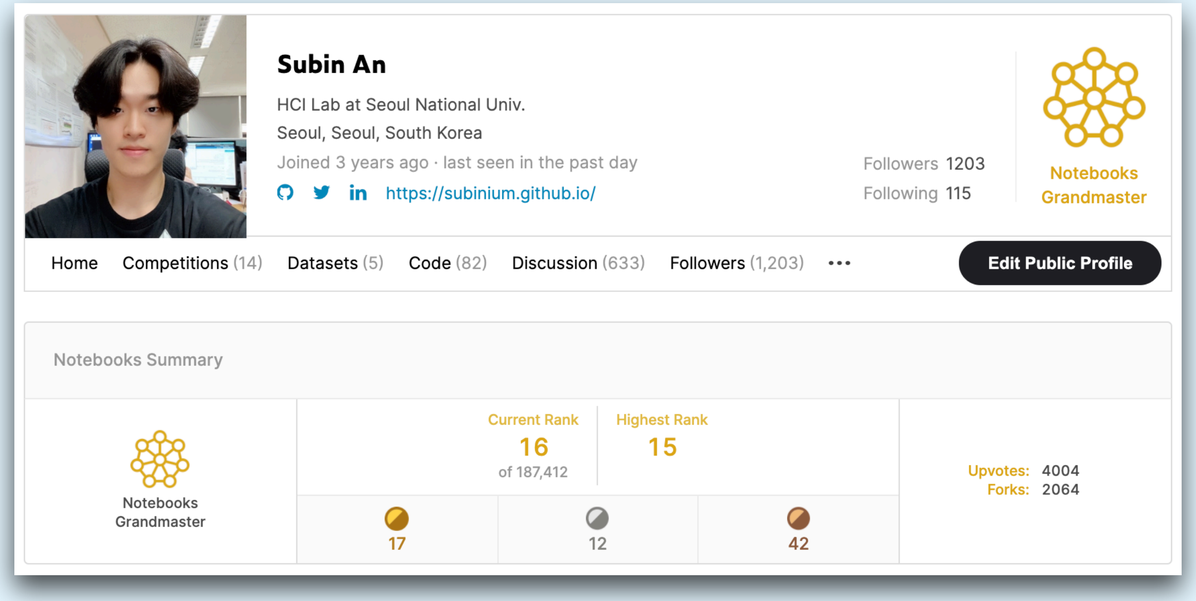
캐글과 데이터 리터러시

데이터분석으로 캐글 그랜드마스터 되기

Kaggle Grandmaster, 안수빈

WHO AM I



서울대학교 컴퓨터공학부 HCI Lab 석박통합과정 (2021~)

카카오 엔터프라이즈 AI Lab R&D intern (2020)

고려대학교 사이버국방학과 졸업 (2016~2020)

서울과학고등학교 졸업 (2013~2016)

Kaggle Notebook Grandmaster / 16위

커뮤니티 Kaggle KR, Data Visualization KR 운영진

Naver AI Boostcamp / Data Visualization 강사

알고리즘 및 코딩테스트 / 기업 데이터 사이언스 강의

캐글 소개

캐글은 어떤 공간인가?

kaggle

Search

- Home
- Competitions
- Datasets
- Code
- Discussions
- Courses
- More

Recently Viewed

- TPS September 2021 E...
- [DataArt] Paint-Style D...
- [Awesome EDA] 2021 ...
- Predicting a Stroke [S...
- All You Need is Tim...

View Active Events

Take the annual Kaggle ML and Data Science Survey!

You'll be contributing to the most comprehensive industry-wide view of the state of data science.

Start here

BEXGBoost • Follow
created this topic 2 days ago

Georgii Vyshnia upvoted this topic 9

How to Write Powerful Code With Custom Sklearn Transformers
in the Tabular Playground Series - Sep 2021 forum

Single `fit`, single `predict` - how awesome would that be?

You get the data, fit your pipeline just one time, and it takes care of everything-preprocessing, feature engineering, modeling, everything. All you have to do is call predict and have the output.

What kind of pipeline is that powerful? Yes, Sklearn has many transformers, but it doesn't ha... See More

Mohamadreza Kariminejad • Follow
created this dataset 9 days ago

Vahideh Dashti upvoted this dataset 24

House Price (Tehran, Iran)
About 3500 Houses with thier complete information (Price in Dollor & Toman)
CSV | 190.2 kB | Data files © Original Authors | 6 | 0 | 1k

Data (1 file)

housePrice.csv (3,479 rows × 8 columns)
Each row represents the information of a house in an apartment.

• Unfollow
created this topic 4 days ago

replied to this topic 29

class activation map again ... for 1d CNN model
in the G2Net Gravitational Wave Detection forum

this is a preview only. more analysis and results coming up soon!

Subin An
Joined 3 years ago

Competitions Contributor

Datasets Expert

Notebooks Grandmaster
Ranked 16th

Discussion Expert

Your Competitions

- Digit Recognizer
- Titanic - Machine Learning fro...
- House Prices - Advanced Regr...
- Predict Future Sales
- Natural Language Processing ...
- I'm Something of a Painter Mys...
- LearnPlatform COVID-19 Impac...
- Optiver Realized Volatility Predi...
- MLB Player Digital Engagement...
- Tabular Playground Series - Au...
- Tabular Playground Series - Se...

Your Datasets +

- anime-face
- High-Resolution Anime Face D...
- Full Emoji Image Dataset

Kaggle

캐글은 어떤 곳일까?

- Data Science Competition 플랫폼
- Competition, Notebook, Dataset, Discussion 등 다양한 활동

- Online community of data scientists and machine learning practitioners
  - 데이터 사이언티스트와 머신러닝 연습자들의 온라인 커뮤니티

Kaggle의 운영 방식

어떻게 온라인 커뮤니티가 순환되는가?

- 3개월 정도 기간의 대회
- 기업에서 일부 금액을 주고 신청제

Competition / Dataset

- 모든 사용자가 만들 수 있음
- 원하는 연습 및 연구를 진행할 수 있음

Notebook

- 대회 관련 인사이트
- 데이터셋 분석과 인사이트
- 튜토리얼 작성

Discussion

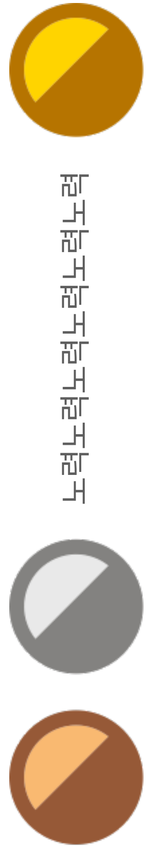
- 코드가 아닌 글로 인사이트 공유
- 대회/데이터셋에 부족한 부분 문의
- 솔루션 공유
- 팀원 모집 등

Kaggler의 목표

어떤 캐글러들이 있을까요?

- 재미, 성취감, 공부, 상금, 취업 등등의 복합적 목표가 될 수 있다!

- **성취감 : 메달과 티어**
  - Competition : **좋은 예측 결과**
  - Dataset : **활용도 높은 데이터셋**
  - Notebook : **도움되는 코드와 설명**
  - Discussion : **도움되는 글**



Kaggler의 Tier

타이틀을 따기 위한 여정

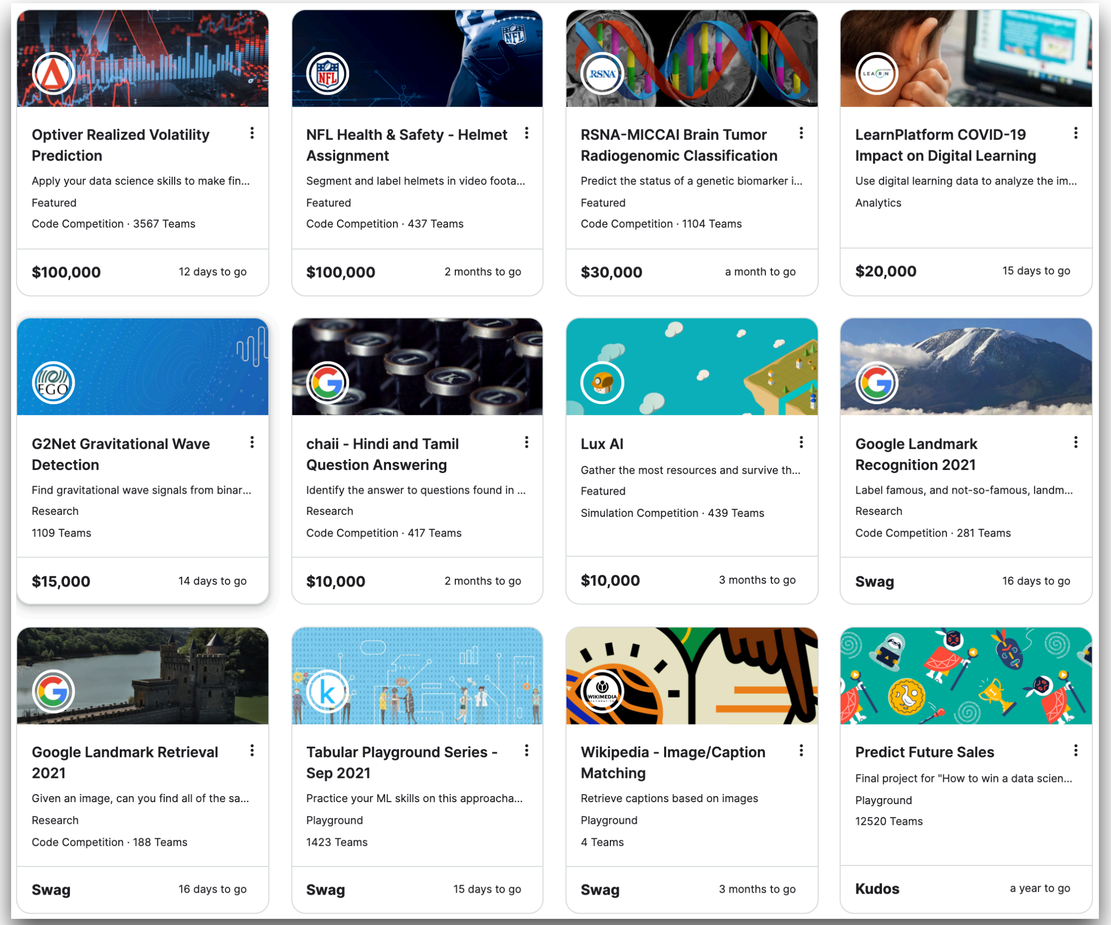
초록색(Novice) 다음은 하늘색(Contributor) 다음은 보라색(Expert) 다음은 주황색(Master) 다음은 금색(Grandmaster)

대회, 노트북, 데이터셋

어떤식으로 활용하고 공부할 수 있을까?

Kaggler의 수 많은 대회들

어떤 대회를 선택하지??



대회 선택 팁 (1)

좋은 결과를 얻고 싶다면 이런 기준으로

- 충분히 기한 내에 대회 프로세스를 거칠 수 있는가?
  - 매일 야근하는데 1주 남은 대회를 참여할 수 있을까?
- 아는 도메인 또는 관심이 있는 도메인인가?
  - e.g., 의료 이미지, 분자 이미지, 주식 가격, 스포츠 결과 예측
- 현재 딥러닝/머신러닝 지식로 충분히 승산이 있는가?
  - 공유된 최신 논문을 빠르게 내 알고리즘에 적용해볼 수 있을까?

대회 선택 팁 (2)

공부를 하고 싶다면…

- 현재 진행중인 대회 외에도 **지난 대회로 연습** 가능
- 가장 최근 종료한 대회들은 최신기술과 solution이 많이 공유되어 있는 공유의 장!
- 최대한 이전 대회를 잘 활용해야 스스로도 실력이 늘고, 좋은 결과를 얻을 수 있다!

대회 시스템의 문제점

세상은 좀 달라…

- 대회에 좋은 성과를 내면 좋은 데이터 사이언티스트인가???
  - 다양한 역량 중, 결과에 초점되어 있다. (결과를 지향하다보면 다른 역량도 같이 오르는 것도 맞다.)

- 하지만 실제 데이터는 수집 단계/전처리 단계/배포 단계까지 다양하다! ~~(MLOps의 세상으로…)~~
- 캐글 시스템을 다양하게 활용하여 데이터 리터러시 역량을 늘리는 것을 추천

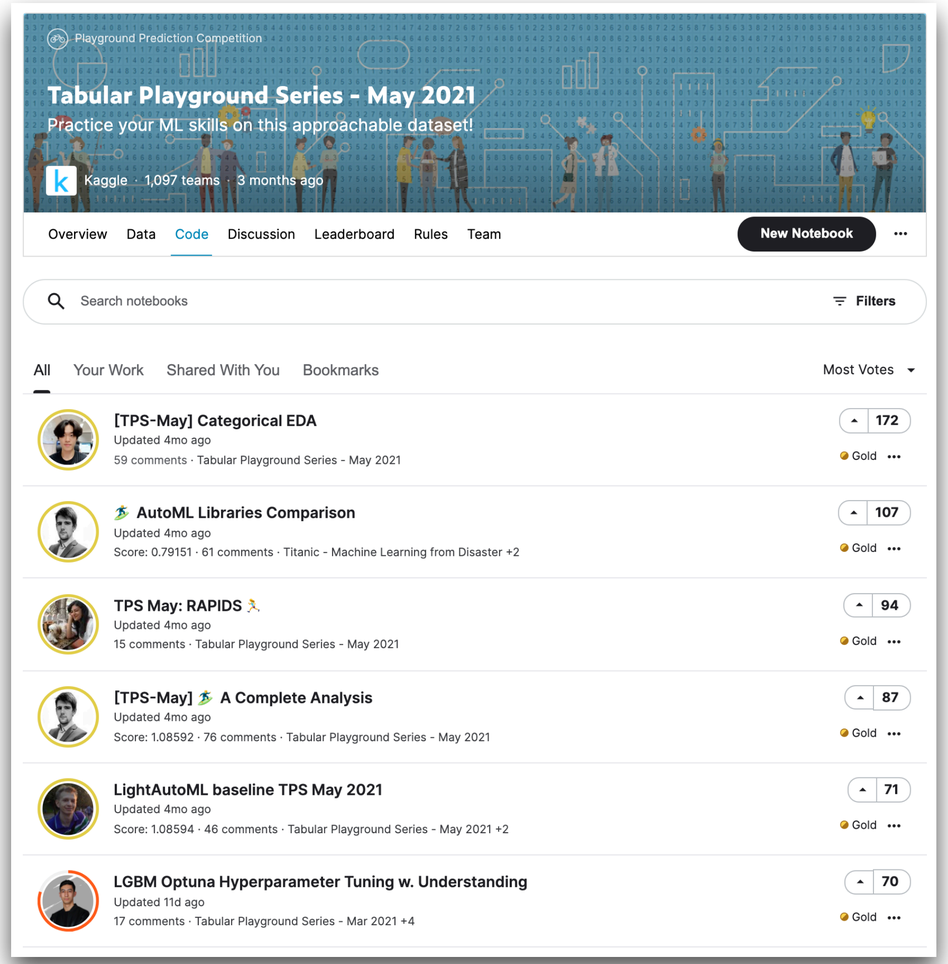
데이터 리터러시를 키우자!

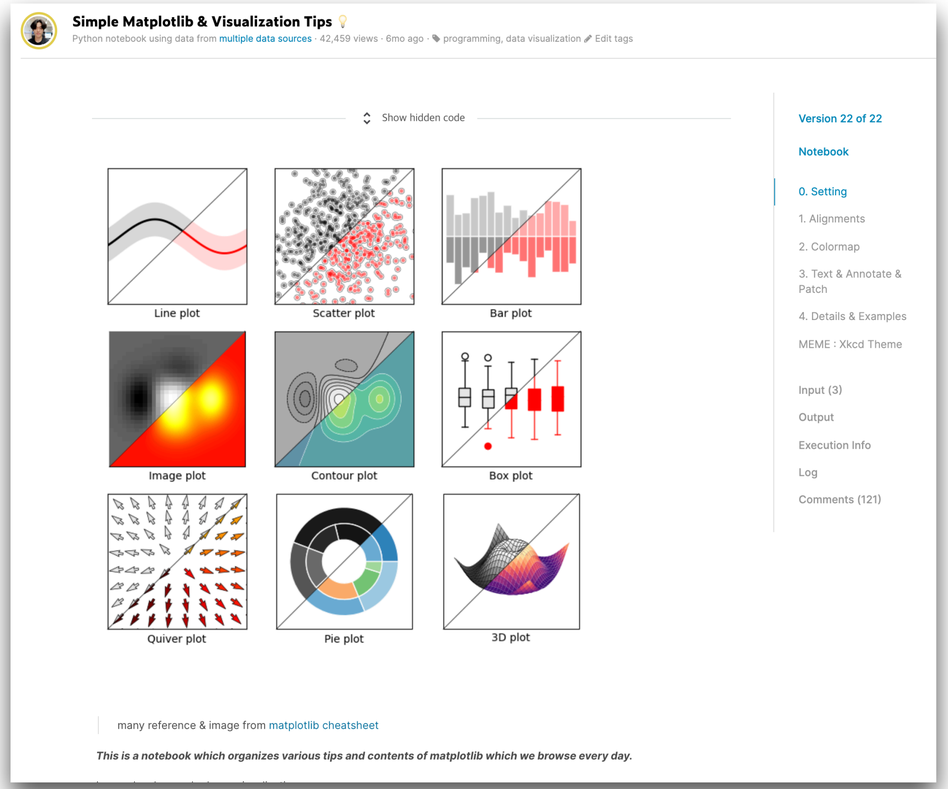
시야를 넓히자

- 데이터 리터러시는 데이터를 건전한 목적과 윤리적인 방법으로 사용한다는 전제 하에, 현실 세상의 문제에 대한 끊임없는 탐구를 통해 질문하고 답하는 능력
  - **좋은 질문**을 할 수 있는 역량
  - 필요한 **데이터를 선별하고 검증**할 수 있는 역량
  - 데이터 **해석 능력을 기반으로 유의미한 결론**을 만들어내는 역량
  - 가설 기반 **A/B 테스트를 수행하여 결과를 판별**할 수 있는 역량
  - **의사결정자**들도 이해하기 쉽게 **분석 결과를 표현**할 수 있는 역량
  - **데이터 스토리텔링**을 통해 의사결정자들이 전체그림을 이해하고 분석 결과에 따라 실행하게 하는 역량

노트북이란?

특정 주제에서 나올 수 있는 질문, 방법론, 코드, 시각화의 집합소





머신러닝 프로세스

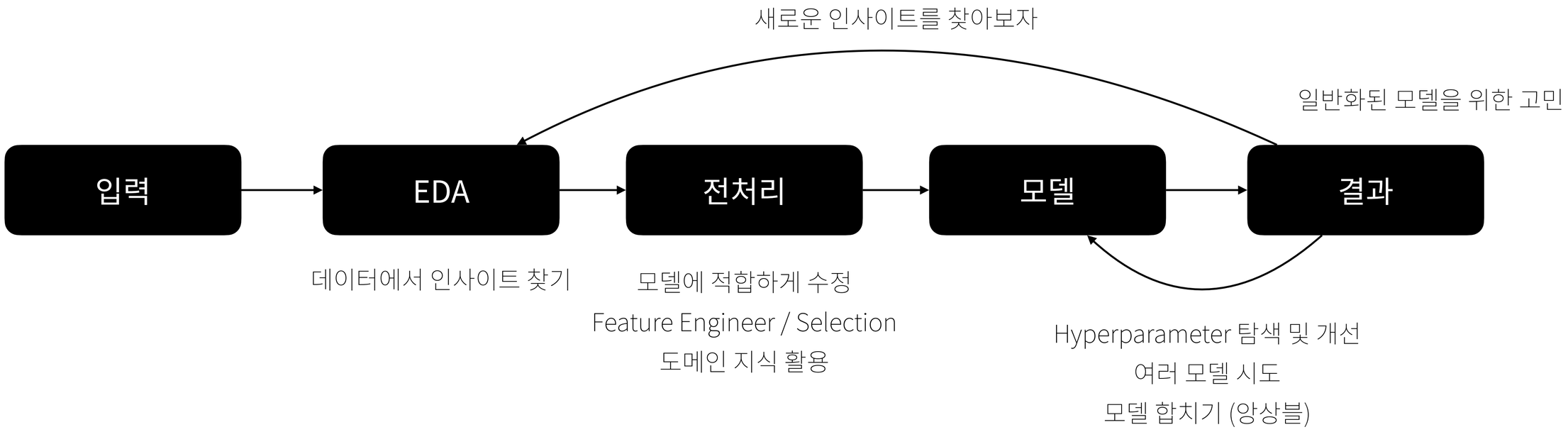
어떤 과정에서 사람들은 아이디어가 나올까?

어느 깊은 가을밤 잠에서 깨어난 제자가 울고 있었다.
그 모습을 본 스승이 기이하게 여겨 제자에게 물었다.
"무서운 꿈을 꾸었느냐?"
"아닙니다."
"슬픈 꿈을 꾸었느냐?"
"...아닙니다. 입력을 넣고 모델을 돌리니 결과가 좋은 꿈을 꾸었습니다."
"그런데 왜 그리 슬피 우느냐?"
제자는 흐르는 눈물을 닦아내며 나지막이 말했다.
"그 꿈은... 이루어질 수 없기 때문입니다."

1. **입력받고**
2. **모델돌리고** 좋겠지만… 그렇게 ML이 호락호락하지 않다!
3. **결과만 나오면**

머신러닝 프로세스

어떤 과정에서 사람들은 아이디어가 나올까?



머신러닝 프로세스

다양한 라이브러리

pandas, numpy
pickle, json
open-cv, pillow, librosa
datable, cudf

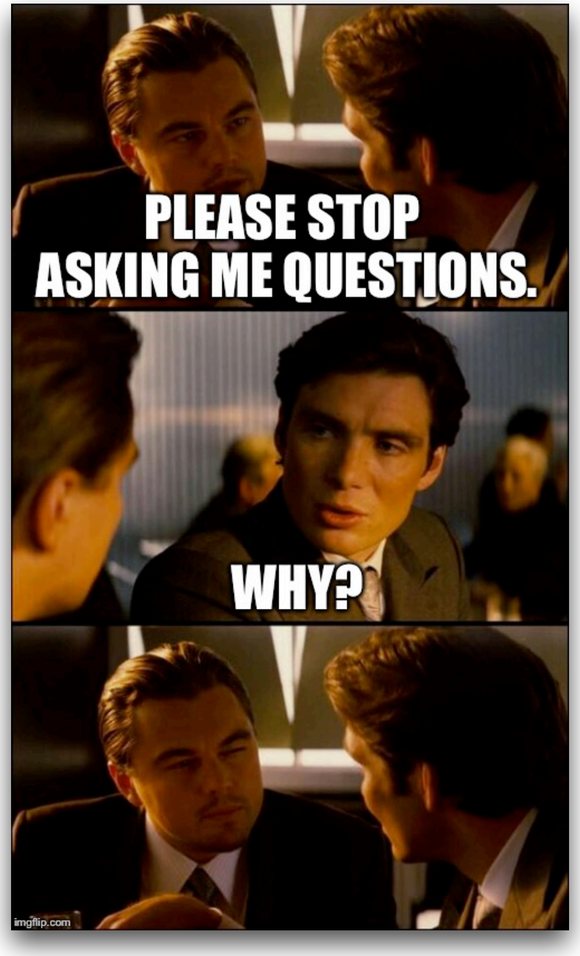
matplotlib
seaborn
plotnine
plotly
altair
Bokeh

sklearn (cuml)
xgboost, lgbm, catboost
tensorflow, keras, pytorch
nni, optuna
automl, pycarrot

일차적으로 정제된 데이터지만 그 데이터 내에서 다양한 시도 가능

질문을 멈추지 말자

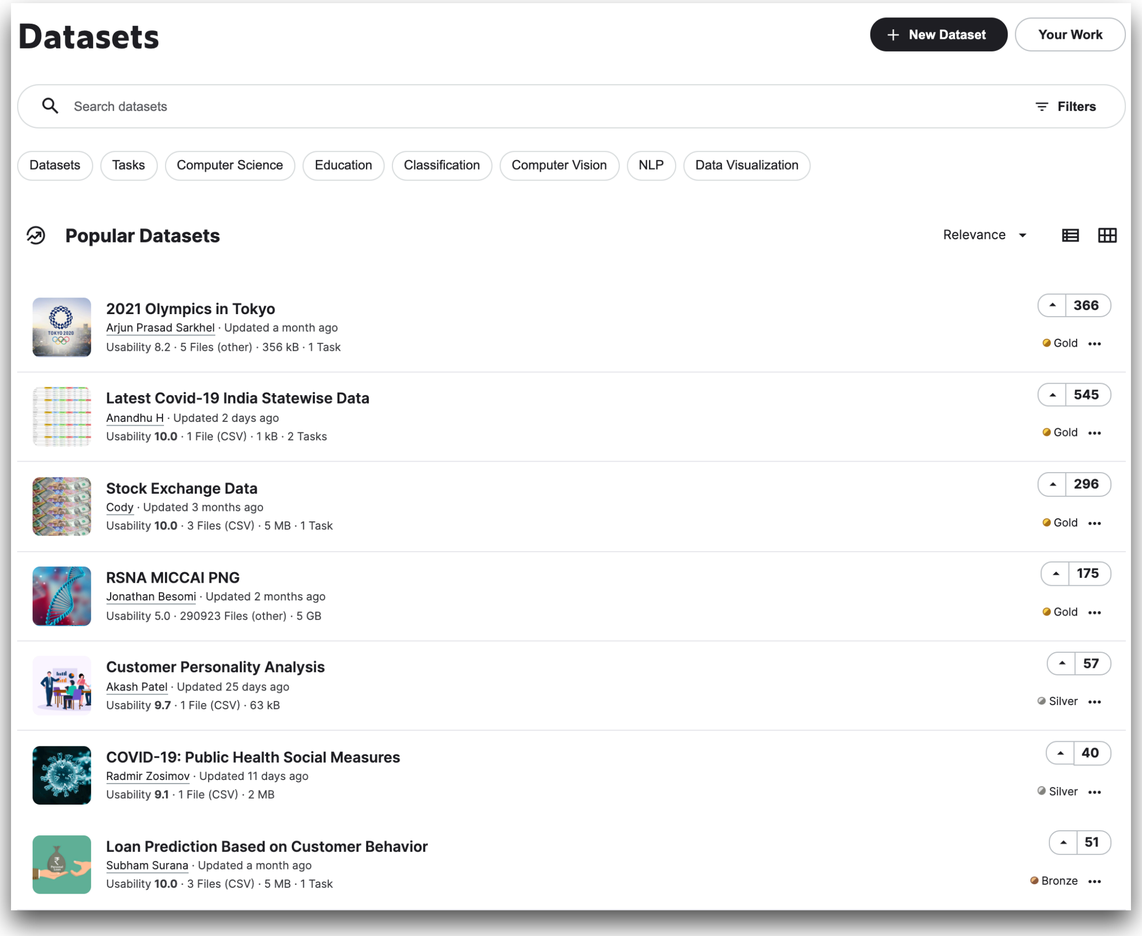
끊임없는 질문은 데이터 사이언티스트의 대표적 역량



- 어떤 문제를 해결하려고 하고 있는가?
- 어떤 종류의 데이터가 있고, 이를 어떻게 처리할까?
- 데이터에 누락값이 무엇이고, 왜 생겼을까?
- outlier가 있다면, 왜 있는 outlier일까?
- 데이터를 최대한 활용하기 위해 어떻게 새로운 피처를 만들까?

데이터셋

방법이 아니라 재료를 만들어보자



데이터 분석과 스토리텔링

잘 분석한 결과를 어떻게 보여줄 것인가?

Designers != Developers != Users

디자이너와 개발자의 의도는 사용자의 이해와 다를 수 밖에 없다.

- 개발자/디자이너의 중심의 사고는 일반적인 사용자와 다르다.

- **심성모형(Mental Model)**
  - Designer Model : 시스템을 만드는 사람 속의 시스템 구조
  - System Image : 실제 시스템의 구조
  - User’s Mental Model : 사용자가 생각하는 시스템의 구조

- 진짜 독자에 대한, 데이터 분석 결과 사용자에 대해 더 고민하자.

User Experience

사용자 경험

- UX의 3요소
  - **유용성 (Usefulness)** : 하고자 하는 일을 효과적으로 달성
  - **사용성 (Usability)** : 사용하는 과정이 효율적
  - **감성 (Affect)** : 느낌

- 여러분의 분석 결과를 효과적으로 스토리텔링하기 위해서는
  - 정보를 (왜곡없이) 잘 전달했는가?
  - 분석 내용을 효과적으로 이해할 수 있는가?
  - 보기 좋은가? (시각화)

유용한 분석인가?

데이터 분석에서 다수 놓칠 수 있는 부분

- **데이터 분석은 다양한 부분에서 미스가 발생할 수 있다.**

- 데이터셋이 수집된 환경과 분석 결과를 적용할 환경에 대한 고려가 충분히 되었는가?
- 인과관계에 대한 해석에 타당성이 있는가? (상관관계에서 그치면 안됨.)
- 가정이 맞다는 것을 어떻게 보였는가?
- 해결책에 대해 장점 외의 한계점 등 보이지 않는 부분을 포함했는가?
- 데이터를 올바르게 시각화하였는가?

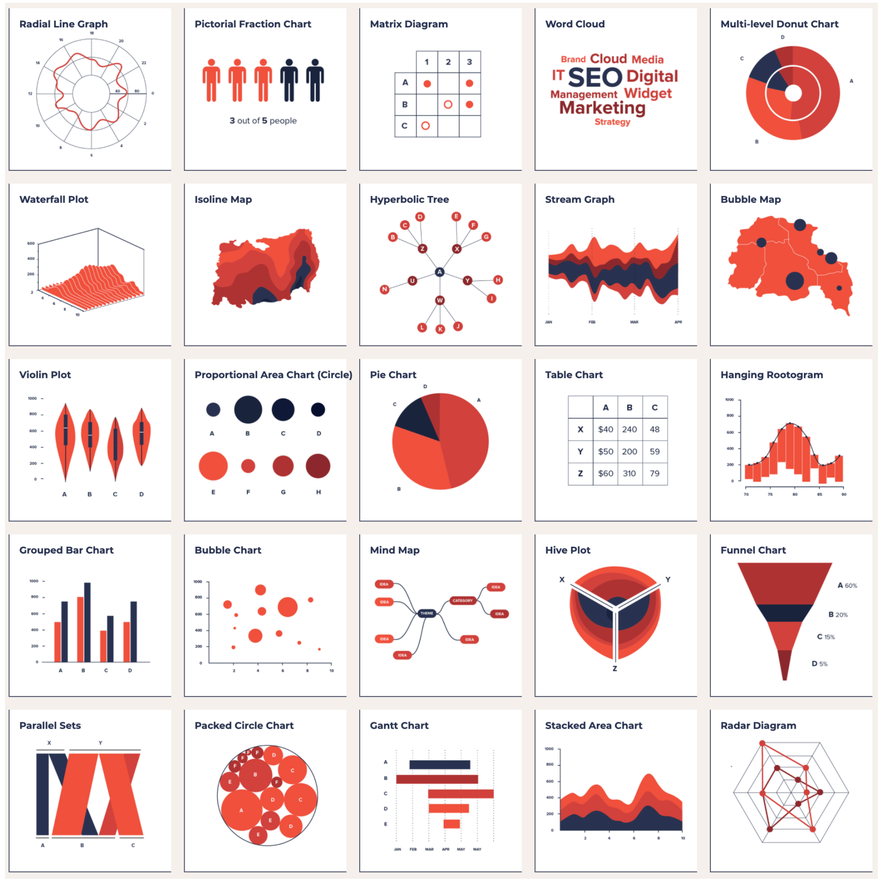
올바른 시각화란?

이쁜 것에 현혹되어 정보를 잃지 말자.

- 시각화에서 가장 중요한 것은 강조는 하지만, 왜곡은 없이 그대로 보여주는 것.
- 이런 고민들을 해보며 시각화를 해야 합니다.
  - 값 또는 값의 변화가 실제 차이보다 크게 보이지는 않는가?
    - 예시1) 사각형/삼각형/원으로 산점도를 그리면 비교하기 좋을까?
    - 예시2) 막대 그래프의 기준점이 0이 아니어도 될까?
  - 순서가 중요한 경우, 이를 오인할 수 있는 요소를 사용하지는 않았는가? (잘못된 색상 배치)
    - 예시 1) 색을 무지개색을 사용해도 될까?

효과적인 방법론

효과적인 방법론에 이쁜 디자인을 섞자



https://datavizproject.com/

시각 요소들의 활용

시각화에서 각 요소가 담는 정보는 무엇일지 고민해보기

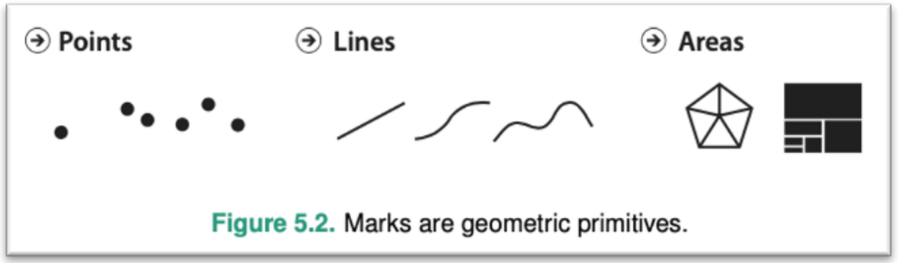


Figure 5.2. Marks are geometric primitives.

점, 선, 면으로 이루어진 데이터 시각화

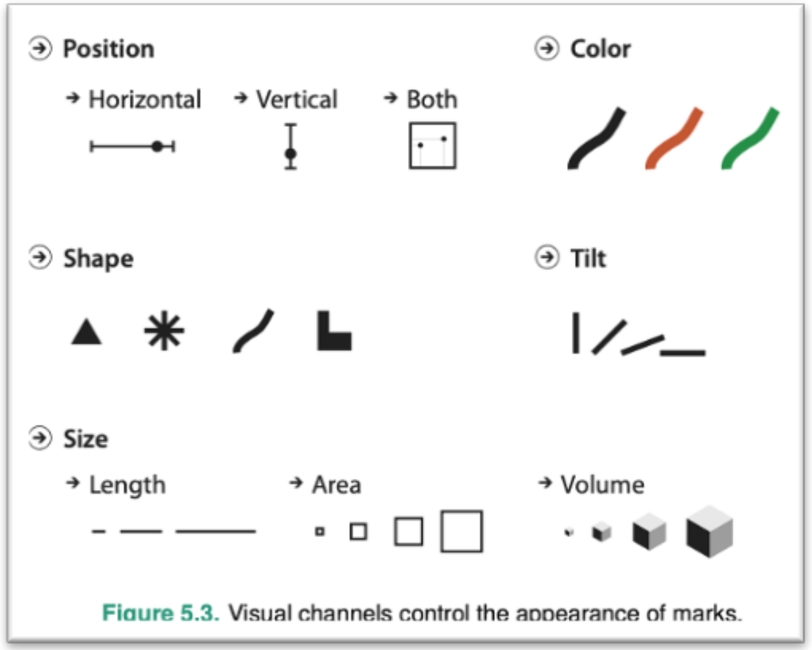


Figure 5.3. Visual channels control the appearance of marks.

각 마크를 변경할 수 있는 요소들

효과적으로 전달하기

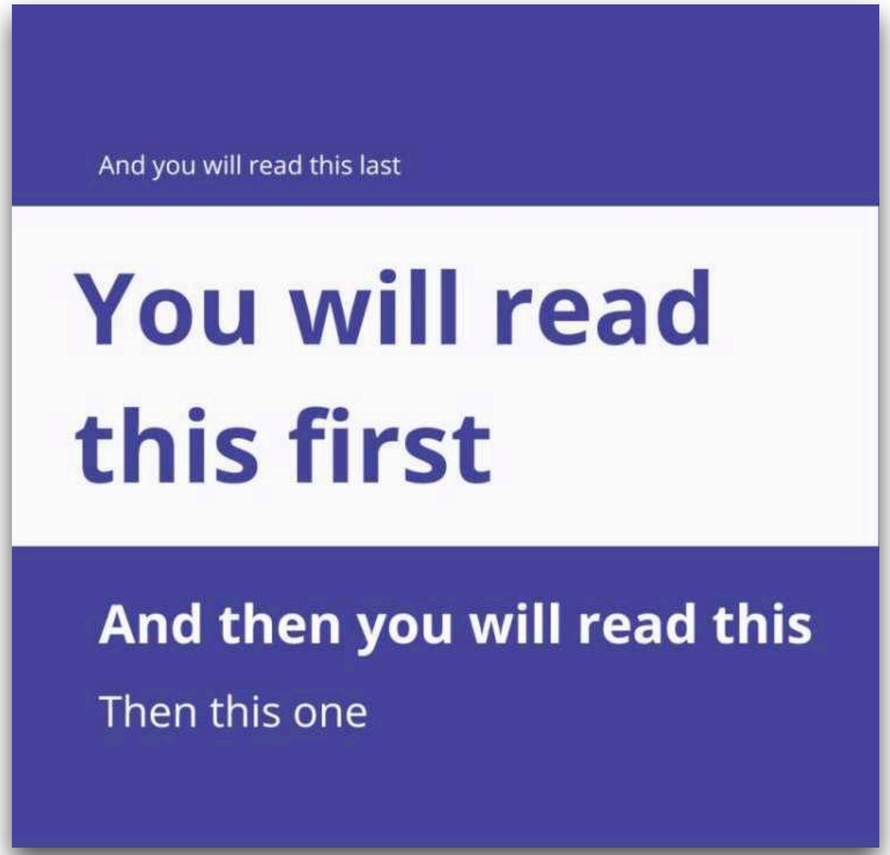
한눈에? 재미있게? 의도에 맞게!

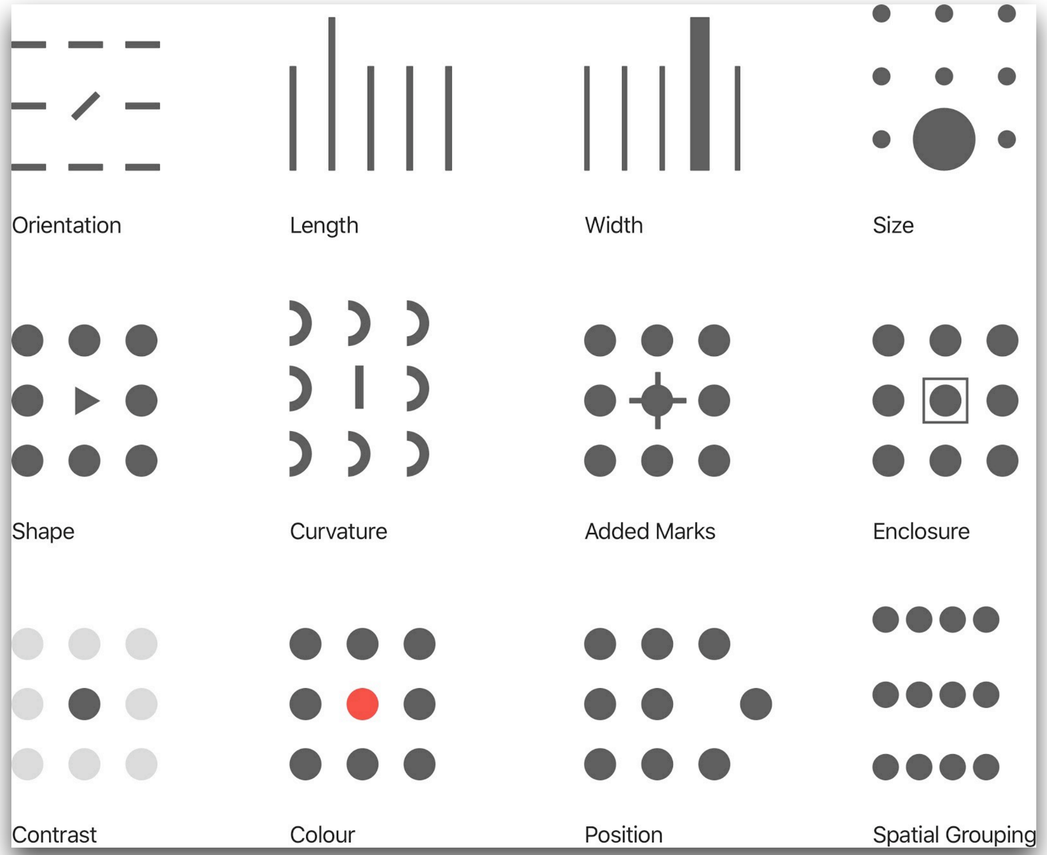
- 결국에 데이터 분석의 결과는 **의사결정자의 마음**에 들어야 한다.

- 흔히 말하는 기승전결이 있는 스토리텔링인가? (포인트가 확실한가)
- 구조화가 되어 있어 독자가 지식을 계층적으로 체계화할 수 있었는가?
- 청자/독자의 이해에 맞는 표현을 사용하였는가? (축약어 등)
- 적절하게 강조를 주어 원하는 부분에 강조를 주었는가?
- 차트 자체만으로도 설명이 충분한가?

사람의 시각적 흐름을 생각하자

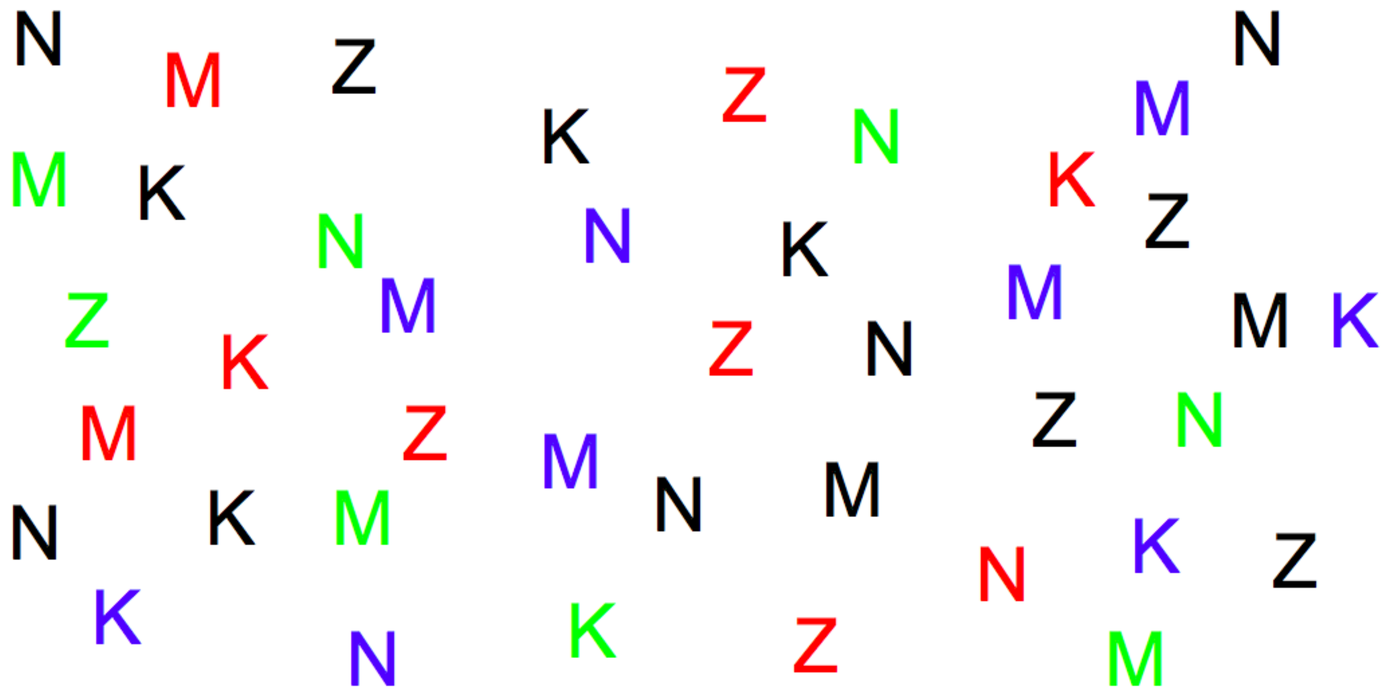
눈은 의도치 않게 따라가고, 비교한다.





재미있는 실험

N의 개수는? 빨간색의 개수는?



심미적 향상을 위한 디자인

디자인이 이쁘면 분석 결과도 좋아보인다.

- 디자인은 정말 다양한 조건이 있다.
- 그 중에서도 대표적인 디자인 이론은 C.R.A.P.
  - Contrast : 시각에서 인지할 수 있는 차이를 통한 강조와 비교
  - Repetition : 디자인 전반에 걸쳐 동일하거나 유사한 요소 재사용 (색상, 테마 등)
  - Alignment : 요소들의 배치
  - Proximity : 유사한 내용은 가까이 그룹화하여 의미전달 높이기
- 의미를 해치지는 말 것. (왜곡은 X)

http://web.mit.edu/6.813/www/sp18/

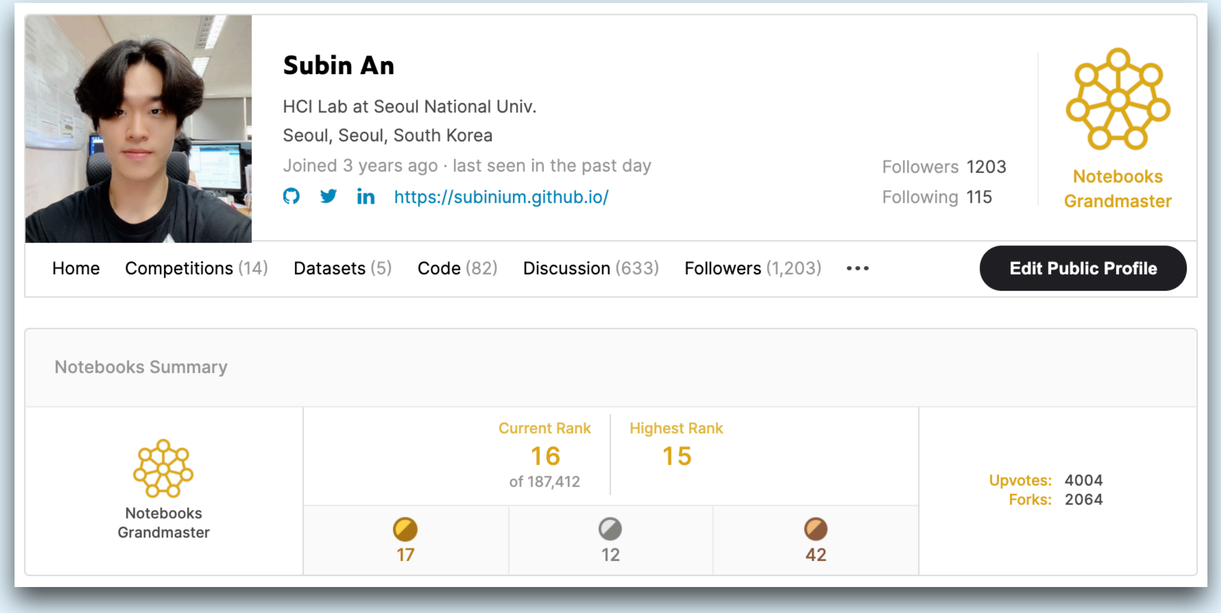
Kaggle Notebooks Gold Medal

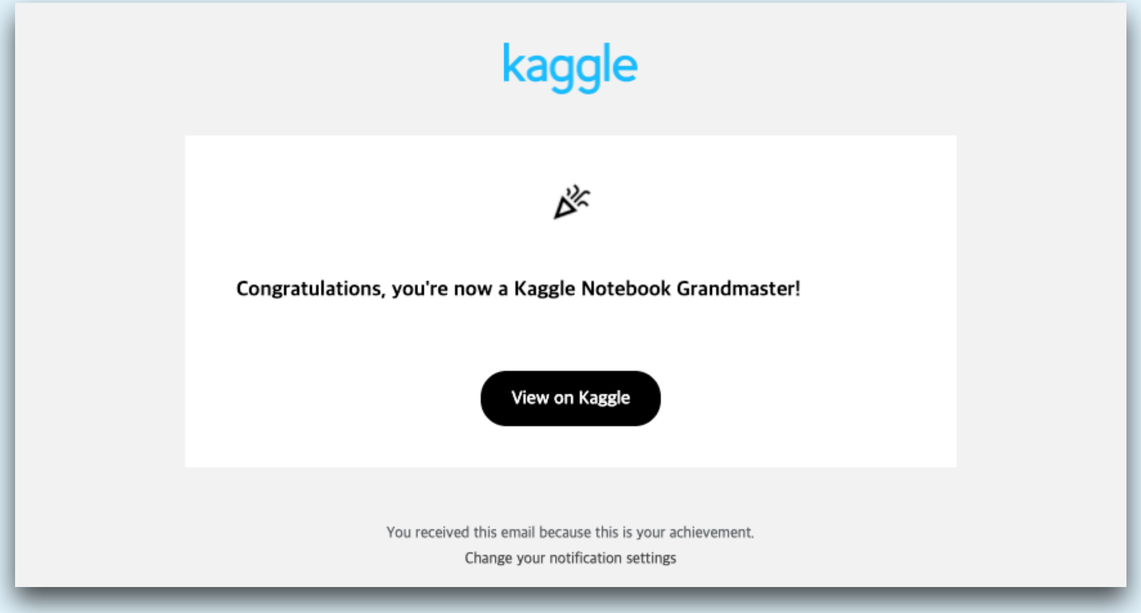
각 고민 끝에 얻을 수 있었던 캐글 노트북 금메달 17개

- **Data Visualization Techniques**
  - Matplotlib Tips
  - Information visualization Tips
  - Python Data Visualization Library Tutorial
  - Plotly Express/Seborn Comparison & Tutorial
- **Survey Competition**
  - 2020 Kaggle Survey
  - 2019 Kaggle Survey
- **Utils**
  - 11 categorical encoder
  - PyCaret Explanation
- **Competition EDA**
  - Tabular Playground Series April EDA
  - Tabular Playground Series May EDA
  - Tabular Playground Series May EDA
  - Tabular Playground Series Jun EDA
  - Categorical Feature Encoding Challenge II EDA
- **Beautiful Visualization**
  - Titanic Dataset Visualization
  - Netflix Dataset Visualization
  - Matplotlib Darkmode Visualization
  - Christmas Tree 3D Animation Visualization

Kaggle Notebooks Gold Medal

각 고민 끝에 얻을 수 있었던 캐글 노트북 메달





감사합니다

QnA