Train In Data

Mean or Median Imputation

Mean / Median imputation: definition

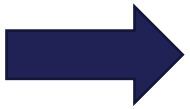
- Mean / median imputation consists of **replacing** all occurrences of **missing values** (NA) within a variable **with the mean or median.**

- Suitable numerical variables.

Mean imputation: example

Price
100
90
50
40
20
100
60
120
200

Mean = 86.66

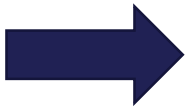


Price
100
90
50
40
20
100
86.66
60
120
86.66
200

Median imputation: example

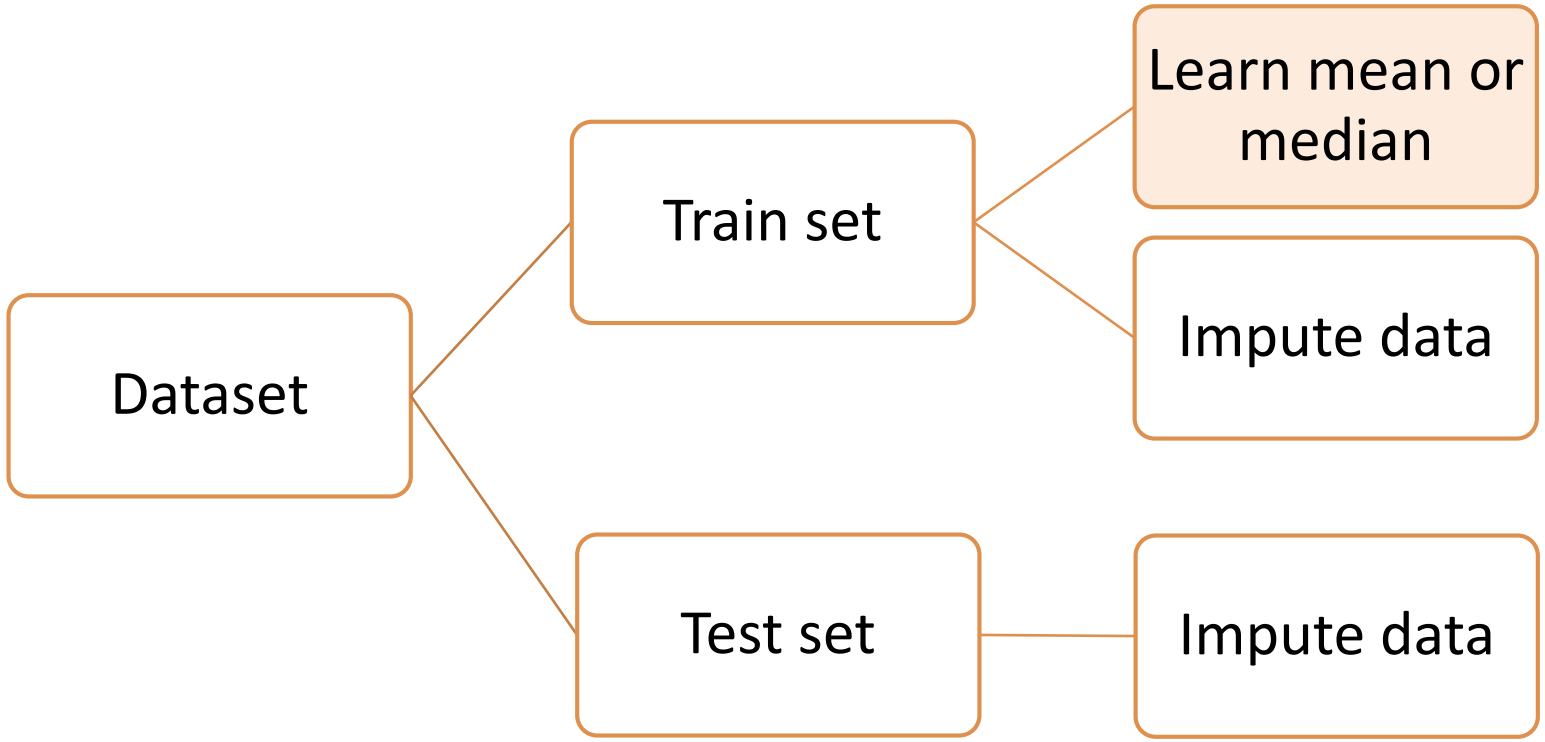
Price
100
90
50
40
20
100
60
120
200

Median = 90



Price
100
90
50
40
20
100
90
60
120
90
200

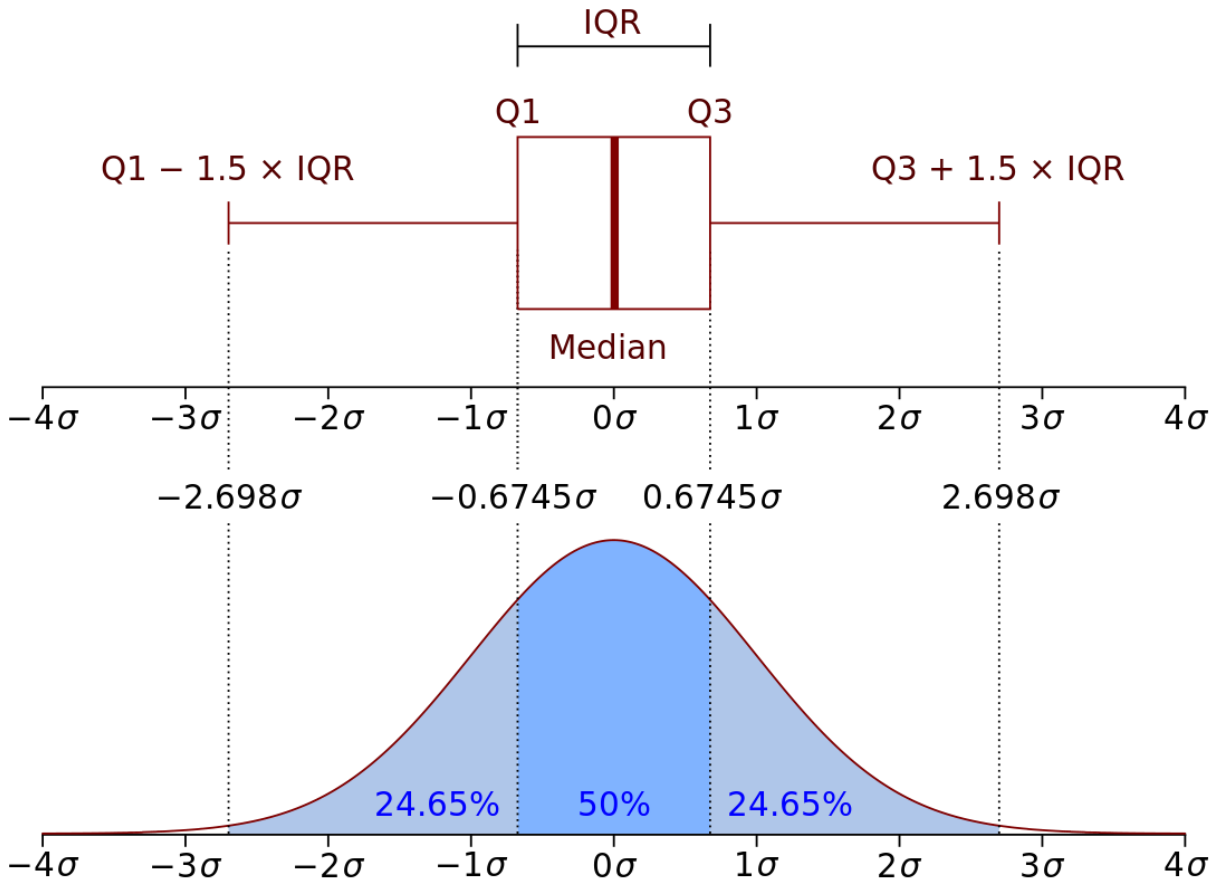
Correct workflow



The mean or the median are "learned parameters", like the coefficients of a linear model, or the splits of a tree.

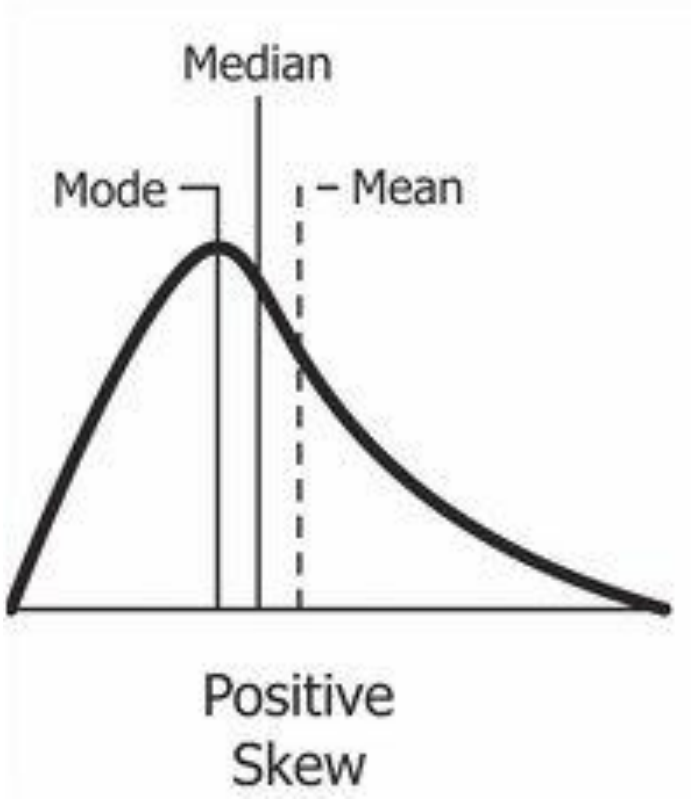
Mean or median?

Mean or Median imputation



- If the variable is normally distributed the mean and median are approximately the same

Mean or Median imputation



- If the variable is skewed, the median is a better representation

Assumptions

- Data is missing at random
- The missing observations, most likely look like most observations
  - The mean / median represent the majority
- Missing data are **blended** with the other values.

Good imputation strategy

Train In Data

THANK YOU

www.trainindata.com