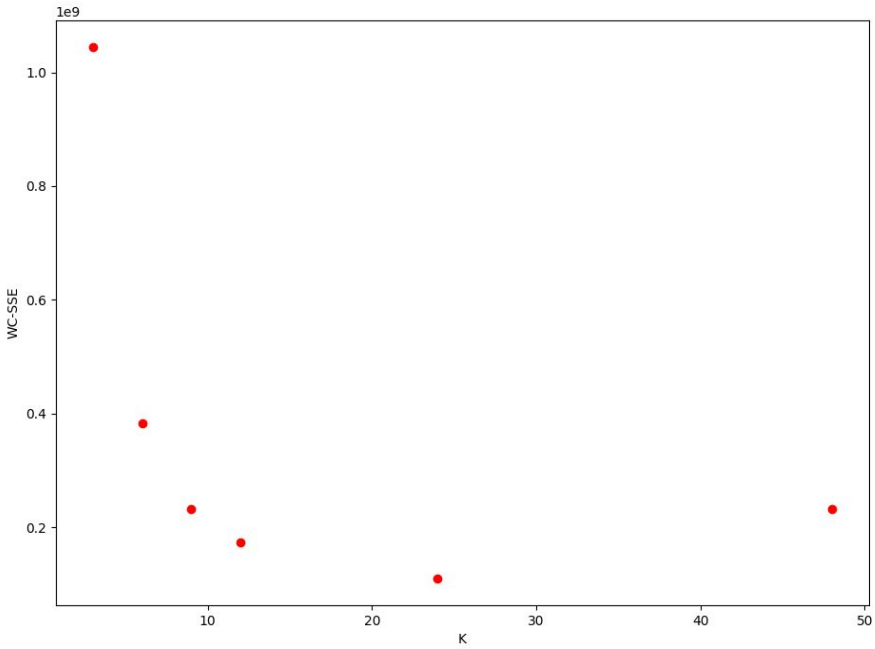
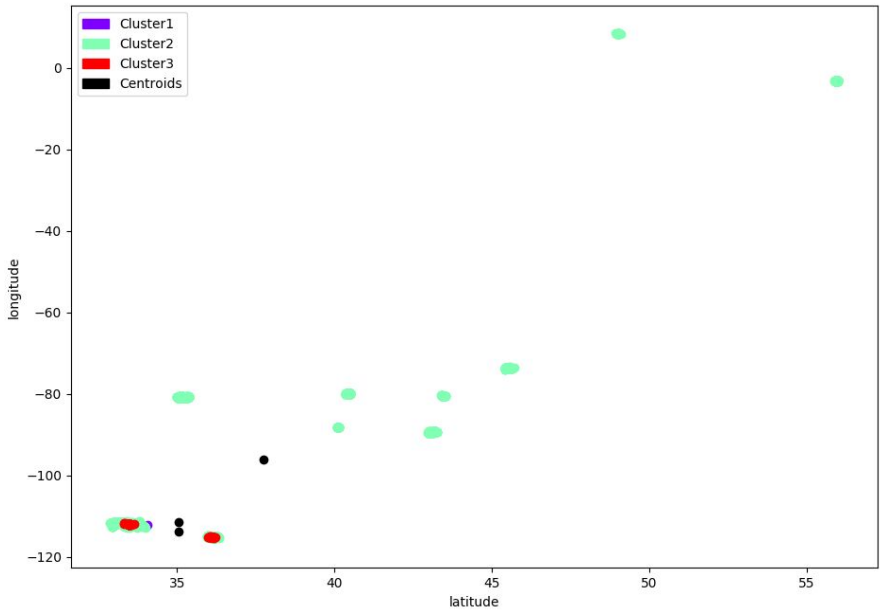
1. Jason Chen - Homework 4
2. 
    i. 

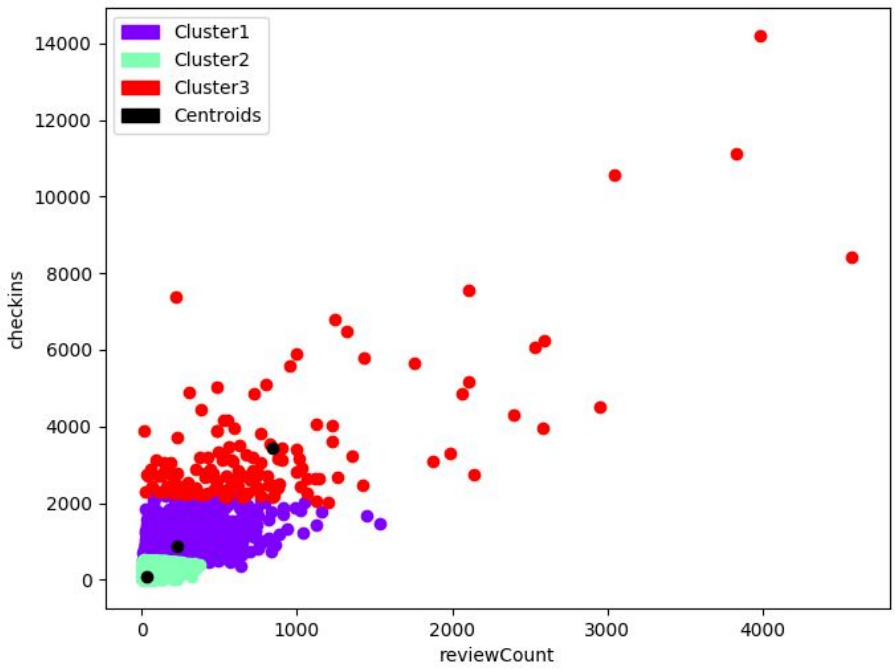



I will choose to use K value of 3 because smaller k values means less centroids and cluster, it is easier to visualize and observe a graph with 3 clusters than 48. However, the most optimal K to choose is K=24 because based off the scoring function, it has the lowest error.

The graph for latitude vs. longitude shows that the clusters aren't as tightly packed. We can see some anomalies where the latitude is near 55 and longitude is near 0. Essentially, one cluster is very loosely distributed while the other two clusters are tightly packed together near 35 latitude and -60 longitude. There really isn't a distinct pattern in the graph, but most of the points are between -60 latitude and 45 longitude.

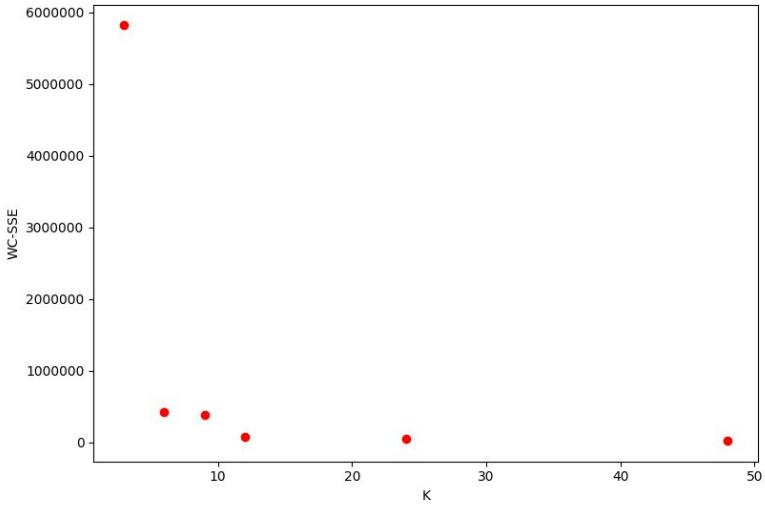



The Graph reviewCount vs checkins shows much more distinct pattern than the graph of latitude vs longitude. We can see that all the clusters are seemingly close together, almost layered on top of each other with horizontal lines splitting each cluster. Luster 3 contains all the outliers/anomalies. We can see most of the points are within 2000 review counts and 4000 checkins.



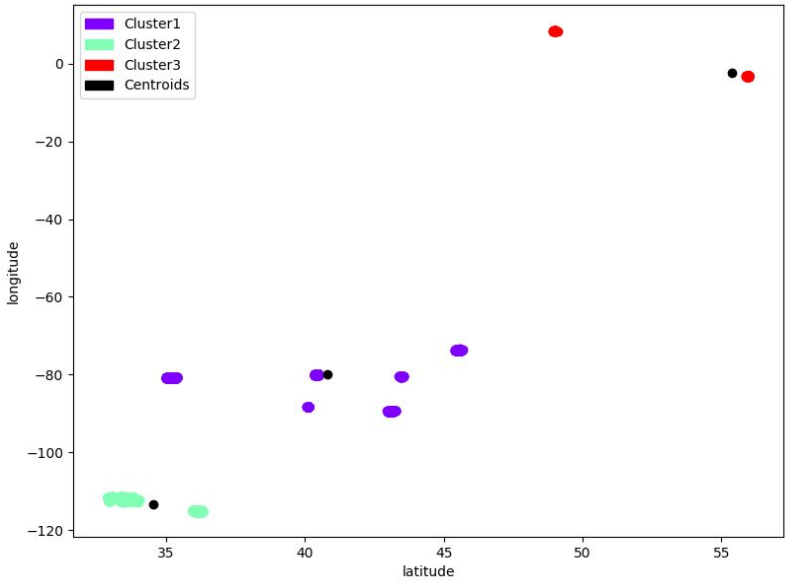

ii.

adding a log transformation should reduce skew in the graphs.

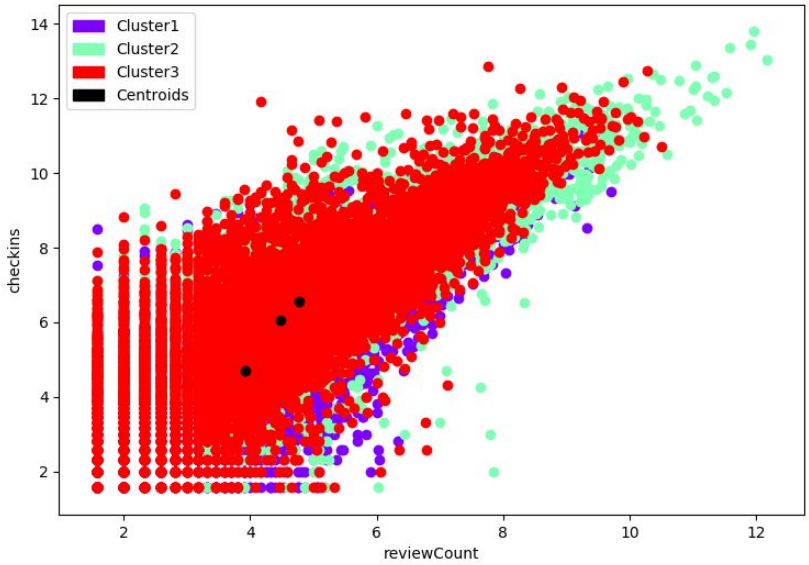



I will choose to use K value of 3 because smaller k values means less centroids and cluster, it is easier to visualize and observe a graph with 3 clusters than 48. However, the most optimal K to choose is K=12 because based off the scoring function, it has the almost the lowest error.

Latitude vs longitude: We can see that the centroids seem to lie along a line, and each cluster is more clearly separated than before.

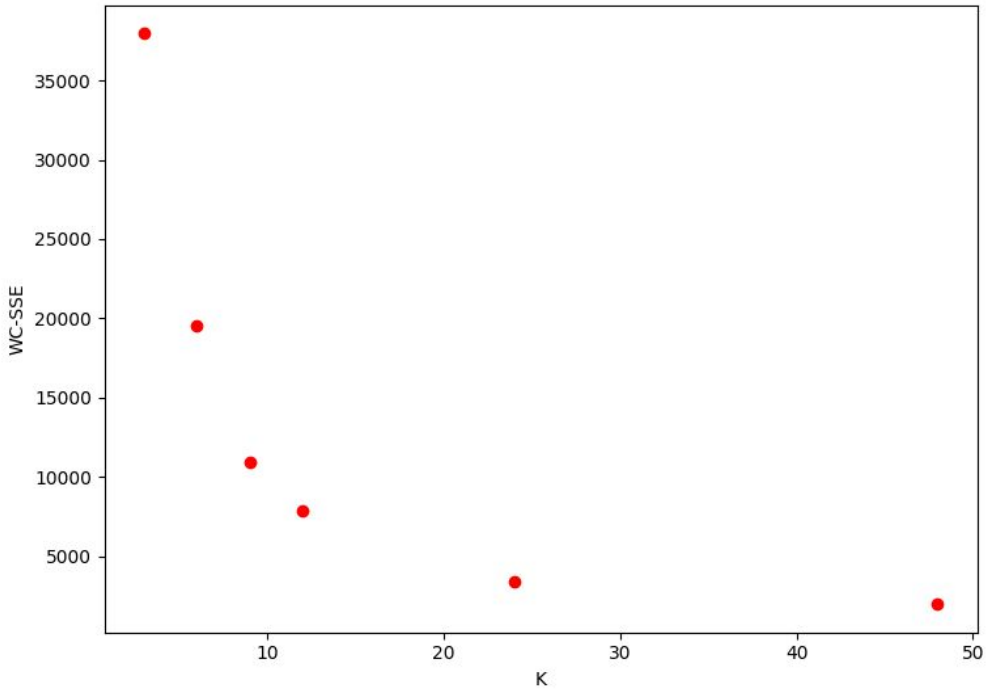



ReviewCount vs checkins: The centroids, similar to the latitude vs longitude graph, seem to lie along a line. The clusters however are less clear than the clusters seen in the graph without a log transformation.



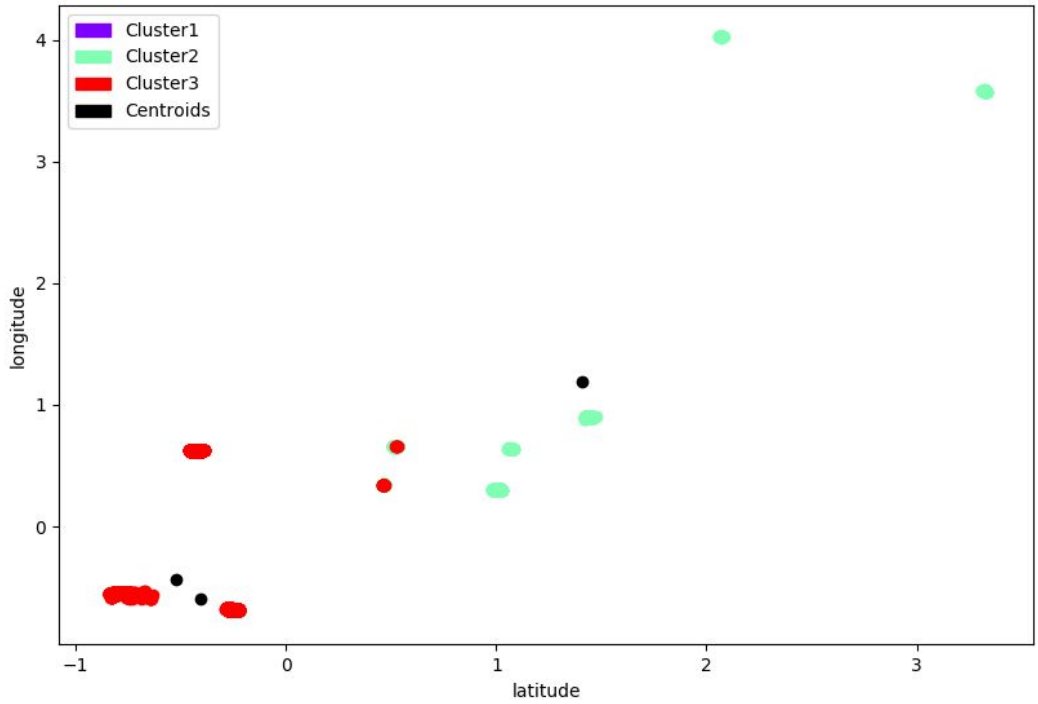

iii.

The standardized data should not change the pattern of data, but rather the ranges of each data range because all values will essentially be a Z score.

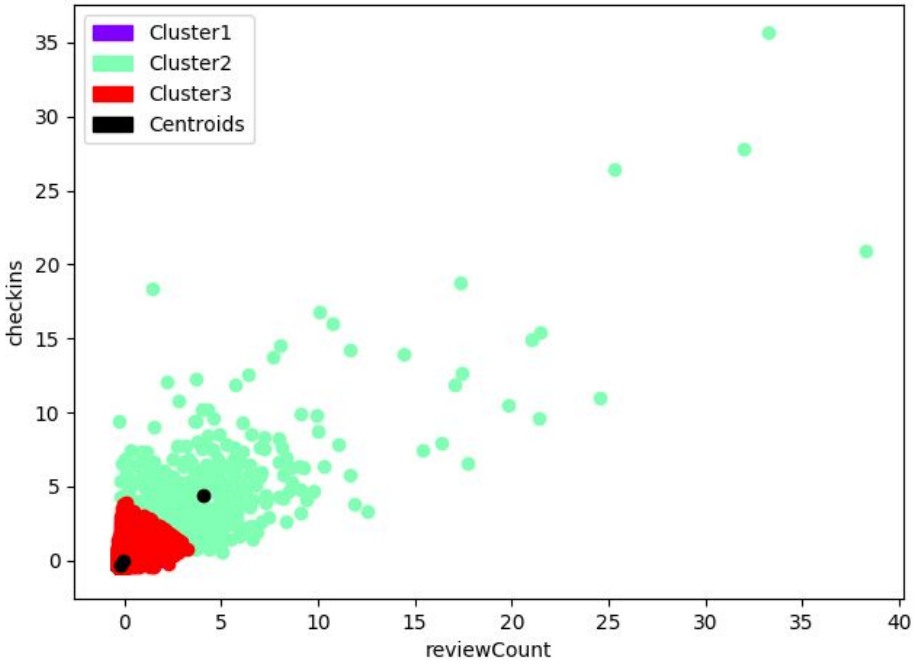


I will choose to use K value of 3 because smaller k values means less centroids and cluster, it is easier to visualize and observe a graph with 3 clusters than 48. However, the most optimal K to choose is K=48 because based off the scoring function, it has the almost the lowest error.

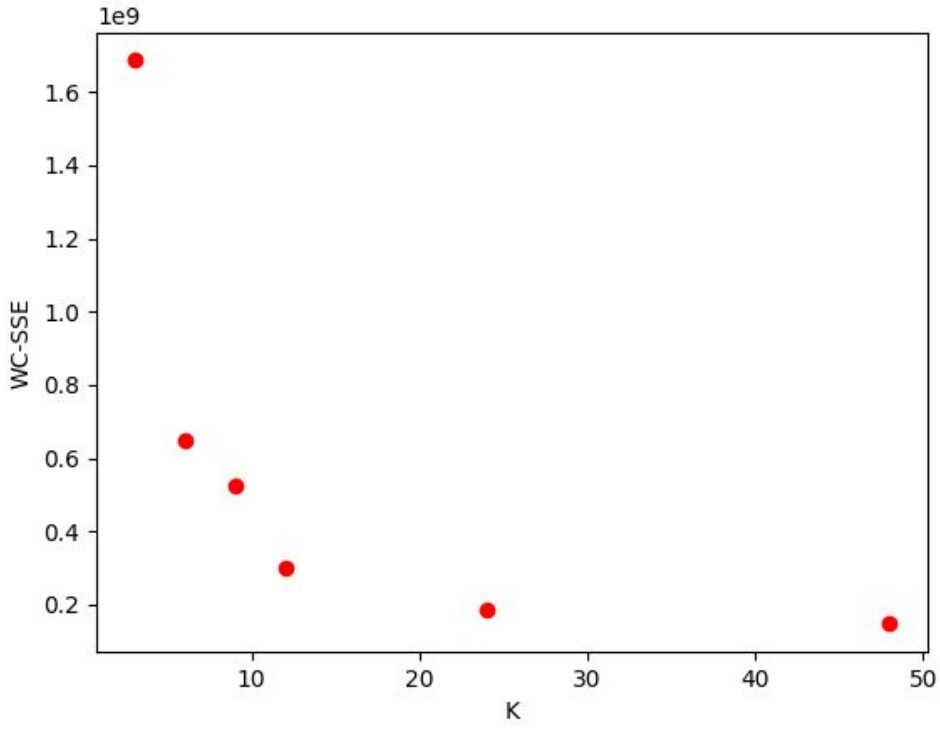
Latitude vs longitude: The pattern did not change from the non-transformed data. Most points lie within a z score of 2 latitude and 2 longitude.




reviewCount vs checkins: Two clusters are very closely packed together, not visible on the graph. One cluster(cluster2) is not as closely packed, with many pounds having a Z score of up to 40. The shape of the distribution of points is still the same as the non-transformed data.

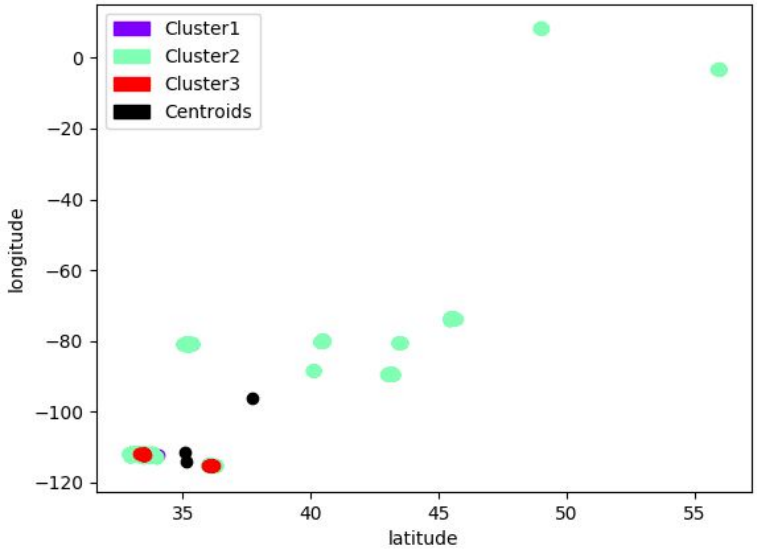


iv. There should be minimal difference between the graphs using manhattan distance and euclidean distance

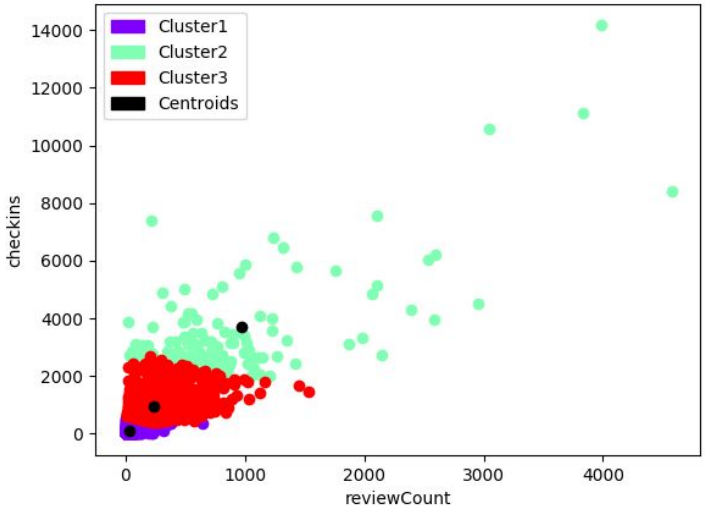



I will choose to use K value of 3 because smaller k values means less centroids and cluster, it is easier to visualize and observe a graph with 3 clusters than 24.However, the most optimal K to choose is K=24 because based off the scoring function, it has the almost the lowest error.

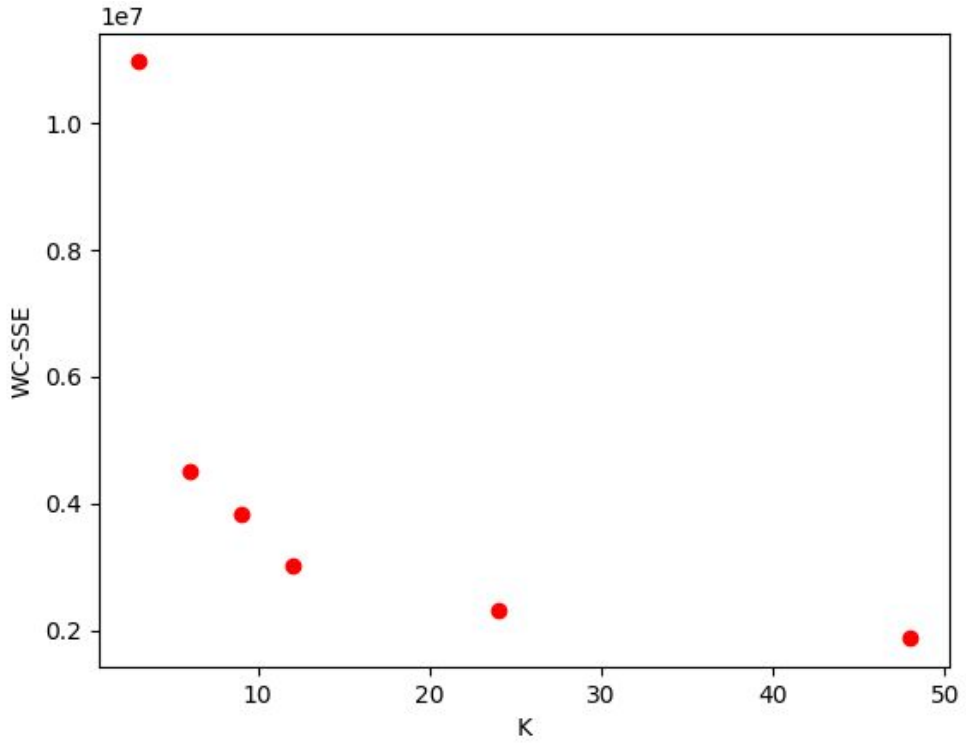
Latitude vs longitude: There seems to be little difference between the analysis using euclidean distance and the analysis using manhattan distance. We can see that one cluster seems to have a very large range of values, whereas the other two clusters are more tightly packed.




reviewCount vs checkins: There seems to be little difference between the analysis using euclidean distance and the analysis using manhattan distance. We can see that each cluster is somewhat separate and all clusters can be seen on the on the graph.

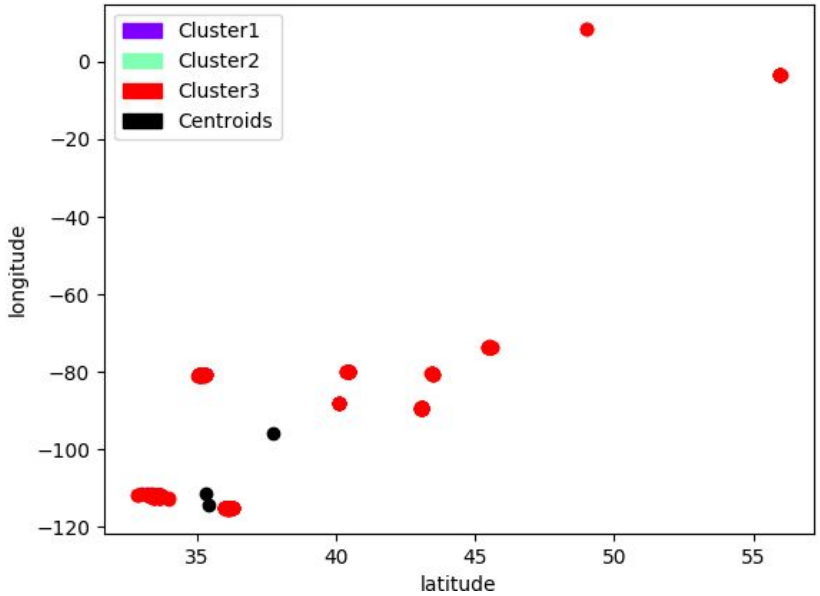



v.



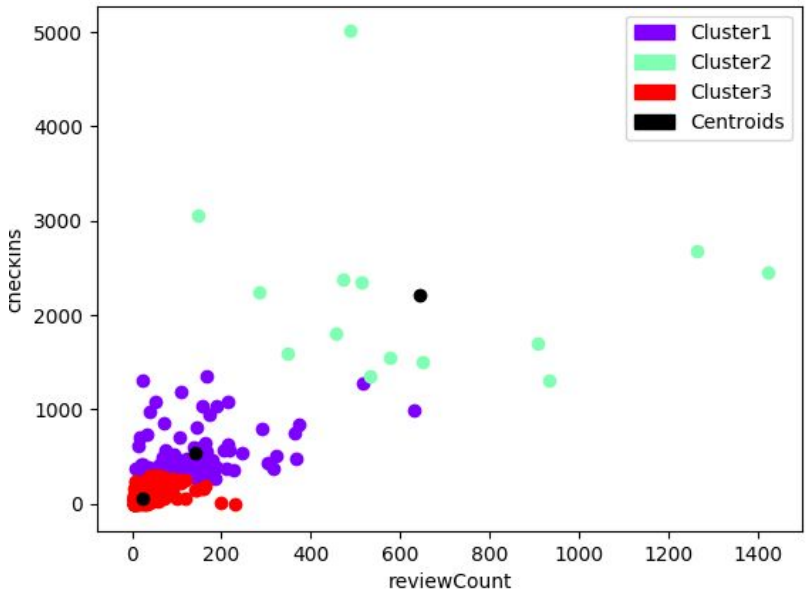

I will choose to use K value of 3 because smaller k values means less centroids and cluster, it is easier to visualize and observe a graph with 3 clusters than 24.However, the most optimal K to choose is K=24 because based off the scoring function, it has the almost the lowest error.

Latitude vs longitude: Graph seems to not change despite the small sample size of only 3% the original data. However this still is around 632 points. One cluster has a range of values while the other two cluster are very tightly packed around 35 latitude and -110 longitude.




reviewCount vs checkins: The points have a similar shape compared to the graph with all data points. The clusters seem to be more separated. There is less overlap between

the clusters. One centroid is farther away from the other two centroids, which is to be expected because the points in that centroid vary largely.




3. F(C) =

$$= \frac{\sum_{k=1}^{K} \sum_{x(i) \epsilon C_k} d(x(i), r_k)^2}{\sum_{k=1}^{K} \sum_{p=1}^{K} d(r_k, r_p)^2}$$

Where $r_k$ and $r_p$ are centroids of cluster $C_k$ and $C_p$

The goal of this scoring function is to reduce the error if the sum of the distances of each cluster is high. The higher the differences between each cluster, the lower the error. The numerator portion of the scoring function is the Within-cluster sum of squared error.

I will be choosing K = 9, this is the point past the “knee” of the graph. This point has a moderate number of cluster that are still visible if graphed, and based off my scoring function has the least number of cluster while also having the most separating of clusters.