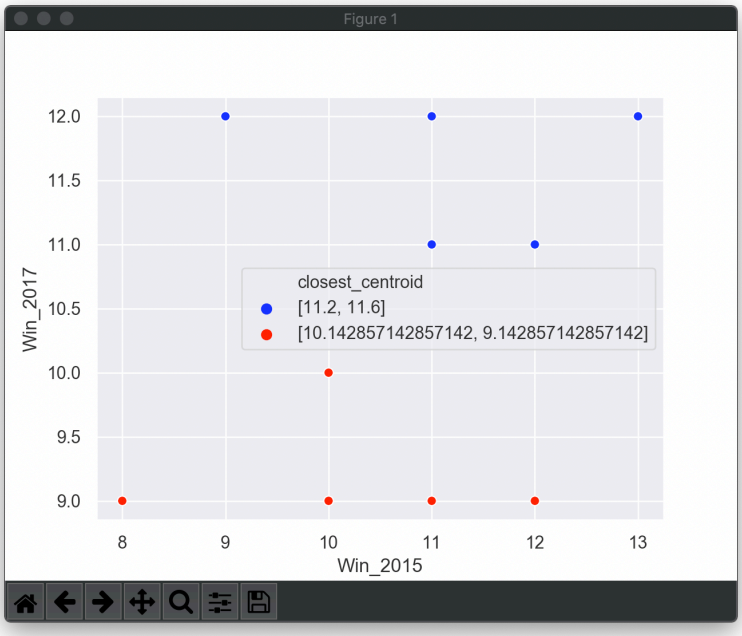
Ryan Karl
CSE-40647 Data Science
Dr. Meng Jiang
2 November 2019

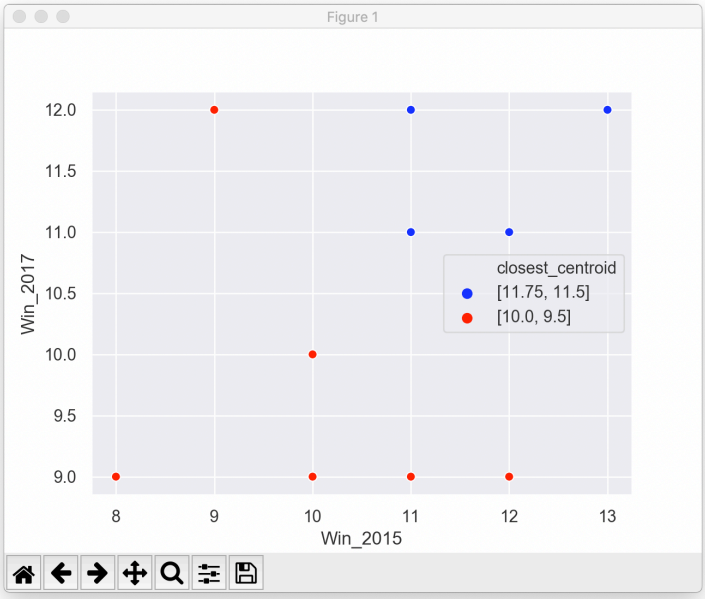
Homework 3 Solutions

Question 1 Solution:

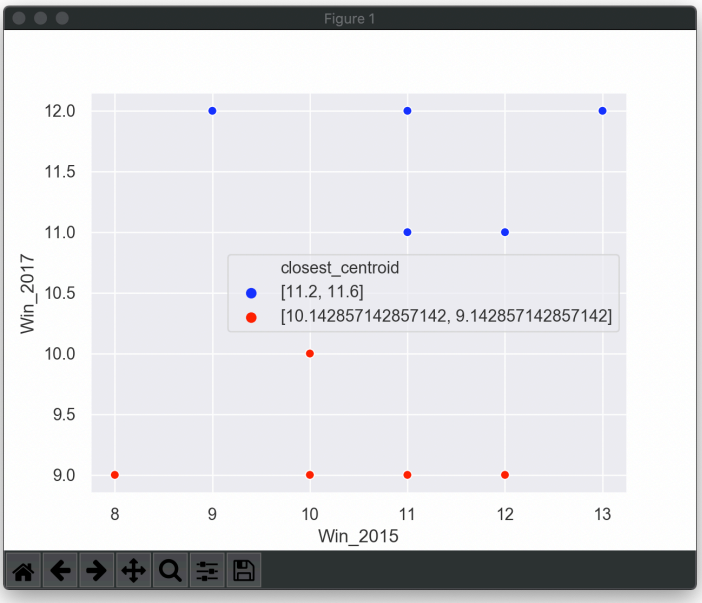
Plot with centroids (7,7) and (14,14) with version 1 of clustering:



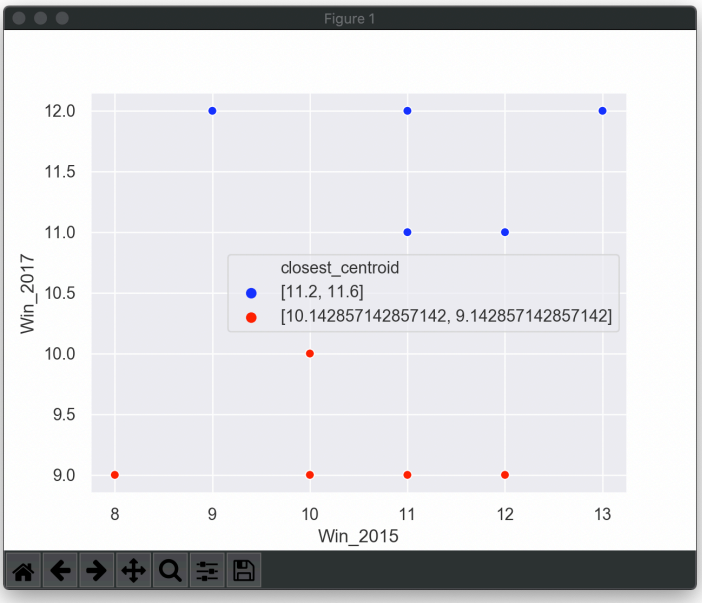
Plot with centroids (7,7) and (14,14) with version 2 of clustering:



Plot with centroids (7,7) and (7,14) with version 1 of clustering:



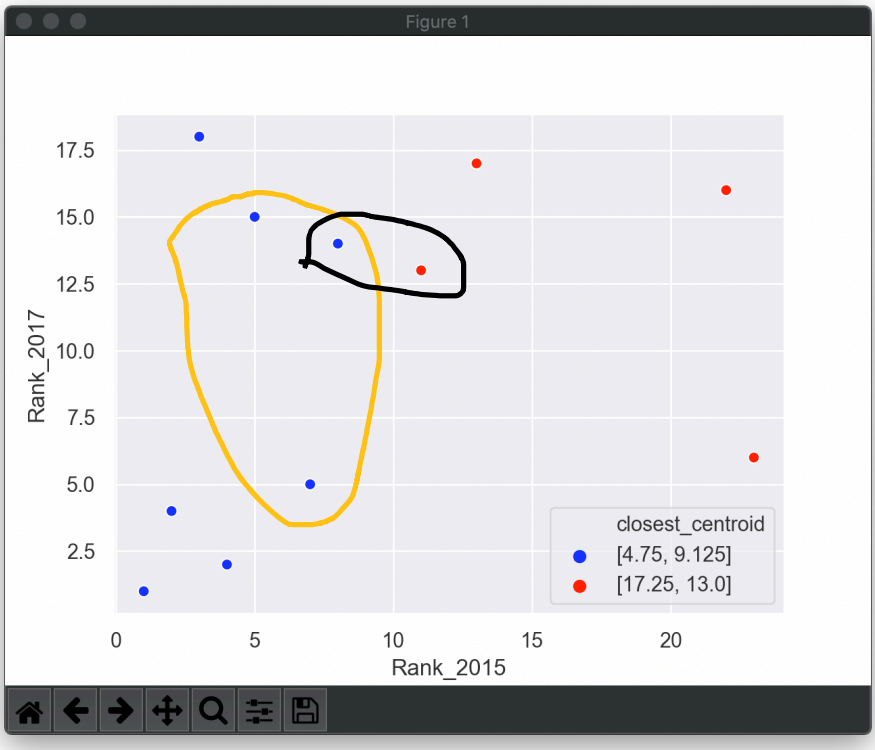
Plot with centroids (7,7) and (7,14) with version 2 of clustering:



Notice that initializing with either set of centroids yields the same result if we use version 1 of the k-means implementation, but a different result occurs if we use version 2 of k-means. Note the positions of the final closest centroids are shown in each image. I prefer the centroids (7,7) and (7,14) because the final SSE, for this pair is 19.7142857143 for both k-means implementations, but it is 19.7142857143 and 21.75 for the initial centroids (7,7) and (14,14) depending on the k-means implementation used; this means that the (7,7) and (7,14) pair of initial centroids produces a more consistent output and on average produces a lower final SSE. Nevertheless, the other pair of centroids are a farther distance from each other than (7,7) and (7,14), and this will promote the formation of clusters that have a farther distance from each

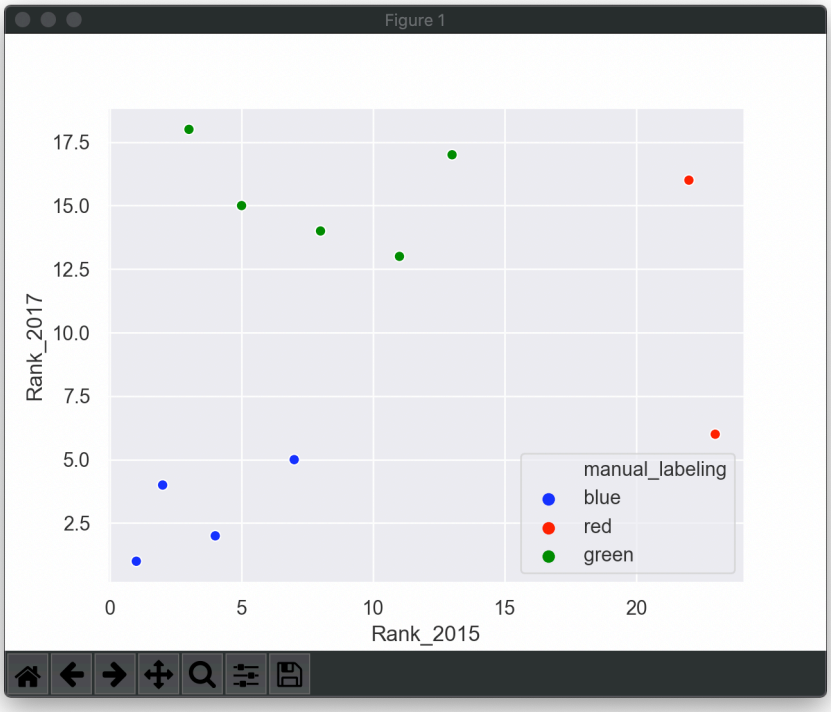
other (this is similar to the strategy promoted by Arthur & Vassilvitskii in the K-Means++ paper from 2007). However, it is interesting to note that this choice can yield different clusters in different implementations of the clustering algorithm in this instance, so the K-Means++ strategy will not always be optimal and seems to be sensitive to outliers in some instances.

Question 2 Solution:
Plot with centroids (1,1) and (25,25):

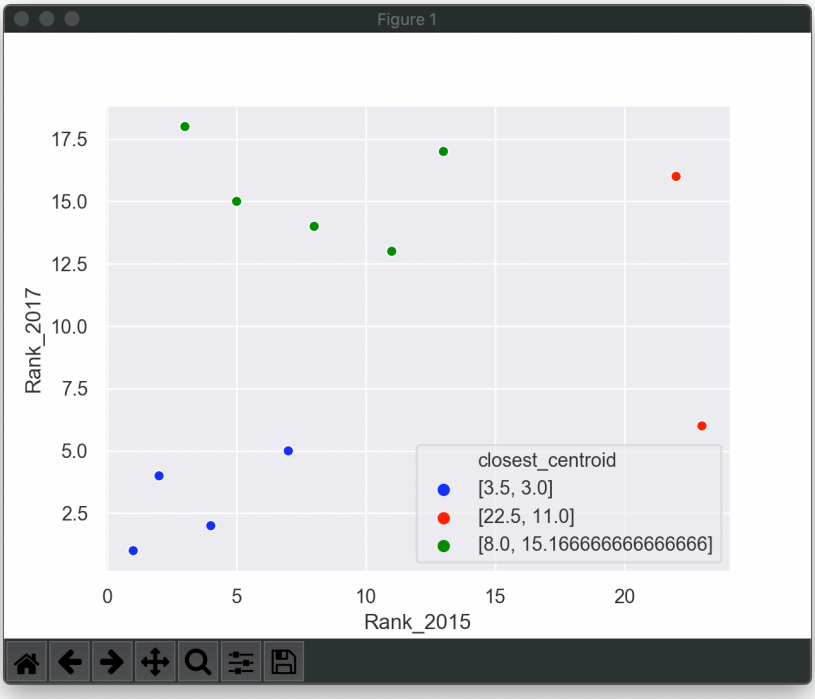


I prefer the clusters generated in Q1 because the maximum distance between each point in the same cluster seemed smaller and the minimum distance between points in different clusters seemed larger. This is apparent when looking at the scatter plot. Notice how the points in the yellow oval are far away from each other, but are still considered to belong to the same cluster. Similarly, notice how the points in the black oval are close together even though they are in different classes. More formally, I prefer Q1 because the best SSE from Q1 was 19.7142857143 which is noticeably less than the SSE of 559.125 in the clustering for Q2.

Question 3 Solution:
Plot with my own manual labels:



Plot using centroids (5,5), (15,10), and (20,10) (note it is the same as above):



I prefer the plot K=3 from Q3 over that from Q2. With the Q3 plot, the maximum distance between each point in the same cluster seems smaller and the minimum distance between points in different clusters seems larger. More specifically, it seems to handle classifying the outliers on the right side of the graph better, and yields a better grouping for the points in the upper left.

More formally, I prefer Q3 because the SSE for Q3 was 168.333333333 which was considerably less than the SSE of 559.125 for Q2.