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 with an **arbitrary value**.

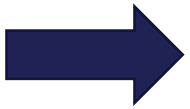
- Typically used arbitrary values are 0, 999, -999 (or other combinations of 9s) or -1 (if the distribution is positive).

- 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 |

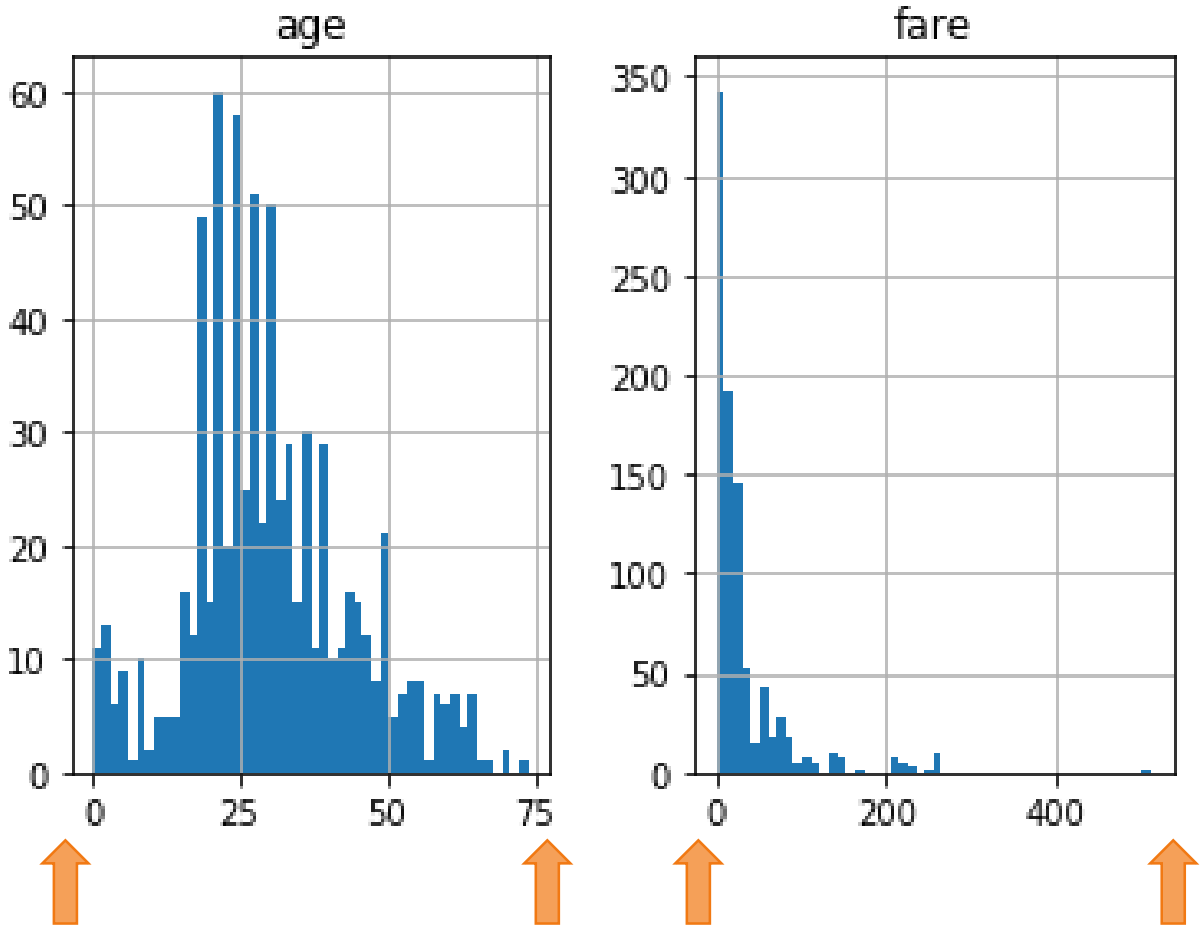
Arbitrary = 99

| Price |
|---|
| 100 |
| 90 |
| 50 |
| 40 |
| 20 |
| 100 |
| **999** |
| 60 |
| 120 |
| **999** |
| 200 |

We impute with a value that is different from most values in the distribution.

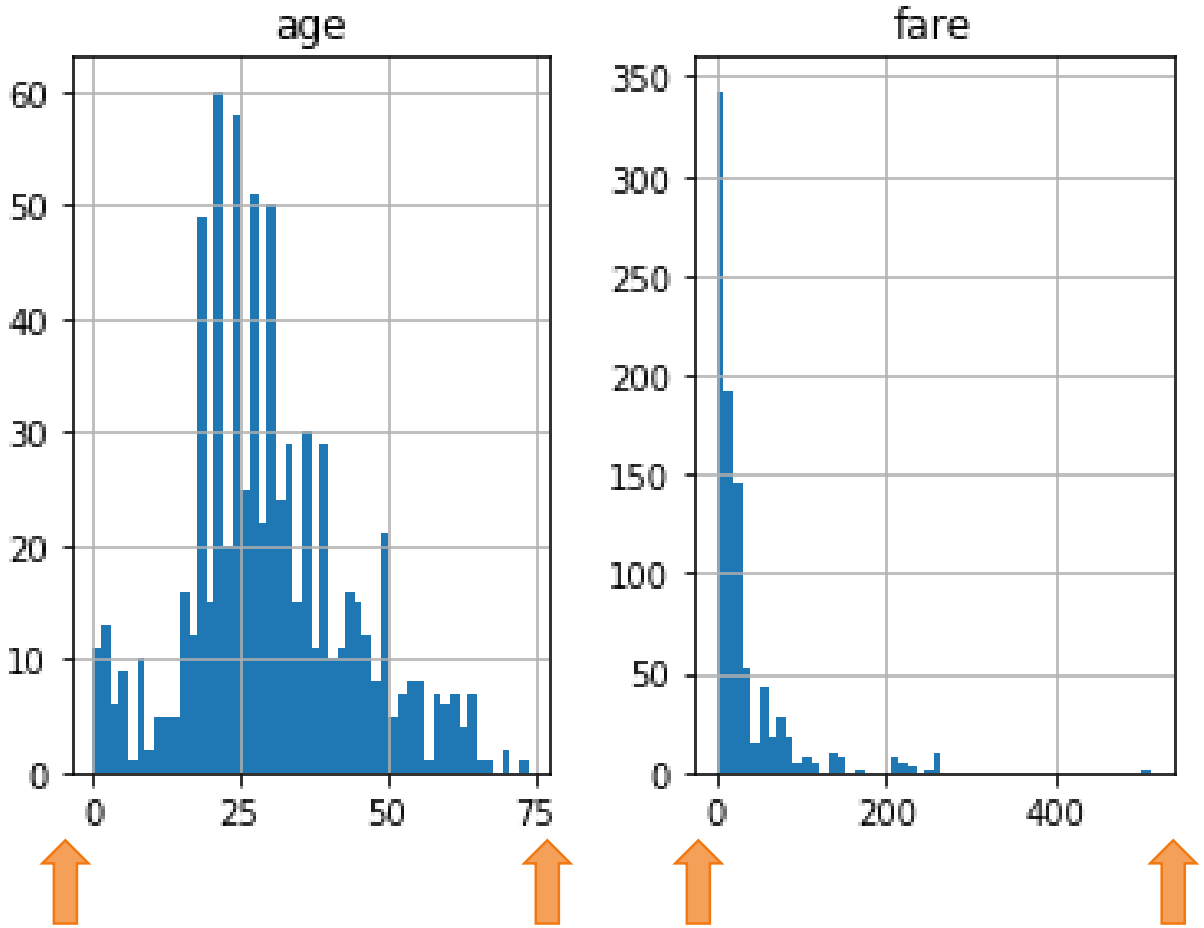
# Which value to use?

# Selecting the value




- Find the variable value range
- Pick a value outside that range.

# Selecting the value




- Find the variable value range.
- Pick a value outside that range.
- **Painful if we have too many variables.**

# Assumptions

Data is not missing at random.

We **flag** the missing values with a value that is different from most values in the distribution.

Train In Data

# THANK YOU

www.trainindata.com