
Introduction To Data Analysis
데이터분석입문
Lecture 29. pandas로 올림픽 메달 집계 분석
인공지능소프트웨어학과 강환수 교수

학습개요

- 위키피디아의 '올림픽 메달 집계' 데이터 내려 받아 특정 테이블 활용
- 하계 올림픽 메달 집계를 위한 데이터프레임 생성
- 특정 메달 순서로 정렬하고 '대한민국'이 있는 행 검색

학습목표

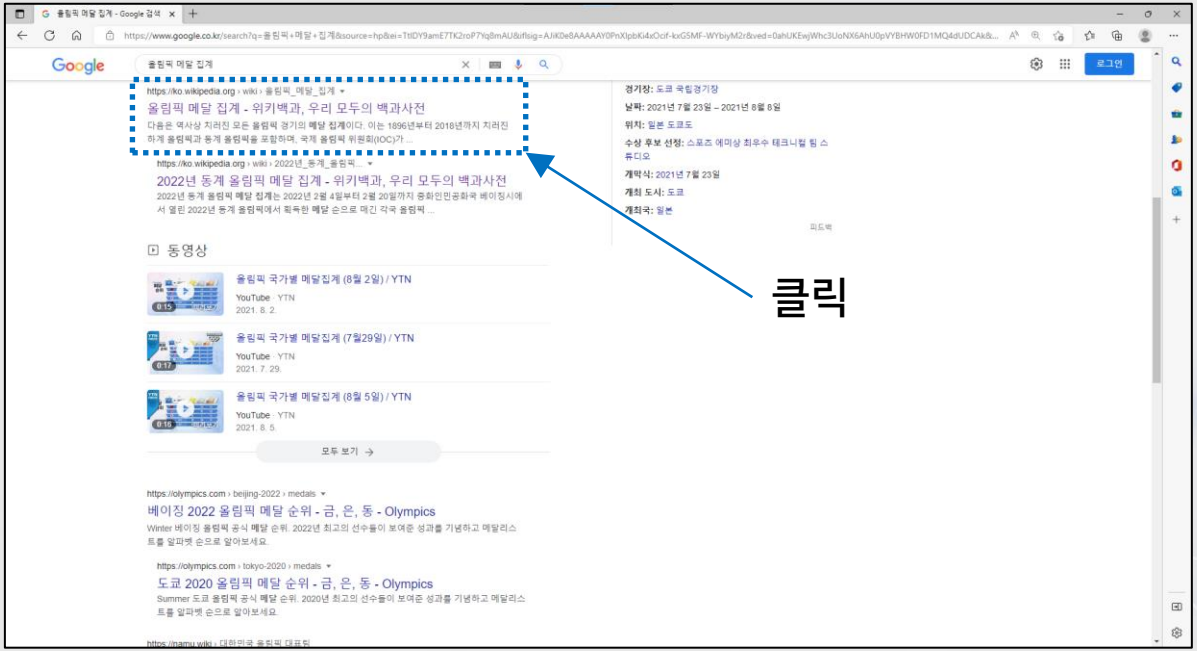
- 위키피디아의 '올림픽 메달 집계' 전체 페이지에서 데이터 읽어와 특정 데이터프레임을 준비할 수 있다.
- 하계 올림픽 메달 집계를 위한 데이터프레임을 생성할 수 있다.
- 특정 메달 순서로 정렬하고 '대한민국'이 있는 행 검색할 수 있다.

LESSON 01

인터넷 자료를 DataFrame에 저장

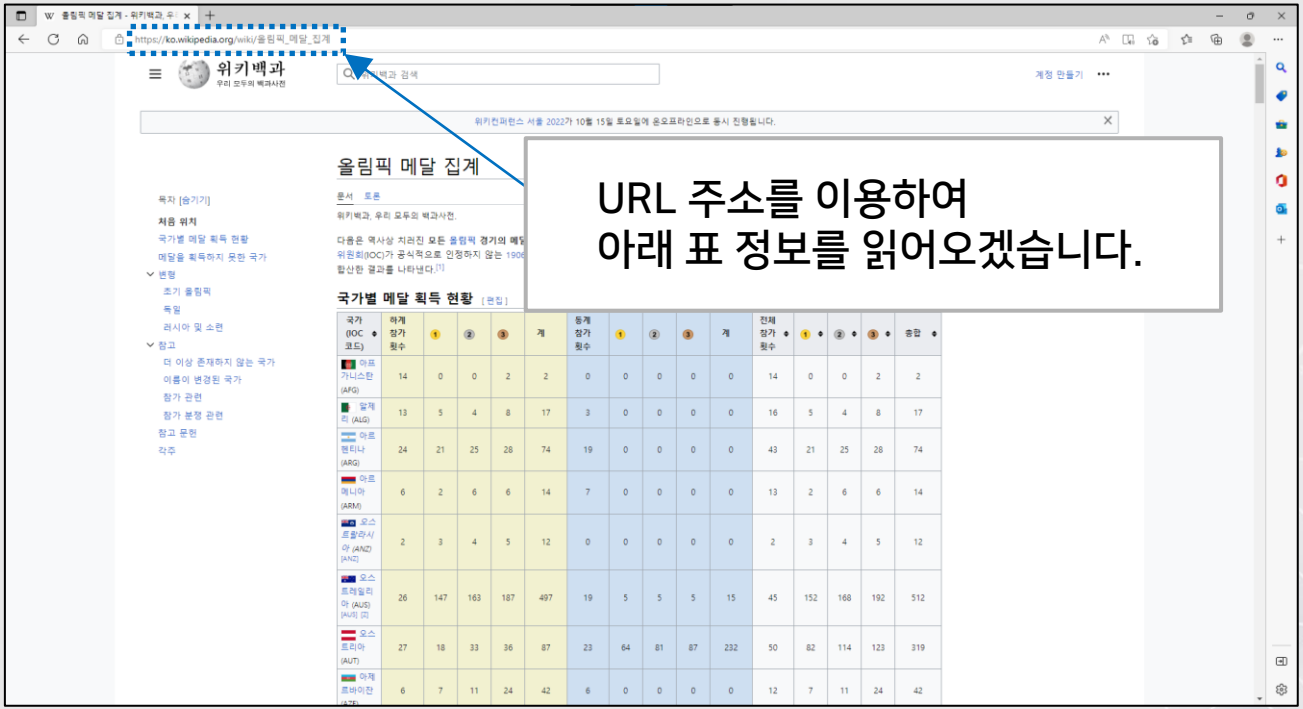
올림픽 메달 기록 데이터 (1/2)

- pandas 라이브러리 실습을 위해 올림픽 메달 기록 데이터를 활용하겠습니다.
- Google에서 “올림픽 메달 집계”라고 검색 → “올림픽 메달 집계 – 위키백과, 우리 모두의 백과사전” 검색 결과 클릭



올림픽 메달 기록 데이터 (2/2)

위키백과(Wikipedia)에서 제공하는 하계 및 동계 올림픽 메달 획득 결과 표



pandas를 이용하여 URL로부터 데이터 읽어 오기

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
print(df)
```

```
[            국가 (IOC 코드)  하계 참가 횟수  Unnamed: 2  Unnamed: 3  Unnamed: 4      계  ₩
0          아프가니스탄 (AFG)         14           0           0           2      2
1              알제리 (ALG)         13           5           4           8     17
2          아르헨티나 (ARG)         24          21          25          28     74
3          아르메니아 (ARM)          6           2
4    오스트랄라시아 (ANZ) [ANZ]          2           3
..                   ...         ...         ...
148        독립 (IOA) [IOA]          3           1
149      독립 참가 (IOP) [IOP]          1           0
150  러시아 출신 올림픽 선수 (OAR)          0
151        혼성 (ZZX) [ZZX]          3           8
152                  총합         28        5116        5082        5450  15666

     동계 참가 횟수  Unnamed: 7  Unnamed: 8  Unnamed: 9  Unnamed: 10  계.1  전체 참가 횟수  ₩
0            0           0           0           0           0    14
1            3           0           0           0           0    16
2           19           0           0           0           0    43
3            7           0           0           0           0    13
4            0           0           0           0           0     2
```

대괄호 [ ]가 있습니다!
그 뜻은 자료형이 리스트라는 거겠죠?
df의 구성을 살펴보겠습니다.

DataFrame의 리스트

[df1, df2, ···]

```
In [2]: print(type(df))
        print(type(df[0]))

        <class 'list'>
        <class 'pandas.core.frame.DataFrame'>
```

데이터 프레임 df[0] 살펴보기

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df[0]
```

	국가 (IOC 코드)	하계 참가 회수	Unnamed: 2	Unnamed: 3	Unnamed: 4	계	동계 참가 회수	Unnamed: 7	Unnamed: 8	Unnamed: 9	계.1	전체 참가 회수	Unnamed: 12	Unnamed: 13	Unnamed: 14	총합
0	아프가니스탄 (AFG)	14	0	0	2	2	0	0	0	0	0	14	0	0	2	2
1	알제리 (ALG)	13	5	4	8	17	3	0	0	0	0	16	5	4	8	17
2	아르헨티나 (ARG)	24	21	25	28	74	19	0	0	0	0	43	21	25	28	74
3	아르메니아 (ARM)	6	2	6	6	14	7	0	0	0	0	13	2	6	6	14
4	오스트랄라시아 (ANZ) [ANZ]	2	3	4	5	12	0	0	0	0	0	2	3	4	5	12
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

df[0]에 우리가 원하는 데이터가 담겨있네요!

인덱스(Index)와 열 이름(Column Name) 확인

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df[0]
```

열 이름

인덱스

	국가 (IOC 코드)	하계 참가 횟수	Unnamed: 2	Unnamed: 3	Unnamed: 4	계	동계 참가 횟수	Unnamed: 7	Unnamed: 8	Unnamed: 9	계.1	전체 참가 횟수	Unnamed: 12	Unnamed: 13	Unnamed: 14	총합
0	아프가니스탄 (AFG)	14	0	0	2	2	0	0	0	0	0	14	0	0	2	2
1	알제리 (ALG)	13	5	4	8	17	3	0	0	0	0	16	5	4	8	17
2	아르헨티나 (ARG)	24	21	25	28	74	19	0	0	0	0	43	21	25	28	74
3	아르메니아 (ARM)	6	2	6	6	14	7	0	0	0	0	13	2	6	6	14
4	오스트랄라시아 (ANZ) [ANZ]	2	3	4	5	12	0	0	0	0	0	2	3	4	5	12
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

인덱스를 '국가 (IOC 코드)'로 변경하기

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df2 = df[0].set_index('국가 (IOC 코드)')
df2
```

국가 (IOC 코드)	하계 참가 횟수	Unnamed: 2	Unnamed: 3	Unnamed: 4	계	동계 참가 횟수	Unnamed: 7	Unnamed: 8	Unnamed: 9	계.1	전체 참가 횟수	Unnamed: 12	Unnamed: 13	Unnamed: 14	총합
아프가니스탄 (AFG)	14	0	0	2	2	0	0	0	0	0	14	0	0	2	2
알제리 (ALG)	13	5	4	8	17	3	0	0	0	0	16	5	4	8	17
아르헨티나 (ARG)	24	21	25	28	74	19	0	0	0	0	43	21	25	28	74
아르메니아 (ARM)	6	2	6	6	14	7	0	0	0	0	13	2	6	6	14
오스트랄라시아 (ANZ) [ANZ]	2	3	4	5	12	0	0	0	0	0	2	3	4	5	12
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

인덱스가 "국가 (IOC 코드)"로 변경되었습니다.

하계 정보만 추출하기

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df2 = df[0].set_index('국가 (IOC 코드)')

summer = df2.iloc[:, :5]
# iloc은 integer location의 약어로, 데이터 프레임의 행이나 열의 순서를 나타내는 정수로 특정 값을 추출합니다.
summer
```

국가 (IOC 코드)	하계 참가 횟수	Unnamed: 2	Unnamed: 3	Unnamed: 4	계
아프가니스탄 (AFG)	14	0	0	2	2
알제리 (ALG)	13	5	4	8	17
아르헨티나 (ARG)	24	21	25	28	74
아르메니아 (ARM)	6	2	6	6	14
오스트랄라시아 (ANZ) [ANZ]	2	3	4	5	12
...	...	...	...	...	...
독립 (IOA) [IOA]	3	1	0	1	2
독립 참가 (IOP) [IOP]	1	0	1	2	3
러시아 출신 올림픽 선수 (OAR)	0	0	0	0	0
혼성 (ZZX) [ZZX]	3	8	5	4	17
총합	28	5116	5082	5490	15688

153 rows × 5 columns

iloc( ) 함수를 활용하여
하계(Summer) 정보만 추출하였습니다.

컬럼 이름 설정하기

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df2 = df[0].set_index('국가 (IOC 코드)')

summer = df2.iloc[:, :5]
summer.columns = ['경기수', '금', '은', '동', '합계']
summer
```

국가 (IOC 코드)	경기수	금	은	동	합계
아프가니스탄 (AFG)	14	0	0	2	2
알제리 (ALG)	13	5	4	8	17
아르헨티나 (ARG)	24	21	25	28	74
아르메니아 (ARM)	6	2	6	6	14
오스트랄라시아 (ANZ) [ANZ]	2	3	4	5	12
...	...	...	...	...	...
독립 (IOA) [IOA]	3	1	0	1	2
독립 참가 (IOP) [IOP]	1	0	1	2	3
러시아 출신 올림픽 선수 (OAR)	0	0	0	0	0
혼성 (ZZX) [ZZX]	3	8	5	4	17
총합	28	5116	5082	5490	15688

153 rows × 5 columns

데이터 프레임의 columns에
컬럼 이름을 설정하였습니다.

내림차순으로 정렬하기

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df2 = df[0].set_index('국가 (IOC 코드)')

summer = df2.iloc[:, :5]
summer.columns = ['경기수', '금', '은', '동', '합계']
summer = summer.sort_values('금', ascending=False)
summer
```

국가 (IOC 코드)	경기수	금	은	동	합계
총합	28	5116	5082	5490	15688
미국 (USA) [P] [Q] [R] [Z] [F]	27	1022	795	706	2523
소련 (URS) [URS]	9	395	319	296	1010
영국 (GBR) [GBR] [Z]	28	263	295	293	851
중화인민공화국 (CHN) [CHN]	10	224	167	155	546
...	...	...	...	...	...
레바논 (LIB)	17	0	2	2	4
세르비아 몬테네그로 (SCG) [SCG]	1	0	2	0	2
지부티 (DJI) [B]	8	0	0	1	1
키프로스 (CYP)	10	0	1	0	1
아프가니스탄 (AFG)	14	0	0	2	2

153 rows × 5 columns

sort_values( ) 함수를 이용하면
원하는 열을 기준으로
데이터 순서를 정렬할 수 있습니다.

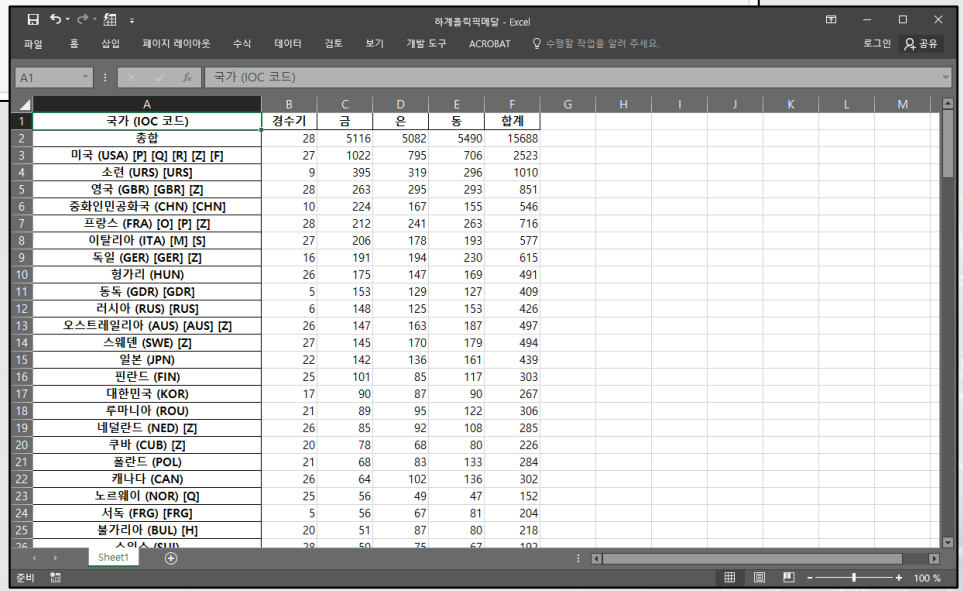
엑셀 파일로 저장하기


```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df2 = df[0].set_index('국가 (IOC 코드)')

summer = df2.iloc[:, :5]
summer.columns = ['경기수', '금', '은', '동', '합계']
summer = summer.sort_values('금', ascending=False)
summer.to_excel('하계올림픽메달.xlsx')
```

to_excel( ) 함수를 이용하여 작업했던 "summer" 데이터 프레임을 엑셀 파일로 저장합니다.



하계올림픽메달 - Excel

	A	B	C	D	E	F
1	국가 (IOC 코드)	경수기	금	은	동	합계
2	총합	28	5116	5082	5490	15688
3	미국 (USA) [P] [Q] [R] [Z] [F]	27	1022	795	706	2523
4	소련 (URS) [URS]	9	395	319	296	1010
5	영국 (GBR) [GBR] [Z]	28	263	295	293	851
6	중화인민공화국 (CHN) [CHN]	10	224	167	155	546
7	프랑스 (FRA) [O] [P] [Z]	28	212	241	263	716
8	이탈리아 (ITA) [M] [S]	27	206	178	193	577
9	독일 (GER) [GER] [Z]	16	191	194	230	615
10	헝가리 (HUN)	26	175	147	169	491
11	동독 (GDR) [GDR]	5	153	129	127	409
12	러시아 (RUS) [RUS]	6	148	125	153	426
13	오스트레일리아 (AUS) [AUS] [Z]	26	147	163	187	497
14	스웨덴 (SWE) [Z]	27	145	170	179	494
15	일본 (JPN)	22	142	136	161	439
16	핀란드 (FIN)	25	101	85	117	303
17	대한민국 (KOR)	17	90	87	90	267
18	루마니아 (ROU)	21	89	95	122	306
19	네덜란드 (NED) [Z]	26	85	92	108	285
20	쿠바 (CUB) [Z]	20	78	68	80	226
21	폴란드 (POL)	21	68	83	133	284
22	캐나다 (CAN)	26	64	102	136	302
23	노르웨이 (NOR) [Q]	25	56	49	47	152
24	서독 (FRG) [FRG]	5	56	67	81	204
25	불가리아 (BUL) [H]	20	51	87	80	218

LESSON 02

pandas의 DataFrame 행 검색

엑셀 파일 읽어 오기 index_col=0

```python
import pandas as pd

df = pd.read_excel('하계올림픽메달.xlsx', index_col=0)
df.head()
```

✓ 0.0s Python

국가 (IOC 코드)	경기수	금	은	동	합계
총합	28	5116	5082	5490	15688
미국 (USA) [P] [Q] [R] [Z] [F]	27	1022	795	706	2523
소련 (URS) [URS]	9	395	319	296	1010
영국 (GBR) [GBR] [Z]	28	263	295	293	851
중화인민공화국 (CHN) [CHN]	10	224	167	155	546

엑셀 파일 읽어 오기

```
import pandas as pd

df = pd.read_excel('하계올림픽메달.xlsx')
df.head()
```

	국가 (IOC 코드)	경기수	금	은	동	합계
0	총합	28	5116	5082	5490	15688
1	미국 (USA) [P] [Q] [R] [Z] [F]	27	1022	795	706	2523
2	소련 (URS) [URS]	9	395	319	296	1010
3	영국 (GBR) [GBR] [Z]	28	263	295	293	851
4	중화인민공화국 (CHN) [CHN]	10	224	167	155	546

'국가 (IOC 코드)' 열에서 '대한민국 (KOR)' 행 검색 출력

```
df['국가 (IOC 코드)'] == '대한민국 (KOR)'
```

✓ 0.0s

```
0      False
1      False
2      False
3      False
4      False
       ...
148    False
149    False
150    False
151    False
152    False
Name: 국가 (IOC 코드), Length: 153, dtype: bool
```

```
df[df['국가 (IOC 코드)'] == '대한민국 (KOR)']
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267

```
df.loc[df['국가 (IOC 코드)'] == '대한민국 (KOR)']
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267

'국가 (IOC 코드)' 열에서 '대한민국 (KOR)' 행 검색 출력

```
df['국가 (IOC 코드)'].isin(['대한민국 (KOR)'])
```

✓ 0.0s

```
0      False
1      False
2      False
3      False
4      False
       ...
148    False
149    False
150    False
151    False
152    False
Name: 국가 (IOC 코드), Length: 153, dtype: bool
```

```
df[df['국가 (IOC 코드)'].isin(['대한민국 (KOR)'])]
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267

```
df.loc[df['국가 (IOC 코드)'].isin(['대한민국 (KOR)'])]
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267

Series.isin([item1, item2, …])
시리즈의 각 요소가 전달된 값 시퀀스의 요소와
정확히 일치하는지 여부를 보여주는 부울 시리즈를 반환

'국가 (IOC 코드)' 열에서 '대한민국' 등 부분 문자열이 있는 행 검색 출력

```
df[df['국가 (IOC 코드)'].str.contains('대한민국')]
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267

```
df[df['국가 (IOC 코드)'].str.contains('KOR')]
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267

```
df[df['국가 (IOC 코드)'].str.contains('러시아')]
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
10	러시아 (RUS) [RUS]	6	148	125	153	426
86	러시아 제국 (RU1) [RU1]	3	1	4	3	8
110	러시아 출신 올림픽 선수 (OAR)	0	0	0	0	0

SUMMARY

학습정리

위키피디다의 '올림픽 메달 집계' 전체 페이지에서 데이터 읽어와 특정 데이터프레임을 준비

하계 올림픽 메달 집계를 위한 데이터프레임을 생성

```
import pandas as pd

df = pd.read_html('https://ko.wikipedia.org/wiki/%EC%98%AC%EB%A6%BC%ED%94%BD_%EB%A9%94%EB%8B%AC_%EC%A7%91%EA%B3%84')
df2 = df[0].set_index('국가 (IOC 코드)')

summer = df2.iloc[:, :5]
summer.columns = ['경기수', '금', '은', '동', '합계']
summer = summer.sort_values('금', ascending=False)
summer
```

'국가 (IOC 코드)' 열에서 '대한민국' 등 부분 문자열이 있는 행 검색

```
df[df['국가 (IOC 코드)'].str.contains('대한민국')]
```

✓ 0.0s

	국가 (IOC 코드)	경기수	금	은	동	합계
15	대한민국 (KOR)	17	90	87	90	267