**Missing Values Treatment** (how to treat missing values?)

1. Remove : after removing missing values if you have >1000 records (then drop)

: percentage of missing values <5% (drop)

2. Replace :

- try to collect original data & replace it (top & best)

- replace statistically

- continuous variables has to be replaced with mean (if no outliers)

- continuous variables has to be replaced with median (if outliers)

- discrete variables has to be replaced with mode

**How to treat outliers**

(3R technique)

1. Remove

2. Replace

3. Retrain (keep the values as it is & do the work)

Remove: we should never do this

why?

- Sometimes, though we remove outliers, still some outliers will be present

- Sometimes, after removing outliers on given data,

if you check box plot on new data

still some outliers will be present

(Reason: Q1, Q3, IQR will be recalculated based on new data)

**2. Replace Outliers**

-- > try to collect original values & replace them (top & best)

-- > replace statistically (Winsorization)

any value >ul --- replace with ul (ul=Q3+1.5\*IQR)

any value <ll --- replace with ll (ll=Q1-1.5\*IQR)

**Steps involved in Data Analytics Project**

Data Analytics : Descriptive Analytics on a Structured Data

: Describe the given data (past data)

1. Business problem Understanding : Understanding Project Requirements

2. Extract data from Database using SQL

3. Data Understanding (Research & Expertise)

-Understanding the meaning of each column

-Data Exploration(EDA)

4. Data Cleaning

5. Data Analysis

- Apply various logics & making observations on the output

- all observations == analysis

6. Presentation

DATA CLEANING IS COMPLETELY DEPENDENT ON BUSSINESS REQUIREMENT