

4차시 데이터 요약하기



I. 기술통계

- 데이터 요약이란?

Data Summary: 자료의 내용과 특징을 압축적으로 요약하여 설명하는 작업

→ 한눈에 데이터의 특징을 파악할 수 있도록 만드는 것은 데이터 분석가의 주요한 과제가 된다!

주요 개념	특징
기술통계 (요약통계)	주어진 데이터의 특성에 대해 압축적으로 설명하는 작업 (기본적 통계량을 제공함)
탐색적 데이터 분석 (EDA, Exploratory DA)	주어진 데이터를 다양한 관점에서 관찰하고 이해하여 숨겨진 패턴을 발견하는 과정 (일반적으로 데이터 시각화를 포함)

- 데이터 시각화란?

Data Visualization: 복잡한 데이터나 통계 정보를 그래프, 차트, 지도, 인포 그래픽, 동영상 등 다양한 형태의 시각적 요소로 표현하는 작업

- 일반적인 기술통계량

유형	정의	구하는 방법
count	누락된 값을 제외하고 헤아린 데이터(샘플)의 개수	데이터 항목의 개수를 센다
mean	데이터의 평균값	데이터 합계 / 데이터 개수
std	평균으로부터 데이터가 흩어져 있는 정도	분산(편차 제곱의 평균)의 제곱근
min	주어진 데이터 중 가장 작은 값	가장 작은 데이터
mid	데이터를 정렬했을 때 한 가운데 값	데이터 정렬 후 가운데 값 (짝수 개일 경우 가운데 두 값의 평균)
max	주어진 데이터 중 가장 큰 값	가장 큰 데이터

Q. Pandas에서 describe() 메서드를 활용한 기술통계 구하기

describe() 사용하기

```
import pandas as pd
import numpy as np
# 샘플 데이터 생성
data = {
    "age": [23, 45, 12, 36, 28, 54, 33],
    "height": [165, 180, 150, 175, 170, 160, 172],
    "weight": [55, 80, 40, 70, 65, 60, 68]
}
df = pd.DataFrame(data)
# describe() 사용
print(df.describe())
```

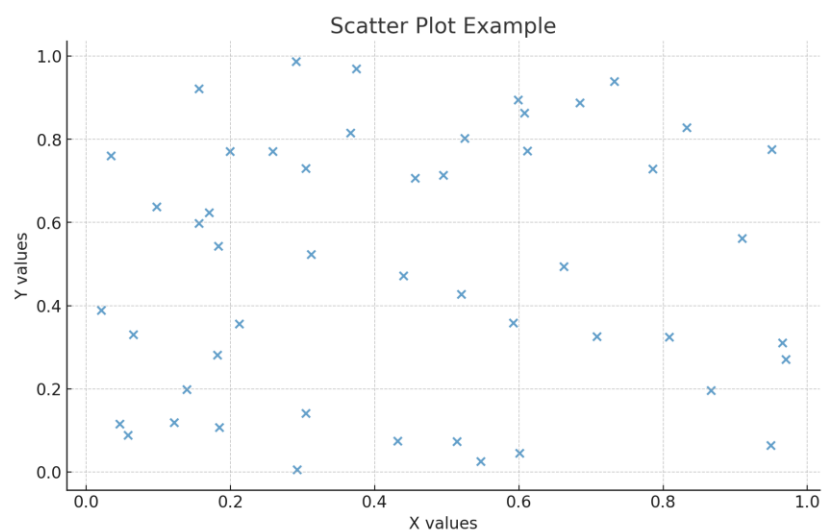
기술통계 출력

	age	height	weight
count	7.000000	7.000000	7.000000
mean	33.142857	167.428571	62.571429
std	13.484314	10.498413	12.658988
min	12.000000	150.000000	40.000000
25%	25.500000	162.500000	58.000000
50%	33.000000	170.000000	65.000000
75%	39.500000	172.750000	69.000000
max	54.000000	180.000000	80.000000

II. 분포 요약하기

- 산점도 그리기

산점도 (scatter plot): 두 변수 간의 관계를 점(좌표)으로 나타낸 그래프
데이터의 분포, 패턴, 상관관계를 직관적으로 확인할 수 있음



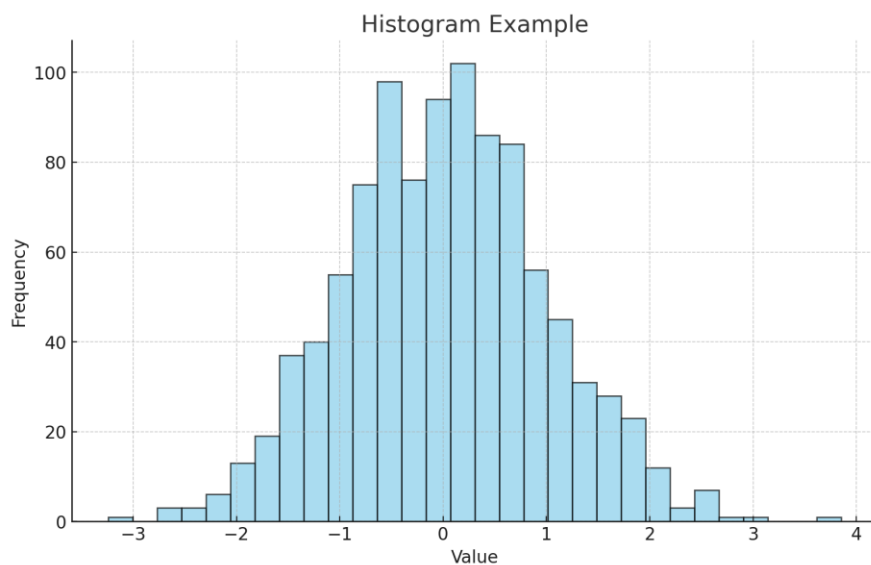
산점도 그리기

```
# 1. 산점도 (Scatter Plot)
import matplotlib.pyplot as plt
import numpy as np
# 샘플 데이터 생성
np.random.seed(42)
x = np.random.rand(50)
y = np.random.rand(50)
# 산점도 그리기
plt.scatter(x, y, alpha=0.7, edgecolors='black')
plt.title("Scatter Plot Example")
plt.xlabel("X values")
plt.ylabel("Y values")
# 결과 출력
plt.show()
```

- 히스토그램 그리기

히스토그램 (histogram): 데이터를 구간(빈, bin)으로 나누고 각 구간에 속하는 데이터 개수를 막대의 높이로 표시한 그래프

데이터의 분포, 빈도, 왜도(비대칭 정도) 등을 파악할 수 있음

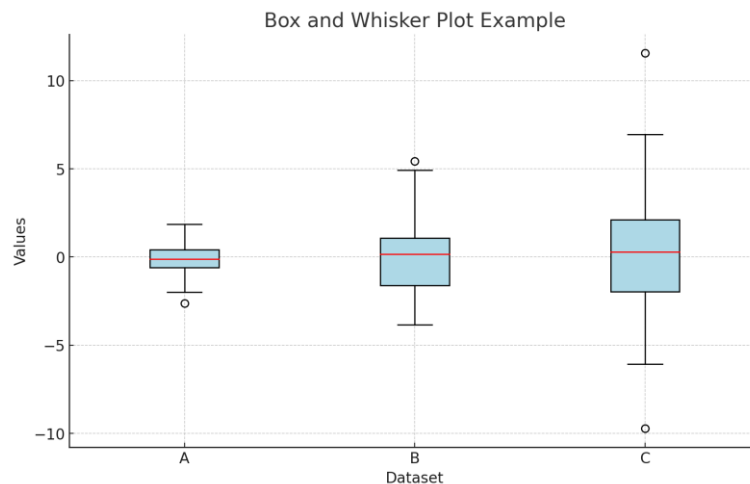


히스토그램 그리기

```
# 2. 히스토그램 (Histogram)
import matplotlib.pyplot as plt
import numpy as np
# 샘플 데이터 생성 (정규분포 따르는 값)
np.random.seed(42)
data = np.random.randn(1000)
# 히스토그램 그리기
plt.hist(data, bins=30, color="skyblue", edgecolor="black", alpha=0.7)
plt.title("Histogram Example")
plt.xlabel("Value")
plt.ylabel("Frequency")
# 결과 출력
plt.show()
```

- 상자 수염 그림 그리기

상자 수염 그림 (box-and-whisker plot): 데이터의 최솟값, 제1사분위수(Q1), 중앙값(Q2), 제3사분위수(Q3), 최댓값을 요약해 상자와 수염 형태로 나타낸 그래프
데이터의 분포 범위와 이상치(outlier)를 쉽게 파악할 수 있음



상자 수염 그림 그리기

```
# 3. 상자 수염 그림 (Box-and-whisker plot)
import matplotlib.pyplot as plt
import numpy as np
# 샘플 데이터 생성
np.random.seed(42)
data = [np.random.normal(0, std, 100) for std in range(1, 4)]
# 박스플롯 그리기
plt.boxplot(data, patch_artist=True,
            boxprops=dict(facecolor="lightblue", color="black"),
            whiskerprops=dict(color="black"),
            capprops=dict(color="black"),
            medianprops=dict(color="red"))
plt.title("Box and Whisker Plot Example")
plt.xlabel("Dataset")
plt.ylabel("Values")
plt.xticks([1, 2, 3], ["A", "B", "C"])
# 결과 출력
plt.show()
```