COVID-19 Dallas District Analysis and Policy Recommendations (2021)

By Lawrence Lim

Table of Contents

Abstract

As COVID-19 continues to spread without any available vaccine or treatment options, communities are relying on smart social policies like quarantines and mask-wearing. However, different communities have a wide variety of policies and strategies based on their unique situation. As the data analyst for the city of Dallas, I wish to understand our county's performance in minimizing COVID-19 cases compared to districts in different states with similar population amounts. I want to understand whether or not population is truly the top factor that influences the spread of COVID. Or does the median income or educational background of the community have a greater influence/correlation with increased success at suppressing the virus?

Business Understanding

COVID-19 is a severe respiratory disease that originated in Wuhan, China, during the Fall/Winter months of 2019. As an airborne transmissible disease, it has spread to other countries through means such as air travel. According to the CDC, the first individual with COVID in the USA was reported on January 20th. On January 30th, the first US case of airborne-transmitted COVID was reported in Illinois, and from this point, the virus continued to spread. Throughout the months of February, the virus would begin spreading to populous states like California and New York, and Texas would report its first covid cases on February 14th. Like most transmissible viruses, COVID spread exponentially, with the months of March seeing an exponential surge in cases in all 50 states.

With an increasing amount of severe cases and limited resources, communities across the country enforced controlled lockdowns, enforced mask-wearing, and limited the occurrence and size of any public events as a means of reducing transmission of the virus. Furthermore, communities also established curfews, recommended individuals practice social distancing in public by standing at least 6 feet from another person, and recommended individuals who were exposed to individuals infected quarantine for 15 days.

The actions above were taken with the purpose of flattening the curve, which means reducing the transmission of the virus and ultimately lowering the infection rate to levels that are manageable by the healthcare system. This is significant from the context of my role as a data health analyst, as according to reporting from The Texas Tribune, over 50 Texas hospitals were overwhelmed with no ICU beds for increasingly sick patients during the winter of 2020-2021. Thus, individuals that needed hospital treatment suffered from COVID without necessary aid, which could have led to unnecessary casualties. Thus, both state and national district data, I can identify commonalities in districts that performed poorly so I can predict which districts might again struggle in the Winter of 2021-2022. Thus, policies can be implemented on a state level to prepare the healthcare system on a district-by-district basis.

The stakeholder for this research is the Texas Department of State Health and Services. Among many responsibilities, this organization establishes and maintains quarantine orders throughout the state and influences state health policy. Thus, when a vaccine is released shortly, the organizations will be responsible for ensuring that all districts get adequate amounts of the vaccine. Furthermore, as stated on their webpage, they are "Driven by Science and data" in the decisions they make, which means they will be more likely to take the inferences from my study seriously.

Data Understanding

❖ Description of the Data

The COVID-19 national census file contains information regarding the number of cases and confirmed deaths as of January 19th 2021 in various counties. However, the data engineer compiling this data set also included demographic information regarding each of the counties including demographic information, socioeconomic information, and educational background. Considering the context of this study is merely to determine whether education, socioeconomic, demographic and population play a role in COVID and the level of role they play in cases, I have decided to subset the dataset to include only a specific number of features as listed in the Data Dictionary below:

NOTE: * means that the particular feature is a feature I derived/calculated from other attributes in the dataset.

Column	Data Type	Description	Example	Potential Use Cases of Features
County Name	String	This is the name of the particular county.	Tarrant County	Needed to identify counties for comparison
State	String	This is the state abbreviation the county resides in.	TX	Needed to identify larger scope regions of multiple counties to see if there are state-related trends.
Population_1_year_and_over	integer	This is the population of a county that is over one year old	1000000	Needed to rank different districts based on population size. Thus, areas with similar populations can be compared based on other variables like education.
Confirmed	integer	This is the number of confirmed cases (as reported on January 19th, 2021).	1201	Needed to see and compare number of cases between different areas.
Deaths	integer	This is the number of deaths (as reported on January 19th, 2021).	321	Needed to see and compare the number of deaths between different areas.

geo_id	integer	This is the geographical ID as identified by the Census Bureau for regions across the nation.	50009	May be essential if I choose to use R to create a geographical map of the United States.
median_year_structure_built	integer	This is the median year that a home was built in for a particular county.	1972	Necessary to compare the overall age of homes to determine if new infrastructure rules could be a factor.
percent_income_spent_on_rent	numeric	This is the average % amount of income spent on rent for residents in the county.	52%	Necessary to determine the influence of income to see if areas with individuals struggling more economically are more prone to covid.
Percent_male *	numeric	This is the percentage of males in the county	50%	Necessary in determining whether one gender or another is at greater risk of COVID.
median_income	integer	This is the median income of an individual in the county.	$60,000	Used in comparing the relative economic performance of a region.
% graduate degree or more *	numeric	This is the % of individuals that have a graduate degree or higher.	41%	Could be used in ranking the level of education per county.
aggregate_time_travel _work	integer	This is the total amount of time (in minutes) spent by individuals traveling to work on January 19th.	1668430	Could be used in ranking the level of “openness” of a particular district. I would expect areas that are more closed/restrictive to have less travel minutes relative to an area with the same population, but more travel

				minutes.
Genie index	numeric	This is an index that measures inequality in a particular county.	0.417	Could add depth to analysis regarding the impact of a populus economic wellbeing on COVID spread.

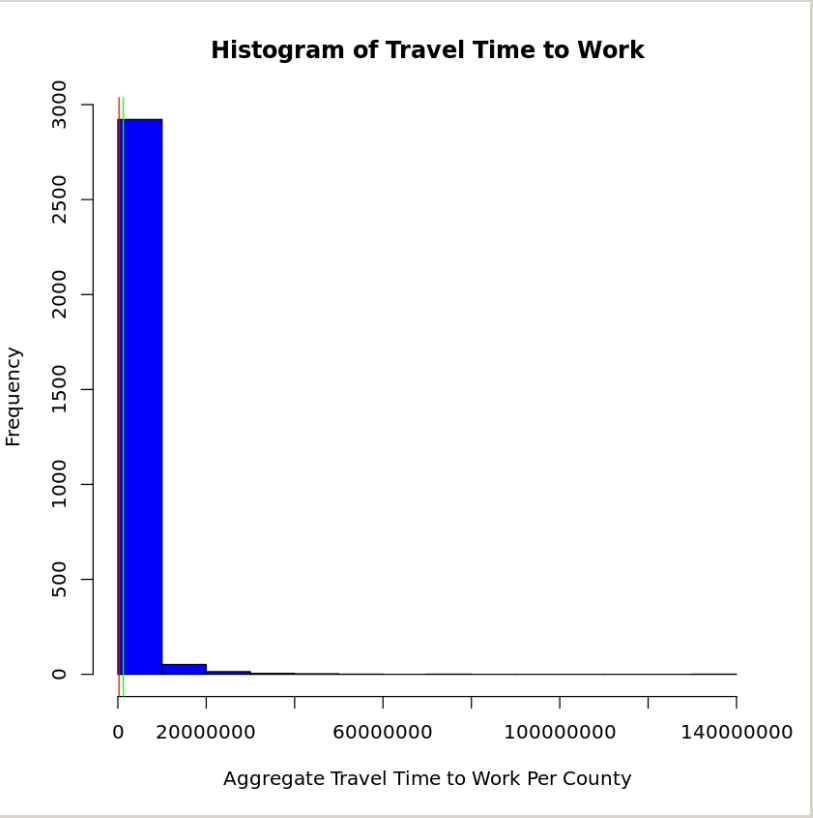
The COVID-19 cases Texas file contains the number of cases and deaths for Texas counties daily, from January 22nd, 2020, to January 25th, 2021. Given that my study is from the perspective of Dallas County in comparison to other counties, I have decided to keep all the features but extract only three counties from this report (Dallas, Harris, and Bexar). These were selected as they are the three largest cities in Texas. Bexar has a very similar population to Dallas, which allows us to see whether other factors might cause cases and deaths to diverge between the two cities. Furthermore, Harris has a larger population. It will be interesting to see how Harris' increased population changes the way cases and deaths occur daily. For instance, does it increase the rate of increase in cases daily?

- ❖ **Analysis of Data Quality**

It is important to consider whether or not there is the presence of blank values in the dataset that could cause issues when comparing different counties. Fortunately, the state dataset does not have any blank values in the attributes pertaining to Dallas, Harris, and Bexar counties. However, in regards to my subset of the national Census Data, the following columns have blank values:

Column	Number of blanks	% of rows blank
median_year_structure_built	1	0.03%
Percent_income_spent_on_rent	1	0.03%
Aggregate_Time_Work_Travel	143	4.54%

We can see that the % income spent on rent and the median year structures are built is missing only one value. Aggregate time to work, a more difficult-to-track statistic, is missing over 100 blanks. Given this, we have the option of either deleting the blank values or imputing them. This is important as performing calculations on NAN values in my later analysis will cause the result also to be NAN, ruining the calculation. Determining whether to impute or delete depends on the mean and the median of each of these columns and a histogram displaying the distributions of values.



The above histogram shows that the aggregate travel time is skewed to the right, with the mean being greater than the median. The scatter plot further zooms in to allow us to better see the difference in the mean and median per county. We note that there are counties that lie in between the mean and median lines, which suggests that one of the lines (the mean and median) is a better representation of the true “middle” of the data compared to the other. To determine this, we determine the number of counties with aggregate time travel above and below the mean and median, respectively to determine whether the mean or median has a ratio nearest to 1:1.

Mean-Ratio (Greater: Less)	508:2491
Median-Ratio (Greater: Less)	1499: 1499

The table above clearly shows that the median is a much better representation of the middle of the data, and thus, is the option we will use for imputing the aggregate time traveled column. Deleting values would not be ideal as the people we would be deleting would skew the median artificially- introducing potential bias. Using this technique of mean and median ratios, I will look at the remaining columns with blank values to determine how we should impute them.

Percent_income_spent_on_rent	
mean-ratio	1578:1421

median-ratio	1470:1492

median_year_structure_built	
mean-ratio	1663:1336
median-ratio	1417:1457

Though the median is not the precise middle for either of the two columns above, we see that it is still much closer to a 1:1 ratio than the mean. Furthermore, the fact that the mean and median have different ratios shows that both percent income and mean year structure built are skewed, which, again, makes deleting the blank rows, not an option. In summary, we will impute each of the columns with the median for the respective column.

❖ Relevant Summary Statistics

The summary statistics for all measurable, quantitative variables are listed in the below matrix:

Variable	Mean	Median	Mode	1st Quartile	3rd Quartile	Min	Max
Population_1_year_and_over	100970	25390	22637	10832	66822	74	9988370
Confirmed	7558.9	1916.5	1147	796.2	4955.0	0.0	1002614.0
Deaths	124.8	31.0	0	12.0	77.0	0.0	13936.0
median_year_structure_built	1974.5	1977.0	1979	1968.0	1983.0	1939	2003.0
percent_income_spent_on_rent	27.82	28.10	28.8	25.30	30.30	10.00	50.00
Percent_male	50.08	49.60	49.37	48.90	50.59	4.90	80.83
median_income	49754	48066	48412	41123	55764	19264	129588
% graduate degree or more	9.479	7.789	5.095541	5.899	11.314	0	58.621
aggregate_time_travel _work	1198322	263630	263630	115416	676734	3040	136962170
Genie_Index	0.4448	0.4423	0.4174	0.4211	0.4665	0.3271	0.5976

The summary statistics above, overall, shows that there is enough variety in the data in order to make comparisons based on each of the different features. For instance, the % of income spent on rent varies from 50% to as low as 10%, which suggests the counties represent communities across the economic spectrum, from those that face economic hardships to more upper-class communities. Furthermore, the community % of individuals with a graduate degree or higher ranges from less than 10% to even more than 50%.

However, there is also an area of concern, highlighted in red. Firstly, in regard to deaths, the most frequent number is the minimum, zero. Though this may not seem surprising, as areas with healthcare infrastructure have the resources to care for more at-risk patients, it does make me question the reason the number is zero. Did a community actually have zero reported cases? Or were there periods when a county simply decided not to record cases, leading to days being reported as zero? Assuming the latter, it may be better to take the mean deaths as artificially skewed downwards. Furthermore, if we make any inferences utilizing deaths, we might want to assume that the deaths reported are conservative in nature.

It is also important to note that the min and max values above do not appear unrealistic or suspicious in any manner, which suggests that though there might be counties with radically-high amounts of cases, deaths, number of males, etc., there is no evidence of erroneous values. For instance, if the minimum value for the median_year attribute were to be 200 or a maximum of 20000, that would suggest a potential typo or issue that could be explored. Furthermore, the percentages all lie between 0 and 100, which is to be expected.

The following shows summary statistics for the measurable statistics (confirmed cases and deaths) as reported in The Texas Subset pertaining to Dallas, Bexar, and Harris:

Variable	Mean	Median	Mode	1st Quartile	3rd Quartile	Min	Max
Confirmed_Dallas	60423.58919	44416.5	0	2703	92683	0	246820
Deaths_Dallas	753.2108108	580	0	66.75	1219.25	0	2637
Confirmed_Harris	87198.53243	62017.5	0	5240.75	158132.25	0	297629
Death_Harris	1420.386486	919	0	80.5	2781	0	4024
Confirmed_Bexar	39299.92703	35311	0	1136.25	64581	0	162108

Death_Bexar	704.8648649	305.5	0	40	1393.5	0	2161

The statistics above appear to suggest that population is certainly a factor in comparing the big three Texas cities. Harris, which is by far, the most populous county, has the largest number of confirmed cases and around double the amount of deaths when compared with Dallas and San Antonio. That being said, It is important to note the stark differences between Bexar and Dallas. Given that the populations of both cities are rather close, one would expect that the mean number of confirmed cases and deaths would also d, not it is proportional, to the number of cases present in a district. This does hint that alternate factors also play a role in whether a person survives COVID or not. When looking at the median and mean for confirmed cases in Dallas and Bexar, we also note that the median and mean differ by much greater in Dallas compared to Bexar. This suggests that there are more days in Dallas that have a radically high number of cases that are causing the mean to skew further from the median. And when we look at the 3rd quartile, we confirm this as the top 25% of confirmed case reports in Dallas are far higher than the top 25% for Bexar.

- **Analysis and Visualization of Important Attributes**

The first attribute I would like to explore is the number of cases in Harris, Dallas, and San Antonio on a time chart. Analyzing this attribute will allow me to see the rise and fall in cases per day to see if there are initial patterns between different cities.

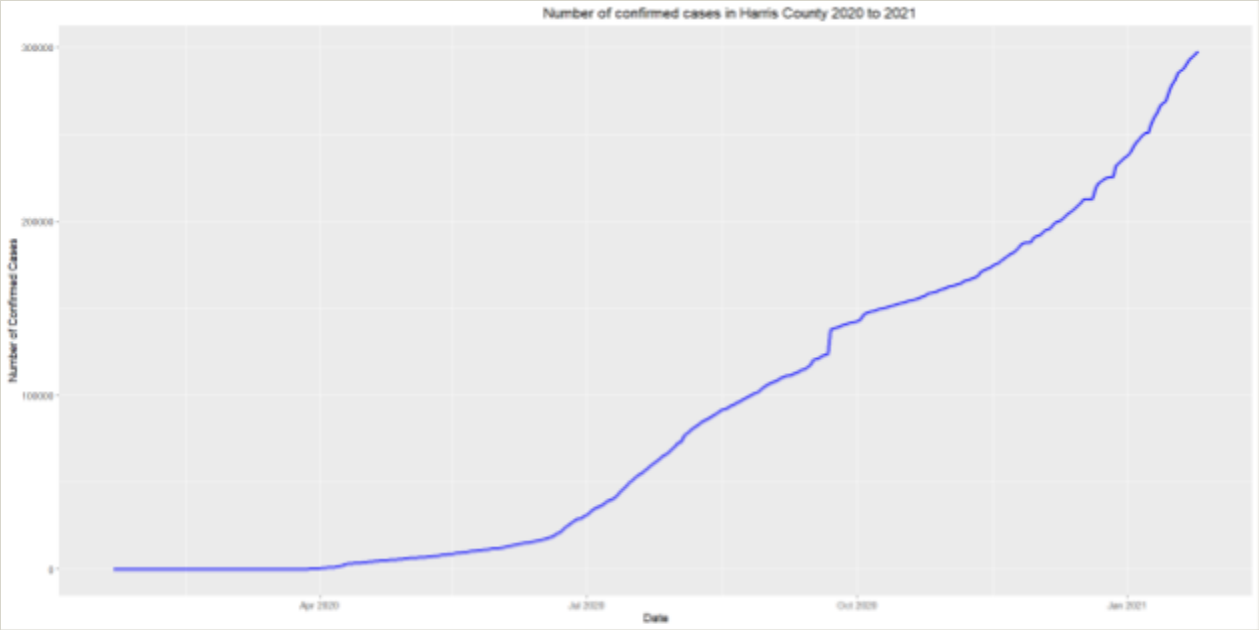


Figure 1: Harris County Confirmed Cases

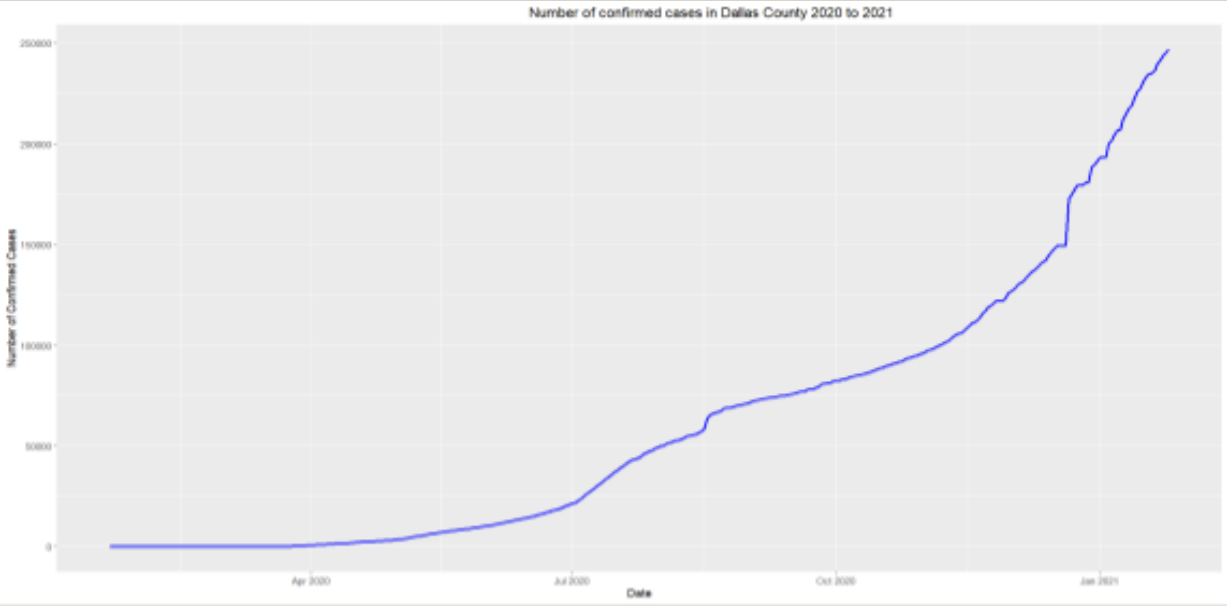


Figure 2: Dallas County Confirmed Cases

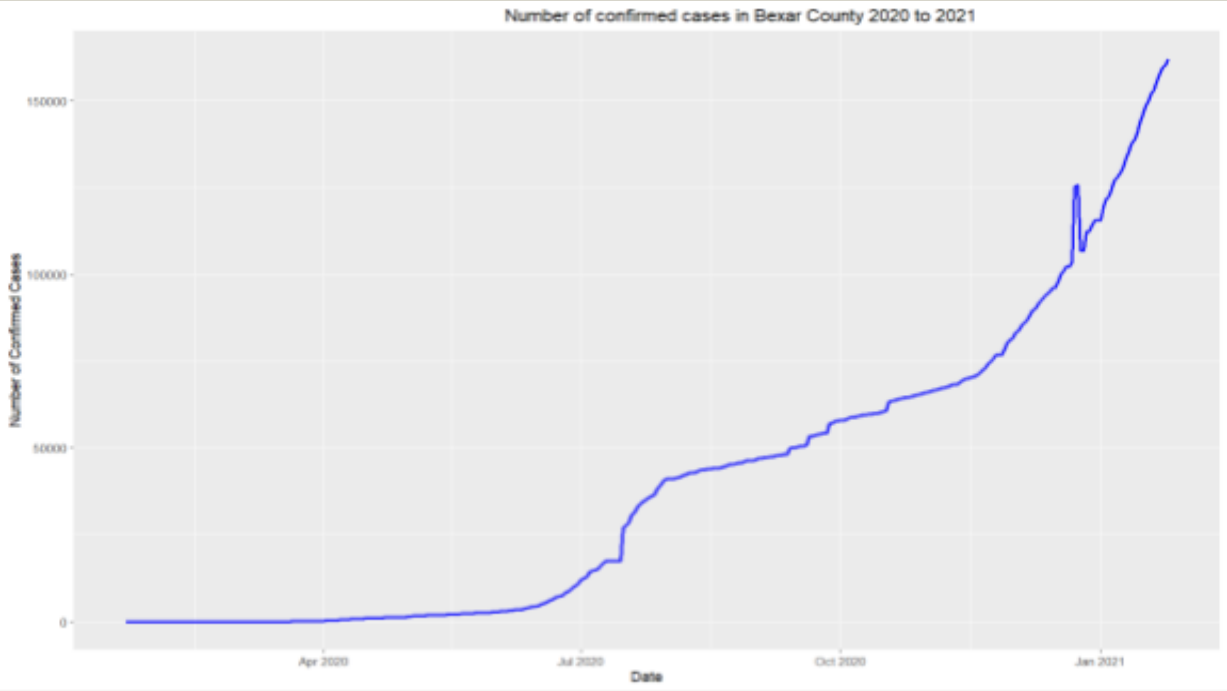


Figure 3: Bexar County Confirmed Cases

Figure 1 shows the number of confirmed cases over time in Houston. When comparing Dallas and Bexar county, we see that all 3 counties had a very similar number of confirmed cases around the month of July (approximately 25000). However, while Dallas and especially Bexar clearly show the curve flatten during the Summer months with the slope decreasing, Harris county shows a relatively linear uptick in cases from July to October. This is important as when cases begin to exponentially increase during the November months, Harris exponentially increases from a higher “starting point”, which causes its number of daily cases to quickly surpass 250,000 after the start of January. When comparing Figure 2 (Dallas) to Figure 3 (Bexar), we see that cases accelerate faster in Dallas compared to Bexar, with Dallas having 25,000 cases by July while Bexar has 12,000.

Another important attribute to consider is the distribution of deaths across the country. This is an important visualization as it allows us to see whether my hypothesis that population is the greatest determiner of whether or not COVID spreads rampant in a particular region. Counties that do not fall into this trend could be explored in relation to the other features to determine if other features are more important. I have decided to use a bar chart to show this as the state values on the independent axis should not be combined For the sake of Comparison.

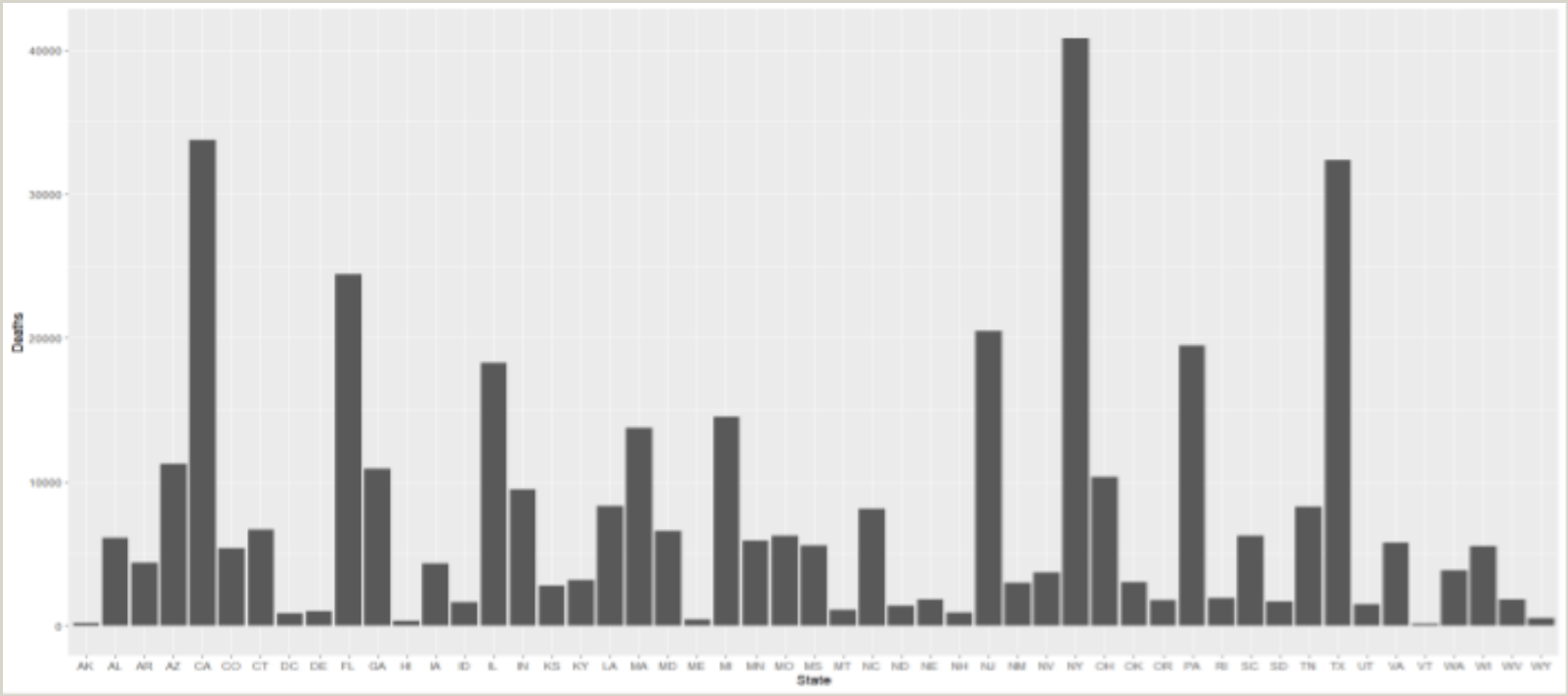


Figure 4: Deaths Depending on State

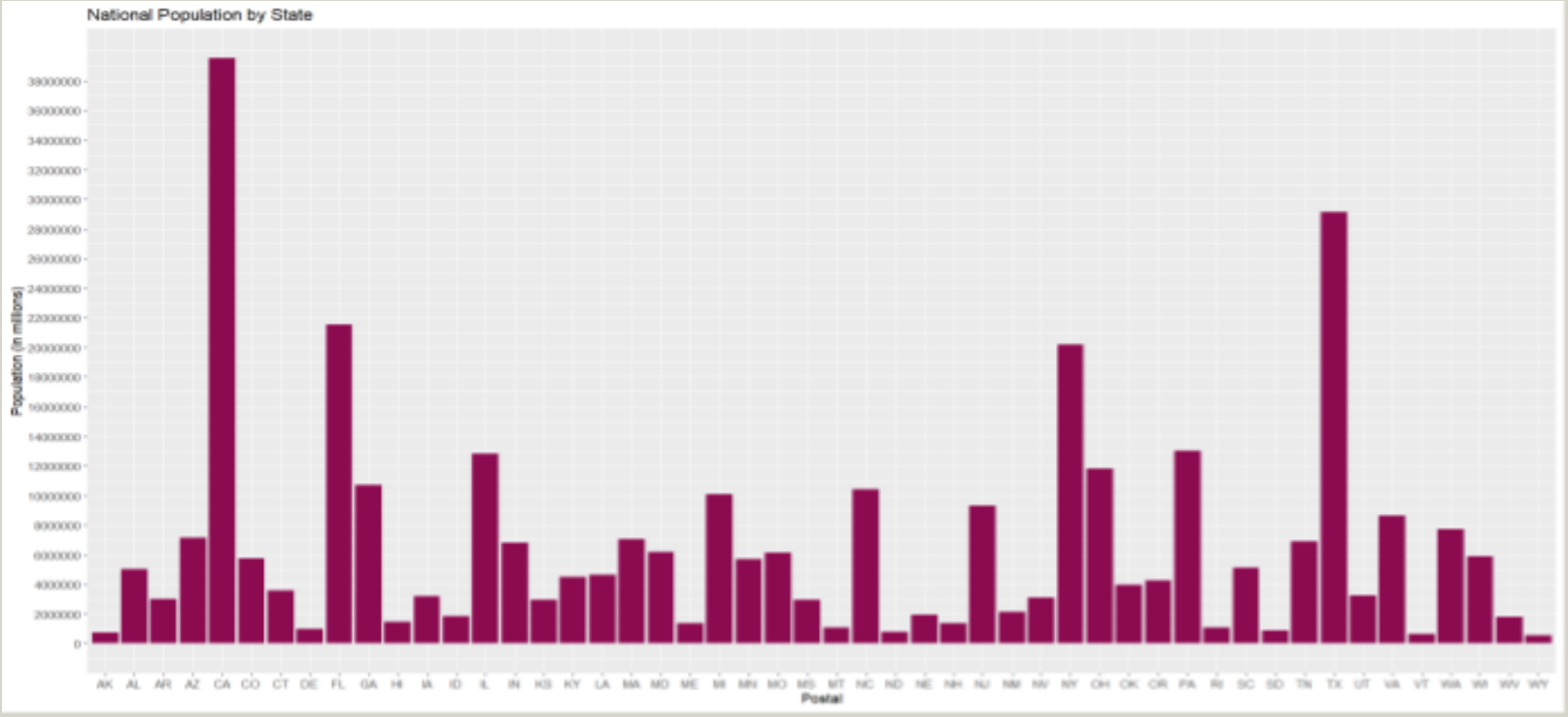


Figure 5: Population per State

Figures 4 and 5 show the number of COVID Deaths as of January 25th 2021 and the state population as reported by the US Census in 2020 online. Since both bar graphs are in alphabetical order, it is easy to see if there are similarities between the two graphs pertaining to populations and deaths based on the overall shape formed by each corresponding state. We see that Texas and California, two of the most populous states, are also the states that have a significant number of COVID deaths. However, the greatest number of COVID deaths is actually from New York, which has the 4th highest population. Furthermore, New Jersey has the 5th most COVID deaths but only the 11th highest population.

❖ Relationships Between Attributes

Figure 6 below shows the correlation of all attributes in the national census. Ignoring the ones along the diagonal, we can see that there is an extremely strong correlation between the population size of a county and the number of confirmed cases and deaths within that county. Furthermore, there are weaker correlations between population and median income and the percent of graduates. It is also important to note that median income and % with a graduate degree have a strong positive correlation. This may explain why for many of the attributes that are correlated with median income, they also have a similar “absolute level” of correlation with percent_graduate. For instance, median_income and Percent_graduate have a weak correlation with both Confirmed cases and Deaths.

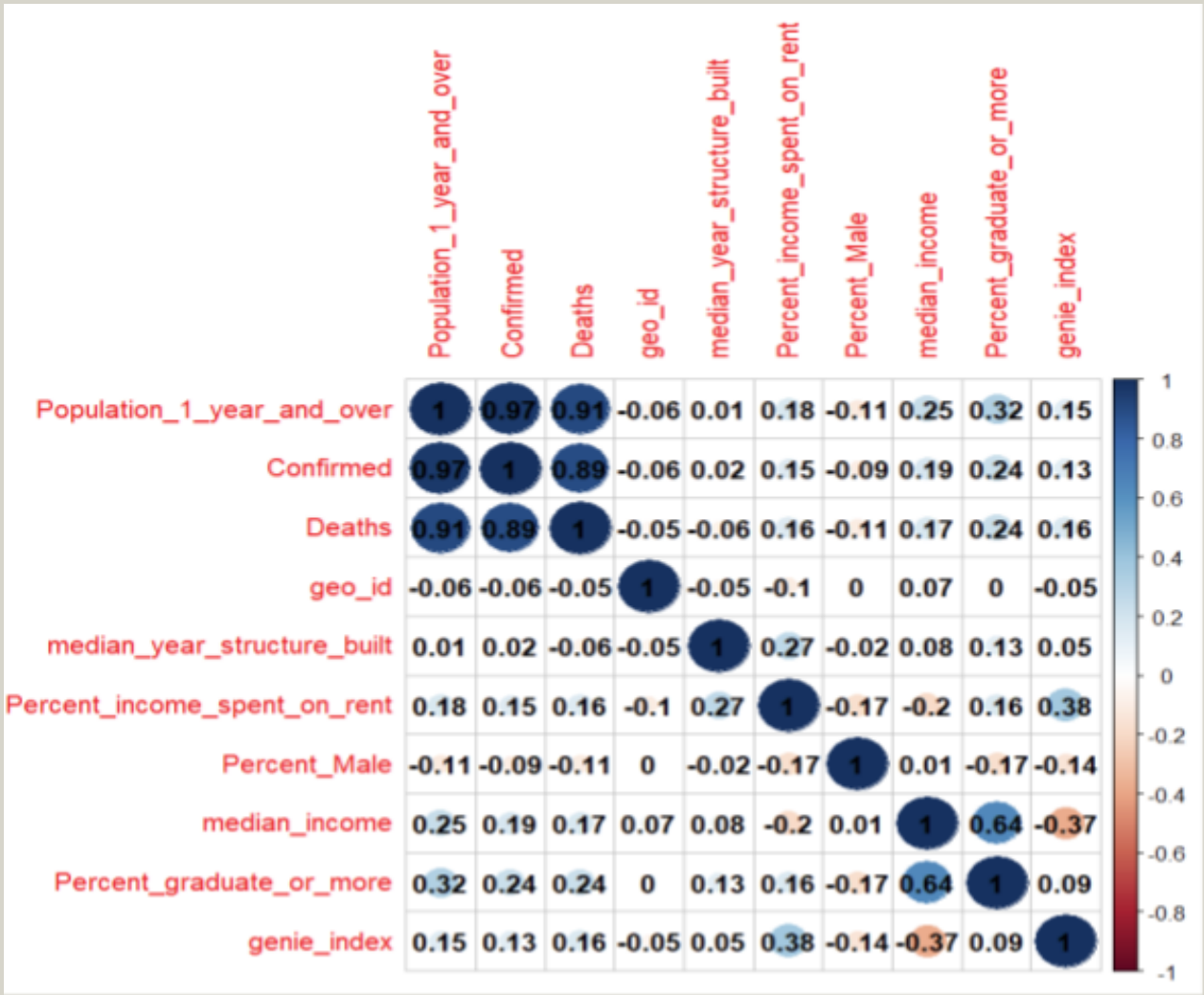


Figure 6: Correlation Matrix of National Census Attributes

Given that Figure 6 only contains quantitative attributes, it is also important to understand the distribution of important numerical features based on categorical variables. Considering my role as a data analyst for Dallas, it would be insightful to compare the distribution of confirmed cases per county on a violin chart for Texas, Florida, California, Georgia, Arizona, and Louisiana. The reason I selected these states is that they have similar seasonal weather patterns to Texas. As we saw in Figures 1-3, COVID cases increase steeply during the winter. Thus, for a more fair comparison, it is best to select states that will rise in COVID cases around the same time as Texas. Furthermore, since we are not looking at the number of confirmed cases, but rather, the distribution of those cases relative to the state mean, population size is not a huge factor.

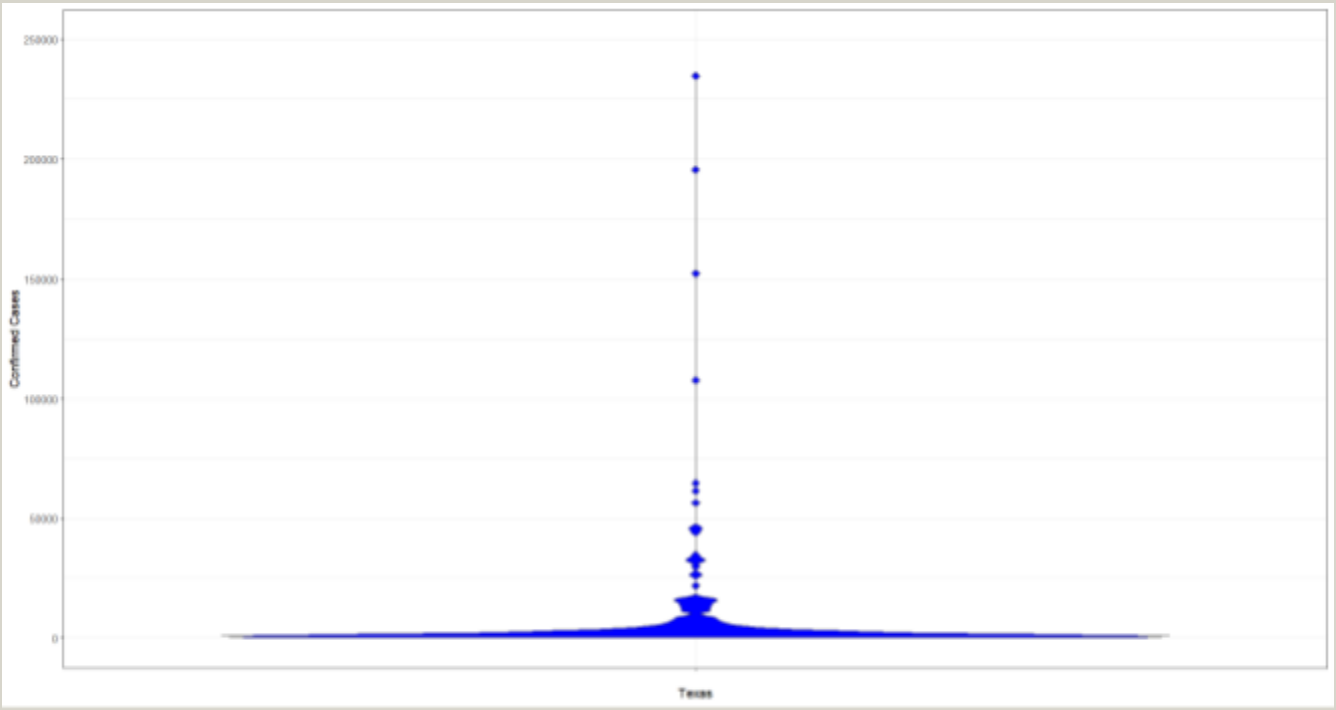


Figure 7: Distribution of Confirmed Cases in Texas

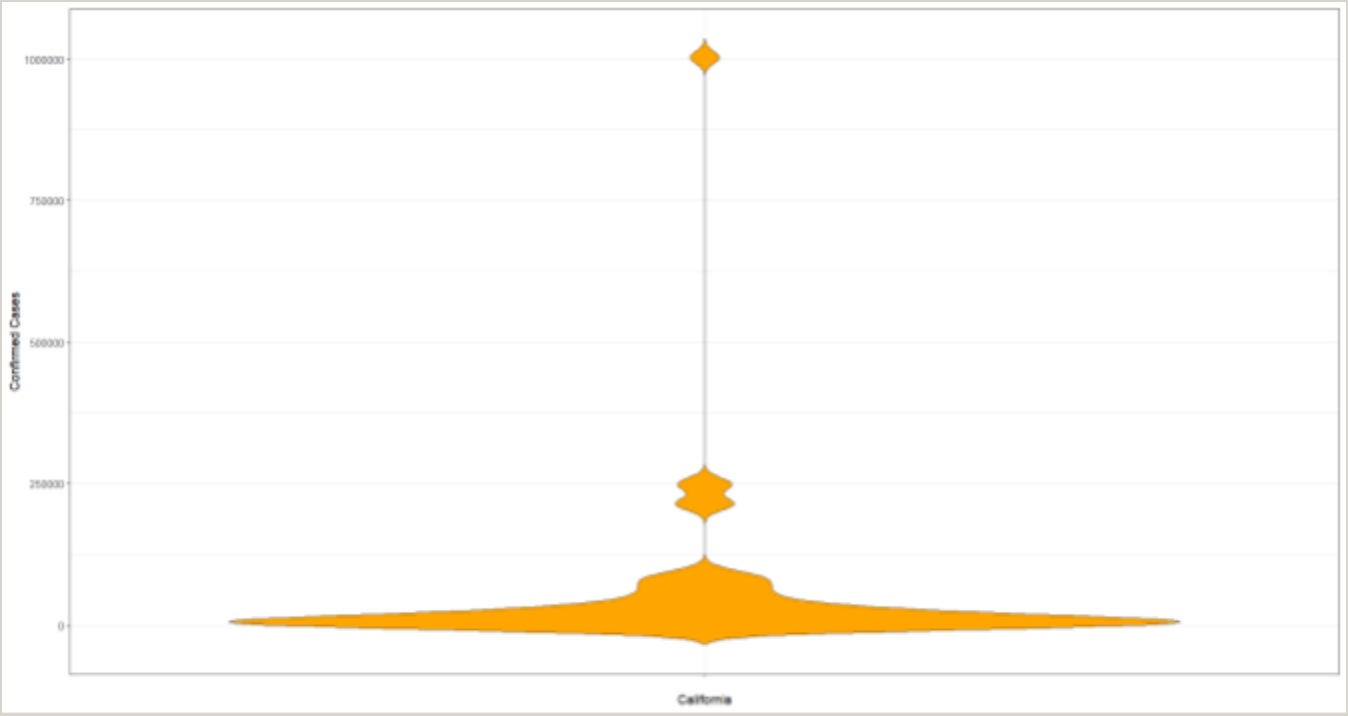


Figure 8: Distribution of Confirmed Cases in Califonia

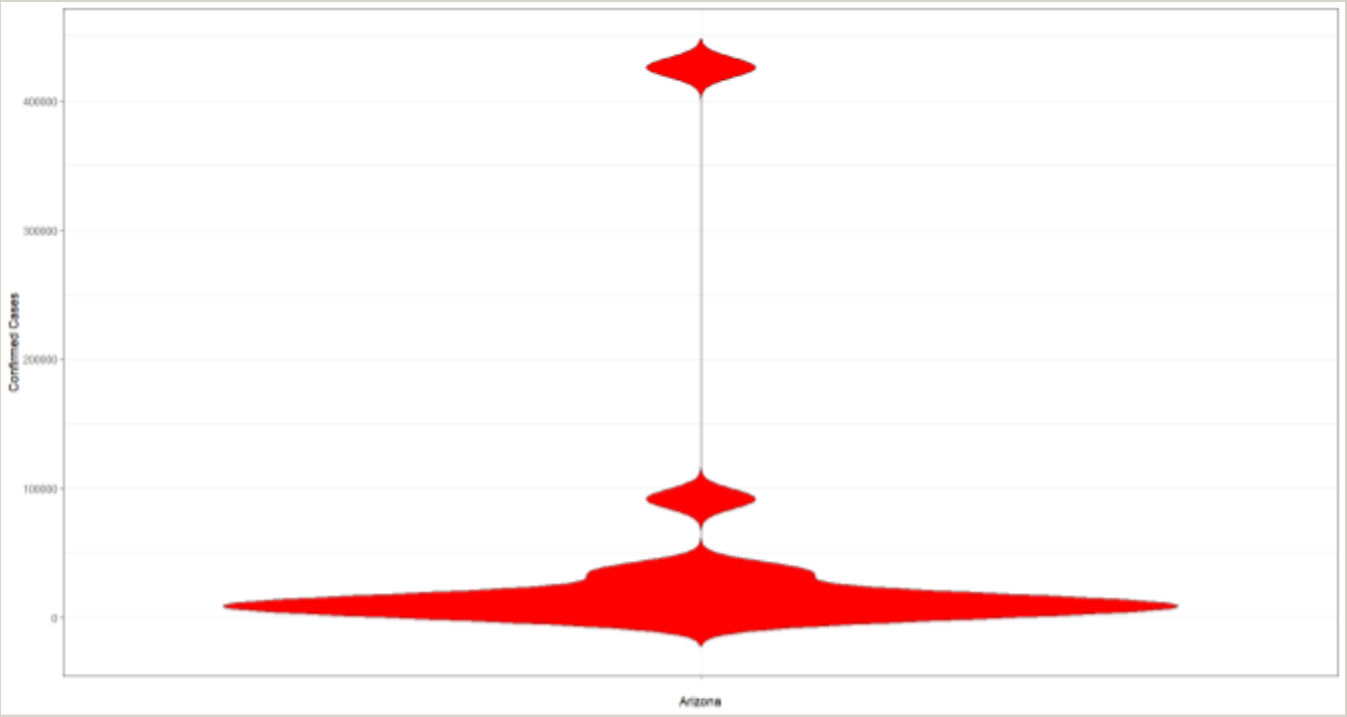


Figure 9: Distribution of Confirmed Cases in Arizona

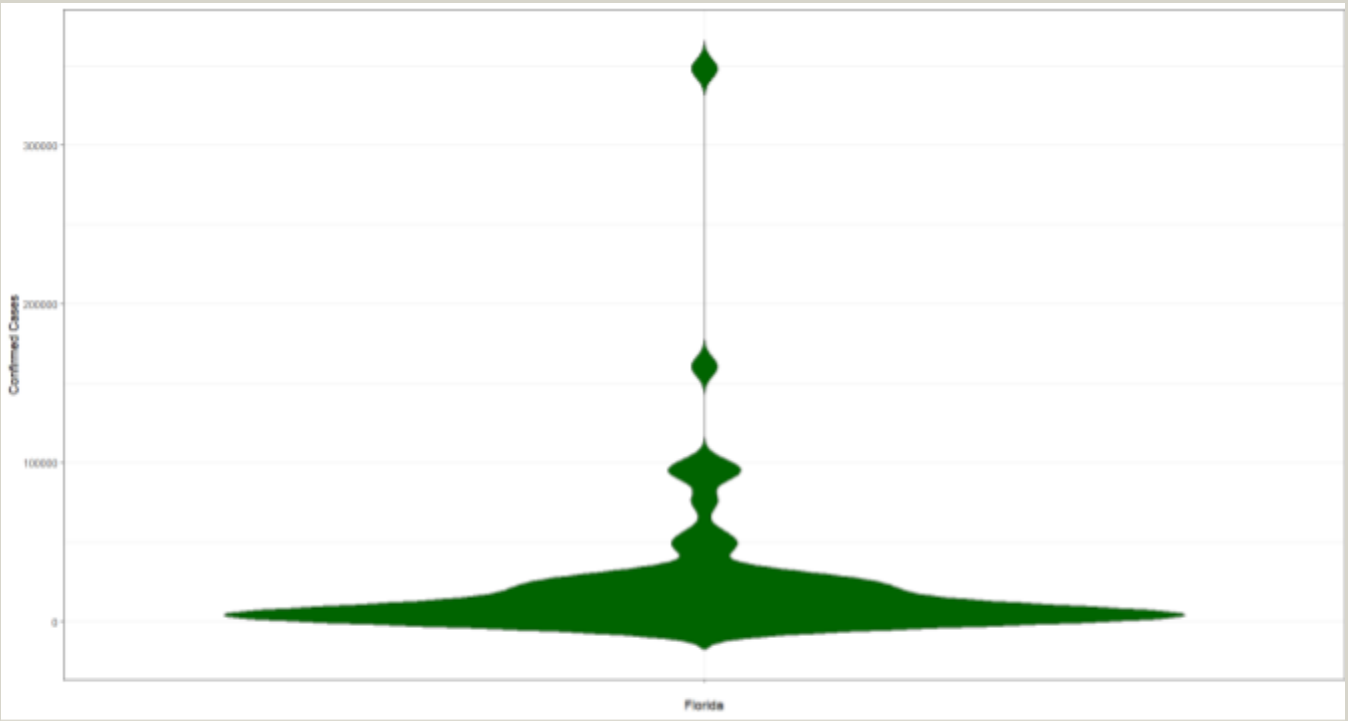


Figure 10: Distribution of Confirmed cases in Florida

State	Confirmed Cases Mode
TX	487
CA	40
AZ	8705
FL	1498

When looking at the figures 6 through 10, we see that all the states share a similar distribution with the large number of counties having less than 9000 cases in all the states, as shown by the mode for each of the states above. However, we see the distribution of the counties with a higher number of cases in different. In the case of Texas, we see each individual county that has greater than 50,000 confirmed cases marked with a blue diamond. The distribution of these counties appears to be distributed in a sporadic manner that is ungrouped/unrelated. However, in California and particularly Arizona, we see that counties with a much higher amount of cases are actually grouped together. In the case of Arizona, there is one group of counties that is centered around 90,000 confirmed cases and another even larger swath of cases that are centered above 400,000 cases. And when looking at Figure 10, Florida's violin chart shows us that there are a significant clustering of counties that have 50,000 to over 100,000 confirmed cases that is connected with the "base counties" in the chart. This clustering of various districts together suggests that perhaps the specific location of a particular county in relation to other counties also plays a role in COVID spread. In essence, in the case of Florida, California, and especially Arizona, particular regions of the state might be performing better or worse as a group due to geographic location, regional policies, etc.

❖ Data Preparation

How fast is the virus spreading in Dallas in relation to other similar counties in other states?

To answer this question, I must firstly choose which counties are considered most similar to Dallas. In the article "What is Your Cities Twin" in the New York Times, the city of Dallas is most similar to Chicago, and Houston due to the job mixes present in the cities.

The subset I will use for this analysis will consist of the following columns with additional data for Atlanta and Chicago being found online. With this dataset, I will create a line chart to compare the spread of the virus:

Column	Data Type	Description
Date	Date	The particular day from 1-22-2020 to 1-25-2021
Confirmed_Cases_Dallas	Integer	The number of confirmed cases on a particular day in Dallas County
Confirmed_Cases_Chicago	Integer	The number of confirmed cases on a particular day in Cook County. (As reported on City of Chicago Site)
Confirmed cases_Houston	Integer	The number of confirmed cases on a particular day in Harris county

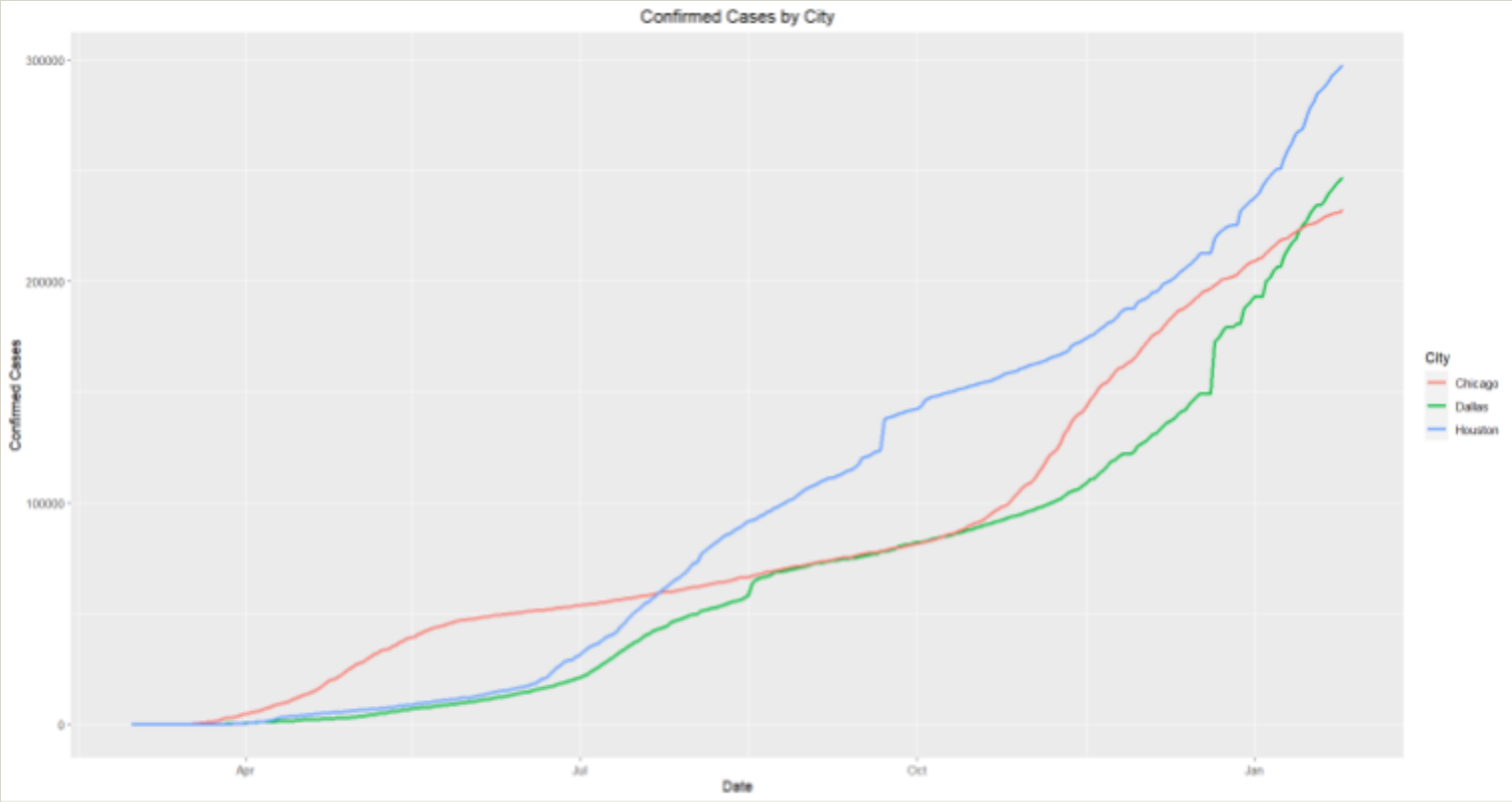


Figure 11: Cities confirmed COVID cases from March 2020 to January 2021

The results of the chart above are interesting as one would anticipate that Chicago, a city with a much larger population than Dallas with harsher Winters, would have had a much more significant increase in cases during the Winter compared to Dallas and Houston. Though Chicago does see an uptick in cases beginning in November, we do see the slope of the graph begin to flatten during December and January months. Upon further research, it is important to consider that Chicago implemented statewide restrictions in November of 2020 that reduced capacity limits for retail shops and shut down casinos and other entertainment centers as a means of reducing expected statewide surges. Meanwhile, during the October and November months, Texas was attempting to reopen bars and restaurants and 50% and 75% capacity. In fact, it wasn't until December 2nd that Texas rolled back restrictions after Hospitals were reaching capacity and being overwhelmed. As the graph depicts, the reactionary rather than proactive policy that Texas implemented made it not possible to reduce the spread of COVID during the Winter months.

Does aggregate time spent driving to have a relationship with median income and confirmed cases?

And is the relationship positive for both of these attributes?

For this question, I will have a subset of the dataframe with aggregate time spent driving, median income, and confirmed cases as the columns. I will create two correlation line charts.

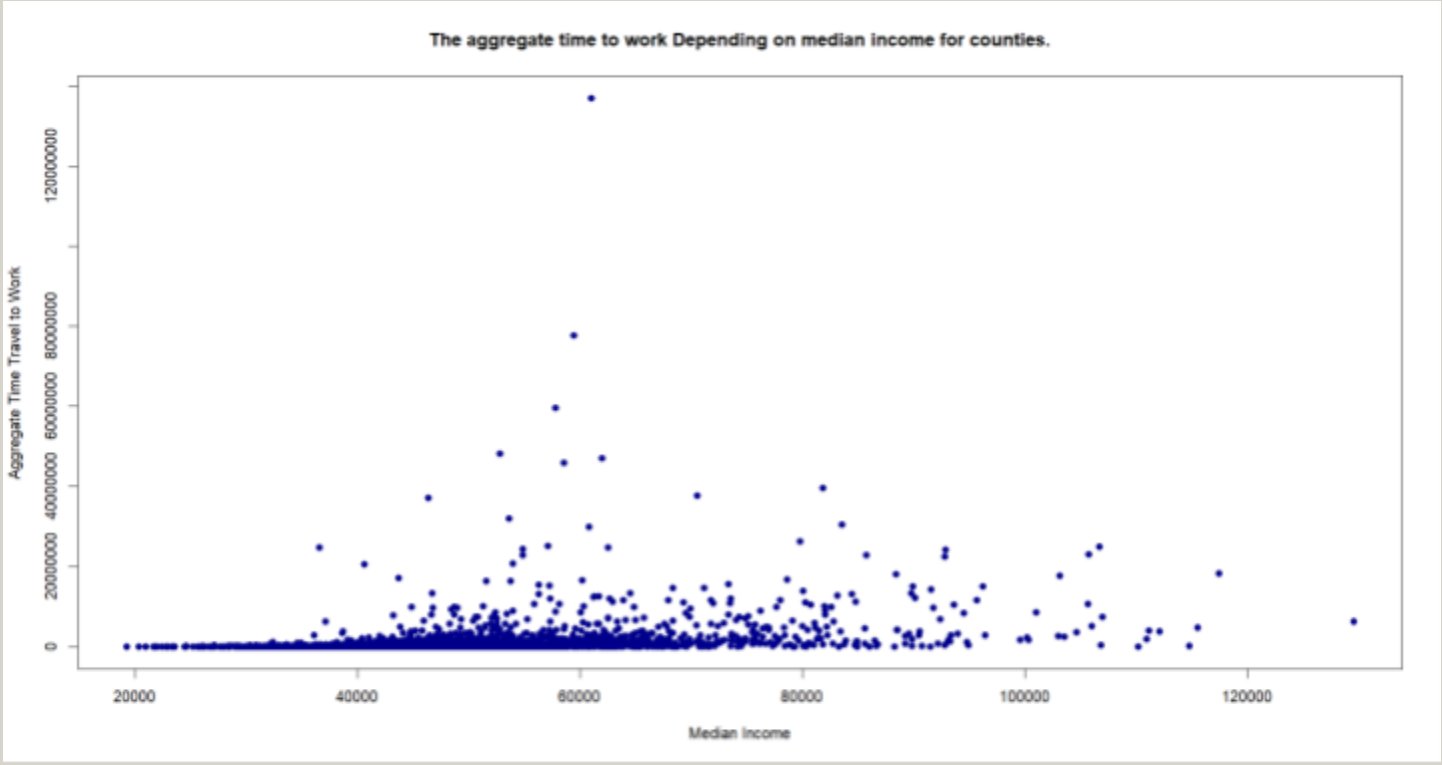


Figure 12: Aggregate Time Travel to Work depending on Median Income

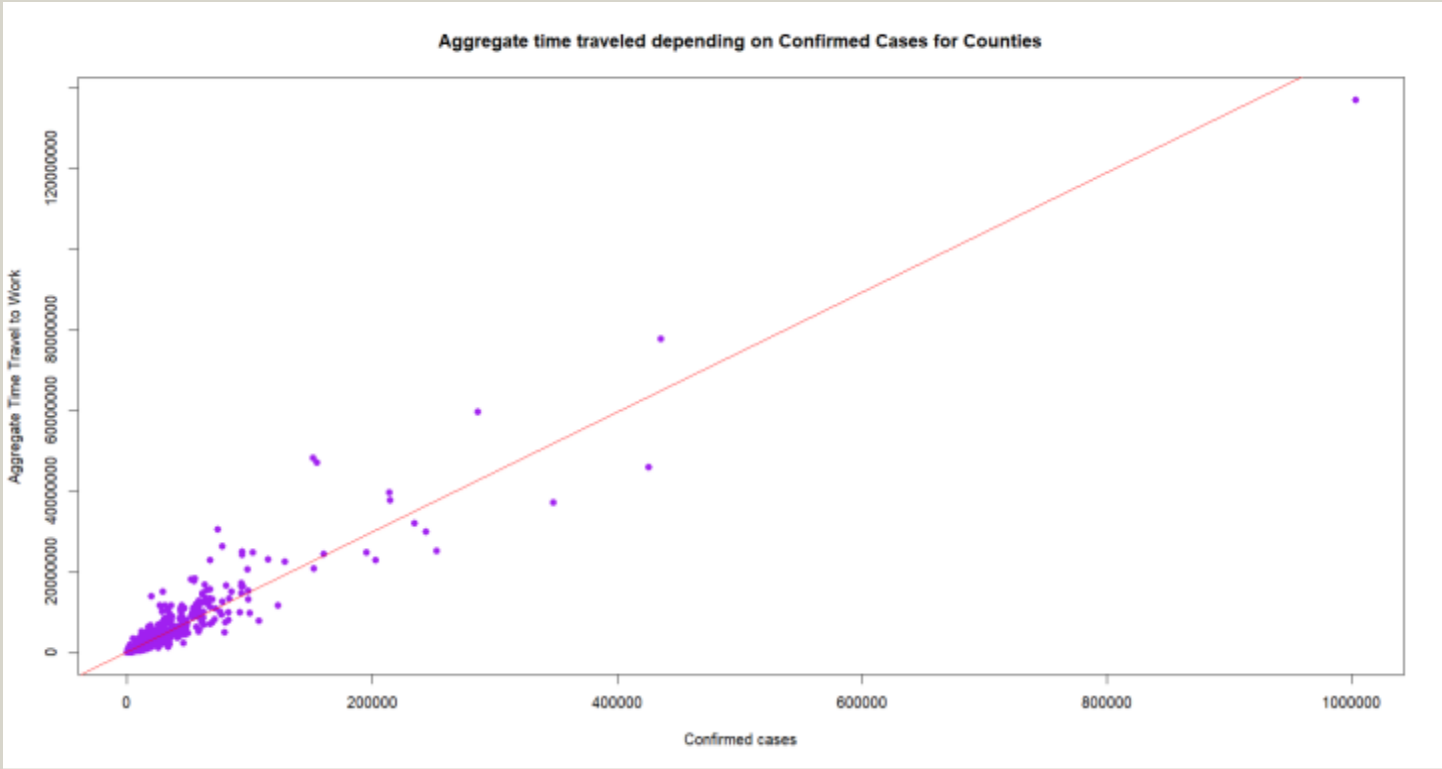


Figure 13: Aggregate Time traveled dependent on Confirmed Cases

Figure 12 shows that there is little linear correlation between median income and aggregated time traveled to work. We also see that most counties are clustered near the bottom of the graph in terms of aggregated time traveled. This does align with the fact that many businesses in January 2021 were utilizing a hybrid or work-from-home model that reduced or eliminated the need to travel to work. We also not that higher income communities are more likely to have an increased amount of aggregated travel hours than low-income communities. This could suggest a few things. Firstly, the higher income counties may have a higher proportion of individuals in higher leadership positions within a firm, and thus, they are required to physically come to work more often. Furthermore, it is important to not forget that many individuals working regular jobs during COVID were suspended and eventually laid off during the pandemic, especially prior to the start of the new year. Thus, the proportion of individuals with a high number of aggregated hours on lower income brackets might be explained by the fact that they didn't have a job to go to.

Figure 13 shows a moderate linear correlation between the number of confirmed cases and the aggregate work travel time for a community. This would be expected, considering that individuals who are physically traveling to work will be in physical contact with others, which increases the risk of getting COVID from someone else.

Conclusions and Recommendations

- Population is the predominant factor in the number of cases and deaths in a particular location, as demonstrated by the high correlation between the attributes shown in figure 6.
- The distribution of COVID cases in counties also may be dependent on the location of a particular county and neighboring counties. This is demonstrated by figures 7-10, which shows how Texas has isolated cities with higher numbers of COVID cases while Florida and Arizona have groups of counties that each have a similar amount of increased cases.
- The number of confirmed COVID cases has an upward trend and is seasonal; however, counties that perform better or worse are also heavily influenced by policy decisions.

- Counties should plan, enact, and implement COVID policies in isolation, but rather, should do so with neighboring counties on a regional and state level. State leaders should be prepared to implement policies on a region-by-region basis rather than doing so for the entire state. Thus, one district in a state might be completely locked down while another is more open.
- COVID policies should be done proactively rather than reactively to limit the extent to which COVID cases increase during a surge period. Thus, we should be ready to implement restrictions in November in anticipation for the Winter surge.

Works Cited

About DSHS | Texas DSHS

CDC Museum COVID-19 Timeline | David J. Sencer CDC Museum | CDC

COVID-19 Daily Cases, Deaths, and Hospitalizations | City of Chicago | Data Portal
Fulton County, Georgia coronavirus cases and deaths | USAFacts

https://www.texastribune.org/2021/01/22/coronavirus-texas-hospital-capacity/

Two years of the COVID pandemic in Chicago: Take a look back – Chicago Tribune

What Is Your City's Twin? - The New York Times (nytimes.com)
deaths | USAFacts