CptS 483-04: Introduction to Data Science, Fall 2017

Assignment 4: Data Wrangling and Unsupervised Learning

Release Date: October 23, 2017 **Due Date:** October 28, 2017 (11:59 pm)

For this assignment you will be using the dplyr package to manipulate and clean up a dataset. The dataset is called msleep (mammals sleep), and is available on the course webpage (at https://scads.eecs.wsu.edu/wp-content/uploads/2017/10/msleep_ggplot2.csv). The dataset contains the sleeptimes and weights for a set of mammals. It has 83 rows and 11 variables. Here is a description of the varaibles:

column name	Description
name	common name
genus	taxonomic rank
vore	carnivore, omnivore or herbivore?
order	taxonomic rank
conservation	the conservation status of the mammal
sleep_total	total amount of sleep, in hours
sleep_rem	rem sleep, in hours
sleep_cycle	length of sleep cycle, in hours
awake	amount of time spent awake, in hours
brainwt	brain weight in kilograms
bodywt	body weight in kilograms

Load the data into R, and check the first few rows for abnormalities. You will likely notice several.

Below are the tasks to perform. You are encouraged to use R Markdown to generate your report (in PDF).

a) Use select() to print the head of the columns with a title including "sleep".
b) Use filter() to count the number of animals which weigh over 50 kilograms and sleep more than 6 hours a day.
c) Use piping (%>%), select() and arrange() to print the name, order, sleep time and bodyweight of the animals with the top 6 sleep times, in order of sleep time.
d) Use mutate to add two new columns to the dataframe; wt_ratio with the ratio of brain size to body weight, rem_ratio with the ratio of rem sleep to sleep time. If you think they might be useful, feel free to extract more features than these, and describe what they are.
e) Use group_by() and summarize() to display the average, min and max sleep times for each order. Remember to use ungroup() when you are done.
f) Make a copy of your dataframe, and use group_by() and mutate() to impute the missing brain weights as the average wt_ratio for that animal's order times the animal's weight. Make a second copy of your dataframe, but this time use group_by() and mutate() to impute missing brain weights with the average brain weight for that animal's order. What assumptions do

these data filling methods make? Which is a better way to impute the data, or do you see a better way, and why? You may impute or remove other variables as you find appropriate. Explain your decisions.

g) Generate a complete linkage clustering of the msleep data using Euclidian distances. Generate a dendogram of your clustering. What does it tell you about the sleep data? Which animals are outliers? Describe how feature selection could change this clustering.
h) Cut the clustering so that there are 4-8 clusters, depending on your dendogram. Print a table that shows the number of animals from each order that appear in each cluster. How do the sleep patterns of rodents, primates and carnivores compare?
i) Perform PCA on the data and print a biplot of the result.
j) Perform PCA, using standard deviation scaling and print a biplot for the data. How does scaling affect the PCA results? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.