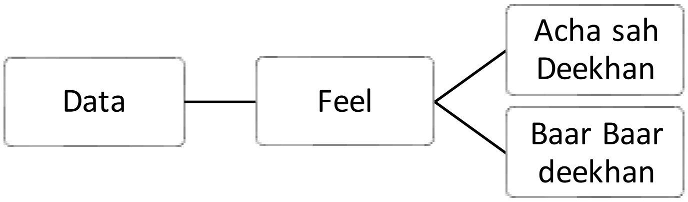
Exploratory Data Analysis (EDA)

In data science, EDA stands for **Exploratory Data Analysis**. It's like detective work of the field, where you **investigate and understand a dataset** before diving into deeper analysis or building models.

Think of it as:

1. **Uncovering:** hidden patterns and trends.
2. **Identifying:** unusual values or errors.
3. **Discovering:** relationships between variables.

It's crucial for building **strong foundations** for any data science project.



Things to keep in mind is:

1. Explore who gathered this data.
2. What is this data about.
3. Meta data of the data such as **sibsp, parch and survived (0,1) why the name of the column is this and why the entries in the data are like this**.
4. Dimension of the data using the **info () and shape ().**
5. See four things of the data:

    I. Composition of data. **(Comparing data in EDA involves analyzing similarities and differences between variables or groups within a dataset. This helps you understand the data better and form hypotheses for further exploration)**

    II. Correlation of data. **(relationship or the dependences of integer or float variable in the dataset with the other integer or float variables)**

    III. Comparison of data. **(It refers to the makeup of your dataset. Such as the data types, missing values, variables, and observations)**

    IV. Distribution of data. **(Check whether the variables in the data are normally distributed or not.)**

- The important point to be considered is the correlation ranges between -1 to +1, if it is higher than 0.5 than it means it is highly correlated.

Descriptive and Diagnostic Analysis in the EDA

Descriptive Analysis:

What it does: Summarizes the basic characteristics of your data through measures like:

1. **Central tendency:** Mean, median, mode.
2. **Spread:** Standard deviation, variance, range.
3. **Frequency:** Counts of unique values and their distributions.
4. **Visualization:** Histograms, boxplots, bar charts.
5. **Purpose:** Gain an initial understanding of the data, identify potential issues (missing values, outliers), and describe key data distributions.

Diagnostic Analysis:

What it does: Goes beyond describing what happened and digs deeper to understand why it happened. It involves:

1. **Univariate Analysis:** Examining individual variables for skewness, outliers, potential transformations.
2. **Bivariate Analysis:** Exploring relationships between pairs of variables using scatter plots, correlation coefficients.
3. **Group Comparisons:** Comparing distributions between different groups (e.g., income by gender) using statistical tests (t-tests, ANOVA).
4. **Purpose:** Uncover potential relationships, identify anomalies, and formulate hypotheses for further investigation.

- EDA is also called data exploration.
  - Data wrangling
  - Data munging
  - Data preprocessing
  - Data cleaning

- Data preprocessing is the combination of **data exploration, data wrangling and data mumming**.