Data Mining Project 1:

COVID-19's Impact on Future Decisions for the State of Texas

Sreshta Ghosh, Muskaan Mahes, and Ridhi Ralli

Executive Summary

Due to the COVID-19 pandemic, many policymakers and public health officials had to make critical decisions to mitigate the virus's impact and efficiently allocate resources. The spread of the virus had severe consequences which varied significantly across Texas counties, influenced by demographic, economic, and healthcare factors. The goal of the analysis is to understand how COVID-19 data can inform future public health funding and policymaking in Texas. Furthermore, this report focuses on the relationship between the state, counties, confirmed cases, deaths, total population, median income, median age, household structures, and different fields of employment. There are several features that were selected and analyzed to determine their impact on COVID-19 outcomes at the county level. The analysis reveals that urban counties had higher case numbers and deaths, while rural counties faced accessibility challenges. Additionally, counties with lower median income and higher employment were more vulnerable to the virus, as also household structure influenced transmission. Counties with a stronger healthcare workforce showed better response demonstrating the importance of allocating healthcare resources. By leveraging data-driven strategies policymakers can enhance public health response and resource distribution for future health challenges.

Table of Contents

1) Problem Description (Business Understanding)

DESCRIBING THE PROBLEM

What is COVID-19?

COVID-19 is a transmissible disease caused by the SARS-CoV-2(severe acute respiratory syndrome coronavirus 2) virus which emerged in December 2019 (John Hopkins). Structurally, coronavirus is characterized by a large RNA genome and spikes that project from the surface of its envelope. Even though, the virus was first identified in 1965, the human virus (HCoV) has become more prominent because of their role in the global outbreak. These strains: HCoV-229E, HCoV-NL63, HCoV-OC43, and HCoV-HKU1 occur in humans and cause is responsible for 15-30% of the common cold in adults and children. There are many different strains of COVID-19 which causes diseases and sometimes death. Those who are infected with COVID-19 will experience mild symptoms, and others have no symptoms at all. COVID-19 can cause respiratory problems, lasting lung and heart muscle damage, nervous system problems, kidney failure, and even death (John Hopkins). The virus is primarily transmitted between individuals. When someone coughs, sneezes, breathes, or even talks their breath can carry infectious droplets or particles from an affected person, especially in close proximity and areas with low air flow. Additionally, if you touch a surface contaminated with respiratory droplets and then touch yourself, you may infect yourself. There are serious risk factors for people over the age of 65 as well as babies that are younger than 6 months if they contract the virus. Those at higher risk have the chance of developing sickle cell disease, heart disease, high blood pressure, and chronic kidney, liver, and lung disease. Additionally, there are complications of COVID-19 once contracted such as loss of tase and smell, skin rashes, pneumonia, acute respiratory distress syndrome, shock caused by the infection, and blot clots (Mayo Clinic, 2024).

What is social distancing and flattening the curve?

Social distancing and flattening the curve were a strategy and a goal implemented to mitigate and reduce the spread of COVID-19. Social distancing refers to measures taken to reduce the close contact between people to slow the spread of the virus. The rules included maintaining at least 6 feet away from others, avoiding large gathering or crowded places, transitioning to remote work, and wearing masks when going in public areas.

Flattening the curve refers to the goal to reduce the number of total COVID-19 cases over time. During the beginning of the pandemic there was a spike in the number of cases which led to hospitals being overwhelmed, shortages of medical staff, equipment, and resources. To address this, efforts to flatten the curve were put in place to slow the spread of cases and bring them back to a manageable level. Social distancing was one of the strategies that helped with flattening the curve by reducing the transmission.

Why is it important to look at data about the virus spread, hospitalizations, and available resources?

Understanding data on COVID-19 cases, hospitalizations, and available resources is crucial for managing the pandemic effectively. Tracking how the virus spreads help public health officials predict outbreaks and take preventive actions, such as implementing social distancing or mask mandates (CDC, 2021). Monitoring hospital admissions ensures that healthcare systems don’t become overwhelmed, as seen in Italy’s early struggles when hospitals ran out of ICU beds. Reliable data also assists governments in making informed policy decisions, as demonstrated by how New Zealand used real-time case numbers to enforce lockdowns and control the virus. Keeping track of medical resources, including hospital beds, ventilators, and vaccines, ensures they are distributed fairly, especially in communities at higher risk (WHO, 2021). Understanding which populations are most vulnerable—such as older adults and people with limited healthcare access—helps design targeted health programs to reduce severe cases and deaths (CDC MMWR, 2021). Additionally, making accurate data publicly available aids in countering misinformation and guiding people to take necessary precautions, as shown by the widespread use of the Johns Hopkins COVID-19 Dashboard. Without a strong focus on data, responses to the pandemic could be delayed or ineffective, leading to more severe outbreaks, hospital shortages, and long-term social and economic challenges.

STAKEHOLDER

Short description of the problem area.

The problem area that we are focusing on are the critical gaps in public health infrastructure, resource allocation, and emergency response capabilities that were exposed during the COVID-19 pandemic. By analyzing case trends, healthcare capacity, and demographic impacts we can inform future public health funding, resource allocation, and policymaking for future health crises.

Who is your stakeholder and what does the stakeholder want?

Our stakeholder is the Commissioner of the Texas Department of State Health Services. Their role is to promote the health of the population by enforcing health laws, monitoring virus trends, educating the public on health issues, and advocating for health policies. (PMC, 2007) Therefore, public health officials are important because after monitoring resources they can allocate them (vaccines, oxygen tanks, hospital beds, etc.) efficiently. Additionally, they can identify populations that need more attention by targeting interventions.

Define some questions that are important for this stakeholder.

Our main question is:

- How can COVID data inform future public health funding and policymaking for the state of Texas?

Additional questions:

- How many people reported have died from COVID in the state of Texas?
- Based on these deaths, where are the most outbreaks occurring, and what factors contribute to their spread?
- Which demographics are most affected, and how should interventions be tailored?
- Where should resources like vaccines and medical staff be distributed?

Understanding how key factors affect the spread and severity of the disease is essential for shaping future public health funding and policy decisions. Using these relationships and questions, we can identify the areas of greatest need and allocate resources effectively.

What decisions can your stakeholders make, and how would they affect COVID-19 outcomes?

The Commissioner of the Texas Department of State Health Services uses COVID-19 data to make critical decisions that impact public safety and healthcare management. One key decision is determining when to enforce or relax restrictions like mask mandates, social distancing rules, and lockdowns. By analyzing case trends and compliance levels, they can adjust measures to slow virus spread without causing unnecessary disruptions. Additionally, hospitalization and demographic data help officials allocate healthcare resources efficiently, ensuring that medical supplies, hospital funding, and personnel are directed to the areas with the highest needs. Public health messaging and campaigns also rely on this data, allowing officials to identify which populations have lower compliance rates or vaccination uptake and tailor outreach strategies accordingly. Another crucial decision involves prioritizing vaccine and treatment distribution by identifying high-risk groups, such as the elderly and immunocompromised, to reduce severe cases and fatalities. Finally, studying past trends in virus spread, the effectiveness of social distancing, and hospital capacity helps officials prepare for future waves and pandemics, strengthening healthcare systems and emergency response plans.

These decisions significantly impact COVID-19 outcomes by shaping policies, improving resource allocation, and implementing targeted health interventions. Effective policymaking, such as enforcing timely lockdowns or mask mandates, directly reduces virus transmission by limiting high-risk interactions. Data-driven policies also ensure that restrictions are lifted at the right time, preventing premature reopening that could lead to case surges. Better resource allocation ensures that hospitals and healthcare systems are not overwhelmed by directing medical supplies, staff, and funding to areas with high infection rates or hospitalizations. This prevents critical shortages of ICU beds and ventilators, ultimately improving patient survival rates. Additionally, targeted health interventions, such as prioritizing vaccines for high-risk populations and launching outreach programs in communities with low compliance, help reduce severe cases and fatalities. Public health campaigns based on demographic data also increased awareness and encouraged safer behaviors, further slowing the virus spread. By making informed, data-driven decisions, public health officials can mitigate outbreaks, protect healthcare capacity, and minimize long-term social and economic disruptions.

2) Data Understanding

The datasets we are using includes COVID-19_cases_plus_census, COVID-19_cases_TX, and Global_Mobility_Report. Our data is reliable as it was taken/sampled in a concise and detailed manner. Each dataset provided was taken from the Google Cloud Platform which includes different datasets from various sources such as the CDC and from other public health sectors of other countries in the world.

Describe what Data is Available

Table 1: Important Variables' Description

Variable (Cleaned Names)	**Scale of Measurement**	**Description**
State	Nominal	The U.S. state (Texas) chosen for where the data was collected.
County	Nominal	The counties in Texas where the data was recorded.
Confirmed_Cases	Nominal	The total number of reported COVID-19 cases in a county.
Deaths	Nominal	The total number of reported deaths due to COVID-19 in a county.
Total_Population	Nominal	The overall population of a county.
Median_Income	Nominal	The average household income in a county.
Median_Age	Nominal	The average age of the population in the county.
Employed_Edu_Health	Nominal	The number of people employed in the education, healthcare, and social assistance fields of a county.

Employed_Science_Admin	Nominal	The number of people employed in the science, management, administrative, and waste management fields of a county.
Nonfamily_Households	Nominal	The number of households in a county where the residents are not related by family ties (example: individuals living alone or with roommates).
Family_Households	Nominal	The number of households in a county where residents are related by family ties. (example: nuclear families, extended families).

After analyzing the three datasets provided, these variables were carefully chosen to capture both the public health impact of the pandemic and the broader socioeconomic context of each county within Texas. All 11 variables included in Table 1 are measured on a nominal scale. The variables include **State** and **County**, which identify the geographic scope of the data; **Confirmed_Cases** and **Deaths**, which quantify the health impact of COVID-19; and **Total_Population**, offering insight into the size of each county. Additionally, **Median_Income** and **Median_Age** shed light on the economic and demographic characteristics of the population. Employment data is represented through **Employed_Edu_Health** and **Employed_Science_Admin**, indicating the number of people working in essential service fields that play a critical role in community response and resilience. Finally, the distinctions between **Nonfamily_Households** and **Family_Households** provide a window into household structures that may influence both the transmission dynamics of the virus and the social support networks available in a community. Collectively, these variables form a multifaceted dataset that supports analyses ranging from epidemiological modeling to policymaking and resource allocation, offering valuable insights into both the health crisis and the underlying social fabric of Texas counties.

Verify Data Quality

The code first verifies data quality by checking for missing values and duplicate rows. Using colSums(is.na(cases_new)), we confirmed that there are 0 missing values across all 11 variables, and by identifying duplicates with the duplicated() function, we found 0 duplicate rows. For outlier detection, the code uses a custom function along with the interquartile range (IQR) method on all numeric columns grouped by county. It calculates Q1, Q3, IQR, and then defines lower and upper bounds to filter out any values outside these limits. The outlier check returns NULL for the State and County variables because these are categorical (nominal) and not subject to numeric outlier analysis. However, for the remaining 9 numeric variables—Confirmed_Cases, Deaths, Total_Population, Median_Income, Median_Age, Employed_Edu_Health, Employed_Science_Admin, Nonfamily_Households, and Family_Households—the code identifies several outliers. Despite this, we opted not to remove these outliers because they represent key important data points; eliminating them could potentially skew our analysis. In summary, the quality checks confirmed that we retained all 3142 objects from the original dataset. However, once filtered for Texas, our working dataset consists of 254 objects across the 11 variables, ensuring that we have a complete and robust dataset for further analysis.

Statistics for the Most Important Variables

Table 2: Statistical Summary of Important Variables

Variables	**Data Type**	**Mean**	**Median**	**Mode**	**St. Dev**	**Variance**	**Min**	**Max**	**Range (Max-Min)**
State	Character	------	---------	TX	-----	----------	-------	----------	----------
County	Character	------	---------	Anderson County	---------	----------	-------	----------	----------
Confirmed_Cases	Double	8419	1310	988	29570.38	874407557	1	286356	286355
Deaths	Double	127.48	30	20	382.1727	146056	0	3825	3825
Total_ Population	Double	107951	18612	57747	389476.9	151692226474	74	4525519	4525445
Median_Income	Double	49894	48311	44601	12132.68	147201815	24794	93645	68851
Median_Age	Double	39.02	38.55	35.4	5.965956	35.59263	25.80	57.50	31.7
Employed_Edu_Health	Double	10784.3	1468	1074	37024.44	1370809146	2	408677	408675
Employed_Science_ Health	Double	5660.3	377.5	88	24945.42	622273825	0	286766	286766
Nonfamily_	Double	11300	1883	420	43496.67	1891959	12	496164	496152

Households						882			
Family_ Households	Double	25828	4474	11611	91572.7	8385560068	19	1066649	1066630

The statistical summary of key variables provides demographic, economic, and COVID- 19 data at the state and county level, as shown in Table 2. The State and County variables are of character type, while all other variables are of type double. Unlike the numerical variables, the State and County variables do not have detailed statistical summaries beyond the mode, which indicates that the most frequently occurring state is Texas (TX) and the most common county is Anderson County. The confirmed cases variable has an average of 8.419 cases per county, with a median of 1,310 and a mode of 988. This indicates that even though some counties report much higher cases there are many that have lower case counts. Additionally, there is a maximum of 286,356 cases, and the minimum is 1, resulting in a range of 286,355 cases. The variance is 874,407,557 which indicates a spread in case numbers across counties. For the deaths variable it follows a similar pattern as the average is 127.38 deaths per county, with a median of 30 deaths, therefore, this shows that the counties have relatively low death counts. Additionally, there is a mode of 20 deaths which is close to the median. The standard deviation is 382.17 and the variance is 146,056 this indicates the variability in death counts.

The demographic variables include total population which varies across counties. Its mean population per is 107,951, and the median is 18,162. This indicates that the distribution is skewed as most counties population is much lower than the mean. The maximum county population is 4,525,519, and the smallest county has just 74 residents. Furthermore, the median income across the counties averages $49,894 and has a median of $48,311. Additionally, there is a mode of $44,601 which suggest that the income levels are stable across most counties. However, there is a relatively huge range between the minimum median income of $24,794, and the maximum being $93,645 suggesting economic disparities between the counties. The median age in counties average at 39.02 years, with a low standard deviation of 5.97 this indicates that most counties have similar age distribution. The youngest county has a median age of 25.8 years, and the oldest county's median age is 57.7 years, which depicts variation.

The employment data highlights the distribution of workers across different fields. There are 10,784 people who are employed in the education and health sectors, with a median of 1,4688 and a mode of 1,074. Additionally, there is a wide range between a minimum of 2 and maximum of 408,677, suggesting that some counties have higher employment rate than in other fields. Similarly, the employed science and healthcare variable has a mean of 5,660.3 people per county, and a median of 377.5. There is a similar pattern within these employment sectors with concentrations in certain areas. Within the household data it reveals disparities in family and nonfamily living arrangements. The nonfamily household variable has an average of 11,300 per county, with a median of 1,883 and a mode of 420.; therefore, this indicates that smaller counties have fewer nonfamily households. Additionally, there is a maximum of 496,164 households per county, and a minimum of 12. In contrast, family households have a higher average of 25,828 per county, with a median of 4,474 and a mode of 11,611. The maximum family households are 1,066,649 which emphasizes the variation in households' structure between counties.

In summary, the COVID-19 dataset demonstrates high variability in population size, economic conditions, employment distribution, and COVID-19 impact across counties. There is a difference in means and median in many variables, which suggest that the distributions are skewed due to populated counties which influences the overall statistics.

Visually Explore the Chosen Attributes Appropriately

Below we created histograms, scatter plots, bar graphs, and maps to visually explore our selected attributes and relationships between the attributes to understand out data better. This allowed us to identify patterns, trends, and potential relationships within the data to understand where funding and resource allocation should occur. However, upon reviewing these visualizations, we did not observe significant insights related to our primary research question in variables such as State, County, Confirmed Cases, Deaths, Median Income, and Median Age. Below, we provide a detailed analysis of the key attributes that we identified as most relevant to our study.

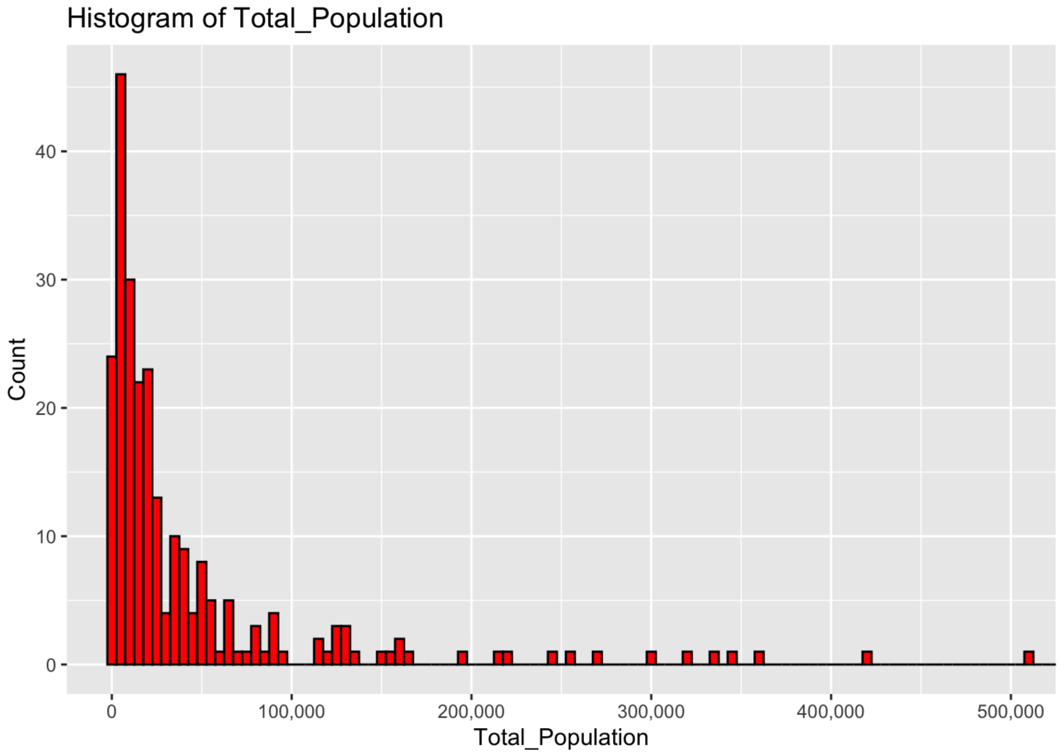
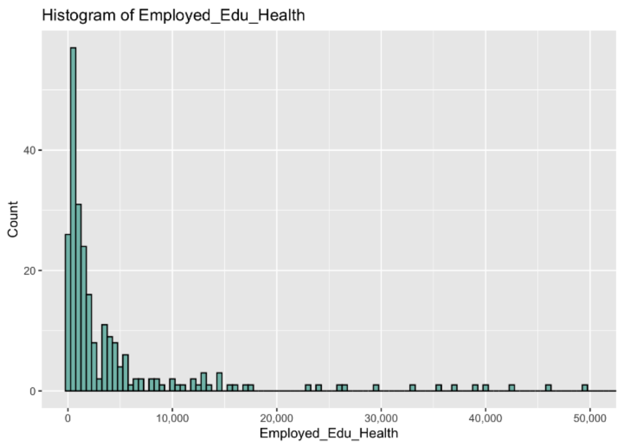


Figure 1: Histogram of Total Population

Figure 1 illustrates the distribution of Total_Population across Texas counties, showing a pronounced right skew. Most counties have relatively small populations, clustered toward the lower end of the histogram, while a handful of counties have populations exceeding 100,000 and even 500,000. From a COVID-19 standpoint, these large-population counties are likely to face higher case numbers and potentially greater demand for healthcare resources, making them key focal points for pandemic response measures. Meanwhile, smaller counties may have fewer medical facilities and less robust healthcare infrastructure but also face fewer overall cases. A histogram is an appropriate choice here because it visually highlights the frequency of different population sizes, helping public health officials and policymakers identify potential hotspots (larger counties) as well as areas that may need additional support due to limited resources.



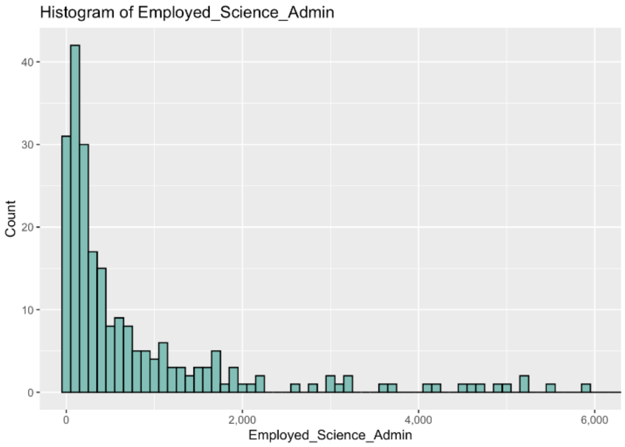
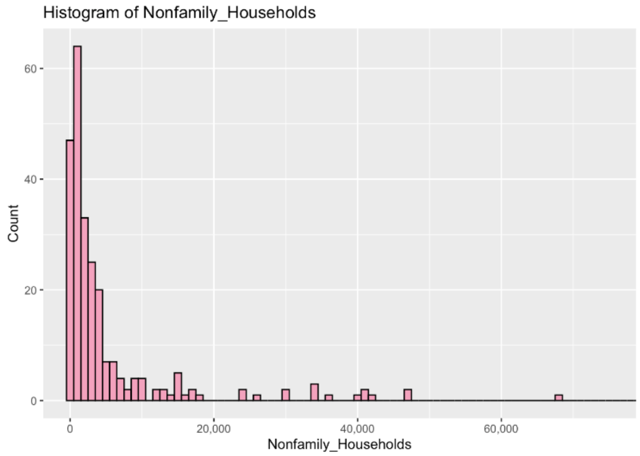


Figure 2: Histogram of Employed Edu Health

Figure 3: Histogram of Employed Science Admin

Placing Figures 2 (Employed_Edu_Health) and 3 (Employed_Science_Admin) side by side reveals that both distributions are heavily right-skewed, with most Texas counties showing relatively small employment counts and only a few counties having significantly higher numbers. However, Employed_Edu_Health spans a larger range—reaching up to around 50,000 employees—compared to Employed_Science_Admin, which tops out closer to 6,000. This contrast highlights that education, healthcare, and social assistance positions generally encompass a larger workforce than science, management, and administrative fields. From a COVID-19 perspective, counties with high Employed_Edu_Health counts may face more immediate frontline challenges, such as staffing schools and healthcare facilities during surges, whereas higher Employed_Science_Admin employment could signal stronger administrative capacities and potential for remote work flexibility. The histogram format is well-suited for showing these distributions side by side, making it easy to compare where each type of employment clusters and how a few outlier counties dominate the upper range of both categories.



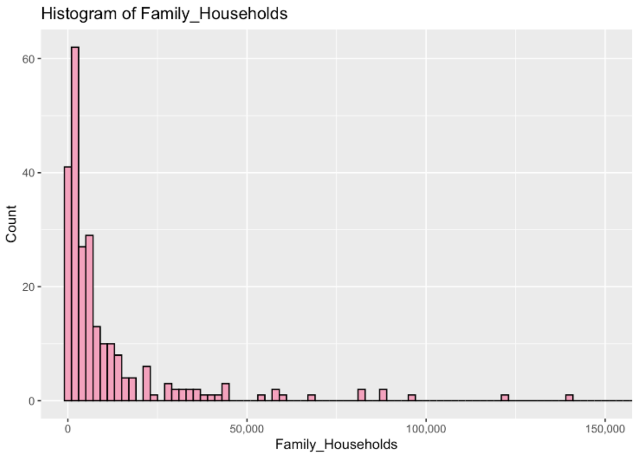


Figure 4: Histogram of Nonfamily Households

Figure 5: Histogram of Family Households

Placing Figure 4 (Nonfamily_Households) and Figure 5 (Family_Households) side by side highlights that both distributions are strongly right skewed, with most Texas counties having relatively small household counts and only a few counties reaching much higher values. While nonfamily households top out around 60,000, family households can exceed 100,000, suggesting that traditional family units make up a larger share of the household landscape in certain populous areas. From a COVID-19 standpoint, counties with more family households might face different transmission risks and support needs, given larger or multigenerational living arrangements, while those with higher nonfamily households may have different social dynamics—such as single-resident or roommate situations—that also affect virus spread and community resource requirements. The histogram format clearly shows where most counties cluster and where outliers exist, helping policymakers and public health officials tailor strategies to the unique household compositions within each region.

Explore Relationships Between Attributes

Understanding the impact of COVID-19 requires analyzing how different attributes interact with one another. In the following section, we explore the relationship between different attributes that we chose. This was used to identify potential trends and correlations. By visualizing and analyzing these relationships, we aimed to uncover patterns that could inform public health strategies and policy decisions.

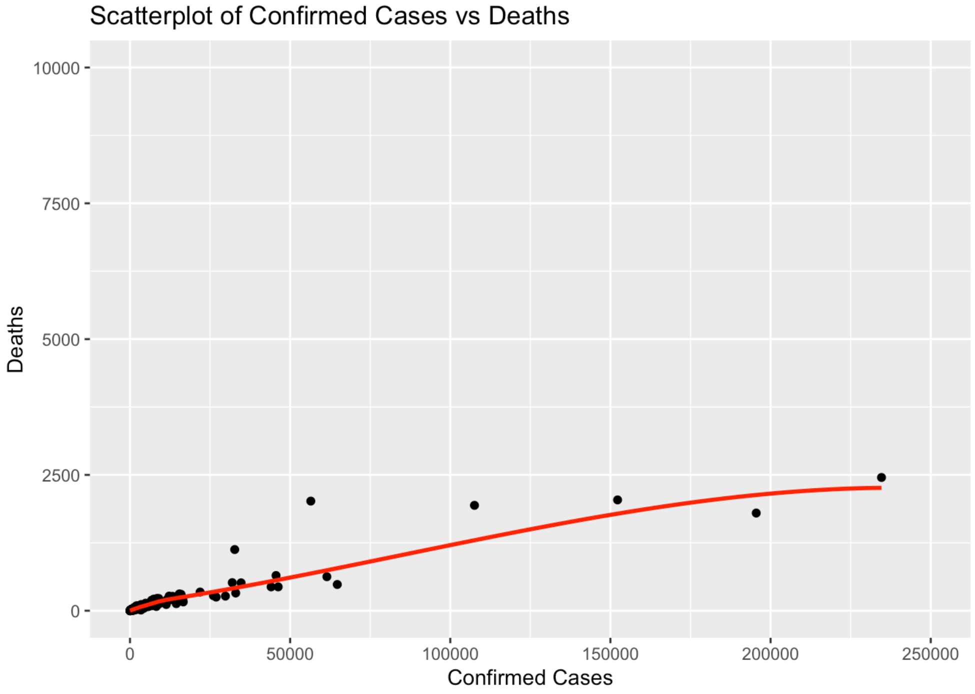


Figure 6: Confirmed Cases vs. Deaths

Figure 6 presents a scatterplot of Confirmed Cases (x-axis) versus Deaths (y-axis), along with a fitted trend line. As expected, there is a positive relationship: counties reporting more confirmed cases also tend to report higher numbers of deaths. While most data points cluster toward the lower-left portion of the graph, a few outliers appear at the higher end, indicating counties with particularly large case counts and corresponding deaths. From a COVID-19 standpoint, these patterns can guide future decision-making in several ways. For instance, identifying counties with disproportionately high death rates relative to their case counts could highlight gaps in healthcare resources or underlying risk factors (e.g., higher elderly populations or comorbidities). Conversely, counties with many confirmed cases but relatively fewer deaths might suggest more effective medical infrastructure or robust public health interventions. Overall, understanding how case counts correlate with fatalities allows policymakers and health officials to allocate resources more effectively, prioritize interventions, and develop targeted strategies for high-risk areas.

Table 3: Counties vs. Deaths

	County (chr)	**Deaths (dbl)**
1	Harris County	3825
2	Dallas County	2453
3	Bexar County	2040
4	Hidalgo County	2018
5	El Paso County	1940
6	Tarrant County	1798
7	Cameron County	1126
8	Lubbock County	648
9	Travis County	626
10	Nueces County	517

```
## # A tibble: 10 × 2
##    County         Deaths
##    <chr>           <dbl>
##  1 HARRIS COUNTY    3825
##  2 DALLAS COUNTY    2453
##  3 BEXAR COUNTY     2040
##  4 HIDALGO COUNTY   2018
##  5 EL PASO COUNTY   1940
##  6 TARRANT COUNTY   1798
##  7 CAMERON COUNTY   1126
##  8 LUBBOCK COUNTY    648
##  9 TRAVIS COUNTY     626
## 10 NUECES COUNTY     517
```

Table 3 was created after first generating a group-wise average table that resulted in a 10 x 2 tibble, which provided the basis for organizing the data into a clear top-ten summary of counties by reported COVID-19 deaths. It highlights the top ten Texas counties by reported deaths, showing that Harris County had the highest count at 3,825, followed closely by Dallas, Bexar, and Hidalgo Counties. These higher death counts often correlate with larger population centers, which face increased transmission risk due to greater population density. From a COVID-19 perspective, such data helps public health officials identify hotspots for allocating medical resources, staffing, and targeted interventions. For instance, counties like Harris and Dallas, with high death tolls, may require additional healthcare support or community outreach programs to mitigate further loss of life. On the other hand, smaller counties—like Lubbock or Nueces—still warrant attention if they exhibit above-average death rates relative to their population sizes. Ultimately, understanding which counties have experienced the greatest impact informs future decisions around vaccine distribution, public health mandates, and preparedness measures to prevent similar surges in future outbreaks.

Table 4: Counties vs. Confirmed Cases

	County (chr)	**Confirmed Cases (dbl)**
1	Harris County	286356
2	Dallas County	234625
3	Tarrant County	195518
4	Bexar County	152231
5	El Paso County	107552
6	Collin County	64721
7	Travis County	61468
8	Hidalgo County	56455
9	Denton County	46272
10	Lubbock County	45600

```
## # A tibble: 10 × 2
##    County         Confirmed_Cases
##    <chr>                    <dbl>
##  1 HARRIS COUNTY          286356
##  2 DALLAS COUNTY           234625
##  3 TARRANT COUNTY          195518
##  4 BEXAR COUNTY           152231
##  5 EL PASO COUNTY         107552
##  6 COLLIN COUNTY           64721
##  7 TRAVIS COUNTY           61468
##  8 HIDALGO COUNTY           56455
##  9 DENTON COUNTY           46272
## 10 LUBBOCK COUNTY           45600
```

Table 4 was created after first generating a group-wise average table that resulted in a 10 x 2 tibble, which provided the basis for organizing the data into a clear top-ten summary of counties by confirmed COVID-19 cases. It highlights the top ten Texas counties by confirmed cases, showing that Harris County has the highest count at 286,356, followed closely by Dallas and Tarrant Counties. A table is especially useful here because it keeps the data straightforward and neatly organized, allowing readers to quickly identify the largest case counts. The top five counties—Harris, Dallas, Tarrant, Bexar, and El Paso—stand out for having substantially higher case numbers than the remaining five. This disparity suggests that counties with large urban centers and high population densities face greater transmission risks. By contrast, smaller or more rural counties typically report fewer cases, likely due to lower population density and reduced person-to-person contact. For policymakers and public health officials, these findings underscore the need for targeted interventions—such as vaccine distribution and healthcare resource allocation—in more populous regions where case counts are highest.

Scatterplot of Median Age vs. Deaths

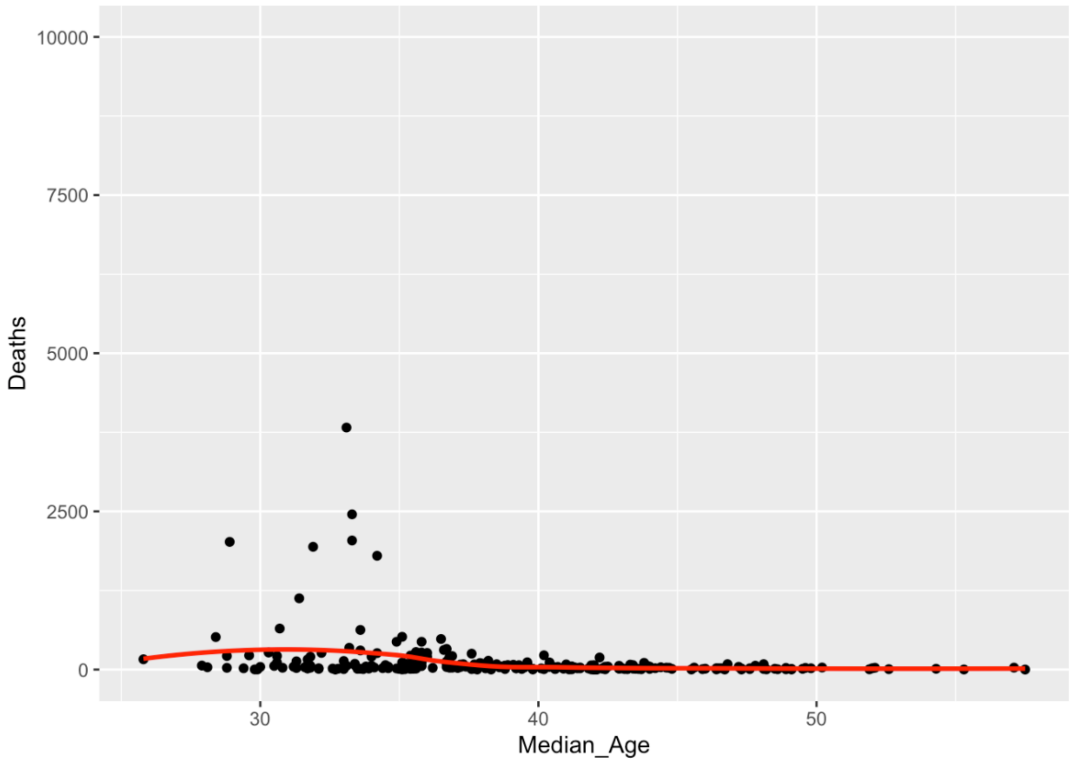


Figure 7: Median Age vs. Deaths

To examine the relationship of median age and deaths a scatterplot would be most useful. As seen in Figure 7, most of the points cluster at low death counts, therefore, most counties had deaths less than 500. This suggest that majority of the clusters had low death counts, whereas more urban areas had higher counts. Additionally, the trend line is flat indicating there is no correlation between median age and death counts. Though, the trend line has a slight increase from age 30 to 35 and then levels off. This could be because populated counties with a younger median age had higher case numbers, resulting to a higher death count. Additionally, even though, older individuals are at a higher risk, the scatterplot suggest that there may be other factors that played a larger role for the death counts at the younger median age level. For example, the counties with a younger median age may have had higher death counts because these individuals were perhaps essential workers. Additionally, there are outliers present where counties have more than 5,000 deaths which are likely to had occur in larger cities. Therefore, these areas are more likely to experience higher COVID-19 outbreaks due to a higher dense population and lack of public health measures.

Scatterplot of Median Age vs. Employment in Education and Health

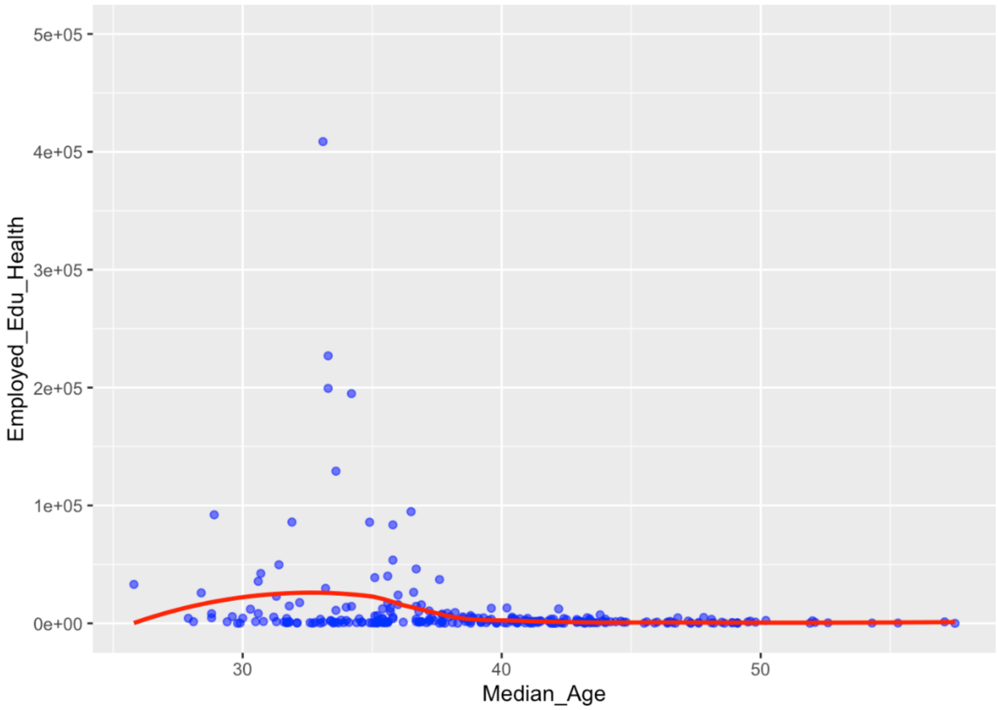


Figure 8: Median Age vs. Employed Edu Health

A scatterplot is most effective to analyze the relationship of median age and employed education and healthcare sectors. In Figure 8, most of the points cluster at the lower level of employment, below the 100,000 employee's mark. The trend line arises initially within the age range of 30 to 35 but then declines from the age of 40 on. This suggest that older counties have less working-age individuals within these fields. Additionally, that counties which has younger populations have a higher employment in these sectors. Additionally, counties that had a high employment was more exposed to COVID-19, and larger hospitals and education institutions were hotspots during this period. Furthermore, outliers are present as there are few counties with more than 400,000 employees, which perhaps represent employees in populated cities and worked in a large healthcare system or universities.

Scatterplot of Median Age vs. Employment in Science and Admin

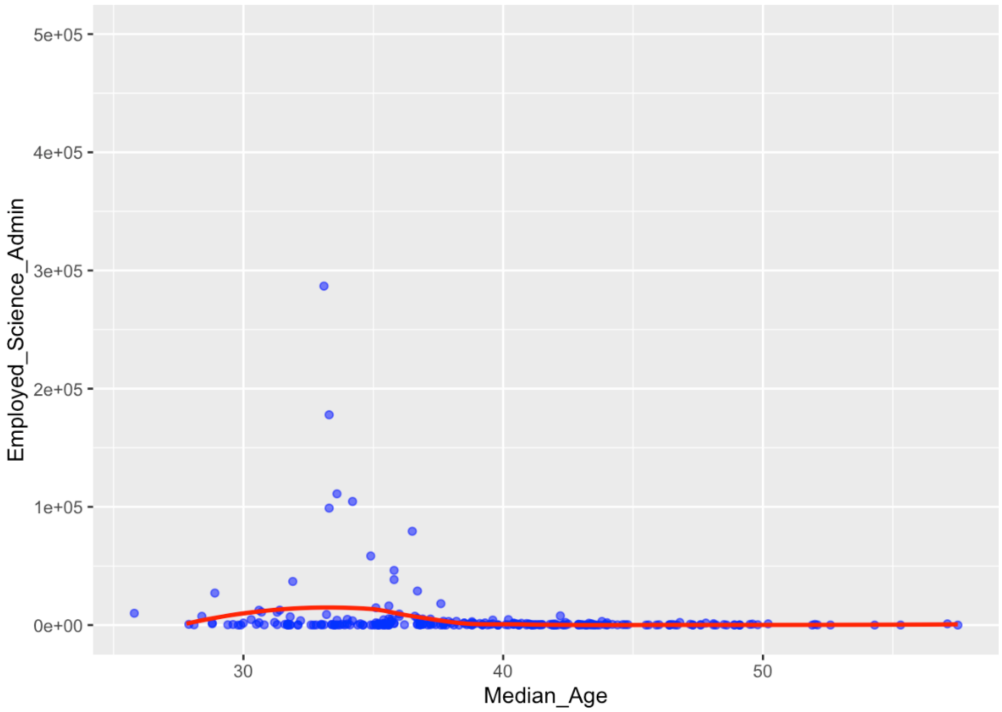


Figure 9: Median Age vs. Employed Science Admin

A scatterplot is most ideal for visualizing the relationship between Median Age and Employed_Science_Admin because both variables are continuous and numeric. In Figure 9, most data points cluster at relatively low employment levels, indicating that most Texas counties have fewer individuals working in science, management, or administrative fields. The lack of a clear correlation suggests that higher employment in these fields does not depend on a county's median age. However, there are notable outliers, including at least one county with an employment figure around 300,000, likely reflecting a larger population or an area with a significant concentration of scientific or administrative industries. From a COVID-19 standpoint, these outlier counties may have distinct advantages in terms of administrative infrastructure, possibly enabling better coordination for remote work or crisis management. Meanwhile, counties with lower employment in these sectors might have fewer resources for organizing pandemic-related initiatives, highlighting a potential area of focus for future public health planning and support.

Scatterplot of Median Incomes vs. Deaths

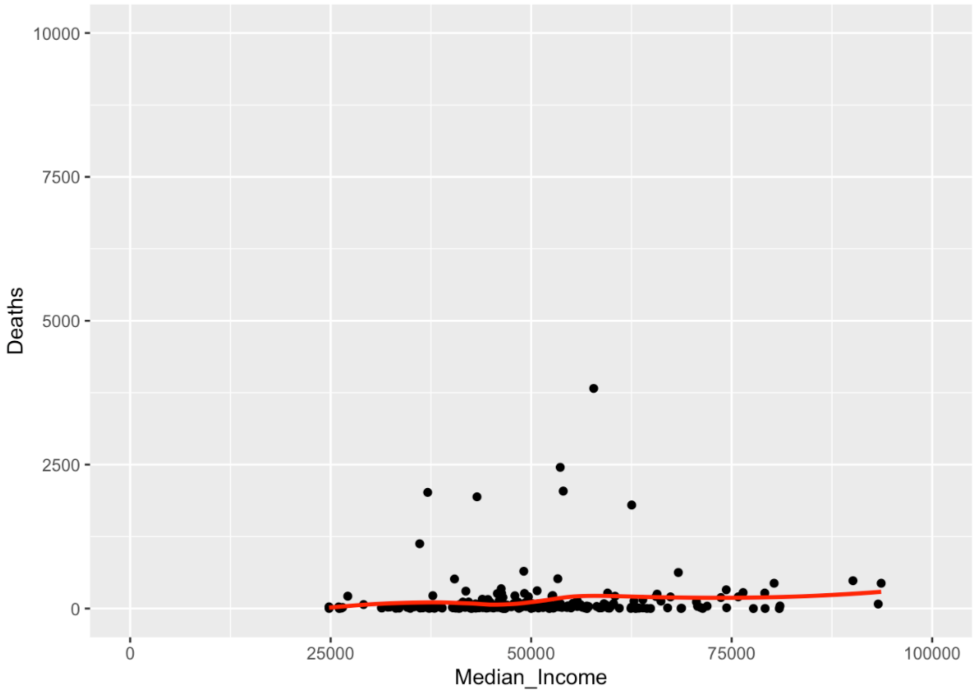


Figure 10: Median Income vs. Deaths

A scatterplot is the most suitable way to illustrate the relationship between Median Income and Deaths since both are numerical variables. In Figure 10, most data points cluster near the lower range of deaths, indicating that the majority of counties experienced relatively few fatalities. The trend line suggests a weak correlation between median income and death counts, though there is a slight upward slope in higher-income counties where death numbers appear to increase. These outliers—counties reporting notably higher deaths—may correspond to larger population centers where overall caseloads and corresponding mortality are naturally higher. From a COVID-19 perspective, the presence of these high-death outliers could highlight areas needing enhanced public health resources or further investigation into specific socioeconomic or healthcare factors. Meanwhile, the weaker correlation across most counties suggests that median income alone does not fully explain death patterns, emphasizing the importance of considering multiple variables (such as population density, age demographics, and healthcare capacity) when planning future pandemic response measures.

Table 5: Family vs Nonfamily Households, Median Income, and Deaths

	Family Households (dbl)	Nonfamily Households (dbl)	Median Income (dbl)	Deaths (dbl)
1	182543	39788	93645	439
2	24526	5839	93269	78
3	240307	83598	90124	483
4	10265	3426	81023	44
5	19	12	80938	0
6	195602	79562	80290	439
7	251	75	79167	1
8	122968	47083	79123	269
9	179	69	77708	0
10	87576	29512	76426	277

```
## # A tibble: 10 × 4
##    Family_Households Nonfamily_Households Median_Income Deaths
##              <dbl>              <dbl>                  <dbl>        <dbl>

##  1          182543             39788          3645                  439
##  2           24526             5839           93269                  78
##  3          240307             83598        90124                   483
##  4           10265             3426         81023                    44
##  5              19            12            80938                     0
##  6          195602             79562        80290                   439
##  7             251            75              79167                   1
##  8          122968             47083        79123                   269
##  9             179            69              77708                   0
## 10           87576             29512        76426                   277
```

Table 5 was created after first generating a group-wise average table that resulted in a 10 x 2 tibble, which provided the basis for organizing the data as it represents the top ten observations of family households, nonfamily household, the median income in those areas, and the total number of deaths recorded in that area. This table provides insights on how COVID-19 impacted on a household level, while also focusing on median income and deaths. From this table it can be noted that family households have a higher proportion than nonfamily households, thus there is a higher population density which suggest that some counties experiences severe outbreaks, while other had minimal case numbers. The median income ranges from $76,426 to $93,645. It can be inferred that households with a higher median income have better access to healthcare, while lower-income counties may be vulnerable to higher case rates due to limited healthcare access. Additionally, the households with a higher population have a higher death count due to more exposure. For example, row 1 has a population of 182,543 family households and 439 deaths whereas, row 10 has a population of 87,576 family households with 277 deaths. Larger family households could have contributed to an increased transmission rate whereas, nonfamily household have fewer deaths likely due to a lower population.

3) Data Preparation

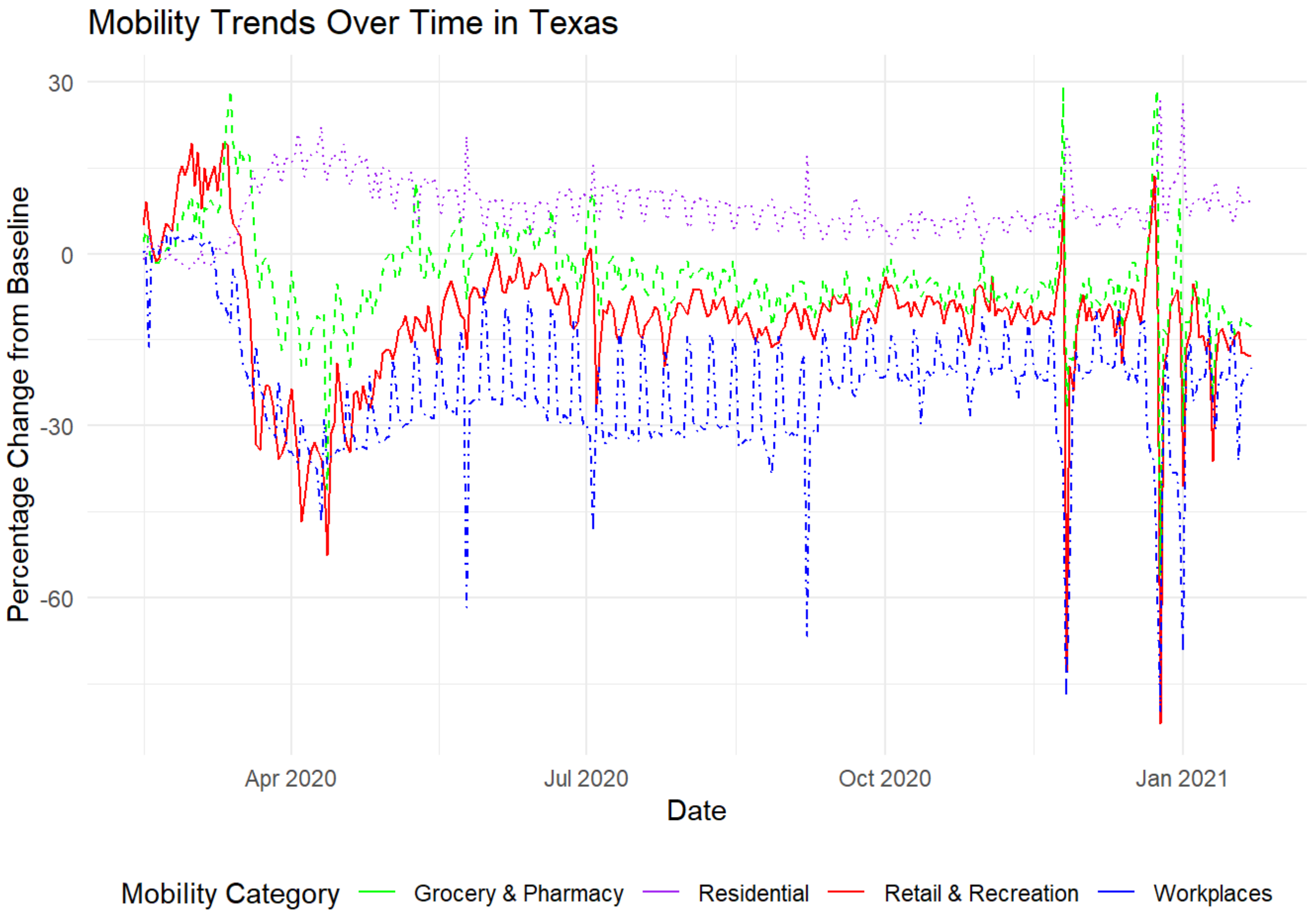


Figure 11: Mobility Trends Over Time in Texas

Figure 11, a time-series graph, illustrates the mobility trends over time in Texas, showing the percentage change for grocery & pharmacy, residential, retail & recreation, and workplaces. From March to April 2020, was when the initial lockdowns occurred. In terms of grocery & pharmacy, there seems to be a spike right at the beginning of the pandemic with visits to essential stores like markets and pharmacies. This likely occurred due to Texas declaring a state of disaster, stay-at-home advisory, and closed non-essential businesses leading to panic buying. From May to June 2020, there were partial reopening causing retail & recreation to begin recovering and residential mobility to slowly decline. In July to August 2020, summer resurgence and the second lockdowns occurred. This can be seen in the decline in workplaces and retail/recreation mobility due to rising cases. During this time there was a statewide mask mandate and a slowdown in mobility reflecting the rise in public caution and renewed restrictions. In January to March 2021, were when vaccine rollouts and gradual recovery started to occur. This is shown by the slight increase in workplaces and retail/recreation mobility. Additionally, during this time, Greg Abbott lifted the mask mandate and fully reopened Texas.

Table 6: Working Dataset

	State	County	Confirmed _Cases	Deaths	Total Population	Median_Income	Median_Age	Employed_Edu _Health	Employed_Science _Admin	Nonfamily_Households	Family_Households
1	TX	Anderson County	5575	75	57747	42313	39.1	4334	1325	4956	11611
2	TX	Andrews County	1606	37	17577	70753	31.6	1093	659	1256	4157
3	TX	Angelina County	6765	193	87700	46472	36.8	10067	2570	8531	22400
4	TX	Aransas County	895	26	24832	44601	49.6	1824	835	3252	6277
5	TX	Archer County	694	10	8793	63192	44.8	1039	347	960	2391
6	TX	Armstrong County	128	6	1929	68750	45.9	234	59	162	540
7	TX	Atascosa County	3781	90	48139	55194	35.4	4329	1379	3690	11819
8	TX	Austin County	1404	18	29292	62614	40.9	2693	1268	2890	8131
9	TX	Bailey County	742	15	7098	43523	35.2	448	178	530	1700
10	TX	Bandera County	820	20	21316	56413	52.0	2114	610	2584	5694

Table 6 showcases our cleaned dataset, representing the first ten variables that we identified as key variables for our analysis. The dataset has been filtered specifically for Texas counties, ensuring a more focus examination of regional trends. Table 6 is derived from a tibble with dimensions 10 x 11, meaning it contains the first 10 rows and all 11 columns, displaying a subset of the most critical variables for our study. This refined dataset allows us to explore key demographic, economic, and COVID-19-related metrics, helping to uncover patterns in population distribution, median income levels, employment sectors, household structures, and the impact of the pandemic at the county level. By cleaning and filtering the data, we ensure higher accuracy, relevance, and consistency in our statistical analyses.

Answering Questions that are Important to our Stakeholder

<u>How many people reported have died from COVID in the state of Texas?</u>

Based on *Table 3: Counties vs. Deaths* as shown above, the top ten counties in Texas collectively account for a substantial number of COVID-19 fatalities—totaling 16,991 when summed (led by Harris County at 3,825, followed by Dallas at 2,453, and Bexar at 2,040). While these ten counties alone represent a significant share of the reported deaths in Texas, the statewide total is undoubtedly higher once all 254 counties are considered. The large numbers in these top counties can be attributed to factors such as population density, urbanization, and higher overall case counts, which naturally increase the likelihood of severe outcomes. For instance, Harris and Dallas counties encompass major metropolitan areas, leading to elevated transmission rates and, consequently, more deaths. By focusing on these high-impact regions, public health officials and policymakers can prioritize resource allocation—such as hospital staffing, testing capacity, and vaccination campaigns—in order to mitigate further loss of life. However, it is equally important to recognize that smaller counties, though not listed in the top ten, may still face significant challenges relative to their healthcare infrastructure. Overall, analyzing these most affected counties sheds light on the critical need for targeted interventions and support in regions that bear the brunt of the pandemic's toll while also acknowledging that the statewide impact extends beyond just the major urban centers.

Based on these deaths, where are the most outbreaks occurring, and what factors contribute to their spread?

Table 4: Counties vs. Confirmed Cases and *Figure 11: Mobility Trends Over Time* in Texas both suggest that major urban counties—such as Harris, Dallas, Tarrant, Bexar, and El Paso—experience the highest number of COVID-19 outbreaks. In Table 4, Harris County alone reports over 286,000 confirmed cases, with Dallas and Tarrant not far behind, which aligns with the fact that these counties also have some of the largest populations in the state. This connection between population density and high case counts underscores the role of crowded living conditions and frequent person-to-person interactions in fueling virus transmission.

The mobility data offers additional insight into how public behavior affects outbreak patterns. Early in the pandemic, substantial declines in visits to workplaces, retail locations, and recreational sites likely contributed to slowing the spread. However, as the data show mobility rebounding over time, in-person interactions and crowding in workplaces and commercial spaces have likely accelerated new outbreaks. In large metropolitan areas, many residents hold jobs that cannot be performed remotely, further driving the potential for viral transmission. Moreover, health disparities and resource constraints play a role: urban centers can quickly see hospitals overwhelmed when cases surge, while smaller counties may have fewer total cases but limited capacity to handle even a moderate influx of patients.

Public policies and local compliance with measures such as mask mandates, social distancing, and vaccination campaigns also influence how outbreaks unfold. In areas where adherence is weaker or enforcement is less consistent, the virus may spread more quickly, especially if mobility remains high. Consequently, these data underscore the need for targeted interventions—such as allocating vaccines and healthcare resources to the most affected urban areas—and adaptive public health measures that can tighten or relax restrictions based on observed mobility trends and rising or falling case counts. By synthesizing the confirmed case data with mobility insights, policymakers and health officials can implement strategies that address the specific challenges of densely populated counties and reduce transmission risks statewide.

Which demographics are most affected, and how should interventions be tailored?

An examination of Figures 7, 8, and 9 suggests that COVID-19 impacts do not hinge on a single demographic factor like median age alone. In *Figure 7: Median Age vs. Deaths*, most counties cluster at lower death counts, indicating that many areas did not see high fatality numbers. However, a few outliers, typically large urban centers, recorded thousands of deaths regardless of their median age. This pattern suggests that population density, occupational exposure, and healthcare capacity may outweigh age as predictors of higher fatality counts. In some cases, younger counties saw more deaths, possibly because a larger portion of their population worked in essential roles, increasing exposure risk.

Turning to *Figure 8: Median Age vs. Employment in Education and Health*, the data show that counties with younger populations tend to have more people employed in education, healthcare, and social assistance, which are sectors closely tied to frontline work. These workers faced higher exposure risks during the pandemic, particularly in large hospitals or educational institutions that became hotspots. Conversely, older counties generally had fewer working-age individuals in these fields. This discrepancy indicates that younger counties might need additional protections (e.g., PPE distribution, vaccine outreach) for essential workers, while older counties may require more comprehensive healthcare services and vaccination strategies to protect vulnerable populations.

Lastly, *Figure 9: Median Age vs. Employment in Science and Admin* reveals minimal correlation between these fields and median age. Counties with substantial science, management, or administrative employment may have been better equipped to transition to remote work or implement efficient pandemic protocols, thereby reducing transmission risk. Meanwhile, counties lacking these administrative infrastructures could struggle to coordinate large-scale public health initiatives, underscoring a need for external support and resource allocation.

Overall, these findings suggest that interventions should be tailored to each county's demographic and occupational profile rather than solely focusing on age. Urban centers with high death counts need robust healthcare infrastructure and targeted outreach for essential workers, while counties with high frontline employment but fewer administrative resources may benefit from external coordination and protective measures. Counties with large older populations also warrant dedicated strategies—such as improving vaccine access and healthcare capacity—to safeguard those at higher risk of severe illness. By customizing interventions based on median age, employment sectors, and population density, public health officials can address the diverse needs of Texas counties more effectively.

Where should resources like vaccines and medical staff be distributed?

Table 5: Family vs Nonfamily Households, Median Income, and Deaths reveals that counties with higher counts of both family and nonfamily households tend to report higher numbers of COVID-19 deaths. For instance, counties in rows 1, 3, and 6—each characterized by large household counts—also record death counts nearing or exceeding 400, suggesting that densely populated areas are experiencing more severe outbreaks. While median income values are relatively similar across most counties, the one outlier with a notably lower median income in row 1 underscores potential vulnerabilities where economic constraints may limit access to healthcare services. Furthermore, it is important to consider how neighborhoods are structured; areas with closely spaced housing, such as apartment complexes or condos, can facilitate easier transmission of the virus compared to more spread-out housing communities. Given these findings, resources such as vaccines and medical staff should be strategically allocated to counties with the highest household densities and death counts, particularly in urban centers where residents live in closer proximity. By focusing on these high-risk areas—especially where housing arrangements may exacerbate transmission—public health officials can better target interventions to curb spread, support overburdened healthcare systems, and ultimately reduce COVID-19 fatalities.

4) Recommendations

How can COVID data inform future public health funding and policymaking for the state of Texas?

Based on the comprehensive COVID-19 data analyses, I recommend that the Commissioner of the Texas Department of State Health Services prioritize targeted public health funding and policy measures focused on urban centers with high case counts, deaths, and densely populated households. The data clearly indicates that major urban counties such as Harris, Dallas, Tarrant, Bexar, and El Paso—notably highlighted in Table 3 and Table 4—are experiencing the most severe outbreaks. These areas, with high mobility trends and densely packed living conditions (as seen in the household data), face increased transmission risks, especially where apartment complexes or condos are prevalent compared to more spread-out housing. Furthermore, the scatterplots comparing median age with deaths and employment in essential sectors reveal that while there is a weak direct correlation between median age and death counts, counties with younger populations tend to have a larger proportion of frontline workers in education, healthcare, and related fields. This suggests that interventions should include enhanced protective measures, such as prioritizing PPE distribution and vaccinations for essential workers as well as bolstering healthcare capacity in these critical regions.

To act on these findings, I recommend that the Commissioner implement a strategy to allocate additional vaccines and medical staff to counties with high household densities and elevated death counts. Policymakers should consider funding programs that enhance remote work infrastructure and telehealth services, particularly in urban areas where administrative capacity—as reflected by employment in science and management—is limited. Additionally, tailored public health messaging and outreach should focus on high-risk demographics, ensuring that younger, essential workers in densely populated areas receive the support they need. These data-driven, actionable strategies will not only help mitigate future outbreaks but also ensure that resources are deployed where they are most needed, ultimately strengthening Texas' overall public health response.

5) Conclusion

In conclusion, the analysis reveals that urban counties such as Harris, Dallas, Tarrant, Bexar, and El Paso bear the brunt of COVID-19's impact, with the highest confirmed cases and deaths. The data further indicate that densely populated areas, particularly those with close housing arrangements, experience higher transmission rates and fatalities, while employment in essential sectors—especially in education, healthcare, and administrative fields—varies with demographic factors. These findings are crucial because they highlight where public health resources and interventions should be prioritized, ensuring that funding, vaccines, and medical support are allocated to regions most at risk. Ultimately, these insights provide a clear, data-driven roadmap for policymakers to enhance Texas' preparedness and response to future public health challenges.

6) List of References

Centers for Disease Control and Prevention. (n.d.). CDC Covid Data tracker. Centers for Disease Control and Prevention. https://covid.cdc.gov/covid-data-tracker

Centers for Disease Control and Prevention. (2025, February 13). Home page for MMWR. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/index.html

Covid-19. Johns Hopkins Medicine. (n.d.). https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

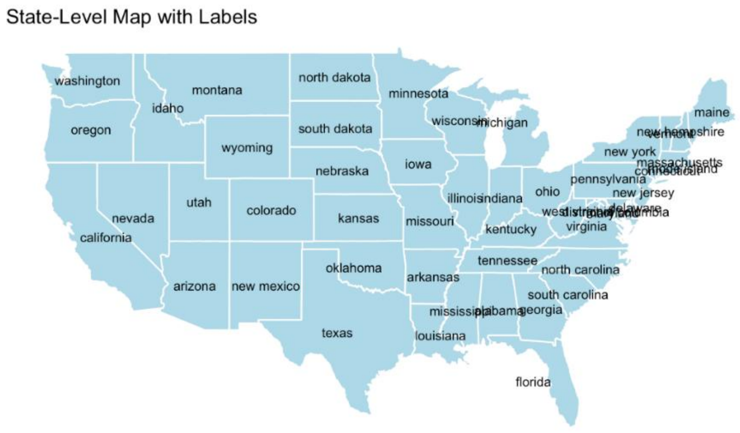
U.S. Department of Health and Human Services. (n.d.). *Coronaviruses*. National Institute of Allergy and Infectious Diseases. https://www.niaid.nih.gov/diseases-conditions/coronaviruses

World Health Organization. (n.d.). Covid-19 cases | WHO COVID-19 Dashboard. World Health Organization. https://covid19.who.int/

“CDC Museum Covid-19 Timeline,” Centers for Disease Control and Prevention, https://www.cdc.gov/museum/timeline/covid19.html (Accessed Mar. 12, 2025).

7) Appendix

Here are the graphs visually exploring some of our chosen attributes. While they provide useful insights, we didn't find them as critical for stakeholder analysis and have prioritized other key factors in our evaluation.



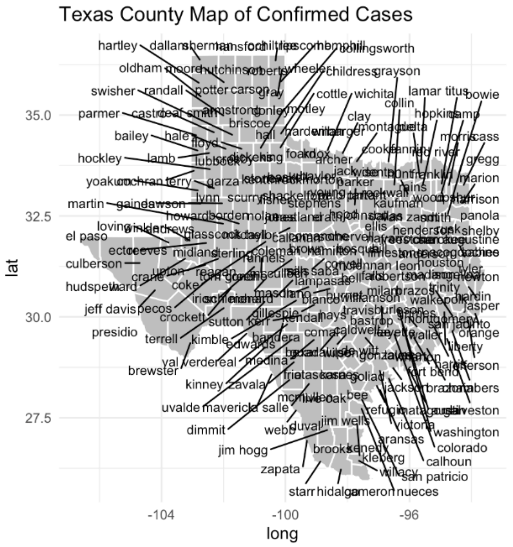
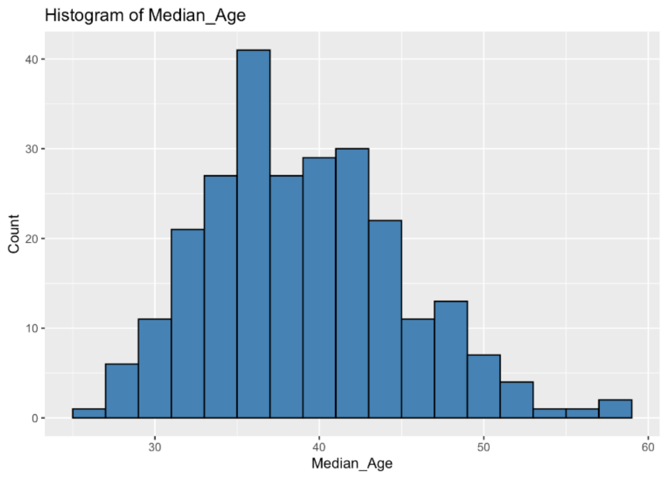


Figure 12: State Level Map with Labels

Figure 13: Texas County Map of Confirmed Cases



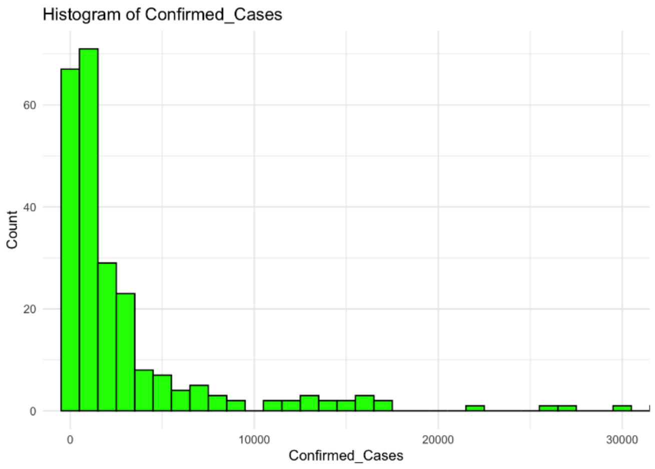


Figure 14: Histogram of Median Age

Figure 15: Histogram of Confirmed Cases

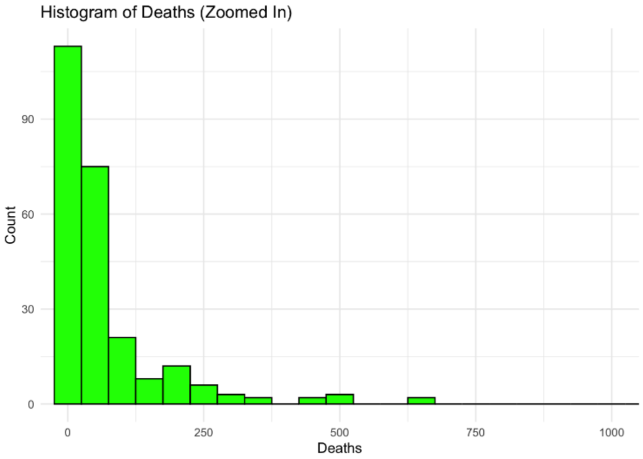


Figure 16: Histogram of Deaths

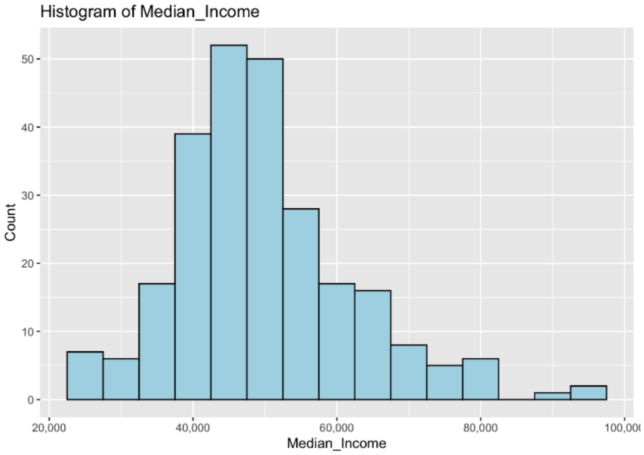


Figure 17: Histogram of Median Income

Student Contributions

All three group members contributed equally to the project. Sreshta and Muskaan took the lead on the coding portion of the report, while Ridhi focused on conducting the report analysis. Each person played a crucial role in ensuring the overall success of the project, with the workload distributed evenly among the team.