

5차시

matplotlib을 활용한 데이터 시각화



I. 데이터 시각화

- 데이터 시각화란?

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

- 데이터 시각화 사례

Gap Minder

[https://www.gapminder.org/tools/#\\$chart-type=bubbles&url=v2](https://www.gapminder.org/tools/#$chart-type=bubbles&url=v2)

스웨덴의 통계학자 한스 로슬링(Hans Rosling)이 만든 인터랙티브 시각화 도구.
세계 각국의 GDP, 기대수명, 인구데이터를 시간의 흐름에 따라 애니메이션으로 시각화.

x축: 1인당 GDP (소득)

y축: 기대 수명 (건강)

버블 크기: 인구 수

시간 축: 1800년~현재까지 애니메이션으로 변화

The Pudding

<https://pudding.cool/>

The Pudding은 인터랙티브 데이터 스토리텔링을 전문으로 하는 미디어
사회문화 이슈를 감각적으로 시각화

How Music Taste Evolved in the U.S.

<https://pudding.cool/projects/music-history/>

미국의 음악 차트 데이터를 통해, 장르의 유행 흐름을 시각화

연도별 인기 곡을 애니메이션 그래프로 보여줌

마우스를 올리면 실제 곡이 재생됨

How do animals sound across languages?

<https://pudding.cool/2025/03/language/>

다양한 동물의 울음소리가 각국의 언어로 어떻게 표현되는지 시각화



- 탐색적 분석(EDA)과 보고(presentation)의 차이

구분	탐색적 분석 (EDA)	보고용 시각화 (Presentation)
목적	데이터의 패턴·관계·이상치를 파악	인사이트를 전달하고 설득
대상	본인 또는 팀원 (분석자 중심)	고객, 의사결정자, 대중 등 (청중 중심)
스타일	자유롭고 실험적인 시각화	간결하고 메시지 중심적인 시각화
그래프 수	많고 복잡해도 OK (한눈에 파악할 수 없어도 됨)	적고 명확해야 함 (한눈에 핵심 전달)
기본 질문	“여기서 뭔가 흥미로운 게 있을까?”	“이건 왜 중요한가요?”

- 정보를 ‘읽는다’는 것

시각화에서 “읽는다”는 것은 단순히 보는 것을 의미하지 않는다.

-> 시선을 따라 의미를 해석하고, 나의 판단으로 이어지는 과정 전반

우리가 그래프를 ‘읽을 때’ 실제로는:

- 무엇이 강조되었는가를 찾기
- 축, 색상, 범례를 이해하기
- 패턴이나 추세를 파악하기
- 왜 이걸 보여주는지를 추론하기

이 과정은 텍스트 독해보다 훨씬 빠르지만,

명확한 의도 없이 설계된 시각화는 오히려 오해를 낳아 데이터 전달과 이해에 방해가 된다.

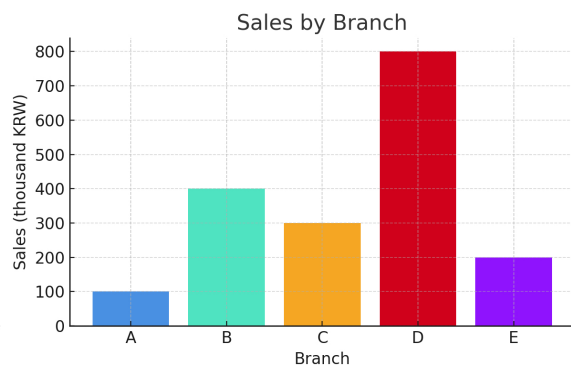
예시:

불필요하게 3D로 만든 bar chart → 왜곡된 비율 인식

색상 대비가 약한 heatmap → 중요한 영역이 안 보임

“정보를 읽히게 만드는 시각화”를 하려면,

메시지를 중심으로 설계된 구조가 필수적이다.



II. matplotlib 구조와 사용법 소개

- matplotlib의 위치와 역할

matplotlib은 파이썬에서 가장 기본이 되는 시각화 라이브러리이다.

- “데이터를 픽셀로 바꿔주는 엔진” 역할을 수행
- 낮은 수준의 설정(좌표축, 눈금, 색상, 폰트 등)까지 세밀하게 제어 가능
- 단순 선 그래프부터 복잡한 여러 축을 가진 레이아웃까지 표현 가능
- 탐색 단계에서는 빠르게 plt.plot()같은 함수로 간단히 그려보고,
- 보고용 단계에서는 세밀하게 스타일을 다듬어 **출판 품질의 그래프**를 만들 수 있음

- Figure, Axes, Axis 구조 모델의 이해

matplotlib의 내부 구조는 이렇게 계층화된다:

Figure → Axes → Axis

요소	설명	비유
Figure	전체 그래프를 담는 “캔버스”이다	그림판, 슬라이드
Axes	실제 데이터가 그려지는 영역(하나의 plot)이다	도화지 위 작은 프레임
Axis	x, y 축(눈금·라벨 포함)을 관리한다	눈금 자

하나의 Figure에 여러 Axes를 배치할 수 있다 → plt.subplots()

각 Axes 안에서 데이터와 제목, 범례 등을 제어한다.

예시 코드:

```
fig, ax = plt.subplots(figsize=(6,4))
ax.plot(x, y)
ax.set_title("Sample Plot")
ax.set_xlabel("X value")
ax.set_ylabel("Y value")
```

- pyplot 함수와 스타일 설정의 흐름

pyplot은 matplotlib을 간단히 조작할 수 있는 상태 기반 인터페이스이다.

plt.plot()은 현재 활성 Figure/Axes에 바로 선 그래프를 그린다.

plt.xlabel(), plt.title()은 그려진 그래프에 레이블과 제목을 추가한다.

스타일 설정은 두 가지 흐름이 있다.

빠른 탐색용: plt.plot()plt.title()plt.show()

코드가 짧고 직관적이지만, 여러 그래프를 동시에 관리하기에는 불편하다.

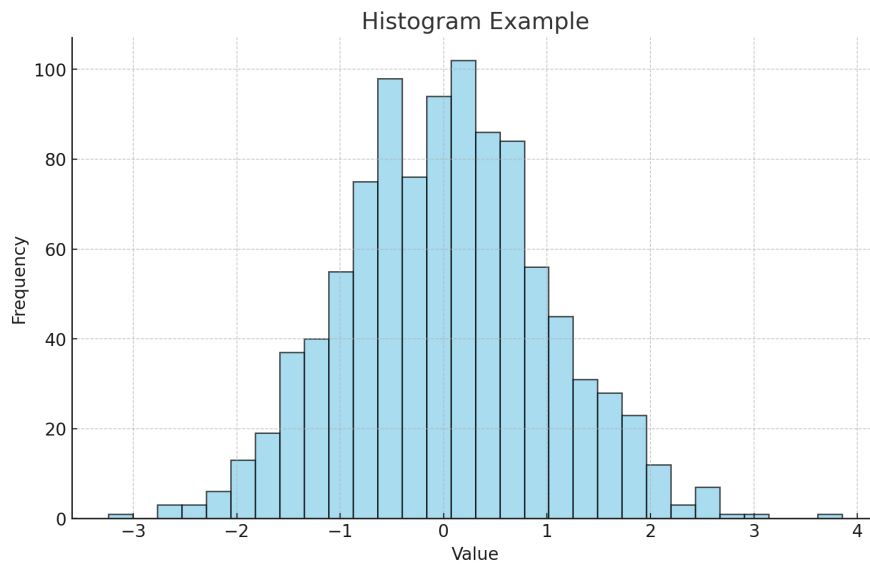
객체 지향형(권장): Figure/Axes를 변수에 담아 세밀하게 제어한다.



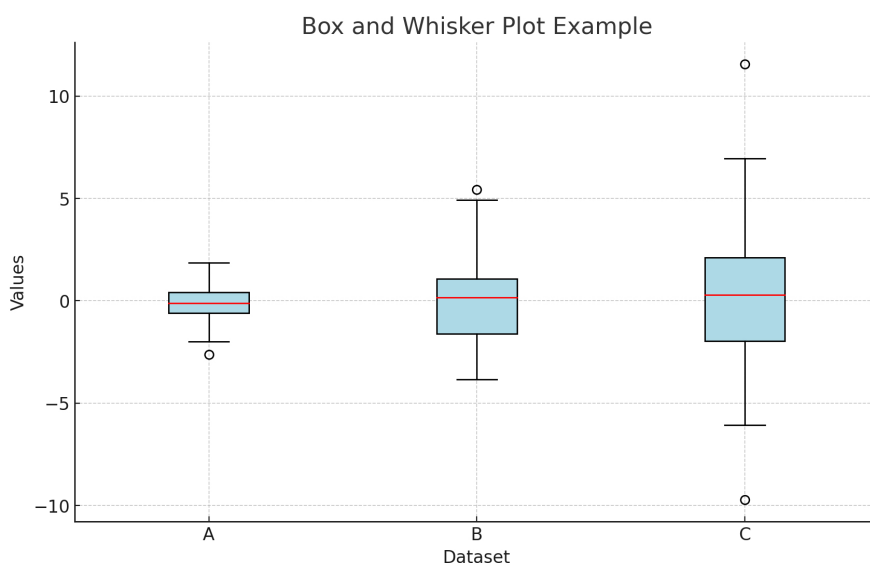
III. 데이터 유형별 시각화

- 수치형 데이터:

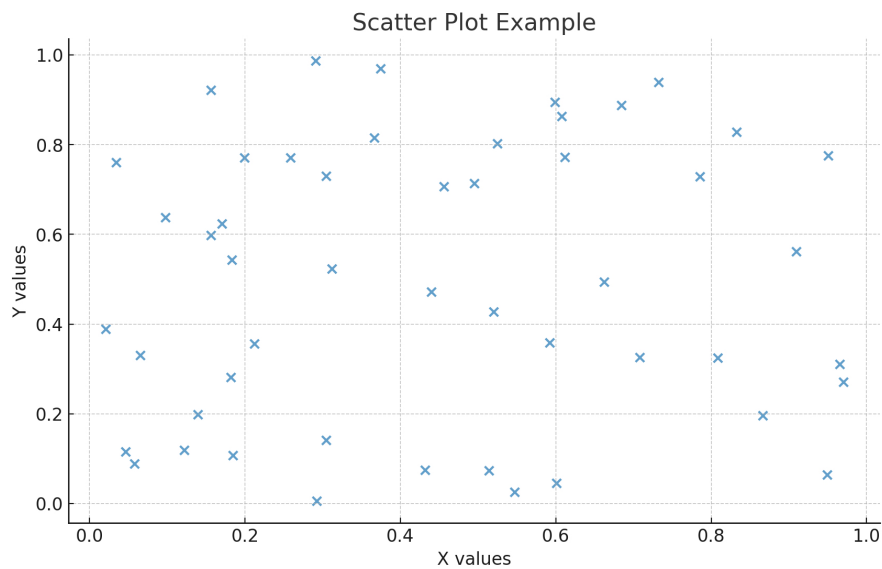
histogram (분포 확인)



box-and-whisker plot (이상치 검출)

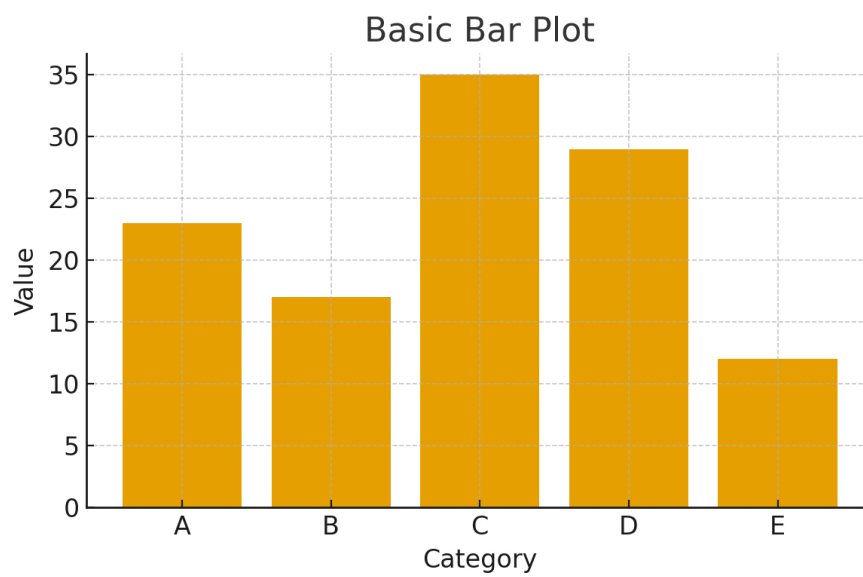


scatter (관계성 확인)

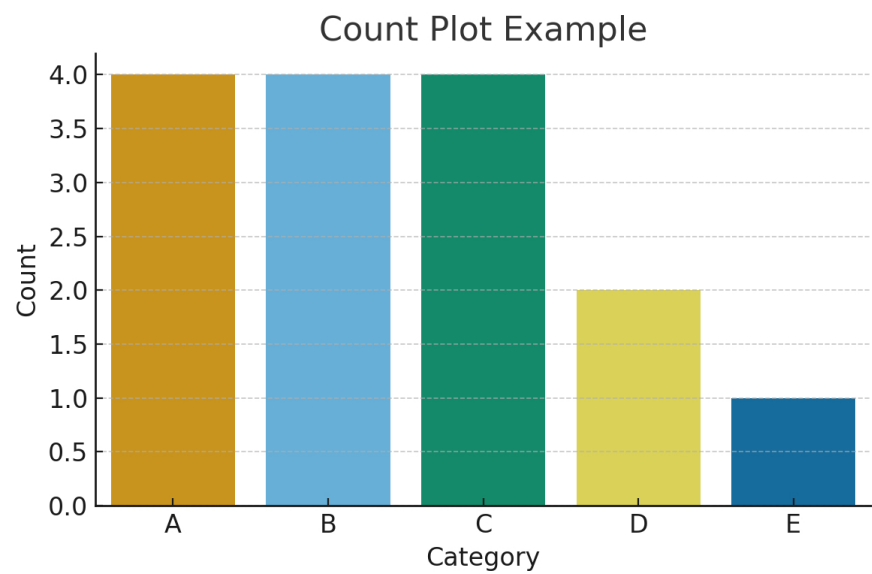


- 범주형 데이터:

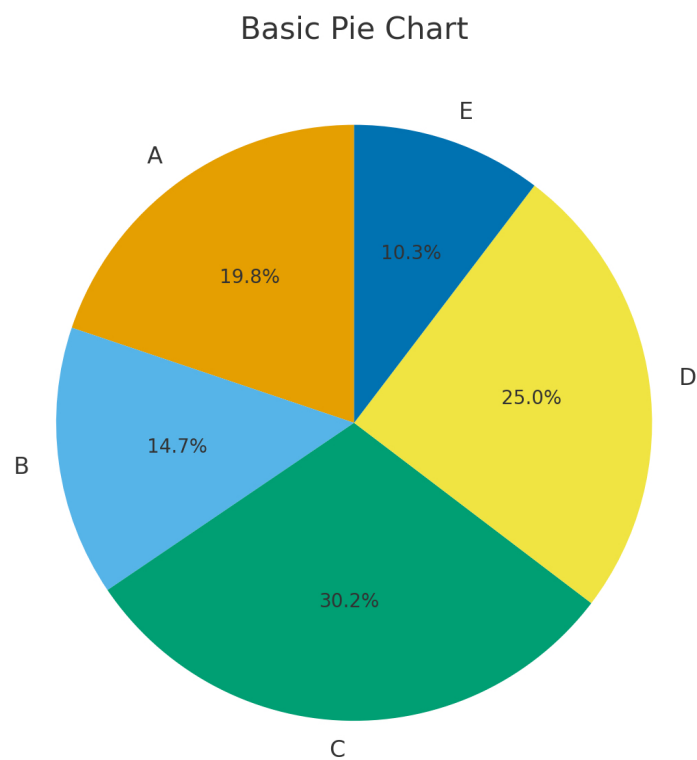
barplot



countplot

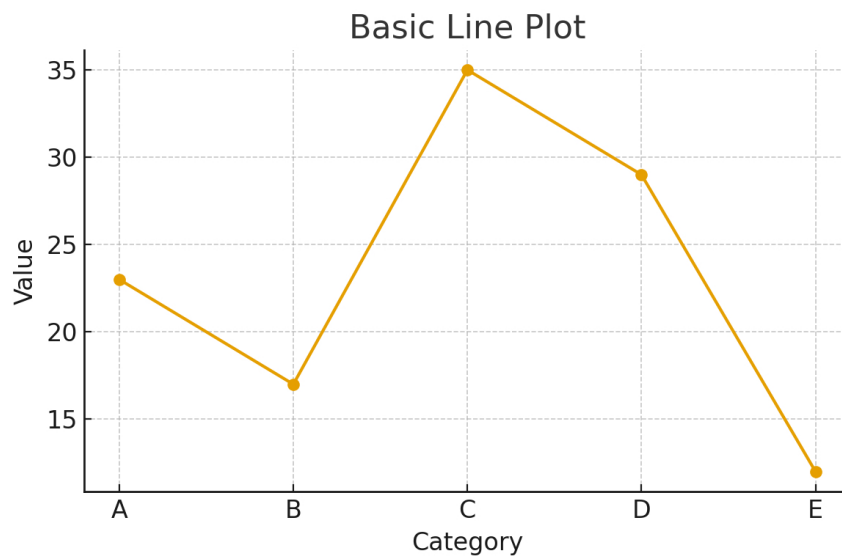


pie chart

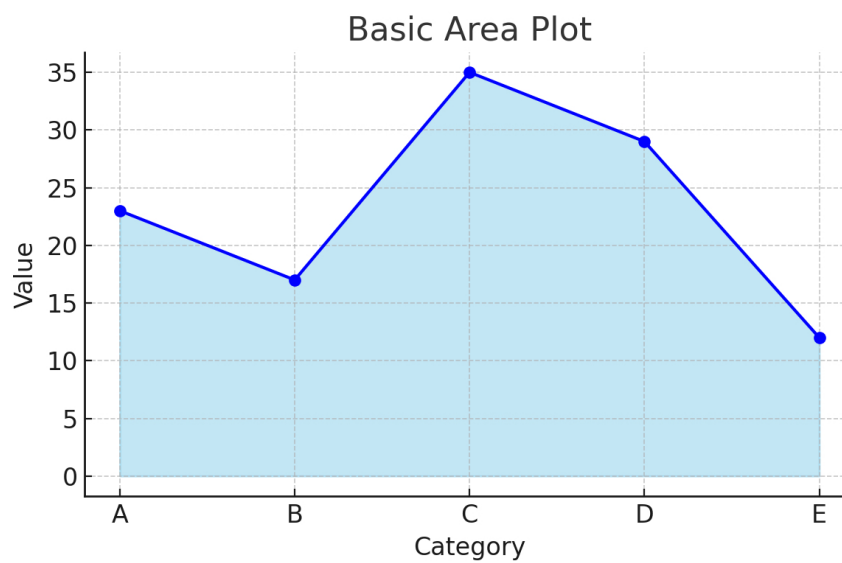


- 시계열 데이터:

lineplot

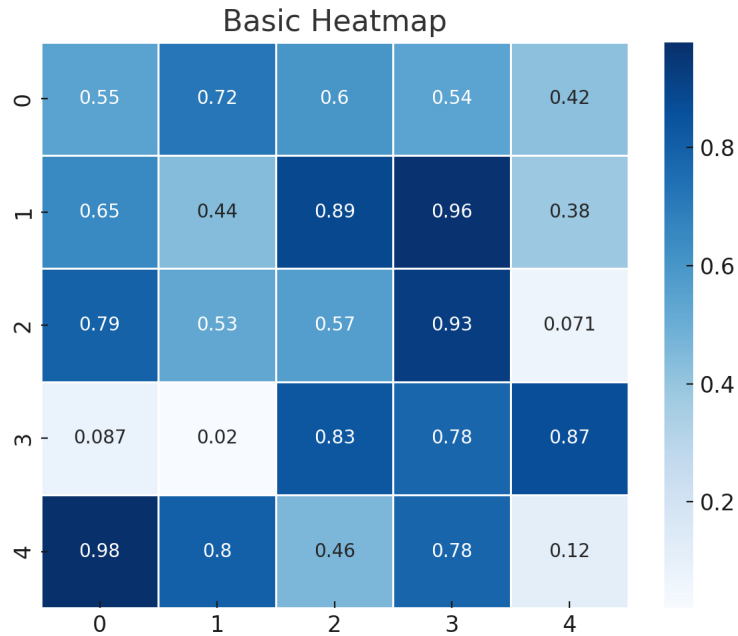


area plot (변화의 흐름 시각화)

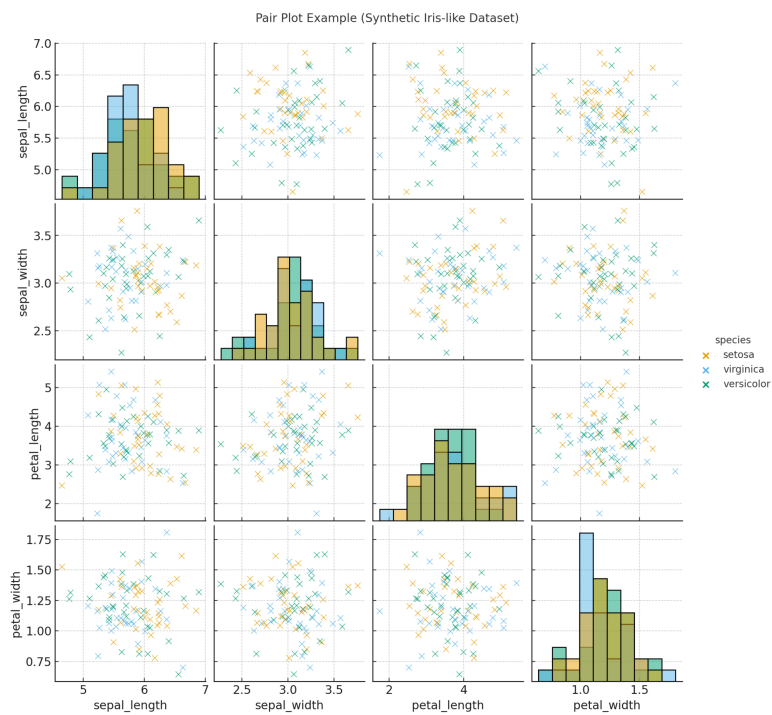


- 상관관계:

heatmap



pairplot (변수 간 관계 시각화)



- 각 시각화 예시와 함께 ‘어떤 질문에 어떤 그래프가 적절한가?’ 연결하기



1. Pairplot이란?

여러 개의 수치형 변수가 있을 때, 그 변수들을 모두 쌍으로 짝지어 산점도(scatter plot)를 그려주는 도구다. 대각선에는 보통 변수 하나의 분포(히스토그램이나 KDE 곡선)가 들어간다.

hue옵션을 주면, 범주형 변수에 따라 색을 다르게 표시해 그룹별 차이를 쉽게 볼 수 있다.

즉, Pairplot은 데이터의 전반적인 구조와 변수 간 관계를 한눈에 스캔하는 지도 같은 역할을 한다.

-> 탐색적 데이터 분석(EDA)단계에서 자주 사용

2. 어떻게 읽는가?

예를 들어 iris데이터셋(붓꽃)에서 4개의 변수(sepal length, sepal width, petal length, petal width)가 있다고 가정한다.

대각선(히스토그램)

각 변수 하나씩의 분포를 보여준다.

→ 예: petal length는 값이 1~6cm 사이에 주로 몰려 있다.

비대각선(산점도)

두 변수 사이의 관계를 보여준다.

→ 예: petal length와 petal width는 양의 상관관계(길수록 폭도 넓다)를 가진다.

→ sepal length와 sepal width는 큰 상관이 없거나 음의 상관이 약하게 나타난다.

색깔(hue)

다른 그룹(예: 꽃 품종 setosa, versicolor, virginica)을 구분해서 보여준다.

→ 예: setosa는 petal length가 짧고, virginica는 길어서 그룹별로 뚜렷이 구분된다.

3. 해석 포인트

변수 간 상관관계 확인

산점도의 패턴이 직선에 가까우면 상관관계가 존재한다.

점이 흩어져 있으면 관계가 약하거나 존재하지 않는다.

클러스터(군집) 여부

특정 그룹이 다른 그룹과 뚜렷하게 떨어져 있으면, 분류(classification)에 잘 활용할 수 있다.

변수의 분포 확인

대각선 히스토그램에서 분포가 한쪽으로 치우쳐 있거나, 여러 봉우리가 있는지 확인한다.



IV. 시각화 스토리텔링과 인사이트

1. 시각화는 메시지를 전달하는 스토리텔링 도구이다

단순히 데이터를 보여주는 것이 목적이 아닌
데이터를 통해 “무엇을 말하고 싶은가”가 핵심이다.

예: “매출이 줄었다” → “특정 제품군의 매출이 급격히 줄어든 것을 보면 시장 구조가 변하고 있다”

2. 그래프에 생명을 불어넣는 요소들

강조

색상: 핵심 데이터 포인트에만 강렬한 색을 넣는다.

크기: 중요도를 크기로 표현한다.

주석(annotation): 숫자만 보여주는 대신 설명을 붙인다.

흐름

배치 순서: 그래프를 보는 사람이 자연스럽게 이해할 수 있도록 정렬한다.

시선 유도: 화살표, 강조 라인 등을 사용해 중요한 부분에 눈이 머물게 한다.

해석

캡션·제목: 그래프만 봐도 메시지가 드러나도록 작성한다.

부가 텍스트: “그래서 이게 의미하는 바는?”에 대한 짧은 답변을 제공한다.

