Excercise 1 - Introduction to Data Science

Ken Benoit and Slava Mikhaylov

Assignments for the course focus on practical aspects of the concepts covered in the lectures. Assignments are based on the material covered in James et al. You will start working on the assignment in the lab sessions after the lectures, but may need to finish them after class.

Please submit your assignments via Moodle by 7pm on the day of the class. We will subsequently open up solutions to the problem sets.

Exercise 1.0

- Data Camp R tutorials (https://www.datacamp.com/courses/free-introduction-to-r)
- Data Camp R Markdown tutorials (https://www.datacamp.com/courses/reporting-with-r-markdown). You can complete the free first chapter.

Exercise 1.1

This exercise relates to the `College` data set, which can be found in the file `College.csv` on the website for the main course textbook (James et al 2013) http://www-bcf.usc.edu/~gareth/ISL/data.html (http://www-bcf.usc.edu/~gareth/ISL/data.html). It 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

Before reading the data into R, it can be viewed in Excel or a text editor, if you find that convenient.

a. Use the `read.csv()` function to read the data into `R`. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data. You can load this in R directly from the website, using:

```
college <- read.csv("http://www-bcf.usc.edu/~gareth/ISL/College.csv")
```

b. Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don’t really want `R` to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
rownames(college) <- college[, 1]
View(college)
```

You should see that there is now a `row.names` column with the name of each university recorded. This means that `R` has given each row a name corresponding to the appropriate university. `R` will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
college <- college[, -1]
View(college)
```

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that `R` is giving to each row.

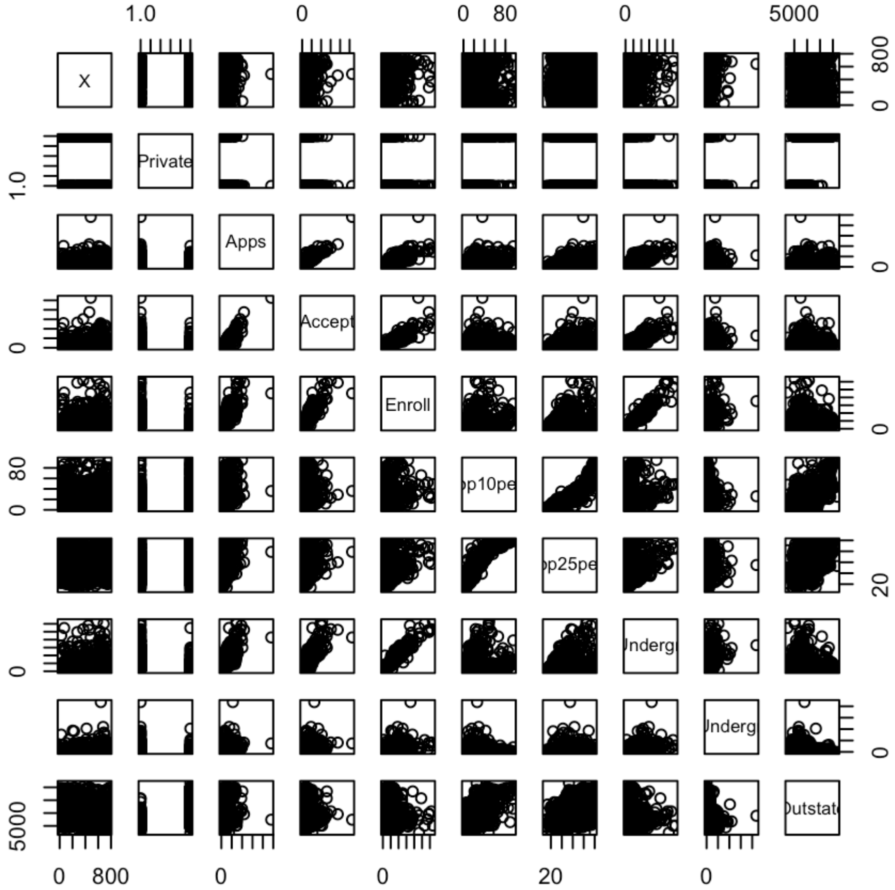
(c) i. Use the `summary()` function to produce a numerical summary of the variables in the data set.

```
summary(college)
```

```
##                              X       Private        Apps
##  Abilene Christian University:  1   No :212   Min.   :   81
##  Adelphi University          :  1   Yes:565   1st Qu.:  776
##  Adrian College              :  1             Median : 1558
##  Agnes Scott College         :  1             Mean   : 3002
##  Alaska Pacific University   :  1             3rd Qu.: 3624
##  Albertson College           :  1             Max.   :48094
##  (Other)                     :771
##      Accept          Enroll       Top10perc       Top25perc
##  Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
##
##   F.Undergrad     P.Undergrad         Outstate       Room.Board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##
##      Books           Personal         PhD            Terminal
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##
##    S.F.Ratio      perc.alumni        Expend        Grad.Rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
##
```

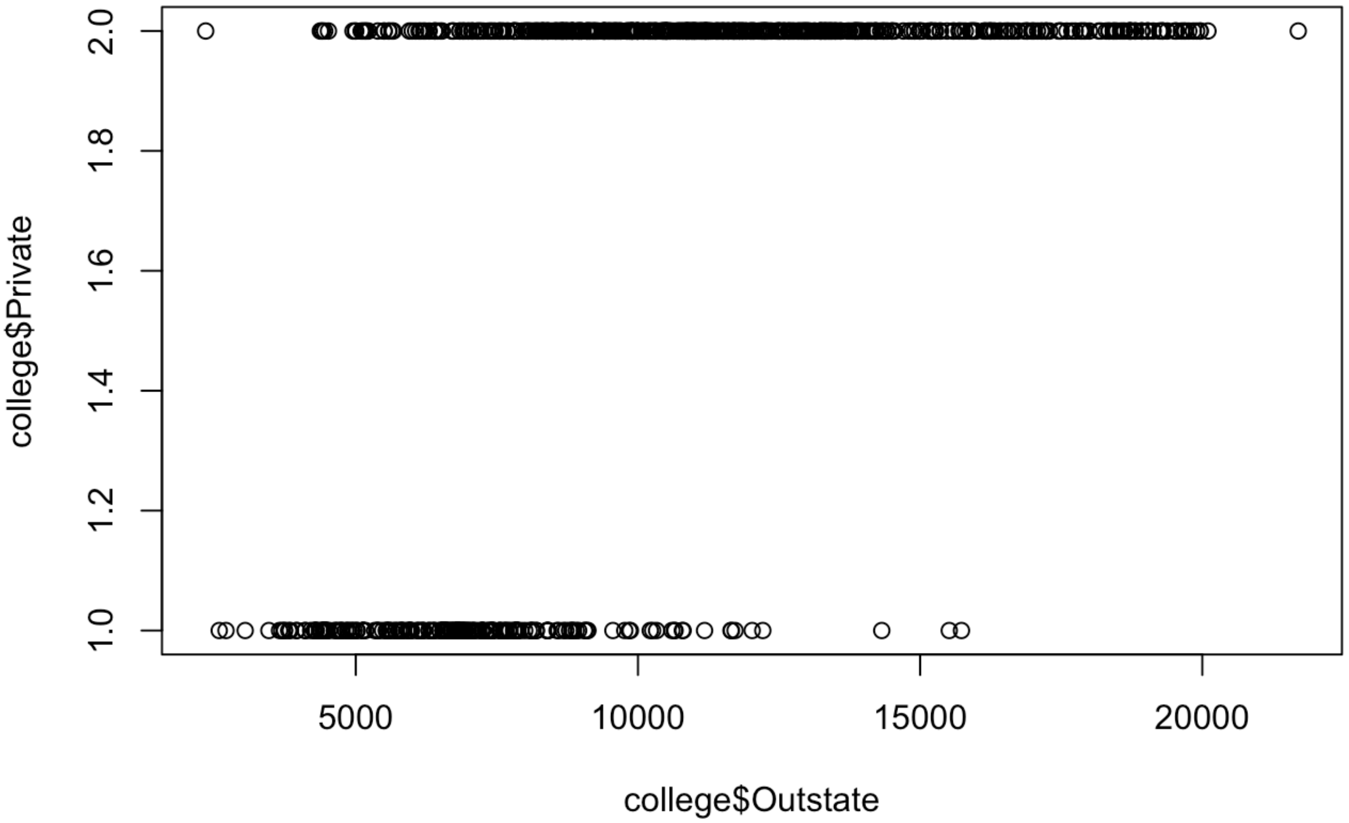
ii. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix `A` using `A[,1:10]`.

```
pairs(college[,1:10])
```



iii. Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.

```
plot(college$Outstate,college$Private)
```



iv. Create a new qualitative variable, called `Elite`, by *binning* the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

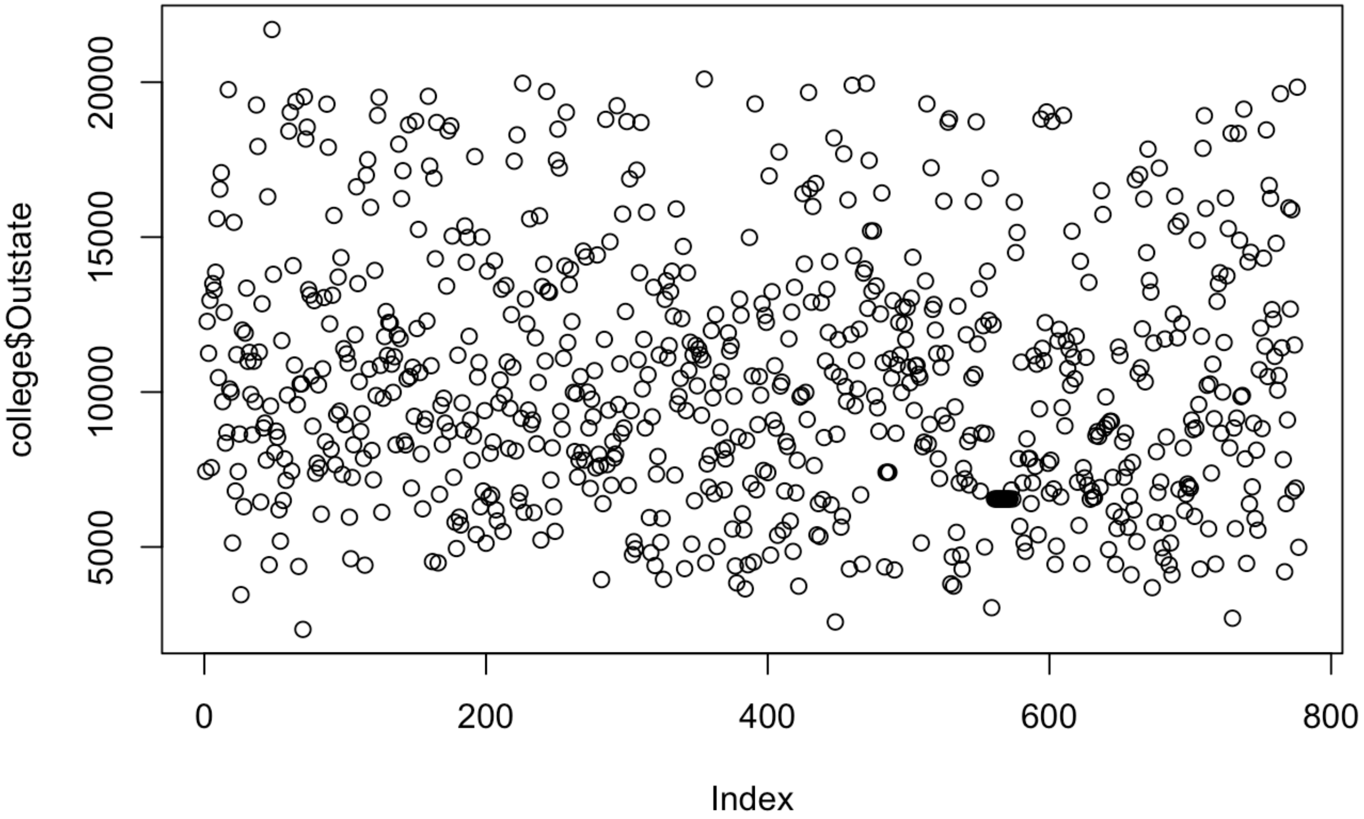
```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
```

Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

```
summary(college)
```

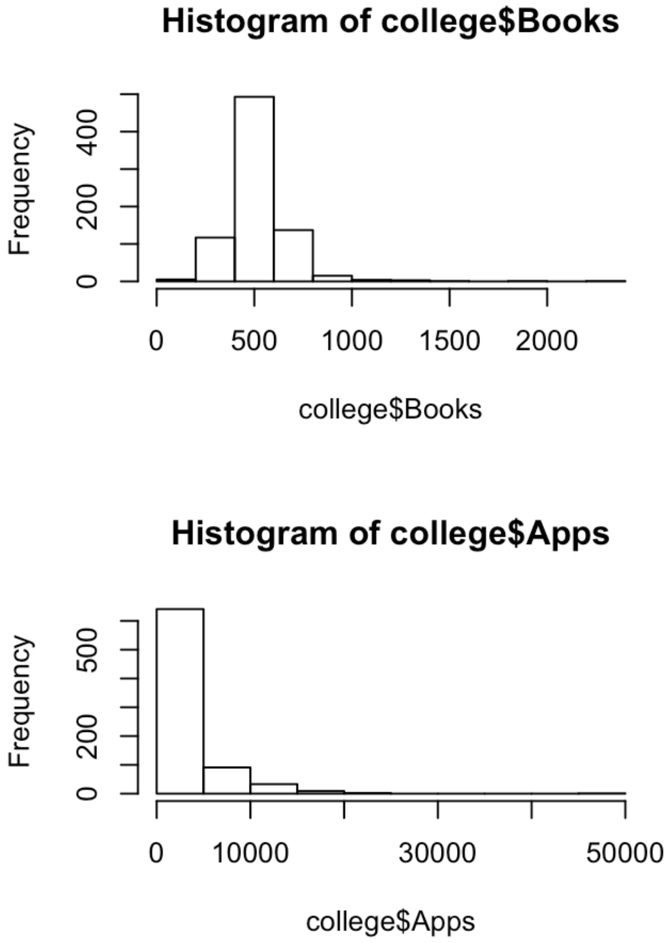
```
##                              X       Private        Apps
##  Abilene Christian University:  1   No :212   Min.   :   81
##  Adelphi University          :  1   Yes:565   1st Qu.:  776
##  Adrian College              :  1             Median : 1558
##  Agnes Scott College         :  1             Mean   : 3002
##  Alaska Pacific University   :  1             3rd Qu.: 3624
##  Albertson College           :  1             Max.   :48094
##  (Other)                     :771
##      Accept          Enroll       Top10perc       Top25perc
##  Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
##
##   F.Undergrad     P.Undergrad         Outstate       Room.Board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##
##      Books           Personal         PhD            Terminal
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##
##    S.F.Ratio      perc.alumni        Expend        Grad.Rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
##
```

```
plot(college$Outstate,college$Elite)
```



v. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow = c(2,2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```
par(mfrow = c(2,2))
hist(college$Books)
hist(college$Outstate)
hist(college$Apps)
hist(college$Enroll)
```



vi. Continue exploring the data, and provide a brief summary of what you discover.

Exercise 1.2

This exercise involves the `Auto` data set available as `Auto.csv` from the website for the main course textbook James et al. http://www-bcf.usc.edu/~gareth/ISL/data.html (http://www-bcf.usc.edu/~gareth/ISL/data.html). Make sure that the missing values have been removed from the data. You should load that dataset as the first step of the exercise.

a. Which of the predictors are quantitative, and which are qualitative?

b. What is the *range* of each quantitative predictor? You can answer this using the `range()` function.

c. What is the mean and standard deviation of each quantitative predictor?

d. Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

e. Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

f. Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.