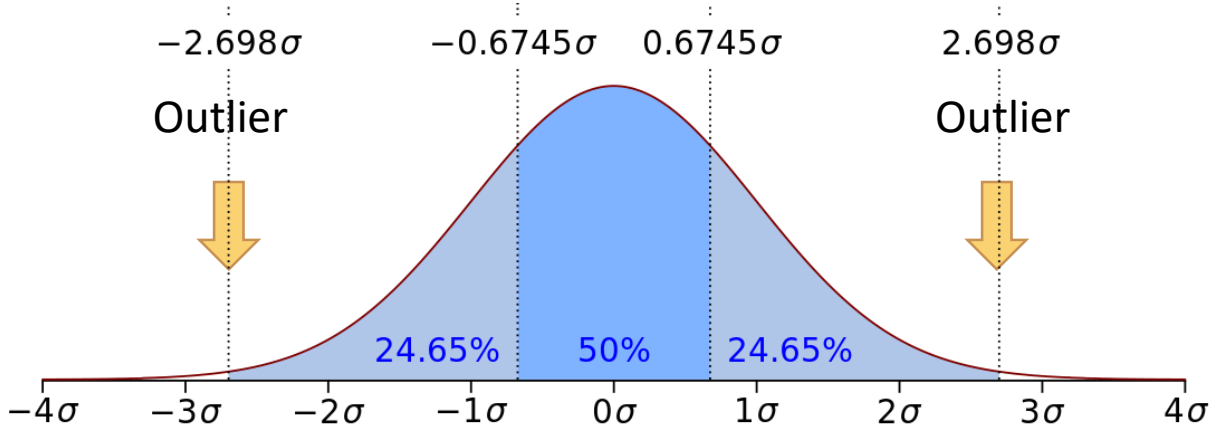
Train In Data

End of Tail Imputation

End of tail imputation: definition

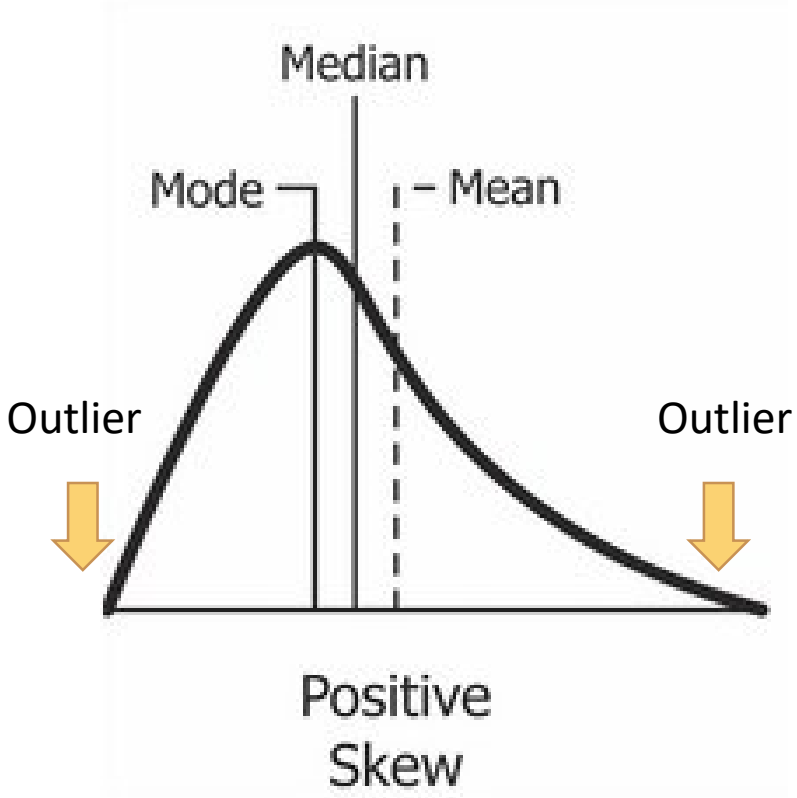
- End of tail imputation is equivalent to arbitrary value imputation, but automatically selecting arbitrary values at the end of the variable distributions.
- If the variable is normally distributed, we can use the mean plus or minus 3 times the standard deviation
- If the variable is skewed, we can use the IQR proximity rule
- Suitable numerical variables

Normal distribution



- ~99% of the observations of a normally distributed variable lie within the mean ± 3 × standard deviations.
- Values outside mean ± 3 × standard deviations are considered outliers

Skewed distributions



- The general approach is to calculate the quantiles, and then the inter-quantile range (IQR), as follows:
- IQR = 75th Quantile – 25th Quantile
- Upper limit = 75th Quantile + IQR × 1.5
- Lower limit = 25th Quantile - IQR × 1.5

Note, for extreme outliers, multiply the IQR by 3 instead of 1.5

Accompanying Jupyter Notebook

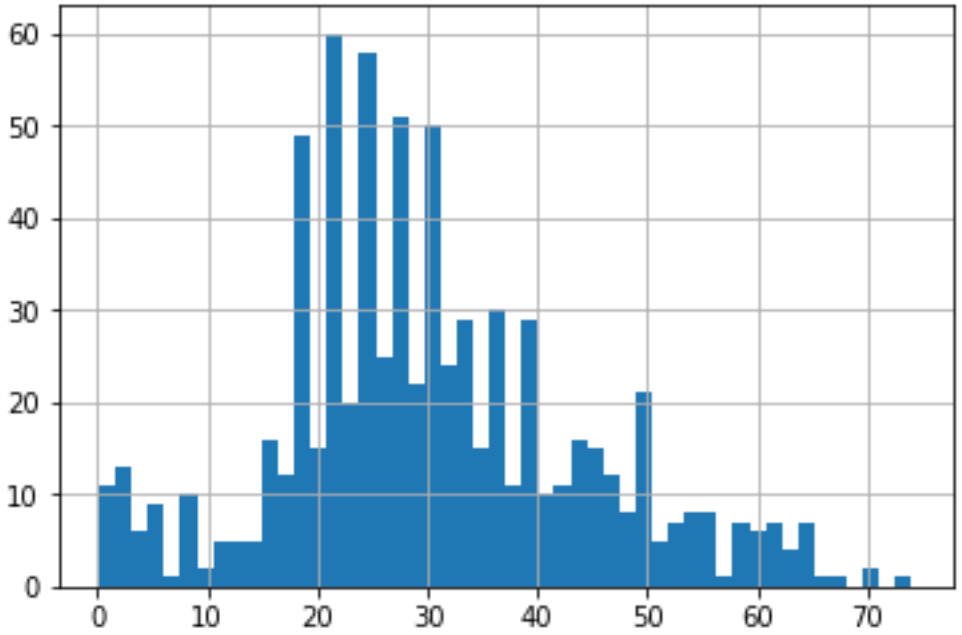
- Read the accompanying Jupyter Notebook
  - End of tail imputation with pandas
  - Effect of the imputation on:
    - Variable distribution - variance
    - Interaction with other variables - covariance
    - Outliers

End of tail imputation: how to do it

- The values to replace missing data should be calculated only on the train set
- We need to divide the data set into train and test before doing the imputation techniques

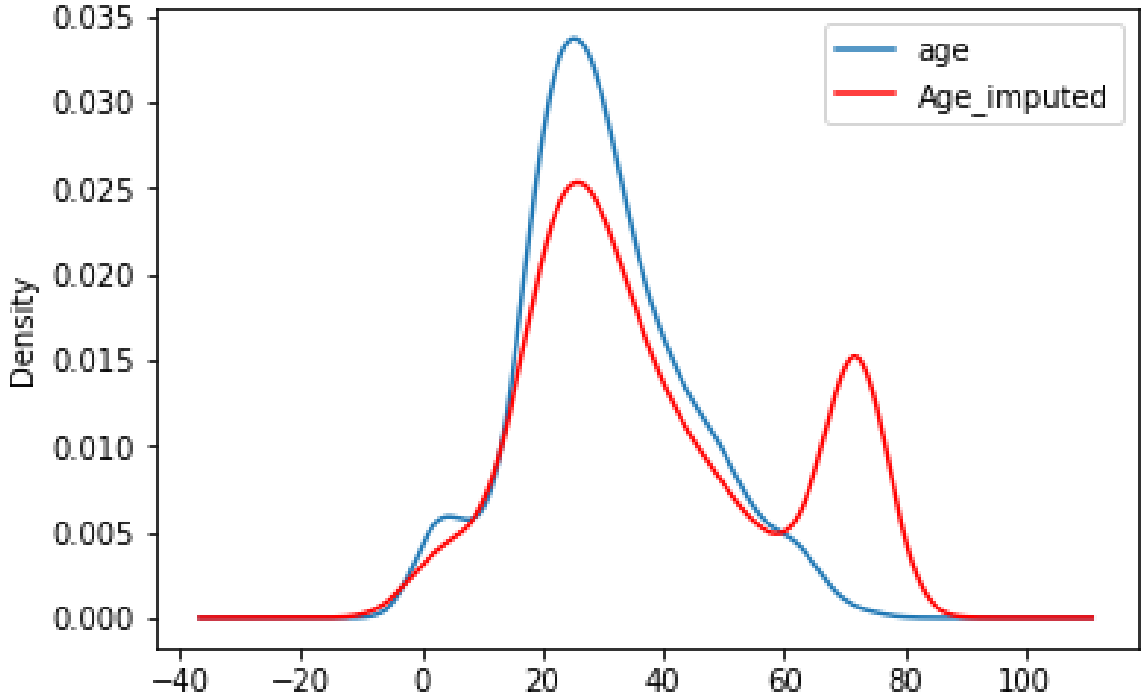
End of tail imputation: how to do it

Histogram of Age from Titanic



Mean(Age) + 3 × std(Age) = 72

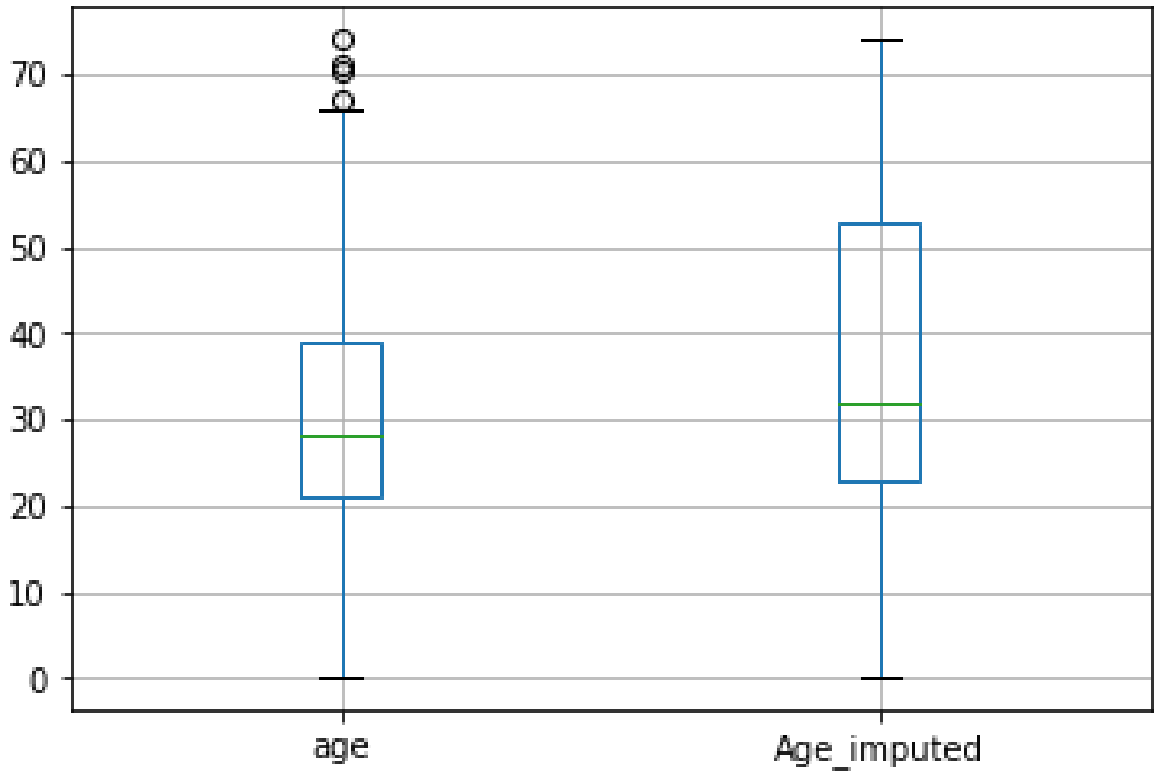
End of tail imputation and distribution



- ~20% of data is missing in Age

Original variable variance: 194
Variance after imputation: 427

End of tail imputation and outliers



Masks outliers

End of tail imputation: effects

	fare
fare	2248.326729
age	136.176223
Age_imputed	19.647139

Train In Data

THANK YOU

www.trainindata.com