

복잡한 데이터를 쉽게 이해할 수 있는 형태로 변환: 데이터 시각화

시각화 전 어떤 메시지를 전달할지 먼저 정의하는 것이 중요하다: O,X

시간에 따른 값의 변화: 꺾은선형

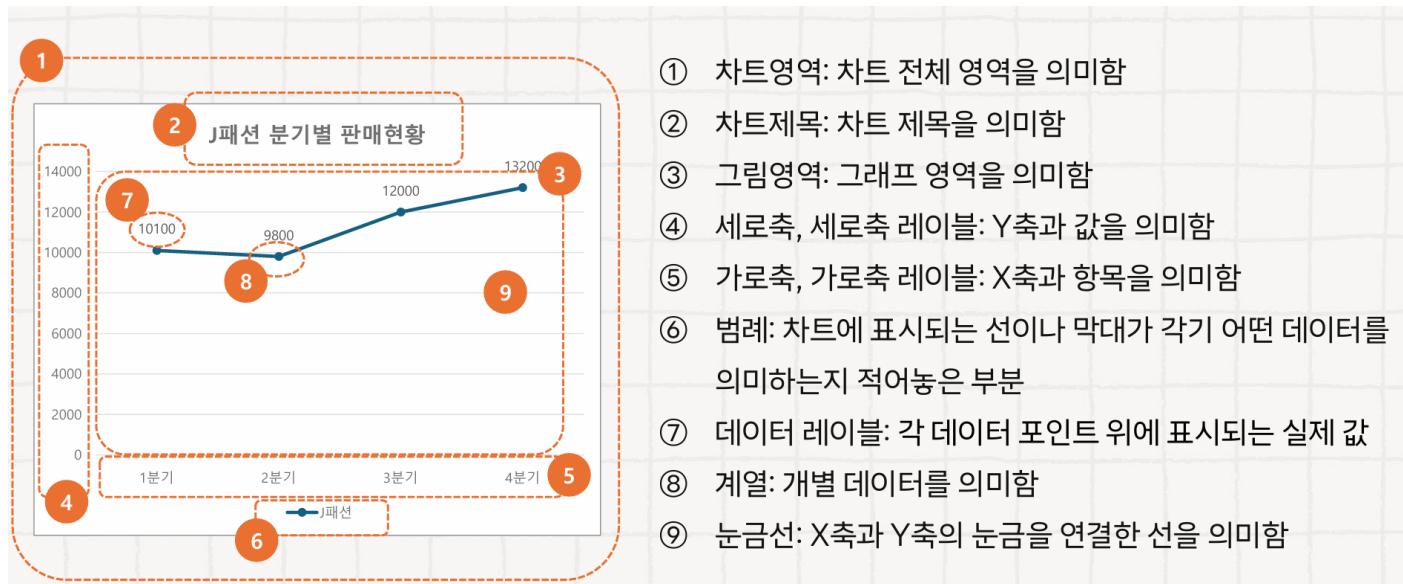
수치의 크고 작음 비교: 세로 막대형

시간에 따른 개별 항목 추이 및 합계 추이, 변동요인 파악: 누적 세로 막대형 차트

시간에 따라 개별 항목이 전체에서 차지하는 비율을 파악하고 싶을 때: 100% 기준 누적 세로 막대형

항목이 길거나 항목수가 많은 데이터 간 비교: 가로 막대형 차트

특정 시점의 개별 항목이 차지하는 비율을 파악하고 싶을 때: 원형 차트



보기 좋은 차트 꾸미기 꿀팁

- 눈금선 개수 조정하기(데이터 범위 안에서 필요한 최소한의 눈금선 표시)
- 글씨는 기본 설정보다는 (크게)
- 세로축에는 (단위 표기)하기
- 꺾은선형 차트에선 범례 대신 (데이터 레이블) 활용하기
- 강조하고 싶은 데이터는 (따뜻한 색상) 사용하기

두 변수 사이 관계 이해 및 예측 위한 통계분석 방법: 단순회귀분석

단순회귀분석 식: $Y = a + bX$

Y: 종속변수, X: 독립변수, a: Y축 절편, b: 회귀계수

회귀 분석은 데이터의 점들을 가장 잘 설명할 수 있는(최적의 직선)을 찾는 과정

실제값 - 예측값 : 잔차

잔차의 제곱의 합을 (최소화)하는 방향으로 회귀방정식을 만듬

독립변수로 설명되지 못한 부분: 오차항

b> 0 : 양의방향, b< 0:음의 방향

통계적으로 신뢰할 수 있는지 판단: 통계적 유의성

실질적으로 중요한지: 경제적 유의성

독립변수가 종속변수에 통계적으로 의미 있는지 지표: p-value, 유의확률

- 일반적 기준 : 0.05

경제적 유의성 판단 기준: 회귀계수 크기

다양한 요인을 동시 고려하여 예측: 다중회귀분석

다중회귀분석 회귀모형: $Y = a + b_1x_1 + b_2x_2 \dots b_nx_n + e$

Y: 종속변수 a: Y축 절편, X: 독립변수, :회귀계수, e:오차항

각 독립변수에 회귀계수는 다른 변수들이 일정하다고 가정한다: O, X

독립변수들이 종속변수의 변동을 얼마나 설명하는지의 비율: R^2

- 한국어: 결정계수

빅데이터의 도전 과제: 빅데이터의 양면성

빅데이터 기회: 효율적 의사결정, 기술혁신 가속화, 사회적 변화 유도

빅데이터 도전: 정보 격차, 차별 발생, 사생활 침해

빅데이터 관리 안전장치: 데이터 윤리

데이터를 수집, 관리, 사용하는 모든 과정에서의 (사회적 책임) 과 (도덕적 기준)을 준수하는 원칙

- 종류: 개인정보 등의 원칙, 최소 수집 원칙

단일 정보로 개인을 못 알아보면 개인정보가 아니다: O, X

데이터를 저장하고 관리할 때 지켜야 할 윤리적 기준: 데이터 관리윤리

- 3개: 데이터 보안, 데이터 보관기간 준수, 접근 제한

데이터를 사용할 때 지켜야 할 윤리적 기준: 데이터 사용윤리

- 4개: 공정성 원칙, 익명성 보장, 목적 준수 원칙, 투명성 원칙

데이터 윤리 필요 이유: 개인 사생활 보호, 신뢰 확보, 부정적 영향 최소화