# Statistical Data Mining II
# Homework 1

**Directions:** Submit all source codes with write up. Please see UB Learns guidelines

1) Consider the "College" data in the ISLR2 package:

   ```
   > library(ISLR2)
   > data(College)
   > head(College)
   ```

   a) Present some visualizations of this data such as pair plots and histograms? Do you think any scaling or transformation is required?
   b) Scale the data appropriately (e.g., log transform) and present the visualizations in part A. Have any new relationships been revealied.
   c) Subset the data into two data frames: "private" and "public". Save them as an *.RData file. Be sure these are the only two objects saved in that file. Submit it with you assignement.

   ** For the remaining parts – use the "private and public" datasets. **

   d) Within each new data frame, sort the observations in decreasing order by number of applications recieved.
   e) Eliminate Universities that have less than the median number of HS students admitted from the top 25% of the class ("Top25perc").
   f) Create a new variable that categorizes graduation rate into "High", "Medium" and "Low", use a histogram or quantiles to determine how to create this variable. Append this variable to your "private" and "public" datasets.
   g) Create a "list structure" that contains your two datasets and save this to an *.RData file. Make sure that your file contains only the list structure.

2) You are going to derive generalized association rules to the marketing data from your book ESL. This data is in the available on UB learns. Specifically, generate a reference sample of the same size of the training set. This can be done in a couple of ways, e.g., (i) sample uniformly for each variable, or (ii) by randomly permuting the values within each variable independently. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability. Compare the results to the results near Table 14.1 (ESL), which were derived using PRIM.

3) Consider the Boston Housing Data in the ISLR2 package.

   a) Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.
   b) Visualize the data using the itemFrequencyPlot in the "arules" package. Apply the apriori algorithm (Do not forget to specify parameters in your write up).
   c) A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by "dis"). What can you advise on this matter through the mining of association rules?
   d) A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through

the mining of association rules?

Extra Credit: Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?