National University of Singapore
School of Computing

SWS3018 Predictive Analytics
Lab 1

Learning Objectives

- Getting help in R
- Read data using R
- Manipulate data in R

Note: You will need R and RStudio for most of the tutorials. Make sure you install it on your current workstation (and install it again if you log onto another workstation in the lab)

1. In this exercise, we will be using the `College` (`College.csv`) data set which can be downloaded from IVLE. This data set contains a number of variables for 777 different universities and colleges in the US. The variables are:

   - **Private** : Public/private indicator
   - **Apps** : Number of applications received
   - **Accept** : Number of applicants accepted
   - **Enroll** : Number of new students enrolled
   - **Top10perc** : New students from top 10% of high school class
   - **Top25perc** : New students from top 25% of high school class
   - **F.Undergrad** : Number of full-time undergraduates
   - **P.Undergrad** : Number of part-time undergraduates
   - **Outstate** : Out-of-state tuition
   - **Room.Board** : Room and board costs
   - **Books** : Estimated book costs
   - **Personal** : Estimated personal spending
   - **PhD** : Percent of faculty with Ph.D.'s
   - **Terminal** : Percent of faculty with terminal degree
   - **S.F.Ratio** : Student/faculty ratio
   - **perc.alumni** : Percent of alumni who donate
   - **Expend** : Instructional expenditure per student
   - **Grad.Rate** : Graduation rate

   Note:
   At any point in the lab exercise, you can type `?func` in R or RStudio to see the help on the functions

   a) Download `College.csv` onto your desktop or any directory. Before reading the data into R, it can be viewed in Excel or a text editor.

   b) Open RStudio and change the working directory to the path where you have placed `College.csv`

c) Use the `read.csv()` function to read the data into a variable `college` in R
Note: did you ensure that you do not read the header row?

d) Use the `head()` function to read the first 3 rows of `college` dataframe.

e) `rownames()` is a function that allows you to retrieve and set the row names of a matrix-like object. What do you see when you execute:
`rownames(college)`

f) What is the R command for getting a vector of the university names?

g) Ideally we want to set the row names to its corresponding university names and remove the existing column of university names. What do you think are the commands to do this?

h) What is the total number of applications received? How about the average number of applications received by a university?

i) How many of the universities received more than 5000 applications? Show the list of universities.

j) What is the average graduation rate for universities that received more than 5000 applications? Is the average graduate rate higher compared with universities that did not receive more than 5000?

k) Which universities have the highest/lowest number of full time undergrads?