# Exercise: Data Exploration, Preparation & Visualization

A dataset “expenses.csv” containing medical expenses of customers is given to you. Carry out the following steps to explore and prepare the data for data analysis.

1. Read the dataset into R.
2. How many observations and variables are there?
3. What are the different types of variables?
4. What values are contained in the variables?
5. For numeric data, generate some summary statistics.
6. Which variables are skewed?
7. Find the frequency distribution of gender in the dataset.
8. Find how many smokers and non-smokers are there in the dataset.
9. Find the frequency distribution of different regions in the dataset.
10. Cross tabulate the columns sex, smoker and region columns
11. Check for missing NA values in different columns of the dataset
12. Create a boxplot to check for outliers in age and bmi columns.
13. At which positions are the outliers lying?
14. Visualize the data - univariate analysis - draw a histogram for bmi.
15. Create a scatter plot for age and charges.
16. Do males generally have higher expenses than females?
17. Is there any relationship between region and medical expenses?
18. Do patients with less or no children have lower medical expenses?
19. Does smoking lead to higher medical expenses?
20. What is the correlation between bmi and medical expenses?
21. Impute the outliers in age variable with 99th percentile.
22. Impute missing values in bmi variable with mean.
23. Transform qualitative data into quantitative data. Create dummy variables for Gender.
24. Partition the data (70/30) into training and validation datasets.

\*\*\*\*\*\*\*\*\*