**R-Programming Exercises for Classroom Practice**

**# Dataset being used is "train.csv" unless explicitly specified**

1. How do we convert "4+" in "Stay\_in\_current\_city\_years" variable to a value of 4?
2. How many rows exist with a marital\_status = 0?
3. How many rows exist within the age\_group of "26-35" with a marital status =0?
4. How many distinct users exist within the age\_group of "26-35" with a marital status =0?
5. How many distinct age groups exist?
6. How many distinct User\_IDs exist?
7. Which product\_ID occurs the most frequently?
8. What is the average purchase rate of gender = f & gender = m?
9. What is the average value of purchase when gender = f or age\_group = "0-17"
10. What is the average value of purchase within the odd rows of train.csv?
11. create a new dataset (train2) that does not have any row in train.csv that has missing value
12. In which city\_category do most of the users within age group "0-17" live?
13. For how many rows is "Product\_category\_2" missing a value?
14. Which value of "Product\_category\_1" occurs the most whenever product\_category\_2 value is missing?
15. Of all the users that exist in "test.csv", how many of them, also exist in "train.csv"
16. Of all the users in "train.csv" how many of them also exist in "test.csv"
17. What is the average purchase of customers who exist in "train.csv" but not in "test.csv"
18. How many distinct combinations of "user\_id" & "Product\_id" are available in train.csv?
19. Among all the variables from "Gender" to "Product\_Category\_3":
    1. Calculate the average purchase rate for all the distinct values of each variable
    2. identify the variable that has the highest lift in purchase rate

(For example, if city has 3 distinct values (A,B,C) with average purchase rate of 100,200,300 & overall average of 200 - lift is (300/200 - 100/200)

i.e., (highest average value/ overall average purchase - lowest average value/ overall average purchase)

1. Write a function that takes variable name as input & gives out the frequency of occurrence table of the distinct values of the variable
2. Write a function that takes variable name & creates dummy variables:

For example, if we give age as input to the function, then the function creates 7 distinct columns named Age\_0-17, Age\_55+ so on.,

For each of the 7 columns, it gives a value of 1 to the row if the value belongs to the column - i.e., Age\_0-17 will be 1 only if Age = "0-17" else it is a 0.