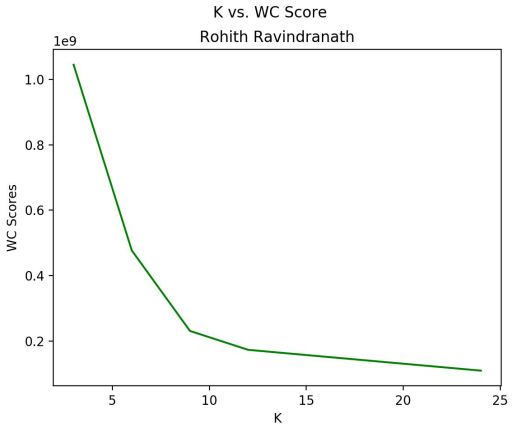
Rohith Ravindranath
Dan Goldwasser
CS 37300
30th April, 2019

Homework 5

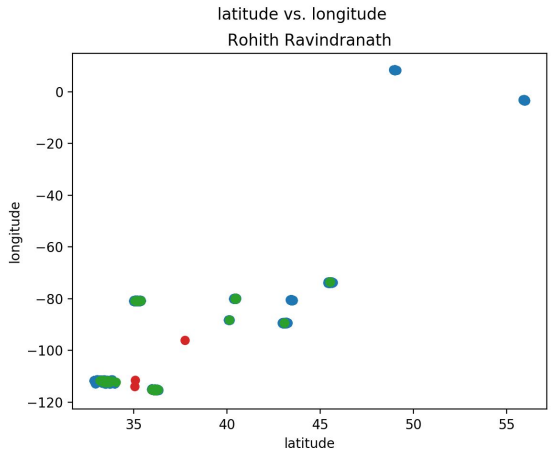
2.1 Theory

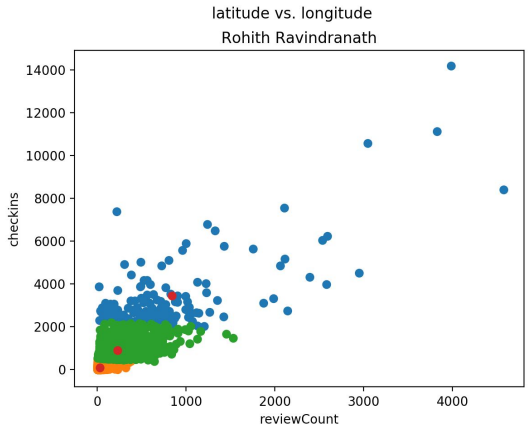
One strength of using KMeans algorithm is due to this efficiency. Since it is a relatively fast algorithm, it doesn't take that long to create clusters given an appropriate value of K. Another strength is that it is able to find spherical clusters compared to other clustering algorithms. Some of the issue is that it is sensitive to initial values of centroids, only applicable when mean is defined , one needs to manual define K, and very susceptible to outliers/noise. One should use KMeans when trying to classify a dataset and also knows how many clusters there are.

3.1



I would choose K = 9, mainly because that is the elbow point of this graph. If we increase K we notice that the change in WC-Score is not significantly significant. If we choose a K less than 9, we notice that there a dramatic change in WC-Score, this alludes that a lower K is also not the optimal choice.

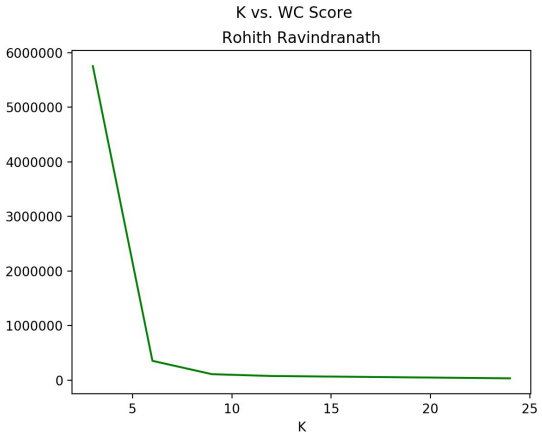




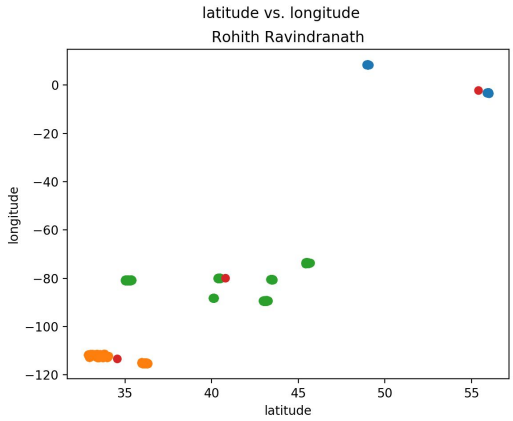
We notice that a lot of points overlap each other in the lat vs long plot, therefore making it harder to see the clusters. However, with the second plot there seems to be a wider distribution in data within the data points. This makes the data easier to see on the graph. We notice that the clusters are not spherical, rather they are closely clumped together. We also notice that the farthest cluster seems to have a lot of outliers.

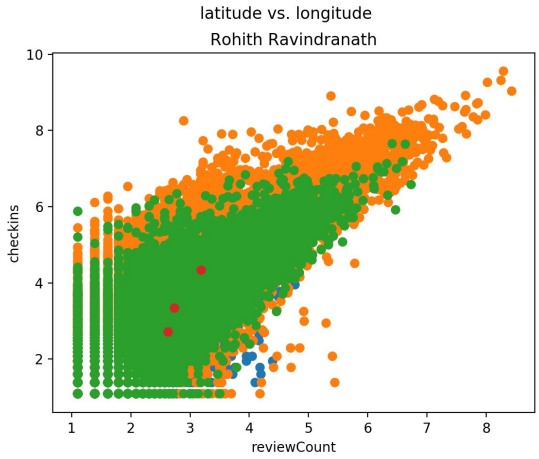
3.2

After the log transformation we would likely see a more spread out plot of the cluster as they log transformation tends to make the values more spread out.



We would choose 9 as our K value since after that there does seem to be a dramatic change in slope in the graph.

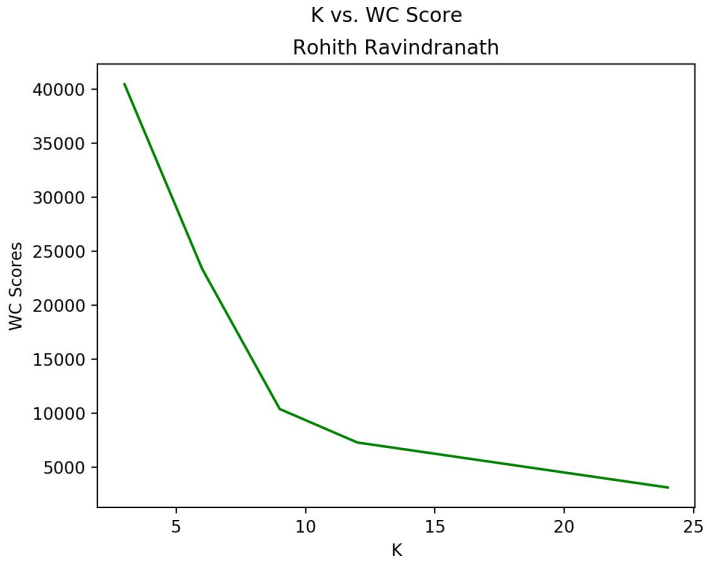




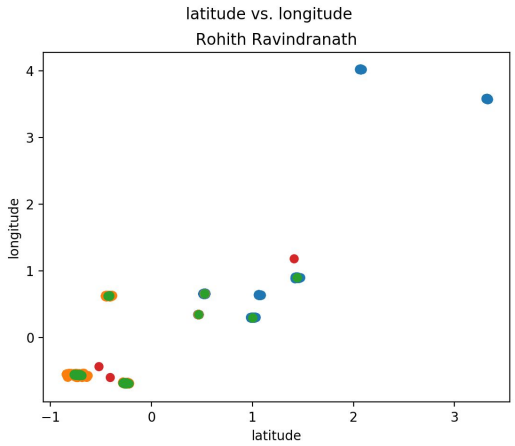
We notice that there really isn't any change in the first plot compared to the vanilla option. However in the second plot we see a drastic change, mainly because these were the values that we had transformed. We notice that the clusters overlay each other and this could be due to the log function and how it tends to decrease in slope as x increases.

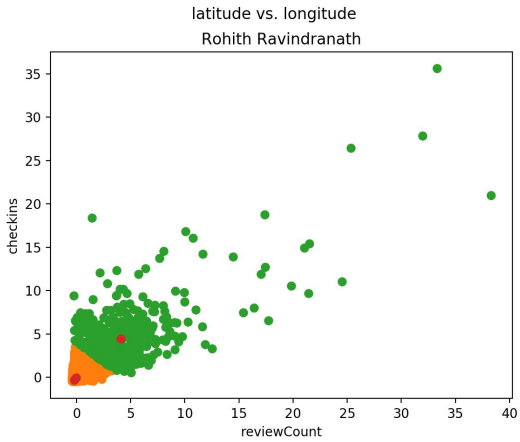
3.3

Since we are using a different method to compute distance I would expect some change in the clusters, but only in the outliers of the clusters. Overall the clusters should still look similar.



I would choose K = 9, mainly because that is the elbow point of this graph. If we increase K we notice that the change in WC-Score is not significantly significant. If we choose a K less than 9, we notice that there a dramatic change in WC-Score, this alludes that a lower K is also not the optimal choice.





We notice that the centroids have moved to different locations compared to the vanilla centroids. However, overall the clusters seem to be relatively similar. This is be due to the fact that although we are using a different distance method, the difference in distance between points should be relatively the same.