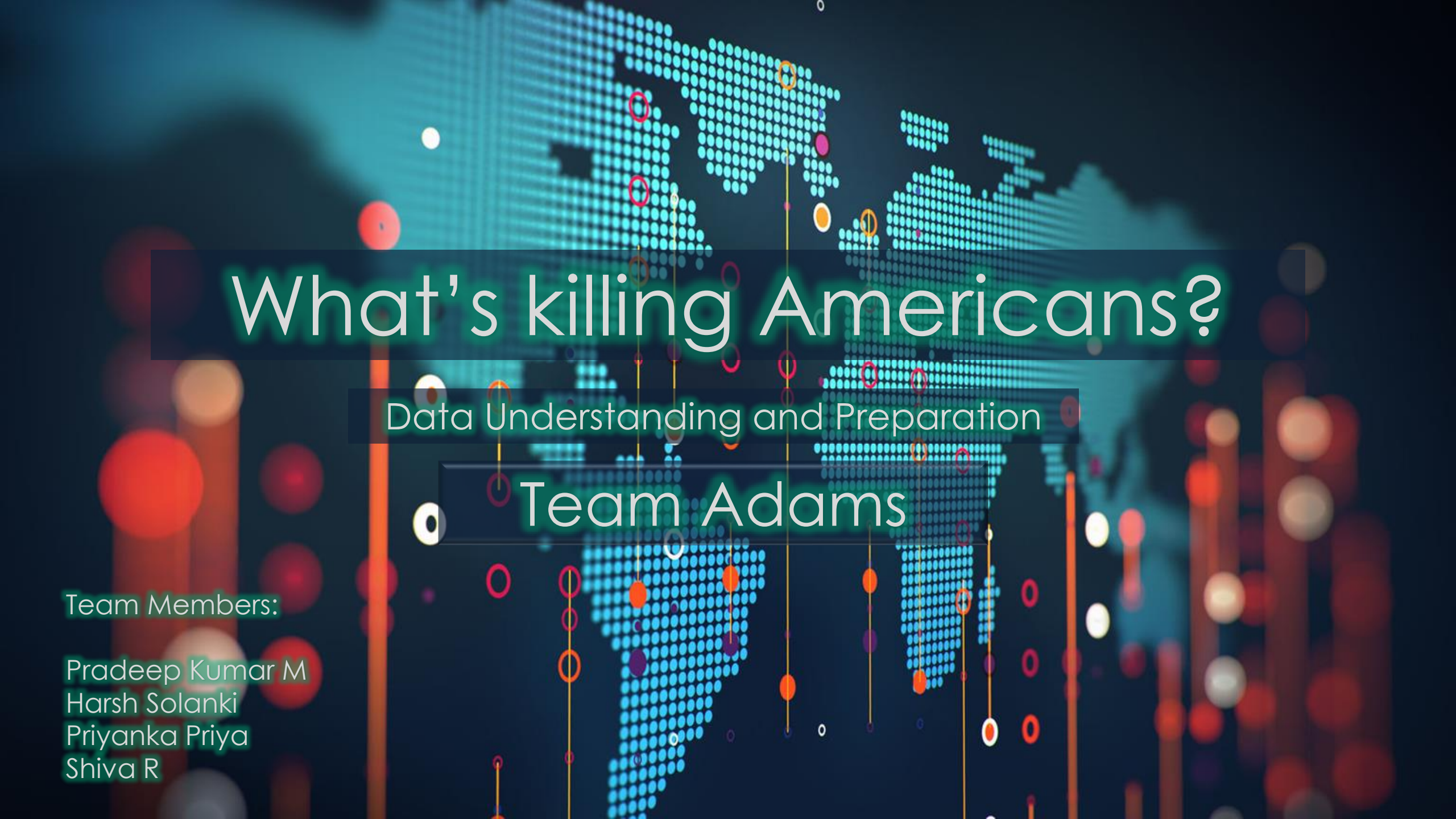


What's killing Americans?

Data Understanding and Preparation

Team Adams

Team Members:

Pradeep Kumar M
Harsh Solanki
Priyanka Priya
Shiva R

Content

- Business Need and Introduction to Dataset
- Data Understanding:
  - Entity Relationship Diagram
  - Normalization
  - Uploading Data
- Analysis with SQL Queries
- Visualization
- Recommendations
- Lessons Learned

Business Need and Introduction to Dataset

Business Need and Introduction to Dataset

This project is an attempt to analyze and recognize the correlation between various living factors and mortality due to major diseases in USA. These correlation will help governments in devising policies and initiatives at a County level

Comprehensive **Census Data** highlighting Demographics, Income, Nature of Employment, Mode of Transport, Population count granular to ethnicity at county level

Mortality Rate Data of Top 21 Categories corresponding to major diseases at County level

A collective Dataset of both Census and Mortality for years **2015 & 2017** to intersperse & capture the 'changing trends'

67,000+ records

A rich and complete collection of mortality rates corresponding to different categories for all counties across the USA

Data Understanding

Data Understanding

Census Data

Location

County, State

Mode of Transport

Drive, Carpool, Transit, Walk, Other Transport, Mean Commute, Work at Home

Demographic

Total Population, Men, Women, Hispanic, White, Black, Native, Asian, Citizen

Employment Sector

Professional, Service, Office, Construction, Production

Employment Type

Private, Public, Family

Income

Income, Income Error, Income Per Capita, Income Per Capita Error, Poverty, Child Poverty

Employment Status

Employed, Self Employed, Unemployed

Mortality Data

Diseases and other mortality categories

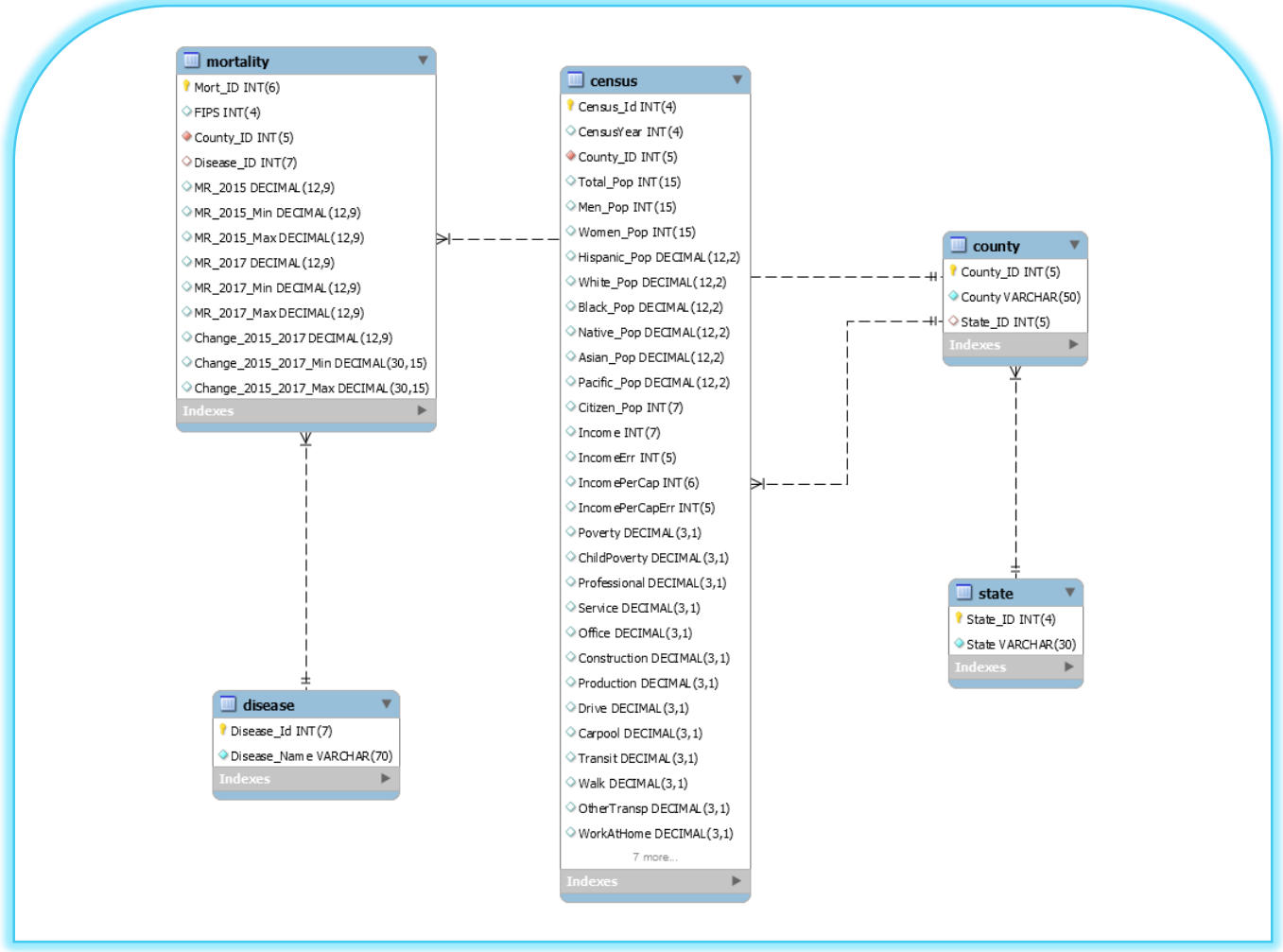
Cardiovascular, Chronic respiratory, Cirrhosis and other chronic liver diseases, Diabetes, urogenital, blood, and endocrine diseases, Diarrhea, lower respiratory, and other common infectious diseases, Digestive diseases, HIV/AIDS and tuberculosis

Maternal disorders, Mental and substance use disorders, Musculoskeletal disorders
Neglected tropical diseases and malaria
Neonatal disorders, Neoplasms, Neurological disorders
Nutritional deficiencies, Other communicable, maternal, neonatal, and nutritional diseases, Other non-communicable diseases

Self-harm and interpersonal violence, Transport injuries, Unintentional injuries, Forces of nature, war, and legal intervention

Data Understanding

1. Entity Relationship Diagram (EER)



Data Understanding

2. Normalization

1st Normal form	
There are no repeating groups	Yes
Data atomic	Yes
Each field has unique name	Yes
Primary Key	Yes

2nd Normal form	
First normal form	Yes
All non-key attributes are dependent	Yes
Eliminate partial dependencies	Yes

3rd Normal form	
First normal form	Yes
2nd normal form	Yes
Eliminate transitive dependencies	Yes

Partial Dependency

Census
CensusId
CensusYear
County_ID
~~County~~
~~State~~
Total_Pop

Morality
Mort_ID
FIPS
County_ID
Disease_ID
~~Disease~~
MR_2015

Transitive Dependency

County
County_ID
County
~~State~~
State_ID

State
State_ID
State

First Normal Form (1NF): Each row is Unique, and no columns contains multiple values. Table for Census was mostly downloaded in first normal form. For mortality table, we had to separate county and states. It was in following form (Autauga County, Alabama)

Second Normal Form (2NF): It is in 1NF and all non-key attributes are fully functional dependent on the primary key (does not contain partial functional dependencies)

Third Normal Form (3NF): It is in 2NF and non key columns do not determine other non key columns

Data Understanding

3. Uploading Data

Data Cleaning *

- Rows with '**NA**' values removed or Treated with mean value
- Removing **spaces**
- Making counties **unique**

Data Sources

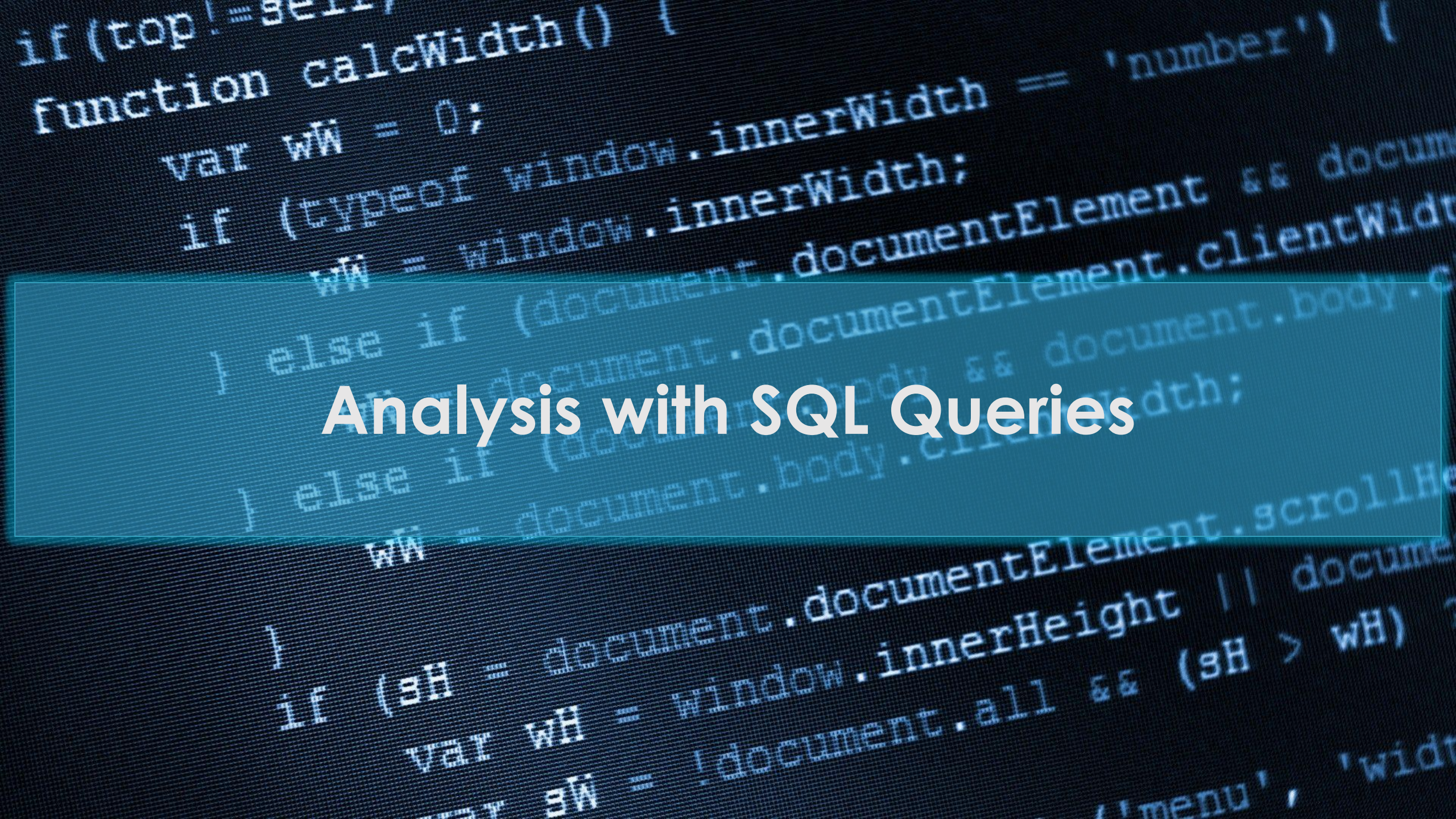
Census Data: US Census Bureau
Mortality Data: Institute for Health Metrics and Evaluation

ETL

- Records with blank spaces treated with mean value
- Imported Data in CSV files

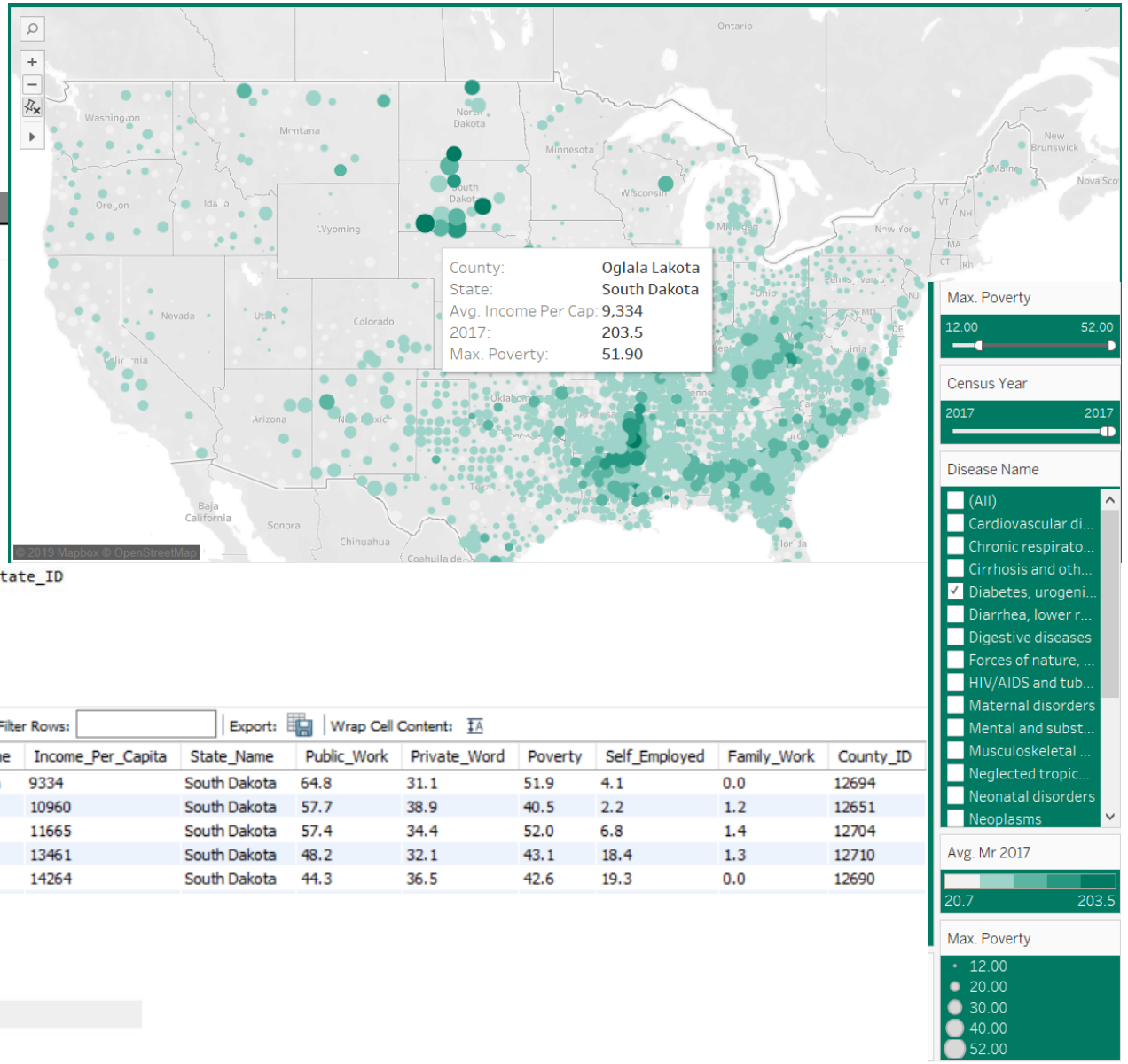
Troubleshooting/ lessons learned*

- Secured file Upload Path : *secure_file_priv*
- sql_mode=*('ONLY_FULL_GROUP_BY','')*


Analysis with SQL Queries

Public Jobs curse Oglala Lakota with Poverty and Diabetes!

In 2017, Oglala Lakota County records the highest deaths due to Diabetes(across USA) having the least per capita income(across South Dakota) with half the population under poverty and majority working for public sector



US_Final_Project_SQL_Prade...

Limit to 2000 rows

```
759 • Select census.CensusYear AS Year,
760       coun.County AS County_Name,
761       census.IncomePerCap AS Income_Per_Capita,
762       state.state AS State_Name,
763       census.PublicWork AS Public_Work,
764       census.PrivateWork AS Private_Word,
765       census.poverty AS Poverty,
766       census.SelfEmployed AS Self_Employed,
767       census.FamilyWork AS Family_Work,
768       census.County_ID
769   from  census censu
770   left join county coun ON censu.County_ID = coun.County_ID Left join state state ON coun.State_ID = state.State_ID
771   where coun.State_ID IN
772   (select* from ( Select c.State_ID
773       From  mortality m
774       left join disease d ON m.disease_id = d.disease_id
775       left join county c ON m.county_id = c.county_id
776       left join census cen ON m.County_ID = cen.County_ID
777       left join state s ON c.State_ID = s.State_ID
778       where cen.County_ID IN
779           (select* from (select County_ID
780                   from mortality
781               where disease_id = '1000004'
782                   group by County_ID
783                   order by MAX(MR_2017) desc Limit 1 )
784           temp_tab )and cen.CensusYear = '2017' limit 1 )temp_tab )
785   and censu.CensusYear = '2017'
786   Order by (censu.IncomePerCap) asc;
```

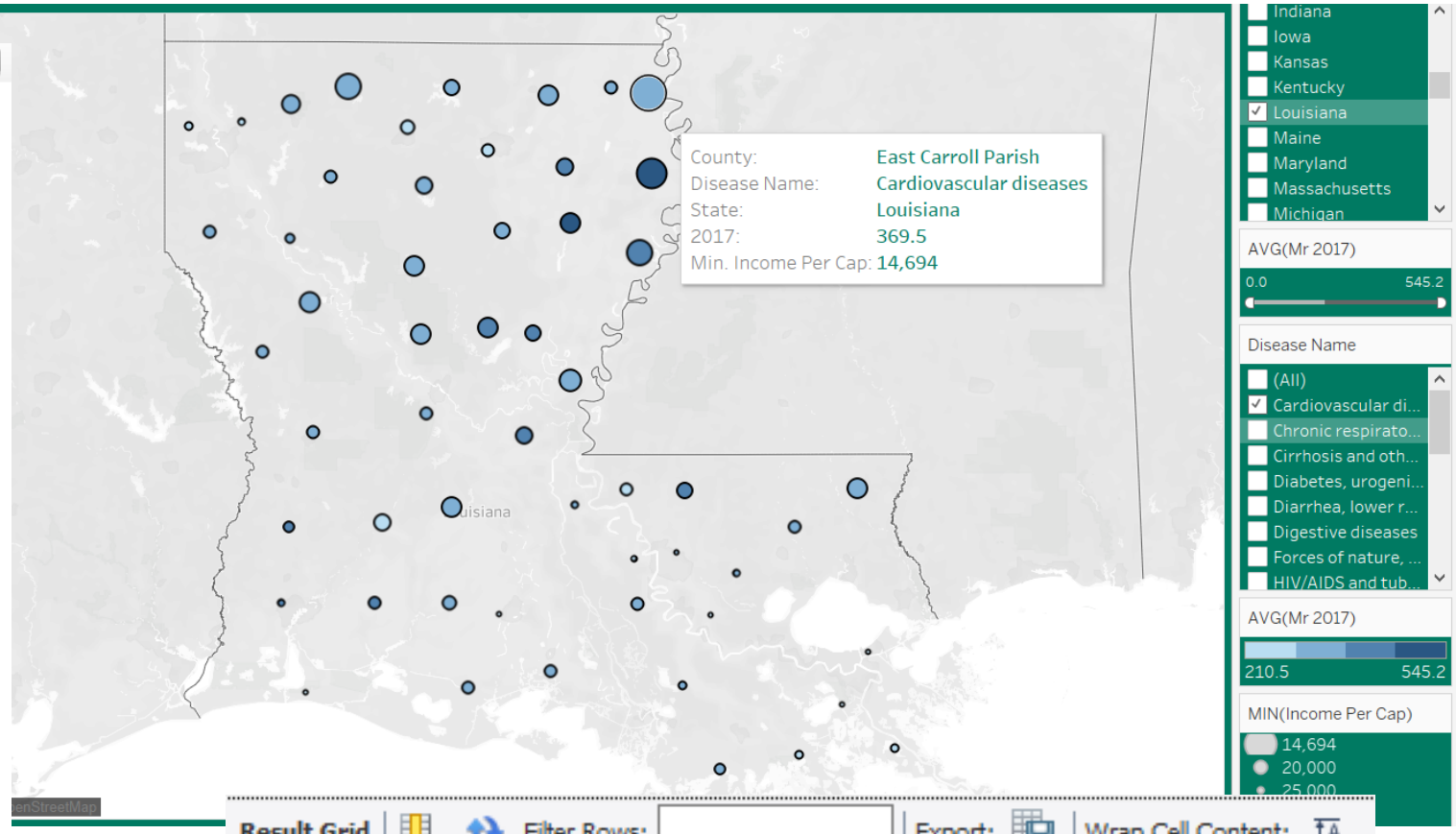
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Year	County_Name	Income_Per_Capita	State_Name	Public_Work	Private_Word	Poverty	Self_Employed	Family_Work	County_ID
2017	Oglala Lakota	9334	South Dakota	64.8	31.1	51.9	4.1	0.0	12694
2017	Buffalo	10960	South Dakota	57.7	38.9	40.5	2.2	1.2	12651
2017	Todd	11665	South Dakota	57.4	34.4	52.0	6.8	1.4	12704
2017	Ziebach	13461	South Dakota	48.2	32.1	43.1	18.4	1.3	12710
2017	Mellette	14264	South Dakota	44.3	36.5	42.6	19.3	0.0	12690

Franklin Parish suffers with Cardiovascular Diseases

In 2017, People in Franklin Parish, Louisiana saw the highest cardiovascular deaths. Interesting to note that nearby counties have similar mortality rate. East Carroll Parish, has the least per capita income in the same state

```sql
Select
    census.IncomePerCap AS PerCapitaIncome,
    census.County_ID AS County_ID,
    coun.County AS County,
    state.state AS State_ID,
    disease.Disease_Name AS Disease_Name
from  census census
left join county coun ON census.County_ID = coun.County_ID
Left join state state ON coun.State_ID = state.State_ID
Left join mortality mortality ON census.County_ID = mortality.County_ID
Left join disease disease ON mortality.Disease_ID = disease.Disease_Id
where coun.State_ID IN
(select* from
     ( Select c.State_ID
    From  mortality m
    left join disease d ON m.disease_id = d.disease_id
    left join county c ON m.county_id = c.county_id
    left join census cen ON m.County_ID = cen.County_ID
    left join state s ON c.State_ID = s.State_ID
    where cen.County_ID IN
        (select* from
            (select County_ID from
            mortality where disease_id = '1000001'
        group by County_ID order by MAX(MR_2017) desc Limit 1
            ) temp_tab )
     and cen.CensusYear = '2017' limit 1 )temp_tab )
and disease.Disease_Name IS NOT NULL AND disease.Disease_Id = '1000001'
Order by census.IncomePerCap asc
;
```



Result Grid | Filter Rows: | Export: | Wrap Cell Content:

PerCapitaIncome	County_ID	County	State_ID	Disease_Name
14694	11168	East Carroll Parish	Louisiana	Cardiovascular diseases
15134	11199	Madison Parish	Louisiana	Cardiovascular diseases

Fallon battles with Nutritional Deficiency

*In the year 2017, **Fallon** from Montana has the greatest number of deaths due to 'Nutritional deficiencies' with **170** people more susceptible to the disease because of poverty*

SQL Query-3

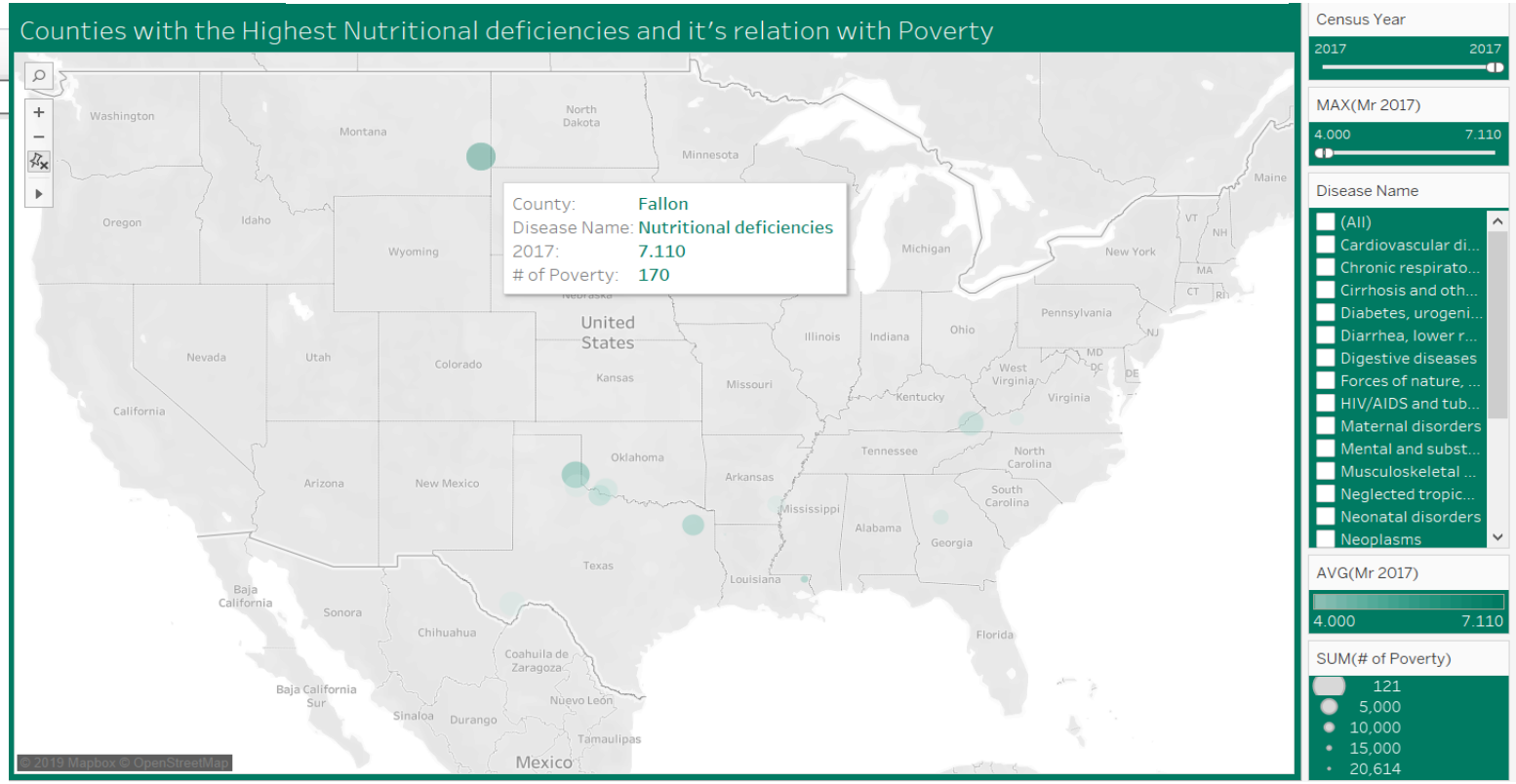
Result Grid | Filter Rows: | Export: | Wrap Cell Content:

	State	County	Population_in_Poverty
▶	Montana	Fallon	170

Limit to 1000 rows

Find

```
174
175 •  Select  s.state AS State,
176            c.County AS County,
177            round((cen.total_pop*cen.poverty)/100) AS Population_in_Poverty
178        From  mortality m
179    left join disease d ON m.disease_id = d.disease_id
180    left join county c ON m.county_id = c.county_id
181    left join census cen ON m.County_ID = cen.County_ID
182    left join state s ON c.State_ID = s.State_ID
183    where cen.County_ID IN
184    (select* from
185        (select County_ID
186            from mortality
187        where disease_id = '1000016'
188            group by County_ID
189        order by AVG(MR_2017)
190            desc Limit 1 ) temp_tab )
191    and cen.CensusYear = '2017'
192        group by c.County_ID;
```



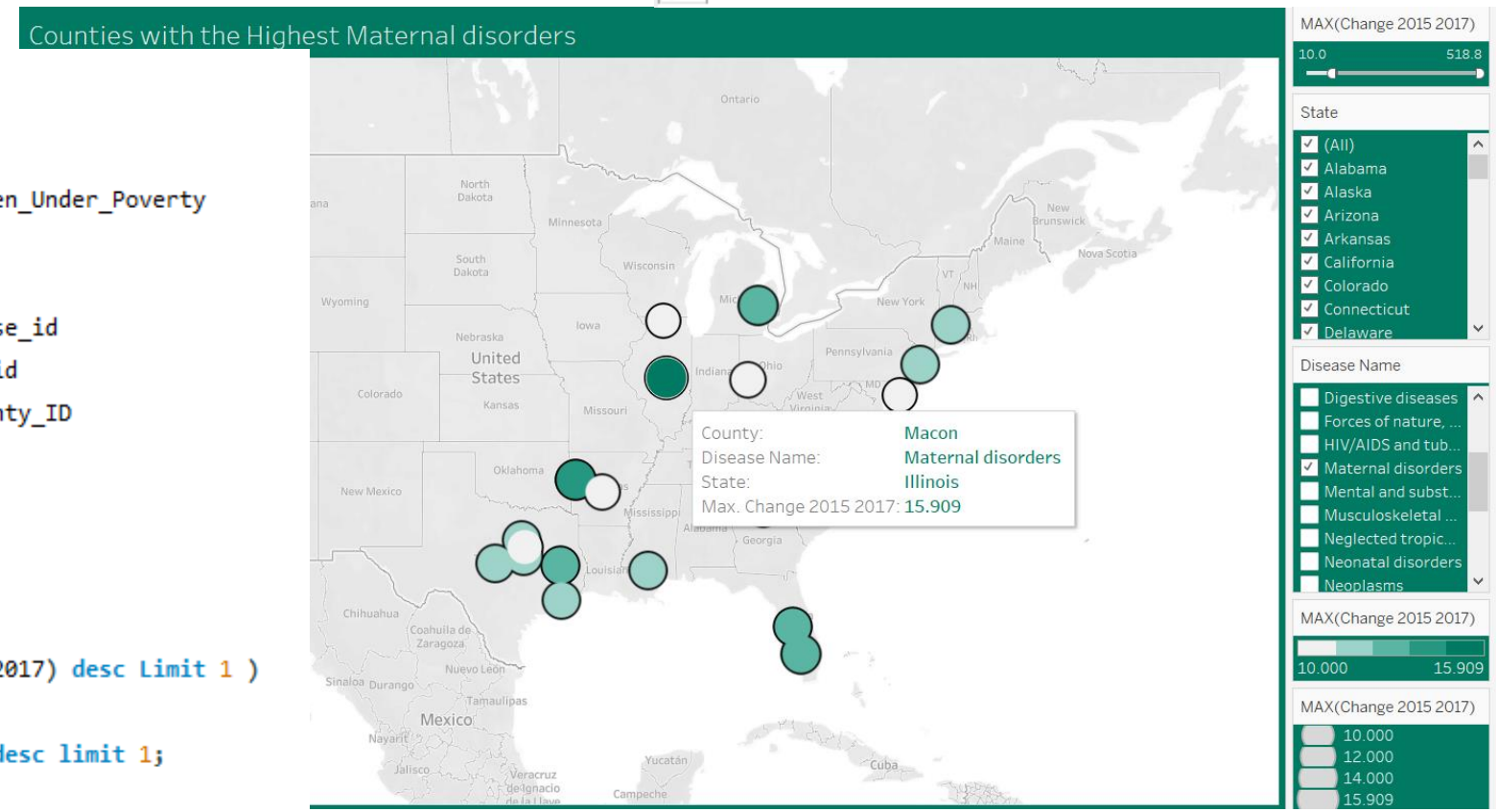
Macon's women in growing risk with Maternal Disorders

Women in Macon, Illinois have seen highest increase in Maternal disorders with around 10.5k women are under poverty

```sql
Select  c.County_id AS County_ID,
        s.state AS State,
        c.County AS County,
        round(cen.Women_Pop*(cen.poverty/100)) AS Women_Under_Poverty

From  mortality m
        left join disease d ON m.disease_id = d.disease_id
        left join county c ON m.county_id = c.county_id
        left join census cen ON m.County_ID = cen.County_ID
        left join state s ON c.State_ID = s.State_ID
where cen.County_ID IN
        (select* from
            (select County_ID from mortality
                where disease_id = '1000009'
            group by County_ID order by (Change_2015_2017) desc Limit 1 )
        temp_tab )
group by cen.County_ID order by (m.Change_2015_2017) desc limit 1;
```

	County_ID	State	County	Women_Under_Poverty
▶	10664	Illinois	Macon	10462

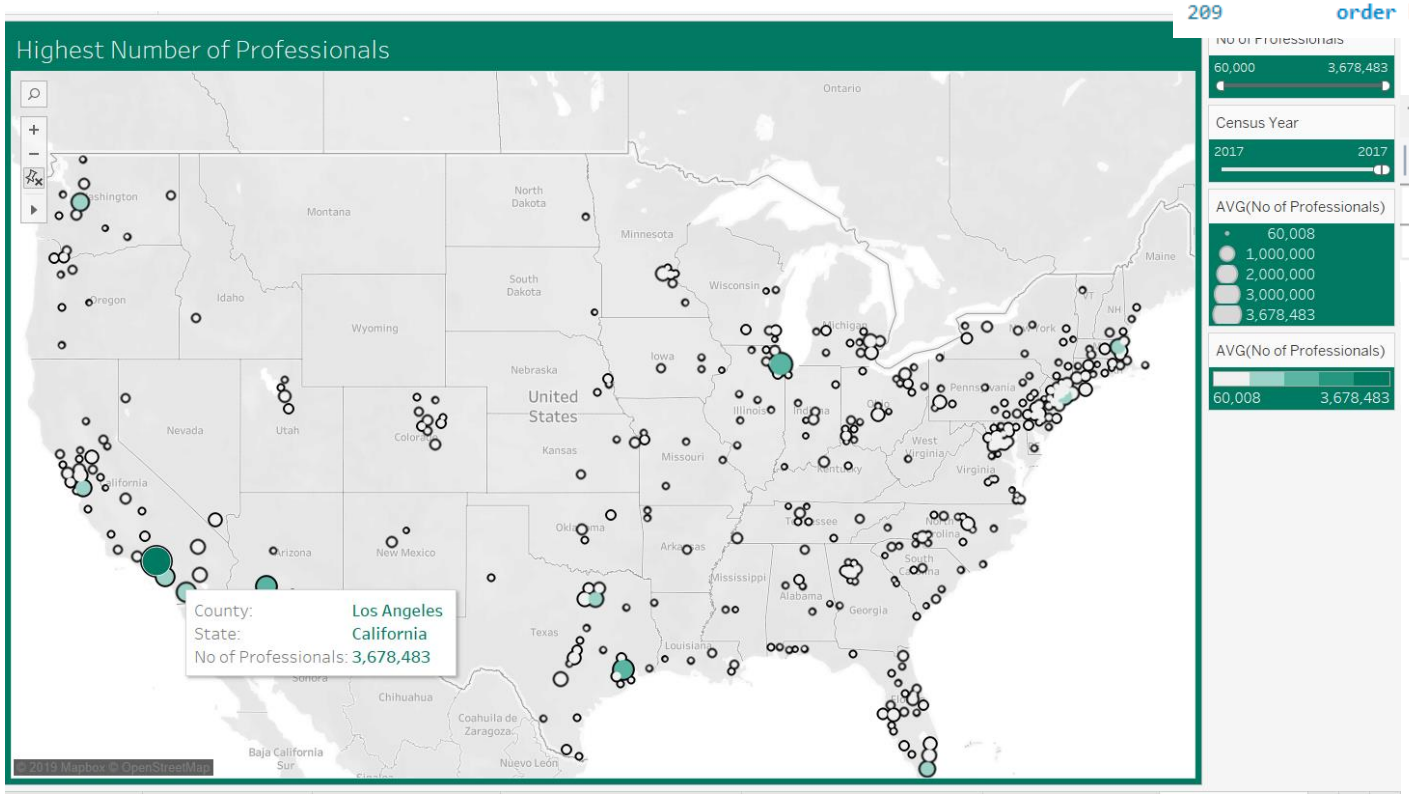


Los Angeles embracing Immigrants

LA from CA is the most welcoming place in USA for opportunity seekers across the globe

SQL Query-5

```
200 •  Select
201        state.state AS State,
202        coun.county AS County,
203        round(((censu.Total_Pop-censu.Citizen_Pop)*(100-censu.Unemployment))/100 ) AS EmployedImmigrants
204    FROM census censu
205        left join county coun ON censu.County_ID = coun.County_ID
206        Left join state state ON coun.State_ID = state.State_ID
207    where censu.CensusYear = '2017'
208        group by censu.County_ID
209        order by ((censu.Total_Pop-censu.Citizen_Pop)*(100-censu.Unemployment))/100 desc limit 1;
```



Result Grid | Filter Rows:

State	County	EmployedImmigrants
California	Los Angeles	3584222

Macon's Asian women suffer the most in maternity

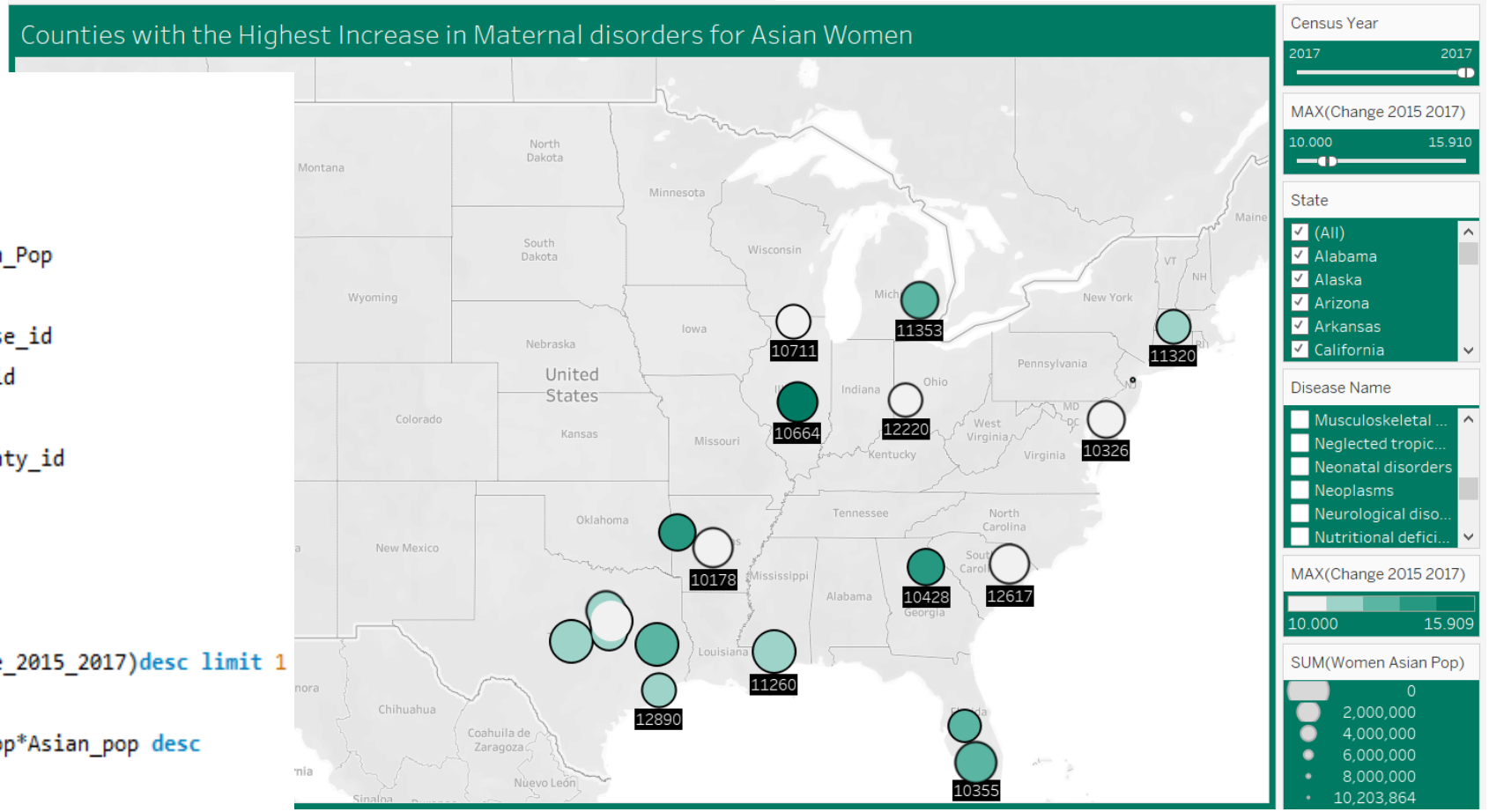
SQL Query-6

Asian Women in Macon, Illinois are recorded with highest Maternal disorders

Result Grid | Filter Rows:

	County	State_ID	State	Asian_Women_Pop
▶	Macon	114	Illinois	68228.40

```
242
243 •  Select  county AS County,
244            s.state_id AS State_ID,
245            s.State AS State,
246            cen.Women_Pop*Asian_Pop AS Asian_Women_Pop
247    From  mortality m
248    left join disease d ON m.disease_id = d.disease_id
249    left join county c ON m.county_id = c.county_id
250    left join state s ON c.state_id = s.state_id
251    left join census cen ON m.county_id = cen.county_id
252    where cen.county_id IN
253    (select* from (select county_id
254            from mortality
255            where disease_id = '1000009'
256            group by county_id order by MAX(Change_2015_2017)desc limit 1
257            ) temp_tab )
258    group by (cen.county_ID) order by cen.Women_Pop*Asian_pop desc
259    limit  3;
```



Team Adams

Visualization

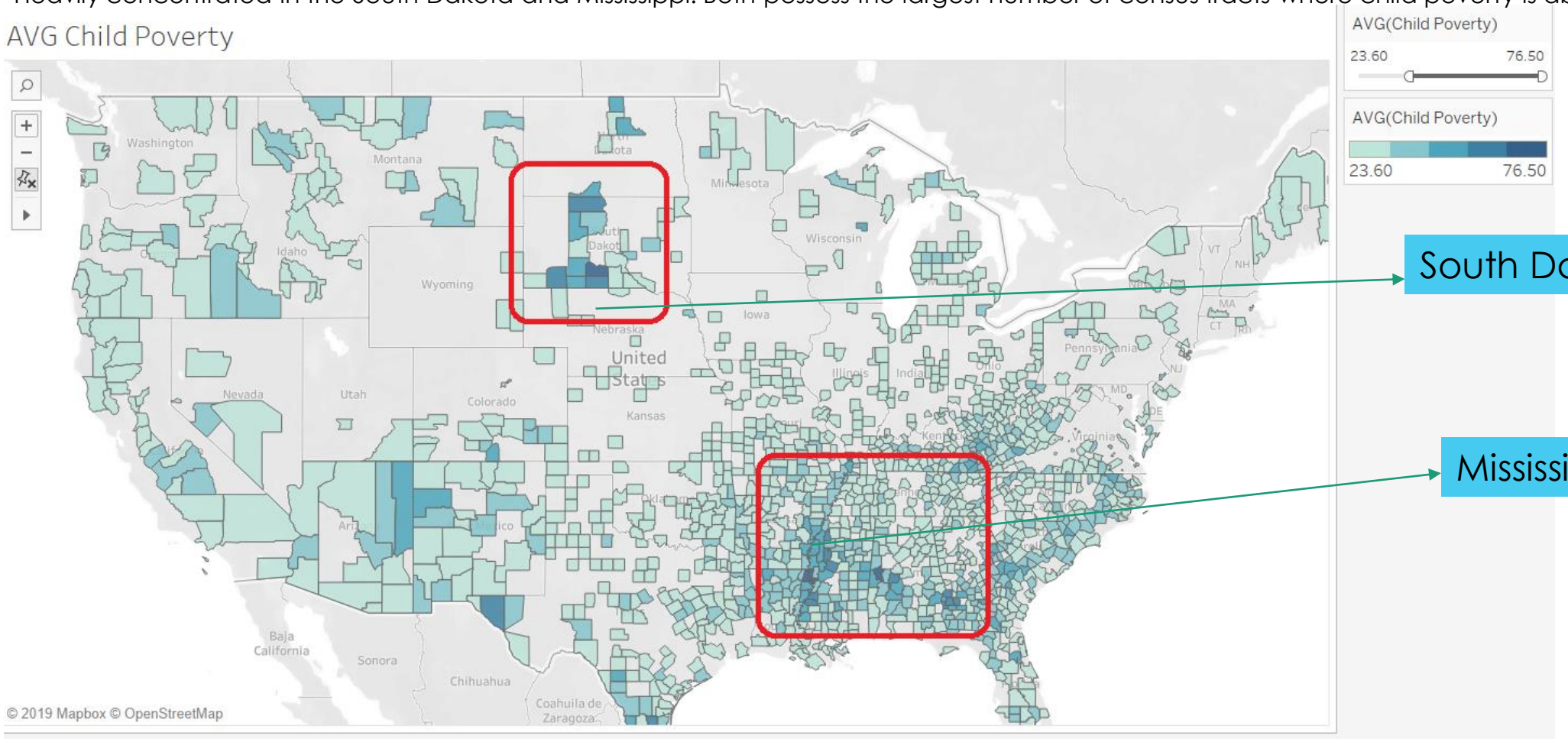
Concentration of Child Poverty VS USA Average:

Viz-1

State wise:
Greatest concern : Mississippi
Least concern: Connecticut.

County wise:
Greatest concern: East Carroll Parish
Least Concern: Mineral County

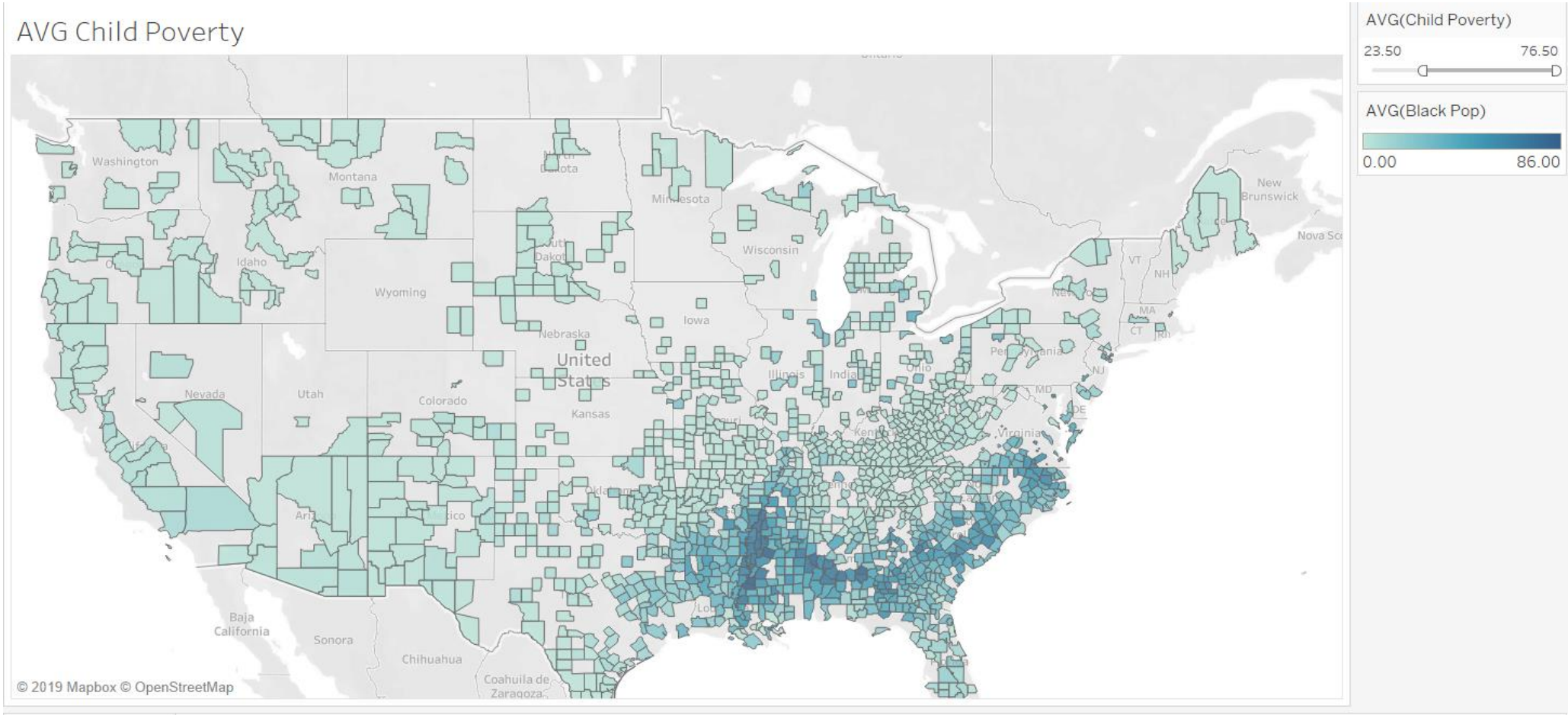
Heavily concentrated in the South Dakota and Mississippi. Both possess the largest number of census tracts where child poverty is above average.



Child Poverty VS Black Population Concentration:

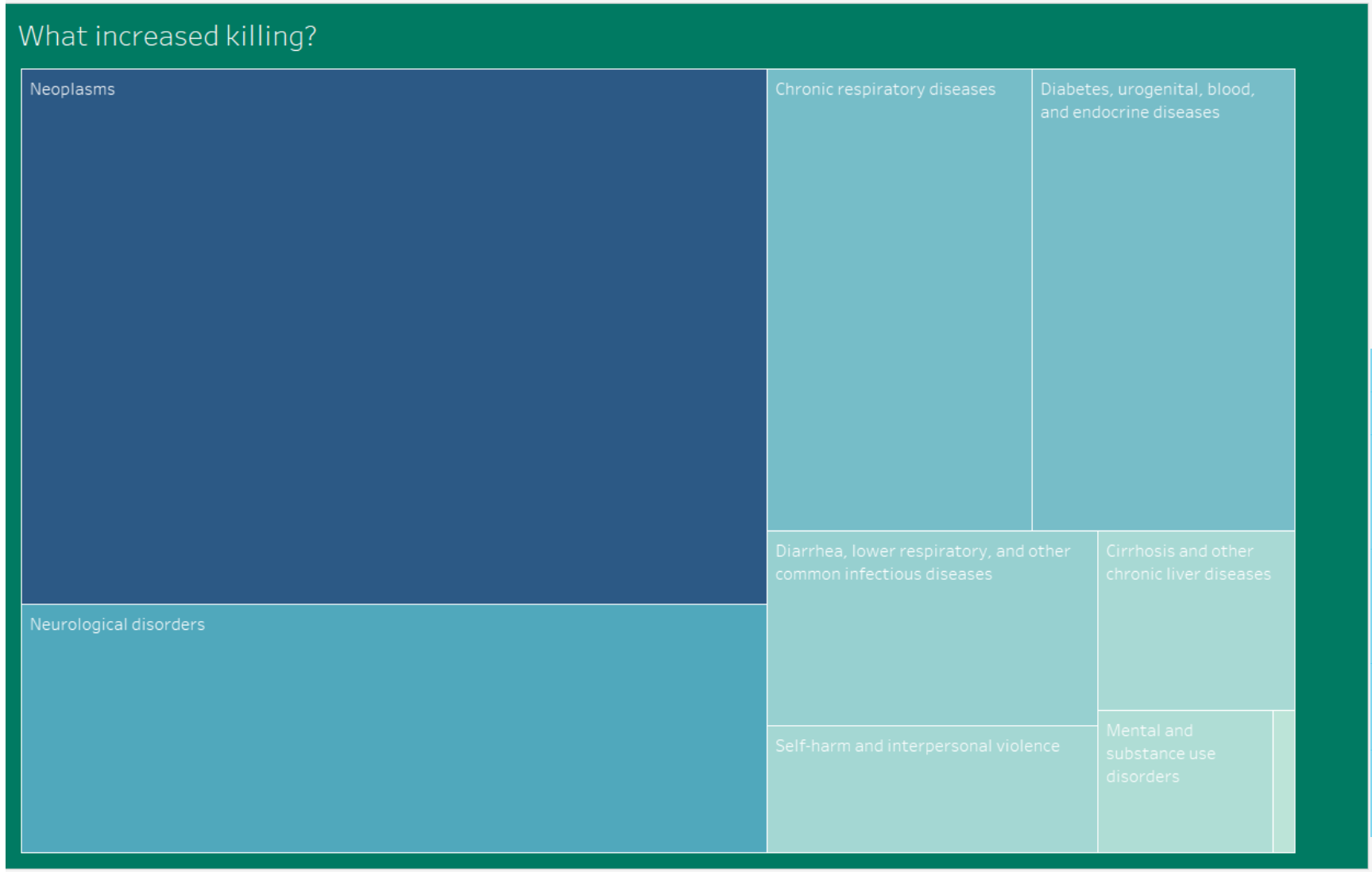
Viz-2

Child Poverty is weakly correlated with the proportion of black residents, showing a positive relationship.



What increased the deaths?

Viz-3



AVG(Change 2015 2017)

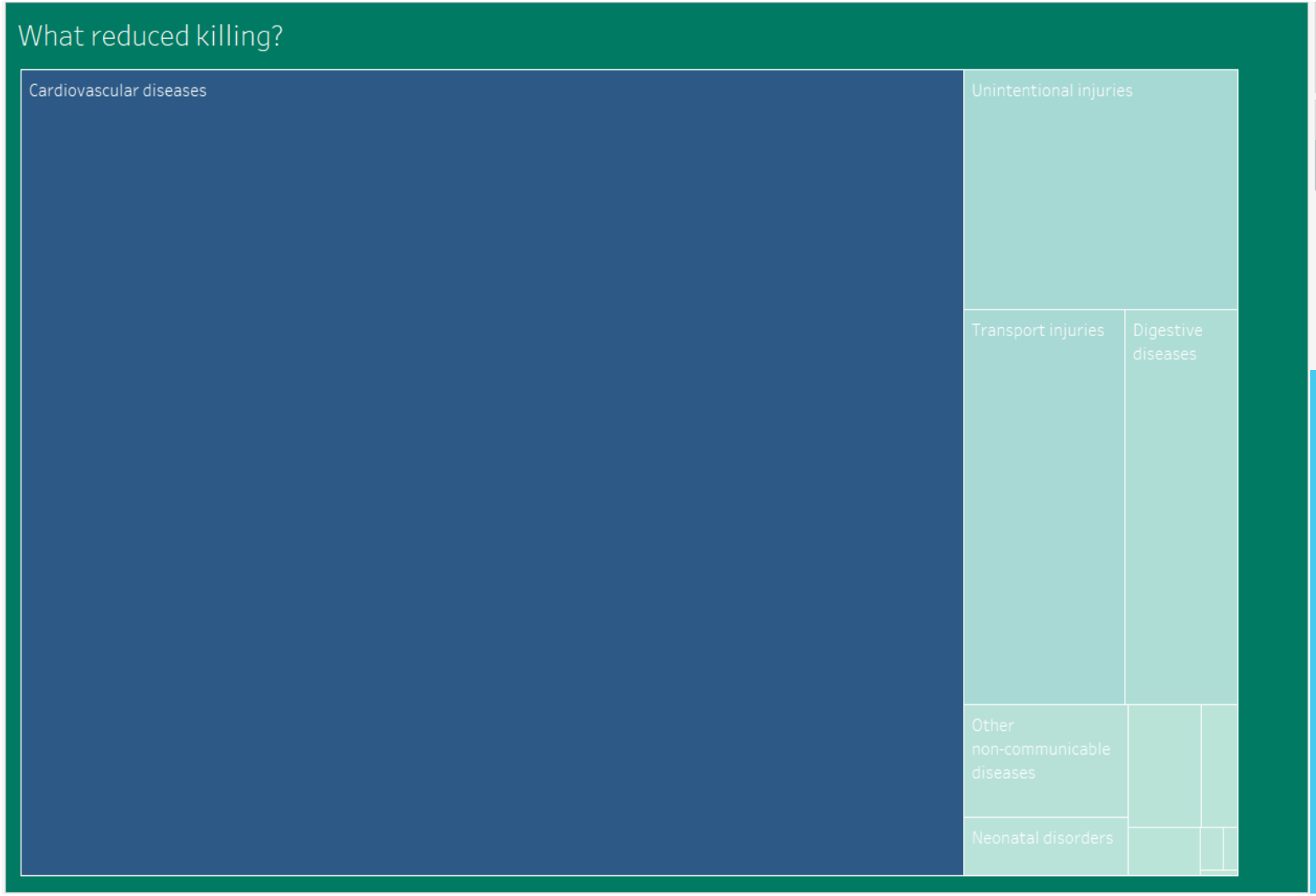
0.000 7.097

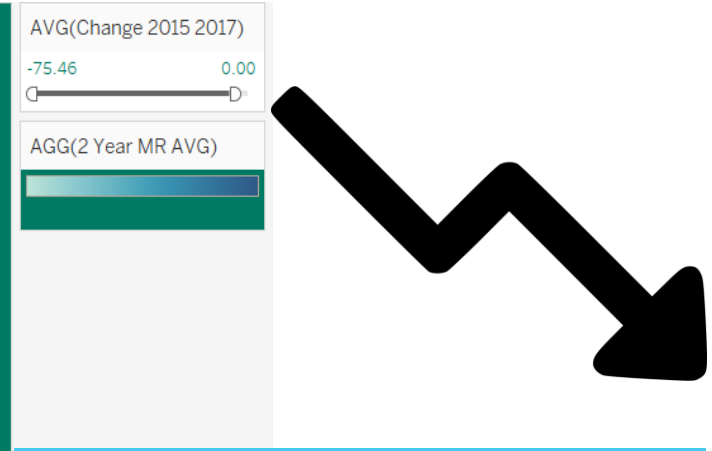
AGG(2 Year MR AVG)

- Neoplasms
- Neurological disorders
- Chronic respiratory diseases
- Diabetes, urogenital, blood, and endocrine diseases
- Diarrhea, lower respiratory, and other common infectious diseases
- Self-harm and interpersonal violence
- Cirrhosis and other chronic liver diseases
- Mental and substance use disorders
- Nutritional deficiencies
- Neglected tropical diseases and malaria

What reduced the deaths?

Viz-4





Cardiovascular diseases

Unintentional injuries

Transport injuries

Digestive diseases

Other non-communicable diseases

Neonatal disorders

Musculoskeletal disorders

HIV/AIDS and tuberculosis

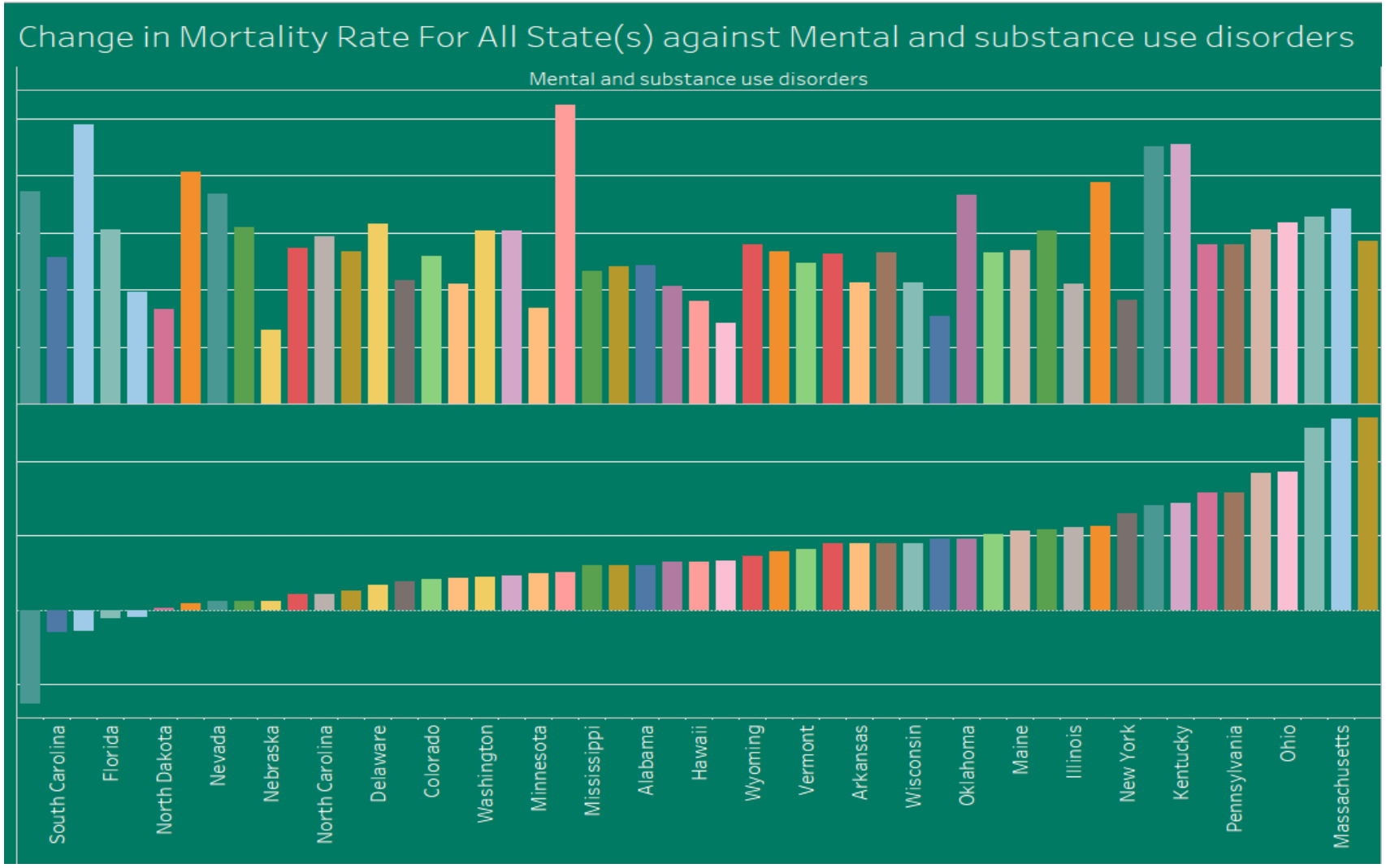
Other communicable, maternal, neonatal, and nutritional diseases

Maternal disorders

Forces of nature, war, and legal intervention

It is interesting to observe that the variation of "Mental and substance use disorders" has almost no negative values. That means that almost no State except DOC, South Carolina, Florida, Alaska, South Dakota has decreased their consumption of substances:

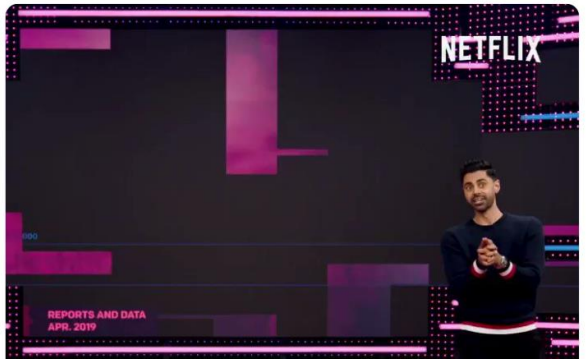
Viz-5



This is a pretty serious problem in USA. Fentanyl-Involved Deaths are Soaring In U.S.
Do watch how serious this is here:
https://www.youtube.com/watch?v=vkhMfgOT9Xs

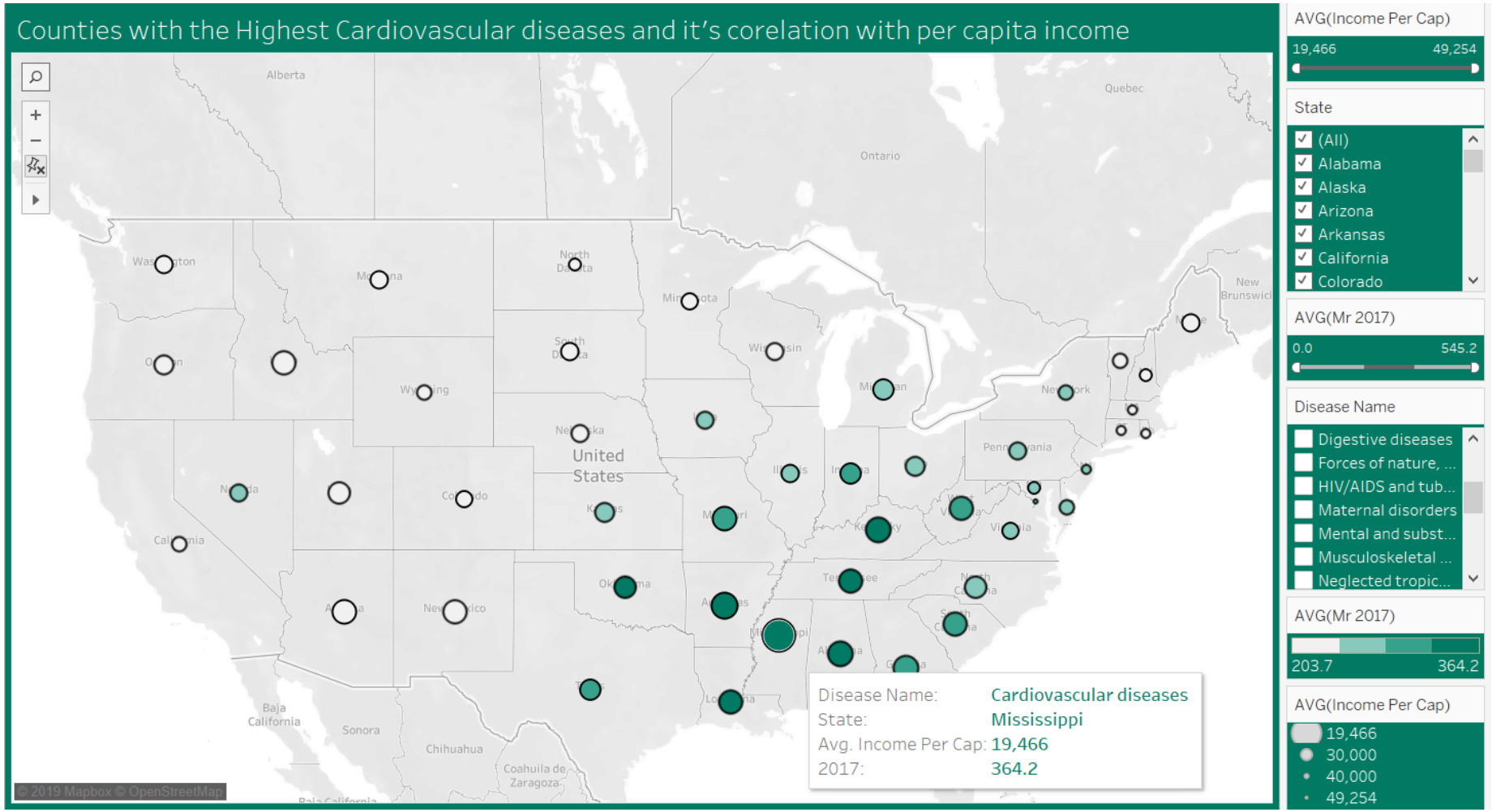
Patriot Act with Hasan Minhaj
@patriotact
Follow

In the last five years, fentanyl became the reason for nearly two-thirds of all opioid deaths, making it the third wave of the opioid crisis.



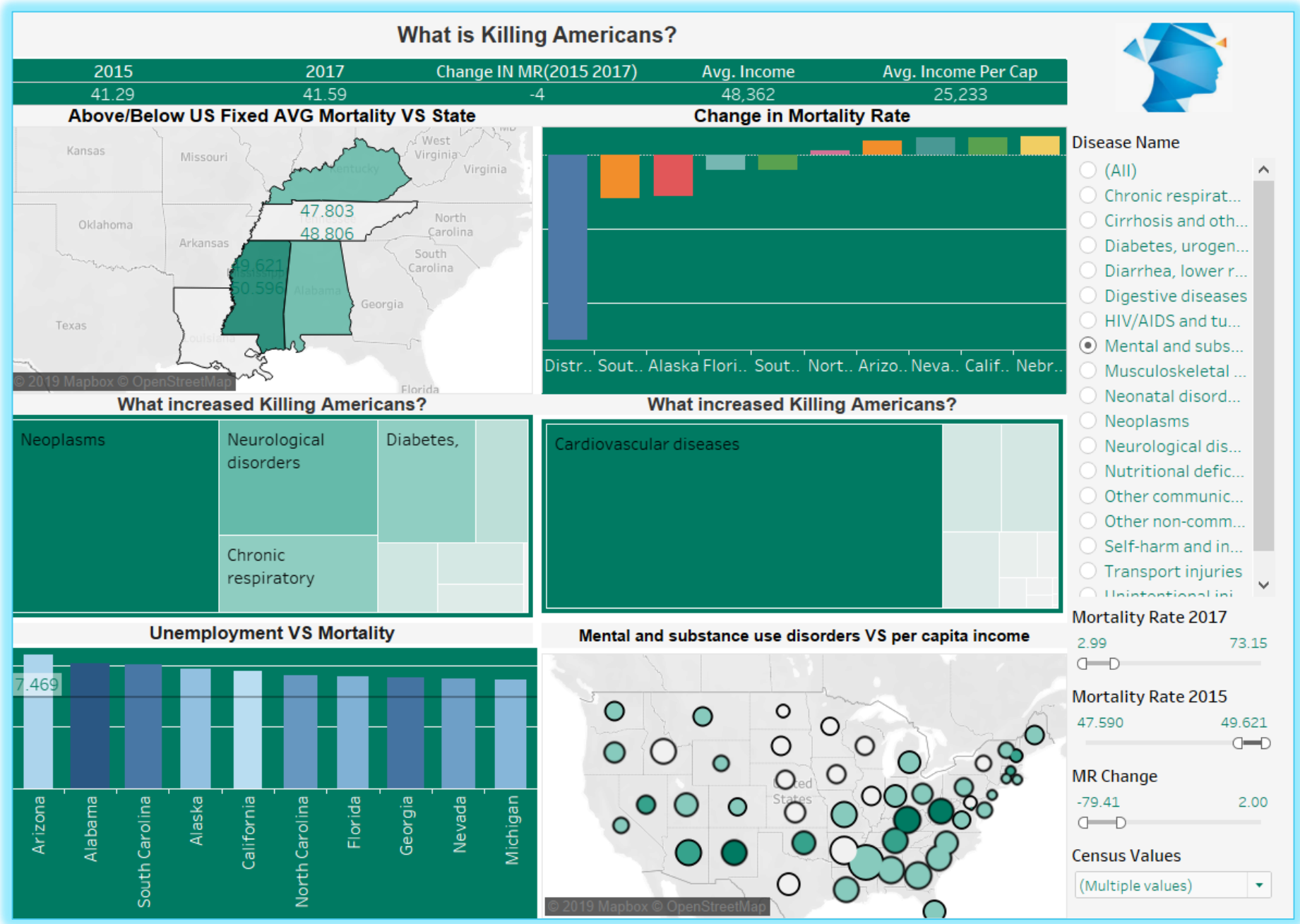
It is interesting to observe that the Mississippi has highest deaths due to Cardiovascular disease and lowest per capita income

Viz-6



Less income
equals
more junk food

Dashboard



Viz-7

Recommendations

Interesting insights Gained from the SQL Analysis and Reports:

Growing unemployment leads to poverty:

Corson, **Unemployment**: 29.4, **Poverty**: 45.6
Kusilvak, Census Area: **Unemployment**: 28.8, **Poverty**: 39.1
Oglala Lakota, **Unemployment**: 28.7, **Poverty**: 53.3

Poverty decreases with higher income
Counties with 6 figure income has poverty within national average

Poverty encourages people to carpool:
Oglala Lakota with highest poverty has the maximum number of people using Carpool

Communities that take Transit have longer commutes:
Bronx, New York and Queens are topping the chart

Longest and Shortest County names:
Prince of Wales-Hyder Census Area: 33
Lee: 3

Natives and Poverty, Unemployment:
Oglala Lakota has highest Native Population and it is 3rd in unemployment and 1st in Poverty.

Following Counties have 100% White Population:
Jones, Webster, Jackson, Magoffin, Wheeler, Highland, Garfield, Loup

Can two states have the same county name?
The answer is a resounding, patriotic yes and coincidently there are 51 counties with names similar to the State

A well thought idea of collecting Census and Mortality data is primarily to understand and improve livelihood. The results/observations shown earlier and on this page are vital inputs to the county/state/federal administrators in the policy making/ driving initiatives around healthcare and medical sectors

Commute:
More people walk in New York City than people who drive, but they do take transit. Clay, Georgia has the highest rate of carpooling.
Alaska likes to walk and, it has the lowest mean commute time as well.

Career and Income:
The highest fraction of people in a 'Professional' career live in Fall Church City, Virginia. This is the same county with third highest income.
Counties with the lowest portion of Construction jobs include Washington D.C., Florida, Rhode Island, New York and California. This is a good sanity check as they are already developed, and any level of development jobs are diluted by their large population.

Population:
The most populated County is Los Angeles, California with a population of 10105722 people The least populated County is: Loving, Texas with a population of: 74 people.
The most populated State is California with a population of 77404311 people The least populated State is: North Dakota with a population of: 1467115 people

Lessons Learned

Lessons Learned

1. Complete Data wrangling process to bring necessary data required (from multiple sources) for the analysis
2. Data cleaning to maintain accurate records for efficient querying and visualizing processes
3. Data Normalization and appropriate categorization of data into respective tables
4. Complex queries and nuisance of SQL with the knowledge of multiple paths of getting results
5. Sequential process in querying, with the quest of finding more business implications from the data
6. Learning from encountered errors like knowing to insert records from correct folder (secure_file_priv), aggregation function (sql_mode=('ONLY_FULL_GROUP_BY',‘’))
7. Garbage 'IN' equals Garbage 'Out': We initially started off with very little cleaning of the data and we realized that we are not getting correct result/visualization.
   Example: Similar County Names leading to incorrect results. We were not aware that USA has various counties with similar names and hence, after realization concatenated the county name with the state name to get unique record
8. Hands on with Tableau taught usage of various kinds of charts that can be associated with a particular form a data

End of Presentation