

Ch20 Web Scraping - \ 추출

웹 스크랩핑(**Web Scraping**)은 웹 페이지의 특정 부분에서 필요한 데이터를 추출하는 기술

주로 정적인 웹 페이지에서 사용되며 \ BeautifulSoup, lxml, requests, Selenium 등의 라이브러리나 프레임워크를 사용

웹 크롤링(**Web Crawling**)은 웹 페이지를 주기적으로 탐색하여 필요한 정보를 수집하는 기술

여러 웹 페이지를 순회하며 링크를 따라가면서 정보를 수집 \ 동적인 웹 페이지에서 사용되며, 스크립트 언어나 API 등을 사용하여 자동화

In []:

웹 스크랩핑(**Web Scraping**) :\

위주로 추출

웹 페이지의 \

에 포함된 데이터들을 데이터 프레임으로 추

출

[방법-Table추출] HTML 스크립트의 특정 <table> 태그에 포함된 데이터 추출하기

(1) 웹 페이지에서 HTML 스크립트 데이터 얻기

import urllib3 라이브러리 사용

PoolManager를 사용하여 연결 관리\ PoolManager 객체로
GET 요청 보내기\ 서버 응답에서 HTML 내용 추출

(2) HTML 문서 파싱

BeautifulSoup() 사용

(3) \

의 text를 데이터 프레임으로 추출
`pandas.read_html()` 사용

(1) 웹 페이지에서 HTML 스크립트 데이터 얻기

URL로 웹 페이지 열어보기

```
In [3]: ## URL로 웹 페이지 열어보기
import webbrowser as wb

url = 'https://www.example.com'
wb.open_new(url) #브라우저에 사이트가 열림
```

Out[3]: True

[B] import urllib.request 라이브러리 사용

```
In [4]: ## [import urllib3] HTML 내용 추출
from urllib.request import urlopen, Request

url = 'https://www.example.com'

# 웹 페이지 요청 및 응답 데이터 읽기
req = Request(url)          #http 연결 요청
html = urlopen(req).read()   #http 연결 정보로 html 문서 요청

print(html)
```

```
b'<!doctype html>\n<html>\n<head>\n    <title>Example Domain\n</title>\n\n    <meta charset\n        = "utf-8" />\n    <meta http-equiv="Content-type" content="text/html; charset=utf-8" />\n\n    <meta name="viewport" content\n        = "width=device-width, initial-\n            scale=1" />\n\n    <style type\n        = "text/css">\n        body {\n            background-color: #f0f0f2;\n            margin: 0;\n            padding: 0;\n            font-family: -apple-system, system-ui, BlinkMacSystemFont, "Segoe UI", "Open Sans", "Helvetica Neue", Helvetica, Arial, sans-serif;\n\n        }\n        div {\n            width: 600px;\n            margin: 5em auto;\n            padding: 2em;\n            background-color: #fdfdff;\n            border-radius: 0.5em;\n            box-shadow: 2px 3px 7px 2px rgba(0,0,0,0.02);\n        }\n        a:link, a:visited {\n            color: #38488f;\n            text-decoration: n
```

```
one; \n      } \n      @media (max-w\n      idth: 700px) { \n          div\n          {\n              margin: 0 aut\n          o; \n              width: auto; \n          } \n      } \n      </style> \n</\nhead>\n<body>\n<div>\n    <h\n1>Example Domain</h1>\n    <p>\nThis domain is for use in illu\nstrative examples in document\ns. You may use this \n    domai\nn in literature without prior\ncoordination or asking for per\nmission.</p>\n    <p><a href\n= \"https://www.iana.org/domain\ns/example\">More information...\n</a></p>\n</div>\n</body>\n</h\nml>\n'
```

[B'] import urllib3 라이브러리 사용

```
In [6]: ## import urllib3 라이브러리 사용 : HTML 내용 추출\nimport urllib3\n## [import urllib3] HTML 내용 추출\nimport urllib3\nurl = 'https://www.example.com'\n\n# PoolManager를 사용하여 page 요청 후 추출\nhttp = urllib3.PoolManager()          # PoolManager를 사용하여 연\n결 관리\nresponse = http.request('GET', url)    # GET 요청 보내기\nhtml = response.data.decode('utf-8')    # 서버 응답에서 HTML 내용\n추출\n\nprint(html)
```

```
<!doctype html>
<html>
  <head>
    <title>Example Domain</title>

    <meta charset="utf-8" />
    <meta http-equiv="Content-type" content="text/html; charset=utf-8" />
    <meta name="viewport" content="width=device-width, initial-scale=1" />
    <style type="text/css">
      body {
        background-color: #f0f0f2;
        margin: 0;
        padding: 0;
        font-family: -apple-system, system-ui, BlinkMacSystemFont, "Segoe UI", "Open Sans", "Helvetica Neue", Helvetica, Arial, sans-serif;
      }
      div {
```

```
width: 600px;
margin: 5em auto;
padding: 2em;
background-color: #fdf
dff;
border-radius: 0.5em;
box-shadow: 2px 3px 7p
x 2px rgba(0,0,0,0.02);
}
a:link, a:visited {
color: #38488f;
text-decoration: none;
}
@media (max-width: 700px)
{
    div {
        margin: 0 auto;
        width: auto;
    }
}
</style>
</head>

<body>
<div>
    <h1>Example Domain</h1>
    <p>This domain is for use
```

in illustrative examples in documents. You may use this domain in literature without prior coordination or asking for permission.</p>

<p>More information...</p>

</div>

</body>

</html>

(2) HTML 문서 파싱

BeautifulSoup() 사용

> HTML 및 XML 문서를 파싱하고 검색, 탐색하는 데 사용

```
from bs4 import BeautifulSoup\nsoup =\nBeautifulSoup(html_content, 'html.parser')\n\n> 원하는 정보 추출\n\nlinks = soup.find_all('a')
```

```
In [11]: ## HTML 파싱 후 원하는 데이터 추출\nfrom bs4 import BeautifulSoup\n\n# BeautifulSoup을 사용하여 HTML 파싱\nsoup = BeautifulSoup(html, 'html.parser')\n\n# 원하는 정보 추출: 모든 링크를 가져오기\nalist = []\nlinks = soup.find_all('a')\nfor x in links:\n    alist.append(x.get('href'))\nalist
```

```
Out[11]: ['https://www.iana.org/domains/example']
```

[연습] \ 에 포함된 데이터 추출

```
In [12]: ##[연습] <Table>가 포함된 HTML 페이지
html = """
<!DOCTYPE html>
<html>
<head>
    <title>Table Example</title>
</head>
<body>

    <h2>Sample Table</h2>
    <table>
        <thead>
            <tr>
                <th>Name</th>
                <th>Age</th>
                <th>City</th>
            </tr>
        </thead>
        <tbody>
            <tr>
                <td>John</td>
                <td>25</td>
                <td>New York</td>
            </tr>
            <tr>
                <td>Alice</td>
                <td>30</td>
                <td>London</td>
            </tr>
        </tbody>
    </table>

</body>
</html>
"""
```

(3) \

의 text를 데이터 프레임으로 추출

pandas.read_html() 사용

read_html() 함수는 HTML 문서에서 표(테이블) 데이터를 추출하여 DataFrame으로 변환하는 함수

> \

수만큼의 데이터 프레임으로 추출됨 원하는 데이터프레임[index]로 원하는 table 을 사용

```
In [13]: # 웹 페이지에서 표 스크래핑
import pandas as pd

# 웹 페이지에서 표 스크래핑
tables = pd.read_html(html) #html 문서에서 테이블 태그(<table>
</table>) 부분의 값을 list 형태(table 개수만큼)로 반환
tables[0] #데이터 프레임
```

C:\Users\ADMIN\AppData\Local\Temp\ipykernel_7216\4201469221.py:5: FutureWarning: Passing 1 literal html to 'read_html' is deprecated and will be removed in a future version. To read from a literal string, wrap it in a 'StringIO' object.

```
tables = pd.read_html(html)
#html 문서에서 테이블 태그(<table> </table>) 부분의 값을 list 형태(table 개수만큼)로 반환
```

Out[13]:

	Name	Age	City
0	John	25	New York
1	Alice	30	London

In []:

부천대학 입학 홈페이지에서 전년도 입시결과 페이지에서
 '학과별 모집단위별 지원자 현황' 데이터 추출
 데이터프레임을 구성하여 그래프로 표현

0. URL로 웹 페이지 열어보기

```
In [14]: ## URL로 웹 페이지 열어보기
import webbrowser as wb

url = 'https://dept.bc.ac.kr/ipsi/susi02/results-application-2023.do'
wb.open_new(url) #브라우저에 사이트가 열림
```

Out[14]: True

1. \의 text를 데이터 프레임으로 추출

```
In [19]: ## [B] 부천대학 입시 경쟁률 페이지에서 표 스크래핑 : import
urllib
import pandas as pd
from urllib.request import urlopen, Request

url = 'https://dept.bc.ac.kr/ipsi/susi02/results-application-2023.do'

# 웹 페이지 요청 및 응답 데이터 읽기
req = Request(url)          #http 연결 요청
html = urlopen(req).read()   #http 연결 정보로 html 문서 요청
tables = pd.read_html(html)   #html 문서에서 테이블 태그(<table>
                             </table>) 부분의 값들을 List 형태(table 개수만큼)로 반환

## 첫 번째 <table>에 대한 데이터 프레임
tables[0] #데이터 프레임
```

Out[19]:

학 과 명	수 업 년 한	주 야 구 분	정원내				... 정원외				북한이탈주민				전문 자			
			일반고	특성화고	특 기 자	농 촌	서해5도	지원 율	모 집	지원 율	모 집	지원 율	모 집	지원 율	모 집		
건축과	3 년	주간	20	81	4.1	2	28	14.0	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1
토목공학과	2 년	주간	20	97	4.9	2	28	14.0	0	...	NaN	1	0.0	0.0	0	NaN	NaN	1

학과명	수업년한	주야구분	정원내						... 정원외											
			일반고			특성화고			특기자	농어촌		서해5도		북한이탈주민			전문자			
			모집	지원	지원율	모집	지원	지원율		모집	지원율	모집	지원	지원율	모집	지원	지원율	모집		
섬유패션비즈니스학과	2년	주간	22	99	4.5	3	31	10.3	0	...	NaN	1	0.0	0.0	0	NaN	NaN	1		
실내건축디자인학과	3년	주간	17	85	5.0	2	27	13.5	1	...	NaN	0	NaN	NaN	0	NaN	NaN	1		
전자공학과	4년	주간	29	126	4.3	9	65	7.2	0	...	3.0	0	NaN	NaN	0	NaN	NaN	2		
정보통신과	2년	주간	25	128	5.1	3	26	8.7	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1		
IT융합비즈니스학과	6년	주간	20	85	4.3	5	29	5.8	4	...	NaN	0	NaN	NaN	0	NaN	NaN	2		
영상&게임콘텐	7년	주간	14	181	12.9	3	62	20.7	3	...	NaN	0	NaN	0.0	1	0.0	0.0	1		

학과명	수업년한	주야구분	정원내				... 정원외																	
											... 특기자		... 농어촌		서해5도		북한이탈주민				전문자			
			일반고		특성화고		... 특기자		농어촌		서해5도		북한이탈주민		전문자		... 지원율		모집지원율		모집지원율		모집지원율	
학과명	수업년한	주야구분	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	지원율	모집	
츠과																								
전기과	3년	주간	14	84	6.0	4	41	10.3	0	...	NaN	0	NaN	NaN	0	NaN	NaN	2						
컴퓨터소프트웨어과	3년	주간	14	166	11.9	4	54	13.5	0	...	NaN	0	NaN	NaN	0	NaN	NaN	2						
컴퓨터정보보안학과	3년	주간	5	50	10.0	3	20	6.7	0	...	NaN	1	0.0	0.0	1	0.0	0.0	2						
경영학과	2년	주간	26	153	5.9	6	58	9.7	1	...	NaN	0	NaN	NaN	0	NaN	NaN	1						
비서사무행정학과	2년	주간	11	68	6.2	6	29	4.8	1	...	NaN	0	NaN	NaN	0	NaN	NaN	2						
호텔관광경영학과	2년	주간	27	215	8.0	13	93	7.2	0	...	NaN	0	NaN	NaN	0	NaN	NaN	2						
세무	2년	주간	26	121	4.7	9	50	5.6	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1						

학과명	수업년한	주야구분	정원내						... 정원외						전문자		
			일반고			특성화고			특기자	... 농어촌	서해5도			북한이탈주민			
			모집	지원	지원율	모집	지원	지원율			모집	지원	지원율	모집	지원	지원율	
회계학과																	
15 항공서비스과	2년	주간	30	587	19.6	3	75	25.0	2	...	10.5	1	0.0	0.0	0	NaN	NaN 1
16 식품영양학과	3년	주간	14	105	7.5	3	24	8.0	1	...	NaN	1	0.0	0.0	1	0.0	0.0 1
17 호텔외식조리학과	2년	주간	20	157	7.9	6	41	6.8	2	...	1.0	1	0.0	0.0	0	NaN	NaN 4
18 뷰티케어과해어디자인전공	2년	주간	7	123	17.6	3	38	12.7	3	...	NaN	0	NaN	NaN	1	0.0	0.0 2
19 뷰티케어과뷰티디자인	2년	주간	7	146	20.9	3	45	15.0	3	...	NaN	1	1.0	1.0	1	0.0	0.0 2

학과명	수업년한	주야구분	정원내						... 정원외						전문자		
			일반고			특성화고			특기자	... 농어촌	서해5도		북한이탈주민				
			모집	지원	지원율	모집	지원	지원율			모집	지원율	모집	지원	지원율		
전공			모집	지원	지원율	모집	지원	지원율	모집	지원율	모집	지원율	모집	지원	지원율	모집	
보건의료행정학과	20	3년 주간	16	186	11.6	4	34	8.5	0	...	2.0	1	0.0	0.0	0	NaN	NaN 1
디지털미디어디자인학과	21	3년 주간	11	65	5.9	6	44	7.3	2	...	NaN 0	NaN	NaN 0	NaN	NaN 0	NaN	NaN 0
재활스포츠과	22	2년 주간	17	165	9.7	6	50	8.3	5	...	2.0	1	0.0	0.0	1	0.0	0.0 1
유아교육과	23	3년 주간	16	97	6.1	2	14	7.0	0	...	1.0 0	NaN	NaN 0	NaN	NaN 1	NaN	NaN 1
아동보육과	24	2년 주간	11	96	8.7	5	28	5.6	0	...	NaN 0	NaN	NaN 0	NaN	NaN 1	NaN	NaN 1
사회복지학과	25	2년 주간	15	84	5.6	5	21	4.2	0	...	NaN 0	NaN	NaN 0	NaN	NaN 1	NaN	NaN 1
간호학과	26	4년 주간	35	472	13.5	1	38	38.0	0	...	4.8 0	NaN	NaN 0	NaN	NaN 12	NaN	NaN 12

학과명	수업년한	주야구분	정원내						... 정원외								
			일반고	특성화고			특기자	농어촌	서해5도	북한이탈주민			전문자				
학과명	수업년한	주야구분	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집		
학과																	
27	치기공과	3년 주간	14	150	10.7	2	21	10.5	0	...	NaN	0	NaN	NaN	0	NaN	NaN 1
28	반려동물과	2년 주간	17	145	8.5	3	40	13.3	0	...	NaN	0	NaN	NaN	0	NaN	NaN 1

29 rows × 27 columns

```
In [20]: ## [B'] 부천대학 입시 경쟁률 페이지에서 표 스크래핑 : import urllib3
import pandas as pd
import urllib3 #HTTP 클라이언트 구현 모듈

url = 'https://dept.bc.ac.kr/ipsi/susi02/results-application-2023.do'

# 웹 페이지에서 표 스크래핑
http = urllib3.PoolManager() #http나 https 연결 관리자 호출
req = http.request('GET', url) #http 연결 관리자로 html 문서 요청
tables = pd.read_html(req.data) #html 문서에서 테이블 태그(<table> </table>) 부분의

## 첫 번째 <table>에 대한 데이터 프레임
tables[0] #데이터 프레임
```

Out[20]:

학과명	수업년한	주야구분	정원내 ...																			
			학과명		수업년한		주야구분		일반고		특성화고		특기자		... 농어촌		서해5도		북한이탈주민		전문대(
			모집	지원	지원율	모집	지원	지원율	모집	...	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	
건축과	3년	주간	20	81	4.1	2	28	14.0	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1	0			
토목공학과	2년	주간	20	97	4.9	2	28	14.0	0	...	NaN	1	0.0	0.0	0	NaN	NaN	1	0			
섬유패션비즈니스학과	2년	주간	22	99	4.5	3	31	10.3	0	...	NaN	1	0.0	0.0	0	NaN	NaN	1	0			
실내건축디자인학과	3년	주간	17	85	5.0	2	27	13.5	1	...	NaN	0	NaN	NaN	0	NaN	NaN	1	1			
전자공학과	2년	주간	29	126	4.3	9	65	7.2	0	...	3.0	0	NaN	NaN	0	NaN	NaN	2	1			
정보통신과	2년	주간	25	128	5.1	3	26	8.7	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1	0			
IT융합비즈니스	2년	주간	20	85	4.3	5	29	5.8	4	...	NaN	0	NaN	NaN	0	NaN	NaN	2	0			

학과명	수업년한	주야구분	정원내 ...												전문대(
			일반고			특성화고			특기자 ...			농어촌			서해5도		
			모집	지원	지원율	모집	지원	지원율	모집	...	지원율	모집	지원	지원율	모집	지원	지원율
7 영상&게임콘텐츠과	2년	주간	14	181	12.9	3	62	20.7	3	...	NaN	0	NaN	0.0	1	0.0	0.0
8 전기과	3년	주간	14	84	6.0	4	41	10.3	0	...	NaN	0	NaN	NaN	0	NaN	NaN
9 컴퓨터소프트웨어과	3년	주간	14	166	11.9	4	54	13.5	0	...	NaN	0	NaN	NaN	0	NaN	NaN
10 컴퓨터정보보호안학과	3년	주간	5	50	10.0	3	20	6.7	0	...	NaN	1	0.0	0.0	1	0.0	0.0
11 경영학과	2년	주간	26	153	5.9	6	58	9.7	1	...	NaN	0	NaN	NaN	0	NaN	NaN
12 비서사무행정학과	2년	주간	11	68	6.2	6	29	4.8	1	...	NaN	0	NaN	NaN	0	NaN	NaN

학과명	수업년한	주야구분	정원내 ...																							
			학과명	수업년한	주야구분	일반고			특성화고			특기자			... 농어촌			서해5도			북한이탈주민			전문대(
						모집	지원	지원율	모집	지원	지원율	모집	...	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	
호텔관광경영학과	2년	주간	13	27	215	8.0	13	93	7.2	0	...	NaN	0	NaN	NaN	0	NaN	NaN	0	NaN	NaN	2	0			
세무회계학과	2년	주간	14	26	121	4.7	9	50	5.6	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1	0						
항공서비스과	2년	주간	15	30	587	19.6	3	75	25.0	2	...	10.5	1	0.0	0.0	0	NaN	NaN	1	1						
식품영양학과	3년	주간	16	14	105	7.5	3	24	8.0	1	...	NaN	1	0.0	0.0	1	0.0	0.0	1	0.0	0.0	1	0			
호텔외식조리학과	2년	주간	17	20	157	7.9	6	41	6.8	2	...	1.0	1	0.0	0.0	0	NaN	NaN	4	2						
뷰티케어과																										
해어디자인전공	2년	주간	18	7	123	17.6	3	38	12.7	3	...	NaN	0	NaN	NaN	1	0.0	0.0	2	0						

학과명	수업년한	주야구분	정원내 ...																							
			학과명	수업년한	주야구분	일반고			특성화고			특기자			... 농어촌			서해5도			북한이탈주민			전문대(
						모집	지원	지원율	모집	지원	지원율	모집	...	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	
뷰티케어과	2년	주간	뷰티디자인전공	2년	주간	7	146	20.9	3	45	15.0	3	...	NaN	1	1.0	1.0	1	0.0	0.0	2	2				
보건의료행정학과	3년	주간	디지털미디어디자인학과	3년	주간	16	186	11.6	4	34	8.5	0	...	2.0	1	0.0	0.0	0	NaN	NaN	1	2				
재활스포츠과	2년	주간	유아교육과	3년	주간	11	65	5.9	6	44	7.3	2	...	NaN	0	NaN	NaN	0	NaN	NaN	0	NaN	0	NaN		
아동보	2년	주간	11	96	8.7	5	28	5.6	0	...	NaN	0	NaN	NaN	0	NaN	NaN	0	NaN	NaN	1	0				

학과명	수업년한	주야구분	정원내 ...																							
			학과명	수업년한	주야구분	일반고			특성화고			특기자			... 농어촌			서해5도			북한이탈주민			전문대(
						모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	지원율	모집	지원	
육과																										
25	사회복지학과	2년 주간	25	주간	15	84	5.6	5	21	4.2	0	...	NaN	0	NaN	NaN	0	NaN	NaN	0	NaN	NaN	1	0		
26	간호학과	4년 주간	26	주간	35	472	13.5	1	38	38.0	0	...	4.8	0	NaN	NaN	0	NaN	NaN	12	204					
27	치기공과	3년 주간	27	주간	14	150	10.7	2	21	10.5	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1	4					
28	반려동물과	2년 주간	28	주간	17	145	8.5	3	40	13.3	0	...	NaN	0	NaN	NaN	0	NaN	NaN	1	3					

29 rows x 27 columns

2. 컬럼명, 데이터 확인

- > 컬럼명 구성을 확인
- > (정원내, 일반고, 지원율) 데이터를 확인

```
In [21]: ##### **2. 컬럼명, 데이터 확인**
##### > 컬럼명 구성을 확인
df = tables[0]
for x in df:
    print(x)
```

```
('학과명', '학과명', '학과명')
('수업 년한', '수업 년한', '수업 년한')
('주야 구분', '주야 구분', '주야 구분')
('정원내', '일반고', '모집')
('정원내', '일반고', '지원')
('정원내', '일반고', '지원율')
('정원내', '특성화고', '모집')
('정원내', '특성화고', '지원')
('정원내', '특성화고', '지원율')
('정원내', '특기자', '모집')
('정원내', '특기자', '지원')
('정원내', '특기자', '지원율')
('정원외', '수급자및차상위', '모집')
('정원외', '수급자및차상위', '지원')
('정원외', '수급자및차상위', '지원율')
('정원외', '농어촌', '모집')
('정원외', '농어촌', '지원')
('정원외', '농어촌', '지원율')
('정원외', '서해5도', '모집')
('정원외', '서해5도', '지원')
('정원외', '서해5도', '지원율')
('정원외', '북한이탈주민', '모집')
('정원외', '북한이탈주민', '지원')
('정원외', '북한이탈주민', '지원율')
('정원외', '전문대이상졸업자', '모집')
('정원외', '전문대이상졸업자', '지원')
('정원외', '전문대이상졸업자', '지원율')
```

```
In [22]: ##### **2. 컬럼명, 데이터 확인**
##### > (정원내, 일반고, 지원율) 데이터를 확인
df[['정원내', '일반고', '지원율']]
```

```
Out[22]: 0      4.1
1      4.9
2      4.5
3      5.0
4      4.3
5      5.1
6      4.3
7     12.9
8      6.0
9     11.9
10     10.0
11     5.9
12     6.2
13     8.0
14     4.7
15    19.6
16     7.5
17     7.9
18    17.6
19    20.9
20    11.6
21     5.9
22     9.7
23     6.1
24     8.7
25     5.6
26    13.5
27    10.7
28     8.5
Name: (정원내, 일반고, 지원율), dtype: float64
```

3. 학과별 일반고 지원율에 대한 컬럼만 데이터 프레임으로 구성하기

> '학과명', '일반고_지원율' 컬럼만으로 새로운 데이터 프레임 구성

>> '학과명' : ('학과명', '학과명', '학과명')

>> '정원내_일반고_지원율' : ('정원내', '일반고', '지원율')

In [23]: ## 원하는 컬럼만 추출 데이터프레임 재구성

```
import pandas as pd

df_rate = pd.DataFrame() #학과별 일반고 경쟁률 데이터프레임 생성
df_rate['학과명'] = df[['학과명', '학과명', '학과명']]
df_rate['정원내_일반고_지원율'] = df[['정원내', '일반고', '지원율']]
df_rate
```

Out[23]:

학과명 정원내_일반고_지원율		
0	건축과	4.1
1	토목공학과	4.9
2	섬유패션비즈니스학과	4.5
3	실내건축디자인학과	5.0
4	전자공학과	4.3
5	정보통신과	5.1
6	IT융합비즈니스학과	4.3
7	영상&게임콘텐츠과	12.9
8	전기과	6.0
9	컴퓨터소프트웨어과	11.9
10	컴퓨터정보보안학과	10.0
11	경영학과	5.9
12	비서사무행정학과	6.2
13	호텔관광경영학과	8.0
14	세무회계학과	4.7
15	항공서비스과	19.6
16	식품영양학과	7.5
17	호텔외식조리학과	7.9
18	뷰티케어과 헤어디자인전공	17.6
19	뷰티케어과 뷰티디자인전공	20.9
20	보건의료행정학과	11.6
21	디지털미디어디자인학과	5.9
22	재활스포츠과	9.7
23	유아교육과	6.1
24	아동보육과	8.7
25	사회복지학과	5.6
26	간호학과	13.5
27	치기공과	10.7
28	반려동물과	8.5

4. 학과별 일반고 지원율 데이터 프레임으로 경쟁률 그리기

> 학과별 일반고-지원율 비교 막대 그래프 그리기

> 경쟁률 내림차 순으로 그리기

In [24]:

```
## 데이터프레임으로 막대 그래프 그리기
import seaborn as sns
import matplotlib.pyplot as plt
```

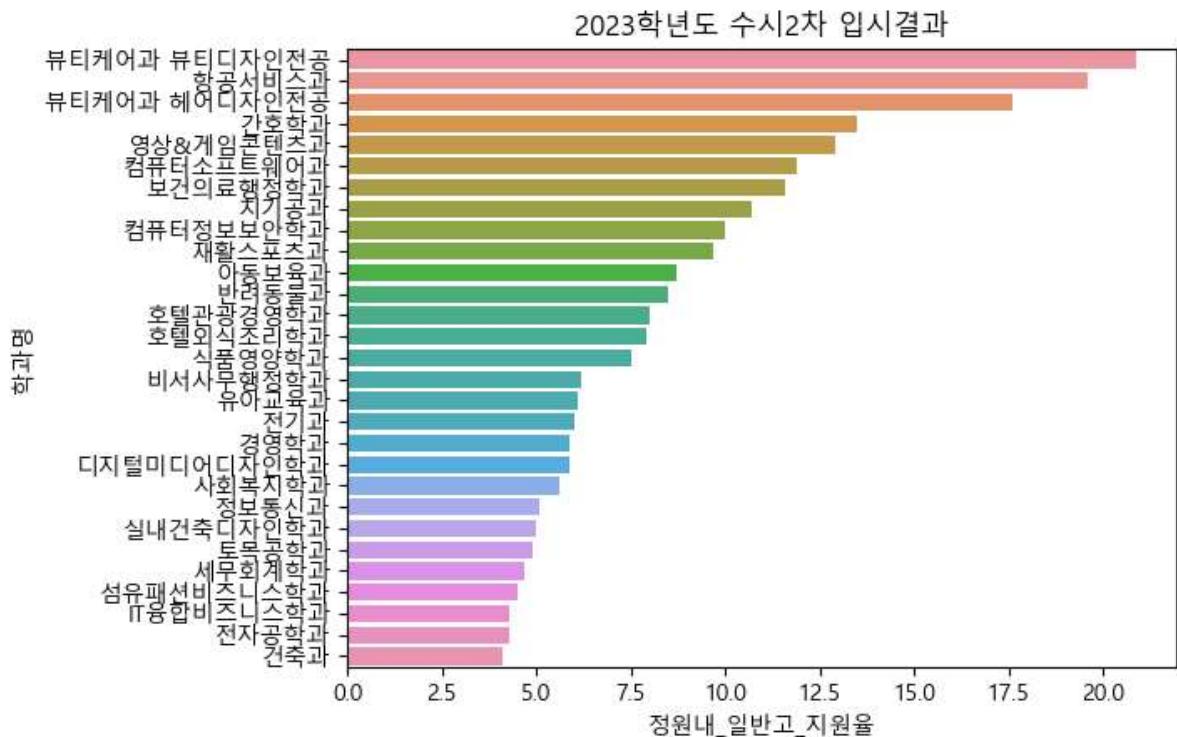
```

#경쟁률 순 정렬
sdf = df_rate.sort_values(by = '정원내_일반고_지원율', ascending=False)

#막대 그래프 그리기
plt.rc('font', family='Malgun Gothic') #폰트 사용
plt.title('2023학년도 수시2차 입시결과')
sns.barplot(data=sdf, y = '학과명', x = '정원내_일반고_지원율')

```

C:\Users\ADMIN\anaconda3\lib\site-packages\scipy__init__.py:155: UserWarning: A NumPy version >=1.18.5 and <1.25.0 is required for this version of SciPy (detected version 1.25.2)
... warnings.warn(f"A NumPy version >={np_minversion} and <{np_maxversion}"
C:\Users\ADMIN\anaconda3\lib\site-packages\seaborn__oldcore.py:1498: FutureWarning:
is_categorical_dtype is deprecated and will be removed in a future version. Use isin
stance(dtype, CategoricalDtype) instead
... if pd.api.types.is_categorical_dtype(vector):
C:\Users\ADMIN\anaconda3\lib\site-packages\seaborn__oldcore.py:1498: FutureWarning:
is_categorical_dtype is deprecated and will be removed in a future version. Use isin
stance(dtype, CategoricalDtype) instead
... if pd.api.types.is_categorical_dtype(vector):
C:\Users\ADMIN\anaconda3\lib\site-packages\seaborn__oldcore.py:1498: FutureWarning:
is_categorical_dtype is deprecated and will be removed in a future version. Use isin
stance(dtype, CategoricalDtype) instead
... if pd.api.types.is_categorical_dtype(vector):
<Axes: title={'center': '2023학년도 수시2차 입시결과'}, xlabel='정원내_일반고_지원
율', ylabel='학과명'>



In []:

5. [도전] 데이터프레임 컬럼명 재구성

> 데이터프레임의 Tuple로 구성된 난해한 컬럼명을 단순하게 재구성

공백을 포함하는 단어는 공백 제거, 예: '수업_년한' > '수업년한'\ 반복되는 단어는 하나로 압축, 예: ('학과명', '학과명', '학과명') > 학과명

> 원하는 컬럼만 추출 재구성

```
{'정원내_일반고_모집' : '일반고_모집', '정원내_특성화고_모집' : '특성화고_모집'},\ '정원내_특기자_모집' : '특기자_모집', '정원내_일반고_지원' : '일반고_지원'},\ '정원내_특성화고_지원' : '특성화고_지원', '정원내_특기자_지원' : '특기자_지원'}
```

문자열에서 공백제거하기

```
In [25]: ##[부천대학 입시] 원하는 컬럼만 추출 데이터프레임 재구성
for x in df: #데이터프레임 컬럼들을 Unpacking
    print(x, end=' > ')
    x1, x2, x3 = x #컬럼명 튜플 Unpacking
    x1 = ''.join(x1.split()) #공백제거
    x2 = ''.join(x2.split())
    x3 = ''.join(x3.split())
    print(x1, x2, x3)
```

```
('학과명', '학과명', '학과명') > 학과명 학과명 학과명
('수업_년한', '수업_년한', '수업_년한') > 수업년한 수업년한 수업년한
('주야_구분', '주야_구분', '주야_구분') > 주야구분 주야구분 주야구분
('정원내', '일반고', '모집') > 정원내 일반고 모집
('정원내', '일반고', '지원') > 정원내 일반고 지원
('정원내', '일반고', '지원율') > 정원내 일반고 지원율
('정원내', '특성화고', '모집') > 정원내 특성화고 모집
('정원내', '특성화고', '지원') > 정원내 특성화고 지원
('정원내', '특성화고', '지원율') > 정원내 특성화고 지원율
('정원내', '특기자', '모집') > 정원내 특기자 모집
('정원내', '특기자', '지원') > 정원내 특기자 지원
('정원내', '특기자', '지원율') > 정원내 특기자 지원율
('정원외', '수급자및차상위', '모집') > 정원외 수급자및차상위 모집
('정원외', '수급자및차상위', '지원') > 정원외 수급자및차상위 지원
('정원외', '수급자및차상위', '지원율') > 정원외 수급자및차상위 지원율
('정원외', '농어촌', '모집') > 정원외 농어촌 모집
('정원외', '농어촌', '지원') > 정원외 농어촌 지원
('정원외', '농어촌', '지원율') > 정원외 농어촌 지원율
('정원외', '서해5도', '모집') > 정원외 서해5도 모집
('정원외', '서해5도', '지원') > 정원외 서해5도 지원
('정원외', '서해5도', '지원율') > 정원외 서해5도 지원율
('정원외', '북한이탈주민', '모집') > 정원외 북한이탈주민 모집
('정원외', '북한이탈주민', '지원') > 정원외 북한이탈주민 지원
('정원외', '북한이탈주민', '지원율') > 정원외 북한이탈주민 지원율
('정원외', '전문대이상졸업자', '모집') > 정원외 전문대이상졸업자 모집
('정원외', '전문대이상졸업자', '지원') > 정원외 전문대이상졸업자 지원
('정원외', '전문대이상졸업자', '지원율') > 정원외 전문대이상졸업자 지원율
```

> 원하는 컬럼만 추출 재구성

```
In [ ]:
```

6. 학과별 일반고 지원율 데이터 프레임으로 경쟁률 그래프 그리기

- > 학과별 일반고-지원율 비교 막대 그래프 그리기
- > 경쟁률 내림차 순으로 그리기

```
In [ ]:
```