

예쁜 막대기 그리기 ?

[시각화, 데이터로 이야기 하는 법]

2021. 06. 03 Jiseong Kim.

Intro.

오늘 할 이야기는 ..

휘황찬란한 그래프 ? 알록달록하고 이쁜 그래프 ?
저도 잘 못합니다 :) 알려주세요

과제를 그지같이 했지만
어쨌든 제출한 게 자랑스러울 때



Intro.

오늘 할 이야기 !

데이터를 이해하기 위한 방법 ‘시각화’

데이터로 이야기하는 가장 직관적인 도구 ‘시각화’

(+) 그리고 이런것도 있습니다..ㅎㅎ

데이터 분석 ?!

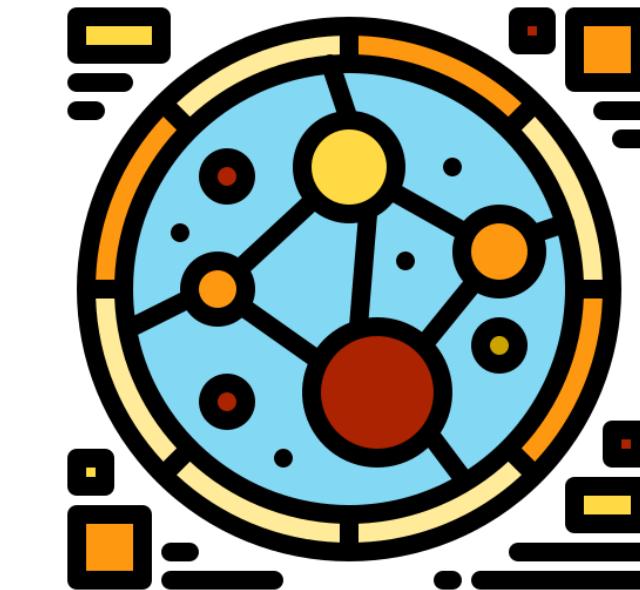
이상과 마주하게될 현실

0110
1001
1010

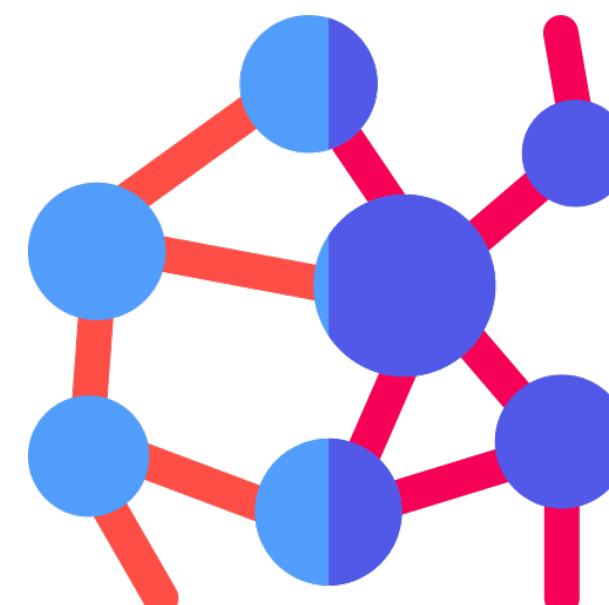
(잘 정돈된 데이터)



(데이터 분석)



(분석 결과)



(복잡하게 연결된 데이터)

0110
1001
1010

(전치리를 거친 데이터)



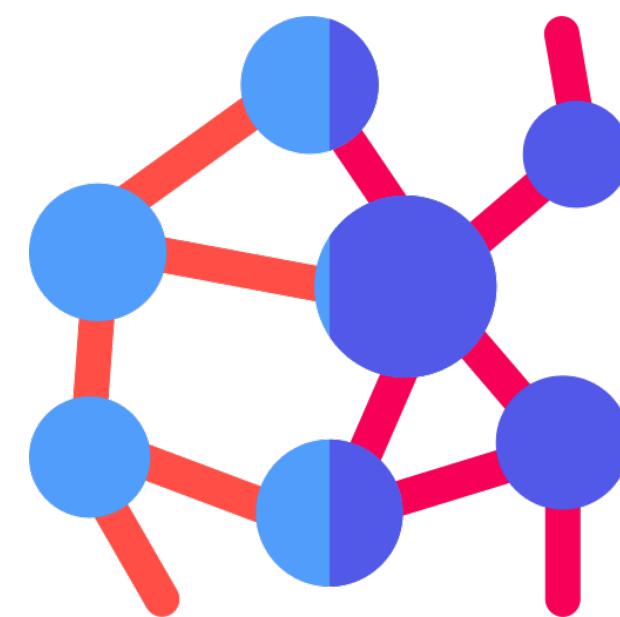
(탐색적 자료조사)



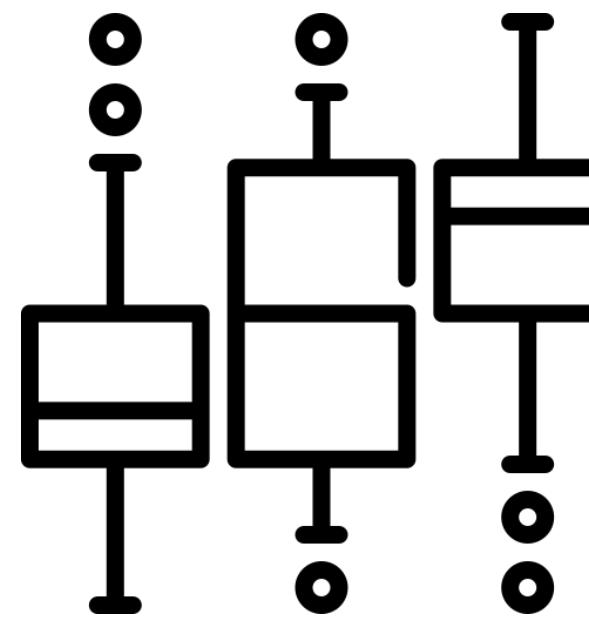
(분석 모형)

(프로젝트 전체 80%)

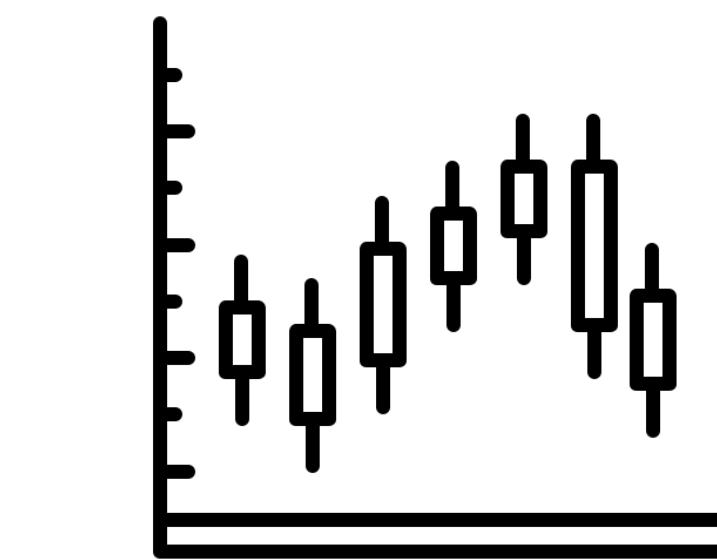
데이터를 이해하는 과정 ‘EDA’



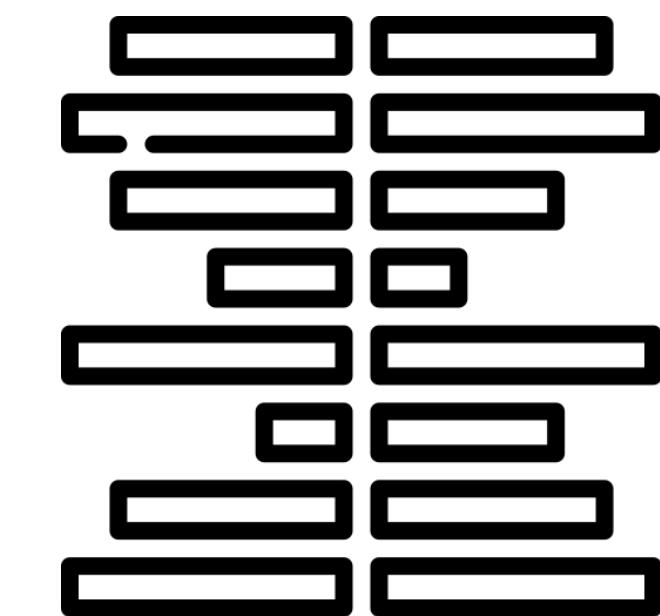
(복잡하게 연결된 데이터)



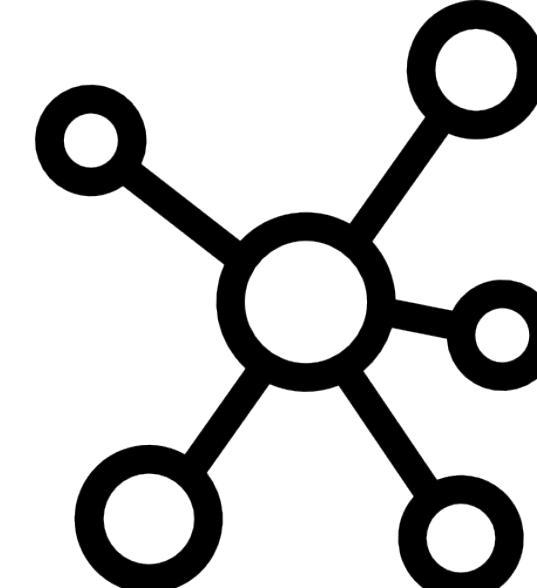
(이상치 파악)



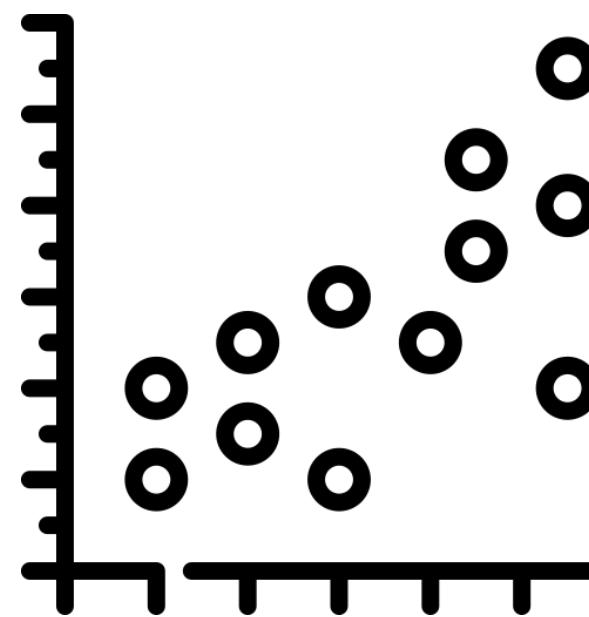
(데이터 분포 파악)



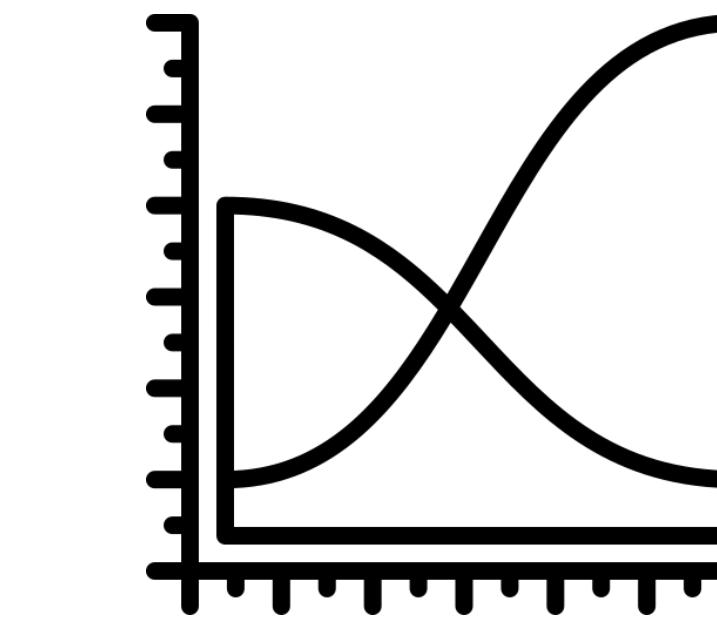
(변수 비교)



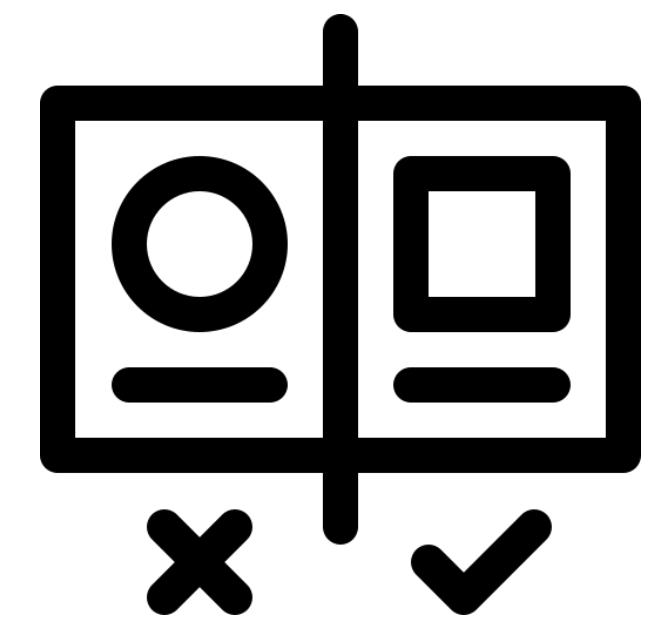
(연관성 파악)



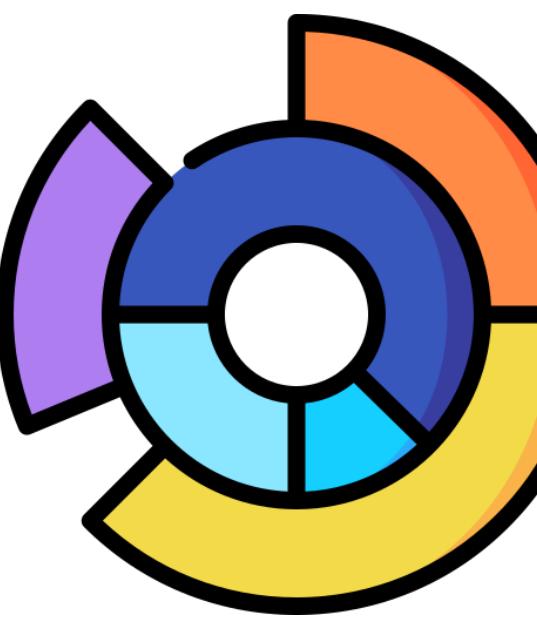
(상관관계 파악)



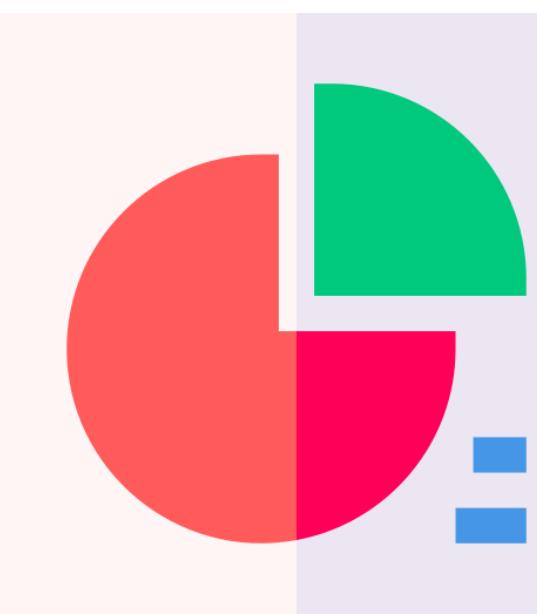
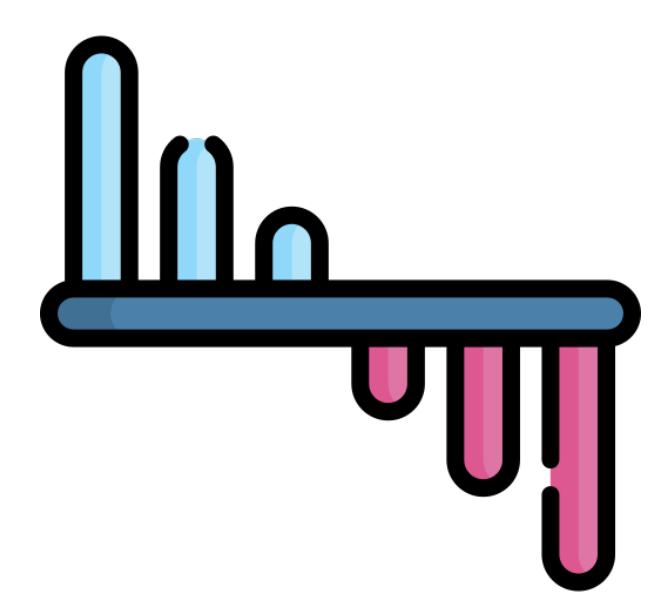
(변수 관계 파악)



(집단간 비교)



(범주간 비교)

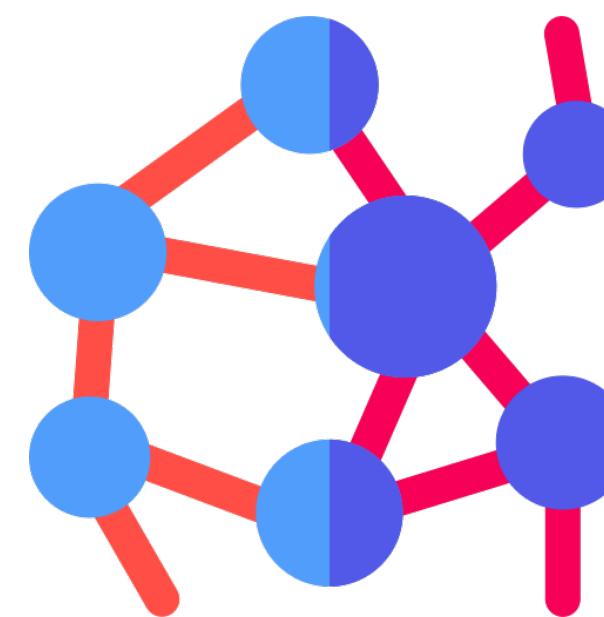


데이터를 이해하는 과정 ‘EDA’

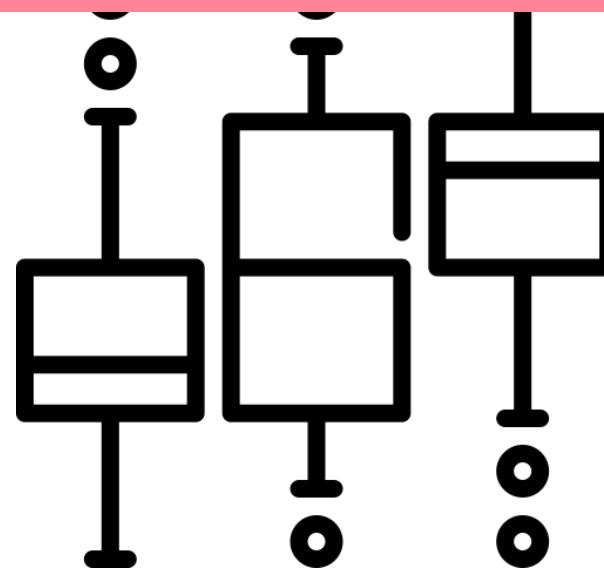
Q : 그래프는 그렸는데 뭘 어떻게 하죠?
A : 설명을 위해 ‘데이터’를 의인화 해보죠!

데이터를 이해하는 과정 ‘EDA’

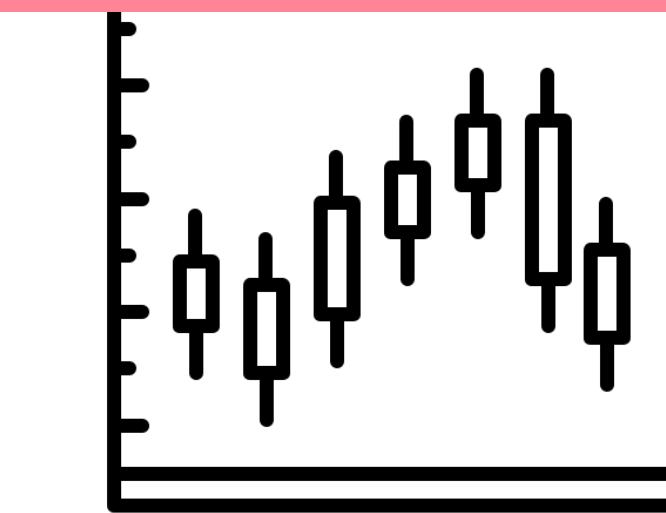
(사람으로 치면 스타일, 성격, 성향, 인성 등등 을 파악하는 일)



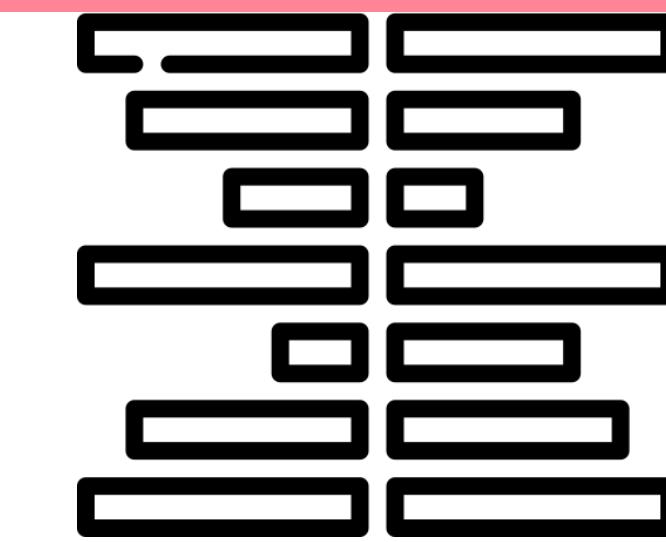
(복잡하게 연결된 데이터)



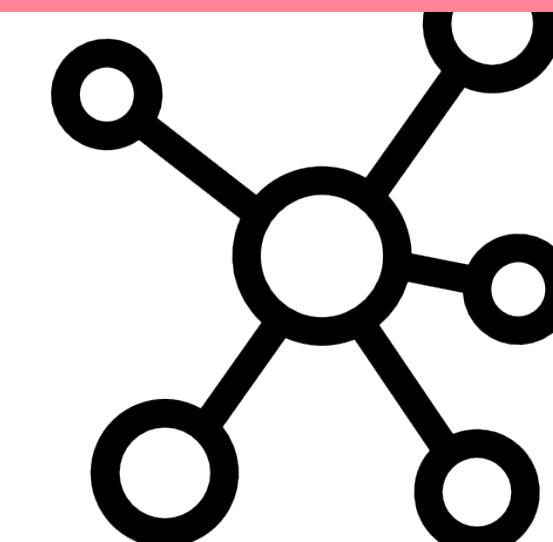
(이상치 파악)



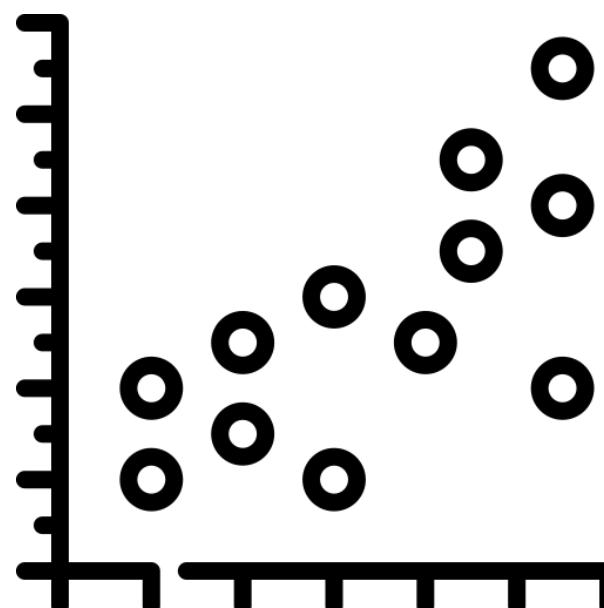
(데이터 밀도 파악)



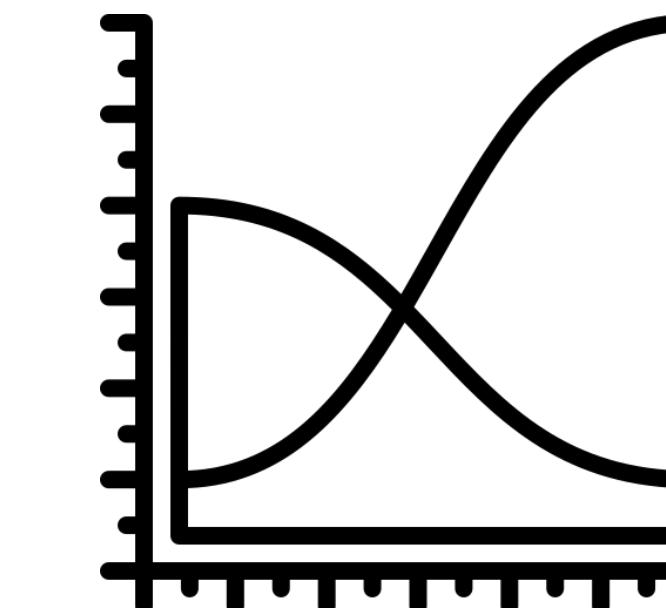
(변수 비교)



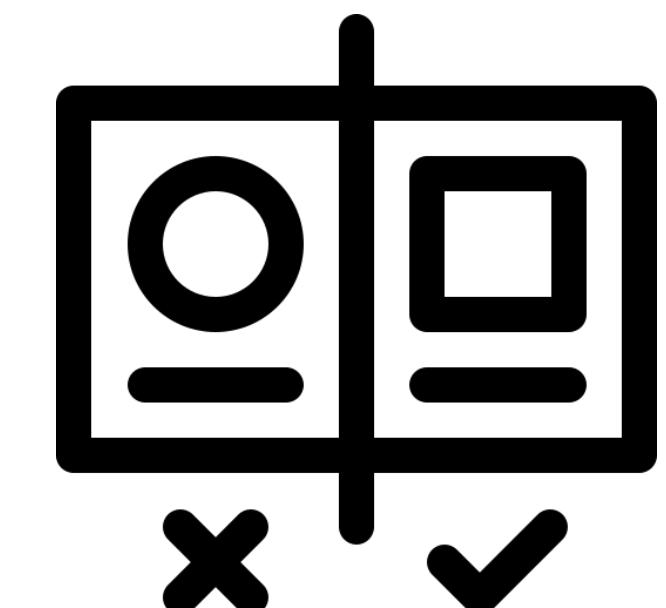
(연관성 파악)



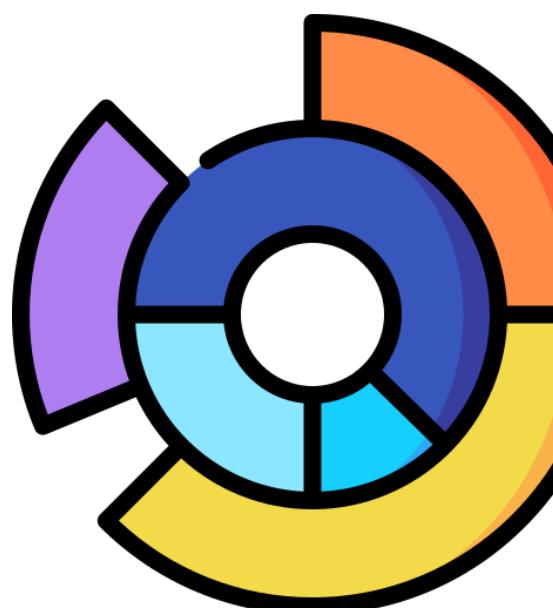
(상관관계 파악)



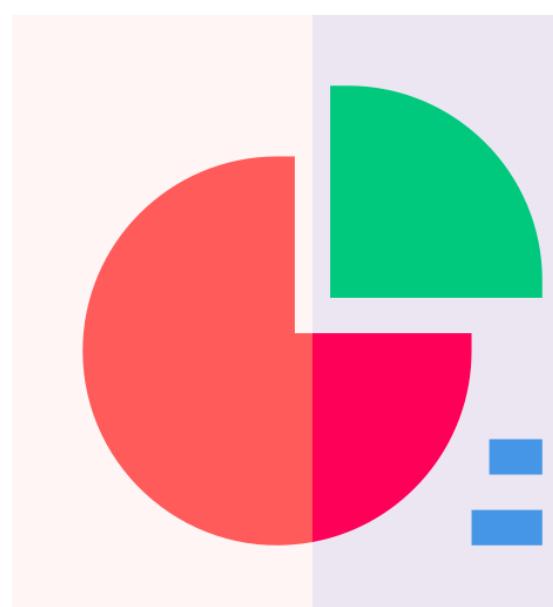
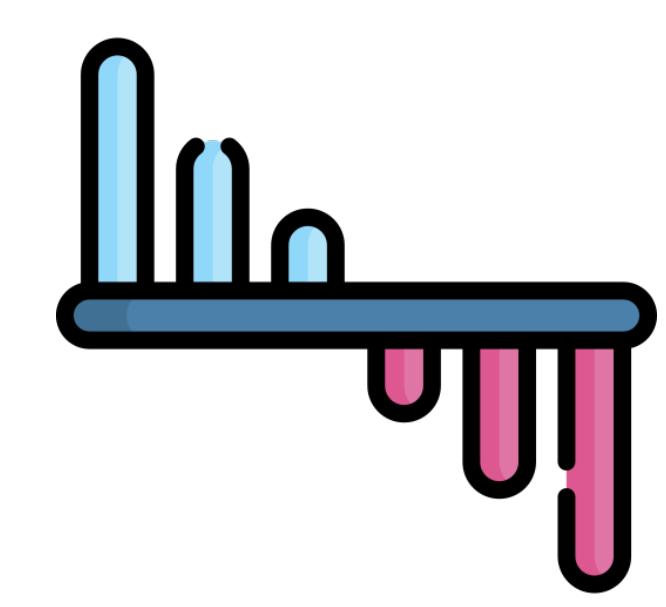
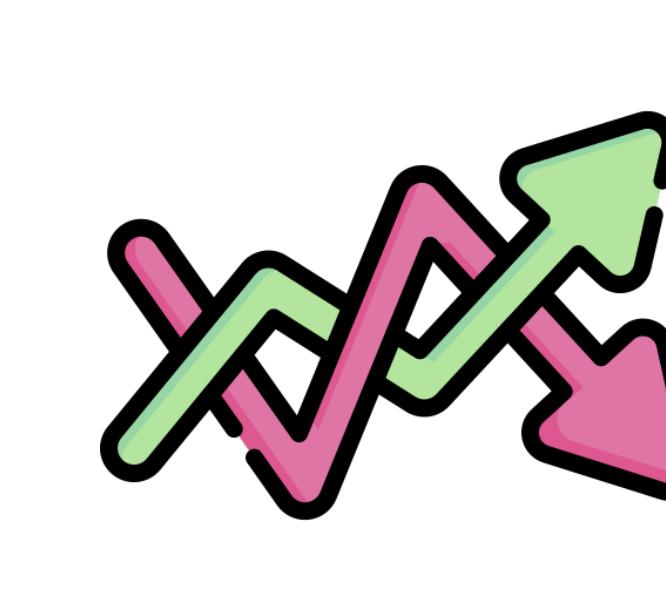
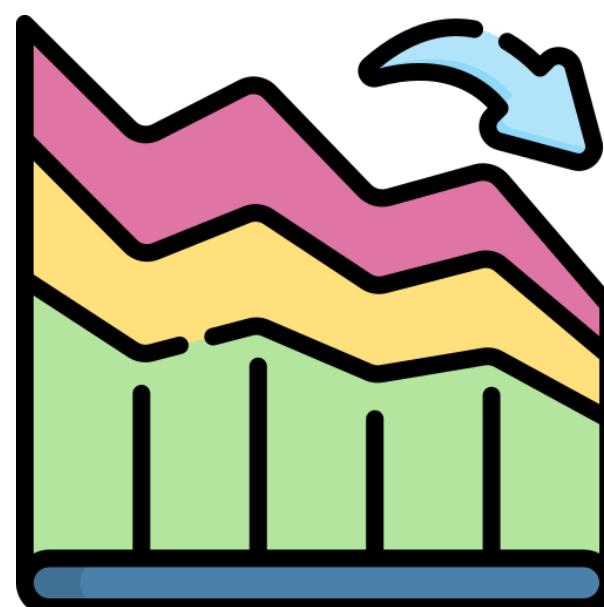
(변수 관계 파악)



(집단간 비교)



(범주간 비교)



데이터를 이해하는 과정 ‘EDA’

Q : 저렇게 많은걸 다 해봐야 하나요 ?

A : 데이터를 이해하기에 왕도는 없습니다
계속 데이터에게 질문을 던져보는 수
대신 직관적인 시각화는 도움이 되죠 !
그리고 다양한 사람들이 질문 던지는
방법을 보세요 !

다양한 사람들이 질문을 던지고 답하는 방법 ?!

Kaggle

The screenshot shows the Kaggle homepage with a sidebar on the left containing links to Home, Competitions, Datasets, Code, Discussions, Courses, and More. The main area features a competition titled 'CommonLit Readability Prize' with a banner image of books and a prize of '\$60,000'. Below the banner, there are tabs for Overview, Data, Code (which is selected), Discussion, Leaderboard, Rules, and Team. A 'New Notebook' button is visible. A search bar at the top and another one for notebooks are present. The main content area shows recent activity, including a post from 'CommonLit Readability Prize EDA' and another from 'Scraping Data Augmentation'.

DACON

The screenshot shows the DACON homepage displaying five AI competition entries. 1. '구내식당 식수 인원 예측 AI 경진대회' by LH, with a total prize of 700만 원. 2. '전력사용량 예측 AI 경진대회' by 한국에너지공단, with a total prize of 1,800만 원. 3. '영어 음성 국적 분류 AI 경진대회' by DACON, with a total prize of 100만 원. 4. '북극 해빙예측 AI 경진대회' by KOPRI, with a total prize of 600만 원. 5. '동서발전 태양광 발전량 예측 AI 경진대회' by 한국동서발전(주), with a total prize of 100만 원. Each entry includes details like category, sponsor, deadline, and number of participants.

So.prize

The screenshot shows the So.prize homepage with a large image of a squirrel. The main text reads '데이터와 글쓰기에 진심인 당신에게 매일 1,000,000 KRW를 드립니다.' Below it, it says '작성중인 답변 236 개, 제출된 답변 73 개' and '누적 상금 지급액 34,000,000 원'. A pink button labeled '더 알아보기' is visible. At the bottom, there's a section titled '쏘프라이즈에 참여하세요' with a note about the purpose of the competition.

다양한 페이스북 커뮤니티

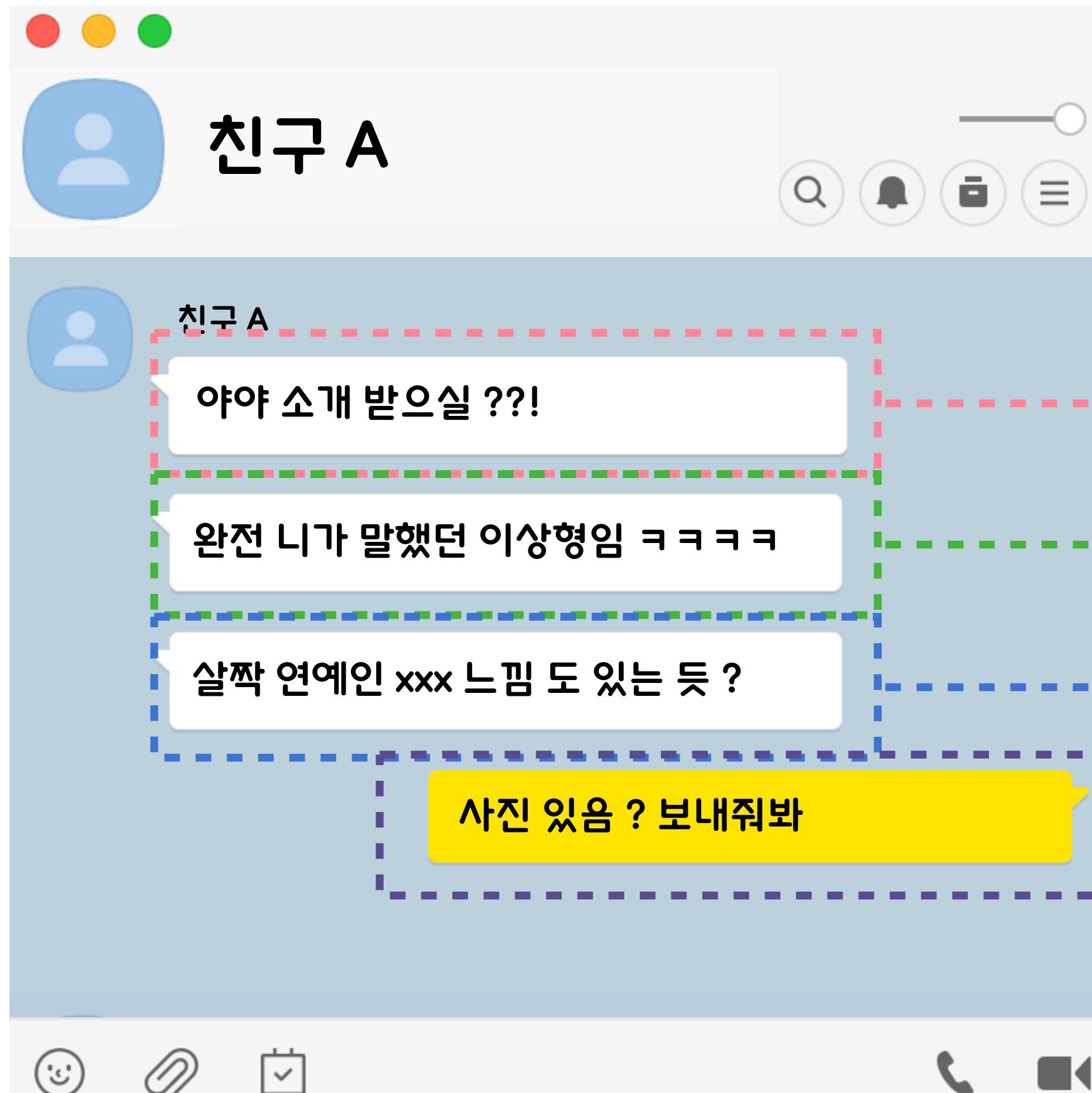
The screenshot shows a Facebook page listing several groups: 1. 'TensorFlow KR' (1개 게시물) 2. 'Pseudo Lab' (마지막 활동: 4일 전) 3. '딥리워드' (마지막 활동: 1일 전) 4. '데이터 분석 커뮤니티' (1개 게시물) 5. 'GNN KR' (마지막 활동: 1일 전) 6. 'PyTorch KR' (1개 게시물) 7. 'Python' (마지막 활동: 약 1시간 전) 8. '데이터 피플 코리아 커뮤니티' (마지막 활동: 5일 전) 9. 'Reinforcement Learning KR' (마지막 활동: 32분 전) 10. '데이터뽀개기' (마지막 활동: 19시간 전) 11. '통계분석연구회(Statistics Analysis Study)' (마지막 활동: 3시간 전) 12. 'NLP KR' (마지막 활동: 1주 전) 13. 'Python Korea' (마지막 활동: 10시간 전). Each group has its logo and a brief description.

데이터를 이해하는 과정 ‘EDA’

Q : EDA 대충 감은 와요 그럼 왜 시각화죠?

A : 다른 예를 하나 들어보죠!

데이터를 이해하는 과정 ‘EDA’



▶ 새로운 데이터

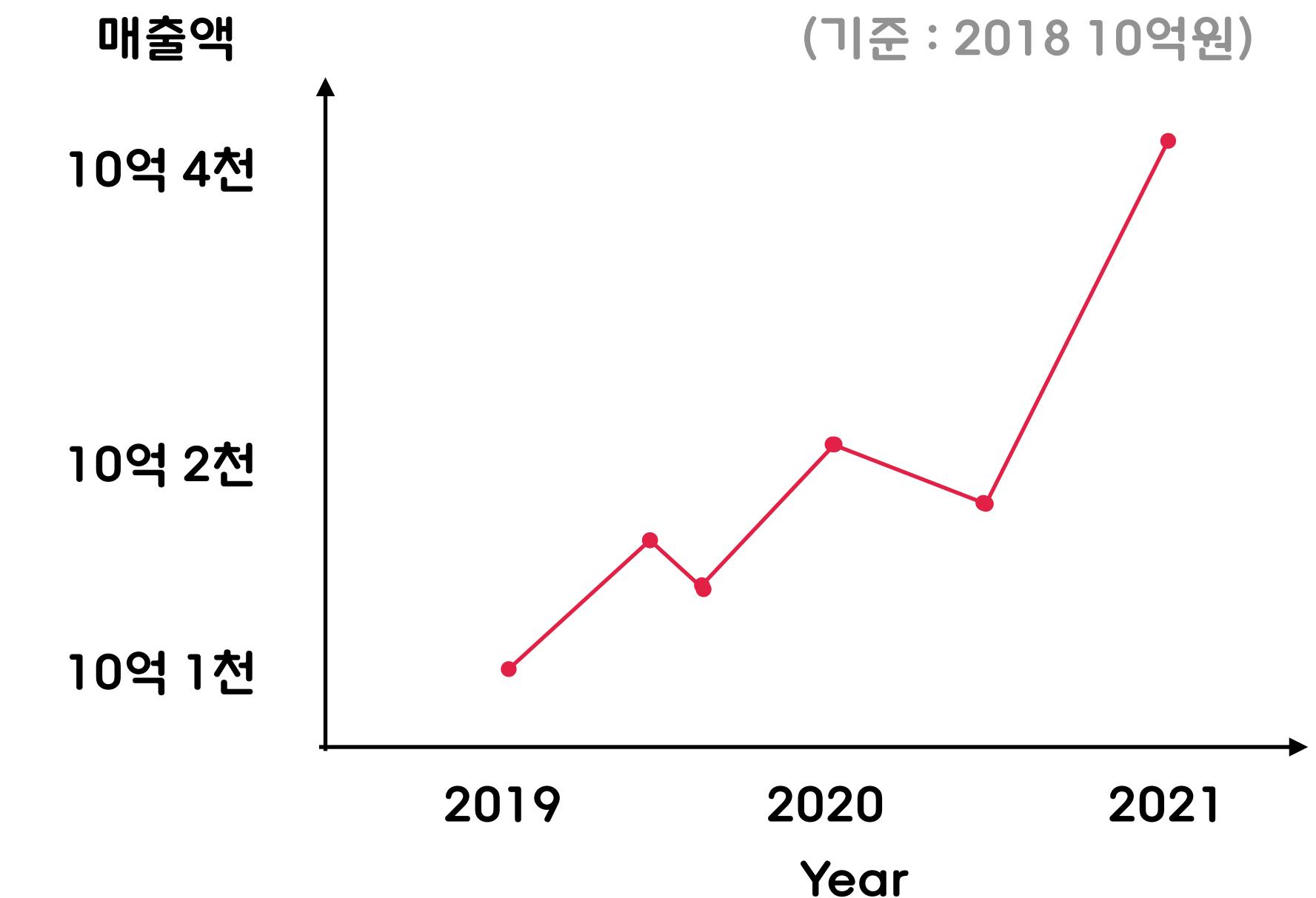
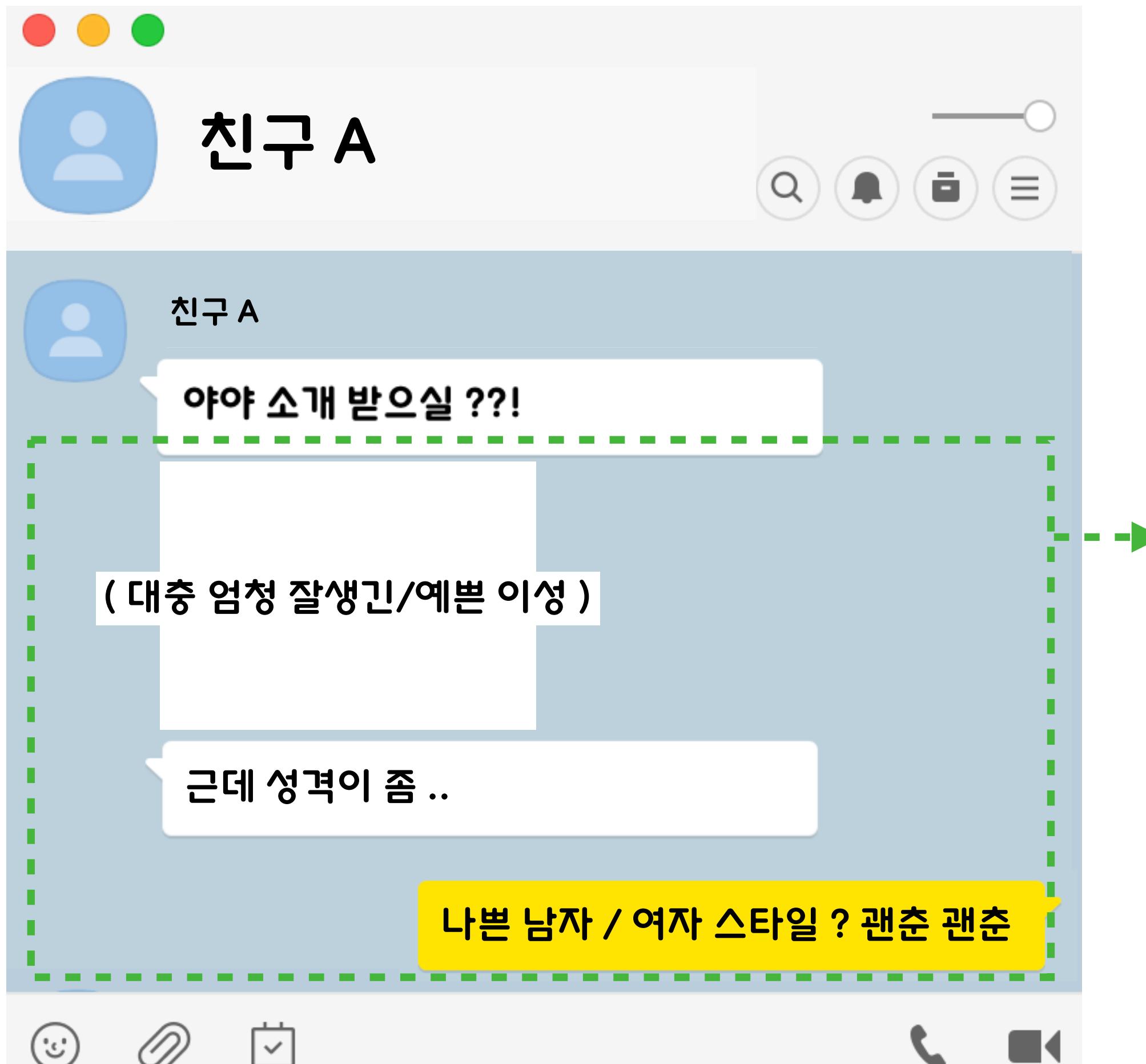
▶ 정규 분포

▶ 요약정보

▶ 시각화

수치로 전달하는 정보에는 한계가 있습니다.
그리고 변화 정도를 인지하는 것이 비교적 상대적입니다.

하지만! 시각화 주의해야 합니다



엄청 매력적인 그래프 이지만 사실 ‘연간 매출액 상승률’은 기준 년도 대비
매년 0.1% 상승 보여지는 것처럼 극적인 변화는 아님!

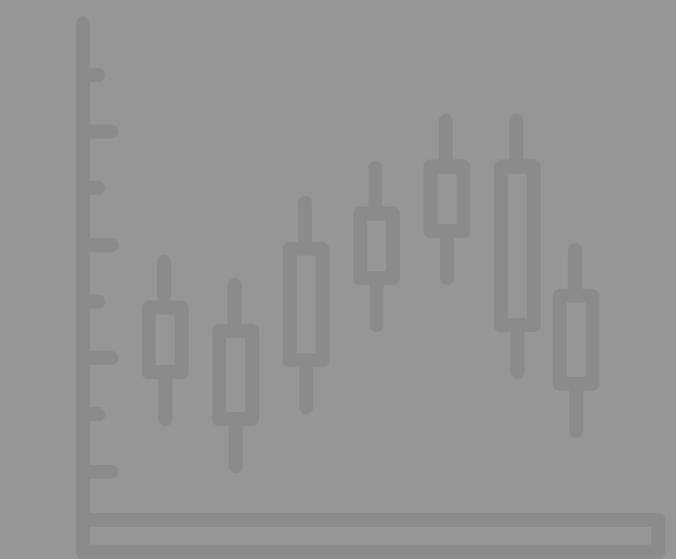
데이터를 이해하는 과정 ‘EDA’



(복잡하게 연결된 데이터)



(이상치 파악)



(데이터 밀도 파악)



(변수 비교)



(연관성 파악)

시각화를 ‘데이터로 이야기 하는 방법’이라 했습니다.
TAD 강연, 스티브 잡스처럼 잘 말하기 위해서는?

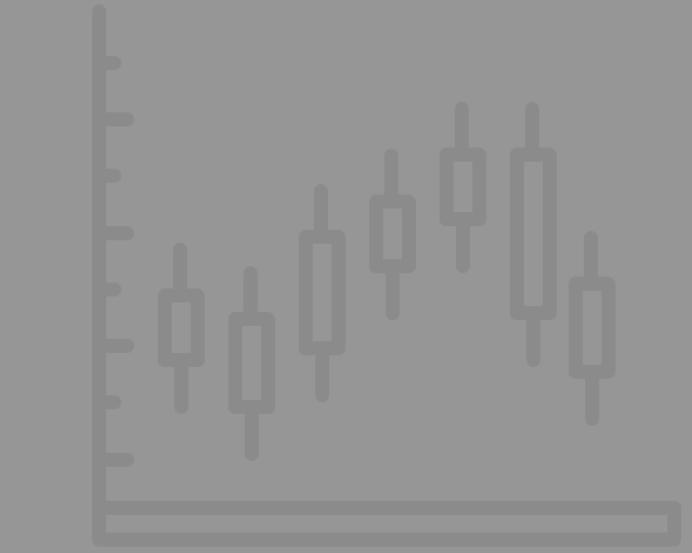
데이터를 이해하는 과정 ‘EDA’



(복잡하게 연결된 데이터)



(이상치 파악)



(데이터 밀도 파악)



(변수 비교)



(연관성 파악)

다양한 강연들을 보고, ‘적용해보기’입니다.
다양한 화법을 보고, 나의 발표에서 실천 해보는 거죠
지금부터는 새로운 시각화 방법을 짹먹 해보겠습니다.

(상관관계 파악) (변수 관계 파악) (집단간 비교) (범주간 비교)



새로운 화법 ! 새로운 시각화 방법 짹먹

(<https://github.com/KimJiSeong1994/lecture>)