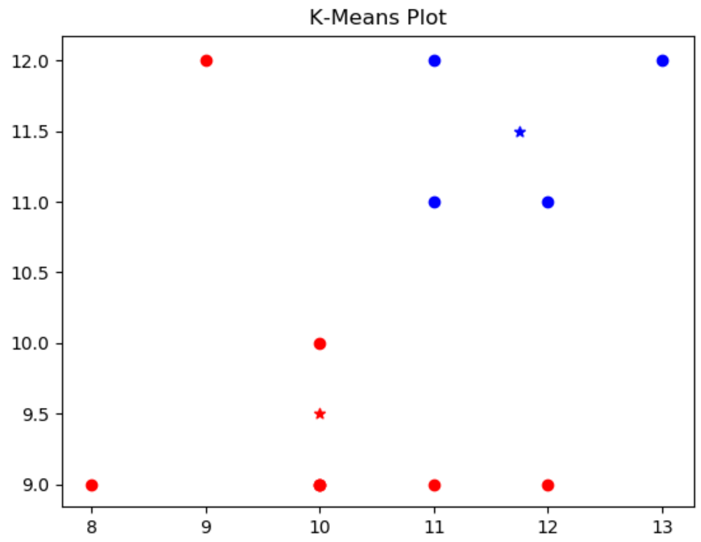
Kailee Madden
Programming Homework 3

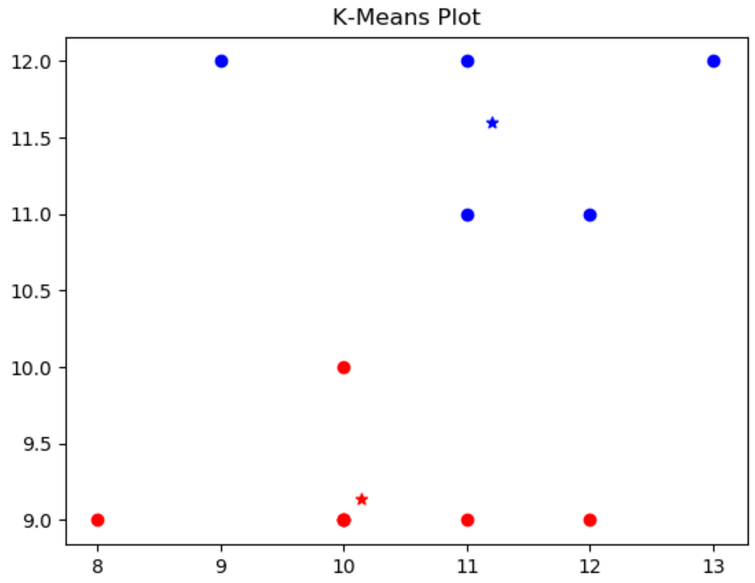
Question 1:

The two different sets of centroids do not give the same result. The coordinates of the ending centroids are different, and the final groupings of data objects differ.
I prefer the initialization using the (7,7) and (7,14) centroids because the data point in the top left of (7,7) and (14,14) seems misclassified.

For (7,7) and (14,14): (centroids are the stars)

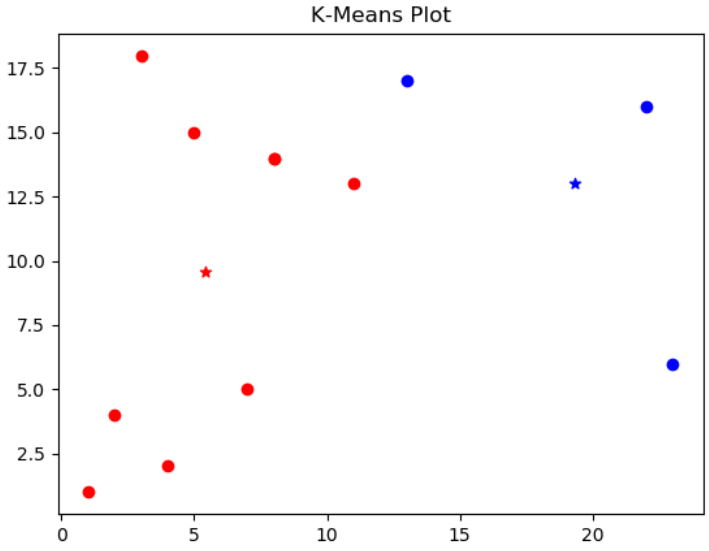


For (7,7) and (7,14): (centroids are the stars)



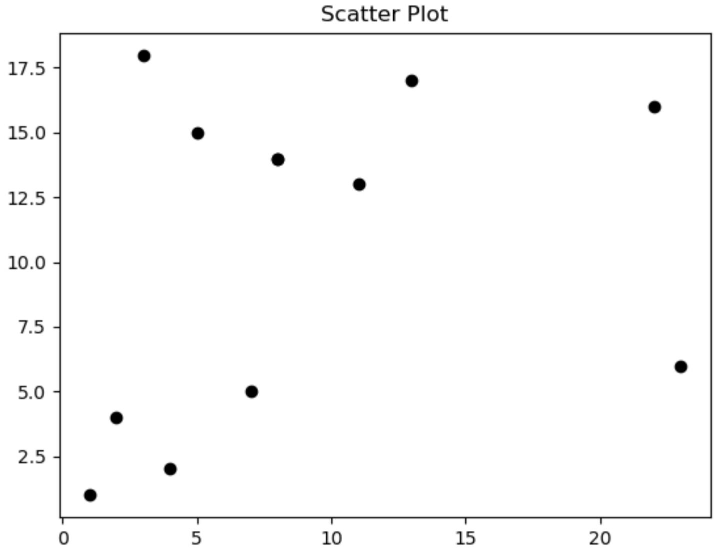
Question 2:

Compared with the cluster results in Q1, I prefer the clustering based on these two new features, using Manhattan distance less than the clustering done in Q1. Specifically, the clustering that I prefer the most is Q1 with starting centroids (7,7) and (7,14).
In the scatter plot the centroids are the stars.

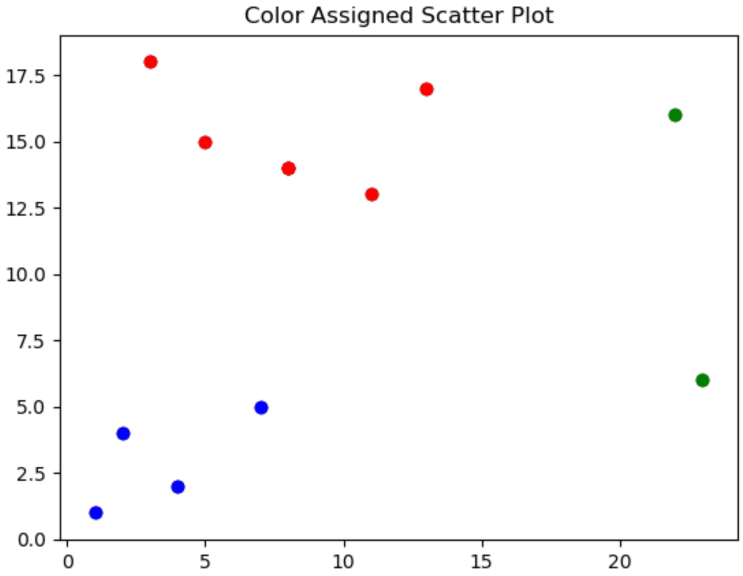


Question 3:

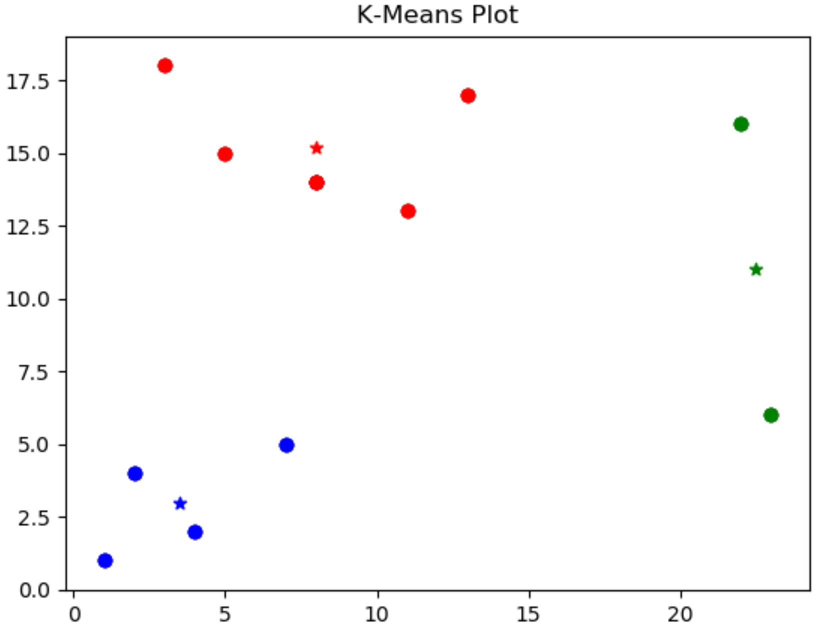
Ungrouped scatter plot



Manually grouped scatter plot.



To generate my preferred grouping (as shown above), I used the following centroids: (9,15), (5,3), and (23,12)



I still prefer the K=2 grouping from Q1 that used centroids (7,7) and (7,14). This is because in Q1 we used #Wins in 2015 and 2017, which looking at the axis and the data points seems more consistent in terms of grouping and number of wins in 2015 being similar to in 2017. On the other hand, when we use the Rankings, there seems to be more variability. For instance, in the green cluster we have one team that was ranked lower than 20 in 2015 but higher than 7 in 2017 while the other data point is lower than 20 and lower than 15 respectively. From a predictive analysis point of view this would be less predictable than looking at the number of wins with K=2. Of course, the K=3 grouping is more visually appealing because it uses more colors and the clusters are more separated from one another.