October 26, 2025

▣ **강의명:** CS109A: 데이터 과학 입문

▣ **주차:** Lecture 02

▣ **교수명:** Pavlos Protopapas, Kevin Rader, Chris Gumb

▣ **목적:** Lecture 02의 핵심 개념 학습

□ 핵심 요약

본 문서는 데이터 과학의 핵심 구성요소인 '데이터'가 무엇인지 정의하고, 데이터를 수집, 저장, 분류하는 방법을 다룹니다. 또한, 데이터의 특성을 요약하는 '기술 통계' 방법(평균, 분산 등)과, 데이터에 숨겨진 패턴을 찾는 '시각화'의 중요성(앤스컴 4중주) 및 다양한 시각화 기법(히스토그램, 산점도, 박스 플롯 등)을 설명합니다. 이 자료는 데이터 과학 프로세스의 2, 3단계(수집 및 탐색)의 기초를 다룹니다.

Contents

1 핵심 용어 정리

데이터 분석을 시작하기 위해 꼭 알아야 할 기본 용어들입니다.

Table 1: 데이터 분석 핵심 용어

용어	쉬운 설명	원어	비고
데이터	관찰을 통해 수집된 사실, 값, 정보의 집합.	Data	단수형은 Datum. (Data는 복수형)
정형 데이터	엑셀 시트처럼 행과 열로 명확히 구조화된 데이터.	Tabular Data	"Tidy Data"라고도 함.
관측치	분석 대상의 개별 단위 (엑셀의 '행').	Observation	예: 한 명의 사람, 한 개의 영화.
변수	측정하려는 특성 (엑셀의 '열').	Variable	예: 나이, 평점. (Feature라고도 함)
모집단	연구 대상이 되는 전체 집단.	Population	예: 이 수업을 듣는 '모든' 학생.
표본	모집단에서 추출한 '일부' 대표 집합.	Sample	예: 오늘 수업에 '출석한' 학생.
EDA	시각화와 통계를 통해 데이터의 패턴을 탐색하는 과정.	Exploratory Data Analysis	"탐색적 데이터 분석"
[중심 측정]			
평균	모든 값을 더해 개수로 나눈 값. (무게 중심)	Mean ($\bar{x}$)	이상치(Outlier)에 매우 민감함.
중앙값	데이터를 순서대로 나열했을 때 딱 중간에 있는 값.	Median	이상치에 둔감함 (Robust).
최빈값	데이터에서 가장 자주 등장하는 값.	Mode	범주형 데이터에서 주로 사용.
[퍼짐 측정]			
분산	데이터가 평균에서 얼마나 멀리 퍼져있는지의 정도.	Variance (s^2)	단위가 원래 단위의 '제곱'이 됨.
표준편차	분산에 제곱근을 씌운 값.	Standard Dev. (s)	원래 데이터와 단위가 동일해 해석이 쉬움.
[시각화]			
히스토그램	수량형 데이터의 '분포'를 막대로 표현.	Histogram	구간(bin) 너비에 따라 모양이 변함.
막대 그래프	범주형 데이터의 '빈도'를 막대로 비교.	Bar Plot	막대 순서를 바꿔도 의미가 통함.
산점도	두 수량형 변수 간의 '관계'를 점으로 표현.	Scatter Plot	방향, 강도, 형태, 이상치를 봄.
박스 플롯	범주형 그룹 간 수량형 데이터의 분포를 '요약' 비교.	Box Plot	중앙값, 사분위수, 이상치를 보여줌.

2 핵심 개념 1: 데이터란 무엇인가? (What is Data?)

데이터 과학(Data Science)은 이름 그대로 '데이터'에서 시작합니다.

2.1 데이터의 정의

- **데이터(Data):** 관찰이나 측정을 통해 얻은 여러 개의 정보 조각들입니다. (복수형)
- **데이터(Datum):** 정보 조각 '하나'를 의미합니다. (단수형)

과거에는 데이터가 주로 숫자(Numeric)였지만, 현대에는 텍스트, 이미지, 소리 등 모든 것이 데이터가 될 수 있습니다.

2.2 데이터 수집 방법 (어디서 오는가?)

데이터는 크게 세 가지 경로로 얻을 수 있습니다.

1. **내부 소스 (Internal Sources):**
   - 조직이나 개인이 직접 수집한 1차 데이터입니다.
   - 예: 과학 실험 결과, 임상 시험 데이터, 회사 내부의 판매 기록.
2. **기존 외부 소스 (Existing External Sources):**
   - 이미 누군가 수집/가공하여 공개한 데이터입니다.
   - 예: 정부 공공 데이터 포털, Kaggle 데이터셋, 스포츠 기록 사이트.
3. **수집이 필요한 외부 소스 (External Sources Requiring Collection):**
   - 외부에 존재하지만, 가져오려면 별도의 노력이 필요한 데이터입니다.
   - 이 강의에서 주목하는 방식이며, 주로 온라인 데이터를 의미합니다.

2.3 온라인 데이터 수집 3가지 방법

온라인에서 데이터를 가져오는 대표적인 3가지 기술입니다.

1. API (Application Programming Interface) • **개념:** 회사가 외부 사용자가 자신의 데이터나 서비스에 "합법적으로" 접근할 수 있도록 열어둔 '공식 창구'입니다.

- **특징:** 보통 사용량 제한이 있거나 유료입니다. 안정적이고 정확한 데이터를 제공받습니다. (예: 구글 지도 API, 스포티파이 API)
- **필요 기술:** 파이썬(Python)과 각 API의 사용 설명서(Dictionary)를 읽는 능력.

2. RSS (Rich Site Summary) • **개념:** 블로그나 뉴스 사이트처럼 '자주 업데이트되는' 콘텐츠를 요약하여 스트림(Stream) 형태로 제공하는 규격입니다.

- **특징:** 무료이며, 주로 새로운 게시물의 제목, 요약, 링크를 받아볼 때 사용됩니다.

3. 웹 스크래핑 (Web Scraping) • **개념:** 웹사이트의 HTML 코드에서 직접 필요한 정보를 '추출'하는 기술입니다.

- **특징:** API가 없거나 유료 API를 우회하고 싶을 때 사용됩니다. (예: 위키피디아의 표(table) 정보를 긁어오는 것)

주의사항

웹 스크래핑의 윤리적/법적 문제

웹 스크래핑은 강력하지만 매우 조심해야 하는 기술입니다.

- **서비스 약관(Terms of Service) 위반:** 많은 웹사이트가 스크래핑을 명시적으로 금지합니다.
- **개인정보 침해:** 사용자의 비공개 정보를 수집하면 안 됩니다.
- **서버 부하:** 과도한 스크래핑은 대상 웹사이트의 서버를 마비시킬 수 있습니다 (DoS 공격과 유사).
- **해악(Harm):** 수집한 데이터를 통해 사생활을 침해하거나 불법적인 용도로 사용해서는 안 됩니다.

항상 데이터를 수집하기 전에 "이 데이터를 사용해도 되는가?"를 먼저 질문해야 합니다.

3 핵심 개념 2: 데이터의 유형과 구조

데이터를 수집했다면, 그 형태와 유형을 파악해야 합니다.

3.1 데이터 유형 (Data Types)

- **원자적 유형 (Atomic Types):** 더 이상 쪼갤 수 없는 기본 단위입니다.
  - **수량형 (Numeric):** 정수(Integers, 예: 109), 실수(Floats, 예: 3.14)
  - **부울형 (Boolean):** 참/거짓 (True/False, Yes/No, 1/0)
  - **문자열 (Strings):** 텍스트 (예: "Hello")
- **복합 유형 (Compound Types):** 원자적 유형들이 모여 구성됩니다.
  - **리스트 (Lists):** 순서가 있는 값의 모음 (예: `[1, 2, 3]`)
  - **사전 (Dictionaries):** '키(Key)'와 '값(Value)'이 짝을 이룬 모음 (예: `{"name": "Kevin"}`)

3.2 데이터 저장 구조

정형 데이터 (Tabular Data) 가장 중요합니다. 엑셀 시트나 CSV 파일처럼 2차원 테이블(표) 형태입니다. 대부분의 데이터 분석 패키지(예: Pandas)는 이 형태를 기본으로 가정합니다.

반정형 데이터 (Semistructured Data) JSON, XML처럼 키-값 쌍으로 이루어져 있지만, 정형 데이터처럼 엄격한 행/열 구조를 따르지 않을 수 있습니다.

비정형 데이터 (Unstructured Data) 텍스트 문서, 이미지, 오디오 파일 등 구조가 없는 데이터입니다.

3.3 정형 데이터 (Tabular Data) 집중 탐구

정형 데이터는 데이터 분석의 표준입니다.

- **관측치 (Observations):** 표의 **행 (Row)**. 분석하려는 개별 대상 하나하나를 의미합니다. (예: 영화 1개, 학생 1명)
- **변수 (Variables):** 표의 **열 (Column)**. 관측치에서 측정한 특정 속성입니다. (예: 영화 평점, 학생 나이)

```
# imdb_top_1000.csv 파일을읽어들임
imdb = pd.read_csv('imdb_top_1000.csv')
# 처음줄 5(head)을 출력
imdb.head()
```

Listing 1: Pandas를 이용한 정형 데이터(CSV) 로딩 예시

[*lst:pandas*의 출력 결과: *IMDB* 영화 목록 표]
각 행(*Row*)은 영화 *1*개(관측치)를 나타내며, 각 열(*Column*)은 *Series_Title, Released_Year, IMDB_Rating* 등(변수)을 나타냅니다.

3.4 변수의 유형: 분석의 첫걸음

주의사항

왜 변수 유형을 구분해야 하나요?
변수의 유형에 따라 사용할 수 있는 **요약 방법(통계)**과 **시각화**가 완전히 달라지기 때문입니다. 예를 들어, '키'는 평균을 낼 수 있지만 '좋아하는 색깔'은 평균을 낼 수 없습니다.

변수는 크게 두 가지로 나뉩니다.

1. **수량형 변수 (Quantitative / Numeric Variable)**
   - 숫자로 측정되며, 산술 연산(+, -)이 의미가 있습니다.
   - **이산형 (Discrete):** 값이 정수처럼 딱딱 떨어져 셀 수 있습니다. (예: 형제자매 수, 주사위 눈금)
   - **연속형 (Continuous):** 값이 특정 범위 내에서 무한히 많은 값을 가질 수 있습니다. (예: 키, 몸무게, 온도)
2. **범주형 변수 (Categorical Variable)**
   - 값이 몇 개의 그룹이나 범주로 나뉩니다.
   - **순서형 (Ordinal):** 범주 간에 자연스러운 순서가 있습니다. (예: 학점 A, B, C / 만족도 '높음', '중간', '낮음')
   - **명목형 (Nominal):** 범주 간에 순서가 없습니다. (예: 혈액형 A, B, O / 좋아하는 애완동물 '개', '고양이', '쥐')

3.5 데이터의 흔한 문제점과 "Tidy Data"

현실의 데이터는 깨끗하지 않습니다.

- **결측치 (Missing values):** 값이 비어있습니다. (이 값을 버릴까요? 아니면 추측해서 채울까요?)
- **오입력값 (Wrong values):** 잘못된 값이 입력되었습니다. (예: 나이 200세)
- **형식 불일치 (Format mismatch):** 두 데이터를 합치려는데 형식이 다릅니다.
- **지저분한 데이터 (Messy Data):** 데이터가 분석하기 어려운 형태로 되어있습니다.

□ 예제: 지저분한(Messy) 데이터 변환하기

다음은 주말 농산물 배송 횟수를 기록한 "지저분한" 표입니다.
Before: (지저분한 형식)

왜 이 형식이 나쁜가요?

- '관측치 1개 = 행 1개' 원칙이 깨졌습니다. 'Morning' 행 하나에 3개의 관측치(금요일 아침, 토요일 아침, 일요일 아침)가 들어있습니다.
- 'Friday'는 변수 이름이어야 하는데, '값'처럼 취급되고 있습니다.
- 이 상태로는 "평균 배송 횟수는?" 또는 "요일별 배송 횟수 합계는?"을 계산하기 어렵습니다.

After: (정형/Tidy 형식) 이 데이터를 분석하기 쉬운 '정형(Tabular)' 또는 'Tidy' 형식으로 바꾸면 다음과 같습니다.

왜 이 형식이 좋은가요?

- **1 관측치 = 1 행:** '금요일 아침'이라는 관측치 하나가 행 하나를 차지합니다.
- **1 변수 = 1 열:** '시간', '요일', '횟수'라는 명확한 변수(열)가 생겼습니다.
- 이제 Pandas 같은 도구로 'Number' 열의 평균을 구하거나, 'Day'별로 그룹화하여 합계를 구하는 것이 매우 쉬워집니다.

4 핵심 개념 3: 탐색적 데이터 분석 (EDA)

탐색적 데이터 분석 (Exploratory Data Analysis, EDA)은 수집한 데이터를 본격적으로 모델링하기 전에, 시각화나 간단한 통계 기법을 통해 데이터의 구조와 패턴을 파악하고, 이상치나 잠재적인 문제점을 발견하는 과정을 말합니다.

EDA는 "데이터와 친해지는 과정"이며, 데이터 과학 프로세스의 3단계에 해당합니다.

4.1 모집단 (Population) vs. 표본 (Sample)

- **모집단 (Population):** 내가 궁극적으로 알고 싶은 대상 '전체'입니다. (예: 하버드의 모든 학생)
- **표본 (Sample):** 모집단 전체를 조사하기는 불가능하므로, 그중 '일부'를 뽑아서 조사한 것입니다. (예: 오늘 CS109A 수업에 온 학생)

우리는 '표본'을 분석해서 '모집단'의 특성을 추측합니다. 이때 가장 중요한 것은 표본이 모집단을 잘 대표해야 한다는 것입니다.

4.2 표본 추출 편향 (Sampling Bias)

표본이 모집단을 잘 대표하지 못할 때 '편향(Bias)'이 발생했다고 말합니다.

선택 편향 (Selection Bias): 특정 하위 그룹이 다른 그룹보다 표본으로 더 잘 선택되는 경우.

무응답/자발적 참여 편향 (Non-response/Volunteer Bias): 응답하기 쉬운 대상만 응답하거나(예: 수업에 출석한 학생), 특정 주제에 열성적인 사람들(예: 앱 얼리 어답터)만 자발적으로 참여하는 경우.

□ 예제: 잘못된 표본 추출 예시

사례 1: 수업 출석률

- **목표:** CS109A 전체 학생의 평균 만족도 조사.
- **표본:** 오늘 수업에 '출석한' 학생들.
- **문제점:** 수업에 결석한 학생들(어쩌면 만족도가 매우 낮아서 안 온 학생들)의 의견이 반영되지 않습니다. (무응답 편향) 이 표본의 만족도 점수는 실제 모집단(전체 학생)의 점수보다 높게 나올 가능성이 큽니다.

사례 2: 신규 앱 기능 테스트

- **목표:** 새로운 앱 기능이 모든 사용자에게 효과가 있는지 테스트.
- **표본:** 신규 기능을 자발적으로 신청한 '얼리 어답터' 그룹.
- **문제점:** 얼리 어답터들은 원래 새로운 기능에 호의적이고 IT 활용도가 높은 집단입니다. (자발적 참여 편향) 이들이 새 기능을 좋아한다고 해서, 변화를 싫어하는 일반 대중(모집단)도 좋아할 것이라고 일반화할 수 없습니다.

5 절차/방법 1: 기술 통계 (Descriptive Statistics)

기술 통계는 수집한 데이터(표본)의 특성을 몇 개의 숫자로 요약하는 방법입니다. 주로 데이터의 '중심'이 어디인지, '얼마나 퍼져있는지'를 봅니다.

5.1 데이터의 "중심" 측정 (Measures of Center)

5.1.1 평균 (Mean)

- **정의:** 모든 값을 더한 뒤, 값의 개수(n)로 나눈 값.
- **직관:** 데이터 분포의 "무게 중심" 또는 "균형점".
- **공식:** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1+x_2+\cdots+x_n}{n}$
- **단점: 이상치(Outlier)**에 매우 민감합니다.
- **예시:**
  - 데이터: `[1, 2, 3, 4, 5]` → 평균: $(1+2+3+4+5)/5 = 3$
  - 데이터: `[1, 2, 3, 4, 100]` → 평균: $(1+2+3+4+100)/5 = 22$
  - '100'이라는 극단값 하나 때문에 무게 중심이 3에서 22로 확 이동했습니다.

5.1.2 중앙값 (Median)

- **정의:** 데이터를 크기순으로 정렬했을 때, 정확히 '가운데'에 위치한 값.
  - 데이터 개수(n)가 홀수: 가운데 1개 값.
  - 데이터 개수(n)가 짝수: 가운데 2개 값의 평균.
- **장점:** 이상치에 거의 영향을 받지 않습니다. (Robust)
- **예시:** (위의 예시를 다시 사용)
  - 데이터 (정렬됨): `[1, 2, 3, 4, 5]` → 중앙값: 3
  - 데이터 (정렬됨): `[1, 2, 3, 4, 100]` → 중앙값: 3
  - '100'이 아니라 '1000'이 되어도 중앙값은 여전히 3입니다.

5.1.3 평균 vs 중앙값: 왜도 (Skewness)

평균과 중앙값의 차이는 데이터 분포의 '비대칭성(왜도)'을 알려줍니다.

- **대칭 분포 (Symmetric):** 평균 ≈ 중앙값 (예: 정규분포)
- **오른쪽 꼬리 분포 (Right-skewed): 평균 > 중앙값**
  - 소수의 매우 큰 값(outlier)이 평균을 오른쪽으로 끌어당깁니다.
  - 예: 개인 소득 분포. (대부분은 중간 소득, 소수의 재벌이 평균을 높임)
- **왼쪽 꼬리 분포 (Left-skewed): 평균 < 중앙값**
  - 소수의 매우 작은 값이 평균을 왼쪽으로 끌어당깁니다.

[이미지 플레이스홀더: 오른쪽 꼬리 분포(*Right-skewed*) 그래프]

대부분의 데이터가 왼쪽에 몰려있고, 긴 꼬리가 오른쪽으로 뻗어 있음.
(중앙값)이 (평균)보다 왼쪽에 위치함.

5.1.4 최빈값 (Mode)

- **정의:** 데이터에서 가장 '자주' 등장하는 값.
- **용도: 범주형 데이터**의 중심을 나타낼 때 사용합니다.
- **예시:** (선호하는 애완동물) `["개", "고양이", "개", "쥐", "고양이", "개"]` → 최빈값: "개"
- 범주형 데이터는 순서가 없으므로 평균이나 중앙값을 계산하는 것이 의미가 없습니다.

□ 예제: 어느 것이 더 빠를까? (Mean vs Median)

질문: 수십억 개의 데이터가 있을 때, 평균과 중앙값 중 무엇이 더 빠를까요?
답변: 평균이 훨씬 빠릅니다.

- **평균 ($O(n)$):** 데이터를 한 번만 훑으면서 합계와 개수만 알면 됩니다.
- **중앙값 ($O(n \log n)$):** '중간' 값을 찾으려면, 먼저 모든 데이터를 **정렬(Sorting)**해야 합니다. 정렬은 매우 비싼 연산입니다.

이러한 연산 속도 차이도 평균이 자주 사용되는 이유 중 하나입니다.

5.2 데이터의 "퍼짐" 측정 (Measures of Spread)

데이터가 중심에 밀집해 있는지, 아니면 넓게 퍼져있는지 측정합니다.

5.2.1 범위 (Range)

- **정의:** 최대값(Max) - 최소값(Min).
- **단점:** 데이터 양 끝의 극단값 2개에만 의존하므로, 분포 전체의 퍼짐을 잘 설명하지 못합니다.

5.2.2 분산 (Variance)

- **정의:** 데이터가 평균($\bar{x}$)으로부터 '평균적으로 얼마나 멀리 떨어져 있는지'를 나타내는 값.
- **계산 (직관):**
  1. 각 데이터가 평균과 얼마나 차이 나는지(편차: $x_i - \bar{x}$) 계산.
  2. 편차를 '제곱'함 (음수를 없애고, 멀리 떨어진 값에 더 큰 가중치를 주기 위해).
  3. 제곱한 값들의 평균을 냄.
- **공식 (표본 분산):** $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
- **단점:** 값을 '제곱'했기 때문에, 원래 데이터와 단위가 맞지 않습니다. (예: 키(cm)의 분산은 cm^2가 됨)

□ **예제: Q&A: 왜 n이 아니라 $n-1$로 나누나요?**

질문: 평균은 n으로 나누는데, 왜 분산은 $n-1$로 나누나요?
답변 (직관): "퍼짐(spread)"을 측정하려면 최소 몇 개의 데이터가 필요할까요? 만약 데이터가 1개([5])만 있다면, 이 데이터가 얼마나 퍼져있는지 말할 수 없습니다. 분산 공식의 분모에 $n=1$을 넣어보면 $\frac{1}{1-1}=\frac{1}{0}$이 되어 정의되지 않습니다. 즉, 이 공식은 "분산을 계산하려면 최소 2개 이상의 데이터가 필요하다"는 직관을 반영하고 있습니다.
답변 (기술): 우리가 가진 '표본'의 분산(s^2)을 가지고 '모집단'의 분산(σ^2)을 추정할 때, n으로 나누면 실제보다 분산이 작게 추정되는 경향(편향)이 생깁니다. $n-1$로 나누면 이 편향이 보정되어 모집단의 분산을 더 잘 추정할 수 있습니다. (통계 용어로 '자유도(Degrees of Freedom)'를 고려한 '불편추정량(unbiased estimator)'이라고 합니다.)

5.2.3 표준편차 (Standard Deviation)

- **정의:** 분산에 제곱근($\sqrt{}$)을 씌운 값.
- **공식:** $s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}$
- **장점:** 분산의 "단위 문제"를 해결합니다. **원래 데이터와 단위가 동일**해져 해석이 매우 직관적입니다.
- **직관:** "데이터가 평균으로부터 '평균적으로' 이 정도(s) 떨어져 있다."

6 절차/방법 2: 기본 시각화 (Basic Visualizations)

EDA의 꽃은 시각화입니다. 숫자는 우리를 속일 수 있지만, 그림은 그렇지 않습니다.

6.1 시각화의 중요성 (앤스컴 4중주)

주의사항

앤스컴의 4중주 (Anscombe's Quartet)는 "왜 통계 요약치만 보면 안 되는지"를 보여주는 고전적인 예시입니다.
여기 4개의 서로 다른 (X, Y) 데이터셋이 있습니다.

Dataset I		Dataset II		Dataset III		Dataset IV	
X	**Y**	**X**	**Y**	**X**	**Y**	**X**	**Y**
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
평균/분산		**평균/분산**		**평균/분산**		**평균/분산**	
X Avg: 9.0		X Avg: 9.0		X Avg: 9.0		X Avg: 9.0	
Y Avg: 7.50		Y Avg: 7.50		Y Avg: 7.50		Y Avg: 7.50	
X Var: 11.0		X Var: 11.0		X Var: 11.0		X Var: 11.0	
Y Var: 4.12		Y Var: 4.12		Y Var: 4.12		Y Var: 4.12	
상관계수: 0.816		상관계수: 0.816		상관계수: 0.816		상관계수: 0.816	

놀랍게도, 이 4개 데이터셋의 X/Y 평균, X/Y 분산, 상관계수가 모두 동일합니다. 숫자만 보면 이 4개 셋은 "똑같은" 데이터처럼 보입니다.

하지만 시각화하면 진실이 드러납니다.

[이미지 플레이스홀더: 앤스컴 *4*중주 산점도]
Dataset I: 점들이 완만한 선형 관계를 보임 (정상)
Dataset II: 점들이 위로 볼록한 *2*차 곡선(포물선) 모양을 보임 (비선형)
Dataset III: 거의 완벽한 직선 위에 있으나, *Y*축 방향의 이상치 *1*개가 존재함
Dataset IV: 모든 점이 *X=8*에 수직으로 있으나, *X*축 방향의 이상치 *1*개가 존재함

교훈: 절대 요약 통계치만 믿지 말고, 항상 데이터를 시각화해야 합니다.

6.2 기본 플롯 유형 비교

시각화는 "내가 무엇을 보고 싶은가?"에 따라 종류가 나뉩니다.

Table 2: 지저분한 데이터 예시 (*Messy Data*)

	Friday	**Saturday**	**Sunday**
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

- **히스토그램 vs. 막대 그래프:** 둘 다 막대를 사용하지만, 히스토그램은 **수량형** 변수를 '구간(bin)'으로 나눠 그리고 (막대들이 붙어있음), 막대 그래프는 **범주형** 변수의 '범주'별로 그립니다 (막대들이 떨어져있음).
- **파이 차트 (Pie Chart)는 왜 별로일까요?** 인간의 눈은 '각도'의 미세한 차이를 '길이'의 차이보다 훨씬 못 알아봅니다. 비슷한 비율을 비교할 때는 파이 차트보다 막대 그래프가 훨씬 효과적입니다.

6.3 3개 이상의 변수 시각화하기

3개 이상의 고차원 데이터를 2D 화면에 표현하는 것은 어렵습니다.

- **나쁜 예: 3D 산점도 (3D Scatter Plot)** 3개의 수량형 변수를 X, Y, Z축에 매핑하는 것은 그럴듯해 보이지만, 2D 모니터에서는 깊이감이 왜곡되어 "산점도 구름(scatter cloud)"처럼 보일 뿐, 관계 파악이 거의 불가능합니다.
- **좋은 예: 미적 매핑 (Aesthetic Mapping) 활용** X축과 Y축 외에, **색상(Color)**, **크기(Size)**, **모양(Shape)**, **애니메이션(Animation)** 등 추가적인 시각 요소를 사용하여 변수를 표현합니다.

□ 예제: 사례 연구: 갭마인더(Gapminder)의 5변수 시각화

한스 로슬링의 "Wealth & Health of Nations" 시각화는 5개의 변수를 하나의 차트에 훌륭하게 녹여냈습니다.

[이미지 플레이스홀더: 갭마인더 차트 (*1950*년 스냅샷)]

*X*축: 소득, *Y*축: 기대 수명. 점들이 분포해 있음.

이 차트는 다음 5가지 변수를 동시에 보여줍니다.

- **변수 1 (수량형):** 1인당 소득 → **X축 위치**
- **변수 2 (수량형):** 기대 수명 → **Y축 위치**
- **변수 3 (수량형):** 국가 인구 수 → **원의 크기 (Size)**
- **변수 4 (범주형):** 대륙 (아시아, 유럽...) → **원의 색상 (Color)**
- **변수 5 (시간형):** 연도 (1800 2020) → **애니메이션 (Animation)**

매핑 전략:

- **수량형 변수(인구) → 크기:** 크고 작음으로 양을 표현하기 좋음.
- **범주형 변수(대륙) → 색상:** 그룹을 구분하기 좋음.

7 부록 1: 데이터 시각화의 역사 (Historical Interlude)

데이터 시각화는 최근 기술이 아닌, 오래된 데이터 과학의 한 분야입니다.

존 스노 (John Snow, 1854) • **시각화:** 런던 콜레라 발병 지도

- **내용:** 콜레라 사망자 발생 위치를 지도에 점(dot)으로 찍었습니다.
- **결과:** 특정 '펌프'(Broad Street Pump) 주변에 사망자가 밀집된 것을 시각적으로 확인하고, 펌프를 폐쇄하여 전염병의 원인이 '오염된 물'임을 증명했습니다.

플로렌스 나이팅게일 (Florence Nightingale, 1858) • **시각화:** 로즈 차트 (Rose Chart / Coxcomb)

- **내용:** 크림 전쟁 당시 사망 원인을 월별로 시각화했습니다. (파란색: 예방 가능한 질병, 빨간색: 부상, 검은색: 기타)
- **결과:** 전투로 인한 사망(빨간색)보다, 열악한 위생으로 인한 질병 사망(파란색)이 압도적으로 많음을 보여주어 병원 위생 개혁을 이끌어냈습니다.

샤를 미나르 (Charles Minard, 1869) • **시각화:** 나폴레옹의 러시아 원정 지도

- **내용:** 단 하나의 차트에 나폴레옹 군대의 규모(선의 굵기), 이동 경로(지리), 방향(진격/후퇴), 시간, 그리고 후퇴 시의 기온 변화(하단 그래프)를 모두 담았습니다.
- **결과:** 42만 대군이 모스크바로 진격했다가 1만 명만 돌아오는 과정을 처참하게 보여주는, 데이터 시각화 역사상 최고의 걸작 중 하나로 꼽힙니다.

[이미지 플레이스홀더: 미나르의 나폴레옹 행군도]

8 부록 2: 효과적인 시각화 원칙 (Effective Visualization)

좋은 시각화를 만들기 위한 5가지 원칙입니다.

8.0.1 1. 그래픽 무결성 (Graphical Integrity)

"데이터로 거짓말을 하지 말아야 합니다."

□ 예제: 잘못된 예: 2020년 미국 대선 지도

- **지리적 면적 지도 (A):** 각 '카운티(County)'의 면적을 기준으로 승리한 정당(빨간색/파란색)을 칠합니다. → 결과: 미국 전역이 빨갛게 보입니다.
- **인구 기반 지도 (B):** 각 카운티의 '인구 수'에 비례하여 원의 크기를 조정한 점(dot) 지도를 만듭니다. → 결과: 인구가 밀집된 해안가와 도시에 파란색 점이 집중되고, 인구가 적은 중부 내륙에 빨간색 점이 흩어져 보입니다.
- **결론:** (A) 지도는 땅이 넓지만 인구가 적은 지역을 과대평가하여 "미국 대부분이 빨간색을 지지한다"는 **잘못된 인상**을 줍니다. (B) 지도가 실제 득표 수에 더 가까운 '무결성'을 가집니다.

8.0.2 2. 단순성 (Keep it simple)

"불필요한 장식을 피해야 합니다. (차트 정크 금지)"

- **차트 정크(Chart Junk):** 데이터 이해에 도움이 되지 않는 모든 시각적 요소를 말합니다.
- **나쁜 예:** 3D 효과가 들어간 막대 그래프, 현란한 배경색, 의미 없는 그림자, 지나치게 복잡한 범례.
- **좋은 예:** 데이터 잉크 비율(Data-Ink Ratio)을 높여, 꼭 필요한 선과 점, 텍스트만 남깁니다.

8.0.3 3. 올바른 표현 (Use the right display)

인간의 뇌가 정보를 더 효율적으로 처리하는 시각적 수단이 있습니다.

[이미지 플레이스홀더: 시각적 표현의 효율성 계층]
(가장 효율적 / 정량적) → ***1.* 위치 (*Position*)** (예: 산점도)
↓→ ***2.* 길이 (*Length*)** (예: 막대 그래프)
↓→ *3.* 기울기 (*Slope*)
↓→ *4.* 각도 (*Angle*) (예: 파이 차트)
↓→ *5.* 면적 (*Area*) (예: 버블 차트)
↓→ *6.* 강도 (*Intensity*) / 색상 (*Color*)
(가장 비효율적 / 범주형) → *7.* 모양 (*Shape*)

교훈: 같은 데이터라도 '면적'이나 '각도'로 표현하는 것보다, '위치'나 '길이'로 표현하는 것이 훨씬 더 정확한 비교를 가능하게 합니다. (이것이 파이 차트보다 막대 그래프가 나은 이유입니다.)

8.0.4 4. 전략적인 색상 사용 (Use color strategically)

색상은 강력하지만, 잘못 사용하면 혼란을 줍니다.

- **정성적 (Qualitative):** 범주를 구분할 때. (예: 대륙별 색상) **5 8개 이하**의 색상 사용을 권장합니다.
- **순차적 (Sequential):** 값이 낮음에서 높음으로 갈 때. (예: 연한 녹색 → 진한 녹색)
- **발산형 (Diverging):** 값이 '0'이나 '평균'을 기준으로 양쪽으로 갈라질 때. (예: 파란색 ← 흰색 → 빨간색)
- **주의 1: 무지개색(Rainbow Colormap) 금지!** 무지개색은 순서가 명확하지 않고(노란색이 녹색보다 높은가?), 특정 부분이 불필요하게 강조됩니다.
- **주의 2: 색맹/색약 고려 (Color Blindness)** 인구의 상당수가 적록색약입니다. 빨간색과 녹색을 동시에 사용한 비교는 피해야 합니다.

8.0.5 5. 청중 이해 (Know your audience)

시각화의 목적이 무엇인지, 청중이 무엇을 알고 싶어 하는지 알아야 합니다.

- **탐색적(Exploratory):** 스스로 데이터를 탐색하기 위한 (중립적인) 시각화. (예: 내부용 대시보드)
- **설명적(Explanatory):** 청중에게 특정 메시지나 주장을 전달하기 위한 (의견이 담긴) 시각화. (예: 신문 기사의 "이라크의 피의 대가" 그래프)

9 학습 체크리스트

이 강의를 올바르게 이해했는지 다음 질문에 답해보세요.

데이터(Data)와 데이텀(Datum)의 차이를 설명할 수 있는가?

API, RSS, 웹 스크래핑의 차이점과 스크래핑 시 윤리적 문제점을 아는가?

'정형 데이터(Tidy Data)'의 3가지 원칙 (1행=1관측치, 1열=1변수)을 아는가?

수량형 변수(이산형/연속형)와 범주형 변수(순서형/명목형)를 구분할 수 있는가?

모집단과 표본의 차이를 알고, '표본 편향'의 예시를 2가지 들 수 있는가?

평균과 중앙값의 차이를 설명하고, '오른쪽 꼬리 분포'에서 둘의 대소 관계(**평균 > 중앙값**)를 아는가?

분산(s^2) 대신 표준편차(s)를 주로 사용하는 이유(단위 문제)를 설명할 수 있는가?

분산을 계산할 때 n이 아닌 $n-1$로 나누는 직관적인 이유를 설명할 수 있는가?

앤스컴 4중주(Anscombe's Quartet)가 주는 교훈(**"항상 시각화하라"**)을 아는가?

히스토그램과 막대 그래프의 차이점(수량형 vs. 범주형)을 아는가?

파이 차트보다 막대 그래프가 권장되는 이유(각도 vs. 길이)를 아는가?

3개 이상의 변수를 시각화할 때 '미적 매핑'(색상, 크기 등)을 활용하는 법을 아는가?

'차트 정크'를 피하고, 색상을 전략적으로 사용해야 함을 이해했는가?

10 빠르게 훑어보기 (1-Page Summary)

title=**1. 데이터 수집 (Getting Data)**, colback=gray!10

- **API:** 공식적이고 안정적인 창구 (유료/제한 있음)
- **RSS:** 블로그/뉴스 스트림 (무료, 요약본)
- **웹 스크래핑:** HTML에서 직접 추출 (강력하지만 법적/윤리적 위험)

title=**2. 데이터 구조 (Data Structure)**, colback=gray!10 **정형 데이터 (Tidy Data)가 목표!**

- 1 행 = 1 관측치 (Observation)
- 1 열 = 1 변수 (Variable)
- 1 테이블 = 1 종류의 데이터

지저분한(*Messy*) 데이터는 이 원칙에 맞게 변형(*Tidying*)해야 함!

title=**3. 변수 유형 (Variable Types) - (중요!)**, colback=gray!10

- **수량형 (Quantitative):** 숫자로 연산 가능.
  - **이산형(Discrete):** 셀 수 있음 (예: 형제 수)
  - **연속형(Continuous):** 측정함 (예: 키)
- **범주형 (Categorical):** 그룹으로 구분.
  - **명목형(Nominal):** 순서 없음 (예: 애완동물 종류)
  - **순서형(Ordinal):** 순서 있음 (예: 학점)

title=**4. 기술 통계 (Descriptive Statistics)**, colback=gray!10

중심 (Center):
- **평균 (Mean):** 무게 중심. (이상치에 민감)
- **중앙값 (Median):** 순서상 중앙. (이상치에 둔감)
- **최빈값 (Mode):** 최고 빈도. (범주형 데이터용)

퍼짐 (Spread):
- **분산 (Variance):** 퍼진 정도 (단위가 ^{2} 됨)
- **표준편차 (Std Dev):** 퍼진 정도 (단위가 원본과 동일 $\rightarrow$ 해석 용이)

title=**5. 핵심 시각화 (Key Visualizations)**, colback=gray!10

앤스컴 4중주 (Anscombe's Quartet) 교훈: "숫자(통계)만 보지 말고, 항상 그래프를 그려라!"

히스토그램 (Histogram) 수량형 변수 1개의 **분포** 확인. (빈(bin) 너비에 민감)

막대 그래프 (Bar Plot) 범주형 변수 1개의 **빈도** 비교.

산점도 (Scatter Plot) 수량형 변수 2개의 **관계** 확인. (방향, 강도, 형태, 이상치)

박스 플롯 (Box Plot) & 바이올린 플롯 (Violin Plot) (범주형) 그룹 간 (수량형) 변수의 **분포 비교**. (바이올린이 더 많은 모양 정보 제공)

5변수 시각화 (Gapminder) X축, Y축, **크기(Size)**, **색상(Color)**, **애니메이션(Time)**