Census tract level of gun violence and the demographic in Boston, Massachusetts

Boheng Lin

December 14, 2019

Abstract

Past studies have suggested correlations between demographic variables and firearm violence on national, state and city level. This paper takes on certain demographic variables and explore any relations with firearm violence on census tract level in Suffolk, Massachusetts. Using K-means, spatial K-means and linear regression analysis, this paper finds that while most of the observations made on national, state or city level are still applicable on census tract level, degree of firearm violence to some extent is also affected by characteristics of nearby tracts. This suggests that attempts to curb or deter firearm violence in a given tract level should go beyond studying the demographic profile of that tract, but also nearby tracts.

Section 1: Introduction

Being one of the few countries who make it a constitutional right to bear firearms, the debate on gun control has never been settled among its people and has thus been one of the most divisive issues in American society (Hughes, 2019). On one hand is the right written in constitution to bear guns on personal level, on the other hand is the prevalence of firearm assaults that is high enough to push the average number of mass shootings in 2019 beyond one per day as of September 1, 2019 (Silverstein, 2019) – and this number does not include isolated cases of firearm homicides or firearm suicides.

While this possesses increasing challenges to the local police department in deterring and prevent gun violence, attempts have been made by many to draw links between firearm assaults and other factors such as demographic, social-economic or other information on individual level or residential region level. For example, while gun owner ship is positively associated with homicides and suicides, the degree of such association differs much by one's race. And this is subject to state difference too (Riddell, et al., 2018)

However, many studies, including ones cited in this paper by Jiao and Riddell, et al – take aggregated data on national, state or at least the city level. This aggregation of data may sometimes fail to capture important detail related to the subject, for example, variables of local residency are often not recorded in state data. They might not provide an understanding to the gun violence as comprehensive as many used to think to be. Similarly, it is interesting for this paper find how the time variable, which is often omitted from analysis affect gun violence in the region.

With a timely understanding of firearm assault in Boston on a more detailed level – census tract level – not only would the pattern between firearm violence and each neighborhood become clearer, but also, with the inclusion of time variable in this paper, would the paper facilitate the police departments, city planners or interested groups to conduct further exploratory researches and programs to better deter and prevent firearm assault from occurring.

In an attempt to study the degree of firearm assaults in Boston on a more local level, this paper examines the 911 dataset at census tract level from 2010 through 2018, as well as the demographic changes at census tract level to determine if certain demographic groups are more likely to experience firearm assault.

To explore the possible factors shaping the prevalence of gun crimes of a region in Boston, this paper will use regressions to find patterns or associations between firearm assault and, with statistically and economically significance, variables including – tract income level, tract minority percentage, time (in terms of year) and proportion of local residents.

By finding associations and resemblances between population demographics and the gun-safety in the region, public safety authorities on city levels such as the Police departments will be encouraged to more effectively utilize their resources in preventing, deterring or curbing crimes. As this paper examines the ratio of local residence, it is expected to provide insights on the effect of immigrants – regardless of inter-state migration or international migration. Organizations on federal level or national level such as National Crime Prevention Council or the Department of Public Safety could be encouraged to refine policies that either encourage or discourage inter-state migration.

Section 2: Literature Review

The conclusion that certain racial groups are more likely to be associated with firearm assault (Beard, et al., 2017) was made after looking into data from Philadelphia Police Department with individual victim's demographic information including gender, age and race.

In particular, Black are associated with higher rate of firearm assault and higher rate to sustain firearm assault across all victim residence incomes and both low-income areas and 'hot spot' with high proportions of Black residents are associated with frequent firearm assault events (Beard, et al., 2017).

While this was a study in Philadelphia, it suggests that racial composition of a residential region could be a factor shaping the safety of the region. It provides a justification to include racial composition as a factor of the analysis model to be presented in this paper. It is found that even though being in a high-income area is protective for the population overall, the effect varies statistically significant between White and Black residents. In relation to this paper, this study provides a critical point to be noted in the definition of safety – a region could be regarded as safe if the firearm assault incidents are low, but not necessarily so if all such incidents are targeting certain demographic groups.

Other possible factors were tested by Jiao upon looking into the demographic of offenders related to gun violence in Boston city (Jiao, 2013). Jiao did not use a complete dataset with individuals' demographics, rather a compilation of selected 250 entries from 2001 to 2007 were selected by the local police department on the basis of file completion status was provided to Jiao for the study.

Jiao tested with statistical significance that local residents tend to commit more serious weapon offenses using revolvers, shotguns and/or riffles. Jiao also find positive correlation between young age, local residency, prior records, and violent offenses. Due to the small data size, as well as other limitations such as the representativeness of the 250-entry sample data, Jiao used 0.1 significance level.

With attempts to look into data on census tract level, other considerations have been justified by scholars. For example, 'neighborhood are not just affected by their internal characteristics but also by those of nearby neighborhoods' (Sampon, 2012). To this finding, the paper will use clustering to testify if such 'nearby influence' is present in Boston as well.

Jiao's study may be limited by the sample data the study is based on, but it at least provides relevant variables to be modelled – age of individuals and local residency. Even though the study takes data on the city level and studies the effect of individual's demographics on likelihood of committing gun crime, it can be inferred that residential regions with higher local residents are more likely to see gun crimes. However, in contrary to the previous study, this paper finds no statistically strong evidence that firearm crimes are associated with individuals' race – and so the demographic of a residential region. This paper thus aims to explore any relationship between certain demographic profile of a region and the level of firearm violence in a census tract.

Section 3: Data

911 call data used is from a database curated by the Boston Area Research Initiative (BARI) based on records by the City of Boston's 911 system and accessed via Harvard

Dataverse.[1]

Among all the data files, this paper takes '911 Econometrics CT Longitudinal, Yearly, External.csv' and only data on the prevalence of Guns, Private Conflicts, Violence collected in between 2010 and 2018 inclusive will be analyzed by this paper. All values in the data file are recorded as number of incidents per 1000 people. And census tracts with population less than 500 are excluded from for more valid interpretation of these rates. For this paper, analysis will be done on census tract level.

Demographics data at census tract level for the city of Boston is obtained from Federal Financial Institutions Examination Council.

Original demographic information includes tract median income as percentage of the state median income, percentage of tract residents identified as racial minorities, percentage of owner-occupied units, percentage of units hosting fewer than 5 families, as well as values recorded in their base unit, such as tract median income in US dollars and population.

Certain tracts are removed during data cleaning process as they do not present a valid phenomenon in the analysis of the relations between demographics and firearm violence[2]. In addition, due to the limited availability of demographic data on census tract level, no data regarding other usual demographic terms are collected for this paper, namely the distributions of education, age, gender et cetera.

Refer to Table 1 for the summary statistics on the data. There are in total 1399 observations of 10 variables. These variables are:

Variable Name	Definition
Guns	911 Calls in which guns are reported, recorded in number of cases per 1000 people
PrivateConflict	911 Calls in which private conflicts are reported, recorded in number of cases per 1000 people
Violence	911 Calls in which violence are reported, recorded in number of cases per 1000 people
FireVio	911 Calls in which are labelled with both 'Guns' and 'PrivateConflict' or with both 'Guns' and 'Violence'. This is the independent variables in regression analysis and has been scaled up by 10E7 for easier analysis.
MedianIncome%	Tract median family income as a percentage of the state median income
Population	Tract population, in unit of 1000 people
Minority%	Percentage of tract population identified as racial minorities
OwenerOccupied%	Percentage of buildings in the tract occupied by owners
1- to 4 – Family Units%	Percentage of buildings in the tract that are of 1- to 4-family units
Year	Year in which the data was recorded

1 The data set is accessed via https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/XTEJRE

2 Reasons include tracts covering non-residential area, and tracts having unestablished population

Section: 4 Model

4.1 K-means clustering and spatial K-means clustering

Two models will be determined to describe any observable relations – regression model and clustering model. As both models use the same data sets, and missing data is present from demographic data set – such as income level of a census tract in a particular year is not available, or only a particular year was the income recorded. For example, income data on tract 1501 is extremely limited[3]. Missing variables for these census tracts will be filled with the mean of the same observations that are not missing for other tracts for both models.

The first model will be based on clustering in the following model:

$$X \rightarrow Z$$

$$V \rightarrow G$$

Where:

X consists of all five demographic variables from FFIEC data, including:

Tract Median Family Income %

Tract Population

Tract Minority %

Owner Occupied Units

1- To 4- Family Units

V consists of all variables other than the census tract ID from 911 call data:

Guns

Private Conflict

Violence

Arrow ($\rightarrow$) denotes a specific mapping from left to right

Z consists of categorical classes that put each census tract into different groups based on demographic data.

G consists of same number of categorical classes that put each census tract into different groups based on public safety data.

Clustering algorithm will capture unseen features of the data and group census tracts with similar characteristics in the same group. In determining the optimal number of clusters to obtain, an elbow plot will be made to assess the change in sum of squared errors with different clusters. The optimal number of clusters should be the one bringing about the largest improvement on sum of squared errors.

By applying clustering techniques before regression analysis justifies the possible

[3] Tract 1501 consists of large water body and Thompson Island, which is currently managed by a non-profit organization.

presence of any link between the demographic information and the firearm violence information both within each census tracts as well as between nearby tracts.

After the census tract are grouped in different classes based on two sets of variables, comparisons will be done between groups in Z and G to evaluate the similarity of the two. Namely, the proportion of total census tracts that are present in both corresponding groups of Z and G. K-means clustering and similarity between clustering will be done using the programming application R.

In addition to ordinary K-means clustering, a spatial K-means clustering will also be used with spatial analysis tools such as ArcMap and GeoDa to capture the geographic location of each census tracts. This is to explore if firearm violence in a census tract is determined by both intra-tract characteristics and inter-tract characteristics, or if firearm violence is determined regardless of demographic profiles in each census tracts.

The purpose of this model is to have a preliminary assessment on the expected effects of each variable on the degree of gun violence, as well as provide visual evidence that difference in demographic profiles of each tract is associated with difference in their firearm violence profiles.

4.2 Linear regression

Beyond clustering techniques, which analyzes continuous measurement of firearm violence as discrete measurement, linear regression will be done to assess the continuous nature of the measurement. The linear regression model will be of the following form:

$$Y = \beta_0 + \Sigma(\beta_i X_i)$$

Y = Firearm Violence, and is computed using

$(A \cap B \cap C) = A*B*C$

A = Guns

B = Violence

C = Private Conflict

β = coefficients of each independent variable, β_0 denotes the error term

X denotes independent variables of interests[4], consisting of:

Variable Type	Variables
Original demographic variables (5)	Minority% MedianIncome% Population OwnerOccupied% FamilyUnits%
Squared demographic variables (3)	Minority%^{2} Population^2 OwnerOccupied%^{2}
Interaction variables (1)	Minority% * MedianIncome%

[4] Demographic data on census tract level is accessed via Federal Financial Institutions Examination Council.

Fixed effect variables (1)	Year

As Blacks and lower-income regions are more likely to be associated with firearm assaults (Beard, et al., 2017), Minority% and MedianIncome% are expected to have a positive and a negative coefficient respectively. OwnerOccupied% is expected positive sign. The intercept, or the error term β_0 should be 0. Refer to Table 2 for detailed expectations of each variable.

As demographic variables have been studied to have correlations with different measures of crimes and violence at national, state and city level, running a linear regression on tracts level explores and testifies similar correlations on a census tract level. In addition, clustering may introduce misclassifications and bias through grouping census tracts into discrete labels as these records are of continuous variables. Running a simple linear regression model explores a more accurate and intuitive assessment on the effect of demographic variables on the degree of firearm violence in a given census tract.

As both models do not specify the role of individuals related to the crimes, minorities – if proven with statistical significance are associated with higher gun violence – could either be the attacker or the victims of gun violence. Therefore, having strong associations between minorities and gun violence does not suggest minorities are more or less likely to possess threat to the public safety of the region.

Due to constraints on the data set, this model fails to capture or control other possible relevant variables such as the prevalence of gun ownership, number of police posts or stations et cetera. Omitting the variable 'Gun Ownership' may create random errors to the error term β_0 and impedes the accuracy of the model. In terms of clustering, omitting a useful dimension will reduce the number of optimal clusters on the data sets, and induce inaccurate associations if the data is not hierarchical.

Section 5: Empirical Analysis

5.1 K-means clustering

K-means clustering is applied on both demographic data and 911 call data. Elbow plots are used to determine the optimal number of clusters in grouping the demographic data by year, and all present clear indication for the optimal number of clusters to be 3(Figures 1).

After applying k-means on 911 calls data defining the intended number of clusters to be 3, clustering similarities are evaluated in Table 3.

Table 3 contains the average clustering similarity through the years is 53%, with a maximum of 55%, and minimum of 50%. While similarity is only slightly better than 50%, it can be accepted due to two major reasons. First is omitting variables bias. Some demographic variables such as gender distribution, education distribution, police force per capital who are likely to be associated with firearm violence are not included in the model, reducing the demographic profile fed to the algorithm to a lower dimension and inducing higher misclassification rates. Second is missing geographic information. The demographic data does not contain the latitude or longitude of each tract but representing them by a string of numbers or census tract identification number (CT_ID). This would omit further useful information that might be important to the analysis of possible relations between demographic profiles and firearm violence profiles of a tract.

To improve the clustering technique, a spatial clustering will be conducted. While spatial clustering works similar to K-means clustering, it captures the geographic location of census tracts as well before clustering. Visual representations are also available for further analysis of the demographic and firearm violence effects. Figure 2 and Figure 3 are the spatial clustering results and Table 4 and Table 5 show the clustering statistics.

From the clustering using Guns, Private Conflicts and Violence, the majority of the Boston county have very low incidences of any firearm violence, certain level of such violence is present in central and south county. This clustering has a fairly small value for sum of squares, suggesting that firearm violence in Boston could generally be classified into three distinct groups, such as safe, moderate and dangerous.

For the clustering using demographic information, the green tracts reflect areas with medium to low income, relatively higher population and minority percentage, as well as lower owners occupied units. Tracts colored in red reflect tracts with high Median income, lower population, lower minority percentage and higher owners occupied units.

Visual comparison suggests that tracts colored red in demographic clustering are not necessarily always colored red in firearm violence clustering. However, tracts colored red in demographic clustering often have neighboring tracts colored red in firearm violence clustering. This seem to suggest that characteristics of a tract also affect neighboring tracts in terms of firearm violence. In other words, tracts are affected by both internal characteristics and those of nearby tracts. Considering both numeric clustering similarity and spatial clustering similarity, the degree of firearm violence in Boston census tracts is both related to individual tracts' demographic profiles, as well as nearby tracts' demographic profiles.

5.2 Linear regression

Most of the variables demonstrated statistical significance, such as MedianIncome%, Population, OwnerOccupied%, FamilyUnits%, Minority%^{2}, OwnerOccupied%^{2}, and Population^2.

The linear regression model enhances some of the initial hypothesis that demographic information at tracts level is correlated with the degree of firearm violence due to the statistical significance of most the variables, in particular, higher population, higher OwnerOccupied%, and higher Minority%^{2} are correlated with higher firearm violence. Both population and OwnerOccupied% echoed with our initial hypothesis, suggesting that at tracts level in Boston, higher population density of a tract is associated with higher firearm violence rates, and given that there are certain percentages of minorities in the tract, an autonomous increase in owners occupied units is also associated with higher firearm violence rates. While Minority% does not have a significant coefficient, it squared term, Minority%^{2} has. This suggests that when minority forms a very small part of the tract residents, they are unlikely to be associated with firearm violence in the tract, but when their proportion goes beyond certain threshold, firearm violence rates will be positively associated with any increase in minority proportion.

Table 7 shows the comparison between actual and expected signs for dependent variables' coefficients. While at higher levels of analysis such as city level or state level, higher income is often correlated with lower incidence of firearm violence, higher median income at census tract might induce higher violence targeted specifically to the tract. This is also suggested from studies using the American Housing Survey that higher mean income reduces disorder but raises crime (Hipp, 2007) and number of attractive targets (Rephann, 2007).

OwnerOccupied%^{2} being a negative sign could suggest that there exists a threshold for local residents proportion, and only beyond such threshold would the firearm violence incidence decrease.

One of the limitations that this model does not capture is the neighboring tract effect suggested from clustering technique. While clustering models suggest that the degree of firearm violence of a tract is determined by both internal characteristics of a tract and that of nearby tracts, linear regression does not consider neighboring tract effect.

In addition, this linear regression model does not take complete demographic profile of each tract for some other possible variables are omitted, such as number of police posts or stations in the tract, age distribution of the tract et cetera. Missing variable would create missing variable bias and this is seen from both the residual plot and Q-Q plot (Figure.4) which show heteroscedasticity in the model as well as major outliers.

The presence of outliers as shown in residual plot and Q-Q plot in Figure.5 could also explain the low similarities between K-means clustering in year 2016, 2017 and 2018. The large outliers and deviations from heteroscedasticity will contribute to misclassification by the algorithm on certain census tracts.

Section 6: Conclusion

Clustering models suggest that in Boston, the degree of firearm violence of a tract is determined by demographic characteristics of both the tract itself and tracts that are nearby. The models also suggested the direction of certain demographic variables on firearm violence in a tract. For example, higher firearm violence rates are often found in regions with high population and owner-occupied units.

The effect of tracts' internal demographic profile on the firearm violence is explored using linear regression model, which suggests that higher firearm violence is often associated with high median income, population and low proportion of family units. In addition, with higher income of tract's median income, minority percentage is less likely to be associated with firearm violence.

Therefore, city planners and public safety authorities, in attempts to curb or deter firearm violence in specific census tract, should put more emphasis on tracts with certain demographic variables such as tract median income, owners-occupied units. These tracts are likely to be victims of high incidence of firearm violence.

In addition, based on the clustering model, the demographic profile of a census tract does only affect the firearm violence rate of that tract, but also tracts around it. Authorities attempting to influence the firearm violence profile of certain tracts should thus go beyond studying the demographic profile of those targeted tract and study demographic profiles of tracts around the targeted tracts.

In further studies, cautions should be made when exploring correlations between demographic profiles and firearm violence of a given tract to capture the nearby effect. In better understanding how demographic influences firearm violence, attempts could be made to identify the role of the tract in firearm violence, either being an attractive target or violence initiator.

Section 7: References and Bibliography

Hughes, Roland (5 Aug 2019) 'US gun debate: Four dates that explain how we got here' *BBC*, accessed via: https://www.bbc.com/news/world-us-canada-42055871

Silverstein, Jason (1 Sep 2019) 'There have been more mass shootings than days this year' *CBS NEWS*, access via: https://www.cbsnews.com/news/mass-shootings-2019-more-mass-shootings-than-days-so-far-this-year/

Riddell CA, Harper S, Cerdá M, et al. 'Comparison of Rates of Firearm and Nonfirearm Homicide and Suicide in Black and White Non-Hispanic Men, by U.S. State.' *Annals of Internal Medicine*. 2018;168:712–720. [Epub ahead of print 24 April 2018]. doi: 10.7326/M17-2976

Beard, Jessica & Morrison, Christopher & Jacoby, Sara & Dong, Beidi & Smith, Randi & Sims, Carrie & Wiebe, Douglas. (2017). 'Quantifying Disparities in Urban Firearm Violence by Race and Place in Philadelphia, Pennsylvania: A Cartographic Study.' *American Journal of Public Health*. 107. e1-e3. 10.2105/AJPH.2016.303620.

Jiao, Allan Y. (2013) 'Gun incidents at the local level: understanding the demographic variables' Criminal Justice Studies, 26:2, 213-227, DOI: 10.1080/1478601x.2012.710614

Sampson, Robert J. (2012) 'Moving and the Neighborhood Glass Ceiling' *Science*, 337(6101), 1464-1465, DOI: 10. 1126/science.1227881

Hipp, John R. (2007) 'Block, Tract, and Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point', *American Sociological Review*, Vol. 72, No. 5, pp.659-680

Rephann, Terance J. (2007) 'Rental Housing and Crime: The Role of Property Ownership and Management' *The Annals of Regional Sciences*, 43(2) DOI: 10.1007/s00168-008-0215-1

Section 8: Tables and Graphs

Table.1 Summary statistics on variables

Variables	Min	1st Qu.	Medain	Mean	3rd Qu	Max
Guns	0	0.86	2.18	3.98	5.64	25.79
PrivateConflict	0	5.82	9.26	10.54	14.53	32.97
Violence	0	10.34	20.35	26.23	37.27	172.54
FireVio(*10^{-4})	0	0.492	3.970	29.900	27.100	104.000
MedianIncome%	15.81	48.74	71.07	83.94	104.69	274.62
Population	0.995	2.620	3.571	3.730	4.760	8.368
Minority%	1.36	25.13	47.63	52.71	83.15	99.7
OwnerOccupied%	0	7.39	12.89	13.97	20.36	34.68
1- to 4- Family Units%	0	16.96	29	25.43	34.21	51.1
Year	2010	2012	2014	2014	2016	2018

Table.2 Expected signs and effect of each demographic variables

Coefficients	Expected Sign	Expected effect
β_1 *Minority%*	+	Higher Proportion of minorities increases the chance of witnessing firearm violence in the tract
β_2 *MedianIncome%*	-	Higher income tracts are less likely to witness firearm violence
β_3 *Population*	+	Tracts with larger population are more likely to witness higher incidences of firearm violence
β_4 *OwnerOccupied%*	+	Tracts with fewer non locals are likely to see higher incidences of firearm violence
β_5 *FamilyUnits%*	-	Tracts with more buildings such as apartments or commercial buildings are more likely to witness higher incidences of firearm violence
β_6 *Year*	+	Fixed effect from year to year
β_7 *Minority%*^{2}	+	Expected coefficient is to be smaller than Minority% for the incidence of firearm violence is expected to level off as minority% increases
β_8 *Population*^{2}	+	Expected coefficient is to be smaller than Population for the incidence of firearm violence is expected to level off as population increases
β_9 *OwnerOccupied %*^{2}	-	Tracts with extremely high proportion of local residents is expected to have fewer incidence of firearm violence
β_{10} *Minority% * MedianIncome*	-	Compared with tracts with similar proportion of minorities, tracts with higher income are expected to witness lower firearm violence

Table.3 K-Means clustering similarity for each year

Year clustered	Similarities between clustering
2010	55%
2011	54%
2012	53%
2013	55%
2014	53%
2015	53%
2016	50%
2017	50%
2018	50%
Average	53%

Clustering similarities between clusters by demographic information and cluster by firearm violence are calculated for each year. On average, there is 53% of similarities between the two clustering where the best similarities occurred in 2010 and 2013, and worst similarities occurred in 2016, 2017, and 2018, with 50% similar.

Table. 4 Spatial clustering output using demographic variables

Method	KMeans				
No. of Clusters	3				
Distance	Euclidean				
Cluster Center	MedianIncome%	Population	Minority	OwnersOccupied%	FamilyUnits
C1	46.3	3.5	64.6	7.9	20.0
C2	99.6	3.3	24.2	19.0	31.0
C3	202.2	3.0	21.3	21.1	16.2
Total sum of squares		733.402			
Within-cluster sum of squares					
C1		167.46			
C2		42.10			
C3		30.10			

Spatial clustering using demographic variables generate higher sum of squares than that using firearm violence data despite optimum number of clusters used. This could suggests that true demographic profiles on tracts level are of more complex dimension than that being used in this paper.

Table.5 Spatial clustering using Guns, Private Conflicts and Violence

Method	KMeans		
No. of Clusters	3		
Distance	Euclidean		
Cluster Center	Guns	Private Conflicts	Violence
C1	0.42	2.06	5.26
C2	1.72	8.64	21.29
C3	4.80	12.76	52.85
Total sum of squares		534	
Within-cluster sum of squares			
C1		55.05	
C2		57.12	
C3		41.80	

Spatial clustering using firearm violence data generate lower sum of squares than that in demographic variables. There is a clear trend in clusters that as the cluster labels goes from 1 to 3, the degree of firearm violence also increases.

Table.6 Linear regression statistics

Coefficients	estimate	Std. Error	T value	Pr(>|t|)	
(Intercept) β_0	- 4962	1204	- 4.121	4E-5	***
Minority Population β_1	0.021	0.397	0.054	0.956	
MedianIncome% β_2	0.329	0.083	2.814	0.0049	**
Population β_3	1.711	5.164	3.314	0.0009	***
OwenrOccupied% β_4	1.764	0.778	2.268	0.0234	*
FamilyUnits% β_5	- 0.478	0.186	- 2.570	0.0102	*
Year β_6	2.445	0.599	4.084	4.69E-5	***
Minority%^{2} β_7	0.013	0.003	4.978	7.23E-7	***
OwnerOccupied% β_8	- 0.056	0.022	- 2.562	0.0104	*
Population^2 β_9	- 2.391	0.615	- 3.887	0.0001	***
Minority% * MedianIncome% β_{10}	- 0.008	0.002	- 3.710	0.0002	***
Signif codees: 0 '***' 0.001'**' 0.01'*' 0.05'.' 0,1' ' 1					
Multiple R-squared: 0.2928		Adjusted R-squared: 0.2877			

Most of the coefficients are significant except for Minority Population. However, its squared term Minority%^{2} is highly significant. This suggests that minority percentage of a census tract is likely to be non-linearly correlated with firearm violence of the tract.

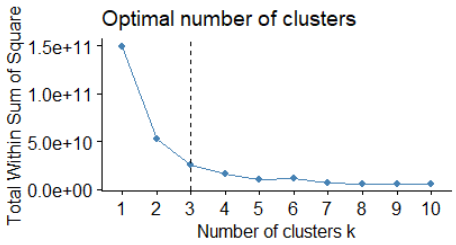
Table.7 Comparison of expected and actual signs for coefficients

Coefficients	Expected Sign	Actual Sign
β_1 Minority Population	+	+
β_2 MedianIncome%	-	+
β_3 Population	+	+
β_4 OwenrOccupied%	+	+
β_5 FamilyUnits%	-	-
β_6 Year	+	+
β_7 Minority%^{2}	+	+
β_8 OwnerOccupied%^{2}	+	-
β_9 Population^2	-	-
β_{10} Minority% * MedianIncome%	-	-

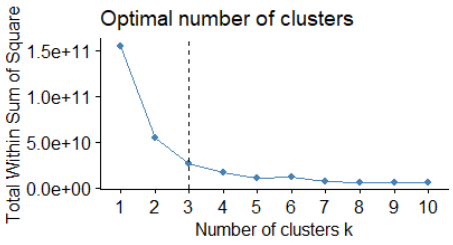
Most of the coefficients have actual signs or direction of effect in line with the expected signs, except for MedianIncome% and OwnerOccupied%^{2}

Graphs

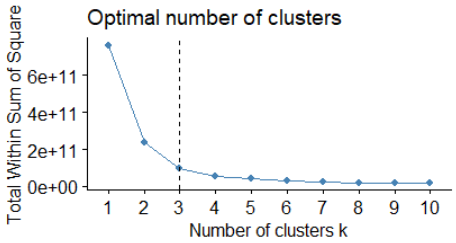
Figures.1



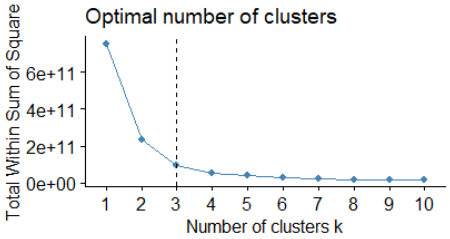
Elbow plot for 2010 demographics data



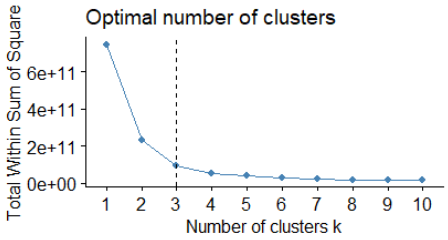
Elbow plot for 2011 demographic data



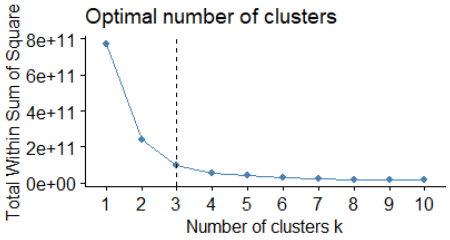
Elbow plot for 2012 demographic data



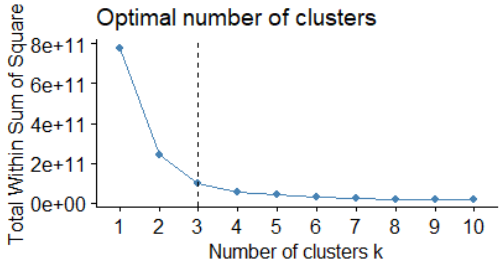
Elbow plot for 2013 demographic data



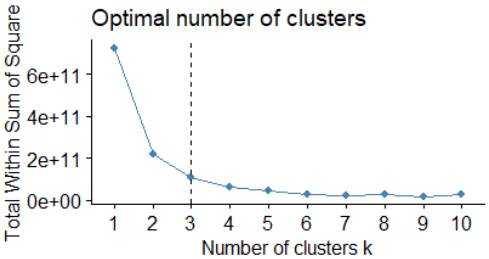
Elbow plot for 2014 demographic data



Elbow plot for 2015 demographic data



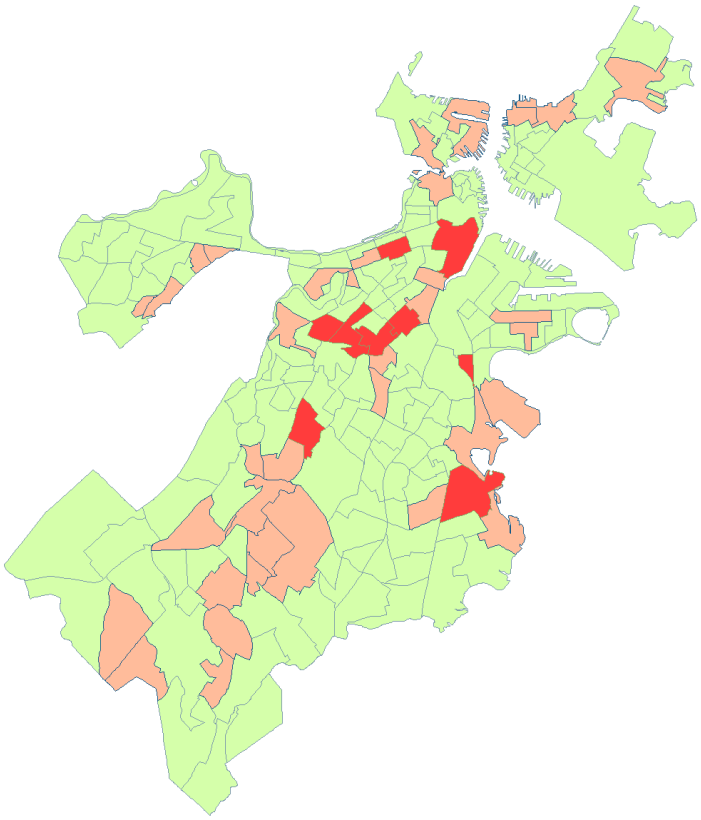
Elbow plot for 2016 demographic data



Elbow plot for 2017 demographic data

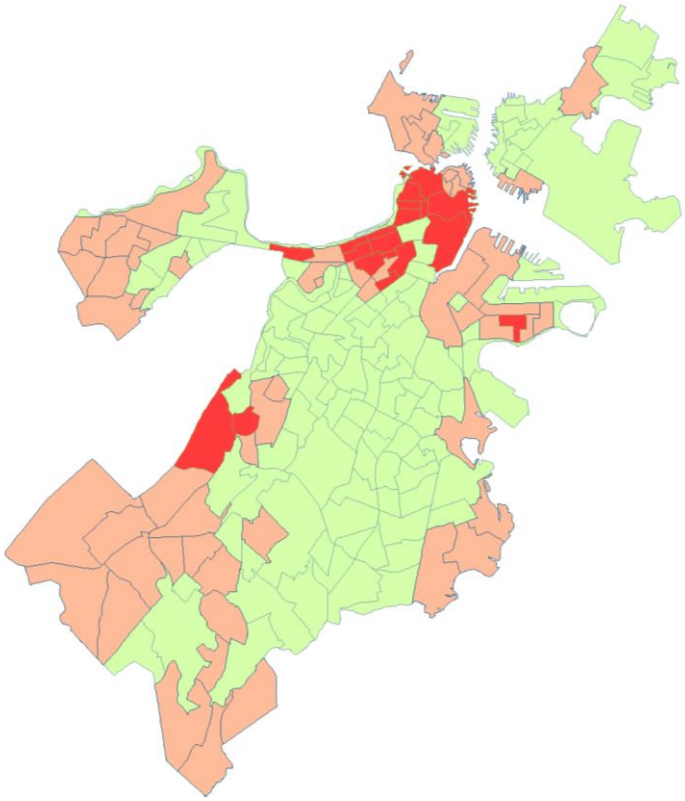
Elbow plots using demographic data through out the years all suggest an optimal cluster number of 3 clusters for K-means algorithm with largest decrease in sum of squared errors.

Figure.2 Spatial clustering using Guns, Private conflicts and Violence. Label 1 denotes lowest level of firearm violence



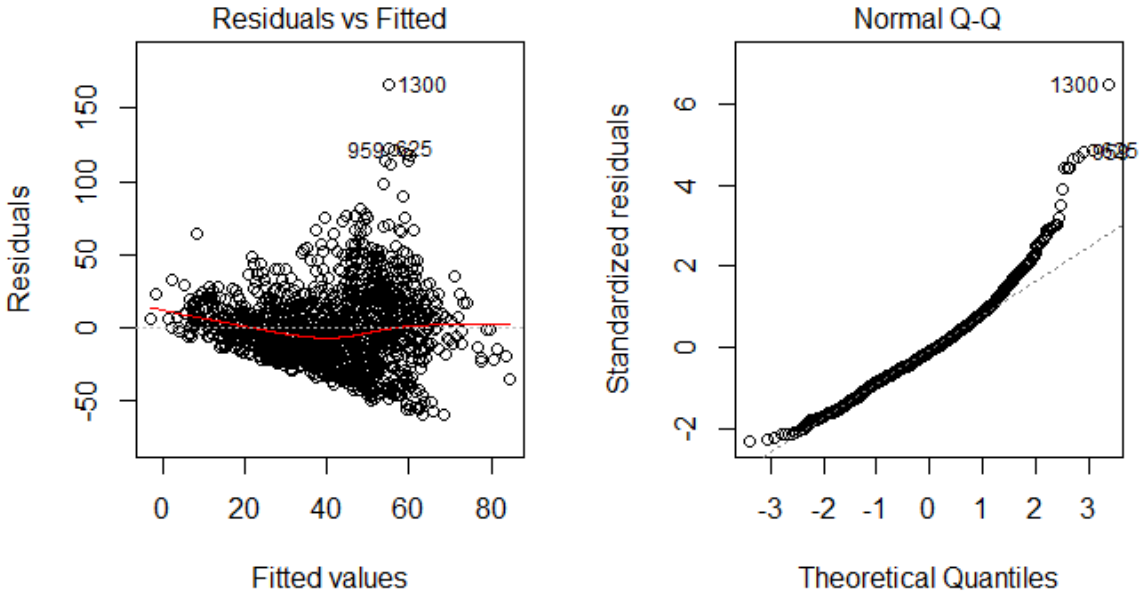
Clustering using firearm violence data including guns, private conflicts, and violence shows that most of the higher incidence tracts – tracts colored orange or red. These higher incidence tracts are present in north, central, southeast and southwest of Boston county.

Figure.3 Spatial clustering using demographic variables. Label 1 denotes higher characteristics of 'dangerous' tracts



Spatial clustering by demographic. Tracts with higher color intensity such as colored orange and red have higher Median Income, lower Minority%, and higher Owner Occupied units This clustering suggests that most of the high characteristic tracts are in north and south west of Boston, and are more accumulated compared to the clustering by firearm violence data.

Figures.4 – Residual plot on the left shows that heteroscedasticity is present in the linear regression, suggesting relevant variables are missing from the regression model. Normal Q-Q plot on the right indicates some large outliers and deviations from the model, suggesting the model is useful to most of the observation but there is also presence of certain extreme outliers that are not captured by the model. .



Section 9: Appendix, the following codes are executed in analyzing the data for this paper on R version 3.6.1 (2019-07-05) with platform x86_64-w64-mingw32

```
#INSTALL PACKAGES
install.packages('readxl')
install.packages('janitor')
install.packages('dplyr')
install.packages('factoextra')
install.packages('clusteval')
install.packages('devtools')
install.packages('rgdal')
install.packages('sf')
install.packages('geosphere')
library(readxl)
library(janitor)
library(dplyr)
library(factoextra)
library(clusteval)
library(wkb)
library(sp)
library(rgdal)
library(rgeoda)
library(sf)
library(geosphere)

#IMPORTING DEMOGRAPHIC DATA
raw10 <- read_excel('2010.xlsx',col_names=TRUE,skip=1)
raw10 <- remove_empty(raw10,which=c("cols"))

raw11 <- read_excel('2011.xlsx',col_names=TRUE,skip=1)
raw11 <- remove_empty(raw11,which=c("cols"))

raw12 <- read_excel('2012.xlsx',col_names=TRUE,skip=1)
raw12 <- remove_empty(raw12,which=c("cols"))

raw13 <- read_excel('2013.xlsx',col_names=TRUE,skip=1)
raw13 <- remove_empty(raw13,which=c("cols"))

raw14 <- read_excel('2014.xlsx',col_names=TRUE,skip=1)
raw14 <- remove_empty(raw14,which=c("cols"))

raw15 <- read_excel('2015.xlsx',col_names=TRUE,skip=1)
raw15 <- remove_empty(raw15,which=c("cols"))

raw16 <- read_excel('2016.xlsx',col_names=TRUE,skip=1)
raw16 <- remove_empty(raw16,which=c("cols"))
```

```
raw17 <- read_excel('2017.xlsx',col_names=TRUE,skip=1)
raw17 <- remove_empty(raw17,which=c("cols"))

raw18 <- read_excel('2018.xlsx',col_names=TRUE,skip=1)
raw18 <- remove_empty(raw18,which=c("cols"))

#IMPORTING 911 DATA
raw_911<-read.csv('911 Ecometrics CT Longitudinal, Yearly, External.csv',header=TRUE)

#SYNCHRONIZE THE TRACT ID
raw10$Tract_ID <- raw10$`State Code`*1000000000 + raw10$`County Code`*1000000 +
raw10$`Tract Code`*100
raw10 <- select(raw10, -c(1,2,3))

raw11$Tract_ID <- raw11$`State Code`*1000000000 + raw11$`County Code`*1000000 +
raw11$`Tract Code`*100
raw11 <- select(raw11, -c(1,2,3))

raw12$Tract_ID <- raw12$`State Code`*1000000000 + raw12$`County Code`*1000000 +
raw12$`Tract Code`*100
raw12 <- select(raw12, -c(1,2,3))

raw13$Tract_ID <- raw13$`State Code`*1000000000 + raw13$`County Code`*1000000 +
raw13$`Tract Code`*100
raw13 <- select(raw13, -c(1,2,3))

raw14$Tract_ID <- raw14$`State Code`*1000000000 + raw14$`County Code`*1000000 +
raw14$`Tract Code`*100
raw14 <- select(raw14, -c(1,2,3))

raw15$Tract_ID <- raw15$`State Code`*1000000000 + raw15$`County Code`*1000000 +
raw15$`Tract Code`*100
raw15 <- select(raw15, -c(1,2,3))

raw16$Tract_ID <- raw16$`State Code`*1000000000 + raw16$`County Code`*1000000 +
raw16$`Tract Code`*100
raw16 <- select(raw16, -c(1,2,3))

raw17$Tract_ID <- raw17$`State Code`*1000000000 + raw17$`County Code`*1000000 +
raw17$`Tract Code`*100
raw17 <- select(raw17, -c(1,2,3))

raw18$Tract_ID <- raw18$`State Code`*1000000000 + raw18$`County Code`*1000000 +
raw18$`Tract Code`*100
raw18 <- select(raw18, -c(1,2,3))
```

```
#SETTING UP SET X, AND SET V FOR CLUSTERING ANALYSIS
#X CONTAINS DEMOGRAPHIC VARIABLES
X10 <- select(raw10, c(3,7,8,10,11,12)) #10 and 11 will be devided by 7 to make it %
X11 <- select(raw11, c(3,7,8,10,11,12))
X12 <- select(raw12, c(3,7,8,10,11,12))
X13 <- select(raw13, c(3,7,8,10,11,12))
X14 <- select(raw14, c(3,7,8,10,11,12))
X15 <- select(raw15, c(3,7,8,10,11,12))
X16 <- select(raw16, c(3,7,8,10,11,12))
X17 <- select(raw17, c(3,7,8,10,11,12))
X18 <- select(raw18, c(3,7,8,10,11,12))

#REPLACING 'OWNER OCCUPIED' AND '1- TO 4- FAMILY UNITS' TO PERCENTAGE
OF TOTAL POPULATION
X10$`Owner Occupied Units`<- X10$`Owner Occupied Units`*100/X10$`Tract Population`
X10$`1- to 4- Family Units` <- X10$`1- to 4- Family Units`*100/X10$`Tract Population`

X11$`Owner Occupied Units`<- X11$`Owner Occupied Units`*100/X11$`Tract Population`
X11$`1- to 4- Family Units` <- X11$`1- to 4- Family Units`*100/X11$`Tract Population`

X12$`Owner Occupied Units`<- X12$`Owner Occupied Units`*100/X12$`Tract Population`
X12$`1- to 4- Family Units` <- X12$`1- to 4- Family Units`*100/X12$`Tract Population`

X13$`Owner Occupied Units`<- X13$`Owner Occupied Units`*100/X13$`Tract Population`
X13$`1- to 4- Family Units` <- X13$`1- to 4- Family Units`*100/X13$`Tract Population`

X14$`Owner Occupied Units`<- X14$`Owner Occupied Units`*100/X14$`Tract Population`
X14$`1- to 4- Family Units` <- X14$`1- to 4- Family Units`*100/X14$`Tract Population`

X15$`Owner Occupied Units`<- X15$`Owner Occupied Units`*100/X15$`Tract Population`
X15$`1- to 4- Family Units` <- X15$`1- to 4- Family Units`*100/X15$`Tract Population`

X16$`Owner Occupied Units`<- X16$`Owner Occupied Units`*100/X16$`Tract Population`
X16$`1- to 4- Family Units` <- X16$`1- to 4- Family Units`*100/X16$`Tract Population`

X17$`Owner Occupied Units`<- X17$`Owner Occupied Units`*100/X17$`Tract Population`
X17$`1- to 4- Family Units` <- X17$`1- to 4- Family Units`*100/X17$`Tract Population`

X18$`Owner Occupied Units`<- X18$`Owner Occupied Units`*100/X18$`Tract Population`
X18$`1- to 4- Family Units` <- X18$`1- to 4- Family Units`*100/X18$`Tract Population`

#G CONTAINS GUN VIOLENCE VARIABLES
G10 <- na.omit(select(raw_911, c(2,11,29,1)))
G11 <- na.omit(select(raw_911, c(3,12,30,1)))
G12 <- na.omit(select(raw_911, c(4,13,31,1)))
```

```
G13 <- na.omit(select(raw_911, c(5,14,32,1)))
G14 <- na.omit(select(raw_911, c(6,15,33,1)))
G15 <- na.omit(select(raw_911, c(7,16,34,1)))
G16 <- na.omit(select(raw_911, c(8,17,35,1)))
G17 <- na.omit(select(raw_911, c(9,18,36,1)))
G18 <- na.omit(select(raw_911, c(10,19,37,1)))

#COMBING THE TWO
FULL10 <- inner_join(G10, X10, by = c("CT_ID_10" = "Tract_ID"))
FULL11 <- inner_join(G11, X11, by = c("CT_ID_10" = "Tract_ID"))
FULL12 <- inner_join(G12, X12, by = c("CT_ID_10" = "Tract_ID"))
FULL13 <- inner_join(G13, X13, by = c("CT_ID_10" = "Tract_ID"))
FULL14 <- inner_join(G14, X14, by = c("CT_ID_10" = "Tract_ID"))
FULL15 <- inner_join(G15, X15, by = c("CT_ID_10" = "Tract_ID"))
FULL16 <- inner_join(G16, X16, by = c("CT_ID_10" = "Tract_ID"))
FULL17 <- inner_join(G17, X17, by = c("CT_ID_10" = "Tract_ID"))
FULL18 <- inner_join(G18, X18, by = c("CT_ID_10" = "Tract_ID"))

#DATA CLEANNING, REMOVING TRACT WITHOUT INCOME
clean10 <- FULL10[!(FULL10$`Tract Median Family Income %`==0 & FULL10$`Tract Population`==0),]
clean11 <- FULL11[!(FULL11$`Tract Median Family Income %`==0& FULL11$`Tract Population`==0),]
clean12 <- FULL12[!(FULL12$`Tract Median Family Income %`==0& FULL12$`Tract Population`==0),]
clean13 <- FULL13[!(FULL13$`Tract Median Family Income %`==0& FULL13$`Tract Population`==0),]
clean14 <- FULL14[!(FULL14$`Tract Median Family Income %`==0& FULL14$`Tract Population`==0),]
clean15 <- FULL15[!(FULL15$`Tract Median Family Income %`==0& FULL15$`Tract Population`==0),]
clean16 <- FULL16[!(FULL16$`Tract Median Family Income %`==0& FULL16$`Tract Population`==0),]
clean17 <- FULL17[!(FULL17$`Tract Median Family Income %`==0& FULL17$`Tract Population`==0),]
clean18 <- FULL18[!(FULL18$`Tract Median Family Income %`==0& FULL18$`Tract Population`==0),]

#ADD A YEAR VARIABLE TO THE DATA FRAME
clean10$year <- 2010
clean11$year <- 2011
clean12$year <- 2012
clean13$year <- 2013
clean14$year <- 2014
clean15$year <- 2015
clean16$year <- 2016
```

```
clean17$year <- 2017
clean18$year <- 2018

#SYNCHRONIZE THE COLUMN NAMES
colnames(clean10) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean11) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean12) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean13) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean14) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean15) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean16) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean17) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")
colnames(clean18) <- c("Guns", "PrivateConflict","Violence","CT_ID","MedianIncome%","Population","Minority%","OwnerOccupied%","1- to 4- Family Units%","Year")

#SUMMARY STATISTICS OF THE ENTIRE DATA SET
CleanFull <- rbind(clean10,rbind(clean11,rbind(clean12,rbind(clean13,rbind(clean14,rbind(clean15,rbind(clean16,rbind(clean17,clean18,by=c("CT_ID")),
                                                                                                                    by=c("CT_ID")),
                                                                                                             by=c("CT_ID")),
                                                                                                      by=c("CT_ID")),
                                                                                               by=c("CT_ID")),
                                                                                        by=c("CT_ID")),
                                                                                 by=c("CT_ID")),
                                                                          by=c("CT_ID")),
                                                                   by=c("CT_ID"))

#WHILE MERGING THE DATA, R TREATS EACH VALUE AS CHARACTER, NOW CONVERTING TO NUMERIC
```

```
CleanFull <- mutate_all(CleanFull, function(x) as.numeric(as.character(x)))
CleanFull <- CleanFull[!(CleanFull$Year==0),]
#WHILE SOME VALUES ARE TRUE 0, CONVERSION WILL COERCE THEM TO NAS, NOW CHANGING THEM BACK TO 0

CleanFull$`MedianIncome%`[CleanFull$`MedianIncome%` == 0 ]<- NA
CleanFull$`MedianIncome%` <- ifelse(is.na(CleanFull$`MedianIncome%`),
                                    mean(CleanFull$`MedianIncome%`, na.rm=TRUE),
                                    CleanFull$`MedianIncome%`)
#CleanFull[is.na(CleanFull)] <- 0
#CleanFull[CleanFull$`MedianIncome%`==0] <- mean(CleanFull$`MedianIncome%`)
summary(CleanFull)


#   USING KMEANS TO CLUSTER THE DATA
#LOOK FOR OPTIMAL NUMBER OF CLUSTERS BASED ON DEMOGRAPHIC FEATURES
fviz_nbclust(select(clean10,c(5,6,7,8,9,10)),kmeans,method='wss')
+geom_vline(xintercept=3,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean11,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean12,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean13,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean14,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean15,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean16,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean17,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
fviz_nbclust(select(clean18,c(5,6,7,8,9,10)),kmeans,method='wss')+geom_vline(xintercept=3
,linetype=2)
  #optimal = 3
```

```
#NOW CLUSTER THE DEMOGRAPHICS AND GUNS INTO 3 CLUSTERS AND FIND THE SIMILARITY FOR YEAR 10
XL10 <- kmeans(select(clean10,c(5,6,7,8,9,10)),3,algorithm = 'Lloyd')
clean10 <- cbind(clean10,cluster=XL10$cluster)
colnames(clean10)[11] <- 'XL'
GL10 <- kmeans(select(clean10,c(1,2,3)),3,algorithm='Lloyd')
clean10 <- cbind(clean10,cluster=GL10$cluster)
colnames(clean10)[12] <- 'GL'
SIM10 <- cluster_similarity(clean10$XL,clean10$GL,similarity = 'rand', method = 'independence')

#REPEAT FOR YEARS 11 THROUGH 18
XL11 <- kmeans(select(clean11,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean11 <- cbind(clean11,cluster=XL11$cluster)
colnames(clean11)[11] <- 'XL'
GL11 <- kmeans(select(clean11,c(1,2,3)),3,algorithm='Lloyd')
clean11 <- cbind(clean11,cluster=GL11$cluster)
colnames(clean11)[12] <- 'GL'
SIM11 <- cluster_similarity(clean11$XL,clean11$GL,similarity = 'rand', method = 'independence')

XL12 <- kmeans(select(clean12,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean12 <- cbind(clean12,cluster=XL12$cluster)
colnames(clean12)[11] <- 'XL'
GL12 <- kmeans(select(clean12,c(1,2,3)),3,algorithm='Lloyd')
clean12 <- cbind(clean12,cluster=GL12$cluster)
colnames(clean12)[12] <- 'GL'
SIM12 <- cluster_similarity(clean12$XL,clean12$GL,similarity = 'rand', method = 'independence')

XL13 <- kmeans(select(clean13,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean13 <- cbind(clean13,cluster=XL13$cluster)
colnames(clean13)[11] <- 'XL'
GL13 <- kmeans(select(clean13,c(1,2,3)),3,algorithm='Lloyd')
clean13 <- cbind(clean13,cluster=GL13$cluster)
colnames(clean13)[12] <- 'GL'
SIM13 <- cluster_similarity(clean13$XL,clean13$GL,similarity = 'rand', method = 'independence')

XL14 <- kmeans(select(clean14,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean14 <- cbind(clean14,cluster=XL14$cluster)
colnames(clean14)[11] <- 'XL'
GL14 <- kmeans(select(clean14,c(1,2,3)),3,algorithm='Lloyd')
clean14 <- cbind(clean14,cluster=GL14$cluster)
colnames(clean14)[12] <- 'GL'
```

```
SIM14 <- cluster_similarity(clean14$XL,clean14$GL,similarity = 'rand', method = 'independence')

XL15 <- kmeans(select(clean15,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean15 <- cbind(clean15,cluster=XL15$cluster)
colnames(clean15)[11] <- 'XL'
GL15 <- kmeans(select(clean15,c(1,2,3)),3,algorithm='Lloyd')
clean15 <- cbind(clean15,cluster=GL15$cluster)
colnames(clean15)[12] <- 'GL'
SIM15 <- cluster_similarity(clean15$XL,clean15$GL,similarity = 'rand', method = 'independence')

XL16 <- kmeans(select(clean16,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean16 <- cbind(clean16,cluster=XL16$cluster)
colnames(clean16)[11] <- 'XL'
GL16 <- kmeans(select(clean16,c(1,2,3)),3,algorithm='Lloyd')
clean16 <- cbind(clean16,cluster=GL16$cluster)
colnames(clean16)[12] <- 'GL'
SIM16 <- cluster_similarity(clean16$XL,clean16$GL,similarity = 'rand', method = 'independence')

XL17 <- kmeans(select(clean17,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean17 <- cbind(clean17,cluster=XL17$cluster)
colnames(clean17)[11] <- 'XL'
GL17 <- kmeans(select(clean17,c(1,2,3)),3,algorithm='Lloyd')
clean17 <- cbind(clean17,cluster=GL17$cluster)
colnames(clean17)[12] <- 'GL'
SIM17 <- cluster_similarity(clean17$XL,clean17$GL,similarity = 'rand', method = 'independence')

XL18 <- kmeans(select(clean18,c(5,6,7,8,9,10)),3,algorithm='Lloyd')
clean18 <- cbind(clean18,cluster=XL18$cluster)
colnames(clean18)[11] <- 'XL'
GL18 <- kmeans(select(clean18,c(1,2,3)),3,algorithm='Lloyd')
clean18 <- cbind(clean18,cluster=GL18$cluster)
colnames(clean18)[12] <- 'GL'
SIM18 <- cluster_similarity(clean18$XL,clean18$GL,similarity = 'rand', method = 'independence')

#TEST OUT IF SIM IS SIGNIFICANTLY GREATER THAN 0.5
SIM <- c(SIM10, SIM11, SIM12, SIM13, SIM14, SIM15, SIM16, SIM17, SIM18)
mean(SIM)
#MOST OF THE CLUSTERING HAVE MORE THAN 50% OF SIMILARITY.


#REGRESSION ANALYSIS
```

```
#USING YEAR 2010, 2012, 2014, 2016, 2018 AS SAMPLE DATA TO BUILD REGRESSION MODEL - TRAINING DATA
#USING YEAR 2011, 2013, 2015, 2017 AS TESTING/VALIDATION SET

#SET UP A VARIABLE CALLED FIREVIO,WHICH IS THE CASES LABELLED GUN AND VIOLENCE, OR GUN AND PRIVATE CONFLICT

clean10$FireVio <- clean10$Guns * clean10$Violence * clean10$PrivateConflict/100


clean11$FireVio <- clean11$Guns * clean11$Violence * clean11$PrivateConflict/100

clean12$FireVio <- clean12$Guns * clean12$Violence * clean12$PrivateConflict/100

clean13$FireVio <- clean13$Guns * clean13$Violence * clean13$PrivateConflict/100

clean14$FireVio <- clean14$Guns * clean14$Violence * clean14$PrivateConflict/100

clean15$FireVio <- clean15$Guns * clean15$Violence * clean15$PrivateConflict/100

clean16$FireVio <- clean16$Guns * clean16$Violence * clean16$PrivateConflict/100

clean17$FireVio <- clean17$Guns * clean17$Violence * clean17$PrivateConflict/100

clean18$FireVio <- clean18$Guns * clean18$Violence * clean18$PrivateConflict/100

#UPDATE THE CLEANFULL DATA SET FOR REGRESSION ANALYSIS
CleanFullFireVio <-
rbind(clean10,rbind(clean11,rbind(clean12,rbind(clean13,rbind(clean14,rbind(clean15,rbind(
clean16,rbind(clean17,clean18,by=c("CT_ID")),
                                                                    by=c("CT_ID")),
                                                                by=c("CT_ID")),
                                                            by=c("CT_ID")),
                                                        by=c("CT_ID")),
                                                    by=c("CT_ID")),
                                                by=c("CT_ID")),
                                            by=c("CT_ID")),
                                        by=c("CT_ID"))

#RUN THE REGRESSION
#AS IT IS EXPECTED THAT SOME USEFUL DEMOGRAPHIC VARIABLES ARE OMITTED FROM THE DATA SET, RIDGE REGRESSION IS USED
#SO THAT LARGE VALUES WILL NOT BE PENALIZED.
CleanFullFireVio <- mutate_all(CleanFullFireVio,function(x) as.numeric(x))
CleanFullFireVio$`MedianIncome%`[CleanFullFireVio$`MedianIncome%` == 0 ]<- NA
CleanFullFireVio$`MedianIncome%` <- ifelse(is.na(CleanFullFireVio$`MedianIncome%`),
                                           mean(CleanFullFireVio$`MedianIncome%`, na.rm=TRUE),
```

```
                                CleanFullFireVio$`MedianIncome%`)
CleanFullFireVio$Population<-CleanFullFireVio$Population/1000

lmod <- lm(FireVio ~ `Minority%` + `MedianIncome%` + Population
        +`OwnerOccupied%`+`1- to 4- Family Units%`+ Year
        +I(`Minority%`**2) +I(`OwnerOccupied%`**2) + I(Population**2)
        +I(`Minority%`*`MedianIncome%`), data=CleanFullFireVio)
summary(lmod)
plot(lmod)
#====================================
#plot each variables distribution

plot(CleanFullFireVio$Guns,CleanFullFireVio$PrivateConflict,main="How firearm reports reveal safety",xlab="Gun reports",ylab="Number of cases")
points(CleanFullFireVio$Guns,CleanFullFireVio$Violence,col='red')


#NEED A PLOT FOR DEMOGRAPHIC DATA
plot(CleanFullFireVio$`MedianIncome%`,CleanFullFireVio$`OwnerOccupied%`,main="Pro
file of different labels",xlab="Median Income",ylab="%")
plot(CleanFullFireVio$`MedianIncome%`,CleanFullFireVio$`Minority%`,main="Profile of different labels",xlab="Median Income",ylab="%",col='red')
points(CleanFullFireVio$`MedianIncome%`,CleanFullFireVio$`OwnerOccupied%`)
```