**Excercise-5**

Title: Missing data handling

**Data Preparation:**

(i) Create a dataset (excel/vector/frame etc) manually (numerical values against single feature) through some function and load in into R/Python (Given some name i.e MAIN\_DATASET). Consider size of dataset around 200 observations.

(ii) Create missing values randomly deleting 10% observations (Named it DIRTY\_DATASET\_1)

**Handle Missing values as:**

1. Make duplicate copy of DIRTY\_DATASET\_1 (named DB\_A) and handle it by deleting all observations having missing values

2. Make duplicate copy of DIRTY\_DATASET\_1 (named DB\_B) and handle it by replacing missing value with global constant.

3. Make duplicate copy of DIRTY\_DATASET\_1 (named DB\_C) and handle it by replacing missing value with mean.

4. Make duplicate copy of DIRTY\_DATASET\_1 (named DB\_D) and handle it by replacing missing value with median.

5. Make duplicate copy of DIRTY\_DATASET\_1 (named DB\_E) and handle it by replacing missing value mean of 10 samples of 20 size each.

6. Calculate the standard deviation of MAIN\_DATASET, DIRTY\_DATASET\_1, DB\_A, DB\_C, DB\_D and DB\_E.

7.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | MAIN\_DATASET | DIRTY\_DATASET\_1 | DB\_A | DB\_B | DB\_C | DB\_D | DB\_E |
| SD |  |  |  |  |  |  |  |

Observe the performance of each missing value imputation.