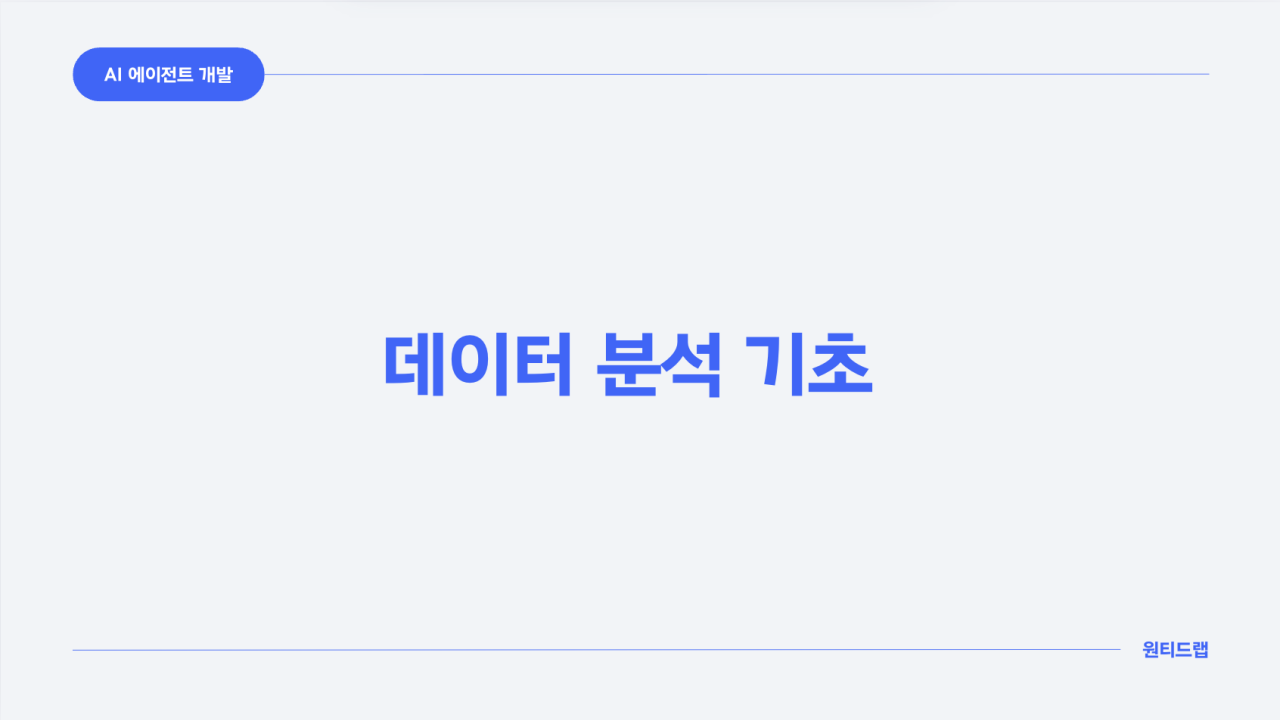
04 데이터 분석 기초 - 데이터 시각화



1. 데이터의 이해

1) 데이터

정형 데이터(Structed Data)

- 표 형태로 구성된 데이터로, 열(Column)에 이름이 있고, 각 열마다 정해진 데이터 형식이 있다.
- 데이터베이스, 엑셀, CSV 파일 등에 많이 사용된다.
- ex. 고객정보, 매출 데이터 등

비정형 데이터(Unstructed Data)

- 형태가 일정하지 않거나 구조화되어 있지 않은 데이터
- 텍스트, 이미지, 음성, 영상 등 다양한 형태를 가짐
- ex. 영화 리뷰, 뉴스 기사, 사진, 음성 녹음 파일

2) 데이터 유형

데이터 유형	설명	예시	추천 시각화	그림을 통해 알 수 있는 것
범주형	정해진 그룹 중 하나, 순서 없음	성별, 지역, 브랜드	막대그래프, 카운트플롯	각 범주의 빈도, 분포
순서형	정해진 순서 있음, 간격은 일정하지 않음	만족도(상중하), 학점	막대그래프, 누적 막대그래프	순위 분포
연속형	소수 포함, 정밀한 수치, 간격 의미 있음	키, 몸무게, 수입	히스토그램, 박스플롯, 산점도	분포, 평균, 이상치, 그룹 간 차이
이산형	개수 등 정수로 표현, 셀 수 있음	자녀 수, 구매 횟수	히스토그램, 막대그래프	분포 형태
시계열	시간 흐름에 따라 측정됨	일별 기온, 주간 매출	선 그래프 (line plot)	추세, 계절성
위치 정보형	위도/경도 등 공간 정보 포함	지점 위치, 사용자 위치	지도 시각화 (folium 등)	집중 지역, 공간적 분포
텍스트	구조 없는 문자 데이터	리뷰, 기사, 댓글	워드클라우드, 네트워크 그래프	핵심 단어, 연관 키워드

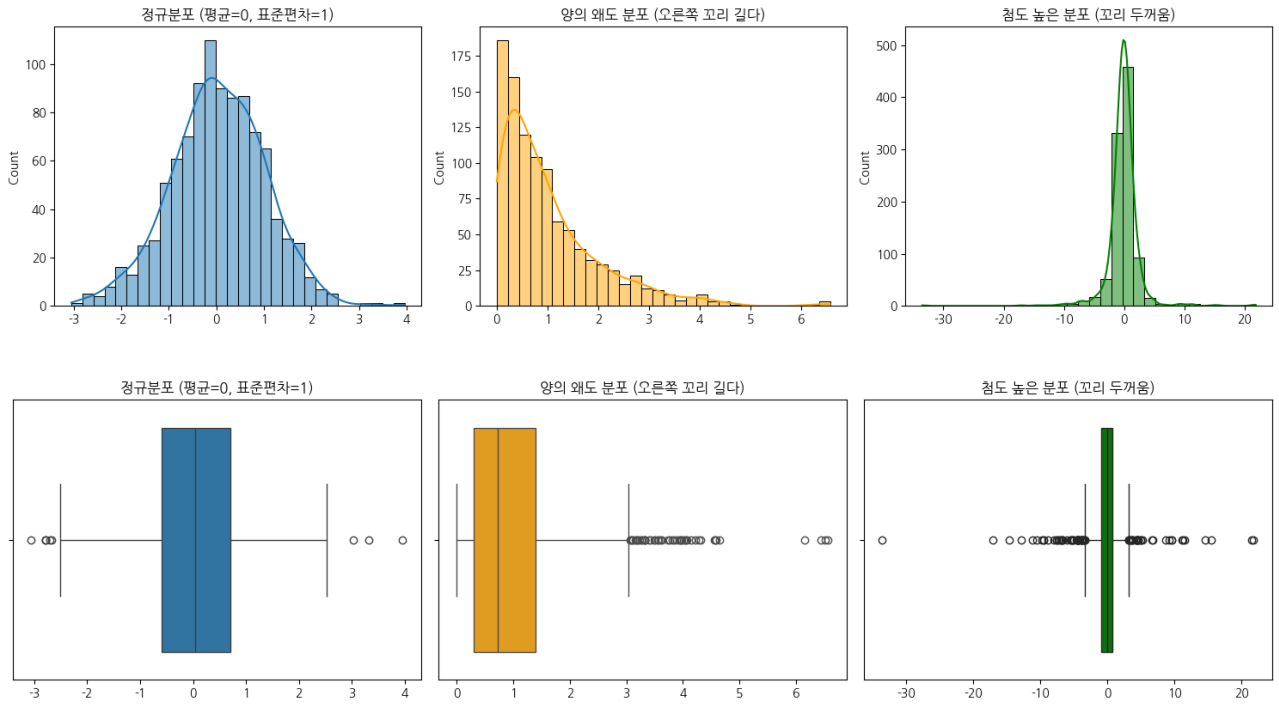
3) 분포의 이해

개념 정리

개념	설명	시각화
평균 (Mean)	전체 데이터를 더한 후 개수로 나눈 값 → 중심 경향성 파악에 유용	히스토그램, 박스플롯
중앙값 (Median)	데이터를 크기순으로 나열했을 때 가운데 위치한 값 → 이상치에 영향을 덜 받음	박스플롯
분산/표준편차	값들이 평균으로부터 얼마나 퍼져 있는지를 나타냄 → 데이터의 **흩어짐 정도**를 설명	박스플롯
이상치 (Outlier)	일반적인 데이터 범위에서 벗어난 값 → 오류일 수도 있고, 중요한 패턴일 수도 있음	박스플롯, 산점도

> ⓘ **분산(Variance)과 표준편차(Standard Devication)의 차이**
>
> - 분산: 데이터가 평균에서 얼마나 퍼져 있는지를 나타내는 값. $(데이터 - 평균)^2$ 의 평균을 의미한다.
> - 표준편차: 분산에 제곱근을 씌운 값. 실제 데이터와 같은 단위를 가지기 때문에 해석이 쉽다.
>
> ex. 키의 분산=100이라고 할 때, 평균적으로 약 $\sqrt{100} = 10$cm 정도 퍼져있다라고 해석할 수 있다.

분포 파악



2. 데이터 시각화 도구

1. Matplotlib

Matplotlib Documentation

기능	예시 코드	설명
기본 그래프	`plt.plot()`	추세 파악
막대 그래프	`plt.bar()`	범주형 데이터 시각화
히스토그램	`plt.hist()`	분포 확인
산점도	`plt.scatter()`	관계 확인
박스플롯	`plt.boxplot()`	이상치/분포 확인

2. Pandas 내장

Pandas Documentation

기능	예시 코드	설명
선 그래프	`df.plot()`	추세 파악
막대 그래프	`df.plot.bar()`	범주형 데이터 시각화
히스토그램	`df.plot.hist()`	분포 확인
박스플롯	`df.plot.box()`	이상치/분포 확인

3. Seaborn

Seaborn Documentation

기능	예시 코드	설명
막대 그래프	`sns.barplot()`	요약 통계 데이터 시각화
카운트 그래프	`sns.countplot()`	범주형 데이터 시각화
산점도	`sns.scatterplot()`	연속형 변수의 그룹별 산점도

기능	예시 코드	설명
상관 관계	`sns.heatmap()`	상관 계수 시각화
박스플롯	`sns.boxplot()`	범주형 그룹별 분포/이상치 확인
페어플롯	`sns.pairplot()`	변수 간 관계 전체 보기

4. Folium

Folium Documentation

기능	예시 코드	설명
지도 생성	`folium.Map()`	중심 좌표 지정
마커 표시	`folium.Marker()`	특정 위치 강조
원형 표시	`folium.Circle()`	범위 시각화
팝업	`popup='설명'`	지도에 설명 표시

3. 한글 폰트 깨짐 이슈

1) 도구

```
uv add matplotlib-koreanize
```

```
import matplotlib_koreanize
```

2) 직접 설정

(1) 내가 가진 폰트 확인하기

```
# 내가 가지고 있는 폰트 목록 확인하기
from matplotlib import font_manager

search_font = "gothic"    # 필터

for font in font_manager.findSystemFonts():
  font_info = font_manager.FontProperties(fname=font)
  font_name = font_info.get_name()
  font_path = font_info.get_file()
  if search_font in font_name.lower():
    print(font_name, font_path)

## 출력 예시
# NanumGothic Eco /usr/share/fonts/truetype/nanum/NanumGothicEcoExtraBold.ttf
# NanumBarunGothic /usr/share/fonts/truetype/nanum/NanumBarunGothicLight.ttf
# NanumGothicCoding /usr/share/fonts/truetype/nanum/NanumGothicCoding.ttf
```

(2) Matplotlib에 내가 가진 폰트 등록하기

1번에서 URL을 복사하여 `font_path`에 넣는다.

```
# 폰트 등록하기
from matplotlib import font_manager

## 폰트 경로
font_path = "/usr/share/fonts/truetype/nanum/NanumGothicCodingBold.ttf"
## 폰트 추가
```

```
font_manager.fontManager.addfont(font_path)
## 폰트 이름 확인
font_name = font_manager.FontProperties(fname=font_path).get_name()
print(font_name)
```

(3) Font Family 설정하기

2번의 `font_name`과 동일하게 작성한다.

```
# 폰트 설정하기
font_name = "NanumGothic Eco"

# 폰트 설정
plt.rcParams["font.family"] = font_name
plt.rcParams['axes.unicode_minus'] = False   # 마이너스 부호 깨짐 방지
```