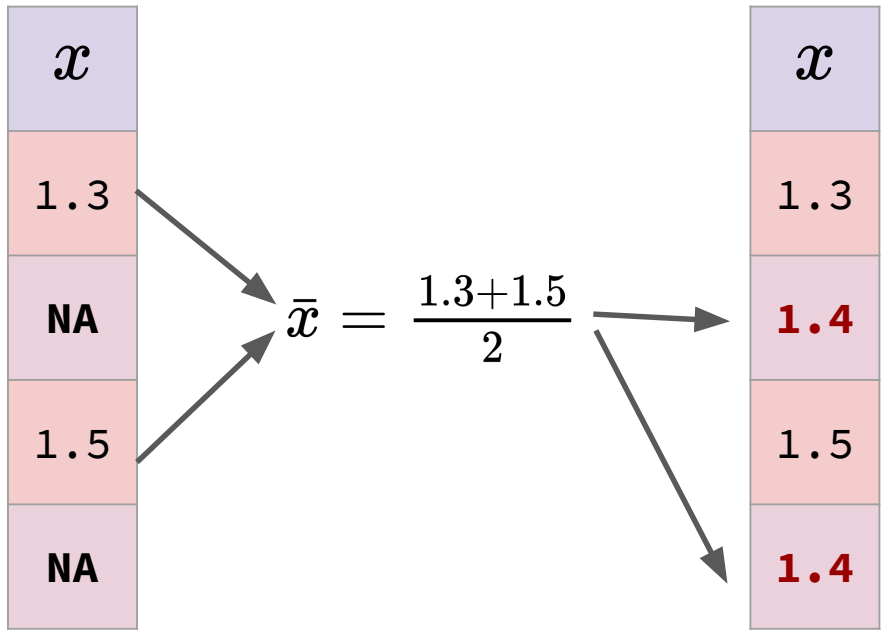
SIMPLE IMPUTATION METHODS

A very simple imputation strategy is to replace missing values with univariate statistics, e.g. mean or median, of the feature:



SIMPLE IMPUTATION METHODS

The statistic used to impute the missing values has to match the type of the feature:

- Numeric features: mean, median, quantiles, mode, ...
- Categorical features: mode, ...

Alternatively missing values can be encoded with new values

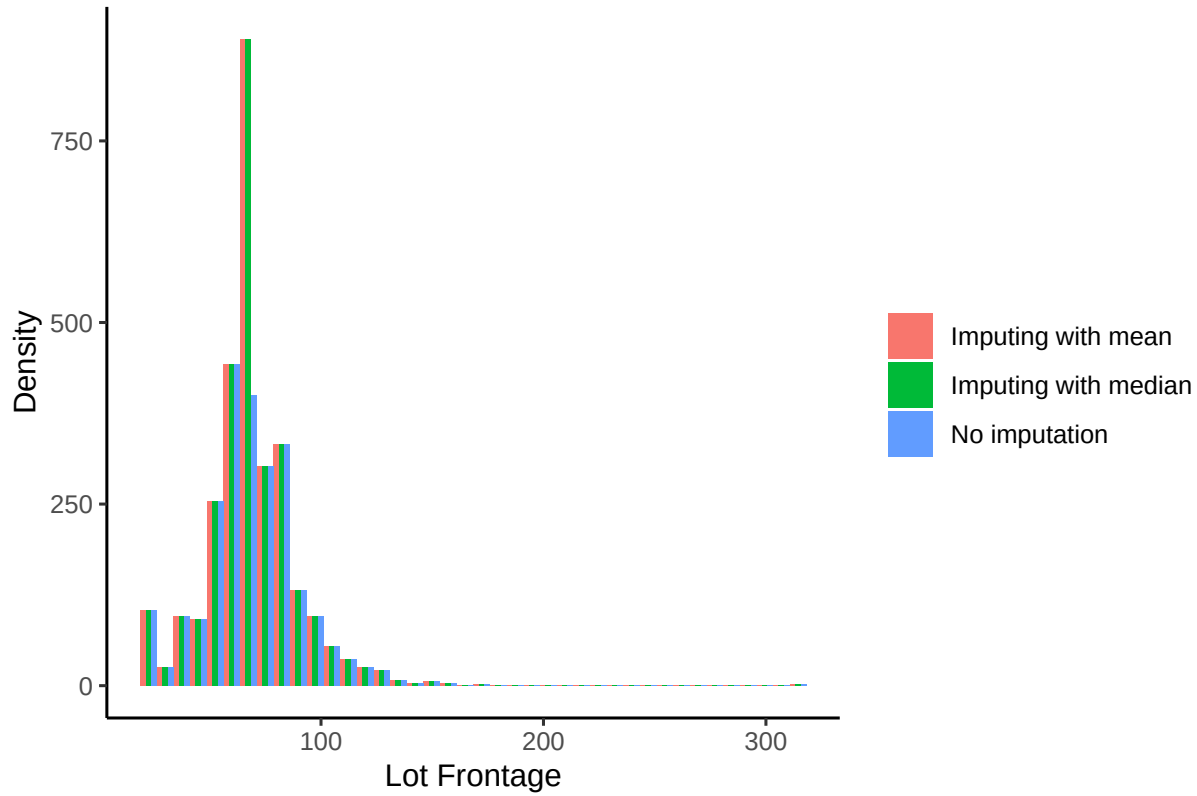
- Numeric features: `2*max`, ...
- Categorical features: `__MISS__`, ...

IMPUTATION - NOTES

- To ensure that the information regarding which values were imputed is not lost, we can add a binary indicator variable.
- Domain knowledge is highly important: Missing *Credit* can mean that the individual has 0 debt.
- Encoding numeric values with *out-of-range* values has been shown to work well in practice for complex ML models.
  - This is especially useful for tree-based methods, as it allows separating observations with missing values in a feature.
  - But using *out-of-range* imputation when estimating global effects (e.g. in linear models) can skew the results

DISADVANTAGE OF CONSTANT IMPUTATION

By imputing a feature with one value we shift the distribution of that feature towards a single value.



IMPUTATION BY SAMPLING

A way out of this problem is to sample values to replace each missing observation from

- the empirical distribution or histogram, for a numeric feature.
- the relative frequencies of levels, for a categorical feature.

This ensures that the distribution of the features does not change much.

BENCHMARK OF SIMPLE IMPUTATION

To illustrate the effect of imputation on the performance we evaluate a linear model on the Ames housing dataset. Evaluation is done with a 10-fold cross-validation: