빅데이터 특강

기초통계 관점으로 보는 데이터분석의 기초

신지은 (서울시립대학교 통계학과)

목차

- 실습자료 다운로드: https://github.com/jiieunshin/high-univ
- 문의메일: jieunstat@gmail.com

1. 데이터의 이해

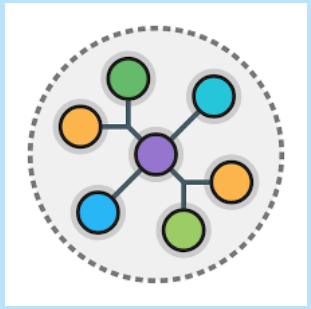
데이터의 분류

데이터

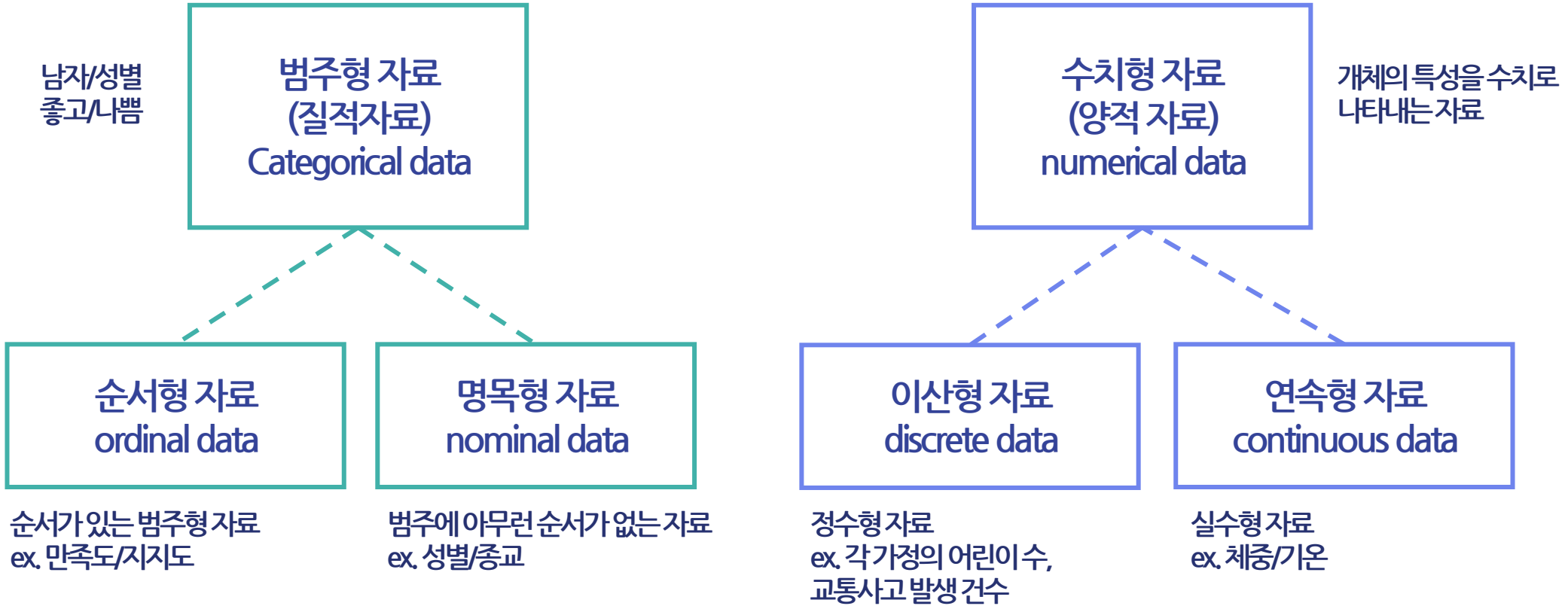
정형 데이터

id	이름	나이	성별
01	Kim	32	M
02	Lee	26	F
03	Park	72	F
04	Choi	15	M

비정형 데이터

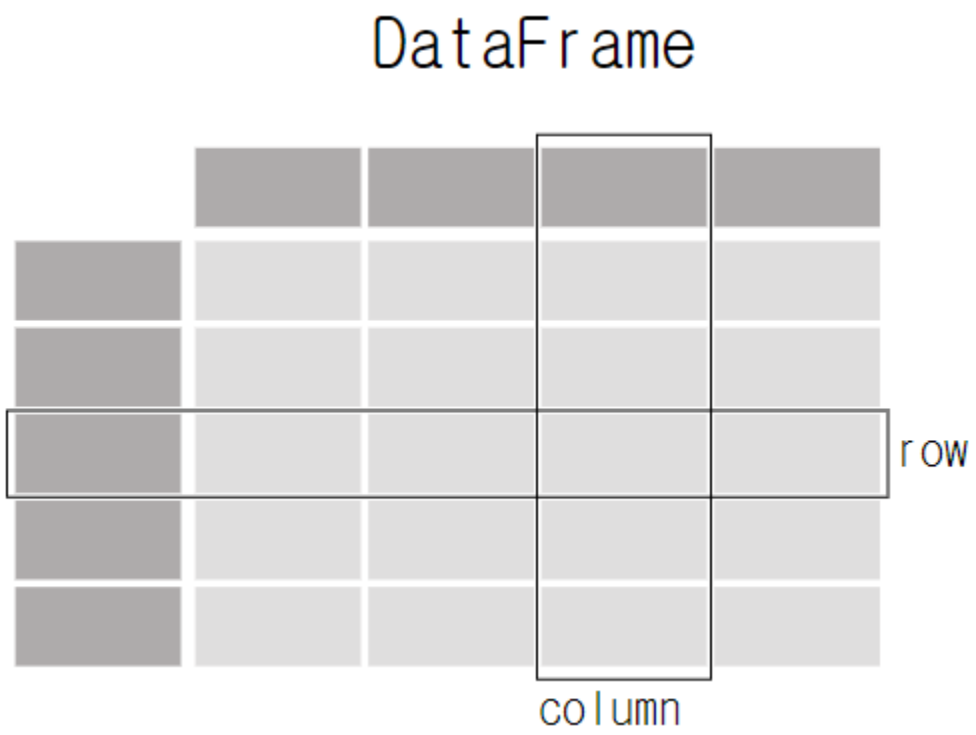


정형 데이터의 분류


남자/성별
좋고/나쁨
범주형 자료
(질적자료)
Categorical data
순서형 자료
ordinal data
명목형 자료
nominal data
순서가 있는 범주형 자료
ex. 만족도/지지도
범주에 아무런 순서가 없는 자료
ex. 성별/종교
수치형 자료
(양적 자료)
numerical data
개체의 특성을 수치로
나타내는 자료
이산형 자료
discrete data
연속형 자료
continuous data
정수형 자료
ex. 각 가정의 어린이 수,
교통사고 발생 건수
실수형 자료
ex. 체중/기온

1. 데이터의 이해

파이썬 데이터프레임



화재발생 데이터.csv

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

1. 데이터의 이해

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

정형 데이터의 집계 방법: 기초 통계량

셈 척도 갯수 (count), 합산 (sum)

중심척도 평균 (mean), 중위수 (median)

산포척도 최댓값 (max), 최솟값 (min), 분산 (variance), 표준편차 (standard deviation), 백분위 (quantile)

범주형 자료

(셈 척도) count, percent

수치형 자료

(셈 척도) sum
중심척도 모두
산포척도 모두

화재발생연도, 시군구

사망자수, 부상자수,
재산피해금액, 출동횟수

1. 데이터의 이해

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	57	42
17	2021	종로구	0.0	12	465499	192	48	54

집계를 위한 문제와 설계

문제 시군구별 평균 재산피해금액과 총 출동횟수

설계

- 그룹화: 시군구
- 계산하고 싶은 열: 재산피해금액, 출동횟수
- 집계함수: sum, mean

파이썬 구현 예시

```
df5.groupby(['시군구']).agg({"재산피해금액" : "mean", "출동횟수" : "sum"})
```

시군구	재산피해금액	출동횟수
은평구	886196.4	921
종로구	711661.8	1129
중구	15976858.0	1042
중랑구	286252.0	1098

2. 데이터 분석 방법

2. 데이터 분석 방법

변수

	화재발생연도	시군구	사망자수	부상자수	재산피해금액	출동횟수	출동횟수_겨울	출동횟수_여름
0	2017	은평구	0.0	3	218200	159	51	32
1	2017	종로구	1.0	3	1077665	234	55	69
2	2017	중구	5.0	14	485392	198	48	47
3	2017	중랑구	2.0	5	332366	196	53	38
4	2018	은평구	5.0	10	419503	214	58	47
5	2018	종로구	14.0	22	574300	254	71	70
6	2018	중구	0.0	23	1257005	275	76	74
7	2018	중랑구	2.0	8	201421	254	72	55
8	2019	은평구	3.0	20	2412769	196	62	34
9	2019	종로구	4.0	16	801094	232	60	63
10	2019	중구	3.0	17	74077097	213	51	39
11	2019	중랑구	1.0	9	322650	210	54	49
12	2020	은평구	2.0	6	504788	192	48	46
13	2020	종로구	2.0	5	639751	217	50	49
14	2020	중구	0.0	10	1284422	185	41	54
15	2020	중랑구	2.0	12	229566	225	54	57
16	2021	은평구	3.0	8	875722	160	37	42
17	2021	종로구	0.0	12	465499	192	48	54

표

그래프

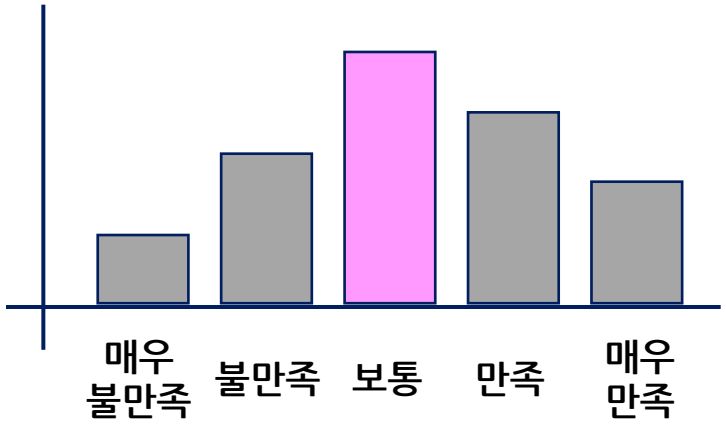
- ✓ 하나의 변수에서?
- ✓ 두 변수의 조합에서?

- ✓ 수치형 변수에서?
- ✓ 범주형 변수에서?

정형 데이터의 시각화

범주형 자료

막대그래프

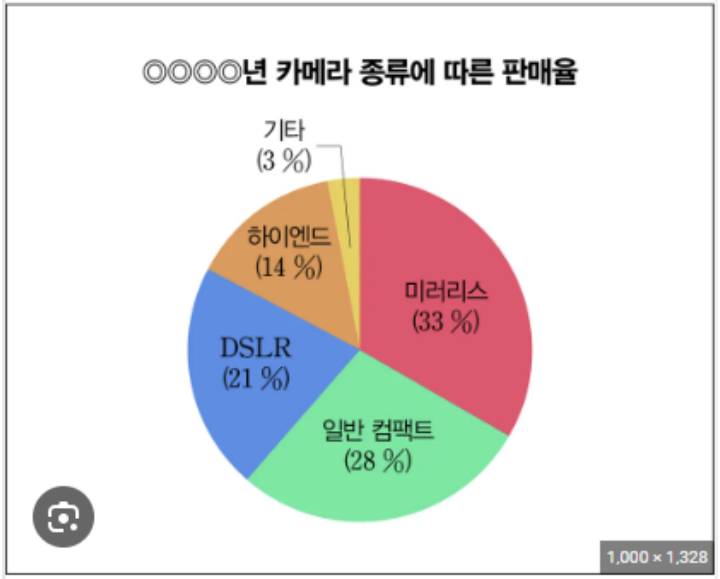


원 그래프

◎◎◎◎년 카메라 종류에 따른 판매율

(단위: %)

종류	미러리스	일반 컴팩트	DSLR	하이엔드	기타
판매율	33	28	21	14	3



분할표

교육수준	결혼생활		
	빈약	원만	대단히 양호
대학	72	112	245
고등학교	65	90	120
중학교	95	103	98

[표] 교육수준과 결혼생활

정형 데이터의 시각화

범주형 자료

정형 데이터의 시각화

수치형 자료

➢ 도수분포표와 히스토그램

아래의 수학 점수를 도수분포표로 나타내보자

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75

1. 관측치의 최댓값과 최솟값의 차이, 즉 범위를 구한다.
   ⇒ 85-7 = 78

2. 구간을 몇 개로 나눌 것인가?
   ⇒ 10개

3. 구간 폭을 정하자
   ⇒ 구간 폭 = (최댓값 - 최솟값) / 구간수 = 78 / 10 = 7.8

4. 도수와 상대도수, 누적도수, 누적상대도수 등을 산출한다.

정형 데이터의 시각화

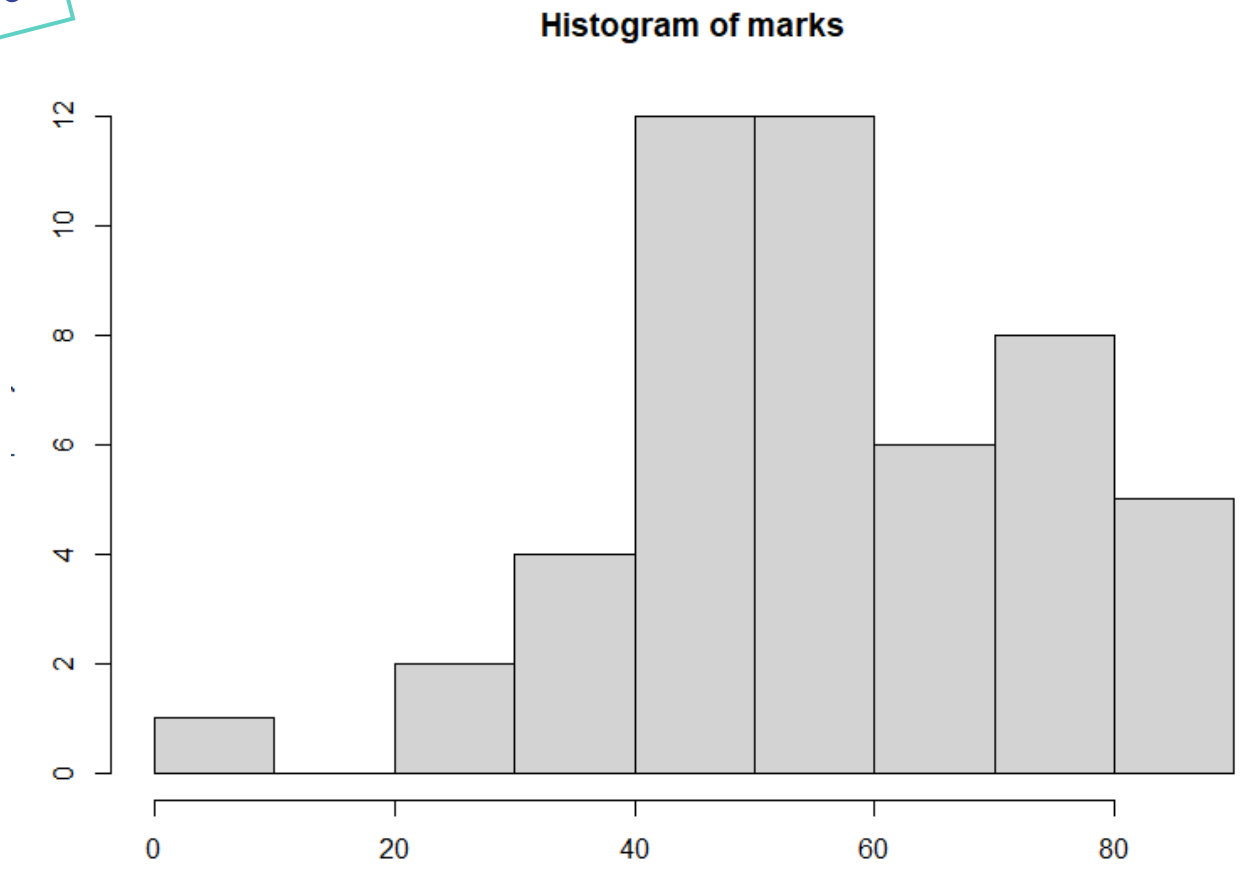
수치형 자료

➢ 도수분포표와 히스토그램

점수	학생 수 (명)
(0, 10]	1
(10, 20]	0
(20, 30]	2
(30, 40]	4
(40, 50]	12
(50, 60]	12
(60, 70]	7
(70, 80]	9
(80, 90]	5
(90, 100]	0
계	50

도수분포표

구간: 10개,
구간 폭: 10



히스토그램

정형 데이터의 시각화

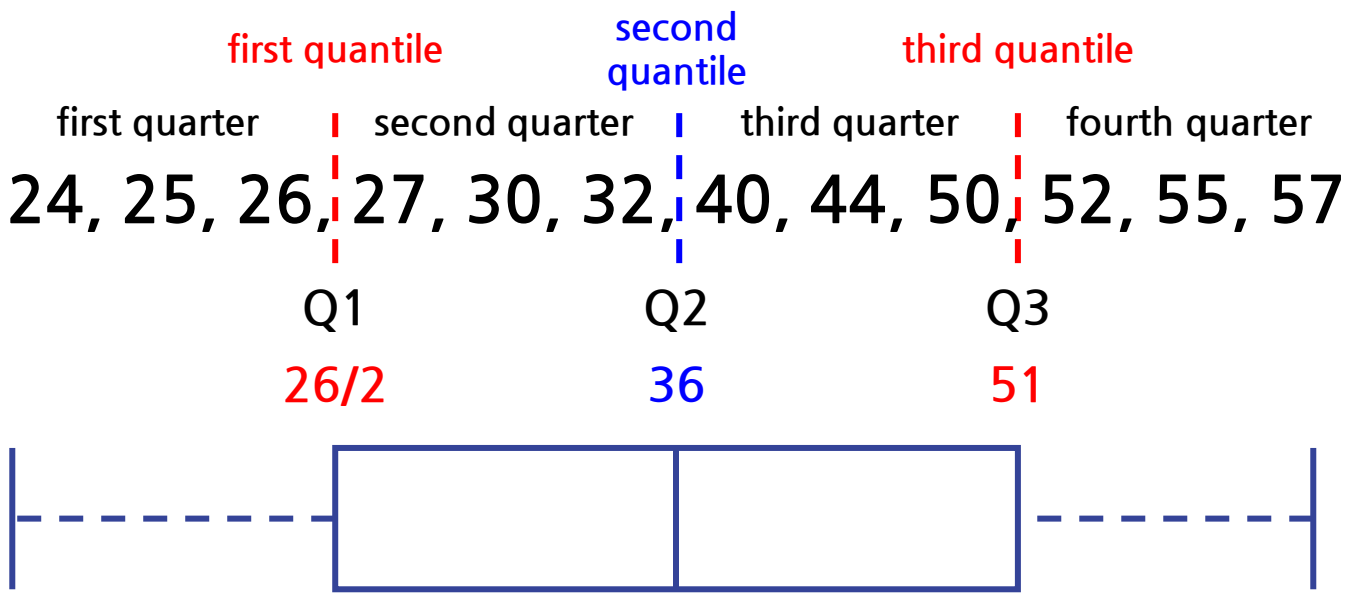
수치형 자료

➢ 평균과 중위수

두 가지 자료 (0, 1, 2, 2, 2, 3, 4)와 (70, 1, 2, 2, 2, 3, 4)의 평균, 중앙값을 비교해보자

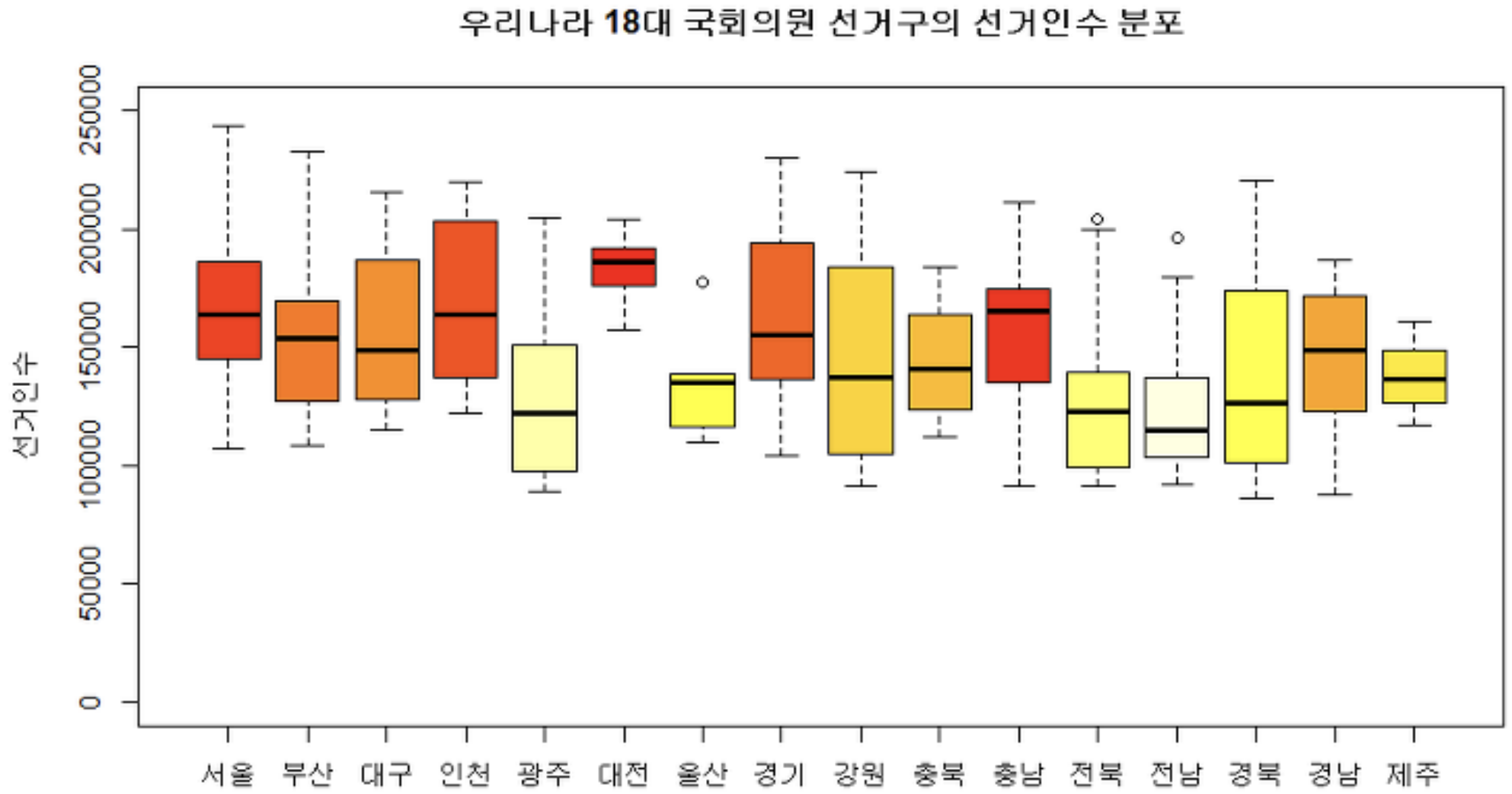
0, 1, 2, 2, 2, 3, 4 → 평균 2, 중앙값 2
70, 1, 2, 2, 2, 3, 4 → 평균 12, 중앙값 2

➢ 백분위수와 상자그림



정형 데이터의 시각화

수치형 자료



정형 데이터의 시각화

수치형 자료

➢ 분산과 표준편차

- 분산 (variance)

: 각 자료값들과 평균과의 차이 $x_i - \bar{x}$로 산포를 나타낸다. 즉, 평균으로부터 멀리 떨어져 있을수록 $x_i - \bar{x}$의 절댓값이 커짐.

표본분산 s^2 은 다음과 같은 식으로 구한다.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 표준편차 (s.d., standard deviation)

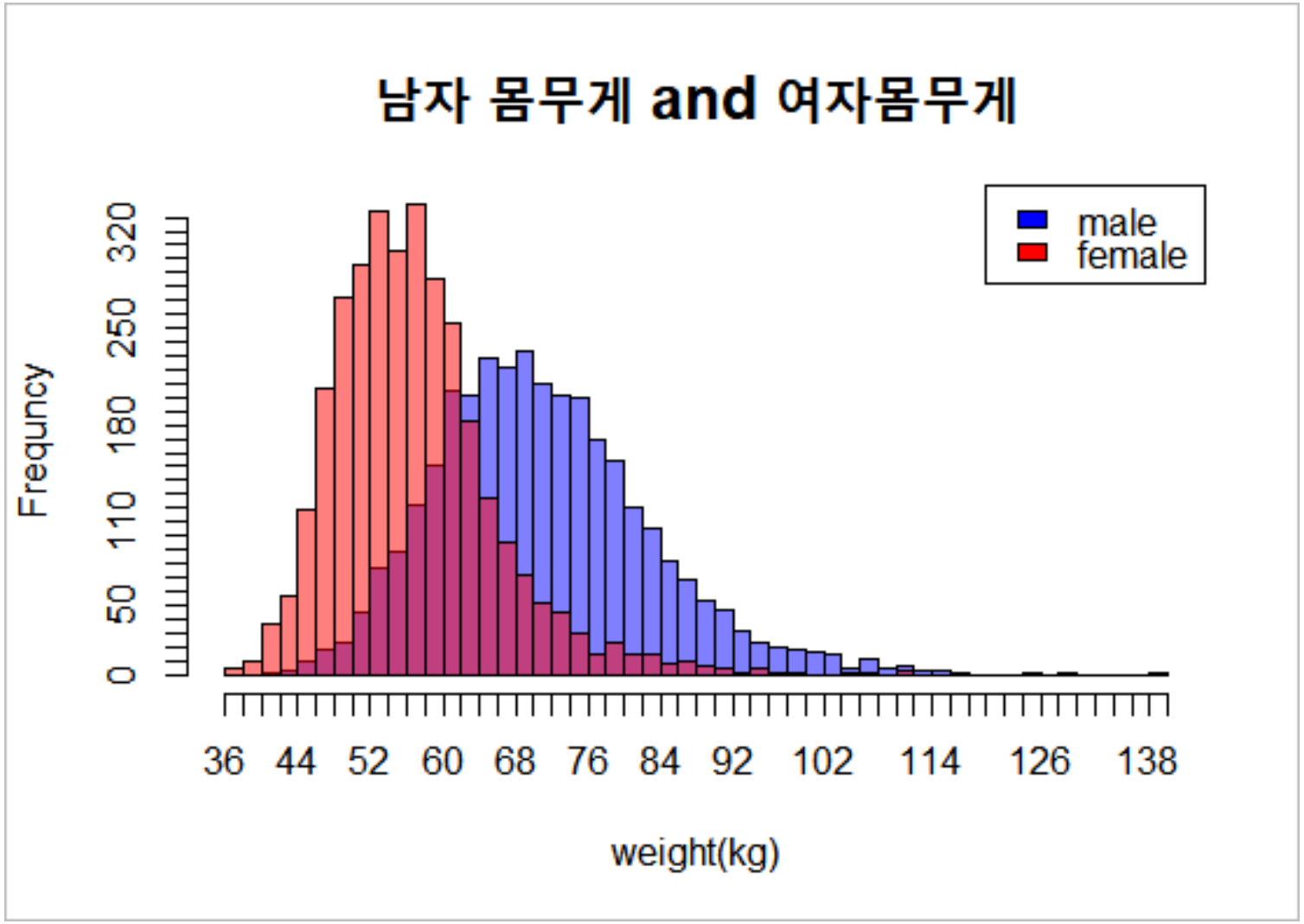
: 분산의 제곱근. 분산을 구할 때 제곱을 취함으로써 원래 자료값의 단위가 달라진 것을 복구한 것이다.

표본표준편차 s은 다음과 같은 식으로 구한다.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

정형 데이터의 시각화

수치형 자료



이변량 데이터

Female과 Male을 동시에 분석할 수는 없을까?

Female	Male
7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51, 55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85	48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86, 42, 51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58, 64, 64, 75

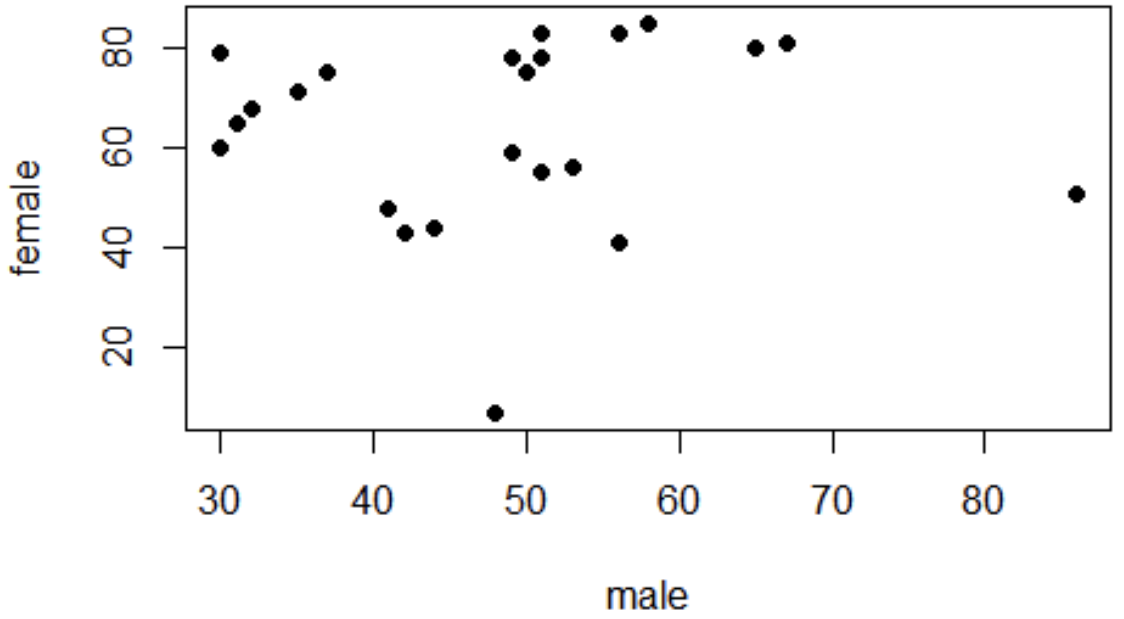
테이블을 재구성하자

obs	Female	Male
1	7	48
2	59	49
3	78	49
4	79	30
5	60	30
6	65	31
…	…	…

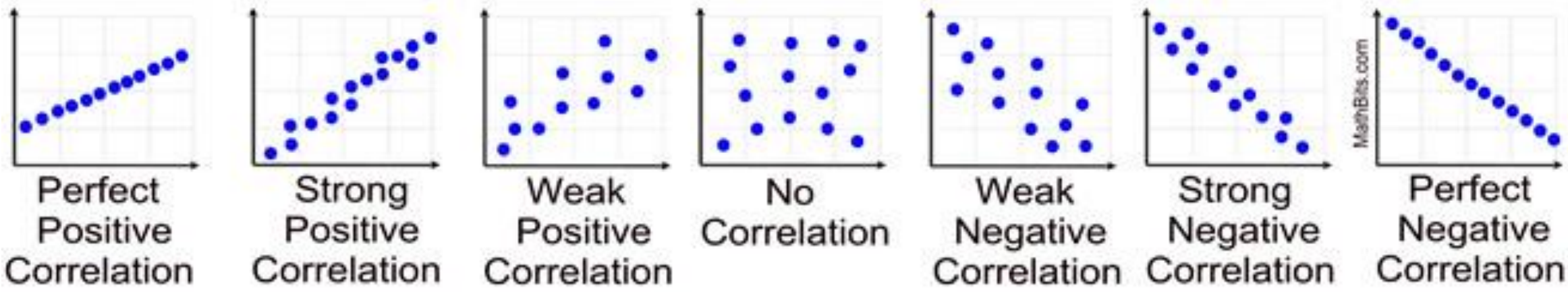
이변량 데이터의 시각화: 산점도

```
female = c(7, 59, 78, 79, 60, 65, 68, 71, 75, 48, 51,
           55, 56, 41, 43, 44, 75, 78, 80, 81, 83, 83, 85)
male = c(48, 49, 49, 30, 30, 31, 32, 35, 37, 41, 86,
         51, 53, 56, 42, 44, 50, 51, 65, 67, 51, 56, 58)

plot(male, female, pch = 16)
```



두 변수간 관련성이 있는가?



➢ 상관계수

- 변수 간의 관계의 강함을 보는 척도
- $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$을 얻어진 표본 (2변량 자료)이라 하자. $\bar{x}$와 $\bar{y}$를 각각 x와 y의 표본평균으로 하였을 때

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

x와 y 표본 공분산(sample covariance)이라고 한다.

- 또 S_x^2와 S_y^2을 각각 x와 y의 표본분산이라고 하면

$$\gamma = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}}$$

를 표본상관계수 (sample correlation coefficient)라고 한다.

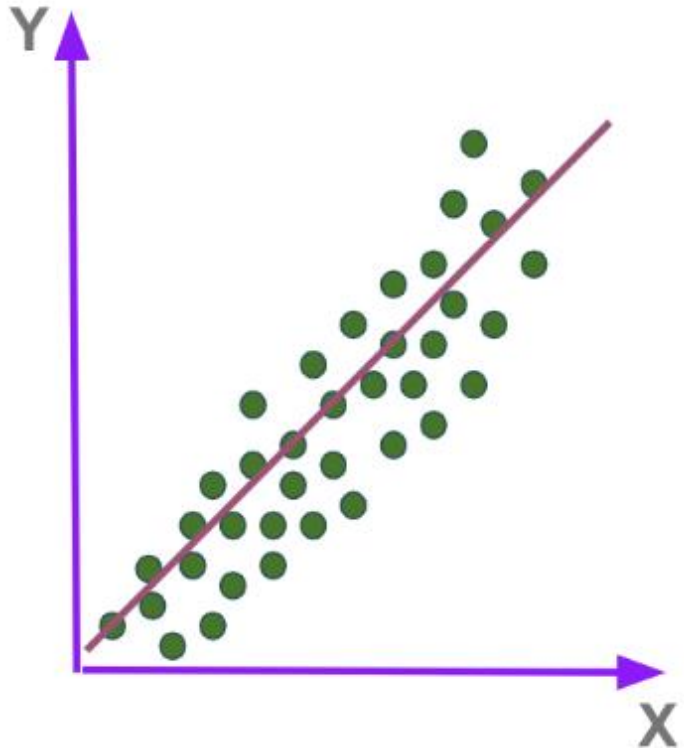
정형 데이터 분석방법

분석 목적에 따라

분석 방법이 다르다

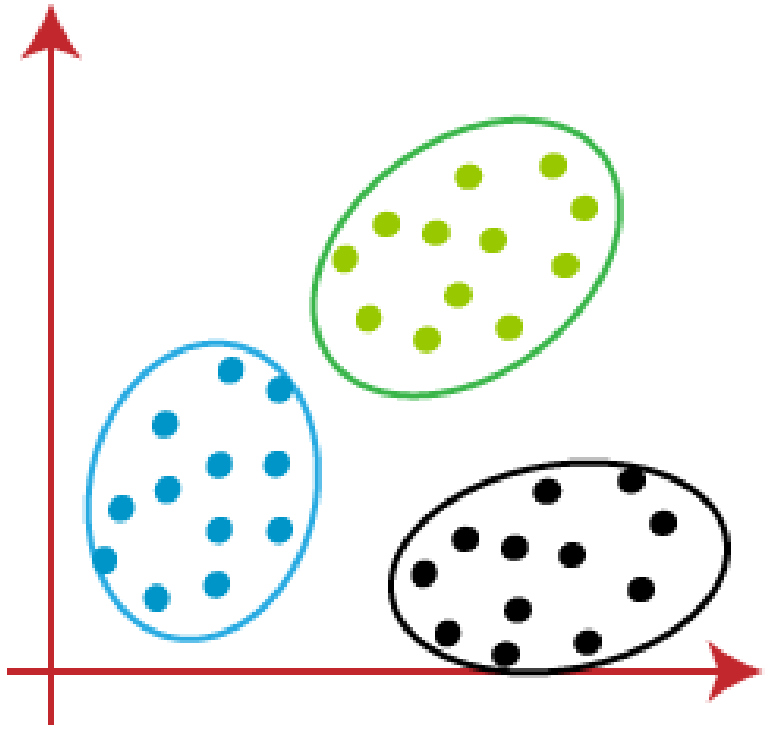
회귀
(regression)

회귀분석



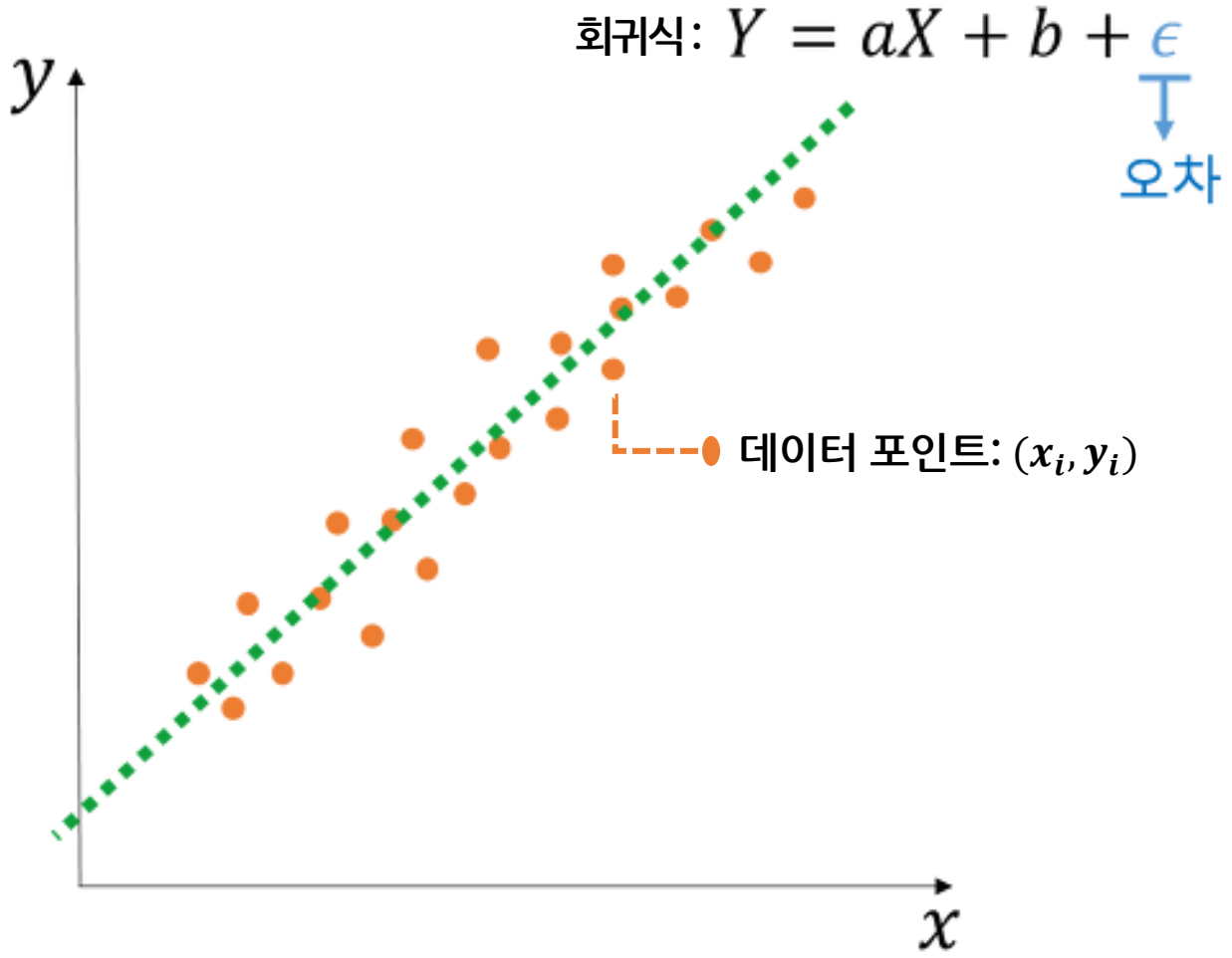
군집
(clustering)

K-means clustering



회귀분석

데이터를 가장 잘 설명하는 선을 찾는 방법



K-means clustering

비슷한 특성을 갖는 데이터를 묶는 방법