha = 1 - 0.95$
$\Rightarrow \alpha = 0.05$
$p - value = 0.006246$

Since $0.006246 < 0.05 \Rightarrow p - value < \alpha$ we reject the null hypothesis that there is no difference in means and conclude that the alternative hypothesis is true.

The alternative hypothesis states that there is a difference in means of the two results. i.e. I am 95% confident that there is a difference in the mean of Positivity Rates of the states with a Republican Governor and the mean of the states with a Democratic Governor.

The 95 confidence interval is

$0.8834677 - 5.0680615$

which is an indication that

$\mu_1(\text{Republican}) - \mu_2(Democratic) > 0 \Rightarrow \mu_1(Republican) > \mu_2(Democratic)$

This indicates that the positivity rate of the Republican states is higher than the positivity rate of the Democratic states. This is an indication that maybe the measures taken by the Democratic governors are more effective in controlling the spread than the measures taken by the Republican governors.

Question 2:

Is there a relationship between the case fatality ratio and the number of ICU beds and Physicians per 10000 of the population and is the interaction significant, if yes then how much of the variability is explained by these factors?

Why this question?

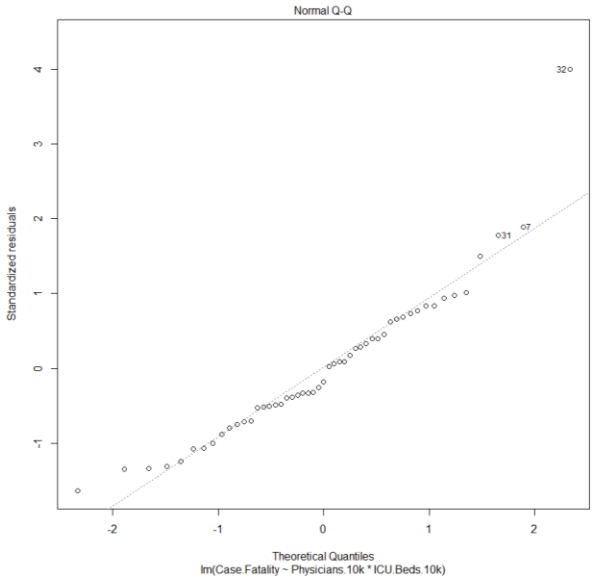
There is no doubt in the fact that the medical infrastructure all over the world including the United States has been strained with hospitalizations increasing. There has been a shortage of medical workers all across the globe. By this question, I will try to analyze a linear model of how the case fatality ratio is affected by the number of physicians and ICU Beds available relative to the population.

Method Used:

I will be using Multiple Regression and design a linear model that uses two numerical predictors. I will also consider the interaction between the two predictors. This method is appropriate because it can give us an idea that whether the lack of a significant number of ICU Beds and physicians is a key factor in the number of deaths.

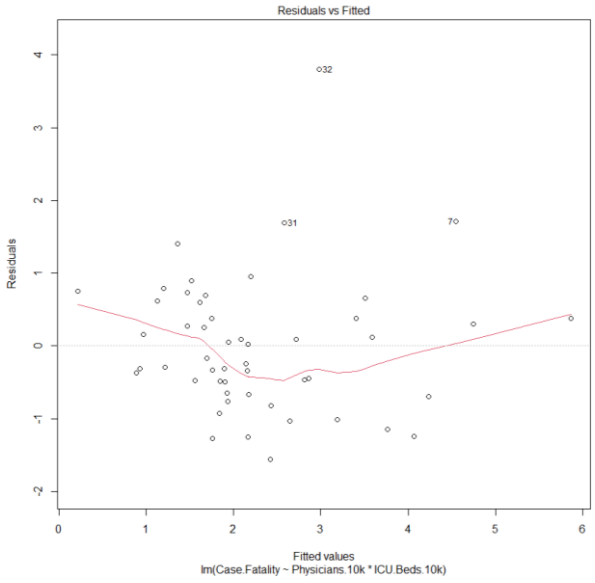
The validity of Assumptions:

Normality Check:



The residuals do look like they follow a normal distribution

Constant Variation Check:



The points look like they are evenly distributed across the mean and there is no observable pattern

The test and Results:

Null Hypothesis(H_0): There is no linear relationship between any predictor and the Case Fatality ratio

Alternate Hypothesis(H_1): There is a linear relationship between any predictor and the Case Fatality ratio

```
> summary(lmodel)

Call:
lm(formula = Case.Fatality ~ Physicians.10k * ICU.Beds.10k)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5591 -0.5722 -0.1694  0.4864  3.8027

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -4.91914    1.31459  -3.742 0.000496 ***
Physicians.10k              0.25690    0.03759   6.835 1.46e-08 ***
ICU.Beds.10k                1.40467    0.44353   3.167 0.002705 **
Physicians.10k:ICU.Beds.10k -0.05358    0.01035  -5.179 4.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9669 on 47 degrees of freedom
Multiple R-squared:  0.5787,    Adjusted R-squared:  0.5518
F-statistic: 21.52 on 3 and 47 DF,  p-value: 6.456e-09
```

$\alpha = 1 - 0.95$
$\Rightarrow \alpha = 0.05$
$p - value = 6.456e - 09$

Since $6.456e - 09 < 0.05 \Rightarrow p - value < \alpha$ we reject the null hypothesis that there is no difference in means and conclude that the alternative hypothesis is true(at 95% confidence). There is some linear relation with at least one of the predictors. Since the $p - values$ of all the predictors are very small and $< \alpha = 0.05$ we can say that all the predictors are significant in explaining some variability in the Case Fatality ratio.

Multiple R-squared: 0.5787 i.e. nearly 57.9 % of the variability in the case-fatality ratio is explained by this model.

Looking at the Anova Table

```
> anova(lmodel)
Analysis of Variance Table

Response: Case.Fatality
                           Df Sum Sq Mean Sq F value    Pr(>F)
Physicians.10k              1 30.738 30.7378 32.8816 6.780e-07 ***
ICU.Beds.10k                1  4.542  4.5421  4.8589   0.03244 *
Physicians.10k:ICU.Beds.10k  1 25.074 25.0737 26.8224 4.581e-06 ***
Residuals                  47 43.936  0.9348
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the Physicians initially explain a decent chunk and adding ICU beds also explains a decent amount and the interaction also explains a good amount. This is an indication that there is some correlation between ICU beds, Physicians, and the interaction is significant. Although the slope cannot be interpreted easily we can see there is a relation.

Question 3:

Is there a correlation between the Positivity rate the average Temperature of the state?

Why this question?

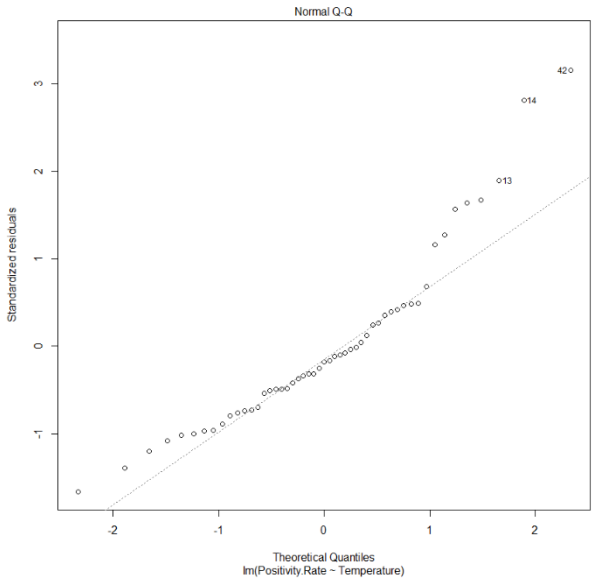
On February 10th, 2020 President Trump said "You know, a lot of people think that goes away in April with the heat — as the heat comes in. Typically, that will go away in April."(source) At this time the United States had a handful of cases, a national emergency was declared on, March 13th, 2020 a national emergency was declared and ever since the pandemic has ravaged the country. At that point, many people did believe that the virus spread will slow down if not disappear by summer. The data includes data from Feb – November and different states have different average temperatures and I try to analyze whether the relatively hotter states have a lower positivity rate or vice versa.

Method Used:

We will try to estimate the correlation using the Hypothesis Test for Population Correlation i.e. using the cor.test() method in R. This method will tell us that if we have significant evidence to say that there is a correlation between the two variables.

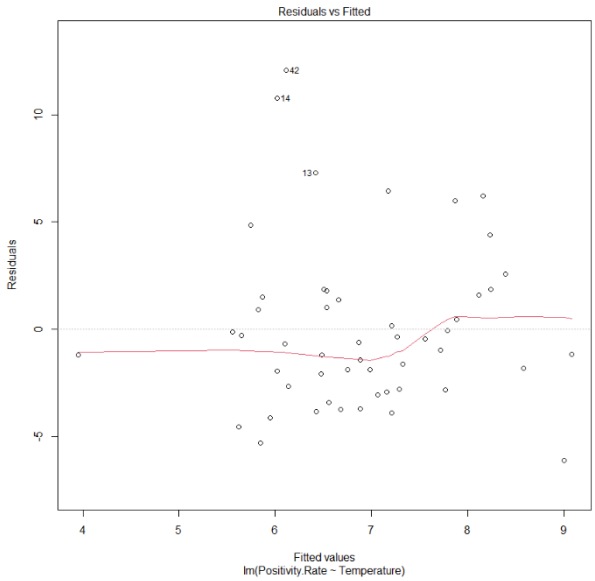
The validity of Assumption

Normality Check for Residuals:



The residuals do look like they somewhat follow a normal distribution

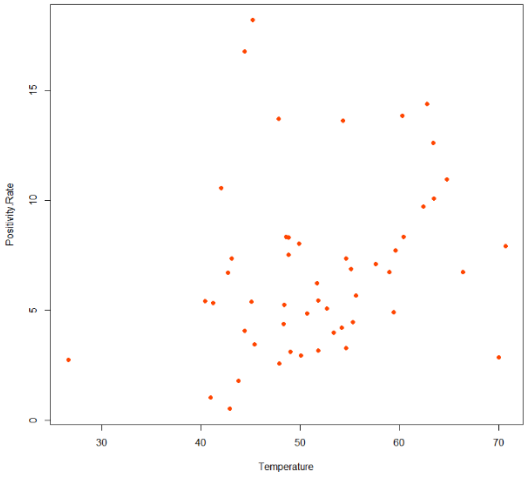
Constant Variation Check:



The points look like they are evenly distributed across the mean and there is no observable pattern

The test and Results:

Null Hypothesis(H_0): There is no correlation between Positivity Rate and Temperature and the correlation is equal to 0
Alternate Hypothesis(H_1): There is a correlation between Positivity Rate and Temperature and the correlation is not equal to 0



This is a plot of Positivity Rate vs Temperature. We can see that there is not a significant correlation as points look randomly scattered. Although there can be a weak positive relationship

```
> cor.test(Positivity.Rate, Temperature, alternative = "two.sided")

        Pearson's product-moment correlation

data:  Positivity.Rate and Temperature
t = 1.823, df = 49, p-value = 0.07441
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0253226  0.4933398
sample estimates:
      cor
0.2520198
```

$\alpha = 1 - 0.95$
$\Rightarrow \alpha = 0.05$
$p - value = 0.07441$

Since $0.07441 > 0.05 \Rightarrow p - value > \alpha$ we fail to reject the null hypothesis since there is no significant evidence to conclude that there is a correlation between the two values. Hence we conclude that the null hypothesis is true and there is no significant relationship between the two.

In the end, we can conclude that the Positivity Rate is related to the average temperature of the state. Although if we were to do a day to day analysis we might find something conclusive.