

Successes:

* The differences in K based on the plots with population included are showing some significant differences.

Shortcomings:

Problems:

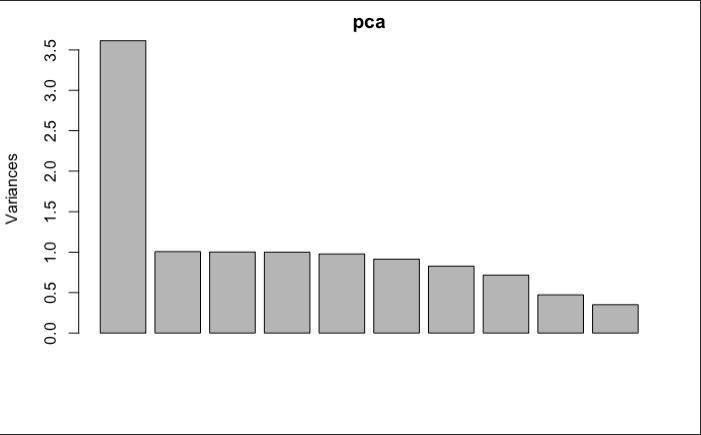
1. What K should be used?
2. Is K-means more predictive with or without population?
3. Is K-means more predictive on centered or un-centered data?
4. What is the definition of *error* between clusters?
5. Why are bad and good cities being clustered together so much?
6. Why are 5000 of my log(x+1) data points being dropped?

Solutions:

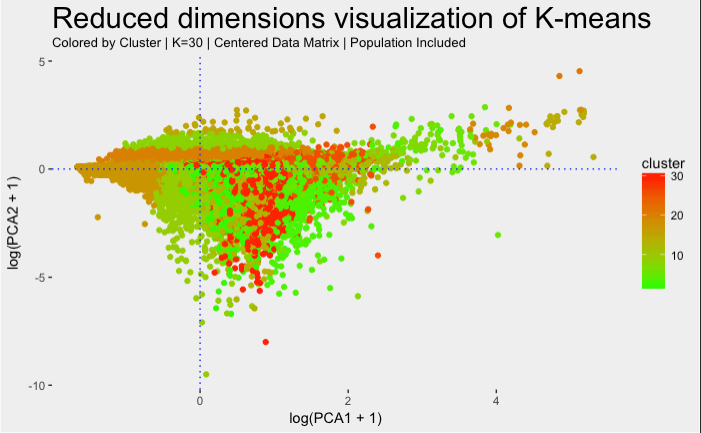
1. Plot and analyze 5 intervals to 25 to determine which gives us the *least* error, from k=5, 10, 15, 20, and 25. This will give an idea of which K is the best.
2. Plot the above 5 K’s with population and see how it changed from the original percentages based on our *least error*.
3. Plot the above 5 K’s with the data centered and see how it changed from the original percentages based on our least error.
4. Considering the data set is capped at 2014 data, I retrieved data from <http://www.usatoday.com> data to determine the 10 worst cities and 10 best cities to live (based on population from a 2014 statistic). With this axiom as the standard of *correct* and *incorrect*, my hope is to obtain a K such that these cities fall into the *most different* clusters. By manually checking to see the differences between clusters, I will select the best K such that the clusters are divided into K groups that are the most different based on crime patterns.
5. They are, but some values of more than others, when population was included, 3/12 figures showed an 80% expected difference ratio, but what about the 20%?

Unique findings:

1. Are safe populations actually as safe as they seem? The statistics I am finding may startle you based on the variables taken into account.
2. For our clusters without population *that are un-centered* we notice a peak at rising up of *expected percentages* up to the peak at K=20, then we see them lower back off again.
3. For our clusters without population that were first centered …
4. After analyzing the graphs, why is 20% of those clusters the same? Maybe *some* of the best cities aren’t as great as they say, or maybe *some* of the worst cities aren’t as bad as they say in media.
5. After my analysis, population is as valid a crime statistic at the population/city level to determining crime patterns when the data is scaled and centered.
6. After dimensionality reduction from 11 to 2, we see clusters are still maintained fairly well.
7. Currently am analyzing each cities cluster assignment over years in the email that Dundar sent me. This was a challenge, but I have a dataframe holding each city as an observation and each column as the year with the cell value as the cluster assignment.

**

1. Turns out a log(x+1) transformation preserved the data best upon visualization. Other attempts involved simple normalization, centering, and just log transforming; again log(x+1) transforming preserved the data and visualized the cluster *fairly* well.

**

New Analysis

***Without population un-centered***

In order to determine a graphs percentage, because we *expect* the two groups of 10 to land in *unique* clusters, we will call this percentage our *expected percentage*, while the groups of cities that land near one another we will call the *unexpected* percentage. The percentage part is obtained simply by dividing each number by the total number of cities in the sample.

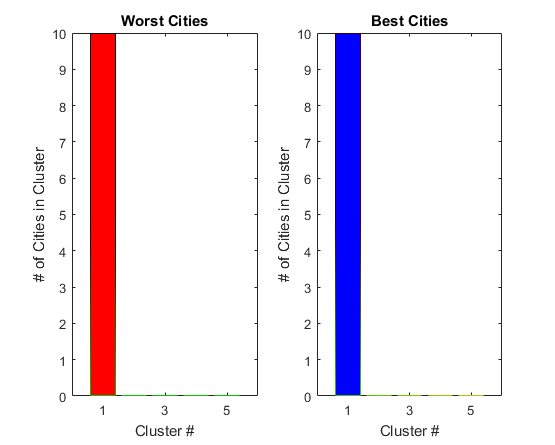
Without Population | Un-centered

K=5 |

0% Expected

100% Unexpected

1. K=5
   1. Clearly we can see that each of the cities landed in the exact same cluster. Not what you would expect from best cities vs. worst cities to live in.



1. K=10

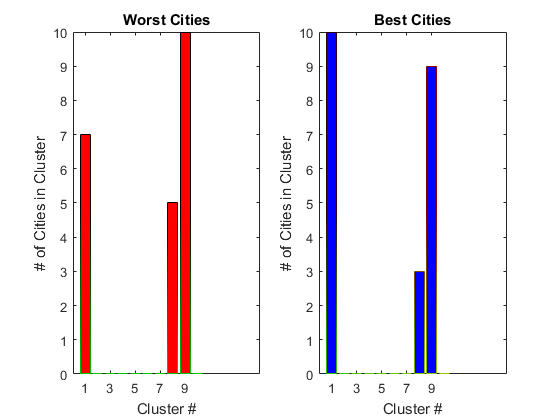
Without Population | Un-centered

K=10 |

33.33% Expected

66.66% Unexpected

* 1. (7/10) 0 0 0 0 0 0 (3/5) (9/10)
  2. We will consider any number *relatively* close to 50% or less as a significant difference, or something that we would *expect*.



1. K=15

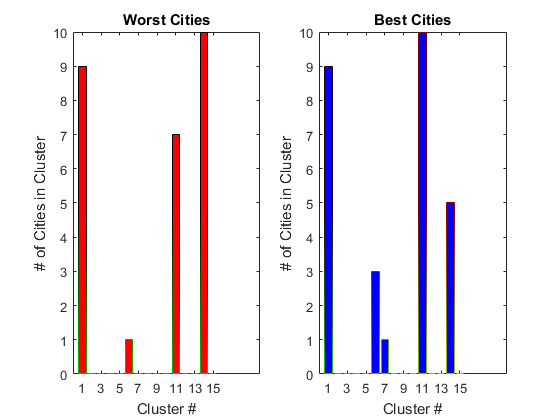
Without Population | Un-centered

K=15 |

40% Expected

60% Unexpected

* 1. (10/10) 0 0 0 0(1/3) (1/1) 0 0 0 (10/10) 0 0 (5/10) 0



1. K=20
   1. (4/8) 0 0 0 0 0 (6/9) 0 0 (0/8) 0 0 (6/10) (2/3) (7/10) 0 0 0 0 (0/2)

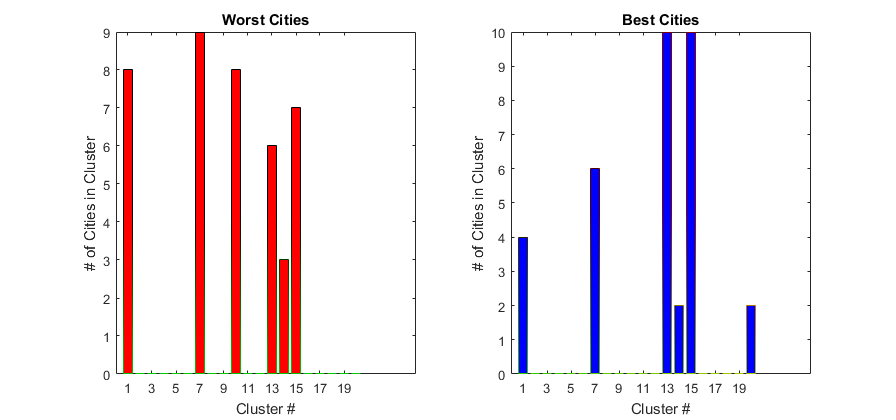
Without Population | Un-centered

K=20 |

50% Expected

50% Unexpected

* 1. Exactly 50% of the observations on are *on the fence*, so we see a solid position of differentiation.



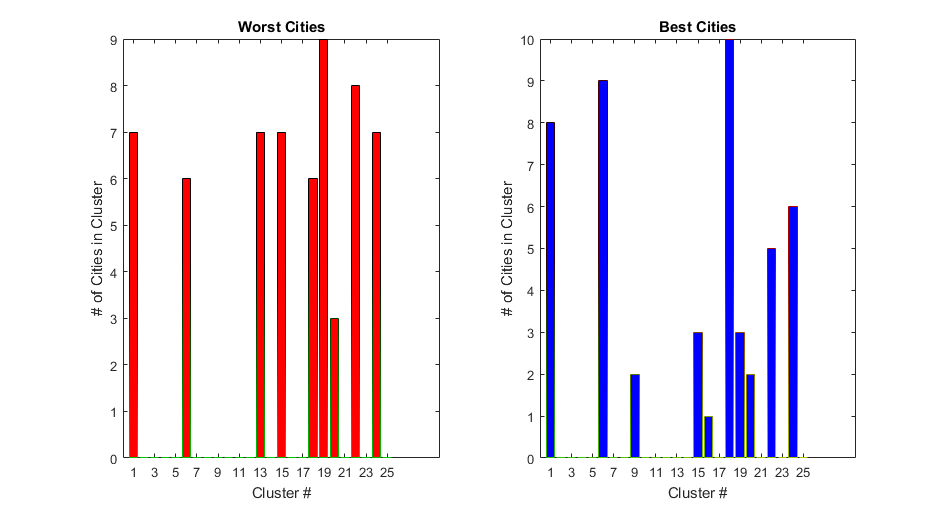
1. K=25
   1. (7/8) 0 0 0 0 (6/9) 0 0 (0/2) 0 0 0 (0/7) 0 (3/7) (0/1) 0 (6/10) (3/9) (2/3) 0 (5/8) 0 (6/7)
   2. Less than 50% of them were *on the fence* so we can start to see a decline.

Without Population | Un-centered

K=25 |

45% Expected

65% Unexpected



1. K=30

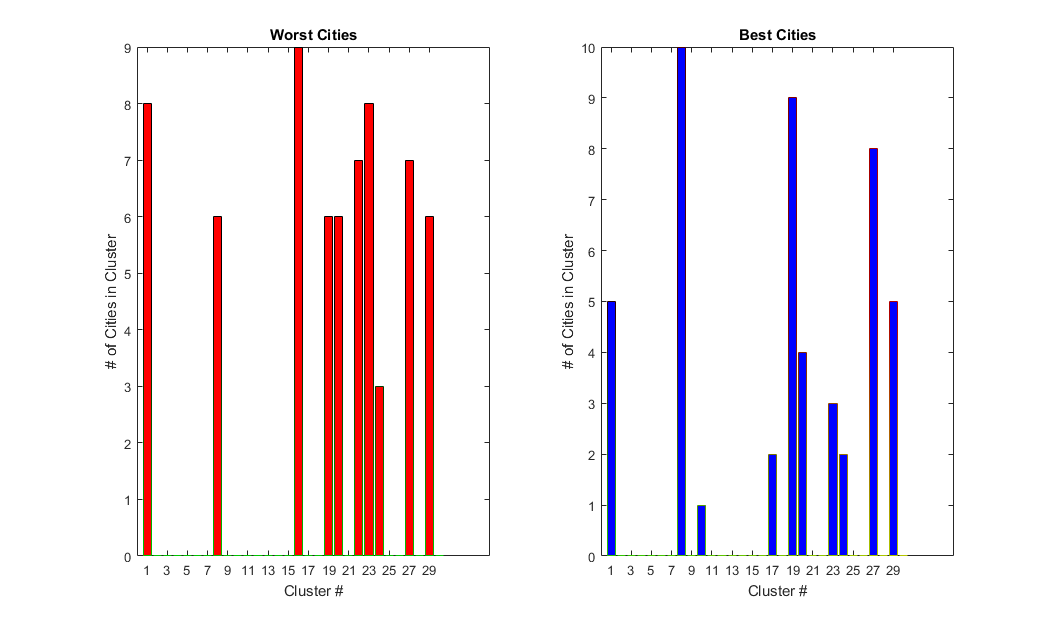
Without Population | Un-centered

K=30 |

41.66% Expected

58.33% Unexpected

* 1. 5/8 0 0 0 0 0 0 6/10 0 0/1 0 0 0 0 0 0/9 0/2 0 6/9 4/6 0 0/7 3/8 2/3 0 0 7/8 5/6 0



***Without population centered***

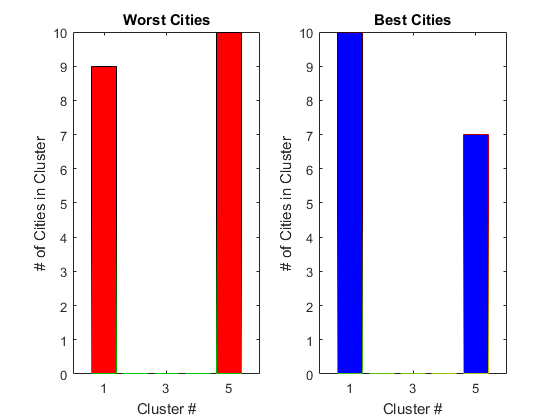
1. K=5
   1. 9/10 0 0 0 7/10

Without Population | Centered

K=5 |

0% Expected

100% Unexpected



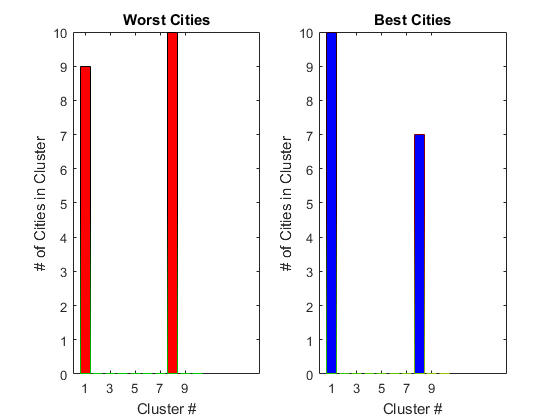
1. K=10
   1. 9/10 0 0 0 0 0 0 7/10 0

Without Population | Centered

K=10 |

0% Expected

100% Unexpected



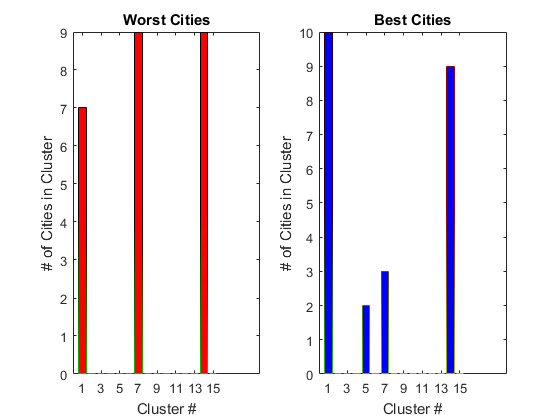
1. K=15
   1. 7/10 0 0 0 0/2 0 3/9 0 0 0 0 0 0 9/10 0

Without Population | Centered

K=15 |

50% Expected

50% Unexpected



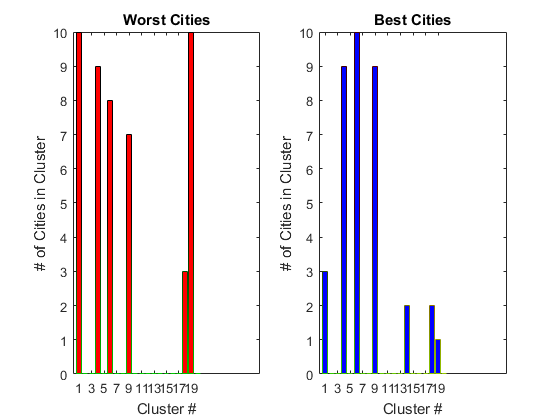
1. K=20
   1. 3/10 0 0 9/9 0 8/10 0 7/9 0 0 0 0 0/2 0 0 0 2/3 1/10

Without Population | Centered

K=20 |

42.86% Expected

57.14% Unexpected



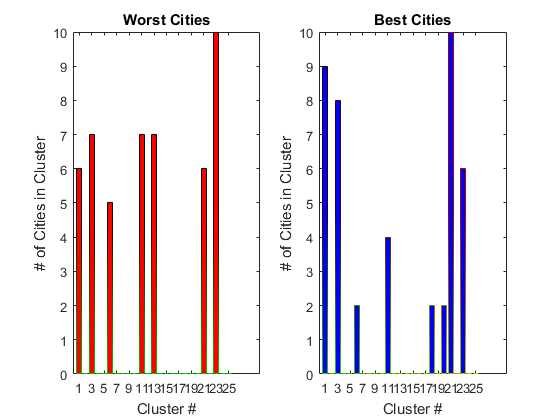
1. K=25
   1. 6/10 0 7/8 0 2/5 0 0 0 0 4/7 0 0/7 0 0 0 0 0 0 0/2 0 0/2 6/10 0 6/10 0 0

Without Population | Centered

K=25 |

44.44% Expected

55.56% Unexpected



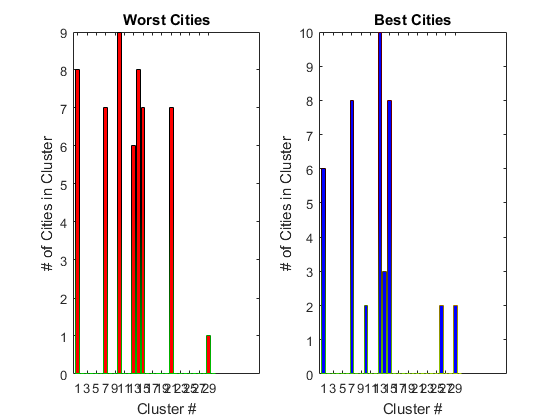
1. K=30
   1. 6/8 0 0 0 0 0 7/8 0 0 2/9 0 0 6/10 3/8 7/8 0 0 0 0 0 0/7 0 0 0 0 0/2 0 0 1/2

Without Population | Centered

K=30 |

55.55% Expected

44.44% Unexpected



***With population centered***

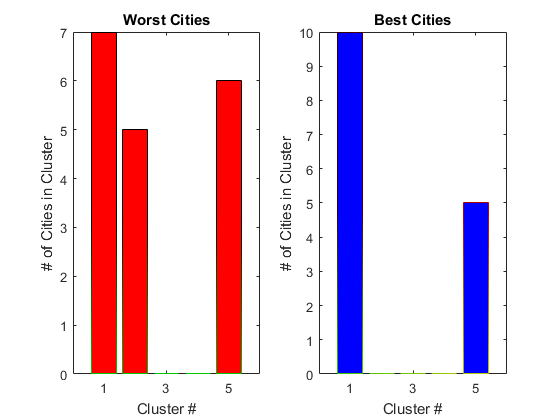
1. K=5
   1. 7/10 0/5 0 0 5/6

With Population | Centered

K=5 |

33.33% Expected

66.67% Unexpected



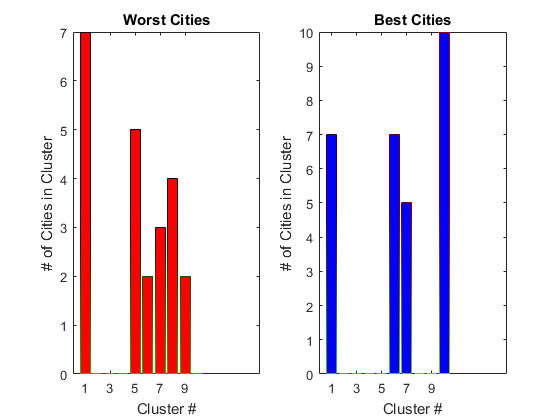
1. K=10
   1. 7/7 0 0 0 0/5 2/7 3/5

With Population | Centered

K=10 |

50% Expected

50% Unexpected



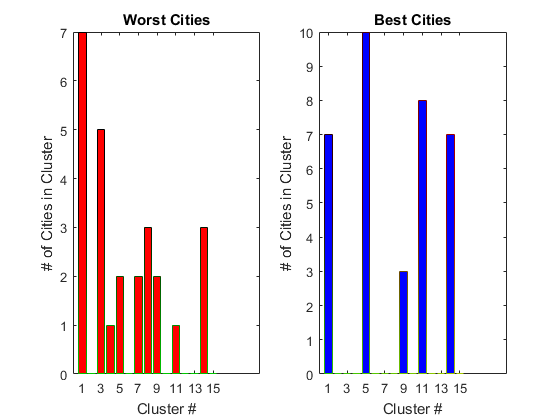
1. K=15
   1. 7/7 0 0/5 0/1 2/10 0 0/2 0/3 0/2 2/3 0 1/8 0 0 3/7 0

With Population | Centered

K=15 |

80% Expected

20% Unexpected



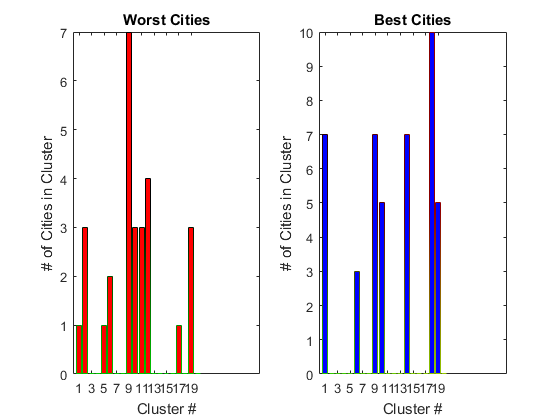
1. K=20
   1. 1/7 0/3 0 0 0/1 2/3 0 0 7/7 3/5 0/3 0/4 0 0 0/7 0 0 0/1 0/10 3/5

With Population | Centered

K=20 |

70% Expected

30% Unexpected



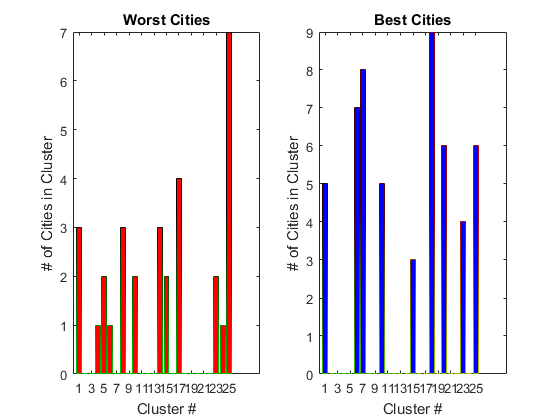
With Population | Centered

K=25 |

78.57% Expected

21.42% Unexpected

1. K=25
   1. 3/5 0 0 0 0/2 1/7 0/8 0/3 0 2/5 0 0 0 0/3 2/3 0 0/4 0/9 0 0/6 0 0 2/4 0/1 6/7



1. K=30

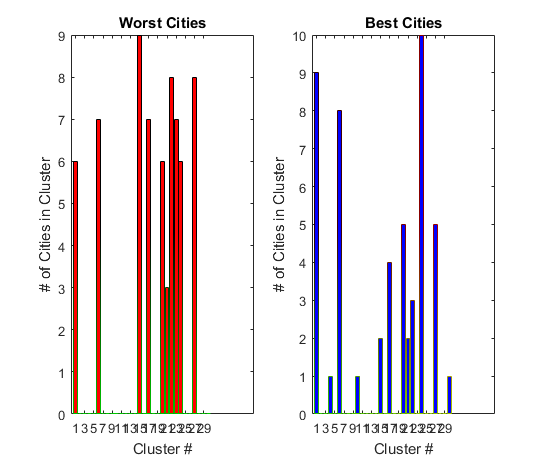
With Population | Centered

K=30 |

46.15% Expected

53.85% Unexpected

* 1. S
  2. 6/9 0 0 0/1 0 7/8 0 0 0 0/1 0 0 0 0 2/9 0 4/7 0 0 5/6 2/3 3/8 0/7 6/10 0 0 5/8 0 0 0/1



***With population un-centered***

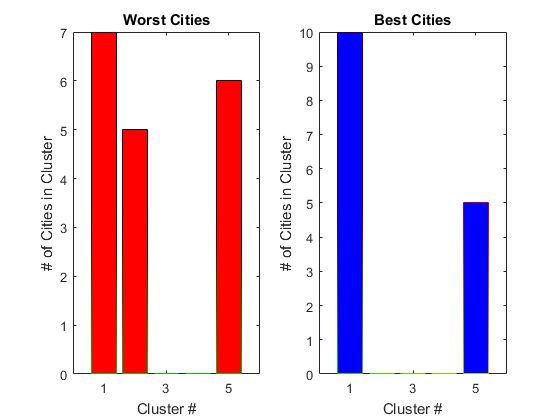
1. K=5
   1. 7/10 0/5 0 0 5/6

With Population | Un-centered

K=5 |

33% Expected

67% Unexpected



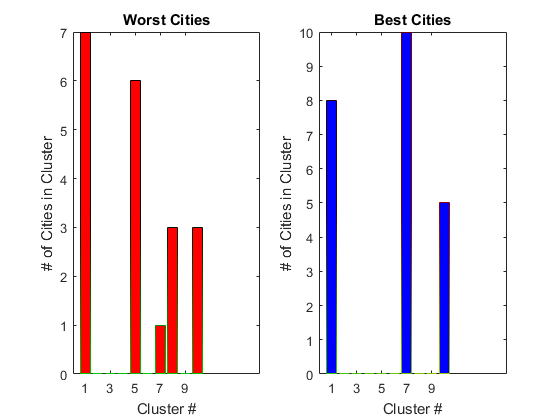
1. K=10
   1. 7/8 0 0 0 0/6 0 1/10 0/3 0 3/5

With Population | Un-centered

K=10 |

60% Expected

40% Unexpected



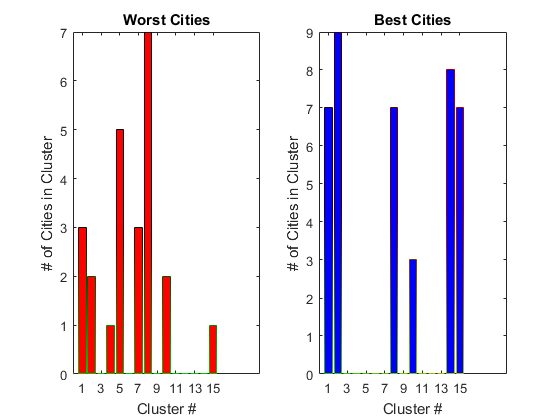
1. K=15
   1. 3/7 2/9 0/1 0/5 0/3 7/7 2/3 0/8 1/7

With Population | Un-centered

K=15 |

77% Expected

23% Unexpected



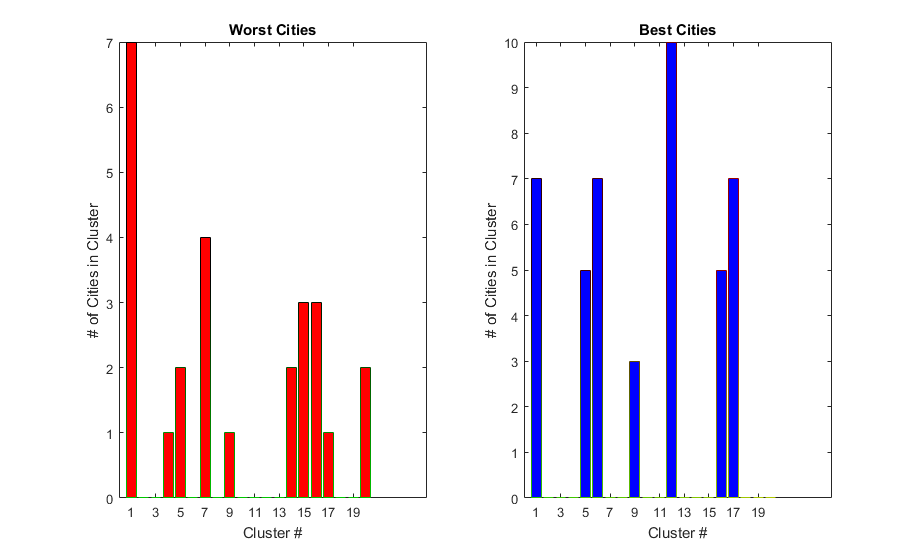
1. K=20
   1. 7/7 0 0 0/1 2/5 0/7 0 1/3 0 0 0/10 0 0/2 0/3 3/5 1/7 0 0 0/2

With Population | Un-centered

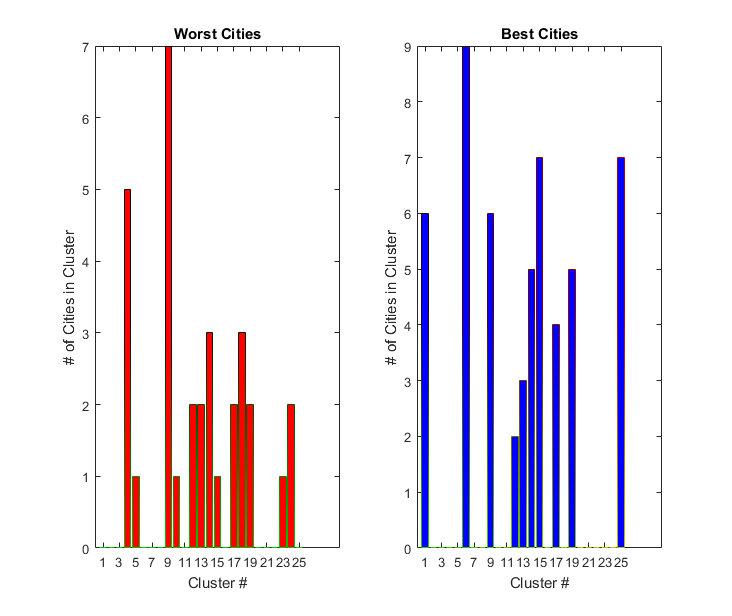
K=20 |

82.82% Expected

18.18% Unexpected



1. K=25
   1. S
   2. 0/6 0 0 0/5 0/1 0/9 0 0 6/7 0/1 0 2/2 2/3 3/5 1/7 0 2/4 0/3 2/5 0 0 0 0 0/1 0/2 0/7



With Population | Un-centered

K=25 |

75% Expected

25% Unexpected

1. K=30
   1. S

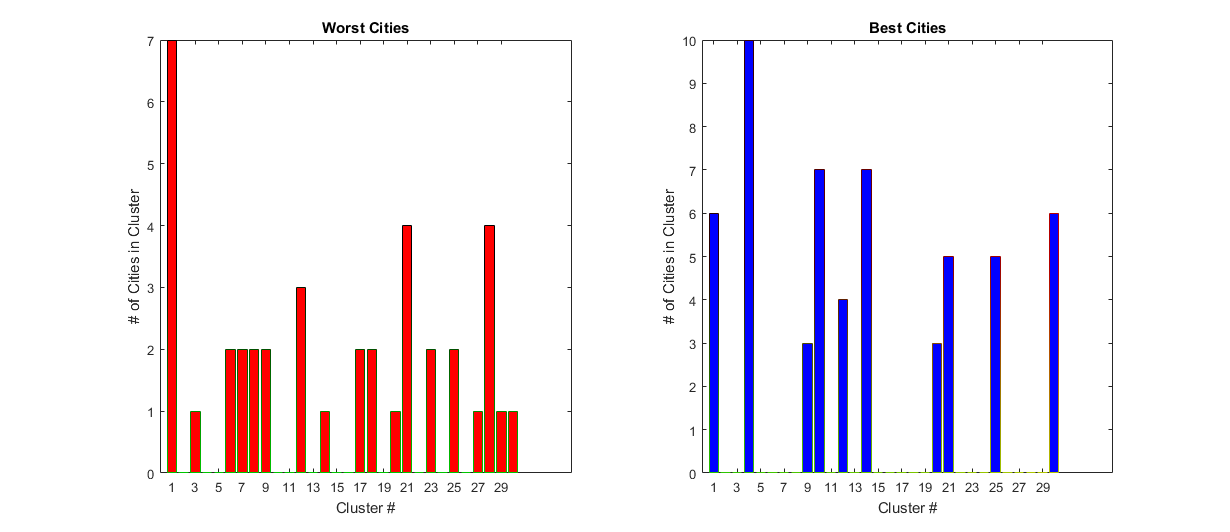
With Population | Un-centered

K=30 |

84.24% Expected

15.76% Unexpected

* 1. 4/10 0 1/10 0/10 0 0/2 0/2 0/2 2/2 0/7 0 ¾ 0 0 1/7 0 0 0 1/3 4/5 0 0/2 0/2 0 2/5 0 0/1 0/4 0/1 1/6



∴ One last plot will be AVG % Expected vs. clusters graph to visualize and give an idea of what’s going on.

∴ So for each graph type, with pop uncentered, with pop centered, without pop centered, without pop uncentered, build a table of 6x2 where the k=5:30 is on the rows and %expected and %unexpected is on the columns, then compare each tables rows to find the best k=5,10,..,30 from the tables, then from those elect ones, choose the best k number.

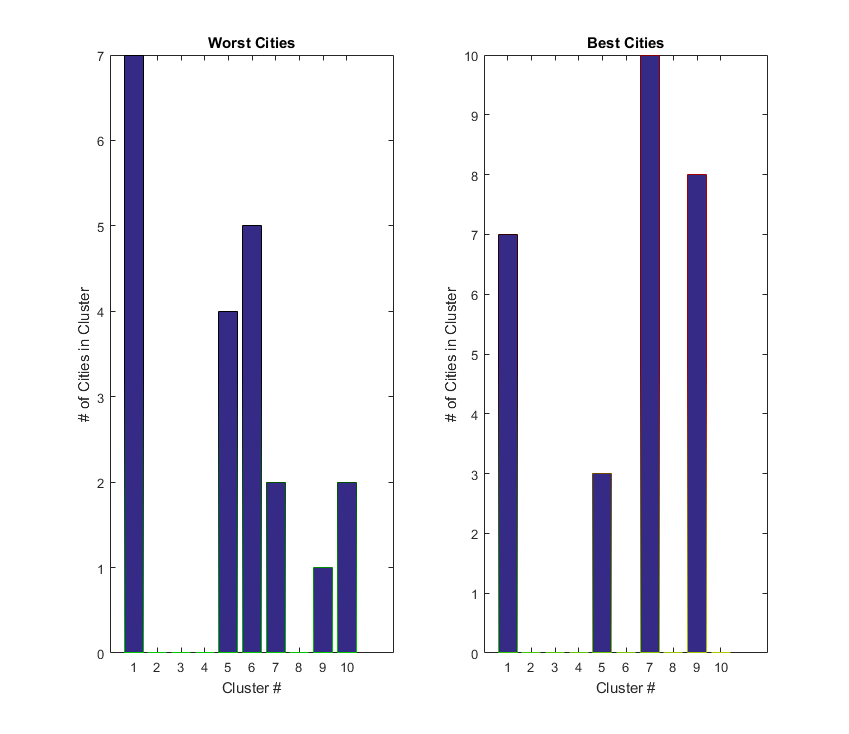
∴Then once the clusters have been decided, example k=20, then find a new way to determine whether a cluster is preferred over another and rank them 🡪 Notice, there may be some assumptions that need to be made, or drawbacks that occur in this process, make note of them and explain what happened.

*Milestones in the Analysis*

1. Understanding why K-Means clustering is putting oddly related crime cities into similar clusters.
   1. The possibility of city’s *safety* is dependent on not only crime patterns, but population. This leads to a potentially interesting find in the data; so called *safer* cities, might not be as safe with respect to some very important variables.
2. After **adding population as a crime** statistic and picking top 10 cities from USATODAY’s 2016 safest cities to live, I used the results from the K-means algorithm to determine how good the K in the clusters is, and with the exception of a couple things, the results made sense:

K=10 | 40% Expected

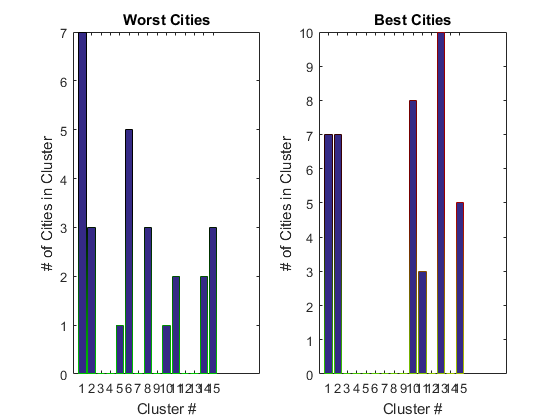
20% Unexpected



* 1. Abs\_Diffs = 0 0 0 0 1 5 8 0 7 6.
     1. **Yellow** represents interesting intersections, why are good and bad so close in these clusters?
     2. **Green** are not issues, neither landed in these clusters and that’s not something I want to study initially.
     3. **Blue** represents good cities occurring significantly in different clusters.
  2. An interesting find in the image above, is the ratios of the worst cities and the clusters they showed up in vs. the best cities and the clusters they showed up in. There are several similarities and several distinctions. Notice bins 1 & 5, why are they so close? Then notice the *expected* distinction between the rest of the bins 6,7,9 & 10. After running the results again, I see that exactly 2 bins have the closeness. Why are these good cities finding themselves in the bad cities cluster? Or is it the bad cities finding themselves in the good cities cluster? Are bad cities, actually not all that bad? Or are good cities not actually all that good based on the statistics from these data? Maybe without population we can see a difference.
  3. Expected find ratio: 4/10 = 40%
  4. Unexpected find ratio 2/10 = 20%

K=15 | 46.67% Expected

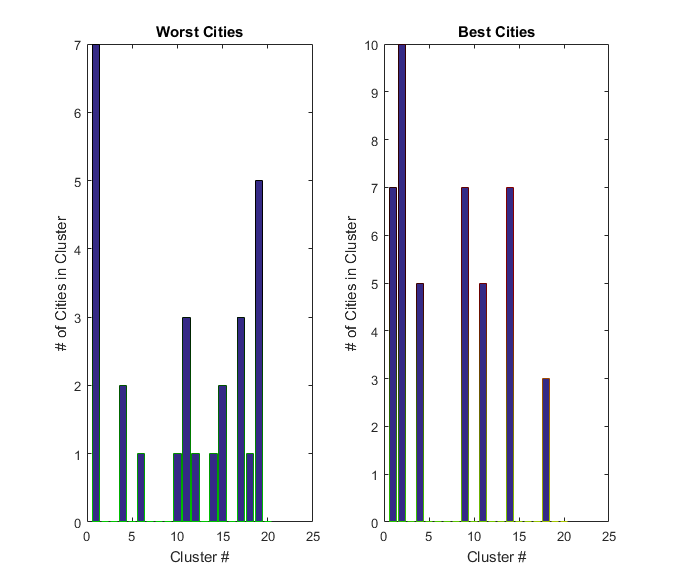
20% Unexpected



1. Abs\_Diffs = 0 4 0 0 1 5 0 3 0 7 1 0 10 2 2.
   1. **Yellow** represents interesting intersections, why are good and bad so close in these clusters?
   2. **Green** are not issues, neither landed in these clusters and that’s not something I want to study initially.
   3. **Blue** represents good cities occurring significantly in different clusters.
   4. Again, when we increase k=15 we see a better ratio 3 *odd* intersections of good and bad cities occurring the same cluster, with about 7 as expected.
   5. Expected find ratio: 7/15=46.67%
   6. Unexpected find ratio: 3/15=20%

K=20 | 50% Expected

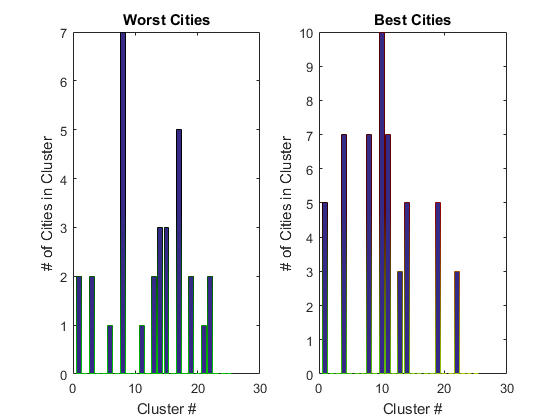
10% Unexpected



1. Abs\_Diffs = 0 10 0 3 0 1 0 0 7 1 2 1 0 6 2 0 3 2 5 0
   1. If we see a city from the worst category appear in cluster\_1 1 time, and there are no cities from the best category that appear in cluster 1, this is not going to be considered a significant *yellow* category because there is no overlap, so the 1’s and 0’s we are looking for are the completely overlapped in 1 cluster or closely, nonzero overlap. Else if the number is above 0,1, or 2 we would
   2. So by the exact same metric of searching to minimize 0’s and 1’s in the yellow we see:
      1. Expected find ratio: 10/20 = 50%
      2. Unexpected find ratio: 1/20 = 10%

K=25 | x% expected

y% unexpected



Abs\_Diffs =

3 0 2 -7 0 1 0 0 0 -10 -6 0 -1 -2 3 0 5 0 -3 0 1 -1 0 0 0

Continued .. I want to do all of the above k5,k10,k15,k20,k25 again but with my new plots and the new data given to me from them then see if the results are consistent.