빅데이터 특강

신지은 (서울시립대학교 통계학과)

목차

#1일차

데이터의 이해

통계학이란?

R 프로그램 소개 및 설치

#2일차

기초통계량

R 실습

#3일차

시각화

R 실습

#4일차

예측

R 실습

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

들어가며

통계학이란?

2024학년도 4차산업 관련 학과 입시 현황

인공지능(AI) & 빅데이터학과, 수도권 대학 1,887명 모집

1. 연도별 주요 10개대 대학별 정시 합격점수 상위 3위권 이내 학과 분석

* 분석대상 : 서울대, 연세대, 고려대, 성균관대, 서강대, 한양대, 중앙대, 경희대, 이화여대, 서울시립대 등 10개대 정시 합격점수 상위 3위 이내 진입학과
* 어디가 대학 발표 기준(국수탐백분위 평균 70% 컷 기준)
* 의약학계열 제외
* 일반전형 기준(고른기회 등 특별전형 제외), 서울대 지역균형, 고려대 교과우수전형 제외

2) 자연(의약학계열 제외)

2021			2022			2023			2024학년도		
학과분류	학과수	비율	학과분류	학과수	비율	학과분류	학과수	비율	학과분류	학과수	비율
전자전기	5	13.9%	컴퓨터	8	24.2%	컴퓨터	5	15.6%	AI	5	12.8%
컴퓨터	4	11.1%	화학공	6	18.2%	반도체	5	15.6%	반도체	5	12.8%
의생명	3	8.3%	전자전기	3	9.1%	AI	4	12.5%	컴퓨터	4	10.3%
기계	3	8.3%	반도체	2	6.1%	화학공	4	12.5%	전자전기	3	7.7%
교육	3	8.3%	AI	2	6.1%	전자전기	3	9.4%	화학	3	7.7%
반도체	2	5.6%	통계	1	3.0%	통계	2	6.3%	화학공	3	7.7%
통계	2	5.6%	의생명	1	3.0%	의생명	1	3.1%	도시공	2	5.1%
생명	2	5.6%	기계	1	3.0%	생명	1	3.1%	생명	2	5.1%
AI	1	2.8%	교육	1	3.0%	디스플레이	1	3.1%	의생명	2	5.1%
디스플레이	1	2.8%	생명	1	3.0%	자유전공	1	3.1%	자동차	2	5.1%
자유전공	1	2.8%	디스플레이	1	3.0%	수학	1	3.1%	제약	2	5.1%
수학	1	2.8%	자유전공	1	3.0%	자동차	1	3.1%	수학	1	2.6%
화학	1	2.8%	수학	1	3.0%	간호	1	3.1%	건축	1	2.6%
공대	1	2.8%	화학	1	3.0%	의류	1	3.1%	뇌인지	1	2.6%
자동차	1	2.8%	공대	1	3.0%	통신	1	3.1%	디스플레이	1	2.6%
간호	1	2.8%	에너지	1	3.0%				바이오	1	2.6%
물리	1	2.8%	도시공	1	3.0%				자유전공	1	2.6%
융합	1	2.8%									
뇌인지	1	2.8%									
과학	1	2.8%									
총합계	36	100.0%	총합계	33	100.0%	총합계	32	100.0%	총합계	39	100.0%

컴퓨터 공학 → 통계 → 인공지능& 데이터사이언스

<표6> 2024학년도 컴퓨터공학(소프트웨어)과 &

대학	컴퓨터공학(소프트웨어)		인공지능 관련
가톨릭대	컴퓨터정보공학부		인공지능학과
경기대	컴퓨터공학전공		인공지능전공
경희대	컴퓨터공학과	소프트웨어융합대학	인공지능학과
국민대	소프트웨어학부		인공지능학부
동국대	컴퓨터공학		AI융합학부(인문/자연)
동덕여대	컴퓨터학과		HCI사이언스전공
삼육대	컴퓨터공학부		인공지능융합학부
상명대	컴퓨터과학전공		휴먼지능공학전공
서울과기대	컴퓨터공학과		인공지능응용학과
서울시립대	전자전기컴퓨터공학부	컴퓨터과학부	인공지능학과
성균관대	소프트웨어학		글로벌융합학부
성신여대	컴퓨터공학과		AI융합학부
세종대	컴퓨터공학과	소프트웨어학과	인공지능
숭실대	컴퓨터학부	소프트웨어학부	AI융합학부
연세대	컴퓨터과학과		인공지능학과
이화여대	컴퓨터공학전공(자연/인문)		인공지능전공(인문/자연)
인하대	컴퓨터공학과		인공지능학과
중앙대	소프트웨어학부		AI학과

<표7> 2024학년도 통계학과 & 빅데이터학과 동시 개설 대학

대학	통계학과	빅데이터 관련
고려대	통계학과(인문)	데이터과학과
동덕여대	정보통계학과	데이터사이언스전공
세종대	수리통계학부	데이터사이언스학과
이화여대	통계학과	데이터사이언스학과
인하대	통계학과	데이터사이언스학과

들어가며

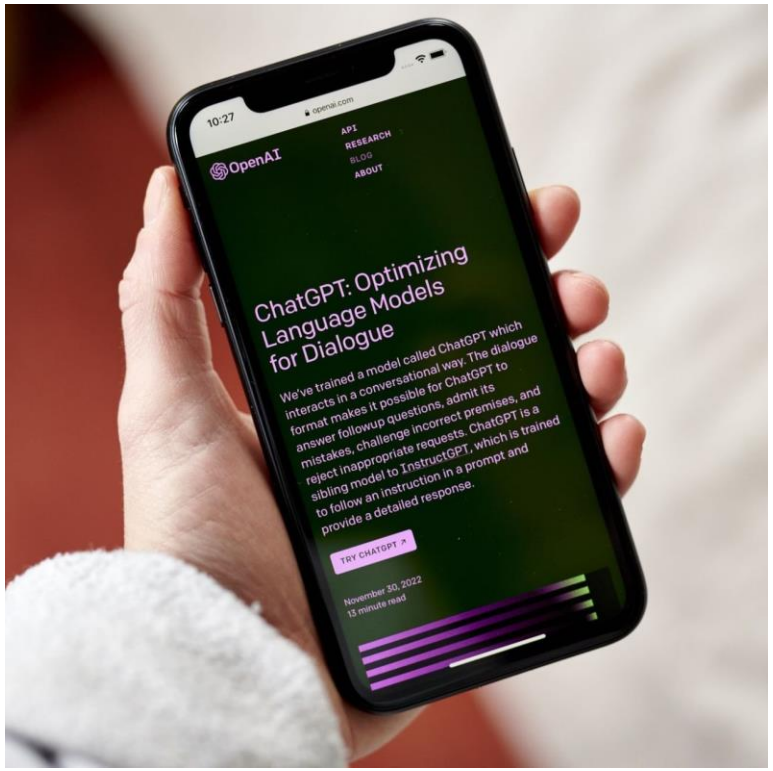
통계학이란?

정보수집 (데이터) → **데이터 탐색** → **학습 (모델링)** → **사용자 (자동화 및 배포)**

- ✓ 분석 목적 세우기
- ✓ 데이터 수집

- ✓ 요약 및 시각화
- ✓ 데이터 정제(이상치/ 결측값 처리)

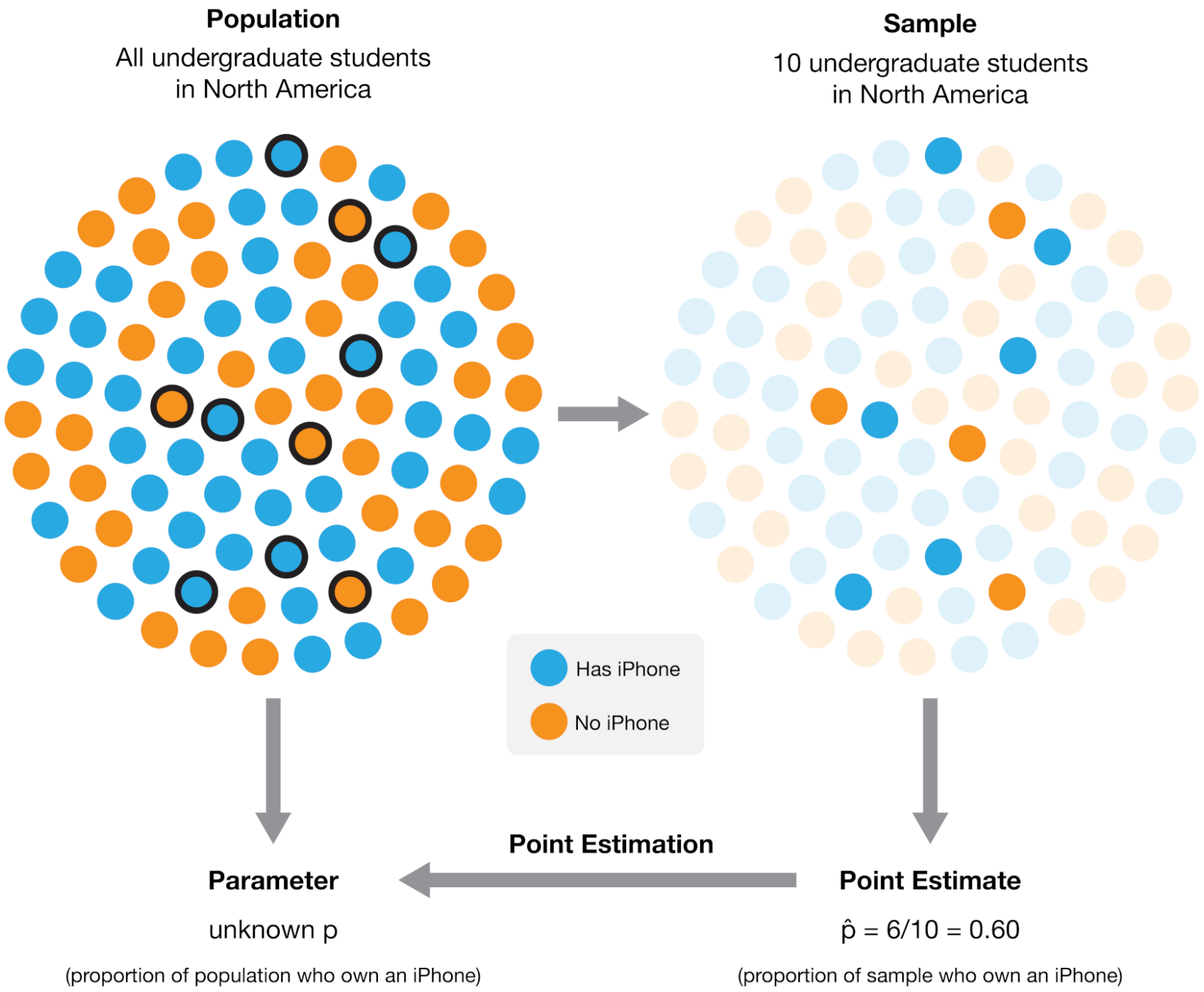
- ✓ 모형 선정
- ✓ 최적의 모형 학습



컴퓨터 공학	개발 환경 구축에 중점
통계학과	정형 데이터 분석 통계적 모형의 수리적 이해, 결과 해석에 중점
인공지능	서비스 제공에 중점, 딥러닝/생성모형
데이터 사이언스	모든 분야를 통합하며 분석 행위 자체에 중점(프로젝트 중심) (개발+통계+모델링+도메인 지식)

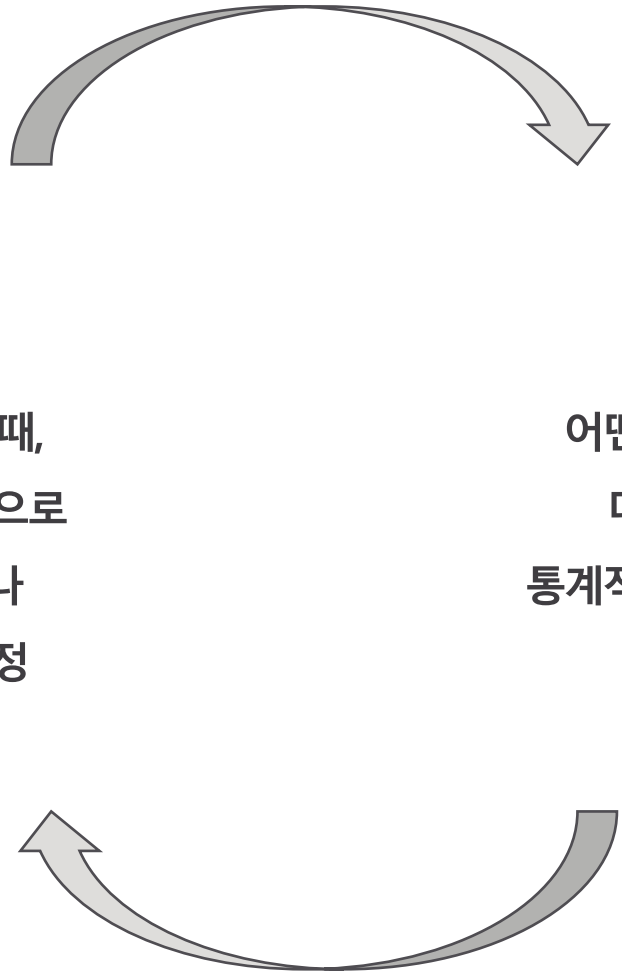
들어가며

통계학이란?


Population
All undergraduate students in North America
Sample
10 undergraduate students in North America
Has iPhone
No iPhone
Point Estimation
Parameter
unknown p
(proportion of population who own an iPhone)
Point Estimate
p̂ = 6/10 = 0.60
(proportion of sample who own an iPhone)

들어가며

통계학이란?



추정

전체를 알 수 없을 때,
일부 데이터를 바탕으로
모수를 예측하거나
값을 계산하는 과정

검정

어떤 주장이 맞는지를
데이터를 이용해
통계적으로 판단하는 과정

들어가며

이번 시간에는..

1

통계적 관점에서의 데이터 분석

2

데이터 프레임(정형 데이터) 이해하기

3

데이터 요약 방법과 여러 지표(통계량)을 살펴보는 것에 초점

4

R 프로그램을 통한 실제 데이터 실습

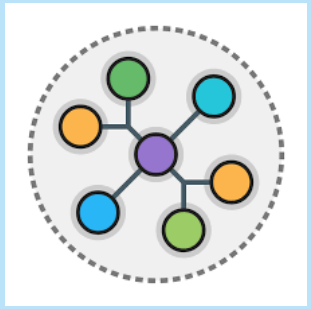
데이터의 분류

데이터

정형 데이터

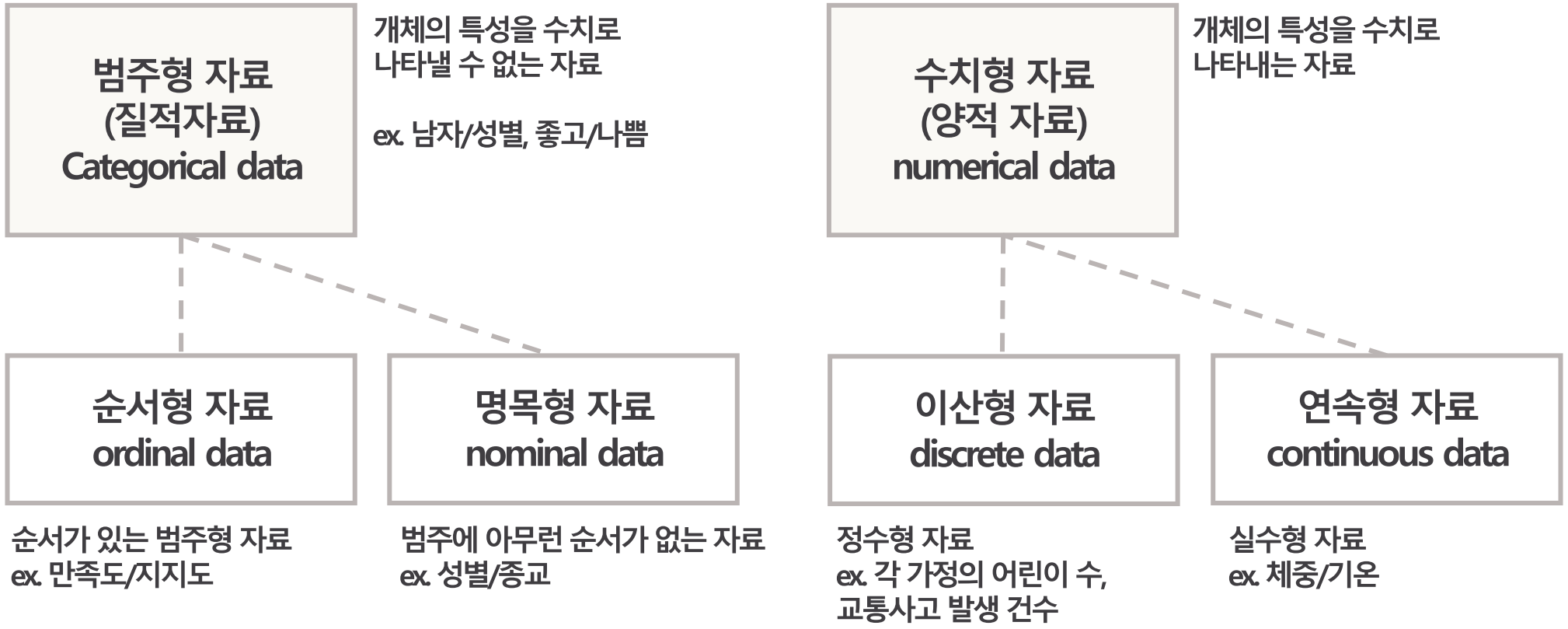
id	이름	나이	성별
01	Kim	32	M
02	Lee	26	F
03	Park	72	F
04	Choi	15	M

비정형 데이터



정형 데이터 분석

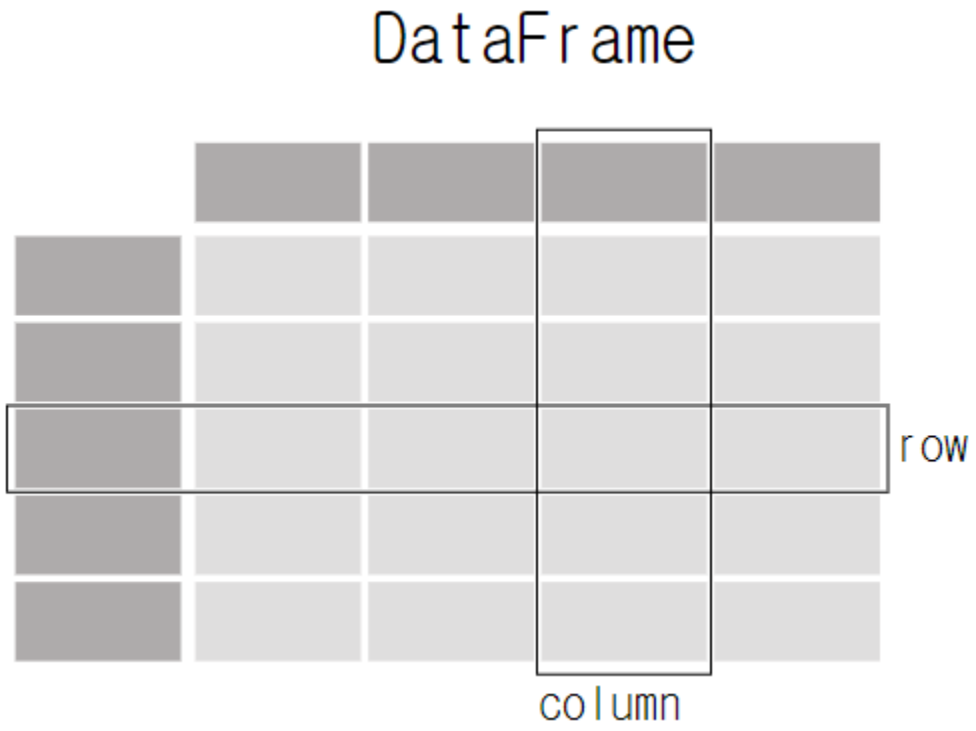
정형 데이터의 분류



정형 데이터 분석

정형 데이터의 표현

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

정형 데이터 분석

정형 데이터의 요약

	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

정형 데이터 분석

정형 데이터의 요약

	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"})
```

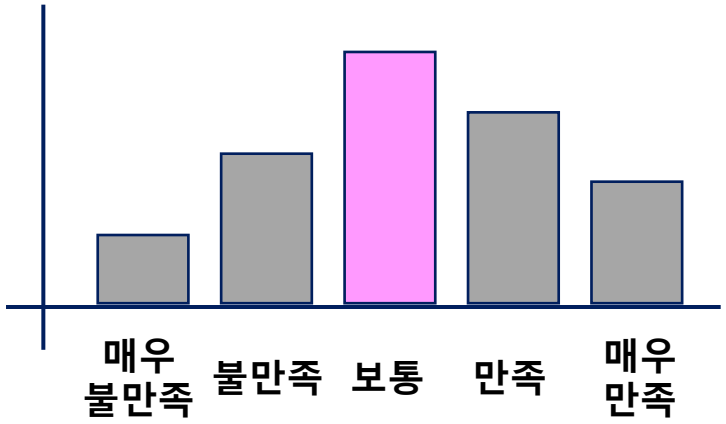
시군구	재산피해금액	출동횟수
은평구	886196.4	921
종로구	711661.8	1129
중구	15976858.0	1042
중랑구	286252.0	1098

정형 데이터 분석

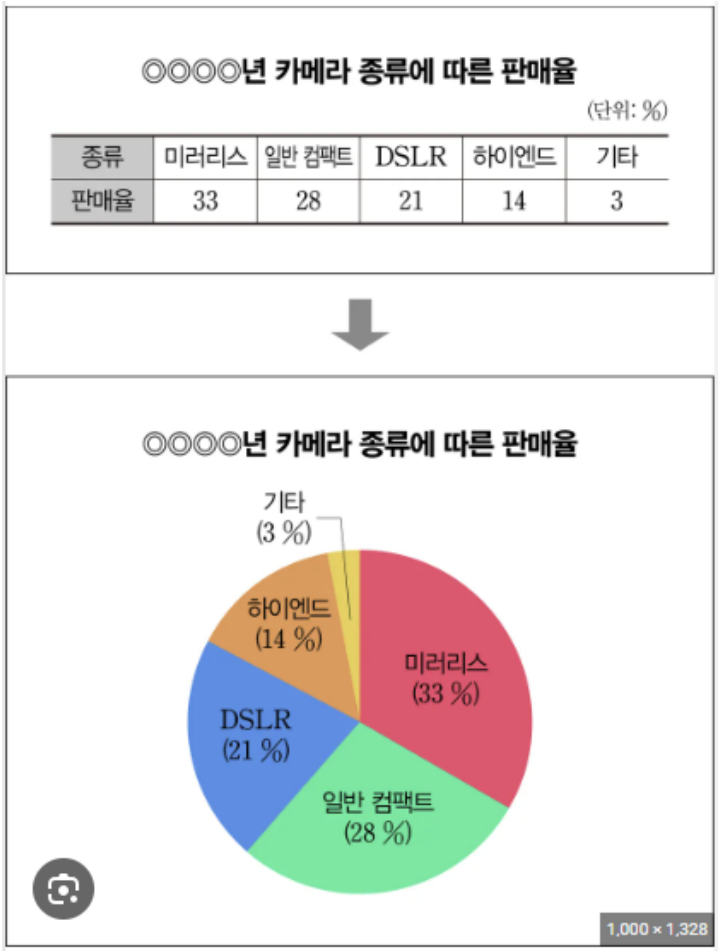
정형 데이터의 시각화

범주형 자료

✓막대그래프



✓원 그래프



✓분할표

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

[표] 교육수준과 결혼생활

정형 데이터 분석

정형 데이터의 시각화

범주형 자료

정형 데이터 분석

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, 8	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 데이터 분석

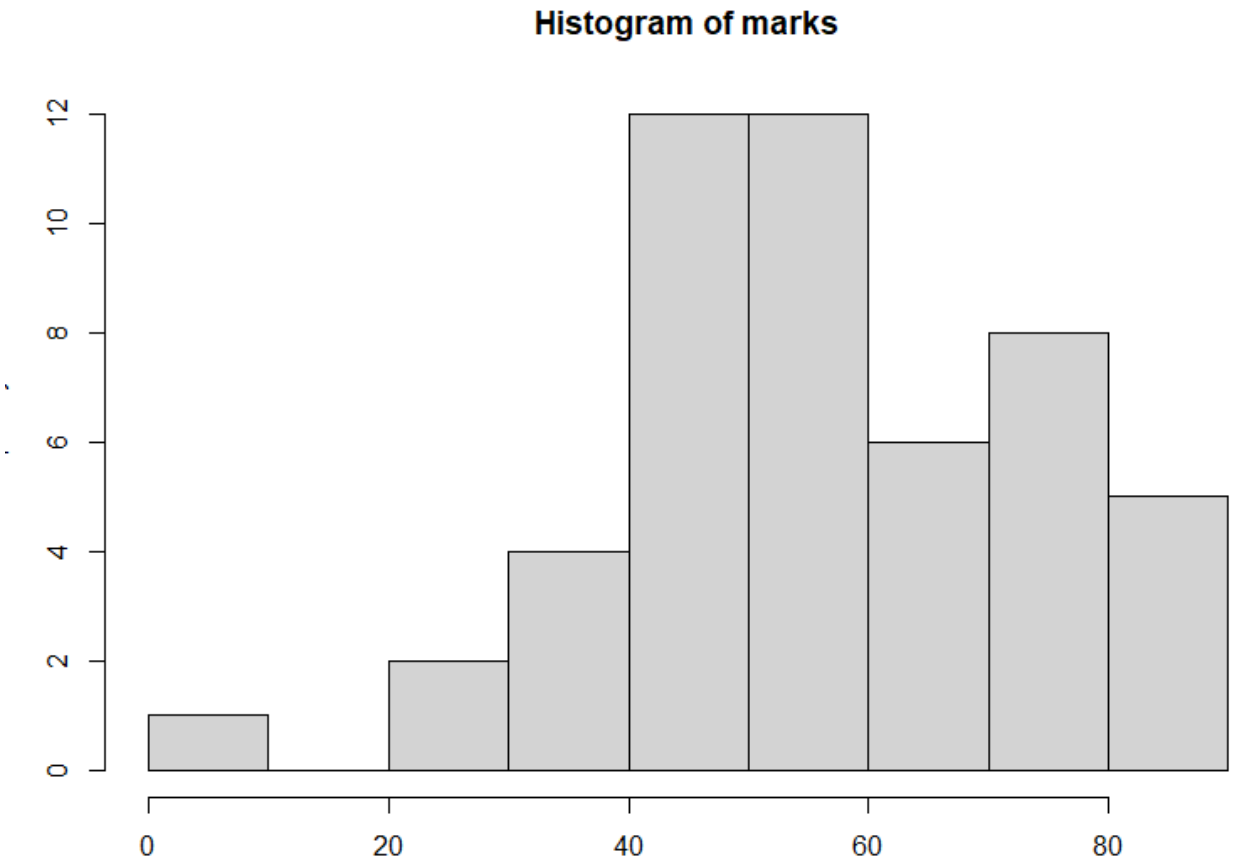
정형 데이터의 시각화

수치형 자료

✓ 도수분포표와 히스토그램

구간: 10개,
구간 폭: 10

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



정형 데이터의 시각화

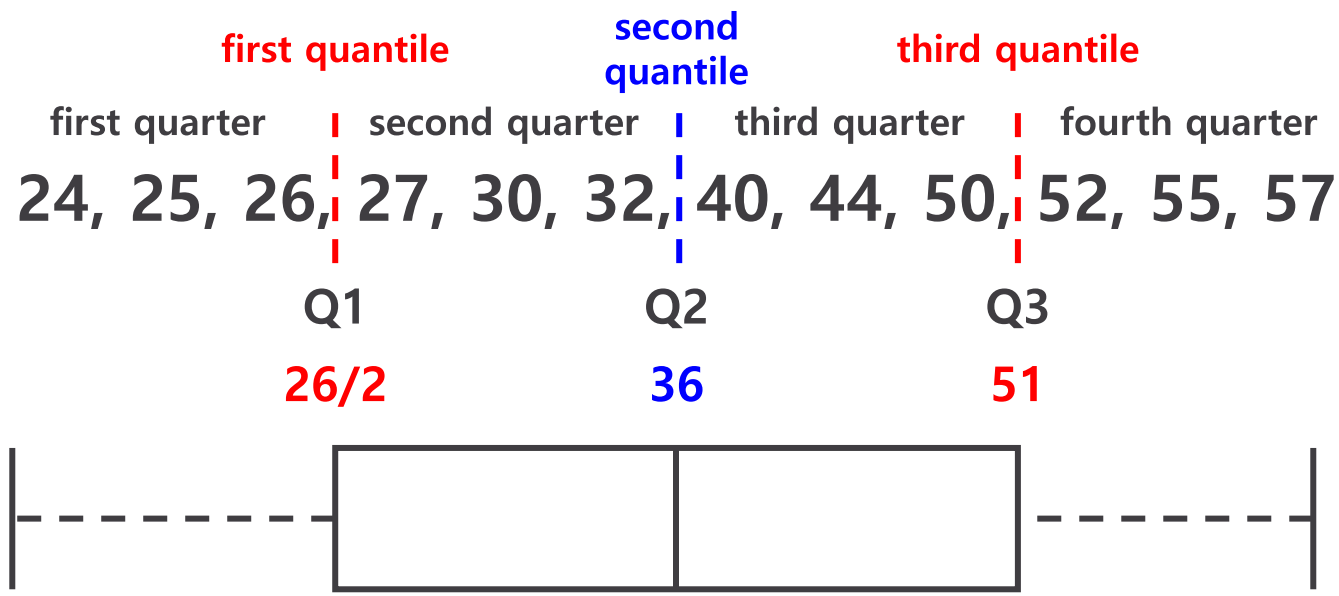
수치형 자료

✓평균과중위수

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

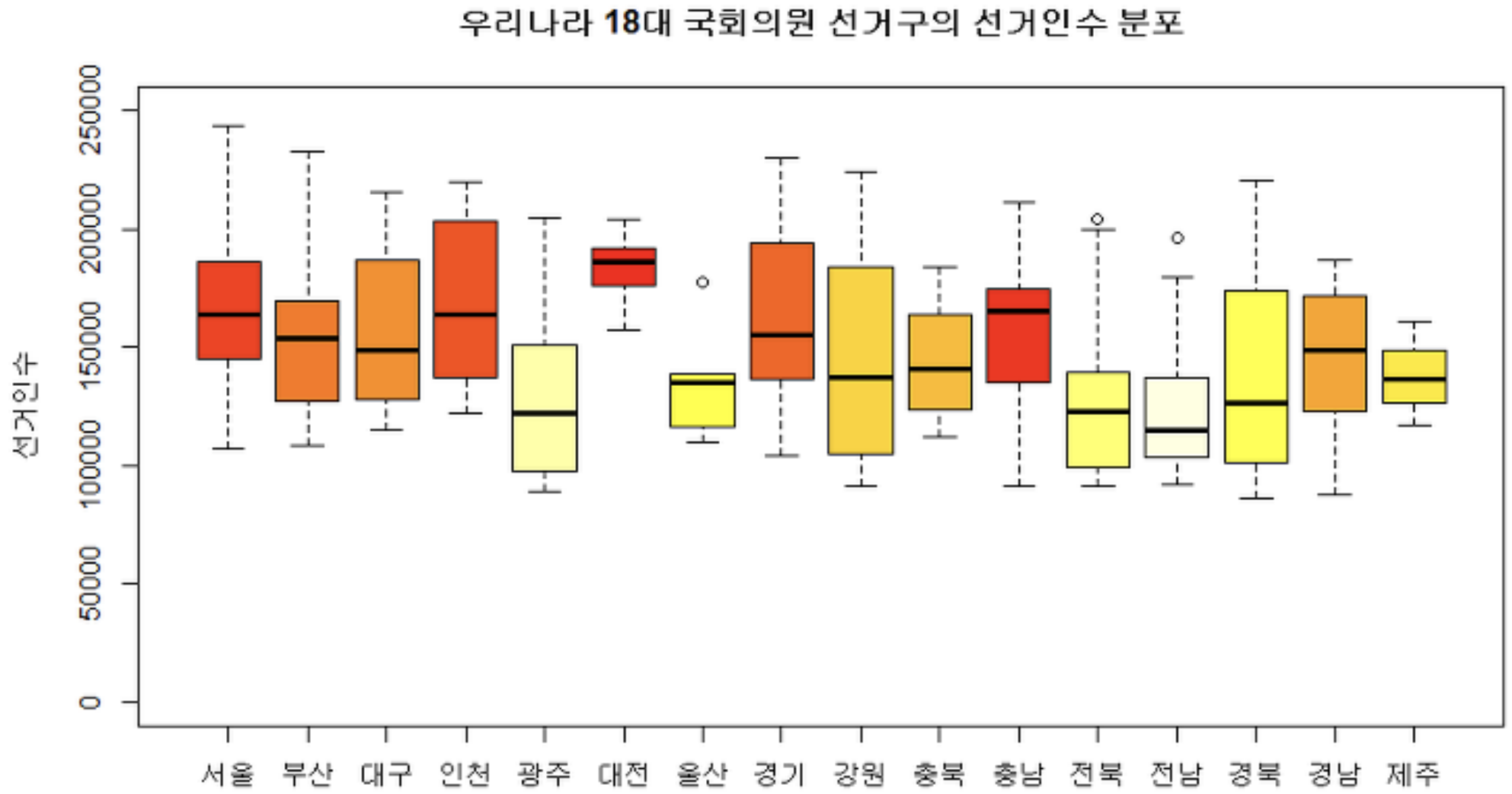
0, 1, 2, 2, 2, 3, 4 → 평균 2, 중앙값 2
70, 1, 2, 2, 2, 3, 4 → 평균 12, 중앙값 2

✓백분위수와상자그림



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.

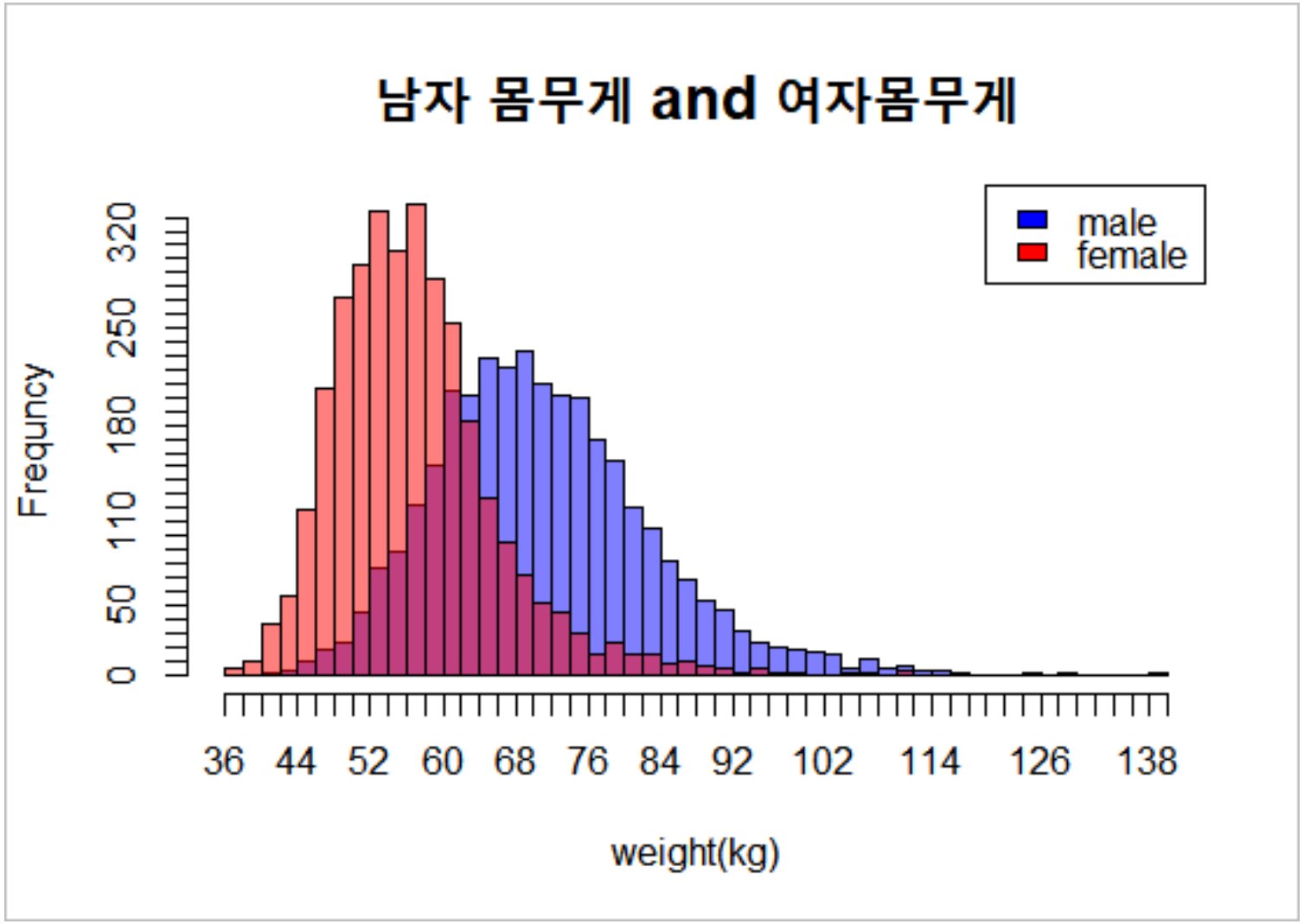
표본표준편차 s은 다음과 같은 식으로 구한다.

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

정형 데이터 분석

정형 데이터의 시각화

수치형 자료



이변량 데이터의 구성

수치형 자료

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, 8	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

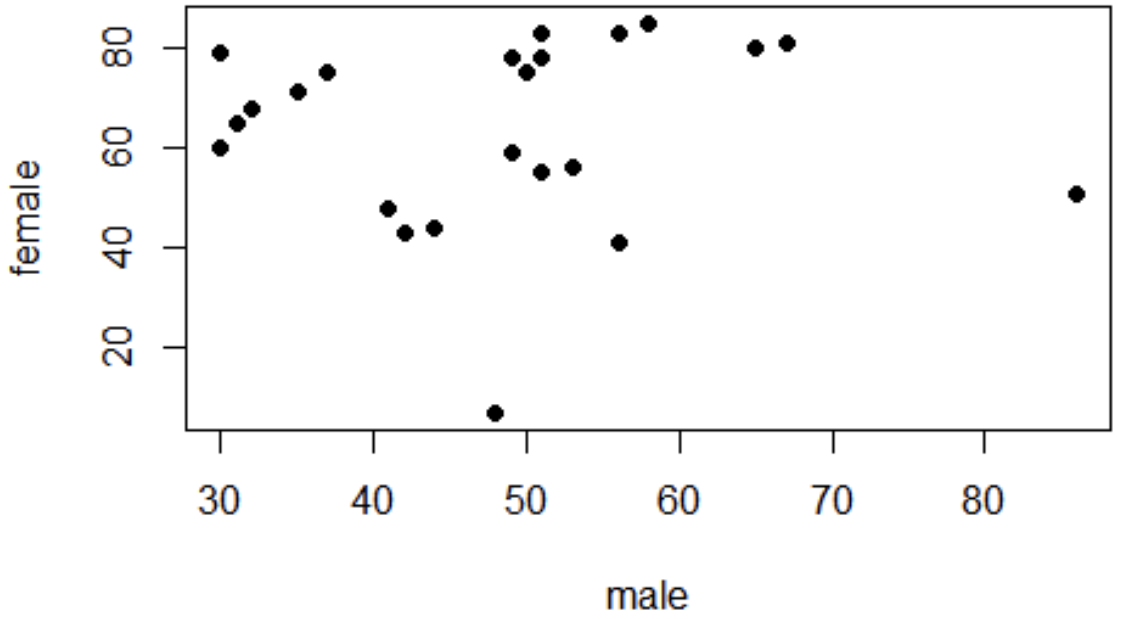
obs	Female	Male
1	7	48
2	59	49
3	78	49
4	79	30
5	60	30
6	65	31
⋮	⋮	⋮

이변량 데이터의 분석방법

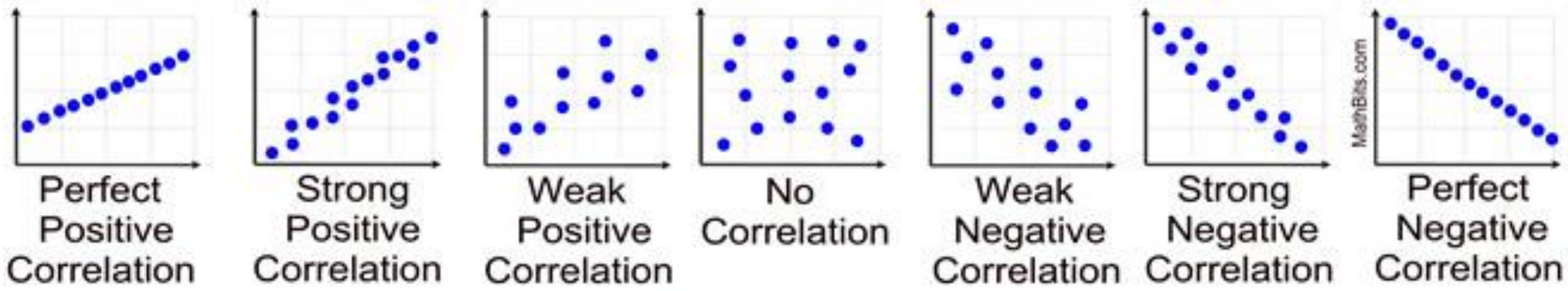
이변량 데이터의 시각화: 산점도

```
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?



이변량 데이터의 분석방법

✓ 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.

- 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.

통계적 예측방법

개요


분석 목적에 따라
분석 방법(모형)이 다르다
회귀
(regression)
선형회귀모형
분류
(classification)
KNN (K-nearest neighborhood)
의사결정나무

통계적 예측방법

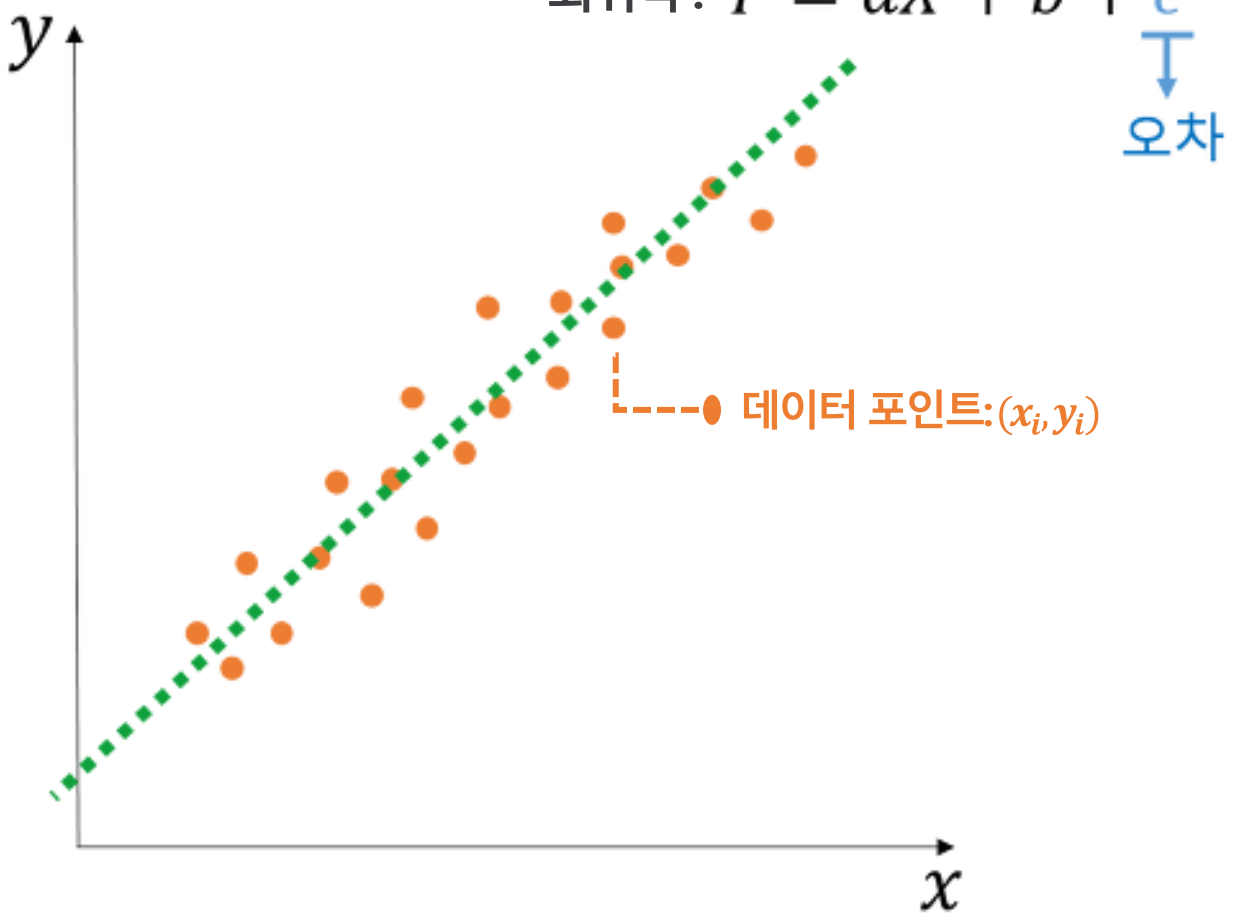
회귀모형

회귀

선형 회귀모형
- 데이터를 가장 잘 설명하는 선을 찾는 방법

회귀식 : $Y = aX + b + \epsilon$

오차

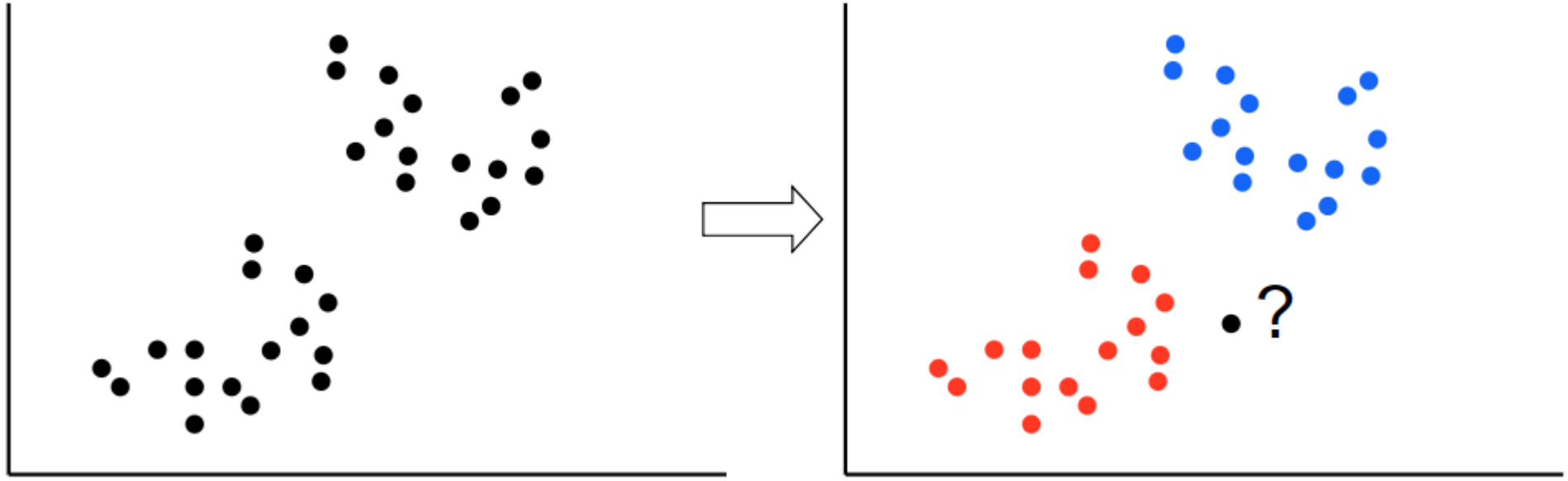


통계적 예측방법

분류모형

분류

KNN
- 내 이웃의 정보를 사용하여 데이터를 나누는 방법

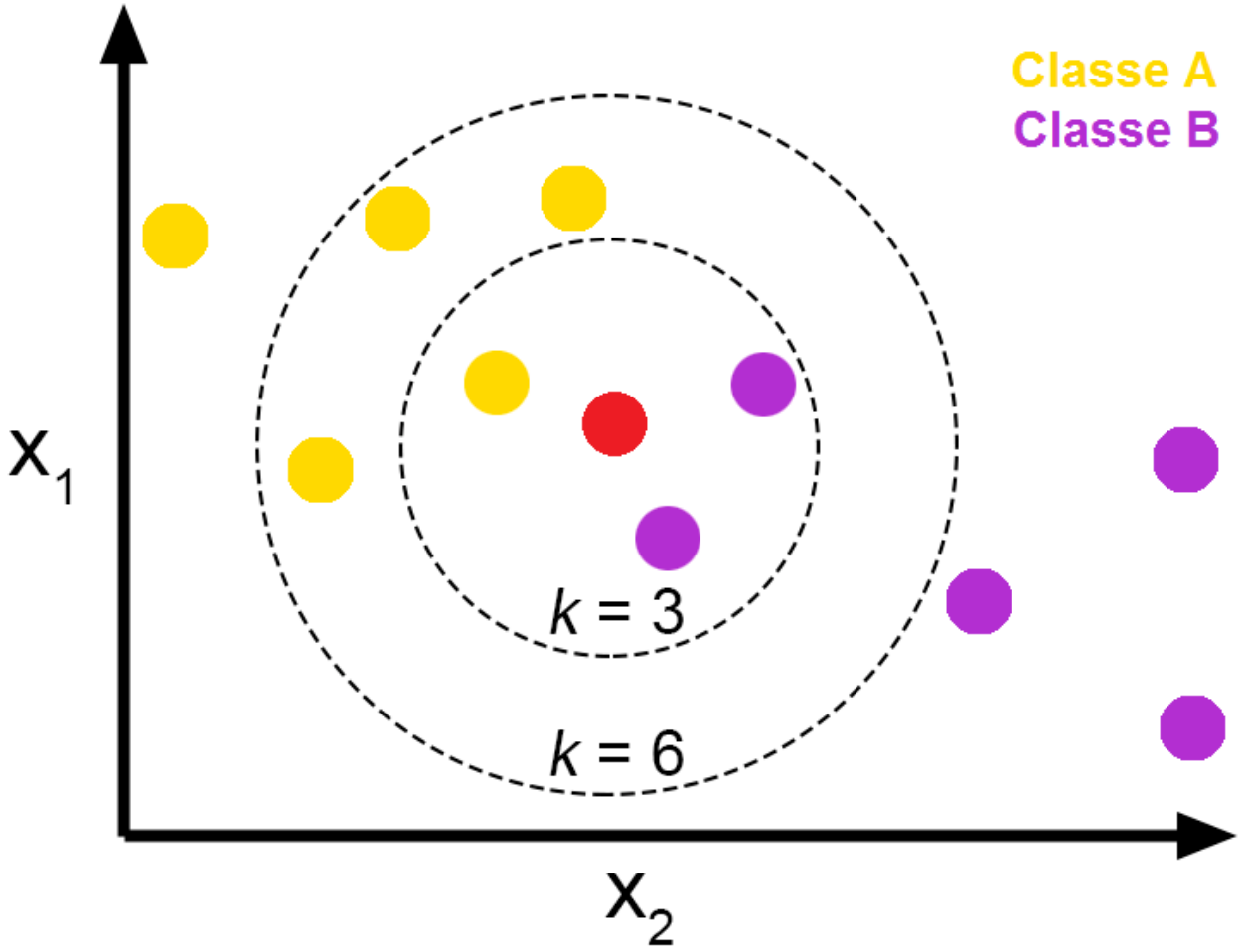


통계적 예측방법

분류모형

분류

KNN
- 내 이웃의 정보를 사용하여 데이터를 나누는 방법



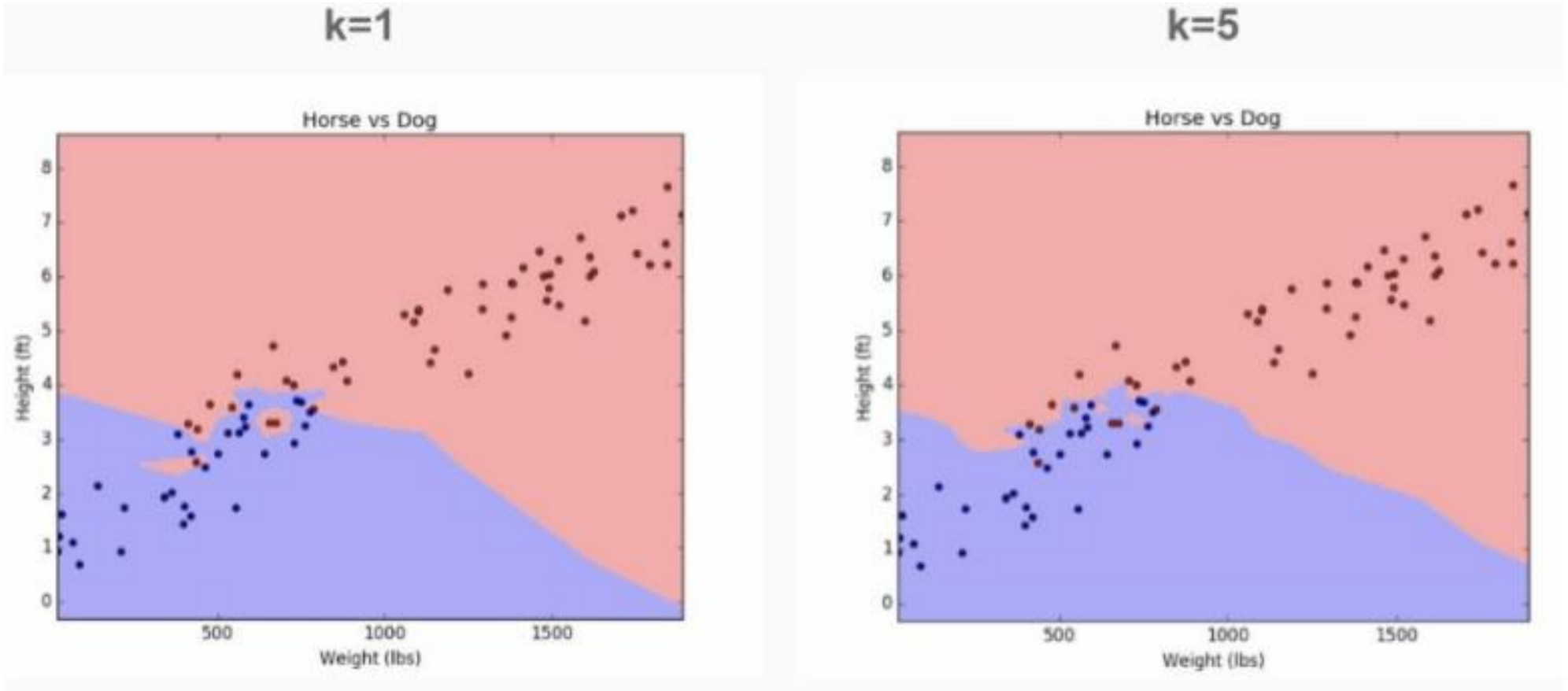
통계적 예측방법

분류모형

분류

KNN

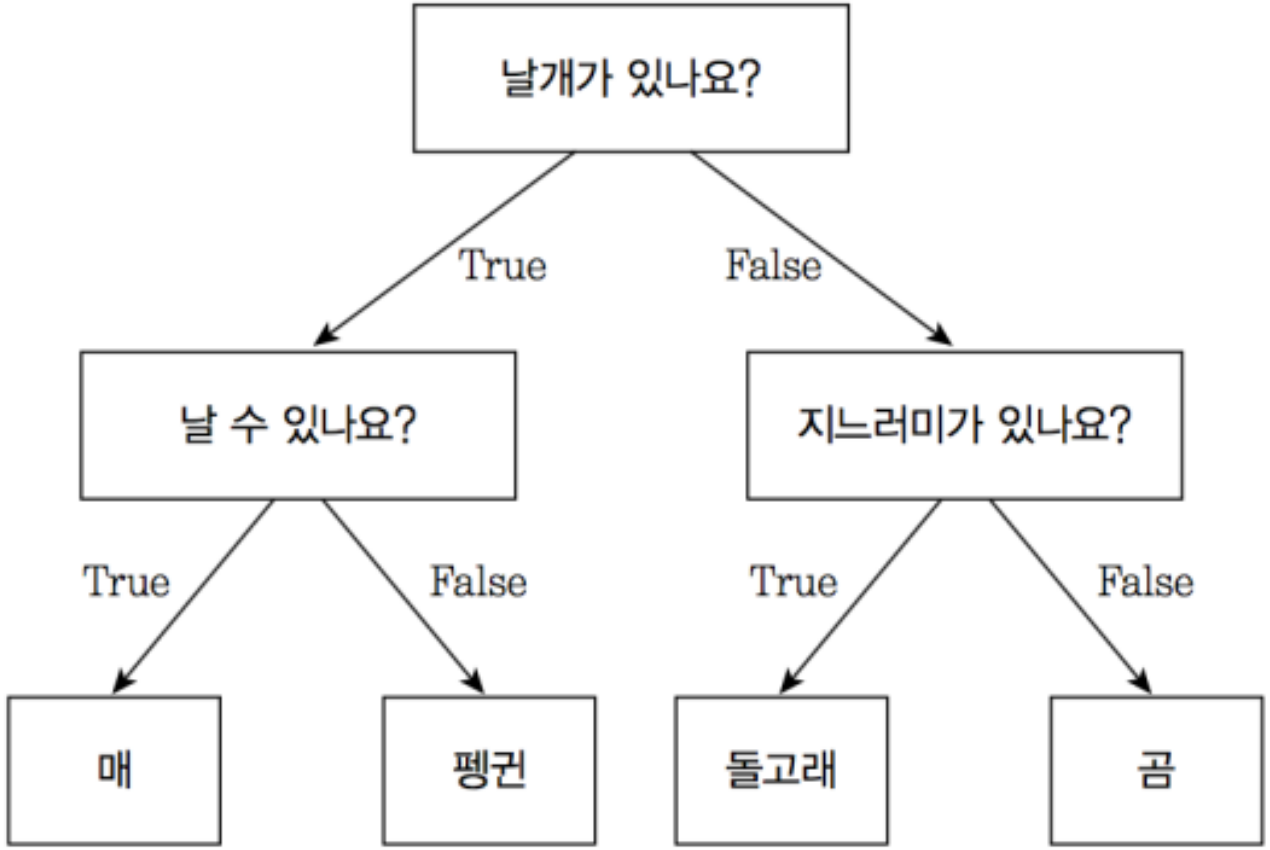
- 내 이웃의 정보를 사용하여 데이터를 나누는 방법

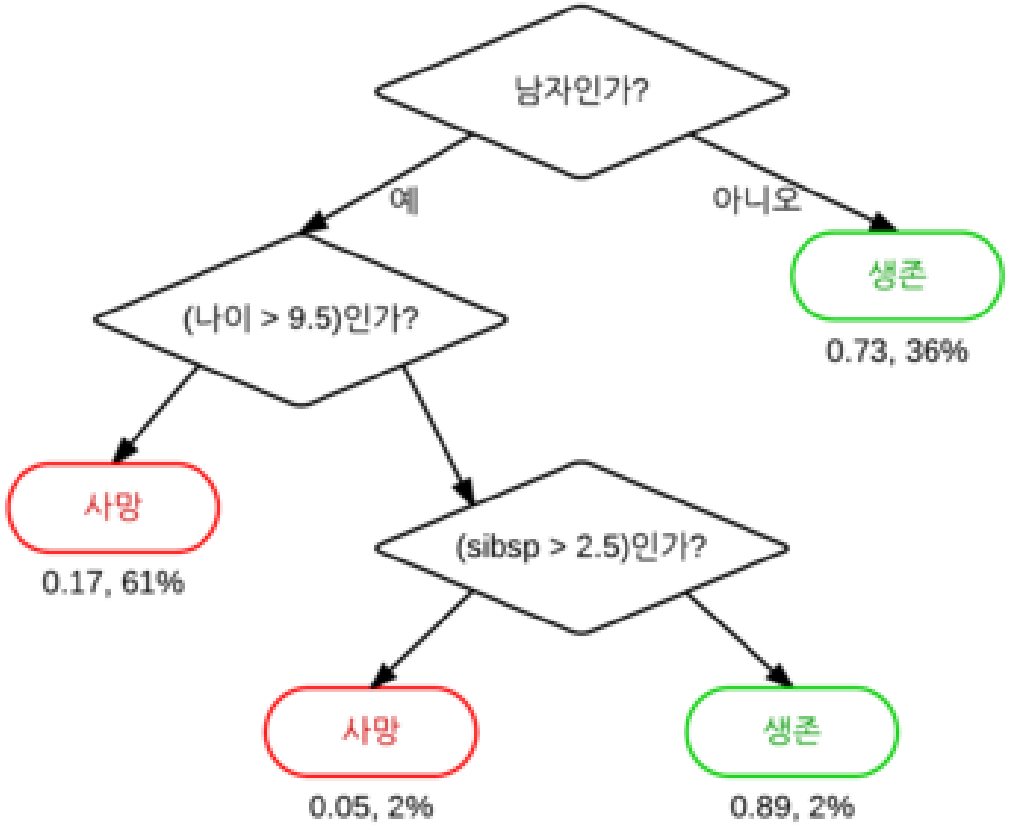


분류모형

분류

의사결정나무
- 데이터의 조합에 대한 의사결정 규칙에 따라 데이터를 분류하는 방법





통계적 예측방법

분류모형

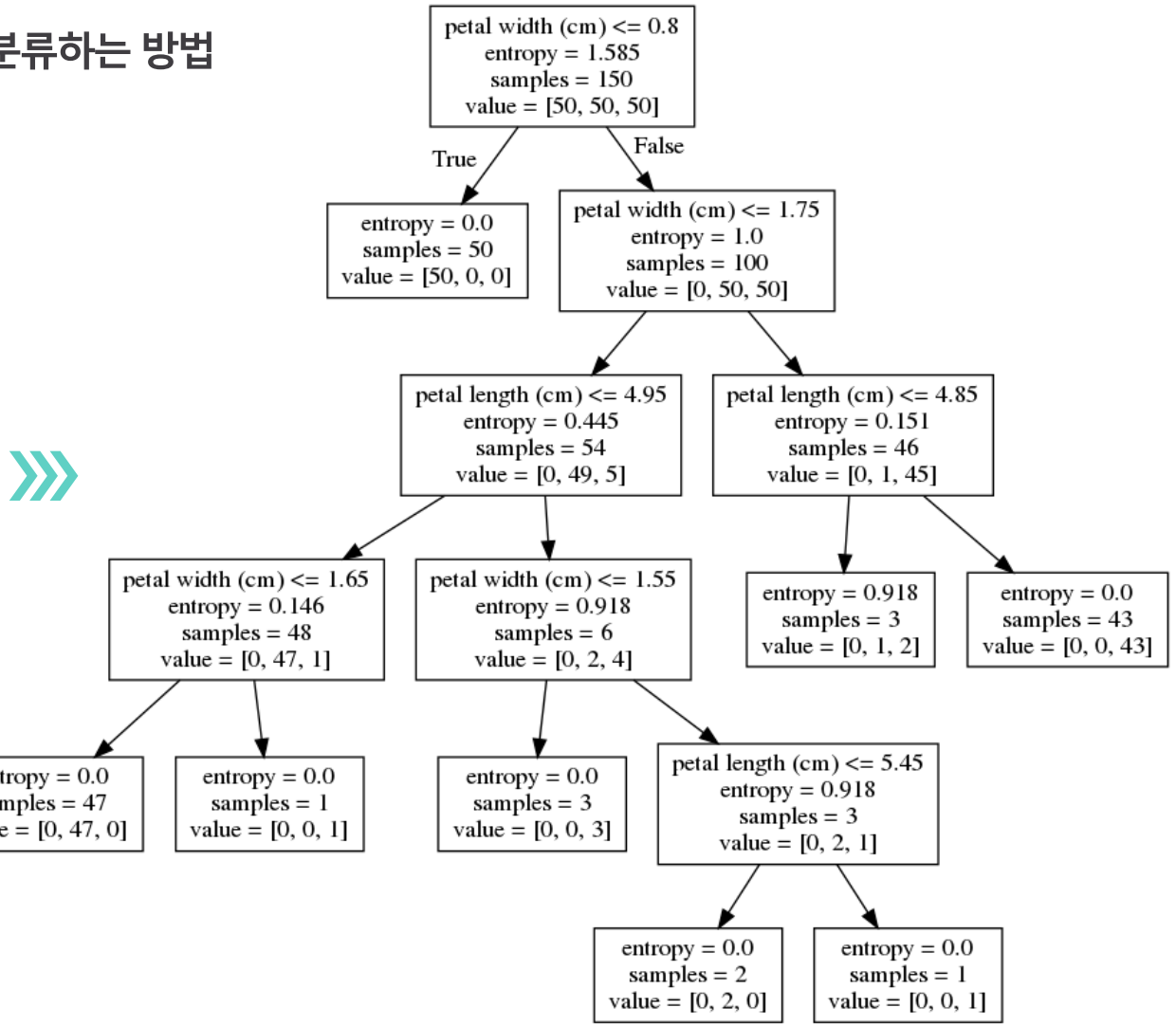
분류

의사결정나무
- 데이터의 조합에 대한 의사결정 규칙에 따라 데이터를 분류하는 방법

- 데이터 과학에서 Iris DataSet
  - 아이리스 품종 중 Setosa, Versicolor, Virginica 분류에 대한 로널드 피셔의 1936년 논문에서 사용된 데이터 셋.

- 꽃받침(Sepal)과 꽃잎(Petal)의 길이 너비로 세개 품종을 분류

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	4.9	3.0	1.4	0.2	setosa
1	4.7	3.2	1.3	0.2	setosa
2	4.6	3.1	1.5	0.2	setosa
3	6.4	3.2	4.5	1.5	versicolor
4	6.9	3.1	4.9	1.5	versicolor
5	5.5	2.3	4.0	1.3	versicolor
6	7.1	3.0	5.9	2.1	virginica
7	6.3	2.9	5.6	1.8	virginica
8	7.6	3.0	6.6	2.1	virginica



통계적 예측방법

데이터의 분할

데이터를 분할시켜 학습하는 것이 기본!

	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

모델링의 목적

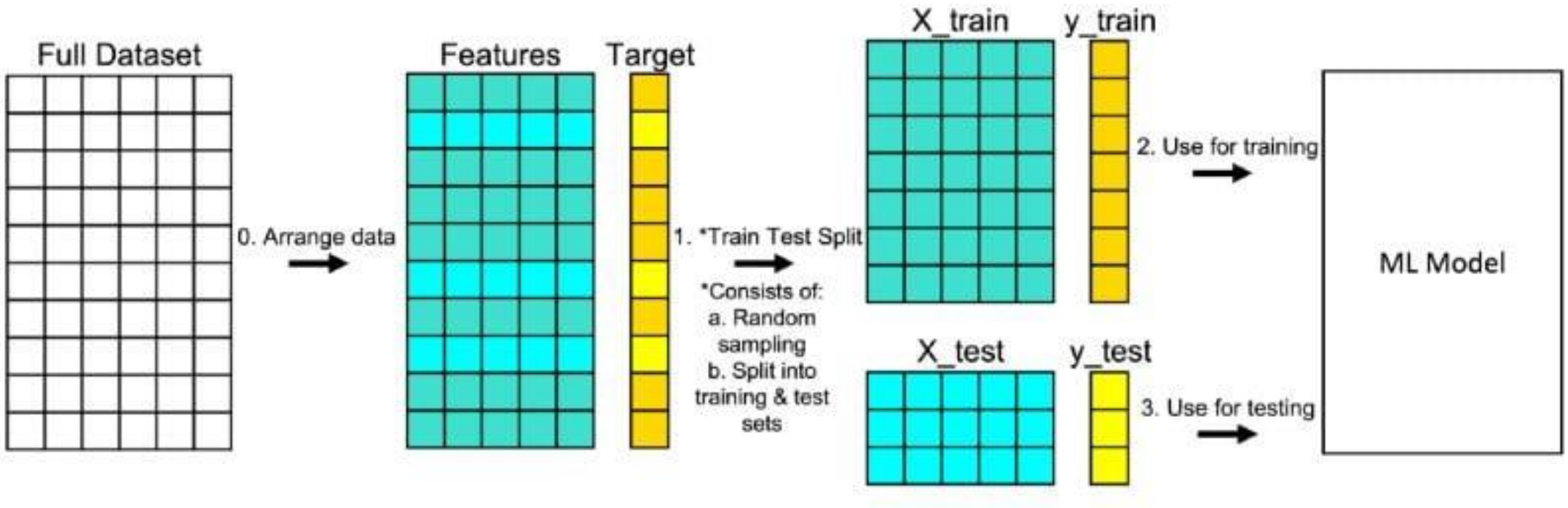
새로운 데이터가 들어왔을 때
이 데이터의 값/라벨을 예측하는 것!

1) 2017~2021년도의 화재발생 데이터로 모델링

2) '2022년도의 화재발생 사망자수' 예측

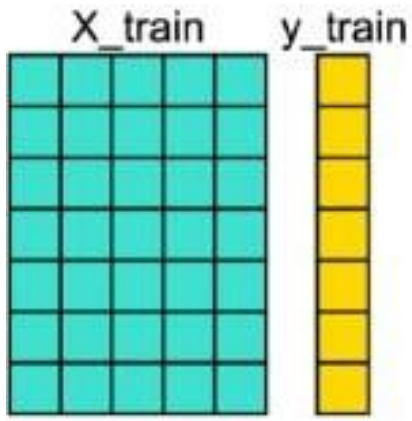
데이터의 분할

데이터를 분할시켜 학습하는 것이 기본!



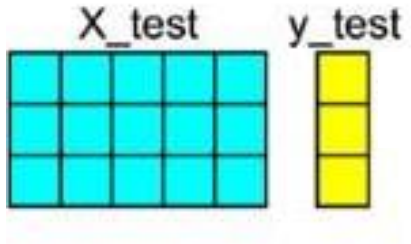
데이터의 분할

데이터를 분할 설계하기



훈련데이터

독립변수(X_train): 2017~2021년도의 시군구/출동건수/부상자수/...
종속변수(y_train): 2017~2021년도의 사망자수



시험데이터

독립변수(X_test): 2022년도의 시군구/출동건수/부상자수/...
종속변수(y_test) = ?