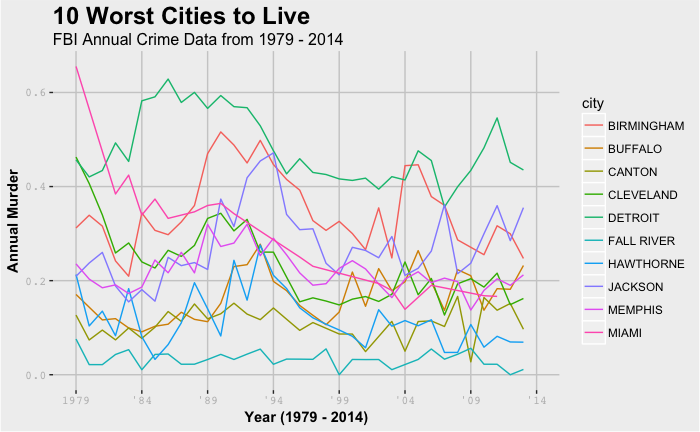
Blake Conrad

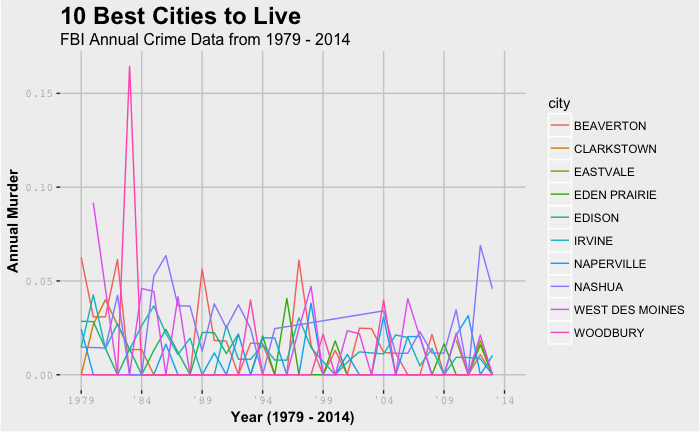
CSCI 49500

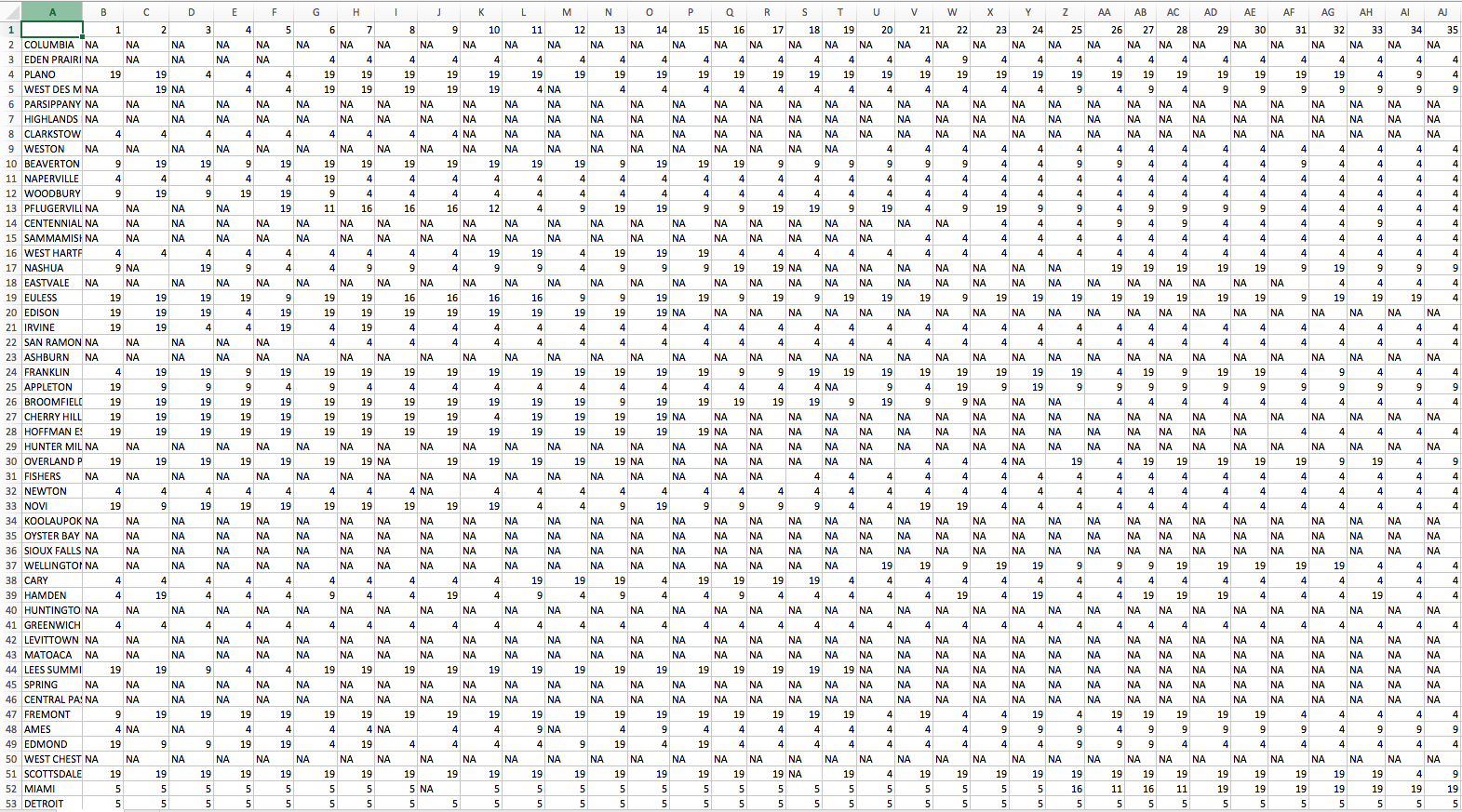
*Midterm Progress Update*

Advisor: Murat Dundar

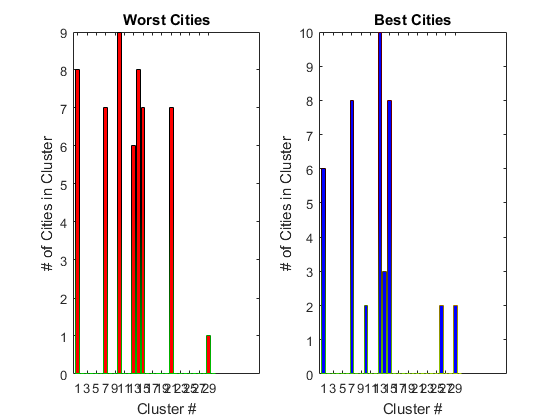
* I am on track as expected with a pretty sound understanding of my data and how to determine my K value
* My initial hump
  + Let’s just use USA Today and do the top 10, include all of my images
    - How Dundar said this is not good because population skews things
    - The media data I found for top 10 worst and best is also population dependent, we don’t want this because there are small cities with bad crime patterns as well as big cities with good crime patterns, the data is already population centered
  + As a way of checking my work I am confirming that the best 50 and worst 50 are in different clusters uniformly
  + As another way of checking my work I want to apply PCA to my data in a high dimensional vector space into a lower dimensional vector space (one that is geometrically interpretable)
* The switch from MATLAB to R because of the flexibility of ggplot2 open source library for plotting.
* The updated analysis links on finding K *[source for 50 best] [source for 50 worst]* both of which are population independent.
* How this is a better strategy of finding legitimate K without any skewedness associated to it





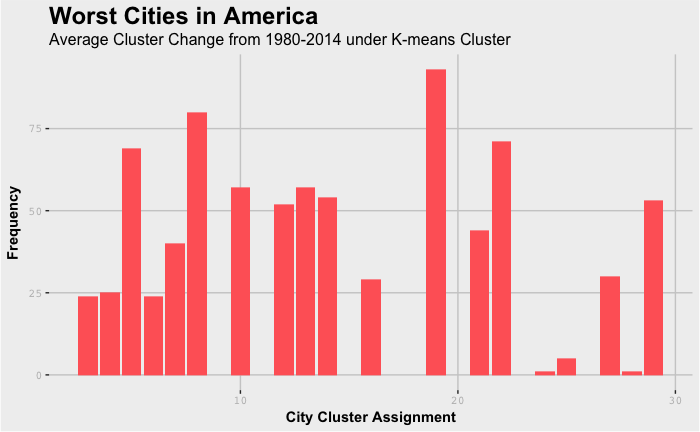
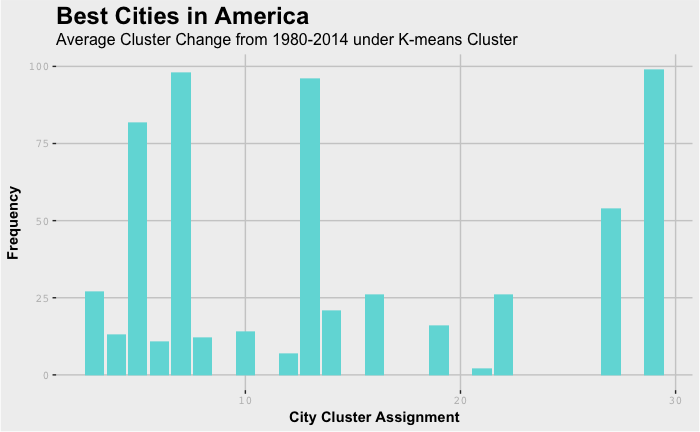


* Top 50 rows are the best cities to live
* Last 50 rows are the worst cities to live in
* Columns from 1-35 represent 1979 to 2014
* Each cell represents the city’s crime pattern cluster assignment for a given year



When I was using matlab, the best I could get for the axis;

Note, harder to read and looks less aesthetically appealing.

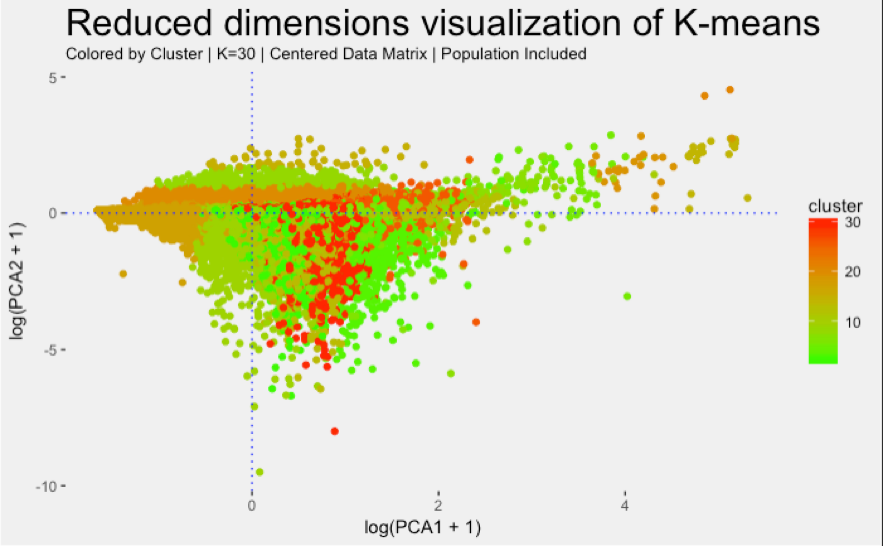


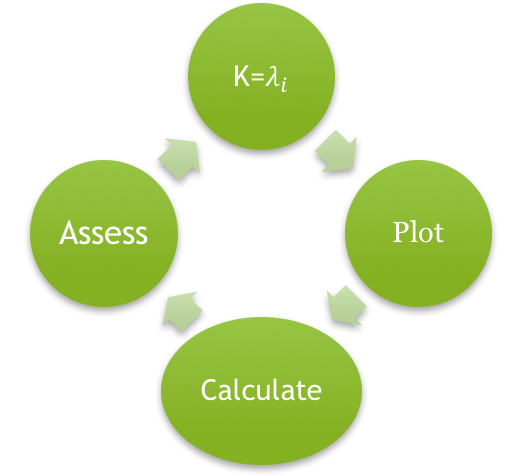
Once converting to R I had a much more broad ability to plot cleaner images that I could scale easier

I have a functional PCA algorithm to reduce my clustering of my 11 dimensions down to 2 just to visualize and see if K is making sense.

I can’t move on with any more cool data mining techniques until I acquire a special K, though PCA helps to make sense of things, I am still experimenting with the data to find this good K, then I can progress with my end goals.

I have also looked into the tSNE dimensionality reduction algorithm which is highly respected from both math and computer science communities for dimensionality reduction, I may look into using that too to compare the results





Above is an image of where I am at in my progress.

I have brought in the data

Cleaned the data

Centered the data

Mingled & Plotted to understand the data itself

Now I am closing in on a good K, from my findings the best is either 20 or 30, but my analysis to back that up is yet to be confirmed. My academic advisor is not persuaded that I have enough evidence to assume that K is good at those values, so I am still digging.