CS 373 Spring 2019: Homework 5

Youngsik Yoon, 0029846135

April 30, 2019

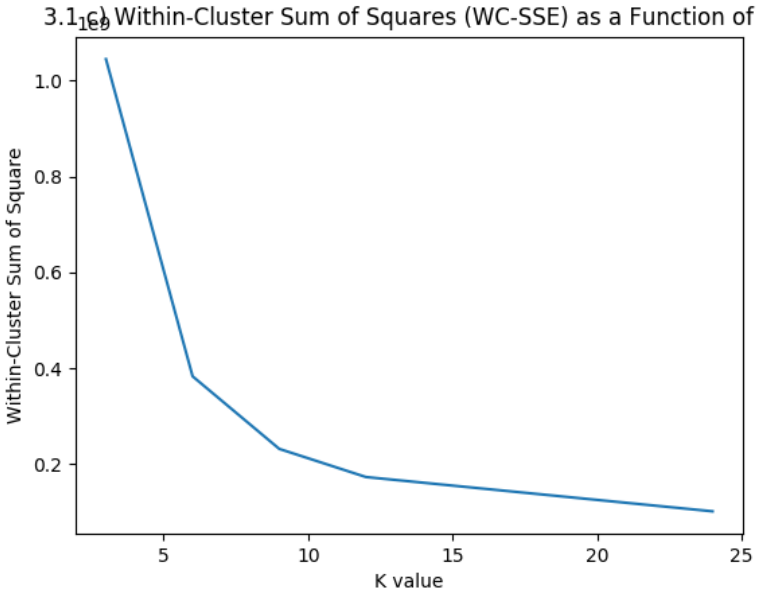
2 Kmeans

2.1 Theory

K-means clustering is a partition-based algorithm with spherical clusters, so it creates a division of the set of data points into non-overlapping clusters where each data point is only in one cluster. The issues with this algorithm is that it terminates at local optimum which is sensitive to initial seeds and the k value needs to be specified. Since calculating the mean is sensitive to outliers, K-means should be used in situations where the mean is defined and there are a lack of outliers and noise.

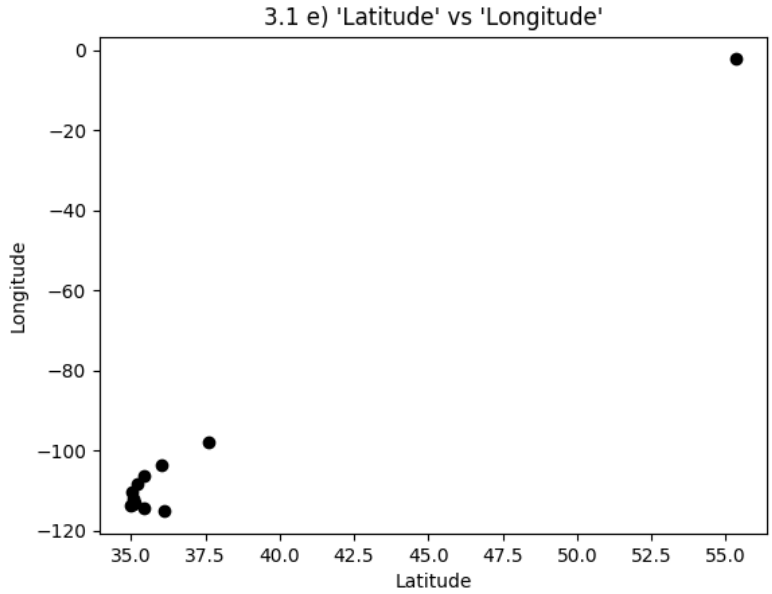
3 Analysis

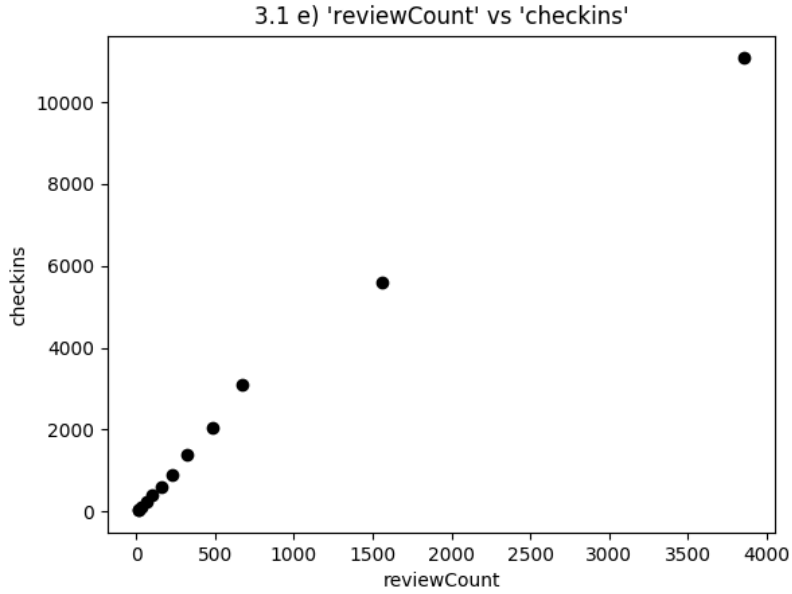
1. (c)



(d) 12; from the plot in part a), this is the *knee* in the plot so that adding another cluster doesn't improve much better the total WS-SSE.

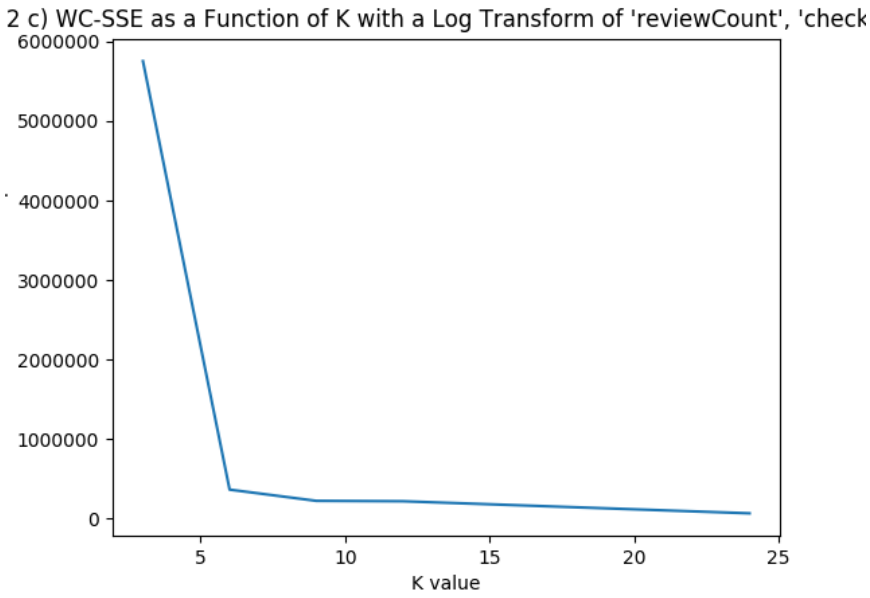
(e) Majority of the centroids were grouped together which means that the between cluster distance wasn't considered at all.





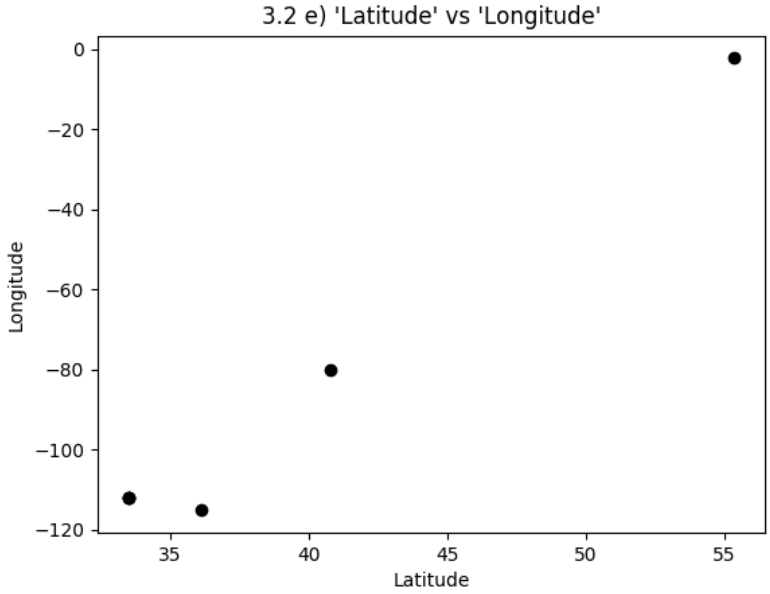
2. With a log transformation, I expect the centroids be better spaced out for the 'reviewCounts' vs 'checkins' graph. Additionally, the WC-SSE should drop.

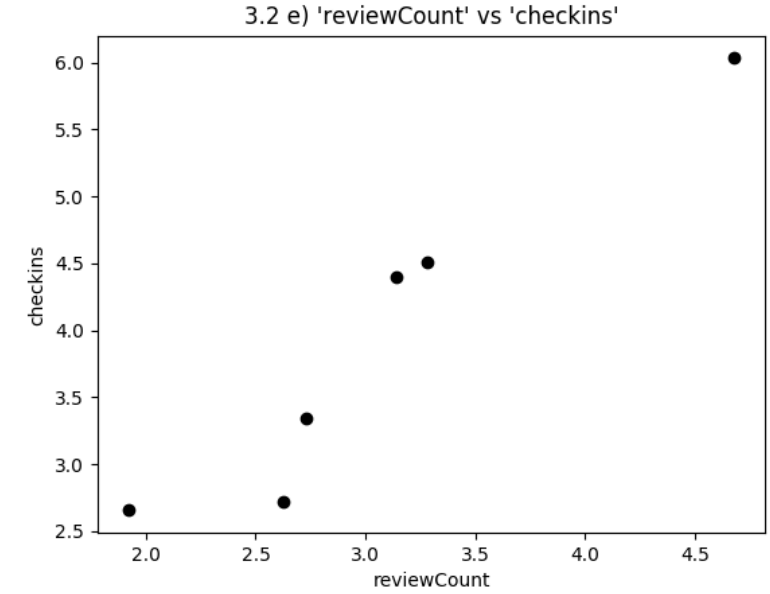
   (c)



   (d) 6; from the plot in part a), this is the *knee* in the plot so that adding another cluster doesn't improve much better the total WS-SSE.

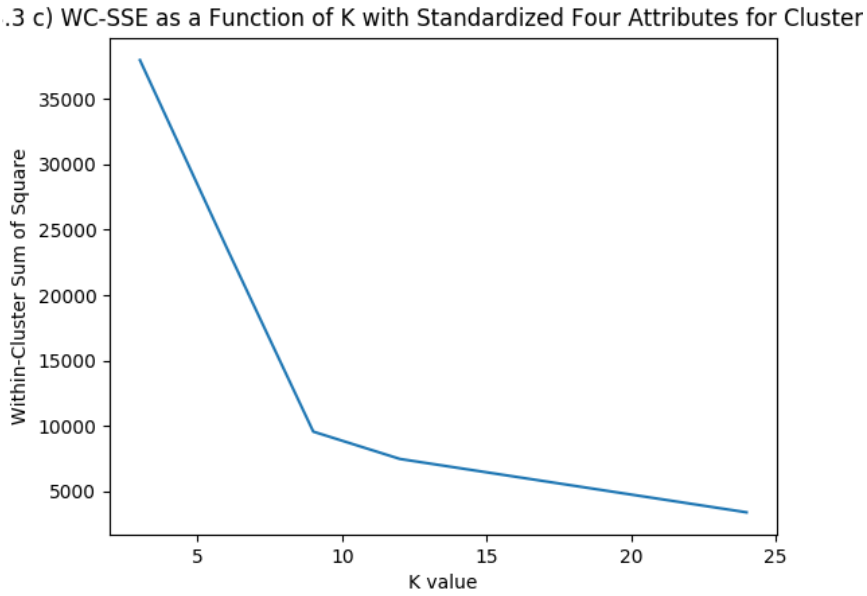
(e) The within cluster distances are a lot better most likely due to the log transformation.





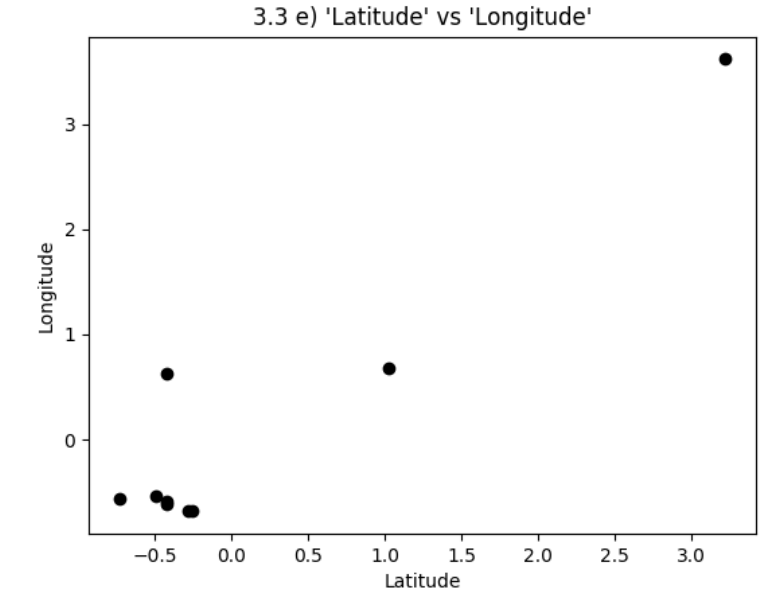
3. Similar to log, the standardization will help space out the centroids.

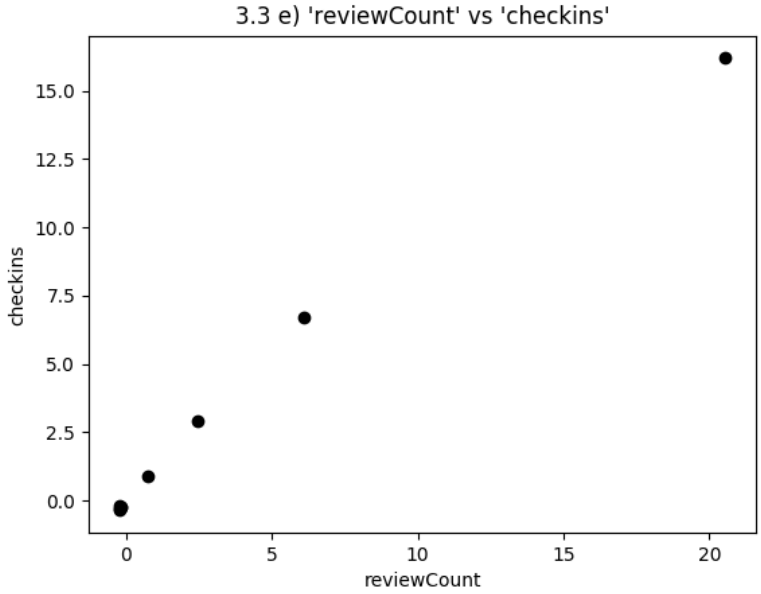
(c)



(d) 9; from the plot in part a), this is the *knee* in the plot so that adding another cluster doesn't improve much better the total WS-SSE.

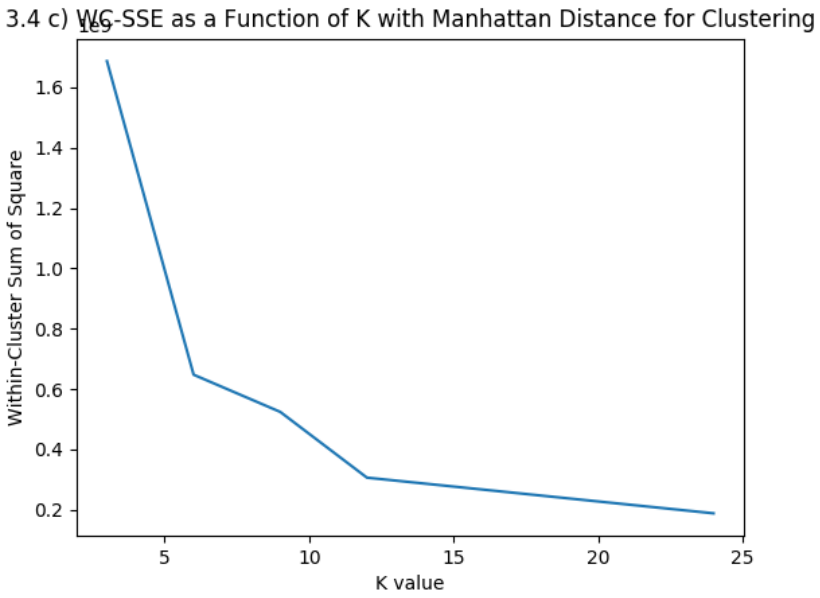
(e) The centroids are better spaced out and consider within cluster distance.





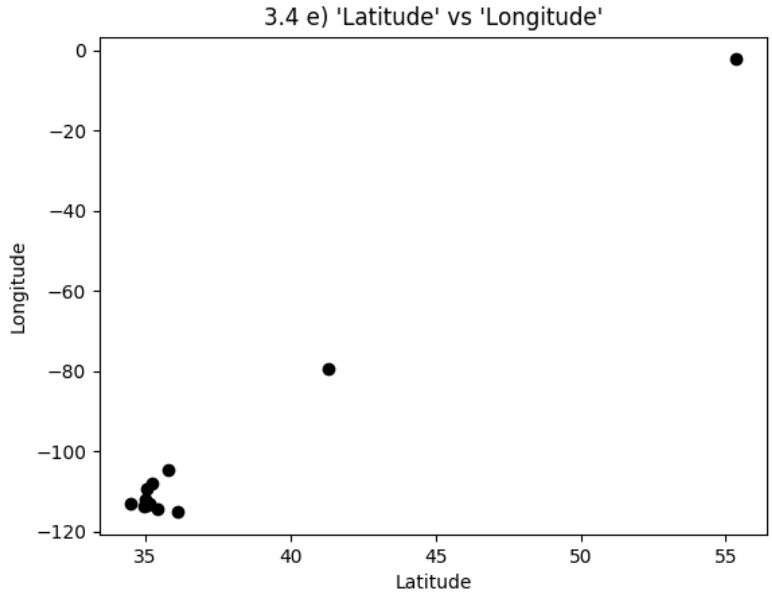
4. By switching the distance metric to Manhattan, the distances are better accounted for. Manhattan works better on data sets with higher dimensions.

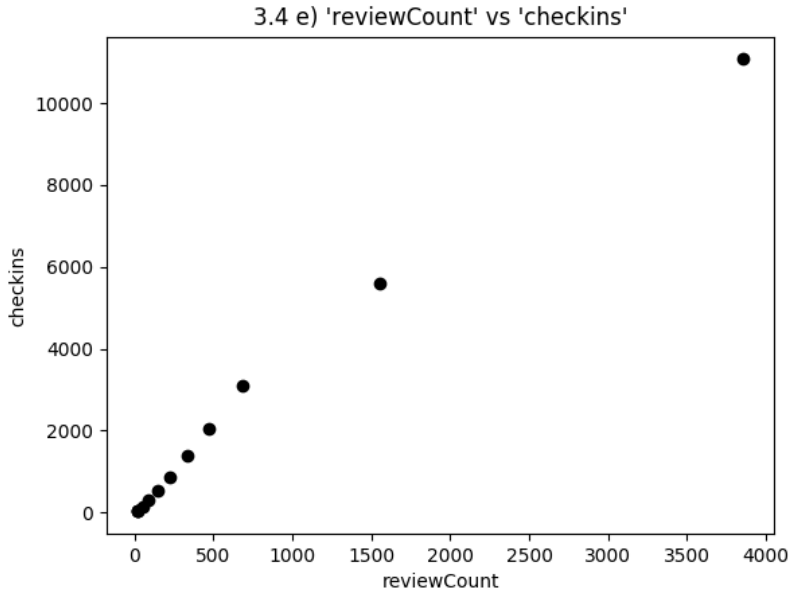
   (c)



   (d) 12; from the plot in part a), this is the *knee* in the plot so that adding another cluster doesn't improve much better the total WSSSE.

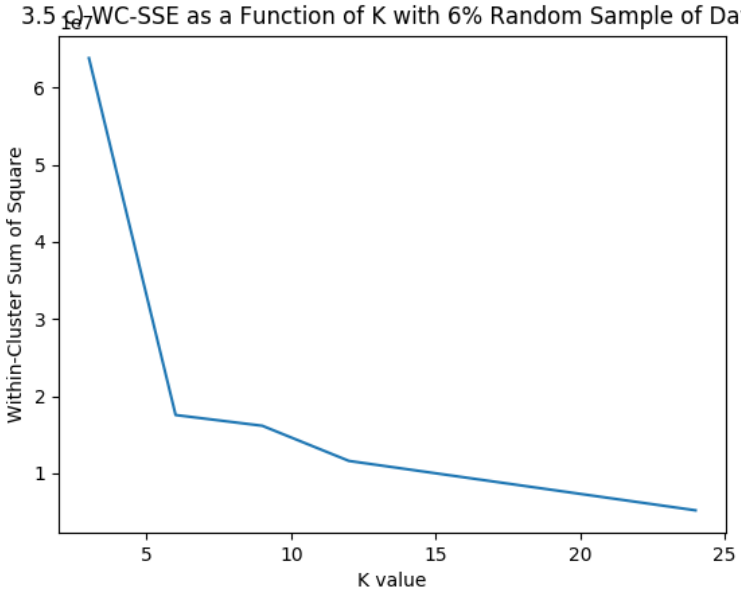
(e) The patterns are similar to the results in part 1.





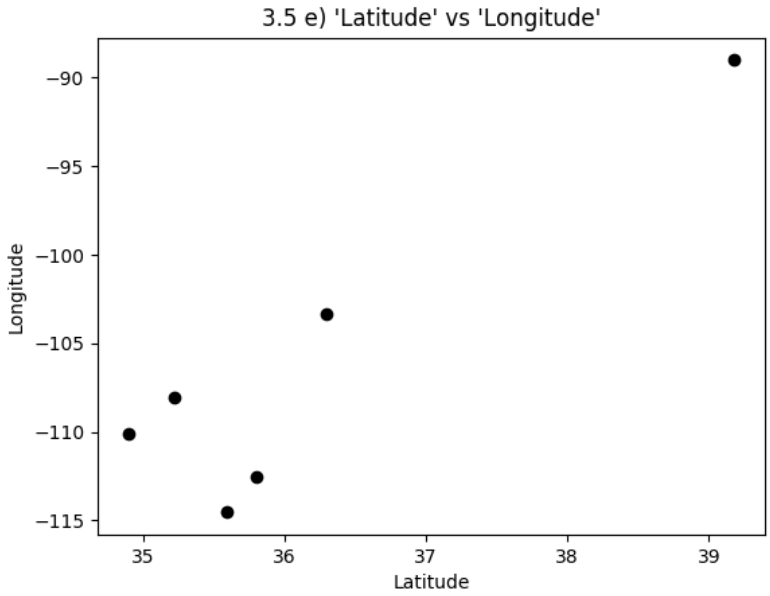
5. By taking only 6 percent of the training set, we should get less accurate clusters.

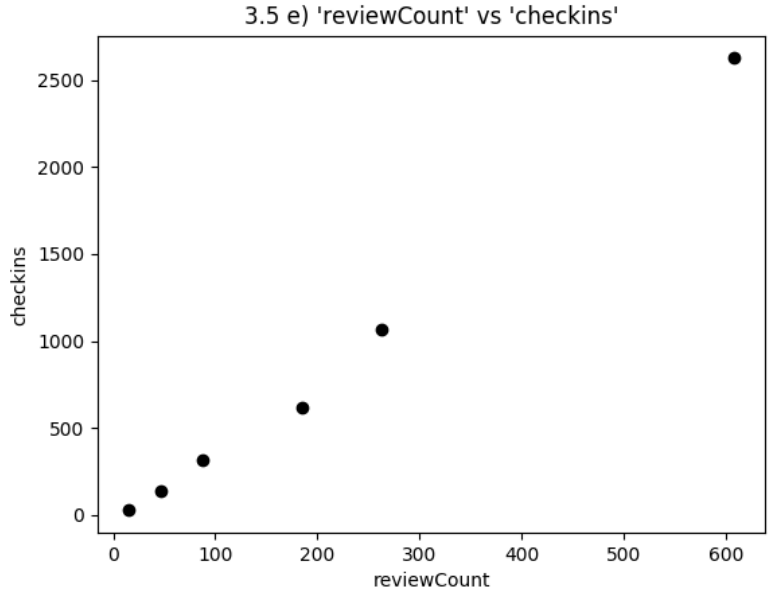
   (c)



   (d) 6; from the plot in part a), this is the *knee* in the plot so that adding another cluster doesn't improve much better the total WS-SSE.

(e) The clusters are relatively the same compared to the other plots which means that the 6% of the dataset was representative.





6.

$$\text{Score}(C, D) = f(wc(C), bc(C)) = wc(C) \cdot bc(C)$$

This scoring function considers both the compactness of the clusters as well as the separation of the clusters.