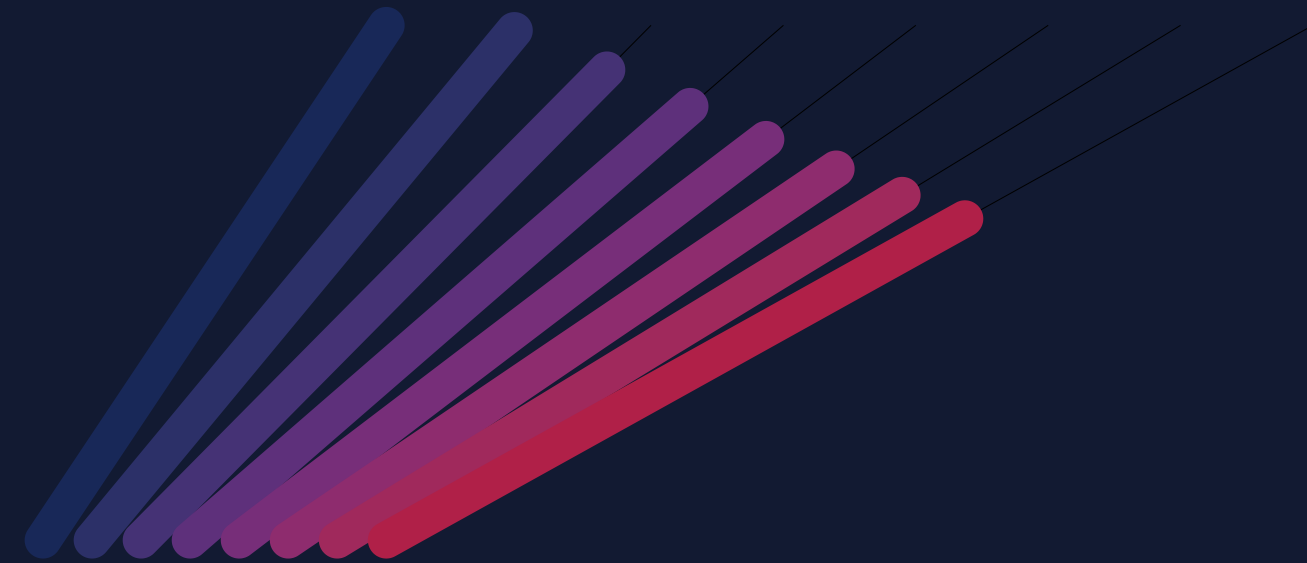




AX 디자인팩토리 부트캠프

청년도약 인재양성 부트캠프 · 온라인 과정

데이터를 구하고, 버린다

2강 · 데이터를 구하고, 버린다

과목	차시	길이
AI 데이터 분석	2 / 5	28분

도입

오늘 정하는 것

이 25분이 끝나면 한 문장에 답할 수 있어야 합니다.

Q 나는 몇 행을 버렸고, 왜 버렸는지 남에게 설명할 수 있는가

답을 못 하면 그 분석은 아무도 다시 만들 수 없습니다.

도입

25분 뒤에 보는 것

데이터 사전 13행 표 + 버린 행 기록 표 (규칙·키·탐지·제거·비율) (재구성)

정제와 데이터 사전

이론

6분

소재 없이 - 어디서든 쓰는 이야기

버리는 것은 쉽다. 왜 버렸는지 적는 것이 정제다.

규칙 없이 눈으로 골라내면 결과는 나오지만 남이 다시 만들 수 없습니다.

같은 파일을 두 사람이 열면

눈으로 고른 사람

결과는 나온다

- 이상해 보이는 행을 지운다
- 몇 행인지 모른다

규칙으로 고른 사람

결과에 근거가 붙는다

- 버릴 조건을 문장으로 쓴다
- 몇 행이 걸렸는지 센다

다섯 가지를 검사한다

- | | |
|---|--------------------------------------|
| 1 | 결측 — 빈칸이 있는가 |
| 2 | 이상치 — 정의를 벗어난 값이 있는가 |
| 3 | 중복 — 같은 행이 두 번 들어 있는가 |
| 4 | 타입과 단위 — 숫자여야 할 열에 문자가 섞였거나, 단위가 다른가 |
| 5 | 논리 모순 — 두 열을 같이 보면 앞뒤가 안 맞는가 |

대전 상가업소정보 원본을 연다

시연

18분
원본 규모부터 확인한다

사전 먼저 16열, 나중에 13열 (재구성)

헤더 39개 중 검사용 16개 (13개 남길 것 + 지점명·동정보·호정보 「검사 후 삭제 후보」 3개, 건물명은 검사하지 않으므로 사전에 안 올림)로 사전 작성 → 검사·규칙 뒤에 13개로 줄인다
(순서 rev9) → 음식 소분류 중 카페 3,836건 (비알코올 중분류) vs 빵/도넛 881건 (기타 간이 중분류) - 분류표는 다른 칸, 갈리는 건 인허가

다섯 검사, 실측 결과 (재구성)

표준산업분류명 빈칸 22 vs 지점명 빈칸 69,924(86.6%) · 중복은 「같은 가게」를 먼저 정의 - 상호+지번주소만 907그룹/1,862 중 225그룹(24.8%)이 소분류 상이(정육점·소고기 구이/찜, 가발 소매업·입시·교과학원도 같은 사례) → 「두 번」이라 부르면 넷 중 하나가 거짓 → 키에 소분류 추가 686그룹·탐지 1,411·제거 725 · 동정보·호정보 100% 빈칸 · 좌표 범위(정의 127.24~127.56/36.18~36.51) 밖 0 · 시군구-지번주소 불일치 0

시연 · 버릴 규칙

버린 행을 규칙으로 남긴다

규칙	키	탐지	제거	비율
표준산업분류명 결측	표준산업분류명	22행	22행	0.03%
중복(같으면 앞 행 남기고 뒤 행 버림)	상호+지번주소+소분류	1,411행(686그룹)	725행	0.90%
동정보·호정보(열 전체)	—	열 2개	열 2개	열 삭제

대전 상가업소정보 2026-06 원본 80,704행 · 오케스트레이터 실측 + 집필자 재확인(perl) · 확인일 2026-09-03 · 표 아래: 39열 중 23열 제외(목록 notes.md)

“결측치 처리해 줘”의 함정 (재구성)

AI 채팅창 - 지점명 결측 69,924건을 "본점"으로, 표준산업분류명 22건을 최빈값 "한식 일반 음식점업" (7,900건)으로 채움 → 없던 지점·분류가 생긴다

버린 행을 기록한다

정리

버린 행을 기록하지 않으면 남이 재현할 수 없습니다.

다음 차시

이제 셀 수 있습니다. 그런데 그냥 세면 서구가 이깁니다.

배웠는지 재는 질문 — 몇 행을 버렸고, 왜 버렸습니까? 그 규칙을 남이 그대로 적용할 수 있습니까?

