Univariate Imputation- Arbitrary Value & End of Distribution Value

14 July 2023 19:31

Arbitrary Value Imputation:

Here, we impute the missing values by an arbitrary value which is not part of the dataset **OR** which is far away from the range of values in column.

Mostly we use values like **0, -1, 99.999, 99999999 or -9999999** and "**Missing**" or "**Not Defined**" for Numerical & Categorical features respectively.

This arbitrary value helps Model differentiate between Known and Missing Values.

- **When to use:-**
  - When Data is **Not Missing At Random**.
- **Advantages:-**
  - Easy to implement.
  - It retains the importance of "missing values" if it exists by making Model learn from missing values as well.
- **Disadvantages:-**
  - It generally **distorts original distribution** of the feature.
  - Arbitrary values can create **outliers**.
  - **Extra caution** required in selecting the Arbitrary value.

End-of-Distribution Value Imputation:

Also called **End-of-Tail Imputation**

This is kind of **similar to Arbitrary Value Imputation** with same assumptions, advantages and disadvantages.

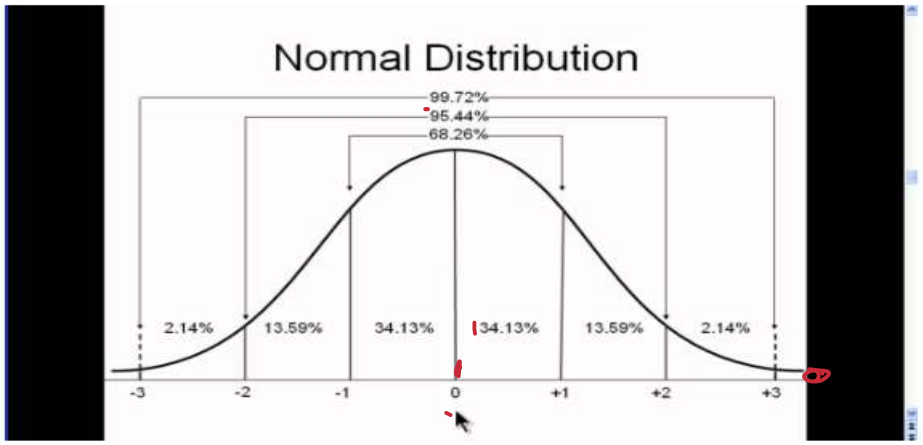
Here we impute the missing values with the **value at the end of the distribution of values in the features**. And this value is decided using a formula based on the type of distribution.

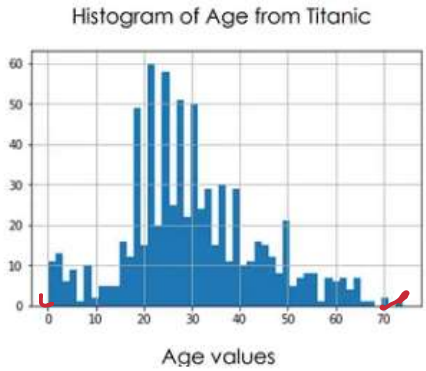
For Normal Distribution:

If the variable is **normally distributed**, we can use the **Mean plus or minus 3 times the Standard Deviation**.

Value for imputation = $\mu + 3\sigma$ (Upper) **or** $\mu - 3\sigma$ (Lower)

Perform End of Distribution imputation





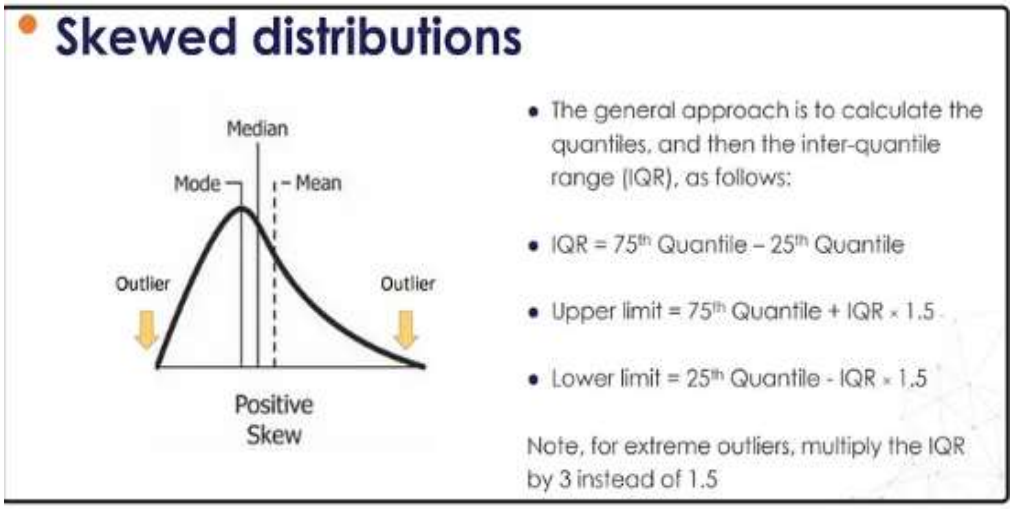
Mean(Age) + 3 × std(Age) = 72

Normal Distribution

For Skewed Distribution:

If the feature is **skewed**, we can use the **IQR Proximity Rule**.

Value = **Q1 – 1.5 IQR** (Lower) **or** **Q3 + 1.5 IQR** (Upper)



Code : Similar to Arbitrary Value