EDA Exercise 9

P. McDonald

September 29, 2016

Navigate to the website

https://archive.ics.uci.edu/ml/datasets/Hepatitis

Read the annotation provided. Download the associated hepatitis.data and the hepatitis.names files.

Prepare an R Markdown file which documents the steps you take in carrying out the following

1. Read the data into memory as a csv file
2. Name the features as described in the names file
3. Write the result to memory as a csv file

Continuing your document, addressing the following questions

4. How many complete cases are there?
5. Subsetting the data on Age, Sex, Bilirubin, ALK, SGOT and Albumin, compute the number of missing values for the Bilirubin feature. Convert the last four features to numeric values. How many complete cases are there for the subsetted frame?
6. Are there any outliers in the Bilirubin and Albumin entries?
7. Bin the age variables in units of decades
8. Aggregate the data to obtain mean readings for the last 4 variables as a function of sex and age, with age as a binned factor.
9. Sort the data on the Bilirubin columns (ascending)
10. Standardize Bilirubin and Albumin and plot the outcome as a scatterplot.
11. Consider the data frame consisting of the complete cases for the variables Bilirubin, ALK, SGOT and Albumin. What fraction of the variance does the first principal component account for?
12. Subsetting the data on Age, Sex, Steroid and Antivirals columns and join the resulting data frame with the data frame of complete cases for Age, Sex, Bilirubin, ALK, SGOT and Albumin. What are the dimensions of the resulting frame?