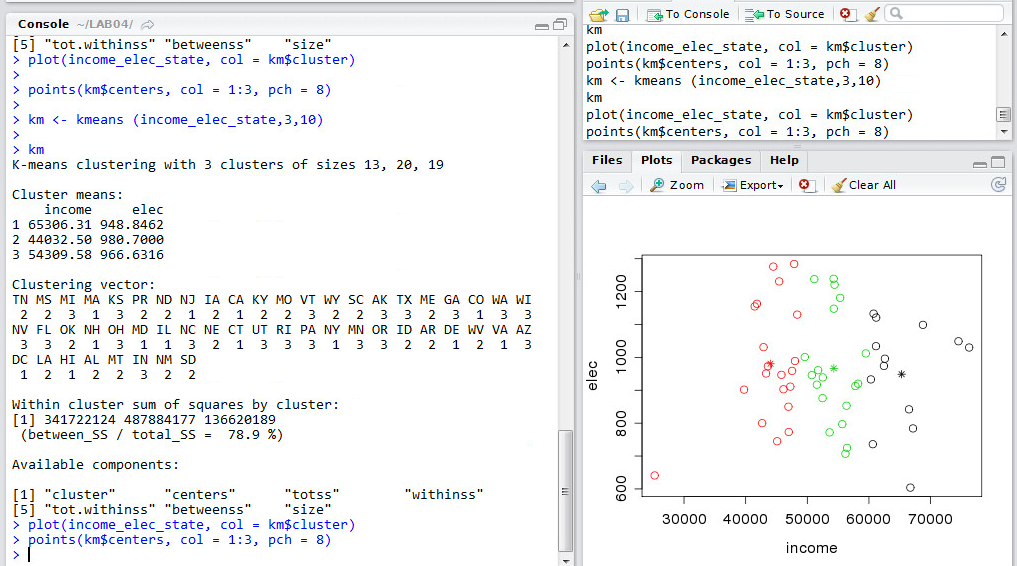
CSC570Q Data Science Essentials

**Module 4 – Advanced Analytics/Methods** **K-means Clustering**

1. Cluster the data and plot all 52 data points, along with the centroids. Mark all data points and centroids belonging to a given cluster with their own color. Here, let k=10.

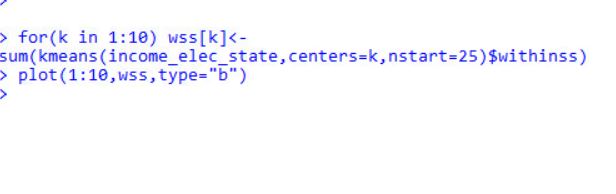


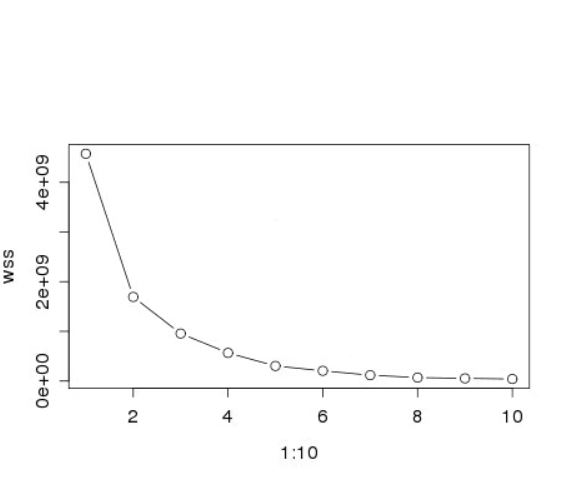
1. Repeat step (a) several times. What can change each time you cluster the data? Why? How do you prevent these changes from occurring?

After multiple repetitions the point where the centroids are plotted are changing for every iteration. This is because the assigned points oscillate back and forth from one iteration to other. The homogeneity of the data changes as we plot the graph many times. This is due to the change in the nstart value. This can be prevented by increasing the nstart value drastically and the homogeneity of data is better when this is done.

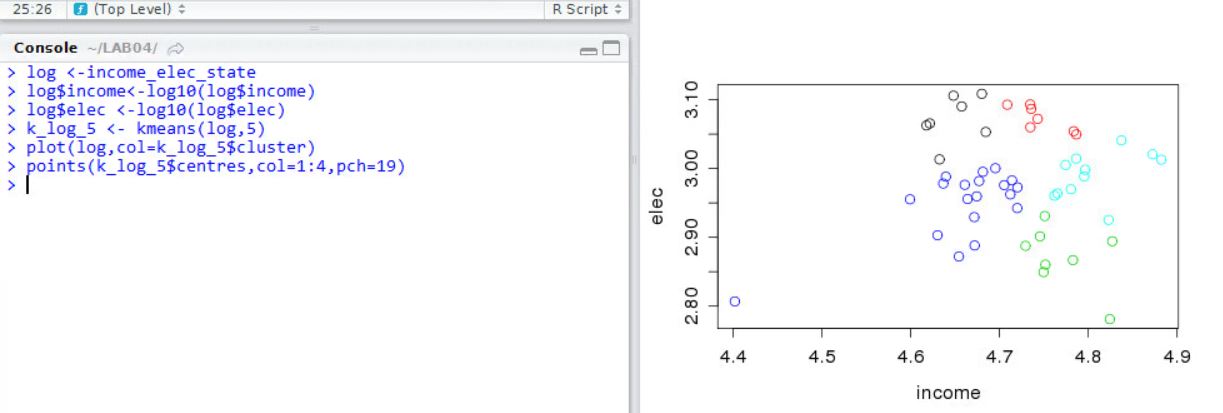
1. Once you’ve accounted for the issues in the previous step, determine a reasonable value of k. Why would you suggest this value of k?

The reasonable value for k I would think is 5 of the cluster section and there are no changes in the graph which is being plotted.





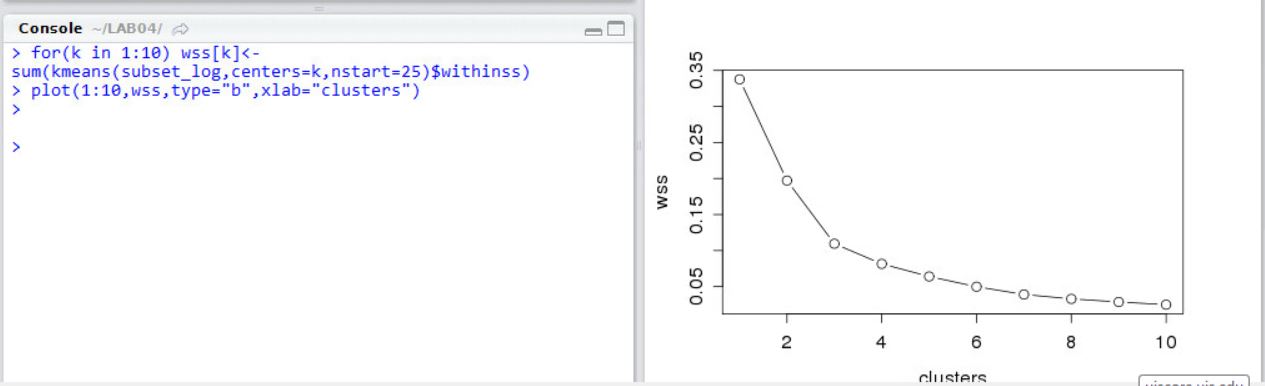
1. Convert the mean household income and mean electricity usage to a log10 scale and cluster this transformed dataset. How has the clustering changed? Why?



The data points are placed at different positions when we convert it into the log10.

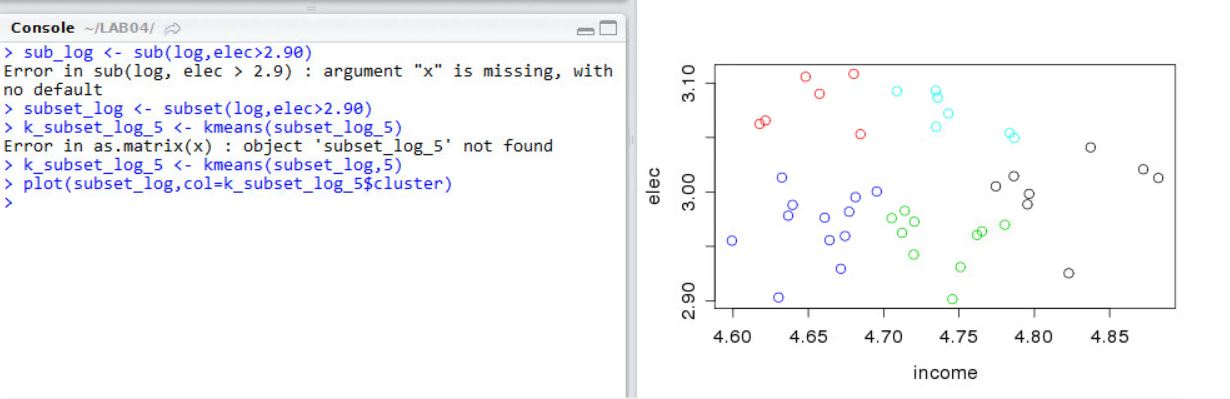
1. Reevaluate your choice of k. Would you now choose k differently? Why or why not?

The value of k I feel should be changed because the values are changing drastically and if the values change the graph would not be proper and we cannot analyze what is happening for the data being plotted.



1. Have you observed an outlier in the data? Remove the outlier and, once again, reevaluate your choice of k.

As I observed there were couple of outliers in the graph. The value of k



1. Color a map of the U.S. according to the clustering you obtained. To simplify this task, use the “maps” package and color only the 48 contiguous states and Washington D.C.

