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 visualizations shows how officer involved shootings occur in clusters and frequently during the summer months occur in rapid succession. The heatmap and stacked bar charts demonstrate how 2017 and 2018 officer involved shootings see a dramatic rise.

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

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

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
- Look at data visualized in different ways (i.e. heatmap/barplot/etc.)

The initial design of the visualization was a simple timeline. This initial version was very difficult to work with any made visualization of clusters a challenge. Clusters were all one shade of blue and fell on one timeline - essentially a straight line. The first real initial visualization was a scatter plot where the colors of the points were their membership to a cluster. This was a good way to visualize the clusters however using different pages in the story made the visualization difficult to track. The final visualization has many enhanced elements compared to the initial version.

Slide One (Dashboard) Having a barchart breakdown by number of officer involved shooting per year was an easy way to illustrate that 2017/2018 have some sort of problem occuring as they are much higher than the rest. The heatmap(hotspot) visualization keys in on two trends that are apparent: summer months have more officer involved shootings and August of 2018 was particularly bad with 7 shootings alone. Finally the timeline serves to display a rolling trend by using a line chart. This line chart illustrates particular spikes that can be used for further analysis/review.

Slide Two (Stacked Bar Chart) The purpose of the stacked bar chart broken down (colorized) by year displays the 2017 and 2018 data to the user in a way that points them out (also utilizing Tableau highlighting). The absense of 2018 Q3 data also shows that these months have the potential to increase the overall officer involved shooting numbers dramatically. A stacked barchart was an easy choice to show the year while also displaying the month and number of shootings without too "busy" of a visualization.

Slide Three (Cluster Colored Scatter Plot) The final visualization in the story is the scatter plot colorized by cluster. This visualization is the heart and mind of the entire analysis/visualization. Changes were made to the initial design to remove the slides and instead add filters depending on the year and length of an OIS cluster. In addition, 6 dense cluster data points were marked/labeled to draw the viewers attention to those points - indicating they are different.

1.3 Feedback

Feedback on Initial Design 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.

Feedback Leading to Finished Visualization The initial visualization was lacking a variety of visualizations to display to the viewer. The final version rectified that by including a stacked barchart, barchart, heatmap, and linechart. In addition, the scatter plot colorized by cluster had too many slides that were essentially displaying the same data. Instead by adding the filters to

the visualization, the user can determine points of interest. To ensure that the key items were not missed by the viewer, labels were added to important data points (mentioned above).

1.4 Analytical Conclusions

With several of the officer involved shootings occuring in the summer, and August of 2018 being a critical example of this, further analysis and research needs to be performed to discover why these shootings increase during the summer. Additional research needs to be performed to discern why officer involved shootings often occur in tight clusters. Does one officer involved shooting beget another? This visualization serves as a stepping-off point and discussion platform to provide to decision makers/police/community leaders to decrease the trend visualized in this product.

1.5 Resources

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