Big Data Analytics Homework Assignment 2
K-means and Hierarchical Clustering

Due: Tuesday, 2/18, by 11:59pm

Deliverables: You will submit an R script (a .R file) with commands to answer the questions below, and comments clearly labelling each answer. You will also submit PDFs with the plots. You can submit these as separate files, or you can compress them into a single zip file.

The Data: The file attached to the assignment in Blackboard provides data on penguins observed in Antarctica.

Preparation:

You can download and save the csv file attached to the assignment in Blackboard OR you can use wget at the terminal to copy the file to your virtual machine:

wget https://cse.sc.edu/~bhipp/587/penguins.csv --no-check-certificate

Additional Instructions of Note:

Your answer for each question must be clearly labelled in your R script and must include the command(s) necessary to generate the answer. Unlabeled answers and answers that don't contain the R command(s) to justify them will not receive credit.

The R commands must be submitted in an R script (.R) file.

The questions / problems:

1. Name a data type that cannot be used for k-means clustering. (Answer this question in a comment in your .R file – your answer does not require an R command)
2. Is k-means clustering deterministic? In other words, if we repeatedly perform k-means clustering on the same data with the same value of k, are we guaranteed to get the same clusters each time? Explain / justify your answer.
3. Use k-means clustering on the four numeric columns in the dataset from penguins.csv. Use a for loop to determine the within-cluster sum of squares error for clusters of size (k) 1 through 20, and plot these sum of squares values (as we did during the in-class lab).
    - Set the random number generator seed to 1801 before **each** call to the kmeans function within the loop (using the command set.seed(1801)).
    - Save your plot to a PDF file. This PDF must be submitted with your assignment.
4. Based on the plot created for the previous question, choose a reasonable value of k. What k did you select? Why did you select this k? (Answer this question in a comment in your .R file)
5. Use the kmeans function with the value of k chosen in question 4. Plot the four-dimensional penguins data (the four numeric fields) with each cluster shown in a different color.
    - To ensure consistent results, set the random number generator seed to 1801 before calling the kmeans function.
    - Save the plot to a PDF file. This PDF must be submitted with your assignment.

6. Normalize the numeric columns in the penguins dataset using the normalization functions we wrote for the in-class k-means lab. Using the normalized data, construct a for loop to determine the within-cluster sum of squares error for k values of 1 through 20, and plot these sum of squares values (as we did during the in-class lab).
    - Set the random number generator seed to 1801 before **each** call to the kmeans function within the loop (set.seed(1801)).
    - Save your plot to a PDF file. This PDF must be submitted with your assignment.
7. Based on the plot created for the previous question, choose a reasonable value of k. What k did you select? Why did you select this k? (Answer this question in a comment in your R script)
8. Use the kmeans function with the value of k chosen in question 7. Plot the four-dimensional penguins data (the four numeric fields) with each cluster shown in a different color.
    - Set the random number generator seed to 1801 before calling the kmeans function.
    - Save the plot to a PDF file. This PDF must be submitted with your assignment.
    - *Graduate students, this may be a good time to answer question 13, below.*
9. Create a distance matrix with the dist function, using Euclidean distance, for the **normalized data** created in question 6.
10. Use the hclust function to create the hierarchy on the distance matrix created in question 9, using the "ward.D2" method. Plot the hierarchy with the plot function.
11. Using the plot created for question 10, determine a reasonable number of clusters. When determining this number, remember that shorter branches denote more similarity between the linked clusters, and longer branches denote more dissimilar / distant clusters.
    - How many clusters did you choose?
    - Use the rect.hclust function to outline the clusters in the plot.
    - Save the plot with the outlined clusters to a PDF file. This file must be submitted with your assignment.

Graduate Students Only:

12. Use the cutree method to determine the cluster to which each penguin record is assigned using the hierarchical clustering analysis you completed in questions 9-11. Plot the four-dimensional penguins data (the four numeric fields) with each of the clusters shown in a different color. Save the plot to a PDF file. This file must be submitted with your assignment.
13. Describe differences between the plots created for problem 5 and problem 8. Explain the reason(s) for those differences.