Feature Engineering

Papangkorn Inkeaw, Ph.D.

Feature Improvement

Chapter 3 (Part III)

Feature Scaling

- ML algorithms perform mathematical operations with features that assume their values are comparable.
- So, we should make features comparable.
- Simple approach, scale the features so all the feature values have the same magnitude are centered on zero.
  - Normalization
  - Standardization

Note that the normalization/standardization parameters computed over the training set. they are applied at runtime (and to the test set)

Feature Scaling

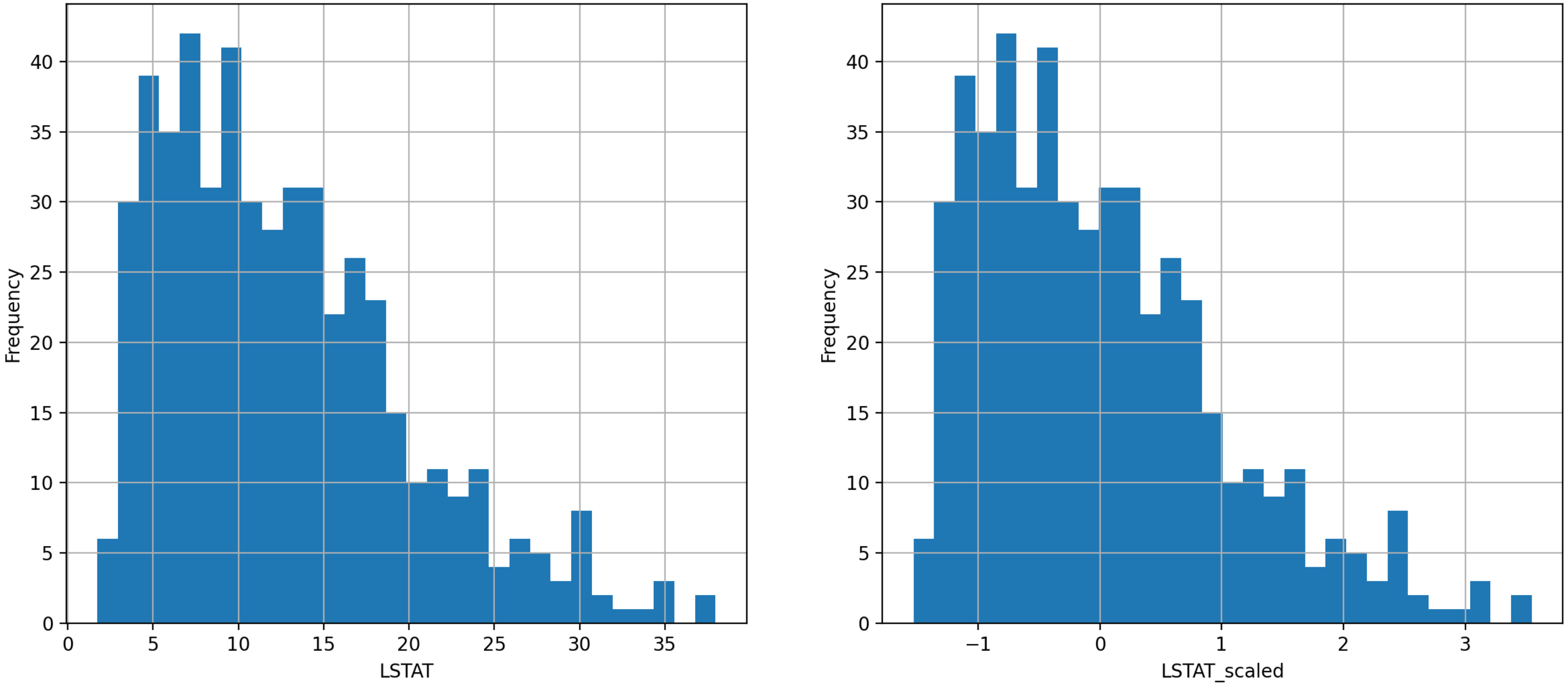
Standardization

- Transforms the features to have zero mean and unit variance.
- A value of feature X can be scaled by:

$$x_{scaled} = \frac{x - \mathrm{mean}(X)}{\mathrm{std}(X)}$$

- Also called the z-score
- This scaling represents how many standard deviations a given observation deviates from the mean.

Feature Scaling



The distribution of LSTAT variable in Boston House Prices dataset before and after standardizing.

Feature Scaling

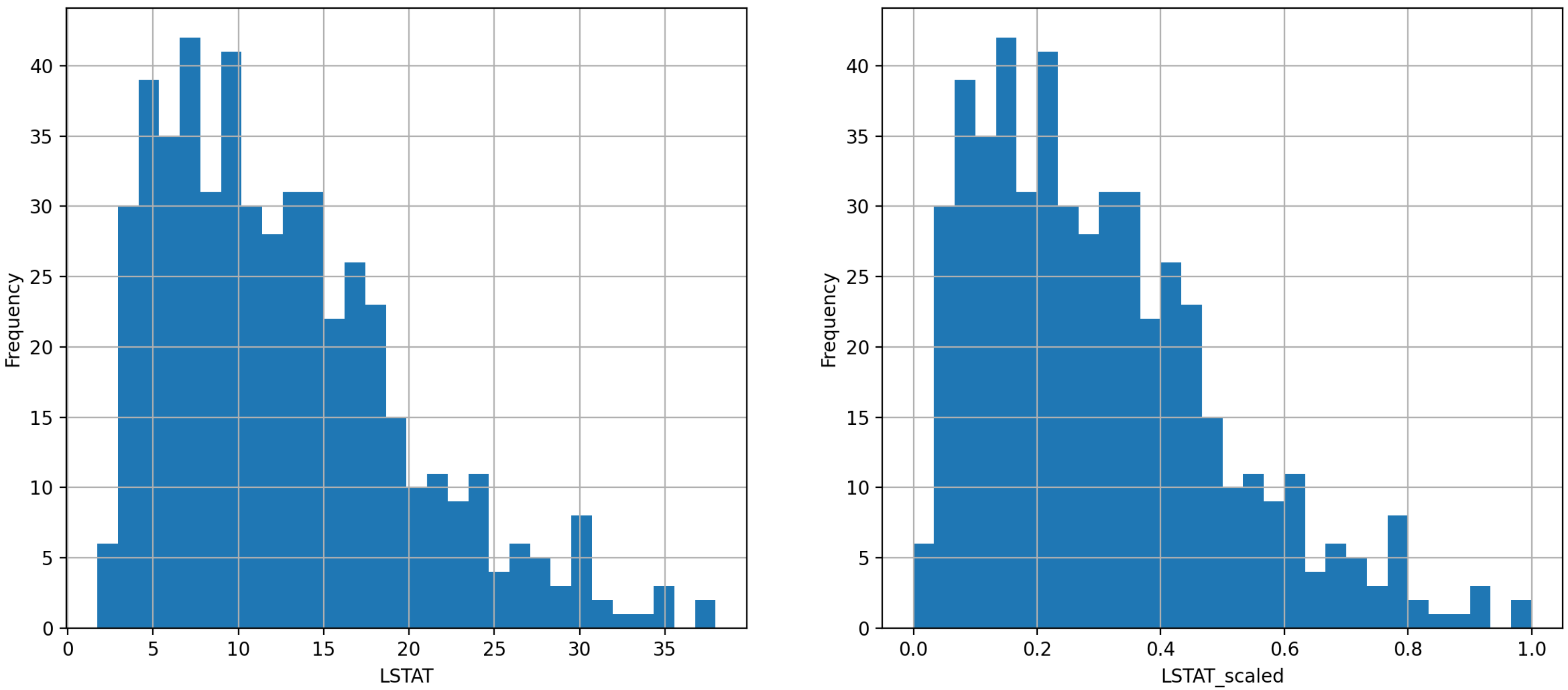
Max-Min Normalization

- Scale to the minimum and maximum values squeezes the values of the variables between 0 and 1.
- A value of feature X can be scaled using the minimum and maximum of X by:

$$x_{scaled} = \frac{x - \min(X)}{\max(X) - \min(X)}$$

- This method has the problem that outliers might concentrate the values on a narrow segment.

Feature Scaling



The distribution of LSTAT variable in Boston House Prices dataset before and after applying max-min normalization.

Feature Scaling

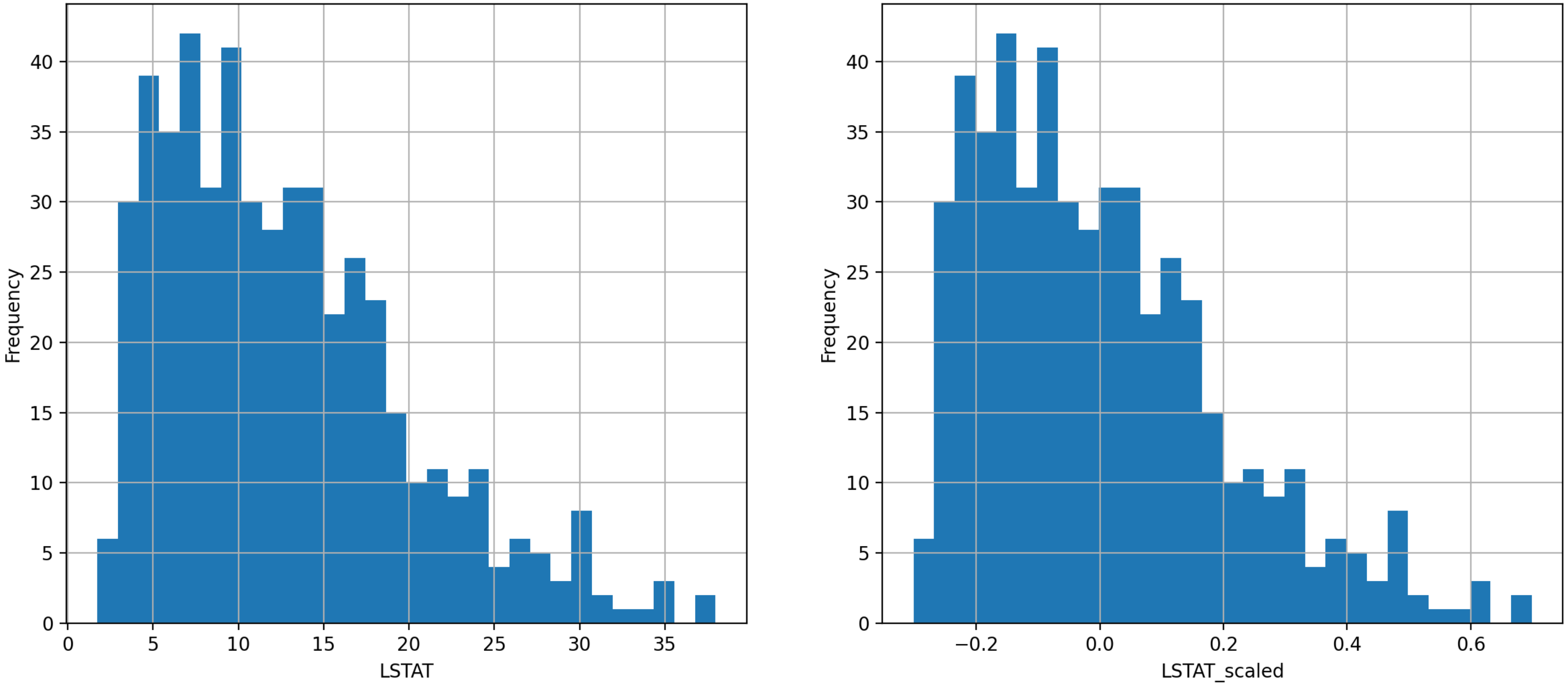
<u>Mean Normalization</u>

- Center the variable at zero and rescale the distribution to the value range.
- This method subtract the mean from each observation and then divide the result by the difference between the minimum and maximum values:

$$x_{scaled} = \frac{x - \mathrm{mean}(X)}{\max(X) - \min(X)}$$

- The distribution of scaled feature is centered at 0, with its minimum and maximum values within the range of -1 to 1.

Feature Scaling



The distribution of LSTAT variable in Boston House Prices dataset before and after applying mean normalization.

Feature Scaling

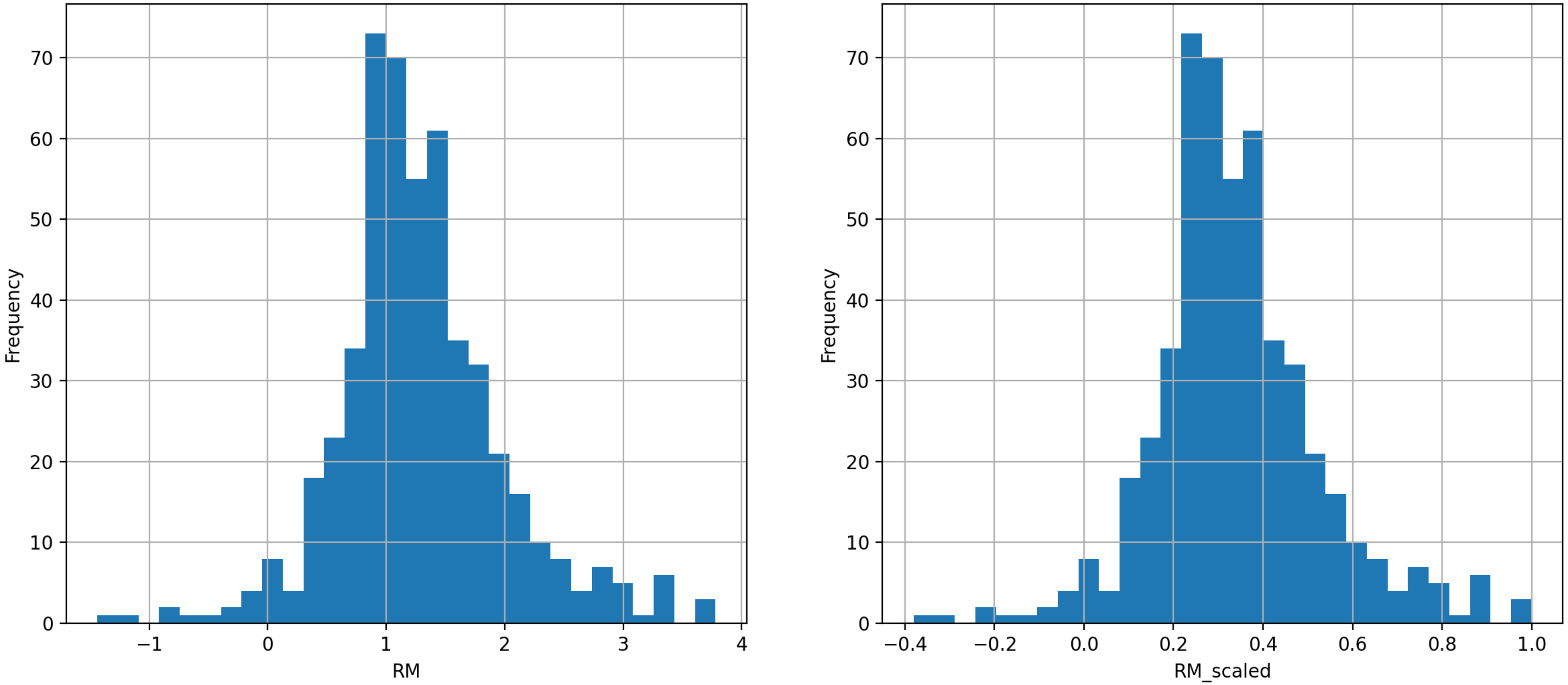
<u>Maximum Absolute Scaling</u>

- Scale a feature to its maximum value
- It divides every observation by the maximum value of the variable:

$$x_{scaled} = \frac{x}{\max(X)}$$

- The scaled values vary approximately within the range of -1 to 1.

Feature Scaling



The distribution of LSTAT variable in Boston House Prices dataset before and after applying maximum absolute Scaling.

Feature Scaling

Normalize to Unit Length

- It applied to multiple features at once.
- Transform the components of a feature vector so that the transformed vector has a length of 1.
- This method is achieved by dividing each observation vector by either:
  - Manhattan distance (l1 norm) is given by:
$$\|\mathrm{x}\| = l1(\mathrm{x}) = |x_1| + |x_2| + \cdots + |x_d|$$
  - Euclidean distance (l2 norm) is given by:
$$\|\mathrm{x}\| = l2(\mathrm{x}) = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$$
- For a feature vector $\mathrm{x} = (x_1, x_2, \ldots, x_d)$, the scaled feature vector is computed by:
$$\mathrm{x}_{scaled} = \left(\frac{x_1}{\|\mathrm{x}\|}, \frac{x_2}{\|\mathrm{x}\|}, \ldots, \frac{x_d}{\|\mathrm{x}\|}\right)$$

Numerical Feature Transformation

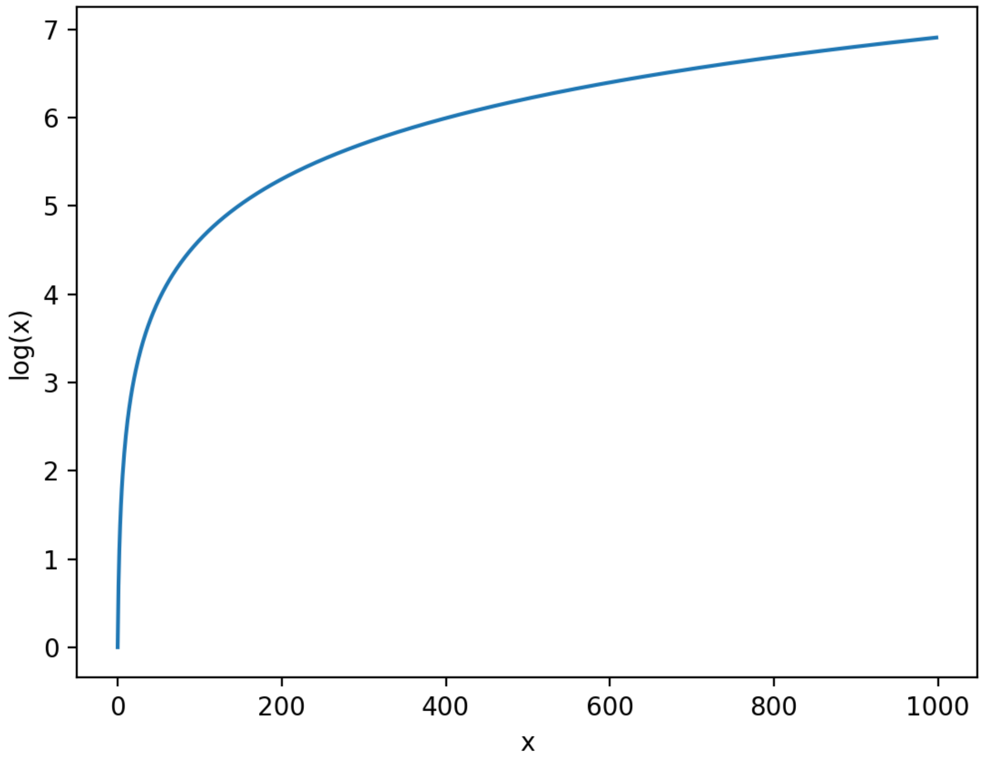
- Some ML models (e.g., linear regression and logistic regression) assume that the variables are normally distributed.
- Mathematical transformation can change the distribution of a variable into normal distribution.
- Common mathematical transformations:
  - Log Transformation
  - Reciprocal Transformation
  - Square-root Transformation
  - Box-Cox Transformation
  - Yeo-Johnson Transformation

Numerical Feature Transformation

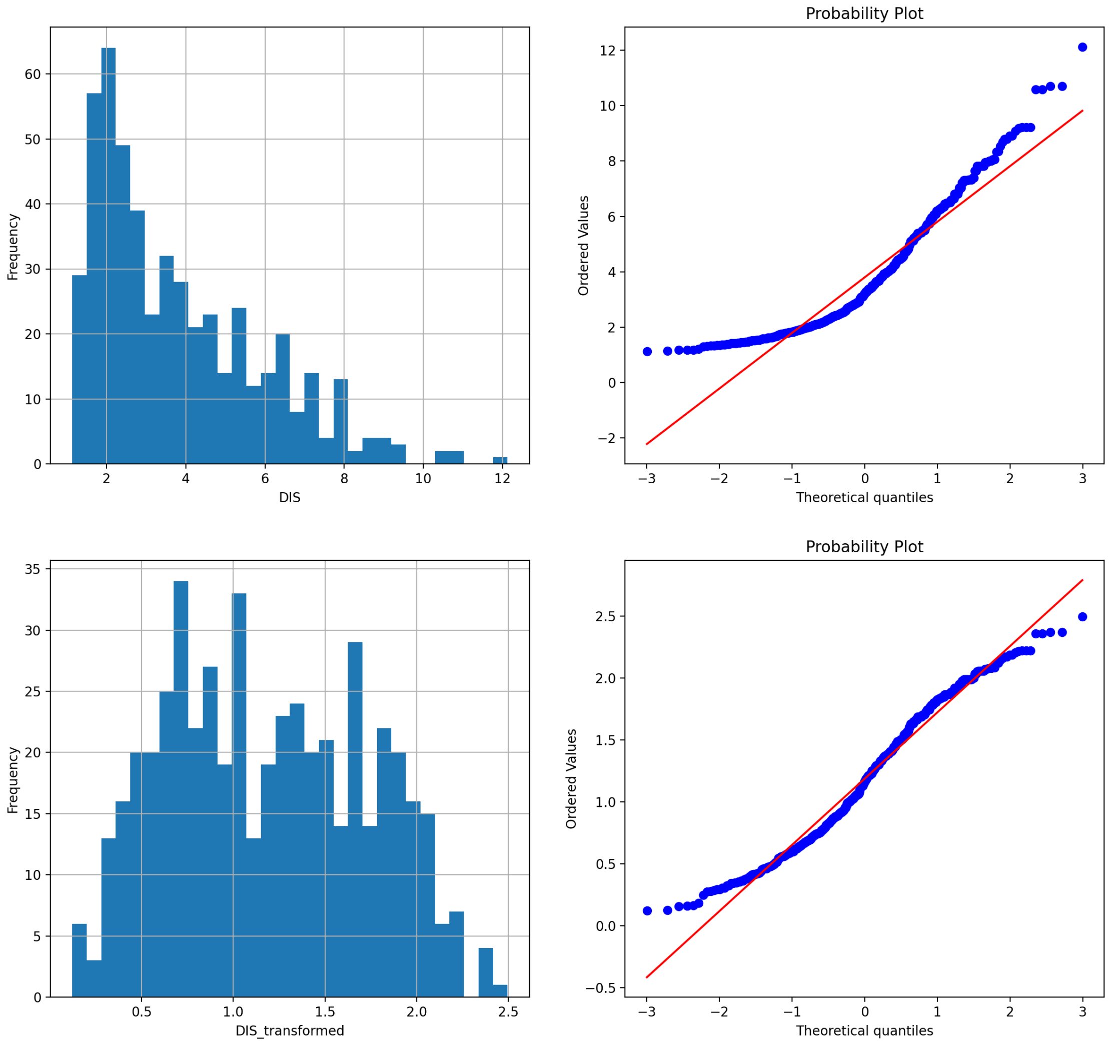
Log Transformation

- A powerful tool for dealing with positive numbers with a heavy-tailed distribution.
- It help to reduce the skewness of the original data.
- It uses a logarithm function to transform a positive numerical feature:

$$x_{tramsformed} = \log(x)$$



Numerical Feature Transformation



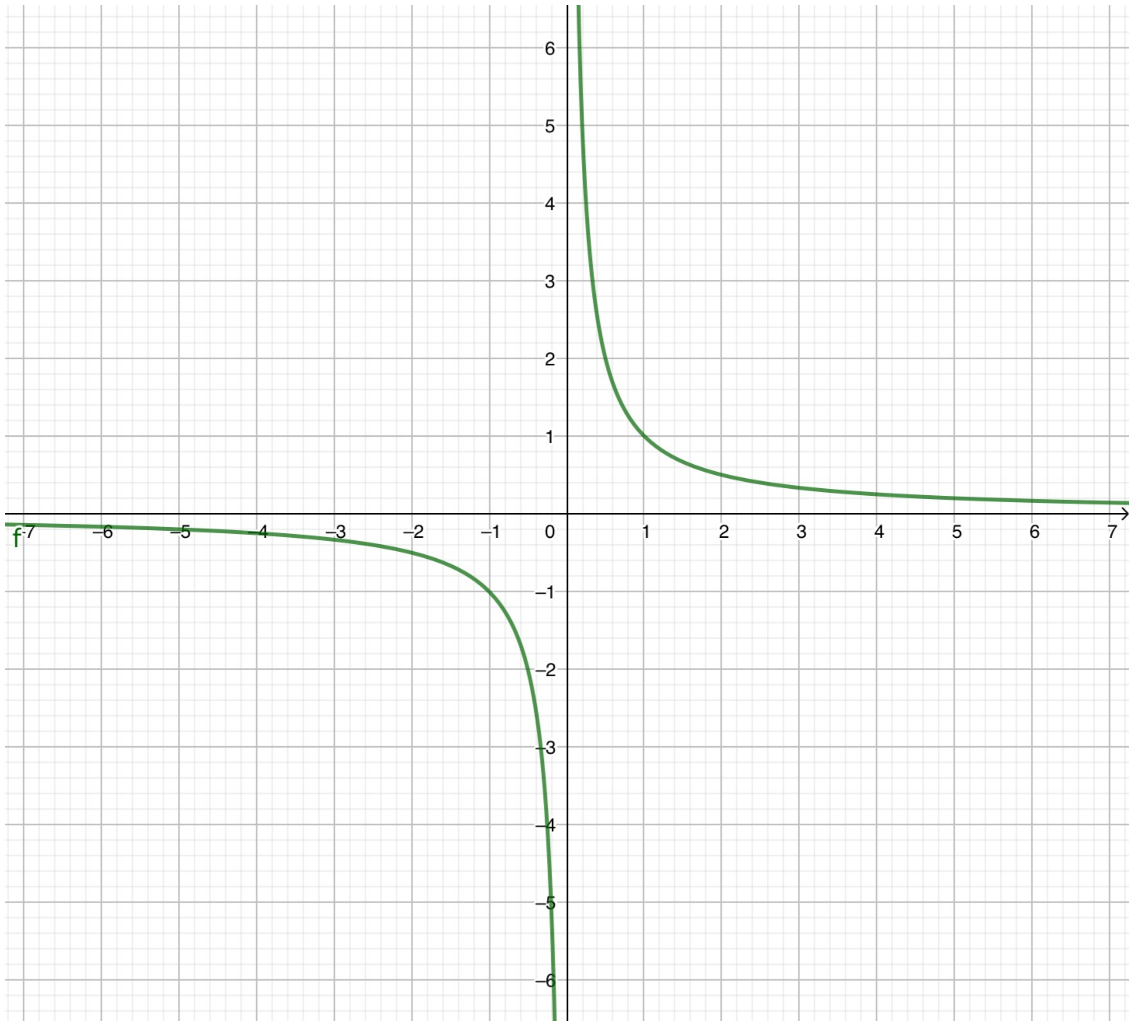
The distribution of DIS variable in Boston House Prices dataset and Q-Q plot before and after applying log transformation.

Numerical Feature Transformation

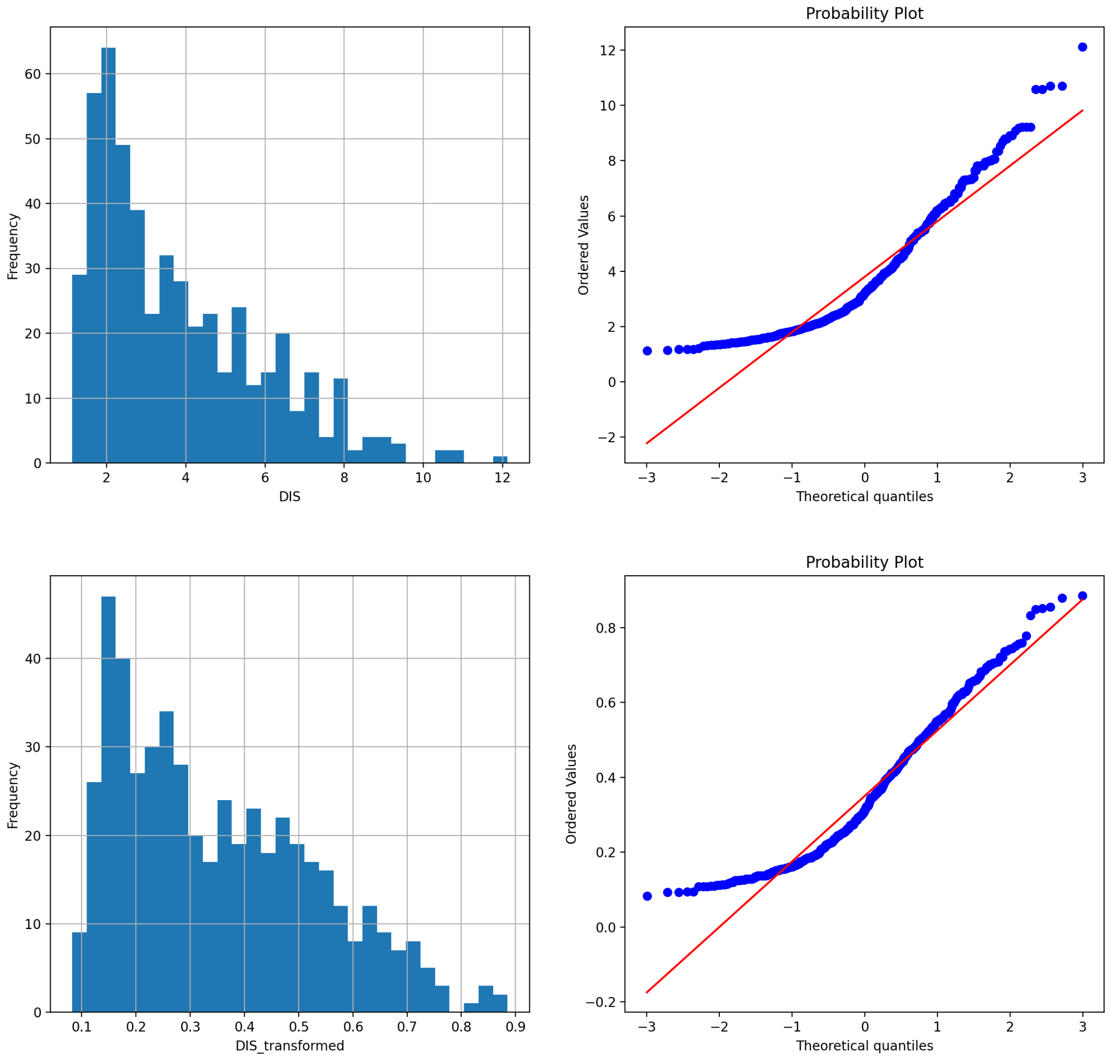
Reciprocal Transformation

- It has a dramatic effect on the shape of the distribution, reversing the order of values with the same sign.
- It is taken for data expressing right skewness; it converts it to a normal distribution.
- Reciprocal transformation maps non-zero values of x to 1/x (or -1/x for negative values):

$$x_{tramsformed} = \frac{1}{x}$$



Numerical Feature Transformation



The distribution of DIS variable in Boston House Prices dataset and Q-Q plot before and after applying reciprocal transformation.

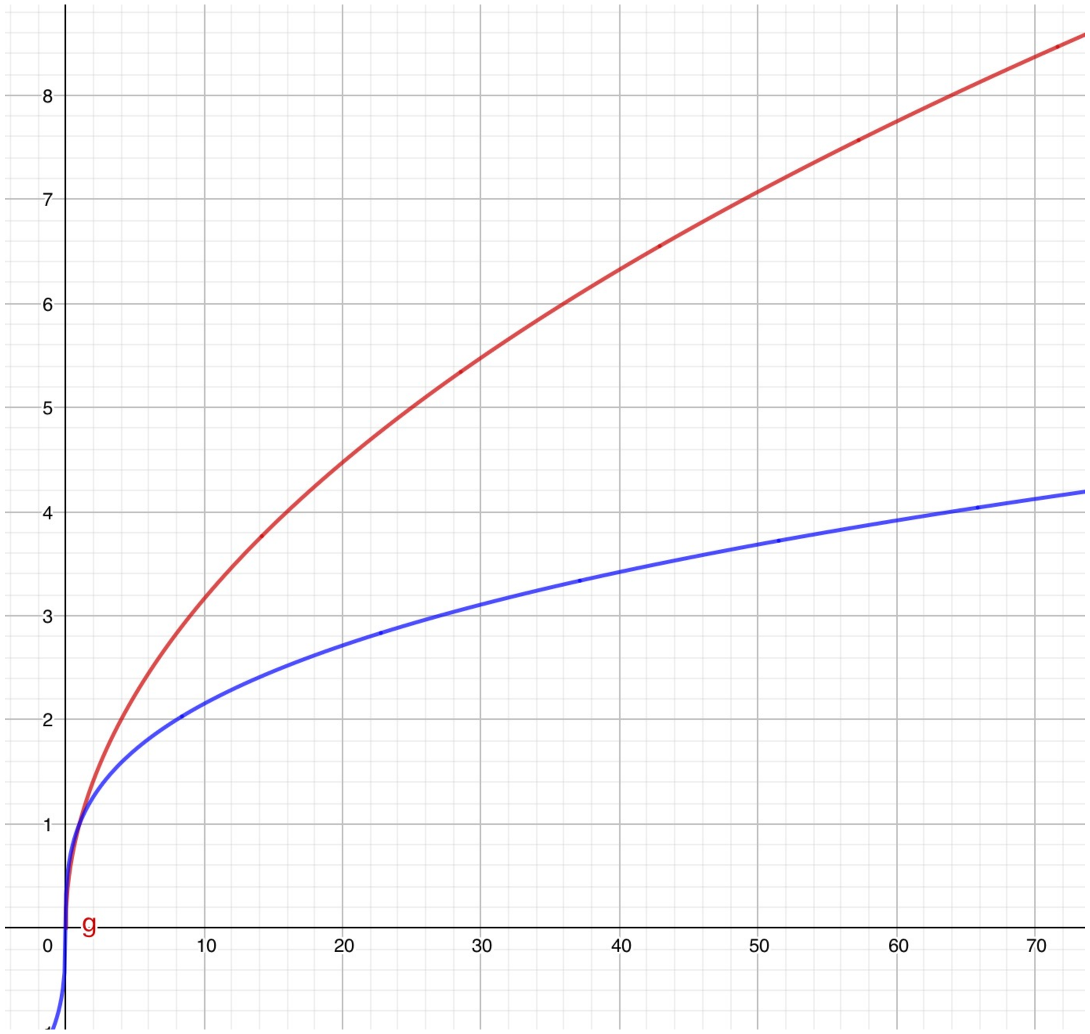
Numerical Feature Transformation

Square-root Transformation

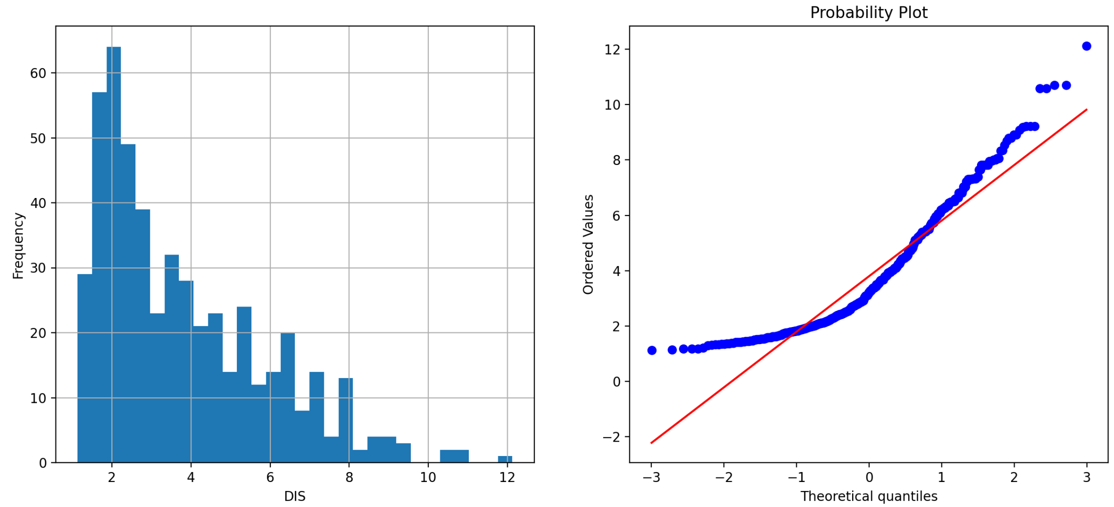
- The square root must be considered when variance is proportional to the mean.
- Stabilizing the variance of the distribution
- It is a special case of power transformations that apply the following transformation:

$$x_{tramsformed} = x^{\lambda}$$

- where $\lambda = \frac{1}{2}\ or \frac{1}{3}$

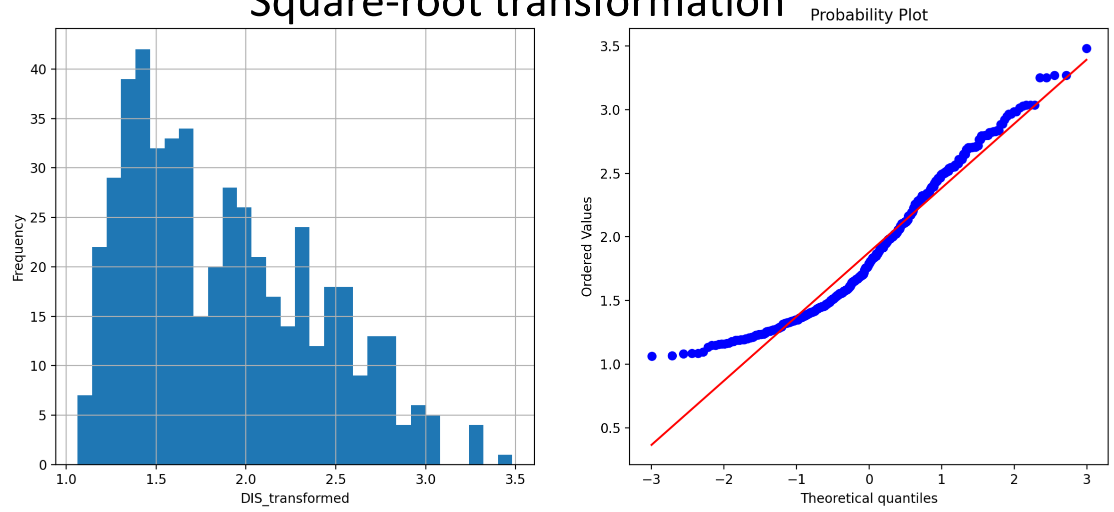


Numerical Feature Transformation

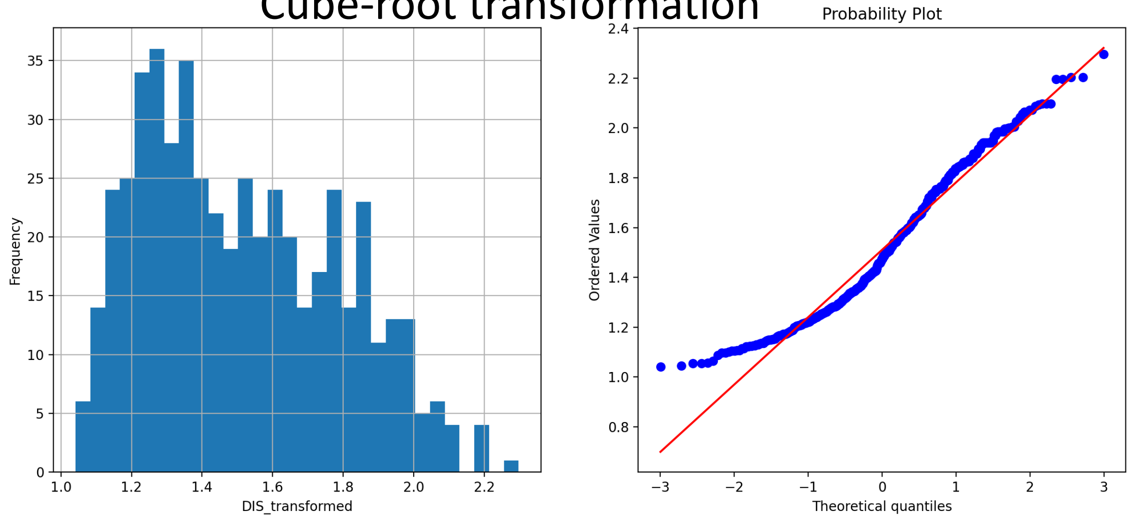


The distribution of DIS variable in Boston House Prices dataset and Q-Q plot before and after applying square-root and cube-root transformations.

Square-root transformation



Cube-root transformation



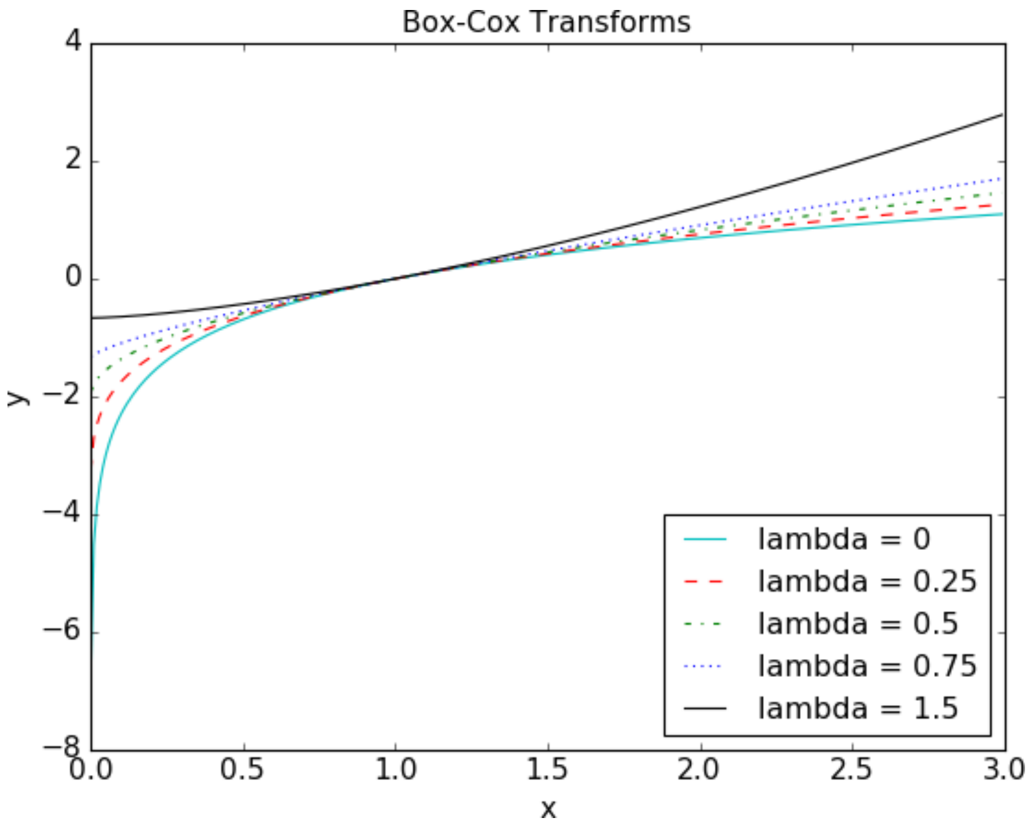
Numerical Feature Transformation

Box-Cox Transformation

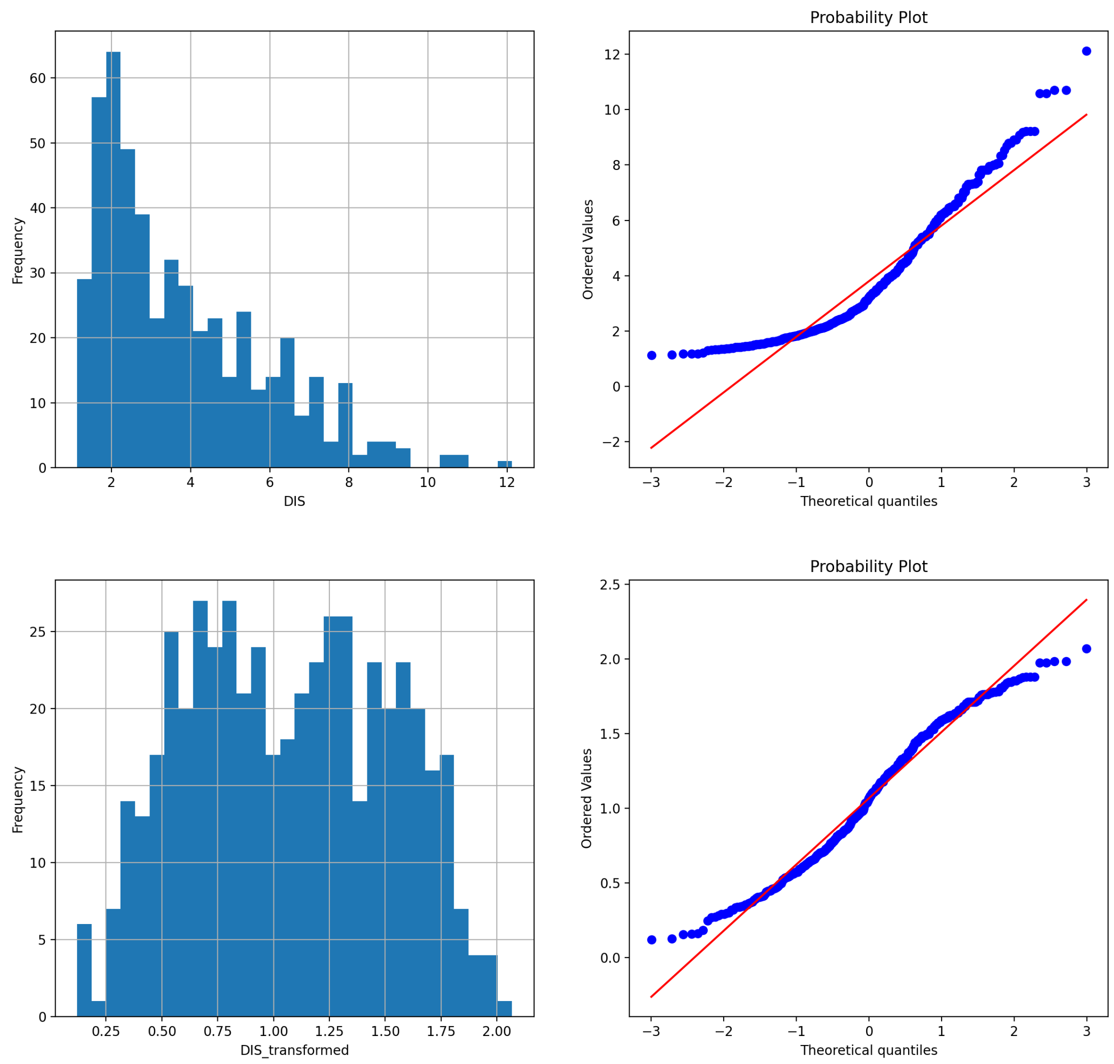
- Box-Cox transformation belongs to the power family of functions.
- It flexible in its ability to address many different data distributions.
- It was defined as

$$x_{tramsformed} = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln(x)\,, & \lambda = 0 \end{cases}$$

- The Box-Cox formulation only works when the data is positive.
- For nonpositive data, one could shift the values by adding a fixed constant.
- Uses maximum likelihood estimation to estimate a transformation parameter λ . (vary between -5 and 5)



Numerical Feature Transformation



The distribution of DIS variable in Boston House Prices dataset and Q-Q plot before and after applying Box-Cox transformations. The optimal $\lambda = -0.1556$.

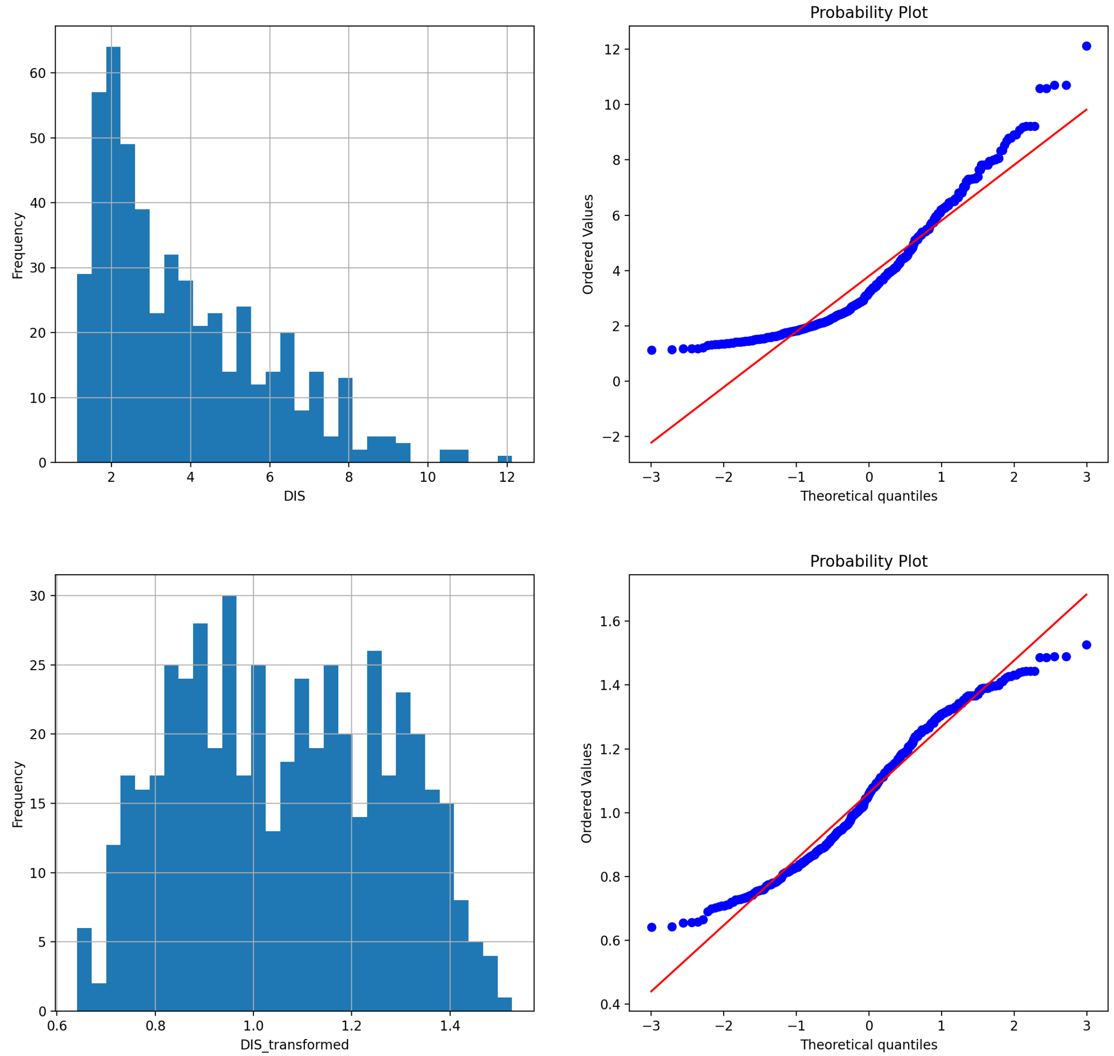
Numerical Feature Transformation

Yeo-Johnson Transformation

- An extension of the Box-Cox transformation.
- Can apply to negative, zero and positive values.
- It was defined as

$$x_{tramsformed} = \begin{cases} \dfrac{(x+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0 \; and \; x \geq 0 \\ \ln(x+1), & \lambda = 0 \; and \; x \geq 0 \\ \dfrac{(-x+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2 \; and \; x < 0 \\ -\ln(-x+1), & \lambda = 2 \; and \; x < 0 \end{cases}$$

Numerical Feature Transformation



The distribution of DIS variable in Boston House Prices dataset and Q-Q plot before and after applying Yeo-Johnson transformations. The optimal $\lambda = -0.4489$.

References & Study Resources

- Pablo Duboue. (2020). The Art of Feature Engineering: Essentials for Machine Learning. Cambridge University Press.
- Alice Zheng and Amanda Casari. (2018). *Feature Engineering for Machine Learning*. O'Reilly Media, Inc.
- Soledad Galli. (2020). *Python Feature Engineering Cookbook*. Packt Publishing.
- https://medium.com/@muhammadibrahim_54071/why-and-which-data-transformation-should-i-use-cfb9e31923cf
- https://towardsdatascience.com/types-of-transformations-for-better-normal-distribution-61c22668d3b9