# ProgDS-2025-10-18 Exercise – 7: Exploratory Data Analysis (EDA) : Part-1

The data file **patient-data.csv** has been uploaded. Perform the following EDA and modelling tasks on this data using Excel (wherever feasible) and Python.

1. Carry out the following data understanding and integrity checks
   a. How many rows and columns are present in the data?
   b. How many rows and how many columns have missing data?
2. Identify the **level of measurement** associated with each column
3. Create appropriate Descriptive Statistics information for every column, based on it's level of measurement, and organize all this data in Tables for easy analysis and decision making.
   a. Analyze the descriptive statistics thus generated and make your initial conclusions
4. Is this data a **time-series** data? Do you expect any trends in each of the columns? Can you make any conclusions by creating a scatter plot of values of each column?
5. Based on the information available so far, what will be your strategy for dealing with the missing values? Make some initial decisions at this stage.
6. Analyze the **Ailment** column with the help of appropriate visualizations. What can you conclude? Are there any missing values in this column? How will you handle them? Why?
7. Create histograms and KDE plots (find out what KDE plots represent) for all the columns and analyze them. Any significant conclusions?
8. Create box plots for all the columns, individually, and analyze them. What are your conclusions?
9. Create box plots for all the numerical columns on a common scale, and in a single plot, and analyze this plot. What do you observe? What are its implications?
10. To understand the relationships between the columns, create the following:
    a. Pairwise scatter plots
    b. Heatmap of pairwise correlation coefficients
    c. What are your conclusions based on these plots?
11. Based on all the analysis so far, what is your final conclusion regarding the handling of **missing values**? Implement your decision!
12. Create a Logistic Regression model using this data. Create the train/test metrics. Hope you have not forgotten to split the data into train and test data!
13. Normalize the numerical columns and re-create the common-scale Box-plot. Interpret the results.
14. Create a Logistic Regression model using the normalized data. Create the train/test metrics and compare with the metrics of step '12'. What are your observations?

BTW, in steps 12 and 14 did you pass the **Ailments** data as is – that is, as text information? Did the Python functions work? Why?

oooOOOooo