DATA MINING PROJECT 1

# Analyzing Covid 19 and its impact on Callifornia

| | |
|---|---|
| Dhruvil Patel | 49375423 |
| Juhi Shah | 49351308 |
| Vishakha Satpute | 49258111 |

# EXECUTIVE SUMMARY

COVID-19 is an illness triggered by a novel virus previously unrecognised by scientists. Proximity to individuals who are infected with COVID-19, when they cough or sneeze, increases the risk of transmission to others. It started in a city called Wuhan in China in December 2019, and then it continued to spread to other countries quickly. As the virus began to spread, many people became sick, and tragically, some lost their lives. Governments worldwide took decisive action to stem the spread, urging people to stay home, wear masks, and avoid large gatherings. This global health crisis profoundly affected communities everywhere, leading to significant changes in how we live our lives and interact with one another.

Now, let's focus on California. We possess abundant data that illustrates the varied impacts of COVID-19 across different regions of California. This data includes information on the number of confirmed cases and deaths in various counties, giving us insights into how severely each region was affected. Additionally, we have demographic, economic, and housing data about the people living in these areas, such as their ages, races, and income levels. Understanding these demographics helps us identify which groups may have been disproportionately affected by the virus.

In our report, we present this data using various visualisations such as charts, graphs, and maps. These visual aids make it easier to comprehend the complex information and identify patterns or trends. By visually representing the data, we can better understand the impact of COVID-19 on different communities within California.

Our analysis of this data allows us to draw important conclusions about the pandemic's effects. For instance, we can identify which demographic groups were most affected by the virus and which regions experienced the highest rates of infection. This information is crucial for policymakers and government officials as they determine where to allocate resources and implement targeted interventions to mitigate the spread of the virus.

Ultimately, this report serves as a valuable tool for the government of California in devising strategies to combat COVID-19 and minimise its impact on the state's residents. By leveraging data-driven insights, policymakers can make informed decisions to protect public health and support communities affected by the pandemic. In essence, this report provides a roadmap for navigating the challenges posed by COVID-19 and guiding California toward a brighter, healthier future.

## Contents

## List of Tables

## List of Figures

# BUSINESS UNDERSTANDING

COVID-19, an abbreviation for Coronavirus Disease 2019, is attributed to the SARS-CoV-2 virus. Originating in Wuhan, China, in December 2019, the outbreak likely stemmed from a seafood market. COVID-19 spreads easily from person to person through respiratory droplets when someone coughs, sneezes, or talks. Since it began, COVID-19 has quickly spread worldwide, leading the World Health Organization (WHO) to declare it a pandemic on March 11, 2020. Millions of people worldwide have been infected, causing a significant number of deaths.

The exact number of infections and deaths keeps changing as new cases are reported. Currently, there have been hundreds of millions of infections globally, resulting in millions of deaths. COVID-19 causes various symptoms, from mild respiratory issues to severe illness and death, especially in older adults and those with underlying health conditions. Common symptoms include fever, cough, shortness of breath, fatigue, body aches, loss of taste or smell, sore throat, congestion, nausea. Some people may not show any symptoms but can still spread the virus. To control the spread of COVID-19 and reduce its impact on public health and healthcare systems, it's crucial to take preventive measures. These include wearing masks, washing hands regularly, keeping a safe distance from others, and getting vaccinated.

Understanding the impact of COVID-19 on California is crucial for government agencies to make informed decisions and effectively manage the situation. Data analysis and visualisations play a pivotal role in this process by providing comprehensive insights into the spread of the virus, identifying high-risk areas, and understanding demographic and socioeconomic factors influencing transmission rates and healthcare outcomes. By analysing data related to confirmed cases, deaths, population demographics, socioeconomic indicators, and mobility patterns, government agencies can develop targeted strategies to allocate resources, implement preventive measures, and prioritise vaccination efforts.

Data analysis allows government agencies to detect emerging trends and patterns, enabling proactive decision-making to mitigate the impact of COVID-19 on communities. Visual representations of data through charts, graphs, and maps make complex information more accessible and understandable for policymakers, facilitating communication and collaboration across various stakeholders. For example, visualising the geographical distribution of COVID-19 cases helps identify hotspots and areas with limited access to healthcare resources, guiding the deployment of medical personnel, testing facilities, and vaccination clinics. Similarly, analysing mobility data provides insights into population movements and adherence to social distancing measures, informing policies to reduce transmission rates and prevent outbreaks.

Furthermore, data analysis enables government agencies to evaluate the effectiveness of interventions and adjust strategies in real-time based on evolving trends. Overall, data analysis and visualisations empower government agencies to make informed decisions, optimise resource allocation, and effectively manage the COVID-19 pandemic in California, ultimately safeguarding public health and minimising socioeconomic disruptions.

# Data Understanding

We'll be working with three different datasets for our analysis:

1. COVID-19 Cases Plus Census Dataset
2. COVID-19 Cases in Texas Dataset
3. Global Mobility Report Dataset

Across all three datasets, our main goal will be as follows:

1. **Check for Missing Values:** We'll examine each dataset to identify any missing values and decide how to handle them to ensure data completeness.
2. **Verify Data Quality:** It's essential to verify the quality of the data to ensure accuracy and reliability. We'll assess data consistency, validity, and potential errors.
3. **Check for Duplicity:** We'll look for duplicate entries within each dataset and eliminate them to maintain data integrity.
4. **Visualise Different Plots:** Visualisations are key to understanding the data better. We'll create various plots, such as bar graphs, line charts, and heatmaps, to visualise trends and patterns in the data.
5. **Explore Relationships:** We'll analyse the relationships between different variables in the datasets to uncover insights and correlations that could inform our analysis.
6. **Identify Outliers:** Outliers can significantly impact our analysis. We'll identify and examine any outliers present in the data and determine their impact on our analysis.
7. **Feature Engineering:** We can derive new features from existing ones to enhance our analysis. For example, we can calculate the mortality rate from COVID-19 cases and population data.

Overall, by performing these tasks across all three datasets, we aim to gain comprehensive insights into the impact of COVID-19, mobility patterns, and demographic factors, which will inform our analysis and decision-making processes.

## Covid-19 Cases Plus Census Dataset

The COVID-19_cases_plus_census dataset contains 259 features, which are discrete pieces of information about various aspects related to COVID-19 cases and census data. These features cover a wide range of topics such as population demographics, housing characteristics, socioeconomic indicators, healthcare access, and commuting patterns, among others.

Out of the many features available in the dataset, a subset of features is considered for analysis and modelling purposes. Factors such as population demographics, healthcare access, housing characteristics, and socioeconomic indicators are often considered crucial in determining the vulnerability of populations to the virus and informing public health interventions. This dataset contains 259 features and a total of 3143 values which might contain duplicate or missing values which we will see in the next phase of the report.

Among the numerous available features, we have considered the ones displayed in the table as the most significant.

*Table 1: Important Features for 'COVID-19 CASES PLUS CENSUS' DATASET*

| Feature Name | Scale of Measurement | Information |
|---|---|---|
| high_school_diploma | Ratio | Number of individuals with a high school diploma |
| high_school_including_ged | Ratio | Number of individuals with a high school diploma or GED |
| households_public_asst_or_food_stamps | Ratio | Number of households receiving public assistance or food stamps |
| no_cars | Ratio | Number of households with no cars |
| male_45_64_high_school | Ratio | Number of males aged 45-64 with a high school education |
| income_10000_14999 | Ratio | Number of individuals with income between $10,000 and $14,999 |
| deaths | Ratio | Number of deaths |
| confirmed_cases | Ratio | Number of confirmed COVID-19 cases |
| commute_45_59_mins | Ratio | Number of individuals with commute time between 45-59 minutes |
| commute_60_more_mins | Ratio | Number of individuals with commute time of 60 minutes or more |
| rent_burden_not_computed | Ratio | Number of households with rent burden not computed |
| dwellings_3_to_4_units | Ratio | Number of dwellings with 3 to 4 units |
| commute_60_89_mins | Ratio | Number of individuals with commute time between 60-89 minutes |
| rent_over_50_percent | Ratio | Number of households with rent over 50% of income |
| income_less_10000 | Ratio | Number of individuals with income less than $10,000 |
| female_80_to_84 | Ratio | Number of females aged 80 to 84 |
| employed_education_health_social | Ratio | Number of individuals employed in education, health, or social services |
| children_in_single_female_hh | Ratio | Number of children in single female-headed households |
| income_15000_19999 | Ratio | Number of individuals with income between $15,000 and $19,999 |
| commute_90_more_mins | Ratio | Number of individuals with commute time of 90 minutes or more |
| housing_built_1939_or_earlier | Ratio | Number of housing units built in 1939 or earlier |
| female_85_and_over | Ratio | Number of females aged 85 and over |
| dwellings_50_or_more_units | Ratio | Number of dwellings with 50 or more units |
| aggregate_travel_time_to_work | Ratio | Aggregate travel time to work |
| commute_35_44_mins | Ratio | Number of individuals with commute time between 35-44 minutes |

| | | |
|---|---|---|
| State | Nominal | The name or identifier of the state (e.g., California, New York) |
| total_pop | Ratio | The total population of the state. |
| male_pop | Ratio | The male population of the state. |
| female_pop | Ratio | The female population of the state. |
| median_age | Ratio | The median age of the state's population. |
| white_pop | Ratio | The population count of individuals identifying as white within the state. |
| black_pop | Ratio | The population count of individuals identifying as black within the state. |
| hispanic_pop | Ratio | The population count of individuals identifying as Hispanic within the state. |
| amerindian_pop | Ratio | The population count of individuals identifying as American Indian or Alaska Native within the state. |
| other_race_pop | Ratio | The population count of individuals identifying with a race other than those listed above, or with multiple races, within the state. |
| gini_index | Ratio | A measure of income inequality within the state. A Gini index of 0 represents perfect equality, while a Gini index of 1 represents perfect inequality. |

We selected these features for our analysis to understanding the spread of the virus, the demographic groups most at risk, and the socioeconomic factors influencing transmission and outcomes. Here's a breakdown of why these features is important:

**Mobility Patterns:**

- Commute times (commute_45_59_mins, commute_60_more_mins, etc.): These attributes provide insights into the mobility patterns of the population, which can influence the spread of COVID-19. Longer commute times may indicate higher exposure to public transportation and increased interaction with others.
- Education Levels: Education attributes (high_school_diploma, high_school_including_ged): Understanding the education levels of the population can help identify groups that may have different levels of access to information about COVID-19 prevention and treatment.
- Economic Impact: Income attributes (income_10000_14999, income_less_10000, etc.): These features provide insights into the economic impact of COVID-19 on different communities, highlighting areas where financial assistance may be needed.

**Tracking the Pandemic:**

- Case and death counts (e.g., deaths, confirmed_cases): These attributes are essential for tracking the number of COVID-19 cases and deaths over time in different states, helping to assess the severity and progression of the pandemic.

**Demographic Composition:**

- Population attributes (e.g., total_pop, male_pop, female_pop, etc.): Understanding the demographic composition of the population affected by COVID-19 can help identify vulnerable groups and tailor public health interventions accordingly.

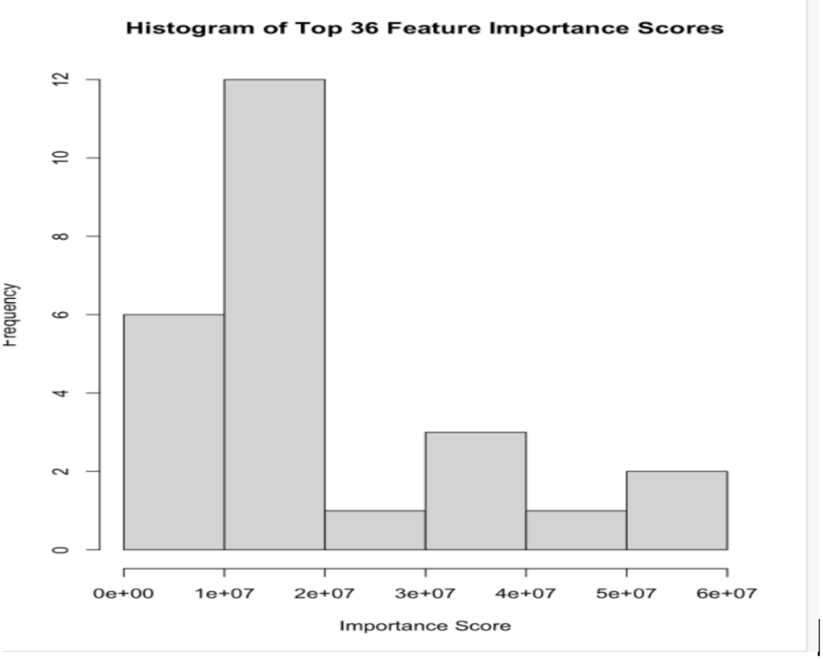
**Economic and Housing Stability Indicators:**

- Features such as households_public_asst_or_food_stamps, no_cars, and rent_over_50_percent: These indicators help identify areas with higher levels of economic vulnerability and housing instability, which can impact the spread and management of COVID-19.

**Demographic and Employment Indicators:**

- Attributes like male_45_64_high_school, employed_education_health_social, and children_in_single_female_hh: These features provide insights into the age, employment sectors, and family structures of the population, which are relevant to understanding the risks and impacts associated with the virus.
- Gini Index: This measure of economic inequality allows for a nuanced understanding of how economic factors intertwine with health outcomes during the pandemic, helping to tailor interventions to those most in need.

By analysing these features, we can identify multi-dimensional view of the pandemic's impact, enabling government agencies to make informed decisions, implement targeted interventions, and support communities effectively. This data-driven approach enhances the ability of government agencies to protect public health and navigate the challenges posed by COVID-19.



## COVID-19_cases_TX Dataset

This dataset contains information about COVID-19 cases specifically in Texas. It includes data such as county names, state information, dates of recorded data, and the number of confirmed COVID-19 cases and deaths in each county. This dataset provides a detailed record of COVID-19 cases and deaths in various counties of Texas over time. This dataset contains 7 features and total of 94351 values which might contains duplicate or missing values which we will see in next phase of the report.

*Table 2: Important features for 'COVID-19_cases_TX' Dataset*

| Feature | Scale | Description |
|---|---|---|
| county_fips_code | Nominal | FIPS code assigned to each county in Texas. |
| county_name | Nominal | Name of the county in Texas. |
| state | Nominal | Name of the state, which is Texas in this dataset. |
| state_fips_code | Nominal | FIPS code assigned to the state of Texas. |
| date | Interval | Date of the recorded data. |
| confirmed_cases | Ratio | Number of confirmed COVID-19 cases in the county. |
| deaths | Ratio | Number of deaths due to COVID-19 in the county. |

## GLOBAL MOBILITY REPORT

The Global Mobility Report dataset provides information on mobility trends during the COVID-19 pandemic across different regions and administrative divisions worldwide. It includes data on various categories such as retail and recreation, grocery and pharmacy, parks, transit stations, workplaces, and residential locations. Each category indicates the percentage change in visits or time spent compared to a baseline period before the pandemic. This dataset helps track how people's mobility patterns have shifted in response to measures implemented to curb the spread of COVID-19, providing valuable insights for public health and policy decision-making. This dataset contains 14 features and total of 1048576 values which might contains duplicate or missing values which we will see in next phase of the report.

*Table 3: Important features for 'Global Mobility Report' dataset*

| Feature | Scale | Description |
|---|---|---|
| country_region_code | Nominal | Code representing the country or region. |
| country_region | Nominal | Name of the country or region. |
| sub_region_1 | Nominal | First-level administrative division within the country or region, such as a state or province. |
| sub_region_2 | Nominal | Second-level administrative division within the country or region, such as a county or district. |

| | | |
|---|---|---|
| metro_area | Nominal | Metropolitan area associated with the sub-region. |
| iso_3166_2_code | Nominal | ISO 3166-2 code representing the sub-region. |
| census_fips_code | Nominal | FIPS code associated with the sub-region. |
| date | Interval | Date of the recorded data. |
| retail_and_recreation_percent_change_from_baseline | Ratio | Percentage change in visits to retail and recreation locations compared to a baseline period before the pandemic. |
| grocery_and_pharmacy_percent_change_from_baseline | Ratio | Percentage change in visits to grocery stores and pharmacies compared to a baseline period before the pandemic. |
| parks_percent_change_from_baseline | Ratio | Percentage change in visits to parks compared to a baseline period before the pandemic. |
| transit_stations_percent_change_from_baseline | Ratio | Percentage change in visits to transit stations compared to a baseline period before the pandemic. |
| workplaces_percent_change_from_baseline | Ratio | Percentage change in visits to workplaces compared to a baseline period before the pandemic. |
| residential_percent_change_from_baseline | Ratio | Percentage change in time spent at residential locations compared to a baseline period before the pandemic. |

# VERIFYING DATA QUALITY

## Covid-19 Cases Plus Census Dataset

### Data Normalisation

Data normalization is achieved using the scale function within a mutate and across call from the ‘dplyr’ package. This method standardizes each selected feature by subtracting the mean and dividing by the standard deviation, ensuring each feature has a mean of 0 and a standard deviation of 1. This process is applied across all specified features, effectively putting them on the same scale without distorting their ranges. This process is applied on only the selected features which are mentioned in Table 1.

## Data Duplicates

Upon analysing the 'Covid-19 Cases Plus Census' dataset for duplicate data, we discovered that none of the selected features contain duplicates.

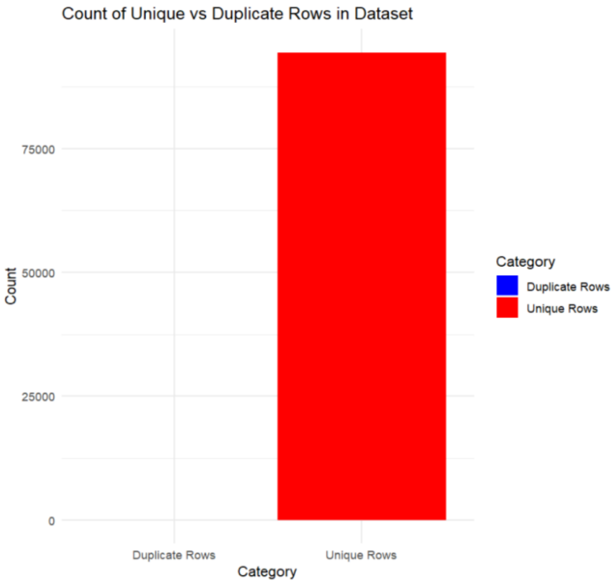


*Figure 1: Count of duplicate and unique rows of Covid 19 cases plus census dataset.*

## Outliers

We can see that there are several outliers above the upper whisker, which means that there are some counties with exceptionally high numbers of confirmed COVID-19 cases compared to most of the other counties. The main body of the box plot, where the bulk of the data lies, seems to be quite compressed, with the median near the bottom of the box, suggesting that most of the data points (the number of confirmed cases across different counties) have lower values. The presence of outliers indicates that there are counties with unusually high case numbers, which may warrant further investigation.

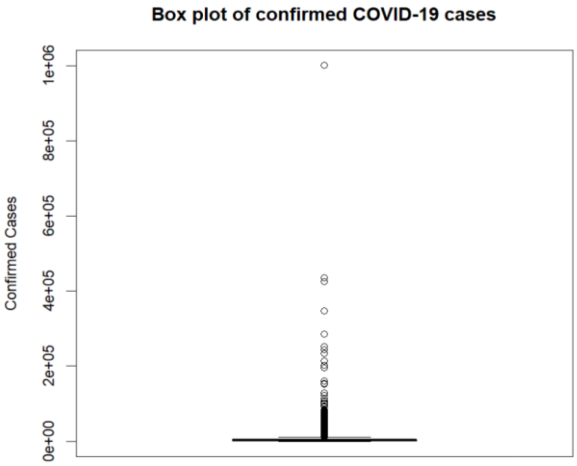


*Figure 2: Outliers for confirmed cases.*

## Missing Values

We analysed and verified the data quality of covid 19 cases plus census dataset for the features we considered which are listed in table 1. We found that all the features have no missing values except ‘aggregate_travel_time_to_work’ which has 143 missing values.

*Table 4: Missing values for Covid 19 Cases Plus Census Dataset*

| Feature | Number of Missing Values |
|---|---|
| high_school_diploma | 0 |
| high_school_including_ged | 0 |
| households_public_asst_or_food_stamps | 0 |
| no_cars | 0 |
| male_45_64_high_school | 0 |
| income_10000_14999 | 0 |
| deaths | 0 |
| confirmed_cases | 0 |
| commute_45_59_mins | 0 |
| commute_60_more_mins | 0 |
| rent_burden_not_computed | 0 |
| dwellings_3_to_4_units | 0 |
| commute_60_89_mins | 0 |
| rent_over_50_percent | 0 |
| income_less_10000 | 0 |
| female_80_to_84 | 0 |
| employed_education_health_social | 0 |
| children_in_single_female_hh | 0 |
| income_15000_19999 | 0 |
| commute_90_more_mins | 0 |
| housing_built_1939_or_earlier | 0 |
| female_85_and_over | 0 |
| dwellings_50_or_more_units | 0 |
| aggregate_travel_time_to_work | 143 |
| commute_35_44_mins | 0 |
| state | 0 |
| total_pop | 0 |
| male_pop | 0 |
| female_pop | 0 |
| median_age | 0 |
| white_pop | 0 |
| black_pop | 0 |
| hispanic_pop | 0 |

| | |
|---|---|
| amerindian_pop | 0 |
| other_race_pop | 0 |
| gini_index | 0 |

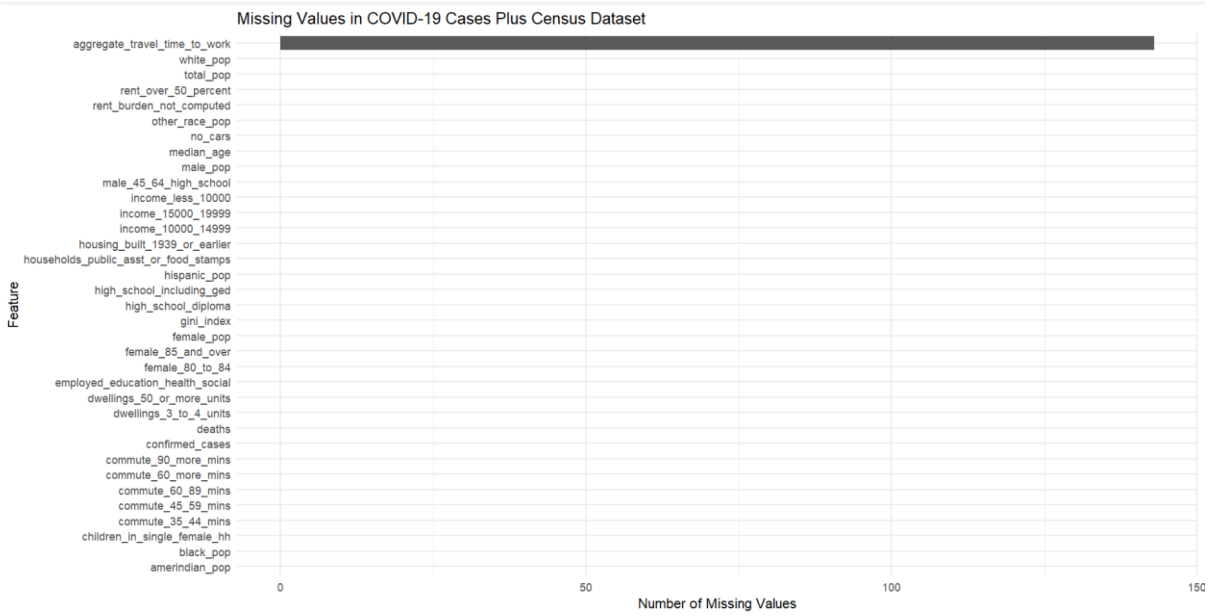


*Figure 3: Missing values of Covid 19 cases plus census dataset.*

# COVID-19_cases_TX Dataset

## Data Normalization

The normalization process scales numeric data in the dataset so that each numeric column will have a mean of 0 and a standard deviation of 1.

## Data Duplicate

After analysing for duplicate values, we found that there are no duplicate values in ‘COVID_19_cases_TX’ Dataset.

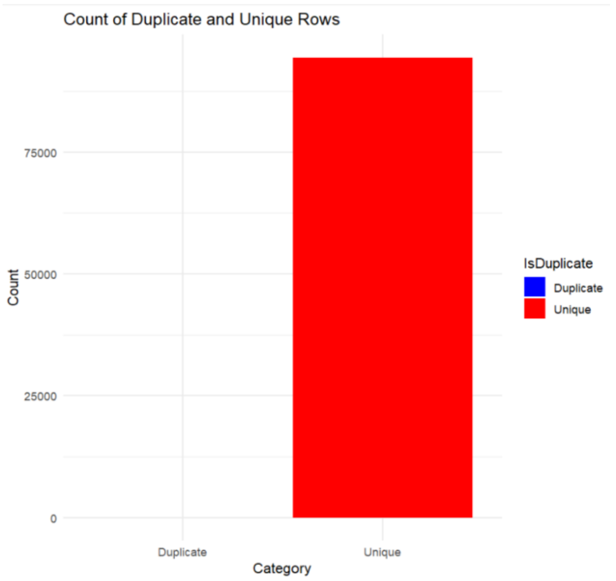


*Figure 4: Count of duplicate and unique rows for 'COVID_19_cases_TX' dataset.*

## Outliers

As seen in below boxplot, there are no outliers in the COVID-19_cases_TX dataset.

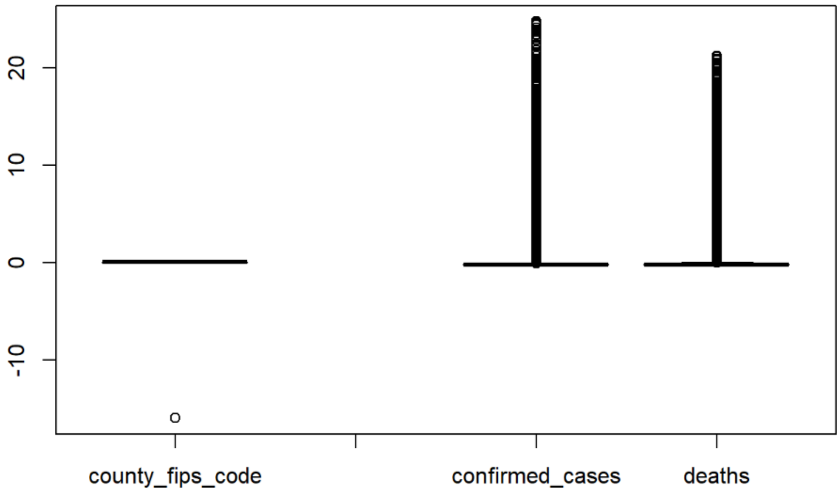


*Figure 5: Outliers of 'COVID-19_cases_TX' dataset.*

## Missing Values

As we run analyse behind this dataset in form of code snippet, we came across that this data is complete 100 % as there are 0 missing values when we trace the output, we check that there were 0 missing values as shown below.

*Table 5: Missing values for Covid-19 cases TX dataset*

| Feature Name | Number of Missing Values |
|---|---|
| county_fips_code | 0 |
| county_name | 0 |

| | |
|---|---|
| state | 0 |
| state_fips_code | 0 |
| date | 0 |
| confirmed_cases | 0 |
| Deaths | 0 |

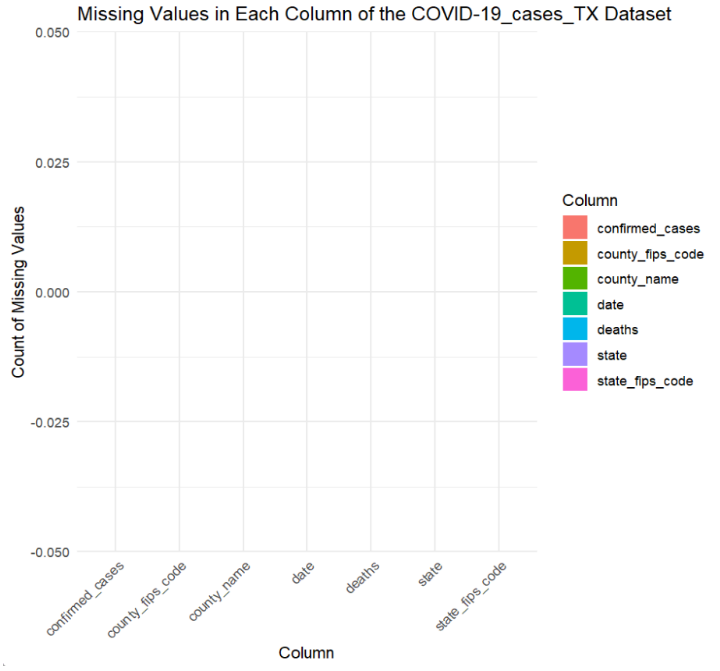


*Figure 6: Missing values in COVID-19_cases_TX Dataset*

# GLOBAL MOBILITY REPORT

## Data Normalization

Data normalization is performed on the numerical columns of the dataset. We rescaled the features to have a mean of 0 and a standard deviation of 1.

## Data Duplicates

After analysing for duplicate values, we found that there are no duplicate values in ‘global mobility report’ Dataset.

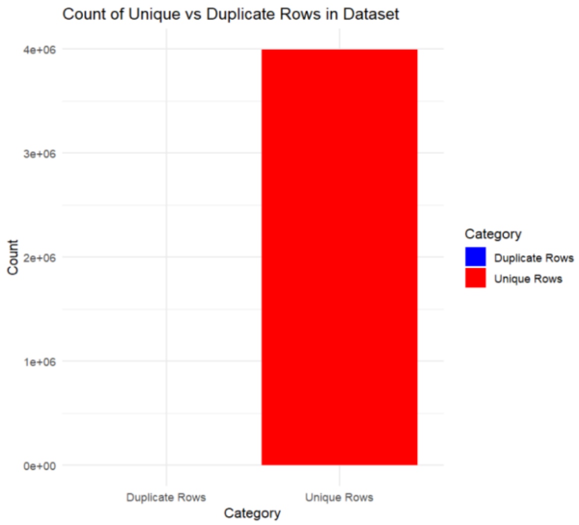


*Figure 7: Count of duplicate and unique rows for 'Global Mobility Report' dataset.*

## Outliers

The below image shows a boxplot titled "Outliers," which is Figure 8 from what is referred to as the 'GLOBAL MOBILITY REPORT' dataset. This boxplot is used to identify outliers in two different variables: census_fips_code and parks_percent_change_from_baseline.

In a boxplot, the central box represents the middle 50% of the data (the interquartile range or IQR), with the horizontal line inside the box denoting the median. The "whiskers" — the lines extending from the top and bottom of the box — typically represent the range of the data excluding outliers, often set to 1.5 times the IQR above and below the box. Points that fall outside of these whiskers are considered outliers and are plotted as individual points.

For the census_fips_code variable, there are no points beyond the whiskers, indicating that there are no outliers in this dataset for this variable. For the parks_percent_change_from_baseline variable, there are numerous points above and below the whiskers, indicating that there are many outliers in this dataset for the variable, showing significant deviation from typical park visitation patterns as compared to the baseline.

The presence of outliers in the parks_percent_change_from_baseline could indicate days when park visitation was unusually high or low, which could be due to various factors like holidays, weather events, or restrictions due to the COVID-19 pandemic.

This visualization is particularly useful for data analysts and researchers who are interested in understanding the variability and extremities in mobility trends during the period the data was collected.

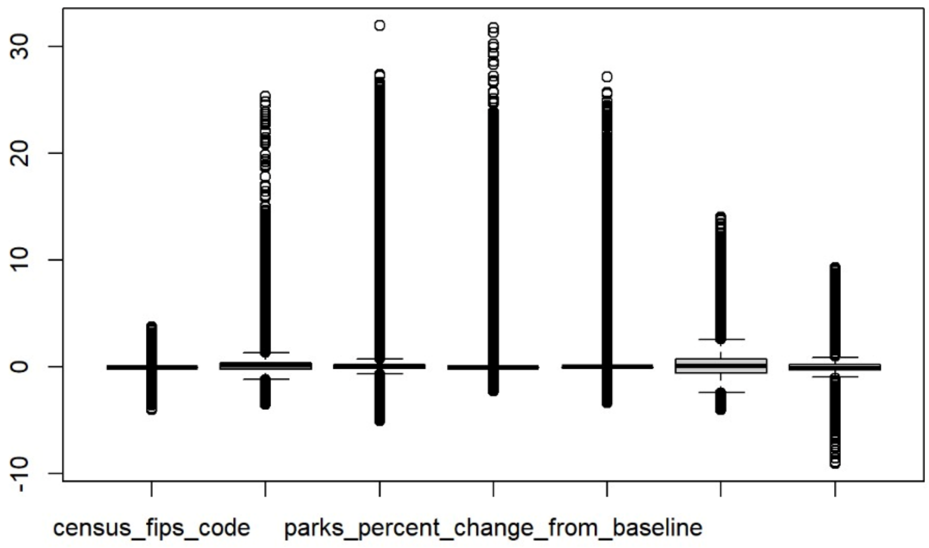


*Figure 8: Outliers for 'GLOBAL MOBILITY REPORT' dataset.*

## Missing Values

When analysing our dataset, we encountered a significant number of missing values across multiple features. To ensure the integrity of our analysis and maintain the dataset's completeness, we decided to address these missing values through imputation techniques. By imputing missing values, we aim to fill in the gaps in our dataset with estimated values based on the available information, thereby allowing us to proceed with our analysis without compromising the overall quality of the data. This process involves carefully selecting and applying appropriate imputation methods to accurately replace the missing values with plausible estimates. Through this approach, we can ensure that our dataset remains robust and suitable for further analysis, enabling us to derive meaningful insights and make informed decisions based on the complete set of data.

Before addressing missing values in our dataset, let's take a closer look at the initial state of the data. The following are the counts of missing values for each feature:

*Table 6: Missing values for 'Global Mobility Report'*

| Feature Name | Number of Missing Values |
|---|---|
| country_region_code | 2544 |
| country_region | 0 |
| sub_region_1 | 0 |
| sub_region_2 | 0 |
| metro_area | 0 |
| iso_3166_2_code | 0 |
| census_fips_code | 3139208 |
| date | 0 |
| retail_and_recreation_percent_change_from_baseline | 1478424 |

| | |
|---|---|
| grocery_and_pharmacy_percent_change_from_baseline | 1564666 |
| parks_percent_change_from_baseline | 2080860 |
| transit_stations_percent_change_from_baseline | 1973496 |
| workplaces_percent_change_from_baseline | 189760 |
| residential_percent_change_from_baseline | 1678955 |

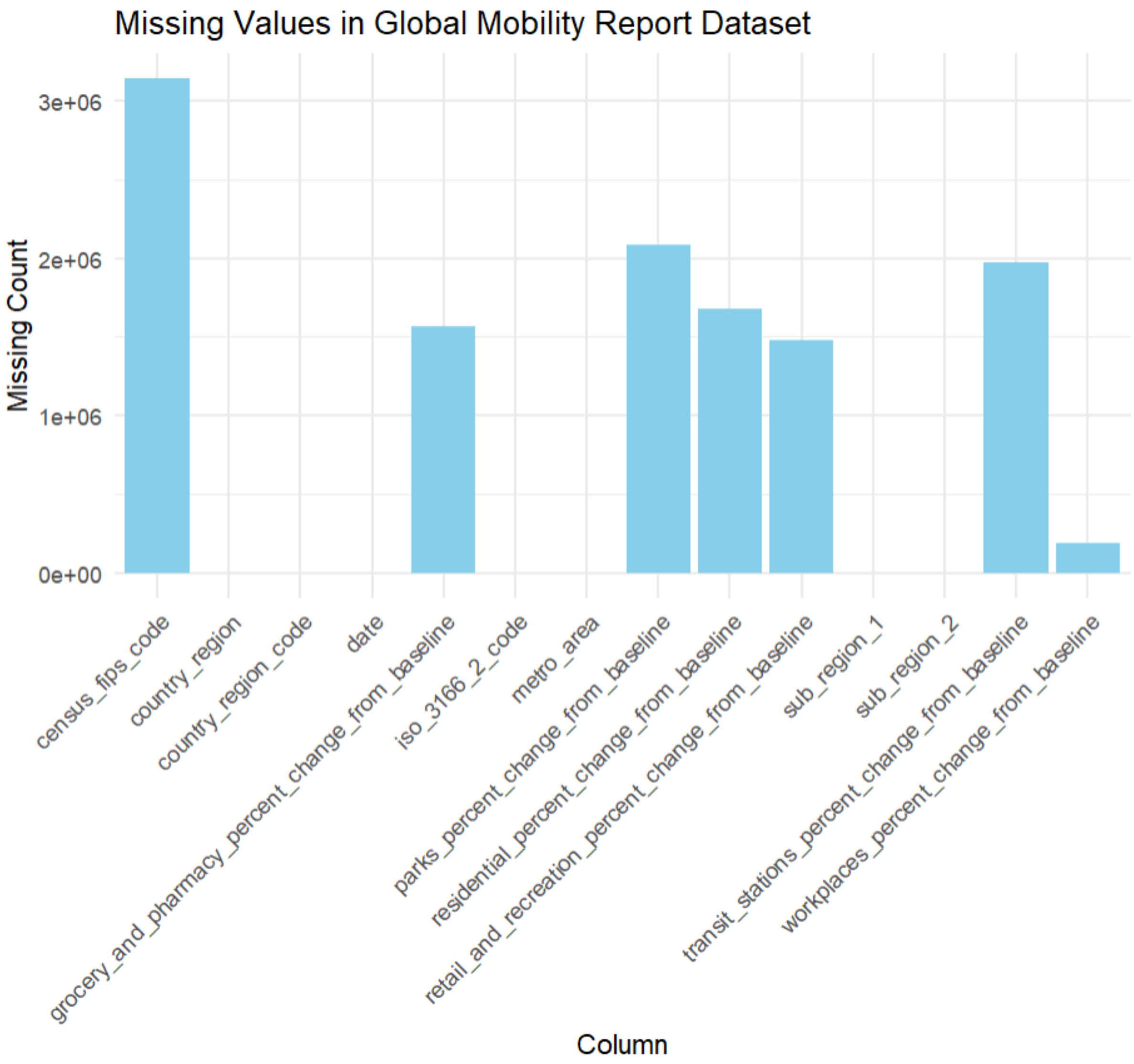


*Figure 9: Missing values in 'Global Mobility Report' dataset.*

After imputing missing values using imputation techniques, such as median imputation, we were able to successfully address the missing values in our dataset. Specifically, we applied imputation to the features with missing values as follows:

1. For the feature `country_region_code`, which initially had 2544 missing values, we applied imputation and filled in these missing values using an appropriate imputation strategy.

2. For the features `census_fips_code`, `retail_and_recreation_percent_change_from_baseline`, `grocery_and_pharmacy_percent_change_from_baseline`, `parks_percent_change_from_baseline`, `transit_stations_percent_change_from_baseline`, `workplaces_percent_change_from_baseline`, and `residential_percent_change_from_baseline`, which had varying numbers of missing values ranging from 189760 to 3139208, we also applied imputation techniques to fill in these missing values.

After performing imputation, we confirmed that all features in the dataset no longer contain any missing values, as indicated by the counts of missing values being reduced to zero for all features.

Overall, imputation allowed us to handle the missing values effectively, ensuring that our dataset is complete and ready for further analysis and modelling in R.

# STATISTICAL SUMARY OF DATASETS

A statistical summary of a dataset typically includes descriptive statistics that provide an overview of the data's central tendency, dispersion, and shape. Common statistical measures included in a summary are:

1. Mean: The average value of the data points.
2. Median: The middle value of the dataset when it is sorted in ascending order.
3. Mode: The value that appears most frequently in the dataset.
4. Range: The difference between the maximum and minimum values in the dataset.
5. Standard Deviation: A measure of the dispersion of data points around the mean.
6. Variance: The average of the squared differences from the mean.
7. Interquartile Range (IQR): The range between the first quartile (25th percentile) and the third quartile (75th percentile).
8. Min: Minimum value
9. Max: Maximum value

Additionally, graphical summaries like histograms, box plots, and scatter plots can provide visual insights into the distribution and relationships within the dataset.

These statistical summaries will help understand the characteristics of the data, identify outliers, assess data quality, and make informed decisions about further analysis or modelling.

## Statistical Summary of Covid-19 Cases Plus Census Dataset

This summary provides key statistical measures for various demographic and economic features across different counties. The data includes measures such as minimum, 1st quartile, median, mean, 3rd quartile, and maximum values for features like confirmed COVID-19 cases, population demographics (e.g., age, race), housing characteristics (e.g., median rent, owner-occupied housing units), and income distribution. These statistics offer insights into the spread of COVID-19, demographic compositions, housing conditions, and income levels within the analysed counties. Such information is valuable for understanding the socioeconomic landscape and potential disparities across different regions, aiding in targeted policymaking and resource allocation efforts.

*Table 7: Statistical Summary for 'COVID-19 cases plus census' DATASET before data normalization*

| Feature | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| high_school_diploma | 16 | 2321 | 5249 | 16071.52 | 12764 | 1293011 |
| high_school_including_ged | 20 | 2754 | 6371 | 18807.64 | 15222 | 1408905 |
| households_public_asst_or_food_stamps | 0 | 574.2 | 1480.5 | 5032.4 | 3635 | 333729 |
| no_cars | 0 | 234.2 | 613 | 3331.77 | 1595.8 | 583560 |
| male_45_64_high_school | 6 | 614.2 | 1405.5 | 3839.38 | 3287 | 259464 |
| income_10000_14999 | 0 | 266.2 | 625 | 1835.81 | 1452.5 | 178737 |
| deaths | 0 | 12 | 31 | 124.83 | 77 | 13936 |
| confirmed_cases | 0 | 796.2 | 1916.5 | 7558.91 | 4955 | 1002614 |
| commute_45_59_mins | 0 | 254 | 696 | 3648.76 | 1900 | 470600 |
| commute_60_more_mins | 0 | 282 | 743 | 4003.56 | 1890 | 603055 |
| rent_burden_not_computed | 0 | 184 | 383 | 1016.4 | 853 | 85384 |
| dwellings_3_to_4_units | 0 | 95 | 283 | 1893.78 | 994 | 236268 |
| commute_60_89_mins | 0 | 168 | 471 | 2775.14 | 1224 | 443923 |
| rent_over_50_percent | 0 | 162 | 486 | 3237.09 | 1576 | 536832 |
| income_less_10000 | 0 | 321 | 763.5 | 2527.77 | 1887.8 | 201863 |
| female_80_to_84 | 0 | 141.2 | 321 | 1094.33 | 801.8 | 95940 |
| employed_education_health_social | 2 | 1035 | 2448 | 11069.81 | 7092 | 989093 |
| children_in_single_female_hh | 0 | 528 | 1389 | 5900.33 | 3707 | 582056 |
| income_15000_19999 | 0 | 260 | 616 | 1822.78 | 1427 | 156089 |
| commute_90_more_mins | 0 | 95 | 255 | 1228.42 | 660.5 | 159132 |
| housing_built_1939_or_earlier | 0 | 261 | 552.5 | 2197.14 | 1316.8 | 374655 |
| female_85_and_over | 0 | 156 | 352.5 | 1283.26 | 873.5 | 113668 |
| no_car | 0 | 82 | 238.5 | 2042.58 | 684.8 | 636105 |
| dwellings_50_or_more_units | 0 | 3 | 59 | 2251.53 | 381 | 483779 |
| aggregate_travel_time_to_work | 3040 | 109998 | 263630 | 1198322.5 | 708848 | 136962170 |
| commute_35_44_mins | 0 | 193.2 | 546.5 | 3066.28 | 1597.8 | 348097 |
| state | 1 | - | - | 27.01 | - | 51 |
| total_pop | 74 | 10945 | 25691 | 102165.63 | 67445 | 10105722 |
| male_pop | 39 | 5514 | 12798 | 50292.41 | 33481 | 4979641 |
| female_pop | 35 | 5460 | 12885 | 51873.22 | 34108 | 5126081 |
| median_age | 21.6 | 37.9 | 41.2 | 41.15 | 44.2 | 66.4 |
| white_pop | 18 | 8093 | 20205 | 62787.33 | 53500 | 2676982 |
| black_pop | 0 | 95 | 758 | 12554.26 | 5396 | 1226134 |

| | | | | | | |
|---|---|---|---|---|---|---|
| hispanic_pop | 0 | 323 | 1025 | 17985.56 | 4868 | 4893579 |
| amerindian_pop | 0 | 24 | 95.5 | 668.01 | 348 | 64102 |
| other_race_pop | 0 | 0 | 10 | 227.67 | 73 | 46396 |
| gini_index | 0.3271 | 0.4211 | 0.4423 | 0.4448 | 0.4665 | 0.5976 |

## Statistical Summary of COVID-19_cases_TX Dataset

The provided summary presents statistical measures for various features at the county level, including the county FIPS code, confirmed COVID-19 cases, deaths, and state FIPS code. The county FIPS code ranges from 0 to 48507, with a median value of 48253. For confirmed COVID-19 cases, the dataset shows a wide range from 0 to 297629, with a median value of 82. Similarly, the number of deaths ranges from 0 to 4024, with a median value of 2. The state FIPS code is consistent throughout the dataset, with all values equal to 48. These metrics offer insights into the distribution and characteristics of COVID-19 cases and related data across different counties within the state.

*Table 8: Statistical Summary of 'COVID-19_cases_TX' Dataset before data normalization*

| Feature | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|---|
| county_fips_code | 0 | 48125 | 48253 | 48065 | 48381 | 48507 |
| county_name | - | - | - | - | - | - |
| state | - | - | - | - | - | - |
| confirmed_cases | 0 | 1 | 82 | 2158.6 | 639.8 | 297629 |
| deaths | 0 | 0 | 2 | 38.19 | 15 | 4024 |
| state_fips_code | 48 | 48 | 48 | 48 | 48 | 48 |
| date | - | - | - | - | - | - |

## Statistical Summary of Global_Mobility_Report

This table summarizes various features related to COVID-19 mobility trends, including changes in retail and recreation, grocery and pharmacy visits, park visits, transit station usage, workplace attendance, and residential movement, across different regions. The statistics include minimum, 1st quartile, median, mean, 3rd quartile, and maximum values for each feature.

*Table 9: Statistical Summary of 'Global_Mobility_Report' dataset before data normalization*

| Feature | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|---|
| country_region_code | - | - | - | - | - | - |
| country_region | - | - | - | - | - | - |
| sub_region_1 | - | - | - | - | - | - |
| sub_region_2 | - | - | - | - | - | - |
| metro_area | - | - | - | - | - | - |
| iso_3166_2_code | - | - | - | - | - | - |

| census_fips_code | 1001 | 29115 | 29115 | 29380 | 29115 | 56045 |
|---|---|---|---|---|---|---|
| date | - | - | - | - | - | - |
| retail_and_recreation_percent_change_from_baseline | -100 | -27 | -19 | -21.67 | -13 | 545 |
| grocery_and_pharmacy_percent_change_from_baseline | -100 | -5 | -2 | -2.613 | 2 | 615 |
| parks_percent_change_from_baseline | -100 | -17 | -17 | -13.43 | -17 | 1206 |
| transit_stations_percent_change_from_baseline | -100 | -29 | -28 | -27.6 | -28 | 554 |
| workplaces_percent_change_from_baseline | -100 | -31 | -19 | -20.02 | -6 | 260 |
| residential_percent_change_from_baseline | -46 | 7 | 8 | 8.798 | 10 | 65 |

# DATA VISUALISATION

## COVID-19 CENSUS PLUS CASES DATASET

The analysis begins by selecting specific features of interest from the dataset, including demographic attributes like education levels, household characteristics, and commuting patterns, along with COVID-19 metrics such as confirmed cases and deaths. This curated dataset is then examined for missing values, and any incomplete rows are removed to ensure data integrity.

Below heatmap of a correlation matrix is used to determine the degree to which different variables in a dataset are linearly related. Here are some key takeaways from this chart:

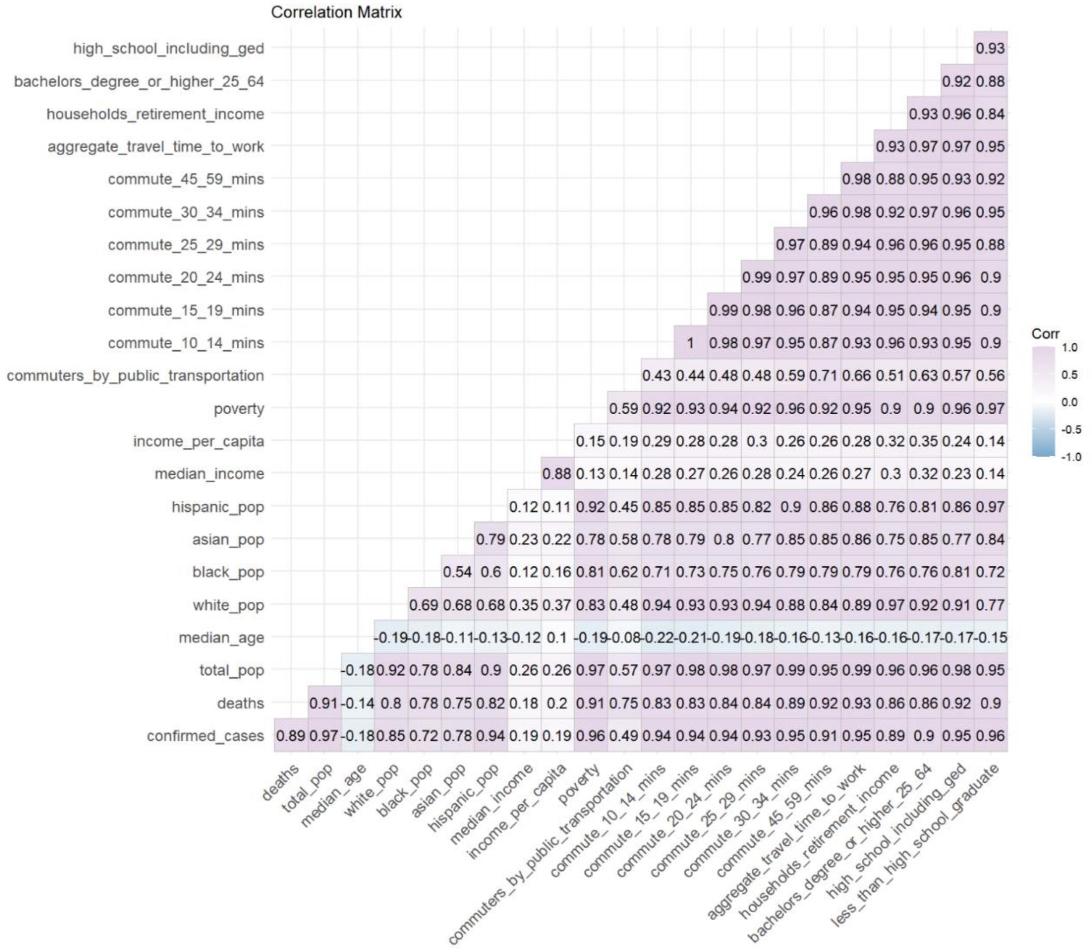
There's a very strong positive correlation between the total population of a state and both the number of confirmed COVID-19 cases and deaths. This suggests that as the population increases, the number of cases and deaths tends to increase as well, which is logical given that a higher population would typically lead to more cases and deaths simply due to more people being exposed.

Certain demographics show strong correlations with each other. For example, the white population is highly correlated with the black population and the Asian population, suggesting that in areas where there are more white people, there are also likely to be more black and Asian individuals. This could be indicative of the diversity in those areas or other socio-economic factors.

Various economic Factors like Median income, income per capita, and poverty have notable correlations with demographic variables like the different ethnic populations. This might indicate that economic factors are unevenly distributed among different ethnic groups, a topic that often garners significant attention in social studies.

There is a strong correlation between higher education levels and commute times, particularly for commutes longer than 30 minutes. This could suggest that higher-educated individuals may be traveling further to work, potentially due to job types or locations that require higher education.

The aggregate travel time to work shows a correlation with the use of public transportation, which may imply that in areas where public transportation is more widely used, the overall travel time for commutes is higher. This could be due to the nature of public transportation systems and their reach, or it might reflect traffic conditions in more densely populated areas.



| | deaths | total_pop | median_age | white_pop | black_pop | asian_pop | hispanic_pop | median_income | income_per_capita | poverty | commuters_by_public_transportation | commute_10_14_mins | commute_15_19_mins | commute_20_24_mins | commute_25_29_mins | commute_30_34_mins | commute_45_59_mins | aggregate_travel_time_to_work | households_retirement_income | bachelors_degree_or_higher_25_64 | high_school_including_ged | less_than_high_school_graduate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| high_school_including_ged | | | | | | | | | | | | | | | | | | | | | | 0.93 |
| bachelors_degree_or_higher_25_64 | | | | | | | | | | | | | | | | | | | | | 0.92 | 0.88 |
| households_retirement_income | | | | | | | | | | | | | | | | | | | | 0.93 | 0.96 | 0.84 |
| aggregate_travel_time_to_work | | | | | | | | | | | | | | | | | | | 0.93 | 0.97 | 0.97 | 0.95 |
| commute_45_59_mins | | | | | | | | | | | | | | | | | | 0.98 | 0.88 | 0.95 | 0.93 | 0.92 |
| commute_30_34_mins | | | | | | | | | | | | | | | | | 0.96 | 0.98 | 0.92 | 0.97 | 0.96 | 0.95 |
| commute_25_29_mins | | | | | | | | | | | | | | | | 0.97 | 0.89 | 0.94 | 0.96 | 0.96 | 0.95 | 0.88 |
| commute_20_24_mins | | | | | | | | | | | | | | | 0.99 | 0.97 | 0.89 | 0.95 | 0.95 | 0.95 | 0.96 | 0.9 |
| commute_15_19_mins | | | | | | | | | | | | | | 0.99 | 0.98 | 0.96 | 0.87 | 0.94 | 0.95 | 0.94 | 0.95 | 0.9 |
| commute_10_14_mins | | | | | | | | | | | | | 1 | 0.98 | 0.97 | 0.95 | 0.87 | 0.93 | 0.96 | 0.93 | 0.95 | 0.9 |
| commuters_by_public_transportation | | | | | | | | | | | | 0.43 | 0.44 | 0.48 | 0.48 | 0.59 | 0.71 | 0.66 | 0.51 | 0.63 | 0.57 | 0.56 |
| poverty | | | | | | | | | | | 0.59 | 0.92 | 0.93 | 0.94 | 0.92 | 0.96 | 0.92 | 0.95 | 0.9 | 0.9 | 0.96 | 0.97 |
| income_per_capita | | | | | | | | | | 0.15 | 0.19 | 0.29 | 0.28 | 0.28 | 0.3 | 0.26 | 0.26 | 0.28 | 0.32 | 0.35 | 0.24 | 0.14 |
| median_income | | | | | | | | | 0.88 | 0.13 | 0.14 | 0.28 | 0.27 | 0.26 | 0.28 | 0.24 | 0.26 | 0.27 | 0.3 | 0.32 | 0.23 | 0.14 |
| hispanic_pop | | | | | | | | 0.12 | 0.11 | 0.92 | 0.45 | 0.85 | 0.85 | 0.85 | 0.82 | 0.9 | 0.86 | 0.88 | 0.76 | 0.81 | 0.86 | 0.97 |
| asian_pop | | | | | | | 0.79 | 0.23 | 0.22 | 0.78 | 0.58 | 0.78 | 0.79 | 0.8 | 0.77 | 0.85 | 0.85 | 0.86 | 0.75 | 0.85 | 0.77 | 0.84 |
| black_pop | | | | | | 0.54 | 0.6 | 0.12 | 0.16 | 0.81 | 0.62 | 0.71 | 0.73 | 0.75 | 0.76 | 0.79 | 0.79 | 0.79 | 0.76 | 0.76 | 0.81 | 0.72 |
| white_pop | | | | | 0.69 | 0.68 | 0.68 | 0.35 | 0.37 | 0.83 | 0.48 | 0.94 | 0.93 | 0.93 | 0.94 | 0.88 | 0.84 | 0.89 | 0.97 | 0.92 | 0.91 | 0.77 |
| median_age | | | | -0.19 | -0.18 | -0.11 | -0.13 | -0.12 | 0.1 | -0.19 | -0.08 | -0.22 | -0.21 | -0.19 | -0.18 | -0.16 | -0.13 | -0.16 | -0.16 | -0.17 | -0.17 | -0.15 |
| total_pop | | | -0.18 | 0.92 | 0.78 | 0.84 | 0.9 | 0.26 | 0.26 | 0.97 | 0.57 | 0.97 | 0.98 | 0.98 | 0.97 | 0.99 | 0.95 | 0.99 | 0.96 | 0.96 | 0.98 | 0.95 |
| deaths | | 0.91 | -0.14 | 0.8 | 0.78 | 0.75 | 0.82 | 0.18 | 0.2 | 0.91 | 0.75 | 0.83 | 0.83 | 0.84 | 0.84 | 0.89 | 0.92 | 0.93 | 0.86 | 0.86 | 0.92 | 0.9 |
| confirmed_cases | 0.89 | 0.97 | -0.18 | 0.85 | 0.72 | 0.78 | 0.94 | 0.19 | 0.19 | 0.96 | 0.49 | 0.94 | 0.94 | 0.94 | 0.93 | 0.95 | 0.91 | 0.95 | 0.89 | 0.9 | 0.95 | 0.96 |

*Figure 10: Correlation Matrix of COVID-19 CENSUS PLUS CASES Data Features*

The output is a graph representing COVID-19 cases per 1000 people in California. Each county in California is depicted on the map, with colors indicating the density of COVID-19 cases per 1000 individuals within that county. Counties with higher case rates are represented by warmer colors such as red, while those with lower case rates are depicted in cooler colors like yellow. This visualization provides a spatial understanding of the distribution of COVID-19 cases across different regions within California, highlighting areas with higher infection rates. The graph aids in identifying COVID-19 hotspots and assessing the geographical spread of the virus within the state. It enables policymakers, healthcare professionals, and the public to make informed decisions regarding resource allocation, public health measures, and targeted interventions to mitigate the spread of the virus in California.

This visual representation of COVID-19 cases per 1000 people in California serves as a crucial tool for understanding the impact of the pandemic at the county level. By identifying areas with higher infection rates through color gradations, stakeholders can prioritize resource allocation and implement targeted interventions where they are most needed. Additionally, this graph facilitates ongoing monitoring and evaluation of the effectiveness of public health measures and vaccination efforts across different regions of California. By continuously analyzing and updating this data, authorities can adapt their strategies to address emerging trends and effectively combat the spread of COVID-19, ultimately safeguarding the health and well-being of communities statewide.

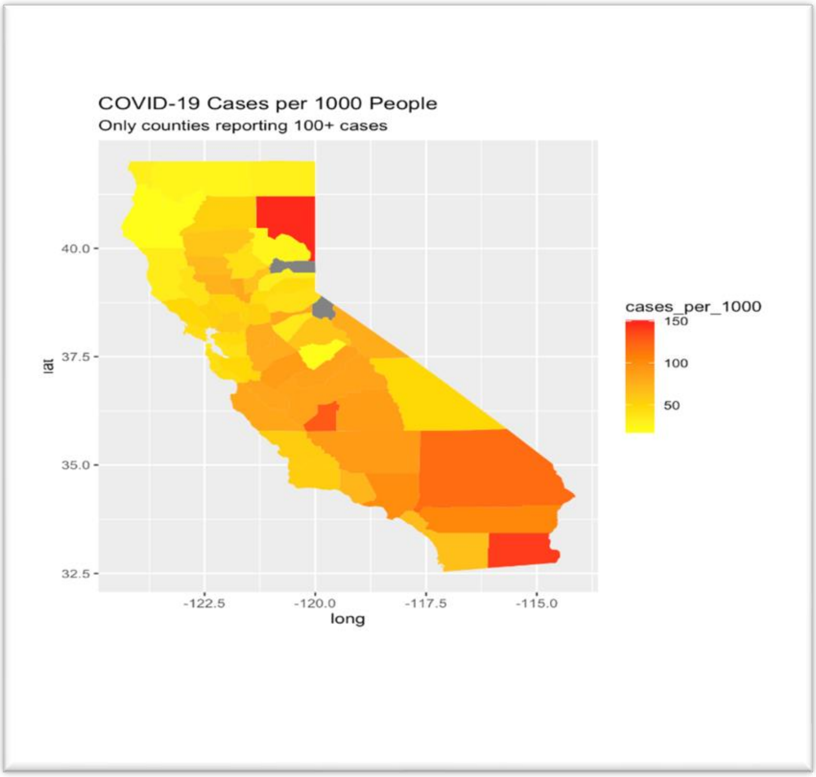


*Figure 11: COVID-19 cases per 1000 People.*

The scatter plot below illustrates a comparison between total population and confirmed COVID-19 cases. There exists a positive correlation between total population and the number of confirmed cases. There are numerous data points densely clustered near the origin, indicating a portion of observations with low values for both population and confirmed cases. As we move to the right along the x-axis (higher population values), the data points become sparser, indicating fewer counties or areas with very high populations. The analysis from Figure 11 leads to the inference that as the total population increases, confirmed cases also rise concurrently.

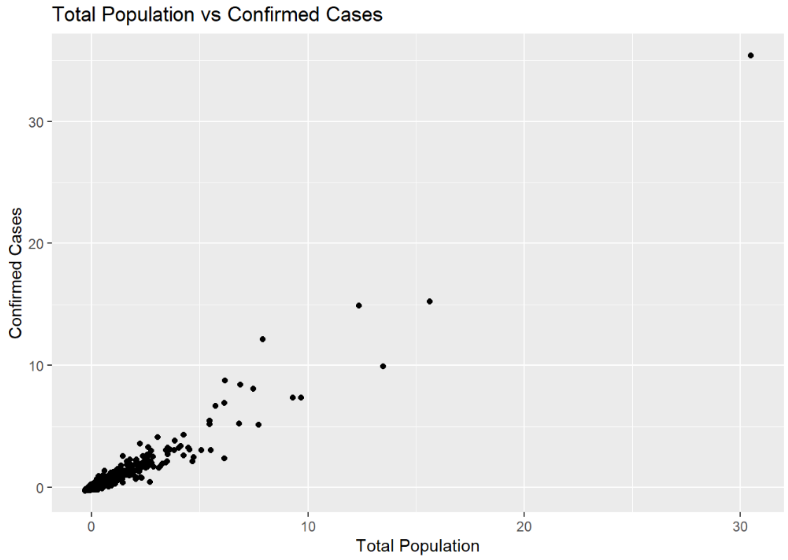


*Figure 12: Scatter plot for Total Population vs Confirmed Cases*

In the scatter plot below, a positive correlation between total population and the number of deaths is evident. This correlation aligns with expectations, as larger populations are to experience more cases and subsequently, more deaths. There is a significant concentration of data points near the origin suggests that most of the observed entities, potentially counties or regions, possess relatively small populations and consequently fewer deaths.

Analysing the below scatter plot (Figure 13), it can be inferred that the increase in population might strain the availability of necessary medical treatment and facilities for effective COVID-19 management, consequently leading to a rise in the number of deaths.

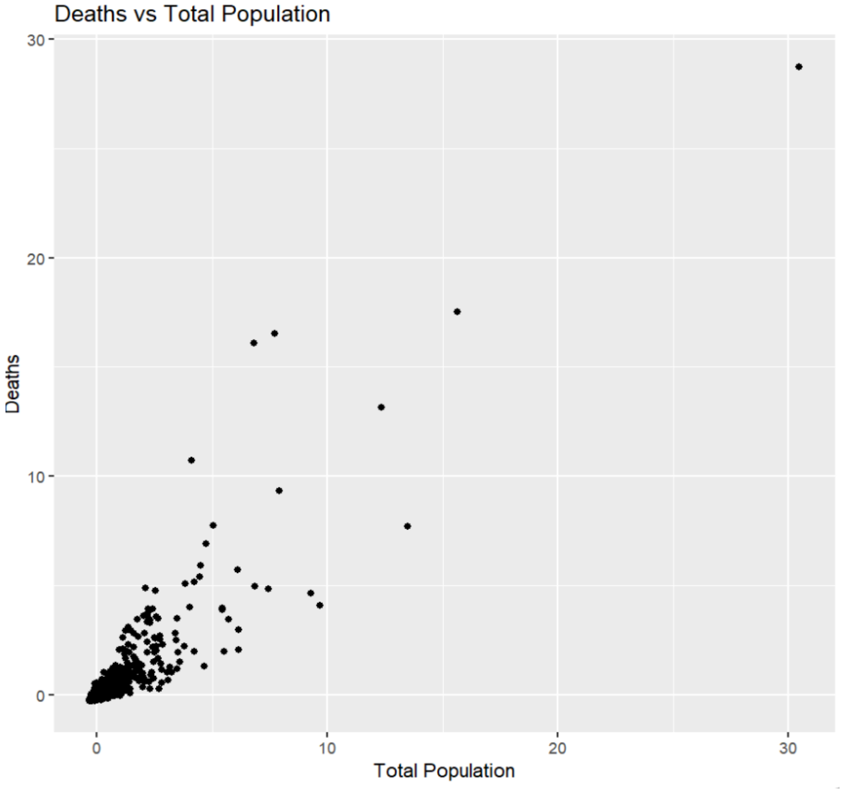


*Figure 13: Scatter Plot of Death vs Total Population*

The pie chart illustrates deaths attributed to COVID-19 across various demographic features. It's evident from the chart that Hispanic population experienced the highest number of deaths due to COVID-19, whereas the Amerindian population faced the lowest number of deaths caused by the virus.

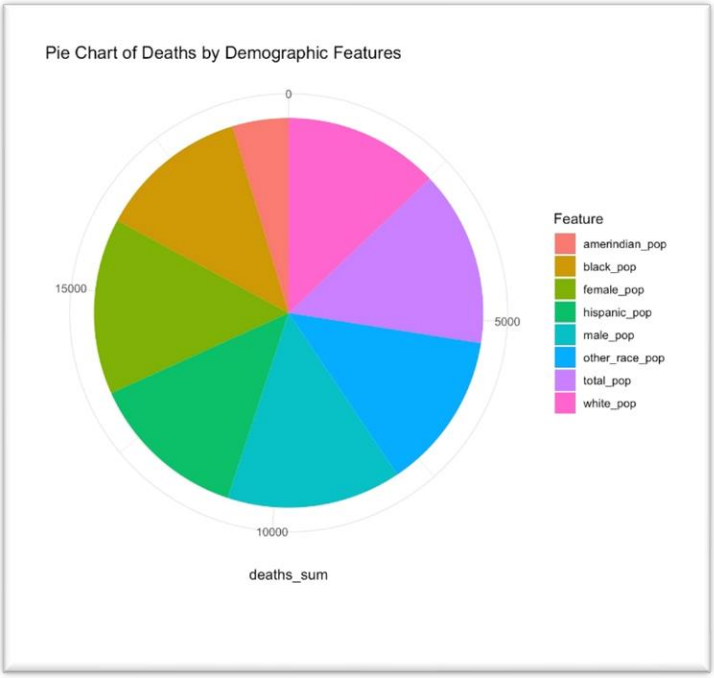


*Figure 14: Pie chart of deaths by demographic features.*

# COVID-19_cases_TX

In the 'confirmed cases vs deaths' curve graph, a positive correlation between confirmed cases and deaths is observed, indicating that as the number of confirmed cases increases, so does the number of deaths. Despite a similar number of confirmed cases, there exists a wide variance in the number of deaths. This discrepancy can be attributed to various factors, including disparities in healthcare quality, differences in reporting accuracy, or the prevalence of comorbid conditions among the patient population. As confirmed cases increase, there is a tendency for the number of deaths to plateau towards the right end of the graph. This levelling off suggests a potential saturation point, wherein the death rate does not increase at the same pace as infections. Possible explanations for this phenomenon include advancements in treatment methods, the development of herd immunity, or other contributing factors.

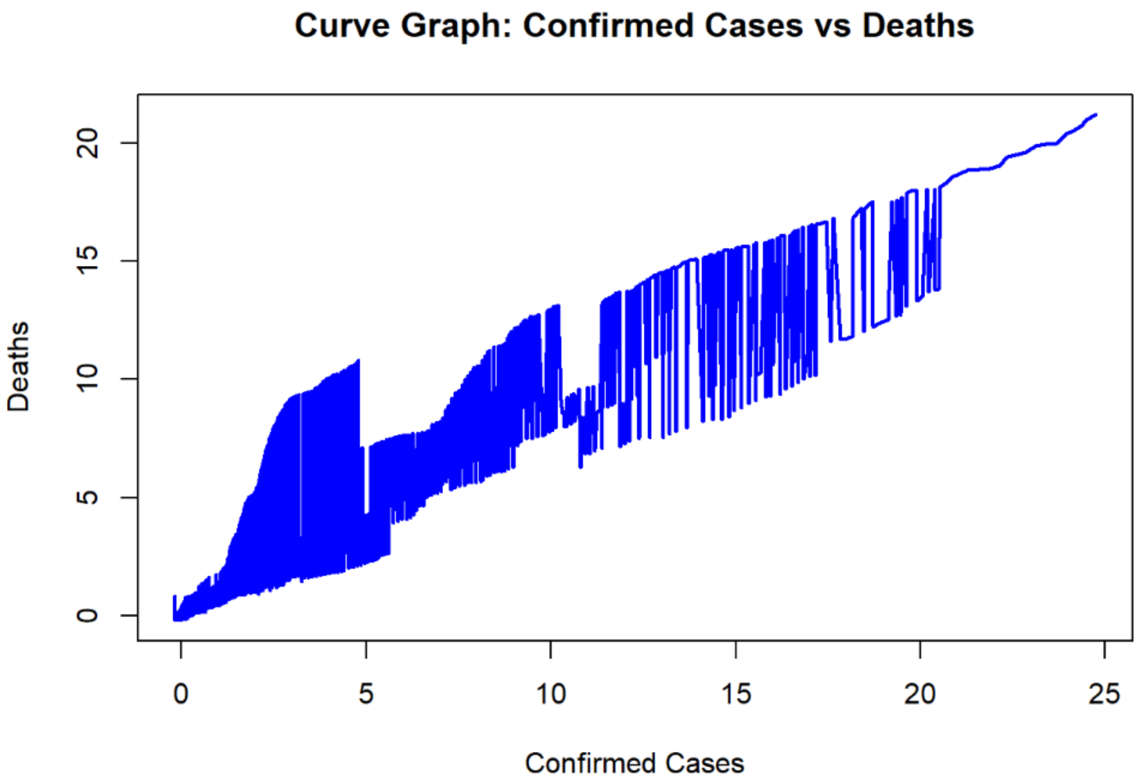


*Figure 15: Curve graph: confirmed cases vs deaths for 'COVID-19_cases_TX' dataset.*

The below time series plot shows the number of confirmed COVID-19 cases over time, from around April 2020 to January 2021. Examining the time series plot below reveals a distinct upward trend in the number of confirmed cases throughout the observed period. This trend indicates a continual rise in COVID-19 infections within Texas during these months. Notably, the slope of the curve becomes notably steeper as time progresses, particularly noticeable starting from around October 2020. This steepening suggests an acceleration in the rate of confirmed cases, possibly indicating the onset of a second wave or surge in infections.

By the conclusion of the observed period in January 2021, the plot demonstrates no signs of plateauing, indicating that the number of confirmed cases was still on the rise at this juncture.

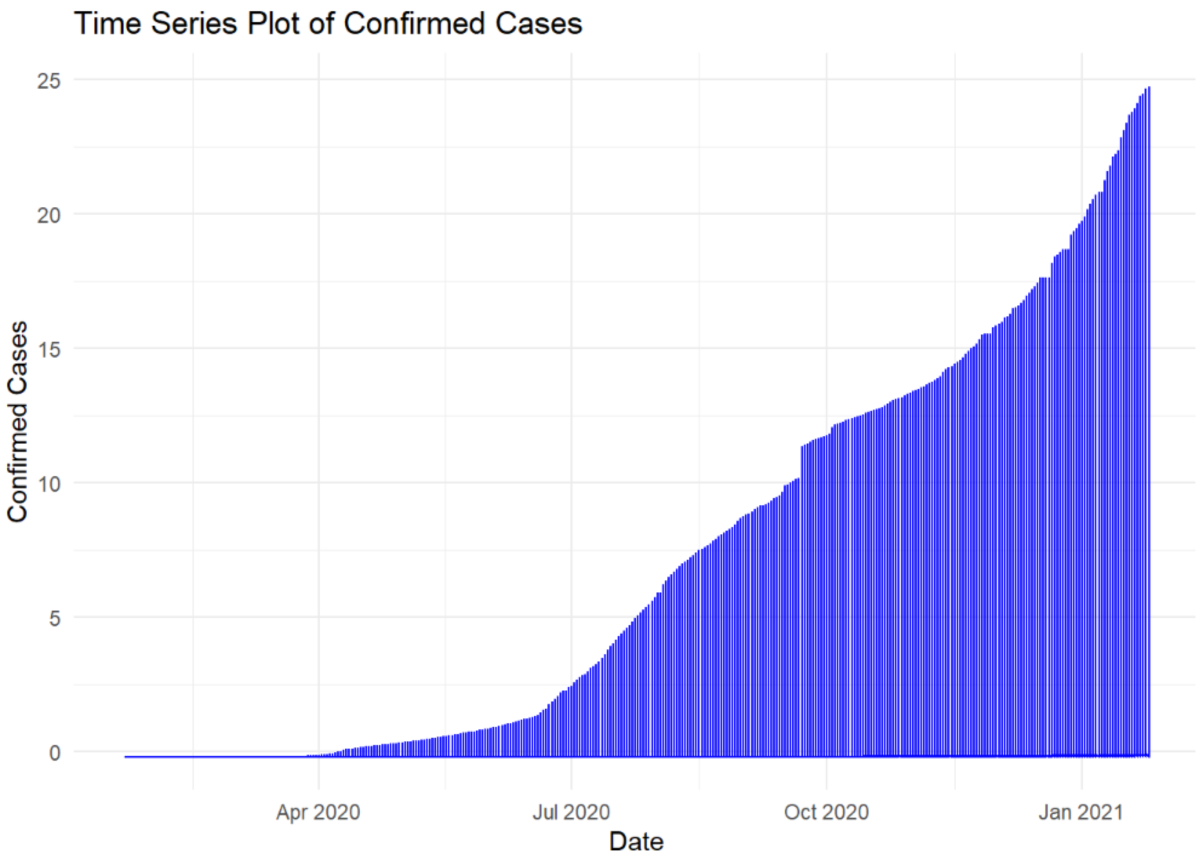


*Figure 16: Time Series plot of Confirmed cases over time for 'COVID_19_cases_TX' dataset.*

The time series plot below shows the number of deaths over time, likely associated with COVID-19, from around April 2020 to January 2021. Over this time, the number of deaths keeps going up, especially as more cases of COVID-19 are confirmed. The line on the graph looks like it's getting steeper, especially from the middle of 2020 to January 2021. This means that the rate of deaths was speeding up during these months.

If we compare figure 17 which figure 16, we notice that the deaths tend to increase after the confirmed cases start going up. This makes sense because there's usually a delay between when someone gets confirmed with COVID-19 and when they might pass away. The rise in deaths shows how serious the COVID-19 pandemic is and highlights the ongoing struggle for public health efforts to control the spread and reduce the number of deaths.

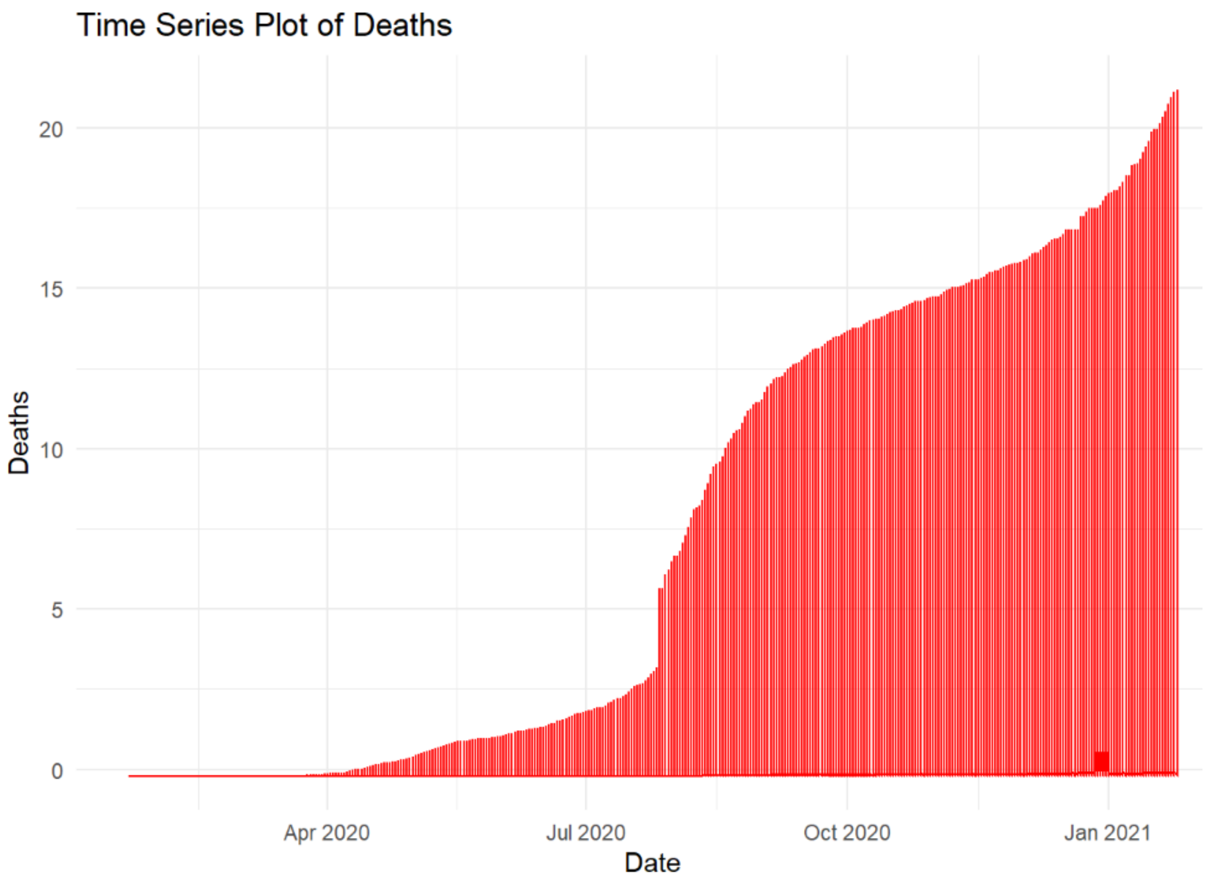


*Figure 17: Time Series plot of deaths based on 'COVID-19_cases_TX' dataset.*

# Global Mobility Report

The below bar graph showing the percentage changes from baseline over time for various activities. The dataset used contains information about COVID-19 mobility data, including metrics such as retail and recreation, grocery and pharmacy visits, park visits, transit station usage, workplace attendance, and residential activity levels. 'Workplaces' have the highest positive change from the baseline, followed by 'Transit Stations', 'Retail & Recreation', 'Residential', 'Parks', and 'Grocery & Pharmacy' having the least positive change among the displayed categories.

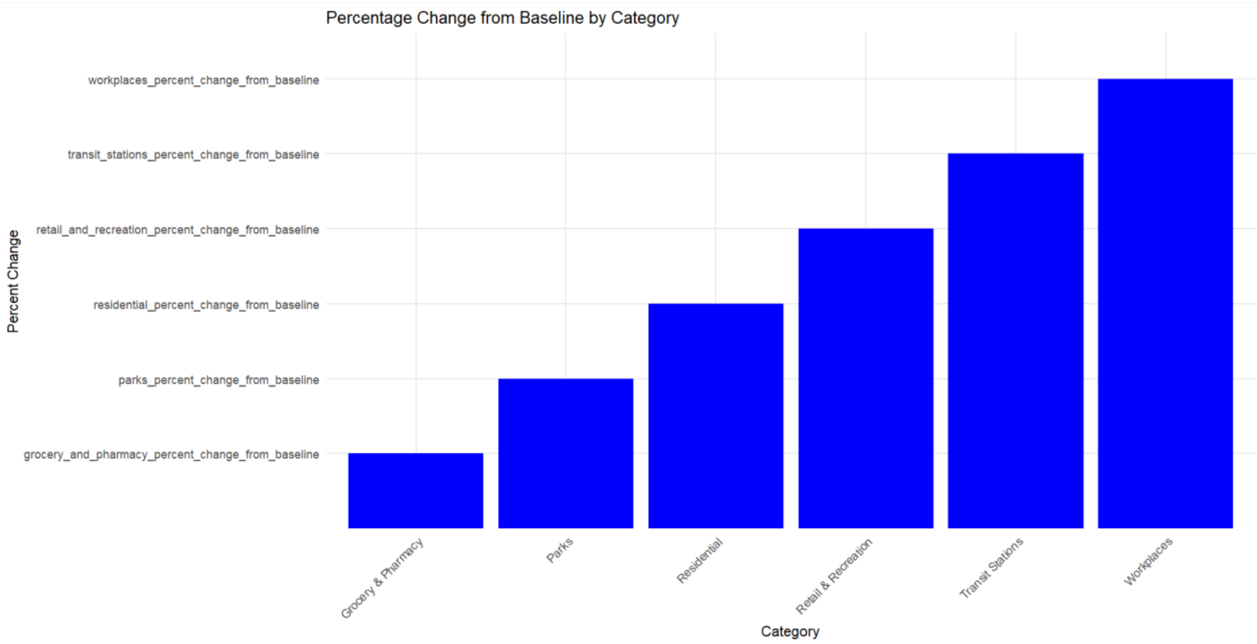


*Figure 18: Percentage changes from baseline over time.*

# COMBINED DATASETS

The combined dataset comprises crucial features pertaining to various states, offering insights into their demographic and socioeconomic landscapes. The "State" column serves as a nominal identifier for each state, facilitating easy reference and analysis. Key demographic indicators like "total_pop," "male_pop," and "female_pop" provide a comprehensive overview of population distribution, while "median_age" offers insight into the age structure of each state's residents.

Socioeconomic factors are also well-represented, with columns such as "white_pop," "black_pop," and "hispanic_pop" highlighting the racial composition within each state. The "gini_index" metric quantifies income inequality, crucial for understanding economic disparities. Moreover, detailed information on income brackets, commute times, and COVID-19 statistics like "deaths" and "confirmed_cases" further enriches the dataset, offering a holistic view of each state's sociodemographic profile. This comprehensive dataset empowers researchers and policymakers alike to explore and address various social and economic challenges across different states effectively.

*Table 10: Important features of all 3 datasets*

| Feature | Scale | Description |
|---|---|---|
| State | Nominal | The name or identifier of the state (e.g., California, New York) |
| total_pop | Ratio | The total population of the state. |
| male_pop | Ratio | The male population of the state. |

| | | |
|---|---|---|
| female_pop | Ratio | The female population of the state. |
| median_age | Ratio | The median age of the state's population. |
| white_pop | Ratio | The population count of individuals identifying as white within the state. |
| black_pop | Ratio | The population count of individuals identifying as black within the state. |
| hispanic_pop | Ratio | The population count of individuals identifying as Hispanic within the state. |
| amerindian_pop | Ratio | The population count of individuals identifying as American Indian or Alaska Native within the state. |
| other_race_pop | Ratio | The population count of individuals identifying with a race other than those listed above, or with multiple races, within the state. |
| gini_index | Ratio | A measure of income inequality within the state. A Gini index of 0 represents perfect equality, while a Gini index of 1 represents perfect inequality. |
| income_10000_14999 | Ratio | Number of individuals with income between $10,000 and $14,999 |
| deaths | Ratio | Number of deaths |
| confirmed_cases | Ratio | Number of confirmed COVID-19 cases |
| aggregate_travel_time_to_work | Ratio | Aggregate travel time to work |
| commute_35_44_mins | Ratio | Number of individuals with commute time between 35-44 minutes |
| commute_45_59_mins | Ratio | Number of individuals with commute time between 45-59 minutes |
| commute_60_more_mins | Ratio | Number of individuals with commute time of 60 minutes or more |
| income_less_10000 | Ratio | Number of individuals with income less than $10,000 |

| | | |
|---|---|---|
| income_15000_19999 | Ratio | Number of individuals with income between $15,000 and $19,999 |
| commute_90_more_mins | Ratio | Number of individuals with commute time of 90 minutes or more |

## FIRST 10 ROWS of COMBINED DATASET

The table provides a snapshot of demographic and socio-economic indicators for different states. Each row represents a state, while columns present various attributes such as population counts, median age, racial demographics, income distribution, and commute times. The decision to use numeric column labels was made for brevity due to short feature names. This format allows for a concise overview of key metrics across states, aiding in comparative analysis and understanding of broad trends in population characteristics and socio-economic factors.

*Table 11: First 10 rows of confirmed dataset*

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VT | 6203 | 6203 | 3068 | 50 | 5929 | 64 | 32 | 83 | 20 | 6 | 0.429 | 224 | 0 | 111 | 65350 | 187 | 256 | 164 | 208 | 241 | 109 | 38767 |
| VT | 160985 | 160985 | 82057 | 37 | 143657 | 4091 | 6144 | 3542 | 374 | 240 | 0.454 | 2755 | 78 | 3636 | 1763730 | 4595 | 3539 | 2474 | 3080 | 2825 | 848 | 66906 |
| DE | 173145 | 173145 | 89601 | 37 | 108627 | 41729 | 3459 | 11820 | 967 | 377 | 0.43 | 2763 | 187 | 11548 | 1975260 | 3999 | 5923 | 8007 | 3732 | 2838 | 2396 | 57647 |
| RI | 126190 | 126190 | 65036 | 44 | 115206 | 1621 | 2436 | 3769 | 1047 | 260 | 0.446 | 2057 | 122 | 5521 | 1551355 | 6059 | 5080 | 4019 | 2337 | 1642 | 1328 | 77862 |
| NH | 60383 | 60383 | 30678 | 47 | 57523 | 285 | 579 | 951 | 130 | 65 | 0.428 | 796 | 79 | 2496 | 756695 | 2307 | 2705 | 2287 | 1121 | 1342 | 860 | 65834 |
| RI | 83204 | 83204 | 42252 | 45 | 71549 | 2596 | 1670 | 4623 | 290 | 159 | 0.469 | 1586 | 6 | 3578 | 911115 | 2397 | 2923 | 2429 | 1921 | 1294 | 928 | 75463 |
| VT | 25191 | 25191 | 12687 | 41 | 23895 | 214 | 161 | 408 | 208 | 0 | 0.469 | 599 | 1 | 312 | 309910 | 815 | 1448 | 1152 | 596 | 641 | 128 | 54899 |
| CT | 151596 | 151596 | 75434 | 38 | 129519 | 4425 | 6690 | 7860 | 38 | 336 | 0.428 | 1575 | 125 | 6255 | 1926600 | 7485 | 5898 | 4581 | 2036 | 1567 | 1399 | 81312 |
| VT | 36825 | 36825 | 18611 | 43 | 34245 | 337 | 711 | 795 | 94 | 57 | 0.417 | 646 | 5 | 527 | 416400 | 1480 | 1466 | 1229 | 590 | 523 | 259 | 61875 |
| VT | 30576 | 30576 | 15210 | 44 | 29070 | 272 | 186 | 460 | 65 | 9 | 0.424 | 605 | 4 | 307 | 339665 | 658 | 888 | 1260 | 685 | 669 | 512 | 47371 |

## DATA VISUALISATIONS OF COMBINED DATASET

The visualisation is a scatter plot graph from a data visualization tool, likely generated using ggplot2 in R. The plot is showing the relationship between confirmed COVID-19 cases (on the x-axis) and the number of deaths (on the y-axis) across various counties.

Each point on the plot represents a county, with the size of the point correlating to the total population of the county—larger points indicate more populous counties. The counties are labelled by name, making it easy to identify specific areas. For example, Los Angeles County, being the point farthest along both axes, appears to have the highest number of confirmed cases and deaths, which is consistent with it being one of the most populous counties.

The blue line represents a linear model (linear regression) fit to the data, indicating the trend or relationship between the number of confirmed cases and deaths. The fact that the line has a positive slope suggests that there is a positive correlation between confirmed cases and deaths: as the number of confirmed cases increases, the number of deaths tends to increase as well.

This kind of visualization helps government agencies to quickly grasp the impact of the pandemic on different population centres and to allocate resources accordingly. It also can also be a tool for the public to understand the severity of the pandemic across different regions.

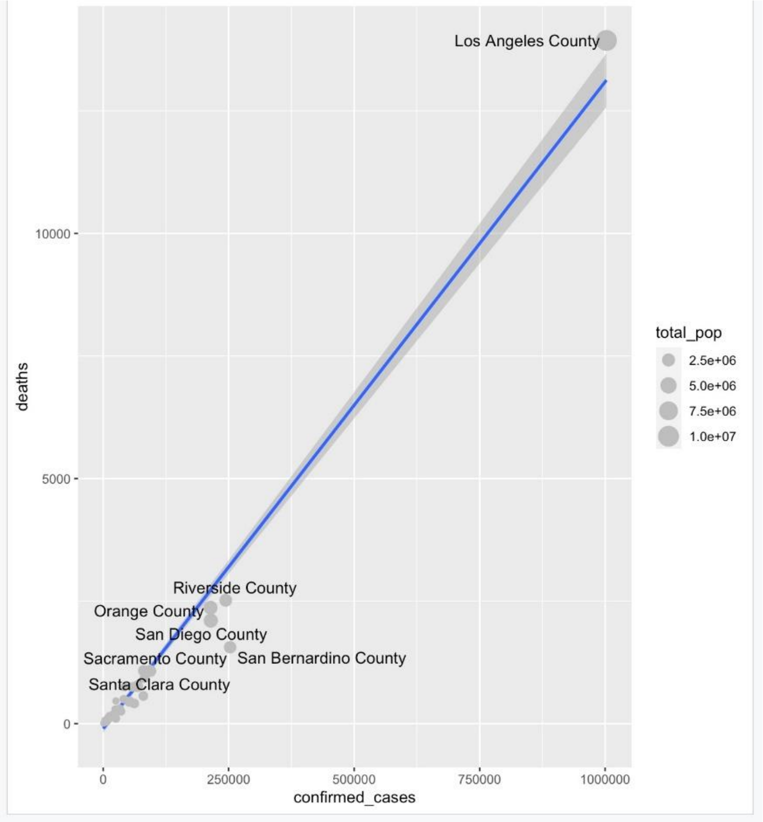


*Figure 19: confirmed cases vs deaths.*

The below plot shows South Dakota and North Dakota having the highest number of Average Deaths per 1000. Average Deaths per 1000 is calculated based on the total population and number of deaths reported in the state, therefore the value could change if number of deaths change in future. However, this increase could be due to older aged people living in the states resulting in more fatalities, or this could also be due to resources issues in medical supplies. Therefore, authorities could plan for providing more medical supplies, implementing strict Covid 19 guidelines and rules and boosting vaccination as a precaution for the next wave.

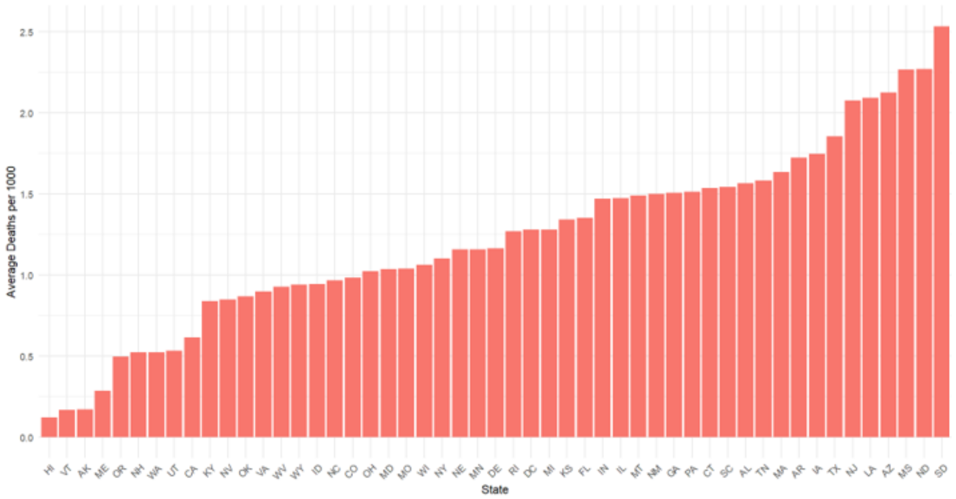


*Figure 20: Average deaths per 1000 by States.*

The plot below depicts the distribution of Gini Index values across various states, offering insights into income inequality. Each peak or bump in the plot represents the Gini Index distribution for a specific state. It's apparent that states exhibit different levels of income inequality, with some clustering around a central range, indicating a common level of inequality. States with peaks to the right tend to have higher Gini Index values, signifying greater income disparity within those states. Conversely, states with peaks to the left typically have lower Gini Index values, indicating less inequality. Additionally, the width of each state's peak reflects the variability of Gini Index values within that state, with wider peaks suggesting more variability and narrower peaks indicating more tightly clustered values.

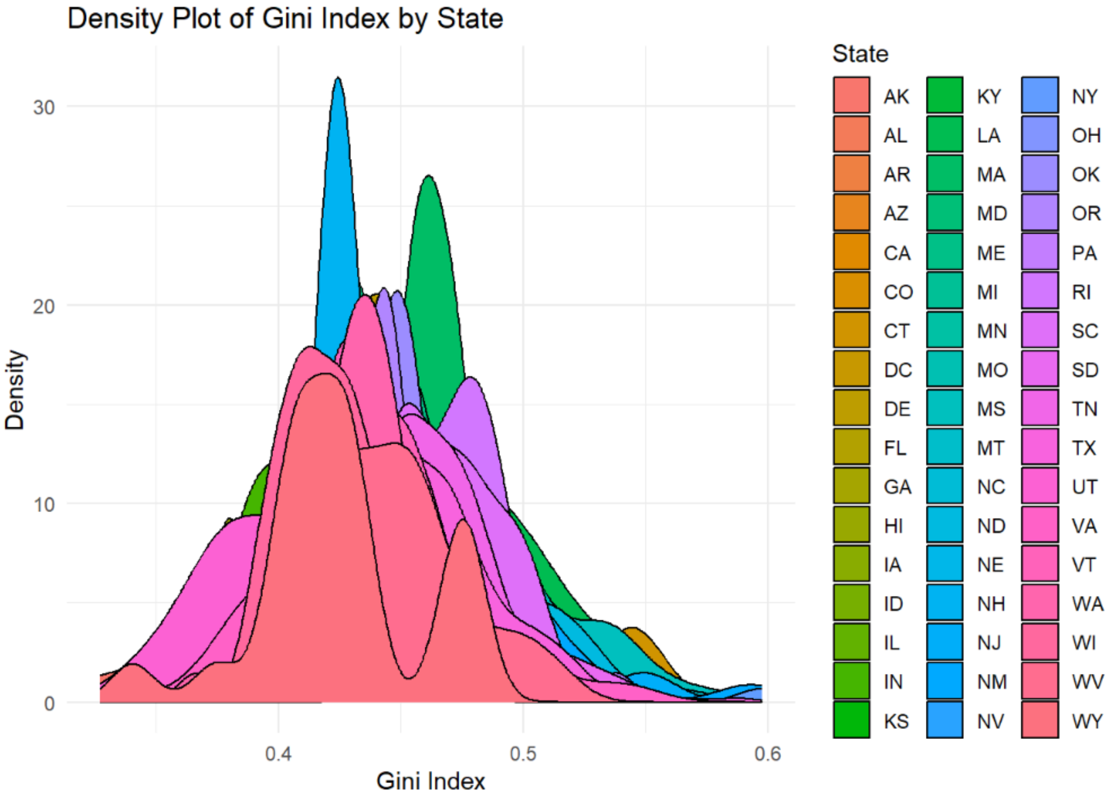


*Figure 21: Density Plot of Gini Index by State*

# How these visualisations done can help government agencies?

Scatter Plots (Total Population vs Confirmed Cases and Deaths vs Total Population): These graphs help to understand the relationship between the population size and the number of confirmed cases and deaths. Agencies can identify if larger populations are correlated with more cases or deaths, which can be critical for resource allocation, such as where to send more medical supplies or where to increase testing and vaccination efforts.

Pie Charts (Demographic Features vs Deaths): These show the breakdown of COVID-19 deaths among different demographic groups. This information is essential for identifying which segments of the population are most affected by the pandemic and may need targeted public health interventions.

Curve Graph (Confirmed Cases vs Deaths): This visualization might be showing the cumulative number of confirmed cases and the corresponding deaths over time. It can be used to assess the lethality of the disease across the confirmed cases and to monitor the progression of the pandemic.

Time Series Plots (Confirmed Cases and Deaths): These plots show how the number of cases and deaths changes over time. They can help in understanding the spread and the peak of the pandemic, and in evaluating the effectiveness of public health measures, like lockdowns or mask mandates.

Bar Graph (Percentage Change from Baseline by Category): This bar chart illustrates the changes in mobility and activity in various categories from a pre-pandemic baseline. It can inform policymakers about the impact of lockdown measures on people's behaviour and the subsequent economic effects.

Density Plot (Gini Index by State): The density plot for the Gini Index, which measures income inequality, can help agencies understand economic disparities between states and target economic relief or support measures where inequality is higher.

# Exceptional Work

## SIGMA-TUKEY PLOT

The histogram showcases the disparities between predictions made by a statistical model and the actual observed values. The bulk of these differences are clustered around zero, indicating the model's generally accurate predictions. However, there's a discernible number of instances where the model forecasted fewer cases than what transpired, evidenced by a rightward tail in the histogram. This implies that the model wasn't flawless and tended to underestimate confirmed cases in certain scenarios. While effective for most predictions, it struggled with accuracy when forecasting higher numbers of COVID-19 cases.

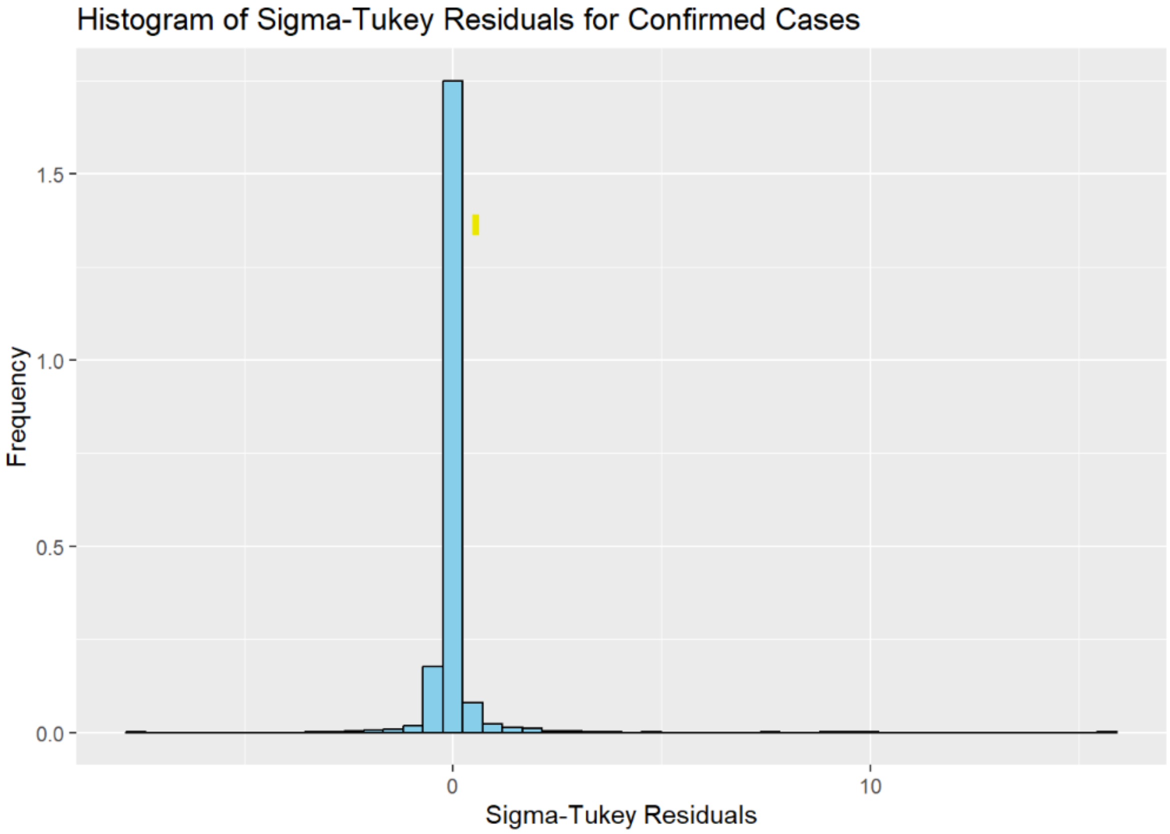


*Figure 22: Histogram of Sigma-Tukey Residuals for Confirmed Cases*

Each point on the biplot represents an observation from the dataset in the reduced dimensionality space defined by the first two principal components. The location of a point is determined by its scores on these two components. The axes of the biplot (Dim1 and Dim2) represent the first and second principal components, respectively. The percentages in the axis labels (e.g., "Dim1 (81.2%)") indicate how much of the variance in the original dataset is captured by that principal component. In this case, the first dimension captures 81.2% of the variance, and the second dimension captures 26.6%. The spread of the points across the plot can give insights into the relationships between observations. Observations that are close to each other in this space have similar profiles in terms of the underlying features that were used in the PCA. Points that are far from the origin or the cluster of other points could be considered outliers. They may represent unique or rare occurrences within the dataset. In summary, the biplot is aiming to visually summarize the PCA by showing how much of the original data's variance is explained by the first two principal components and how the observations relate to each other in this reduced space.

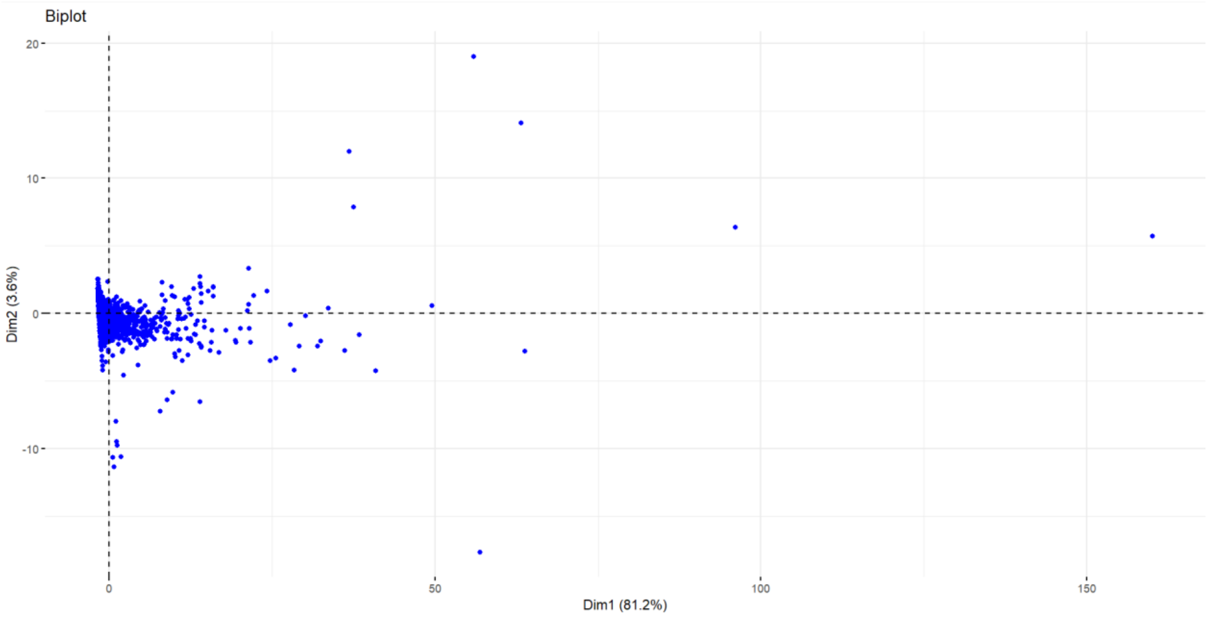


*Figure 23: Biplot.*

PC1 Histogram: The distribution of scores on PC1 is very skewed, with most of the scores concentrated near the origin and a long tail extending to the right. This indicates that most observations have similar low scores on PC1, but there are a few observations with very high scores.

PC2 to PC9 Histograms: These show more normally distributed scores around the origin, which suggests that the remaining principal components capture dimensions of the data that have a more symmetric distribution of the variance.

Each principal component (PC) is a linear combination of all the original features, with coefficients (loadings) that define the contribution of each feature to that component. The first principal component typically captures the largest amount of variance in the data, with each subsequent component capturing progressively less. The percentages of variance captured by each principal component would typically be displayed in a scree plot, which is not present in the image you uploaded. The aim of PCA is often to reduce the dimensionality of a dataset by identifying a small number of principal components that capture most of the variance. In practice, one might choose to focus on the first few principal components for further analysis if they capture a significant portion of the total variance.

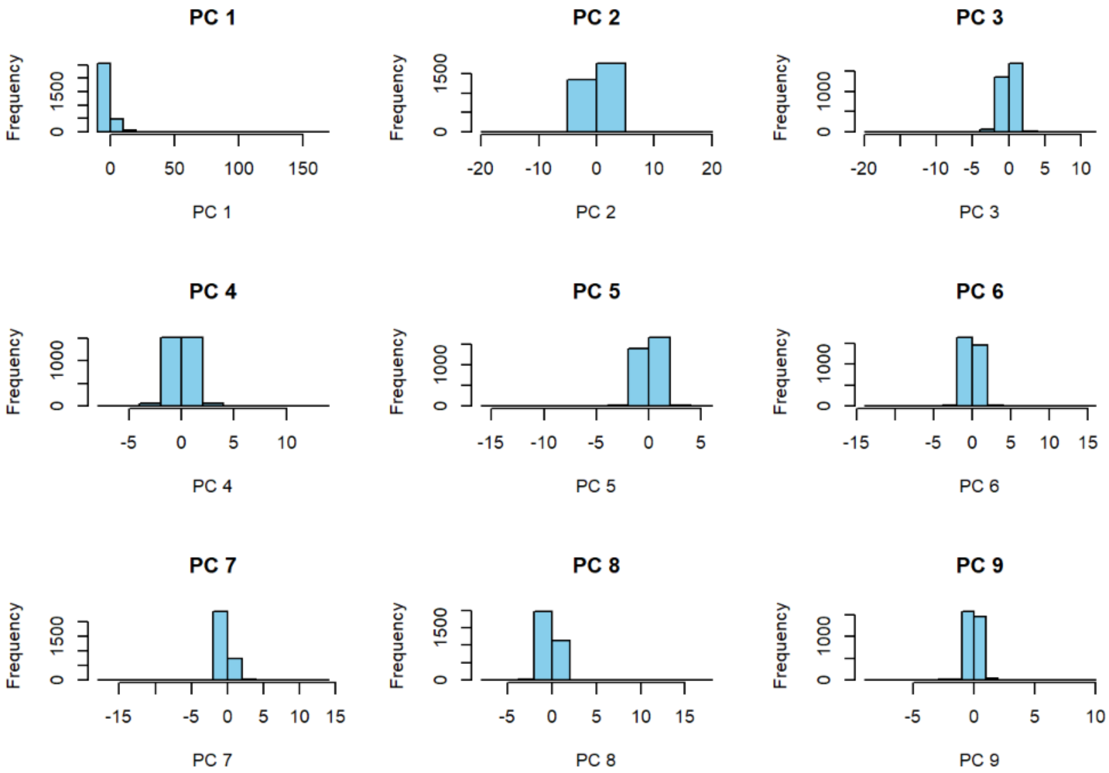


*Figure 24: Histograms for principal component.*

# CONCLUSION

This report is beneficial for government agencies as it offers a comprehensive analysis of COVID-19's impact on California, utilizing datasets that provide insights into confirmed cases, demographic distribution, and mobility trends. The detailed statistical summaries and visualizations help identify regions most affected by the pandemic and demographic groups at higher risk. By understanding these patterns, government agencies can allocate resources more effectively, implement targeted interventions, and develop strategies to combat the virus. Overall, the report aids in informed decision-making to protect public health and support communities during the pandemic.

# DISTRIBUTION OF WORK

| Task | Contribution |
|---|---|
| Document Writing/Formatting | Dhruvil, Juhi, Vishakha |
| Executive Summary | Dhruvil, Juhi |
| Business Understanding | Dhruvil, Vishakha |
| Data Understanding | Dhruvil, Juhi, Vishakha |
| Data Preparation | Dhruvil, Juhi, Vishakha |

| | |
|---|---|
| Data Preparation | Dhruvil, Juhi, Vishakha |
| Data Visualization | Dhruvil, Juhi, Vishakha |
| Exceptional Credit | Dhruvil, Juhi, Vishakha |

All members contributed equally across the project.

# REFERENCES

[1] https://console.cloud.google.com/marketplace/browse?filter=solution-%20type:dataset&filter=category:covid19

[2] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_California

[3] https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Texas

[4] https://smu.instructure.com/courses/119142/assignments/926947

[5] https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/california/

[6] https://www.who.int/health-topics/coronavirus#tab=tab_1