E0410 Fundamentals of Statistics for Scientific Data Using R

by Daria Sapunova, PhD student, RECETOX

daria.sapunova@recetox.muni.cz

Bohunice, D29, room 123

Theoretical part

Data preparation – cleaning, harmonizing and structuring

- ✓ Upload your data.
- ✓ Choose variables you are going to work with.
- ✓ Inspect the variables (do descriptive statistics, check unique values, check data type, check amount of missing data).
- ✓ Harmonize your variables (esp. categorical variables).
- ✓ Structure your data.

Preliminary analysis

- ✓Build histograms in case of numerical variables.
- ✓Build box plots in case of categorical variables.
- ✓Build scatterplots to see associations between numerical variables.
- ✓Check data distributions in case of numerical variables (normality tests).
- ✓Transform the data if necessary.

You may proceed to the analysis

Hypothesis testing

Question: Is there a difference in female and male heights?

Null hypothesis (H0)

Hypothesis:

There is NO difference

What is the probability that there is no difference?

The probability is low

Alternative hypothesis (H1)

Hypothesis:

There is a difference

Reject H0, accept H1

Hypothesis testing

The probability is low…

alpha value (the threshold for statistical significance)

Question: Is there a difference in female and male heights?

Null hypothesis (H0) ✗

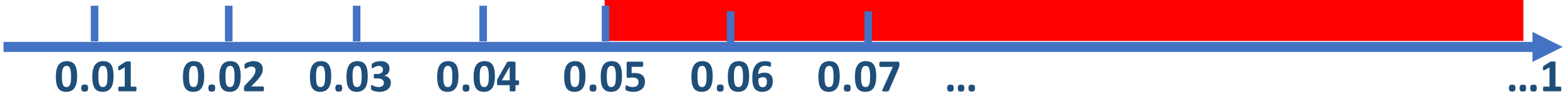
Alternative hypothesis (H1) ✓

→ test → p-value (probability value)

Hypothesis testing

The probability is low…

alpha value (the threshold for statistical significance)



Question: Is there a difference in female and male heights?

Null hypothesis (H0) ✓

Alternative hypothesis (H1) ✗

→ test → p-value (probability value)

Question: Is there a difference in female and male heights?

Null hypothesis (H0)

Hypothesis:

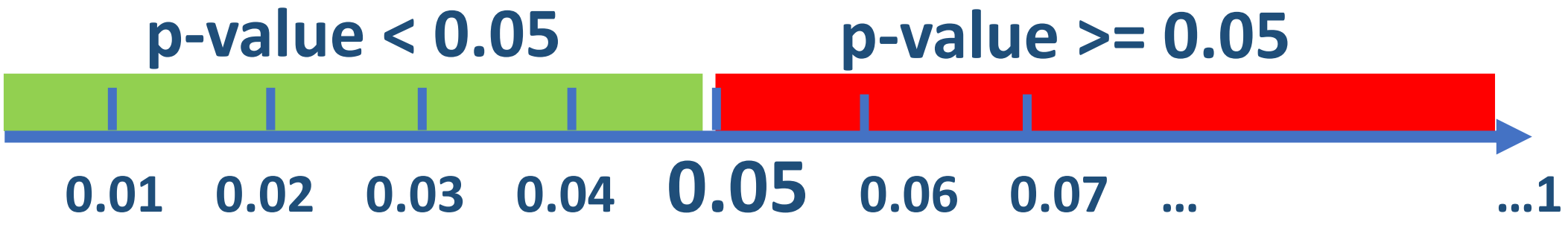
There is NO difference

Alternative hypothesis (H1)

Hypothesis:

There is a difference

Test -> p-value (probability of H0)



H0 (no difference) ✗

H1 (is difference) ✓

H0 (no difference) ✓

H1 (is difference) ✗

Task!

Question:?

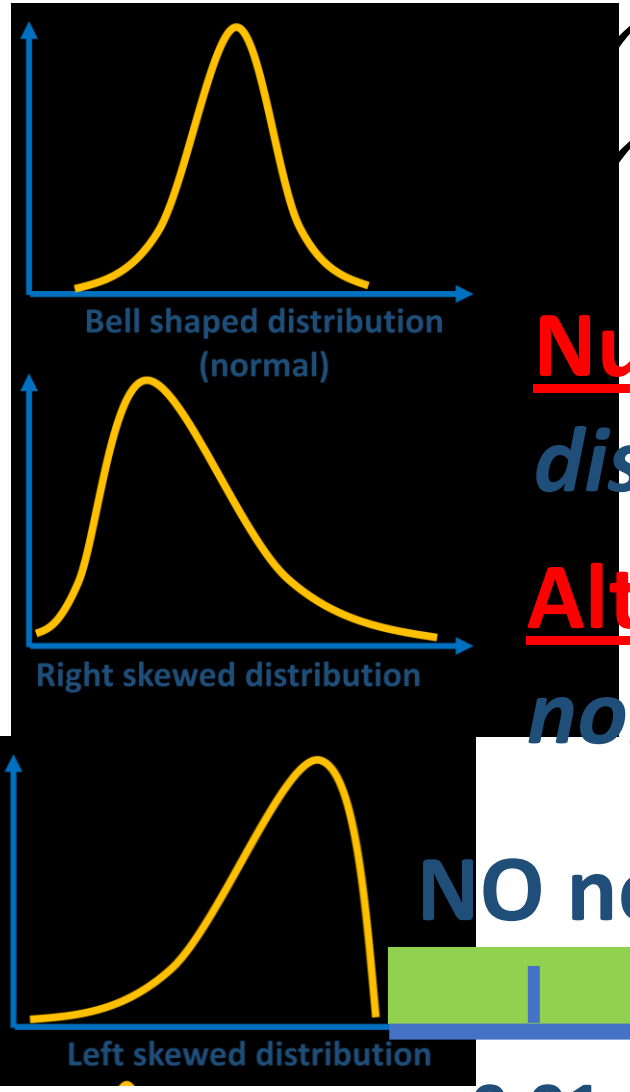
Null hypothesis (H0)

Alternative hypothesis (H1)

p-value = 0.003

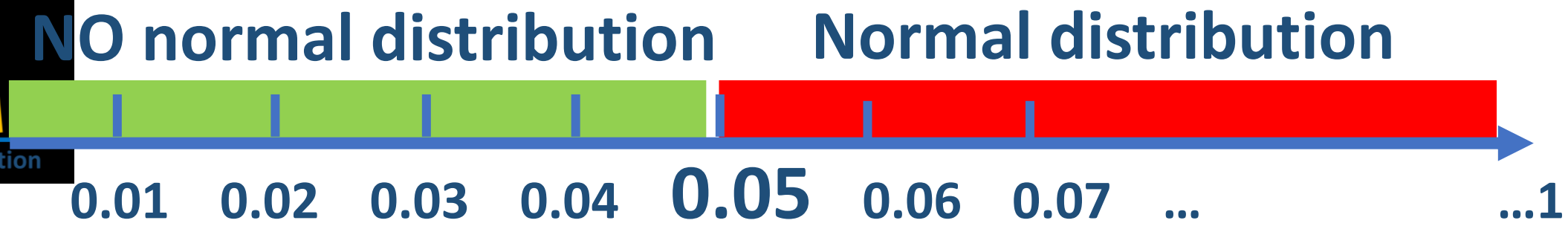
Normality tests

- ✓Kolmogorov-Smirnov Test
- ✓Shapiro-Wilk Test
- ✓Anderson-Darling Test



Null hypothesis (H0): The data *is normally distributed*

Alternative hypothesis (H1): The data *is NOT normally distributed*



Normality tests

Shapiro-Wilk Test: The Shapiro-Wilk test is widely regarded as one of the most powerful normality tests. Developed in 1965 by Samuel S. Shapiro and Martin B. Wilk, **it is specifically designed for small to moderate sample sizes.** The test calculates a W statistic, which compares the observed data to the expected data if it follows a normal distribution. A small W value indicates that the data deviates significantly from a normal distribution. **It is sensitive to deviations from normality in the tails of the distribution.**

Kolmogorov-Smirnov Test: The Kolmogorov-Smirnov (K-S) test is a non-parametric test that compares the empirical distribution function of a dataset to a specified theoretical distribution, typically the normal distribution. The test calculates the maximum difference (D) between the two cumulative distribution functions. A large D value indicates a significant deviation from normality. **One limitation of the K-S test is that it is less sensitive to deviations in the distribution's tails.**

Anderson-Darling Test: The Anderson-Darling test, developed by Theodore Anderson and Donald Darling in 1952, is another powerful normality test. Like the K-S test, it compares the empirical distribution function of a dataset to a specified theoretical distribution. **Still, it places more weight on the distribution's tails.** This test calculates an A^2 statistic, with larger values indicating a greater deviation from normality.

In summary, while all three tests can be used to assess normality, they have different sensitivities and are suitable for different purposes and sample sizes. The Shapiro-Wilk test is often preferred for testing normality specifically, especially with smaller sample sizes, while **the Kolmogorov-Smirnov and Anderson-Darling tests are more general** and suitable for comparing any distribution to a theoretical distribution.

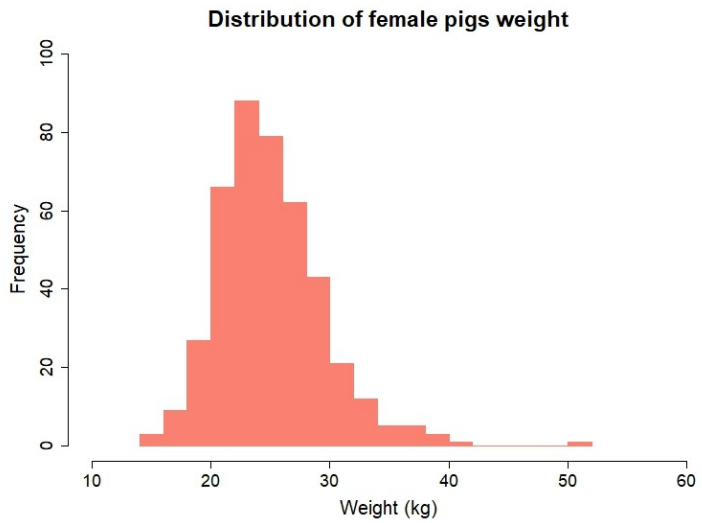
https://statisticseasily.com/normality-test/

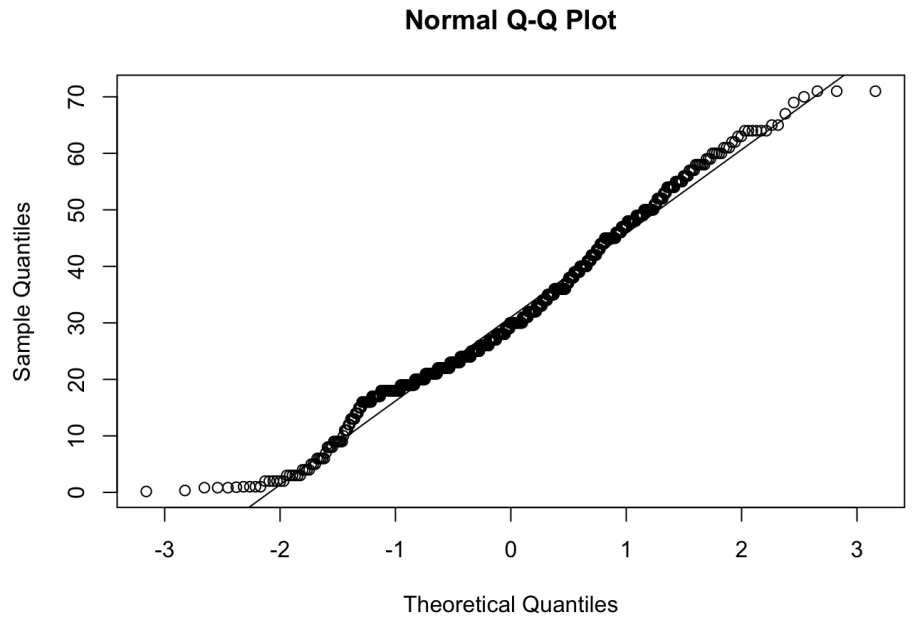
Normality tests

Analytical

- ✓Kolmogorov-Smirnov Test
- ✓Shapiro-Wilk Test
- ✓Anderson-Darling Test

Graphical





https://whitlockschluter3e.zoology.ubc.ca/Tutorials%20using%20R/R_tutorial_Normal_and_sample_means.html#:~:text=The%20QQ%20plot%20should%20follow,a%20QQ%20plot%20for%20you.

Normality tests: disadvantage

- ✓ Depends on sample size: too large or too small sample size can bias the normality test result

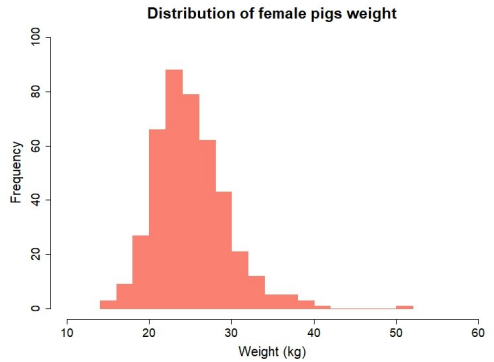
Especially when your distribution *is slightly deviates* from the normal distribution!

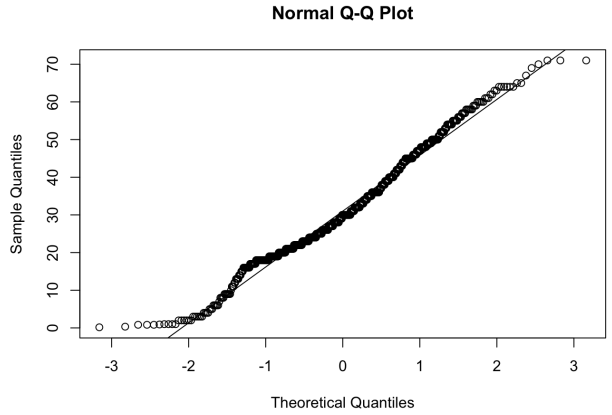
✓Check data distributions in case of numerical variables (normality tests).

Analytical

Graphical

- ✓Kolmogorov-Smirnov Test
- ✓Shapiro-Wilk Test
- ✓Anderson-Darling Test





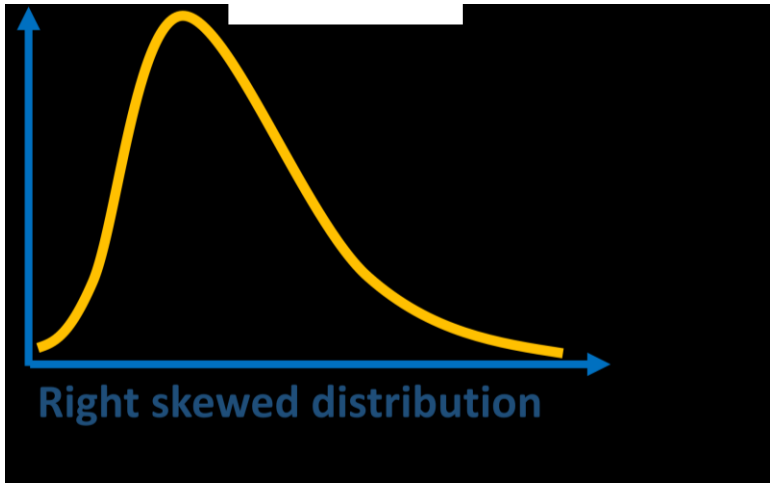
Data transformation:
to reach normal distribution

✓Transform the data if necessary.

- ✓Logarithmic Transformation
- ✓Square Root Transformation
- ✓Reciprocal Transformation
- ✓Box-Cox Transformation
- ✓Yeo-Johnson Transformation
- ✓Rank-Based Transformation
- ✓Arcsine Transformation
- ✓Johnson Transformation

Data transformation

✓Logarithmic Transformation



Logarithmic Transformation:

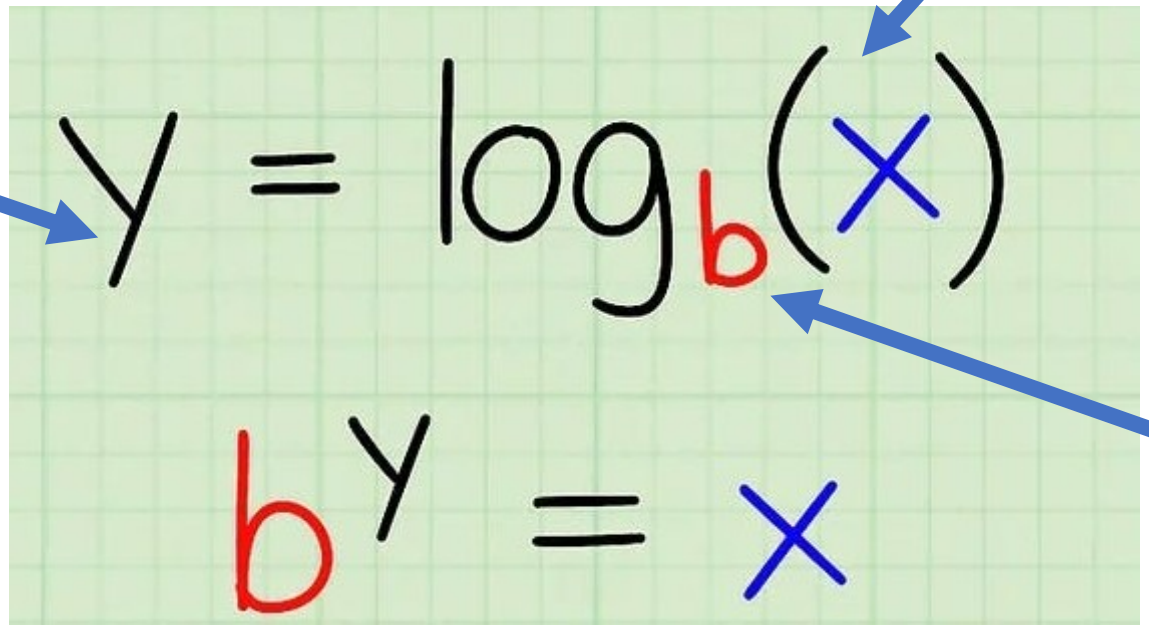
1. Logarithmic transformation **is often applied to right skewed data.**
2. It is particularly useful for data that exhibit exponential growth patterns.
3. Common transformations include **natural logarithm (ln) or base-10 logarithm.**

Artificial intelligence tools were used for generating descriptions of the data transformation types.

Data transformation:
Logarithmic Transformation

Our data we want to transform

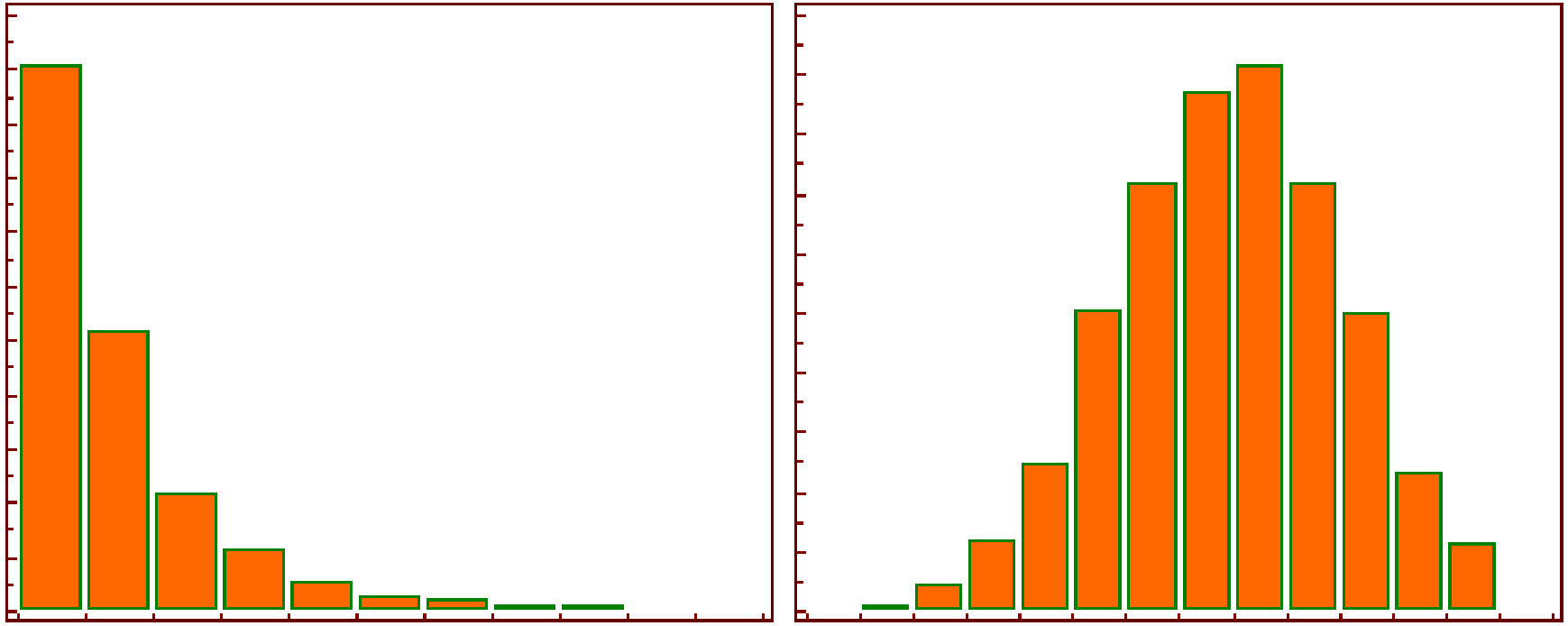
New values of transformed data



Base,
we will use 10

https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9

Data transformation:
Logarithmic Transformation



https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9

Practical part