Data Exploration
- Standardization & Normalization

Dept. of Mechanical System Design Engineering,
Seoul National University of Science and Technology

Prof. Ju Yeon Lee
(jylee@seoultech.ac.kr)

Review

Descriptive Statistics

Index	mpg	cylinders	displacement	weight	acceleration	model_year	origin
count	398	398	398	398	398	398	398
mean	23.5146	5.45477	193.426	2970.42	15.5681	76.0101	1.57286
std	7.81598	1.701	104.27	846.842	2.75769	3.69763	0.802055
min	9	3	68	1613	8	70	1
25%	17.5	4	104.25	2223.75	13.825	73	1
50%	23	4	148.5	2803.5	15.5	76	1
75%	29	8	262	3608	17.175	79	2
max	46.6	8	455	5140	24.8	82	3

- **Count : the number of available data**
- **Mean : arithmetic mean value**
- **Min : minimum value**
- **Max : maximum value**
- **Q1 : ~25%**
- **Q2 : ~50% (median)**
- **Q3 : ~75%**
- **Q4 : ~max**
- **Mode: most frequent value**
- **Std : standard deviation**
- **Min – Max : a range of values**

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

σ = population standard deviation

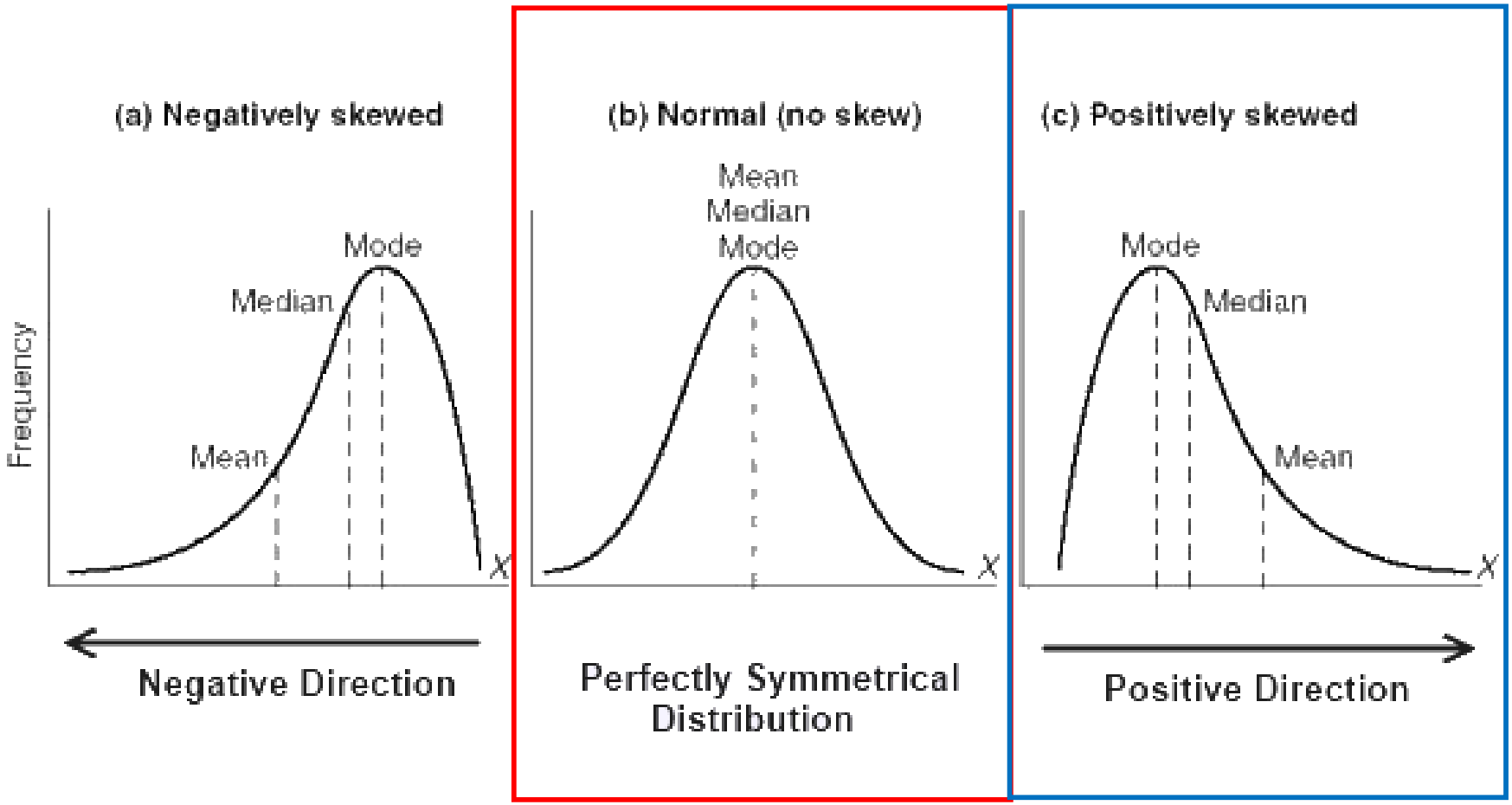
N = the size of the population

x_i = each value from the population

μ = the population mean

Skewness

Mean = Median = Mode



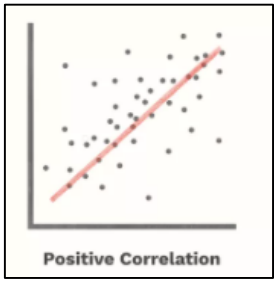
Auto MPG Dataset

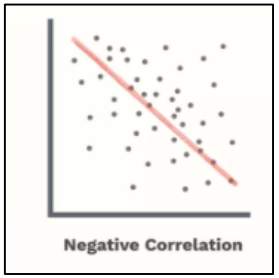
Correlation Analysis

Index	mpg	cylinders	displacement	weight	acceleration	model_year	origin
mpg	1	-0.775396	-0.804203	-0.831741	0.420289	0.579267	0.56345
cylinders	-0.775396	1	0.950721	0.896017	-0.505419	-0.348746	-0.562543
displacement	-0.804203	0.950721	1	0.932824	-0.543684	-0.370164	-0.609409
weight	-0.831741	0.896017	0.932824	1	-0.417457	-0.306564	-0.581024
acceleration	0.420289	-0.505419	-0.543684	-0.417457	1	0.288137	0.205873
model_year	0.579267	-0.348746	-0.370164	-0.306564	0.288137	1	0.180662
origin	0.56345	-0.562543	-0.609409	-0.581024	0.205873	0.180662	1

- **Positive correlation :**
  the variables move in the same direction

- **Negative correlation :**
  the variables move in opposite directions





- **No correlation**

Standardization & Normalization

Standardization

- **Standardization (표준화, Z-score Normalization)**
  - ✓ 입력 변수(X)의 정규 분포를 평균이 0이고 표준 편차가 1인 표준 정규 분포로 재조정

$$\mu = 0, \sigma = 1$$

$$Z-score = \frac{x - \mu}{\sigma}$$

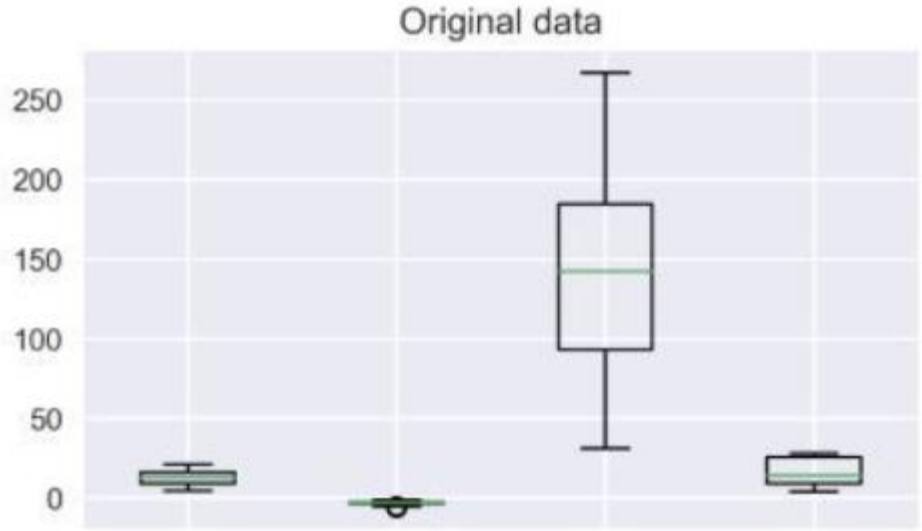
  - ✓ Z-score : 특정 데이터가 평균에서 멀리 떨어진 정도 → outliers (이상치)

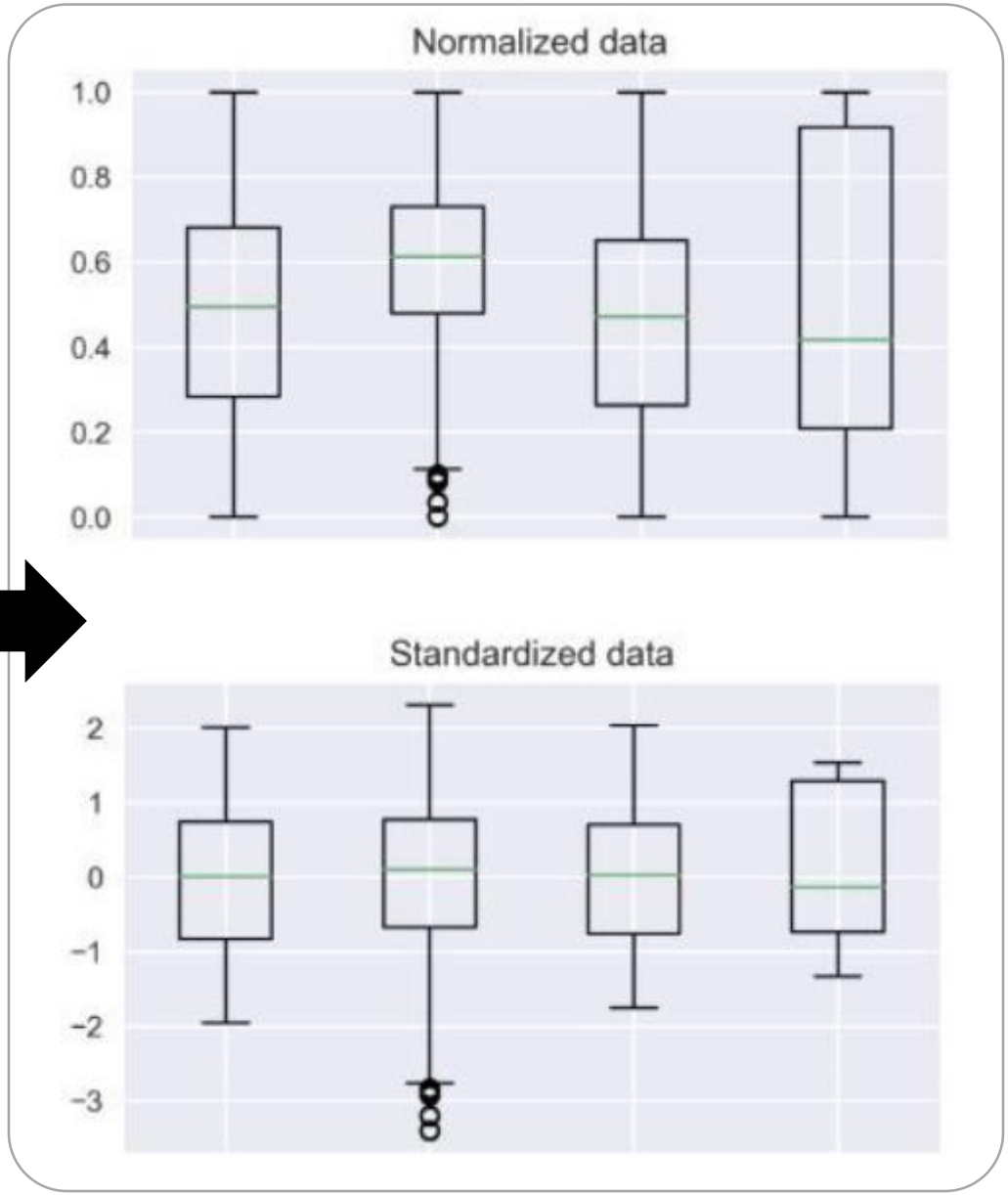
Normalization

- **Normalization (정규화)**
  - ✓ **모든 입력 변수를 0과 1 사이의 값으로 변환**
  - ✓ **Min-Max Scaling**

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Standardization & Normalization





Source: https://bskyvision.com/849


Thank you
Q & A