Train In Data

Arbitrary Value Imputation

Arbitrary value imputation: definition

- Arbitrary value imputation consists of replacing all occurrences of missing values (NA) within a variable by an arbitrary value.

- Typically used arbitrary values are 0, 999, -999 (or other combinations of 9s) or -1 (if the distribution is positive).

- Suitable numerical and categorical variables
  - Categorical ➔ "Missing"

Arbitrary value imputation: example

Price
100
90
50
40
20
100
60
120
200

Arbitrary = 999

Price
100
90
50
40
20
100
999
60
120
999
200

Arbitrary value imputation: example

Price
100
90
50
40
20
100
60
120
200

Arbitrary = 99

Price
100
90
50
40
20
100
999
60
120
999
200

Arbitrary value imputation: Assumptions

Data is not missing at random.

If this is the case, we want to flag the missing values with a different (arbitrary) value, instead of replacing those occurrences with the mean or the median, which represent the most common value.

Mean / Median imputation: Advantages

- Easy to implement
- Fast way of obtaining complete datasets
- Can be integrated in production (during model deployment)
- Captures the importance of being "missing" if there is one

Mean / Median imputation: Limitations

- Distortion of the original variable distribution
- Distortion of the original variance
- Distortion of the covariance with the remaining variables of the dataset
- If the arbitrary value is at the end of the distribution it may mask or create outliers
- Need to be careful not to chose an arbitrary value too similar to the mean or median (or any other common value of the variable distribution)
- **The higher the percentage of NA, the higher the distortions**

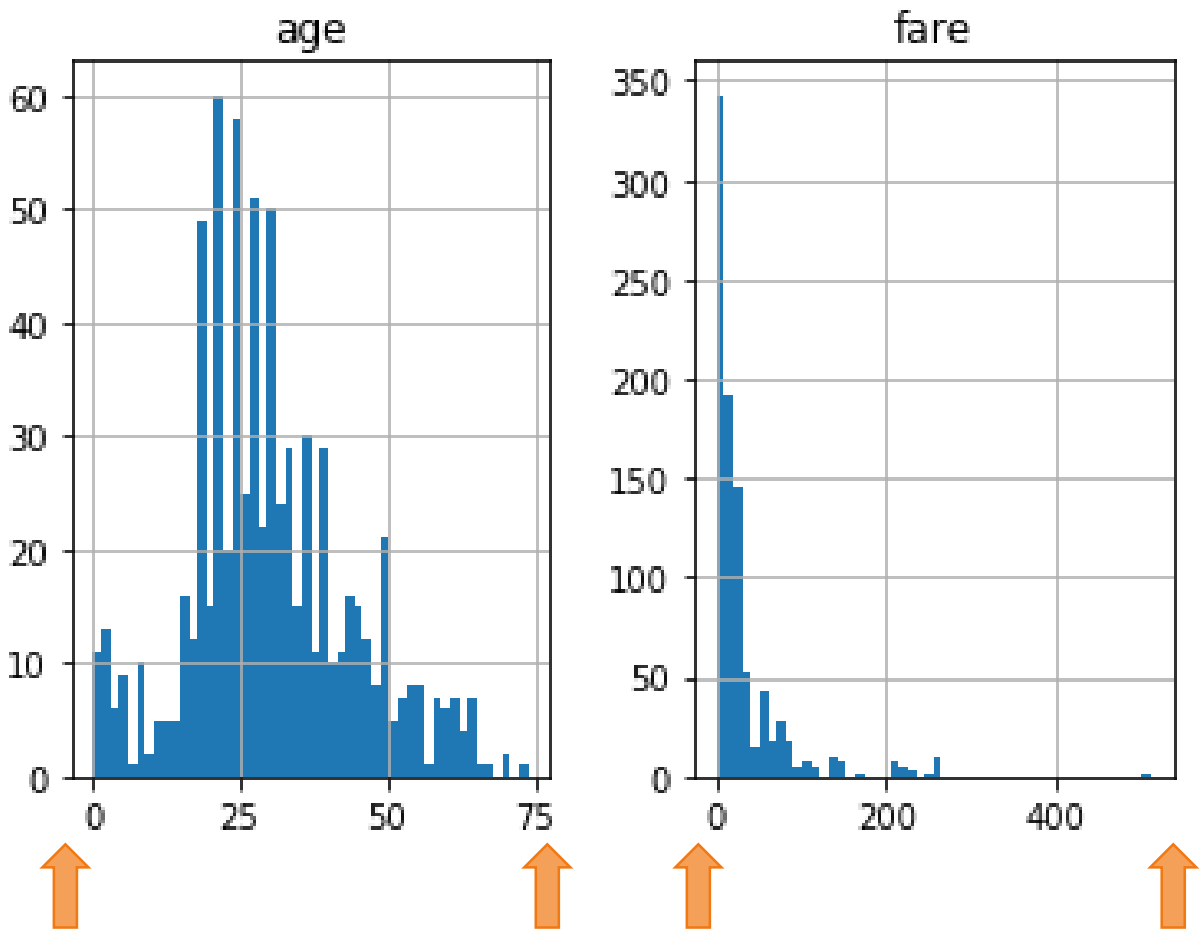
When to use arbitrary value imputation

- Data are not missing at random

Accompanying Jupyter Notebook

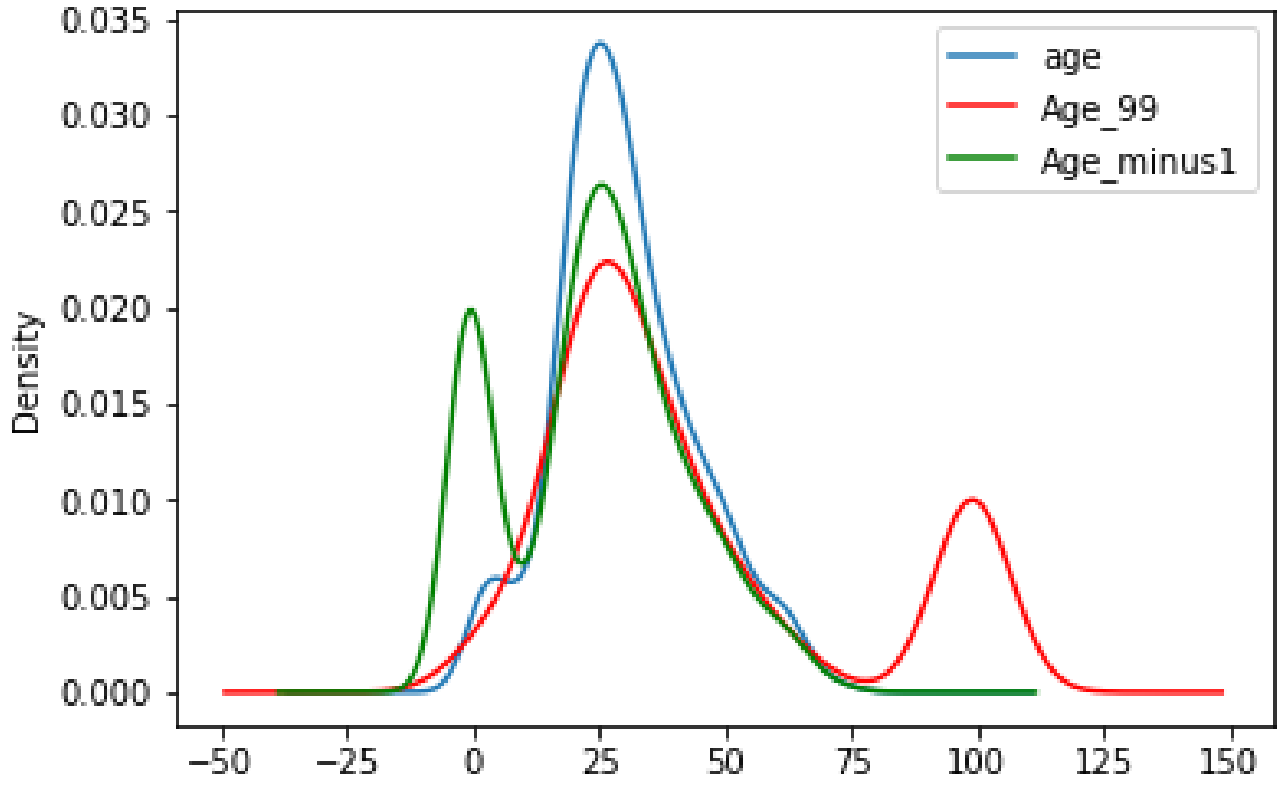
- Read the accompanying Jupyter Notebook
  - Arbitrary value imputation with pandas
  - Effect of the imputation on:
    - Variable distribution - variance
    - Interaction with other variables - covariance
    - Outliers

Which arbitrary value to use?



Let's compare the effect of using 99 or -1 for Age

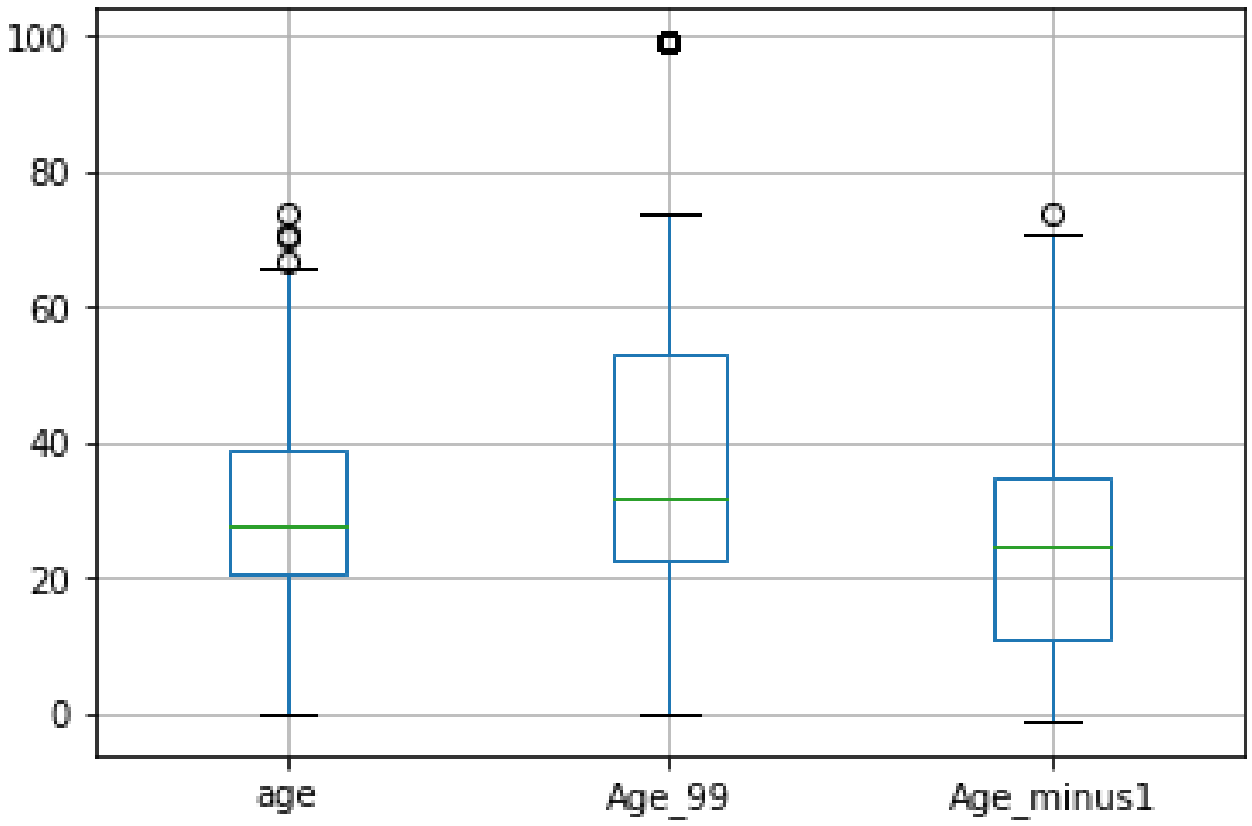
Arbitrary value imputation and distribution



- ~20% of data is missing in Age

Original variable variance: 194
Variance after 99 imputation: 888
Variance after -1 imputation: 307

Arbitrary value imputation and outliers



Masks outliers

Arbitrary value imputation: effects

	fare
fare	2248.326729
age	136.176223
Age_99	-38.722001
Age_minus1	177.733891

Train In Data

THANK YOU

www.trainindata.com