Abstract:

The global epidemic of the 2019 Novel Coronavirus disease (COVID-19) is one of the most pressing scientific issues of our time. COVID-19 is a highly transmissible, rapidly spreading virus, that has placed enormous strain of the healthcare systems of even the wealthiest nations. In this study, I will examine statistics regarding COVID-19 Infection rates and deaths in various U.S. States. Developing an understanding of the infection rate and fatality rate of the virus in different areas is critical and allows us to answer basic yet critically important questions about the virus. Does the virus spread at the same rate everywhere? If not, how statistically significant are the differences in infection rate? At what rate does the virus kill those it infects? Does where you live affect your chances of dying of COVID-19 after being infected? This study demonstrates that not only are infection rates and death rates drastically different in Different U.S. States, but that you have dramatically different odds of survival if infected with COVID-19 in different regions.

Introduction:

In early 2020, the emergence of the 2019 Novel Coronavirus disease (COVID-19) in Wuhan China began to transform from an endemic local disease outbreak into a global pandemic. The highly infectious SARS-COV-2 virus spread rapidly across international borders, binging societies across the globe to a near standstill. "The 2019 novel-coronavirus (COVID-19) has affected 181 countries with approximately 1197404 confirmed cases by 5th April" (1). In response to this emerging threat, scientists, researchers, and medical professionals from around the world have made tracking, treating, and containing the virus their top priority. In the campaign waged by the scientific and medical professionals against the danger of viral spread, geography has played an important role, aiding public health authorities track the spread of COVID-19. "Understanding the transmission dynamics of the infection in each country which got affected on a daily basis and evaluating the effectiveness of control policies are critical for our further actions" (1). Medical, statistical, and geographic analysis have been combined for the purposes of tracking where the proliferation or control of the virus is occurring. Using geographically based statistics

on Covid 19, public health officials can depict infection rates and death rates in different areas. Depicting these factors spatially allows public health officials and infectious disease experts to better understand where, how, and at what rate the virus is spreading. (2) This knowledge is critical, as it allows politicians and other public decision makers to better understand the health risks to their constituencies, and to make more informed choices on what policies should be put in place to reduce risk. (3)

This project consists of a study of COVID-19 cases and deaths in the United States. The data used in this study was compiled by the U.S. Centers for Disease Control based on daily reports of COVID-19 cases and deaths in each U.S. state and territory from 01 January 2020 to 12 April 2021. U.S. States vary widely from one another in population size, population density, and other factors, and thus have widely varying numbers of COVID-19 cases. In this study, I have elected to compare COVID-19 cases and death rates in Oklahoma with six other States with mid-size populations (between three million and five million people). The states included in this study are Louisiana, Kentucky, Oregon, Oklahoma, Connecticut, Utah, and Iowa. These areas are diverse in terms of geographical region, total land area, population density, and demographics. Additionally, these areas took a diverse array of approaches to implementing shutdowns, social distancing, and other COVID-19 mitigation techniques. This study will examine the reported case counts and death rates of these states over time to see how they compare against one another.

Data summary

Records within the CDC's United States COVID-19 Cases and Deaths Dataset are listed by state, territory, or other Jurisdictions (NYC has its own entry record due to high case volumes in the city). The dataset has recorded entries submitted by every jurisdiction in the U.S. for every day of the pandemic starting with discovery of the first case in the Unites States on 1/22/2020 and continuing to 4/12/2021. The dataset is continuously updated each day, but I will analyze data through 04/12 for the purposes of this study. Each record lists the total number of cases so far in the pandemic, total number of deaths, new

cases, and new deaths per day. Entries are dependent on the accurate and timely reporting of COVID-19 infection and death rates to the CDC. This may be difficult in smaller jurisdictions which are not sufficiently resourced to manage mass infection and mass casualty events. This lack of resources leads to delays in testing and reporting. (4) Additionally, tracking the exact number of COVID-19 cases in any jurisdiction is not possible. COVID-19 may only cause mild illness in some individuals, especially young people. Symptoms might not appear immediately or at all during the period where virus is transmissible. Individuals with a mild case of COVID-19 may not suspect they are infected with the virus and may not seek testing. (4) This is especially problematic in communities of young people, as it can lead to undetected clusters of the virus that can spread to their more vulnerable relatives and contacts. Not every individual infected with COVID-19 will get tested or receive medical treatment. This leads to cases going unreported to the CDC. Furthermore, because COVID-19 is one of 120 diseases for which reporting to the CDC is voluntary, there are differences in how completely States and Territories report their COVID-19 cases and fatalities. (4). Health departments generally update their COVID-19 case data whenever they receive up to date information from hospitals and other medical bodies within their jurisdiction. This is sometimes done on an irregular basis by smaller jurisdictions at the county level, as their hospitals, treatment centers, and other health infrastructure are often operating above their normal capacity, and it takes time and personnel to compile and submit data.

Results and analysis, Part 1: Analyzing Individual states

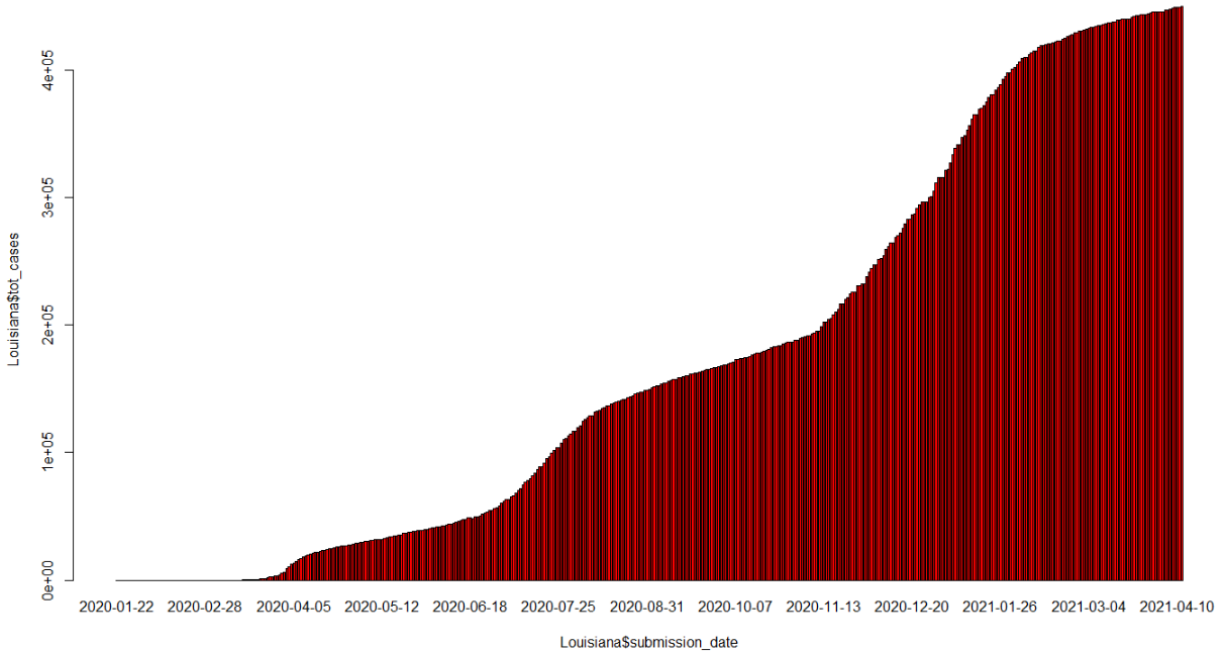
I will start my analysis by giving a detailed summary of the Covid statistics of each state over time. This will allow the opportunity to examine the details of how the virus effected each state over time throughout the course of the pandemic and will allow for more meaningful comparison between the states later on in the study. This section will consist primarily of graphs and charts of individual states with numerical descriptions performed in Rstudio. Analysis of these results is listed in the Part 1 Summary at the end of the section. States in Part 1 are listed by population size in descending order. Comparative statistical

analysis of the relationship between cases and death rates in different states will be conducted in Part 2 of the study.

Louisiana: Population: 4,648,794 as of July 1, 2019 (U.S. Census Bureau) (5)

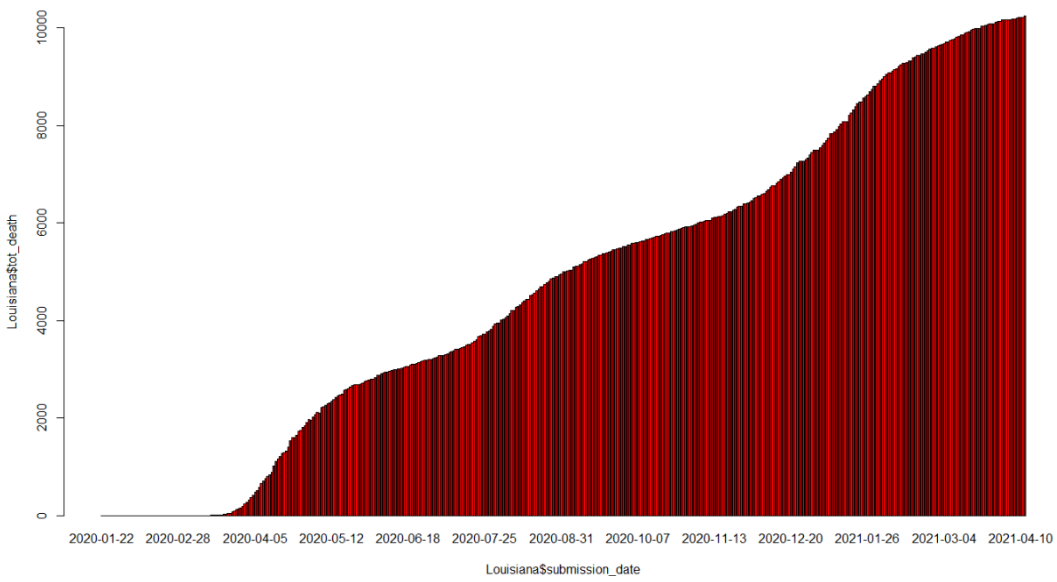
Louisiana total cases over time:

```
> summary(Louisiana$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   32422  149583  172223  289111  449827
> stat.desc(Louisiana$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
1.495830e+05 1.722227e+05 7.228543e+03 1.420623e+04 2.335657e+10 1.528286e+05 8.873891e-01
> describe(Louisiana$tot_cases)
   vars   n      mean       sd median  trimmed      mad min    max  range skew kurtosis      se
X1    1 447 172222.7 152828.5 149583 160459.1 177670.3   0 449827 449827 0.56    -1.06 7228.54
>
```

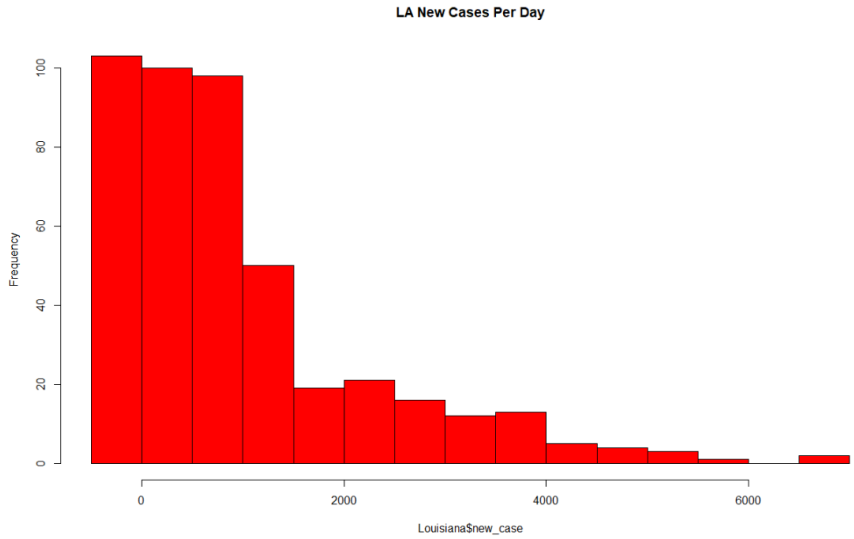


Louisiana total deaths over time:

```
> summary(Louisiana$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    2364    5004    4759    7132   10241
> stat.desc(Louisiana$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
5.004000e+03 4.759011e+03 1.529749e+02 3.006411e+02 1.046038e+07 3.234252e+03 6.796058e-01
> describe(Louisiana$tot_death)
   vars   n    mean       sd median trimmed     mad min   max range skew kurtosis     se
X1    1 447 4759.01 3234.25   5004 4702.67 3552.31   0 10241 10241 0.06     -1.1 152.97
>
```

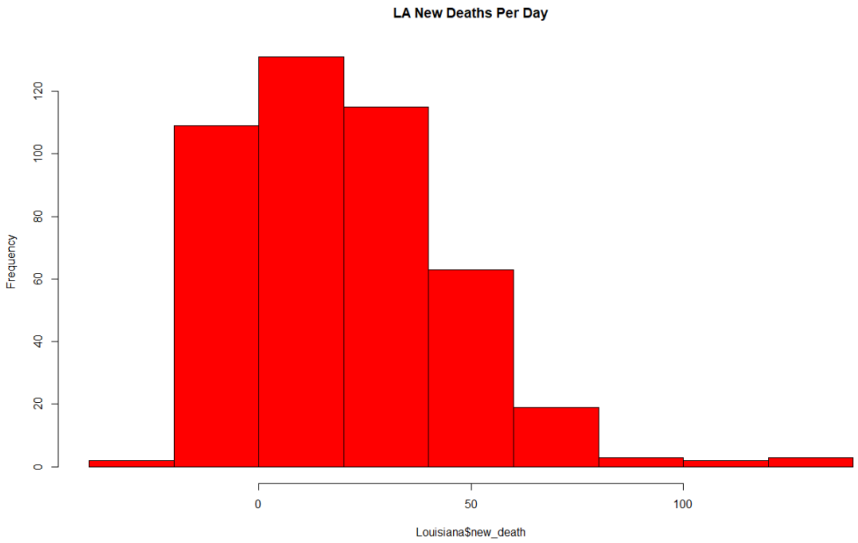


Histogram of Louisiana new cases per day:



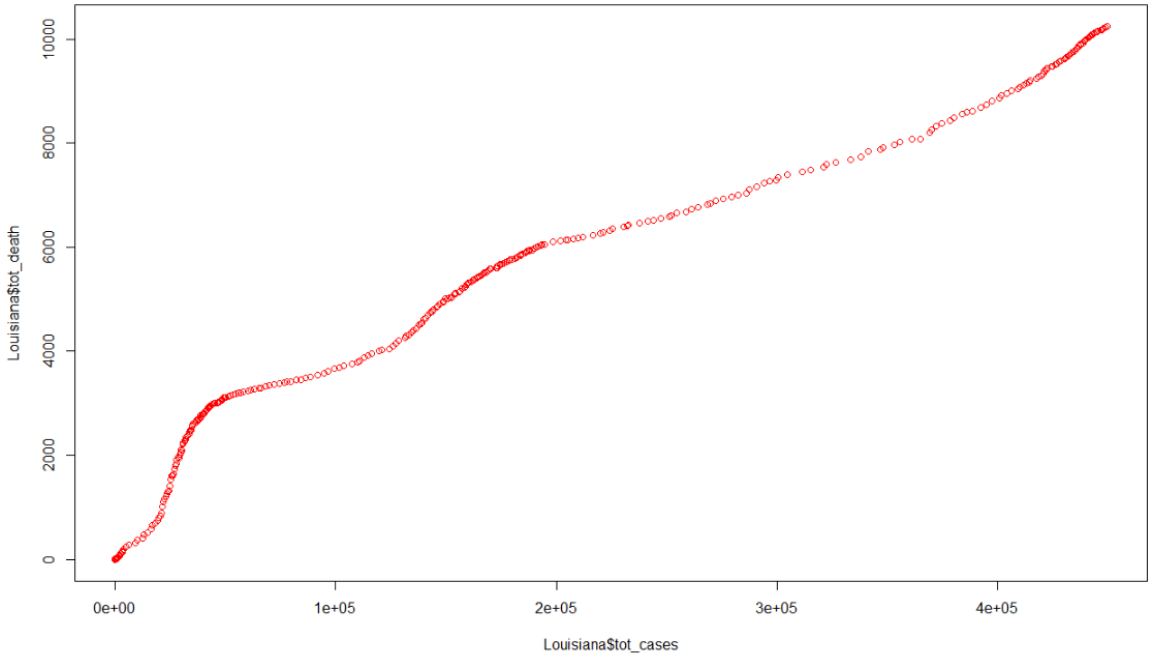

```
> describe(Louisiana$new_case)
   vars   n    mean      sd median trimmed    mad  min  max range skew kurtosis    se
X1    1 447 1006.32 1231.45    558  764.38 827.29 -119 6876  6995 1.83     3.46 58.25
> |
```

Histogram of Louisiana new deaths per day:




```
> describe(Louisiana$new_death)
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 447 22.91 22.74     19   19.98 23.72 -21 129   150 1.32     2.62 1.08
> |
```

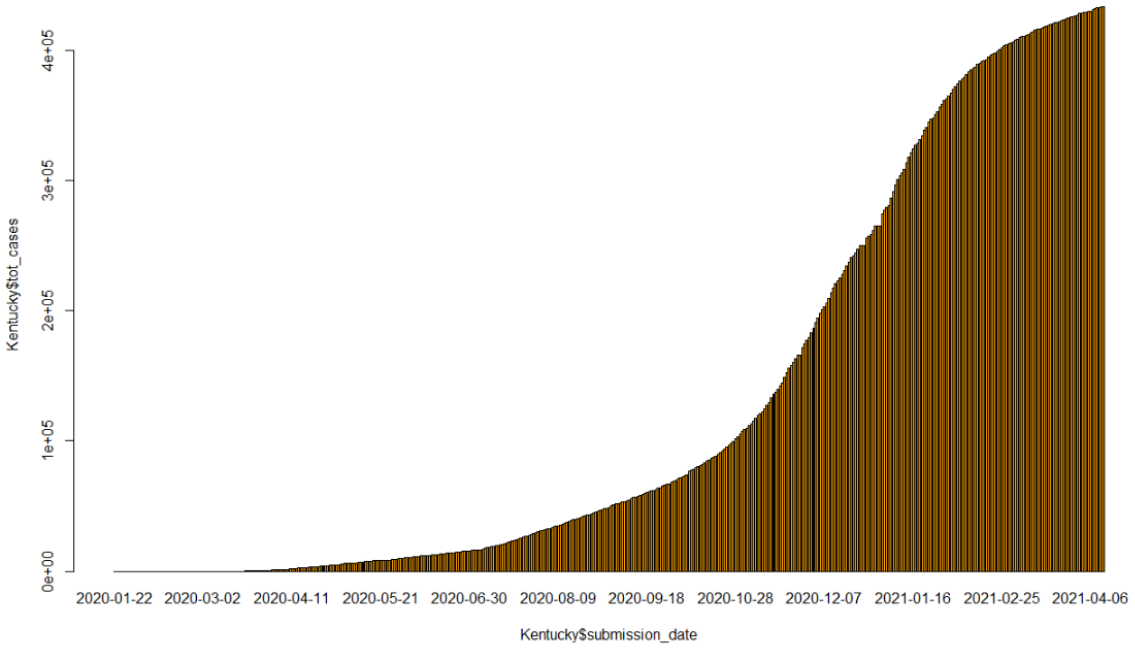
Plot of the relationship between total cases and total deaths in Louisiana:



Kentucky: Population: 4,467,673 as of July 1, 2019 (U.S. Census Bureau) (5)

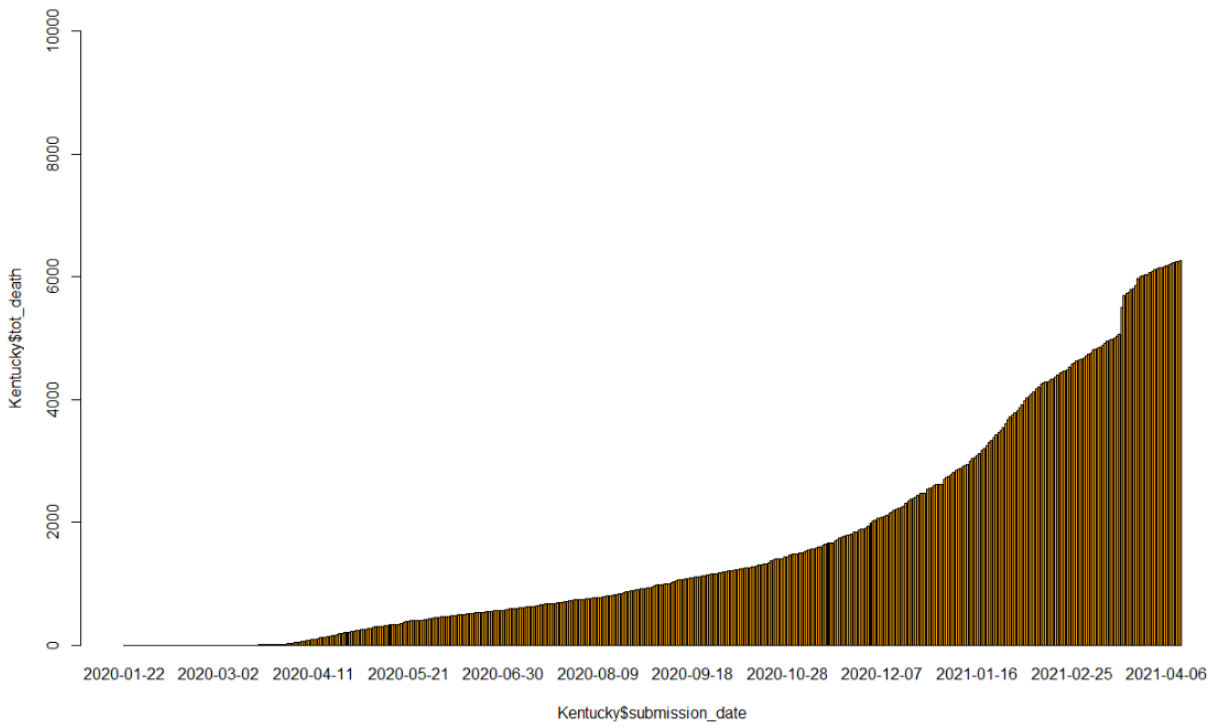
Kentucky total cases over time:

```
> #Kentucky central tendency
> summary(Kentucky$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    6966   49185  127998  245821  433352
> stat.desc(Kentucky$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
4.918500e+04 1.279982e+05 7.211610e+03 1.417296e+04 2.324727e+10 1.524706e+05 1.191193e+00
> describe(Kentucky$tot_cases)
   vars   n     mean       sd median  trimmed      mad min    max  range skew kurtosis      se
X1    1 447 127998.2 152470.6  49185 107839.7 72883.13   0 433352 433352 0.93    -0.75 7211.61
>
```

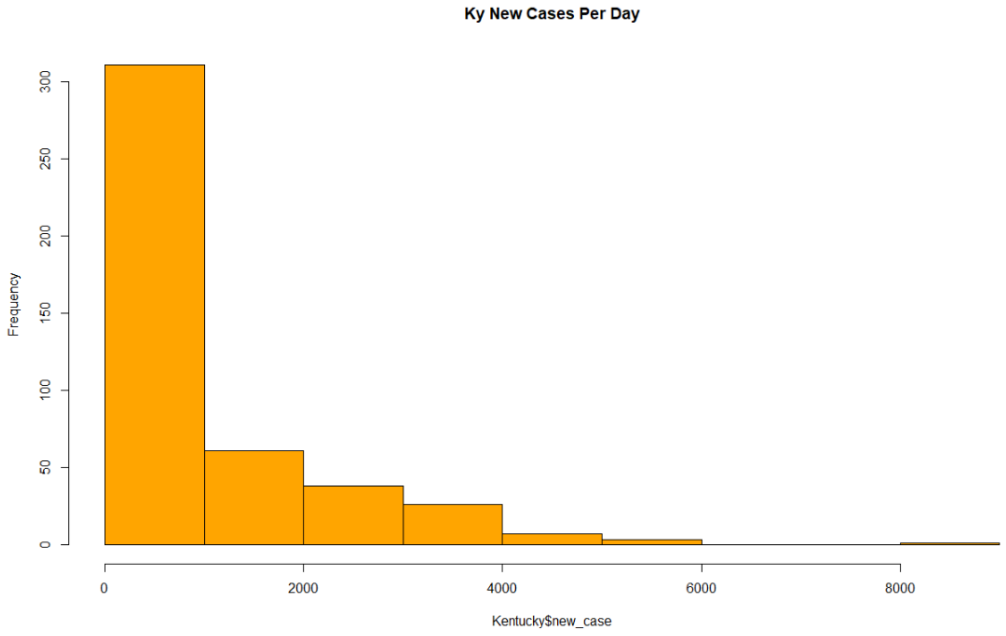


Kentucky total deaths over time:

```
> summary(Kentucky$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   323.5   948.0  1623.8  2426.0  6257.0
> stat.desc(Kentucky$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
9.480000e+02 1.623799e+03 8.293613e+01 1.629941e+02 3.074645e+06 1.753467e+03 1.079855e+00
> describe(Kentucky$tot_death)
   vars   n    mean       sd median trimmed     mad min  max range skew kurtosis    se
X1    1 447 1623.8 1753.47    948  1342.5 1251.31   0 6257  6257 1.21     0.36 82.94
```

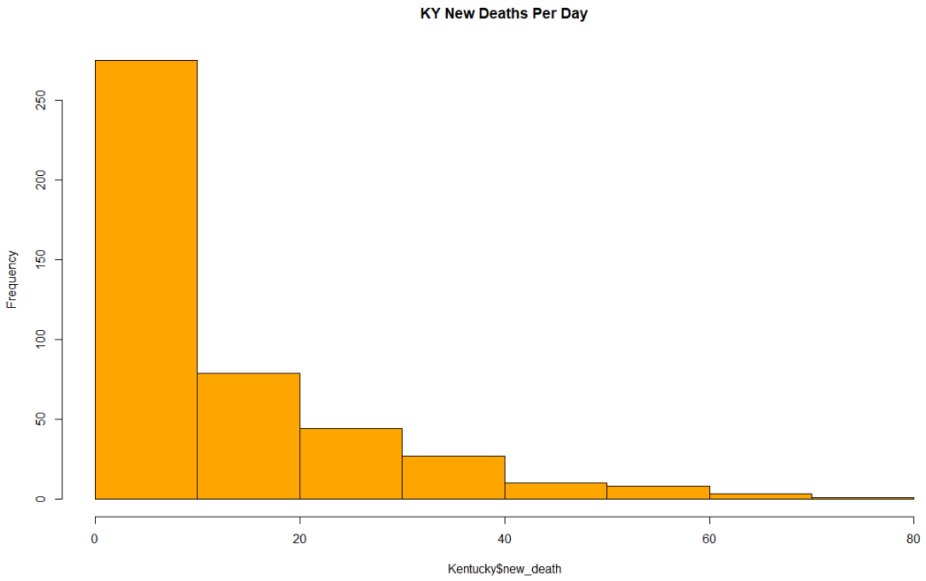


Histogram of new cases per day in Kentucky:



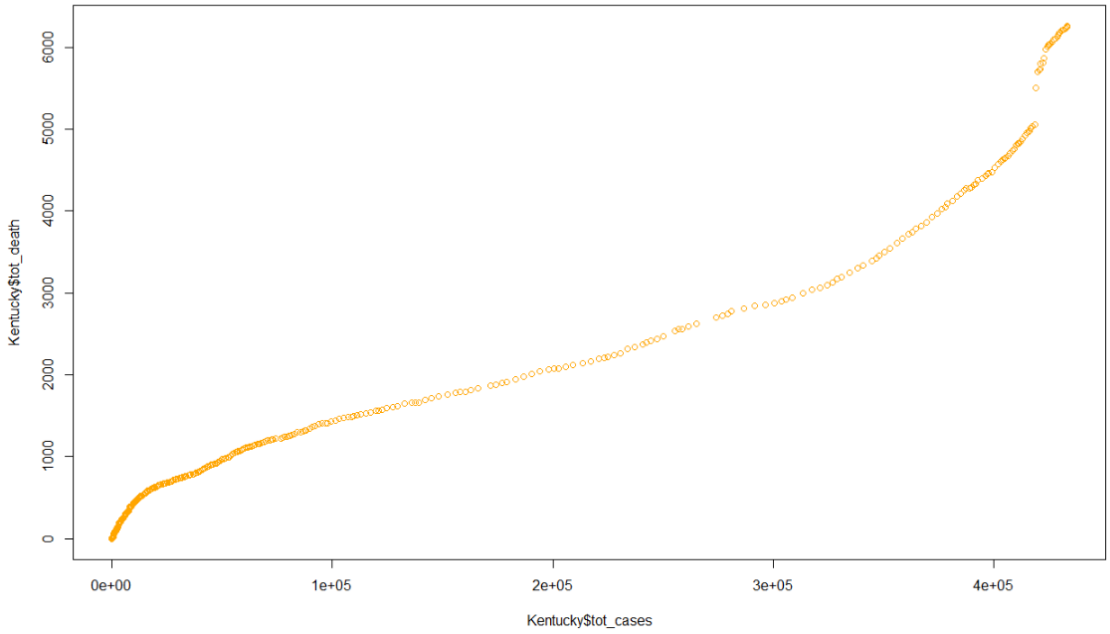

```
> describe(Kentucky$new_case)
   vars   n   mean      sd median trimmed    mad min  max range skew kurtosis    se
X1    1 447 969.47 1198.34    542  741.82 693.86   0 8709  8709 1.95     5.05 56.68
```

Histogram of new deaths per day in Kentucky:



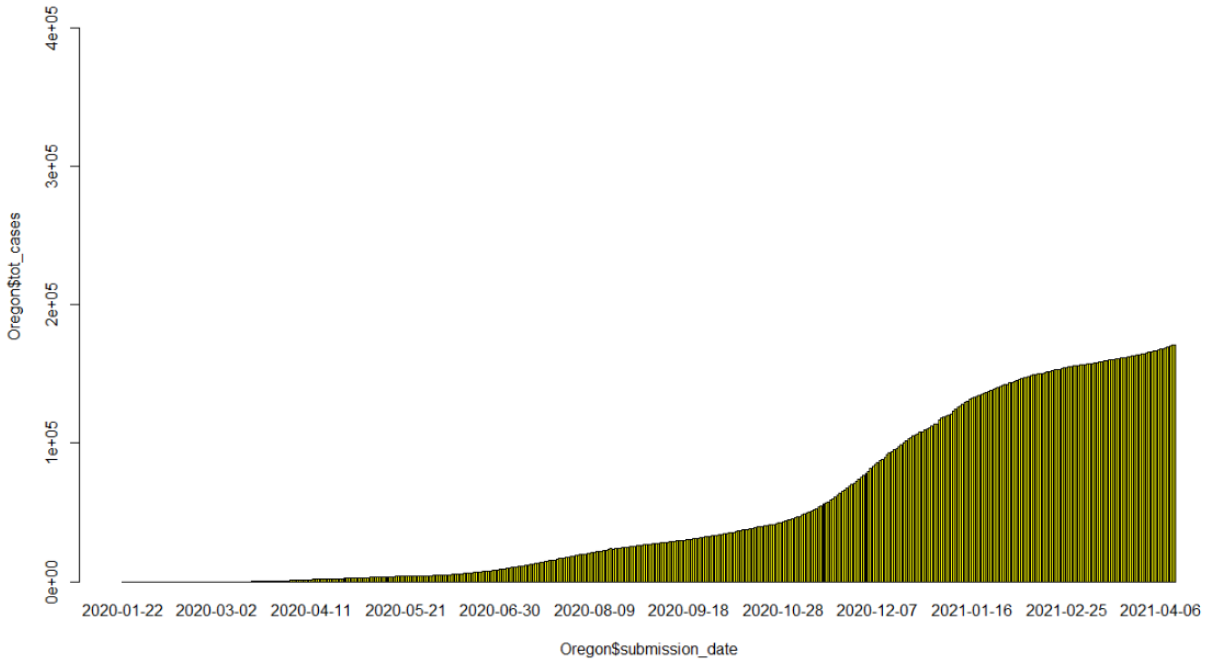

```
> describe(Kentucky$new_death)
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 447 12.25 13.67      8    9.79 10.38   0  75    75  1.7     2.96 0.65
```

Plot of the relationship between total cases and total deaths in Kentucky:



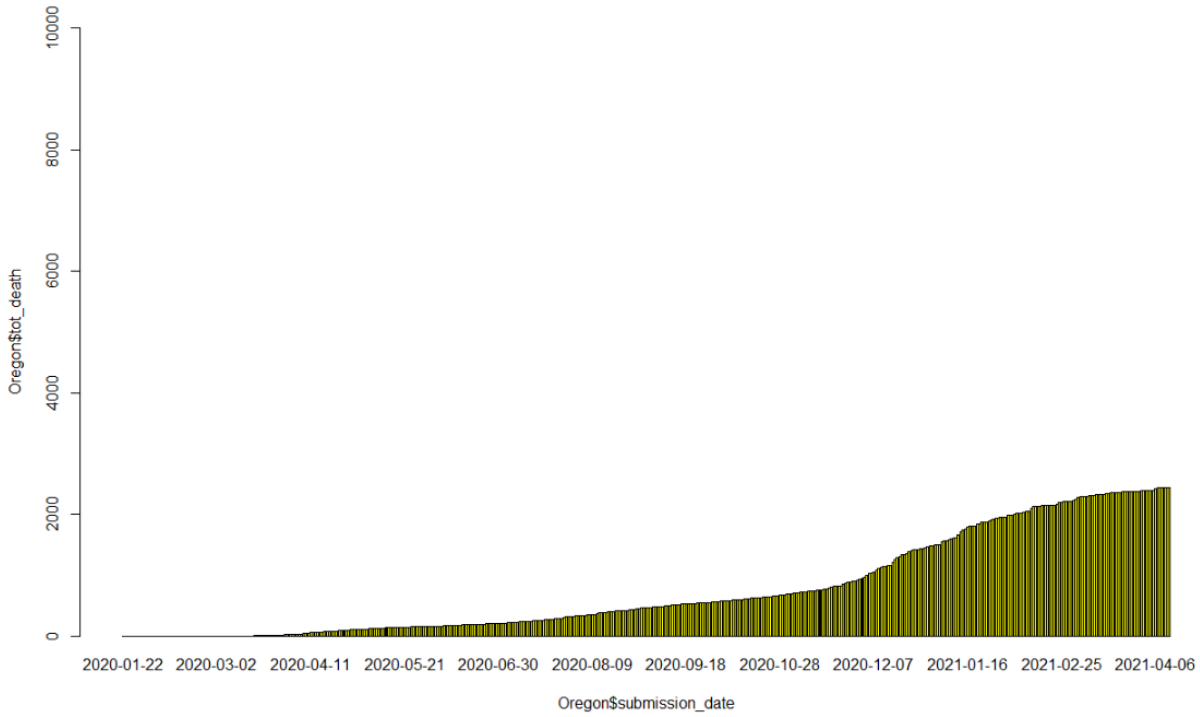
Oregon: Population: 4,217,737 as of July 1, 2019 (U.S. Census Bureau) (5)

```
> #Oregon central tendency
> summary(Oregon$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3387   26946   52739  104414  170850
> stat.desc(Oregon$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
2.694600e+04 5.273864e+04 2.789302e+03 5.481807e+03 3.477752e+09 5.897247e+04 1.118202e+00
> describe(Oregon$tot_cases)
   vars   n     mean       sd median  trimmed      mad min    max  range skew kurtosis     se
X1    1 447 52738.64 58972.47  26946 45783.78 38725.51   0 170850 170850 0.84     -0.9 2789.3
```

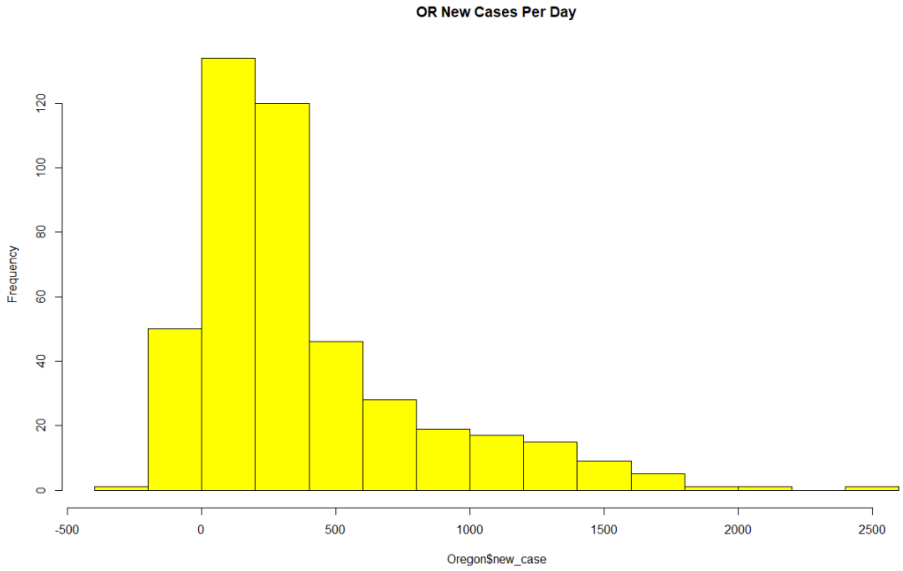


Oregon total deaths over time:

```
> summary(Oregon$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   132.0   465.0   764.8  1364.5  2441.0
> stat.desc(Oregon$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
4.650000e+02 7.647606e+02 3.817386e+01 7.502298e+01 6.513878e+05 8.070860e+02 1.055345e+00
> describe(Oregon$tot_death)
   vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 447 764.76 807.09    465  664.08 596.01   0 2441  2441 0.91    -0.65 38.17
```

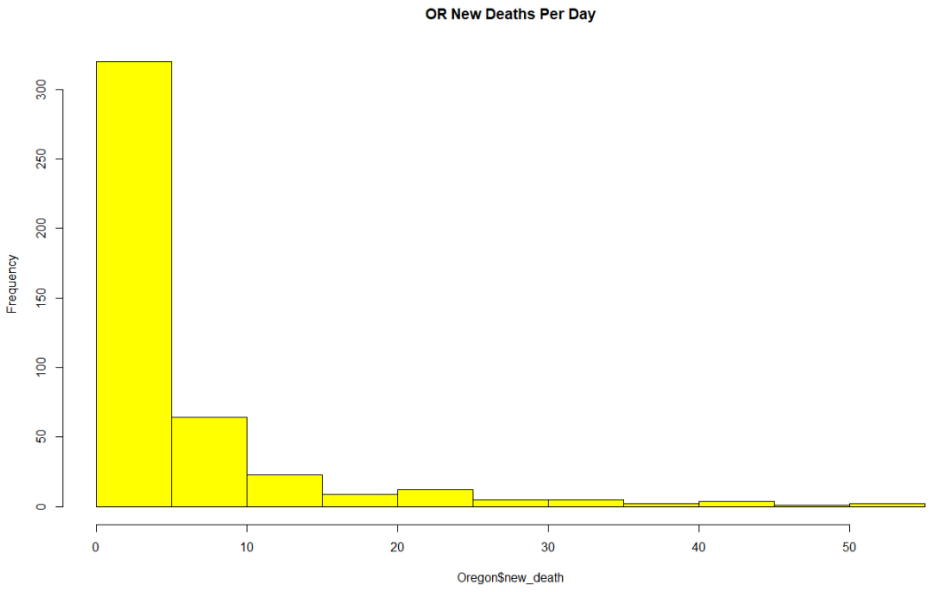


Histogram of new cases per day in Oregon:



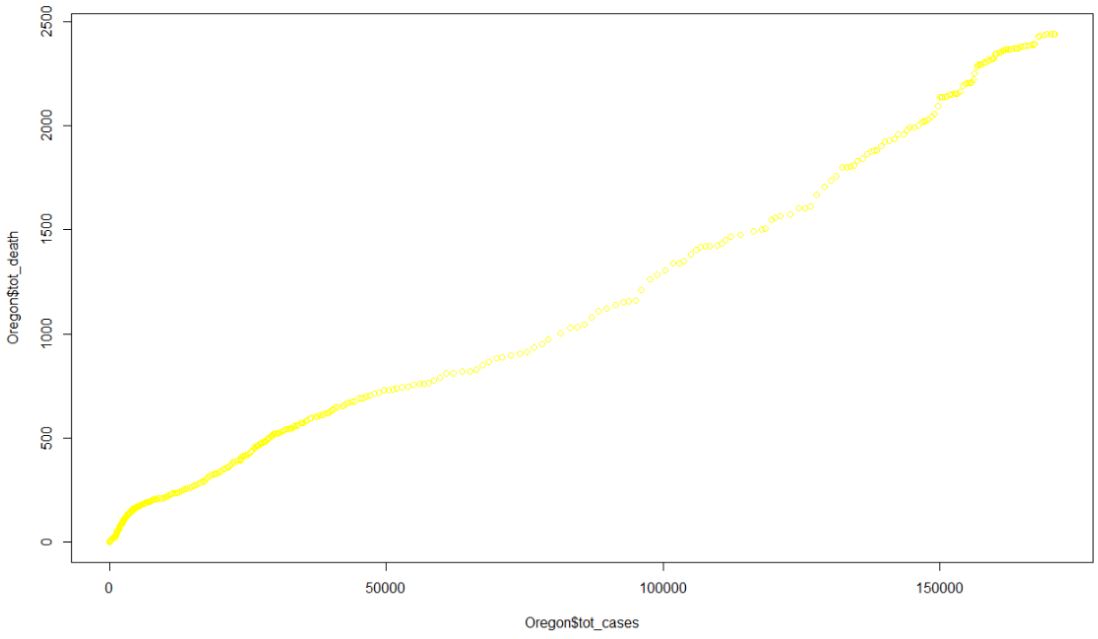

```
> describe(Oregon$new_case)
   vars   n   mean  sd median trimmed    mad  min  max range skew kurtosis    se
X1    1 447 382.21 424    261  305.96 300.97 -293 2439  2732 1.62     2.51 20.05
```

Histogram of new deaths per day in Oregon:




```
> describe(Oregon$new_death)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis  se
X1    1 447 5.46 8.51      3    3.45 4.45   0  54    54  2.9      9.5 0.4
```

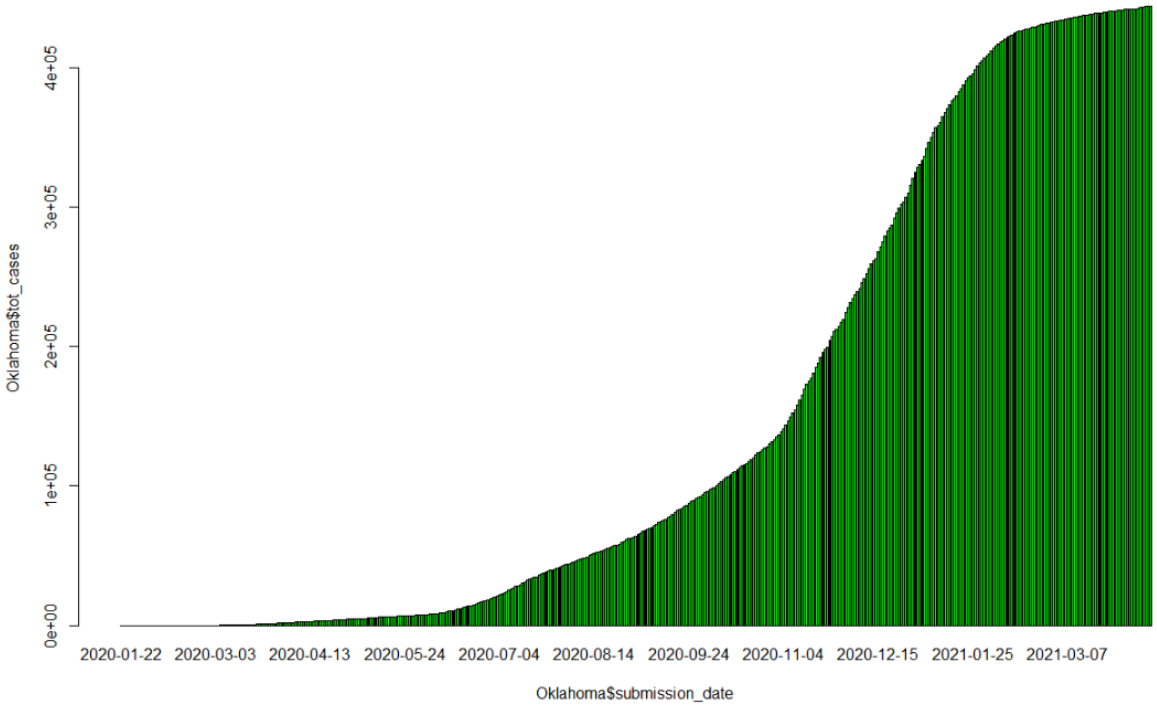
Plot of the relationship between total cases and total deaths in Oregon:



Oklahoma: Population: 3,956,971 as of July 1, 2019 (U.S. Census Bureau) (5)

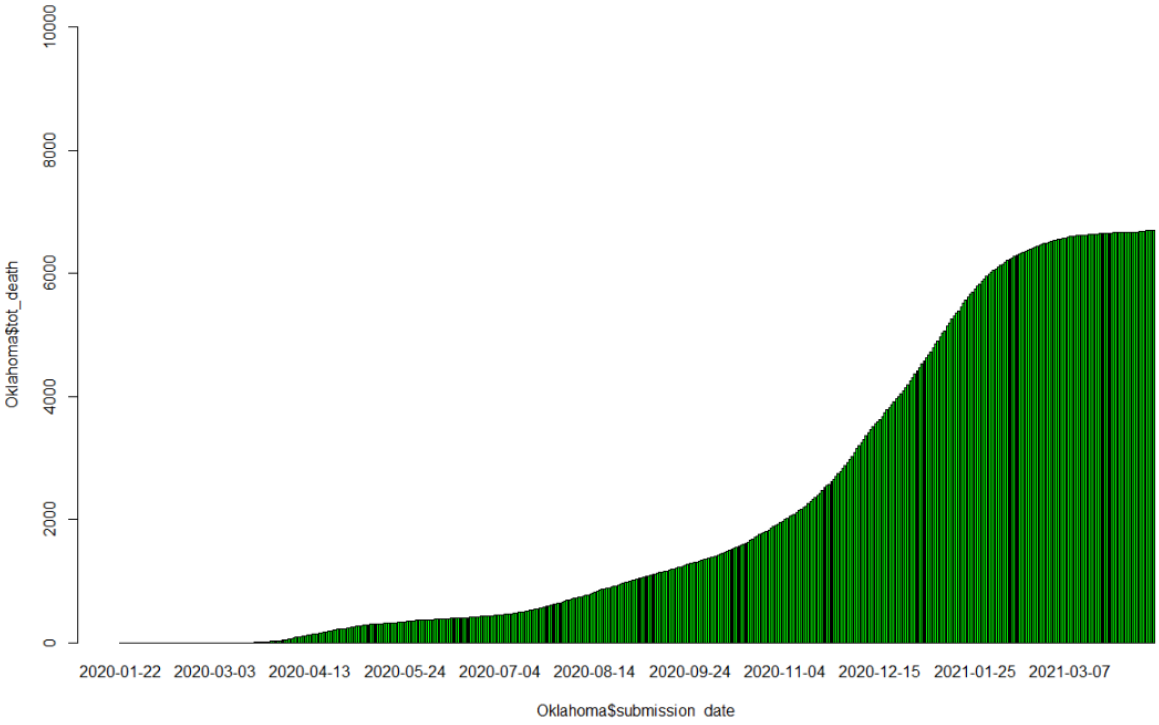
Oklahoma total cases over time:

```
> summary(Oklahoma$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    5654   64105  145809  289548  443882
> stat.desc(Oklahoma$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
6.410500e+04 1.458089e+05 7.803106e+03 1.533542e+04 2.721714e+10 1.649762e+05 1.131455e+00
> describe(Oklahoma$tot_cases)
   vars   n     mean       sd median  trimmed      mad min    max  range skew kurtosis      se
X1    1 447 145808.9 164976.2  64105 127783.3 94201.44   0 443882 443882 0.79    -1.01 7803.11
```

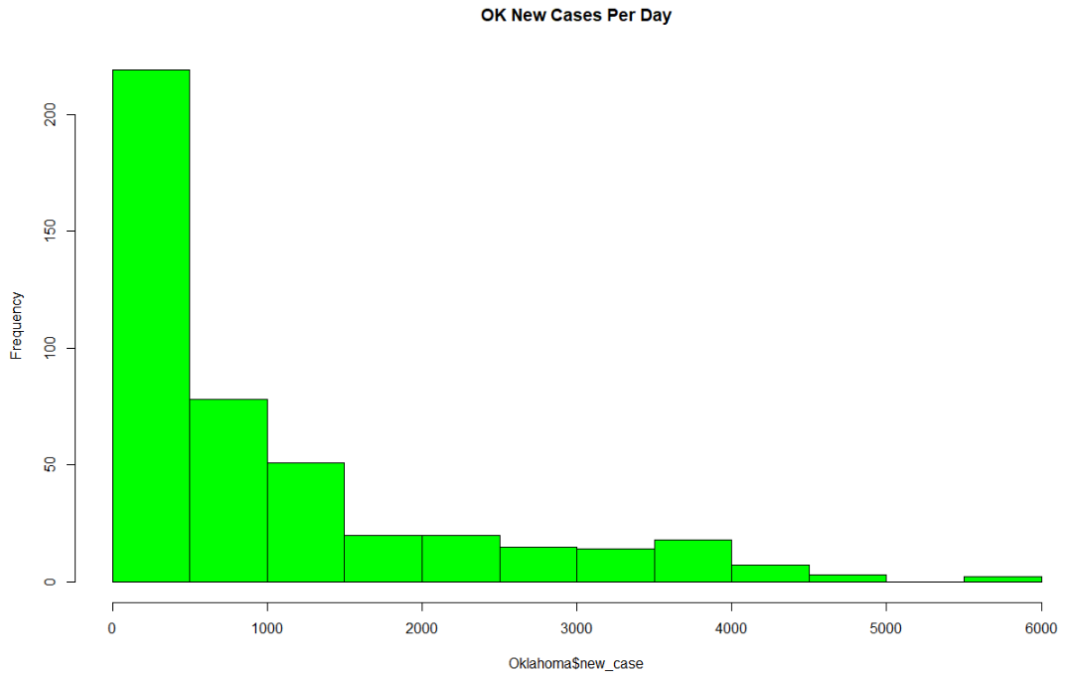


Oklahoma total deaths over time:

```
> summary(Oklahoma$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0     303    1027    2156    3934    6697
> stat.desc(Oklahoma$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
1.027000e+03 2.155893e+03 1.132654e+02 2.226002e+02 5.734585e+06 2.394699e+03 1.110769e+00
> describe(Oklahoma$tot_death)
   vars   n    mean     sd median trimmed     mad min  max range skew kurtosis     se
X1    1 447 2155.89 2394.7   1027 1870.78 1494.46   0 6697  6697 0.89     -0.8 113.27
```

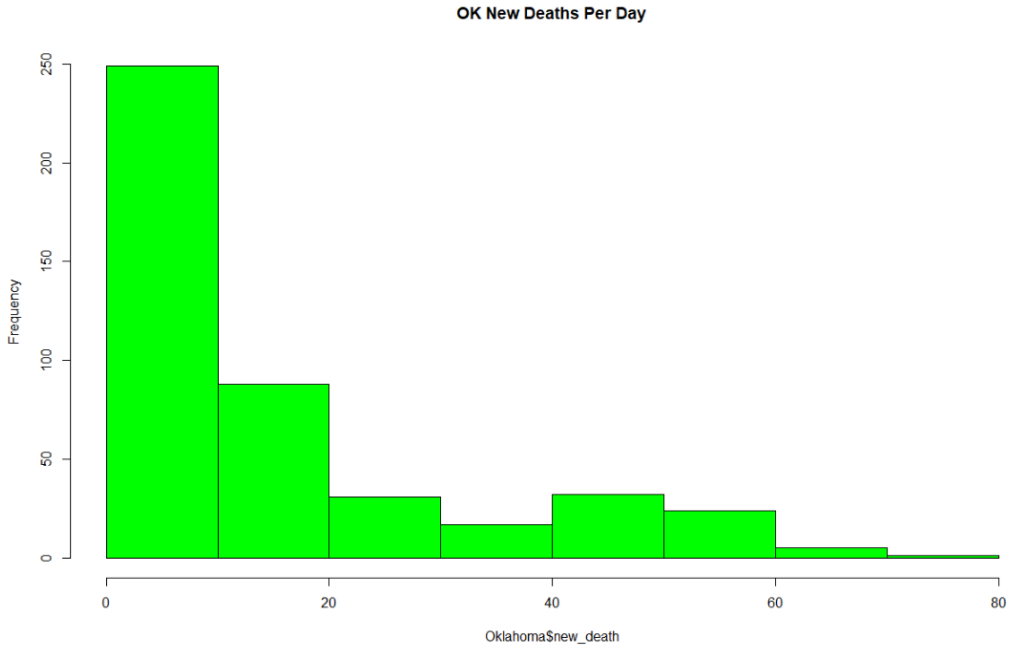


Histogram of new cases per day in Oklahoma:



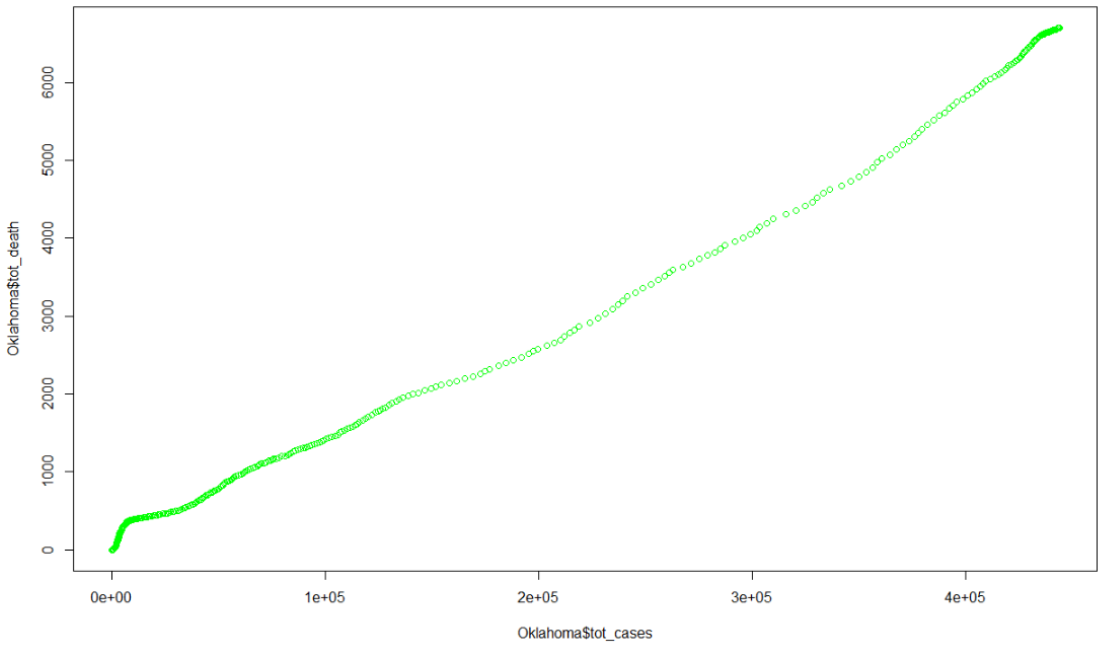

```
> describe(Oklahoma$new_case)
   vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 447 993.02 1182.6    535  768.89 690.89   0 5931  5931 1.54     1.75 55.93
```

Histogram of new deaths per day in Oklahoma:




```
> describe(Oklahoma$new_death)
   vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 447 14.98 17.13      9   11.99 10.38   0  72    72 1.36     0.76 0.81
```

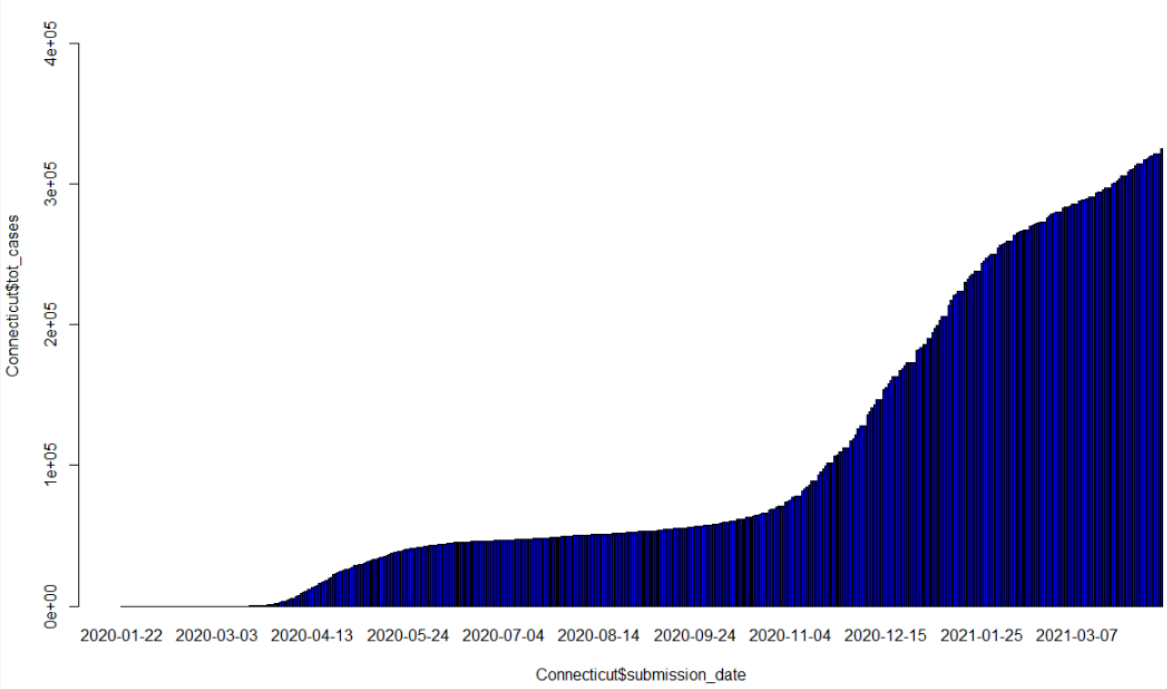
Plot of the relationship between total cases and total deaths in Oklahoma:



Connecticut: Population 3,565,287 as of July 1, 2019 (U.S. Census Bureau) (5)

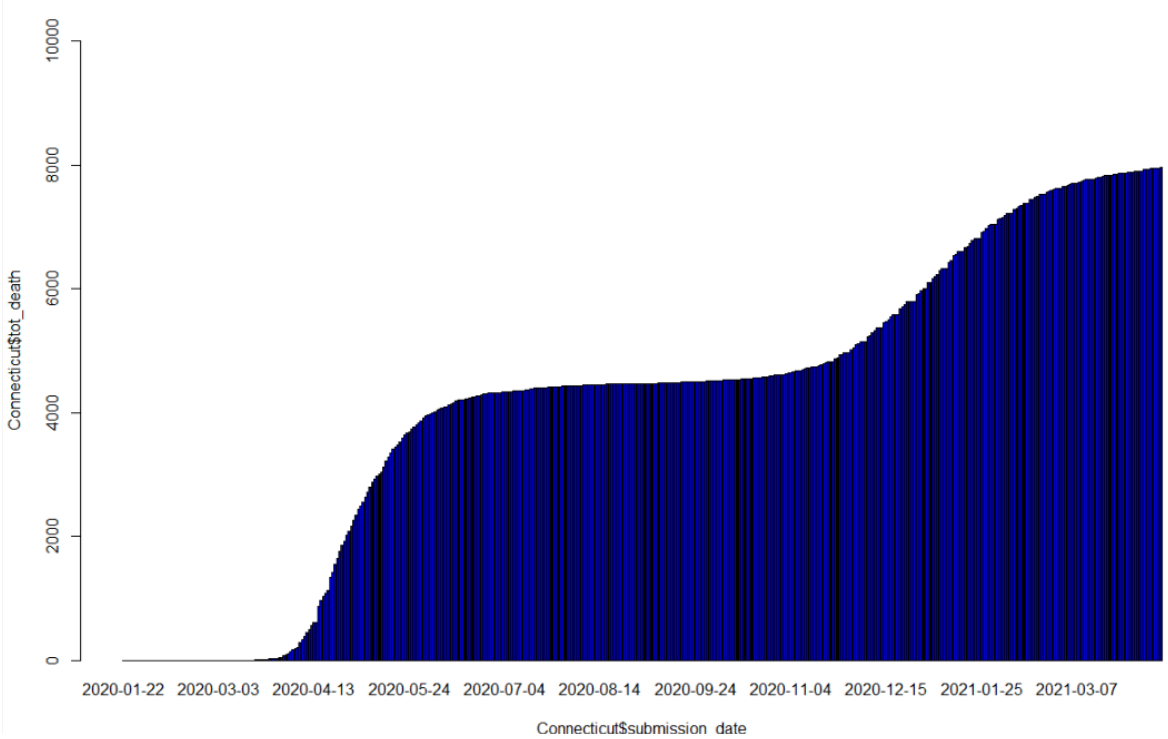
Connecticut total cases over time:

```
> #Connecticut central tendency
> summary(Connecticut$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   34594   53006   99738  168169  324571
> stat.desc(Connecticut$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
5.300600e+04 9.973829e+04 4.764557e+03 9.363771e+03 1.014735e+10 1.007341e+05 1.009984e+00
> describe(Connecticut$tot_cases)
   vars   n     mean       sd median  trimmed      mad min    max  range skew kurtosis      se
X1    1 447 99738.29 100734.1  53006 87335.84 65267.02   0 324571 324571 0.98    -0.52 4764.56
```

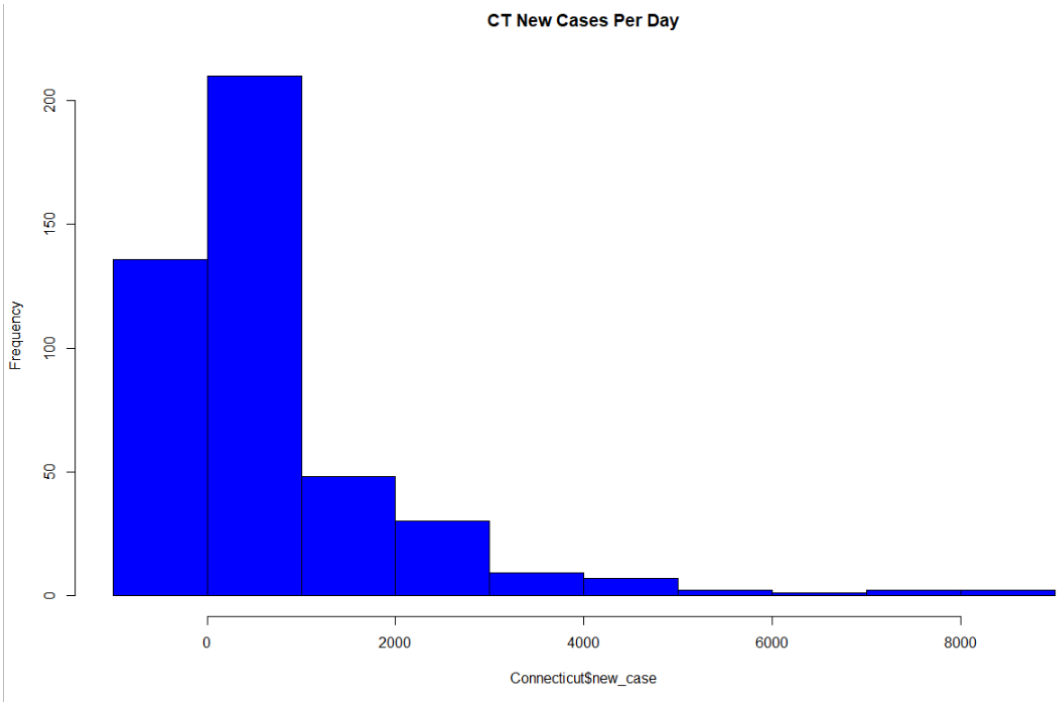


Connecticut total deaths over time:

```
> summary(Connecticut$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3083    4466    4178    5690    7957
> stat.desc(Connecticut$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
4.466000e+03 4.178434e+03 1.163643e+02 2.286905e+02 6.052672e+06 2.460218e+03 5.887895e-01
> describe(Connecticut$tot_death)
   vars   n    mean      sd median trimmed    mad min  max range skew kurtosis     se
X1    1 447 4178.43 2460.22   4466 4243.97 1882.9   0 7957  7957 -0.4    -0.72 116.36
```

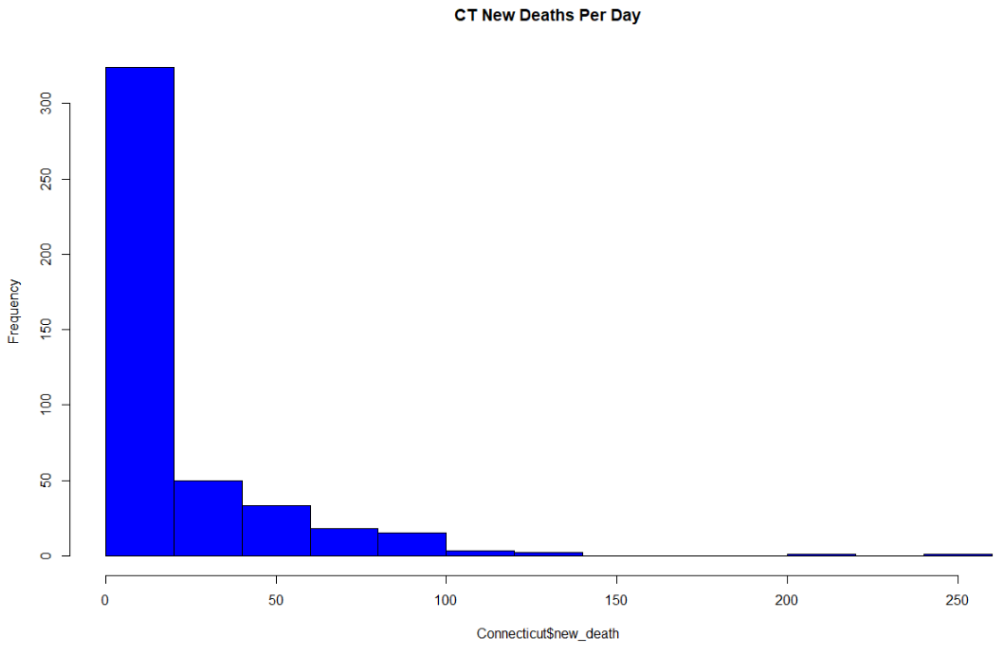


Histogram of new cases per day in Connecticut



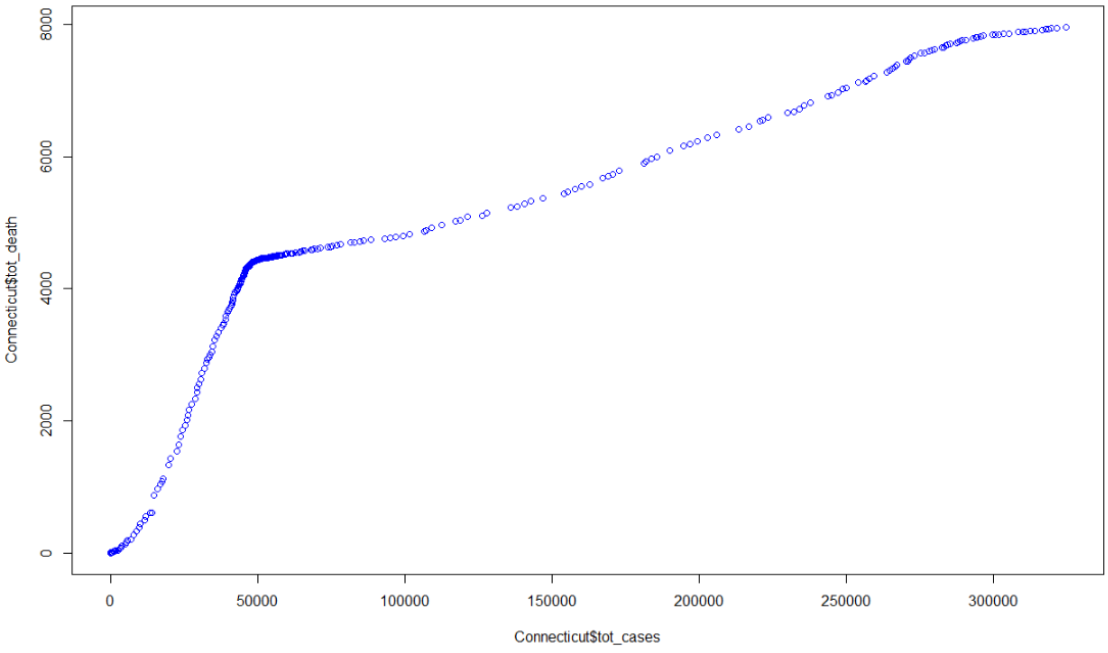

```
> describe(Connecticut$new_case)
   vars   n   mean      sd median trimmed    mad min  max range skew kurtosis    se
X1    1 447 726.11 1242.52    179  441.39 265.39 -15 8457  8472    3    11.21 58.77
```

Histogram of new deaths per day in Connecticut:




```
> describe(Connecticut$new_death)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 447 17.8 29.3      4   11.32 5.93   0 260   260    3    14.34 1.39
```

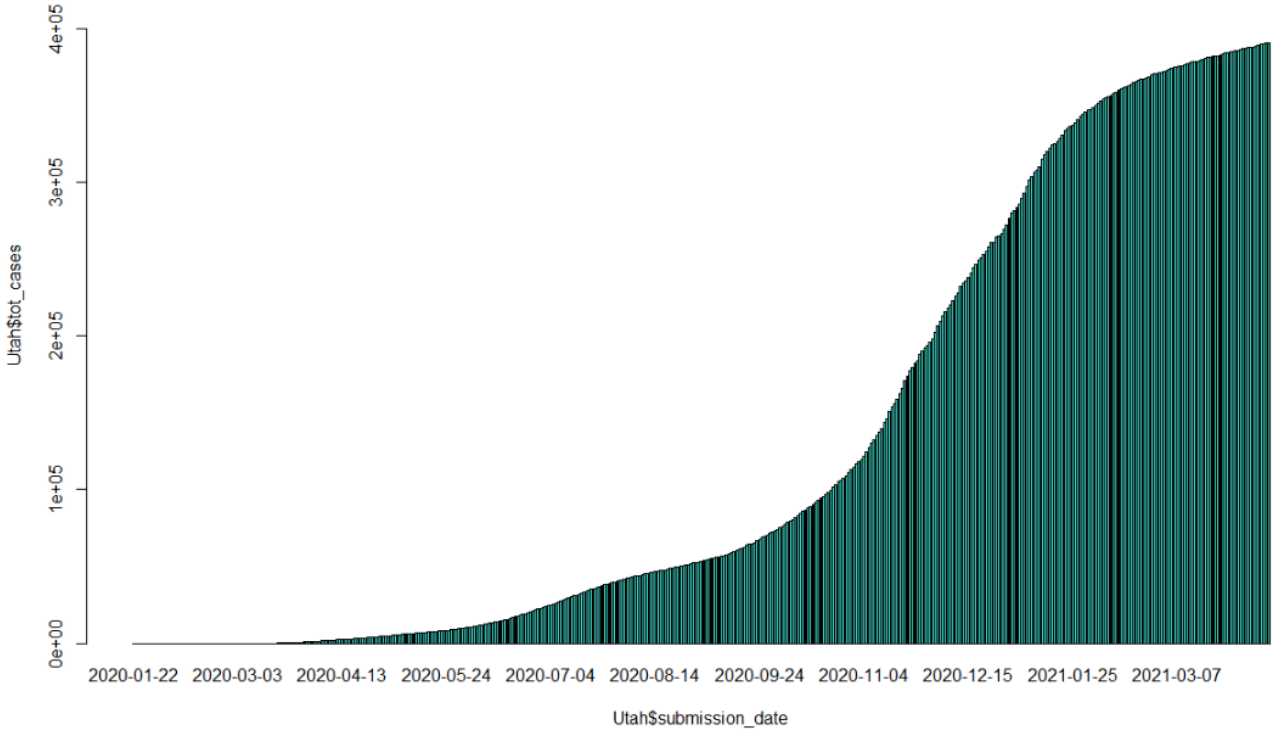
Plot of the relationship between total cases and total deaths in Connecticut:



Utah: Population 3,205,958 as of July 1, 2019 (U.S. Census Bureau) (5)

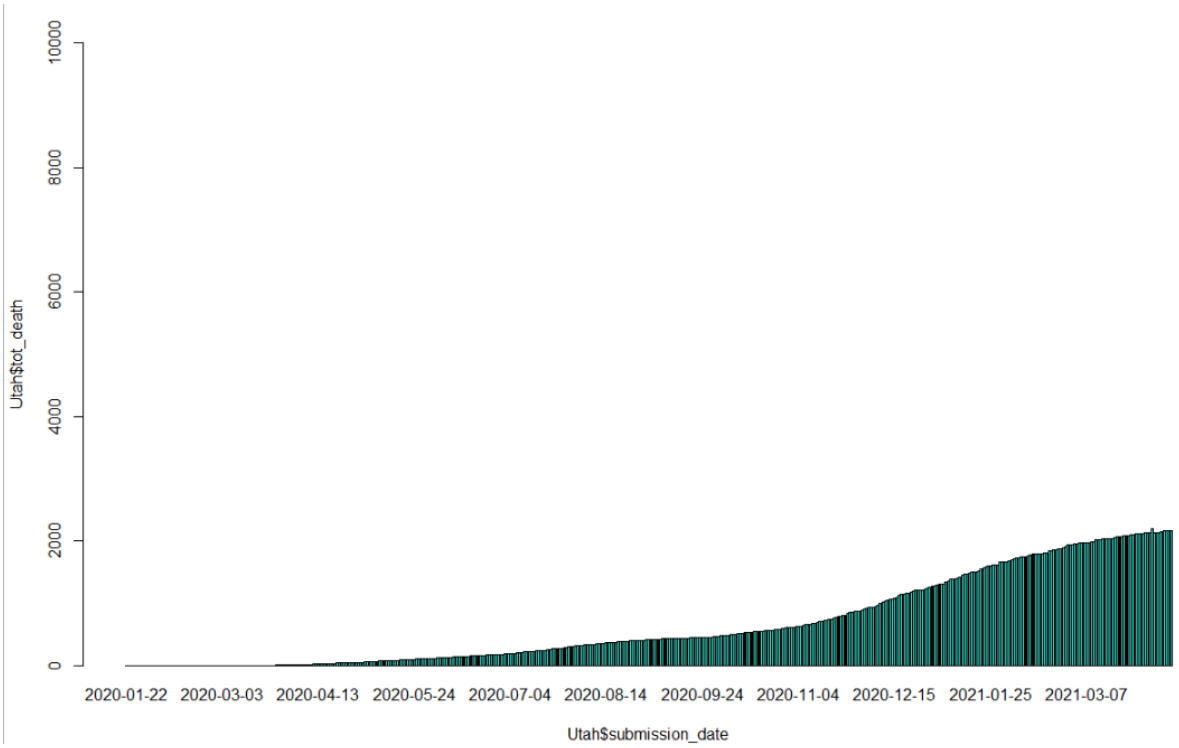
Utah total cases over time:

```
> summary(Utah$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    6612   53150  126211  253934  390289
> stat.desc(Utah$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
5.315000e+04 1.262108e+05 6.706297e+03 1.317987e+04 2.010356e+10 1.417870e+05 1.123414e+00
> describe(Utah$tot_cases)
   vars   n      mean     sd median  trimmed      mad min    max  range skew kurtosis     se
X1    1 447 126210.8 141787  53150 110406.2 78757.19   0 390289 390289  0.8    -1.01 6706.3
```

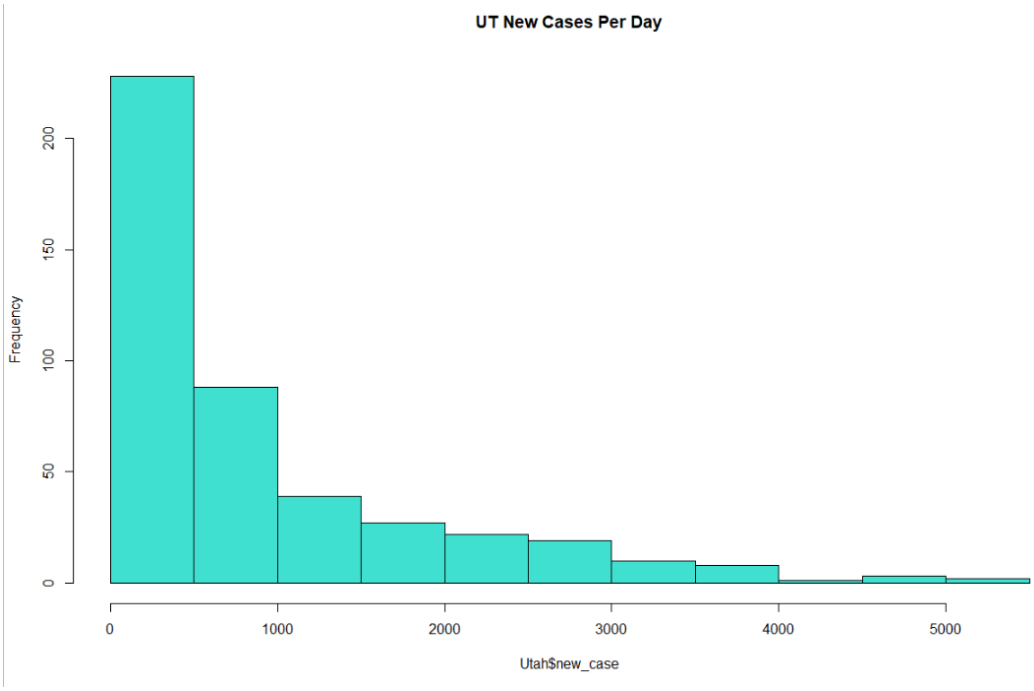


Utah total deaths over time:

```
> summary(Utah$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    75.0   408.0   663.3  1167.0  2192.0
> stat.desc(Utah$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
4.080000e+02 6.632528e+02 3.325838e+01 6.536261e+01 4.944357e+05 7.031612e+02 1.060171e+00
> describe(Utah$tot_death)
   vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 447 663.25 703.16    408  573.01 576.73   0 2192  2192  0.9    -0.63 33.26
```

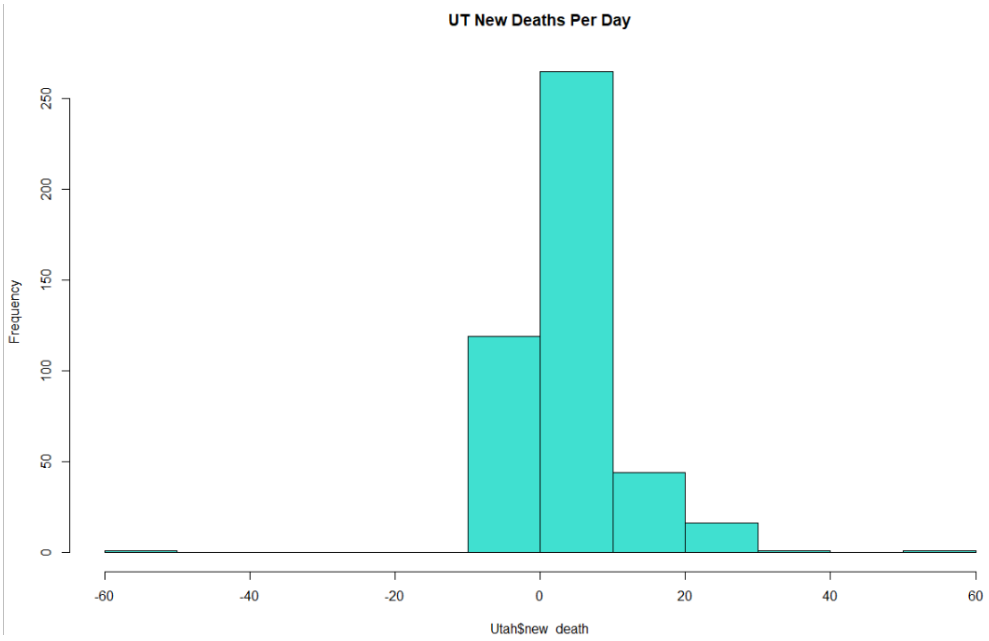


Histogram of new cases per day in Utah:



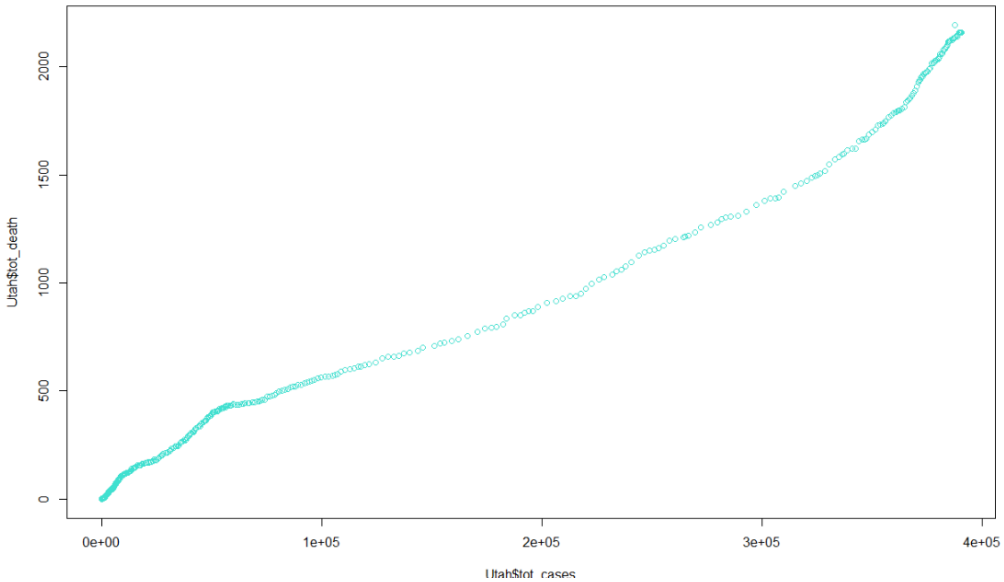

```
> describe(Utah$new_case)
   vars   n   mean      sd median trimmed    mad min  max range skew kurtosis   se
X1    1 447 873.13 1019.04    483   683.2 578.21   0 5352  5352 1.71      2.7 48.2
```

Histogram of new deaths per day in Utah:




```
> describe(Utah$new_death)
   vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
X1    1 447 4.83 7.26      3     3.7 4.45 -59  60   119 0.56    21.18 0.34
```

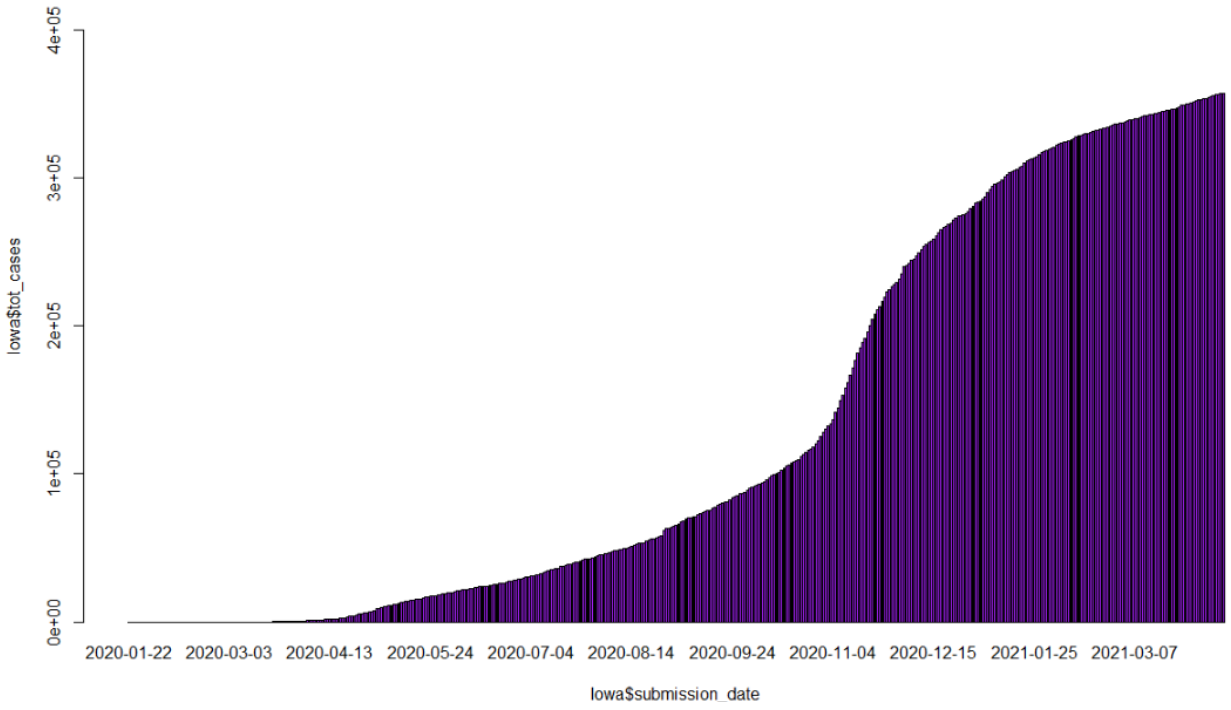
Plot of the relationship between total cases total deaths in Utah:



Iowa: Population 3,155,070 as of July 1, 2019 (U.S. Census Bureau) (5)

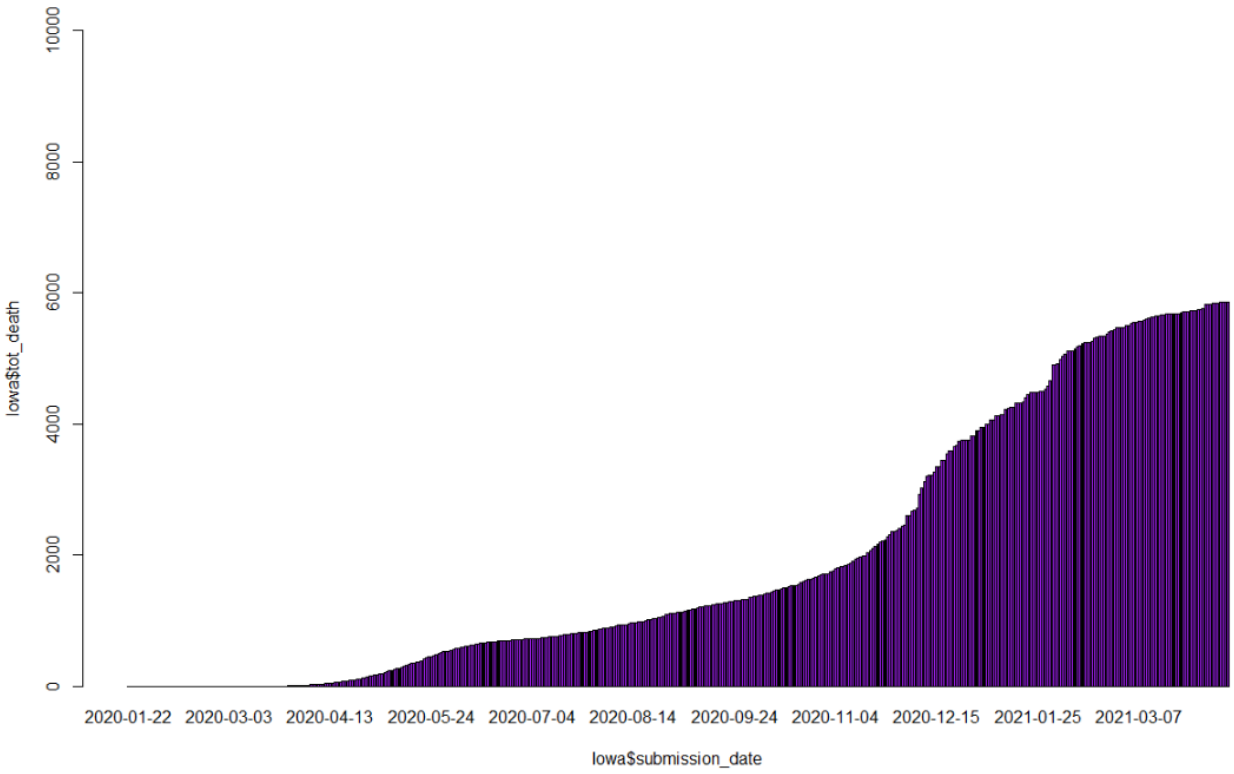
Iowa total cases over time:

```
> summary(Iowa$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   13100   65448  126777  268669  357178
> stat.desc(Iowa$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
6.544800e+04 1.267771e+05 6.205646e+03 1.219594e+04 1.721399e+10 1.312021e+05 1.034904e+00
> describe(Iowa$tot_cases)
   vars   n     mean       sd median  trimmed      mad min    max  range skew kurtosis      se
X1    1 447 126777.1 131202.1  65448 115384.9 96877.53   0 357178 357178 0.62    -1.29 6205.65
```

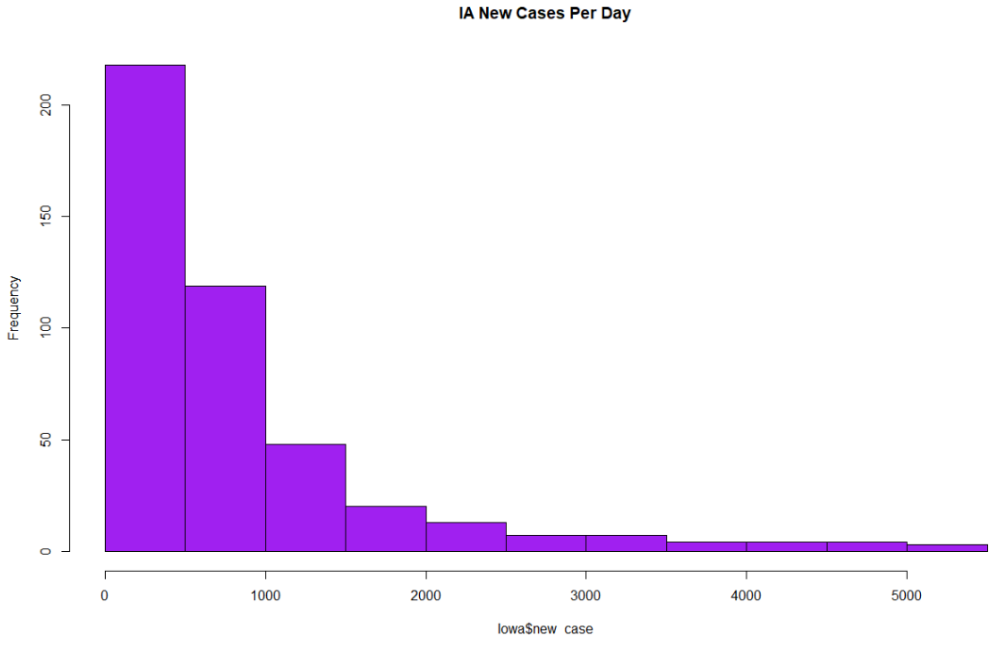


Iowa total deaths over time:

```
> summary(Iowa$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   297.5  1122.0  1920.7  3621.0  5857.0
> stat.desc(Iowa$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
1.122000e+03 1.920671e+03 9.332073e+01 1.834030e+02 3.892815e+06 1.973022e+03 1.027256e+00
> describe(Iowa$tot_death)
   vars   n    mean      sd median trimmed     mad min  max range skew kurtosis    se
X1    1 447 1920.67 1973.02   1122 1694.89 1599.73   0 5857  5857 0.85    -0.78 93.32
```

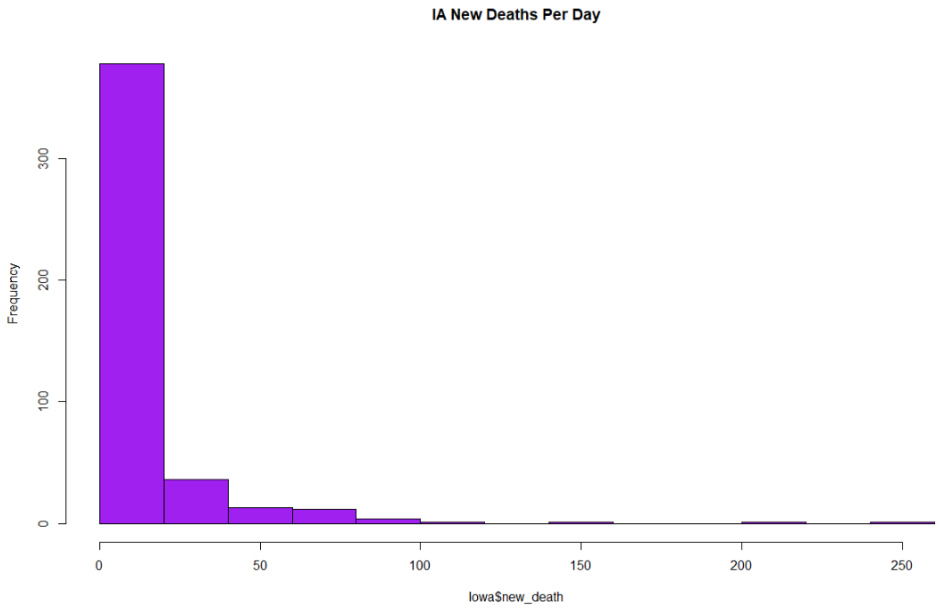


Histogram of new cases per day in Iowa:



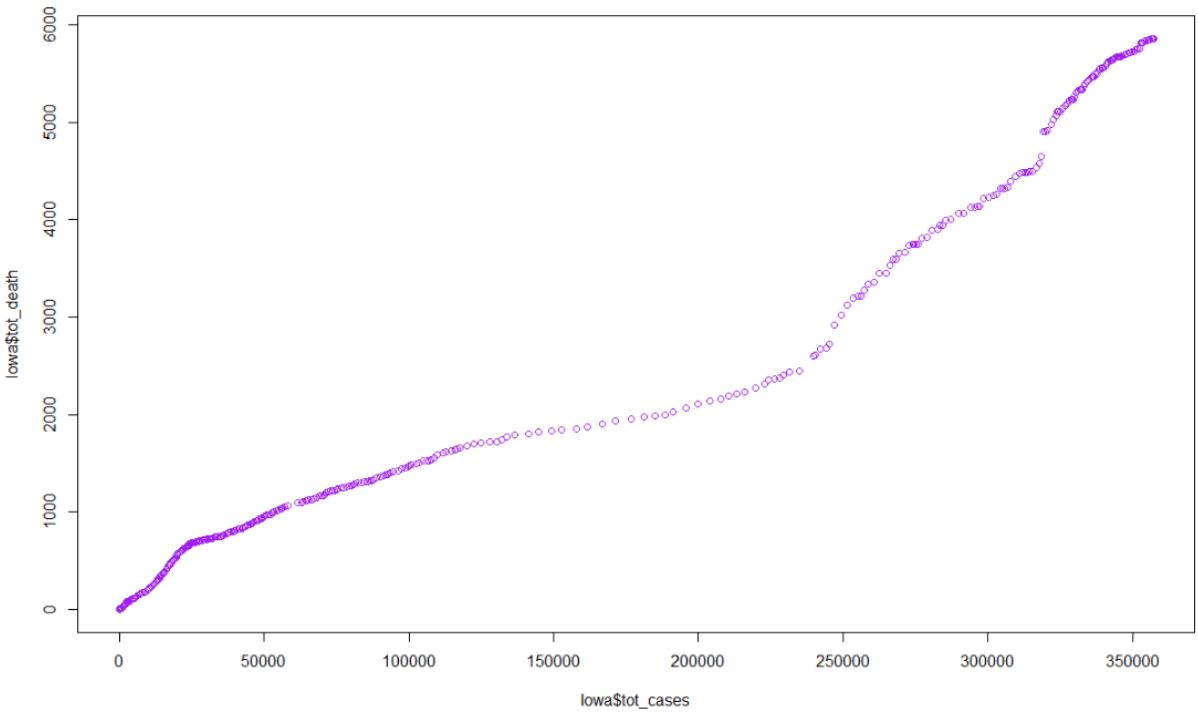

```
> describe(Iowa$new_case)
   vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 447 799.06 956.63    508  601.09 504.08   0 5452  5452 2.38     6.22 45.25
```

Histogram of new deaths per day in Iowa:




```
> describe(Iowa$new_death)
   vars   n mean    sd median trimmed mad min max range skew kurtosis   se
X1    1 447 13.1 23.07      6    8.18 8.9   0 250   250 5.07    37.67 1.09
```

Plot of the relationship between total cases and total deaths in Iowa:



Part 1 Summary:

Mean new case rates and death rates per day, included for reference:

State	Mean, New Cases per Day	Mean, New Deaths per Day
LA	1006.32	22.91
OK	993.02	14.98
KY	969.47	12.25
UT	873.13	4.83
IA	799.06	13.1
CT	726.11	17.8
OR	382.21	5.46

Based on a detailed analysis of the data, we can see that the COVID-19 pandemic affected each of the states included in this study differently over time. In Louisiana, COVID began to spread rapidly, and remained relatively deadly over the course of the pandemic. Louisiana had the highest infection rate overall, with a mean of 1006 new confirmed cases per day. Louisiana also had the highest daily death rate, with a mean of 22 fatalities per day. The plot of cases over time shows multiple surges in cases in Louisiana, especially during the holiday season, November, December, and January of 2020. This pattern of infections is repeated to various extents in each of the other states, except Oregon, where the number of total infections increased at a lower rate than in other states for the duration of the pandemic. Oregon had the lowest infection rate overall, with a mean of 382 new cases per day, and the second lowest fatality rate, with a mean 5.46 new deaths per day. Utah had the lowest fatality rate with a mean 4 new fatalities per day, however, Utah had a relatively high case rate, with a mean of 873 new cases per day.

Other notable findings:

Connecticut had a uniquely high fatality rate per covid case. CT had the second lowest infection rate, at 726 new cases, per day, and also the second highest fatality rate at 17.8 new deaths per day. COVID was deadliest in CT early in the pandemic, with the daily death rate increasing rapidly in April and May 2020

then stabilizing from July to October 2020. COVID deaths began to increase again at a lower rate in November 2020. The inverse case is the state of Utah, which had a notably low fatality rate per covid case. Utah has a moderately high infection rate of 873 new cases per day, which puts it at rank 4 out of 7 states in this study for infection rate. The infection rate in Utah remained relatively low until November of 2020, then began to increase rapidly. The death rate increased at a relatively low rate during this period.

Results and Analysis Part 2: Comparing States

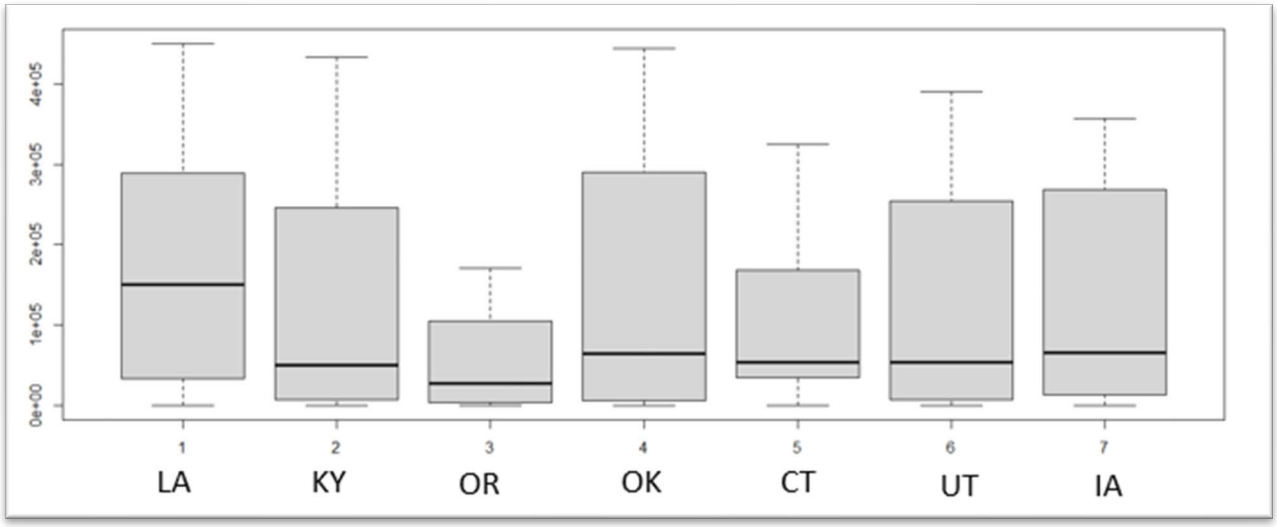
Part two of this study will consist of a detailed statistical examination of differences in COVID-19 Cases and deaths between the states selected for this survey. This will constitute the bulk of the analysis conducted within this study. Whereas part one sought to detail the narrative of how the Covid-19 Pandemic grew and spread in various parts of the country over time, Part two will analyze the cumulative differences in COVID-19 infection rate and lethality rate in various U.S. States.

Analysis of Variance:

Summary of Total cases by state: Measures of Central tendency

```
> #Central tendency cases
> summary(Covid_Cases_Over_Time_selected_states$tot_cases)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    6846   55343  125293  214227  449827
> stat.desc(Covid_Cases_Over_Time_selected_states$tot_cases,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
5.534300e+04 1.252927e+05 2.756856e+03 5.405778e+03 2.038388e+10 1.427721e+05 1.139508e+00
> describe(Covid_Cases_Over_Time_selected_states$tot_cases)
   vars    n     mean       sd median  trimmed      mad min    max  range skew kurtosis      se
X1    1 2682 125292.7 142772.1  55343 104514.9 82021.14   0 449827 449827 0.97    -0.48 2756.86
> |
```

Boxplot: Total cases by state



ANOVA: Total cases by State

Null hypothesis: The mean values of total COVID cases are the same across all of the states, or the differences are small enough to be statistically insignificant.

Alternate hypothesis: The differences in the mean values of COVID cases are different enough to be statistically significant.

Analysis of variance summary: The F value of the ANOVA between total cases in the selected States was calculated to be 36.96. The probability of this F value is less than 2e-16. In this case we can reject the null hypothesis.

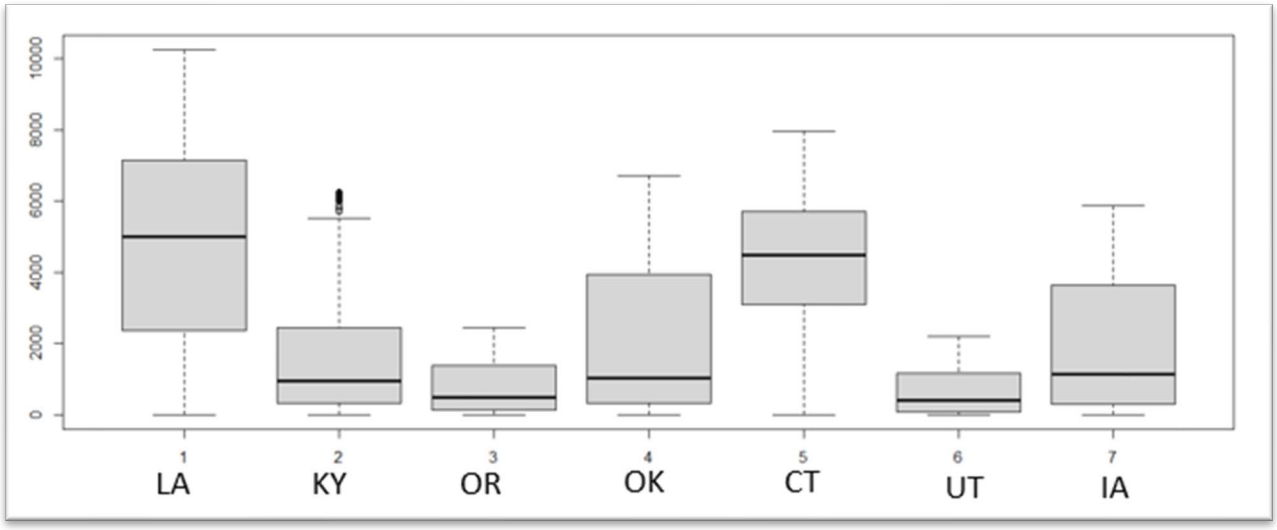
```
> summary(ANOVA_COVID_cases_selected_States)
             Df    Sum Sq   Mean Sq F value Pr(>F)
state         5 3.530e+12 7.061e+11   36.96 <2e-16 ***
Residuals  2676 5.112e+13 1.910e+10
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

<u>Summary of Total deaths by state:</u>

Summary of total deaths by state: Measures of Central tendency

```
> #Central tendency deaths
> summary(Covid_Cases_Over_Time_selected_states$tot_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   188.0   908.5  1981.2  2832.0 10241.0
> stat.desc(Covid_Cases_Over_Time_selected_states$tot_death,basic = F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
9.085000e+02 1.981231e+03 4.687733e+01 9.191938e+01 5.893653e+06 2.427685e+03 1.225341e+00
> describe(Covid_Cases_Over_Time_selected_states$tot_death)
   vars    n    mean      sd median trimmed     mad min   max range skew kurtosis    se
X1    1 2682 1981.23 2427.68  908.5 1539.46 1335.08   0 10241 10241 1.48     1.44 46.88
>
```

Boxplot: Total deaths by state



ANOVA: Total deaths by state

Null hypothesis: The mean values of total COVID deaths are the same across all of the states, or the differences are small enough to be statistically insignificant.

Alternate hypothesis: The differences in the mean values of COVID deaths are different enough to be statistically significant.

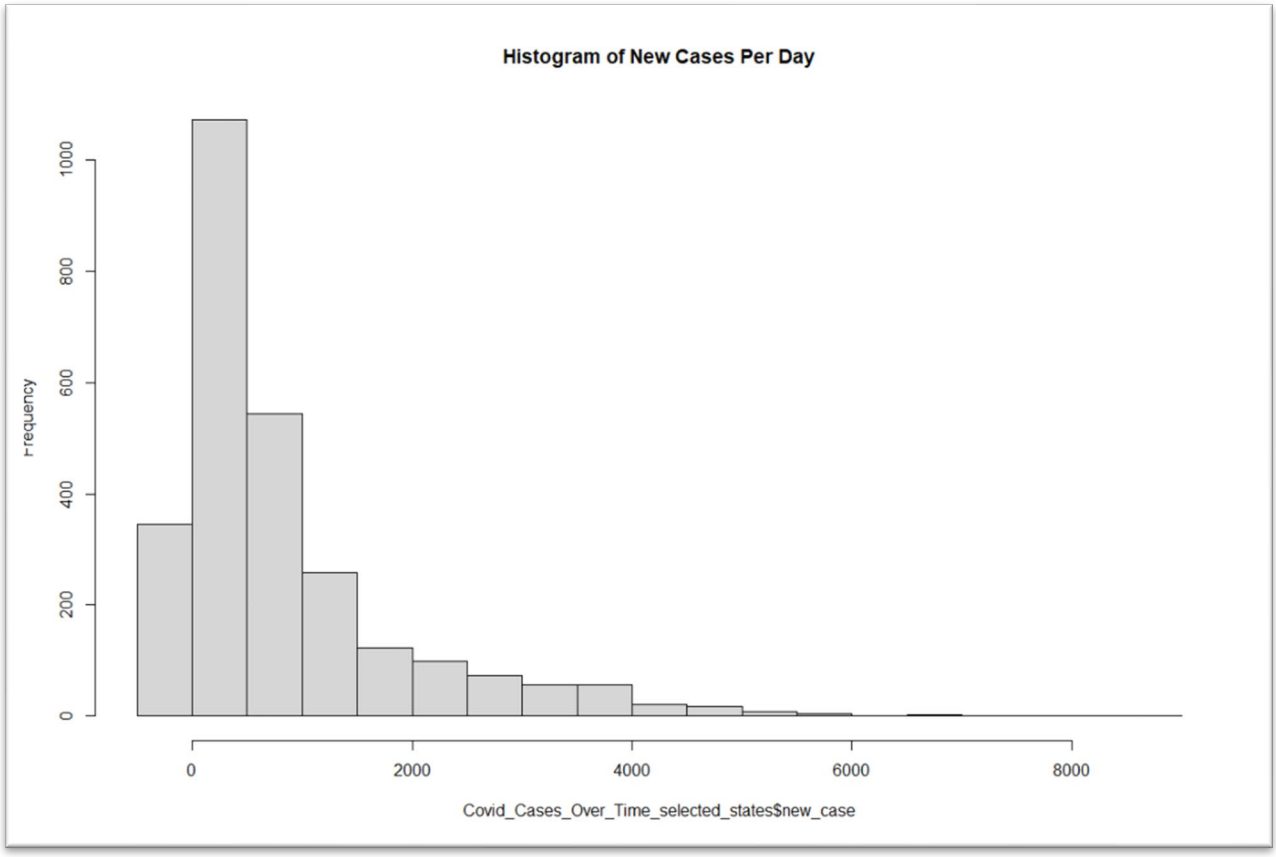
Analysis of variance summary: The F value of the analysis of variance between total COVID deaths in the selected states came to 244.8. The probability of this F value is less than 2e-16. We can reject the null hypothesis in this case.

```
> summary(ANOVA_COVID_DEATHS_Selected_States)
             Df    Sum Sq   Mean Sq F value Pr(>F)
state         5 4.959e+09 991880621   244.8 <2e-16 ***
Residuals  2676 1.084e+10   4051376
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis testing, One sample T-test: New cases per day

This T test will evaluate the rate of new cases per day in all of the states selected for this study. The mean value for new cases per day in each state is 837.5. Because the histogram is left skewed, I will examine values below the mean.

```
> summary(Covid_Cases_Over_Time_selected_states$new_case)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -293.0   118.0   454.0   837.2  1086.0  8709.0
> |
```



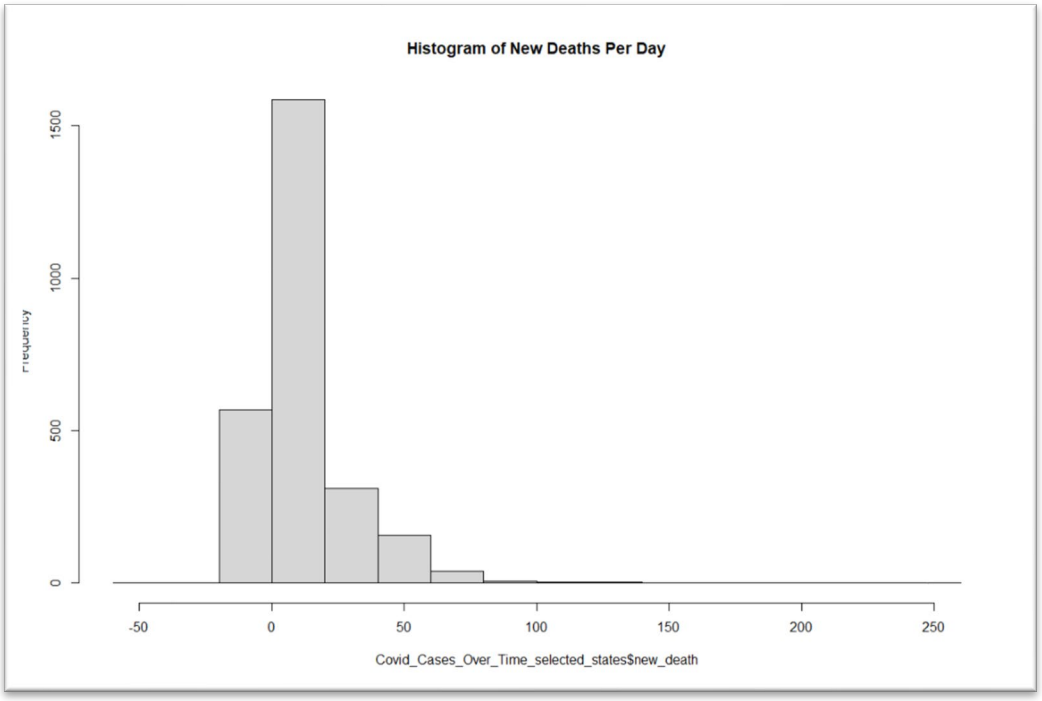
Results: The T value for this result is 0.00013833, the P value is 0.5001. There is a 95% confidence interval for all negative values and positive values through 870.9.

```
> t.test(Covid_Cases_Over_Time_selected_states$new_case,mu=837.2,alternative = "less")

        One Sample t-test

data:  Covid_Cases_Over_Time_selected_states$new_case
t = 0.00013833, df = 2681, p-value = 0.5001
alternative hypothesis: true mean is less than 837.2
95 percent confidence interval:
     -Inf 870.9087
sample estimates:
mean of x
 837.2028
```

Hypothesis Testing, One sample T-test: New deaths per day

Like the previous test, This T test will also evaluate all of the states in this study but will evaluate the rate of new deaths per day. The mean number of new deaths per day in each state is 12.6. because the histogram is right skewed, I will test values less than the mean.

```
> summary(Covid_Cases_Over_Time_selected_states$new_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -59.00    1.00    6.00   12.26   16.00  250.00
```



Results: The T value returned by this test is -0.71666. The P value for this result is 0.2368. There is a 95% confidence interval for all negative values, and positive through 12.81699.

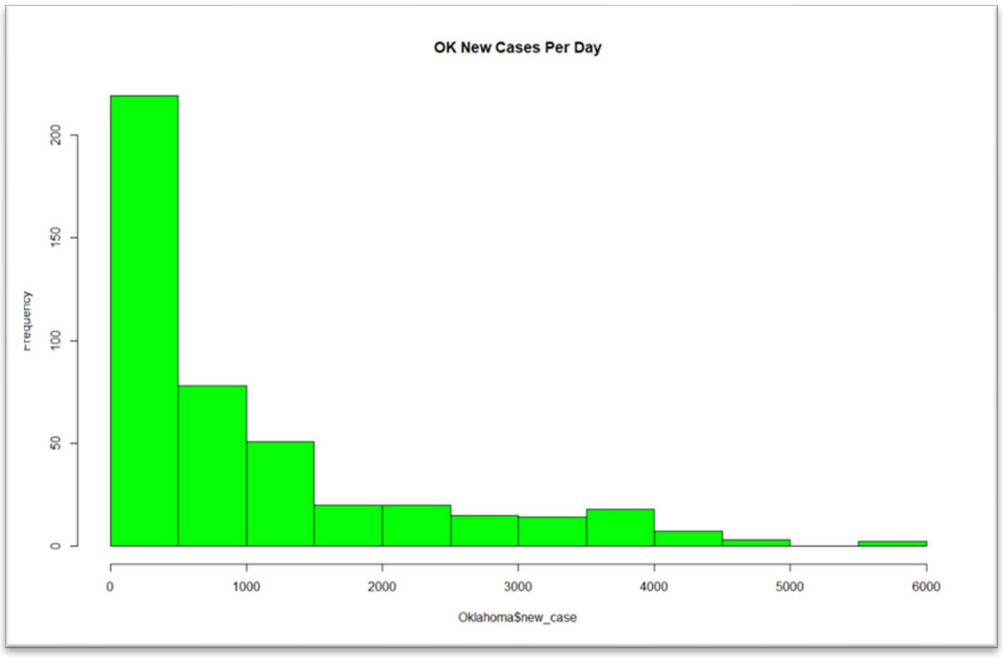
```
> t.test(Covid_Cases_Over_Time_selected_states$new_death,mu=12.5,alternative = "less")

        One Sample t-test

data:  Covid_Cases_Over_Time_selected_states$new_death
t = -0.71666, df = 2681, p-value = 0.2368
alternative hypothesis: true mean is less than 12.5
95 percent confidence interval:
     -Inf 12.81699
sample estimates:
mean of x
 12.25541
```

Hypothesis testing, Two Sample T-Test: New Cases per day OK and LA

This test will evaluate any differences in new daily cases between the two states with the highest mean infection rates in this study, Louisiana and Oklahoma.

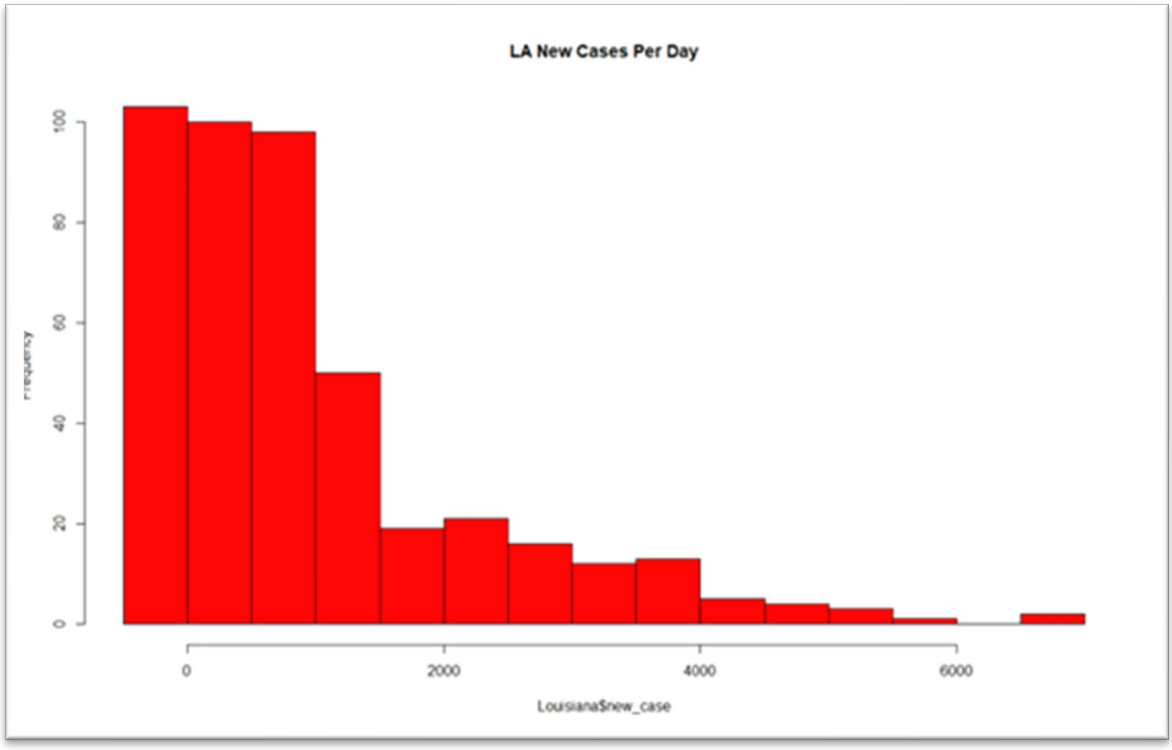
Histogram of daily new cases, Oklahoma:

```
> summary(Oklahoma$new_case)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   105.5   535.0   993.0  1342.5  5931.0
```



Histogram of daily new cases, Louisiana:

```
> summary(Louisiana$new_case)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -119      93     558    1006    1304    6876
```



Results: For my first two-sample test I compared new daily cases in Oklahoma and Louisiana. The mean of new cases in OK is 993.02. The mean of new cases in LA is 1006.32. The T value is-0.16469, with a p-value of o.8692.

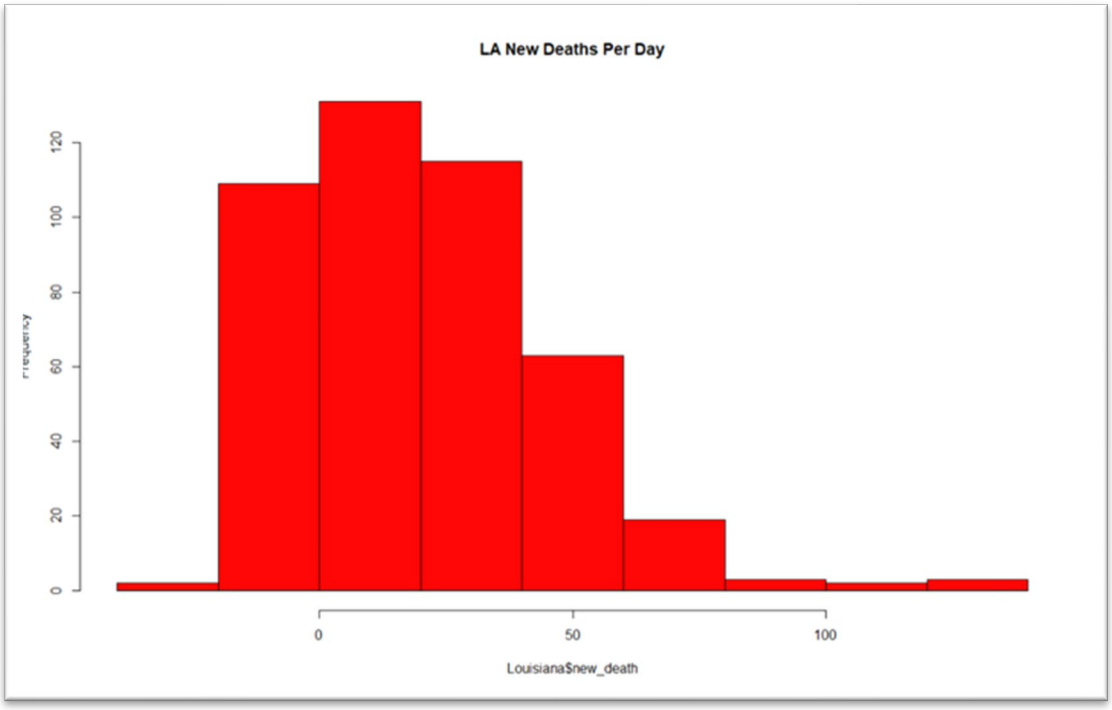
```
        Welch Two Sample t-test

data:  Oklahoma$new_case and Louisiana$new_case
t = -0.16469, df = 890.54, p-value = 0.8692
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -171.7910  145.1915
sample estimates:
mean of x mean of y
 993.0246 1006.3244
```

Hypothesis testing, Two Sample T-Test: New Deaths per day LA and CT

My second Two sample T-Test will evaluate the daily death rates in the two states with the highest mean COVID-19 fatality rates in this study, Louisiana and Connecticut.

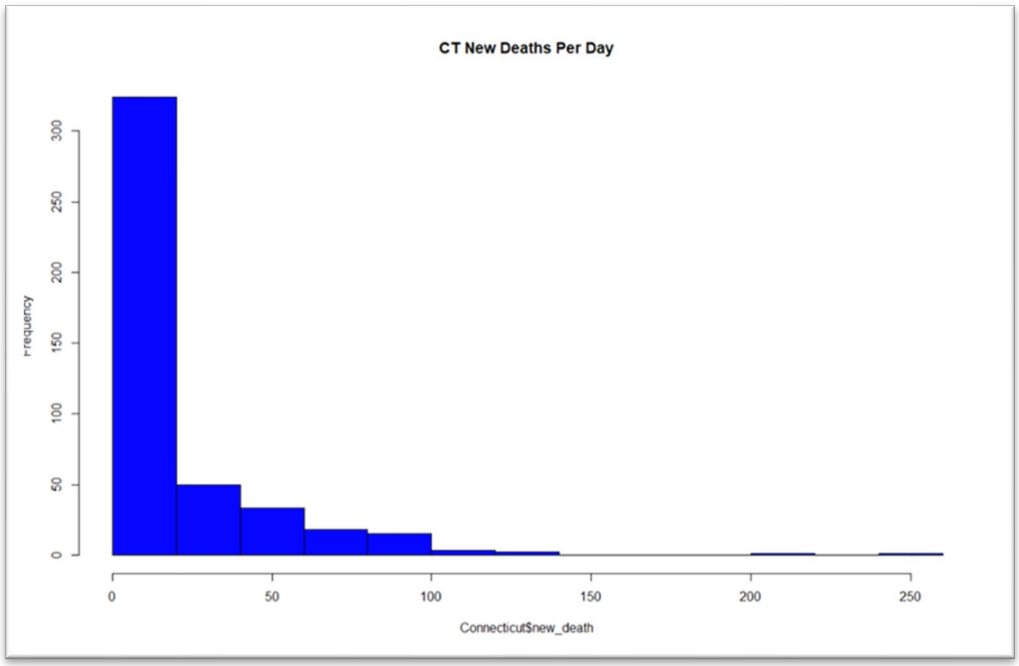
Histogram of new deaths per day, Louisiana

```
> summary(Louisiana$new_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -21.00    1.00   19.00   22.91   34.00  129.00
```



Histogram of new deaths per day, Connecticut:

```
> summary(Connecticut$new_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0     4.0    17.8    23.0   260.0
```



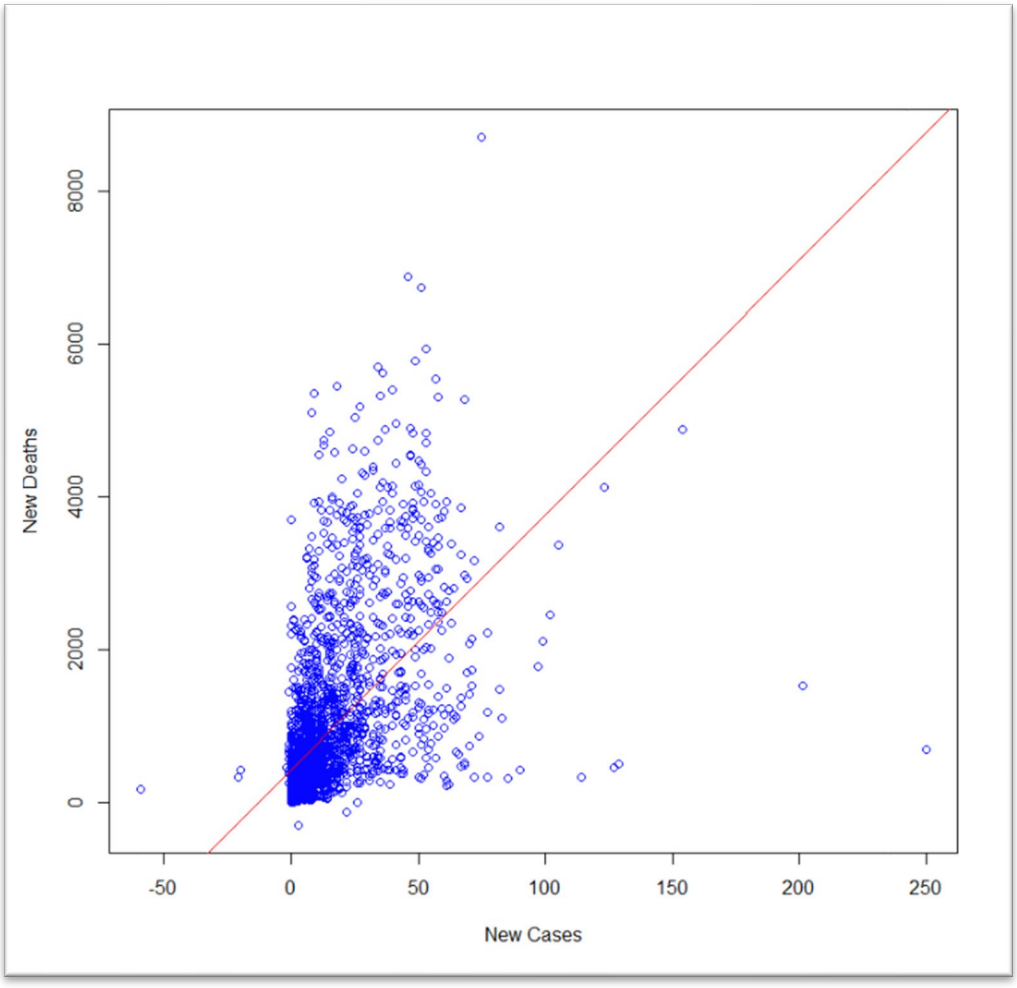
Results: For my first two-sample test I compared new daily cases in Oklahoma and Louisiana. The mean of new deaths in LA is 22.91 The Mean of New Deaths in in CT is 17.80. The T value is 2.91, with a P-value of 0.003678.

```
> t.test(Louisiana$new_death, Connecticut$new_death)

        Welch Two Sample t-test

data:  Louisiana$new_death and Connecticut$new_death
t = 2.9127, df = 840.14, p-value = 0.003678
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.666352 8.552888
sample estimates:
mean of x mean of y
 22.91051  17.80089
```

Scatterplot with best fit line:

New Cases Vs New Deaths

The Best fit line shows that there is a strong positive correlation between the rate of new COVID-19 cases per day and the rate of new COVID-19 deaths per day in all of the states included in this study.



Correlation Coefficient: The correlation coefficient between new cases and new deaths in the selected states is **0.556**

```
> cor.test(Covid_Cases_Over_Time_selected_states$new_case,Covid_Cases_Over_Time_selected_states$new_death)

        Pearson's product-moment correlation

data:  Covid_Cases_Over_Time_selected_states$new_case and Covid_Cases_Over_Time_selected_states$new_death
t = 34.666, df = 2680, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5297121 0.5819981
sample estimates:
      cor
0.5564057

> Covid_new_cases_new_deaths_reg <-lm(Covid_Cases_Over_Time_selected_states$new_case~Covid_Cases_Over_Time_selected_states$new_death)
> Covid_new_cases_new_deaths_reg

Call:
lm(formula = Covid_Cases_Over_Time_selected_states$new_case ~
    Covid_Cases_Over_Time_selected_states$new_death)

Coefficients:
                                  (Intercept)  Covid_Cases_Over_Time_selected_states$new_death
                                        427.9                                            33.4
```

Regression Summary: the r^2 value for this data is 0.3096, as detailed in my results, which have been transcribed below.

```
> summary(Covid_new_cases_new_deaths_reg)

Call:
lm(formula = Covid_Cases_Over_Time_selected_states$new_case ~
    Covid_Cases_Over_Time_selected_states$new_death)

Residuals:
    Min      1Q  Median      3Q     Max
-8085.8  -427.9  -253.0   169.2  5776.4

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  427.9301    20.7173   20.66   <2e-16 ***
Covid_Cases_Over_Time_selected_states$new_death  33.3953     0.9633   34.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 881.6 on 2680 degrees of freedom
Multiple R-squared:  0.3096,    Adjusted R-squared:  0.3093
F-statistic:  1202 on 1 and 2680 DF,  p-value: < 2.2e-16
```

Multiple Regression:

Total cases over time

The first linear regression model I have included is based on determining the feasibility of creating a predictive model of COVID totals and new cases based on submission date.

```
> Covid_Cases_Over_Time_LM

Call:
lm(formula = submission_date ~ tot_cases + tot_death + new_case +
    new_death, data = Covid_Cases_Over_Time_selected_states)

Coefficients:
(Intercept)    tot_cases    tot_death     new_case    new_death
  1.590e+09    7.034e+01   -3.339e+02    1.453e+03   -4.137e+04

> ANOVA_Covid_Cases_Over_Time_LM <-aov(Covid_Cases_Over_Time_LM)
> summary(ANOVA_Covid_Cases_Over_Time_LM)
              Df    Sum Sq   Mean Sq F value   Pr(>F)
tot_cases      1 2.528e+17 2.528e+17 8976.75  < 2e-16 ***
tot_death      1 1.229e+15 1.229e+15   43.64 4.75e-11 ***
new_case       1 3.060e+15 3.060e+15  108.68  < 2e-16 ***
new_death      1 8.877e+14 8.877e+14   31.52 2.18e-08 ***
Residuals   2677 7.539e+16 2.816e+13
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> coef(Covid_Cases_Over_Time_LM)
  (Intercept)     tot_cases     tot_death      new_case     new_death
 1.590058e+09  7.033810e+01 -3.339487e+02  1.452649e+03 -4.136616e+04
>
```

The ANOVA has determined that all of these variables in this dataset are statistically significant when compared to submission date. This makes sense based on the pattern of overall growth in cases and death rates which we observed in part one of this study.

Forwards stepwise model: Total Cases

This stepwise model uses total cases as the comparative value to which the other variables are compared.

```
> ANOVA_TOTAL_CASES_STEPWISE
Call:
   aov(formula = TOTAL_CASES_FORWARD_STEPWISE)

Terms:
                   tot_death     new_case    new_death    Residuals
Sum of Squares  3.893778e+13 1.499556e+12 8.340010e+10 1.412844e+13
Deg. of Freedom            1            1            1         2678

Residual standard error: 72634.32
Estimated effects may be unbalanced
> summary(ANOVA_TOTAL_CASES_STEPWISE)
              Df    Sum Sq   Mean Sq F value  Pr(>F)
tot_death      1 3.894e+13 3.894e+13 7380.53 < 2e-16 ***
new_case       1 1.500e+12 1.500e+12  284.24 < 2e-16 ***
new_death      1 8.340e+10 8.340e+10   15.81 7.2e-05 ***
Residuals   2678 1.413e+13 5.276e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

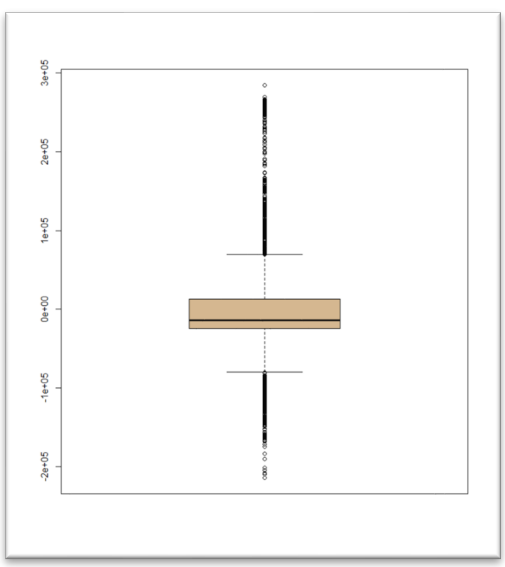
```
> coef(ANOVA_TOTAL_CASES_STEPWISE)
(Intercept)   tot_death    new_case   new_death
14247.11574    47.29294    26.57290  -399.78692
> confint(ANOVA_TOTAL_CASES_STEPWISE)
                2.5 %      97.5 %
(Intercept) 10355.84888 18138.38259
tot_death      46.03964    48.54625
new_case       23.43709    29.70871
new_death    -596.95272  -202.62112
```

The ANOVA values of this model demonstrate that the total number of cases are statistically significant to the total number of deaths, as well as the rate of new cases and new deaths per day. This pattern make sense based on our previous model correlating submission date of COVID reports with greater total numbers of infections and deaths.

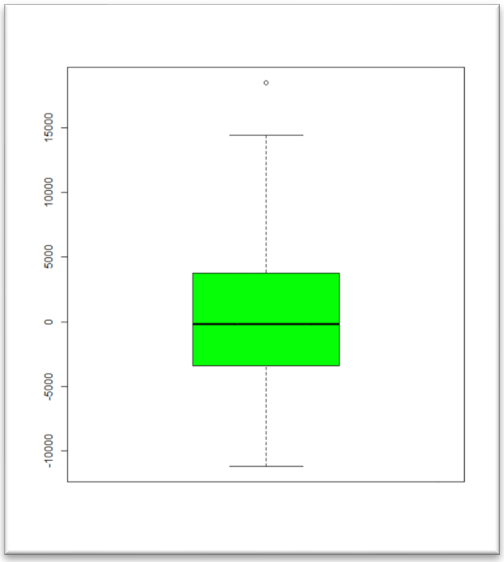
Boxplot, Total cases stepwise model:

The boxplot of these values shows a large number of outliers both above and below the mean value. This pattern was not observed when running this pattern for individual states (as observed in the Oklahoma Total Cases Stepwise model below.) This discrepancy may be accounted for by the differences in case and death values among states during the pandemic.

Boxplot, Total cases stepwise (all states):



Boxplot, Total cases Stepwise (Oklahoma) included for reference.



Summary of forward stepwise model: The Multiple R-squared value of the model shows that the total case rate explains 74.15% of the overall variability in total deaths, new deaths, and new cases.

```
> summary(TOTAL_CASES_FORWARD_STEPWISE)

Call:
lm(formula = tot_cases ~ tot_death + new_case + new_death, data = Covid_Cases_Over_Time_selected_states)

Residuals:
    Min      1Q  Median      3Q     Max
-214074  -24087  -14247   13295  284274

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14247.1157  1984.4794   7.179 9.04e-13 ***
tot_death      47.2929     0.6392  73.992  < 2e-16 ***
new_case       26.5729     1.5992  16.616  < 2e-16 ***
new_death    -399.7869   100.5512  -3.976 7.20e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72630 on 2678 degrees of freedom
Multiple R-squared:  0.7415,    Adjusted R-squared:  0.7412
F-statistic:  2560 on 3 and 2678 DF,  p-value: < 2.2e-16
```

CHI Squared Tests:

These Chi squared tests will examine new cases and new deaths per day among the states selected for this study. Cases and deaths per day have been broken down into their respective quartile ranges for analytical purposes.

New Cases per day:

Null hypothesis: The number of new COVID-19 cases per day is the same or similar enough across the states for differences to be statistically insignificant.

Alternate hypothesis: The number of new COVID-19 cases per day is different enough across the states to be statistically significant.

Summary of new cases per day: the following is a summary of the quartile ranges of new cases per day across the selected states.

```
> summary(Covid_Cases_Over_Time_selected_states$new_case)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -293.0   118.0   454.0   837.2  1086.0  8709.0
```

Contingency table: New cases per day in selected states, listed by quartile

```
> Chi_table_New_Case

   1st Qu. 2nd Qu. 3rd  Qu. 4th Qu. Max Min
IA      38      113      151       97   0  48
KY      42       99      114      125   1  66
LA      10       79      114      141   0 103
OK      98       86       93      142   0  28
OR      97      175       82       42   0  51
UT      41      118      116      122   0  50
>
```

```
> round(prop.table(Chi_table_New_Case,1),2)*100

   1st Qu. 2nd Qu. 3rd  Qu. 4th Qu. Max Min
IA       9       25        34       22   0  11
KY       9       22        26       28   0  15
LA       2       18        26       32   0  23
OK      22       19        21       32   0   6
OR      22       39        18        9   0  11
UT       9       26        26       27   0  11
>
```

Chi-squared test: The Chi squared test for this dataset returned an X^2 value of 318 with 25 degrees of freedom. The P value in this case is sufficiently low to reject the null hypothesis.

```
> CHi_Test_New_Case

        Pearson's Chi-squared test

data:  Chi_table_New_Case
X-squared = 318.56, df = 25, p-value < 2.2e-16
```

Expected Values: The following are the expected values in each state if the null hypothesis were accurate. These values are sufficiently different from the observed values to reject the null hypothesis.

```
> CHi_Test_New_Case$expected

      1st Qu.  2nd Qu. 3rd  Qu. 4th Qu.       Max      Min
IA 54.33333 111.6667 111.6667   111.5 0.1666667 57.66667
KY 54.33333 111.6667 111.6667   111.5 0.1666667 57.66667
LA 54.33333 111.6667 111.6667   111.5 0.1666667 57.66667
OK 54.33333 111.6667 111.6667   111.5 0.1666667 57.66667
OR 54.33333 111.6667 111.6667   111.5 0.1666667 57.66667
UT 54.33333 111.6667 111.6667   111.5 0.1666667 57.66667
```

Analysis: The number of new cases per day during this time period ranged from a minimum value of 0 at the start of the pandemic, to a maximum value of 8709 new confirmed cases in one day. The number of new cases recorded each day seems to be correlated with the state submitting records to the CDC. The maximum value of 8709 new cases in one day was recorded in Louisiana, which was the hardest hit by COVID-19 of all the states in this study. Louisiana and Oklahoma both recorded 32% of their daily covid records in the 4th quartile, however Oklahoma had a greater percentage of its records in the lower quartiles than Louisiana. Both Oklahoma and Oregon had 22% of their daily records in the first quartile, however the case rate was lower in Oregon overall, with the remaining entries skewed towards the lower quartiles, whereas Oklahoma's daily case records were relatively balanced across all four quartiles, with a larger concentration in the fourth quartile, as stated previously.

New Deaths per day:

Null hypothesis: The rate of new deaths per day is the same or similar enough across the selected states for the differences to be statistically insignificant.

Alternate hypothesis: The number of new COVID-19 deaths per day is different enough across the states to be statistically significant.

Summary of New deaths per day: the following is a summary of the quartile ranges of new deaths per day across the selected states.

```
> #New Death Chi
> summary(Covid_Cases_Over_Time_selected_states$new_death)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -59.00    1.00    6.00   12.26   16.00  250.00
> |
```

Contingency Table: New deaths per day in the selected states, listed by quartile

```
> Chi_table_New_Death

   1st Qu. 2nd Qu. 3rd Qu. 4th Qu. Max Min
IA     114     113     130      89   1   0
KY      94      97     140     116   0   0
LA     112      13      72     248   0   2
OK      87     101     129     130   0   0
OR     163     180      65      39   0   0
UT     153     164      90      29   0  11
```

```
> round(prop.table(Chi_table_New_Death,1),2)*100

   1st Qu. 2nd Qu. 3rd Qu. 4th Qu. Max Min
IA      26      25      29      20   0   0
KY      21      22      31      26   0   0
LA      25       3      16      55   0   0
OK      19      23      29      29   0   0
OR      36      40      15       9   0   0
UT      34      37      20       6   0   2
```

Chi-squared test: The Chi squared test for this dataset returned an X^2 value of 588 with 25 degrees of freedom. The P value in this case is sufficiently low to reject the null hypothesis.

```
> Chi_Test_New_Death

        Pearson's Chi-squared test

data:  Chi_table_New_Death
X-squared = 588.05, df = 25, p-value < 2.2e-16
```

Expected values: The following are the expected values in each state if the null hypothesis were accurate, these values are sufficiently different from the observed values to reject the null hypothesis.

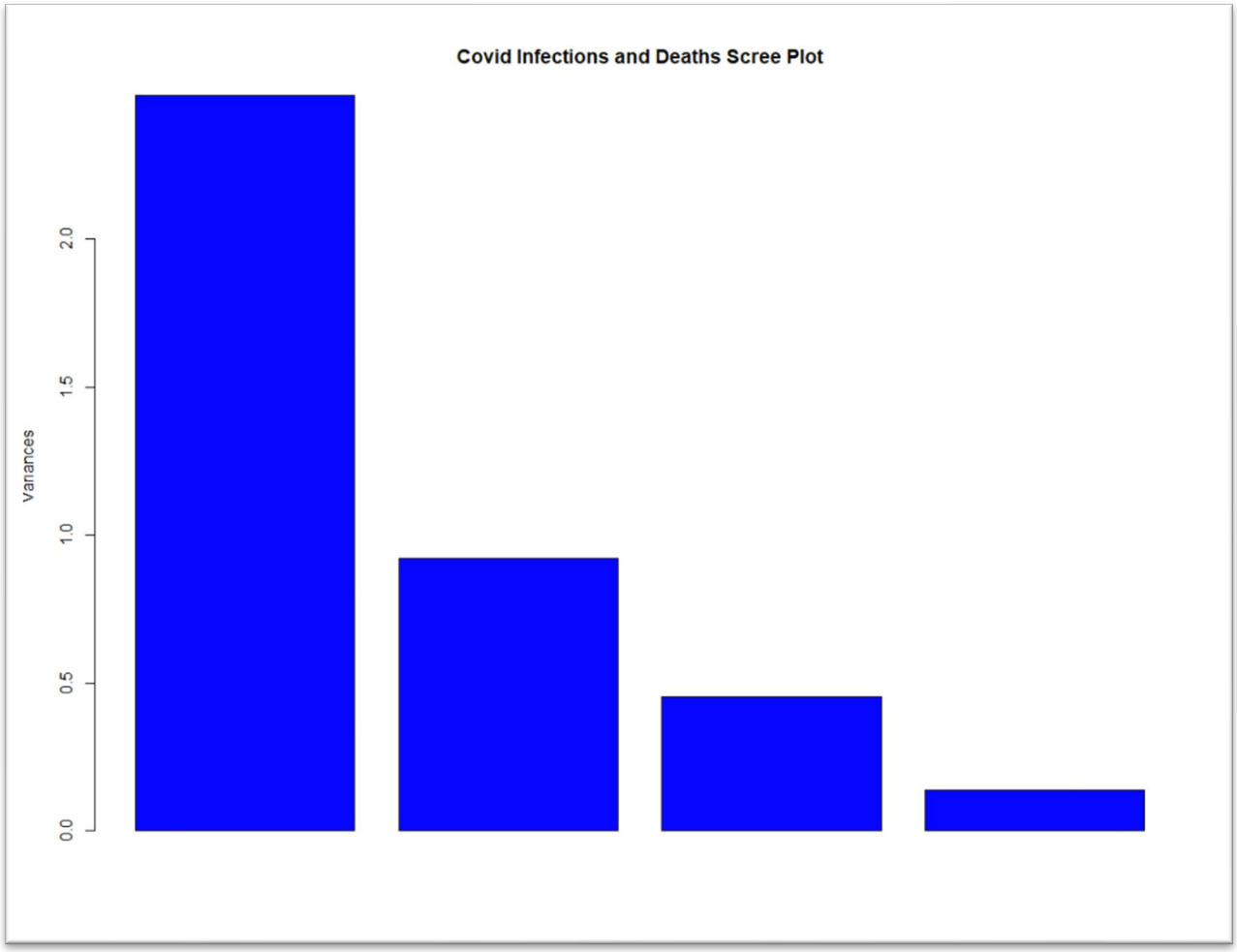
```
> Chi_Test_New_Death$expected

     1st Qu.  2nd Qu.  3rd Qu. 4th Qu.       Max      Min
IA     120.5 111.3333 104.3333   108.5 0.1666667 2.166667
KY     120.5 111.3333 104.3333   108.5 0.1666667 2.166667
LA     120.5 111.3333 104.3333   108.5 0.1666667 2.166667
OK     120.5 111.3333 104.3333   108.5 0.1666667 2.166667
OR     120.5 111.3333 104.3333   108.5 0.1666667 2.166667
UT     120.5 111.3333 104.3333   108.5 0.1666667 2.166667
```

Analysis: The number of new COVID deaths per day ranged from a minimum value of 0 to a maximum value of 250. The highest number of COVID deaths in a single day was recorded in Iowa on January 31st 2021. Louisiana had the highest percentage of its daily records in the 4th Quartile, with 55% of it's daily record falling within the highest quartile. Utah had the 34% of its daily records in the lowest quartile, with the remaining entries skewed towards the lower quartiles.

Principal Components analysis:

This Principal Components analysis examines the relationship between Total Cases, Total Deaths, New Cases, and New Deaths in each of the selected states throughout the course of the pandemic.

Scree Plot: COVID Total Infections, Total deaths, New Infection per day, and new deaths per day.



PCA Summary: Based on the Scree Plot for Covid Statistics, we can see the variance between the total number of infections and deaths the selected states, and the number of new infections and deaths per day. The Eigenvalues of these variables are 1.5759, 0.9599, 0.6744, and 0.37439. These variables are highly correlated with one another, and the daily infection and death rates serve as the primary drivers of the overall totals. These correlations are not uniform in every state, some have greater or lower fatality rates in comparison to their infection rates, and vice versa. However, this principal components analysis will serve to give a generalized idea of the degree of variance between these variables overall.

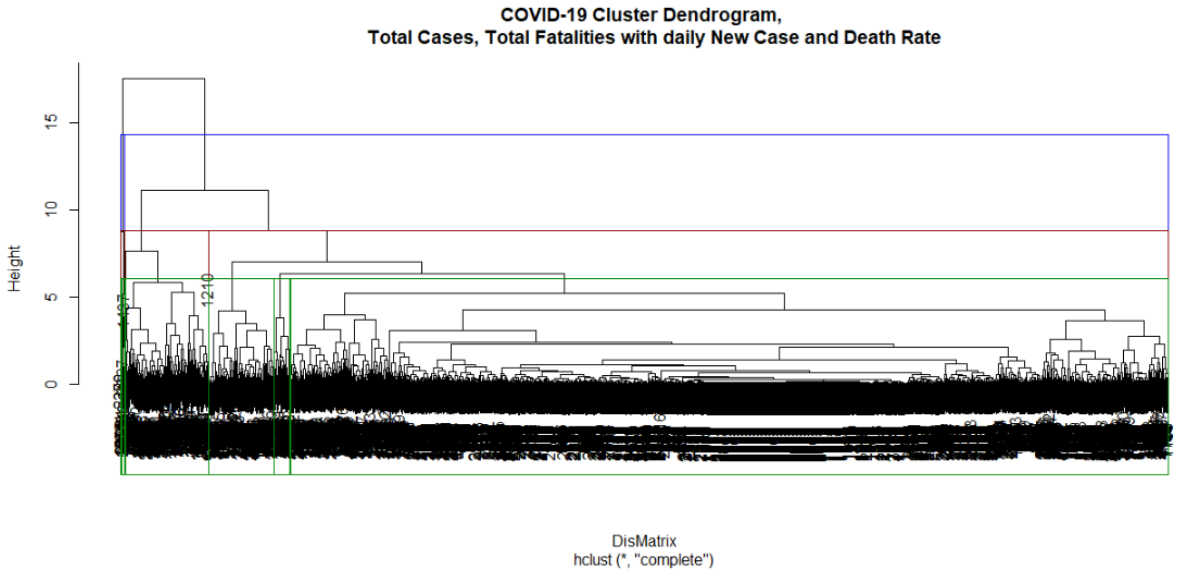
Cluster Analysis/ K-means clustering: The cluster analysis for these variables contain a large number of objects. There are 2682 individual records within this study. These consist of all daily COVID-19 record submissions to the CDC by each of the seven states included in the study.

```
> summary(Covid_PCA_2)
Importance of components:
                          PC1    PC2    PC3     PC4
Standard deviation     1.5759 0.9599 0.6744 0.37439
Proportion of Variance 0.6209 0.2304 0.1137 0.03504
Cumulative Proportion  0.6209 0.8513 0.9650 1.00000
```

Variability is weighted highly for all of the measures in the first column. New Cases and New deaths are weighted highly in the second column, and total cases and new deaths are weighted in the third column.

```
> Covid_PCA_2
Standard deviations (1, .., p=4):
[1] 1.5759458 0.9599296 0.6743591 0.3743929

Rotation (n x k) = (4 x 4):
                PC1        PC2        PC3        PC4
tot_cases 0.5521683 -0.4144450  0.1881917  0.6985194
new_case  0.4397724  0.5898645  0.6545131 -0.1739906
new_death 0.4633931  0.4948994 -0.7247838  0.1225963
tot_death 0.5357026 -0.4851492 -0.1043311 -0.6832043
```

Cluster Dendrogram:

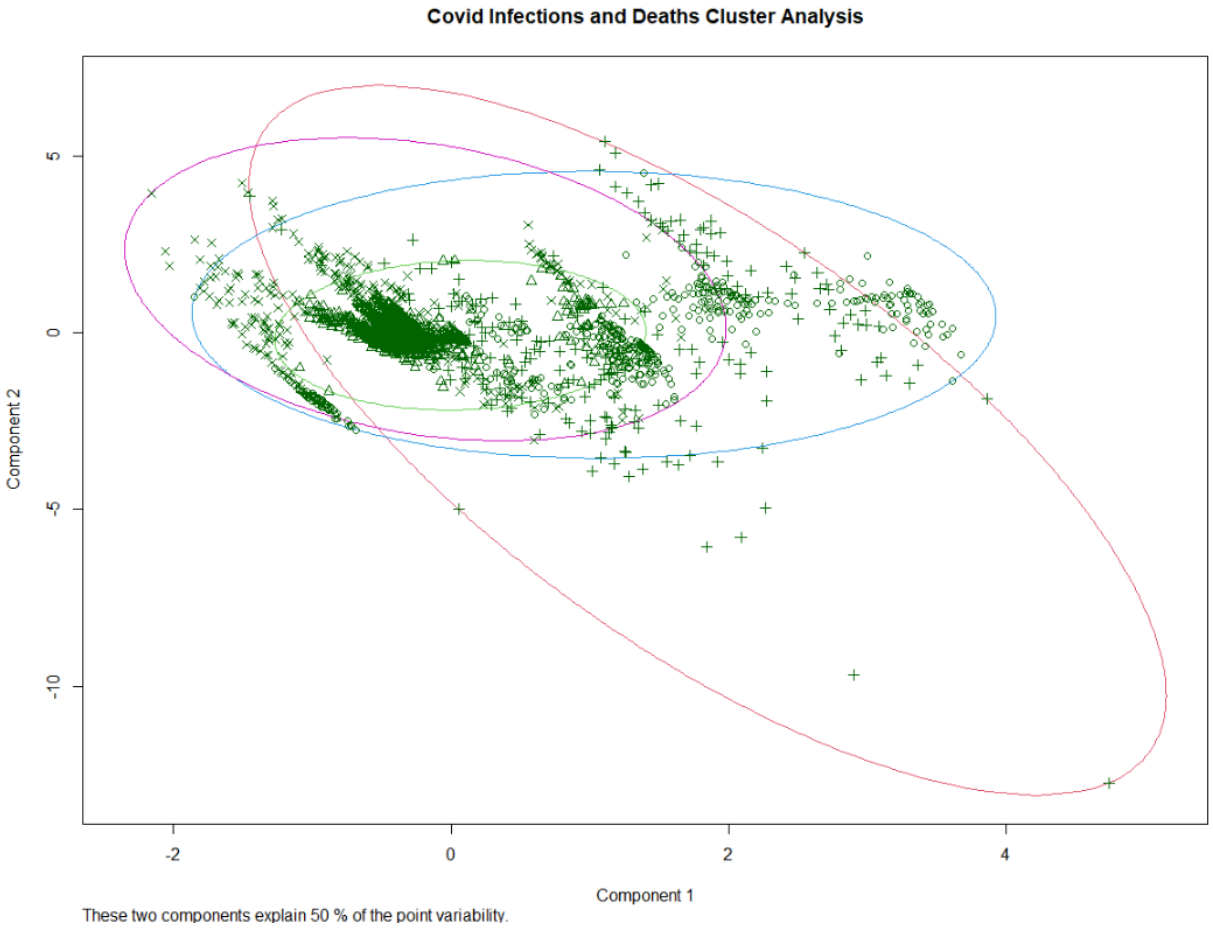


Blue = 2 clusters, **Red** = 4 clusters **Green** = 8 clusters

Analysis: upon examination of the dendrogram, we can see a long period of low values, followed by a short period of rapidly increasing values. This is demonstrated most clearly by the blue rectangular cluster markers, which divides the dendrogram into two parts. This is consistent with the growth pattern observed in the virus over time, with many states experiencing relatively low case numbers early in the pandemic, followed by a period of rapid growth during the Nov 2020- January 2021 time period.

```
> summary(Covid_Kmeans)
             Length Class  Mode
cluster      2682   -none- numeric
centers        16   -none- numeric
totss           1   -none- numeric
withinss        4   -none- numeric
tot.withinss    1   -none- numeric
betweenss       1   -none- numeric
size            4   -none- numeric
iter            1   -none- numeric
ifault          1   -none- numeric
```

Cluster Plot:



Analysis: When the Total number of Cases, Total number of Deaths, New Daily Cases and New daily deaths are plotted with 4 clusters, we get the above result. The variability between clusters was calculated to be 69.5% There are 16 centers within the cluster plot as computed within Rstudio.

```
Clustering vector:
   [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [50] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [99] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [148] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [197] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [246] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [295] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2
 [344] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [393] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2
 [442] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [491] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [540] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [589] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [638] 2 2 2 2 2 2 2 2 2 4 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 4 4 2 4 4 4 2 2 2 2 2 2 2 2
 [687] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [736] 2 2 2 2 2 2 2 2 2 4 4 4 2 2 2 2 2 2 2 2 2 4 4 4 4 2 2 2 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 2 2 2 2 2 2
 [785] 2 2 2 2 2 2 4 2 2 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 4 4 4 2 2 2 2 2 2 2 2 2 2 2
 [834] 4 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 4 2 2 2 2 2 2
 [883] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 4 2 2 2 2 2 2 2 2 4 2 2 4 2 2 2 4 2 2 2 2 2 4 2 2 4 2 2
 [932] 2 2 2 2 2 2 2 2 2 4 4 2 2 2 2 2 2 4 2 2 2 2 2 2 2 4 4 4 2 4 4 2 4 4 2 2 4 4 4 4 4 2 4 4 4 4 2 4 2
 [981] 4 4 4 4 4 2 4 4 4 4 2 2 2 2 2 2 4 4 4 2
 [ reached getOption("max.print") -- omitted 1682 entries ]
```

```
Within cluster sum of squares by cluster:
[1] 1048.1731  637.0609  869.0444  716.5004
 (between_SS / total_SS =  69.5 %)
```

Part 2: Summary

Part two of this study indicates that the variability of COVID-19 Cases rates and fatality rates can be effectively explained by the state or region within the United States. An analysis of variance between the states selected for this study determined that the differences in the mean values of recorded total cases and total deaths per day were different enough among the states to be statistically significant. Two sample T tests of the states with the highest Infection rates (Louisiana and Oklahoma) as well as the States with the Highest fatality rates (Louisiana and Connecticut) determined that the differences between these states were statistically significant. CHI squared analysis of the rate of new cases and new deaths per day in all of the states included in this study showed a high degree of variance in the rate of new cases and new deaths per day as compared to expected values. Louisiana was shown to have the highest rate of both new cases and new deaths per day of the states included in this study. Of daily records submitted to the CDC, 55% of Louisiana's daily death numbers were in the 4th quartile. Louisiana also had a high infection rate, with 32% of daily new case records in the 4th Quartile. Oregon had the lowest rate of daily new infections, whereas Utah had the lowest daily death rate. A scatterplot analysis of new COVID-19 cases

deaths displayed a strong positive correlation between the rate of new confirmed cases of COVID 19 per day, and the rate of new COVID-19 deaths per day. This phenomenon is also shown in the plots of the relationships between total cases and total deaths in part one. It should be noted that although total cases and total deaths are correlated in all states, the degree of correlation varies. Further analysis of factors such as population density, available hospital space, and other variables which contribute to COVID survivability is needed to determine why such discrepancies exist.

Conclusion

The findings of this study are largely consistent with other statistical analyses of COVID-19 rates in the U.S. It has been statistically demonstrated that COVID-19 Infection rates vary based on location. Infection rates in Oregon, for example remained lower than in Oklahoma, Connecticut, Utah, and Iowa throughout the duration of the pandemic, despite Oregon having a larger population than those states. Many factors contribute to differences in Infection rate, with prodigious amounts of ongoing statistical analysis of these factors ongoing in medical, academic and public health circles. "Higher risks of clustering and incidence of COVID-19 were consistently observed in metropolitan versus rural counties, counties closest to core airports, the most populous counties, and counties with the highest proportion of racial/ethnic minorities. However, geographic differences in incidence have shrunk since early April, driven by a significant decrease in the incidence in these counties (EWPC range: −2.0%, −4.2%) and a consistent increase in other areas l" (6). This tendency for greater COVID clustering in more urbanized regions with more transit hubs may be partly responsible for the rapid increase in COVID cases in Connecticut early in the pandemic, as well as the disproportionally high fatality rate in that state relative to others in this study. The inverse may be the case in Utah, where fatality rates remained low despite high infection rates. The rate of COVID Infection grew rapidly during the holiday from November 2020 – January 2021. This Increase can be seen especially prominently in Louisiana, Kentucky, Utah, and Iowa. Several of these states, (Kentucky, Utah, and Iowa) had relatively low transmission rates before this period. Louisiana maintained the highest infection and death rate for the duration of the pandemic among

the states in this study. Although this study has effectively demonstrated that there are statistically significant differences in infection and death rates between different states, further analysis is needed to account for these differences. An effective follow up study would examine how factors such as population density, hospital space per capita, access to healthcare/ health insurance, and other variables affected the spread of the virus. It would also be helpful to view the total case increase over time graphics with an overlay depicting when lockdowns were instituted, and when restrictions were relaxed. I believe this would provide an effective means of determining how these policies affected health outcomes, and to what degree.

Works Cited:

1. Amirhoshang Hoseinpour Dehkordi, Majid Alizadeh, Pegah Derakhshan, Peyman Babazadeh, Arash Jahandideh, *Understanding epidemic data and statistics: A case study of COVID-19,* Journal of Medical Virology, Volume 92, Issue 7, Special Issue on New Coronavirus (2019-nCoV or SARS-CoV-2) and the outbreak of the respiratory illness (COVID-19): Part-IV, July 2020Pages 868-882. Accessed 04/12/21. https://onlinelibrary.wiley.com/doi/full/10.1002/jmv.25885

2. Neil Pearce PhD, Jan P. Vandenbroucke MD, PhD, Tyler J. VanderWeele PhD, and Sander Greenland DrPH, MS, *Accurate Statistics on COVID-19 Are Essential for Policy Guidance and Decisions* AJPH, A publication of the American Public Health Association, July 2020, Accessed 04/12/21. https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2020.305708

3. Ivan Franch-Pardoa,∗, Brian M. Napoletanob,∗, Fernando Rosete-Vergesa, Lawal Billa, *Spatial analysis and GIS in the study of COVID-19. A review*, Science of the Total Environment Volume 739, 2020, Accessed 04/12/21. https://reader.elsevier.com/reader/sd/pii/S0048969720335531?token=87F7AECB2B6F57634DD4CC19E3227208B3983208FCDA365B020B26241930F2ADCD305F91F7E30B2A36854DFA561B457A&originRegion=us-east-1&originCreation=20210412194558

4. CDC*, United States COVID-19 Cases and Deaths by State over Time,* U.S. Centers for Disease Control and Prevention. Metadata updated May 8, 2021. Accessed 4/6/2021 https://catalog.data.gov/dataset/united-states-covid-19-cases-and-deaths-by-state-over-time-7e14c

5. Census Bureau, *State Population Totals and Components of Change*: 2010-2019, U.S. Census Bureau, Last revised 04/20/21. Accessed 04/23/21. *https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html*

6. Yun Wang, Ying Liu, James Struthers, Min Lian, *Spatiotemporal Characteristics of the COVID-19 Epidemic in the United States* Clinical Infectious Diseases, Volume 72, Issue 4, 15 February 2021, Pages 643–651. https://academic.oup.com/cid/article/72/4/643/5868985?login=true

7. Lauren M. Andersena, Stella R .Harden, Margaret M. Sugg Ph.D.,Jennifer D. Runkle Ph.D., .Taylor E.Lundquist, *Analyzing the spatial determinants of local Covid-19 transmission in the United States,* Science of the Total Environment *Volume 754, 1 February 2021. https://www.sciencedirect.com/science/article/pii/S0048969720359258*