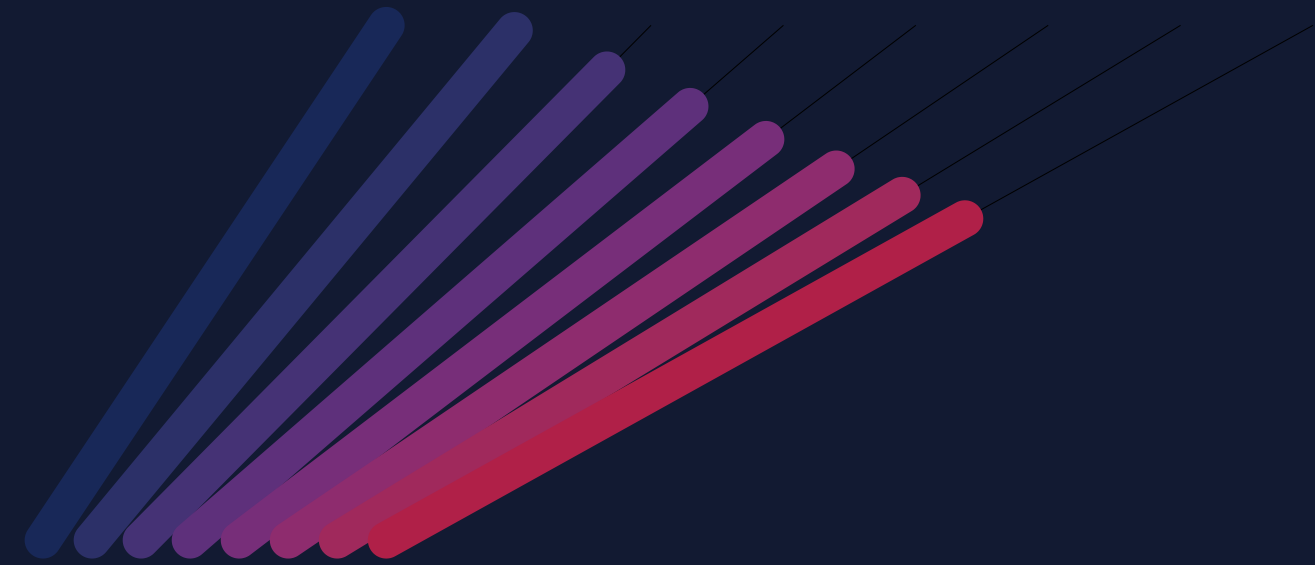




# AX 디자인팩토리 부트캠프

청년도약 인재양성 부트캠프 · 온라인 과정

# 상식과 반대로 나올 때



4강 · 상식과 반대로 나올 때

과목  
**AI 데이터 분석**

차시  
**4 / 5**

길이  
**29분**

도입

# 오늘 정하는 것

이 편이 끝나면 한 문장에 답할 수 있어야 합니다.

Q 상관과 인과, 잔차와 코호트를 구분해서 설명할 수 있는가

답을 못 하면 방금 만든 산점도가 무엇을 보여주는지 아무도 믿을 수 없습니다.

도입

# 이 편이 끝나면 보는 것

2019 코호트 2년 내 폐업률 표 (구별: 동구가 앞서는 편, 서구·중구·유성·대덕 순으로 뚜렷하게 낮아짐) + 밀도-폐업률 산점도 (재구성)

인허가업소 정보 — author-c/probe-dj/3b/data/license.csv(260,856행) · 밀도 — 04-3 인구 만명당 표 · 확인일 2026-09-05

# 비교의 정석 — 상관·잔차·코호트

이론

6분

소재 없이 — 어디서든 쓰는 이야기

# 같이 움직인다고 해서 하나가 다른 하나를 일으킨 것 은 아니다.

상관과 인과의 차이입니다. 둘 다 다른 원인 때문에 같이 움직였을 수 있습니다.

# 같은 상관계수를 본 두 사람

숫자만 보는 사람

## 상관계수 하나로 끝낸다

- 같이 움직이면 관계가 있다고만 본다
- 원인이 무엇인지 묻지 않는다

그림까지 그리는 사람

## 흠어진 모양을 눈으로 본다

- 같은 상관계수라도 그림은 다를 수 있다 — 앤스컴의 예
- 추세선에서 벗어난 잔차를 찾는다

# 두 표를 겹칠 때 확인할 것

- 1 상관이 있다고 인과까지 단정하지 않는다
- 2 같은 상관계수라도 그림을 실제로 그려 확인한다
- 3 추세선에서 크게 벗어난 잔차를 찾는다
- 4 비교하는 집단이 같은 출발선(코호트)에서, 같은 기간만큼 관측됐는지 확인한다

# 대전 인허가로 다시 세는 폐업률

시연

19분

코호트를 맞추지 않으면 벌어지는 일

# 더러운 원천 다시 — 이상치 13종과 「대전광역시」 행 (재구성)

인허가업소 정보 260,856행 × 17열 (CP949, 지번주소 따옴표 쉼표) → 인허가일자 이상값 열세 갈래 87행 (형식 자체가 낱자 아님 70행: 리터럴 null111  
64·2000000×3·42424·0·1997-8-78 / 형식은 맞지만 달력에 없음 16행: 2000-00-00×9·1990-05-00×2·1998-15-28·1992-04-31·1991-05-53·1989-12-00·1999-  
03-00 / 달력엔 있지만 파일 기준일 이후 1행: 2996-12-28)을 정의(달력 낱자로 해석 + 파일 기준일 이전)로 제거 → 「개방자치단체명」이 「대전광역시」로만 찍힌 3,731행 (구 없음,  
2019 코호트 내 109행)을 분자·분모 모두에서 제외

인허가업소 정보 — author-c/probe-dj/3b/data/license.csv(260,856행 × 17열) · 확인일 2026-09-05

# 코호트를 섞으면 그럴듯한 거짓이 나온다 (재구성)

2019~2021 인허가 45,887건을 한 분모에 넣고 폐업일자가 있는지만 보면 17.0%(창 없는 단순 비율, AI가 이 값을 "폐업률"이라 답함) → 코호트별로 나누면 2019년 27.6%·2020년 17.4%·2021년 8.1%(최근에 열수록 낮게 나올 뿐 안 망한 게 아님) → 창을 2년으로 고정해 다시 세면 2018년 12,295건 중 3,066건(24.9%)·2019년 13,665건 중 3,326건(24.3%)·2020년 15,218건(2022년 폐업 기록 없음, 비교 불가) → 2019 코호트 구별 동구28.5%(동구-서구 마진 5.96%, 헤지)·서구26.8%·중구23.5%·유성21.7%·대덕 18.3%(나머지 세 경계는 전부 6% 넘게 뚜렷)

인허가업소 정보 — author-c/probe-dj/3b/data/license.csv · 확인일 2026-09-05

# 산점도·잔차 — 잔차가 가장 큰 곳은 동구다 (재구성)

x=인구 만명당 업소 수 (04-3, 중구·서구는 마진 0.77%로 위 묶음 - 1위 단정 금지) · y=2019 코호트 2년 내 폐업률 → 추세선, 상관계수  $r=0.50$  (약한 양의 상관) → 잔차 절대값 순 동구+5.3·대덕-3.7·중구-2.5·서구+0.9·유성-0.1 → 밀도가 중간쯤인 동구가 다섯 구 중 가장 크게 벗어나 예상보다 훨씬 많이 문을 닫는다

밀도 - 04-3 인구 만명당 표 · 폐업률 - 위 코호트 표 (license.csv) · 확인일 2026-09-05

# “생존율 계산해 줘”의 함정 + 전공이 드러나는 곳 (재구성)

AI에게 "생존율 계산해 줘"만 시키면 창을 묻지 않고 2019~2021 인허가 45,887건을 그대로 뒤집어 생존율 83.0%로 답함(코드블록, 재구성) → "코호트를 말하지 않으면, 에이전트는 출발선이 다른 것을 같은 줄에 세웁니다" → 개방서비스명 일반음식점 60,899·휴게음식점 14,545·제과점영업 1,685으로 인허가 업종을 나눠 보는 인물 대사

인허가업소 정보 — author-c/probe-dj/3b/data/license.csv · 확인일 2026-09-05

# 출발선이 다른 것을 한 분모에 넣지 않는다

정리

# 출발선이 다른 것을 한 분모에 넣으면, 그럴듯한 거짓이 나옵니다.

다음 차시

다음은 4-5 「고르고, 틀릴 수 있는 곳을 적는다」 입니다.

배웠는지 재는 질문 — 이 폐업률의 분모는 몇 년 개업분이고, 몇 년까지 관측됐습니까?

