Executive Summary

Describes the Data Set

Summary:

The Hopkins data is available at the county level in the United States. The AP has paired this data with population figures and county rural/urban designations, and has calculated caseload and death rates per 100,000 people. Be aware that caseloads may reflect the availability of tests, and the ability to turn around test results quickly, rather than actual disease spread or true infection rates.

Data information:

Johns Hopkins' county-level COVID-19 case and death data, paired with population and rates per 100,000. Source: https://data.world/associatedpress/johns-hopkins-coronavirus-case-tracker https://data.world/associatedpress/johns-hopkins-coronavirus-case-trackerhttps://data.world/resources/coronavirus/

Variables:

This data set has the variables, annd Ten (10) columns and 3,269 rows:

Date: last update: In which the database has been updated by The Associated Press johns-hopkins. Highlighting this database is automatically updated from the data.world page since it is directly connected to this project.

Ubication Geographic: Data set collected in the United Stated by Location type, State, County_name, County name long.

Coordinate: Fips code, Latitude, Longitude.

Size of population by geographic: NCHS Urbanization: Medium metro, small metro, non core, large fringe metro, micropolitan.

Population segmentation: Total Population subdivide by Confirmed (covic) and Confirmed per 100.000.

Population segmentation by Deaths and Deaths per 100.000 (covic).

Summarizes the Goal of project:

This project has the objective to analyze the actual pandemic issue that has affected the society, families and the economy, the emphasis is to identifique what are the places with more impact in the state of the United States. To accomplish this objectives we performed the next steps: data cleaning, exploration, visualization, model approach and finally prediction.

Double-click (or enter) to edit

Installing libraries

```
install.packages('caret')
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)

also installing the dependencies 'numDeriv', 'SQUAREM', 'lava', 'prodlim', 'iterators', 'data.table', 'gower', 'ipred', 'timeDate', 'foreach'
```

```
install.packages('mlbench')
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

Importing libraries

```
library(dplyr)
library(magrittr)
library(knitr)
library(MASS)
library(lattice)
library(tidyverse)
library(mlbench)
library(tidyverse)
library(ggplot2)
library(caret)
```

How many CPU cores are there?

```
library(parallel)
detectCores(all.tests = FALSE, logical = TRUE)
```

2

```
#Load data set directt from data world page.
#Read File and Columns name. and deleted unnecesary columns.
df <- read.csv('https://query.data.world/s/e7co64e3e47t3sdnrviomq3sasf6g2')
df$location_type <- NULL
df$county_name_long <- NULL
df$fips_code <- NULL
df$county_name <- NULL
head(df)
```

A data.frame: 6 × 10

	last_update	state	lat	lon	NCHS_urbanization	total_population	confirmed	confirmed_per_100000	deaths	deaths_per_100000
	<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>	<int>	<dbl>	<int>	<dbl>
1	2021-02-12 00:23:16 UTC	Alabama	32.53953	-86.64408	Medium metro	55200	5970	10815.22	81	146.74
2	2021-02-12 00:23:16 UTC	Alabama	30.72775	-87.72207	Small metro	208107	18960	9110.70	240	115.33
3	2021-02-12 00:23:16 UTC	Alabama	31.86826	-85.38713	Non-core	25782	2030	7873.71	46	178.42
4	2021-02-12 00:23:16 UTC	Alabama	32.99642	-87.12511	Large fringe metro	22527	2377	10551.78	54	239.71

```
#Create a data sett My data and replace nan from original data set
my_data <- as_tibble(df)
my_data <- replace(my_data, is.na(my_data), 0)
```

```
#filter only integer data to developmente the differents analysis.
new_data = my_data %>% select_if(is.numeric)
```

The principal variable that we will analyze is, how many cases of infected by Covic have been given in different states of the United States, now we have that the average of confirmed by covic is 834 the median is 2073. In the other hands, the mean for the deaths cause by covic is 8324 and the median is 37, standard deviation 557. Note that this data set is automatically update every day, this data can be change.

```
#Mean, Media and Standard deviation of confirmed by Covic 19
mean=mean(new_data[['confirmed']])
median=median(new_data[['confirmed']])
standard_deviation= sd(new_data[['confirmed']])
print((paste0('Mean of Confirmed  cases cause for Covic:', mean)))
print((paste0('Median of Confirmed  cases cause for Covic:', median)))
print((paste0('Standar Deviation of Confirmed  cases cause by Covic:', standard_deviation)))
```

```
[1] "Mean of Confirmed  cases cause for Covic:8373.55858060569"
[1] "Median of Confirmed  cases cause for Covic:2098"
[1] "Standar Deviation of Confirmed  cases cause by Covic:31571.8279383445"
```

```
meean=mean(df[['deaths']])
median=median(df[['deaths']])
sd=sd(df[['deaths']])
print((paste0('Mean of deaths cases cause for Covic:', mean)))
print((paste0('Median of deaths  cases cause for Covic:', median)))
print((paste0('Standar Deviation of deaths  cases cause by Covic:', sd)))
```

```
[1] "Mean of deaths cases cause for Covic:8373.55858060569"
[1] "Median of deaths  cases cause for Covic:38"
[1] "Standar Deviation of deaths  cases cause by Covic:563.256882265914"
```

Covariance data set Covic

For a sample of cases confirmed of 100,000 inhabitants, the covariance is of the 3% and the correlation is the 1% deaths by covic.

```
cov(new_data)
```

A matrix: 7 × 7 of type dbl

	lat	lon	total_population	confirmed	confirmed_per_100000	deaths	deaths_per_100000
lat	65.29220	-101.2963	5.238697e+03	-3.504861e+03	7084.046	-4.563559e+01	8.262083e+01
lon	-101.29632	340.8037	-2.280191e+05	-1.820301e+04	-19946.458	1.614169e+02	-2.880707e+02
total_population	5238.69725	-228019.0918	1.049150e+11	9.878597e+09	1351737.672	1.678893e+08	-1.217842e+06
confirmed	-3504.86116	-18203.0148	9.878597e+09	9.967803e+08	5773206.100	1.642123e+07	-3.436224e+04
deaths_per_100000	82.62083	-288.0707	-1.217842e+06	-3.436224e+04	196588.062	3.522653e+03	1.115755e+04

Correlacion data set Covic

```
cor(new_data)
```

A matrix: 7 × 7 of type dbl

	lat	lon	total_population	confirmed	confirmed_per_100000	deaths	deaths_per_100000
lat	1.000000000	-0.67906429	0.002001584	-0.01373854	0.267204308	-0.01002690	0.09679980
lon	-0.679064288	1.00000000	-0.038132937	-0.03123140	-0.329311132	0.01552353	-0.14772789
total_population	0.002001584	-0.03813294	1.000000000	0.96599911	0.001271939	0.92023298	-0.03559488
confirmed	-0.013738536	-0.03123140	0.965999112	1.00000000	0.055732701	0.92342045	-0.01030380
confirmed_per_100000	0.267204308	-0.32931113	0.001271939	0.05573270	1.000000000	0.03513731	0.56723818
deaths	-0.010026896	0.01552353	0.920232977	0.92342045	0.035137311	1.00000000	0.05920780
deaths_per_100000	0.096799798	-0.14772789	-0.035594882	-0.01030380	0.567238183	0.05920780	1.00000000

```
#Read File and Columns name str.
str(new_data)
```

```
tibble [3,269 × 7] (S3: tbl_df/tbl/data.frame)
 $ lat                 : num [1:3269] 32.5 30.7 31.9 33 34 ...
 $ lon                 : num [1:3269] -86.6 -87.7 -85.4 -87.1 -86.6 ...
 $ total_population    : num [1:3269] 55200 208107 25782 22527 57645 ...
 $ confirmed           : int [1:3269] 5970 18960 2030 2377 5955 1136 1886 12539 3305 1738 ...
 $ confirmed_per_100000: num [1:3269] 10815 9111 7874 10552 10330 ...
 $ deaths              : int [1:3269] 81 240 46 54 116 32 64 257 92 37 ...
 $ deaths_per_100000   : num [1:3269] 147 115 178 240 201 ...
```

Summary show us the minimu, max and median and mean the all data set.

```
summary(new_data)
```

```
      lat             lon          total_population     confirmed
 Min.   : 0.00   Min.   :-174.16   Min.   :       0   Min.   :      0
 1st Qu.:34.02   1st Qu.: -97.73   1st Qu.:   10447   1st Qu.:    845
 Median :38.04   Median : -89.52   Median :   25324   Median :   2098
 Mean   :37.18   Mean   : -89.66   Mean   :   99546   Mean   :   8374
 3rd Qu.:41.65   3rd Qu.: -82.51   3rd Qu.:   65558   3rd Qu.:   5475
 Max.   :69.31   Max.   :   0.00   Max.   :10098052   Max.   :1155491
 confirmed_per_100000     deaths          deaths_per_100000
 Min.   :    0        Min.   :    0.0   Min.   :  0.00
 1st Qu.: 6564        1st Qu.:   13.0   1st Qu.: 80.77
 Median : 8497        Median :   38.0   Median :142.46
 Mean   : 8310        Mean   :  145.3   Mean   :156.02
 3rd Qu.:10276        3rd Qu.:   90.0   3rd Qu.:211.01
 Max.   :33481        Max.   :18519.0   Max.   :788.57
```

```
str(df)
```

```
'data.frame':	3269 obs. of  10 variables:
 $ last_update         : chr  "2021-02-12 00:23:16 UTC" "2021-02-12 00:23:16 UTC" "2021-02-12 00:23:16 UTC" "2021-02-12 00:23:16 UTC" ...
 $ state               : chr  "Alabama" "Alabama" "Alabama" "Alabama" ...
 $ lat                 : num  32.5 30.7 31.9 33 34 ...
 $ lon                 : num  -86.6 -87.7 -85.4 -87.1 -86.6 ...
 $ NCHS_urbanization   : chr  "Medium metro" "Small metro" "Non-core" "Large fringe metro" ...
 $ total_population    : num  55200 208107 25782 22527 57645 ...
 $ confirmed           : int  5970 18960 2030 2377 5955 1136 1886 12539 3305 1738 ...
 $ confirmed_per_100000: num  10815 9111 7874 10552 10330 ...
 $ deaths              : int  81 240 46 54 116 32 64 257 92 37 ...
 $ deaths_per_100000   : num  147 115 178 240 201 ...
```

```
summary(df)
```

```
 last_update           state                lat             lon
 Length:3269        Length:3269        Min.   :17.98   Min.   :-174.16
 Class :character   Class :character   1st Qu.:34.34   1st Qu.: -97.93
 Mode  :character   Mode  :character   Median :38.19   Median : -89.92
                                       Mean   :37.96   Mean   : -91.54
                                       3rd Qu.:41.71   3rd Qu.: -82.95
                                       Max.   :69.31   Max.   : -65.29
                                       NA's   :67      NA's   :67
 NCHS_urbanization  total_population    confirmed       confirmed_per_100000
 Length:3269        Min.   :     102   Min.   :      0   Min.   :    0
 Class :character   1st Qu.:   11309   1st Qu.:    845   1st Qu.: 6713
 Mode  :character   Median :   26212   Median :   2098   Median : 8562
                    Mean   :  101884   Mean   :   8374   Mean   : 8505
                    3rd Qu.:   66842   3rd Qu.:   5475   3rd Qu.:10323
                    Max.   :10098052   Max.   :1155491   Max.   :33481
                    NA's   :75                           NA's   :75
     deaths        deaths_per_100000
 Min.   :    0.0   Min.   :  0.00
 1st Qu.:   13.0   1st Qu.: 85.45
```

```
df$deaths <- factor(df$deaths,
            levels=c(0, 1),
            labels=c('confirmed', 'deaths'))
```

```
                    NA's   :75
```

```
summary(new_data$deaths)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    13.0    38.0   145.3    90.0 18519.0 
```

```
which.max(df[['deaths']])
which.max(df[['confirmed']])
```

75
204

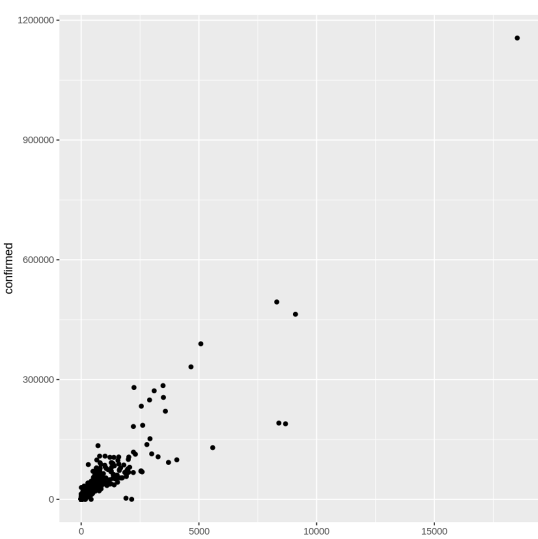
Group by NCHS Urbanization and State from United States, in Medium, small, large, micropolitan and non-core

```
dataset <- df %>% group_by(NCHS_urbanization, state)
```

Analysis cases confirmed and deaths group by states.

```
dataset %>% summarise(
  confirmed=mean(confirmed),
  deaths=mean(deaths)
)
```

```
ggplot(new_data, aes(deaths, confirmed)) + geom_point()
```




```
dataset %>% filter(total_population==max(total_population))
```

A grouped_df: 250 × 10

last_update	state	lat	lon	NCHS_urbanization	total_population	confirmed	confirmed_per_100000	deaths	deaths_per_100000
<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>	<int>	<dbl>	<fct>	<dbl>
2021-02-12 00:23:16 UTC	Alabama	30.72775	-87.72207	Small metro	208107	18960	9110.70	NA	115.33
2021-02-12 00:23:16 UTC	Alabama	34.45947	-85.80783	Non-core	71200	8332	11702.25	NA	227.53
2021-02-12 00:23:16 UTC	Alabama	33.55555	-86.89506	Large central metro	659892	69117	10473.99	NA	191.70
2021-02-12 00:23:16 UTC	Alabama	34.36976	-86.30487	Micropolitan	95145	11057	11621.21	NA	200.75
2021-02-12 00:23:16 UTC	Alabama	30.78472	-88.20842	Medium metro	414659	34555	8333.35	NA	160.86
2021-02-12 00:23:16 UTC	Alabama	33.26880	-86.66233	Large fringe metro	211261	21098	9986.70	NA	90.88
2021-02-12 00:23:16 UTC	Alaska	61.14998	-149.14270	Medium metro	296112	26011	8784.18	NA	51.33
2021-02-12 00:23:16 UTC	Alaska	64.80726	-146.56927	Small metro	99653	6018	6038.96	NA	26.09
2021-02-12 00:23:16 UTC	Alaska	58.45032	-134.20044	Micropolitan	32330	1205	3727.19	NA	15.47
2021-02-12 00:23:16 UTC	Alaska	60.24430	-151.53889	Non-core	58220	3958	6798.35	NA	37.79
2021-02-12 00:23:16 UTC	Arizona	35.39465	-109.48924	Non-core	71522	10036	14032.05	NA	480.97
2021-02-12 00:23:16 UTC	Arizona	33.34836	-112.49182	Large central metro	4253913	494345	11620.95	NA	195.26
2021-02-12 00:23:16 UTC	Arizona	35.39977	-110.32190	Micropolitan	108705	15060	13854.01	NA	426.84
2021-02-12 00:23:16 UTC	Arizona	32.09713	-111.78900	Medium metro	1019722	105909	10386.07	NA	198.39
2021-02-12 00:23:16 UTC	Arizona	32.90526	-111.34495	Large fringe metro	419721	44439	10587.75	NA	165.82
2021-02-12 00:23:16 UTC	Arizona	34.59934	-112.55386	Small metro	224645	16616	7396.56	NA	193.19
2021-02-12 00:23:16 UTC	Arkansas	36.34039	-93.54270	Non-core	27887	2690	9646.07	NA	143.44
2021-02-12 00:23:16 UTC	Arkansas	35.83018	-90.63236	Small metro	105701	12563	11885.41	NA	164.62
2021-02-12 00:23:16 UTC	Arkansas	35.21247	-90.30839	Large fringe metro	49013	5580	11384.73	NA	183.62
2021-02-12 00:23:16 UTC	Arkansas	34.77054	-92.31355	Medium metro	393463	36192	9198.32	NA	140.80
2021-02-12 00:23:16 UTC	Arkansas	35.25688	-91.74908	Micropolitan	78804	6964	8837.11	NA	126.90
2021-02-12 00:23:16 UTC	California	39.66728	-121.60053	Small metro	227075	10594	4665.42	NA	65.18
2021-02-12 00:23:16 UTC	California	38.20537	-120.55291	Non-core	45235	1850	4089.75	NA	55.27
2021-02-12 00:23:16 UTC	California	36.75734	-119.64670	Medium metro	978130	91876	9393.03	NA	131.17
2021-02-12 00:23:16 UTC	California	40.69923	-123.87604	Micropolitan	135768	3009	2216.28	NA	23.57
2021-02-12 00:23:16 UTC	California	34.30828	-118.22824	Large central metro	10098052	1155491	11442.71	NA	183.39
2021-02-12 00:23:16 UTC	California	34.84060	-116.17747	Large fringe metro	2135413	280068	13115.40	NA	105.04
2021-02-12 00:23:16 UTC	Colorado	39.64977	-104.33536	Large fringe metro	636671	46525	7307.54	NA	99.89
2021-02-12 00:23:16 UTC	Colorado	38.86246	-107.86313	Non-core	30346	2404	7921.97	NA	184.54
2021-02-12 00:23:16 UTC	Colorado	39.76018	-104.87257	Large central metro	693417	57283	8260.97	NA	107.73
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2021-02-12 00:23:16 UTC	Utah	40.11667	-111.66577	Medium metro	590440	87084	14749.00	NA	51.32

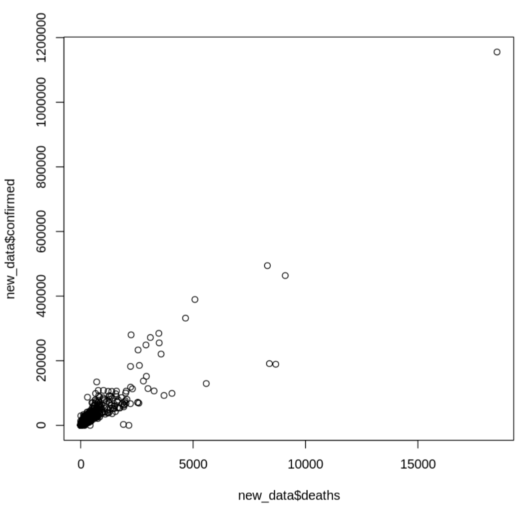
2021-02-12 00:23:16 UTC	Vermont	44.46323	-73.08359	Small metro	162052	4301	2654.09	NA	50.60
2021-02-12 00:23:16 UTC	Vermont	43.57724	-73.03742	Micropolitan	59273	1109	1871.00	NA	15.18
2021-02-12 00:23:16 UTC	Vermont	42.98698	-72.71269	Non-core	43150	847	1962.92	NA	25.49
2021-02-12 00:23:16 UTC	Virginia	38.02081	-78.55481	Small metro	106355	4164	3915.19	NA	36.67
2021-02-12 00:23:16 UTC	Virginia	38.83678	-77.27566	Large fringe metro	1143529	62854	5496.49	NA	72.23

```
dataset %>% filter(confirmed_per_100000==max(confirmed_per_100000))
```

A grouped_df: 250 × 10

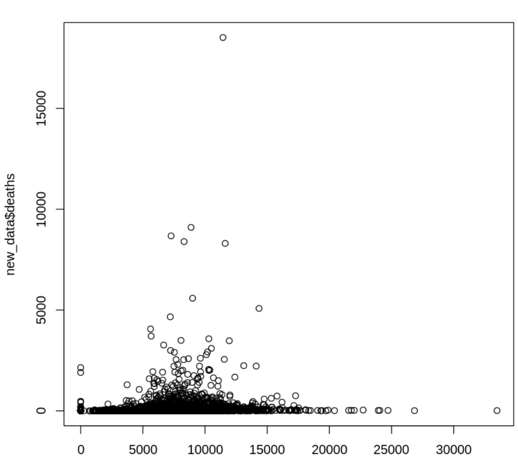
last_update	state	lat	lon	NCHS_urbanization	total_population	confirmed	confirmed_per_100000	deaths	deaths_per_100000
<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>	<int>	<dbl>	<fct>	<dbl>
2021-02-12 00:23:16 UTC	Alabama	32.99642	-87.12511	Large fringe metro	22527	2377	10551.78	NA	239.71
2021-02-12 00:23:16 UTC	Alabama	31.68100	-87.83549	Non-core	24387	3367	13806.54	NA	176.32
2021-02-12 00:23:16 UTC	Alabama	32.76039	-87.63285	Small metro	14887	2011	13508.43	NA	382.88
2021-02-12 00:23:16 UTC	Alabama	34.78144	-85.99750	Micropolitan	52094	6401	12287.40	NA	165.09
2021-02-12 00:23:16 UTC	Alabama	33.55555	-86.89506	Large central metro	659892	69117	10473.99	NA	191.70
2021-02-12 00:23:16 UTC	Alabama	32.15973	-86.65158	Medium metro	10236	1283	12534.19	NA	410.32
2021-02-12 00:23:16 UTC	Alaska	61.14998	-149.14270	Medium metro	296112	26011	8784.18	NA	51.33
2021-02-12 00:23:16 UTC	Alaska	60.90980	-159.85618	Non-core	18040	3498	19390.24	NA	94.24
2021-02-12 00:23:16 UTC	Alaska	64.80726	-146.56927	Small metro	99653	6018	6038.96	NA	26.09
2021-02-12 00:23:16 UTC	Alaska	58.45032	-134.20044	Micropolitan	32330	1205	3727.19	NA	15.47

```
plot(new_data$deaths, new_data$confirmed)
```

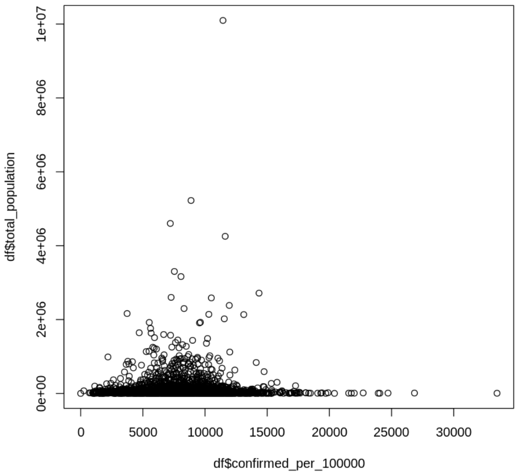


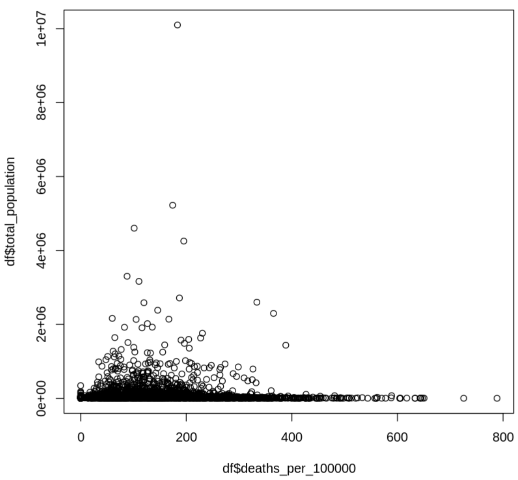
Visualization cases confirmed compared with the deaths.

```
plot(new_data$confirmed_per_100000, new_data$deaths)
```



```
plot(df$confirmed_per_100000, df$total_population)
title(plot(df$deaths_per_100000, df$total_population))
```






```
dataset %>% filter(deaths_per_100000==max(deaths_per_100000))
```

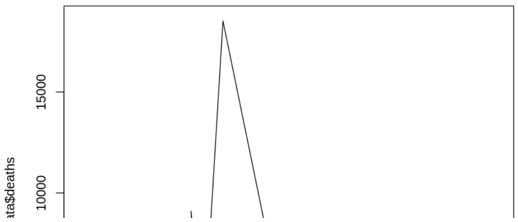
A grouped_df: 250 × 10

last_update	state	lat	lon	NCHS_urbanization	total_population	confirmed	confirmed_per_100000	deaths	deaths_per_100000
<chr>	<chr>	<dbl>	<dbl>	<chr>	<dbl>	<int>	<dbl>	<fct>	<dbl>
2021-02-12 00:23:16 UTC	Alabama	33.26984	-85.85836	Non-core	13378	1386	10360.29	NA	396.17
2021-02-12 00:23:16 UTC	Alabama	32.32688	-87.10867	Micropolitan	40029	3269	8166.58	NA	314.77
2021-02-12 00:23:16 UTC	Alabama	32.76039	-87.63285	Small metro	14887	2011	13508.43	NA	382.88
2021-02-12 00:23:16 UTC	Alabama	33.55555	-86.89506	Large central metro	659892	69117	10473.99	NA	191.70
2021-02-12 00:23:16 UTC	Alabama	32.15973	-86.65158	Medium metro	10236	1283	12534.19	NA	410.32
2021-02-12 00:23:16 UTC	Alabama	33.80271	-87.30027	Large fringe metro	64493	6325	9807.27	NA	362.83
2021-02-12 00:23:16 UTC	Alaska	61.14998	-149.14270	Medium metro	296112	26011	8784.18	NA	51.33
2021-02-12 00:23:16 UTC	Alaska	64.80726	-146.56927	Small metro	99653	6018	6038.96	NA	26.09
2021-02-12 00:23:16 UTC	Alaska	58.45032	-134.20044	Micropolitan	32330	1205	3727.19	NA	15.47
2021-02-12 00:23:16 UTC	Alaska	59.89098	-140.36015	Non-core	689	66	9579.10	deaths	145.14
2021-02-12 00:23:16 UTC	Arizona	35.39465	-109.48924	Non-core	71522	10036	14032.05	NA	480.97
2021-02-12 00:23:16 UTC	Arizona	33.34836	-112.49182	Large central metro	4253913	494345	11620.95	NA	195.26
2021-02-12 00:23:16 UTC	Arizona	35.39977	-110.32190	Micropolitan	108705	15060	13854.01	NA	426.84
2021-02-12 00:23:16 UTC	Arizona	32.09713	-111.78900	Medium metro	1019722	105909	10386.07	NA	198.39
2021-02-12 00:23:16 UTC	Arizona	32.90526	-111.34495	Large fringe metro	419721	44439	10587.75	NA	165.82
2021-02-12 00:23:16 UTC	Arizona	32.76896	-113.90667	Small metro	207829	35910	17278.63	NA	360.87
2021-02-12 00:23:16 UTC	Arkansas	35.21247	-90.30839	Large fringe metro	49013	5580	11384.73	NA	183.62
2021-02-12 00:23:16 UTC	Arkansas	36.38177	-91.81729	Non-core	12139	1052	8666.28	NA	362.47
2021-02-12 00:23:16 UTC	Arkansas	33.70376	-94.23469	Small metro	12417	1102	8874.93	NA	322.14
2021-02-12 00:23:16 UTC	Arkansas	35.91947	-93.21613	Micropolitan	7848	669	8524.46	NA	331.29
2021-02-12 00:23:16 UTC	Arkansas	35.19606	-94.27163	Medium metro	127461	14140	11093.59	NA	194.57
2021-02-12 00:23:16 UTC	California	33.03931	-115.36690	Small metro	180216	26589	14753.96	NA	324.06
2021-02-12 00:23:16 UTC	California	36.51112	-117.41120	Non-core	18085	1176	6502.63	NA	188.00
2021-02-12 00:23:16 UTC	California	34.30828	-118.22824	Large central metro	10098052	1155491	11442.71	NA	183.39
2021-02-12 00:23:16 UTC	California	34.84060	-116.17747	Large fringe metro	2135413	280068	13115.40	NA	105.04

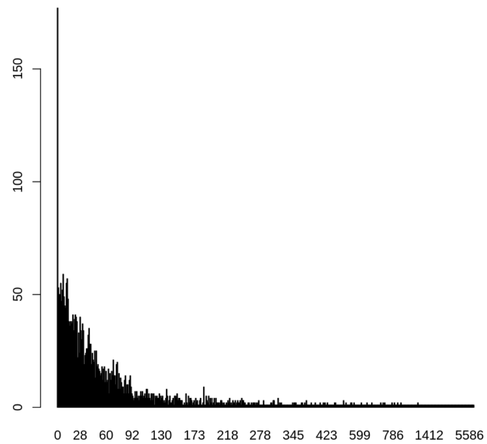
Visualization compared total population with cases confirmed and deaths.

2021-02-12

```
1 plot(new_data$confirmed_per_100000, new_data$deaths, type='l')
```



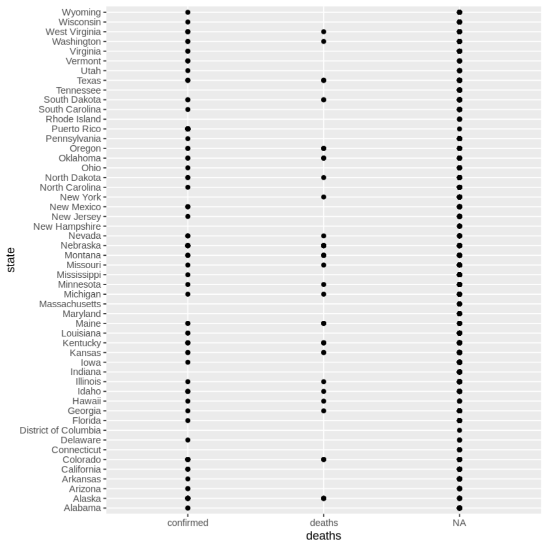

```
#Barplot compared deaths
barplot(table(new_data$deaths))
```




```
hist(new_data$total_population)
hist(new_data$confirmed_per_100000, breaks=10)
```

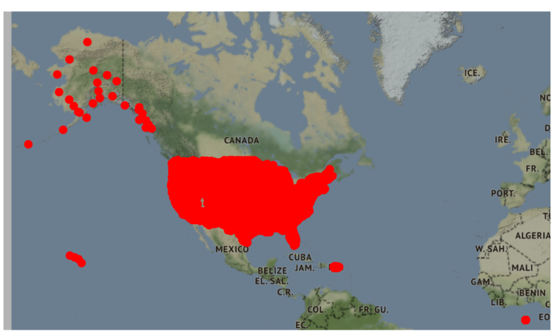
Visualization show differents the states with most deaths us that california is one the most deaths had.

```
ggplot(df, aes(x=deaths, y=state)) + geom_point()
```




```
geopoin <- subset(df, state == 'deaths')
qmplot(lon, lat, data=df, colour=I('red'), size=I(3), darken=.3)
```

```
Using zoom = 3...
```



Prediction

The prediction show us that for 24 people infected 2 would death and this prediction has 0.915 of accuracy and the precision is 0.913, data that confirmed that the average of death for this virus can oscillate into 2% or 3% of impacted in the all population infected.

```
set.seed(0)
  actual = c('confirmed', 'deaths')[runif(100, 1, 4)]
  predicted = actual
  predicted[runif(30,1,100)] = actual[runif(30,1,100)]
  cm = as.matrix(table(Actual=actual, Predicted=predicted))
  cm
```

```
           Predicted
Actual      confirmed deaths
  confirmed        24      2
  deaths            3     30
```

```
num_instances = sum(cm)
num_class = nrow(cm)
diag = diag(cm) #classified
```

```
rowsums = apply(cm, 1, sum)
colsums = apply(cm, 2, sum) #predictions
p = rowsums / num_instances # distribution of instances over the actual classes
q = colsums / num_instances # distribution of instances over the predicted classes
```

```
accuracy = sum(diag) / num_instances
accuracy
```

0.915254237288136

```
precision = diag / colsums
recall = diag / rowsums
f1 = 2 * precision * recall / (precision + recall)
```

```
data.frame(precision, recall, f1)
```

A data.frame: 2 × 3

	precision	recall	f1
	<dbl>	<dbl>	<dbl>
confirmed	0.8888889	0.9230769	0.9056604
deaths	0.9375000	0.9090909	0.9230769

```
macroPrecision = mean(precision)
macroRecall = mean(recall)
macroF1 = mean(f1)
```

```
data.frame(macroPrecision, macroRecall, macroF1)
```

A data.frame: 1 × 3

macroPrecision	macroRecall	macroF1
<dbl>	<dbl>	<dbl>
0.9131944	0.9160839	0.9143687

Machine Learning in R:

Building a Linear Regression Model 1

We build the model Linear Regression to analyze the relationship between differents variable, as size of states more infectioned and deaths or analysis what is the most relevant variante. The coefficient and residuals give us the R-squared is 0.8478 and .8557 with p value of 2.2 a tendeced to increase the cases of deaths.

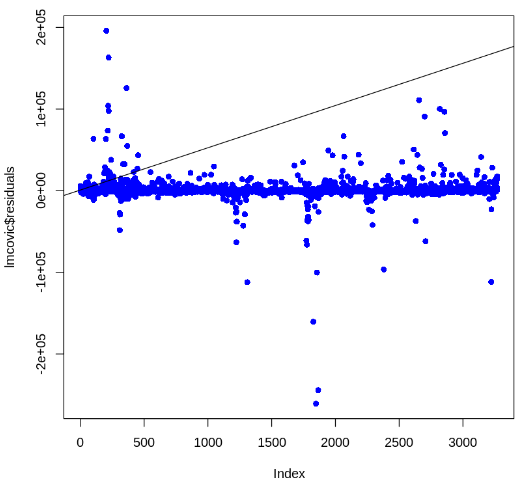
```
lmcovic = lm(confirmed~deaths, data = df)
summary(lmcovic)
```

```
Call:
lm(formula = confirmed ~ deaths, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-260937   -1201    -717     141  196026 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 856.5516   218.9690   3.912 9.35e-05 ***
deaths       51.7635     0.3767 137.410  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12120 on 3267 degrees of freedom
Multiple R-squared:  0.8525,	Adjusted R-squared:  0.8525 
F-statistic: 1.888e+04 on 1 and 3267 DF,  p-value: < 2.2e-16
```

```
lmcovic2 = lm(confirmed~deaths + confirmed_per_100000, data=df)
summary(lmcovic2)
```

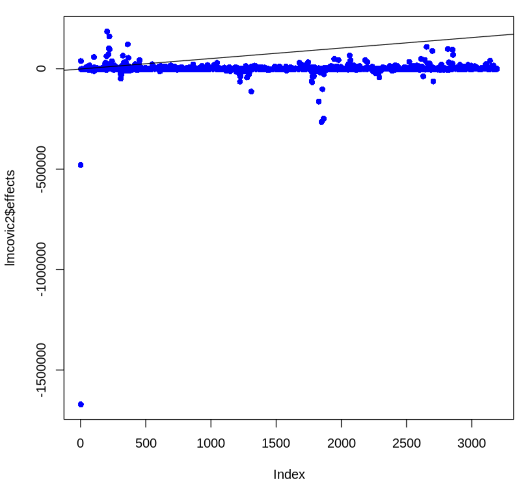
```
Call:
lm(formula = confirmed ~ deaths + confirmed_per_100000, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-263662   -1514    -590     504  188963 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          -1048.3445   624.7193  -1.678  0.09342 .  
deaths                  52.1098     0.3723 139.981  < 2e-16 ***
```

```
plot(lmcovic$residuals, pch=16, col='blue')
abline(lmcovic)
```




```
plot(lmcovic2$effects, pch=16, col='blue')
abline(lmcovic2)
```

```
Warning message in abline(lmcovic2):
“only using the first two of 3 regression coefficients”
```



Linear Regression Model 2

The second model we created a validation data set divide to analysis in 652 and 2617, also a training and test set the prediction show us confirmed the cases of deaths have been increase.

```
set.seed(100)
validationIndex <- caret::createDataPartition(df$deaths, p=0.80, list=FALSE)
validation <- df[-validationIndex,]
dataset <- df[validationIndex,]
dim(validation)
dim(dataset)
```

652 · 14
2617 · 14

```
sum(is.na(dataset))
```

```
sum(is.na(dataset))
```

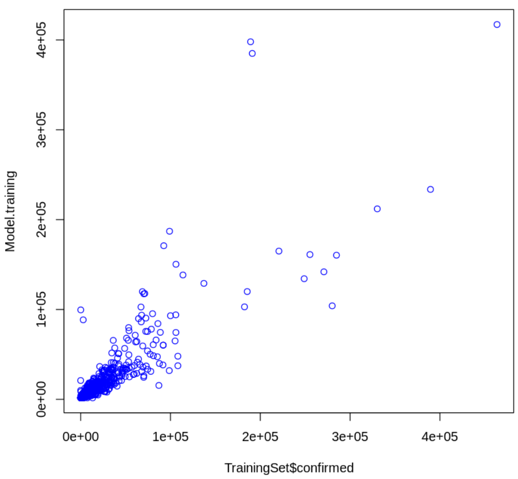
300

```
#Stratified random split of the data set
TrainingIndex <- createDataPartition(df$deaths, p=0.8, list = FALSE)
TrainingSet <- df[TrainingIndex,] # Training Set
TestingSet <- df[-TrainingIndex,] # Test Set

```

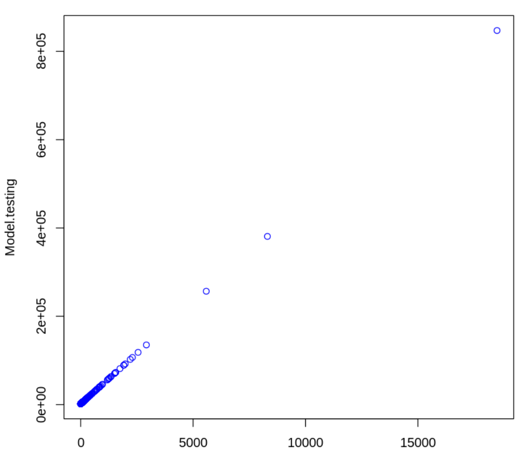
```
# Build Training model
Model <- train(deaths ~ confirmed, data = TrainingSet,
               method = "lm",
               na.action = na.omit,
               preProcess=c("scale","center"),
               trControl= trainControl(method="none")
)
```

```
# Apply model for prediction
Model.training <-predict(Model, TrainingSet) #model to make prediction on Training set
Model.testing <-predict(Model, TestingSet) #model to make prediction on Testing set
```

```
# Scatter plot of Training set performance matrics
plot(TrainingSet$confirmed,Model.training, col = "blue" )

```

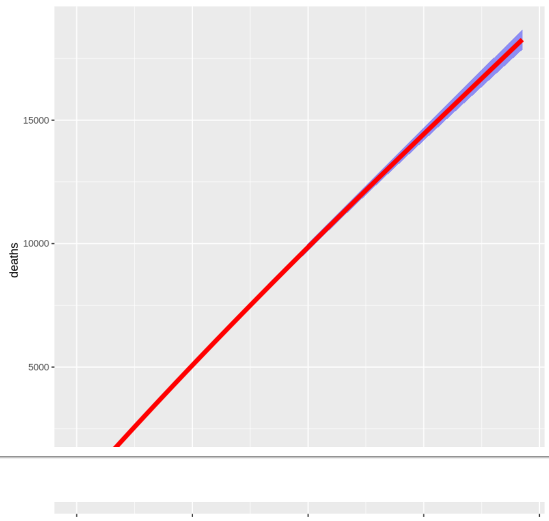



```
plot(TestingSet$deaths,Model.testing, col = "blue" )
```




```
ggplot(data = new_data) +
  geom_smooth(mapping = aes(x = confirmed, y = deaths), color="red", fill="blue", size=2)
```

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Conclusion:

This project shows us in real time how the cases of infection by the Civic 19 virus are evolving, it presents the most relevant analyzes either by states, infected and deaths, which is the maximum and minimum of each one, which is the probability of increasing or decreasing. Persuading self-care turns out to be the greatest impact that this type of information can have on society, as it raises awareness of the risks of not paying attention to biosafety care, the greatest limitation that was found is the few variables to analyze. In the future, different databases could be merged to obtain more information that can assist society in keeping it more informed.