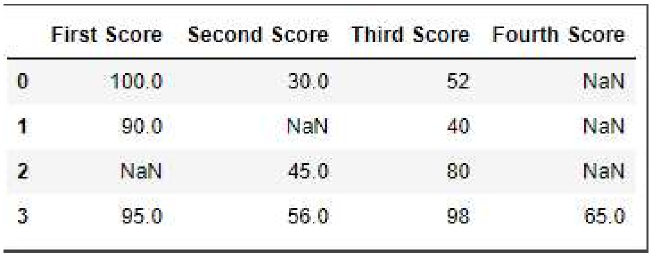
All Techniques for Handling Missing Values – Brief

06 July 2023 19:50

1. Removing Missing Values:

Missing values can be handled **by deleting the rows or columns** having null values.
If **columns** have **more than half of the values as null** then the entire column can be dropped. The **rows** which are having **one or more columns values as null** can also be dropped.



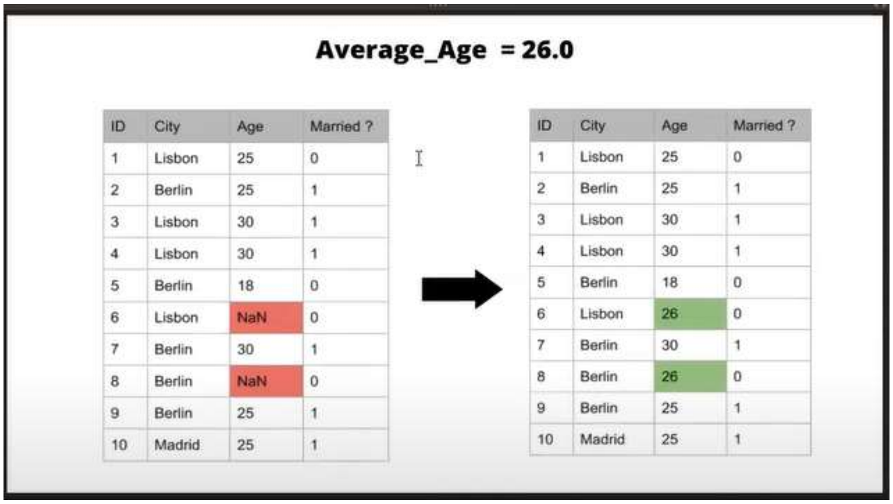
	First Score	Second Score	Third Score	Fourth Score
0	100.0	30.0	52	NaN
1	90.0	NaN	40	NaN
2	NaN	45.0	80	NaN
3	95.0	56.0	98	65.0

2. Imputing (Replacing) Missing Values:

a. Univariate Imputation:

This is a method of replacing missing values in a feature with a value that is estimated from the non-missing values in the same feature.
This estimated value depends on the type of feature.

a. **Numerical Feature:-** Mean, Median, Any Random Value, End of Distribution Value etc.
b. **Categorical Feature:-** Mode – most frequent value/category, "Missing" word.



Average_Age = 26.0

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	NaN	0
7	Berlin	30	1
8	Berlin	NaN	0
9	Berlin	25	1
10	Madrid	25	1

ID	City	Age	Married ?
1	Lisbon	25	0
2	Berlin	25	1
3	Lisbon	30	1
4	Lisbon	30	1
5	Berlin	18	0
6	Lisbon	26	0
7	Berlin	30	1
8	Berlin	26	0
9	Berlin	25	1
10	Madrid	25	1

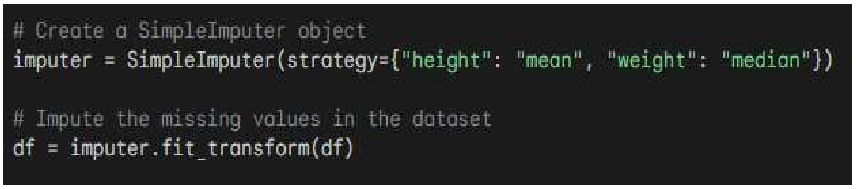
The **SimpleImputer Class** in **scikit-learn** is **a simple univariate imputation class** that can be used to replace missing values in a dataset.

It can be used to impute missing values with a variety of input strategies, including the mean, median, most frequent, and constant values.

```
from sklearn.impute import SimpleImputer
```

```
imputer = SimpleImputer(strategy="mean")

New_df = imputer.fit_transform(df)
```



```
# Create a SimpleImputer object
imputer = SimpleImputer(strategy={"height": "mean", "weight": "median"})

# Impute the missing values in the dataset
df = imputer.fit_transform(df)
```

The SimpleImputer class is **designed to work with numerical data**, but **can also handle categorical data** represented as strings.

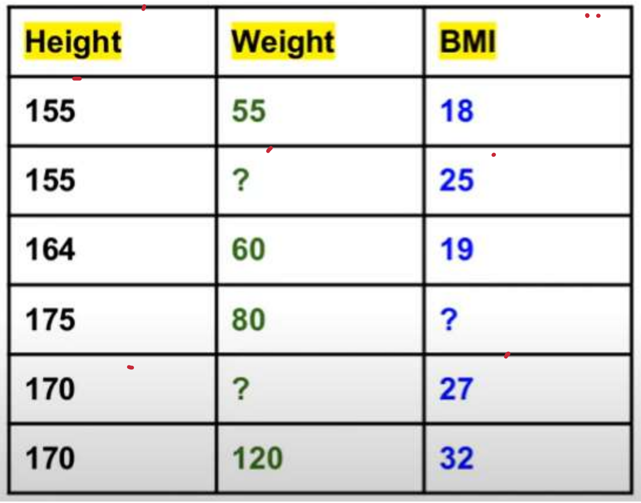
b. Multivariate Imputation:

This is a method of replacing missing values in a feature with values which are estimated on the basis of relationship among the different features of the dataset.

Multivariate Imputation is done using **kNN Imputer and Iterative Imputer Classes** available in scikit-learn.

1. **KnnImputer** – The **knnImputer Class** uses the k-Nearest Neighbours Algorithm to impute missing values. This means that it imputes each missing value with the value of same feature in most similar row/observation, as determined by the kNN algorithm.

2. **Iterative Imputer** – The **IterativeImputer Class** uses a more sophisticated imputation algorithm called Chained Multiple Imputation **(MICE)**. MICE works by iteratively imputing the missing values in a feature by predicting on the basis of imputed values in other features using regression. After imputing the values in current target feature, the next feature missing values will be imputed the same way using regression. We can set the number of iterations for iterative imputer object.



Height	Weight	BMI
155	55	18
155	?	25
164	60	19
175	80	?
170	?	27
170	120	32