# Exploratory Data Analysis (EDA)

CCDATSCL | Data Science

## *What is Exploratory Data Analysis (EDA)?*

- The analysis of datasets based on various numerical methods and graphical tools.
- Exploring data for patterns, trends, underlying structure, deviations from the trend, anomalies and strange structures.
- Techniques for visualizing and summarizing data.
- Created by statistician ***John Tukey***

- Seminal book is "Exploratory Data Analysis" by **Tukey**

## *Aims of EDA*

- ❖ Maximize insight into a dataset
- ❖ Uncover underlying structure
- ❖ Extract important variables
- ❖ Detect outliers and anomalies
- ❖ Test underlying assumptions
- ❖ Develop valid models
- ❖ Determine optimal factor settings (Xs)

## *Exploratory vs Confirmatory Data Analysis*

| Exploratory Data Analysis | Confirmatory Data Analysis |
|---|---|
| No hypothesis at first | Starts with hypothesis |
| Generate hypothesis | Test the null hypothesis |
| Uses graphical methods | Use statistical models |

## *Steps of EDA*

1. Generate good research questions
2. Data restructuring
3. Based on the research question, use appropriate graphical tools and obtain descriptive statistics. Try to understand the data structure, relationships, anomalies, unexpected behaviors.
4. Identify confounding variables, interaction relations and multicollinearity
5. Handle missing observations
6. Decide on the need of transformation
7. Decide on the hypothesis based on your research questions.

## *After Exploratory Data Analysis*

1. Do confirmatory data analysis. Verify the hypothesis by statistical analysis.
2. Get conclusions and present your results nicely.

## *Classification of EDA*

1. Exploratory data analysis is generally cross-classified in two ways.
2. First, each method is either **non-graphical** or **graphical**.
3. And second, each method is either **univariate** or **multivariate**.
4. **Non-graphical methods** generally involve calculation of summary statistics.
5. **Graphical methods** obviously summarize the data in a diagrammatic or pictorial way.
6. **Univariate methods** look at one variable (data column) at a time.
7. **Multivariate methods** look at two or more variables at a time to explore relationships.
8. Usually, multivariate EDA will be bivariate (looking at exactly two variables), but occasionally it will involve three or more variables.