데이터 분석 과정

기술통계 및 그래프 분석

SK 주식회사 C&C

Agenda

1. 기술통계(Descriptive Statistics)

2. 그래프 분석

데이터 구성 – 변수(Variable)와 관측치(Observation)

13개의 행과 5개의 열로 이루어진 데이터

Table
= Data Set
= Data Frame

변수(Variable) / 열(column) / 특성(Feature)

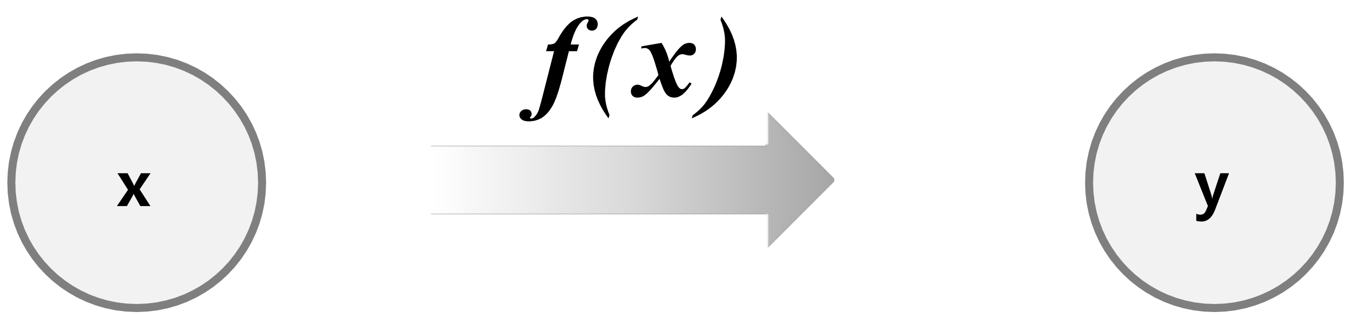
관측치(Observation) / 행(row) / 케이스

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa

독립/설명변수 vs. 타겟/종속/반응변수

X를 입력하여 Y를 추론하는 것이 모델 f(x)

- 분석가가 알고자 하는 값은 y (종속변수, 타겟변수, 반응변수)
- y를 추론하기 위해 활용할 수 있는 값은 x (독립변수, 설명변수, 특성, 피쳐…)



독립변수 (Independent Variable)
설명변수 (Explanatory Variable)

종속변수 (Dependent Variable)
반응변수 (Response Variable)

데이터 구성 – 변수(Variable)와 관측치(Observation)

13개의 행과 5개의 열로 이루어진 데이터

setosa versicolor virginica

설명변수, 특성, 피쳐..

타겟변수
종속변수

변수(Variable) / 열(column) / 특성(Feature)

관측치(Observation) /
행(row) /
케이스

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa

Table
= Data Set
= Data Frame

변수의 데이터 유형(척도)

변수의 데이터 유형/특성에 따라 사용 가능한 분석기법이 달라지므로 구분할 줄 알아야 함

【 범주형 】

범주형/이산형
(Categorical/Discrete)

이진척도(Binary)

- 2개의 서로 다른 상태를 구분
- 예: 합격 여부 (합격=1, 불합격=0), 양불 여부

명목척도(Nominal)

- 데이터 특성을 분류하기 위해 수치로 기호 부여
- 수치 간의 양적인 의미는 없음
- 예: 상품분류(패션=1, 뷰티=2, 식품=3), 품종 등

서열척도(Ordinal)

- 데이터간 순서 존재
- 수치 간의 양적인 의미 있음
- 예: 순위(1등 > 2등 > 3등)

【 수치형 】

수치형/연속형
(Numerical/Continuous)

등간척도(Interval)

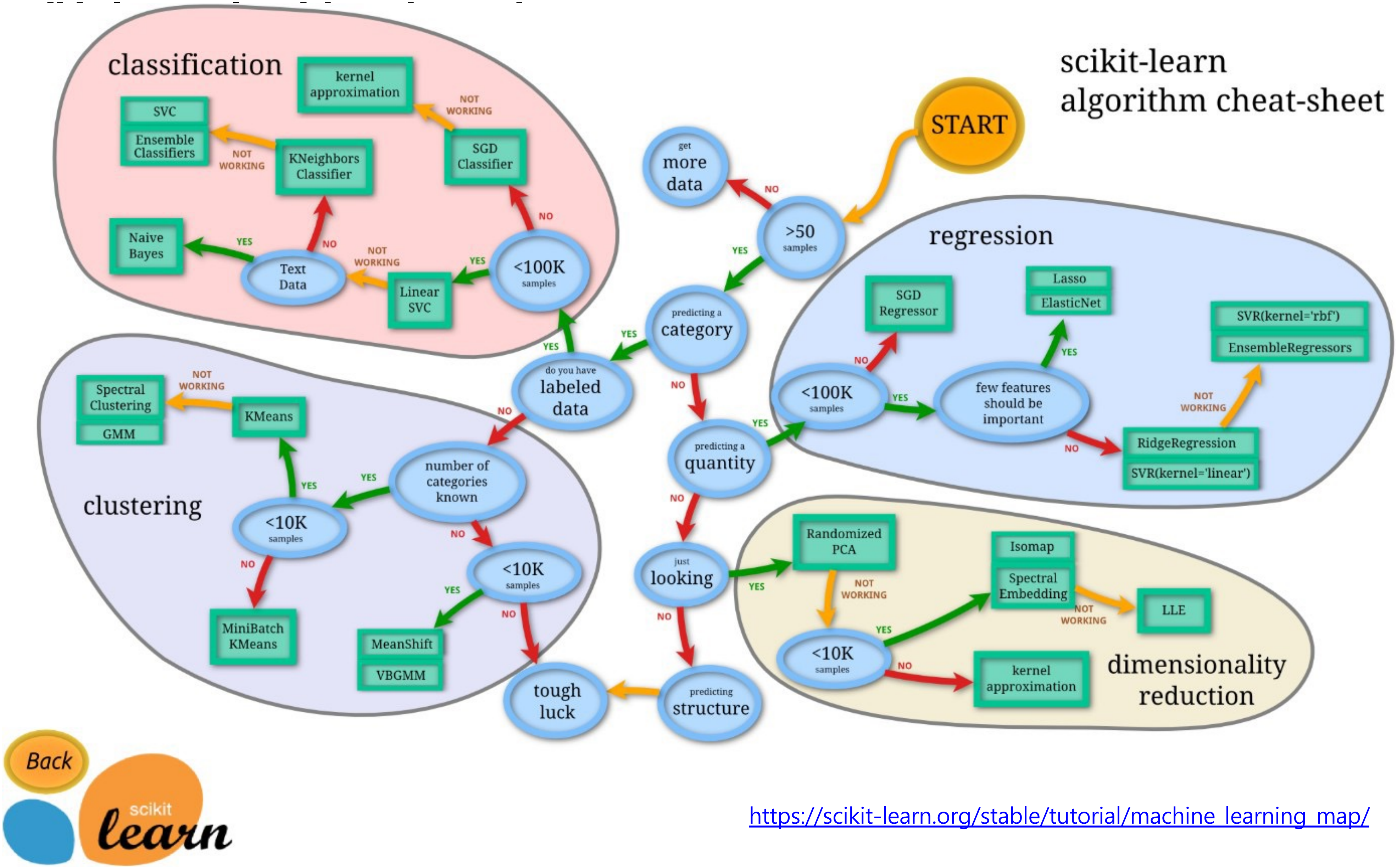
- 명목, 서열척도의 특징을 모두 가지고 있으면서 크기가 어느정도나 되는지, 특성간의 차이가 어느정도 나 되는지 파악이 가능한 척도
- 숫자 간의 간격(interval)이 균등
- 더하기 빼기 가능
- 없음(無)의 의미는 가지지 못함
- 예: 온도(임의의 온도가 0도)
  서기 년도(서기 0년은 예수님 태어나신 해로 구분)

비율척도(Ratio)

- 가장 높은 수준의 척도
- 등간 척도의 특성을 가지고 있으면서 측정치간 **사칙연산이 가능함**
- 없음(無)의 의미 가짐(절대영점)
- 예: 몸무게(0kg), 소요시간(0초), 키(0cm)

척도(Scale) : 어떠한 대상의 특성을 단위를 사용하여 정량화 한 것

예시) 데이터 유형/특성에 따른 분석기법 선택



https://scikit-learn.org/stable/tutorial/machine_learning_map/

기술통계량(Descriptive Statistics)

대표값, 퍼짐의 정도, 쏠림으로 데이터의 특성을 표현함

중심위치 (대표성)

- 평균
- 중앙값
- 최빈값

퍼짐 (유사도, 응집도)

- 분산
- 표준편차
- 변동계수
- 범위(Max-Min)

쏠림 (편향)

- 왜도
- 첨도
- 사분위수
- 사분위수 범위

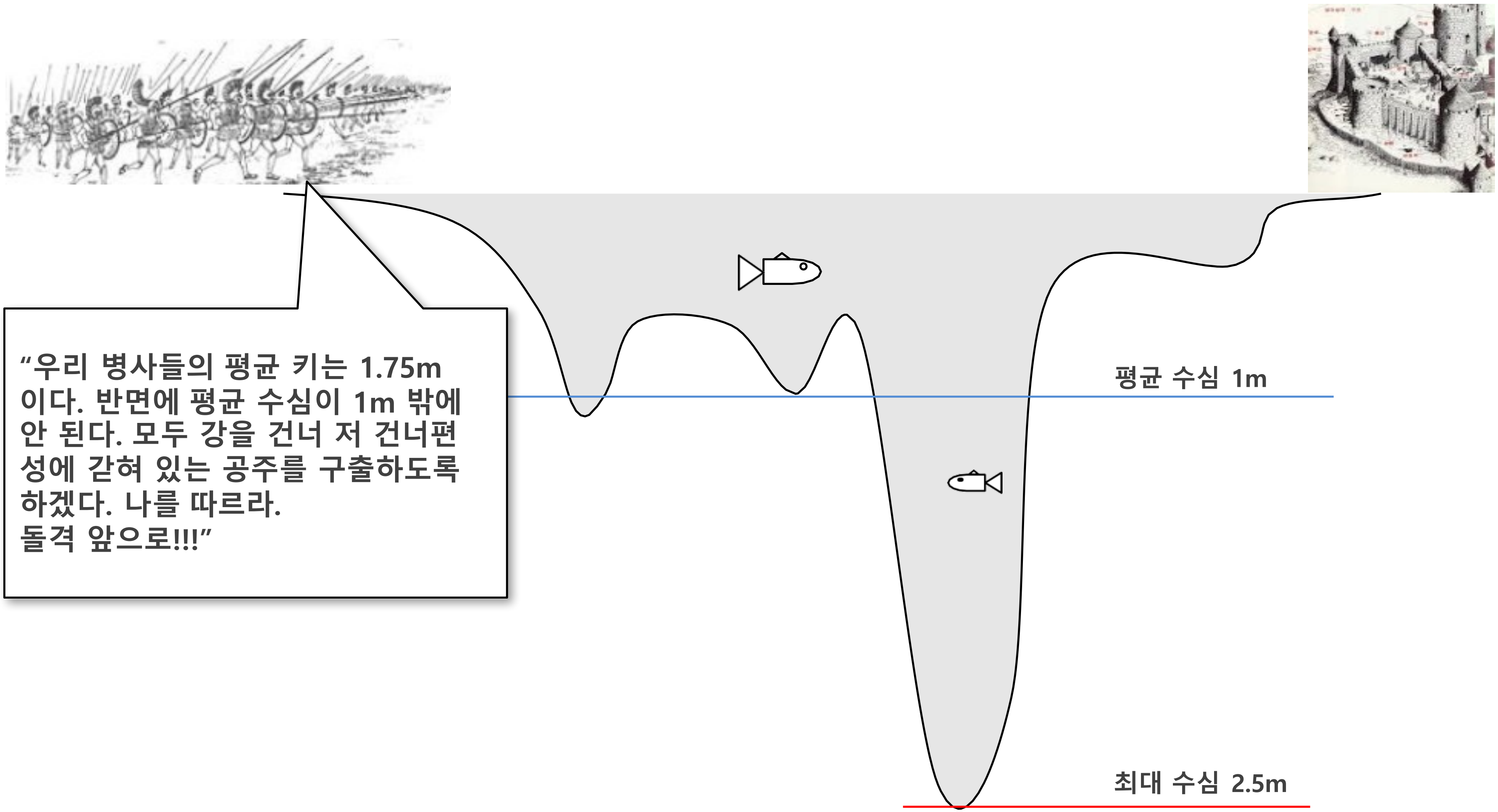
기술통계 - 중심 위치(대표성)

중심 위치 측도로 평균, 중앙값, 최빈값이 있으며, 분석 목적 및 데이터 속성 특성에 따라 적합한 대표값 선택 필요함

통계량	설명	특징
평균 (mean)	(산술평균) 관측값의 합을 관측값의 개수로 나눈 값 $\frac{1}{n}\sum_{I=1}^{n} x_i$	■ 특이값이 없는 정규분포 데이터에 최적 ■ 표본집단의 평균은 모집단의 평균 추정에 활용 ■ **특이값(outlier)에 민감함 : 그래프 확인 및 특이값 제거 후 분석하거나 중앙값 활용** ■ **Data 개수가 적고, 특이값 존재 시 : 중앙값 활용**
중앙값 (median)	데이터를 순서대로 나열할 때 가운데 있는 값	■ 관측값의 개수가 짝수인 경우 가운데 두 값의 평균 ■ **특이값(outlier)에 영향 덜 받음** ■ 사분위수에서 2사분위수와 같음
최빈값 (mode)	데이터 중 빈도가 가장 높은 값	■ 범주형 변수에서 가장 빈도수가 높은 값을 도출 시 사용 ■ **연속형 데이터의 경우 먼저 구간으로 범주화한 후 활용**

기술통계 활용 - 평균의 함정

『평균의 함정』을 모르는 지도자를 만나면?



기술통계 - 중심위치

키(cm) 데이터의 대표값을 구해보자.

	키(㎝)	정렬
1	178	170
2	173	170
3	180	173
4	173	173
5	181	173
6	170	176
7	178	178
8	170	178
9	173	180
10	176	181

평균 $= \frac{\text{관측값 합}}{\text{관측값 수}} = \frac{178+174+\ldots+176}{10} = \frac{1752}{10} = 175.2$

중앙값 $= \frac{173+176}{2} = \frac{349}{2} = 174.5$

값	빈도수
170	2
173	3
176	1
178	2
180	1
181	1

최빈값 = 173

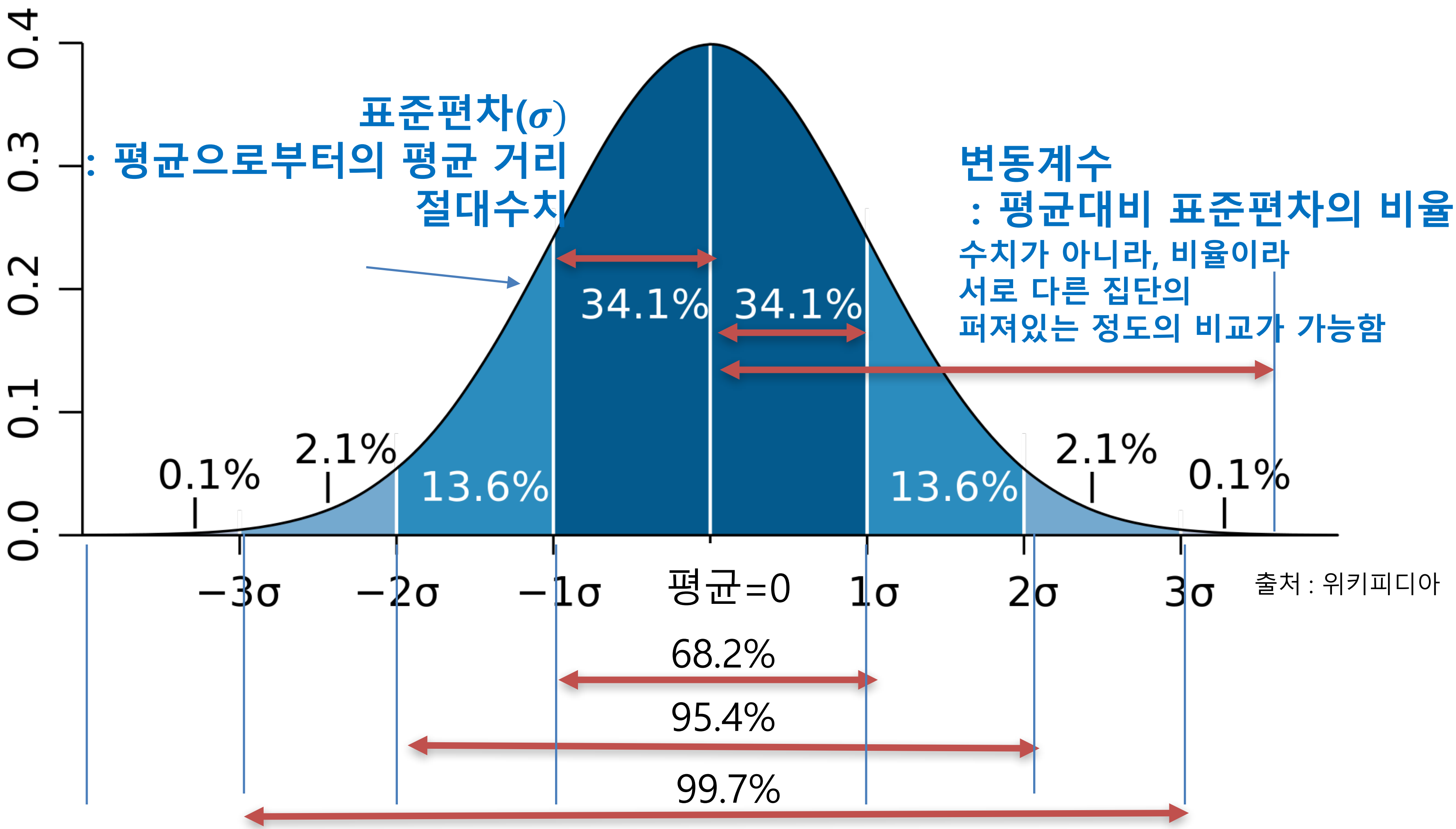
기술통계 - 퍼짐(유사도, 응집도)

데이터가 (평균으로부터) 얼마나 흩어져 있는지 하나의 값으로 표현할 때 사용함

통계량	설명	특징
분산 (Variance)	$\sigma^2 = \frac{\sum(Xi - \mu)^2}{N}$ (σ^2 (시그마) : 분산 X_i : 측정치, μ(뮤) : 평균, 0≤ I ≤N)	데이터간의 편차가 얼마나 들쑥날쑥한지 그 정도를 표현 (**편차 = 각 관찰값 – 평균**) 분산이 클 수록 퍼져있는 정도가 큼 **편차^2 이라 활용이 어려움(단위:Scale문제)** **표준편차 사용**
표준편차 (Standard Deviation)	분산의 제곱근(루트) $\sqrt{\sigma^2} = \sigma$	**평균에서의 평균 거리**를 의미함 **절대 크기가 현저하게 달라서 평균이 크게 다르거나 측정 단위가 다른 변수를 비교하기에 부적합** **변동계수 사용**
변동계수 **(Coefficient of Variation)**	표준편차를 평균으로 나눈 비율 $\frac{\sigma}{\mu}$	평균에 대비한 표준편차의 비율 (단위: %) **두 변수를 비교할 때, 절대 크기가 현저하게 달라서 평균이 크게 다르거나 측정단위가 다른 두 변수를 비교할 때**
범위 (Range)	최대값 – 최소값	**범위는 특이치가 있을 경우 왜곡 존재 사분위수 범위(IQR)를 활용한 특이치 탐색**

분산과 표준편차

예시) 표준정규분포



출처 : 위키피디아

기술통계 – 퍼짐

키(cm) 데이터의 평균, 편차, 분산, 표준편차를 구해보자.

	키(㎝)	편차	편차2
1	170	-5.2	27.04
2	170	-5.2	27.04
3	173	-2.2	4.84
4	173	-2.2	4.84
5	173	-2.2	4.84
6	176	0.8	0.64
7	178	2.8	7.84
8	178	2.8	7.84
9	180	4.8	23.04
10	181	5.8	33.64

평균 = 175.2

편차 = 관측값 – 평균

$$\textbf{분산} = \frac{\text{편차}^{E}\text{의 합}}{\text{관측값 수}} = \frac{(27.04+27.04+\ \dots\ +33.64)}{10} = \frac{141.6}{10} = 14.16$$

$$\textbf{표준편차} = \sqrt{\text{분산}} = \sqrt{14.16} = 3.76$$

평균에서의 평균 차이는 3.76 (cm)임

기술통계 - 퍼짐

만약에 데이터의 단위가 바뀐다면 표준편차는 어떻게 될까?

(1cm = 0.032808ft)

	키(㎝)	키(ft)
1	170	5.577360
2	170	5.577360
3	173	5.675784
4	173	5.675784
5	173	5.675784
6	176	5.774208
7	178	5.839824
8	178	5.839824
9	180	5.905440
10	181	5.938248

단위	cm	ft
평균	175.2	5.747962
분산	14.16	0.01524133
표준편차	3.76	0.1234558
변동계수	0.02146119	0.02147819
범위	11	0.360888

$$\text{변동계수} = \frac{\text{표준편차}}{\text{평균}}$$

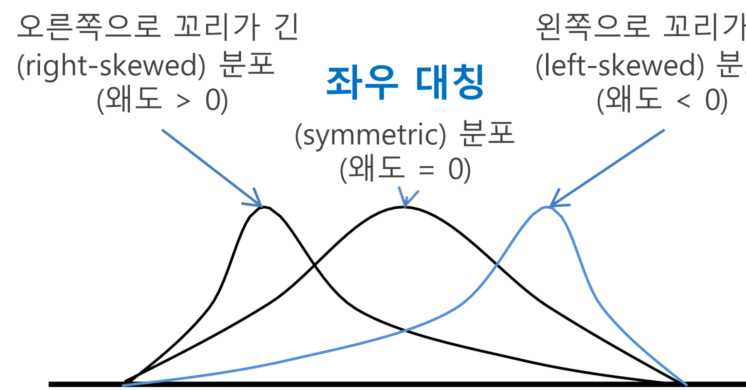
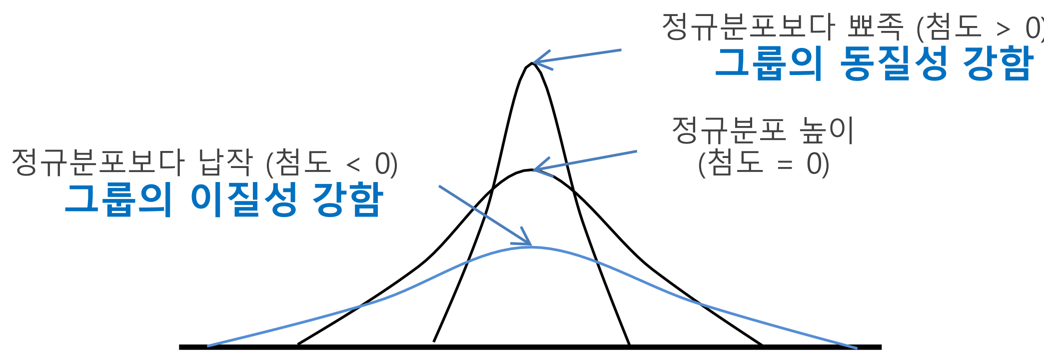
$$\text{변동계수(cm)} = \frac{3.76}{175.2} = 0.02146119$$

$$\text{변동계수(ft)} = \frac{0.1234558}{5.747962} = 0.02147819$$

**변동계수는 단위나 크기가 다른 값을
비교할 때 쓸 수 있다.**

기술통계 – 쏠림

데이터가 얼마나 어느 쪽으로 쏠려 있는지 하나의 값으로 표현할 때 사용함

<table>
<tr><th>통계량</th><th>설명</th><th>특징</th></tr>
<tr><td>왜도
(Skewness)</td><td>데이터의 분포 모양이
어느 쪽으로 얼마나
기울었는지
즉, 대칭성을 알아보는
척도</td><td>
</td></tr>
<tr><td>첨도
(Kurtosis)</td><td>정규분포와 비교하여
봉오리가 얼마나 높은지
알아보는 측도,
평균을 중심으로 넓게 or
좁게 분포 되어 있는지의
비율</td><td>
</td></tr>
</table>

기술통계 - 쏠림

데이터가 얼마나 어느 쪽으로 쏠려 있는지 하나의 값으로 표현할 때 사용함 (중앙값 기준)

통계량	설명	특징
사분위수 (Quartile)	데이터를 순서대로 나열할 때 25, 50, 75, 100%번째에 있는 값	제1사분위수(Q1) : 25%번째 값 제2사분위수 (Q2) : 50%번째 값 (중앙값과 같음) 제3사분위수(Q3) : 75%번째 값 제4사분위수(Q4) : 100%번째 값 (최대값과 같음)
사분위수 범위 (IQR, Interquartile Range)	제3사분위수 – 제1사분위수	Q1 또는 Q3에서 사분위수 범위의 1.5배 멀리 떨어진 관측치는 **이상치로 판단**할 수 있는 기준이 됨 '사분위수 범위'가 작고, '범위(MAX-MIN)'가 크면 이상치가 많다는 의미 '범위' 대비 '사분위수 범위'의 위치에 따라 데이터가 어느 쪽으로 쏠려있는지 알 수 있음

기술통계 > 쏠림

사분위수와 사분위수 범위를 확인하여 데이터가 약간 아래로 쏠려있음을 알 수 있음

	키(㎝)
1	170
2	170
3	173
4	173
5	173
6	176
7	178
8	178
9	180
10	181

0% (= 최소값)

25% (Q1)

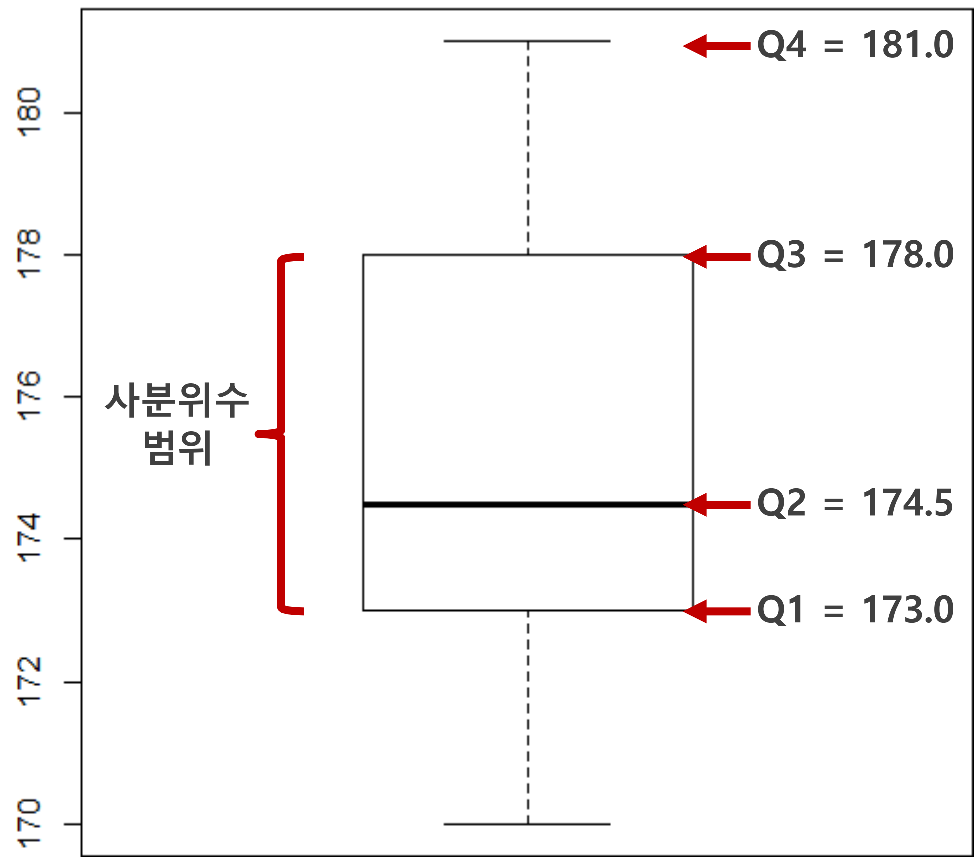
50% (Q2 = 중앙값)

75% (Q3)

100% (Q4 = 최대값)

사분위수 범위(IQR) = Q3 – Q1
= 178 – 173 = 5

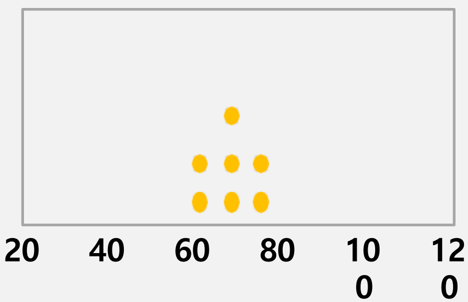
박스플롯 (Box Plot)



Q) A社의 2017년, 2018년 승진 시험 결과 점수입니다.
기술통계량을 활용하여 시험 결과를 비교하시오.

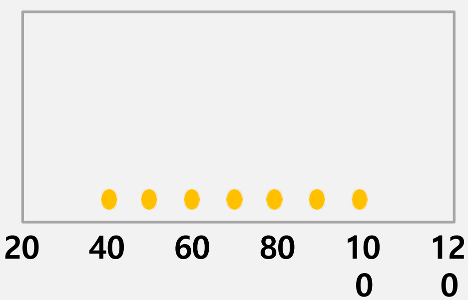
								평균
2017년	65	65	70	70	70	75	75	70
2018년	40	50	60	70	80	90	100	70

2017년



평균 : 70
표준편차 : 4.08

2018년



평균 : 70
표준편차 : 21.6

평균은 70으로 동일

표준편차는 2018년 시험이 크게 나타남

2018년도 시험의 변별력이 더 큼

기술통계 활용 > 두 집단 비교 (평균, 분산 활용)

당신이라면 A매장과 B매장 중에서 어디에 투자하겠는가? 그 이유/근거는?

어느 매장에 투자하면 좋을까?

날짜	A점 매출액 (단위: 만 원)	편차 (=하루 매출액–평균)	편차 제곱
1	49	-1	1
2	53	3	9
3	48	-2	4
4	47	-3	9
5	49	-1	1
6	43	-7	49
7	49	-1	1
8	52	2	4
9	50	0	0
10	51	1	1
11	52	2	4
12	57	7	49
13	48	-2	4
14	52	2	4
15	50	0	0
평균	**50**	**편차 합계 = 0**	**편차 제곱합 = 140**
분산	**9.3 (= 140/15)**	**표준편차 3.1 (= sqrt(9.33))**	

날짜	B점 매출액 (단위: 만 원)	편차 (=하루 매출액–평균)	편차 제곱
1	55	5	25
2	48	-2	4
3	49	-1	1
4	60	10	100
5	45	-5	25
6	44	-6	36
7	43	-7	49
8	55	5	25
9	44	-6	36
10	44	-6	36
11	61	11	121
12	60	10	100
13	54	4	16
14	48	-2	4
15	40	-10	100
평균	**50**	**편차 합계 = 0**	**편차 제곱합 = 678**
분산	**45.2 (= 678/15)**	**표준편차 6.7 (= sqrt(45.2))**	

* 데이터 출처: 퇴근시간이 빨라지는 비즈니스 통계입문

기술통계 활용 > 두 집단 비교 (사분위수 활용)

당신이라면 A매장과 B매장 중에서 어디에 투자하겠는가? 그 이유/근거는?

A SHOP

A Shop의 매출 범위

날짜	A점 매출액 (단위: 만 원)
1	49
2	53
3	48
4	47
5	49
6	43
7	49
8	52
9	50
10	51
11	52
12	57
13	48
14	52
15	50

매출 기준 Sort (오름차순)

순서	A점 매출액 (단위: 만 원)
1	43
2	47
3	48
4	48
5	49
6	49
7	49
8	50
9	50
10	51
11	52
12	52
13	52
14	53
15	57

Min.

1st Qu.

Median

3rd Qu.

Max

B Shop의 매출 범위

BSHOP

OPEN

날짜	B점 매출액 (단위: 만 원)
1	55
2	48
3	49
4	60
5	45
6	44
7	43
8	55
9	44
10	44
11	61
12	60
13	54
14	48
15	40

매출 기준 Sort (오름차순)

순서	B점 매출액 (단위: 만 원)
1	40
2	43
3	44
4	44
5	44
6	45
7	48
8	48
9	49
10	54
11	55
12	55
13	60
14	60
15	61

Min.

1st Qu.

Median

3rd Qu.

Max

A Shop 매출 범위

Min	1st Qu.	Median	3rd Qu.	Max
43	48.5	50	52	57

B Shop 매출 범위

Min	1st Qu.	Median	3rd Qu.	Max
40	44	48	55	61

기술통계 활용 > 변동 계수

절대 크기가 다른 두 집단, 측정단위가 다른 두 변수간의 산포를 비교할 때는 『변동계수』를 쓰자!

규모가 다른 두 가게의 매출 변동 비교

A 슈퍼마켓 매출 현황

- 하루 평균 매출 : 100만원
- 하루 매출 표준편차 : 30만원

$$변동계수 = \frac{표준편차}{평균} = \frac{30}{100}$$

$= 0.3$

B 편의점 매출 현황

- 하루 평균 매출 : 30만원
- 하루 매출 표준편차 : 10만원

$$변동계수 = \frac{표준편차}{평균} = \frac{10}{30} =$$

0.33

B 편의점의 매출 변동이 더 크다.

규모가 다른 두 회사의 주가 변동 비교

A 회사 주가 현황

- 6개월 평균 주가 : 500원
- 6개월 주가 표준편차 : 100원

$$변동계수 = \frac{표준편차}{평균} = \frac{100}{500}$$

$= 0.2$

B 회사 주가 현황

- 6개월 평균 주가 : 3,000원
- 6개월 주가 표준편차 : 300원

$$변동계수 = \frac{표준편차}{평균} = \frac{300}{3000}$$

$= 0.1$

A 회사의 주가 변동이 더 크다.

* 데이터 출처: 퇴근시간이 빨라지는 비즈니스 통계입문, 우치다 마나부 저

확률분포 > 정규분포

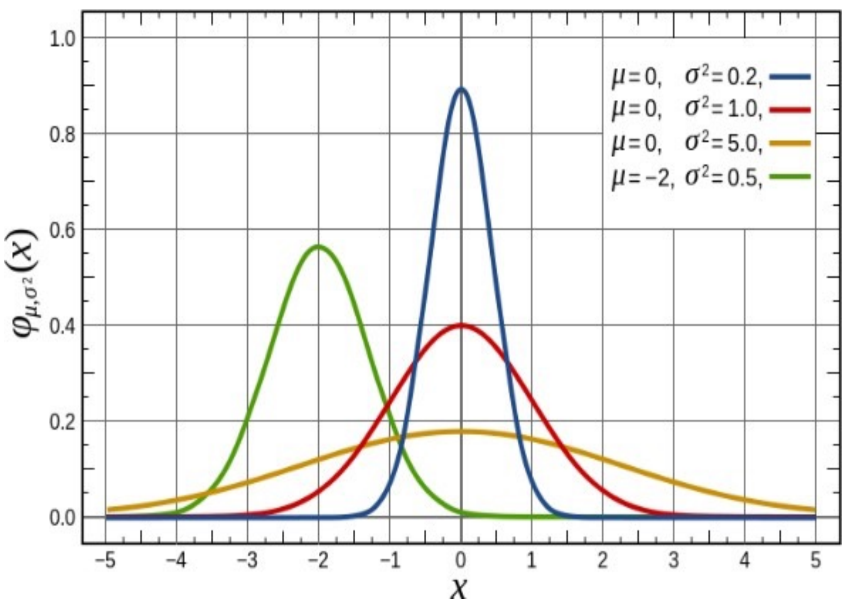
정규분포는 가장 많이 사용되는 연속형 확률분포이며, 많은 통계기반 알고리즘이 정규분포를 가정하고 만들어졌고, ML에서도 (데이터 분포에 의존하지 않지만 데이터가 특정한 특성을 가지고 있다면) 잘 동작하므로 정규분포 변수로 작업하는 것이 바람직함

정규분포 정의

확률밀도함수로

간단하게 N(μ, σ²)로 표기함

μ : 평균
σ : 표준편차
σ² : 분산



* 정규분포 그래프 출처 : http://nakyungpapa.tistory.com/215

정규분포의 특징 및 성질

평균(μ)을 중심으로 좌우 대칭인 종 모양 곡선

x = 평균에서 값이 최대(발생확률이 가장 높음)

평균값 = 중앙값 = 최빈값

곡선과 x축 사이의 면적은 1 (100%를 의미)

확률변수 X가 정규분포 N(μ, σ2)을 따른다면,
X의 1차함수 aX+b는 정규분포 N(aμ+b, a2σ2)을 따름

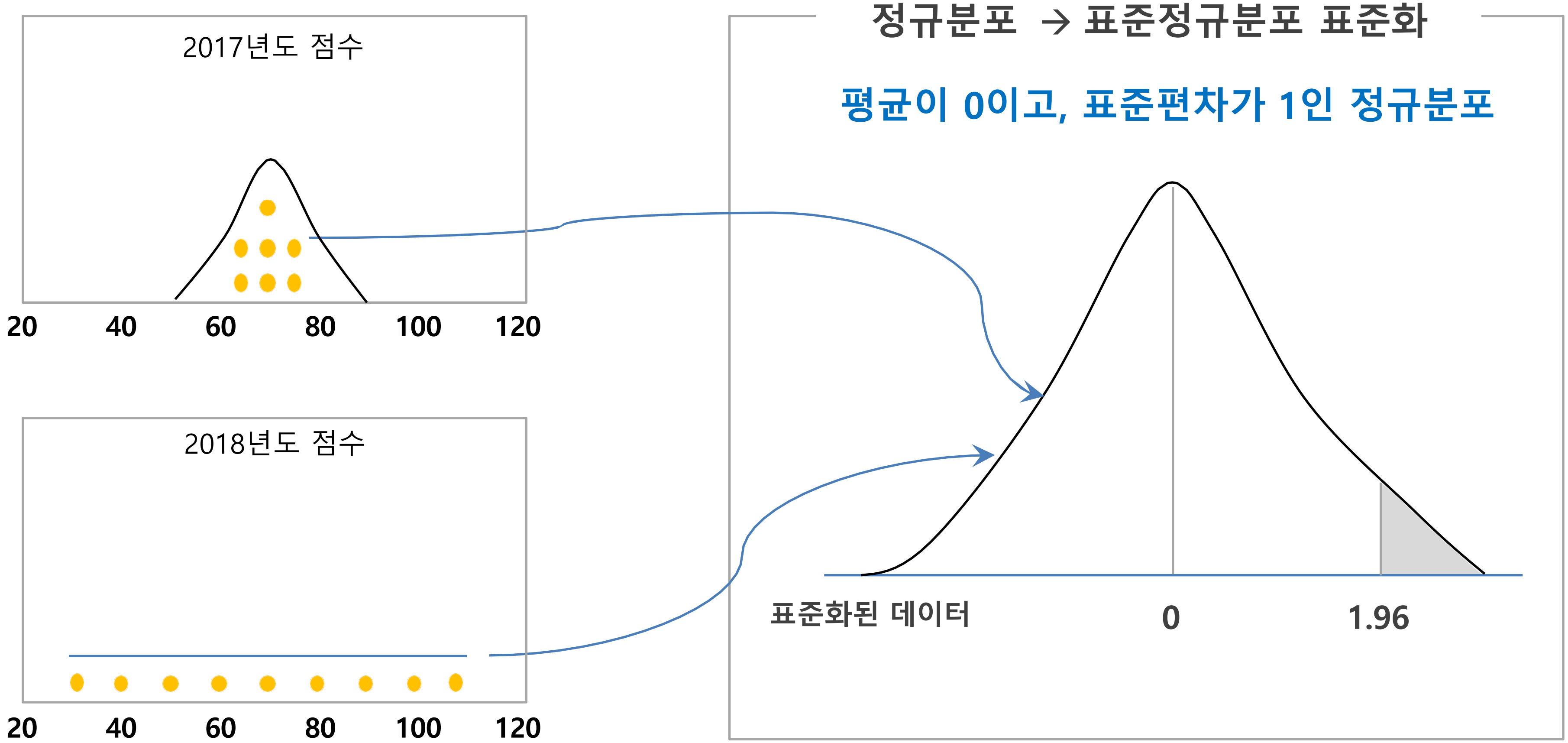
서로 독립인 두 확률변수 X, Y가 각각 정규분포 $N1(\mu_1, \sigma_1)$, $N2(\mu_2, \sigma_2)$를 따른다면,
aX+bY는 정규분포 $N(a\mu_1+b\mu_2, a\sigma_1+b\sigma_2)$를 따름

Scaling > 표준화(Standardization)

표준화는 평균과 표준편차가 다른 두 개의 집단을 서로 비교 가능하도록 단위를 통일 시켜주는 작업. 평균과 표준편차를 알면 표준화 할 수 있음

$$표준점수(Z\ Score) = \frac{(x - \mu)}{\sigma}$$

(μ: 평균, σ: 표준편차)



표준화

Q) 아래는 빅데이터팀 A/B 두 파트장의 팀원 1차 평가 점수입니다. 팀장은 전체 10명 중 6명을 승진 시키고자 합니다.

A파트	70	75	80	85	90
B파트	10	65	70	75	80

평균 : 80, 표준편차 : 7.9

평균 : 60, 표준편차 : 28.5

1 안

A 파트 70 75 80 85 90

B 파트 10 65 70 75 80

표준화 점수 $= \dfrac{(x - \text{평균})}{\text{표준편차}}$

2 안

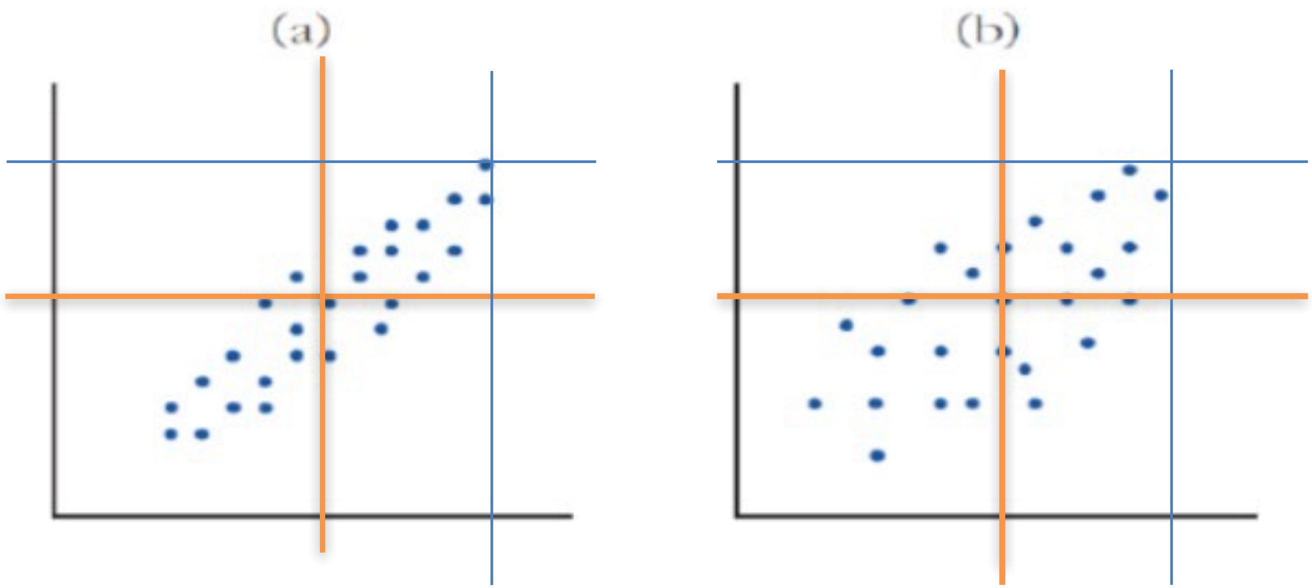
A 파트 -1.3 -0.6 0 0.6 1.3

B 파트 -1.8 0.2 0.4 0.5 0.7

상관분석

변수간의 관계(x의 변화에 따른 y의 변화)를 파악하는 데 있어서 평균과 표준편차만으로는 충분하지 않음

가로든 세로든 평균과 표준편차가 동일해도 두 변수 관계는 상이



위의 두 산포도는 가로든 세로든 중심과 퍼진 정도가 동일하지만 (a)가 (b)보다 더 강한 선형관계를 보임

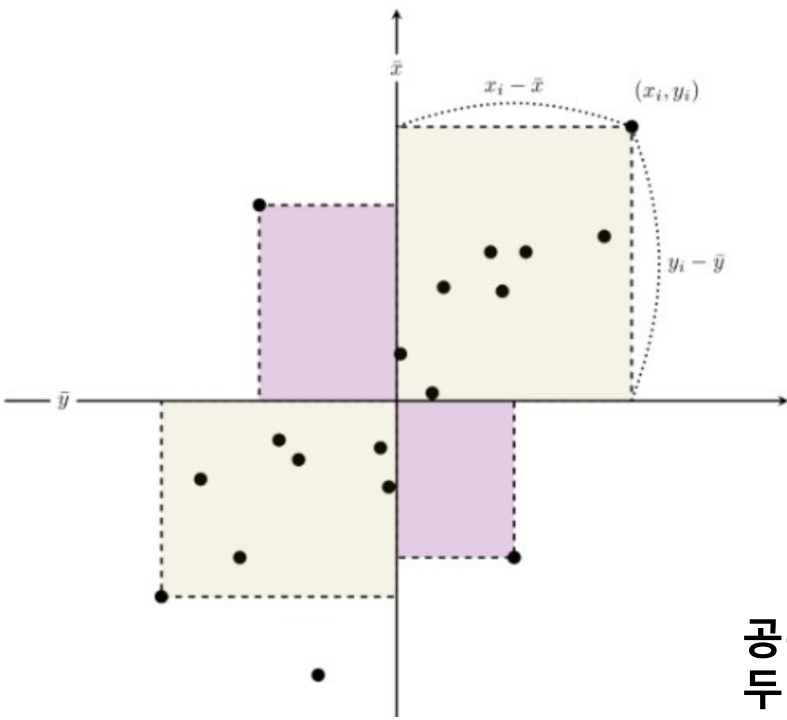
두 변수간 선형관계의 방향과 강도가 얼마나 되는지 측정할 필요성 있음

공분산(Covariance)과 상관계수(Correlation)

두 수치형 변수의 관계를 계산한 기술 통계량

공분산(Covariance)

두 변수가 각각의 평균에 대해서 얼마나 떨어져 있는지를 수치화 한 것



$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

(μ_x : x의 평균, μ_y: y의 평균)

- ✓ $Cov(X,Y) > 0$ X가 증가 할 때 Y도 증가 (양의 상관관계)
- ✓ $Cov(X,Y) < 0$ X가 증가 할 때 Y는 감소 (음의 상관관계)
- ✓ $Cov(X,Y) = 0$ 두 변수간에는 아무런 선형관계가 없음

공분산은 값의 범위가 정해져 있지 않고,
두 변수의 단위에 의존하여 다른 데이터와 비교시 불편
è 상관계수 활용

공분산(Covariance)과 상관계수(Correlation)

두 수치형 변수의 관계를 계산한 기술 통계량

상관계수(Correlation Coefficient)

공분산을 정규화 하여 그 결과가 -1 ~ 1 범위의 값을 가지도록 한 것

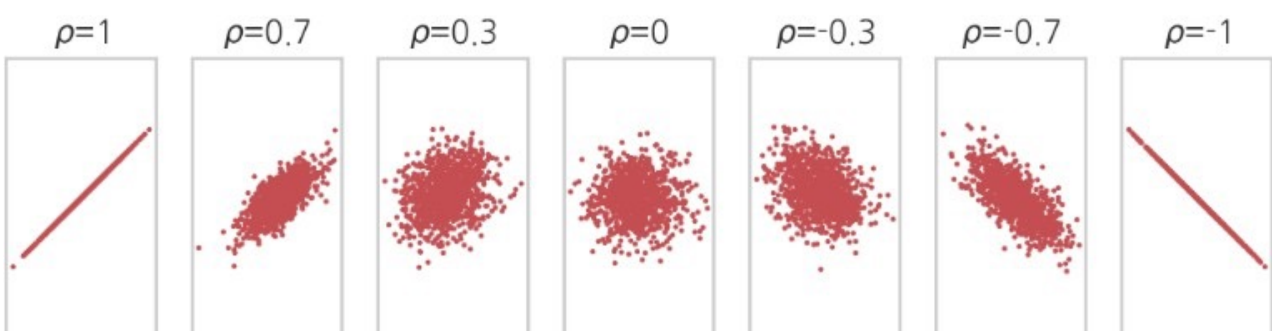
가장 많이 사용하는 값이 피어슨 상관계수임

$$Corr(X,Y) = \frac{Cov\ (X,Y)}{\sigma_x \sigma_y}$$

(σ_x: x의 표준편차, σ_y: y의 표준편차)

- ✓ 양의 값이면 두 변수가 같은 방향으로 움직임
- ✓ 음의 값이면 두 변수가 반대 방향으로 움직임
- ✓ 0이면 선형관계가 없음

상관계수와 스캐터 플롯의 모양



공분산과 상관계수 예제

공부시간(x)과 점수(y)의 상관관계

공부시간(x)	점수(y)	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)(y_i - \mu_y)$
0	60	-5	-20	100
4	78	-1	-2	2
3	83	-2	3	-6
6	74	1	-6	-6
6	100	1	20	20
7	80	2	0	0
8	90	3	10	30
8	85	3	5	15
3	70	-2	-10	20
평균 : 5	평균 : 80			합계 : 175
표준편차 : 2.693	표준편차 : 11.587			

공분산

$Cov(X,Y) = \frac{175}{9-1} = 21.875$ (**시간*점수**)

크기와 단위의 문제로 다른 데이터와 비교시 불편!

상관계수

$Corr(X,Y) = \frac{21.879}{2.693 * 11.587}$ = **0.7**

공부시간과 점수는 양의 상관 관계를 가지고 있다.

Agenda

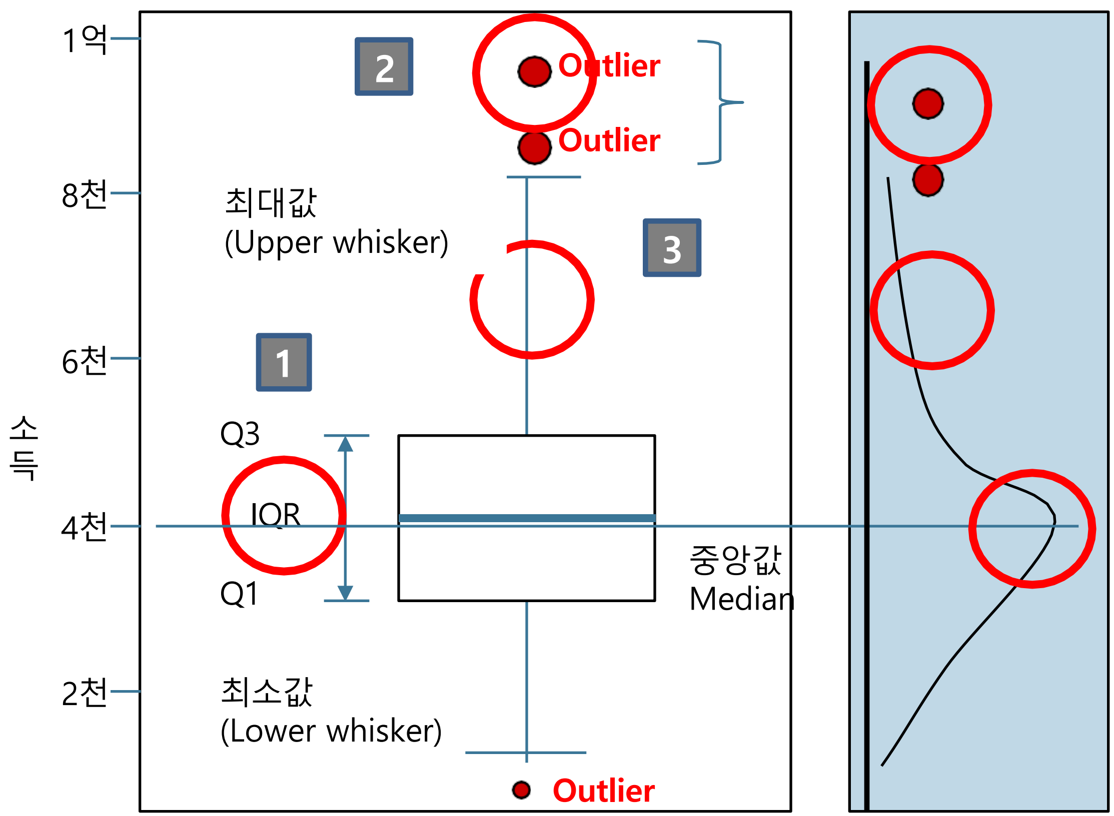
1. 기술통계(Descriptive Statistics)

2. 그래프 분석

Box Plot

사분위수를 통해 데이터 분포를 상자 형태로 나타내 중심화와 분포 파악

Box-and-whiskers plot (예시)



Box-and-whiskers plot으로 알 수 있는 것

1 중심경향 파악

중앙값 (median) 으로 중심 경향 파악

2 특이값/이상치(outlier) 발견

특이값(outlier): 중앙값에서 편차가 많은 값

IQR(Inter Quartile Range)*1.5배 벗어나는 상/하한 값 은 이상치

3 대칭성 및 편차의 발견

대칭성 (symmetry): 최대값과 최소값의 길이 비교 길이가 길수록 비대칭, 편차 큼

왜도 (skew): 중앙값으로 부터 편차 큼

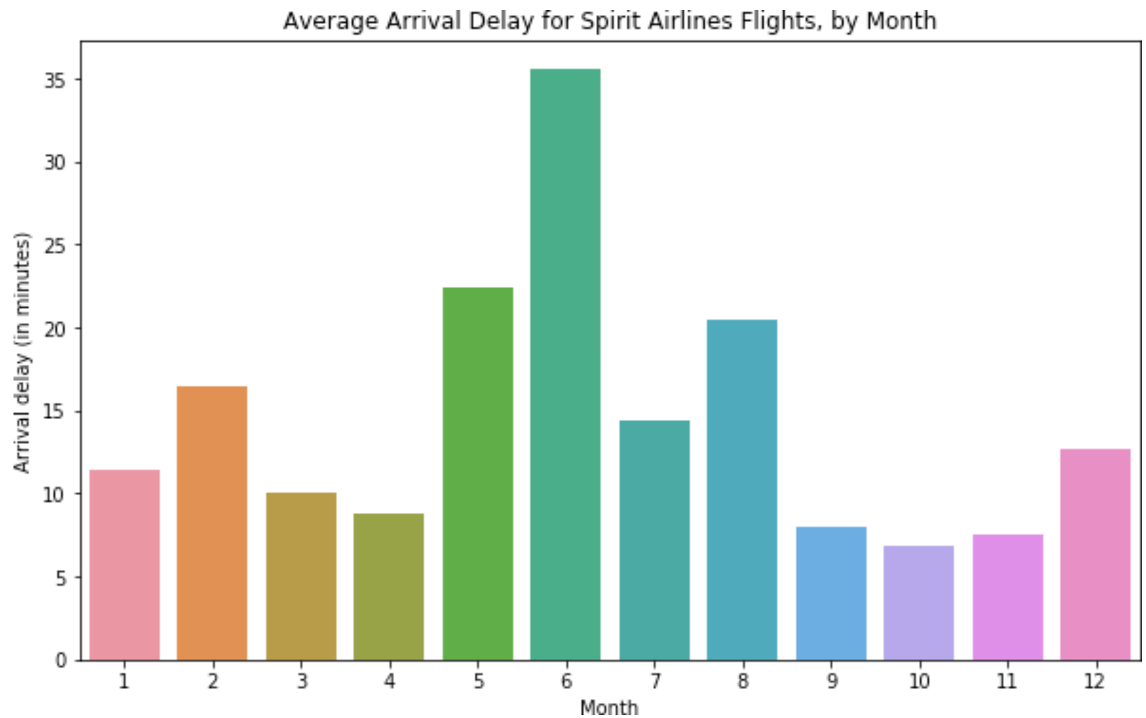
Bar Plot

범주형 데이터의 값의 크기(height)를 막대모양으로 나타난 그래프 (절대 크기 비교)

	Delay 수
1월	12
2월	17
3월	10
4월	9
5월	23
6월	35
	. . .
12월	15

<도수분포표>

월별 평균 Delay 수

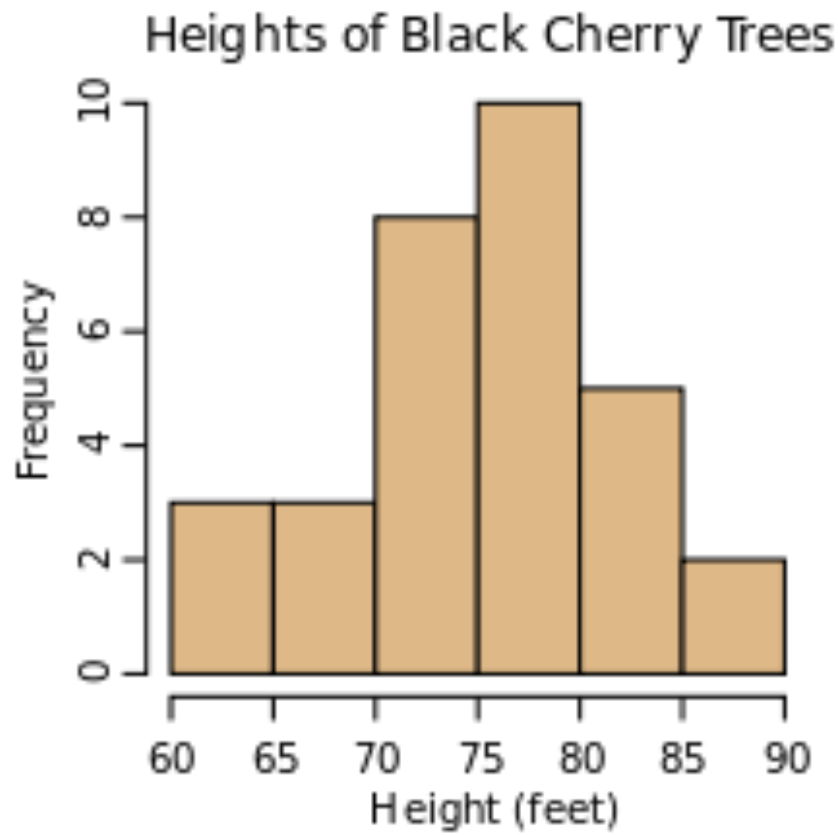


Histogram

연속형 데이터의 값의 범위 마다 데이터 빈도수(frequency)를 막대모양으로 표현
연속형 데이터를 계급으로 나누어 계급별 도수를 막대로 나타냄

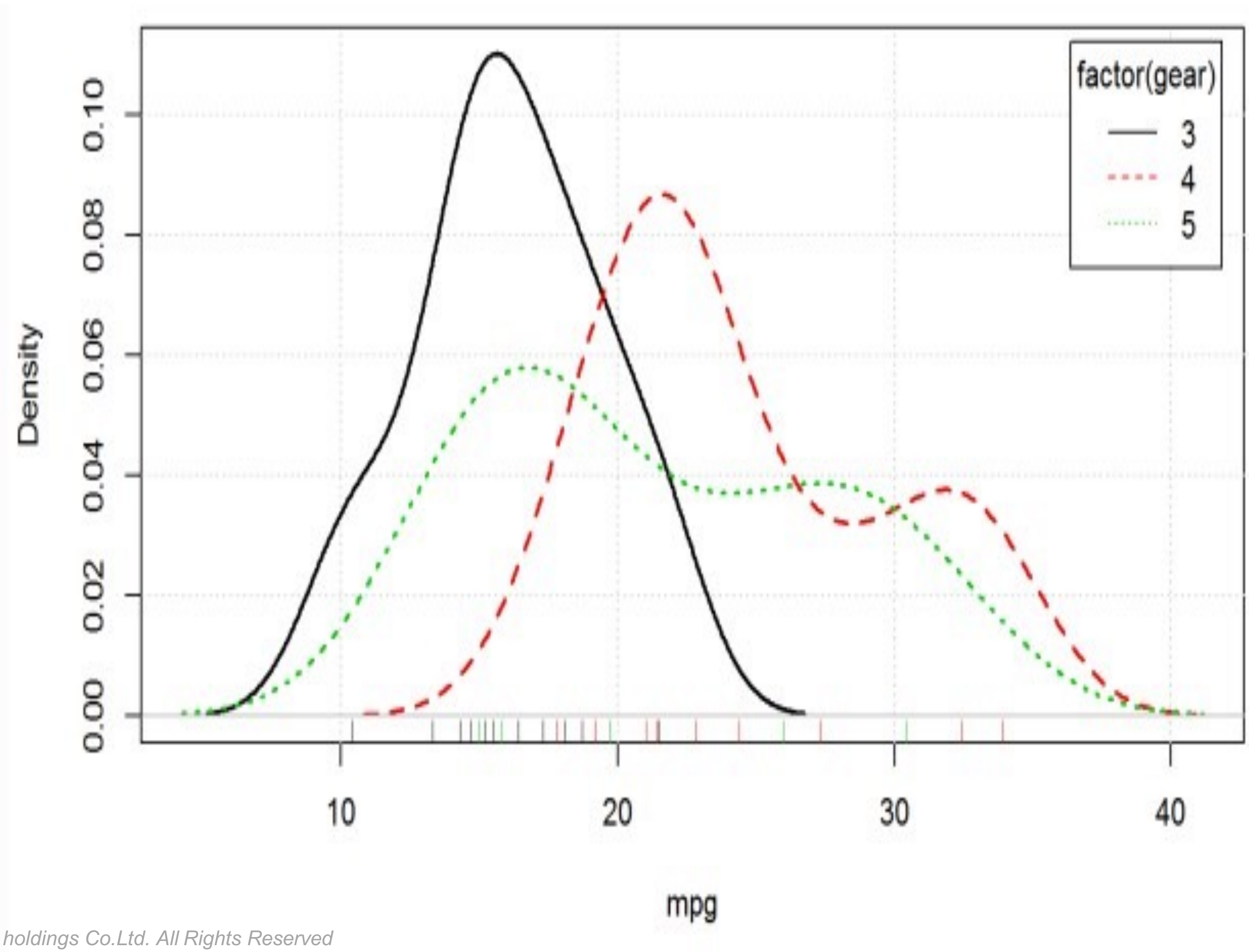
	빈도수
60~65	3
65~70	3
70~75	8
75~80	10
80~85	5
85~90	2

<도수분포표>



Density plot

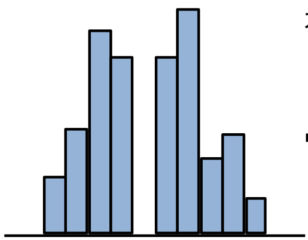
데이터 분포를 밀도함수로 나타낸 그림으로 히스토그램을 곡선화한 형태
변수의 분포, 평균 등을 시각화 함



분포를 통한 데이터 이해

데이터 분포는 정규분포가 이상적이나 다양한 유형의 분포가 발생하고 있으며, 분포 유형별 데이터 이해가 필요함

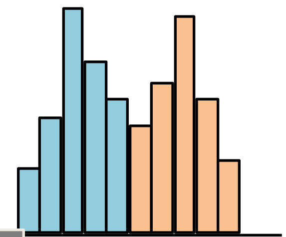
이빠진형



구간이 빠져 하나건너 뛰어 돗수가 적어지지는 경우 발생

➡구간폭의 측정단위 문제

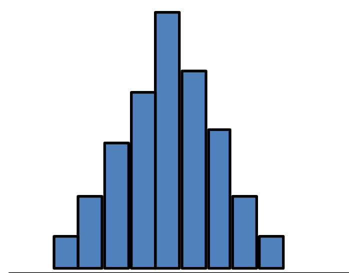
쌍봉우리 형



평균치가 다른 두 개의 분포가 혼합되어 있는 경우 발생
예) 두 대의 장비간에 차이가 있는 경우.

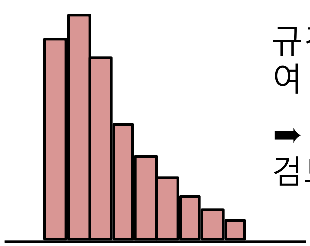
➡ 장비 종류에 따라 분리하여 두 개 장비 유형별 분석 필요

정규분포형



돗수는 중심부근이 가장 많으며 중심에서 멀어질수록 작아짐.

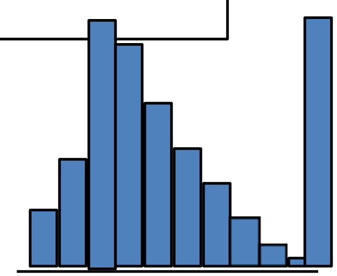
절벽형



규격이하의 것을 전수 선별하여 제거했을 경우 등에 발생

➡ 측정의 오류, 측정오차 등 검토 필요

좌우 비대칭 형



평균치가 분포의 중심에서 좌측으로 치우쳐있음

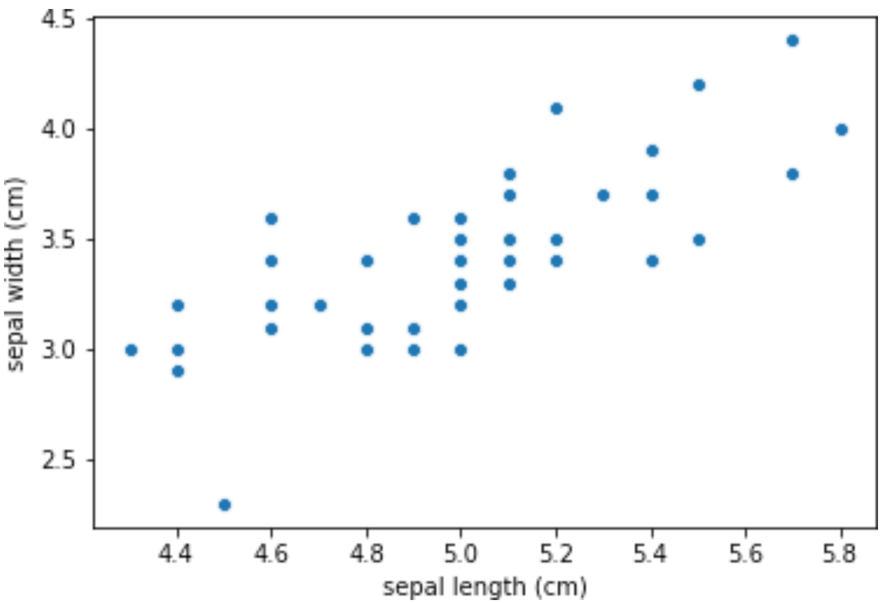
➡하한이 억제되고 있고 특정 값 이하는 취하지 않는 경우

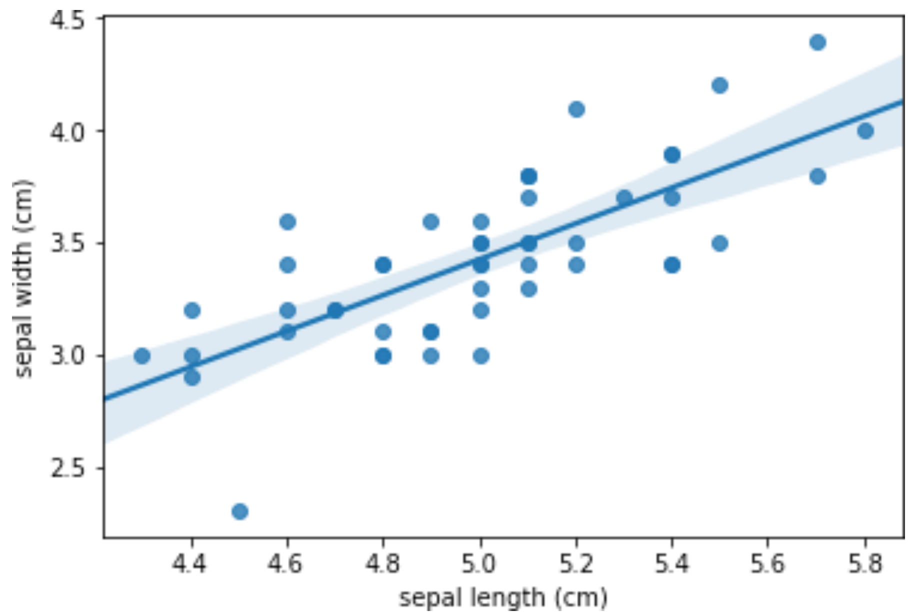
오른쪽 바는 상한선을 의미함

Scatter Plot

데이터 분포를 점으로 표현하여 파악. 두 변수간 상관 관계 탐색에 유용

예측/분류 모델링을 위한 유효 변수 탐색에 유용함

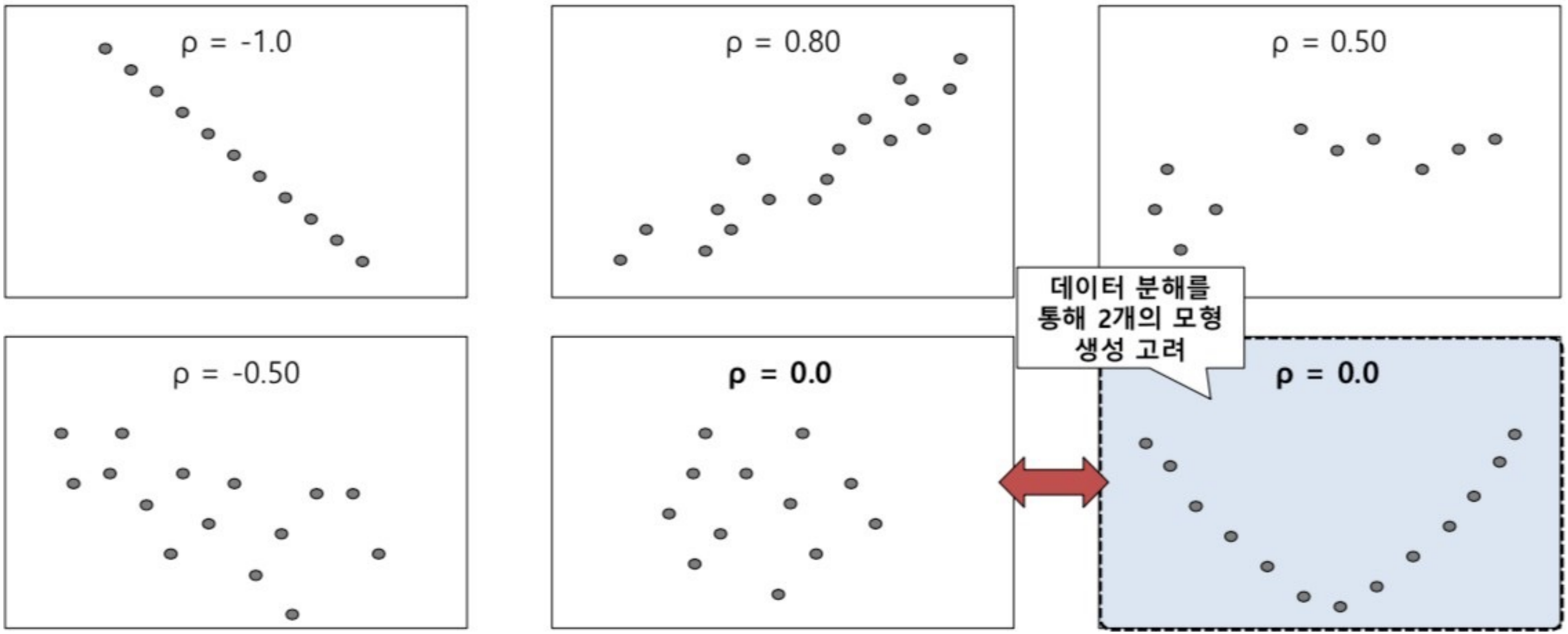




Scatter Plot과 상관계수

상관계수는 선형적인 관계만을 알 수 있으며 곡선관계는 알 수 없으므로, 상관계수 수치만 보고 판단하지 말고 그래프 분석을 병행할 필요가 있음

여러 가지 상관계수 ρ 값에 대한 산포도 (예시)



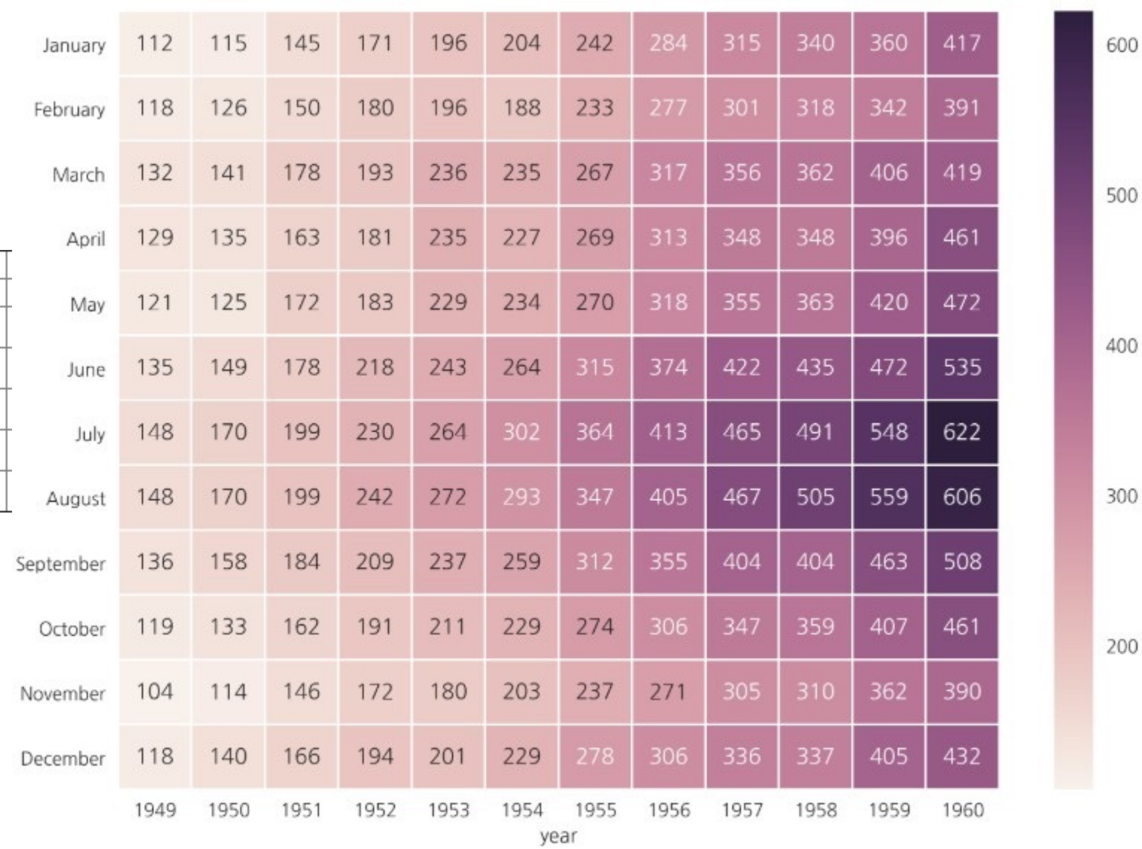
Heatmap

데이터의 상하 위계 구조를 열분포 형태의 비주얼한 그래픽으로 표현

항목이 많을 때 직관적으로 파악하기 좋음 변

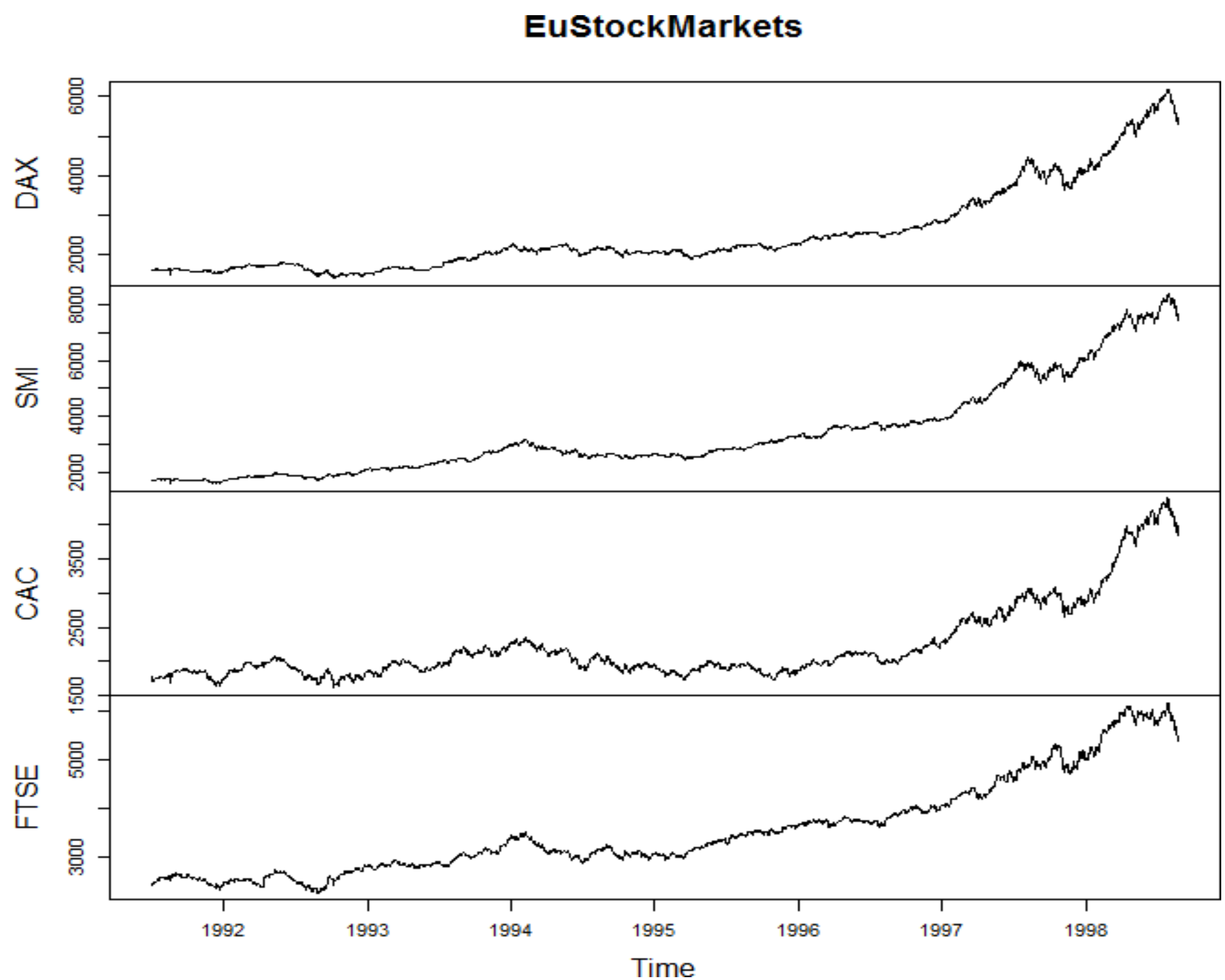
수간 상관관계 비교 분석에도 사용할 수 있음

year	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958
month										
August	148	170	199	242	272	293	347	405	467	505
September	136	158	184	209	237	259	312	355	404	404
October	119	133	162	191	211	229	274	306	347	359
November	104	114	146	172	180	203	237	271	305	310
December	118	140	166	194	201	229	278	306	336	337



Time-series plot (Line plot)

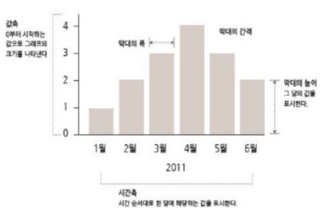
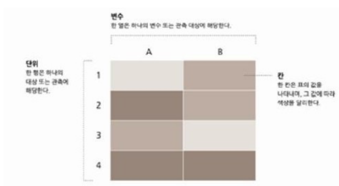
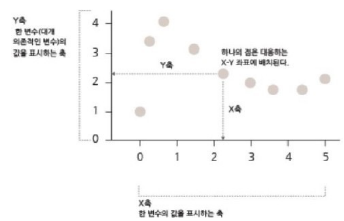
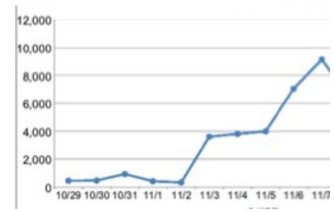
시간의 흐름에 따른 흐름, 추세, 계절성 등을 비교



데이터 시각화 유형

데이터를 통해 특정 유형으로 시각화 할 수 있으며, 시각화 목적을 고려하여 적용

<u>데이터 시각화 유형</u>

	분포시각화	비교시각화	관계시각화	시간 시각화	공간 시각화
목적	전체 그룹 중에서 부분이 차지하는 비율	데이터의 상하 위계(크기) 구조를 계층적으로 표현	X, Y 두개의 요인간의 관계를 표현	시간에 흐름에 따른 변화 표현 꺾은 선을 통해 경향 또는 추세 파악 가능	위치
시각화 기법	박스그래프/히스토그램 파이/도우넛 차트 누적 연속 그래프 	히트맴 스타 차트 	스케터 플롯 히스토그램 	시계열 도표(점 그래프) 막대 그래프 	지도상에 버블차트 등을 겹쳐서 표시 

고려사항	
	v 시각화 위계 구조 : 전체를 파악하고, 집중 하여할 요소를 고려하여 시각화
	v 시각화 요소의 우선순위 : 데이터 시각화 요소들이 사용자들이 명확히 데이터를 이해 할 수 있도록 "정확한 요소"부터 "부정확한"요소를 고려하여 시각화
	v 가독성 : 데이터의 비교가능성, 데이터의 맥락, 데이터의 의미를 고려한 시각화

Appendix. matplotlib

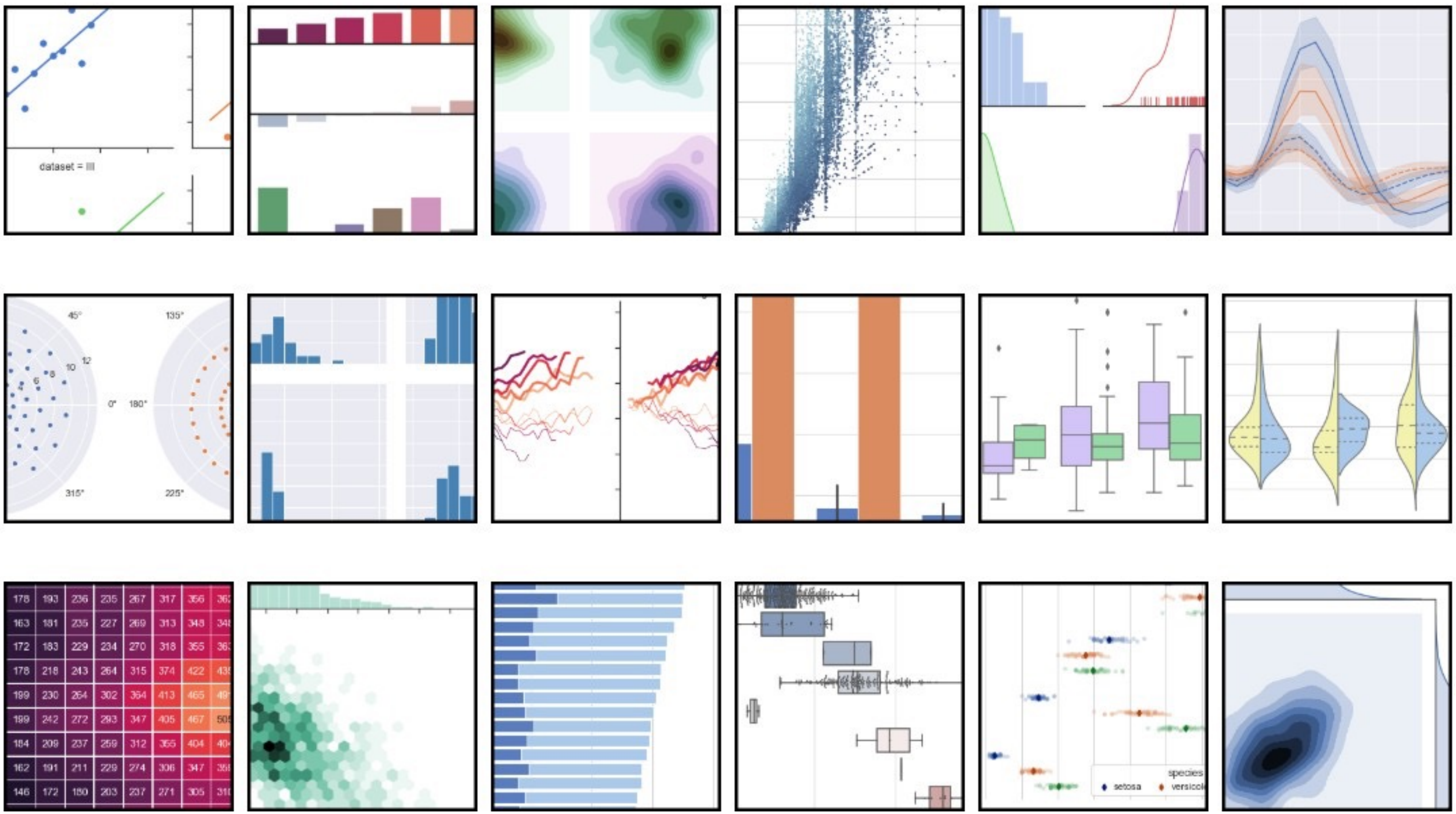
https://matplotlib.org/gallery/index.html

Appendix. Seaborn

https://seaborn.pydata.org/examples/index.html

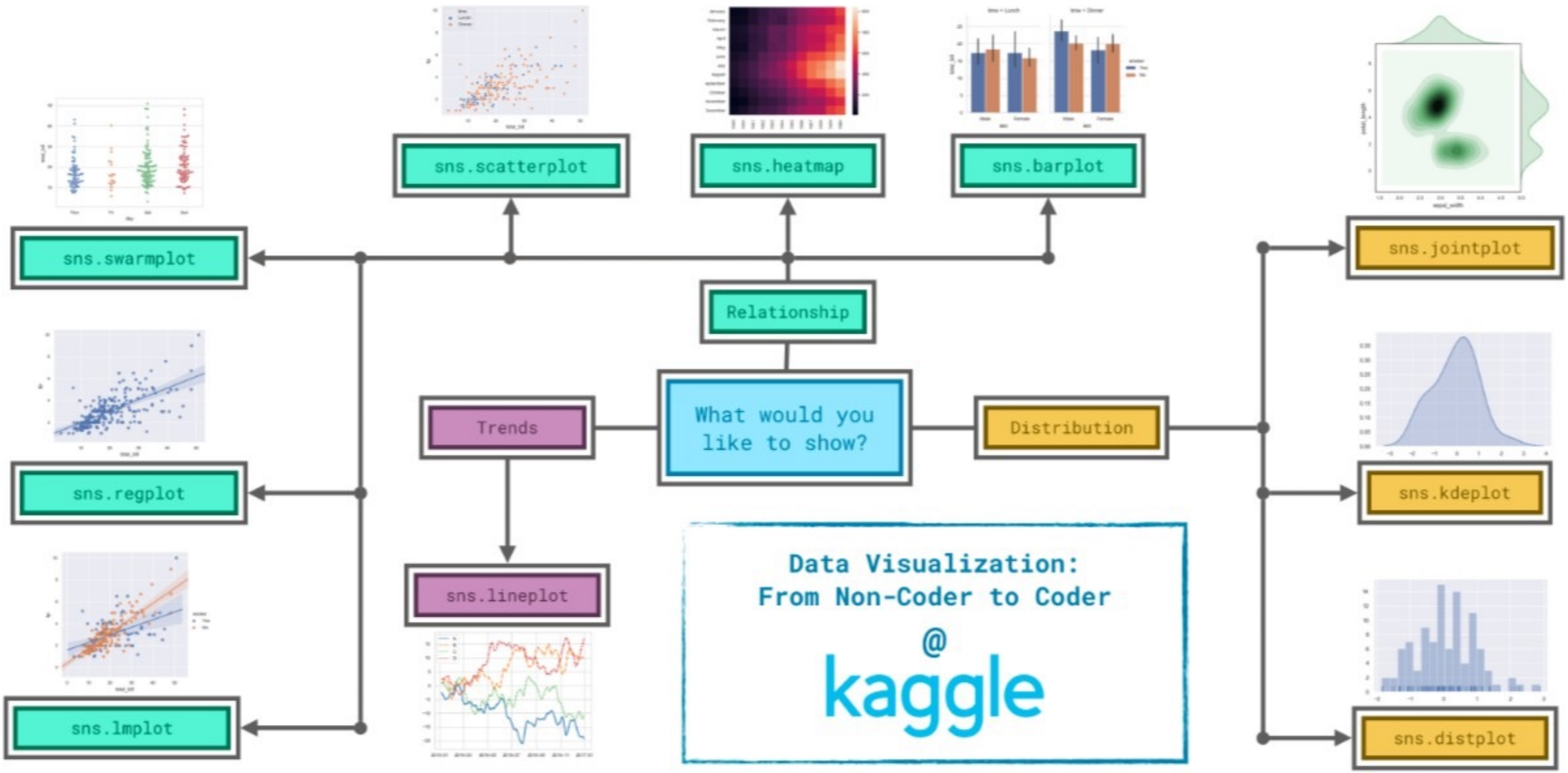
seaborn 0.9.0 Gallery Tutorial API Site Page Search

Example gallery



Appendix. Kaggle Micro-Learning > Data Visualization w/Seaborn

[Kaggle Micro-Learning](https://www.kaggle.com/learn/data-visualization-from-non-coder-to-coder)
https://www.kaggle.com/learn/data-visualization-from-non-coder-to-coder



Thank you