

# 부동산 데이터➔ : 집값 차이 수집

DROP THE BIT



김동규 · 박성훈 · 정선우 · 윤상하

**DROP**

**THE BIT**



윤상하



김동규



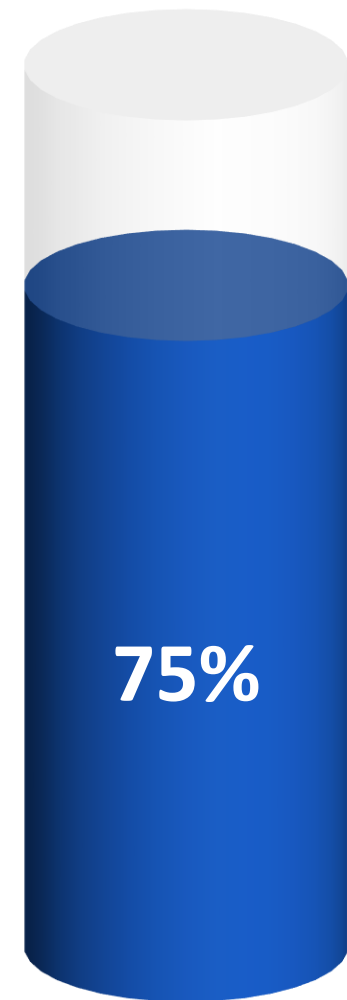
박성훈



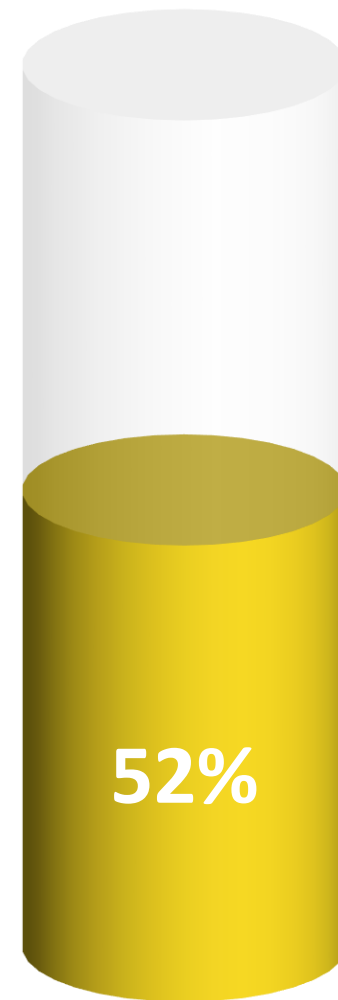
정선우

# 1등 당첨되면 '내 집 마련 꿈꿔'

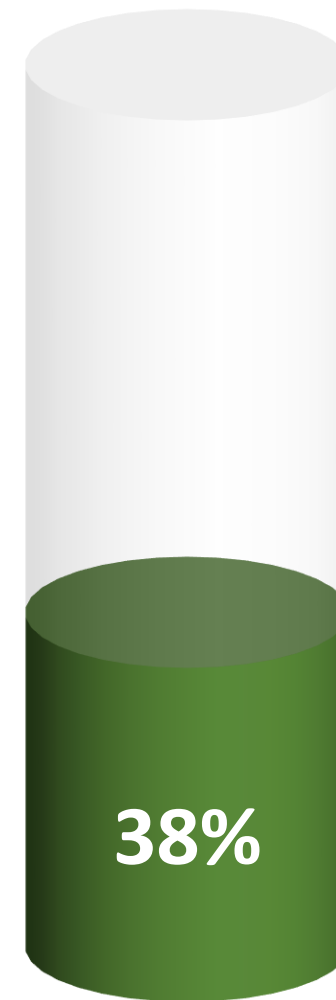
20대 1,094명 대상 설문 조사 결과 ( 복수응답 ) / 자료제공 : 알바천국



내 집 마련



저축



여행



대출 상환



주식/코인 및 재테크





• DROP THE BIT

# • 부동산 데이터를 선택한 이유



## 부동산 가격 변동 추이에 대한 궁금증

부동산은 많은 사람들의 주요한 자산의 역할을 하기 때문에 중장기적인 자산 관리를 위해 부동산 가격 변동 추이를 알아보고자 합니다.



## 장기적으로 이용 가능한 데이터

장기적인 데이터 분석 수집을 통해 부동산 전망에 대해 예측하고 다양한 산업, 경제지표에 연결 시킬 수 있습니다.



# 아실

아파트별 실거래가 수집

서울시 기준 25개구 467동 3,846개

아파트 별 실거래가 수집

아파트 실거래가



분양

아파트/오피스텔

평형 50세대

아파트명을 검색해보세요.

시도 선택 > 시군구 선택 > 읍면동 선택

서울시	경기도	부산시
대구시	인천시	광주시
대전시	울산시	세종시
강원도	충청북도	충청남도
전라북도	전라남도	경상북도
경상남도	제주도	

부동산 스터디

new  
최근하락  
최고가  
최고상승  
가격변동  
가격비교

장미유

명칭  
세대수

공급

인구

가격

심리

거래

비교

정책

학군

랭킹

# HTML IFRAME 태그

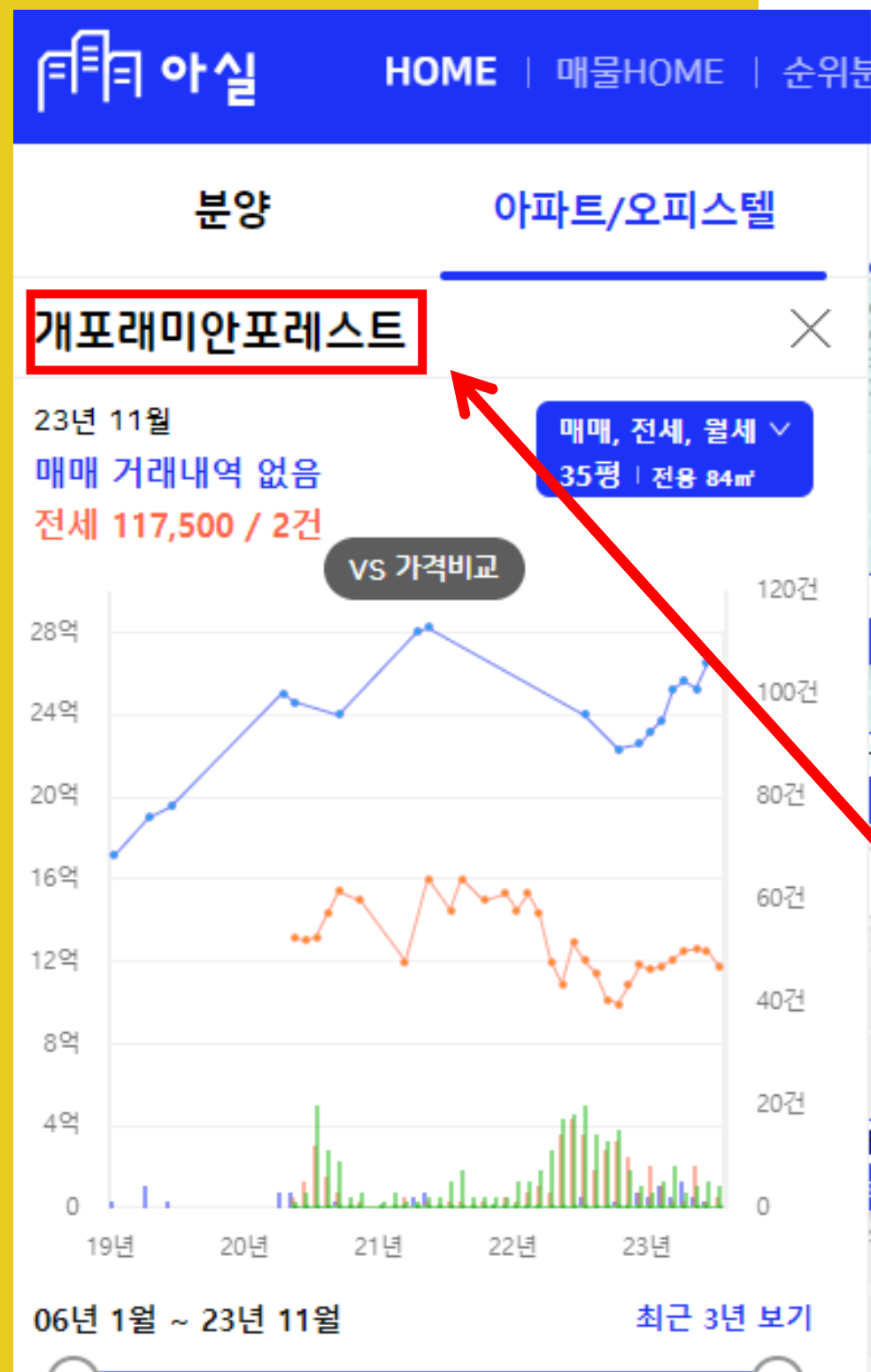
```
<div class="apt_type">...</div>
<iframe id="sub1" src="/app/area_list.jsp?os=pc" class="ranking" style="f
#document
<!DOCTYPE html>
<html lang="kr">
  <head>...</head>
  <body style="overflow:hidden;">
    <div id="wrap" style="padding-top:76px; -webkit-overflow-scrolling
      <form name="frm" method="post" style="height:100%;"> == $0
        <input type="hidden" name="area" value>
        <input type="hidden" name="dong" value>
```

Iframe 코드

## ➔ Iframe 이란?

Inline Frame의 약자로, 웹 페이지 안에 또 다른 웹 페이지를 삽입하는 것을 의미합니다. 현재 브라우저에서 렌더링 되고 있는 문서 안에 또 다른 HTML 페이지를 삽입할 수 있도록 하는 기능입니다. 현재는 보안이슈와, 모바일 호환성, 대체 명령어 등장 등의 이유로 사용이 감소하고 있는 추세입니다.

# Iframe 내의 데이터 추출하기



```
Elements Console Sources Network Performance Memory Application Security Lighthouse Recorder
<!DOCTYPE html>
<html lang="kr">
  <head>
  </head>
  <body>
    <div id="cover"></div>
    <div id="wrap">
      <!-- header -->
      <div id="header">
      </div>
      <!--// header -->
      <!-- 회사소개 : 가맹점가입안내 : 앱다운로드 -->
      <div class="util_cont">
      </div>
      <!--// 회사소개 : 가맹점가입안내 : 앱다운로드 -->
      <div class="app_pc_login_mask" style="display:none;">
      </div>
      <div id="div_bigdata" style="display:none; position:absolute; top:64px; left:0; right:0; z-index:3102; width:100%; height:100%;>
      <!-- iframe_home 시작 -->
      <div class="filter_area">
      </div>
      <div id="container">
        <h2 class="hide">본문 영역</h2>
        <!-- filter (시작) -->
        <div id="filter" class="filter_wrap">
        </div>
        <div id="contents">
          <div class="ctn_wrap open">
            <div class="apt_type">
            </div>
            <iframe id="sub1" src="/app/apt_info.jsp?os=pc&move=no&building=apt&apt=1500060160&min_py=0&max_py=70" class="r">
              #document
              <!DOCTYPE html>
              <html lang="kr">
                <head>
                </head>
                <body style="padding-top:45px;">
                  <div style="position:fixed; top:0; left:0; z-index:1001; width:100%; height:45px; padding:10px 0 0 16px;">
                    <h2 class="h2" style="font-size:20px;">개포래미안포레스트</h2> == $0
                    <span class="apt_info_close" onclick="self.location.href='/app/apt_list.jsp?os=pc&dong=1168010300'">X</span>
                  </div>
                  <div class="asilScroll">
                  </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </body>
</html>
```

```

<!DOCTYPE html>
<html lang="kr">
  <head>...</head>
  <body style="padding-top:45px;">
    <div style="position:fixed; top:0; left:0; z-index:1001; width:100%; height:45px; padding:10px 0 0 16px;">
      <h2 class="h2" style="font-size:20px;">개포래미안포레스트</h2> == $0
      <span class="apt_info_close" onclick="self.location.href='/app/apt_list.jsp?os=pc&dong=1168010300'">X</span>
    </div>
  </body>
</html>

```

1

해당 소스의 XPATH인 /html/body/div[2]/div[1]/div[1]/div[3]/a[1] 탐색 시 브라우저 탐색 실패 오류가 발생합니다.

```

<div class="apt_type">...</div>
<iframe id="sub1" src="/app/apt_info.jsp?os=pc&move=no&building=apt&apt=1500060160&min_py=0&max_py=70" class="ranking" style="width:100%; height:45px; border-bottom:1px solid #ccc;">
  #document
  <!DOCTYPE html>
  <html lang="kr">
    <head>...</head>
    <body style="padding-top:45px;">
      <div style="position:fixed; top:0; left:0; z-index:1001; width:100%; height:45px; padding:10px 0 0 16px; border-bottom:1px solid #ccc;">
        <h2 class="h2" style="font-size:20px;">개포래미안포레스트</h2> == $0
        <span class="apt_info_close" onclick="self.location.href='/app/apt_list.jsp?os=pc&dong=1168010300'">X</span>
      </div>
    </body>
  </html>

```

2

해당 소스는 id가 'sub1'인 iframe에 들어있습니다. 해당 html에 내부에 또 다른 html이 생성된 경우로 해당 iframe 내부로 진입해야합니다.



```
div class="apt_type">...</div>
frame id="sub1" src="/app/area_list.jsp?os=pc" class="ranking" st
#document
<!DOCTYPE html>
<html lang="kr">
  <head>...</head>
  <body style="overflow:hidden;">
    <div id="wrap" style="padding-top:76px; -webkit-overflow-scr
      <form name="frm" method="post" style="height:100%;" == $0
        <input type="hidden" name="area" value>
        <input type="hidden" name="dong" value>
```

```
iframe1 = browser.find_element(By.ID, value: "sub1")
browser.switch_to.frame(iframe1)
time.sleep(2)
apt_name = browser.find_element(By.XPATH, value: '/html/body/div[1]/h2').text
browser.switch_to.default_content()
```

3

진입할 iframe을 변수로 지정해주고 Browser.switch\_to.frame()에 적용해서 해당 iframe으로 진입해서 기존의 방법으로 데이터를 추출합니다. 추출이 끝나면 Browser.switch\_to.default\_content()를 통해 상위의 html로 돌아갑니다. 해당 작업을 해야 다른 iframe에 있는 데이터를 추출 가능합니다.

# 테이블 데이터

분양

아파트/오피스텔

개포래미안포레스트

23.11	21	6	월세 8억 /160	84	110동 11층
	15	7	갱신 전세 10억	84	110동 11층
	14	6	전세 13억 5,000	84	114동 33층
	11	2	월세 1억 /450	84	114동 12층
	10	7	월세 5,000/480	84	110동 5층
	07		월세 1억 /500	84	114동 10층
23.10	27		월세 7억 /180	84	114동 30층
	21	3	전세 12억 5,000	84	115동 4층
	21	7	매매 26억 5,000	84	109동 13층

거래현황 더보기

거래현황

✓ 매매

✓ 전세

✓ 월세

전체년도

22평  
118세대

26평  
662세대

30평  
226세대

35평  
679세대

38평  
99세대

평풀기

최고가 매매 18.5억

전세 10억

?

등 선택

그래프

계약	일	경과	체결가격	타입	거래 등층
	전세 7억 (22평)				
23.09	21		매매 17억	49	128동 12층
23.06	30		매매 16억 1,000	49	111동 14층
	19		매매 15억	49	128동 10층
	13		매매 16억	49	111동 8층
	08		매매 15억 4,000	49	102동 7층
23.02	22		매매 14억 5,000	49	102동 14층
21.10	13		최고가	49	125동 5층

테이블  
row 개수  
추출

## 추출예시-1

```
for t in table_element:
    table_list.append(t.text)
print(len(table_list))
for i in range(2, len(table_list)) :
    mm = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[1]/span').text
    dd = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[2]/span').text
    price = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[4]/span[2]').text
    type = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[5]/span').text
```

1 테이블의 row 갯수 만큼 for 문을 반복해준다. 이때 테이블 안에 들어있는 값을 각각 추출합니다.

23.09, 21, 17억, 49

Traceback (most recent call last):

File "C:\git\python\python\_partice01\01\_print\_hello.py", line 152, in <module>

```
mm = browser.find_element(By.XPATH,f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[1]/span').text
```

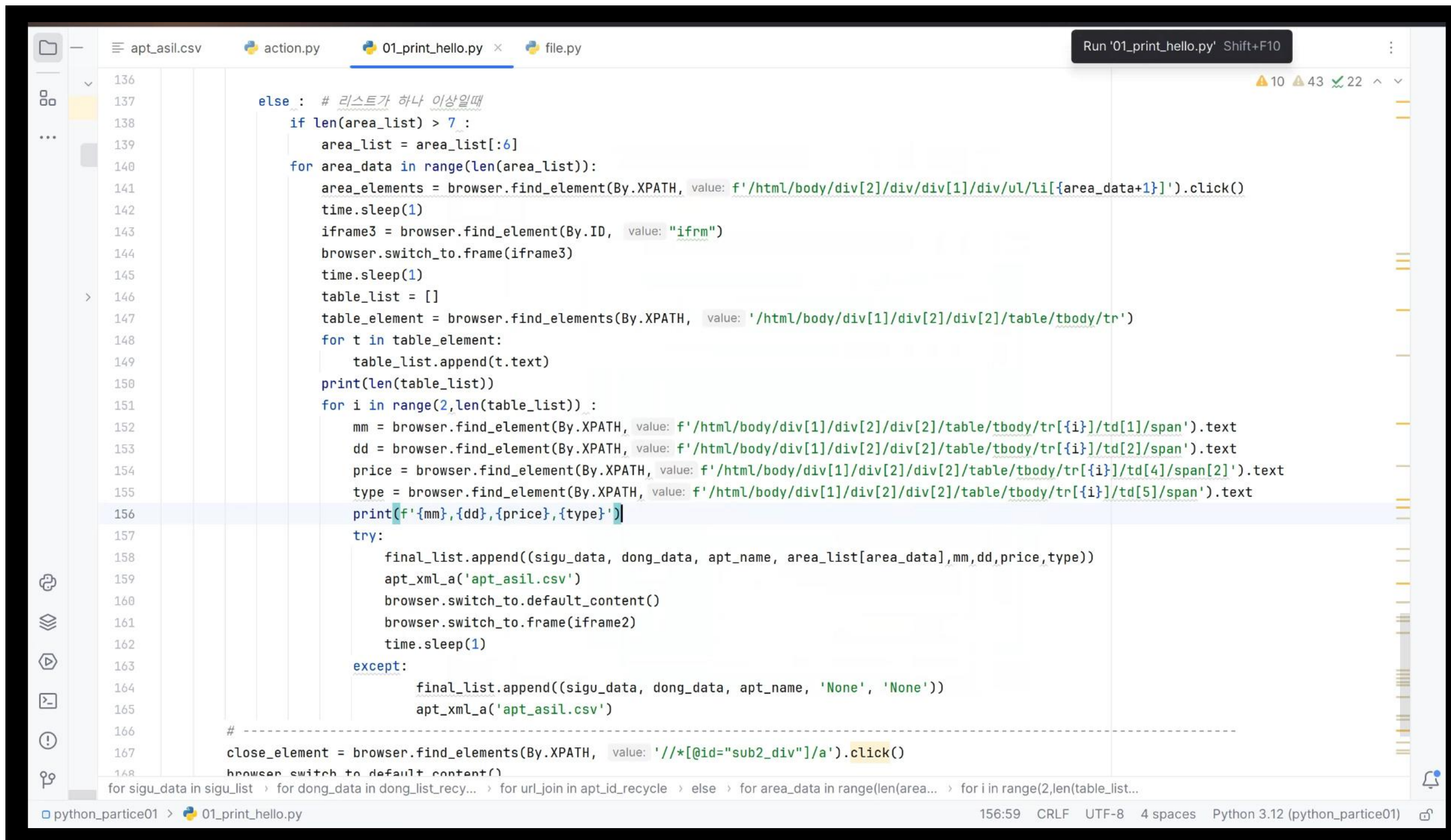
[illegible]

File "C:\git\python\python\_partice01\venv\Lib\site-packages\selenium\webdriver\remote\webdriver.py", line 741, in find\_element

2 하지만 해당 코드는 `selenium`이 데이터를 1 회만 추출하고  
이후의 `for` 문은 경로를 찾지 못하고 오류가 발생합니다.



# 추출예시 -1



```
136
137
138     else : # 리스트가 하나 이상일때
139         if len(area_list) > 7 :
140             area_list = area_list[:6]
141         for area_data in range(len(area_list)):
142             area_elements = browser.find_element(By.XPATH, value: f'/html/body/div[2]/div/div[1]/div/ul/li[{area_data+1}]').click()
143             time.sleep(1)
144             iframe3 = browser.find_element(By.ID, value: "iframe")
145             browser.switch_to.frame(iframe3)
146             time.sleep(1)
147             table_list = []
148             table_element = browser.find_elements(By.XPATH, value: '/html/body/div[1]/div[2]/div[2]/table/tbody/tr')
149             for t in table_element:
150                 table_list.append(t.text)
151             print(len(table_list))
152             for i in range(2, len(table_list)) :
153                 mm = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[1]/span').text
154                 dd = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[2]/span').text
155                 price = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[4]/span[2]').text
156                 type = browser.find_element(By.XPATH, value: f'/html/body/div[1]/div[2]/div[2]/table/tbody/tr[{i}]/td[5]/span').text
157                 print(f'{mm},{dd},{price},{type}')]
158             try:
159                 final_list.append((sigu_data, dong_data, apt_name, area_list[area_data], mm, dd, price, type))
160                 apt_xml_a('apt_asil.csv')
161                 browser.switch_to.default_content()
162                 browser.switch_to.frame(iframe2)
163                 time.sleep(1)
164             except:
165                 final_list.append((sigu_data, dong_data, apt_name, 'None', 'None'))
166                 apt_xml_a('apt_asil.csv')
167
168     # -----
169     close_element = browser.find_elements(By.XPATH, value: '//*[@id="sub2_div"]/a').click()
170     browser.switch_to.default_content()
171     for sigu_data in sigu_list :
172         for dong_data in dong_list_recycle :
173             for url_join in apt_id_recycle :
174                 else :
175                     for area_data in range(len(area_list)):
176                         for i in range(2, len(table_list)):
```

python\_partice01 > 01\_print\_hello.py 156:59 CRLF UTF-8 4 spaces Python 3.12 (python\_partice01)

## 추출예시-2

try:

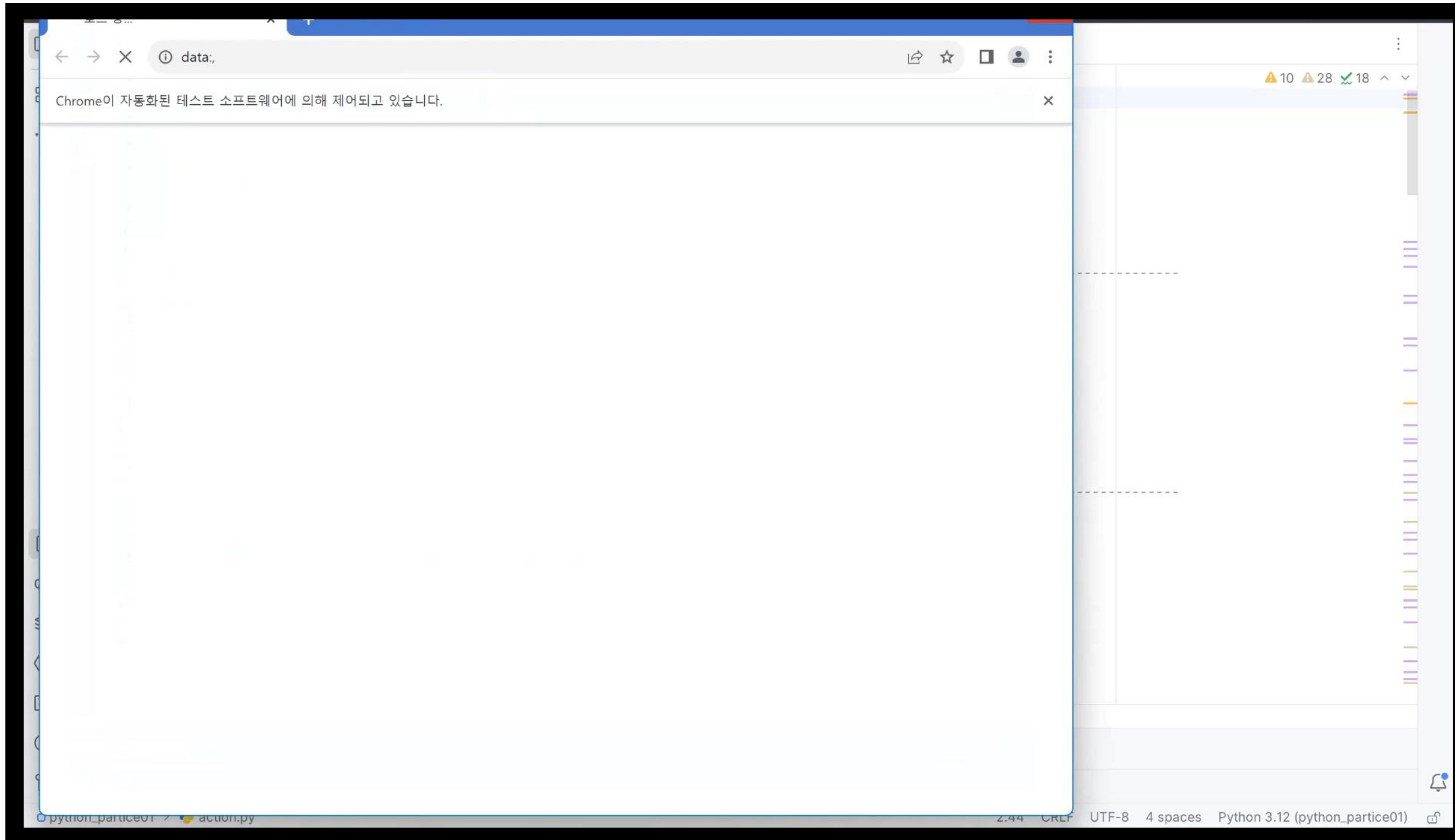
```
data_elements = browser.find_element(By.XPATH, value: '/html/body/div[1]/div[2]/div[2]/table/tbody').text  
print(data_elements)
```

1

테이블의 전체 텍스트를 불러올 수는 있지만 줄바꿈 삽입으로  
데이터 가공에 어려움이 큼니다.

```
SH대치1단지/APT 공급  
향후 아파트공급이 너무 많으면 위험해요./계약 일 경과 체결가격 타입 거래 동층  
실거래 내역이 없습니다.개포래미안포레스트/주소 : 서울 강남구 개포동 1282/계약 일 경과 체결가격 타입 거래 동층  
매매 17억 (22평)  
23.09 21 매매17억 49  
128동  
12층  
23.06 30 매매16억 1,000 49  
111동  
14층  
19 매매15억 49  
128동  
10층
```

# 추출예시-2





# 네이버 지도 N

아파트별 상한가 데이터 수집

서울시 기준 25개구 467동 3,846개  
아파트 별 상한가 데이터 수집

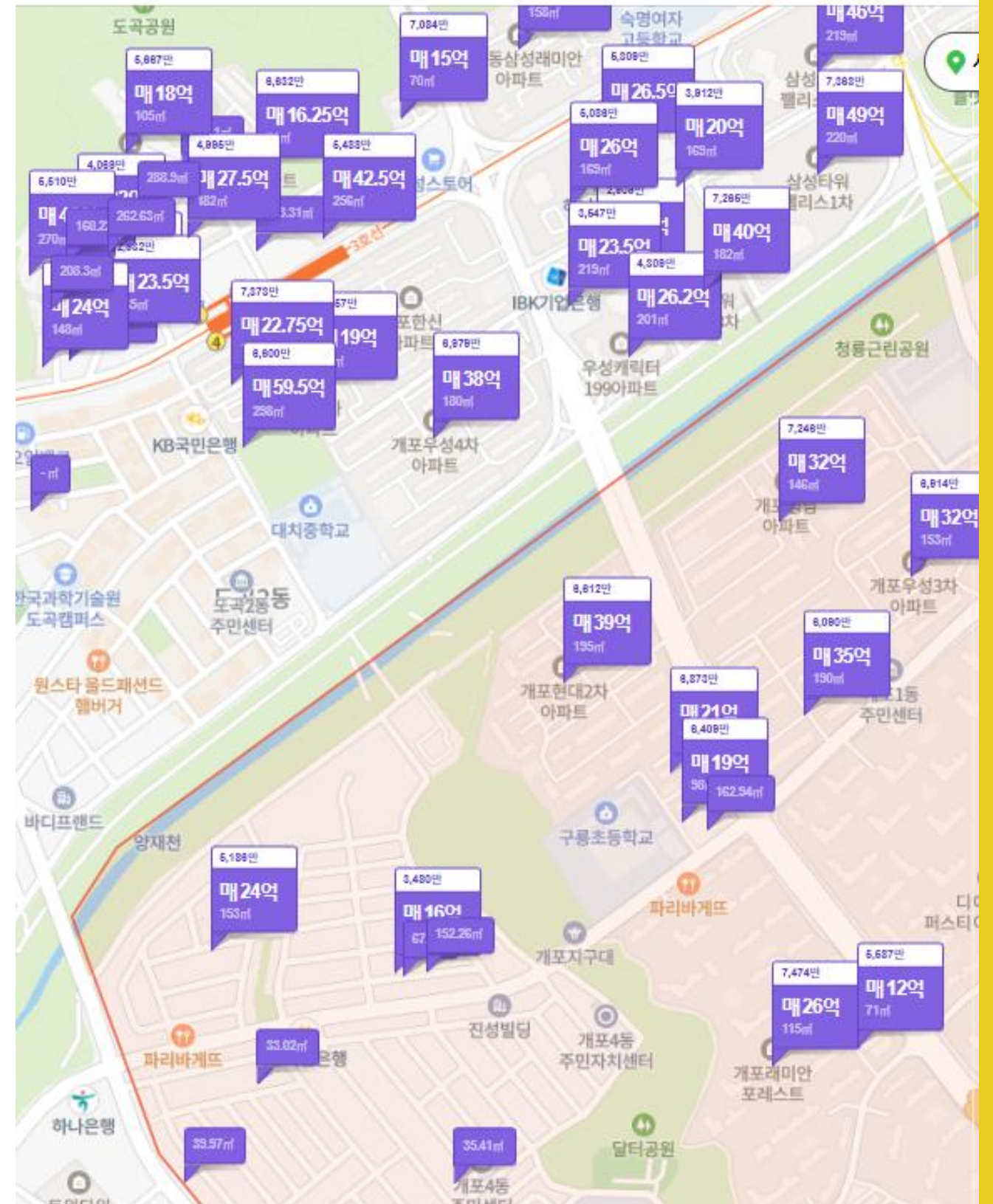
네이버 지도



N pay 부동산

아파트·오피스텔 빌라·주택 원룸·투룸 상가·업무·공장·토지 분양 MY관심 우리집

아파트 아파트분양권 재건축 오피스텔 오피스텔분양권 재개발 분양중·예정



# 네이버 부동산 데이터 탐색



네이버 지도 메인

1



지역 탐색

2



해당 지역에  
아파트 리스트 추출

3



아파트의 매물  
상한가 파악

4

# 네이버 부동산 데이터추출

```
browser = webdriver.Chrome()
browser.get('https://new.land.naver.com/complexes?ms=37.3595704,127.105399,16&a=APT:ABYG:JGC:PRE&e=RETAIL')
time.sleep(1)

element = browser.find_element(By.XPATH, value: '//*[@id="region_filter"]/div/a/span[1]').click()
time.sleep(1)

element = browser.find_element(By.XPATH, value: '//*[@id="region_filter"]/div/div/div[2]/ul/li[1]/label').click()
time.sleep(1)
```

1 브라우저 진입 및 데이터 크롤링을 위한 사이트 링크 파도타기를 진행합니다.

```
sigu_elements = browser.find_elements(By.CLASS_NAME, value: 'area_list--district')
sigu_list = []

for sigu_search in sigu_elements:
    sentences = sigu_search.text.split('\n')
    sigu_list.extend(sentences)
```

2 줄바꿈을 기준으로 문장을 분리하고 서울시의 모든 '구'를 리스트화 시켜서 추가해줍니다.



# 네이버 부동산 데이터추출

```
for dong_data in range(len(dong_list)):
    dong_element = browser.find_element(By.XPATH, f'//*[@id="region_filter"]/div/div/div[2]/ul/li[{dong_data+1}]/label').click()
    time.sleep(1)
    apt_elements = browser.find_elements(By.CLASS_NAME, 'complex_title')
    apt_list_recycle = []
    for apt_search in apt_elements:
        sentences = apt_search.text
        if sentences:
            apt_list.append(sentences)
            apt_list_recycle.append(sentences)
```

3 '구' 리스트화 이후에는 '구'에 포함된 '동' 리스트도 만듭니다.

```
for apt_data in range(len(apt_list_recycle)-1):
    button = browser.find_element(By.XPATH, f'//*[@id="region_filter"]/div/div/div[3]/ul/li[{apt_data+1}]/a')
    browser.execute_script('arguments[0].click()', button)
    time.sleep(1)
    try:
        apt_won = browser.find_element(By.LINK_TEXT, '시세/실거래가')
        apt_won.click()
        time.sleep(1)
```

4 '동'에 포함된 모든 '아파트' 리스트를 만듭니다.

```

try:
    arrow = browser.find_element(By.XPATH, '//*[@id="detailContents2"]/div[1]/div/div[2]/button')
    if arrow.text == '더보기':
        arrow.click()
        time.sleep(1)
        area_elements = browser.find_elements(By.CLASS_NAME, 'detail_sorting_tab')
        time.sleep(1)
        area_list = []
        for area_search in area_elements:
            sentences = area_search.text
            if sentences:
                area_list.append(sentences)
        for area_data in area_list[:-3]:
            browser.find_element(By.LINK_TEXT, f'{area_data}').click()
            time.sleep(1)

        try:
            price = browser.find_element(By.XPATH, '//*[@id="tabpanel1"]/div[3]/div[1]/div/div[2]/strong')
            price_list.append(price)
            final_list.append((sigu_list[sigu_data], dong_list[dong_data], apt_list_recycle[apt_data], area_data))
            time.sleep(1)
        except:
            final_list.append((sigu_list[sigu_data], dong_list[dong_data], apt_list_recycle[apt_data], area_data))

    if arrow.text == '더보기':
        arrow.click()
        time.sleep(1)

except:
    area_elements = browser.find_elements(By.CLASS_NAME, 'detail_sorting_tab')
    time.sleep(1)
    area_list = []
    for area_search in area_elements:
        sentences = area_search.text
        if sentences:
            area_list.append(sentences)
    for area_data in area_list[:-3]:
        browser.find_element(By.LINK_TEXT, f'{area_data}').click()
        time.sleep(1)

    try:
        price = browser.find_element(By.XPATH, '//*[@id="tabpanel1"]/div[3]/div[1]/div/div[2]/strong').text
        price_list.append(price)
        final_list.append((sigu_list[sigu_data], dong_list[dong_data], apt_list_recycle[apt_data], area_data))
        time.sleep(1)
    except:
        final_list.append((sigu_list[sigu_data], dong_list[dong_data], apt_list_recycle[apt_data], area_data))

```



## try ~ except

‘평수’, ‘매물’ 이 아예 없는 경우에는 데이터를 크롤링 하는 과정에서 오류가 발생할 수 있기 때문에, try를 활용해서 없으면 예외처리를 하여, 해결하였습니다.



## area

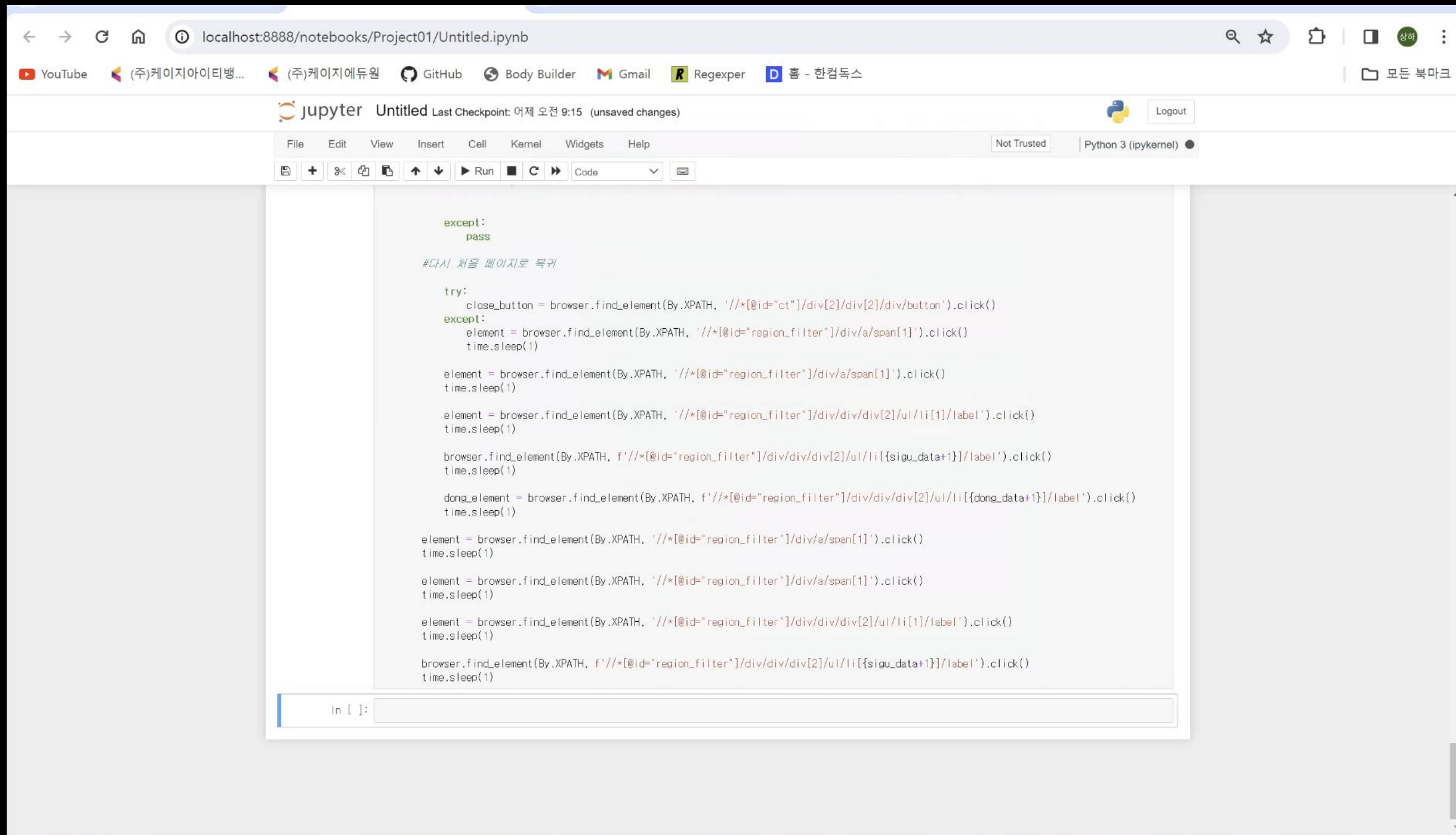
‘평수’ 구분을 하기위해서 리스트 처리하여, 각 아파트를 분류한 후 아파트의 평수별로 ‘상한가’ 출력을 위한 데이터를 찾습니다.



## final\_list

마지막에 CSV 파일로 입력을 위해서 일괄적으로 데이터를 저장할 리스트를 만들어 주면서 결과값이 나오는 데이터를 추가 합니다.

# 네이버 추출 예시



```
except:
    pass

#다시 처음 페이지로 복귀

try:
    close_button = browser.find_element(By.XPATH, '//*[@id="ct"]/div[2]/div[2]/div/button').click()
except:
    element = browser.find_element(By.XPATH, '//*[@id="region_filter"]/div/a/span[1]').click()
    time.sleep(1)

element = browser.find_element(By.XPATH, '//*[@id="region_filter"]/div/a/span[1]').click()
time.sleep(1)

element = browser.find_element(By.XPATH, '//*[@id="region_filter"]/div/div/div[2]/ul/li[1]/label').click()
time.sleep(1)

browser.find_element(By.XPATH, f'//*[@id="region_filter"]/div/div/div[2]/ul/li[{sign_data+1}]/label').click()
time.sleep(1)

dong_element = browser.find_element(By.XPATH, f'//*[@id="region_filter"]/div/div/div[2]/ul/li[{dong_data+1}]/label').click()
time.sleep(1)

element = browser.find_element(By.XPATH, '//*[@id="region_filter"]/div/a/span[1]').click()
time.sleep(1)

element = browser.find_element(By.XPATH, '//*[@id="region_filter"]/div/a/span[1]').click()
time.sleep(1)

element = browser.find_element(By.XPATH, '//*[@id="region_filter"]/div/div/div[2]/ul/li[1]/label').click()
time.sleep(1)

browser.find_element(By.XPATH, f'//*[@id="region_filter"]/div/div/div[2]/ul/li[{sign_data+1}]/label').click()
time.sleep(1)
```



## • CSV

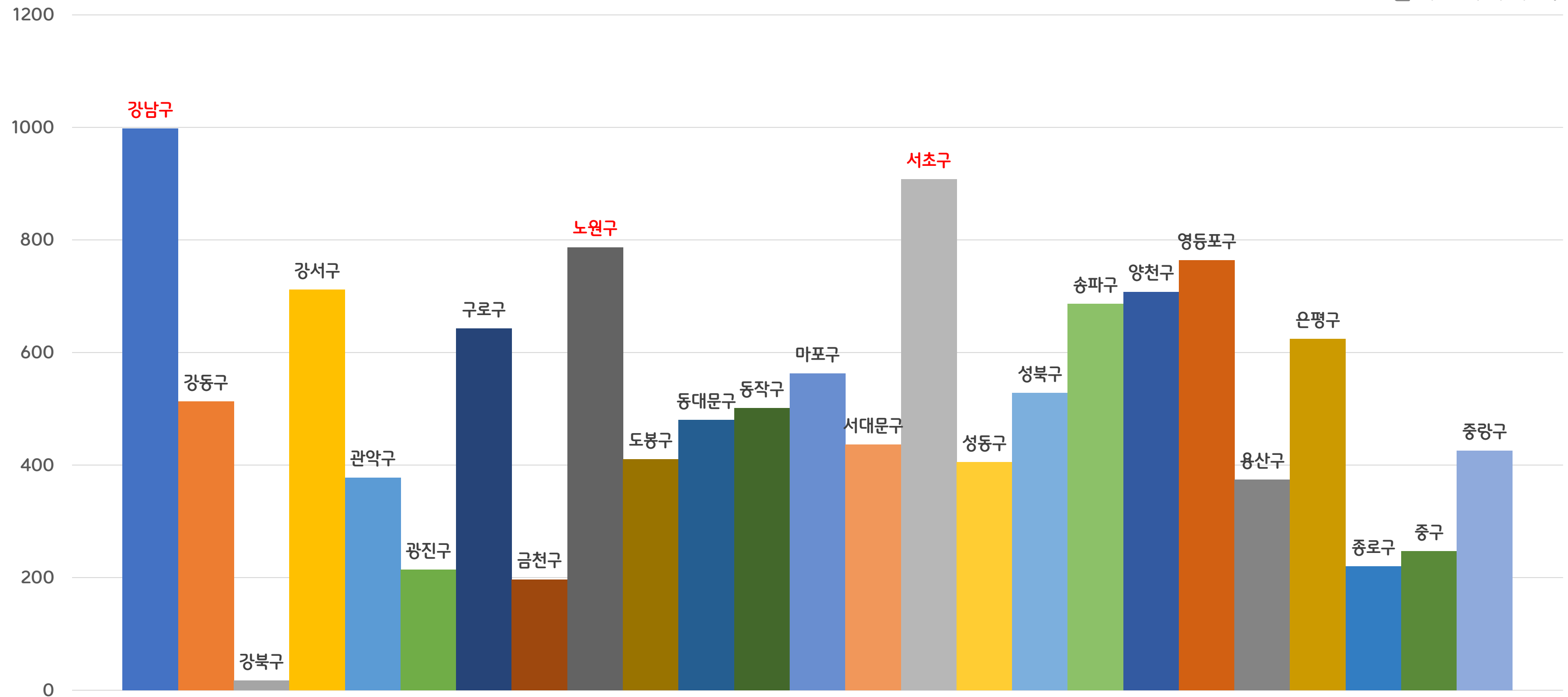
### CSV 파일 EXPORT 결과

각 구에 동 별로 데이터를 정리하여, 면적별로 '상한가' 데이터를 입력 받아서 넣고, 해당 매물이 없는 경우에는 'None' 데이터를 삽입해서 구분 합니다. 최종적으로 만 이천개의 상한가 데이터를 가지고 출력 하여, CSV 파일로 정렬 합니다.

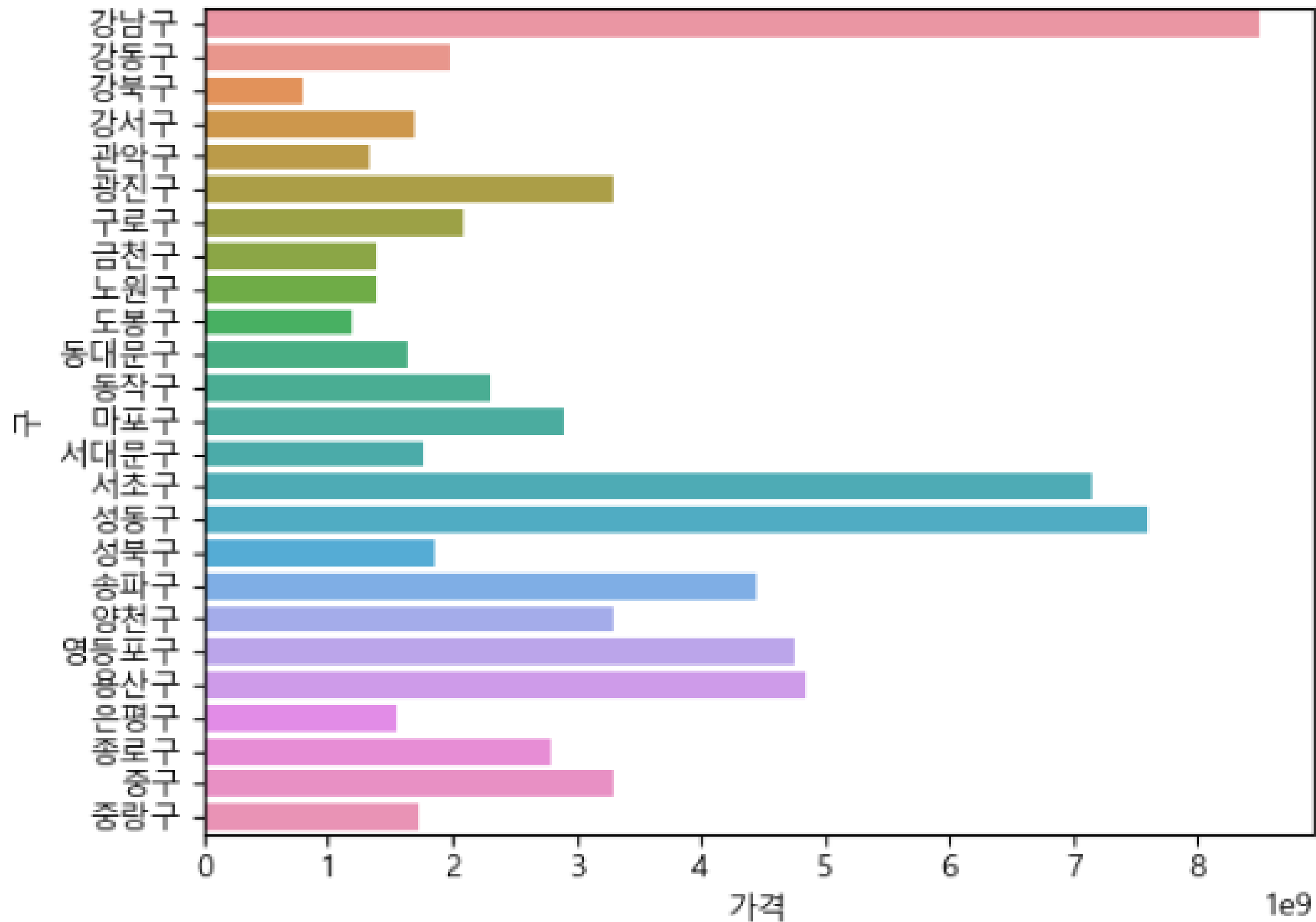
4	강남구	개포동	개포래미안	86A m <sup>2</sup>	18억 5,000	
5	강남구	개포동	개포래미안	86B m <sup>2</sup>	18억 5,000	
6	강남구	개포동	개포래미안	101 m <sup>2</sup>	23억 5,000	
7	강남구	개포동	개포래미안	115A m <sup>2</sup>	25억	
8	강남구	개포동	개포래미안	115B m <sup>2</sup>	25억	
9	강남구	개포동	개포래미안	115C m <sup>2</sup>	25억	
10	강남구	개포동	개포래미안	115D m <sup>2</sup>	25억	
11	강남구	개포동	개포래미안	128A m <sup>2</sup>	None	
12	강남구	개포동	개포래미안	129B m <sup>2</sup>	None	
13	강남구	개포동	개포래미안	135A m <sup>2</sup>	None	
14	강남구	개포동	개포래미안	136B m <sup>2</sup>	None	
15	강남구	개포동	개포래미안	148A m <sup>2</sup>	None	
1091	노원구	하계동	학여울청구	58 m <sup>2</sup>	4억 4,000	
1092	노원구	하계동	한신	132 m <sup>2</sup>	11억	
1093	노원구	하계동	한신	161 m <sup>2</sup>	11억	
1094	노원구	하계동	한신동성	104 m <sup>2</sup>	10억	
1095	노원구	하계동	한신동성	123 m <sup>2</sup>	10억	
1096	노원구	하계동	한신동성	162 m <sup>2</sup>	12억 5,000	
1097	노원구	하계동	한신코아	87 m <sup>2</sup>	9억	
1098	노원구	하계동	한신코아	87B m <sup>2</sup>	9억	
1099	노원구	하계동	한신코아	87A m <sup>2</sup>	9억	
1100	노원구	하계동	한신코아	101A m <sup>2</sup>	9억	
1101	노원구	하계동	한신코아	102B m <sup>2</sup>	10억	
1102	노원구	하계동	한신코아	103C m <sup>2</sup>	10억	

# 서울시 구 별 아파트 수 비교

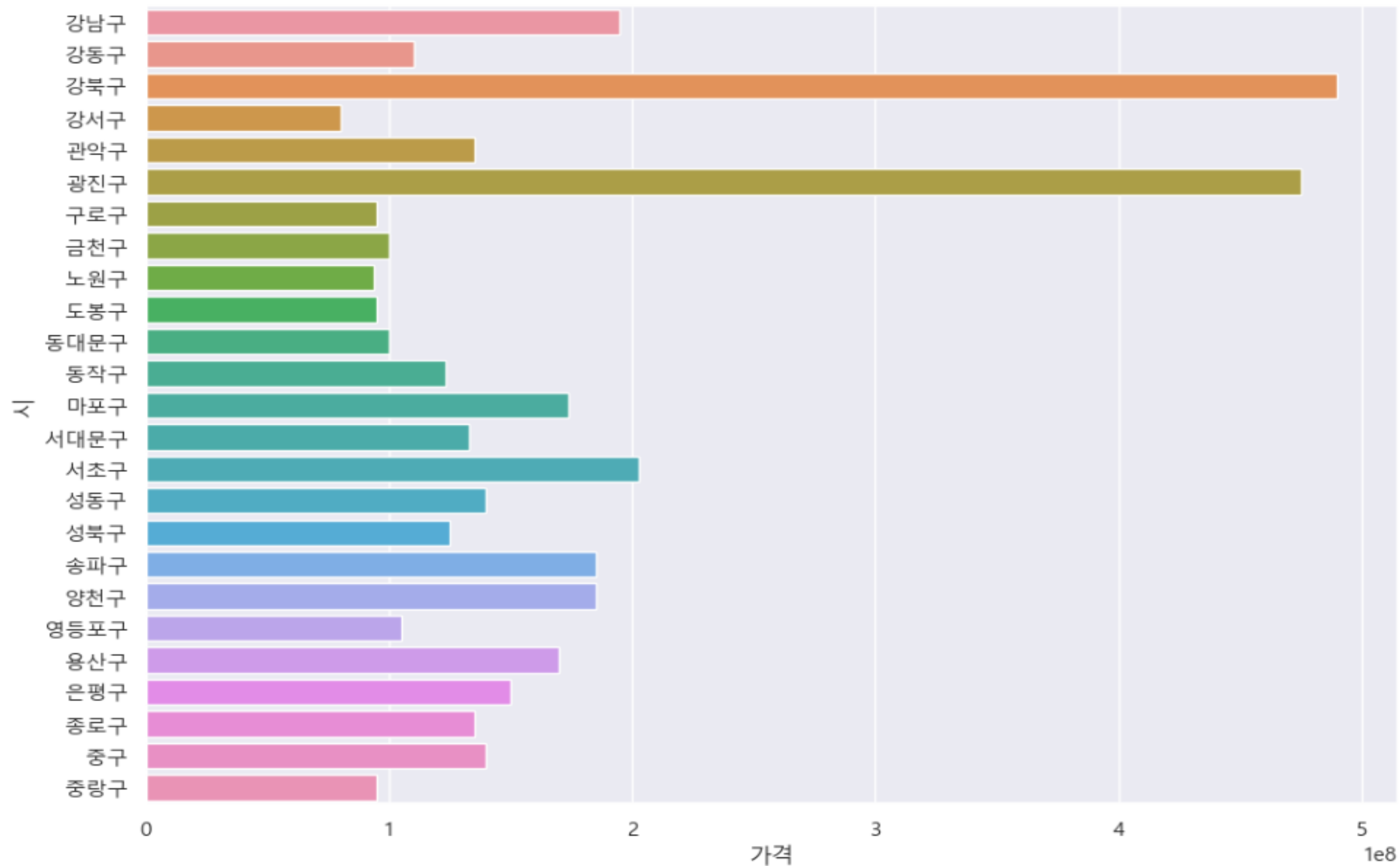
출처 : 네이버 부동산



# 서울시 지역별 최고가

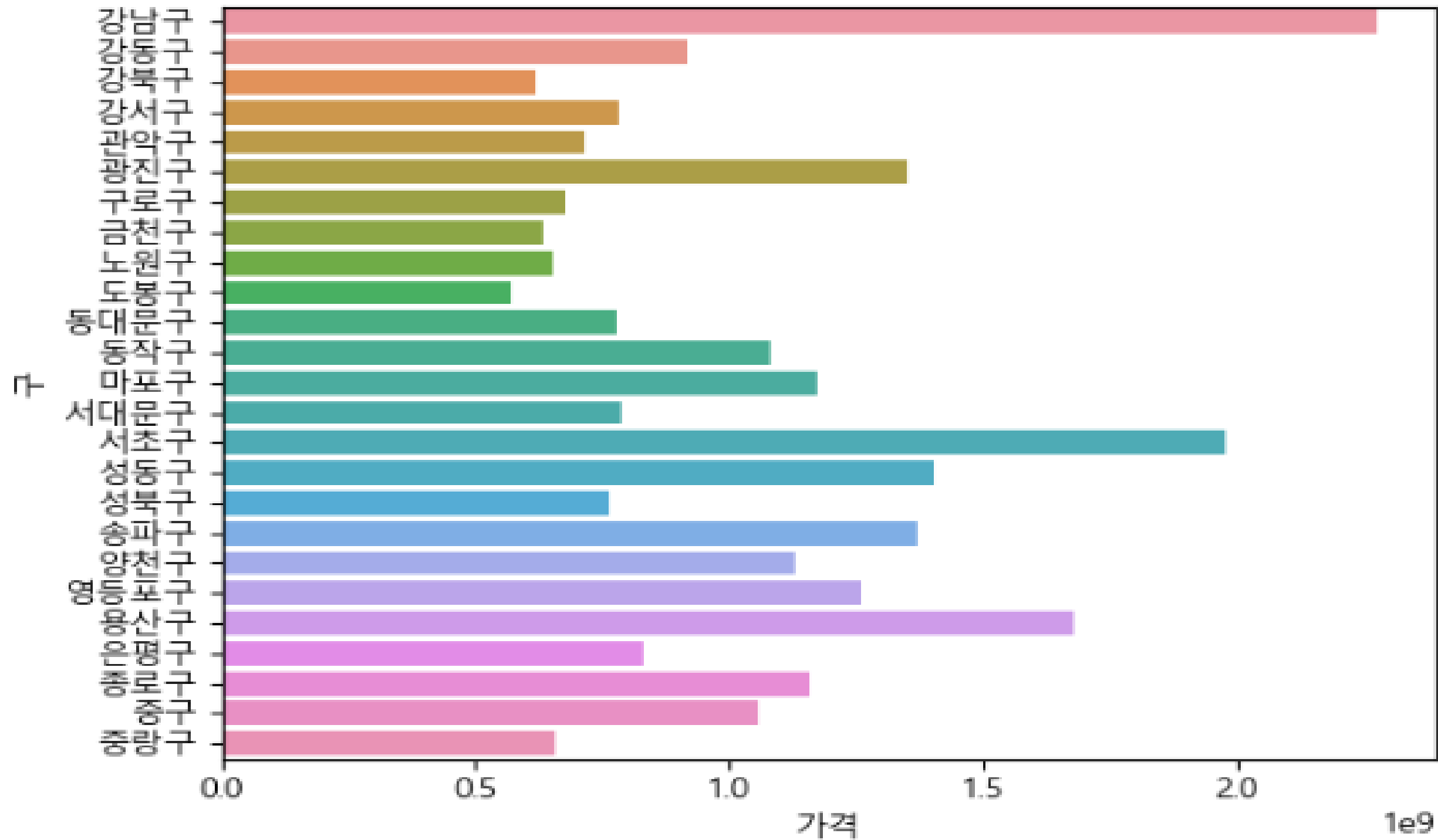


# 서울시 지역별 최저가





# 서울시 지역별 평균가



# 감사합니다



DROP THE BIT

\*

김동규 · 박성훈 · 정선우 · 윤상하