STAT330/430 - Introduction to Statistical Learning

Assignment 4: Chapter 10
Due date - May 23rd

Note: You must submit two files for this assignment:

- a pdf document of your solutions, and
- a complete and concisely annotated *R Script* file of all **R** analysis that was undertaken to produce your results.

In addition, to receive marks for Question 1 you are required to engage in the Topic 4 moodle discussion forum.

Please refer to the Assignment assessment criteria document for additional guidance.

Question 1 [5 marks]

In each assignment you will be able to earn up to 5 marks based on your engagement on the moodle Discussion forums from the topics associated with the given assignment. See the Assignment assessment criteria document for more details.

Question 2 [25 marks]

Spinner dolphins are a small delphinid that reside in subtropical and tropical waters. I collected data on spinner dolphins in Fiji to explore characteristics of their whistles. The type of whistle (e.g., concave, constant, convex, downsweep, sine, and upsweep) along with a variety of other acoustic properties of individual whistles was measured and is summarised in the *Spinner_dolphin* data-set.

Variable	Description
Whistle	Whistle classification (Concave, Constant, Convex, Downsweep, Sine or Upsweep)
Duration	The time span of the whistle (seconds)
Centre_Freq	The frequency recorded at the midpoint (centre) of the whistle (kHz)
Low_Freq	The lowest frequency recorded during a whistle (kHz)
Delta_Freq	The range in frequency recorded during a whistle (kHz)
Max_Freq	Maximum frequency (kHz) recorded during a whistle (kHz)
Range_50	Centre frequency [minus] minimum frequency
Range_100	Maximum frequency [minus] centre frequency
Inflections	Number of points of change in whistle curvature

(a) Conduct an exploratory data analysis and present a summary of the *Spinner_dolphin* data-set.

(b) Find a low-dimensional representation of the numerical variables in the *Spinner_dolphin* data-set. Provide carefully labeled and well-formatted tables and/or figures - as well as an informative summary of your model and results. In addition, provide a visualisation and brief summary of whistle type in the context of your low-dimensional representation.

Question 3 [15 marks]

The Pacific Islands region stretches from the eastern coast of Australia all the way to French Polynesia and is comprised of a wide variety of island countries and territories. Exploring the similarities and characteristics of these nations is useful for environmental management and planning. A summary of the variables contained in the *Pacific_Islands* data-set is listed below.

Variable	Description
PICT	Pacific Island Country or Territory
Land_area	Total land area (km^2)
Coastline	Total distance of coastlines (km)
EEZ	Economic Exclusive Zone area (km^2)
Shelf_Area	Continental shelf area (km^2)
Inshore_fishing_area	Continental shelf area accessible for inshore fishing (km^2)
Coral_reefs	Percentage of the world's coral reefs contained in this PICT
Seamounts	Percentage of the world's seamounts contained in this PICT
Primary_production	Estimated primary production for the marine environment (mg of carbon/m^2/day)
MPA	"Marine protected areas" (km^2)
PA	Terrestrial "protected areas" (km^2)
Pop	Population size of the PICT
Total_spp	Total number of animal species listed for this PICT
Region	Melanesia, Polynesia or Micronesia

(a) Use k-means clustering to identify the optimal number of clusters for the *Pacific_Islands* data-set. Visualise your results, provide some broad insights into cluster membership in relation to Region, and provide a brief summary of your results.

(b) Create at least two different types of dendrograms based on the *Pacific_Islands* data-set. Use *PICT* as your label on each leaf for these plots. Using the same number of clusters in each dendrogram compare and contrast your results. Provide a general interpretation of groupings seen in each of the dendrograms you produce.