The Project Structure

1. Start analyzing the data
   1. “Play with the data”, remember that cities and states can have the same name so you can find the uniquely with the joint condition when searching.
2. Learn Kmeans clustering

My proposal will be typically a document I submit which is a couple pages in length explaining the path I want to take in the course of the project. **With slightly more detail on each point**, the proposal should go along the lines of the following email he sent me:

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 will also discuss “*open source*” tools that I will use to accomplish my data science related project. He suggested vL-Flats for this in MATLAB. Then in a conclusion paper at the end of the semester he basically said that an analysis of alternatives to Kmeans that might do better might take place in what went well and predicted nicely and what didn’t. This is where he cut me off and didn’t want to give me any more information.

Data Set Structure

Five different variables named:

* A (Agency – City; Note, some cities and states have the same names.)
* G (Year)
* P (Population – The integer sequence is as follows:)
  + 1 🡪1980
  + 2 🡪1981
  + …
  + 35 🡪 2014
* S (States – See the codes in the handout)
* X (Crime Stats – The listing of values for this are in order as the following:)
  + Murder, manslaughter, raped total, robbery total, assault total, burglary total, larceny total, auto theft, Number of offices killed, Number of officers assaulted.