**Data and Visualization**

Team 10: Jeevan Rai & Abhilash Narayanan

**Executive Summary:**

This report is on analyzing various aspects of social structures that were collected during the COVID-19 pandemic. These social structures are referred as variables for the four datasets (COVID-19\_global\_mobility, COVID-19\_cases\_TX, COVID-19\_cases\_plus\_census, COVID Vaccination Report ) that go through various data mining processes to discover any possible relationships with the number of cases and/or deaths for various geographic regions. There could be many possible reasons that could depict the observed confirmed cases and/or deaths for a particular region during the pandemic. Discovering such insightful relationships and building an effective predictive model are the two primary problems described in the report. From the general public to leaders of various government and non-government institutions, the information described in the report can be beneficial in many ways. For example, the report shows that there is a strong correlation between the number of confirmed cases and the people aged 0-20 yrs for New York and California. Based on this observation, the government officials could implement necessary preventative measures that are aimed at this age-group of people. Not only that, even the educational institutions and families that are responsible to care for this age-group people could provide more attention to these people during a similar pandemic. Similarly, the report also shows that there are more number deaths observed in Texas counties where there were people that took fewer booster shots. This information could be used by the Texas health officials to convince the general public on the significance of booster shots on preventing COVID-19 death.

**Table of Contents**

[1](#_gjdgxs) Problem Description 3

[2](#_30j0zll) Data Collection and Data Quality 4

[2.1](#_1fob9te) Dataset 1: COVID-19\_global\_mobility 4

[2.2](#_3znysh7) Dataset 2: COVID-19\_cases\_TX 4

[2.3](#_3dy6vkm) Dataset 3: COVID-19\_cases\_plus\_census 5

[2.4](#_1t3h5sf) Dataset 4: COVID Vaccination Report 6

[3](#_4d34og8) Data Exploration 9

[3.1](#_2s8eyo1) Dataset 1: COVID-19\_global\_mobility 9

[3.1.1](#_17dp8vu) Feature Processing (Dataset 1) 11

[3.1.2](#_3rdcrjn) Data Analysis (Dataset 1) 13

[3.2](#_26in1rg) Dataset 2: COVID-19\_cases\_TX 20

[3.2.1](#_lnxbz9) Feature Processing (Dataset 2) 20

[3.2.2](#_35nkun2) Data Analysis (Dataset 2) 20

[3.2.2.1](#_1ksv4uv) Death Percentages for each County 20

[3.2.2.2](#_44sinio) Is there any relation between Death Percentage and Median Income? 24

[3.2.2.3](#_2jxsxqh) How does total population affect the death percentage compared to median income? 25

[3.2.2.4](#_z337ya) Are we flattening the curve? 26

[3.2.2.5](#_3j2qqm3) Finding Correlation between some features 27

[3.3](#_1y810tw) Dataset 3: COVID-19\_cases\_plus\_Census 28

[3.3.1](#_4i7ojhp) Data Analysis (Dataset 3) 30

[3.4](#_2xcytpi) DataSet 4:- COVID Vaccination Report for TX 42

[3.4.1](#_1ci93xb) Feature Processing(Dataset 4) 42

[3.4.2](#_3whwml4) Data Analysis( Dataset 4) 42

[3.4.2.1](#_2bn6wsx) How is the vaccination trend looking for TX? 42

[3.4.2.2](#_qsh70q) How is the Booster Dose per thousand spread across TX county? 43

[3.4.2.3](#_3as4poj) Is there any relation between Booster Vaccines and Income per Capita? 44

[3.4.2.4](#_1pxezwc) Is there any relationship between the Death percentage and the Booster Shots? 45

[4](#_49x2ik5) Modeling and Evaluation 45

[5](#_2p2csry) Recommendations 46

[6](#_147n2zr) Conclusion 46

[7](#_3o7alnk) List of References 46

[8](#_23ckvvd) Appendix 47

[8.1](#_ihv636) 7.1 Student Contributions 49

# Problem Description

A widespread disease COVID-19 started at the end of 2019, and it quickly spread throughout the entire world impacting every aspect of human society. Since then, there have been several kinds of studies conducted on the impact of this pandemic. The data on those studies are available for the general public. Among those datasets, we will be looking at three different datasets. The primary focus of the analysis is based on understanding the impact of this pandemic on different aspects of human society all around the world since 2019. For an instance, the dataset “Global\_Mobility\_Report.csv” provides insights into the change in public visits to different locations (like parks, grocery stores, recreational locations, etc.) in response to the policies that the governments of different countries implemented to combat against COVID-19 between 2020 and 2021. Additionally, data mining is performed on these datasets to understand any existing relationships between entities. This can be used to predict recurrence of similar pandemic in the near future and the consequences of such situations.

The insights generated from these analyses can be utilized by the leaders of various government functions to establish necessary preventative measures. For example, organizations such as Centers for Disease Control and Prevention (CDC) could allocate resources in necessary preventative measures and contingency plans if a similar pandemic occurs in the future. Such measures not only save the lives of the people, they could save the billions or trillions of government spending that come with such pandemics. For example, the total spending by the US Department of Health and Human Services was $4.7 trillion as of October 2024 (see reference). The dataset “Global\_Mobility\_Report.csv “ can be used by public health officials to understand the impact of public visits to different places due to the pandemic. Such observations can help them to further understand the effectiveness of the policies that they implemented to combat against the pandemic. If we go into a more granular level, this sort of analysis can be very helpful for business owners of all sizes. These business institutions could implement necessary preventative and contingency measures to protect the business and their employees.

As there is a large amount of information embedded within these datasets, several meaningful and important questions could be answered. For example, using the dataset “Global\_Mobility\_Report.csv” the health officials of a specific country could answer if their policies were effective based on the public visits to different places. We could compare the public movement between 2020 and 2021. The government officials not only can understand the effectiveness of their policies, they can also compare themselves to other countries’ policies to combat against the pandemic.

# Data Collection and Data Quality

## Dataset 1: COVID-19\_global\_mobility

The dataset “Global\_Mobility\_Report.csv” is from the report collected by COVID-19 Community Mobility Reports (see source link below). Although there are some missing values in many feature columns, the overall data quality is high as it was collected by a reliable platform. Depending on the nature of the data, the missing values could be replaced or removed while maintaining the quality of the data for data mining. All 14 variables in the dataset have been labeled appropriately for a user to understand the nature of the dataset.

## Dataset 2: COVID-19\_cases\_TX

This dataset holds the confirmed COVID cases and deaths for each county in Texas for the timeframe between

The features that are available in the data set are noted in the below table:

| **Feature Name** | **Feature Summary** |
| --- | --- |
| county\_fips\_code | Min. : 0  1st Qu.:48125  Median :48253  Mean :48065  3rd Qu.:48381  Max. :48507 |
| county\_name | Length:94350  Class :character  Mode :character |
| state | Length:94350  Class :character  Mode :character |
| state\_fips\_code | Min. :48  1st Qu.:48  Median :48  Mean :48  3rd Qu.:48  Max. :48 |
| date | Min. :2020-01-22  1st Qu.:2020-04-23  Median :2020-07-24  Mean :2020-07-24  3rd Qu.:2020-10-25  Max. :2021-01-25 |
| confirmed\_cases | Min. : 0.0  1st Qu.: 1.0  Median : 82.0  Mean : 2158.6  3rd Qu.: 639.8  Max. :297629.0 |
| deaths | Min. : 0.00  1st Qu.: 0.00  Median : 2.00  Mean : 38.19  3rd Qu.: 15.00  Max. :4024.00 |

The quality of the data is overall good. Below are the cleaning activities that would be performed on this dataset:

* For some of the observations, the County could not be identified, As we are performing our analysis at county level, These records are invalid for this analysis purpose. Hence the dataset will be cleaned by removing these observations.
* The data type of the date column will be converted to date so that analysis is easier with the correct data type.

## Dataset 3: COVID-19\_cases\_plus\_census

The dataset “COVID-19\_cases\_plus\_census” is extracted from USAFacts US Coronavirus Database (USAFacts). The dataset has 259 feature columns and 3142 observations. These observations represent US COVID-19 cases and death counts for all US states and counties. As per the site, the data is collected from the CDC, and state and local health agencies. This is made available for the general public and is hosted in Google BigQuery. Since the source of the dataset is from the government functions, the quality is reliable and is of high quality.

To maintain the focus of the analysis limited to specific aspects of the dataset, there were many feature columns removed or updated. Shown below are some variable cleaning processes implemented.

* All the values in the feature columns, representing less than 50% spent on rent, were added to update a single value in the new column “rent\_under\_50\_percent”. The values in the “rent\_over\_50\_percent” were left unchanged. So, the previous features columns representing the counts for less than 50% on rent were removed.
* To limit focus on the financial structure of the families, the feature columns that represent various structures of the families (for example, count of children, count of parents, etc.) were removed from the dataset
* The features columns containing data on various income ranges were combined to form four new columns. These columns divided the income data from 0 to 200K and above in the 50K range.

Refer to the “DataMining\_Project1\_plus\_census” file for the entire feature processing performed on the dataset. The table above shows the remaining feature columns with their statistics.

## Dataset 4: COVID Vaccination Report

This data set holds the vaccination report filter for all Texas Counties. The report is downloaded from the CDC (refer “COVID Vaccination Report source” in the reference section for the source)

This data includes the cumulative number of vaccines administered per county in Texas. The data set also includes the number of further doses that were Administered. According to the CDC report this data is considered as approximately 98% accurate. This accuracy is considered as appropriate for the analysis that we are going to perform. Below are the columns that are available in this dataset.

Below table indicates the feature and a brief explanation of these features:

| **Feature Name** | **Feature Description** |
| --- | --- |
| Date | Date for the Observation |
| FIPS | FIPS |
| MMWR\_week | Week Number |
| Recip\_County | County for which the Observation is recorded |
| Recip\_State | State to for which the Observation is recorded |
| Completeness\_pct | Completeness percentage of the data |
| Administered\_Dose1\_Recip | Administered Dose1 count and percentage and the split across different age groups |
| Administered\_Dose1\_Pop\_Pct |
| Administered\_Dose1\_Recip\_5Plus |
| Administered\_Dose1\_Recip\_5PlusPop\_Pct |
| Administered\_Dose1\_Recip\_12Plus |
| Administered\_Dose1\_Recip\_12PlusPop\_Pct |
| Administered\_Dose1\_Recip\_18Plus |
| Administered\_Dose1\_Recip\_18PlusPop\_Pct |
| Administered\_Dose1\_Recip\_65Plus |
| Administered\_Dose1\_Recip\_65PlusPop\_Pct |
| Series\_Complete\_Yes | Series Complete Dose count and percentage and the split across different age groups |
| Series\_Complete\_Pop\_Pct |
| Series\_Complete\_5Plus |
| Series\_Complete\_5PlusPop\_Pct |
| Series\_Complete\_5to17 |
| Series\_Complete\_5to17Pop\_Pct |
| Series\_Complete\_12Plus |
| Series\_Complete\_12PlusPop\_Pct |
| Series\_Complete\_18Plus |
| Series\_Complete\_18PlusPop\_Pct |
| Series\_Complete\_65Plus |
| Series\_Complete\_65PlusPop\_Pct |
| Booster\_Doses | Booster Dose count and percentage and the split across different age groups |
| Booster\_Doses\_Vax\_Pct |
| Booster\_Doses\_5Plus |
| Booster\_Doses\_5Plus\_Vax\_Pct |
| Booster\_Doses\_12Plus |
| Booster\_Doses\_12Plus\_Vax\_Pct |
| Booster\_Doses\_18Plus |
| Booster\_Doses\_18Plus\_Vax\_Pct |
| Booster\_Doses\_50Plus |
| Booster\_Doses\_50Plus\_Vax\_Pct |
| Booster\_Doses\_65Plus |
| Booster\_Doses\_65Plus\_Vax\_Pct |
| Second\_Booster\_50Plus | Second Booster |
| Second\_Booster\_50Plus\_Vax\_Pct |
| Second\_Booster\_65Plus |
| Second\_Booster\_65Plus\_Vax\_Pct |
| SVI\_CTGY | Series Complete Dose count and percentage and the split across different age groups |
| Series\_Complete\_Pop\_Pct\_SVI |
| Series\_Complete\_5PlusPop\_Pct\_SVI |
| Series\_Complete\_5to17Pop\_Pct\_SVI |
| Series\_Complete\_12PlusPop\_Pct\_SVI |
| Series\_Complete\_18PlusPop\_Pct\_SVI |
| Series\_Complete\_65PlusPop\_Pct\_SVI |
| Metro\_status |
| Series\_Complete\_Pop\_Pct\_UR\_Equity |
| Series\_Complete\_5PlusPop\_Pct\_UR\_Equity |
| Series\_Complete\_5to17Pop\_Pct\_UR\_Equity |
| Series\_Complete\_12PlusPop\_Pct\_UR\_Equity |
| Series\_Complete\_18PlusPop\_Pct\_UR\_Equity |
| Series\_Complete\_65PlusPop\_Pct\_UR\_Equity |
| Booster\_Doses\_Vax\_Pct\_SVI |
| Booster\_Doses\_12PlusVax\_Pct\_SVI | Booster Dose count and percentage and the split across different age groups |
| Booster\_Doses\_18PlusVax\_Pct\_SVI |
| Booster\_Doses\_65PlusVax\_Pct\_SVI |
| Booster\_Doses\_Vax\_Pct\_UR\_Equity |
| Booster\_Doses\_12PlusVax\_Pct\_UR\_Equity |
| Booster\_Doses\_18PlusVax\_Pct\_UR\_Equity |
| Booster\_Doses\_65PlusVax\_Pct\_UR\_Equity |
| Census2019 | Population Census data for every county |
| Census2019\_5PlusPop |
| Census2019\_5to17Pop |
| Census2019\_12PlusPop |
| Census2019\_18PlusPop |
| Census2019\_65PlusPop |
| Bivalent\_Booster\_5Plus | Bivalent Booster Dose count and percentage and the split across different age groups |
| Bivalent\_Booster\_5Plus\_Pop\_Pct |
| Bivalent\_Booster\_12Plus |
| Bivalent\_Booster\_12Plus\_Pop\_Pct |
| Bivalent\_Booster\_18Plus |
| Bivalent\_Booster\_18Plus\_Pop\_Pct |
| Bivalent\_Booster\_65Plus |
| Bivalent\_Booster\_65Plus\_Pop\_Pct |

# Data Exploration

## Dataset 1: COVID-19\_global\_mobility

'data.frame': 3991405 obs. of 14 variables:

$ country\_region\_code : chr "AE" "AE" "AE" "AE" ...

$ country\_region : chr "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" ...

$ sub\_region\_1 : chr "" "" "" "" ...

$ sub\_region\_2 : chr "" "" "" "" ...

$ metro\_area : chr "" "" "" "" ...

$ iso\_3166\_2\_code : chr "" "" "" "" ...

$ census\_fips\_code : int NA NA NA NA NA NA NA NA NA NA ...

$ date : chr "2020-02-15" "2020-02-16" "2020-02-17" "2020-02-18" ...

$ retail\_and\_recreation\_percent\_change\_from\_baseline: int 0 1 -1 -2 -2 -2 -3 -2 -1 -3 ...

$ grocery\_and\_pharmacy\_percent\_change\_from\_baseline : int 4 4 1 1 0 1 2 2 3 0 ...

$ parks\_percent\_change\_from\_baseline : int 5 4 5 5 4 6 6 4 3 5 ...

$ transit\_stations\_percent\_change\_from\_baseline : int 0 1 1 0 -1 1 0 -2 -1 -1 ...

$ workplaces\_percent\_change\_from\_baseline : int 2 2 2 2 2 1 -1 3 4 3 ...

$ residential\_percent\_change\_from\_baseline : int 1 1 1 1 1 1 1 1 1 1 …

Internal structure of dataset “COVID\_19\_global\_mobility”

country\_region\_code country\_region sub\_region\_1 sub\_region\_2

135 135 1861 9916

metro\_area iso\_3166\_2\_code census\_fips\_code

66 2225 2838

Count of unique values in the columns

Shown above are the internal structures of the dataset “COVID-19\_global\_mobility” and the counts of unique values in each of the 14 feature columns or variables. The dataset has 3991405 observations. Following are brief details on the variables:

1. **country\_region\_code**: Two letters country code, character type, 135 unique codes
2. **country\_region**: Country name, character type, 135 unique country names
3. **sub\_region\_1**: character type, 1861 unique sub regions of different countries, contains the name of a primary administrative subdivision within the country, such as a state, province, or region.
4. **sub\_region\_2**: character type, 9916 unique sub regions of different countries, contains the name of a secondary administrative subdivision within the primary subdivision, such as a county, district, or municipality
5. **metro\_area**: character type, 66 unique metropolitan areas of different countries, areas typically encompass a central city and its surrounding suburbs and exurbs.
6. **iso\_3166\_2\_code**: character type, 2225 unique codes (two letter country code followed by two letter province code)
7. **census\_fips\_code**: character type, 2838 unique codes (two letters code for US state followed by two letters code for its county)
8. **date**: character type which represents date of observations, observations from 2020 to 2021
9. **retail\_and\_recreation\_percent\_change\_from\_baseline**: integer data type which simply indicate the percentage changes in visits to retail and recreation sectors to the baseline (i.e. before COVID-19 pandemic)
10. **grocery\_and\_pharmacy\_percent\_change\_from\_baseline**: integer data type which indicates the percentage changes in visits to grocery and pharmacy sectors to the baseline
11. **parks\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in the visits to parks to the baseline
12. **transit\_stations\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in the visits to transit stations (like public bus stations, train stations, etc.) to the baseline
13. **workplaces\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in the visits to workplaces to the baseline
14. **residential\_percent\_change\_from\_baseline**: integer data type which indicates the percentage change in amount of time people spent in residential locations compared to the baseline

| **Attribute Name** | **Attribute Summary** |  | **Attribute Name** | **Attribute Summary** |
| --- | --- | --- | --- | --- |
| census\_fips\_code | Min. : 1001  1st Qu.:18105  Median :29115  Mean :30356  3rd Qu.:45051  Max. :56045  NA's :3139208 |  | retail\_and\_recreation\_percent\_change\_from\_baseline | Min. :-100.0  1st Qu.: -41.0  Median : -19.0  Mean : -23.2  3rd Qu.: -4.0  Max. : 545.0  NA's :1478424 |
| grocery\_and\_pharmacy\_percent\_change\_from\_baseline | Min. :-100  1st Qu.: -14  Median : -2  Mean : -3  3rd Qu.: 9  Max. : 615  NA's :1564666 |  | parks\_percent\_change\_from\_baseline | Min. :-100.0  1st Qu.: -44.0  Median : -17.0  Mean : -9.5  3rd Qu.: 11.0  Max. :1206.0  NA's :2080860 |
| transit\_stations\_percent\_change\_from\_baseline | Min. :-100.0  Median : -28.0  Mean : -27.2  3rd Qu.: -7.0  Max. : 554.0  NA's :1973496 |  | workplaces\_percent\_change\_from\_baseline | Min. :-100.0  1st Qu.: -32.00  Median : -19.00  Mean : -20.07  3rd Qu.: -5.00  Max. : 260.00  NA's : 189760 |
| workplaces\_percent\_change\_from\_baseline | Min. : -46.0  1st Qu.: 4.0  Median : 8.0  Mean : 9.4  3rd Qu.: 14.0  Max. : 65.0  NA's :1678955 |  |  |  |

Shown above is basic statistical information on all the feature columns that contain values of numerical data type. For example, we can see that there are census fips code 1001 to 56045. The values in the column **“**residential\_percent\_change\_from\_baseline” range from -46 to 65. We can see that there are missing values in all feature columns. See below for the percentage of missing values in each feature column.

country\_region 0.00

iso\_3166\_2\_code 0.00

census\_fips\_code 78.65

retail\_and\_recreation\_percent\_change\_from\_baseline 37.04

grocery\_and\_pharmacy\_percent\_change\_from\_baseline 39.20

parks\_percent\_change\_from\_baseline 52.13

transit\_stations\_percent\_change\_from\_baseline 49.44

workplaces\_percent\_change\_from\_baseline 4.75

residential\_percent\_change\_from\_baseline 42.06

year\_month 0.00

Here is a quick summary of the variables that have non-character data type. Let’s take the variable “census\_fips\_code” where we can see that the code for “census\_fips\_code” goes from 1001 to 56045. The mean of the values is 30356. We can also see it has 3139208 (out of 3991405)observations as “NA”. This means it has about 78% of the values not available. Similar interpretations can be made for the remaining variables.

| **country\_region** | **iso\_3166\_2\_code** | **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** | **residential\_percent\_change\_from\_baseline** | **year\_month** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| United Arab Emirates |  | NA | 0 | 4 | 5 | 0 | 2 | 1 | 2020-02 |
| United Arab Emirates |  | NA | 1 | 4 | 4 | 1 | 2 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -1 | 1 | 5 | 1 | 2 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -2 | 1 | 5 | 0 | 2 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -2 | 0 | 4 | -1 | 2 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -2 | 1 | 6 | 1 | 1 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -3 | 2 | 6 | 0 | -1 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -2 | 2 | 4 | -2 | 3 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -1 | 3 | 3 | -1 | 4 | 1 | 2020-02 |
| United Arab Emirates |  | NA | -3 | 0 | 5 | -1 | 3 | 1 | 2020-02 |

**Table 3.1**: First 10 rows in the dataset 1

### Feature Processing (Dataset 1)

Since the variable “country\_region” already has names of the countries, we can ignore the variable “country\_region\_code”. The metropolitan areas in the “metro” column represent only the areas of certain countries in the “country\_region” column. Additionally, the focus of the analysis will be limited to the province of certain countries. Thus, the variable “metro” can be ignored. Similarly, the variables “sub\_region\_1” and “sub\_region\_1” can be ignored from the dataset.

For the variable “date”, the data type is of “char” which means they will need to be converted to date type. Since the focus on the trend analysis month-month and year-year, the converted values will be extracted to keep only the year and month.

1. 'country\_region'
2. 'iso\_3166\_2\_code'
3. 'census\_fips\_code'
4. 'retail\_and\_recreation\_percent\_change\_from\_baseline'
5. 'grocery\_and\_pharmacy\_percent\_change\_from\_baseline'
6. 'parks\_percent\_change\_from\_baseline'
7. 'transit\_stations\_percent\_change\_from\_baseline'
8. 'workplaces\_percent\_change\_from\_baseline'
9. 'residential\_percent\_change\_from\_baseline'
10. 'year\_month'

After removing irrelevant features from the dataset, here are the feature columns that will be used to make analysis on the dataset. We can see that we removed 4 features from the initial dataset.

Since there are about 78% missing values in the column “census\_fips\_code”. These values are for countries other than the United States since other countries do not have states. This has been validated by looking at the missing values in this column for other countries as shown below. This feature column will not be utilized when analyzing any statistical trends related to other countries than the United States. Since the majority of these missing values are for countries other than the United States, dummy codes can be assigned to these remaining 134 countries in the “census\_fips\_code” column. Shown below are 20 dummy fips codes assigned to the 20 countries. Now the remaining missing values are the United States that are not accounted for in the dataset. This is only about 0.43% of the dataset which is a very small portion of the dataset. So these missing values are removed from the dataset.

The feature column “retail\_and\_recreation\_percent\_change\_from\_baseline” has about 37% missing values which is a significant portion of the observations in the dataset. These missing values for specific regions based on the “iso\_3166\_2\_code” column are replaced with the corresponding region’s average value. There are certain regions that have only missing values (about 5000 observations) within this feature column. Since these remaining missing values are not that significantly big, they are removed completely from the dataset.

Similar feature processing technique is applied to “grocery\_and\_pharmacy\_percent\_change\_from\_baseline”, “parks\_percent\_change\_from\_baseline”, “transit\_stations\_percent\_change\_from\_baseline”, and “residential\_percent\_change\_from\_baseline” since they contain about 39%, 52%, 49%, and 42% of missing values respectively. There are only about 4% missing values in the feature column “workplaces\_percent\_change\_from\_baseline”. So, the observations with these missing values are removed from the dataset.

tibble [3,639,804 × 10] (S3: tbl\_df/tbl/data.frame)

$ country\_region : chr [1:3639804] "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" "United Arab Emirates" ...

$ iso\_3166\_2\_code : chr [1:3639804] "" "" "" "" ...

$ census\_fips\_code : int [1:3639804] 1 1 1 1 1 1 1 1 1 1 ...

$ retail\_and\_recreation\_percent\_change\_from\_baseline: num [1:3639804] 0 1 -1 -2 -2 -2 -3 -2 -1 -3 ...

$ grocery\_and\_pharmacy\_percent\_change\_from\_baseline : num [1:3639804] 4 4 1 1 0 1 2 2 3 0 ...

$ parks\_percent\_change\_from\_baseline : num [1:3639804] 5 4 5 5 4 6 6 4 3 5 ...

$ transit\_stations\_percent\_change\_from\_baseline : num [1:3639804] 0 1 1 0 -1 1 0 -2 -1 -1 ...

$ workplaces\_percent\_change\_from\_baseline : int [1:3639804] 2 2 2 2 2 1 -1 3 4 3 ...

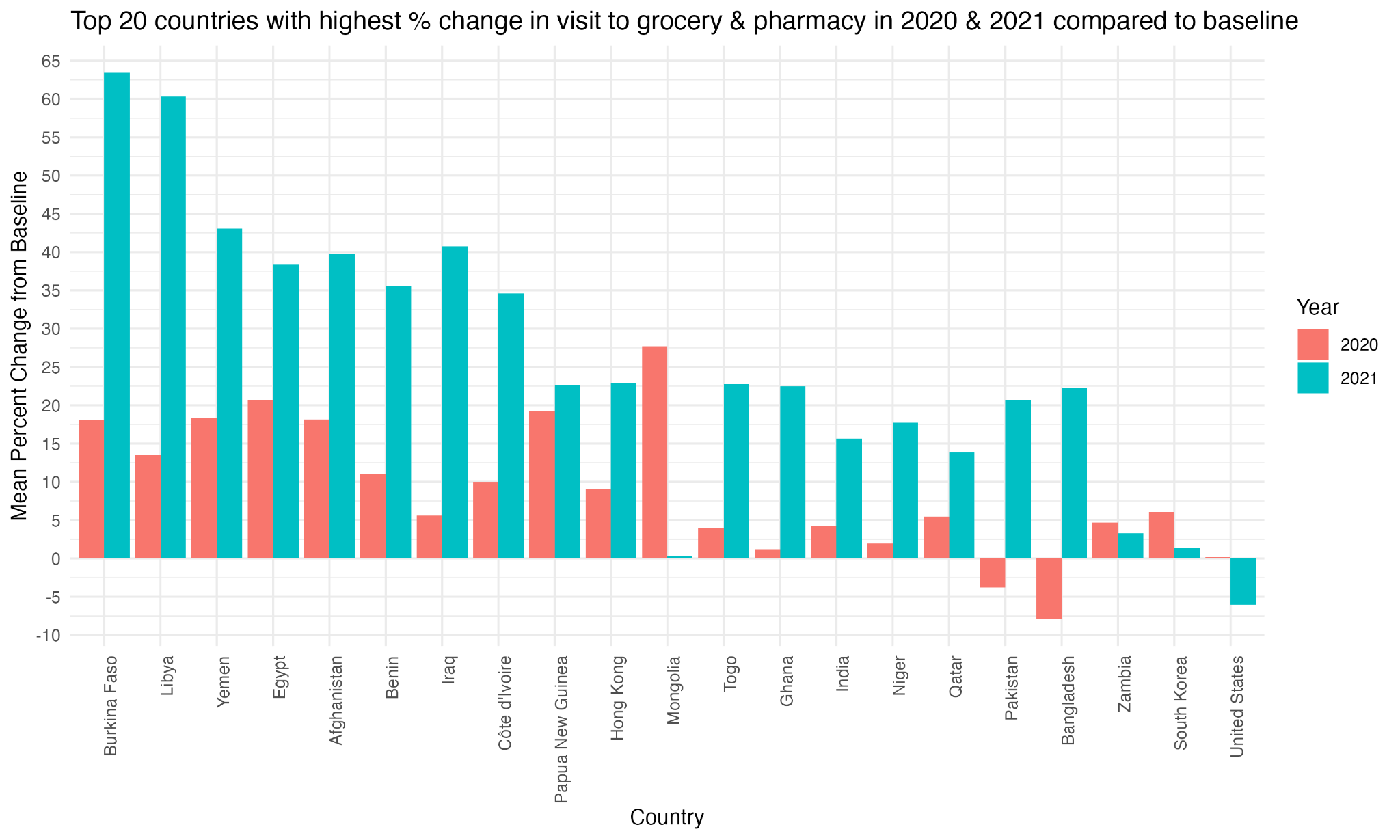
$ residential\_percent\_change\_from\_baseline : num [1:3639804] 1 1 1 1 1 1 1 1 1 1 ...

$ year\_month : Date[1:3639804], format: "2020-02-01" "2020-02-01" ...

After all the feature processing, here is the quick overview of the updated dataset. There are now 3639804 observations and 10 variables as compared to 3991405 observations and 14 variables in the original dataset. This means that about 8% of the observations were removed from the original dataset as they were missing values. However, there is still a sufficient amount of data left to perform various kinds of data mining.

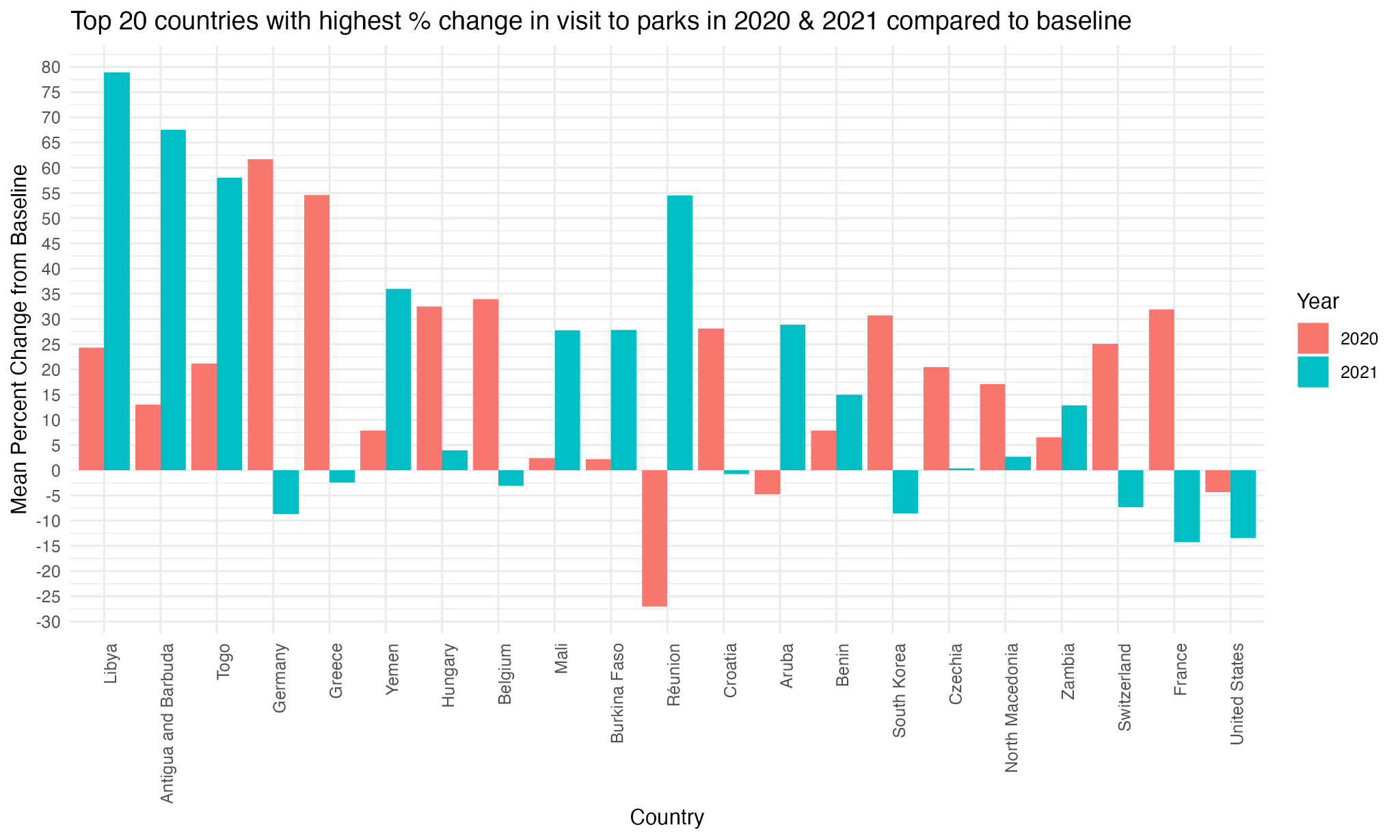
The above plot shows change (in percentage) in people's visits to retail and recreation locations for various countries between 2020 and 2021 when compared to baseline. The baseline is a value (normal) given to a day of the week. This day is the median value from the 5-week period (i.e. Jan 3 to Feb 6, 2020). As the analysis for an entire year for a specific country, the baseline is the aggregate of the individual observations recorded for the given year. The chart only focuses on the 20 countries (and the United States) with the highest change. We can clearly see that the visits to these locations in 2020 were significantly lower as compared to the year before COVID-19. For example, Reunion and New Zealand saw the slowdown by about 34% and 18% respectively. But only after a year, there were many countries that saw an increase in the number of visits to these locations again. For example, Reunion and Libya saw an increase of these visits by about 33% and 30% respectively.

### Data Analysis (Dataset 1)



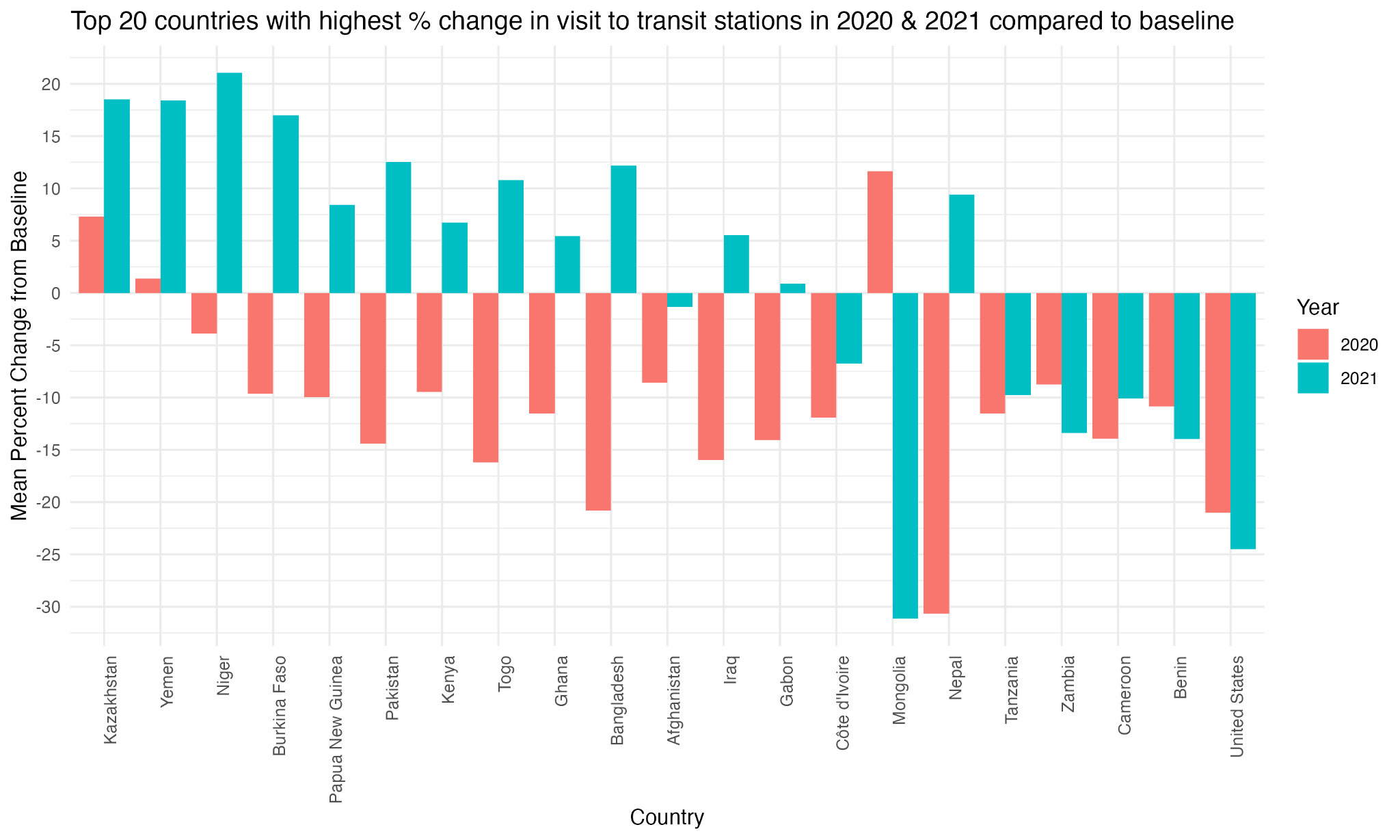
**Fig 3.1.1**: Top 20 countries (with USA) with highest % change in visit to grocery stores & pharmacies

The above chart shows 20 countries (with addition of the United States) that have the highest change (percentage) in the public visits to grocery and pharmacy locations for 2020 and 2021 from the baseline. Compared to the public visits to retail and recreational locations, we can see that the people visited more grocery and pharmacy locations in these countries for both years. This does make sense since the locations such as grocery stores are a more essential part of people's lives than the retail or recreational locations. It also makes sense that the people visited pharmacies more than the retail or recreational locations. It’s clear that the people visited these locations more in 2021 compared to 2020 which indicates that the policies were more relaxed in 2021. It’s interesting to see that the people in the United States made fewer visits to these locations in 2021 than in 2020. It could be due to late response to the pandemic compared to other countries. It could also mean that the United States policies did not relax or be made more stringent in 2021. It could also mean that the people were being treated more quickly than other countries which reduced the number of visits to pharmacies. On the other hand, the visits to these locations in Bangladesh and Pakistan were lower in 2020 than in 2021 indicating that the people there perhaps reacted slower to the pandemic than the other countries. Additionally, we can see that the people in Mongolia visited these locations the most amongst all other 20 countries right after the pandemic started (i.e. 2020). This number decreased significantly, from 27% to 1%, for Mongolia in 2021. Overall, this chart gives us a sense of people's visits to grocery stores and pharmacies in these countries between 2020 and 2021 in response to the policies imposed by the respective government officials.



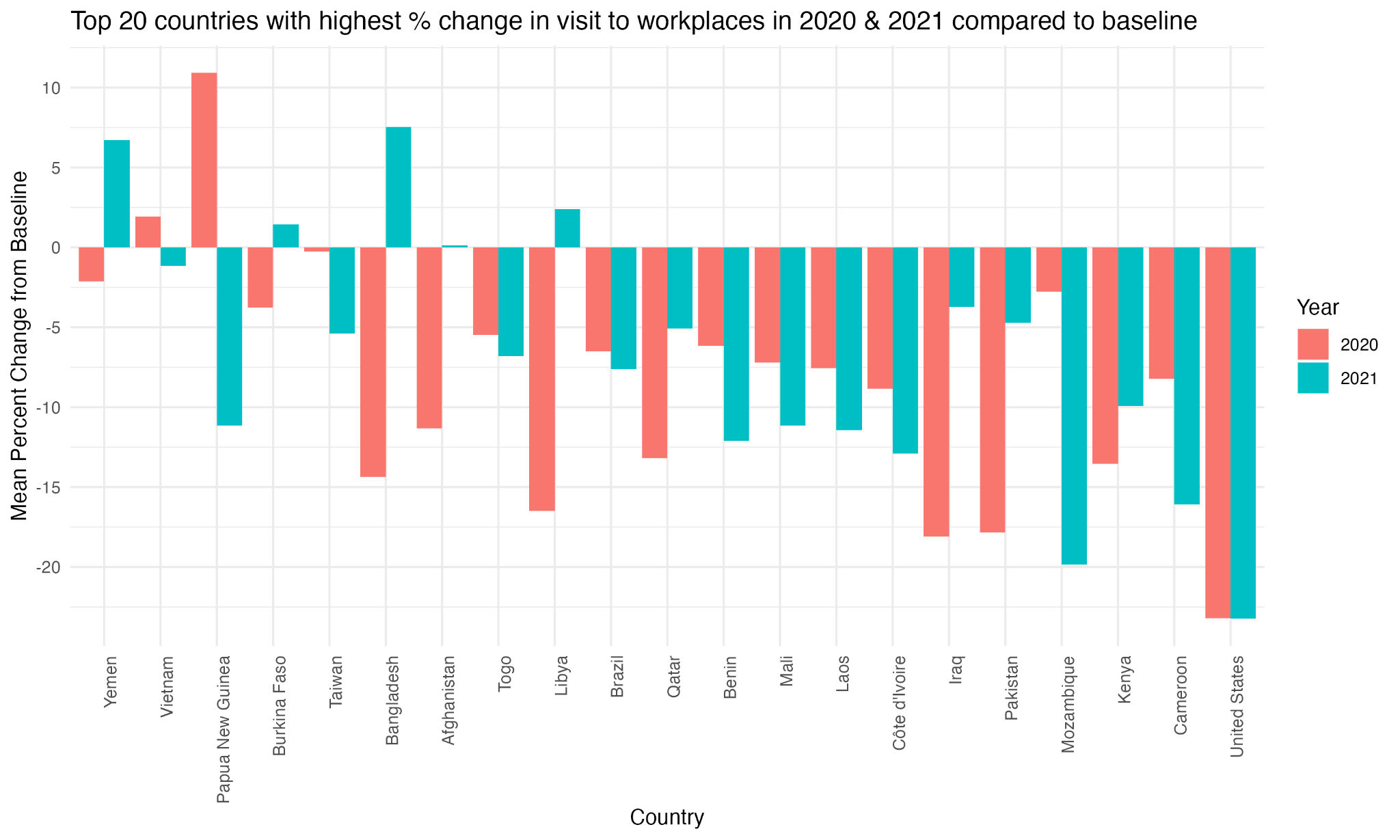
**Fig 3.1.2**: Top 20 countries (with USA) with highest % change in visit to parks

The chart above shows the percentage change in the number of public visits to the parks in the top 20 countries (in addition to the United States) between 2020 and 2021. Similar to the observations made on the previous chart (visits to grocery stores and pharmacies), we can see that the people in most of these countries were still making high visits to the parks in both years. In the year 2020, people in Reunion were making the least number of visits to the parks. But that number increased dramatically in 2021 indicating that either the pandemic policies were relaxed or the public started to cope with the pandemic. It is insightful to see that the people in Germany, France, Switzerland, South Korea, Belgium, and Greece visited the parks more in 2020 than in 2021. There could be several reasons for such behavior. Interestingly, the people in the United States visited the parks less in both years even though they visited the parks less in 2021.



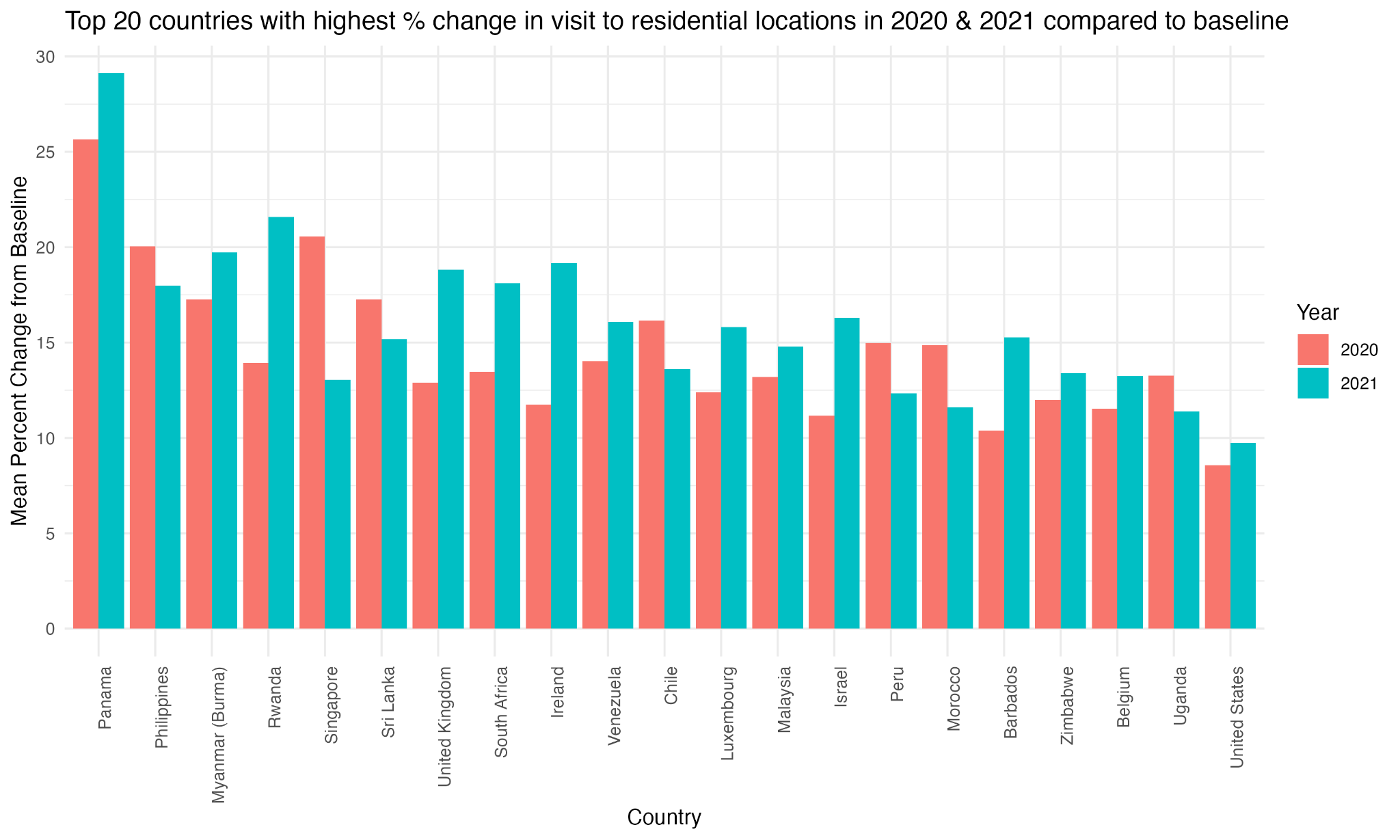
**Fig 3.1.3**: Top 20 countries (with USA) with highest % change in visit to parks

The graph above shows the average change in the number of public visits to public transit stations for the top 20 countries (in addition to the United States) in 2020 and 2021 when compared to the baseline. Similar to what we saw in the graph showing the public visits to the retail and recreational locations, we can see that the people in the majority of the countries visited the transit stations less in 2020. Only the people in Mongolia, Kazakhstan, and Yemen still visited these locations more than the baseline. Similar to previous reasoning, this behavior could be due the respective government policies not being strict enough or the people could have simply ignored those policies. Again, it is interesting to see that the people in Mongolia visited these locations more in 2021 than in 2020. On the contrary, the people in Nepal made significantly less visits to these locations in 2020, and there was slight recovery seen in 2021. On the other hand, the people in the USA still made less visits to these locations in both years. In fact, they visited these locations even less in 2021 as compared to 2020. This indicates that the government restriction policies made the people make less visits to these public transit locations.



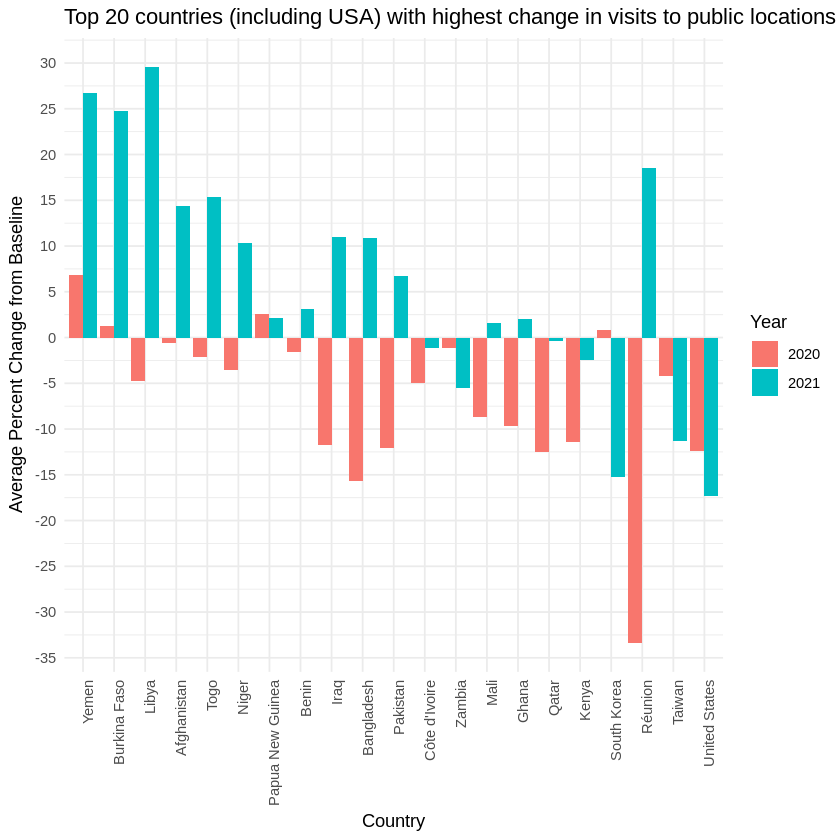
**Fig 3.1.4**: Top 20 countries (with USA) with highest % change in visit to parks

The above graph shows the percentage change of public visits to workplaces in top 20 countries (including USA) for 2020 and 2021 when compared to the baseline. Except for Papua New Guinea, the people in the rest of the other countries made fewer visits to workplaces in both years. It seems like the restriction policies in Papua New Guinea only started to show impact in 2021 as there is negative change in the visits to the workplaces in 2021. It is interesting to see that the people in Bangladesh started going to workplaces more in 2021 than in 2020. It is also interesting to see that the trend remained the same for the USA as the percentage change is almost the same for both years for the USA. This could indicate that the restriction policies remained unchanged from 2020 to 2021 in the USA. All the remaining 20 countries had the people go to workplaces more or less in 2021 compared to 2020.



**Fig 3.1.5**: Top 20 countries (with USA) with highest % change in people staying home

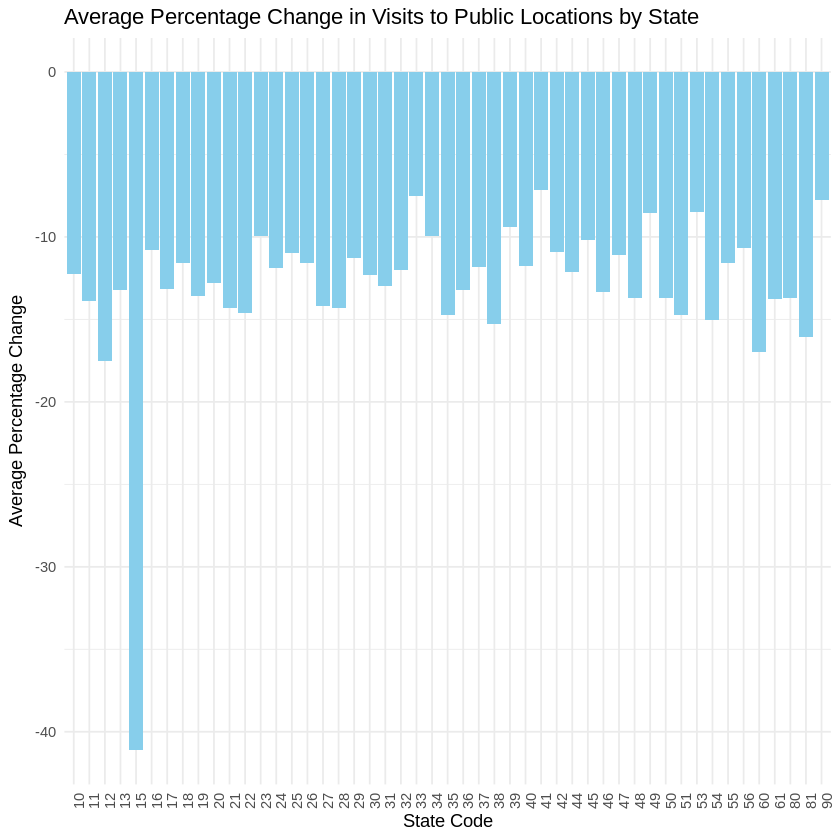
The above chart shows the percentage change in people staying home for the top 20 countries (in addition to the USA) in 2020 and 2021 when compared to the baseline. As expected, people were staying home more after the pandemic started all around the world. Out of these 21 countries, the people in 13 countries are staying home more in 2021 than in 2020. This indicates that the restrictions were not relaxed or remained effective to keep the people at their residential locations in 2021 too. This observation applies to the people in the USA too.



**Fig 3.1.6**: Top 20 countries (with USA) with highest % change in visits to public locations

​​

The above chart shows the average change of public visits to the public locations described in the previous charts in top 20 countries (including USA) for 2020 and 2021 when compared to the baseline. As predicted, we can see that the majority of these countries saw less visits to these locations in 2020, and saw some recovery in 2021. We can see that there were still fewer visits to these locations even in 2021 in the USA , supporting the previous observations on individual public locations previously.



**Fig 3.1.7:** Top 20 countries (with USA) with highest % change in visits to public locations by US states

This chart shows the % average change in public visits to public locations (2020 and 2021 combined) for all US states. We can see that the state 15 (i.e. Hawaii) saw the highest change in such visits compared to the rest of the states. Looking at statistics for Texas (code 48), we can see that the average % change is about -13%. See the appendix B “State Level Fips Code” for the state fips codes.

## Dataset 2: COVID-19\_cases\_TX

### Feature Processing (Dataset 2)

The below activities were performed to clean the data set and add additional features. (See Appendix A for code)

1. Removing the observations that were not tagged to a county
2. Correcting the datatype of the date fields.
3. Adding a new feature – Death Percentage – which is death/confirmed cases

### Data Analysis (Dataset 2)

This dataset is used in conjunction with Dataset 3 for some of the analysis performed in the below slides. Details about dataset 3 are mentioned in section 3.3.

#### Death Percentages for each County

Death percentage is “total deaths” divided by “confirmed cases” . A low death percentage indicates that a specific county was able to treat the confirmed cases better than other counties. A high death percentage indicates that these counties had more deaths w.r.t the confirmed cases. From the below map plot, We can understand the below:

* The highest death percentage for a county is close to 7.5%
* Majority of the counties have death percentage less than 2.5%
* There are some counties with relatively higher death percentage which is close to each other

A map of the state of texas

Description automatically generated

**Fig 3.2.2.1:** Death percentages of counties in TX

The below figure shows Top 10 counties with the highest and lowest death percentages.

A graph of different colored bars

Description automatically generated

**Fig 3.2.2.2:** Top 10 counties in TX with highest death percentage

* From the above bar graph can understand that Sherman county and Motley County have the highest death percentage slightly above 7.5%
* All the other counties have death percentage between 5 and 7.5%

A graph with different colored bars

Description automatically generated

**Fig 3.2.2.3:** Top 10 counties in TX with the lowest death percentage

* Loving county, King county, Borden County have death percentages close to 0%.
* All the other counties with the lowest death percentage is only below 1%
* Loving county , for example, is the least populated county in the United States with a very small population of 64. Hence it is reasonable to see such low deaths in this county.

#### Is there any relation between Death Percentage and Median Income?

A graph with blue dots

Description automatically generated

**Fig 3.2.2.4:** Relation between Death Percentage and Median Income

* The above graph indicates that the death percentage is negatively correlated to median Income.
* The counties with the highest death percentages are the counties with lowest median income.
* For example, Kenedy County has one of the highest death percentages and is one of the counties with the lowest Median Income.
* We can also see that no counties with the highest median income have the highest death percentage.

From this we can infer that the death percentage is inversely proportional to median income.

#### How does total population affect the death percentage compared to median income?

A graph with numbers and a line

Description automatically generated

**Fig 3.2.2.5:** Death Percentage w.r.t to total Population and Median Income

* We can infer from the above graph that the total population has weaker correlation with death percentage as compared to the median income.
* If the Median Income of a county is lower, irrespective of the total population, these counties seem to have a higher death percentage
* It can be noticed that the counties with the highest death percentages are spread across the county population.

#### Are we flattening the curve?

For evaluating whether we have flattened the curve or not, The daily new cases need to be calculated from the dataset. Below code snippet shows the calculation of new cases from the dataset.

tx\_covid\_cases\_date = tx\_covid\_cases **%>%** **group\_by**(date) **%>%**  
**summarise**(total\_confirmed\_cases = **sum**(confirmed\_cases),  
total\_deaths = **sum**(deaths),  
.groups = 'drop')  
tx\_covid\_cases\_date <- **mutate**(tx\_covid\_cases\_date, prev\_day\_case=**lag**(total\_confirmed\_cases, order\_by = date))  
tx\_covid\_cases\_date**$**daily\_increase<- tx\_covid\_cases\_date**$**total\_confirmed\_cases**-**tx\_covid\_cases\_date**$**prev\_day\_case

A graph with numbers and lines

Description automatically generated

**Fig 3.2.2.6:** New case per day

From the above graph, at least with the timeframe for which the dataset is available, We do not seem to be flattening the curve for the state of Texas

#### Finding Correlation between some features

A screenshot of a graph

Description automatically generated

**Fig 3.2.2.7:** Correlation chart for selected features (per 1000)

Based on the above correlation graph, We can identify that income\_less\_than \_1000, commute\_45\_49\_mins, are highly correlated features. These are good candidates for feature reduction in future.

## Dataset 3: COVID-19\_cases\_plus\_Census

| **Attribute Name** | **Attribute Summary** |
| --- | --- |
| state | Length Class Mode  3142 character character |
| state\_fips\_code | Min. 1st Qu. Median Mean 3rd Qu. Max.  1.00 18.00 29.00 30.28 45.00 56.00 |
| county\_name | Length Class Mode  3142 character character |
| confirmed\_cases | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 796.2 1916.5 7558.9 4955.0 1002614.0 |
| deaths | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 12.0 31.0 124.8 77.0 13936.0 |
| median\_age | Min. 1st Qu. Median Mean 3rd Qu. Max.  21.60 37.90 41.20 41.15 44.20 66.40 |
| total\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  74 10945 25692 102166 67445 10105722 |
| male\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  39 5514 12798 50292 33481 4979641 |
| female\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  35 5460 12885 51873 34108 5126081 |
| white\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  18 8093 20205 62787 53500 2676982 |
| black\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  0 95 758 12554 5396 1226134 |
| asian\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 31.0 138.0 5407.2 712.5 1442577.0 |
| hispanic\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  0 323 1025 17986 4868 4893579 |
| amerindian\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 24.0 95.5 668.0 348.0 64102.0 |
| other\_race\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 24.0 95.5 668.0 348.0 64102.0 |
| median\_income | Min. 1st Qu. Median Mean 3rd Qu. Max.  19264 41123 48066 49754 55764 129588 |
| income\_50K\_100K | Min. 1st Qu. Median Mean 3rd Qu. Max.  19 1244 2971 11342 8038 927390 |
| income\_100K\_150K | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 396.2 1017.5 5315.7 3134.0 477403.0 |
| income\_150K\_more | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 183.0 483.5 4582.0 1756.5 501413.0 |
| rent\_under\_50\_percent | Min. 1st Qu. Median Mean 3rd Qu. Max.  6.0 685.2 1751.5 9429.8 5166.8 1160618.0 |
| rent\_over\_50\_percent | Min. 1st Qu. Median Mean 3rd Qu. Max.  0 162 486 3237 1576 536832 |
| median\_age | Min. 1st Qu. Median Mean 3rd Qu. Max.  21.60 37.90 41.20 41.15 44.20 66.40 |
| male\_0\_20 | Min. 1st Qu. Median Mean 3rd Qu. Max.  1 1476 3524 14139 9245 1378157 |
| male\_21\_49 | Min. 1st Qu. Median Mean 3rd Qu. Max.  10 1900 4604 19741 12309 2161991 |
| male\_50\_above | Min. 1st Qu. Median Mean 3rd Qu. Max.  25 2312 5163 17760 13163 1544545 |
| female\_0\_20 | Min. 1st Qu. Median Mean 3rd Qu. Max.  4 1363 3277 13514 8837 1322003 |
| female\_21\_49 | Min. 1st Qu. Median Mean 3rd Qu. Max.  8 1737 4334 19567 11846 2131518 |
| female\_50\_above | Min. 1st Qu. Median Mean 3rd Qu. Max.  23 2452 5710 20243 14641 1795247 |
| unemployed\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 286.2 745.5 3361.0 2099.8 406426.0 |
| employed\_pop | Min. 1st Qu. Median Mean 3rd Qu. Max.  39 4550 10695 47931 29515 4805817 |
| commute | Min. 1st Qu. Median Mean 3rd Qu. Max.  140 15410 36238 154635 98145 15285110 |
| walked\_to\_work | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 98.0 242.5 1288.8 678.5 181289.0 |
| walked\_to\_work | Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0 98.0 242.5 1288.8 678.5 181289.0 |

**Table 3.3:** Summary of the attributes in the final “COVID-19\_cases\_plus\_census” dataset

| **state** | **state\_fips\_code** | **county\_name** | **confirmed\_cases** | **deaths** | **total\_pop** | **male\_pop** | **female\_pop** | **white\_pop** | **black\_pop** | **asian\_pop** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **VT** | 50 | Essex County | 111 | 0 | 6203 | 3135 | 3068 | 5929 | 64 | 32 |
| **VT** | 50 | Chittenden County | 3636 | 78 | 160985 | 78928 | 82057 | 143657 | 4091 | 6144 |
| **DE** | 10 | Kent County | 11548 | 187 | 173145 | 83544 | 89601 | 108627 | 41729 | 3459 |
| **RI** | 44 | Washington County | 5521 | 122 | 126190 | 61154 | 65036 | 115206 | 1621 | 2436 |
| **NH** | 33 | Belknap County | 2496 | 79 | 60383 | 29705 | 30678 | 57523 | 285 | 579 |
| **RI** | 44 | Newport County | 3578 | 6 | 83204 | 40952 | 42252 | 71549 | 2596 | 1670 |
| **VT** | 50 | Lamoille County | 312 | 1 | 25191 | 12504 | 12687 | 23895 | 214 | 161 |
| **CT** | 9 | Tolland County | 6255 | 125 | 151596 | 76162 | 75434 | 129519 | 4425 | 6690 |
| **VT** | 50 | Addison County | 527 | 5 | 36825 | 18214 | 18611 | 34245 | 337 | 711 |
| **VT** | 50 | Caledonia County | 307 | 4 | 30576 | 15366 | 15210 | 29070 | 272 | 186 |

| **hispanic\_pop** | **amerindian\_pop** | **other\_race\_pop** | **median\_income** | **income\_less\_50K** | **income\_50K\_100K** | **income\_100K\_150K** |
| --- | --- | --- | --- | --- | --- | --- |
| 83 | 20 | 6 | 38767 | 1650 | 802 | 193 |
| 3542 | 374 | 240 | 66906 | 24184 | 20318 | 11615 |
| 11820 | 967 | 377 | 57647 | 27659 | 21746 | 8704 |
| 3769 | 1047 | 260 | 77862 | 16334 | 14204 | 9953 |
| 951 | 130 | 65 | 65834 | 9365 | 8472 | 3886 |
| 4623 | 290 | 159 | 75463 | 12051 | 10443 | 6530 |
| 408 | 208 | 0 | 54899 | 4803 | 3240 | 1411 |
| 7860 | 38 | 336 | 81312 | 16508 | 16815 | 10705 |
| 795 | 94 | 57 | 61875 | 5965 | 5122 | 2363 |
| 460 | 65 | 9 | 47371 | 6387 | 3736 | 1339 |

| **income\_150K\_more** | **rent\_under\_50\_percent** | **rent\_over\_50\_percent** | **median\_age** | **male\_0\_20** | **male\_21\_49** | **male\_50\_above** | **female\_0\_20** | **female\_21\_49** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 75 | 411 | 97 | 50 | 677 | 900 | 1701 | 603 | 924 |
| 8789 | 16148 | 6620 | 36.6 | 20796 | 32947 | 27110 | 20668 | 32285 |
| 5272 | 13483 | 4408 | 37.3 | 24873 | 31121 | 29975 | 23801 | 33109 |
| 9117 | 8510 | 3429 | 44.1 | 15755 | 20438 | 27238 | 16106 | 20580 |
| 2856 | 4211 | 1351 | 46.7 | 6834 | 9663 | 14520 | 6541 | 9882 |
| **income\_150K\_more** | **rent\_under\_50\_percent** | **rent\_over\_50\_percent** | **median\_age** | **male\_0\_20** | **male\_21\_49** | **male\_50\_above** | **female\_0\_20** | **female\_21\_49** |
| 6397 | 10033 | 2525 | 44.6 | 9057 | 15262 | 18187 | 9040 | 14232 |
| 947 | 1814 | 720 | 40.5 | 3293 | 4659 | 5045 | 3317 | 4489 |
| 10850 | 10632 | 3650 | 37.9 | 21806 | 28436 | 27876 | 21270 | 26221 |
| 1251 | 2975 | 665 | 43.4 | 4747 | 6130 | 7914 | 4614 | 6228 |
| 632 | 2290 | 645 | 43.7 | 4099 | 5144 | 6640 | 3647 | 4865 |

| **female\_50\_above** | **unemployed\_pop** | **employed\_pop** | **commute** | **worked\_at\_home** | **walked\_to\_work** |
| --- | --- | --- | --- | --- | --- |
| 1657 | 149 | 2784 | 4820 | 167 | 81 |
| 30864 | 3827 | 90054 | 148357 | 4783 | 6859 |
| 35761 | 5563 | 78078 | 145421 | 3270 | 1379 |
| 30387 | 4142 | 64576 | 114755 | 3110 | 1725 |
| 15317 | 1589 | 30674 | 55288 | 1487 | 449 |
| 20449 | 2349 | 41471 | 73711 | 2711 | 2749 |
| 5218 | 526 | 13437 | 23083 | 906 | 536 |
| 30288 | 4962 | 80613 | 137743 | 3821 | 3533 |
| 8340 | 907 | 19872 | 31572 | 1811 | 1466 |
| 7125 | 741 | 14475 | 24724 | 1106 | 535 |

**Table 3.4:** First 10 rows of “COVID-19\_cases\_plus\_census” dataset

Table 3.3 shows the attributes names and basic statistics of the final dataset COVID-19\_cases\_plus\_census. Table 3.4 shows an overview of the first 10 rows of the dataset.

### Feature Processing (Dataset 3)

The dataset, before feature processing, contained 259 variables that provided information on various aspects of the pandemic. There were many variables that could be merged and still held necessary insights on the pandemic. For example, the variables providing information on the various income levels could be reduced to income ranges that would still provide enough insights on the financial status of the general public. Thus, the data of those variables were merged into four new income ranges: less than 50K, 50-100K, 100-150K, and 150K-above. This allowed us to drop the previous feature columns and reduce the dimension of the dataset. Following are some more feature processing performed on the dataset:

* 8 variables, containing information on certain percentage of income spent on rent, were merged into 2 new variables (i.e. rent\_under\_50\_percent and rent\_over\_50\_percent)
* variables containing various structures of families were dropped because they were less significant than people’s financial structures
* variables containing counts of males and females for various age groups were merged into new variables that contain counts of males and females for age groups 0-20, 21-49, and 50 above
* dropped variables that contained the counts of people belonging to multiple races and rather maintained emphasis on the counts of people belonging to only one specific race
* dropped variables that contained the counts of people holding various academic qualifications and rather maintained emphasis on income, race, working environment, gender and age
* grouped all the variables that contained the counts of people commuting using various methods and for certain time to work into simply a new group “commute” to have all the counts commuting to work

After all the feature processing completed, there were 33 variables left in the dataset. The information in these variables were used to discover the following insights on the pandemic.

### Data Analysis (Dataset 3)

The table above shows the statistics of the dataset with only 29 variables that were left after removing irrelevant or less significant variables and adding new variables. Following is brief details on some of the variables:

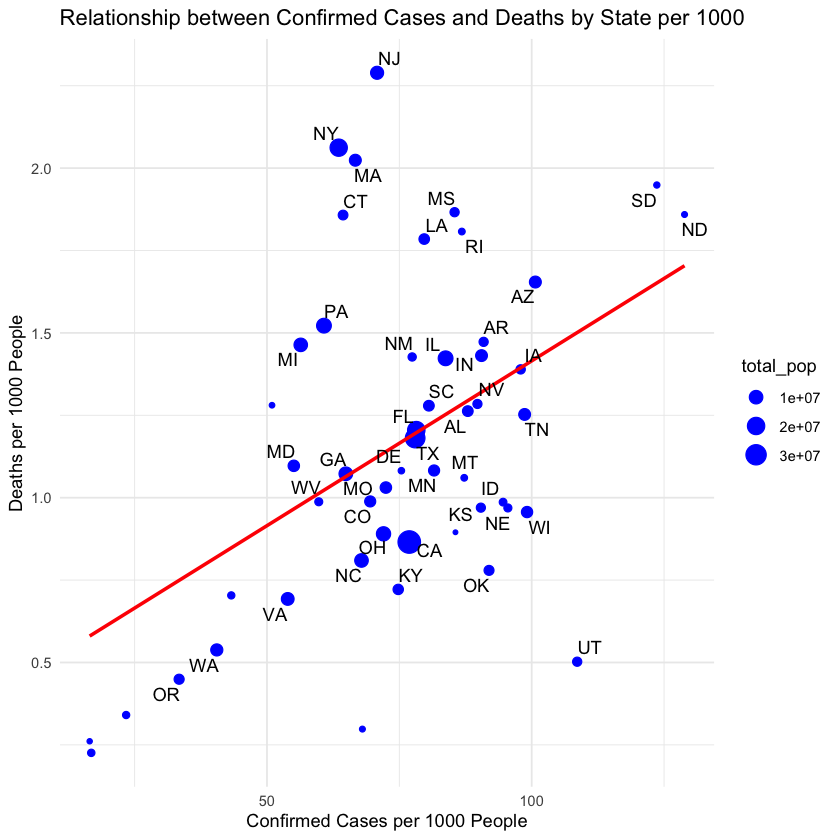
* state : contains two letters code for the states, 51 unique codes, data type is character
* state\_fips\_code: contains two digits fips code for the states, 51 unique fips codes, data type is integers
* county\_name: contains the names of the US counties, 1878 unique county names, data type is character
* confirmed\_cases: contains the number of confirmed cases per geographic region, the maximum confirmed cases is 1002614, data type of integers
* total\_pop: contains total population in a geographic region, the range of total population range from 74 to 10105722, data type of numeric

Upon further inspecting each variable, there were no missing values found in the entire dataset. Also, each of the variables only contained only one type of data (i.e. numeric or character or integers).



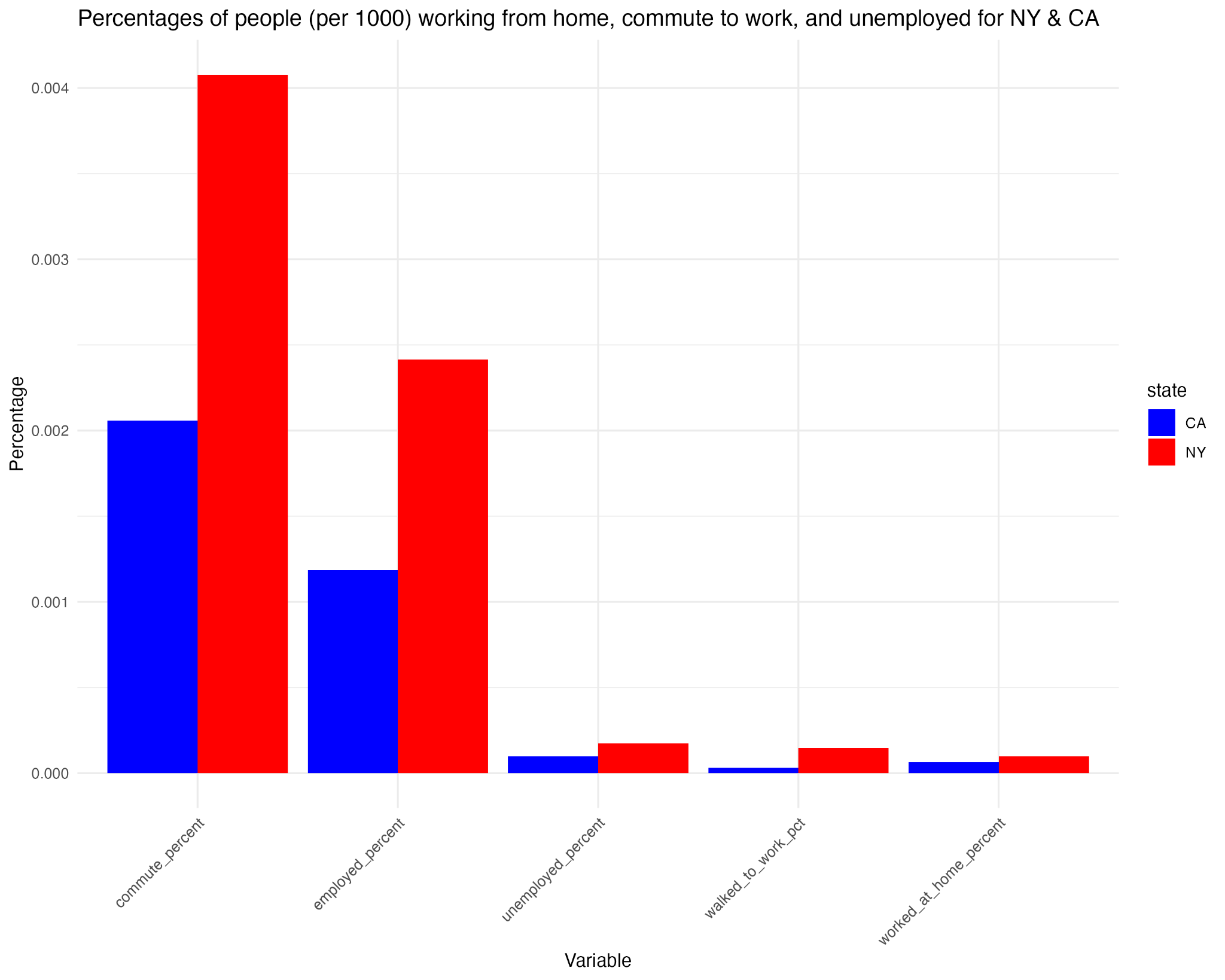
**Figure 3.3.1**: Relationship between confirmed cases and deaths by state

The chart above shows the relationship between the number of confirmed cases of COVID-19 and the number of the deaths for the US states. The analysis only includes the death cases that are 1000 or more. The size of the bubbles represent the size of the state's total population. We can see that there is a linear relationship between the number of the confirmed cases and the number of the deaths. For example, it’s clear that this relationship is very strong for Florida (FL) and Texas (TX). We can also see that the states with higher population saw a higher number of confirmed cases. One interesting observation is that California (CA), with a higher population and number of confirmed cases than New York (NY), still saw a lower number of deaths than NY. There is somewhat similar behavior that can be seen with Pennsylvania (PA) and New Jersey “NJ”. Both of these states have similar population sizes but there were fewer confirmed cases in PA than in NJ. However, there were slightly more deaths in NJ. There could be many possible reasons behind such behavior.



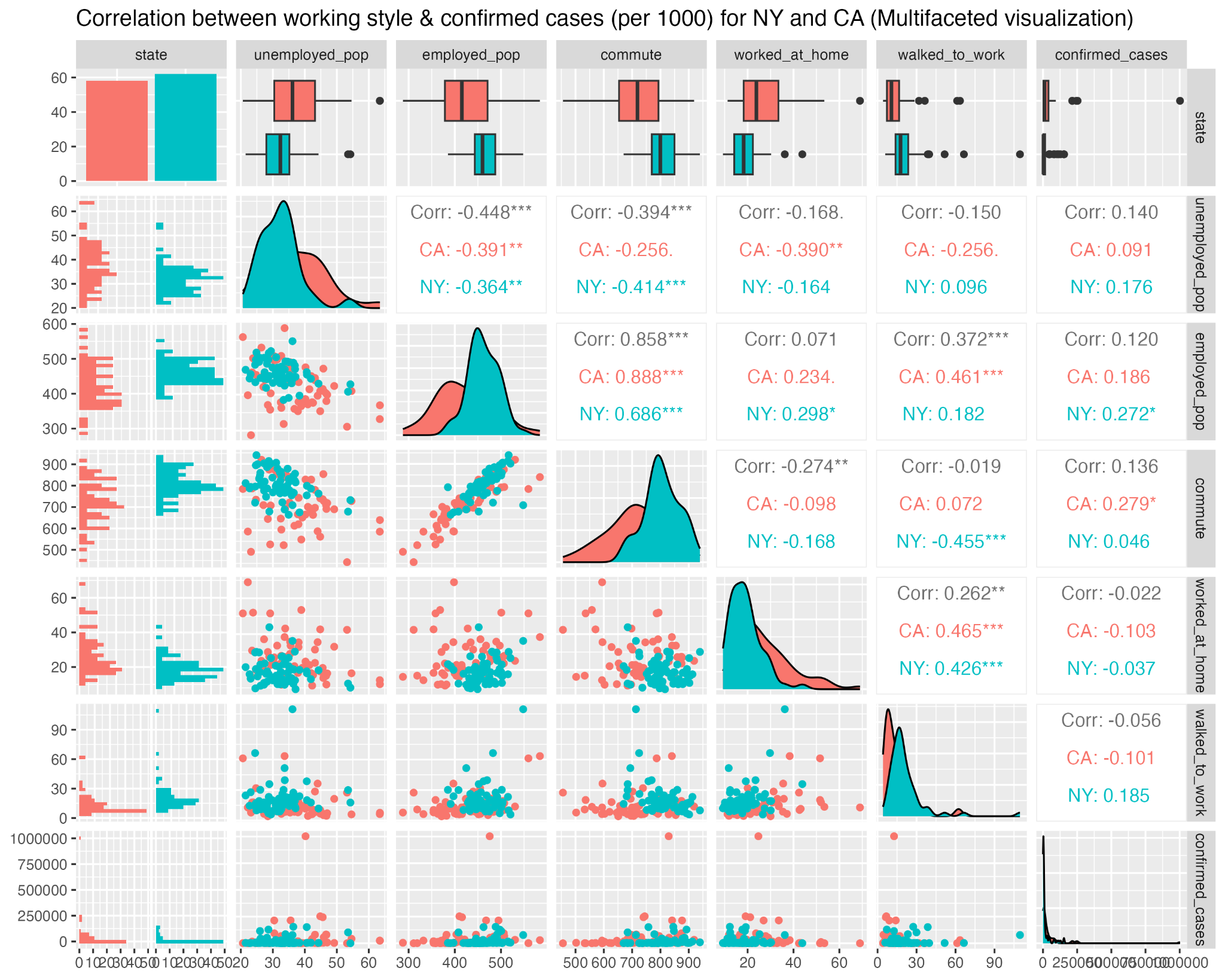
**Figure 3.3.2**: Relationship between confirmed cases and deaths 1000 by state

Now looking at the relationship between the confirmed cases and deaths per 1000 people, we can still see that there were more number of confirmed cases in CA than in NY but there were significantly more number of deaths in NY than in CA. In fact, there were more than twice the number of deaths in NY than in CA. This means that the observation made on the previous section still remains. Since we are more interested in looking into different variables that could provide insights on confirmed cases for a state, we will continue to look for other variables that could be a good indicator for the confirmed cases in a state.



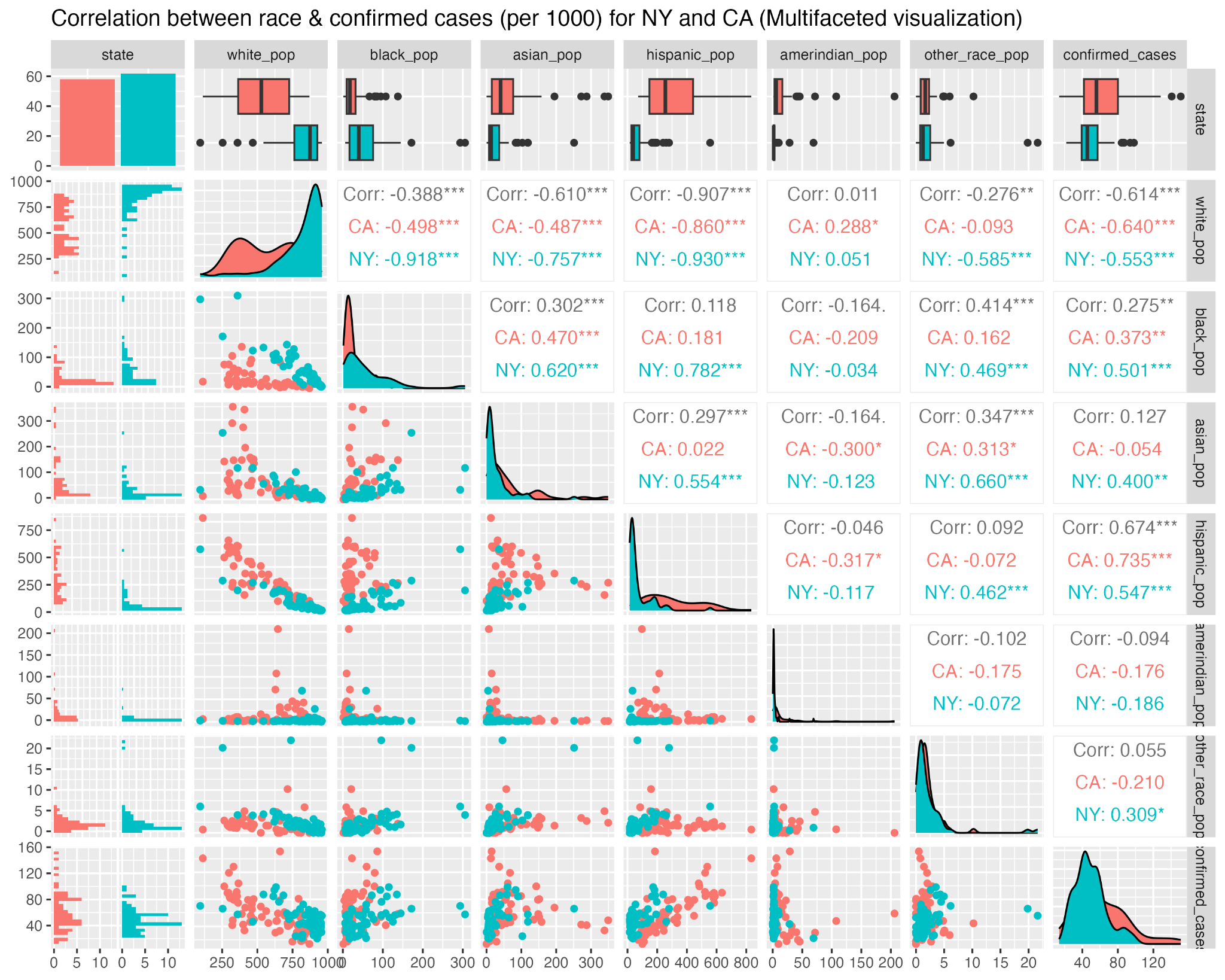
**Fig 3.3.3:** Percentages of population (per 1000) commuting to work, worked from and unemployed (NY & CA)

The above bar chart shows the comparison between the percentages of the people commuting to work, unemployed, and working from home between CA and NY. We can see that significantly more people were commuting to work and were employed in NY than in CA during the pandemic. At the same time, if we look at the percentage of people who stayed at home or walked to work or were unemployed, there is not much difference between CA and NY . To understand if any of these work related variables is a strong indicator for the confirmed cases in a state, we will compute correlation values between these variables against confirmed cases for NY and CA.



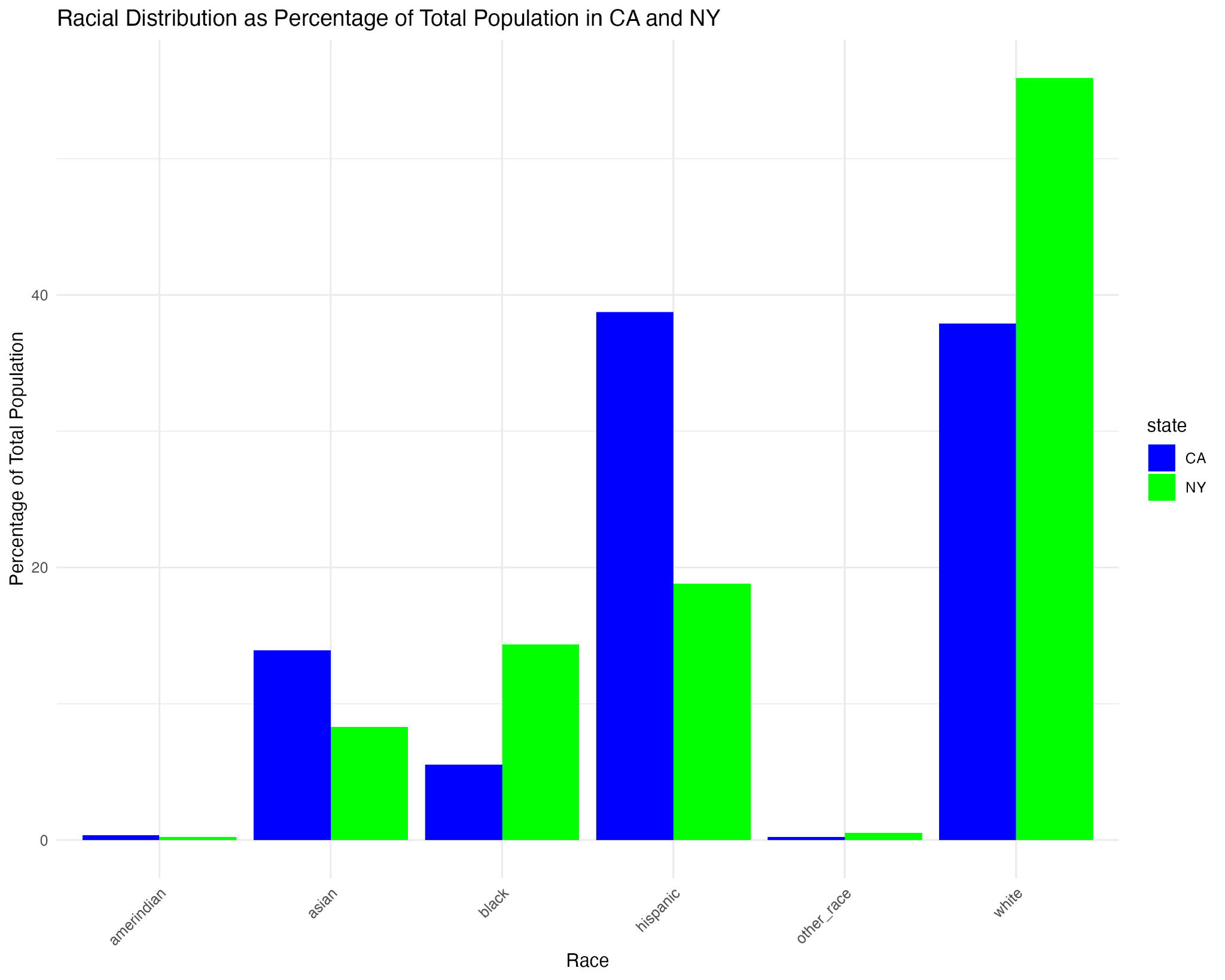
**Fig 3.3.4**: Correlation plot (work style against confirmed cases) for NY and CA per 1000 (Multifaceted Visualization)

The above multifaceted chart shows correlation between different variables related to employment and number of COVID confirmed cases for CA and NY population for 1000 people. Looking at the overall correlation, we can see that the variable “employment\_pop” has the highest correlation value. We also know that there are more employed people in NY than in CA. This should mean that there should be more confirmed cases in NY than in CA. However, in figure 3.3.2, we see that there are more confirmed cases in CA than NY. Thus, we could say that the working style alone is not a significant indicator for confirmed cases for a state.



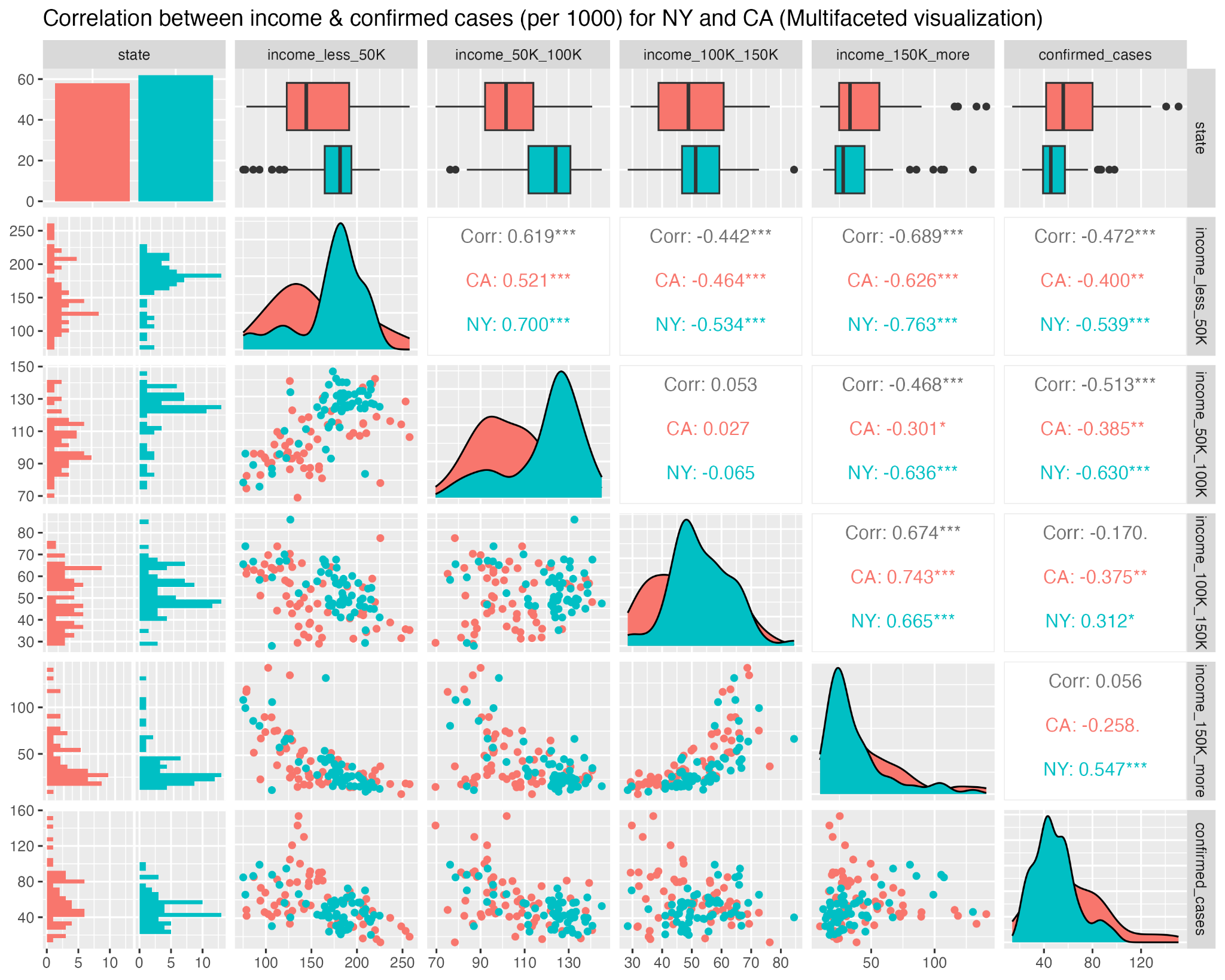
**Fig 3.3.6**: Correlation plot (race against confirmed cases) for NY and CA per 1000 (Multifaceted Visualization)

Based on the above correlation plot, we can see that the variable “hispanic\_pop” has the highest overall correlation value (i.e. 0.674). This indicates that the state with a higher hispanic population could have a higher number of confirmed cases. Let’s validate this hypothesis by looking at the proportion of different races proportions for CA and NY.



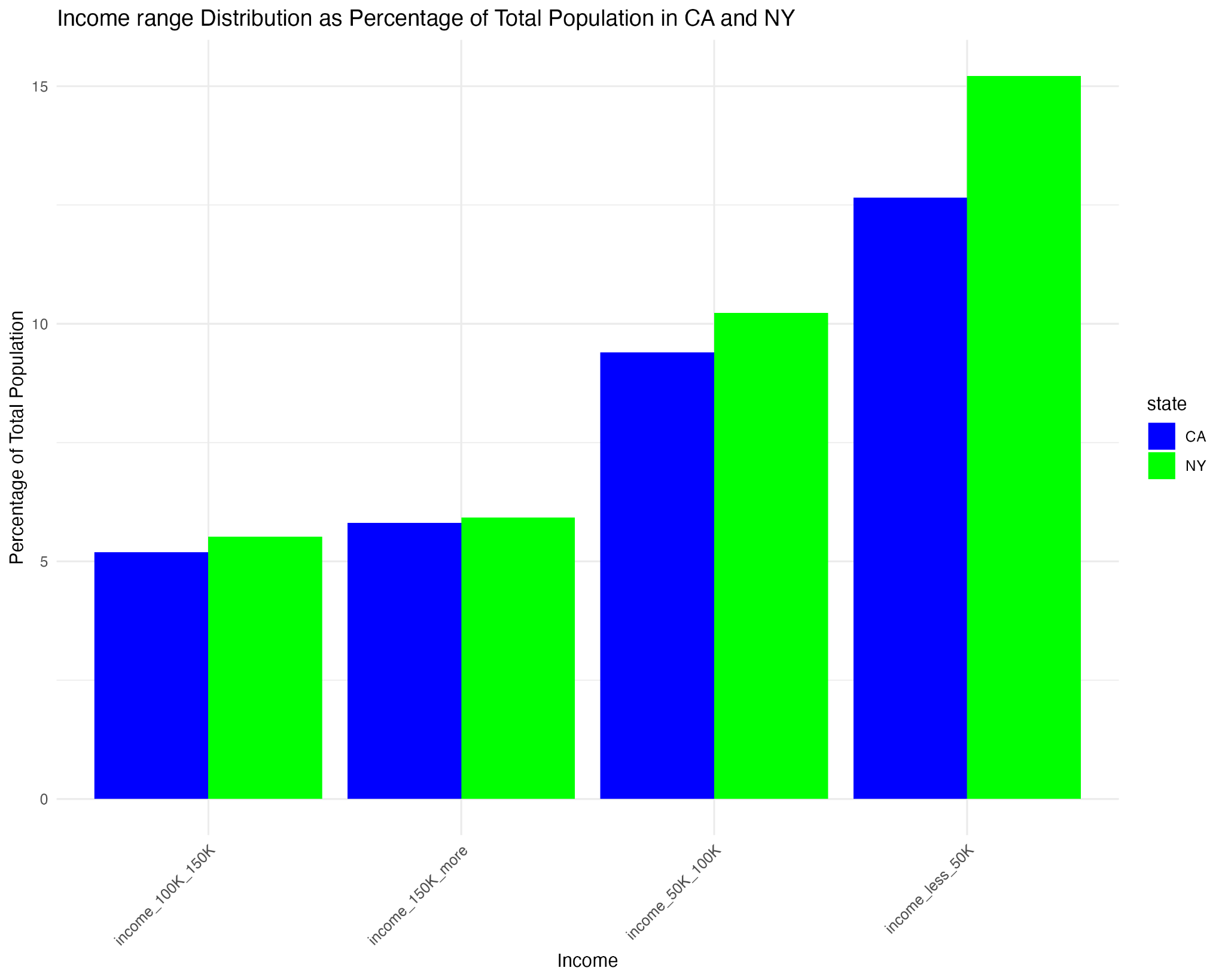
**Fig 3.3.7**: Percentages of population (per 1000) of various race (NY & CA)

The bar chart above we can see that there are more Hispanic people in CA than in NY. This could be a reason where there were more confirmed cases in CA than in NY. Similarly, we can also see that there are more White people in NY than in CA. Since there is a strong negative correlation between white race and confirmed cases, this could be why there are less number of confirmed cases in NY than in CA. Overall, we could say that race is a good indicator for the confirmed cases for CA and NY.



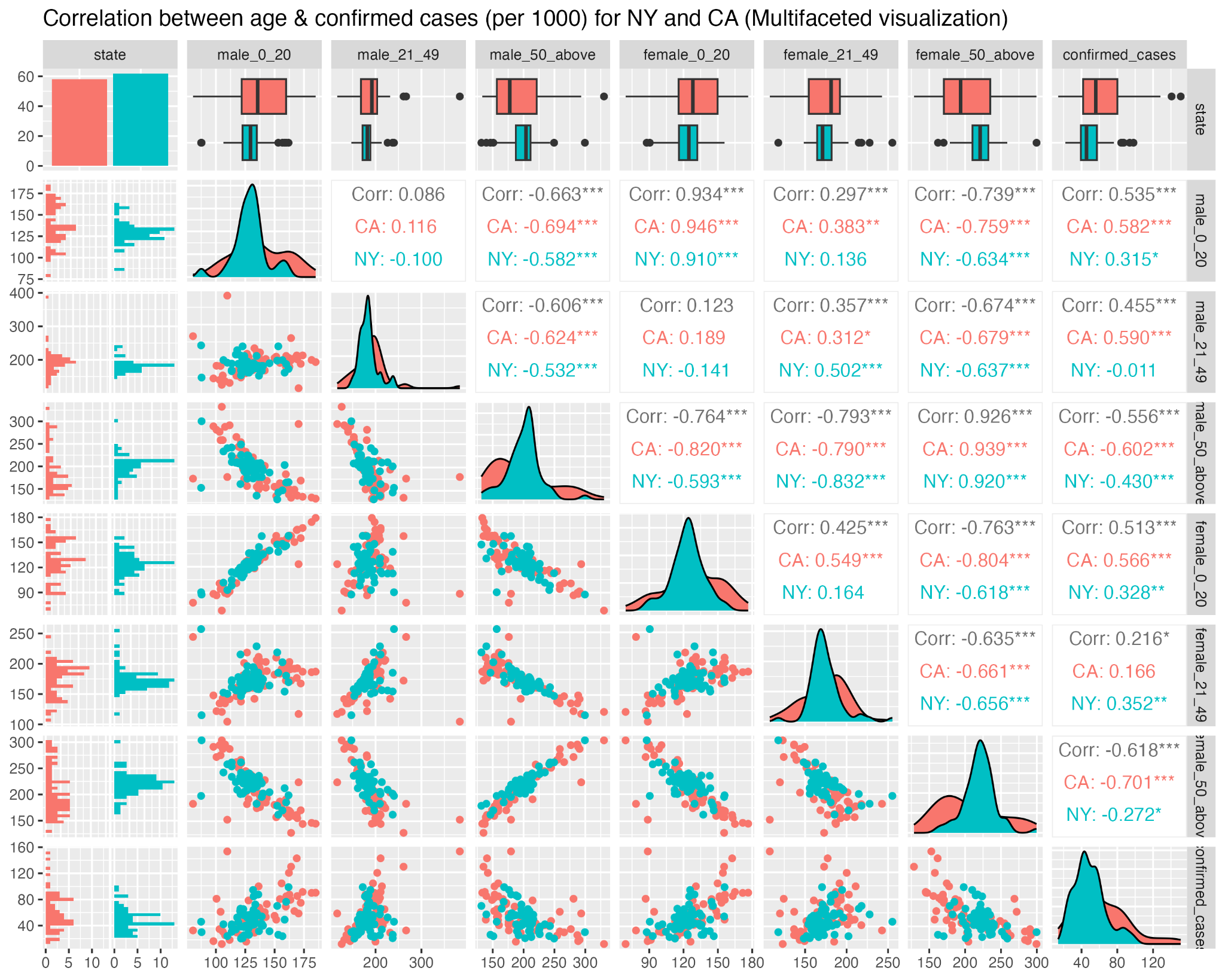
**Fig 3.3.8:** Correlation plot (income against confirmed cases) for NY and CA per 1000 (Multifaceted Visualization)

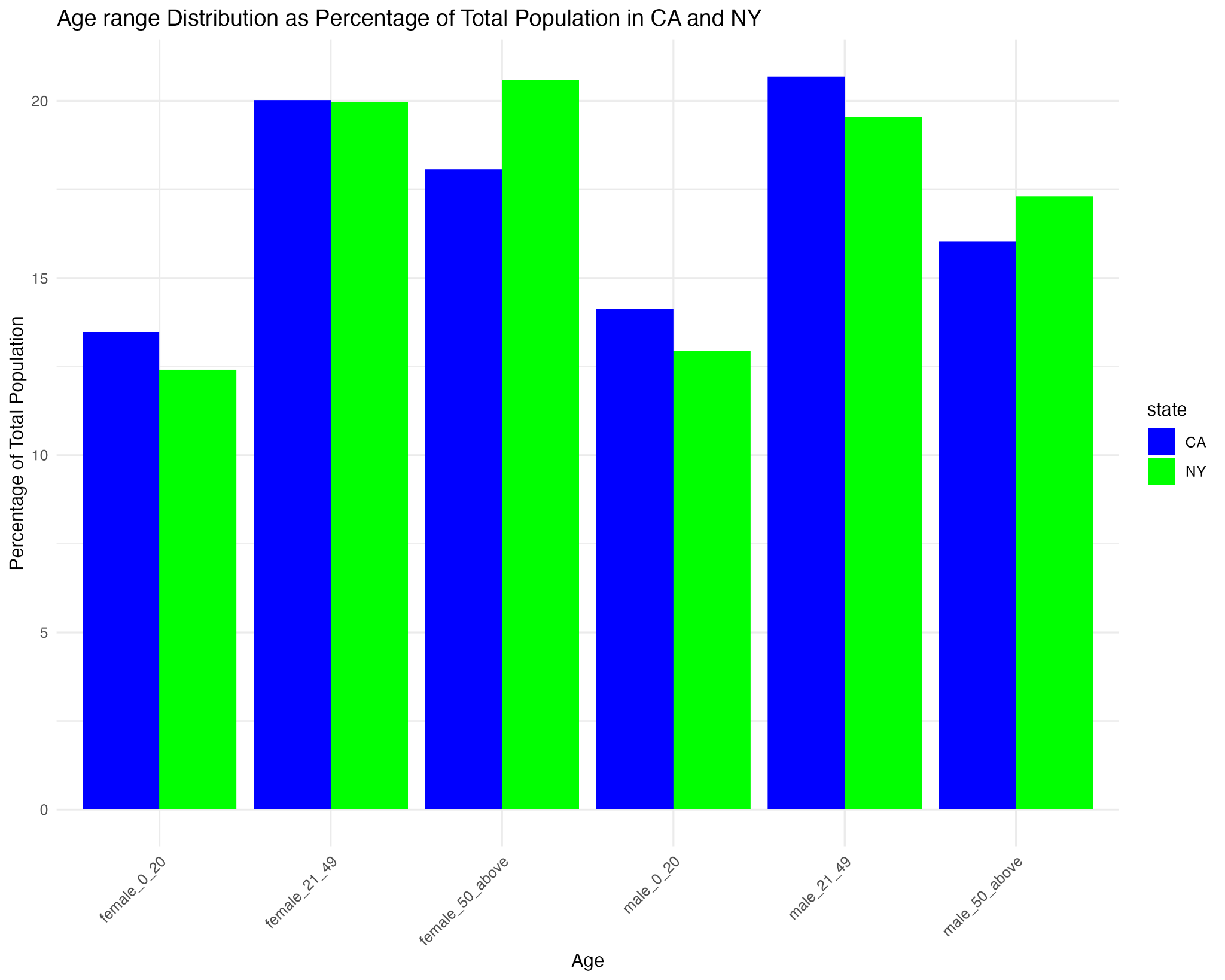
From the above correlation plot, we case that the income of 150K and more has the strongest positive correlation with confirmed cases. Likewise, the income of 50-100K has the strongest negative correlation with the confirmed cases. This means that if there are more people making between 50-100K in CA, it could be the indicator of its higher number of confirmed cases. We are emphasizing more on the negative correlation value since the people making less income generally are more susceptible to being hit by a pandemic.



**Fig 3.3.9**: Percentages of population (per 1000) of various income ranges (NY & CA)

From the bar chart above, we can see that there are fewer people making between 50-100K in CA than in NY. Previously we saw that there is a strong negative correlation between confirmed cases and this income range. This could be why there are more confirmed cases seen in CA than in NY. Thus, we could say that income range is a good indicator of the confirmed cases in NY and CA.





**Fig 3.3.10**: Correlation between age and confirmed cases per 1000 (CA & NY)

From the correlation plot, we can see that the age group 0-20 has the strongest positive correlation for both males and females. We can see that there are more males and females between 0-20 yrs in CA than in NY from the bar chart. Thus, we can say that this age group is a good indicator for the number of confirmed cases in CA and NY.





**Fig 3.3.11**: Correlation between rent spending and confirmed cases

From the correlation plot above, we can see that the variable “rent\_over\_50\_percent” has the stronger correlation for confirmed cases. This should mean that there are more people spending over 50% on rent in CA than in NY to confirm that there are more confirmed cases in CA than in NY. But we can also see that there are more people spending over 50% on rent in NY than in CA. Similar observations can be made if we consider negative correlation. Thus, we can say that rent is not a determining factor for the confirmed cases in these states.

## DataSet 4:- COVID Vaccination Report for TX

### Feature Processing(Dataset 4)

The below cleaning and feature processing was performed for the Dataset 4

### Data Analysis( Dataset 4)

We tried to look at the Dataset 4 and join the Dataset 1 and county map data to perform some analysis and answer some questions based on our understanding of the data.

#### How is the vaccination trend looking for TX?

The below graph shows the vaccination trend for TX state. The line shows the cumulative vaccination trend for the applicable dates. The chart includes the rends for first dose, series completion dose and the Booster dose

A graph of a covid-19 vaccination

Description automatically generated

**Fig 3.4.2.1:** COVID Vaccinations Weekly Trend

This graph shows that the trend for booster shots is far much lesser than the initial and the series complete doses. This probably indicates that the not everyone is taking the Booster shots as the

#### How is the Booster Dose per thousand spread across TX county?

Since the data is cumulative, We need to take the data from the latest date to make sure that we are not overcounting. This vaccination data is joined with the map data to plot the same in a map.

A green and white map

Description automatically generated

**Fig 3.4.2.2:** Booster Doses per 1000 in TX Counties

With this we can find that Booster doses have only been taken by maximum around 400 per thousand. That is close to 40-45% of the population.

#### Is there any relation between Booster Vaccines and Income per Capita?

A graph with green dots and numbers

Description automatically generated

**Fig 3.4.2.2:** Booster Doses per 1000 with Income per capita

From the above graph, It seems like there is no correlation between the Income Per Capita and the Booster Shot per 1000. One of the reasons for this is that COVID vaccines were provided by the US government for free which enabled everyone irrespective of their income to get vaccinated.

#### Is there any relationship between the Death percentage and the Booster Shots?

A graph with green dots and numbers

Description automatically generated

**Fig 3.4.2.2:** Booster Doses per 1000 with Death Percentage

From the above graph we can understand that most of the counties with the highest death percentage are in the lower range of Booster shot. We can infer that the Booster shots have reduced deaths due to COVID

# Modeling and Evaluation

[**Modeling and Evaluation**: What type of model do we apply to the data?

Describe why you chose the particular model, model assumption and limitations, what variable you use for the model, and how well the model works. ]

# Recommendations

[Deployment: Describe how to interpret the model and what **recommendations** you can make based on the findings. How would the stakeholder use the findings and why is the recommendation useful to the stakeholder.]

# Conclusion

[Does the project answer the initial questions? Repeat the key findings and why they are important.]

# List of References

Google. (n.d.). *Global mobility report*. Retrieved October 1, 2024, from <https://www.google.com/covid19/mobility/index.html>

U.S. Department of Health and Human Services. (2024). *U.S. spending on COVID-19 as of October 2024*. Retrieved October 1, 2024, from USASPENDING.gov: The Federal Response to COVID-19.

USAFacts. (n.d.). COVID-19 case and death counts by state and county. In USAFacts Public Data - COVID-19 US Cases. Retrieved from

<https://console.cloud.google.com/bigquery?p=bigquery-public-data&d=covid19_usafacts&page=dataset&project=crucial-cycling-338005&ws=!1m4!1m3!3m2!1sbigquery-public-data!2scovid19_usafacts>

Centers for Disease Control and Prevention. (n.d.). COVID-19 vaccinations in the United States, county. Retrieved from<https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data>

# Appendix

tx\_covid\_cases <- **subset**(tx\_covid\_cases,county\_name **!=**'Statewide Unallocated')   
tx\_covid\_cases**$**date <- **as.Date**(tx\_covid\_cases**$**date,format = "%Y-%m-%d")  
**options**(max.print=10)  
tx\_covid\_cases

## county\_fips\_code county\_name state state\_fips\_code date  
## 371 48001 Anderson County TX 48 2020-01-22  
## confirmed\_cases deaths  
## 371 0 0  
## [ reached 'max' / getOption("max.print") -- omitted 93979 rows ]

Summary of the COVID cases and Death by County

tx\_covid\_cases\_county = tx\_covid\_cases **%>%** **group\_by**(county\_name) **%>%**  
 **summarise**(total\_confirmed\_cases = **sum**(confirmed\_cases),   
 total\_deaths = **sum**(deaths),   
 .groups = 'drop')  
tx\_covid\_cases\_county**$**death\_perc=tx\_covid\_cases\_county**$**total\_deaths**\***100**/**tx\_covid\_cases\_county**$**total\_confirmed\_cases   
tx\_covid\_cases\_county <- tx\_covid\_cases\_county[**order**(tx\_covid\_cases\_county**$**death\_perc,  
 decreasing = TRUE), ]

**Appendix A**: Data preprocessing for dataset “COVID-19\_cases\_TX.csv”

| **State Level Fips Code** | **State** |  | **State Level Fips Code** | **State** |  | **State Level Fips Code** | **State** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | ALABAMA |  | 28 | MISSISSIPPI |  | 54 | WEST VIRGINIA |
| 2 | ALASKA |  | 29 | MISSOURI |  | 55 | WISCONSIN |
| 4 | ARIZONA |  | 30 | MONTANA |  | 56 | WYOMING |
| 5 | ARKANSAS |  | 31 | NEBRASKA |  |  |  |
| 6 | CALIFORNIA |  | 32 | NEVADA |  |  |  |
| 8 | COLORADO |  | 33 | NEW HAMPSHIRE |  |  |  |
| 9 | CONNECTICUT |  | 34 | NEW JERSEY |  |  |  |
| 10 | DELAWARE |  | 35 | NEW MEXICO |  |  |  |
| 11 | DISTRICT OF COLUMBIA |  | 36 | NEW YORK |  |  |  |
| 12 | FLORIDA |  | 37 | NORTH CAROLINA |  |  |  |
| 13 | GEORGIA |  | 38 | NORTH DAKOTA |  |  |  |
| 15 | HAWAII |  | 39 | OHIO |  |  |  |
| 16 | IDAHO |  | 40 | OKLAHOMA |  |  |  |
| 17 | ILLINOIS |  | 41 | OREGON |  |  |  |
| 18 | INDIANA |  | 42 | PENNSYLVANIA |  |  |  |
| 19 | IOWA |  | 44 | RHODE ISLAND |  |  |  |
| 20 | KANSAS |  | 45 | SOUTH CAROLINA |  |  |  |
| 21 | KENTUCKY |  | 46 | SOUTH DAKOTA |  |  |  |
| 22 | LOUISIANA |  | 47 | TENNESSEE |  |  |  |
| 23 | MAINE |  | 48 | TEXAS |  |  |  |
| 24 | MARYLAND |  | 49 | UTAH |  |  |  |
| 25 | MASSACHUSETTS |  | 50 | VERMONT |  |  |  |
| 26 | MICHIGAN |  | 51 | VIRGINIA |  |  |  |
| 27 | MINNESOTA |  | 53 | WASHINGTON |  |  |  |

A**ppendix B**: State Level Fips Code

## 7.1 Student Contributions

Add a list with who contributed to what part of this report.