

## Practical Data Science (R Basics)

---

**Solve the following problems.**

### Problem1: Working with Numeric Vectors

Create the following vectors using R:

- (1, 2, 3, ..., 19, 20)
- (20, 19, ..., 2, 1)
- (1, 2, 3, ..., 19, 20, 19, ..., 2, 1)
- (4, 6, 3, 4, 6, 3, ..., 4, 6, 3) where there are 10 occurrences of 4.
- (4, 6, 3, 4, 6, 3, ..., 4, 6, 3, 4) where there are 11 occurrences of 4, 10 occurrences of 6 and 10 occurrences of 3.
- (4, 4, ..., 4, 6, 6, ..., 6, 3, 3, ..., 3) where there are 10 occurrences of 4, 20 occurrences of 6 and 30 occurrences of 3.
- Create a vector of the values of  $\exp(\cos(x))$  at  $x = 3, 3.1, 3.2, \dots, 6$

### Problem2: Working with Character Vectors

Use the function paste to create the following character vectors of length 30:

- ("label 1", "label 2", ..., "label 30") Note that there is a single space between label and the number following.
- ("fn1", "fn2", ..., "fn30") In this case, there is no space between fn and the number following.

### Problem3: Working with Random Vectors

Execute the following lines which create two vectors of random integers which are chosen with replacement from the integers 0, 1, ..., 999. Both vectors have length 250.

```
set.seed(50)
x_vec = sample(0:999, 250, replace=T)
y_vec = sample(0:999, 250, replace=T)
```

- Suppose  $x = (x_1, x_2, \dots, x_n)$  denotes the vector `x_vec` and  $y = (y_1, y_2, \dots, y_n)$  denotes the vector `y_vec`. Create the vector  $(y_2 - x_1, \dots, y_n - x_{n-1})$ .
- Pick out the values in `y_vec` which are  $> 600$ .
- What are the index positions in `y_vec` of the values which are  $> 600$ ?
- What are the values in `x_vec` which correspond to the values in `y_vec` which are  $> 600$ ? (By correspond, we mean at the same index positions)
- How many values in `y_vec` are within 200 of the maximum value of the terms in `y_vec`?
- How many numbers in `x_vec` are divisible by 2? (Note that the modulo operator is denoted `%`.)

## Practical Data Science (R Basics)

---

- g. Sort the numbers in the vector `x_vec` in the order of increasing values in `y_vec`.
- h. Pick out the elements in `y_vec` at index positions 1, 4, 7, 10, 13,....

### Problem4: Working with Data Frames

The data file `rainfall.dat` (available at algorithmica github repository) records hourly rainfall at a certain location in Canada, every day from 1960 to 1980. Answer the following questions:

- a. Load the data set into R and make it a dataframe called `rain.df`. What command did you use?
- b. How many rows and columns does `rain.df` have? How do you know? (If there are not 5070 rows and 27 columns, you did something wrong in the first part of the problem.)
- c. What command would you use to get the names of the columns of `rain.df`? What are those names?
- d. What command would you use to get the value at row 2, column 4? What is the value?
- e. What command would you use to display the whole second row? What is the content of that row?
- f. What does `names(rain.df) = c('year', 'month', 'day', 0:23)` do?
- g. Create a new column called `daily` which is the sum of the 24 hourly columns
- h. Make a histogram of the daily rainfall amounts.

### Problem5: A First attempt at Predictive Analytics Problems

Go through the following problems of kaggle:

<https://www.kaggle.com/c/sf-crime/data>

<https://www.kaggle.com/c/digit-recognizer/data>

Do the following tasks:

- a. Apply random predictions to each observation and find out how much accurate your predictions are by submission to kaggle?
- b. Find out the pattern in the data manually and then hard-code the logic. Resubmit your predictions and compare with random predictions?