**5. Data Understanding**

**What Happens in Data Understanding?**

After collecting the data, you **explore and assess** its quality before processing it.

**Steps in Data Understanding:**

1. **Initial Data Exploration**
   * Load the data into a tool (Python, Excel, SQL, etc.).
   * Check the **number of rows and columns** (to see if the dataset is complete).
   * Identify **data types** (numerical, categorical, text, dates, etc.).
   * Sample a few records to understand how the data looks.
2. **Checking Data Quality**
   * **Missing Values:** Find empty or null values.
   * **Duplicates:** Identify and remove repeated records.
   * **Inconsistencies:** Detect incorrect data (e.g., negative ages, dates in the future).
3. **Exploratory Data Analysis (EDA)**
   * **Descriptive statistics** (mean, median, standard deviation, etc.).
   * **Visualizations** (histograms, boxplots, scatter plots).
   * **Correlations** (see relationships between variables).

**Example:**

If you're analyzing customer purchase behavior:

* You might check the **average order value** and **most common purchase category**.
* You could create **bar charts** showing which products sell the most.

After understanding the data, the next step is to prepare it for modeling.

**6. Data Preparation**

**What Happens in Data Preparation?**

This phase involves **cleaning, transforming, and structuring** the data for analysis.

**Steps in Data Preparation:**

1. **Handling Missing Data**
   * Remove rows with too many missing values.
   * Fill missing values using averages, medians, or predictions.
2. **Removing Outliers**
   * Identify extreme values using boxplots, z-scores, or percentiles.
   * Remove or adjust outliers if they are errors.
3. **Feature Engineering (Creating New Features)**
   * Convert categorical variables into numbers (e.g., one-hot encoding).
   * Create new variables from existing ones (e.g., extract year from a date column).
   * Normalize/scale numerical values for better performance in models.
4. **Splitting Data (for Machine Learning Models)**
   * **Training set:** Used to train the model.
   * **Test set:** Used to evaluate the model’s performance.

**Example:**

If you're building a model to predict customer churn:

* You might replace missing income values with the **median income** of similar customers.
* Convert **"Subscription Type" (Basic, Premium, VIP)** into numbers (0, 1, 2).
* Normalize the **"Monthly Spend"** column so values are on a similar scale.