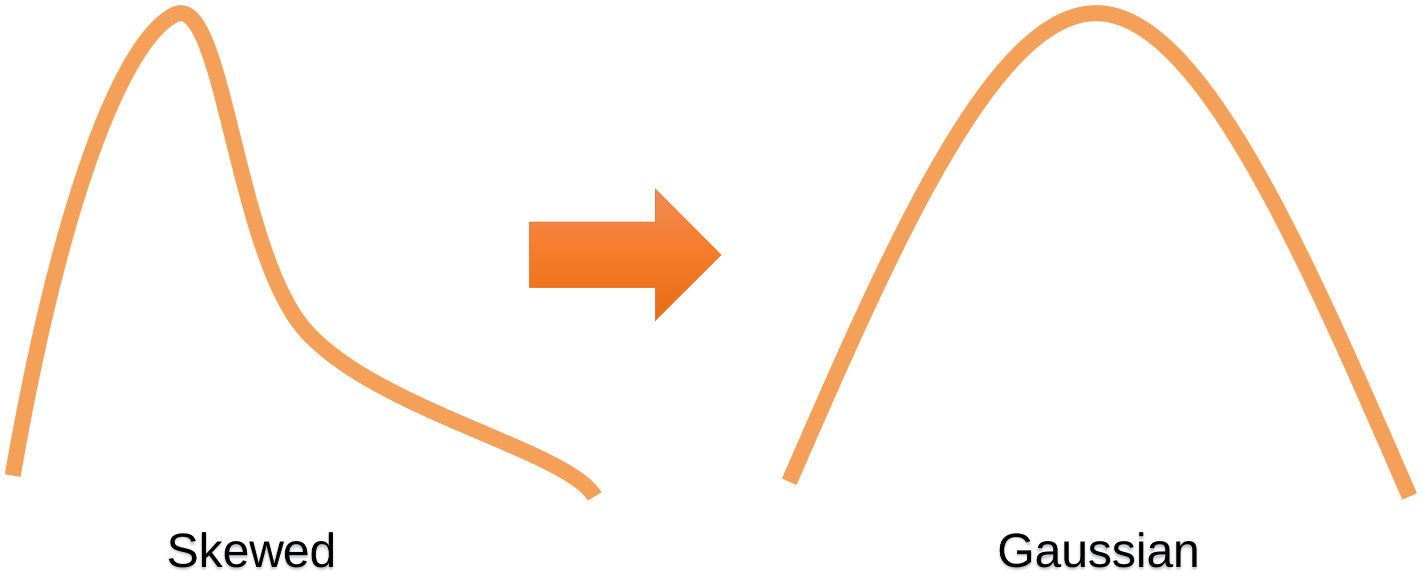
Train In Data

Variable Transformation

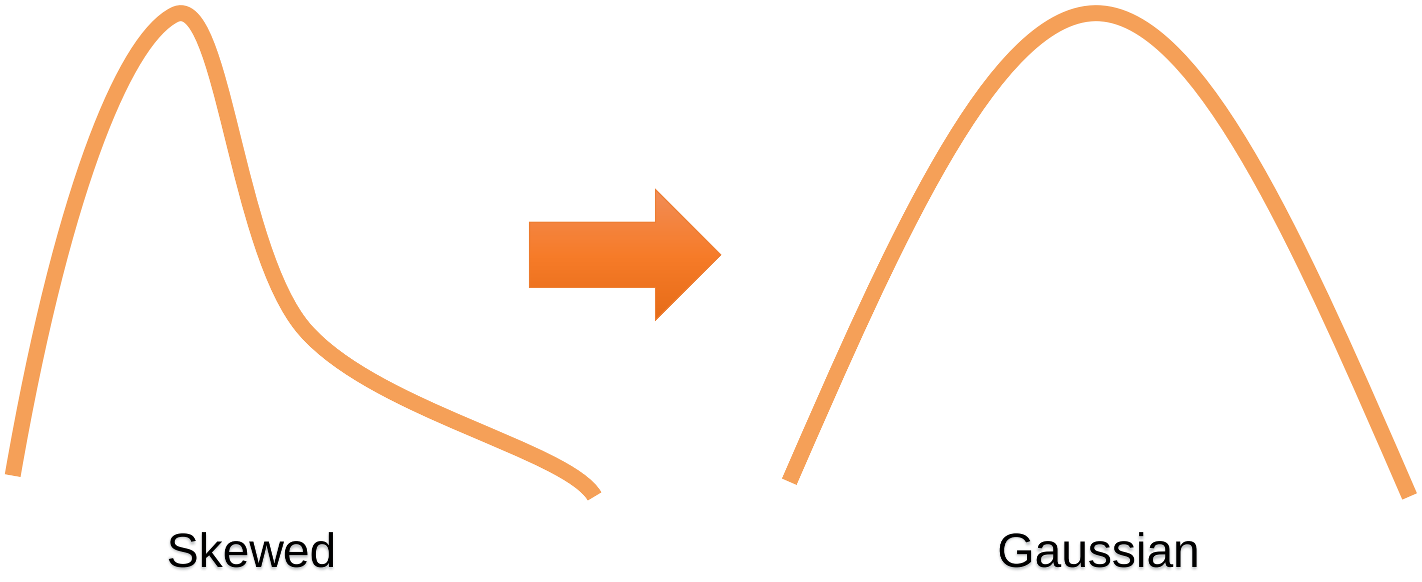
Variable transformation



Variable transformation

- Logarithmic
- Reciprocal
- Square-root
- Arcsin
- Power
- Box-Cox
- Yeo-Johnson

Variable transformation



Variable transformation

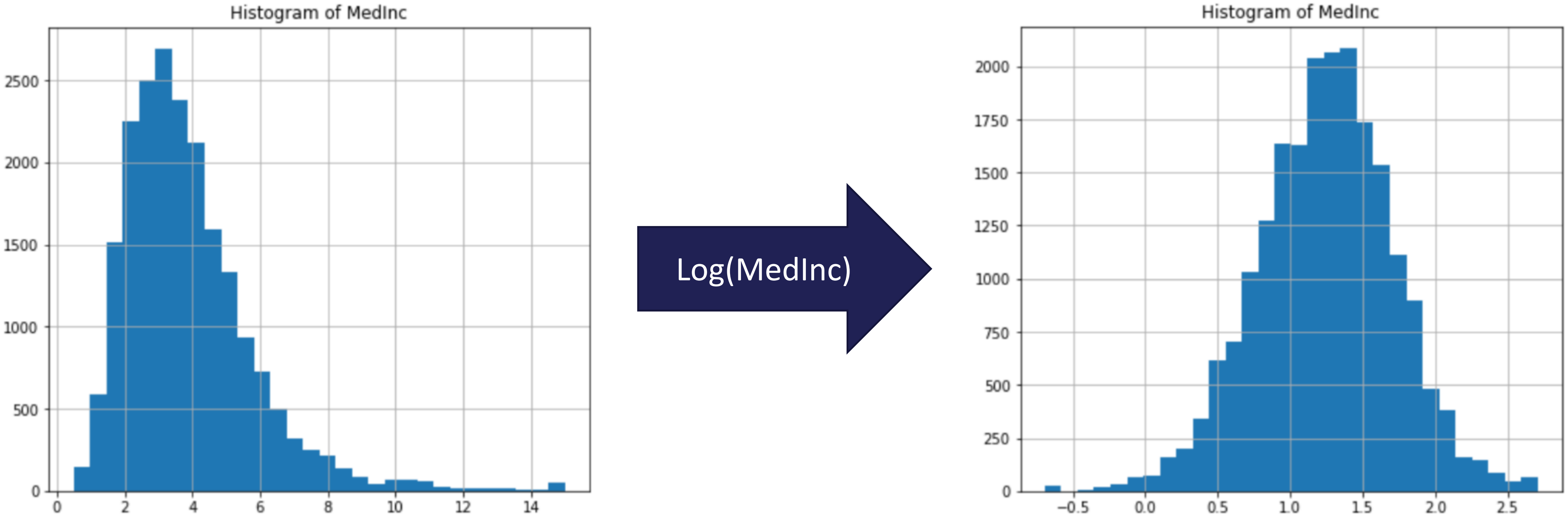
- **Logarithmic**
- **Reciprocal**
- **Square-root**
- **Arcsin**
- **Power**
- Box-Cox
- Yeo-Johnson

Logarithm

The logarithm deals with **positive data** with a **right-skewed distribution** (observations accumulate at lower values of the variable).

X_new = log(X)

Logarithm



California housing dataset.

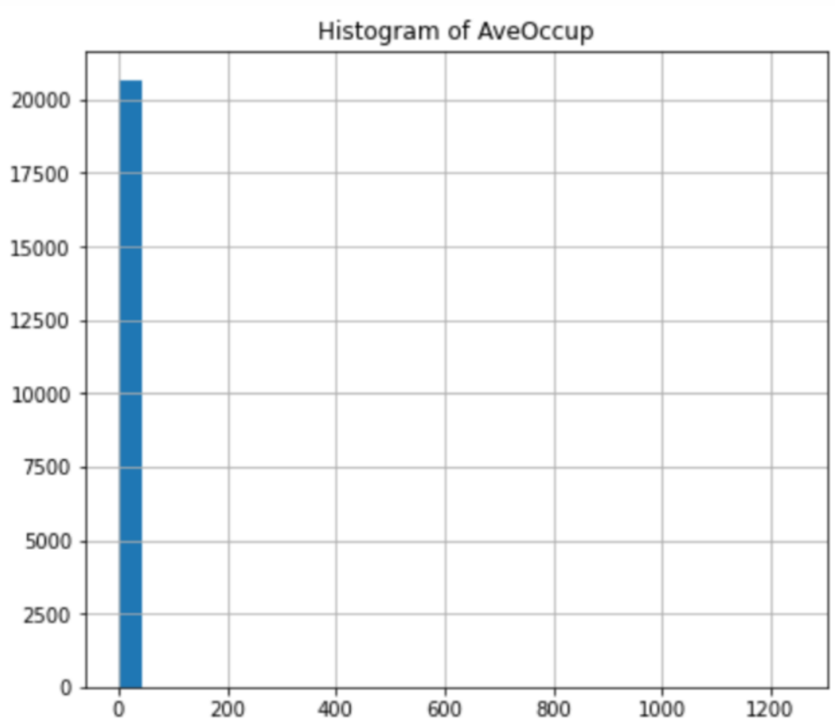
Reciprocal

The reciprocal transformation is useful when we have **ratios**, that is, values resulting from the division of two variables.

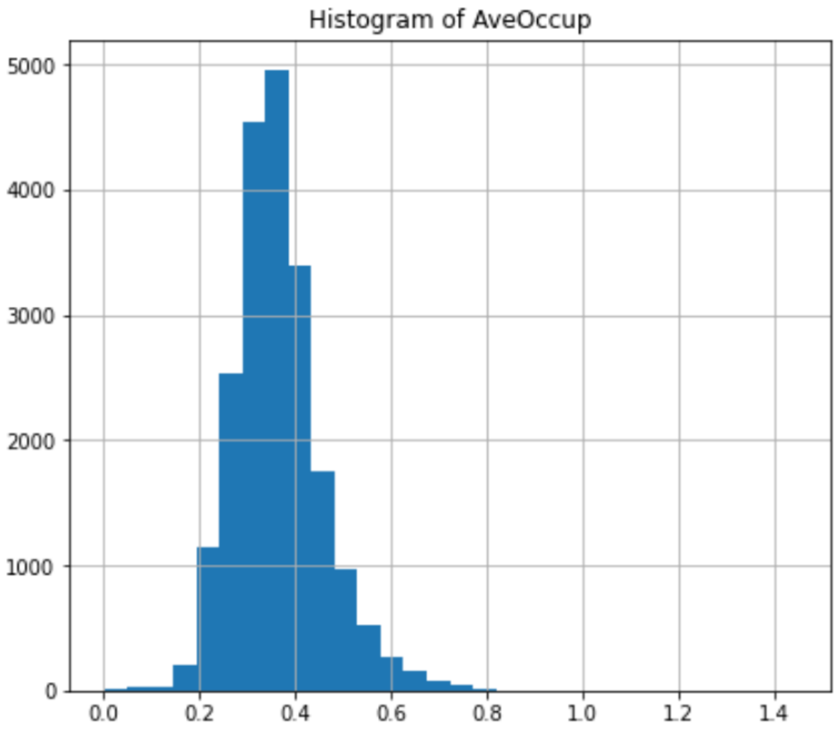
Classical examples are **population density**, that is, people per area, or **house occupancy**, that is, the number of occupants per house.

X_new = 1 / X

Reciprocal



1 / AveOccup



California housing dataset.

Square-root

The square root transformation is suitable for variables with a Poisson distribution **(counts)**. It transforms them into variables with an approximately standard Gaussian distribution.

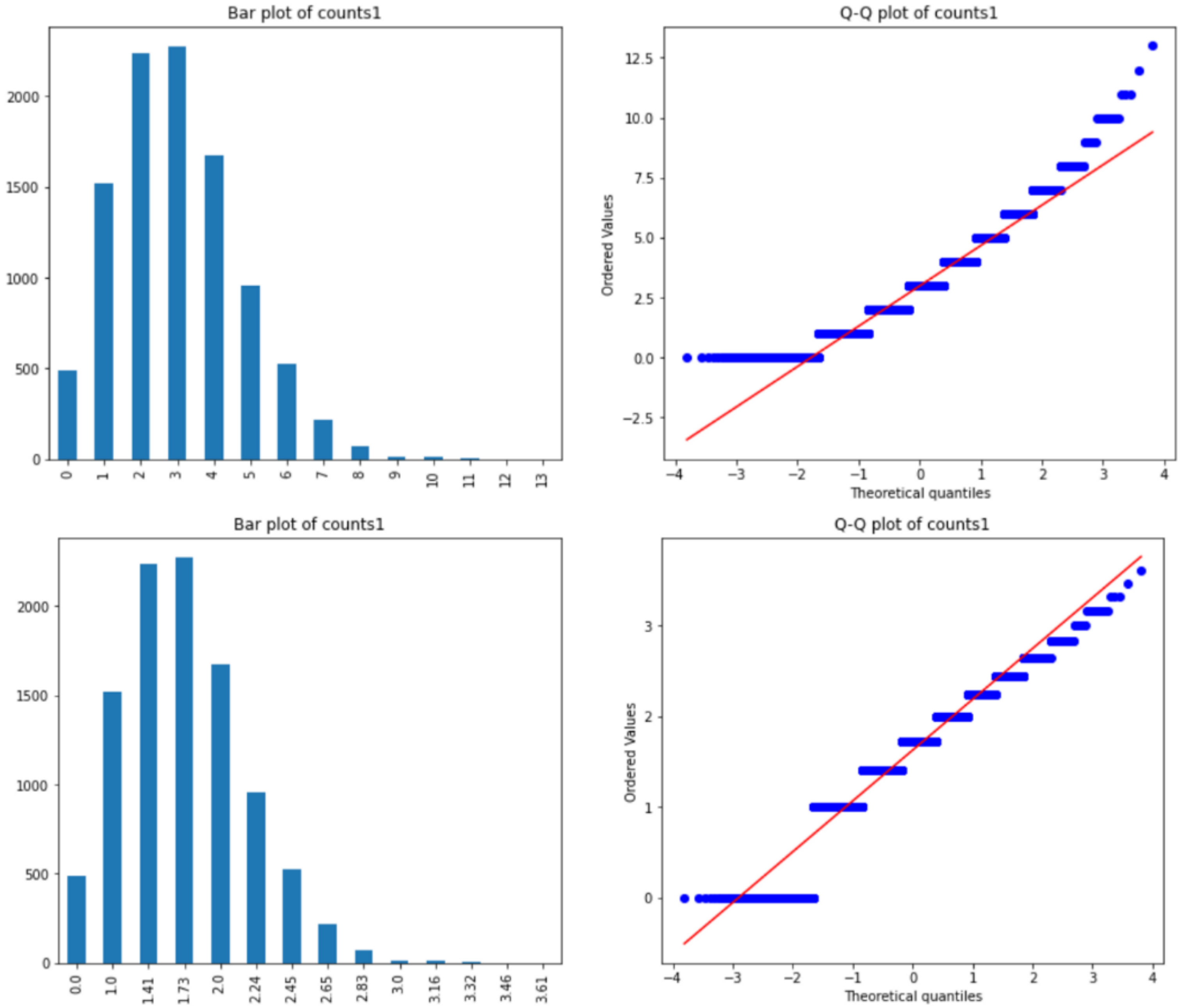
The square root transformation is a form of **power transformation** where the exponent is 1/2 and is only defined for **positive values**.

$$\text{X\_new} = \sqrt{X}$$

$$\text{X\_new} = X^{1/2}$$

Square-root

Example with a toy variable with a Poisson distribution.



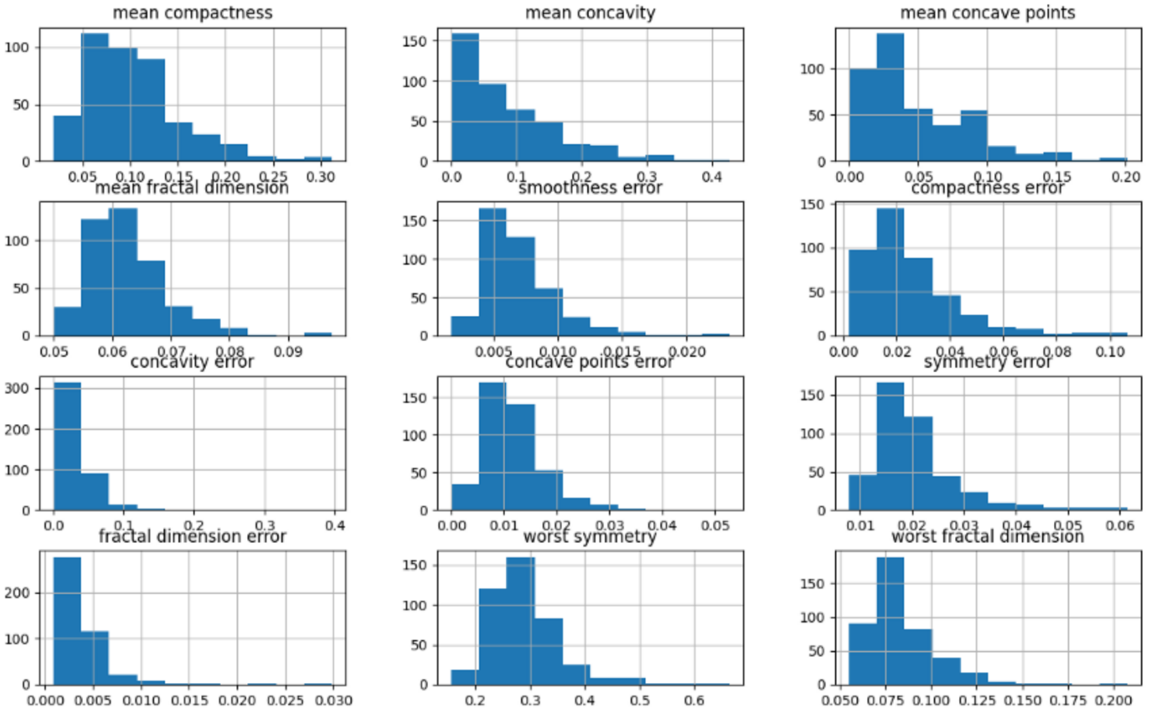
Arcsin

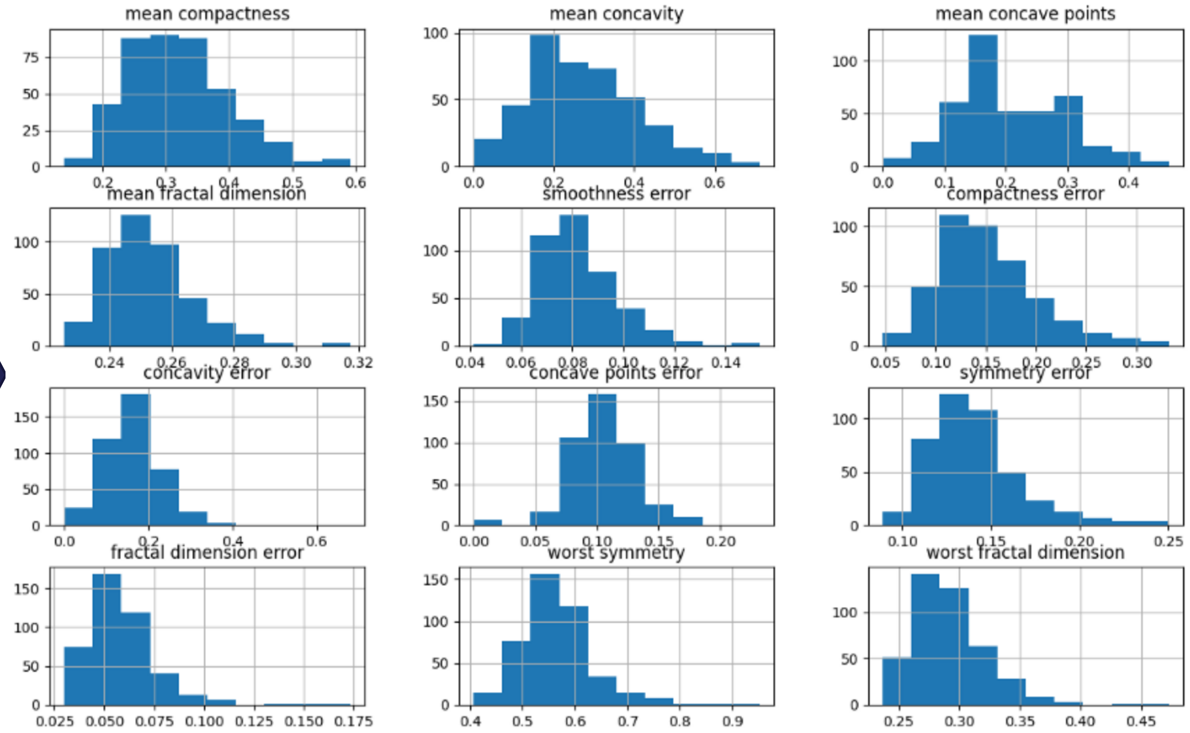
The arcsin square root transformation helps in dealing with probabilities, percentages, and proportions.

The variable (X) varies between 0 and 1.

$$X\_new = \arcsin(\sqrt{x})$$

Arcsin





Breast cancer dataset.

Power

$$X\_new = X^{lambda}$$

Lambda needs to be optimized.

As general guidance:

- If data is right-skewed (i.e. more observations around lower values), use lambda <1.
- If data is left-skewed (i.e. more observations around higher values), use lambda >1.

Train In Data

THANK YOU

www.trainindata.com