Data Visualization Write-Up by Patrick Flynn

June 1, 2019

1 Officer Involved Shooting Analysis

1.0.1 By Patrick Flynn

1.1 Summary

Three datasets comprised of officer involved shooting data from three separate neighboring counties (in an anonymous US midwest state) were gathered and processed. A K-Means clustering algorithm (with n=15 clusters) was applied to the dataset. The datapoints were laid out on a timeline for each year, and given a color depicting what cluster they belong to. The resulting visualization shows how officer involved shootings occur in clusters and frequently during the summer months occur in rapid succession.

Final Visualization: - Tableau Public - Patrick Flynn - Officer Involved Shooting Clusters - https://public.tableau.com/profile/patrick.flynn5461#!/vizhome/Officer_Shootings/Story

Initial Visualization (VERY ROUGH!) - Tableau Public - Patrick Flynn - Officer Involved Shooting Timeline - https://public.tableau.com/profile/patrick.flynn5461#!/vizhome/OfficerShootingsInitialTimeline

1.2 Design

Determining how to visually demonstrate clusters of single dates presented quite a challenge. Presenting the clusters had to accomplish the following:

- Group officer involved shootings by year
- Easily identify what events were part of a particular cluster
- Highlight key clusters of interest to the viewer of the visualization
- Present data (such as averages, etc.) to the viewer without the use of interaction for key clusters

An initial sketch was created that did not differ that much from the final product. It was during the actual development of the visualization that other types of visualizations were tested and presented with my team. The team and I agreed that that the initial sketch illustrated and visualized the clusters in a way that made the most sense to a person viewing the data for the first time.

The design went through several iterations, one concept was to originally group the data by season or hour in the day. This proved to not be as effective as presenting the data in the final version. What was also important in the design of the visualization of the product was to show that, while many officer involved shootings occur rapidly after another, there were also times in the dataset where a LARGE gap was present between shootings.

Because this visualization was also used for exploratory data analysis, many of the callouts on the visualization were discovered as the visualization was built and investigated. The intention of this visualization was to determine if, by applying K-Means clustering, a discernable pattern would emerge from the dataset. After visualizing the data, it can be concluded that there appears to be a pattern in officer involved shootings. Further research will be performed.

1.3 Feedback

The data was initially presented to members of my analytical team with the expressed intention of being able to accurately model and demonstrate whether officer involved shootings fell into some sort of pattern. Initially one of the problems with my visualization was the color choices. I initially used a "blue" spectrum to display the clusters. My team expressed that this made the distinction between clusters very difficult to determine.

The very first iteration of the visualization had all of the datapoints on one singular timeline. While I printed this visualization on an 11x17 sized paper, members of my team explained that it was easier to see clusters that shared two years (i.e. events in the end of 2016 and the beginning of 2017) however the datapoints were too small and long. This is when I made the decision to split the points/clusters by year instead of laying them out on one single timeline. By doing this, it also illuminated some trends that occured in the summer months of the years in regards to shootings. Doing this also highlighted the large gaps that occured in 2014 and 2016.

Finally, the clusters were labeled based on the "labs" column derived by the K-Means model. This was initially the label given to each cluster on the visualization. Members of the team indicated that this label was nothing more than a placeholder and that instead a date range would be far more beneficial to readers/viewers. This was time consuming because it required re-manipulating the original dataframe in Pandas during the execution of the code, but in the final product it is far easier to determine exactly when clusters occured.

1.4 Resources

Python Pandas Package Documentation
Python SciKit Learn Clustering Documentation
Altair Visualization Platform (used for inspiration)