

Project 1

Henry Lambson, Jonas Moros, Blake Miller



February 21, 2023

Table of Contents

[Abstract 2](#_Toc127355399)

[1. Business Understanding 2](#_Toc127355400)

[2. Data Understanding 3](#_Toc127355401)

[2.1 Data Description 3](#_Toc127355402)

[2.2 Data Quality 5](#_Toc127355403)

[Data Preparation 5](#_Toc127355404)

[Exceptional Work 6](#_Toc127355405)

# Abstract

# 1. Business Understanding

The COVID-19 outbreak began in late 2019 and early 2020. It is a contagious disease caused by a virus SARS-CoV-2. The first case of COVID-19 was documented in Wuhan, China, and due to the virus’ highly contagious nature, the disease spread worldwide in the ensuing months. This spread led to the COVID-19 Pandemic.

Once the United States Federal Government saw how widespread the disease had become, it began taking steps to “flatten the curve,” or lessen the spread. This took the form of “social distancing,” where citizens were expected to limit their exposure to people outside their immediate household, and when that could not be done, stay at least six feet apart. Due to these regulations, places of public gathering, such as churches, gyms, and restaurants, began closing.

In this report, the data on the spread of the virus and its affects will be analyzed to answer questions like: How does the virus effects people differently in different parts of the country? How effective is social distancing? Are there regions that do better or worse against the virus? Can we predict the virus’s impact in one region given the data in another region? (We should probably come up with our own set of questions) Answering these questions through observing data is important because it allows people to make decisions with evidence instead of guessing or using their own “common sense”.

This report is intended for policy makers, doctors, educators, or any concerned citizens wondering about the virus. With the information in this report, policy makers could make decisions about whether to put social distancing or other spread prevention policies in place; doctors could learn more about how the virus spreads and those who are at highest risk; educators could make decisions about in-person vs. distance learning; and concerned citizens can learn more about how the virus might affect their lives.

# 2. Data Understanding

## 2.1 Data Description

The COVID-19\_cases\_plus\_census dataset contains data from the U.S. Census as well as data relating to the COVID outbreak, such as confirmed cases and deaths. The dataset contains 259 features and 3142 attributes. A truncated description of the dataset is provided in Table 1.

Table 1 The Description of the Data

|  |  |
| --- | --- |
| $ county\_fips\_code | Factor w/ 3142 levels "01001","01003",.. |
| $ county\_name | Factor w/ 1878 levels "Abbeville County",.. |
| $ state | Factor w/ 51 levels "AK","AL","AR",.. |
| $ state\_fips\_code | Factor w/ 51 levels "01","02","04",.. |
| $ date | Date, format |
| $ confirmed\_cases | num |
| $ deaths | num |
| $ geo\_id | Factor w/ 3142 levels "01001","01003",.. |
| $ nonfamily\_households | num |
| $ family\_households | num |
| $ median\_year\_structure\_built | num |
| $ rent\_burden\_not\_computed | num |
| $ rent\_over\_50\_percent | num |
| $ rent\_40\_to\_50\_percent | num |
| $ rent\_35\_to\_40\_percent | num |
| $ rent\_30\_to\_35\_percent | num |
| $ rent\_25\_to\_30\_percent | num |
| $ rent\_20\_to\_25\_percent | num |
| $ rent\_15\_to\_20\_percent | num |
| $ rent\_10\_to\_15\_percent | num |
| $ rent\_under\_10\_percent | num |
| $ total\_pop | num |
| $ male\_pop | num |
| $ female\_pop | num |
| $ median\_age | num |
| $ white\_pop | num |
| $ black\_pop | num |
| $ asian\_pop | num |
| $ hispanic\_pop | num |
| $ amerindian\_pop | num |
| $ other\_race\_pop | num |
| $ two\_or\_more\_races\_pop | num |

Due to the large size of the dataset, this report will only focus on a few important features, rather than taking a broad look at many features. The focus of this report will be to visualize the correlation between confirmed cases, deaths, and population by state, sex, and race. Table 2 shows a description of the updated dataset with only the features that will be analyzed in more detail.

Table 2 Description of Smaller Dataset

|  |  |
| --- | --- |
| $ state | Factor w/ 51 levels "AK","AL","AR",.. |
| $ confirmed\_cases | num |
| $ deaths | num |
| $ total\_pop | num |
| $ male\_pop | num |
| $ female\_pop | num |
| $ white\_pop | num |
| $ black\_pop | num |
| $ asian\_pop | num |
| $ hispanic\_pop | num |
| $ amerindian\_pop | num |

## 2.2 Data Quality

To continue the preprocessing of the data, it is necessary to verify the quality of the data. Duplicates and missing data must be accounted for in order to see quality results once the analysis begins.

The small dataset was verified to find any missing or duplicated values, and there were none of either. The data is clean and ready to be worked with. Just to be safe, the full dataset will also be checked for duplicates and missing data so that it can be worked with if need be.

The full dataset also had no duplicated values, but it did contain missing values. The rows with these missing values have been dropped from the dataset, so now the full dataset has been cleaned and can also be worked with. Now that both datasets have been cleaned, the small dataset will be split into the two states that will be analyzed further: Florida and California. Table 3 contains a summary of the California data, and Table 4 contains a summary of the Florida data.

Table 3 Summary of California Dataset

|  |  |  |
| --- | --- | --- |
| state  CA: 58  AK: 0  AL: 0  AR: 0  AZ: 0  CO: 0  (Other): 0 | confirmed\_cases  Min. : 40  1st Qu.: 2673  Median : 11114  Mean : 51668  3rd Qu.: 39784  Max. : 1002614 | deaths  Min. : 0.0  1st Qu.: 27.5  Median : 119.0  Mean : 581.7  3rd Qu.: 440.5  Max. : 13936.0 |
| total\_pop  Min. : 1203  1st Qu.: 47268  Median : 182486  Mean : 672118  3rd Qu.: 677036  Max. : 10105722 | male\_pop  Min. : 664  1st Qu.: 23726  Median : 92134  Mean : 333907  3rd Qu.: 336391  Max. : 4979641 | female\_pop  Min. : 539  1st Qu.: 23541  Median : 91938  Mean : 338212  3rd Qu.: 340644  Max. : 5126081 |
| white\_pop  Min. : 777  1st Qu.: 30575  Median : 101253  Mean : 254786  3rd Qu.: 289446  Max. : 2676982 | black\_pop  Min. : 4.0  1st Qu.: 545.8  Median : 3695.5  Mean : 37266.5  3rd Qu.: 16494.5  Max. : 799579.0 | asian\_pop  Min. : 0  1st Qu.: 732  Median : 9925  Mean : 93585  3rd Qu.: 54587  Max. : 1442577 |
| hispanic\_pop  Min. : 117  1st Qu.: 9684  Median : 49339  Mean : 260446  3rd Qu.: 248389  Max. : 4893579 | amerindian\_pop  Min. : 28.0  1st Qu.: 550.2  Median : 1218.5  Mean : 2376.1  3rd Qu.: 2651.5  Max. : 19915.0 |  |

Table 4 Summary of Florida Dataset

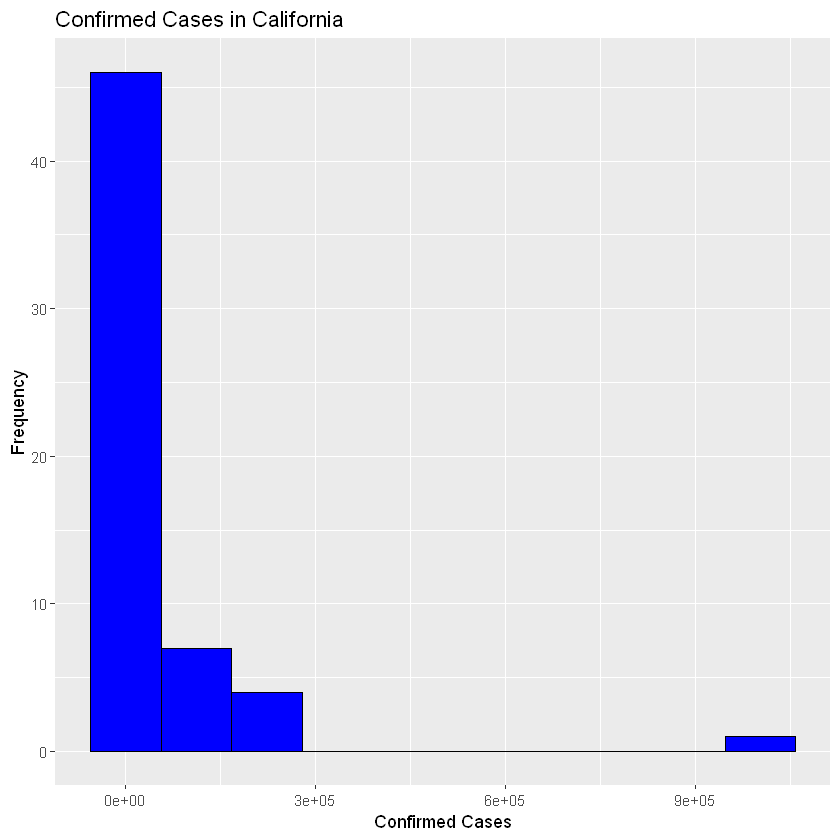
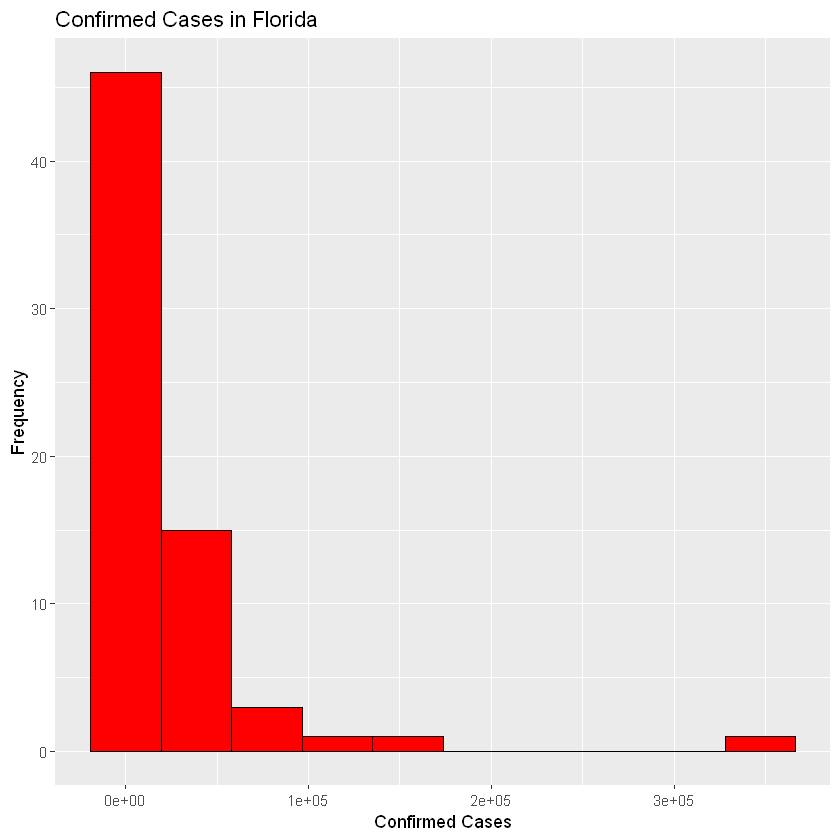
|  |  |  |
| --- | --- | --- |
| state  FL: 67  AK: 0  AL: 0  AR: 0  AZ: 0  CO: 0  (Other): 0 | confirmed\_cases  Min. : 845  1st Qu.: 2466  Median : 6886  Mean : 23668  3rd Qu.: 23409  Max. : 347965 | deaths  Min. : 4.0  1st Qu.: 34.0  Median : 156.0  Mean : 364.7  3rd Qu.: 414.5  Max. : 4622.0 |
| total\_pop  Min. : 8347  1st Qu.: 27432  Median : 116754  Mean : 302663  3rd Qu.: 334997  Max. : 2702602 | male\_pop  Min. : 5104  1st Qu.: 14735  Median : 58883  Mean : 147976  3rd Qu.: 162600  Max. : 1311997 | female\_pop  Min. : 3131  1st Qu.: 12862  Median : 57871  Mean : 154688  3rd Qu.: 173556  Max. : 1390605 |
| white\_pop  Min. : 5933  1st Qu.: 17984  Median : 98930  Mean : 166044  3rd Qu.: 229463  Max. : 799018 | black\_pop  Min. : 998  1st Qu.: 4116  Median : 9616  Mean : 46708  3rd Qu.: 37357  Max. : 518277 | asian\_pop  Min. : 0  1st Qu.: 139  Median : 1785  Mean : 7963  3rd Qu.: 7414  Max. : 66304 |
| hispanic\_pop  Min. : 406  1st Qu.: 1784  Median : 11801  Mean : 74851  3rd Qu.: 50581  Max. : 1823038 | amerindian\_pop  Min. : 11.0  1st Qu.: 101.0  Median : 347.0  Mean : 622.5  3rd Qu.: 903.0  Max. : 3156.0 |  |

## 2.3 Visualization of Attributes

In this section, the individual attributes of both the California and Florida datasets will be analyzed and compared.

1. *Confirmed Cases (Ratio)*

The data for confirmed cases in California and Florida are represented in the histograms in Figures 1 and 2. Despite the difference in population between the two states, which will be discussed in detail further below, the histograms share a similar structure. The vast majority of the data is on the very low end on the graph, with one outlier causing the graph to be skewed right.

 Figure 1 Confirmed Cases CA Figure 2 Confirmed Cases FL

# Data Preparation

# Exceptional Work