

## **ABSTRACT**

The Covid-19 pandemic has affected people's lives in many ways all over the world since 2020. Data sources, policy options, decision making tools and other related recourses are increasing rapidly. Analyzing these data helps people to better understand the spread of the virus and make further prediction on when the pandemic will be over. For this project, we will look at different data sets collected by Google about the spread of the Covid-19 virus. The focus of this project is on the cleaning and understanding of the data. Several features in different data sets are selected and then individually analyzed by data visualization. Furthermore, the relationship of various features in different data sets are also explored to get a deeper understating of data. The analysis shows the trend of total confirmed cases and daily new cases in the US, especially in Texas. It also reveals the impact of social distancing in business and transportation. Lastly, a data set is constructed with important features such as spread rate, death rate, social distancing response, and hospital capacity to analyze the impact of Covid-19 pandemic in Texas counties.

# Content

<b>1</b>	<b>BUSINESS UNDERSTANDING.....</b>	<b>1</b>
<b>2</b>	<b>DATA UNDERSTANDING .....</b>	<b>5</b>
2.1	DATA DESCRIPTION .....	5
2.2	VERIFYING DATA QUALITY .....	7
2.2.1	<i>US covid-19 cases plus census data .....</i>	<i>7</i>
2.2.2	<i>Texas covid-19 cases data .....</i>	<i>7</i>
2.2.3	<i>Global mobility data.....</i>	<i>8</i>
2.3	STATISTICAL SUMMARY .....	8
2.3.1	<i>Statistics on US covid-19 cases plus census data.....</i>	<i>9</i>
2.3.2	<i>Statistics on TX covid-19 cases data .....</i>	<i>10</i>
2.3.3	<i>Statistics on Global mobility data .....</i>	<i>10</i>
2.4	VISUALIZATION OF SELECTED DATA .....	11
2.4.1	<i>US covid-19 cases plus census data .....</i>	<i>11</i>
2.4.2	<i>Texas covid-19 cases data .....</i>	<i>13</i>
2.4.3	<i>Global mobility data (How the Covid-19 pandemic affects public on Dallas country).....</i>	<i>16</i>
2.5	EXPLORING RELATIONSHIPS BETWEEN ATTRIBUTES.....	20
<b>3</b>	<b>DATA PREPARATION .....</b>	<b>24</b>
3.1	DATA SET WITH IMPORTANT FEATURES .....	24
3.2	VISUALIZATION OF IMPORTANT FEATURES.....	26
3.2.1	<i>Covid-19 spread rate in each county in Texas .....</i>	<i>26</i>
3.2.2	<i>Hospital capacity in each county in Texas .....</i>	<i>27</i>
3.2.3	<i>Medical staff in each county in Texas .....</i>	<i>30</i>
3.2.4	<i>Correlation between medical staff, hospital capacity, death rate, and population in Texas.....</i>	<i>33</i>
3.2.5	<i>Relationship between the number of deaths and confirmed cases.....</i>	<i>34</i>
3.2.6	<i>Social distancing response for each county in Texas .....</i>	<i>36</i>

3.2.7	<i>Relationship Correlation among deaths, spread rate, population, confirmed cases, median age and social distancing response .....</i>	<i>38</i>
<b>4</b>	<b>CONCLUSION.....</b>	<b>40</b>
	<b>REFERENCES .....</b>	<b>41</b>

# 1 Business Understanding

Covid-19 is a new emerging virus disease stems from corona virus family that has not been previously identified in humans. The first case of Covid-19 was reported in Wuhan china in 2019 [1] and it rapidly spread in countries all over the world. The Covid-19 pandemic affects the business owners, schools, and endangers people lives around the world. Different resources show the growing speed of this pandemic. For example, the daily total overall number of tests, total positives, total lives lost can be found in the [2].

The main problem in the Covid19 outbreak is how big and how fast it will get. What make fear most is sudden explosion of illness and healthcare system becomes overwhelmed because of the need for hospitalization. This problem can be somewhat handled by protective measures like avoiding crowds, cancelling mass gathering, working from home, wearing mask. Even if it doesn't reduce the number of total cases, it slows down the rate of its spread. [3]

Medical scientists believe that people should keep a minimum distance of 6 feet from each other to reduce the spread of the virus [4]. This is known as a social distancing or physical distancing. This trend is known as flatten the curve and is shown in the Figure 1 provided in [5] where less cases are affected over longer period rather than a spike of cases in which patients might not be hospitalized and treated due to the shortage of medical stuff and hospital capacity.

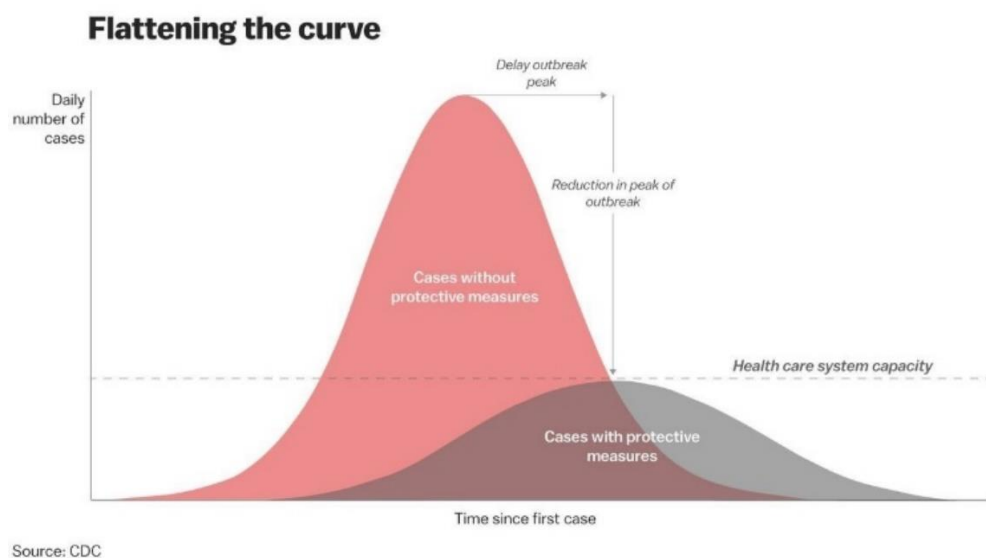


Fig. 1: Flattening the curve

Covid19 need data collection for states, providers and healthcare decision makers. Reporting daily Covid19 cases and current hospital resources will assist local governments and medical personnel to project the essential medical supplies such as ventilators, hospital beds, and testing equipment etc. For instance, the available and occupied Intensive Care Unit (ICU) beds by Covid patients and non-Covid patients in the Texas hospital are reported in [5]. They keep track of this data so they can transform the surgical beds to ICU beds and postpone some procedures [5]. These data are not only useful to prepare the medical needs, but also important to take proper responses to slow down the spreading of Covid-19. These responses might be by enforcing a lockdown on the most affected areas or adding restrictions on the places where a high number of confirmed cases reported. They update and analyze this data in a daily basis. Tracking the spread of Covid cases will also be a good measure to answer the questions on how people follow the medical guidelines, why the number of positive Covid cases increases in certain area, and where the large numbers of Covid cases are. In short, analyzing Covid-19 data can assist local government and medical experts to enforce or lift restrictions on certain areas, anticipate the medical needs, and response to fight Covid-19 pandemic.

Data sources, policy options, decision making tools and other related recourses are increasing rapidly. In the table1 some available resources reporting the cases are listed [6]. Based on the released data, lots of the studies and reports are published. For example, in [7], the Covid19 published researches are reported. These topics included in the database and the data visualization include, but are not limited to, budgeting and revenue, child welfare, commerce, criminal justice, education, elections, employment, finance, health access and coverage, housing and homelessness, labor and retirement, legislative operations, public health, workforce and more [8].

Since the beginning of this pandemic in US, most states have imposed lockdown measures restricting gathering and social contact which reduced the speed of its spread but it has disrupted the lives of most of people and the operations of businesses, so some states, have plans to relax restrictions [9]. These policies and restrictions have caused different trend in the number of cases in states. There are some available reports and studies showing trends for states. For example, in [10] the trend for the number of cases and number of deaths are reported and

the trend for number of deaths are shown in Figure 2. Another resource is [11], where the trend in the number of cases is reported and in Figure 3 an table showing the trends is presented. Based on these reports, one can see how these policies have been effective in different states.

The researchers have analyzed the effect of social distancing across all the 50 states and they have seen that the spread of the virus is significantly decreased by social distancing when they compare different states during the same time periods and two of the states without such policies had the least changes in transmission of all states [12]. They have found that there is a strong correlation between lower community mobility and reduced spread of the Covid19, suggesting that it is crucial to practice social distancing to see transmission rates drop [12].

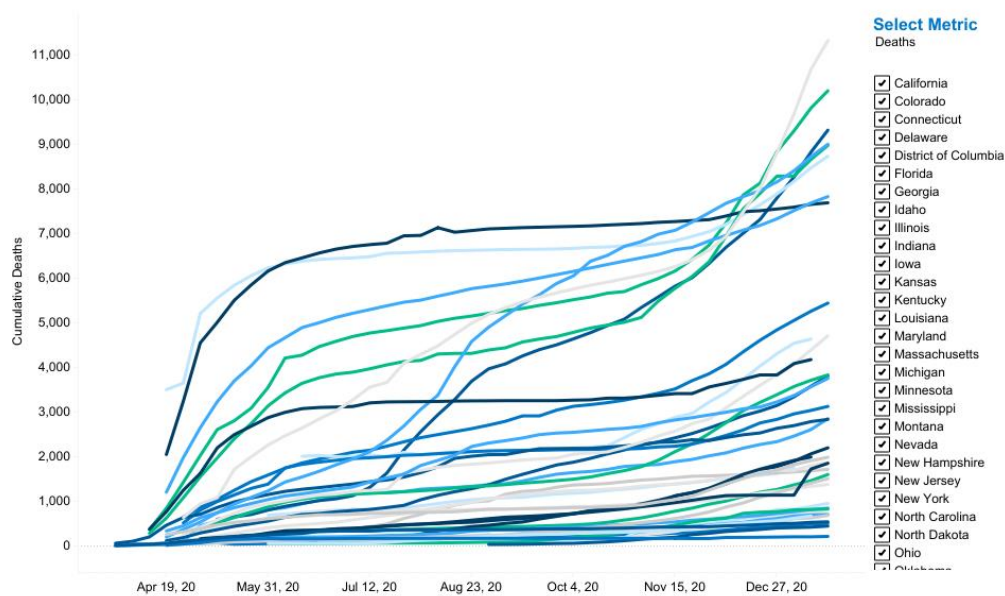


Fig. 2: Cumulative deaths in different states


	STATE	14-DAY TREND OF NEW CASES	14-DAY NEW CASES	14-DAY NEW CASES PER 100,000	CONFIRMED CASES	CASES PER 100,000	CONFIRMED DEATHS	% POSITIVE TESTS	TOTAL TESTS	CURRENTLY HOSPITALIZED	TOTAL HOSPITALIZED	TOTAL VS NEW CASES	
✓	Texas		-38.2%	96,339	336	2.61M	9,107	41,980	100%	2.61M	6,738	N/A	
✓	California		-42.2%	92,380	234	3.46M	8,735	49,877	100%	3.46M	6,764	N/A	
✓	Florida		-14.6%	85,492	401	1.85M	8,691	30,878	16.9%	10.9M	4,077	79,675	
✓	North Carolina		-23.5%	43,732	421	849,630	8,182	11,074	8.38%	10.1M	1,530	N/A	
✓	Georgia		-16%	43,155	410	994,061	9,450	17,064	100%	994k	2,615	55,394	
✓	New Jersey		-13.4%	42,712	479	775,386	8,704	23,077	7.39%	10.5M	2,070	63,252	
✓	Pennsylvania		-23.4%	40,343	315	920,634	7,188	23,787	19.3%	4.76M	1,972	1,145	
✓	South Carolina		-21.2%	35,904	706	509,044	10,012	8,398	10.4%	4.91M	968	20,117	

Fig. 3: Trends in the number of cases for different states

Table 1: some of Covid-19 report sources (**Exceptional work included**)

Data sources	Provided Data
<a href="#">Tracking COVID at U.S. Colleges and Universities</a>	COVID-19 case counts at U.S. colleges and universities since the start of the pandemic
<a href="#">COVID-19 Nursing Home Public File</a>	Nursing home-level information on confirmed and suspected COVID-19 cases and deaths among nursing home residents and staff, as well as information on facility capacity, PPE supply, and ventilator capacity.
<a href="#">Coronavirus Locations: COVID-19 Map by County and State</a>	COVID-19 cases and deaths by county and state
<a href="#">CV19 Lab Testing Dashboard</a>	State- and county-level COVID-19 viral and antibody testing rates, positive and negative test results, and testing results disaggregated by age and gender
<a href="#">USA Coronavirus (COVID-19) Case Tracker</a>	COVID-19 case and test counts, deaths, and state-level reopening status
<a href="#">Coronavirus in the U.S.: Latest Map and Case Count</a>	Various metrics related to case counts
<a href="#">Global COVID-19 Outlook</a>	Case counts of COVID-19 around the world, by country and region. Information about world-wide travel restrictions
<a href="#">COVID Tracking Project</a>	State-level case information for states across the United States
<a href="#">Global Pandemic Real-Time Report</a>	Counts of confirmed cases, daily new cases, and deaths, by continent and country
<a href="#">Covid-19</a>	Case counts by region across the world
<a href="#">JHU/ESRI Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering</a>	Confirmed cases, deaths, and recoveries from COVID-19 globally
<a href="#">Coronavirus Disease 2019 (COVID-19) Cases in U.S.</a>	The number of COVID-19 cases in the U.S., overall and by state
<a href="#">COVID-19 in the US and Canada</a>	COVID-19 cases and testing at the county level in the United States and province level in Canada

## 2 Data Understanding

### 2.1 Data Description

The data file contains three different data sets including US covid-19 cases plus census data, Texas covid-19 cases and Google global mobility report. US covid-19 cases plus census data provides total US COVID-19 case and death counts by state and county until 2021-03-03. This data is sourced from the CDC, and state and local health agencies. Additionally, census data provides useful information about the population. The data has 3,142 observations and 259 columns. A description of some important features in US covid-19 cases plus census data is given in Table 2.

Texas covid-19 cases data set provides COVID-19 case and death counts by county in Texas from 2020-01-22 to 2021-03-06. It also provides changes and trends by county. The data has 104,550 observations and 7 columns. A description of all features in Texas covid-19 cases data is given in Table 3.

Google global mobility report aims to provide insights into what has changed in response to policies aimed at combating COVID-19. The reports chart movement trends over time by geography, across different categories of places such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. For each category in a region, reports show the changes in one way: Headline number. It compares mobility for the report date to the baseline day. Calculated for the report date and reported as a positive or negative percentage. It contains different mobility metrics for countries and their states from 2020-02-15 to 2021-01-22. A description of all features in global mobility report is given in Table 4.

Table 2: The description of some important features in US covid-19 cases plus census data

Feature	Data Type	Description
County Name	Nominal	Name of each county
State	Nominal	Name of the state
Confirmed Cases	Ratio	Number of confirmed cases
Deaths	Ratio	Number of deaths
Median Age	Ratio	Median age



Total Population	Ratio	Total population
Population over 16	Ratio	Population over 16 years old
Female Population	Ratio	Female population
Male Population	Ratio	Male population
Family Households	Ratio	Number of households contain families
Median Income	Ratio	Median Income
Employed Population	Ratio	Population that are classified as employed
Unemployed Population	Ratio	Population that are classified as unemployed
Master's Degree	Ratio	Number of people with master's degree
Poverty	Ratio	Population that are classified as poverty
Gini Index	Ratio	A summary measure of income inequality, the coefficient ranges from 0 to 1

Table 3: The description of all features in Texas covid-19 cases data

Feature	Data Type	Description
County FIPS Code	Nominal	Numbers which uniquely identify counties in United States
County Name	Nominal	Name of each county
State	Nominal	Name of the state
State FIPS Code	Nominal	Numbers which uniquely identify states in United States
Date	Nominal	Date from 2020-01-22 to 2021-03-06
Confirmed Cases	Ratio	Number of confirmed cases on a given day
Deaths	Ratio	Number of deaths on a given day

Table 4: The description of all features in Global mobility data

Feature	Data Type	Description
Country Region Code	Nominal	Two letters code for each country
Country Region	Nominal	Name of each country
Sub-Region 1	Nominal	Name of each state in each country
Sub-Region 2	Nominal	Name of each county in each state
Date	Nominal	Date from 02-15-2020 to 01-22-2021
Percent Changes in Retail and Recreation	Ratio	Retail and recreational percent changes from the baseline
Percent Changes in Grocery and Pharmacy	Ratio	Grocery and pharmacy percent changes from the baseline

Percent Changes in Parks	Ratio	Parks percent changes from the baseline
Percent Changes in Transit Station	Ratio	Transit station percent changes from the baseline
Percent Changes in Workplace	Ratio	Workplace percent changes from the baseline
Percent Changes in Residential	Ratio	Residential percent changes from the baseline

## 2.2 Verifying Data Quality

It is essential to verify data quality before attempting to conduct the analyses. The quality of the data can be explored in following aspects: Are there missing values? Are there duplicate data? Are there outliers and are those mistakes? The quality of each data set is discussed following.

### 2.2.1 US covid-19 cases plus census data

Most of the variables had no missing data. However, some variables (*'median rent'*, *'percent income spent on rent'*, *'owner occupied housing units median value'* etc.) had a few missing data. Checking the data showed that those data were missing completely at random. Thus, there is no problem with those missing values. There were some variables (*'pop 5years over'*, *'speak only English at home'*, *'speak Spanish at home'*, *'pop separated'*, *'pop widowed'*, *'pop divorced'* etc.) that had no data values. All the data were missing. Thus, we can simply delete those variables. Duplication of data will not be a problem in this data set. In addition, there were no extreme outliers.

### 2.2.2 Texas covid-19 cases data

There were no missing values in this data set. Duplication of data will not be a problem in this data set. In addition, there were no extreme outliers.

### 2.2.3 Global mobility data

Table 5 shows the missing values in global mobility data. The reason is that some countries did not identify the states and the country names. Also, some countries did not report the percent change in some variables. However, these missing values are not an issue in this report since this report is focus on the mobility data in Dallas country in Texas which there are no missing values in Dallas mobility data.

Table 5: The missing values in Global Mobility data

Feature	Number of Missing Values	Description
Sub-Region 1	68,243	These missing values show that name of states in some country does not exist
Sub-Region 2	666,546	These missing values show that name of counties in some country does not exist
Percent Changes in Retail and Recreation	1,478,424	Some countries did not report retail and recreational Percent Changes from the baseline
Percent Changes in Grocery and Pharmacy	1,564,666	Some countries did not report grocery and pharmacy percent changes from the baseline
Percent Changes in Parks	2,080,860	Some countries did not report Parks Percent Changes from the baseline
Percent Changes in Transit Station	1,973,496	Some countries did not report transit station percent changes from the baseline
Percent Changes in Workplace	189,760	Some countries did not report workplace percent changes from the baseline
Percent Changes in Residential	1,678,955	Some countries did not report residential percent changes from the baseline

## 2.3 Statistical Summary

The statistical summary of some features listed in Table 2 are given in Table 6. For nominal variables such as County Name and State, there is not much of statistics to be addressed. Thus, only mode and frequency are provided for these nominal variables. For Texas cases data, only mode, frequency, minimum and maximum for variables confirmed cases and deaths are

included in the statistical summary Table 7. For global mobility date, minimum, maximum, median and mean are included for all the numeric variables in the statistical summary Table 8.

### 2.3.1 Statistics on US covid-19 cases plus census data

Some of the ratio variables in US data set like confirmed cases, median age, total population, population over 16, female population, employed population etc., mode and frequency are not addressed in the table because there is no data value that occur most frequently. Notice that variables median age, median income and Gini index have close mean and median, this may indicate that the distributions of those variables are normal. Further analysis will be provided in the part of Visualization of Selected Data.

Table 6: The statistical summary of some important features in US covid-19 cases plus census data

Feature	Mode	Frequency	Mean	Median	St.Dev	Min	Max
County Name	Washington County	30	-----	-----	-----	-----	-----
State	TX	254	-----	-----	-----	-----	-----
Confirmed Cases	-----	-----	8,961	2,269	32,978	0	1,156,826
Deaths	0	82	162	42	630	0	21,554
Median Age	-----	-----	41	41	5.38	22	66
Total Population	-----	-----	102,166	25,692	328,292	74	10,105,722
Population over 16	-----	-----	81,412	20,735	261,086	67	8,102,402
Female Population	-----	-----	51,873	12,885	167,145	35	5,126,081
Male Population	10028	4	50,292	12,798	161,185	39	4,979,641
Family Households	-----	-----	24,920	6,604	75,140	15	2,203,922
Median Income	48412	4	49,754	48,066	13,154	19264	129,588
Employed Population	-----	-----	47,931	10,695	157,622	39	4,805,817
Unemployed Population	0	11	3,361	746	12,406	0	406,426
Master's Degree	-----	-----	5,778	790	20,906	0	492,924

Poverty	-----	-----	14,529	4,120	51,702	10	1,688,505
Gini Index	0.417	11	0.445	0.442	0.036	0.327	0.598

### 2.3.2 Statistics on TX covid-19 cases data

Harris county has the maximum total number of confirmed cases, 357558, and deaths, 5296, in Texas on 03/06/2021. The mode for confirmed and death cases is zero since there was no cases in Texas from 1/22/2020 to 3/4/2020.

Table 7 The statistical summary of features in TX covid-19 cases data

Feature	Mode	Frequency	Min	Max
Confirmed Cases	0	21,051	0	357,558
Deaths	0	39,107	0	5,296

### 2.3.3 Statistics on Global mobility data

The minimum of percent change from the baseline for retail and recreation, grocery and pharmacy, transit station, parks, and workplace are -100 and for the residential percent change from the baseline is -46. The mean for the percent changes in these variables is between -3 and 9. The negative mean indicates that the most mobility variable impacted by the Covid-19 pandemic and go down below the zero most of the time. It shows that the most impacted variable that fall down below zero is transit station with -27.2% changes from the baseline followed by retail and recreation with -23.2% and workplace with -20.07. The positive mean in the residential percent changes (9%) indicates that the residential go up above the baseline most of the time. The maximum values in these variables range from 65 to 1206 which might indicate some outliers in these data.

Table 8: The statistical summary of features in global mobility data

Feature	Min	Median	Mean	Max
Percent Changes in Retail and Recreation	-100.0	-19.0	-23.2	545.0
Percent Changes in Grocery and Pharmacy	-100	-2	-3	615
Percent Changes in Parks	-100	-17	-9.5	1,206
Percent Changes in Transit Station	-100	-28	-27.2	554
Percent Changes in Workplace	-100	-19	-20.07	260
Percent Changes in Residential	-46	8	9	65

## 2.4 Visualization of Selected Data

### 2.4.1 US covid-19 cases plus census data

#### ➤ *Confirmed Cases and Deaths*

The total number of cases and number of deaths for each state can be calculated. The death percentage as death-per-cases, the number of cases in 1000 as the cases\_per\_1000 and the number of deaths in 1000 as death-per-1000 is calculated which is shown in Table 9. Table 9 shows the top 10 states with highest COVID-19 cases in United States. From the table above, we can tell that as of March 3, 2021, the state with the highest number of COVID-19 cases was California. Almost 11 million cases have been reported across the United States, with the states of California, Texas, Florida, New York and Illinois reporting the highest numbers. Besides that, New Jersey and New York have relatively high death percentage rate among US, and AZ has the highest confirmed cases rate based on its total population.

Table 9: Top 10 states with highest COVID-19 cases in United States as of March 3, 2021

	state	confirmed	population	deaths	cases_per_1000	deaths_per_1000	death_per_case
1	AZ	819952	6809946	16091	120.4051	2.3629	1.96%
2	CA	3484934	38982847	52775	89.3966	1.3538	1.51%
3	FL	1920622	20278447	31273	94.7125	1.5422	1.63%
4	IL	1191437	12854526	20629	92.6862	1.6048	1.73%
5	NC	865554	10052564	11366	86.1028	1.1307	1.31%
6	NJ	798353	8960161	23449	89.1003	2.6170	2.94%
7	NY	1650184	19798228	47242	83.3501	2.3862	2.86%
8	OH	968874	11609756	17351	83.4534	1.4945	1.79%
9	PA	938419	12790505	24171	73.3684	1.8898	2.58%
10	TX	2664897	27419612	43544	97.1894	1.5881	1.63%

➤ *Median age, median income and Gini index*

As mentioned before, the variables median age, median income and Gini index have close mean and median, this may indicate that the distributions of those variables are normal. The distribution plot for those variables is shown in Figure 4. From the plot we can tell that the distribution of median age and Gini index resembles a bell. The bell curve distribution is also called normal distribution. For example, the mean of median age is 41 and it is almost at the center of the plot. Half of the data falls to the left of the mean; half falls to the right, and majority of the data falls with 30 and 50. The distribution of median income is slightly right skewed.

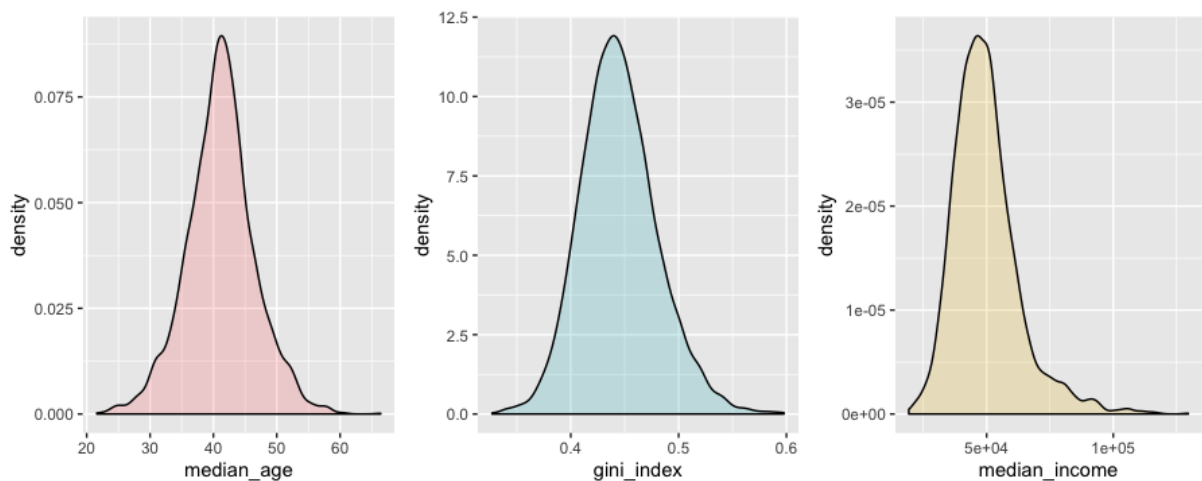


Fig. 4: Distribution of median age, Gini index and median income

### ➤ Population

The following two plots show the population of United States and Texas by race and ethnicity as of Mar 3, 2021, respectively.

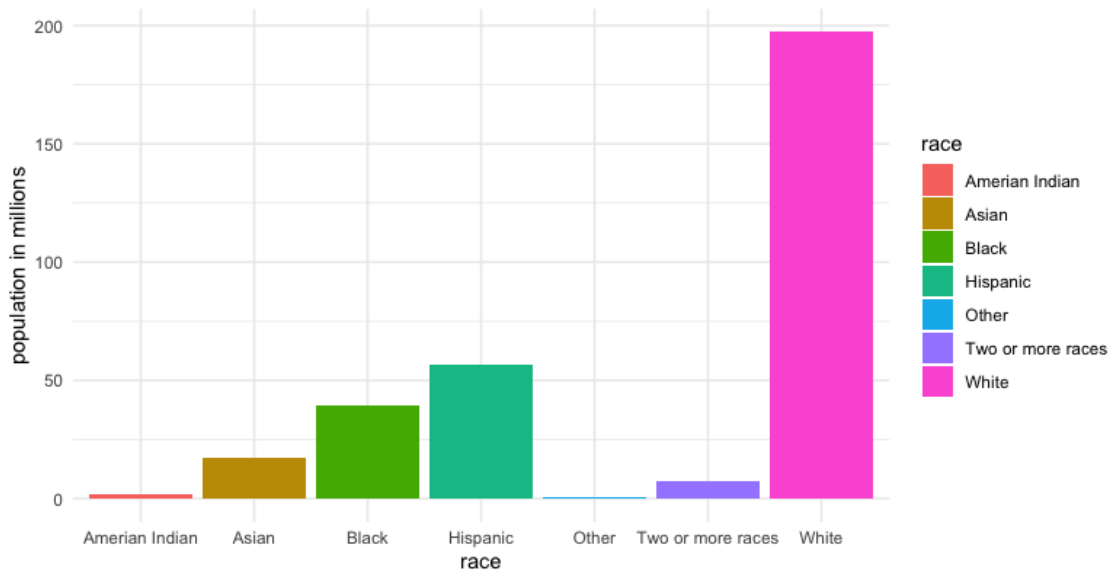


Fig. 5: US total population by race and ethnicity

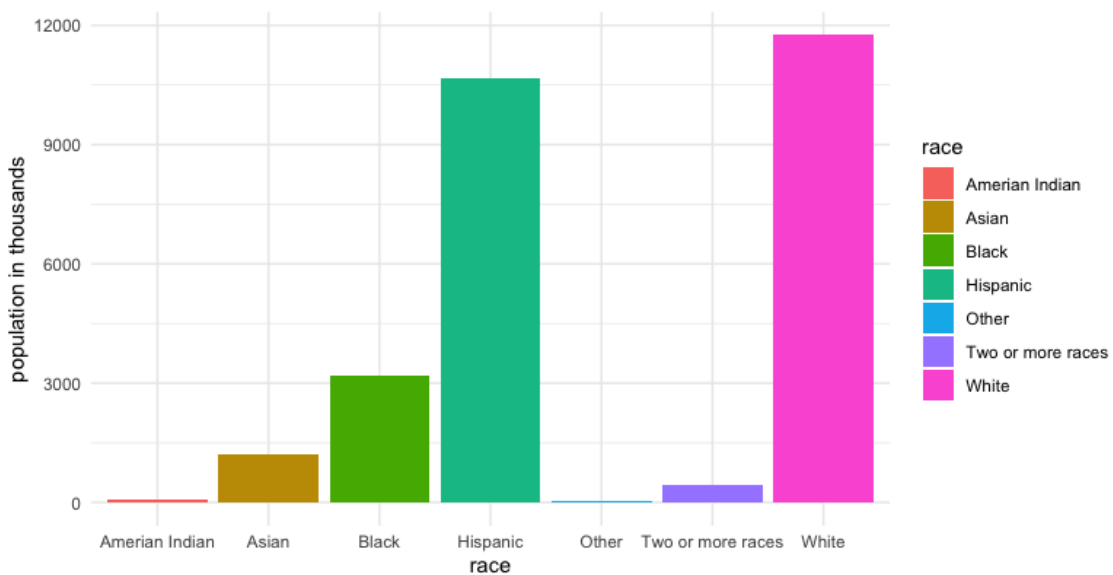


Fig. 6: Texas total population by race and ethnicity

#### 2.4.2 Texas covid-19 cases data

Next, we focus on the cases data in Texas particularly. The top 10 states with highest COVID-19 cases in Texas is shown in Figure 7. As of March 6, 2021, Harris county has the



highest number of confirmed cases, which is almost 360,000. The second-most populous county, Dallas County, has about 280,000 confirmed cases. Furthermore, visualizing the trend of some specific variables is an effective way for better understanding the data with time frame. In this data set, confirmed cases is the important feature that we need to explore. Total confirmed cases and daily new confirmed cases in Texas by date from 2020-01-22 to 2021-03-06 are shown in Figure 8 and 9, respectively. We can tell that before March 2020, there were no cases in Texas. Total confirmed cases dramatically increased after July 2020. In the Figure 9, there is a significant increase in daily new cases, about 180,000 from December 20<sup>th</sup> to December 21<sup>st</sup>. In addition, we can tell that the trend of daily new cases has been decreasing from January 2021 to March 2021. Furthermore, we are interested in the trend of total confirmed cases in DFW area, which is shown in Figure 10. We can tell that confirmed cases in Dallas County and Tarrant County after July 2020 increased much more than Collin County and Denton County.

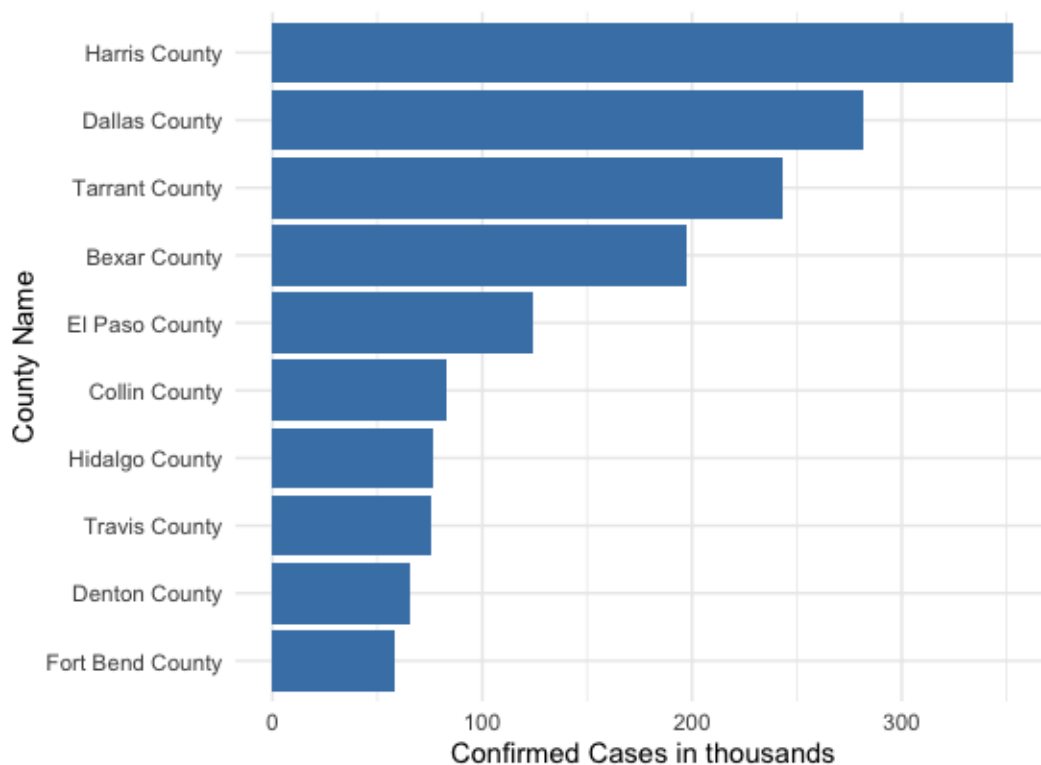


Fig. 7: Top 10 states with highest COVID-19 cases in Texas

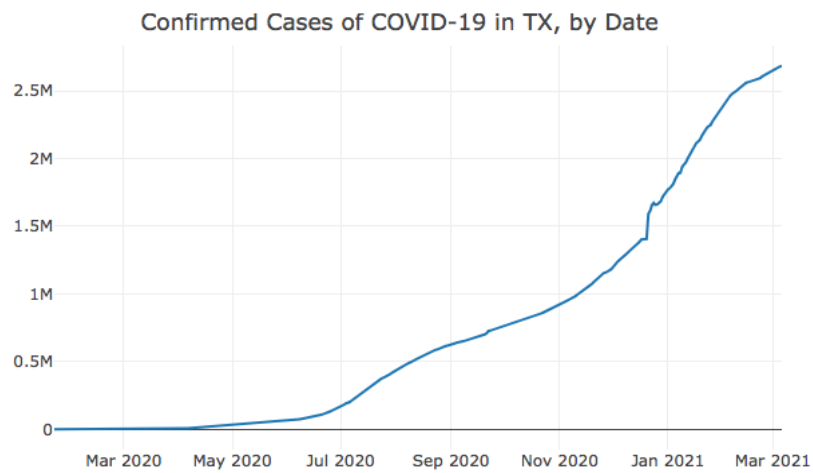


Fig. 8: Total confirmed cases in Texas by date

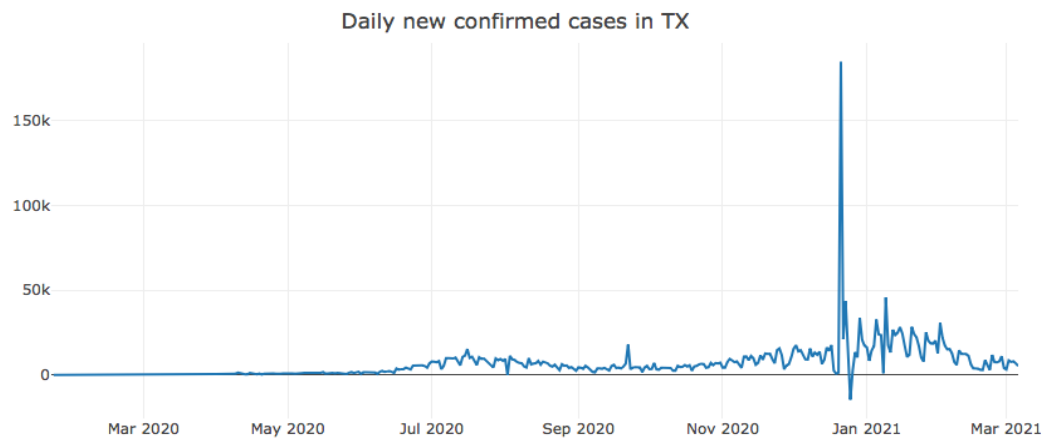


Fig. 9: Daily new confirmed cases in Texas by date

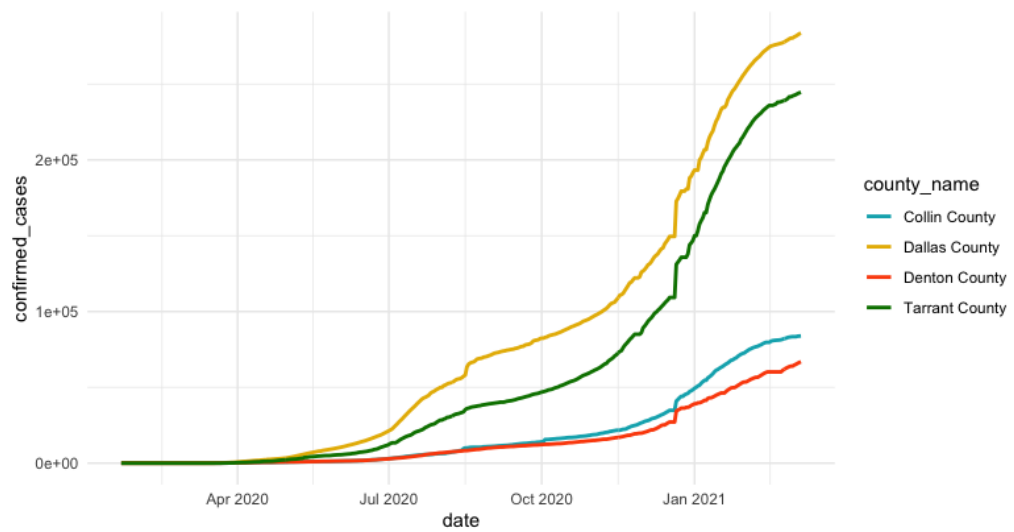


Fig. 10: Total confirmed cases in Collin, Dallas, Denton and Tarrant County by date

### 2.4.3 Global mobility data (How the Covid-19 pandemic affects public on Dallas country)

Dallas county issued a shelter-in-place order as a response to the rise in number of confirmed cases in the city. This order mandates all non-essential business to temporarily closed including bars, recreation centers, theaters, museums etc. Essential business such as grocery stores, Banks, pharmacy and other essential business is allowed to operate. It was effective from March 23<sup>rd</sup> to April 3<sup>rd</sup> and later the order was extended to May 15<sup>th</sup> [13]. In this section, the impacts of Covid-19 pandemic and Dallas county restrictions to slow down the spread of Covid-19 are analyzed.

#### ➤ Impact on recreation and retail mobility:

The percent change in retail and recreation business from the baseline in Dallas county is shown in Figure 11. Dallas county enforced lockdown during April 2020. As it can be seen from the figure, the retail and recreation mobility went down to -40% from the normal state which illustrates the impact of a shelter-in-place order. As the recreation centers and retail industry started to open on May 18<sup>th</sup> [13], the mobility went up to -25%, but it still did not go back to the normal state. The retail and recreation mobility stayed steadily between -25% and -15% from May to December 2020. It seems from this analysis that people avoid going to recreational centers and non-essential business due to Covid-19 pandemic.

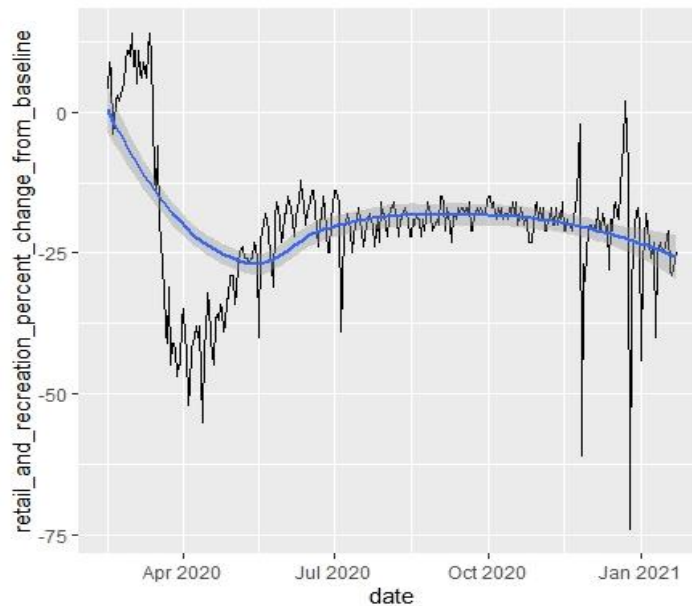


Fig. 11: Percent change of retail and recreation business from the baseline during Covid-19 pandemic in

Dallas county

➤ Impact on pharmacy and grocery mobility:

The percent change in grocery and pharmacy from the baseline in Dallas county is shown in Figure 12. It can be seen that the grocery and pharmacy mobility is increased by 30% from the baseline during March. It seems that the grocery and pharmacy mobility is increasing before the shelter-in-place order being effective on March 23. As soon the restrictions started, the pharmacy and grocery mobility went down by -20% in April. It stayed steadily at -10% from the baseline from May to November 2020. The pharmacy and grocery mobility have peaked to 20% and 30% from the baseline during December 2020 which it seems that the impact of celebrating the end of the year 2020 and Christmas Day on the pharmacy and grocery mobility.

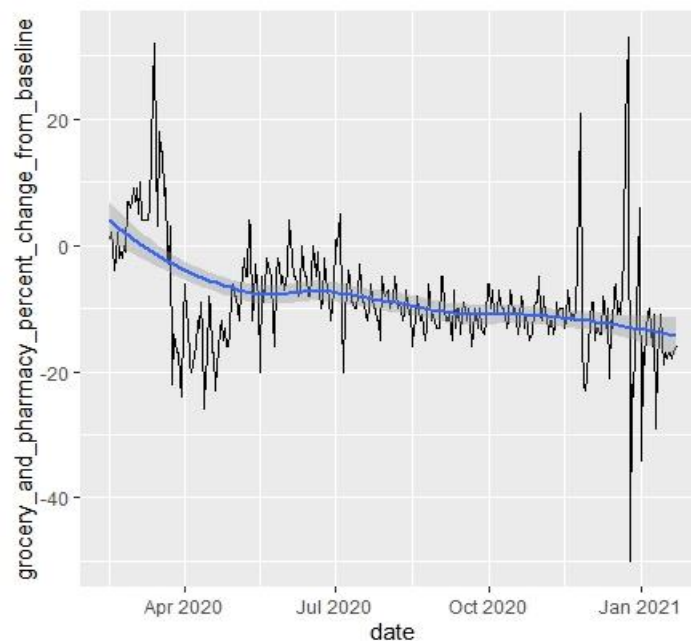


Fig. 12: Percent change of grocery and pharmacy from the baseline during Covid-19 pandemic in Dallas county

➤ Impact on parks mobility:

The percent change in parks from the baseline in Dallas county is shown in Figure 13. It can be seen from Figure 13. The parks mobility is increased to 55% from the baseline in middle of March before the lockdown restrictions. Subsequently, it declines to -45% from the baseline by April. The parks mobility went up and down between -55% to 40% from May to September. It started to decline from October 2020 until it reached -40% changes from the baseline by January 2021 which may be the effect of the cold weather on the parks mobility. It seems that the

lockdown restriction dramatically affects the parks mobility which causes the parks mobility to stay below the zero from April to May. In the winter season (November to January), the parks mobility continues to decline below the baseline. In the other periods, the parks mobility fluctuates and cannot be predicted.

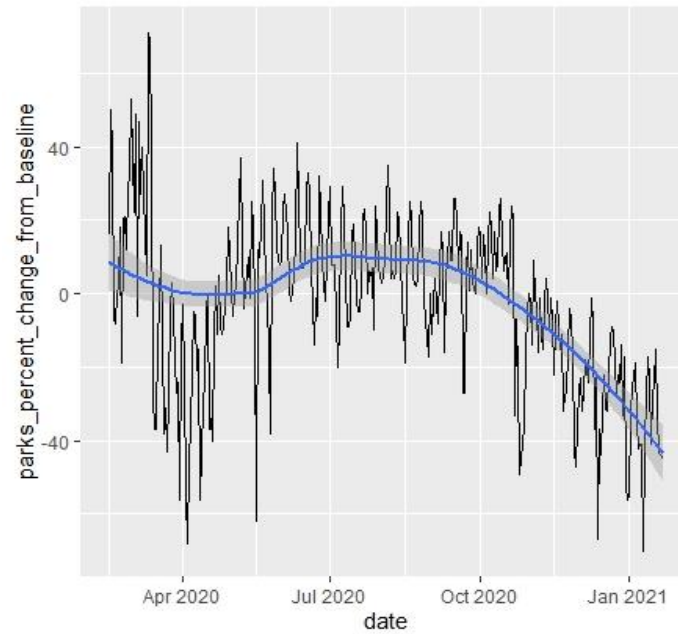


Fig. 13: Percent change of parks from the baseline during Covid-19 pandemic in Dallas county

➤ Impact on transit station mobility:

Dallas Area Rapid Transit (DART) who operates the buses and commuter rail in Dallas county enforces health measures such as requiring face mask, social distancing, and cleaning protocols as a response to slow down the spread of Covid-19 [14]. To see the impact of the covid-19 pandemic in the transit station mobility, Figure 14 shows the transit station mobility percent changes from the baseline in Dallas county Texas. As it can be seen from this figure, the transit station mobility was fluctuating around the base line between 5% and -5% before activating the shelter-in-place order on March 23<sup>rd</sup>. As soon as the order was effective at the end of March 2020, the transit station mobility went down to -50% from the baseline in April. The transit station mobility started to go up and fluctuates between -20% and -40% from May to October. Subsequently, the transit station mobility declined to -60% by January 2021. It seems that Covid-19 pandemic dramatically affects the transit station mobility in Dallas county.

Since the end of March where the shelter-in-place was issued for Dallas residents, the transit station mobility went down from the baseline and continue to decline due to social distancing where people avoid commuting using public transportation.

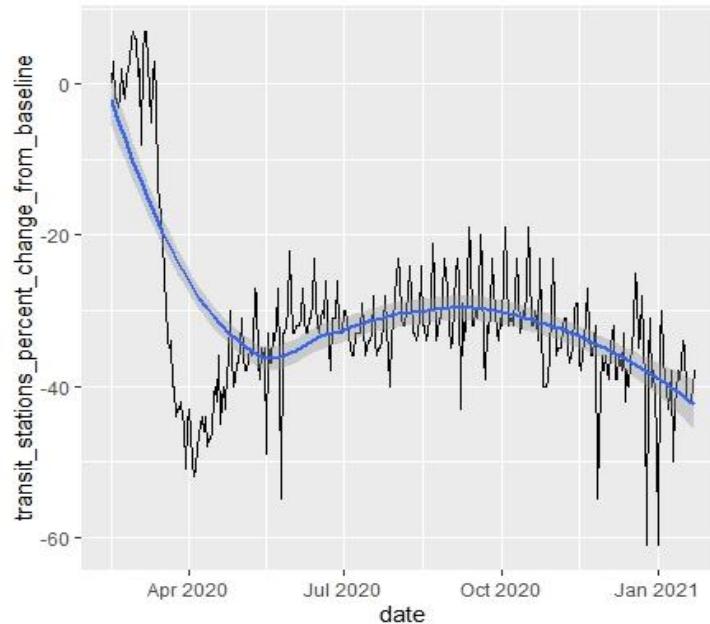


Fig. 14: Percent change of transit stations from the baseline during Covid-19 pandemic in Dallas county

➤ Impact on workplace mobility:

Covid-19 pandemic and shelter-in-place orders affect the workplace. Instead of face-to-face work atmosphere, the pandemic shifts the workplace toward working from home and utilizing a virtual work atmosphere such as Zoom, Skype etc. To see the impact of the covid-19 pandemic in the workplace mobility, Figure 15 shows the workplace mobility percent changes from the baseline in Dallas county Texas. As it can be seen from this figure, the workplace mobility started to go down as the shelter-in-place order took place on March 23<sup>rd</sup> 2020. The workplace mobility reached to -50% and -70% changes from the baseline by April and June, respectively. The workplace mobility was slightly increased and fluctuated during July to November. In December 2020 and January 2021, the mobility went down to -80% changes from the baseline. It seems that Covid-19 pandemic has a huge impact on the workplace. The Covid-19 pandemic and social distancing changed the normal work atmosphere and a large number of workers are working from home.

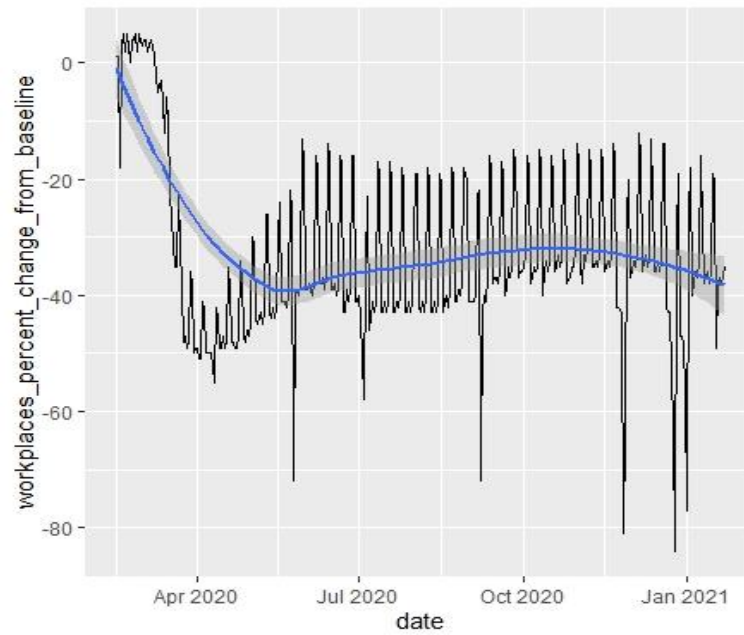


Fig. 15: Percent change of workplace from the baseline during Covid-19 pandemic in Dallas county

## 2.5 Exploring Relationships between Attributes

The correlation plot of some features given in Figure 16 implies that there is a strong relationship between confirmed cases, deaths, unemployed population, number of family households, poverty population and total population.

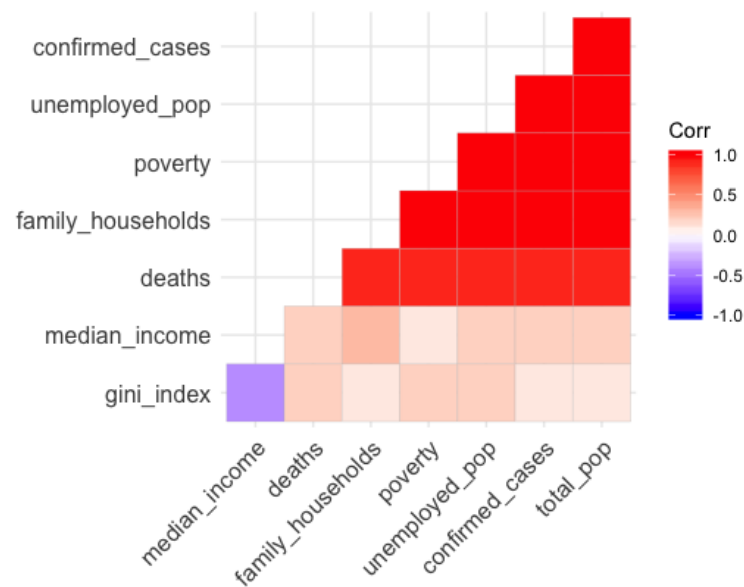


Fig. 16: Correlation plot

In Figure 17, the number of deaths\_per\_1000 and number of cases\_per\_1000 is presented. This figure is considering the population of state and has scaled the number of deaths and cases. As it can be seen in the figure, New York, New Jersey and Massachusetts sates are the three states with highest death rates.

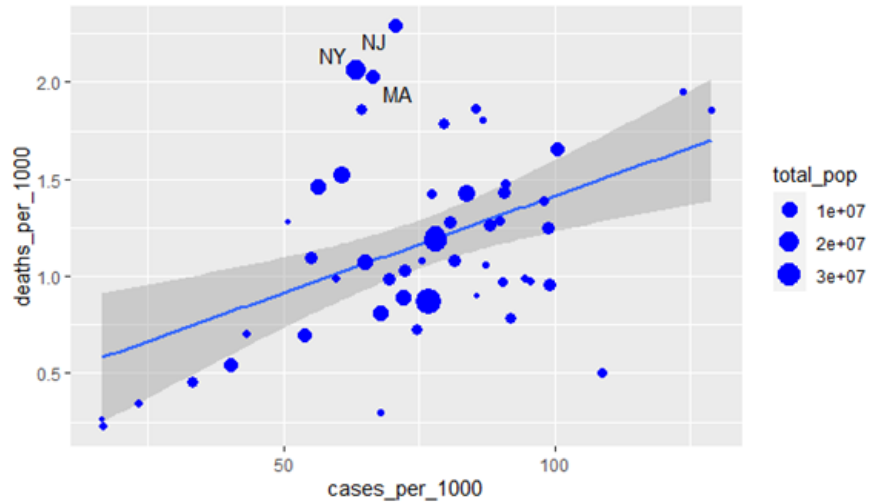


Fig. 17: deaths\_per\_1000 in cases\_per\_1000

In Figure 18, the relation between the cases per 1000 and total population is shown. We see that with increasing the population, the number of cases per 1000 does not change and we can conclude that there is no relation between the population and the relative number of cases in the states. New York, New Jersey, and Massachusetts sates that has the highest rate of deaths per cases based on the previous figure, have less relative number of cases.

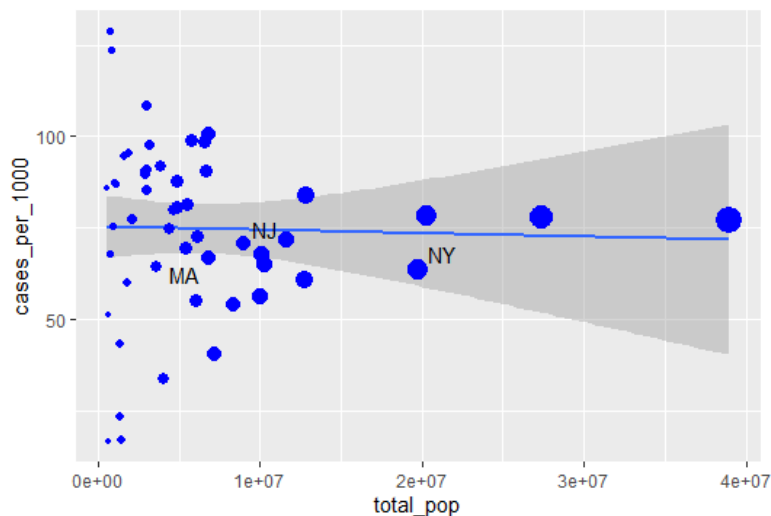


Fig. 18: cases\_per\_1000 in total\_pop



Figure 19 is figured based on the state wise data. It seems that the relative number of cases in population increased as the number of married households increases. The reason may be that the more members in the household, the more likely they will get infected.

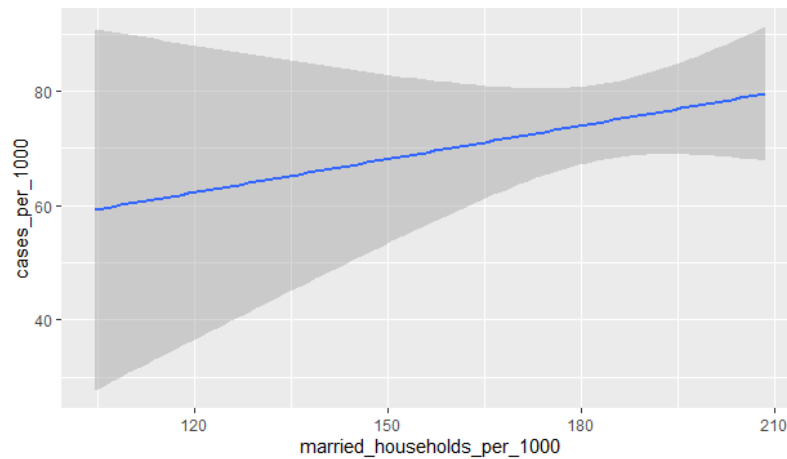


Fig. 19: cases\_per\_1000 in married\_households\_per\_1000

Figures 20 is showing the relative number of cases which is increasing when the number of white populations including the Hispanic population increases.

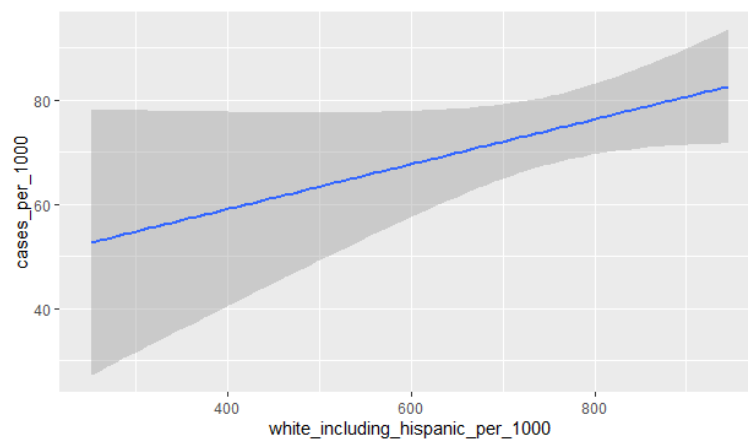


Fig. 20: cases\_per\_1000 in white\_including\_hispanic\_per\_1000

Figure 21 is showing the relative number of the cases compare to the relative graduated people. As the relative graduated people increases, the relative number of cases decreases. This is because graduate people have more information about the virus spreading and they have a better life level and can be more careful about themselves. They also do not need to use the

public transportation and mostly commute with their private cars which is safer. It can also be explained by their job type. They have the job that can be done at home online, but lower-level job cannot be done by work from home.

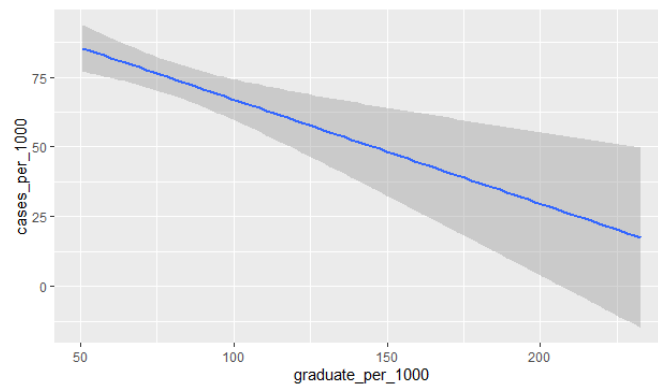


Fig 21: cases\_per\_1000 in graduate\_per\_1000

In Figure 22, the relation between the relative number of teenagers and the relative number of cases is shown. The male and female in age range of (15,20) summed up and considered as the total number of teenagers in the state. As the number of teenagers increases, the number of cases increases fast. One reason might be that teenagers are less careful about the virus and they are more willing to hang out with their friends and they can transmit the virus and bring it to the house and all the family will be exposed to the virus.

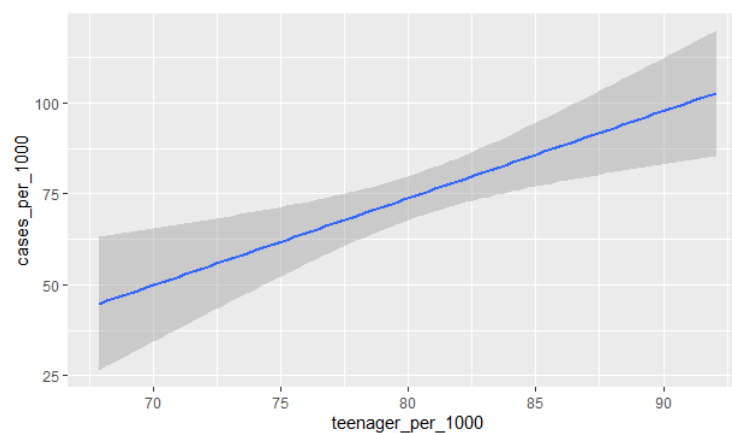


Fig 22: cases\_per\_1000 in teenager\_per\_1000

### 3 Data Preparation

#### 3.1 Data Set with Important Features

For this section, first we need to find some useful data about Covid-19 in US counties. Our focus is on the Texas counties. By searching through different data set exist on the internet, some useful data sets are found, but not all the data that are provided in these data sets are useful for our work. We need data set about the population, number of cases, facilities, and resources in each of these counties. In this regard, we extract useful information in the data sets we found and then merged them together. One challenge we faced was how to merge these data. If we used the county names, because of different spelling in data sets, we could not merge them properly. Therefore, we had to use a feature that is the same for all counties which is the `county_fips_code` which is a unique feature for each county. The `county_fips_code` is calculated using the “US\_Fips\_Codes” data set which provides the state and the `county_fips_code`. Then, we assigned this `county_fips_code` to the data set “Texas COVID-19 Data by County” data set. “Texas COVID-19 Case Count Data by County” data set provides the daily data about the number of covid-19 cases collected over one year in each county in Texas. The date for the first case and the spread rate is found using this data set.

The data set contains 15 features for each county in Texas. This data set aims to compare and analyze the covid-19 data in each county in Texas. The features are selected to analyze the impact of Covid-19 in each county in term of first case reported, Covid-19 spread rate, social distancing response, median age, confirmed cases, and the number of deaths. Moreover, hospital capacity and medical staff in each county in Texas are considered in the data set including total hospital beds, isolation rooms, total physician and dentist, registered nurses, and laboratory technicians [15]. The description of each variable is given in Table 10. The first ten rows in the data set is shown in Table 11.

Table 10: The description of important features in the data set (**Exceptional work included**)

Feature	Data Type	Description
County Name	Nominal	Name of each county in Texas
First Case Reported	Nominal	Date of the first case reported

Spread Rate	Ratio	The slope of the line derived from daily number of cases over one year
Total Population	Ratio	Total population in each county
Total physicians and dentists	Ratio	Total physicians and dentists in each county
Laboratory technicians	Ratio	Total laboratory technicians in each county
Registered nurses	Ratio	Total registered nurses in each county
Total hospital beds	Ratio	Total hospital beds in each county
Total isolation rooms	Ratio	Total number of airborne infection isolation rooms in each county
Median Age	Ratio	Median age
Deaths	Ratio	Number of deaths
Confirmed Cases	Ratio	Number of confirmed cases
Death Rate	Ratio	Percentage of number of deaths per total confirmed cases in each county in Texas
Deaths per 1000	Ratio	Number of deaths per 1000
Confirmed Cases per 1000	Ratio	Number of confirmed cases per 1000
Social Distancing Response	Ratio	The overall average changes in mobility from the baseline in each county in Texas

Table 11: The first ten rows in the data set

	County_Name	first_case	spread_rate	pop2021	physicians_and_dentists	laboratory_technicians	registered_nurses	hospital_beds
1	Anderson	4/1/2020	15.17885147	57747	5	10.0000	125.0000	12,000.00%
2	Andrews	04-04-2020	5.413299548	17577	8	10.0000	42.0000	3,400.00%
3	Angelina	3/26/2020	13.96753216	87700	7	55.0000	457.0000	36,800.00%
4	Aransas	04-05-2020	2.47703786	24832				
5	Archer	4/5/2020	1.990005574	8793				
6	Armstrong	5/15/2020	0.307911331	1929	0			
7	Atascosa	3/25/2020	10.90176001	48139	0	5.0000	48.0000	5,100.00%
8	Austin	3/25/2020	4.243127892	29292	0	3.0000	5.0000	1,000.00%
9	Bailey	5/5/2020	1.97588253	7098	0	5.0000	8.0000	2,500.00%
10	Bandera	4/9/2020	2.296859295	21316				
	isolation_rooms	median_age	deaths	confirmed_cases	death_rate_in_percentage	cases_per_1000	deaths_per_1000	distancing_response
	7	39.1	107	6063	1.764802903	104.9924671	1.852910108	-7.1372549
	3	31.6	46	1694	2.715466352	96.37594584	2.61705638	-14.660305
	15	36.8	260	7984	3.256513026	91.03762828	2.964652223	-6.34601
		49.6	35	1140	3.070175439	45.90850515	1.409471649	-9.00464
		44.8	11	773	1.423027167	87.91083817	1.25099511	-21.9867
		45.9	6	149	4.026845638	77.24209435	3.110419907	
	2	35.4	139	5395	2.576459685	112.0712935	2.887471697	-12.4272
	1	40.9	30	1842	1.628664495	62.88406391	1.024170422	-8.4729167
		35.2	17	794	2.141057935	111.8624965	2.395040857	-17.836991
		52	24	1272	1.886792453	59.67348471	1.125914806	-12.40566

## 3.2 Visualization of Important Features

In this section, the visualizations of the most important features in the data set are reported.

### 3.2.1 Covid-19 spread rate in each county in Texas

The frequency of spread rate is shown in figure 23. As it can be seen, most counties in Texas have a small spread rate. In particular, most counties in Texas has a spread rate less than 250. The relation between Covid-19 spread rate and total population in each county in Texas is shown in Figure 24. As shown in Figure 24, five counties in Texas (Harris, Dallas, Tarrant, Bexar, and El Paso) have spread rate more than 300. Four of these five counties have more population than the other 249 counties with more than 2 million population. However, El Paso county is an extreme situation because it has lower population (674,826 population) but the spread rate is high. It can be concluded that the county with more population will be most likely to have a higher Covid-19 spread rate.

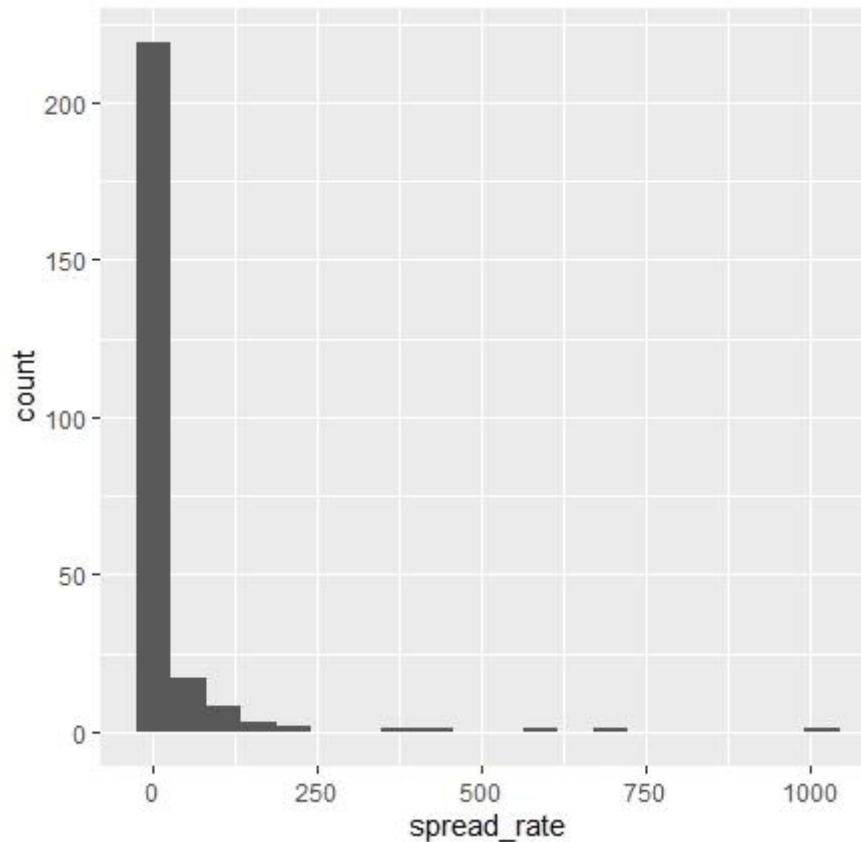


Fig 23: Spread rate histogram

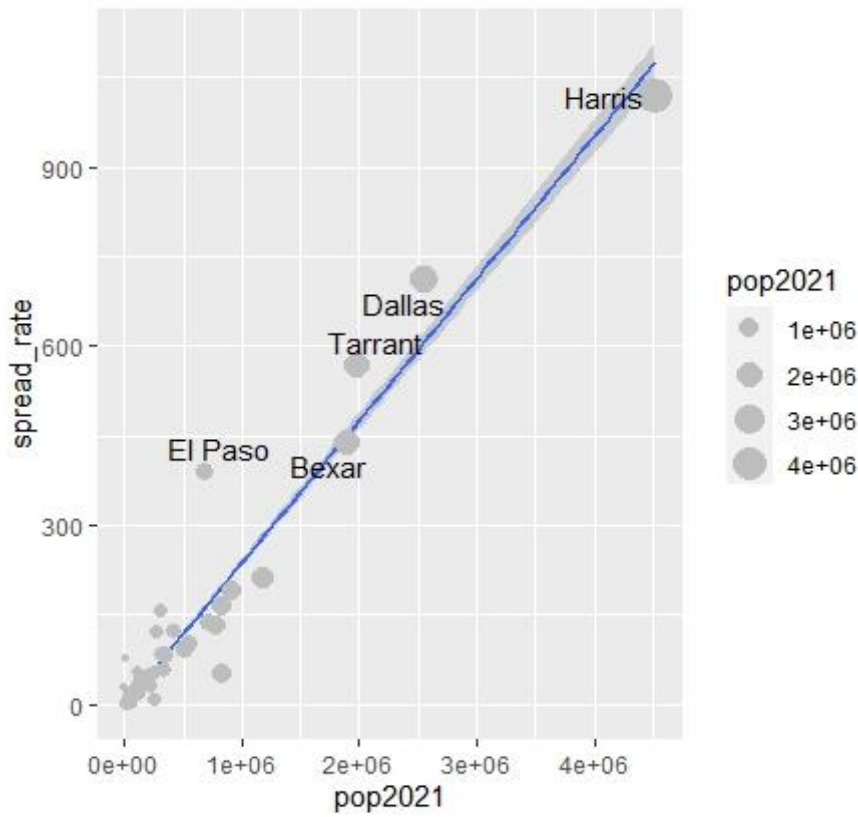


Fig 24: Relation between Covid-19 spread rate and total population in each county in Texas

### 3.2.2 Hospital capacity in each county in Texas

In this section, the hospital capacity for each county in Texas is analyzed. The histogram of hospital beds in Texas is shown in figure 25. Most counties in Texas has a total hospital beds lower than 1500 beds. The largest number of hospital beds is around 8000 beds.

The total hospital beds and isolation rooms have some missing values. The NAs in total hospital beds and isolation rooms are 77 NA's and 103 NA's respectively. The missing values are removed to plot the relation between these variables and total population.

The relation between total hospital beds and total population in each county in Texas is shown in figure 26. As it can be seen from figure 26, the number of total hospital beds is large for high populated county. The highest hospital beds are in Dallas, Bexar, and Tarrant with 7923, 7326, and 5331 hospital beds, respectively. The relation between isolation rooms for airborne infection and total population in each county in Texas is shown in figure 27. As it can be seen from figure 27, the number of total isolation rooms is large for high populated county. The highest isolation rooms are in Dallas, Bexar, and Tarrant with 493, 429, and 318 isolation

rooms beds, respectively. It can be concluded that total hospital capacity (hospital beds and isolation rooms) is high for high populated county, although hospital capacity for high populated county such as Harris is missing in this data that obtained from [15].

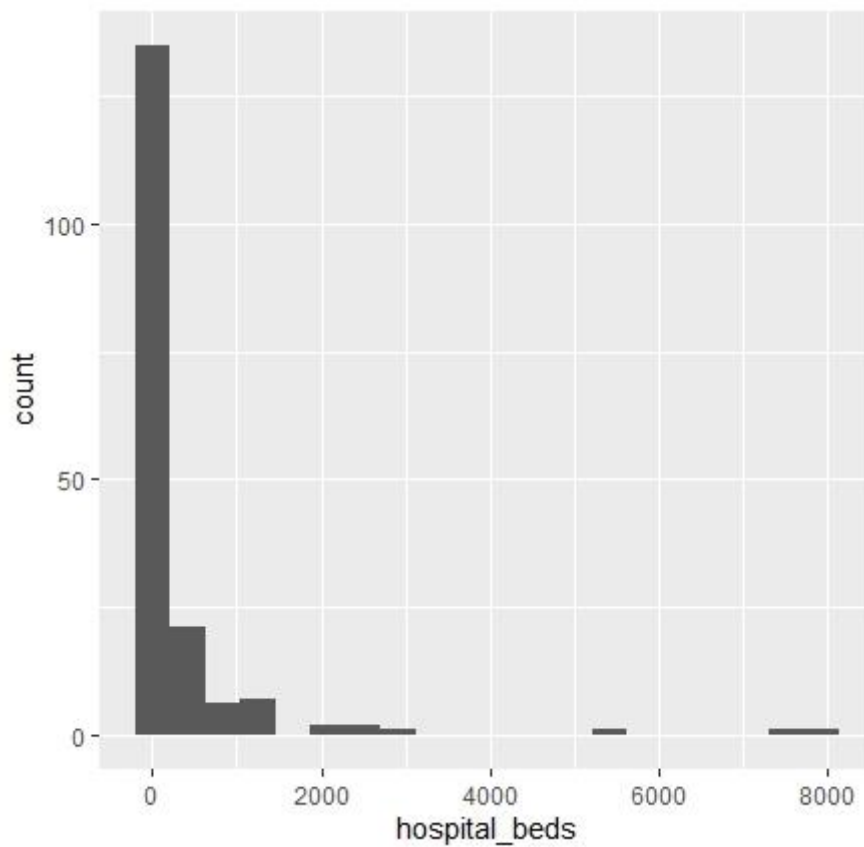


Fig 25: Total hospital beds histogram

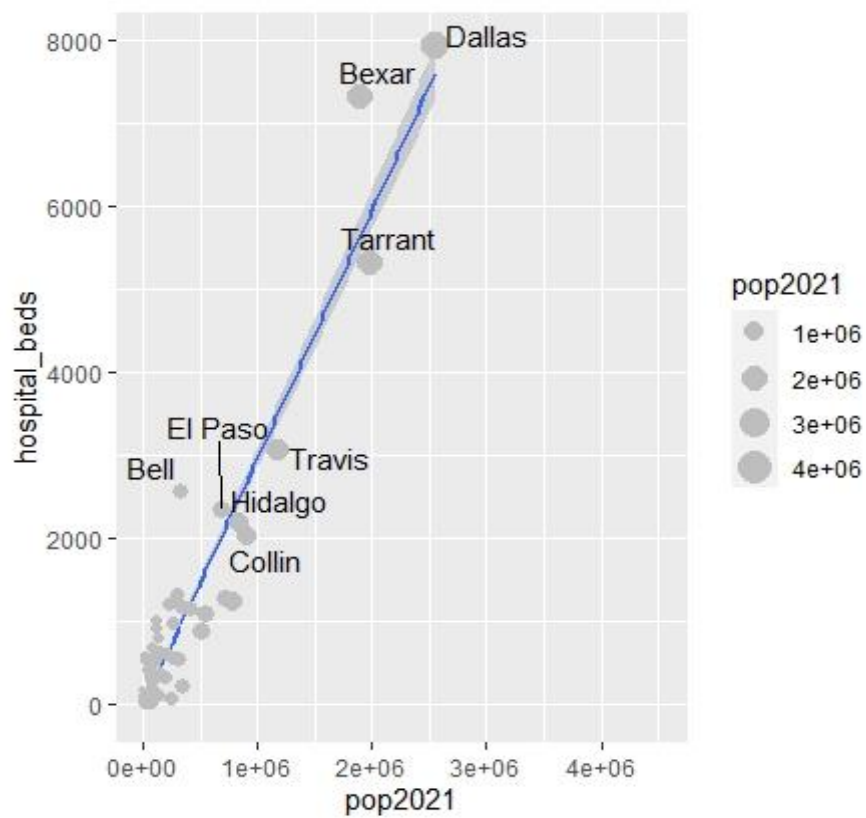


Fig 26: Relation between total hospital beds and total population in each county in Texas

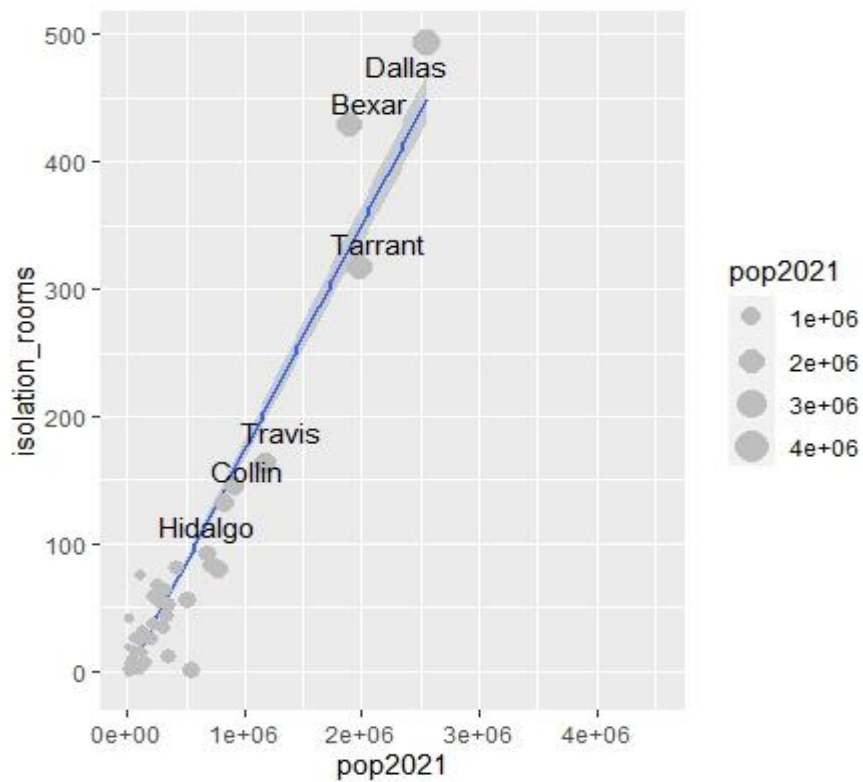


Fig 27: Relation between total isolation rooms and total population in each county in Texas



### 3.2.3 Medical staff in each county in Texas

The medical staff in each county in Texas is analyzed in this section. The data for the three variables (total physicians and dentists, registered nurses, and laboratory technicians) are obtained from [15]. There are 75 missing values for each of these variables. The histogram for the total physicians and dentists in Texas is shown in figure 28. The maximum total physicians and dentist in Texas is 2159. More than 150 counties have reported few numbers of physicians and dentists.

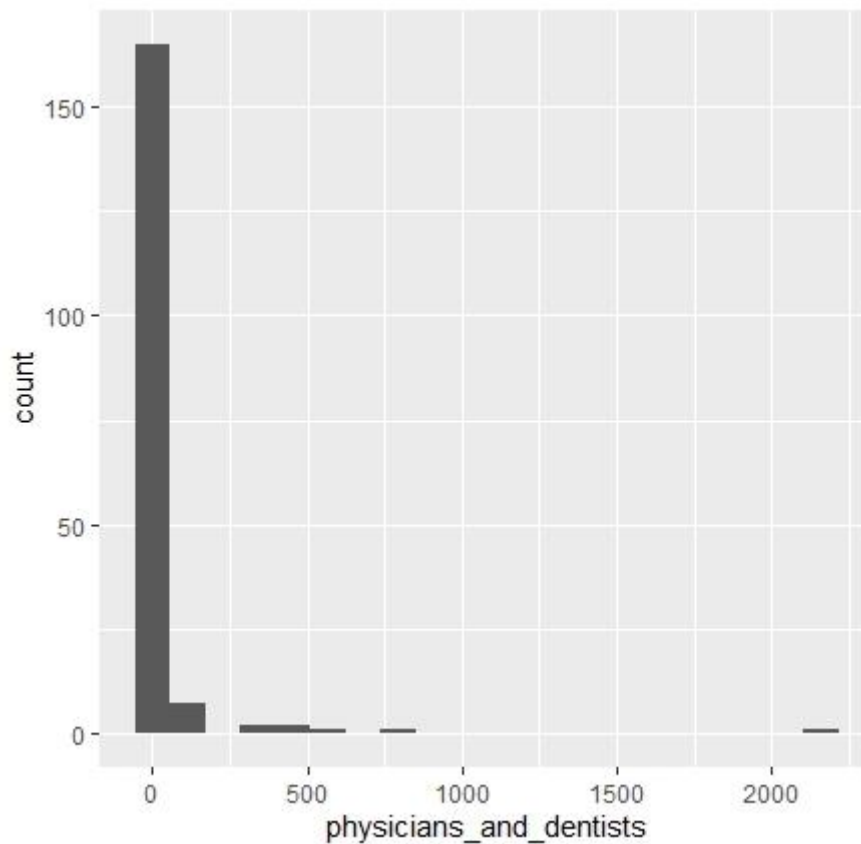


Fig 28: Histogram of the total physicians and dentists variable

The missing values are removed to plot the relation between these variables and total population. The relation between total physicians and dentists and total population in each county in Texas is shown in figure 29. Harris county has the highest number of physicians and dentists with total of 2159. Bexar county has the second highest number of physicians and dentists with total of 843 followed by Dallas county with a total of 489. Although more population in Dallas county than Bexar county, the total number of physicians in Bexar is nearly twice the number of physicians in Dallas county. Therefore, these counties can be considered

as outliers. From figure 29, we can conclude that large populated county is most likely to have higher number of physicians and dentists.

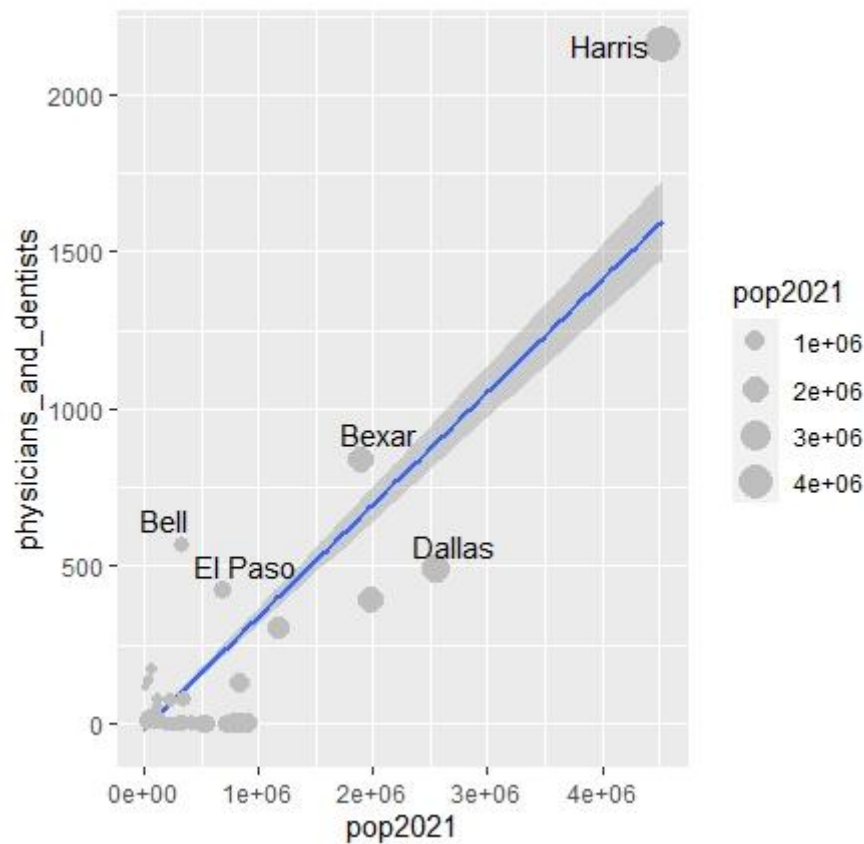


Fig 29: Relation between total physicians and dentists and total population in each county in Texas

The relation between total registered nurses and total population in Texas is shown in figure 30. Similarly, Harris county has the highest number of registered nurses with total of 28,094. Dallas and Bexar counties have the second and third highest number of registered nurses with totals of 15,309 and 10,875, respectively. As it can be seen from figure 30, the number of registered nurses is most likely high for high populated county in Texas. Figure 31 shows the relation between total laboratory technicians and hospital beds in Texas. Harries has the highest number of laboratory technicians with total of 2355. It can be concluded from Figure 31 that the number of laboratory technicians is large for county with large hospital bed capacity.

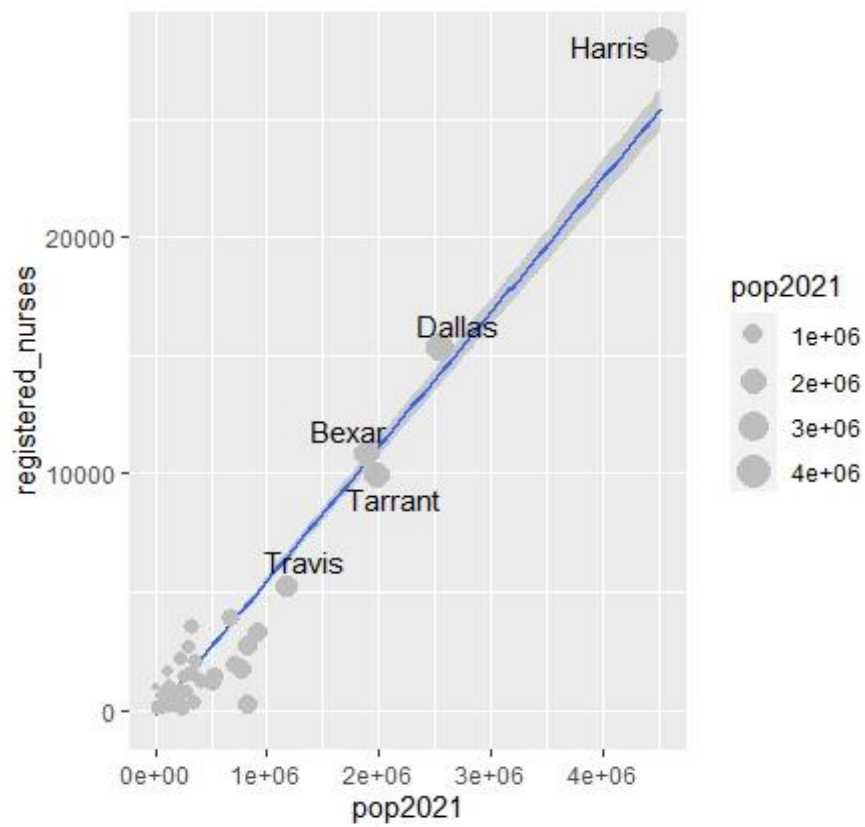


Fig 30: Relation between total registered nurses and total population in each county in Texas

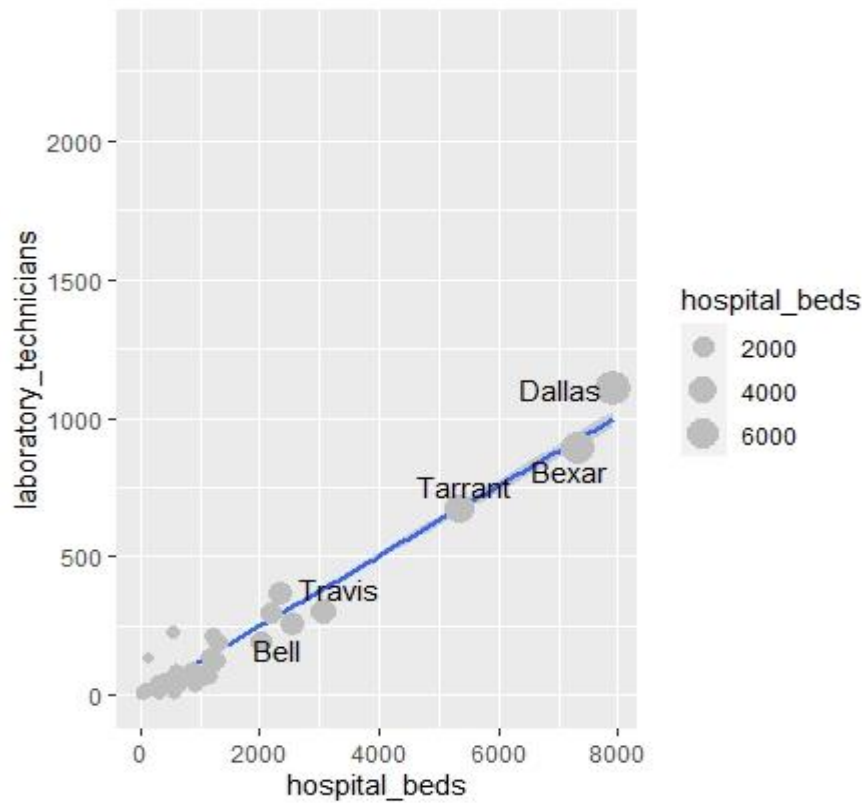


Fig 31: Relation between total laboratory technicians and hospital beds in each county in Texas

### 3.2.4 Correlation between medical staff, hospital capacity, death rate, and population in Texas

The correlations between hospital capacity, medical staff variables, death rate, and population are shown in figure 32. The total physicians and dentists have around +0.5 correlation with isolation rooms, population, registered nurses, and hospital beds. However, total physicians and dentists have around -0.5 correlation with death rate. This negative correlation between total physicians and death rate shows that the death rate is decreasing as more total physicians available in counties. Meanwhile, hospital beds have +1 correlation with registered nurses, laboratory technicians, isolation rooms, and population. Registered nurses have +1 correlation between hospital beds, isolation rooms, laboratory techniques, and population. Lastly, death rate has a -0.5 correlation with hospital capacity and medical staff. It can be concluded that as hospital capacity and medical staff are increased, the death rate due to covid-19 infection will be most likely decreased.

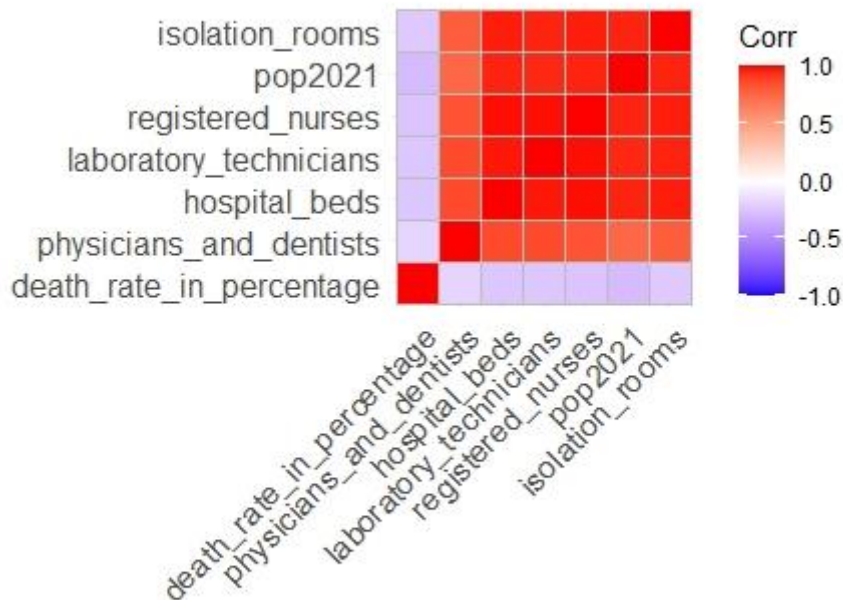


Fig 32: Correlation among hospital capacity, medical staff, death rate, and population variables

### 3.2.5 Relationship between the number of deaths and confirmed cases

In figure 33 the relation between the number of deaths and total number of cases is shown. As total number of cases increases, the total number of deaths also increases and that is natural, but we see that in some counties, the death is more than other counties. Harris county, Ellis county and Hidalgo counties are in this category. Some counties have lower deaths than the average death rate like Dallas and Tarrant counties. Based on this figure, we cannot derive much valid conclusions because the number of the cases and the number of deaths should be compared relative to their population which is presented in the next figure.

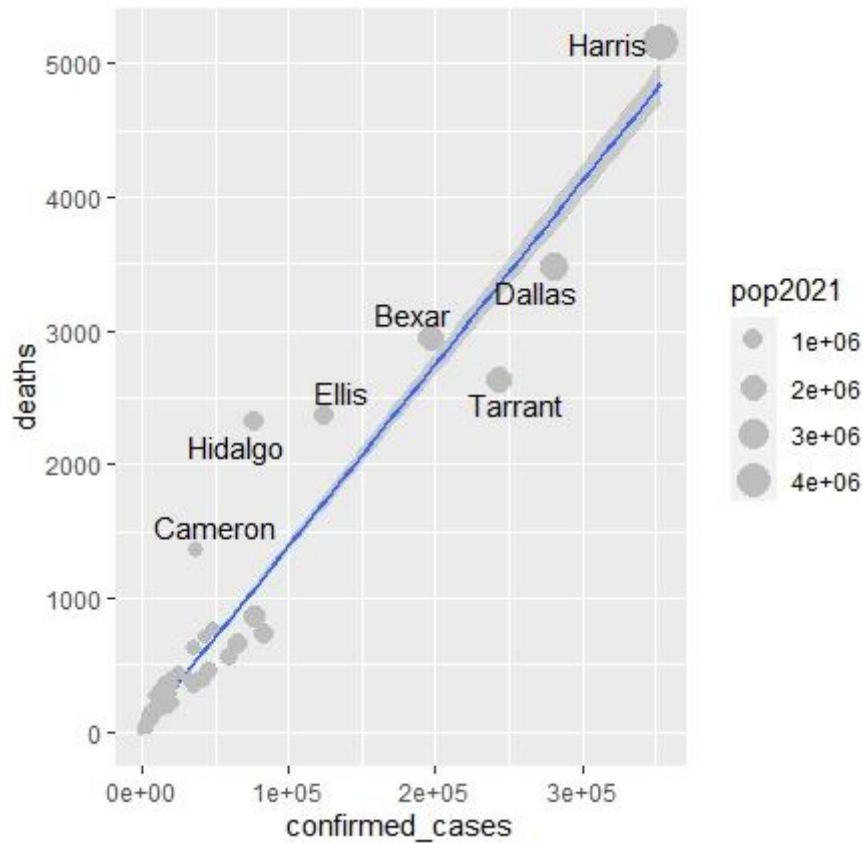


Fig 33: The relation between confirmed cases and deaths in Texas

In the previous figure, we considered the relation between the death and total number of cases and comparing these features, it is hard to make conclusion because they are related to the total population. In this part, we consider their relative numbers. In this regard, the relation between death\_per\_1000 and cases\_per\_1000 is presented in figure 34. By increasing the relative number of the cases, the relative number of deaths is increasing. It is because when the

relative number of cases increases, the number of relative deaths increases because of lack of testing and facilities. Also, the relative number of deaths is very high in some counties and it can be seen that higher rate of relative deaths corresponds to the counties with less populations and this is because of lower income and far and less available healthcare facilities for these counties.

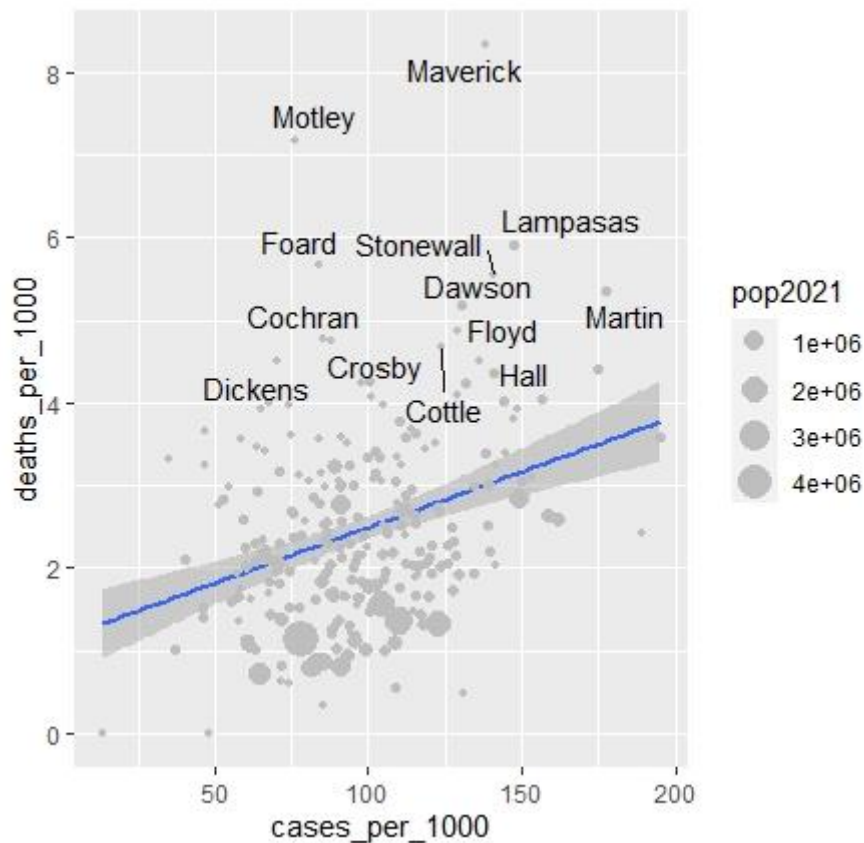


Fig 34: The relation between confirmed cases per 1000 and deaths per 1000 in Texas

The relation between deaths per 1000 and the population is shown in the figure 35. Counties with higher population have the highest relative death and the relative death is increasing as the number of populations increases and that is because in larger counties, people have more knowledge about the Covid19 and how to deal with the illness, but in smaller counties, people do not have access to information. Also, they have less available healthcare facilities and hospitals, and this makes their illness progress and lead to their death. Another reason is their lower income, and the costs are not affordable for them.

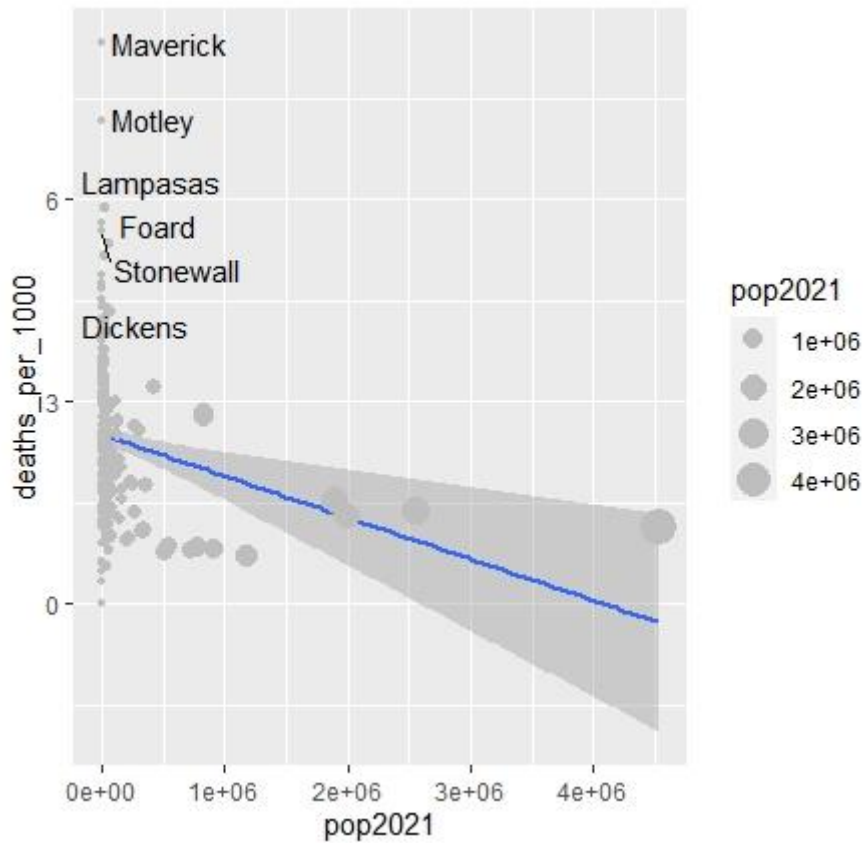


Fig 35: The relation between death per 1000 and population in Texas

### 3.2.6 Social distancing response for each county in Texas

In this section, the social distancing response for each county in Texas is analyzed. The overall social distancing response is obtained by averaging all mobility variables in the mobility data sets. There are 35 missing values in the social distancing variable. The histogram of the social distancing response in Texas is shown in Figure 36. The median of social distance response is -10.64% changes from the baseline in the Texas mobility. The minimum and maximum percent changes in the Texas mobility are -54.928% and 18.516%, respectively. The distribution of the social distancing response variable is shown in Figure 37. The social distancing response in the most counties in Texas is between -15.56% and -5.344% mobility changes from the baseline. The relation between the social distancing response and population in Texas is shown in Figure 38. The horizontal line in Figure 38 shows that there is no relation between the social distancing response and total population in Texas.

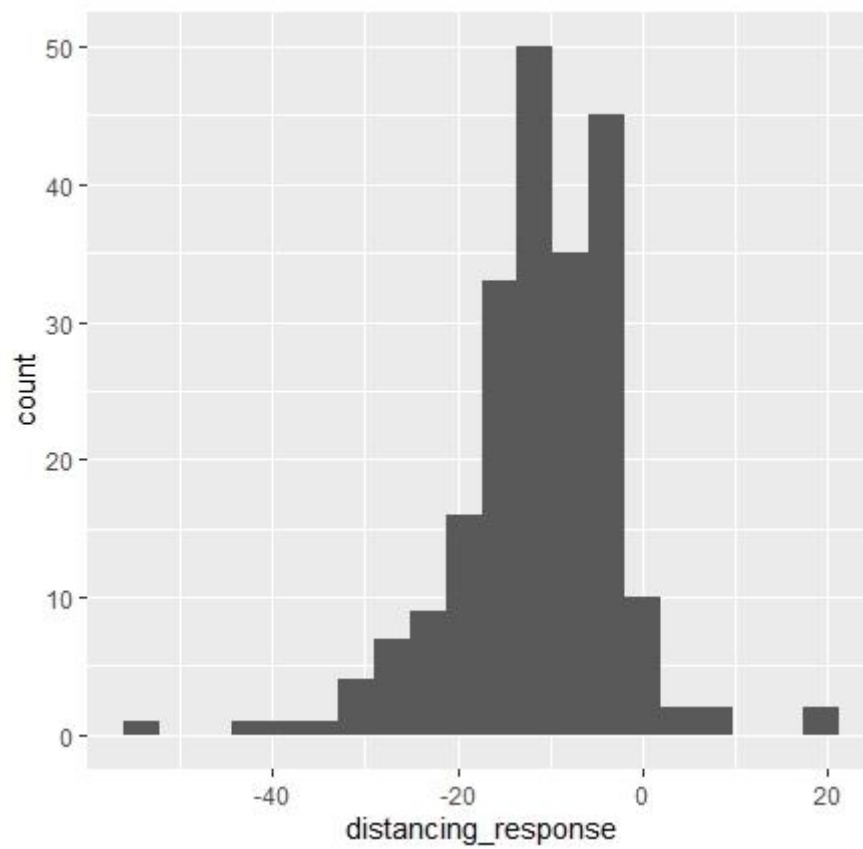


Fig 36: Histogram of the social distancing response

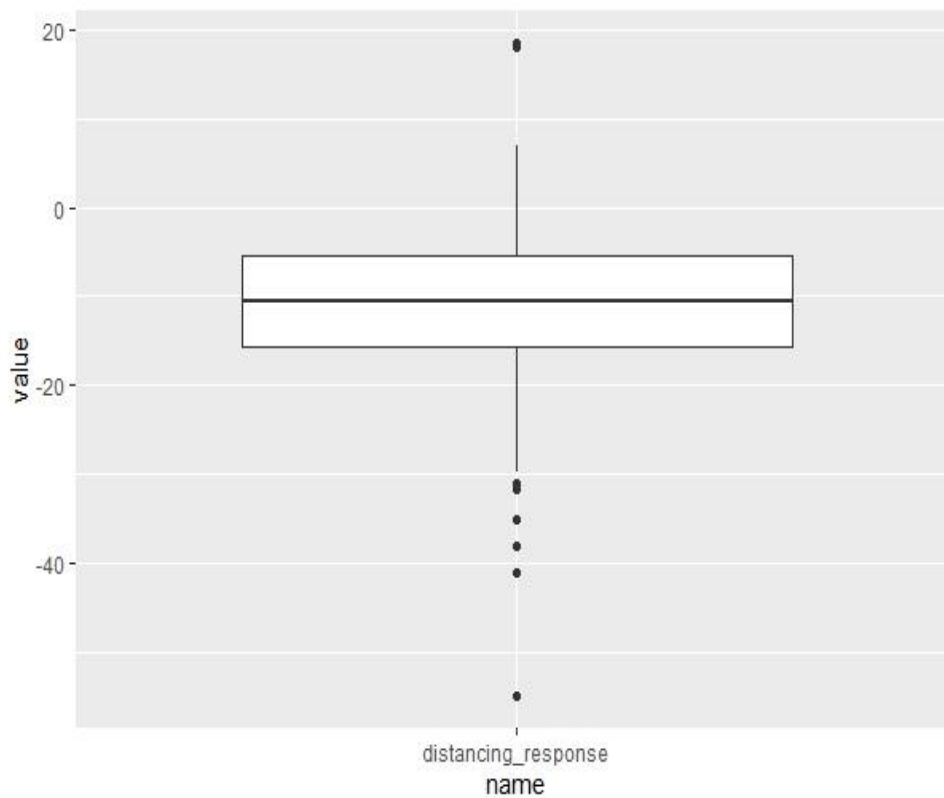


Fig 37: The distribution of the social distancing response variable



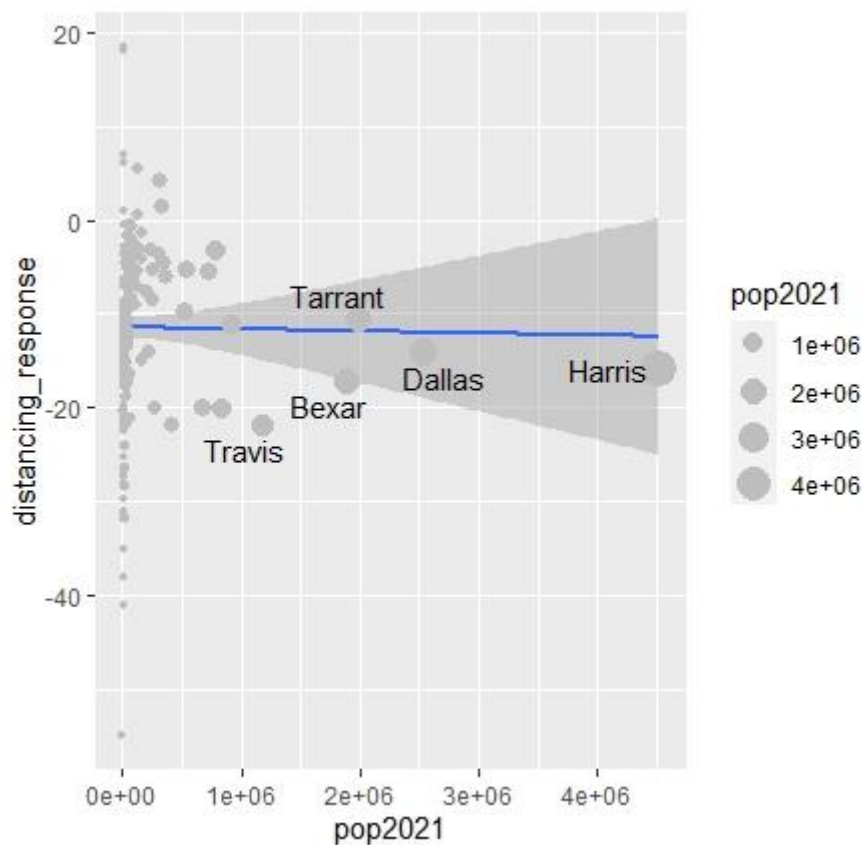


Fig 38: The relation between the social distancing response and population in Texas

### 3.2.7 Relationship Correlation among deaths, spread rate, population, confirmed cases, median age and social distancing response

Correlations among deaths, spread rate, social distancing response, population, confirmed cases, median age, and death rate are shown in figure 39. There is no correlation between social distancing response and other variables as shown in figure 39. The number of deaths has a high positive correlation with spread rate, population and confirmed cases and that is obvious because when we have larger population, the number of people infected are more in that county and by having higher number of cases, the number people who lose their lives also increases. Also, by having larger spread rate, the number cases also increase which leads to higher number of deaths in that county. Meanwhile, the number of deaths has negative correlation with median age. This negative correlation between number of deaths and median ages shows that the older people will most likely to lose their live due to covid-19 infection.

The spread rate has high positive correlation with deaths, total population, and number of confirmed cases. This virus is mainly spreading by the people contacting each other and if a county has higher population, then more people are most likely to be in contact with each other and this leads to higher spread rate. Meanwhile, spread rate has a negative correlation with the death rate and median age. Median age has a negative correlation with spread rate, population, and number of death and confirmed cases. This negative correlation shows that older people most like to get infected and die due to the covid-19 pandemic. The percentage of death rate has a negative correlation with confirmed cases, population, and spread rate. As more people get infected with covid-19, the rate of death decreases. This shows that the death rate due to covid-19 is small comparing to the total confirmed cases.

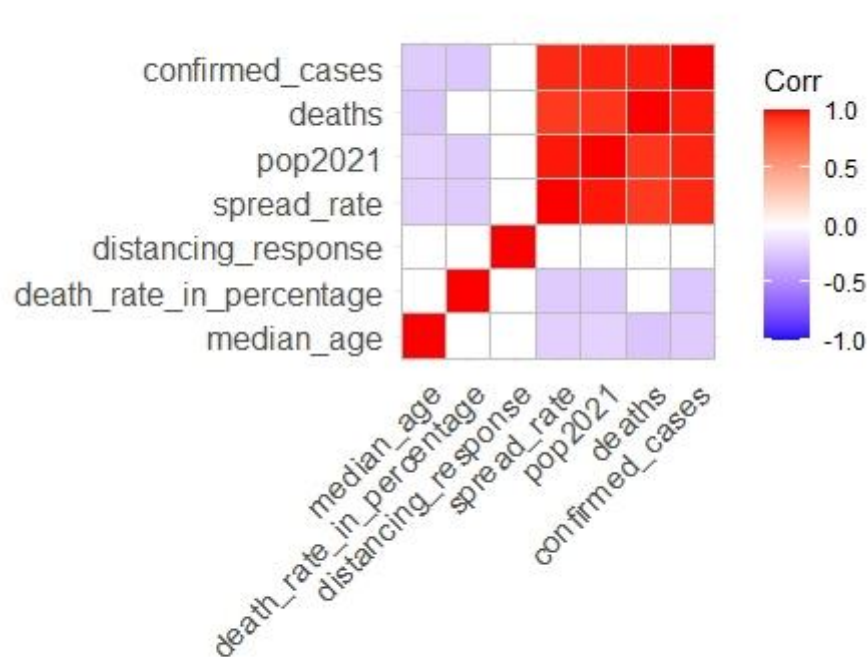


Fig 39: Correlations among deaths, spread rate, social distancing response, population, confirmed cases, median age, and death rate

## 4 Conclusion

The Covid-19 pandemic influenced lives of people in the world and led to loss of many lives and presented a dramatic challenge to food system and jobs worldwide. Since the beginning of this pandemic, lots of researchers have tried to find the factors affecting the spread of this virus and different tools are used for answering these questions. At the first months of this pandemic, there was less data available about this virus and its spread and governors and healthcare systems did not know how they can control the spreading this virus and when it will be over, but after passing time and collecting data about this pandemic more analysis have been done to answer these questions. Data mining is one of the tools dealing with this complicated data and it able to analyze and find strong conclusions. In this project we focused on the US specially Texas state and using the data available, we answered to some of these questions. The results show that lock down and social distancing which led to the less mobility in US helped a lot in lowering the spread rate specially in counties with large population and governors should be serious about considering it in denser aeras. Also, analysis show that different counties has respond to this pandemic differently and in counties with less healthcare facilities, the death rate is very high, and more testing and bed and ventilation should be provided for this group of counties. The correlation between the number of cases and deaths and other factors is also investigated in this project and a strong correlation between the number of cases and number of unemployed populations, poverty and number of family household can be seen which show that these groups of people should be taken into account in proving more healthcare facilities during the pandemic. The correlations among hospital capacity, medical staff and death rate show that the death rate due to covid-19 infection will be most likely decreased.

## References

- [1] <https://www.cdc.gov/coronavirus/2019-ncov/faq.html#Basics>
- [2] <https://coronavirus.dc.gov/data>
- [3] <https://www.vox.com/2020/3/10/21171481/coronavirus-us-cases-quarantine-cancellation>
- [4] <https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html>
- [5] <https://www.tmc.edu/coronavirus-updates/overview-of-tmc-icu-bed-capacity-and-occupancy/>
- [6] <https://www.mathematica.org/features/covid-19-curated-data-modeling-and-policy-resources>
- [7] [https://www.cdc.gov/mmwr/Novel\\_Coronavirus\\_Reports.html](https://www.cdc.gov/mmwr/Novel_Coronavirus_Reports.html)
- [8] <https://www.ncsl.org/research/health/state-action-on-coronavirus-covid-19.aspx>
- [9] <https://www.wsj.com/articles/a-state-by-state-guide-to-coronavirus-lockdowns-11584749351>
- [10] <https://www.kff.org/coronavirus-covid-19/issue-brief/state-covid-19-data-and-policy-actions/>
- [11] <https://datausa.io/coronavirus>
- [12] <https://www.mdanderson.org/cancerwise/does-social-distancing-help-prevent-coronavirus-covid-19-spread.h00-159383523.html>
- [13] <https://www.huschblackwell.com/texas-state-by-state-covid-19-guidance>
- [14] <https://www.dart.org/health/>
- [15] <https://console.cloud.google.com/marketplace/product/aha-public-data/hospital-capacity?project=smiling-diode-307114&folder=&organizationId=>