Gun Violence and Mass Shootings

Tools and Languages Used

- Rapidminer
- R
- Weka
- Java
- Tableau
- Python
- Excel

Related Work

- **Why gun control can help prevent gun violence? (Loria, 2018)**
- Firearm regulations in the U.S.: trend of gun violence. (Bouchet, 2017)
- US Mass Shootings Analysis 1966-2017: do demographics correlate to shooters in mass shootings? (Smith, 2018)
- A Guide to Mass Shootings in America. (Follman et al., 2018)
- **Mental Illness, Mass Shootings, and the Politics of American Firearms. (Metzl et al., 2015)**
- **The impact of the Orlando mass shooting on fear of victimization and gun-purchasing intentions. (Stroebe et al., 2017)**
- Socioeconomic factors and mass shootings in the US. (Kwon et al., 2017)
- Columbine Revisited: Myths and Realities About the Bullying–School Shootings Connection (Mears et al., 2017)

Zip Codes

Business Understanding

The Setting:

- Lack of research in predicting mass shooters and mass shootings
- Mass shootings happen sporadically, but there may be recurring patterns

We want to examine the problem of predicting mass shootings from a data analysis/data mining perspective, answering questions like:

- What cities/states/zip codes are more prone to such attacks?
- Is there any correlation between an area's demographic data (earnings, age, education level), guns licenses, guns manufacturing, etc. and whether or not there has been a mass shooting there?

Datasets Used - Data Understanding

- Firearms Data - ATF (The Bureau of Alcohol, Tobacco, Firearms, and Explosives)
  - Listing of Federal Firearms Licensees
  - Listing of Firearms Manufacturers
  - Challenge: Lack of data
    - ATF Records only from 2014 and onwards
- Census Data - US Census Bureau
  - Demographic Data
- Gun Violence Datasets, Mother Jones, Kaggle + others
  - Data is sparser the further back you go
- As a result, we will only look data from recent years

Data Preparation

Handling missing values

- Replace all missing values with the average value of the feature

Identifying and removing highly correlated features

- Correlation matrix
- Cutoff = 0.75
- Remove features with a correlation coefficient beyond the cutoff

Challenge: Proximity Calculation

- Not only calculate Gun licenses and Manufacturing based on zip codes, but also spread numbers to all nearby zip codes within a 50 mile radius

Data Preparation - Zip Codes

- 33120 observations (rows): each row pertains to a ZIP code
- 91 variables / features (columns):
  - Income per household (2)
  - Demographic data (55)
  - Educational attainment (31)
  - Guns manufacturing (1)
  - Guns licences (1)
  - Guns purchases (1)
- Predicted variable (label): violence
  - Problem type: binary classification
  - 1 = record of at least one gun-related violent incident, 0 = no violent incidents
- Data matrices for three years: 2014, 2015 and 2016

Data Preparation

```
 [1] "X86..Median.age"
 [2] "X82..Population.percentage..60.years.and.over"
 [3] "X21..Total.population.35.to.44.years"
 [4] "X36..Population.total"
 [5] "X57..Population.total"
 [6] "X83..Population.percentage..62.years.and.over"
 [7] "X18..Total.population.25.to.34.years"
 [8] "X10..Total.population.25.years.and.over"
 [9] "X24..Total.population.45.to.64.years"
[10] "X03..Household.median.income.in.dollars"
[11] "X04..Household.mean.income.in.dollars"
[12] "X93..Gun.purchases.approximate"
[13] "X16..Percentage.of.10..bachelor.s"
[14] "X81..Population.percentage..18.years.and.over"
[15] "X26..Percentage.of.24..bachelor.s.or.higher"
[16] "X47..Population.percentage..Hispanic"
[17] "X80..Population.percentage..16.years.and.over"
[18] "X37..Population.percentage..not.Hispanic"
[19] "X76..Population.percentage..5.to.14.years"
[20] "X78..Population.percentage..18.to.24.years"
[21] "X89..Old.age.dependency.ratio"
[22] "X56..Population.percentage..Hispanic.two.or.more.races.excluding.some.other.race"
[23] "X46..Population.percentage..not.Hispanic.two.or.more.races.excluding.some.other.race"
[24] "X58..Population.percentage..under.5.years"
```

Data Preparation - Feature Ranking

Learning Vector Quantization (LVQ) model

- Special case of an artificial neural network
- Competitive learning: the winner neuron has the greatest similarity to the input
- 10-fold cross validation

Variable importance

- The importance of each predictor is evaluated individually
- Method: ROC (receiver operating characteristic) curve analysis
- The sensitivity and specificity are computed for each cutoff
- The area under the ROC curve is used as the measure of variable importance

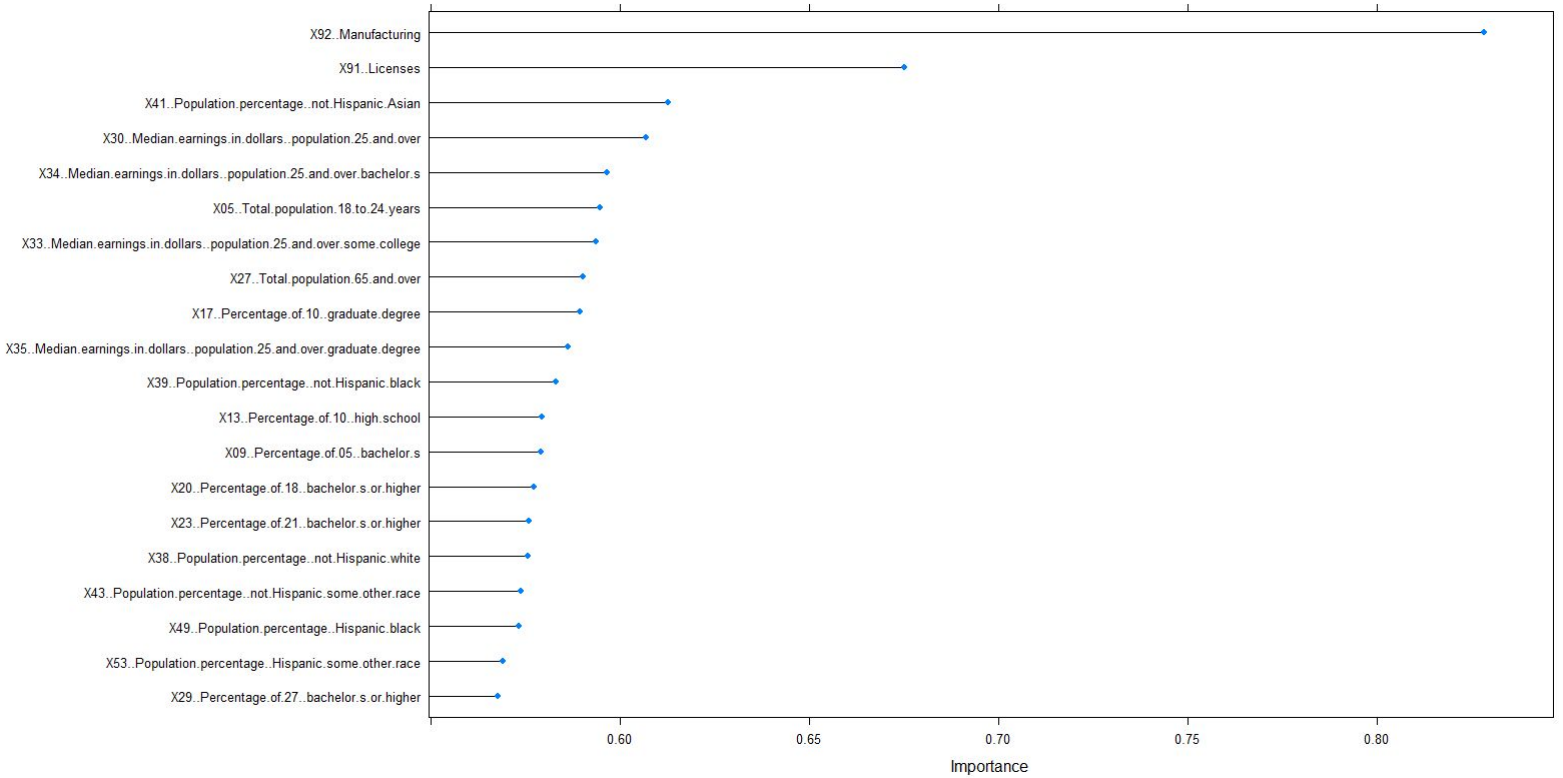
Data Preparation - Feature Ranking

ROC curve variable importance

only 20 most important variables shown (out of 67)

	Importance
X92..Manufacturing	0.8283
X91..Licenses	0.6752
X41..Population.percentage..not.Hispanic.Asian	0.6127
X30..Median.earnings.in.dollars..population.25.and.over	0.6070
X34..Median.earnings.in.dollars..population.25.and.over.bachelor.s	0.5967
X05..Total.population.18.to.24.years	0.5946
X33..Median.earnings.in.dollars..population.25.and.over.some.college	0.5937
X27..Total.population.65.and.over	0.5902
X17..Percentage.of.10..graduate.degree	0.5895
X35..Median.earnings.in.dollars..population.25.and.over.graduate.degree	0.5863
X39..Population.percentage..not.Hispanic.black	0.5831
X13..Percentage.of.10..high.school	0.5794
X09..Percentage.of.05..bachelor.s	0.5790
X20..Percentage.of.18..bachelor.s.or.higher	0.5773
X23..Percentage.of.21..bachelor.s.or.higher	0.5759
X38..Population.percentage..not.Hispanic.white	0.5757
X43..Population.percentage..not.Hispanic.some.other.race	0.5738
X49..Population.percentage..Hispanic.black	0.5733
X53..Population.percentage..Hispanic.some.other.race	0.5690
X29..Percentage.of.27..bachelor.s.or.higher	0.5677

Data Preparation - Feature Ranking



Data Preparation - Feature Ranking

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7599 1436
         1  324  578

               Accuracy : 0.8229
                 95% CI : (0.8152, 0.8303)
    No Information Rate : 0.7973
    P-Value [Acc > NIR] : 6.428e-11

                  Kappa : 0.3099
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9591
            Specificity : 0.2870
         Pos Pred Value : 0.8411
         Neg Pred Value : 0.6408
             Prevalence : 0.7973
         Detection Rate : 0.7647
   Detection Prevalence : 0.9092
      Balanced Accuracy : 0.6230

       'Positive' Class : 0
```

Data Preparation - Feature Ranking

Weka - Ranker + Information Gain Ratio

Across all years, the feature with the highest information gain was gun manufacturing

Other high ranked features included: licenses and approximate # of gun purchases

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 92 VIOLENT?):
        Information Gain Ranking Filter

Ranked attributes:
 0.460891    90 92) Manufacturing
 0.233628    89 91) Licenses
 0.140561    91 93) Gun purchases approximate
 0.021006    12 14) Percentage of 10) some college
 0.020755    28 "30) Median earnings in dollars, population 25 and over"
 0.019041     1 03) Household median income in dollars
 0.01901      2 04) Household mean income in dollars
 0.017089    39 41) Population percentage: not Hispanic Asian
 0.015193    15 17) Percentage of 10) graduate degree
 0.014503    32 "34) Median earnings in dollars, population 25 and over bachelor's"
 0.013599    38 40) Population percentage: not Hispanic American Indian or native
 0.012346    31 "33) Median earnings in dollars, population 25 and over some college"
 0.012342    24 26) Percentage of 24) bachelor's or higher
 0.011418    37 39) Population percentage: not Hispanic black
 0.010221    18 20) Percentage of 18) bachelor's or higher
```

Data Preparation - Feature Ranking

Weka: Ranker + Pearson Correlation w/ Predictive Attribute

Across all years, highest correlated features were gun licenses and manufacturing

The other high-scoring attributes were consistently: Median earnings in dollars, population 25 and over, Household mean/median income in dollars

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 92 VIOLENT?):
        Correlation Ranking Filter
Ranked attributes:
 0.50235    89 91) Licenses
 0.2802     90 92) Manufacturing
 0.16735    28 "30) Median earnings in dollars, population 25 and over"
 0.15501     1 03) Household median income in dollars
 0.15193    32 "34) Median earnings in dollars, population 25 and over bachelor's"
 0.14835     2 04) Household mean income in dollars
 0.13479    12 14) Percentage of 10) some college
 0.13179    33 "35) Median earnings in dollars, population 25 and over graduate degree"
 0.12941    39 41) Population percentage: not Hispanic Asian
 0.12856    31 "33) Median earnings in dollars, population 25 and over some college"
 0.12642    15 17) Percentage of 10) graduate degree
 0.11196    24 26) Percentage of 24) bachelor's or higher
 0.10803    18 20) Percentage of 18) bachelor's or higher
 0.10292    91 93) Gun purchases approximate
 0.10127    21 23) Percentage of 21) bachelor's or higher
 0.09644    22 24) Total population 45 to 64 years
 0.09364    25 27) Total population 65 and over
 0.0917      8 10) Total population 25 years and over
 0.0915     30 "32) Median earnings in dollars, population 25 and over high school"
```

Data Preparation - Feature Selection

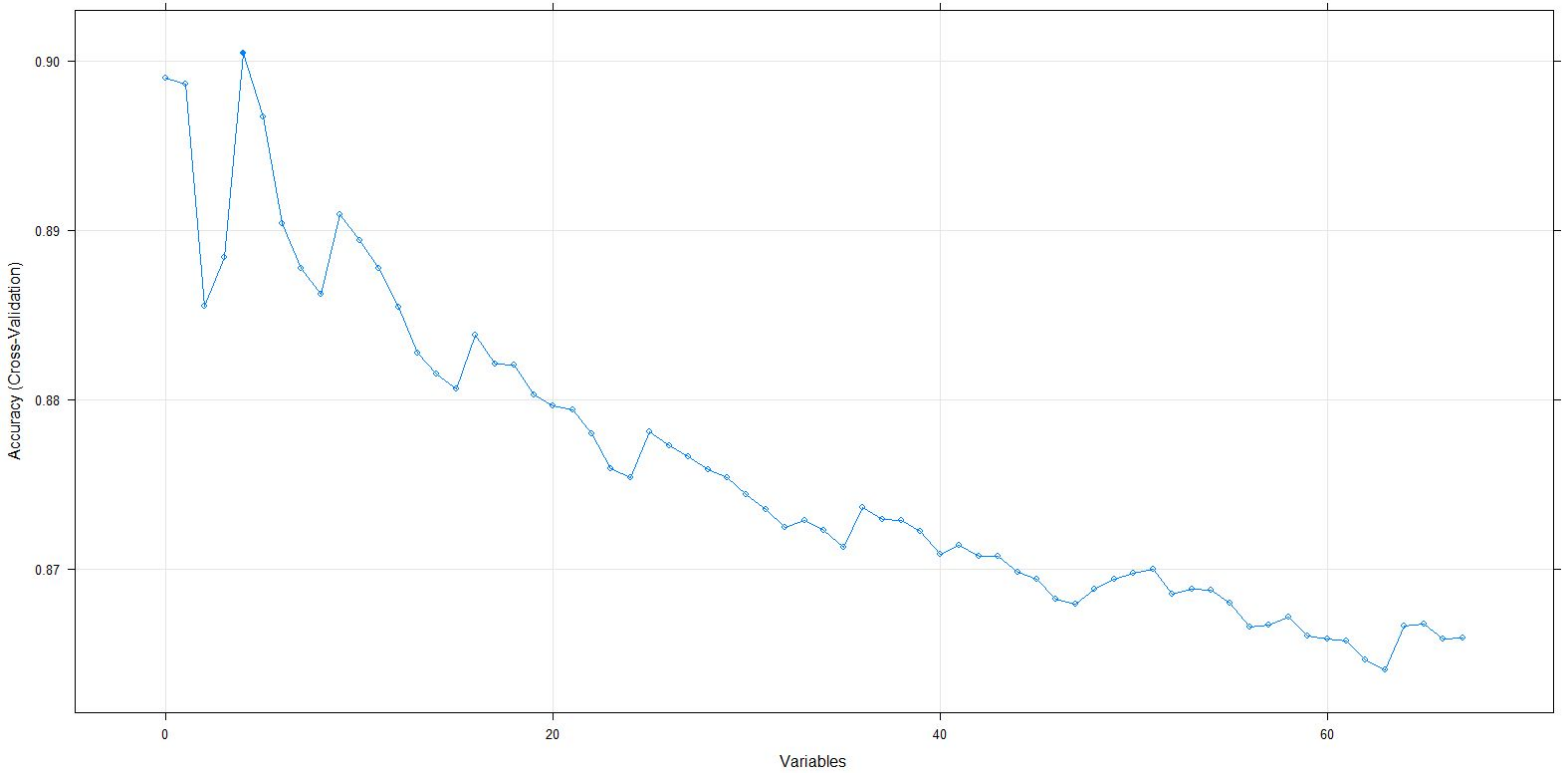
Recursive Feature Elimination (RFE)

- 10-fold cross validation
- Backwards selection algorithm
- First, the algorithm fits the model to all predictors.
- Each predictor is ranked using its importance to the model.
- At each iteration of feature selection, only top ranked predictors are retained, the model is refit and performance is assessed.
- The number of predictors with the best performance is determined and they are used to fit the final model.

Data Preparation - Feature Selection

```
> print(predictors2014)
[1] "X92..Manufacturing"
[2] "X91..Licenses"
[3] "X38..Population.percentage..not.Hispanic.white"
[4] "X41..Population.percentage..not.Hispanic.Asian"
```

Data Preparation - Feature Selection

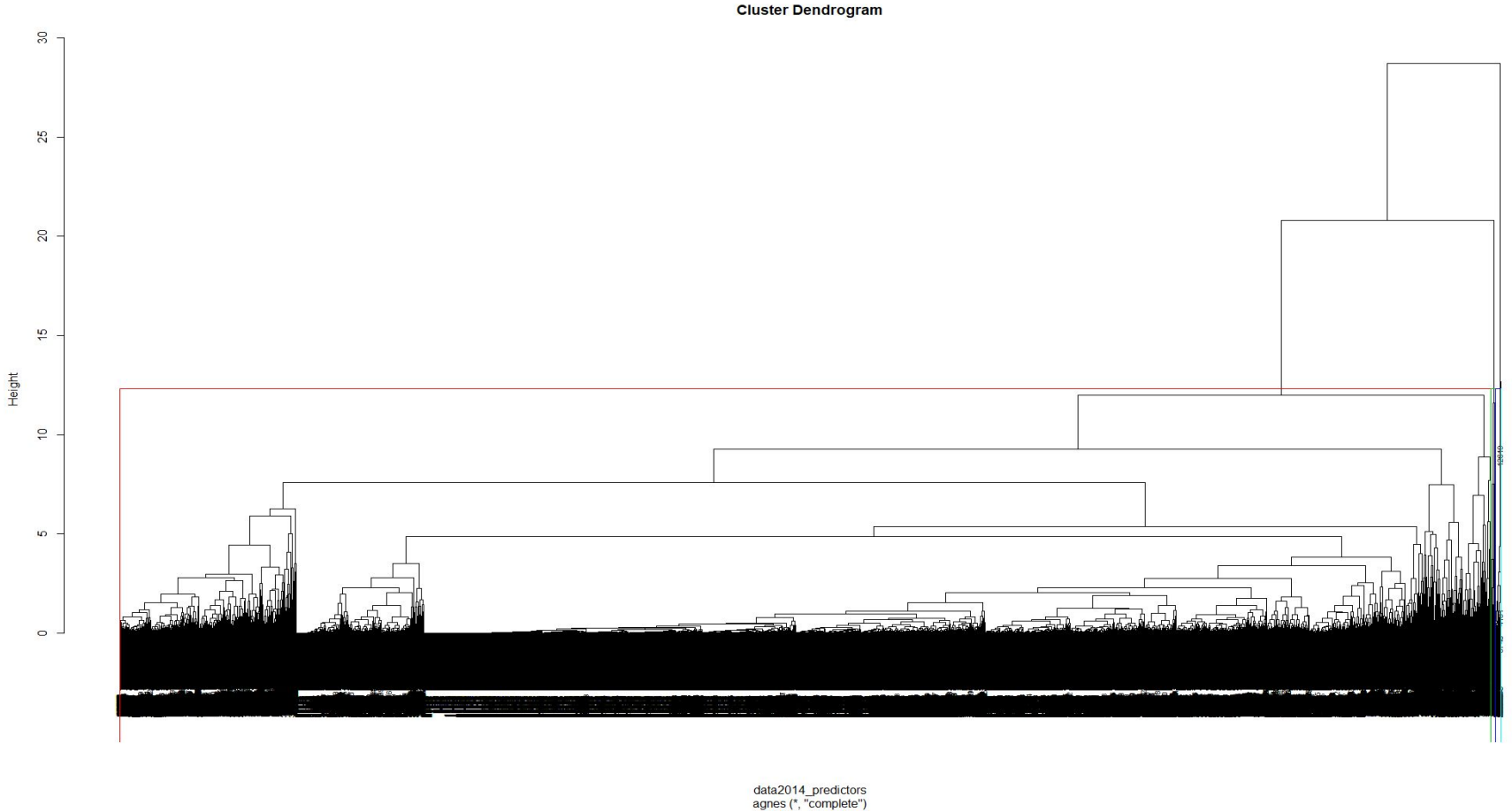


Modeling - Hierarchical Clustering

Agnes (Agglomerative Nesting)

- Hierarchical clustering algorithm
- Groups the data one by one on the basis of the nearest distance measure of all the pairwise distance between the data points
- It yields the agglomerative coefficient, which measures the amount of clustering structure found

Modeling - Hierarchical Clustering



Modeling - Hierarchical Clustering

Cluster assignment:

	X00..Zip.Code	Cluster
1	29634	1
2	71377	1
3	97414	1
4	84112	1
5	86433	1
6	25203	1
7	98859	1
8	48411	1
9	4852	1
10	87749	1

Agglomerative coefficient:

```
> agnes2014$ac
[1] 0.9986162
```

Examples per cluster:

```
> table(clusters2014)
clusters2014
    1     2     3     4
32891   129    99     1
```

Clusters vs. Labels:

	0	1
1	21451	10509
2	6	296
3	14	843
4	0	1

Since clusters 2 and 3 have a vast majority of 1s (zip codes that have experienced violent attacks), this may point to the fact that they have some similarities with each other

Modeling - Classification

Adaboost using 5-fold cross validation, # of weak learners/iterations determined by grid search, using decision stumps, and testing on a hold-out set consisting of 20% of our data

Using all features, our precision was around ~86%

Using only the best features, our precision showed no noticeable difference

Modeling - Classification

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.meta.AdaBoostM1
Cross-validation Parameter: '-I' ranged from 100.0 to 500.0 with 5.0 steps
Classifier Options: -I 500 -P 100 -S 1 -W weka.classifiers.trees.DecisionStump

Correctly Classified Instances        5667               85.5525 %
Incorrectly Classified Instances       957               14.4475 %
Kappa statistic                          0.4625
Mean absolute error                      0.2101
Root mean squared error                  0.3193
Relative absolute error                 65.3828 %
Root relative squared error             80.0666 %
Total Number of Instances             6624

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.961    0.570    0.872      0.961    0.914      0.485    0.880     0.962     0
                 0.430    0.039    0.731      0.430    0.541      0.485    0.880     0.668     1
Weighted Avg.    0.856    0.465    0.844      0.856    0.840      0.485    0.880     0.904

=== Confusion Matrix ===

    a    b   <-- classified as
 5102  208 |    a = 0
  749  565 |    b = 1
```

Modeling - Classification

Neural Nets, running grid search on the # of hidden layers with 5-fold cross validation (for quicker evaluation times), and testing on a hold-out set of comprised of 20% of our data

We ran this model using all features, hoping that neural nets would naturally select the best ones

Our precision was around ~80% as well

Modeling - Classification

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.functions.MultilayerPerceptron
Cross-validation Parameter: '-H' ranged from 1.0 to 4.0 with 4.0 steps
Classifier Options: -H 4 -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20

Correctly Classified Instances        5481               82.7446 %
Incorrectly Classified Instances      1143               17.2554 %
Kappa statistic                          0.3226
Mean absolute error                      0.2418
Root mean squared error                  0.3561
Relative absolute error                 75.2786 %
Root relative squared error             89.2903 %
Total Number of Instances             6624

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.699    0.847      0.958    0.899      0.354    0.790     0.920     0
                 0.301    0.042    0.638      0.301    0.409      0.354    0.790     0.533     1
Weighted Avg.    0.827    0.569    0.806      0.827    0.802      0.354    0.790     0.844

=== Confusion Matrix ===

    a    b   <-- classified as
 5086  224 |    a = 0
  919  395 |    b = 1
```

Modeling - Classification

Decision Trees (REP-Tree) 10-fold cross validation, max depth determined by grid search, final precision obtained by testing on a hold-out test containing 20% of our data

Using all features, our precision was ~88%

Using only the best features, our precision was again ~88%

Tree Structure: The first attribute to split on in the tree is always Manufacturing and/or Licenses, followed by either Population percentage: not Hispanic Asian, Population percentage: not Hispanic Black, and/or Gun purchases approximate

Modeling - Classification

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.trees.REPTree
Cross-validation Parameter: '-L' ranged from 10.0 to 100.0 with 10.0 steps
Classifier Options: -L 20 -M 2 -V 0.001 -N 3 -S 1 -I 0.0

Correctly Classified Instances        5772               87.1377 %
Incorrectly Classified Instances       852               12.8623 %
Kappa statistic                          0.5592
Mean absolute error                      0.1685
Root mean squared error                  0.3061
Relative absolute error                 52.4613 %
Root relative squared error             76.7598 %
Total Number of Instances             6624

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.947    0.434    0.898      0.947    0.922      0.566    0.904     0.970     0
               0.566    0.053    0.725      0.566    0.636      0.566    0.904     0.713     1
Weighted Avg.  0.871    0.358    0.864      0.871    0.865      0.566    0.904     0.919

=== Confusion Matrix ===

    a    b   <-- classified as
 5028  282 |    a = 0
  570  744 |    b = 1
```

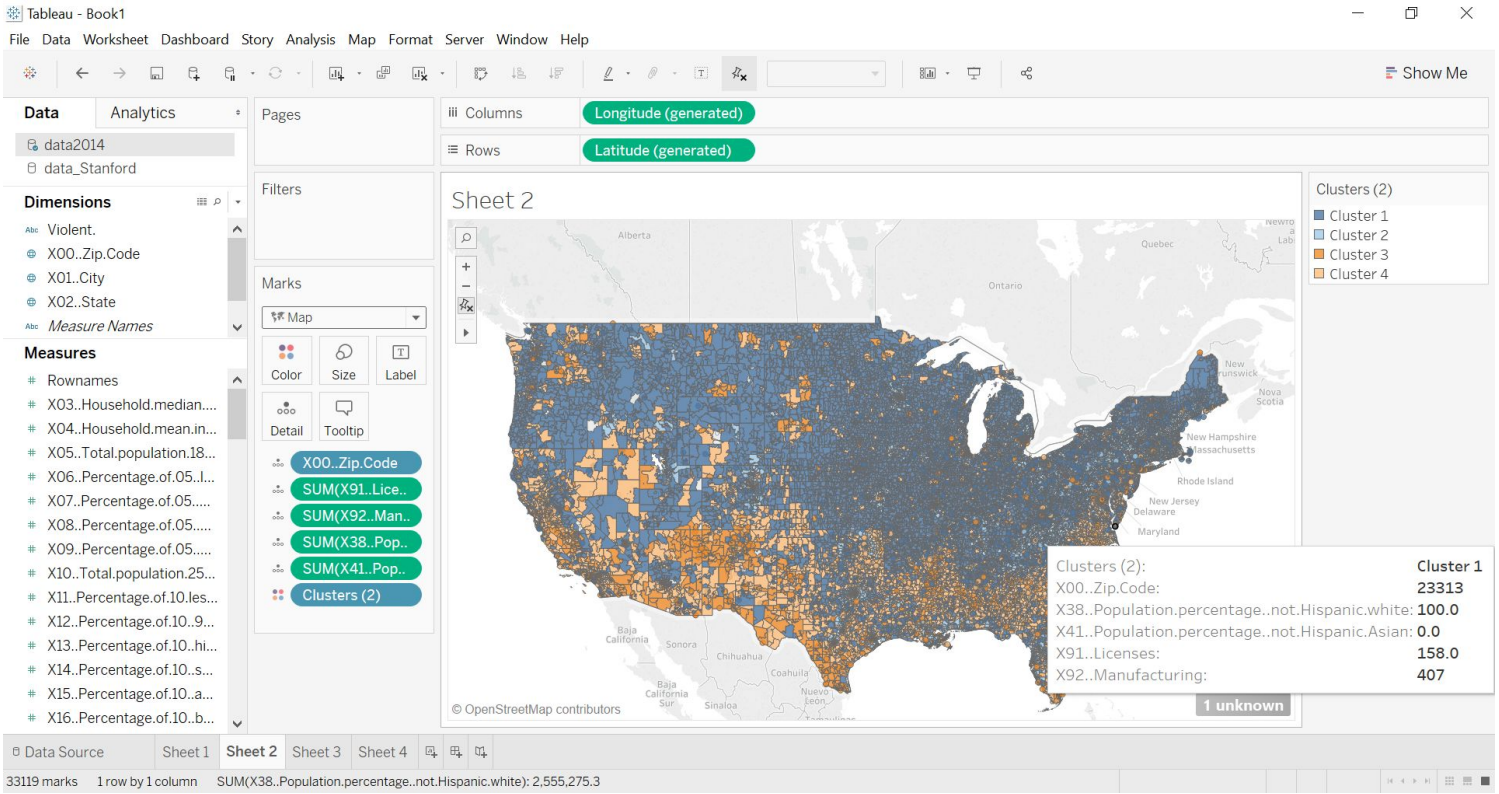
Modeling - Classification

```
92) Manufacturing < 0.5
|   91) Licenses < 54.5 : 0 (11141/159) [5631/77]
|   91) Licenses >= 54.5
|   |   93) Gun purchases approximate < 281011.2
|   |   |   39) Population percentage: not Hispanic black < 4.34
|   |   |   |   76) Population percentage: 5 to 14 years < 14.35
|   |   |   |   |   91) Licenses < 108.5
|   |   |   |   |   |   "33) Median earnings in dollars, population 25 and over some colleg
|   |   |   |   |   |   "33) Median earnings in dollars, population 25 and over some colleg
|   |   |   |   |   91) Licenses >= 108.5 : 0 (48/10) [27/3]
|   |   |   |   76) Population percentage: 5 to 14 years >= 14.35 : 0 (66.27/0) [31/4]
|   |   |   39) Population percentage: not Hispanic black >= 4.34
|   |   |   |   93) Gun purchases approximate < 159189.45 : 0 (23/0) [17/1]
|   |   |   |   93) Gun purchases approximate >= 159189.45
|   |   |   |   |   55) Population percentage: Hispanic two or more races including some ot
|   |   |   |   |   |   28) Percentage of 27) high school or higher < 69.75
|   |   |   |   |   |   |   91) Licenses < 82 : 0 (8/0) [4/1]
|   |   |   |   |   |   |   91) Licenses >= 82 : 1 (12/4) [3/1]
|   |   |   |   |   |   28) Percentage of 27) high school or higher >= 69.75 : 1 (10/0) [7/
|   |   |   |   |   55) Population percentage: Hispanic two or more races including some ot
|   |   93) Gun purchases approximate >= 281011.2
|   |   |   93) Gun purchases approximate < 432082.6
|   |   |   |   93) Gun purchases approximate < 409078.95
|   |   |   |   |   73) Population percentage: 75 to 79 years < 2.05
|   |   |   |   |   |   72) Population percentage: 70 to 74 years < 2.5
|   |   |   |   |   |   |   13) Percentage of 10) high school < 40.8 : 1 (10/3) [5/1]
|   |   |   |   |   |   |   13) Percentage of 10) high school >= 40.8 : 0 (5/0) [1/0]
|   |   |   |   |   |   72) Population percentage: 70 to 74 years >= 2.5 : 1 (12/0) [13/6]
|   |   |   |   |   73) Population percentage: 75 to 79 years >= 2.05 : 0 (96/30) [51/15]
|   |   |   |   93) Gun purchases approximate >= 409078.95 : 1 (22/2) [4/0]
|   |   |   93) Gun purchases approximate >= 432082.6 : 0 (67/8) [41/3]
92) Manufacturing >= 0.5
|   41) Population percentage: not Hispanic Asian < 1.71
|   |   91) Licenses < 132.5
```

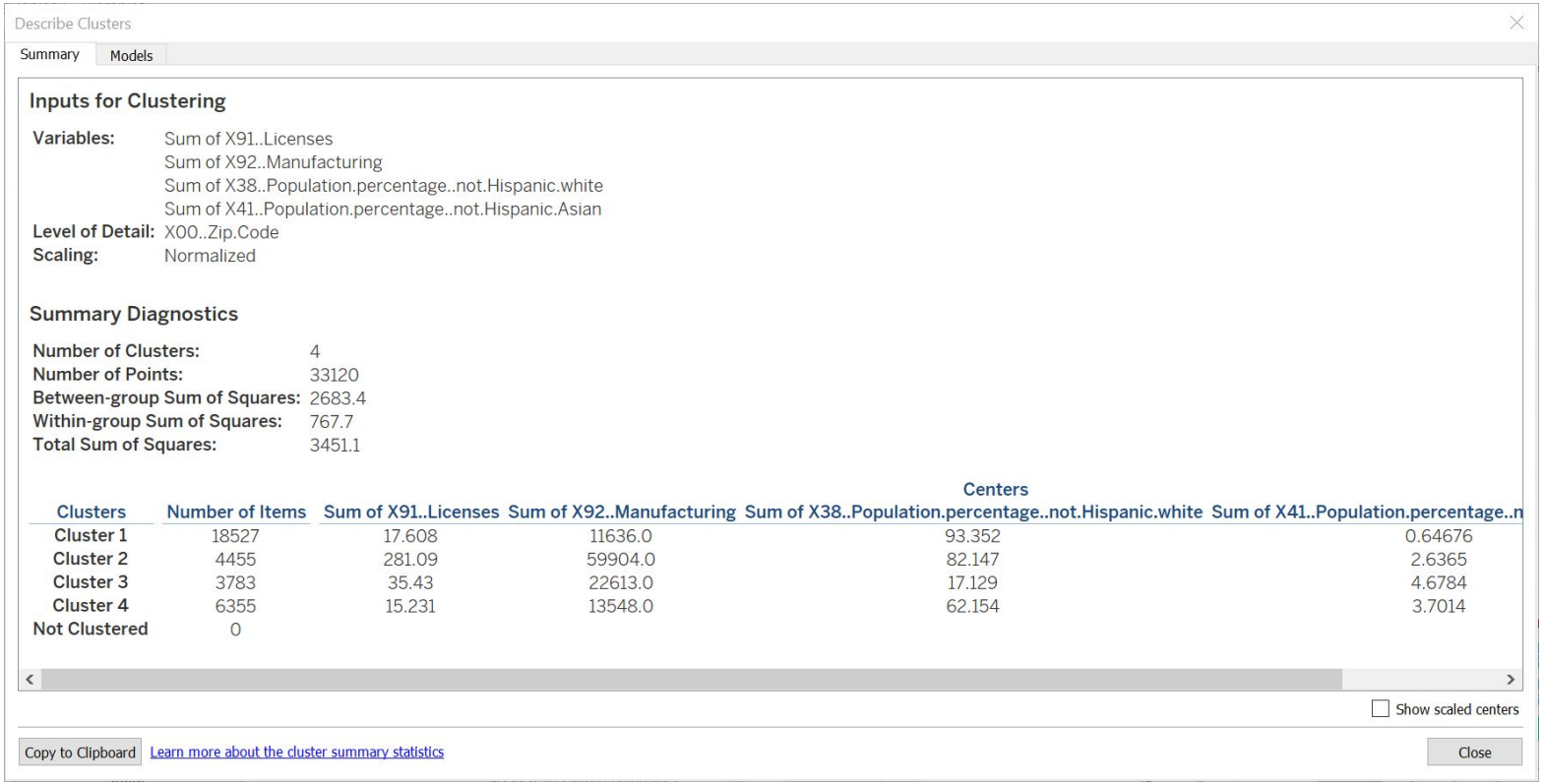
Data Visualization with Tableau

- Tableau is a business intelligence (BI) tool that helps create beautiful and visually-appealing reports, charts, graphs and dashboards using data.
- These reports are interactive and can easily be shared with anyone.
- "Visual Analytics" application: used not only to visualize data, but also to conduct analysis through seeing the data in visuals.
- Unlike other visualization tools, where the dashboard or a graph is the endpoint, Tableau leverages the visual process to develop better understanding of the data.

Data Visualization with Tableau



Data Visualization with Tableau



Conclusions

- What cities/states/zip codes are more prone to such attacks?
- Is there any correlation with demographic data, (earnings, age, education level), guns licenses, guns manufactured, etc. ?

Our results show that places have more distributors and manufacturers of guns are much more prone to violent attacks

Our results through hierarchical clustering show that there are definitely some similarities between certain clusters of zip codes that have experienced violent attacks

In the classification scenario, using decision trees produced the best results, with 88% accuracy

Conclusions

With the data available to us now, we believe that it easy to do a statistical analysis of past events, but predicting future events is very difficult

But we hope that we've exposed some common underlying threads

Stanford MSA Dataset

Business Understanding

- Stanford Mass Shootings in America Project
- Began in 2012 at Stanford, in reaction to the mass shooting in Sandy Hook, CT
- Set out to create a single point repository for as many mass shooting events as could be collected via online media
- Attempt to facilitate research on gun violence in the US by making raw data more accessible

Questions:

- Are there characteristics that make a shooting deadlier than others?
- What role does mental health play?
- Is there any correlation between shooters and race, gender, income, educational attainment, etc.?

Dataset - Data Understanding

- 336 observations (rows): each row pertains to a mass shooting event
- 26 variables / features (columns):
  - # Guns
  - Location
  - Mental Illness (yes or no)
  - Age, Sex
  - Military experience (yes or no)
- Predicted variable (label):
  - Problem type: binary classification
  - 1 = this incident was particularly deadly (# deaths was over median), 0 = all other incidents

Data Preparation

- Highly correlated features:

```
Targeted.Victim.s...General
Total.Number.of.Guns
Number.of.Semi.Automatic.Guns
Fate.of.Shooter
```

- Confusion matrix:

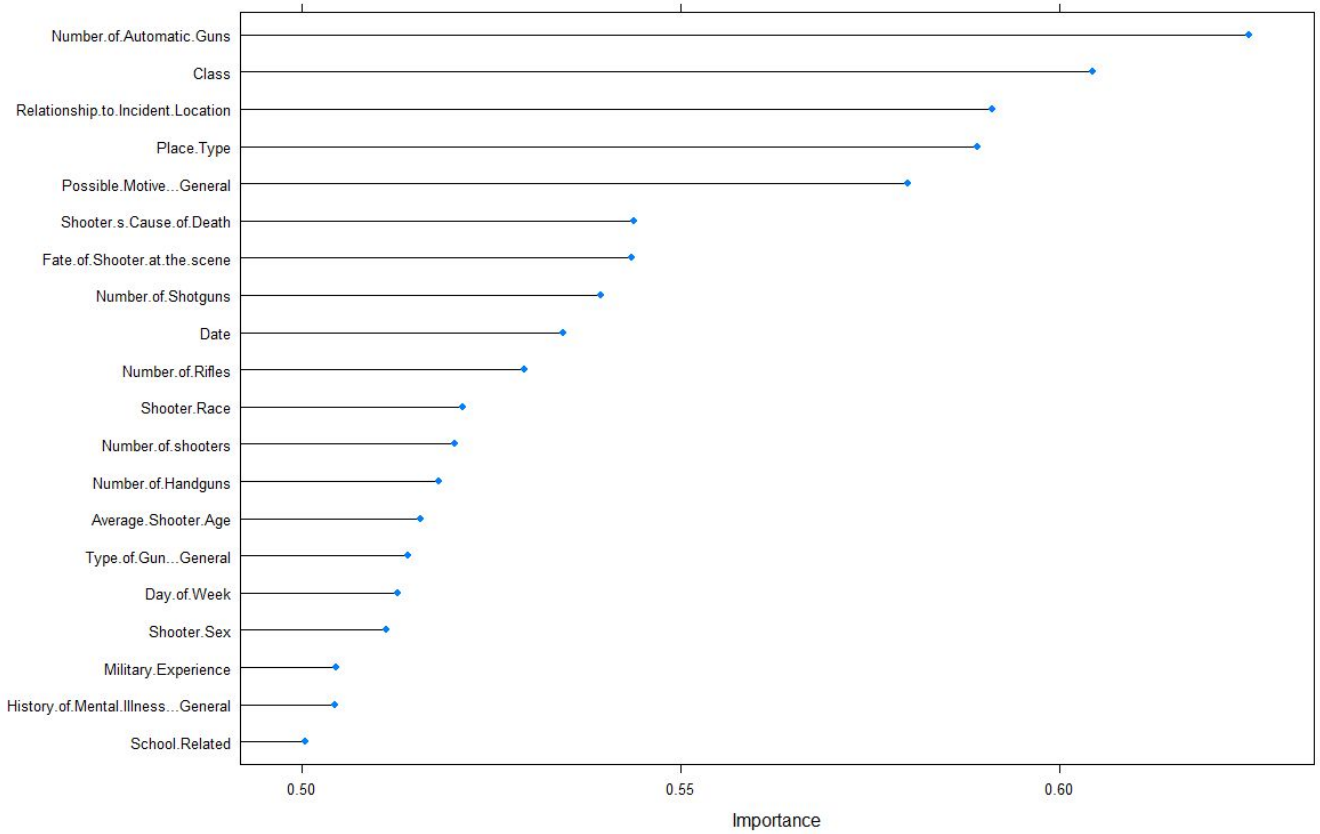
```
          Reference
Prediction  0  1
         0 30 13
         1 15 43

               Accuracy : 0.7228
                 95% CI : (0.6248, 0.8072)
    No Information Rate : 0.5545
    P-Value [Acc > NIR] : 0.0003744

                  Kappa : 0.4364
 Mcnemar's Test P-Value : 0.8501067

            Sensitivity : 0.6667
            Specificity : 0.7679
         Pos Pred Value : 0.6977
         Neg Pred Value : 0.7414
             Prevalence : 0.4455
         Detection Rate : 0.2970
   Detection Prevalence : 0.4257
      Balanced Accuracy : 0.7173

       'Positive' Class : 0
```

Data Preparation

- Feature Ranking:

	Importance
Number.of.Automatic.Guns	0.6250
Class	0.6044
Relationship.to.Incident.Location	0.5911
Place.Type	0.5892
Possible.Motive...General	0.5800
Shooter.s.Cause.of.Death	0.5439
Fate.of.Shooter.at.the.scene	0.5436
Number.of.Shotguns	0.5394
Date	0.5344
Number.of.Rifles	0.5294
Shooter.Race	0.5212
Number.of.shooters	0.5201
Number.of.Handguns	0.5181
Average.Shooter.Age	0.5157
Type.of.Gun...General	0.5140
Day.of.Week	0.5126
Shooter.Sex	0.5111
Military.Experience	0.5045
History.of.Mental.Illness...General	0.5044
School.Related	0.5005

Data Preparation



Data Preparation

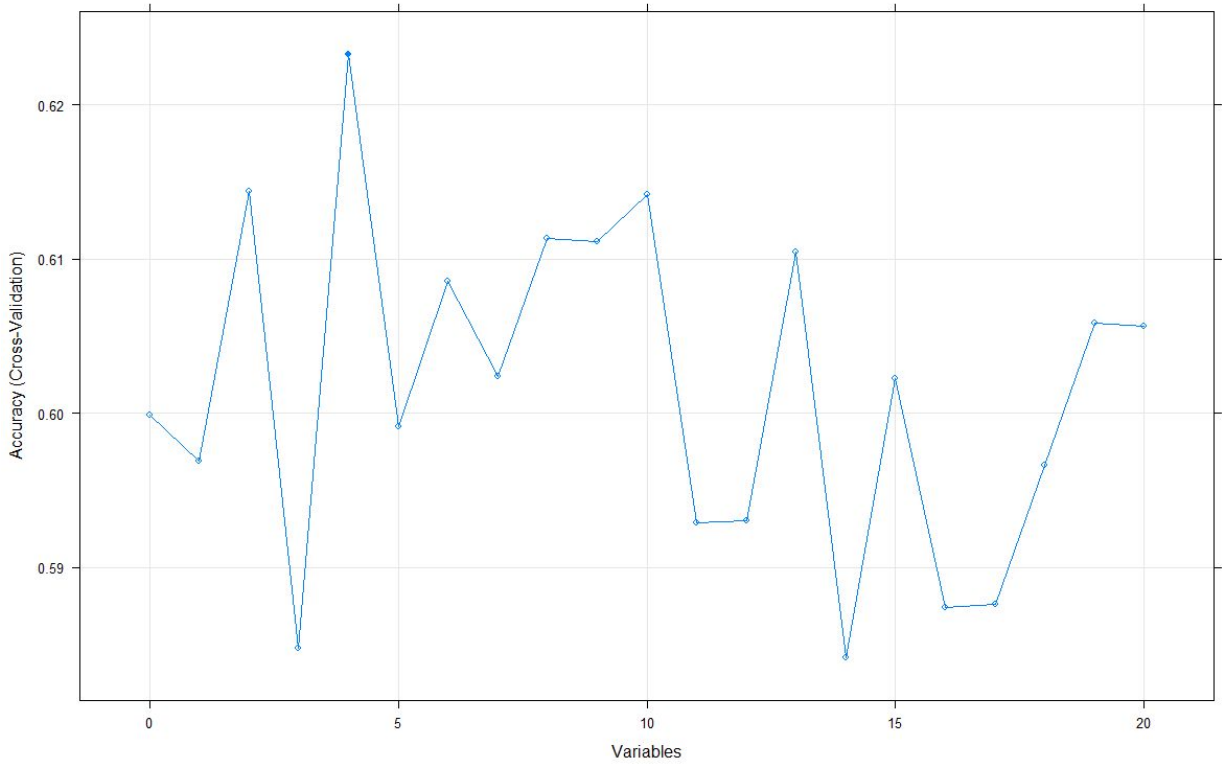
Feature Ranking using Pearson Correlation with predictive attribute/label (Weka)

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 27 LABEL):
        Correlation Ranking Filter
Ranked attributes:
 0.25888   14 Number of Automatic Guns
 0.18509   24 History of Mental Illness - General
 0.17093   15 Number of Semi-Automatic Guns
 0.14701   10 Number of Shotguns
 0.13299   13 Total Number of Guns
 0.12254    9 Type of Gun - General
 0.11677   21 Relationship to Incident Location
 0.11086   11 Number of Rifles
 0.10597   23 Possible Motive - General
 0.10137   20 Place Type
 0.09932   22 Targeted Victim/s - General
 0.09702   18 Shooter's Cause of Death
 0.08815   12 Number of Handguns
 0.07265   26 Class
 0.06847    8 Shooter Race
 0.05899    2 State
 0.05434    5 Number of shooters
 0.05339    7 Shooter Sex
 0.0519     1 City
 0.0514     3 Date
```

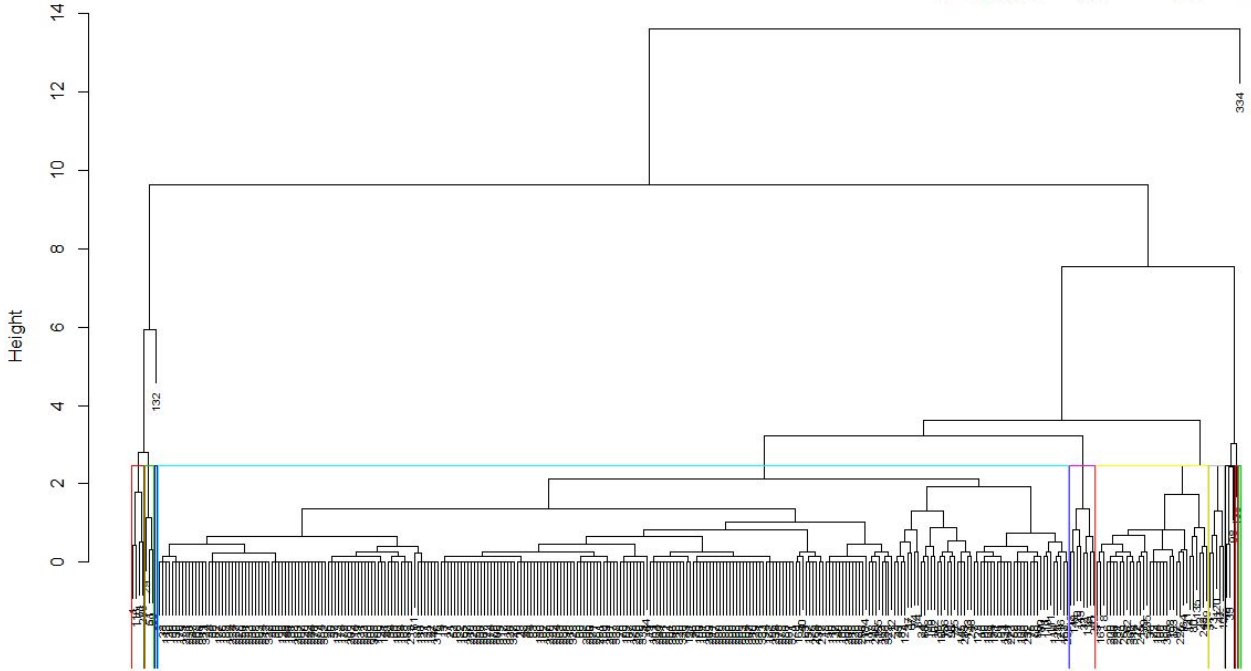
Data Preparation

- Feature Selection:

```
Class
Number.of.Handguns
Number.of.Automatic.Guns
Relationship.to.Incident.Location
```



Modeling - Hierarchical Clustering



Examples per cluster:

```
clusters_Stanford
  1   2   3   4   5   6   7   8   9  10
  4 275  34   8   3   3   5   1   1   1
```

Clusters vs. Labels:

	0	1
1	0	4
2	149	126
3	0	34
4	0	8
5	0	3
6	0	3
7	0	5
8	0	1
9	0	1
10	0	1

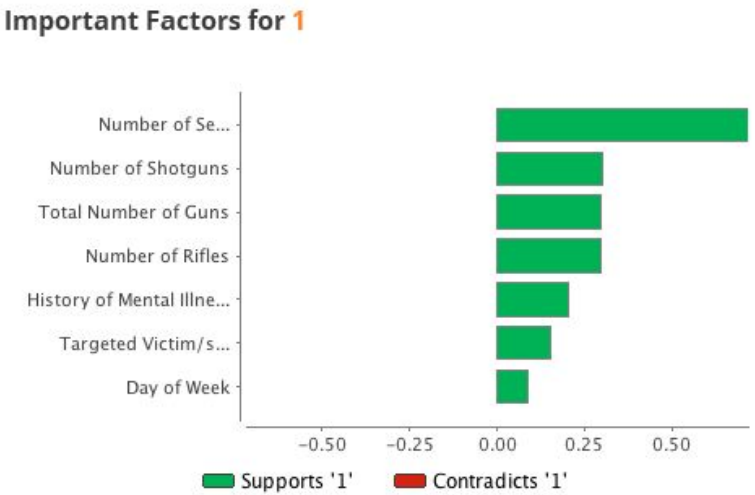
Modeling

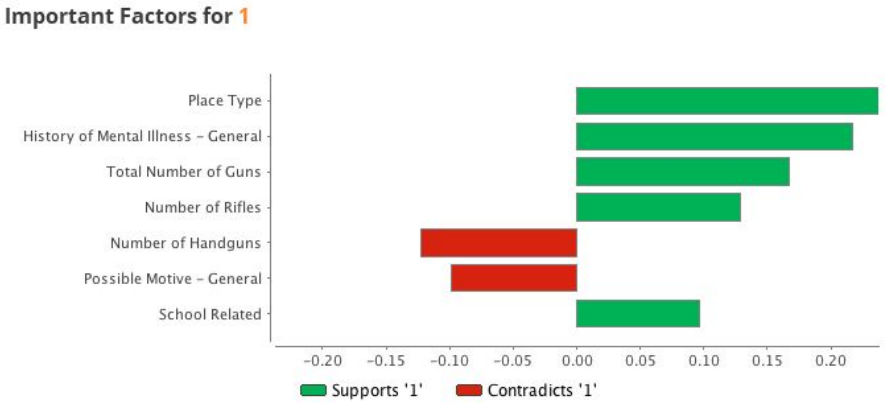
```
Cross-validated Parameter selection.
Classifier: weka.classifiers.meta.AdaBoostM1
Cross-validation Parameter: '-I' ranged from 100.0 to 500.0 with 51.0 steps
Classifier Options: -I 444 -P 100 -S 1 -W weka.classifiers.trees.DecisionStump
```

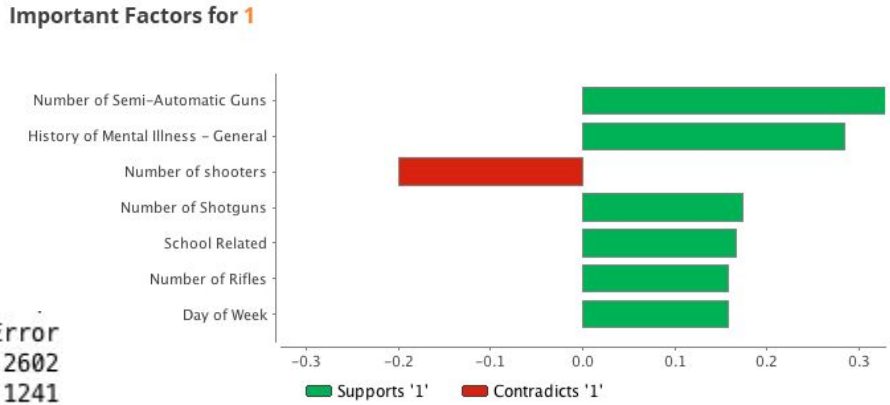
```
Correctly Classified Instances          42               62.6866 %
Incorrectly Classified Instances        25               37.3134 %
Kappa statistic                          0.2445
Mean absolute error                      0.4394
Root mean squared error                  0.4825
Relative absolute error                 89.2915 %
Root relative squared error             97.5972 %
Total Number of Instances               67

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.607    0.359    0.548      0.607    0.576      0.245    0.663     0.615     0
                 0.641    0.393    0.694      0.641    0.667      0.245    0.663     0.729     1
Weighted Avg.    0.627    0.379    0.633      0.627    0.629      0.245    0.663     0.682

=== Confusion Matrix ===

  a  b   <-- classified as
 17 11 |  a = 0
 14 25 |  b = 1
```

Modeling - RapidMiner. MSA - Weights

Right: Logistic Regression (Top) and Deep Learning
Left: Generalized Linear Model







Confusion Matrix for Deep Learning

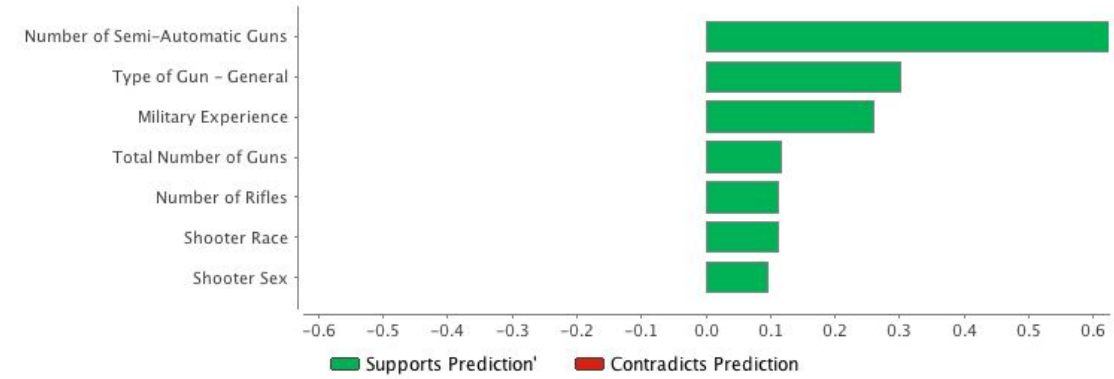
	0	1	Error
0	91	32	0.2602
1	18	127	0.1241

Prediction Modeling - Random Forests

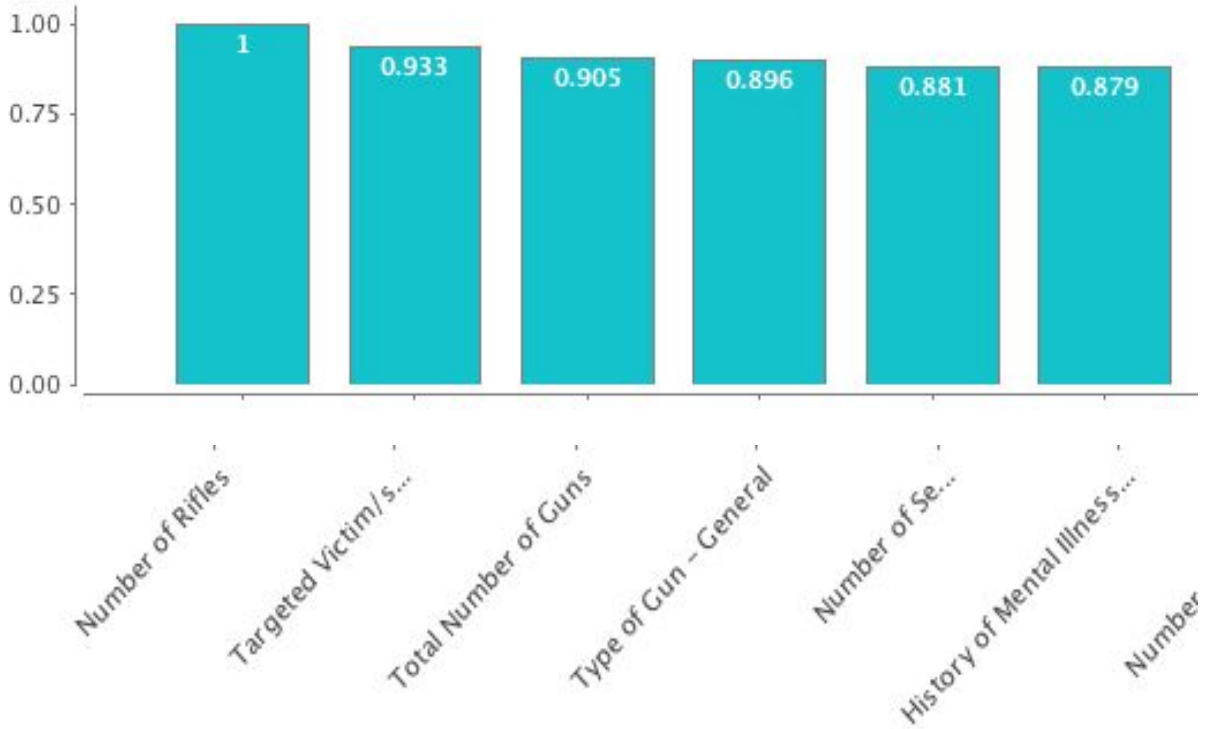
Prediction

42.391

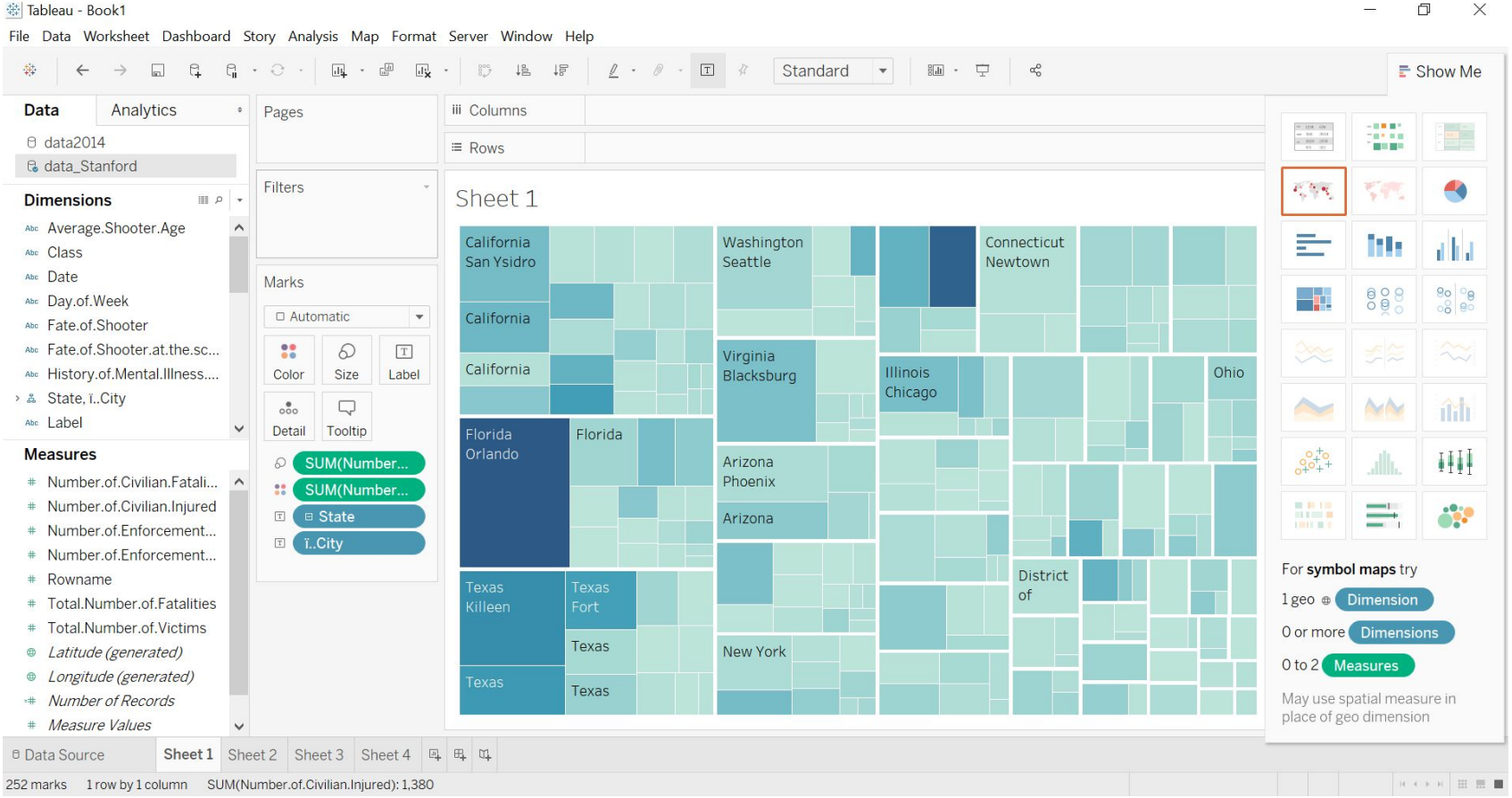
Important Factors for Prediction



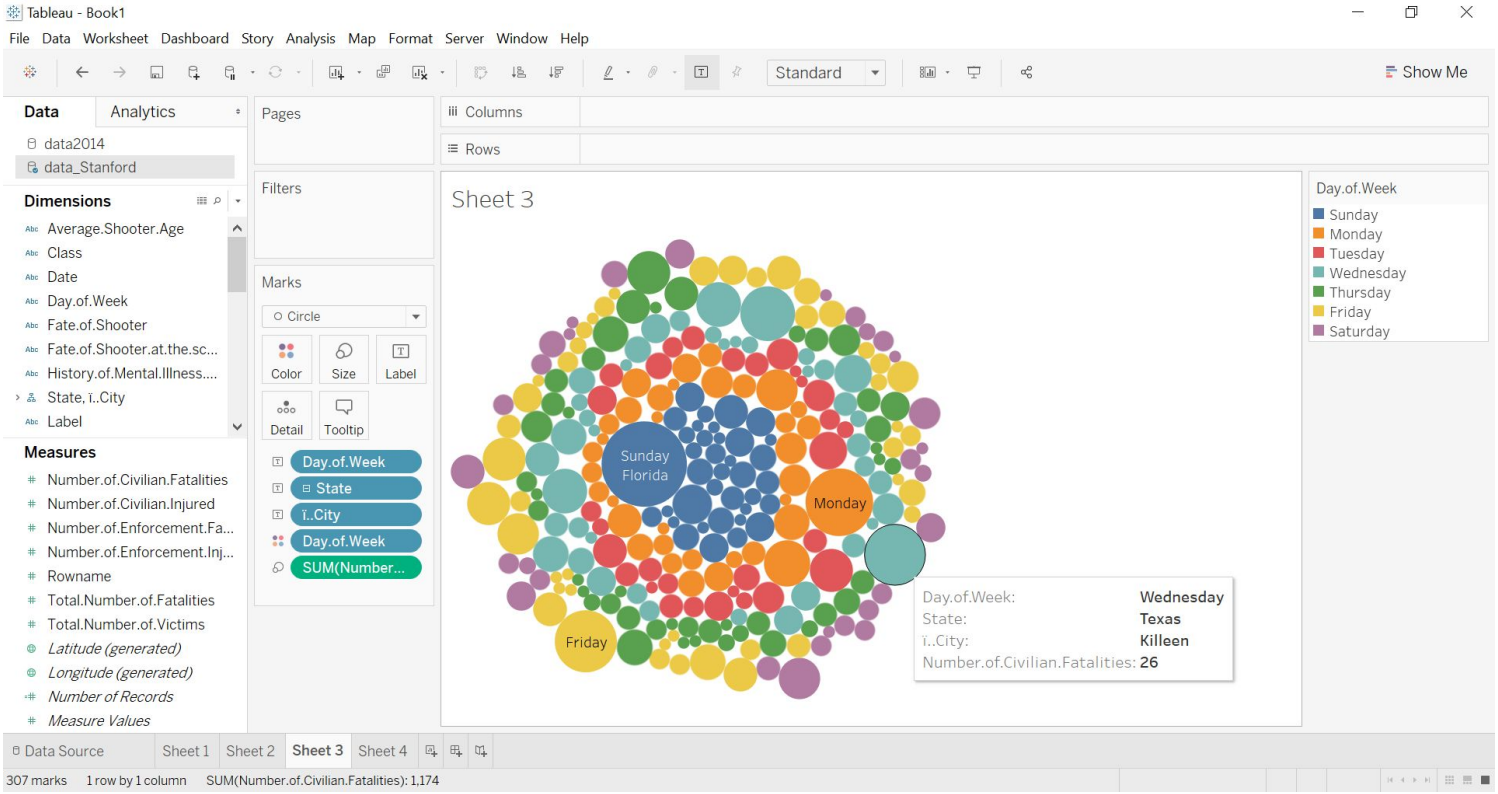
Modeling - RapidMiner Weights Summary



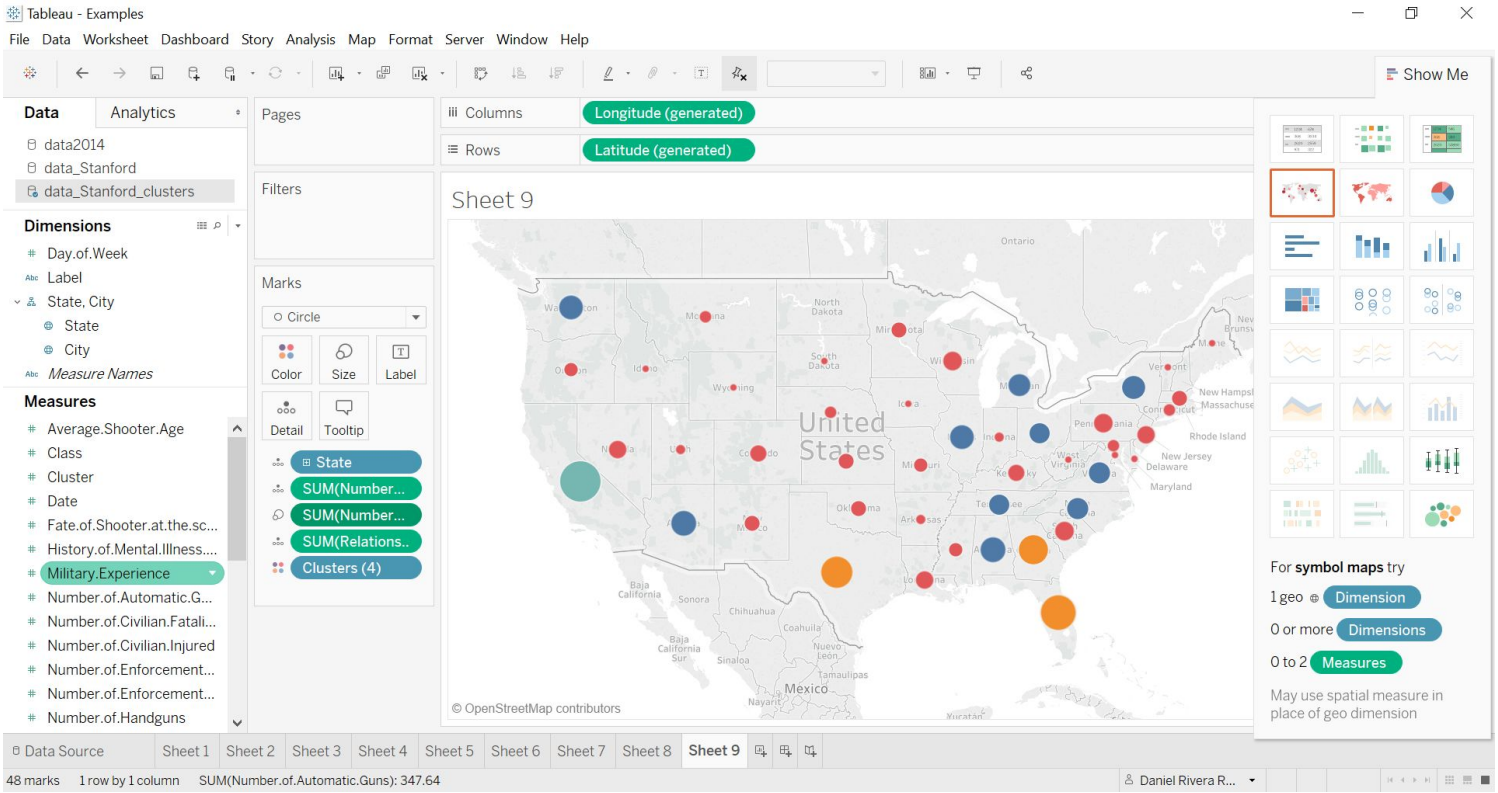
Data Visualization with Tableau



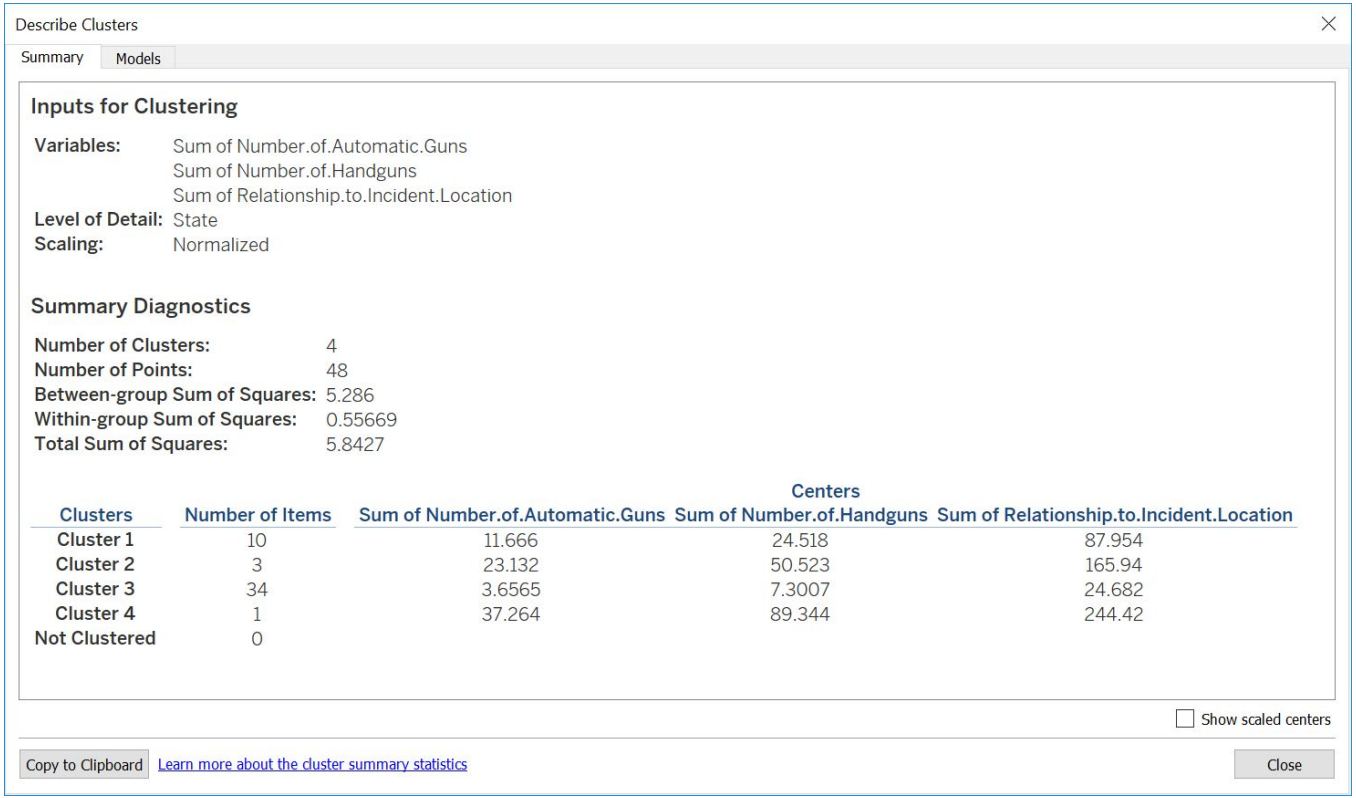
Data Visualization with Tableau



Data Visualization with Tableau



Data Visualization with Tableau



Conclusions

- Are there characteristics that make a shooting deadlier than others? For example, his/her race, gender, income, educational attainment, etc.?
- What role does mental health play?

The only definitive thing we can say is that deadlier weapons+type of place = a deadlier shooting

Through clustering, we saw that there are some events that share similar features

It’s inconclusive if mental illness plays a role (one experiment said it did, another said it didn’t and the resulting accuracies were about the same)

Questions?