K-Means Clustering – Firearm Background Checks and Gun Violence

Patrick Weatherford

Bellevue University

DSC 500 T301 2215-1 Spring 2021

Abstract

K-Means Clustering is an unsupervised model used to dynamically classify data. In this exercise, I found data related to Firearm Background Checks and Gun Violence in the year 2020 and performed K-Means Clustering analysis to see if any sort of group or pattern could be determined. The analysis shows that clusters were indeed formed which showed t and the K-Means Clustering methodology was effective for this problem, but also there were issues common with using this model that had to be taken into consideration.

K-Means Clustering – Firearm Background Checks and Gun Violence

# Data Source

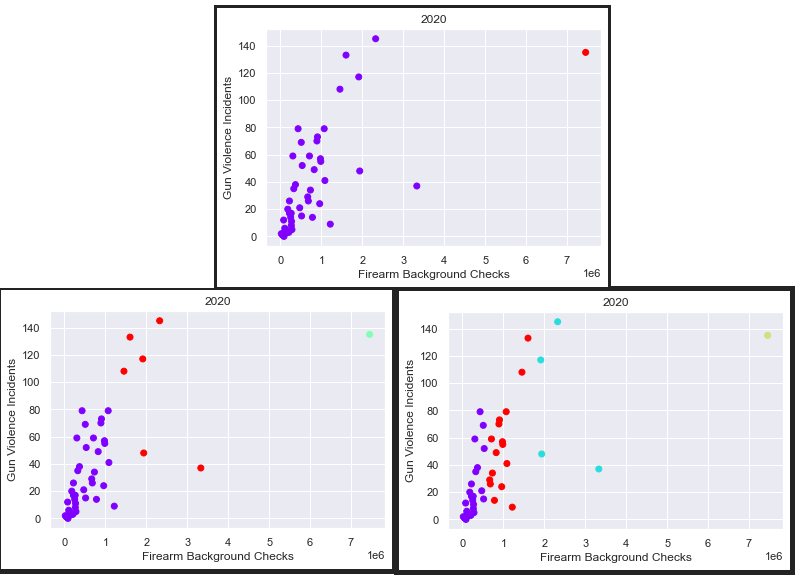
The data source was located using data from ‘Gun Violence Archive’ and from public FBI documentation that shows firearm background checks per year/month/state. The data was then reconciled into a .csv file. From there the data was read using a python script and various data analysis libraries.

**Initial Analysis**

Based on certain criteria fed into the model, the algorithm will automatically group data based on the data points’ relationship to centroids which are first chosen at random by the model and then recalculated as the model iterates through the algorithm. The number of centroids and clusters are determined using a single parameter, labeled “K” in the algorithm formula. Figure 1 below, shows the groups formed using K=2, K=3, and K=4.

**Figure 1**

*K-Means Clustering Plot Chart*

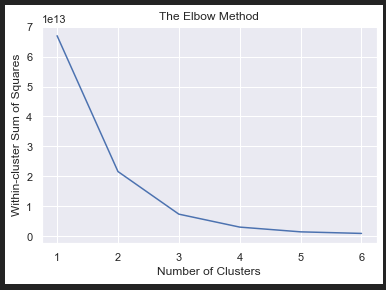


**Determining Number of Clusters “K”**

To determine the appropriate number of clusters to use for the model, the Elbow Method was utilized which compares the Sum of Squares, or variation, of data points within the group for models using different inputs for “K”. The best input to choose for “K” was 3 because it was the point where increasing anymore would no longer drastically affect the variation within the clusters. Figure 2 below represents this the plotting of the Elbow Method.

**Figure 2**

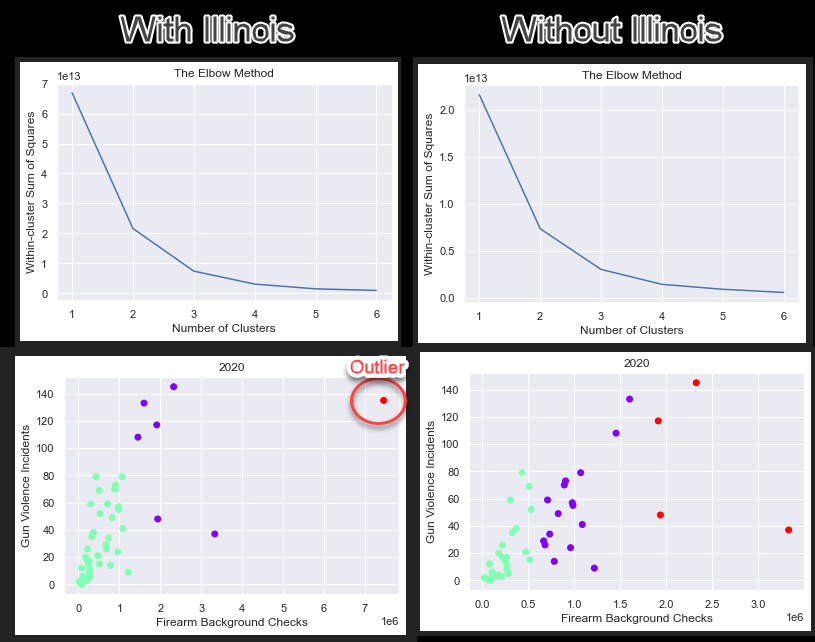
*Elbow Method Plot to Determine “K”*



**Removing Outliers**

One of the caveats to using the K-Means Clustering model is that outliers can drastically affect the prediction. For example, in Illinois, the number of Gun Violent Incidents and Background checks is considerably higher compared to other states within the U.S. Because of this, it would probably be better to remove this state from the data frame as shown in Figure 3 below.

**Figure 3**

*K-Means Cluster Comparison Both With and Without Outlier*

References

NICS Firearm Checks: Month/Year by State — FBI. (2021) <https://www.fbi.gov/file-repository/nics_firearm_checks_-_month_year_by_state.pdf/view>

SEARCH INCIDENTS. (2021) <https://www.gunviolencearchive.org/query>