Explanatory Data Analysis

US Counties: COVID19 + Weather + Socio/Health

ABOUT DATASET

Dataset is from the following link: https://www.kaggle.com/datasets/johnjdavisiv/us-counties-covid19-weather-sociohealth-data. This dataset contains three files that show us county-level data on health, socioeconomics, and weather. This dataset can help us identify which populations are at risk for COVID-19 and help prepare high-risk communities. By understanding the risk factors for COVID-19 in each county, we can better target our efforts to prevent the spread of the virus and protect the most vulnerable people.

First step we will load datasets from 3 different .csv and observe the data using the head() function

```
library(readr)
health_weather <- read_csv("/Users/shafaqonitatingmail.com/Documents/Semester 4/Data Mining and Visualization/Assignment/LAB/USCounties_COVID19_Weather_SocioHealth/archive/US_counties_COVID19_health_weather_data.csv", show_col_types = FALSE)
head(health_weather
```

A tibble: 6 × 227

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

6 rows | 1-7 of 227 columns

```
sociohealth <- read_csv("/Users/shafaqonitatingmail.com/Documents/Semester 4/Data Mining and Visualization/Assignment/LAB/USCounties_COVID19_Weather_SocioHealth/archive/us_county_sociohealth_data.csv", show_col_types = FALSE)
head(sociohealth)
```

A tibble: 6 × 181

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

6 rows | 1-7 of 181 columns

```
geometry <- read_csv("/Users/shafaqonitatingmail.com/Documents/Semester 4/Data
Mining and
Visualization/Assignment/LAB/USCounties_COVID19_Weather_SocioHealth/archive/us_coun
ty_geometry.csv", show_col_types = FALSE)
head(geometry)
```

A tibble: 6 × 7

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

6 rows | 1-3 of 7 columns

Now we print the dimensions of our dataset

```
print(dim(health_weather))
print(dim(sociohealth))
print(dim(geometry))
```

```
[1] 790331    227
[1] 3144  181
[1] 3142    7
```

We can see that the dimensions owned by our datasets are very many, especially in the health_weather dataset. Therefore, because we need to visualize it in Tableau Public, where Tableau Public needs a maximum dataset of only 1 GB, we delete all null values owned by our three datasets.

```
# delete null values
health_weather <- na.omit(health_weather)
sociohealth <- na.omit(sociohealth)
geometry <- na.omit(geometry)
```

```
print(dim(health_weather))
print(dim(sociohealth))
print(dim(geometry))head(sociohealth)
```

```
[1] 34307   227
[1] 427 181
[1] 3142    7
```

And finally we get dimensions that are much smaller than before, the next step is we will save our dataset into a .csv file so we can use it in Tableau.

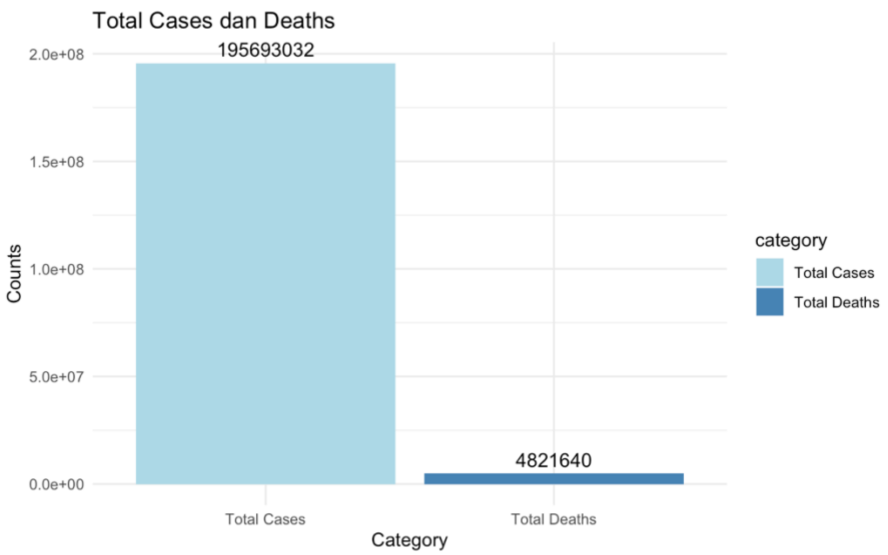
```
write.csv(health_weather, "/Users/shafaqonitatingmail.com/Documents/Semester 4/Data
Mining and
Visualization/Assignment/LAB/USCounties_COVID19_Weather_SocioHealth/health_weather.
csv", row.names = FALSE)

write.csv(sociohealth, "/Users/shafaqonitatingmail.com/Documents/Semester 4/Data
Mining and
Visualization/Assignment/LAB/USCounties_COVID19_Weather_SocioHealth/socio_health",
row.names = FALSE)

write.csv(geometry, "/Users/shafaqonitatingmail.com/Documents/Semester 4/Data
Mining and
```

```
Visualization/Assignment/LAB/USCounties_COVID19_Weather_SocioHealth/geometry",
row.names = FALSE)
```

Because this dataset is generally about COVID-19, the most important thing is to know how many cases and the number of deaths

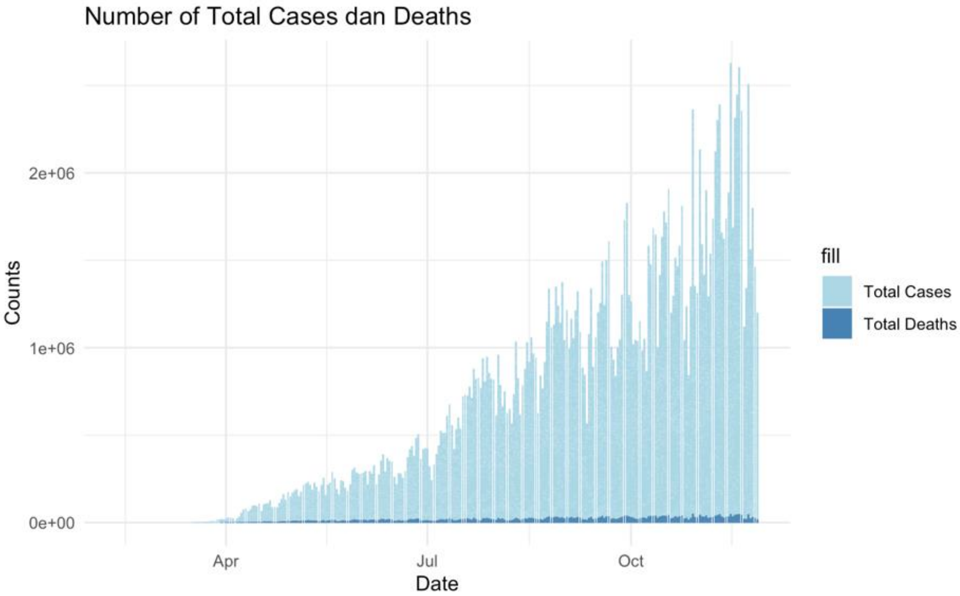
```
total_cases <- sum(health_weather$cases)
total_deaths <- sum(health_weather$deaths)
```

```
library(ggplot2)

data <- data.frame(
  category = c("Total Cases", "Total Deaths"),
  sums = c(total_cases, total_deaths)
)

ggplot(data, aes(x = category, y = sums, fill = category)) +
  geom_bar(stat = "identity") +
  labs(title = "Total Cases dan Deaths", x = "Category", y = "Counts") +
  theme_minimal() +
  geom_text(aes(label = sums), vjust = -0.5, size = 4) +
  scale_fill_manual(values = c("lightblue", "steelblue"))
```



The plot shows the distribution of Total Cases and Total Deaths, where only about 0.02% of people died from all COVID-19 cases in the US.

```
health_weather$date <- as.Date(health_weather$date)

ggplot(health_weather, aes(x = date)) +
  geom_bar(aes(y = cases, fill = "Total Cases"), stat = "identity") +
  geom_bar(aes(y = deaths, fill = "Total Deaths"), stat = "identity") +
  labs(title = "Number of Total Cases dan Deaths", x = "Date", y = "Counts") +
  scale_fill_manual(values = c("Total Cases" = "lightblue", "Total Deaths" =
"steelblue")) +
  theme_minimal()
```



Obtained from the plot above, the distribution of Total Cases and Total Deaths during the period February to November 2020 and we can see that at a glance the distribution of COVID-19 continues to increase during this period.

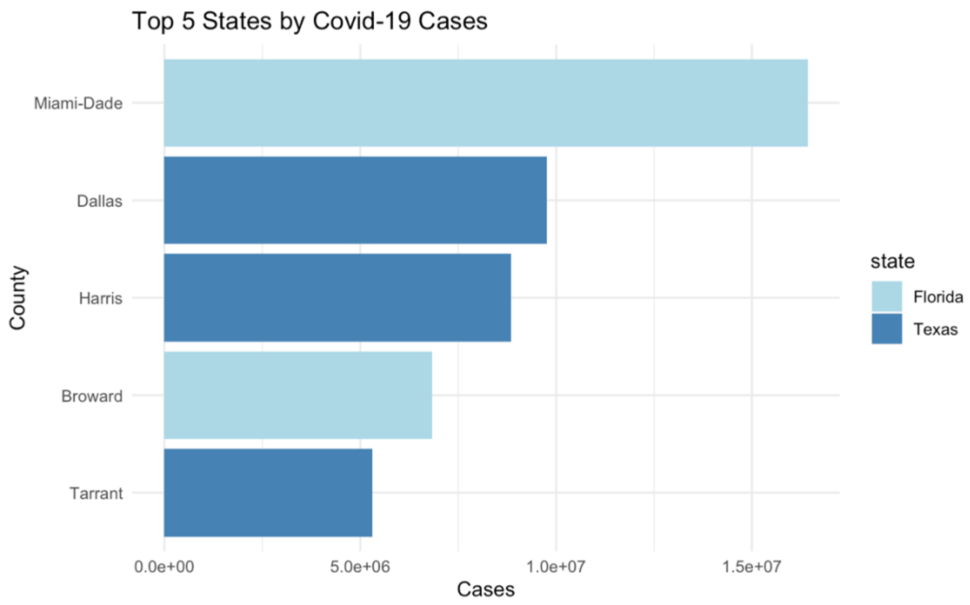
```
cases_counties <- aggregate(cases ~ state + county , data = health_weather, sum)
print(cases_counties)
```

Description: df [223 × 3]

state <chr>	county <chr>	cases <dbl>
Colorado	Adams	1153896
Illinois	Adams	106403
Florida	Alachua	616072
Michigan	Allegan	122796
Indiana	Allen	689132
Ohio	Allen	154048
Texas	Angelina	165024
Colorado	Arapahoe	1234270
Virginia	Arlington	377333
Ohio	Ashtabula	77199

1-10 of 223 rows Previous 1 2 3 4 5 6 ... 23 Next

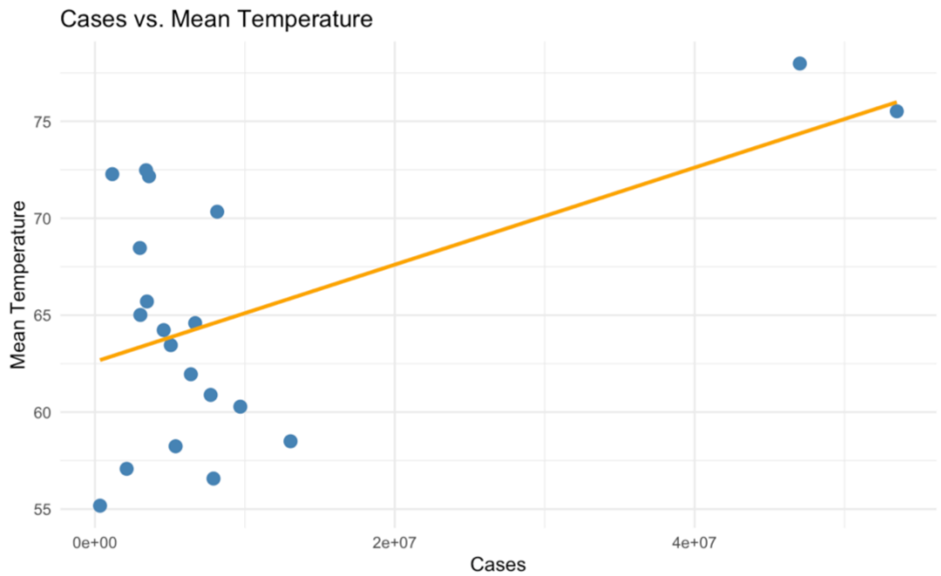
```
cases_counties <- cases_counties[order(-cases_counties$cases), ]
top5 <- head(cases_counties, 5)
top5$county <- factor(top5$county, levels = top5$county[order(top5$cases)])

ggplot(data = top5, aes(x = cases, y = county, fill=state)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 5 States by Covid-19 Cases", x = "Cases", y = "County") +
  scale_fill_manual(values = c("lightblue", "steelblue"))
```



And we can know together, that the top 5 counties are held by Miami-Dade with 16,422,376 cases, Dallas with 9,754,092 cases, Harris with 8,843,599 cases, Boward with 6,826,842 cases, and Tarrant with 5,301,429 cases. It can also be noted that 2 of the 5 counties are from the State of Florida and the rest are from the State of Texas.

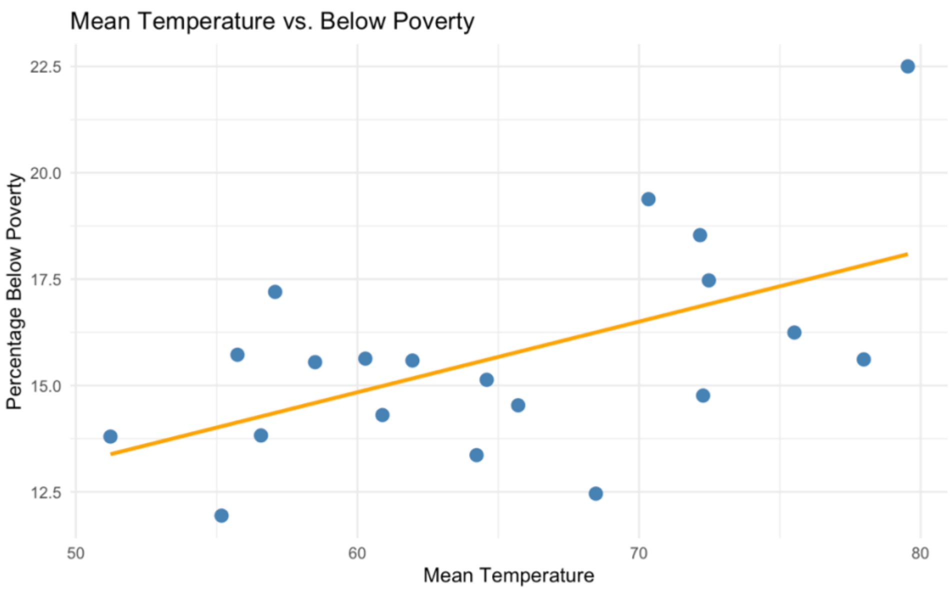
People have speculated about the effects of temperature on the spread of COVID19, motivated by the fact that other coronaviruses are much more prevalent during the winter than the summer. Is it True? Now, we want to plot to observe the relationship between two variables, namely Cases and Mean Temp

```
library(dplyr)

top_state <- health_weather %>%
  group_by(state) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE),
            mean_temp = mean(mean_temp, na.rm = TRUE)) %>%
  top_n(20, total_cases)

plot <- ggplot(top_state, aes(x = total_cases, y = mean_temp)) +
  geom_point(color = "steelblue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(x = "Cases", y = "Mean Temperature", title = "Cases vs. Mean Temperature") +
  theme_minimal()

print(plot)
```



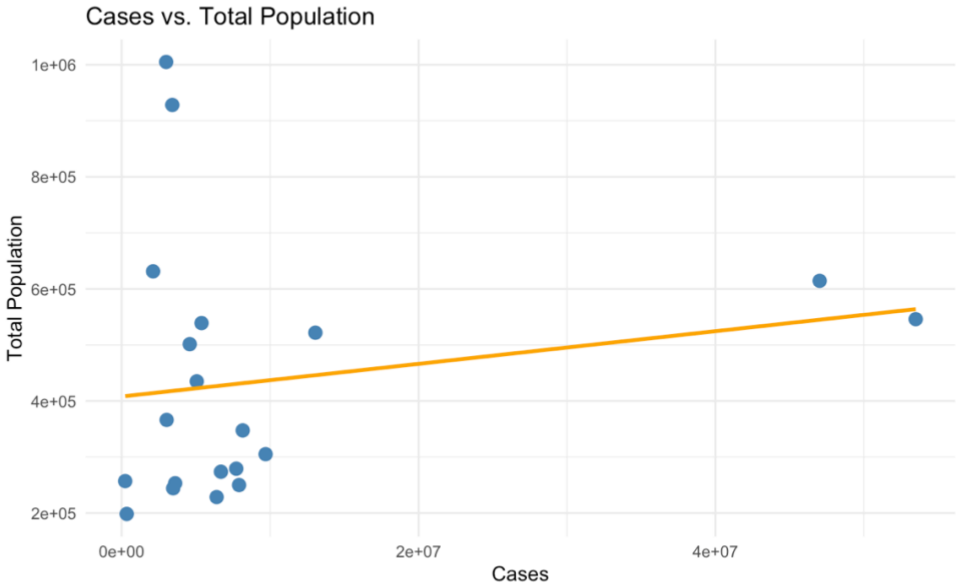
In the plot above, it is clear that there is a moderate but positive correlation between temperature and COVID-19 cases. But if we drill down deeper, this could also be because it turns out that warmer parts of the US (e.g. the Deep South) tend to have very different social, economic and health profiles to colder regions, such as the Pacific Northwest, Midwest and East Coast.

```
top_state <- health_weather %>%
  group_by(state) %>%
  summarize(mean_temp = mean(mean_temp, na.rm = TRUE),
            below_poverty = mean(percent_below_poverty, na.rm = TRUE)) %>%
  top_n(20)

plot <- ggplot(top_state, aes(x = mean_temp, y = below_poverty)) +
  geom_point(color = "steelblue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(x = "Mean Temperature", y = "Percentage Below Poverty", title = "Mean
Temperature vs. Below Poverty") +
  theme_minimal()

print(plot)
```



In the plot above, it is clear that there is a moderate but positive correlation between temperature and poverty rates. In hotter regions, it can be said that more people fall below the poverty.

```
top_state_population <- health_weather %>%
  group_by(state) %>%
  summarize(cases = sum(cases, na.rm = TRUE),
            total_population = mean(total_population, na.rm = TRUE)) %>%
  top_n(20)

plot <- ggplot(top_state_population, aes(x = cases, y = total_population)) +
  geom_point(color = "steelblue", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  labs(x = "Cases", y = "Total Population", title = "Cases vs. Total Population") +
  theme_minimal()

print(plot)
```



And also, there is also a relationship between Cases and Total Population. The more Total Population in a State, the more COVID-19 Cases there will be.
Now we will explore the risk factors in COVID-19, namely Diabetes, Obesity, and HIV.

```
percent_diabetes <- aggregate(percent_adults_with_diabetes ~ state, health_weather,
median)
percent_diabetes <-
percent_diabetes[order(percent_diabetes$percent_adults_with_diabetes), ]
percent_diabetes$percent_adults_with_diabetes <-
round(percent_diabetes$percent_adults_with_diabetes)
percent_diabetes <- percent_diabetes[, c("state", "percent_adults_with_diabetes")]
percent_diabetes <- percent_diabetes[rev(seq_len(nrow(percent_diabetes))), ]
percent_diabetes$percent_adults_with_diabetes <-
as.integer(percent_diabetes$percent_adults_with_diabetes)

print(percent_diabetes)
```

Description: df [23 × 2]

	state <chr>	percent_adults_with_diabetes <int>
1	Alabama	14
18	Ohio	13
19	Oregon	12
14	Mississippi	12
11	Maryland	12
15	Missouri	11
5	Florida	11
8	Indiana	11
6	Georgia	11
9	Kansas	11

1–10 of 23 rows   Previous 1 2 3 Next

From the output above, we can see that the highest percentage of diabetes is at 14% and is in Alabama State, where the total COVID-19 Cases in Alabama State are 3,603,940.


```
percent_obesity <- aggregate(percent_adults_with_obesity ~ state, health_weather,
median)
percent_obesity <-
percent_obesity[order(percent_obesity$percent_adults_with_obesity), ]
percent_obesity$percent_adults_with_obesity <-
round(percent_obesity$percent_adults_with_obesity)
percent_obesity <- percent_obesity[, c("state", "percent_adults_with_obesity")]
percent_obesity <- percent_obesity[rev(seq_len(nrow(percent_obesity))), ]

print(percent_obesity)
```

Description: df [23 × 2]

	state <chr>	percent_adults_with_obesity <dbl>
1	Alabama	37
11	Maryland	36
9	Kansas	36
14	Mississippi	35
2	Arizona	35
8	Indiana	34
12	Michigan	34
18	Ohio	33
7	Illinois	33
19	Oregon	33

1–10 of 23 rows   Previous 1 2 3 Next

From the output above, we can see that the highest percentage of obesity is at 37% and is still in Alabama State (same as the percentage of diabetes), where the total COVID-19 Cases in Alabama State is 3,603,940.

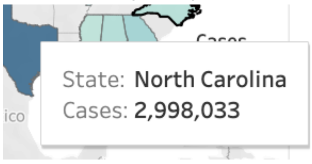
```
num_hiv <- aggregate(num_hiv_cases ~ state, health_weather, median)
num_hiv <- num_hiv[order(num_hiv$num_hiv_cases), ]
num_hiv$num_hiv_cases <- round(num_hiv$num_hiv_cases)
num_hiv <- num_hiv[, c("state", "num_hiv_cases")]
num_hiv <- num_hiv[rev(seq_len(nrow(num_hiv))), ]

print(num_hiv)
```

Description: df [23 × 2]

	state <chr>	num_hiv_cases <dbl>
17	North Carolina	3211
20	Rhode Island	1751
13	Minnesota	1301
3	California	1265
5	Florida	1025
6	Georgia	970
4	Colorado	958
22	Virginia	796
11	Maryland	469
21	Texas	408

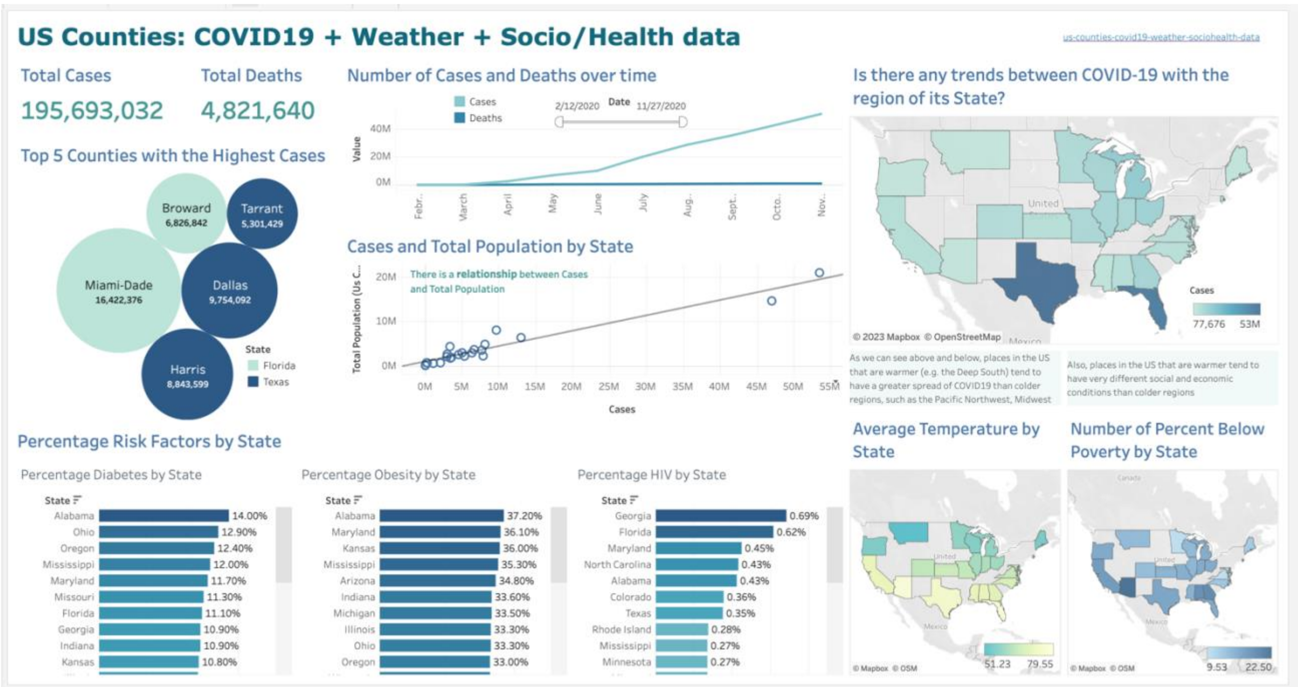
1–10 of 23 rows Previous 1 2 3 Next

From the output above, we can see that the highest percentage of obesity is at 37% and is still in Alabama State (same as the percentage of diabetes), where the total Cases in Alabama State is 2,998,033.



Visualize using Tableau

US Counties: COVID19 + Weather + Socio/Health



https://public.tableau.com/app/profile/shafa.amira.qonitatin/viz/USCounties_16862401966560/Dashboard