일본과 중국의 kaggler 비교

두 나라의 Kaggler 수와 2년동안의 트렌드 분석

Contents

프로젝트 소개

주제 선정 배경 및 목적

1.

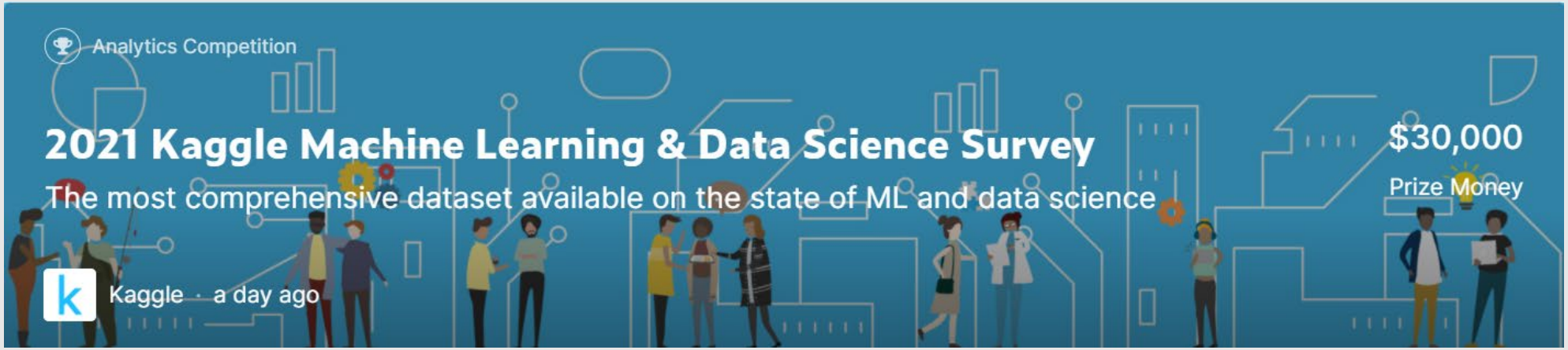
Kaggle은 무엇인가??

kaggle

캐글(kaggle)이란 2010년 설립된 예측모델 및 분석 대회 플랫폼으로, 기업 및 단체에서 데이터와 해결과제를 등록하면, 데이터 과학자들 및 다양한 직군의 개발자들이 이를 해결하는 모델을 개발하고 경쟁한다.

2017년 3월, 구글에 인수되어 알파벳을 모기업으로 두고 있으며, 전 세계 약 500만명의 캐글러가 활동하고 있을 정도로 데이터분야의 종사자에게 많은 인기를 얻고있는 플랫폼이다.

Kaggle 대회 개요



Kaggle Machine Learning & Data Science Survey

2017년부터 시작되어 매년 전 세계의 Kaggler(캐글사용자)들을 대상으로 진행한 설문조사 결과를 바탕으로 데이터를 분석하여 각각의 주제를 선정하고 그에 대한 데이터를 다양한 차트로 시각화하는 대회이다.

주제의 목적 및 필요성

목 적 | 캐글러의 규모와 국가의 위치가 비슷한 동아시아 2국의 데이터분석 트렌드 비교

필요성 | 빅데이터, 머신러닝의 중요성은 점점 커지고 있음

주변국의 데이터 분석 트렌트를 살피고 그에 맞춘 역량강화가 필요

트렌드 비교를 바탕으로 취업준비생의 관점까지 고려할 필요가 있음

프로젝트 소개

사용 데이터 설명

	Time from Start to Finish (seconds)	Q1	Q2	Q3	Q4	Q5	Q6	Q7_Part_1	Q7_Part_2	Q7_Part_3	...
0	Duration (in seconds)	What is your age (# years)?	What is your gender? - Selected Choice	In which country do you currently reside?	What is the highest level of formal education ...	Select the title most similar to your current ...	For how many years have you been writing code ...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	What programming languages do you use on a reg...	...
1	910	50-54	Man	India	Bachelor's degree	Other	5-10 years	Python	R	NaN	...
2	784	50-54	Man	Indonesia	Master's degree	Program/Project Manager	20+ years	NaN	NaN	SQL	...
3	924	22-24	Man	Pakistan	Master's degree	Software Engineer	1-3 years	Python	NaN	NaN	...
4	575	45-49	Man	Mexico	Doctoral degree	Research Scientist	20+ years	Python	NaN	NaN	...
5	781	45-49	Man	India	Doctoral degree	Other	< 1 years	Python	NaN	NaN	...
6	1020	25-29	Woman	India	I prefer not to answer	Currently not employed	< 1 years	Python	NaN	NaN	...
7	141	18-21	Woman	India	Some college/university study without earning ...	Student	1-3 years	NaN	NaN	NaN	...
8	484	30-34	Man	India	Bachelor's degree	Data Scientist	5-10 years	Python	NaN	NaN	...

캐글에서 2019년과 2021년 각각 캐글러들을 대상으로 성별, 연령, 국적, 프로그래밍언어, 사용 툴 등의 여러항목에 대한 설문조사 데이터

사용 데이터 : multiple_choice_responses.csv (2019)
kaggle_survey_2021_responses.csv (2021)

각각의 국가, 나이 등
조건별로 데이터 추출

	Q1	Q3
0	What is your age (# years)?	In which country do you currently reside?
1	50-54	India
2	50-54	Indonesia
3	22-24	Pakistan
4	45-49	Mexico
...	...	...
25969	30-34	Egypt
25970	22-24	China
25971	50-54	Sweden
25972	45-49	United States of America
25973	18-21	India

25974 rows × 2 columns

분석도구

pandas kaggle NumPy plotly

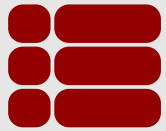
데이터 분석

요인 확인 및 비교 분석

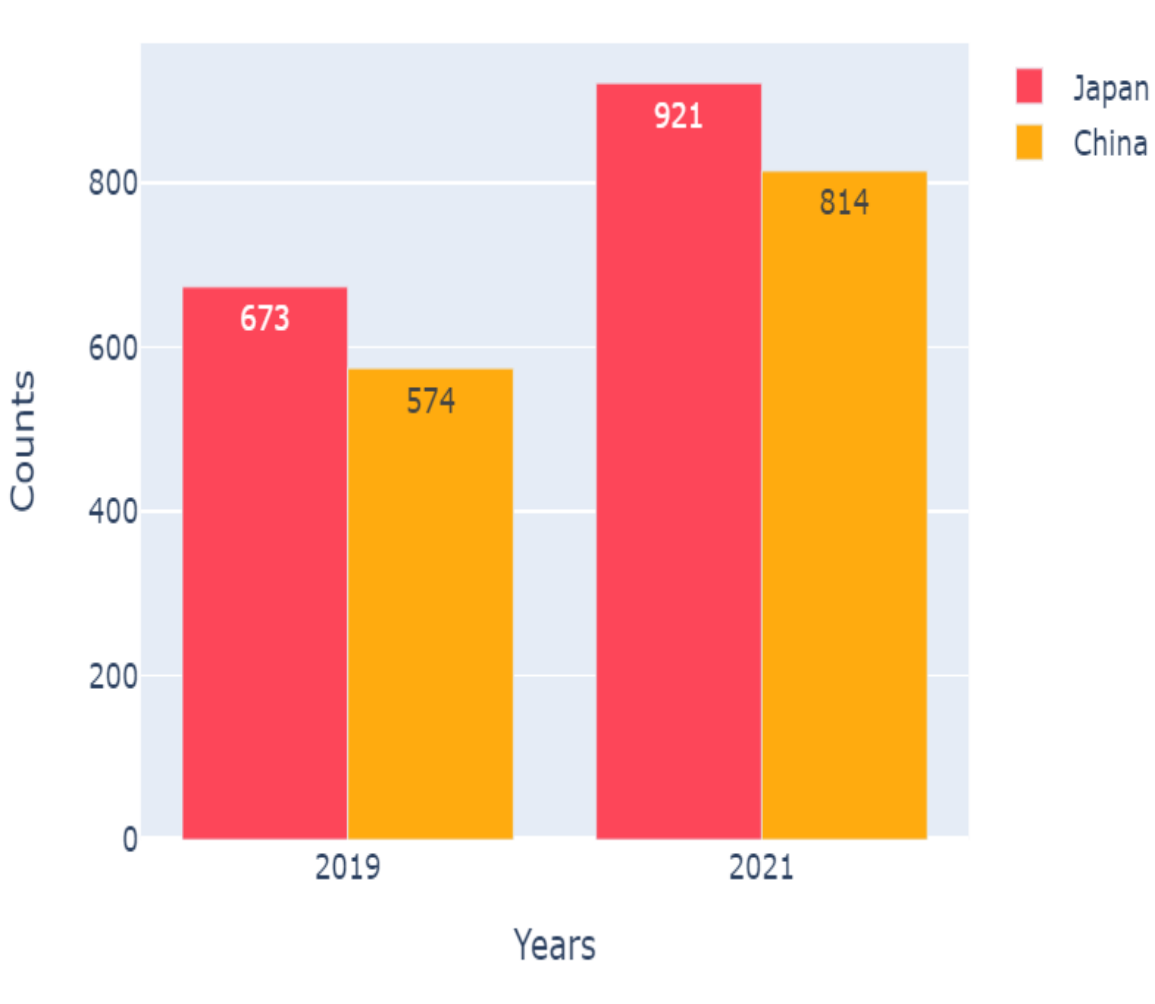
2.

Contents

캐글러 비교를 위한 데이터분석 목차

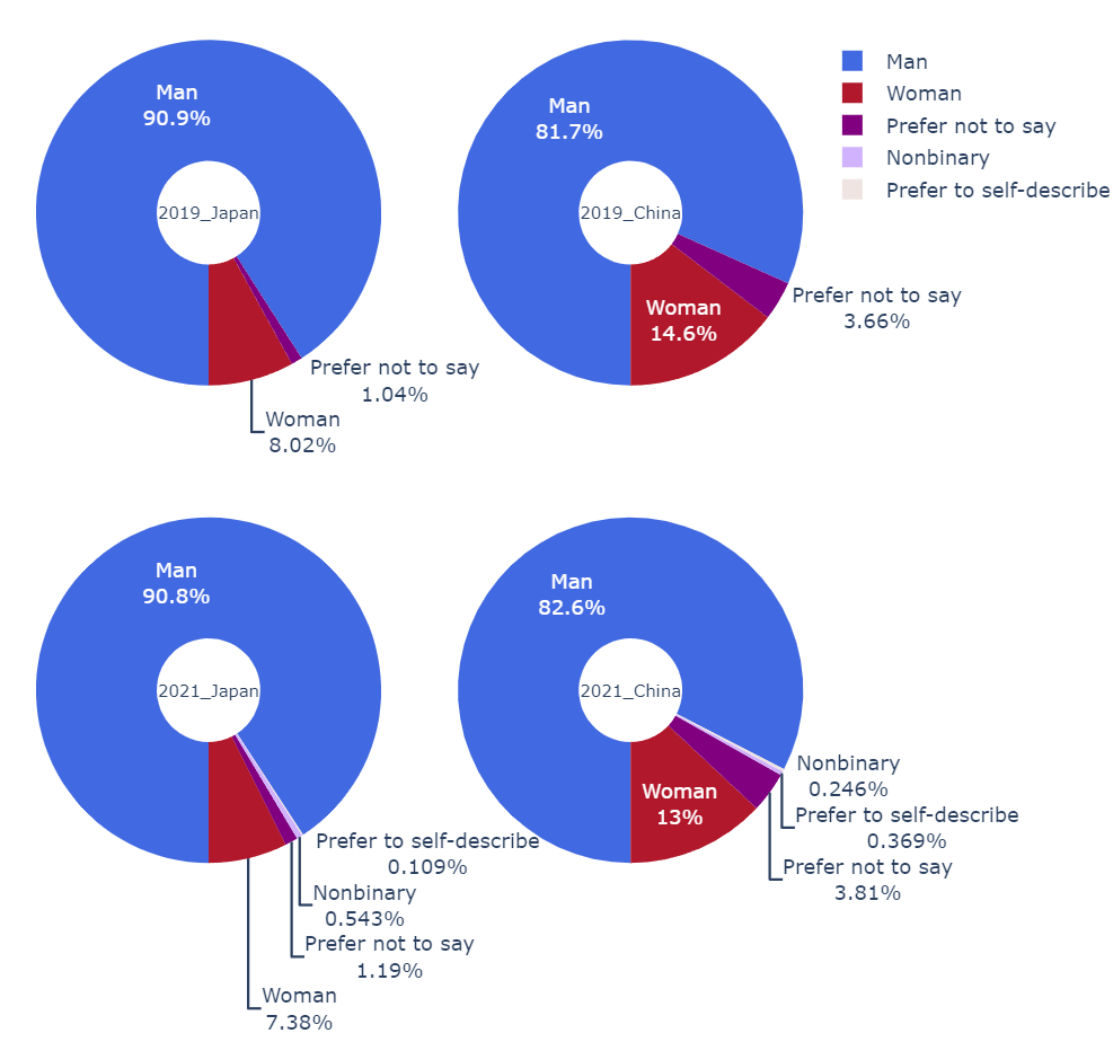


The number of Kaggler



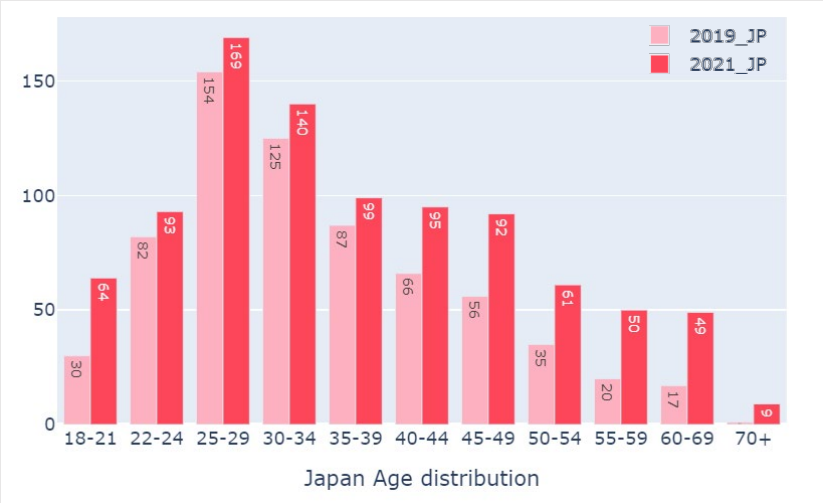
- 양국 모두 시간이 지남에 따라 캐글러가 **증가하는 추세**였으며 수치로는 양국 모두 **약 40%증가**하였다.
- 중국이 일본에 비해 인구는 압도적으로 많으나 캐글러의 수는 일본이 앞섰다.

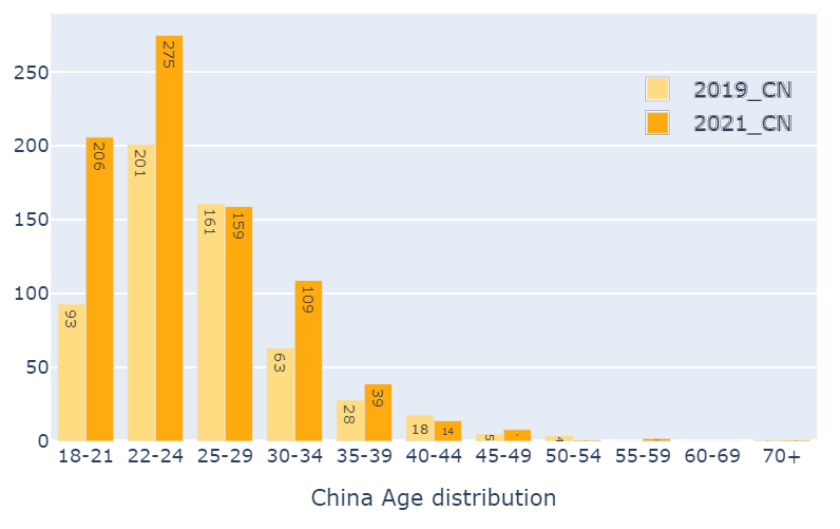
Gender & Age



- 두 나라 모두 **남성비가 압도적**이나 일본은 중국보다 남성비가 더 높아 일본의 여성 캐글러 인구비는 중국의 절반 정도였다.
- 21년에는 'Prefer not to say', 'Nonbinary', 'Prefer to self-describe' **항목이 추가**되었고 5% 미만의 낮은 비율을 보여주고 있었다.

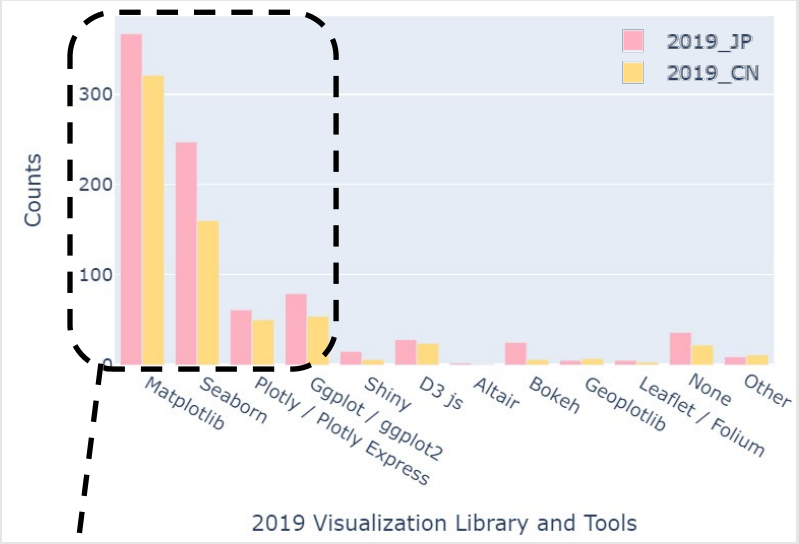
Gender & Age



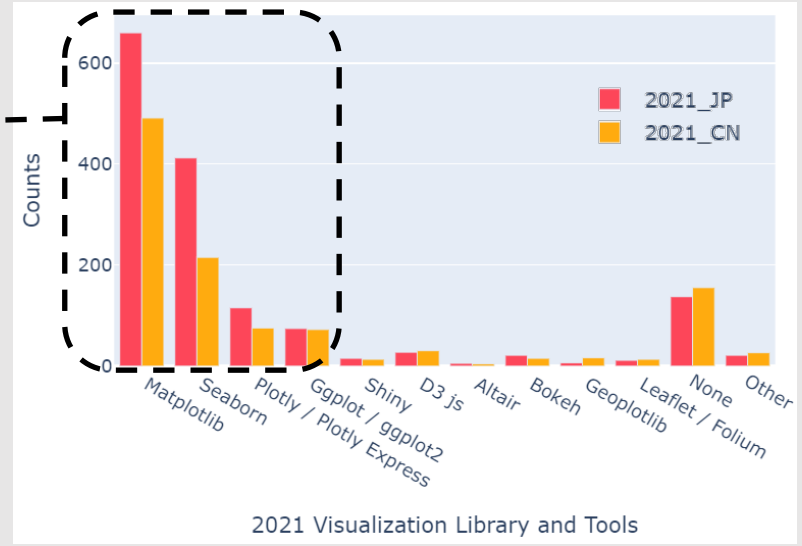


- 두 그래프의 공통점을 꼽자면 젊은층(18~34)의 비율이 높은 점이 공통적이였다.
- 전반적으로 **모든연령층**에서 캐글러의 수가 **증가**한 것을 볼 수 있다.
- 일본은 모든 연령층에서 캐글러의 수가 증가했으나, 중국은 그 수가 **감소한 연령층이 존재**한다.
- 양국을 비교하며 보면 **중국의 젊은층 비율**이 압도적으로 높았다, 비율로 보면 **전체 연령층의 90%** 정도로 일본(50%)에 비해 매우 높은 편이였다.

Libraries and Tools



주요 4개 라이브러리



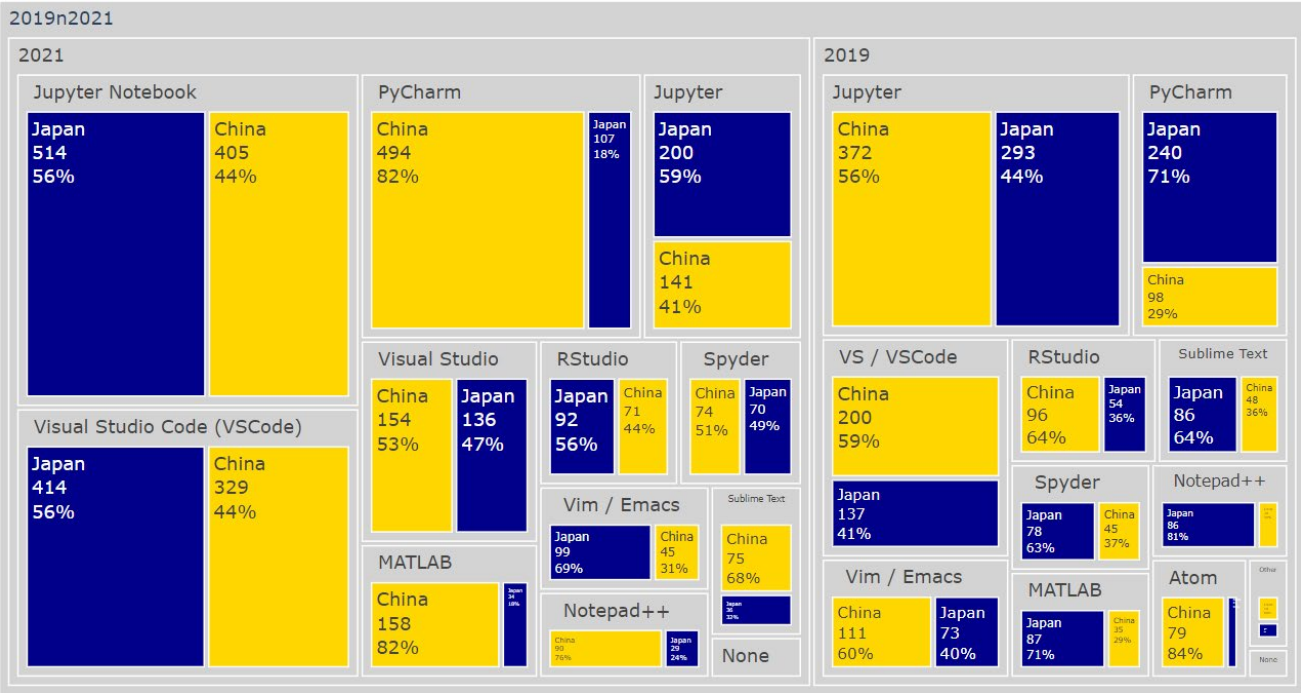
- 데이터 분석과 파이썬 전반에 사용되는 **Matplotlib**가 양국 두 연도에 걸쳐 **가장 높은 점유율**을 보였다.
- 주요 사용되는 4개의 라이브러리의 사용자 수는 ggplot의 일본지표를 제외한 모두가 증가하였다.
- 양국 모두 **라이브러리를 사용하지 않는 캐글러가 증가**하였으며, 일본(+280%)에 비해 중국(+604%)이 더 크게 증가하였다.

Programming Languages

< 2019년과 2021년의 프로그래밍언어 점유율 트리맵 >

- 2019년, 일본과 중국 모두 **Python**의 점유율이 가장 높았으며 **SQL**이 그 뒤를 이었다.
- 일본의 경우 **R의 점유율이 크게 감소**하여 2019년 점유율 3위에서 2021년 7위로 크게 하락했다.
- 중국은 2019년도에 점유율 2위가 SQL이었으나 2021년도에 **C++과 C의 점유율이 급증**하여 4위로 밀려났다.
- 일본 중국 모두 **모든 언어에서의 증가세**가 보였다.

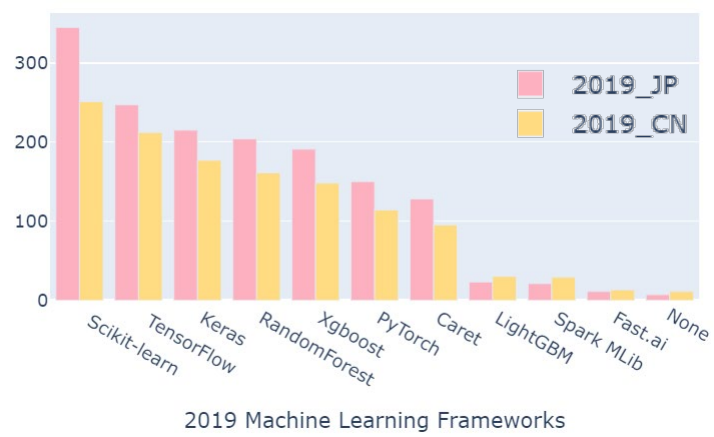
IED’s & Frameworks



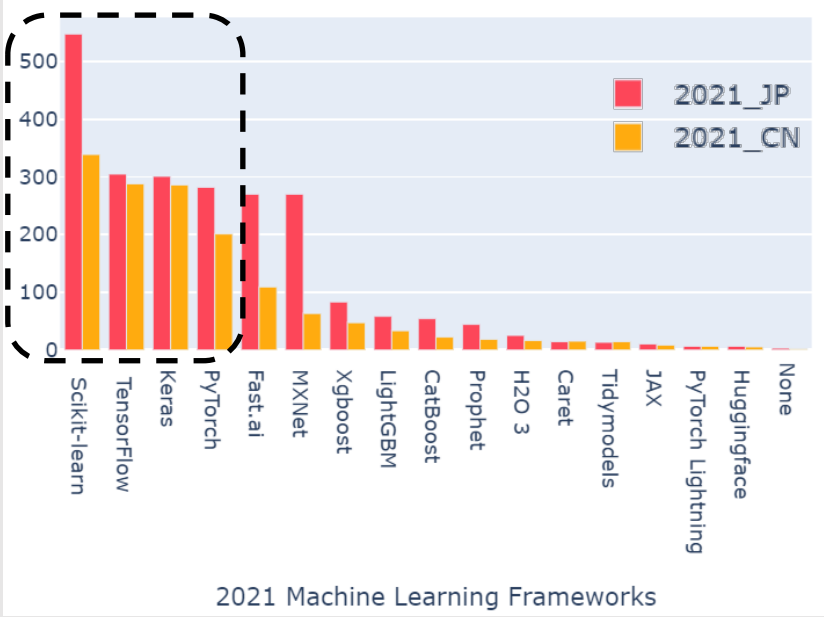
< 2019년과 2021년의 IDE 점유율 트리맵 >

- 양국 모두 **Jupyter Notebook, Visual Studio** 사용자가 **증가** 하였고 **Jupyter**의 사용자가 **감소**하였다.
- 일본은 **PyCharm, Jupyter, MATLAB**의 사용자가 감소하였고, 중국은 **Jupyter, Rstudio, Vim**의 사용자가 감소하였다.
- 21년 중국의 **PyCharm** 사용 캐글러가 494명으로 19년 98명 대비 **404%의 증가율**을 보였다.
- 21년, 일본은 Jupyter Notebook가 중국은 PyCharm의 점유율이 높았다.

IED's & Frameworks



21년상위 4개 프레임워크



- 연도, 국가 모든 차트에서 **Scikit-learn**의 점유율이 **가장 높았다.**
- **Fast.ai**와 **MXNet**의 점유율이 **큰 폭으로 증가**하였다.
- 그중에서도 **일본의 Fast.ai사용자**가 11명에서 270명으로 **증가폭이 매우 컸다.**
- 반면, **XGBoost**와 **Caret**의 점유율은 **크게 하락**하였다.

결론

양국 kaggler의 트렌드 비교

3.

결론

/ 공통점 및 차이점

공통점

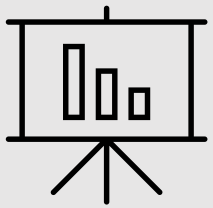
- 캐글러의 수 증가, 젊은층의 가장 높은 점유율
- 대부분의 연령층에서 캐글러의 수가 증가
- 각각 일본 90% 중국 82%로 남성
- 모든 언어 항목에서 증가
- Python사용자의 가장 높은 점유율
- Jupyter Notebook의 사용자 급증
- Scikit-learn의 가장 높은 점유율
- Fast.ai와 MXNet의 점유율이 증가
- 반면, Xgboost와 Caret의 점유율은 크게 하락

차이점

- 2021년 기준, 중국(92%)의 젊은층 비율이 일본(50%)과 비교해 매우 높음
- 일본은 모든 연령층에서 캐글러의 수가 증가했으나, 중국은 그 수가 감소한 연령층 존재
- 일본에서는 JAVA(+144%)가, 중국에서는 C(+240%)의 사용자 증가율이 높았음
- 일본에서는 VS, 중국에서는 PyCharm과 MATLAB 사용자의 증가율이 높았음
- 일본에서는 Scikit-learn의 점유율과 증가율이 매우 높음
- 일본의 Fast.ai 점유율 증가율이 중국의 3배에 달함

다량의 캐글러 신규 유입

다량의 캐글러 신규 유입으로 Kaggle 플랫폼의 확대와 대회의 신뢰도상승 및 시장규모 증가에 따른 일자리 수요 증가 가능성



높은 점유율 항목

점유율이 높거나 점유율이 높아지는 추세가 가파른 항목들이 실무에서 사용될 가능성이 있음

참고문헌

참고문헌

Plotly 막대그래프 튜토리얼

Plotly 막대그래프 속성

Plotly 원그래프 튜토리얼

Plotly 원그래프 속성

[Pandas 기초] 데이터프레임 합치기

[Python] 데이터프레임 열 이름/컬럼명 변경

pandas.DataFrame.insert — pandas 1.3.4 documentation

THANK YOU