DEALING WITH MISSING VALUES

Name: Deepthi M

Batch number: 05

Serial number: 172

The dataset which has no outliers, no missing values are rarely available. If the dataset has outliers, missing values, they need to be treated.

Causes of missing values:

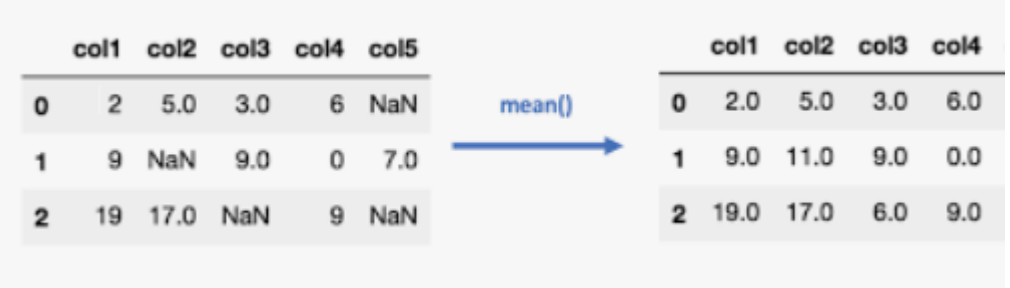
- Improper in data entry.
- Non-availability of data.

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX		Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX		Mid
6	CA	76600	High
7	NY	85000	High
8	CA		Entry
9	CT	45000	Entry

Missing values

Following ate the ways of handle missing values in the dataset:

1.Deleting the rows which has missing values
2.Impute missing values
3.Other Imputation Methods
4.Using Algorithms that support missing values
5.Prediction of missing values
6.Imputation using Deep Learning Library — Datawig



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

	col1	col2	col3	col4
0	2.0	5.0	3.0	6.0
1	9.0	11.0	9.0	0.0
2	19.0	17.0	6.0	9.0

- Deleting the rows which has missing values has the disadvantage and the disadvantage is we might tend to miss the data which might be important or which might play a vital role in the model.
- Impute missing values:
- There are three imputation ways on broadly:

Mean imputation.
Median imputation.
Mode imputation.

- Mean imputation and median imputation are used for numerical values are present in dataset.
- Mode imputation is used when categorical values are present in dataset.
- Mean imputation has the disadvantage when outliers are present in dataset. So, when data has the outliers also missing values then median imputation is used.
- Mode imputation is used when missing values are categorical

Disadvantage of having missing values:

- It reduces statistical power.
- It can cause bias in the estimation of parameters.
- It can reduce the representativeness of the samples.

Advantage of treating missing values:

- Model can perform with accurate results.
- Reduces the computational burden or complexity within the dataset.