기초 통계와 데이터 시각화

sprint by codeit:

학습 목차

01 통계의 종류

통계는 크게 기술 통계와 추론 통계가 있어요

기술 통계
(Descriptive Statistics)

추론 통계
(Inferential Statistics)

통계는 크게 기술 통계와 추론 통계가 있어요

기술 통계
(Descriptive Statistics)

숫자를 이용하여

자료의 정보를 요약 기술하는 것

추론 통계
(Inferential Statistics)

통계는 크게 기술 통계와 추론 통계가 있어요

기술 통계
(Descriptive Statistics)

숫자를 이용하여

자료의 정보를 요약 기술하는 것

추론 통계
(Inferential Statistics)

실험이나 관찰을 통하여 얻은 자료를 분석하여

표본을 바탕으로 모집단에 대한 정보를 유추하는 것

통계는 크게 기술 통계와 추론 통계가 있어요

기술 통계
(Descriptive Statistics)

숫자를 이용하여

자료의 정보를 요약 기술하는 것

- 대표값
- 변동성

추론 통계
(Inferential Statistics)

실험이나 관찰을 통하여 얻은 자료를 분석하여

표본을 바탕으로 모집단에 대한 정보를 유추하는 것

1) 추정
2) 검정

통계는 크게 기술 통계와 추론 통계가 있어요

기술 통계
(Descriptive Statistics)

숫자를 이용하여

자료의 정보를 요약 기술하는 것

- 대표값
- 변동성

추론 통계
(Inferential Statistics)

실험이나 관찰을 통하여 얻은 자료를 분석하여

표본을 바탕으로 모집단에 대한 정보를 유추하는 것

1) 추정
2) 검정

02 대표값과 변동성

대표값

데이터 집합의 중심 경향을 나타내어 데이터의 특성을 전체 데이터를 대표하는 하나의 값으로 요약

대표값

데이터 집합의 중심 경향을 나타내어 데이터의 특성을 전체 데이터를 대표하는 하나의 값으로 요약

평균
(Mean)

중앙값
(Median)

최빈값
(Mode)

대표값

데이터 집합의 중심 경향을 나타내어 데이터의 특성을 전체 데이터를 대표하는 하나의 값으로 요약

72	85	68	73	94	76	88	70	63

평균
(Mean)

$$\frac{72+85+68+73+94+76+88+70+63}{9} = 76.56$$

대표값

데이터 집합의 중심 경향을 나타내어 데이터의 특성을 전체 데이터를 대표하는 하나의 값으로 요약

72 85 68 73 94 76 88 70 63

평균
(Mean)

$$\frac{72+85+68+73+94+76+88+70+63}{9} = 76.56$$

63 68 70 72 73 76 85 88 94

중앙값
(Median)

데이터 오름차순 정렬 시 중간에 있는 값

대표값

데이터 집합의 중심 경향을 나타내어 데이터의 특성을 전체 데이터를 대표하는 하나의 값으로 요약

72	85	68	73	94	76	88	70	63

평균 (Mean)

$$\frac{72+85+68+73+94+76+88+70+63}{9} = 76.56$$

63	68	70	72	73	76	85	88	94

중앙값 (Median)

데이터 오름차순 정렬 시 중간에 있는 값

C	B	D	C	A	C	B	C	D

최빈값 (Mode)

가장 빈번하게 등장하는 값

A : 1개, B : 2개, C : 4개, D : 2개

변동성

데이터의 변동성은 데이터들이 대표값을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

변동성

데이터의 변동성은 데이터들이 대표값을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

분산
(Variance)

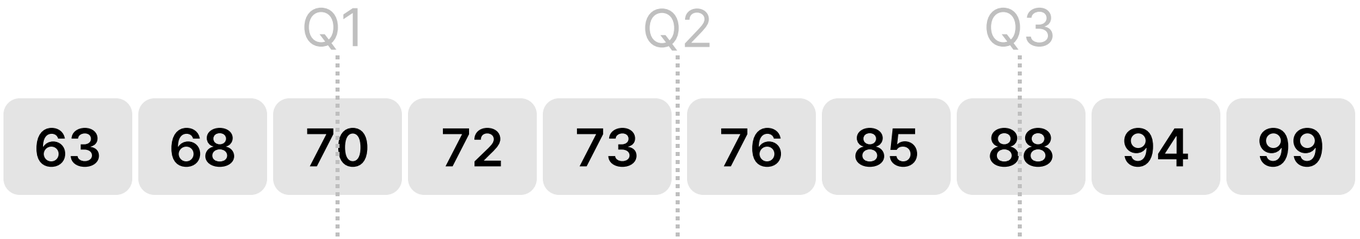
표준편차
(Standard Deviation)

범위
(Range)

사분위범위
(IQR, Interquartile Range)

변동성

데이터의 변동성은 데이터들이 대표값을 중심으로 얼마나 퍼져 있는지를 나타내는 지표



평균 : 78.8

분산
(Variance)

편차 제곱의 평균

$$= \frac{(63 - 78.8)^2 + \ \ldots \ + (99 \ - 78.8)^2}{10} = 129.36$$

* 표본분산 : 편차 제곱 합을 n-1로 나누어 계산

변동성

데이터의 변동성은 데이터들이 대표값을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

Q1 Q2 Q3

63	68	70	72	73	76	85	88	94	99

평균 : 78.8

분산 (Variance) 편차 제곱의 평균

$$= \frac{(63 - 78.8)^2 + \ \ldots \ + (99 \ - 78.8)^2}{10} = 129.36$$

표준편차 (Standard Deviation) 분산의 제곱근

$$= \sqrt{129.36} = 11.37$$

* 표본분산 : 편차 제곱 합을 n-1로 나누어 계산

* 표본표준편차 : 표본분산의 제곱근

변동성

데이터의 변동성은 데이터들이 대표값을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

Q1 Q2 Q3

63	68	70	72	73	76	85	88	94	99

평균 : 78.8

분산 (Variance) 편차 제곱의 평균

$$= \frac{(63 - 78.8)^2 + \ \ldots \ + (99 \ - 78.8)^2}{10} = 129.36$$

표준편차 (Standard Deviation) 분산의 제곱근

$$= \sqrt{129.36} = 11.37$$

범위 (Range) 최댓값 - 최솟값

$$= 99 - 63 = 36$$

* 표본분산 : 편차 제곱 합을 n-1로 나누어 계산

* 표본표준편차 : 표본분산의 제곱근

변동성

데이터의 변동성은 데이터들이 대표값을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

Q1 Q2 Q3

63	68	70	72	73	76	85	88	94	99

평균 : 78.8

분산 (Variance) — 편차 제곱의 평균

$$= \frac{(63 - 78.8)^2 + \ \ldots \ + (99 \ - 78.8)^2}{10} = 129.36$$

표준편차 (Standard Deviation) — 분산의 제곱근

$$= \sqrt{129.36} = 11.37$$

범위 (Range) — 최댓값 - 최솟값

$$= 99 - 63 = 36$$

사분위범위 (IQR, Interquartile Range) — Q3 – Q1

$$= 88 - 70 = 18$$

* 표본분산 : 편차 제곱 합을 n-1로 나누어 계산
* 표본표준편차 : 표본분산의 제곱근

03 이상치 탐색

이상치

데이터의 집합에서 대부분의 데이터에 비해 현저히 차이나는 값

이상치

데이터의 집합에서 대부분의 데이터에 비해 현저히 차이나는 값

IQR 기반 이상치 탐지
(Interquartile Range(IQR)-Based Outlier Detection)

Z-점수 기반 이상치 탐지
(Z-Score-Based Outlier Detection)

이상치

데이터의 집합에서 대부분의 데이터에 비해 현저히 차이나는 값

IQR 기반 이상치 탐지
(Interquartile Range(IQR)-Based Outlier Detection)

IQR(사분위 범위)를 활용하여 데이터의
중앙 집중 부분에서 벗어난 값을 이상치로 탐지

- IQR = Q3 – Q1
- Lower Limit(하한) = Q1 – 1.5 X IQR
- Upper Limit(상한) = Q3 + 1.5 X IQR
- 데이터가 하한과 상한의 범위를 벗어나는 경우 이상치로 간주

이상치

데이터의 집합에서 대부분의 데이터에 비해 현저히 차이나는 값

IQR 기반 이상치 탐지
(Interquartile Range(IQR)-Based Outlier Detection)

IQR(사분위 범위)를 활용하여 데이터의 중앙 집중 부분에서 벗어난 값을 이상치로 탐지

- IQR = Q3 – Q1
- Lower Limit(하한) = Q1 – 1.5 X IQR
- Upper Limit(상한) = Q3 + 1.5 X IQR
- 데이터가 하한과 상한의 범위를 벗어나는 경우 이상치로 간주

Z-점수 기반 이상치 탐지
(Z-Score-Based Outlier Detection)

데이터가 평균에서 몇 표준편차만큼 떨어져 있는지를 측정하는 Z-점수로 이상치를 탐지

- $Z = \frac{(X - \text{평균})}{\text{표준편차}}$
- Z가 ± 3 범위를 벗어나는 경우 이상치로 간주

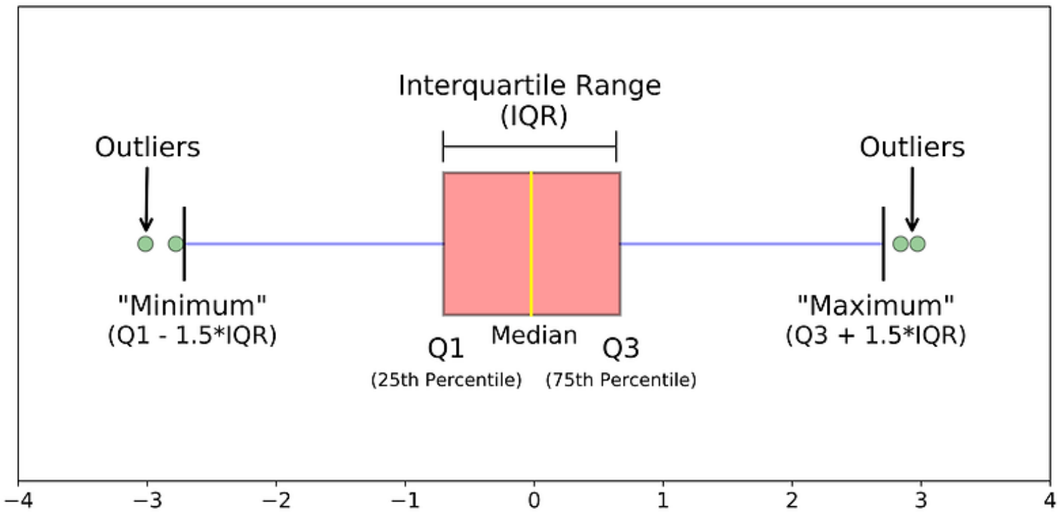
이상치

데이터의 집합에서 대부분의 데이터에 비해 현저히 차이나는 값

IQR 기반 이상치 탐지
(Interquartile Range(IQR)-Based Outlier Detection)

IQR(사분위 범위)를 활용하여 데이터의
중앙 집중 부분에서 벗어난 값을 이상치로 탐지

- IQR = Q3 – Q1
- Lower Limit(하한) = Q1 – 1.5 X IQR
- Upper Limit(상한) = Q3 + 1.5 X IQR
- 데이터가 하한과 상한의 범위를 벗어나는 경우 이상치로 간주



04 분포와 형태 이해

데이터의 분포와 비대칭성 이해

데이터가 평균을 중심으로 얼마나 치우쳐 있고 퍼져 있는지를 파악하여 데이터의 분포와 특성을 이해

왜도
(Skewness)

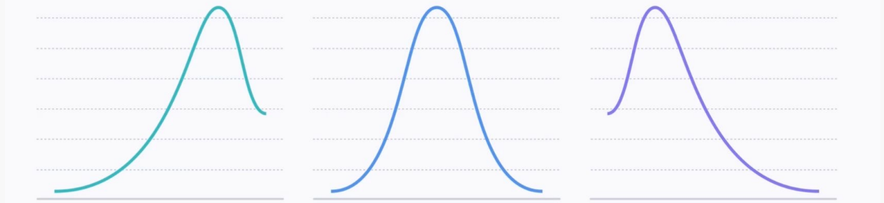
첨도
(Kurtosis)

데이터의 분포와 비대칭성 이해

데이터가 평균을 중심으로 얼마나 치우쳐 있고 퍼져 있는지를 파악하여 데이터의 분포와 특성을 이해

왜도
(Skewness)

데이터의 분포가 평균을 중심으로
얼마나 비대칭적인지 나타내는 측도



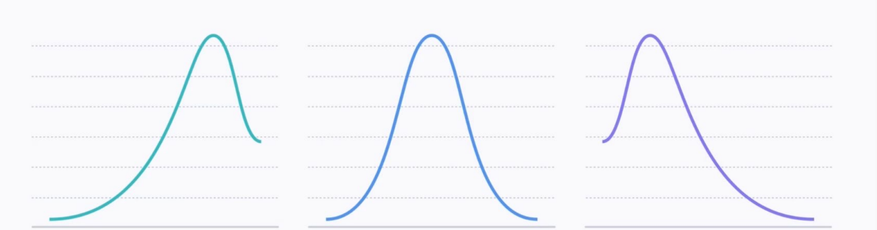
Left-skewed(왜도 < 0)	정규분포(왜도=0)	Right-skewed(왜도 > 0)
평균 < 중앙값 < 최빈값	평균 = 중앙값 = 최빈값	최빈값 < 중앙값 < 평균

데이터의 분포와 비대칭성 이해

데이터가 평균을 중심으로 얼마나 치우쳐 있고 퍼져 있는지를 파악하여 데이터의 분포와 특성을 이해

왜도
(Skewness)

데이터의 분포가 평균을 중심으로
얼마나 비대칭적인지 나타내는 측도



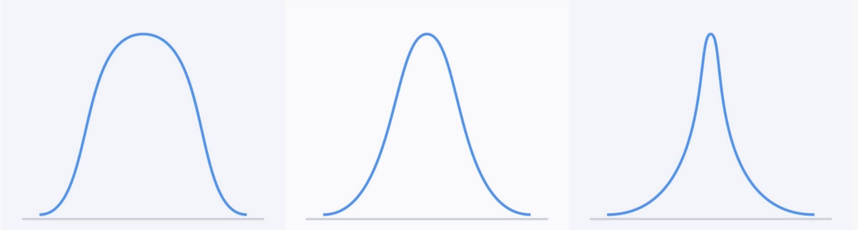
Left-skewed(왜도 < 0)
평균 < 중앙값 < 최빈값

정규분포(왜도=0)
평균 = 중앙값 = 최빈값

Right-skewed(왜도 > 0)
최빈값 < 중앙값 < 평균

첨도
(Kurtosis)

데이터 분포의 뾰족함이나
평평함의 정도를 나타내는 측도



평평한 분포
(첨도 < 0)

정규분포
(첨도=0)

뾰족한 분포
(첨도 > 0)

첨도는 교재/도구에 따라 3 또는 0을 기준으로 한다.

05 변수 간 관계 분석

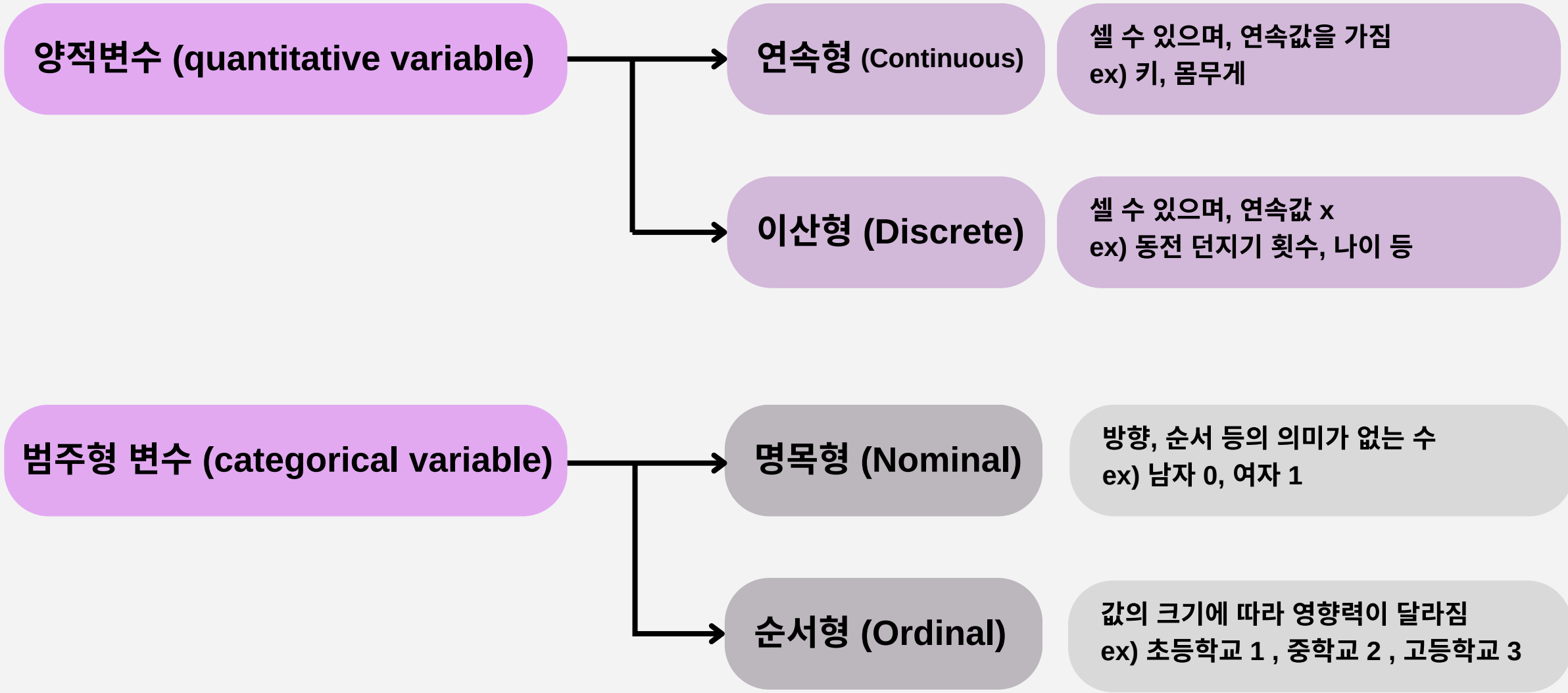
상관계수

두 변수 간의 선형적 관계의 방향과 강도를 -1에서 1 사이의 값으로 나타낸 지표

상관계수 절댓값	해석
0.50 ~ 1.00	강한 상관관계
0.30 ~ 0.49	중간 상관관계
0.10 ~ 0.29	약한 상관관계

Cohen(1988)

데이터 유형


양적변수 (quantitative variable)
연속형 (Continuous)
셀 수 있으며, 연속값을 가짐
ex) 키, 몸무게
이산형 (Discrete)
셀 수 있으며, 연속값 x
ex) 동전 던지기 횟수, 나이 등
범주형 변수 (categorical variable)
명목형 (Nominal)
방향, 순서 등의 의미가 없는 수
ex) 남자 0, 여자 1
순서형 (Ordinal)
값의 크기에 따라 영향력이 달라짐
ex) 초등학교 1 , 중학교 2 , 고등학교 3

상황별 데이터 시각화

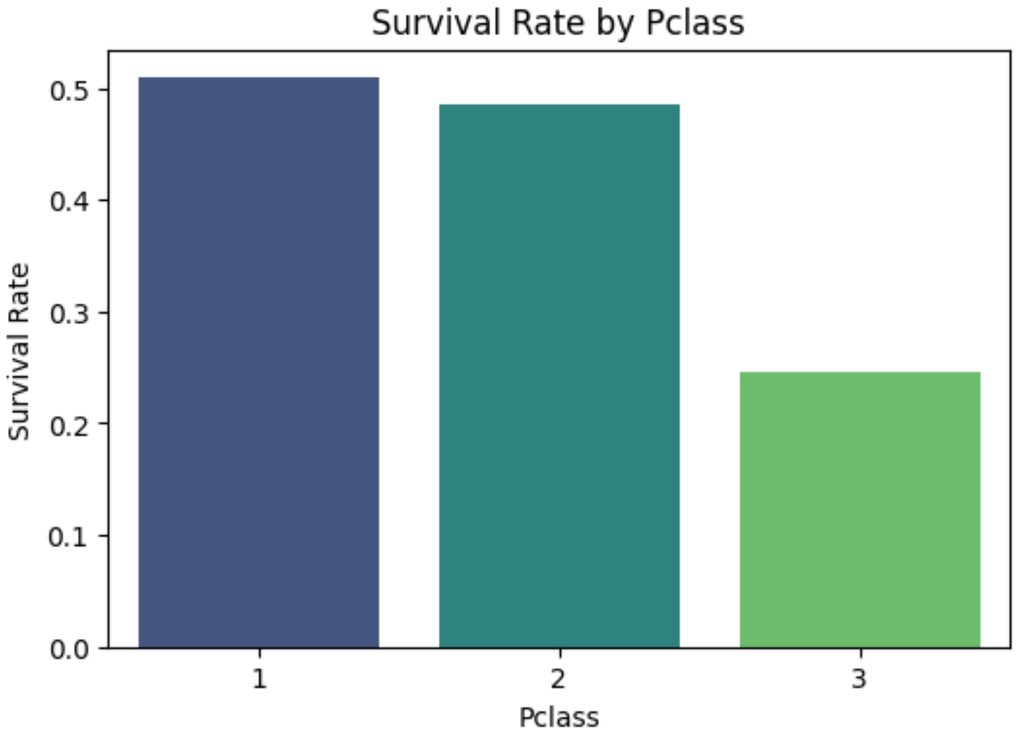
학습할 내용

1. 범주형 데이터 탐색
2. 연속형 데이터 탐색
3. 이상치 탐색
4. 변수 간 관계 탐색
5. 시간에 따른 변화 탐색

범주형 데이터 탐색

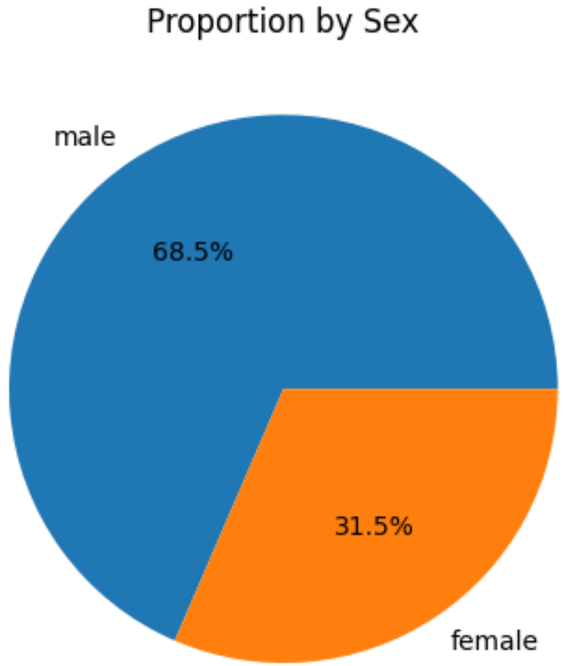
막대 그래프(Bar Plot)

🎯 목적 : 각 범주별 빈도 및 비율 비교



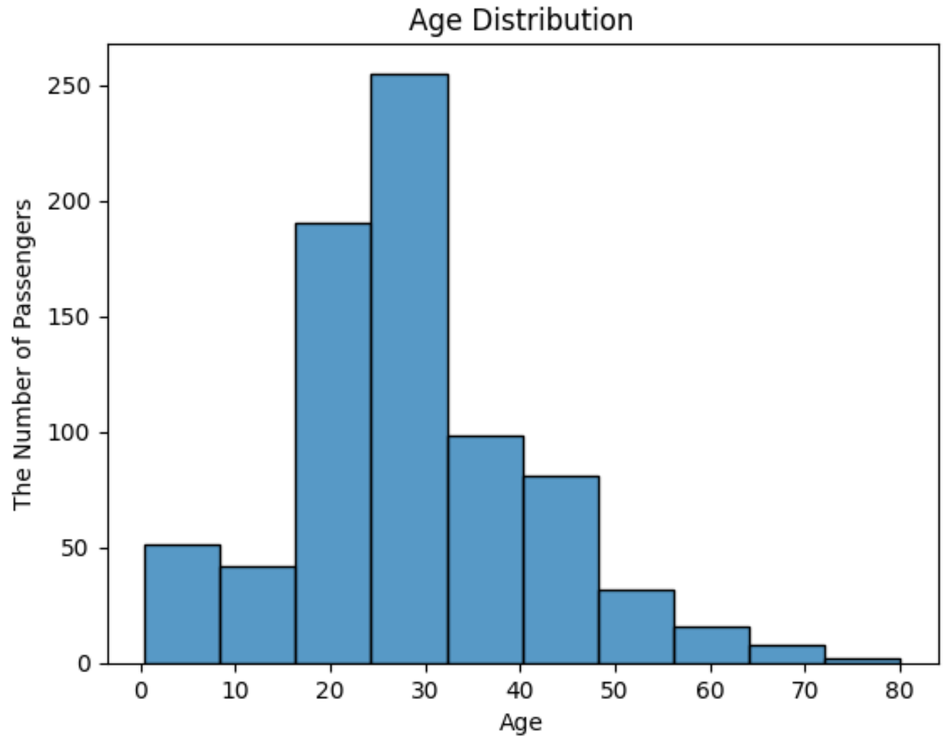
파이 차트(Pie Chart)

🎯 목적 : 전체 대비 각 범주의 비율



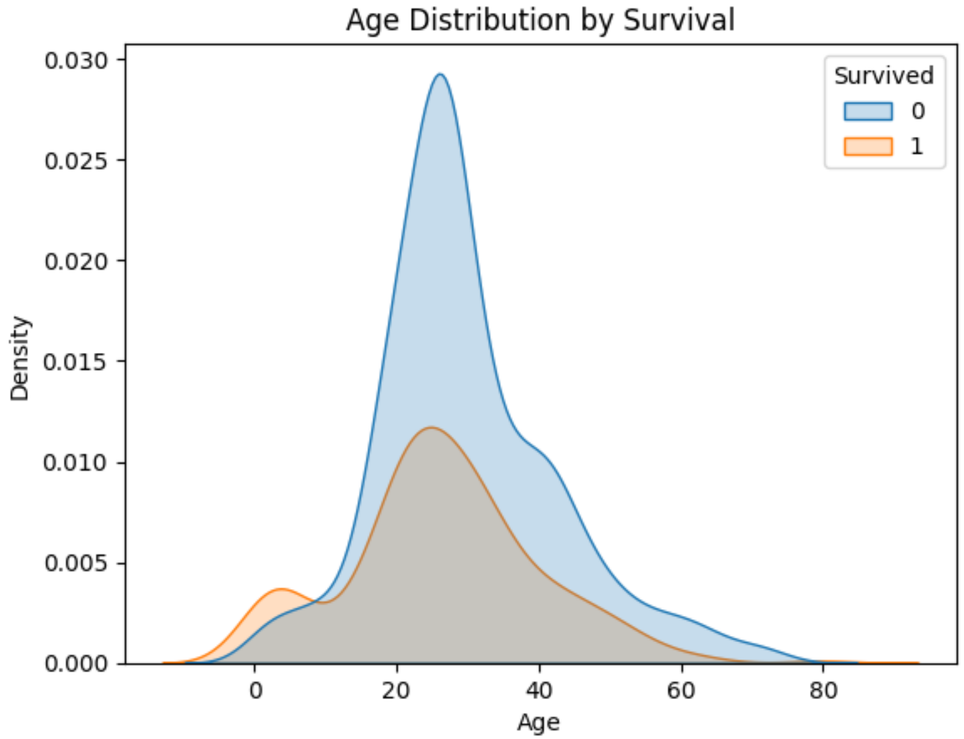
히스토그램(Histogram)

🎯 목적 : 데이터의 분포 형태 및 빈도 탐색



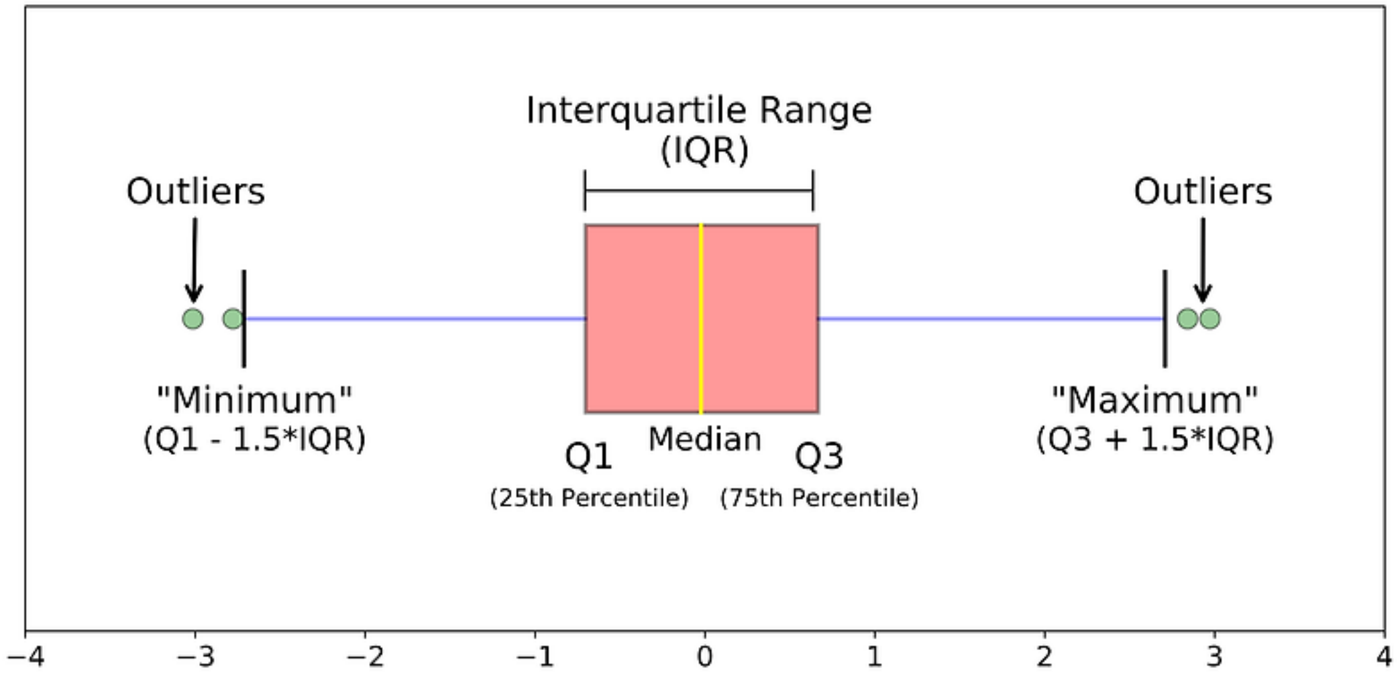
밀도 플롯(KDE Plot)

🎯 목적 : 분포를 부드러운 곡선 형태로 시각화



박스 플롯(Box Plot)

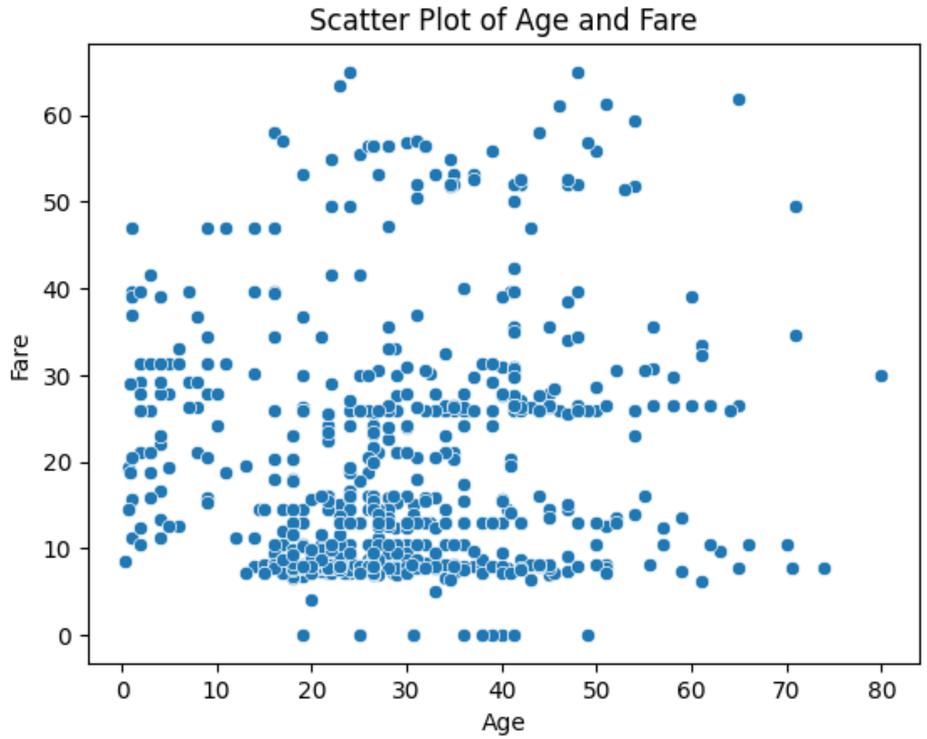
🎯 목적 : 이상치, 중앙값, 사분위수 탐색



변수 간 관계 탐색

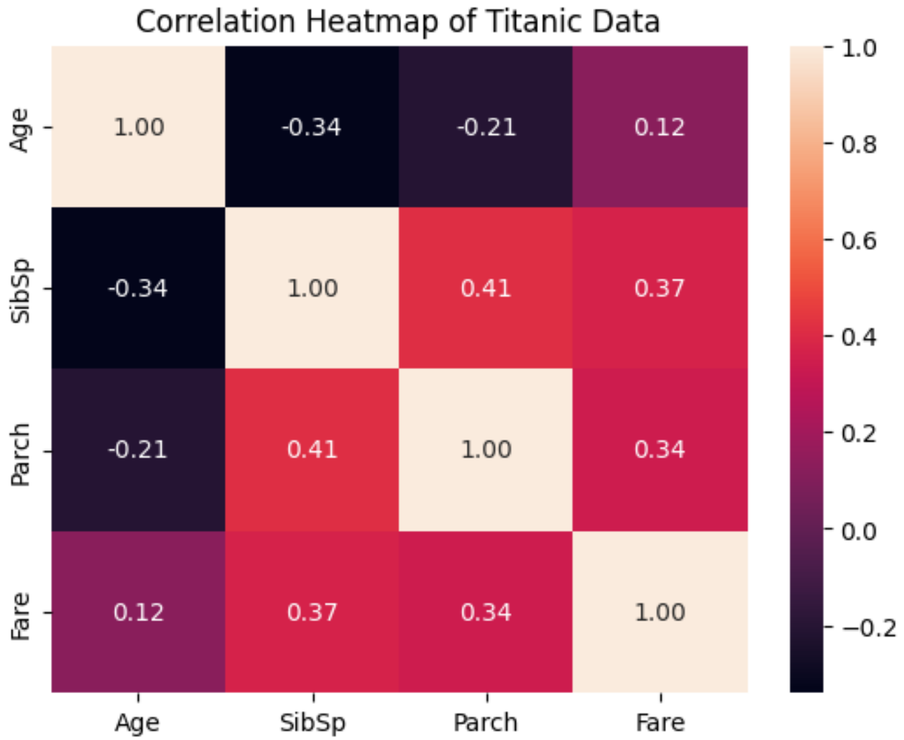
산점도(Scatter Plot)

🎯 목적 : 두 변수 간 상관관계 분석



히트맵(Heatmap)

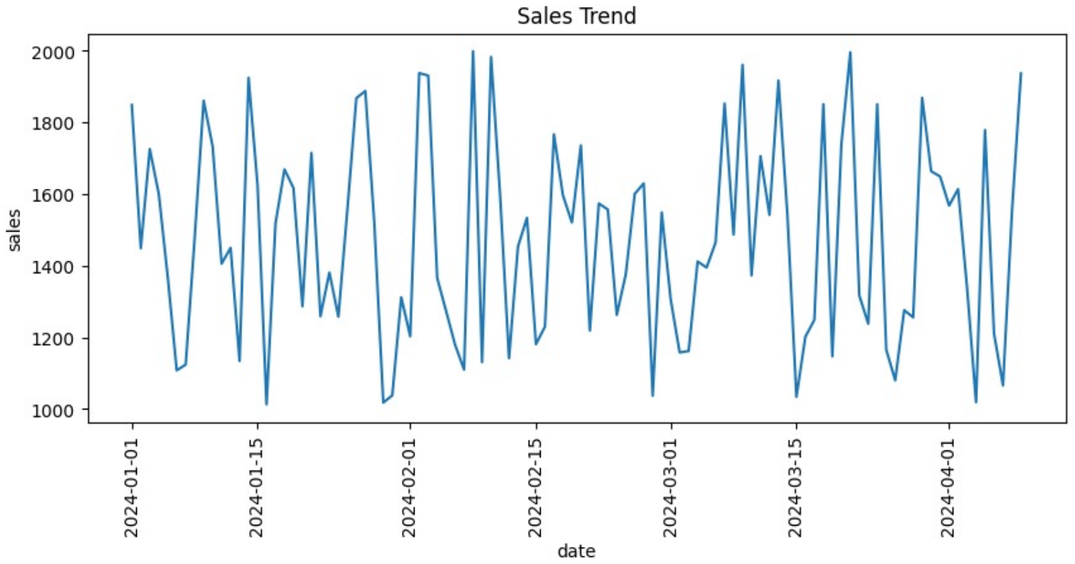
🎯 목적 : 다변량 데이터의 상관관계 파악



시간에 따른 변화 탐색

선 그래프(Line Plot)

🎯 목적 : 시간에 따른 변화 시각화



면적 그래프(Area Chart)

🎯 목적 : 누적 변화 및 각 요소의 변화량 분석