영등포구

영등포구 청년을 위한

생성형AI 활용 데이터 시각화 교육

데이터 시각화 개요

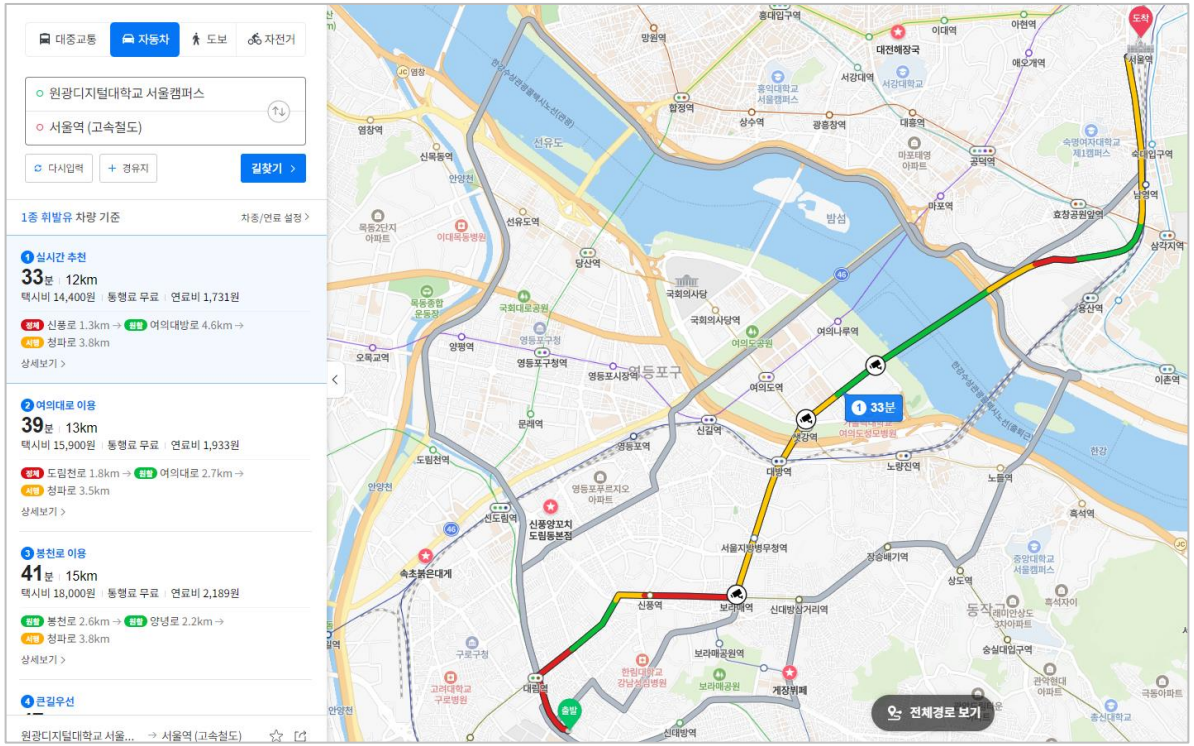
목 차

01 데이터 시각화

- 데이터 시각화
- 데이터 시각화 필요성
- 데이터 시각화 활용사례

데이터 시각화

- 데이터에서 발견한 정보를 시각적으로 이해할 수 있도록 그래픽 형태로 표현한 것
- 방대한 양의 데이터들을 살펴보는 것은 현실적으로 어렵기 때문에 데이터를 한 눈에 이해할 수 있도록 시각적 형식을 활용
- 데이터로부터 유의미한 인사이트를 도출하고, 더 나은 의사결정을 지원



데이터 시각화 필요성

- 1970년, 미국 성인의 상위 10%는 오늘날 달러로 약 135,000달러의 평균 수입을 받은 반면 하위 50%에 해당하는 사람들은 16,500달러를 받았습니다. 세계 불평등 데이터베이스에 따르면 지난 50년간 이 격차는 가파르게 늘어났는데, 상위 10% 계층의 소득은 350,000달러로 급격히 증가했고, 하위 50% 계층의 소득은 19,000달러로 소폭 증가했습니다.

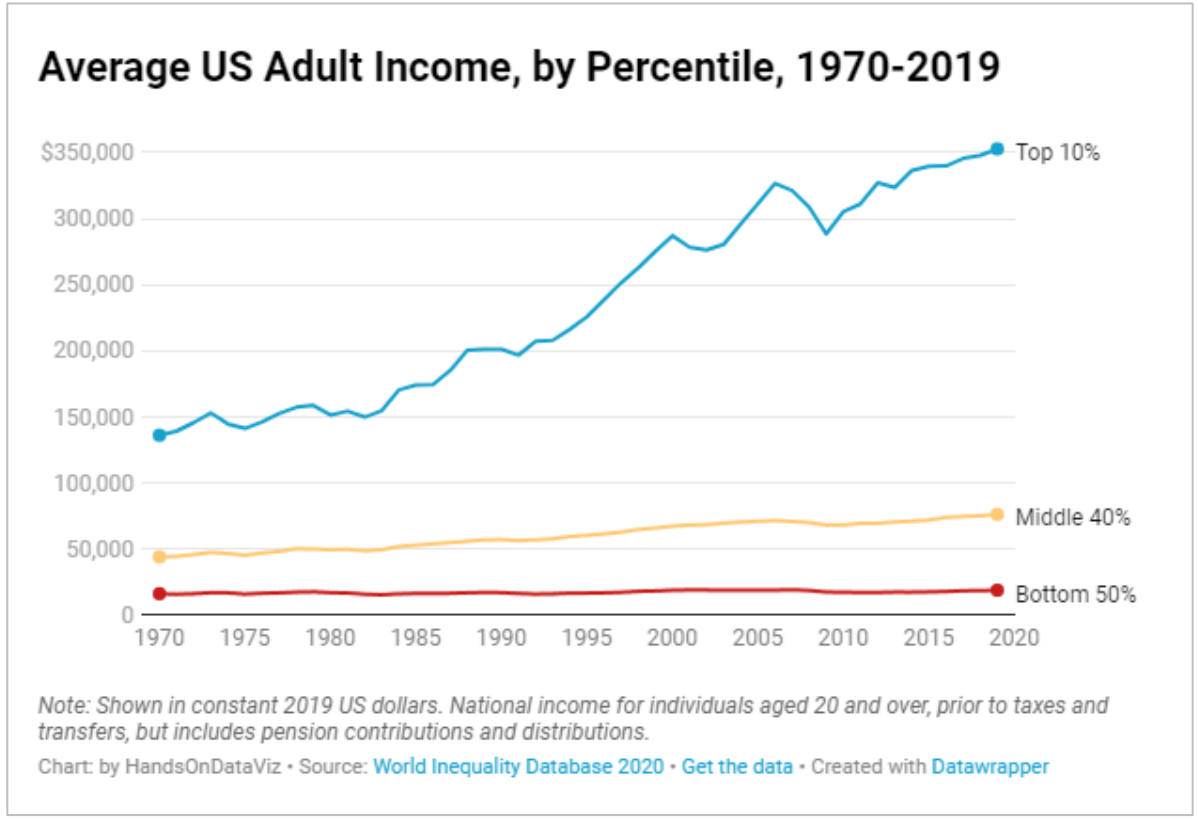
데이터 시각화 필요성

- 1970년, 미국 성인의 상위 10%는 오늘날 달러로 약 135,000달러의 평균 수입을 받은 반면 하위 50%에 해당하는 사람들은 16,500달러를 받았습니다. 세계 불평등 데이터베이스에 따르면 지난 50년간 이 격차는 가파르게 늘어났는데, 상위 10% 계층의 소득은 350,000달러로 급격히 증가했고, 하위 50% 계층의 소득은 19,000달러로 소폭 증가했습니다.

미국 소득 계층	1970	2019
상위 10%	$136,308	$352,815
중위 40%	$44,353	$76,462
하위 50%	$16,515	$19,117

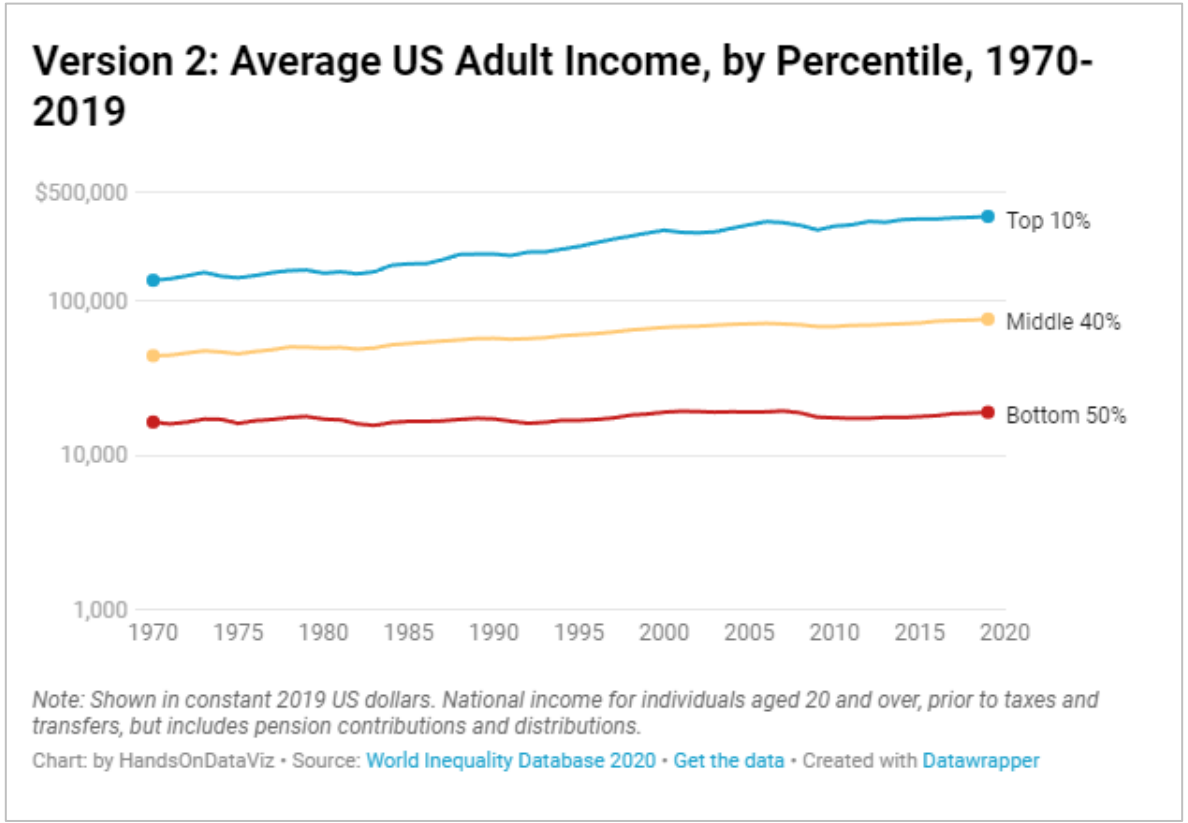
데이터 시각화 필요성

- 1970년, 미국 성인의 상위 10%는 오늘날 달러로 약 135,000달러의 평균 수입을 받은 반면 하위 50%에 해당하는 사람들은 16,500달러를 받았습니다. 세계 불평등 데이터베이스에 따르면 지난 50년간 이 격차는 가파르게 늘어났는데, 상위 10% 계층의 소득은 350,000달러로 급격히 증가했고, 하위 50% 계층의 소득은 19,000달러로 소폭 증가했습니다.



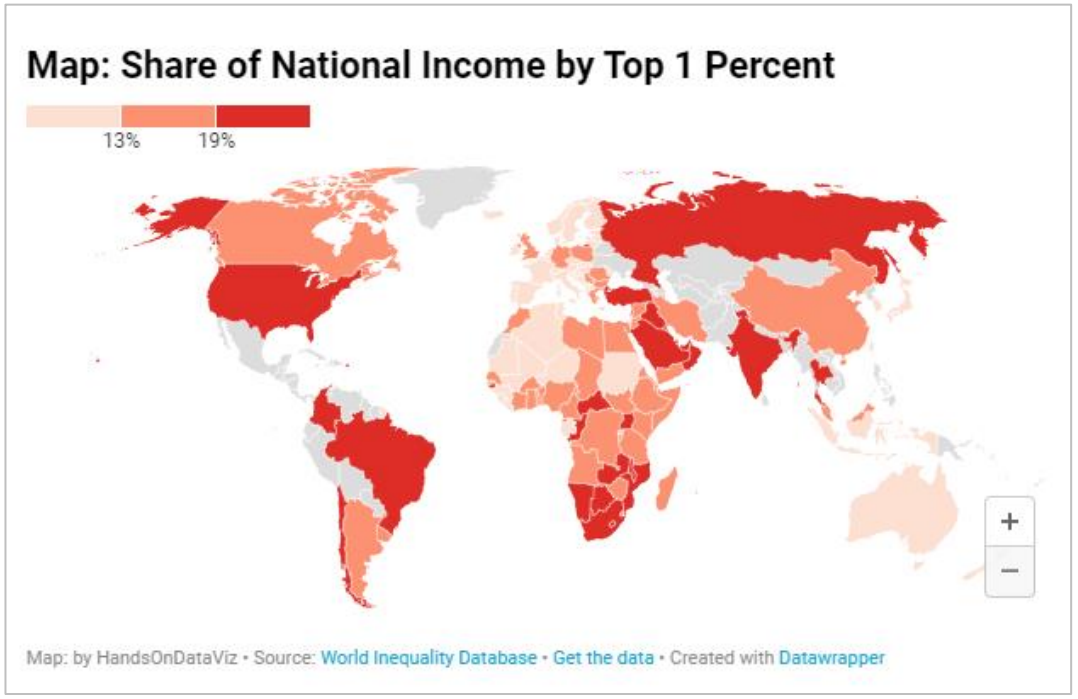
데이터 시각화 필요성

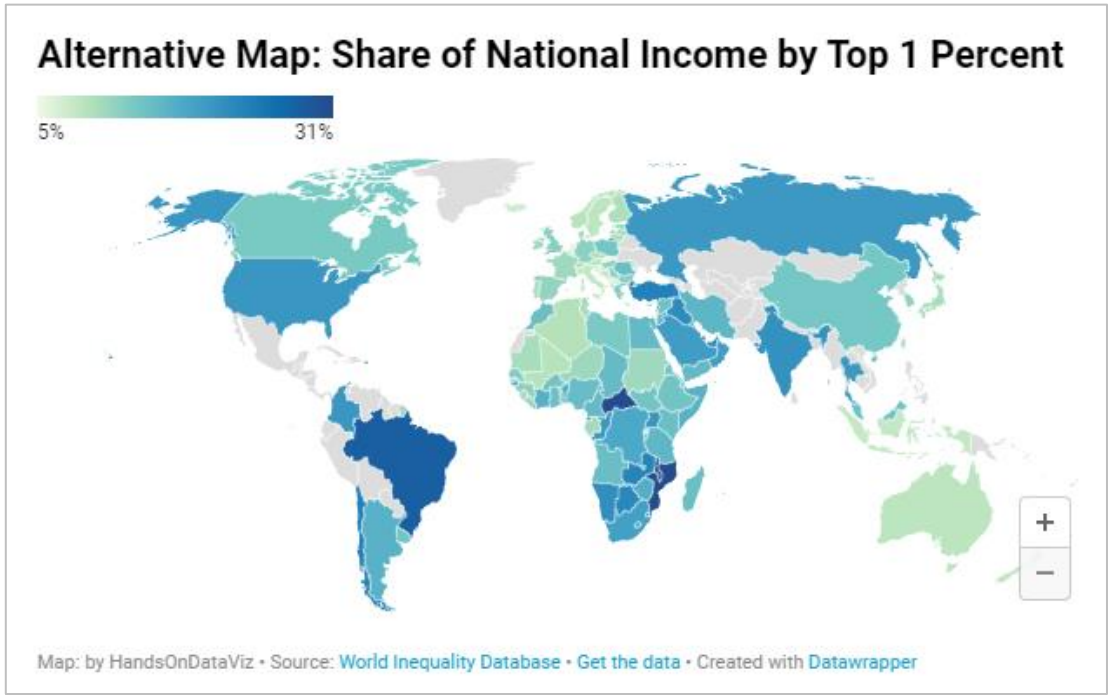
- 1970년, 미국 성인의 상위 10%는 오늘날 달러로 약 135,000달러의 평균 수입을 받은 반면 하위 50%에 해당하는 사람들은 16,500달러를 받았습니다. 세계 불평등 데이터베이스에 따르면 지난 50년간 이 격차는 가파르게 늘어났는데, 상위 10% 계층의 소득은 350,000달러로 급격히 증가했고, 하위 50% 계층의 소득은 19,000달러로 소폭 증가했습니다.



데이터 시각화 필요성

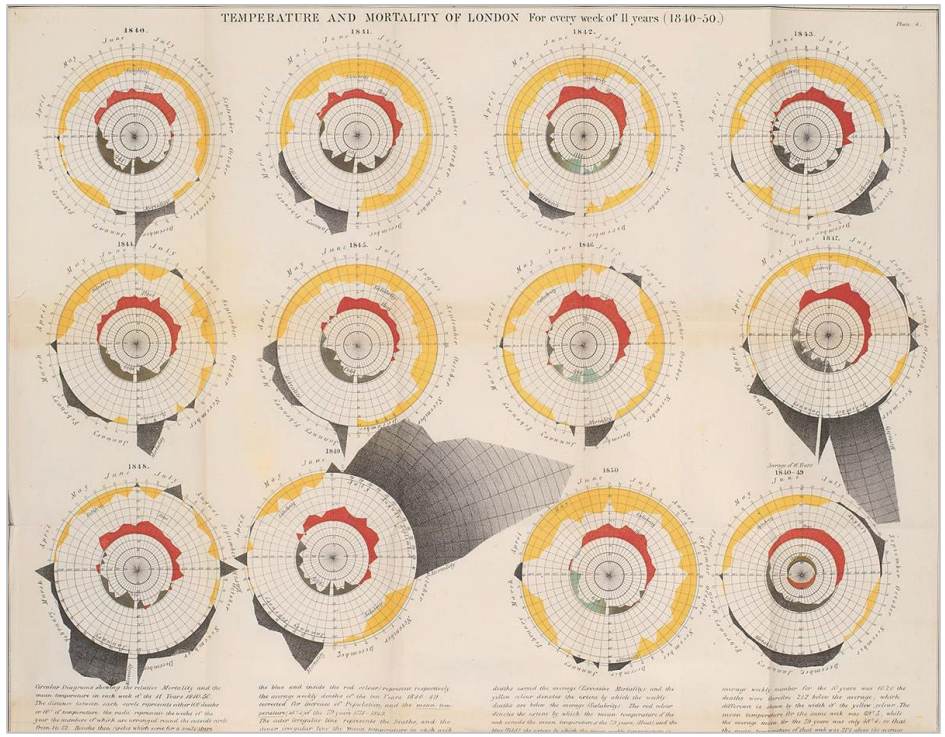
- 소득 상위 1%가 전체 국내 소득의 20%를 차지하고 있는 미국의 소득 불평등은 매우 심각한 상황입니다. 이에 반해 대부분의 유럽 국가에서는 상위 1%가 국내 소득의 6%에서 15%정도만 차지하고 있을 뿐입니다.

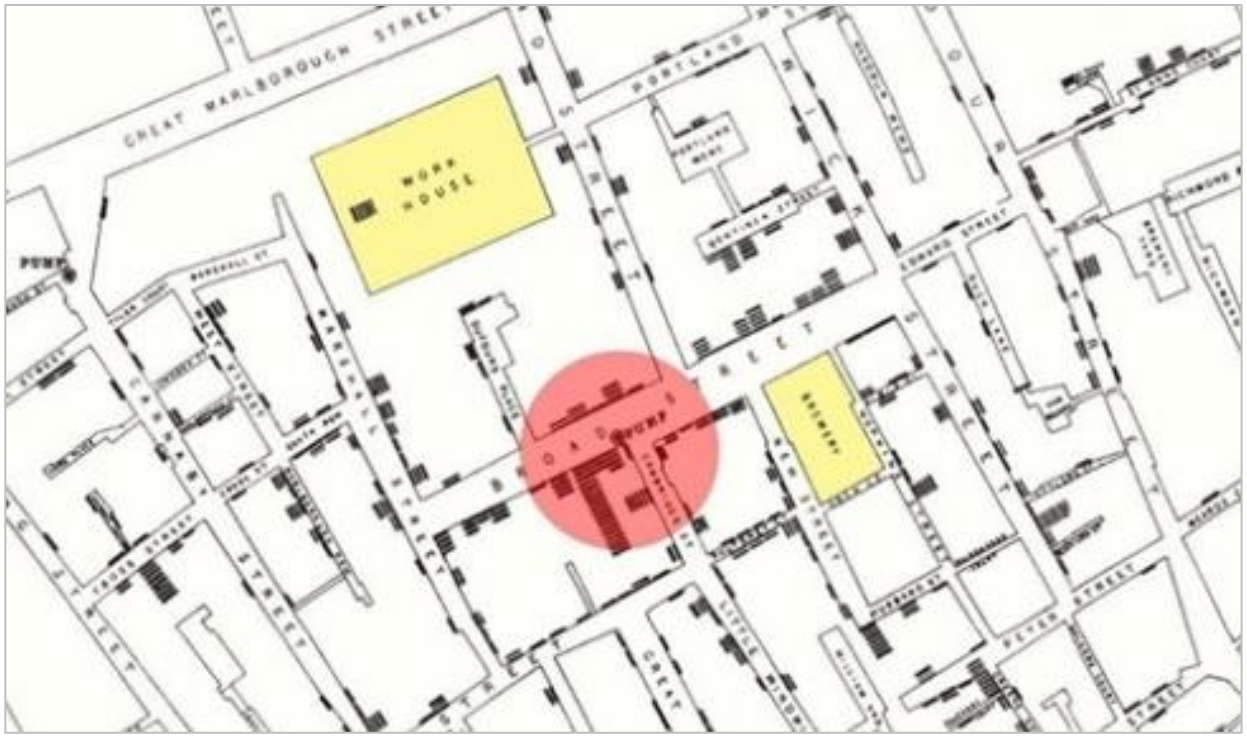




데이터 시각화 활용사례

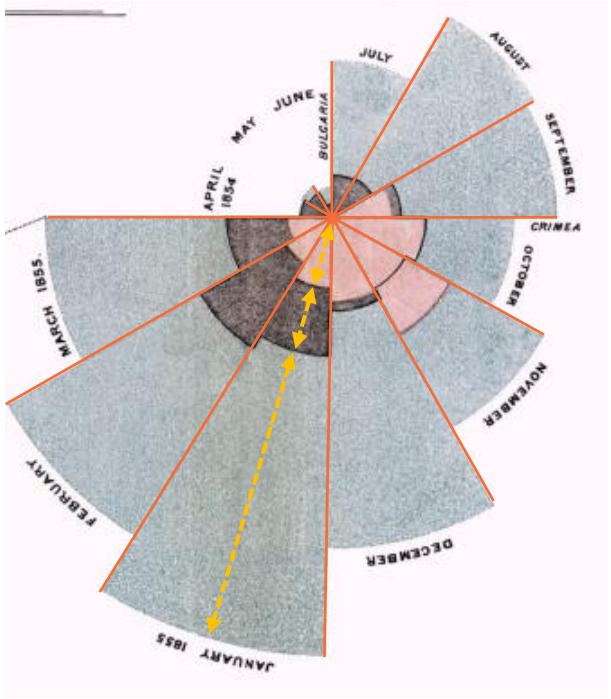
- **콜레라 역학지도(존 스노)**
  - 19세기까지 사람들은 콜레라도 다른 전염병처럼 공기를 통해 전염된다고 생각(윌리엄 파 – 기온과 관련된 그래프)
  - 콜레라 환자들을 일일이 방문해 사망 날짜와 장소를 지도에 표시
  - '브로드 스트리트의 식수 펌프' 를 중심으로 발병 장소와 사망자가 집중 사실 확인





데이터 시각화 활용사례

- **사망 원인 도표(플로렌스 나이팅 게일)**
  - 터키 병원에서 간호사로 일하면서 많은 영국인과 연합군이 죽어가는 것을 보며 원인 확인
  - 기존에 사용되던 파이 차트를 발전 시켜 “나이팅게일 로즈 차트“ 라고 불리는 새로운 시각화 고안
  - 시각화와 설명을 이용하여 의회를 설득시켜 조립식 병원을 건설하여 사망률을 급격히 떨어뜨림



12개월로 분할 후 길이로 빈도를 나타냄

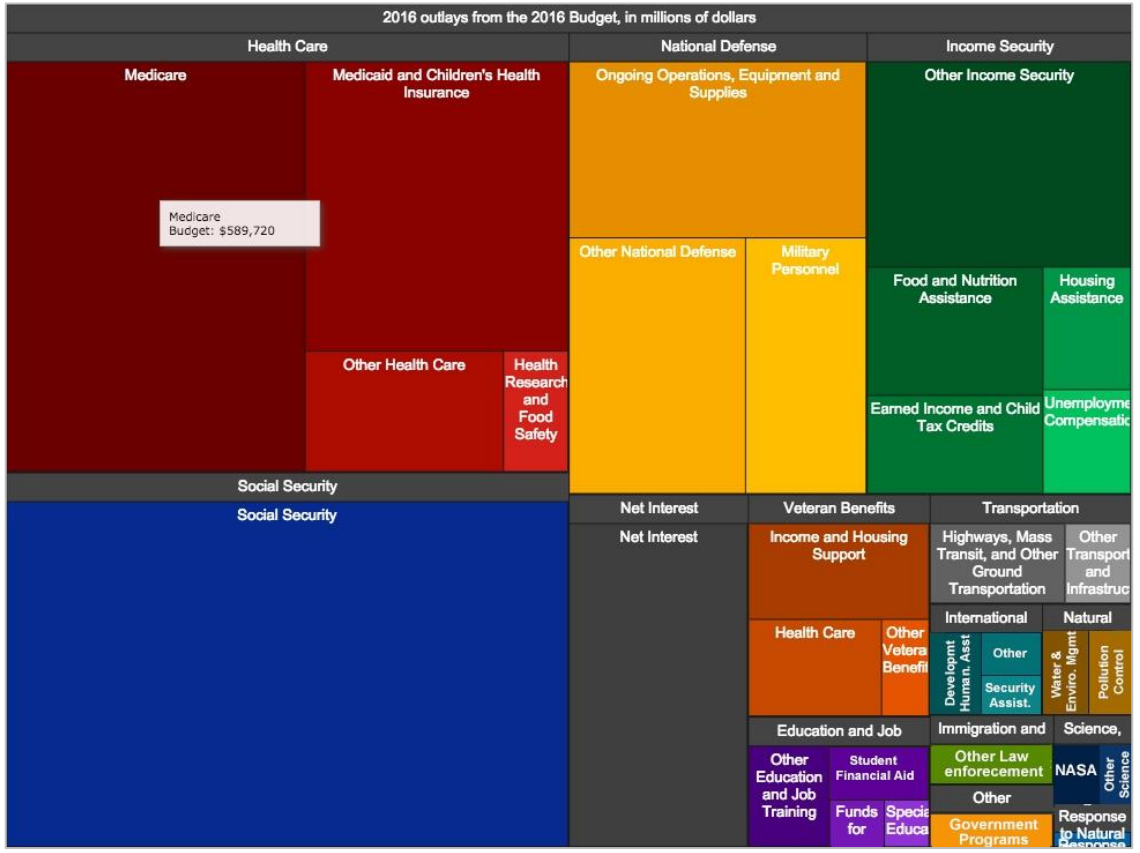
✓ 빨간색 : 부상

✓ 검은색 : 기타 원인

✓ 회색 : 예방 가능한 질병

데이터 시각화 활용사례

- **미국 예산 트리맵(미국 관리예산실)**
  - 미국의 2016년도 예산을 정부 프로그램 맥락에 맞도록 시각적으로 분류
  - 복잡하고 모호한 주제가 간단하고 명확한 시각화를 통해 설명



02 데이터의 이해

- 데이터
- DIKW 모델
- 데이터의 종류
- 데이터 대푯값
- 데이터 분석

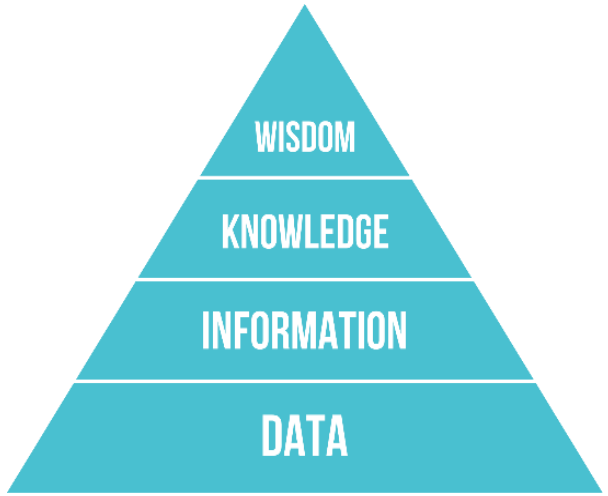
데이터(Data)

- **세상에 존재하는 모든 것은 데이터(Data)가 될 수 있음**
  - 이론을 세우는데 기초가 되는 사실. 또는 바탕이 되는 자료관찰이나 실험, 조사로 얻은 사실이나 자료
  - 컴퓨터가 처리 할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료

- **정보(Information)는 데이터(Data)를 가공 · 처리해서 얻을 수 있는 결과**

DIKW 모델

구분	내용
데이터(Data)	개별 데이터 자체로는 의미가 중요하지 않은 객관적 사실
정보(Information)	데이터를 가공, 처리하여 데이터 간의 연관 관계와 함께 의미가 도출 된 것
지식(Knowledge)	다양한 정보를 구조화 하여 유의미한 정보로 분류하고 대인적인 경험을 결합시켜 고유의 지식으로 내재화
지혜(Wisdom)	지식의 축적과 아이디어가 결합된 창의적인 산물



기존의 지식을 바탕으로 의사결정(action)

정보를 바탕으로 경험적 요소를 결합시켜 이해(experience)

기존 데이터를 보충하는 정보를 담고 있음(context)

객관적인 수치나 사실(data)

미래

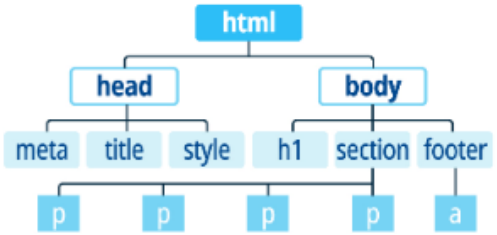
과거

데이터의 종류

- **데이터는 형태에 따라 종류를 나눔**

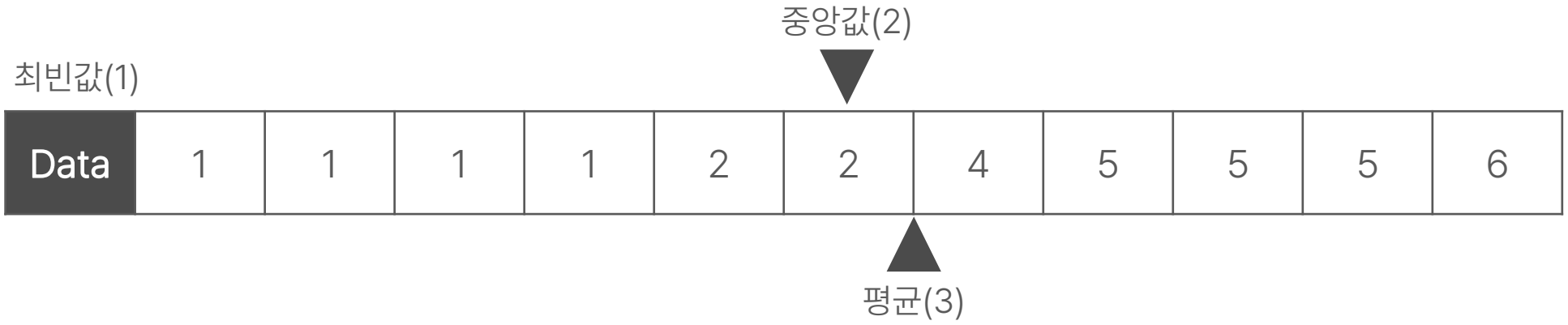
정형데이터(Structured Data)	반정형데이터(Semi-Structured Data)	비정형데이터(Unstructured Data)
테이블(table)에 행(row)과 열(column)으로 구조화	스키마가 정의되어 있음	구조가 일정하지 않음
이용하기 쉬움, 비용 낮음	보통	이용하기 어려움, 비용 높음
나이, 몸무게 등	XML, HTML, JSON 등	뉴스기사, 음악 등

ID	Name	AGE	SEX
01	KIM	32	M
02	LEE	26	F
03	PARK	72	F
04	CHOI	15	M



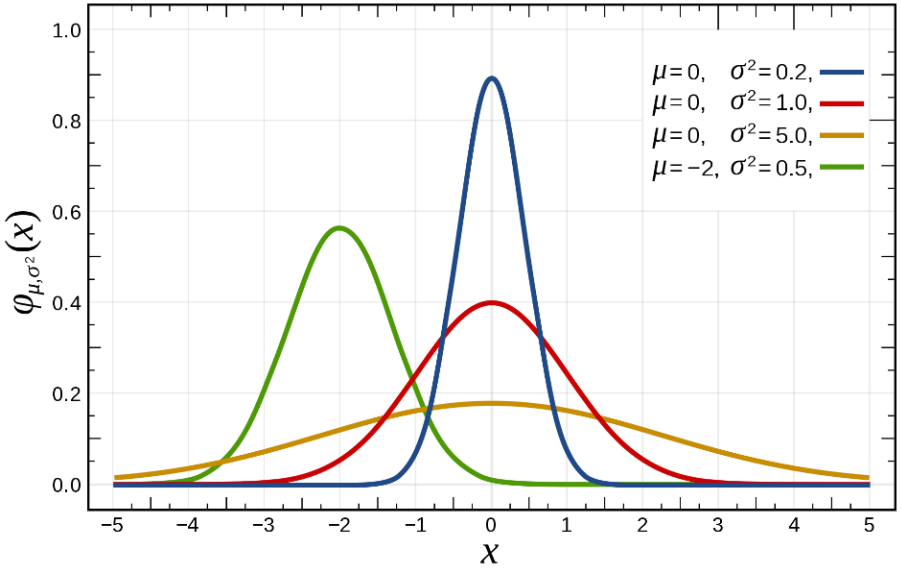
데이터 대푯값

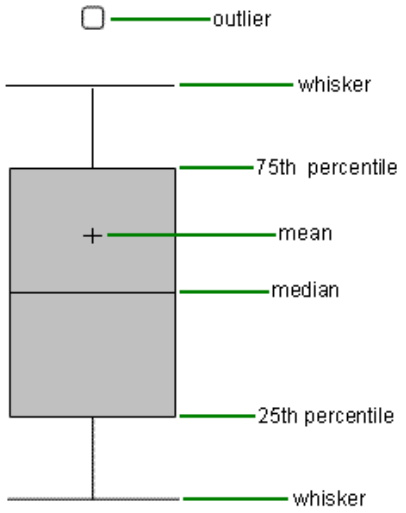
- **하나의 데이터로 전체를 대표하는 값**
  - 평균 : 전체의 수를 더하고 이를 데이터 수로 나눈 값. 데이터 전체의 중심에 해당. 이상치(특이값)에 약함
  - 중앙값 : 데이터를 작은 값부터 순서대로 나열 했을 때 한가운데 위치하는 값. 이상치(특이값)에 영향을 받지 않음
  - 최빈값 : 데이터 중에서 가장 많이 나타나는 값



데이터 대푯값

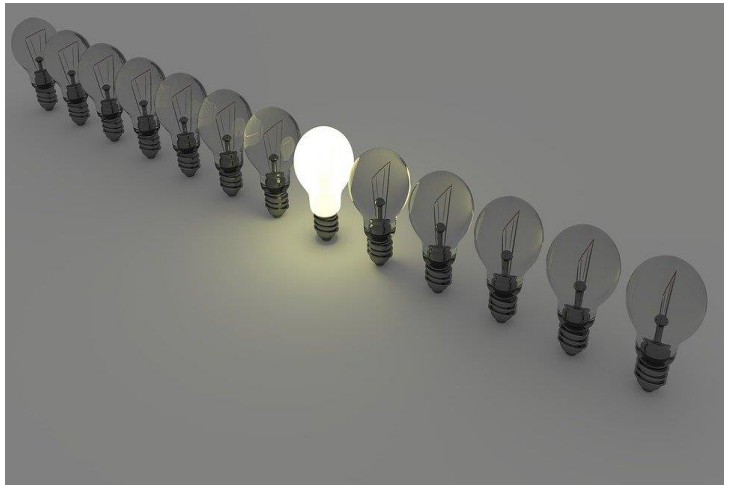
- **데이터 전체가 어떻게 퍼졌는지, 흩어짐은 어느정도 인지를 나타냄**
  - 분산(표준편차) : 데이터의 흩어짐 정도를 나타내는 값. 분산과 표준편차는 같은 내용
  - 사분위범위 : 중심 근처의 데이터 흩어짐 정도를 보는 지표
  - 범위 : 데이터가 위치하는 폭(최대-최소)을 나타내는 값





데이터 분석

- 데이터를 이해하고 원하는 정보를 뽑아내는 능력
- 유용한 정보를 찾기 위한 데이터를 조사하고, 정제하고, 변환하고, 모델링하며, 이 과정에서 도출한 결론을 의사 결정에 활용하는 전반적인 과정



데이터 분석

- **데이터를 이용하여 유용한 정보를 얻고, 이를 통해 의사결정에 필요한 통찰을 얻는 행위**
  - 분석가의 역량에 따라 데이터 분석의 결과가 바뀔 수 있음
  - 단순히 데이터를 탐색하는 것에 그치지 않고, 해당 결과의 실체를 꿰뚫어볼 수 있어야 함
  - 따라서, 분석기술의 역량보다 해당 데이터(도메인)에 대한 깊은 이해가 훨씬 중요

출처 : 가우스전자 시즌 3 441화 오독(https://comic.naver.com/webtoon/detail?titleId=675554&no=442&weekday=mon)

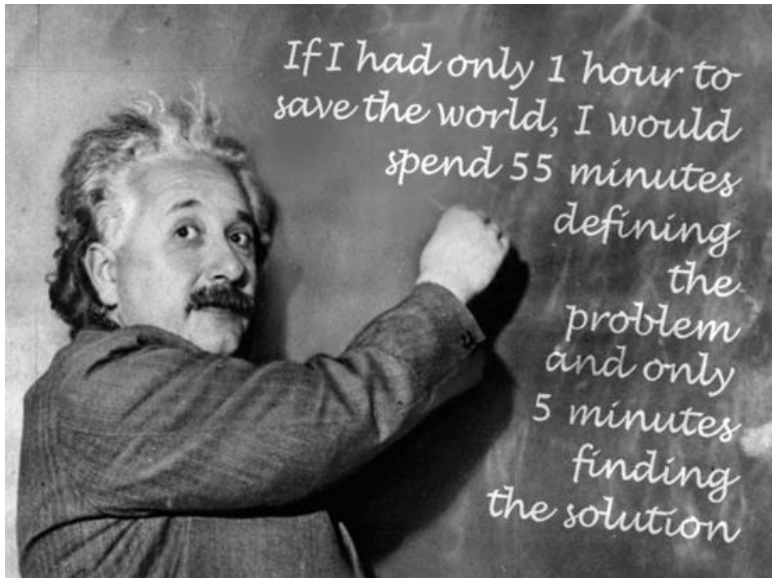
03 데이터 시각화 프로세스

- 문제 정의
- 데이터 수집
- 데이터 탐색 및 전처리
- 데이터 시각화

문제 정의

↗ 문제 정의

문제 정의 → 데이터 수집 → 데이터 탐색 및 전처리 → 데이터 시각화



나에게 1시간이 주어진다면,

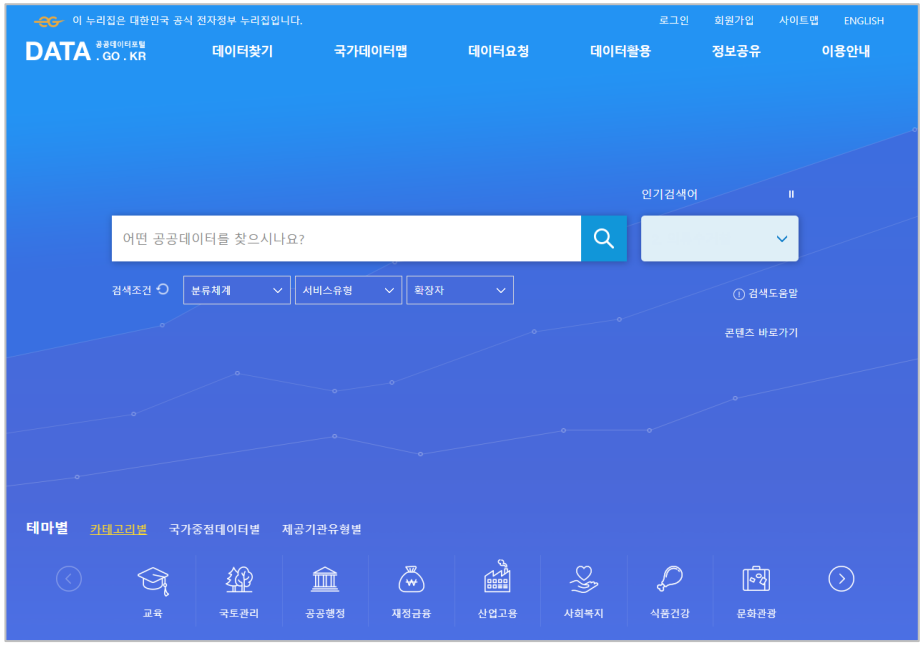
문제가 무엇인지 정의하는데 55분의 시간을 쓰고,

해결책을 찾는데 나머지 5분을 쓸 것이다.

- 알베트 아인슈타인

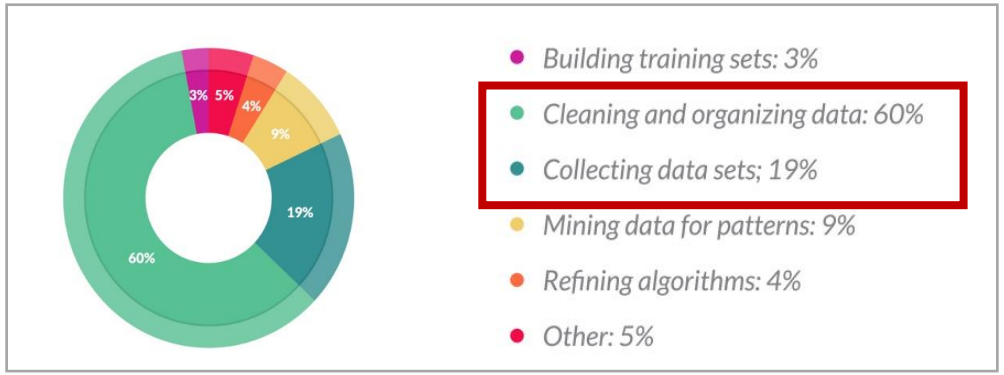
데이터 수집

- **어디서, 어떻게 데이터를 수집할 것인가?**
  - 온라인을 통한 데이터 수집(공공데이터, 크롤링, API 등)
  - 설문 조사 등을 통한 데이터 수집(구글 설문지, 오프라인 설문지 등)
  - 이미 가지고 있는 데이터(회사 내부 데이터, 고객 데이터 등)
  - 데이터 수집, 저장을 위한 개발이 필요



데이터 탐색 및 전처리

- **데이터를 탐색(EDA, Exploratory Data Analysis)하고 분석에 용이하게 변경**
  - 분석이 불가능할 정도로 지저분한 데이터를 처리하는 것
  - 머신러닝에 적용할 수 있도록 가공하는 것
  - 데이터 셋을 만들고 정제하는데 약 80%

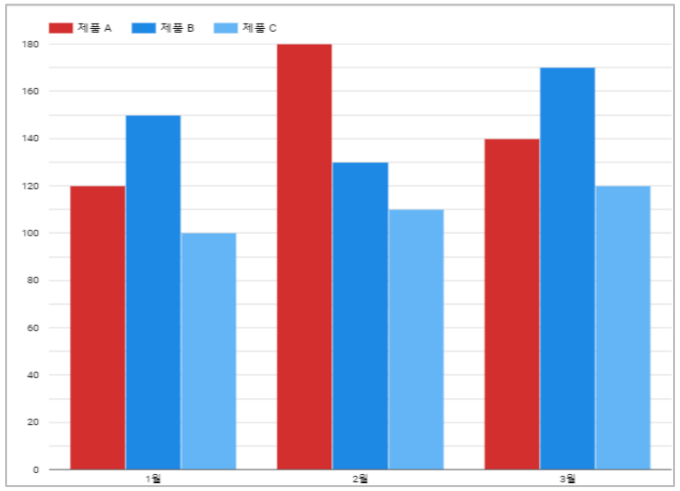


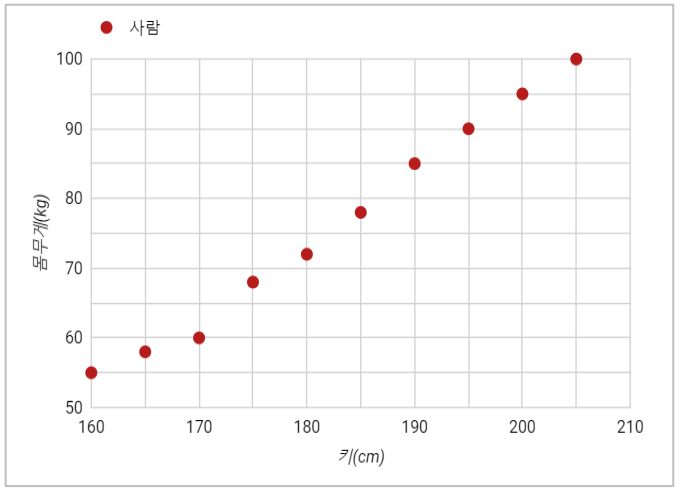
- **데이터 전처리의 중요성**
  - 다듬어지지 않은 데이터는 우수한 알고리즘의 성능을 저하시킴
  - 데이터 분석의 신뢰도를 떨어트림과 동시에 잘못된 결정 유도(비용 낭비)
  - GIGO(Garbage In Garbage Out) : 쓰레기가 들어가면 쓰레기가 나옴

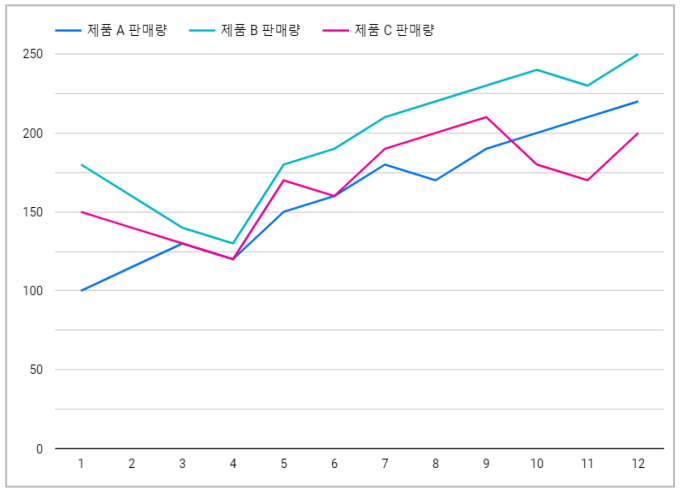
데이터 시각화

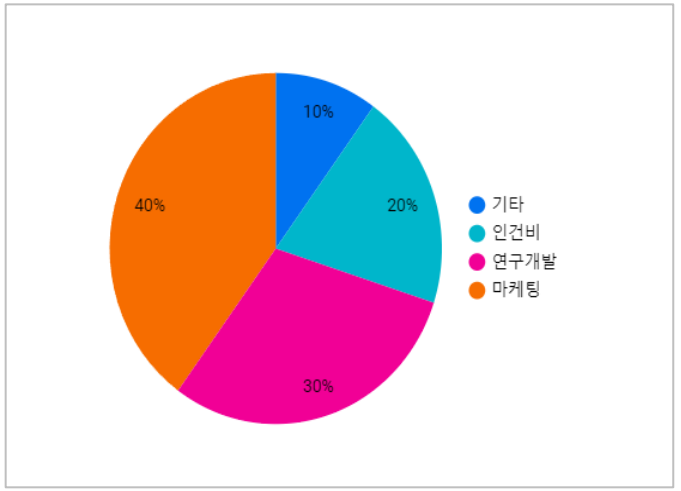
- **적절한 차트 선택하기**
  - 항목 간 상대 수치 비교하기 – 막대
  - 항목 간 관계 파악하기 – 산점도
  - 데이터 패턴 포착하기 – 꺾은선
  - 데이터 구성 요소 비율 파악하기 – 원형

누가 스듀디오









04 (실습) 노 코드 기반 데이터 시각화

감사합니다.