
회사에서
진짜 쓰는 통계

강사 소개

개발 → 증설 → 생산 → 생산관리 → IT 직무 경험

- 2005년 고려대 화공 생명 공학과 공정 시스템 석사 과정 수료
- 2006년 입사
  - 자동차용 Battery 연구소 팩 공정 개발팀 (대전)
- 2008년 자동차 Battery Pack 증설PJT → 팩 생산P (오창)
  - 자동차 Battery Pack 양산 라인 증설 (품질 Sys. 생산 Sys.)
  - 팩 생산관리 및 기획
- 2009년 자동차 Battery 생산관리 P (오창)
  - 자동차 Battery 생산 시스템 구축, 관리
- 2010년 자동차 Battery Global SCM 팀 (오창)
- 2011년 Battery PI(Process Innovation) 팀 (오창)
- 2013년 남경 소형 Battery PI팀 (남경, 주재원)
- 2018년 폴란드 자동차 배터리 PI 팀 (폴란드, 주재원)
- 2024년 ~ 현재 남경 자동차 배터리 PI팀 (남경, 주재원)
  "업무의 잔머리" 유튜브 & 블로그 운영 중

※ 저서 : 업무의 잔머리, 나는 땡땡이다. 공돌이 선배들의 해외생활 이야기,
일잘러의 UiPath 업무 자동화, AI 코딩을 위한 최소한의 파이썬

강의 순서

날짜	Chapter명	시간(분)	주요 내용
Day 1	1. 통계란?	10:00~10:50	통계에 대한 기본 이해, 사용되는 도구
	2. 기초 통계	11:00~11:50	기초 통계 지식
	점심	12:00~13:00	
	3. 통계 응용	13:00~13:50	데이터 시각화, 머신러닝
	4. 통계 실습	14:00~14:50	통계 분석 실습
	5. 통계 분석시 유의점	15:00~15:50	통계 분석시 유의점

강의에 앞서

참고 영상 리스트 : https://tricks-offic.notion.site/mp4links
구글 드라이브 링크 : https://drive.google.com/drive/folders/1cmagzHRbEWkUwqFlbNHBFMwJuW27Gf5l?usp=sharing

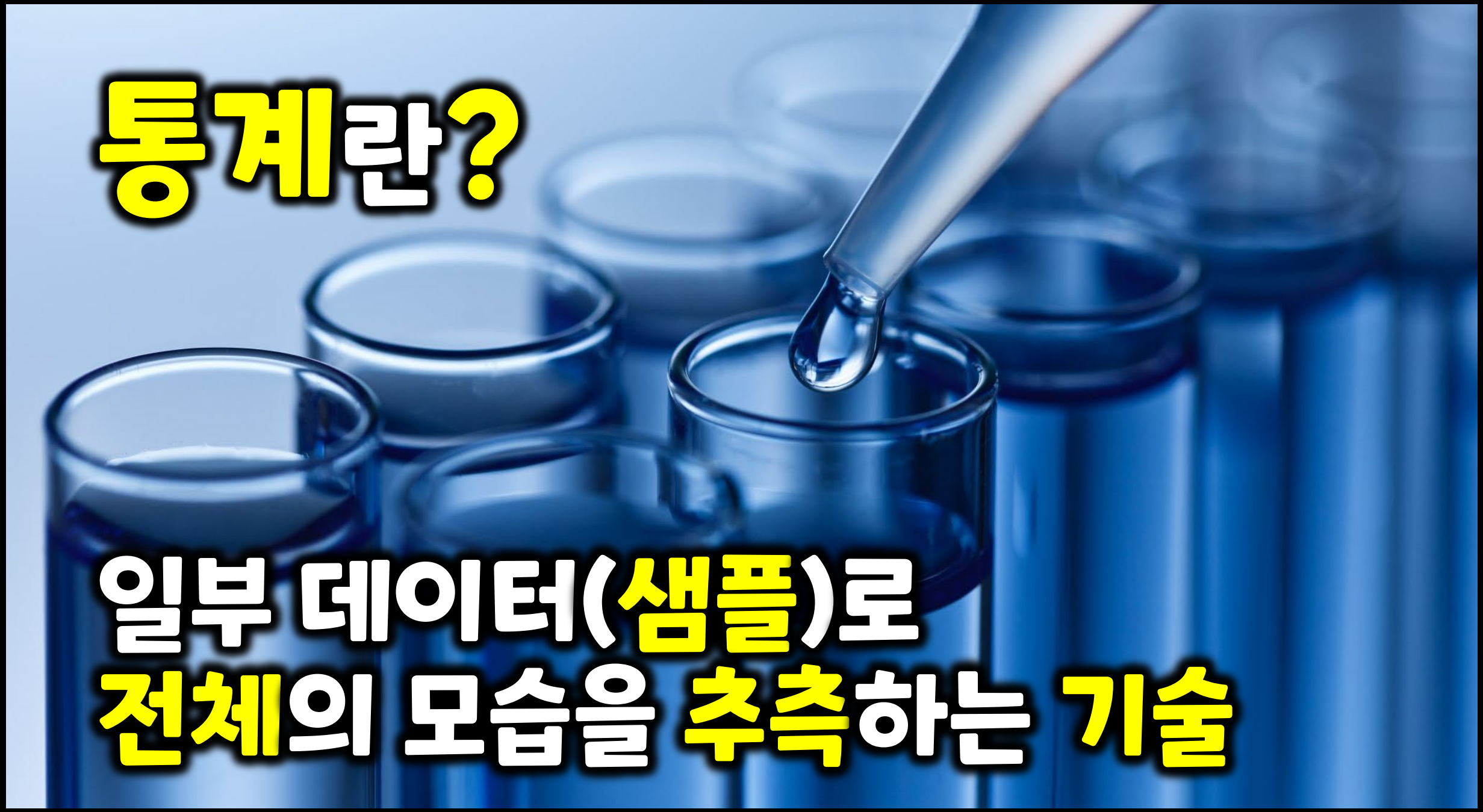
Orange Datamining : https://orangedatamining.com/

R :
- R : https://cran.r-project.org/bin/windows/base/
- R Studio : https://posit.co/download/rstudio-desktop/
- 시각화 갤러리 : https://r-graph-gallery.com/

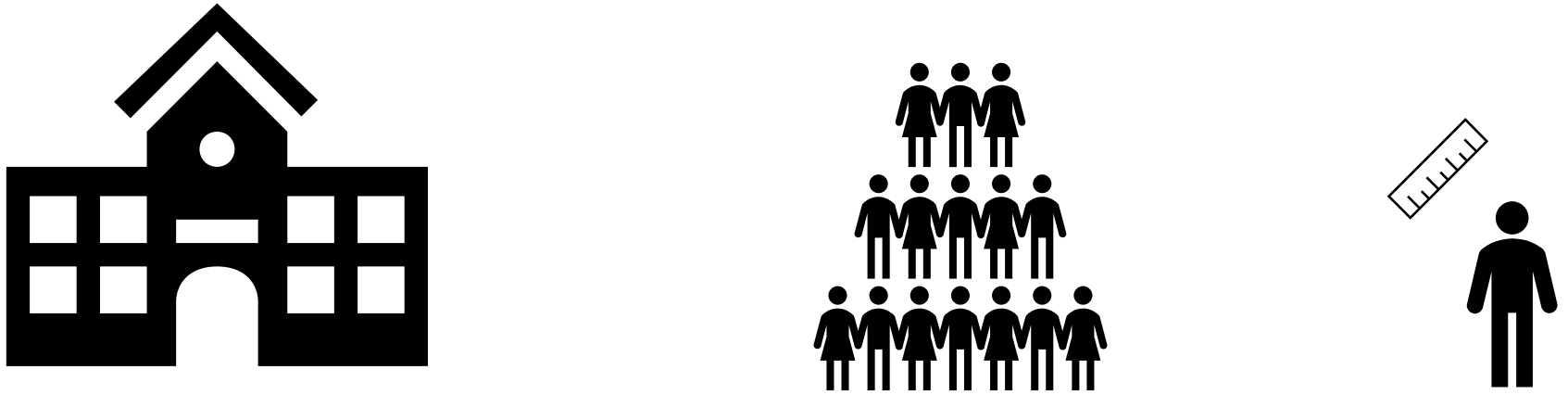
Python :
- Python : https://www.python.org/downloads/
- VC : https://code.visualstudio.com/download
- PyCharm : https://www.jetbrains.com/pycharm/download
- Colab : https://colab.research.google.com/?hl=ko
- 시각화 갤러리 : https://python-graph-gallery.com/

색상표 : https://color.adobe.com/

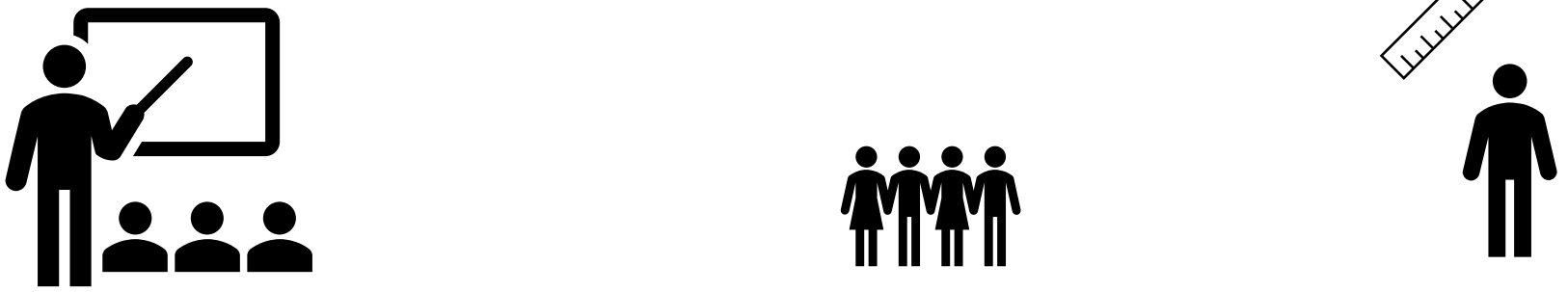
AI :
- ChatGPT : https://chatgpt.com/
- Claude : https://claude.ai/new
- Cursor : https://www.cursor.com/
- Perplexity : https://www.perplexity.ai/
- Genspark : https://www.genspark.ai/


통계란?
일부 데이터(샘플)로
전체의 모습을 추측하는 기술

진짜 알고 싶은 것을 빠르게(싸게) 알고 싶으니까



알고 싶은 것 : 학교 전체 인원의 키 분포



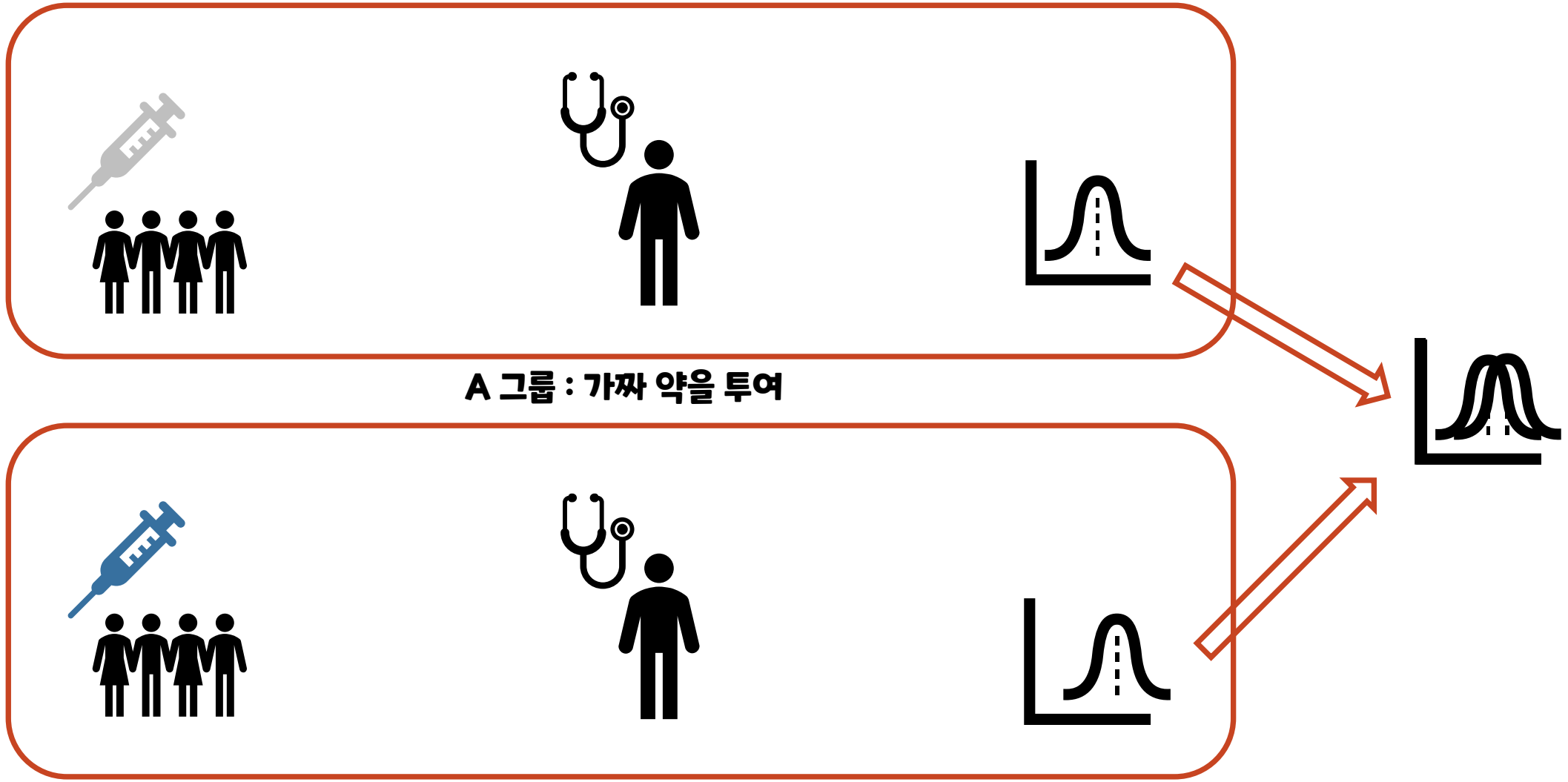
샘플 조사 : 한 반 인원의 키 분포

차이점을 빠르게(싸게) 비교하고 싶으니까

A 그룹 : 우유를 마시지 않는 학생들의 키

B 그룹 : 우유를 마시는 학생들의 키

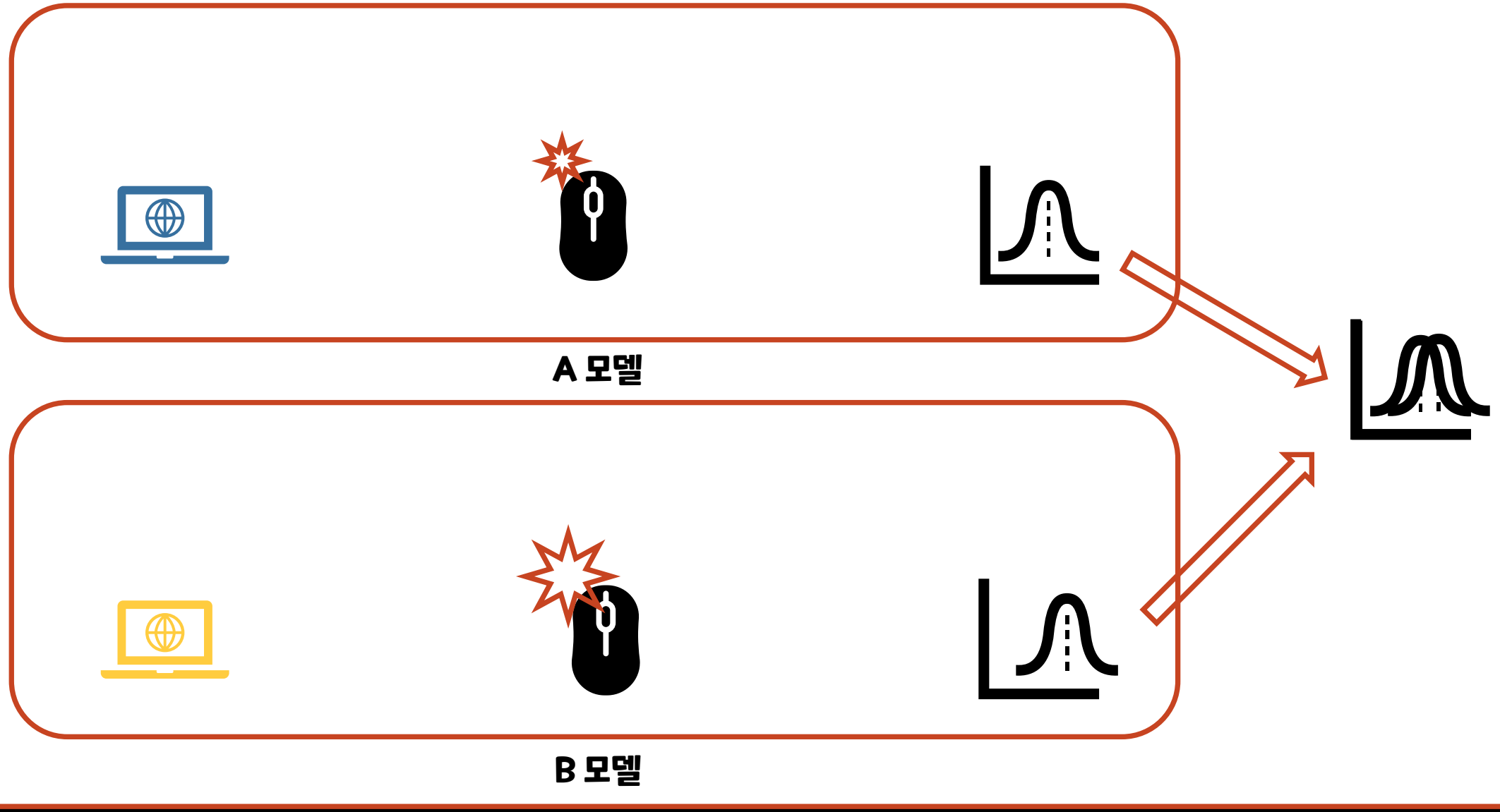
의학에서… (임상실험)



A 그룹 : 가짜 약을 투여

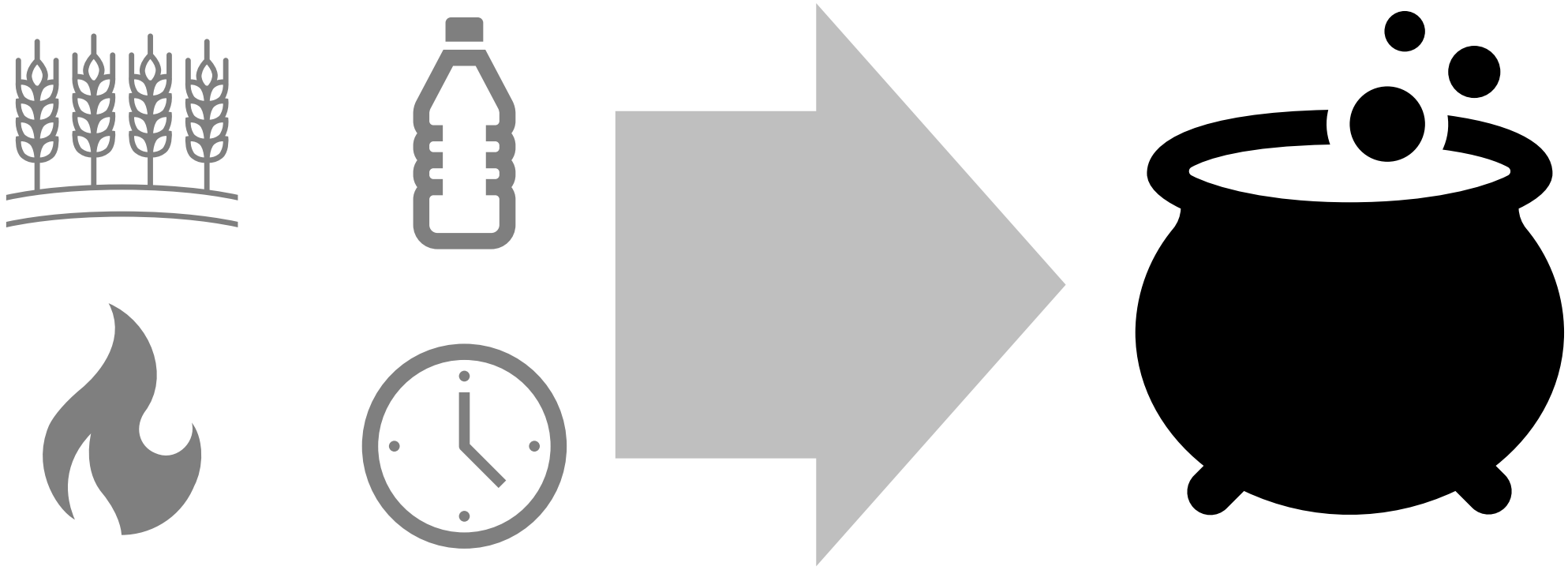
B 그룹 : 진짜 약을 투여

산업에서… (AB테스트)

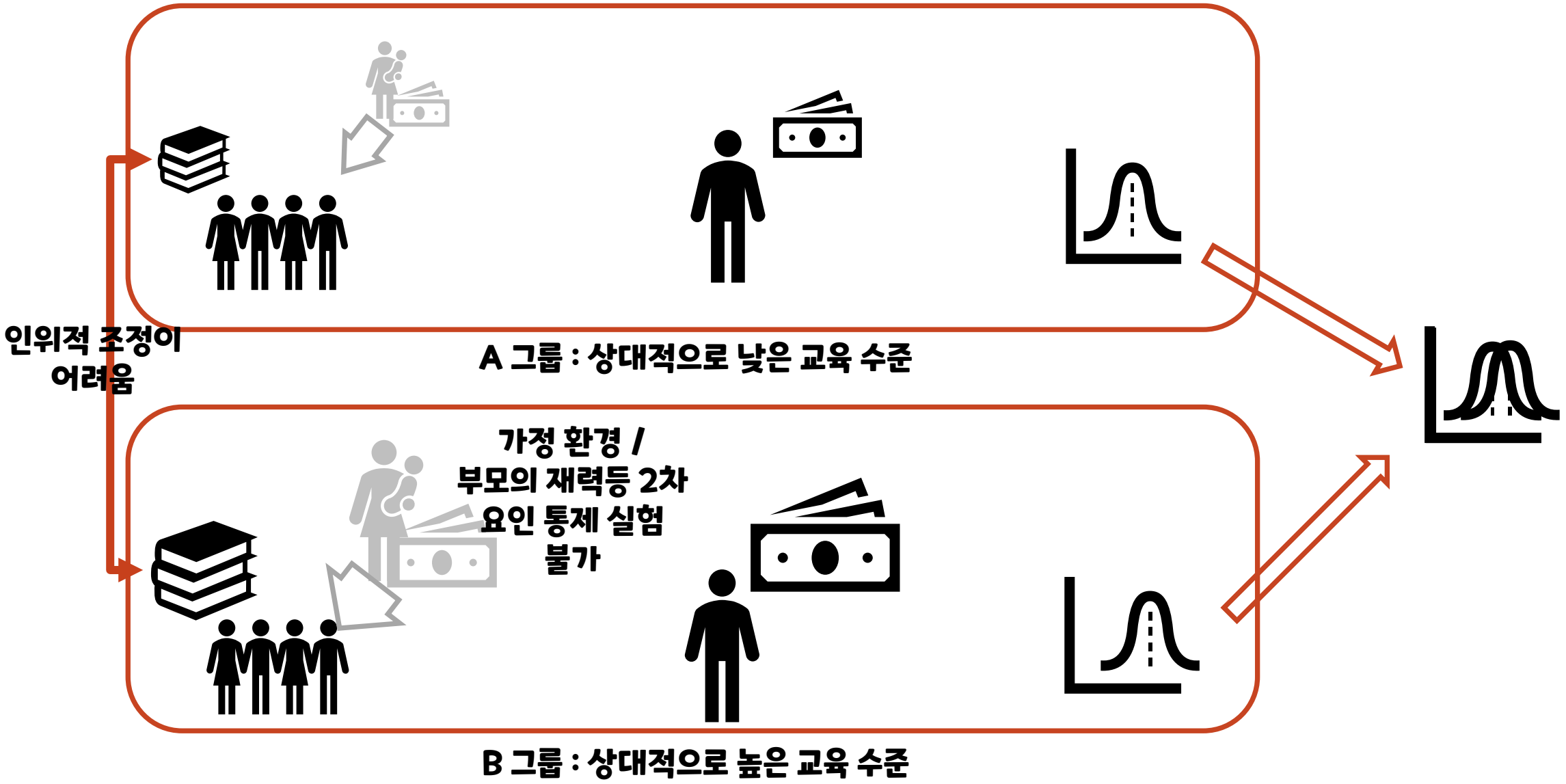



통계는
문과?
이과?

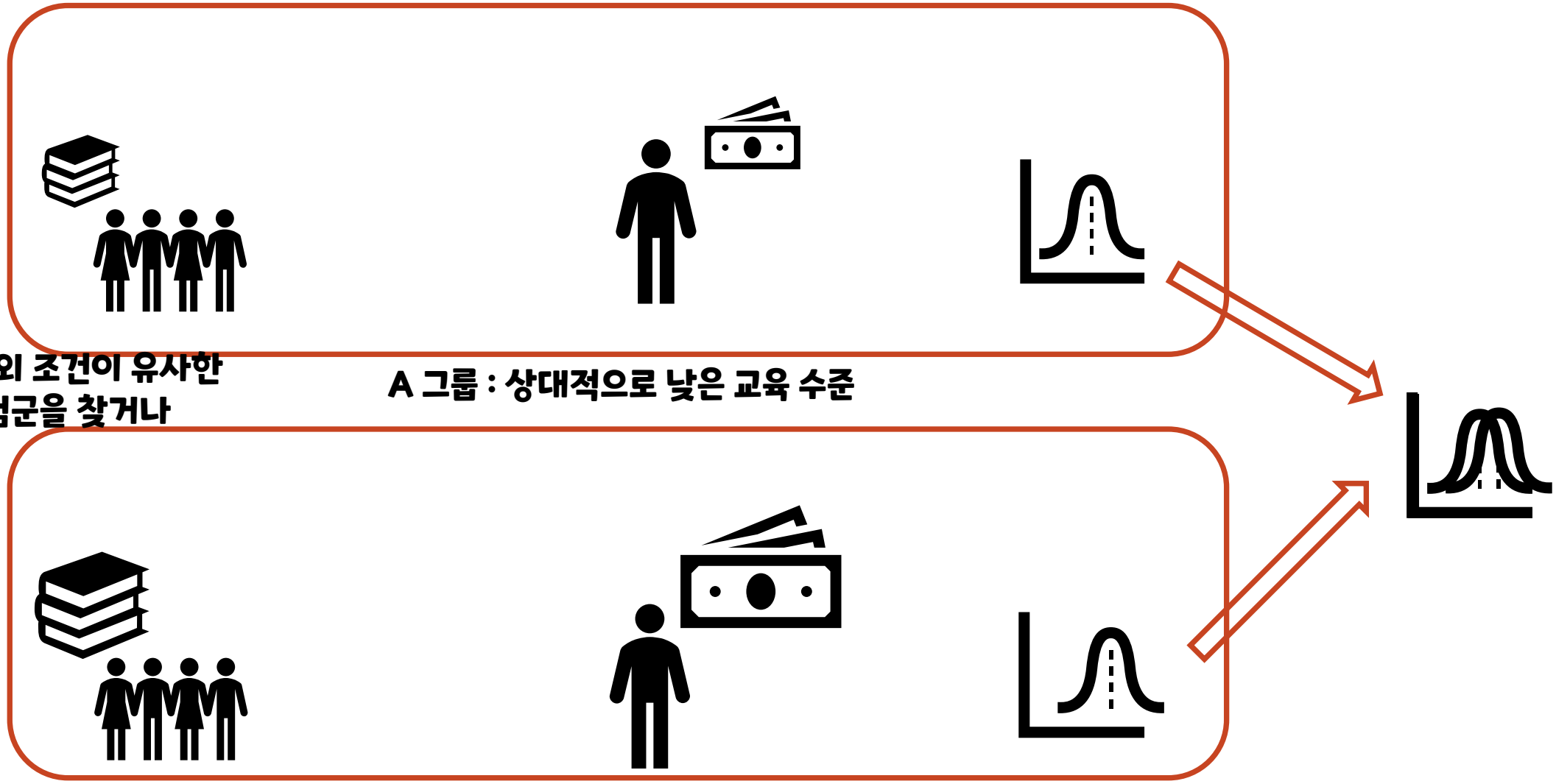
자연 과학에서는 통계가 없더라도…



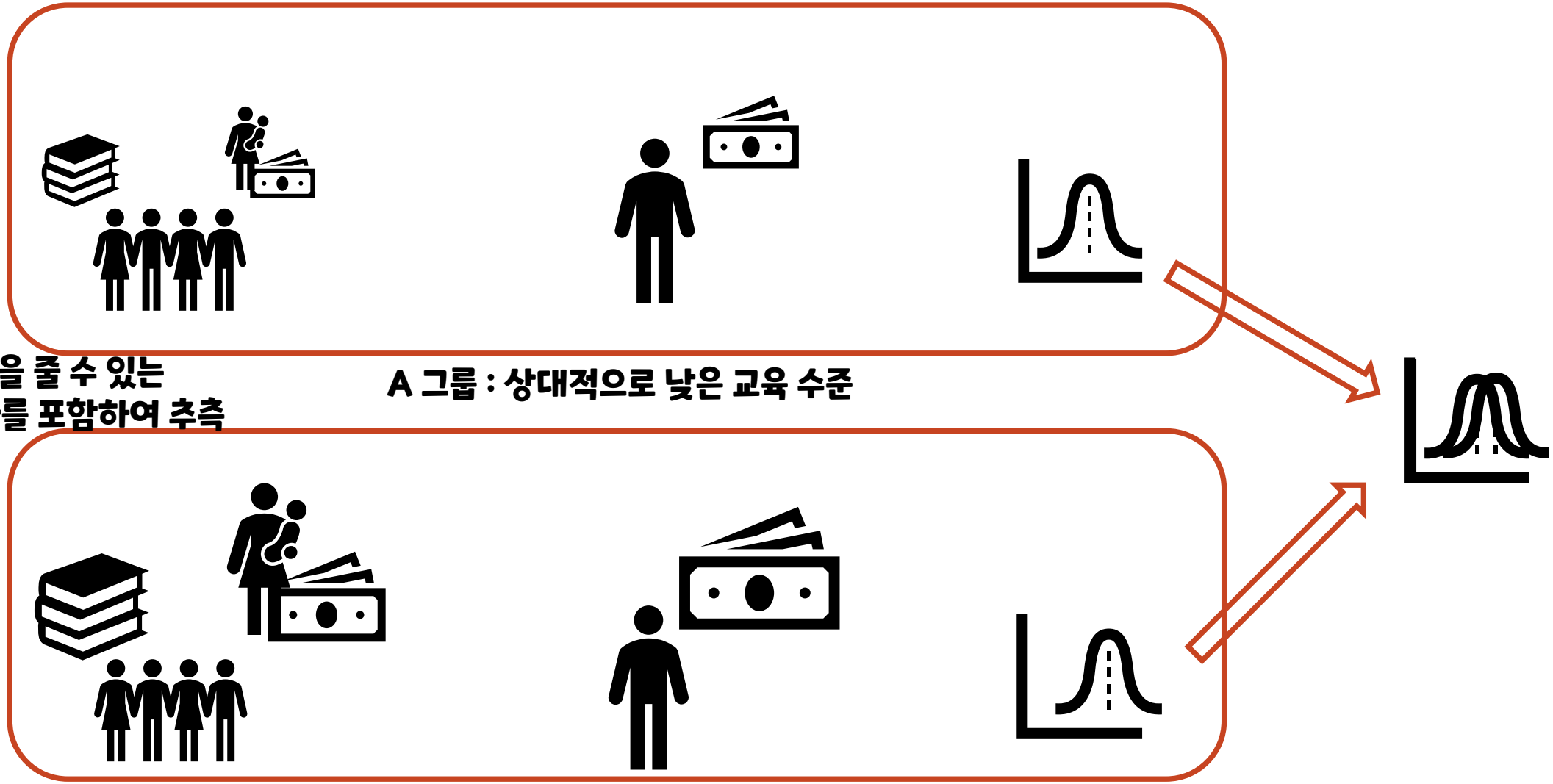
인문 과학에서는 ?


인위적 조정이
어려움
A 그룹 : 상대적으로 낮은 교육 수준
가정 환경 /
부모의 재력등 2차
요인 통제 실험
불가
B 그룹 : 상대적으로 높은 교육 수준

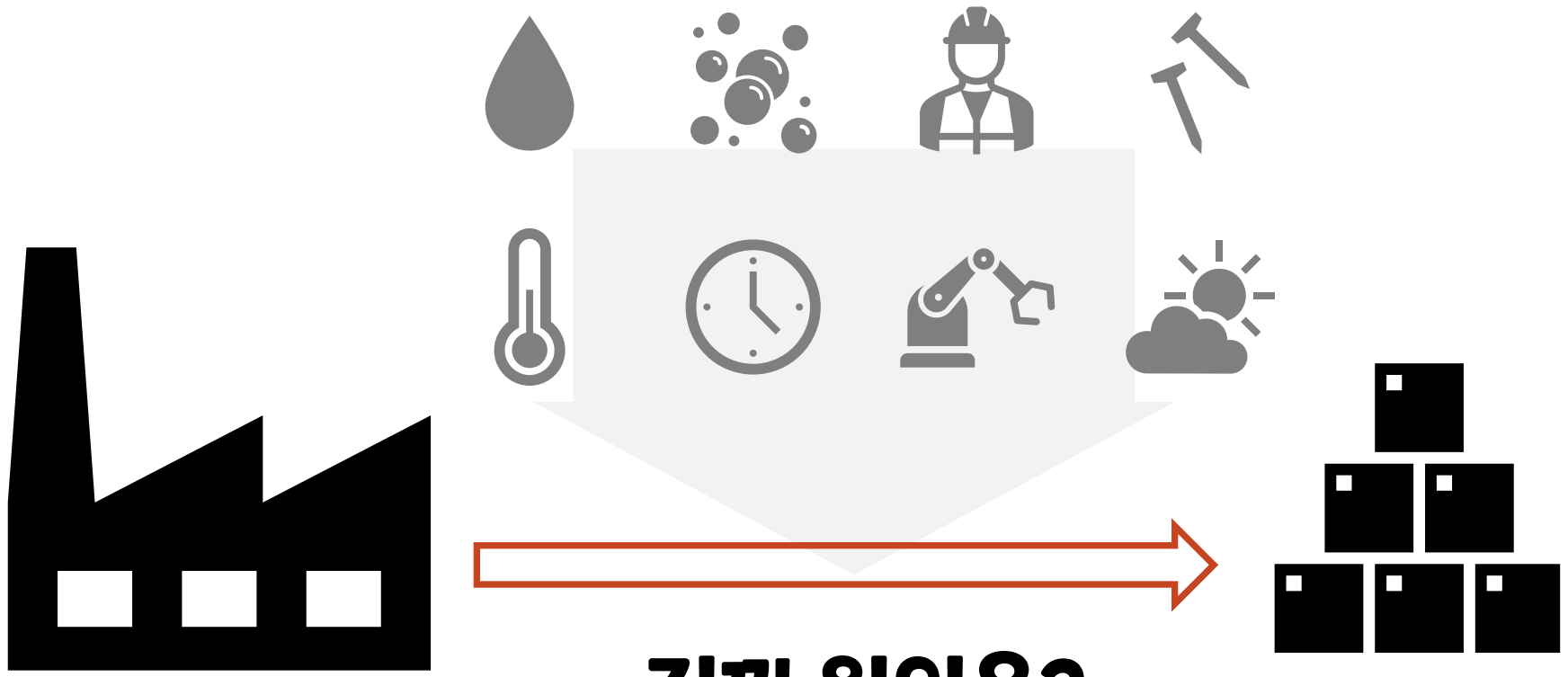
인문 과학 연구 방법(1)



인문 과학 연구 방법(2)



회사에서는 왜 갑자기 통계? (1)

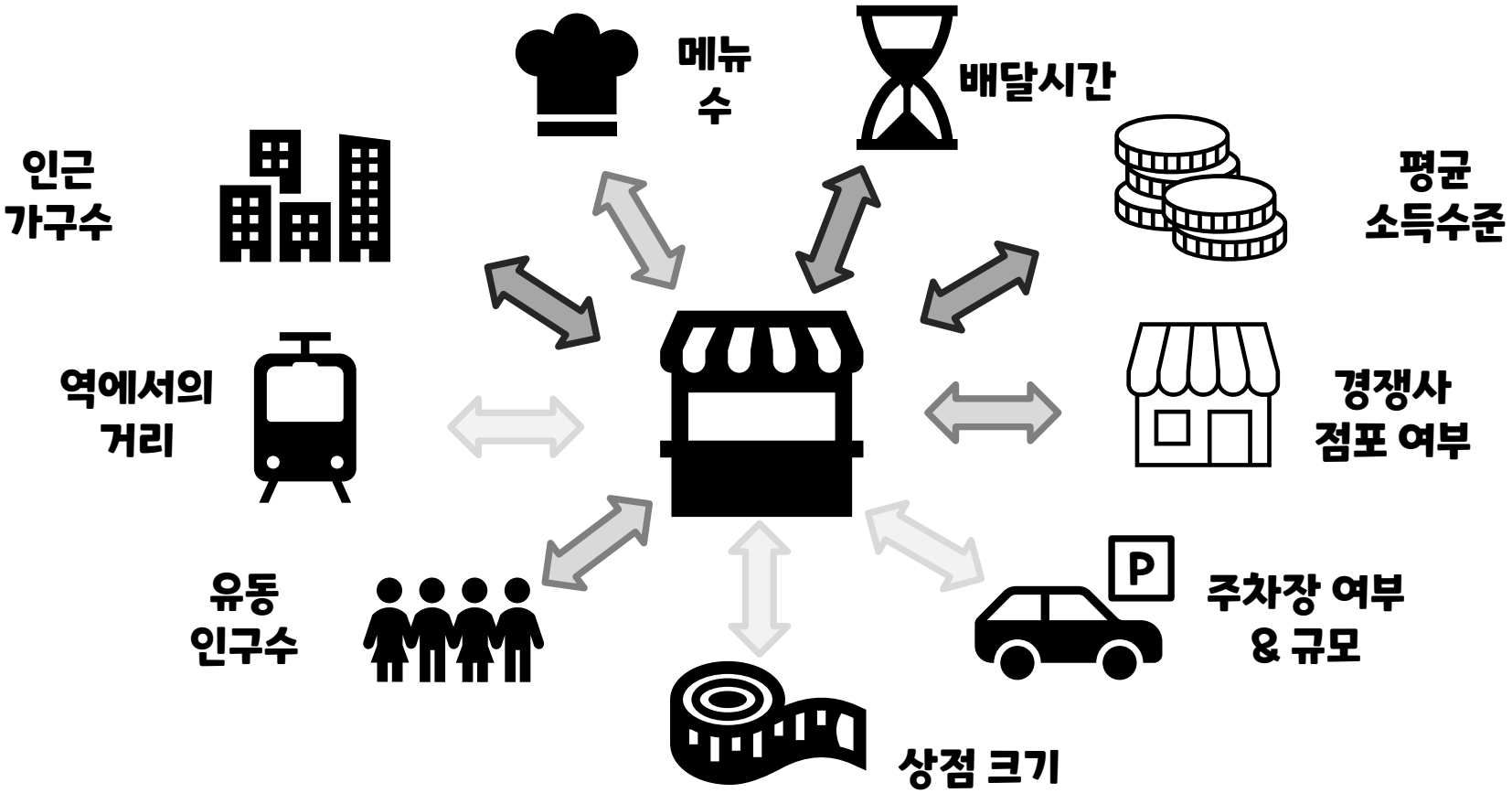


진짜 원인은?

회사에서는 왜 갑자기 통계? (2)

가게를 차리면 잘 될까? (예측)

회사에서는 왜 갑자기 통계? (3)



어디에? / 매출을 높이려면? (추론)

회사에서는 왜 갑자기 통계? (4)

효과가 있어? (검증)

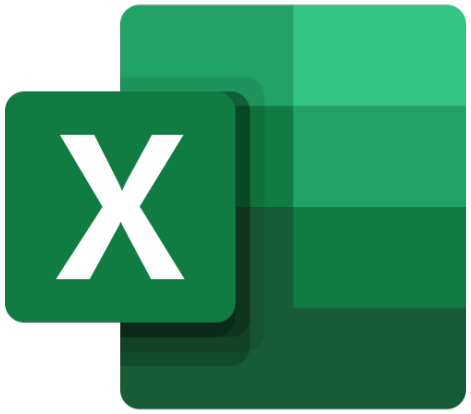

자주쓰는
통계도구

대표적 통계 분석 툴






대표적 통계 분석 툴







대표적 통계 분석 툴 비교

엑셀	구글 Sheets	미니탭	R	파이썬	Orange Datamining	SPSS	SAS
데이터 분석의 기본 개별 데이터 가공이 용이 기본으로 깔려 있다 (추가비용 없음)	엑셀의 경쟁자 개별 데이터 가공이 용이 비교적 저렴하다	가볍다 (통계툴 치고) 쉽다 (통계툴 치고)	공짜다 다된다 참고자료가 많다 인기 ※ 통계학자가 만든 언어	공짜다 다된다 참고자료 매우 많다 가장 인기 ※ 언어에 통계기능 확장	공짜다 쉽다 ※ 파이썬 기반 일부 기능 시각화	다된다 ※ IBM 인수 후 그다지…	다된다 축적된 전문가 풀 금융 & Health 영역 강자
개별 데이터 가공이 가능 통계기능은 약함 (상대적) 대용량 데이터 약함	개별 데이터 가공이 가능 통계기능은 약함 (상대적) 대용량 데이터 약함	유료다 대용량 데이터 약함	어렵다	어렵다	기능이 한정적	비싸다	비싸다 (아주 많이)

Orange Datamining 설치해 보기

orange DATA MINING

Screenshots Download Blog Docs Workshops Donate Search

Suggested Download

Orange 3.37.0 for Windows

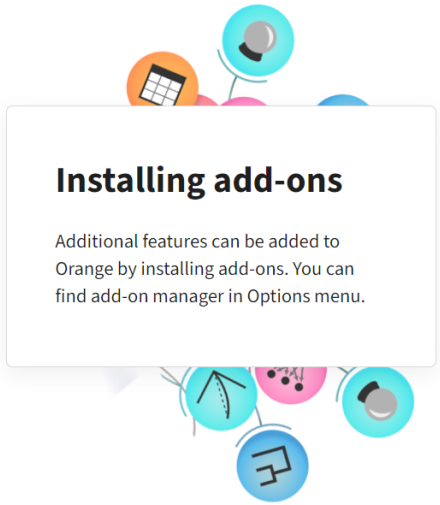
Windows

Standalone installer (default)

Orange3-3.37.0-Miniconda-x86_64.exe
Can be used without administrative priviledges.

Portable Orange

Orange3-3.37.0.zip
No installation needed. Just extract the archive and open the shortcut in the extracted folder.



macOS

Orange for Apple silicon

Orange3-3.37.0-Python3.11.8-arm64.dmg

https://orangedatamining.com/download/

R 설치해 보기

R

R-4.2.0 for Windows

Download R-4.2.0 for Windows (79 megabytes, 64 bit)

README on the Windows binary distribution
New features in this version

This build requires UCRT, which is part of Windows since Windows 10 and Windows Server 2016. On older systems, UCRT has to be installed manually from here.

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the md5sum of the .exe to the fingerprint on the master server.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

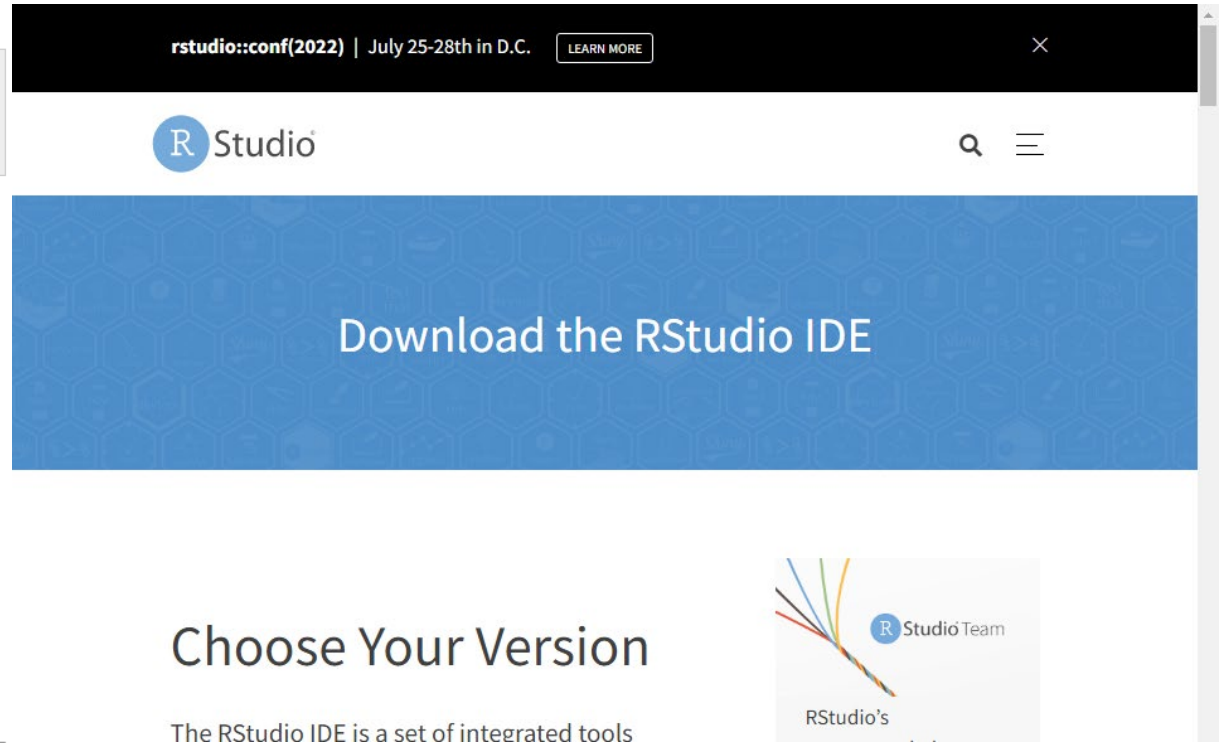
Other builds

- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.html.

https://cran.r-project.org/bin/windows/base/

R Studio



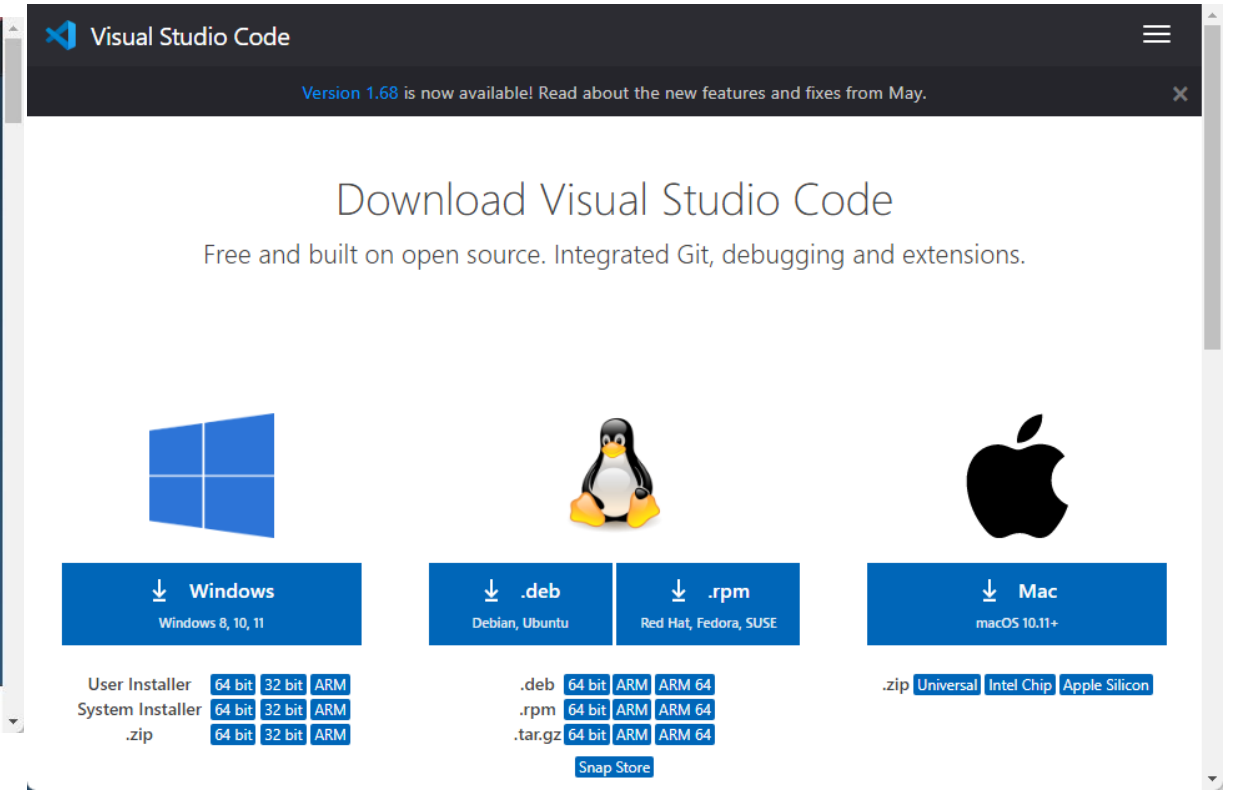
https://www.rstudio.com/products/rstudio/download/

파이썬 설치해 보기

Python

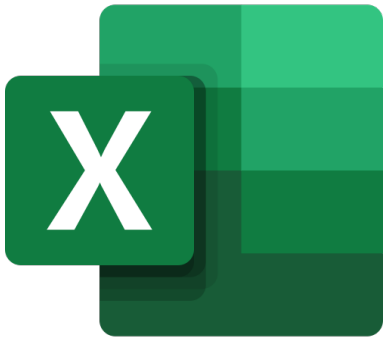
https://www.python.org/downloads/

Visual Studio Code



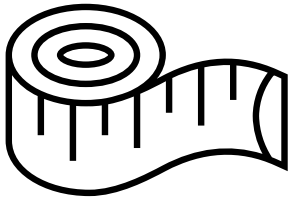
https://code.visualstudio.com/download

※ 대표적 데이터 시각화 툴

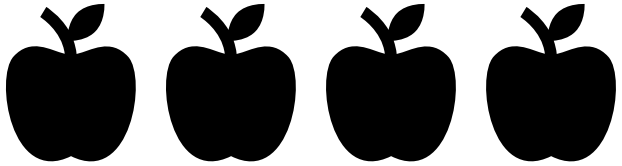



기초 통계를
알아보자

데이터의 종류



연속형 : 연속되는 숫자로 표현되며
정수로 떨어지지 않음 (길이, 무게, 온도 등)

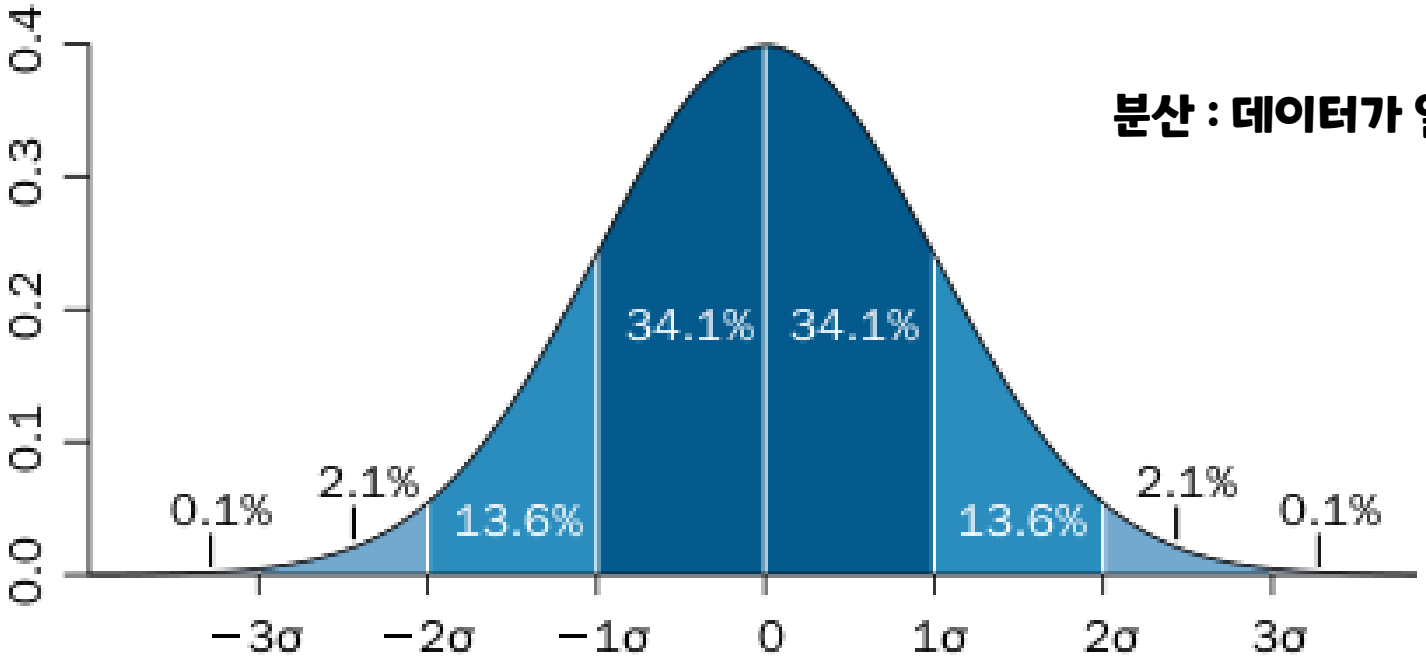


이산형 : 숫자로 표현되며 정수로 떨어짐
(EA, 개 등)

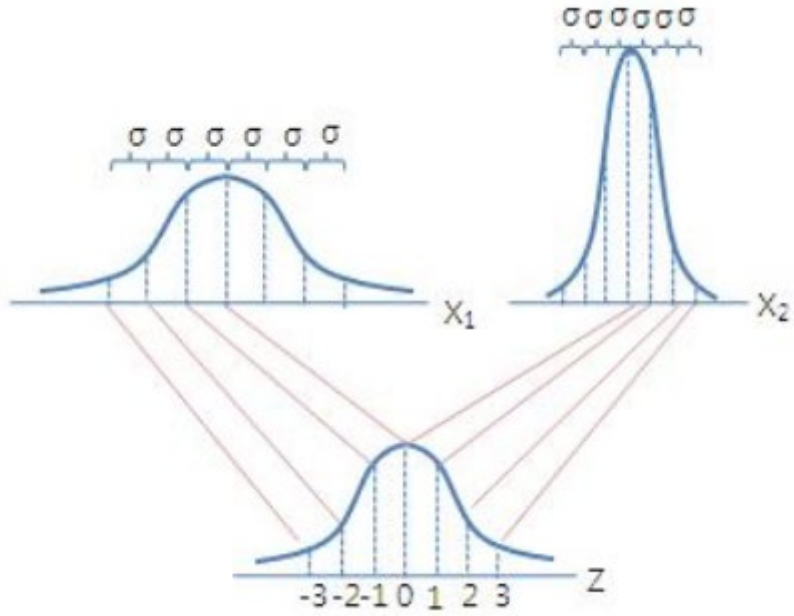
범주형 : 대 / 소의 구분이 없음. 다른 종류/범주를 의미
(남 / 여, 사과 / 복숭아 / 체리 등)

대표적인 대표값 (1)

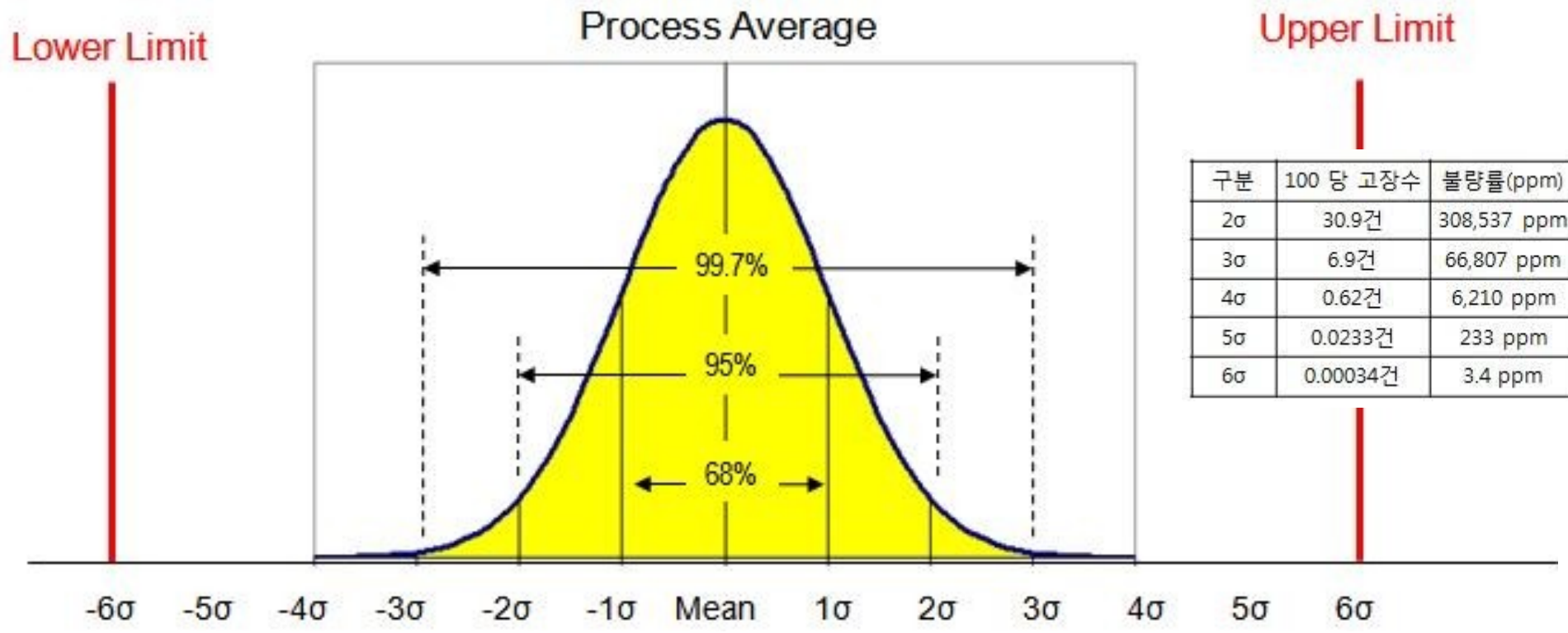
평균 : 데이터를 대표하는 값



분산 : 데이터가 얼마나 퍼져있나

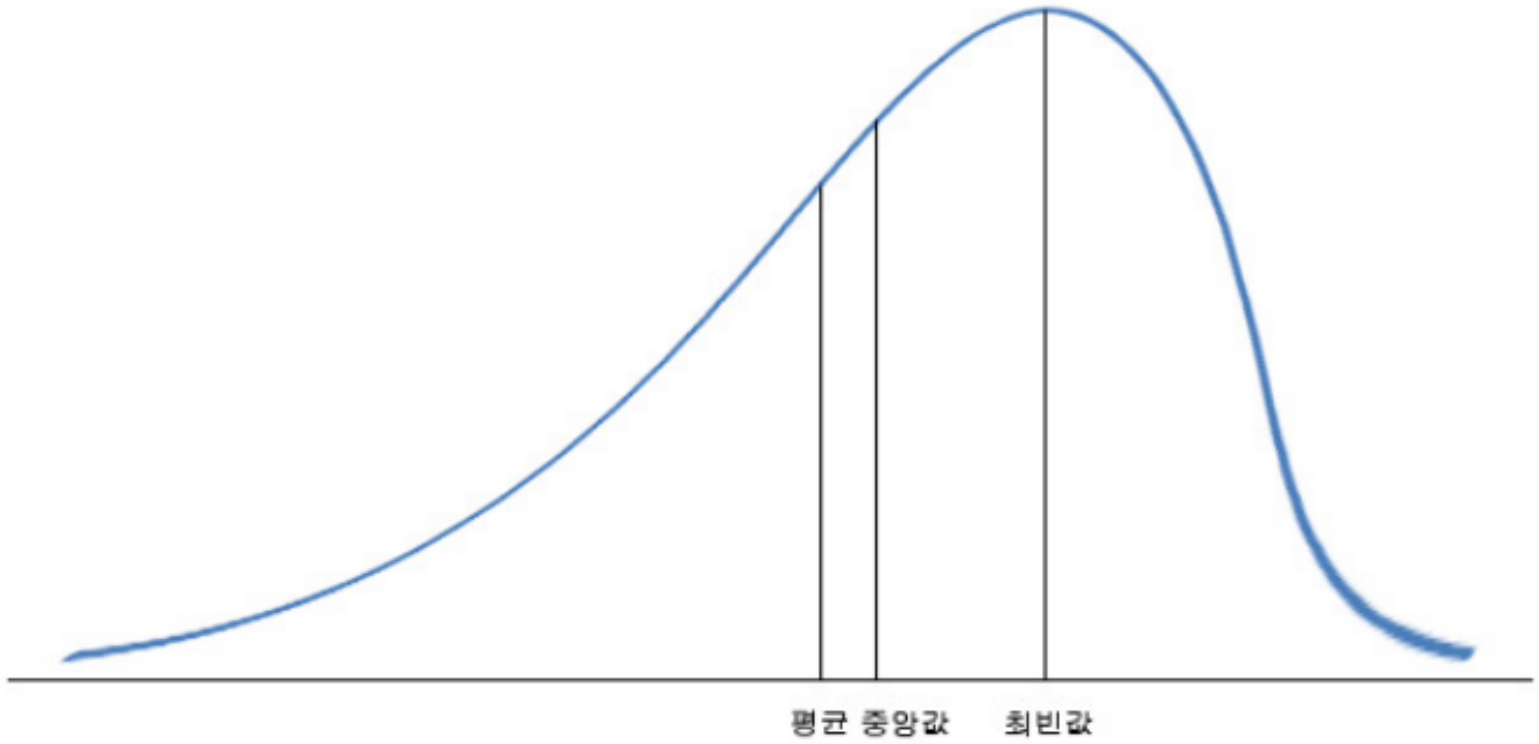


대표적인 대표값 (2)

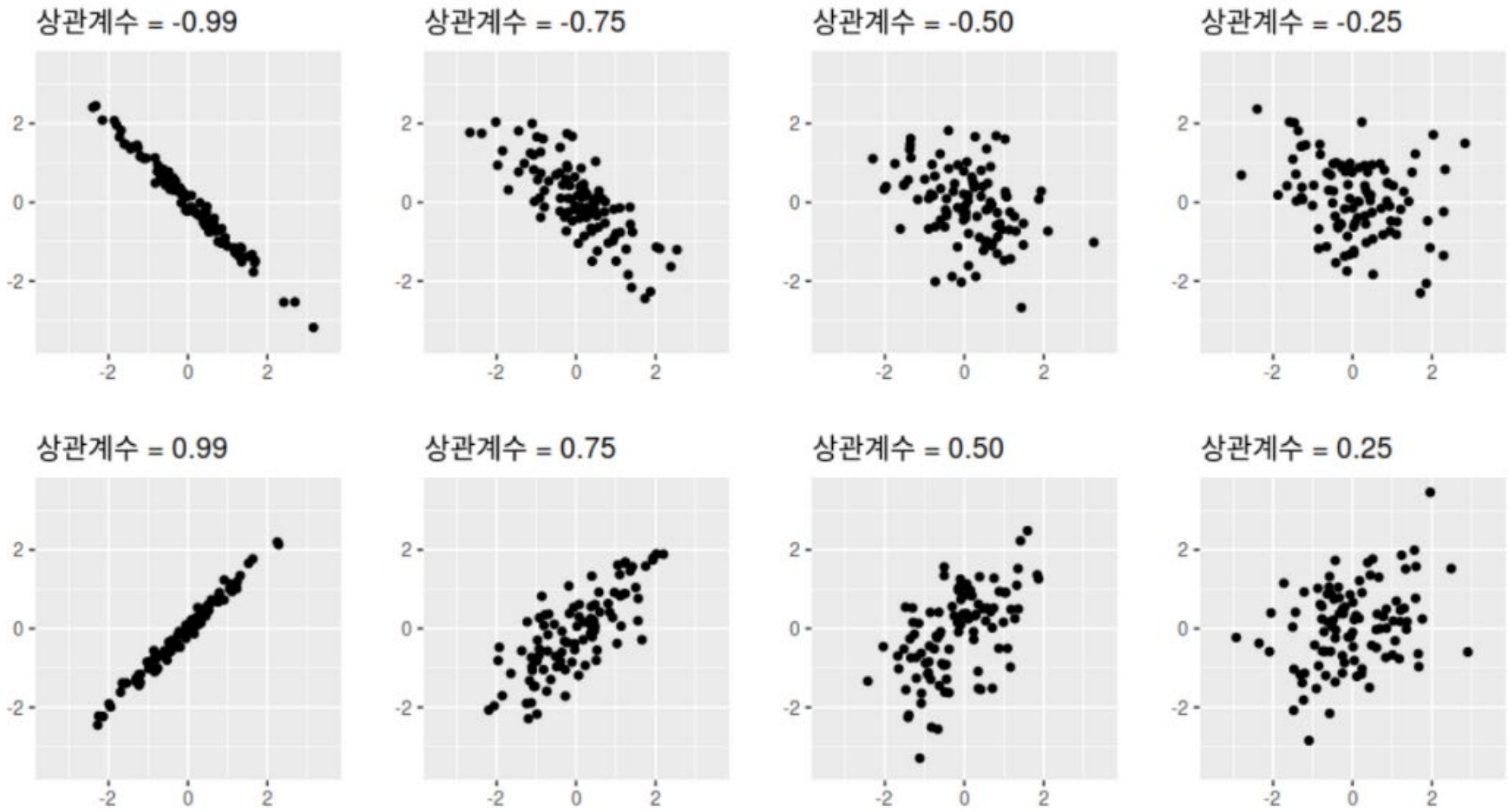


구분	100 당 고장수	불량률(ppm)	양품률
2σ	30.9건	308,537 ppm	67.15 %
3σ	6.9건	66,807 ppm	93.32 %
4σ	0.62건	6,210 ppm	99.48 %
5σ	0.0233건	233 ppm	99.98 %
6σ	0.00034건	3.4 ppm	99.99966 %

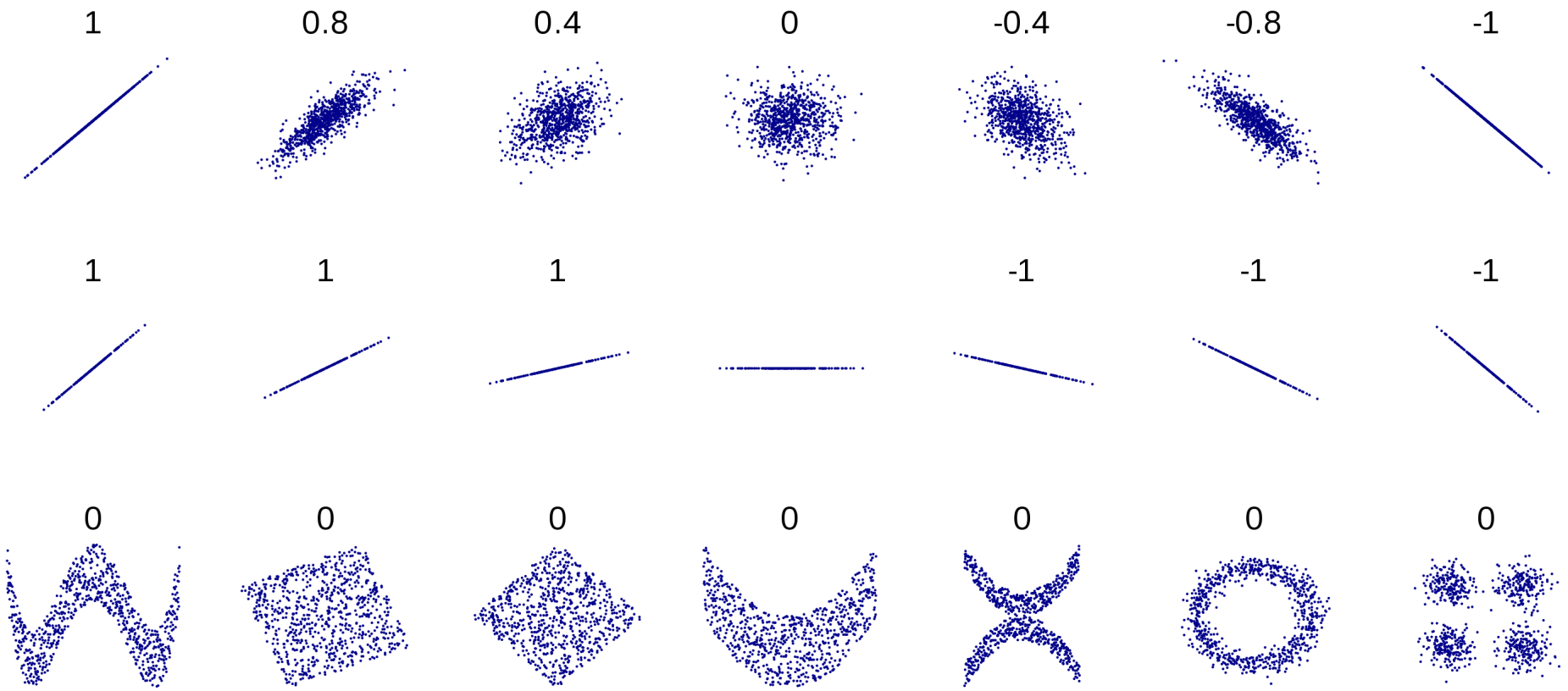
대표적인 대표값 (3)



대표적인 대표값 (4)

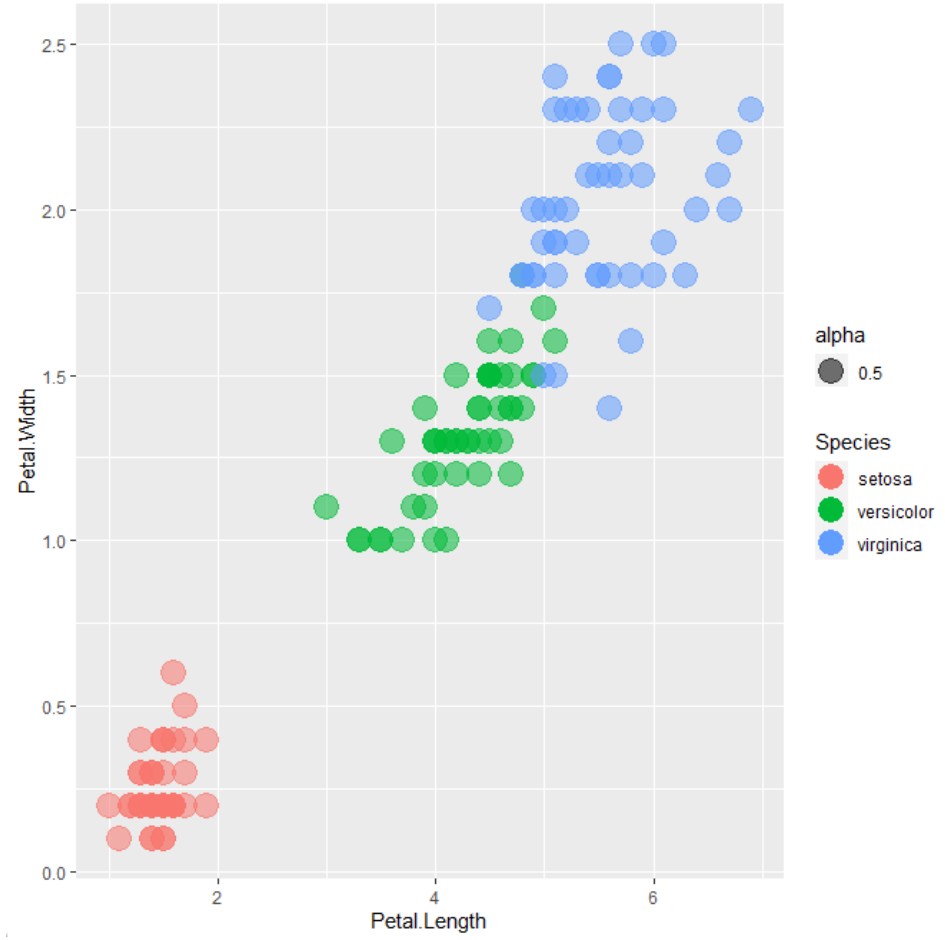

상관계수 = -0.99
상관계수 = -0.75
상관계수 = -0.50
상관계수 = -0.25
상관계수 = 0.99
상관계수 = 0.75
상관계수 = 0.50
상관계수 = 0.25
2
0
-2
-2
0
2

대표적인 대표값 (5)

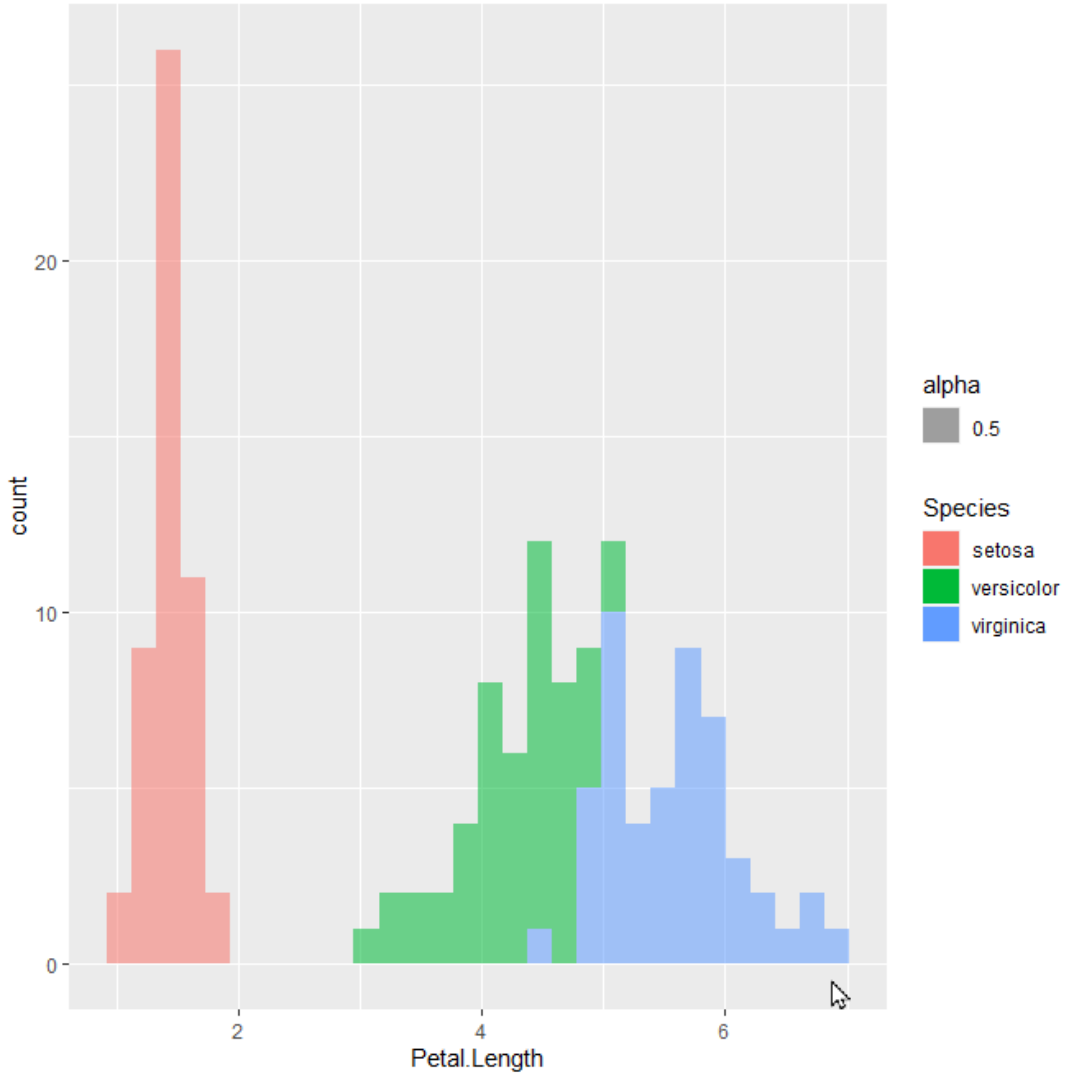


※ 상관계수가 0이라고 상관관계가 없는 것은 **아니다.**

대표적인 표현방법 (1)

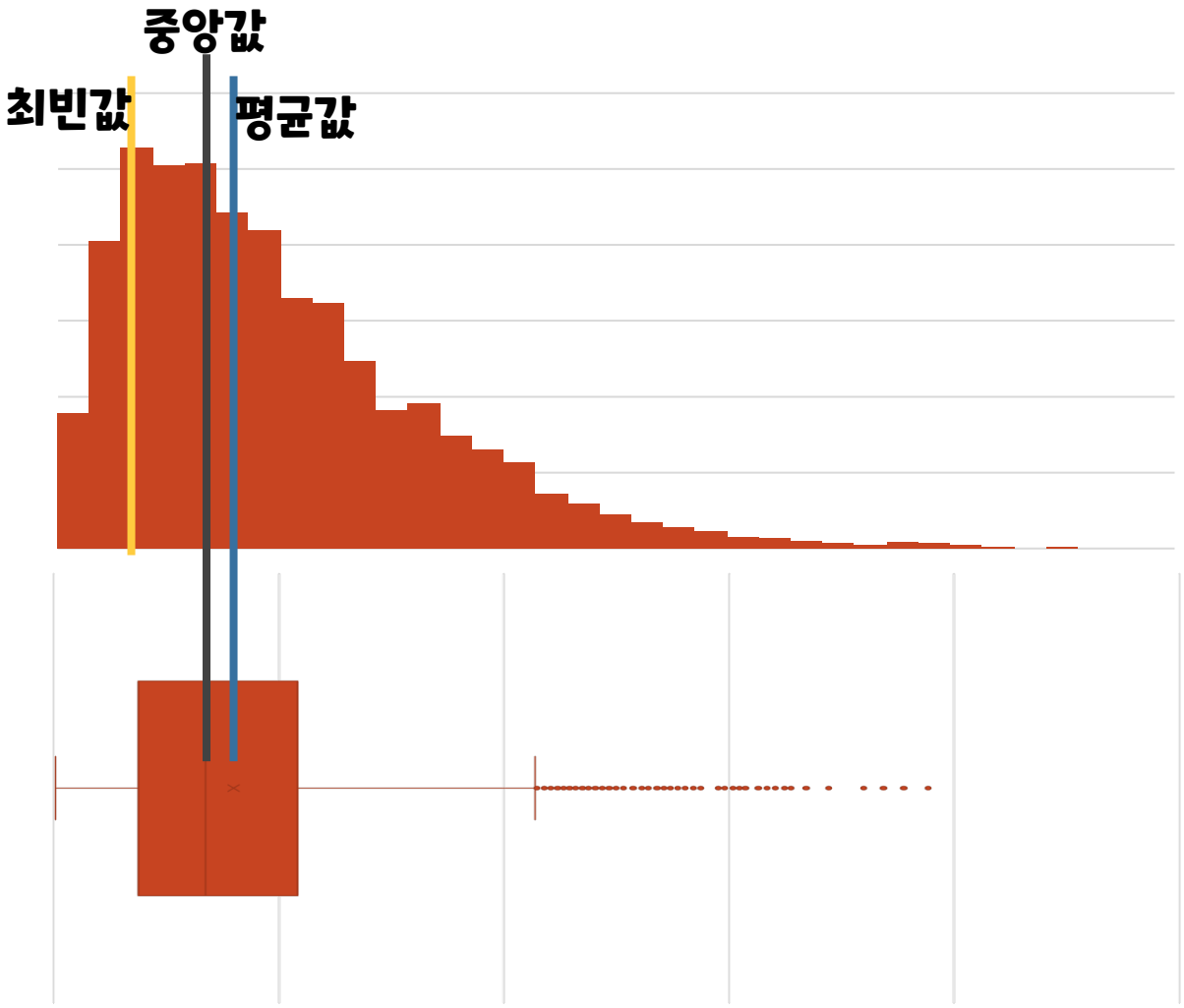


산점도



히스토 그램

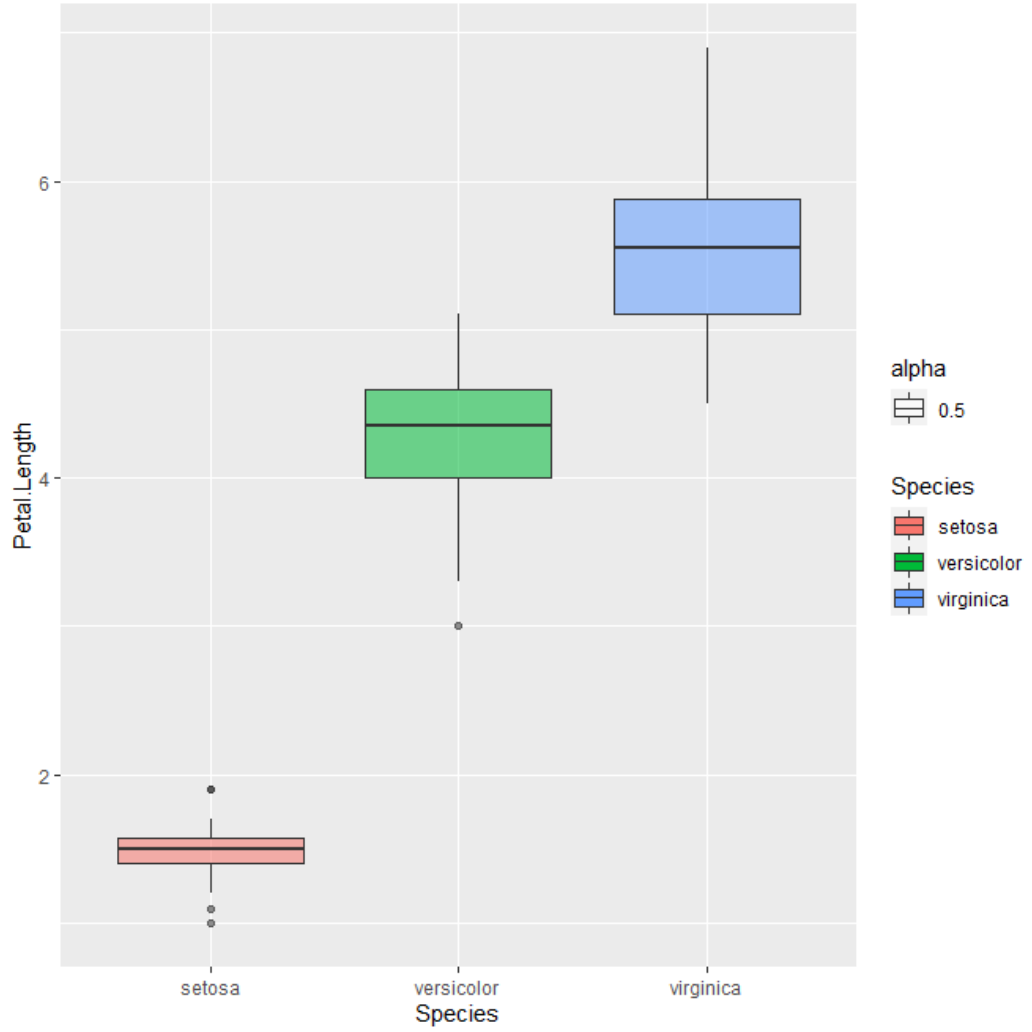
대표적인 표현방법 (2)



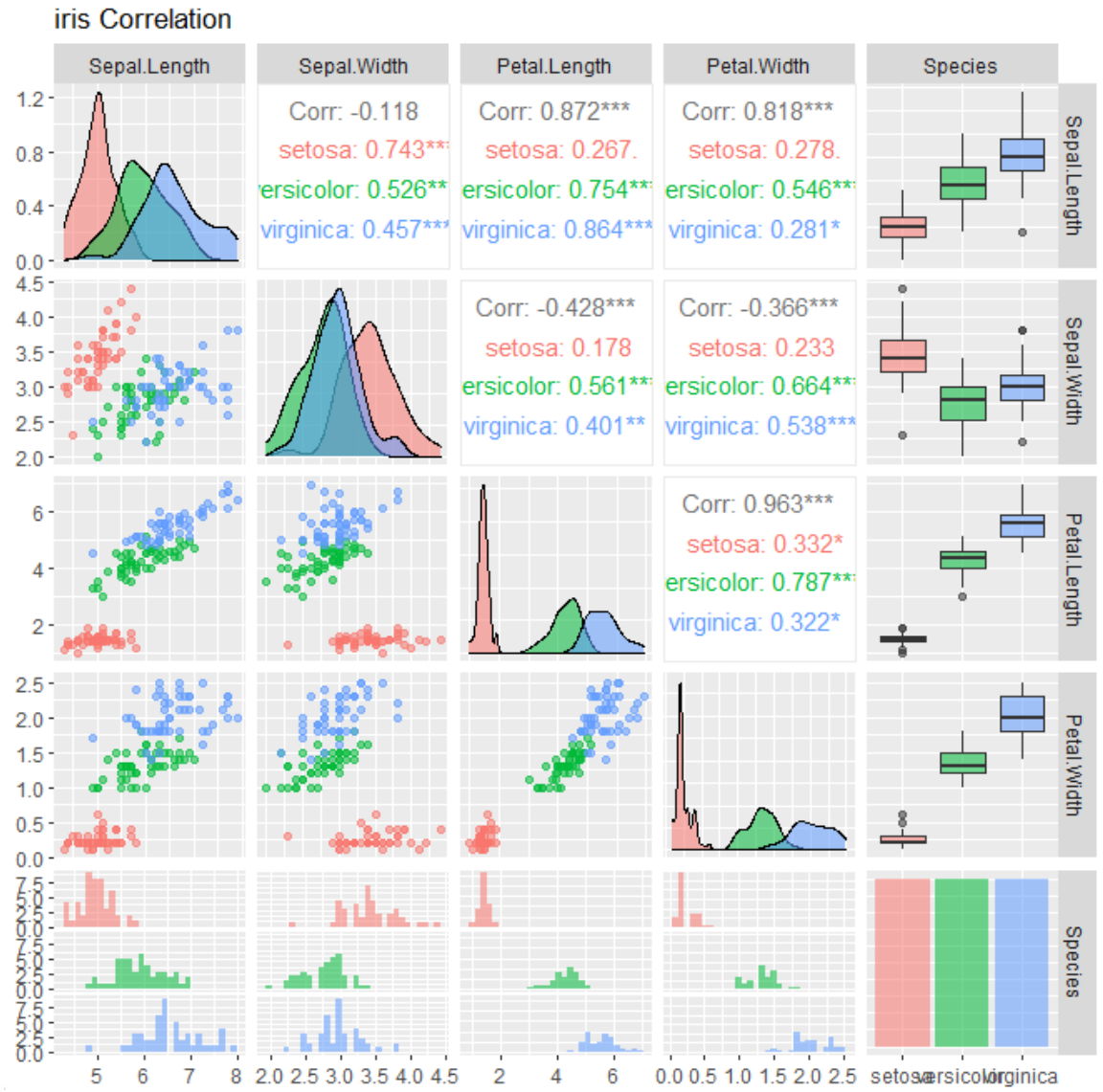
히스토그램

상자수염그림

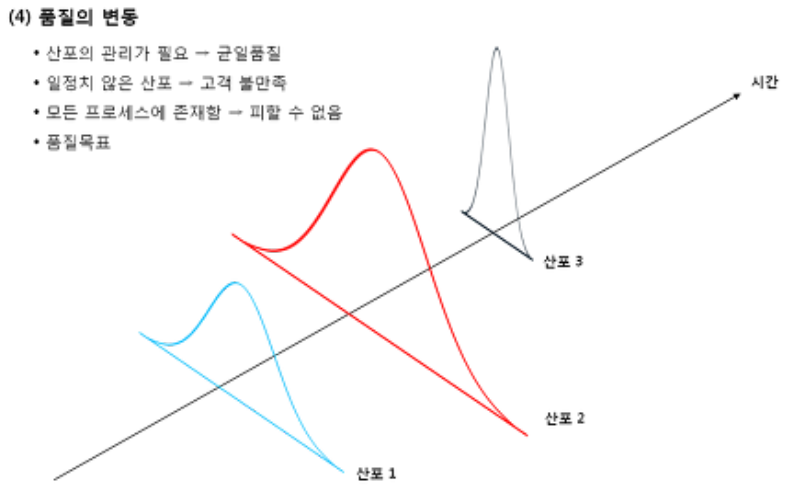
대표적인 표현방법 (3)


6
4
2
Petal.Length
setosa
versicolor
virginica
Species
alpha
0.5
Species
setosa
versicolor
virginica

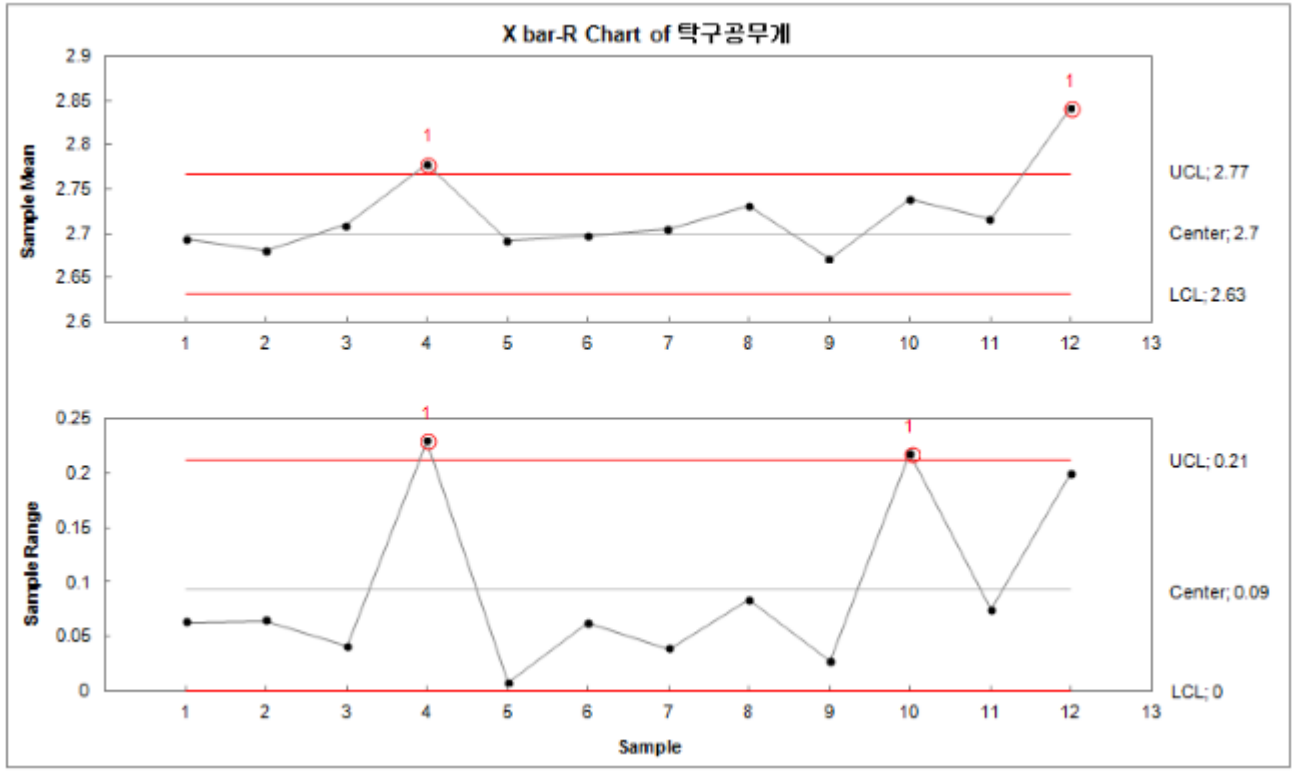
대표적인 표현방법 (4)


iris Correlation
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
Corr: -0.118
setosa: 0.743***
ersicolor: 0.526**
virginica: 0.457***
Corr: 0.872***
setosa: 0.267.
ersicolor: 0.754***
virginica: 0.864***
Corr: 0.818***
setosa: 0.278.
ersicolor: 0.546***
virginica: 0.281*
Corr: -0.428***
setosa: 0.178
ersicolor: 0.561***
virginica: 0.401**
Corr: -0.366***
setosa: 0.233
ersicolor: 0.664***
virginica: 0.538***
Corr: 0.963***
setosa: 0.332*
ersicolor: 0.787***
virginica: 0.322*
setosa
versicolor
virginica

대표적인 표현방법 (5)

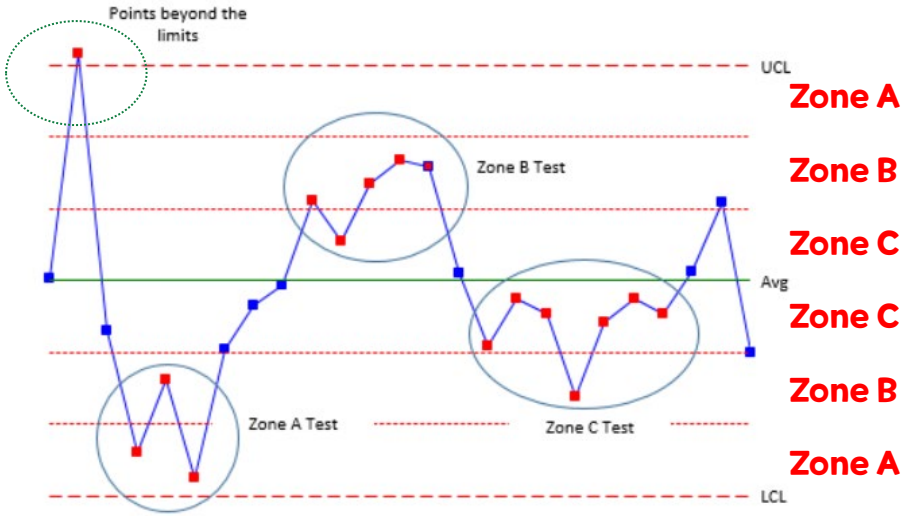


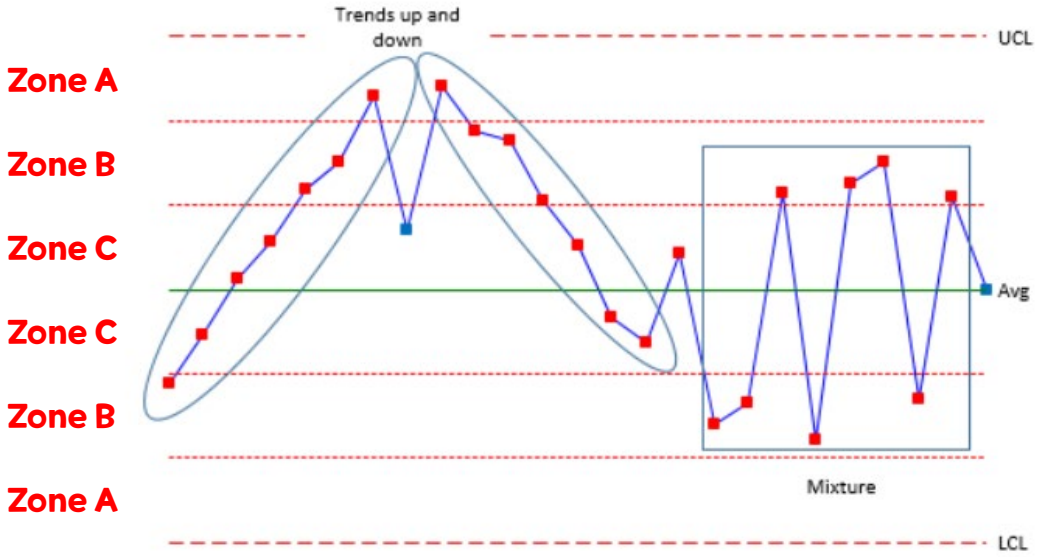
※ SPC 관리도는 변화 발생 여부를 빨리 알아채고자 사용하는 툴

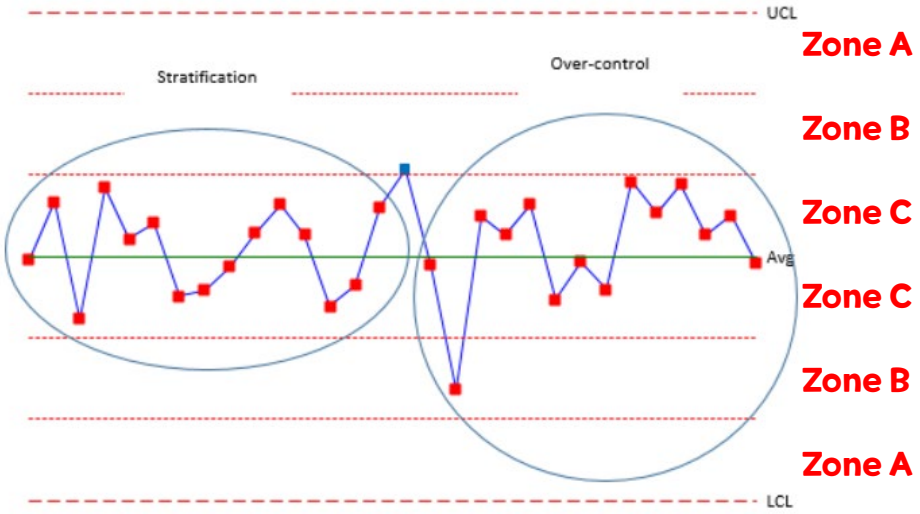


SPC rules

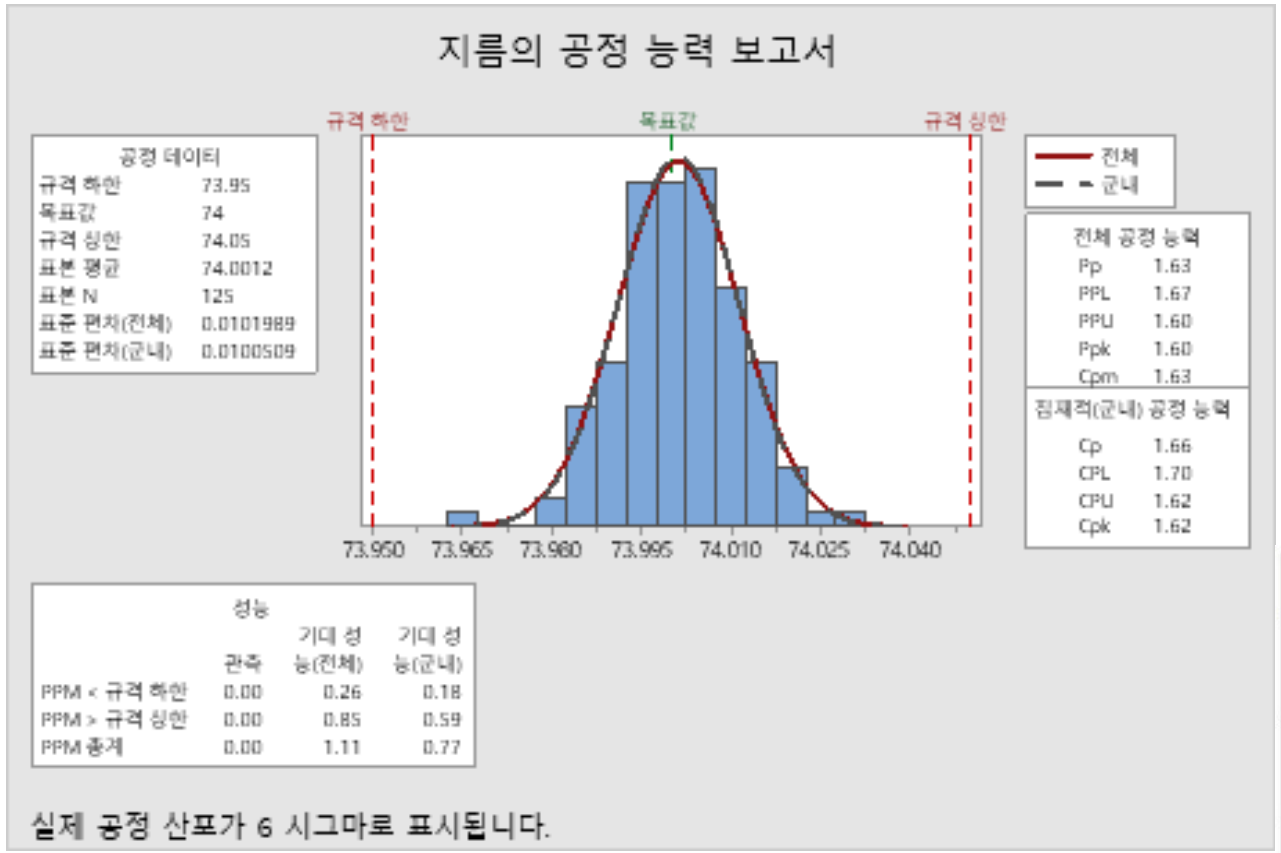
Rule	Rule Name	Pattern
1	Beyond Limits	One or more points beyond the control limits
2	Zone A	2 out of 3 consecutive points in Zone A or beyond
3	Zone B	4 out of 5 consecutive points in Zone B or beyond
4	Zone C	7 or more consecutive points on one side of the average (in Zone C or beyond)
5	Trend	7 consecutive points trending up or trending down
6	Mixture	8 consecutive points with no points in Zone C
7	Stratification	15 consecutive points in Zone C
8	Over-control	14 consecutive points alternating up and down







대표적인 표현방법 (6)

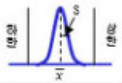
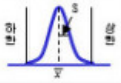
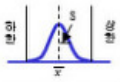
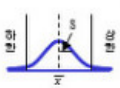
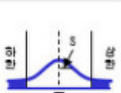


공정 데이터	
규격 하한	73.95
목표값	74
규격 상한	74.05
표본 평균	74.0012
표본 N	125
표준 편차(전체)	0.0101989
표준 편차(군내)	0.0100509

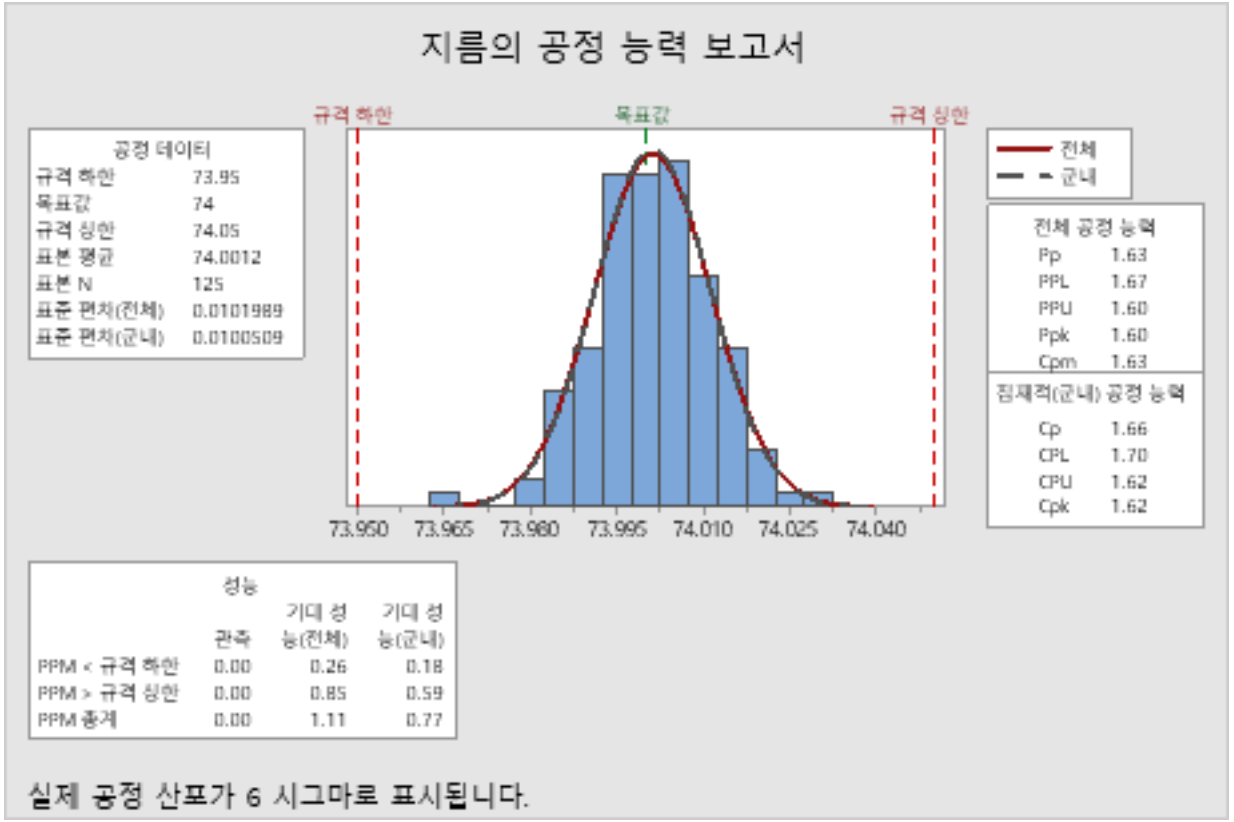
전체 공정 능력	
Pp	1.63
PPL	1.67
PPU	1.60
Ppk	1.60
Cpm	1.63

잠재적(군내) 공정 능력	
Cp	1.66
CPL	1.70
CPU	1.62
Cpk	1.62

성능	관측	기대 성능(전체)	기대 성능(군내)
PPM < 규격 하한	0.00	0.26	0.18
PPM > 규격 상한	0.00	0.85	0.59
PPM 총계	0.00	1.11	0.77

분포현상	공정능력 지수	등급	공정능력 유무 판단	시정조치	비고	
					Cp값	σ
	Cp ≥ 1.67	0	공정능력은 매우 충분.	• 등쭉날쭉이 약간 커져도 걱정할 필요가 없다. • 비용절감이나 관리의 간소화를 생각하도록 한다.	Cp = 1.67	± 5σ
	1.67 > Cp ≥ 1.33	1	공정능력은 충분.	• 아주 이상적인 공정상황이므로 현재의 상태를 유지한다.	Cp = 1.33	± 4σ
	1.33 > Cp ≥ 1.00	2	공정능력이 충분하지는 않지만 그 정도면 괜찮다	• 공정관리를 확실하게 하여 관리상태를 유지할 것. • Cp가 1에 가까워지면 불량발생의 가능성이 있으므로 주의해야 한다.	Cp = 1.00	± 3σ
	1.00 > Cp ≥ 0.67	3	공정능력이 모자란다.	• 불량품이 생기고 있다. • 전체 선별, 공정의 개선, 관리가 필요하다.	Cp = 0.67	± 2σ
	0.67 > Cp	4	공정능력이 매우 부족하다.	• 품질이 전혀 만족스럽지 않다. 서둘러 현황조사, 원인규명, 품질개선 같은 긴급 대책을 펴야 한다. • 상한·하한 규격 값의 재검토도 해야 한다.	Cp = 0.33	± 1σ

공정 능력 지수

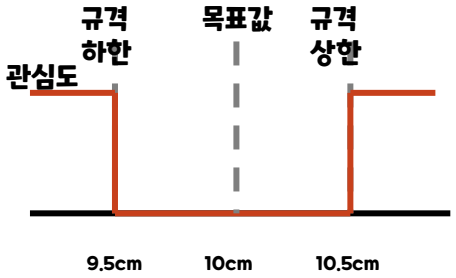


※ 이미지 출처 : 미니탭

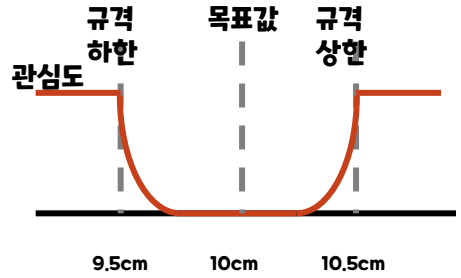
불량 발생 전, 사전 조치를 위함

1) 평균이 목표값으로 오게하거나

2) 분산을 줄이는 활동의 근거



기존의 품질 개념의 관리 관심도



공정 능력 관점의 관리 관심도


데이터
시각화하기

데이터 시각화의 목적

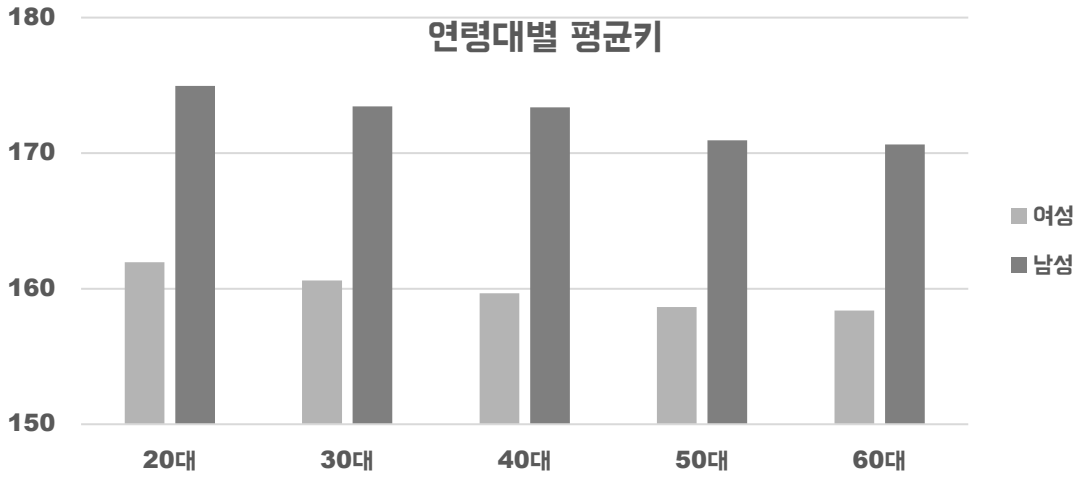
- 데이터에서 읽어낼 수 있는 사실의 효과적인 전달
- 핵심 메시지가 있어야 함
- 핵심 메시지와 시각화의 방향이 일치 해야함

하나. 목적에 따라 자주 사용하는 차트가 있다

-반드시 그래야만 하는 것은 아니나,

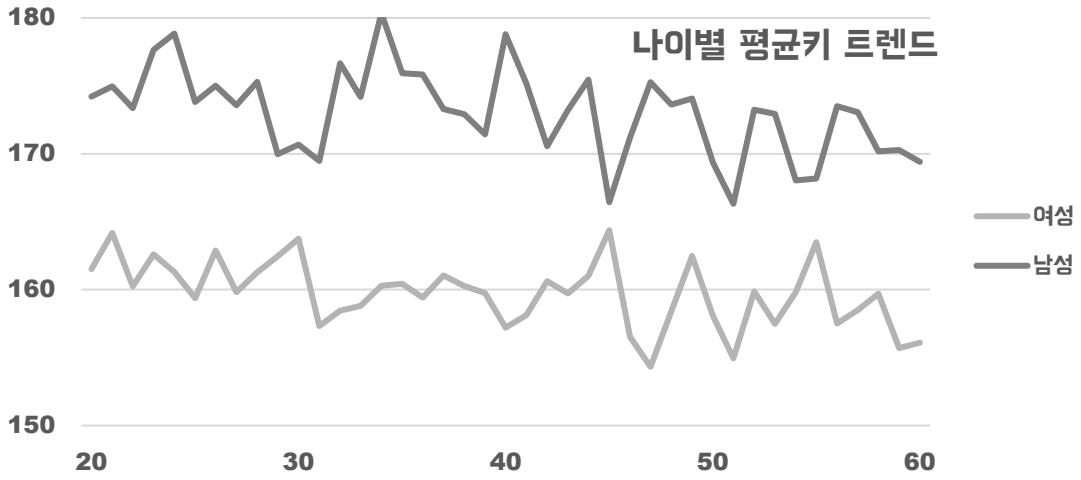
일반적으로 이해가 편하다.

속성 값의 비교

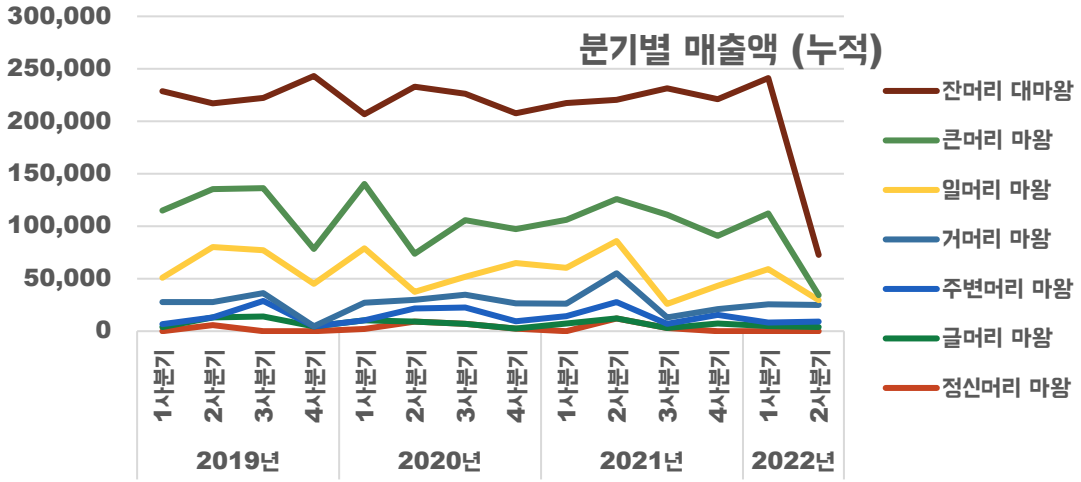


-막대 그래프 (세로)

시간에 따른 트렌드 분석

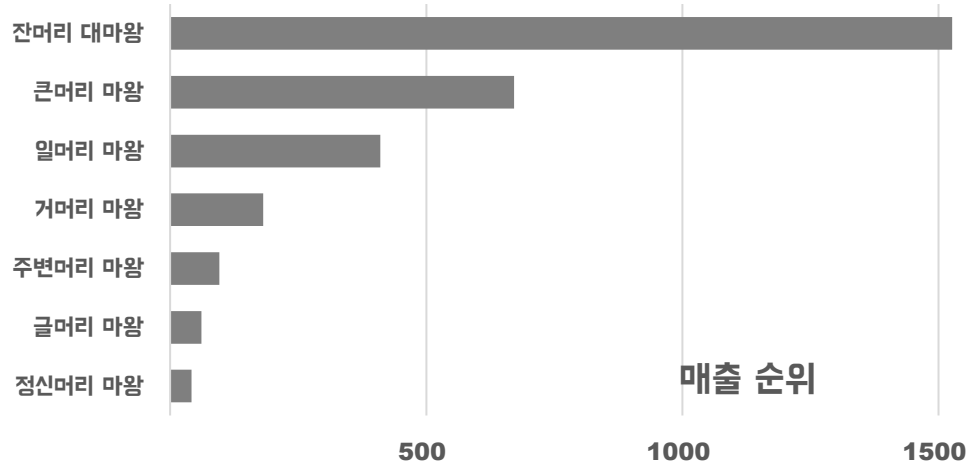


-꺾은선 그래프

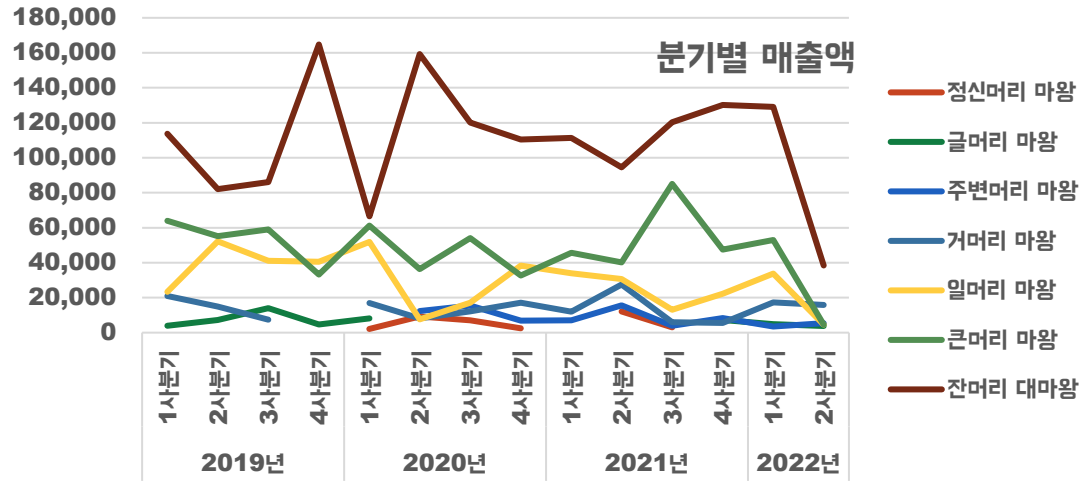


-영역 그래프

순위 비교



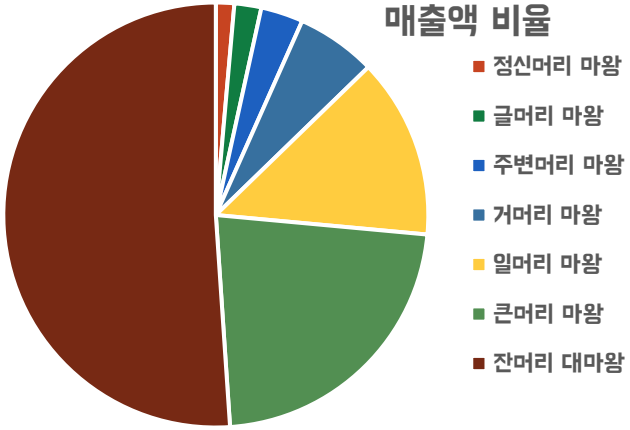
-막대 그래프 (가로, 현시점)

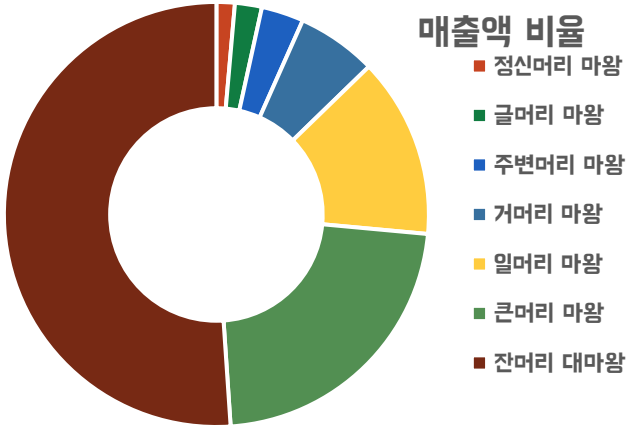


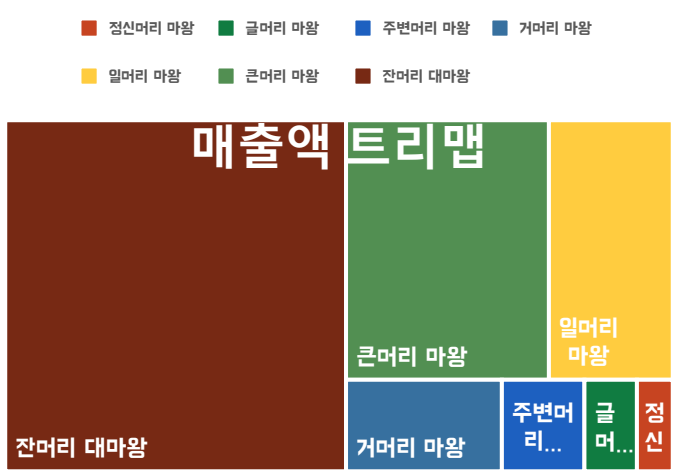
-꺾은선 그래프

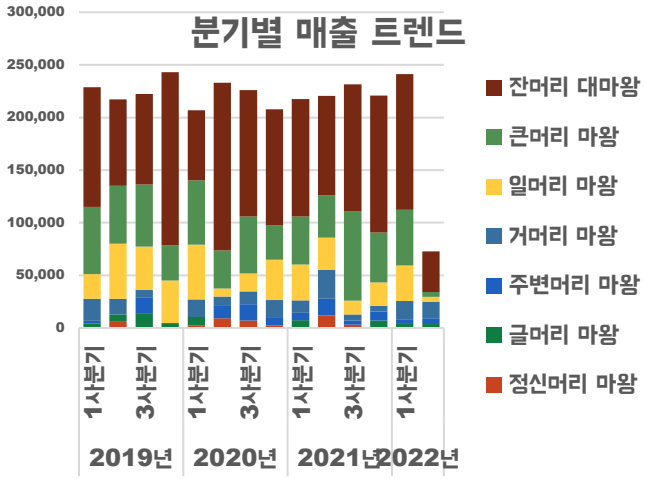
(시간 트렌드 포함)

비율 비교







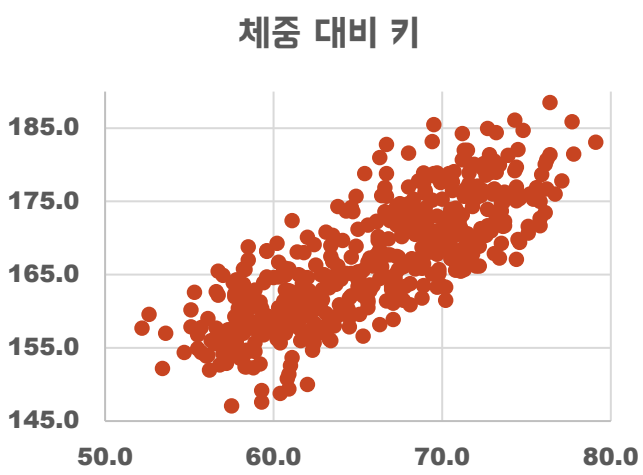


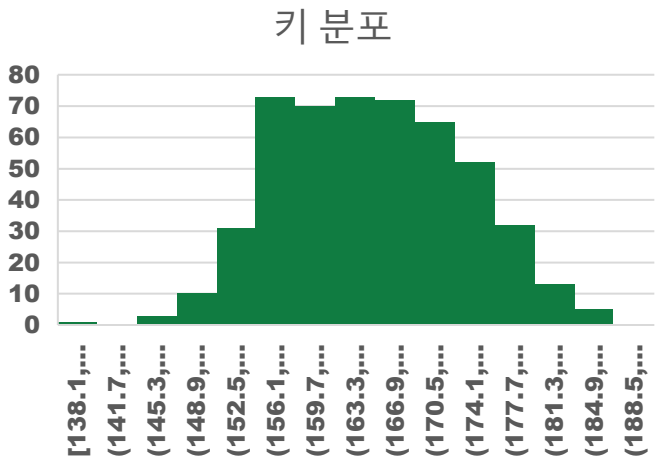
-파이 그래프 or
도넛 그래프

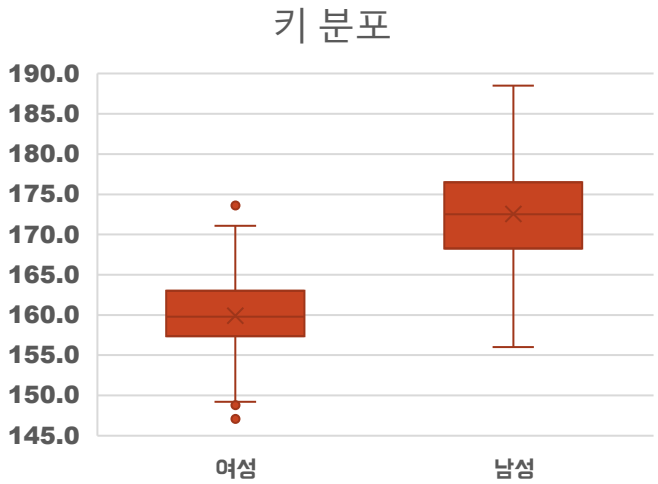
-트리 그래프

-누적 막대 그래프
(시간 트렌드 포함)

데이터 분포 분석





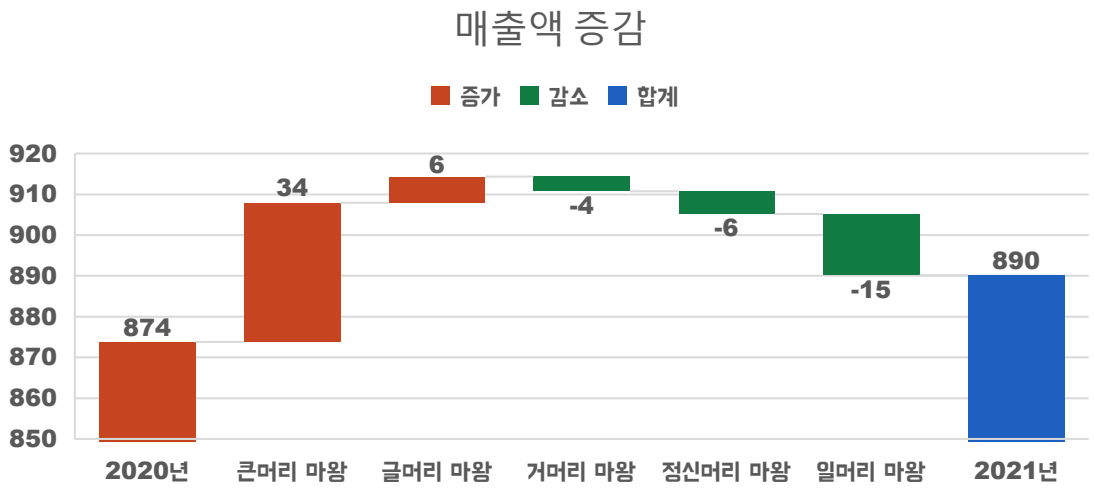


-산포도

-히스토 그램

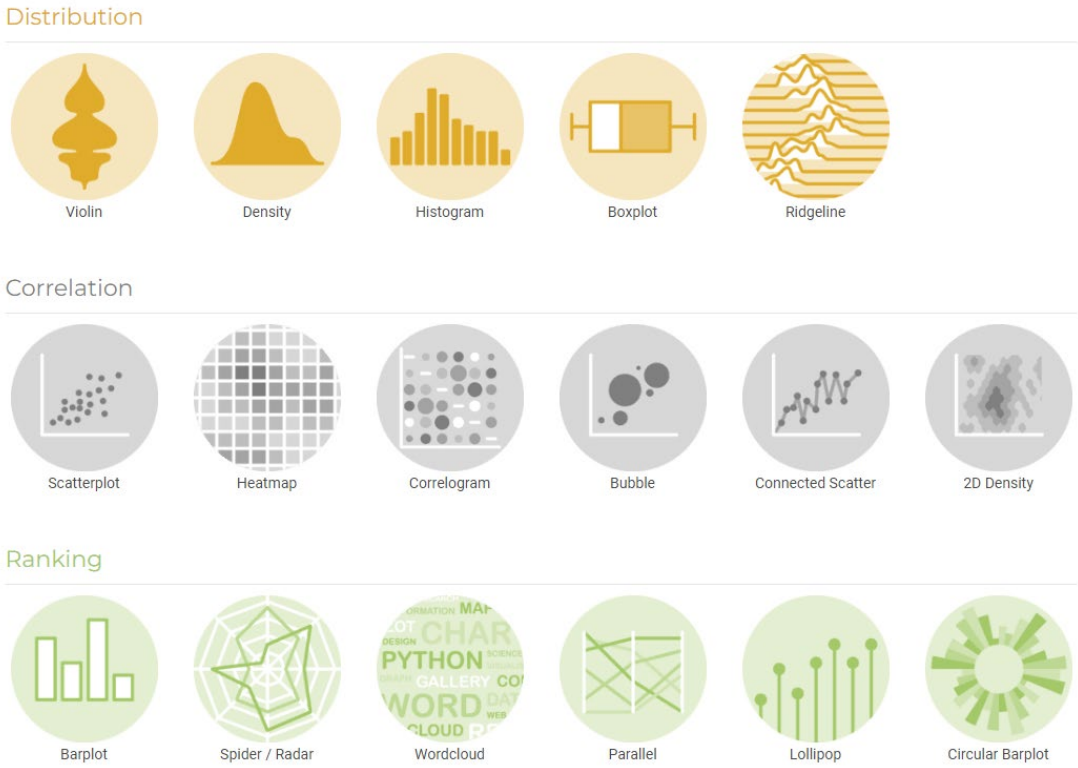
-박스 수염 플롯

증감 분석


매출액 증감
증가
감소
합계
920
910
900
890
880
870
860
850
874
34
6
-4
-6
-15
890
2020년
큰머리 마왕
글머리 마왕
거머리 마왕
정신머리 마왕
일머리 마왕
2021년

-폭포 차트

그 외…



-https://python-graph-gallery.com/

-https://r-graph-gallery.com/

둘. 데이터 잉크 비율 (Edward Tufte)

-데이터를 나타내기 위해 사용된 잉크의 양이

전체 잉크 총량에 비해 높을 수록 좋은 시각화

데이터 잉크 비율을 높이려면

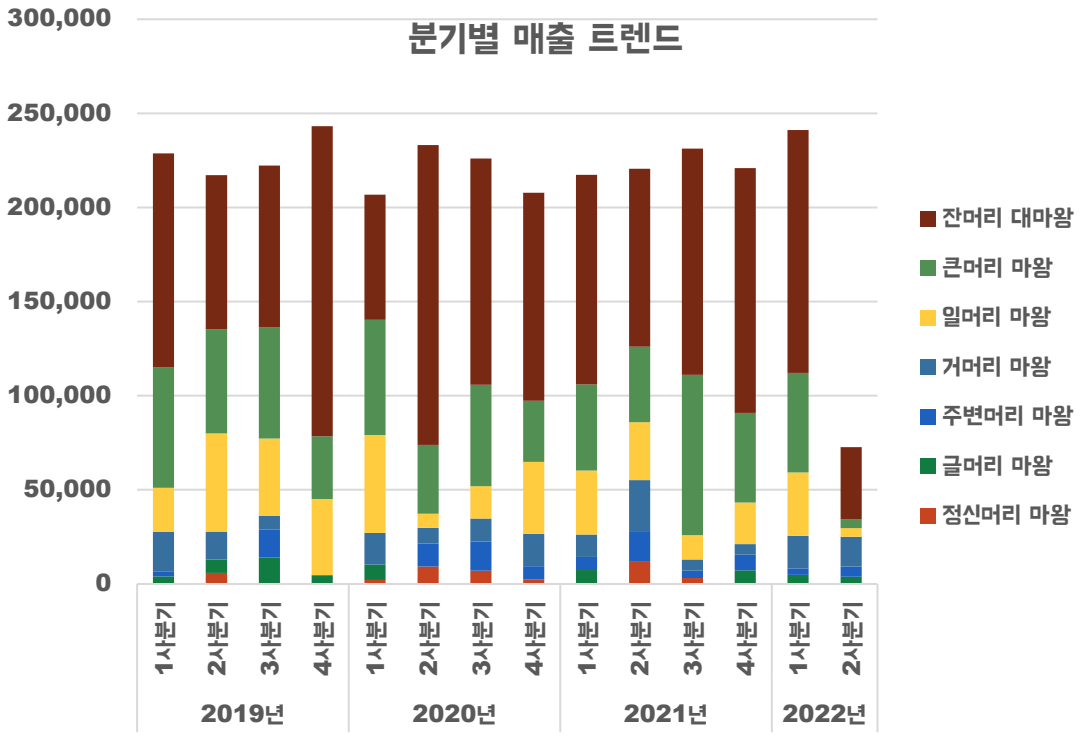
-예쁘게 꾸미기 보다는 간결하게 필요한 정보만.

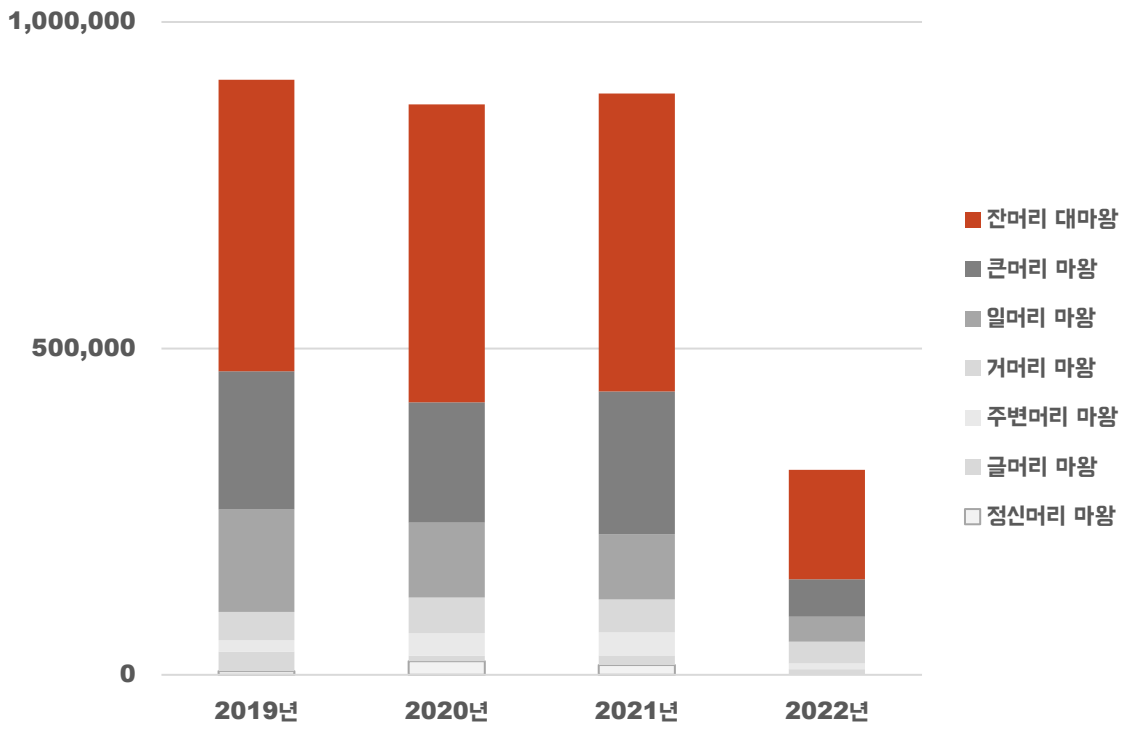
→ 3D 금지, 보조선 최소화, 칼라도 최소화

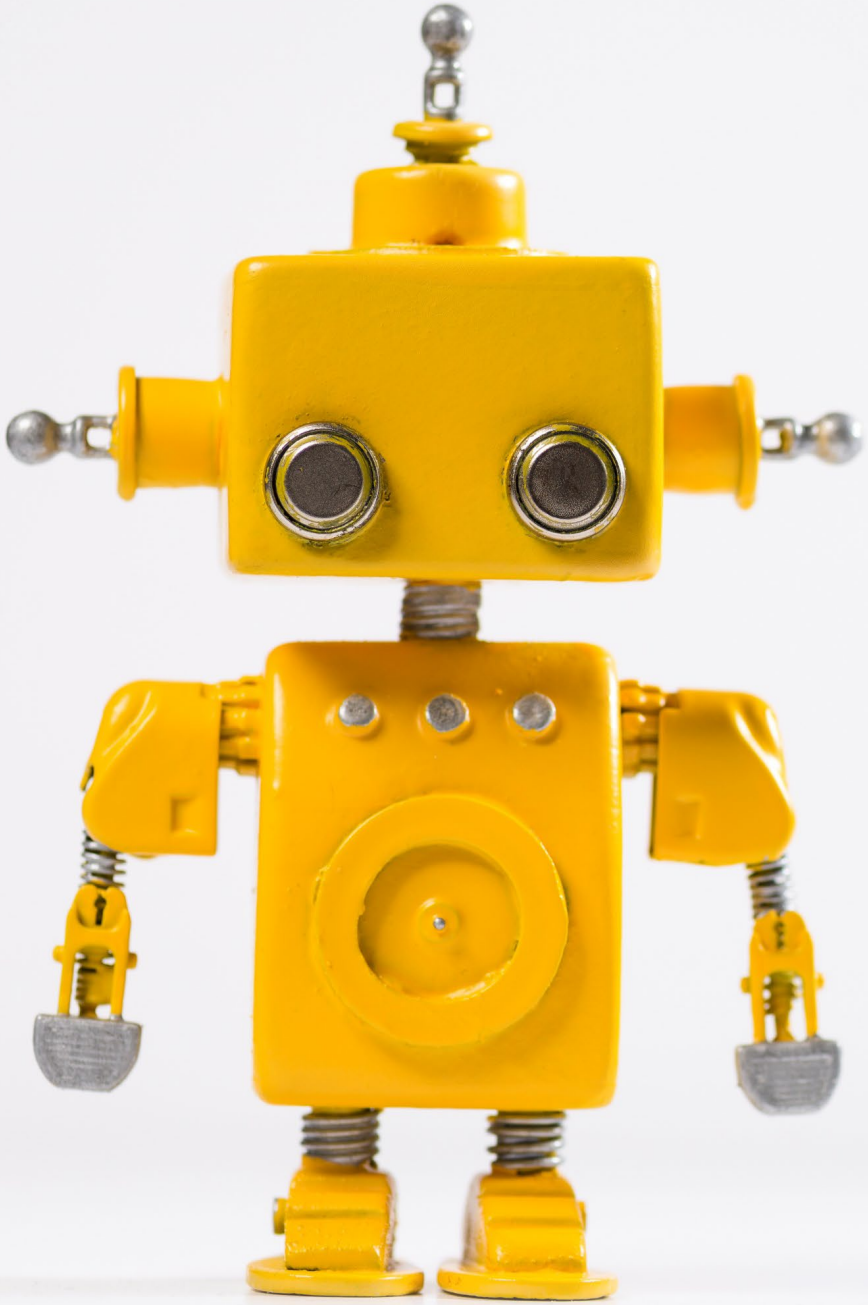
-목적에 맞는 내용만 강조

→ 분석 결과로 이야기 하고 싶은 내용에 집중

데이터 잉크 비율 높인 예

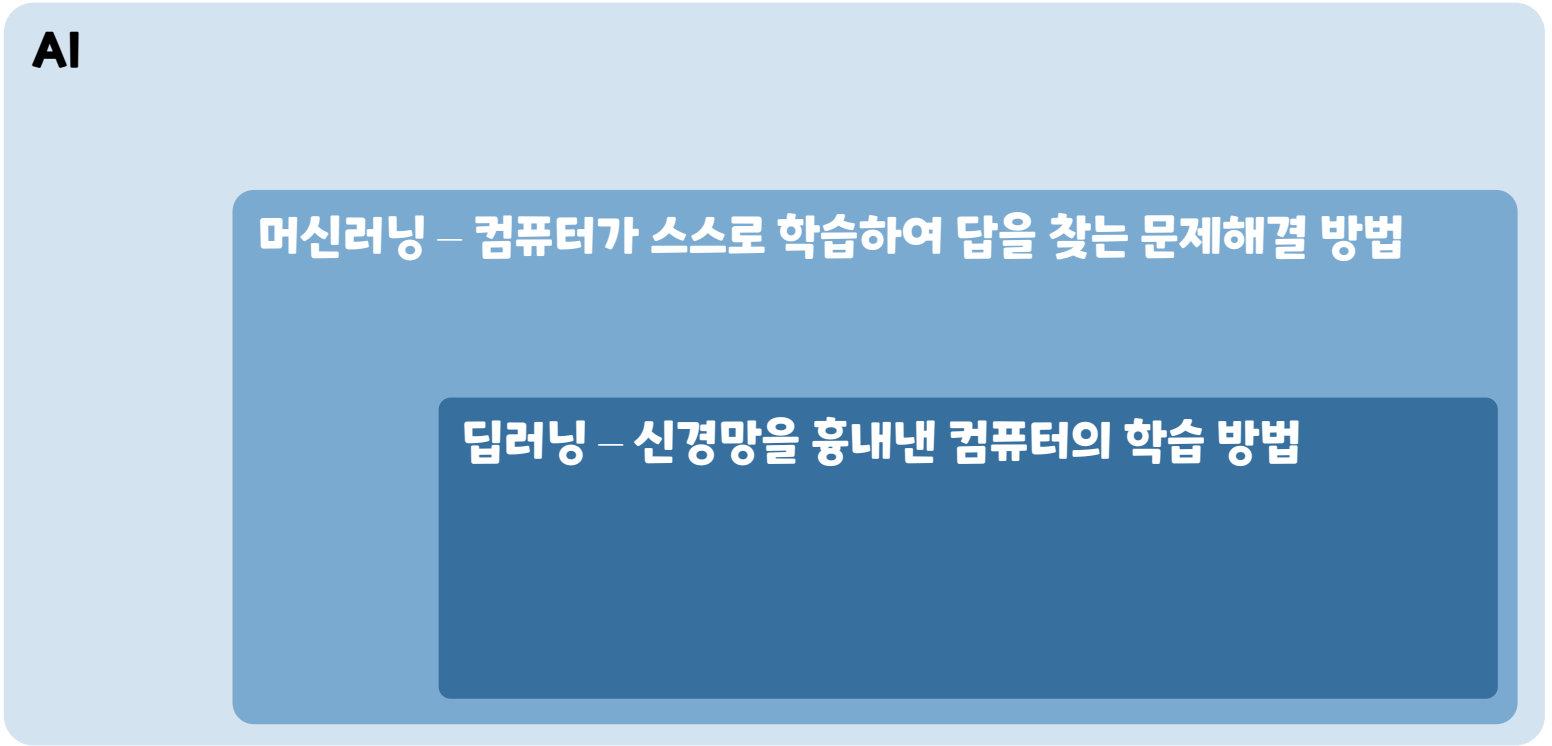

분기별 매출 트렌드
300,000
250,000
200,000
150,000
100,000
50,000
0
1사분기
2사분기
3사분기
4사분기
2019년
2020년
2021년
2022년
잔머리 대마왕
큰머리 마왕
일머리 마왕
거머리 마왕
주변머리 마왕
글머리 마왕
정신머리 마왕


1,000,000
500,000
0
2019년
2020년
2021년
2022년
잔머리 대마왕
큰머리 마왕
일머리 마왕
거머리 마왕
주변머리 마왕
글머리 마왕
정신머리 마왕



머신러닝과 모델링

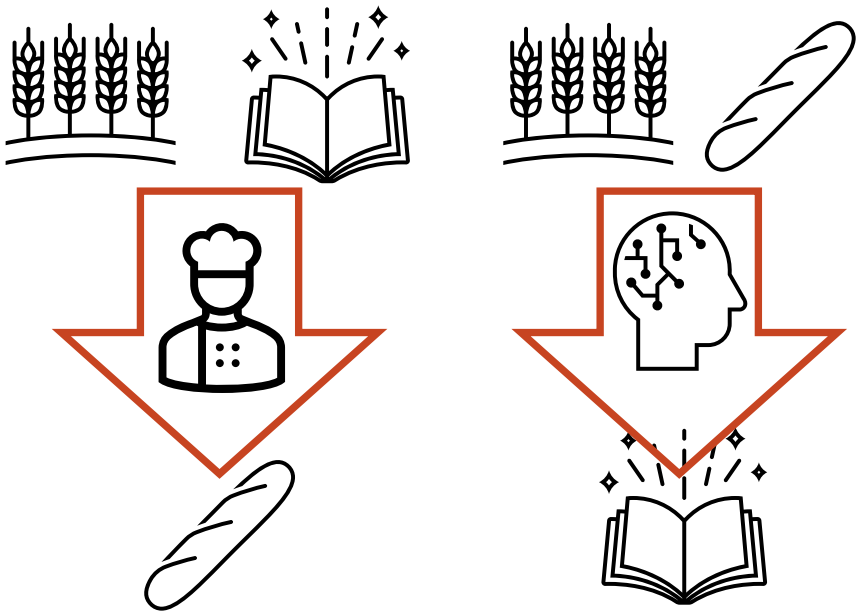
머신러닝?
딥러닝?



머신러닝

- 답과 답에 영향을 주는 인자 정보를 제공하고
  → 컴퓨터가 풀이법을 만들게 하는 알고리즘

  - 회귀 분석
  - 다항 로지스틱 회귀 모형
  - 의사결정나무
  - Random Forest
  - 서포트 벡터 머신
  - 딥러닝 (신경망모델)
  - …



https://youtu.be/CA5Ggqg5x6o

회귀 분석(Regression analysis)

- 관찰된 연속형 변수들에 대해 두 변수 사이의 모형을 구한 뒤 적합도를 측정하는 방식

선형 회귀 $$\mathrm{y_i} = \beta_1 x_{i1} + \ \ldots + \beta_p x_{ip} + \varepsilon_i$$

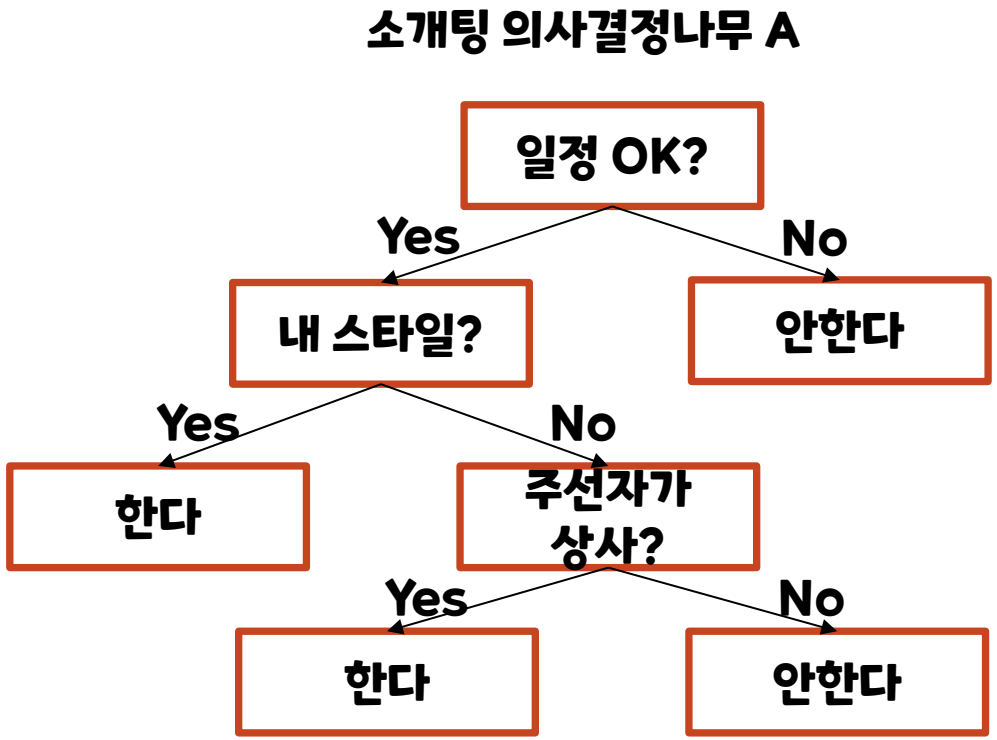
다항 로지스틱 회귀
(Multinomial Logistic Regression)

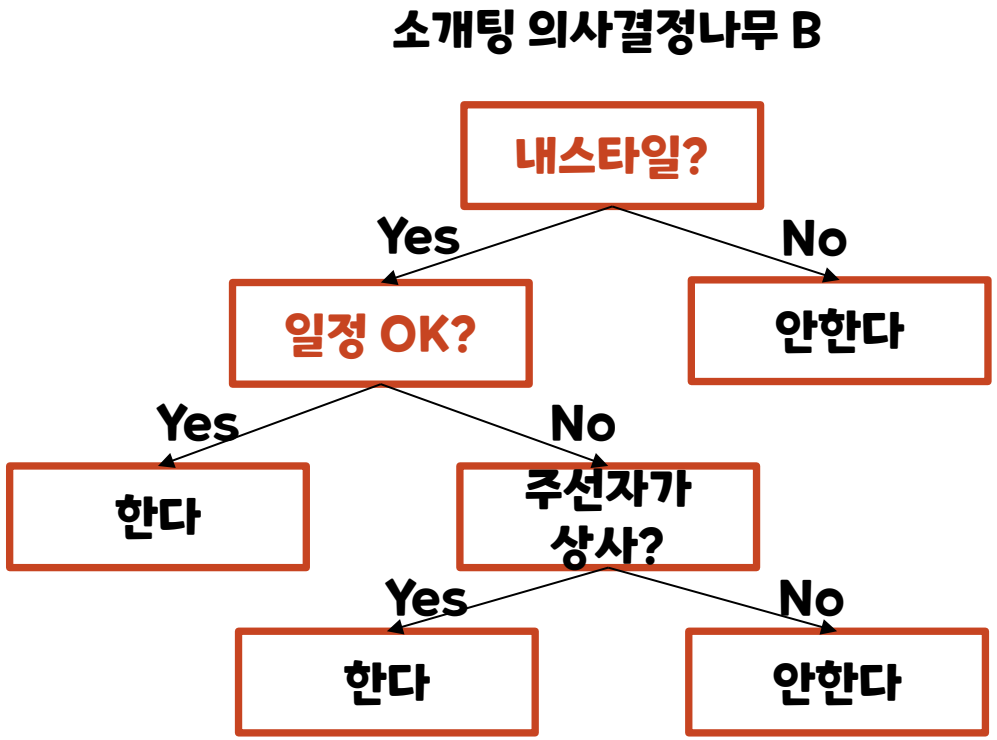
- Log를 활용한 계산식 형태로 확률 구하기

$$\mathrm{P}(Y = J) = \frac{1}{1 + \sum_{j=1}^{j=1} e^{\alpha_j + \beta_j x}}$$

의사 결정 나무(Decision tree)

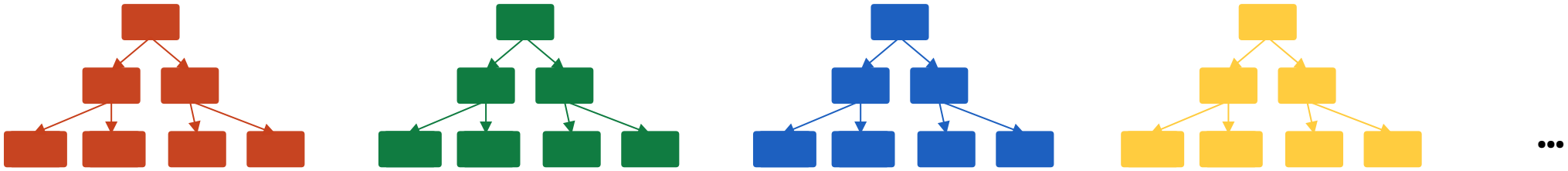
- 조건에 따라 의사 결정을 내리는 모델링 기법





Random Forest

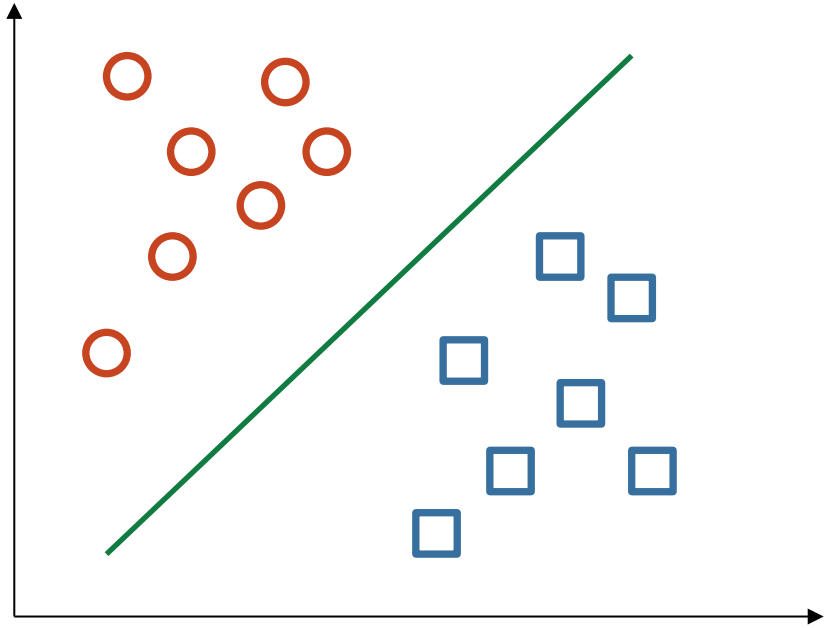
- 여러 개의 의사결정나무의 조합으로 결정



결론

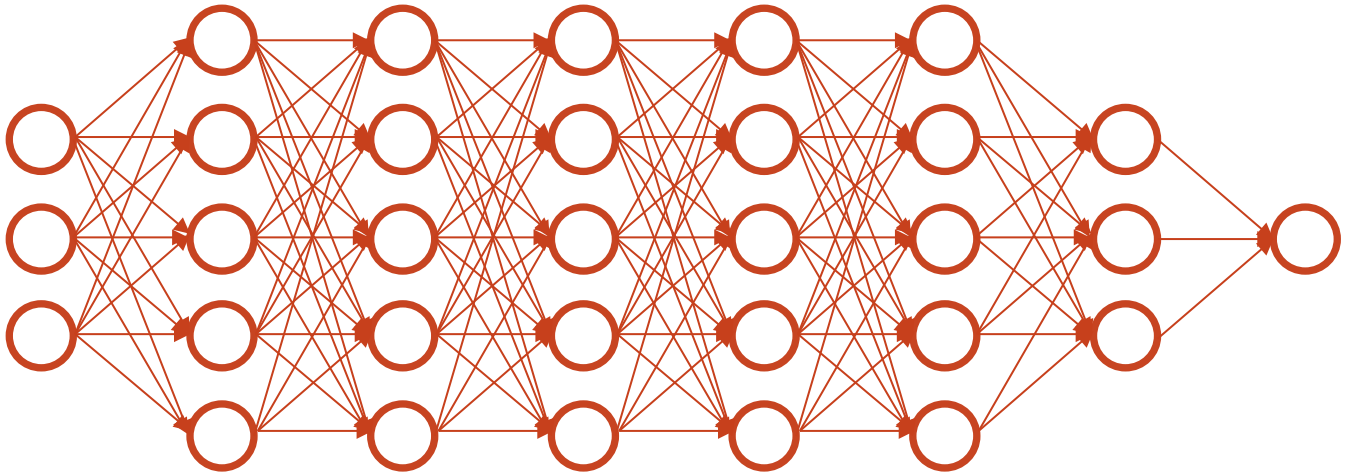
서포트 벡터(Support Vector)

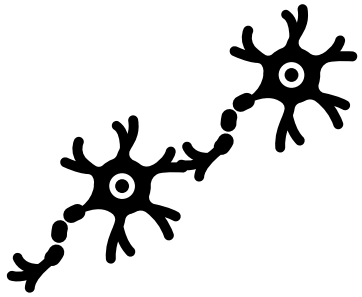
- 데이터 집단의 경계를 만들어 분류



딥러닝(Deep Learning)

- 신경망 구조를 모방한 학습 모델



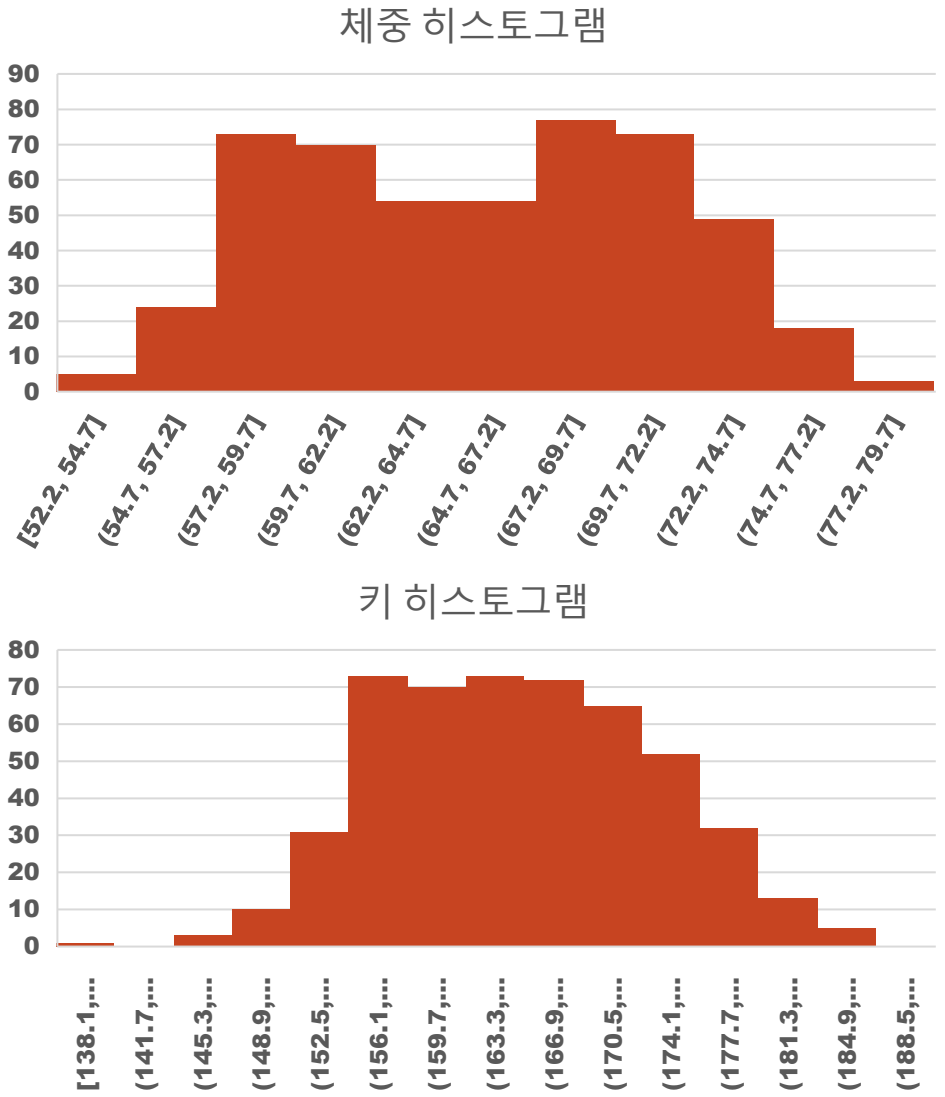



기초 통계 분석
시작해 보기

엑셀로 기본 통계 분석을 해보자

	A	B	C	D	E
1	ID	체중	키	성별	나이
2	1	62.9	161.6	0	44
3	2	61.8	168.0	1	42
4	3	60.7	158.3	0	34
5	4	67.6	174.2	1	22
6	5	68.6	177.8	1	25
7	6	58.9	157.6	0	56
8	7	76.1	173.5	1	56
9	8	71.2	167.0	1	53
10	9	68.3	175.3	1	52
11	10	72.4	180.1	1	27
12	11	73.7	172.3	1	58

남/여, 연령대별
체중 & 키 분포
(가상 Data, 500)



엑셀로 기본 통계 분석을 해보자

	A	B	C	D
1	체중		키	
2				
3	평균	65.6316	평균	166.6448
4	표준 오차	0.258075044	표준 오차	0.369880859
5	중앙값	66.1	중앙값	166.2
6	최빈값	60.9	최빈값	173.5
7	표준 편차	5.770733417	표준 편차	8.270787436
8	분산	33.30136417	분산	68.40592481
9	첨도	-1.042653911	첨도	-0.504611269
10	왜도	-0.01270554	왜도	0.081708473
11	범위	26.9	범위	50.4
12	최소값	52.2	최소값	138.1
13	최대값	79.1	최대값	188.5
14	합	32815.8	합	83322.4
15	관측수	500	관측수	500
16	신뢰 수준(95.0%)	0.507047624	신뢰 수준(95.0%)	0.726715794

기본 기술 통계

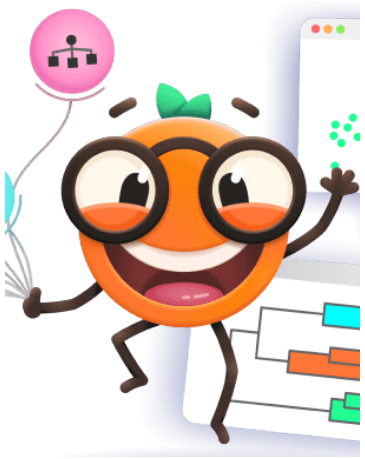
	A	B	C	D	E
1		체중	키	성별	나이
2	체중	1			
3	키	0.793278	1		
4	성별	0.82299	0.764911	1	
5	나이	0.206245	-0.10872	0.0669	1

상관 관계 분석

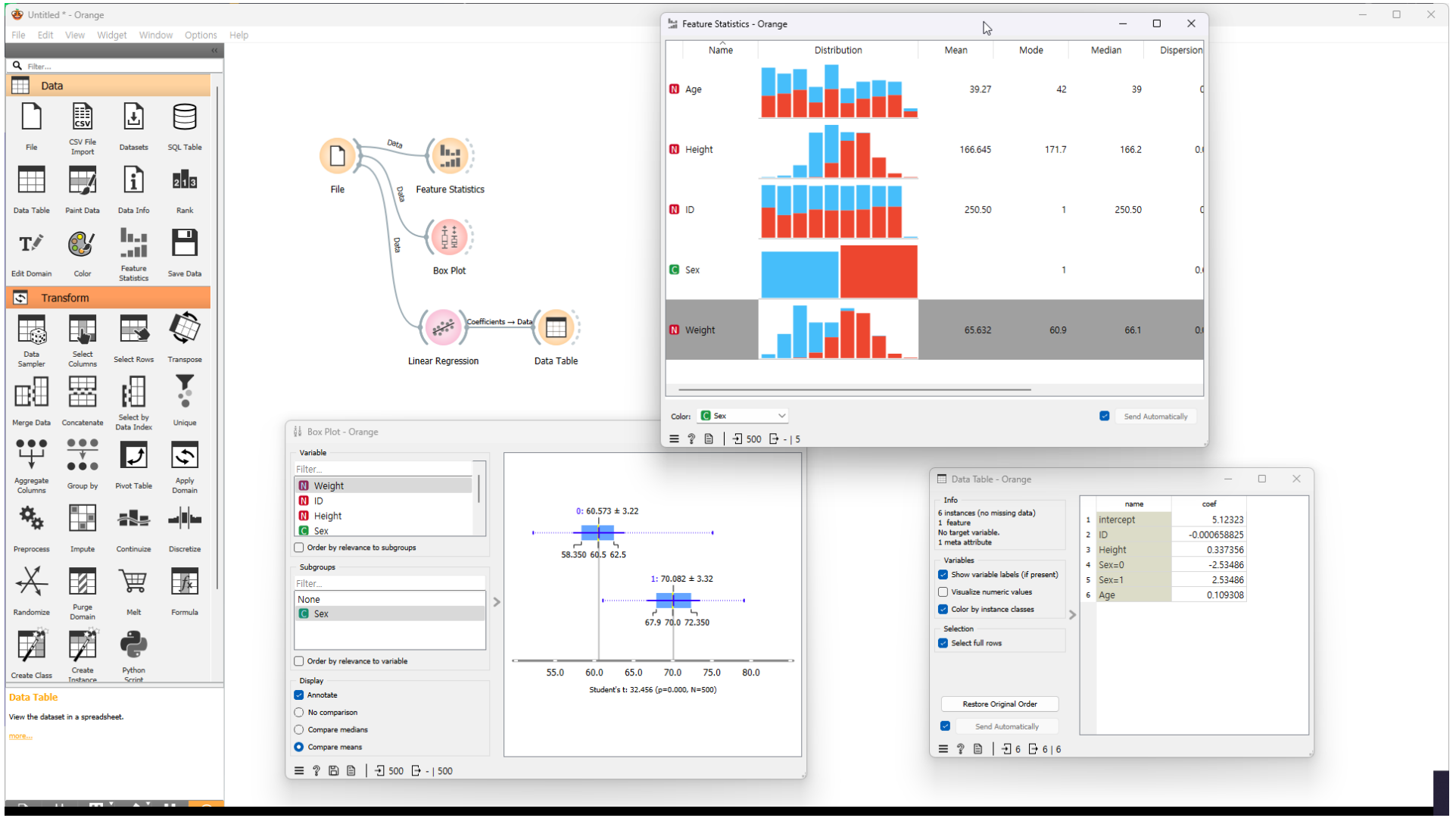
엑셀로 기본 통계 분석을 해보자

	A	B	C	D	E	F	G	H	I
1	요약 출력								
2									
3	회귀분석 통계량								
4	다중 상관계수	0.889450779							
5	결정계수	0.791122689							
6	조정된 결정계수	0.789859318							
7	표준 오차	2.64536791							
8	관측수	500							
9									
10	분산 분석								
11		자유도	제곱합	제곱 평균	F 비	유의한 F			
12	회귀	3	13146.38692	4382.128972	626.199899	3.4E-168			
13	잔차	496	3470.993805	6.99797138					
14	계	499	16617.38072						
15									
16		계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
17	Y 절편	2.455109238	3.7897836	0.647823068	0.517399141	-4.9909	9.901118	-4.9909	9.901118
18	키	0.337087847	0.022950171	14.68781399	8.24485E-41	0.291996	0.382179	0.291996	0.382179
19	성별	5.061538336	0.378626788	13.36814639	4.99928E-35	4.317628	5.805448	4.317628	5.805448
20	나이	0.109748296	0.010156378	10.80585018	1.38245E-24	0.089793	0.129703	0.089793	0.129703

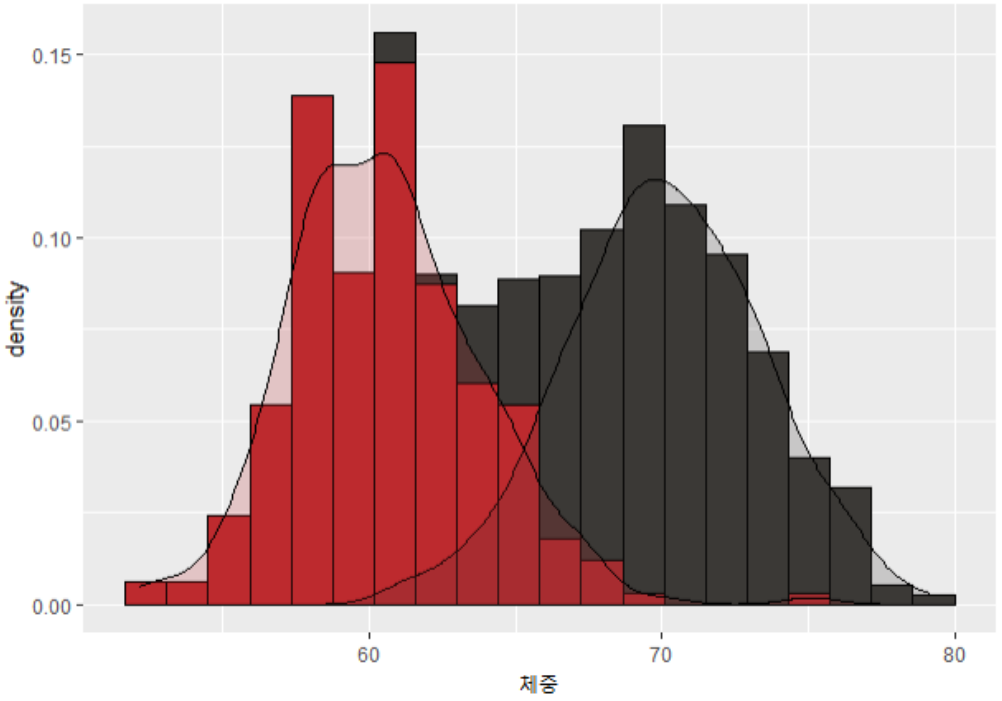
회귀 분석

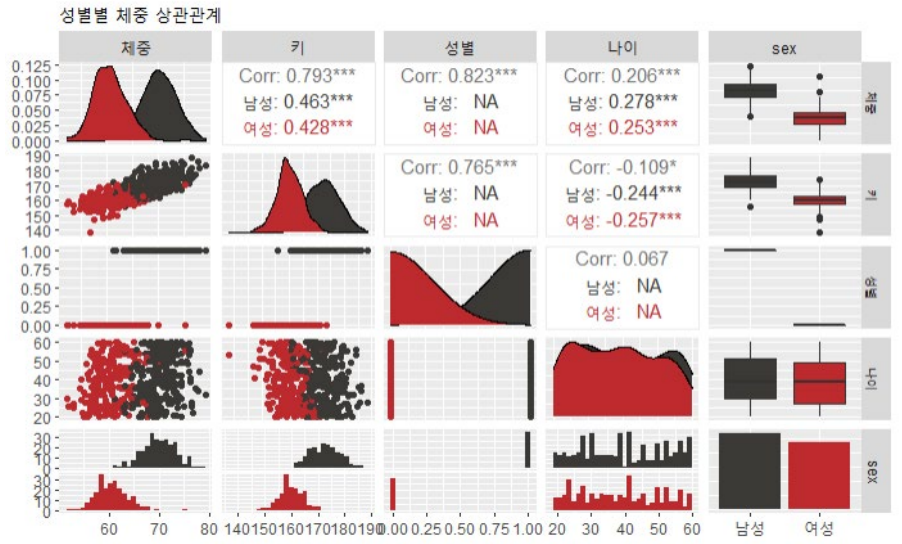


Orange로 기초 통계 분석을 해 보자



R로 기초 통계 분석을 해 보자

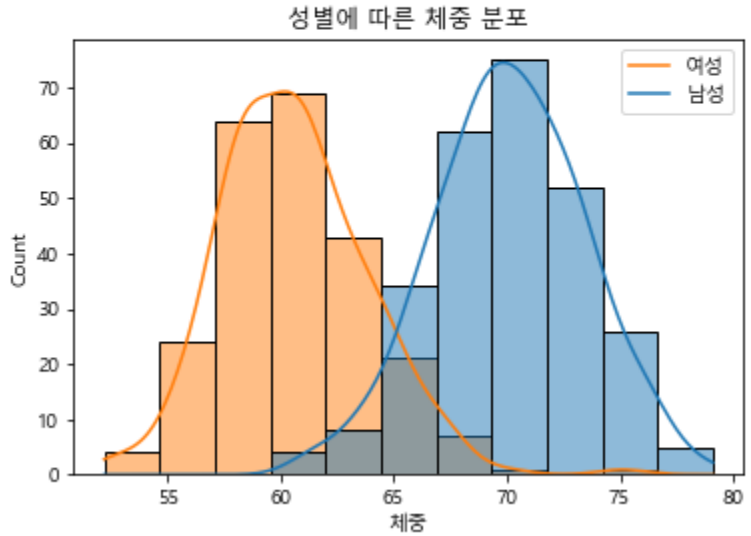




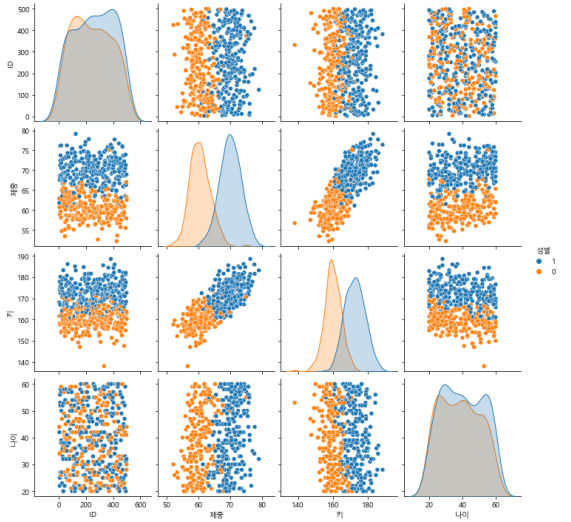

```
Call:
lm(formula = 체중 ~ 키 + 성별 + 나이, data = dfAnal)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3840 -1.7161 -0.0144  1.8531  9.0678

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.45511    3.78978   0.648    0.517
키           0.33709    0.02295  14.688   <2e-16 ***
성별         5.06154    0.37863  13.368   <2e-16 ***
나이         0.10975    0.01016  10.806   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.645 on 496 degrees of freedom
Multiple R-squared:  0.7911,    Adjusted R-squared:  0.7899
F-statistic: 626.2 on 3 and 496 DF,  p-value: < 2.2e-16
```

파이썬으로 기초 통계 분석을 해 보자

```
In [7]:
sns.histplot(data = df,x="체중",hue="성별",hue_order=[1,0])
plt.title("성별에 따른 체중 분포")
plt.legend(['여성', '남성'])
plt.show()
```




```
In [20]:
sns.pairplot(data = df,hue ="성별",hue_order=[1,0])
plt.show()
```



	coef	std err	t	P>|t|	[0.025	0.975]
const	2.4551	3.790	0.648	0.517	-4.991	9.901
x1	0.3371	0.023	14.688	0.000	0.292	0.382
x2	5.0615	0.379	13.368	0.000	4.318	5.805
x3	0.1097	0.010	10.806	0.000	0.090	0.130

Omnibus:	1.105	Durbin-Watson:	1.960
Prob(Omnibus):	0.576	Jarque-Bera (JB):	0.904
Skew:	0.067	Prob(JB):	0.636
Kurtosis:	3.160	Cond. No.	5.51e+03

AI에게 모든 과정을 시켜 보자

데이터를 분석하고 시각화 하는 파이썬 코드를 만들어줘.

1. 분석할 데이터는 "Weight_Sample.xlsx"라는 이름의 파일이야
2. "Sheet1"에 분석할 데이터가 있어.
3. 첫번째 행은 머리글행이고, "ID", "Weight", "Height", "Sex", "Age"의 데이터가 있어.
4. "Weight"와 "Height"의 히스토그램을 그려줘
5. "Weight", "Height", "Sex", "Age"의 상관관계를 분석하고 시각화 하여 보여줘
6. 성별을 색깔로 그룹핑하고, "Weight", "Height", "Age"의 상관관계를 분석하고 시각화해서 보여줘
7. "Weight"를 Y값으로 하고 나머지를 X값으로 하여 선형회귀식을 구해줘

※ 샘플 데이터는

- 가상의 데이터
- 성별, 나이는 Random 함수로 무작위 취출
- 나이별 평균 키 = 0.1 * (나이 – 20)
- 체중 = 0.35 * 키 – 0.12 * 나이 + 4.8(남자일경우)
- 상기 기준으로 임의 값을 구하되 노이즈를 섞는 형태로 생성

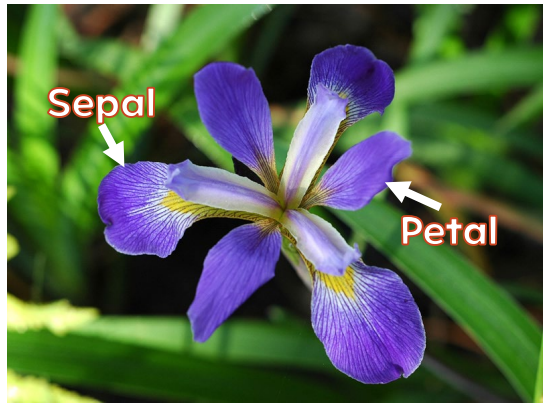

Advanced 통계 분석
도전해 보기

Iris flower Dataset

- 머신러닝 학습용 데이터셋

Iris versicolor

Iris setosa

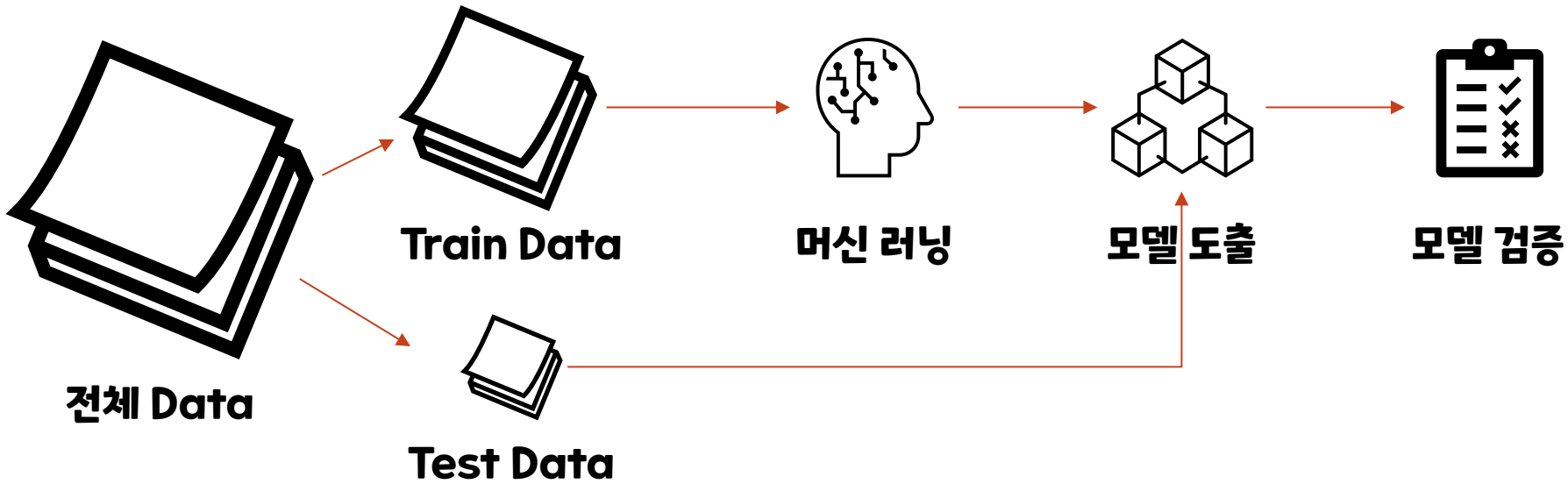


Iris virginica

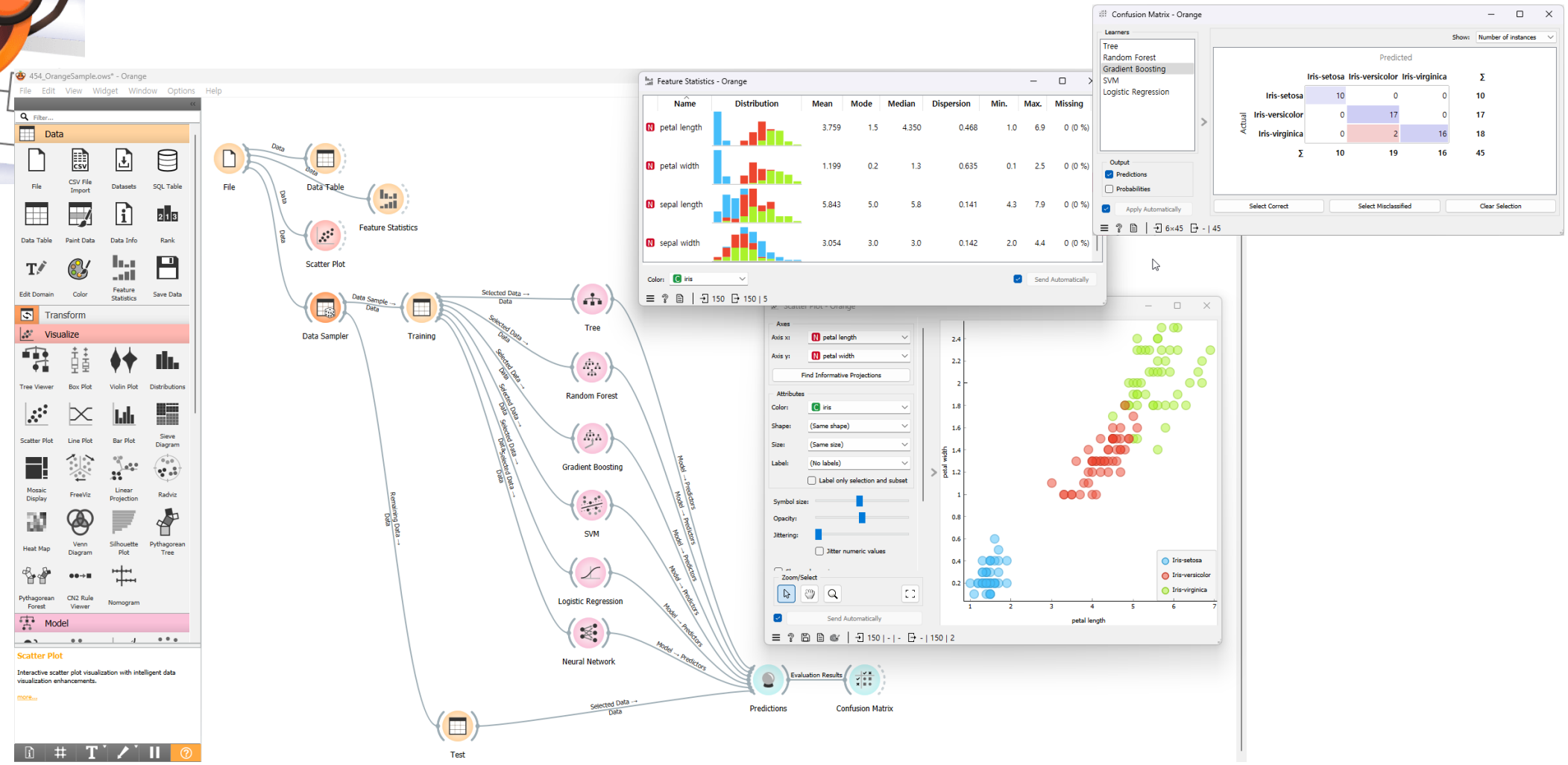
- 3종의 붓꽃의 꽃잎, 꽃 받침 길이와 폭을 정리한 데이터표
- 150개의 측정값이 있으며 이를 활용하여 머신러닝을 연습할 수 있다.

일반적인 모델링 방법론

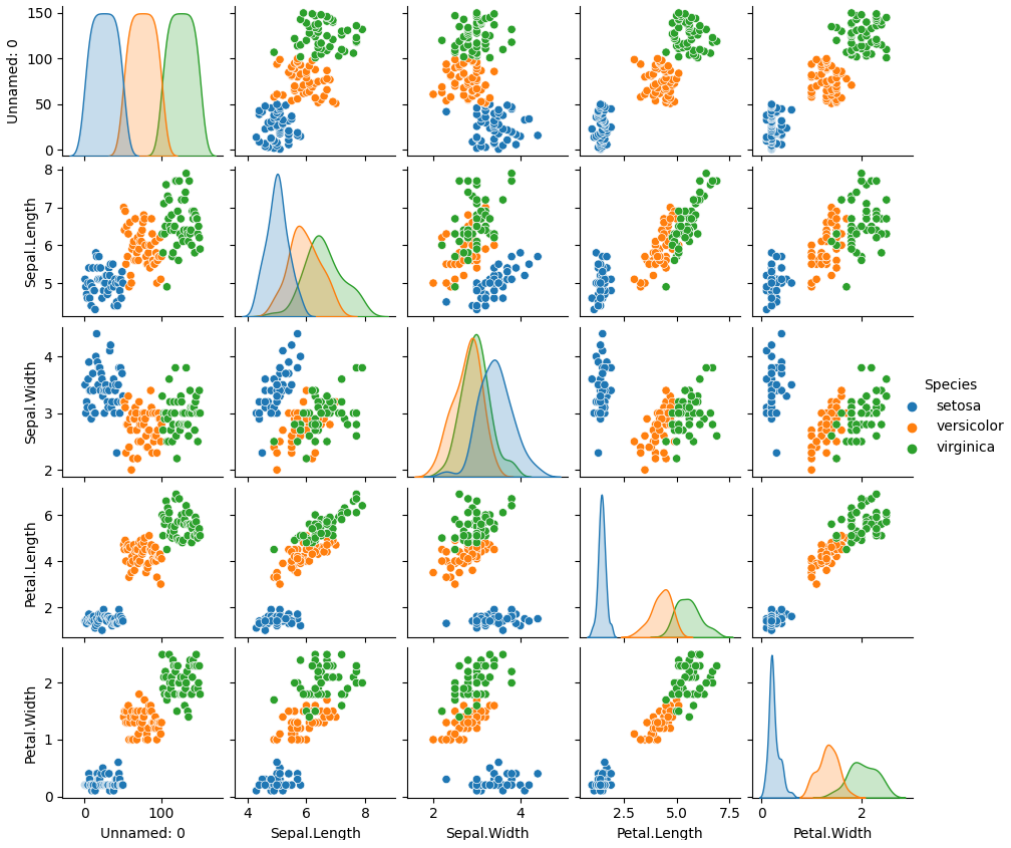
- 전체 데이터를 Train data 와 Test Data로 분리
- 모델 Train 후 Test Data로 결과 확인

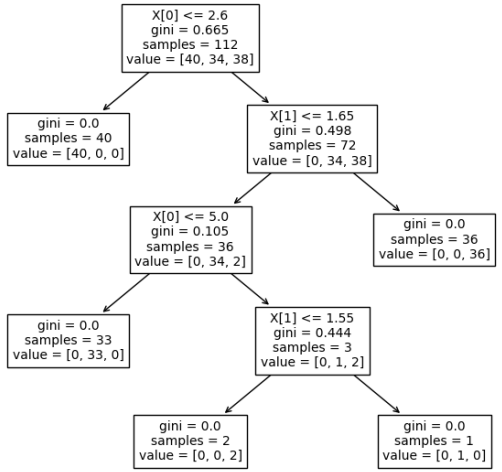


Orange로 모델링을 해 보자



파이썬으로 모델링을 해 보자






```
다항 로지스틱 회귀
Species     setosa  versicolor  virginica
row_0
setosa          10           0          0
versicolor       0          14          1
virginica        0           2         11
의사 결정 나무
Species     setosa  versicolor  virginica
row_0
setosa          10           0          0
versicolor       0          14          2
virginica        0           2         10
Random Forest
Species     setosa  versicolor  virginica
row_0
setosa          10           0          0
versicolor       0          14          1
virginica        0           2         11
Support Vector
Species     setosa  versicolor  virginica
row_0
setosa          10           0          0
versicolor       0          16          0
virginica        0           0         12
인공 신경망
Species     setosa  versicolor  virginica
row_0
setosa          10           0          0
versicolor       0          15          0
virginica        0           1         12

Process finished with exit code 0
```

AI에게 모든 과정을 시켜 보자

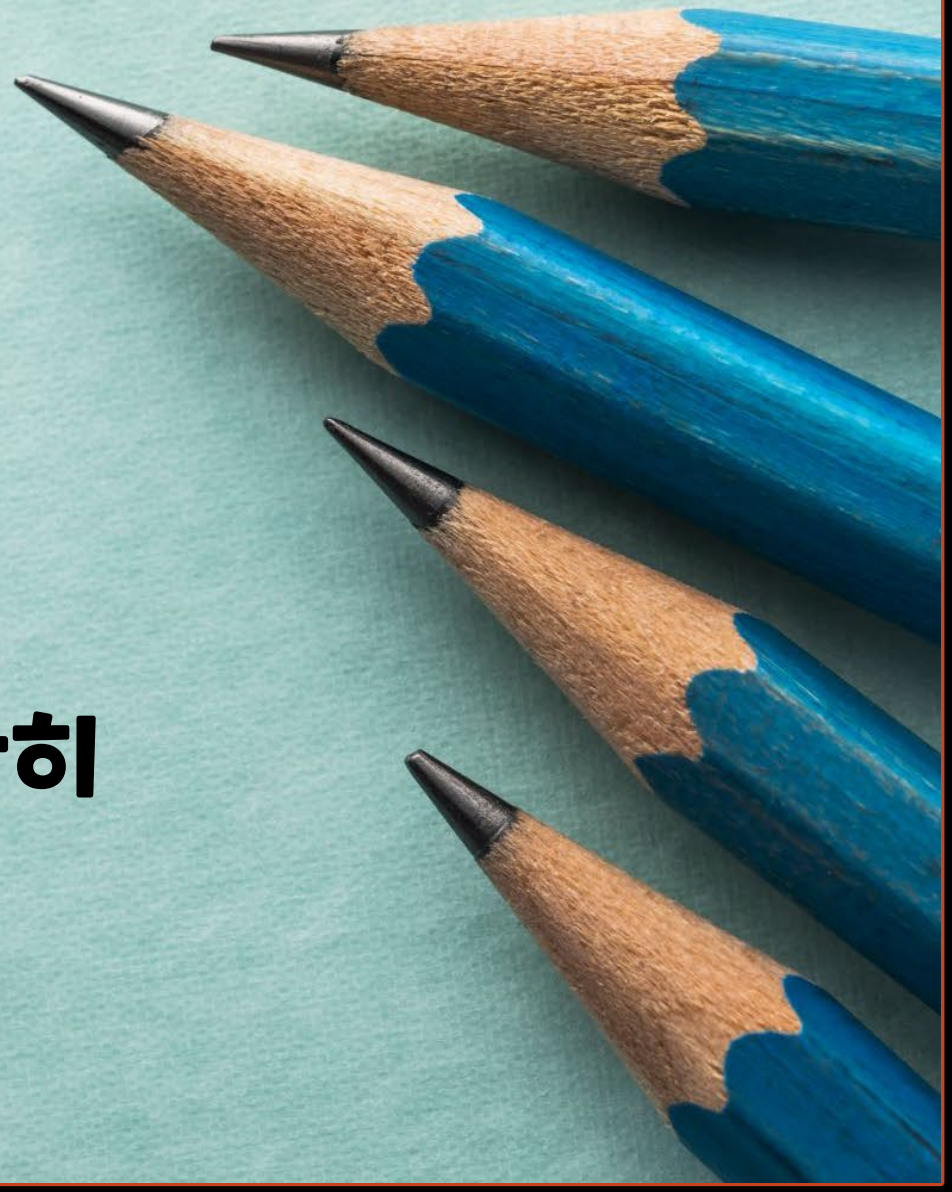
머신러닝에 대한 대표적인 데이터중의 하나인 IRIS 데이터를 활용해서 머신러닝 모델을 비교하는 코드를 만들어줘.

1. IRIS 데이터를 생성한 다음
2. 꽃의 종류에 따라 다른 색깔로 표시하고 각 변수간의 산점도를 한눈에 볼 수 있게 비교해서 그려줘
3. 각 변수간의 상관계수도 한번에 보여줘
4. 70%의 데이터를 training 데이터로, 30%의 데이터를 test 데이터로 분리하고
5. training 데이터에 대해 Logic Tree, Random forest, SVM, CNN 모델로 꽃의 종류를 예측하는 모델을 모델링해줘
6. 만들어진 모델을 test 데이터로 평가해줘
7. 평가 결과를 confusion matrix로도 그려줘

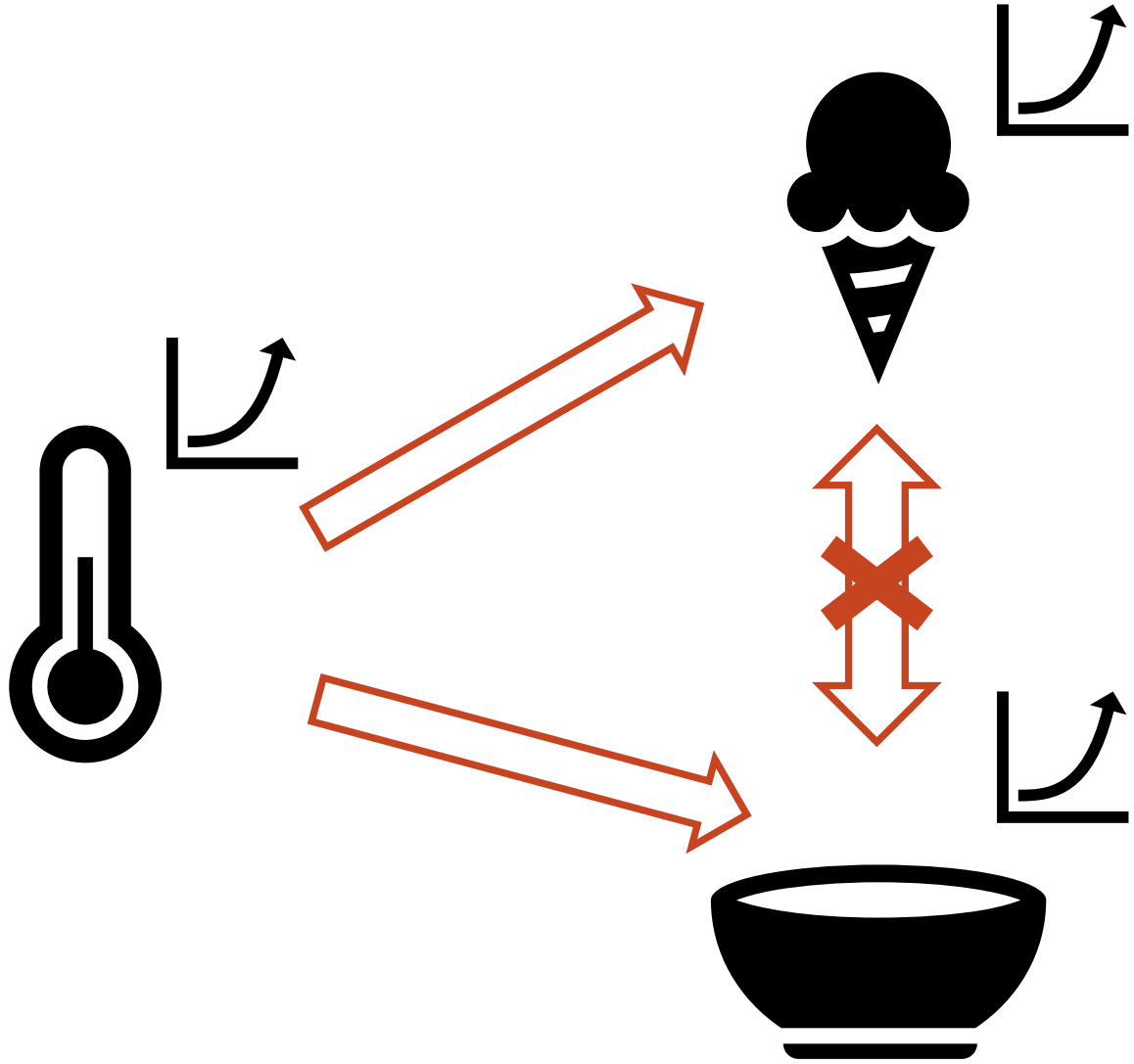

데이터
분석시
주의할 점

실제 회사에서는…

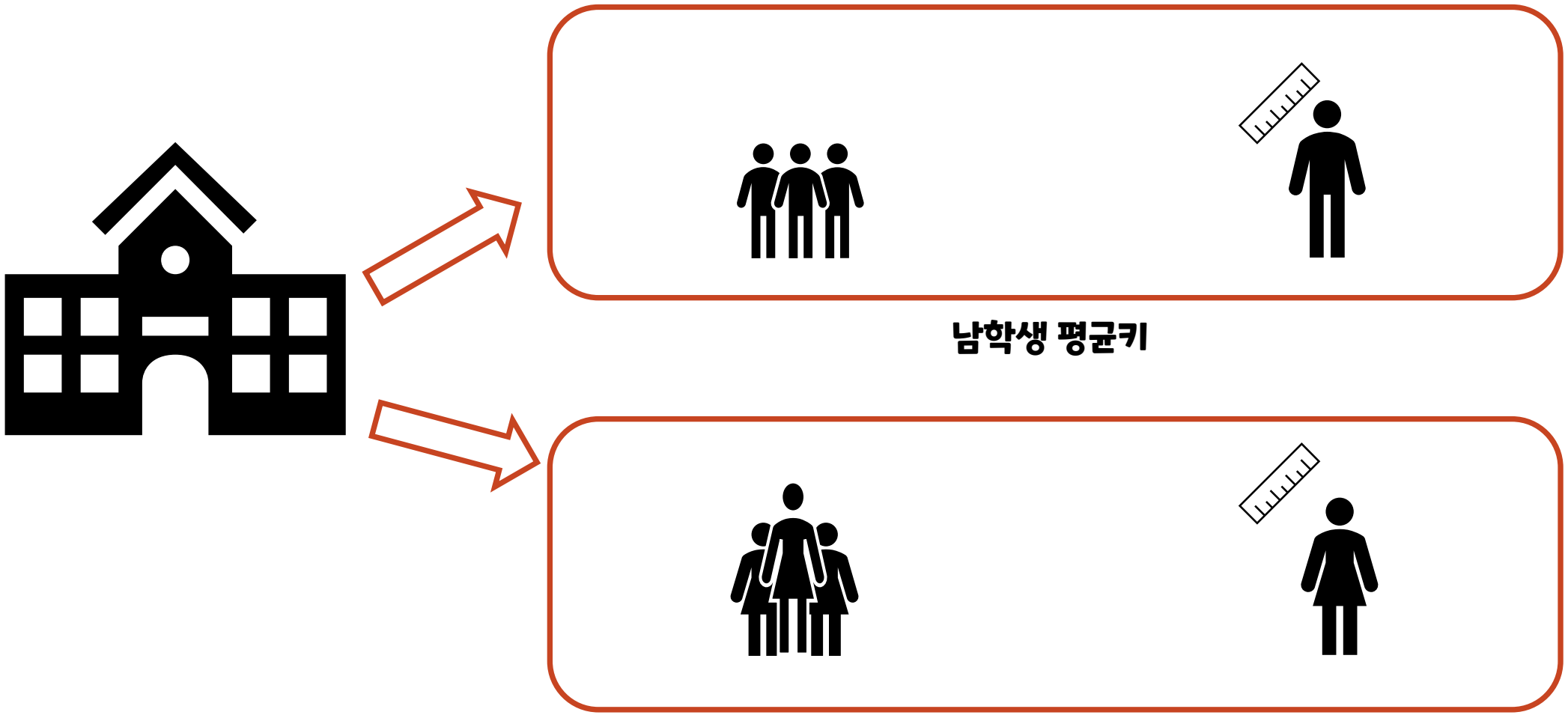
- 선진 기법의 활용 여부 보다
- 실제 필요에 맞는 사용이 중요
- 개념을 이해하고, 목적을 명확히
- 기본을 먼저 !



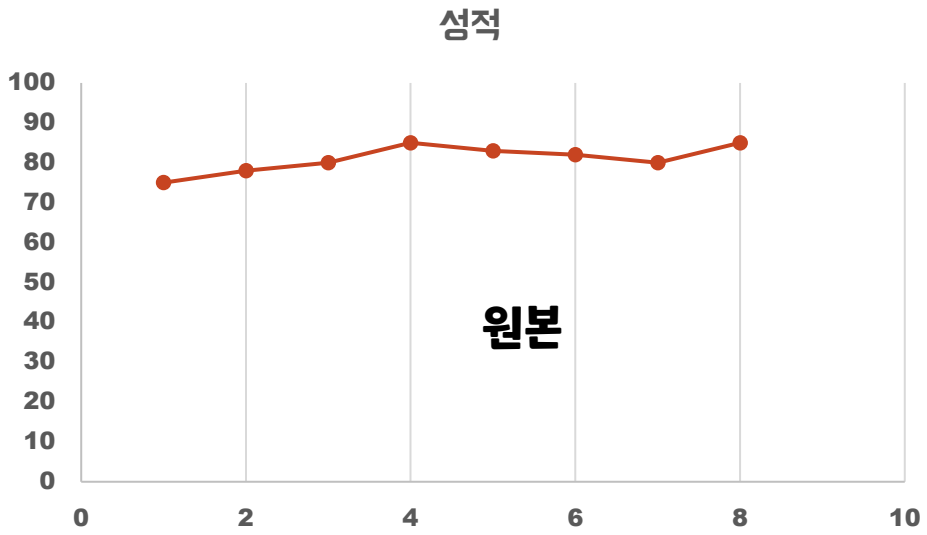
헷깔리지 말자! 상관관계와 인과관계

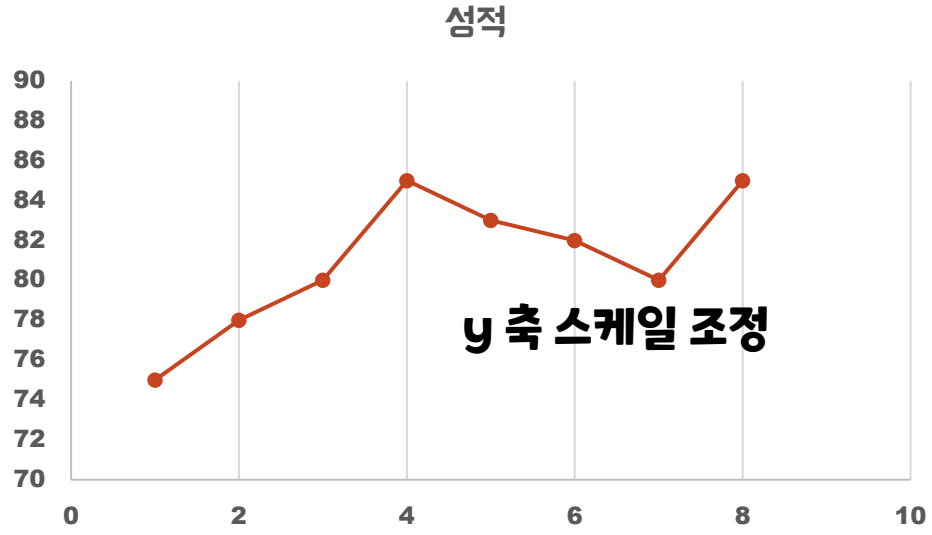


한번더 확인하자! 샘플 적합성

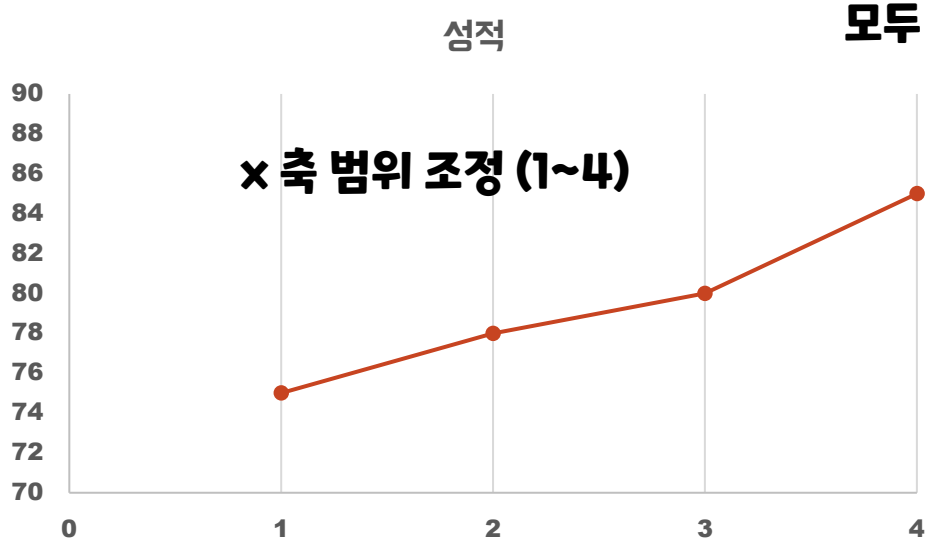


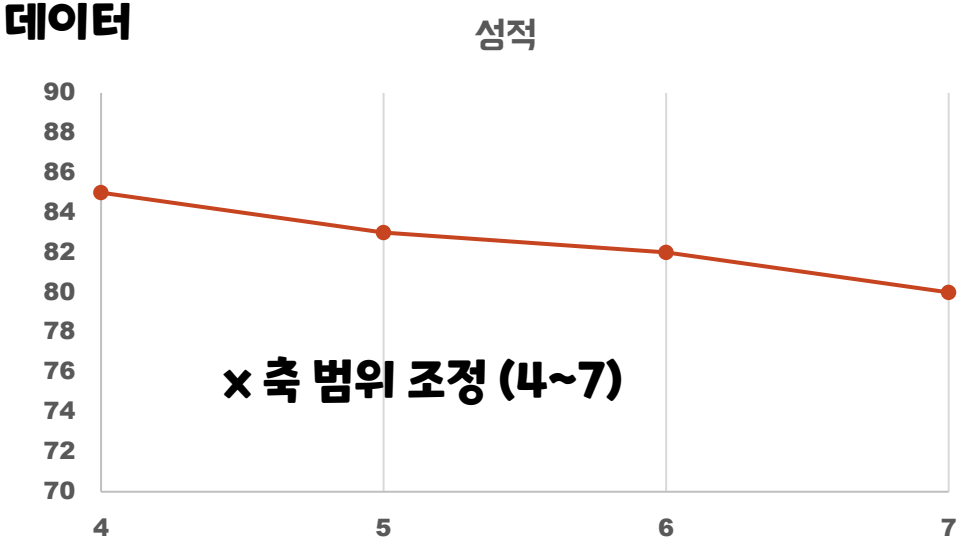
속지 말자! 스케일과 범위



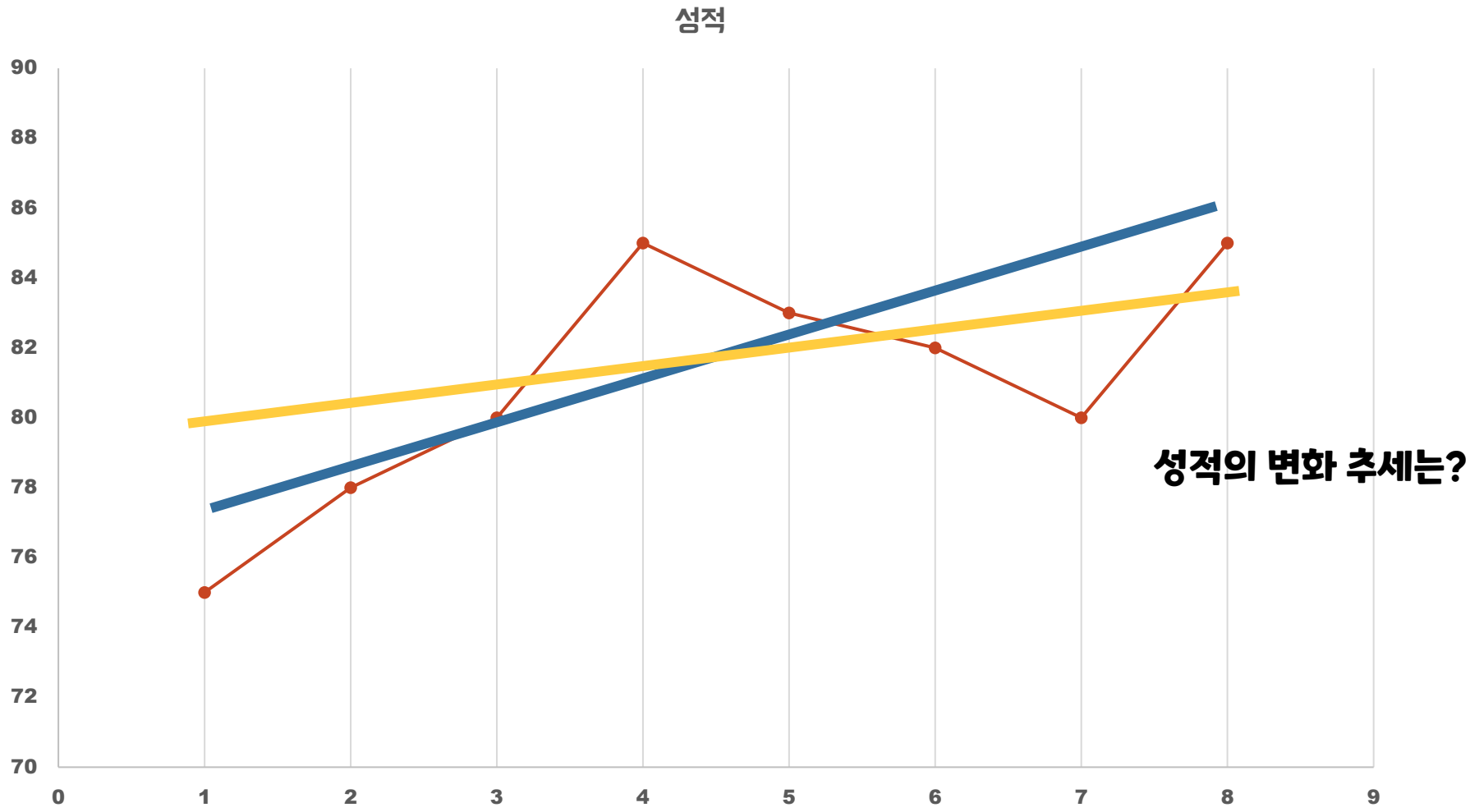


모두 같은 원본 데이터





다시 보자! 추세선



성적의 변화 추세는?

회사에서 진짜 데이터를 쓰려면

- 목적이 명확해야 함
- 오히려, 기본이 강력하다!
- 데이터 분석을 지시할 때

  → 기초 통계 체크는 하고, 기본 원리는 이해하자
- 데이터를 분석할 때

  → 리더의 Wants를 헤아리자


추천
도서 & Site

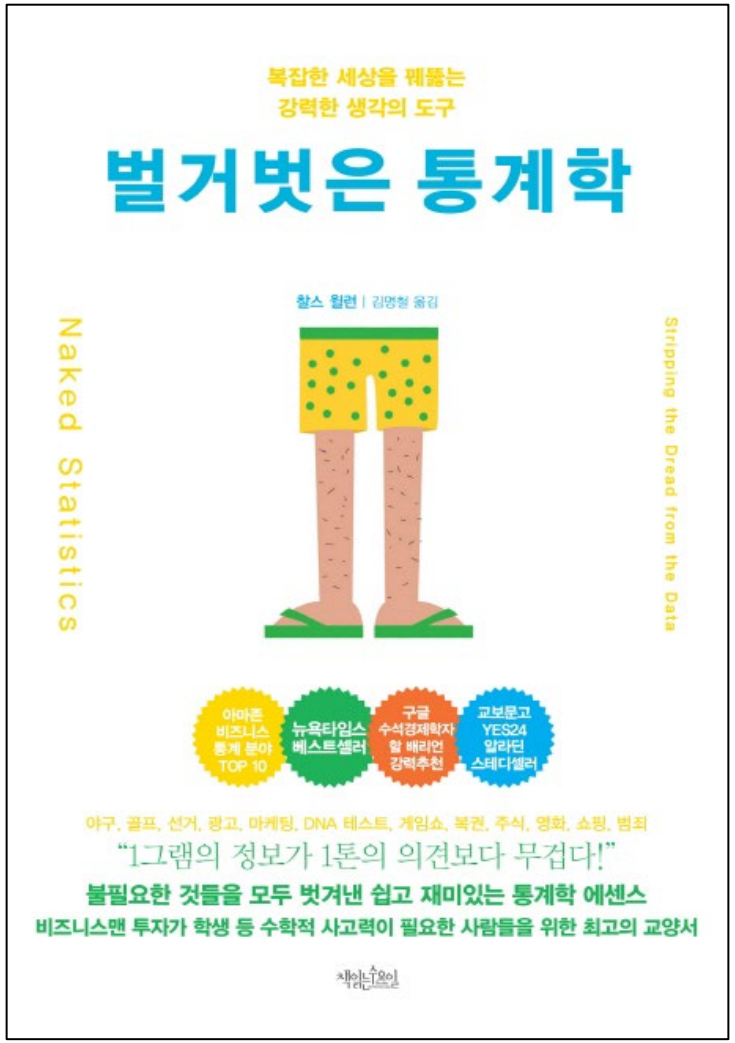
추천 도서 (감 잡기)


데이터 문해력 소양서
직장인이여, 데이터 문해력을 키워라
데이터 분석가의
숫자유감
만화로 배우는 업무 데이터 분석 상식
글 권정민 · 만화 주형
리디북스 베스트 아티클
신규 에피소드 추가
통계 미니 사전
GOLDEN RABBIT

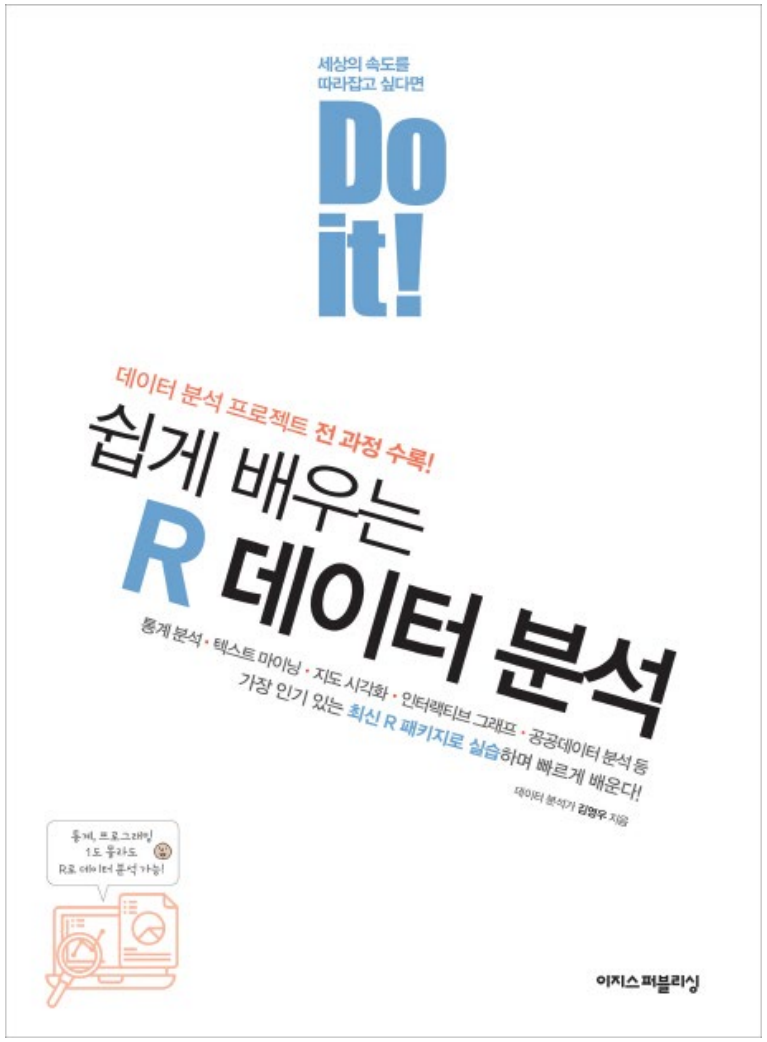

감으로만 일하던 김 팀장은 어떻게 데이터 좀 아는 팀장이 되었나
비전공자를 위한 데이터 분석 속성 스쿨
황보현우 · 김철수 지음
"조직에서 데이터를 이용해 문제를 해결하는 과정을 쉽게 풀어냈다!"
— 이준기 연세대 교수, 전 한국빅데이터학회장
실무에서 가장 많이 쓰는 분석법 Best 17 수록!
한빛비즈

추천 도서 (속지말자! 통계)


새빨간
거짓말,
통계
대릴 허프 지음 · 박영훈 옮김
빌 게이츠가 두 번이나
강력히 추천한 책!!
(TED 강연, 휴가 필독서)
1950년 이후로 출간된
최고의 책 중 하나로 추천한다.
-빌 게이츠 TED 강연 중에서

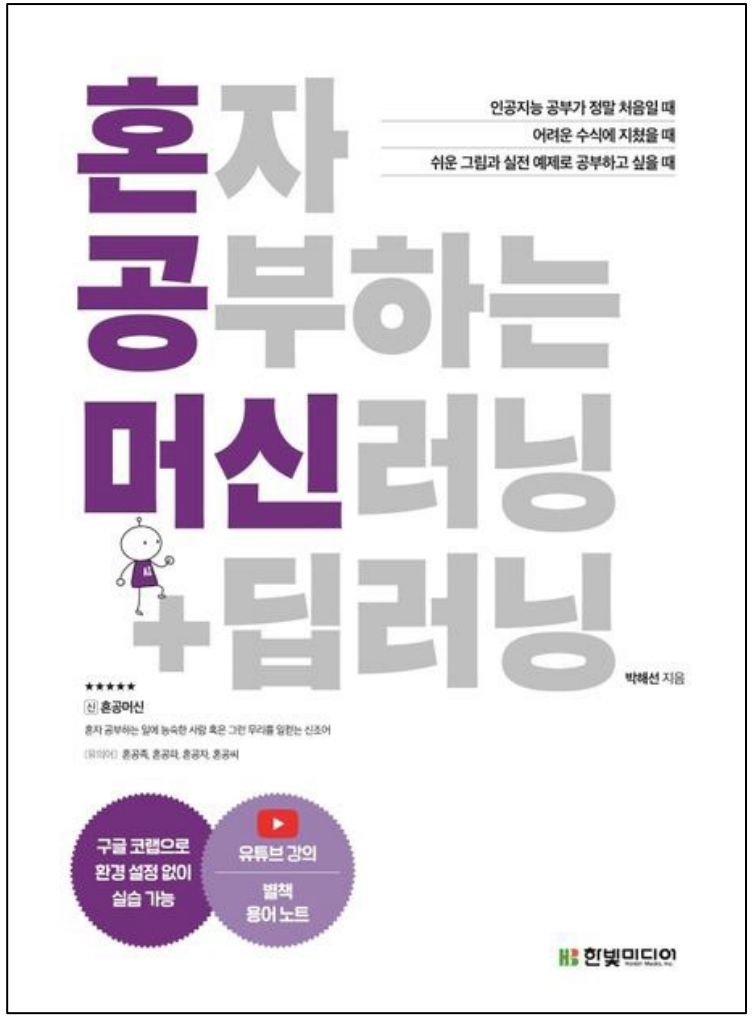

복잡한 세상을 꿰뚫는
강력한 생각의 도구
벌거벗은 통계학
찰스 윌런 | 김명철 옮김
Naked Statistics
Stripping the Dread from the Data
아마존 비즈니스 통계 분야 TOP 10
뉴욕타임스 베스트셀러
구글 수석경제학자 할 배리언 강력추천
교보문고 YES24 알라딘 스테디셀러
야구, 골프, 선거, 광고, 마케팅, DNA 테스트, 게임쇼, 복권, 주식, 영화, 쇼핑, 범죄
"1그램의 정보가 1톤의 의견보다 무겁다!"
불필요한 것들을 모두 벗겨낸 쉽고 재미있는 통계학 에센스
비즈니스맨 투자가 학생 등 수학적 사고력이 필요한 사람들을 위한 최고의 교양서

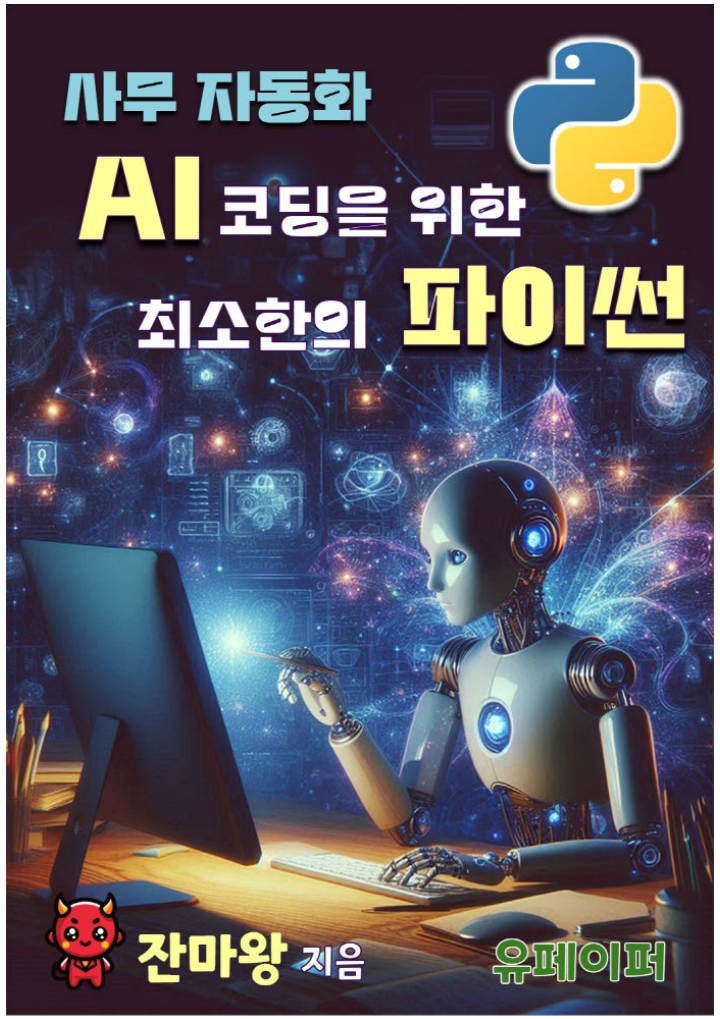
추천 도서 (심화 - Tool)


세상의 속도를
따라잡고 싶다면
Do it!
데이터 분석 프로젝트 전 과정 수록!
쉽게 배우는
R 데이터 분석
통계 분석 · 텍스트 마이닝 · 지도 시각화 · 인터랙티브 그래프 · 공공데이터 분석 등
가장 인기 있는 최신 R 패키지로 실습하며 빠르게 배운다!
데이터 분석가 김영우 지음
통계, 프로그래밍
1도 몰라도
R로 데이터 분석 가능!
이지스퍼블리싱


세상의 속도를
따라잡고 싶다면
Do it!
데이터 분석 프로젝트 전 과정 수록!
쉽게 배우는
파이썬 데이터 분석
통계 분석 · 머신러닝 모델링 · 텍스트 마이닝 · 공공데이터 분석 등
가장 많이 쓰는 파이썬 패키지로 데이터 분석에 입문한다!
데이터 분석가 김영우 지음
통계학, 파이썬 몰라도
바로 실습 가능!
필요한 건 이 책에
다 있으니까!
이지스퍼블리싱

추천 도서 (심화 - Tool)


혼자 공부하는 머신러닝 + 딥러닝
인공지능 공부가 정말 처음일 때
어려운 수식에 지쳤을 때
쉬운 그림과 실전 예제로 공부하고 싶을 때
박해선 지음
★★★★★
신 혼공머신
혼자 공부하는 일에 능숙한 사람 혹은 그런 무리를 일컫는 신조어
(유의어) 혼공족, 혼공파, 혼공자, 혼공씨
구글 코랩으로 환경 설정 없이 실습 가능
유튜브 강의
별책 용어 노트
한빛미디어


사무 자동화
AI 코딩을 위한
최소한의 파이썬
잔마왕 지음
유페이퍼


잔마왕의 사무자동화
파이썬 실습

추천 도서 & 채널 (심화 & 응용)

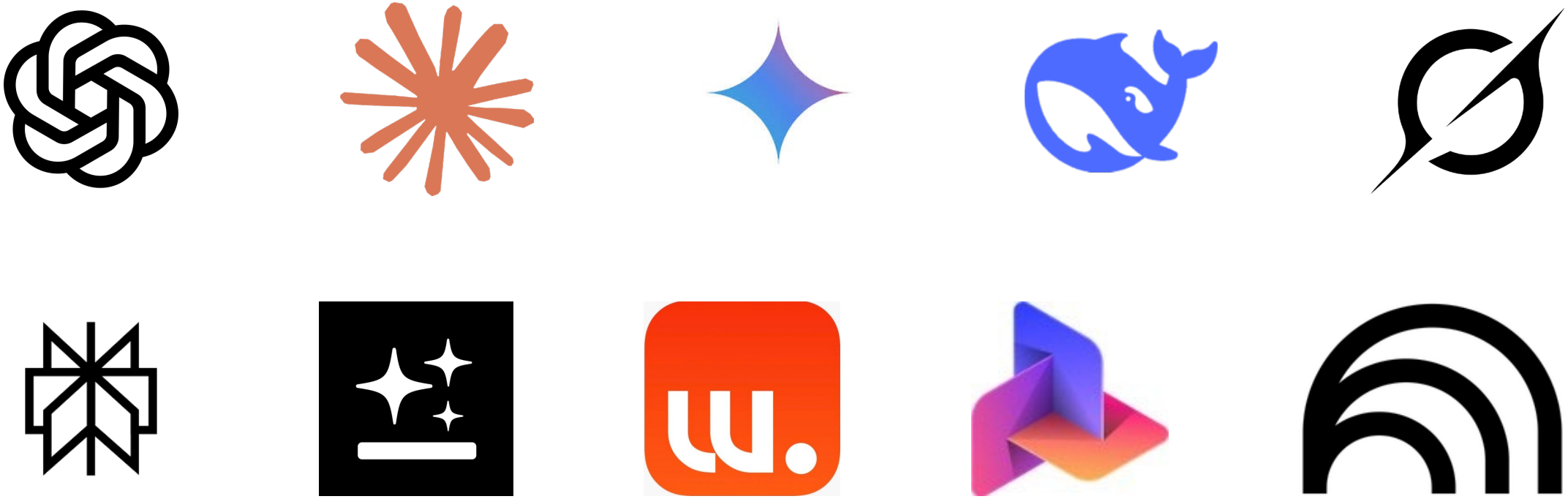

도쿄대학교
기초 통계학
강의를
한 권에
담았다!
Percentage
Population & Sample
Distribution
Chart
Graph
30분
통계학
구라타 히로시 지음 · 김소영 옮김


모두가 궁금해하는 데이터의 모든 것
데이터홀릭!
박박사
김팀장
앨리스


업무의 잔머리
소소한 사무 DX
업무의 잔머리


매주 토요일 오후 9시 엑셀 LIVE 방송
강의 대표 채널
No.1 엑셀
오빠두 엑셀
강사.
전 진 권
2,400만 직장인을 위한
진짜 엑셀 강의

그리고… Study & work with AI



모든 것을 알 필요는 없지만, 전문용어 & 기본 개념을 알고

알고 싶은 것 /시키고 싶은 것을 정확히 지시할 수는 있어야 한다!