**Teamwork Covid**

[**1. Introduction** 1](#_Toc98511451)

[**2. Exploratory Data Analysis** 1](#_Toc98511452)

[**2.1 Missing values and accuracy** 2](#_Toc98511453)

[**2.2 Sampling data** 2](#_Toc98511454)

[**2.3 Climate Features** 3](#_Toc98511455)

[**2.4 Demographic Features** 6](#_Toc98511456)

[**2.5 Morbility Features** 7](#_Toc98511457)

[**2.6 Immigration Features** 8](#_Toc98511458)

[**2.7 Educational Features** 8](#_Toc98511459)

[**2.8 Economical features** 11](#_Toc98511460)

[**3. Summary** 13](#_Toc98511461)

[**Bibliography** 14](#_Toc98511462)

**1. Introduction**

The aim of this study is to conduct a descriptive analysis on a Covid 19 dataset on the United States available in the GitHub repository <https://github.com/network-and-Data-Science-IUT/covid-19>, and identify possible key driving factors that helps us to create a predict model for new Covid 19 cases or deaths. The group of features analyzed include Geolocation, climate, public health, political, economic, educational attainment, demographic, mobility and morbidity.

The dataset provides information that related to the outbreak of COVID-19 disease in the United States, including data from each of 1961 counties from the beginning of the outbreak from January 2020 until June 2021. This dataset has various numerical and categorical variables that relate to COVID in specific states and counties in the USA. There is a date component to this set as well. The details of various features are described in the article written by Haratian et al. (2021).

# **2. Exploratory Data Analysis**

We used a dataset of Covid 19 cases in the USA and its potential features[[1]](#footnote-1). In total we found 992.266 observations and 64 variables. This analysis is focused on the features education level, climate, demographic and morbidity, which are typify in the Appendix 1.

The computing platform used were Jupyter Notebook (Python) and SAS Studio. The codes use for the analysis in Juptyter are found in the Appendix A (file Appendix\_A.pdf) and the SAS Code, in Appendix B (file Appendix\_B.pdf).

## **2.1 Missing values and accuracy**

According to the article Dataset of Covid-19 outbreak and potential predictive features in USA, the data was already processed, imputing the missing values, as well as fixing the abnormal observations like negative counties values.[[2]](#footnote-2) For this reason, the variables were validate and it was not found any missing value or observation to be fixed.

The variable data was also check to validate missing data and it was found that it was continuous:

Chart, line chart

Description automatically generated

## **2.2 Sampling data**

To understand the behavior of the features already selected on the Covid 19 cases and deaths, we filtered the data with the top states in the USA based on the population density:

Table

Description automatically generated

**Chart, bar chart

Description automatically generated**

A **T-Test** was conducted to validate if the sample qualify as a representation of the population. It was found that the p-value was less than the alpha value, that is equal to 1 minus the confidence interval, meaning that we reject the Null Hypothesis of no difference and say that with a high degree of confidence (95%) the true difference in means is not equal to zero.

The mean in covid cases for the sample with the 10 states is equal to **63.04** and for the total dataset the mean is equal to **29.7**

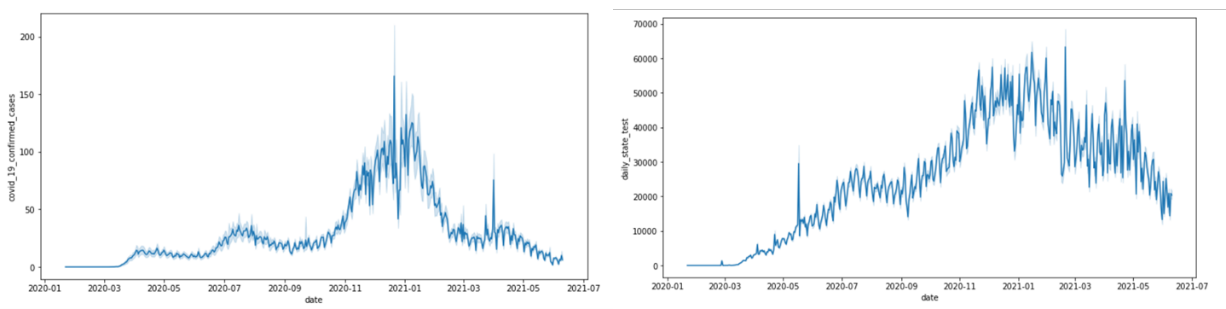
## **2.3 Climate Features**

The majority of the cases are detected when temperature is between -15 to 40, as the graphic shows:

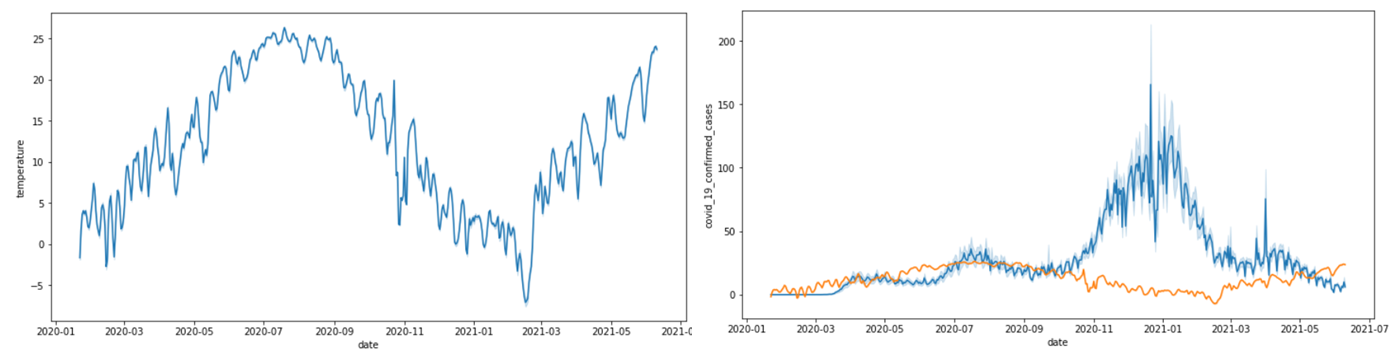
Chart, scatter chart

Description automatically generated

It was checked also that the daily tests increased based on Covid cases increased:



During the winter the cases increased, maybe because people staying mostly inside public places which was a major factor of case spread, directly they are not correlated. It was also found that precipitation has no relation.



The covid cases are not accumulated. Also all states have a similar pattern stating that major spread during winters.Chart, histogram

Description automatically generated

With the Covid 19 deaths the patterns are similar to covid cases. There is also a spike in death during initial spread as well as during the winter season:

Chart, bar chart, histogram

Description automatically generated

It was also validated the relationship between the Covid Cases and the Covid Deaths, and it was found that there is a correlation. However, the relationship is low but periodical.

Chart, scatter chart

Description automatically generated

It was found a correlation between Covid Cases with Covid Test. the relationship is high in some cases and low in others. it maybe because of the difference in the population density:

Chart, scatter chart

Description automatically generated

There is no direct relationship between temperature vs covid cases:

Chart, scatter chart

Description automatically generated

There is a negative correlation of covid cases with covid deaths . the relationship is low as it is periodical:

Chart, scatter chart

Description automatically generated

It was not found any relationship between precipitation and daily test:  
Chart

Description automatically generated

## **2.4 Demographic Features**

It was found a very high correlation between Virus Pressure and Covid Cases:

Chart, scatter chart

Description automatically generated

There is also a significant relation but not directly correlation between Virus Pressure and Covid Deaths:

Chart, scatter chart

Description automatically generated

There was found a strong correlation between Virus Pressure and Daily Tests:

Chart, scatter chart

Description automatically generated

**Total population, Death ratio and Covid Cases**

Chart, bar chart

Description automatically generated

## **2.5 Morbility Features**

Death ratio, smokers, and diabetes have relationship among themselves.

Chart, bar chart, histogram

Description automatically generated

## **2.6 Immigration Features**

It was found that the more distance to the airport, the less the immigration ratio:

Chart, bar chart, histogram

Description automatically generated

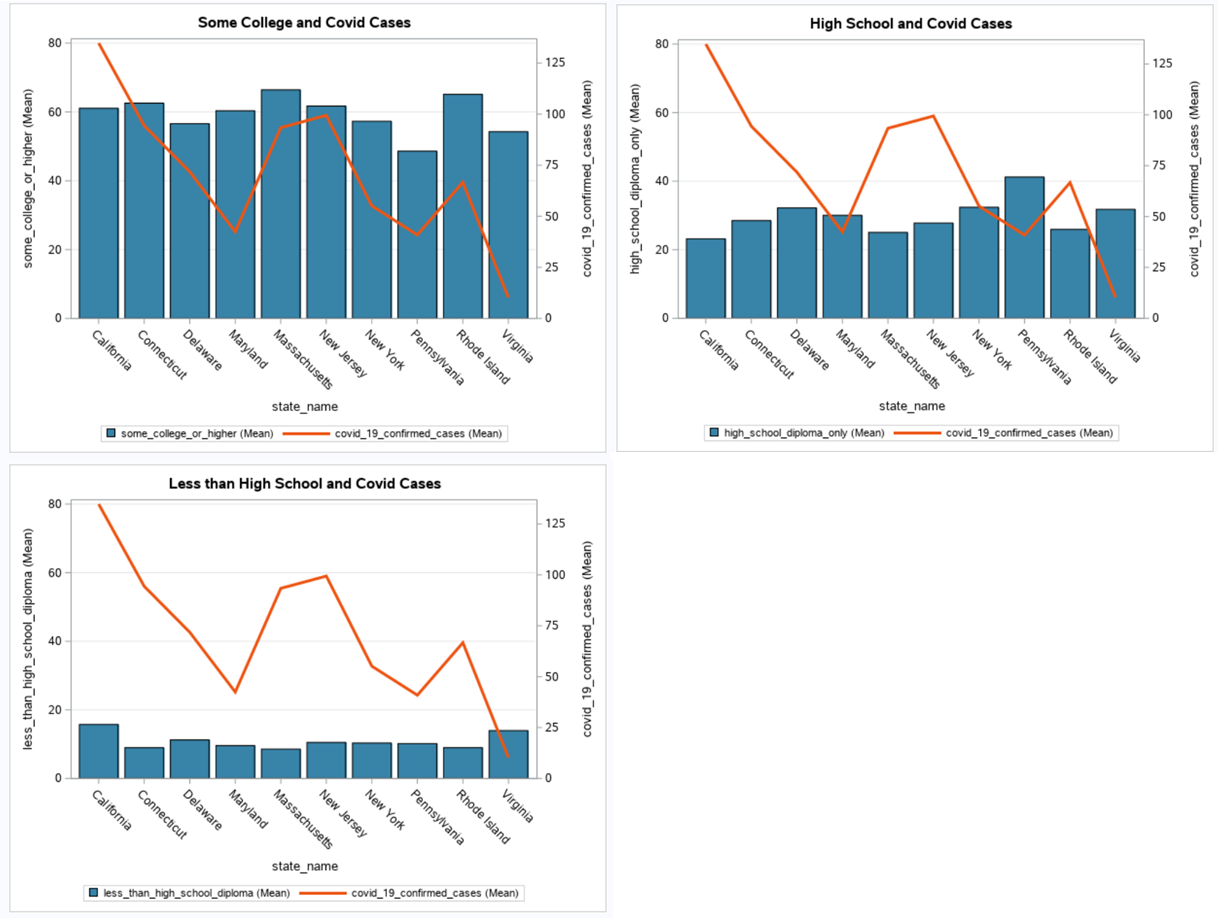
## **2.7 Educational Features**

The means for the group education level are shown below:

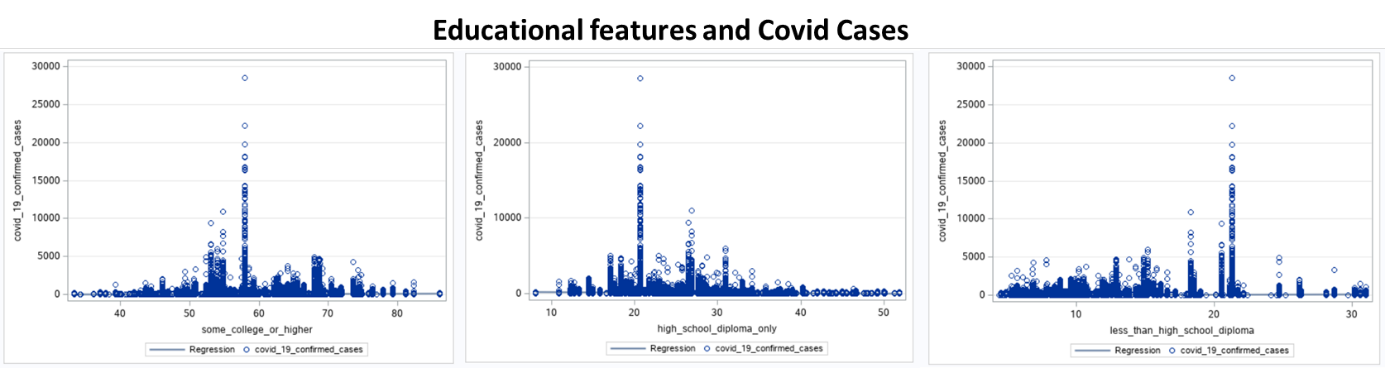
Table

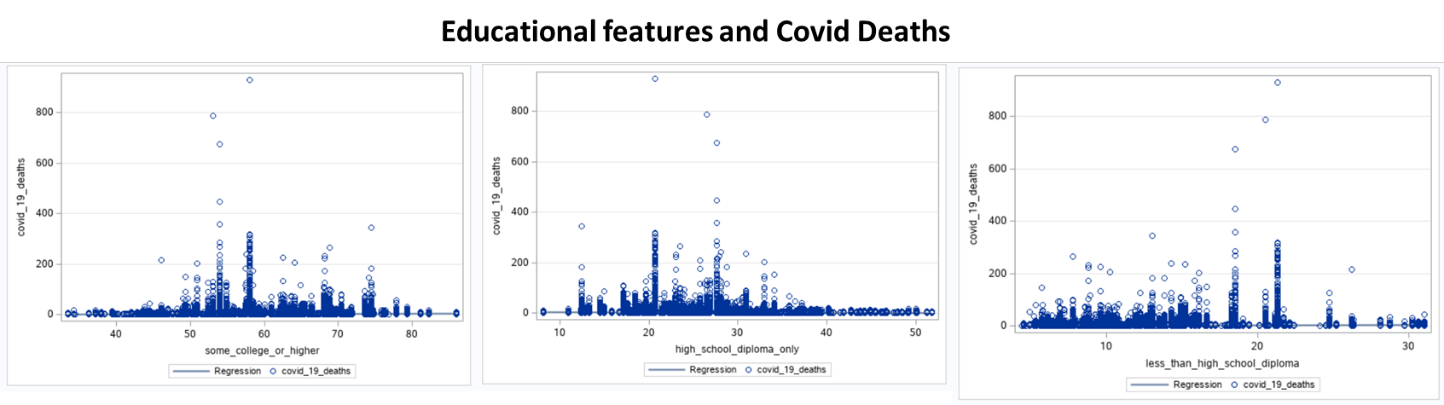
Description automatically generated

The following graph shows the educational level and the Covid Cases by each state of the sample, indicating that there is not a clear relationship among these variables. However, the Pearson Correlation presented a P-Value less than .05 and a r coefficient between -0.13 and 0.07, suggesting there is a weak relationship among the educational features and Covid cases.

**Table

Description automatically generated**

****

****

The distributions of these educational features are shown below:

The variable “less than high school diploma” shows the percentage of people in the county that has an educational attainment without completing a high school diploma. It is possible to see a distribution skewed to the right with fewer outliers, this means that most counties have people with lower educational level mainly between 5% and 20% of the population.

Chart, histogram

Description automatically generated

The mean of this low-level education in the counties in United States is 12.9% showing a high dispersion of the values as well with a range of 46.8%. Meaning that the counties present a big difference between the level education related to high school completion.

Chart, histogram

Description automatically generated

The histogram related to the frequency of the people that have completed just a high school diploma is higher in each county than those without high school. It is possible to see a distribution with a slightly skewed to the left and with a mean of 33.3%. The graph shows also a higher dispersion among the counties with a standard deviation of 7.3% and a range of 45%.

Chart, histogram

Description automatically generated

It seems to be the percentage of the population with some college or higher degree in each county follow a Normal Distribution with a mean (53.8%) almost equal to the median (53.7%) and the mode (54.7%). The observations present high dispersion with a range of 63.6%, this could indicate that there are some counties with percentage with lower education and some of them with higher. Apparently, there are not many outliers in this variable.

The higher percentage of people in the country have higher education attainment, with whiskers in the variable “some college or higher” in 24.4% and 82.8%. The variable “Only high school level” has a minimum value in 13.9% and maximum in 53.1%, while the variable “Less than high school level” presents lower values, with a minimum in 1.7% and maximum in 27.3%.

Chart, box and whisker chart

Description automatically generated

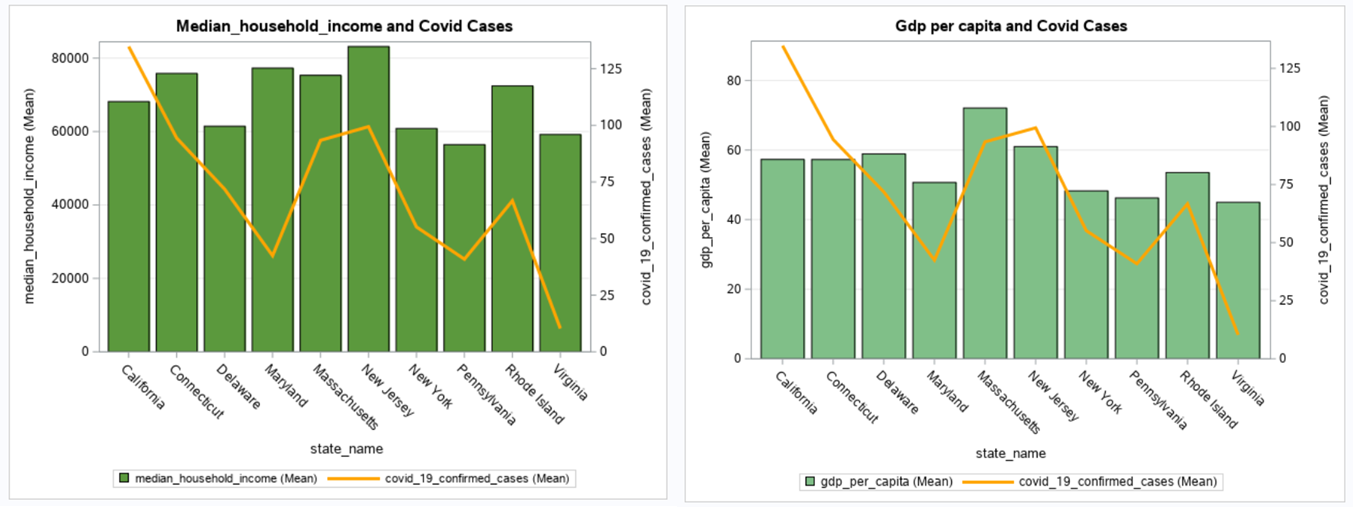
## **2.8 Economical features**

The main statistics for the economical features are shown in the following table:

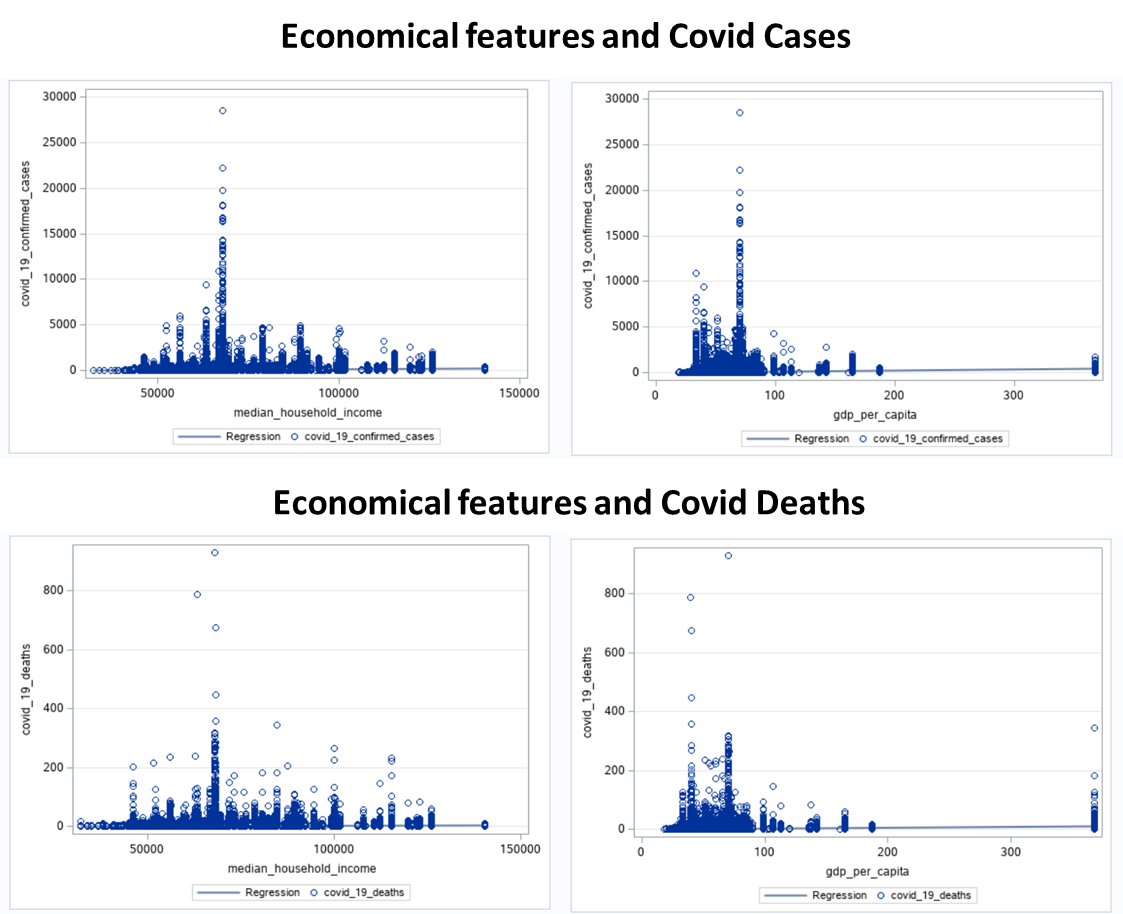
Table

Description automatically generated

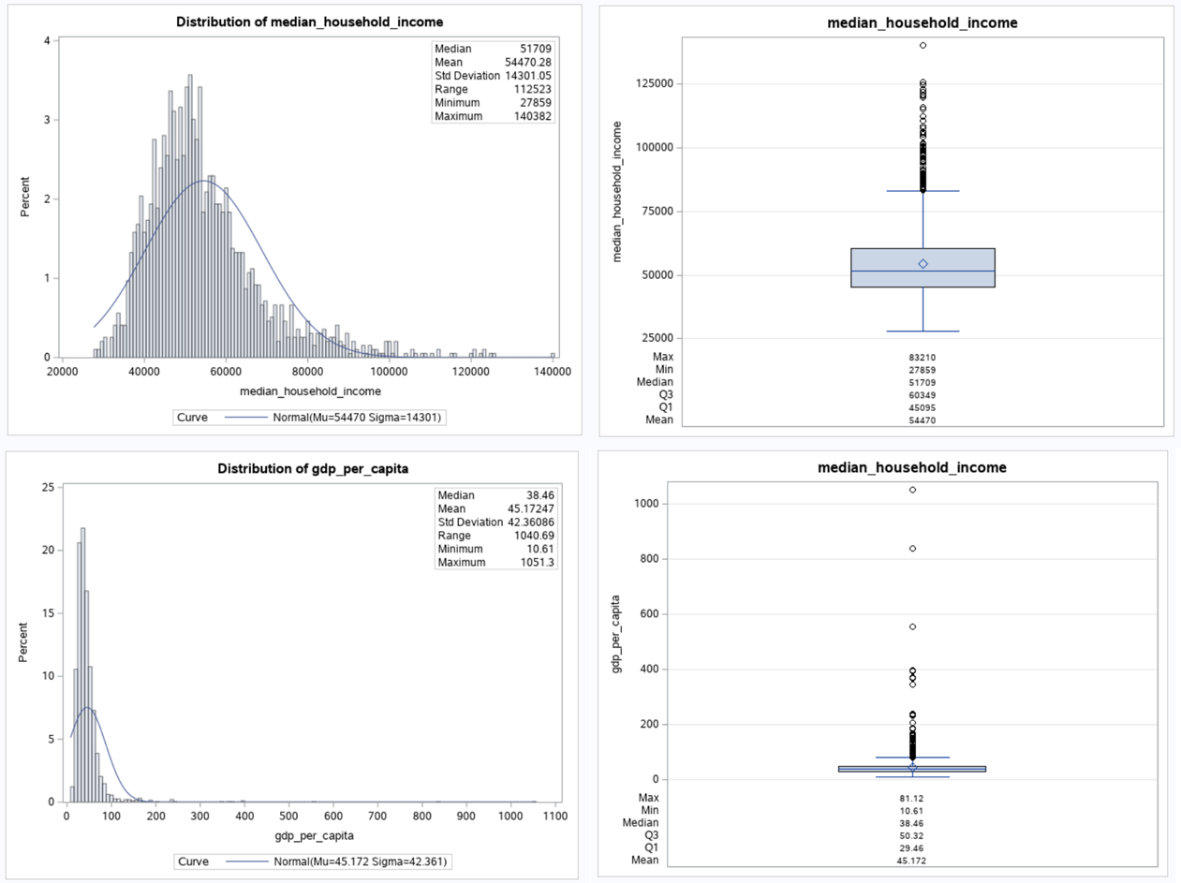
The r coefficients, obtaining among the economical features and Covid Cases and Deaths, take values between 0.10 and 0.08, showing a weak and positive correlation among those variables, with p-value less than .05. The graphs also show a relationship among these variables.

**Graphical user interface, table

Description automatically generated with medium confidence**



The distribution of the variables median household income and gdp are presented below:



## **Normalized data in gdp and media household to smooth outliers and conduct again the correlation with covid cases and deaths**

## **2.9 Demographic features (age)**

It was grouping the age according to the following: Percentage of population with age between 0 and 19, 20 and 39, 40 and 59, 60 and 79, and 80 and higher, by each county.

# **3. Summary**

According to our analysis, the features that we plan to consider in our predictive model related to the Covid Cases and Deaths will be the following:

* Climate
* Educational
* Economical
* Demographic (Virus pressure, total population)
* Morbility

It is possible that we find new variables to include in the predictive model, if they improve its accuracy.

# **Bibliography**

Haratian, Arezoo, Hadi Fazelinia, Zeinab Maleki, Pouria Ramazi, Hao Wang, Mark A. Lewis, Russell Greiner, and David Wishart. 2021. “Dataset of COVID-19 Outbreak and Potential Predictive Features in the USA.” Data in Brief 38 (October): 107360. https://doi.org/10.1016/j.dib.2021.107360.

**APPENDIX 1**

**Dataset variables**



1. **(Haratian et al., 2021)** [↑](#footnote-ref-1)
2. Data in Brief 38 (2021) [↑](#footnote-ref-2)