DAMI 2019

Homework Assignment 1

This assignment comes with an R file named "HW1_R_codes.R", where you can find the guidelines for the tasks to be performed as a part of this assignment. You can either do the implementation on your own coding style, or you can follow the code structure and suggestions provided in the file for your convenience. Read the instructions, comments, and hints in the R file carefully.

What to submit:

- ONLY: complete R code file containing all the codes you have used to implement the tasks of this assignment. Please include any additional comments or answers you want to write in text in the .R file as comments at the appropriate location.
- DO NOT SUBMIT: any plots, pdf, docx, or any other file.

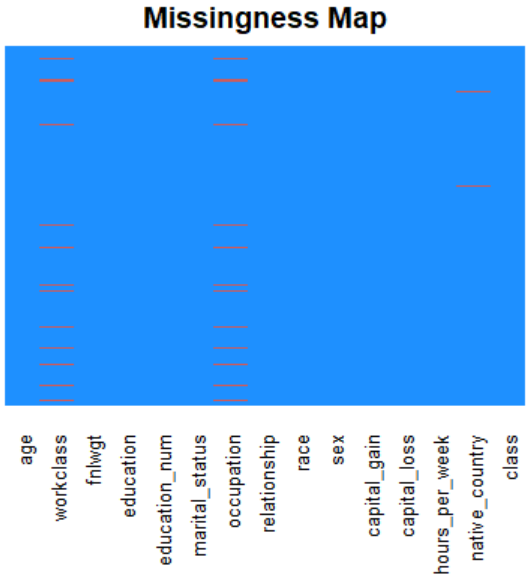
For this assignment you are going to use the "Adult" dataset. The dataset is publically available at the **UCI Machine Learning Repository**. This is an example of a real-world dataset for identifying and predicting whether an adult earns more than fifty thousand dollars annually.

1. Download the dataset from the following URL:
   http://archive.ics.uci.edu/ml/datasets/Adult

   After you follow the above link, go to "Data Folder", and download the following two files into your local working directory: "adult.data" and "adult.names".

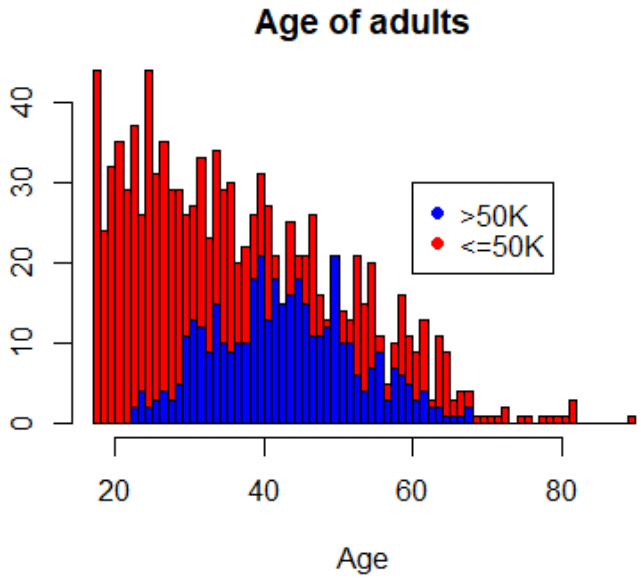
   Missing values in the file are represented as "?". While importing the dataset into the R environment, make sure that "?"s are recognized as missing values, i.e., NAs in R. [3 points]

2. Write some codes to plot missingness map from the data and find how many missing values each attribute has. Remove all the instances with missing values. [5 points]

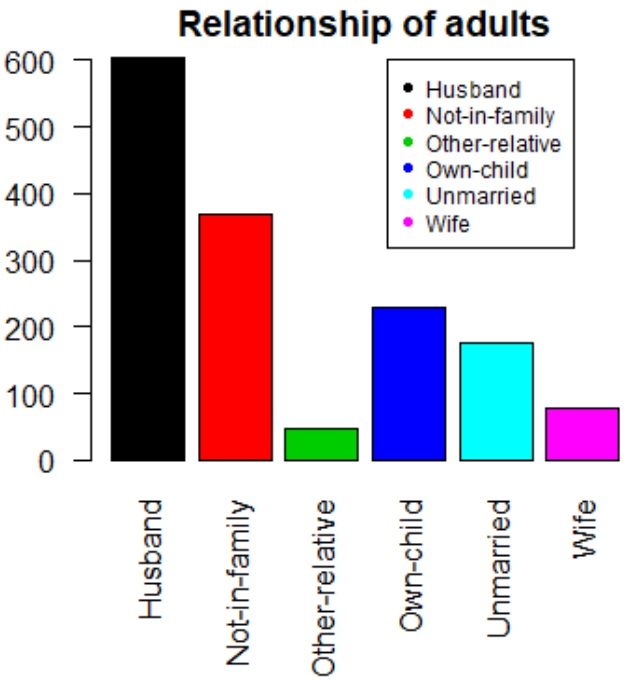


3. Select a small data sample of 1500 examples (rows) for this exercise. You should use the same random seed as in the accompanying R file "HW1_R_codes.R". We shall be using this smaller dataset from now onwards.

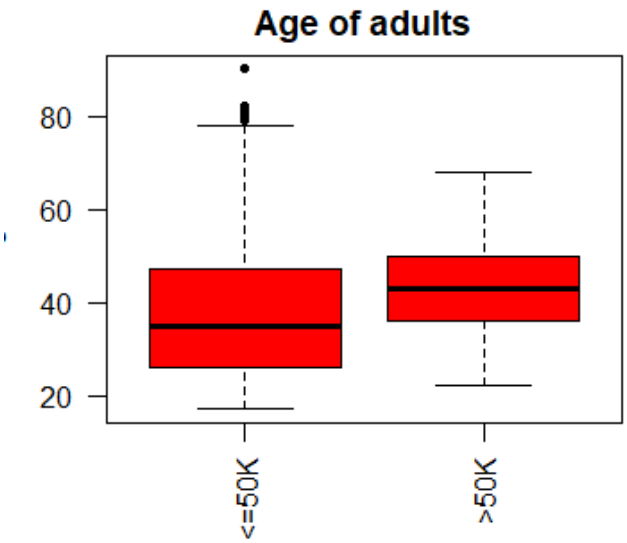
   a) Examine the attribute "age" by plotting a histogram, which should look like the one shown below. [3 points]



   b) Plot a bar chart for attribute "relationship", which should look like the one shown below. [3 points]

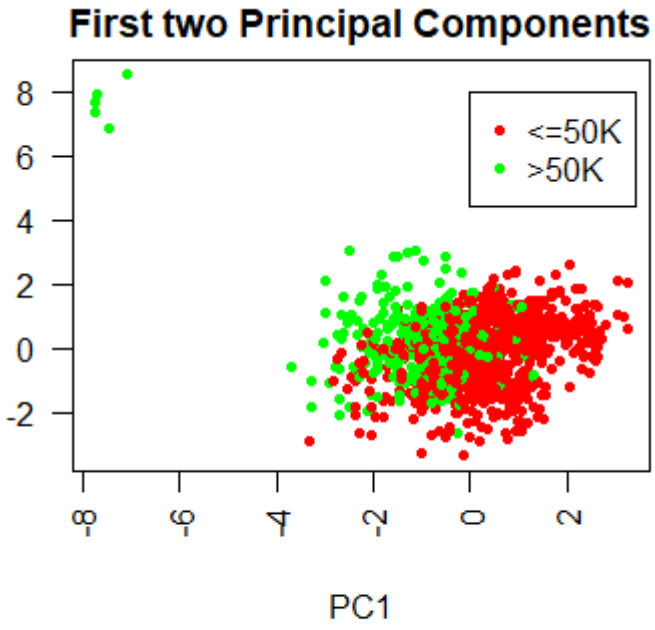


c) Plot a boxplot for “age” attribute for two group of adults who earn >50K and those who earn <=50K. [3 points]

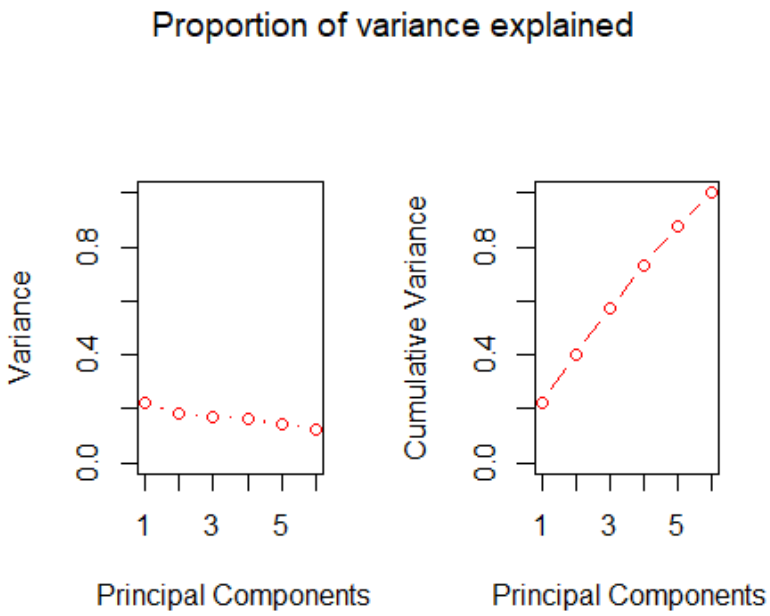


4. Consider only the numeric attributes from the dataset and **standardize** them, such that mean is zero and the standard deviation is one, for all of these numeric attributes. [3 points]

5. Consider **only the numeric attributes** from the previous task (**standardized**) and transform them into **principal components**. Note that you should have as many principal components as the selected attributes.
   a) Plot the first two principal components, which should look similar to the ones shown below. [3 points]



b) Calculate the **proportion of variance** and **cumulative variance** explained by each principal component (as shown in class), and plot it. Note that you should be considering six principal components. Your result should look similar to the figure below. [5 points]



c) Based on your results, how many principal components would you use to capture at least 50% of the total variance in the dataset? How many would you use to capture at least 90% of the variance? Please include your answer as a comment into the .R file you submit. [2 points]