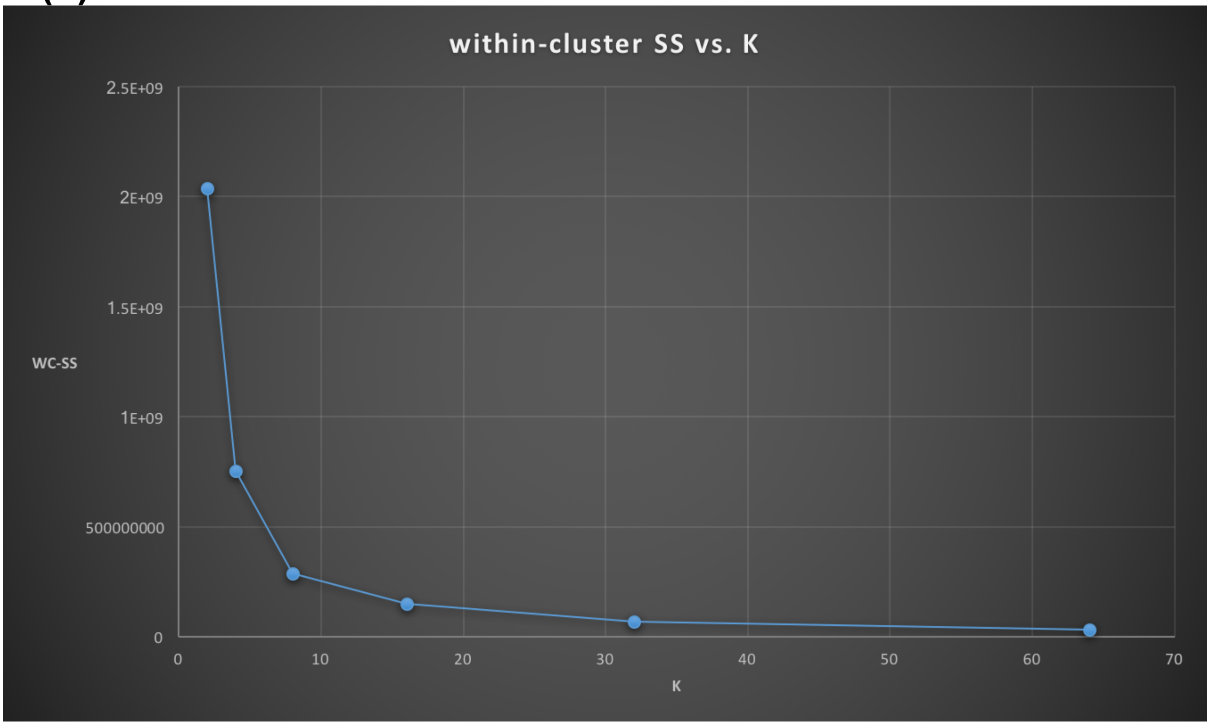
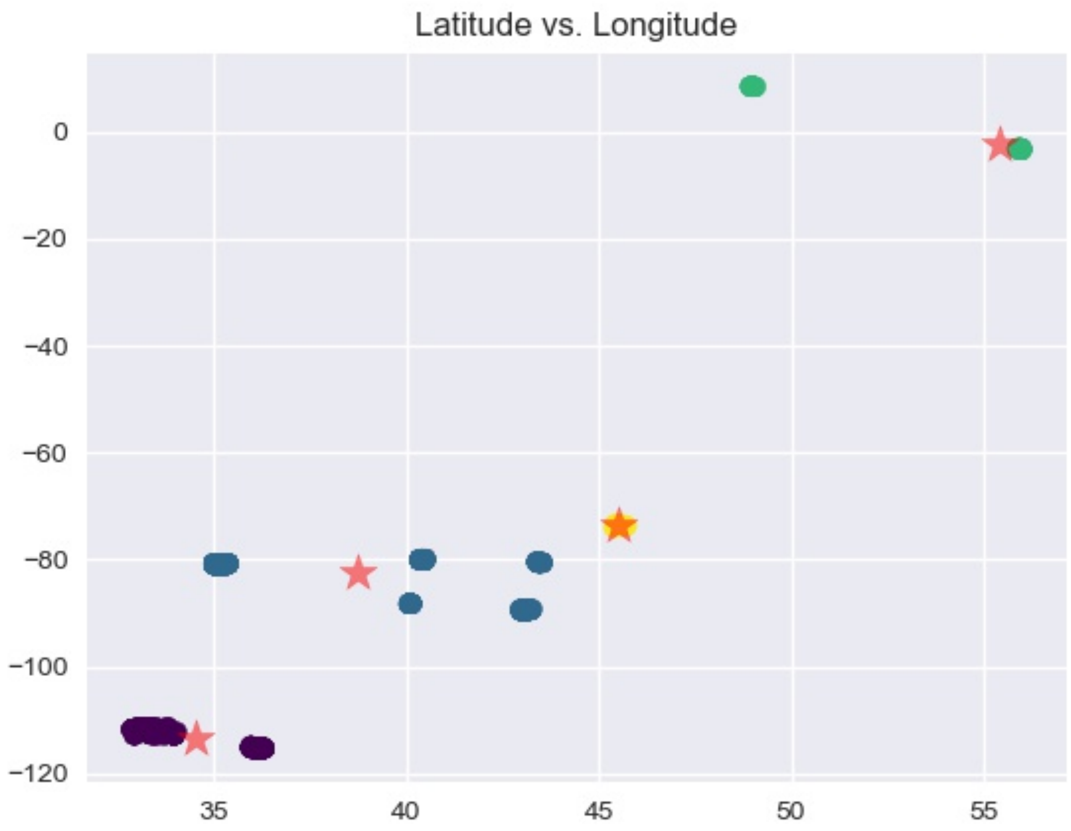
2/(a)

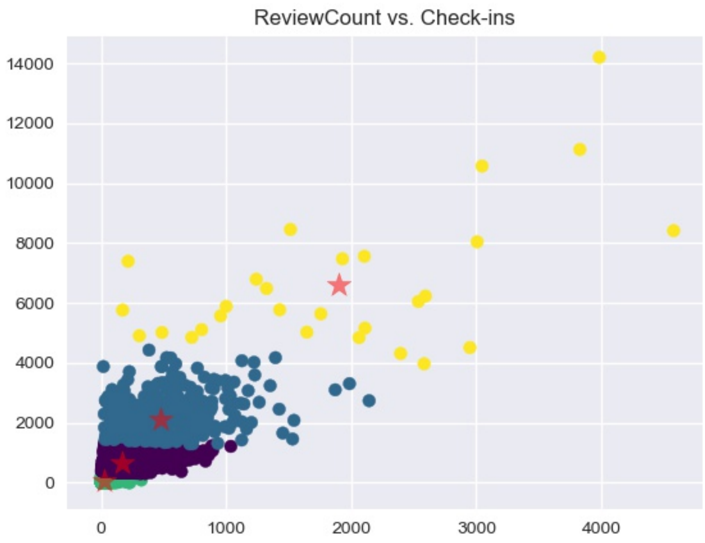


I would choose k=8 because it is the “knee point” in the plot. Meaning that adding another cluster will not improve the modeling of the data by a significant amount.

/(b)



(Note: the 3rd centroid(orange) overlaps with the points(yellow) in its cluster. It can be hard to tell from the graph)



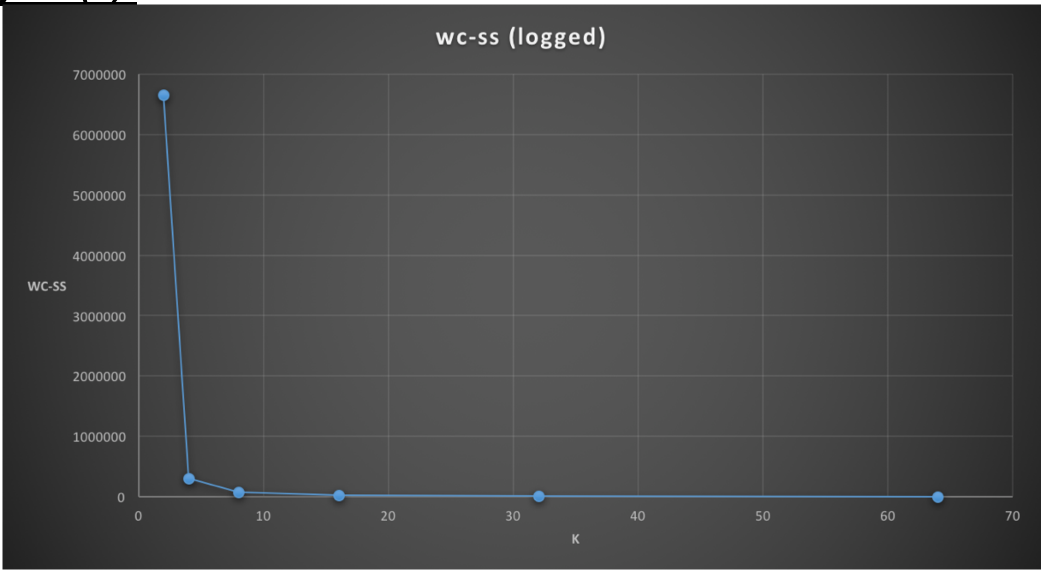
reviewCount and checkins are most taken into account by the clustering model.

From the graph, we can see latitude and longitude features are mostly already stay in a “cluster” before clustering. The data points are grouped

in a specific area with very low variance. Therefore, those two features will not affect the clustering model significantly.
However, reviewCount and checkins have a large variance and the dataset cannot be easily clustered before the clustering process. Therefore, those two features will determine the clustering modal.

/(c)

analysis (a):

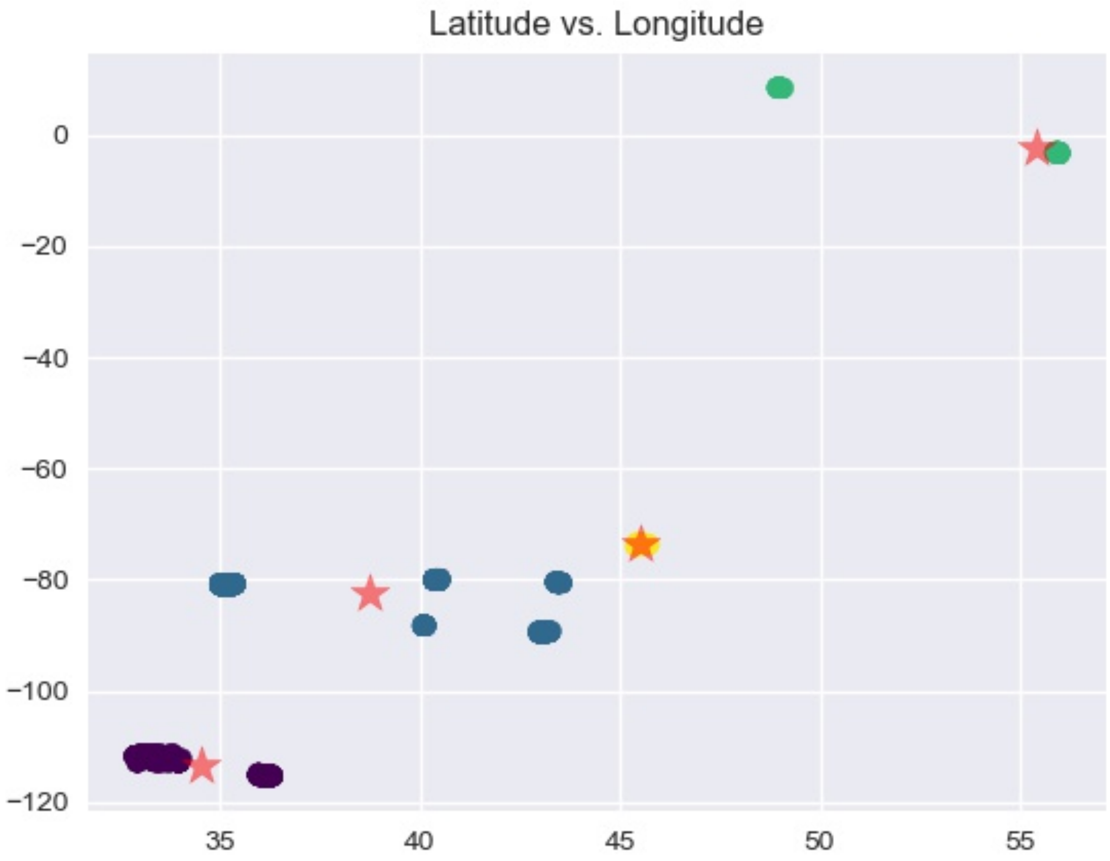


I would choose k=4 because it is the “knee point” in the plot. Meaning that adding another cluster will not improve the modeling of the data by a significant amount.

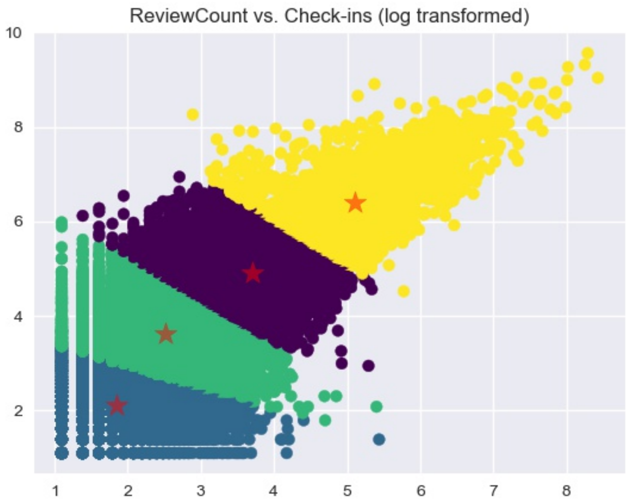
Difference:

The result is different from last one. The “knee point” is much smaller; The variance after “knee point” is also smaller; The “drop” before “knee point” is steeper. The reason for the differences is that we performed log transformation on “reviewCount” and “checkins” attributes. That affects the clustering model. The distribution is closer to a normal distribution after perform log transformation.

analysis (b):



(Note: the 3rd centroid(orange) overlaps with the points(yellow) in its cluster. It can be hard to tell from the graph)



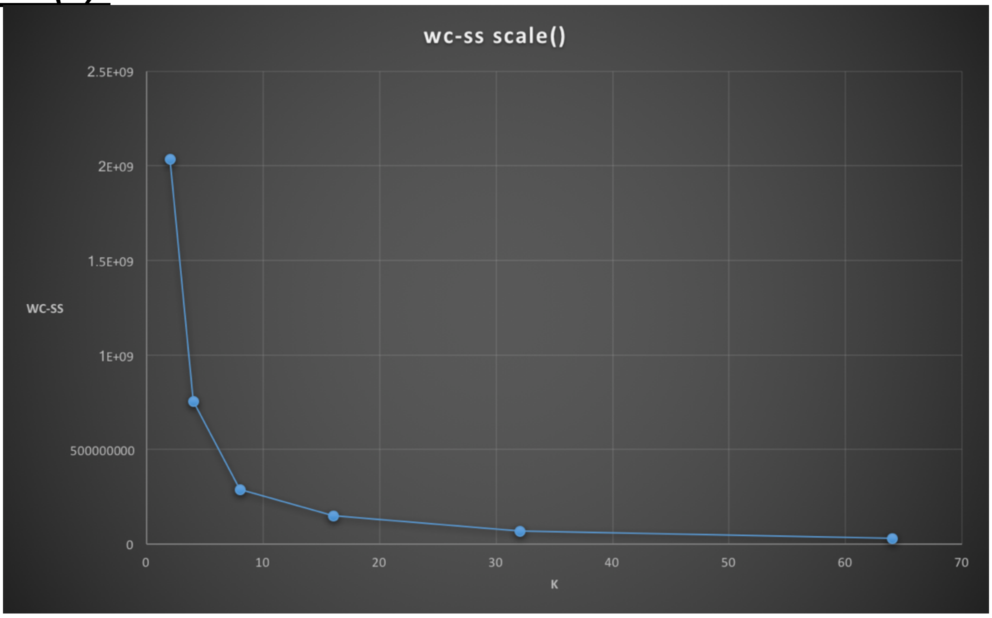
reviewCount and checkins are most taken into account by the clustering model.

From the graph, we can see latitude and longitude features are mostly already stay in a “cluster” before clustering. The data points are grouped

in a specific area with very low variance. Therefore, those two features will not affect the clustering model significantly.
However, reviewCount and checkins have a large variance and the dataset cannot be easily clustered before the clustering process. Therefore, those two features will determine the clustering modal.
Difference:
“Latitude vs. Longitude” stays the same since we didn’t not change the data set.
“reviewCount vs. checkins” was very noisy and was heavily skewed, the clusters were not optimized. After the log transformation, the dataset is less skewed, closer to normal distribution and therefore the clusters are optimized.

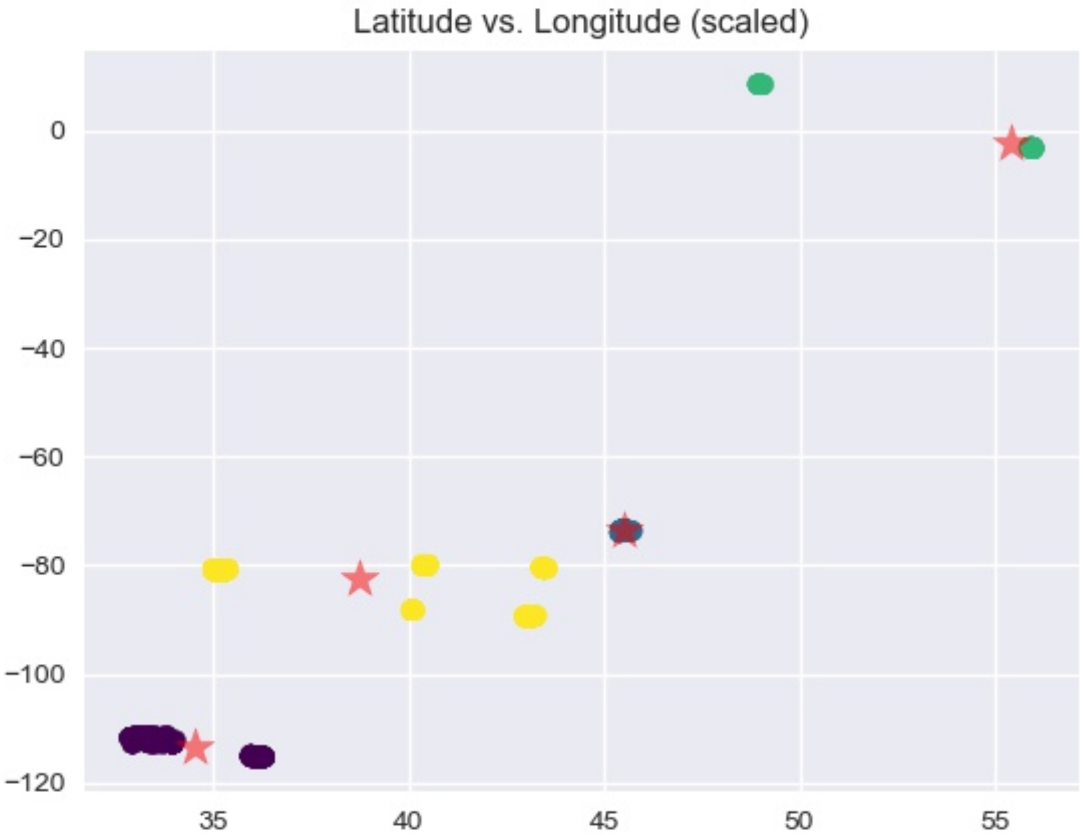
/(d)

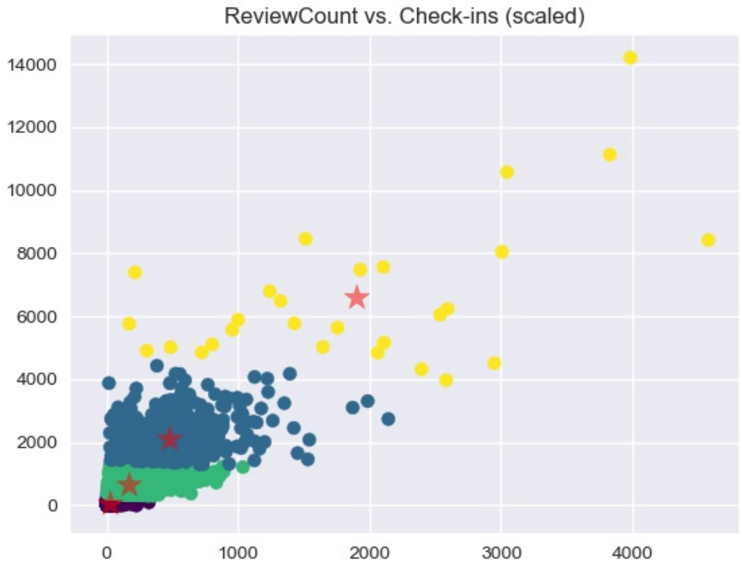
analysis (a):



I would choose k=8 because it is the “knee point” in the plot. Meaning that adding another cluster will not improve the modeling of the data by a significant amount.

analysis (b):



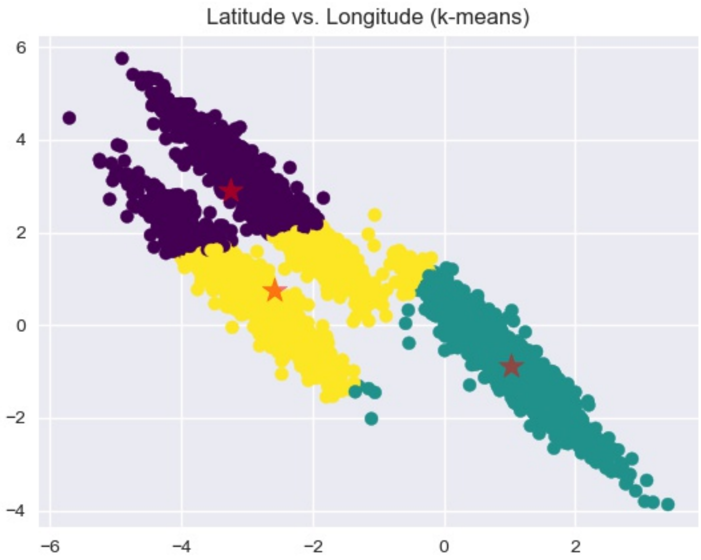


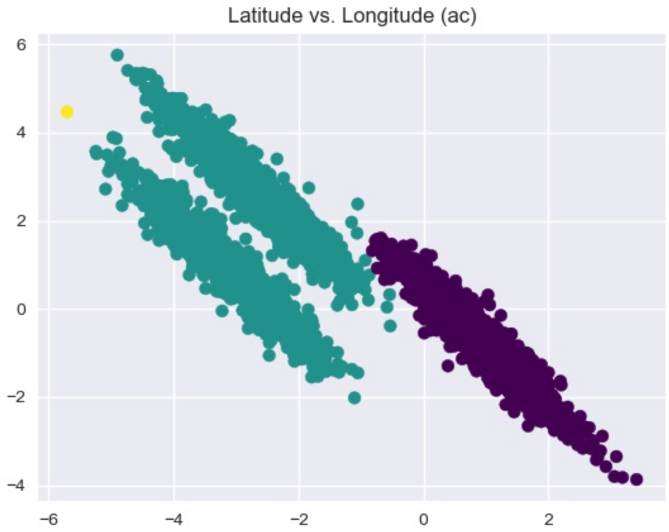
reviewCount and checkins are most taken into account by the clustering model.

From the graph, we can see latitude and longitude features are mostly already stay in a “cluster” before clustering. The data points are grouped

in a specific area with very low variance. Therefore, those two features will not affect the clustering model significantly.
However, reviewCount and checkins have a large variance and the dataset cannot be easily clustered before the clustering process.
Therefore, those two features will determine the clustering modal.
Difference:
The scale() transformation made mean=0, stdev=1, it optimizes data and represent a good measure of distance, especially for Latitude vs. Longitude.

3/(a)





K-means performs a lot better.
The distribution between each cluster in K-means is approximately even, which is not the case in Agglomerative Clustering.

/(b)

No. K-means does not always yield the same result if applied multiple times because we start with k random points as our initial clusters.
Agglomerative Clustering yields the same result if applied multiple times.

/(c)

Agglomerative clustering is going to take more time.
The time complexity for AC (average-linkage) is generally O(n^2).
The time complexity for k-means is O($k * n * i$) where i is the number of iterations.