PROJECT 1: DATA AND VISUALIZATION

Data Analysis on the COVID-19 Virus

Prepared by: William Chavula, Hareish Raghupathy, Kevin Sanchez

Table of Contents

Table of Tables

Table of Figures

Abstract

This report is prepared for the Department of Treasury, the Bureau of the Fiscal Service, and the Internal Revenue Service (IRS). This document offers a preliminary analysis and assessment of the people groups, age ranges, and counties most susceptible to financial burdens due to the pandemic. To this end we analyzed a total of 3 datasets from which we selected and derived 23 features. We applied a set of transformation techniques to our data, performed correlation analyses, normalization calculations to uncover trends and insights in confirmed cases of Covid in relation to various features within the data. We also performed a series of bivariate analyses on features such as percent change in retail and recreation in a geographic area to see how the Pandemic influenced people going out to retail shops, recreation centers, and to workplaces. Some of our findings include: a particular age range that is most likely to contract covid, a relationship between number of cases and counties where most residents worked from home, and relationships between retail and recreation mobility to Covid induced death. We used several visualizations are used to communicate our findings and insights throughout the report.

Business Understanding

COVID-19, also known as the Coronavirus Disease 2019, is a highly infectious respiratory illness caused by the novel coronavirus SARS-CoV-2. It was first detected in Wuhan, China, in December 2019 and quickly spread to other countries, leading to a pandemic declaration by the World Health Organization (WHO) in March 2020 [1].

Social distancing is a public health practice aimed at reducing close contact of sick people with healthy people. It involves maintaining physical distance from others (about 1m from others), preventing large gatherings, and wearing masks. The goal of social distancing is to slow down person-to-person transmission of COVID-19, thereby slowing its spread within the community [2].

Flattening the curve refers to another public health practice aimed at reducing the spread of the virus to a level where healthcare systems are not overwhelmed. By practicing things like social distancing, testing, contact tracing, and quarantine, the curve can get flatter when the number of cases spread out over a longer period, preventing overwhelming high, narrow curves due to the increase in infections [3].

It is important to look at data about the spread of the virus, hospitalizations, and available resources to evaluate the current state of the pandemic, predict future trends, and make informed decisions to protect public health. Governments, public health officials, healthcare providers, researchers, businesses, and the public are all interested in this information.

Data on virus spread, hospitalizations, and resource availability can inform decisions such as figuring out when and where to implement measures like lockdowns, mask mandates, and restrictions on social gatherings. This data can also aid in the distribution of medical supplies, vaccines, personnel, and equipment to areas that are most affected by COVID 19. Furthermore, this data can help different governments produce different strategies to support affected industries and mitigate economic recessions that may occur because of the pandemic. The U.S. Department of the Treasury will use COVID-19 data to inform several key decisions regarding stimulus checks such as determining who is eligible to receive stimulus checks, deciding on the amount of stimulus payments individuals or households will receive, determining how stimulus checks are distributed to eligible recipients, and deciding on the timing of stimulus payments.

For this project, we want our stakeholder to be the U.S. Department of the Treasury who specializes in Economic Impact Payments [4]. These are the stakeholders in charge of sending out Stimulus checks during the COVID-19 crisis. The U.S. Department of the Treasury will use COVID-19 data to inform several key decisions regarding stimulus checks such as determining who is eligible to receive stimulus checks, deciding on the amount of stimulus payments individuals or households will receive, determining how stimulus checks are distributed to eligible recipients, and deciding on the timing of stimulus payments. We will take a closer look at the stimulus checks sent in the TX counties.

The effectiveness of stimulus checks in stimulating economic recovery and alleviating hardship amidst the COVID-19 pandemic is a subject of ongoing debate. When the U.S. began to shut down in the wake of the onset of Covid-19, it meant millions of families were suddenly without the income they needed. Within weeks, Congress passed a massive emergency aid package aimed at providing relief, including "economic impact payment" checks of up to $1,200 per eligible adult. There has been a total of three rounds of such checks, including additional payments of up to $600 and $1,400 per person in 2021, referred to as "stimulus checks”. While the government had deployed stimulus checks before, particularly in the wake of the Financial Crisis, the size and scope of the direct checks were unprecedented. With each stimulus check, the IRS and the U.S. Department of the Treasury became faster and more efficient at deploying the money. However, "there were glitches along the way, including some initial checks sent to deceased Americans," and concerns about the money reaching "well-to-do taxpayers who were unaffected financially by the pandemic." While the payments were not as targeted as they could have been, "lawmakers prioritized speed in getting relief out fast because of the nature of the pandemic." Despite efforts to reach non-tax filers and those most vulnerable, "some individuals may have still fallen through the cracks, facing difficulties accessing the funds." Moreover, as the payments were deployed, there was a shift in how they were spent, with a decrease in household spending and an increase in saving or paying down debt, likely influenced by changing economic conditions. Regarding inflation, there are debates about whether the stimulus money may have fueled inflationary pressures. Some argue that "the size of the relief overall, including stimulus payments, contributed to higher inflation," but others suggest that "inflation has been driven more by supply constraints and other factors rather than excessive stimulus."[5]

Data Understanding

For this project, we will be focusing on three datasets that we are collecting COVID 19 data from. These data sets are collected from Google which contains datasets from the USA and around the world. For the sake of our project, we are only focusing on data from the state of Texas. We will be analyzing the following three datasets:

- COVID-19 cases plus census
- COVID-19 cases TX
- Global Mobility Report

We want to be able to describe the type of data of the most important variables of the data set, verify the quality of the data, give some sample statistics for these important variables, visualize these prominent features, and explore relationships between attributes by seeing their correlation.

COVID-19 cases plus census Data Set

For the first dataset we analyzed the Covid-19 and census dataset. This dataset contains 250 features and 3,142 samples. For the purposes of this analysis, however, we only consider 30 of the features. These are described in the table below:

Table 1: Description of Features of Interest their respective data types for Covid Cases plus census

Features	Data Type	Description
Country name	Nominal	Name of the county in the state
State	Nominal	Name of the State
Date	Interval	Date when recordings or observations were made
Confirmed cases	Ratio	Number of people that tested positive for Covid-19
deaths	Ratio	Number of people whose death was reported as caused by Covid-19
rent_over_50_percent	Ratio	Housing units spending over 50% income on rent
rent_40_to_50_percent	Ratio	Housing units spending 40% - 49.9% income on rent
rent_35_to_40_percent	Ratio	Housing units spending over 35% - 39.9% income on rent

rent_30_to_35_percent	Ratio	Housing units spending over 30% - 34.9% income on rent
total_pop	Ratio	Total Population. The total number of people living in a given geographic area
Median Income	Ordinal	Median household income. Within a geographic area, the median income received by every household on a regular basis
Median rent	Ordinal	The median contract rent within a geographic are.
percent_income_spent_on_rent	Ratio	Percent of household income spent on rent
households_public_asst_or_food_stamps	Ratio	Households on cash public assistance or receiving food stamps
pop_determined_poverty_status	Ratio	The number of people living in a geographic area who could be identified as living in poverty or not. It includes people whose status could not be determined
poverty	Ratio	Income in the past 12 months below poverty level. The number of people in a geographic area who are part of a family determined to be in poverty
median_age	Ratio	The median age of all people living in a given geographic area.
commuters_by_public_transportation	Ratio	The number of workers aged 16 years and over within a geographic area who primarily traveled to work by public transportation.

walked_to_work	Ratio	The number of workers aged 16 years and over within a geographic area who primarily walked to work
worked_at_home	Ratio	The count within a geographical area of workers over the age of 16 who worked at home.

Data Quality

One of the selected features, ‘median rent,’ for our analysis has missing data. Since only 2 out of 3,142 samples, representing a 0.064%, contain missing values, we decided to drop these rows concluding that this will not hamper the analysis in any way.

Table 2: Displaying the number of missing values per feature.

Feature	Number of Missing
median_rent	2
county_name	0
State	0
Date	0
confirmed_cases	0
Deaths	0
rent_over_50_percent	0
rent_40_to_50_percent	0
rent_35_to_40_percent	0
rent_30_to_35_percent	0
total_pop	0
median_income	0
households_public_asst_or_food_stamps	0
pop_determined_poverty_status	0
poverty	0

Summary Statistics

To better understand the data, we investigate the nature of the relationships that may exist in the data. Furthermore, we calculate correlations between the features to calculate the strength of the relationship and whether it is significant.

Table 3: Descriptive statistics for Covid Cases plus census

Variable	N	Missing	Mean	SD	Min	Q25	Mdn	Q75	Max	Skewness	Kurtosis
confirmed_cases	3140	0	7,563.70	28,108.18	0.00	797.75	1,917.00	4,955.25	1,002,614.00	18.79	553.34
deaths	3140	0	124.91	481.00	0.00	12.00	31.50	77.00	13,936.00	14.70	314.58
rent_over_50_percent	3140	0	3,239.14	14,632.42	0.00	162.00	486.50	1,576.75	536,832.00	19.67	605.93
rent_40_to_50_percent	3140	0	1,167.94	4,996.10	0.00	60.00	193.50	615.25	177,284.00	18.28	534.27
rent_35_to_40_percent	3140	0	849.09	3,558.21	0.00	42.00	136.00	449.50	124,762.00	17.80	510.80
rent_30_to_35_percent	3140	0	1,157.30	4,766.05	0.00	67.00	194.00	618.00	161,522.00	16.83	455.75
percent_income_spent_on_rent	3140	0	27.82	4.40	10.00	25.30	28.10	30.30	50.00	0.01	4.45
total_pop	3140	0	102,229.93	328,386.87	74.00	10,967.00	25,704.00	67,500.75	10,105,722.00	13.92	326.21
median_income	3140	0	49,737.40	13,139.89	19,264.00	41,121.00	48,048.50	55,761.25	129,588.00	1.26	6.08
median_rent	3140	0	563.36	214.81	140.00	424.00	510.50	642.00	1,879.00	1.82	7.91
median_age	3140	0	41.15	5.38	21.60	37.90	41.20	44.20	66.40	0.04	3.70
commuters_by_public_transportation	3140	0	2,422.88	24,171.51	0.00	6.00	33.00	145.25	735,534.00	20.76	507.43
walked_to_work	3140	0	1,289.49	6,012.14	0.00	98.00	243.00	679.50	181,289.00	17.04	393.93
worked_at_home	3140	0	2,238.01	8,078.83	0.00	178.75	420.00	1,228.50	249,490.00	14.08	331.09
households_public_asst_or_food_stamps	3140	0	5,035.59	15,783.80	0.00	575.75	1,481.00	3,635.25	333,729.00	10.85	167.79
pop_determined_poverty_status	3140	0	99,696.26	323,062.17	70.00	10,580.50	24,804.00	65,151.00	9,955,473.00	13.97	328.20
poverty	3140	0	14,538.26	51,717.03	10.00	1,631.75	4,125.50	9,912.25	1,688,505.00	16.00	406.99

This table gives us a picture of how the data is shaped. Looking at the Skewness column specifically, we see that all the values are positive (above zero). This non-symmetric shape of the data can also be seen by comparing columns **M** and **Q50**. We observe that for all the columns, the mean is larger than Q50, the median. Figure 1 below shows histograms that visually show the skewness of the features in the data. The data is skewed right, meaning, a lot of the observations are collected around the small values and a relatively small number have high values.

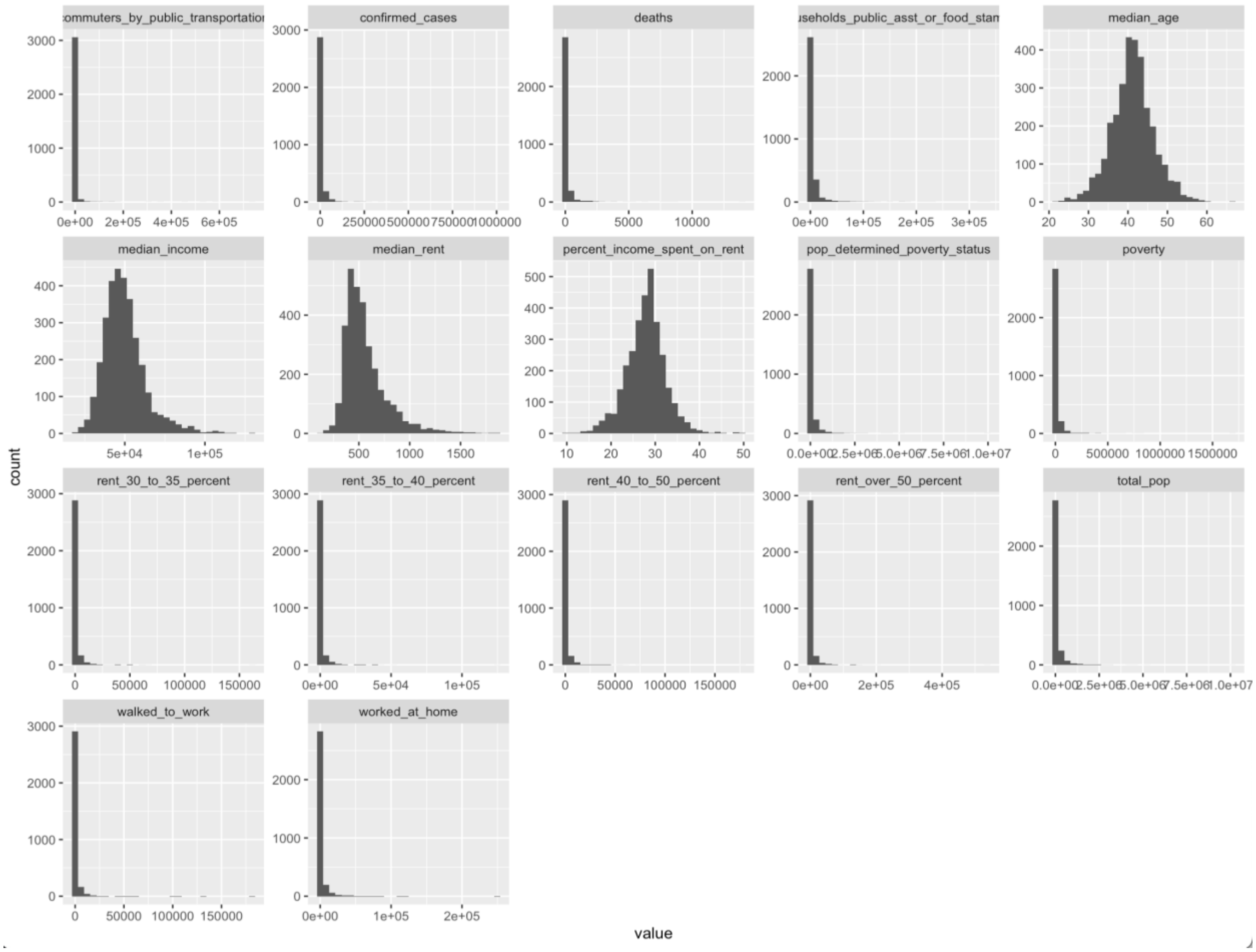


Figure 1 Distribution of each numeric feature in the dataset.

We will have to deal with the skewness in our data since most statistical and machine learning methods assume the data is distributed symmetrically. Two ways we can change the shape of the data are either transforming the data by standardizing it with z-scores or by using logarithmic transformation.

Aside from skewness, we also observe that the difference between the minimum and maximum values is quite large. Meaning that the features in the data have high variance. And that some if not all features have outliers as can be seen from the plots below.

Data Correlation

Next, we investigate the sort of relationship 2 variables have with each other. That is, are the features in our dataset correlated, and if so, how are they correlated, positively, negatively, or neutral. We used a heatmap to visualize these correlations. Note that the numeric variables have been normalized by dividing the values with the total population and multiplying by 1000. This gives us the metric per 1000 people.

Some highlights in this plot are the inverse relationship that can be observed between median income and the features poverty_per_1000 and hh_asst_or_food_stamps_per_1000. This is in line with our expectations, as a person's income increases, we expect them not to be categorized as being in poverty status and vice versa. Another interesting relationship to observe is the strong correlation between poverty_per_1000 and hh_asst_or_food_stamps_per_1000. This makes sense because as the number of people receiving public assistance and living on food stamps increases, we expect that the poverty level in that individual or household to also increase. We also observed an inverse correlation between confirmed_cases_per_1000 and worked_at_home_per_1000 which is expected.

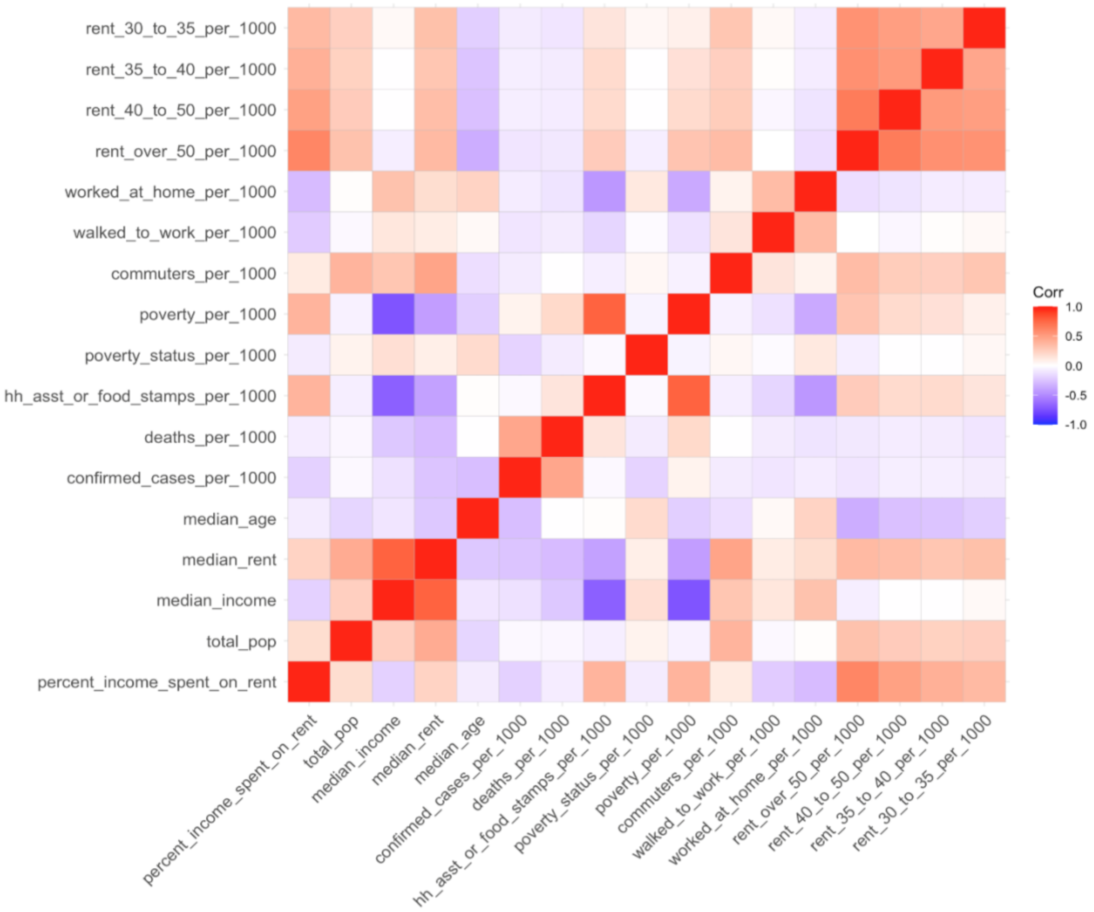


Figure 2 Correlation heatmap.

Data Visualization

How many people receive assistance and are on food stamps in each state?

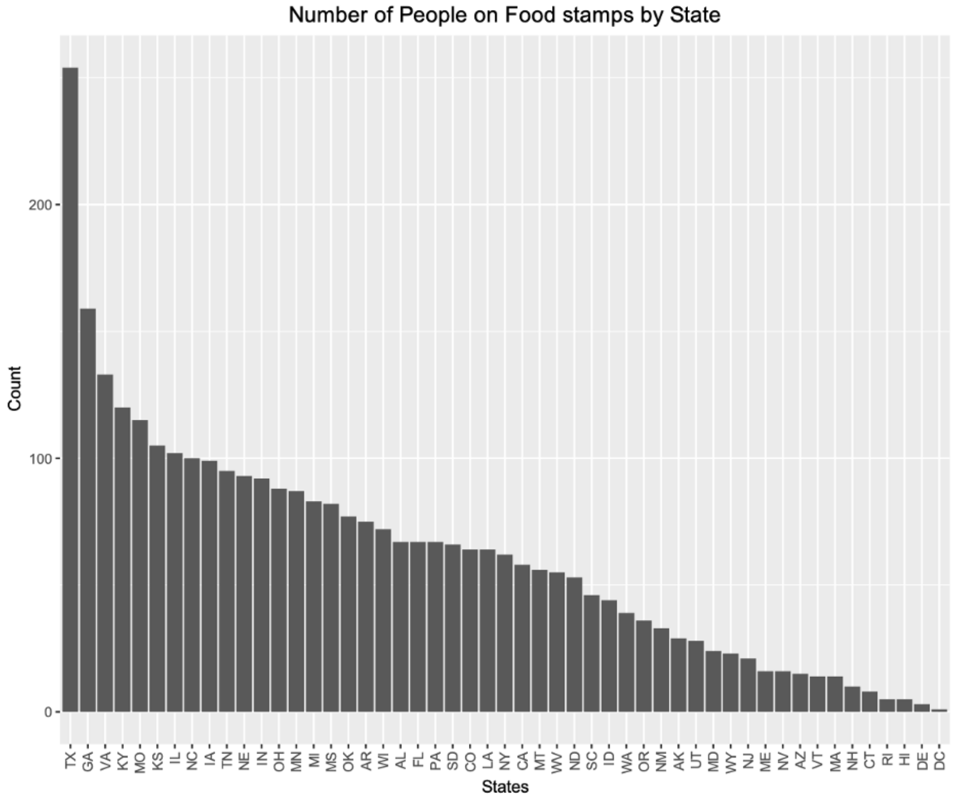


Figure 3 Number of people on food stamps by state.

This graph shows that Texas has the largest number of people that are receiving assistance and living on food stamps.

By State, what is the median income versus what people pay for rent?

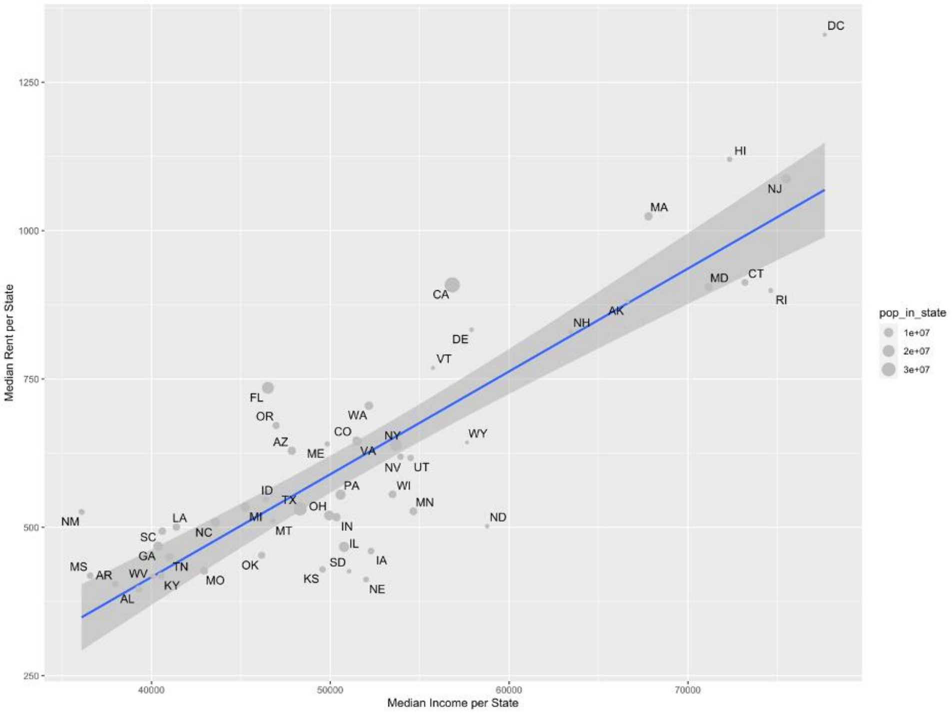


Figure 4 Median income versus median rent by state.

Unsurprisingly, Washington DC has the highest median income but also the highest rent in the nation. The point size is the total population in the state. Relative to other states like California, Texas, and New Jersey, DC's population size is quite small.

What is the distribution of total confirmed cases versus deaths in US States?

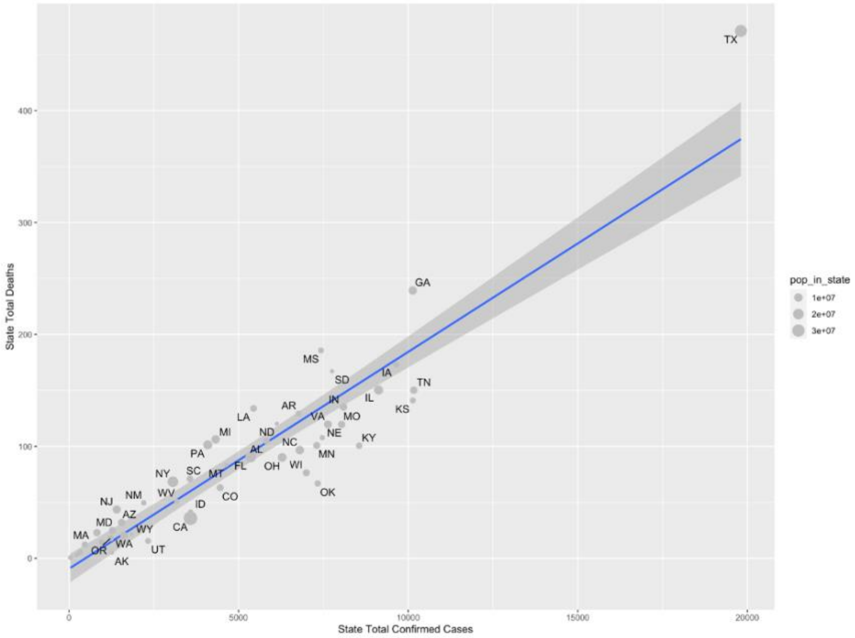


Figure 5 Total confirmed cases versus deaths by state.

From this visualization, we see that Texas had the highest number of confirmed cases as well as recorded deaths. This is an interesting observation because California, which has a similar population size (reflected in the size of the points) to Texas, is less than a quarter the number of cases in Texas. We'll need to investigate how well preventive measures such as social distancing and wearing masks were being practiced in Texas.

COVID-19 cases TX Data Set

This set of data sourced from CDC, state and health agencies highlights the day-to-day case count and number of deaths due to COVID-19 across various counties within Texas. There are 7 unique features consisting of 94350 rows of observation in here out of which we selected the most informative and key features to answer questions pertaining to our area of interest i.e., the economic stimulus check.

Post eliminating the unwanted features, we could see that the total observation has been reduced from 660,450 to 377,400, effectively cutting it down by half.

The total number of counties across Texas has been identified as 254 with an extra category for the State-Wide Unallocated data making the total 255. To give more insight into the time frame at which this data has been captured during covid, we could see that each of these counties have data recorded across a full year starting from 1/22/20 (Jan of 2020) to 1/25/21 (Jan of 2021).

It is especially important to note that our data is recorded each and every day rather than having a total or average of a particular county. Hence, we are required to group the data based on metrics like county date or severity etc. for timeline computation which we will see in the upcoming feature visualizations to understand the dataset better. This time-related data will prove to be valuable when we try to analyze the trend for the spread of covid throughout the year. Further, this will enable us to identify the key factors that have led to the accelerated spread of covid if the respective data is given.

The features extracted from the above set are: “county_name,” “date,” “confirmed_cases” and “deaths.”

These factors have been deemed to be useful as they can help researchers and scientist to derive insights into the economic impact of covid in Texas counties based upon the shift in economic trend ‘when there is a confirmed case’ for a person or ‘when there is an unfortunate death recorded’ in that area with respect to its timeline.

This can help us study and accurately narrow down the impact of COVID-19 cases and deaths on the economic fluctuations at a larger scale.

The table below shows the finalized features form the Texas Covid-19 data set:

Table 4: Description of Features of Interest their respective data types for Covid cases in TX

Feature	Scale of Measurement	Description/Information
county_name	Nominal	This column provides the names of different counties in Texas. They are considered nominal in this case as they represent categories with no inherent order or ranking.
confirmed_cases	Ratio	The number of confirmed cases in Texas with respect to each and every county is provided by this feature. This enables us to assess the impact of covid spread in the counties.
deaths	Ratio	The number of confirmed deaths in Texas with respect to each and every county is provided by this feature. Death as a ratio allows us to calculate the relative risk impact of covid across the different counties.
date	Interval	The date of each and every case data recorded across the various counties for a year. This proves to be useful to provide a timeline of every recorded data regarding the

		above variable and how they change with time.

Summary Statistics

Table 5: Descriptive statistics for Covid Cases in TX

	Described Variable	N	N/A	Mean	SD	SE_M	Min	Max	Q25	Q50	Q75	Skewness	Kurtosis
1	confirmed_cases	94,350	0	2,158.5704	11,940.338	38.8727	0	297,629	1.00	82.00	639.75	12.4927	196.6726
2	deaths	94,350	0	38.1899	187.987	0.6120	0	4,024	0.00	2.00	15.00	10.6570	139.1378

Some of the statistics discussed above are:

- N or Number of values
  - This shows the total number of values present in the dataset.
  - Both the respective variables have around 94,350 values in this case.
- N/A or Missing
  - This tells us the number of missing values in the dataset in the respective column.
  - We do not have any missing values in this dataset which proves to be helpful while doing the analyses as we do not have to do separate imputations or other processes to ensure the reliability of the data for computation.
- M or Mean
  - Shows us the computed average value across 94,350 values in each variable.
  - Calculated by (sum of all values/total number of values), it shows what is the average value you could find in the column while accounting for all the other values in the same column.
- SD or Standard Deviation
  - It is the measure of dispersion/variability of the values across the dataset.
  - The standard deviation here is high due to inherent variability in recorded covid case counts in different counties across a full year every day.
- SE_M or Standard Error Mean
  - It is the measure of precision or accuracy of the dataset. It quantifies the variability of sample statistics like mean in this case.
  - While the SD is high, the SE provides a more precise estimate of the Mean as the difference between them is quite negligible.
- Min or Minimum
  - It is the minimum value found across the respective variable.
  - The minimum covid cases or deaths recorded here is 0.
- Max or Maximum
  - It is the minimum value found across the respective variable.

    - The maximum cases and deaths recorded in the dataset are 297,629 and 4,024 respectively on a single day.
- Q25 or First Quartile(Q1)
    - This separates the lowest 25% of the data from the rest, therefore known as lower quartile.
- Q50 or Second Quartile(Q2)
    - This separates the dataset into two equal parts by 50%.
- Q75 or Third Quartile(Q2)
    - This is the value below which 75% of the data falls under and separates the lowest 75% from the highest 25%, hence known as upper quartile.
- Skewness
    - It is the measure of the asymmetry of a data distribution. It indicates if the data is positively skewed, negatively skewed or symmetric in nature.
    - A skewness close to zero is generally considered balanced, however in our case it is significantly skewed to one side (12.49 and 10.65) showing us the increasing numbers recorded across the year. This can further be seen in the upcoming visualizations where we will be able to find the data to be heavily concentrated on one side in this dataset.
- Kurtosis
    - It is the measure of peakedness of a distribution. A sharp peak and heavy tail indicated high kurtosis showing the presence of more extreme outliers and vice versa.
    - In our case, the high kurtosis number means that there are certain counties with exceptionally high cases and exceptionally low cases across the timeline, both accounting for the sharpness of peaks and heavy tails in the distribution. It is important to further investigate these extreme values to understand their causes to arrive at implications for the analyses.

Data Quality

With the help of ‘diagnose' function from the 'dlookr' package in R, we assessed the Texas dataset and attained the following quality metrics. The' missing count' column shows the count of variables that had missing values. The ‘Type’ column describes the type of data present in the column. The' missing percent' column denotes the proportion of missing variables. The' unique count' column indicates the number of distinct values within a specific variable.

Table 6: Missing data information for Covid Cases in TX

	Variable	Types	Missing Count	Missing %	Unique Count
1	county_name	Factor	0	0	255
2	confirmed_cases	Numeric	0	0	8,123

3	deaths	Numeric	0	0	1,383
4	date	Date	0	0	370

Overall, the dataset had no missing values. The uniqueness of the values for every variable account for the differences in the covid growth rate based on other key factors prominent in those respective regions. On top of that, even the ’unallocated data ‘county category (the county category for unallocated details recorded in the year) was categorized separately making it easier for us to include or neglect that group if necessary, during a visual or logical interpretation of the data.

Outliers:

This operation is performed with the help of a function named ‘diagnose_outlier’ from ‘dlookr’ package which gives various detailed insight into the outliers present in the dataset.

There is the total number of outliers calculated along with its ratio; we have the parameter ‘outliers_mean’ which says whether the calculation of outliers included the mean or not during the analysis. Hence, we have the two options named, With Mean and Without Mean.

The former gives us a balanced and optimized insight into dispersion of outliers form the mean based on factors such as standard deviation and mean of the variable, while the latter gives us the depiction of raw outliers solely based on deviation from the other data points.

In our context here, since we have a lot of outliers, going with the ‘With Mean’ parameter makes the most sense as it helps us understand the data in a more insightful way and gives us more information on how the data is spread across; the reason being it has been processed by computation against balancing metrics while dealing with the outliers at the same time.

We will visualize the following dataset to understand how the actual raw outliers are spread and how they are spread across after removing or resolving the outliers in the further sections.

Table 7: Outlier information for Covid Cases in TX

	Variable	Outliers Count	Outliers Ratio	Outliers_Mean	With Mean	Without Mean
1	confirmed_cases	13,397	14.19926	13,872.9165	2,158.57046	219.950601
2	deaths	13,732	14.55432	232.8165	38.18992	5.038366

Outlier Visualization:

Here is an illustration of how the confirmed cases and deaths for each county are distributed. This will further help us understand why the overall outliers are as shown in the subsequent graphs. This also proves the statistical skewness and kurtosis of the dataset as we can see that most counties have seen a steady increase while quite a lot of them had an exponential burst of deaths due to covid across the year. The outliers shown with lower death rates represent the fact that the counties have gone from being without covid or deaths due to covid to outlier on the opposite end of the spectrum accounting for the mass increase in covid cases/deaths as the days pass by in the year.

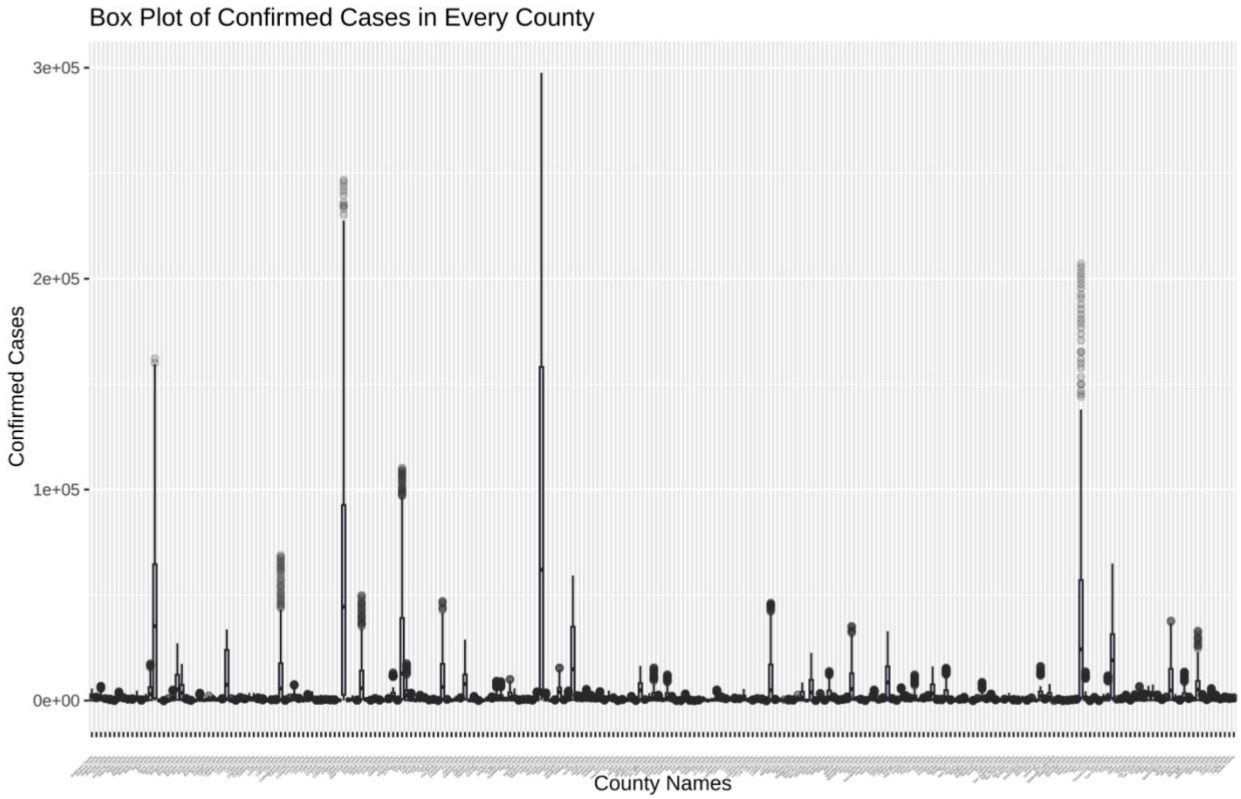


Figure 6 Confirmed cases by county.

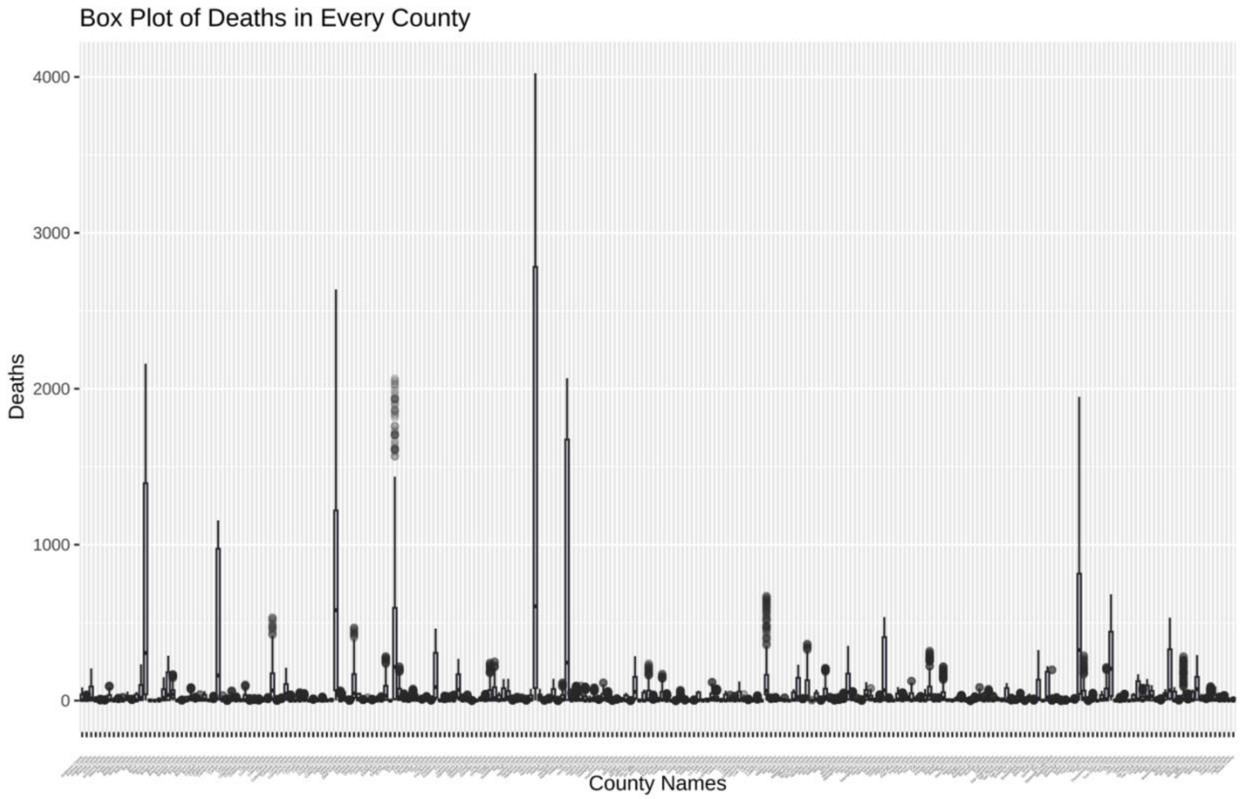


Figure 7 Confirmed deaths by county.

The ‘dlookr’ package also helps us to visualize the same data by removing or modifying the outliers. After having a look at the above figure which shows the outlier spread pattern across all counties, we can clearly understand the common trend and why the data with outliers are focused on two extremes. This further strengthens the claim that covid has seen a drastic increase from a baseline of 0 in a single year. Though similar in characteristic increase of the cases or deaths, we can see that the rate and pattern of increase is different for different counties, this is due to the influence of other complex factors pertaining to that region which ultimately decides the spread rate or death rate.

We can clearly distinguish the data set with and without outliers in the plot below.

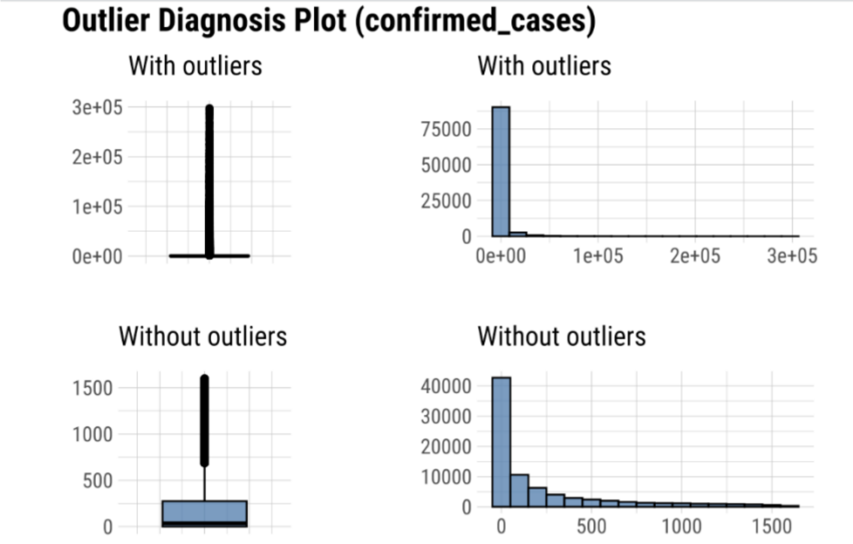


Figure 8 Univariate analysis - Outlier Diagnosis of Confirmed Cases.

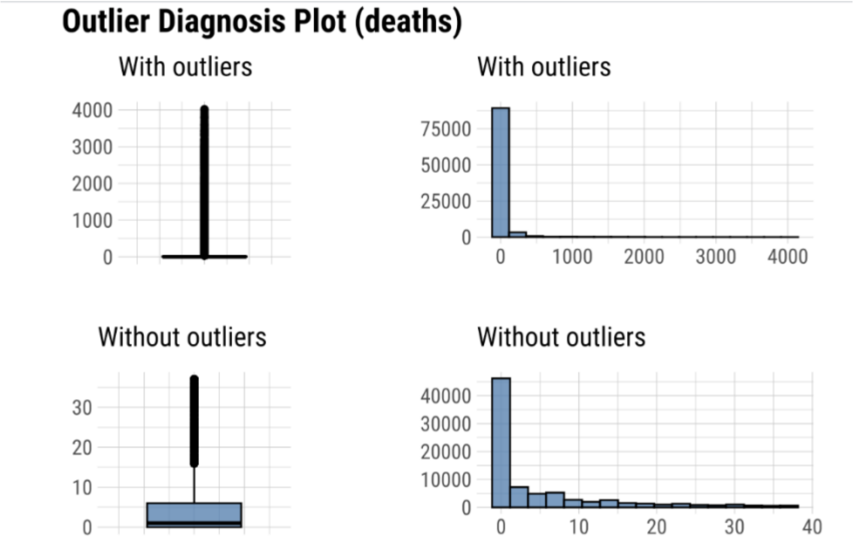


Figure 9 Univariate analysis - Outlier Diagnosis of Covid Deaths.

Handling Outlier:

The diagnose_outlier() function in the dlookr package utilizes the Tukey's fences method to detect and remove outliers from a dataset. Tukey's fences method identifies outliers based on the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3) of the data distribution.

Here is how the method works:

- Calculate the interquartile range (IQR) of the data.
- Define the lower fence (LF) as Q1 - k * IQR and the upper fence (UF) as Q3 + k * IQR, where k is a user-specified multiplier (often set to 1.5).
- Identify any data points that fall below the lower fence or above the upper fence as potential outliers.
- Optionally, remove or flag the identified outliers based on the analysis requirements.

Data Visualization

The distribution of total cases in every county and total deaths across the year has been shown below.

- The top 10 counties with the highest total deaths across the entire year are mentioned here with names.
- The total number of cases identified for these counties are represented by the size of the dots plotted for that county.

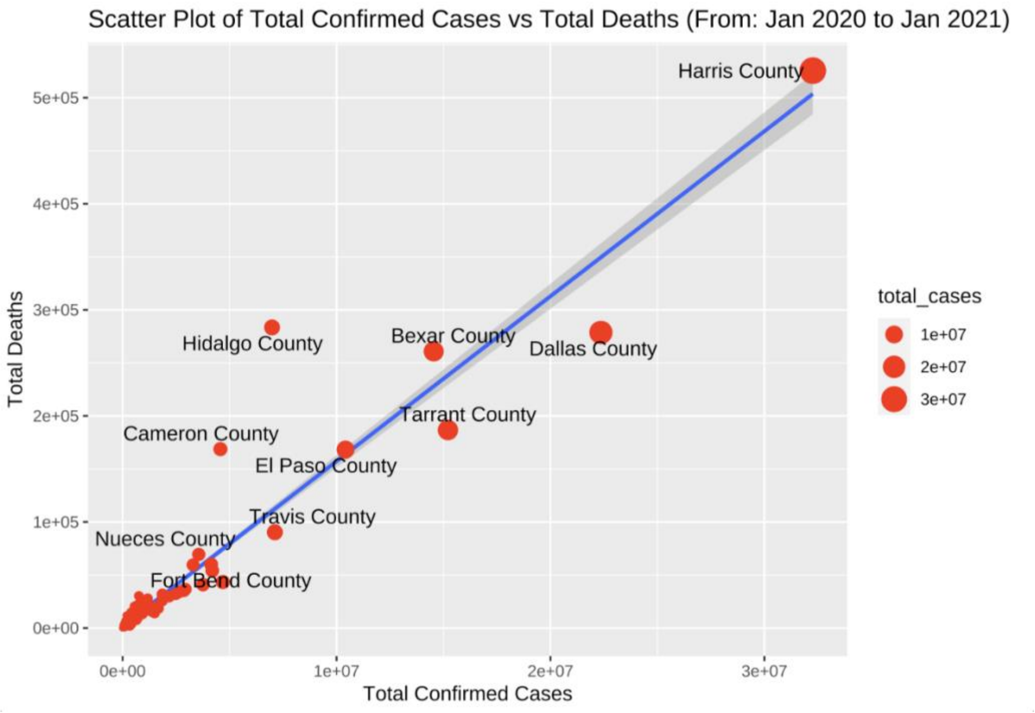


Figure 10 Total confirmed cases versus total deaths.

The above representation gives us counties with top 10 highest death count along with the number of cases identified. It is important to note that this data is based on the raw inference from the data set rather than a normalized one.

In this case, we can see Harris County leading in terms of the total number of cases and in total number of deaths. The difference between the first and the second county seems to be considerably huge. This could be due to various other factors. For instance, the top ten counties for the same metric might change if we take in the population and normalize the variables accordingly.

We are only able to deduce top counties affected with Covid and their death rates here with the data available. We will further see how to create new features by associating different columns to answer key questions in our 'data preparation' section.

Covid in relation with the date/time metric:

Now, Let is we can also look at the way in which the cases and death counts varies across the year to understand the impact of covid during different months.

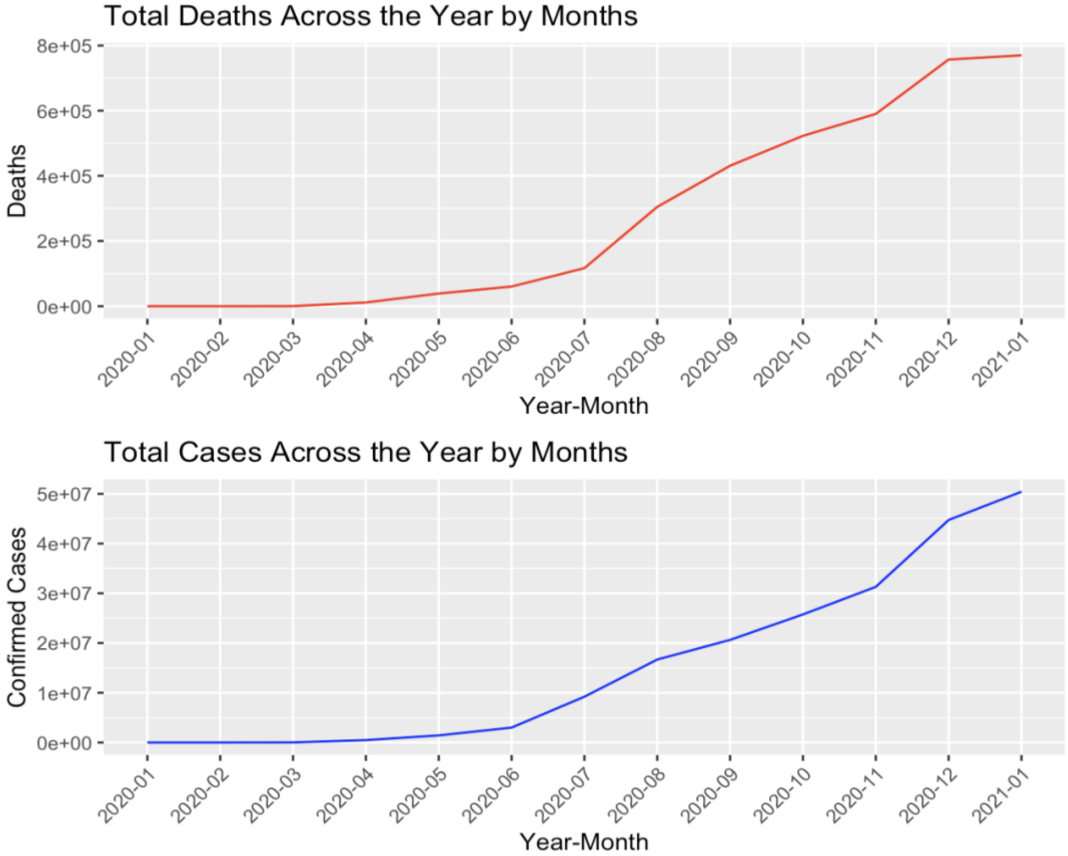


Figure 11 Monthly analysis of cases and deaths.

Though the scales are different, we can see similar increasing characteristics being displayed by both the number of 'cases confirmed' across the year and the 'deaths' accounted for from the above data.

To conclude with "COVID-19_cases_TX" dataset, these findings are some of the most prominent feature understandings that we could derive from this stand-alone dataset without associating it with other datasets available. The key features in this dataset that proved insightful were ‘confirmed cases across counties, death rates and the recorded time frame for the above two variables ‘. The former addresses ‘what is happening in the counties,’ while the latter helped us pinpoint ‘when it occurred’ in the given time frame of 1year and 3 days, giving us the top view of the overall increase of cases in Texas Counties.

Note: We might be able identify new or changed patterns and trends which could be different from these observations as we would be having a lot of other underlaying factors that would give more information on the data. This is because we are going to be working with our combined features to identify the trends across Texas rather than working with just 4 out of the 7 features that were available in the current dataset.

Data Correlation

It is extremely clear from the previous visualizations and tables that the ‘confirmed_cases’ and ‘deaths’ are highly correlated to one another (+1.0), meaning that deaths always increase when there is an increase in confirmed_cases. If you notice further, you can also find that the confirmed cases lead the deaths by a considerable margin. Thus, giving us more logical proof into the difference between the scales at which both display same characteristics while maintaining their correlation properties.

Here is a graph plotted with the help of correlation function from ‘R’ to depict the same using the data available across both the columns.

Additionally, we have included the date column here to show an interesting relationship between the date (i.e.) the time frame and the confirmed_cases and deaths recorded across the counties. Though not a numerical value, the increase in date (as 2021 comes nearby) is proportional to the increase in number of cases and hence the number of deaths. This proves that the date or the time frame has a valid positive correlation (+0.5) with them as shown here while also solidifying the claim that covid has been increasing across the year as mentioned earlier in the deaths/cases by month plot.

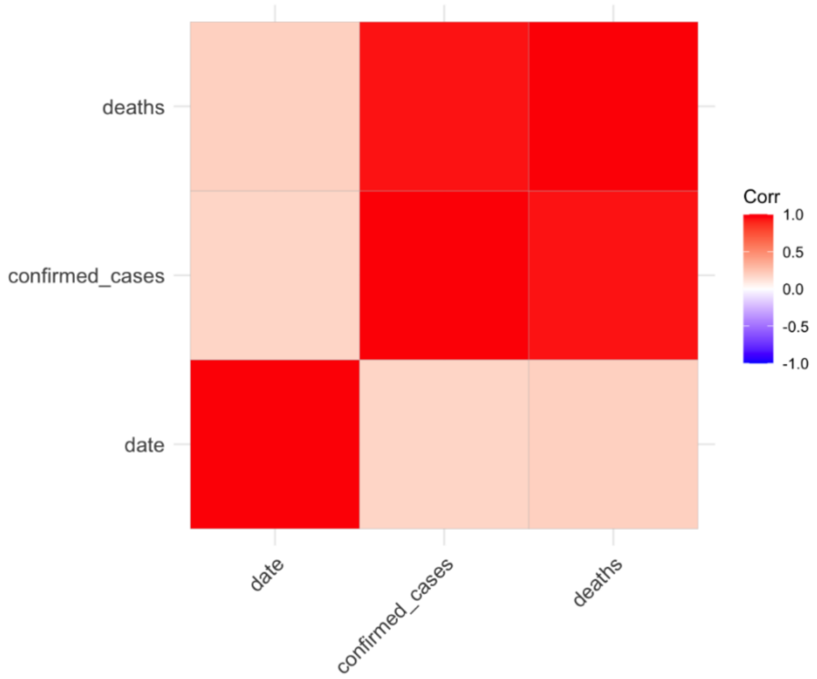


Figure 12 Correlation of cases and deaths in Texas.

Global Mobility Report Data Set

The third data set we will be observing is the global mobility Report data set which included almost 4 million observations but only 14 features. Because of the vast number of observations, we wanted to only focus on counties in the state of Texas. This reduced the data set to 64,162 observations. This data set included data from February 15, 2020, to January 22, 2021. It provides insights into how mobility patterns have changed worldwide in response to the COVID-19 pandemic. This data set allows researchers, policymakers, and public health officials to analyze trends in movement and understand the impact of social distancing measures, lockdowns, and other interventions aimed at controlling the spread of the virus. Key metrics often included in the data set may involve changes in mobility to several types of locations (e.g., retail and recreation, grocery stores, workplaces, parks, and residential areas) relative to a baseline period before the pandemic. By examining these mobility patterns, our stakeholders can assess economic recovery efforts and resource allocation. Since this dataset has 14 features, we wanted to narrow this number to only include important ones that are useful to our stakeholders. We filtered the data to only include counties in the state of TX and information that is useful for our stakeholders. Table 8 shows the features we want to include from this dataset along with the scale of measurement and information of each.

Table 8: Description of Features of Interest their respective data types for Mobility Report

Feature	Scale of Measurement	Information
Sub Region 2	Nominal	Second geographic sub-region in which the data is aggregated. This varies by country/region to ensure privacy and public health value in consultation with local public health authorities
Date	Interval	Changes for a given date as compared to baseline. Baseline is the median value for the corresponding day of the week during the 5-week period Jan 3–Feb 6, 2020.
Retail and Recreation percent change from baseline	Ratio	Mobility trends for places like restaurants cafes shopping center's theme parks museums libraries and movie theaters.
Workplaces percent change from baseline	Ratio	Mobility trends for places of work.

Data Quality

To verify the quality of the data, we checked for missing values in our dataset, and we observed the following amount:

- Sub region 2 # of missing values: 343
- Date # of missing values: 0.
- Retail and Recreation percent change from baseline # of missing values: 23368
- Workplaces percent change from baseline # of missing values: 4789

To fix this problem, we simply removed these missing values from the data. There were no duplicates in our data set. This further reduced our dataset to 39,955 observations. Next, we noticed that there were a lot of outliers in our data set for Retail and Recreation percent change from baseline and Workplaces percent change from baseline (see Figure 13 and 14). To show we fixed this problem, please check our exceptional work section on how we used Winsorization to remove these outliers. This lowered the number of observations for this data set to 39,612.

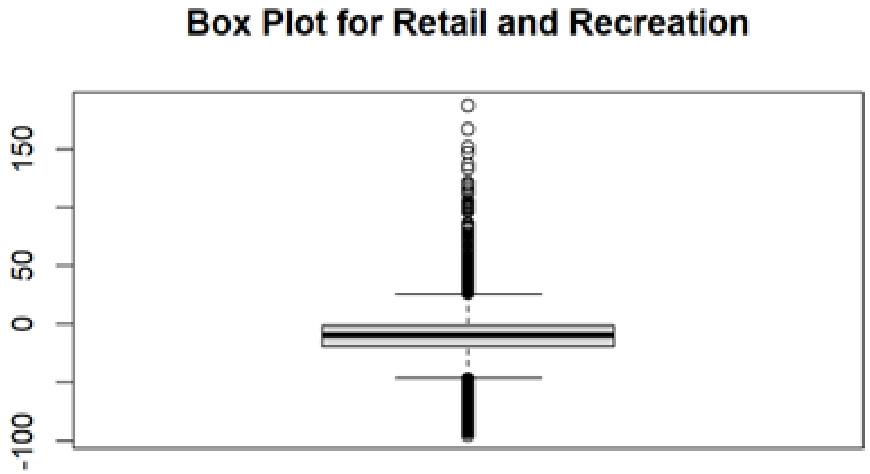


Figure 13 Outliers for Retail and Recreation.

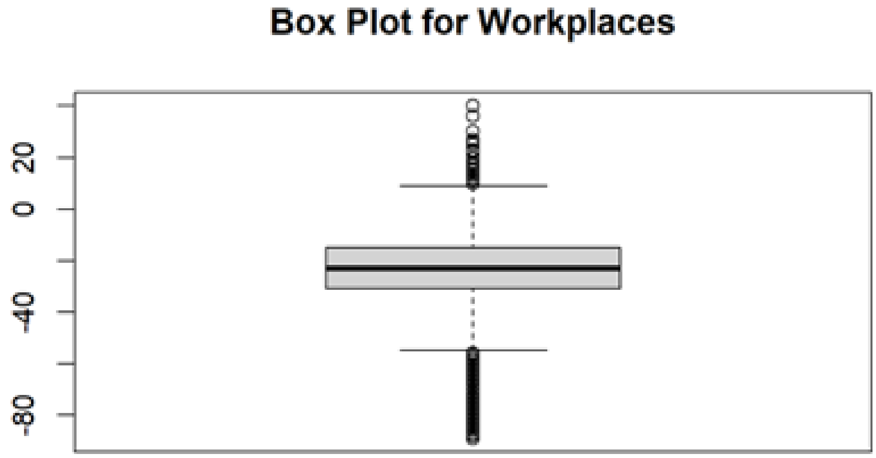


Figure 14 Outliers for workplaces.

Summary Statistics

The following are the statistics we have chosen for this data set includes the minimum, Q1, Median, Mean, Q3, and Max. Here is description of what each mean:

- **Minimum**: The smallest value observed in the data. For example, if the minimum value for the variable "retail_and_recreation_percent_change_from_baseline" is -39, it means that the lowest observed percentage change in retail and recreation mobility was a decrease of 39%.
- **1st Quartile (Q1)**: This value represents the 25th percentile of the data. It separates the lowest 25% of the observations from the rest. For example, if the 1st quartile for "workplaces_percent_change_from_baseline" is -20, it means that 25% of the observations had a percentage change in workplace mobility less than or equal to -20%.
- **Median (Q2)**: This value represents the middle value of the dataset when it is ordered from smallest to largest. It divides the data into two equal halves. For example, if the median for "retail_and_recreation_percent_change_from_baseline" is -10, it means that half of the observations had a percentage change in retail and recreation mobility less than or equal to -10%.
- **Mean**: The arithmetic average of all the values in the dataset. For example, if the mean for "workplaces_percent_change_from_baseline" is -15, it means that on average, the percentage change in workplace mobility was -15%.
- **3rd Quartile (Q3)**: This value represents the 75th percentile of the data. It separates the lowest 75% of the observations from the highest 25%. For example, if the 3rd quartile for "retail_and_recreation_percent_change_from_baseline" is 5, it means that 75% of the observations had a percentage change in retail and recreation mobility less than or equal to 5%.
- **Maximum**: The largest value observed in the data. For example, if the maximum value for "workplaces_percent_change_from_baseline" is 20, it means that the highest observed percentage change in workplace mobility was an increase of 20%.

Table 9 shows these stats. The interquartile range (IQR) for retail and recreation is 18 (from -19 to -1), indicating a widespread in mobility changes within the middle 50% of observations.

The IQR for workplaces is 18 (from -32 to -14), suggesting a similar spread in mobility changes within this category. Retail and recreation locations generally experienced smaller decreases in mobility compared to workplaces, as indicated by the quartile values and median. The median values for both categories suggest that, on average, mobility decreased during the specified period. However, the extent of the decrease varied between retail and recreation locations and workplaces.

Table 9: Descriptive statistics for Mobility Report after Outliers Removed

Feature	Min	Q1	Median	Mean	Q3	Max
Retail and Recreation percent change from baseline	-39	-19	-9	-10.44499	-1	14

Workplaces percent change from baseline	-47	-32	-23	-22.88993	-14	2

This summary statistics of mobility changes in Texas offer valuable insights into the economic impact of the COVID-19 pandemic and its potential relationship with the distribution of stimulus checks. These statistics, highlighting significant decreases in mobility, suggest economic challenges such as decreased consumer spending and employment disruptions. Such trends emphasize the potential need for financial assistance, like stimulus checks, to support affected individuals and families facing job losses or income reduction. By analyzing mobility data alongside stimulus check distribution, our stakeholders can better target assistance to areas and populations most affected by economic downturns, ensuring that financial relief reaches those in need.

Data Correlation

After exploring the attributes of this data set, we noticed that the correlation between Retail/Recreation and Workplaces Mobility gives a correlation of 0.54 as Shown in Figure 15 below.

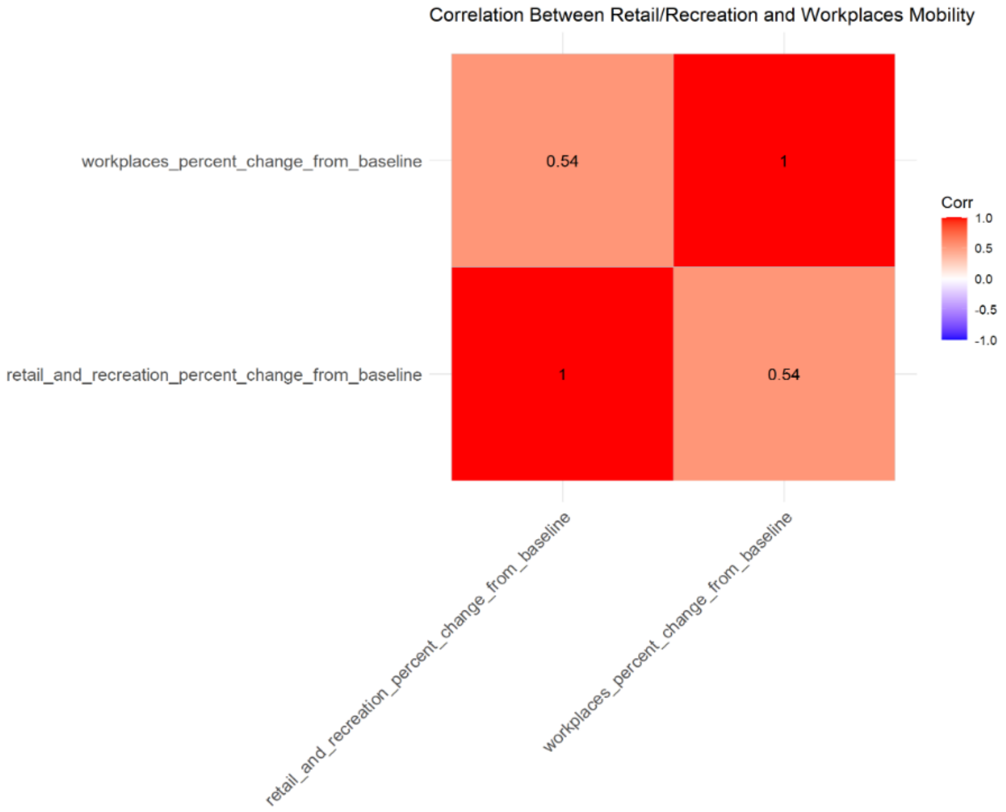


Figure 15 Correlation between Retail/Recreation and Workplaces Mobility.

The correlation coefficient of 0.54 indicates a moderate level of association between the variables. It implies that when retail_and_recreation_percent_change_from_baseline increases (or decreases), workplaces_percent_change_from_baseline also tends to increase (or decrease). In other words, there is a tendency for both variables to move in the same direction. This makes sense given that during the COVID-19 epidemic less people were going out shopping and working making both variables positively correlated. However, the reason for a moderate level of association was because the only people for the most part who were able to work were the essential workers and you also had more people work from home. This could be the reason there wasn't a higher level of association between both variables.

Data Visualization

To visualize the data better, a line plot would be suitable to show the trend in changes in mobility over time for retail and recreation locations and workplaces. Each point on the plot represents the percentage change from baseline for a specific date. This visualization allows us to observe any trends in mobility changes over time. A line plot is chosen because it effectively displays the trend in mobility changes over time, which is particularly relevant for retail and recreation locations and workplaces where there may be seasonal or temporal patterns in mobility (See Figure 16 and 17).

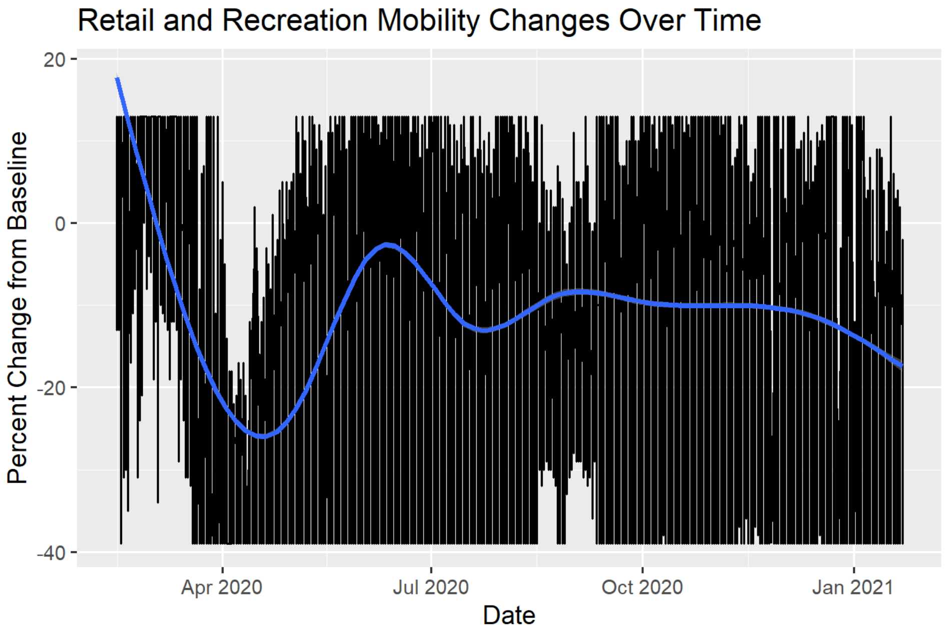


Figure 16 Retail and Mobility Changes over time.

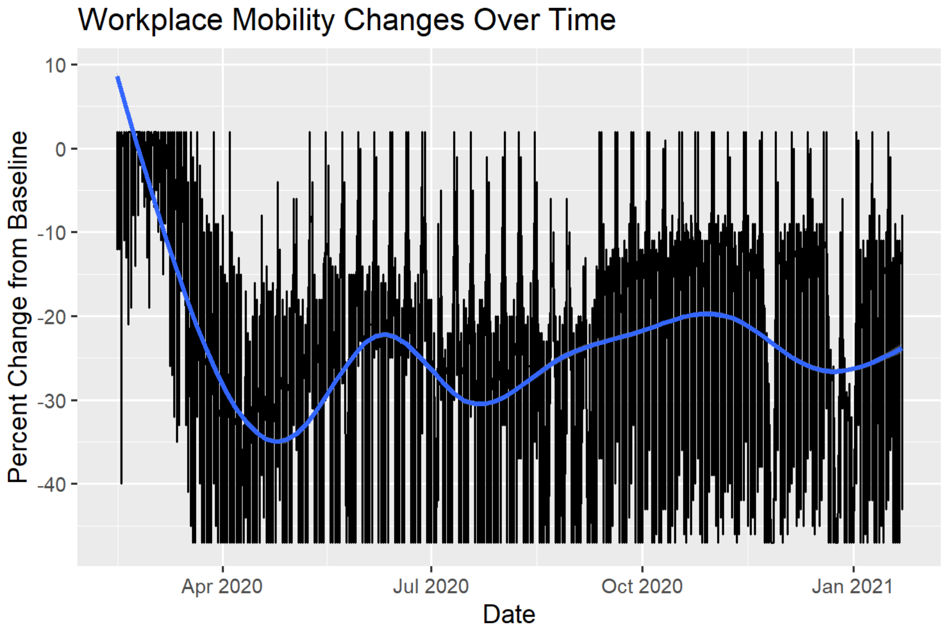


Figure 17 Workplace mobility changes over time.

When considering the timing, amount, and necessity of distributing stimulus checks, the interpretation of mobility data, as demonstrated in line plots depicting changes in retail and recreation mobility and workplaces mobility, is crucial. These visualizations offer insights into consumer spending patterns and employment dynamics, enabling our stakeholders to gauge the economic impact of potential stimulus measures. Decreases in retail and recreation mobility may signal reduced consumer confidence and spending, while declines in workplaces mobility indicate employment challenges and income instability. We see this happening in the beginning of the pandemic from Feb 2020 to Apr 2020 and then again in Aug 2020 and Dec 2020 for both graphs. Using mobility trends as indicators, our stakeholders can determine the optimal timing for stimulus check distribution. Furthermore, the magnitude of stimulus checks can be tailored based on the severity of economic conditions reflected in mobility data. By aligning stimulus check distribution with fluctuations in economic activity and consumer behavior, policymakers can implement targeted measures to support individuals, businesses, and the broader economy during challenging times.

According to as.com, the first round of stimulus was sent in April of 2020. People started seeing their money in their bank accounts on April 11-12 [7]. Below in Figure 18, we can see the day-to-day Retail/Recreation Mobility Changes in the state of Texas for the month of April 2020.

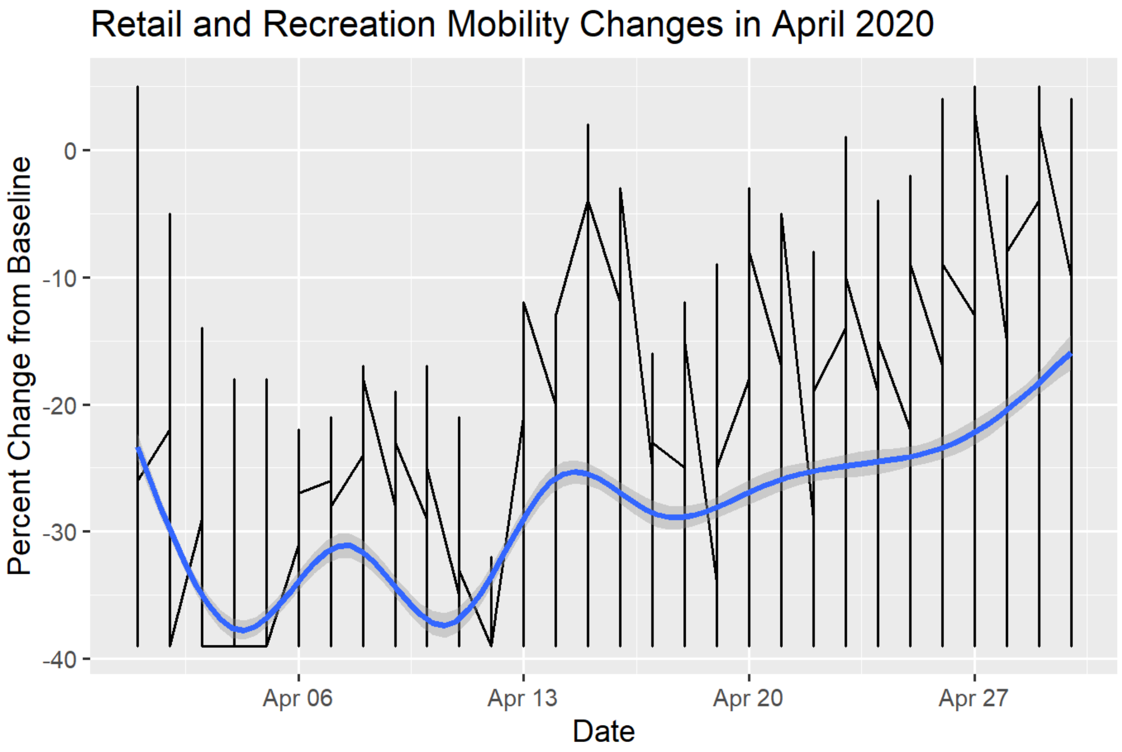


Figure 18 Retail and Recreation mobility changes in April 2020.

As we can see in this Figure, on April 11-12, we can see an increase in Texans going out to retail stores and doing more recreation activities. This accurately proves the effects of the first round of stimulus checks on Texans. However, in Figure 19, we can see the day-to-day Workplaces Mobility Changes in the state of Texas for the month of April 2020 not have the same effect.

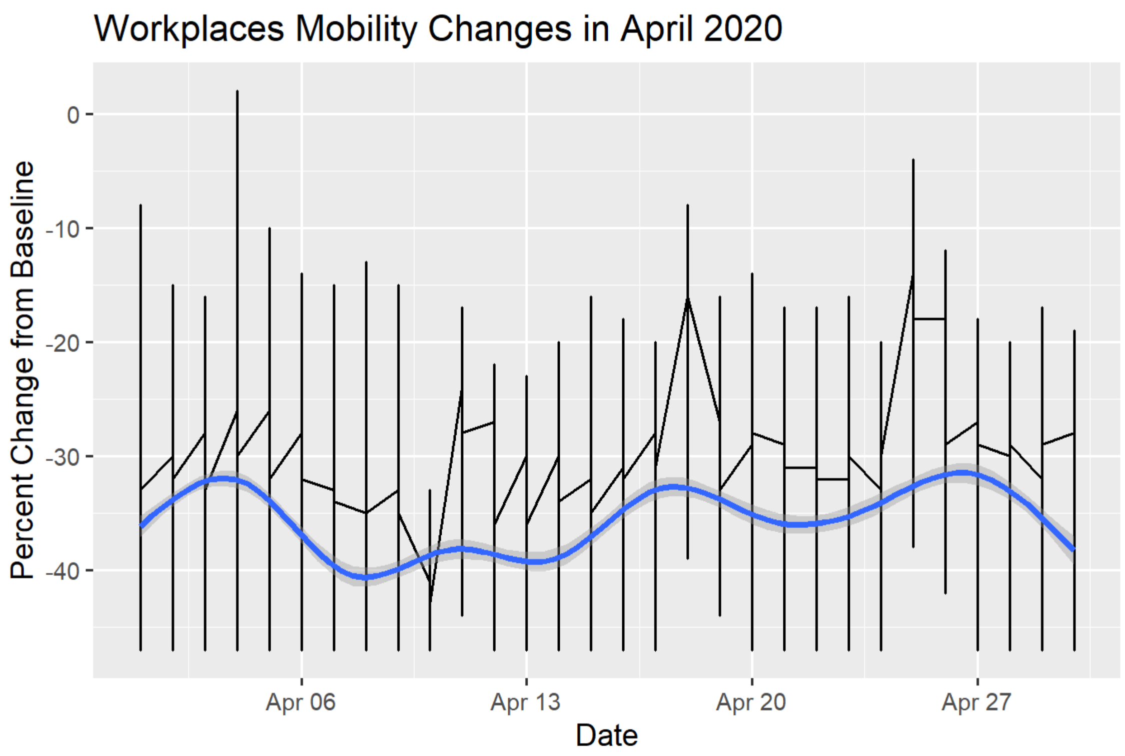


Figure 19Workplaces mobility changes in April 2020.

On April 11-12, we don't see a spike here like we did for Retail/Recreation mobility. For the most part the mobility for workplaces wasn't really affected by the stimulus checks. This makes sense given that essential workers still had to work, and people didn't necessarily live off the stimulus checks.

According to as.com, the second round of stimulus was sent in December 2020-January 2021. People started seeing their money in their bank accounts on December 29, 2020-January 15, 2021 [7]. Below in Figure 20, we can see the day-to-day Retail/Recreation Mobility Changes in the state of Texas for the months of December 2020-January 2021.

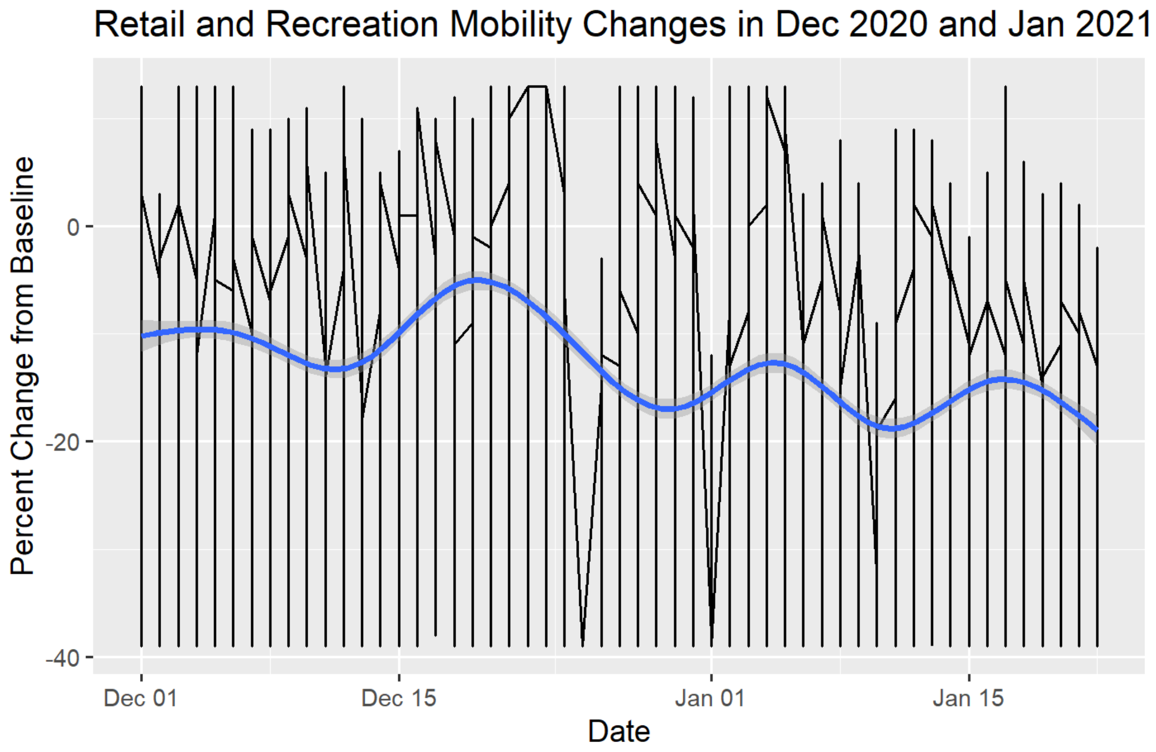


Figure 20 Retail and Recreation Mobility Changes in December 2020 and January 2021.

As we can see in this Figure, between December 29, 2020, and January 15, 2021, we can see an increase and decrease in Texans going out to retail stores and doing more recreation activities. The second stimulus check wasn't as nearly impactful as the first stimulus check, most likely since the check was less than the first one ($1200 vs $600). However, in Figure 21, we can see the day-to-day Workplaces Mobility Changes in the state of Texas for the months of December 2020-January 2021 not have the same effect.

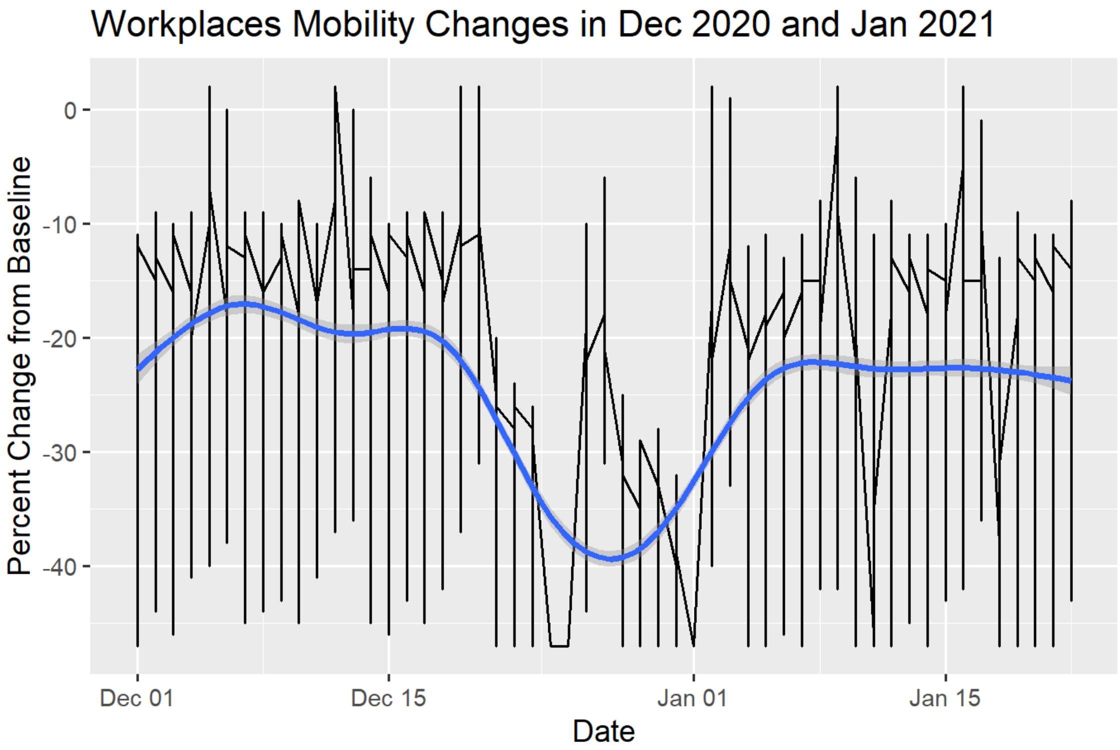


Figure 21 Workplace mobility changes in December 2020 and January 2021.

Between December 29, 2020, and January 15, 2021, we see a spike. One could argue that the second round of stimulus did have a positive impact on workplaces but given the time of year, this spike could have been a result of essential workers going back to work after the holidays. Nonetheless, these essential workers still didn't necessarily live off these stimulus checks.

Data Preparation

Final Combined Data Set

After doing individual analysis for each of the three datasets, we were able to combine the features of each dataset into a final dataset with combined features as seen below. We decided to make modifications of features from the Covid-19 Census plus cases datasets for normalization, as well as features from both the Covid-19 cases TX dataset and the global mobility report data set (for just one day on January 19, 2021) that best suits our stakeholders specifically interested in the state of Texas. We decided to add an extra feature that listed whether a Texas county is a rural or urban county based on the information from the Texas Department of Housing and Community Affairs [8]. This will give a better feel of the socioeconomic factors that can play into deciding the distribution of stimulus checks. Furthermore, we wanted to fill in the missing values we had for our dataset through mean imputation. Having the counties separated into rural and urban helped make the fill-in values more accurate to other counties of the same type. Table 10 will list the features of the final dataset and Table 11 will list the first 10 rows of this dataset with the Texas counties listed as the rows and the features as columns.

Table 10: Description of Features of Interest their respective data types for Final Data Set

Feature	Scale	Description
1. county_name	Nominal	Name of the county
2. retail_and_recreation_percent_change_from_baseline	Ratio	Mobility trends for places like restaurants cafes shopping center's theme parks museums libraries and movie theaters.
3. workplaces_percent_change_from_baseline	Ratio	Mobility trends for places of work.
4. confirmed_cases	Ratio	Number of confirmed cases of Covid-19
5. Deaths	Ratio	Number of deaths as a result of Covid-19
6. percent_income_spent_on_rent	Ratio	Percent of household income spent on rent
7. **total_pop**	Ratio	The total number of all people living in a given geographic area
8. median_income	Ordinal	Median Household Income in the past 12 Months.
9. median_rent	Ordinal	The median contract rent within a geographic area.
10. median_age	Ratio	The median age of all people living in a given geographic area.
11. confirmed_cases_per_1000	Ratio	Number of confirmed cases of Covid-19 by 1000 people
12. deaths_per_1000	Ratio	Number of deaths as a result

		of Covid-19 by 1000 people
13. hh_asst_or_food_stamps_per_1000	Ratio	Households on cash public assistance or receiving food stamps (SNAP) by 1000 people
14. poverty_status_per_1000	Ratio	The number of people in each geography who could be identified as either living in poverty or not by 1000 people
15. poverty_per_1000	Ratio	The number of people in a geographic area who are part of a family determined to be "in poverty" by 1000 people
16. commuters_per_1000	Ratio	The number of workers aged 16 years and over within a geographic area who primarily traveled to work by public transportation by 1000 people
17. walked_to_work_per_1000	Ratio	The number of workers aged 16 years and over within a geographic area who primarily walked to work by 1000 people
18. worked_at_home_per_1000	Ratio	The count within a geographical area of workers over the age of 16 who worked

		at home by 1000 people.
19. rent_over_50_per_1000	Ratio	Housing units spending over 50% income on rent by 1000 people.
20. rent_40_to_50_per_1000	Ratio	Housing units spending 40% - 49.9% income on rent by 1000 people
21. rent_35_to_40_per_1000	Ratio	Housing units spending over 35% - 39.9% income on rent by 1000 people
22. rent_30_to_35_per_1000	Ratio	Housing units spending over 30% - 34.9% income on rent
23. Rural/Urban	Nominal	Determines if a county is rural or urban

Table 11: First 10 Rows of Final Data Set

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Bastrop County	-15	-30	4373	57	26.5	80306	59185	724	38.8	54.45421	0.709785	45.7002	969.2676	128.5334	0.958832	5.603566	15.90168	12.48973	4.806615	5.703185	4.706996	Urban
Bell County	-23	-24	16512	214	27.7	336506	52583	722	30.6	49.06896	0.635947	48.34089	970.1878	138.6929	1.973219	9.515432	15.76198	26.80784	12.5971	11.02804	13.9106	Urban
Bexar County	-25	-39	1522231	2040	29.7	1892004	53999	786	33.3	80.46019	1.078222	46.8535	981.9689	161.0948	12.00632	7.83825	19.20345	30.08027	12.48465	9.070805	11.46192	Urban

Bowie County	-14	-20	4876	139	30.7	93635	46283	579	38.2	52.07454	1.484488	53.03572	940.7807	168.2063	1.334971	3.417525	9.750627	28.03439	10.91472	7.32632	10.34869	Urban
Brazoria County	-11	-28	26046	277	26	345995	76426	825	35.6	75.27854	0.80059	30.71432	965.7307	95.50138	0.710993	3.904681	13.13603	16.90198	6.03766	5.211058	7.098368	Urban
Brazos County	-20	-33	16634	162	36.8	214231	43907	735	25.8	77.64516	0.756193	33.37986	928.7358	241.347	11.49694	11.72099	15.94074	69.57443	16.64558	10.02189	11.98239	Urban
Caldwell County	-17	-30	2738	55	29.5	40544	51346	623	35.8	67.53157	1.356551	49.67443	948.5744	168.1876	0.86326	15.90864	17.80781	21.38418	9.791831	6.21547	9.249211	Urban
Cameron County	-28	-35	32698	1126	32.4	420201	36095	529	31.4	77.81514	2.67967	76.71567	989.5003	309.1663	1.987144	4.840541	8.695838	23.31503	8.905262	4.983329	7.936678	Urban
Chambers County	-17	-24	3431	13	23	39283	74368	712	35.4	87.34058	0.330932	22.40155	993.2286	128.0707	0.763689	3.818446	8.833338	9.800677	3.691164	1.731029	3.207494	Urban
Collin County	-26	-44	64721	483	26.6	914075	90124	1079	36.5	70.80491	0.528403	13.65533	995.3319	68.3084	5.792741	4.152832	43.68241	22.0704	9.354812	7.683177	10.34488	Urban

Visualizations of Combined Features

What is the trend in different areas of Texas?

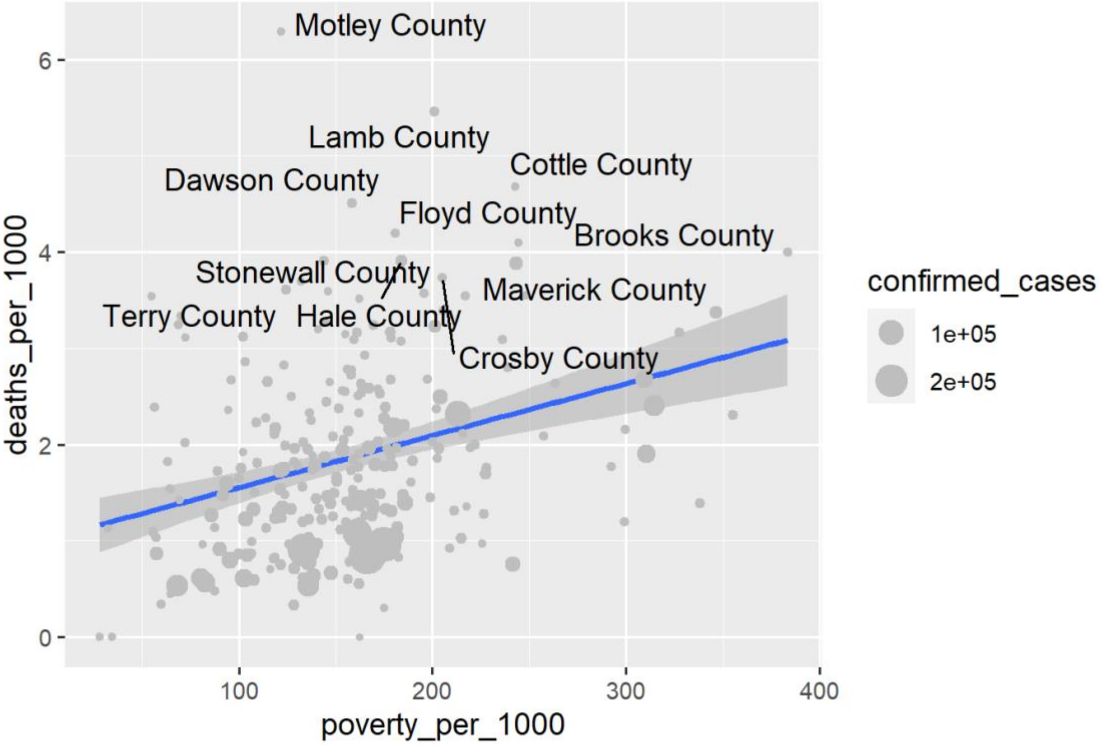


Figure 22 Poverty versus Covid induced deaths.

Figure 22 highlights a positive correlation between poverty rate and death rate due to COVID-19, suggesting that counties with higher poverty rates may face greater challenges in managing the pandemic. Stimulus checks can play a crucial role in mitigating the socioeconomic impact of the pandemic. By targeting poor communities with higher poverty rates, stimulus checks can help improve financial hardships, reduce economic instability, and potentially improve health outcomes.

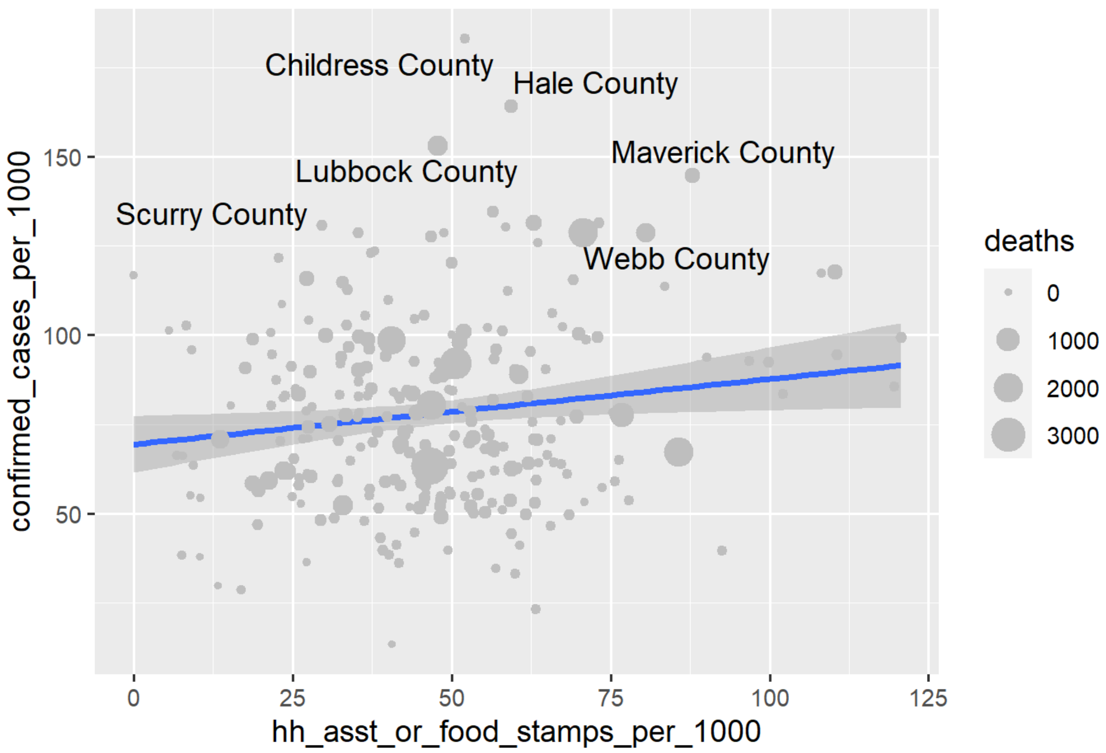


Figure 23 Count of households on public assistance versus confirmed cases per 1000.

Furthermore, the scatter plot in Figure 23 illustrates a potential association between households receiving assistance or food stamps per 1000 population and confirmed COVID-19 cases per 1000 population. The positive slope of the linear regression trend line suggests that counties with higher rates of households receiving assistance or food stamps tend to experience higher rates of confirmed COVID-19 cases. This observation underlines the vulnerability of poor communities to the impacts of the pandemic, possibly due to factors such as limited access to healthcare, crowded living conditions, and the inability to practice social distancing. Targeted stimulus check distribution can contribute to reducing disparities, promoting equity, and supporting the resilience of vulnerable populations in Texas and beyond.

Is social distancing done and is it working?

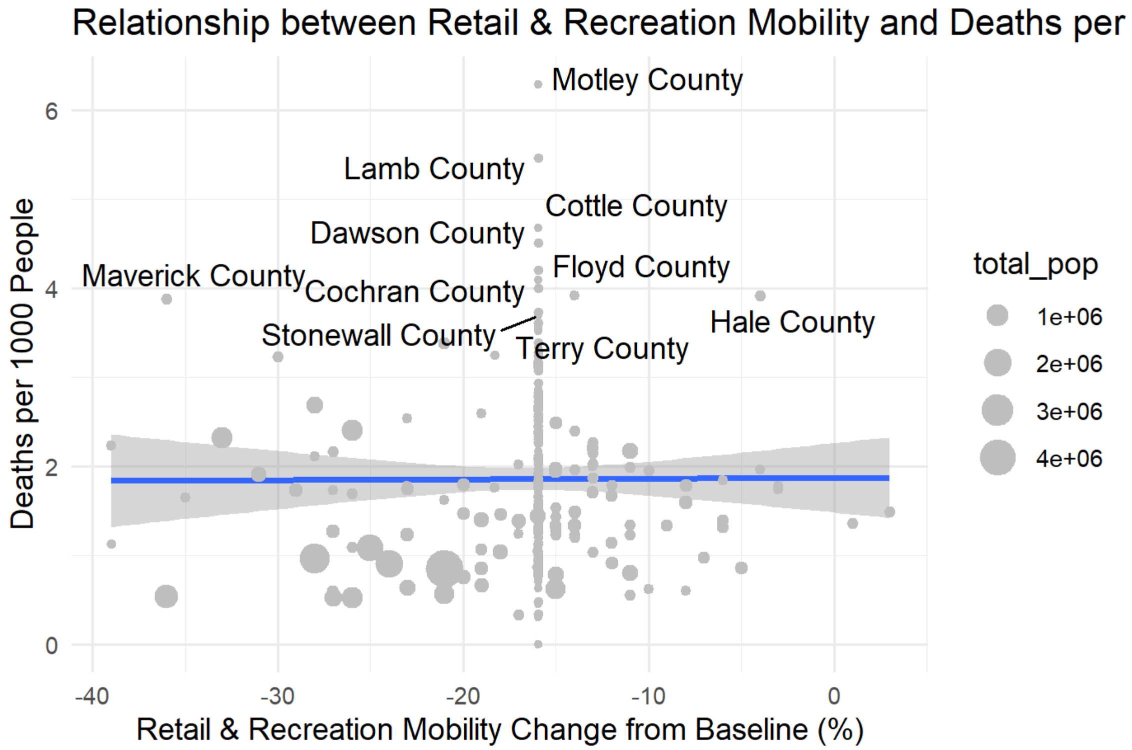


Figure 24 Relationship between retail & recreation and covid induced death.

The linear regression trend line suggests a slight negative relationship between changes in retail and recreation mobility and deaths per 1000 people, although the relationship appears to be relatively weak. Counties with higher mortality rates are labeled to highlight potentially concerning areas.

This analysis provides insights into the potential impact of social distancing measures on COVID-19 outcomes. The negative correlation observed suggests that areas with greater reductions in retail and recreation mobility may experience lower mortality rates. This implies that social distancing measures, as indicated by decreased mobility in public spaces, may be associated with reduced transmission and mortality rates. However, this doesn’t give a clear indication that social distancing is working, since factors, such as healthcare capacity, population density, and adherence to other preventive measures, may also influence COVID-19 outcomes.

Regions that do particularly well

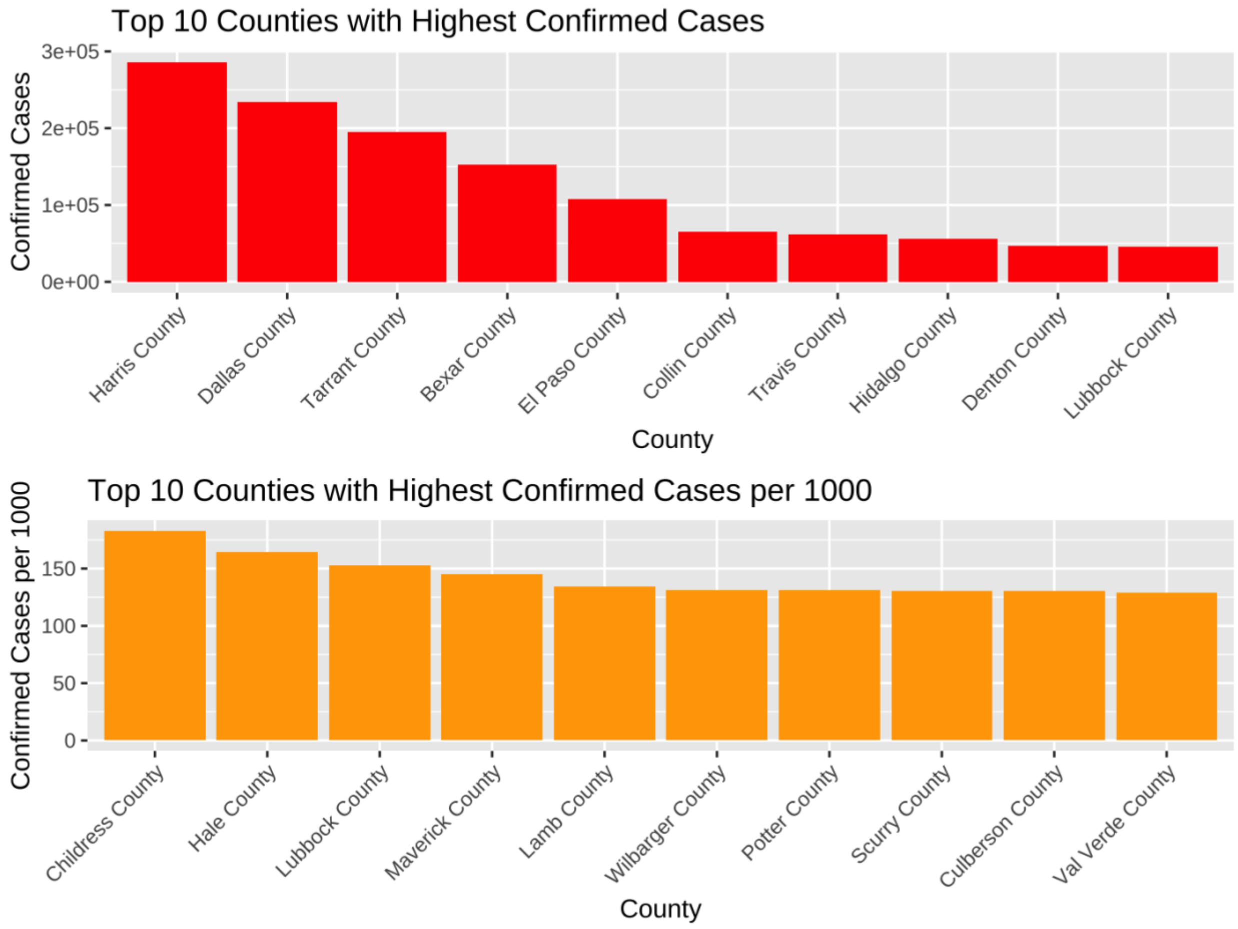


Figure 25 Counties with highest confirmed cases.

Table 12: Top 10 counties with highest cases per 1000

county_name	Rural/Urban	confirmed_cases	confirmed_cases_per_1000	total_pop
Childress County	Rural	1,292	182.90	7,064
Hale County	Rural	5,668	164.16	34,527
Lubbock County	Urban	45,600	153.00	298,042
Maverick County	Rural	8,320	144.77	57,471
Lamb County	Rural	1,799	134.58	13,368
Wilbarger County	Rural	1,707	131.59	12,972
Potter County	Urban	15,947	131.54	121,230
Scurry County	Rural	2,270	130.87	17,346
Culberson County	Rural	294	130.26	2,257
Val Verde County	Rural	6,331	129.27	48,976

Here, in the red graph we can see the top ten counties with highest Confirmed Cases across the state of Texas. The trend here changes when we normalize the visualization by the population per 1000 factor. Initially, we can see a lot of counties with a higher population density constitute the top ten survey. This means that a much larger county like Harris will be compared directly to a small county like Rockwall while not involving a scaling factor in-between. This way a respective highly affected county might not be accounted for in the top 10.

When we normalize the same using cases per 1000 population factor, the top 10 highly affected counties change completely in a rather interesting fashion where in none of the previous counties are now the most affected with respect to its relative population size. Therefore, showing us that these (represented in orange) are actually the highly affected counties on a proper scaling platform.

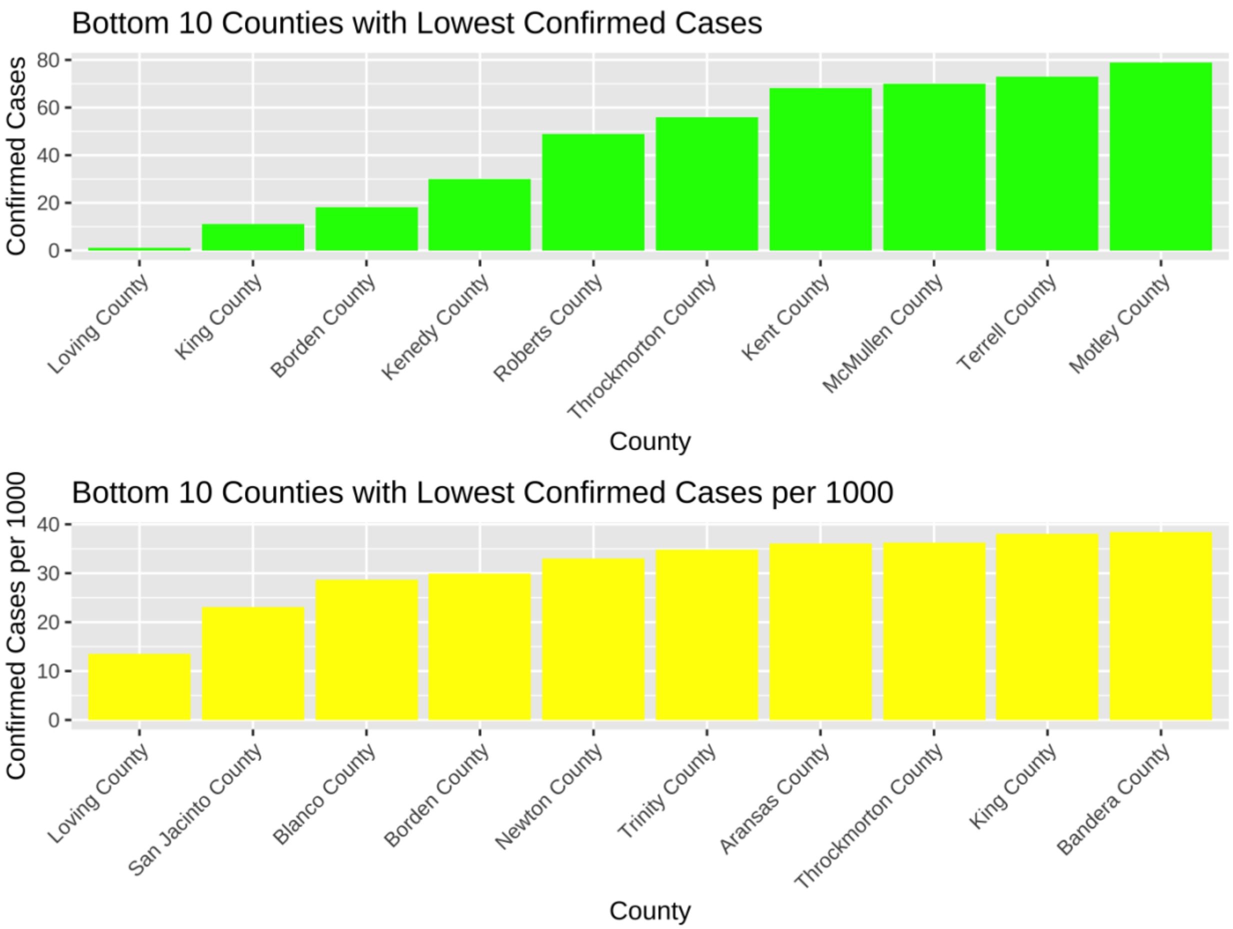


Figure 26 Counties with lowest confirmed cases.

The same trend carries forward to the bottom 10 counties with the least number of cases confirmed, where the correlated death rate is naturally down by a large margin in those countries. Key factors regarding what has caused the death rate to be low will be studied in further analyses.

The confirmed cases recorded in green, i.e., before normalizing by population shows us the counties with the least number of cases overall without taking their population into account whereas the counties in yellow shows us the counties with the least cases after factoring in their respective population per 1000.

We can see a significant trend here as the same counties namely - Loving, King, Borden, and Throckmorton County has appeared in both the graphs. This shows us two things, one proving some small counties did have a lesser number of covid cases and that population alone is not a deciding factor for the covid spread. The loving county is the least affected one with or without bringing in the population factor into consideration. While this proves that there can be special cases like this due to other factors like medical facilities available, overall hygiene followed or the various measures of covid control undertaken; on a high level, many of the regions exhibit correlation of cases with respect to the population density.

This is one such example of a scenario where the effect of other factors overshadows the effect of correlation. Also, studying what helped the county to overcome the correlation and maintain a lower cases across would prove to be beneficial in exploiting the weakness of COVID-19 spread. This solution factor could also be replicated in the other affected areas provided they satisfy the requirements for it.

Table 13: Top 10 counties with highest deaths per 1000

county_name	**Rural/Urban**	**deaths**	**deaths_per_1000**	**total_pop**
Motley County	Rural	7	6.2837	1,114
Lamb County	Rural	73	5.4608	13,368
Cottle County	Rural	7	4.6729	1,498
Dawson County	Rural	59	4.5055	13,095
Floyd County	Rural	25	4.1996	5,953
Cochran County	Rural	12	4.0928	2,932
Brooks County	Rural	29	3.9994	7,251
Hockley County	Rural	91	3.9101	23,273
Hale County	Rural	135	3.9100	34,527
Maverick County	Rural	223	3.8802	57,471

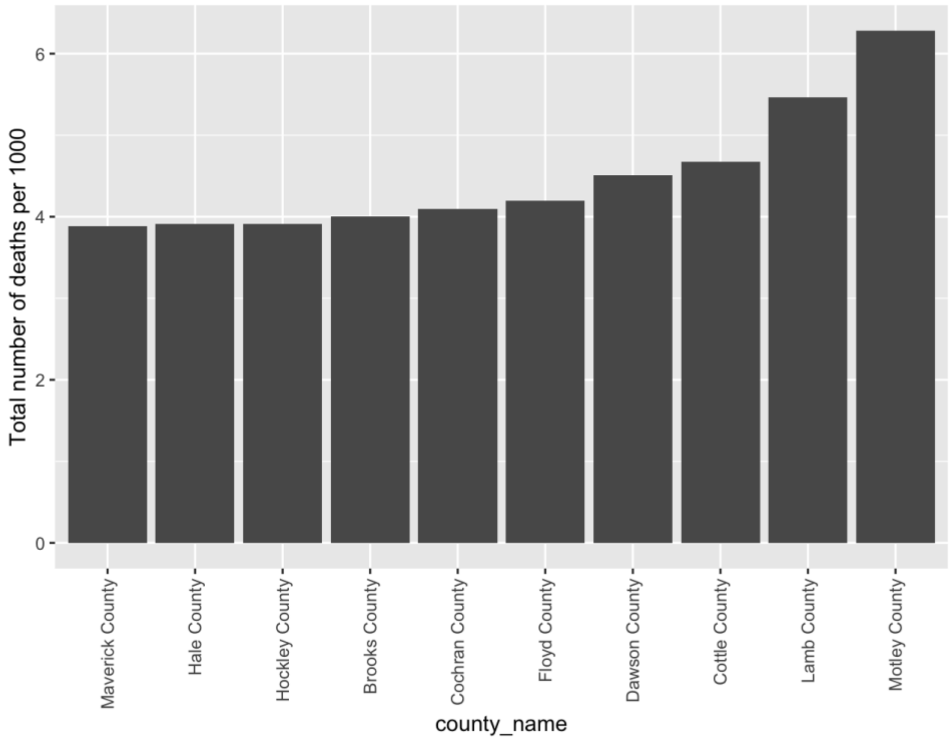


Figure 27 Counties with highest deaths - Normalized by population.

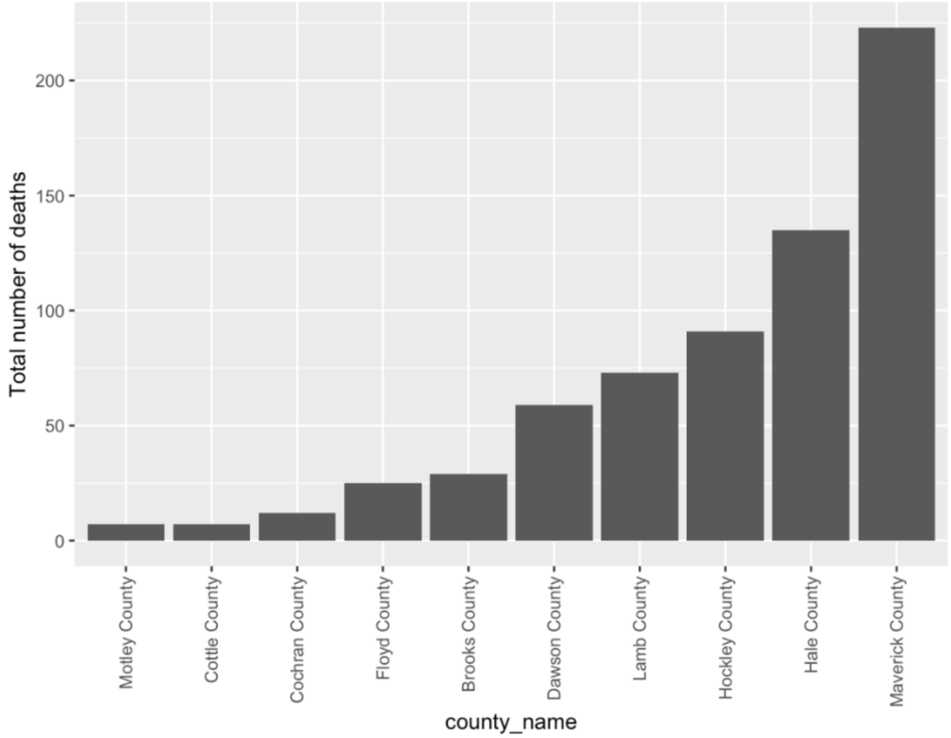
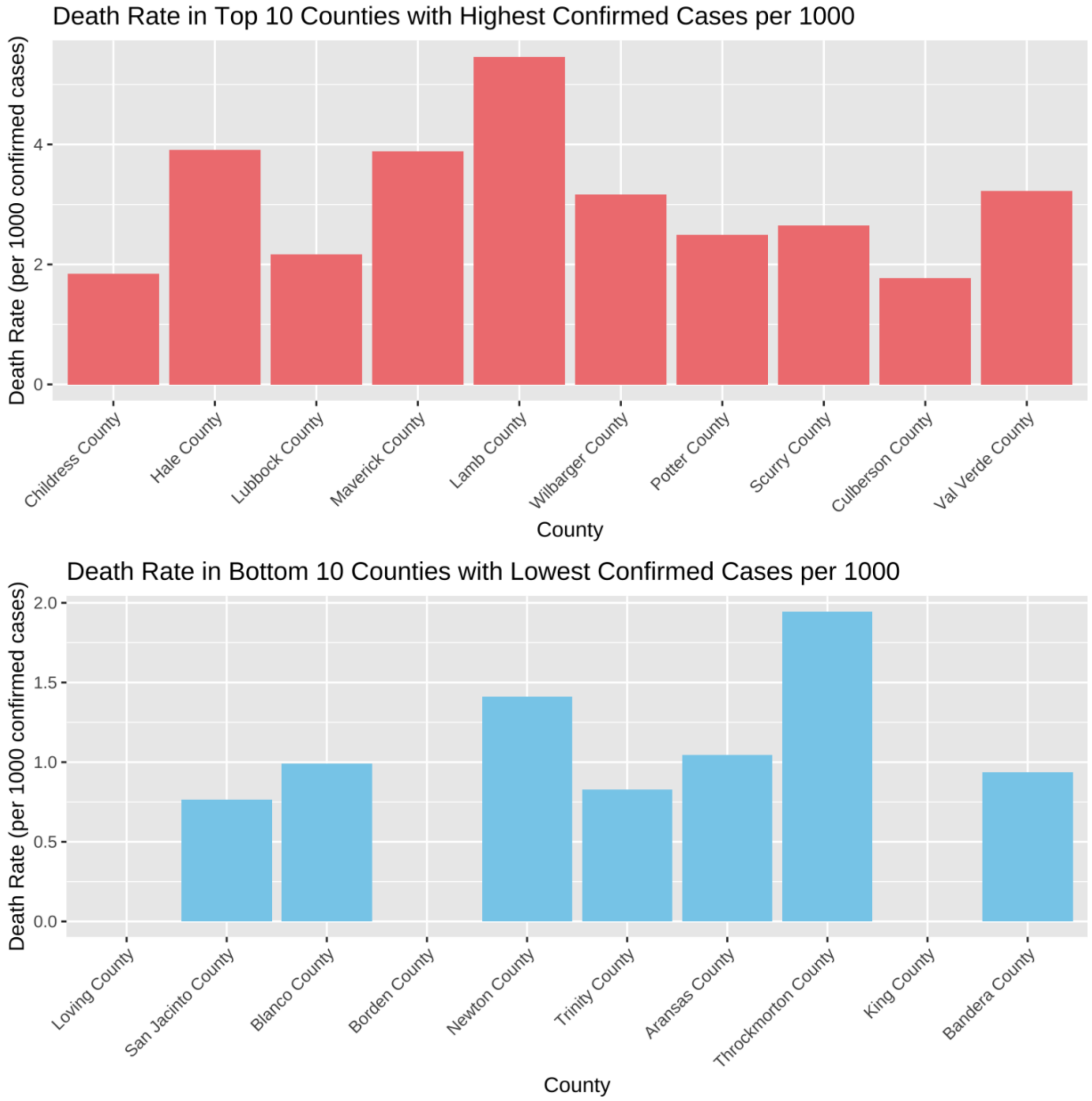


Figure 28 Counties with highest deaths - Not normalized by population.

The two plots above reveal a hidden phenomenon, which is, Maverick County, despite recording the highest number of deaths had the lowest number of deaths per 1000 among the top 10 counties in Texas. Motley County recorded the lowest number of deaths among the top 10 counties but had the highest number of deaths per 1000.

Case to Death Ratio or Fatality Measure of COVID in top states.



Death Rate in Top 10 Counties with Highest Confirmed Cases per 1000

Death Rate in Bottom 10 Counties with Lowest Confirmed Cases per 1000

Figure 29 Case to Death Ratio or Fatality Measure of COVID in top states.

Fatality rate is the measure of confirmed cases that do or do not result in death. A lower fatality rate indicates less probability of death if a person is affected by it in that region.

From the above figure, we can see that the Death rate or the Fatality rate of COVID varies across the top 10 and the bottom 10 counties per 1000 confirmed cases. The order of the counties in the x-axis is the same as the one we have seen in our previous normalized graph indicating the arrangement in decreasing order of cases confirmed per 1000 as it moves from left to the right.

While the counties in x-axis are ordered that way, we can see that the scale of death rate differs. Though cases and number of deaths are correlated at a high level, there are differences in death rate of the affected people in the above graph. This indicates the effect of other factors such as availability of medication, affordability due to poverty (like food, shelter etc.), change in expenditure due to covid, general health level of people in that area, covid protocols followed, retails and recreation closed, immunization done and the success rate of immunization etc.

Hence correlation of deaths alone cannot be a sole deciding factor for the quantification of Fatality or Mortality rate. It could also be affected due to various other factors like the ones mentioned above. Finding such key elements that tip the fatality to a lower level could prove to be beneficial in replicating similar strategies across counties to bring down death rate.

How does median income affect rural areas and urban areas?

The counties are categorized based on rural/urban status for this context as the economy and income in the respective places differ by a considerable margin. Hence, it would be more insightful to compare rural to rural and urban to urban to understand the economic trend during covid with considerations given to its overall economy. Our goal for the visualization shown below is to find if people had the affordability to stay home even after all the economic fluctuations during COVID to follow social distancing during crucial times. When median income of a particular county is visualized vs month in a year, it would also give us insight into the unemployment rates due to covid at each point in a year. However, we neglected it as there is no proof in our finalized data set to consider the decrease is absolutely due to unemployment and no other factors like a decrease in the total economy which could bring the cost of living/affordability down as well; at which point there is no relationship between income affecting covid spread or death rate.

We have computed the 'median rent times x10' and visualized it against their 'median income per month'. Median income per month was calculated as follows: median income in a year/12 to compare with the median rents people are paying on a monthly basis.

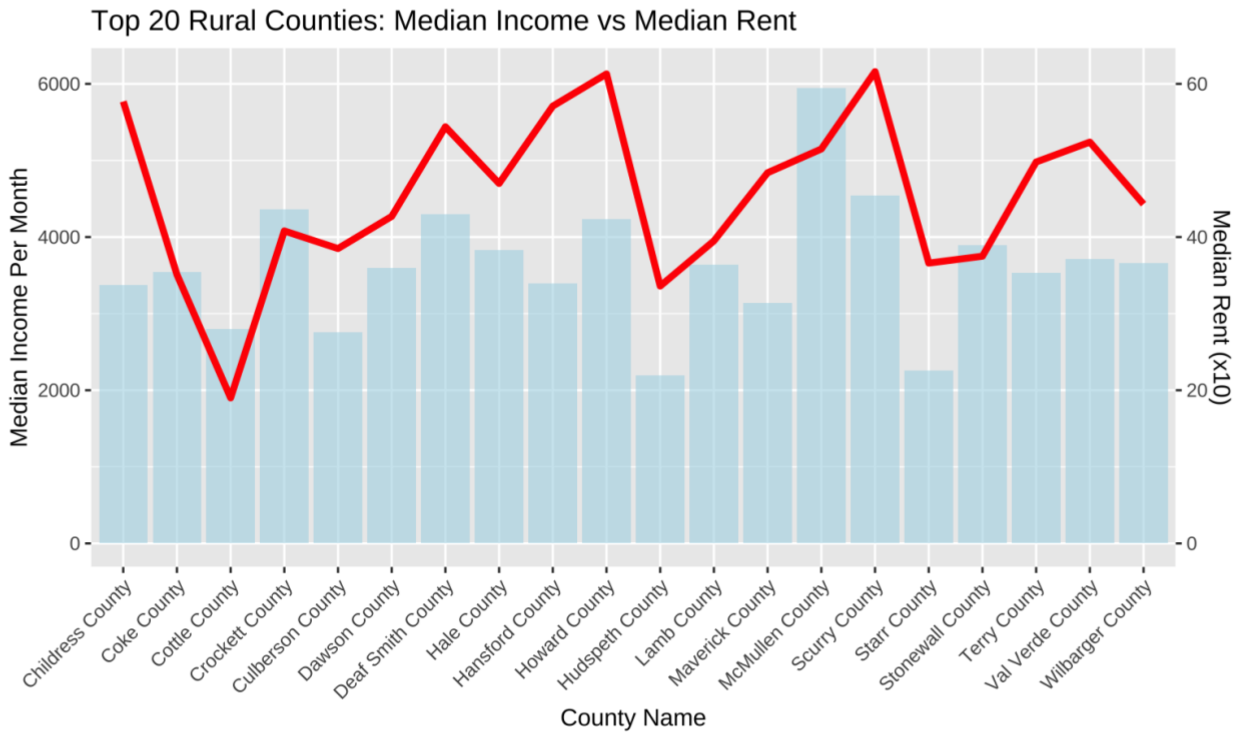


Figure 30 Comparison of median income and median rent in top 20 Texas rural counties.

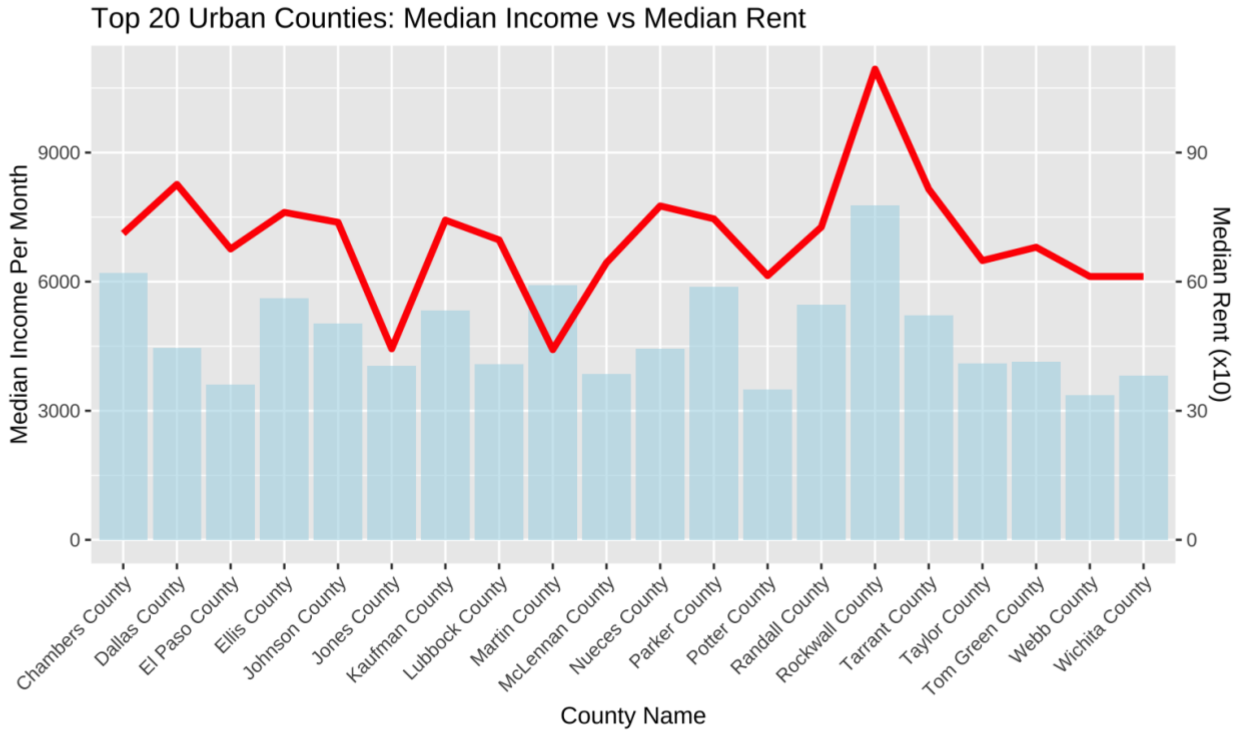


Figure 31 Comparison of median income and median rent in top 20 Texas urban counties.

As expected from the above comparison, we can see the average median income per month in the Urban areas is significantly higher than the ones observed across the Rural counties. When we calculate the median rent x10, we can see rents ranging from 380 to 600 dollars in the Rural areas with their average median income of around 3,000 to 4,500 dollars, while the average rents range from 450 to 750 dollars in the Urban areas with average incomes from around 4,500 to even 7,500 at the highest.

The ratio of the salary amount that the Urban people paid as rent during the covid time frame is considerably less than what portion of their income the Rural people had to pay. This further solidifies the claim that people in Rural area struggle relatively more than the ones in Urban area when it comes to affordability of safety measures and medical expenses with their median wage during covid.

Are the number of cases driving percent change in retail and recreation in the counties? Would Counties with high cases have high or low retail and recreation deviation from baseline?

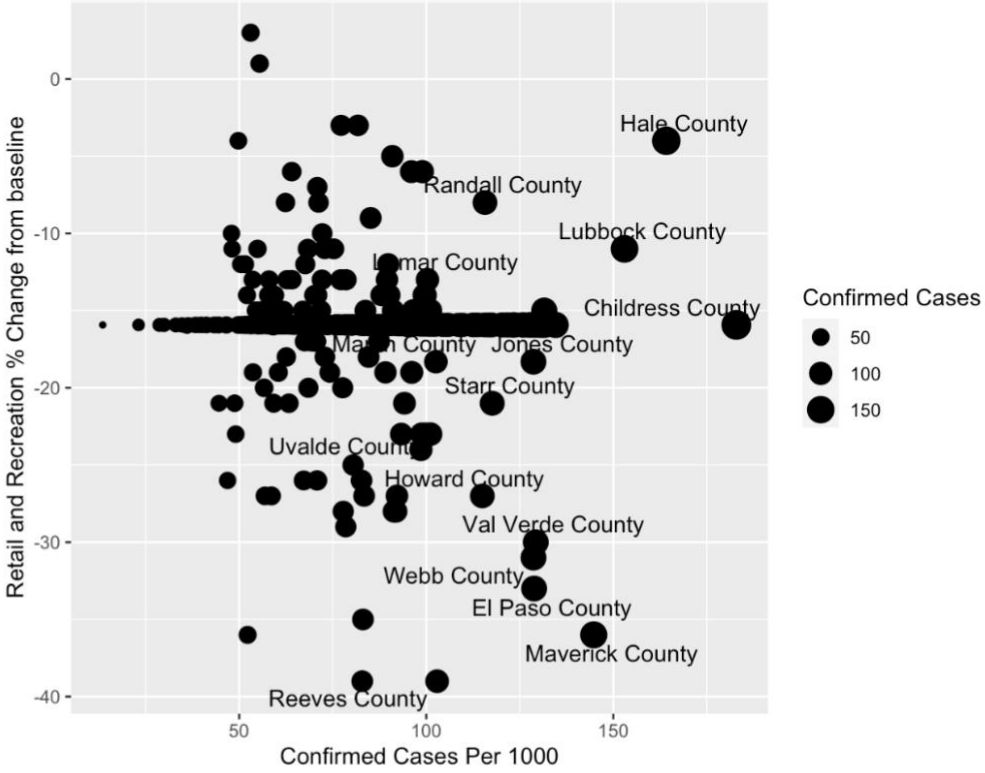


Figure 32 Tracking effect of recreation and retail changes on number of confirmed covid cases.

We see no clear patterns from the chart that suggest that the number of cases in a County affected Retail and recreational baseline.

Are the number of cases driving percent change in people going to work?

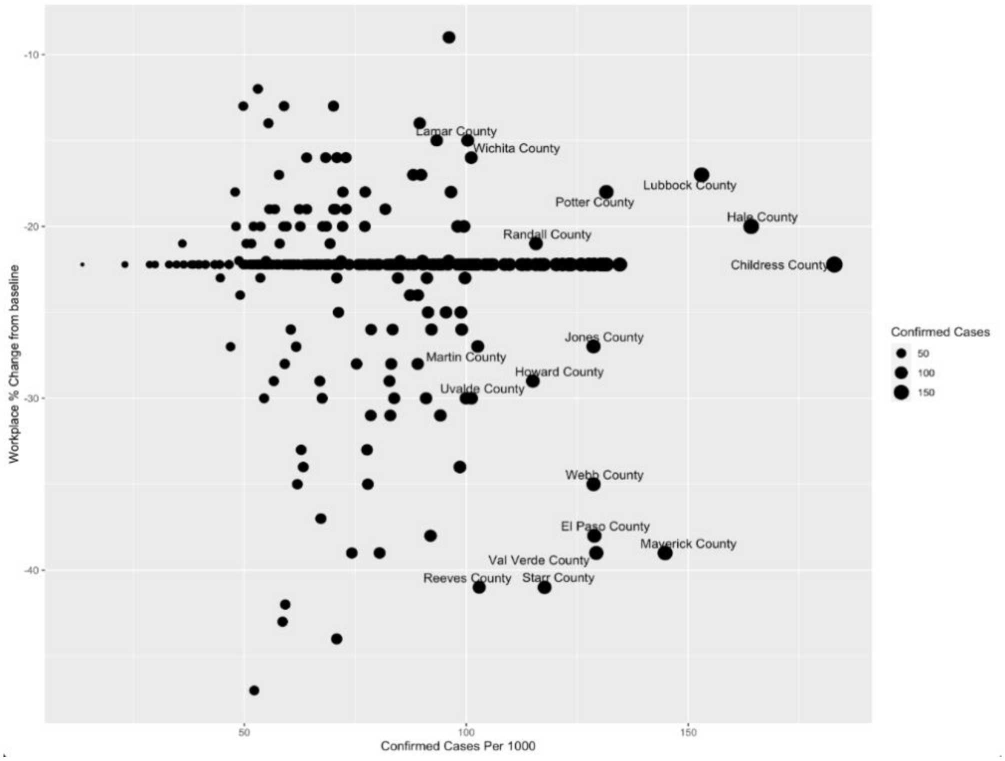


Figure 33 Tracking effect of Workplace % change from baseline on number of confirmed covid cases.

Similarly, we see no clear pattern or relationship between the number of cases and the percentage change in people going to work.

Did employees working at home during the Covid pandemic reduce the number of cases?

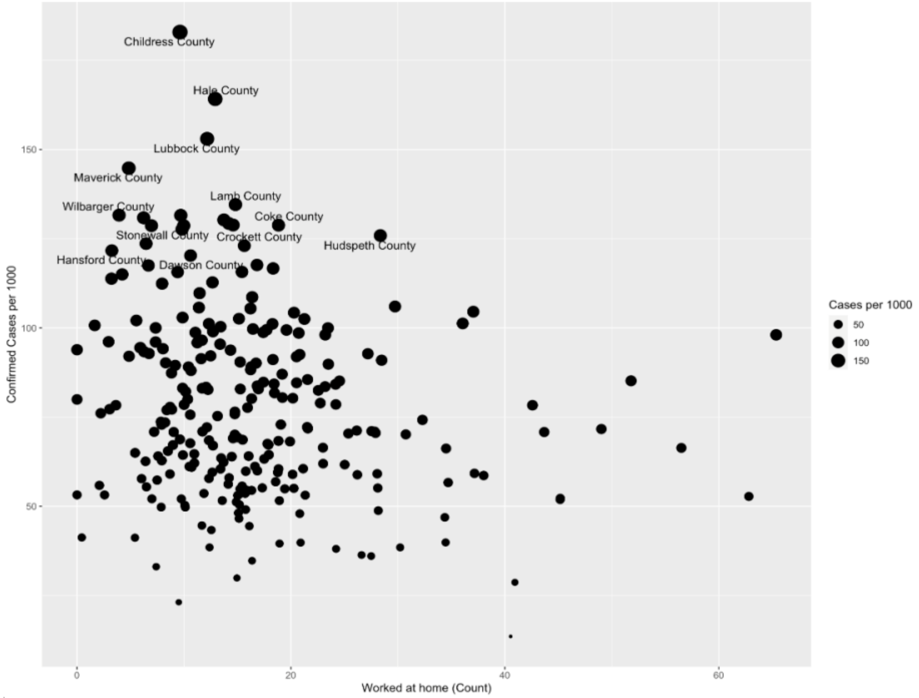


Figure 34 Tracking effect of working from home on number of confirmed covid cases.

We see from this plot that most Counties where few people worked from home had higher numbers of confirmed covid cases reported. Therefore, we can conclude that government shutdowns and mandating that people work from home is an effective measure in reducing the number of covid cases.

Were people who commute using public transportation more susceptible to contracting covid?

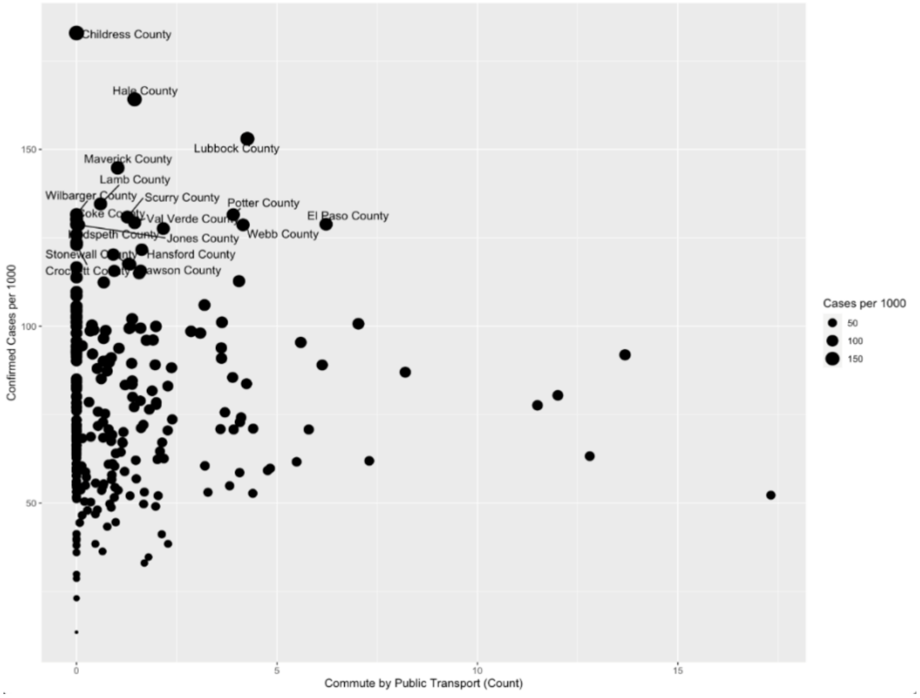


Figure 35 Tracking effect of public transport commute on number of confirmed covid cases.

We observe in this plot that most cases of covid are not in any way related to commuting by public transportation. We believe the reason for this could be because public transport operators have implemented restrictions where commuters do not sit next to each other. This form of social distancing could be the reason the number of confirmed cases is not high.

How were different age groups affected by Covid-19?

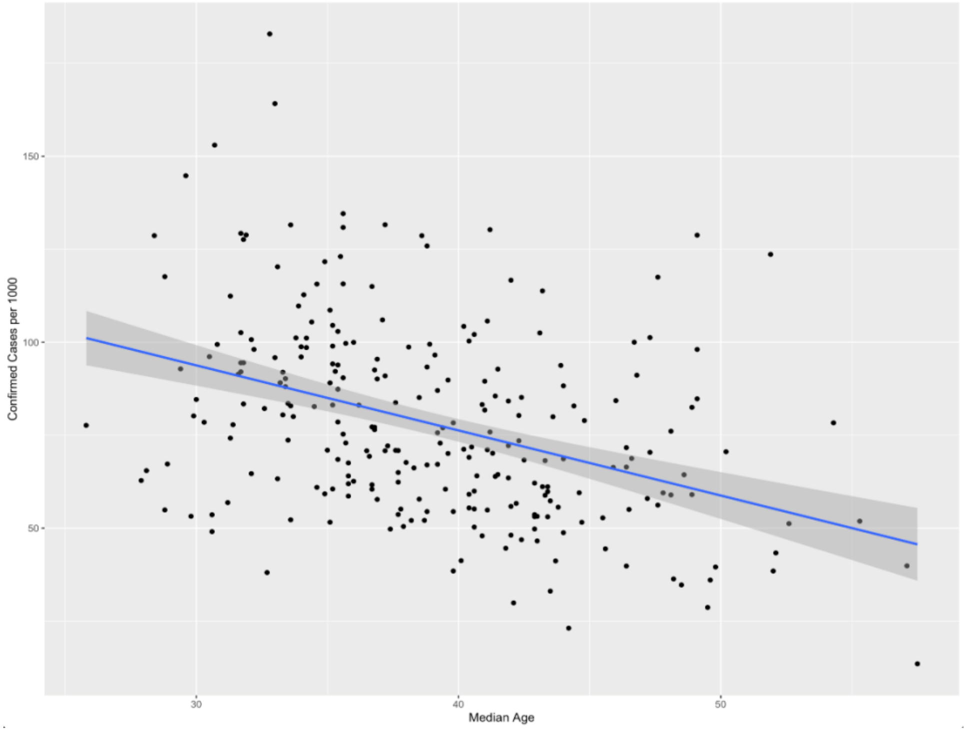


Figure 36 Tracking Covid cases among age groups in geographic area.

We observe that most confirmed cases affected the age group of approximately 35 – 45 years old. This is probably because this age group is the bulk of the workforce, and they are the most active in society hence are more at risk of exposure.

Which age group had the highest mortality rate due to Covid?

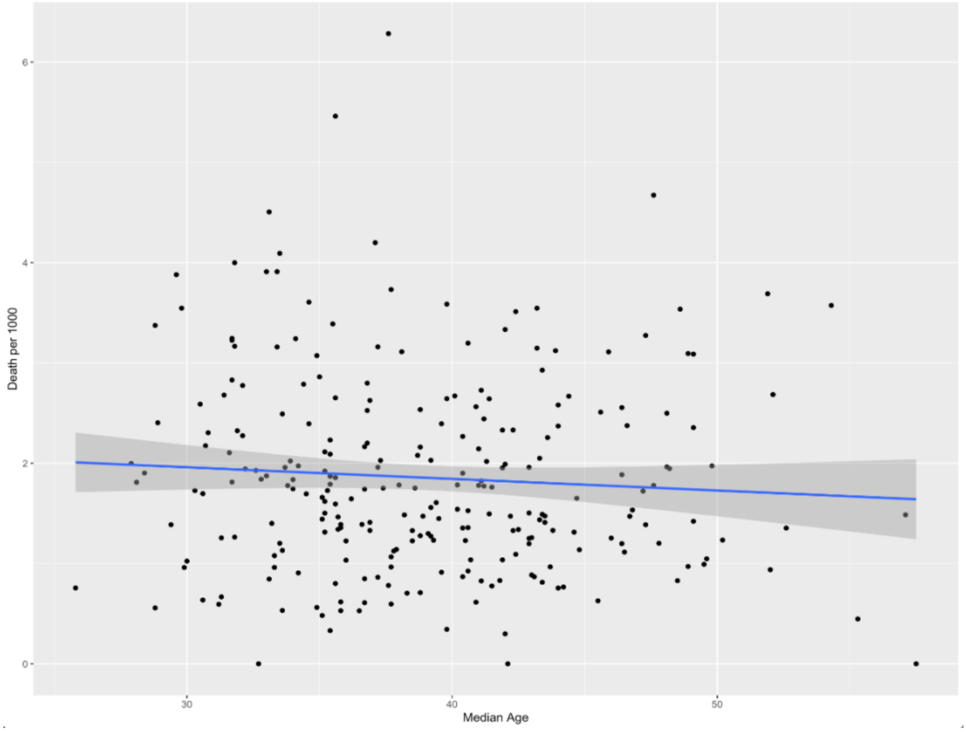


Figure 37 Tracking Covid deaths among age groups in geographic area.

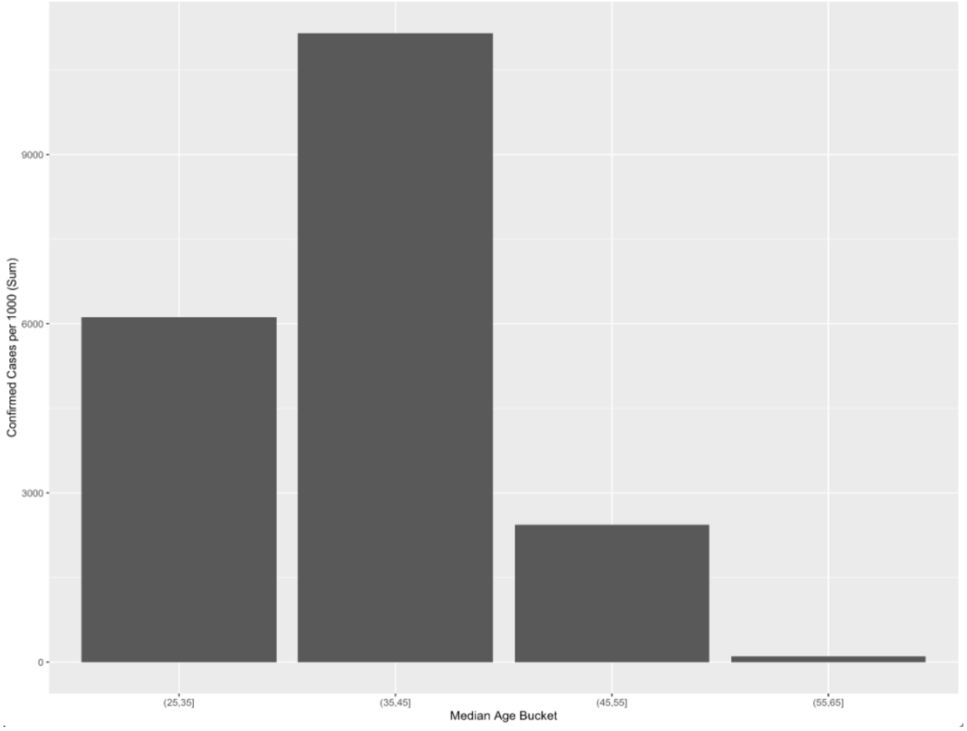


Figure 38 Tracking Covid deaths among age groups in geographic area.

We see that most people who died due to covid are in the 35 – 45 years age group. This is consistent with the observation that this age group also had the highest confirmed cases. We also note that there is no strong correlation with this observation as seen from the regression line being almost perfectly horizontal.

Exceptional Work

To remove the outliers from our data set, we implemented a process called Winsorization. It is a statistical technique used to address outliers in a dataset while preserving the integrity of the data. It involves replacing extreme values with less extreme values, typically by setting outliers equal to the values at predefined cutoff points. By doing so, Winsorization reduces the influence of outliers on statistical analyses without entirely discarding potentially valuable data points. The process begins by identifying outliers, followed by determining cutoff points beyond which values are considered extreme. Outliers are then replaced with values at the cutoff points, effectively "pulling in" the extreme values towards the center of the distribution. This iterative process can be repeated multiple times, gradually reducing the impact of outliers. [6] After performing this method, our outliers were removed.

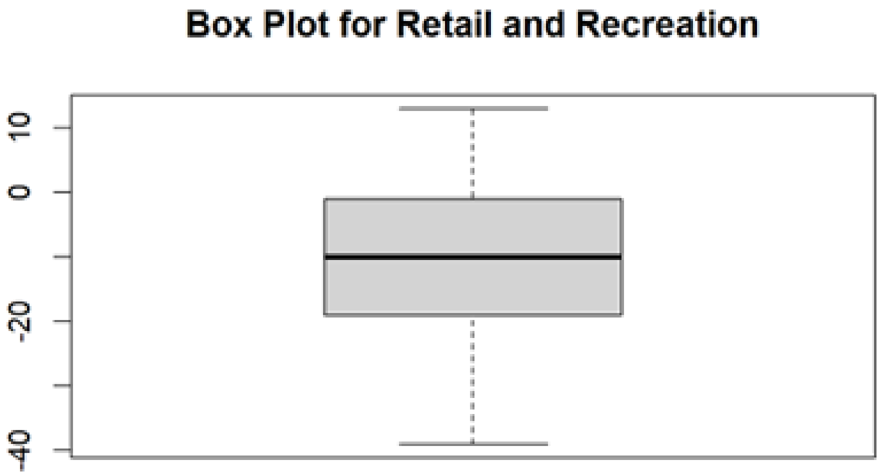


Figure 39 Post outlier treatment distribution of Retail and Recreation.

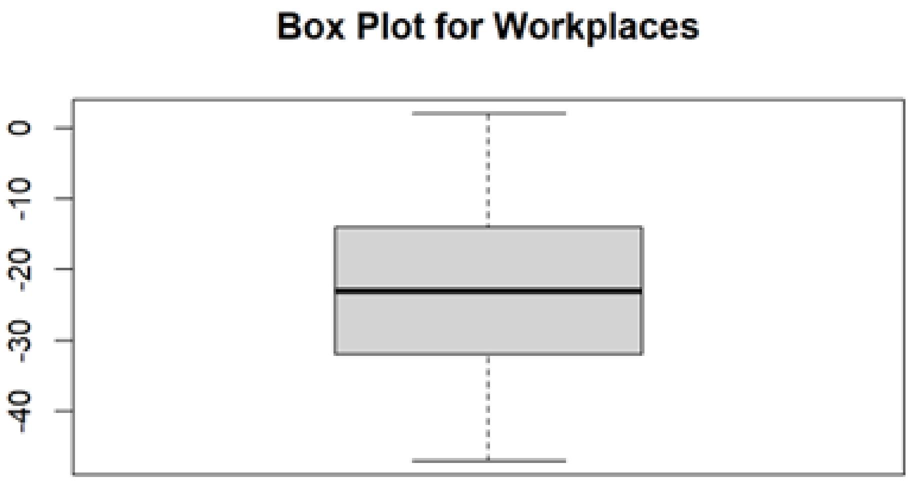


Figure 40 Post outlier treatment distribution of Workplaces feature.

Before implementing this technique, the outliers heavily impacted the statistics and graphs. Below are the stats/graphs before this technique was applied:

Table 14: Descriptive statistics for Mobility Report before Outliers Removed

Feature	Min	Q1	Median	Mean	Q3	Max
Retail and Recreation percent change from baseline	-95	-19	-9	-10.61681	-1	187
Workplaces percent change from baseline	-89	-32	-23	-23.32064	-14	40

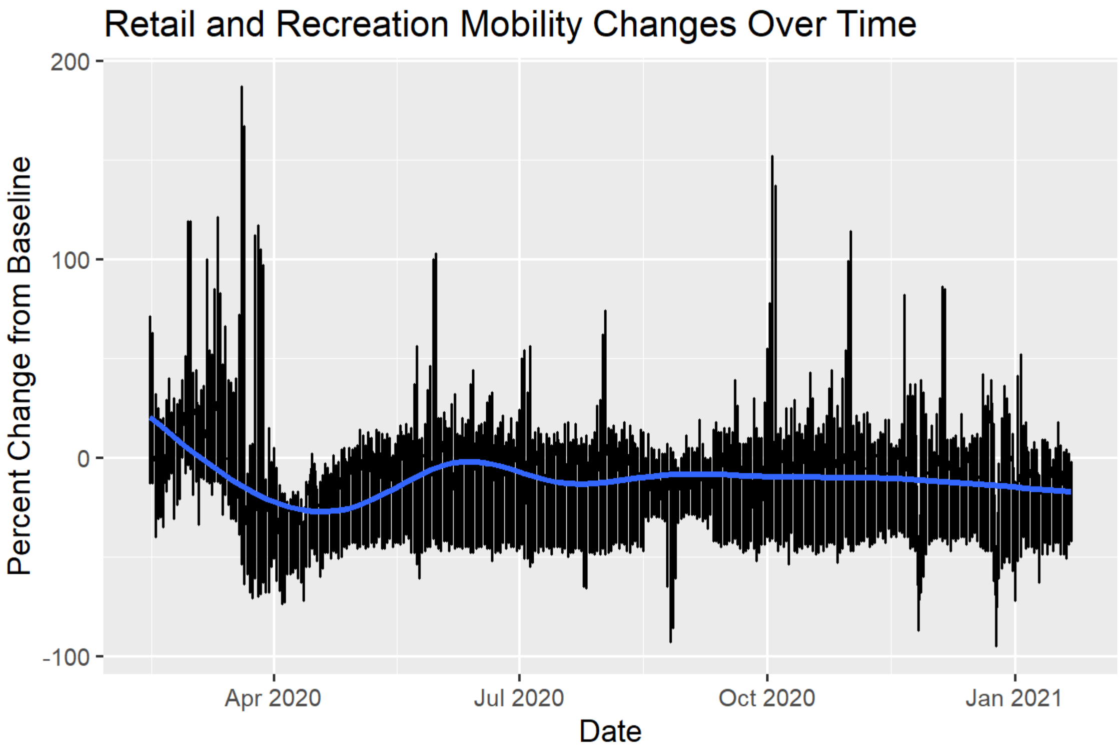


Figure 41 Distribution of Retail and Recreation mobility changes before Outliers remove.

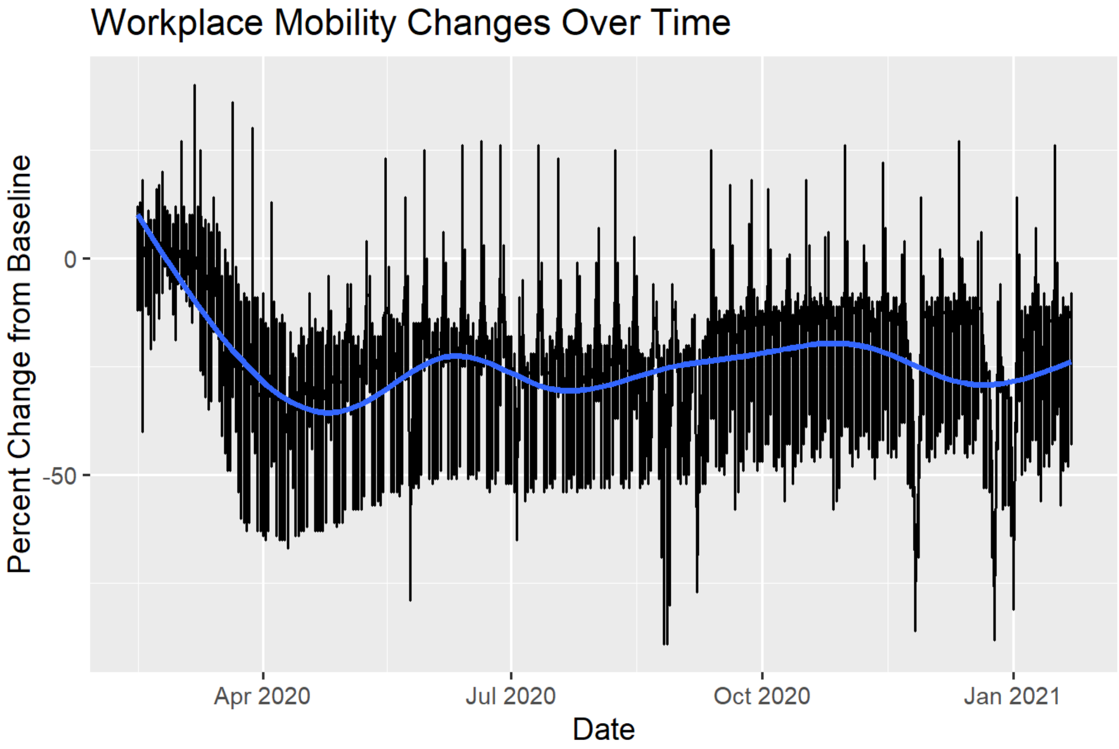


Figure 42 Distribution of Workplace mobility changes before Outliers Removed.

As you can see, the stats are similar to how it looks after applying this technique as indicated in Table 9. The main difference is in the min, max, and mean values. Since the outliers are present on both the upper end and the lower end of the data, this technique works perfectly and improves the quality of the data.

Conclusion

In summary, our analysis has shown that the cases and death rates in Texas Counties underscore the significant influences of factors such as Age, Mode of commute, Mode of Work, Retail and Recreations, Amount of Income, Expenditure for people in Rural and Urban areas, their Affordability of Medicine, Effect of Immunization and so on. It is also interesting to note that the correlation of numerous factors affecting our economic stimulus check is different at various levels of views.

The findings strongly suggest that individuals with lower incomes, those dependent on public transportation, and so on are disproportionately affected by COVID-19, especially the Rural areas. In response, we advocate for an appropriate response form the Department of Treasury, the Bureau of the Fiscal Service, and the Internal Revenue Service (IRS) to address the counties with more economical burden by exploring providing the facilities, policies, and infrastructures necessary to curb the effects of covid, thus fortifying the counties with alarming numbers based on our data-driven insights.

We have derived various trends exhibited by the given data along with the identification of special cases like counties performing good while belonging to a bad region. Leveraging these insights derived to aid in the formulation of plans for other affected counties can help us replicate the identified defensive phenomenon across the board; therefore, reducing the impact levels and fatality rates. We should also note that our insights are only constrained within our 23 finalized features across the given datasets pertaining to the economic sector for the academic purpose; working with the full covid datasets to include as many features as possible to analyze trends at a deeper level will be appropriate to identify patterns and implement processes with much higher accuracy at a practical level.

References

[1] - https://www.who.int/europe/emergencies/situations/covid-19

[2] - https://hub.jhu.edu/2020/03/13/what-is-social-distancing/

[3] - https://www.discovery.co.za/corporate/covid19-flatten-curve#:~:text=This%20curve%20shows%20a%20slow,illness%20at%20the%20same%20time.

[4] - https://home.treasury.gov/policy-issues/coronavirus/assistance-for-american-families-and-workers/economic-impact-payments

[5] - https://www.cnbc.com/2022/06/11/the-pandemic-stimulus-checks-were-a-big-experiment-did-it-work.html

[6] - https://www.statisticshowto.com/winsorize/

[7] - https://en.as.com/en/2022/03/17/latest_news/1647514558_032514.html

[8] - https://www.tdhca.state.tx.us/home-division/docs/22-IndexCounties.pdf

Distribution of Work

- Document Writing/Formatting: Everyone
- Abstract: William
- Business Understanding: Kevin
- Data Understanding
  - COVID-19 cases plus census Data Set: William
  - COVID-19 cases TX Data Set: Hareish
  - Global Mobility Report Data Set: Kevin
- Data Preparation: Everyone
- Conclusion: Hareish
- Exceptional Credit: Kevin