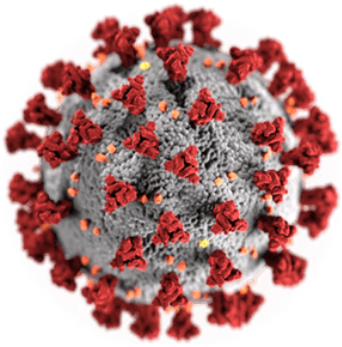


Project 1
Data and Visualization
COVID-19 Dataset

CS 5331 - Data Mining

Authors:

Liam Lowsley-Williams

Emily Fashenpour

Harrison Noble

I. Executive Summary

In early 2020, the United States was introduced to the novel COVID-19 virus which rapidly swept across the nation, infecting and killing thousands of US citizens. States and counties quickly began documenting case numbers and the death toll on a day to day basis to better track the effects of this virus. When infection and death numbers are cross referenced with census data such as total population, median population age, and income, interesting conclusions can be made regarding how the virus spreads in different parts of the country. The focus of this report is to examine COVID-19 and census data from the state of Texas in order to analyze the spread of the virus in counties of different population densities and economic status. Particularly, the focus will be on per county COVID-19 cases and deaths, total population, racial breakdown of the population, median age, and income. To get a deeper understanding of the COVID-19 effects in Texas, relationships and trends must be found between these attributes. The analysis reveals that although small or rural counties in Texas have a much smaller rate of transmission, the likelihood of dying from COVID-19 in these areas is significantly higher when compared to big cities. Additionally, it was found that age is not strongly correlated with death by the virus, suggesting that underlying health conditions are the major factors tying into death from COVID-19.

Table of Contents

Executive Summary **1**

Business Understanding **2**

Data Understanding **3**

Dataset Description 3

Verifying Data Quality 4

Statistical Summary 8

Attribute Visualizations 11

Attribute Relationships 16

Data Preparation **22**

What is The Total Population in Each County? 22

What Does the Spread Rate Look Like per County? 23

When Was The First Reported Case in Each County? 24

What is The Median Age in Each County? 25

What is The Case Fatality Rate in Each County? 26

Conclusion **27**

References **28**

II. Business Understanding

COVID-19 is a new disease, caused by a novel (or new) coronavirus that has not previously been seen in humans. It stems from a coronavirus called SARS-CoV-2. Because it is a new virus, scientists are learning more each day about it and its properties. Although most people who have COVID-19 have mild symptoms, COVID-19 can also cause severe illness and even death. Some groups, including older adults and people who have certain underlying medical conditions, such as heart or lung disease or diabetes, are at increased risk of severe illness [1].

One of the ways health officials say help limit the spread of the COVID-19 virus is to practice social distancing. Social distancing, or physical distancing, is the practice of keeping a safe distance, around 6 feet or more, from others that are not in your household when you are out in public. Other ways to limit the spread are to wear a mask, avoid touching your eyes, mouth, and nose, and to wash your hand with soap and water frequently [2]. The idea of 'flattening the curve' refers to a graph on the number of people who are sick with COVID-19. The curve on the graph that health officials wanted to avoid was a tall and narrow curve, which would indicate many people were sick at once. This would lead to hospitals being overloaded with sick patients, meaning some sick people would not be able to be treated. The ideal graph would have a curve that is very short and wide. This would indicate fewer people have the virus at once and more people would get it over a longer period of time [3]. Essentially, the same number of people would get sick but the infections would happen over a longer time span and hospitals would not be overloaded with patients. By slowing the spread of the virus by implementing strategies like social distancing and mask wearing, the curve would be flattened, the spread of the virus would slow, and hospitals would be able to treat all sick patients.

It is important for us to look at data regarding the virus spread, hospitalizations, and available resources because these factors help us understand how we can better protect ourselves from potentially obtaining the virus and prevent overloading of our healthcare systems. Looking at virus spread will help us understand how quickly the virus spreads and how we can contain it if possible. Bringing down the spread of the virus is imperative to our survival as a society as if less people obtain the virus then less people who fall into a higher risk group can avoid obtaining the virus and we can potentially save lives. Additionally, understanding the trend behind hospitalizations and how, when, or why they occur can prevent people who need intensive care from obtaining the hospital care they need. A big issue we have seen thus far with COVID-19 is overloaded hospitals and running out of room, utilizing the data we have can help us better predict why and when these hospitals have become overloaded. Furthermore, looking at the available resources we have to combat the virus and how such resources are being utilized can promote more efficient use of them so that combatting the virus can be more easily achieved. One of the first major issues we saw with COVID-19 was a lack of supply in PPE which helped significantly suppress the spread. Overall, looking at these key pieces of information in the data

will help us understand exactly what we need to do to efficiently and adequately squash the virus and prevent further fatalities from occurring.

Information regarding COVID-19 spread, hospitalizations, and medical resources can be useful to a wide range of individuals and organizations. State and local governments are interested in this COVID-19 data to gain insights on how their population is handling the virus or how the policies put in place by the government are affecting the spread. Hospitals are interested in this data to determine the amount of medical resources needed in the near future to assist individuals affected by the virus. Local businesses could be interested in localized COVID-19 data in order to see how their policies or operation are impacting the spread of this virus. Organizations researching the impact of masks and social distancing can utilize this COVID-19 data to determine the effectiveness of these policies at the local, state, or country wide level.

III. Data Understanding

Dataset Description

The initial "COVID-19 cases TX" dataset provides the raw data of Texas COVID-19 cases and deaths broken down by county. The timespan for this data begins January 22, 2020, and ends January 25, 2021. Generally, the total number of cases and deaths were updated daily except on Sundays and holidays. Several features of this reported data require clarification, thus, a description of the feature data is provided in Table 1. Additionally, the feature data type is also provided for a more in-depth look at each feature. Nominal data types are used for features without any quantitative value, such as labels. Interval data types are used for features where the difference between values provide meaningful insight, such as dates. Ratio data types are used for features that tell us about exact value, order, and an absolute-zero value which allows for the application of a multitude of statistical inferences. It should be noted that this dataset also includes "state" and "state_fips_code" features, however we decided to remove them because we are only analyzing Texas and do not need the specification.

Table 1: Description of features in "COVID-19 cases TX" dataset

Feature	DataType	Description
county_fips_code	Nominal	Identification number that uniquely identifies geographical area for county
county_name	Nominal	Name of county in US
date	Interval	Analysis horizon
confirmed_cases	Ratio	Total number of confirmed COVID-19 cases in county at the given date

deaths	Ratio	Total number of confirmed COVID-19 deaths in county at the given date

Additionally, we wanted to extract some extra data points from the "COVID-19 cases plus census" dataset in order to gain extra insight on our Texas COVID-19 dataset. Because this dataset accounts for every county in the United States and contains over 200 features for each of those counties, we decided to only look at a select group of features from all the counties in the state of Texas. Table 2 outlines each feature we decided to extract from the dataset. Additionally, the table lists the data type and gives a description of each feature. It should be noted that we only kept census data and excluded the COVID-19 data from this dataset as our first dataset provides a more in depth look at case numbers and deaths.

Table 2: Description of select features from "COVID-19 cases plus census" dataset

Feature	DataType	Description
county_fips_code	Nominal	Identification number that uniquely identifies geographical area for county
county_name	Nominal	Name of county in US
total_pop	Ratio	Total population of county
median_age	Ratio	Median age of county
white_pop	Ratio	Total number of population identifying as White in county
black_pop	Ratio	Total number of population identifying as African American in county
asian_pop	Ratio	Total number of population identifying as Asian in county
hispanic_pop	Ratio	Total number of population identifying as Hispanic in county
amerindian_pop	Ratio	Total number of population identifying as American Indian in county
other_race_pop	Ratio	Total number of population identifying as race other than White, African American, Asian, Hispanic, or American Indian in county
median_income	Ratio	Median income of county
income_per_capita	Ratio	Income per-capita of county

Verifying Data Quality

It is absolutely necessary that we visit verifying data quality prior to actually using the data to make inferences upon the data. It is a crucial process and without performing any verification on the data or preprocessing mistakes are bound to occur as a result of missing data, duplicate entries, NaN values, outlier values, and other poor data quality characteristics. These poor

characteristics will drastically detract from the quality of the data mining and make inferences pulled from such data somewhat useless or biased.
With our dataset, we are looking at COVID-19 data between the dates of January 22, 2020 and January 25, 2021. The reason we are looking at such data is to attempt to understand how the virus spreads geographically and the speed at which the virus spreads so we can draw upon such information to make predictions and analysis to better combat the spread of the virus.

Before we begin looking at our feature data quality descriptions, we found two issues with the first data set that we had to eliminate before moving forward. This first issue we found was in regards to several instances (around 370) with a "county_name" of "Statewide Unallocated". For these particular instances we decided to remove these data rows since we were unsure as to what such a value implied and could not map the data to any individual county in Texas. The second issue we found was in regards to when the data recording started and when the first COVID-19 case in Texas was confirmed. We found the first confirmed case in Fort Bend County, outside of Houston, on March 5, 2020. Therefore, we decided to eliminate all instances of data in each county preceding this date to avoid skewing our statistical data since we are focused on the spread of the virus in Texas alone.

Feature quality and cleaning methods are detailed below:

County_fips_code:
Because this data feature is a part of both datasets, this section will be broken down by individual dataset.

Regarding the first dataset, (outlined above in Table 1) there are many duplicate values, however this is due to the data being a time series of daily cases. As a result, we expect there to be duplicated county FIPS codes. Other than the expected duplication, we have no issues with NaN entries. In fact there were exactly 254 unique values which is exactly how many counties there are in Texas.

Regarding the second dataset, (outlined above in Table 2) there were no issues with duplicate data or NaN entries. Since this dataset only contains one entry per county, we expect to see 254 rows for our selected data for Texas. After inspecting the data, there were exactly 254 unique values which reflects the total number of counties in Texas.

County_names:
Because this data feature is a part of both datasets, this section will be broken down by individual dataset.

Regarding the first dataset, there are many duplicate county names, however we expect this since each county was updating their cases and deaths daily. Because of this, duplicate county names are not an issue for this dataset. We saw no missing or NaN values for this feature.

Regarding the second dataset, there were no issues with duplicate data, NaN entries, or missing values. In fact there were exactly 254 unique values which is exactly how many counties there are in Texas, just like with the fips code.

Date:
When looking at the date, each value occurred approximately 254 times as each entry reflected one day ranging from January 22, 2020 to January 25, 2021 for each county. As mentioned earlier, the first confirmed case in Texas was on March 5, 2020. We updated the range of dates to March 5, 2020 to January 25, 2021 because we are looking at the spread of cases in Texas alone and data from before this date would skew the statistical data. When we checked the data, we saw no missing values and no duplicate dates within the same county, meaning no county recorded the same date twice.

Confirmed_cases:
When looking at the confirmed cases, these values are cumulative according to the date specified for that instance, this meant that the values would build on each other for each county. We were not concerned with duplicate data here as there could be a situation where two counties had the same number of cases or no new cases were confirmed. All of these values were valid and we had no missing values.

Deaths:
When looking at the deaths, these values are cumulative with the corresponding date, meaning over time the values would increase. There was no concern for duplicate data here because it is possible for there to be no new deaths to report, meaning two days would show the same death count. There were no missing values and all were valid.

Total_pop:
When looking at the total population attribute we saw a minimum of 74 and maximum of ~4.5 million. These numbers are valid for the dataset as Harris County (the largest county in Texas) has that maximum number for its population and Loving County has the minimum. Therefore we saw no issues with these numbers. We also checked for NaN values and found none. We were not concerned with duplicate values here as any two counties could have the same population.

Median_age:
When looking at the median age attribute, we saw that the minimum median age in a county was 25.8 years old and the maximum median age found in a county was 57.5. We saw no issues with

these numbers after closer examination and determined there were no outliers present within this feature. We checked for NaN values and found none. There is no concern for duplicate values here as any two counties could have the same median age.

White_pop:
When looking at the White population feature, we found that the minimum value of any county was 55 and the maximum was ~1.38 million. When we went in and checked these counties, we found that the minimum was tied to the county with the lowest population and the maximum was tied to the county with the highest population. Thus, we can conclude that these entries are correct. Additionally, we found no missing data or malformed entries for this feature.

Black_pop:
When looking at the Black population feature, we found that the minimum value for some counties was 0. After going in and checking these counties manually, we discovered that these counties have very low total populations. Thus, we can conclude that these entries are correct. We found no missing data or malformed entries for this feature.

Asian_pop:
Similar to the Black population, we found that the minimum Asian population for some counties was 0. After going in and checking these counties manually, we discovered that these counties have very low total populations. Thus, we can conclude that these entries are correct. Again, we found no missing data or malformed entries for this feature.

Hispanic_pop:
When looking at the Hispanic population feature, we found a minimum value of 12 and maximum value of ~2 million. After going in to check on these numbers we identified that the maximum number was in Harris County which is quite far south and the minimum was in Loving County which was the smallest county in Texas. Thus, we can conclude that these entries are correct. Again, we found no missing data or malformed entries for this feature.

Amerindian_pop:
When looking at the American Indian population feature, we found that the minimum value for some counties was 0. After going in and checking these counties manually, we discovered that these counties have very low total populations. Thus, we can conclude that these entries are correct. Again, we found no missing data or malformed entries for this feature.

Other_race_pop:
When looking at the other Race population feature, we found that the minimum value for some counties was 0. We were not concerned with this as the majority of the most prevalent races in

the world were covered above. Thus, we can conclude that these entries are correct. Again, we found no missing data or malformed entries for this feature.

Median_income:
For the median income feature, we found the minimum value to be $24,794 and the maximum value to be $93,645. Given the vast socioeconomic differences between rural areas and the large cities in Texas, we do not consider these values to be outliers. Additionally, we found no missing data within this feature. We are not worried about duplicate data for this feature as it is entirely possible for two counties to have the same median income.

Income_per_capita:
For the income per capita feature, we found the minimum value to be $12,543 and the maximum value to be $41,609. Like with median income, we do not consider these values to be outliers due to the vast socioeconomic differences between rural areas and the large cities in Texas. Additionally, we found no missing data within this feature. We are not worried about duplicate data for this feature as it is entirely possible for two counties to have the same income per capita.

Statistical Summary

The statistical summary of all features for the COVID-19 cases TX and our selected features in the COVID-19 cases plus census datasets are given in Table 3 and Table 4 respectively. By separating the datasets into two different statistical summaries, we can better understand the differences between the overlapping features in both our datasets. A majority of the data in the first dataset is classified as nominal, meaning the only statistical inferences we can draw from these features are mode and frequency. Table 4 also contains some nominal features, all of which overlap with the Table 3 nominal features, however the majority of the feature data is ratio. We will be able to perform much more statistical analysis on the ratio data.

In regards to Table 3 shown below, **this dataset specifically looked at cumulative case and death counts across counties starting with the first confirmed case at 3/5/20 and up to the date 1/25/21.** Starting with the county_fips_code and county_name the two are related, one is an ID number and the other is the name of the county. The mode for both of these is simply the first county we see in the dataset with a frequency of 327 which is how many instances we saw for each county with recorded data. For the date, we saw a mode of 3/5/20 since that was the first date at which we saw a confirmed case and a frequency of 254 since there is only one entry per county per date specified. **We also saw a minimum here of 3/5/20 and a maximum of 1/25/21 since that was the start and end date at which the data collection occurred**. This specified that there were approximately 327 entries for each county inclusive between those dates. Looking at the confirmed cases we started to see some more interesting data. We saw a mean of 2,452.03, a median of 143, a mode of 0, a frequency of 9,766, a standard deviation of nearly 12,697.85, a variance of 161,235,386, a minimum of 0 and a maximum of nearly 297,629. When

analyzing these statistics more in depth, two common characteristics are tied across them all. **Firstly, this data is cumulative data, which means that rural counties may not see as many cases as large cities would earlier on in the pandemic. Secondly, the values of the mean, median, standard deviation, and maximum imply that there is likely a skew to the data across counties.** This is because the median is far lower than the mean and the maximum is far higher than the mean. **We likely see skew here because some rural counties with very low populations may have not seen as much spread as higher population counties located in densely populated cities. These inferences should be further analyzed to see if such a correlation can be identified, which is something we should expect when dealing with a virus that thrives on person-to-person transmission**. Moving on to the last attribute of deaths, we saw a mean of 43.37, a median of 3, a mode of 0, a frequency of 27,838, a standard deviation of 199.79, a variance of 39,917.28, a minimum of 0, and a maximum of 4,024. When looking at these statistics more in depth, we saw similar characteristics as we did with confirmed cases, just on a lower scale. **Again, we can make the same inferences regarding the characteristics mentioned above about cumulative data and a skew between rural counties and non-rural counties.** Lastly, **It may be worthwhile to further analyze the death data between counties to see if counties with more advanced medical systems (larger cities) were able to prevent deaths from occurring more effectively than rural counties in relation to confirmed cases.**

Table 3: Statistical Summary of features in COVID-19 cases TX dataset

Feature	DataType	Mean	Median	Mode	Frequency	St. Dev	Variance	Min	Max	Range
county_fips_code	Nominal	---	---	48001	327	---	---	---	---	---
county_name	Nominal	---	---	Anderson County	327	---	---	---	---	---
date	Interval	---	---	3/5/20	254	---	---	3/5/20	1/25/21	---
confirmed_cases	Ratio	2,452.03	143	0	9,766	12,697.85	161,235,386	0	297,629	297,629
deaths	Ratio	43.37	3	0	27,838	199.79	39,917.28	0	4,024	4,024

As shown below, **Table 4 specifically looks at census data regarding population size, age, race, and income across all Texas counties.** Starting with county_fips_code and county_name, the mode for each is simply the first county and its respective ID number we see in the dataset. The frequency reveals that each county is only recorded once within the dataset. Looking at total_pop, we can see quite a spread between mean and median which is expected as our range of the population across counties in texas is roughly 4.5 million, this is due to the large cities and rural areas. The frequency of the mode is simply 1 which we are not very concerned with as no two counties are likely to have the exact same number of people. The standard deviation value is around 390,000 which is also expected as the range between the minimum and maximum is so large. Now looking at the median age feature, we can see a mean of 39.02 and median of 38.55.

We expect this because there are a lot of young people in Texas and especially in cities and it is not necessarily considered a retirement state. We are not too concerned with the mode here as identifying counties with the exact same median age is not super important to our objective. What is interesting though is the minimum and maximum median age. **We have a minimum median age of 25.8 and a maximum median age of 57.5. It would be interesting to see how the deaths over time in counties with a higher median age compare with those of a lower median age, as one might suspect that older persons are more at risk of dying.** When looking at the white_pop feature, we can see a mean of around 46,000 and a median of about 9,400. Like the total population feature, this difference is expected as our white population range is 1,386,521 due to large cities and rural areas. This difference is also reflected in the large standard deviation of roughly 136,000 which helps explain the spread of white population across counties. However, we can see some surprising numbers regarding the minority populations within our dataset. Looking at black_pop, asian_pop, amerindian_pop, and other_race_pop, we can see that the mode of each of these features is 0. What's even more interesting is the frequency at which this mode appears. There are 12 counties with 0 Black individuals, 37 counties with 0 Asian individuals, 40 counties with 0 American Indians, and 121 counties with nobody identifying as other race. The only minority group that did not fall such criteria were the hispanic population. **Thus it may be interesting to see if counties with a higher percentage of minority groups experienced a higher number of confirmed cases or deaths or vice versa. It may also be interesting to identify in which counties these minority groups reside in higher populations, as the population size of a given county may be the determining factor of such a relation.** Lastly, looking at median income and income per capita we can see that based on the standard deviations and means, these are the least skewed data of the data set we have seen yet. We see that our minimums for both features are roughly around 24,000 and the maximum for the median income feature is 93,645 while the maximum for the income per capita is 41,609. These numbers are expected as we could imagine that large cities would have higher incomes due to large corporations and rural areas would have lower incomes. **It will be interesting to take a look at how COVID-19 affects lower income counties versus how it affects higher income counties.**

Table 4: Statistical summary of select features in COVID-19 cases plus census dataset

Feature	DataType	Mean	Median	Mode	Freq.	St. Dev	Variance	Min	Max	Range
county_fips_code	Nominal	---	---	48195	1	---	---	---	---	---
county_name	Nominal	---	---	Hansford County	1	---	---	---	---	---
total_pop	Ratio	107,951.2	18,612.5	5532	1	389,476.9	151,692,226,474	74	4,525,519	4,525,445
median_age	Ratio	39.02	38.55	35.4	5	5.97	35.59	25.8	57.5	31.7
white_pop	Ratio	46,281.47	9,404	1528	2	136,770.4	18,706,145,863	55	1,386,576	1,386,521

black_pop	Ratio	12,594.57	676	0	12	67,723.32	4,586,448,514	0	838,285	838,285
asian_pop	Ratio	4,814.86	73.5	0	37	25,904.23	671,029,182	0	307,109	307,109
hispanic_pop	Ratio	42,023.26	5,068.5	490	2	172,210.5	29,656,465,741	12	1,910,535	1,910,523
amerindian_pop	Ratio	259.38	41	0	40	766.95	588,213.2	0	8,078	8,078
other_race_pop	Ratio	154.15	2.5	0	121	760.46	578,293.2	0	9,681	9,681
median_income	Ratio	49,859.34	48,311	42500	2	12,132.68	147,201,815	24,794	93,645	68,851
income_per_capita	Ratio	24,859.02	24,284.5	21938	2	5,240.75	27,465,478	23,543	41,609	29,066

Attribute Visualizations

We decided to look into confirmed cases and deaths as a whole, confirmed cases and deaths per 1000 people by county, total populations by county, populations of race groups, median age, and median income attributes to gain a more detailed understanding about the effects of COVID-19 on the Texas population.

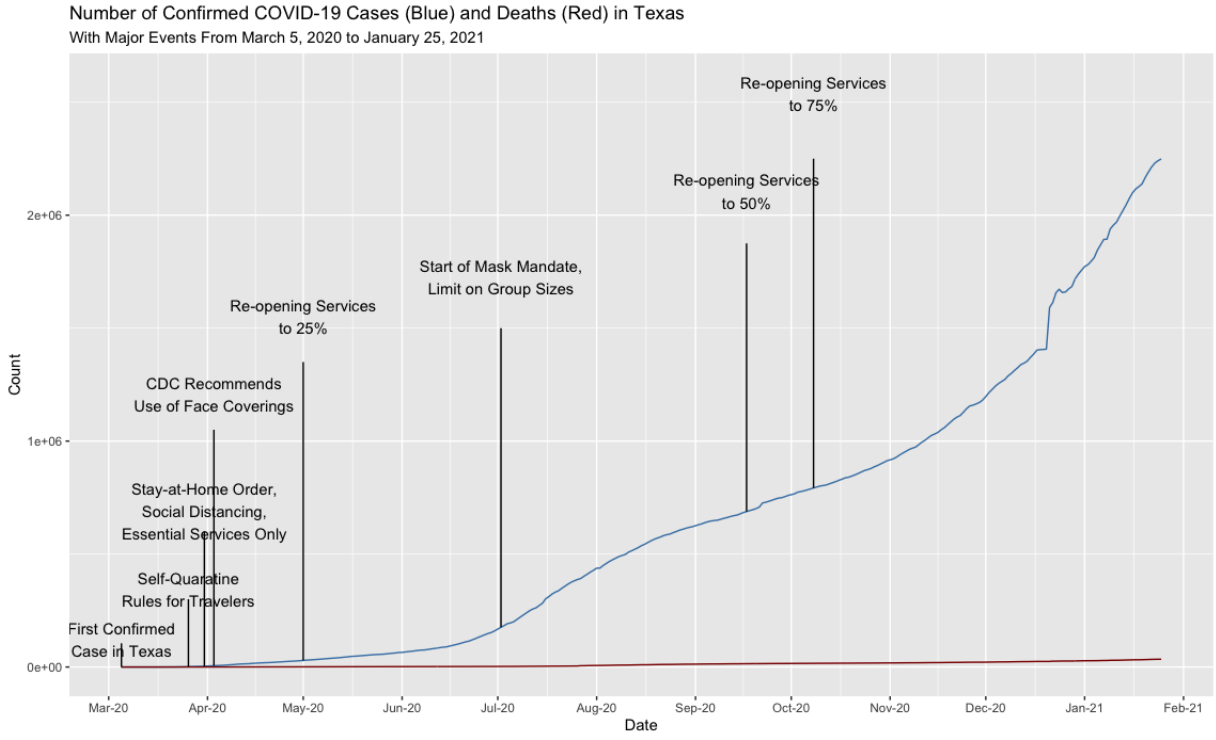
A. Total Confirmed Cases and Deaths (Ratio)

We first wanted to take a look at the confirmed COVID-19 cases and deaths in the entirety of Texas as well as some major events pertaining to COVID-19 policy. In Figure 1, shown below, the blue line represents the total number of COVID-19 cases and the red line represents the total number of COVID-19 related deaths. Additionally, major events such as the beginning of the mask mandate and the reopening of businesses to a certain threshold are labeled along the x-axis.

One assumption that we made in regards to the placement of events on this graph and their relation to the data is that the **effects of major events are not realized until a few weeks to a month after the policies were put in place indicating a lag. Therefore, the following inferences look not at the exact date at which the event occurred but the surrounding area after the fact.** First looking at the reopening of services to 25% we see a gradual increase in the number of cases. This could be attributed to people starting to go out and get near each other. Then, **once the mask mandate begins, we can see that the slope of the graph slowly decreases, meaning the spread starts to slow down**. When services reopen to 50% we see a small jump in the number of cases likely due to the increase in crowding at places like restaurants, bars, and shopping centers. When services open to 75% we see the slope increase even more meaning the virus starts to spread faster. We can also see a large jump in cases between December 2020 and January 2021, likely due to holiday travel and colder months, which ultimately leads to a steeper slope in the graph towards the last recorded date. **Although we see sharp increases in total case numbers throughout this graph, it is not necessarily**

reflected in the total death count, which follows a relatively consistent slope throughout the timeline.

Figure 1: Confirmed Cases and Deaths in Texas with Major Events



B. Confirmed Cases and Deaths per 1000 (Ratio)

Below we can see 3 figures all pertaining to the confirmed cases and deaths per 1000 people grouped by counties. First looking at Figure 2 we see a somewhat normal distribution with a small skew to the left. This skew is likely explained by the larger and more densely populated cities such as Dallas, Houston and San Antonio having higher case counts due to their size. However, for **the vast majority of counties we can see that they fall between 50 and 100 cases per 1000.** Looking at Figure 3 we see another somewhat normal distribution with another skew to the left. Here the skew is a little more significant than what we saw with case counts and that is again likely caused by the larger more densely populated cities where spread rates are likely to be higher. **One major difference we see here is that the deaths per 1000 people does not really exceed more than 6 people per 1000.** This is interesting because it tells us that even though we have counties with a high number of cases per 1000, the deaths per 1000 are relatively low compared to that number. Perhaps we can deduce from this information that the likelihood of dying from COVID-19 once getting it is not a high percentage. To better understand just how great this difference is between deaths and cases, we can look at Figure 4 which contains the two prior figures plotted on one chart. Here we can see just how segmented the deaths are versus the

cases. **This information allows us to make a conclusion that even though case counts may be high, that does not necessarily mean that death counts will be high with it.** It also allows us to make the assumption that the **likelihood of dying from COVID-19 once obtaining it is not as high as one might suspect making the recovery quite high**. We will further analyze the relationship between deaths and cases later on in this report.

Figure 2: Cases per 1000 by County

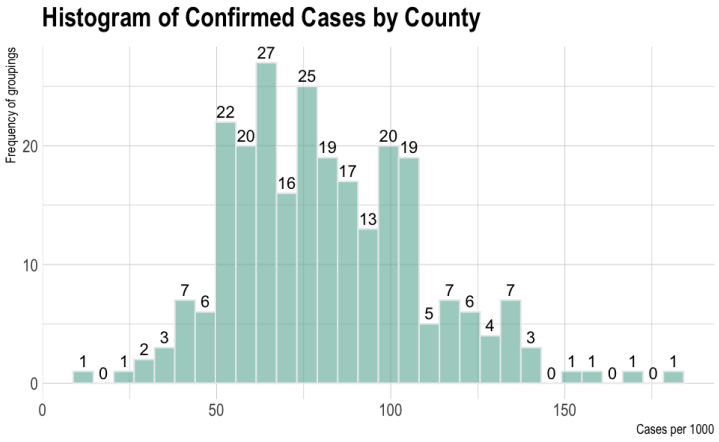


Figure 3: Deaths per 1000 by County

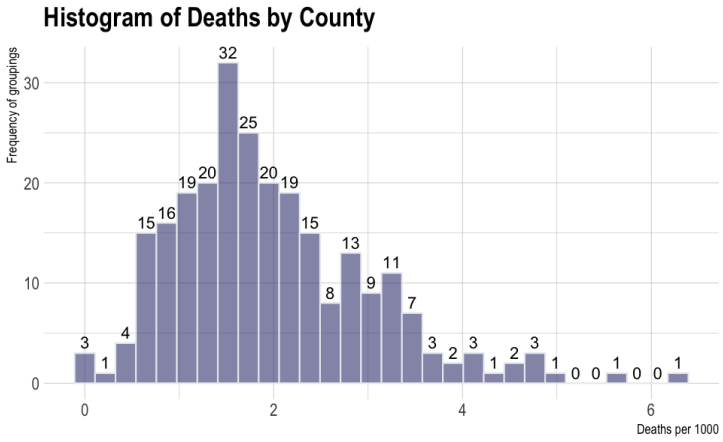
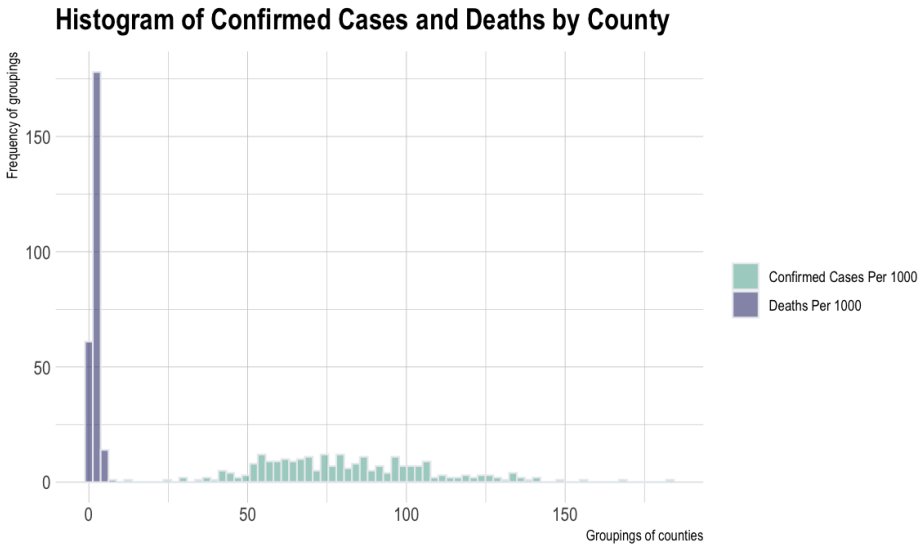


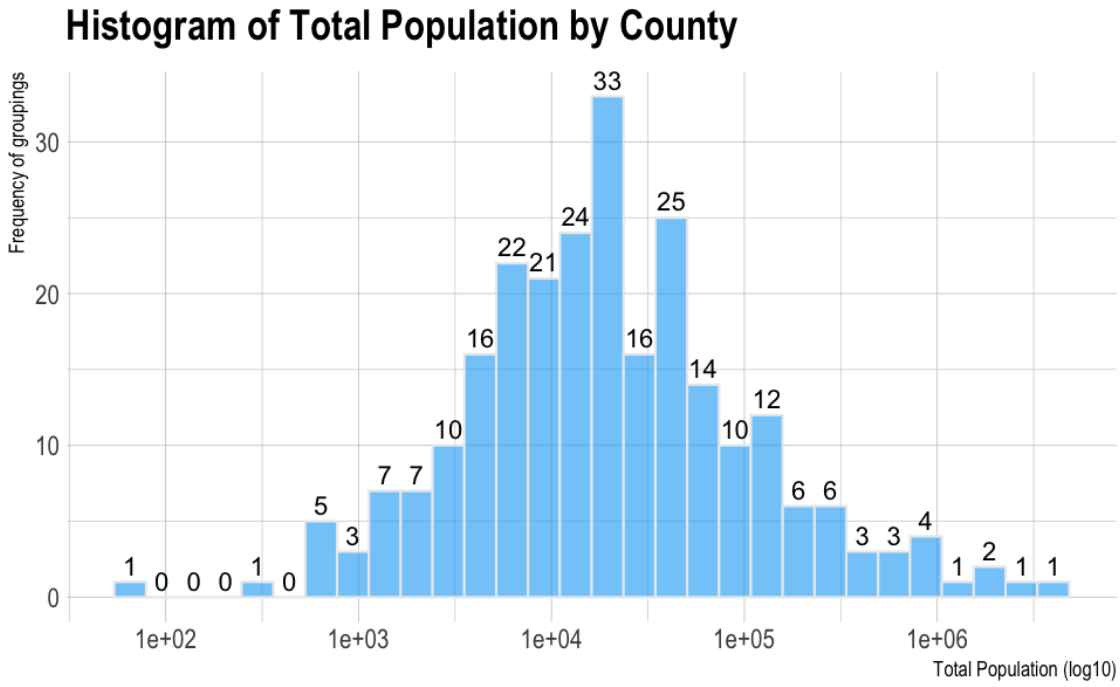
Figure 4: Confirmed Cases and Deaths per 1000 (Combined)



C. Total Population (Ratio)

Below we can see Figure 5 depicting the total population grouped by counties and the frequency at which that group occurs. To make this graph more readable we applied the log base 10 scale to the x axis for total populations. We did this due to the fact that the large cities within texas are so densely populated that the smaller rural counties were all grouped near 0. Here we can see normal distribution which is expected due to the log scale we applied. We can see that the vast majority of counties fall between 10,000 and 100,000 which is exactly where our median and mean values lie.

Figure 5: Population per County (Log Scale)

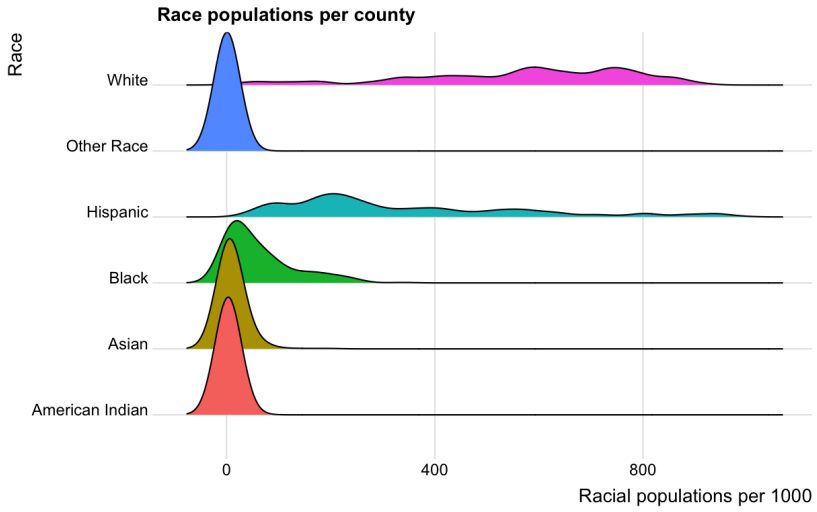


D. Race Populations (Ratio)

Below we can see Figure 6 depicting the different racial populations per 1000 people by county split by the different races we saw in our dataset. Here we used a ridges plot to depict the concentrations of races throughout Texas to see what races had the highest populations. As seen in the figure, the race with the highest concentrations is White followed by Hispanic, Black, Asian, American Indian and finally other races are last. It is interesting to see this demographic data as it may be useful to observe the relationship of race to other features later on in this report.

It should also be noted that a high concentration of hispanic people does make sense given the proximity of Texas to Mexico.

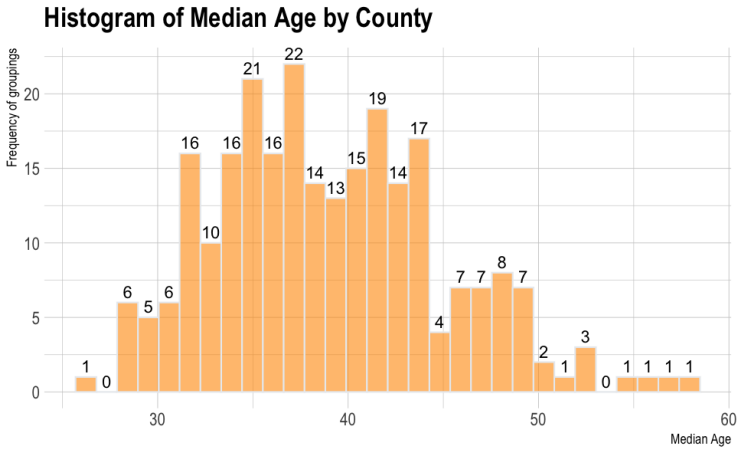
Figure 6: Race populations per county



E. Median Age (Ratio)

Below in Figure 7 we can see the median age grouped by counties and the frequency at which the grouping occurs. We see a somewhat normal distribution again with a slight skew to the left if even a skew at all. Our highest concentration falls between the ages of ~ 33 to 45 which makes sense considering Texas is not really considered a retirement state where you would find many old people. It will be interesting to see how this feature relates to other features such as case count or income later on in this report.

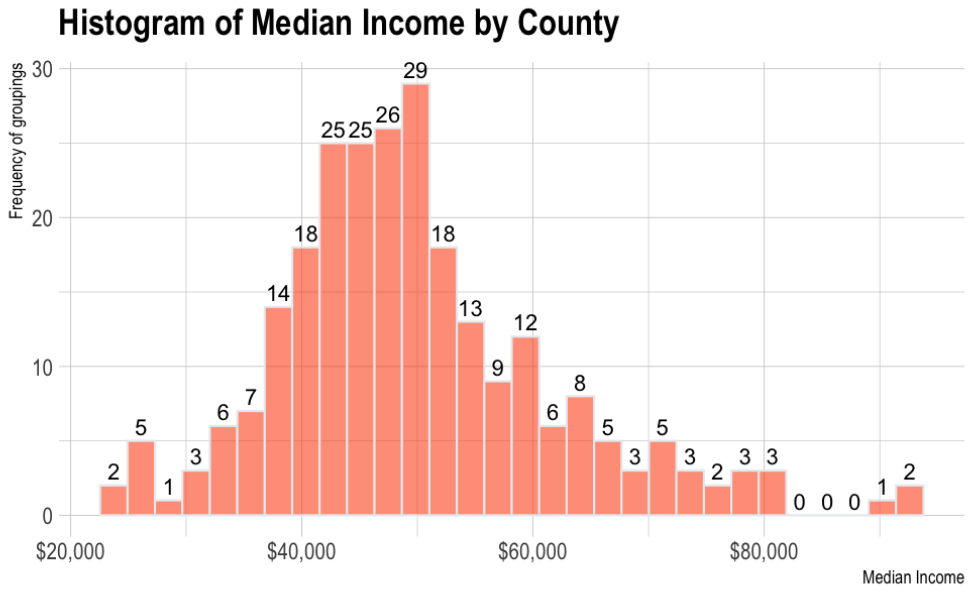
Figure 7: Median Age by County



F. Median Income (Ratio)

Below in Figure 8 we can see the median income grouped by counties and the frequency at which the grouping occurs. We see a somewhat normal distribution again with a slight skew to the left. What's interesting here is that we have a few outliers with median incomes close to $90,000 while the vast majority of counties had a median income between $40,000 and $50,000. We will also be taking a look at median income and how it relates to other features later on in this report.

Figure 8: Median Income by County



Attribute Relationships

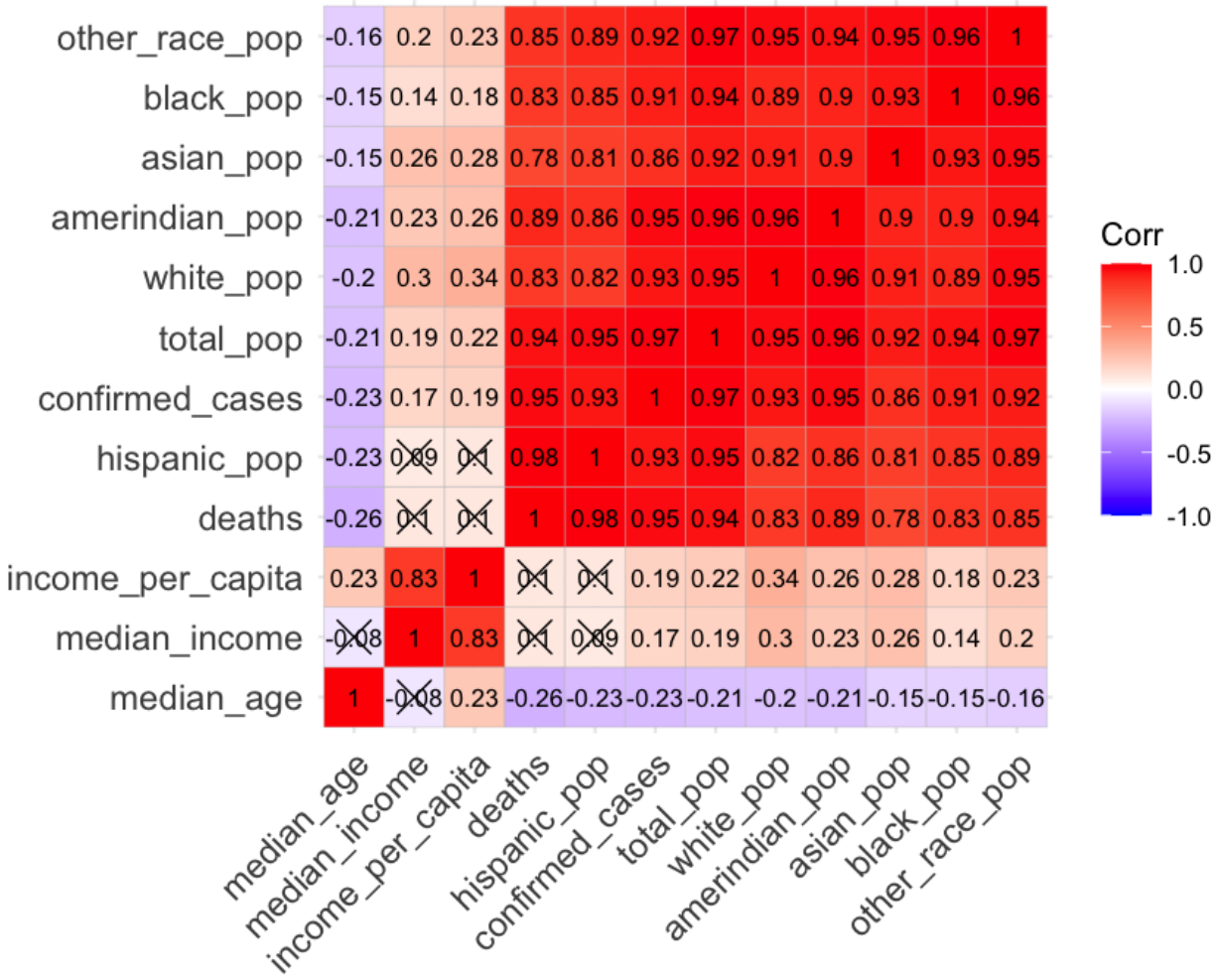
Now that we have a better understanding of the attributes in our dataset, we can begin analyzing the relationships between census data and COVID-19 data to draw some conclusions pertaining to how the virus spreads.

A. Correlation Heatmap of all Data Attributes

Below in Figure 9 we see a heatmap of all our data attributes depicting the correlation between them. As you can see the diagonal is equal to one as the correlation between the same attributes will undoubtedly be very high. Boxes marked with an X signify that the correlation coefficient is insignificant and between -0.1 to 0.1 inclusive. What is interesting about this correlation heatmap is that there seems to be a strong correlation between racial populations and deaths. This

is expected as with larger populations you would have more deaths and vice versa. We can confirm this by looking at the correlation between deaths/cases and total population and we see high coefficients there as well. However, one racial population that has a lower coefficient than the rest is the Asian population. **The correlation between Asian populations and confirmed cases/deaths is lower than all the rest. This could imply that the Asian population might be doing something different than other populations in protecting themselves from contracting the virus.** One thought could be that Asian cultures around the world more frequently wear masks when sick whereas other cultures, such as american or european, are less likely to do so. Additionally we see **there is a negative correlation between median age and deaths** which will be described further in the next section.

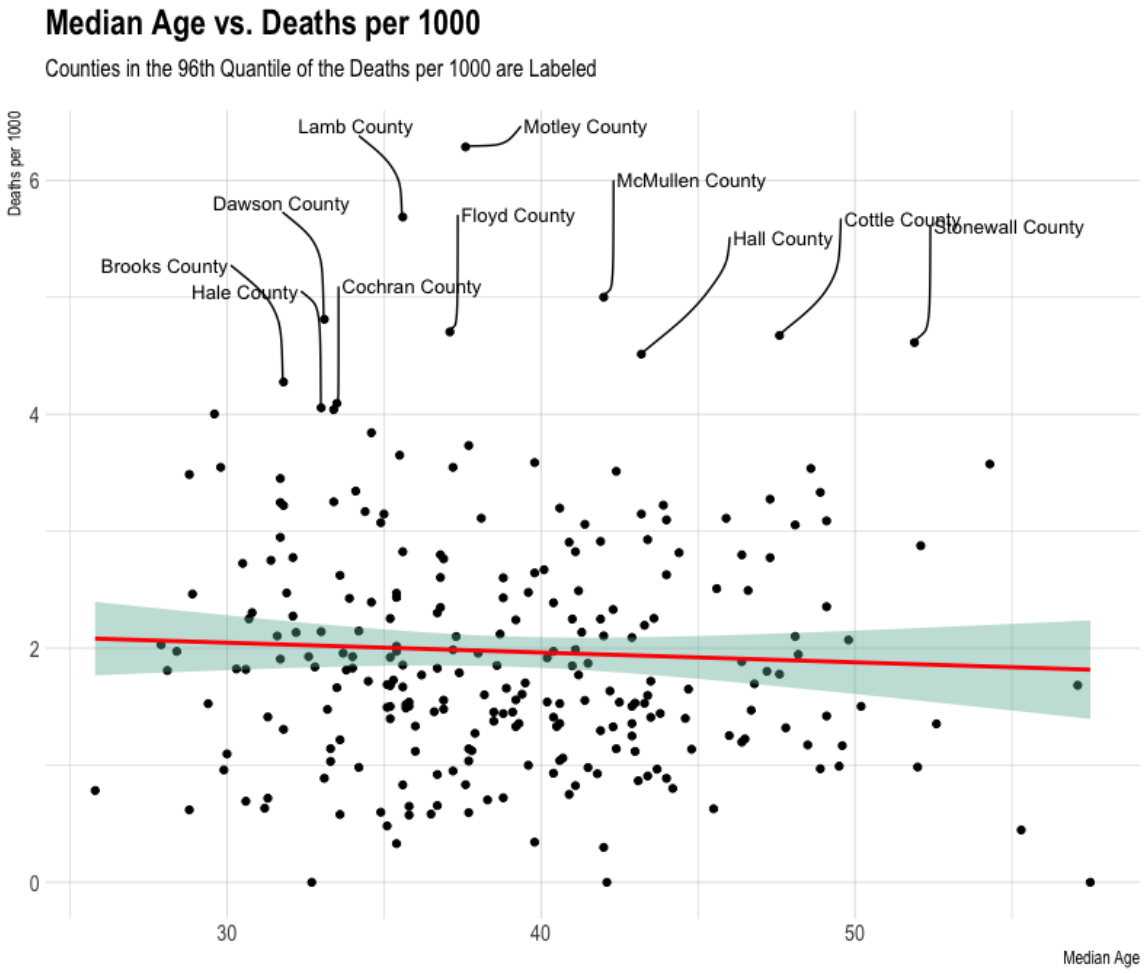
Figure 9: Correlation Heatmap of Attributes



B. Median Age vs. Deaths per 1000 People

The scatter plot, Figure 10 below, contains a trend line to help illustrate the relationship between median age and deaths per 1000 people in each county in Texas. We found the slope of the trendline or correlation coefficient to equal -0.05. This value is very near zero which means there is no relationship or correlation between median age and deaths per 1000 people. Health officials have warned that older people are more at risk from perishing from COVID-19 and we expected to see this relationship once we graphed the attributes. However, this is not the case. Instead, we see that there is a relatively even spread of points across the whole graph. **This suggests that the biggest risk of death with COVID-19 is not age**. This also makes sense because health officials have said people with pre-existing conditions, like diabetes, obesity, etc., are also very at risk.

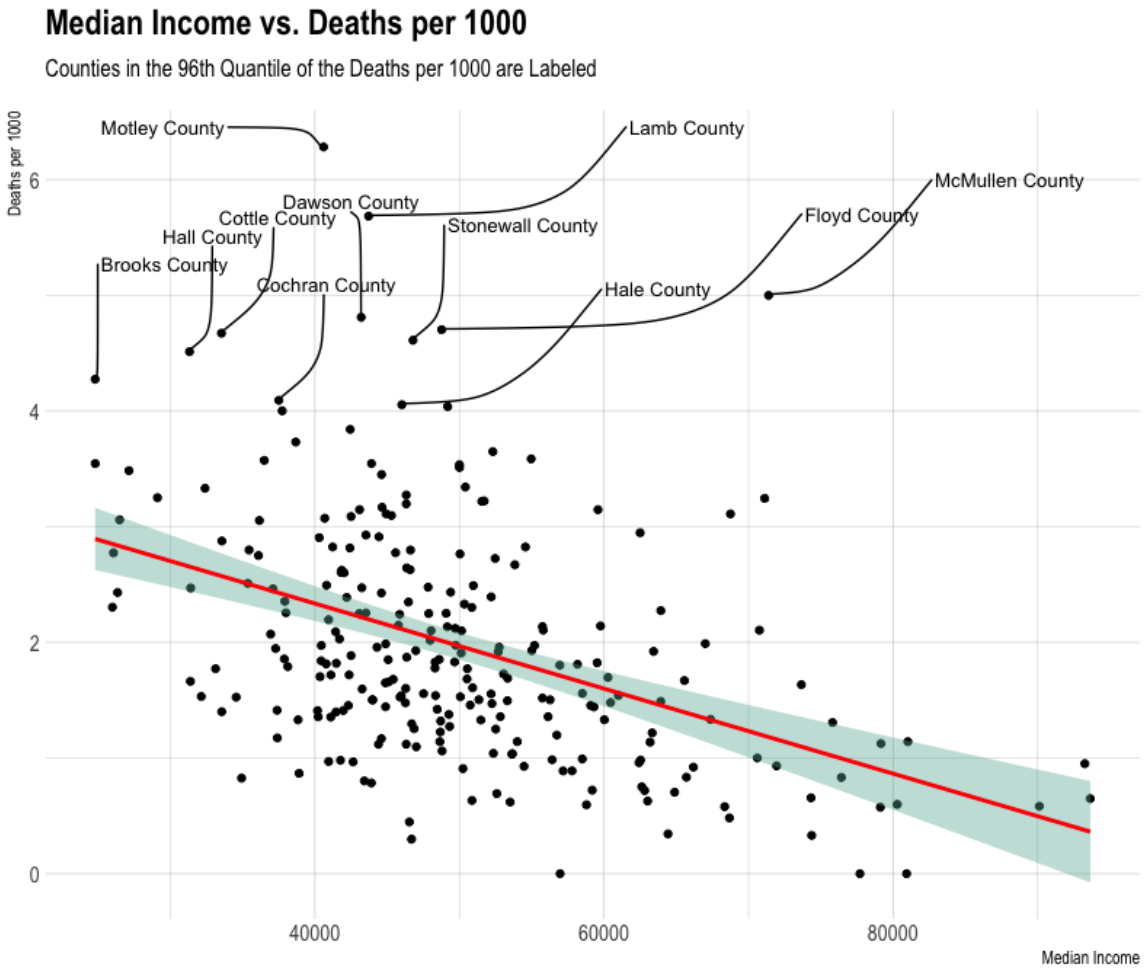
Figure 10: Scatterplot of Median Age and Death per 1000 People



C. Median Income vs. Deaths per 1000 People

The scatter plot from Figure 11 contains a trend line to help illustrate the relationship between median income and deaths per 1000 people in each county in Texas. We found the slope of the trendline or correlation coefficient to equal -0.43. This means there is a somewhat strong negative relationship or correlation between median income and deaths per 1000 people. This negative correlation means as median income in a county increases, the deaths per 1000 people decreases. This could indicate a variety of things. For instance, in a **county that has a higher median income, there is more money available from taxes to fund hospitals and the building of hospitals in that county, meaning that people who get COVID-19 have access to more and better funded hospitals for treatment. This could show how counties with a higher median income have more people who can easily pay for treatment.**

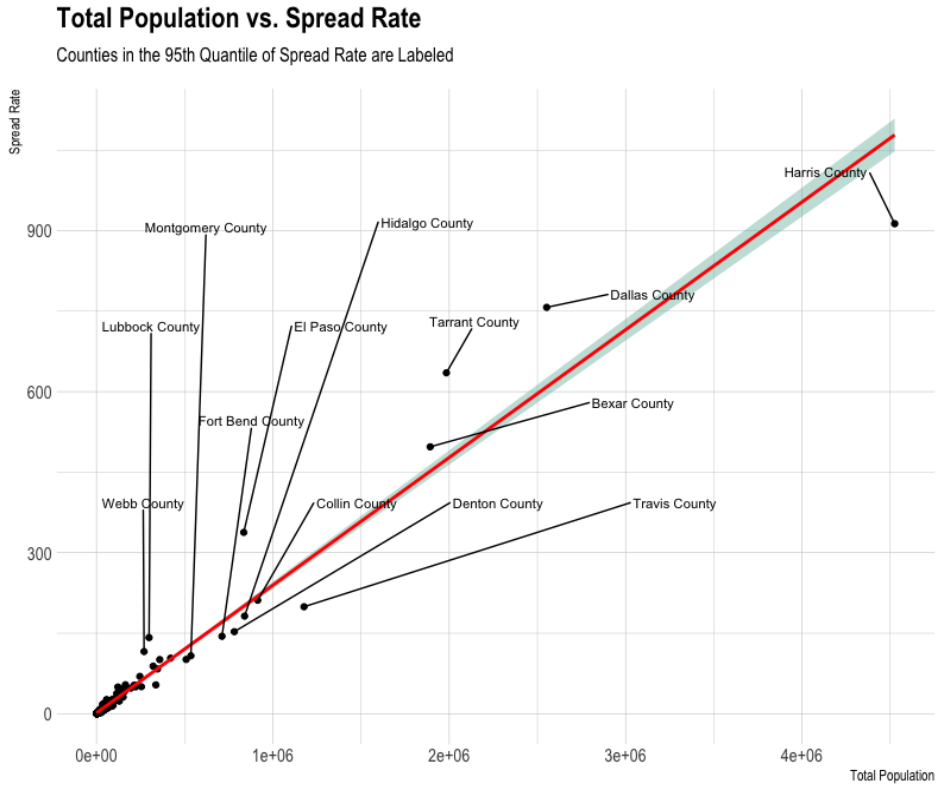
Figure 11: Scatterplot of Median Income and Death per 1000 People



D. Spread Rate vs. Total Population

Below, we can see Figure 12. The scatter plot contains a trend line to help illustrate the relationship between COVID-19 spread rate and the total population in each county in Texas. **Spread rate is the linear slope or rate at which COVID-19 cases accumulated over time.** For example, if the spread rate is near 5, then that county had an average of 5 new confirmed cases per day each day between March 5, 2020 and January 25, 2021. Spread rate was found by finding the difference in the number of cases on March 5, 2020 and the number of cases on January 25, 2021 and dividing that value by the total time in days between those two dates. We found the slope of the trendline or correlation coefficient to equal 0.97. Since the correlation coefficient is very close to 1, this means there is **a very strong positive relationship between spread rate and total population**. This positive correlation means as the total population in a county increases, the rate at which the virus spreads also increases. This strong positive relationship is not a surprise. Larger cities were hit a lot harder with this virus because there are more people occupying a smaller space so the virus spread a lot quicker, especially once services were allowed to open again. In Dallas county, which has a population of nearly 2.5 million people, has a spread rate of nearly 757. Comparing this number to the smaller Anderson county, which has a population around 57,000 people and a spread rate around 17.

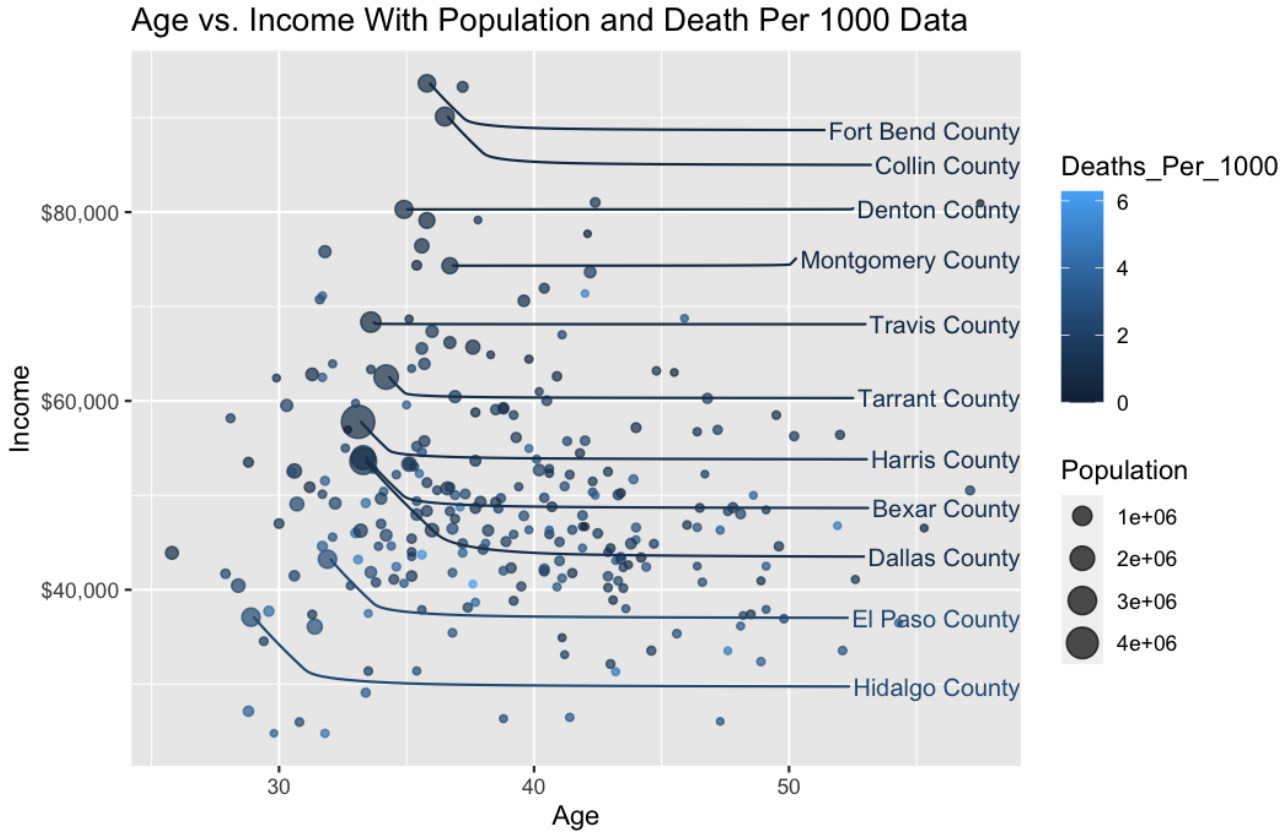
Figure 12: Scatterplot of Spread Rate and Total Population



E. Relationship Between Age, Income, Death per 1000, and Population Data

Below in Figure 13 we can see the Age vs. Income bubble chart. Here we have age on the X-axis, income on the Y-axis, and population depicted by bubble size. We also can see the deaths per 1000 people through the shading of the bubble. A darker shade indicates a lower death per 1000 count while brighter shades indicate the opposite. This figure provides us with some interesting data regarding the socioeconomic status of certain cities and their populations. As you can see the vast majority of larger populations contain younger median ages while smaller populations contain older ages. Additionally, we can see that some of the larger populations also have higher incomes. This is expected as more high paying corporations should reside in such populations. **Furthermore, we can see that the vast majority of counties with a lower median income experience higher death rates per 1000 which could indicate less access to advanced health care.**

Figure 13: Age vs. Income with Population and Death per 1000 Data per County



IV. Data Preparation

Now that we have visualized and found multiple relationships between the attributes, we can now ask some interesting questions regarding the features. Using the datasets from the data understanding section, we compiled a new dataset with some new features to help us answer the questions asked in this section. In Table 5 below, we can see our newly created dataset:

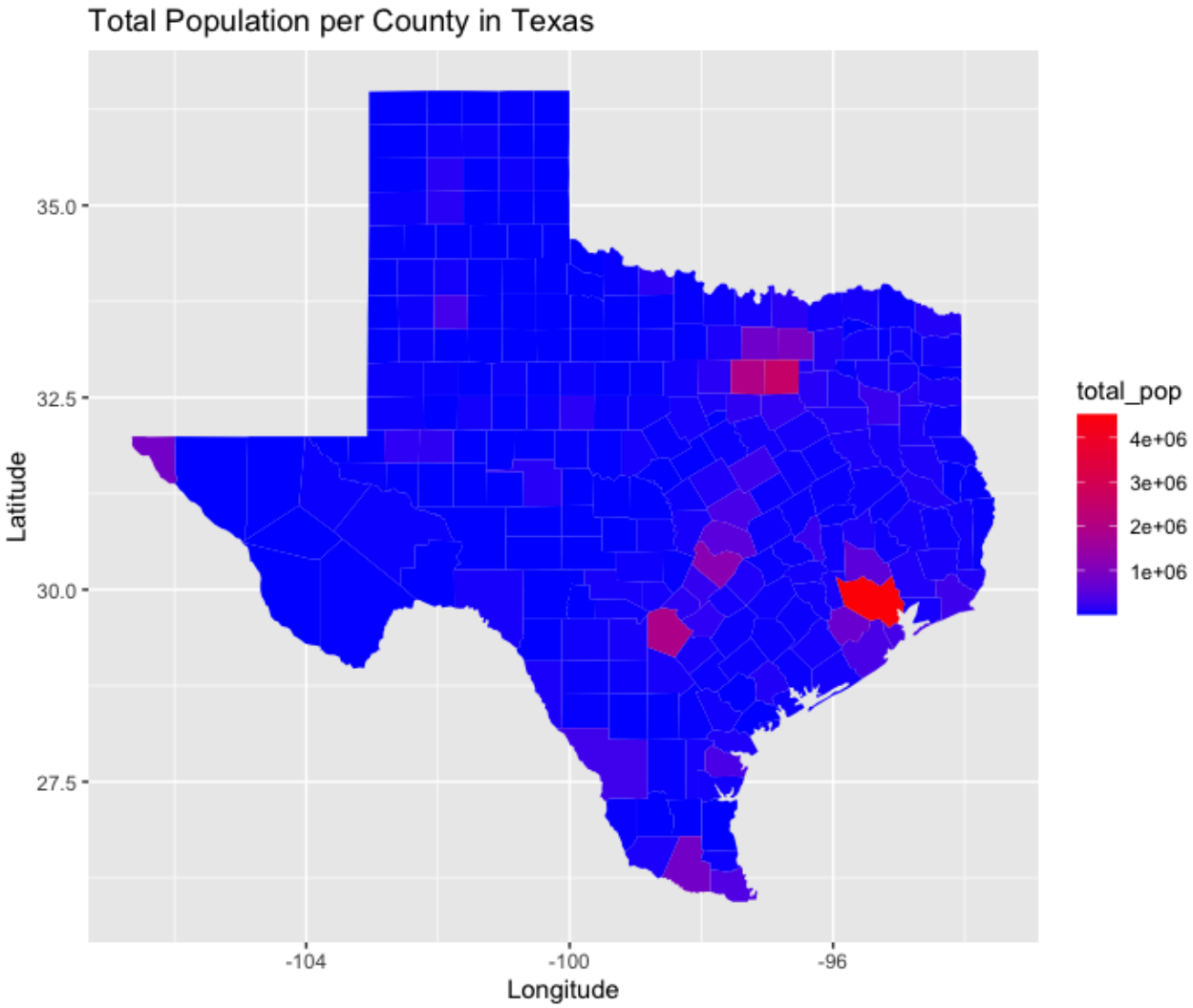
Table 5: New Dataset Compiled from Data Understanding

Feature	DataType	Description
county_fips_code	Nominal	Identification number that uniquely identifies geographical area for county
county_name	Nominal	Name of county in US
total_pop	Ratio	Total population of county
date	Interval	Analysis horizon
confirmed_cases	Ratio	Total number of confirmed COVID-19 cases in county at the given date
deaths	Ratio	Total number of confirmed COVID-19 deaths in county at the given date
cases_per_1000	Ratio	1,000 * (total confirmed cases / total population)
deaths_per_1000	Ratio	1,000 * (total deaths / total population)
spread_rate	Ratio	(Cases as of last recorded date - cases as of March 5, 2020) / total time period
fatality_rate	Ratio	Total deaths / total cases
first_case_date	Interval	Date when the first case of COVID-19 occurred

What is The Total Population in Each County?

Below in Figure 14, you can see a texas map of the population densities. As you can see, we have high populations in the major cities within texas and low populations in rural areas. The majority of counties in Texas can be considered rural due to the shading on the map. This map looks very similar but is different from the map of the spread rate for each county seen below in this section. This makes sense since it is likely that **there would be a higher spread rate in the large cities due to high populations.**

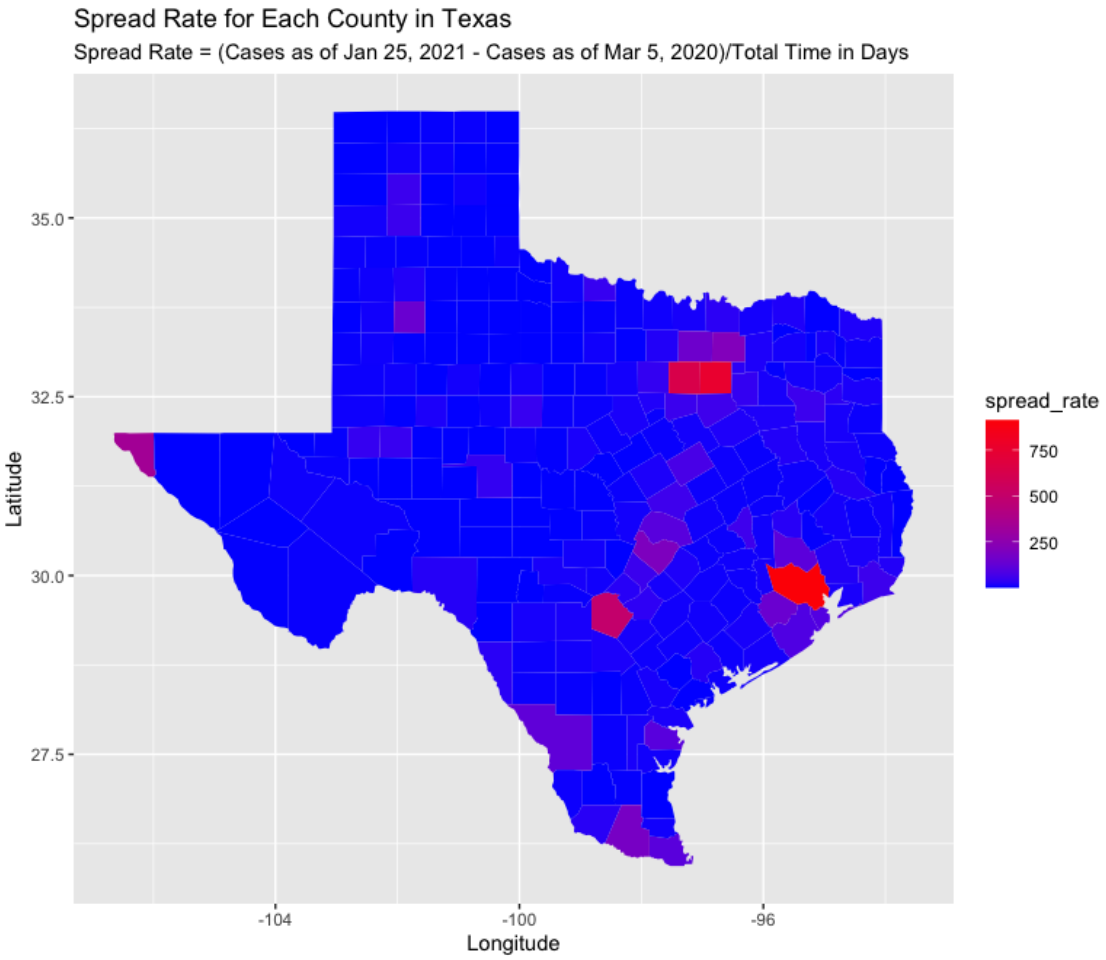
Figure 14: Total Population per County in Texas



What Does the Spread Rate Look Like per County?

Below in Figure 15, we can see the spread rate for each county depicted through a visual map of Texas. As we can see when referencing a map of cities in Texas, the spread rate was relatively low in rural areas but very high in high population cities. **This is expected as high population cities are more likely to have higher spread due to being more densely populated while rural areas are likely to have lower spread due to the natural distance between individuals.**

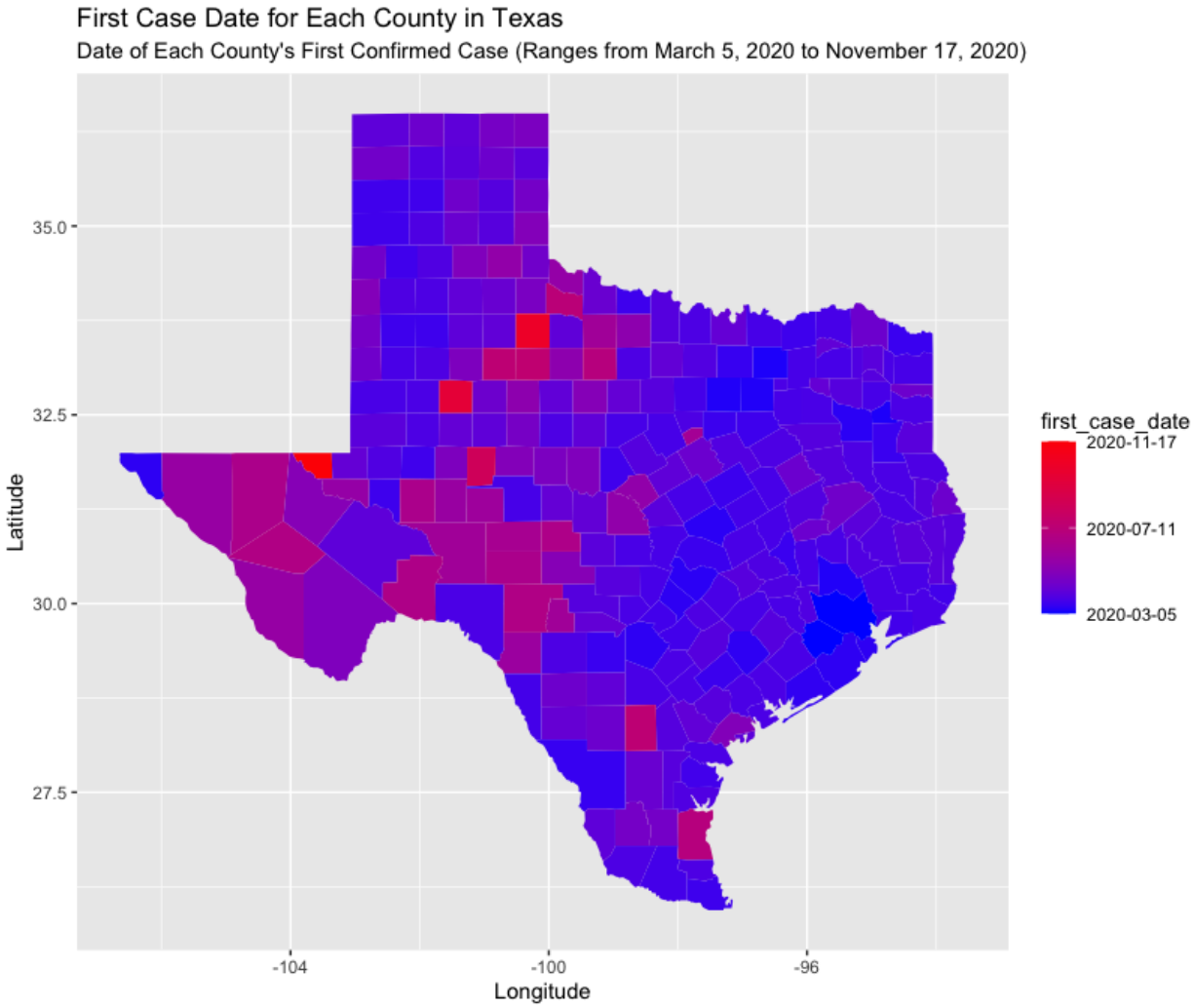
Figure 15: Spread Rate for Each County in Texas



When Was The First Reported Case in Each County?

Below in Figure 16, we can see the date of each county's first confirmed case depicted through a visual map of Texas. **From the map we can see how smaller population counties and more rural counties experience their first confirmed case very late, with some counties not experiencing their first case until November**. We can also see how more populated and urban cities like Houston were the first to experience cases. **Loving county, the bright red county on the map not far from El Paso, saw their first case on November 17, 2020 while Harris county, the dark blue county on the map near the coast, saw their first case on March 5, 2020. This is expected because more people travel and live in high population and large cities while fewer people travel to small rural cities.**

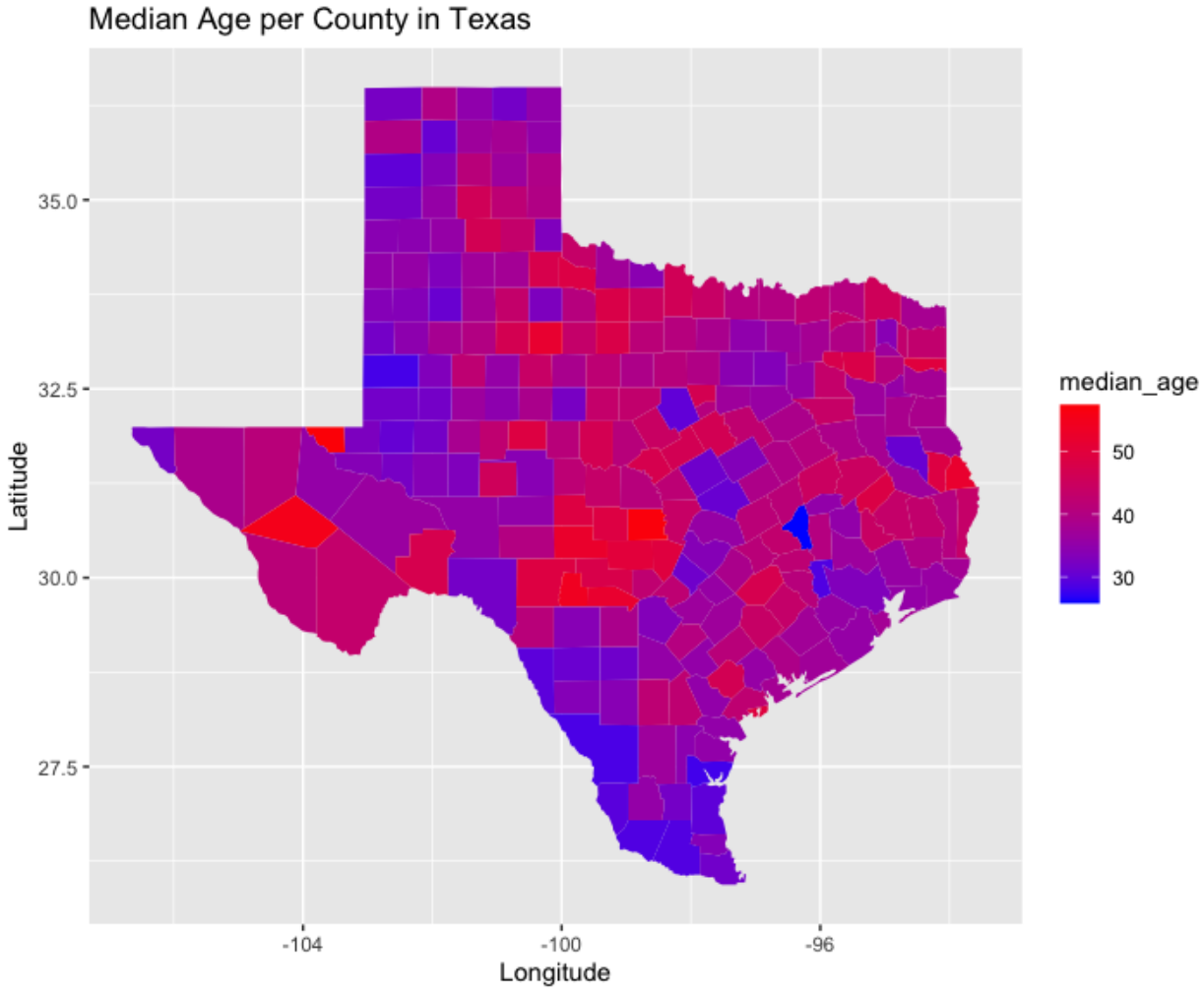
Figure 16: First Reported Case in each County



What is The Median Age in Each County?

As you can see below in Figure 17, we can see the median age per county depicted through a visual map of Texas. What is interesting about this map is that the majority of older people seem to live in the more rural areas. **When cross referenced with the fatality map seen in Figure 18, some of the areas with the highest fatality also have a higher median age. This could indicate that the older a person is, the higher they are at risk from contracting COVID-19 fatally.** This is an inference that we predicted could be present as stated prior in the report.

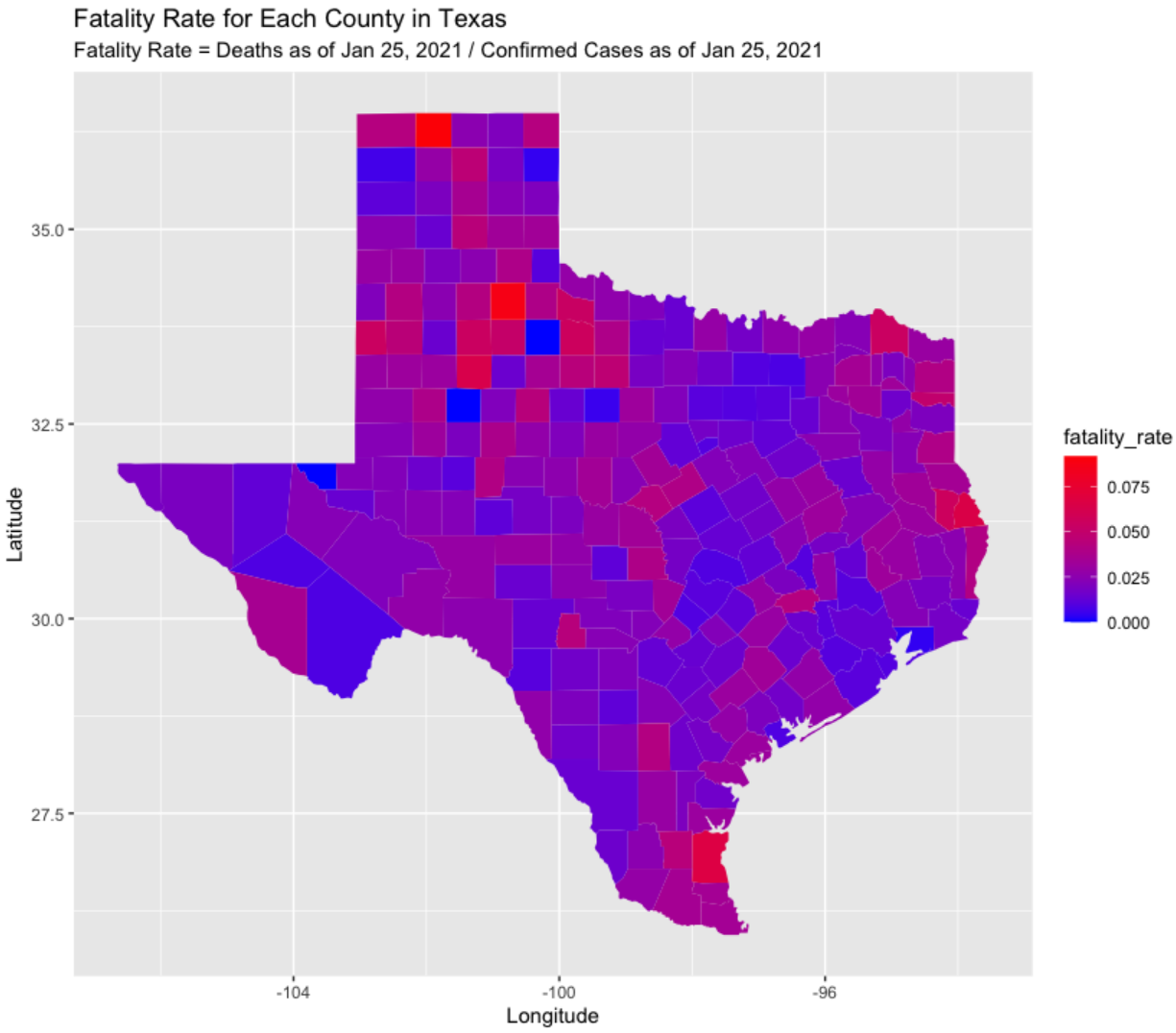
Figure 17: Median Age per County in Texas



What is The Case Fatality Rate in Each County?

Below in Figure 18, we can see the case fatality rate for each county depicted through a visual map of Texas. **The case fatality rate is calculated by simply dividing the number of deaths by the number of cases for each county**. What is interesting to see here is that the case fatality rate was actually higher in more rural counties and not as high near large cities. **This could be attributed to the fact that better health care is available in larger cities and therefore serious cases could be more effectively treated. It could also suggest that individuals in rural areas could have pre existing underlying health conditions due to the lack of healthcare, ultimately making them more at risk of dying from COVID-19.**

Figure 18: Case Fatality Rate in each County



V. Conclusion

County based census data is a good potential indicator for the spread of COVID-19 within the respective county. In this report, an in-depth analysis on various features of each county located within the state of Texas. The following conclusions can be made from the numerous graphs throughout this report:

- Due to a lower correlation between deaths/cases and the Asian population compared to other racial groups, Asians might be doing something different than other populations in protecting themselves from contracting the virus related to their culture. In Asian

countries it is common to wear a mask for any illness one might have so this may have to do with the spread/fatalities.

- When the correlation between median age and deaths per 1000 people in the population was calculated, the value was very close to zero, indicating there is no relationship between the two. This suggests that old age is not the biggest risk for dying from COVID-19 and instead suggests other underlying or pre-existing conditions puts an individual more at risk.
- There is a very strong positive relationship between spread rate and total population.
- The vast majority of counties with a lower median income experience higher death rates per 1000 which could indicate less access to advanced health care.
- Smaller population counties and more rural counties experience their first confirmed case very late because there is significantly less travel into and out of the area.
- High population cities are more likely to have higher spread due to being more densely populated while rural areas are likely to have lower spread due to the natural distance between individuals.
- Case fatality rates were higher in more rural counties than in large cities likely due to the fact that better health care is available in larger cities and therefore serious cases could be more effectively treated.
- Rural areas could have pre existing underlying health conditions due to the lack of healthcare, ultimately making them more at risk of dying from COVID-19.

Analysis over other states in the US would definitely be interesting and would provide better insights as to how Texas fared during this pandemic compared to other states. That being said, in general, Texas is a very rural state with few large cities so comparing it with states like California or New York could reveal fascinating information.

VI. References

1. https://www.cdc.gov/coronavirus/2019-ncov/faq.html#Basics
2. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html
3. https://www.webmd.com/lung/qa/what-is-flattening-the-curve
4. https://www.dshs.state.tx.us/coronavirus/execorders.aspx
5. https://www.texastribune.org/2020/07/31/coronavirus-timeline-texas/