Data Mining II
Homework 1

(1) (R programming + Data Processing) (- 20 points)
Consider the "College" dataset in the package "ISLR".

a) Use the function summary() to produce a numerical summary of the variables in the dataset.
b) Use pairs() to produce a scatterplot of the continuous variables in the data set.
c) Create a new qualitative variable called "Elite", by "binning" the variable "Top10perc". We are going to divide universities into two groups based on whether or not the proportion of students coming from the two 10% of their high school exceeds 50%. Add this variable to your dataset.
d) Use the table function to figure out how many Elite schools there are.
e) Use the table function to figure out how many of the Elite schools are private.
f) Do elite schools tend to have higher graduation rates?

(2) (R programming + Data Processing – 20 points)
This exercise uses the "Auto" dataset in the package "ISLR".

a) Remove missing values from the data.
b) What variables are numerical (continuous) or factors (categorical)?
c) Report the mean and standard deviation for each continuous variable in the data.
d) Remove the 5^{th} through 55^{th} observation. What is the range, mean and standard deviation?
e) In the full Auto dataset, are there any variables you would consider removing, or representing differently? Why?
f) In the full Auto dataset, graphically explore the relationships between the variables in the data set.
g) In the full Auto dataset, consider the variable mpg. You are going to create a new categorical variable for MPG, which has the categories: {low, med, high}. Call this variable "my_mpg", and create a new_Auto dataset, which contains all of the Auto variables, and your new variable "my_mpg". Save the dataset as an *.RData file and submit it with your assignment.