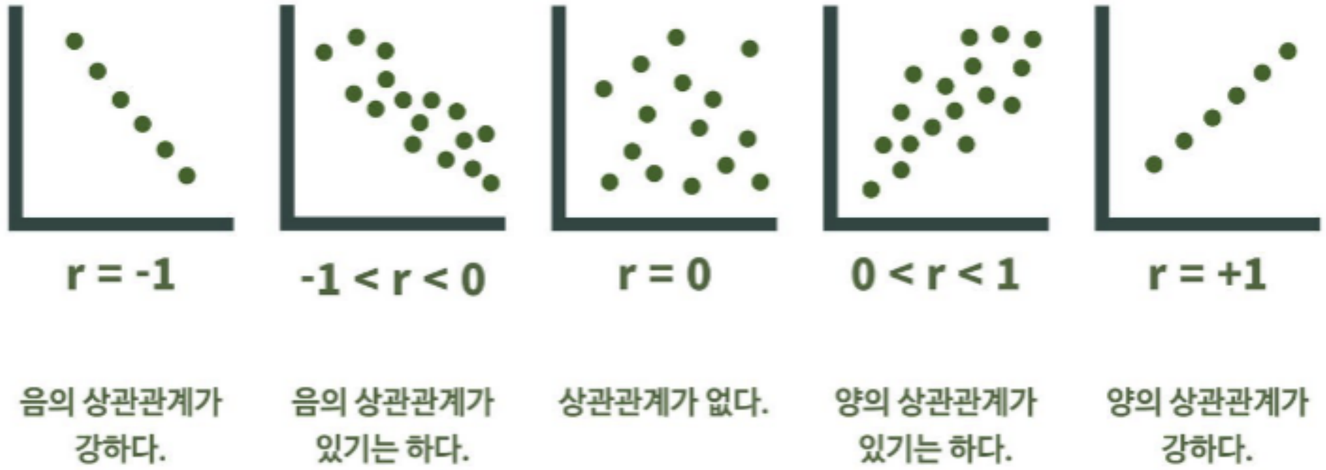
Correlation

상관계수(Correlation)

두 변수 사이에 관계를 나타내는 수치

- [-1, +1] 사이의 값을 가짐
- -1에 가까울수록 **음의 상관관계**
- +1에 가까울수록 **양의 상관관계**
- 0은 **관계가 없음**



Correlation

장점

- 간단하고 직관적인 방법
- 변수 간 관계 파악이 쉬움
- 계산이 단순

단점

- 비선형 관계 파악이 어려움
- 연관성을 볼 뿐, 인과관계를 설명하지 않음
- 결측치나 이상치에 민감하게 반응

Correlation

$$corr(A,B) = \frac{Cov(A,B)}{\sqrt{Var(A) \cdot Var(B)}}$$

$$Cov(A,B) = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})$$

$$Var(A) = \frac{1}{n-1}\sum_{i=1}^{n}(A_i - \bar{A})^2$$

표준화 ↓

$$corr(A,B) = \frac{Cov(A,B)}{\sqrt{Var(A)} \cdot \sqrt{Var(B)}} = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})^2} \cdot \sqrt{\sum_{i=1}^{n}(B_i - \bar{B})^2}}$$

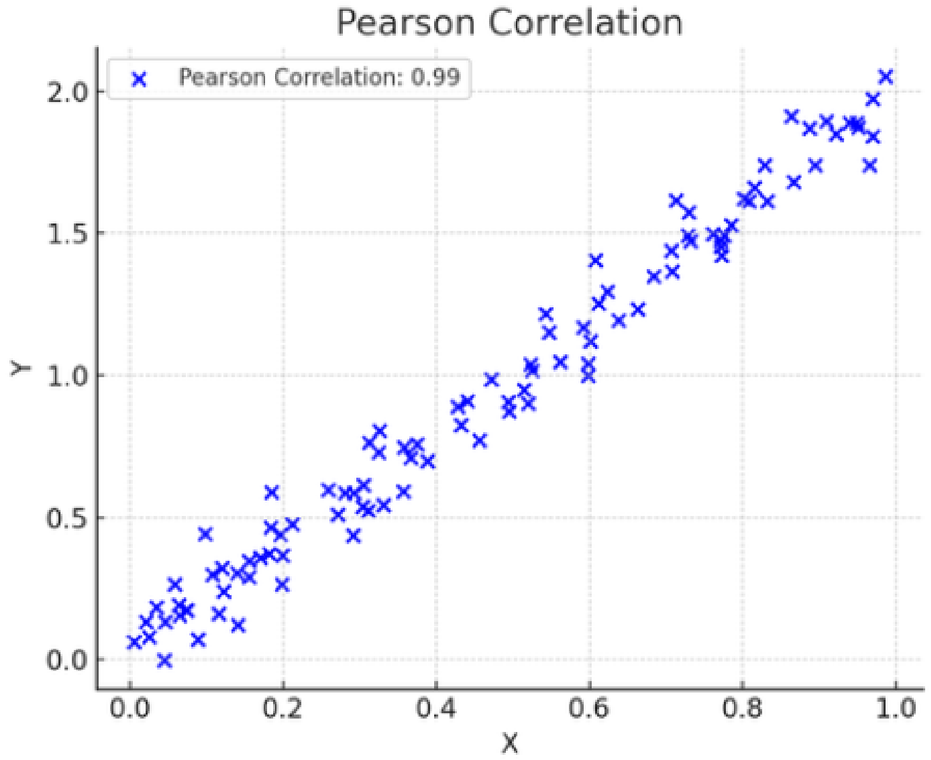
Correlation

이걸로 뭐하는거임?

차원의 저주를 해결 → **차원을 줄여야 함.**

비슷한 feature가 있으면 **두 feature 중 선택**

하나는 Drop



Correlation

x1	X2
3	6
4	9
8	21
2	6
7	8
2	3
4	5

$\overline{x_1} = 4.28$ $\overline{x_2} = 8.28$

≈ 4 ≈ 8

Correlation

x1	X2
3	6
4	9
8	21
2	6
7	8
2	3
4	5

$\overline{x_1} = 4.28 \approx 4$ $\overline{x_2} = 8.28 \approx 8$

X'1	X'2	X'1*x'2	∑
-1	-2	2	
0	1	0	
4	13	52	
-2	-2	4	68
3	0	0	
-2	-5	10	
0	-3	0	

Correlation

x1	X2
3	6
4	9
8	21
2	6
7	8
2	3
4	5

$\overline{x_1} = 4.28 \approx 4$ $\overline{x_2} = 8.28 \approx 8$

X'1	X'2	X'1*x'2	$\sum$
-1	-2	2	68
0	1	0	
4	13	52	
-2	-2	4	
3	0	0	
-2	-5	10	
0	-3	0	

A=(X'1)^2	B=(X'2)^2	$\sum A * \sum B$	$\sqrt{}$A*B
1	4	34*212= 7208	84.89
0	1		
16	169		
4	4		
9	0		
4	25		
0	9		

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{68}{84.89} = 0.80$$

Bostion dataset (보스턴 주택 가격 데이터)

```
1  import statsmodels.api as sm
2
3  boston = sm.datasets.get_rdataset("Boston", "MASS").data
4  print(boston.head())
```

✓ 1.6s

Bostion dataset (보스턴 주택 가격 데이터)

[01]	CRIM	자치시(town) 별 1인당 범죄율
[02]	ZN	25,000 평방피트를 초과하는 거주지역의 비율
[03]	INDUS	비소매상업지역이 점유하고 있는 토지의 비율
[04]	CHAS	찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0)
[05]	NOX	10ppm 당 농축 일산화질소
[06]	RM	주택 1가구당 평균 방의 개수
[07]	AGE	1940년 이전에 건축된 소유주택의 비율
[08]	DIS	5개의 보스턴 직업센터까지의 접근성 지수
[09]	RAD	방사형 도로까지의 접근성 지수
[10]	TAX	10,000 달러 당 재산세율
[11]	PTRATIO	자치시(town)별 학생/교사 비율
[12]	B	1000(Bk-0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함.
[13]	LSTAT	모집단의 하위계층의 비율(%)
[14]	MEDV	본인 소유의 주택가격(중앙값) (단위: $1,000)

Bostion dataset (보스턴 주택 가격 데이터)

```
import matplotlib.pyplot as plt
import seaborn as sns

cor = boston.corr()

plt.figure(figsize=(10, 10))
sns.heatmap(cor, annot = True, cmap=plt.cm.Blues)
plt.show()
```

Correlation