COVID-19 Dataset Descriptive Analysis

Submitted to:

Dr Thi - Quynh Nguyen

Predictive Analytics - Quantitative Data

Report Prepared by:

Kyra Nicole Melenciano Feliz

March 23rd, 2022

Table of Contents

Abstract

This report centers around the descriptive analysis for the data set of COVID-19 outbreak and potential predictive features in the USA. The data set provides information related to the outbreak of COVID-19 disease in the United States, including data from its counties and states. The data set includes confirmed cases, deaths, and different features that may prove relevant to the pandemic dynamics.

The report is organized as follows. Section I contains the descriptive analysis, including a general description of the dataset, state and county wise analysis, distribution analysis and trends of the target variables, as well as correlation analysis and mean comparison between target groups. The results from this analysis will then be employed to aid the prediction of COVID-19 confirmed cases and deaths in the USA.

I. Descriptive Analysis

1.1 Description of Dataset

The data under consideration has COVID-19 outbreak related data for the USA for the period between Jan 2020 - Sep 2020. The data contains cases and deaths reported daily.

The dataset has a total of 56 columns and 562,129 observations. There are 46 features pertaining to various characteristics such as: Demographic, geographic, climatic ,traffic, public-health, social distancing norms and political characteristics.

Below is brief description of each variable:

Variable	**Brief Description**	**Data Type**
date	The date on which case was reported	Date
county_fips	Unique identifier for county name	Numerical
county_name	Name of the county	Categorical
state_fips	Unique identifier for state name	Numerical
state_name	Name of the state	Categorical
covid_19_confirmed_cases	Number of confirmed cases of Covid 19	Numerical
covid_19_deaths	Number of deaths due to covid_19	Numerical
social_distancing_total_grade	Social distancing grade at overall level	Categorical
social_distancing_encounters_grade	Social distancing grade based upon human encounter	Categorical
social_distancing_travel_distance_grade	Social distancing grade based upon distance traveled	Categorical

daily_state_test	Number of Covid-19 tests conducted at each day in state of the county	Numerical
precipitation	Daily precipitation	Numerical
temperature	Daily Temperature	Numerical
virus_pressure	Measure of virus transmission from neighboring counties	Numerical
total_population	Total Population	Numerical
female_percent	Total number of females divided by the total population	Numerical
area	Area in square miles	Numerical
population_density	Population per square mile	Numerical
latitude	latitude of the country	Numerical
longitude	longitude of the country	Numerical
hospital_beds_ratio	Number of beds in hospital w.r.t total population	Numerical
ventilator_capacity_ratio	Number of ventilators in hospital w.r.t total population	Numerical

icu_beds_ratio	Number of ICU beds in hospital w.r.t total population	Numerical
houses_density	No. of houses per square mile	Numerical
less_than_high_school_diploma	Distribution of education level of residents	Numerical
high_school_diploma_only		Numerical
some_college_or_higher		Numerical
total_college_population		Numerical
percent_smokers	percentage of residents who smoke	Numerical
percent_diabetes	percentage of residents who have diabetes	Numerical
Religious_congregation_ratio	Ratio of religious congregation	Numerical
political_party	Political party of each state(1 as democratic , 0 as republic)	Numerical
airport_distance	Distance from nearest airport	Numerical
passenger_load_ratio	Avg daily load of passengers at nearest airport	Numerical
meat_plants	Number of meat processing plants	Numerical
median_household_income	Household income of the residents	Numerical

percent_insured	Percentage of residents with insurance	Numerical
deaths_per_100000	Number of deaths reported per 100000	Numerical
gdp_per_capita	Gross domestic product per capita	Numerical
age related ratio	Ratio of cases among various age groups	Numerical
immigrant_student_ratio	Total number of immigrant students w.r.t total population	Numerical

To begin with analysis, we imported all the data into R studio and checked for summary statistics (see output at Appendix).

There were no null values in our dataset, so we further began to check for data accuracy issues. For the daily State test we could see some decimal values, as the number should be a whole number we rounded it to achieve a whole number.

For the age ratio, there were many observations where the sum total of all the categories was either more than 100 or less than 100, but this does not represent an issue as the ones over 100 did not go over 110 and the ones under 100 did not go under 90. This issue was ignored.

To reduce the number of age variables, the age ratio was grouped between those from 0-49 years old and 50 or more years old. This was implemented as there were many observations where the ratio across different age categories was the same thus for making a valid comparison we narrowed down the age into 2 categories which will help us in analyzing.

As our dataset was very large and there were a lot of parameters that were of fixed nature based on counties, we decided to aggregate the data for each county and state to obtain a summarized dataset on which we will be performing the predictive analysis. The final dataset had unique combinations of state and county.

Below is a summary of the tasks we did to achieve our aggregated data. These tasks were performed to obtain data aggregated on a county basis to account for those features that were fixed for each county

Variable	Aggregation Method
Confirmed cases	SUM
Deaths	SUM
Social_distancing_total_grade	MODE
Daily_state_test	SUM
Precipitation	AVG
Temperature	AVG
Virus_pressure	AVG

The variable age_cat was created to represent the age group in which the majority of the population from a county falls in. For the counties that have the majority of the population less than 50 years old, age_cat will be equal to “less than 50”. For the counties that have the majority of the population 50 years old or more, age_cat will be equal to “50 or more”. For the cases that the proportion is equal for the two groups, age_cat will be equal to “Same Proportion”.

Below is the list of variables that were taken into consideration for the analysis.

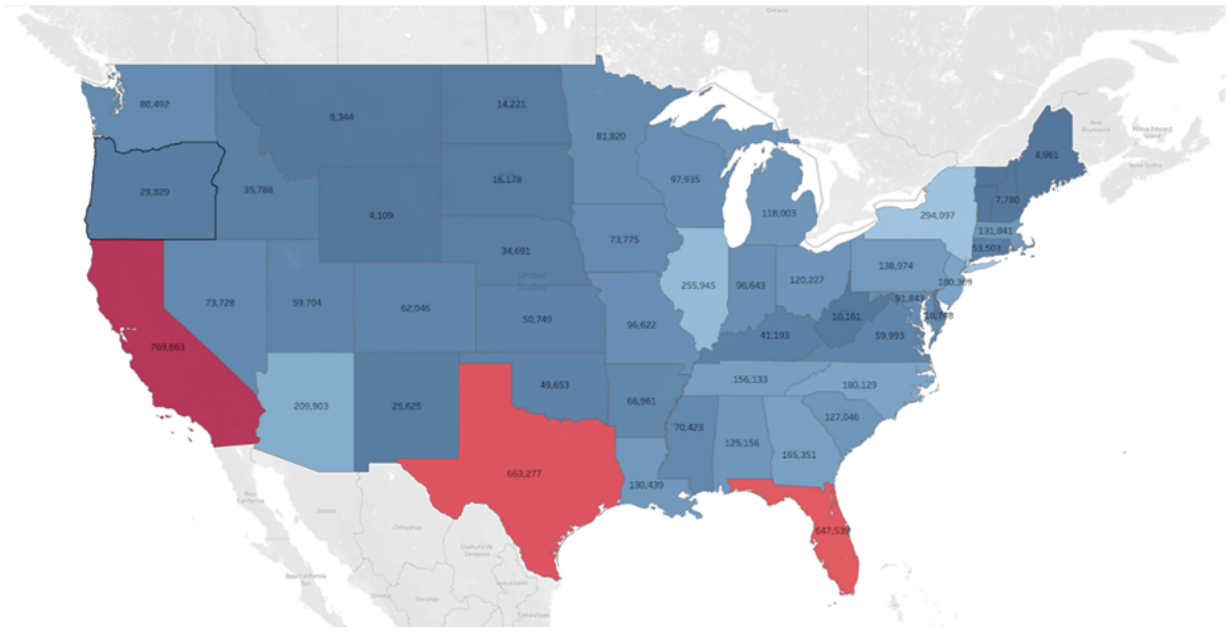
total_confirmed_cases	total_deaths
state_name	county_name
social_distancing_grade	total_state_test
precipitation	percent_smokers
temperature	percent_diabetes
virus_pressure	Religious_congregation_ratio
population_density	political_party
hospital_beds_ratio	meat_plants
ventilator_capacity_ratio	median_household_income

icu_beds_ratio	percent_insured
less_than_high_school_diploma	death_ratio
high_school_diploma_only	gdp_per_capita
some_college_or_higher	age_0_49
age_50_over	age_cat

The aggregated data set was left with 28 columns and 2,352 observations. This data is going to be used in the prediction of the total number of confirmed cases and total number of deaths related to COVID-19 (See Appendix for summary statistics and diagnostic analysis of the aggregated data set).

1.2 State Analysis

Since the first COVID case was reported in Washington State, on January 22, 2020, the virus spread rapidly. Ten months later, the three states that showed the highest level of confirmed cases were California, Texas and Florida.

Confirmed cases by state


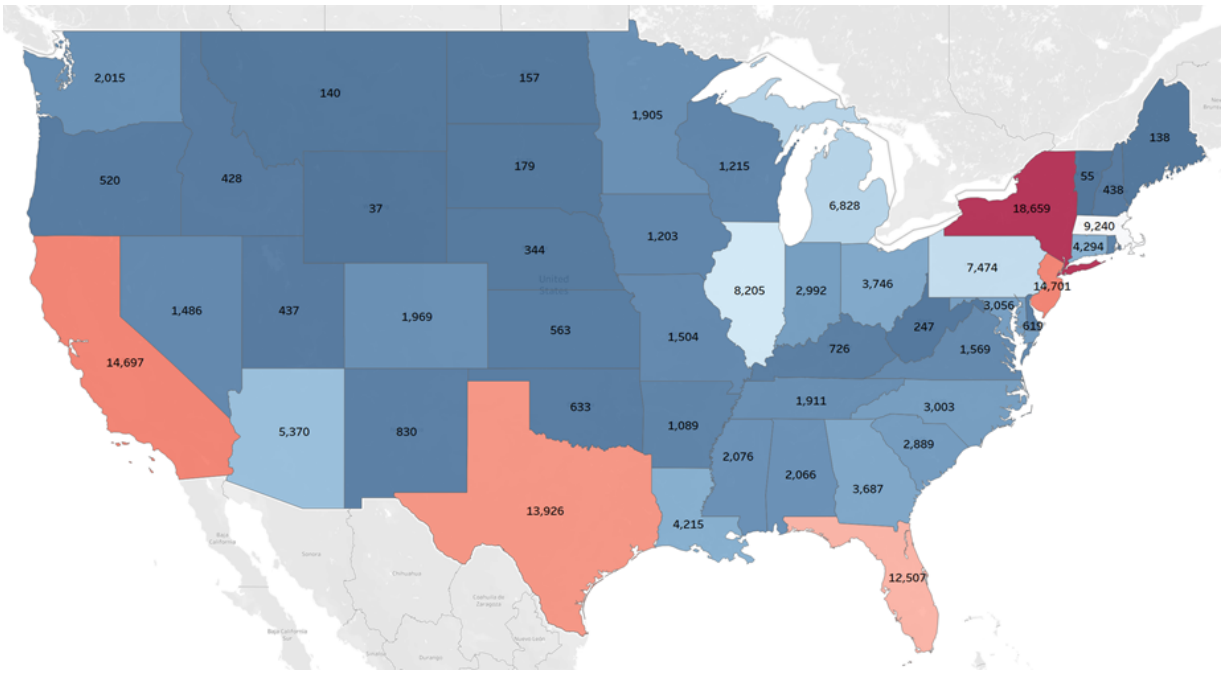
Confirmed Cases

State Name	Month of Date: January-2020	February-2020	March-2020	April-2020	May-2020	June-2020	July-2020	August-2020	September-2..	Grand Total
California	4	18	8,761	41,486	62,365	117,524	269,472	210,981	59,074	769,685
Florida	0	0	6,583	26,222	21,877	93,613	307,823	146,214	45,207	647,539
New Jersey	0	0	13,862	96,455	37,642	9,989	8,726	9,070	4,625	180,369
New York	0	0	49,971	147,225	43,527	15,341	15,589	13,902	8,542	294,097
Texas	0	0	3,557	23,854	35,528	93,727	259,678	186,150	60,783	663,277

During the first months of covid, the east coast was hit by a high number of cases, such as New York, that in April 2020 reached 147,225 confirmed cases, nearly five times California at that time. But, In August 2020, Florida and California were positioned as the states with more confirmed cases in the country. California is the State with the highest number of confirmed cases considering the whole dataset.

The next graph shows the number of deaths by State, highlighting the States that show a high number of deaths compared with the country. So, we can see that some states had been challenged to face the pandemic differently, considering that by March 27, 2020, the USA reached the highest number of COVID cases in the world[1].

Deaths by State



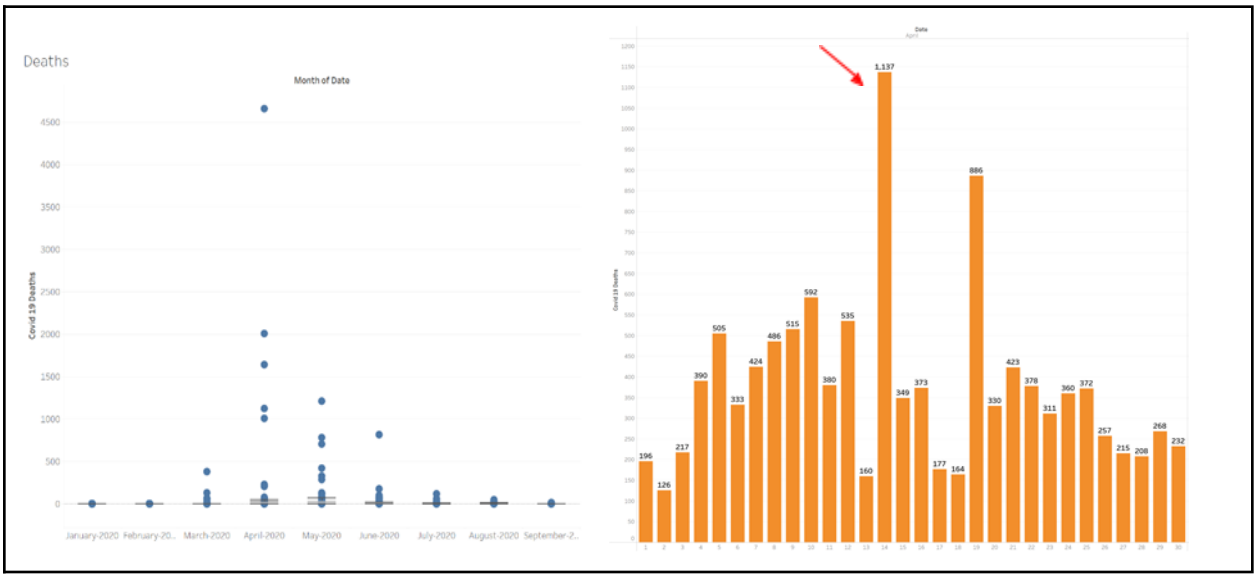
[1] https://www.investopedia.com/historical-timeline-of-covid-19-in-new-york-city-5071986

Death by State

State Name	January-2020	February-2020	March-2020	April-2020	May-2020	June-2020	July-2020	August-2020	September-2..	Grand Total
	Month of Date									
California	0	2	183	1,844	2,140	1,909	3,140	3,791	1,690	14,699
Florida	0	0	78	1,150	1,144	1,038	3,219	4,184	1,694	12,507
New Jersey	0	0	189	6,586	3,997	3,049	670	121	89	14,701
New York	0	0	699	11,299	4,501	1,584	334	175	67	18,659
Texas	0	0	48	711	850	713	4,020	5,733	1,851	13,926

Talking about the State of New York, which had faced the worst performance over all states in the USA. When the pandemic hit the USA, the first case of laboratory-confirmed COVID-19 in New York was diagnosed on February 29[2]; from that moment on, the State of New York quickly reached a number of 11.299 deaths by April 2020. According to Reuters[3], this dramatic increase was due to a massive recognition of people who died due to COVID-19 but were not tested and had the same symptoms. Additionally, there are other states, such as California, Texas, and Florida, which later on in 2020 started to report a high number of COVID deaths.

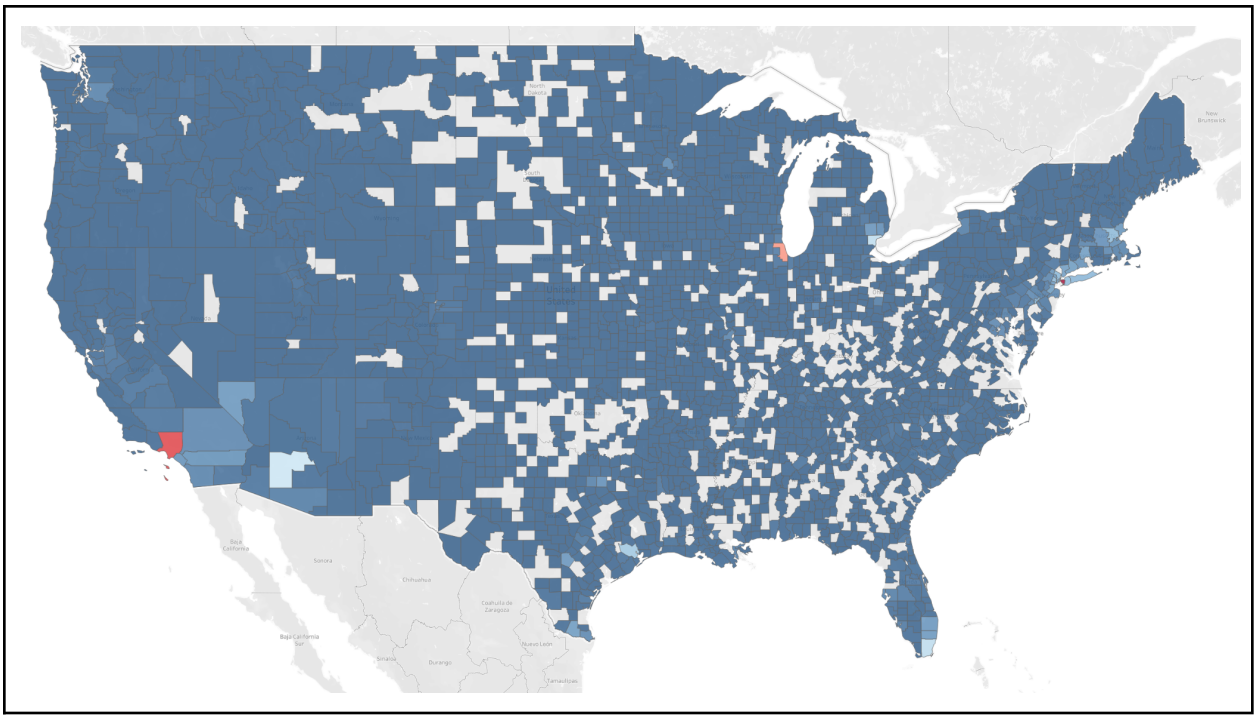
The graph on the left shows deaths by month, showing how covid deaths hit the State of New York in April 2020. On the right, we can see that only on April 14, 2020, more than 1.000 deaths were reported.



[2] https://www.cdc.gov/mmwr/volumes/69/wr/mm6946a2.htm
[3] https://www.reuters.com/article/us-health-coronavirus-usa-nyc-idUKKCN21W330

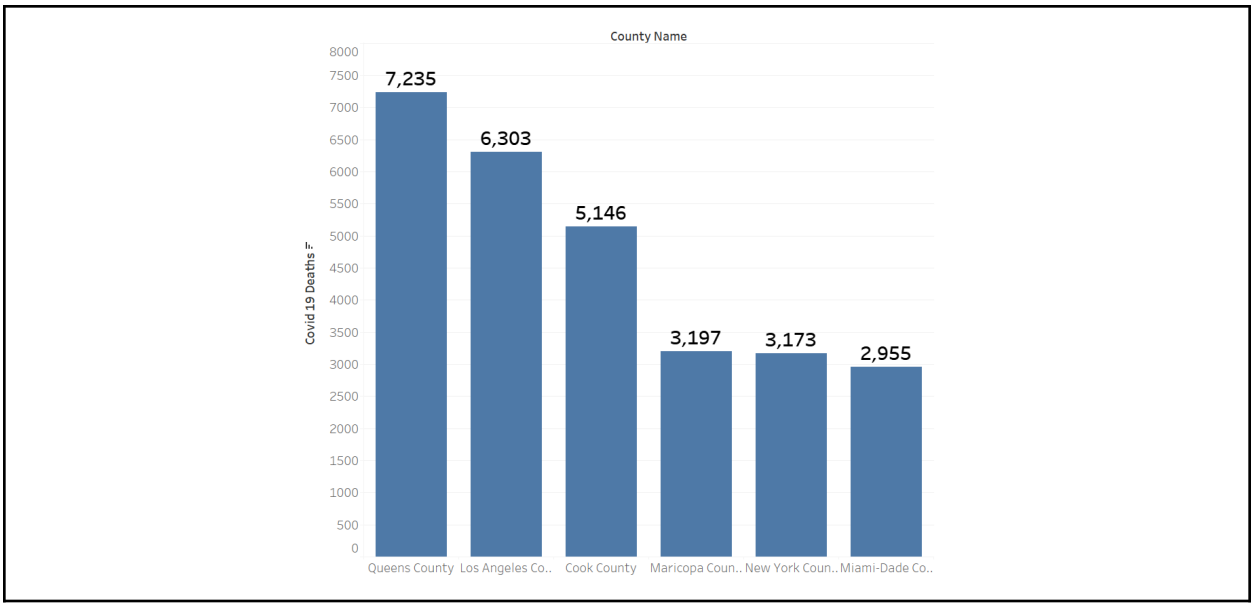
1.3 County Analysis

Based on the data, some USA counties are not present in the dataset. Those counties are represented in gray on the next graph. Besides this, most counties are present, and the data for the main states is accurate.



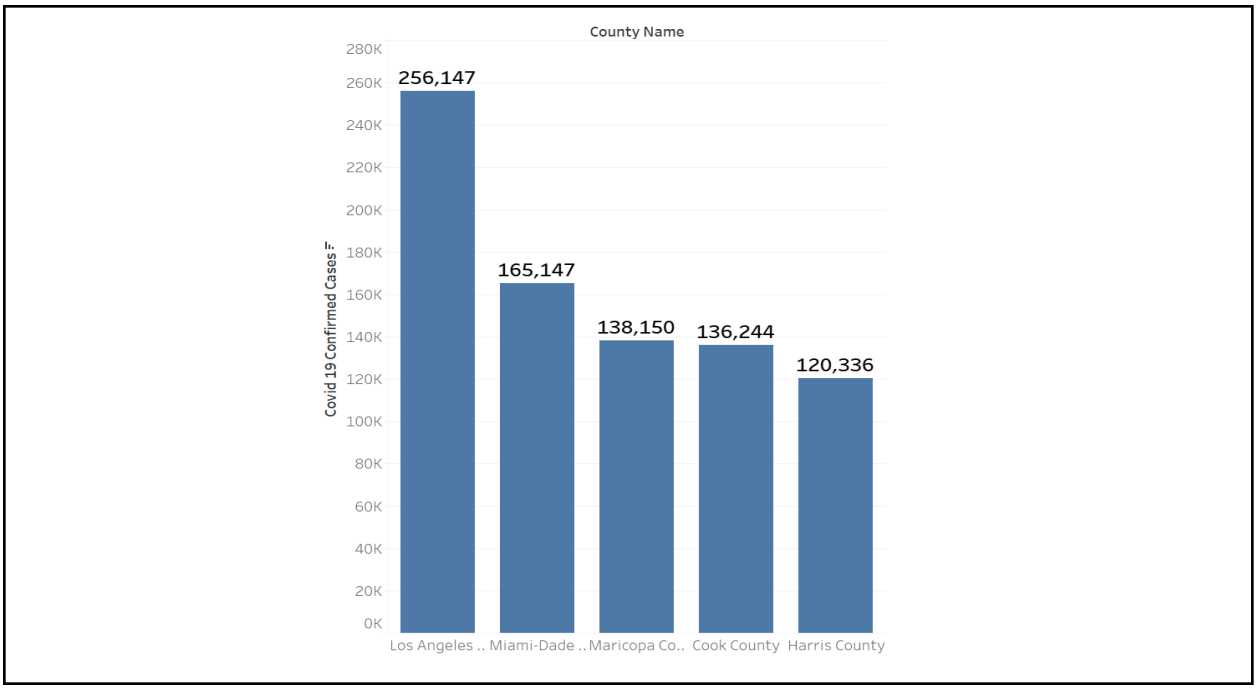
Deaths by county

The six counties with the highest number of deaths are Queens, Los Angeles, Cook, Maricopa, New York, and Miami. The names are repeated based on state, and that is due to these counties being the principal cities for the States that show the highest numbers in total, as mentioned before, for example, New York, Florida, and California.



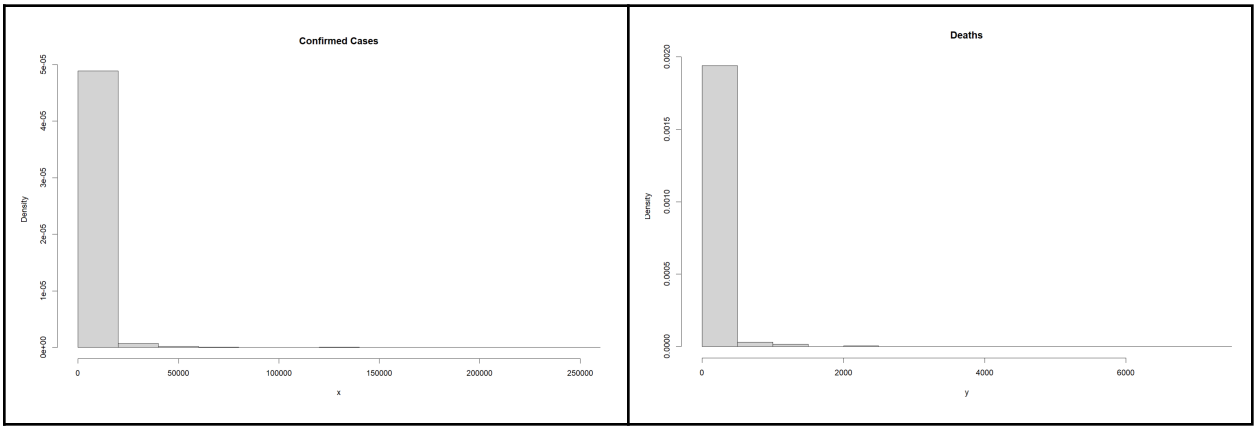
Confirmed Cases by county

Analyzing confirmed cases, we can see the same patterns between counties, showing that Los Angeles has the higher number of cases, followed by Miami. Both cities belong to the worst managed states in relation to case control, California and Florida.



1.4 Distribution Analysis

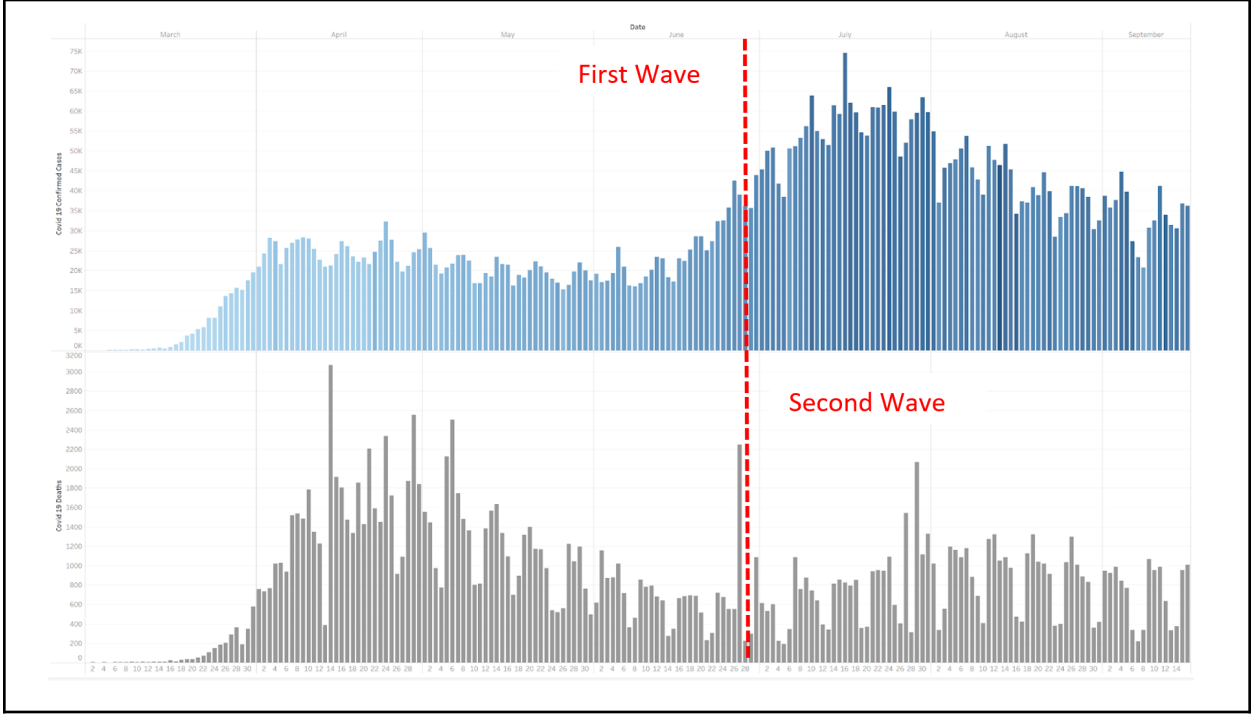
As it was expected, the variables confirmed cases and deaths, being a count, show a highly right-skewed distribution, which can be seen in the next graphs. This is due to most of the confirmed cases being concentrated before 5,000 and most of the deaths being concentrated before 2,000.



1.5 Trends

The covid cases and deaths were classified by waves, which can be seen in the next graph, where the First Wave covers from February 2020 to June 2020, and the second wave can be considered from July 2020 to onwards.

First and second wave in USA by month



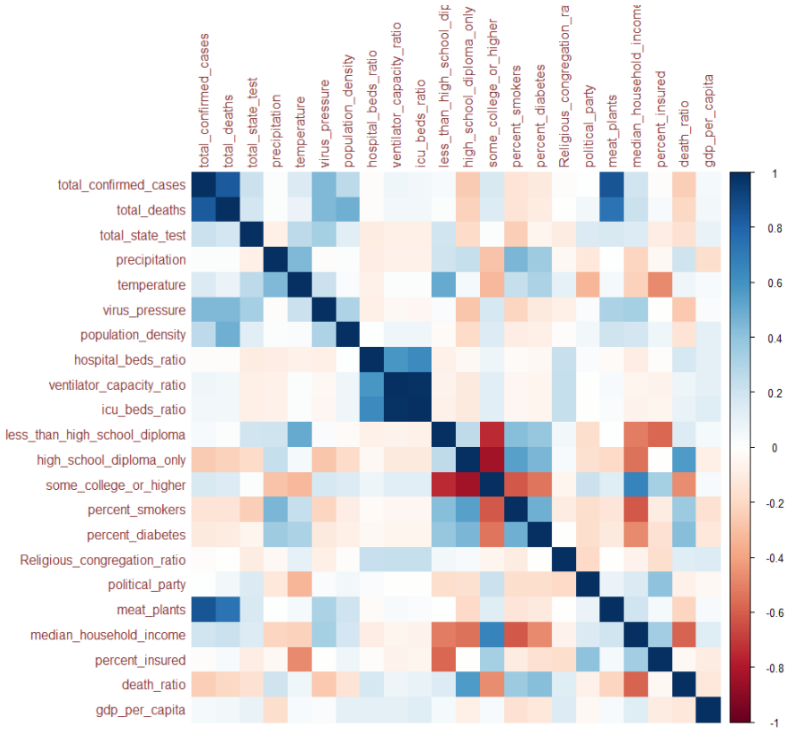
1.6 Correlation Analysis

Since the predictive analysis will concentrate on predicting the total number of confirmed cases and the total number of deaths, the correlation analysis was performed based on these two variables and calculating correlation between them and the rest of the variables.

Threshold for the correlation analysis can be seen below:

Absolute value of r	Strength of relationship
r <0.3	No relationship
0.3 < r < 0.5	Weak relationship
0.5 < r < 0.7	Moderate relationship
r > 0.7	Strong relationship

Below is a correlation plot for all the variables. The deeper blues and reds represent the most correlated variables.



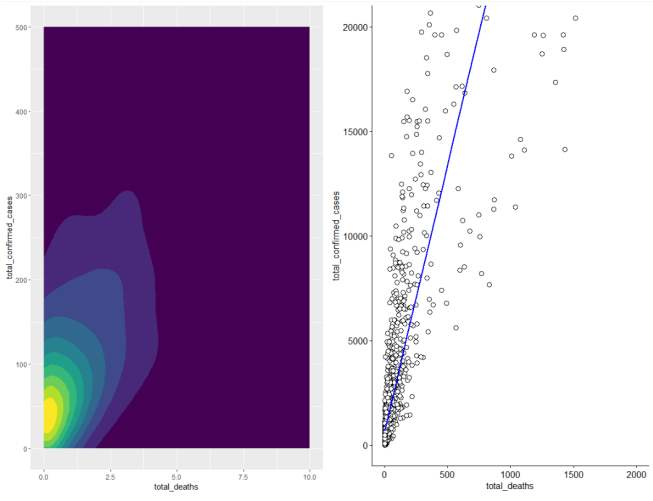
Confirmed Cases vs Predictors

Response Variable	Predictors	Correlation Coefficien
total_confirmed_cases	total_confirmed_cases	1
total_confirmed_cases	total_deaths	0.836986336
total_confirmed_cases	total_state_test	0.219533582
total_confirmed_cases	precipitation	0.018856855
total_confirmed_cases	temperature	0.150828397
total_confirmed_cases	virus_pressure	0.446364283
total_confirmed_cases	population_density	0.265403587
total_confirmed_cases	ventilator_capacity_ratio	0.061533147
total_confirmed_cases	less_than_high_school_diploma	0.031661636
total_confirmed_cases	icu_beds_ratio	0.059389383
total_confirmed_cases	high_school_diploma_only	-0.259495385
total_confirmed_cases	some_college_or_higher	0.160240336
total_confirmed_cases	percent_smokers	-0.146201051
total_confirmed_cases	percent_diabetes	-0.116143325
total_confirmed_cases	Religious_congregation_ratio	-0.013686015
total_confirmed_cases	meat_plants	0.853068903
total_confirmed_cases	median_household_income	0.195965533
total_confirmed_cases	percent_insured	-0.019193986
total_confirmed_cases	death_ratio	-0.245980009
total_confirmed_cases	gdp_per_capita	0.04115798
total_confirmed_cases	hospital_beds_ratio	-0.017183518

Based on the threshold presented in the last section and the table presented above, we can conclude that total_confirmed_cases is strongly correlated with total_deaths and meat_plants. Also, total_confirmed_cases has a weak relationship with the virus_pressure.

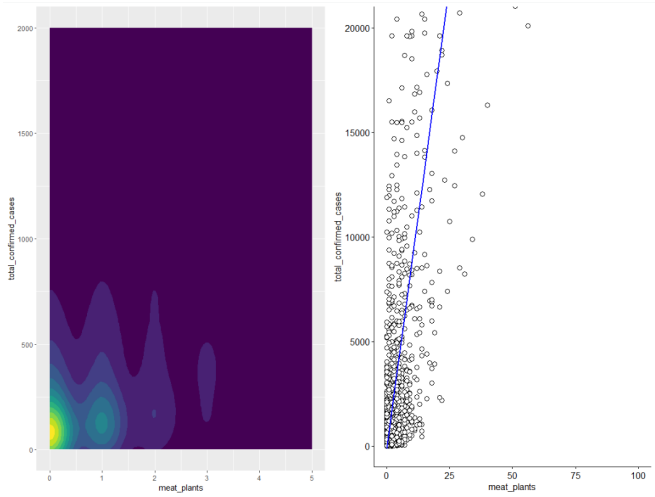
Since the correlation coefficient can only identify linear relationships, let's also look at the joint distribution of confirmed cases with the rest of the variables to confirm if there exists any other type of relationship between them. The graphs showcased here are those that show a relationship between the variables (See Appendix for the rest of the graphs).

Confirmed Cases vs Deaths



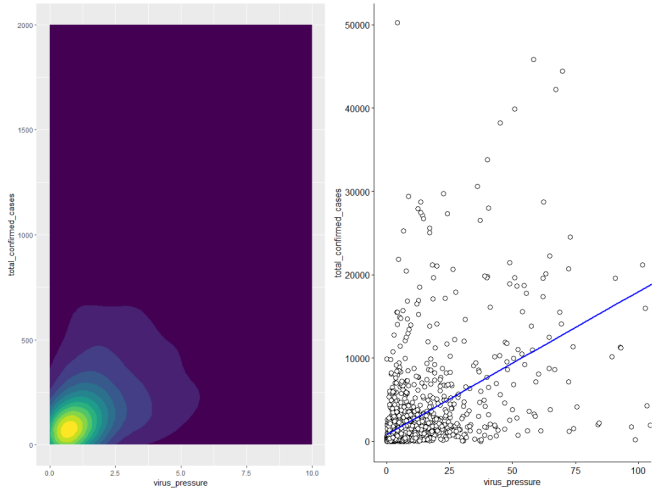
As expected from the correlation coefficient between total_confirmed_cases and total_deaths being 0.8, the strong linear relationship can also be appreciated from the plots above.

Confirmed Cases vs Meat Plants



The plots above showcase the strong positive linear relationship between total_confirmed_cases and meat_plants. This makes sense given that the correlation coefficient between the two variables was **0.85**.

Confirmed Cases vs Virus Pressure



From the plots above, it can be confirmed that the two variables have a weak linear relationship as was seen with the correlation coefficient of **0.4**.

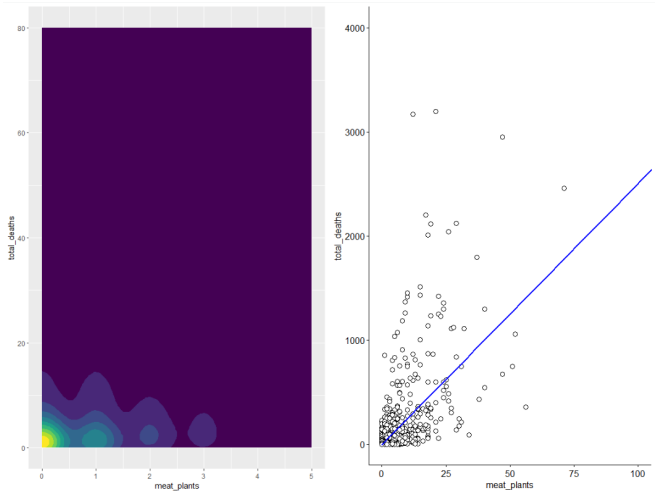
Based on the distribution of the observations for all the plots between the response variable the and predictors, we can say that none of the predictors has a quadratic impact on the response "**total_confirmed_cases**" variable.

Deaths vs Predictors

Response Variable	Predictors	Correlation Coefficient
total_deaths	total_confirmed_cases	0.836986336
total_deaths	total_deaths	1
total_deaths	total_state_test	0.187708417
total_deaths	precipitation	0.014316913
total_deaths	temperature	0.085825482
total_deaths	virus_pressure	0.442492651
total_deaths	population_density	0.484401291
total_deaths	ventilator_capacity_ratio	0.051892001
total_deaths	less_than_high_school_diploma	0.017681159
total_deaths	icu_beds_ratio	0.051259923
total_deaths	high_school_diploma_only	-0.223471553
total_deaths	some_college_or_higher	0.143498774
total_deaths	percent_smokers	-0.148502234
total_deaths	percent_diabetes	-0.102209951
total_deaths	Religious_congregation_ratio	-0.000349008
total_deaths	meat_plants	0.733851399
total_deaths	median_household_income	0.218373131
total_deaths	percent_insured	0.03063514
total_deaths	death_ratio	-0.203979085
total_deaths	gdp_per_capita	0.053595886
total_deaths	hospital_beds_ratio	-0.013803436

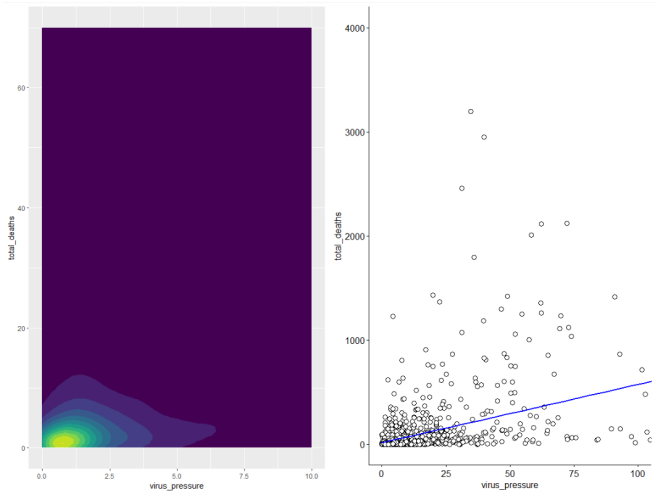
Based on the threshold presented in the last section and the table presented above, we can conclude that total_deaths is strongly correlated with total_confirmed_cases and meat_plants. Also, total_deaths has a weak relationship with virus_pressure and population_density.

Deaths vs Meat Plants



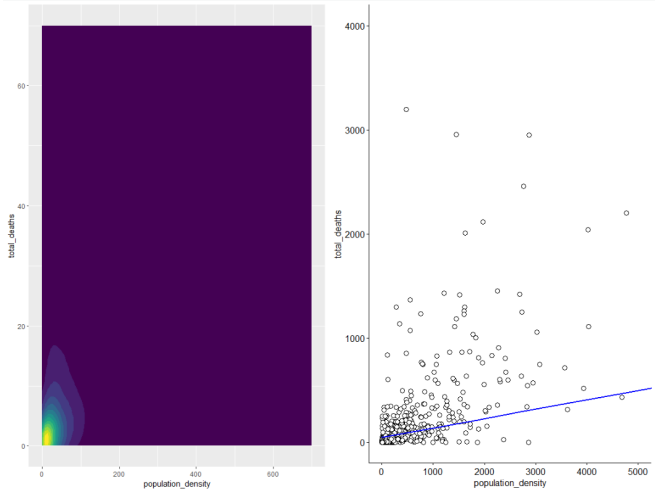
The correlation coefficient between total_deaths and virus_pressure is **0.73** which indicates that there is a strong positive linear relationship between the two variables. The plots above confirm that the relationship between the two variables is linear.

Deaths vs Virus Pressure



The plots above show a weak linear relationship between total_deaths and virus_pressure. This is confirmed by the correlation coefficient between the variables being **0.44**.

Deaths vs Population Density



Same as the case before, the correlation coefficient between total_deaths and population_density is **0.48** which indicates that there is a weak positive linear relationship between the two variables. The plots above confirm that the relationship between the two variables is linear.

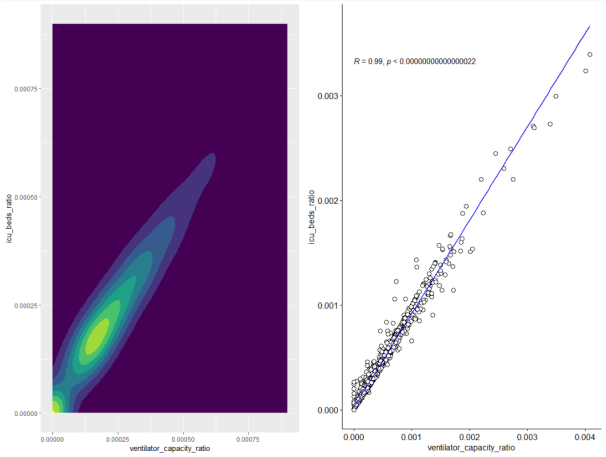
Based on the distribution of the observations for all the plots between response variable and predictors, we can say that none of the predictors has a quadratic impact on the response "**total_deaths**" variable.

Early Multicollinearity Detection

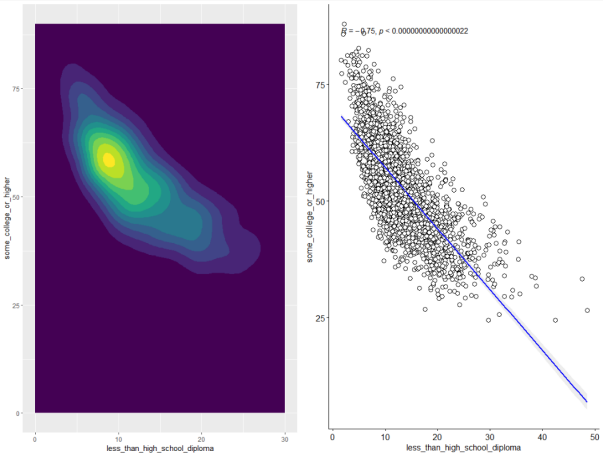
Multicollinearity occurs if independent variables in a model are highly correlated with each other. This means that one independent variable can be predicted by another independent variable in a regression model. Multicollinearity can be problematic in a regression model since one would not be able to recognize the individual effects of the independent variables on the dependent variable. One way of early detection of multicollinearity is by performing correlation analysis between predictors and identifying which predictors share high correlation and could later represent a multicollinearity issue. For this analysis, we will consider highly correlated pairs of predictors as those that have a correlation coefficient equal to or greater than **0.7**.

The plots below to the left show that ventilator_capacity_ratio and icu_beds_ratio are highly correlated. Their correlation coefficient is **0.98**. Another pair of highly correlated variables is shown in the plots to the right. The variables less_than_high_school_diploma and some_college_or_higher share a correlation coefficient of **-0.74.**

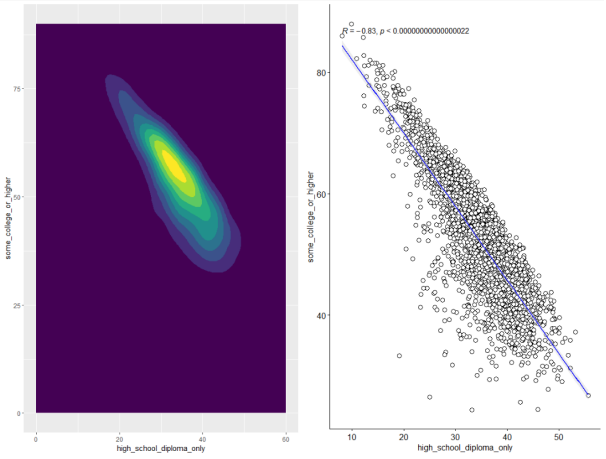
Ventilator Capacity vs ICU Beds



Less than High School vs Some College



High School vs Some College



The last pair of variables that shows a high correlation is shown in the plots above. The variables high_school_diploma_only and some_college_or_higher share a correlation coefficient of **-0.83.**

From this analysis, we could identify three pairs of predictors that could represent an issue with multicollinearity when included in regression predictive models. If the need arises, we will deal with the multicollinearity when performing the predictive analysis.

First Variable	**Second Variable**	**Correlation Coefficient**
ventilator_capacity_ratio	icu_beds_ratio	0.98
less_than_high_school_diploma	some_college_or_higher	-0.74
high_school_diploma_only	some_college_or_higher	-0.83

1.7 Mean Comparison

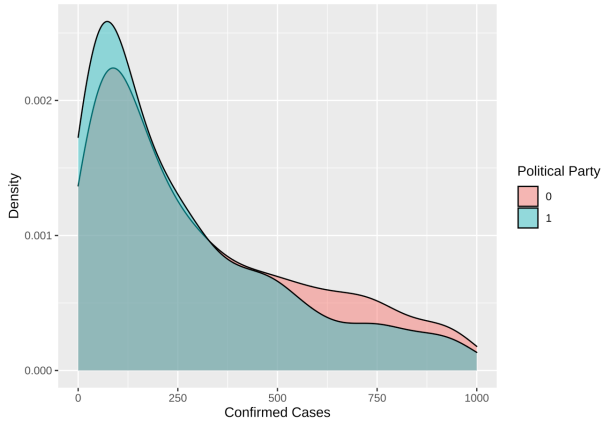
This descriptive analysis is meant to help in the later prediction of COVID-19 confirmed cases and deaths in the United States counties. To be able to know which variables have an effect on total_confirmed_cases and total_deaths, we performed mean comparison on categorical variables.

To test whether the State being from the democrat or republican party represented a difference in confirmed cases or deaths, we performed Wilcoxon Rank Test, a non-parametric alternative to the t-test used for data that is not normally distributed. The difference between the tests is that t-test takes the null hypothesis as equal means, whereas wilcoxon rank test takes the null hypothesis as equal medians.

To test whether the majority of the population from a county being in a specific age group represented a difference in confirmed cases or deaths, we first created a new variable that contained the age group in which the majority of the population from that county was in (less than 50 years old, more than 50 years old and same proportion for those counties were the proportion was 50-50). After creating that variable, we performed Kruskal-Wallis Test, a non-parametric alternative to ANOVA used for data that does not meet the assumptions for ANOVA.

Confirmed Cases vs Political Party

Before performing the test, we took a look at the distribution of confirmed cases based on political party, we can appreciate a small difference. However, the test is going to tell us whether this is a statistically significant difference.



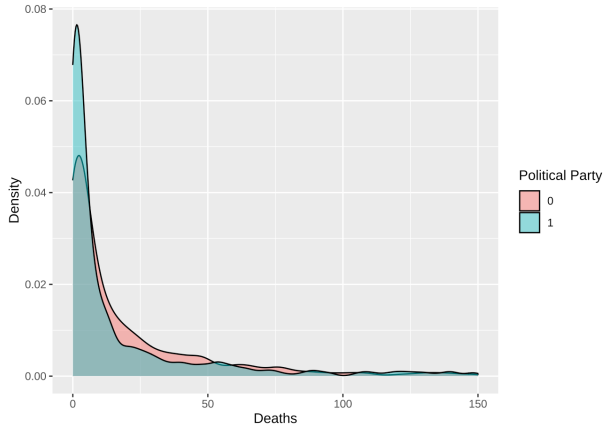

```
        Wilcoxon rank sum test with continuity correction

data:  total_confirmed_cases by political_party
W = 741868, p-value = 0.0007377
alternative hypothesis: true location shift is not equal to 0
```

H0: The medians of the two groups are equal.
Ha: The medians of the two groups are different.

The p-value of the test is 0.0007377, which allows us to reject the null hypothesis and confirms that the median of confirmed cases for democrat states is significantly different from the median of confirmed cases for republican states.

Deaths vs Political Party

Same as before, we took a look at the distributions of total deaths for both groups; we can appreciate a small difference that the test is going to validate as being statistically significant or not.

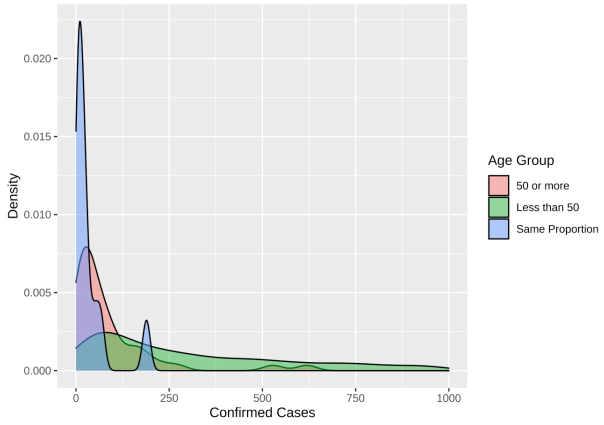



```
        Wilcoxon rank sum test with continuity correction

data:  total_deaths by political_party
W = 737474, p-value = 0.00183
alternative hypothesis: true location shift is not equal to 0
```

Again, the p-value of the test is less than 0.05, which allows us to reject the null hypothesis and confirms that the median of total deaths for democrat states is significantly different from the median of total deaths for republican states.

Confirmed Cases vs Age

As we did for the wilcoxon tests, we looked at the distribution of confirmed cases for the age groups. As can be seen on the graph below, the distributions are very different, however, the test is going to confirm whether this difference is statistically significant or not.



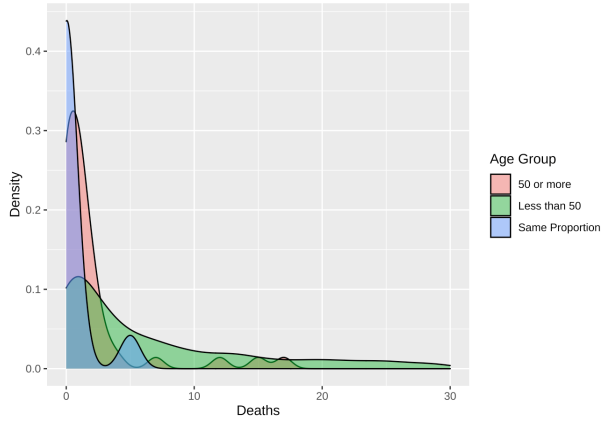

```
        Kruskal-Wallis rank sum test

data:  total_confirmed_cases by age_cat
Kruskal-Wallis chi-squared = 80.241, df = 2, p-value < 0.00000000000000022
```

H0: The mean ranks of the groups are equal.
Ha: The mean ranks of the groups are different.

The p-value of the test is less than 0.05, which allows us to reject the null hypothesis and confirms that the means of confirmed cases for counties with the majority of the population less than 50 years old, more than 50 years old and the same proportion, are different.

Deaths vs Age

Once again, we took a look at the distribution of total deaths for the counties with the majority of the population in the age groups of less than 50, 50 or more and the same proportion. The graph shows a different distribution for each group, the test is going to confirm whether that difference is statistically significant.




```
        Kruskal-Wallis rank sum test

data:  total_deaths by age_cat
Kruskal-Wallis chi-squared = 51.086, df = 2, p-value = 0.000000000008068
```

Again, the p-value of the test is less than 0.05, which allows us to reject the null hypothesis and confirms that the mean of total deaths is significantly different for each of the age groups.

Now that we know that the political party and the majority of the population from a county being in a specific age group have an effect on confirmed cases and deaths, these two variables will be considered to be included in our prediction models.

II. Appendix

Summary Statistics

```
> summary(imputed_data)
     date            county_fips    county_name          state_fips     state_name        covid_19_confirmed_cases
 Length:562128      Min.   : 1003   Length:562128      Min.   : 1.00   Length:562128      Min.   :    0.00
 Class :character   1st Qu.:19041   Class :character   1st Qu.:19.00   Class :character   1st Qu.:    0.00
 Mode  :character   Median :29162   Mode  :character   Median :29.00   Mode  :character   Median :    0.00
                    Mean   :30309                      Mean   :30.22                      Mean   :  10.44
                    3rd Qu.:45046                      3rd Qu.:45.00                      3rd Qu.:    3.00
                    Max.   :56039                      Max.   :56.00                      Max.   :9589.00
 covid_19_deaths    social_distancing_total_grade social_distancing_encounters_grade social_distancing_travel_distance_grade
 Min.   :  0.0000   Length:562128                 Length:562128                       Length:562128
 1st Qu.:  0.0000   Class :character              Class :character                    Class :character
 Median :  0.0000   Mode  :character              Mode  :character                    Mode  :character
 Mean   :  0.2973
 3rd Qu.:  0.0000
 Max.   :674.0000
 daily_state_test precipitation      temperature     virus_pressure     total_population   female_percent        area
 Min.   :     0   Min.   :   0.00   Min.   :-42.20   Min.   :   0.000   Min.   :    1227   Min.   :0.2684   Min.   :    2.5
 1st Qu.:    32   1st Qu.:   0.00   1st Qu.:  8.00   1st Qu.:   0.000   1st Qu.:   13119   1st Qu.:0.4946   1st Qu.:  471.7
 Median :  3556   Median :   0.00   Median : 17.50   Median :   1.000   Median :   32398   Median :0.5031   Median :  651.7
 Mean   :  9905   Mean   :  29.24   Mean   : 15.49   Mean   :   9.932   Mean   :  124755   Mean   :0.4994   Mean   : 1086.4
 3rd Qu.: 11616   3rd Qu.:  18.20   3rd Qu.: 23.90   3rd Qu.:   6.143   3rd Qu.:   87010   3rd Qu.:0.5100   3rd Qu.:  974.7
 Max.   :187926   Max.   :3048.00   Max.   : 65.60   Max.   :3178.667   Max.   :10105518   Max.   :0.5687   Max.   :35572.6
 population_density    latitude       longitude       hospital_beds_ratio ventilator_capacity_ratio icu_beds_ratio
 Min.   :    0.22   Min.   :19.60   Min.   :-159.75   Min.   :0.0000000   Min.   :0.0000000         Min.   :0.0000000
 1st Qu.:   18.39   1st Qu.:35.05   1st Qu.: -98.47   1st Qu.:0.0008365   1st Qu.:0.0001048         1st Qu.:0.0001188
 Median :   48.13   Median :38.87   Median : -91.15   Median :0.0017630   Median :0.0002189         Median :0.0002285
 Mean   :  257.72   Mean   :38.60   Mean   : -92.50   Mean   :0.0025211   Mean   :0.0003105         Mean   :0.0003046
 3rd Qu.:  136.96   3rd Qu.:42.25   3rd Qu.: -83.67   3rd Qu.:0.0030786   3rd Qu.:0.0003991         3rd Qu.:0.0003918
 Max.   :71340.39   Max.   :67.05   Max.   : -67.63   Max.   :0.0399348   Max.   :0.0040732         Max.   :0.0033943
 houses_density     less_than_high_school_diploma high_school_diploma_only some_college_or_higher total_college_population
 Min.   :    0.08   Min.   : 1.60                 Min.   : 8.10            Min.   :24.40          Min.   : 0.000000
 1st Qu.:    9.28   1st Qu.: 8.50                 1st Qu.:29.20            1st Qu.:45.80          1st Qu.: 0.000000
 Median :   23.22   Median :11.60                 Median :33.90            Median :53.30          Median : 0.005462
 Mean   :  113.82   Mean   :12.91                 Mean   :33.73            Mean   :53.35          Mean   : 0.387971
 3rd Qu.:   60.83   3rd Qu.:16.30                 3rd Qu.:38.80            3rd Qu.:60.60          3rd Qu.: 0.506748
 Max.   :38819.49   Max.   :48.50                 Max.   :55.60            Max.   :88.00          Max.   :10.586403
 percent_smokers  percent_diabetes Religious_congregation_ratio political_party  airport_distance  passenger_load_ratio
 Min.   : 5.909   Min.   : 1.80    Min.   :  5.00               Min.   :0.0000   Min.   :  2.675   Min.   : 0.00002
 1st Qu.:14.801   1st Qu.: 9.10    1st Qu.: 39.00               1st Qu.:0.0000   1st Qu.: 53.906   1st Qu.: 0.00157
 Median :16.673   Median :11.40    Median : 50.00               Median :0.0000   Median : 87.143   Median : 0.00610
 Mean   :17.169   Mean   :11.87    Mean   : 51.14               Mean   :0.4575   Mean   : 98.660   Mean   : 0.80283
 3rd Qu.:19.341   3rd Qu.:14.10    3rd Qu.: 62.00               3rd Qu.:1.0000   3rd Qu.:133.886   3rd Qu.: 0.04684
 Max.   :41.491   Max.   :31.00    Max.   :141.00               Max.   :1.0000   Max.   :383.144   Max.   :93.58695
  meat_plants      median_household_income percent_insured deaths_per_100000 gdp_per_capita       age_0_4
 Min.   : 0.000   Min.   : 26278          Min.   :66.25   Min.   : 235.4    Min.   :  10.61   Min.   : 2.000
 1st Qu.: 0.000   1st Qu.: 44565          1st Qu.:86.24   1st Qu.: 919.1    1st Qu.:  29.31   1st Qu.: 6.000
 Median : 1.000   Median : 51121          Median :89.78   Median :1109.2    Median :  39.12   Median : 6.000
 Mean   : 2.963   Mean   : 53410          Mean   :88.94   Mean   :1103.0    Mean   :  47.70   Mean   : 6.273
 3rd Qu.: 3.000   3rd Qu.: 59243          3rd Qu.:92.83   3rd Qu.:1287.8    3rd Qu.:  52.15   3rd Qu.: 7.000
 Max.   :288.000   Max.   :140382          Max.   :97.74   Max.   :2790.7    Max.   :2027.95   Max.   :13.000
```

```
   age_40_44        age_45_49        age_50_54        age_55_59        age_60_64        age_65_69        age_70_74
 Min.   : 3.000   Min.   : 3.00   Min.   : 3.000   Min.   : 3.000   Min.   : 2.000   Min.   : 2.000   Min.   : 1.000
 1st Qu.: 6.000   1st Qu.: 7.00   1st Qu.: 7.000   1st Qu.: 6.000   1st Qu.: 5.000   1st Qu.: 4.000   1st Qu.: 3.000
 Median : 6.000   Median : 7.00   Median : 8.000   Median : 7.000   Median : 6.000   Median : 5.000   Median : 4.000
 Mean   : 6.274   Mean   : 7.34   Mean   : 7.572   Mean   : 6.989   Mean   : 6.156   Mean   : 4.847   Mean   : 3.729
 3rd Qu.: 7.000   3rd Qu.: 8.00   3rd Qu.: 8.000   3rd Qu.: 8.000   3rd Qu.: 7.000   3rd Qu.: 5.000   3rd Qu.: 4.000
 Max.   :10.000   Max.   :11.00   Max.   :12.000   Max.   :12.000   Max.   :12.000   Max.   :11.000   Max.   :10.000
  age_75_79        age_80_84      age_85_or_higher      ...58      immigrant_student_ratio
 Min.   :1.000   Min.   :0.000   Min.   :0.00      Min.   : 96   Min.   :0.0000000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00      1st Qu.: 99   1st Qu.:0.0000000
 Median :3.000   Median :2.000   Median :2.00      Median :100   Median :0.0002023
 Mean   :2.923   Mean   :2.233   Mean   :2.14      Mean   :100   Mean   :0.0157857
 3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.00      3rd Qu.:101   3rd Qu.:0.0201942
 Max.   :7.000   Max.   :6.000   Max.   :8.00      Max.   :104   Max.   :0.5400094
>
```

Summary Statistics: Aggregated Data

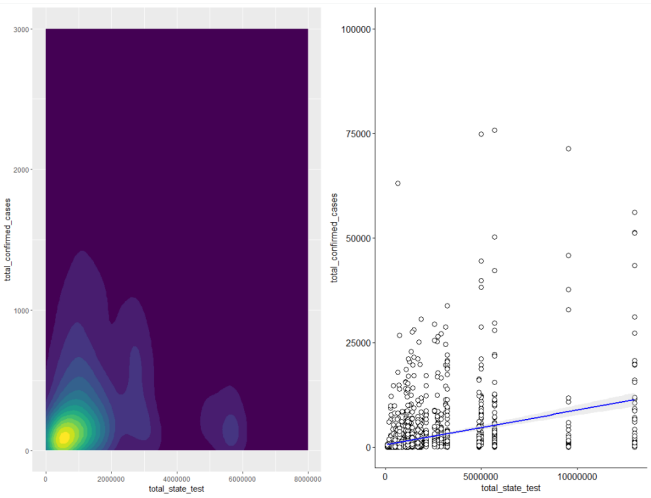
```
> summary(covid_aggregated_1_)
     ...1                state_name              county_name   total_confirmed_cases  total_deaths
 Min.   :   1.0   Texas         : 183   Washington County:  24   Min.   :     2.0      Min.   :   0.00
 1st Qu.: 588.8   Missouri      :  99   Franklin County  :  22   1st Qu.:   124.8      1st Qu.:   1.00
 Median :1176.5   Kansas        :  96   Jefferson County :  20   Median :   429.0      Median :   7.00
 Mean   :1176.5   Iowa          :  92   Jackson County   :  18   Mean   :  2494.1      Mean   :  71.06
 3rd Qu.:1764.2   North Carolina:  87   Lincoln County   :  18   3rd Qu.:  1414.8      3rd Qu.:  32.00
 Max.   :2352.0   Tennessee     :  75   Montgomery County:  15   Max.   :256148.0      Max.   :7235.00
                  (Other)       :1720   (Other)          :2235
 social_distancing_grade total_state_test   precipitation     temperature      virus_pressure
 F      :1116            Min.   :  100859   Min.   : 1.268   Min.   :-0.4213   Min.   :  0.000
 C      : 551            1st Qu.:  695099   1st Qu.:19.907   1st Qu.:12.2471   1st Qu.:  1.226
 D      : 353            Median : 1395231   Median :28.802   Median :15.3210   Median :  3.196
 D+     : 162            Mean   : 2367394   Mean   :29.245   Mean   :15.4895   Mean   :  9.932
 C-     : 122            3rd Qu.: 2895425   3rd Qu.:38.006   3rd Qu.:18.7345   3rd Qu.:  8.216
 D-     :  40            Max.   :13000035   Max.   :74.126   Max.   :26.2937   Max.   :425.780
 (Other):   8
 population_density hospital_beds_ratio ventilator_capacity_ratio icu_beds_ratio
 Min.   :    0.22   Min.   :0.0000000   Min.   :0.0000000         Min.   :0.0000000
 1st Qu.:   18.39   1st Qu.:0.0008365   1st Qu.:0.0001048         1st Qu.:0.0001188
 Median :   48.13   Median :0.0017630   Median :0.0002189         Median :0.0002285
 Mean   :  257.72   Mean   :0.0025211   Mean   :0.0003105         Mean   :0.0003046
 3rd Qu.:  136.96   3rd Qu.:0.0030786   3rd Qu.:0.0003991         3rd Qu.:0.0003918
 Max.   :71340.39   Max.   :0.0399348   Max.   :0.0040732         Max.   :0.0033943

 less_than_high_school_diploma high_school_diploma_only some_college_or_higher percent_smokers
 Min.   : 1.60                 Min.   : 8.10            Min.   :24.40          Min.   : 5.909
 1st Qu.: 8.50                 1st Qu.:29.20            1st Qu.:45.80          1st Qu.:14.801
 Median :11.60                 Median :33.90            Median :53.30          Median :16.673
 Mean   :12.91                 Mean   :33.73            Mean   :53.35          Mean   :17.169
 3rd Qu.:16.30                 3rd Qu.:38.80            3rd Qu.:60.60          3rd Qu.:19.341
 Max.   :48.50                 Max.   :55.60            Max.   :88.00          Max.   :41.491

 percent_diabetes Religious_congregation_ratio political_party  meat_plants      median_household_income
 Min.   : 1.80    Min.   :  5.00               0:1276          Min.   :  0.000   Min.   : 26278
 1st Qu.: 9.10    1st Qu.: 39.00               1:1076          1st Qu.:  0.000   1st Qu.: 44565
 Median :11.40    Median : 50.00                               Median :  1.000   Median : 51121
 Mean   :11.87    Mean   : 51.14                               Mean   :  2.963   Mean   : 53410
 3rd Qu.:14.10    3rd Qu.: 62.00                               3rd Qu.:  3.000   3rd Qu.: 59243
 Max.   :31.00    Max.   :141.00                               Max.   :288.000   Max.   :140382

 percent_insured  death_ratio     gdp_per_capita       age_0_49      age_50_over
 Min.   :66.25   Min.   : 235.4   Min.   :  10.61   Min.   :39.00   Min.   :14.00
 1st Qu.:86.24   1st Qu.: 919.1   1st Qu.:  29.31   1st Qu.:60.00   1st Qu.:32.00
 Median :89.78   Median :1109.2   Median :  39.12   Median :64.00   Median :36.00
 Mean   :88.94   Mean   :1103.0   Mean   :  47.70   Mean   :63.45   Mean   :36.59
 3rd Qu.:92.83   3rd Qu.:1287.8   3rd Qu.:  52.15   3rd Qu.:68.00   3rd Qu.:40.00
 Max.   :97.74   Max.   :2790.7   Max.   :2027.95   Max.   :86.00   Max.   :61.00

           age_cat
 50 or more     :  49
 Less than 50   :2291
 Same Proportion:  12
```

Diagnostic Analysis: Aggregated Data

	variables	types	missing_count	missing_percent	unique_count	unique_rate
1	...1	numeric	0	0	2352	1.0000000000
2	state_name	character	0	0	50	0.0212585034
3	county_name	character	0	0	1479	0.6288265306
4	total_confirmed_cases	numeric	0	0	1363	0.5795068027
5	total_deaths	numeric	0	0	294	0.1250000000
6	social_distancing_grade	character	0	0	10	0.0042517007
7	total_state_test	numeric	0	0	50	0.0212585034
8	precipitation	numeric	0	0	2282	0.9702380952
9	temperature	numeric	0	0	2332	0.9914965986
10	virus_pressure	numeric	0	0	2277	0.9681122449
11	population_density	numeric	0	0	2352	1.0000000000
12	hospital_beds_ratio	numeric	0	0	1955	0.8312074830
13	ventilator_capacity_ratio	numeric	0	0	1946	0.8273809524
14	icu_beds_ratio	numeric	0	0	1954	0.8307823129
15	less_than_high_school_diploma	numeric	0	0	290	0.1232993197
16	high_school_diploma_only	numeric	0	0	357	0.1517857143
17	some_college_or_higher	numeric	0	0	466	0.1981292517
18	percent_smokers	numeric	0	0	2352	1.0000000000
19	percent_diabetes	numeric	0	0	204	0.0867346939
20	Religious_congregation_ratio	numeric	0	0	98	0.0416666667
21	political_party	numeric	0	0	2	0.0008503401
22	meat_plants	numeric	0	0	45	0.0191326531
23	median_household_income	numeric	0	0	2284	0.9710884354
24	percent_insured	numeric	0	0	2352	1.0000000000
25	death_ratio	numeric	0	0	2100	0.8928571429
26	gdp_per_capita	numeric	0	0	1919	0.8159013605
27	age_0_49	numeric	0	0	44	0.0187074830
28	age_50_over	numeric	0	0	44	0.0187074830
29	age_cat	character	0	0	3	0.0012755102

Correlation Analysis

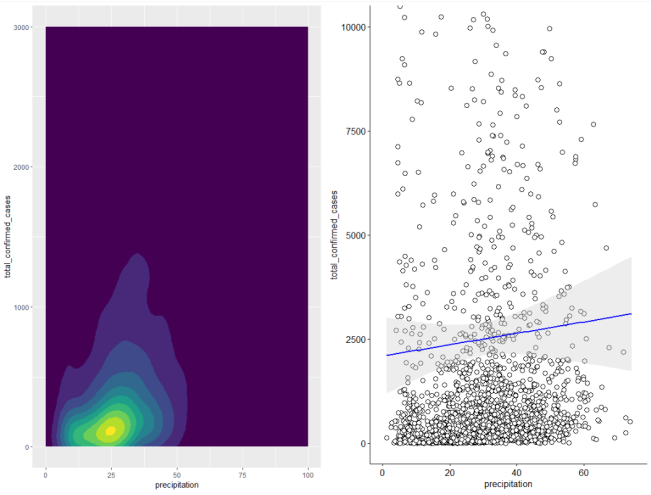
Total_confirmed_cases and total_state_test:



Correlation coefficient value between total_confirmed_cases and total_state_test is **0.21** which indicates that there is no linear relationship between the two variables. Also, from the above plots, no relationship between the two variables can be seen.

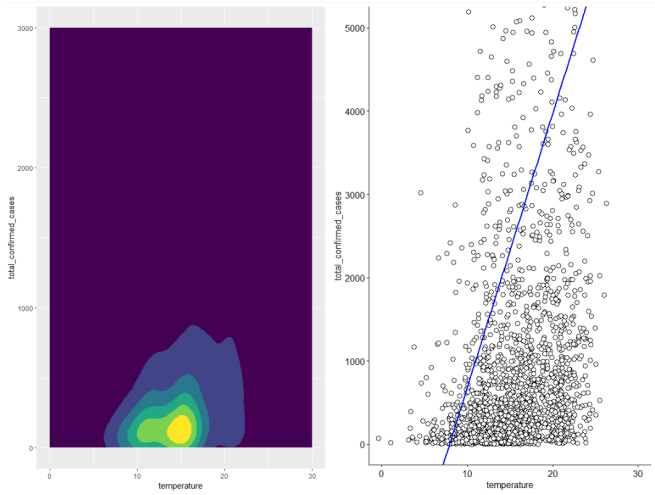
Total_confirmed_cases and precipitation:

Correlation coefficient value between total_confirmed_cases and precipitation is **0.018** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



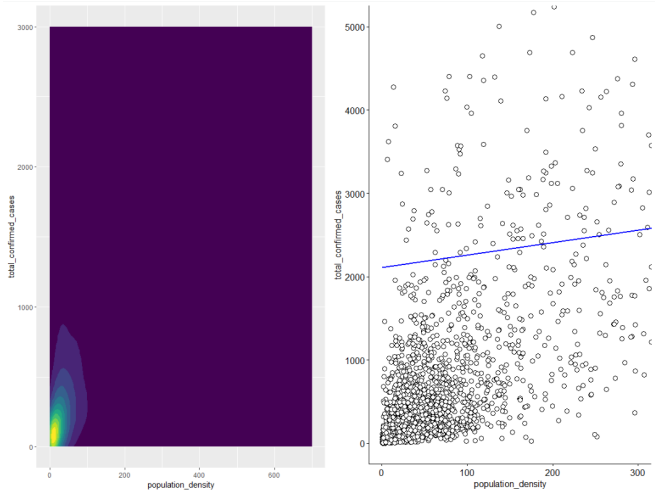
Total_confirmed_cases and temperature:

Correlation coefficient value between total_confirmed_cases and temperature is **0.15** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



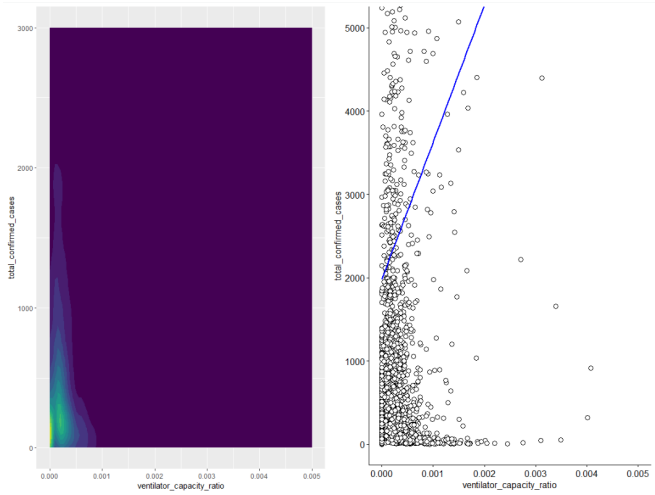
Total_confirmed_cases and population_density:

Correlation coefficient value between total_confirmed_cases and population_density is **0.26** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



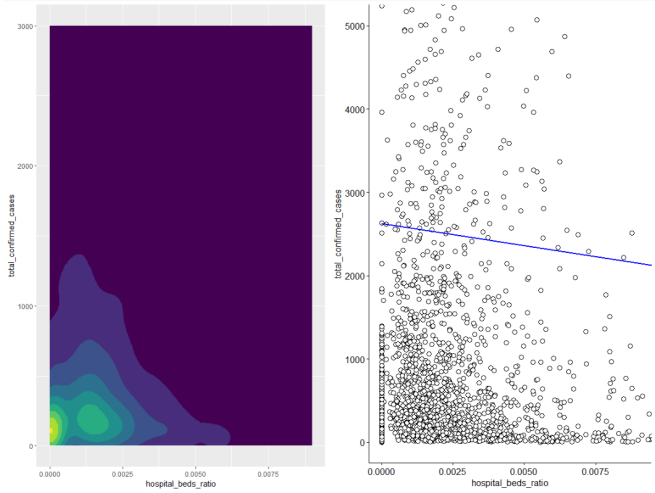
Total_confirmed_cases and ventilator_capacity_ratio:

Correlation coefficient value between total_confirmed_cases and ventilator_capacity_ratio is **0.06** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



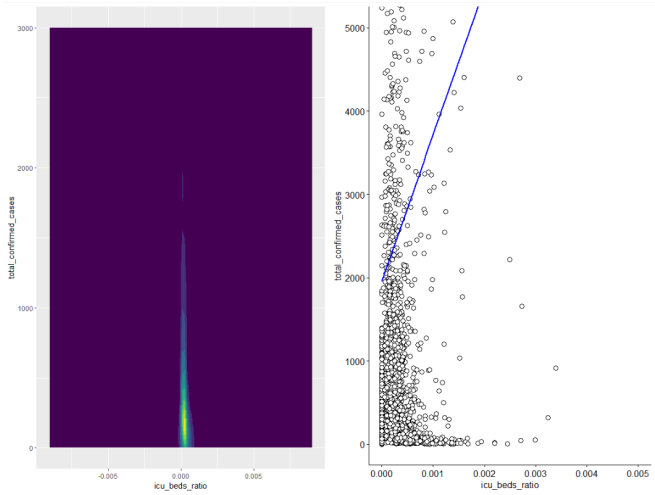
Total_confirmed_cases and hospital_beds_ratio:

Correlation coefficient value between total_confirmed_cases and hospital_beds_ratio is **-0.01** which indicates that variables are impacting negatively to each other, but as the coefficient value is less than 0.3, there is no association between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



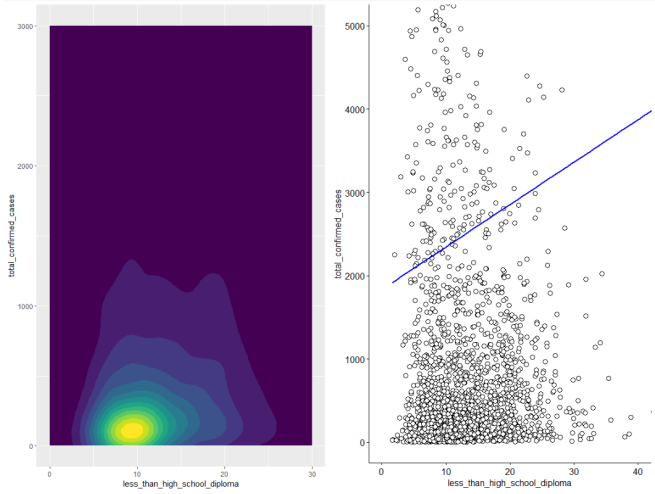
Total_confirmed_cases and icu_beds_ratio:

Correlation coefficient value between total_confirmed_cases and icu_beds_ratio is **0.05** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



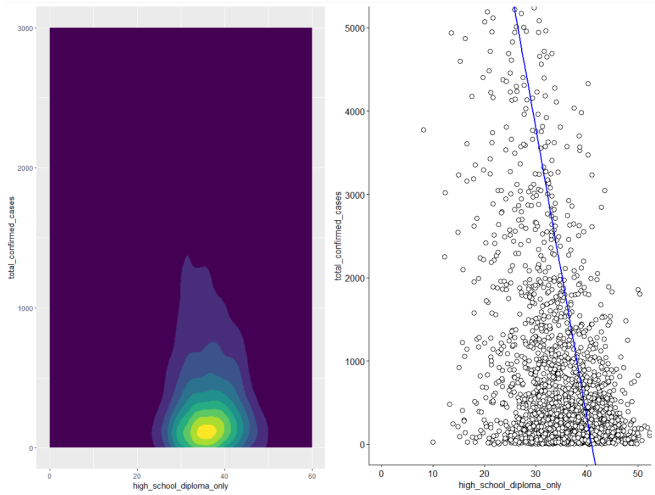
Total_confirmed_cases and less_than_high_school_diploma:

Correlation coefficient value between total_confirmed_cases and less_than_high_school_diploma is **0.03** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



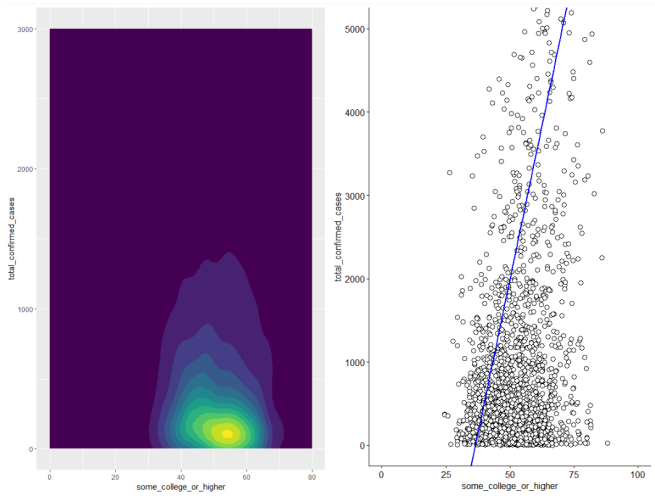
Total_confirmed_cases and high_school_diploma_only:

Correlation coefficient value between total_confirmed_cases and high_school_diploma_only is **-0.25** which indicates that variables are impacting negatively to each other, but as the coefficient value is less than 0.3, there is no association between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



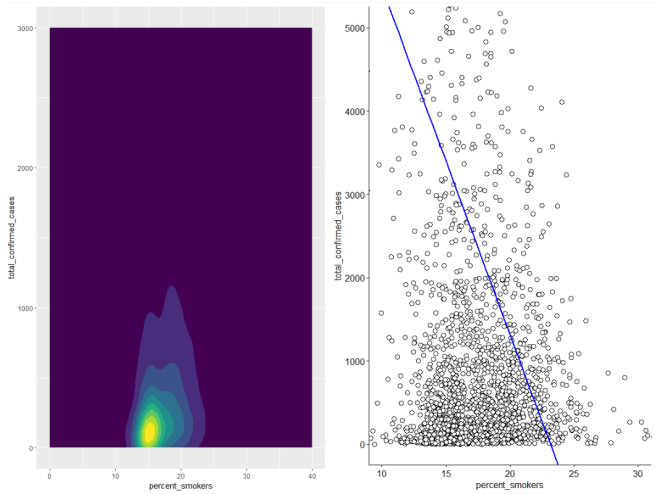
Total_confirmed_cases and some_college_or_higher:

Correlation coefficient value between total_confirmed_cases and less_than_high_school_diploma is **0.16** which indicates that there is no linear relationship between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



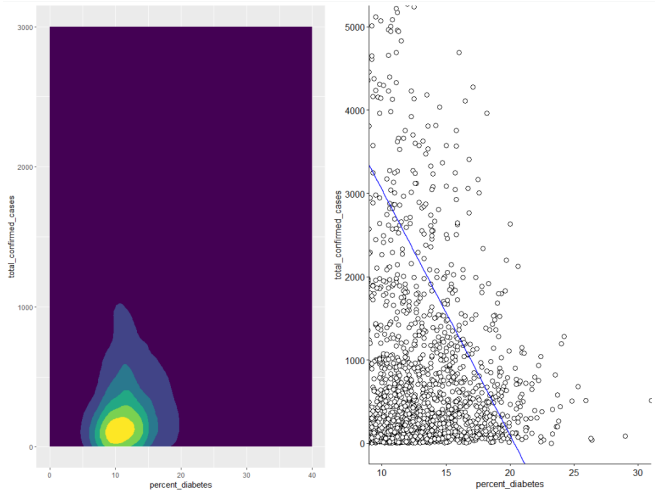
Total_confirmed_cases and percent_smokers:

Correlation coefficient value between total_confirmed_cases and percent_smokers is -0.14 which indicates that variables are impacting negatively to each other, but as the coefficient value is less than 0.3, there is no association between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



Total_confirmed_cases and percent_diabetes:

Correlation coefficient value between total_confirmed_cases and percent_diabetes is **-0.11** which indicates that variables are impacting negatively to each other, but as the coefficient value is less than 0.3, there is no association between the two variables. Also, from the below plots, no relationship between the two variables can be seen.



Total_confirmed_cases and Religious_congregation_ratio:

Correlation coefficient value between total_confirmed_cases and Religious_congregation_ratio is **-0.01** which indicates that variables are impacting negatively to each other, but as the coefficient value is less than 0.3, there is no association between the two variables. Also, from the below plots, no relationship between the two variables can be seen.