Practical 1: Data manipulation in R

1 Getting started

Log-in to a PC in the usual way. Now open Rstudio – if you've not done this before, you can click on the "Start" button in the bottom left-hand-corner and search for "rstudio". Once you've opened Rstudio you should see four panels – if there is no text editor open in the top-left panel, click on the green button with a white plus symbol and open a new `R Script`. Remember – to execute a line of code in the text editor, hit `Ctrl-Enter`.

In the questions below, the important part is **understanding** what's going on rather than just typing the **R** commands. Make sure you answer the questions under each piece of **R** code. If you can't answer the question, then ask! Where possible, write your answer down on the worksheet. You should create the folder `Practical1` in a directory called `MAS1343` – this is where you can save your **R** session and any other files associated with this practical.

Work through *all* of the questions – the practice questions **and** the questions in the assignment!

2 Assessment during the practical

During this practical we will assess that you can correctly use Rstudio.

3 Practice questions

You should try these questions before attempting the assignment.

1. Why does the code in listing 1 produce an error message?

Listing 1: Simple error message

```
x = "1"
y = 2
x*y
```

2. The **seq** command:

Listing 2: The sequence command

```
x = seq(0, 10)
y = seq(0, 1001, by=0.5)
z = seq(1000, 0, by= -exp(1))
```

   a) For each line in listing 2, describe what **R** is doing.

   b) What is the length of each vector?

   c) Generate the vector: -1, -0.5,0.0, . . . , 10.

   d) Generate the vector: 100, 102, . . . , 206.

3. Vectors

Listing 3: Vectors

```
#You will find out more about the runif command in a few weeks.
x = runif(5000, 1, 8)
```

a) What is the length of `x`?

b) How many values in your vector `x` are below 2?

c) How many values in your vector `x` are above 7?

d) How many values in your vector `x` are below 3 or above 8?

e) How many times does the value 2 occur?

f) What are the $3400^{\text{th}} - 3402^{\text{th}}$ elements in your vector?

g) What is the smallest value in `x`?

h) What is the largest value in `x`?

i) What are the first and third quartiles of `x`?

j) Create a new vector called `y` which contains all values of `x` that are between 2 and 6.

k) **Tricky:** Which value in `x` is the closest to 5, i.e.

$$\min(|5 - x_i|)?$$

You can use the **abs** to work out absolute values.

4. Data frames.

```
1  app_ex1 = data.frame(c1=runif(50), c2=rnorm(50), c3=runif(50))
```

a) What does line 1 do?

b) How many rows in `app_ex1` are there where `c1` > 0.2?

c) How many rows in `app_ex1` are there where `c1` > 0.2 and `c2` > 0.2?

d) How many rows in `app_ex1` are there where `c1` > 0.2 and `c2` > 0.2 and `c3` > 0.2?

e) How many rows in `app_ex1` are there where `c1` > 0.2 or `c3` < 0.5?

Assignment

Preliminaries

Before starting, load the `mas1343` package:

```
> library(mas1343)
```

For questions 1 to 4, submit your answers using Blackboard.

- Log onto BB
- Go to MAS1343
- On the left hand menu click assignments
- Click "Practical 1 - Questions 1 to 4"

Warning: Only go to BB once you have answers for all four questions.

Question difficulty

Some of the questions are fairly straightforward, others are fiendishly difficult.

- **Straight forward:** Q1: $\{1-5\}$; Q2: $\{1,4\}$; Q3: $\{1-2\}$; Q4: $\{1-3,7,8\}$.
- **Doable but tricky**: Q1: $\{6,7\}$; Q2: $\{2,3\}$; Q4: $\{4,5,6\}$.
- **Fiendish:** Q2: 5; Q3: 3; Q4: 9.

Look at the commands in table 2.2 of your notes for ideas. The `table` command may also be useful.

Question 1

Run the following **R** code

```
> x1 = GetNumericVector(STUDENT_ID)
> #Where STUDENT_ID is your student id. For example
> #x1 = GetNumericVector("b1234567")
```

1. What is length of `x1`?
2. What is the 55^{th} element of `x1`?
3. What is the final element of `x1`?
4. What is the smallest value of `x1`?
5. How many values are greater than –11 but less than 2.2?
6. How many values fall in the region: $-17.86, 22.26$?
7. **Tricky:** What is the 4000^{th} smallest value in `x1`?

Question 2

Run the following R code

```
> x2 = GetLogicalVector(STUDENT_ID)
```

1. How many times does `TRUE` appear in `x2`?
2. What is the position in the vector of the first occurrence of `TRUE`?
3. What is the position in the vector of the last occurrence of `TRUE`?
4. What is the value of the 4000th element of `x2`?
5. **Tricky:** What position in the vector is the 2406th occurrence of `TRUE`?

Question 3

Run the following R code

```
> x3 = GetCharacterVector(STUDENT_ID)
```

1. How many times does "A" appear in `x3`?
2. Which letter appears the least? If more than one letter appears, just give the first letter (if the letters were sorted in alphabetical order).
3. **Very tricky:** How many pairs of letters are there in the `x3`? For example, in `AABCCC` we would have 3 pairs of letters.

Question 4

Run the following R code

```
> y = GetDataFrame(STUDENT_ID)
```

The data frame `y` is a subset of the movie data we use in the lectures.

1. How many rows does `y` have?
2. How many columns does `y` have?
3. How many movies are there where the budget is known, i.e. where `Budget` $!= -1$?
4. How many movies are there where the rating is less than 2.5 or greater than 7.5 (including 2.5 and 7.5)?
5. How many movies are there where the length is greater than 120 and have a rating of more than 7.5 (not including 120 and 7.5)?
6. How many movies were made in 1980 and have a rating above 5.0?
7. How many movies are classified as Action?
8. How many movies were classified as both Action and Animation?
9. **Very tricky:** How many movies were made in even years?