Data Mining in R
Assignment 1

a) Short answer questions [18 points]

Q.1 - What is the goal of a histogram? **[2 points]**

Q.2 - What is the purpose of a Q-Q plot? **[2 points]**

Q.3 - What is the meaning of the limits of the rectangle in the middle of a box plot? **[2 points]**

Q.4 - What is a conditioned box plot? **[2 points]**

Q.5 - What are the pros and cons of the strategy of removing the observations with unknown values? **[2 points]**

Q.6 - When do you think the substitution of an unknown value on a variable by the variable mean is a bad idea? **[2 points]**

Q.7 - What is the meaning of the distance between two observations of a data set? **[2 points]**

Q.8 - What is the main advantage of the Mean Absolute Error over the Mean Squared Error? **[2 points]**

Q.9 - What is the basic idea of relative error metrics? **[2 points]**

b) Assignments [12 points]

i) The coefficient of variation is a statistic of the dispersion (or spread) of a continuous variable. You can check its definition on the Wikipedia:
http://en.wikipedia.org/wiki/Coefficient_of_variation

The goal of this exercise is to obtain a vector with the values of the coefficient of variation for all 7 algae of the Algae data set of Chapter 2. What can you say regards the dispersion of all algae? **[6 points]**

ii) The data set algae contains mostly numeric variables (with the exception of the first 3 columns). The goal of this exercise is to check how many outliers each of these variables has on the algae data set. A value should be considered and outlier if it is outside the interval [Q1-1.5*IQR, Q3+1.5*IQR], where IQR=Q3-Q1, and Q3 (Q1) is the third (first) quartile of the variable. Note that this is the definition used in the box plot graphs (c.f. page 47 on the text book), where outliers are show as dots according to the above rule.

Finally, if you want to complement your exercise, try to show the obtained information (number of outliers for each variable) as bar plot where each bar represents the number of outliers of the respective variable. In order to obtain bar plots in R you may want to check the help page of the function barplot().**[6 points]**

Hints:

i) There are many ways of obtaining what we want. However, one of the easiest is to use the function boxplot(). If you check the help page of this function, you will notice that the function can be used to return many information on the variable, among which there are the outlying values (check the parameters out an group in the help page).

ii) Moreover, you may want to check that it is also possible to apply the boxplot function to a data set and not only to a single column. This second hint may simplify even more the task of obtaining the number of outliers for all numeric variables…