Blake Conrad

CSCI 49500

*Midterm Progress Update*

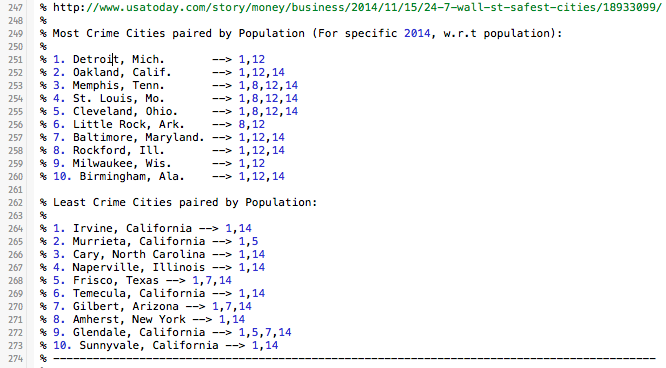
Advisor: Murat Dundar

With the *FBI Crime Data* provided by my faculty advisor Dr. Murat Dundar for this project, I have become more engaged and aware of the problems and goals of my project. I began my project in MATLAB as suggested my Dr. Dundar. Beginning my project, with respect to my deadlines, my hope was to have the *K* for my K-means algorithm established. With some roadblocks along the way, this has not yet been accomplished, however Dr. Dundar and I have agreed I am not far from finding a good *K choice* that is justifiable. With the nearness to the *K* being established, I am excited to say I am *almost* on the exact track I was expecting for my syllabus created in the first assignment. I am going to discuss how I chose to attack this problem initially, the techniques and pipelines I was able to set up, corrections to the initial setbacks, and finally the current state and direction of the project at large.

My initial problem for my project was to establish a good *K* for K-means, which is a non-trivial task. This means I needed evidence to support that I chose a good *K*. There were only a couple valid options on how I was to choose such a *K*. The first option was to confirm it with cities I know, some of these are *Carmel, Fishers, Greenwood,* and *Indianapolis*. This was an alright option, but I really don’t know enough cities to draw good relationships from. The second option was to look at what the media had to say about the best and worst cities and run my algorithm checking for the maximum differences. Applying the second strategy, I leveraged media data from [USA Today](http://www.usatoday.com/story/money/business/2014/11/15/24-7-wall-st-safest-cities/18933099/) which had the 50 safest and 50 most unsafe cities in the U.S. A view at the top 10 best cities and worst cities are in *Figure 1.*

*Figure 1*

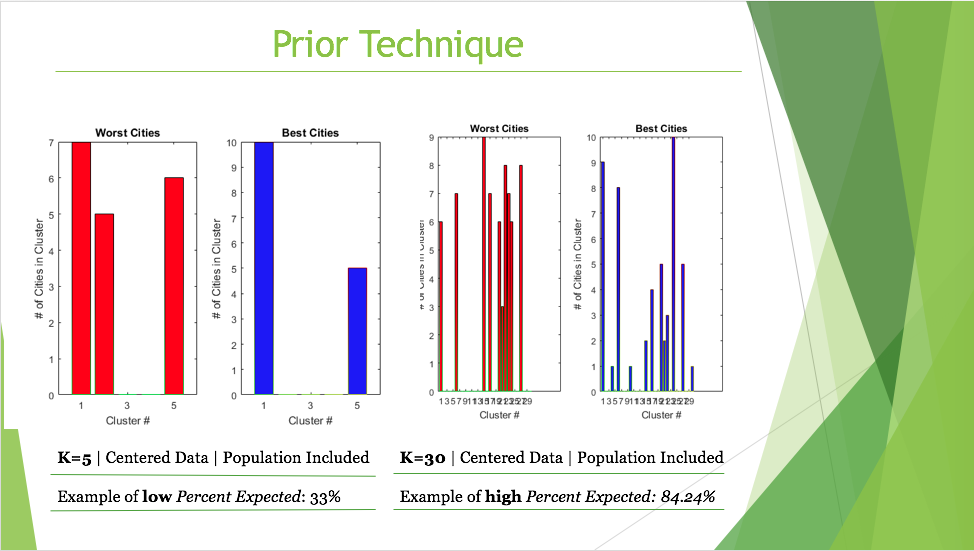
The top 10 best and worst cities to live in the U.S.



With this, I applied K-means with *K=5, 10, 15, 20, 25, and 30*. I recorded images of how many of the best and worst cities landed in each cluster and how many, then found the ratio between how many of the best cities in cluster and the worst cities in that same cluster . This allowed for a simple calculation to show the *Expected Number* of overlapping cities. My baseline expectation is to have overlap, because we know that the cities are from entirely different end goals; the best and the worst. An example of one of many of the calculations done by the algorithm is in *Figure 2.* I have also recently switched from MATLAB to R primarily due to the plotting flexibilities in R and vast literature in the open source community. An image of the identical analysis done in R is illustrated in *Figure 3.*

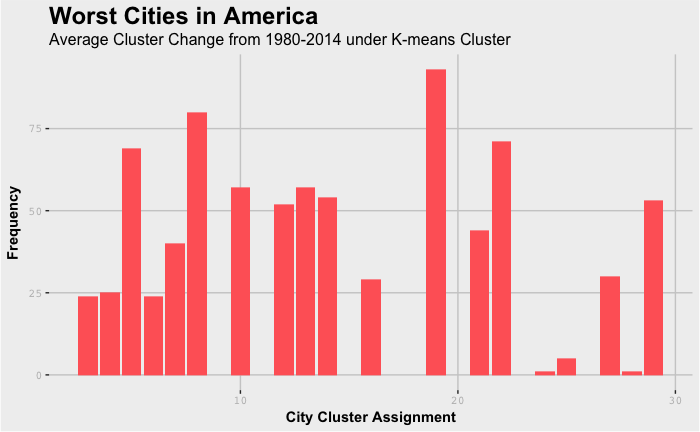
*Figure 2*

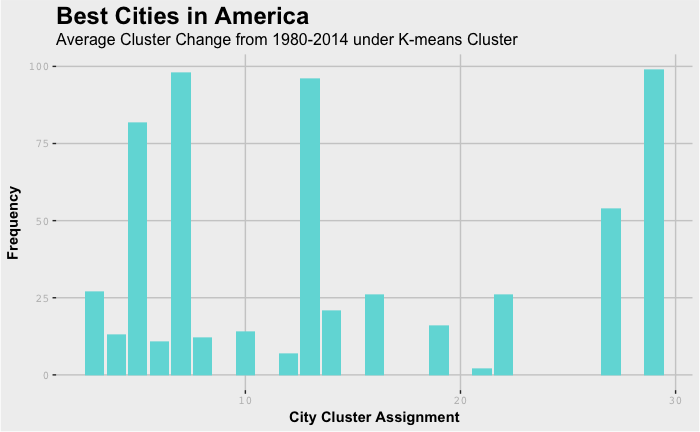
An example of 2 of the 10 different graphs done in the analysis.



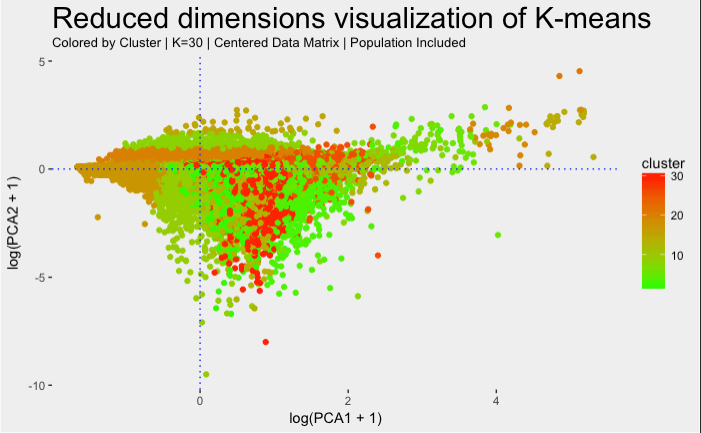
*Figure 3*

An example of analysis done in R using the ggplot library instead of the MATLAB plotting libraries. The hope was to make more eloquent illustrations in my final formal proposal.





This approach led to some great results, which led me to decide *K=*30 as my best choice. I was also excited to set up a *PCA* algorithm to confirm that *K* was a good selection in our high dimensional space. An image of such findings are illustrated in *Figure 4*.



*Figure 4*

A transformation on the collapsed dimension space in order to project the cluster crime data into a geometrically interpretable subspace of the original data set.

The project at large has been very successful in cleaning, formatting, visualizing, and testing. My only initial setback is all of the time I put in deciding a good *K* was skewed based on population while my analysis techniques have been updated as well. The two updated sources I have for the [top 50 best places to live](http://time.com/money/collection/best-places-to-live/) and [top 50 worst places to live](http://247wallst.com/special-report/2016/06/28/the-worst-cities-to-live-in/11/) are listed in the hyperlinks attached. These both give me a better listing on non-population skewed ways to check my work when applying the K-means. I have also opted to confirm my work with an excel file showing each city over the rows and years from 1979 – 2014 along the columns. This has a set up of 100 cities to view over time. By viewing the first 50 rows as the best cities clusters and the last 50 rows as the worst cities, I can track both of their cluster assignments over time and see if they make sense. I am currently dealing with just a few anomalies that shouldn’t exist, but am optimistic that within a week (or sooner) I will have this leveled out. Looking forward, I primarily want to diagnose cities as having healthy or unhealthy growth or decay. A critical element to this is the clustering assignments. My goal is to do this by looking at population and determining more favorable and less favorable crime patterns. Once this is established I will have a good system in place to detecting unintuitive patterns in the data.