

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, Table 3 will show a summary of the data in the small cases dataset.

Table 3 Summary of Small Dataset

|  |  |  |
| --- | --- | --- |
| state  TX: 254  GA: 159  VA: 133  KY: 120  MO: 115  KS: 105  (Other):2256 | confirmed\_cases  Min. : 0.0  1st Qu.: 796.2  Median : 1916.5  Mean : 7558.9  3rd Qu.: 4955.0  Max. : 1002614.0 | deaths  Min. : 0.0  1st Qu.: 12.2  Median : 31.0  Mean : 124.8  3rd Qu.: 77.0  Max. : 13936.0 |
| total\_pop  Min. : 74  1st Qu.: 10945  Median : 25692  Mean : 102166  3rd Qu.: 67445  Max. : 10105722 | male\_pop  Min. : 39  1st Qu.: 5514  Median : 12798  Mean : 50292  3rd Qu.: 33481  Max. : 4979641 | female\_pop  Min. : 35  1st Qu.: 5460  Median : 12885  Mean : 12885  3rd Qu.: 34108  Max. : 5126081 |
| white\_pop  Min. : 18  1st Qu.: 8093  Median : 20205  Mean : 62787  3rd Qu.: 53500  Max. : 2676982 | black\_pop  Min. : 0  1st Qu.: 95  Median : 758  Mean : 12554  3rd Qu.: 5396  Max. : 1226134 | asian\_pop  Min. : 0  1st Qu.: 31  Median : 138  Mean : 5407  3rd Qu.: 712  Max. : 1442577.0 |
| hispanic\_pop  Min. : 0  1st Qu.: 323  Median : 1025  Mean : 17986  3rd Qu.: 4868  Max. : 4893579 | amerindian\_pop  Min. : 0  1st Qu.: 24  Median : 96  Mean : 668  3rd Qu.: 348  Max. : 64102.0 |  |

## 2.3 Visualization of Attributes

# Data Preparation

# Exceptional Work