

Business Analytics 1 - Assignment 1 (15 points)

Submission Deadline: Sunday, the 4th of February, at midnight

This exercise relates to the **colleges** data set, which can be found in the file **colleges.csv**. It contains a number of variables for 777 different universities and colleges in the US. The variables are defined as follows:

- **Name** : University/College names
- **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

Before reading the data into R, it can be viewed in Excel or a text editor.

- 1) Please use the `read.csv()` function to read the data into R. Name the loaded data `colleges`. Make sure that you have the directory set to the correct location for the data.
- 2) Please produce a scatterplot matrix for these five variables (3 points):
`Private + Top10perc + Top25perc + Expend + Grad.Rate`
- 3) In particular, we would like to study the relationship between `Private` and `Expend`. Please first generate a separate scatterplot for them ($x = \text{Private}$, $y = \text{Expend}$). (3 points)
- 4) By investigating the scatterplot, how would you summarize the difference between the instructional expenditure per student made by public and private universities/colleges? (3 points)
- 5) Please identify the university/college that has made the maximum instructional expenditure per student on the scatterplot you have generated in Question 3), and also extract all the information we have on it. (3 points)
- 6) Has this university/college you have just identified also taken in the most students from top 10% and top 25% of high school class? (3 points)

Please first create an R script in RStudio and then do the following:

- Write down your answers in the R script as comments (*i.e.*, all textual explanation should be prefaced with `##`).
- For each question, write down the R code you have used (not as comments), and make sure that every piece of conclusion you have made can be verified by the code you have provided.

You can either finish the assignment individually or in groups of two persons. Please also include your name(s) and matriculation number(s) in your submission.