THE COMPARISON OF TOTAL FATALITIES AND ALCOHOL IMPAIRMENT 2016 AND 2017

AUTHORS:
Arsadur Rahman
Ka Hang Kwong
Naveena Anbu
Tejasri Surapaneni

OCTOBER 29, 2018

GEORGE MASON UNIVERSITY

Table of Contents

Abstract

Bad graphs are often misleading. This paper tries to redesign bad graphs/ table, creating and adding all necessary elements in new graphs which helps to visualize the data clearly. This paper used data from “2017 Fatal Motor Vehicle Crashes: Overview” report that was originally published by the U.S. Department of Transportation. The original report gave poor graphical representation in a tabular format and falls short in offering proper quantitative graphics for the state wise Total Fatalities and Alcohol – Impaired Fatalities. This paper provides various visualization techniques which includes accurate comparison of quantitative graphics, offers enough support interpretation within the graph to make it easy for a reader to analyze. The redesigned graphs in the paper contains a Micromap which helps to geovisualize the data, clustering in shiny, a Scatterplot, a Combo Graph and a Bar Graph. All these graphs display the comparison which helps a reader to envision the data points very quickly and includes how we overcame few of the challenges faced in this project.

Key words: Visualization, Redesigned, Micromap, Clustering, Geovisualize.

1.0 Introduction

In many cases, an Illustration of the data which is in the form of tables, text, spreadsheets etc., is not helpful to understand and interpret the comparisons. Later, many started visual explanation of data in form of pie chart, bar graph etc., in the rudimentary stages. As those are not helpful to explain the various contents of data, they came up with many visualization techniques which are used to present versatile data. This paper includes few of the techniques which involves the redesigning of the bad graph (Figure 1.1) by providing the essentials such as a best detailed graph helps in visualizing the differentiating the data contents like adding appropriate labels, color encoding, clear text in labels etc. to make sure the quantitative comparisons are simply understood by looking at the graphs. As the report of the data has more tables and less graphs it couldn't explain the data visually by designing appropriate graphs but represented a basic several daunting tables which holds different numbers and statistics. This redesigned paper includes interesting and useful graphs in which it compares and visualizes the data between Total Fatalities and Alcohol – Impaired Fatalities by State 2016 – 2017. The table would take a lot of time to answer some of the questions such as a) Which state had highest number of fatalities in the year of 2016 & 2017, b) Which state experienced highest number of Alcohol related fatalities in 2016 & 2017, c) Is there any correlation between the number of Total Fatalities and Alcohol related fatalities? Although this data includes few graphs, neither the data table nor the graphs is helpful to answer the above questions. This paper includes graphical representation of data points which was populated by the latest version of R (R Development Core Team, 2018) & R studio (Version 3.5.1).

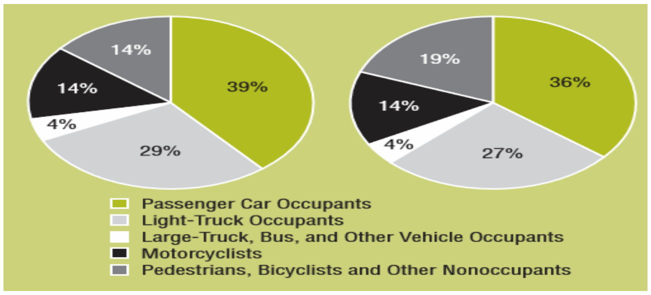

Figure 1.1: Total fatalities of different occupants in the year 2016 and 2017.

The graph (Figure 1.1) is the pie graph lacked in representing geographical specifications, total number of fatalities, accurate info., specification of years in the graph, title within the graph and failed in providing informative presentation.

2.0 Micromap

According to Daniel Carr "Micromap has the significance of introducing the visualization techniques of linking the statistical data and geographical distributions". Figure 2.1 and 2.2

illustrates the geovisualization of the data, comparing the number of alcohol impaired driving fatalities in 2016 and 2017 representing the attributes in columns for different states in US.

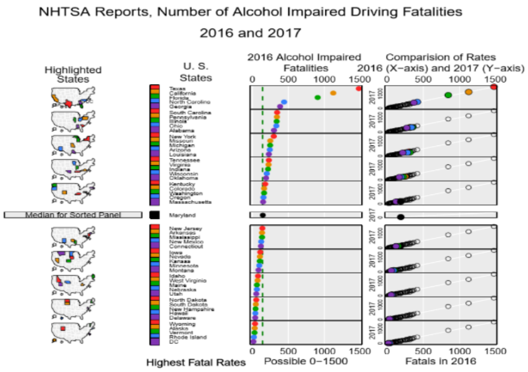

Figure 2.1: Dots with confidence interval line and scatter plot displaying 2016 vs. 2017 total number of alcohol impaired driving fatal rates.

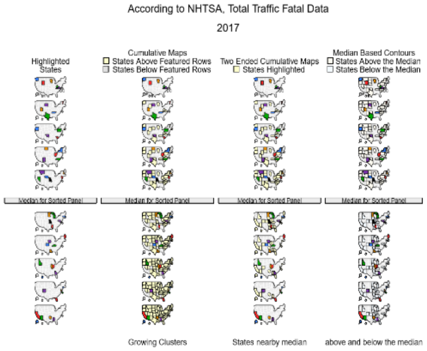

Figure 2.2: Different types of Cumulative maps and median based contours.

(Figure 2.1 & Figure 2.2: Produced with R version 3.5.1; R Studio version 1.1.456; R packages – ggplot2_3.0.0, 2016; tidyverse_1.2.1, 2017 & micromapST_1.1.1, 2014)

In the second column (Figure 2.1) a comparison is made between the two variable sets 2016 fatal rates on X-axis and 2017 fatal rates on Y-axis using scatter plot representation. The interesting point with the scatter plots in micromap is that the background dots are printed first, the median next and finally the colored dots are printed so that the important information within each panel is displayed first. Figure 2.2 displays different cumulative maps where the second column displays the growing clusters of states, third column represents the states which are nearby median, and the fourth column helps in visualizing the differentiation between the states above and below the median.

2.1 Clustering of nodes (Shiny)

Despite of the advantages of using scatterplot, there is an overlapping of the points which hides the data. To make every point clear and visible with an appropriate popup and navigating labels, seeking help of shiny made worth. Clustering concept in shiny helped in grouping the nodes into clusters so that it avoids the ambiguity and misinterpretation. The number on each cluster represents the number of the nodes grouped into cluster (Figure 2.1.1 and Figure 2.1.2).

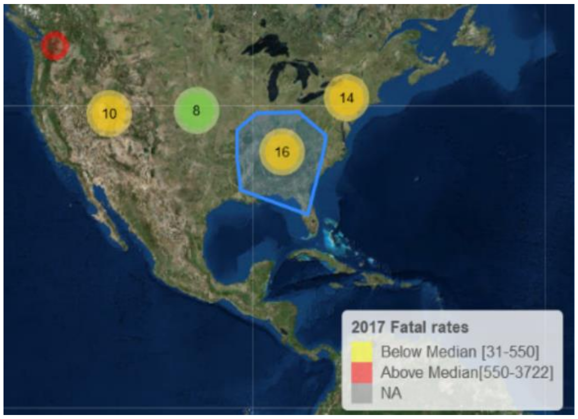

Figure 2.1.1: Clustering of nearby nodes, number on the cluster showing the number of nodes it connected blue line represents the region from which the nodes are clustered.

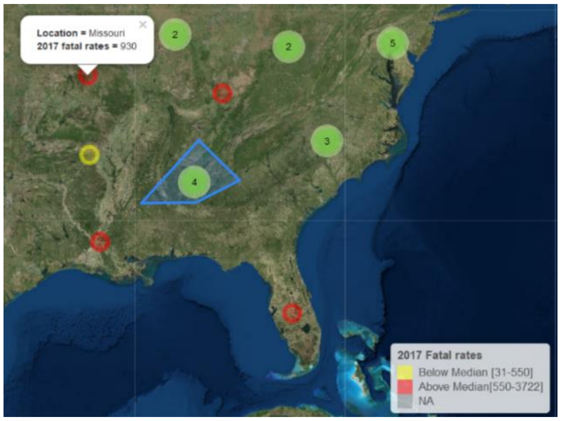

Figure 2.1.2: Each cluster having mini clusters and the yellow nodes represent the states below the median value and the red nodes represent the states above the median value.

(Figure 2.1.1 & Figure 2.1.2: Produced with R version 3.5.1; R Studio version 1.1.456; R packages – ggplot2_3.0.0, 2016; tidyverse_1.2.1, 2017; leaflet_2.0.2, 2018; sp_1.3-1, 2013; Shiny_1.1.0, 2018 & micromapST_1.1.1, 2014)

Usually in the micromap different colors are given to different states and same set of colors are repeated in each panel. Here, while clustering of nodes the total fatal rates data is categorized into 2 parts (above and below the median value) and all the nodes belonging to each part has the same color. The states having total fatal rates above the median value is given red color and states having rates less than the median value is given yellow color. Similarly for clusters, yellow clusters represent the clusters with the nodes connecting more than 9 and green represents the nodes with less than 10.

3.0 Relationship Comparison

The original data report does not have a graphical representation to show the relationship between the total fatalities accidents versus the total fatalities accidents impaired with alcohol. As a result, it did not generate enough persuasion to the reader about the importance of driving responsibly.

In order to provide more contents to the readers, the team has decided to create a graph by using the 2017 National Highway Traffic Safety Administration data to show the relationship between the total fatalities accidents versus the total fatalities accidents impaired with alcohol (need source data reference). To accomplish this task, we have chosen to use a scatterplot graph with a linear regression method to show the relationship between these 2 variables shown in Figure 4 (reference point, R version, R studio and Package). Each of the scatterplot point indicate the number of fatalities (x-axis) and the number of fatalities accidents with alcohol impaired (y-axis) for each state. As Figure 1 has shown, the fatalities accidents impaired with alcohol increases as the fatalities accidents increase. Therefore, we can suggest that there is a positive linear relation between the two variables (James et al, 2013).

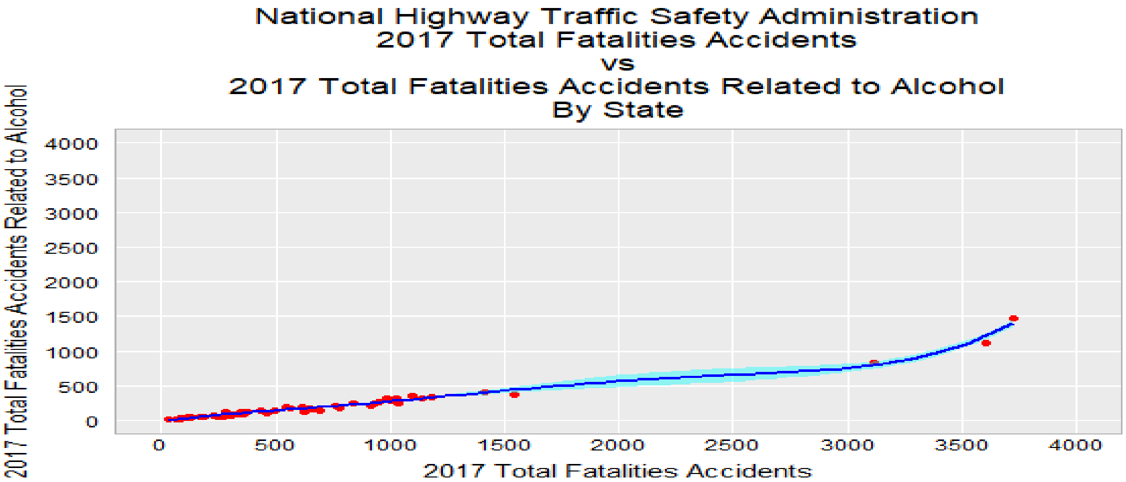

Figure 3: 2017 Total Fatalities Accidents versus 2017 Total Fatalities Accidents Related to Alcohol Impairment. The scatterplot and the smothing suggests that there is a positive linear relationhip between the two variables (James et al, 2013) (Produced with R version 3.5.1, R Studio version 1.1.456, R package – ggplot2, 2016 and tidyverse, 2017)

4.0 Combo Graph

The original table (2017 Fatal Motor Vehicle Crashes: Overview, 2018) has data for % change in total fatalities and % change in alcohol impaired driving fatalities between the years 2016 and 2017 for each state. However, since this is in table format, it is not easily visible if a %

increase in alcohol related fatalities from 2016 to 2017 directly causes a corresponding increase in the total fatalities. In order to represent this in graphical format, a combo graph was generated. The data for % change in total fatalities was sorted in descending order (largest increase to largest decrease) and plotted as a column graph.   The corresponding % change in alcohol impaired driving fatalities was plotted as a line graph.

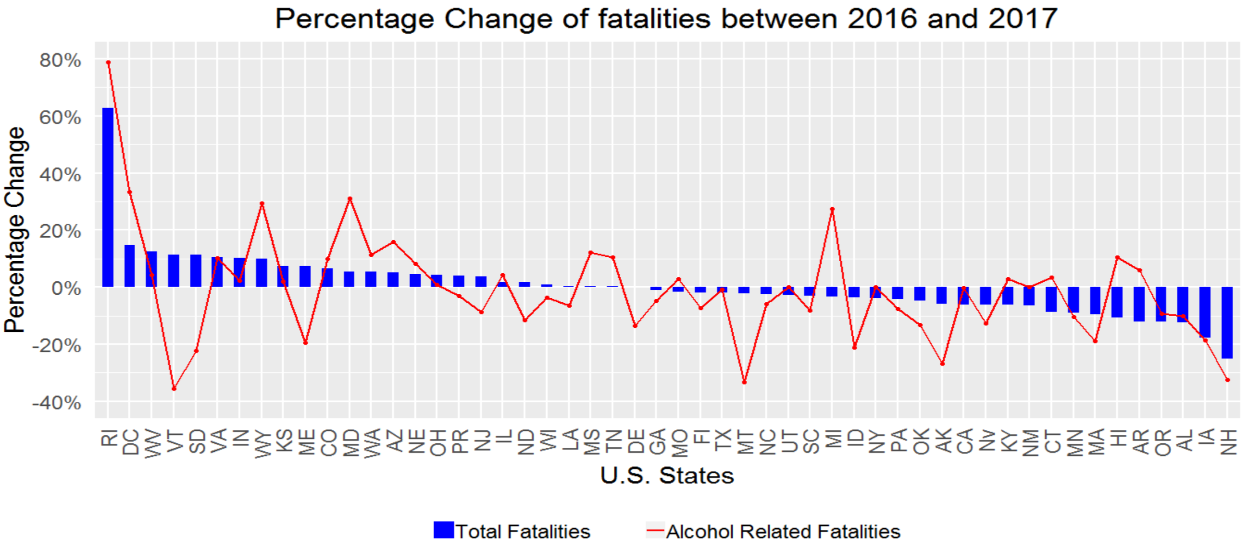

Figure 4: Comparison of percentage change in Total Fatalities and Alcohol impaired driving Fatalities between 2016 and 2017 (Produced with R version 3.5.1; R Studio version 1.1.456; R packages – ggplot2, 2016 & tidyverse, 2017)

With this visualization, a few inferences can be made. For the majority of the states, a % increase or decrease in alcohol impaired fatalities also corresponds to similar increase or decrease in total fatalities. However, for states like Vermont and Maine, the alcohol impaired fatalities have decreased significantly, whereas the total fatalities have gone up. This kind of inference can help these states in investigating other causes of the fatalities. For states like Michigan, the alcohol impaired fatalities have gone up significantly whereas the total fatalities have gone down. This kind of inference can help these states in identifying that they have a problem with drinking and driving and can help them address these issues.

Finally, in states like Maryland, both alcohol fatalities and total fatalities have gone up. But the % increase in alcohol fatalities is much more than the % increase in total fatalities. This is another case where the state should address drinking and driving issues more effectively. Therefore, with proper visualization, the states can more easily determine the trend of total fatalities and alcohol related fatalities and take suitable action.

5.0 Staked Bar Plot for data value of Total Fatalities for 2016 – 2017

The table in the original report shows the no. of fatalities in two columns for each year and the name of the state in each row. Although the states were in an alphabetic order on the main table, it was daunting to see which state the highest and which state had had the lowest no. of fatalities in both the yeas of 2016 and 2017. It was also not clear the rank of states based on the no. of fatalities each state was having. The table did not give us a good idea on the how each state is improving the road safety by reducing the no. of fatalities than the last year, which in this case,

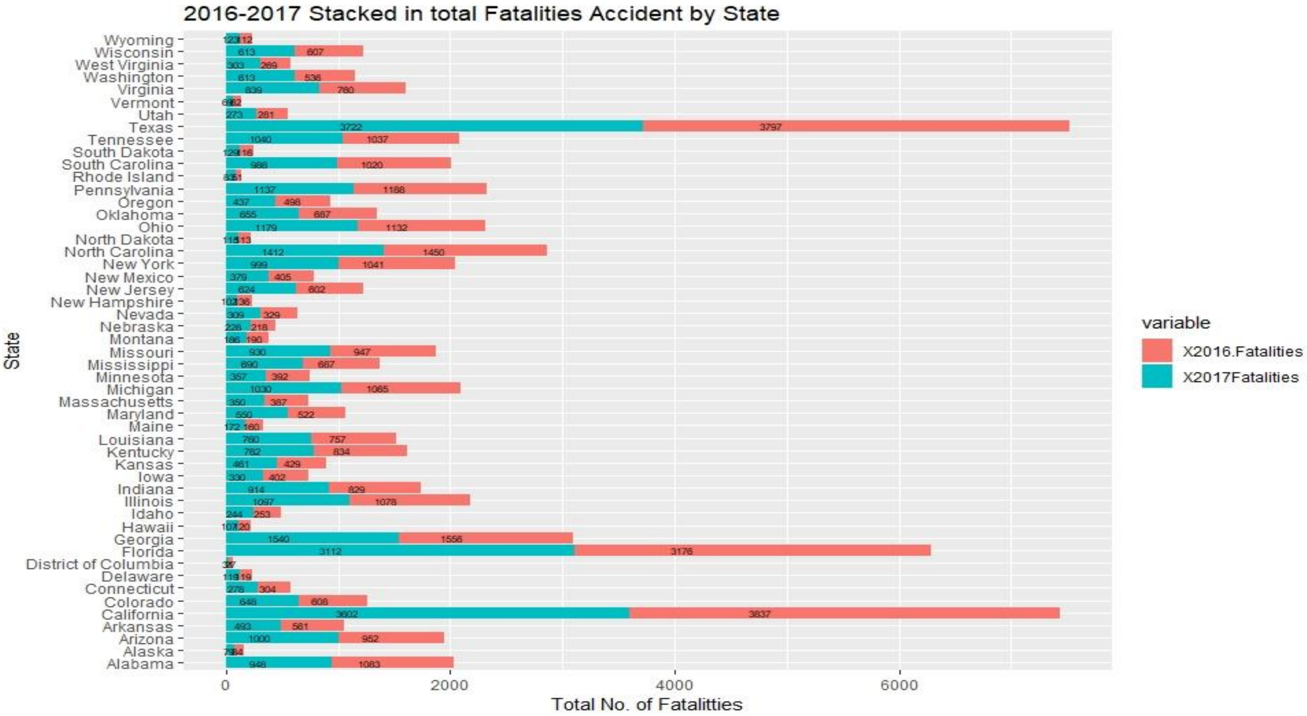

Figure 5: Stacked position for the data point of Total No. of Fatalities for the year of 2016-2917 (Produced with R version 3.5.1; R Studio version 1.1.456; R packages – ggplot2, & tidyverse,)

how the no. of fatalities increased or decreased in the year of 2017 compared to the last year 2016. Figure 5, in this paper shown the cities with highest no. fatalities in a longer stacked bar than the other sate which had less no. of total fatalities. It shows value for each length of bar the stack to differentiate the numbers and uses colors to separate the years from each other. The graphical representation which is the sorted bar plot (shown in appendix) in X axis shows the real-time rank of each state from highest to lowest by no. of fatalities in Y Axis. The bar plot shows in 2016 California was experiencing the highest number of fatalities in the country among all other states but in 2017 the state of Texas took the place of California by having the highest number of fatalities among all other states. From the Stacked bar plot, we can observe that only three states are showing big fatalities at a glance.

The stacked plot articulates an easy visual encoding for a reader to know the major facts of the given table in original report. In terms of ranks, states in the top ten each year remained the top ten for both years, and the same scenario has been observed about the bottom ten states.

6.0 Conclusion

To conclude the redesign of this project, the original contents did not provide a simplistic view of the graphical representation to the readers. As a result, it did not deliver an impactful message regarding the dangers of driving with alcohol impairment. To overcome this gap, the team has redesigned the original graph into multiple statistical graphs with detail description to help the users to comprehend the data.

Special Efforts and Challenges

During this study, there were several challenges that the team has overcome. The original data contents was in a PDF format where a conversion is needed in order for R to consume. Extensive research was done for the appropriate R codes in order to make the statistical graphs more understandable such as: datatype conversion, data sorting, grid size readjustment, trial and error different R packages, etc. In addition, the sorting orders in the micromaps with different sorting values in the same panel was difficult to manage. It was difficult to view the exact values of the data apart from the estimation based on the grid values in micromap. For the scatter plots, although micromap prints the foreground points to be visualized first, there is still an overlap where the points are not visualized clearly.

As to the special efforts, the shiny code took a lot of time and patience to see the desired outcome. During the development, it had to replicate the data first and then made trials on executing each part of the code so that it wouldn't affect the original data values like redundancy etc. Although the initial clusters were created successfully, it had an issue with duplicate data along with fake assigned values. The data cleansing process took couple days to complete along with an extensive research by reading through few articles such as (Journal of Statistical Software) and books on shiny and micromap. The research has helped to create such visualizations and clusters with right popups and navigating labels.

While designing the combo graph, % symbol had to be removed from the data before converting the datatype to numerical or else the negative % change in data were not properly visualized. Additionally, states name in the combo graph have been abbreviated through R (R Core Team, 2018; R studio, 2016) or else the state name occupied one-third of graph area by reducing the size of graph.

Overall, this project has helped us to become more familiar with the statistical graphs, R coding, overcome an unexpected situation, collaborate with teammates across the world and the preparation of the future.

Reference

2017 fatal motor vehicle crashes; overview research note. (2018, October). Retrieved from https://www.nhtsa.gov/press-releases/us-dot-announces-2017-roadway-fatalities-down

Carr DB, Pearson Jr JB, Pickle LW (2013). micromapST: State Linked Micromap Plots. R packa ge version 1.02, URL http://CRAN.R-project.org/package=micromapST.

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

James, G., Witten, D., Hastie, T. & Tibshirani R. (2013). An Introduction to Statistic Learning. New York: Springer.

Joe Cheng, Bhaskar Karambelkar and Yihui Xie (2018). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.2. https://CRAN.R-project.org/package=leaflet.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio, Inc (2014). shiny: State Linked Micromap Plots. R package version 0.10.2.1, URL http://CRAN.R-project.org/package=shiny.

RStudio (2016). RStudio: Integrated development environment for R (Version 1.1.456) [Computer software]. Boston, MA. Retrieved July 19, 2018. http://www.rstudio.com/.

Payton Q, McManus M, Weber M, Olsen T, Kincaid T (2015). “micromap: A Package for Linked Micromaps.” Journal of Statistical Software, 63, 1–16. URL http://www. jstatsoft.org/v63/i02/.

Wickham H (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York. URL http://had.co.nz/ggplot2/book.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.1.0. https://CRAN.R project.org/package=shiny

Appendix

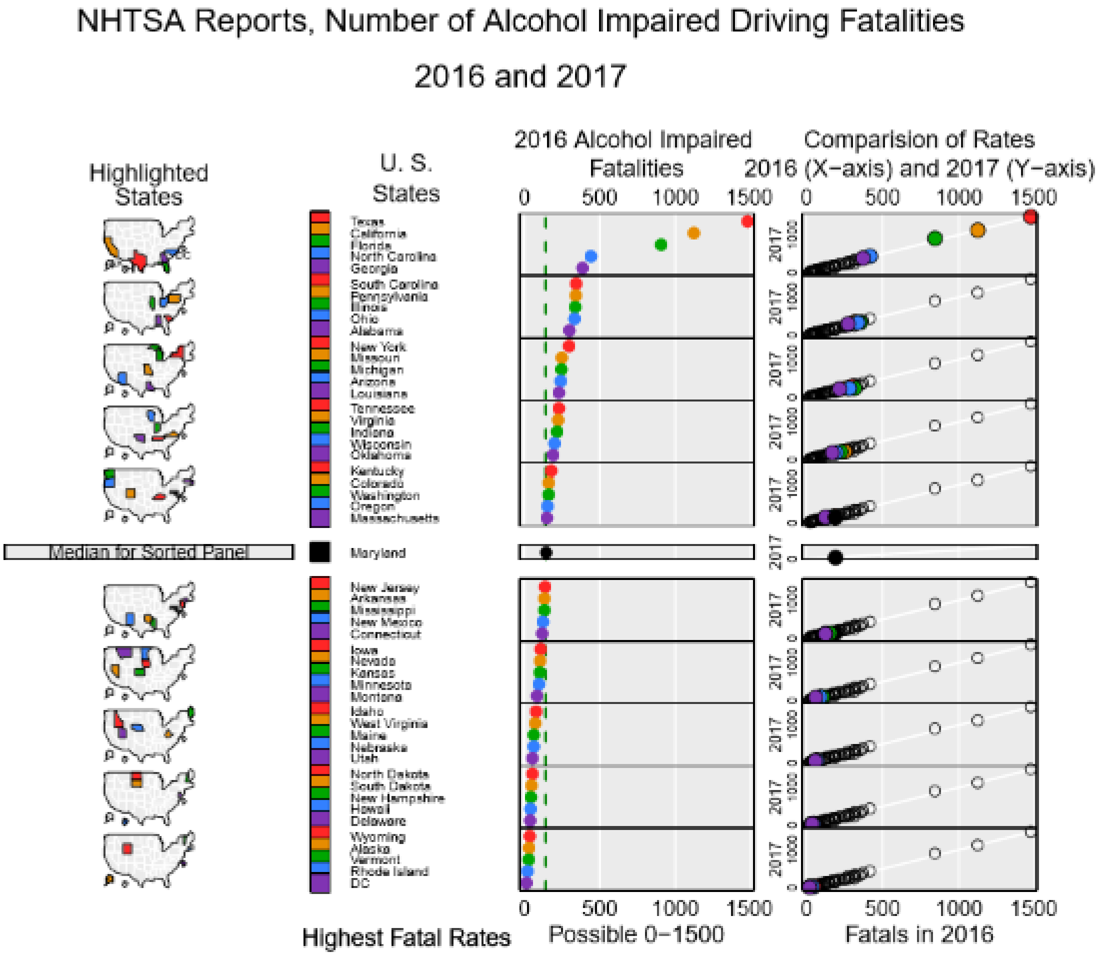

Figure 2.1: Dots with confidence interval line and scatter plot displaying 2016 vs. 2017 total number of alcohol impaired driving fatal rates.

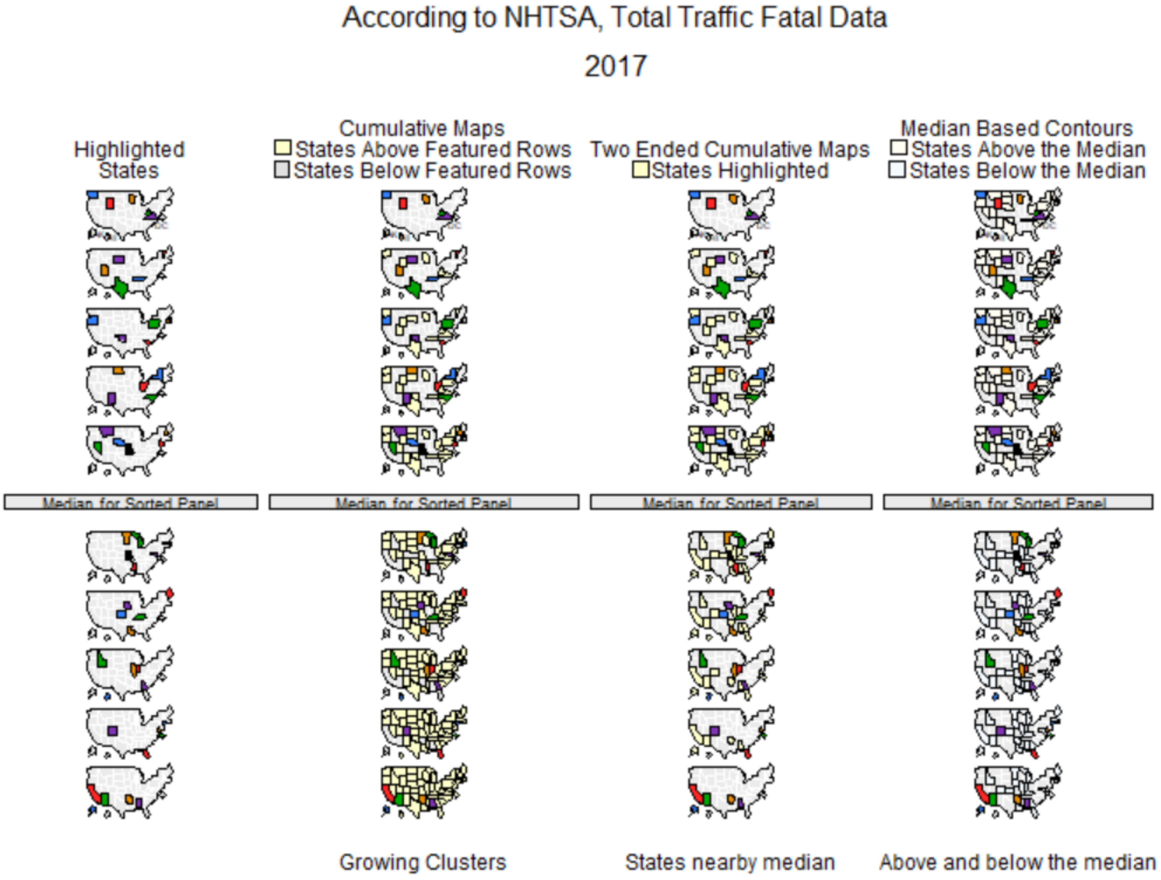

Figure 2.2: Different types of Cumulative maps and median based contours.

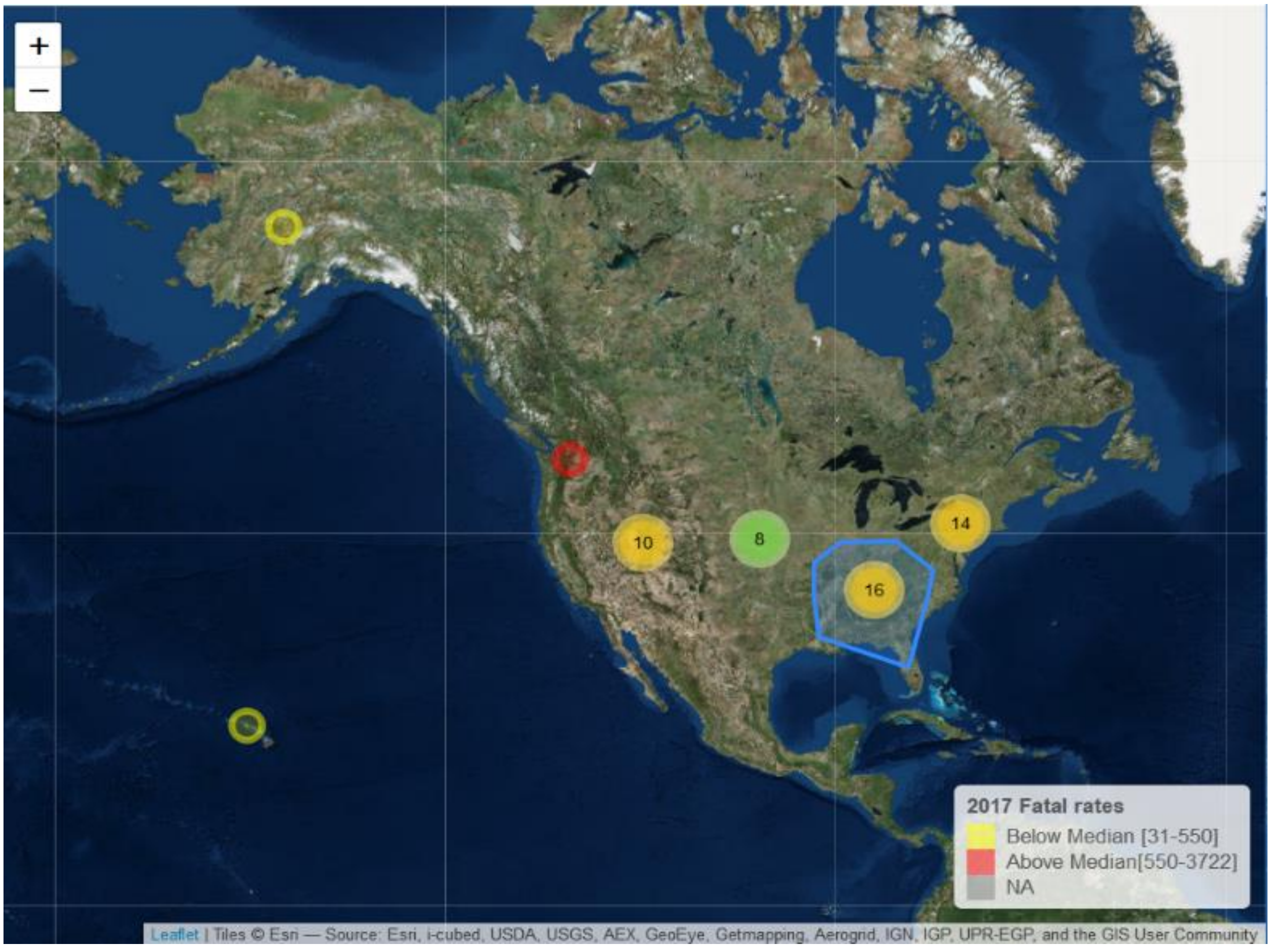
+
−
10
8
14
16
2017 Fatal rates
🟨 Below Median [31-550]
🟥 Above Median[550-3722]
⬜ NA
Leaflet | Tiles © Esri — Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community

Figure 2.1.1: Clustering of nearby nodes, number on the cluster showing the number of nodes it connected blue line represents the region from which the are clustered. Clusters which connects more than 9 nodes are represented with yellow color and cluster which connects less than 10 are represented with green color.

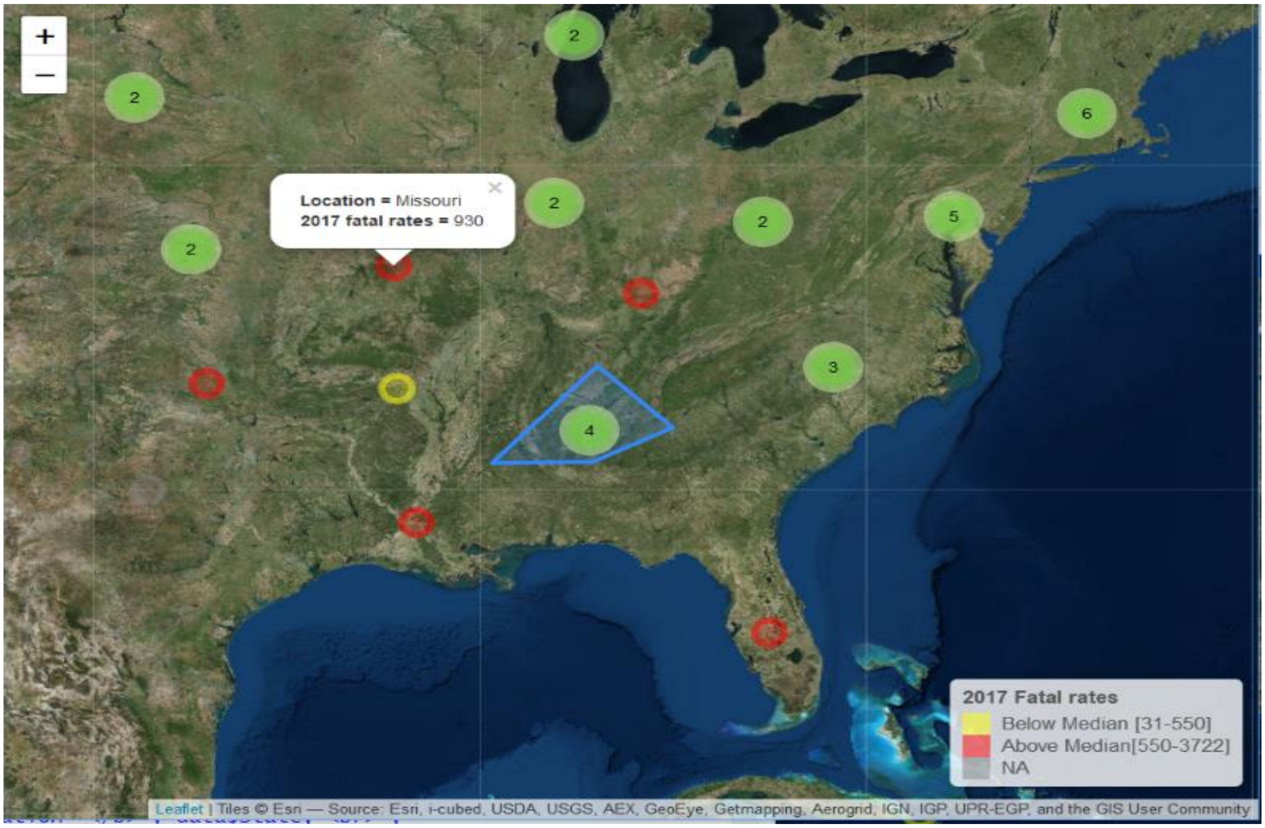
+
−
2
2
6
Location = Missouri
2017 fatal rates = 930
2
2
5
2
3
4
2017 Fatal rates
🟨 Below Median [31-550]
🟥 Above Median[550-3722]
▨ NA
Leaflet | Tiles © Esri — Source: Esri, i-cubed, USDA, USGS, AEX, GeoEye, Getmapping, Aerogrid, IGN, IGP, UPR-EGP, and the GIS User Community

Figure 2.1.2: Each cluster having mini clusters and the yellow nodes represent the states below the median value and the red nodes represent the states above the median value.

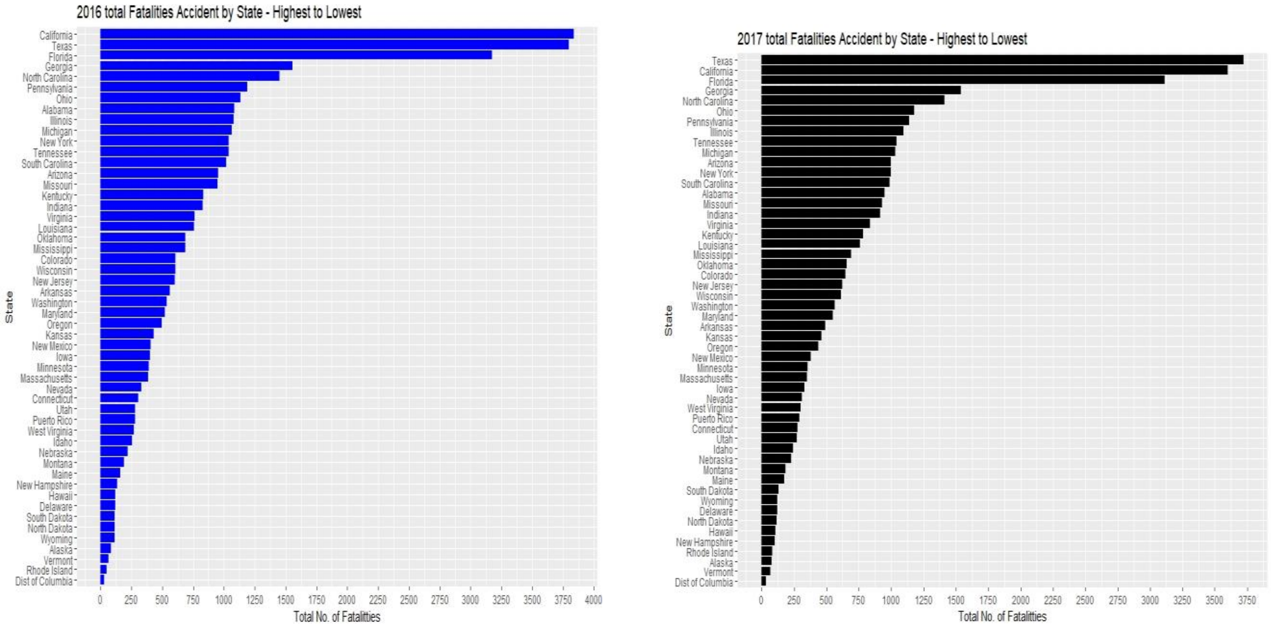

Figure 5.1: Side by side comparison of 2016 Total Fatalities Accidents & 2017 Total Fatalities Accidents; Best & worst perform