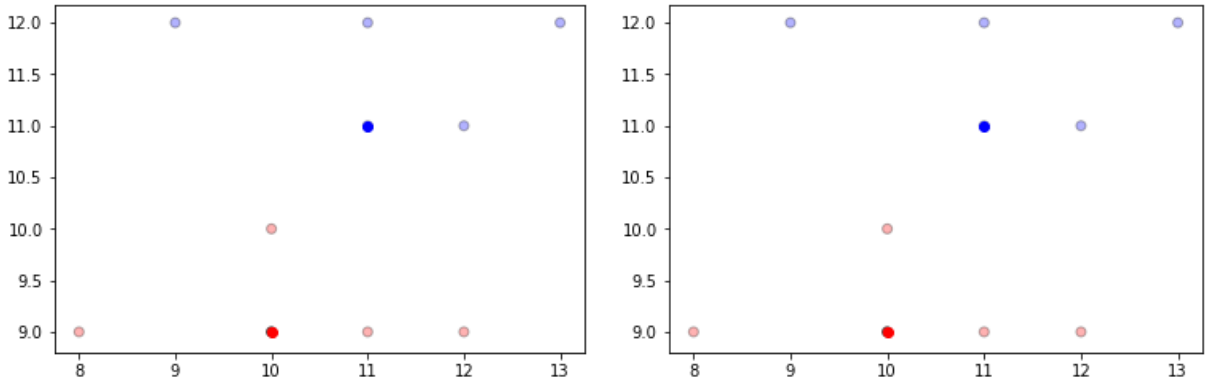
Patrick Tinsley
Data Science - HW3
Due Date: 4/3/2018

This assignment was completed using Python 3.6.4.

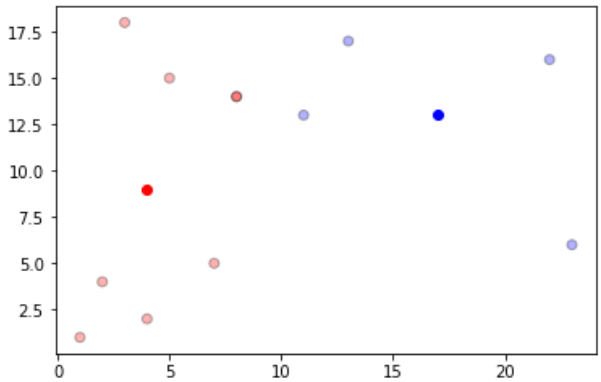
Problem 1

Below are the plots for the different initial centroids. The graphs are the same; the initializations generate the same result. However, we prefer the second initialization to the first because it converges after only one iteration while the other converges after two.



Problem 2

Below is the plot for the K Means clustering on Ranking_2015 and Ranking 2017. Between the first and second problems, we prefer the first two features because it takes fewer iterations to converge, and the plot suggests K=3 might be more appropriate; this can be seen in the blue cluster with the centroid rather far from all points in said cluster.



Problem 3

Below are plots to describe the K=3 clustering algorithm. Upon inspecting the first plot, we choose the initial centroids to be (5,5), (8,15), and (20,15); the resulting clusters can be seen in the second plot. Compared to K=2, we prefer K=3 because it converges after one iteration, and the clusters more appropriately describe the trends in the data. By adding a third cluster, we see centroids that are much closer to the data points, especially in the green cluster, which spread out the blue cluster in the plot from Problem 2.