R ACTUAIRE DATA SCIENCE

ARTHUR CHARPENTIER

In this series of exercices, the files are in the exR.zip file. Extract them in a folder named

...\exRdatascientist\dataset

(1) Get the location of the working directory.

(2) Define ...\exRdatascientist\dataset as the working directory.

(3) Import the data1.csv dataset in a data frame object called `data1`.

(4) Import the data1.csv dataset, without the 6th line.

(5) Import the data1.csv dataset, without the 6th variable.

(6) Import the data1.csv dataset, without non-numeric variables.

(7) Sort the data frame data1 according to the value of variable `data1$var1`.

(8) What is the mean of the `data1$var2` variable ? (make sure that it is a numeric variable, first).

(9) What is the mean of variable `data1$var3` when `data1$var4=="a"` ?

(10) Print a table that contains all the means of `data1$var3` for all possible values of `data1$var4` ?

(11) Using

```
> library(xtable)
```

generate an html page that contains the previous table.

(12) Is the size of the `data1` object smaller or larger than 30 kilobytes ?

(13) Import the data2.csv dataset, in the data2.zip file, without unzipping it first.

(14) Import the dataset in the sheet Sheet2 of the data3.xlsx, by saving the file as data3.csv

(15) Import the dataset in the sheet Sheet2 of the data3.xlsx, using function `read.xls()` in `library(gdata)` ?

(16) Import the dataset in the sheet Sheet2 of the data3.xlsx folder, using function `read.xlsx()` in `library(xlsx)` ?

(17) Import the dataset in the sheet Sheet2 of the data3.xlsx folder, using function `readWorksheet()` in `library(XLConnect)` ?

(18) Import the dataset in the sheet Sheet2 of the data3.xlsx folder, using function `sqlQuery` in `library(RODBC)` ? [hint : use function `odbcConnectExcel` first]

(19) Use function `with()` to compute the mean of variable `var3` of data frame `data1` (without using the `$` symbol).

(20) Using function `tempfile()` and `download.file()`, import the data5.dat dataset from http ://freakonometrics.free.fr/data5.zip

(21) Import datasets data1.csv and data4.csv. Add a column named `counts` in `data1` that counts the number of appearences of variable `Id` in `data4`.

(22) `data1` is a data.frame object. Using function `as.data.table` in `library(data.table)`, create a data.table object, named `data5`.

(23) What doe the following commands returns,

```
> setkey(data5,var4)
> data5["a",]
> setkey(data5,var4,var7)
> data5[J("a",1),]
> data5["b",sum(var1)]
> data5[J("b",1),sum(var1)]
> data5["b",sum(var1),by=var7]
```

Reproduce those outputs using the data.frame object `data1`.

(24) What is the function `intersect` used for? What would the following code return?

```
> intersect(seq(4,28,by=7),seq(3,31,by=2))
```

(25) What would the following code return?

```
> c(TRUE,TRUE,FALSE,FALSE) & c(TRUE,FALSE,FALSE,TRUE)
```

(26) What would the following code return?

```
> n<<--1
> if (n==0) "yes" else "no"; n
> if (n < - 0) "yes" else "no"; n
> if (n<-0) "yes" else "no"; n
> if (n=0) "yes" else "no"; n
> if (n<-2) "yes" else "no"; n
```

(27) What will `1:10*1:5` return?

(28) Given a numeric vector `x`, write a function which returns only elements of `x` larger than `mean(x)`

(29) Write a function `seqrep(n)` which returns vector $(1, 2, 2, 3, 3, 3, \ldots, n)$ where integer k is repeated k times. How long is this vector when $n = 50$?

(30) Write a function that counts the number of `NA`'s in a vector.

(31) Create a function `secdiag(M)` which returns the second diagonal of squared matrix `M`.

(32) What will `M[,2]` return, if

```
> M <- matrix(1:5,3,3)
```

(33) Get the help page on command `%%`. What does that mean for `x` if `x %% 3 == 0` is `TRUE`?

(34) Given matrix

```
> m <- matrix(1:20,5,4)
```

what will the following line return

```
> which(m %% 3 == 0, arr.ind=TRUE)
```

(35) Write a function that computes the power of any square matrix, `power(M,n)`.

(36) Which function should you use to compute M^{-1} ?

(37) Create the identity matrix, of size 5×5.

(38) Using commands `!` and `%in%` return the subvector of

```
> x <- sample(1:15)
```

where values `c(3,7,12)` are removed.

(39) Given a matrix M, write a function which returns the following Kronecker product,

$$\begin{pmatrix} 1 & 3 & 4 \\ 2 & 0 & 5 \end{pmatrix} \otimes M = \begin{pmatrix} M & 3M & 4M \\ 2M & 0 & 5M \end{pmatrix}$$

(40) Compute $\sum_{i=10}^{20}(i^2 + 4/i)$.

(41) Solve numerically the following system

$$\begin{cases} 3x + 2y - z = 1 \\ 2x - 2y + 4z = -2 \\ -x + \frac{1}{2}y - z = 0 \end{cases}$$

(42) Given a vector $\boldsymbol{x}$ in $\mathbb{R}^n$ and a function $f : \mathbb{R} \to \mathbb{R}$, create a function `sum.function(x,f)` which computes $\sum_{i=1}^{n} i \cdot f(x_i)$.

(43) Compute $\sum_{i=1}^{10}\sum_{j=i}^{10} i^2/(5 + i * j)$.

(44) Create a function `mat(n)` which returns the $n \times n$ matrix, such that $M_{i,i} = 2$, $M_{i+1,i} = M_{i,i+1} = 1$ and 0 elsewhere.

(45) Are `sqrt(7)` and `7^.5` equal ? What about `sqrt(7)^2` and `7` ?

(46) Given

```
> a <- c(-0.2,0.2,0.49,0.5,0.51,.99,1.2)
```

what is the difference between `trunc(a)`, `floor(a)`, `ceiling(a)` and `round(a)` ?

(47) Given a vector `x`, use function `ifelse()` to generate a vector with the same length as `x` with the logarithm of elements of `x` that are positive, and `NA` when elements are negative.

(48) Create a function `anagram(word1,word2)` which returns `TRUE` if `word1` and `word2` are anagrams.

(49) Find roots of polynomial $x^2 + x = 1$.

(50) Given a vector `x`, write a function `which.closest(x,x0)` which returns the element in `x` which is the closest to `x0`.

(51) Given two vectors `x` and `y` write a function `subcount(y,x,k)` which returns the number of elements of `y` small than `x[k]`.

(52) Create vector of length 100 `'Ins1'`,`'Ins2'`, ..., `'Ins100'`.

(53) Create vector `c("London (2012)","Beijing (2008)","athens (2004)","Sydney (2000)")` from `c("London","Beijing","Athens","Sydney")`.

(54) What will the two following lines return ?

```
> paste("a", c("b c","d"), sep="")
> paste("a", c("b c","d"), collapse="")
```

(55) What will the following command return ?

```
> grep("ab",c("abc","b","a","ba","cab"))
```

(56) Given a vector `x`, find some functions can be used to return the location of the largest element ? and the location of the second largest ?

(57) Define

```
> Z <- ts(rnorm(240), start=c(1960,3), frequency=12)
```

Compute the sum of elements of time series `Z` related to `January`.

(58) Create a matrix with 4 columns, that contains all combination of 4 terms in `c(1,2,7,6,12,37,59)`, each row being a combination.

(59) Generate a vector

```
> set.seed(1)
> x <- rpois(9,4)
```

What does `unique(sort(x))[1:3]` will return ?

(60) Given a matrix `M` write a function `range.row` that returns a vector whose ith entry the the difference between the largest and the smallest value on the ith row of `M`.

(61) Given two numeric vectors `a` and `b`, write a function that returns the matrix $M = [m_{i,j}]$ with $m_{i,j} = a_i \cdot b_j$.

(62) Write a code to estimate the maximum likelihood estimator of a mixture of two Gaussian distributions for sample `sample.x`,

```
> set.seed(1)
> sample.x <- rnorm(427,mean=sample(c(-2,+1),size=427,
+ replace=TRUE))
```

(63) Create a function `which.min.mat(M)` which returns the the indices of the minimum of the matrix.

Contacts : charpentier.arthur@uqam.ca.