Exploring Shooting Incident Rates with Demographic and Congressional Seat Party Affiliation Data

Alex ONeill
Seidenberg School of CSIS
Pace University
New York, NY
ao24521p@pace.edu

***Abstract*—By exploring the relationships between demographics, political control, and shooting incident rates, we can better understand how to address issues of crime in our society. This paper will explore the relationships between education levels, poverty rates, race, age, gender, voter engagement, political affiliation of congressional seats, and rates of shooting incidents. Key correlations identified during the exploratory analysis will be evaluated for strength with linear regression and k-nearest neighbor models.**

Keywords—crime, shooting incident, congressional district, education, poverty, linear regression, k-nearest neighbor

I. INTRODUCTION

Trying to address and understand the root cause of crime has been a topic that has been researched for years by politicians, data scientists, criminology experts, and those in the social sciences community. Poverty and a lack of education are typically the main factors that many people attribute high crime rates to, but recently, politicians have taken a more partisan stance on crime. Politicians have used crime rate statistics to influence gun legislation that lands on opposite sides of the spectrum and can frequently be quoted blaming the rival party for high crime rates. In an effort to provide clarity and better understand the validity of some of these claims, this project will explore gun violence rates grouped by congressional district and the relationship to political party and district level demographic data. It is expected that there will be correlations between many of the demographic and political data and the rate of shooting incidents per district.

By doing an exploratory analysis in the data, it will be shown that there is a relationship between education level and the rate of shooting incidents. Furthermore, poverty rates and political engagement (through voting rates of eligible voters) will also be shown to have a correlation with the rates of shooting incidents. Lastly, it will be shown that there is strong evidence to support the claim that, “Democrat run districts are more dangerous than Republican run districts”. This evidence will be corollary in nature and care should be taken to avoid making assumptions related to causality.

After demonstrating these corollary relationships, the strength of these correlations will be put to the test by exploring their predictive power with linear regression and k-nearest neighbor algorithms. A attempt will be made to predict how dangerous a district will be (rates of shooting incidents) by combining education, poverty, voting rate, and political affiliation data.

II. METHODS

A. Data Sources

Data was collected for this experiment from three different sources online. Shooting incident data was provided by kaggle.com [1]. This dataset is a collection of 260k plus reports of shooting incidents in the United States between the years of 2013 and 2018. It includes various details of the incident including date, location, congressional district, weapons used, etc. Congressional Representative data was collected from everypolitican.org and includes congressional district, political party, term, and name of each Representative from the terms selected [2]. Demographic data was sourced from the census.gov website [3]. This subset of census data was collected during the 2018 midterm elections and is organized by congressional district. It includes data regarding education level, race, gender, poverty rate, and voting rate.

B. Preprocessing

Although not complex, there was a considerable amount of manual preprocessing required to create features that were consistent and able to used as keys for joining datasets. The shooting incident dataset included the "State" that the incident occurred, but it was shown as the full state name. This required a conversion from "State" to "State_Abbreviation". The shooting incident dataset also included the congressional district that the shooting incident occurred, but it was not joined with any other attribute. Both the census dataset and representative dataset included a district feature that was formatted as the state abbreviation combined with the district number, e.g. (CA-4). Once a conversion to state abbreviation was completed in the shooting incident dataset, it was concatenated with the congressional district. Some additional adjustment was required to standardize district numbers for states that only have one district. These were labeled as "At-large".

Some additional preprocessing was required to combine multiple smaller datasets into a master sheet that could be used for analysis. The raw census data was provided based on subtopic, e.g. (education, age and race). Representative data was also provided in smaller datasets that were broken down by congressional term. As these smaller datasets were read-in, duplicate and unnecessary features were dropped, then combined to form a master sheets for census data and representative data.

The final phase of preprocessing was to combine shooting incident data with census data. This was completed by dropping

shooting incidents that occurred during the years 2013 and 2018 as they were incomplete, grouping incident counts by congressional district, then joining this grouped data with the census master data. This provided a combined sheet that included the count of shooting incidents alongside census features grouped by district.

C. Exploratory Analysis

An exploratory analysis was conducted to identify feature correlations through visualization techniques.

1) Congressional seats over time: A clustered bar chart was produced to visualize party affiliation of congression seats over time between the years 2003 and 2019. The clusters of these bars were grouped by congressional term which last two years and run from January to January.

2) Shooting incidents by month: To better understand shooting incident trends, a line chart was created based on the number of shooting incidents by month and year. Since full shooting incident data was only available for the years 2014-2017, only these four years were plotted. Each year was plotted as a different line with the month number remaining constant.

3) Shooting incidents over time: While the visualization of shooting incidents by month was useful for identifying month-to-month trends, it lacked in its ability to show trends year-over-year. To better understand the year-over-year trends, a line graph with a trend analysis line was created. Shooting incident counts were grouped by months, and all months were laid out sequentially with no break for subsequent years.

4) Shooting incidents per capita by percentages of race: A stacked bar graph was created to visualize the percentages of certain races in each district compared to the rate of shooting incidents per capita.

5) Shooting incidents per capita by percentages of education levels: A stacked bar graph was created to visualize the percentages of increasing education levels in each district compared to the rate of shooting incidents per capita.

6) Shooting incidents per capita by percentages of age brackets: A stacked bar graph was created to visualize the percentages of different age brackets in each district compared to the rate of shooting incidents per capita.

7) Shooting incidents per capita by percentages of gender: A stacked bar graph was created to visualize the percentages of each gender in each district compared to the rate of shooting incidents per capita.

8) Shooting incidents per capita by percentage of the population in poverty: A stacked bar graph was created to visualize the percentage of the populating in poverty in each district compared to the rate of shooting incidents per capita.

9) Shooting incidents per capita by percentage of voter engagement: A stacked bar graph was created to visualize the percent of the population who exercised their right to vote during the 2018 midterm election in each district compared to the rate of shooting incidents per capita.

10) Shooting incidents per capita by party affiliation of congressional seats: An overlay area chart was created to visualize the standard deviation of shooting incidents per capita in each district grouped by the party affiliation of the congressional seat of each district during the 114th Congress which ran from 2015 to 2017.

D. Predicting Rate of Shooting Incidents

To further evaluate the strength of the correlations observed in the exploratory data analysis phase of this project, attributes that were observed to have a corollary relationship with the rate of shooting incidents were processed through the linear regression and k-nearest neighbor models to better understand the predictive value of these features.

1) Linear regression: To perform a linear regression analysis, all continous features were normalized to the standard deviation of that feature and the single categorical feature (political party) was one-hot encoded. Once this normalization was completed, all entries that included any variable which was greater than ***two*** standard deviations was removed from the dataset. The remaining independent variables consisted of education levels, poverty rate, voting rate, and political party. The dependent variable was the standard deviation of the per capita rate of shooting incidents.

2) Classifying districts with k-nearest neighbor: Developing a more user-friendly classification system was done by discretizing the dependent variable (standard deviation of shooting incidents per capita) into five bins. These bins would represent the safest, safe, neutral, dangerous, and most dangerous districts. Prior to discretizing, all independent and dependent variables were normalized to the standard devation of that feature and the single categorical feature (political party) was one-hot encoded. Once this normalization was completed, all entries that included any variable which was greater than ***three*** standard deviations was removed from the dataset. This widening of tolerance from ***two*** standard deviations in linear regression to ***three*** was decided in an effort to retain a larger number of data points who may have had a slight outlier in only one variable. All continuous independent variables were then discretized into five bins and one-hot encoded, and the classifier was set with n_neighbors=15 and weights="distance".

III. Results

There are a few areas in both the exploratory analysis and prediction analysis where the results warrant further discussion. The results in this case will be evaluated on an individual basis as developing a broader picture as to how the results connect with one another is outside the scope of this analysis and would require additional data and social science expertise.

A. Exploratory Analysis

1) Congressional seats over time: The number of congressional seats by political party over time indicated two behavioral trends. While there is variation in the number of seats by party during each election, the changes are generally very small. Even though the changes are generally small and ownship is typically stable, there are two times that a large number of seats change hands. Democrats picked up a large number of seats for the 2007-2009 session, and more for the 2009-2011 session. Republicans then reclaimed full ownership for the 2011-2013 session. Although the timing of the elections

does not directly coincide with the 2008 financial crisis and 9/11, these extreme events may contribute to volatility.

2) Rates of shooting incidents: There are two primary trends identified during this section of exploratory analysis. Shooting incidents sharply decrease during the month of February and, in general, during the winter months. This can likely be explained by the difference in the number of people outdoors in each season and due to February having less days in the month. Year-over-year, shooting incidents steadily increased from 2014 to the start of 2018. In Fig. 1, you can observe the trend-line of increasing shooting incidents.

3) Demographic data with no correlation to shooting incidents: When visualizating how race, gender, and age correlated with rates of shooting incidents in each district, there was no observable trend. Age bracket and gender percentages by district were fairly consistent across the board. As the rate of shooting incidents increased, there was no change in these percentages. On the other hand, the percentages of each race per district did vary greatly. There were a large number of districts with missing data on certain races, but even so, no pattern was noticed in the difference of race percentages when compared to the rate of shooting incidents.

4) Demographic data with correlation to shooting incidents: When visualizating how education level, poverty rates, and voting rates correlated with rates of shooting incidents in each district, there was an observable trend. As the highest-level of education percentage decreased, the rate of shooting incidents appears to increase. The districts with the highest rate of shooting incidents have the highest cumulative percentage of highest education level being lower than "some college". As the rate of poverty increased, the rate of shooting incidents appears to increase. As the rate of ballots cast per eligible voter decreases, the rate of shooting incidents increases. This demonstrates that districts with more crime likely report much lower voter engagement rates.

5) Shooting incidents per capita by party affiliation: As shown in Fig.2, it appears that in discricts which are safer than

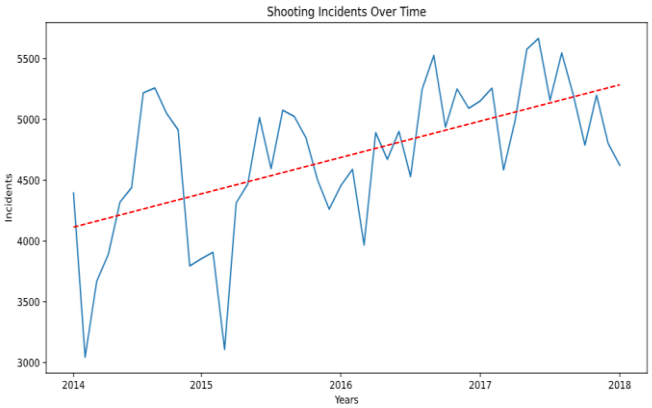


Figure 1: Shooting Incidents Over Time

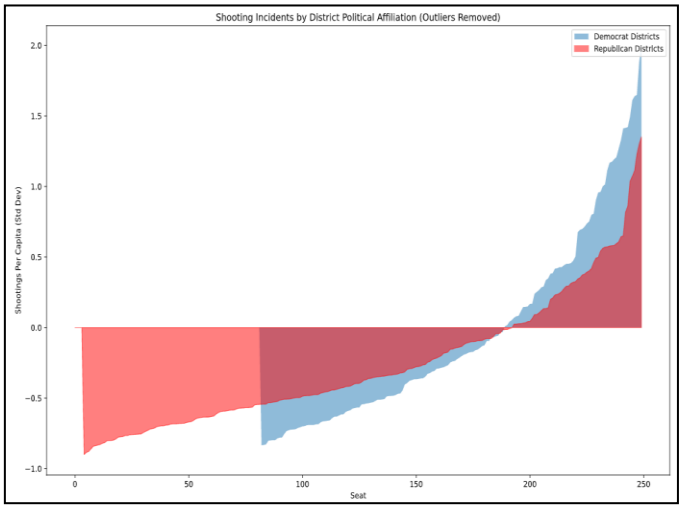


Figure 2: Rates of Shooting Incidents by Political Party Affiliation

average seem to be equally as safe regardless of which political party is affiliated with the district's seat. For districts that are more dangerous than average, it appears that districts where the seat is Democrat affiliated report much higher rates of shooting incidents than Republican districts. This data may be skewed based on the type of setting of each district causing a mismatch in comparisons. Urban environments are typically associated with higher crime and Democratic voters, yet this analysis does not take into consideration the environment of each district when making a comparison.

B. *Prediction Analysis*

Neither linear regression nor k-nearest neighbor were overly successfully in accurately predicting the rates of shooting incidents or safety level of each district, but they did provide some interesting insights and questions for future research.

1) Linear regression: When trying to predict the standard deviation of the rate of shooting incidents by district, most models scored with percentages in the mid 30s to low 40s. This would indicate that education, poverty level, voting rate, and political affiliation of the district seat are unable to explain a large portion of the variation in shooting incident rates. Overall, this is not entirely surpising as there are many external factors and societal differences that influence crime rates which are not captured in these few demographics. The coefficients for each independent variable were also checked. A bit surprising was the fact that the model showed three different education variables with a positive coefficient (high school, graduate, and bachelor's degree percentages). Based on the exploratory analysis and the education visualization, we would expect these coefficients to be negative. This discrepancy warrants further investigation.

2) K-nearest neighbor: The k-nearest neighbor appeared to be a much more friendly way to evaluate the safety of a district. After removing all outliers that were greater than three standard devations from the mean and by discretizing all continuous variables into five bin each, we were able to capture more variation across the different features. Using n_neighbor=15 and a weighting method of weights="distance", the k-nearest

neighbor model was able to predict the district safety classification with a percentage score in the mid 50s – low 60s. By investigating the confusion matrix shown in Fig. 3, we can see that most of the districts had a true label of 0 or 1 which are ratings of safer than the mean. Because our test data included so many districts that were safer than the mean, it is difficult to evalute how successful our model is at predicting dangerous districts. Future work on this topic should include some experiementation with balancing the data to better interpret the model's strengths and weaknesses. Doing so may better highlight the predictive nature of education level, poverty levels, and voting rates with relation to the rate of shooting incidents.

IV. CONCLUSION

The results from the exploratory analysis likely reinforce beliefs that many people have already assumed. That is the fact that lower levels of education and higher poverty rates correlate with higher rates of shooting incidents. Although the linear regression model outputs a confusing result stating that some education levels positively correlate with rates of shooting incidents, our visual analysis shows the opposite. An understanding of societal trends supported by the visual analysis would favor validating that education levels are a contributing factor to shooting incident rates. Both of our visual analysis and linear regression models support the claim that higher poverty rates positively correlate with higher shooting incident rates. It was also found in the visual analysis and regression analysis that lower rates of voter engagement also had a positive correlation to shooting incident rates. This is also supported by an understanding of societal norms that higher crime areas are less likely to have high rates of social and political engagements than safer areas.

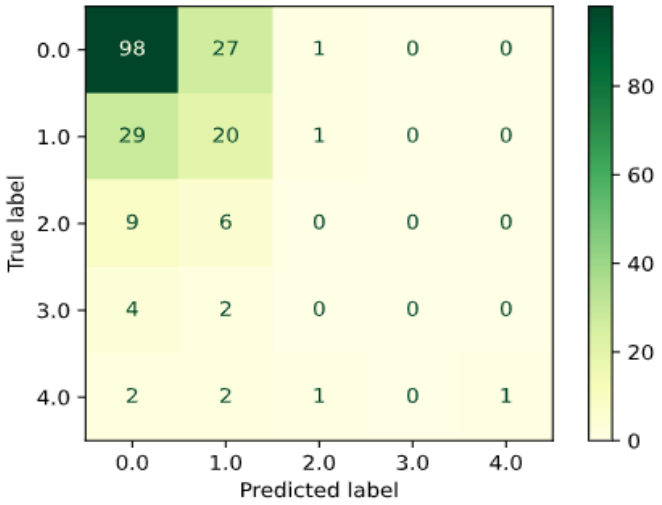


Figure 3: KNN Confusion Matrix (District Safety Bin)

Our exploratory analysis also reinforced that race, gender, and ages do not individually correlate with rates of shooting incidents. Gender and age percentages are very consistent from district to district. Although race percentages do vary from one district to another, there was no visual correlation between changes in the race percentage and shooting incident rate.

Even though it is a disputed claim in politics, it does appear that the political party affiliation of the congressional seat for each district does correlate with the rate of shooting incidents. When sorted from safest to most dangerous, it appears that Democrat and Republican districts are equally as safe. For districts that are more dangerous than average, Democrat districts appear to be more dangerous than Republic districts. This is further supported by our linear regression model which shows Republican districts with a negative coefficient, whereas Democrat districts had a positive coefficient. It is important to understand that this simple analysis does not take into consideration the environment of the district (urban, suburban, rural) and makes no attempts to compare these side by side.

Predictions in using both linear regression and k-nearest neighbor were not very accurate. After evaluating the coefficients and confusion matrices, there are two primary conclusions. The independent variables of education level, poverty rate, voter engagement, and political affiliation are not substantial enough to account for the many societal issues that may impact the rates of violence in a district. The dataset also includes many more safe districts than dangerous districts. This is good for society, but it likely contributed to weakness in the predictive capabilities of the model.

Additional technical details and the python notebooks used to perform this analysis can be found on Github under username "alex-oneill" (www.github.com/alex-oneill).

REFERENCES

[1] Gun Violence Archive, "Gun Violence Data". James Ko. https://www.kaggle.com/jameslko/gun-violence-data

[2] Every Politician, Everypolitician.org. https://everypolitician.org/united-states-of-America/house/download.html

[3] United States Census Bureau, "Congressional Voting Tables". https://www.census.gov/data/tables/time-series/demo-voting-and-registration/congressional-voting-tables.html

Alex ONeill – ao24521p@pace.edu
CS658 – Analytics Computing (Fall 2020)
Prof. Barabasi

EXPLORING SHOOTING INCIDENTS

A look at census data, political affiliation, and shooting incidents in the United States between 2014-2017

INTRODUCTION & DATA SOURCES

CONCEPT & HYPOTHESIS

The original hypothesis was that there would be a relationship between gun violence incidents and congressional seats flipping from one party to another.

--

Unfortunately, the dataset was unable to support an in-depth investigation into this hypothesis.

Two of the five years in the shooting incident dataset were incomplete. Congressional seats do not flip hands very often, therefore data for shooting incidents would be required for many more years to do a proper analysis.

--

The data did support an exploratory analysis into shooting incident trends, relationships with demographic data, and political party ownership at the time of the shooting incident. Exploring this data will can better equip policy makers and help citizens identify whether they will be living in a safe area.

DATA SOURCES

Gun Violence Data: 260,000+ shooting reports from 2013-2018

https://www.kaggle.com/jameslko/gun-violence-data

--

Congressional Representative Data: Congressional Representative information sheet by term.

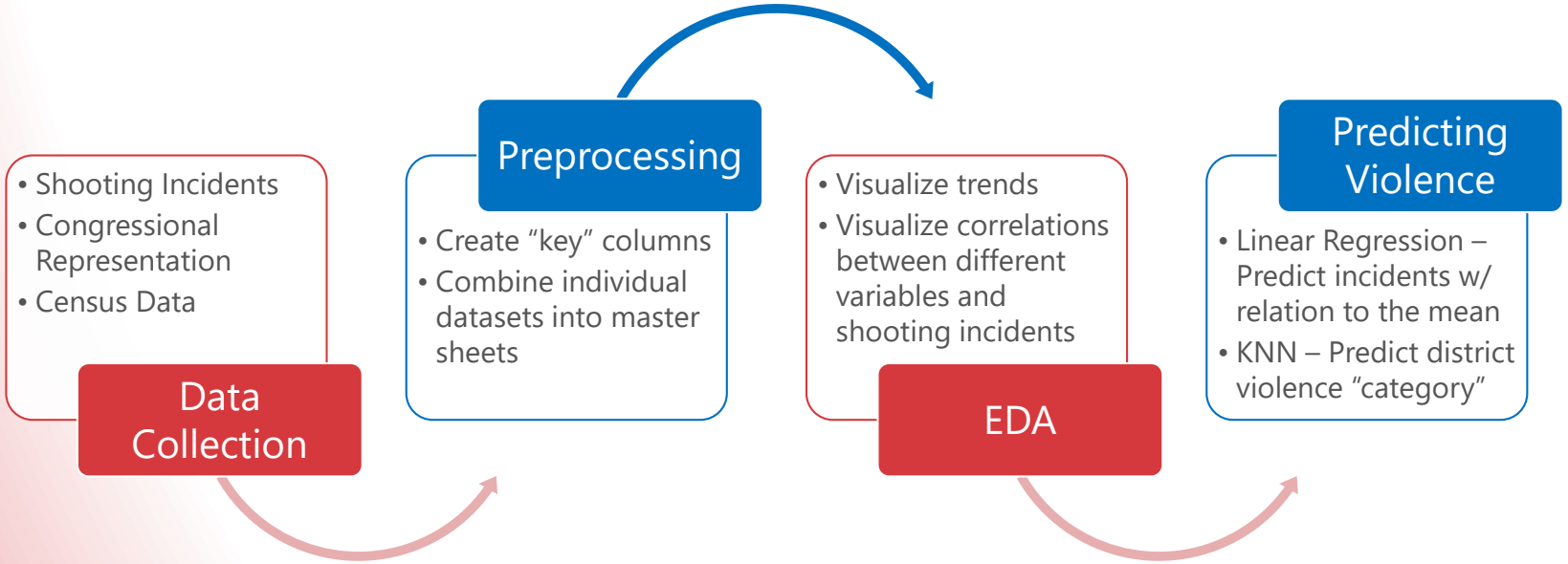
https://everypolitician.org/united-states-of-America/house/download.html

--

Census Data: Demographic information based on voter data from the 2018 midterm election.

https://www.census.gov/data/tables/time-series/demo-voting-and-registration/congressional-voting-tables.html

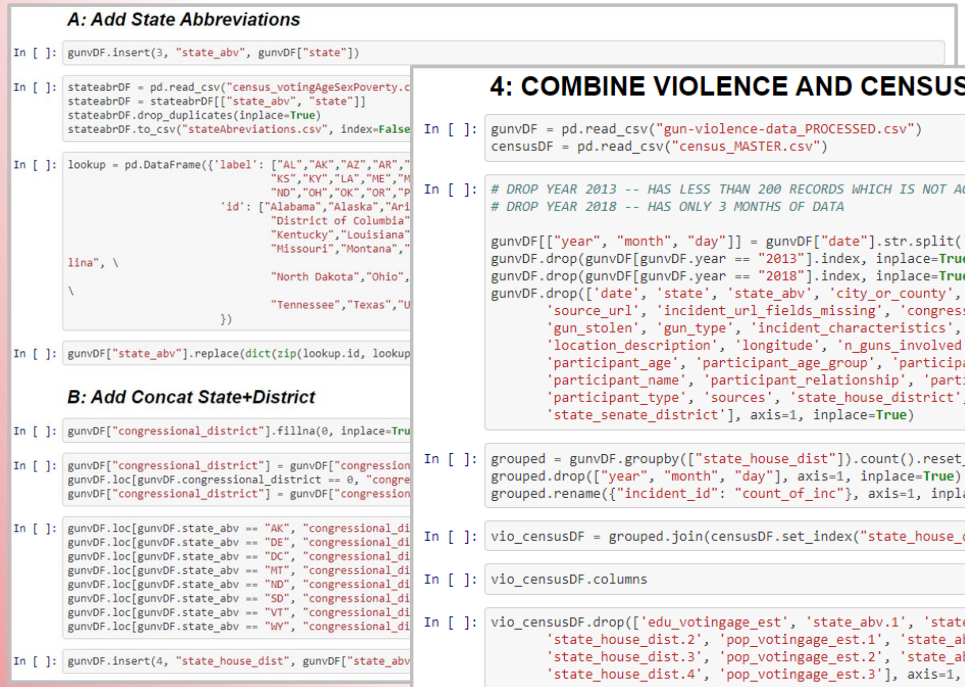
METHOD


• Shooting Incidents
• Congressional Representation
• Census Data
Data Collection
Preprocessing
• Create "key" columns
• Combine individual datasets into master sheets
• Visualize trends
• Visualize correlations between different variables and shooting incidents
EDA
Predicting Violence
• Linear Regression – Predict incidents w/ relation to the mean
• KNN – Predict district violence "category"

PRE-PROCESSING

There was a considerable amount of pre-processing required to do an exploratory analysis.

- Add state abbreviations and combine with congressional district
- Combine relevant stats from multiple census datasets into a master sheet
- Combine multiple congressional representative datasets into a master sheet
- Combine census data with incident report data by congressional district



EXPLORATORY ANALYSIS (CONGRESSIONAL SEATS)

Congressional Seats Appear to Remain Stable Except After Major Events

- During/After the 2008 stock market crash, Democrats pick up seats
- After 9/11, Republicans seats remained stable or increased

1: CONGRESSIONAL REPRESENTATION

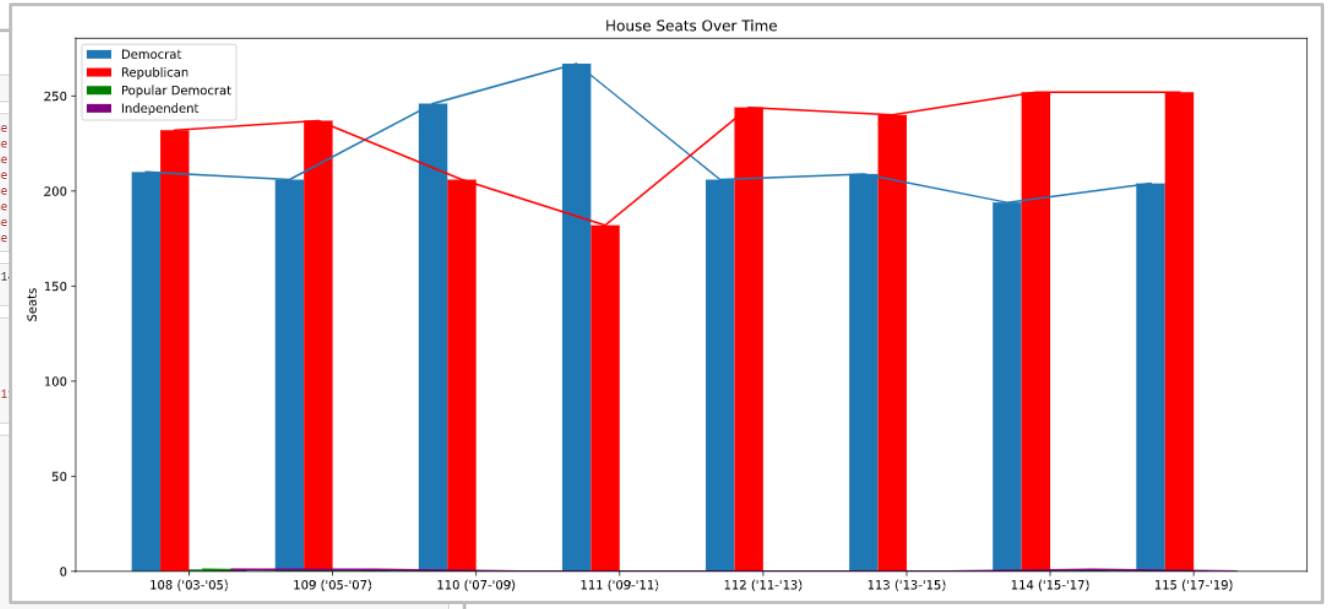
In [2]:
```
repDF = pd.read_csv("representatives_MASTER.csv")
```

In [3]:
```
term108 = repDF.loc[(repDF["term"]) == 108].groupby(["group"]).count()["name
term109 = repDF.loc[(repDF["term"]) == 109].groupby(["group"]).count()["name
term110 = repDF.loc[(repDF["term"]) == 110].groupby(["group"]).count()["name
term111 = repDF.loc[(repDF["term"]) == 111].groupby(["group"]).count()["name
term112 = repDF.loc[(repDF["term"]) == 112].groupby(["group"]).count()["name
term113 = repDF.loc[(repDF["term"]) == 113].groupby(["group"]).count()["name
term114 = repDF.loc[(repDF["term"]) == 114].groupby(["group"]).count()["name
term115 = repDF.loc[(repDF["term"]) == 115].groupby(["group"]).count()["name
```

In [4]:
```
dictionaries = [term108, term109, term110, term111, term112, term113, term11
groups = ['Democrat', 'Republican', 'Popular Democrat', 'Independent']
```

In [5]:
```
demSeats = [210, 206, 246, 267, 206, 209, 194, 204]
repSeats = [232, 237, 206, 182, 244, 240, 252, 252]
popDemSeats = [1, 0, 0, 0, 0, 0, 0, 0]
indepSeats = [1, 1, 0, 0, 0, 0, 1, 0]
congressionalTerms = ["108 ('03-'05)", "109 ('05-'07)", "110 ('07-'09)", "11
7)", "115 ('17-'19)"]
```

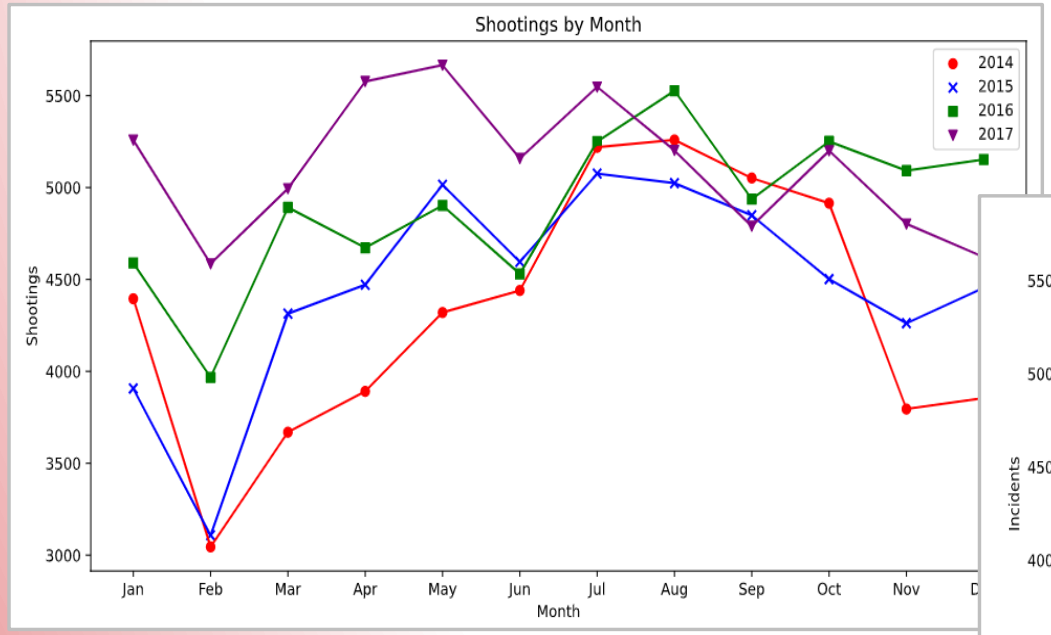
In [6]:
```
plt.figure(figsize=(18, 8))
width = 0.2
firstbar = np.arange(len(congressionalTerms))
secondbar = [i + width for i in firstbar]
thirdbar = [i + width for i in secondbar]
fourthbar = [i + width for i in thirdbar]

plt.bar(firstbar, demSeats, width, label="Democrat")
plt.plot(firstbar, demSeats)

plt.bar(secondbar, repSeats, width, label="Republican", color="Red")
plt.plot(secondbar, repSeats, color="Red")

plt.bar(thirdbar, popDemSeats, width, label="Popular Democrat", color="Green")
plt.plot(thirdbar, popDemSeats, color="Green")

plt.bar(fourthbar, indepSeats, width, label="Independent", color="Purple")
plt.plot(fourthbar, indepSeats, color="Purple")

plt.xticks((firstbar + (width*1.5)), congressionalTerms)
plt.title("House Seats Over Time")
plt.ylabel("Seats")
plt.legend()
```

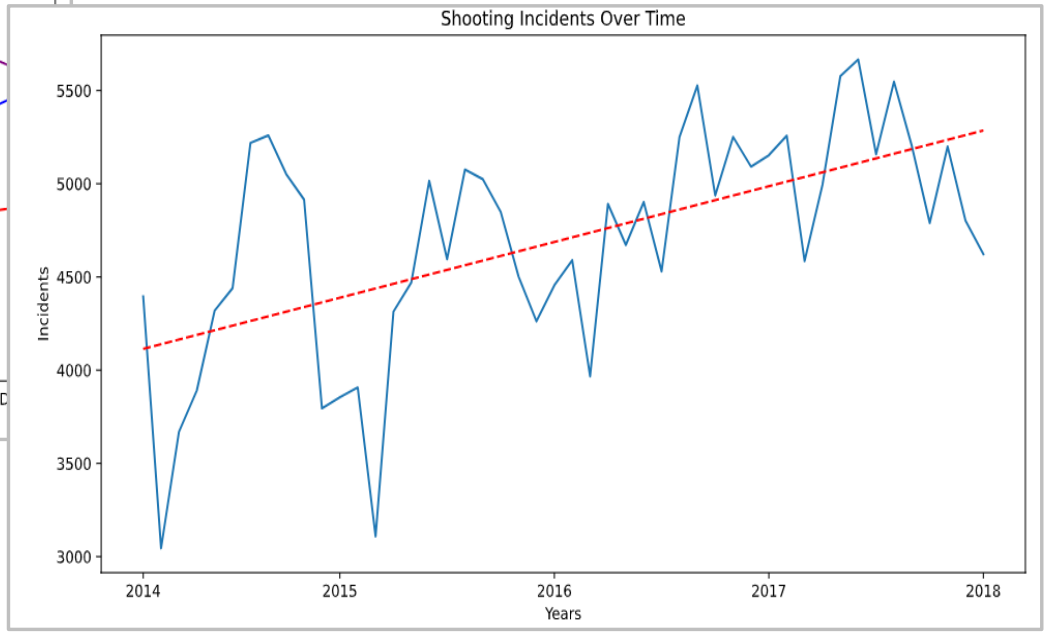


EXPLORATORY ANALYSIS (SHOOTING INCIDENTS)

Shooting Incidents Continued to Increase

- Shooting incidents occur less often during the winter
- Shooting incidents are increasing over time





EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

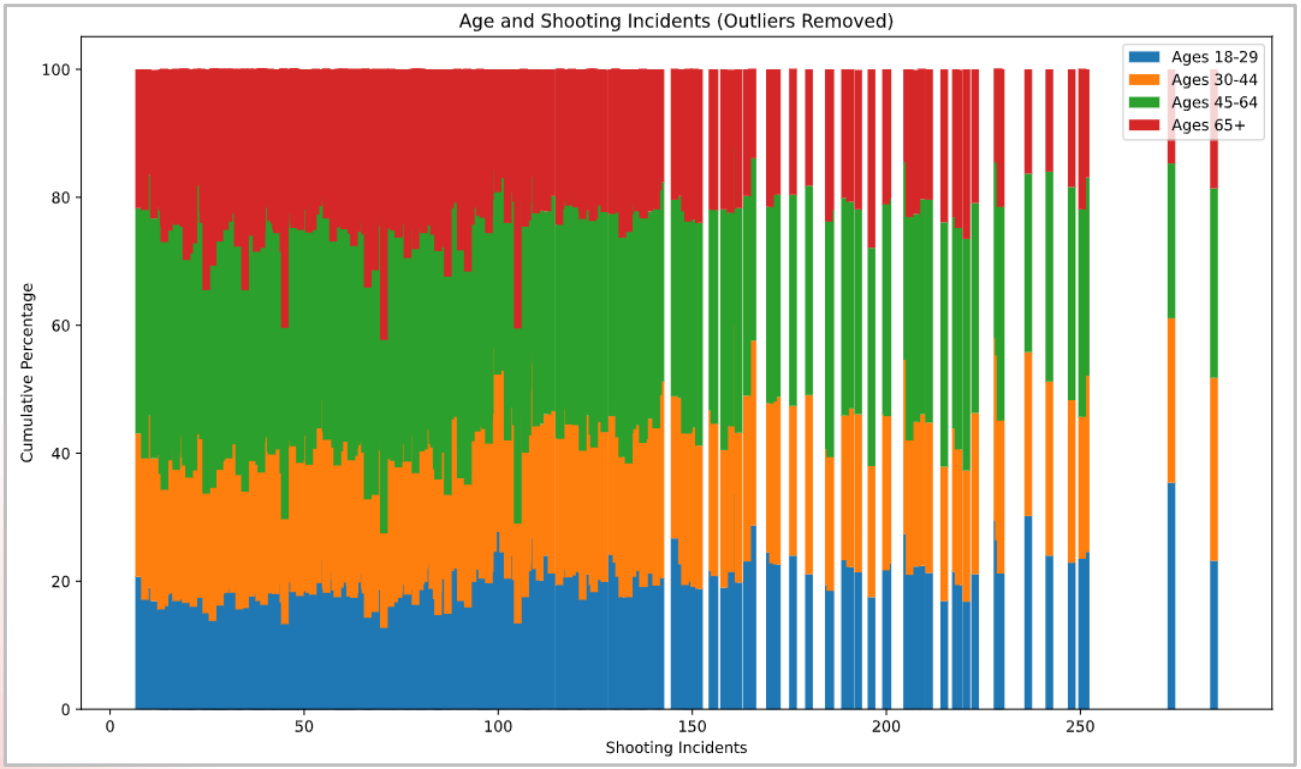
Incidents per Capita (100k Residents of Voting Age)

- No visual correlation between shooting incidents and race, gender, or age percentages in each district

EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

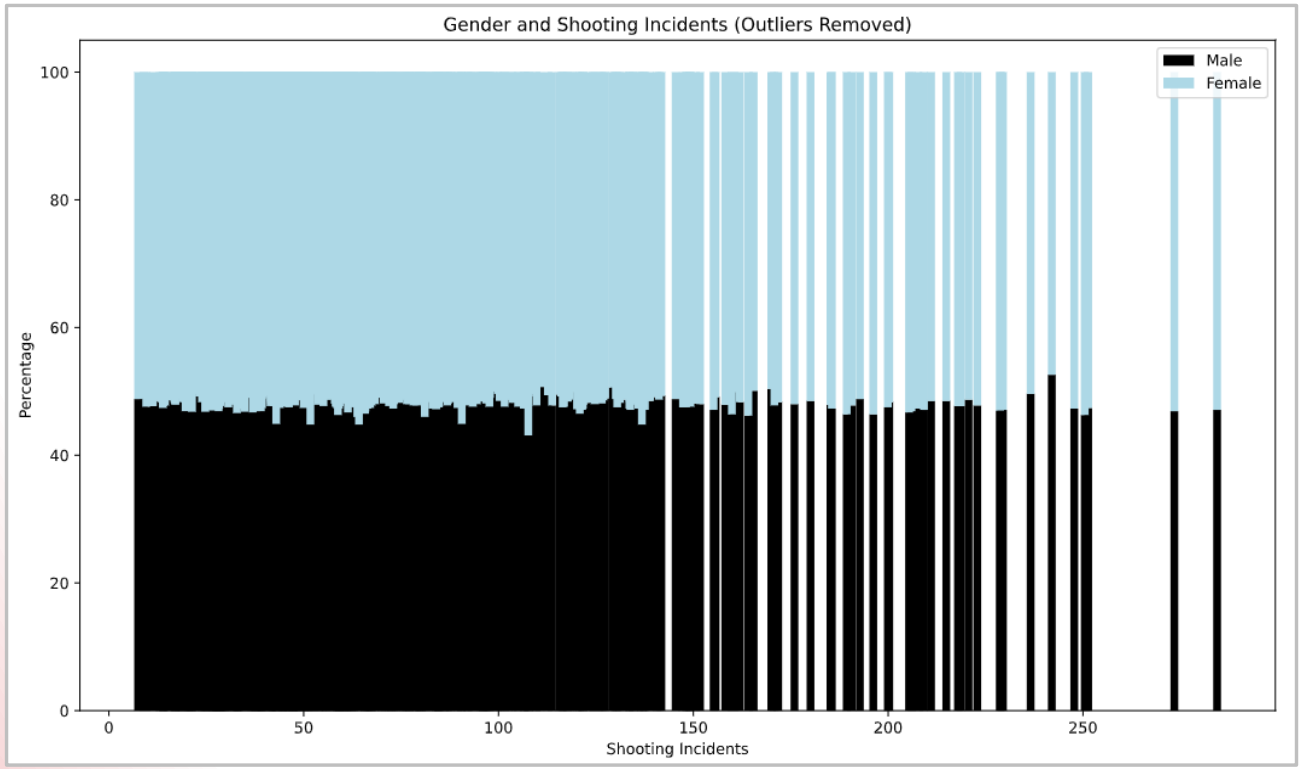
- No visual correlation between shooting incidents and race, gender, or age percentages in each district



EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

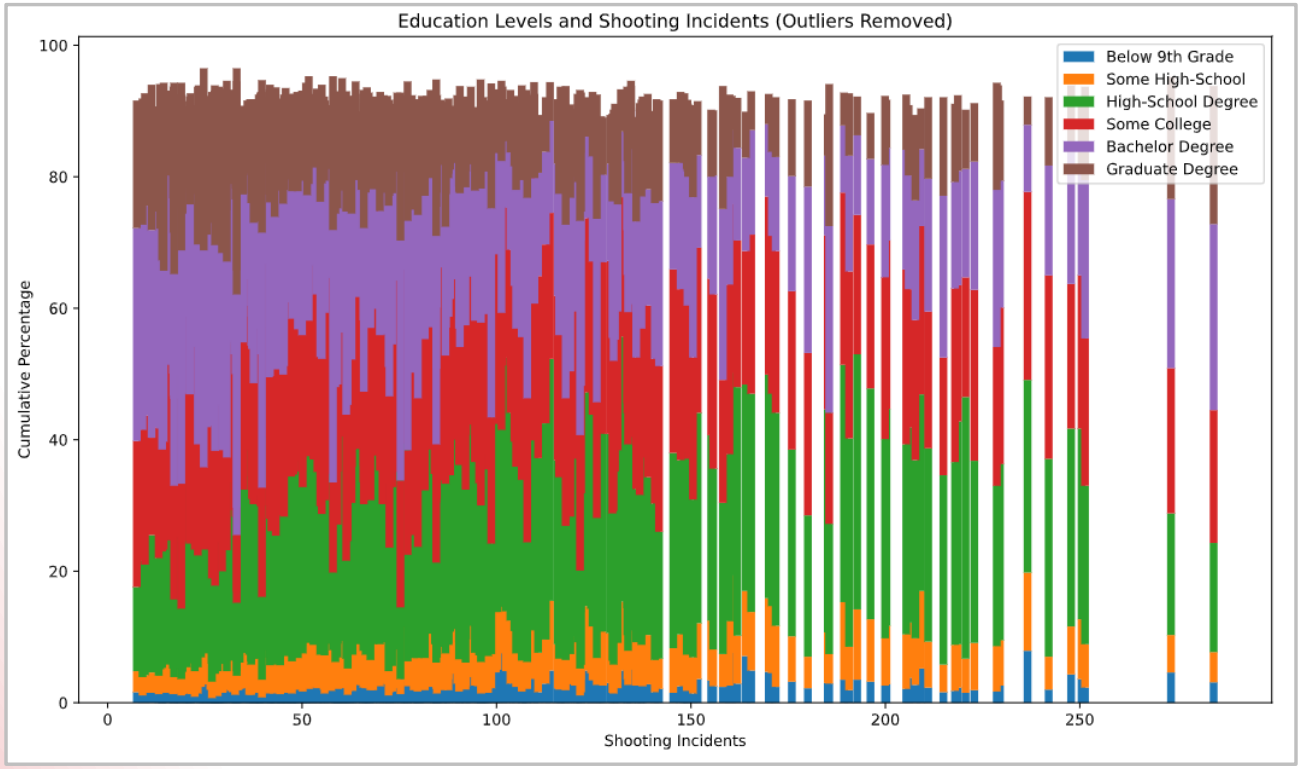
- No visual correlation between shooting incidents and race, gender, or age percentages in each district



EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

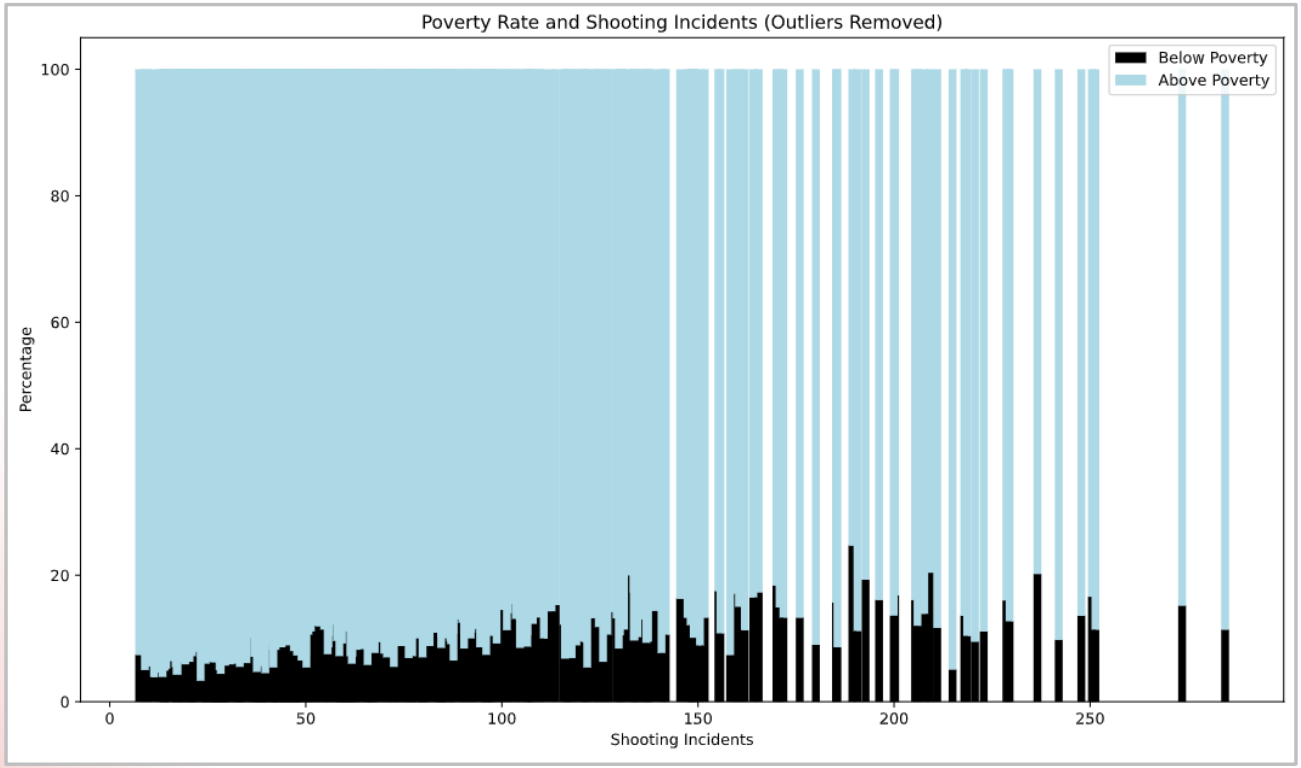
- Visual correlation between shooting incidents and education level, poverty percentage, and voting rate in each district is noticed



EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

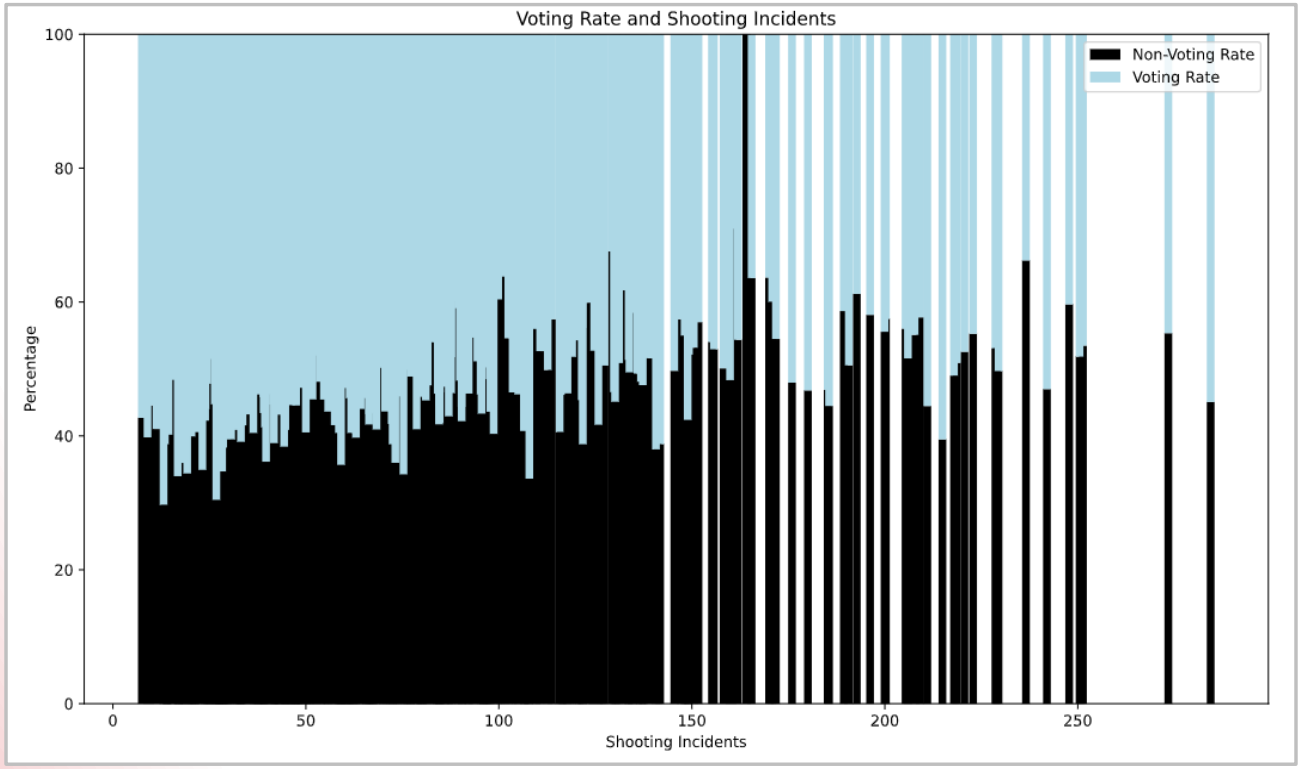
- Visual correlation between shooting incidents and education level, poverty percentage, and voting rate in each district is noticed



EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

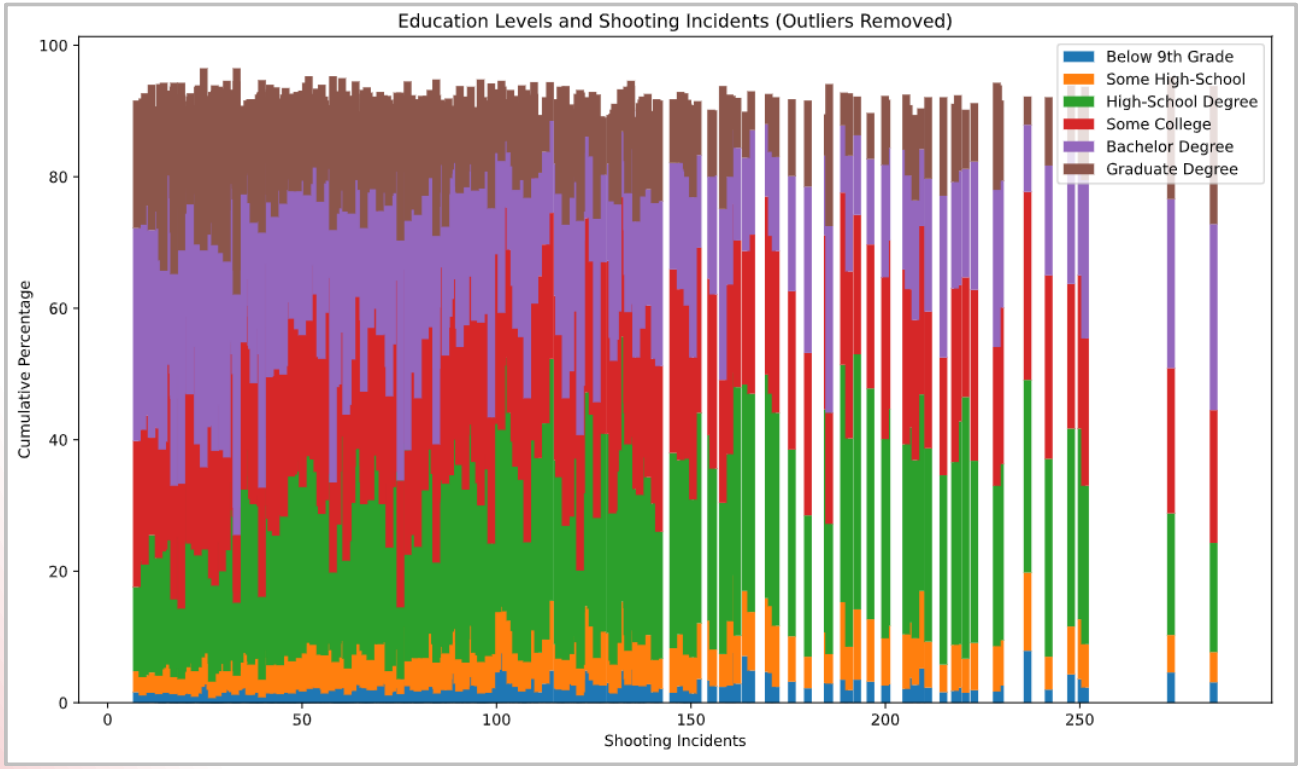
- Visual correlation between shooting incidents and education level, poverty percentage, and voting rate in each district is noticed



EXPLORATORY ANALYSIS (DEMOGRAPHICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

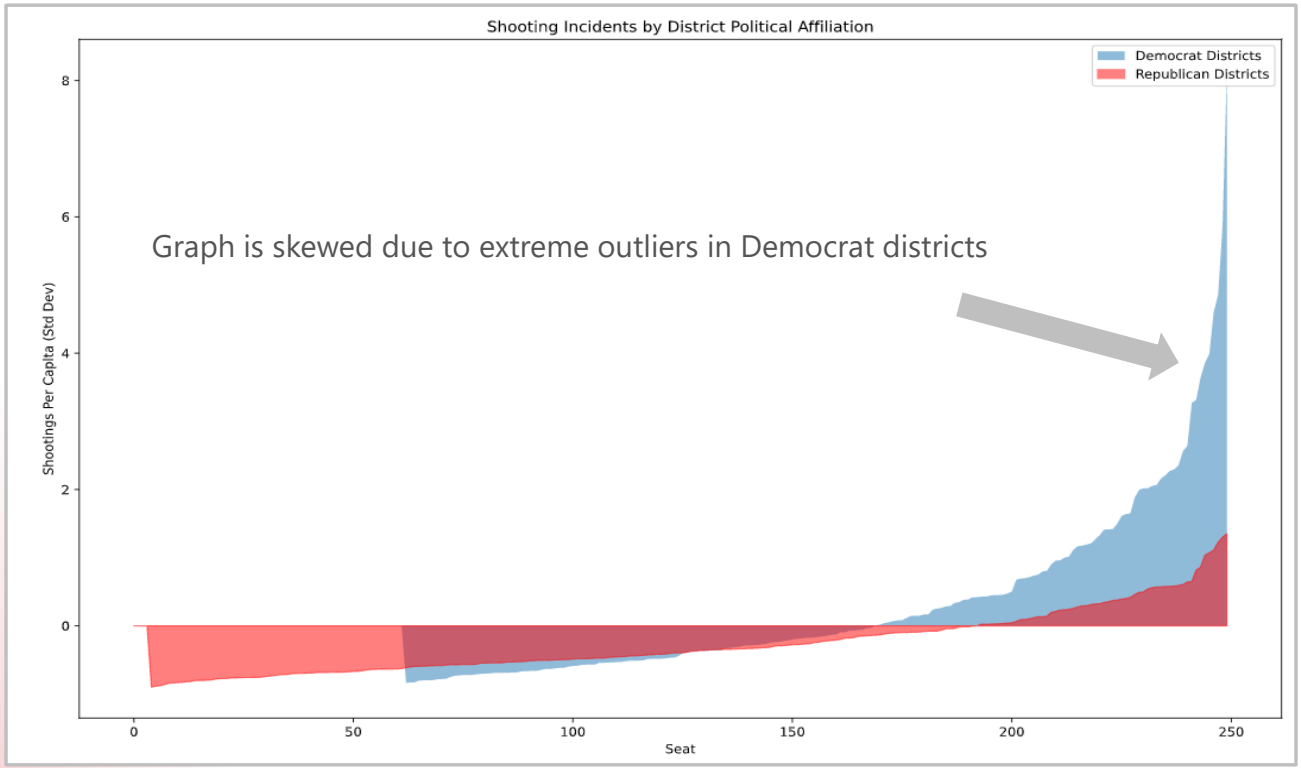
- Visual correlation between shooting incidents and education level, poverty percentage, and voting rate in each district is noticed



EXPLORATORY ANALYSIS (POLITICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

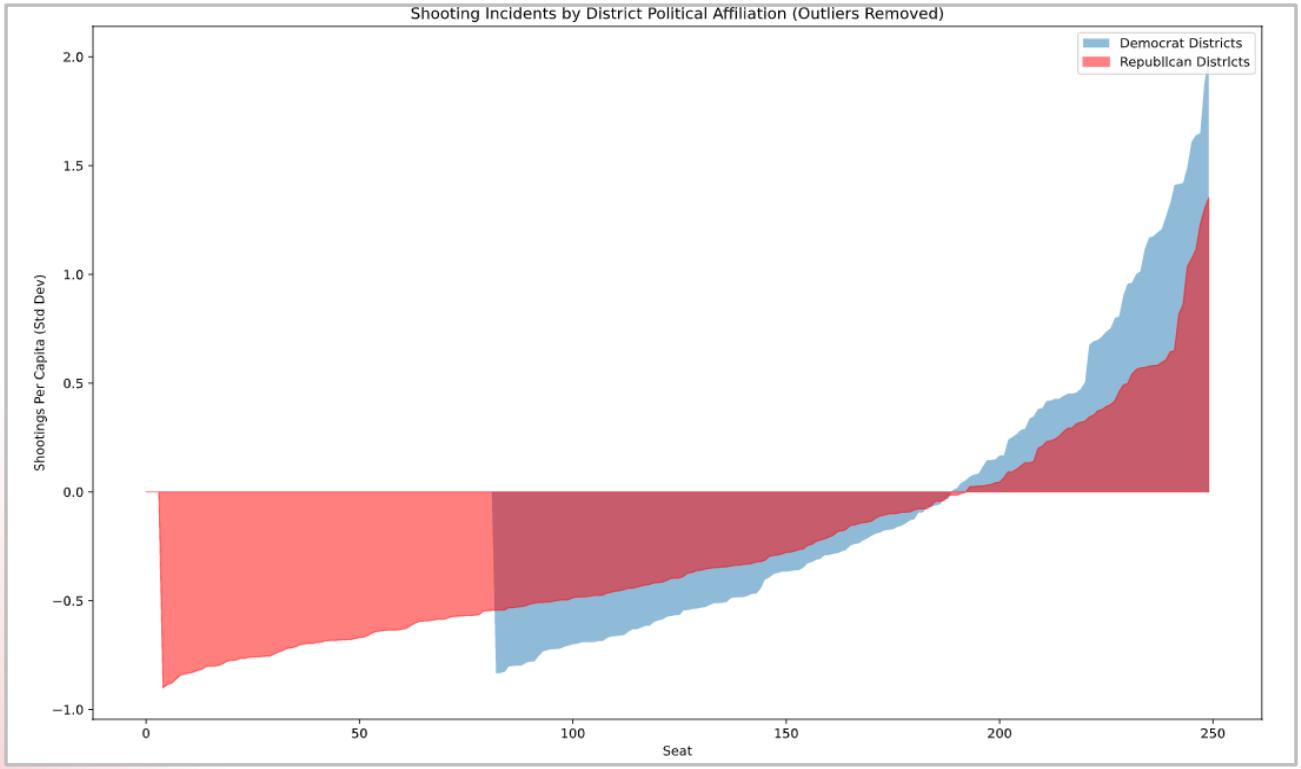
- When the district is safer than average, Democrat and Republican districts appear equally as safe.
- When the district is more dangerous than average, Democrat districts appear more dangerous than Republican districts



EXPLORATORY ANALYSIS (POLITICS & INCIDENTS)

Incidents per Capita (100k Residents of Voting Age)

- When the district is safer than average, Democrat and Republican districts appear equally as safe.
- When the district is more dangerous than average, Democrat districts appear more dangerous than Republican districts



PREDICTING INCIDENTS (LINEAR REGRESSION)

Using Linear Regression, We Can Attempt To Predict How Close To The Mean Of Incidents A District Will Be

- All values set to the standard deviation; political party one-hot encoded
- Outliers > 2 std dev removed
- Train/test split of 50% to allow for more test samples
- The fit is not great; most independent variables had a small coefficient; some education variables had surprisingly positive coefficients

USING STANDARD DEVIATION TO NORMALIZE VALUES

In [2]:
```
reps = pd.read_csv("representatives_MASTER.csv")
vio = pd.read_csv("violence_census_MASTER.csv")
```

In [3]:
```
vio["capita"] = vio["count_of_inc"] / (vio["pop_votingage_est"] / 100000)
vio["capitaSTD"] = (vio["capita"] - vio["capita"].mean()) / vio["capita"].std()
```

In [4]:
```
# REMOVE ALL ENTRIES FOR TERMS OTHER THAN 114 (2015-2017)
reps.drop(reps[reps["term"] != 114].index, inplace=True)
```

In [5]:
```
combo = reps.merge(vio, on="area_id")

# REMOVE OUTLIERS > 2 STANDARD DEVIATION
comboNO = combo.drop(combo[combo["capitaSTD"] > 2].index)
```

In [6]:
```
columns = ['edu_ltgrade9_perc', 'edu_9to12_perc', 'edu_hs_perc', 'edu_somecollege_perc', 'edu_bachelor_pe
    'edu_graduate_perc', 'edu_hsormore_perc', 'edu_bachormore_perc', 'poverty_perc', 'pop_votingrate_perc

for item in columns:
    comboNO["std_"+item] = (comboNO[item] - comboNO[item].mean()) / comboNO[item].std()

# REMOVE OUTLIERS > 2 STANDARD DEVIATION
for outlier in columns:
    comboNO.drop(comboNO[comboNO["std_"+outlier] > 2].index, inplace=True )

comboNO = comboNO.join(pd.get_dummies(comboNO["group"]))
```

In [7]:
```
linreg = LinearRegression()
```

In [8]:
```
indepArray = comboNO[['std_edu_ltgrade9_perc', 'std_edu_9to12_perc', 'std_edu_hs_perc', 'std_edu_somecollege_perc', 'std_edu_bachelor_per
c', 'std_edu_graduate_perc',
    'std_edu_hsormore_perc', 'std_edu_bachormore_perc', 'std_poverty_perc', 'std_pop_votingrate_perc', 'Democrat', 'Republican']]
depArray = comboNO["capitaSTD"]
```

In [9]:
```
X_train, X_test, y_train, y_test = train_test_split(indepArray, depArray, test_size=0.5)
linreg.fit(X_train, y_train)
linreg.score(X_test, y_test)
```

Out[9]:
```
0.3403010117450609
```

In [10]:
```
coeff_df = pd.DataFrame(linreg.coef_, indepArray.columns, columns=["Coefficient"])
coeff_df.Coefficient.sort_values()
```

Out[10]:
```
std_edu_bachormore_perc    -1.577175
std_edu_hsormore_perc      -1.050480
std_edu_ltgrade9_perc      -0.533149
std_edu_9to12_perc         -0.466476
Republican                 -0.169488
std_pop_votingrate_perc     0.016603
std_edu_somecollege_perc    0.116402
Democrat                    0.169488
std_poverty_perc            0.353976
std_edu_hs_perc             0.369488
std_edu_graduate_perc       1.020335
std_edu_bachelor_perc       1.251499
Name: Coefficient, dtype: float64
```

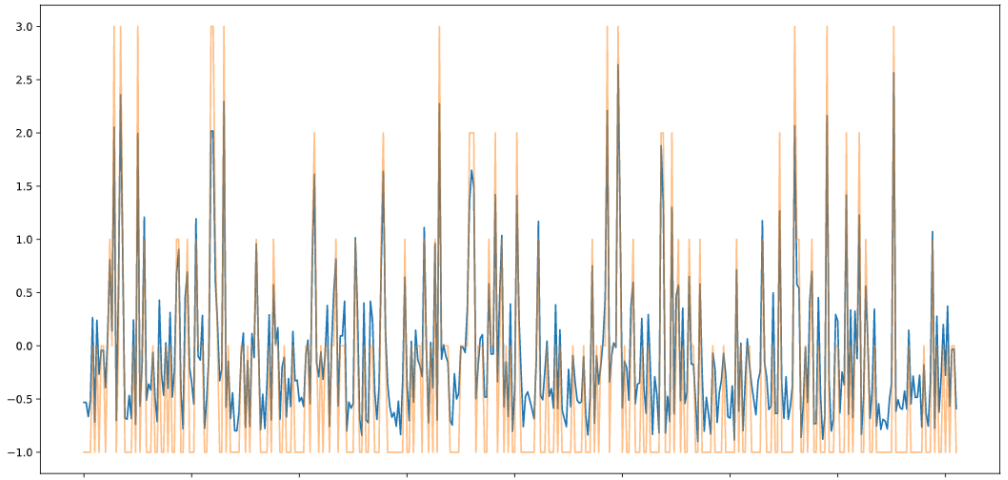
PREDICTING DANGEROUS DISTRICTS (KNN)

Using KNN, We Can Try To Predict Whether A District Will Be "Most Safe", "Safe", "Neutral", "Dangerous", & "Most Dangerous"

- All dependent and independent variables were normalized by std dev; outliers > 3 std deviation removed
- All variables discretized into 5 groups (0-5); 2 = closest to the mean; political party one-hot encoded

```
In [11]: reps = pd.read_csv("representatives_MASTER.csv")
         vio = pd.read_csv("violence_census_MASTER.csv")

In [12]: vio["capita"] = vio["count_of_inc"] / (vio["pop_votingage_est"] / 100000)
         vio["capitaSTD"] = (vio["capita"] - vio["capita"].mean()) / vio["capita"].std()

In [13]: # REMOVE ALL ENTRIES FOR TERMS OTHER THAN 114 (2015-2017)
         reps.drop(reps[reps["term"] != 114].index, inplace=True)

In [14]: combo = reps.merge(vio, on="area_id")

         # REMOVE OUTLIERS > 3 STANDARD DEVIATION
         comboNO = combo.drop(combo[combo["capitaSTD"] > 3].index)

In [15]: discret = KBinsDiscretizer(n_bins=5, encode="ordinal", strategy="uniform")

In [16]: columns = ['edu_ltgrade9_perc', 'edu_9to12_perc', 'edu_hs_perc', 'edu_somecollege_pe
             'edu_hsormore_perc','edu_bachormore_perc', 'poverty_perc', 'pop_votingrate_perc'

         for item in columns:
             comboNO["std_"+item] = (comboNO[item] - comboNO[item].mean()) / comboNO[item].st

         comboNO = comboNO.join(pd.get_dummies(comboNO["group"]))

         # REMOVE OUTLIERS > 3 STANDARD DEVIATION
         for outlier in columns:
             comboNO.drop(comboNO[comboNO["std_"+outlier] > 3].index, inplace=True )

In [17]: columns = ['edu_ltgrade9_perc', 'edu_9to12_perc', 'edu_hs_perc', 'edu_somecollege_perc',
             'edu_bachelor_perc', 'edu_graduate_perc', 'edu_hsormore_perc',
             'edu_bachormore_perc', 'poverty_perc', 'pop_votingrate_perc', "capitaSTD"]

         for bin_cat in columns:
             comboNO["bin_"+bin_cat] = discret.fit_transform(comboNO[[bin_cat]])
```

```
In [18]: plt.figure(figsize=(16,8))
         plt.plot(np.arange(comboNO["bin_capitaSTD"].count()), comboNO["capitaSTD"])
         plt.plot(np.arange(comboNO["bin_capitaSTD"].count()), comboNO["bin_capitaSTD"]-1, alpha=0.5)

Out[18]: [<matplotlib.lines.Line2D at 0x20f2ff15af0>]
```



Visual verification that the discretized bins align properly to the std dev of incidents/capita

PREDICTING DANGEROUS DISTRICTS (KNN)

Using KNN, We Can Try To Predict Whether A District Will Be "Most Safe", "Safe", "Neutral", "Dangerous", & "Most Dangerous"

- KNN set with n_neighbors=15 & weights="distance"; train/test split set at 50% to allow for greater variation in the test set
- Scoring results between mid 50s-60s; disproportionate amount of safer districts is seen

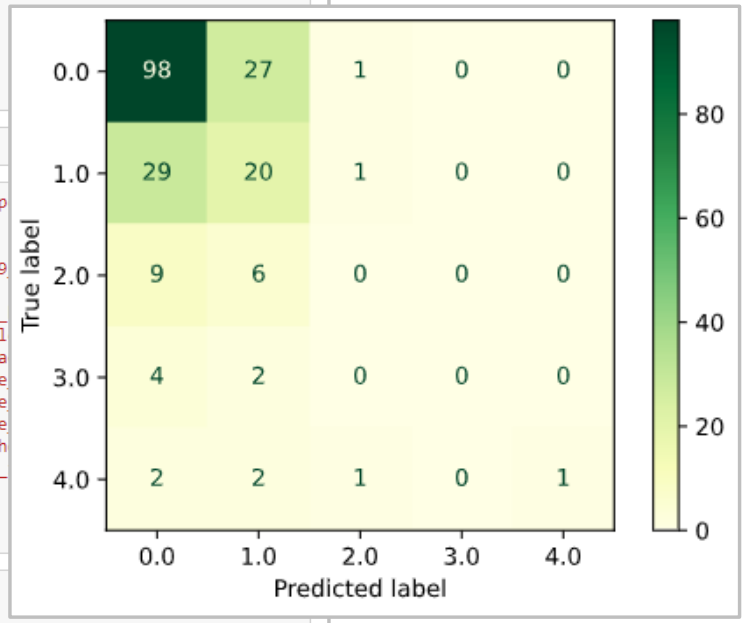
In [19]:
```
binned_columns = ['bin_edu_ltgrade9_perc', 'bin_edu_9to12_perc', 'bin_edu_hs_perc', 'bin_edu_somecollege_perc',
    'bin_edu_bachelor_perc', 'bin_edu_graduate_perc', 'bin_edu_hsormore_perc', 'bin_edu_bachormore_perc',
    'bin_poverty_perc', 'bin_pop_votingrate_perc']

for column in binned_columns:
    comboNO = comboNO.join(pd.get_dummies(comboNO[column], prefix=column))
```

In [20]:
```
knn = KNeighborsClassifier(n_neighbors=15, weights="distance")
```

In [21]:
```
indepArray = comboNO[['Democrat', 'Republican', 'bin_edu_ltgrade9_perc', 'bin_edu_9to12_perc', 'bin_edu_hs_p
    'bin_edu_somecollege_perc', 'bin_edu_bachelor_perc', 'bin_edu_graduate_perc', 'bin_edu_hsormore_perc',
    'bin_edu_bachormore_perc', 'bin_poverty_perc', 'bin_pop_votingrate_perc', 'bin_edu_ltgrade9_perc_0.0',
    'bin_edu_ltgrade9_perc_1.0', 'bin_edu_ltgrade9_perc_2.0', 'bin_edu_ltgrade9_perc_3.0', 'bin_edu_ltgrade9
    'bin_edu_9to12_perc_0.0', 'bin_edu_9to12_perc_1.0', 'bin_edu_9to12_perc_2.0', 'bin_edu_9to12_perc_3.0',
    'bin_edu_9to12_perc_4.0', 'bin_edu_hs_perc_0.0', 'bin_edu_hs_perc_1.0', 'bin_edu_hs_perc_2.0', 'bin_edu_
    'bin_edu_hs_perc_4.0', 'bin_edu_somecollege_perc_0.0', 'bin_edu_somecollege_perc_1.0', 'bin_edu_somecoll
    'bin_edu_somecollege_perc_3.0', 'bin_edu_somecollege_perc_4.0', 'bin_edu_bachelor_perc_0.0', 'bin_edu_ba
    'bin_edu_bachelor_perc_2.0', 'bin_edu_bachelor_perc_3.0', 'bin_edu_bachelor_perc_4.0', 'bin_edu_graduate
    'bin_edu_graduate_perc_1.0', 'bin_edu_graduate_perc_2.0', 'bin_edu_graduate_perc_3.0', 'bin_edu_graduate
    'bin_edu_hsormore_perc_0.0', 'bin_edu_hsormore_perc_1.0', 'bin_edu_hsormore_perc_2.0', 'bin_edu_hsormore
    'bin_edu_hsormore_perc_4.0', 'bin_edu_bachormore_perc_0.0', 'bin_edu_bachormore_perc_1.0', 'bin_edu_bach
    'bin_edu_bachormore_perc_3.0', 'bin_edu_bachormore_perc_4.0', 'bin_poverty_perc_0.0', 'bin_poverty_perc_
    'bin_poverty_perc_2.0', 'bin_poverty_perc_3.0', 'bin_poverty_perc_4.0', 'bin_pop_votingrate_perc_0.0',
    'bin_pop_votingrate_perc_2.0', 'bin_pop_votingrate_perc_3.0', 'bin_pop_votingrate_perc_4.0']]
depArray = comboNO["bin_capitaSTD"]
```

In [22]:
```
X_train, X_test, y_train, y_test = train_test_split(indepArray, depArray, test_size=0.5)
knn.fit(X_train, y_train)
knn.score(X_test, y_test)
```

Out[22]: 0.5862068965517241



CONCLUSION & FUTURE WORK

CONCLUSION

- We were unable to see whether changes in congressional seats impacted the rate of violence
- Age, Race, and Gender makeup of the district DO NOT appear to correlate with shooting incident rates
- Voting rates, education levels, and poverty rates in a district DO appear to correlate with shooting incident rates
- Dangerous Democrat districts appear to be more dangerous than dangerous Republican districts
- No causality was able to be determined
- External attributes other than the variables used in this analysis likely contribute to shooting incident rates

FUTURE WORK

- Expand dataset to include additional years
- Maximize the code to be more concise and pythonic
- Apply boosting or other synthetic data methods to increase the size of the dataset used for predictions
- Modify prediction algorithms to be more accurate
- Expand dataset to include additional variables that may independently (or in combination with current variables) increase the predictability of shooting incident rates. Density, income, school quality, and family dynamics should also be included.
- Compare Democrat and Republican districts to each other with more scrutiny based on locality type (urban, suburban, rural)

Notebooks and processed data sheets are viewable at (excluding the shooting incident data):
https://github.com/alex-oneill/