Practical Data Science
(R Basics)

Solve the following problems.

Problem1: Working with Numeric Vectors

Create the following vectors using R:

a. (1, 2, 3,…..19, 20)

b. (20, 19,….. 2, 1)

c. (1, 2, 3,…..19, 20, 19,….. 2, 1)

d. (4, 6, 3, 4, 6, 3,……4, 6, 3) where there are 10 occurrences of 4.

e. (4, 6, 3, 4, 6, 3,……4, 6, 3, 4) where there are 11 occurrences of 4, 10 occurrences of 6 and 10 occurences of 3.

f. (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.

g. Create a vector of the values of ex cos(x) at x = 3,3.1, 3.2,…..,6

Problem2: Working with Character Vectors

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

a. ("label 1", "label 2", ..., "label 30") Note that there is a single space between label and the number following.

b. ("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)
```

a. Suppose x = (x_1, x_2, …., x_n) denotes the vector x_vec and y = (y_1,y_2,…..,y_n) denotes the vector y_vec. Create the vector (y_2 - x_1,….., y_n - x_{n-1}).

b. Pick out the values in y_vec which are > 600.

c. What are the index positions in y_vec of the values which are > 600?

d. 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)

e. How many values in y_vec are within 200 of the maximum value of the terms in y_vec? f. How many numbers in x_vec are divisible by 2? (Note that the modulo operator is denoted %%.)

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 kaggle problem at this link:
https://www.kaggle.com/c/otto-group-product-classification-challenge
Do the following tasks:

a. Apply random predictions to each test observation and find out how much accurate your predictions are by submitting to kaggle?
b. Perform EDA on train data and write the logic based on your observations. Predict the category for test observations based on your discovered logic and submit it to kaggle?
c. Try to infer the pattern or logic with more EDA and change your logic and check how much does it improves your prediction accuracy?