Exploratory analyses on gun deaths in the USA using Self-Organizing feature Maps

Christiaan Neil Burger

18974139
School of Computer Science
Stellenbosch University
Student Number, 18974139
Email: 18974139@sun.ac.za

November 1, 2019

Contents

1 Introduction

A Self-Organizing Feature Map (SOM) is an unsupervised neural network which projects high-dimensional data onto a lower-dimensional grid. Data is then clustered due to the similarities found by the neural network and prediction, inference and exploratory analyses can be done on the data using this SOM model. In this report, we will make use of the SOM for exploratory analyses on gun deaths.

2 The problem

The purpose of this paper is to do exploratory analyses on a dataset of gun violence in the United States of America. We are analyzing without any prior knowledge of the data given to us.

In this paper tried to find similarities between different states and state gun laws, based on some attributes such as victim information, state information and the gun laws in a state by using a SOM.

The goal of this paper is to do feature selection, get the optimal grid size and architecture for our SOM and analyze the clusters formed by the SOM.

3 The Dataset

Description	Count
Total observations	8306
Total features	34
Total observations with incomplete data	3865

Table 1: Dataset Characteristics

Variable Name	Description
Date_killed	Date of victim death
City	City where the death occurred
State	State where the death occurred
Victim_Name	Name of the victim
Victim_AgeGrp	Age group child, teen & adult
Victim_Gender	Gender of the victim
Source	Source of information
Long	Death Longitude
Lat	Death Latitude
State_Pop	Population of each state
Back_CHK	Background Checks per 100K
LG_SP	Rifle - State permit to purchase
LG_FR	Rifle- Firearm registration
LG_AL	Rifle - Assault weapon law
LG_OL	Rifle - Owner licence required
LG_CP	Rifle - Carry permits issued
LG_OC	Rifle - Open Carry
LG_SLR	Rifle - State Preemption of local restrictions
LG_NFAR	Rifle - NFA wapons restricted
LG_PJ	Rifle - Peaceable Journey laws
HG_SP	Handgun - State permit to purchase
HG_FR	Handgun - Firearm registration
HG_AL	Handgun - Assault weapon law
HG_OL	Handgun - Owner licence required
HG_CP	Handgun - Carry permits issued
HG_OC	Handgun - Open Carry
HG_SLR	Handgun - State Preemption of local restrictions
HG_NFAR	Handgun - NFA wapons restricted
HG_PJ	Handgun - Peaceable Journey laws
Killed_year	Year of the kill
Killed_month	Month of the kill
Killed_day	Day of month of kill
Killed_day_week	Day of the week killed

Table 2: Data Variables and Names

4 Feature Selection

Feature selection is implemented to reduce the dimensionality of the data and to remove features that do not provide any more information to the problem. In this section of the report, we will be reporting on the features removed and the reason for the removal.

Features that were removed without doing any analysis:

- Date killed
- Victim Name
- Source

4.1 Date Killed

The date killed was removed, because other features in the dataset already capture this information such as Month killed and day of Month killed.

4.2 Victim Name

The victim name feature is very sparse, the number of unique occurrences is 7365 out of a dataset with a total of 8306 observations which amounts to 88.671% of this feature that is unique.
This feature does not contribute to the problem that we are trying to solve, in terms of the clustering of the states.

4.3 Source

The source was removed, due to the fact it has nothing to do with the data and just the location where the data was sourced from. It contains 7190 unique values, which amounts to 86.564% of this feature is unique. Upon further inspection of the dataset, there is some variables where more than one victim died in an incident, and it is reported in the same article. This feature does not contribute to the problem that we are trying to solve.

4.4 Longitude & Latitude

In Figure 1 we can see that all the deaths took place in the USA making sure all the observations is in the USA.
The number of unique longitude and latitude pairs in the dataset is 2703, which translates to 31.06% of unique values. The longitude and latitude data are dependent on one another in each observation because they represent a single point on the map, together. They are observed as a pair, because there can be

values in the dataset with the same longitude values and not the same latitude. Location information is captured in the state variable that we are investigating in this problem. Longitude and latitude have been dropped so that the neural network does not find similarities between the longitudes and latitudes of the observations.

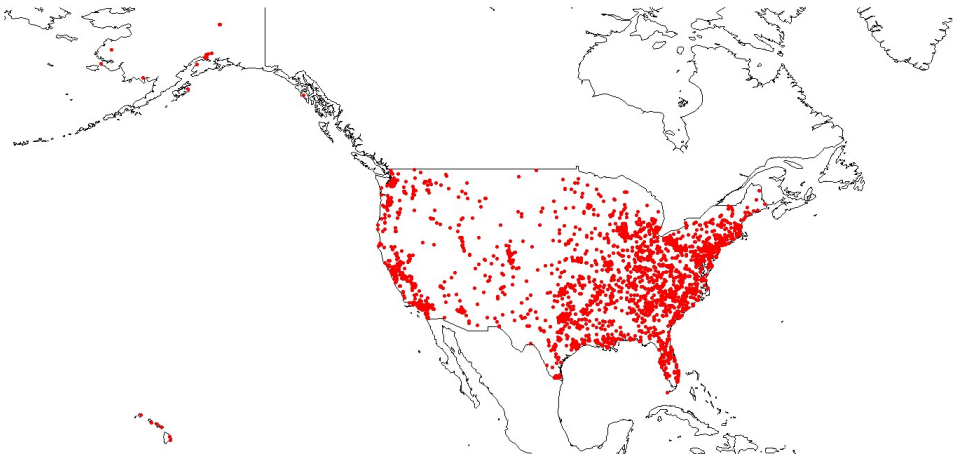


Figure 1: Longitude & Latitude of data observation in the dataset

4.5 City

The number of unique state and city pairs amounts to 2546 observations. 29.25% of the observations are unique. State and city pairs refer to a unique pairing of a state and each of its cities. There are states in the dataset that have the same city names. We decide to drop this variable, to avoid the neural network finding similarities between city names, which are not in the same state, and the problem is to find similarities between states. Data extraction of the same city names in multiple states:

City Name	State Name
Abbeville	Louisiana, South Carolina
Canton	Ohio, Texas. Georgia, Pennsylvania
Columbia	South Carolina, Pennsylvania, Maryland, Missouri
Lakewood	California, Colorado, Washington, New Jersey
Maywood	Illinois, Missouri, California
Miami	Florida, Oklahoma, Ohio

Table 3: Different states same city name - Example

4.6 Age Group

We can see in figure 2 how this feature where binned. The adult bin had the most significant amount of observations. The binning per group range not equal and is skewed. Ages 18 and older are classified as adult, 0-12 child and 13-17 as a teen. We drop this feature because the age information is already included in the numerical age feature provided.

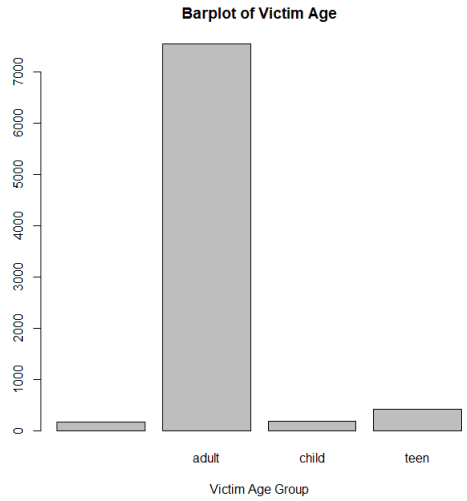


Figure 2: Barplot - Age group

4.7 Year killed

The observations in the dataset span over a period of 9 months, from December 2012 to September 2013. We omit this variable so that year killed in does not play a factor and divide the data. Since we only have observations from the 12^{th} of December for 2012. We see that the is feature is very skewed, as depicted in Figure 3 in a histogram.

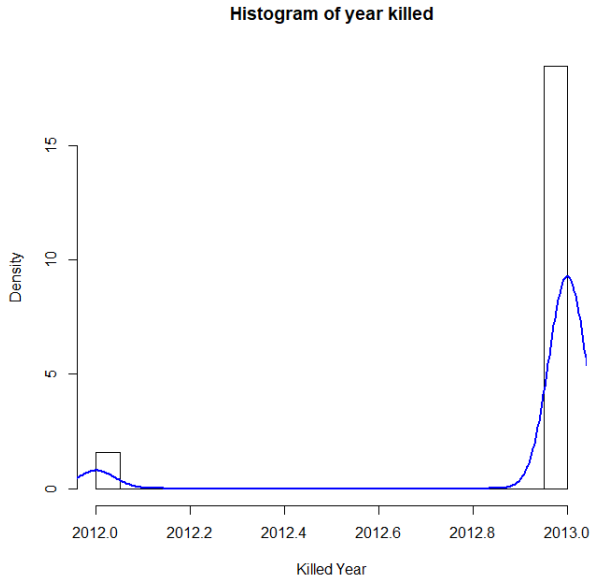


Figure 3: Histogram - Year killed

4.8 Day of the month:

We are dropping the variable day of the month because we have the day of the week and the month a death took place.
We are investigating the deaths occurring in each month as well as the day of the week, deaths occur and not the specific day of the month. This may require sinusoidal or Fourier transformations of the feature, as some literature suggests. The numeric values 28, 30, 31 is the same distance from 1 depending on the month because each month the value resets to 1 we have a cyclic feature. The distribution is also quite uniform, as shown in the histogram in Figure 4, indicating that the variable will also not contribute a lot information to the dataset because there are not a lot of separating factors in the day of the month killed.
The density is tailing to the bottom at either end at of the histogram, because the month resets and the density does not know that 28, 30, 31 is a distance of 1 from the beginning of the month.

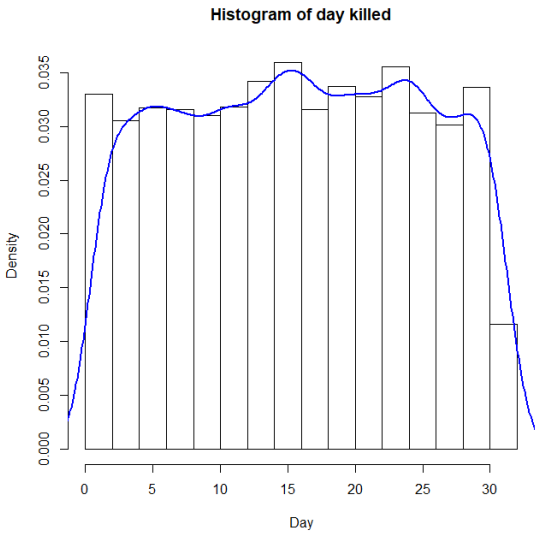


Figure 4: Histogram and density of Day of the Month killed

4.9 Summary of dropped features

1. Victim Name
2. Age group
3. Source
4. City
5. Longitude
6. Latitude
7. Date Killed
8. Year Killed
9. Day of the month killed

We effectively reduce the dimensions of the dataset by 9. In section 5.2, we will see that we reduced the dimensionality significantly due to dummy encoding of features.

5 Data Pre-Processing

Pre-processing refers to how the dataset and its features are processed before it is presented to the neural network so that our algorithms can make sense of the data presented to it.

5.1 Transforming numerical attributes to categorical

The attributes: day of week killed and month killed has been transformed from numerical values to categorical values. We transform these variables due to the fact they are cyclical, after Sunday (7) the variable resets to Monday (1). Where the distance from Monday to Sunday is one day apart but if it is numerically encoded the difference is 7. The same applies to the month of the year, January is close to December, with a difference of 1 month. The encoded solutions can be seen in table **??**. We do this encoding to categorical so that when we make use of dummy encoding this feature in section 5.2 the numeric distances will not play a factor and a dummy variable for each day of the week and month will be created.

Attribute	Original Value	Transformed Values
Day of the week killed	[1,2,3,4, 5,6,7]	[Mon, Tue, Wed, Thu, Fri, Sat, Sun]
Month killed	[1,2,3,4, 5,6,7,8,9,12]	[Jan, Feb, March, Apr, May, Jun, Jul, Aug, Sept, Dec]

Table 4: Converting from Numeric to categorical

5.2 Transformation of Categorical features

We make use of dummy encoding to transform the categorical features to numerical. Dummy encoding works by creating a variable for each unique attribute that is present in a categorical variable [1].

By dummy encoding, we transform the categorical features to numeric values. All the categorical features in the dataset have been changed to numerical dummy variables.
Dummy encoding works by giving the observation a value of 1 if it belongs to the feature otherwise 0. Each categorical variable with K unique values has been transformed to a set of $K + 1$ dummy variables in this dataset, where the extra dummy variable represents the missing values present in the attribute that is dummy encoded, page 324 [1]. Our dimensions of the dataset increases with dummy encoding, but it is not a problem for a Self-Organizing feature map.

An example of how the variables are transformed, the variable Victim_Gender

has been transformed into:

Dummy Variable Name	Description
Victim_Gender.Male	1 if observation is male else 0
Victim_Gender.Female	1 if observation is female else 0
Victim_Gender.Missing	1 if observation value is missing else 0

Table 5: Dummy encoding on gender- Example

If an observation is male, the male dummy variable will have a 1 and 0's for the female and missing dummy values.

The dimensional of the dataset after dummy encoding grew to 135 dimensions.

5.3 Removing outliers

While looking at the data, we observed that the maximum age present in the dataset is 107 and the minimum age is 0, where 0 indicates a child younger than one year old. We decided to remove the lower and upper 0.5% percentile of the data which amounts to 1% of the total observations were omitted from the dataset, a total of 83 observations in the dataset.
In Figure 5, we see that the difference in the density plot of the histogram is minimal, and the mean shifted 0.1 from 32.97 to 32.87. The density with grew marginally.
We can see that the age distribution is skewed to the right.

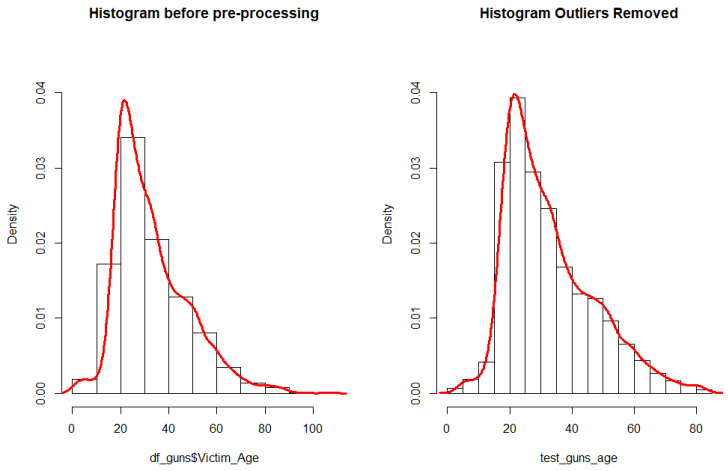


Figure 5: Victim Age Histogram, before and after outliers where removed

5.4 Missing values

The dataset contained a lot of missing values. Self-Organizing feature maps works with missing values if correctly encoded as in section 5.2. However, during the import of data in R, some of the observations contained rows of NA values that were not able to be re-encoded as they had no information to very little. The observations where missing values were present was omitted from the dataset, which resulted in a reduction of 7.39% of observations of the original dataset. A total of 614 observations was removed.

5.5 Total Removed values

Removing outliers and observations that were empty NA values, we reduced the total number of observations to 7602 observations which is a total reduction of 9.15% of the original dataset. The results where promising when comparing the distance Self-Organizing map before removing the outliers and NA values and after. It resulted in a more clear distance map.

5.6 Normalising the data

Normalising the data means that we make the features unit-less and scale all columns to have the same minimum and maximum value. We do this because we are making use of a distance measure while training the SOM, the euclidean distance. Not scaling the features can lead to some features being seen as more important than they are.
When we are not normalising the data, for example, having two columns where one is measured in meters and the other in millimetres. On the distance scale, the meters may be seen as smaller values compared to millimetres because 0.5 meter is equal to 500mm. Scaling the features between 0 and 1 will make the distance measure effective and possible to use. None of the dummy variables needed normalisation, the values present was either a 1 or 0 in a dummy variable column.
The variables that was scaled between 0 and 1 where:

- Victim Age
- Background Checks
- State Population

Summary of the dataset after pre-processing	
Number of observations	7603
Unique patterns	7138
Independent patterns	69
Number of dimensions	135
Total real attributes	3
Total numeric attributes	132
Normalised/ Scaled between 0-1	True

Table 6: Final data set Characteristics

6 Architecture and parameterization

The Kohonen and several other clustering and statistics measure packages was used in R to create the SOM and to generate the results obtained in this report.

The architecture refers to what the SOM consists out of.
The SOM function used in this report had the following required the following and inputs:

Parameter	Value
Data	Scaled version of the training set
Grid	Grid with weights and the corresponding size
Iterations	1000-10000
Learning Algorithm	Batch (parallelized)
Learning rate	Not used due to the Learning Algorithm
Radius	Neighbourhood radius, covers 2/3 of all unit distances

Table 7: SOM R function inputs & characteristics

6.1 Iterations

The iterations refer to the number of times the dataset was presented to the SOM so that the SOM can be updated. Each iteration updates the algorithm. The learning algorithm used is the batch learning algorithm explained in section 6.2.

The number of iterations chosen, was done by fitting a SOM with the desired grid size multiple times until the mean distance between nodes converges, and we had the desired quantization error that was stable.

6.2 Learning Algorithm - Parallel Batch

The parallel batch used in the R package is the same as the standard batch algorithm parallelized making use of all the available cores on a system decreasing the time to compute the final SOM.
As mentioned on page 65 in [2], the stochastic SOM training algorithm is slow due to the updates of the weights after each pattern representation.
The batch map training algorithm developed by Kohonen [3] updates all the weight values after all the patterns have been presented to the SOM after each iteration.

Algorithm 1 Batch Self-Organising map, page 65 in [2]

Initialize the codebook vectors by assigning the first KJ training patterns to them, where KJ is the total number of neurons on the map.
while *max number of iterations not reached or threshold* **do**
  for *each neuron, kj* **do**
    Collect a list of copies of all patterns z_p whose nearest codebook vector belongs to the topological neighbourhood of that neuron;
  end
  for *each codebook vector* **do**
    Compute the codebook vector as the mean over corresponding list of patterns
  end
end

The R-package has three available learning algorithms to choose from the online, batch learning or parallel batch algorithm. Since we have the whole training set, we make use of the batch learning algorithm, presenting the SOM algorithm the entire dataset and then updating the parameters.

6.3 Learning rate:

The batch learning algorithm does not make use of a learning rate parameter, and therefore has no convergence problems and yields stabler asymptotic values than the original SOM, as mentioned on page 140 [3].
The advantage is that we have one less parameter to estimate for the SOM. A possible disadvantage is if the available cores and memory on a system are low, the algorithm may still take long to finish after the iterations.

6.4 Quantization Error

The euclidean distance is for all the patterns to the codebook vector of the winning neuron:

$$\epsilon_T = \sum_{p=1}^{No\,of\,patterns} \|z_p - w_{mn}(t)\|_2^2$$

This measure the error of the map, smaller error the better the map. With batch learning, we can improve this by increasing the amount of iterations that the algorithm gets to train on the dataset [2].

6.5 Neighbourhood function

The package allows us to choose between a bubble neighbourhood function or the Gaussian [4].
The Gaussian neighbourhood function is used in this report. With the Gaussian function, all the neurons will be in the neighbourhoods of the best matching neuron [2], the neuron that is the winner for the frequency of patterns associated with it.

With the Gaussian function, the codebook vectors are updated even if they are far from the best matching neuron [2].
We decided to use the **Gaussian** neighbourhood function.

6.6 Radius of the neighbourhood function

The value for the radius in this package decreases linearly from the radius to zero. As soon as the neighbourhood gets smaller than one, only the winning unit will be updated. The starting radius size is chosen as follows:

$$radius = quantile(neighbourhood\ distance,\ 2/3)$$

Where the 66.667% quantile value of the neighbourhood distances [4].

6.7 Grid size

The optimal grid size was determined by looking at different grid sizes, ranging from 10x10 to 85x85. We started small and grew the size of the map in increments of 5 neurons in the width and length.

We made use of a hexagonal topology, which means that each neuron has six neighbours.

In this report, we chose a square grid, but any grid size could be used. We tested non-square dimensions and the mapping would still be the same just adjusted for the new dimension's.

Ideally, the number of neurons should be equal to the number of independent training patterns, page 62 of [2].In linear algebra, we refer to this as the number of linear independent rows in the dataset. We took the transpose of our data matrix and got the rank of the transposed matrix (number of linear independent columns), which will give us the number of independent patterns that are not linear combinations of each other. The result can be seen below.

$$rank(DataMatrix^T) = 69 = \sqrt{69} = 8.306 \approx 9$$

Which results in a grid size of 10x10 to have enough neurons for the independent samples present in the dataset, which amounts to 100 neurons.

As seen the grid images to follow below the natural boundaries that form during the training of the SOM a be seen in the plot of the U-matrix otherwise known as the neighbourhood distances. We did not include all the outputs of the distance plots to save space, and it would be redundant. Areas with a low neighbourhood distance (red) indicate similar nodes, and areas with large distances indicate nodes that are much more dissimilar. All the figures were generated by the batch algorithm and are the results from table 8 where the number of iterations was equal to 10000.
It is difficult to see the natural boundaries with the human eye that occur on the 10x10 plot, but they do exist.

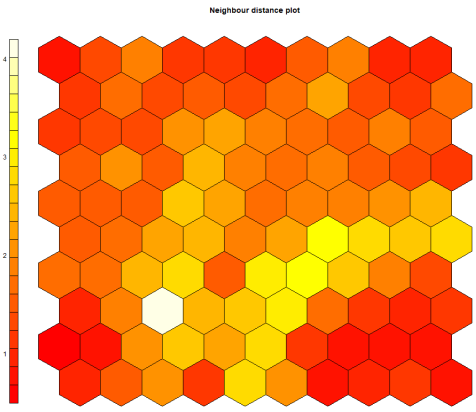


Figure 6: Grid size 10x10 - 10000 iterations

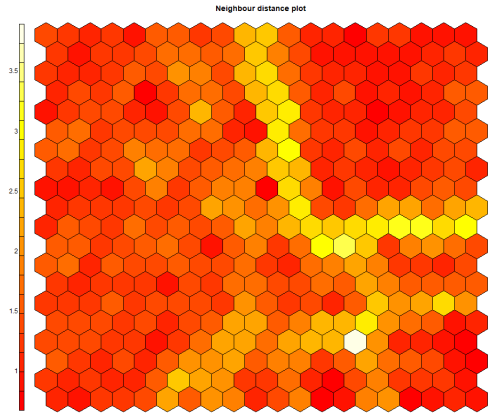


Figure 7: Grid size 20x20 - 10000 iterations

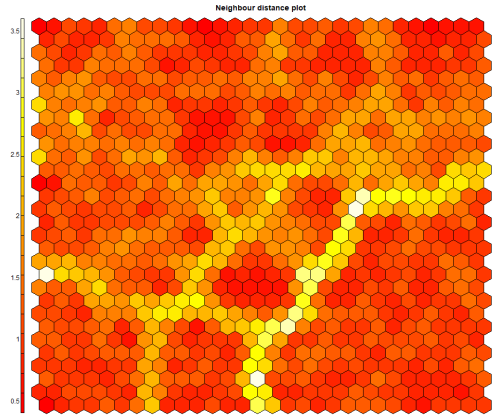


Figure 8: Grid size 30x30 - 10000 iterations

We see that all the figures has more or less the same structure just scaled and some are rotated version of the other. With the structure we are talking about the similar natural clusters that occurs in the distance. Table 8 has the different model sizes and the quantization errors.

Dimensions	**Number of training Iterations**	**Quantization Error**
40 x 40	10000	0.717
40 x 40	7500	0.707
40 x 40	5000	0.713
40 x 40	2500	0.724
30 x 30	10000	0.903
30 x 30	7500	0.896
30 x 30	5000	0.900
30 x 30	2500	0.900
20 x 20	10000	1.093
20 x 20	7500	1.110
20 x 20	5000	1.113
20 x 20	2500	1.105
10 x 10	10000	1.398
10 x 10	7500	1.401
10 x 10	5000	1.395
10 x 10	2500	1.401

Table 8: Iterations and Quantization error of different grid sizes

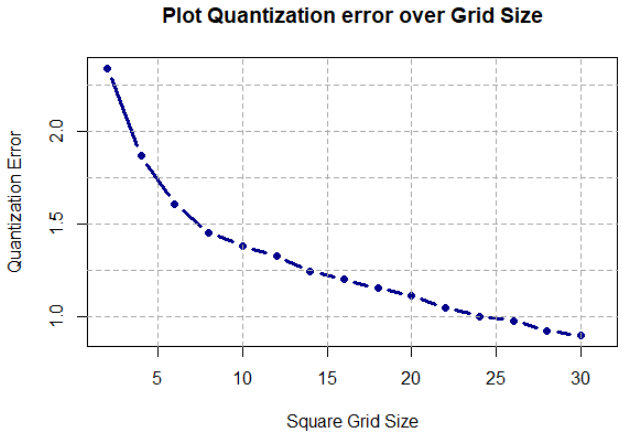


Figure 9: Quantization error over different grid sizes

During this report, careful consideration was given to the optimal number of neurons present in the dataset, evaluating numerous grid sizes and clustering thereof. **We decided on a final size of 10x10, 100 neurons**, results from the 10x10 model can be seen in section 8. We concluded that for larger grid sizes, the mapping and the classification of all of the patterns remained the same.

The quantization error improved linearly from 10x10 on-wards as seen in Figure 9, with small changes in improvement in the quantization error this indicates that the optimum will only be improved with the grid size. The optimum grid size is the point where the quantization error starts to decline linearly. A larger grid size will always lead to a better quantization error.

We chose a simple **10x10 grid**, to represent our data as it had the same clustering as larger grid sizes. *"Why use a 10-pound hammer to kill a fly?".*

Reasons for choosing this grid:

1. The quantization error is not too small, with a value of 1.4 on average. A tiny quantization error means that the map is too large and that convergence takes to long, page 66 in [2] of it is to big the map is too small. Quantization error is the mean distance, increasing or decreasing the map will automatically affect the value of the quantization error.

2. The quantization does not rapidly improve after this grid size, and it only improves due to larger grid sizes.

3. Convergence is fast, and computation time is short.

4. Training time does not take as long as for the larger grids. The training

takes exponentially longer as the grid size increase since the number of neurons increases exponentially.

5. All 69 linear independent patterns are represented in this map.

7 Clustering

In this section, we are going to use different measures to find the optimal number of clusters for our model. The cluster method that we are implementing is Wald's Hierarchical clustering, page 69 [2]. Ward clustering follows a bottom-up approach where each neuron initially forms its own cluster. At consecutive iterations, two clusters that are closest to one another are merged, until the optimal or specied number of clusters has been constructed[2]. The result of Ward clustering is a set of clusters with a small variance over its members, and a large variance between separate clusters [2].

To get the optimal number of clusters, we start by fitting different cluster sizes and then measuring the goodness of the fit with the selected size and get the best optimal number of clusters this way by different measures.

7.1 Within-cluster Sum of Squares

Also known as the "elbow" method. This method works by looking at the within-cluster sum of squares graph. The optimum number of clusters is when there is a decrease from one WSS size to another that is not too significant and the graph start plateauing after this cluster value, this will be optimum. By looking at figure 10, and reading the graph from left to right, we see that the plateauing (transitioning from exponential to linear) starts at 4 or 5. The WSS statistic measure indicates 5 clusters to be the optimum.

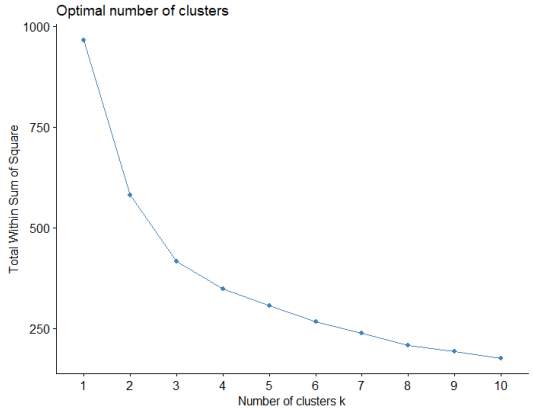


Figure 10: WSS plot for 10 different cluster sizes

7.2 Silhouette

The silhouette statistic plot displays the measure of how close each observation is to its own cluster compared to the neighbouring clusters. It measures how well each observation is classified. The silhouette measures the distance with the euclidean distance.
As seen in figure 11, this measure indicates 2 clusters to be optimum with 5 clusters as a close runner up. We do not consider 2 clusters as the data can be broken up in more clusters with hierarchical clustering.
The silhouette recommends 5 clusters to be optimum.

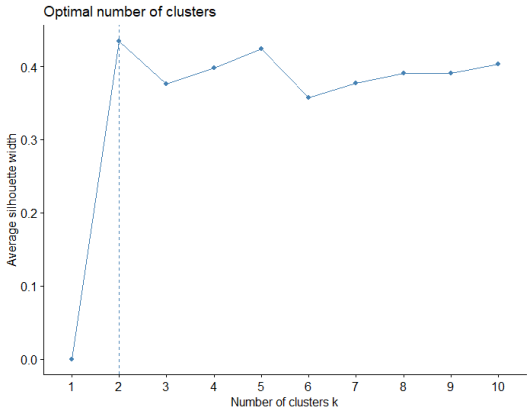


Figure 11: Silhouette plot for 10 different cluster sizes

7.3 Dunn Index

The Dunn index (DI) is a measure that evacuates the clustering. Where a higher value indicates better clustering, that means that the clusters are compact and well separated from the other clusters.
The DI is the minimum inter-cluster distance divided by the maximum cluster size.

$$DI = \frac{\min\limits_{1<i\leq j<m} \delta(C_i,C_j)}{\max\limits_{1\leq k\leq m} \Delta_k}$$

By using a function called clValid() in R. The best Dunn index value corresponded to having four of five clusters as seen in table 9.

7.4 Davies-Bouldin Index

This metric $R_{i,j}$ for evaluating clustering measures, takes into account the separation between clusters which has to be as large as possible and the within-cluster scatter which has to be as low as possible. This index is defined as the

ratio of S_i (within-class scatter) and $M_{i,j}$ such that the following properties are conserved:

- $R_{i,j} \geq 0$
- $R_{i,j} = R_{j,}$
- When $S_j \leq S_k$ and $M_{i,j} = M_{i,k}$ then $R_{i,j} > R_{i,k}$
- When $S_j = S_k$ and $M_{i,j} \leq M_{i,k}$ then $R_{i,j} > R_{i,k}$

If $R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$ then the DB index is defined as $DB = \frac{1}{N} \sum_{i=1}^{N} \max_{j \neq j} R_{i,j}$

The DBI recommends 7 clusters to be the optimum amount of clusters, as seen in Table 9.

Clusters	**DBI**	**Silhouette**	**DI**
2	0.5555187	0.4343	0.4963
3	0.8091609	0.3759	0.3268
4	0.8224999	0.3985	**0.3509**
5	0.7846066	**0.4246**	**0.3509**
6	0.7620004	0.3577	0.2304
7	**0.7598827**	0.3697	0.2304
8	0.7788678	0.3905	0.2304
9	0.7944177	0.3910	0.2304
10	0.8033926	0.3933	0.2304

Table 9: Cluster measures results, DBI Sillhouette, DI

7.5 Conclusion

As seen in Table 10, Silhouette, WSS and the Dunn Index says that 5 clusters will be the optimum and the DBI says 7. We will, therefore, choose to have **five clusters** in our final model based on the most popular amount of recommendations for the optimal number of clusters metrics.

Cluster Measure	**Optimal Clusters**
WSS	5
Silhouette	5
DBI	7
DI	5

Table 10: Optimal number of clusters summary for each measure

8 Final Self-Organizing Map Model

Summary of the final SOM	
Shape	Square
Dimensions	10 x 10
Total neurons	100
Learning Algorithm	Batch
Learning Rate	-
Iterations	10000
Neighbourhood function	Gaussian
Distance Measure	Euclidean
Topology	Hexagonal
Number of clusters	5
Clustering method	Wards Hierarchical

Table 11: Final SOM Architecture

8.1 Convergence

We can see in figure 12 that the SOM converged after 9500 iterations.

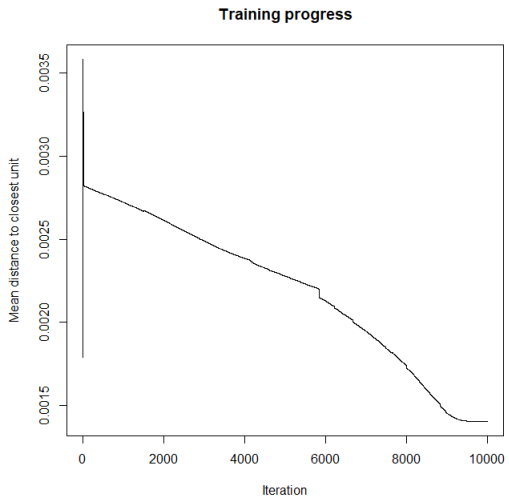


Figure 12: Graph of Mean distance to to closest unit over iterations 10x10 grid

8.2 Neighbourhood Distance Heat-map

In Figure 13, we can see the natural boundaries that occurred after training the SOM.

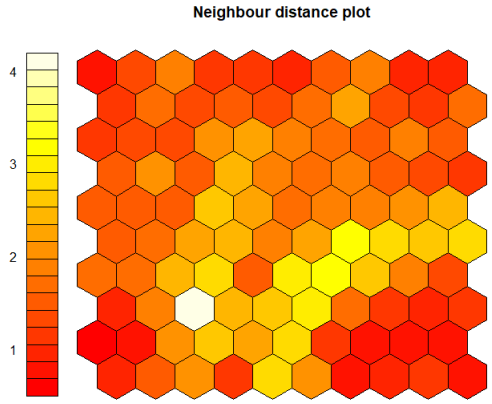


Figure 13: Neighbourhood distance plot, 10x10 grid

8.3 Clusters and Mapping

In Figure 14, we can see how the data is divided into 5 clusters and how the observations mapped to each neuron in our Self-Organizing Map.

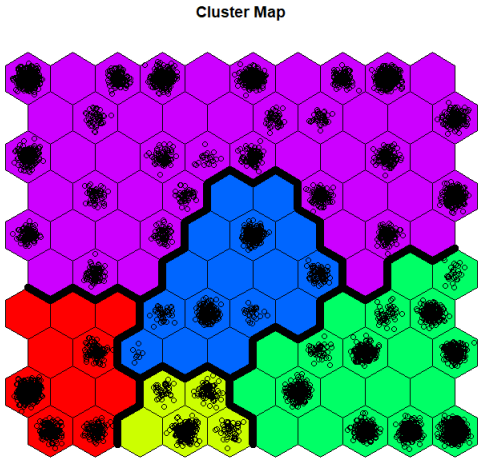


Figure 14: Clustered map and observations mapped to each neuron

8.4 Node Counts

With this graph, we can see the counts in each node of our model. We can see that many nodes are still open and that our mapping is quite uniform. This suggests that our grid size is optimum.

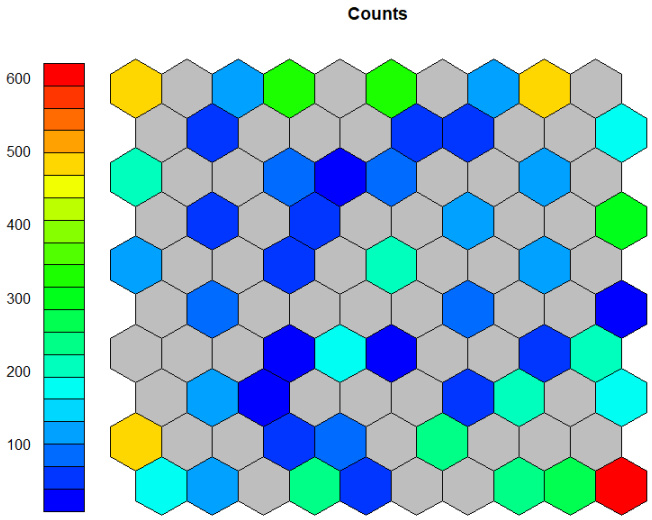


Figure 15: Observations mapped to each neuron counts

9 Exploratory Analysis

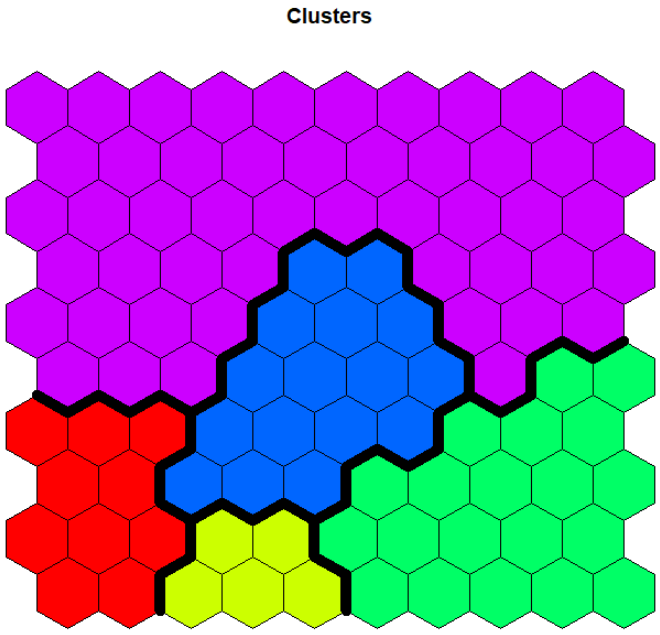


Figure 16: Clustered map

Cluster	Colour	No. observations	% Deaths	No. Features	No. States
1	Red	870	11.44	41	1
2	Yellow	428	5.63	41	1
3	Green	2128	27.99	57	16
4	Blue	523	6.88	61	4
5	Purple	3653	48.05	91	29

Table 12: Cluster statistics & information

9.1 Problem Statement

We want to investigate each cluster that formed in our SOM in figure 16 and explore the reasons for the clustering and the attributes that contribute to the clustering. Describing each feature and explore the relationships within clusters. By doing this, we want to explore the relationships that exist between states and the gun laws present in each state.

9.2 Feature Analysis

9.2.1 Victim Age

In our analysis, we saw that the victim age did not play a role in the clustering of our data. Looking at the corresponding component maps of age, the neurons in all of the clusters are equally spread. Not playing a big role in creating dissimilarity between the clusters. Looking at the age, we see that their densities for all the clusters looks equally spread, looking at Figure 17. The ages is skewed to the right, meaning that most of the deaths are in lower age groups. In table 13 we see that the medians are more or less the same for all clusters with the biggest difference of 4 years. We are not looking at the mean value, since there is still outlying values present in the dataset and the mean heavily affected by outliers.

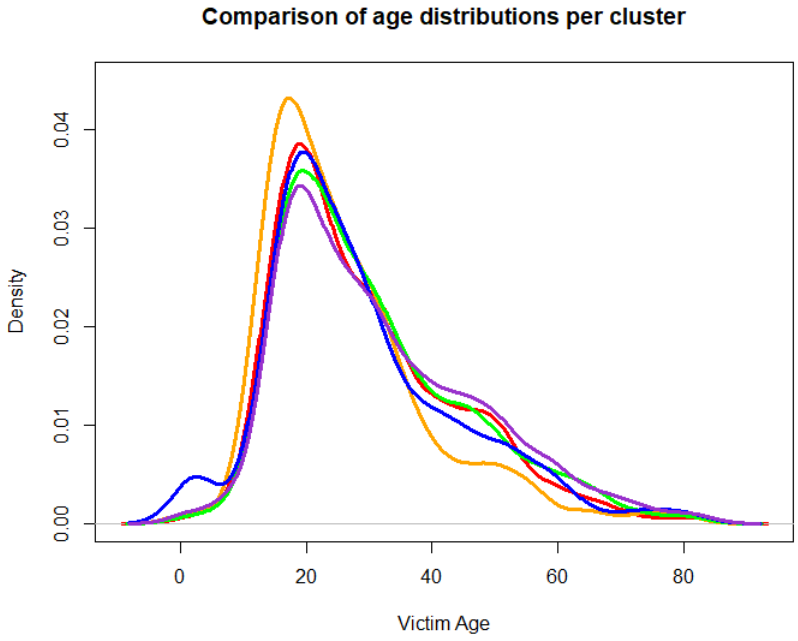


Figure 17: Distribution of victim age for each cluster, each colour represents the corresponding colour

Age Statistics						
Cluster	**Minium**	**Q1**	**Median**	**Mean**	**Q3**	**Maximum**
1	**0**	**18.91**	**26.27**	**29.19**	**36.77**	**83**
2	**1**	**16.81**	**23.114**	**25.883**	**31.519**	**80**
3	**0**	**18.9**	**26.27**	**30.26**	**38.87**	**83**
4	**2**	**17.861**	**25.215**	**28.484**	**35.722**	**81**
5	**0**	**18.91**	**27.32**	**31.44**	**40.97**	**83**

Table 13: Age statistics per cluster

9.2.2 Victim Gender

In this section, we will be analysing the victim gender per cluster.

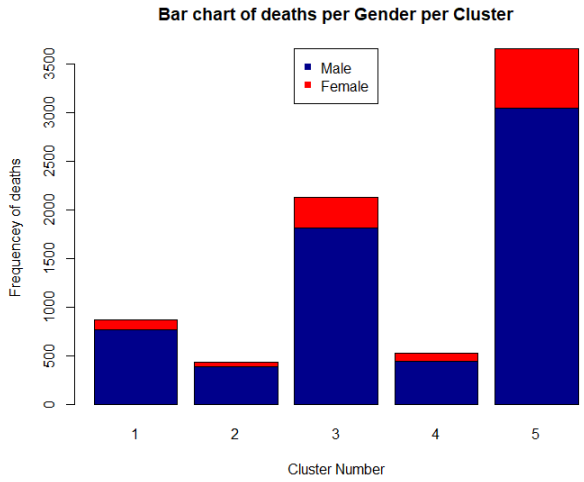


Figure 18: Deaths per Gender for each cluster

We see that there are more male victims than female victims for each cluster. Looking at the bar plot in Figure 18 we see that there are more male victims. In table 14, we see that the percentage of female deaths per state is close to each other with the biggest difference of 9.19% between cluster two (yellow) and cluster one (red). These are the two clusters that only contain one state.

Gender Statistics		
Cluster	**% Males**	**% Females**
1	**88.16**	**11.84**
2	**78.97**	**21.03**
3	**85.23**	**14.77**
4	**84.70**	**15.30**
5	**83.25**	**16.75**

Table 14: Gender Statistics per Cluster

We conclude by looking at the component maps of the victim gender and that the difference between the victim genders are minimal that the gender did not play a significant role in the clustering of the states, this feature can be removed.

9.2.3 Month of Death

We can see by looking at the bar chart below, figure 19, that the frequency of deaths per month is close to uniform per cluster. We see that the frequency of deaths is more for certain clusters, as some clusters contain more states than others.
Looking at this statistics we saw that the number of deaths for cluster one is a lot compared to cluster three and four, which have 16 and 4 states respectively, but the state in cluster one has the largest population which results in a higher frequency, California.

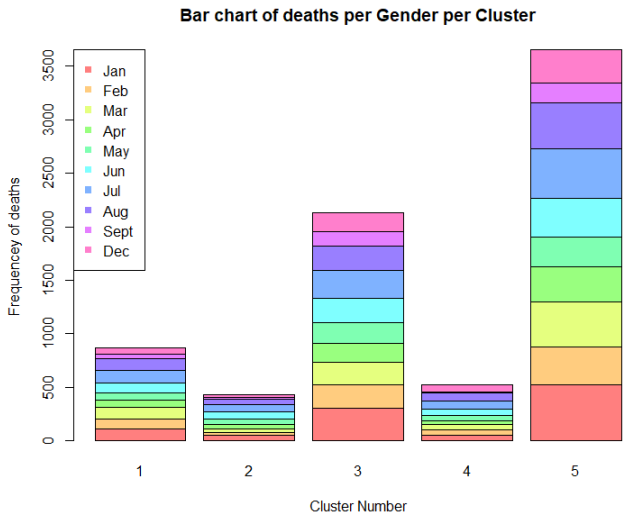


Figure 19: Deaths per month for each cluster

9.2.4 Day of the week of Death

Looking at Figure 20, we see that the frequency of deaths per day of the week are uniform. There is no clear indication that there is more deaths on one day than another. This is also indicated by the components maps of our model.

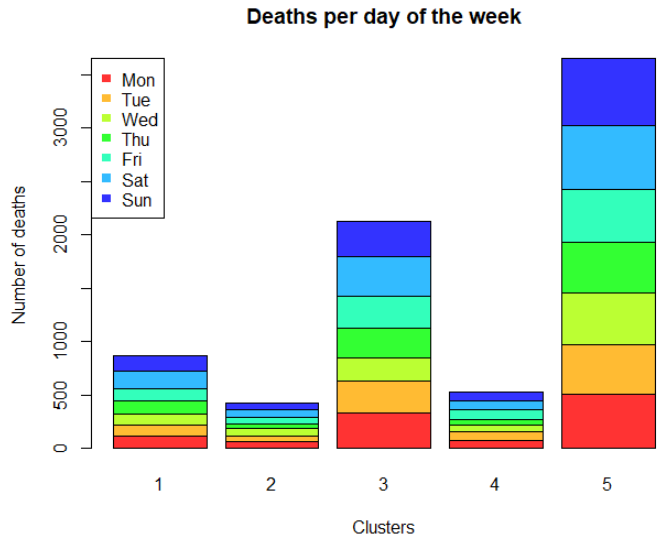


Figure 20: Deaths per day of the week for each cluster

9.2.5 State Population

We saw that the state population feature only contributed to the red cluster, which only contains the state California. California had the highest state population of 36'961'664. This was not an outlying value, as indicated by the box plot in figure 21. We can see that this value is within the 90% quantile of the observations in the whole dataset.

The state population variable was kept in the analysis initially, to not bias the amount of death per state having a distance measure of the total people present. A state with a higher population will have a larger number in deaths than a state with a smaller population.

During our analysis, we saw that we had made a possible error, including Washington D.C., in our model as a state. Washington D.C. is a city within the Washington State, but it is referred to as a Federal district.

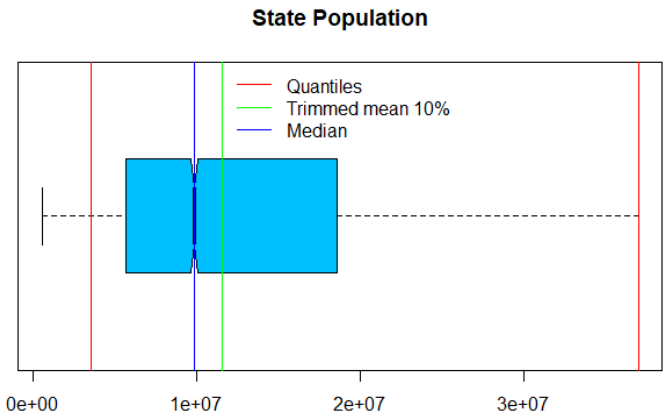


Figure 21: Boxplot of the State population with 10% and 90% quantiles, median and trimmed mean.

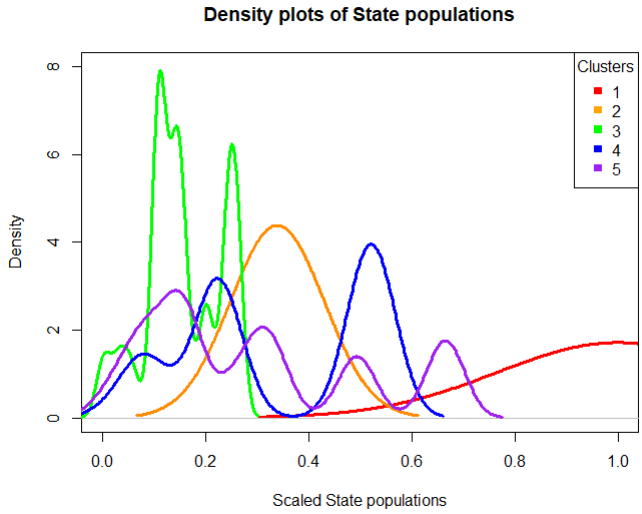


Figure 22: Density plot of the state populations

By looking at the density plots of the different clusters in Figure 22 we can see that cluster one differs the most from the rest of the clusters, and we see that the state population played a role in differentiating this cluster from the rest. We see that the rest of the clusters density plot varies a lot, but all of them is centred to the right and overlap with one another.

9.3 Cluster Analysis

In Table 15 in the appendix, we can see how each state has clustered and each of the features belonging to each cluster.
A large number of states clustered within the purple and green cluster. In this section we will be discussing each cluster's characteristics and what makes them unique/different from the other clusters.

We are going to investigate in-depth why there are six states that is not following the majority of all the other clusters that is grouped in their own clusters. These states must have attributes unique to them that caused them not being clustered with the other states. We are most interested in the clusters that are different. We will investigate similarities and dissimilarities between clusters.

9.3.1 Cluster 1 - Red

Only one state belongs to the red cluster, **California**. As we saw in Section 9.2.5, the state population for this cluster was different from the rest of the variables. One possible reason for California forming its own cluster is due to the large value of its population. We saw in Table 12 that 11.44% of the deaths was represented by this cluster i.e. California.

Features that made this cluster unique from the rest is:

- Partial state permit to purchase a handgun

California has a large population compared to the other states, but in terms of all the observations in the dataset it was not an outlier as seen in the box-plot in Figure 21.

9.3.2 Cluster 2 - Yellow

Only one state belong to the second cluster, **Illinois**. Attributes that made this cluster unique from the rest is:

- No handgun carry permits issued.

We see that this cluster had strict gun laws in general with the following laws belonging to this cluster:

- State permits to buy a rifle and handgun is required.
- Owner license is required for both handguns and rifles.
- No carry permits is issued for both handguns and rifles.
- Open carry is not allowed for both handguns and rifles.
- There is not State preemption of local restrictions for both handguns and rifles.

- NFA weapons are restricted for both handguns and rifles.
- Peaceable journey laws does not apply.

We see that the states in this cluster have strict laws on guns overall. In Table 12, we saw that this cluster represented 5.63% of the deaths in the dataset. The most strict compared to all other states. We believe that this is the reason for this cluster forming.

9.3.3 Cluster 3 - Green

Many states where clustered in this cluster. We saw that what made this cluster unique is that this it contains all the states with attributes that had missing state law values. Looking at the table and the component plots in the appendix, we can see this relationship.

Our dataset had a large number of missing attributes present with 45% of the observations containing at least one missing value. To easy identify states with attributes that had missing values, one can look at the states in this green cluster for possible further analysis to research the reasons why these states had missing values.

One can try to find sources for the missing values and then make a new model.

We conclude that this was the main reason for the separation in clusters, no other clusters contained states that had missing values in their attributes.

9.3.4 Cluster 4 - Blue

Only 4 states belongs to this cluster:

1. Connecticut
2. Hawaii
3. New Jersey
4. New York

Features that were unique to this cluster was:

- Firearm registration for Rifles is necessary.
- Assault weapon law for Rifles is present.
- Partial - State preemption of local restrictions on Handguns.
- Partial - State preemption of local restrictions on Rifles.

These are the only states where there are strict laws on rifle laws. We saw that each law regarding rifles was present. These states has the following regulations on rifles:

- Firearm registration
- Assault weapon Law
- Carry Permits needs to be issued
- Open carry is not allowed.

We conclude that the states in these clusters were clustered together due to their laws on Rifles and that they had more strict rifle laws compared to other states. We saw in Table 12 that this cluster represented 6.88% of the deaths. We can assume that stricter gun laws lead to less gun-related deaths, keeping in mind that correlation does not mean causation.

9.3.5 Cluster 5 - Purple

The majority of states belonged to this cluster. We saw that the background checks played a role in this cluster by looking at the component maps. We will now list the features unique to the states within this cluster that were not a determining factor for other clusters by looking at the component maps.

- No - State permit to purchase a handgun.
- Partial handgun firearm registration.
- Shall issue carry permits for handguns.
- Yes - Open carry for rifles

States that are present in this cluster has less strict gun laws. The majority states within this class has the following laws applied to them.

- No firearm registration is required
- No assault weapon laws are implemented
- No owner licences are required
- Carry permits are issued for guns.
- Carry permits are issued for rifles in some states.
- Shall issue carry permits also
- Open carry is allowed only in some states.
- State preemption of local restrictions applies.

9.4 Redundant features

We saw that many features included in our dataset were redundant and did not contribute to the clustering of our data as they were present in each cluster.

The date which a victim was killed did not play a role in clustering the states in our SOM. We could have dropped the month and day of the week a victim was killed on. The victim information also did not play a role in clustering the data.

The attributes that could have been dropped:

- Victim Age
- Victim Gender
- Month Killed
- Day of the week killed

These variables will not be considered in the analysis of each cluster.

10 Conclusion

We saw that a SOM is a useful tool to visualise multidimensional data on a two dimensional. This mapping makes it easy to identify relationships that may exist between different features of the dataset. SOM's has the ability to work with attributes that contain missing data. In this dataset 44% of our observations each had at least one missing value.

During our analysis of the SOM, we identified more features that was redundant and did not contribute to the unique clustering of features.
We saw the laws that separate states from one another as well as the state population and background checks. States with stricter gun laws were grouped and states with less stricter laws.

For better analysis and relationship exploration between the states, we think that filling in the dataset's missing values will improve the model a lot since a large number of states was clustered together in cluster three (green) that was due to attributes containing missing values. We would then be able to more extensively explore the relationship between the clustering of states and the corresponding laws.

We feel that the data that we have to work with is only limited to crimes that are reported in news outlets, as indicated by the dropped source variable. The USA had a population of 316.2 million people in 2013, and we had only 8703 observations which is a very small amount of gun-related crimes.

To get a better understanding of the impact of the gun laws present in each state, a dataset is required, which does not only contain crimes that were reported in news articles online but also unreported crimes. We do note that this will increase the cost of obtaining a dataset as where information present in the dataset was sourced online. More man-hours and work will be needed to improve the dataset.

SOM are a useful tool for exploratory analysis, and a lot of information can be gathered from them.

References

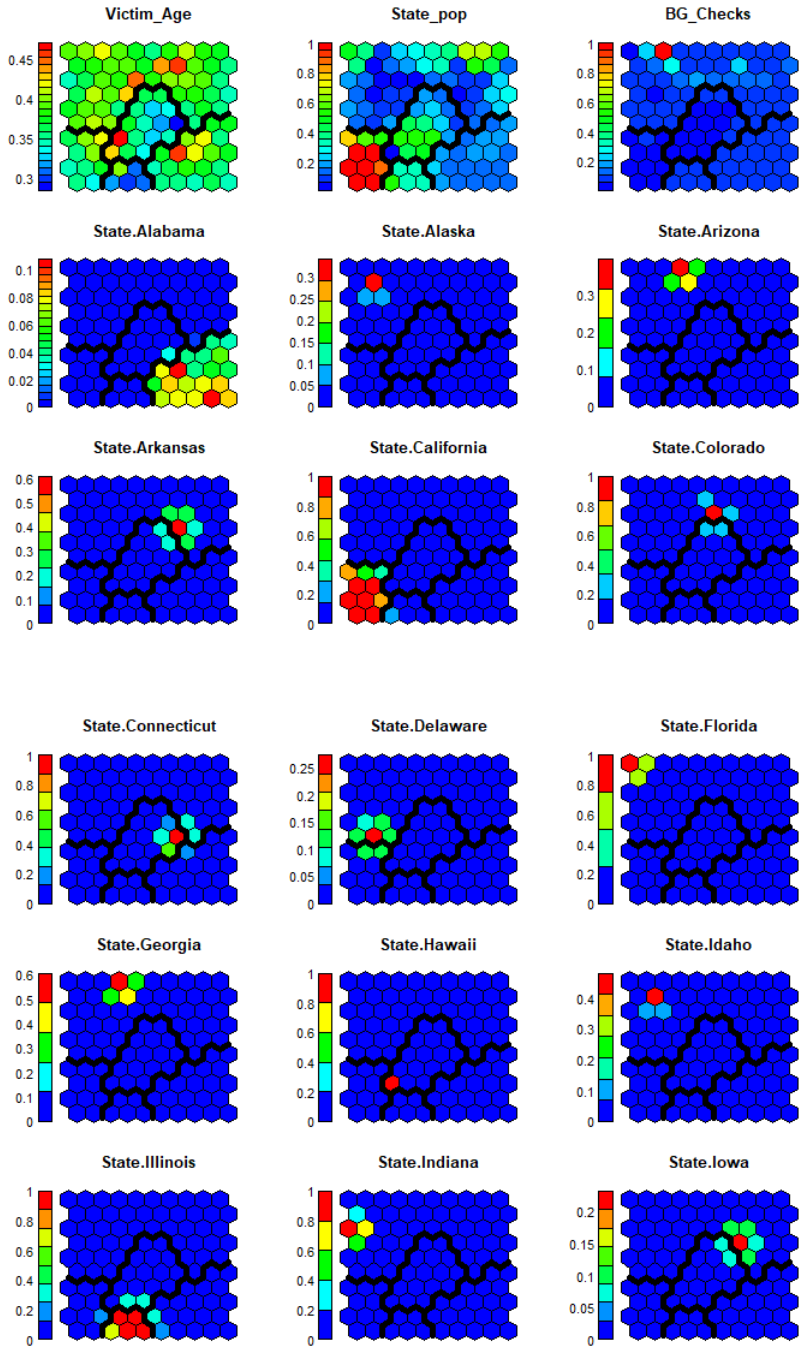
[1] *Applied linear statistical models.* eng. 5th ed. / Michael H. Kutner ... [et al.]. The McGraw-Hill/Irwin series operations and decision sciences. Boston: McGraw-Hill Irwin, 2005. ISBN: 0072386886.

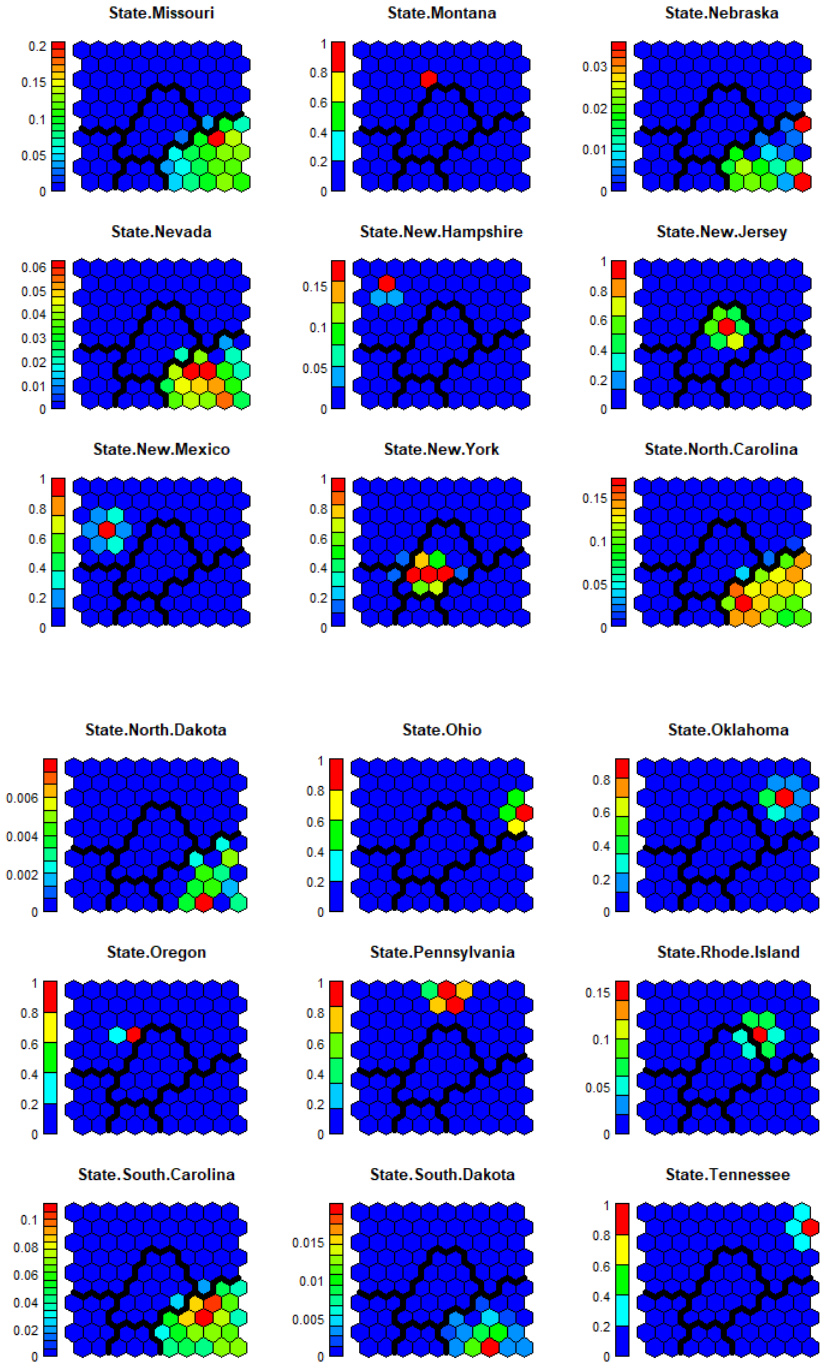
[2] Andries P Engelbrecht. *Computational intelligence : an introduction.* eng. Chichester, England: J. Wiley Sons, 2005. ISBN: 0470848707.

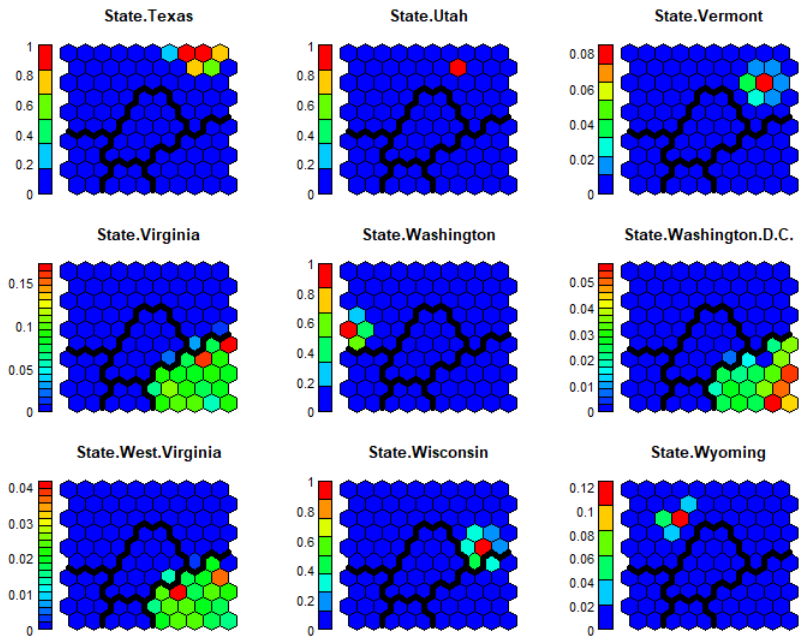
[3] Teuvo Kohonen. *Self-organizing maps.* eng. Springer series in information sciences ; 30. Berlin: Springer, 1995. ISBN: 3540586008.

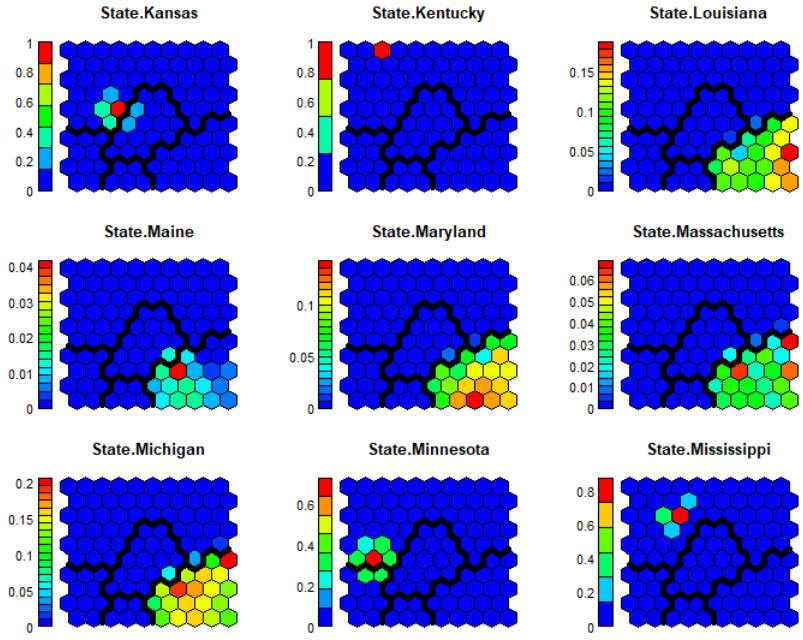
[4] Ron Wehrens and Lutgarde M. C. Buydens. "Self- and Super-Organizing Maps in R: The kohonen Package". In: *Journal of Statistical Software* 21.5 (2007), pp. 1–19. DOI: 10.18637/jss.v021.i05.

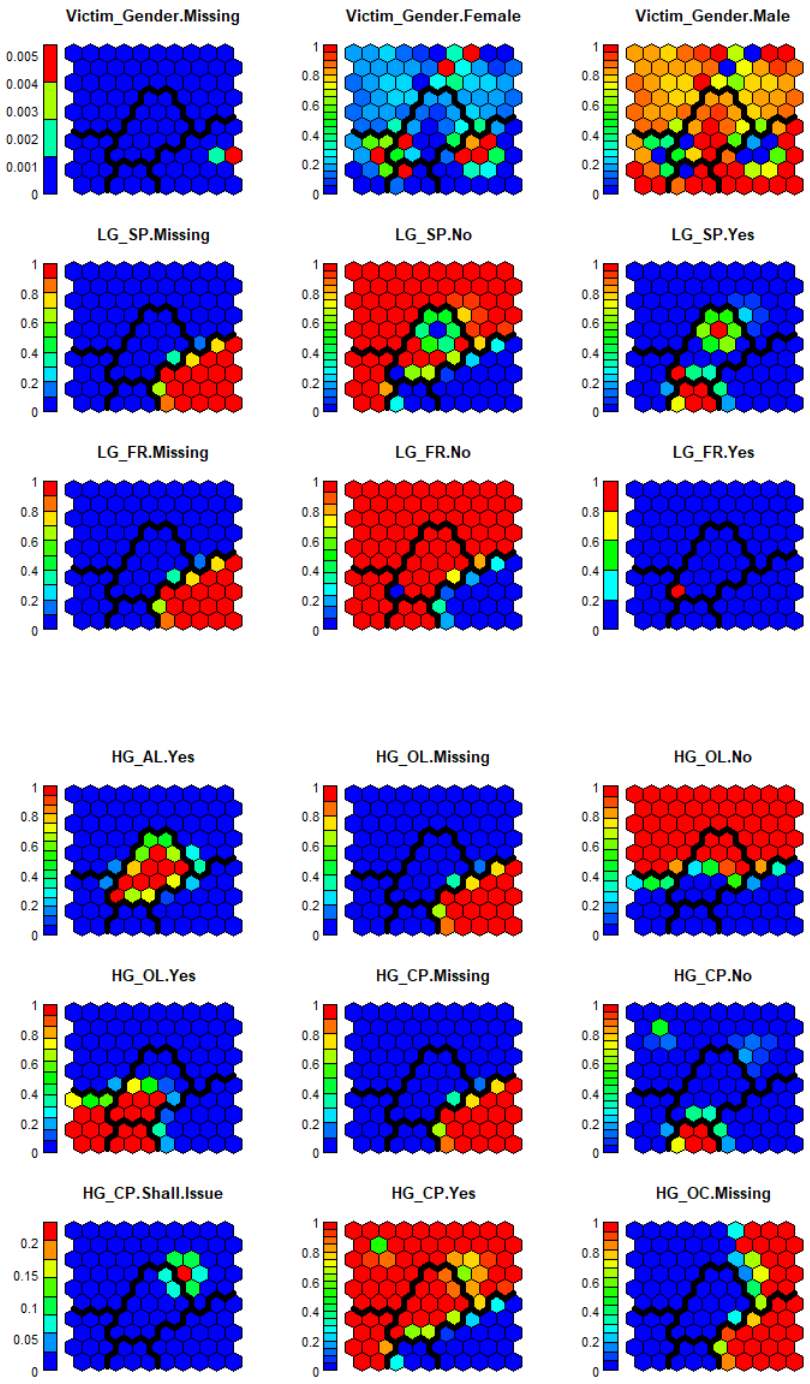
A Component Map Outputs

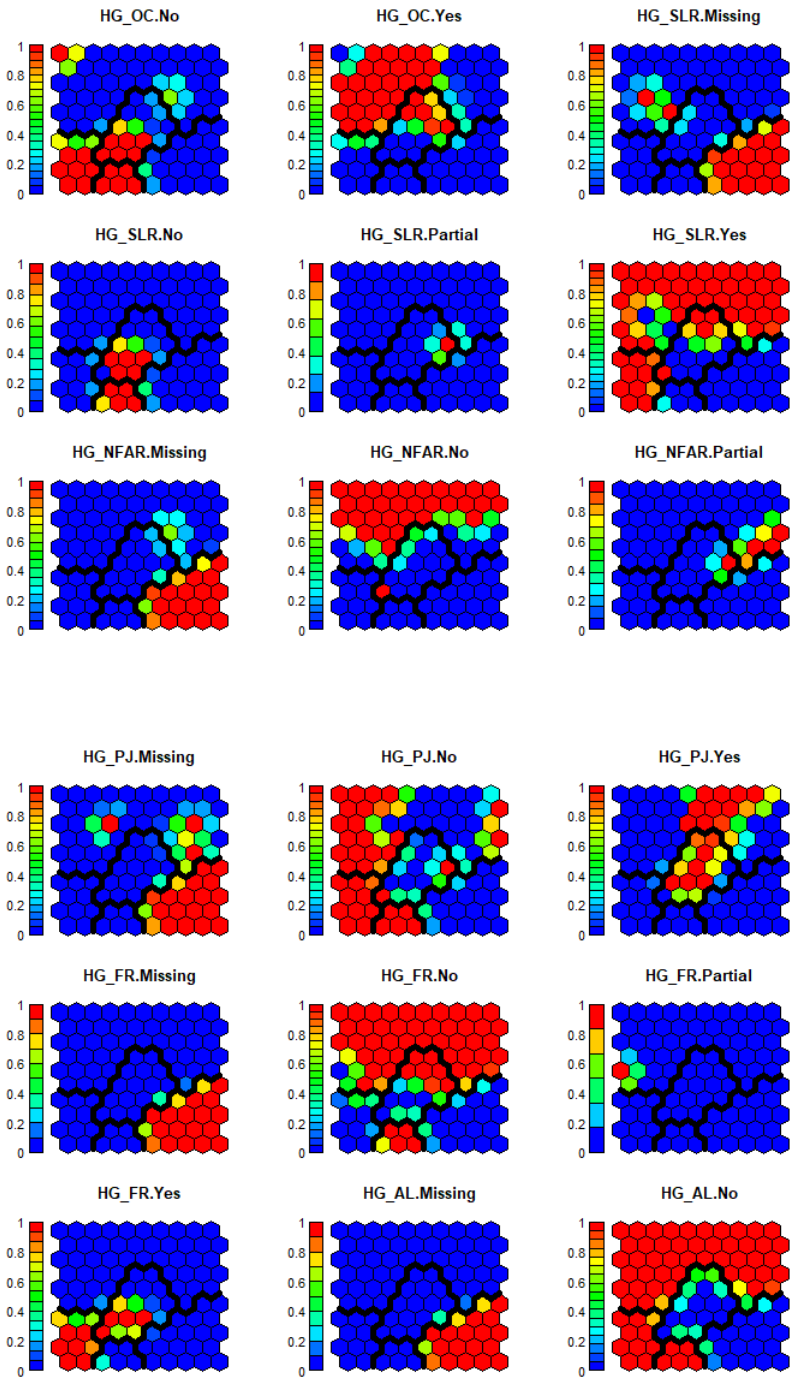


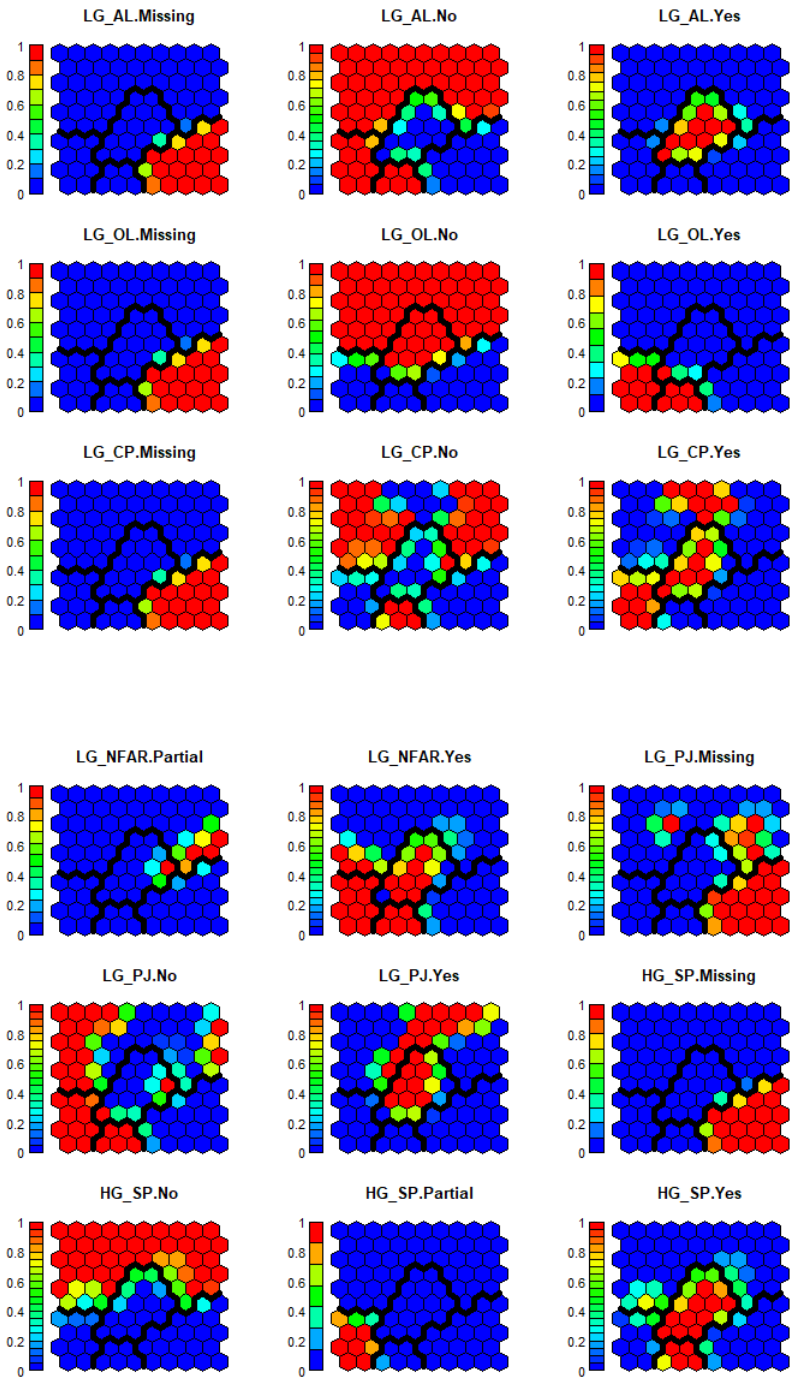

State.Missouri
State.Montana
State.Nebraska
State.Nevada
State.New.Hampshire
State.New.Jersey
State.New.Mexico
State.New.York
State.North.Carolina
State.North.Dakota
State.Ohio
State.Oklahoma
State.Oregon
State.Pennsylvania
State.Rhode.Island
State.South.Carolina
State.South.Dakota
State.Tennessee

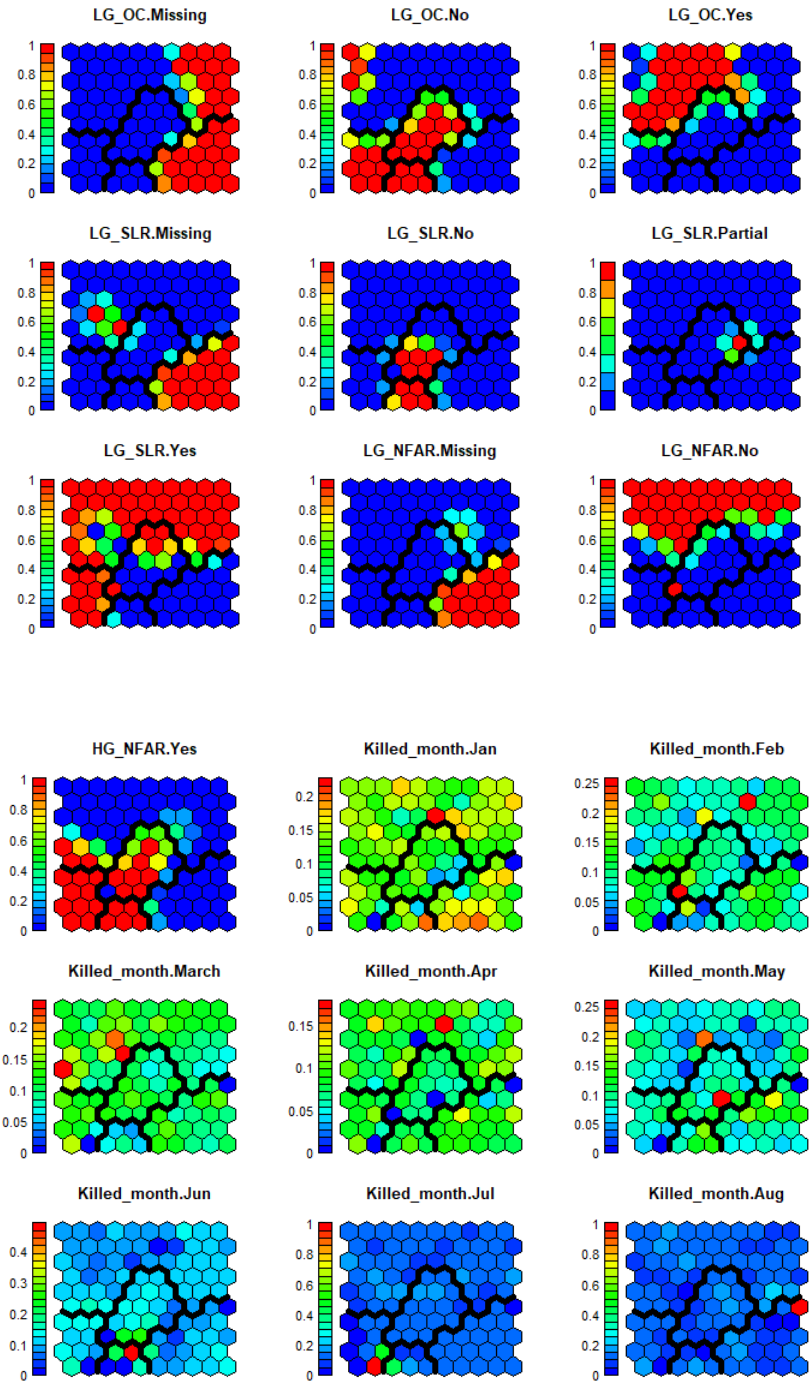

State.Texas
State.Utah
State.Vermont
State.Virginia
State.Washington
State.Washington.D.C.
State.West.Virginia
State.Wisconsin
State.Wyoming

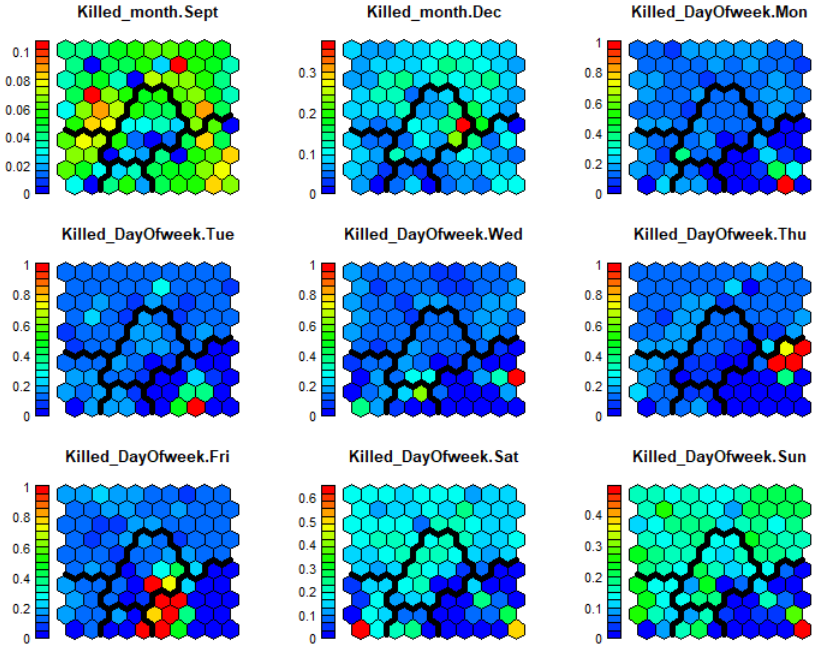

State.Kansas
State.Kentucky
State.Louisiana
State.Maine
State.Maryland
State.Massachusetts
State.Michigan
State.Minnesota
State.Mississippi


Victim_Gender.Missing
Victim_Gender.Female
Victim_Gender.Male
LG_SP.Missing
LG_SP.No
LG_SP.Yes
LG_FR.Missing
LG_FR.No
LG_FR.Yes
HG_AL.Yes
HG_OL.Missing
HG_OL.No
HG_OL.Yes
HG_CP.Missing
HG_CP.No
HG_CP.Shall.Issue
HG_CP.Yes
HG_OC.Missing


HG_OC.No
HG_OC.Yes
HG_SLR.Missing
HG_SLR.No
HG_SLR.Partial
HG_SLR.Yes
HG_NFAR.Missing
HG_NFAR.No
HG_NFAR.Partial
HG_PJ.Missing
HG_PJ.No
HG_PJ.Yes
HG_FR.Missing
HG_FR.No
HG_FR.Partial
HG_FR.Yes
HG_AL.Missing
HG_AL.No
1
0.8
0.6
0.4
0.2
0


LG_AL.Missing
LG_AL.No
LG_AL.Yes
LG_OL.Missing
LG_OL.No
LG_OL.Yes
LG_CP.Missing
LG_CP.No
LG_CP.Yes
LG_NFAR.Partial
LG_NFAR.Yes
LG_PJ.Missing
LG_PJ.No
LG_PJ.Yes
HG_SP.Missing
HG_SP.No
HG_SP.Partial
HG_SP.Yes
1
0.8
0.6
0.4
0.2
0


LG_OC.Missing
LG_OC.No
LG_OC.Yes
LG_SLR.Missing
LG_SLR.No
LG_SLR.Partial
LG_SLR.Yes
LG_NFAR.Missing
LG_NFAR.No
HG_NFAR.Yes
Killed_month.Jan
Killed_month.Feb
Killed_month.March
Killed_month.Apr
Killed_month.May
Killed_month.Jun
Killed_month.Jul
Killed_month.Aug


Killed_month.Sept
Killed_month.Dec
Killed_DayOfweek.Mon
Killed_DayOfweek.Tue
Killed_DayOfweek.Wed
Killed_DayOfweek.Thu
Killed_DayOfweek.Fri
Killed_DayOfweek.Sat
Killed_DayOfweek.Sun

B States and Corresponding Clusters

Attribute		Clusters				
		Purple	**Red**	**Blue**	**Yellow**	**Green**
	Victim Age	-	-	-	-	-
Victim Gender	Missing	-	-	-	-	-
	Male	-	-	-	-	-
	Female	-	-	-	-	-
State Statistics	State Population					
	Background Checks					
	Alaska					
	Arizona					
	Arkansas					
	Colorado					
	Delaware					
	Florida					
	Georgia					
	Idaho					
	Indiana					
	Iowa					
	Kansas					
	Kentucky					
	Minnesota					
	Mississippi					
	Montana					
	New Hampshire					
	New Mexico					
	Ohio					
	Oklahoma					
	Oregon					
	Pennsylvania					
	Rhode Island					
	Tennessee					
	Texas					
	Utah					
	Vermont					
	Washington					
	Wisconsin					
	Wyoming					
	California					
	Connecticut					
	Hawaii					
	New Jersey					
	New York					
	Illinois					

	Alabama					
	Louisiana					
	Maine					
	Maryland					
	Machusetts					
	Michigan					
	Missouri					
	Nebraska					
	Nevada					
	North Carolina					
	North Dakota					
	South Carolina					
	South Dakota					
	Virginia					
	Washington D.C.					
	West Verginia					

Table 15: States and Victim information and their clusters

C Gun Laws & Corresponding Clusters

			Clusters				
	Attribute Laws		Purple	Red	Blue	Yellow	Green
State Permit to Purchace	Handgun	Missing					x
		Partial		x			
		Yes			x	x	
		No	x				
	Rifle	Missing					x
		Yes			x	x	
		No	x	x	x		
Firearm Registration	Handgun	Missing					x
		Partial	x				
		Yes		x	x		
		No	x		x	x	
	Rifle	Missing					x
		Yes			x		
		No	x	x	x	x	
Assult Weapon Law	Handgun	Missing					x
		Yes		x	x		
		No	x	x		x	
	Rifle	Missing					x
		Yes			x		
		No	x	x		x	
Owner License Required	Handgun	Missing					x
		Yes		x	x	x	
		No	x		x		
	Rifle	Missing					x
		Yes		x	x	x	
		No	x		x		
Carry Permits Issued	Handgun	Shall Issue	x				
		Missing					x
		Yes	x	x	x		
		No				x	
	Rifle	Missing					x
		Yes	x	x	x		
		No	x			x	
Open Carry	Handgun	Missing	x				x
		Yes	x		x		
		No	x	x	x	x	
	Rifle	Missing	x				x
		Yes	x				
		No	x	x	x	x	
		Missing	x				x

State Preemption of local restrictions	**Handgun**	**Partial**			x		
		Yes	x	x	x		
		No			x	x	
	Rifle	**Missing**	x				x
		Partial			x		
		Yes	x	x	x		
		No			x	x	
NFA weapons restricted	**Handgun**	**Missing**					x
		Partial	x		x		
		Yes	x	x	x	x	
		No	x		x		
	Rifle	**Missing**					x
		Partial	x		x		
		Yes	x	x	x	x	
		No	x		x		
Peaceable Journey Laws	**Handgun**	**Missing**	x				x
		Yes	x		x		
		No	x	x	x	x	
	Rifle	**Missing**	x				x
		Yes	x		x		
		No	x	x	x	x	

Table 16: Gun laws and to which clusters they belong

D Date information & Corresponding Clusters

Attribute		Clusters				
		Purple	Red	Blue	Yellow	Green
Month Killed	Jan	-	-	-	-	-
	Feb	-	-	-	-	-
	Mar	-	-	-	-	-
	Apr	-	-	-	-	-
	May	-	-	-	-	-
	Jun	-	-	-	-	-
	Jul	-	-	-	-	-
	Aug	-	-	-	-	-
	Sep	-	-	-	-	-
	Dec	-	-	-	-	
Day of the week killed	Mon	-	-	-	-	-
	Tue	-	-	-	-	-
	Wed	-	-	-	-	-
	Thu	-	-	-	-	-
	Fri	-	-	-	-	-
	Sat	-	-	-	-	-
	Sun	-	-	-	-	-

Table 17: Date Deaths Occurred