OK here is the data I have. Please take a look. I am not familiar with Caret. I don’t know what it is … Weka or MATLAB by itself could be more than enough for you. You just need a package that has clustering algorithms in them. From my perspective a simple K-means clustering algorithm would do the job. However, this project is interesting mainly because of the potentially interesting patterns you could discover.

I have spent two weeks on this data a few months ago. Already aware of some interesting things we could do with this data. Three things I want to do:

1.      Cluster all crime vectors across all years and cities to identify crime patterns and manually analyze them to see if they make sense. Which crime pattern is least desirable, which one is relatively more tolerable? Do the crime patterns assigned to cities we know make sense? If not then can we adjust clustering to more sensible assignments?

2.      Once we make sure the number and the type of crime patterns make sense we can investigate temporal crime trends across specific cities to identify interesting patterns.

a.      For example there are big cities with 100K+ populations that has the same crime patterns as they did 20 years ago when their populations was less than 10K. These are cities that developed healthily. On the other hand there are cities which switched to worse crime patterns as they develop. We can identify these cities if we look into temporal pattern assignments. What method can we use there?

b.      There are big cities (population flat) which changed for better or worse after certain years. Can we identify those cities? Can we look into some external factors that may cause this change?

I think this would a good project. It has a nice data science flavor but more interestingly it opens the door for you to discover potentially interesting crime trends.

First steps from Professor Tooman:

I think before you jump into the mean's clustering you will want to start with calculating some summary statistics so that you get an idea of the center/shape and range/variability of each variable. Then you can start adding plots. Individually you can create histograms to assess the shape and look for normality/skewness/outliers. The side-by-side box plots can help you look at differences between the means of the variables. But I think what may be more helpful is to look at scatter plots for correlation between your variables.  
  
Then once you start having an idea of the relationships between the variables, begin working on clustering...then potential new variables or prediction.