데이터 분석,
의심에서 전달까지

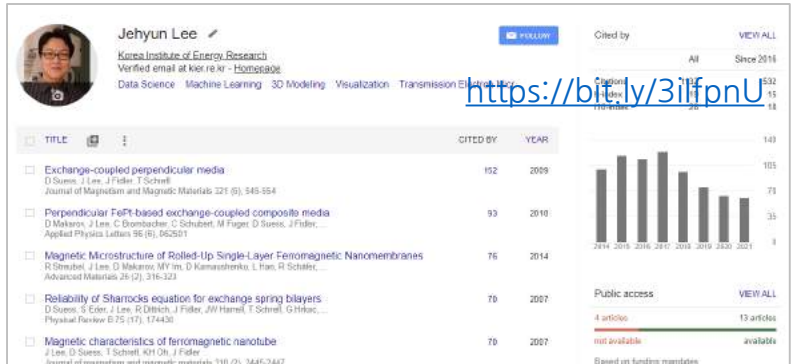
2021. 10. 08.

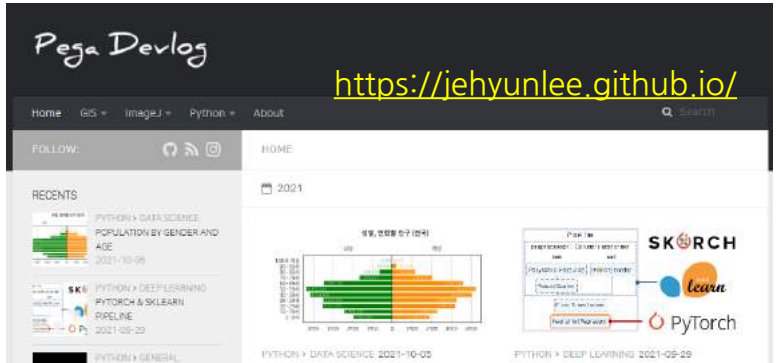
한국에너지기술연구원 이제현

안녕하세요

이름 : 이제현

소속 : 한국에너지기술연구원	2018.01. -	신재생E, 신소재, etc.
삼성전자 반도체연구소	2013.10. - 2017.12.	공정 sim, 3D 모델링, AI
삼성전자 종합기술원	2013.01. - 2013.09.	자성소재 개발
서울대학교 연구교수	2011.09. - 2012.12.	재료공학부
흔적 : Google Scholar	2004. -	44 SCI papers (h-index 19), ~3/year
Pega Devlog	2019.12. -	114 posts, ~5/month
책읽고 글쓰기	2020.01. -	22 posts, ~1/month






사 먹는 사람 vs 파는 사람

식당에서 음식을 **사 먹는** 사람

- 배가 고프다
- 어느 정도 식사를 할 시간이 있다.
- 이 식당에서 판매하는 음식을 먹고 싶다
- 이 식당의 위생을 신뢰한다
- 이 식당의 음식을 신뢰한다
- 이 식당의 서비스를 신뢰한다

https://www.fsnews.co.kr/news/articleView.html?idxno=35428

식당에서 음식을 **파는** 사람

- **배고픈 사람을 유인해야 한다.**
- **너무 오래 기다리게 하면 안된다.**
- **먹음직스러운 음식을 판매해야 한다.**
- 이 식당의 위생을 신뢰한다 **청결 : 기본**
- 이 식당의 음식을 신뢰한다 **맛 : 기본**
- **기분이 불쾌하지 않아야 한다.**

http://www.fsnews.co.kr/news/articleView.html?idxno=11722

식당에서 음식을 파는 사람의 **자세**

http://www.fsnews.co.kr/news/articleView.html?idxno=11722

- **우리 가게는 다른 가게랑 뭐가 다를까?**
- **muscle memory**
- **우리 음식을 어떻게 궁금하게 할까?**
- 이 식당의 위생을 신뢰한다 **청결 : 기본**
- 이 식당의 음식을 신뢰한다 **맛 : 기본**
- **어떻게 기분을 좋게 모실까?**

이 모든 것을 가능하게 하는 것 : **의심**

의심을 해소해서 **신뢰**를 만든다

“dubito, ergo cogito, ergo sum”

의심의 범위 : “의심할 수 있는 모든 것”

https://www.hankookilbo.com/News/Read/A2021070510310004511 http://news.kmib.co.kr/article/view.asp?arcid=0014434852&code=61181911&sid1=ent&cp=du1

음식을 만드는 사람이 의심해야 할 대상

나

- 왜 이 음식을 만들어 팔지?
- 이게 최선인가?
- 손님에게 뭘 전달하려고 하지?
- 이 음식을 드신 손님이 어쩌기를 바라지?

재료

- 원산지는 어디지?
- 상하지는 않았나?
- 저거 말고 이거 맞지?
- 필요한 거 다 있나?

조리 방법

- 감자는 이렇게 써는 거지?
- 이 팬에다 하는 거 맞지?
- 초벌구이를 해야 되던가?
- 너무 안 맵게, 안 싱겁게!

손님

- 배가 고프신건가,
  맛이 고프신건가?
- 알러지는 없으신가?
- 맛있게 드시고 계신가?

- **이게 정량이긴 한데 양에 차시려나?**
- **중년 남성이시니 얼큰한게 좋으실까?**
- **오늘 닭이 이상하게 질기네. 평소보다 잘게 자를까?**

데이터를 분석하는 사람이 의심해야 할 대상

나

- 왜 이 데이터를 뒤지고 있지?
- 이게 최선인가?
- 고객이 진짜 원하는 게 뭐지?
- 고객이 이 결과를 받아서 뭘 하길 바라지?

데이터

- 레퍼런스는 어디지?
- 빠진 column, row 없나?
- 저거 말고 이거 맞지?
- 필요한 거 다 있나?

분석 방법

- 이 데이터는 이렇게 전처리하는 거지?
- 이 모델에다 넣는 거 맞지?
- 정규화를 해야 되던가?
- 언더피팅 안되게, 오버피팅 안되게!

고객

- 분석이 궁금하신가, 대안을 요구하시나?
- 이 분 금기어가 뭐더라?
- 결과물 마음에 들어하시나?

- **진짜 원하는 건 따로 있는 듯 한데?**
- **통계는 모르시니 어려운 말은 뺄까?**
- **이 데이터 앞뒤가 안 맞네. 레퍼런스 한번 들어가볼까?**

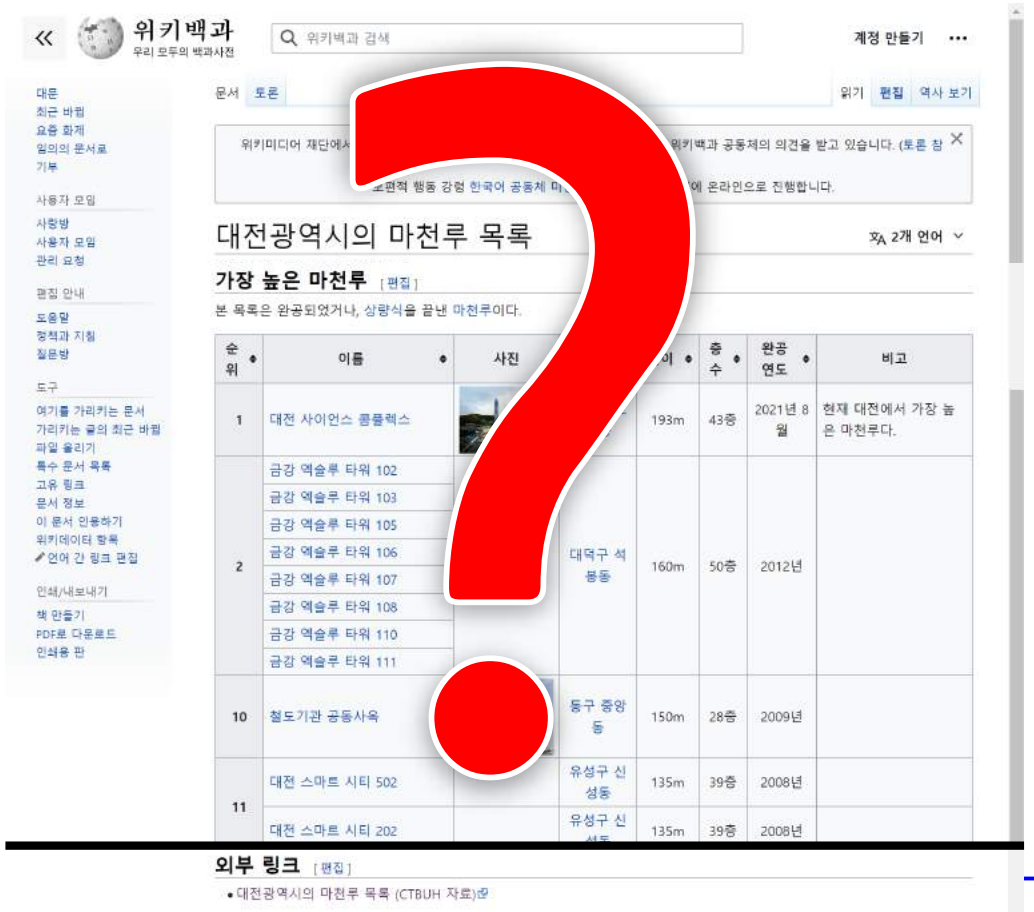
1. 데이터 의심하기

- 레퍼런스는 어디지?
- 빠진 column, row 없나?
- 저거 말고 이거 맞지?
- 필요한 거 다 있나?

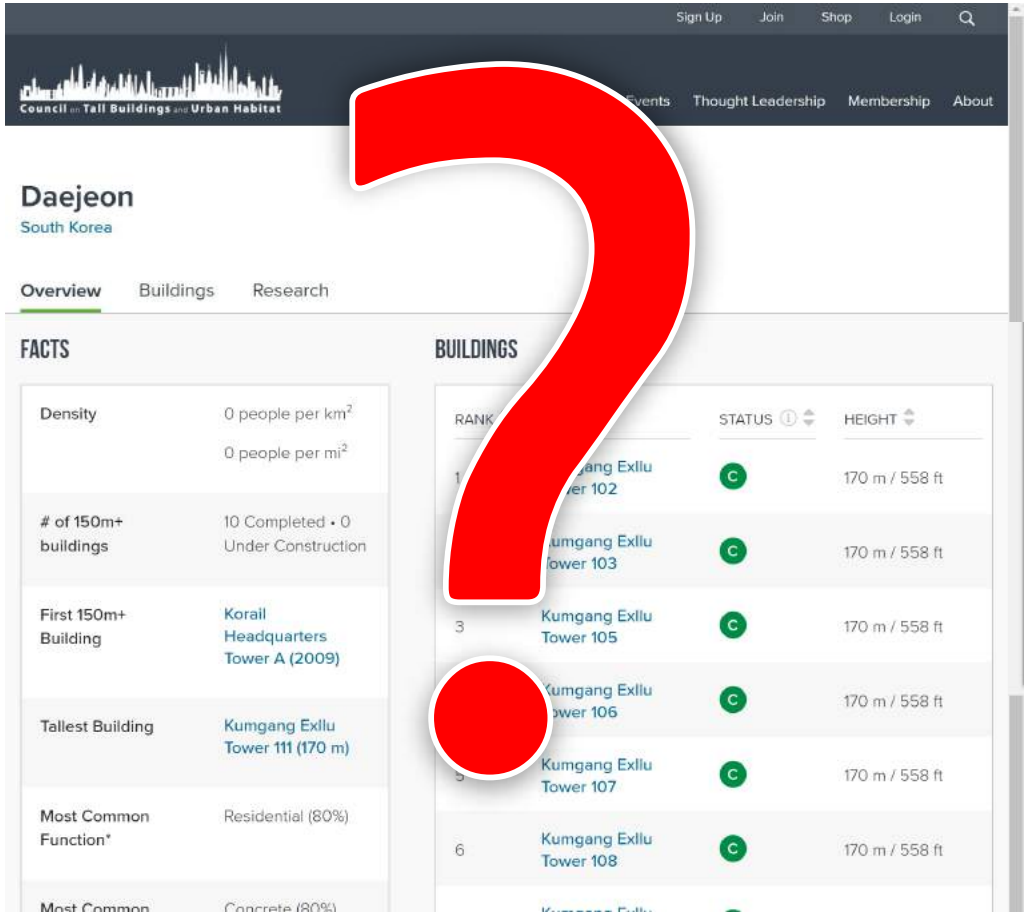
① 레퍼런스 : bad case

- “대전광역시에서 가장 높은 건물들”

https://ko.wikipedia.org/wiki/대전광역시의_마천루_목록



https://www.skyscrapercenter.com/city/daejeon



① 레퍼런스 : good case

- “대전광역시에서 가장 높은 건물들”

https://blcm.go.kr/

② row, column : 건전성 확인

이미지= “맛의 달인”. 대사 수정=본인

체크리스트 : 있나? 몇 개나 있나? 비율은? 어디에 있나?

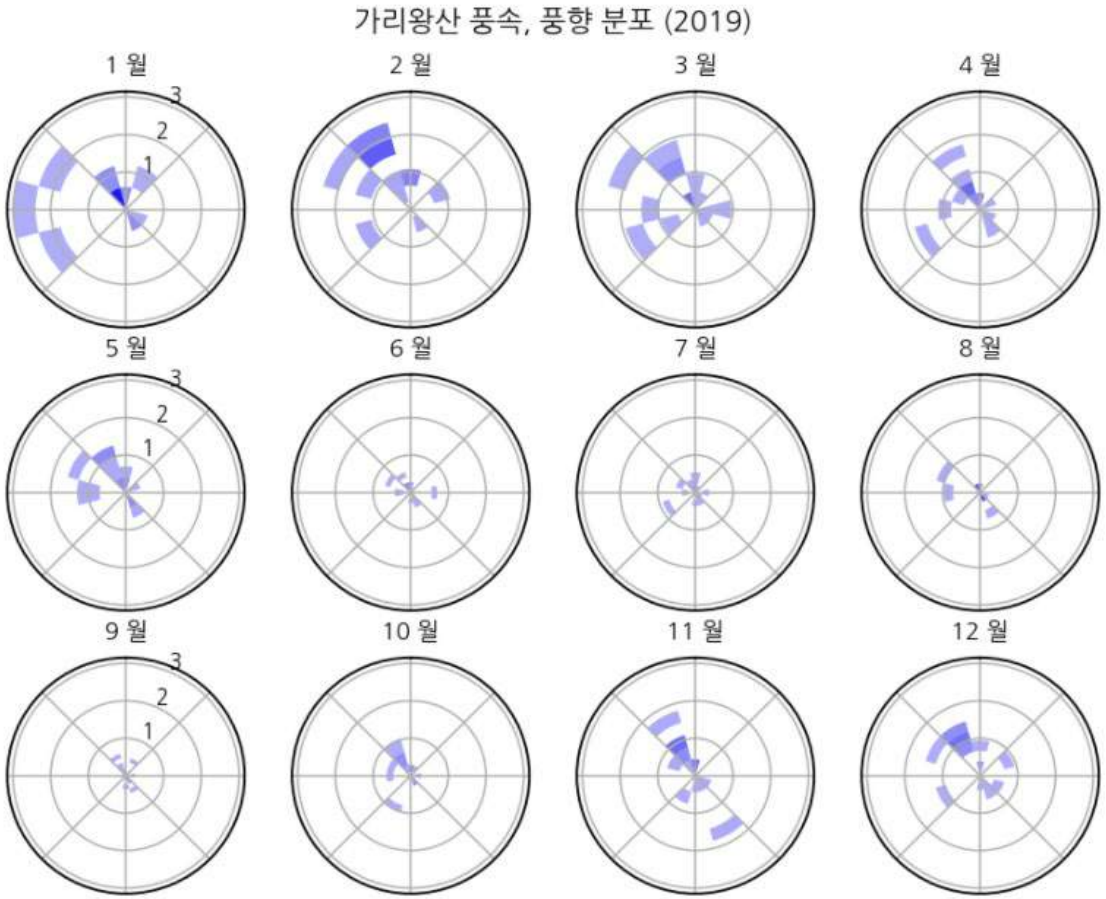
- **결측치 :** 데이터가 없음.
  가끔은 “데이터가 없음”이 중요한 메시지
  이 데이터는 왜 없을까? 단순 누락일까? 이유가 있을까?

- **중복 데이터 :** 같은 데이터가 여러 개
  Key feature를 중심으로 논리적으로 판단해야 함.
  같은 자리에 건물이 절반 겹쳐서 두 개: 비정상
  같은 시간 같은 가게에 손님이 두 명: 정상

- **이상치 :** 정상적인 범위에서 벗어나는 데이터
  통계 분석을 통해 이상치 후보군을 추리고,
  도메인 접근을 통해 진짜 이상치인지 판별해야 함.
  대학 중퇴자가 왜 이렇게 소득이 높아?
  빌 게이츠, 스티브 잡스, 마크 저커버그, 잭 도시 (…)

② row, column : 건전성 확인

- “이건 결측치일까, 아닐까?”



산림 빅데이터 플랫폼 : 산림 풍향풍속 구간별 관측 정보

(지상기상관측용 자동기상관측장비의 표준규격 고시)

[별표 4]

관측센서의 신호 및 자료처리의 표준규격(제9조관련)

관측요소	신호 및 자료처리 표준규격
기온(초상, 지면, 지중온도), 습도, 기압	○ 자료 단위 : 0.1 ℃(기온, 초상, 지면, 지중온도), 0.1 %(습도), 0.1 hPa(기압) ○ 샘플링 시간 : 10초 ○ 자료처리 시간간격 : 1분 - 10초마다 전기적 신호를 수신하여 디지털값으로 변환한다. - 10초 간격의 6개 자료를 평균하여 1분 자료를 산출한다.
바람자료	○ 자료 단위 : 풍향(0.1°), 풍속(0.1 ㎧) ○ 샘플링 시간 : 0.25초 ○ 자료처리 시간간격 : 1분 ○ 풍향, 풍속센서로부터 0.25초마다 전기적 신호를 수신하여 디지털 값으로 변환 후 벡터 환산한다. ○ 순간풍향·풍속(gust) - 매 0.25초 간격으로 3초 동안 12개의 샘플링된 자료를 평균하고 1초 간격으로 이동평균하여 순간풍향·풍속을 산출한다. - 1분 동안 수집된 지난 240개의 자료 중 최대값을 1분 최대 순간풍향·풍속을 산출한다. - 매 1분마다 지난 10개의 1분값 중에서 최대값을 10분 최대 순간풍향·풍속을 산출한다. - 하루 동안 수집된 1분 최대순간풍향·풍속 1440개 중에서 최대값을 일 최대순간풍향·풍속을 산출한다. ○ 1분(10분) 평균 풍향·풍속 - 0.25초 간격의 바람벡터 자료를 10초 동안 평균을 구한 후 1분(10분) 동안 6개(60개)의 자료를 다시 평균하여 매분(10분)자료를 산출한다.

기상청 : 기상측기규격기준고시

③ 너무 믿지 말아야 할 데이터

- 영화 장르 데이터 http://www.kobis.or.kr/kobisopenapi/homepg/apiservice/searchServiceInfo.do
  - 주관적인 판단에 의한 데이터 : 일관성을 상실할 소지가 큼
  - 큰 수의 법칙(?) : “많이 모으면 전반적인 경향은 비슷하겠지”

영화 제목	개봉년도	장르
해리포터와 마법사의 돌	2001	가족, 판타지
해리포터와 비밀의 방	2002	드라마
해리포터와 아즈카반의 죄수	2004	드라마
해리 포터와 불의 잔	2005	판타지, 액션
해리 포터와 불사조 기사단	2007	판타지, 액션, 어드벤처
해리 포터와 혼혈 왕자	2009	액션, 어드벤처, 가족
해리 포터와 죽음의 성물1	2010	미스터리, 판타지
해리포터와 죽음의 성물2	2011	미스터리, 판타지

④ 내게 필요한 그 데이터인지

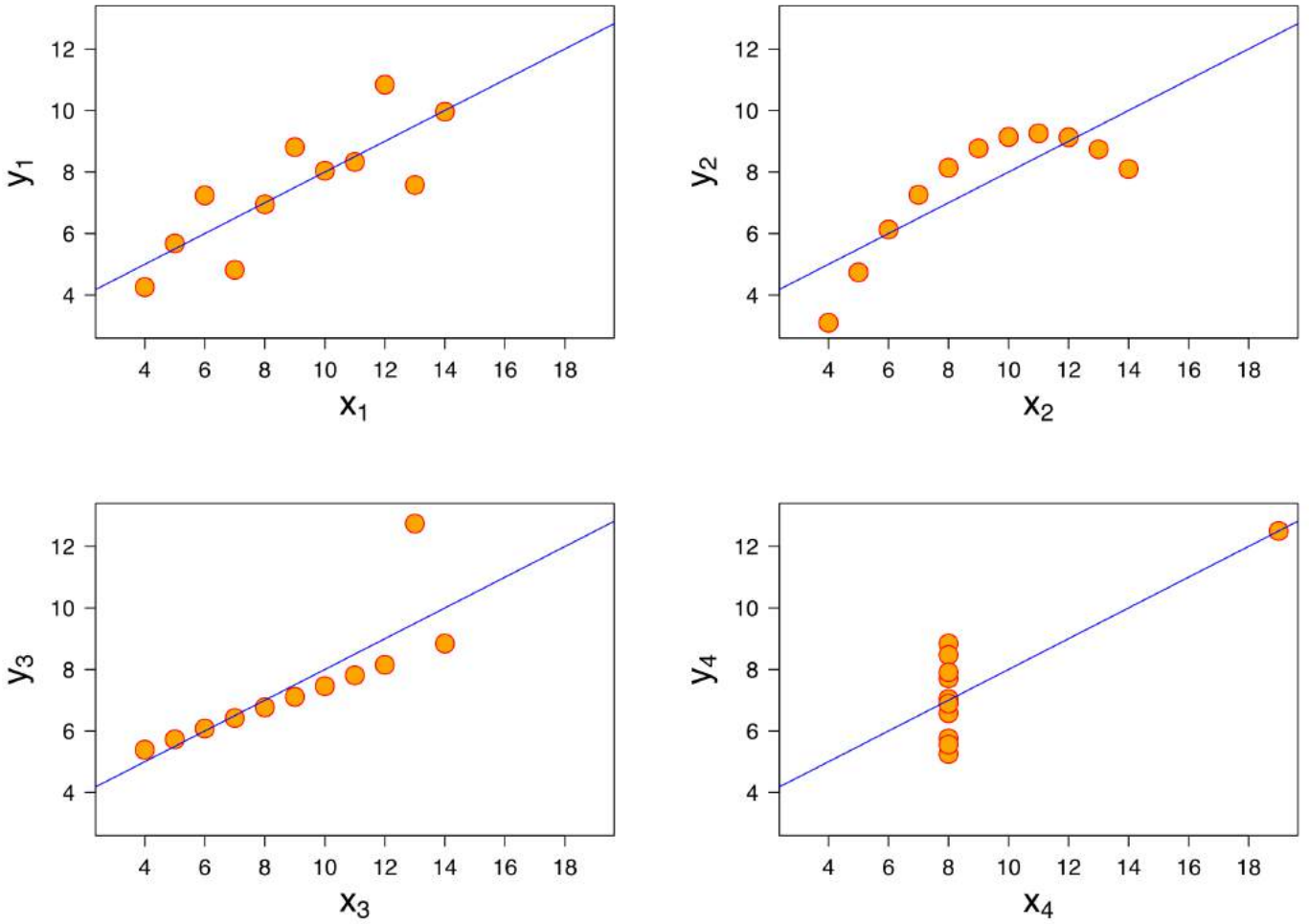
- 엄마 : “어? 두부가 없네? 아들, 마트에서 두부 사와라.”

⑤ 데이터 파악

- “통계치 잘 보면 되겠지?”

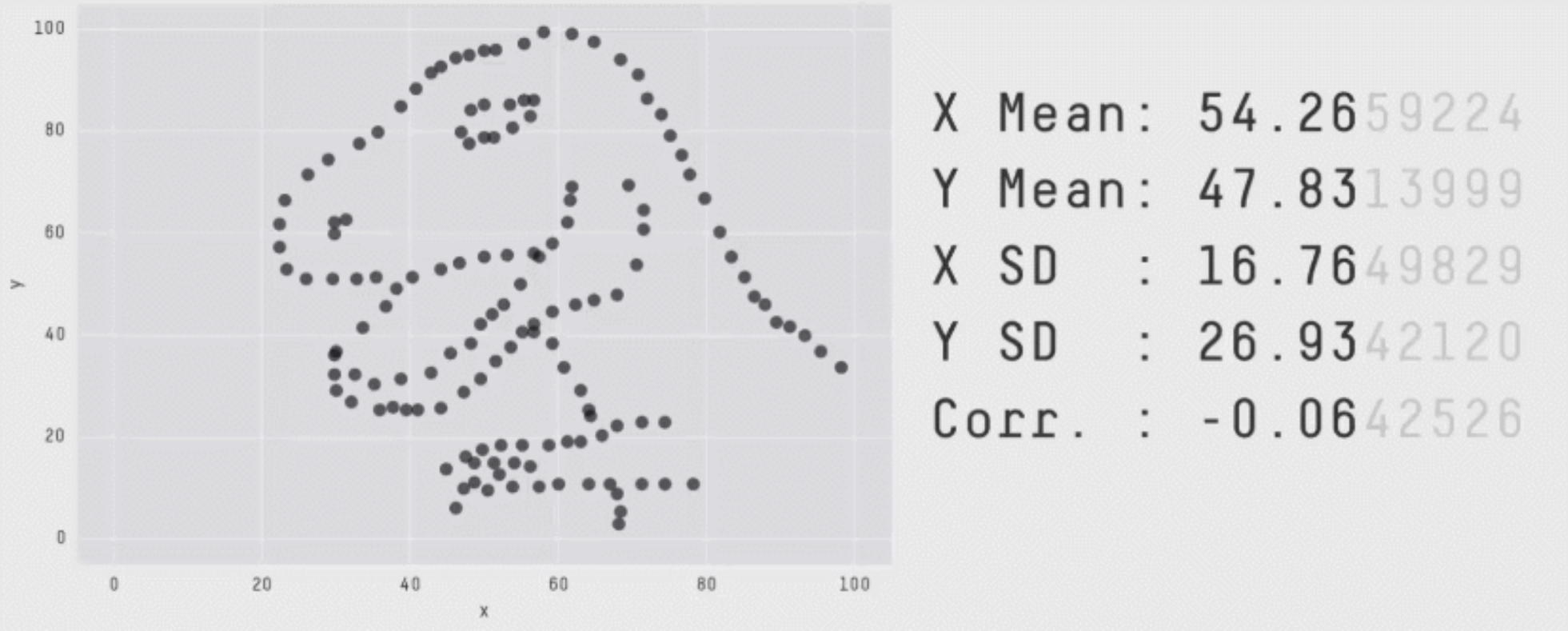
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

https://en.wikipedia.org/wiki/Anscombe%27s_quartet



⑤ 데이터 파악

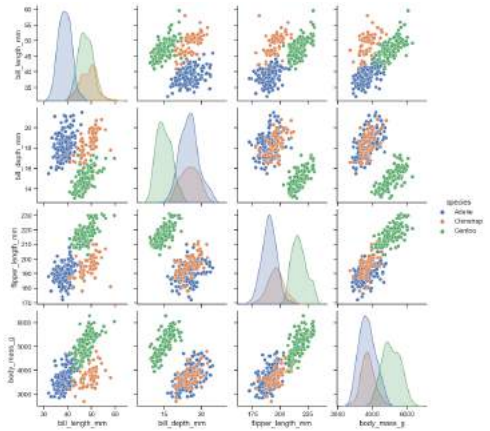
- “무조건 그린다.”

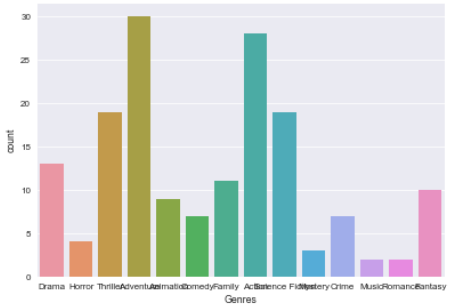


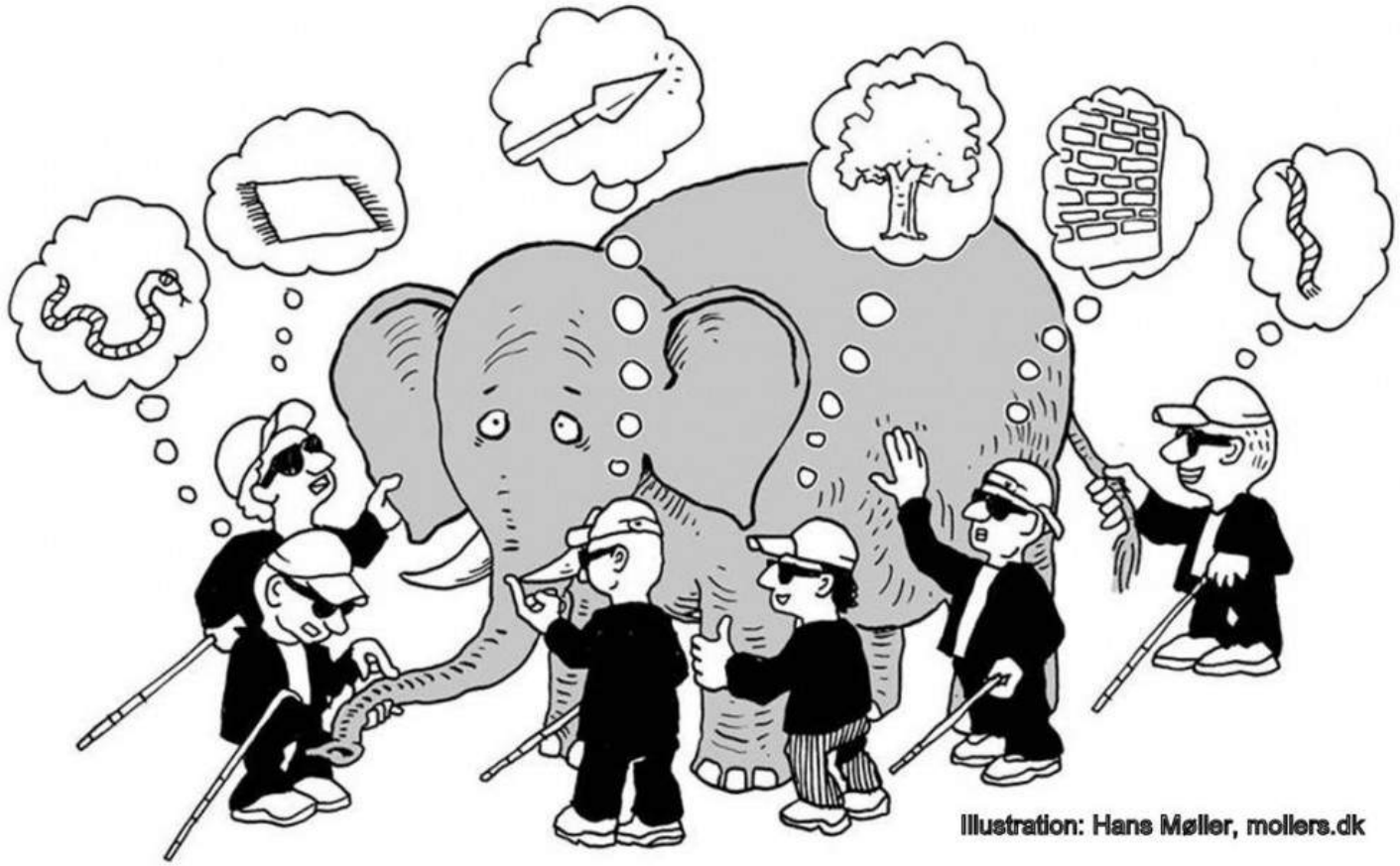
http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

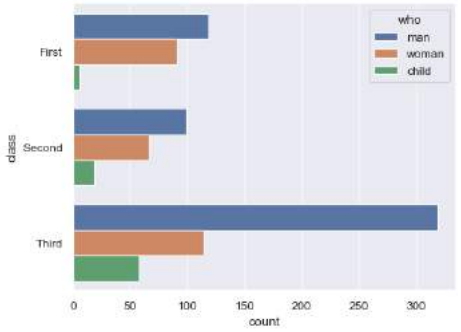
데이터를 제대로 의심하는 방법

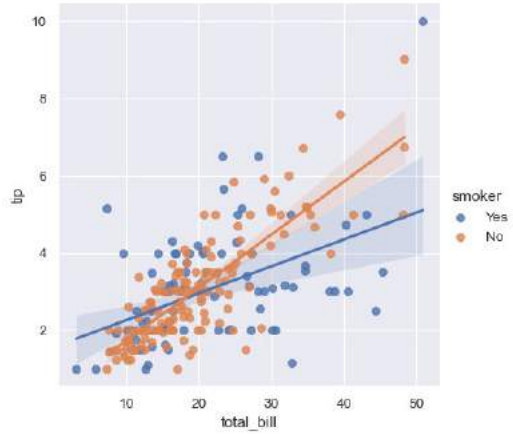
- Exploratory Data Analysis







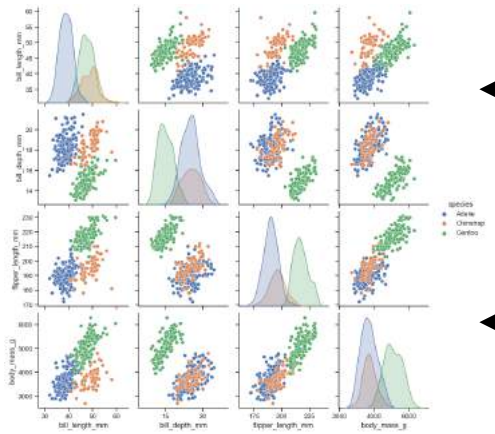




https://enfocussolutions.com/the-six-blind-men-and-the-requirements-part-one/

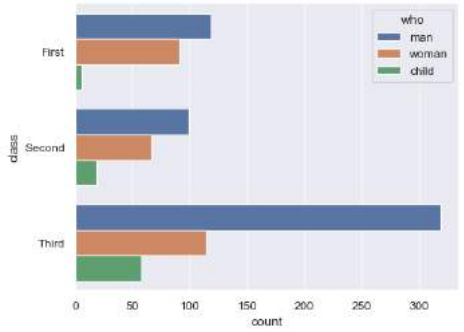
데이터를 제대로 의심하는 방법

- Exploratory Data Analysis **+ Hypothesis = self-강화학습**



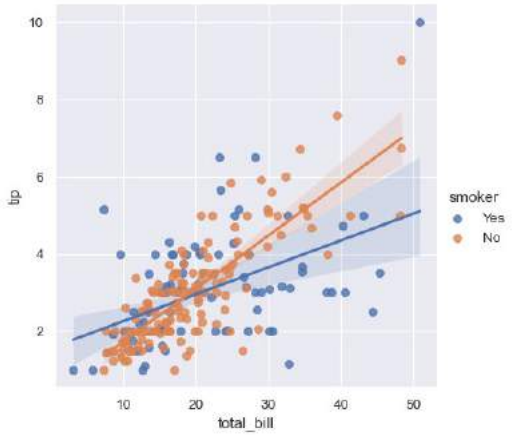
← 이거 보면 OOO가 영향을 미치는 것 같은데?

근데 이건 또 다르네? 왜 앞뒤가 안 맞지? →

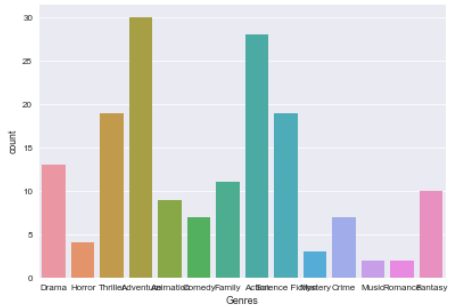


← 이 이상한 건 또 뭐야? 데이터 단위가 안 맞나?

잠깐, 이상치 제거 먼저 하고 다시 하자. →



← OO형 전공이 이거지? 해석 방법 좀 물어 봐야지.



"아 이게 그래서 이런 이야기구나. 대강 알겠다."

https://enfocussolutions.com/the-six-blind-men-and-the-requirements-part-one/

2. 분석 방법 의심하기

- 이 데이터는 이렇게 전처리하는 거지?
- 이 모델에다 넣는 거 맞지?
- 정규화를 해야 되던가?
- 언더피팅 안되게, 오버피팅 안되게!

사장님이 감자를 잘라 놓으란다. 어떡하지?

https://www.manzlab.com/news/articleView.html?idxno=1459

감자 썰기의 관건: “무엇을 만들 것인가”

촉촉함 & 부드러움: 프라이팬!

↑

감자 볶음

https://happytime.tanz.xyz/entry/백종원-감자볶음-만들기-작은-팁

바삭함: 튀김기!

↑

감자칩

https://m.blog.naver.com/cshee32/221990287069

적당한 바디감과 온기: 오븐!

↑

구운 감자

https://steptohealth.co.kr/delicious-baked-potato-5-recipes/

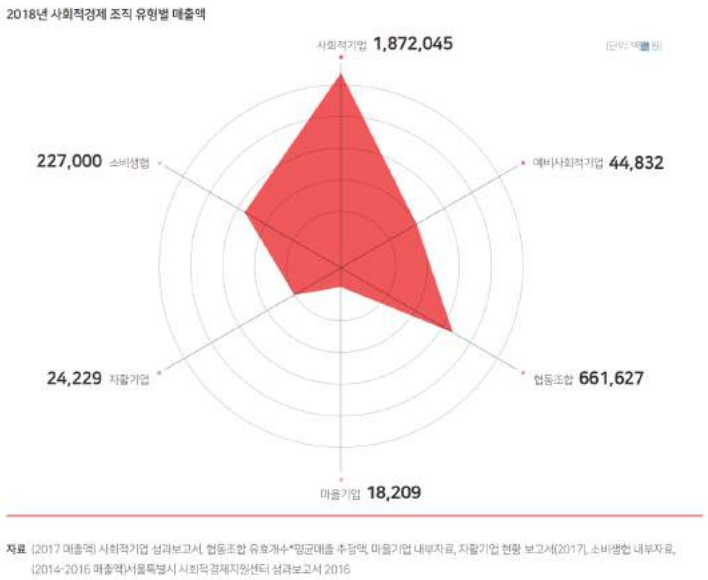
이 밖에도 수많은 감자 감자 감자 감자…

데이터 분석의 관건: “무엇을 할 것인가”

내용 전달: 글 + 그림

↑

현황 분석

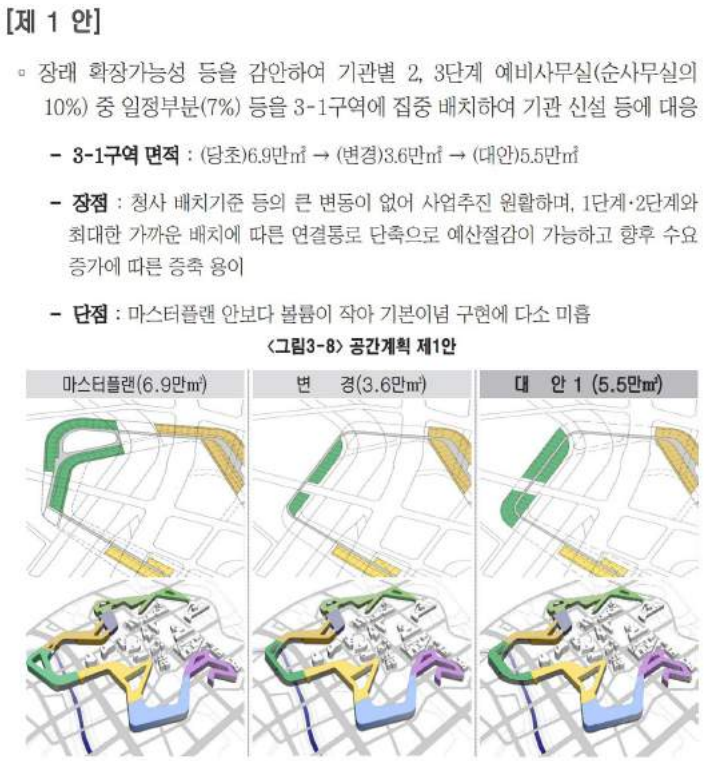


[서울시 사회적경제 조직 현황 분석 및 주요 성과 연구 보고서]

설득력: 대안의 장점과 단점

↑

대안 제시

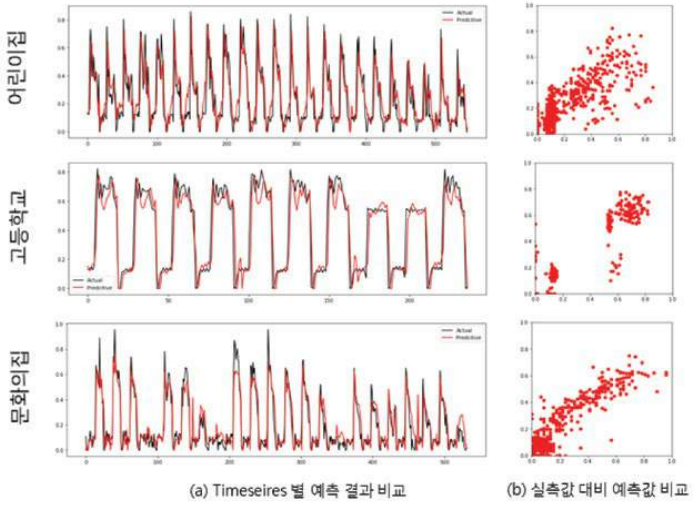


[정부세종청사 건립백서]

신뢰성: 검증 결과, 예상 오차

↑

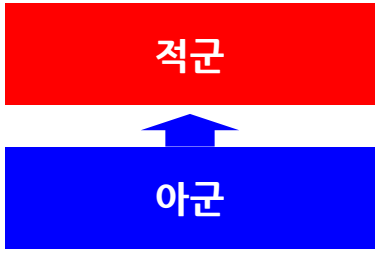
예측모델 개발

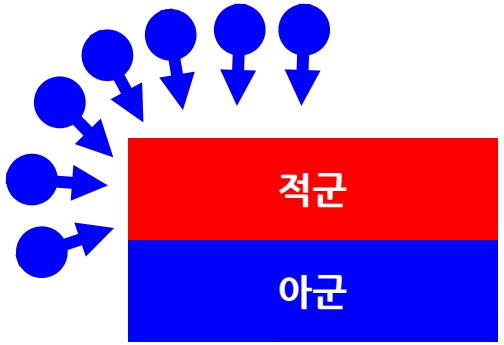


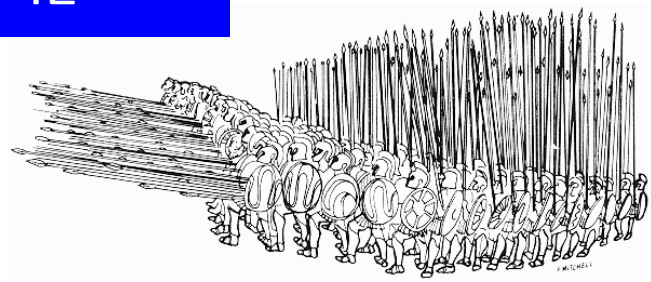
[건물 열에너지 수요 예측모델]

망치와 모루 전략 Hammer and Anvil Tactic

모루가 버티는 동안 망치가 때린다







https://www.medievalists.net/2015/04/the-men-behind-the-metal/
pixabay.com
https://en.wikipedia.org/wiki/Macedonian_phalanx
https://en.wikipedia.org/wiki/Companion_cavalry

① 모루: 수학적 엄밀함

- “허튼 소리 피하기” = 온갖 의심의 집합체

교차검증

카이제곱

정규화

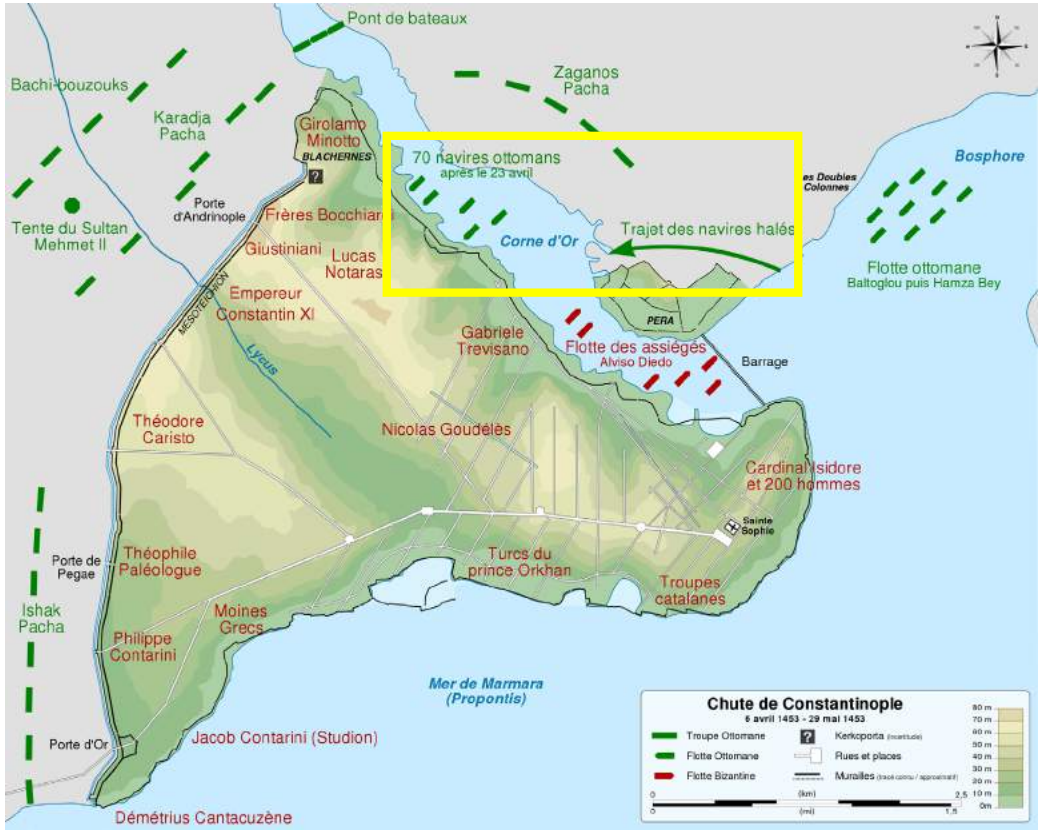
T-검정

층화추출

온갖 메트릭

② 망치: 나만의 인사이트

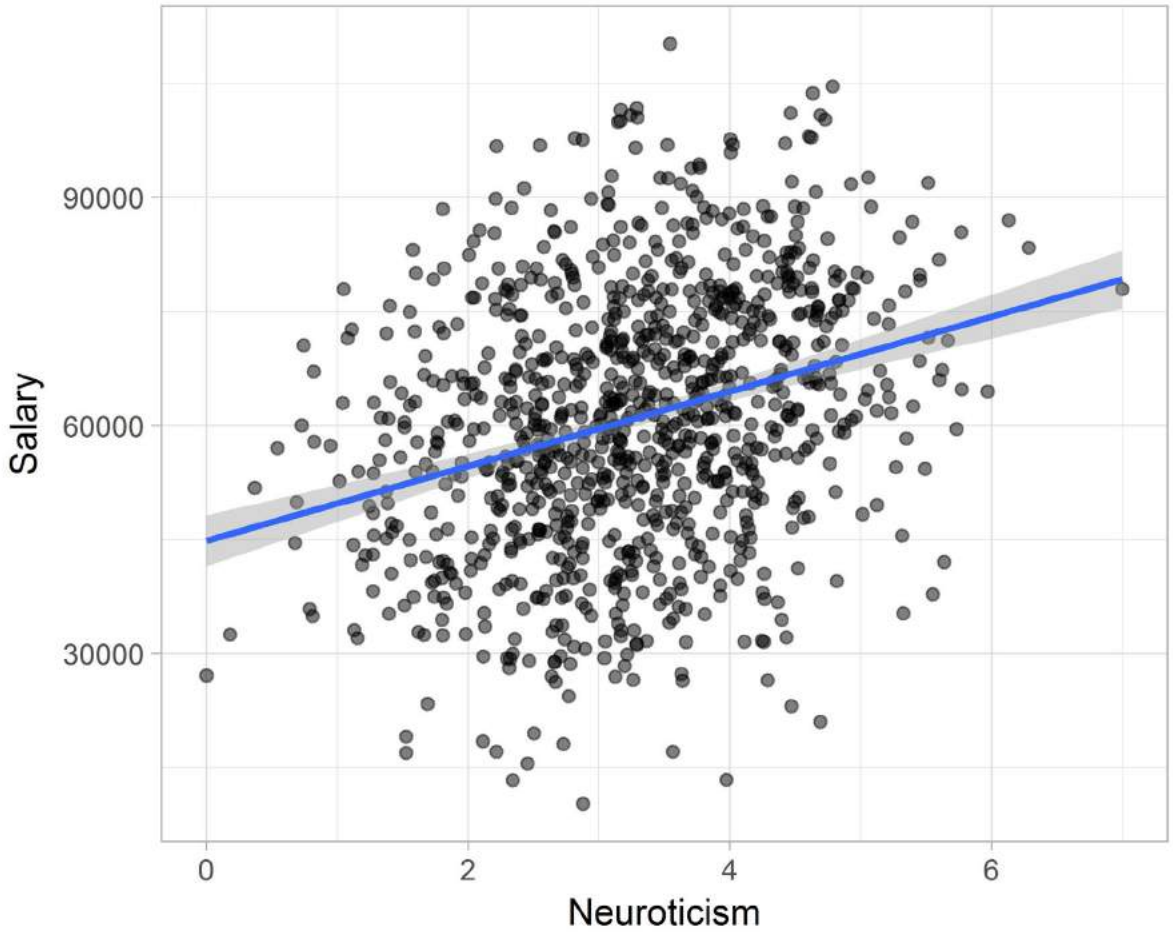
- “아무도 못한 생각을 해내기”

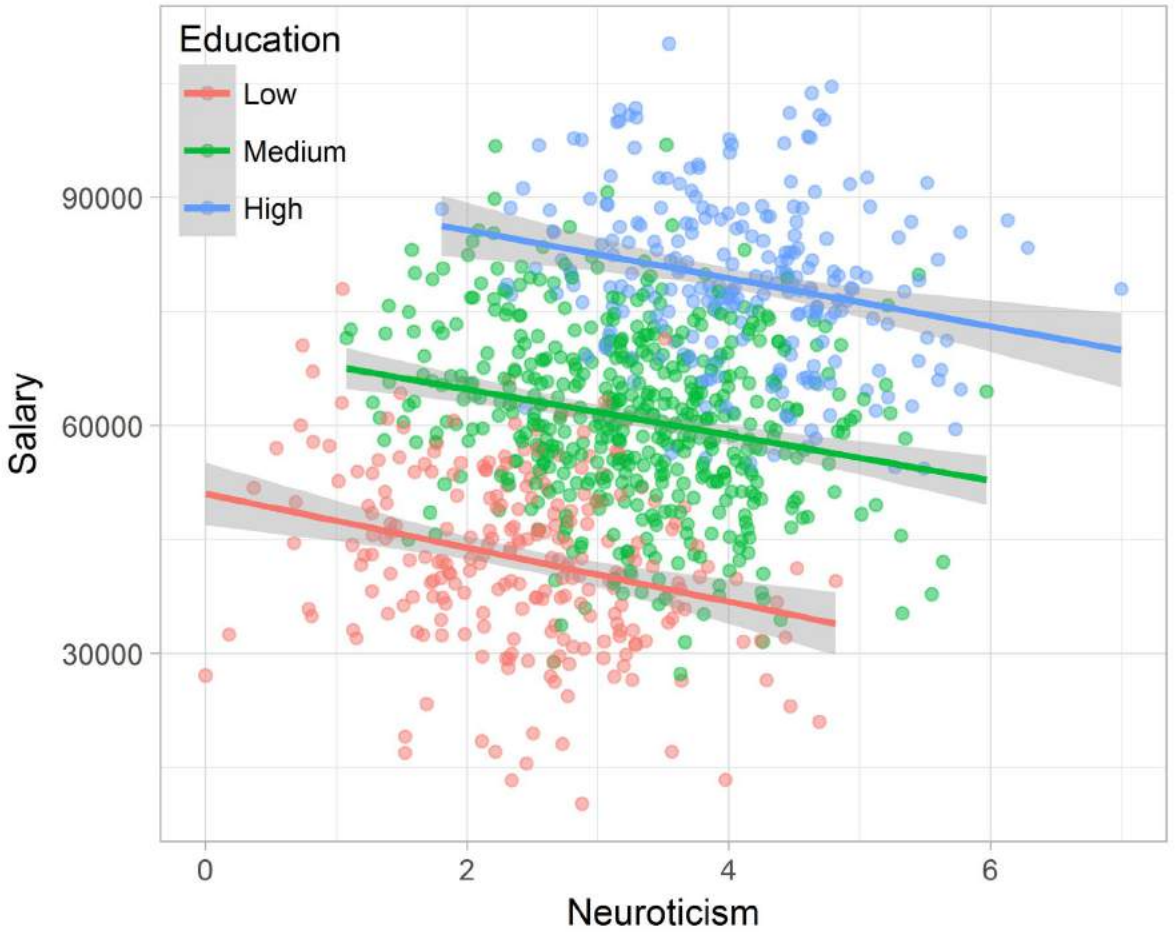


https://ko.wikipedia.org/wiki/콘스탄티노폴리스의_함락

② 인사이트 도출 방법: 데이터 자르기 Segment

- Simpson's paradox





https://paulvanderlaken.com/2017/09/27/simpsons-paradox-two-hr-examples-with-r-code/

Airbnb “국내, 300마일”

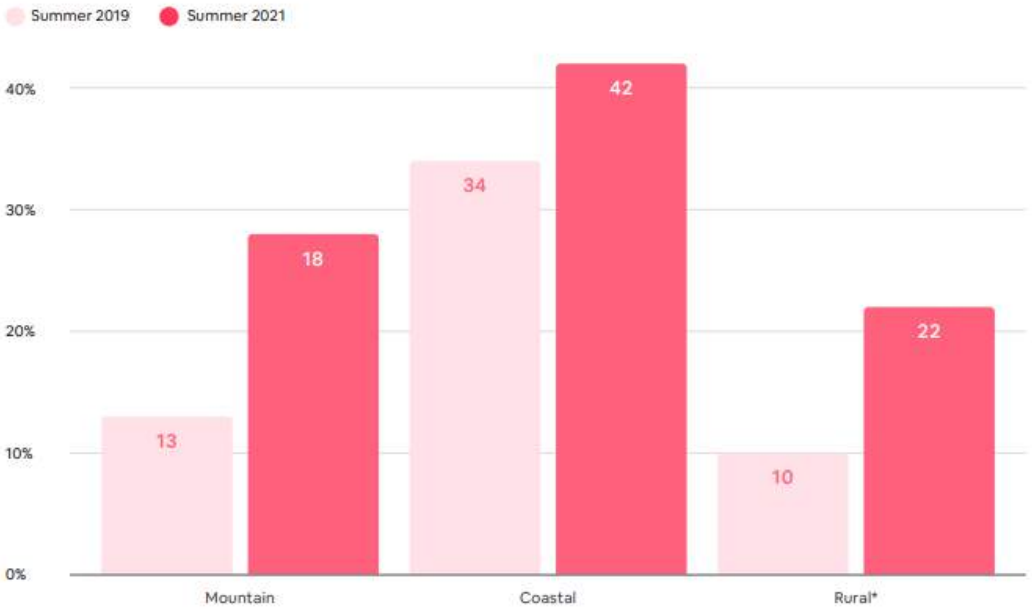
- 코로나19 이후 여행 패턴 분석

2. People are traveling everywhere

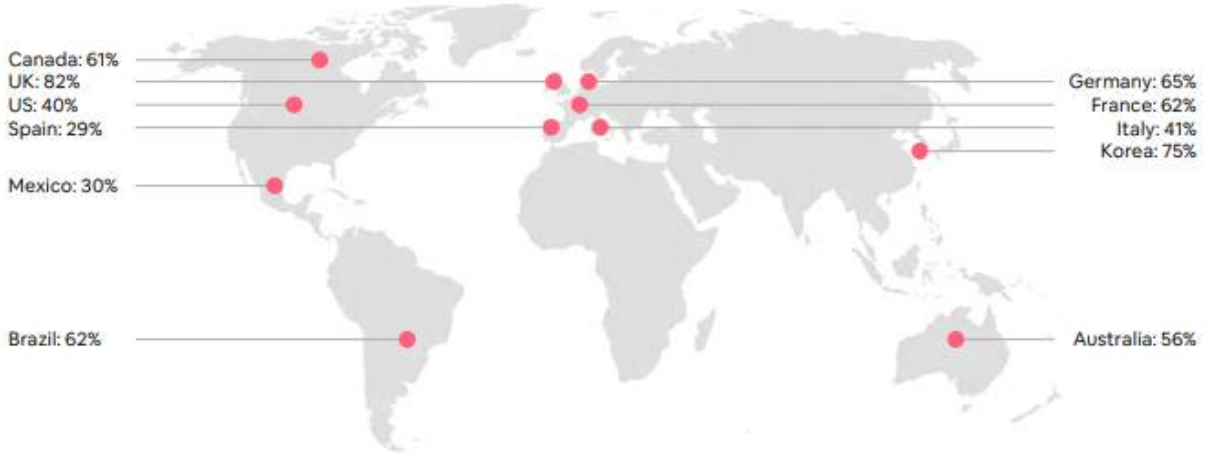
The pandemic has sped up a trend of people thinking more flexibly and further afield about where to travel, shifting away from mass tourism focused on an iconic set of cities and other popular destinations. On Airbnb, 27% of all searchers in April were flexible on location without use of any tools. At the same time, travel has spread more diffusely over the past several years as it has become less about the same places and more about people and connection, to the benefit of more communities and those who live there.

2020년 1월		2021년 1월
50%	국내여행 비중	80%
30%	300마일 이내	45%

Share of travel to natural settings

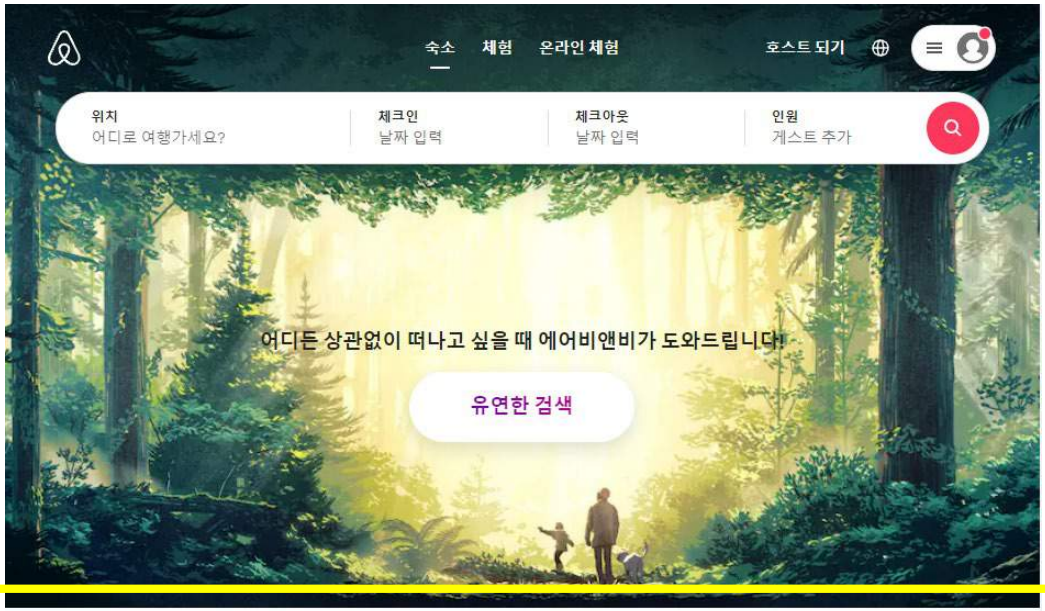


Nearby travel (within 300 miles), share of 2021 nights booked



https://news.airbnb.com/wp-content/uploads/sites/4/2021/05/Airbnb-Report-on-Travel-Living.pdf

Airbnb “300마일 이내 아무데나”



'하이퍼 솔로' 전환 통했다...에어비앤비, IPO 대박

> 코로나로 호텔업계 고전 불구
> 내집 같은 편안함에 고객 선호
> 상장 첫날 주가 113% 급등

박용범 기자 | 입력 : 2020.12.11 17:44:45 수정 : 2020.12.11 17:47:21 0

'내 주변에 300마일 이내 어디라도 가고 싶은데 추천해줄래?'

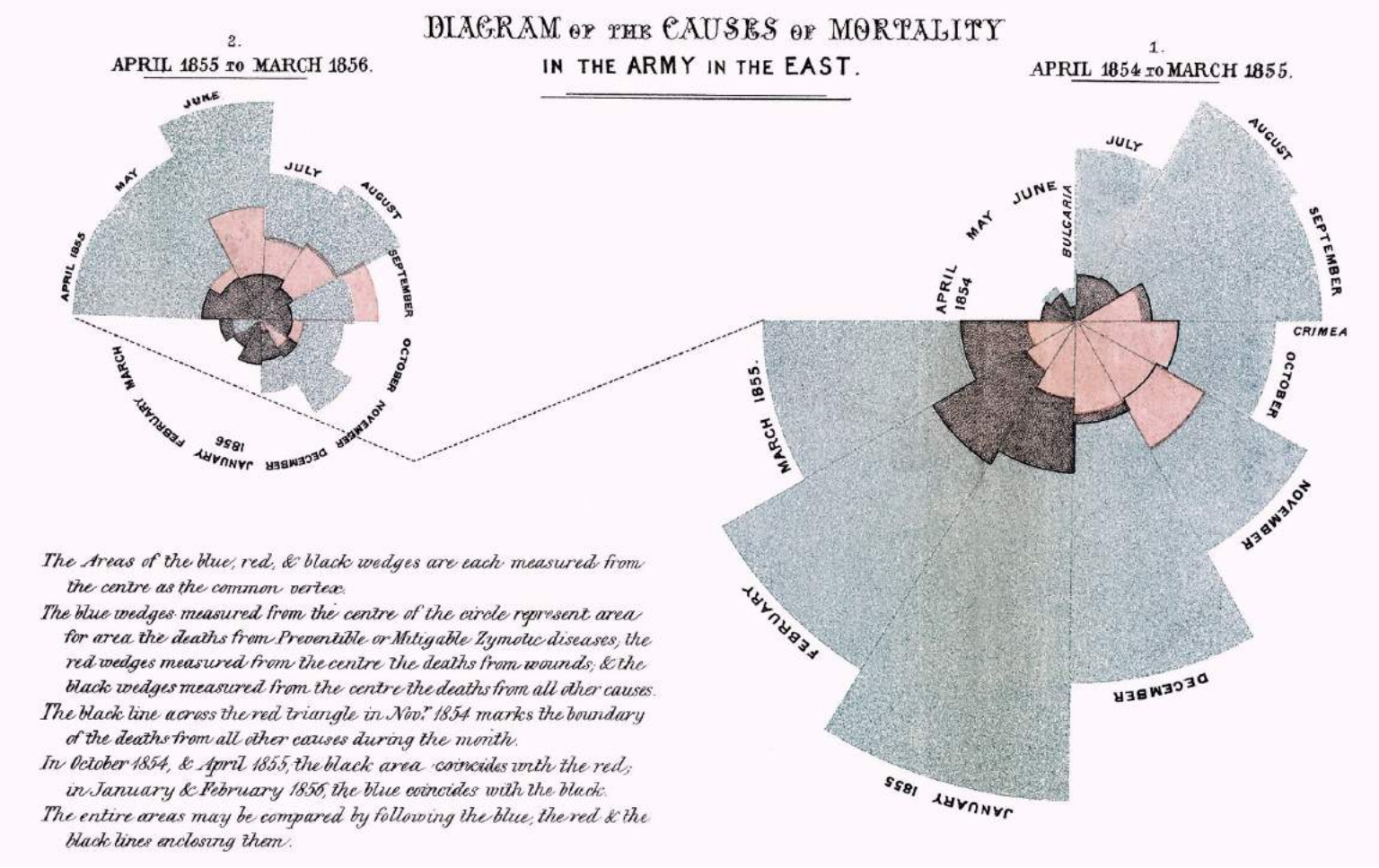
브라이언 체스키 에어비앤비 최고경영자(CEO)가 소개한 새로운 여행 트렌드다. 과거에는 목적지, 날짜를 넣고 검색을 시작했다면, 팬데믹 이후에는 이런 검색 조건을 넣지 않는다는 뜻이다. 에어비앤비는 올해 천국과 지옥을 오갔다. 지난 4~5월에는 예약이 80% 감소했다. 에어비앤비가 역사 속으로 사라질 것이라는 비관적 전망도 나왔다. 하지만 사람들은 에어비앤비를 다시 찾기 시작했다.

팬데믹 이후 호텔처럼 사람들이 붐비는 대형 숙박시설에 대한 거부감이 생긴 탓이다. 좀 더 '내 집'처럼 편한 곳을 찾기 시작했다. 이른바 '하이퍼 소셜(Hyper-Social·극단적 사회화)'에서 '하이퍼 솔로(Hyper-Solo·극단적 비접촉, 1인 전용)' 시대로 전환이다.

https://news.airbnb.com/wp-content/uploads/sites/4/2021/05/Airbnb-Report-on-Travel-Living.pdf

② 인사이트 도출 방법: 독창적 시각화 Visualization

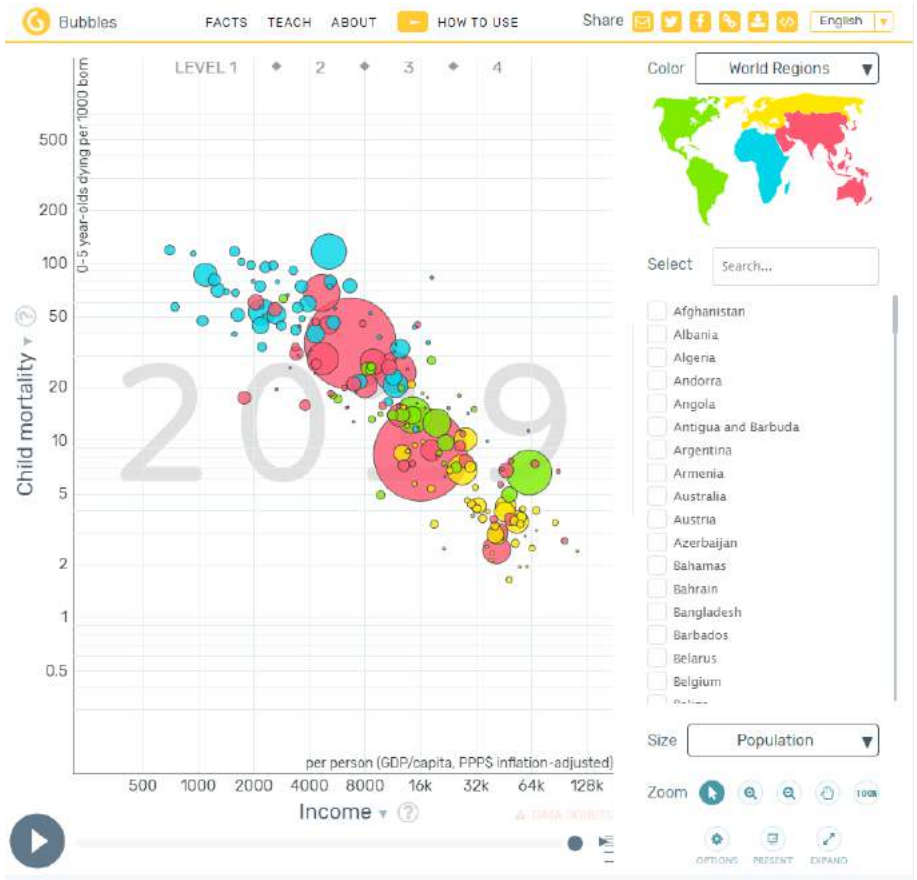
- “다쳐서가 아니라 더러워서 사람들이 죽어간다고!” (F. Nightingale, 19C)



https://www.tableau.com/learn/articles/best-beautiful-data-visualization-examples

독창적 시각화 (1) “animated bubble chart”

- “세상은 점점 나아지고 있다.” (Hans Rosling, TED, 2007)

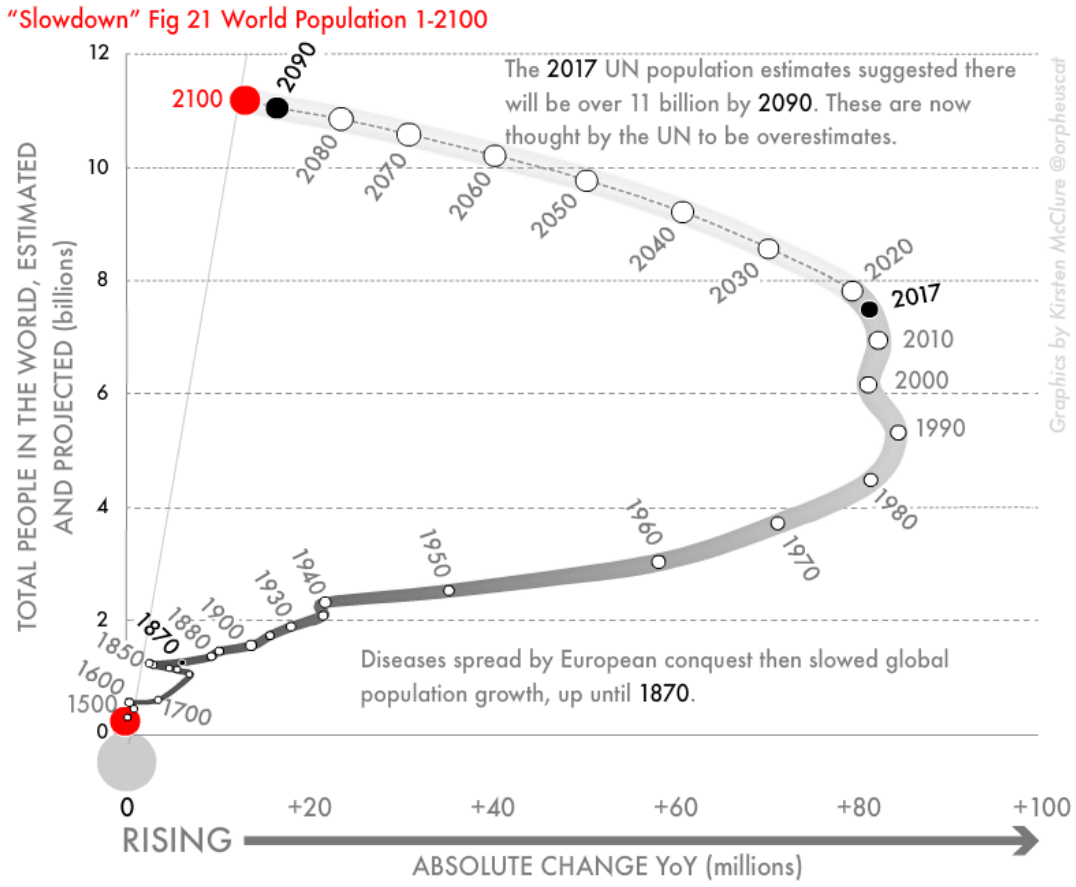


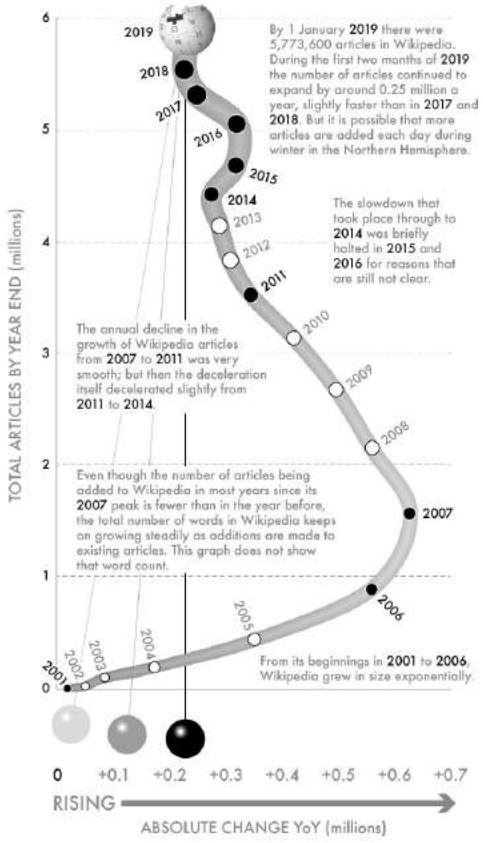
https://www.gapminder.org/tools/#$model$markers$bubble$encoding$y$data$concept=child_mortality_0_5_year_olds_dying_per_1000_born&space@=country&=time;;&scale$domain:null&zoomed:null&type:null;;;;;;&chart-type=bubbles&url=v1

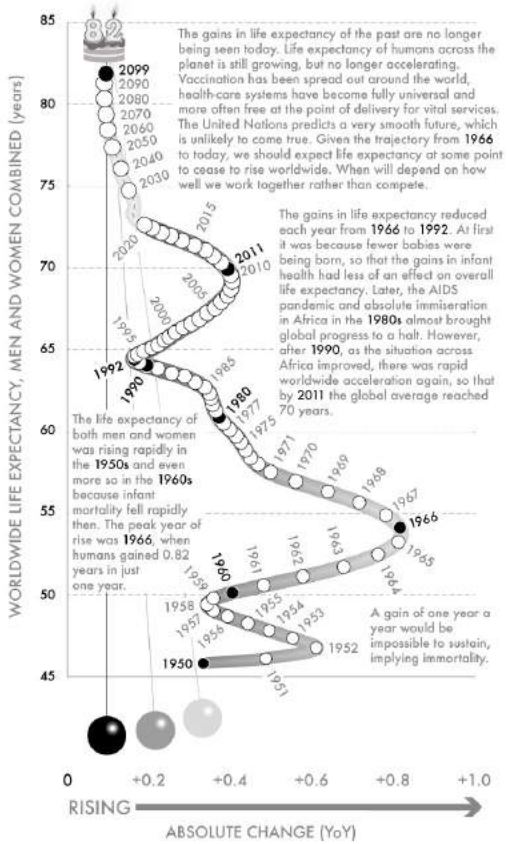
https://youtu.be/hVimVzgtD6w

독창적 시각화 (2) “slow down”

- “가속 성장의 시대는 끝났다.” (Danny Dorling, 2020)







http://www.dannydorling.org/books/SLOWDOWN/

나만의 인사이트 도출 방법

"송강호 배우는 답안지에 정답이 아닌 답을 적는데, 그게 더 정답일 때가 있다."

박찬욱 감독의 송강호 배우에 대한 평가 (https://bit.ly/2YsCNPv)

책에 나오는 분석 방법 반복

이 기법

저 기법

저렇게

이렇게

요렇게

대충 백만가지. 다 못함. 남들도 다 함.

세상에 없는 분석 방법 조합

이 기법 + 이 파생변수

+ 저렇게 자르기

+ 그 데이터

+ 요렇게 그리기

대충 수 억 가지. **나랑 같은 거 드물걸?**

어쩌면 기대 이상의 답일 수도 있음.

나만의 인사이트 도출 방법

- “남들보다 더 많은 시간과 정성 쏟기”

넷플릭스, 오징어 게임

그럼에도 불구하고, “이게 최선인가?”

1. 와 안 뭉쳐지는 모래로 이만큼 했다. 나 좀 짱인 듯.
   조개껍데기 걸러내느라 죽을 뻔.

2. 근데 옆집 철수는 더 잘 하던데. 걔가 했으면 어땠을까.
   어디서 상도 받고 인스타에 멋진 거 올라가 있던데.

3. 기사 보니까 모래성 찍기틀 새로 릴리즈 된 거 있던데.
   그거 쓰면 벽돌무늬 같은 거 낼 수 있던데.

4. 나 따위... 그저 빛을 가리는 먼지같은 존재일 뿐...

← 알고 보면 성장하는 과정

멋진 거.jpg

벽돌무늬 같은 거.jpg

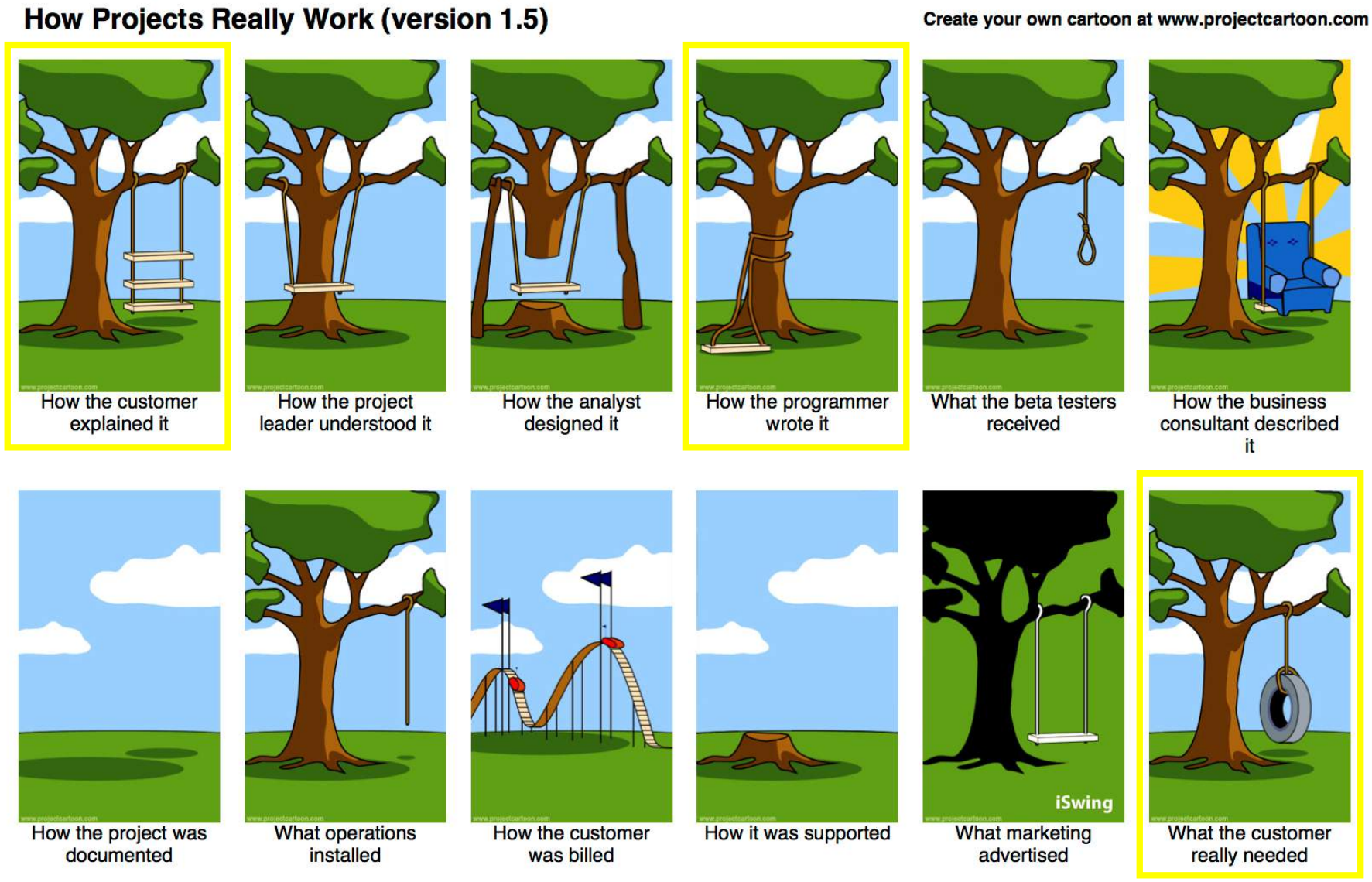
https://bluzelle.com/blog/managing-stored-data-load-is-like-building-a-sand-castle
https://tricycle.org/magazine/castles-made-sand/

https://watchandlearn.scholastic.com/videos/physical-science-and-engineering/physical-science/build-an-epic-sandcastle.html

3. 고객에게 잘 전달하기

- 분석이 궁금하신가,
  대안을 요구하시나?
- 이 분 금기어가 뭐더라?
- 결과물은 마음에 들어하시나?

소프트웨어 개발 프로젝트



실제 경험담 (2017)

- 빈 칸에 들어올 단어는?

Clouds are in the 

같은 결과도 상대방에 따라 다르게

Discussion

청자 : 나만큼 전문가

목적 : 더 나은 결과 도출

상세 : 최대한 자세히

특기사항 : 운명 공동체

손으로 대충 그린 그림

복잡한 수식

전문용어 난무

사내 업무보고

청자 : 세부사항 잘 모름

목적 : 매출 증대

상세 : 용건만 간단히

특기사항 : 내 고과 결정자

다시 사용하고 싶은 예쁜 그림

수식 = “복잡하지? 보지 마”

사내용어 난무

대중매체 게시

청자 : 문제부터 잘 모름

목적 : 先관심 後전달

상세 : 용건만++, 흥미 필수

특기사항 : 다른 볼거리 많음

일단 예쁜 그림

수식 = “나 잘났음 ㅋㅋㅋ”

첫째도, 둘째도, 셋째도 쉽게.

결과를 보고하는 데이터 분석가의 주의사항

1. 내 업무의 시간 순이 아니라 상대방의 논리에 따라 보고하기.
2. 결론 없이 사실만 나열하지 않기.
3. 경영 용어가 아닌 통계 용어를 남발하지 않기.

- **나쁜 예)**
  "상관관계 분석 결과 상관성이 약했고, 그래서 파생변수 OO를 만들고 DBSCAN으로 클러스터링을 해서 추가 분석을 했더니 상관성에 유의차가 확인되었는데 p-value는…"

- **좋은 예)**
  "웹페이지 개편 이후 매출이 N% 상승하였으며, ← 가장 알고 싶어하는 결론
  이는 OO 고객층의 재방문률 증가에 기인한 것으로 판단됩니다. ← 결론이 발생한 원인 해석
  OO 고객층 중에서도 XX 연령대의 유입이 늘었습니다. ← 조금 더 자세하게 설명 추가
  이들을 대상으로 한 마케팅 이벤트를 제안 드립니다." ← 나름의 제안. 거절당할 수 있음.

진짜 원하는 것 찾아내기

- “자신들이 원하는 게 뭔지도 정확히 모른다.” - 스티브 잡스

관리자	분석가
“우리 회사 고객 데이터를 분석해 주세요 ”	• 성별, 연령대별 통계치 뽑기 • 매출액 기준 고객 등급 분류 • 작년까지 고객 vs 올해 신규고객 수 비교 **… 다 했는데요?**
“작년보다 장사 안되는 거 누가 몰라요?”	

“충성 고객에게 집중해야 되나?”
“신사업을 발굴해야 되나?”
“공장을 이전해야 하나?”
“SNS 홍보를 강화해야 되나?”

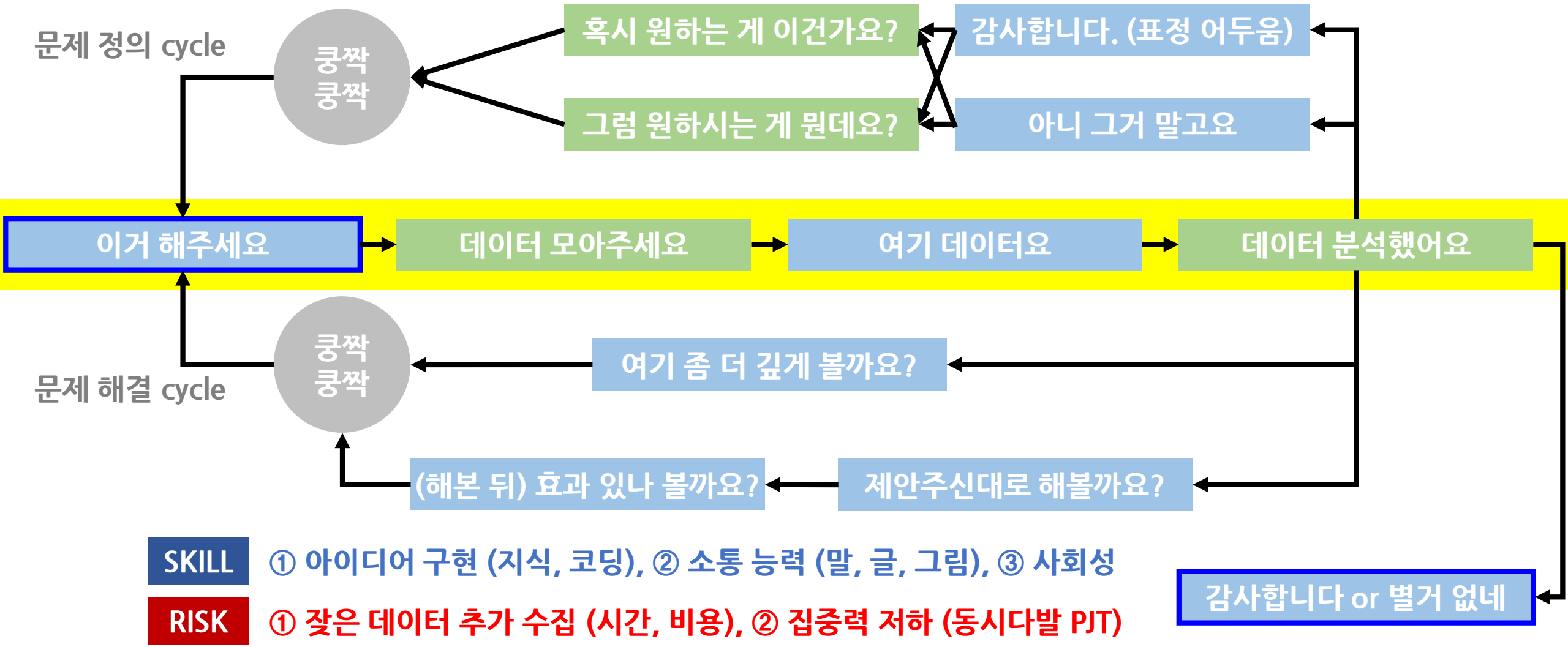
의사소통 책임소재
1. 명확하게 설명하지 않은 관리자
2. 목적을 물어보지 않은 분석가
비중은 다를 지 모르겠지만 결과적으로 쌍방과실

데이터 vs 도메인

- “우리 영문과는 영어가 모국어처럼 입에 붙어야
  비로소 국문과랑 같은 출발선에 서는 거야.” - 모 영어영문학과 교수님

- **초벌 데이터 분석으로 알아내는 것 = 도메인에서는 모두 알고 있는 것.**
  - (특) 역사가 길고, 책에 … 복잡한 업종 ~ 제조업
  - 데이터를 분석해서 “이… 리자에게 가져간다 **(X)**
  - 데이터를 분석해서 “이… 도메인 실무자에게 공유한다. **(O)**
    + 흥미를 보이는 지점이… 있는 부분을 캐치,
    심층 분석(인자간 관계,…)을 해서 관리자에게 가져간다.
    + 피드백을 주고 받으면… 로운 것을 알아낸다.

Agile process: 잦은 상호 Feedback



SKILL ① 아이디어 구현 (지식, 코딩), ② 소통 능력 (말, 글, 그림), ③ 사회성

RISK ① 잦은 데이터 추가 수집 (시간, 비용), ② 집중력 저하 (동시다발 PJT)

4. 끊임없이 나를 의심하기

- 왜 이 데이터를 뒤지고 있지?
- 이게 최선인가?
- 고객이 진짜 원하는 게 뭐지?
- 고객이 이 결과를 받아서 뭘 하길 바라지?

Keep in Mind

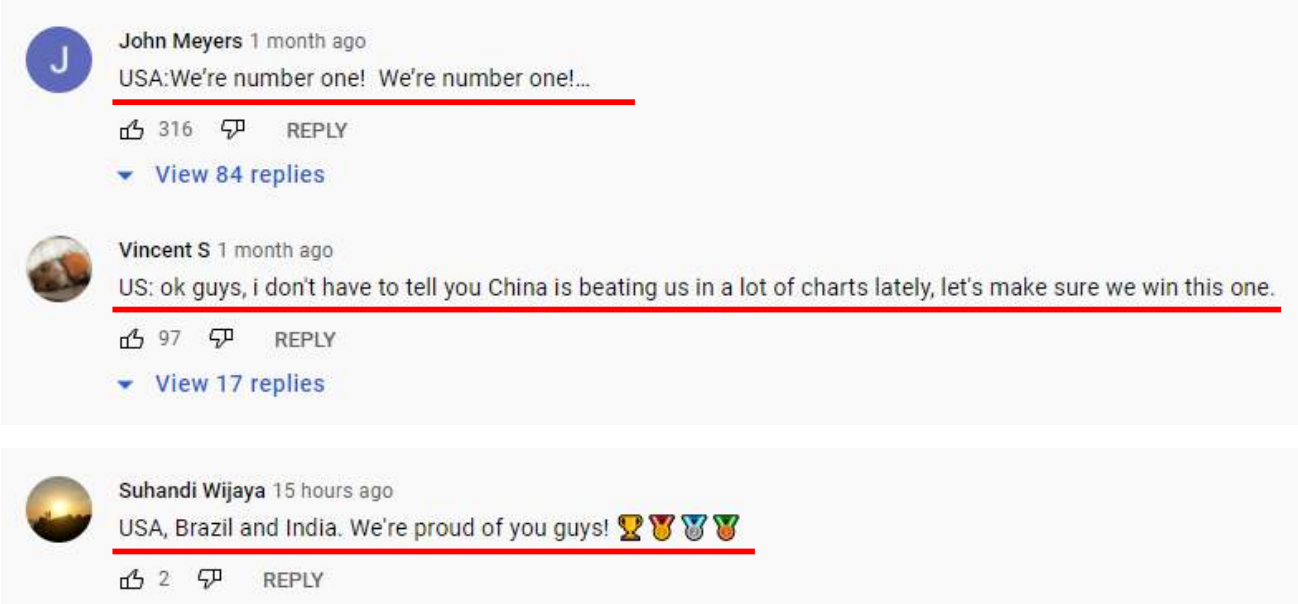
http://www.fsnews.co.kr/news/articleView.html?idxno=11722

- 내 분석 결과를 봐야 하는 이유 = insight
- 아이디어를 빠르게 적용하고 피드백 = agile
- 호기심을 어떻게 불러일으킬까? = 흥미요소
- 이 식당의 위생을 신뢰한다 **신뢰 : 기본**
- 이 식당의 음식을 신뢰한다 **품질 : 기본**
- 이성적, 감성적 충만함 = 진짜 목적 찾기

나는 데이터로부터 무엇을 찾고 있는가?

① 생명에 대한 예의

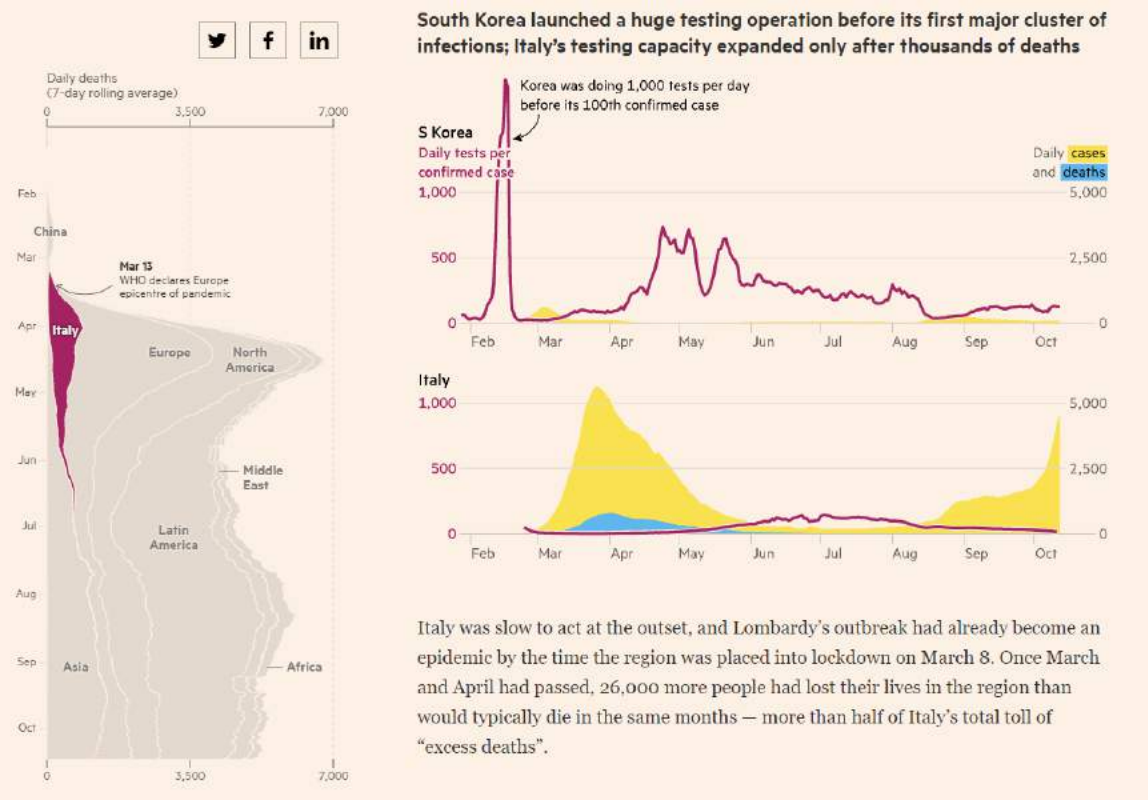
- 국가별 코로나19 사망자

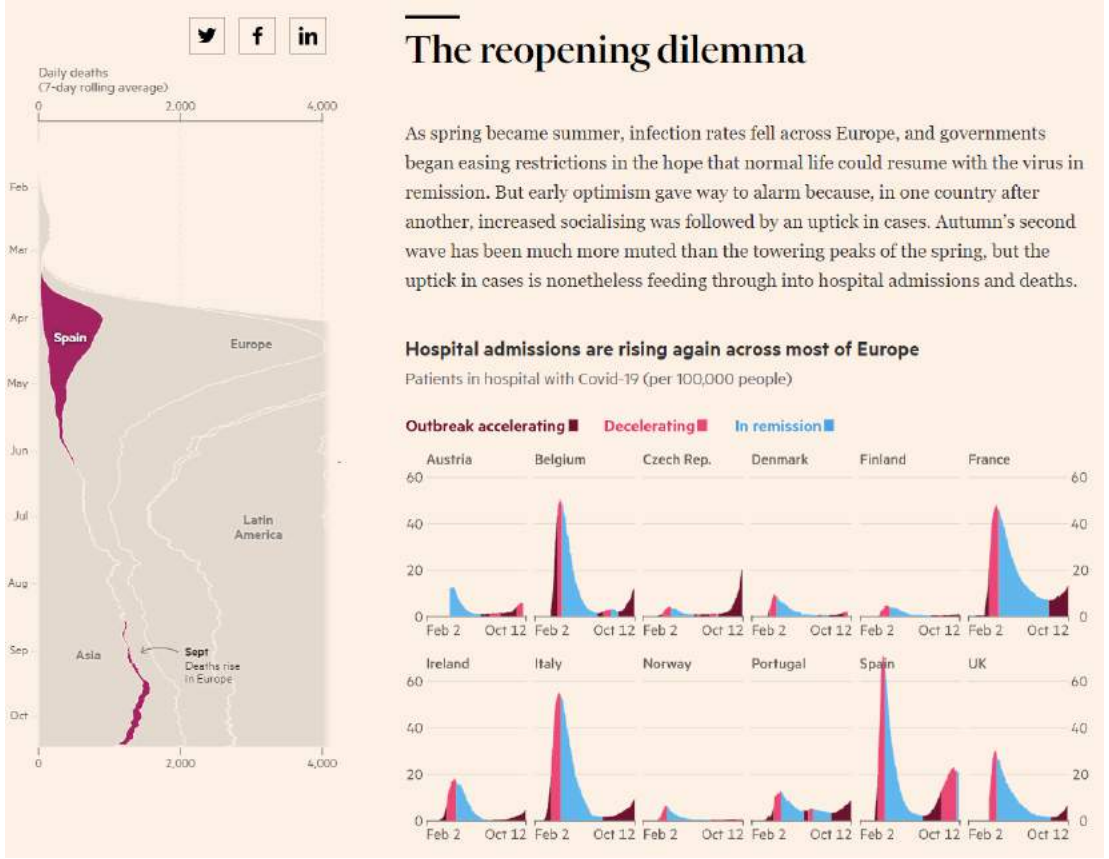


https://www.youtube.com/watch?v=Q1fm9qVls-0

① 생명에 대한 예의

- 코로나19 경과 정리



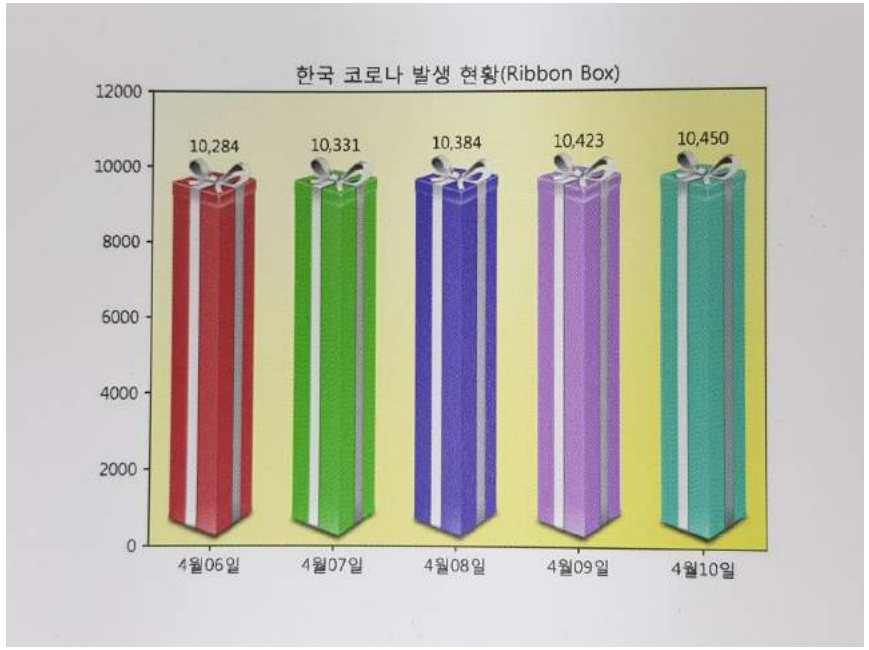


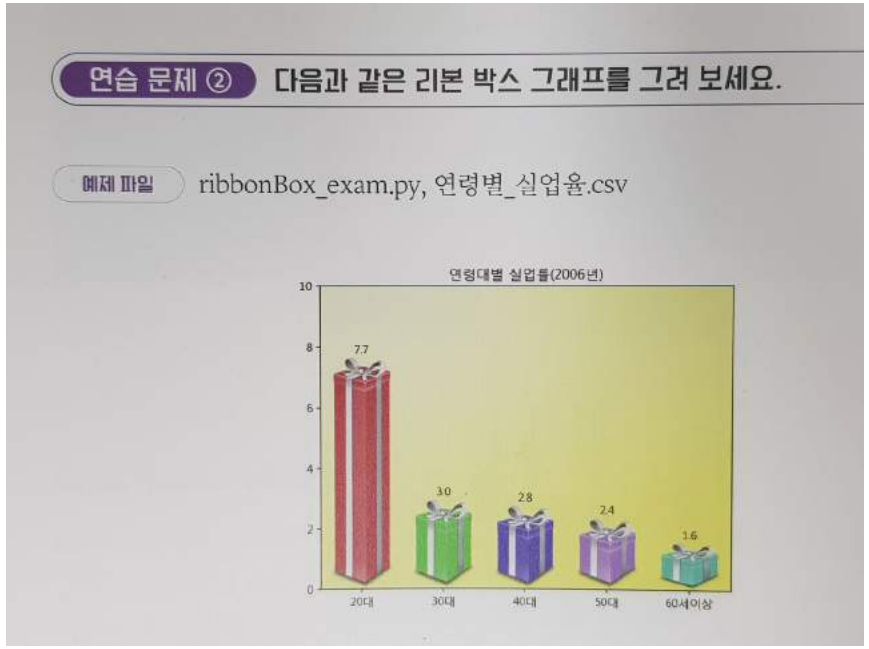
https://ig.ft.com/coronavirus-global-data/

① 생명에 대한 예의

- 코로나19, 실업률

https://matplotlib.org/stable/gallery/misc/demo_ribbon_box.html





굳이 서적의 이름은 밝히지 않겠습니다.

② 데이터 밖의 세계

- “이 양반들은 컴퓨터 속에서만 사나?” vs “어차피 숫자야”

현장의 중요성..

게시일: 2017년 8월 22일 Soo-Yong Shin님이 작성

지난 주에 정말 간만에 AI관련 학회인 KDD (Knowledge Discovery and Data Mining)에 다녀왔다. 한동안 의료정보 관련학회랑 ISO표준 미팅만 가다가.. 이쪽에 간 건 거의 10년만인가? 앞으론 몇 년에 한번이라도 CS 전공 학회를 하나씩 가야할 듯 하다.. 내년엔 GECCO나 가볼까..

KDD에서 학술적인 내용보다 더 크게 느낀 건..

“현장에 답이 있다”

라는 거다.

KDD에도 많은 Healthcare 관련 논문들이 나왔는데.. 솔직히 2~3편을 빼곤.. 임상현장에서 유의미하게 쓰일 수 있는 건 거의 못 봤다. 심지어 artificially generated data를 가지고 쓴 논문이 oral presentation을 하는 정도였다.. 나중에 물어보니 정말 현장을 모르더라.. 지금 setting이 좀 비현실적이라고 했더니 자기로썬 그게 최선이었다고.. MIT에서 나온 논문은 젤 중요한 결과를 생략하고 발표해서 황당했는데.. 나중에 poster도 있어서 물어볼려고 한참을 기다렸는데.. poster만 붙여놓고 안 나타나더라 -_-.. 만약 내가 Reviewer였다면 대부분의 논문은 reject인 논문들이었는데..

어느 닭 농장에서 닭들이 거품을 물고 쓰러지기 시작했다.
그래서 닭 농장 주인은 생물학자, 화학자, 물리학자를 각각 불렀다.
먼저 생물학자가 와서 죽은 닭 한 마리를 챙기더니 "1주일 동안 조사해본 뒤 결과를 보내겠습니다." 라며 떠났다.
그 다음 화학자가 와서 죽은 닭의 피를 뽑더니 "1주일 동안 조사해본 뒤 결과를 보내겠습니다." 라며 떠났다.
그 다음 물리학자가 와서 죽은 닭을 자세히 관찰했다.
그러더니 방 안에 들어가 종이를 펴고 여러가지 계산을 하더니 30분 뒤 농장 주인에게 다가왔다.
"이렇게 하면 닭이 병에 걸리지 않을 것입니다.
다만 진공 상태의 구형 닭(어느 방향에서 봐도 닭)에만 적용됩니다.

https://sooyongshin.wordpress.com/2017/08/22/%ED%98%84%EC%9E%A5%EC%9D%98-%EC%A4%91%EC%9A%94%EC%84%B1/
https://namu.wiki/w/%EA%B3%B5%EB%8C%80%EA%B0%9C%EA%B7%B8

맺음말: 식사를 마친 손님

저희 식당은 프랑스 OO에서 수학하고 뉴욕 OO의 OO 레스토랑에서 다년간 경험을 쌓은 셰프가 손님 여러분들의 건강을 위해 토종닭과 최고급 버터를 토마토 페이스트와 함께 넣고 끓인 스튜를 대표 메뉴로 하여 MSG는 일절 첨가하지 않음으로써…

“아우, 여기 맛이 왜 이래?”

“한 시간 기다렸는데 언제 나와요?”

“알러지 있으니까 계란 빼 달라고 했죠? 근데 왜 굳이 넣으세요?”

주문하고, 음식이 제때 나오고, 맛있고 배부르다.
주인장은 계속 고민하고, 손님은 점점 더 만족한다.
음식을 어떻게 만들었는지는 잘 모르겠다.
궁금해지면 계산하면서 여쭤 보기도 한다.

“맛있게 잘 먹었습니다.”

“여기 음식 잘 하네요?”

“내가 맛집 한 군데 아는데 갈래?”

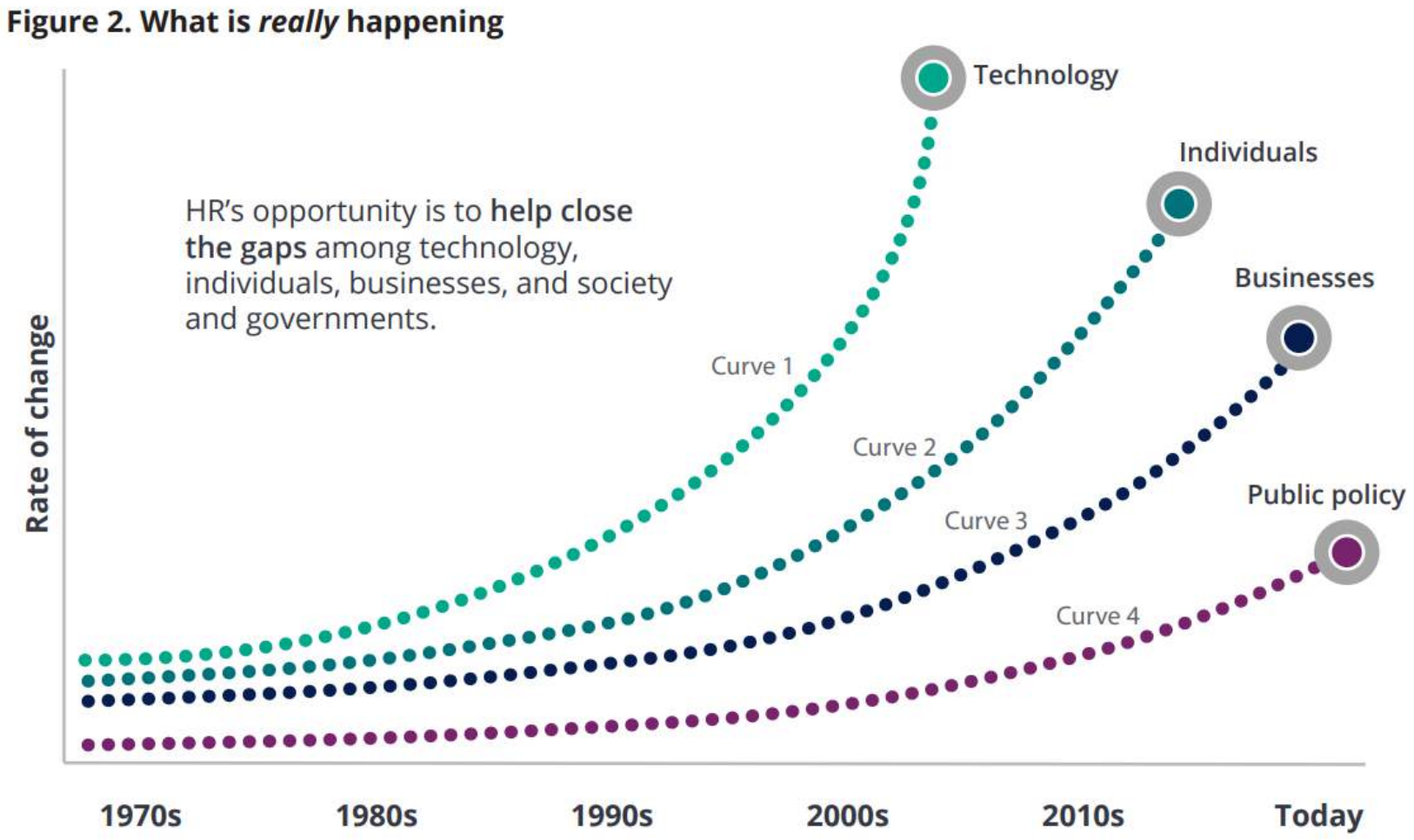
사전질문

1. 데이터분석 4차산업혁명의 발전 어떻게 이루어지나요?

- 적지 않은 분야가 현재 2차 산업혁명 상태라고 생각합니다.
  - 분야마다 다른 속도로 찾아오고 있다고 느낍니다.

https://www.daejeon.go.kr/drh/DrhContentsHtmlView.do?menuSeq=6700

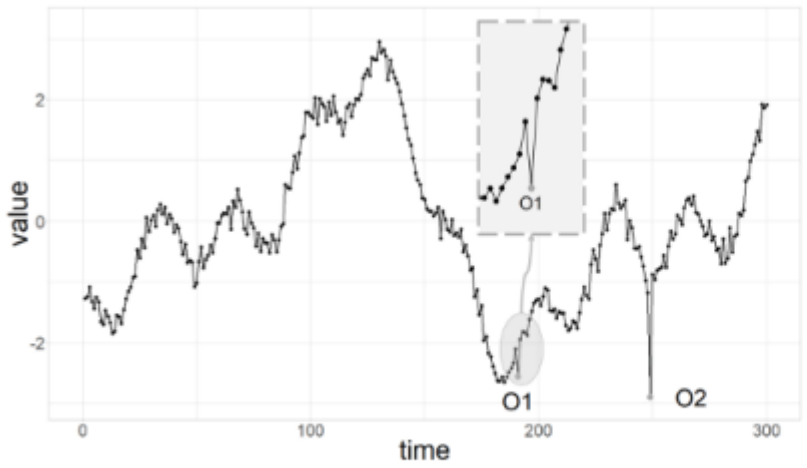
1. 데이터분석 4차산업혁명의 발전 어떻게 이루어지나요?



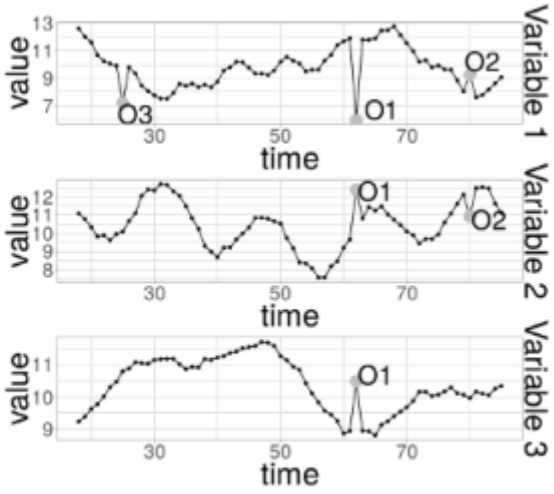
https://sloanreview.mit.edu/article/digital-disruption-is-a-people-problem/
"Rewriting the rules for the digital age", Deloitte Global Human Capital Trends (2017)

2. 수많은 로그들로부터 이상치(anomaly) 검출을 위해 학습 전 데이터 정제 과정에서 유용한 시각화 방법이 있다면 무엇이 있는지 궁금합니다.

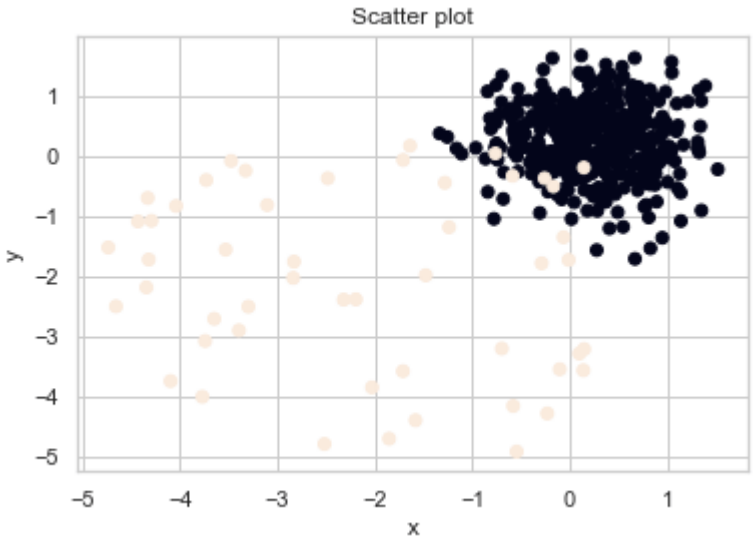
- 시계열 데이터를 거의 다루어 보지 않았습니다.
- 아래 글을 추천드립니다. 단변수 plot부터 오토인코더까지 소개합니다.
  - https://neptune.ai/blog/anomaly-detection-in-time-series



(a) Univariate time series.



(b) Multivariate time series.

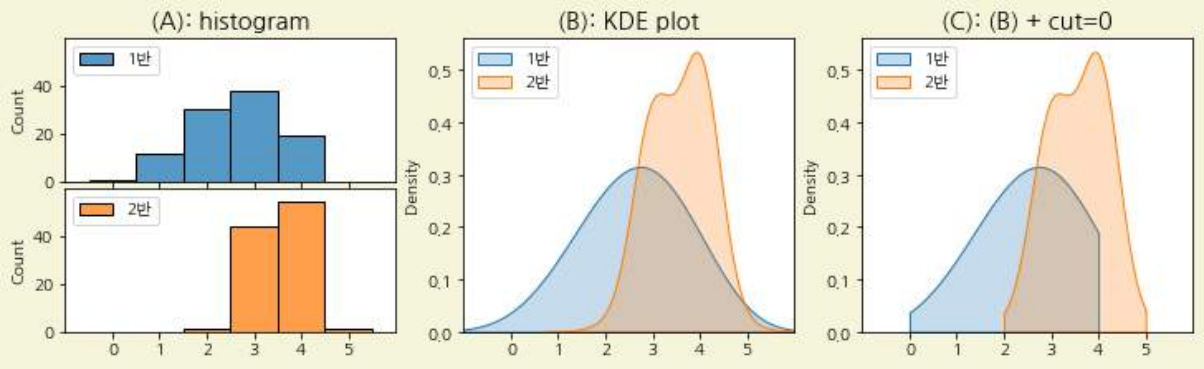


3. 데이터 분석 실무에서 자주 빠지게 되는 혹은 유의해야하는 함정은 무엇이 있을지 궁금합니다!

- 저는 목적을 잊어버리고 기법에 매몰된 경우가 많았습니다.
- 데이터 분석의 목적을 의도적으로 상기하면서 분석하려 합니다.
- 데이터에 갇히는 것도 유의해야 할 것입니다.
- 데이터를 해석하기 위해 데이터 밖의 정보를 적극적으로 구해야 할 때가 많습니다.

4. 데이터의 왜곡된 해석을 최소화하려면 어떻게 해야 하는가요?

- 저는 도메인 지식을 정확히 알고 분석기법의 개념을 확립하려고 노력합니다.
- 어떤 기법이든 장점을 활용해야 겠지만 늘 단점을 염두에 두려고 합니다.
  - 예를 들어, KDE plot은 데이터 값에 Gaussian Kernel을 씌워 확률 분포를 추정합니다.
  - 이 때 존재하지 않는 데이터 범위가 생성되는데 때에 따라 치명적일 수 있습니다.
  - seaborn.kdeplot()의 경우 cut=0 옵션으로 예방할 수 있습니다.



  - 그러나 이런 옵션을 줄 수 없는 경우가 많습니다. 여러 방식으로 그려보고 정리하면서 모순점을 찾으려고 노력합니다. 눈으로 본다는 것은 생각이 실체화된다는 면에서 중요합니다.

5. 데이터 분석 직무를 희망하는 비전공자인데 대학원이 필수일지 궁금합니다!

- 산공, 통계 등 대학원을 말씀하시는 것으로 판단됩니다.
- 어떤 조직이건 고유의 목표가 있고, 해당 분야의 전문가로서 데이터 분석을 수행하는 편이 장점이 많다고 생각하고 있습니다.
- 흔히 '비전공'이라고 표현하는, 데이터를 주 업으로 하지 않는 전공도 훌륭한 데이터 분석가가 될 수 있다고 생각합니다. 실제로 문과 출신 능력자도 많고요.
- 그러나 스스로 두각을 나타내기 어렵다면 대학원을 거치는 편이 확률적으로 유리하다고 생각합니다.
- 다만 이 경우 대학원 진학에 따른 기회비용(시간, 소득 상실)까지 판단에 포함하시기 바랍니다.

6. 최근에 데이터분석과 관련한 교육과 강의가 많아 수강중인데 이게 큰 도움이 될지도 궁금합니다!

- 요즘 데이터 분석 관련해서 학습자료가 넘쳐나고 있습니다.
- 그러나 모든 공부가 다 그렇지만 수업을 듣는다고 다 도움이 되진 않습니다. 듣는 것만으로도 도움이 되면 모두 다 수능부터 만점을 받아야 할 겁니다.
- 강의를 듣는 것보다 스스로 해서 손가락에 코드를, 눈썹에 개념을 스며들게 하는 것을 추천드립니다.
- 그리고 강의를 하나 들으면 그 시간 만큼 다른 것(다른 강의, 운동, 인간관계, 자율학습)을 포기할 수 밖에 없습니다. 무턱대고 듣는다고 도움이 되지 않을 것은 자명합니다.
- 자신의 갈 길을 다듬어간다는 생각으로 학습 아이템을 결정하시면 좋겠습니다. 예를 들어 저는 몇 년 전 공부를 시작할 때 과감히 딥러닝을 미뤘다가 다른 것들보다 늦게 시작했습니다.

7. 데이터분석 직무로 취업하기 위해서 어떤 파트를(파이썬, r, 머신러닝 등) 집중적으로 공부해야하며 어떠한 스펙이 있어야 할까요(현실적으로 궁금합니다)

- 말이 데이터분석이지 그 안에 채워야 할 일들이 너무 많습니다. 백엔드부터 웹의 프론트엔드까지, 통계학부터 딥러닝까지, 파이썬이라면 print부터 함수, 클래스, 데커레이터 등 열거하기도 어렵습니다.
- 목적을 정하고 거기에 필요한 기술을 역으로 찾아보는 것이 좋을 것 같습니다. 예를 들어 머신러닝 주식거래 알고리즘을 만들어 큰 돈을 벌고 싶다면 그 길을 가고 있는 다른 분들을 어깨 너머로 보면서 따라해볼 수 있습니다.
- 취업 스펙은 회사마다 다를 겁니다. SQL이 필수인 곳, 태블로가 우대받는 곳, R은 필수이지만 파이썬은 필요 없는 곳과 그 반대인 곳이 있을 겁니다. 내가 택한 분야의 한 회사를 정하고 거기에 맞춘다는 생각으로 알아보시면 어떨까요? 커뮤니티 활동을 통해서 그 회사 분을 알게 되면 더 좋을 겁니다.