강원대학교
AI 소프트웨어학과

머신러닝1

– 기초통계 –

추론통계(분산의 분포– **상관관계**)

KNU 강원대학교

추론통계

상관관계 분석

- 상관관계 테스트(Correlation Test) : 상관 테스트는 두 연속 변수 사이의 연관성이 있는지 확인하는 데 사용됨
- 상관 분석: 상관 분석은 두 연속 변수 간의 선형 관계의 강도와 방향을 측정하는 데 사용됨
- Pearson 상관관계는 선형 관계에 적합하고
- Pearson의 상관 계수는 -1과 1 사이의 값을 가지며, 여기서 -1은 완벽한 음의 선형 관계를 나타내고, 1은 완벽한 양의 선형 관계를 나타내고, 0은 선형 관계가 없음을 나타냄
- Spearman 순위 상관관계는 비선형 관계에 사용됨

상관관계 분석(Pearson)

- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 두 변수는 독립이라고 할 순 없음
- H_0 : 상관계수가 0이다.
- H_1 : 상관계수가 0이 아니다.

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

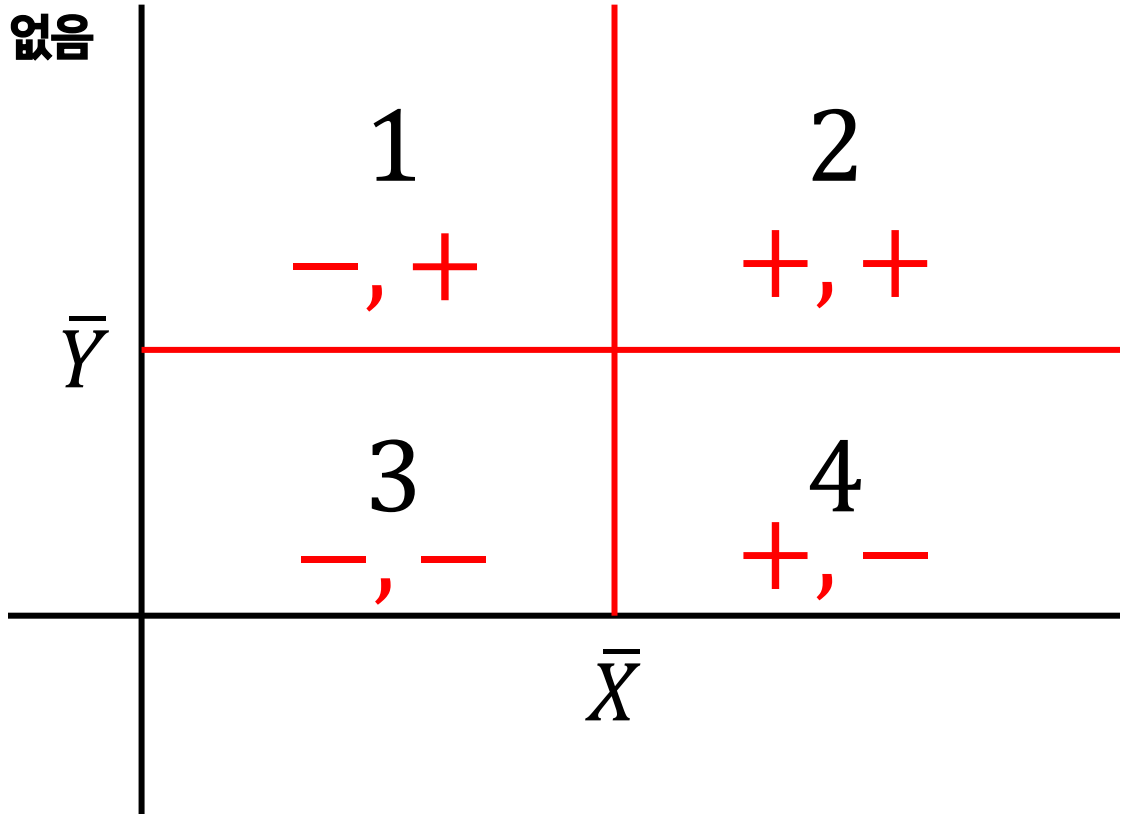
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

상관관계 분석(Pearson)

- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 두 변수는 독립이라고 할 순 없음

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$



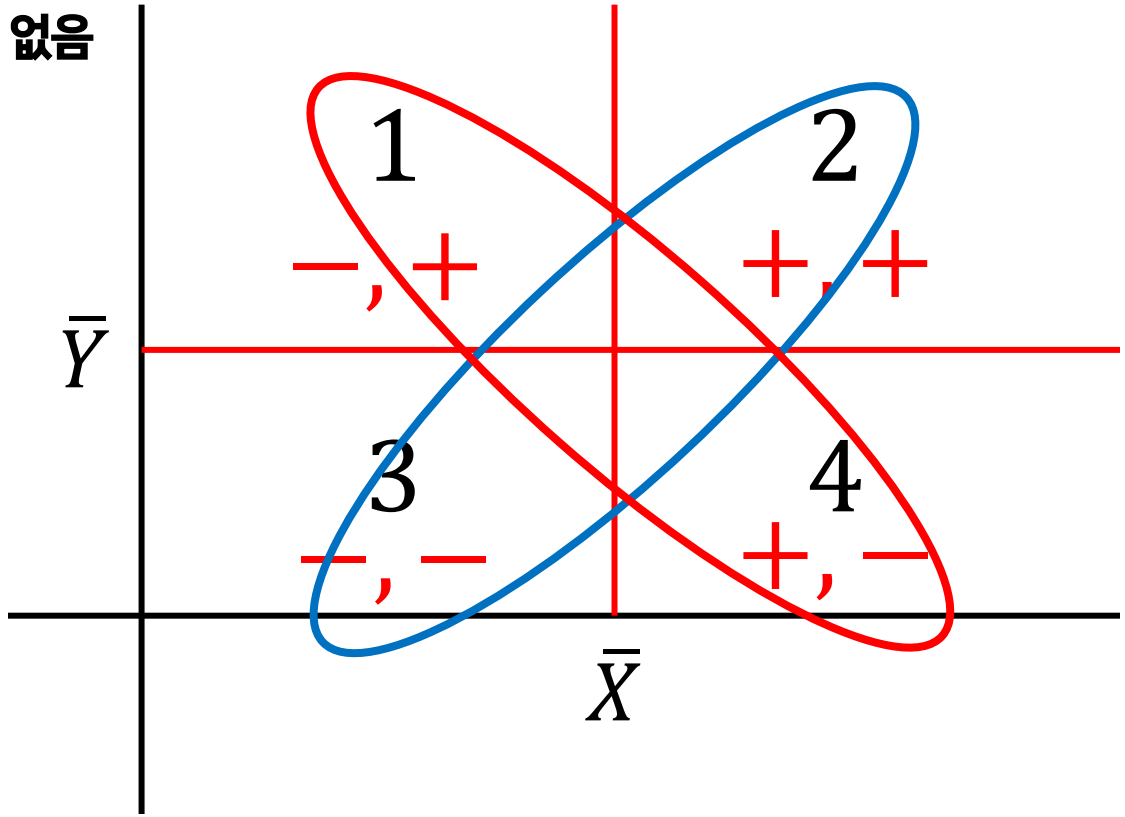
상관관계 분석(Pearson)

- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 두 변수는 독립이라고 할 순 없음

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$Cov(x,y) > 0(2,3구간)$
$Cov(x,y) < 0(1,2구간)$



상관관계 분석(Pearson)

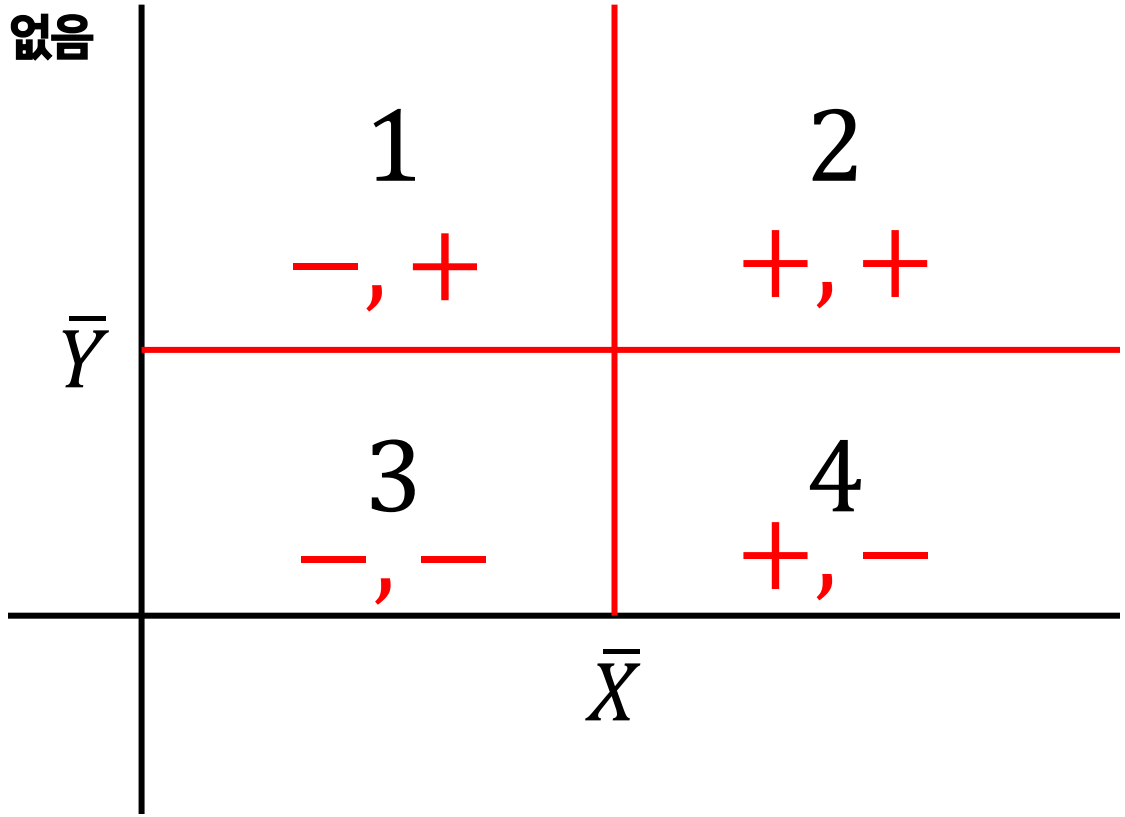
- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 두 변수는 독립이라고 할 순 없음

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

극단치에 영향을 많이 받을 수 있음

$$\sqrt{Var(X)Var(Y)}$$



상관관계 분석(Pearson)

- 상관계수 : Cov(x,y)에서 표준화 된 개념의 등장 -1과 1에서 분포의 모양을 판단함

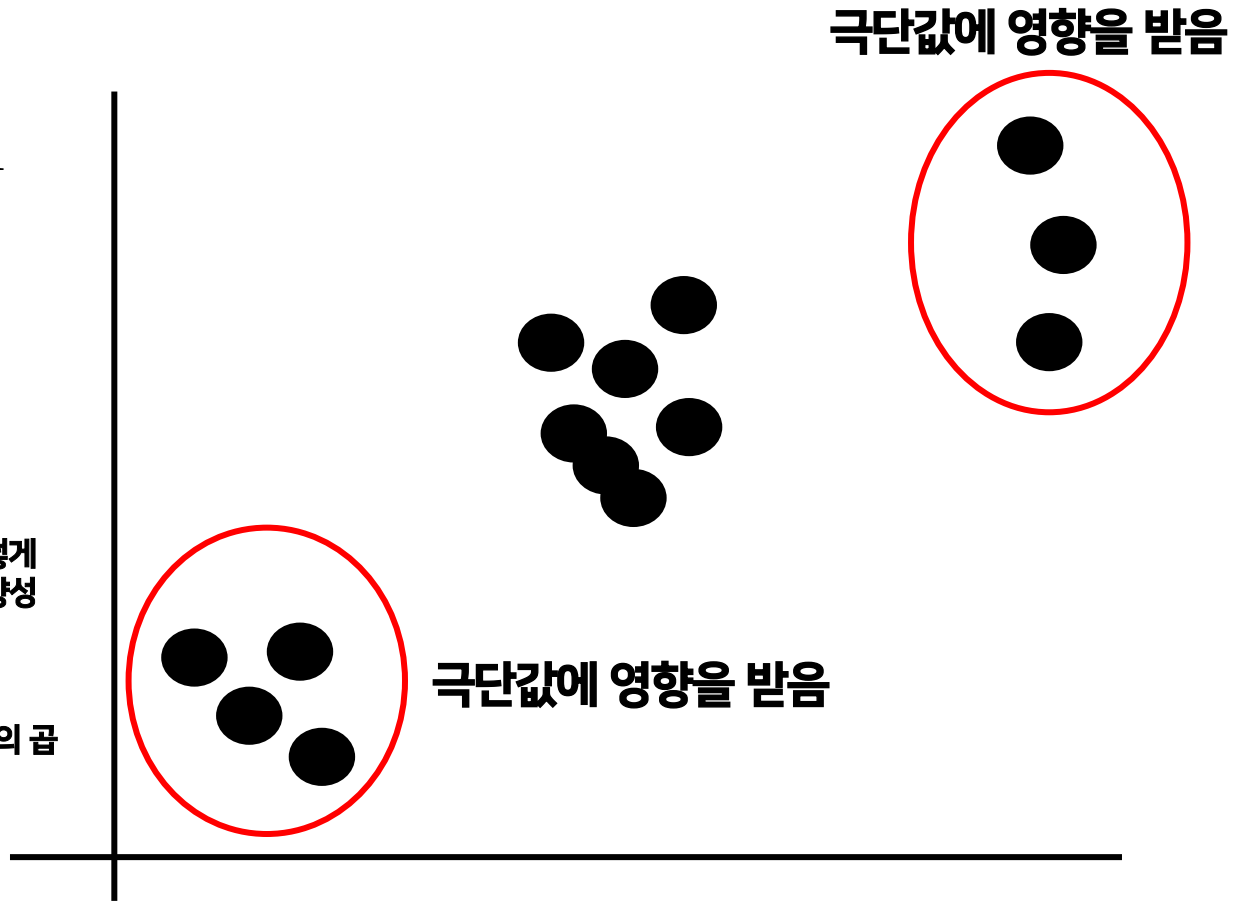
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \qquad -1 \le \rho(X,Y) \le 1$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{\sigma_X}{\sqrt{n-1}} \times \frac{\sigma_Y}{\sqrt{n-1}}}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

→ 두 변수 X와 Y가 함께 어떻게 변하는지를 측정함 → 방향성 (공변량)

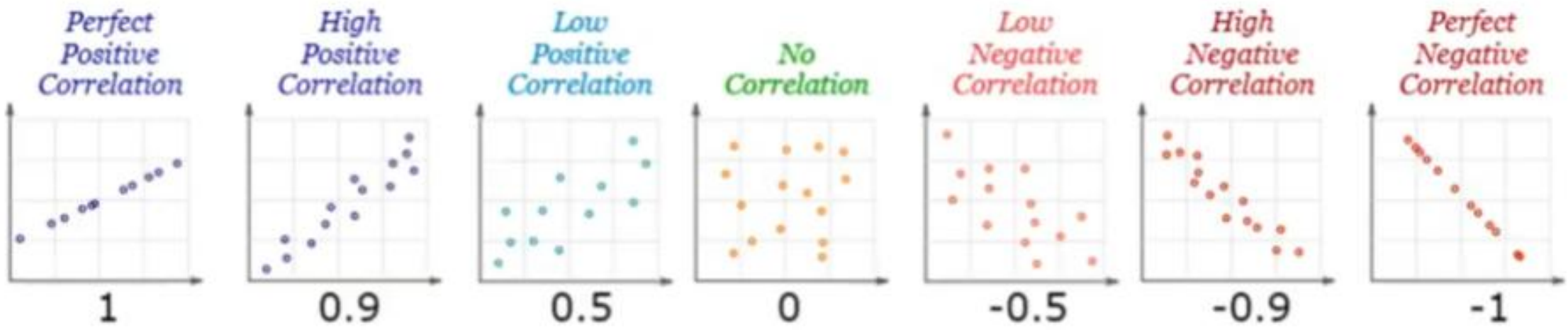
→ X와 Y 각각의 표준편차의 곱 → 퍼짐 정도 (변동성)



추론통계

상관관계 분석(Pearson)

- 상관계수 : Cov(x,y)에서 표준화 된 개념의 등장 -1과 1에서 분포의 모양을 판단함
- 귀무가설(H_0): 변수 간에 상관관계가 H_0 : ρ = 0
- 대립가설(H_1): 양측(H_1 : ρ ≠ 0) 또는 단측(H_1 : ρ > 0 또는 H_1 : ρ < 0)



추론통계

(1) 표본상관계수의 범위는 $-1 \le r \le 1$이다.

(2) $0 < r \le 1$이면 양의 직선적 상관관계를 갖는다.

(3) $-1 \le r < 0$이면 음의 직선적 상관관계를 갖는다.

(4) $r = 0$이면 직선적 상관관계를 갖지 않는다.

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

공분산을 정규화(normalize)하기 위해 두 변수의 표준편차로 나누면 상관계수는 -1에서 1까지의 범위를 가지게 됨

상관관계 분석(Pearson)

```
# 키와 몸무게 데이터 생성
heights <- c(160, 162, 155, 180, 170, 175, 165, 171, 177, 172)
weights <- c(55, 60, 53, 72, 70, 73, 62, 64, 69, 65)

library(psych)
# 피어슨상관계수 도출 및 p-value
result_pearson=corr.test(heights, weights, method="pearson")

result_pearson$p #p-value
result_pearson$r #상관관계 계수
```

상관관계 분석(Spearman)

- 두 변수 간의 단조(monotonic) 관계에 사용함 → 즉, 한 변수가 증가하면 다른 변수도 일정한 방향으로 증가 or 감소
- 범주형 변수에서 순위형 변수에 해당되는 값의 비선형 상관관계를 파악할 때 사용함
- 상관 계수는 -1과 1 사이의 값을 가지며, 여기서 -1은 완벽한 음의 단조 관계를 나타내고, 1은 완벽한 양의 단조 관계를 나타내고, 0은 단조 관계가 없음을 나타냄
- 순위형 vs 순위형, 순위형 vs 수치형 변수 가능
  → 고객 만족도(1~5점) ↔ 서비스 재이용 횟수(수치형)
  → 공부 강도(1~5점) ↔ 학교 시험 등수(순위형)
- 이진 범주형 변수 vs 수치형 변수 가능(불가능)
  → 성별 ↔ 시험점수
- 일반 명목형 변수 대해서는 불가능(불가능)
  → 혈액형 ↔ 키, 지역 ↔ 수입 등

추론통계

상관관계 분석(Spearman)

- 순위로 변환하여 계산
- 선형이 아닌 단조관계 포착
- 이상치에 강함
- 시험 성적 순위 vs 자기 만족도 순위
- 건강상태 평가척도(1~5점) vs 우울감 순위

순위 1	순위 2	d=X-Y
1	4	1-4=-3
2	3	2-3=-1
3	1	3.5-1=2.5
3	2	3.5-2=1.5

$$\rho_s = corr(rank(X), rank(Y))$$

순위가 동률일 경우 동률의 순위의 평균 값으로 대체

$$d_i = R(x_i) - R(y_i)$$

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

순위가 동률일 경우 순위의 평균으로 동률 순위를 계산

n = 데이터 수(비교 쌍의 수)

상관관계 분석(Spearman)

```
# 학생의 자기평가 (순위형: 1=매우 낮음 ~ 5=매우 높음)
s_eval <- c(3, 4, 2, 5, 1, 4, 3, 2, 5, 3)

# 평가자의 평가 (순위형: 1=매우 낮음 ~ 5=매우 높음)
e_eval <- c(2, 5, 2, 4, 1, 4, 3, 2, 5, 3)

# 스피어만 상관계수 도출 및 p-value
result_pearson=corr.test(s_eval, e_eval, method="spearman")

result_pearson$p #p-value
result_pearson$r #상관관계 계수
```

상관관계 분석(Spearman)

```
data= read.csv(“pearson.csv”)
data= read.csv(“spearman.csv”)

indep_vars <- data[, c(" 독립변수1", " 독립변수2", " 독립변수3", “독립변수4”)]

# 종속변수 (연속형)
target_var <- data_pearson[, " 종속변수", drop = FALSE] #drop = FALSE(데이터프레임 유지)
```

추론통계

상관관계 분석

독립변수의 유형	종속변수의 유형	분석방법	분석 목적
범주형	범주형	카이제곱검정(독립성)	두 범주형 변수 간의 독립성 검정
범주형	연속형	ANOVA분석 or t-test	범주에 따른 평균차이 검정
연속형	연속형	피어슨 상관분석	선형관계 측정
순위형	순위형	스피어만 상관분석	단조관계 측정

추론통계

상관관계 분석

- 데이터를 기계 학습 알고리즘에 제공하기 전에 데이터를 이해하고 전처리하는 것이 중요함
- T-test, Z-test, 카이제곱, 상관 분석 및 ANOVA와 같은 통계 테스트는 데이터 탐색 및 준비를 위한 유용한 도구 역할을 하여 데이터의 전반적인 이해를 알리는 데 도움됨
- 기술통계는 데이터세트의 특성을 요약하고 정리하는 것
- 추론통계는 데이터의 기본 구조를 이해하고, 데이터들 간의 관계를 통해 모집단에 대해 추론
- 추론통계는 데이터에서 통찰력을 얻고, 예측을 수행하기 위한 강력한 접근 방식