Alex Chen
August 7, 2017

Data Visualization and Analysis through R and Tableau in 2017 VAST Challenge

The VAST(Visual Analytics Science and Technology) Challenge is a competition designed for researchers to use different types of software in order to answer data analysis questions. Each year, the challenge consists of some realistic context and data for researchers to analyze and draw conclusions through data visualizations to answer complex questions. These complex questions pertain to human behavior backed up by patterns shown through data. Researchers submit one or two pages of analysis and visualizations to help support their response to each question. In the 2017 VAST Challenge, researchers were challenged to analyze environmental damage in a park. The challenge was split up into three different sections, all examining a different possible source of damage. To answer these questions on possible human behavior affecting the environment, data has to be looked at manipulated and organized in different ways in order to observe patterns both spatially and temporally. Only after that process, can we begin to make hypotheses on human behavior and attempt to validate them with data visualizations. In this paper, we'll be looking at the data visualization portion of the challenge and analyze two different tools of data visualization: R and Tableau. To answer the questions of effectiveness of each tool, examples will be drawn through the 2017 VAST Challenge.

To preface both methods, R is an open source programming language that is often used as a statistical data analysis and visualization environment. Statisticians and data-miners most commonly use R for statistical computing. One of the more unique aspects of R is the availability of packages to use. Packages give users more diversity in the uses of R beyond the base programming availability. Each package contains different functions that extend R's capabilities. In Tableau, the focus on the application is data visualization. Tableau allows users to input a data set and then create different types of graphs. The main focus in Tableau is in exploring and analyzing relational databases in a visual format. In addition, Tableau contains a mapping function that allows users to input coordinates in order to map out data based on location. A general comparison between the two is: R is a sandbox where users can make almost anything whereas Tableau is a classroom where users can use selected tools.

The first step of analyzing a data set for human behavior is to explore the data for both temporal and spatial patterns. For these types of exploration, Tableau allows users to very easily import the data and begin playing around with variables. Although Tableau is very limited on the number of different types of graphs it offers, it is very easy to start using for both spatial and temporal patterns. R, however, requires a little more knowledge in order to begin using as well as some data cleaning after importing the data set. On the other hand, R can offer almost every type of graph imaginable but still requires that pre-existing knowledge in order to begin to investigate temporal patterns. While both analytical tools offer capability to show spatial patterns in pre-existing maps, importing a map image and drawing over it is still proving difficult in both R and Tableau. All of this is assuming the data sets were complete and without missing data. Because R is a programming language, data sets can be cleaned and manipulated in order to create additional temporal data. To be most time-efficient, users could use R to clean and manipulate the data, then Tableau to create data visualizations very quickly.

After the creation of data visualizations for patterns, users analyze the data visualizations for outliers and surprises. While looking for outliers, Tableau visualizations allow for a mouse-over function that can identify the points where there might be an outlier or surprise.

Depending on the package in R, the same can be done. In analyzing the data, users will want to identify specific outliers that could be potentially errors in the data. In removing those data points, R allows more easy capability in removing the errors with one function while Tableau users would have to remove the data from the data set manually one by one.

As I mentioned previously, Tableau's strengths lie in the data visualizations of the data set while R is stronger in statistical computing and manipulating data sets. After removing outliers, pattern analysis and making hypotheses on the reasoning behind patterns is most effectively done through R. When creating more in-depth graphs to show clearer patterns, R can more easily separate a data set into different data sets based on a factor such as car type. Other than factor, users can look into seasonal, monthly, or even daily patterns. Those different data sets provide different possible patterns users could analyze. After creating the different types of data sets, data visualization, once again, can be made through Tableau to show additional patterns. Beyond these data visualizations, it's up to the users to find specific patterns and draw conclusions from those data visualizations.

When confirming those patterns, users should look to creating different types of graphs based on the same pattern to cross-reference each data visualization in order to check whether the hypothesis holds true among all the graphs. If one pattern holds true, another pattern can usually be drawn and tested against the hypothesis. By creating the different graphs in R and cross-checking with Tableau, users can re-affirm or reject their hypothesis.

In conclusion, both R and Tableau are very effective at doing their intended jobs. In Tableau, the creation of data visualizations is extremely quick and easy which makes it a very good tool for exploration of general patterns. R excels in looking more in-depth into the data and being able to draw and confirm hypotheses. By using both tools concurrently, users can greatly reduce the amount of time it takes to go through this kind of data analysis process. Once again, R is a sandbox while Tableau is a classroom. In most aspects, R is the tool to use, but has a much higher learning curve than Tableau which provides a very user-friendly environment for users to work with preselected tools. When looking at two relatively similar environments for making data visualizations, both have their benefits and faults when it comes to depth and efficiency, but using both R and Tableau has proven effective for the data analysis in the 2017 VAST Challenge.