

“US Counties: COVID19 + Weather + Socio/Health data” Report in R

1. Introduction to the dataset

The United States consists of 3,142 counties, each with its unique social, economic, health, and weather characteristics. Among these counties, more than 2,400 have already reported cases of COVID-19 due to the ongoing pandemic. By combining data on health, socioeconomic factors, and weather at the county level, we can identify populations that are particularly vulnerable to COVID-19 and take proactive measures to support high-risk communities. Temperature and humidity potentially impact the transmission of COVID-19. However, in the United States, regions with higher temperatures also exhibit distinct socioeconomic and health demographics. Therefore, it is crucial to account for factors such as obesity, diabetes, access to healthcare, and poverty rates. These factors themselves likely contribute to the transmission and fatality rates of COVID-19

2. Exploratory Data Analysis

We have 3 types of data, namely :

- Covid 19
- SocioHealth
- Geometry data

let's do EDA for the following 3 datasets, then we will select 1 data among these 3 data for further analysis.

Data	
data_Covid19	790331 obs. of 227 variables
data_Geometry	3142 obs. of 7 variables
data_SocioHealth	3144 obs. of 181 variables

2.1 Analysis Covid-19 Dataset

- Data Set

Data	
data_Covid19	790331 obs. of 227 variables

Data Insight :

Dataset Covid-19 include 79.0331 sample dan 227 variable

- Data Head

date <date>	county <chr>	state <chr>	fips <chr>	cases <dbl>	deaths <dbl>	stay_at_home_announced <chr>
2020-01-21	Snohomish	Washington	53061	1	0	no
2020-01-22	Snohomish	Washington	53061	1	0	no
2020-01-23	Snohomish	Washington	53061	1	0	no
2020-01-24	Cook	Illinois	17031	1	0	no
2020-01-24	Snohomish	Washington	53061	1	0	no
2020-01-25	Orange	California	06059	1	0	no

Data Insight :

Here is the data content looks like as we can see.

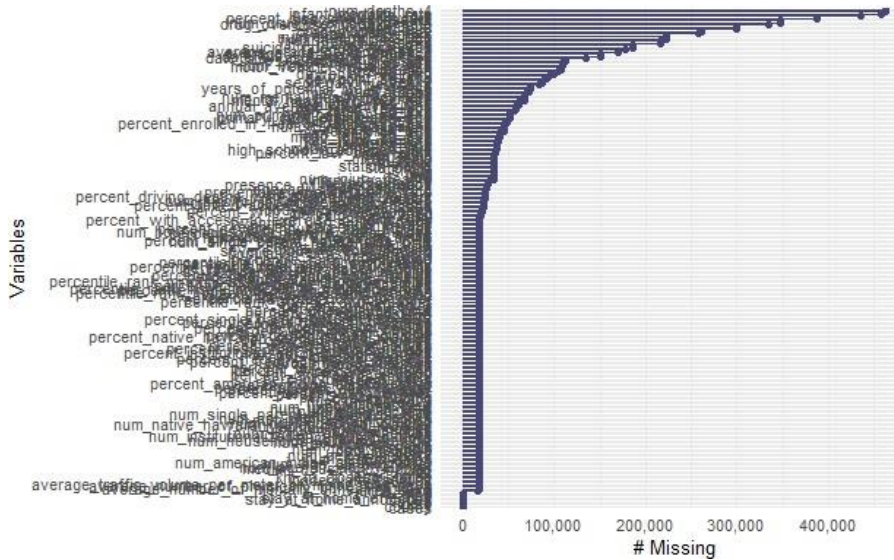
- **Checking Duplicated Data**

```
[1] 0
```

Data Insight :

There is no duplicated data in this dataset. Now let's check Missing Value.

- **Check Missing Values**



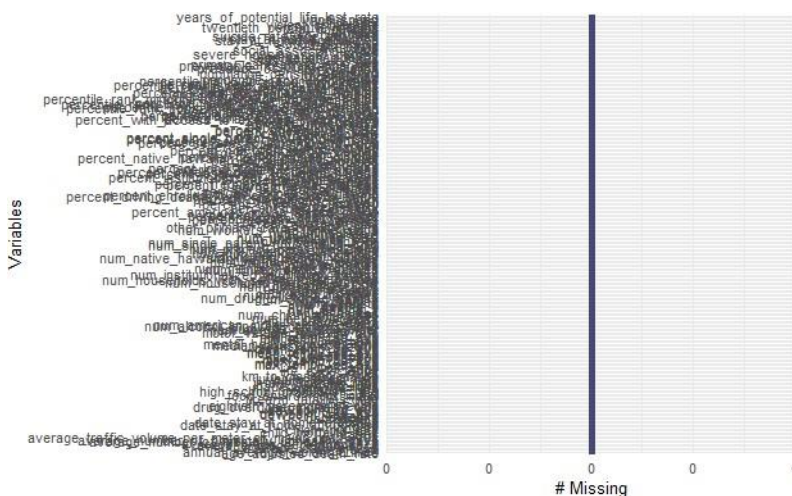
```
[1] 12225014
```

Data Insight :

- From this visualization we can conclude that there is a lot of missing values in this data set.
- The number of missing values is 12.225.014

- **Dealing With Missing Values**

Because there are so many missing values, we will clean them with `na.omit`, here are the results

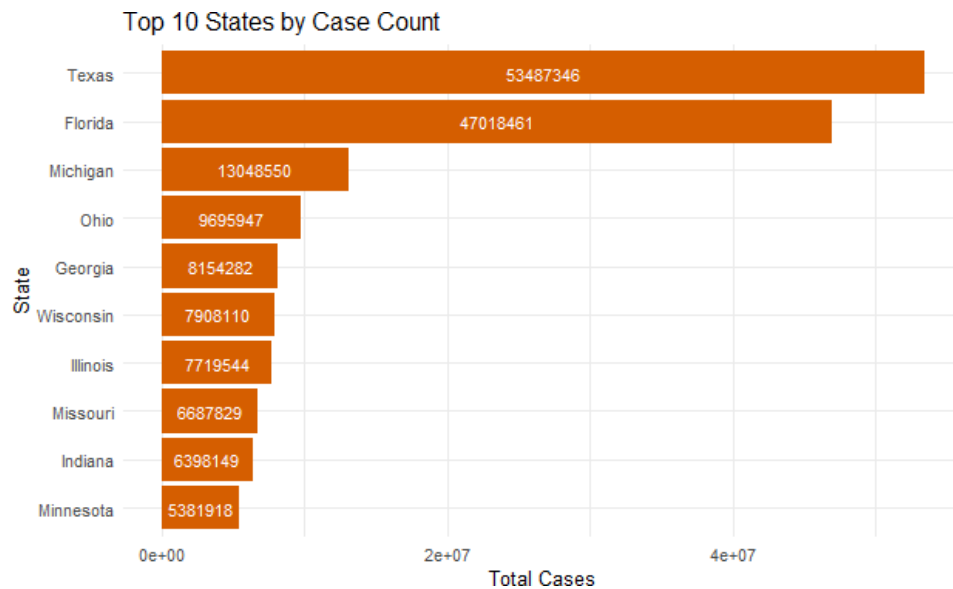


```
[1] 0
```

Data insight :

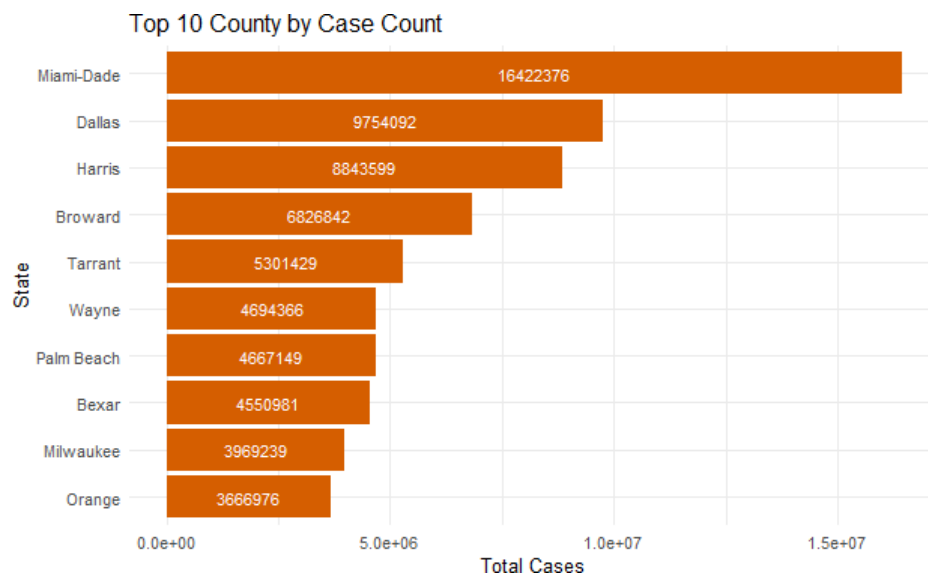
- After cleaning it can be seen from the results above that the missing values has been gone, then I will explore the dataset for futher anlysis with new dataset namely “ Covid19_clean”

- **Visualization**



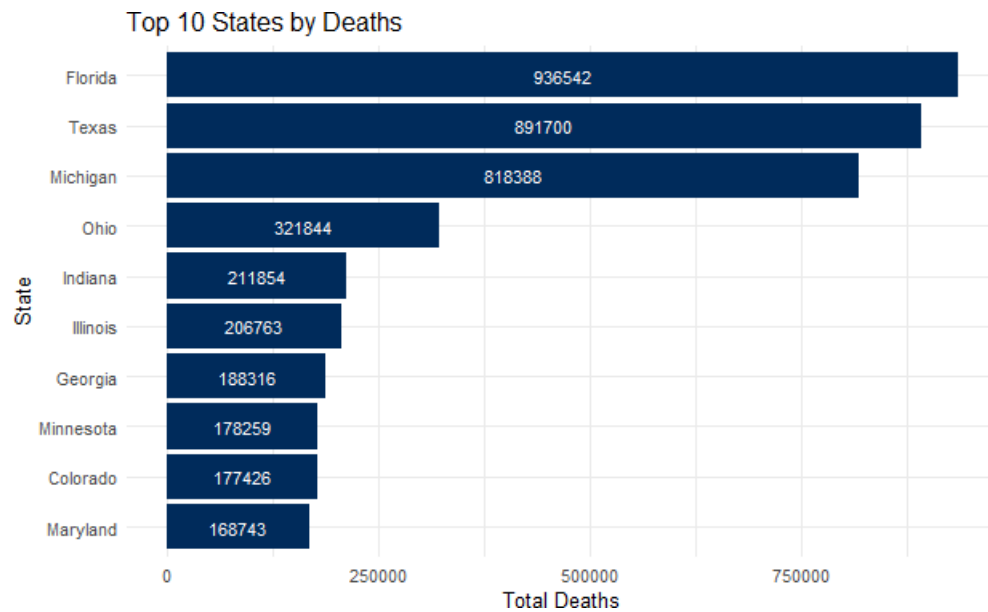
Data Insight :

As we can see top 10 most highest case based on states is lead by texas with total 53487346 case in total.



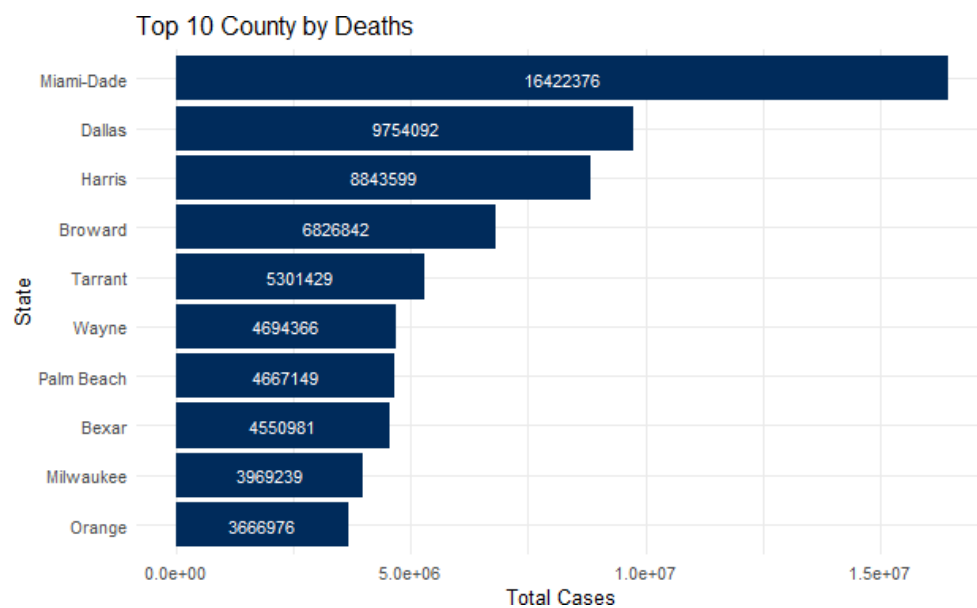
Data Insight :

As we can see top 10 most highest case based on county is lead by Miami-dade with total 16422376 case in total.



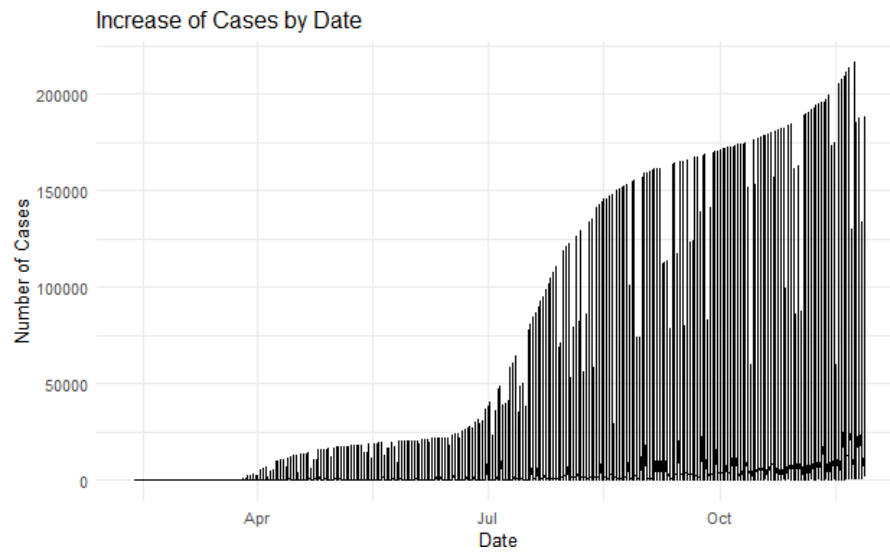
Data Insight :

As we can see top 10 most highest death rate based on states is lead by florida with total 936.542 deaths case.



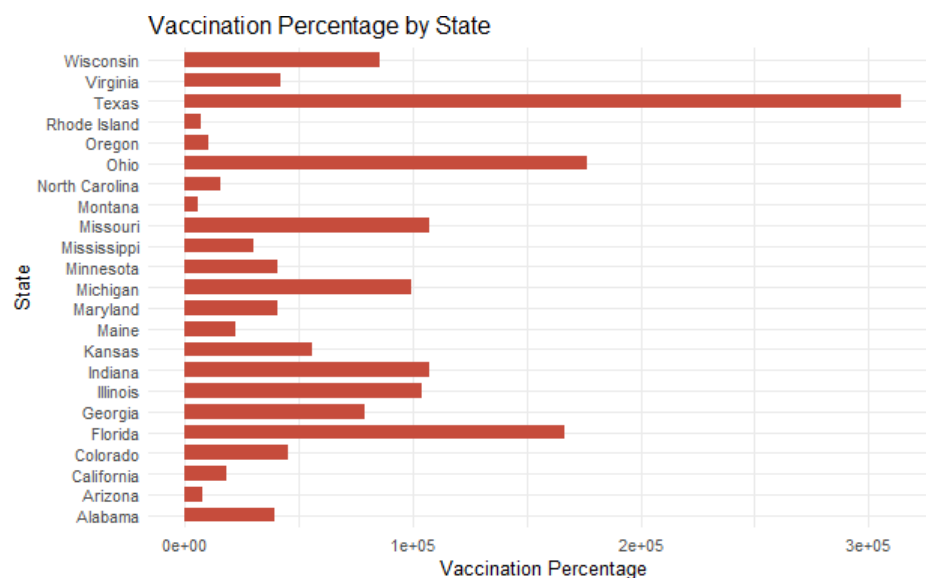
Data Insight :

As we can see top 10 most highest death rate based on states is lead by Miami-Dade with total 16.422.376 deaths case.



Data Insight :

From what we can see that cases started in April and tended to increase steadily until October with more than 2000 cases.




Data Insight :

Most state that have been vaccinated is “Texas “

2.2 Data Geometry

- **Data Set**

 **data_Geometry** | 3142 obs. of 7 variables

Data Insight :

Dataset Covid-19 include 3142 sample dan 7 variable

- **Data Head**

state <chr>	county <chr>	fips <chr>
ALABAMA	Autauga	01001
ALABAMA	Blount	01009
ALABAMA	Chambers	01017
ALABAMA	Coffee	01031
ALABAMA	Colbert	01033
ALABAMA	Covington	01039

Data Insight :

Here is the data content looks like as we can see, this data set almost have the same format as Data Covid 19 but the the different is this dataset having less column and sample.

- **Checking Duplicated Data**

```
[1] 0
```

Data Insight :

There is no duplicated data in this dataset. Now let's check Missing Value.

- **Check Missing Values**

```
[1] 0
```

Data Insight :

There is ni Missing values in this dataset

2.3 Data SocioHealth

- **Data Set**

```
data_SocioHealth 3144 obs. of 181 variables
```

Data Insight :

Dataset Covid-19 include 3144 sample dan 181 variable

- **Data Head**

fips <chr>	state <chr>	county <chr>	lat <dbl>	lon <dbl>	total_population <dbl>	area_sqmi <dbl>
01001	Alabama	Autauga	32.53493	-86.64275	55049	594.4461
01003	Alabama	Baldwin	30.72749	-87.72258	199510	1589.8074
01005	Alabama	Barbour	31.86959	-85.39321	26614	884.8758
01007	Alabama	Bibb	32.99863	-87.12648	22572	622.5824
01009	Alabama	Blount	33.98088	-86.56738	57704	644.8065
01011	Alabama	Bullock	32.10053	-85.71569	10552	622.8054

Data Insight :

Here is the data content looks like as we can see, this data set almost have the same format as the two dataset before.

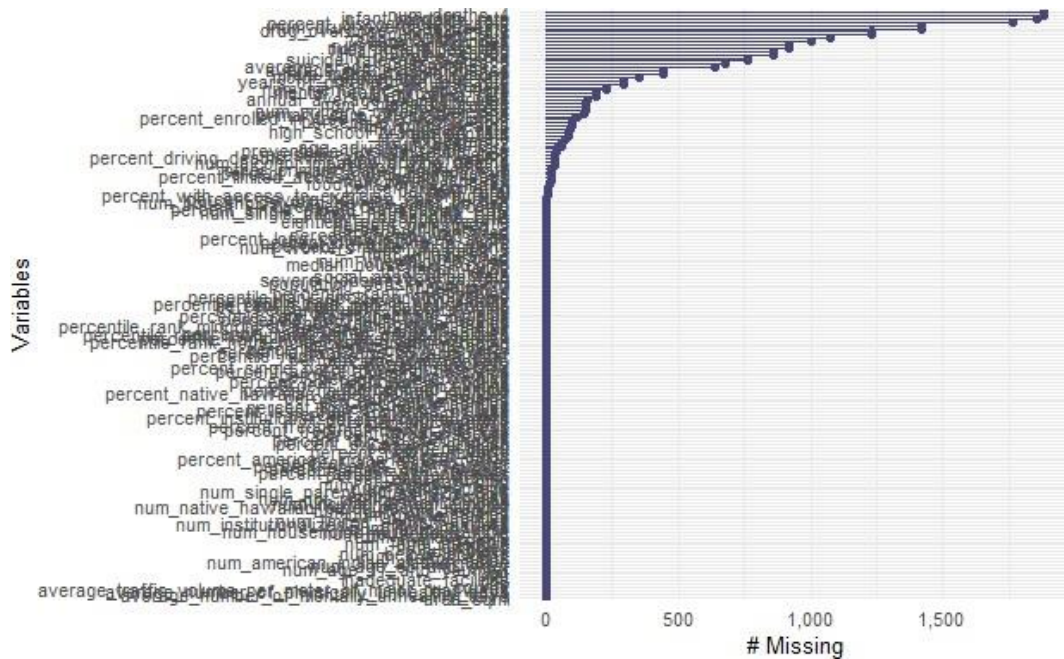
- **Checking Duplicated Data**

```
[1] 0
```

Data Insight :

There is no duplicated data in this dataset. Now let's check Missing Value.

- **Check Missing Values**



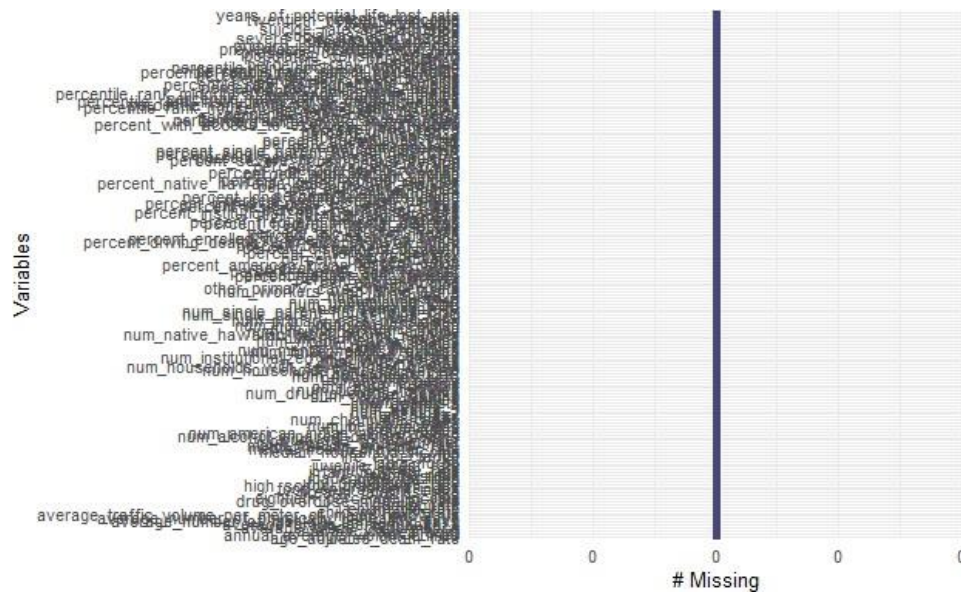
```
[1] 25871
```

Data Insight :

- From this visualization we can conclude that there is a lot of missing values in this data set.
- The number of missing values is 25.871

- **Dealing With Missing Values**

Because there are so many missing values, we will clean them with `na.omit`, here are the results



[1] 0

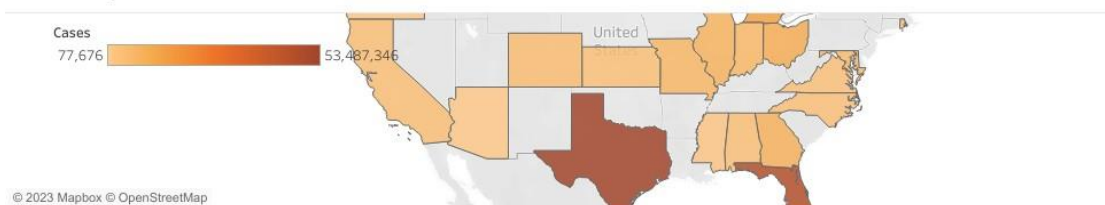
Data insight :

- After cleaning it can be seen from the results above that the missing values has been gone, then I will explore the dataset for futher anlysis with new dataset namely “SocioHealth_clean”

Conclsn

Based on the exploratory data analysis above on the 3 datasets that I have done, here I choose the Covid 19 dataset to be visualized using tableau, here is the result :

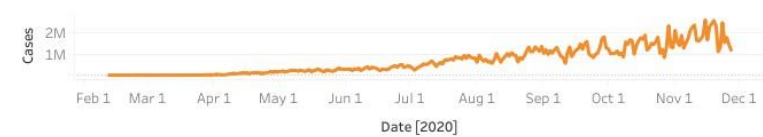
Cases Map Distribution



Confirm Cases

195,693,032

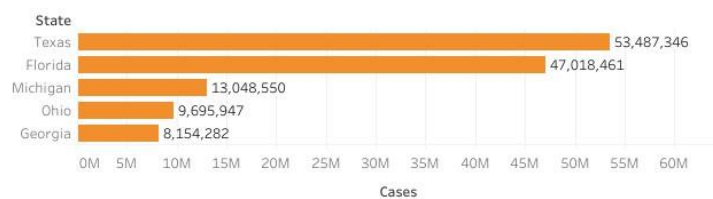
Increase Cases in a Year



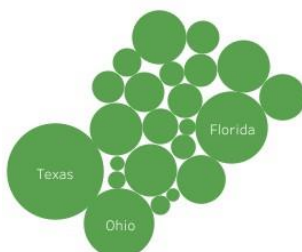
Confirm Deaths

4,821,640

Top 5 Cases By State



Most Vaccinated States



Increase Deaths in a Year



