Q4:

K-means clustering is faster. K-means clustering running time is O(n), whereas Hierarchical clustering running time is O(n^2).

Q7:

hierarchical\_clustering

1.75163886916e+11

kmeans\_clustering

2.71254226924e+11

Q8:

The cluster centers generated by the two methods are similar. However, a major differences are the clusters centers located on the West coast. Kmeans-clustering generated three cluster centers in California, while hierarchical clustering generated only 2 centers in this state. Since this method initiate centers using counties with the highest population, and California is the most populous state, it is expected that cluster centers tend to localize in this state.

Q9:

Hierarchical clustering produces clusters with lower distortion.

Q11:

For smaller data set (111-county set), hierarchical clustering consistently produces lower distortion. However, in larger data set (290-, 896-county set), there distortions are large for both methods, and there is no advantage (in terms of minimize distortions) in choosing one over another. However, k-means clustering tend to be much faster, hence, for large data set, this is method superior.

For smaller data set, use hierarchical clustering since it is less biased. For k-means clustering, the initial centers are arbitrary. However, if there is sufficient iteration, I image this approach will provide better performance. In general, hierarchical clustering is very slow, while k-means clustering is a lot faster, and for large data sets, the distortion error are similar for both approaches. Hence, for large data set, k-means clustering is preferred as it is much faster.