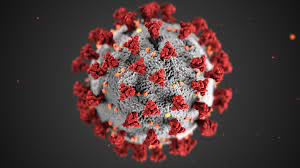
COVID-19 Data & Visualization



**Ryan Capó, James Zhai, Will Clark**

15 March 2021

CSE 5331

# 

# 

# EXECUTIVE OUTLINE

This study was conducted to explore the data surrounding the COVID-19 virus. After sifting through data and generating visualizations as well as comparing the different government orders set to control the virus, we saw numerous relationship between different variables that allow us to make predictions of how the virus might spread throughout a county based on the median age, population density, wealth levels, and education levels of that county.

The study concluded that the wealthier the population and the higher education of the population, the less likely it was to have high case numbers and deaths, but the lower the wealth and education levels, the more likely it was to have higher cases and deaths. It also concluded that because of the response to limit the elderly’s contact with the virus, the older the population’s median age means the lower the chances of high COVID case numbers and deaths.

# 

# 

# TABLE OF CONTENTS

[**EXECUTIVE OUTLINE**](#_heading=h.2et92p0) **1**

[**BUSINESS UNDERSTANDING**](#_heading=h.op9u2vio2k4x) **2**

[**DATA UNDERSTANDING**](#_heading=h.tyjcwt) **3**

[**DATA PREPARATION**](#_heading=h.3dy6vkm) **13**

[**CONCLUSION**](#_heading=h.1t3h5sf) **18**

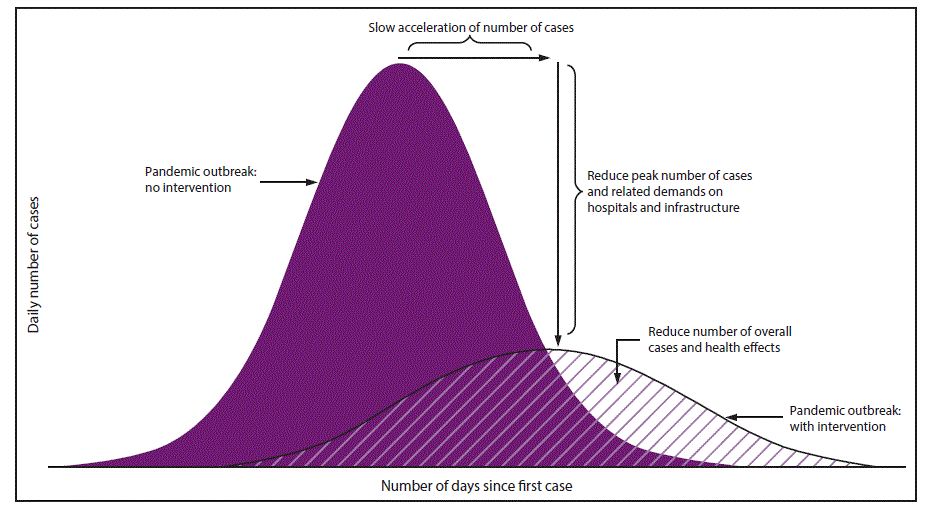
[**REFERENCES**](#_heading=h.2s8eyo1) **18**

# 

# BUSINESS UNDERSTANDING

COVID-19 is a respiratory disease that was discovered in 2019. It is caused by a new coronavirus called SARS-CoV-2. Scientists believe the virus mainly spreads from person to person when an infected person coughs, sneezes, or talks [1]. However, not every person that becomes infected through respiratory droplets develops symptoms. On the other hand, those that have symptoms can range from mild to severe. There are high risk and low risk categories for the population. An example of the high risk category is people with underlying medical conditions or adults 65 years and older [1].

Social distancing is one of the key components of slowing the spread of the virus. It means you keep at least six feet away from people that are not from your household. This holds true for indoor and outdoor spaces [2]. The virus spreads via respiratory droplets, so it is important to practice social distancing to prevent the spread. Another reason that social distancing is important is because it assists in flattening the curve. Figure 1 explains the “curve” scientists refer to. The dark purple curve represents the projected amount of people who will contract the disease over a certain amount of time without safety precautions implemented. The lined purple graph represents social dinstancing’s effect on the curve — hence, “flattening the curve” [3].



**Figure 1: Describing “Flattening the Curve”**

It is important to look at the spread of the virus data because it is essential in predicting how it spreads over time. This is useful because health care capacity is limited. If scientists are able to predict the spread of the virus in areas where health care capacities are at their limits, then necessary resources can be produced and allocated ahead of time so capacity is not exceeded. Agencies like the Center for Disease and Control Prevention collect this information but hospitals might be interested in looking at the data. If they are able to see that over time cases will rise exponentially, then they can take measures to allocate resources so they do not exceed capacity. At the end of the day, health care workers know lives will be lost, yet the goal is to limit this. Therefore, the data gathered in predicting the spread of the virus can be very useful in limiting the death rate.

# DATA UNDERSTANDING

DATA SET DESCRIPTION

The data sets we worked with broke down confirmed COVID-19 cases and deaths by each county in Texas. The data sets are discussed in more detail in the *Data Quality* section. This section describes the most important variables we used. The data types are also specified. A nominal data type means data that labels variables that provide no quantitative value. An ordinal data type means data that can be measured or ordered. Interval scales give us the order of values and the ability to quantify the difference between each one. Ratio scales give us order, interval values, and the ability to calculate ratios. Table 1 provides the necessary data set descriptions.

Note that this section *only* describes the most important variables featured in all of the graphs. The *Data Preparation* section further breaks down variables like median\_income by providing info on every given income range. Another example of this would be the variable total\_pop being broken down by age ranges for males and females 60 years old and up.

**Table 1: The description and data types of each important feature**

|  |  |  |
| --- | --- | --- |
| Feature | Data Type | Description |
| County\_fips\_code | Nominal | The individual zip code for each county in Texas |
| County\_name | Nominal | The name of each county in Texas |
| Confirmed\_cases | Ratio | The number of cases confirmed per date per county. |
| Deaths | Ratio | The number of deaths per date per county |
| State | Nominal | The state in which the data comes from (only Texas) |
| Total\_pop | Ratio | The total population for each county in Texas |
| Median\_age | Ordinal | The median age for each county in Texas |
| Median\_income | Ratio | The median income for each county in Texas |
| Date | Nominal | The data set contains a year's worth of confirmed cases and deaths for each county. The *Date* variable indicates which day over that year period. |

DATA QUALITY

In dealing with data sets, it is very important to be working with a clean and quality data set. If we are working with an incomplete set, then we are susceptible to skewed data because of outliers, duplicates, or NA values. This is a necessary step in data mining because we have to be able to understand the data in order to be able to present anything meaningful.

For this project, we pulled from the COVID-19\_cases\_TX data set and the COVID-19\_cases\_plus\_census data set. The COVID-19\_cases\_TX data set describes confirmed cases and deaths for every county in Texas, which is outlined by county zip codes. The COVID-19\_cases\_plus\_census data set gives us a specific breakdown of each county. For example, the data set breaks down confirmed cases and deaths by income ranges. This section describes the data quality of the most important variables we used.

**County\_fips\_code:**

The zip code by county was not missing any data because we are looking at every county in Texas and our data set has that available to us. However, the COVID-19\_cases\_TX set contained many duplicates for zip codes because this file is broken down over a time period of one year for every county. Therefore, a year’s worth of data on confirmed cases and deaths by county will have duplicates. Yet, the file COVID-19\_cases\_plus\_census set fixes this for us. The date on this file is the same for each county, so there are only 255 unique zip codes represented.

**County\_name:**

The county name variable also contains duplicates, but it is also for the same reason that county\_fips\_code has duplicates. The census file only has 255 rows representing each county, so there are no duplicates in that file as each unique county is represented once.

**Confirmed\_cases:**

This variable was mostly clean and accurate, therefore it did not need a lot of attention in terms of cleaning. However, some outliers do exist. For example, Loving County (total population of 74) only had 1 confirmed case as of January 19, 2021. On the other hand, Dallas County (total population of 2,552,213) had 234,625 confirmed cases on that same day. This outlier does not necessarily skew our data, but rather it gives us a scope into how covid affected small and large populations.

**Deaths:**

This variable is essentially the same as confirmed\_cases. Using the same example, Loving County has had 0 deaths whereas Dallas County has had 2,453 deaths. Again, outliers like Loving County does not skew out data but rather gives us a better understanding of it.

**State:**

State contains duplicates only because the state is the same for every entry since we are only looking at Texas. This does not affect our data.

**Total\_pop:**

Given that we have access to a census file, the total population variable per country is accurate. There are no missing values.

**Median\_age:**

This variable represents the median age for each county in Texas. It is a rough estimate, but there are no missing values.

**Date:**

The date contains duplicates in the census file. This is because the file is showing us the total confirmed cases across an entire year. The date duplicates do not affect our data but give us a broader scope on confirmed cases and deaths per county.

STATISTICS

The following section provides a statistical summary of variables we looked to find correlations in. Comparing all of the data, it is important to look at the number of cases and the number of deaths. This data is extremely helpful in finding correlations between different data types. We chose to look at education levels, wealth, and age as it related to the number of cases and deaths. Some of these variables seemed to be connected. The percentage of wealthy people in a county are often correlated to the level of education in the county. This makes sense because people with a higher level of education typically make more money. Not only do they make more money, but it also allows them to live in areas with better healthcare if they do become infected. This higher level of healthcare allows them to be cured more easily because the facilities they are in are often better funded and have better equipment and treatment options. Concerning age, we assumed the older the population of the county, the more deaths per case. This is simply due to the fact that older people have weakened immune systems. They also tend to have more underlying conditions that decrease their ability to fight the virus.

As we analyzed the data, our initial hypothesis seemed to be correct. These variables proved to be effective in predicting the impact COVID would have on a county. Wealthier counties had fewer confirmed cases, which agrees with our hypothesis. Even in wealthy counties with more cases, the death rate was much lower due to the higher quality medical facilities. The older the county was on average meant a higher number of cases and of deaths

**Table 2: The statistical summary of various educational backgrounds**

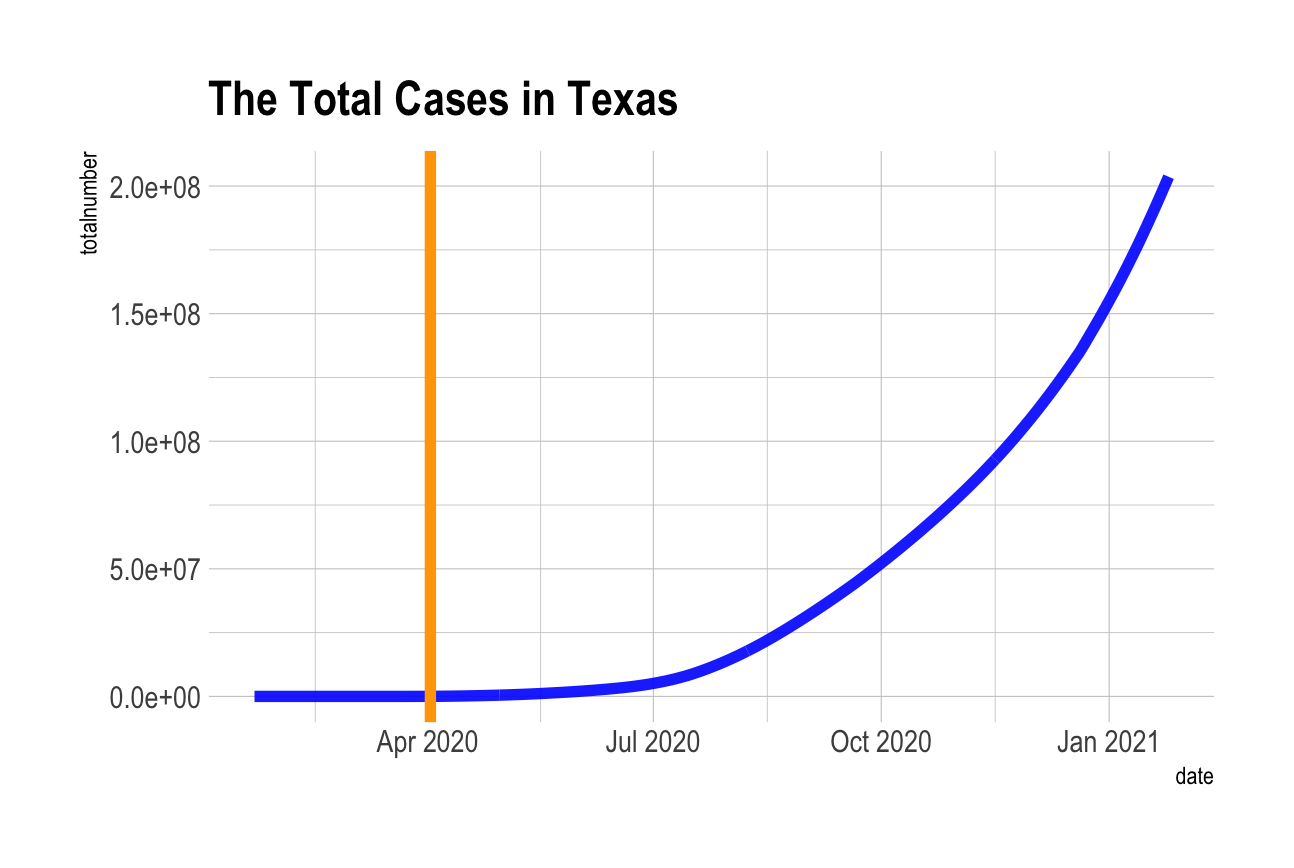
**(the ages for the various educational backgrounds are males 45-64)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Data Type | Range | Median | Mean | Variance | Std. Dev. | Max | Min |
| LowEdu% | Ratio | 41.88 | 8.25 | 9.97 | 48.98 | 7.00 | 41.88 | 0 |
| MidEdu% | Ratio | 66.11 | 68.03 | 66.81 | 89.16 | 9.44 | 100.0 | 33.89 |
| HighEdu% | Ratio | 63.56 | 21.91 | 23.22 | 88.20 | 9.39 | 63.56 | 0 |
| Case% | Ratio | 16.94 | 7.39 | 7.80 | 6.94 | 2.63 | 18.29 | 1.35 |
| Death% | Ratio | .6284 | .1727 | .1855 | .0098 | .0989 | .6284 | 0 |
| Associates\_degree | Ratio | 29,530 | 134 | 846 | 7,750,461 | 2784 | 29,530 | 0 |
| Bachelors\_degree | Ratio | 96,857 | 210.5 | 2275.4 | 79,468,326 | 8914.5 | 96,857 | 0 |
| Graduate\_degree | Ratio | 60,236 | 93 | 1337.5 | 30,451,095 | 5518.3 | 60,236 | 0 |
| Less\_than\_9\_grade | Ratio | 67,074 | 227.5 | 1194 | 28,385,351 | 5327.8 | 67,074 | 0 |
| Some\_college | Ratio | 95,642 | 513.5 | 2012.5 | 73,979,292.7 | 8501.1 | 3 | 0 |
| Male\_60\_to\_61 | Ratio | 46,922 | 242.5 | 1125.7 | 15,368,199.7 | 3920.2 | 46,922 | 0 |
| Female\_60\_to\_61 | Ratio | 48,638 | 247 | 1234.7 | 35.59 | 4241.1 | 48,638 | 0 |
| Male\_70\_to\_74 | Ratio | 49,831 | 390 | 1453 | 18,985,602 | 4357 | 49,831 | 0 |
| Female\_70\_to\_74 | Ratio | 57,421 | 392 | 1700.3 | 27,337,975.7 | 5228.6 | 57,421 | 0 |
| Male\_80\_to\_84 | Ratio | 19,416 | 186.5 | 629.4 | 3,146,740.6 | 1773.9 | 19.416 | 0 |
| Female\_80\_to\_84 | Ratio | 29,974 | 219.5 | 892.1 | 7,504,377.1 | 2739.4 | 29,974 | 0 |
| Median\_age | Ratio | 31.70 | 38.55 | 39.02 | 35.59 | 5.97 | 57.50 | 25.80 |

INTERPRETING AND VISUALIZING THE DATA

After generating many different charts, relationships between different variables become very clear. The relationship between wealth and education level is very strong, and it is shown consistently in the graphs as well as the relationship between population density and the rate of the spread.

In looking to see the impacts different government orders have on the virus, we can look at a graph of the total number of cases in Texas.

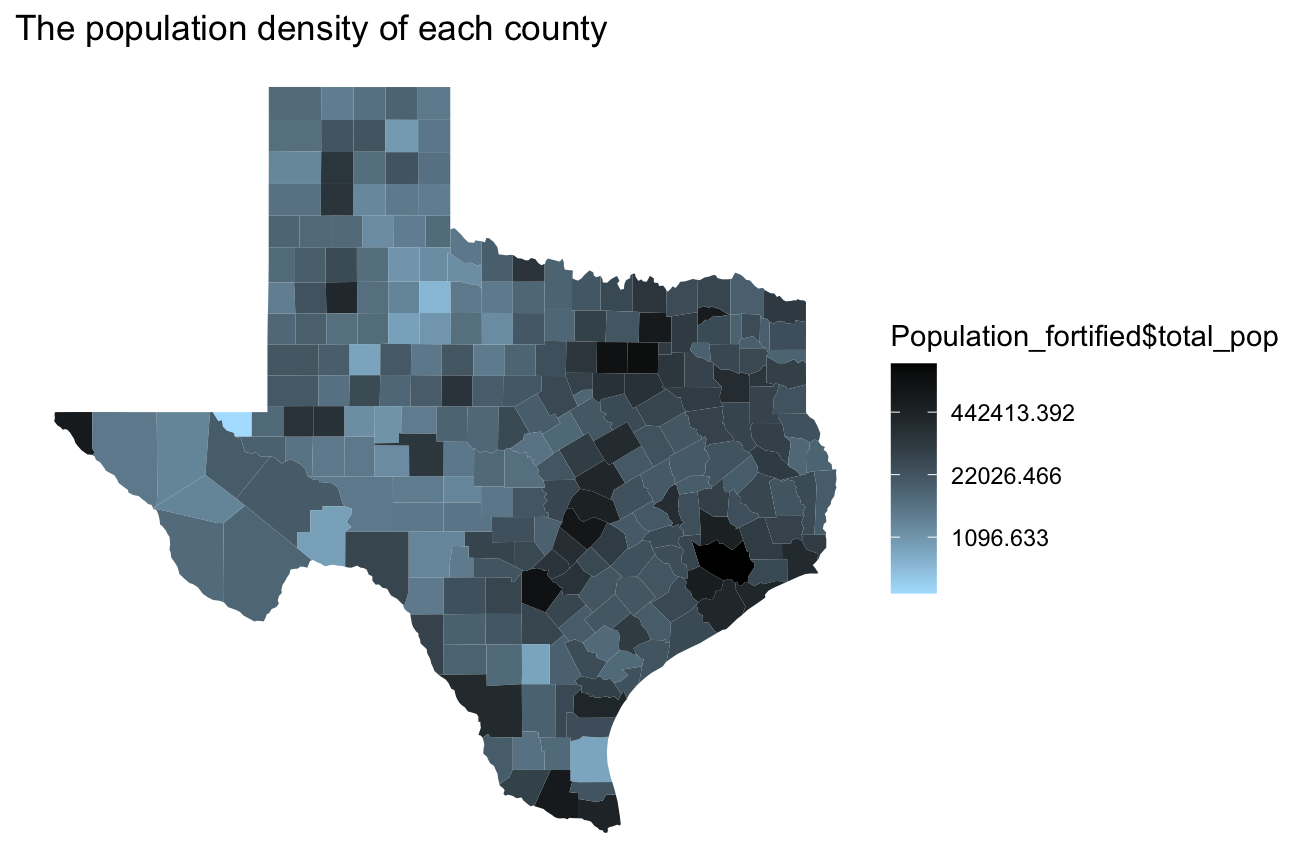


**Figure 2: The Total Number of Cases in TX**

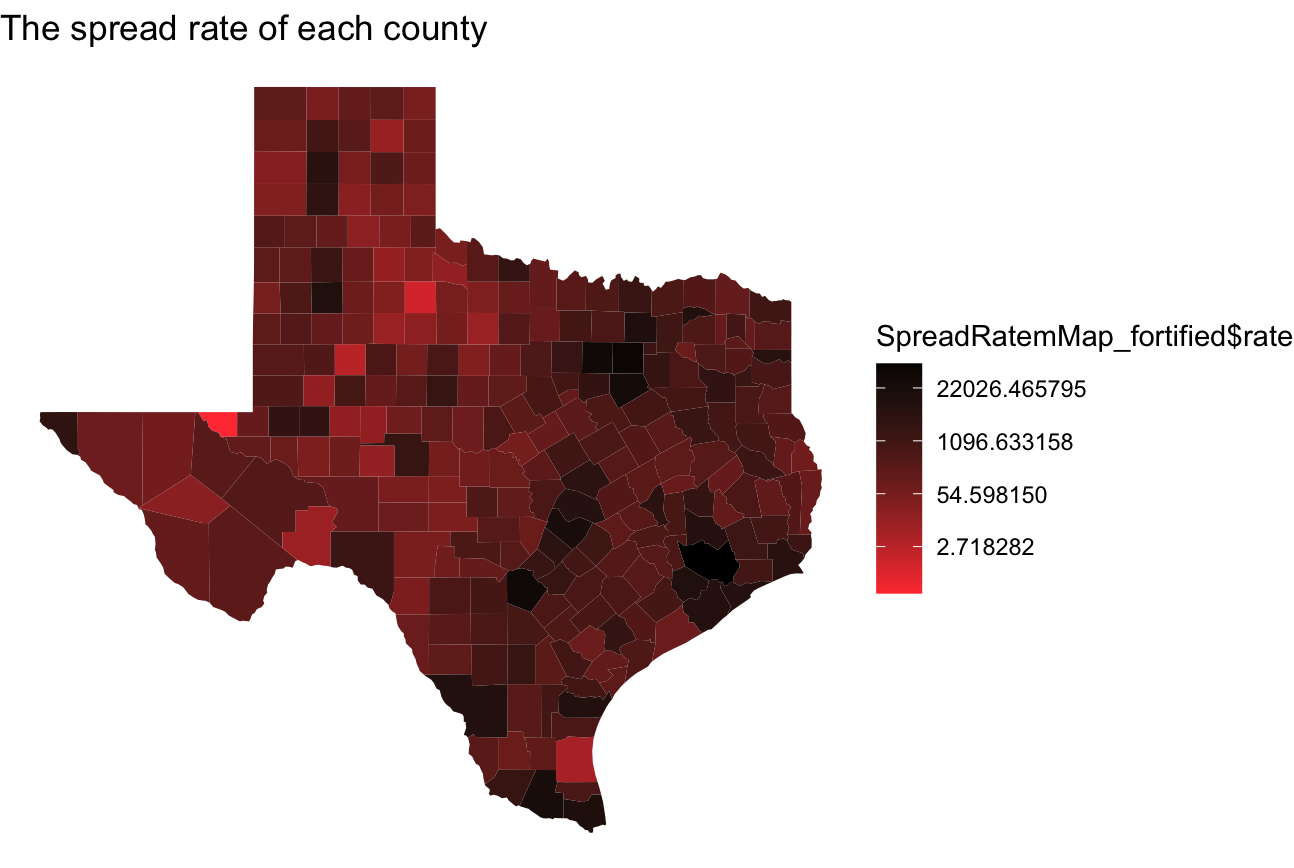
You can see the orange line in Figure 2 marking when Gov. Abbott passed an executive order no longer allowing people to eat in restaurants. After this was passed, the number of cases remained low until the beginning of May when the governor began opening up restaurants and bars. As you can see by the blue line, around the beginning of May the number of cases began to rise. From then, no matter what the governor did, the case numbers continued to rise despite mask mandates and social distancing.

*Population Density and Spread Rate*

One important thing to note is the relationship between the population density and spread rate of the virus.



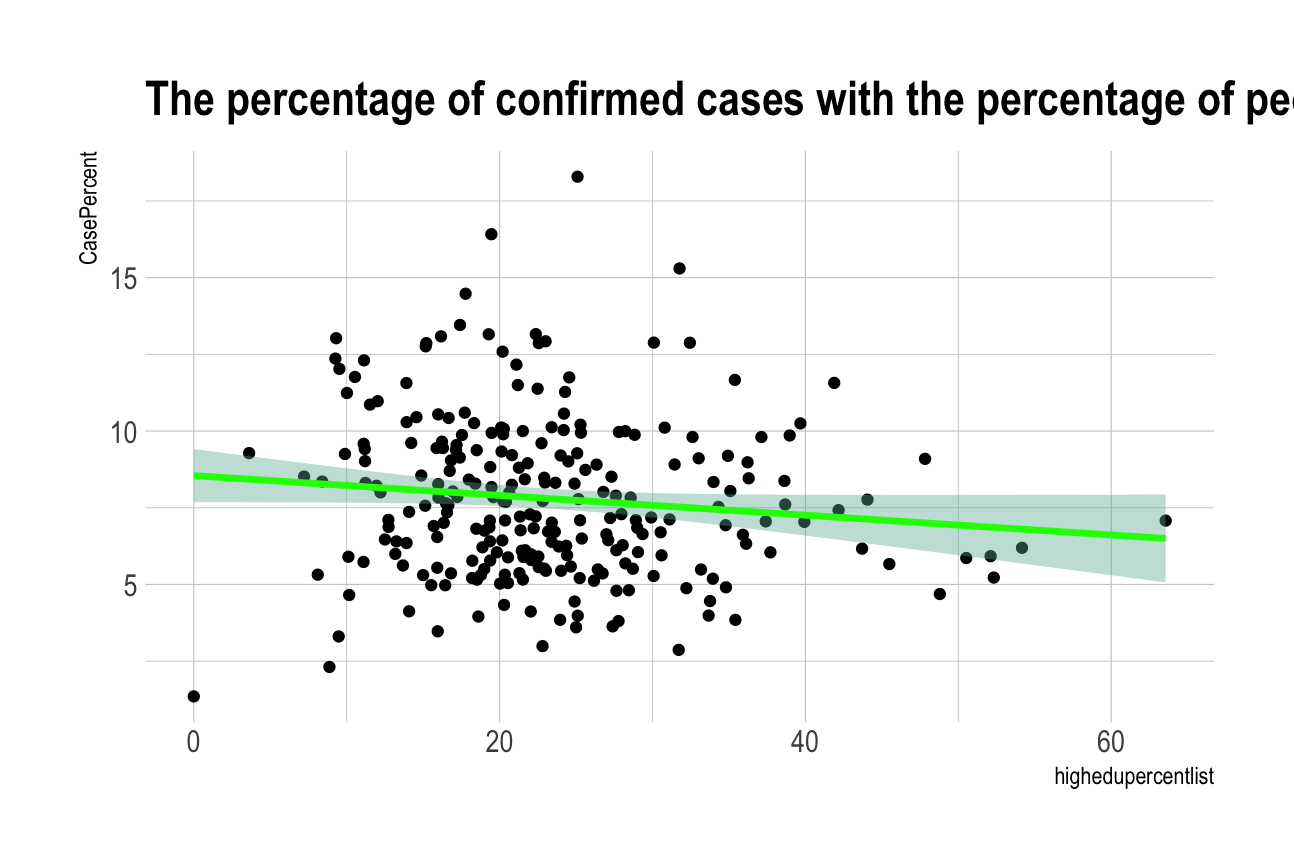
**Figure 3: The Population Density of Each Texas County**

****

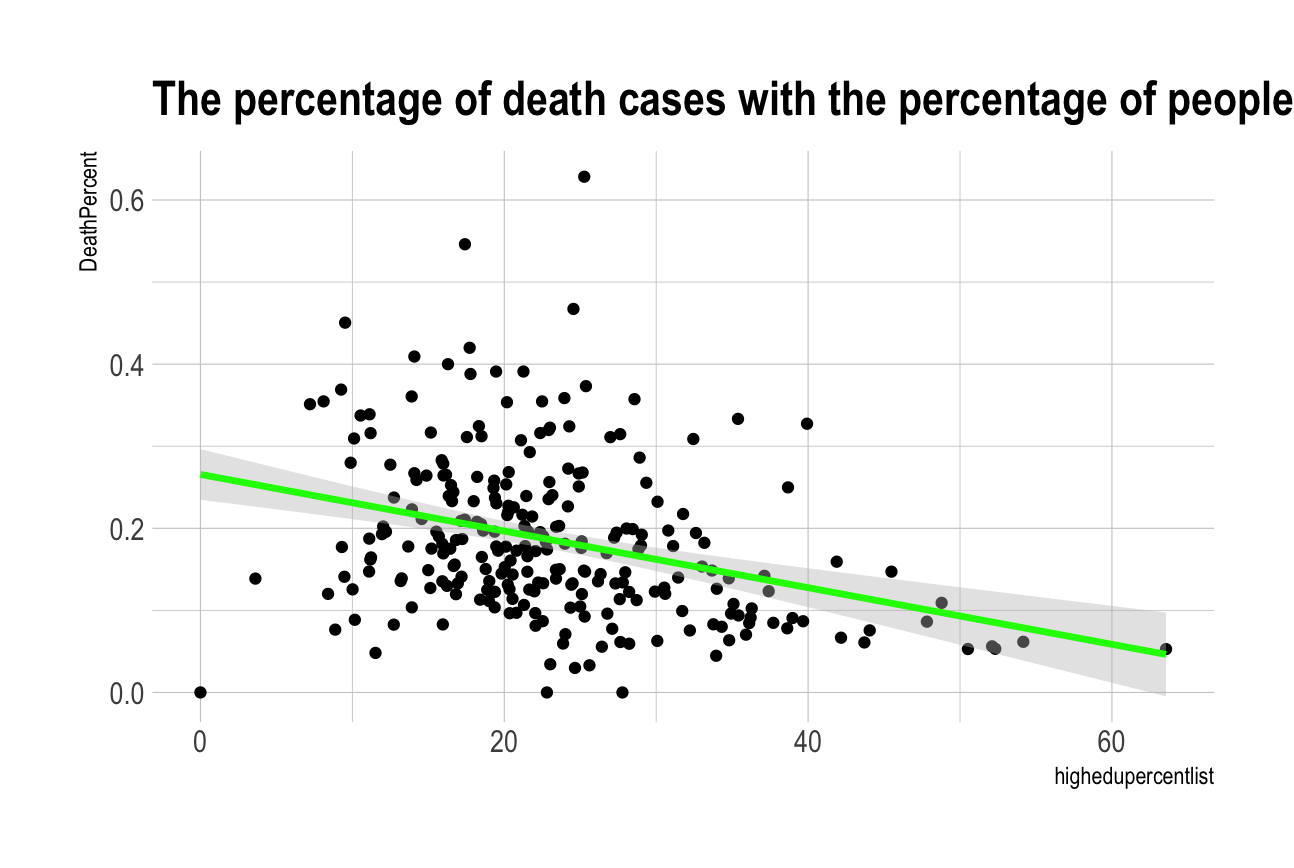
**Figure 4: The Spread Rate of Each Texas County**

In these graphs, it is easy to see that population density has an extremely strong correlation to the rate of spreading. In counties that contain large cities, such as Dallas, Houston, and Austin, the population density is larger and so is the spread rate. These are shown by darker counties on Figure 3 to show more dense populations, and darker red counties on Figure 4 to show a higher rate of spread. The figures are almost interchangeable. We see that the darker areas on Figure 3 correspond greatly to the darker areas on Figure 4.

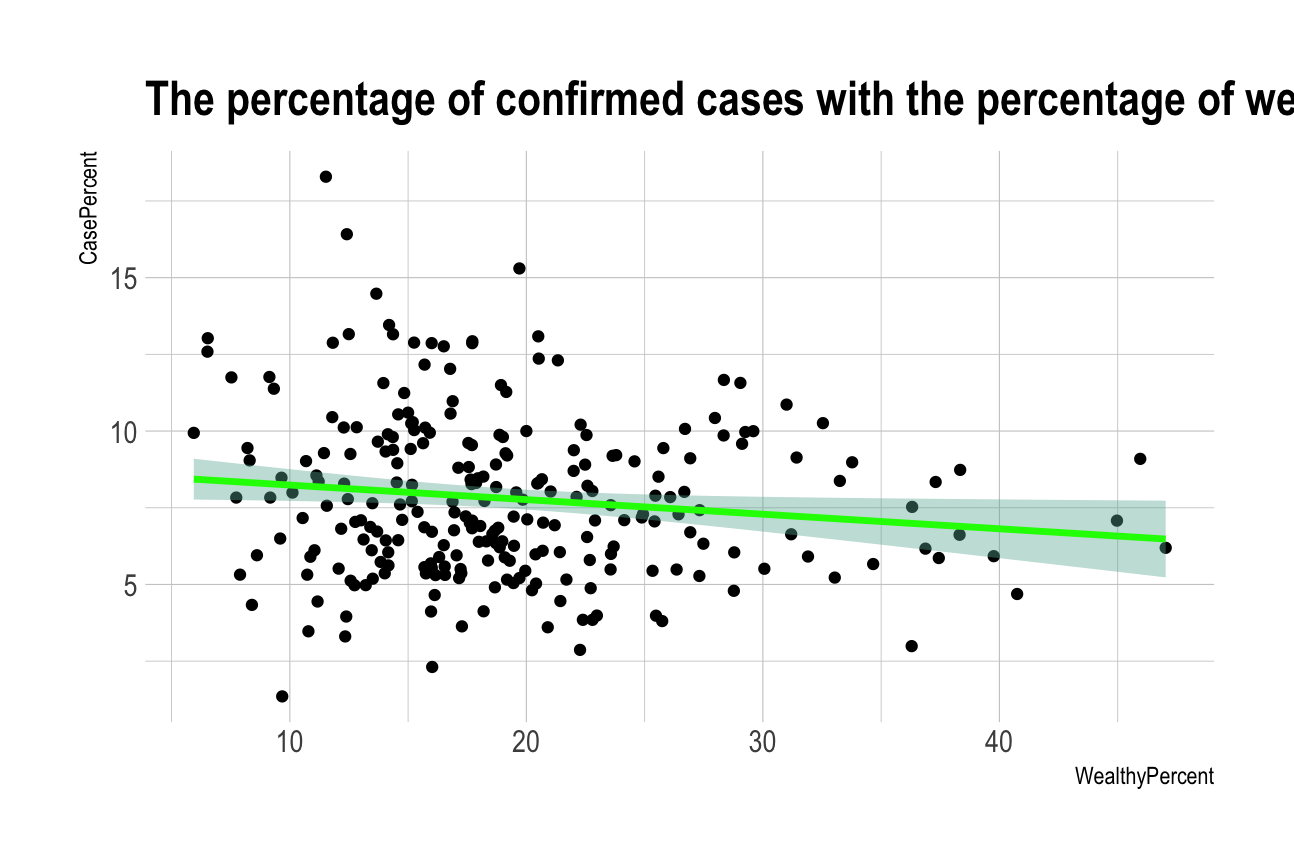
*COVID-19 vs. Wealth and Education*



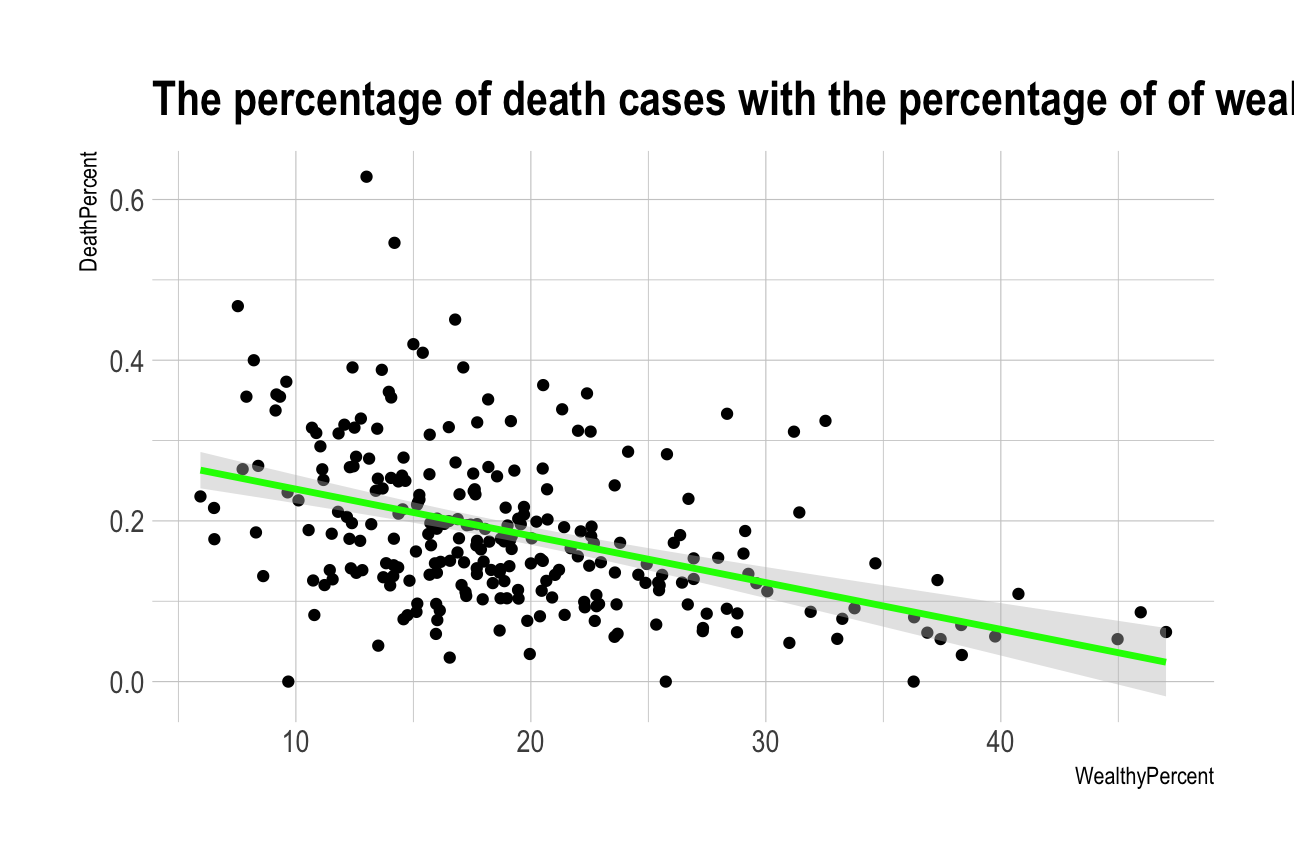
**Figure 5: Percent of Confirmed Cases vs. Percent of Population with Higher Education**



**Figure 6: Percent of Deaths vs. Percent of Population with Higher Education**

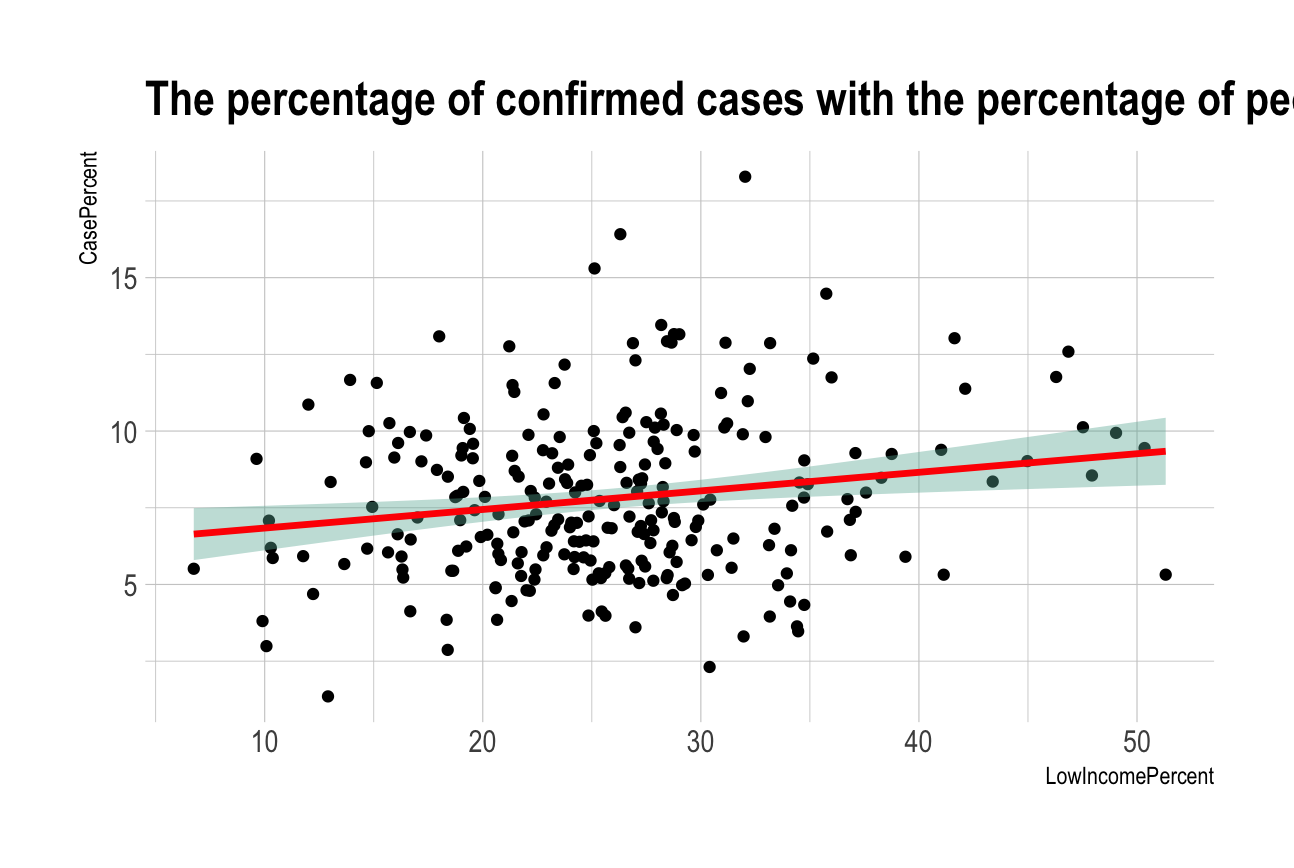
****

**Figure 7: Percentage of Confirmed Cases vs. Percent of Population that is Wealthy**

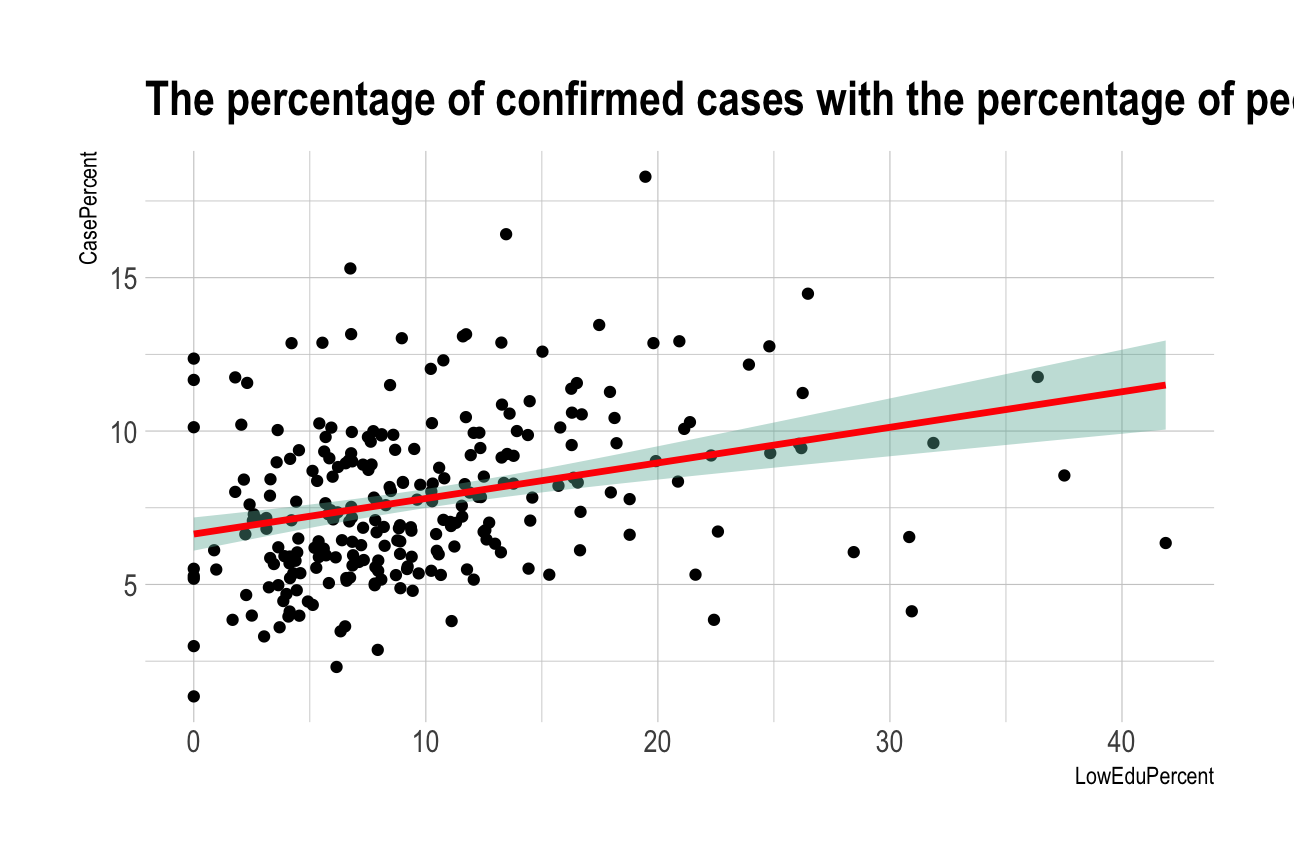
****

**Figure 8: Percentage of Deaths vs. Percent of Population that is Wealthy**

These charts are four of the most fascinating. It seems it could easily be inferred after thinking about the regular relationship between wealth and education, but it is interesting nonetheless. Looking at all four of these figures, Figure 5 and 6 look extremely similar to Figures 7 and 8. Looking at Figure 6, the higher the percentage of higher education in a county signifies a lower death percentage. This easily relates to Figure 7. This is expected for a few reasons. One reason is the higher the education level, the more income on average. These people tend to live in nicer places with more easily accessible healthcare. Not only is healthcare more accessible, it is also probably more cutting edge. This allows doctors to help patients even more. This means more aspirators and more doctors. These people likely have a better understanding of all of the different ways the virus spreads. This includes knowledge of how easily the virus spreads during different activities. It also includes the knowledge of what lifestyle changes they can make to decrease their risk of getting COVID. Some of these things are only available to wealthier people. For example, grocery delivery and pickup and food delivery are great options to limit your exposure to other people and the virus. However, these options are not cheap. This means they are better suited to prevent their exposure, which leads to less infections and less death. This is portrayed in figures 5 through 8.



**Figure 9: Percent of Confirmed Cases vs. Percent of Low Income Population**

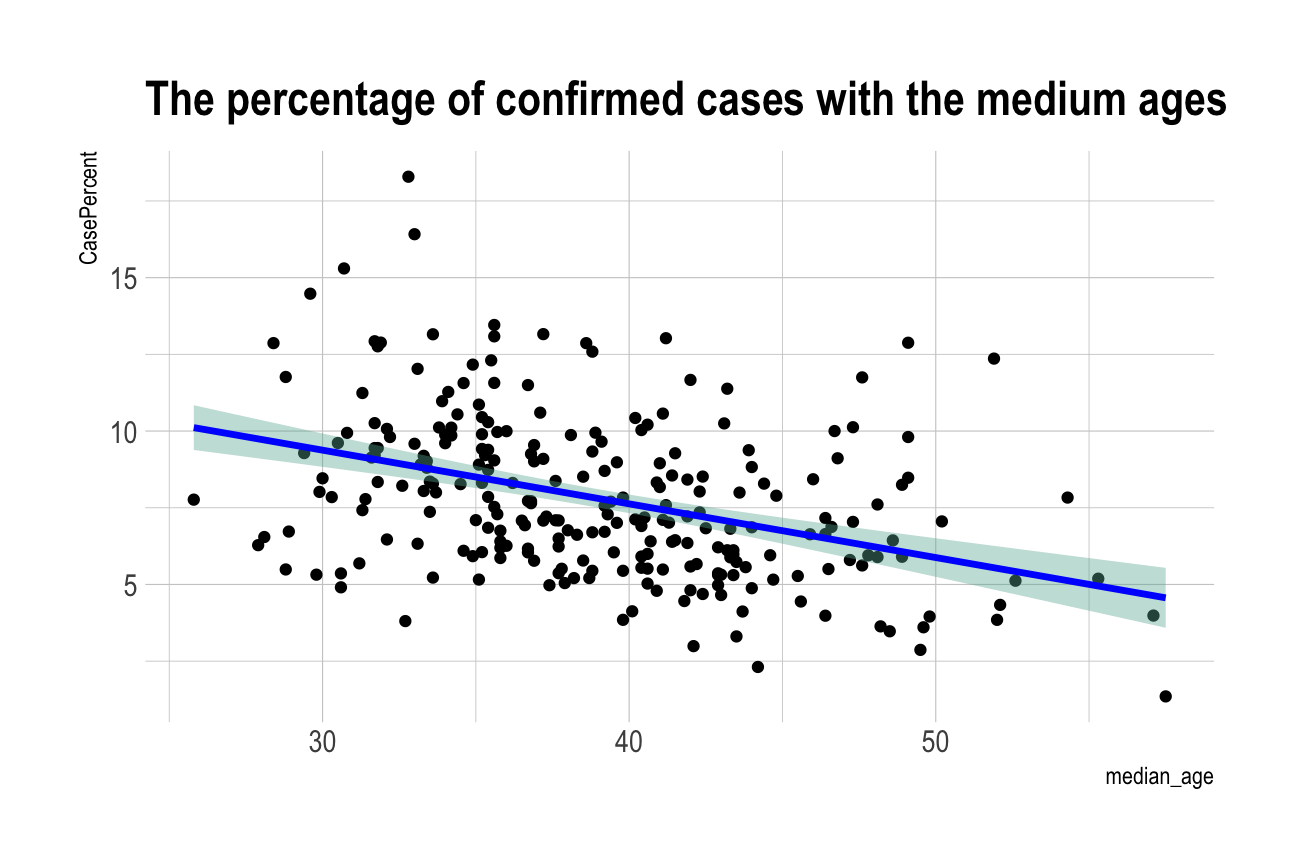
****

**Figure 10: Percent of Confirmed Cases vs. Percent of Population with Low Education**

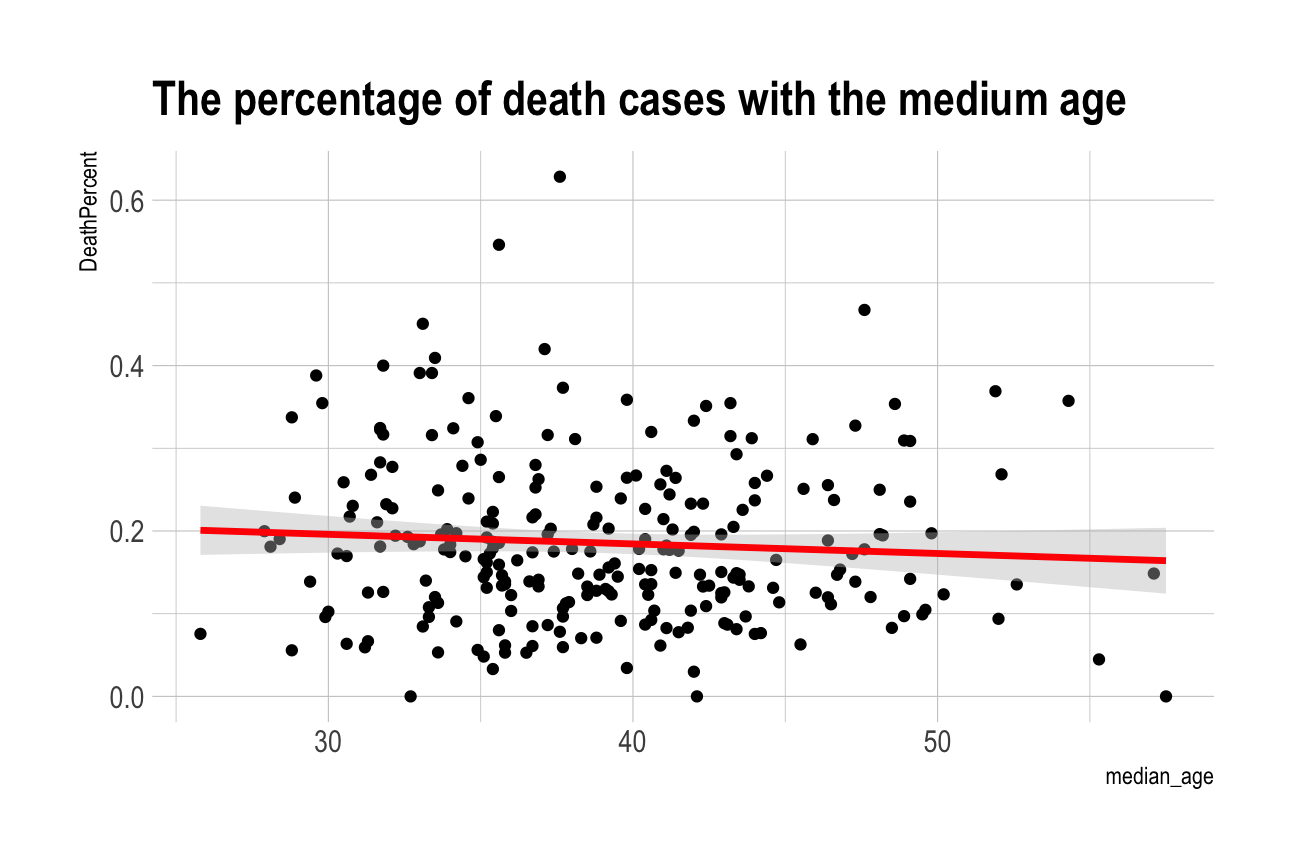
As one would expect, all graphs relating to Wealth and Education follow this same pattern. Lower education typically corresponds to lower income, and the confirmed cases for each of these (shown in Figure 9 and Figure 10) look extremely similar.

*COVID-19 vs. Age*

One of the last comparisons we made compared cases and deaths to the median age of each county, as shown in the following figures.



**Figure 11: Percent of Confirmed Cases vs. Median Age of County**

****

**Figure 12: Percent of Death Cases vs. Median Age of the County**

The data shown on these figures is initially shocking. Figure 11 exemplifies that counties with older median ages had less COVID cases and deaths. This is shocking because we believe the older populations are more at risk for catching the virus, yet they actually got the virus less often than counties with younger median ages. This makes sense when we look at the people’s response to COVID. Many people stopped visiting elderly people in order to protect them from the virus. This means exposure to the virus was limited, and thus it made them less likely to catch COVID. Since they had less likelihood of getting COVID, Figure 12 shows that they had less deaths from COVID.

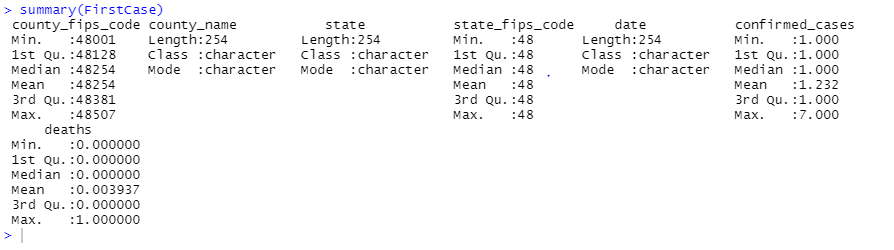
# DATA PREPARATION

Source Code Introduction

There are six sections in the source code, which refer to all the sections in the report. The program will first read in "COVID-19\_cases\_TX.csv" as the data frame called TXCase. It will then read in "COVID-19\_cases\_plus\_census.csv" as a data frame called CensusData.

Section 1: When did the first case happen in each county in Texas.

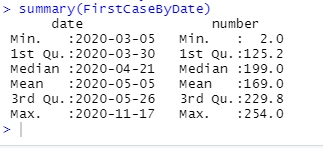
To find out the date of the first case in each county in Texas, it needs to use TXCase, which shows daily confirmed cases by date in each county. It will store the first date for each county that shows a confirmed case into a data frame called FirstCase. In that container, it will have a list of county names, a list of county codes, a list of dates with the first case, and how many cases in that day for each county. The project will create a curved graph with the date and the counties' names to visualize that dataset.



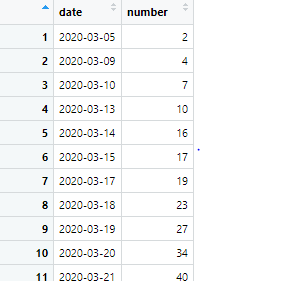
**Figure 13: Statistical Summary of “FirstCase”**

Figure 13 is the statistical summary of “FirstCase.” It has seven columns, in which four of them are numeric. The data frame shows that the dates showing the first case in each county typically have one new case on that day since the mean and the median is all 1. However, the maximum number could reach 7, reflecting that a county could have seven new cases on its “first day.” For the data of death cases, most counties do not have death cases on their “first day,” but some counties do; it may be caused by some cases that are transported from other counties to those counties’ hospitals for further treatment.

To find out how the COVID-19 spread in Texas, the project will keep analyzing the FirstCase; the project will create a new data frame called “FirstCaseByDate.” It is created from FirstCase, but it uses a *ddply* function, which categorizes it by using date. Therefore, FirstCaseByDate will have a list of dates, and it also shows the accumulated number of counties, which already have confirmed cases. In a word, as time passes, how many counties have shown their first case will be reflected on the data frame.



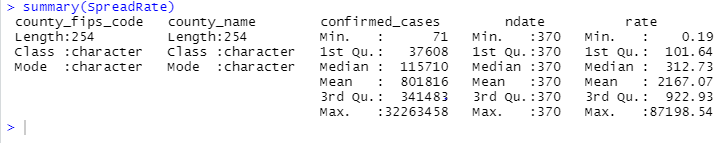
**Figure 14: Statistical Summary of “FirstCaseByDate”**



**Figure 15: Table of “FirstCaseByDate”**

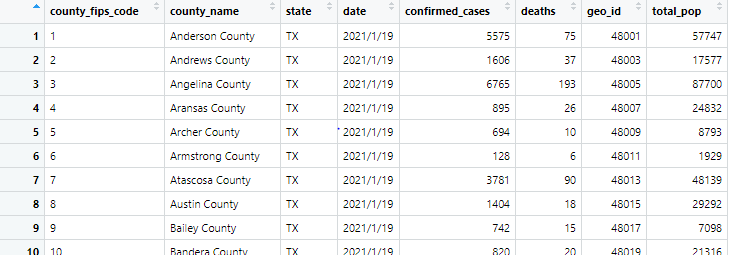
Section 2: How fast did the virus spread in Texas

In this section, the project will create a new data frame called SpreadRate, it has all-255 counties’ names, and it has the total confirmed cases for each county, which is combined from TXCase. Moreover, it also has the total days that count the confirmed cases to get the average new cases per day for each county, which is the spread rate.



**Figure 16: Statistical Summary of “SpreadRate”**

To further analyze the spread rate, the project will compare it with the city’s density (the population of the county & the geographical size of the county). The first step is to get the counties’ total population, and it will be imported from CensusData. The population data frame has the counties’ names with geo\_id, and it also has the total confirmed case and the total population.

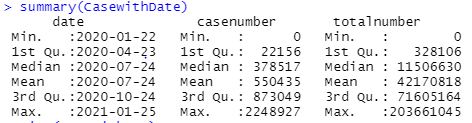


**Figure 17: Table of CensusData**

The second step is to visualize the result by using cartograms. A cartogram is a map in which the geometry of regions is distorted to convey the information of an alternate variable. The region area will be inflated or deflated according to its numeric value. This project will use the map of Texas, “TX-48-texas-counties.geojson.” [4] This *geojson* file is modified so that the county’s list sorts in ascending order. The first map is about the geographical size and each county’s population, and the second map is about how the geographical size relates to the spread rate.

Section 3: What is the social distancing response, and how long did it take after the first case?

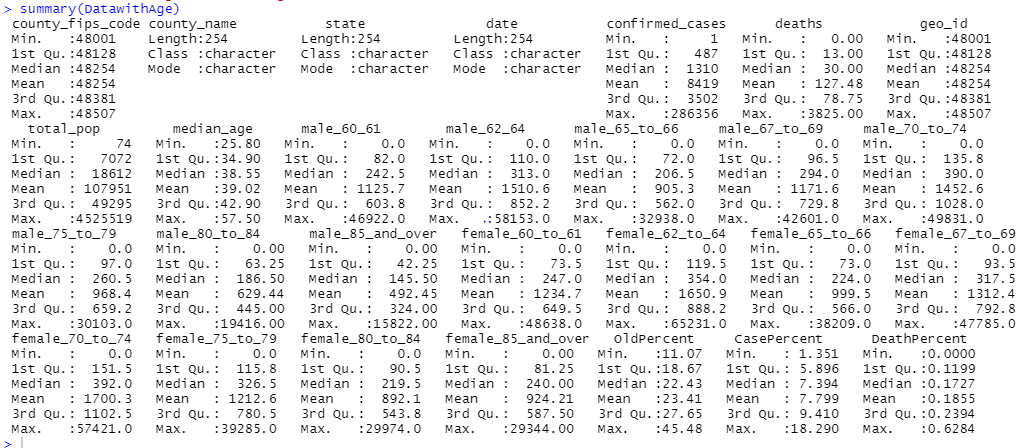
To analyze the impact of social distancing policy in this section, the project uses April 2nd, 2020, which is when the Texas government announced a stay-at-home order and identify how it relates to the spread rate in Texas. The graph has another vertical curve (the x-intercept) with the date announcing the stay-at-home policy. Furthermore, to find its effect on the total number of cases in Texas, the project has created a new data frame called CasewithDate; it has all the dates that are recorded in the TXCase, but compared to FirstCaseByDate, it also has the cumulative number of new cases and total cases in Texas.



**Figure 18: Statistical Summary of “CasewithDate”**

Section 4: Does the spread of COVID-19 relate to ages?

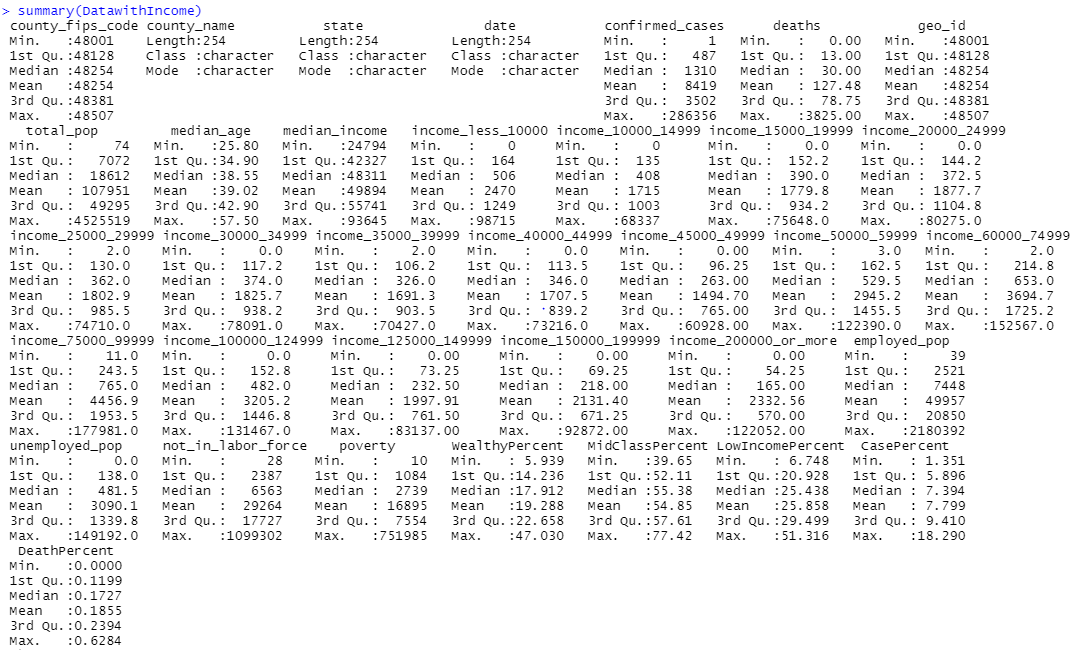
In this section, the project will mainly focus on the CensusData; it has a DatawithAge data frame with a list of counties and with their COVID-19 conditions and civilians' age distribution. Using the total population, the total number of confirmed cases and deaths, and the population under age distribution, we can determine how age could relate to the spread of COVID-19. In the data frame, OldPercent refers to the percentage of older adults (age is greater than 60) in that county; it calculates the total number of people with ages greater than 60 in each county and divides it by the total population of that county. CasePercent simply refers to the percentage of confirmed cases with that country's total population, and the same is true for DeathPercent. The project will also analyze the connection between COVID-19 and age distribution with the data of median age in each county. It is an item from CensusData.



**Figure 19: Statistical Summary of “DatewithAge”**

Section 5: Does the spread of COVID-19 relate to civilians' income?

In this section, instead of using the age distribution, the project will analyze the income condition in each county. The project has created a new data frame called *DatawithIncome*, which imported the income data from the CensusData. It has the total population, total confirmed cases and deaths, the number of employees, and income distribution in each county. The project will mainly analyze the income, but it will also collect the number of people in each class. Poor condition refers to the people whose income is smaller than 25000. The wealthy condition refers to the people whose income is higher than 100,000, and the middle class is the people in between.

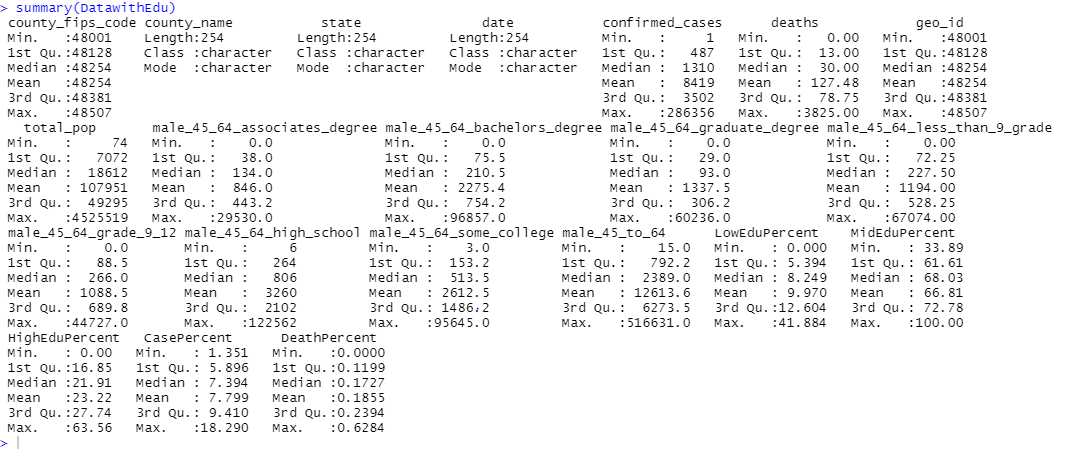


**Figure 20: Statistical Summary of “DatewithIncome”**

This data frame also has three lines: WealthyPercent, MidClassPercent, and LowIncomePercent. These are the percentages of people in each class, and they are collected by dividing the numbers of people in each class with the total population in each county.

Section 6: Does the spread of COVID-19 relate to the educational background? (case select male 40-65)

In this last section, there is a new data frame called *DatawithEdu* and it collects the numbers of people between 40 and 65 with different educational backgrounds. This data also comes from CensusData, but three new lines are added to the data frame: LowEduPercent, MidEduPercent, and HighEduPercent. Like those lines in the *DatawithIncome*, LowEduPercent is the percentage of people with a primary school diploma or does not have any educational background. The HighEduPercent is the percentage of people with a college diploma.



**Figure 21: Statistical Summary of “DatewithEdu”**

# CONCLUSION

It is possible to predict the potential effect of the COVID-19 virus in a county by using the data regarding the county’s demographics. In the report, we mined COVID-19 data in order to come up with the following conclusions:

* The time to prevent the spread of the virus was at the beginning, and no matter what the government did after reopening businesses could significantly reduce the spread
* The greater the wealth in a county means the less likely they were to have a large number of COVID cases and deaths, and vice versa
* The higher the education level in a county means the less likely they are to have a large number of COVID cases and deaths
* The lower the education level in a county means the more likely they are to have a large number of COVID cases and deaths
* The larger the population density in a county means the faster COVID is spread
* The smaller the population density means the slower COVID is spread
* The median age of a county could become a predictor of COVID-19’s effects depending on the actions taken within the county

# REFERENCES

[1] “Coronavirus Disease 2019 (COVID-19).” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 25 Feb. 2021, www.cdc.gov/dotw/covid-19/index.html.

[2] “Social Distancing.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html.

[3] Specktor, Brandon. “Coronavirus: What Is 'Flattening the Curve,' and Will It Work?” *LiveScience*, Purch, 16 Mar. 2020, www.livescience.com/coronavirus-flatten-the-curve.html.

[4] Deldersveld. “Deldersveld/Topojson.” *GitHub*, github.com/deldersveld/topojson/blob/master/countries/us-states/TX-48-texas-counties.json.