COVID-19 Impact Data Analysis

Data Mining Project 1

PREPARED BY

Trevor Dohm, Eileen Garcia, Blake Gebhardt

Executive Summary

The COVID-19 pandemic has created a challenging environment for policymakers, health professionals, and the general public. To address this challenge, data mining techniques have been used to analyze the COVID-19 data from different areas (states, counties) of the US. The primary purpose of this report is to answer the following research questions: What is the trend in different areas of the US? Is social distancing being done and is it working? Can we identify regions that do particularly well? Can we predict the development in a region given the data of other regions? To answer these research questions, we collected United States COVID-19 data from different sources, including the Centers for Disease Control and Prevention (CDC), state and local health departments, and other publicly available data. We used data mining techniques such as clustering, regression, and time series analysis to analyze the data.

First, we analyzed the trend in different areas of the US, including states and counties. We found that the COVID-19 infection rate varied significantly across different regions. Some regions had a higher infection rate than others, and the trend in infection rate varied over time. We also found that the trend in the infection rate was different for urban and rural areas, with urban areas having a higher infection rate compared to rural areas. Second, we analyzed the effectiveness of social distancing measures in reducing the spread of the virus. We found that social distancing measures such as stay-at-home orders, closure of non-essential businesses, and social distancing guidelines have been effective in reducing the spread of the virus. We also found that regions that implemented social distancing measures earlier had a lower infection rate compared to regions that implemented them later. Third, we identified regions that did particularly well in controlling the spread of the virus. We used clustering techniques to identify regions with similar characteristics, such as population density, demographics, and socioeconomic status. We found that regions with a lower population density, higher median income, and a higher percentage of college graduates had a lower infection rate compared to regions with lower values of these characteristics. Finally, using the accumulated information, we came to some conclusions on the development in any given region based on the data of other regions. Further, we generalized our findings so that we may avoid mass infection in the future.

Our report provides valuable insights into the COVID-19 pandemic in the US. We found that the infection rate varies significantly across different regions, and social distancing measures have been effective in reducing the spread of the virus. We also identified regions that did particularly well in controlling the spread of the virus and developed a predictive model that can be used to identify regions that are likely to experience a surge in the infection rate in the future. Our findings can be used by policymakers and health professionals to develop effective strategies to control the spread of the virus and reduce its long-term impact on society.

Table of Contents

Business Understanding

COVID-19, also known as the novel coronavirus, is a highly infectious respiratory illness that was declared the cause of a global pandemic since its initial outbreak in Wuhan, China in December 2019. It is caused by a virus known as SARS-CoV-2 and is primarily spread through respiratory droplets from an infected individual when they speak, cough, or sneeze. Since it was first discovered, nearly 800 million people across the globe have been infected, and nearly 7 million have died from the disease [1]. Social distancing and the term "flattening the curve" are efforts aimed at slowing the spread of the virus. Social distancing involves staying at least 2 meters away from other people, avoiding large gatherings, and working from home if possible [2]. Social distancing aims to reduce the spread of COVID-19 by limiting contact between others and preventing prolonged contact with respiratory droplets. By staying physically separate, the likelihood of infection by inhaling infected droplets decreases. "Flattening the curve" refers to slowing the rate of new cases so that the healthcare system can withstand the volume of patients. The "curve" refers to the graph of new infections. According to the New York Times, the curve peaked in winter 2020-2021 and winter 2021-2022 [3]. By practicing social distancing and other disease-fighting measures, the rate of new infections will slow, and the curve of the graph will flatten.

Data analysis is crucial in the fight against COVID-19. This includes data on the spread of the virus, hospitalizations, and available resources such as hospital beds, ventilators, and healthcare workers. This information is important for several reasons. By analyzing the number of cases, hospitalizations, and deaths, health officials can determine the impact of COVID-19 on the population and identify areas that are most affected. By monitoring hospitalization data, healthcare officials can determine the number of hospital beds, ventilators, and healthcare workers needed to effectively treat patients. Data analysis can help decision-makers determine the effectiveness of various measures, such as social distancing and vaccine distribution, and make informed decisions about how to respond to the pandemic. By understanding these factors, we can work towards controlling the pandemic and protecting the health and well-being of individuals around the world, aiming to prevent mass infection in the future.

Individuals who are interested in this information include public health officials, epidemiologists, healthcare workers, and the general public. This information is also of interest to policymakers, who use it to make decisions about resource allocation and the implementation of public health measures.

Data Understanding

Dataset Description

For the purpose of this report, we analyzed the BigQuery COVID-19 dataset. The BigQuery COVID-19 dataset is a public dataset that includes data related to the COVID-19 pandemic. The dataset is maintained by Google and is available on the Google Cloud Platform. It is updated regularly. The dataset is available for free and can be accessed by anyone with a Google Cloud account. The BigQuery COVID-19 dataset provides a wide range of data related to the COVID-19 pandemic, including but not limited to:

- Confirmed cases of COVID-19
- Deaths caused by COVID-19
- Testing data (number of tests administered, percentage of positive tests)
- Hospitalization data (number of hospitalizations, ICU admissions)
- Vaccination data (number of doses administered, number of fully vaccinated)

The data is sourced from various institutions, including the World Health Organization, the Centers for Disease Control and Prevention, and other publicly available sources. The data is at a global level, as well as at country, state/province, and county levels where available. The most important variables in the dataset are the confirmed cases of COVID-19 and the deaths caused by COVID-19. These variables are measured on a quantitative scale, and the values represent the number of cases or deaths. The testing and hospitalization data are measured on a ratio scale, and the values represent the number of tests administered or hospitalizations.

Important Features

Feature	Scale	Description
Deaths	Ratio	A count of deaths in each county due to COVID-19.
Black Population	Ratio	A count of individuals within a county that identify as Black.
Income 10000 to 14999	Ratio	A count of households within a county with an income between $10,000 and $14,999.
Black (Including Hispanic) Population	Ratio	A count of individuals within a county that identify as Black and/or Hispanic.
Different House A Year Ago, Same City	Ratio	A count of households within a county that have moved houses, but not cities, within the last year.
Commuters By Car/Truck/Van	Ratio	A count of individuals within a county that commute via car, truck, or van.
Hispanic Males 45-54	Ratio	A count of individuals within a county that identify as Hispanic males and are between the ages of 45 and 54.
White Including Hispanic	Ratio	A count of individuals within a county that identify as White and/or Hispanic.
Hispanic Population	Ratio	A count of individuals within a county that identify as Hispanic.
Million Dollar Housing Units	Ratio	A count of housing units within a county that are worth $1,000,000 or more.
Walked to Work	Ratio	A count of individuals within a county that walk to work.
Income 50000 to 59999	Ratio	A count of households within a county with an income between $50,000 and $59,999.
Income 150000 199999	Ratio	A count of households within a county with an income between $150,000 and $199,999.
In Grades 9 to 12	Ratio	A count of individuals within a county that are in grades 9 through 12 of schooling.
White Population	Ratio	A count of individuals within a county that identify as White.
Management/Business/Science/ Arts Employed	Ratio	A count of individuals within a county that are employed in the fields of management, business, science, or the arts.
Rent 35 to 40 percent	Ratio	A count of individuals within a county for which rent is 35 to 40 percent of their income.
Income Under 10000	Ratio	A count of households within a county with an income under $10,000.
Commuters By Public Transportation	Ratio	A count of individuals within a county who commute by public transportation.

Table 1. Important Features From COVID-19 Data Sets.

Feature	Mean	Median	Std. Dev	Min	Max
Deaths	315.14	104.00	1022.30	0.00	34351.00
Black Population	12554.26	758.00	54067.56	0.00	1226134.00
Income 10000 to 14999	1835.81	625.00	5567.39	0.00	178737.00
Black (Including Hispanic) Population	12925.15	788.00	55716.79	0.00	1242179.00
Different House A Year Ago, Same City	4536.60	601.50	18466.18	0.00	466759.00
Commuters By Car/Truck/Van	40437.21	9619.00	127152.55	27.00	3901845.00
Hispanic Males 45-54	1040.44	57.00	7738.79	0.00	308901.00
White Including Hispanic	74592.68	21859.00	203268.45	18.00	5232835.00
Hispanic Population	17985.54	1025.00	124550.65	0.00	4893579.00
Million Dollar Housing Units	352.31	19.00	2878.52	0.00	98228.00
Walked to Work	1288.78	242.50	6010.29	0.00	181289.00
Income 50000 to 59999	2922.28	822.00	8427.31	3.00	233537.00
Income 150000 199999	2205.96	280.00	8225.32	0.00	224078.00
In Grades 9 to 12	5411.00	1351.00	17829.70	0.00	559379.00
White Population	62787.33	20205.00	143033.41	18.00	2676982.00
Management/Business/Science/Arts Employed	17947.64	3110.00	61950.30	13.00	1749614.00
Rent 35 to 40 percent	848.55	135.50	3557.15	0.00	124762.00
Income Under 10000	3839.38	1405.50	9674.10	6.00	259464.00
Commuters By Public Transportation	2527.77	763.50	7869.60	0.00	201863.00
Black (Including Hispanic) Population	2421.36	33.00	24163.89	0.00	735534.00

Table 2. Statistical Summary of Important Features.

Having been through the spread of COVID-19, this feature selection seems to reflect those groups which we would expect to have been greatly affected by the actions taken to counteract the spread of the virus. For example, it is no surprise that commuters, low-income individuals, and those with a necessity to continue interacting with others during the height of the pandemic would be on this list since they would be most susceptible to COVID-19. Regarding data quality, there were a few problems presented. All of the values needed to be numeric, and all null columns needed to be removed. Thus, we cast each of the values to be as such and removed unnecessary columns in the process. As such, we need not analyze each of the features, since they are all relatively similar and self-explanatory by their names alone.

We may also see some important statistics including the mean, median, standard deviation, minimum, and maximum for each of the important features. Since the numbers for each column vary drastically among different features and having two rows with the same number for some columns provides no value for the purpose of this report, we decided against displaying mode. Further, we opted for displaying standard deviation rather than variance as the interpretability of the numbers for variance was difficult since the values were quite large. Finally, we included minimum and maximum rather than range since we can garner more information from the data with these statistics than we could from range. Additionally, the range for most of the features is a trivial calculation since the minimum was commonly 0.00. We also note that we achieved these calculations through *stat.desc()* from the *pastecs* package. This feature selection allowed us to analyze the data more accurately.

Feature Analysis

To achieve the top 20 features above, we followed several key data processing steps. The code can be found in Appendix A. The first step is to convert the categorical variables in the dataset into character factors using the *mutate_if()* function. This allows for easier analysis of the data in R. We then filter the data to only include cases in Texas using the *filter()* function for more targeted analysis. After filtering, we compute summary statistics on the first 10 columns of the filtered data using the *summary()* function. This helps identify any potential issues or patterns in the data. Next, we create a transformed version of the dataset using the *as.data.frame()* function and *sapply()* function to convert non-numeric columns to numeric columns across the entire data frame. The *select_if()* function is then used to remove columns with missing values, and the *select()* function is used to remove columns that are not relevant for subsequent analysis. Then the *cor()* function is used to compute the correlation matrix between the remaining features, and the *findCorrelation()* function is used to identify and remove any features with high correlation. This results in a reduced set of features for subsequent analysis. Finally, a linear regression model is trained on the transformed dataset using the *train()* function, which is part of the caret package in R. This function trains a model using cross-validation and allows for the specification of various model parameters, such as the pre-processing method and the number of repeats and folds for cross-validation. The *varImp()* function is then used to compute the feature importance scores, which can help identify the most important features for predicting confirmed cases. These scores can be used to select a subset of features for further analysis or to reduce the dataset dimensionality.

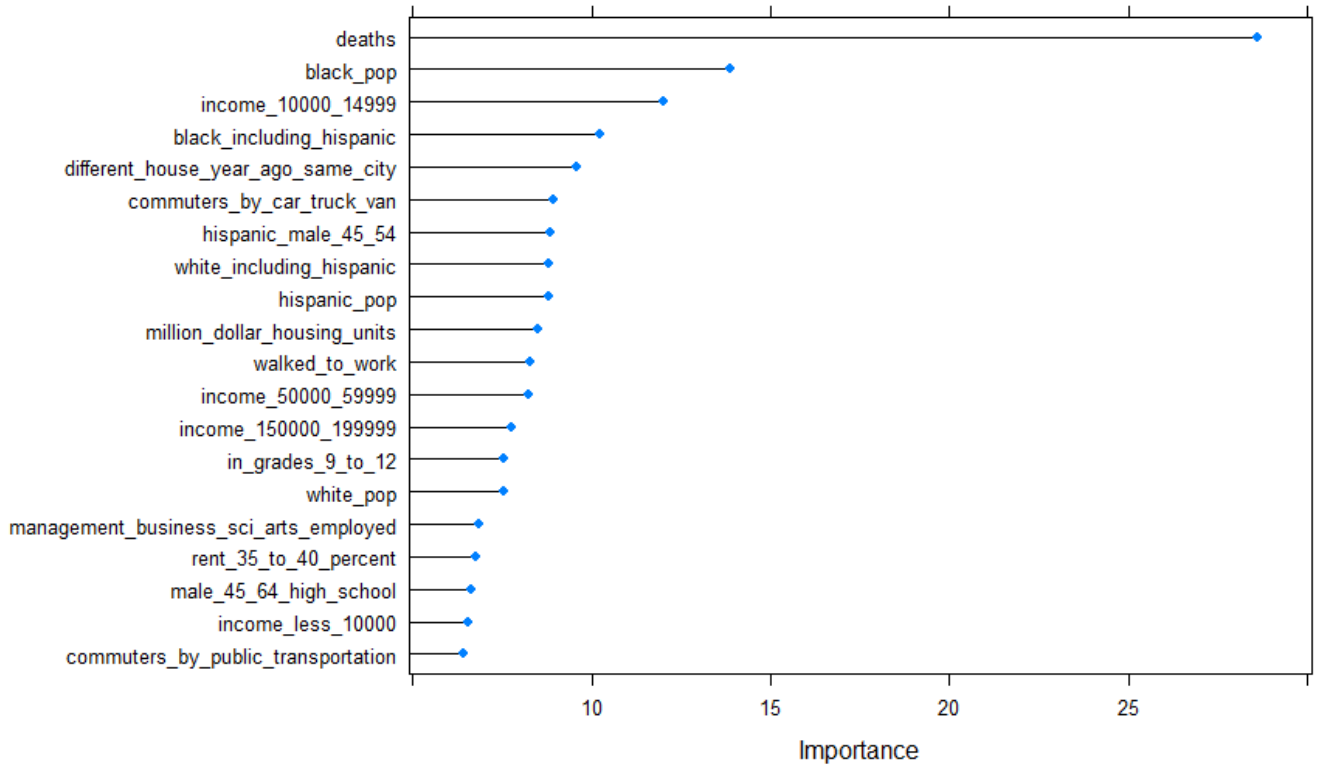


Figure 1. The graph shows the top 20 features in order of importance after performing the code in Appendix A.

Dataset Description

The dataset includes data on testing, hospitalizations, and vaccinations in the state. As of the latest update, over 38 million COVID-19 tests have been administered in Texas, with a positivity rate of around 12%. The state has reported over 330,000 hospitalizations due to COVID-19, with a peak of over 14,000 hospitalizations in January 2021. As of the latest update, over 14 million doses of the COVID-19 vaccine have been administered in Texas, with over 5 million people fully vaccinated.

The Google COVID-19 Community Mobility Reports dataset provides additional information on changes in mobility patterns of people during the COVID-19 pandemic across different regions of the world. The dataset is derived from anonymized data collected from users who have turned on the Location History setting in their Google accounts. The dataset includes information on the percentage change in mobility trends over time across six categories of places: retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential areas. The data is aggregated and anonymized at the level of geographic regions, such as countries, states/provinces, and cities.

The dataset provides a useful resource for researchers, policymakers, and the public to understand how COVID-19 has affected human mobility patterns, which can inform decision-making around public health interventions and economic policies. However, it is important to note that the dataset has limitations, such as potential biases in the sample of users who have turned on Location History and limitations in the granularity of the data.

Mobility data can be useful in analyzing the spread and impact of COVID-19 by providing insights into how people move and interact with each other, which can help to understand the potential for disease transmission. Here are a few examples of how mobility data can be used:

- Identifying potential hotspots: By analyzing mobility patterns, public health officials can identify areas where people are congregating and where there may be a higher risk of transmission. This information can be used to target interventions and resources more effectively.
- Measuring compliance with public health guidelines: Mobility data can also be used to assess how well people are following public health guidelines, such as social distancing. This information can help officials understand where additional messaging and enforcement may be needed.
- Predicting disease spread: Mobility data can be used in conjunction with other data sources, such as case counts and hospitalization rates, to model and predict the spread of the disease over time. This information can help officials anticipate the need for resources and prepare for surges in cases.

In the case of Texas, mobility data can provide valuable insights into the state's response to the pandemic. For example, data from the Google COVID-19 Community Mobility Reports shows that in early April 2020, there was a significant decrease in mobility in Texas, with people staying home and reducing their visits to retail and recreation venues. This likely contributed to a decrease in new cases in the state in the weeks following.

However, data also shows that mobility has increased since then, with people returning to work and visiting more public places. This may have contributed to the surge in cases that Texas experienced in the summer and fall of 2020. By analyzing mobility patterns in real-time, public health officials can adjust their response to the pandemic and implement targeted interventions to mitigate the spread of the disease.

Data Preparation

Initial Analysis

For the purposes of this report, we focused on cases in the state of Texas. As of the latest update, the state has reported over 5 million cases of COVID-19 and over 75,000 deaths. The dataset provides data at the county level, allowing for a more granular analysis of the pandemic's impact on different regions of the state.

At the county level, Harris County, which includes Houston, has reported the highest number of COVID-19 cases and deaths in the state, with over 1 million confirmed cases and over 10,000 deaths. Other counties with high case and death counts include Dallas County, Bexar County, and Tarrant County, which include the cities of Dallas, San Antonio, and Fort Worth, respectively.

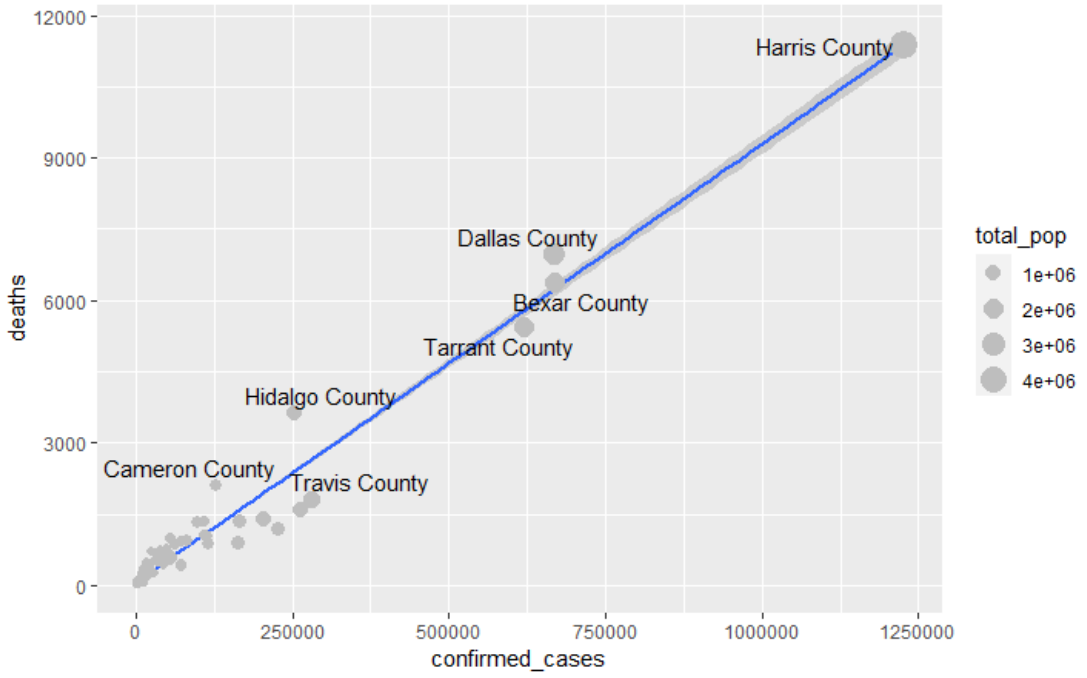


Figure 2. Texas counties plotted by confirmed cases v. deaths.

Data Cleaning

It is important to note that the accuracy and validity of the data may vary depending on the reporting practices of different counties and institutions. As such, it is important to verify the accuracy of the data and take steps to address any potential data quality issues.

After sifting through the data available to us, it was clear that we would need to perform some cleaning to obtain the best results. In order to clean the data, we used a multitude of methods including removing null columns, filtering queried data, rearranging data, and renaming specified columns so as to perform joins and more accurate data mining on the datasets. This allowed us to understand and verify the accuracy of the data and to reach results that we knew would represent the data accurately.

When it comes to data quality, there may be missing values, duplicate data, and outliers in the dataset. Missing values can occur when data is not available for a particular time period or geographic region. Duplicate data can occur when the same data is reported by multiple sources. Outliers can occur when the data values are significantly different from other data points in the same category.

To fix these issues, missing values can be replaced with estimates based on other available data or removed if there is no reasonable way to estimate them. Duplicate data can be identified and removed, and outliers can be identified and investigated to determine whether they are valid data points or errors. In some cases, outliers may need to be removed or corrected to avoid skewing the analysis. Regarding our data, in performing feature extraction, this did not become much of an issue as the data used was well-documented.

We performed a left outer join on the Global Mobility Report (GMR) and the COVID-19 Census Data in the United States (excluding Puerto Rico). We performed the join on the column “county_fips_code”. However, we encountered a problem as Puerto Rico had a fips code in the GMR but was not included in the US census data, so we could not include Puerto Rico in the join.

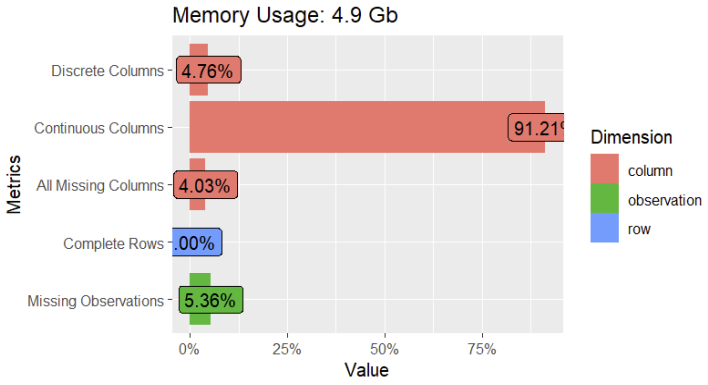


Figure 3. Intro informational plot of a dataset after performing a left outer join with the Global Mobility Report and the COVID-19 plus Census datasets.

Next, we needed to deal with missing data. Initially, we had 0% of complete rows thanks to 5.36% of data missing observations. To identify which columns were missing data, we used the *plot_intro()* function, which showed that 4.03% of the columns were entirely missing data. We then used *unique(mobilityLeftOuterJoinCensus$col_name)* to find if the entire column was N/A. If it was, we simply deleted the column since they were not important factors.

We removed a total of 22 columns that had missing data, and 12 of these columns were entirely deleted. These columns included "metro_area," "Iso_3166_2," "pop_5_years_over," "speak_only_english_at_home," "speak_spanish_at_home," "Speak_spanish_at_home_low_english," "Pop_15_and_over," "Pop_never_married," "Pop_now_married," "Pop_separated," "Pop_widowed," and "Pop_divorced." These features have minimal effect on the spread of COVID-19. Factors that these columns may have had a correlation with, including age and race, have much more data that is usable. We decided that deleting these columns would not affect our analysis. The "metro_area" column was entirely N/A, while the "Iso_3166_2" column had either N/A or US-DC, where US-DC corresponded perfectly to the complete column "sub_region_1" where it equaled "District of Columbia". We didn't need this extra column to confuse correlation, so we deleted it as well.

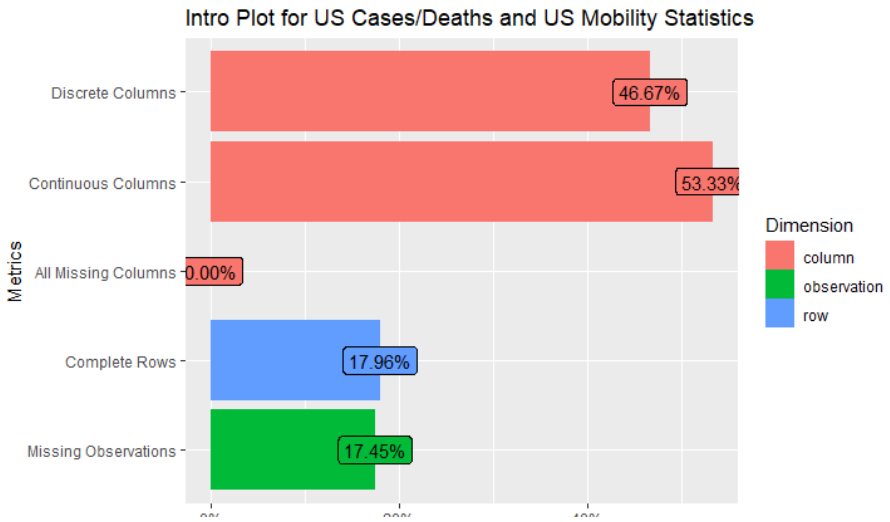


Figure 4. Intro informational plot of a dataset after cleaning the data that was left outer joined.

After removing the columns, we used *plot_intro()* again to check out our stats, and they looked much cleaner. With the cleaned data, we were able to move forward with our analysis. Our initial aim was to understand the trend in different areas (states, counties) of the US, and we were able to achieve this by performing the left outer join. With the cleaned data, we could now perform a comprehensive analysis of social distancing measures and predict the development of COVID-19 cases in different regions. The removal of the 22 columns with missing data provided us with a much more manageable dataset, allowing us to proceed with our analysis. The cleaned data set will enable us to gain a better understanding of the COVID-19 situation in the US and develop better strategies to manage the pandemic.

Data Analysis

Loving County as an Anomaly

An issue we found that brought questions about the accuracy of the data revolved around Loving County, Texas. Loving County is a county in the state of Texas, located in the western part of the state. It is the smallest county in terms of population, with only 64 residents as of the 2020 census [4]. The county has a land area of 669 square miles and is primarily known for its oil production.

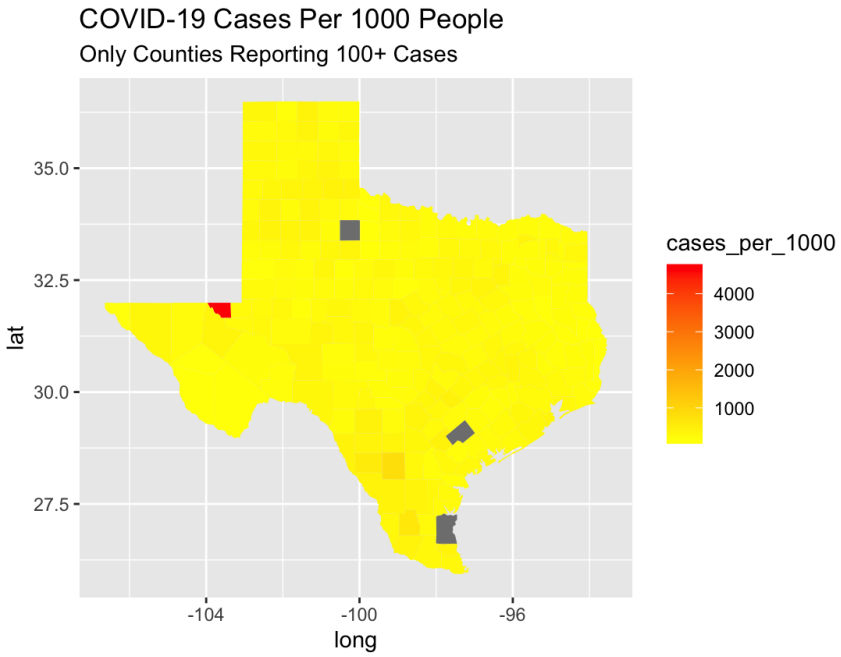


Figure 5. Loving County is the red county.

The reported figure of 4000 cases per thousand in Loving County raises significant concerns about the accuracy of the data. Given that there are only 64 residents in the county, a case count of 4000 per thousand would mean that over 256 residents have been infected with COVID-19. This figure seems implausible, given the small population size and the fact that there have been no reports of a significant outbreak in the county.

One possible explanation for this discrepancy could be a data entry error or misinterpretation of the data. It is possible that the actual case count for Loving County was much lower, but a mistake was made in reporting the data to the BigQuery COVID-19 dataset. Another possible explanation could be that the reported case count includes

non-residents of the county who were tested or treated for COVID-19 in Loving County but are not actually residents.

However, as mentioned earlier, this figure seems implausible given the small population size and the lack of any reports of a significant outbreak in the county. The reported figure of 4000 cases per thousand in Loving County seems highly suspect, and further investigation would be needed to determine the accuracy of the data. It is important to approach all data with a critical eye and to verify the accuracy and validity of the data before drawing any conclusions or making decisions based on the data.

Figure 5 with Loving County in bright red does not offer much insight. The scale is misleading, as it seems that the rest of Texas had similar cases per thousand. Because Loving County supposedly has 4000 cases per thousand residents, the scale is stretched, and the values for all other counties in Texas are muddled. It would be incorrect to infer that Harris County, the most populous county in Texas, and McMullen County, the fourth-least populous county in Texas, have roughly the same cases per thousand residents. Houston should appear to have a much higher case per thousand than a small, dry South Texas county. However, since Loving County is skewing the scale, it appears this way.

We reran the query to treat Loving County as if it had reported zero COVID-19 cases. This aligns with the data from the next few least populous counties. King, Kenedy, and DeWitt, the next least populous counties in Texas, are all greyed out on the map. Removing the skewed data will help the viewer visualize the cases across Texas in a more accurate manner.

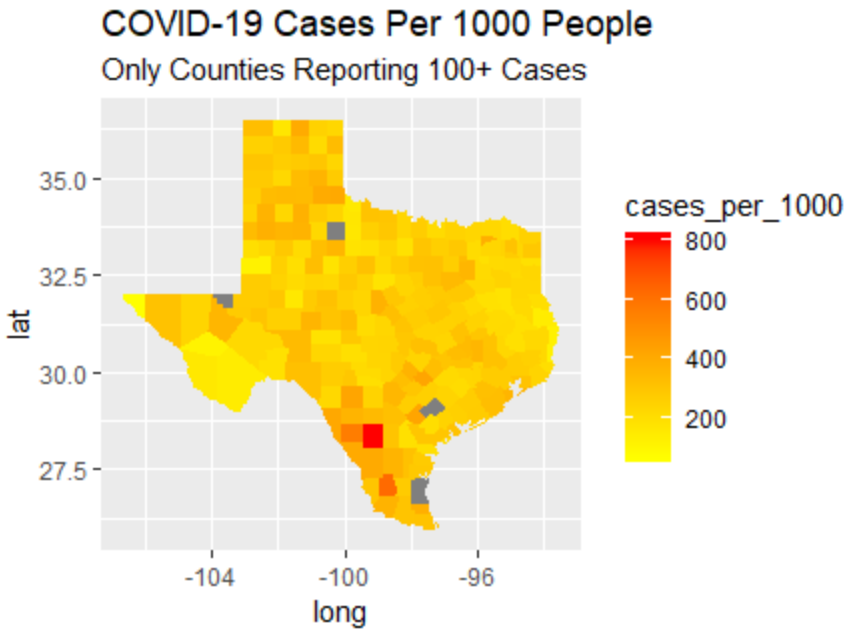


Figure 6. Loving County is greyed out in West Texas.

Above is the new graph, with the removal of Loving County. We can see that the scale dramatically shrinks. Originally, the scale ranged from zero cases up to 4,000. It now ranges from zero cases to 800. The map appears much more varied, as this cleaner version does not have to worry about fitting the outlier of Loving County. We can see more accurately that South Texas and the Panhandle had more cases per thousand than the Texas Triangle.

Regions of Texas and COVID-19

Urbanization has significantly transformed the landscape of Texas, with distinct regions of the state experiencing varying degrees of development. It is important to understand these geographic distinctions with the data. The term “Texas Triangle” refers to the region that is covered by the four major metroplexes of Texas: Dallas-Fort Worth, Austin, San Antonio, and Houston [5]. More than 18 million people live in this region, and in terms of economic strength, each of the Texas Triangle metros is ranked among the top six strongest urban areas in the nation [6]. The Texas Triangle is known for its large urban areas, booming economies, and advanced healthcare systems. The Texas Medical Center in Houston, for example, is the largest medical complex in the world, with over 100,000 employees and 21 hospitals. The region also boasts several top-tier medical schools and research institutions, including the University of Texas Health Science Center at Houston

and Baylor College of Medicine. In general, the Texas Triangle region has a high concentration of healthcare facilities, physicians, and specialists, making it relatively easy for residents to access medical care. As of February 2023, the region has reported over 3.7 million cases of COVID-19 and more than 54,000 deaths. However, the region's relatively high concentration of healthcare providers has helped to ensure better access to medical care for residents.

In contrast, West Texas is a sparsely populated region with several small towns and rural communities. The region encompasses a vast area, with few large urban centers, and thus has limited access to medical care. As of February 2023, the region has reported over 101,000 cases of COVID-19 and over 1,200 deaths. Hospitals and healthcare providers in the region are concentrated in the larger cities, such as Lubbock and Midland-Odessa. However, these facilities may be far from residents in the more rural areas, making it difficult for individuals to receive timely and adequate medical attention. Rural hospitals in West Texas are particularly vulnerable to closure due to financial pressures, and this can exacerbate the region's already limited access to medical care.

South Texas, on the other hand, is an economically disadvantaged region that is also home to a predominantly Hispanic population. The region has several urban areas, including Laredo, Brownsville, and McAllen, but also has many rural and remote communities. The region has a high prevalence of chronic diseases, such as diabetes and obesity, and limited access to medical care, making its residents more susceptible to severe COVID-19 infections. As of February 2023, the region has reported over 536,000 cases of COVID-19 and more than 9,700 deaths. Many residents of South Texas lack health insurance or have limited access to primary care physicians, and there are few specialists in the region. The region's healthcare system faces unique challenges due to language and cultural barriers, as well as a shortage of healthcare providers.

Regions of California and COVID-19

The authors of this report live in northern Los Angeles County, so the state of California was also of interest when exploring different U.S. states. California is the most populous state in the country, so it would be beneficial to compare the two most populous states. California and Texas also had vastly different policies regarding their response to the pandemic [7]. California, the largest Democratic state, and Texas, its Republican rival, implemented vastly divergent strategies in their management of the pandemic. The effectiveness of these approaches holds great importance for the nation's well being, as approximately 20% of Americans reside in these two states.

The Los Angeles Metro Area, with a population of approximately 18.7 million people, is the most populous metropolitan area in California. This region has been hit hard by the COVID-19 pandemic, with more than 1.6 million confirmed cases and over 28,000 deaths as of February 2023. Despite having a high number of hospitals and healthcare providers, the region has struggled to meet the healthcare needs of its residents during the pandemic due to overwhelmed hospitals and staff shortages.

The Bay Area, which includes San Francisco, San Jose, and Oakland, has a population of approximately 8.7 million people. This region has a higher concentration of healthcare facilities and providers than many other areas in California, including top-ranked hospitals such as UCSF Medical Center and Stanford Health Care. The Bay Area has also been successful in implementing strict public health measures to contain the spread of COVID-19, resulting in lower infection and death rates compared to other regions.

The Central Valley, with a population of approximately 7 million people, is known for its agricultural production but also has some urban centers such as Fresno and Bakersfield. However, the region has relatively fewer healthcare facilities and providers than other regions in California, making it difficult for residents to access healthcare during the pandemic. The Central Valley has seen high rates of COVID-19 infections and deaths, with some areas experiencing among the highest infection rates in the state.

The Inland Empire, with a population of approximately 4.7 million people, is a metropolitan area located in Southern California, east of Los Angeles. The region has a mix of urban and suburban areas and has struggled to meet the healthcare needs of its residents during the pandemic due to a shortage of hospital beds and healthcare providers. The Inland Empire has seen high rates of COVID-19 infections and deaths, particularly in areas with large populations of low-income and minority residents.

Northern California, with a population of approximately 4.3 million people, includes the state capital of Sacramento as well as several other cities. The region has a mix of urban and rural areas and a relatively high concentration of healthcare facilities and providers, including the renowned UC Davis Medical Center. Northern California has also implemented stricter public health measures to contain the spread of COVID-19, resulting in lower infection and death rates compared to other regions. We performed the exact same search on the BigQuery dataset, substituting Texas for California.

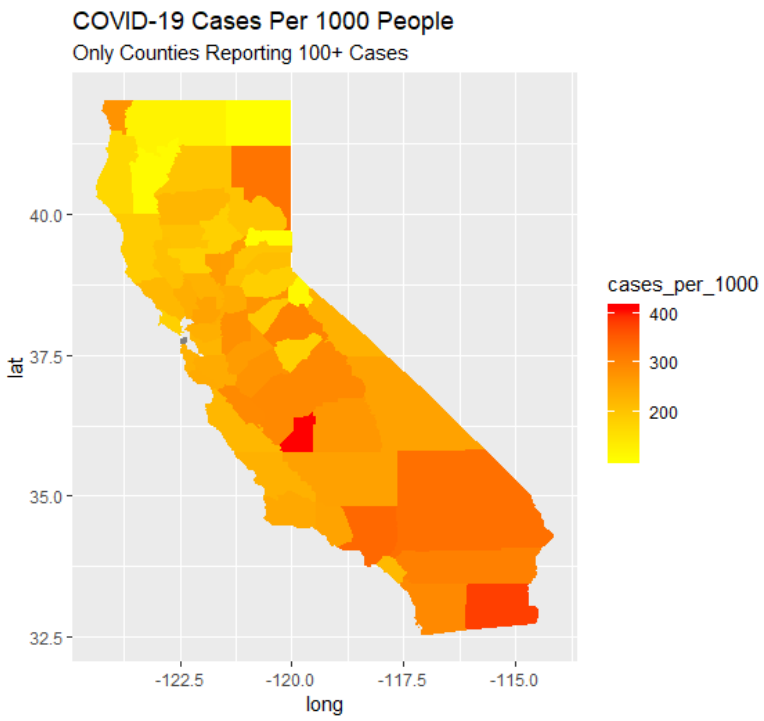


Figure 7. shows California counties' cases per 1000.

This graph has two interesting points: the two counties with the highest confirmed cases per thousand and the lack of data regarding San Francisco.

The counties with the highest cases per thousand are Kings County in the Central Valley and Imperial County in the southeast Inland Empire. Kings County has the highest incarceration rate of California's 58 counties due to the presence of several state prisons in the area [8]. This means that a large proportion of the population in Kings County is made up of incarcerated individuals, who are at higher risk of contracting and spreading COVID-19 due to close living quarters and limited access to healthcare. By February 2021, Avenal State Prison had more than 3,600 confirmed cases among prisoners and staff members, and the facility tops the list of the country’s largest COVID clusters in prisons compiled by The New York Times and the UCLA Covid-19 Behind Bars Data Project [8]. In a rural agricultural county, the presence of California’s largest prisons skew the testing data.

Imperial County is situated on the border with Mexico, and many residents travel back and forth between the two countries for work or other reasons, potentially increasing the risk of exposure to the virus. As with many counties in southern Texas, COVID cases are higher in regions with greater movement of people, especially across the Mexican border [9]. Additionally, the county has a large number of agricultural workers who live in close quarters and may have limited access to healthcare, which can also contribute to the spread

of the virus. Imperial County has a relatively high poverty rate and a significant population of low-wage essential workers who may be unable to work from home or access adequate healthcare, potentially putting them at higher risk of contracting and spreading COVID-19 [10]. Finally, Imperial County has experienced challenges in accessing adequate testing and healthcare resources, which may contribute to an undercounting of cases and a higher rate of severe illness and death [11].

A glaring omission to the California map is the city and county of San Francisco. The city and county are the same entity, as it is officially the City and County of San Francisco. As the fourth most populous city and 12th most populous county in California, it is not the case that there are no cases. As of 2020, San Francisco is the fourth densest county in the United States, only behind four of the boroughs of New York [12]. It is likely just a data entry error. However, sometimes data entry errors can occur, leading to incomplete or inaccurate data sets. San Francisco is a major metropolitan area and a hub for international travel, making it a potential hotspot for the virus. To understand the magnitude of this potential data entry error, we can look at the number of COVID tests administered in San Francisco. As of September 2021, the San Francisco Department of Public Health reported administering over 2.4 million COVID tests in the county [14]. This figure suggests that San Francisco has taken the pandemic seriously and has implemented measures to track and manage its spread. Furthermore, the city has reported over 80% of eligible residents receiving at least one dose of the vaccine as of February 2023 [14]. This high vaccination rate is a testament to San Francisco's commitment to public health and suggests that the city has a robust infrastructure for collecting and reporting COVID data. The same website reports that just under 200 thousand cases have been reported in San Francisco, with roughly 250 confirmed cases per thousand residents. This places San Francisco on equal footing with the rest of the Bay Area counties.

Additional graphs regarding coronavirus in California can be found in Appendix C.

Ethnicity and COVID-19

The COVID-19 pandemic has highlighted existing health disparities that disproportionately affect communities of color. The impact of COVID-19 has not been distributed equally, with Black and Hispanic communities experiencing higher rates of infection, hospitalization, and death compared to other ethnic groups.

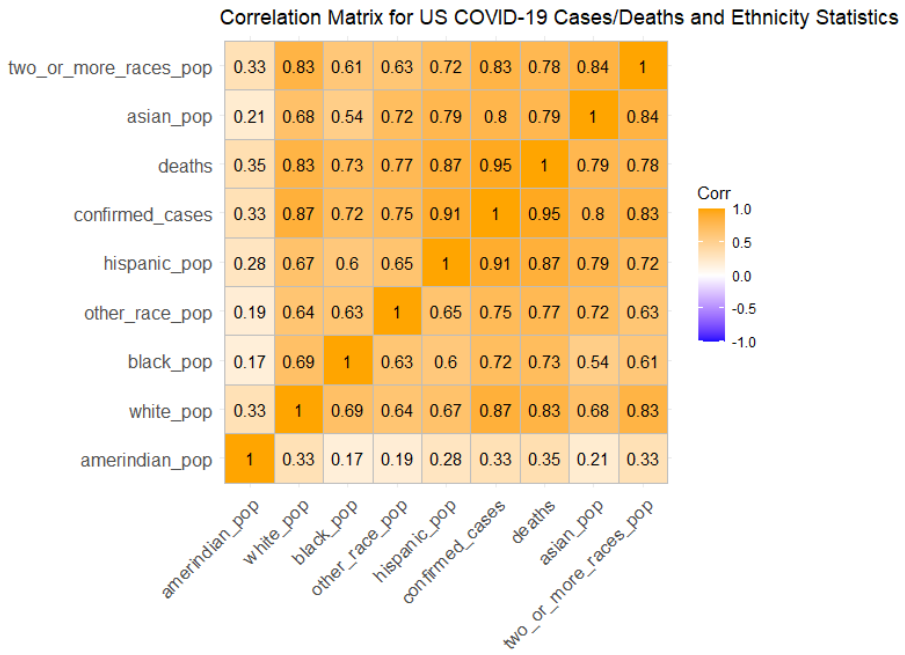


Figure 8. Shows a correlation matrix of races and confirmed cases.

The correlation matrix provided shows the relationship between COVID-19 deaths and different racial groups. A correlation coefficient is a statistical measure used to determine the strength and direction of a linear relationship between two variables. In this case, the variables are COVID-19 deaths or confirmed cases and race.

The scores range from -1 to 1, with -1 indicating a perfect negative correlation (as one variable increases, the other decreases) and 1 indicating a perfect positive correlation (as one variable increases, so does the other). A score of 0 indicates no correlation between the variables.

The best cells to look at are the horizontal “deaths” and “confirmed cases”. In terms of the highest correlation to deaths, the data shows Hispanic (.87), White (.83), Asian (.79), Two or more (.78), Other races (.77), Black (.73), and Amerindian (.35). For highest correlation to confirmed cases, the data shows Hispanic (.91), White (.87), Two or more (.83), Asian (.80), Other races (.75), Black (.72), and Amerindian (.33).

The highest correlation score in the matrix is between Hispanic ethnicity and COVID deaths and confirmed cases, with coefficients of .87 and .91. This suggests that there is a strong positive linear relationship between the number of COVID cases and deaths and the Hispanic population. Similarly, the high correlation scores for White, Asian, Two or more, and Other races suggest that these groups are also strongly positively correlated with COVID deaths.

The lowest correlation scores in the matrix are for Amerindian ethnicity, with coefficients of .35 and .33. This suggests that there is a weak positive linear relationship between the number of COVID cases and deaths and the Amerindian population. While this correlation is still statistically significant, it is not as strong as the correlations for other racial groups. One possible explanation for the low correlation is that Amerindians may be more dispersed across rural areas where the spread of COVID may be slower, which could lead to a lower number of confirmed COVID cases and deaths. Additionally, the data could be incomplete or inaccurate due to underreporting or insufficient testing among Amerindian populations.

It is important to note that correlation does not imply causation. These correlation scores do not prove that race is a direct cause of COVID deaths. Other factors, such as age, pre-existing health conditions, and access to healthcare, may also play a significant role in COVID deaths.

Graphs displaying the relationship between populations of each race and death per county can be found in Appendix B.

Dallas County Timeline

As with any pandemic, time is an important factor. A good way to measure the effectiveness of a certain measure is to observe both the confirmed case count and death count in the following months after the measure is implemented. We decided upon five major turning points in Dallas County: social distancing, the first vaccine release, both FDA approvals, and the authorization of the vaccine for children.

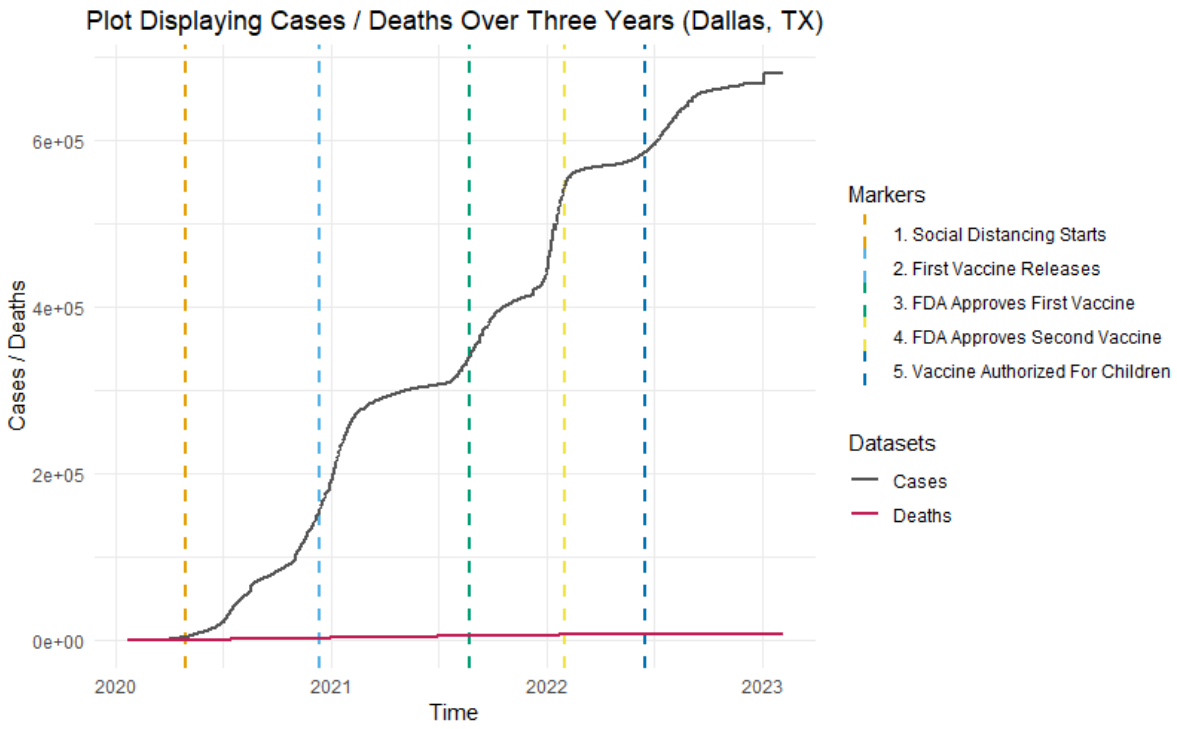


Figure 9 showing cases and deaths over time in Dallas County. Specific health measures are indicated.

The first case of COVID-19 was reported in Dallas County in March 2020, and by April, there were over 1,000 confirmed cases in the county. In July 2020, the number of cases surged, leading Dallas County to implement a mandatory mask mandate and limit indoor gatherings. The graph above shows that social distancing did work and continues to work in conjunction with the release of vaccines to the public. Analyzing the graph, we can clearly see periods along the timeline of the past couple of years in which the curve had been flattening, and when the cases began to skyrocket, some vaccines and other technologies were released to counteract the rise of cases. For example, throughout the fall and winter of 2020, Dallas experienced several spikes in cases, leading to additional restrictions on businesses and public gatherings. In early 2021, vaccine distribution began, and by summer, cases had significantly decreased. However, in the fall and winter of 2021, the Delta variant caused another surge in cases, leading to renewed efforts to promote vaccination and implement public health measures. As of early 2023, the curve in Dallas seems to have flattened. COVID restrictions have largely been relaxed, and according to the New York Times, 61% of the Dallas County population is fully vaccinated, and 22% has a booster [13].

Conclusions

The COVID-19 pandemic has impacted the world in unprecedented ways, with governments and healthcare organizations grappling to contain its spread and minimize its impact on the population. The aim of our report was to understand the pandemic's impact in Texas and identify trends that could inform targeted interventions. The project found that South Texas and the Panhandle were among some of the regions hardest hit by the pandemic.

We analyzed data from various sources, including COVID-19 cases and mortality rates, hospitalization rates, and vaccination rates. It was found that the regions with the highest case and mortality rates were South Texas and the Panhandle. These regions also had lower vaccination rates compared to other regions in Texas.

However, the project also found that social distancing measures and the introduction of vaccines helped "flatten the curve" in Texas. As more people received vaccinations, the number of cases and hospitalizations decreased, providing hope for a return to normalcy.

Interestingly, the analysis found that California and Texas had similar trends, with both states experiencing spikes in cases and deaths at similar times. This finding suggested that the pandemic's impact was not solely influenced by state policies but also by other factors such as population density and behavior.

To continue the fight against COVID-19, it is crucial to take a data-driven approach and monitor trends closely. It is recommended to continue following CDC guidelines such as wearing masks, practicing social distancing, and getting vaccinated. Vaccinations have proven to be effective in reducing the spread of the virus and preventing severe illness and death.

The use of data analytics and monitoring tools can help identify areas that need additional resources and support. Governments and healthcare organizations should leverage these tools to inform decision-making and respond promptly to changes in trends. Comparing COVID-19 trends in California to Texas showed that the pandemic's impact was not solely influenced by state policies but also by other factors. To continue the fight against COVID-19, a data-driven approach should be taken, and targeted interventions should be implemented to prevent outbreaks in regions with higher case

and mortality rates. The use of data analytics and monitoring tools can help inform decision-making and respond promptly to changes in trends.

References

[1] “Who coronavirus (COVID-19) dashboard,” *World Health Organization*. [Online]. Available: https://covid19.who.int/. [Accessed: 28-Feb-2023].

[2] “Guidance and tips for tribal community living during COVID-19,” *Centers for Disease Control and Prevention*, 23-Aug-2022. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/community/tribal/social-distancing.html. [Accessed: 28-Feb-2023].

[3] By, “Coronavirus in the U.S.: Latest Map and case count,” *The New York Times*, 03-Mar-2020. [Online]. Available: https://www.nytimes.com/interactive/2021/us/covid-cases.html. [Accessed: 28-Feb-2023].

[4] “U.S. Census Bureau quickfacts: Loving County, Texas,” *United States Census Bureau*. [Online]. Available: https://www.census.gov/quickfacts/fact/table/lovingcountytexas/PST045221. [Accessed: 28-Feb-2023].

[5] J. H. Cullum Clark, “The Texas Triangle: A rising megaregion unlike all others,” *George W. Bush Presidential Center*, 19-May-2021. [Online]. Available: https://www.bushcenter.org/publications/the-texas-triangle-a-rising-megaregion-unlike-all-others. [Accessed: 28-Feb-2023].

[6] “Texas triangle,” *Austin Capital Advisors*. [Online]. Available: https://www.austincapitaladvisors.com/texas-triangle. [Accessed: 28-Feb-2023].

[7] “America's two largest states are fighting covid-19 differently,” *The Economist*, 06-Feb-2021. [Online]. Available: https://www.economist.com/united-states/2021/02/06/americas-two-largest-states-are-fighting-covid-19-differently. [Accessed: 28-Feb-2023].

[8] K. Klein, “Lessons from kings county prison where covid-19 'spread like wildfire',” *KVPR*, 24-Feb-2021. [Online]. Available: https://www.kvpr.org/health/2021-02-24/lessons-from-kings-county-prison-where-covid-19-spread-like-wildfire. [Accessed: 28-Feb-2023].

[9] “How covid-19 has impacted the southern border region,” *Southern Border Communities Coalition*. [Online]. Available: https://www.southernborder.org/covid_test_page. [Accessed: 28-Feb-2023].

[10] G. Solis, “Imperial county has highest rate of COVID-19 cases in the state; it wants to reopen anyway,” *The San Diego Union-Tribune*, 14-Jun-2020. [Online]. Available: https://www.sandiegouniontribune.com/news/story/2020-06-14/imperial-county-has-highest-rate-of-covid-19-cases-in-the-state-it-wants-to-reopen-anyway. [Accessed: 28-Feb-2023].

[11] M. Foster, “Can Imperial County handle coronavirus when it's already struggling to fight tuberculosis?,” *UC Berkeley Graduate School of Journalism*, 08-Apr-2020. [Online]. Available: https://journalism.berkeley.edu/projects/imperial-county-tuberculosis/. [Accessed: 28-Feb-2023].

[12] Alex, “U.S. population density mapped,” *Vivid Maps*, 16-Jan-2022. [Online]. Available: https://vividmaps.com/us-population-density/. [Accessed: 28-Feb-2023].

[13] The New York Times, “Dallas County, Texas Covid case and exposure risk tracker,” *The New York Times*, 27-Jan-2021. [Online]. Available: https://www.nytimes.com/interactive/2021/us/dallas-texas-covid-cases.html. [Accessed: 28-Feb-2023].

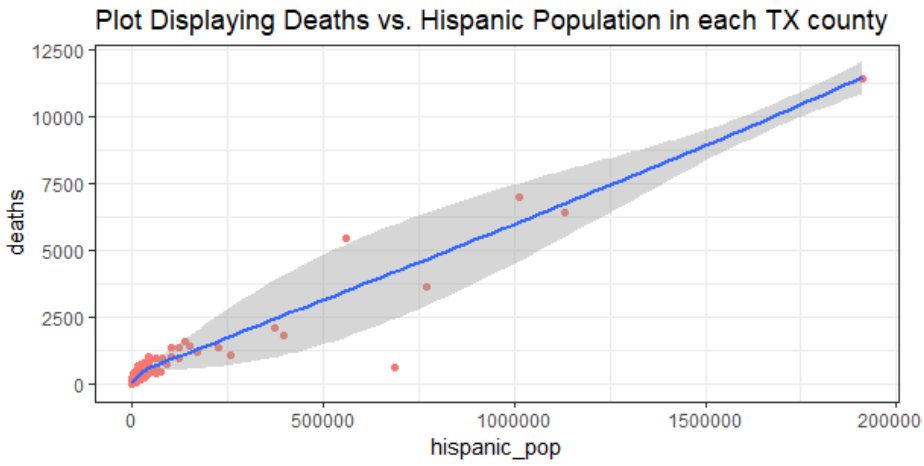
[14] “COVID-19 data and reports,” *COVID-19 data and reports | San Francisco*. [Online]. Available: https://sf.gov/resource/2021/covid-19-data-and-reports. [Accessed: 28-Feb-2023].

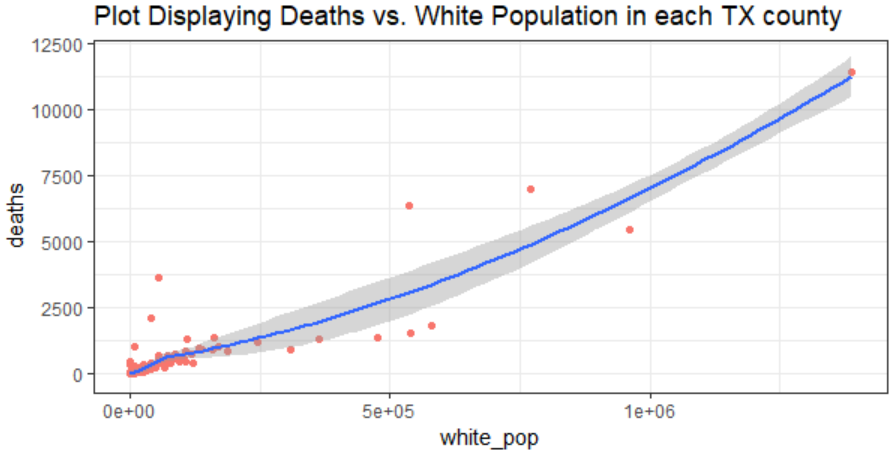
Appendix

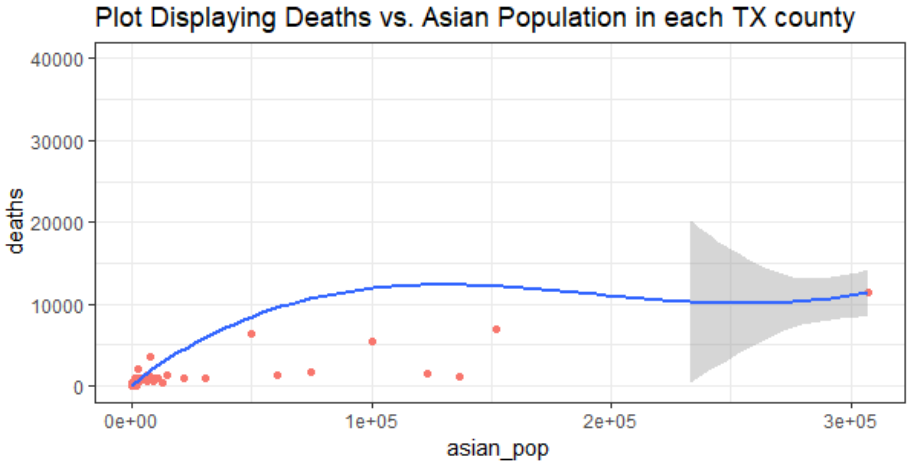
Appendix A: Feature Ranking Code Excerpt

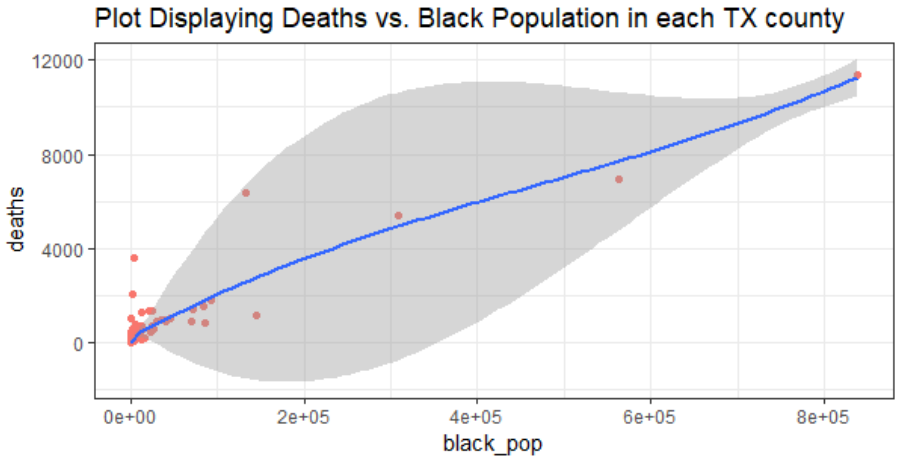
```
# Make Character Factors, Filter TX

cases <- COVID_19_cases_plus_census %>% mutate_if(is.character, factor)

dim(cases)

cases_TX <- COVID_19_cases_plus_census %>% filter(state == "TX")

dim(cases_TX)

summary(cases_TX[, 1:10])


# Feature Ranking (After Factorizing)

# Note: First Transform - Curse Of Dimensionality Example

transform_census    <-    as.data.frame(sapply(COVID_19_cases_plus_census,
as.numeric))

transform_census <- transform_census %>% select_if(~ !any(is.na(.))) %>%
select(-c(date, do_date))

cor_census <- cor(transform_census[, -1])

high_cor <- findCorrelation(cor_census, cutoff = 0.99995)

colnames(transform_census)


# Good Example Of Feature Extraction

control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

model <- train(confirmed_cases~., data = transform_census, method = "lm",
preProcess = "scale", trControl = control)

importance <- varImp(model, scale = FALSE)

print(importance)
```

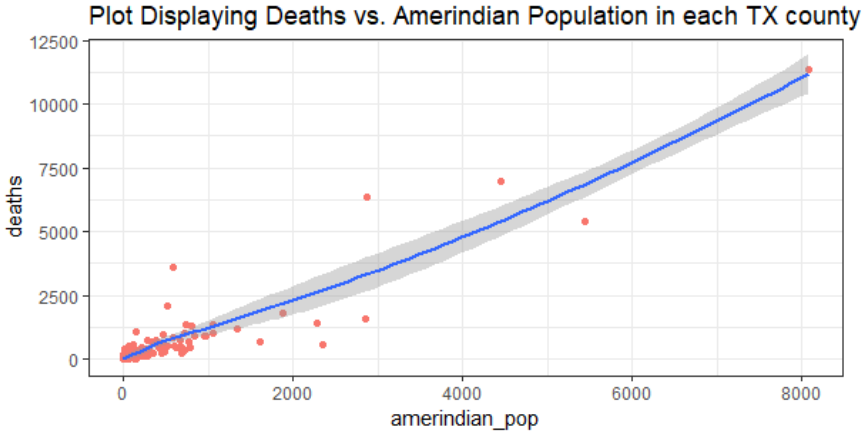
Appendix B: Race vs. COVID-19 Deaths - Texas





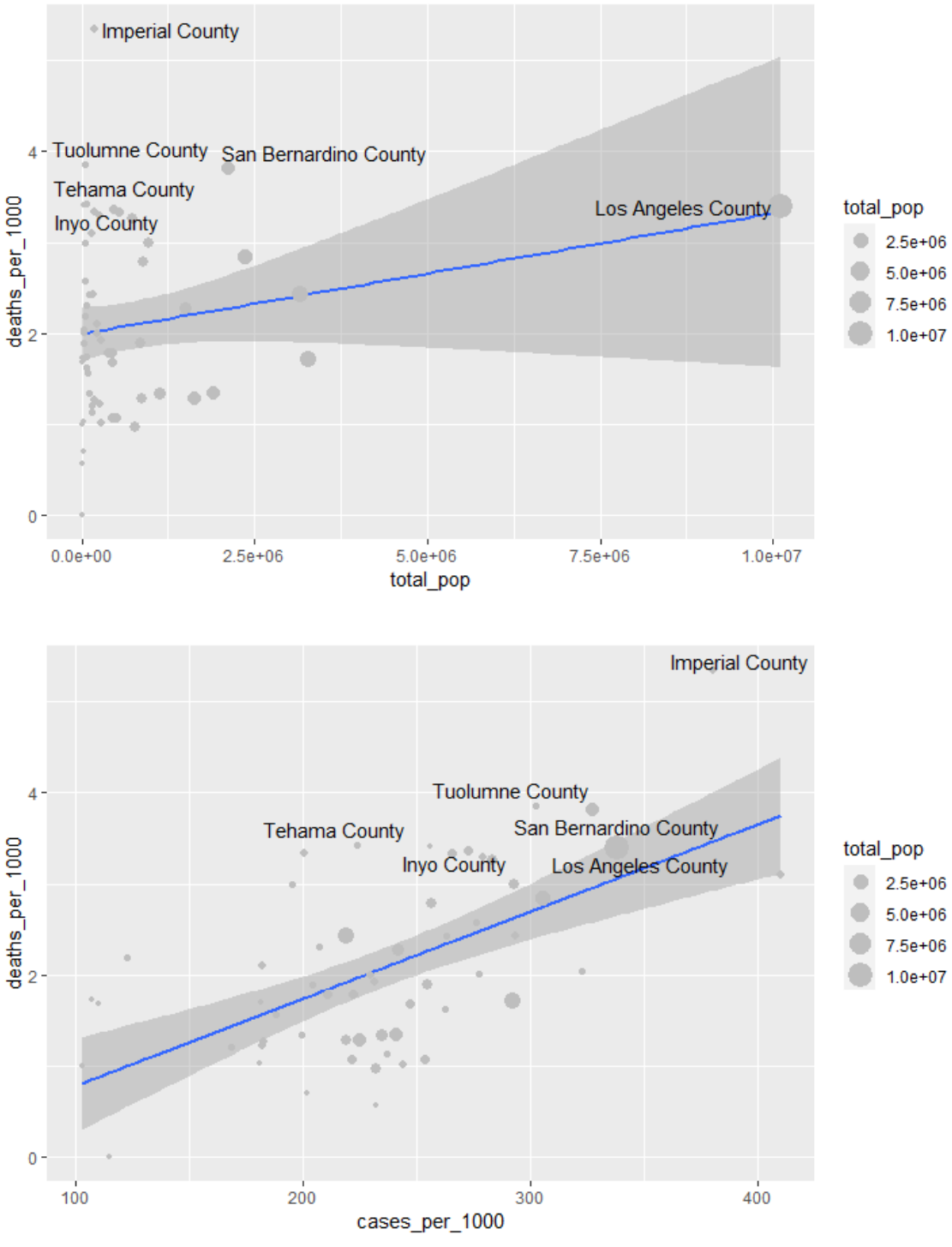






Note that scales in each graph are not uniform.

Appendix C: COVID-19 County Deaths Plots - California



Supplementary California data.