Data Pre-processing

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

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

Review

Standardization

- Standardization (표준화, Z-score Normalization)

✓ 입력 변수(X)의 정규 분포를 평균이 0이고 표준 편차가 1인 표준 정규 분포로 재조정

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

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

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

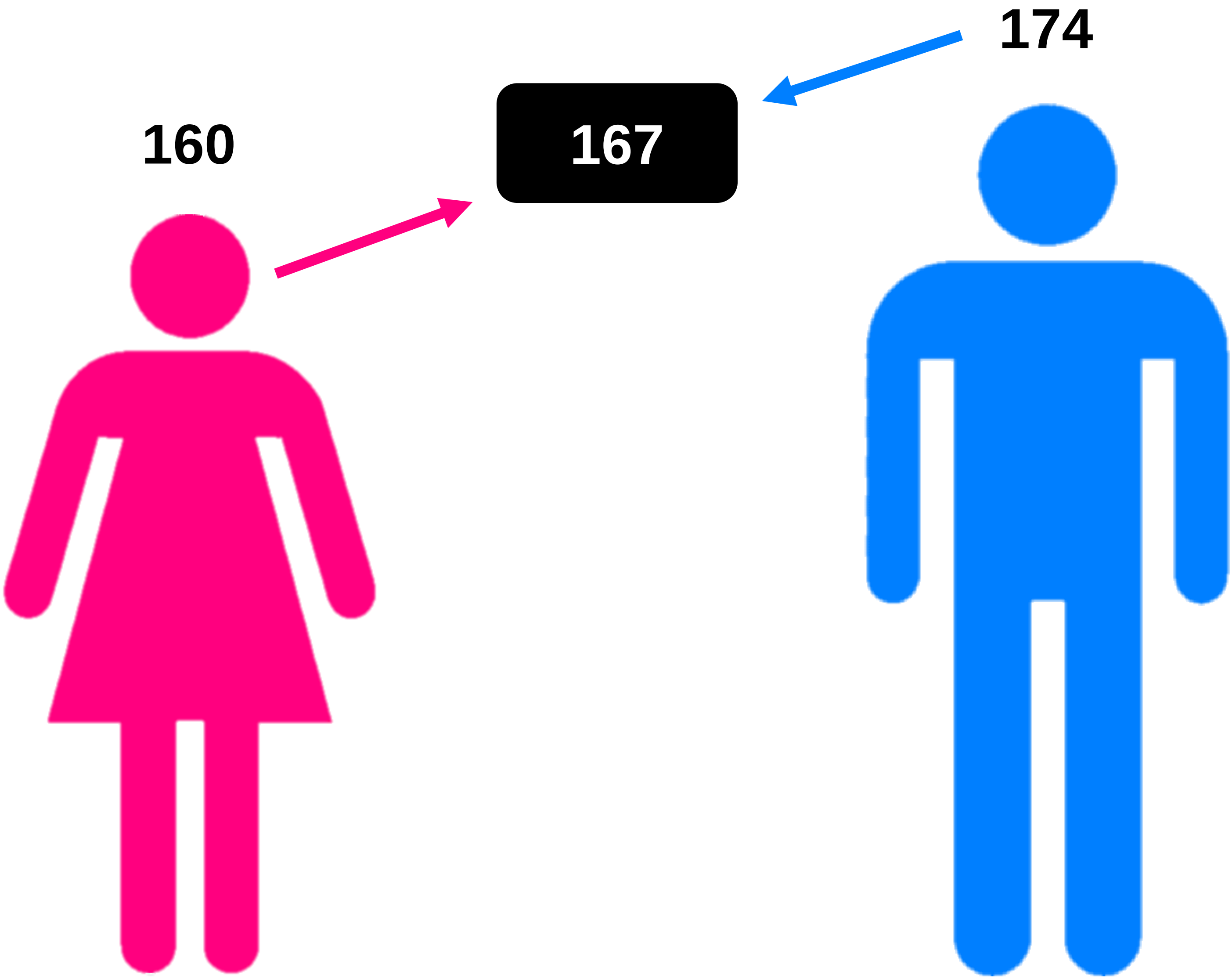
Normalization

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

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

Data Pre-processing

Reflecting the Characteristics of Each Group



174

160

167

lambda

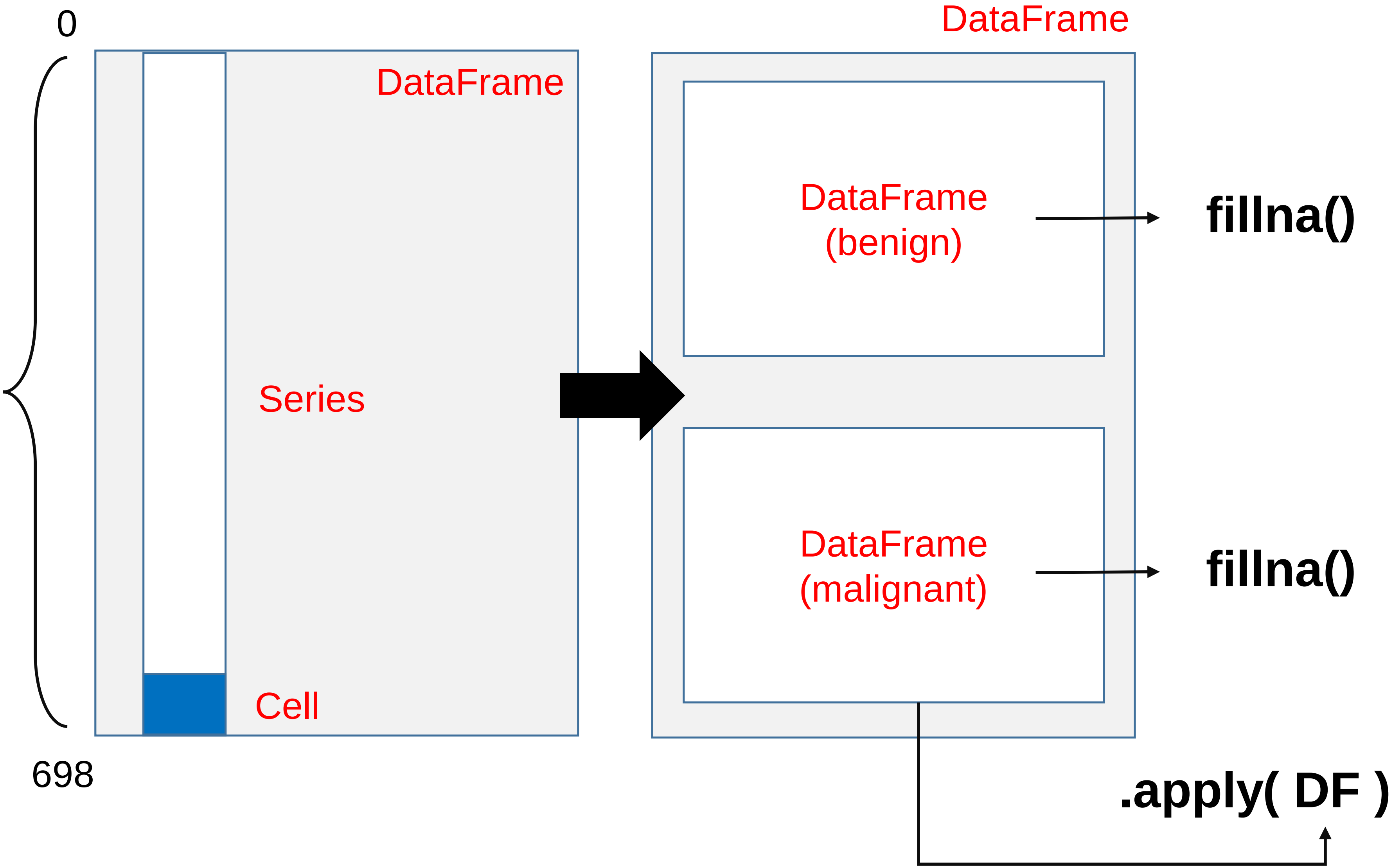
lambda parameters: indented statement

```
def test_ex(x, y):
        return x + y


test_ex(10, 20)
```

```
(lambda x, y: x + y)(10, 20)
```

groupby()


0

DataFrame

Series

Cell

698

DataFrame

DataFrame (benign)

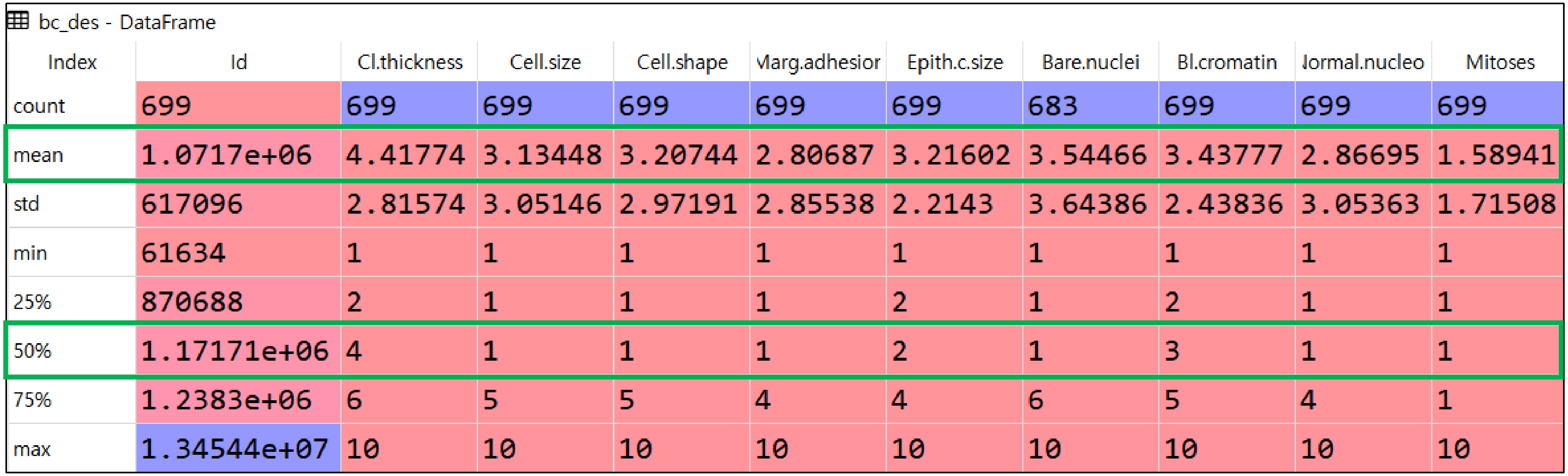
fillna()

DataFrame (malignant)

fillna()

.apply(DF)

Descriptive Statistics



bc_des - DataFrame

Index	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesior	Epith.c.size	Bare.nuclei	Bl.cromatin	Iormal.nucleo	Mitoses
count	699	699	699	699	699	699	683	699	699	699
mean	1.0717e+06	4.41774	3.13448	3.20744	2.80687	3.21602	3.54466	3.43777	2.86695	1.58941
std	617096	2.81574	3.05146	2.97191	2.85538	2.2143	3.64386	2.43836	3.05363	1.71508
min	61634	1	1	1	1	1	1	1	1	1
25%	870688	2	1	1	1	2	1	2	1	1
50%	1.17171e+06	4	1	1	1	2	1	3	1	1
75%	1.2383e+06	6	5	5	4	4	6	5	4	1
max	1.34544e+07	10	10	10	10	10	10	10	10	10

- **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

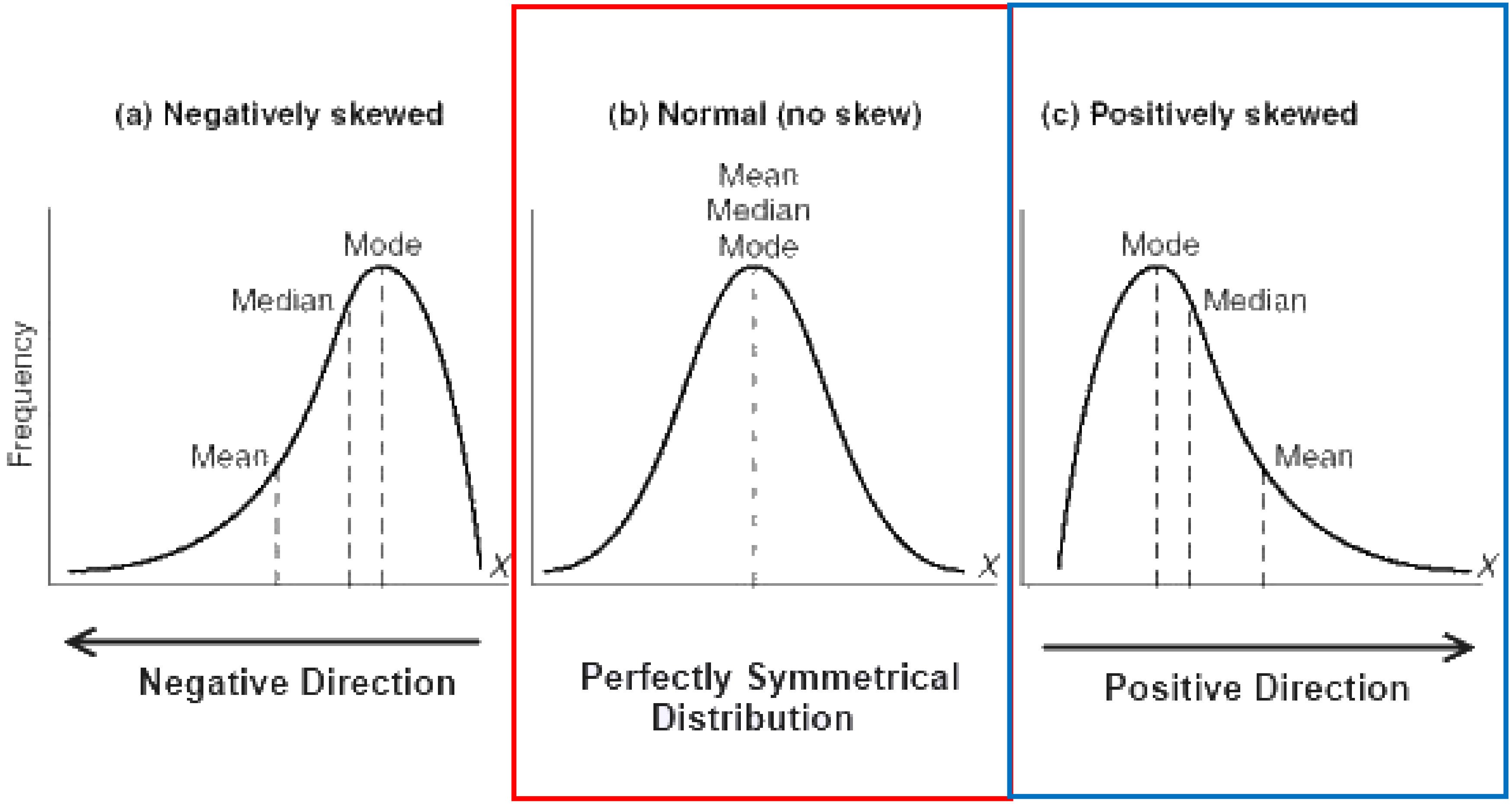
N = the size of the population

x_i = each value from the population

μ = the population mean

Skewness

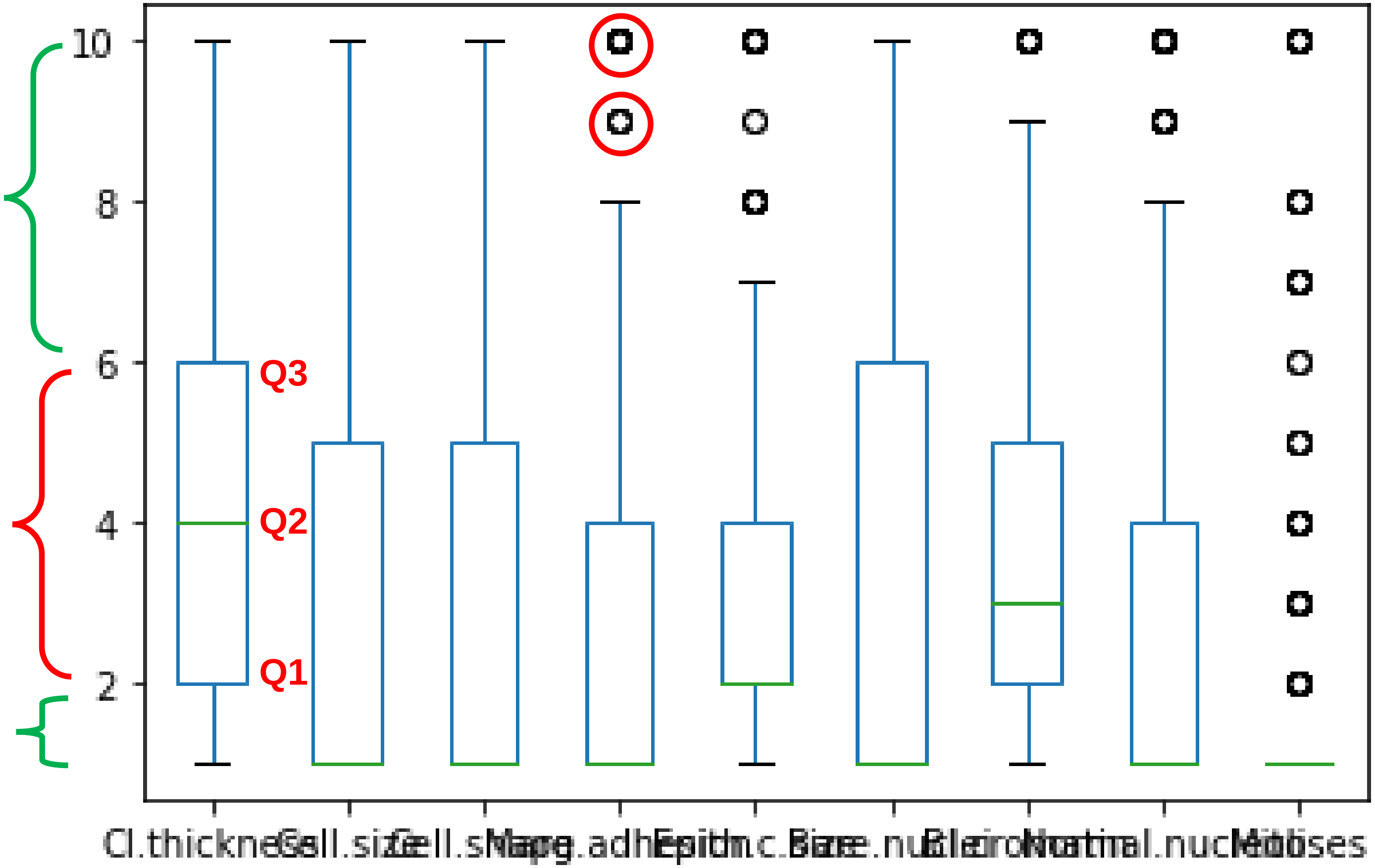
Mean = Median = Mode



BreastCancer Dataset

Boxplot

Outliers



Q3+(1.5*IQR)

Q3

Q2

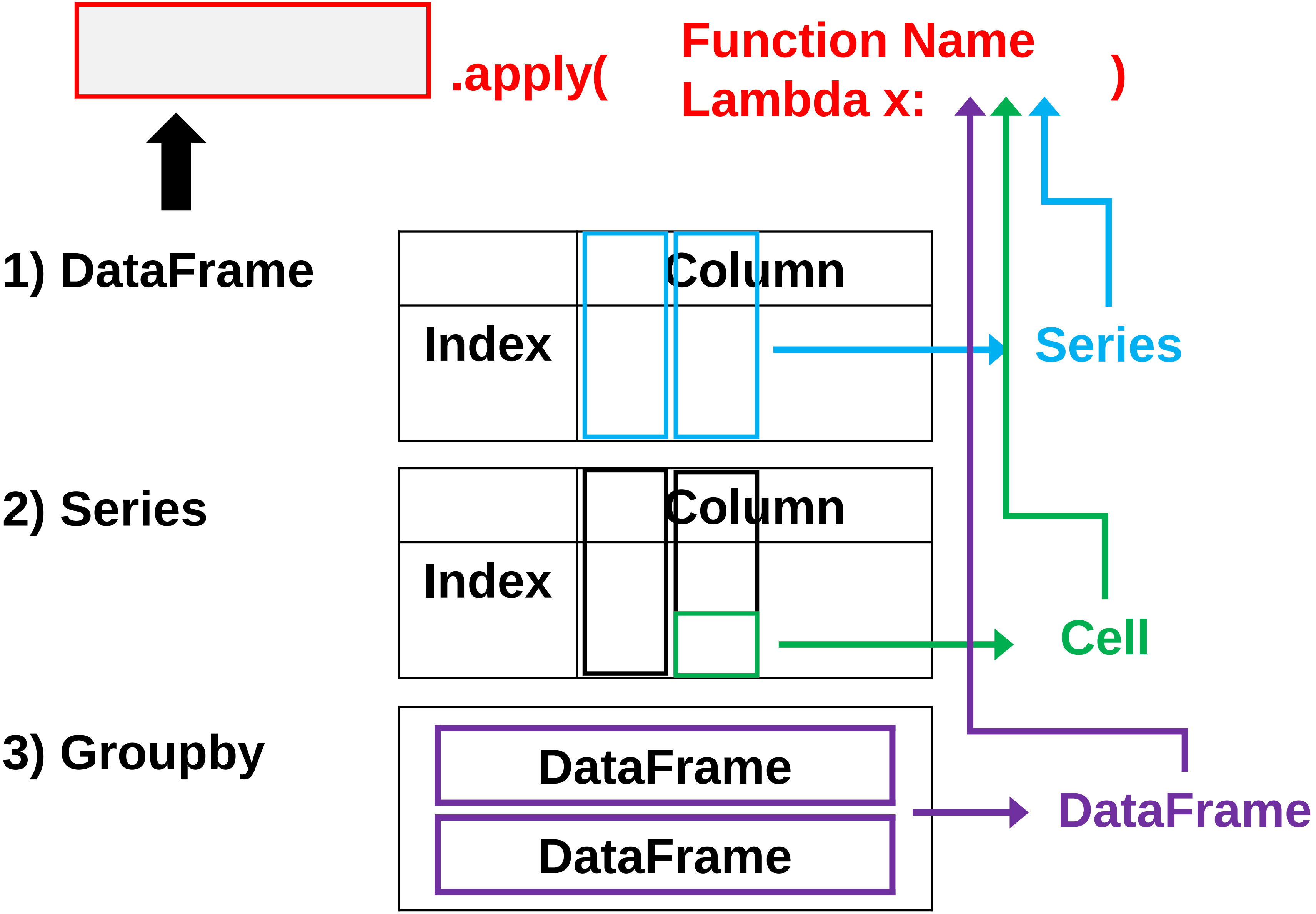
Q1

IQR :

(50%) InterQuartile Range

Q1-(1.5*IQR)

Apply()



.apply(Function Name
Lambda x:)

1) DataFrame

Column

Index

Series

2) Series

Column

Index

Cell

3) Groupby

DataFrame

DataFrame

DataFrame

Thank you

Q & A