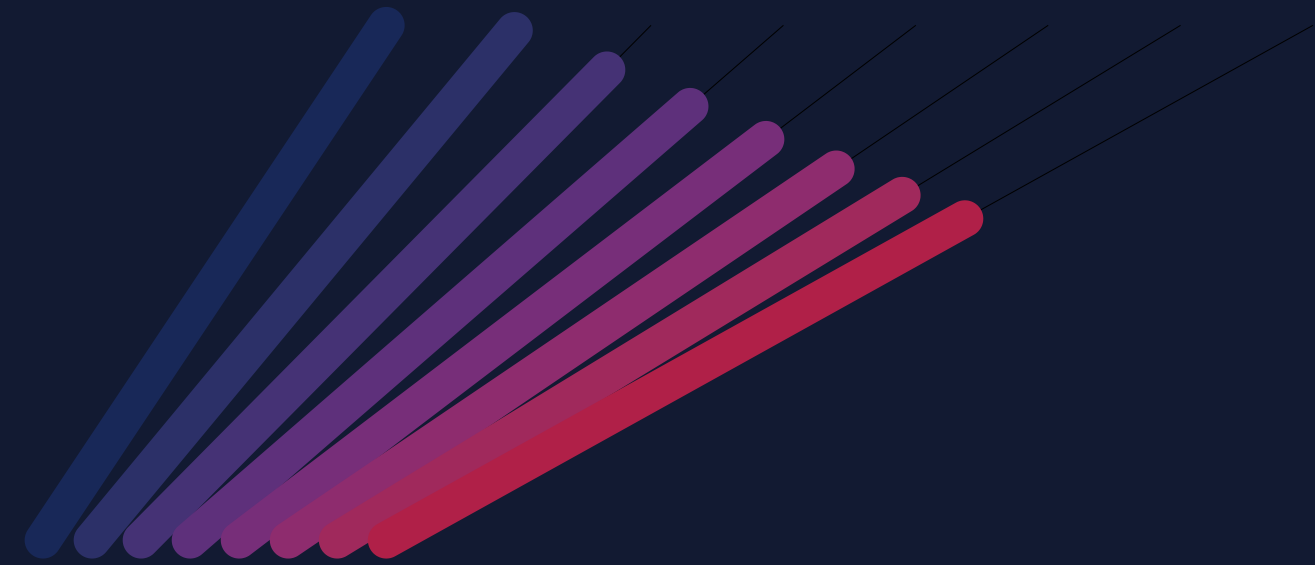




AX 디자인팩토리 부트캠프

청년도약 인재양성 부트캠프 · 온라인 과정

무엇으로 나눌 것인가 — 비교의 기준

3강 · 무엇으로 나눌 것인가 — 비교의 기준

과목
AI 데이터 분석

차시
3 / 5

길이
26분

도입

오늘 정하는 것

이 편이 끝나면 한 문장에 답할 수 있어야 합니다.

Q 나는 왜 이 분모를 골랐는지 설명할 수 있는가

답을 못 하면 방금 만든 순위표를 아무도 믿을 수 없습니다.

도입

이 편이 끝나면 보는 것

절대값 순위표 (서구·유성구·중구·동구·대덕구) + 인구 만명당 순위표 (중구·서구가 위 묶음, 동구는 그 사이쯤, 대덕구·유성구가 아래 묶음) 나란히 (재구성)

업소 수 - 정제 후 집계(04-2 규칙) · 인구 - 대전광역시 「대전의 통계」 <https://www.daejeon.go.kr/sta/index.do> · 확인일 2026-09-04

정규화와 분모 선택

이론

6분

소재 없이 - 어디서든 쓰는 이야기

큰 쪽이 항상 이기지 않으려면 같은 잣대로 나누어 비교한다.

분모를 고르는 기준은 그 질문이 누구 단위로 일어나는가입니다.

같은 표를 두 사람이 보면

절대값만 보는 사람

큰 쪽이 항상 이긴다

- 집단 크기가 다르다는 걸 안 본다
- 1등은 그냥 가장 큰 집단이다

분모로 나눠 보는 사람

밀도가 드러난다

- 질문에 맞는 공통 기준으로 나눈다
- 다른 분모로도 다시 계산해 순위가 바뀌는지 본다

분모를 고르는 절차

- 1 그 질문이 누구 단위로 일어나는지 묻는다
- 2 분모 후보를 나열한다 — 인구·세대 수·면적·사업체 수
- 3 질문에 가장 가까운 것을 고른다
- 4 다른 분모로도 계산해 순위가 바뀌는지 확인한다

대전 구별, 같은 데이터 다른 순위

시연

17분
정제한 표에서 시작

“주관식 응답”의 정체 — 999 사고 (재구성)

「지역경제활성화를 위한 노력」 열 - 우리 탐침 보고서는 「주관식 응답」이라 적었지만 파일에 그런 열 이름은 없다, 실제 고유값 35개 (선택형, 999 1 + 단일 6 + 조합 28) · 999 (무응답 코드) 17,638행 (55.8%) · 행당 평균 6.18개 열이 999 → 우리가 붙인 이름이 아니라 파일의 실제 값을 먼저 본다, 오늘 분모 계산엔 이 열을 쓰지 않는다

소상공인 설문조사 - author-c/probe-dj/3b/data/survey.csv (CP949, 31,635행) · 확인일 2026-09-04

절대값 → 만명당, 원표로 바꾸니 마 진이 두 배 넘게 움직인다 (재구성)

절대값 (2026-06, 정제 후 79,957행) 서구 27,509·유성구 18,949·중구 13,373·동구 11,727·대덕구 8,399 → 인구 원표 (2026-07) 만명당 중구 596.9·서구 592.3 (차이 4.6·0.8% - 「조금 앞선다」, 1위 헤드라인은 안 둠)·동구 537.8 (동구-대덕 차이 25.1·4.67% - 경계도 뚜렷하진 않음, 동구는 그 사이쯤)·대덕구 512.7·유성구 509.6 (대덕·유성 차이 3.1·0.6% - 「아래 묶음」, 꼴찌 단정 안 함) → 음식만 만명당 중구 175.7 (마진 8.1·4.61% - 「뚜렷한 편」)·유성구 167.6·대덕구 165.4 (둘 차이 2.2·1.3% - 「같은 묶음」)·서구 158.2·동구 154.6 (둘 차이 3.6·2.3% - 「같은 묶음」, 대덕구 165.4-서구 158.2 경계, 차이 7.2·4.35%도 뚜렷하진 않음) → 분모를 근사값에서 원표로 바꾸니 중구-서구 마진이 2.0→4.6, 두 배 넘게 움직였다 - 얇은 마진은 1위든 꼴찌든 똑같이 흔들린다 (마진% = (큰값-작은값)÷큰값×100, §0.18-a)

세대 표는 뺐다 + 25칸 타일 지도 + "구 별로 비교해 줘"의 함정 (재구성)

세대 수 원표를 못 구해 표는 만들지 않는다 - "출처 없는 값 다섯 개를 표에 올리는 것은 미확인보다 나쁘다" · "반려동물 용품점이었다면 분모는 세대 수였을 것" 말로만 → 지도 API · 경계 파일 없이 25칸 격자에 인구 만명당 값을 색으로 (위 묶음 중구·서구가 가장 진함, 동구는 그 사이쯤, 아래 묶음 대덕구·유성구가 가장 옅음) → AI에게 "구별로 비교해 줘"만 시키면 절대 값 표 (서구 27,509 · 유성구 18,949 등)로 응답, 분모가 필요한지 되묻지 않는다

분모를 밝히지 않은 순위는 순위가 아니다

정리

분모를 밝히지 않은 순위는 순위가 아닙니다.

다음 차시

'1등은 중구'라고 말하고 싶었는데, 숫자는 그렇게 말해 주지 않았습니다. 다음은 4-4 「상식과 반대로 나올 때」 입니다.

배웠는지 재는 질문 — 왜 그 분모입니까? 다른 분모로 하면 순위가 바뀔니까?

