Mass Shootings in America

Teresa Gerhold, Josh Anderson, Nick O’Neill, Merzia Cutlerywala

Basic Details

- Datasets:
  - Stanford_MSA_Database - contains data about mass shootings in America from 1966 to 2016.
  - NYTimes API - contains various metadata about every news article from 1966 to 2016 containing the keyword “mass shooting.”
- Significant data fields:
  - City/state, date, fatalities, injured, venue, mental health history, possible motive, type of weapon, number of weapons, shooter name/age/race/gender, fate of the shooter, cause of death, and latitude and longitude of where the shooting occurred.
    - Average shooter age
    - Class
    - Count (Article Count)
- Dimensions post cleaning:
  - Stanford_MSA - 325 observations of 40 variables

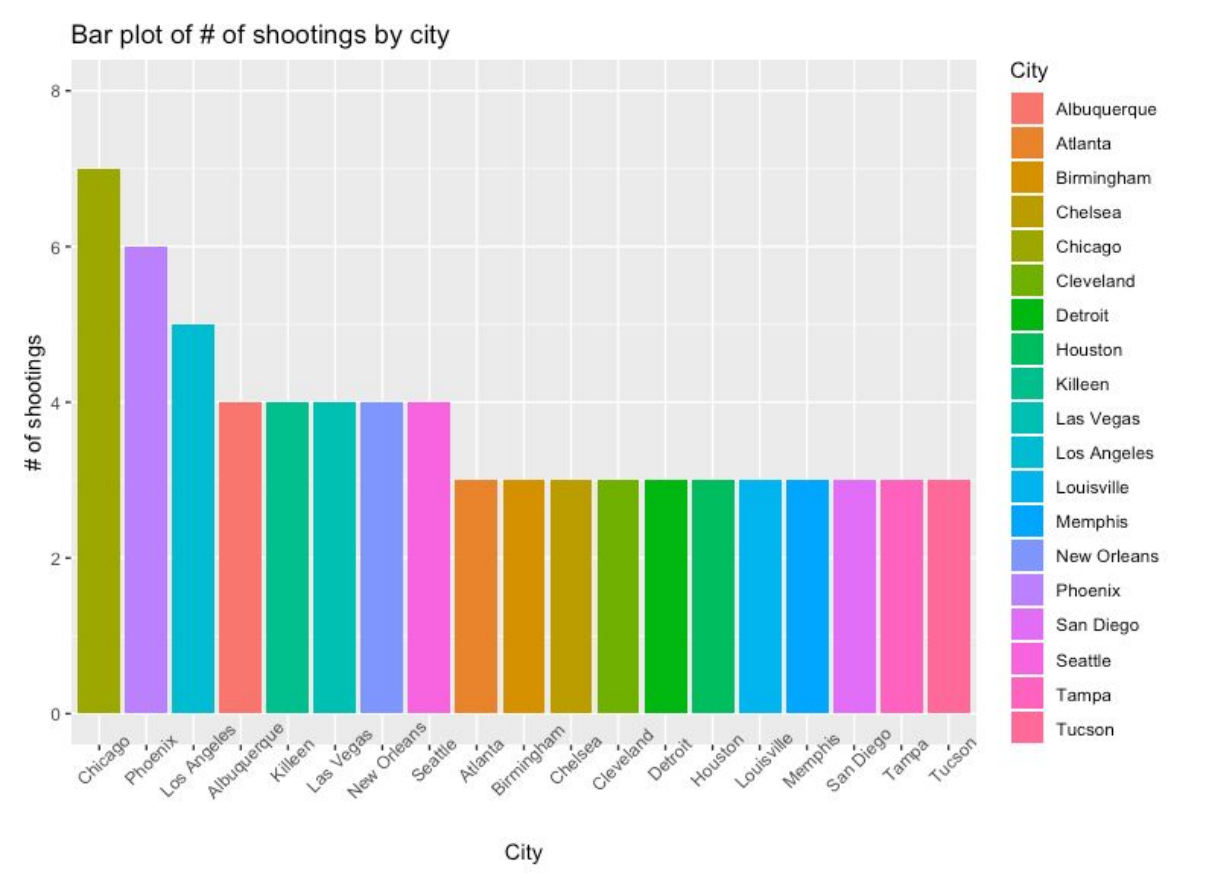
Data Cleaning

- Mutate to aggregate similar factors
  - Shooter race
  - Fate of the shooter at the scene
  - Type of gun
  - Shooter's cause of death
  - Type of place
  - Targeted victims
  - Possible motive
- Mutate to correct column types
- Select to remove description columns - CaseID, Description, Possible Motive Detailed, History of Mental Illness Detailed, Date Detailed, Targeted Victims Detailed, Type of Gun Detailed, Notes, Data Source 1 to 7
- Replaced NAs for numeric values with mean
- Corrected column types in the NYTimes news dataset
- Inner_join Stanford_MSA_Database and NYTimes API

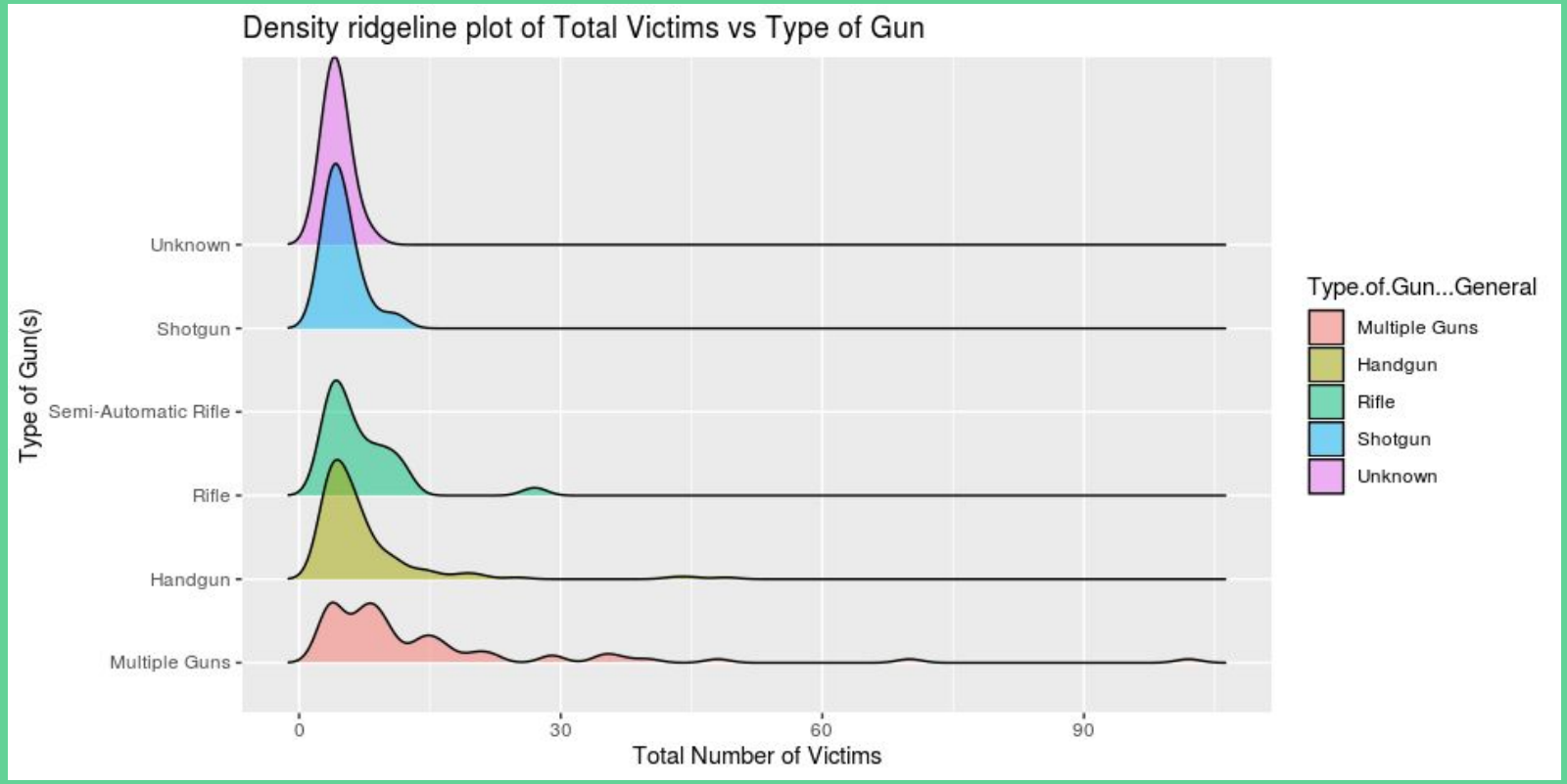
Summary Stats

Variable <fctr>	mean <dbl>	sd <dbl>	max <dbl>	min <dbl>
Total Number of Victims	7.904615	9.451312	102	3
Total Number of Fatalities	4.036923	4.814317	50	0
Shooter Age	31.824462	11.755356	70	12
Total Guns	1.704615	1.085316	10	0
Number of Shooters	1.104615	1.104615	4	1

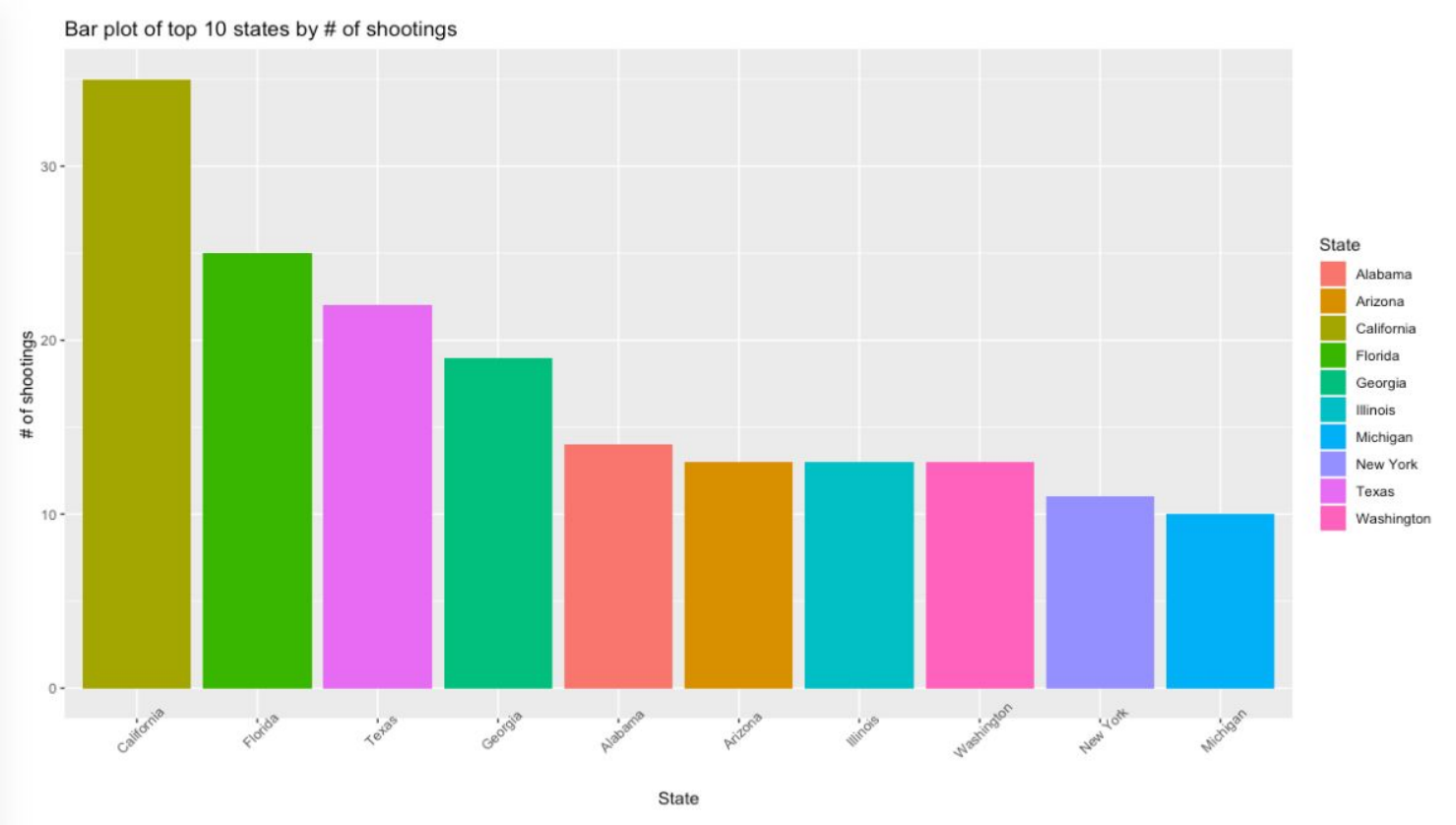
Summary Plots - Histogram based on city



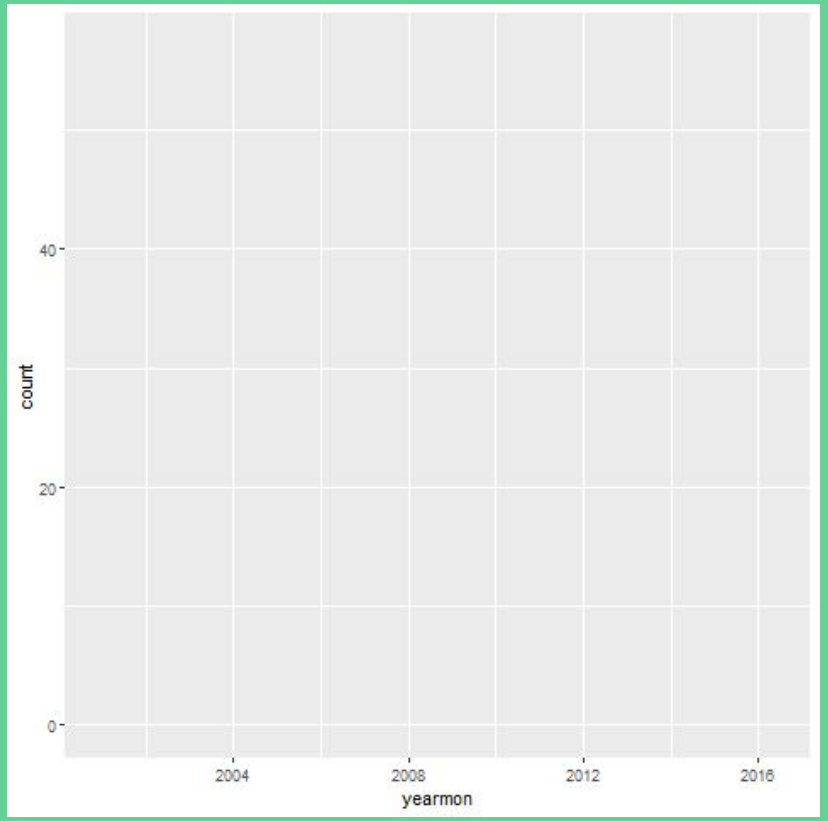
Summary Plots - Victims by Type of Gun

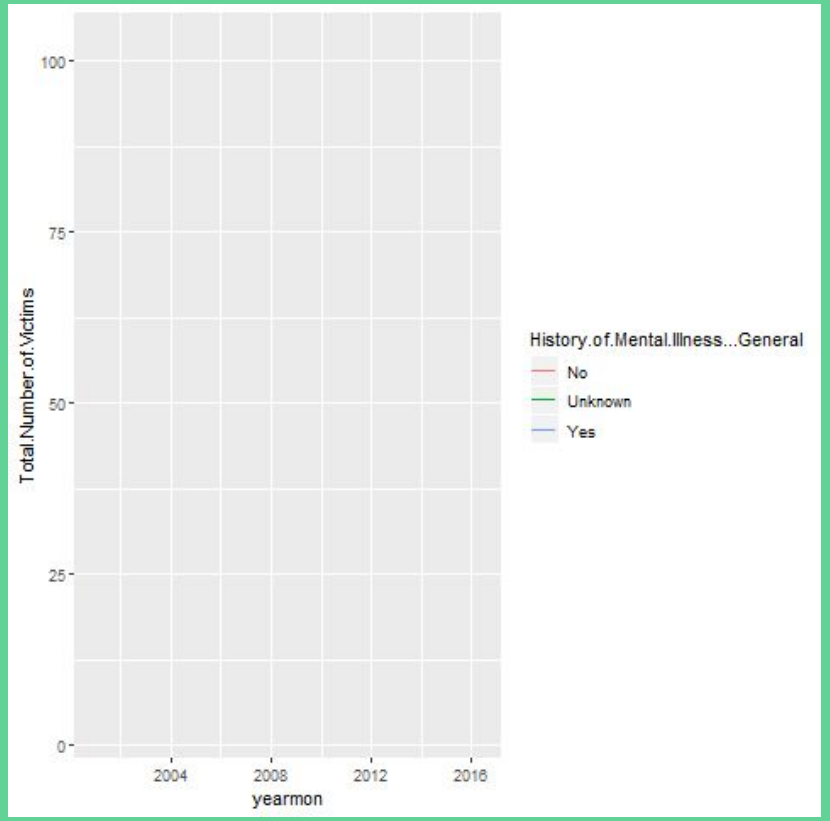


Summary Plots - Bar plot of top 10 states by # of shootings

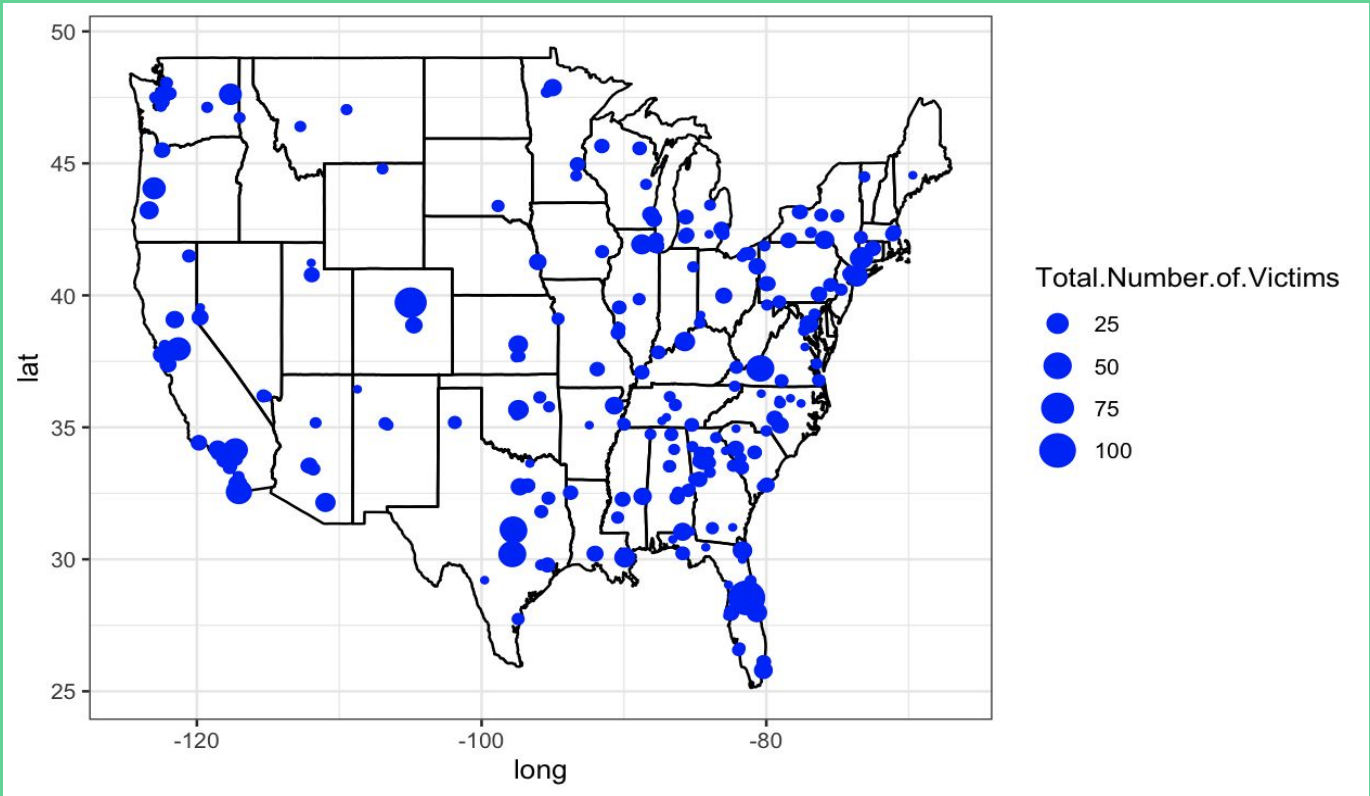


Summary Plots - Line Charts





Summary Plots- Location and Number of Victims



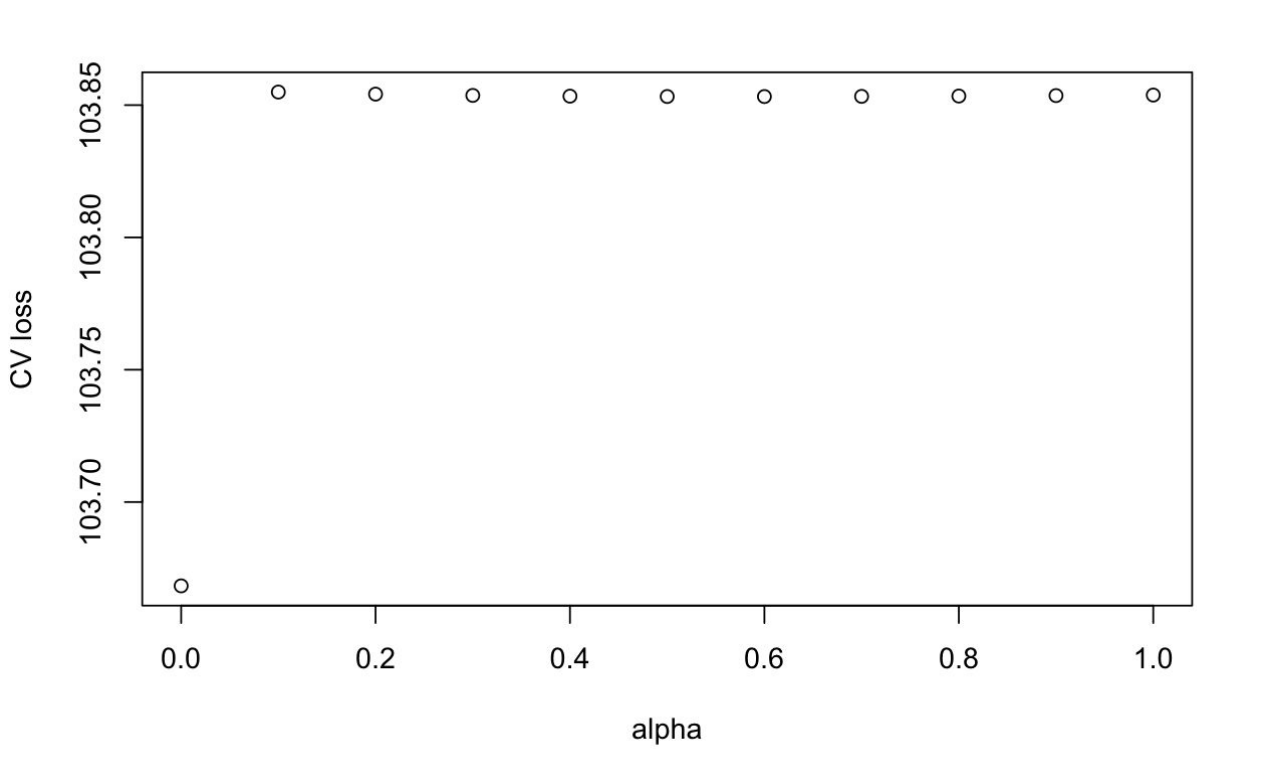
Observations so far ...

- What parts of America have higher severity shootings
- The type of guns impact the number of victims
- How many people are at risk in a given mass shooting

Predictive models used

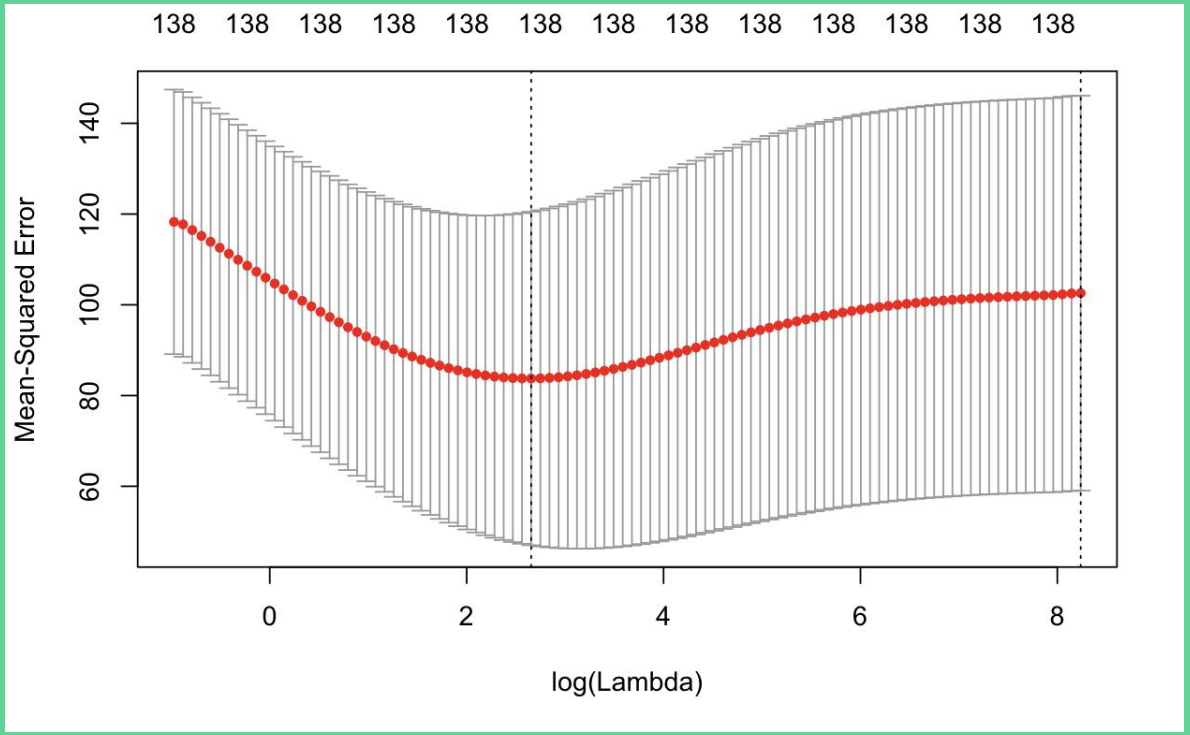
- ElasticNet - what predictors cause the severity of a mass shooting

Minloss Plot



- Optimal alpha is alpha = 0 because that's where cross validated loss is minimized
- Should be ridge instead

Ridge Diagnostics



Training:
RMSE - 7.158553
MAE - 3.723099
R2 - 0.5912334

Testing:
RMSE - 5.9811
MAE - 4.034448
R2 - 0.178672

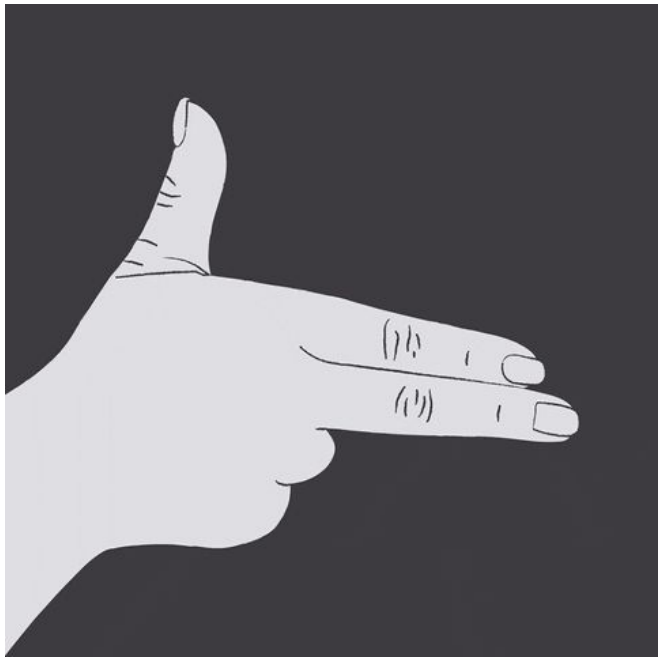
Factors more likely to lead to more victims

	1
(Intercept)	7.993782
StateColorado	7.545096
StateConnecticut	5.486067
Place.TypeMilitary facility	5.118114
Possible.Motive...GeneralRace	3.687715
Military.ExperienceNo	8.570086

Factors that would likely have fewer total victims

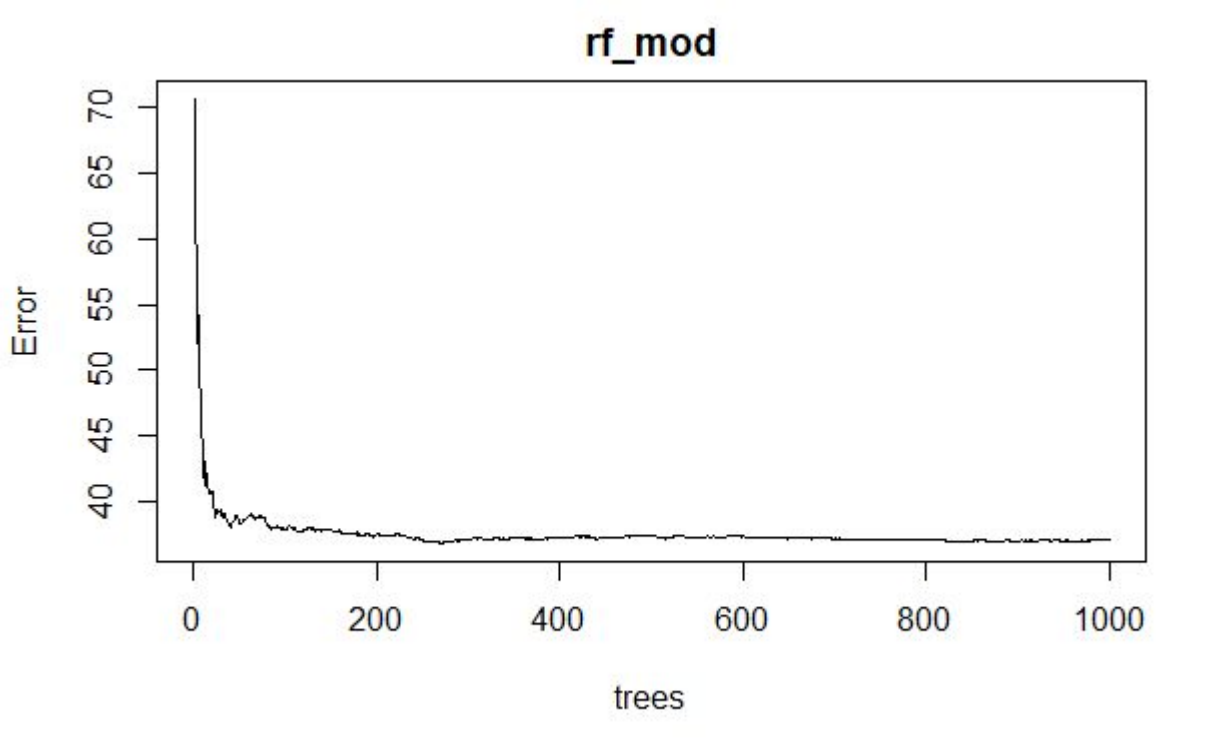
	1
StateAlaska	-1.558130
StateIowa	-4.747437
StateNew Jersey	-1.880130
StateNorth Carolina	-1.848597
StateOhio	-1.788287
StateTennessee	-1.586375
StateUtah	-2.263871
StateWisconsin	-1.546384
Place.TypePark/Wilderness	-1.605217

Predictive models used



- RandomForest - compare the performance metrics to see which model is better

Random Forest Diagnostics



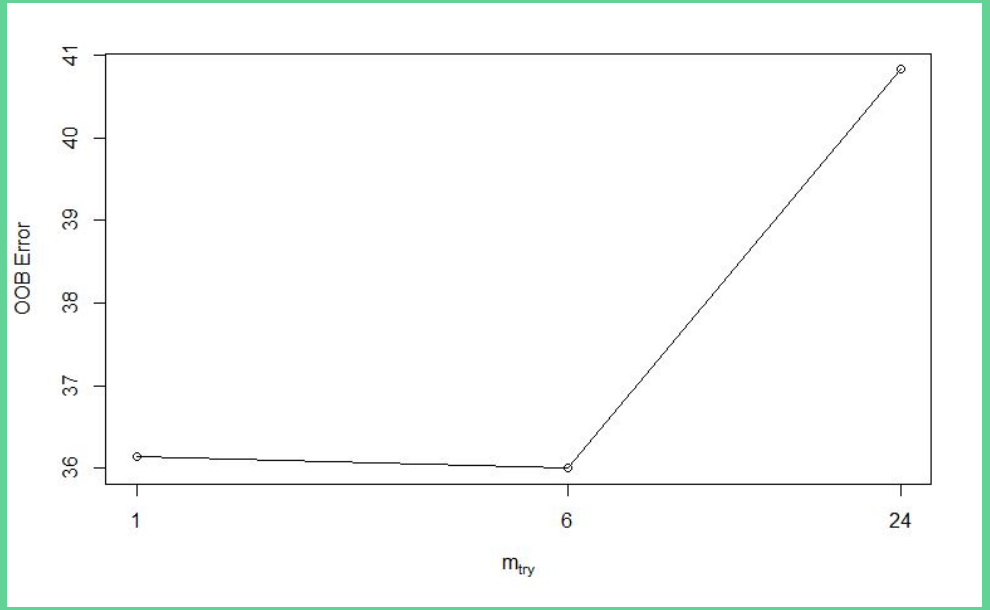
Ntree = 275
Mtry = 6
Type = regression

Training:
RMSE - 2.7437
MAE - 1.451888
R2 - 0.9475252

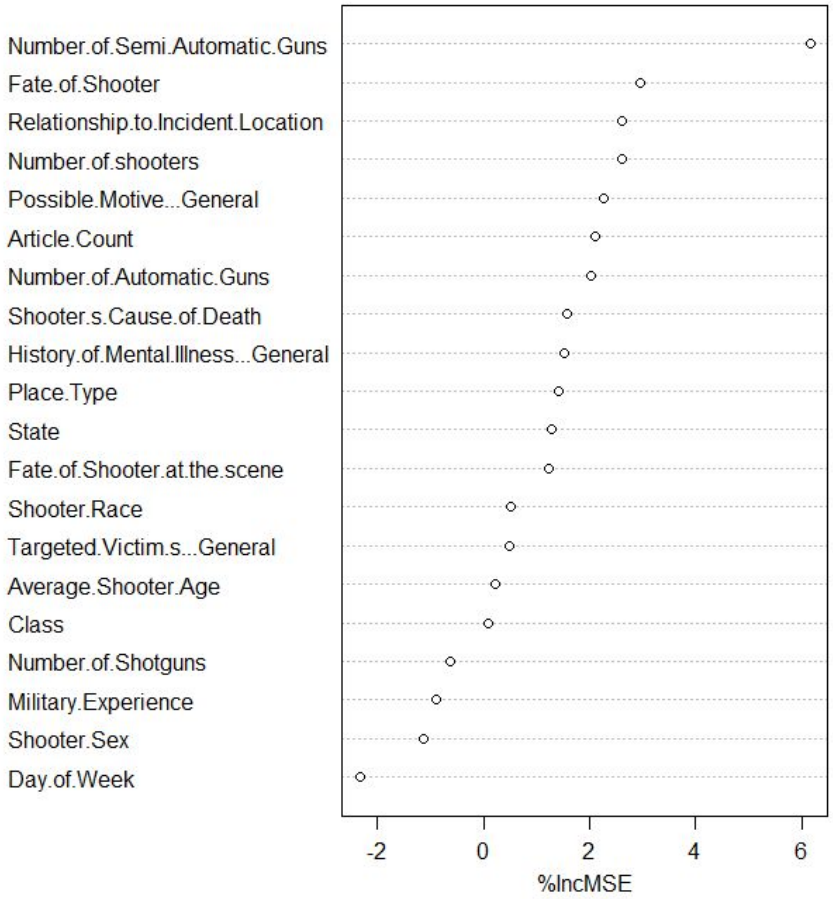
Testing:
RMSE - 15.21373
MAE - 6.124129
R2 - 0.2661599

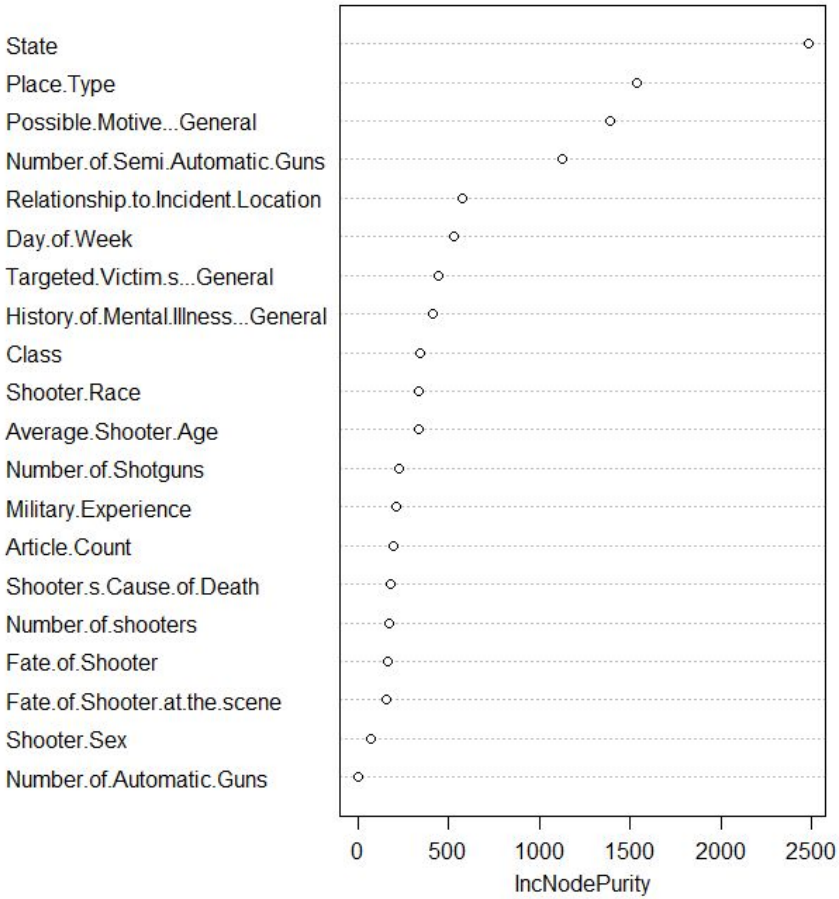
Random Forest Diagnostics - mtry

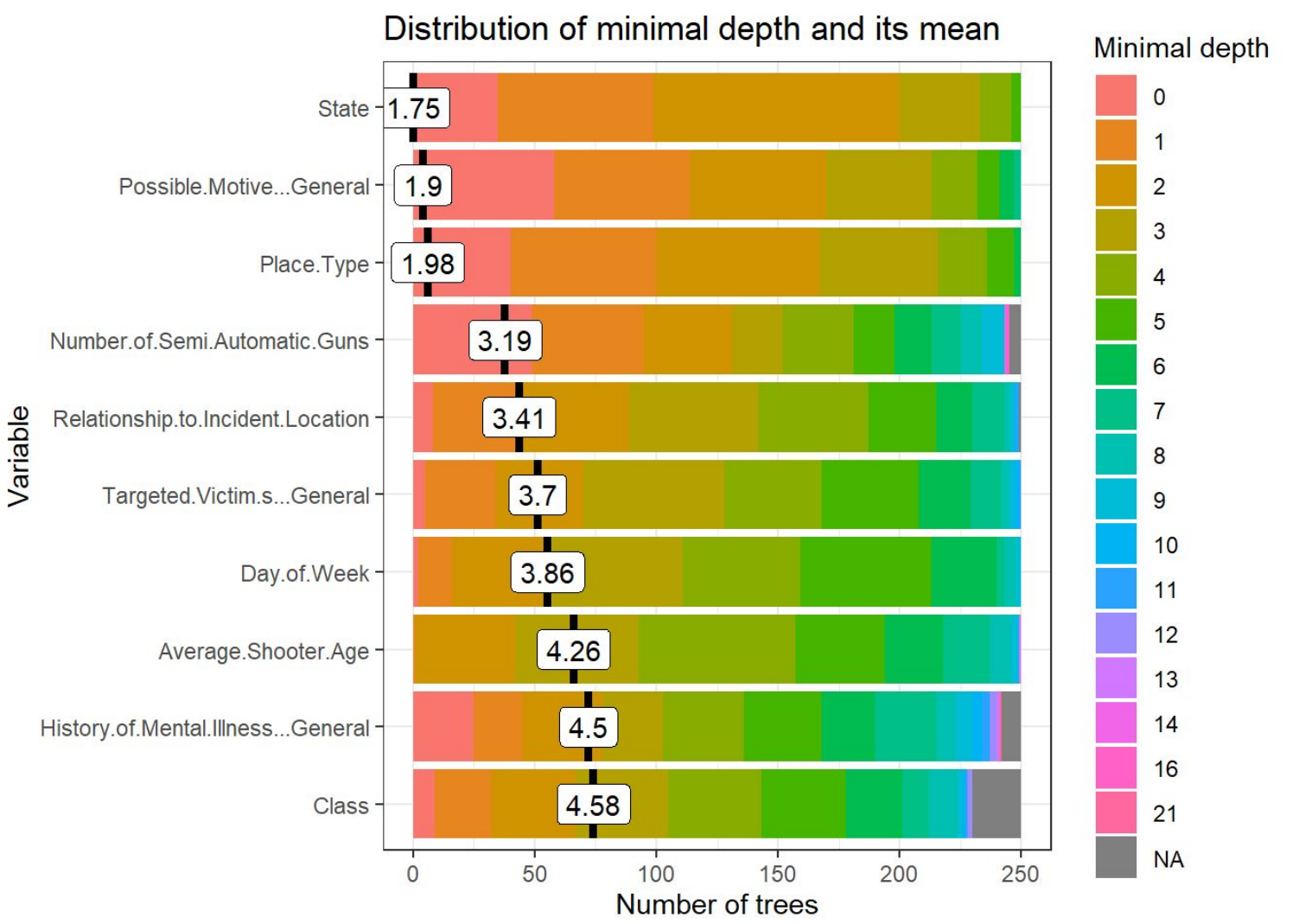
	mtry	OOBError
1	1	36.14507
6	6	36.00280
24	24	40.82990

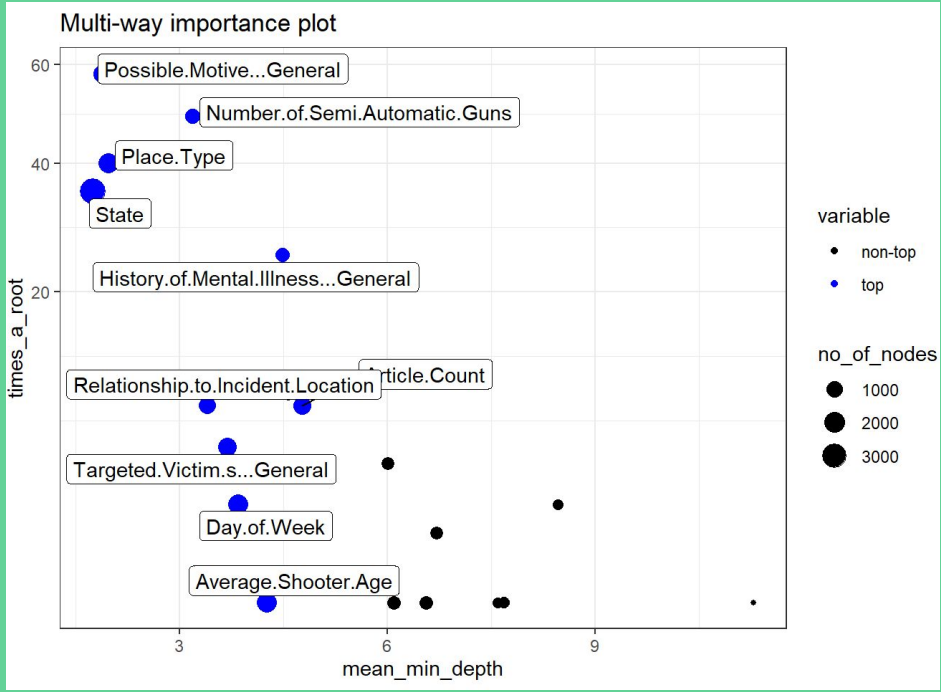


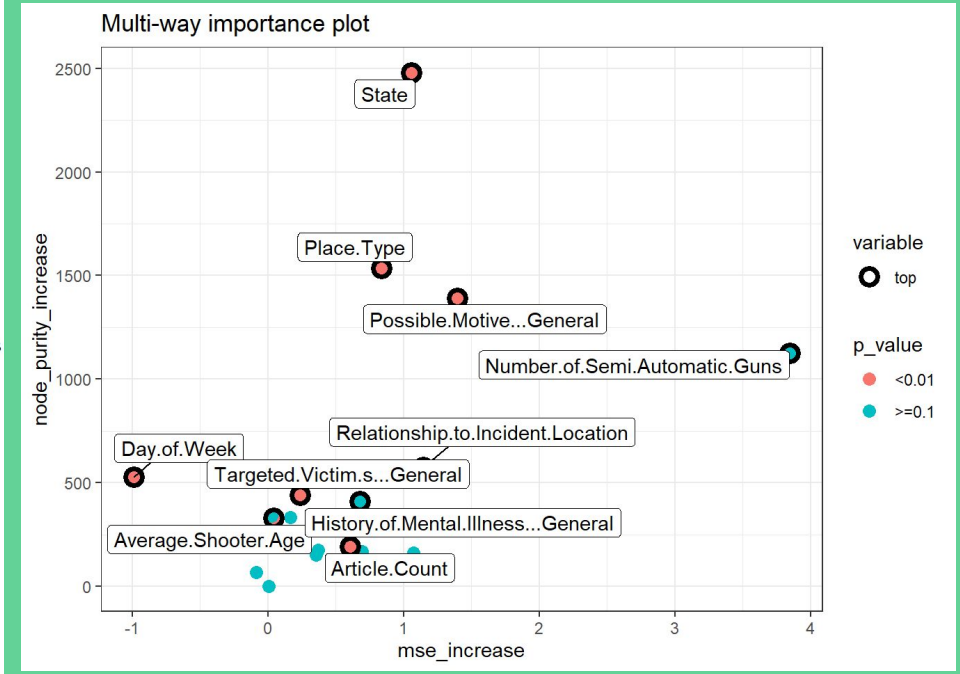
rf_mod


Number.of.Semi.Automatic.Guns
Fate.of.Shooter
Relationship.to.Incident.Location
Number.of.shooters
Possible.Motive...General
Article.Count
Number.of.Automatic.Guns
Shooter.s.Cause.of.Death
History.of.Mental.Illness...General
Place.Type
State
Fate.of.Shooter.at.the.scene
Shooter.Race
Targeted.Victim.s...General
Average.Shooter.Age
Class
Number.of.Shotguns
Military.Experience
Shooter.Sex
Day.of.Week
-2
0
2
4
6
%IncMSE


State
Place.Type
Possible.Motive...General
Number.of.Semi.Automatic.Guns
Relationship.to.Incident.Location
Day.of.Week
Targeted.Victim.s...General
History.of.Mental.Illness...General
Class
Shooter.Race
Average.Shooter.Age
Number.of.Shotguns
Military.Experience
Article.Count
Shooter.s.Cause.of.Death
Number.of.shooters
Fate.of.Shooter
Fate.of.Shooter.at.the.scene
Shooter.Sex
Number.of.Automatic.Guns
0
500
1000
1500
2000
2500
IncNodePurity


Distribution of minimal depth and its mean
Minimal depth
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
21
NA
Variable
State
Possible.Motive...General
Place.Type
Number.of.Semi.Automatic.Guns
Relationship.to.Incident.Location
Targeted.Victim.s...General
Day.of.Week
Average.Shooter.Age
History.of.Mental.Illness...General
Class
1.75
1.9
1.98
3.19
3.41
3.7
3.86
4.26
4.5
4.58
0
50
100
150
200
250
Number of trees





- mean depth of first split on the variable,
- number of trees in which the root is split on the variable,
- the total number of nodes in the forest that split on that variable

- with the additional information on the p-value based on a binomial distribution of the number of nodes split on the variable
- (i.e. if a variable is significant it means that the variable is used for splitting more often than would be the case if the selection was random).

Predictive models used

- K-means - similar characteristics in the different clusters could help us classify the different types of mass shootings that occur

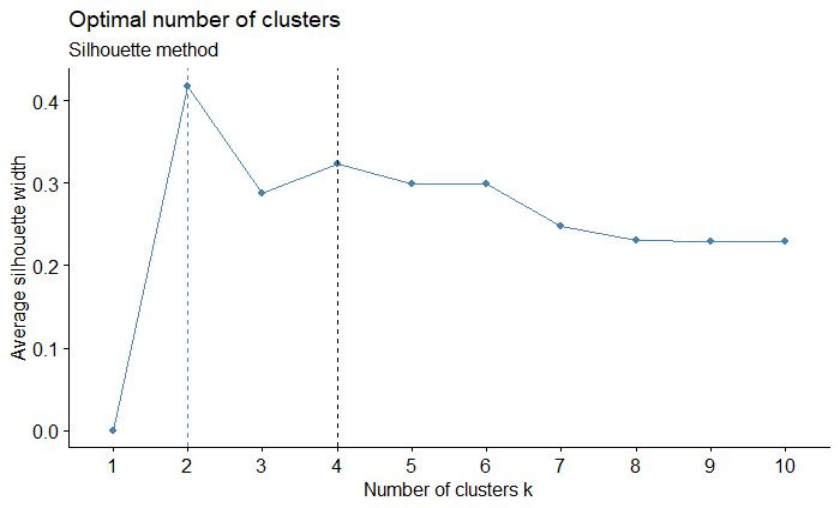
Selecting Variables

- Limited the data to only numeric variables
- Noticed that some of variables shown in the random forest were either insignificant or caused a lot of variance
  - Number of Automatic
  - Number of Semi Auto
  - Number of Shotguns
  - Count

- Variables Used
  - Number of shooters
  - Total Number of Victims
  - Average Age

Selecting K

- Used NbClust and Silhouette to find number of cluster

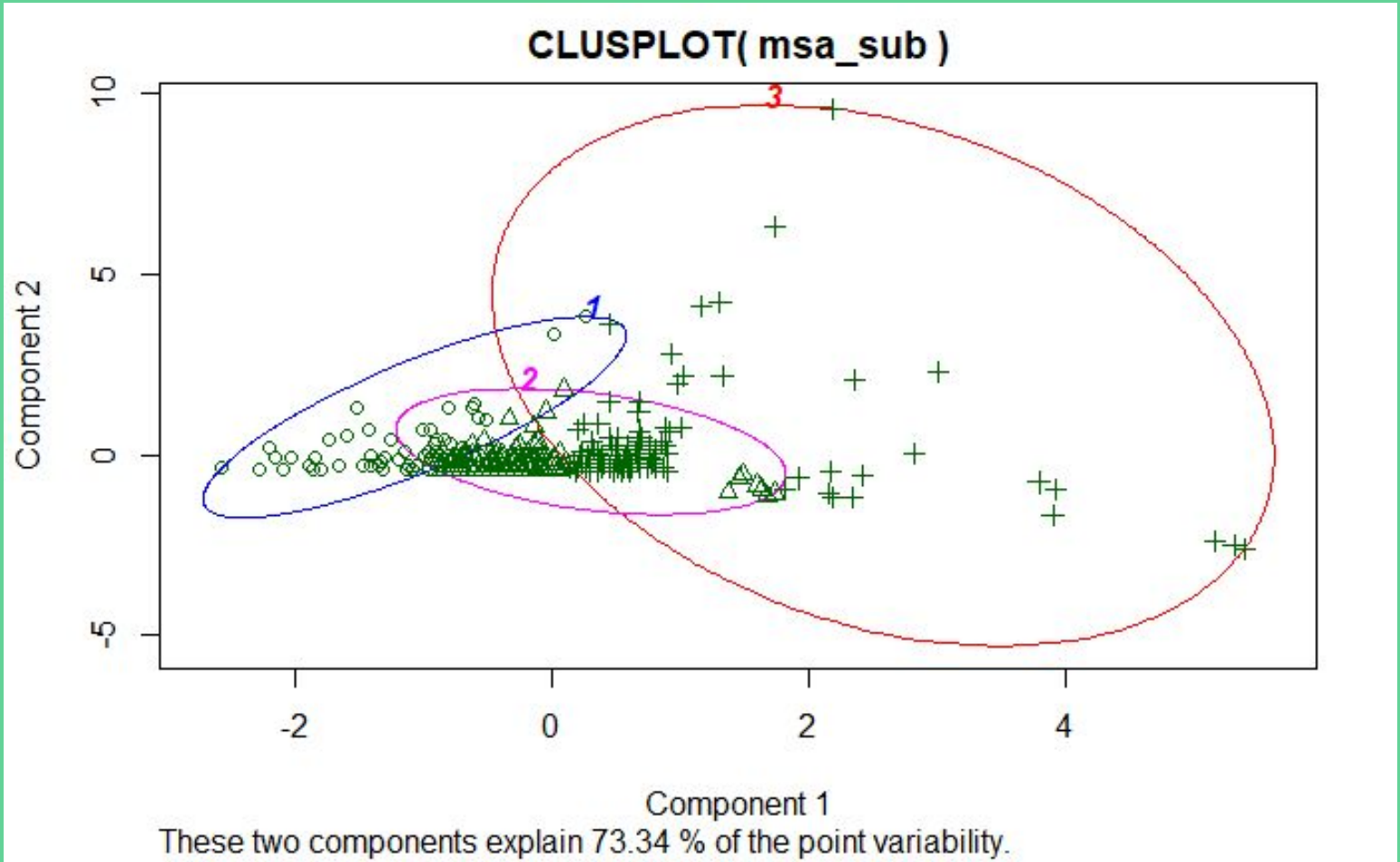



```
* Among all indices:
* 6 proposed 2 as the best number of clusters
* 12 proposed 3 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 11 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 14 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

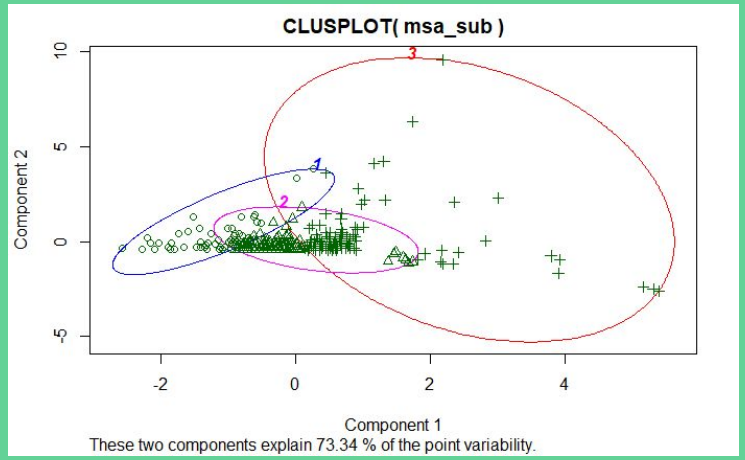
```
Nb_cl <- NbClust(msa_sub,
         diss = NULL,
         distance = "euclidean",
         min.nc = 2,
         max.nc = 21,
         method = "kmeans")
```


CLUSPLOT(msa_sub)
Component 2
10
5
0
-5
-2
0
2
4
Component 1
These two components explain 73.34 % of the point variability.
1
2
3

```
K-means clustering with 3 clusters of sizes 67, 141, 117

Cluster means:
  Total.Number.of.Victims Number.of.shooters Average.Shooter.Age
1                8.492537           1.000000            49.14925
2                5.326241           1.063830            33.41489
3               10.675214           1.213675            19.98675
```

- Even though a majority of the shooters are around 30 years of age that demographic seems to have significantly less victims then the other cases.
- Cluster 3 has very high variability with having a larger number of the outliers potentially skewing the number to be larger
- Despite the potential skew the average seems to generally be larger for cluster 3



Business and practical management

- Mass shootings have become a major problem in the United States
- Useful for organizations to track information on students, employees, etc. to be aware of how many people are at risk in a given situation and what factors they could change to reduce that number
- While not being able to predict whether a shooting will happen or the severity accurately, we have found variables that clearly have a significant impact if a shooting were to occur.
- Addressing these factors can limit the people at risk in a mass shooting:
    - State
    - Place Type
    - Possible Motive
    - Relationship to Incident Location

Thank you!