Exploratory Data Analysis

Topic of Study: Safety in states compared to Murders in USA-2012

Background:

Exploratory data analysis involves discovering patterns in the dataset and visualization helps in doing the same. There are many factors which affect the way you interpret the data and this is the reason why I wanted to analyse a dataset which can provide some insight. Moving to the US, I was just inclined towards understanding which states are safe and which regions in a particular state are safe. It is a very broad question to answer and there are various dimensions to answer this so I selected one dimension (Murder) and wanted to analyse the states which are safe with respect to murders.

Data Selection:

The data has been selected from the FBI database. The FBI maintains an extensive dataset for all the crimes that take place in the United States of America. So I decided to take the data for murders that took place in 2012. There are different databases which formed my source. The first FBI database gave me the count of murders that took place in USA with respect to the states. Another source provided me the split for murders that took place in the metropolitan and non-metropolitan areas of the states. I joined the 2 datasets and also collected data such as the population of the state and the area of the state. I joined all these 4 dimensions and tried to explore if I could find any trends within the same. Since the data was in a raw format, I had to clean the same and transform the same into excel usable files before I begin the visualization.

Data Analysis and Terms used:

With the dataset created, I had to explore if I could find certain trends and also if the inclusion of different parameters affected the final result. I have introduced certain calculated fields like the Murder Ratio which is the number of murders for every 100000 people and Murders/SqKM which states the number of murders per square kilometre in a particular state. The data analysis and the visualization has been created in Tableau. I have different visualizations created for the different analysis explored below and also created a Final story to which contains all the visualization in a Tableau Story.

For all the visualization, I have mentioned the interpretation of the graphs and the inference which can be drawn from the visualization.

Final Question posed:

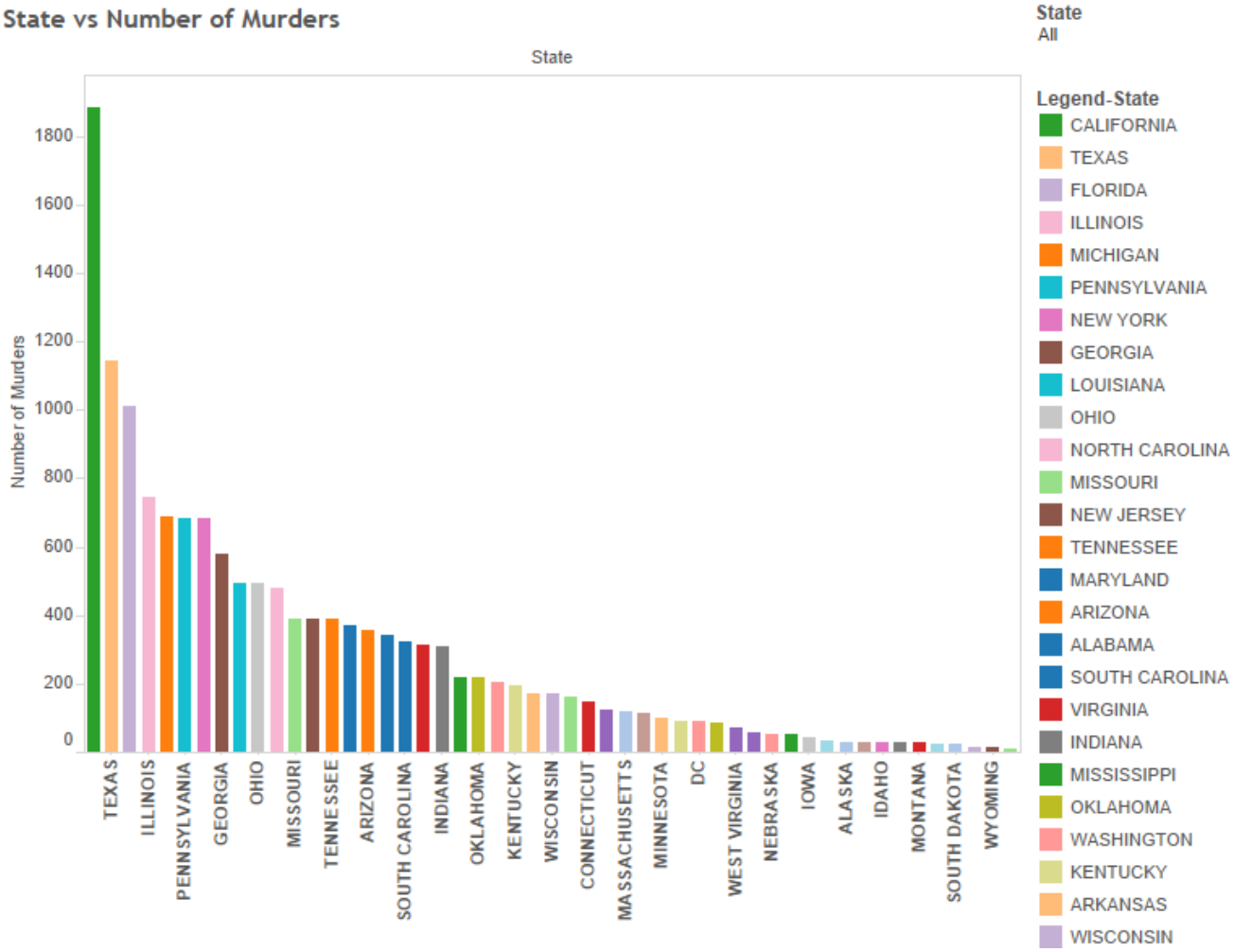
Explore which states and regions within the states are safe in terms of murders in the USA?

Initial Exploration:

I extracted the data containing the murder count for all the 50 states for the year 2012. Below is my first graph-a histogram-which has been plotted for all the states on the x-axis against the count of murders on the y-axis.

Interpretation:
Every bar represents a state and every state has a distinct colour.

Inference:
California tops the list with the most number of murders followed by Texas and Florida with Wyoming and Vermont as the states with the lowest murders.



Second Exploration:

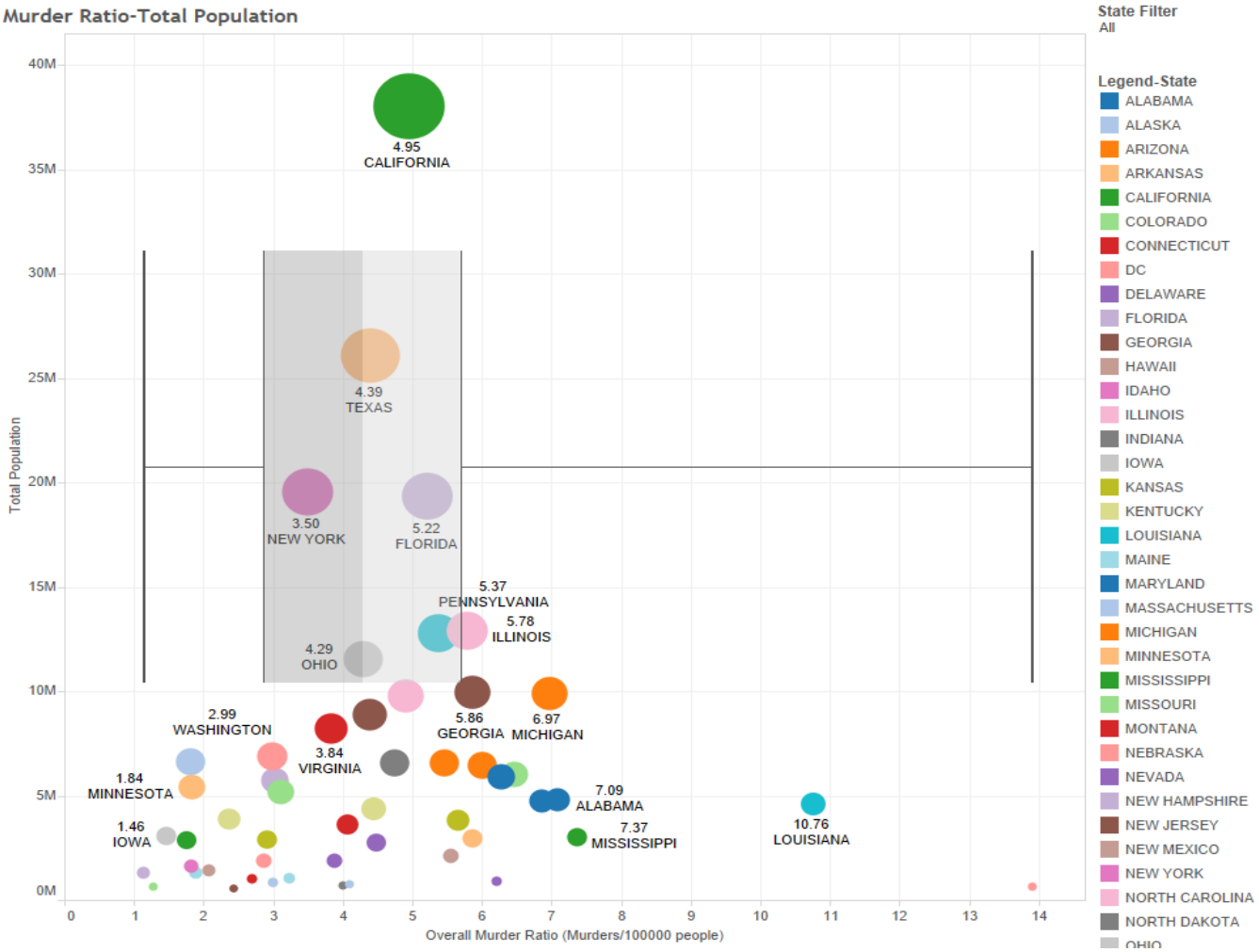
But the graph above simply gave us the basic idea of how many murders took place in a particular state without considering other factors like area size or population so I further decided to drill down in the data. So I decided to understand what changes when I include other factors such as population into the picture. I created a murder ratio index to calculate how many deaths occur per 100000 people in a particular state. Just having murder number doesn't define how safe a state is but this gives a better picture. Below is the graph plotted for the murder ratio index against the population (for the entire state). I also included a box and whisker plot to see the median and the range of the death ratio. California and Texas which actually have the highest deaths actually fall in the median range with DC having the highest death ratio.

Interpretation:

Size represents the state population and colour represents the different states.

Inference:

DC has more deaths per 100000 people as compared to any other state and quite a few states fall under within the 25% and 75% quartiles.



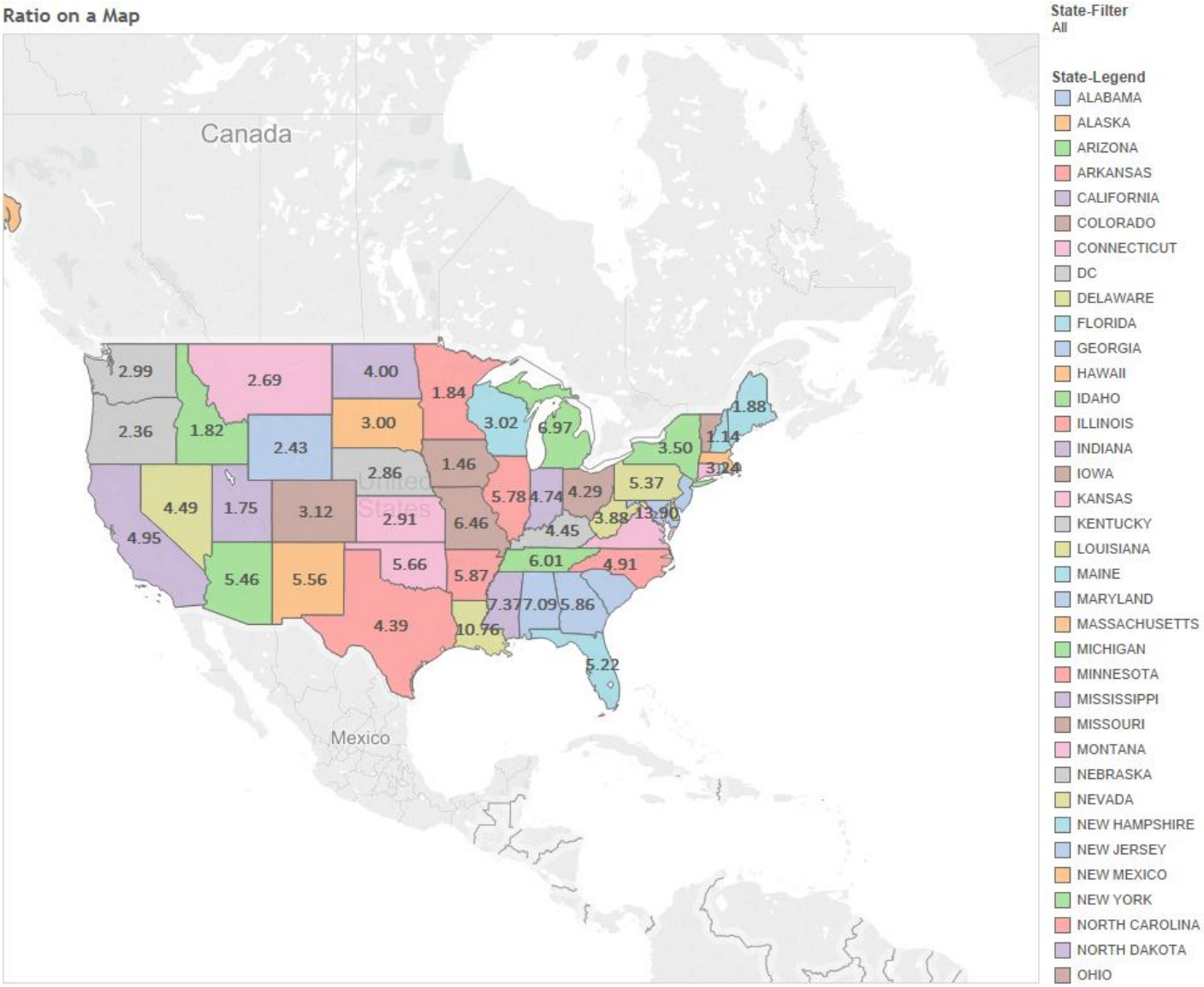
http://kunalseth.github.io/

Third Exploration:

I also wanted to analyse if there are any other trend with respect to position on the map so I plotted all the states on the map along with the murder index per 100000 people.

Interpretation:
Location on the map represents actual location of the state and colour is the distinguishing factor between states.

Inference:
A very interesting trend followed this visualization. States along the south border had a much higher murder ratio as compared to northern states. Almost all states along the south border had a ratio above 5 as compared to 2 in the northern states.



Further Exploration:

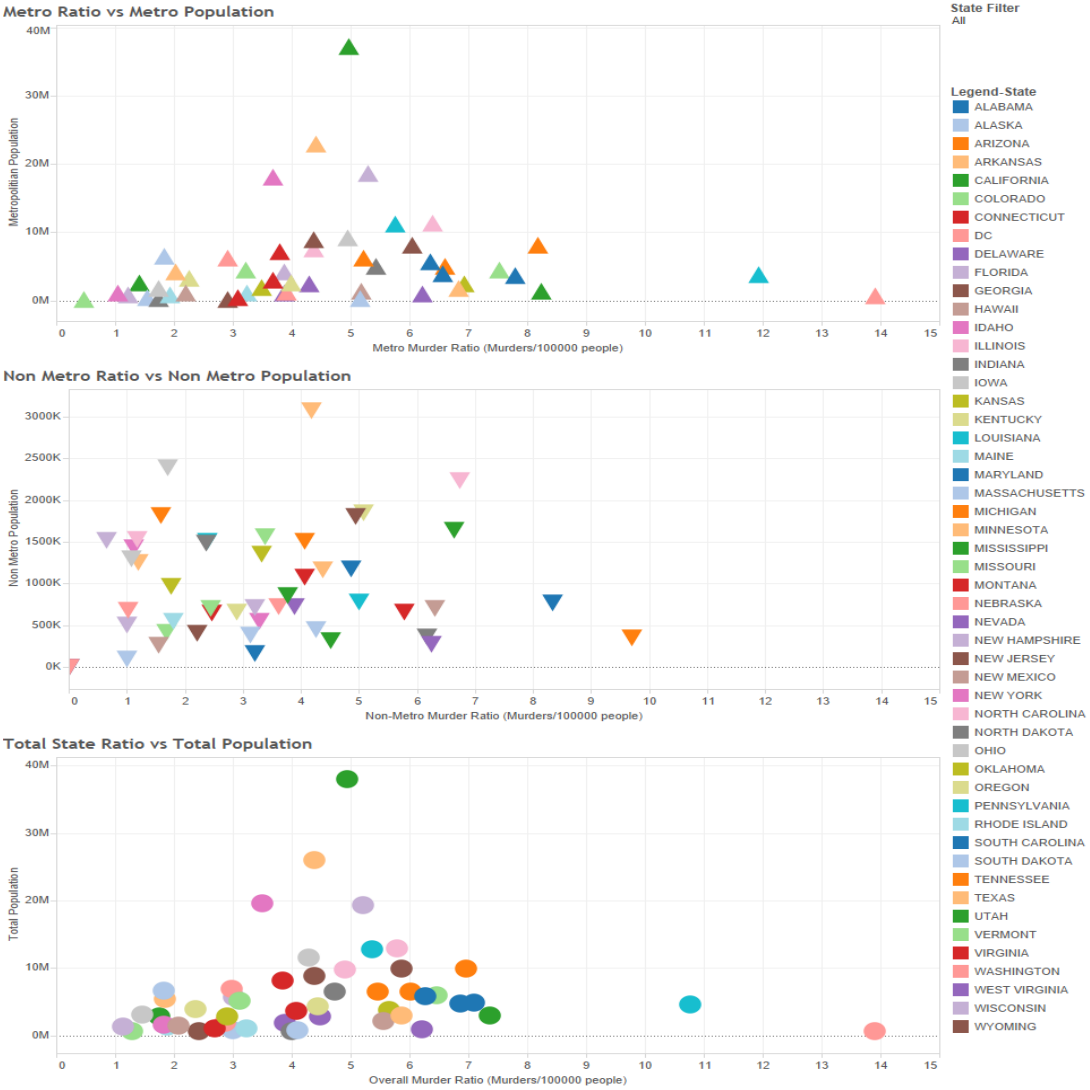
I further decided to see if there is some trend with where people stay in a particular state, so I split the population into metropolitan population and non-metropolitan population based on the counties they live in. Below I have 2 additional graphs created- Graph 1 has Metro murder ratio (per 100000) plotted against the metropolitan population and Graph 2 has Non-metro murder ratio plotted against non-metro population and these 2 graphs have been compared to Overall state murder ratio vs total state population.

Interpretation:

All the 3 graphs are connected to each other and selecting a state in any one of the graph will filter that state in the other 2 graphs too. The ratio axis are synced to each other so position is a direct comparison.

Inference:

There are many states which are actually much safer in the non-metro areas and others are safe in the metro areas. This relation can help in making a safe decision!!



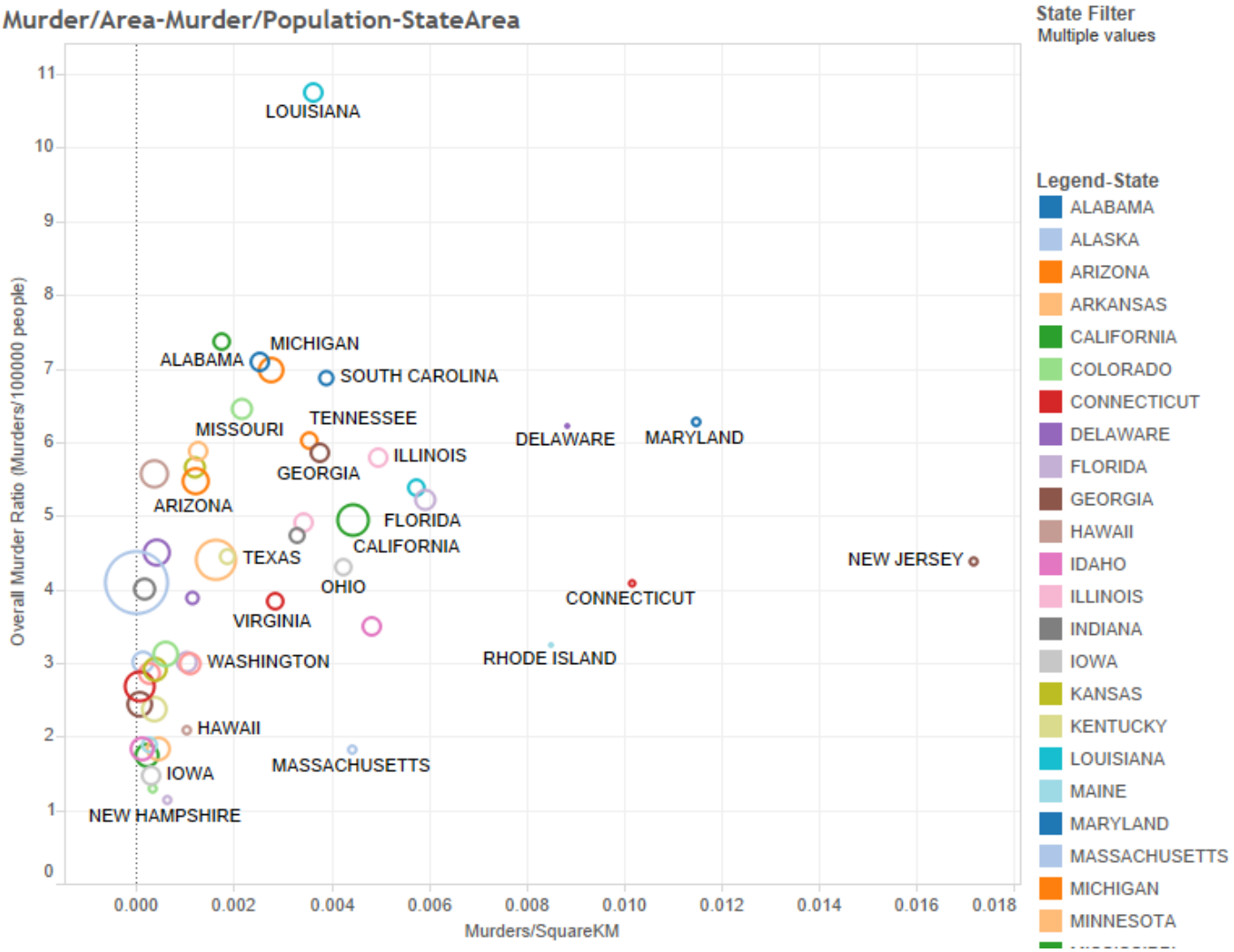
http://kunalseth.github.io/

Final Exploration:

Finally I also wanted to include a 4th dimension to our exploration and I included the area of the states. In the below visualization I have included 4 dimensions-namely murders/square kilometre, Murder/100000 which represent the x and y axis respectively; the states are represented by different colours and the area (in square km) are represented by the size.

Interpretation:
State-Ordinal Variable- Shown by position
Murders/Sqkm and Murders/100000- Quantitative Variable- Shown by Position
Area- Quantitative Variable- Shown by Size/Area

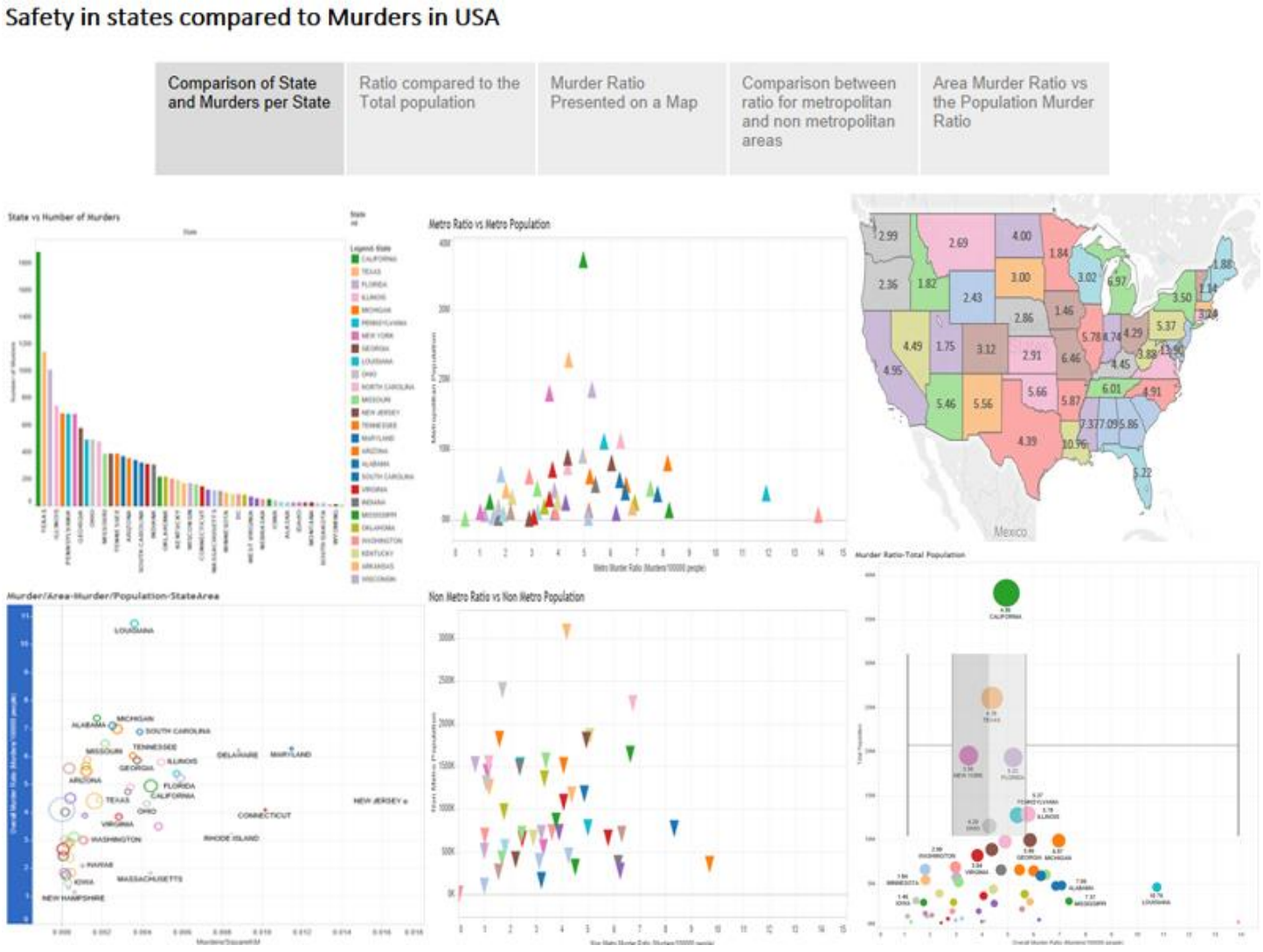
Inference:
DC comes out on top as the most murders per square km and also murders per 100000 people. The values are huge! Also other states which did not really come up when compared in murders/100000 have come up now and can help in making a more responsible decision.



http://kunalseth.github.io/

Final Conclusion:

The dataset which started with just the murders in the states of USA was explored further and compared with other factors such as state population and state area and the interpretation changed as we drilled down further. Although California had the most murders, when compared to the population and area, it actually was pretty much safe and DC came out to be the most unsafe state in terms of murder. The Final Story has been posted on Tableau public. Please view the different visualization using the tabs on the story.



http://kunalseth.github.io/