COVID-19 Cases Data Analysis and Visualization: A look at how COVID-19 affected people by race and in the state of Texas.

Authors:

Zahra Hoobakht Ph.D student in Electrical Engineering,

Jason Brown Graduate Student in Computer Science,

BhargavaSharabha Pagidimarri Graduate Student in Data Engineering,

Lyle School of Engineering,

Southern Methodist University

Project 1:
CS7331- Data Mining
Instructor : Dr. Maya Al Dayeh
October, 2023.

Contents

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION: Understanding Coronavirus (COVID-19) Disease

The year 2019 witnessed the emergence of a novel coronavirus, SARS-CoV-2, which has since triggered a global pandemic known as COVID-19. This pandemic has not only reshaped daily life but has also generated an unprecedented level of scientific and public interest. Understanding COVID-19, both as a disease and as a global phenomenon, has become a critical priority for healthcare professionals, researchers, policymakers, and the public [1].

Covid-19 has swept across societies around the world, affecting individuals and societies in profound ways. It has made people in all corners of the world sick, caused tragic loss of life, and challenged healthcare systems and governments on an unparalleled scale [2]. In the United States, the pandemic has had a significant impact, with 23,750,108 confirmed cases and 392,204 deaths reported on January 19, 2021 [3].

In this report, we will review COVID-19 data for different U.S. states and countries, aiming to discover the influencing factors in the disease's spread. We will investigate various factors, including gender, race, age, and other relevant factors, to understand their potential impact on the spread of the disease.

CHAPTER 1: Business Understanding

In this chapter, we introduce some essential context and foundational knowledge necessary for a comprehensive understanding of COVID-19 data.

1.1 COVID-19

Coronavirus disease (COVID-19) is an infectious respiratory illness caused by a coronavirus known as SARS-CoV-2. It was first identified in Wuhan, China, in late 2019 and quickly spread to become a global pandemic [4]. COVID-19 can cause a wide range of symptoms, from mild to severe, including fever, cough, difficulty breathing, and loss of taste or smell. In severe cases, it can lead to pneumonia, acute respiratory distress syndrome (ARDS), and, tragically, death [5].

1.2 Social Distancing and Flattening the Curve

Social distancing is a public health measure that involves reducing close physical contact between people to limit the spread of infectious diseases. It typically includes measures like staying at least 6 feet away from others, avoiding large gatherings, working from home, and practicing good hand hygiene. Flattening the curve is a strategy to slow the spread of a contagious disease, like COVID-19, to avoid overwhelming healthcare systems. By implementing social distancing and other preventive measures to reduce the peak number of cases at any given time, and spreading out the cases over a longer period, helps hospitals manage the patient load more effectively.

1.3 Significance, Stakeholders, and Decision-Making: COVID-19 Data Insights

Monitoring data on virus spread, hospitalizations, and available resources plays a critical role in effectively managing and mitigating the impact of the COVID-19 pandemic. This practice allows for early detection of outbreaks similar to COVID-19, enabling targeted responses to the virus's spread. The study of the data also allows for the prediction of essential resources, including hospital beds, ventilators, and medical staff, ensuring that healthcare systems can effectively meet the treatment and care demands of severe pandemic diseases such as COVID-19 without becoming

overwhelmed. Understanding the availability of essential resources enables informed decision-making by public health officials and policymakers. Furthermore, it contributes to flattening the curve, preventing healthcare system overload, and raising public awareness, collectively reducing the impact of the virus on individuals and communities.

Stakeholders typically refer to individuals or groups within the government or organizations that are directly involved in managing and responding to the crisis. These internal stakeholders play crucial roles in coordinating, implementing, and monitoring various aspects of the pandemic response. Some key internal stakeholders in the context of COVID-19 include:

Government Stakeholders: Governments at the national level are responsible for developing and implementing public health policies, coordinating resources, and managing the overall response to the pandemic. State and local governments play a crucial role in implementing public health measures, managing healthcare resources, and responding to localized outbreaks.

Public Health Agencies: Public health departments and agencies are essential stakeholders responsible for disease surveillance, contact tracing, testing, and the dissemination of public health guidelines.

Healthcare Stakeholders: Healthcare providers, hospitals, and clinics are on the front lines of treating COVID-19 patients and managing healthcare resources.

Pharmaceutical and Vaccine Manufacturers: Companies involved in vaccine and drug development are key stakeholders in the development, production, and distribution of vaccines and treatments.

Research and Academic Stakeholders:Research Institutions: Research organizations and academic institutions are involved in COVID-19 research, epidemiological studies, and vaccine development.

In my opinion, Healthcare providers are stakeholders in research related to the COVID-19 pandemic. As healthcare providers have a vested interest in staying informed about the latest research, because it directly impacts their ability to provide effective care to COVID-19 patients. Healthcare providers are on the front lines of patient care and are greatly impacted by the latest research findings, treatment protocols, safety measures, and the development of vaccines.

CHAPTER 2: Data Understanding

The datasets available were the COVID-19 Cases Plus Census file, COVID-19 cases TX, and the Global Mobility Report. These csv datasets were provided with the assignment and originally pulled from the Google Cloud platform. Google Cloud's marketplace hosts many other COVID-19 datasets that they make available free of charge. These datasets were provided to Google by governmental, medical, and research organizations as a means to assist in responding to the COVID-19 epidemic.

2.1 COVID-19 Plus Census Dataset Data Description

2.1.1 COVID-19 Plus Census Dataset The COVID-19 plus Census dataset was queried from the USAFacts Dataset repository in Google Cloud's Marketplace. The Covid-19_cases_plus_census.csv dataset has 259 features and 3142 observations. The dataset is in effect normal census data, accompanied by COVID-19 related data which was source by the CDC, and state and local heatlh agencies. The COVID-19-related features hold the total number of confirmed cases and the total number of COVID-19-related death counts from across all US cities up until the date of January 1, 2021. The demographic data accompanied by COVID-19 data gives a real insight into how differently these communities were more naturally equipped to respond to the pandemic and how some communities felt the effects worse than others. The data available can be used to identify racial and financial disparities as well as educational ones that exist in cities reported in the Census data.

2.1.2 Verifying COVID-19 with Census Data Quality

There were missing values in the dataset and exactly 10 columns that were titled but held no data under those features. Those missing features pertained to the population's marital statuses and Spanish-speaking populations. There were no duplicate rows.

Data UnderstandingTable1: appropriate Summary Statistics

	Min	Median	Mean	Max
Confirmed Cases	0.0	1916.5	4955.0	1002614
Total Population	74	25692	102166	10105722
Deaths	0.0	31.0	102166	10105722
Median Age	21.60	41.20	41.15	66.40
Median Income	19624	48066	49754	129588

2.1.3 The COVID-19 with Census Data and Race demographics

The COVID-19 with Census Data allows us to use demographic information like Race in conjunction with confirmed cases and deaths. The below table shows the majority of Black, majority of White, and majority of Hispanic counties in conjunction with the number of COVID-19-related cases and deaths.

2.1.3.1 US Counties Where the Black Population is the Majority

There are 129 US counties where the Black population is the majority. The county with the most COVID-19 related deaths and confirmed cases was Philadelphia County in Pennsylvania. The maximum number of deaths of Black Majority counties was 2,732

Important Summary Statistics:

Data Understanding Table2: Summary Statistics COVID and Black Majority Pop

	Min	Max	Mean	Median
Confirmed Cases	62	98,479	7,087	1,654
Median Income	20,330	78,607	35,566	33,545
Deaths	0	2,732	35,566	42
Total Population	1,296	1,569,657	103,509	20,506

Top 10 Majority Black Counties confirmed cases (in descending order) and deaths:

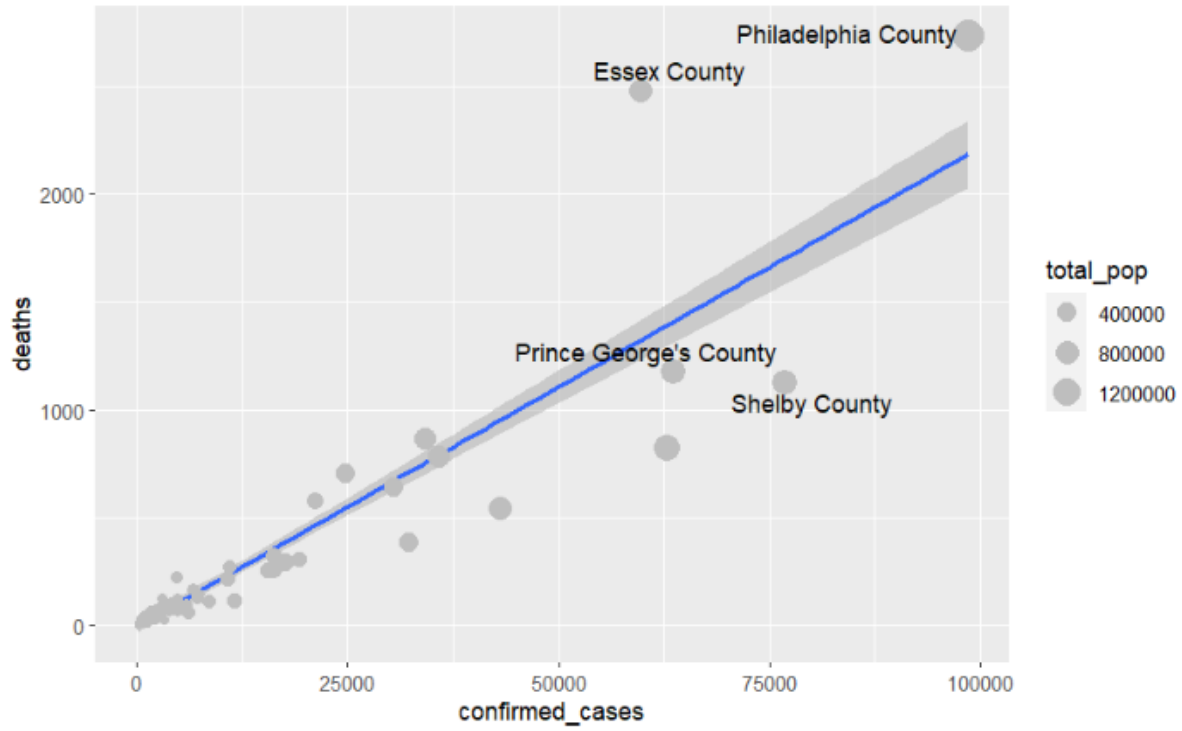
Data Understanding Table3: Top Ten Majority Black Counties COVID-19 Confirmed Cases Descending

county_name	state	confirmed_cases	deaths	total_pop	black_pop
Washington	DC	34259	861	672391	315159
Marengo	AL	2008	29	19743	10689
Lowndes	AL	1115	35	19743	7708
Greene	AL	762	23	8533	6851
Perry	AL	982	18	9680	6735
Hale	AL	1774	42	14995	8795
Macon	AL	1168	35	19358	15772

Barbour	AL	1738	36	26201	12528
Bullock	AL	997	28	10478	7925
Sumter	AL	895	26	13084	9256

Counties where Black people are the majority had outliers that experienced a higher-than-normal number of deaths per confirmed cases and per total population can be visualized from the scatter plot below. Here we can see two counties, the Philadelphia and Essex counties experienced significantly higher rates of deaths per confirmed cases and total population than other black-majority counties. While Shelby and Prince George's counties were outliers in the number of confirmed_cases but below the average number of deaths per total population.

Data Understanding Figure 1: Black Majority County deaths over confirmed cases graph



Black Majority Counties per case and 1000 people table:

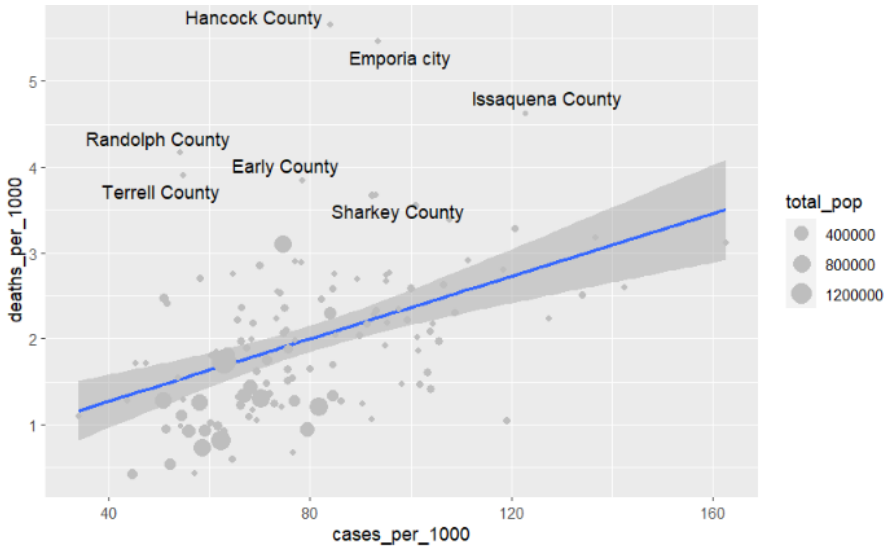
Data Understanding Table 4: Top Six Majority Black Counties COVID-19 deaths and Cases per 1000 Descending Confirmed Cases

state	county_name	confirmed_cases	deaths	total_pop	cases_per_1000	deaths_per_1000	death_per_case
PA	Philadelphia County	98479	2732	1569657	62.73	1.74	0.028
TN	Shelby County	76747	1128	937847	81.83	1.20	0.015
MD	Prince George's County	63592	1179	905161	70.25	1.30	0.019
GA	Fulton County	62813	824	1010420	62.17	0.81	0.013
NJ	Essex County	59727	2480	800401	74.62	3.10	0.042
GA	DeKalb County	43121	543	736066	58.58	0.74	0.013

Outliers in Black majority counties:

Outlier counties that experienced a higher than the normal number of deaths per 1000 people over cases per 1000 people and per total population can be visualized from the scatter plot below. We can see here several locations ranging from small to medium to large counties. Most of the counties are in the south like Virginia, Mississippi, Georgia, North Carolina but there are some Central US counties like Hancock in Ohio. Thus, the outliers in black-majority counties that experienced higher than normal deaths per case are Hancock Count, Emporia city, Issaquena County, Randolph County, Early County, Terrell Count, and Sharkey County.

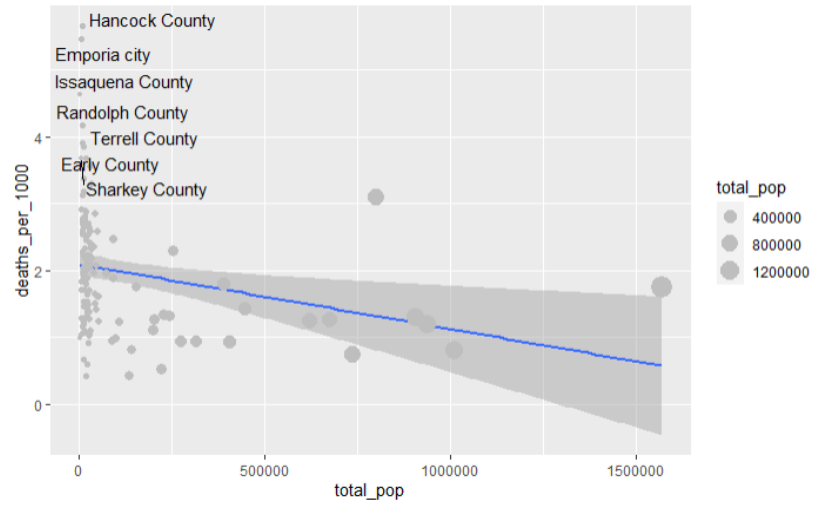
Data Understanding Figure 2: Black Majority Counties deaths per 1000/cases per 1000



How does death per case depend on the size of the population in the US majority Black Counties?

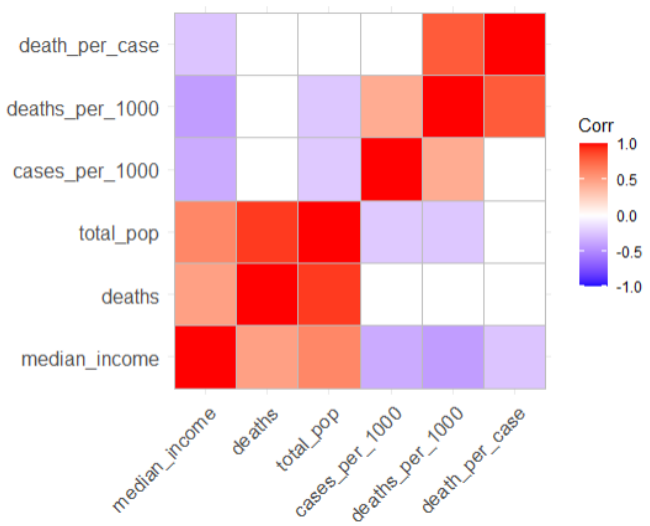
It can be observed that total population size if as a major factor in deaths per case. It can be seen here that Hancock County exhibited a very large number of deaths per case compared with other counties that have the same approximate population size. The is most likely a deeper story there like poorly equipped medical care resources or some other mitigating issue that would take a deeper investigation to find out.

Data Understanding Figure 3: deaths per 1000 over total population



Correlating variables in majority Black counties:

Data Understanding Figure 4: Majority Black Counties Heat Map



In the figure Data Understanding, Figure 4 Heatmap visualizes how total population and median income are correlated with deaths per 1000 and cases per 1000. It can be observed there is a strong correlation between deaths, total population and income. It can also be observed that a moderate correlation between cases per 1000 and total population and a stronger correlation

between deaths per 1000 and cases per 1000. Which would be expected. We can see that median income and total population play a significant role in the number of deaths per case.

2.1.3.2 US Counties Where the Hispanic Population is the Majority

The results show that the COVID-19 virus had a strong effect on the Hispanic Community with a large number of COVID-19-related deaths in Los Angeles. The Hispanic population is about approximately 48% of Los Angeles County so representation wise that would put the Hispanic number of COVID-19-related deaths approximately around 6749 COVID-19-related deaths.

Top 10 Majority Hispanic Counties confirmed cases (in descending order) and deaths:

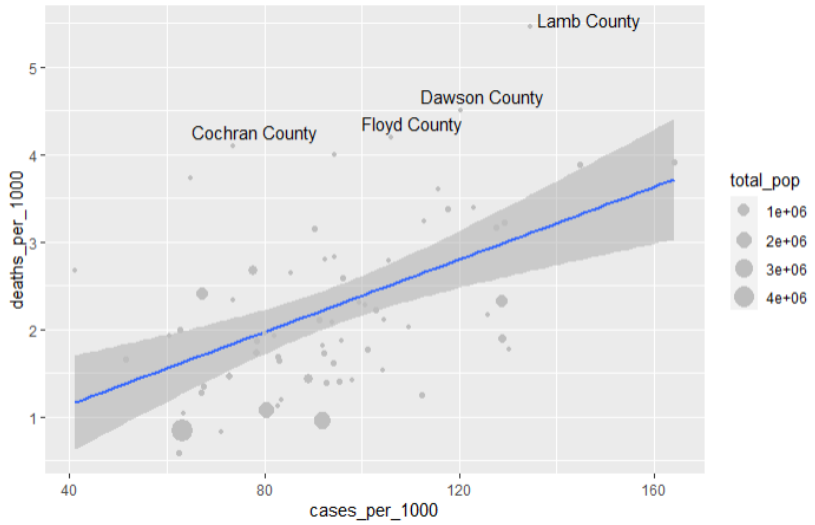
Data Understanding Table 5: Top Ten Majority Hispanic Counties COVID-19 Confirmed Cases Descending

county_name	state	confirmed_cases	deaths	total_pop	hispanic_pop
Los Angeles	CA	1002614	13936	10105722	4893579
Riverside	CA	244151	2517	2355002	1130033
San Bernardino	CA	252808	1560	2121220	1108996
Miami-Dade	FL	347965	4622	2702602	1823038
Hudson	NJ	56285	1792	679756	293465
Bronx	NY	102227	5275	1455846	810549
Queens	NY	154688	7871	2339280	654793
Hidalgo	TX	56455	2018	839539	770794

Cameron	TX	32698	1126	420201	375256
Bexar	TX	152231	2040	1892004	1130949

Outliers in Hispanic majority counties:

Data Understanding Figure 5: Hispanic Majority Counties deaths per 1000/cases per 1000



It can be observed *Data Understanding figure 5* that in the smaller Hispanic-majority counties that there were many cases per 1000 and deaths per 1000. In *Data Understanding Figure 5* Lamb County stands out as most dramatically impacted counties of these sparsely populated counties. The reason for this is most likely that it is such a rural county that it lacked medical resources. The same can be said about Dawson, Floyd, and Cochran County. It can be observed in *Data Understanding Figure 5* that there are some large cities reported, however, those larger cities show that are tackling the curve appropriately.

Hispanic Majority Counties per case and 1000 people table:

Data Understanding Table 6: Top Six Majority Hispanic Counties COVID-19 deaths and Cases per 1000 Descending Confirmed Cases

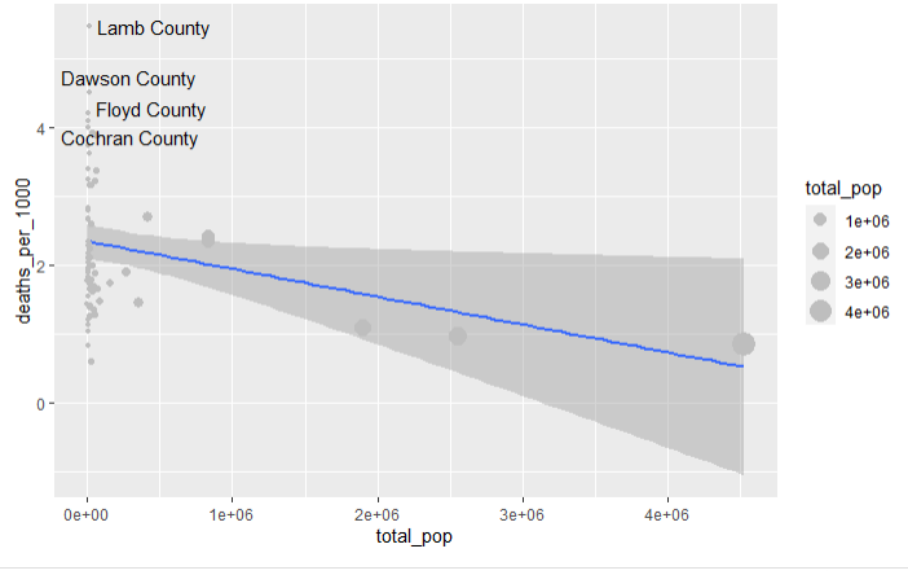
state	county_name	confirmed_cases	deaths	cases_per_1000	deaths_per_1000	deaths_per_case	total_pop
CA	Los Angeles	1002614	13936	99.21	1.38	0.014	4893579
FL	Miami-Dade	347965	4622	128.75	1.71	0.013	1823038
TX	Harris	286356	3825	63.28	0.84	0.013	1910535
CA	San Bernardino	252808	1560	119.18	0.74	0.006	1108996
CA	Riverside	244151	2517	103.67	1.07	0.01	1130033
TX	Dallas	234625	2453	91.93	0.96	0.01	1011374

A notable observation about the table Data Understanding Table 6 is that Miami-Dade which had the highest number of deaths per case per 1000 people was not as many deaths per case as Philadelphia County experienced a death per case rate of 1.74.

How does death per case depend on the size of the population in the US majority Hispanic Counties?

Once again, we can see here that sparsely populated counties were more dramatically affected than densely populated counties.

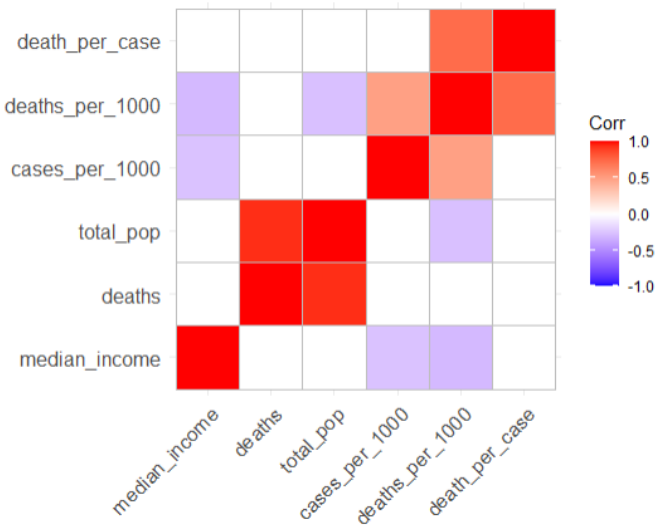
Data Understanding Figure 6: Hispanic Majority Counties deaths per 1000/cases per 1000



Correlating variables in majority Hispanic counties:

The Heatmap scene in Data Understanding Figure 7 indicates correlations with median income and case per 1000 and deaths per 1000 in Hispanic majority counties.

Data Understanding Figure 7: Majority Hispanic Counties Heat Map



2.1.3.3 US Counties Where the White Population is the Majority

There are a significantly larger number of White majority counties in the United States with 2,894 majority-white counties. The maximum number of deaths per county in this dataset was 8455 deaths in Cook County. This would be where Chicago is. Thus, what we see is typical for substantial largely populated counties. However, the total number of deaths was less than that of Los Angeles but that can be explained by the fact that more people live in Los Angeles than in Chicago. The White population makes up approximately 42% of Cook Count therefore by representation the approximate amount of COVID-19 related at approximately 3,646 deaths in Cook County's White population.

Top 10 Majority White Counties confirmed cases (in descending order) and deaths:

Data Understanding Table 7: Top Majority White Counties COVID-19 deaths and total population with Descending Confirmed Cases

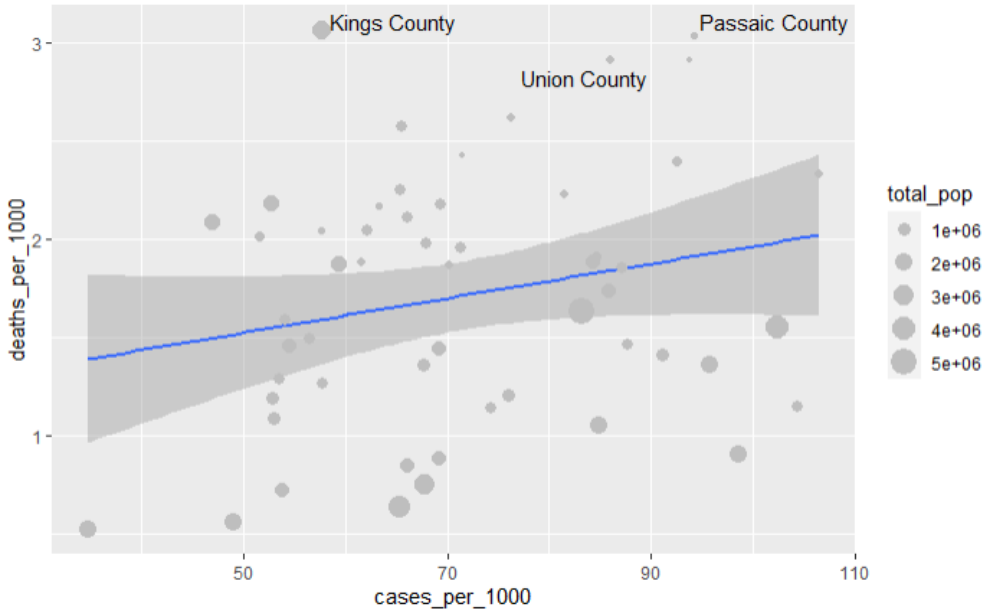
county_name	state	confirmed_cases	deaths	total_pop	white_pop
Cook	IL	435888	8544	5238541	2235598
Maricopa	AZ	425844	6443	4155501	2340105
San Diego	CA	214335	2103	3283665	1517153
Orange	CA	213866	2367	3155816	1306398
Clark	NV	202471	2873	2112436	931891
Tarrant	TX	195518	1798	1983675	959103
Broward	FL	160514	1990	1890416	721241
Kings	NY	151973	8073	2635121	947519
Suffolk	NY	128580	2603	1497595	1025705
Nassau	NY	114969	2563	1363069	836384

Outliers in White majority counties

Keeping with the trend we can see that sparsely populated counties still experience a higher number of deaths per Case when looking at *Data Understanding Figure 8*. However,

there is a unique outlier with Kings County being a more densely populated county experiencing a substantial number of deaths per cases yet their cases per 1000 is still relatively low.

Data Understanding Figure 8: Hispanic Majority Counties deaths per 1000/cases per 1000



Observe the number of deaths per 1000 in Data Understanding Table 8, we can see the White population experienced fewer deaths per 1000 than the Black and Hispanic majority counties.

White Majority Counties per case and 1000 people table:

Data Understanding Table 8: Top Six Majority White Counties COVID-19 deaths and Cases per 1000 Descending Confirmed Cases

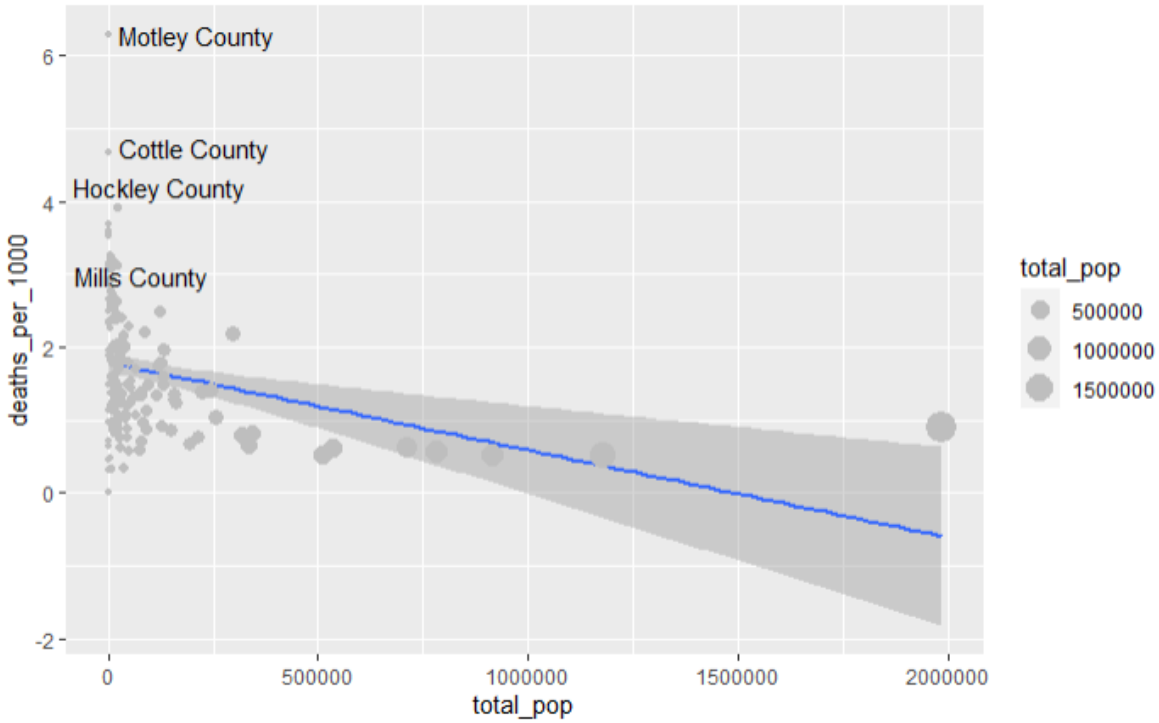
state	county_name	confirmed_cases	deaths	total_pop	cases_per_1000	deaths_per_1000	death_per_case
IL	Cook County	435888	8544	5238541	83.21	1.63	0.0196
AZ	Maricopa County	425844	6443	4155501	102.48	1.55	0.0151

CA	San Diego County	214335	2103	3283665	65.27	0.64	0.0098
CA	Orange County	213866	2367	3155816	67.77	0.75	0.0111
NV	Clark County	202471	2873	2112436	95.85	1.36	0.0141
TX	Tarrant County	195518	1798	1983675	98.56	0.91	0.0092

How does death per case depend on the size of the population in the US majority White Counties in Texas? *The reason this is solely looking at Texas is because the number of counties could overload the visualization if it took into account all the White majority counties across the US.

Observed in the figure *Data Understanding Figure 9*, it can be seen the same trend of sparsely populated counties experiencing a high number of deaths per case. Motley County stands out in this group of counties with under 50,000 people living there. Again, these details are most likely due to a medical resource shortage.

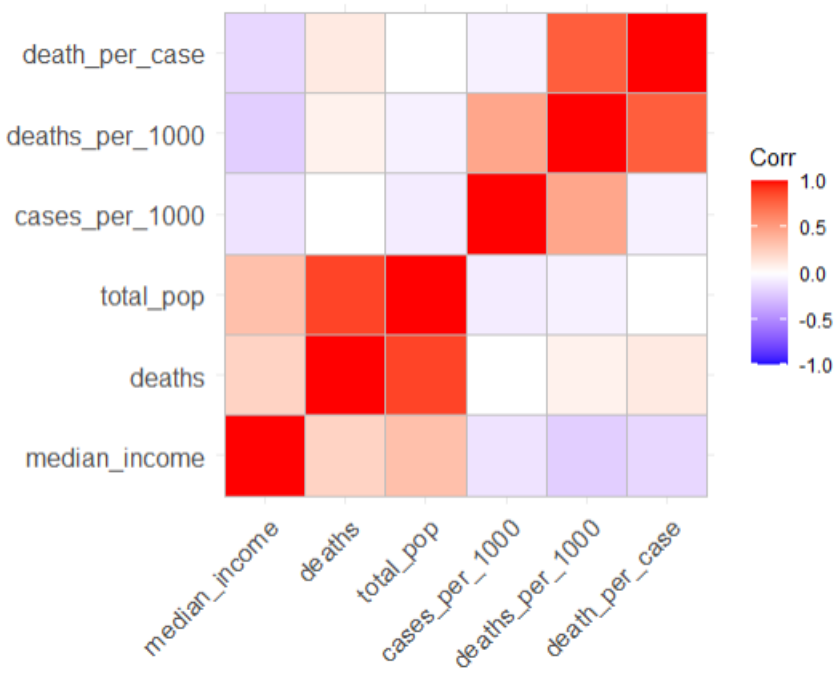
Data Understanding Figure 9: White Majority Counties deaths per 1000/cases per 1000



Correlating variables in majority White counties:

Observed in figure Data Understanding Figure 10, can be seen the same type of correlations per median income and deaths per case. There a strong correlation between total populations and deaths amongst the majority White counties. These are the same correlations that were seen when comparing the Majority Black and Majority Hispanic populations.

Data Understanding Figure 10: Majority White Counties Heat Map



2.1.4 Selecting Important Features for COVID-19 with Census Data

When selecting features of the COVID-19 with Census data set we want to capture a well- rounded number of features, however, we also want to focus on societal stressors could exacerbate an already embattled community that experience more poverty and disparity when it comes to gender, wealth, and race. The below table Data Understanding Table 9 indicate the selected the features that we will use in our predictions of confirmed cases and deaths once we start implementing a solution for tracking vulnerable areas in the United States that could be more adversely affected by another COVID-19 like outbreak that happened during 2020.

Data Understanding Table 9: Feature Selection of the 259 features

Feature	Scale of Measurement	Description
county_name	Nominal	Name of US county
state	Nominal	Name of US state

confirmed_cases	Ratio	Number of positive results from a COVID-19
deaths	Ratio	Number of covid related deaths per county
total_pop	Ratio	Total population per county
white_pop	Ratio	Total white population per county
black_pop	Ratio	Total black population per county
asian_pop	Ratio	Total Asian population per county
amerindian_pop	Ratio	Total American Indian population per county
hispanic_pop	Ratio	Total Hispanic population per county
income_under_40	Ratio	The count of population with income under $40,000.
households_public_asst_or_food_stamps	Ratio	The count of population on assistance per county.
children_in_single_female_hh	Ratio	The number of single female mother households.
children	Ratio	The count of children per county.
employed_pop	Ratio	The count of employed per county.
unemployed_pop	Ratio	The count of unemployed per county.
pop_determined_poverty_status	Ratio	The portion of population considered in poverty.
not_us_citizen_pop	Ratio	The count of non-US citizens per county.

two_parent_families_with_young_children	Ratio	The count of two parent household families in county.
poverty	Ratio	The count of population living in poverty per county.

Due to complications of dealing with the size restraint of using the Random Tree Forest algorithm in R, we did not programmatically select features. The issue encountered with programmatically selecting features with R indicated that there were too many features in the dataset to perform the operation. Due to time constraints, another technique for selecting features was used. That is, we selected the features manually, and since we are focusing on social factors that would affect a region's response to COVID-19 we selected features that pertain to race, income, total population size, median income, number of people enrolled in an assistance programs, number of people who group commute together like carpooling or riding a bus, and etc...In order to keep well rounded data we extract some pertinent features in regards to k-12 students, income and commuter demographic.

Selected Features from the COVID-19 with Census Data Dataset plus extracted features table:

Extracted features from COVID-19 with Census Dataset Table:

Data Understanding Table 10: Extracted features from the original 259.

Feature	Scale of Measurement	Description
income_over_40_under_100	Ratio	The count of population income between $40,000 and $100,000.
income_over_100	Ratio	The count of population with income over $100,000.

income_200000_or_more	Ratio	The count of population with income over $200,000.
commuters_in_groups	Ratio	The count of commuters that commute in groups.
local_commuters	Ratio	The count of commuters that commute locally.
out_of_town_commuters	Ratio	The count of commuters that commute from out of town.
cases_per_1000	Ratio	The ratio of confirmed COVID-19 cases per 1000 people in county.
deaths_per_1000	Ratio	The ratio of deaths per 1000 people in county.
death_per_case	Ratio	The ratio of deaths over cases.
K12_students	Ratio	The count of K-12 students.

2.1.4 Summary COVID-19 with Census Data

The COVID-19 with Census Dataset was a very rich and robust dataset. Using the Census data in conjunction with COVID-19 case data provides an instrument to drill down into the demographic data contained to find insights and patterns as it pertains to the spread and treatment of COVID-19. Some of the remarkable correlations we noticed were how remarkably different the spread and treatment of the COVID-19 virus affected the 3 major US races, Blacks, Whites, and Hispanics. We also see how median income affected the ratio of deaths per COVID-19 cases, and how although total population did show a correlation between an increase in the number of confirmed cases and deaths with an increase in population. It was actually the smaller communities that reflected more deaths. Of course, this is in part due to the large number of counties with small populations but also because of the lack of resources to treat COVID-19 in the more rural counties. In closing, there are many aspects that can be looked into using this dataset from looking at how COVID-19 affected families with children, single-parent families, counties with a large number of children, commuters, and renters, are all aspects that can be more finely viewed in conjunction with the COVID-19 data to expose trends and relationships

between the amount of confirmed COVID-19 cases and deaths that resulted from the COVID-19 virus.

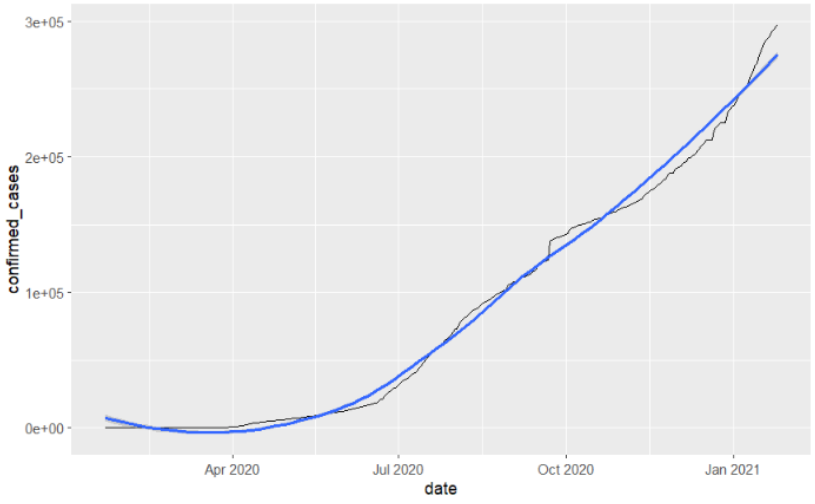
2.2 Texas COVID-19 Cases Data Description

The cases_tx.csv dataset provides a day-to-day report of reported COVID-19 related confirmed cases and deaths starting from January 22nd, 2020, until January 25th, 2021. This day-to-day report provides a means to visualize the rate at which COVID-19 spread through Texas's many various counties. When used in conjunction with the COVID-19 Cases with Census Data Set it can help us understand how the different Texas demographics experienced the spread of COVID.

In the graph figures below, *Data Understanding Figure 11*, and *Data Understanding Figure 12*, that the White and Hispanic majority counties experienced the spread of COVID-19 differently. We can see that COVID-19 spread in the majority of Hispanic county of Harris County at a fast and almost linear rate. We can see that in the White majority county, Tarrant county that there is a more gradual rise in the spread of COVID-19. My thoughts on this is that according to the US Bureau of Labor Statistics that Hispanics are more hat Hispanics are more likely to have a labor force job which would mean they did not start work for home schedule or some other alternative to avoid COVID-19, instead the Hispanic population stayed on the frontline and kept going into work [7].

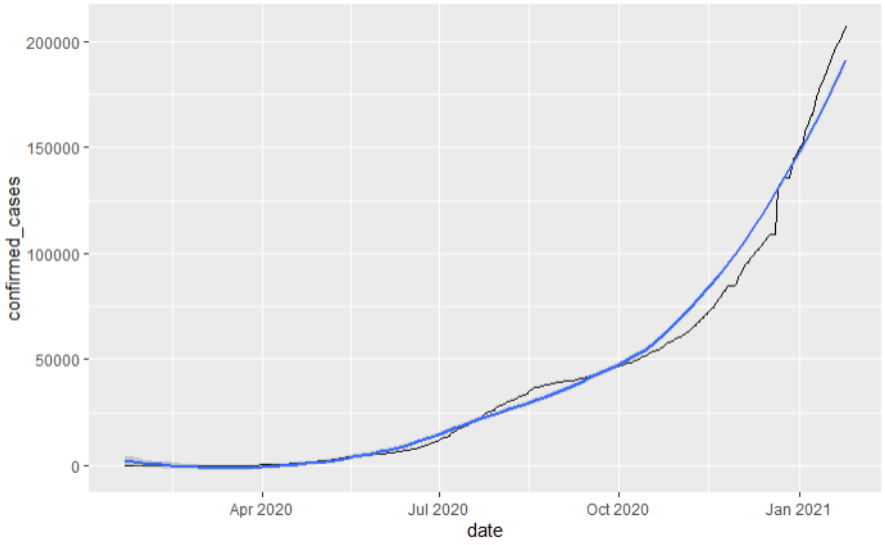
Spread of COVID-19 in Hispanic majority Harris County:

Data Understanding Figure 11: Cases Texas Data Spread of COVID-19 2020 Harris County



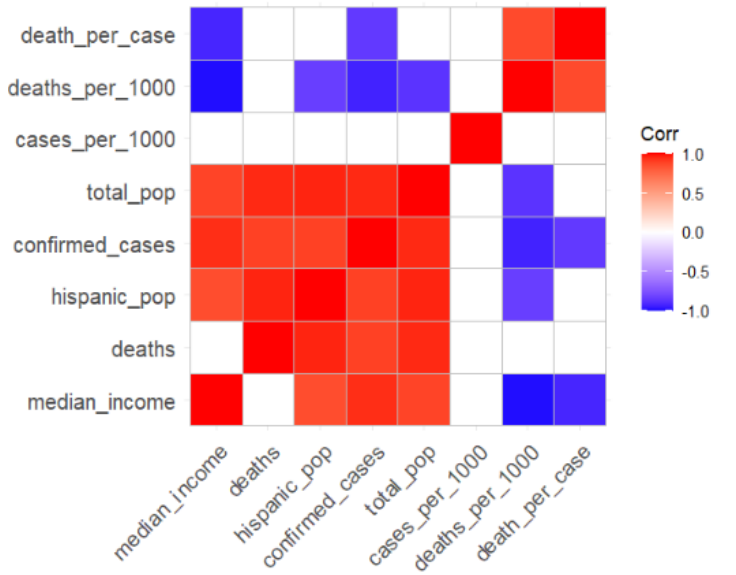
Spread of COVID-19 in White majority Tarrant County:

Data Understanding Figure 12: Cases Texas Data Spread of COVID-19 2020 Tarrant County



Hispanic Majority Texas Cities Heatmap:

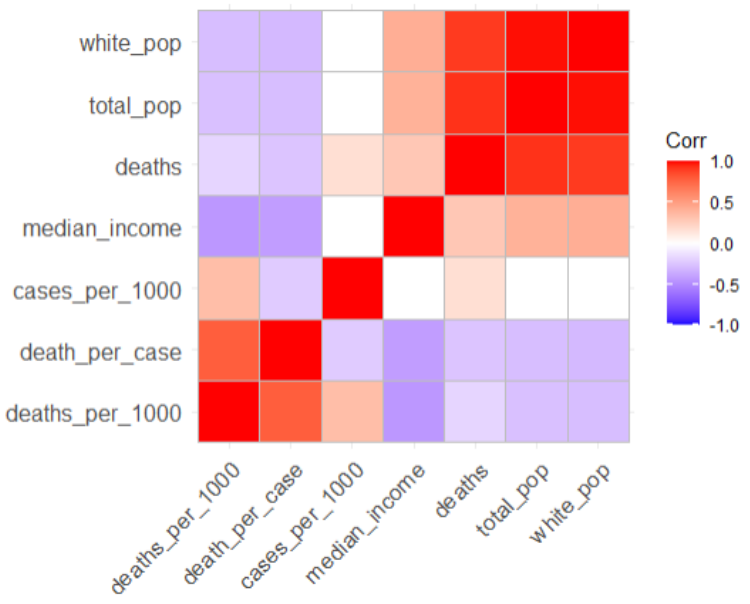
Data Understanding Figure 13: Majority Hispanic Texas Counties Heat Map



In the heatmap *Data Understanding Figure 13,* it can be observed a strong - negative correlations between median income and deaths per case and case per 1000. There is a strong positive correlation of over 50% between the features total population and confirmed cases and deaths in this population. There is also a very strong positive correlation between median income and confirmed cases and total and a moderate positive correlation with median income and total population. Finally, the blank areas of no correlations between deaths and deaths per case, deaths per 1000, and cases per 1000 indicate that there may be some issues with this dataset. Cases per 1000 should at least correlate with cases per 1000.

White Majority Texas Cities Heatmap:

Data Understanding Figure 14: Majority White Texas Counties Heat Map



It can be observed in the heat map *Data Understanding Figure 14*, that there is only a weak 50% negative correlation for deaths per case and total population and deaths per 1000 and total population. Significantly less than what was seen in the previous heatmap for Hispanic-majority counties. We see a strong negative correlation still with median income and deaths per case, and also median income and case per 1000. However, there is a moderate positive correlation with median income and deaths which is interesting because the previous heatmap showed no correlations at all between those

two features in the majority Texas Hispanic populations.

2.2.2 Verifying Texas COVID-19 Cases Data Quality

The Texas COVID-19 dataset contained 7 features and 94350 observations. There were no duplicate columns or rows. However, there was a substantial amount of duplicate data since the report focuses on daily confirmed cases and deaths. Also, there were no NANs found in this dataset. There were two irrelevant features for our project in this dataset and those were the county_fips_code and state_fips_code. These features could be removed for computing efficiency and in order to get the correct output we need to avoid nominal values that could lead to mistakes by using an ID number in a dataset meant to generate predictions.

Appropriate Statistics about import features from Cases Texas dataset:

Data Understanding Table 11: Summary Statistics from the Cases Texas dataset

	Min	Median	Mean	Max
Confirmed Cases	0.0	82.0	2158.6	4024
Date	2020-01-22	2020-07-24	2020-07-24	2021-01-25
Deaths	0.0	2.0	38.19	4024.00

2.2.3 Selecting Important Features for Cases Texas Dataset

The goal here was to select features that we can use to extrapolate meaningful information when used in conjunction with the COVID-19 census data report and the global mobility report.

Selected features from the Cases Texas Data:

Data Understanding Table 12: Selected features from the Cases Texas dataset

Feature name	Scale of Measurement	Description
county_name	Nominal	The name of Texas county
date	Interval	The date cases were reported in Texas by county
deaths	Ratio	The amount of daily reported deaths due to COVID-19 per Texas County
confirmed_cases	Ratio	The amount of daily reported COVID-19 cases in Texas

2.2.4 Summary for Cases Texas Data Set

In summary, this dataset benefits our pursuit to understand the spread of COVID-19 in the state of Texas by viewing the spread of the COVID-19 virus across Texas counties over the time interval of January 21st, 2021, to January 22, 2020.

2.3 Google's Global Mobility Report

This dataset was created by Google using aggregated and anonymous reports from public health officials around the world in conjunction with Google Maps. Global Mobility Report shows how communities around the world are moving around differently due to the outbreak of the COVID-19 virus. The goal of the report was to provide clarity into what has changed in response to policies aimed at fighting COVID-19. The report charts the movement trends over time by geography and across certain spectrums of what we call daily life like retail shopping, enjoying recreation, grocery and pharmacy shopping, and visiting parks, transit stations, workplaces, and residential locations.

2.3.2 Verifying the data of the Global Mobility Report:

The Global Mobility Report data consisted of 14 features and over 3 million rows of data. There was missing data throughout the entire dataset, yet there were no columns that were entirely NA. There were some irrelevant attributes like census_fips_code and country_region_code, and iso_3166_2_code, as the features serve no purpose for our project. Also, the features country_region and metro_area, were removed because they were unneeded since we can derive the same information from the features sub_region_1 and sub_region_2.

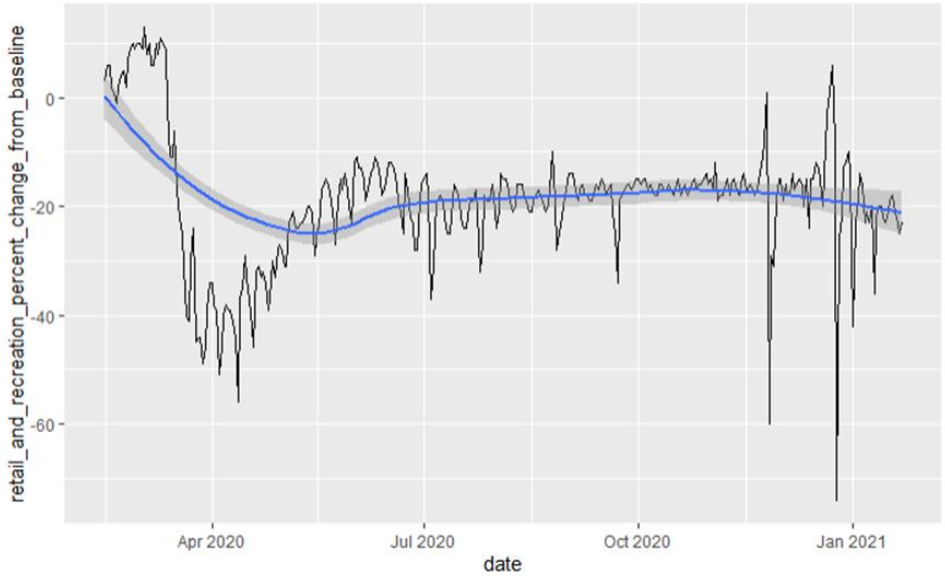
Appropriate Statistics about import features from Global Mobility Report's dataset:

Data Understanding Table 13: Appropriate Statistics from the Global Mobility Report dataset

	Min	Median	Mean	Max
Confirmed Cases	0.0	82.0	2158.6	4024
Date	2020-01-22	2020-07-24	2020-07-24	2021-01-25
Deaths	0.0	2.0	38.19	4024.00

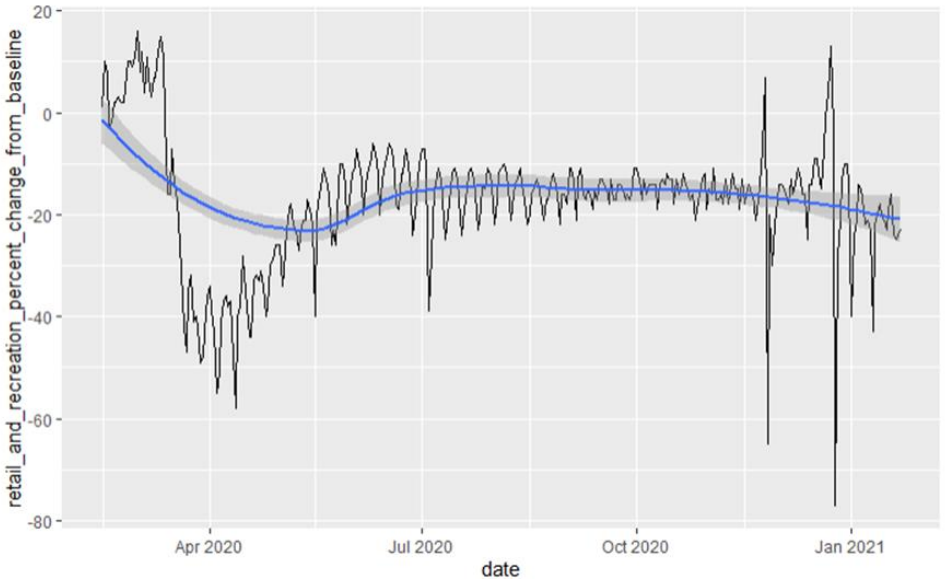
Mobility in majority Hispanic Texas County Harris County:

Data Understanding Figure 15: Loss of Mobility 2020 Harris County - Global Mobility Report



Mobility in white majority Tarrant County in Texas:

Data Understanding Figure 16: Loss of Mobility 2020 Tarrant County - Global Mobility Report



2.3.3 Summary for Global Mobility Report

In summary, the dataset contained data on the level of mobility that occurred until January 01, 2023. This data which is broken down by region, reported the change from baseline statistics for lifestyle attributes like going shopping, to the park, grocery shopping, transiting, coming back and forth from work, and leaving the house.

2.4 What is the trend in number of confirmed/deaths COVID-19 cases in different TX counties?

In this Section, we aim to find a trend in number of confirmed/deaths COVID-19 cases in different TX counties. At first, we want to investigate which county has the highest number of confirmed/death COVID-19 cases. Here we decide to use horizontal bar chart. A horizontal bar chart allows us to represent each state on the y-axis and the number of confirmed cases on the x-axis. Each state is treated as a separate category, and the height of the bar represents the number of cases. This makes it easy to compare the number of cases across states.

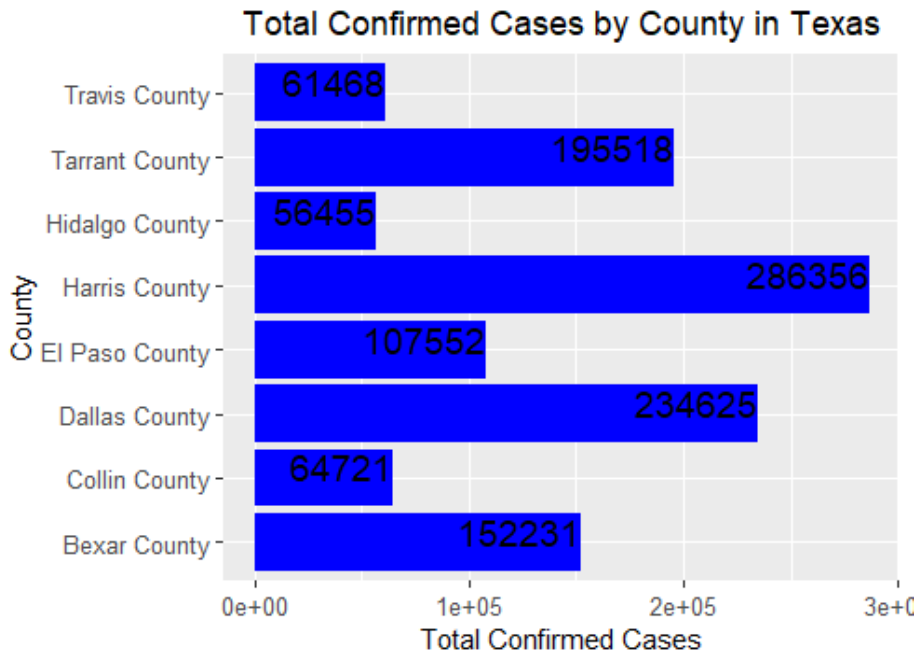


Figure 1 Total number of confirmed cases based on the counties in Texas, the values greater than 500K have been chosen for illustration purpose

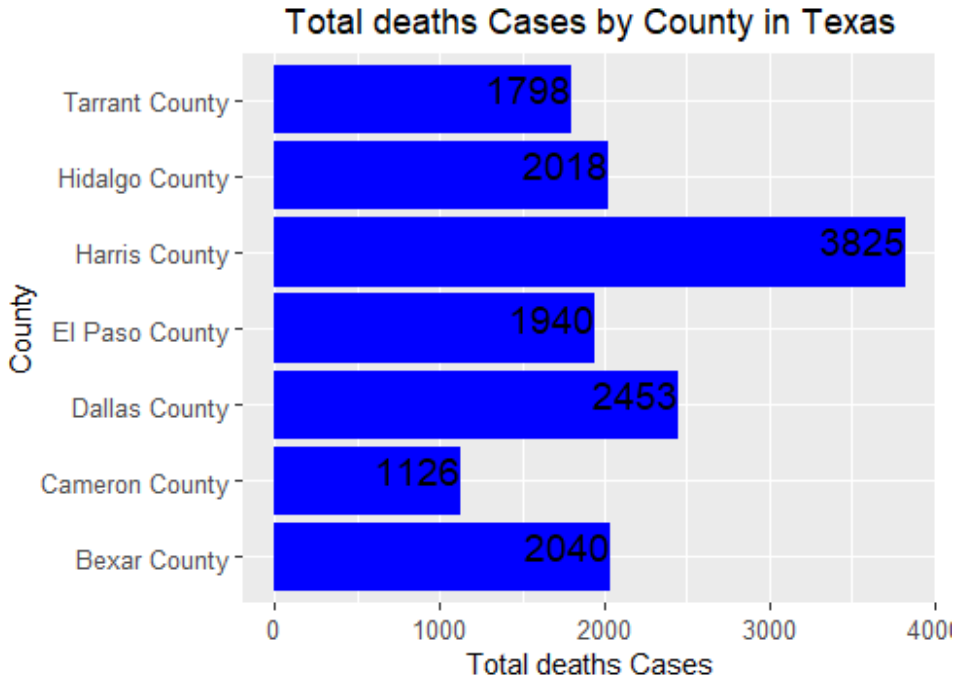


Figure 2 Total deaths cases based on county in Texas, the values greater than 1000 have been chosen for illustration purposes

As it can been seen in the Figure 1 and 2, Harris and Dallas counties have the most number of cases and deaths among all counties in Texas. As Harris and Dallas are the most populous counties in Texas, we conducted an experiment to assess the reliability of our results. We decide to examine the trend in the number of confirmed COVID-19 cases based on population in Texas counties using census data. This data represents the rate of spreading COVID-19 in each Texas county per 1000 people on the specific date of January 19th, 2021. For this experiment, we decide to use geospatial mapping visualization for several compelling reasons. It employs a precise and intuitive color-coding system, rendering it easily comprehensible. With counties color-coded according to their respective COVID-19 case counts, it offers a clear visual insight into the data. This representation keeps the description concise, focusing on the essential information without undue complexity, ensuring it remains user-friendly and accessible. Beyond its visual appeal, the figure aids in analysis and comprehension by enabling straightforward comparisons between counties, thereby enhancing the viewer's understanding of the data.

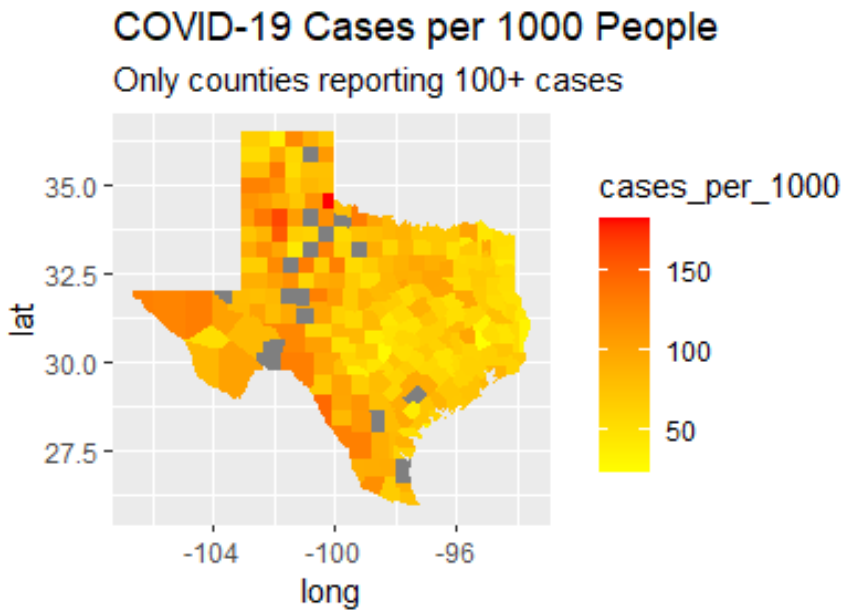


Figure 3 The rate of confirmed COVID-19 cases based on population in each Texas county per 1000 people on the specific date of January 19th, 2021

In figure 3, we can observe that counties with over 50,000 cases are represented in yellow, those exceeding 100,000 in orange, and those surpassing 150,000 in red. Counties with fewer than 1,000 confirmed cases are displayed in gray. Consequently, it becomes evident that nearly all

Texas counties are grappling with a significant number of COVID-19 cases. However, the new results indicate that Childress and Hale counties now rank among the top counties. In my opinion, this new metric also reflects the rate of COVID-19 spread, which has been exceptionally high. In contrast, some counties like Loving and San Jacinto have experienced an almost negligible rate of confirmed COVID-19 cases based on the population.

In the same manner we can compare the deaths based on population trends across the Texas counties.

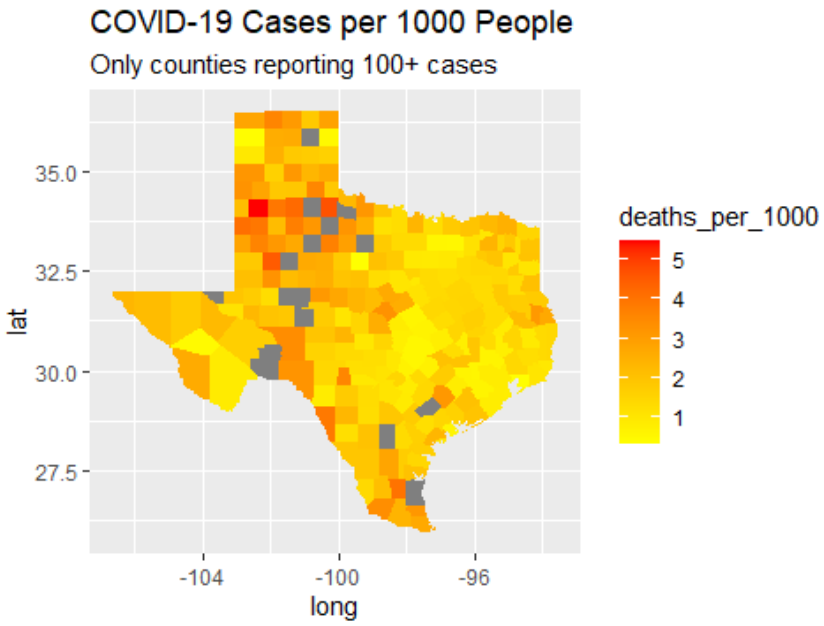


Figure 4 The rate of COVID-19 death cases in each Texas county per 1000 people on the specific date of January 19th, 2021

The figure 4 shows that Motley, Lamb and Cottle counties have the highest deaths per population rates and Borden, Loving and King counties don't have deaths at all. The database in the Table. 3 shows that Harris, Dallas counties, with the highest populations, experienced near to average rates of COVID-19 spreading and lower than the average deaths. This result highlights the importance of normalization in data and implies that these counties, likely have more healthcare facilities due to their larger populations, or residents may better adhere to social distancing rules.

County name	population	cases_per_population	average_cases	max_cases	deaths_per_population	average_deaths_per_population	Max_deaths_per_population
Borden County	602	2.99	7.80	18.29	0.00	0.19	0.63
Cottle County	1498	11.75	7.80	18.29	0.47	0.19	0.63
Dallas County	2552213	9.19	7.80	18.29	0.10	0.19	0.63
Harris County	4525519	6.33	7.80	18.29	0.08	0.19	0.63
King County	289	3.81	7.80	18.29	0.00	0.19	0.63
Lamb County	13368	13.46	7.80	18.29	0.55	0.19	0.63
Loving County	74	1.35	7.80	18.29	0.00	0.19	0.63
Motley County	1114	7.09	7.80	18.29	0.63	0.19	0.63

Table 3 The database is used for comparing the number of confirmed and death cases due to COVID-19 spread.

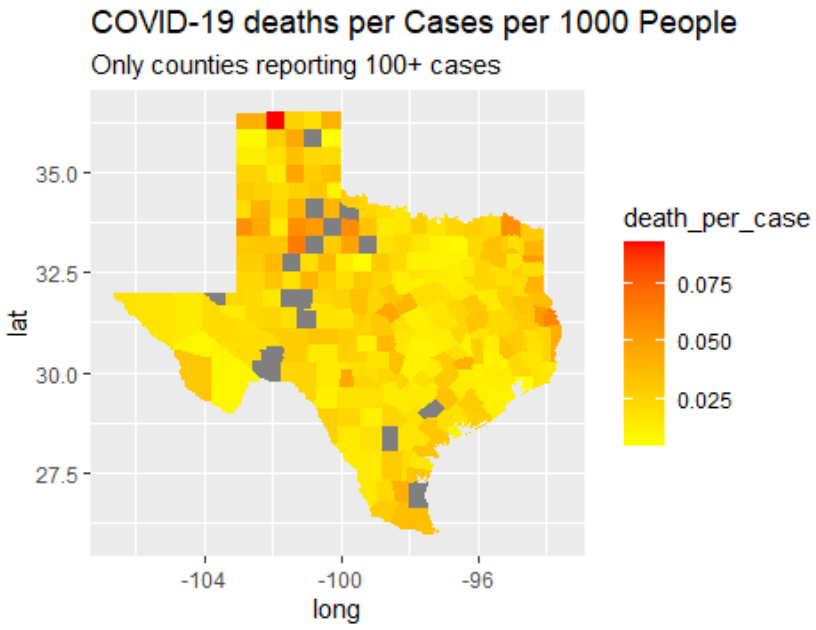


Figure 5 The rate of deaths per confirmed COVID-19 cases in each Texas county per 1000 people on the specific date of January 19th, 2021.

This Fig. 5 shows that Sherman County has the highest death rate per confirmed case, implying inadequate healthcare facilities and less stringent adherence to social distancing rules in the county. In order to ascertain the effectiveness of social distancing measures in Harris and Dallas counties and their impact on flattening the curve, we are conducting an analysis of mobility data and will compare the results.

Figures 6 and 7 display the number of confirmed cases on a daily basis in Dallas and Harris counties, respectively. Holidays are indicated with blue labels, and we have marked January 19th in red for reference. We can observe that number of peaks in Harris County is higher than Dallas County and overally, it seems that Dallas County has a smaller number of cases than Harris County. We can observe that, in Dallas County, confirmed cases are mostly below the baseline, but there is an increase in the number of cases before or after holidays (see Christmas holiday as an example). Figure 7 also illustrates that, during the observed period, the number of confirmed cases fluctuates around the baseline.

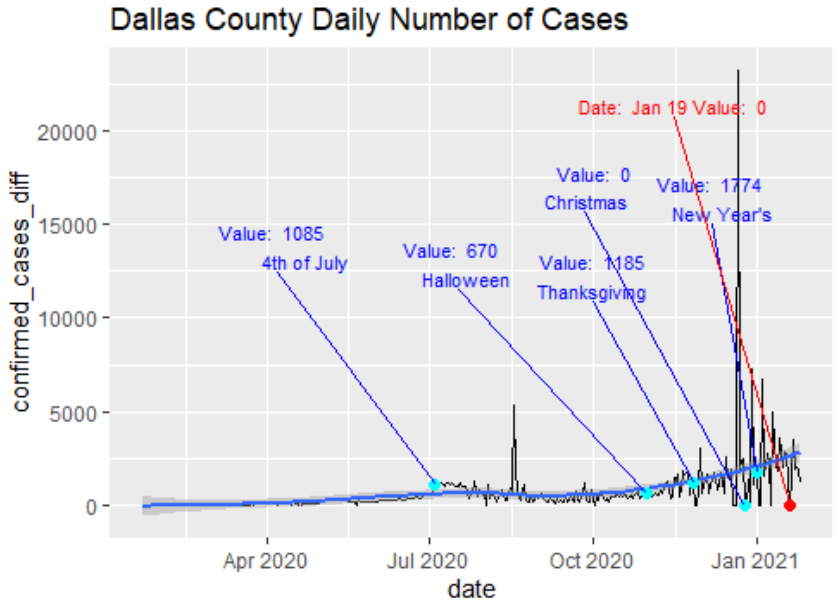


Figure 6 Daily number of confirmed cases in Dallas county

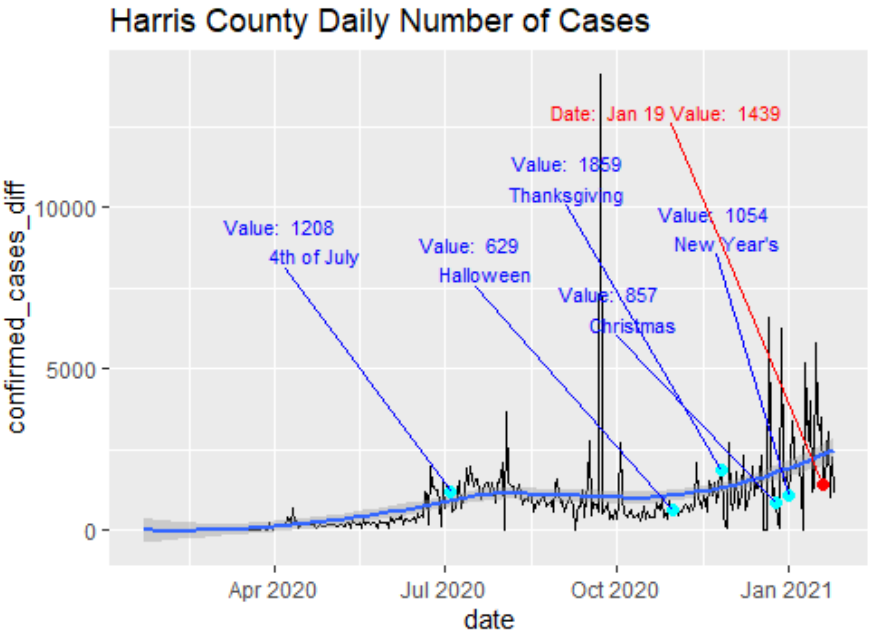


Figure 7 Daily number of cases in Harris County

In Figure 8 to 13, We can observe that some social distancing rules have contributed to flattening the curves. We can observe some falling peaks in holidays and short growths around holidays that it implies that people are moving or interacting more during those periods.

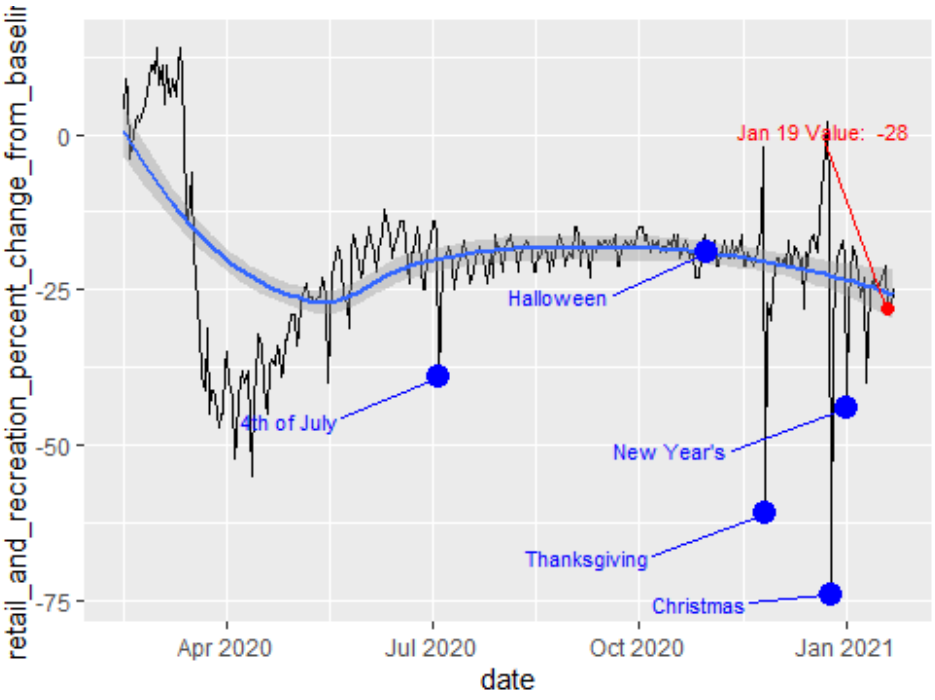


Figure 8 Dallas retails and recreation

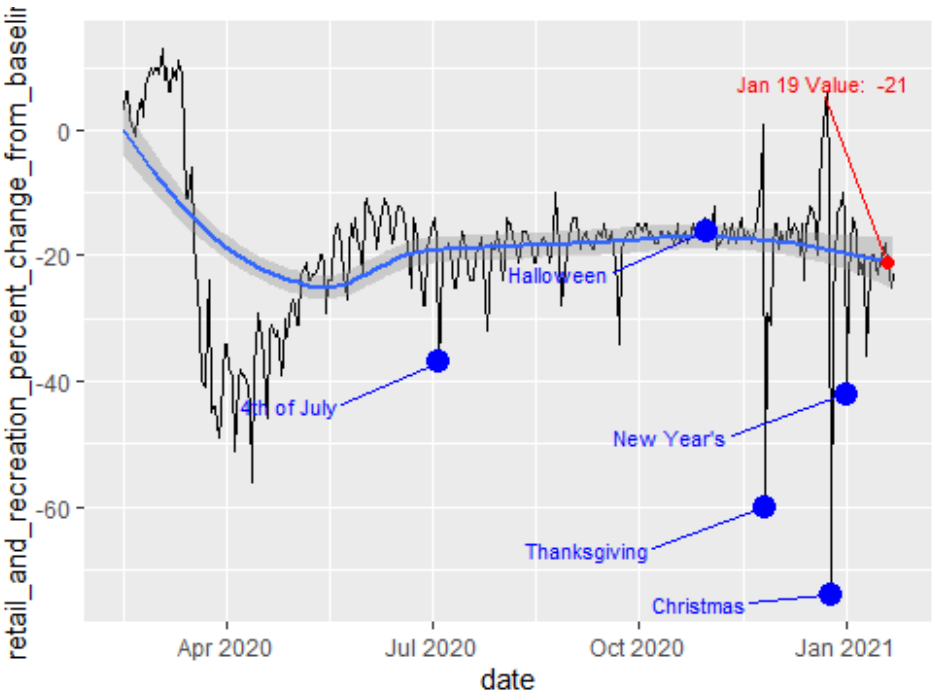


Figure 9 Harris retails and recreation

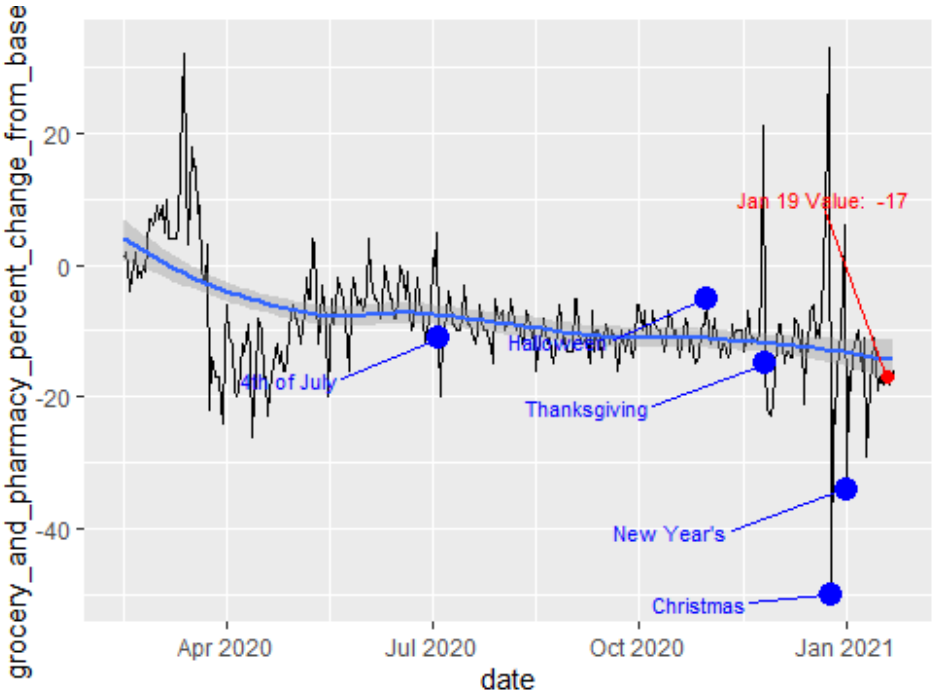


Figure 10 Dallas grocery and pharmacy

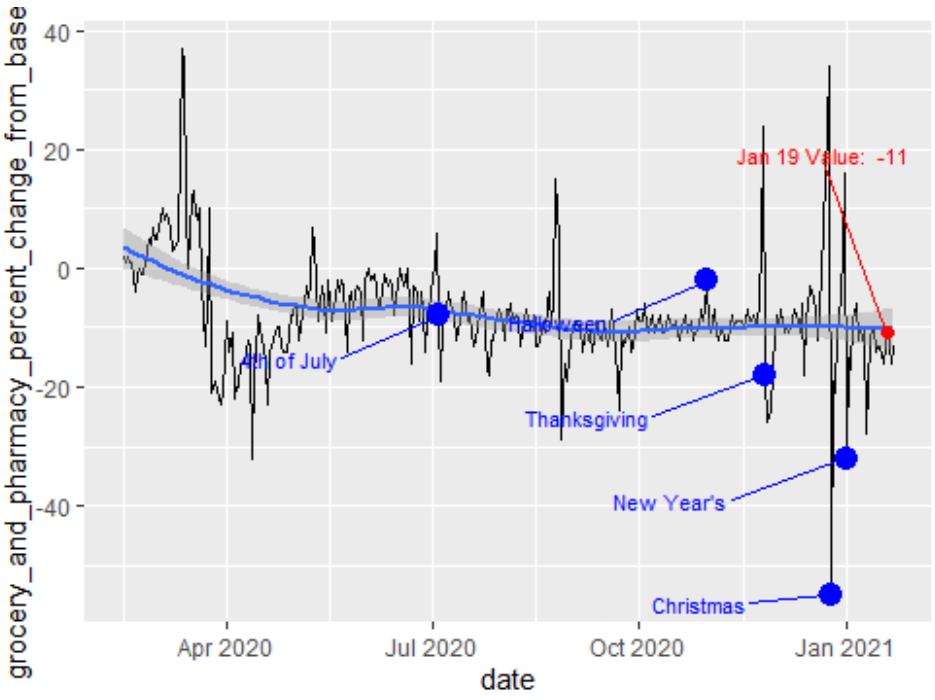


Figure 10 Harris grocery and pharmacy

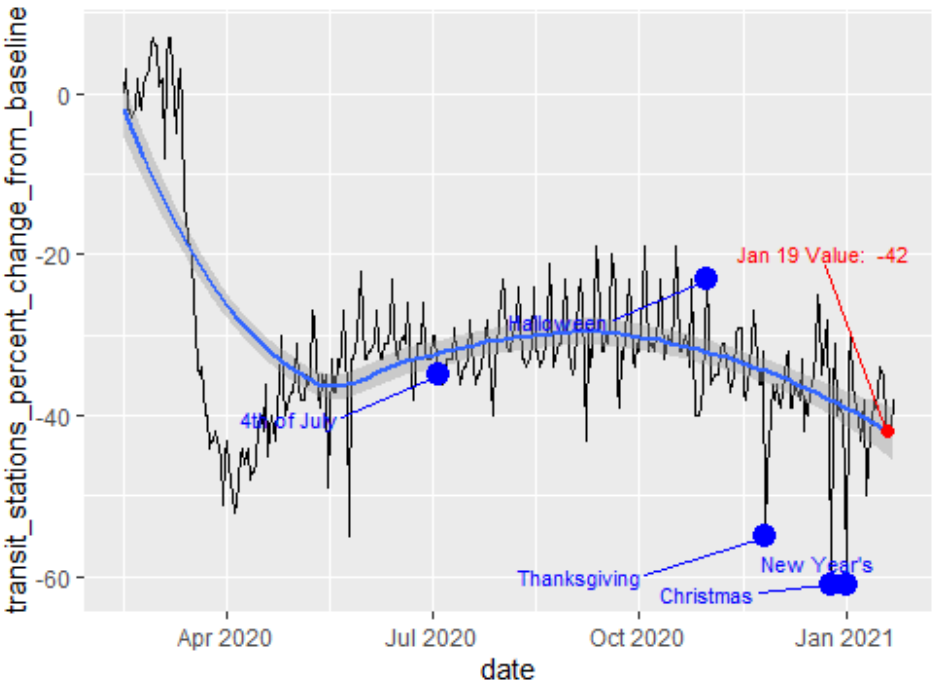


Figure 11 Dallas transit stations

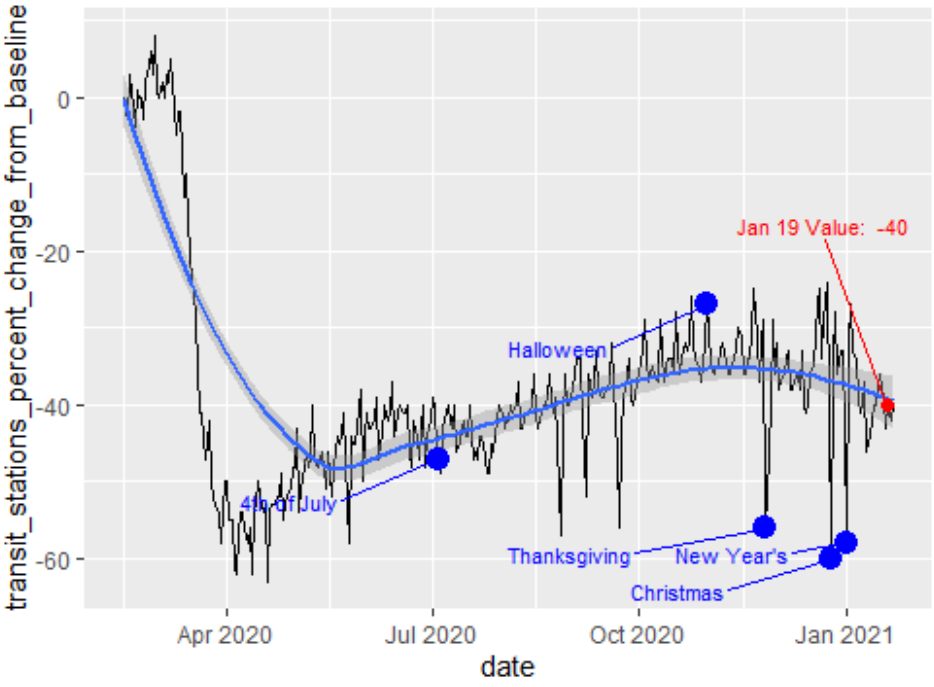


Figure 11 Harris transit stations

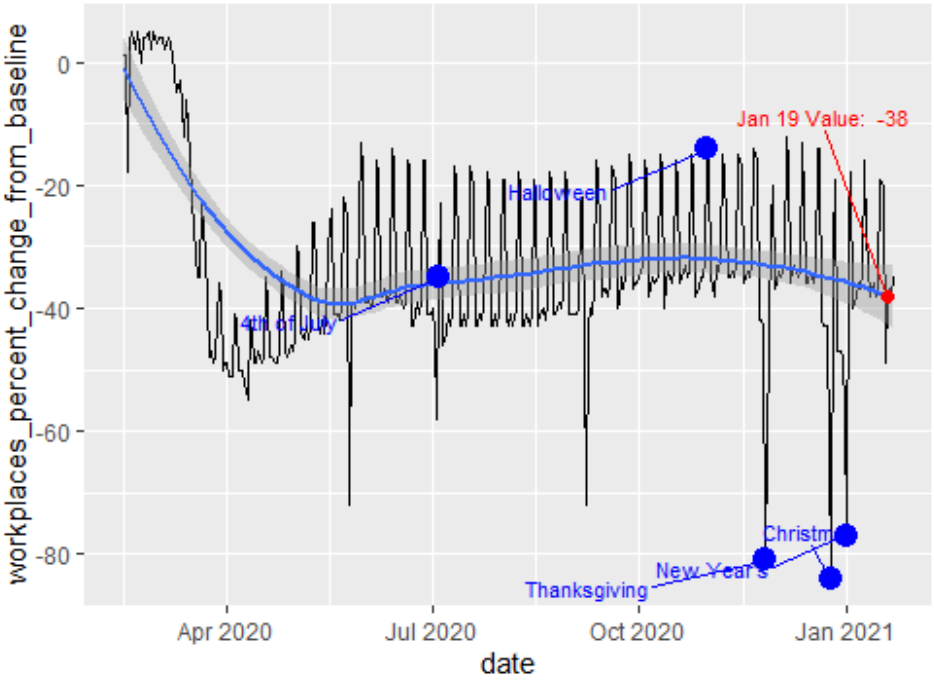


Figure 12 Dallas workplace

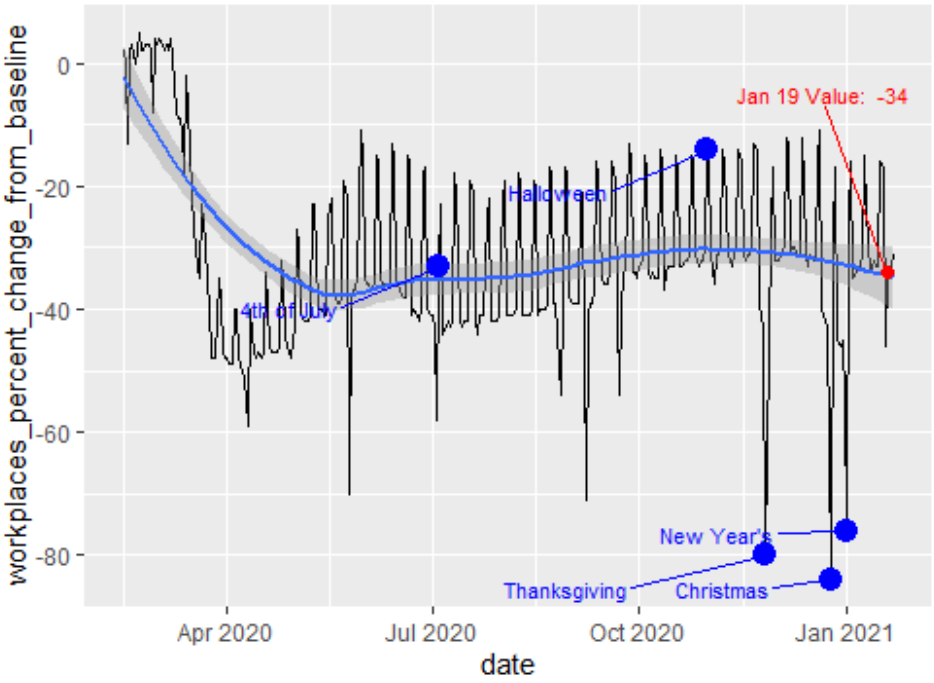


Figure 13 Harris workplace

CHAPTER 3: Data Preparation

To create the dataset, We first merged the two datasets, COVID-19_cases_plus_census and COVID-19_cases_TX, on the county_name column. This ensured that each row in the resulting dataset represented a single county, with data from both datasets. Next, we removed any rows from the dataset that were missing values for any of the four columns: county_name, First_reported_case, population, or income. This ensured that the dataset was complete and could be used for analysis. Finally, you saved the dataset to a CSV file, which is a common format for storing data. This makes the dataset easy to share with others and to use with different software programs is mentioned in this Report below.

Table 1: First Reportded Covid Case Reported

county_name	First_reported_case	population	income	confirmed_cases	deaths
Anderson County	04/01/2020	57747	17466	5575	75
Andrews County	04/04/2020	17577	29903	1606	37
Angelina County	03/26/2020	87700	21974	6765	193
Aransas County	04/05/2020	24832	29999	895	26
Archer County	05/15/2020	8793	31103	694	10
Armstrong County	04/14/2020	1929	31219	128	6
Atascosa County	03/25/2020	48139	23973	3781	90
Austin County	03/25/2020	29292	30101	1404	18
Bailey County	05/05/2020	7098	18662	742	15
Bandera County	04/09/2020	21316	29177	820	20
Bastrop County	03/25/2020	80306	25172	4373	57
Baylor County	06/05/2020	3602	30820	274	9
Bee County	04/08/2020	32729	17504	3082	53

Bell County	03/13/2020	336506	25017	16512	214
Bexar County	03/14/2020	1892004	26158	152231	2040
Blanco County	03/24/2020	11089	31249	318	11
Borden County	09/23/2020	602	38923	18	0
Bosque County	04/15/2020	17955	25763	988	20
Bowie County	03/18/2020	93635	24761	4876	139
Brazoria County	03/14/2020	345995	32343	26046	277
Brazos County	03/18/2020	214231	25337	16634	162
Brewster County	05/01/2020	9220	26073	945	8
Briscoe County	05/02/2020	1591	23199	114	3
Brooks County	04/22/2020	7251	13549	685	29
Brown County	03/21/2020	37787	24040	3382	81
Burleson County	03/28/2020	17596	27112	1093	22
Burnet County	03/24/2020	45017	29247	2196	34
Caldwell County	03/28/2020	40544	23366	2738	55
Calhoun County	03/26/2020	21821	26909	1361	13
Callahan County	04/07/2020	13660	22205	931	28
Cameron County	03/20/2020	420201	16085	32698	1126
Camp County	04/01/2020	12670	21069	969	32
Carson County	04/18/2020	6033	31788	331	11
Cass County	03/24/2020	30118	22145	1499	59
Castro County	03/21/2020	7891	22292	832	22
Chambers County	03/22/2020	39283	31412	3431	13

Cherokee County	03/27/2020	51594	21102	3491	92
Childress County	04/19/2020	7064	18694	1292	13
Clay County	04/01/2020	10369	27593	874	13
Cochran County	04/23/2020	2932	19195	216	12
Coke County	05/05/2020	3238	24623	417	10
Coleman County	05/02/2020	8422	26436	579	20
Collin County	03/09/2020	914075	41609	64721	483
Collingsworth County	05/05/2020	3026	21356	237	8
Colorado County	04/01/2020	20913	26689	1251	17
Comal County	03/22/2020	129100	35841	7314	190
Comanche County	04/01/2020	13492	22751	1118	36
Concho County	04/10/2020	3858	17513	245	4
Cooke County	04/10/2020	39064	29067	2806	48
Coryell County	03/24/2020	75818	21171	4312	45
Cottle County	04/27/2020	1498	20566	176	7
Crane County	03/20/2020	4836	24582	487	11
Crockett County	06/05/2020	3836	23296	472	13
Crosby County	04/06/2020	5895	20057	383	22
Culberson County	06/20/2020	2257	16763	294	4
Dallam County	04/06/2020	7207	25221	466	20
Dallas County	03/10/2020	2552213	29810	234625	2453
Dawson County	03/26/2020	13095	21360	1575	59

DeWitt County	03/19/2020	20474	28116	2418	60
Deaf Smith County	03/21/2020	18947	21209	213	5
Delta County	04/08/2020	5172	22732	46272	439
Denton County	03/17/2020	781321	37928	1553	50
Dickens County	04/08/2020	2224	24171	136	7
Dimmit County	04/11/2020	10822	17939	904	13
Donley County	04/02/2020	3433	23212	289	8
Duval County	04/14/2020	11434	19085	1058	32
Eastland County	03/24/2020	18278	20433	972	23
Ector County	03/29/2020	155744	27728	12222	269
Edwards County	06/25/2020	2111	28968	207	3
El Paso County	03/14/2020	834825	19950	107552	1940
Ellis County	03/20/2020	164092	28612	16403	201
Erath County	03/24/2020	41016	23511	3470	42
Falls County	03/23/2020	17289	17755	1308	22
Fannin County	03/20/2020	33787	23212	2439	66
Fayette County	03/24/2020	24963	30405	1447	43
Fisher County	05/15/2020	3875	27750	266	10
Floyd County	04/08/2020	5953	24347	631	25
Foard County	07/10/2020	1414	26034	91	5
Fort Bend County	03/05/2020	711421	38382	44067	439
Franklin County	03/29/2020	10639	23642	565	16
Freestone County	04/15/2020	19646	24060	988	30

Frio County	04/08/2020	19110	16833	2148	24
Gaines County	03/25/2020	19889	22656	1302	36
Galveston County	03/14/2020	321184	33870	26898	251
Garza County	04/30/2020	6739	20635	278	18
Gillespie County	04/02/2020	25939	32557	1830	32
Glasscock County	04/30/2020	1420	32885	94	1
Goliad County	04/02/2020	7510	30075	299	9
Gonzales County	04/01/2020	20558	23635	1962	29
Gray County	04/01/2020	22962	23457	1773	40
Grayson County	03/24/2020	126146	26535	8983	225
Gregg County	03/09/2020	123402	25144	8451	221
Grimes County	03/24/2020	27358	23585	1889	52
Guadalupe County	03/22/2020	150889	29300	9123	128
Hale County	03/27/2020	34527	19205	5668	135
Hall County	05/20/2020	3102	20022	353	11
Hamilton County	04/07/2020	8220	26522	546	21
Hansford County	04/06/2020	5532	21989	673	17
Hardeman County	05/22/2020	3990	21517	319	9
Hardin County	03/25/2020	55993	29693	4080	69
Harris County	03/05/2020	4525519	30856	286356	3825
Harrison County	03/26/2020	66606	25123	3575	71
Hartley County	04/20/2020	5821	20676	317	2
Haskell County	05/15/2020	5806	21120	355	17

Hays County	03/14/2020	194843	29253	14460	130
Hemphill County	04/04/2020	4152	29470	451	2
Henderson County	04/01/2020	79687	24315	4434	106
Hidalgo County	03/23/2020	839539	15883	56455	2018
Hill County	03/30/2020	35098	23342	2398	47
Hockley County	03/20/2020	23273	22673	2049	91
Hood County	03/27/2020	55418	32578	5050	85
Hopkins County	03/24/2020	35929	24236	2518	86
Houston County	04/19/2020	22849	17884	1360	30
Howard County	04/10/2020	36491	22994	4196	79
Hudspeth County	05/26/2020	3702	12543	466	8
Hunt County	03/24/2020	90322	23942	4557	103
Hutchinson County	04/04/2020	21704	25154	1253	57
Irion County	06/02/2020	1592	32307	84	1
Jack County	04/09/2020	8839	25553	530	12
Jackson County	03/26/2020	14756	26809	1284	23
Jasper County	04/03/2020	35444	21402	1965	48
Jeff Davis County	06/26/2020	2236	25167	116	1
Jefferson County	03/25/2020	254574	25370	15937	263
Jim Hogg County	04/23/2020	5262	17761	494	11
Jim Wells County	04/02/2020	41318	20631	3417	70
Johnson County	03/19/2020	160173	26574	14435	213
Jones County	04/08/2020	19969	17960	2569	35

Karnes County	03/26/2020	15051	27011	1387	26
Kaufman County	03/24/2020	114852	26631	11450	154
Kendall County	03/25/2020	40306	39517	1890	44
Kenedy County	07/03/2020	564	13705	30	2
Kent County	07/14/2020	680	27515	68	1
Kerr County	04/01/2020	50761	28484	3020	61
Kimble County	05/08/2020	4432	26982	227	6
King County	10/13/2020	289	29918	11	0
Kinney County	05/30/2020	3631	21395	258	3
Kleberg County	03/29/2020	31540	19806	1981	63
Knox County	04/07/2020	3753	21046	207	12
La Salle County	04/16/2020	7418	26268	814	15
Lamar County	03/24/2020	49401	23625	4956	112
Lamb County	03/28/2020	13368	21760	1799	73
Lampasas County	04/03/2020	20473	26405	913	17
Lavaca County	03/14/2020	19859	29946	1862	62
Lee County	04/02/2020	16850	26740	1182	34
Leon County	03/30/2020	16966	27096	875	28
Liberty County	03/25/2020	79884	22153	5116	111
Limestone County	03/24/2020	23480	21093	1420	34
Lipscomb County	04/28/2020	3495	29995	248	10
Live Oak County	04/01/2020	12043	22847	967	16
Llano County	03/25/2020	20195	35680	805	30

Loving County	11/17/2020	74	35530	1	0
Lubbock County	03/20/2020	298042	26196	45600	648
Lynn County	03/28/2020	5785	26758	571	18
Madison County	04/15/2020	13979	17436	1162	21
Marion County	04/16/2020	10140	25933	401	20
Martin County	03/27/2020	5547	28560	569	18
Mason County	04/04/2020	4122	24519	340	4
Matagorda County	03/15/2020	36744	23294	2602	72
Maverick County	03/23/2020	57471	16658	8320	223
McCulloch County	04/09/2020	8145	23398	467	12
McLennan County	03/19/2020	245720	24273	21894	344
McMullen County	07/14/2020	600	33472	70	2
Medina County	03/18/2020	48548	25572	3253	62
Menard County	06/22/2020	2123	23613	180	5
Midland County	03/21/2020	159883	38545	13334	202
Milam County	03/25/2020	24479	22911	1577	19
Mills County	05/19/2020	4887	24858	344	16
Mitchell County	04/07/2020	8774	19741	535	21
Montague County	03/27/2020	19406	26278	1713	46
Montgomery County	03/10/2020	535187	38012	33006	326
Moore County	03/30/2020	22016	21372	2116	57
Morris County	03/25/2020	12530	22803	672	15

Motley County	04/13/2020	1114	25908	79	7
Nacogdoches County	03/26/2020	65411	22589	3506	111
Navarro County	03/27/2020	48239	22152	4798	71
Newton County	04/01/2020	14187	20800	469	20
Nolan County	04/24/2020	14990	23686	1399	38
Nueces County	03/22/2020	358484	26780	31926	517
Ochiltree County	04/24/2020	10484	24157	965	19
Oldham County	03/21/2020	2083	25461	167	2
Orange County	03/22/2020	83909	27938	5946	81
Palo Pinto County	04/02/2020	28109	24840	2298	50
Panola County	04/03/2020	23574	26205	1228	49
Parker County	03/23/2020	125963	33367	11314	115
Parmer County	04/19/2020	9871	21876	1113	32
Pecos County	04/06/2020	15804	19088	1313	26
Polk County	03/29/2020	46974	23023	2492	70
Potter County	03/21/2020	121230	21941	15947	302
Presidio County	05/25/2020	7191	15329	615	19
Rains County	04/10/2020	11246	23976	632	20
Randall County	03/28/2020	130552	32922	15104	208
Reagan County	05/27/2020	3735	25273	358	7
Real County	06/05/2020	3358	20873	263	12
Red River County	04/16/2020	12353	21177	549	31

Reeves County	05/08/2020	14791	18992	1522	33
Refugio County	05/04/2020	7293	23959	536	17
Roberts County	04/16/2020	889	34555	49	1
Robertson County	03/24/2020	16727	23337	1069	25
Rockwall County	03/26/2020	90414	38933	8222	78
Runnels County	04/28/2020	10266	22190	1085	28
Rusk County	03/18/2020	53026	23521	3065	65
Sabine County	04/16/2020	10429	20876	452	28
San Augustine County	04/01/2020	8403	21066	496	26
San Jacinto County	03/29/2020	27436	22308	634	21
San Patricio County	03/26/2020	66867	24613	3450	111
San Saba County	05/19/2020	5851	22481	487	15
Schleicher County	06/13/2020	3122	28112	189	6
Scurry County	04/16/2020	17346	24140	2270	46
Shackelford County	05/07/2020	3348	24296	187	1
Shelby County	03/27/2020	25689	20686	1278	45
Sherman County	04/16/2020	3067	25358	118	11
Smith County	03/13/2020	222277	26270	15406	309
Somervell County	06/09/2020	8650	27095	883	8
Starr County	03/27/2020	63420	13167	7460	214
Stephens County	04/10/2020	9365	23044	629	19
Sterling County	08/13/2020	1139	25675	97	4

Stonewall County	07/18/2020	1084	28063	134	4
Sutton County	06/19/2020	3894	31603	406	6
Swisher County	03/29/2020	7541	18878	682	14
Tarrant County	03/10/2020	1983675	30857	195518	1798
Taylor County	03/27/2020	135371	25419	13275	263
Terrell County	06/23/2020	721	21204	73	1
Terry County	03/25/2020	12755	21938	1475	46
Throckmorton County	07/03/2020	1541	27732	56	3
Titus County	04/03/2020	32664	21090	3137	60
Tom Green County	03/24/2020	116906	27513	11547	204
Travis County	03/13/2020	1176584	38820	61468	626
Trinity County	04/04/2020	14481	20369	503	12
Tyler County	04/04/2020	21456	21172	999	19
Upshur County	03/24/2020	40506	24088	2595	42
Upton County	06/19/2020	3575	25290	286	7
Uvalde County	03/27/2020	27015	19146	2733	48
Val Verde County	03/26/2020	48976	20160	6331	158
Van Zandt County	03/25/2020	53607	25394	3154	77
Victoria County	03/26/2020	91518	28181	6672	134
Walker County	03/25/2020	70818	17194	7008	93
Waller County	03/29/2020	48443	23888	2659	27
Ward County	05/23/2020	11498	26860	953	13

Washington County	03/28/2020	34667	28517	1668	69
Webb County	03/17/2020	269624	16316	34690	513
Wharton County	03/25/2020	41430	25867	2988	84
Wheeler County	04/22/2020	5599	25809	431	9
Wichita County	03/20/2020	131778	23263	13325	260
Wilbarger County	04/14/2020	12972	21938	1707	41
Willacy County	03/27/2020	21839	13369	1970	69
Williamson County	03/19/2020	508313	34575	29796	269
Wilson County	03/26/2020	47205	29862	2790	41
Winkler County	04/10/2020	7777	23483	639	15
Wise County	03/30/2020	63247	27447	5384	84
Wood County	04/01/2020	43315	25955	2553	85
Yoakum County	04/18/2020	8481	23681	801	24
Young County	03/27/2020	18166	25661	1685	32
Zapata County	04/06/2020	14415	17817	1338	20
Zavala County	04/17/2020	12152	13105	1208	28

The resulting dataset is a valuable resource for researchers and public health officials who are interested in studying the spread of COVID-19 in Texas. The dataset can be used to answer a variety of questions, such as:

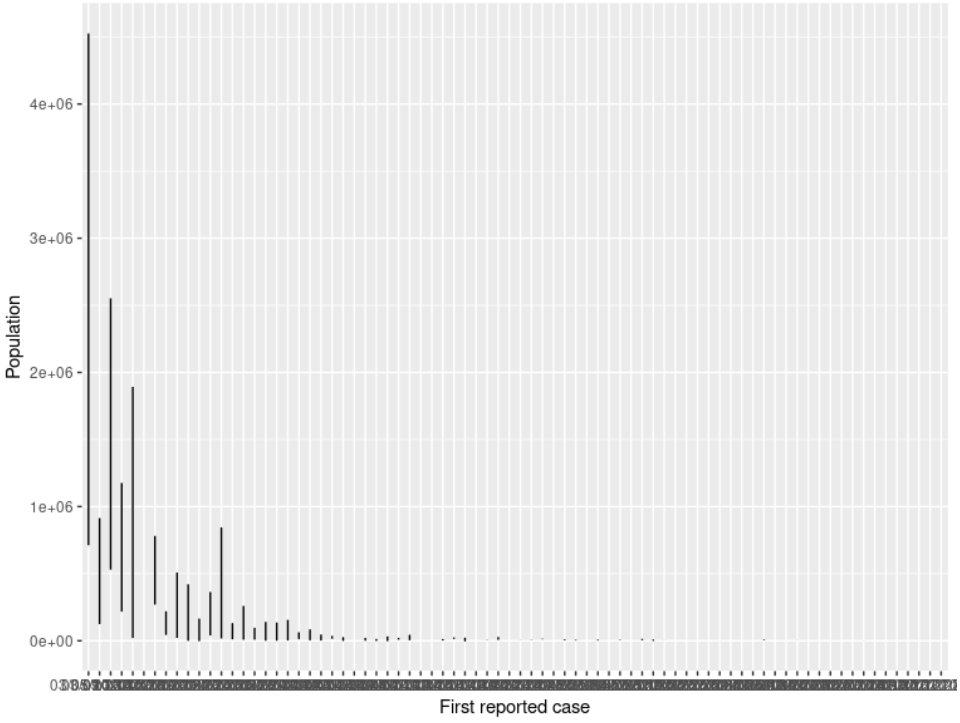
- Which counties in Texas have the highest rates of COVID-19 cases?
- How is the spread of COVID-19 related to factors such as population density and income?

- How effective have social distancing measures been in reducing the spread of COVID-19 in Texas?

Visualize the created attributes

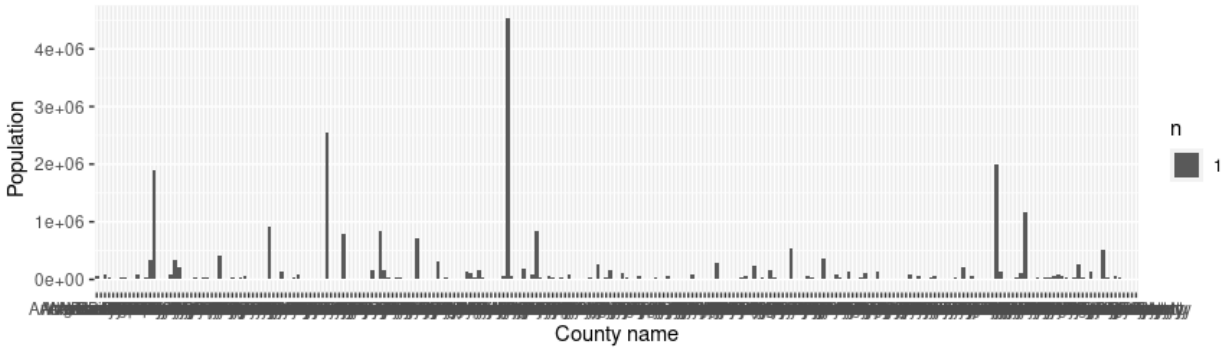
Once the data set is created, we can visualize the created attributes using various data visualization tools, such as:

- Line charts: Line charts can be used to show the trend of a variable over time. For example, we can use a line chart to show the trend of the number of cases over time in a particular county.



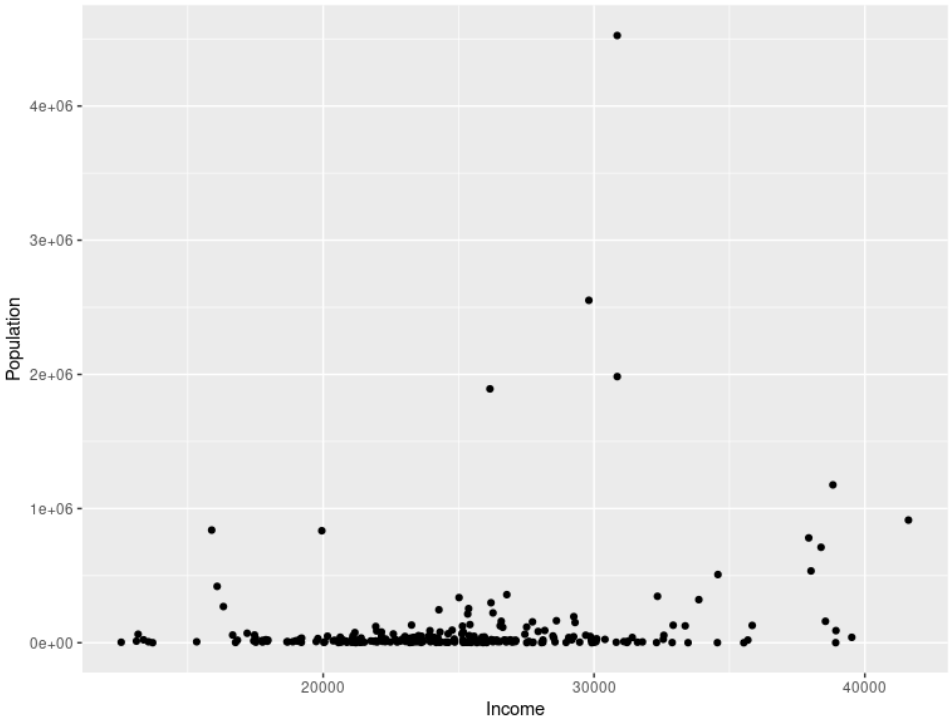
Data Preparation 1 Line chart to show the trend of the First_reported_case over time.

- Bar charts: Bar charts can be used to compare the values of a variable across different groups. For example, we can use a bar chart to compare the number of cases in different counties in a particular state.



Data Preparation 2 A bar chart to compare the population in different counties.

- Choropleth maps: Choropleth maps can be used to visualize the distribution of a variable across a geographic area. For example, we can use a choropleth map to visualize the distribution of the number of cases across the United States.
- Scatter plots: Scatter plots can be used to show the relationship between two variables. For example, we can use a scatter plot to show the relationship between the number of cases and the population density of a county.



Data Preparation 3 A scatter plot to show the relationship between the population and the income.

We can also use more advanced data visualization techniques, such as:

- Heatmaps: Heatmaps can be used to visualize the relationship between multiple variables. For example, we can use a heatmap to visualize the relationship between the number of cases, the population density, and the social distancing measures in different counties.
- Parallel coordinates: Parallel coordinates can be used to visualize the distribution of multiple variables in a single plot. For example, we can use parallel coordinates to visualize the distribution of the number of cases, the population density, the social distancing measures, and the demographics of different counties.

By visualizing the created attributes, we can gain insights into the spread of COVID-19 and the effectiveness of social distancing measures. For example, we can identify counties that have been particularly successful in containing the spread of the virus. We can also identify counties that are at high risk of a surge in cases.

This information can be used by public health officials to develop targeted interventions to reduce the spread of COVID-19 and protect the public.

Data Wrangling the COVID-19 with Census Data Set

In order to use the COVID-19 with Census, the Cases Texas, and Global Mobility Report data sets we need to clean it up so to speak, extract important features, and overall condense the dataset so we can us it for predictive modeling in the next project. First, thing to clean up the dataset was to remove all columns that had titles but held no data, we could find these columns using the command in the figure below.

Data Wrangling Figure 1

```
```{r}
library(purrr)
cols_with_na <- cases %>% keep(~all(is.na(.x))) %>% names
cols_with_na
```

 [1] "pop_5_years_over"   "speak_only_english_at_home"   "speak_spanish_at_home"   "speak_spanish_at_home_low_english"
 [5] "pop_15_and_over"   "pop_never_married"   "pop_now_married"   "pop_separated"
 [9] "pop_widowed"   "pop_divorced"
```

Next, we want to select and extract features. The following command pulls all features with the word “commute in it”. We will use this query to select features to combine into local commuters, out_of_town_commuters (commuters who travel over 50 minutes for work), and group commuters (commuters that ride bus, subway, or carpool). The rest of the column commuter features are dropped from the dataset.

Data Wrangling Figure 2

```
```{r}
commute_cols <- cases[ , grepl( "commute" , names( cases ) ) ]
commute_cols


```
```

Next, we use a similar command to query all features that had the word “income” in it. Then, we stratified the income categories into four different groups, income_under_40k, income_over_40_under_100, income_over_100_under_200, and income_over_200. The rest of the income fields except median_income was dropped from the dataset.

Data Wrangling Figure 3

```
```{r}
income_cols <- cases[ , grepl( "income" , names( cases ) ) ]
income_cols

```
```

Next, we wanted to capture the number of K-12 students in the county. The reason for this is because children are known to be silent spreaders of the virus so counties with large amounts of

children are likely to have more cases. Thus, we need to maintain those features as they are good indicators of the likely spread of COVID-19. Therefore, we command all student numbers from Kindergarten to High School into a K-12 feature. Finally, the last three features want to extract and add to the dataset was per county features for cases_per_1000, deaths_per_1000, and death_per_case. The below command was used to create those features and add them to the dataset.

Data Wrangling Figure 1

```r
cases <- cases %>% mutate(cases_per_1000 = confirmed_cases/total_pop*1000,
  deaths_per_1000 = deaths/total_pop*1000,
  death_per_case = deaths/confirmed_cases)
summary(cases)
```

The final step for now in preparing our COVID-19 with Census dataset for predictive modeling is to combine our features all into one dataset. The below command completed that step. It is important to note here that State and County will need to be removed before we normalize the data as nominal values as such cannot be normalized but they are good here to identify states and counties that may have exhibited extraordinary circumstances during the pandemic.

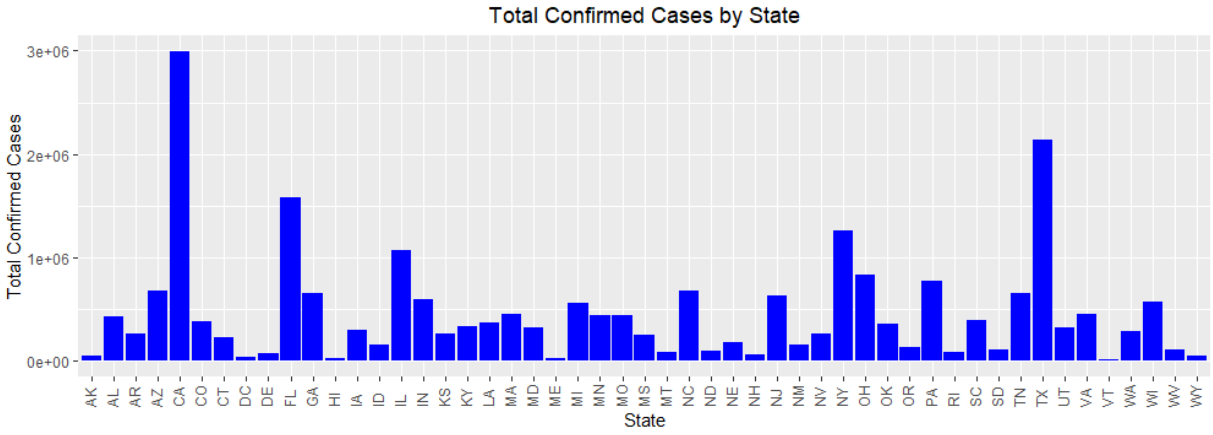
```r
cases_final <- cases[c('county_name', 'state', 'confirmed_cases', 'deaths', 'total_pop', 'white_pop', 'black_pop', 'asian_pop', 'amerindian_pop',  'hispanic_pop',
'median_income', 'households_public_asst_or_food_stamps', 'children_in_single_female_hh', 'employed_pop', 'unemployed_pop', 'pop_determined_poverty_status',
'not_us_citizen_pop', 'two_parent_families_with_young_children', 'poverty', 'income_under_40', 'income_over_40_under_100', 'income_over_100', 'income_200000_or_more',
'commuters_in_groups', 'local_commuters', 'out_of_town_commuters', 'cases_per_1000', 'deaths_per_1000', 'death_per_case', 'k12_students','children')]
cases_final
```

A tibble: 3,142 × 31

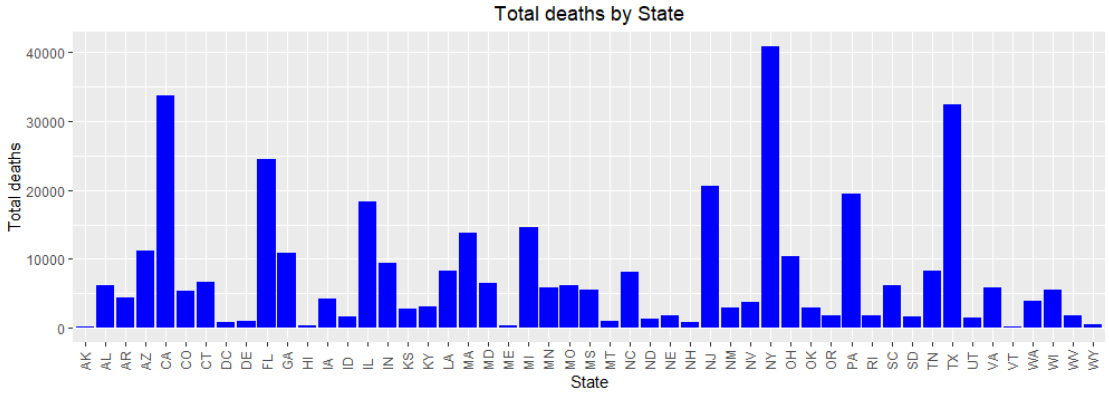
county_name <fctr>	state <fctr>	confirmed_cases <dbl>	deaths <dbl>	total_pop <dbl>	white_pop <dbl>	black_pop <dbl>	asian_pop <dbl>	amerindian_pop <dbl>	hispanic_pop <dbl>
Essex County	VT	111	0	6203	5929	64	32	20	83

CHAPTER 4: Exceptional works

What US states have the most confirmed/death cases?



Total confirmed cases by state

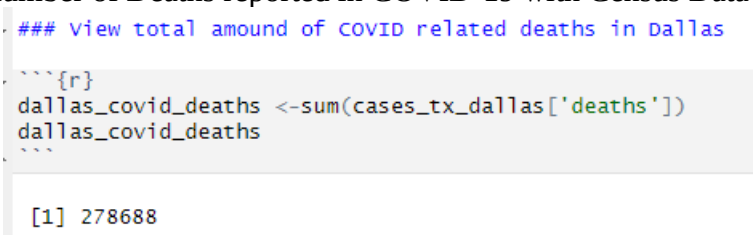


How many deaths did Dallas have during 2020 using our DataSet compared to the number of deaths that have been reported today?

Using the COVID-19 with Census dataset we show the result is 278688 up until January 21st, 2021. According to Coronavirus (COVID-19) statistics, we are now at 1,127,152 deaths.

,

Total number of Deaths reported in COVID-19 with Census Data



```
### View total amound of COVID related deaths in Dallas

```{r}
dallas_covid_deaths <-sum(cases_tx_dallas['deaths'])
dallas_covid_deaths
```

 [1] 278688
```

CONCLUSION

In this report, we have focused on cleaning and understanding the data. We have removed rows with missing values and identified any inconsistencies in the data. We have also visualized the data using a variety of methods, including line charts, bar charts, scatter plots, choropleth maps, heatmaps, and parallel coordinates plots.

The data collected by Google about the spread of the COVID-19 virus is a valuable resource for researchers and public health officials. The data can be used to answer a variety of questions about the spread of the virus and the effectiveness of interventions. The visualizations show that the trend in the number of cases varies across different areas of the US. Some areas have experienced a surge in cases, while others have seen a decrease in cases. Social distancing measures appear to be effective in reducing the spread of the virus, but they need to be implemented and maintained consistently.

The data can be used to identify regions that are at high risk for a surge in cases. This information can be used to target interventions to these regions. The data can also be used to predict the development of the virus in a region given the data of other regions.

This information is helpful for public health officials who are making decisions about how to respond to the pandemic. The data can be used to allocate resources, develop policies, and communicate with the public about the virus. In addition to the methods used in this report, there are a variety of other methods that can be used to visualize and analyze the data. For example,

researchers can use machine learning algorithms to identify patterns in the data and to predict the future spread of the virus.

Overall, the data collected by Google about the spread of the COVID-19 virus is a valuable tool for researchers and public health officials. The data can be used to answer a variety of questions about the virus and to develop effective interventions.

CONTRIBUTIONS:

Zahra Hoobakht:

- Introduction
- Business Understanding
- Data Understanding (Texas and States Trends section)
- Data Preparation (section 3.1-table)

Jason Brown:

- Data Understanding
- Data Preparation (Data Wrangling section)

Bhargava Sharabha Pagidimarri

- Abstract
- Data Preparation
- Conclusion

BIBLIOGRAPHY

[1] Fauci, A. S., & Lane, H. C. (2020). "Redfield R. R. Covid-19 - Navigating the Uncharted." New England Journal of Medicine

[2] Emanuel, E. J., Persad, G., Upshur, R., Thome, B., Parker, M., Glickman, A., ... & Phillips, J. P. (2020). "Fair allocation of scarce medical resources in the time of Covid-19." New England Journal of Medicine

[3] Google Cloud. COVID-19 Datasets. Google Cloud Marketplace. URL: https://console.cloud.google.com/marketplace/browse?filter=solution%20type:dataset&filter=category:covid19

[4] Centers for Disease Control and Prevention. (2021). “Symptoms of COVID-19.” URL: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html

[5] World Health Organization. (2023). “COVID-19: Clinical features. World Health Organization”. URL: https://www.who.int/news-room/fact-sheets/detail/coronavirus-disease-(covid-19)

[6] Johns Hopkins Medicine. (2020). “Coronavirus Social Distancing and Self-Quarantine.” URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/coronavirus-social-distancing-and-self-quarantine#:~:text=The%20practice%20of%20social%20distancing,instead%20of%20in%2Dperson%20contact.

[7] US Bureau of Labor Statistics. (2023). “Labor force characteristics by race and ethnicity, 2021” URL: https://www.bls.gov/opub/reports/race-and-ethnicity/2021/home.htm#:~:text=Among%20adult%20men%20%2820%20years%20and%20older%29%20in,least%20likely%2C%20with%20a%20rate%20of%2066.5%20percent.