

What are the streets with the most crime? What type of crime and when do they happen?

Removed all records with missing location data because the impact of these missing locations was so significant that all the results ended in missing data.

To find the location with the most crime, by type of crime, hour, and day of the week a Simple K Means algorithm was used. Multiple runs were made using different numbers of clusters. The result using 3 clusters was that Washington St had the most crime for all UCR parts. Surprisingly most incidents happen during daylight between 11 am and 2 pm and between Sunday, Tuesday, and Wednesday. Part 1 crimes are the most serious type of crime happening Tuesdays around 1 pm.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer

Choose SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots

Cluster mode

Use training set
 Supplied test set Set...
 Percentage split % 66
 Classes to clusters evaluation
(Nom) STREET
 Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 12:23:55 - SimpleKMeans
- 12:24:46 - SimpleKMeans
- 12:27:03 - SimpleKMeans
- 12:27:17 - SimpleKMeans
- 12:27:53 - SimpleKMeans
- 12:45:51 - SimpleKMeans
- 12:47:17 - SimpleKMeans**
- 12:47:35 - SimpleKMeans
- 12:47:59 - SimpleKMeans
- 12:52:43 - SimpleKMeans
- 12:55:11 - SimpleKMeans
- 12:55:35 - SimpleKMeans

Clusterer output

```
=== Clustering model (full training set) ===

kMeans
=====

Number of iterations: 4
Within cluster sum of squared errors: 130352.45027549368

Initial starting points (random):

Cluster 0: Wednesday,11,'Part Three','PARK PLZ'
Cluster 1: Sunday,3,'Part One','HUNTINGTON AVE'
Cluster 2: Tuesday,8,'Part One','BLACKWOOD ST'

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data      Cluster#
                    (69260.0)      (33845.0)      (19826.0)      (15589.0)
=====
DAY_OF_WEEK        Friday  Wednesday  Sunday  Tuesday
HOUR                13.0837  14.1563   11.314   13.0056
UCR_PART            Part Three  Part Three  Part Two  Part One
STREET              WASHINGTON ST WASHINGTON ST WASHINGTON ST WASHINGTON ST

Time taken to build model (full training data) : 0.13 seconds

=== Model and evaluation on training set ===

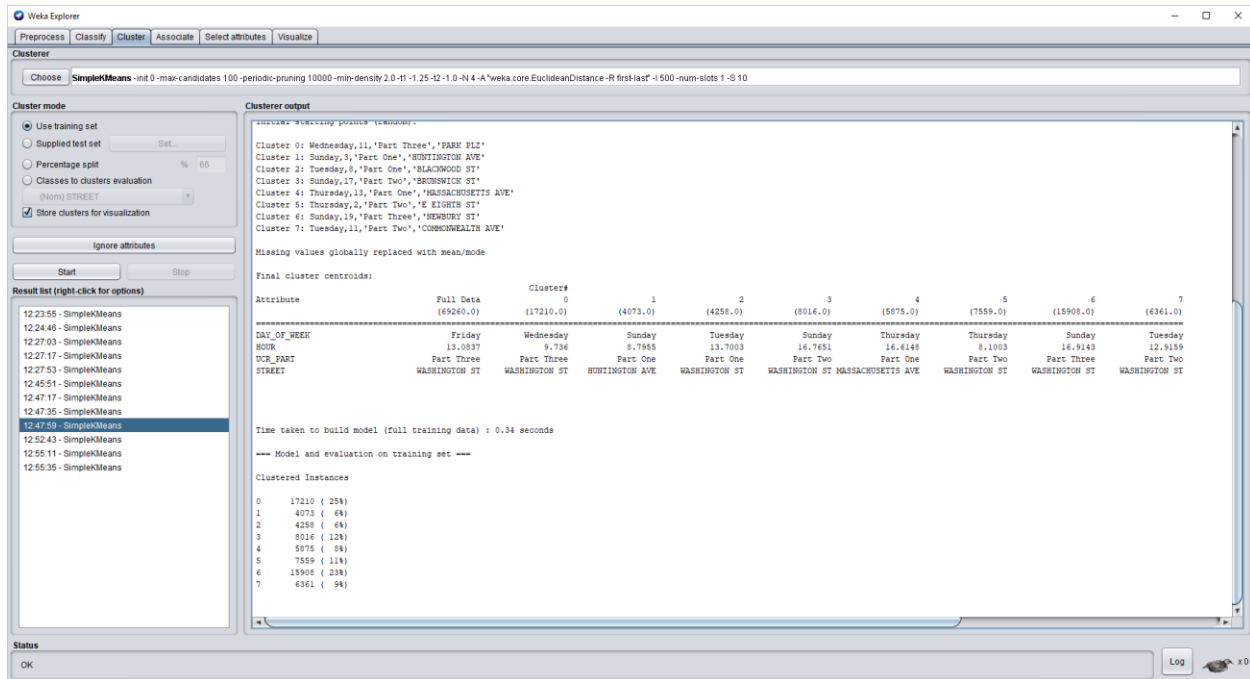
Clustered Instances

0      33845 ( 49%)
1      19826 ( 29%)
2      15589 ( 23%)
```

Status

OK Log x0

After running the algorithm with 8 clusters, the results were somewhat similar since Washington St appeared in 6 out of the 8 clusters with the highest number of records. Again, most crimes seem to happen during daylight and on weekdays.



When and where do crimes involving shootings happen?

***After removing all records with No shooting, the dataset was reduced to 210. This was done to only focus our attention on records where there was a shooting and find out the day, hour, month, and street where the shootings occurred. By keeping the data where there was no shooting, all clusters would result in no shooting clusters.

Clusters for the shooting data with 5 clusters were distributed thru 5 different streets. Most shootings occurred during the months of March to June, the shootings happened mostly during daylight and were distributed between weekdays and weekends. It is of no surprise that the shootings mostly happened during months of warmer weather, rather than during the winter.

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Clusterer

Choose: SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance -R first-last" -l 500 -num-slots 1 -S 10

Cluster mode

Use training set
 Supplied test set (Set...)
 Percentage split % 66
 Classes to clusters evaluation (Nom) STREET
 Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 12:23:55 - SimpleKMeans
- 12:24:46 - SimpleKMeans
- 12:27:03 - SimpleKMeans
- 12:27:17 - SimpleKMeans
- 12:27:53 - SimpleKMeans
- 12:45:51 - SimpleKMeans
- 12:47:17 - SimpleKMeans
- 12:47:35 - SimpleKMeans
- 12:47:59 - SimpleKMeans
- 12:52:43 - SimpleKMeans
- 12:55:11 - SimpleKMeans
- 12:55:35 - SimpleKMeans
- 13:07:15 - SimpleKMeans
- 13:10:21 - SimpleKMeans
- 13:10:37 - SimpleKMeans

Clusterer output

```

=== Clustering model (full training set) ===

KMeans
=====
Number of iterations: 5
Within cluster sum of squared errors: 272.0919101796333

Initial starting points (random):

Cluster 0: 5, Friday, 5, 'WAYLAND ST'
Cluster 1: 9, Sunday, 3, 'W NEWTON ST'
Cluster 2: 6, Friday, 23, 'BORDER ST'
Cluster 3: 7, Monday, 21, 'MASSACHUSETTS AVE'
Cluster 4: 9, Thursday, 17, 'QUINCY ST'

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute          Full Data          Cluster#          1          2          3          4
(209.0)          (50.0)          (24.0)          (45.0)          (64.0)          (26.0)
-----
MONTH              5.4258             6.14             4.9167          4.9333          6.3594          3.0769
DAY_OF_WEEK        Saturday           Saturday         Sunday           Friday           Monday           Thursday
HOUR               14.2488            8.44             11.4583         16.4667         17.2656         16.7308
STREET             WASHINGTON ST      BRACKETT ST      MILLET ST        WASHINGTON ST     NORFOLK ST       WINTER ST

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      50 ( 24%)
1      24 ( 11%)
2      45 ( 22%)
3      64 ( 31%)
4      26 ( 12%)
  
```

Status

Classifying shooting data

Using hour, month, day of the week, and street to classify the possibility of a shooting showed a 75% correct classification but the rules were very hard to obtain because of the large number of Street names. The tree is difficult to read.

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: Choose **J48 - C 0.1 - M 2**

Test options

Use training set
 Supplied test set (Set...)
 Cross-validation Folds **10**
 Percentage split % **66**
 More options...

(Nom) HOUR

Start Stop

Result list (right-click for options)

- 13:16:48 - trees.J48
- 13:17:07 - trees.J48
- 13:17:24 - trees.J48
- 13:17:34 - trees.J48
- 13:17:52 - trees.J48
- 13:18:21 - trees.J48
- 13:19:10 - trees.J48
- 13:19:22 - trees.J48
- 13:19:28 - trees.J48
- 13:19:40 - trees.J48
- 13:19:47 - trees.J48
- 13:19:54 - trees.J48
- 13:20:08 - trees.J48
- 13:24:10 - trees.J48
- 13:24:21 - trees.J48
- 13:26:51 - trees.J48
- 13:26:55 - trees.J48
- 13:26:58 - trees.J48

Classifier output

```

STREET = MCNULTY CT: Sunday (2.05/0.05)
STREET = BORDER ST: Friday (5.12/0.12)

Number of Leaves : 110
Size of the tree : 123

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 158 75.5981 %
Incorrectly Classified Instances 51 24.4019 %
Kappa statistic 0.703
Mean absolute error 0.0866
Root mean squared error 0.2184
Relative absolute error 36.0105 %
Root relative squared error 62.988 %
Total Number of Instances 209

=== Detailed Accuracy By Class ===

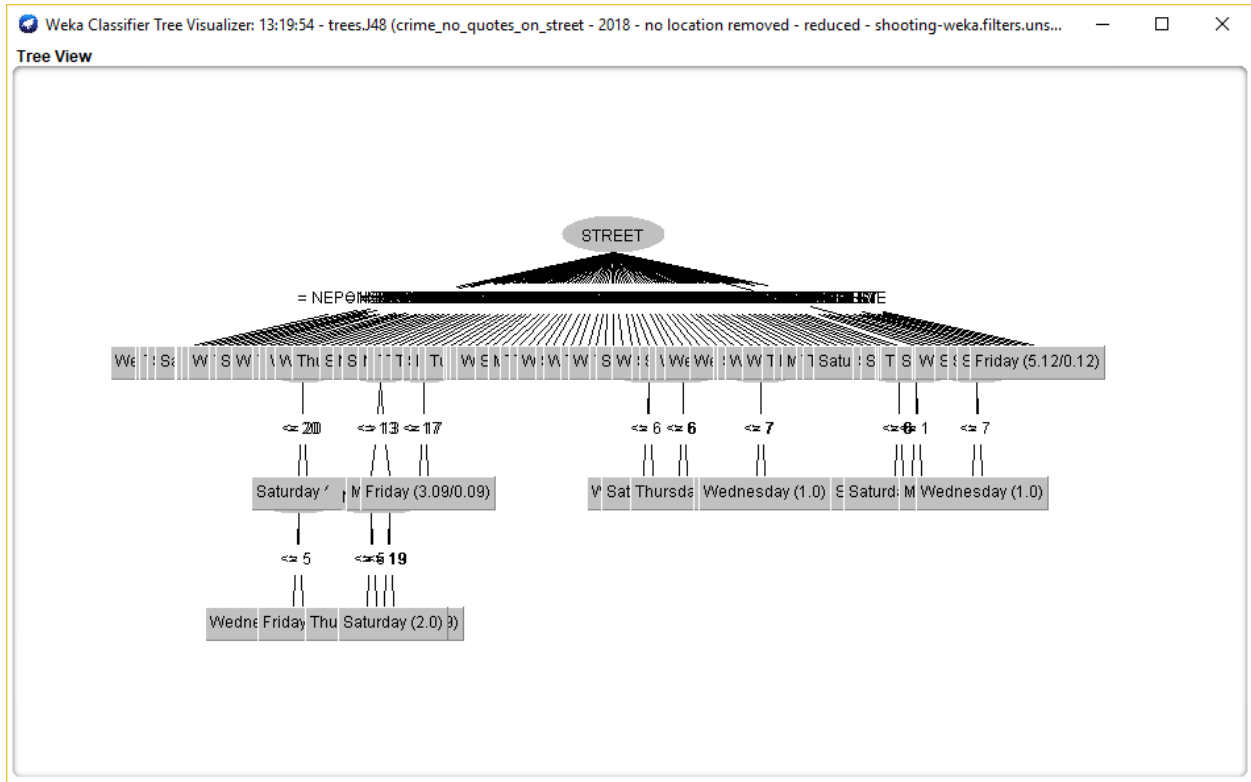
          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
          -----  -----  -
          0.958  0.255  0.529  0.958  0.681  0.600  0.931  0.816  Saturday
          0.750  0.012  0.931  0.750  0.831  0.807  0.902  0.816  Wednesday
          0.591  0.005  0.929  0.591  0.722  0.719  0.933  0.670  Sunday
          0.444  0.005  0.889  0.444  0.593  0.607  0.910  0.529  Tuesday
          0.935  0.011  0.935  0.935  0.935  0.924  0.986  0.917  Friday
          0.657  0.011  0.920  0.657  0.767  0.743  0.911  0.774  Monday
          0.632  0.011  0.857  0.632  0.727  0.714  0.930  0.755  Thursday

Weighted Avg. 0.756 0.066 0.827 0.756 0.760 0.731 0.929 0.778

=== Confusion Matrix ===

 a  b  c  d  e  f  g  <-- classified as
46  0  0  0  0  0  2 | a = Saturday
 5 27  1  0  2  1  0 | b = Wednesday
 9  0 13  0  0  0  0 | c = Sunday
10  0  0  8  0  0  0 | d = Tuesday
 2  0  0  0 29  0  0 | e = Friday
10  1  0  1  0 23  0 | f = Monday
 5  1  0  0  0  1 12 | g = Thursday
  
```

Status



Results without the Street data and using only hour, day of the week, and month, resulted in 58% correct classification, therefore these are not reliable.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose **J48 -C 0.1 -M 2**

Test options

Use training set
 Supplied test set (Set...)
 Cross-validation Folds **10**
 Percentage split % **66**
 More options...

(Nom) HOUR

Start Stop

Result list (right-click for options)

- 13:16:48 - trees.J48
- 13:17:07 - trees.J48
- 13:17:24 - trees.J48
- 13:17:34 - trees.J48
- 13:17:52 - trees.J48
- 13:18:21 - trees.J48
- 13:19:10 - trees.J48
- 13:19:22 - trees.J48
- 13:19:28 - trees.J48
- 13:19:40 - trees.J48
- 13:19:47 - trees.J48
- 13:19:54 - trees.J48
- 13:20:08 - trees.J48
- 13:24:10 - trees.J48
- 13:24:21 - trees.J48
- 13:26:51 - trees.J48**
- 13:26:55 - trees.J48
- 13:26:58 - trees.J48

Classifier output

```

Number of Leaves : 110
Size of the tree : 122
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 123      58.8517 %
Incorrectly Classified Instances 86      41.1483 %
Kappa statistic 0.5104
Mean absolute error 0.1237
Root mean squared error 0.2848
Relative absolute error 51.4334 %
Root relative squared error 82.1376 %
Total Number of Instances 209

=== Detailed Accuracy By Class ===
          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
          0.688  0.106  0.660  0.688  0.673  0.574  0.834  0.661  Saturday
          0.593  0.075  0.618  0.593  0.600  0.520  0.897  0.647  Wednesday
          0.455  0.080  0.400  0.455  0.426  0.354  0.796  0.409  Sunday
          0.333  0.031  0.500  0.333  0.400  0.364  0.869  0.472  Tuesday
          0.710  0.084  0.595  0.710  0.647  0.582  0.944  0.767  Friday
          0.571  0.052  0.690  0.571  0.625  0.561  0.834  0.676  Monday
          0.579  0.058  0.500  0.579  0.537  0.488  0.851  0.507  Thursday
Weighted Avg.  0.589  0.075  0.592  0.589  0.587  0.515  0.862  0.620

=== Confusion Matrix ===
 a b c d e f g <-- classified as
33 2 5 0 5 1 2 | a = Saturday
 4 21 1 0 2 7 1 | b = Wednesday
 4 2 10 3 0 1 2 | c = Sunday
 2 2 5 6 2 0 1 | d = Tuesday
 3 0 0 2 22 0 4 | e = Friday
 3 6 3 1 1 20 1 | f = Monday
 1 1 1 0 5 0 11 | g = Thursday
  
```

Status

OK Log x0

To run the Apriori algorithm the Month attribute was changed from Numeric to Nominal using the NumericToNominal filter in Weka.