HEARTCOUNT 2023 월간 웨비나 시리즈

1. 하트카운트 시작하기 - 데이터셋 연동
2. 시계열 데이터 분석
3. 여러 차원을 한 화면에 시각화하기
4. 상관관계와 분포 - 개별 레코드 수준에서의 시각화 기법들
5. 드릴다운과 트리맵 - 전체를 구성하는 개별 요소들 확인하기
6. 분석하기 좋은 데이터셋을 구성하는 요소들
7. 특정 행동을 보인 집단의 특성을 이해하기
8. 회귀 분석을 통해 지표 차이의 요인 이해하기
9. Causal Inference: 상관관계로 인과성에 대해 이야기하는 법
10. LLM(거대언어모델)과 데이터 분석의 자동화
11. (tentative) 데이터 보고서 잘 쓰는 법
12. 월은 웨비나 쉽니다. 😊

양 승 준 / sidney.yang@idk2.co.kr

Causal Inference (인과적 추론)

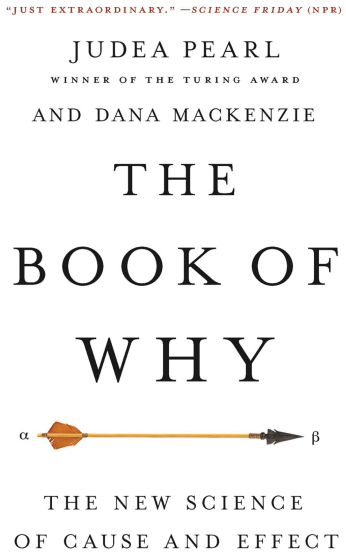
- Why Causality?
  - 실행(Intervention)을 위해 필요
  - 마스크 가격 두배로 올리면 판매량은?

- Statistics - “Causal Inference” 위해 발명 X
  - Statistics 101: Correlation is not Causation
  - 이론/도메인 지식 없이 데이터만으로 인과성 찾는 건 어려움

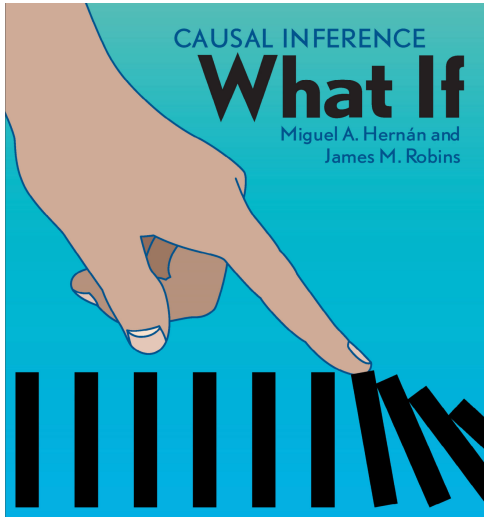
- 실험(RCT, A/B)없이 관측 데이터에서 인과적 패턴 찾기
  - Ceteris Paribus: 하나 빼고 다 같다면 그 하나가 원인
    - 도메인 지식에 기반한 “Causal Graphs”
    - “하나” 빼고 다른 걸 유사하게 하는 Conditioning 기법

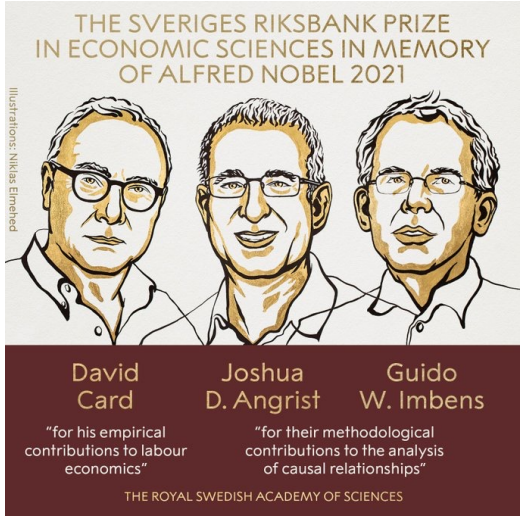
- 상관관계, 인과관계
- 상관관계 업무에 활용하기
- 상관관계로 인과성 주장하기

참고하면 좋은 내용들



https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

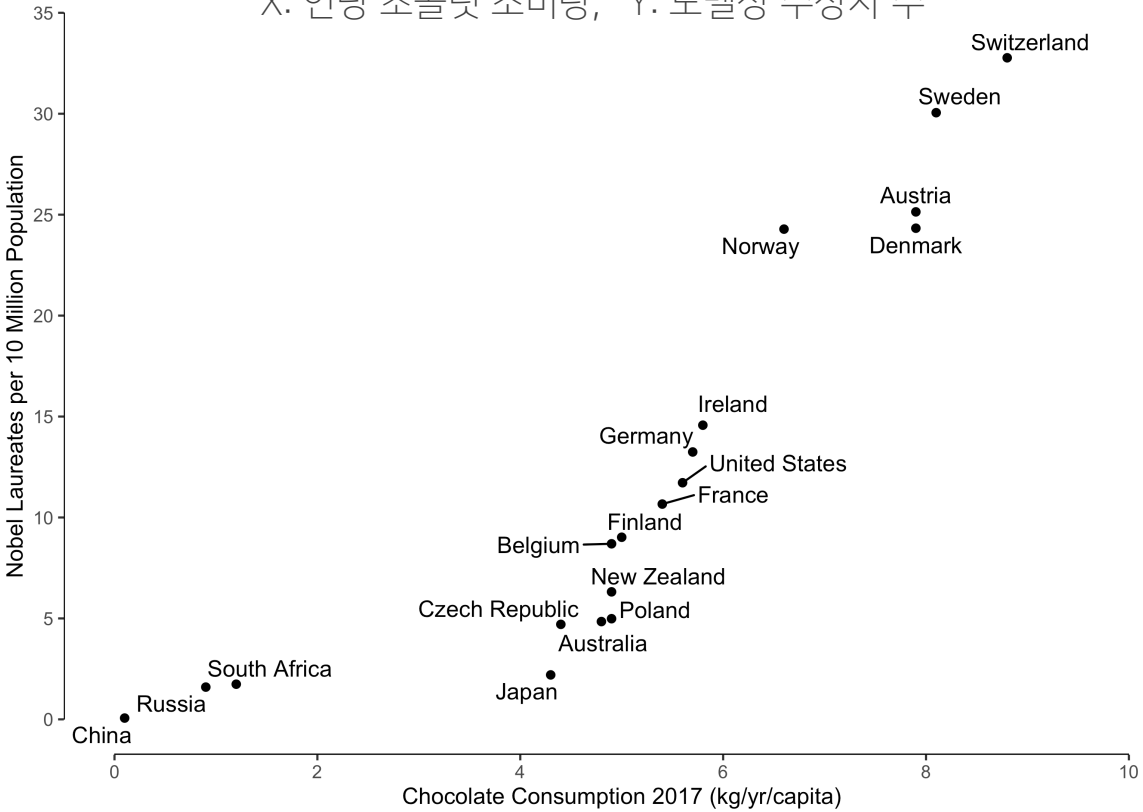




https://www.nobelprize.org/prizes/economic-sciences/2021/imbens/lecture/

Correlation, Cause and Effect

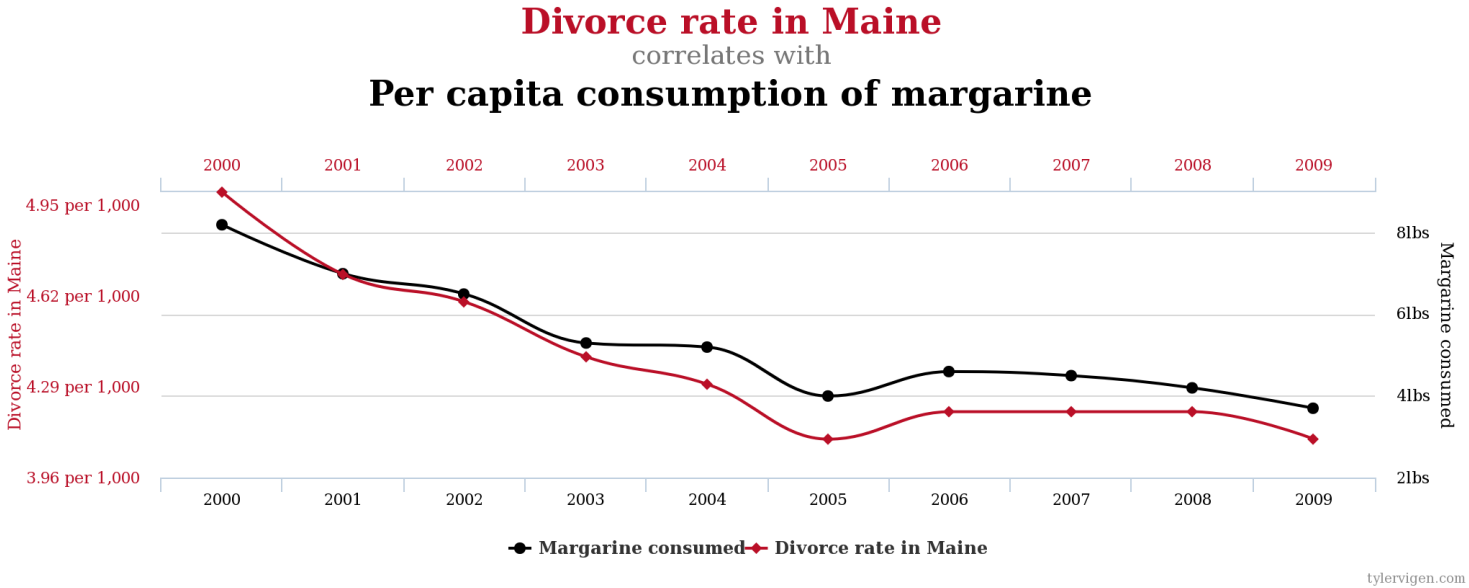
X: 인당 초콜릿 소비량, Y: 노벨상 수상자 수



1. X가 Y의 원인이거나
2. Y가 X의 원인이거나
3. Z가 X,Y의 원인이거나

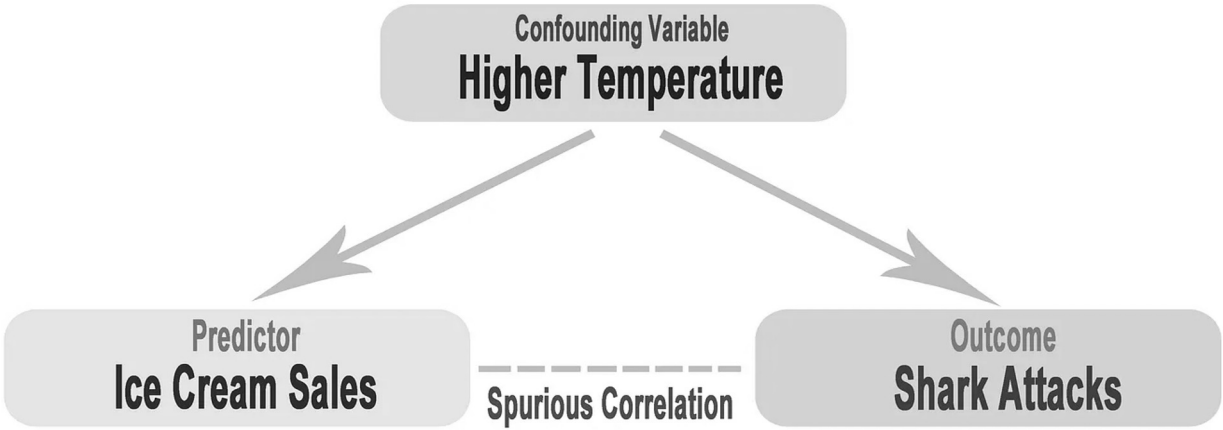
참고:
https://fabiandablander.com/r/Causal-Inference.html

Spurious Correlation (허위상관)
마요네즈 덜 먹으면 이혼을 덜 할까?



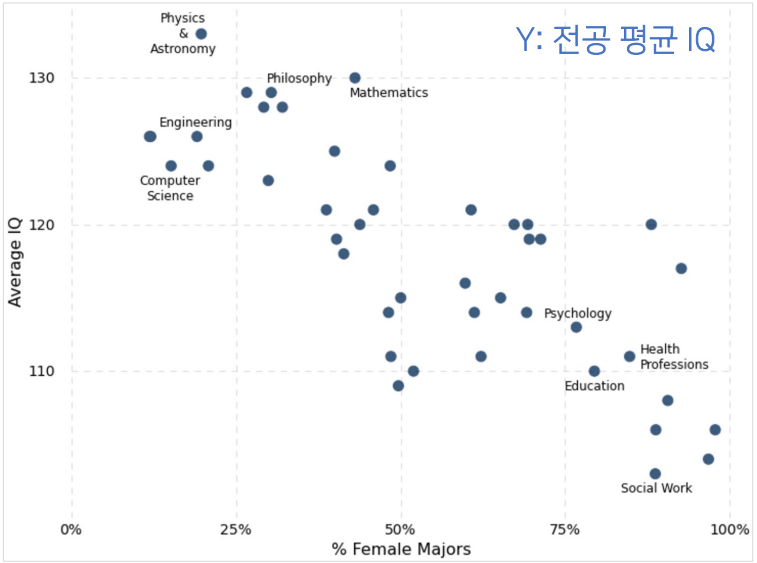
Correlation vs. Causation

Confounding Variable (교란변수)

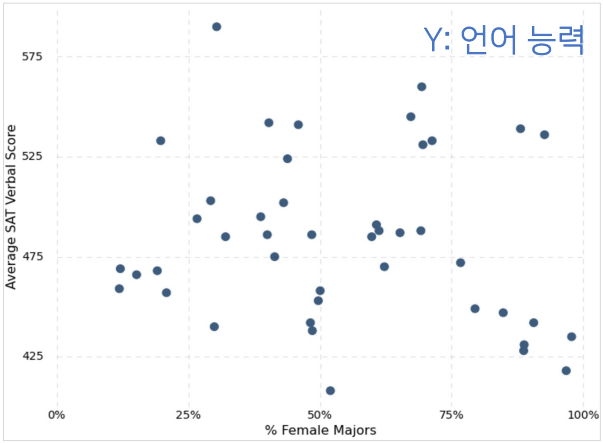


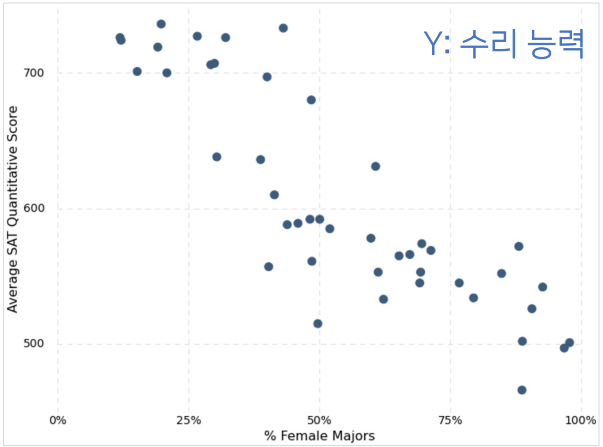
Seeing is NOT Always Believing

함께 움직이는 걸 보면
X가 Y의 원인이라 믿고 싶다.
IQ는 남녀 차이 없음. Then Why?



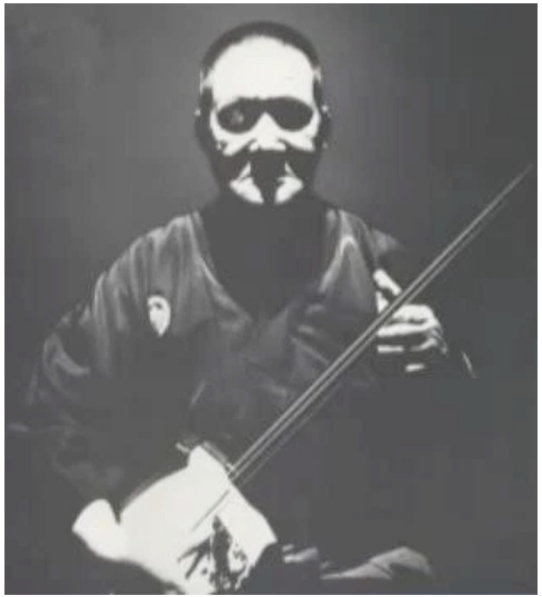
X: 여학생 비율





Correlation vs. Causation – Everything is Endogenous

바람이 많이 불면
나무통 가게가 돈을 번다.

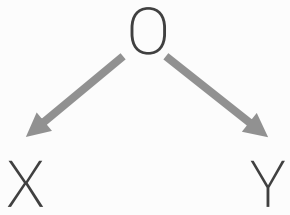


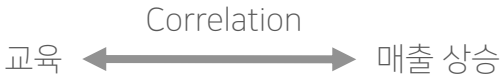
Correlation vs. Causation

$$P(Y \mid X) \neq P(Y \mid do(X))$$

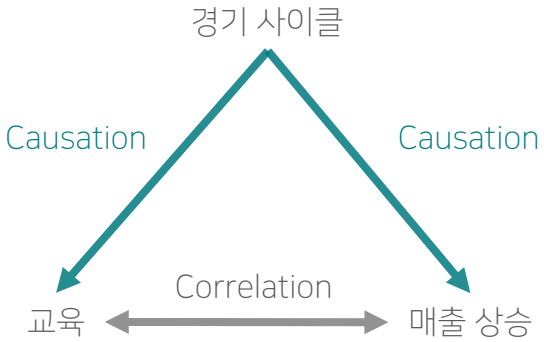
대표적인 오류

Omitted Variable Bias



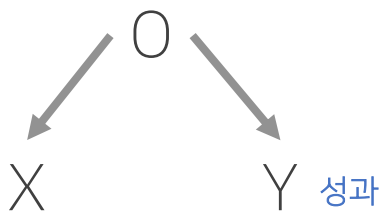


Correlation helps you predict the future
Causality lets you change the future



Two Types of Causality Problems

Omitted Variable Bias



- 교육 받아서 성과가 좋아 졌나?
- (경기 탓) 성과가 안 좋아서 교육 받았나?

Reverse Causality

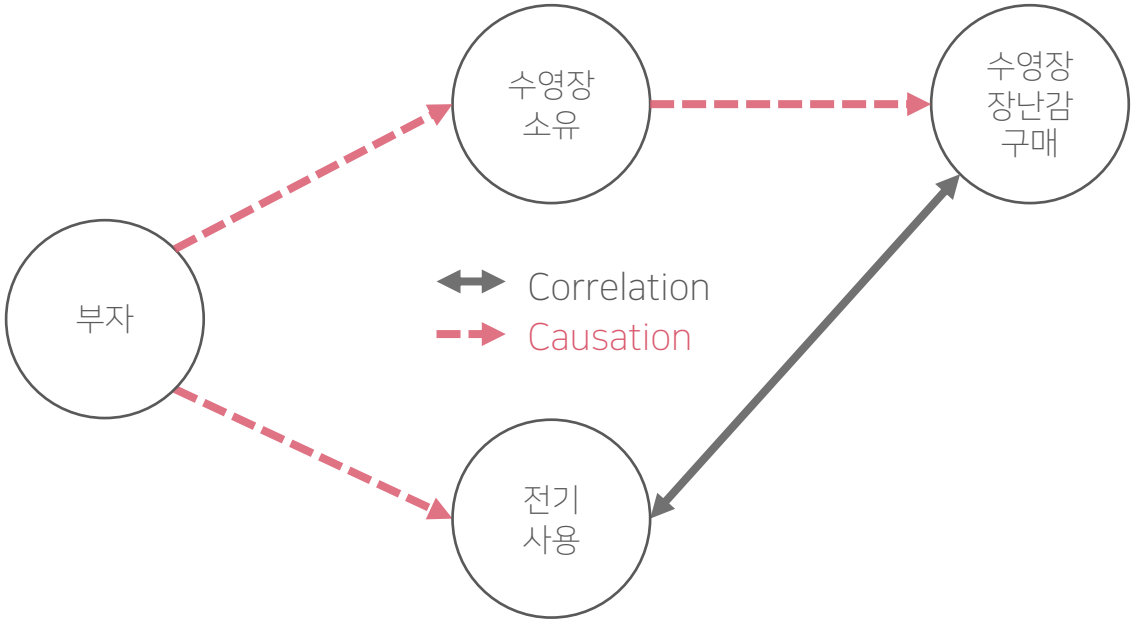
- 한 업무에 오래 있어서 성과가 낮나?
- 성과가 낮아서 한자리에 오래 있었나?

When Knowing Correlation is Enough

수영장 장난감 광고를 하려고 하는데
누가 수영장을 소유했는지 모름

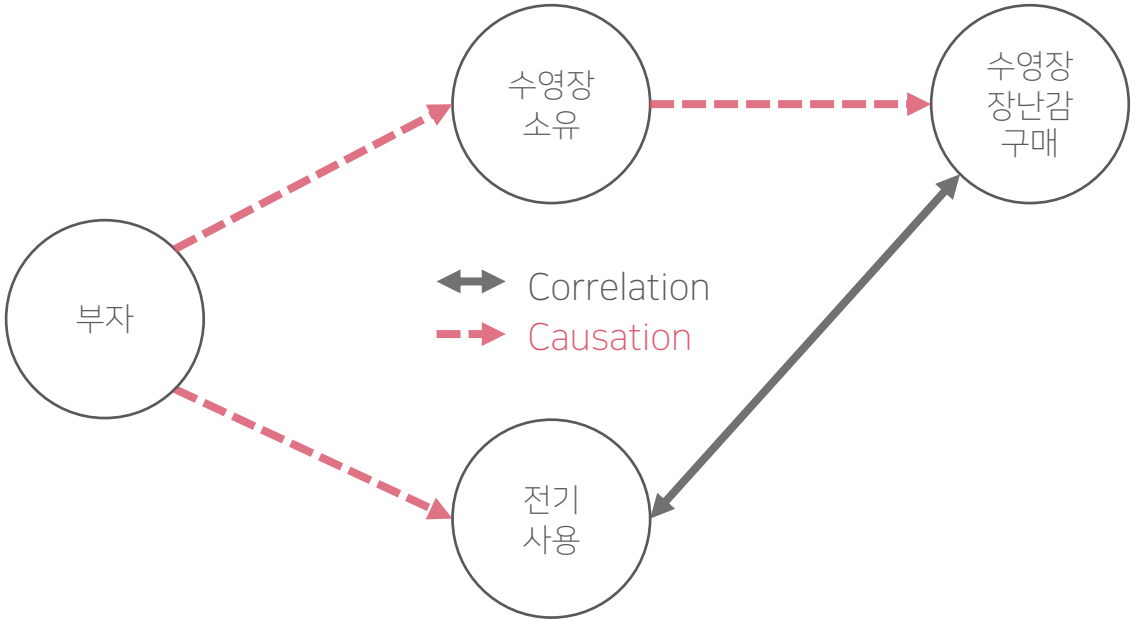
When Knowing Correlation is Enough

수영장 장난감 광고를 하려고 하는데
누가 수영장을 소유했는지 모름



When Knowing Correlation is Enough

Targeted Ad
전기회사와 제휴, 전기요금 사용량이 큰 가정에 광고 집행



참고] Correlation을 언제 의사결정에 활용해야 하나?

데이터를 통해 [소고기와 우유]를 구매한 고객들의 자동차 사고 발생률이
[라면과 소주]를 구매한 고객보다 높은 걸 확인했다. 보험회사가 취할 행동은?

가. 구매패턴에 따른
보험료 차등 적용

나. 저위험군 고객들
타겟 마케팅

참고] Correlation을 언제 의사결정에 활용해야 하나?

데이터를 통해 [소고기와 우유]를 구매한 고객들의 자동차 사고 발생률이
[라면과 소주]를 구매한 고객보다 높은 걸 확인했다. 보험회사가 취할 행동은?

가. 구매패턴에 따른
보험료 차등 적용

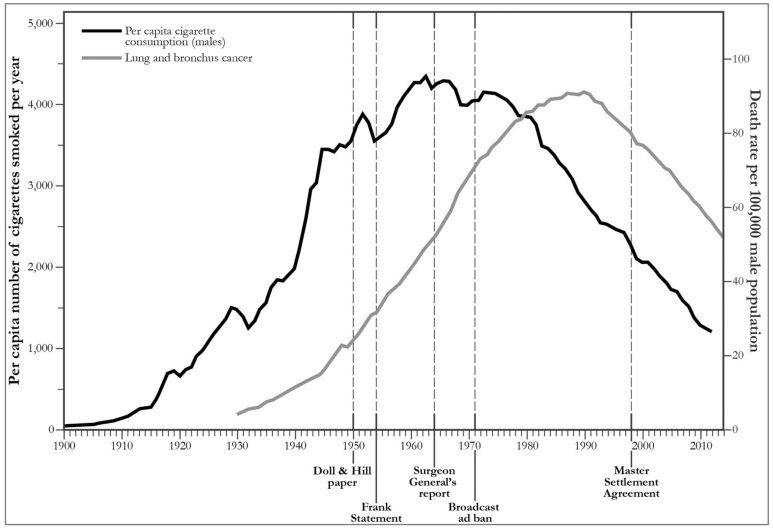
X

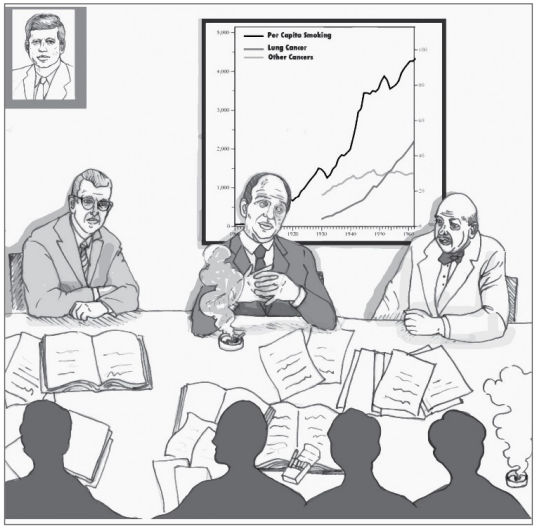
나. 저위험군 고객들
타겟 마케팅

O

RCT (Randomized Controlled Trials)

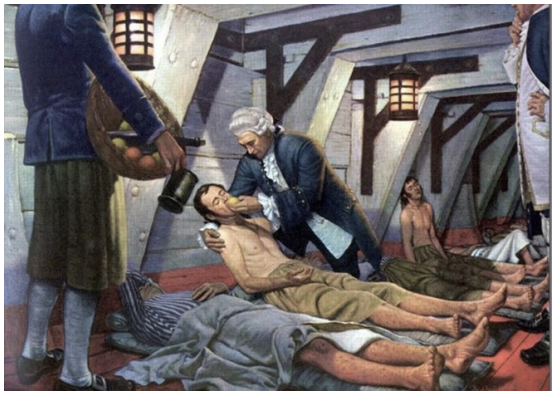
- X → Y 영향(인과성)을 측정하기 위해 실험을 설계하는 방식
- Control: 결과에 영향을 주는 알고 있는 외부 변인(나이)을 유사하게 샘플링
- Random: 모르는 외부 요인의 영향을 추가로 제거하기 위해 무작위로 샘플링
- Mostly Neither Feasible nor Ethical

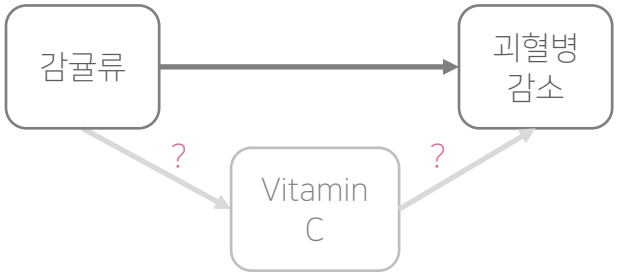




감귤 → 괴혈병

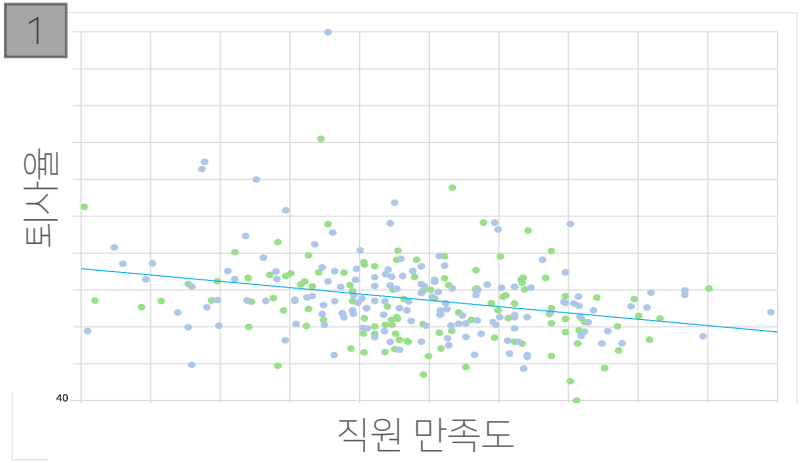
통제된 실험(RCT)이 아니라 관찰 데이터만으로 인과적 관계 발견

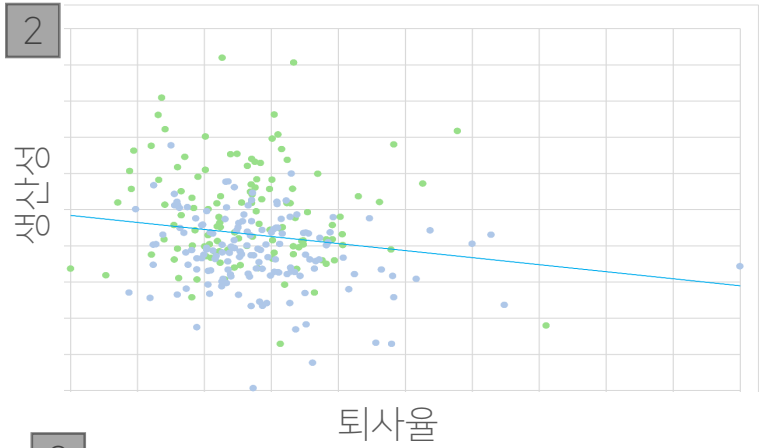


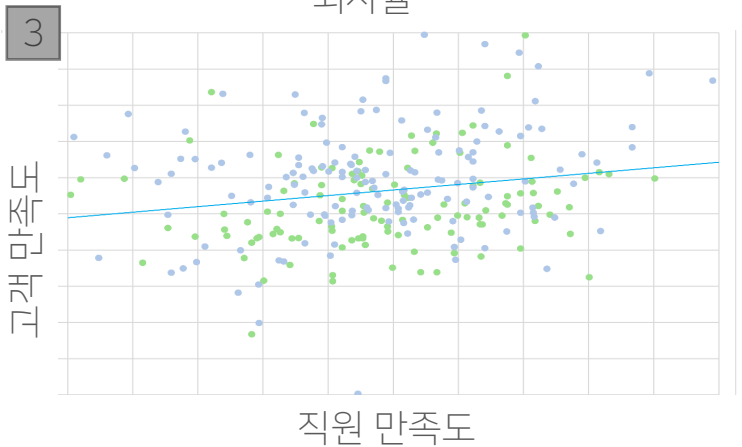


문제점	해결책
Causal Direction	Domain Knowledge
Confounding	Conditioning (변수 조건 통제; 유사 그룹별 분석)

Case Study – Causal Graph for 생산성 요인





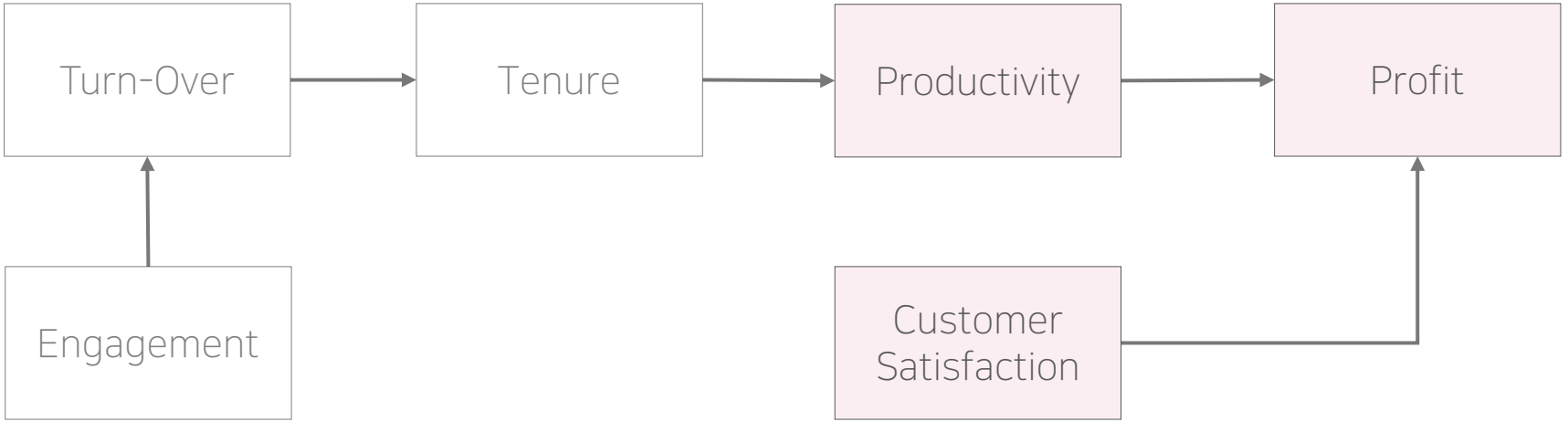


직원 만족도가 중요한가?

① 직원 만족도↗ → 퇴사율↘
② 퇴사율↘ → 생산성↗
③ 직원 만족도↗ → 고객 만족도↗

Case Study – Causal Graph for 생산성 요인

상관관계 정보만 사용해서 Causal Graph(DAG) 그리기



Case Study – Conditioning for 성과 요인

Individual Performance = f(능력, 노력)

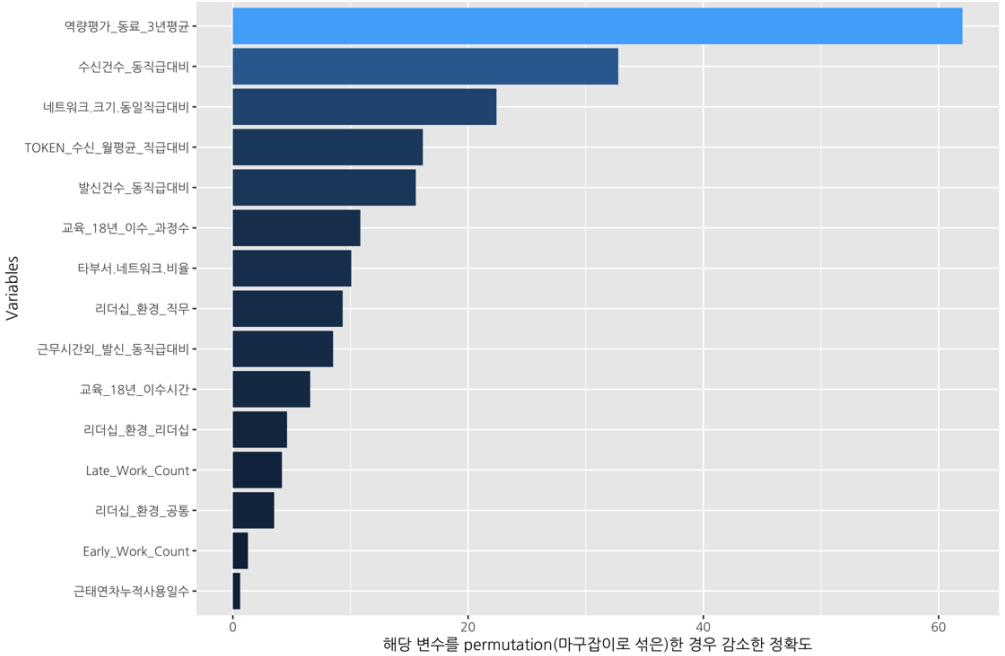
고성과자들은
저성과자들에
비해

업무 역량이
뛰어나고

리더십에 대한
안 좋은 인식에도
불구하고

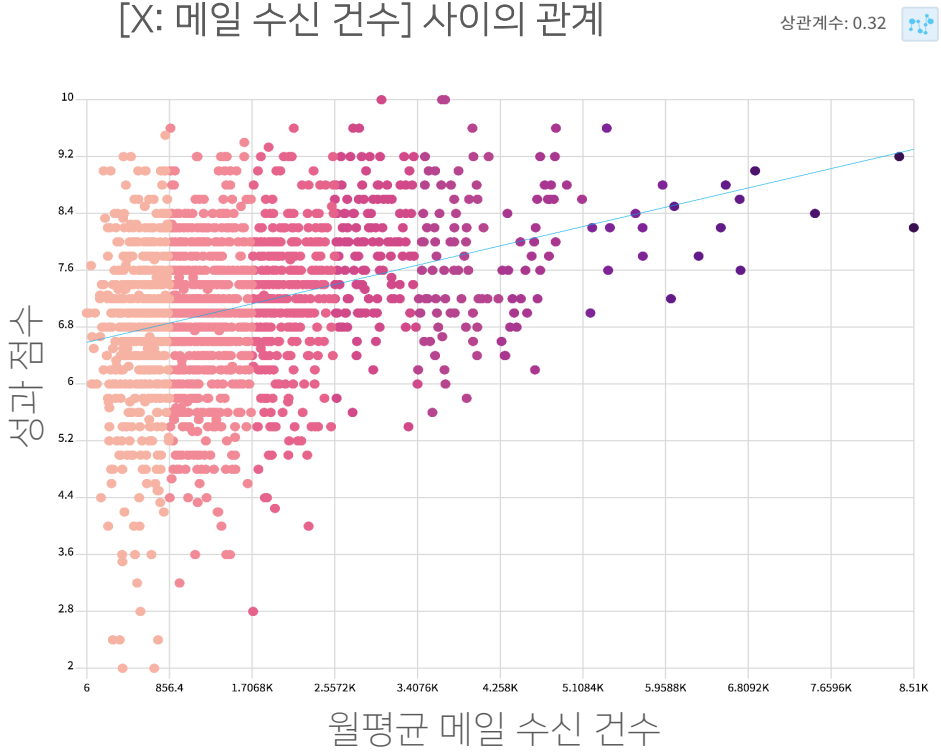
일을 많이 함

- **메일 수발신 ↗**
- 타부서 업무 비중 < 20%
- 업무시간 외 근무 ↗
- 교육 ↘



Case Study – Conditioning for 성과 요인

[Y: 성과 점수]와
[X: 메일 수신 건수] 사이의 관계

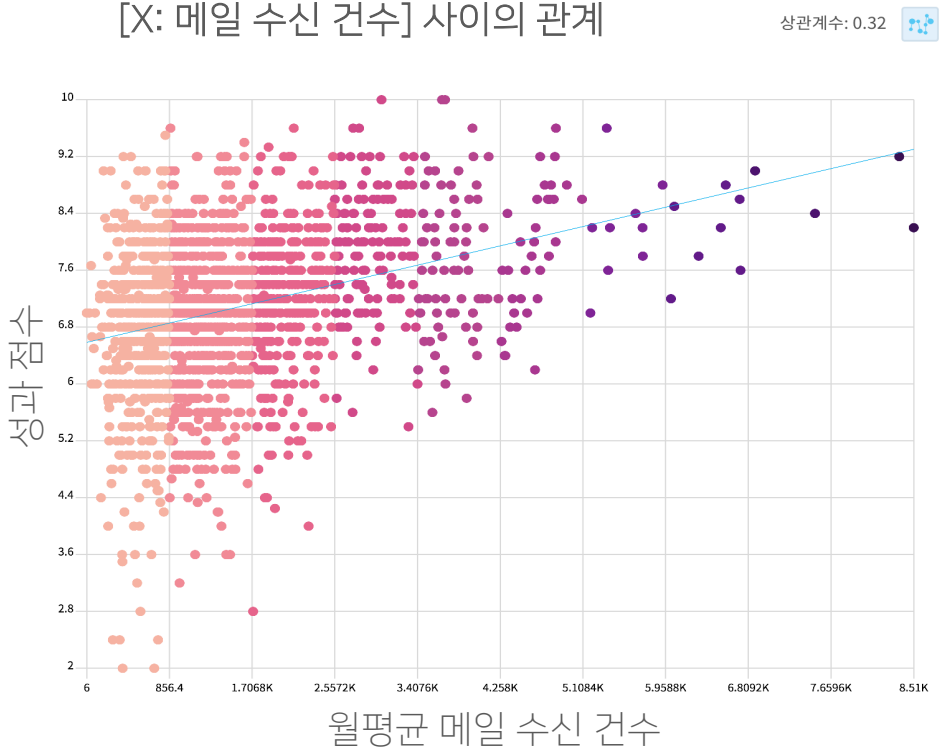


현실을 바꿀 수 있는 패턴의 조건

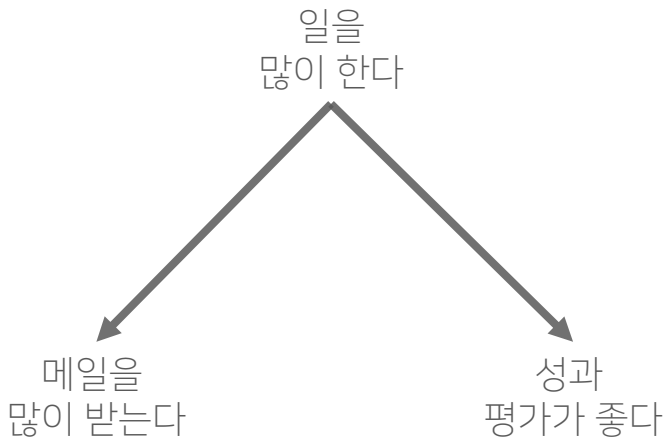
① X와 Y 사이 인과성?
② X에 개입 가능?
③ Y의 개선 정도 측정?

Case Study – Conditioning for 성과 요인

[Y: 성과 점수]와
[X: 메일 수신 건수] 사이의 관계



① X와 Y 사이 인과성?
② X에 개입 가능?
③ Y의 개선 정도 측정?

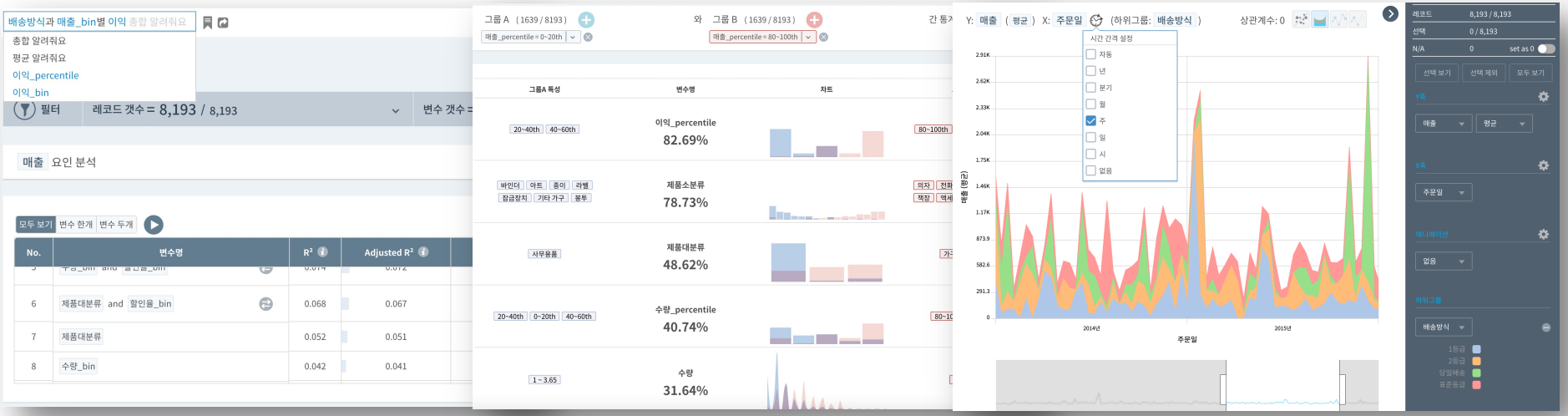


다정한 데이터 도구, HEARTCOUNT

- HEARTCOUNT(하트카운트)는 비전문가도 쉽게 엑셀 데이터셋을 업로드하여 시각화하고 분석할 수 있는 SaaS 솔루션입니다.
- Google 계정만 있다면, 홈페이지에서 바로 사용을 시작할 수 있어요!

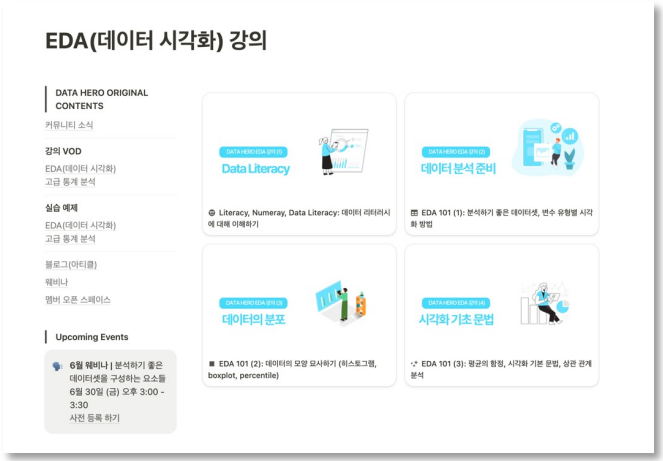
다정한 데이터 도구, HEARTCOUNT

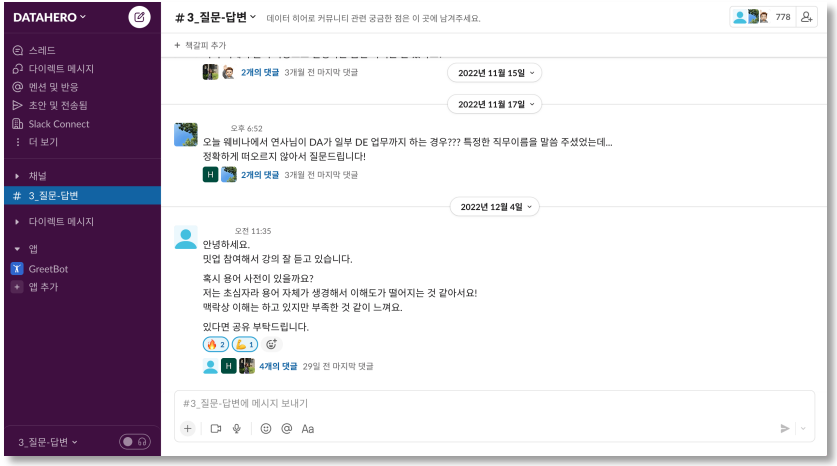
- 특장점으로는 '개별 레코드 수준의 시각화, '파생 변수 자동 생성', '패턴 자동 발견', '자연어 검색 + 설명' 등이 있습니다.



학습하고 소통하는 공간, DATA HERO

- 데이터의 기초부터 실전까지, 전용 페이지에서 무료로 학습 가능
- 하트카운트팀은 물론 다양한 실무자들과의 소통 공간
- 다양한 집중 교육 캠프, 오프라인 밋업 등 이벤트





실습 시간

- EDA 도구: https://www.heartcount.io/login