



2021 TOKYO Olympic Predict

비타민 7기 구연지, 황예원





INDEX

1

2021 도쿄 올림픽

2

도쿄 올림픽 키워드

3

사용 데이터 소개

4

EDA 및 전처리

5

모델 생성 및 예측



1. 2021 도쿄 올림픽

















The image shows the Tokyo 2020 Olympic logo, which consists of the word "TOKYO" in a stylized font above the year "2020". The logo is made of metallic, three-dimensional blocks. Below the logo are the five Olympic rings, also made of metallic blocks. The background is a dark, textured surface.

TOKYO 2020

TOKYO 2021

- ❑ 2021년 7월 23일 ~ 8월 8일, 총 16일 간 진행되는 32회 올림픽
- ❑ 2020년 7월 24일 ~ 8월 9일에 개최될 예정이었으나 코로나 19의 전세계적인 유행으로 개최가 1년 미뤄짐
- ❑ 테니스, 트라이애슬론, 레슬링, 수영, 양궁, 배드민턴 등 47개 종목
- ❑ 205개국의 1만 1천여명의 선수가 출전

도쿄올림픽 현황

순위	NOC	🏅	🥇1	🏅	🥈2	🥉3	합계	합계 순위
1	 중국		15		7	10	32	2
2	 일본		15		4	6	25	4
3	 미국		14		14	10	38	1
4	 ROC		8		12	9	29	3
5	 오스트레일리아		8		2	10	20	5
6	 영국		5		7	6	18	7
7	 대한민국		4		3	5	12	10
8	 프랑스		3		5	3	11	11
9	 독일		3		3	7	13	8
10	 캐나다		3		3	5	11	11
11	 이탈리아		2		7	10	19	6
12	 네덜란드		2		7	4	13	8
13	 뉴질랜드		2		3	1	6	14
14	 헝가리		2		1	2	5	17

~ 2021.07.30



2. 도쿄 올림픽 키워드

: WordCloud, text_to_word_sequence



도쿄 올림픽 키워드

Tokyo Olympics 2020 Tweets

해시태그 (#Tokyo2020) 가 붙은 트위터
게시글들의 텍스트 데이터



tweets_2020.head()

	id	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	retweets	favorites	is_retweet
0	141888645105356803	Abhishek Srivastav	Udupi, India	Trying to be mediocre in many things	2021-02-01 06:33:51	45	39	293	False	2021-07-24 10:59:49	Let the party beginWn#Tokyo2020	['Tokyo2020']	Twitter for Android	0.0	0.0	False
1	141888377680678918	Saikhom Mirabai Channuin	Manipur, India	Indian weightlifter 48 kg category. Champion 🏆	2018-04-07 10:10:22	5235	5	2969	False	2021-07-24 10:58:45	Congratulations #Tokyo2020 https://t.co/8OFKMs...	['Tokyo2020']	Twitter for Android	0.0	0.0	False
2	141888260886073345	Big Breaking	Global	All breaking news related to Financial Market...	2021-05-29 08:51:25	3646	3	5	False	2021-07-24 10:58:17	Big Breaking Now WnWnTokyo Olympic Update WnWn...	NaN	Twitter for Android	0.0	1.0	False
3	141888172864299008	International Hockey Federation	Lausanne	Official International Hockey Federation Twitt...	2010-10-20 10:45:59	103975	2724	36554	True	2021-07-24 10:57:56	Q4: g83-1zAWnWnGreat Britain finally find a wa...	NaN	Twitter Web App	1.0	0.0	False
4	1418886894478270464	Cameron Hart	Australia	Football & Tennis Coach	2020-10-31 08:46:17	6	37	31	False	2021-07-24 10:52:51	All I can think of every time I watch the ring...	['Tokyo2020', 'ArtisticGymnastics', '7Olympics...']	Twitter for iPhone	0.0	0.0	False

Wordcloud, text_to_word_sequence 이용한 시각화

도쿄 올림픽 키워드

```
[108] BoW = []
      from tensorflow.keras.preprocessing.text import text_to_word_sequence
      for text in text_list:
          for word in text_to_word_sequence(text):
              BoW.append(word)

      print(len(BoW), BoW[:5])

531388 ['let', 'the', 'party', 'begin', 'congratulations']
```

text_to_word_sequence를 이용해 제공된 텍스트를 단어로 분리

=> WordCloud를 통해 키워드 시각화



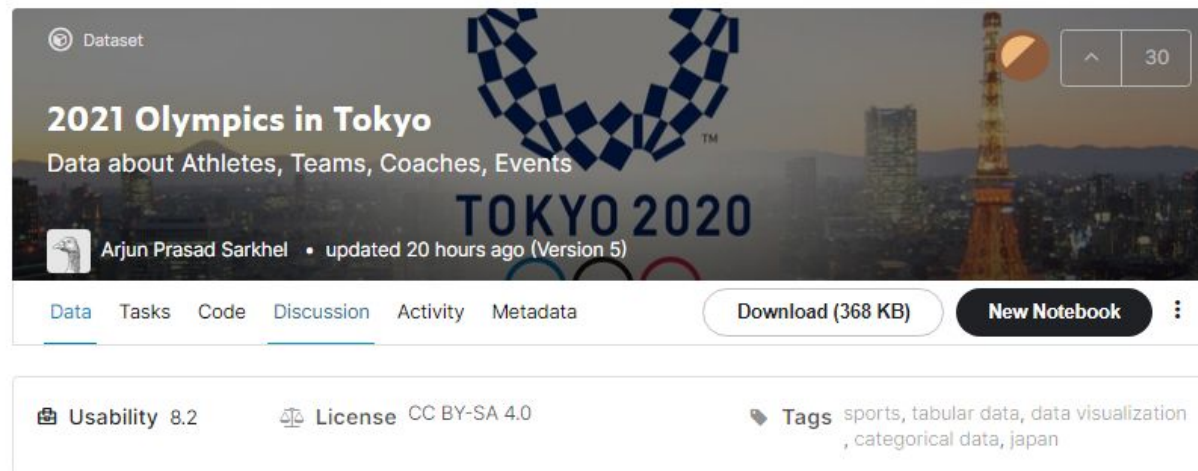
3. 사용 데이터 소개



사용 데이터 소개

1. 2021 Olympics in Tokyo

2021 도쿄 올림픽에 출전한 선수, 국가, 코치, 종목에 관한 데이터셋



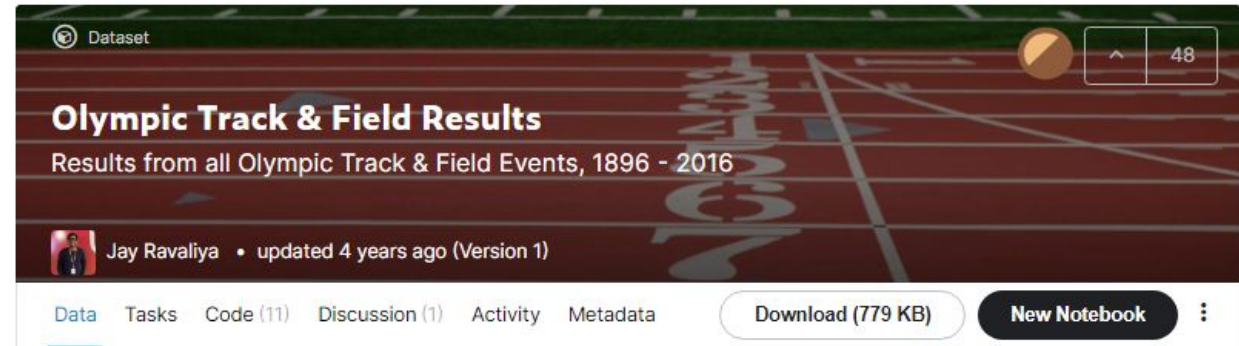
2. 120 years of Olympic history

1896 - 2016 올림픽 사이 국가, 선수, 종목별 기록

사용 데이터 소개

3. Olympic Track & Field Results

1896 - 2016 올림픽 사이 육상 경기 기록



4. Olympics Athlete Events Analysis

1896 - 2016 올림픽 사이 국가, 선수,
종목별 기록

사용 데이터 소개

5. Olympic Games Medals

1986 - 2018 올림픽 사이 메달 수상자 기록



6. Olympic Games Hosts

1896 - 2028 올림픽 사이 개최 국가, 종목 개수, 개최일 데이터

사용 데이터 소개

7. GDP, 1인당 GDP, 인구

1980 - 2024 국가별 GDP, 1인당 GDP, 총
인구 데이터

IMF DATA

What's New

The IMF has released end-2019 results of the Coordinated Direct Investment Survey (CDIS). The new data show that the US holds both the largest inward direct investment position with \$4.5 trillion and the largest outward direct investment position with \$6.0 trillion. The CDIS is a global survey of cross-border direct investment holdings with immediate counterpart country information.

[SEE CDIS ►](#)

Popular Data

[World Economic Outlook Latest Update](#)

[Global Financial Stability Report Latest Update](#)

[Fiscal Monitor Latest Update](#)

[IMF Data Mapper®](#)

[IMF Finances](#)

[SDRs per Currency Unit](#)



4. EDA 및 전처리



EDA 및 전처리

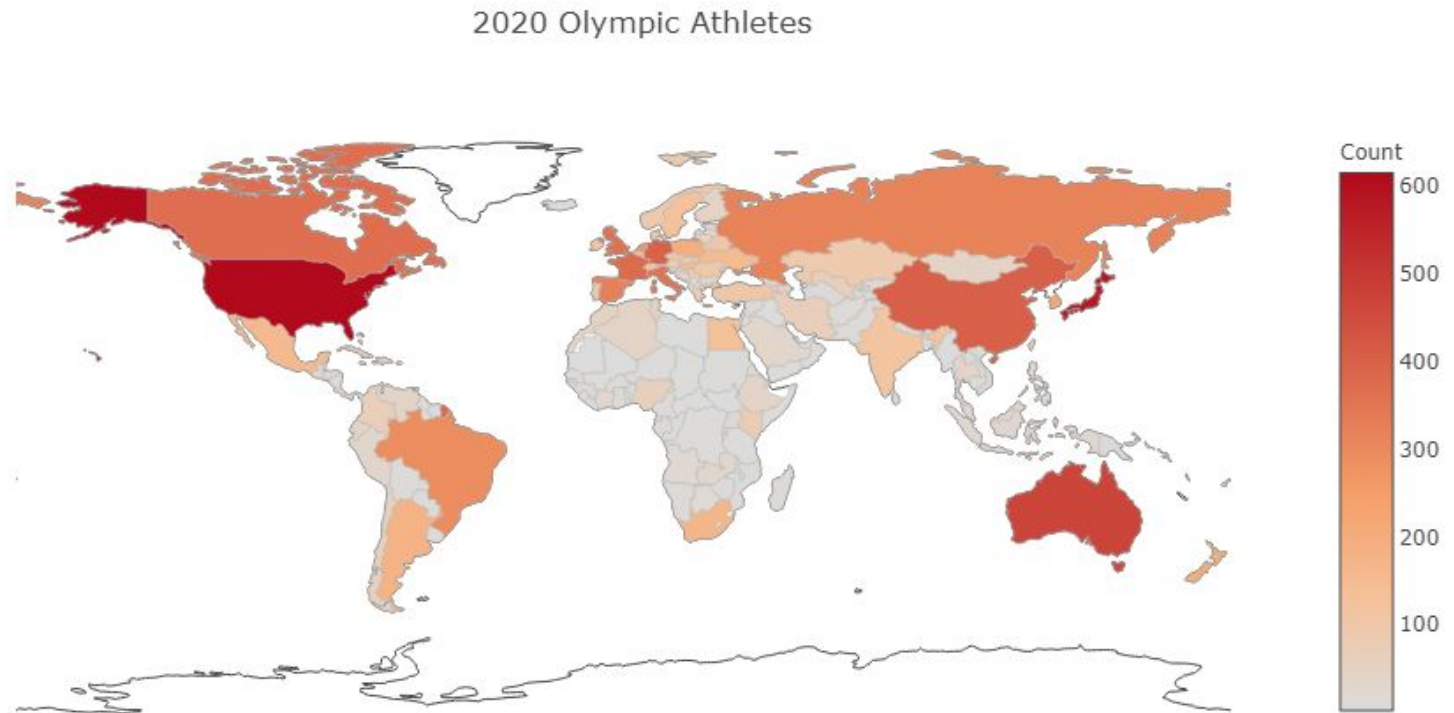
2020 Tokyo Olympic 시행 종목별 인원수

	Discipline	Female	Male	Total
0	3x3 Basketball	32	32	64
1	Archery	64	64	128
2	Artistic Gymnastics	98	98	196
3	Artistic Swimming	105	0	105
4	Athletics	969	1072	2041
5	Badminton	86	87	173
6	Baseball/Softball	90	144	234
7	Basketball	144	144	288
8	Beach Volleyball	48	48	96
9	Boxing	102	187	289
10	Canoe Slalom	41	41	82
11	Canoe Sprint	123	126	249
12	Cycling BMX Freestyle	10	9	19
13	Cycling BMX Racing	24	24	48
14	Cycling Mountain Bike	38	38	76
15	Cycling Road	70	131	201
16	Cycling Track	90	99	189
17	Diving	72	71	143
18	Equestrian	73	125	198
19	Fencing	107	108	215
20	Football	264	344	608
21	Golf	60	60	120
22	Handball	168	168	336
23	Hockey	192	192	384
24	Judo	192	201	393
25	Karate	40	42	82
26	Marathon Swimming	25	25	50
27	Modern Pentathlon	36	36	72
28	Rhythmic Gymnastics	96	0	96
29	Rowing	257	265	522
30	Rugby Sevens	146	151	297
31	Sailing	175	175	350
32	Shooting	178	178	356
33	Skateboarding	40	40	80
34	Sport Climbing	20	20	40
35	Surfing	20	20	40
36	Swimming	361	418	779
37	Table Tennis	86	86	172
38	Taekwondo	65	65	130
39	Tennis	94	97	191
40	Trampoline Gymnastics	16	16	32
41	Triathlon	55	55	110
42	Volleyball	144	144	288
43	Water Polo	122	146	268
44	Weightlifting	98	99	197
45	Wrestling	96	193	289

EDA 및 전처리

	country_name	count
0	Afghanistan	5
1	Albania	8
2	Algeria	41
3	American Samoa	5
4	Andorra	2
...
201	Virgin Islands, British	3
202	Virgin Islands, US	4
203	Yemen	3
204	Zambia	29
205	Zimbabwe	5

206 rows × 2 columns



각 나라의 출전인원을 iplot 통해 시각화

EDA 및 전처리

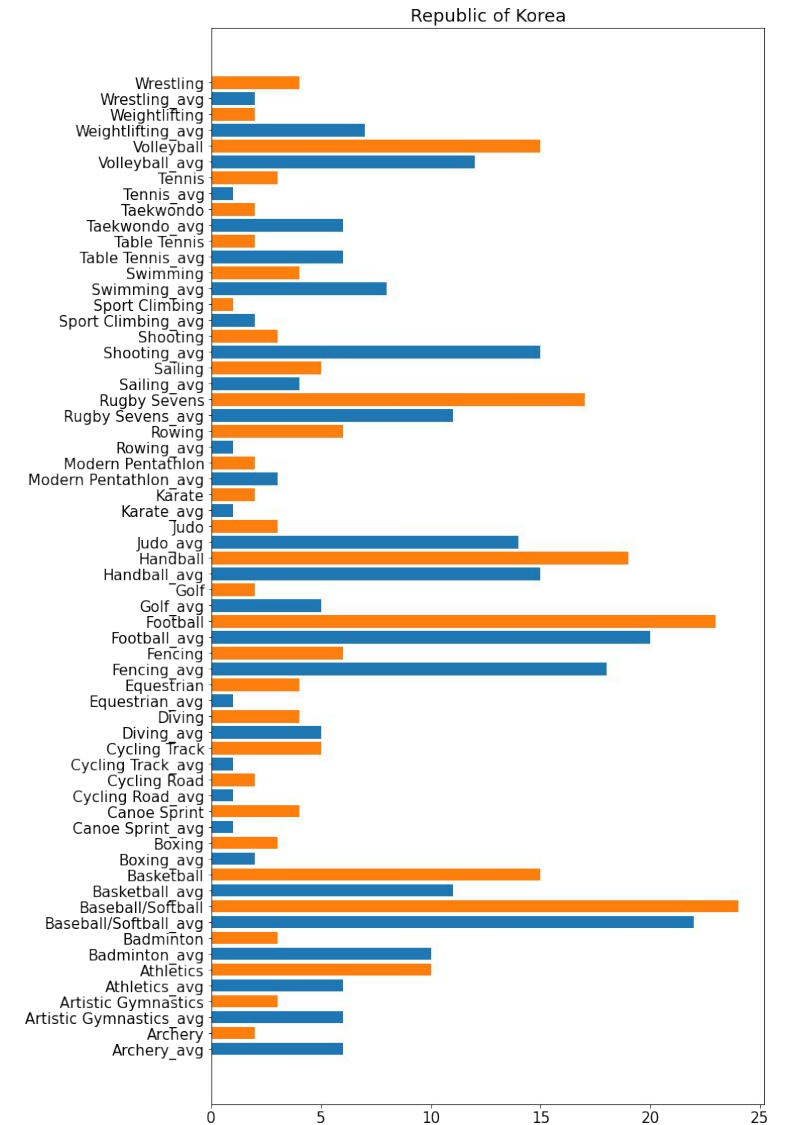
Discipline	count	21	Sailing	4	
0	Archery	6	22	Shooting	15
1	Artistic Gymnastics	6	23	Sport Climbing	2
2	Athletics	6	24	Swimming	8
3	Badminton	10	25	Table Tennis	6
4	Baseball/Softball	22	26	Taekwondo	6
5	Basketball	11	27	Tennis	1
6	Boxing	2	28	Volleyball	12
7	Canoe Sprint	1	29	Weightlifting	7
8	Cycling Road	1	30	Wrestling	2
9	Cycling Track	1			
10	Diving	5			
11	Equestrian	1			
12	Fencing	18			
13	Football	20			
14	Golf	5			
15	Handball	15			
16	Judo	14			
17	Karate	1			
18	Modern Pentathlon	3			
19	Rowing	1			
20	Rugby Sevens	11			

한국의 종목별 인-

전 세계 종목별 인-



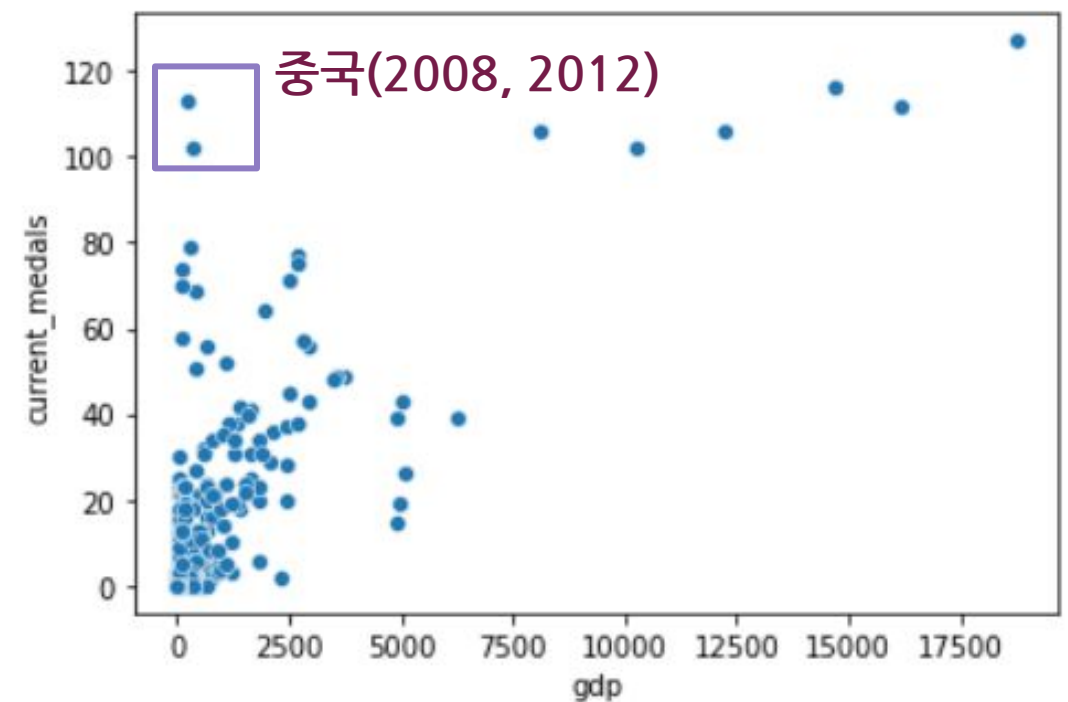
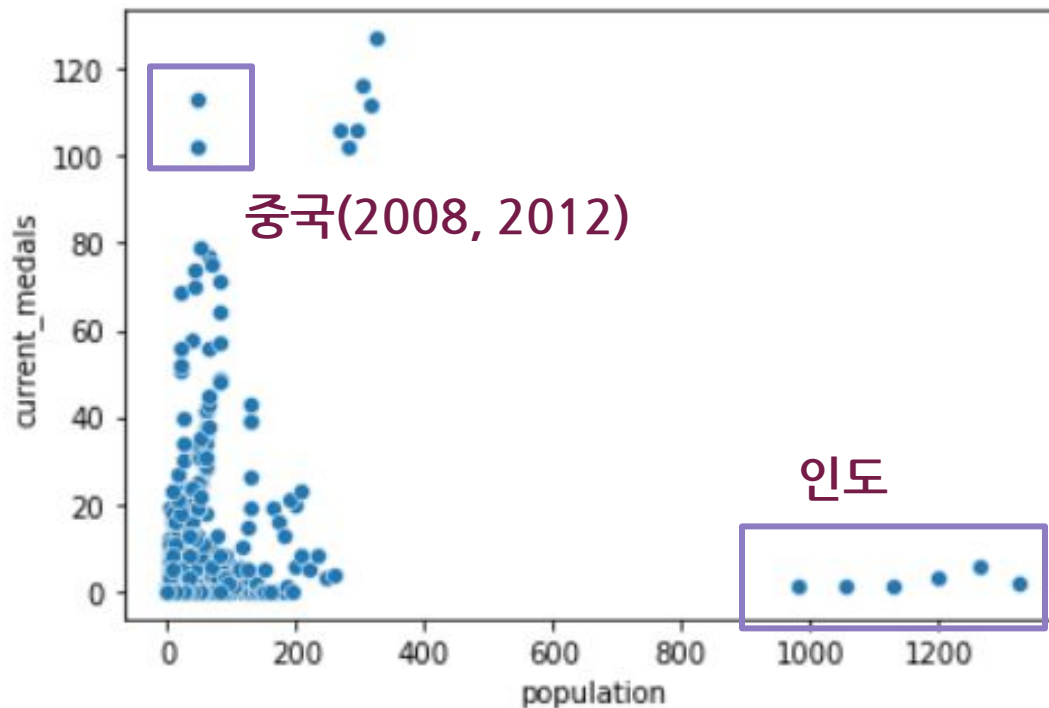
한국의 종목별 인원수와
전 세계 종목별 인원수 평균 비교



EDA 및 전처리

GDP: 경제력이 좋지 않으면 스포츠에 투자를 하지 못함

인구수: 인구수가 적으면 운동에 재능이 있는 사람들의 수도 적을 가능성이 높음



EDA 및 전처리

	Afghanistan	Albania	Algeria	Angola	Antigua and Barbuda	Argentina	Armenia	Aruba	Australia	Austria	Azerbaijan
1980	no data	1.946	42.346	6.639	0.131	233.696	no data	no data	162.628	80.923	no data
1981	no data	2.229	44.372	6.214	0.148	189.802	no data	no data	188.067	70.121	no data
1982	no data	2.296	44.78	6.214	0.164	94.25	no data	no data	186.709	70.111	no data
1983	no data	2.319	47.529	6.476	0.182	116.267	no data	no data	179.151	71.032	no data
1984	no data	2.29	51.513	6.864	0.208	130.544	no data	no data	196.777	67.007	no data
5 rows x 227 columns											

GDP

	Afghanistan	Albania	Algeria	Angola	Antigua and Barbuda	Argentina	Armenia	Aruba	Australia
1980	no data	2.672	18.666	8.91	0.068	27.95	no data	no data	14.802
1981	no data	2.726	19.246	9.151	0.068	28.45	no data	no data	15.039
1982	no data	2.784	19.864	9.393	0.067	28.93	no data	no data	15.289
1983	no data	2.844	20.516	9.639	0.066	29.34	no data	no data	15.483
1984	no data	2.904	21.175	9.894	0.065	29.84	no data	no data	15.677

Population

	year	country	gdp	population
0	1996	Afghanistan	no data	no data
1	1996	Albania	3.2	3.168
2	1996	Algeria	46.941	28.566
3	1996	Angola	7.994	15.214
4	1996	Antigua and Barbuda	0.634	0.068

국가 - 연도별 GDP와 인구수 추가

EDA 및 전처리

```
medals_total.head()
```

	discipline_title	event_title	event_gender	medal_type	participant_type	participant_title	athlete_name	athlete_surname	country_name	country_code
0	Skeleton	Women	Women	GOLD	Athlete	NaN	Lizzy	YARNOLD	Great Britain	GB
1	Skeleton	Women	Women	SILVER	Athlete	NaN	Jacqueline	LOELLING	Germany	DE
2	Skeleton	Women	Women	BRONZE	Athlete	NaN	Laura	DEAS	Great Britain	GB
3	Skeleton	Men	Men	GOLD	Athlete	NaN	Sungbin	YUN	Republic of Korea	KR
4	Skeleton	Men	Men	SILVER	Athlete	NaN	Nikita	TREGUBOV	Olympic Athletes from Russia	NaN

```
medals = medals_total.groupby(['country_name', 'year']).agg('count')[['medal_type']]
medals.head()
```

medal_type		
country_name	year	
Afghanistan	2008	1
	2012	1
	1992	2
Algeria	1996	3
	2000	5



```
medals_ = medals.reset_index(drop = False)
medals_.columns = ['country', 'year', 'medals']
```

```
medals_.head()
```

	country	year	medals
0	Afghanistan	2008	1
1	Afghanistan	2012	1
2	Algeria	1992	2
3	Algeria	1996	3
4	Algeria	2000	5

해당 연도의 나라별 메달 취득 누적 개수 데이터프레임으로 변환

EDA 및 전처리

```
medals_before = medals_.copy()
medals_before['year'] = medals_before['year'] + 4
medals_before.head()
```

	country	year	medals
0	Afghanistan	2012	1
1	Afghanistan	2016	1
2	Algeria	1996	2
3	Algeria	2000	3
4	Algeria	2004	5

```
medal_num = pd.merge(medals_, medals_before, on=['country', 'year'], how = 'outer').fillna(0)
medal_num.head()
```

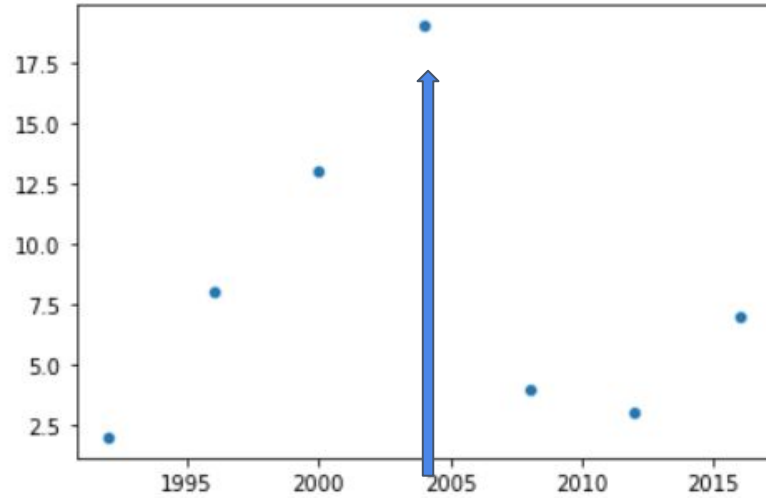
	country	year	current_medals	before_medals
0	Afghanistan	2008	1.0	0.0
1	Afghanistan	2012	1.0	1.0
2	Algeria	1992	2.0	0.0
3	Algeria	1996	3.0	2.0
4	Algeria	2000	5.0	3.0

지난 동계올림픽에서의 누적 메달 수 (전적) 를 before_medals로 추가

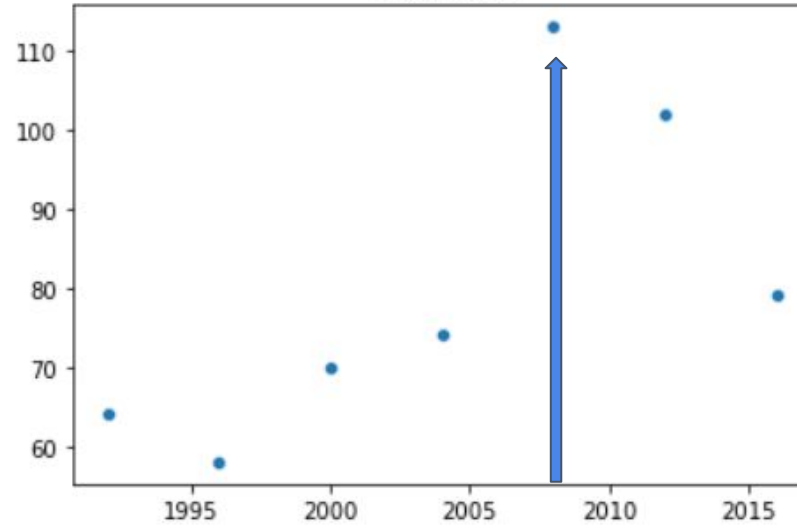
	year	country	gdp	population	current_medals	before_medals
0	1996	Afghanistan	NaN	NaN	0.0	0.0
1	1996	Albania	3.2	3.168	0.0	0.0
2	1996	Algeria	46.941	28.566	3.0	2.0
3	1996	Angola	7.994	15.214	0.0	0.0
4	1996	Antigua and Barbuda	0.634	0.068	0.0	0.0
5	1996	Argentina	304.282	35.196	3.0	2.0
6	1996	Armenia	1.597	3.17	2.0	0.0
7	1996	Aruba	1.38	0.083	0.0	0.0
8	1996	Australia	423.544	18.33	51.0	30.0
9	1996	Austria	237.343	7.959	3.0	24.0

EDA 및 전처리

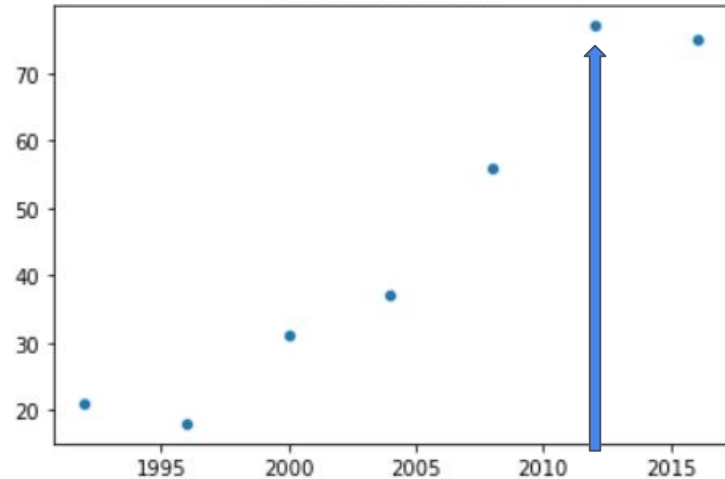
Greece-2004



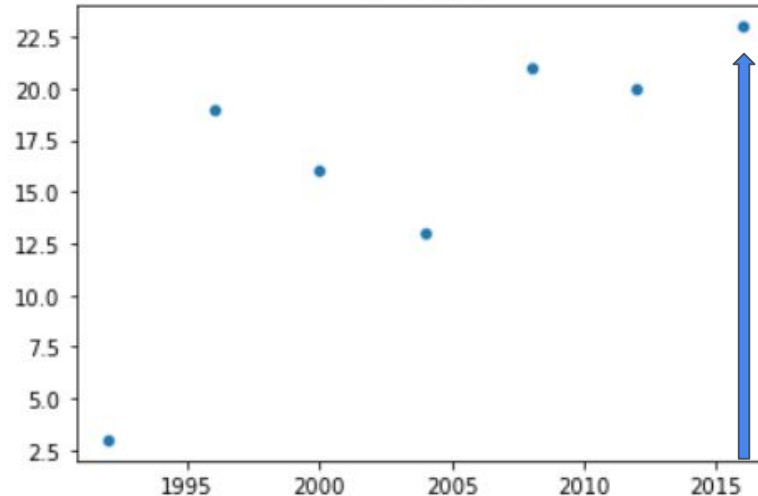
China-2008



Great Britain-2012



Brazil-2016



	country	year	host
43	Greece	2004	1
45	People's Republic of China	2008	1
49	Great Britain	2012	1
53	Brazil	2016	1
56	Japan	2020	1

	country	year	next_host
41	Australia	1996	1
43	Greece	2000	1
45	People's Republic of China	2004	1
49	Great Britain	2008	1
53	Brazil	2012	1

개최지, 다음 개최지:
개최국 프리미엄
→ 개최지 또는 다음
개최지이면 1, 아니면 0

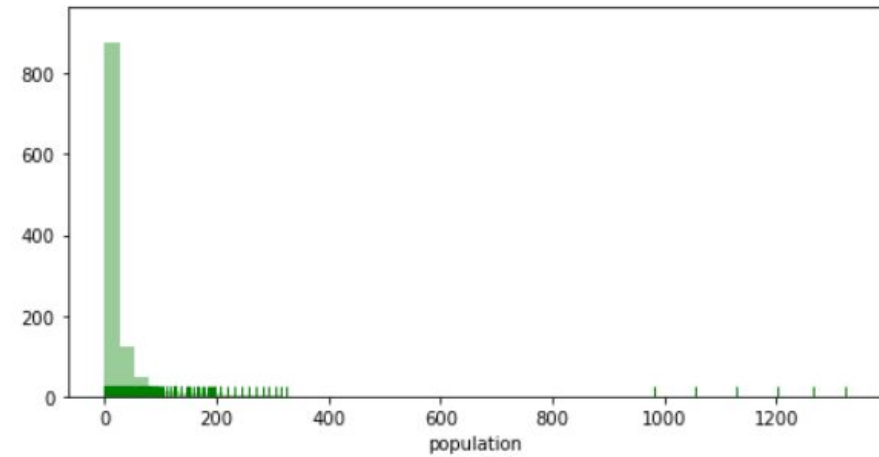
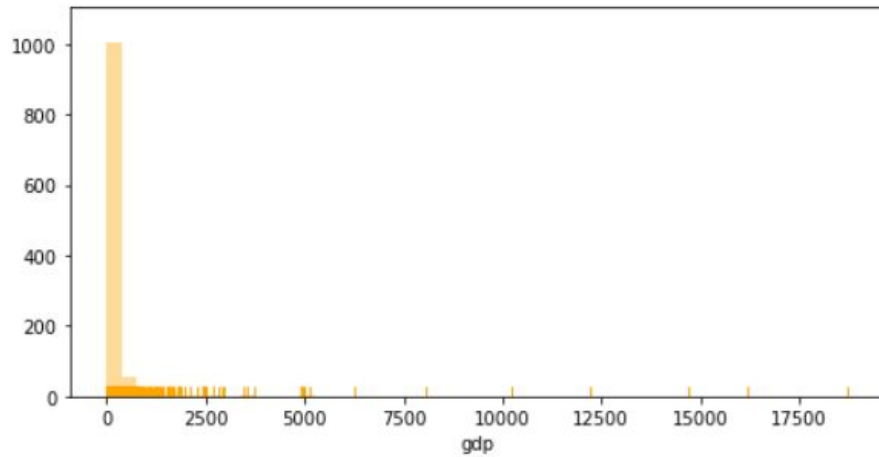
EDA 및 전처리

이번 대회 메달 개수 = $\beta_0 + \beta_1 * \text{GDP} + \beta_2 * \text{인구수} + \beta_3 * \text{연도} + \beta_4 * \text{개최지 변수} + \beta_5 * \text{다음 개최지 변수} + \beta_6 * \text{이전대회 획득 메달 개수} + \epsilon$

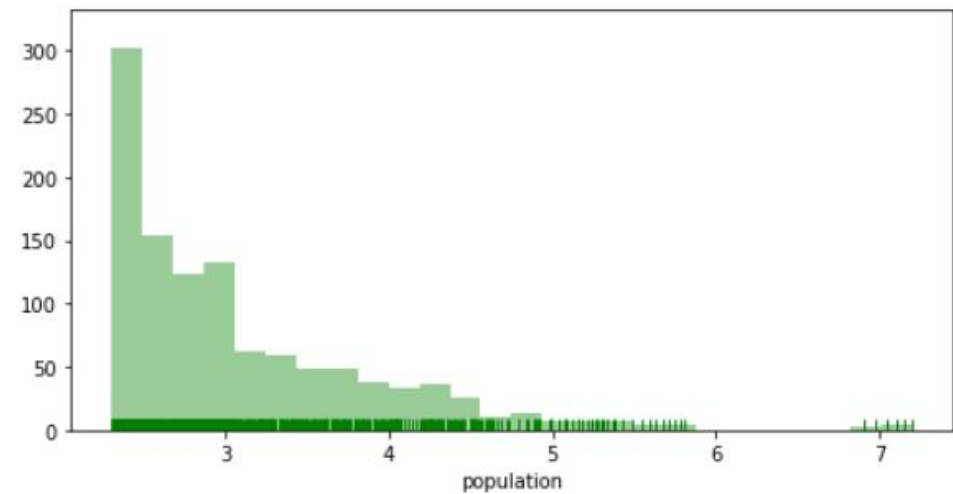
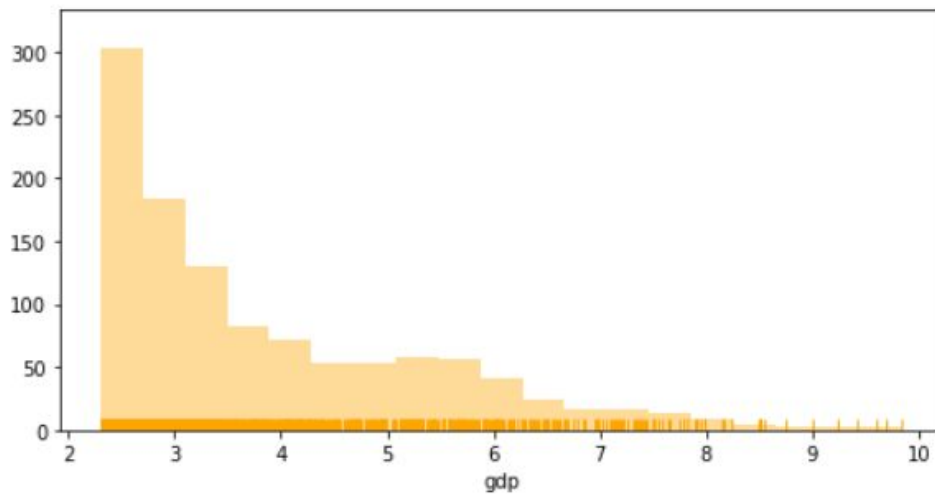


	country	year	next_host	gdp	population	current_medals	before_medals	host
0	Australia	1996	1.0	423.544	18.33	51.0	30.0	0.0
1	Greece	2000	1.0	131.082	10.776	13.0	8.0	0.0
2	People's Republic of China	2004	1.0	116.335	42.368	74.0	70.0	0.0
3	Great Britain	2008	1.0	2952.326	61.824	56.0	37.0	0.0
4	Brazil	2012	1.0	2464.054	198.315	20.0	21.0	0.0
...
1631	Ukraine	2010	0.0	0	0	0.0	3.0	0.0
1632	Ukraine	2022	0.0	0	0	0.0	1.0	0.0
1633	Unified Team	1996	0.0	0	0	0.0	141.0	0.0
1634	United States of America	2022	0.0	0	0	0.0	26.0	0.0
1635	Uzbekistan	1998	0.0	0	0	0.0	1.0	0.0

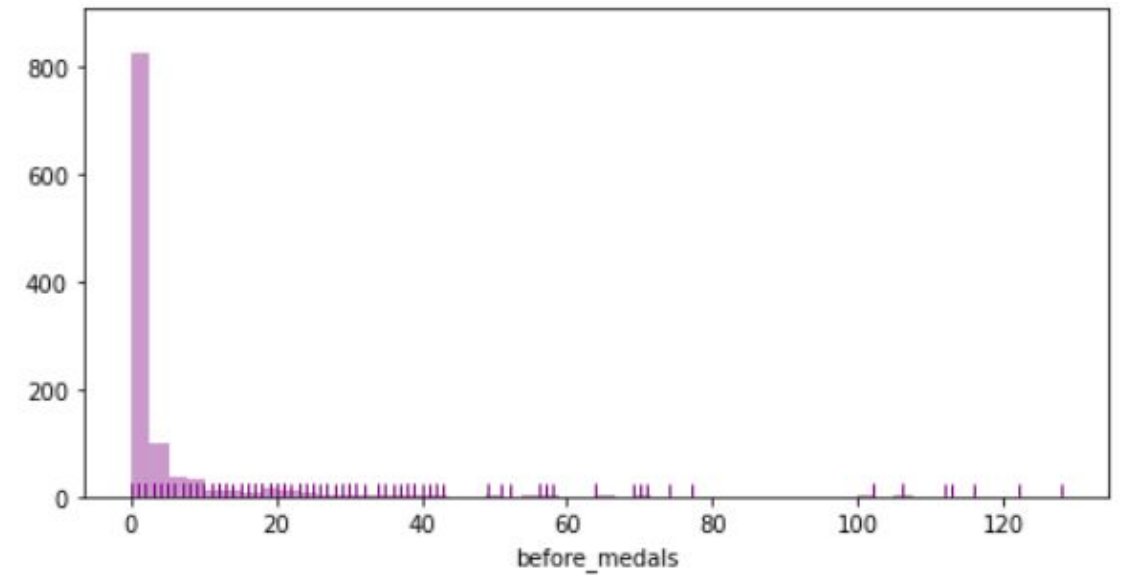
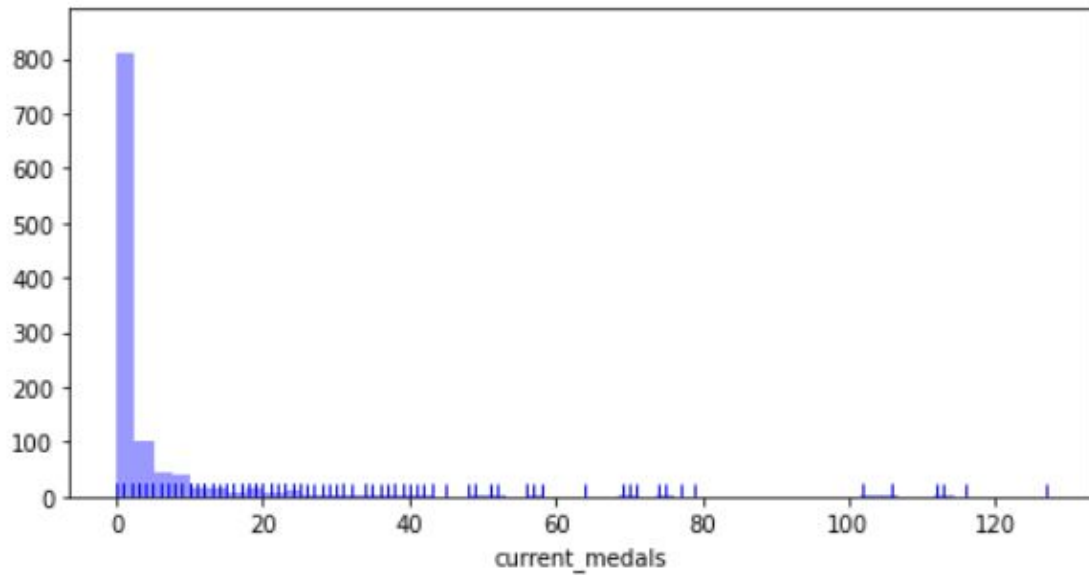
EDA 및 전처리



로그 변환
 $\log(x+10)$



EDA 및 전처리



로그 변환이 필요한가?
메달 '개수'에 의미가 있음

EDA 및 전처리

	year	next_host	gdp	population	current_medals	before_medals	host
year	1.000000	0.000908	0.141255	0.057689	-0.059485	0.072162	0.000908
next_host	0.000908	1.000000	0.137694	0.101943	0.158644	0.131392	-0.004184
gdp	0.141255	0.137694	1.000000	0.751176	0.360964	0.422275	0.146931
population	0.057689	0.101943	0.751176	1.000000	0.233362	0.278439	0.113636
current_medals	-0.059485	0.158644	0.360964	0.233362	1.000000	0.791044	0.267137
before_medals	0.072162	0.131392	0.422275	0.278439	0.791044	1.000000	0.236189
host	0.000908	-0.004184	0.146931	0.113636	0.267137	0.236189	1.000000

종속변수 로그 변환O

	year	next_host	gdp	population	current_medals	before_medals	host
year	1.000000	0.000908	0.141255	0.057689	-0.066292	0.130550	0.000908
next_host	0.000908	1.000000	0.137694	0.101943	0.164198	0.150940	-0.004184
gdp	0.141255	0.137694	1.000000	0.751176	0.399265	0.461242	0.146931
population	0.057689	0.101943	0.751176	1.000000	0.227275	0.270710	0.113636
current_medals	-0.066292	0.164198	0.399265	0.227275	1.000000	0.808371	0.208225
before_medals	0.130550	0.150940	0.461242	0.270710	0.808371	1.000000	0.197488
host	0.000908	-0.004184	0.146931	0.113636	0.208225	0.197488	1.000000

종속변수 로그 변환X



5. 모델 생성 및 예측

: 회귀 분석



모델 생성 및 예측

회귀분석

- XGBRegressor
- LGBMRegressor
- RandomForestRegressor
- SVR
- LinearRegression

```
from sklearn.model_selection import train_test_split
from xgboost import XGBRegressor
from lightgbm import LGBMRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.svm import SVR
from sklearn.linear_model import LinearRegression
```

```
rf = RandomForestRegressor()
svr = SVR()
lr = LinearRegression()
xgb = XGBRegressor()
lgbm = LGBMRegressor()
```

```
target = total_train['current_medals']
feature = total_train.drop('current_medals', axis=1)
```

```
X_train, X_test, y_train, y_test = train_test_split(feature, target, test_size=0.2)
```

메달 '개수'에 대한 예측 → MAE 사용

모델 생성 및 예측

하이퍼 파라미터 튜닝 후 MAE

	종속변수 $\log O$	종속변수 $\log X$
RandomForest Regressor	3.253	2.4615
SVR	5.178	5.512
XGBRegressor	3.146	2.503
LGBMRegressor	3.074	2.594
Linear Regression	3.950	3.557

모델 생성 및 예측

예측 결과값

	2016	RF _log	XGB _log	LGBM _log	Linear _log	RF	XGB	LGBM	Linear
미국	127	108	112	92	111	108	82	95	98
중국	79	86	76	82	68	87	69	79	59
영국	75	83	79	80	65	79	70	78	58
러시아	59	67	62	76	51	51	38	63	42
독일	48	48	33	50	41	46	38	45	38
프랑스	45	42	29	48	49	35	33	41	49
일본	43	42	25	52	52	36	38	46	55
호주	34	23	8	20	29	24	17	25	27
이탈리아	31	25	9	29	26	23	15	30	24
캐나다	24	20	4	20	19	20	11	18	19

모델 생성 및 예측

	2016	RF _log	XGB _log	LGBM _log	Linear _log	RF	XGB	LGBM	Linear
브라질	23	21	3	25	18	22	12	22	17
뉴질랜드	23	19	1	16	19	16	5	14	18
한국	22	21	4	25	18	21	13	20	18
네덜란드	21	21	4	18	17	19	10	17	17
스페인	19	17	4	21	15	20	11	22	15
아제르바이잔	18	15	1	13	14	15	4	14	12
카자흐스탄	18	15	0	13	14	16	5	14	14
덴마크	18	14	0	12	14	14	5	12	13
헝가리	16	14	4	11	12	15	3	11	12
케냐	13	13	4	11	9	13	2	13	8



감사합니다