

# How Happy are you?

-Happiness in Seoul

HAPPINESS AND HEALTH

김민경  
엄민섭  
정현서  
한동규

TEAM 4



# Topics

1. 행복이란?
2. 주제 선정 WHY?
3. Data used
4. 코로나와 행복지수
5. 행복의 요인들
6. 결론

# What is happiness?

-행복이란?-

-

희망을 그리는 상태에서의 좋은 감정.  
심리적인 상태 또는 자신이 원하는 욕구와  
욕망이 충족되어 만족하거나 즐거움과  
여유를 느끼는 상태.



SIMPLY PUT,  
A LIFE LIVED IN VIRTUE

WHY?

04



There can be no happiness  
without Good Health

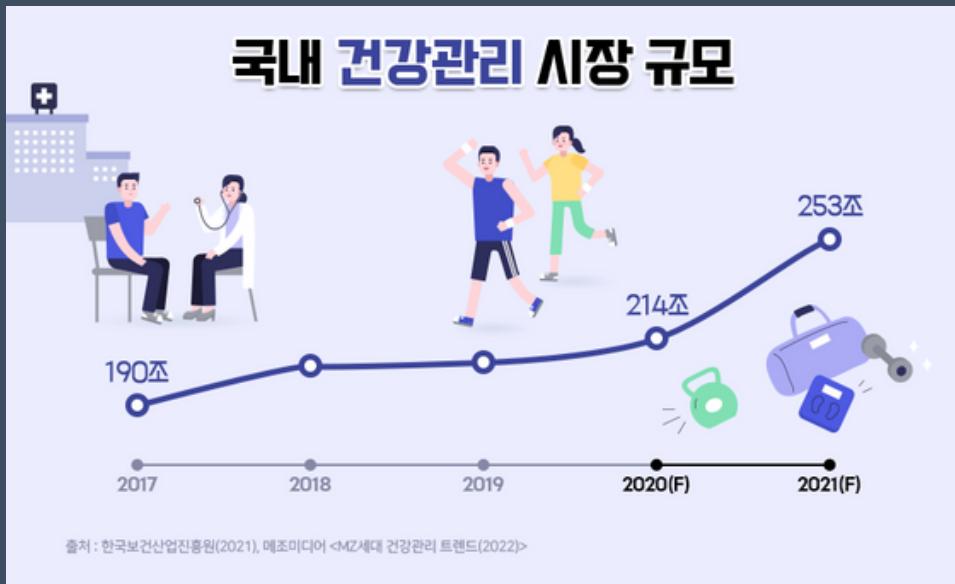
Happiness  
and Health

지역주민의 행복도 결정요인에 관한 연구 -  
안산시를 중심으로 - 김상곤 & 김상중 (2008)

## 7) 주관적인 건강도와 행복도

자신의 건강상태를 보통이상으로 응답한 응답비율은 88.8%에 이른다. 반면 나쁘다고 응답한 비율은 10.3%에 불과하였다. 건강이 나쁘다고 응답한 응답자들의 행복정도는 스스로 행복하다는 응답과 행복하지 않다는 응답이 거의 비슷한 비율로 나타나고 있다. 반면, 건강이 좋다고 응답한 응답군에서는 스스로 행복하다는 응답이 훨씬 많음을 알 수 있다. 이는 자신의 주관적 건강도가 어느 정도 행복도에 영향을 미치고 있음을 의미하는 것이다. 이러한 연구결과는 전반적인 건강상태가 개인의 행복도를 결정하는 요인이라는 선행연구(추홍규, 2004)와 일치하는 결과이다.

# Happiness and Health



## 건강 관련 시장 규모의 증가

WHY?

2020년 택배 물량 비중 2위

# Happiness and Health



# Data

---

## 1. 서울 서베이 데이터

- 매해 서울시 거주 5만명 대상 조사
- 2018 ~ 2021년



## 2. 네이버 뉴스 크롤링

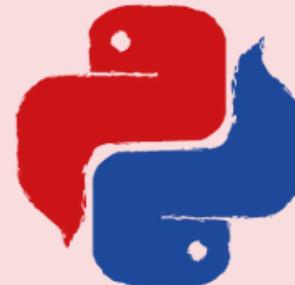
- NAVER NEWS CRAWLING



## 3. 서울시 구별 통계자료

- 녹지면적, 약국, 병원, 공원수 등





KoNLPy



DATA  
ANALYSIS  
LIBRARY

# Coding

---



```
raw2018 = pd.read_excel(f'{src}/2018 서울서베이 가구원_데이터.xlsx')
raw2019 = pd.read_excel(f'{src}/2019 서울서베이 가구원_데이터.xlsx')
raw2020 = pd.read_excel(f'{src}/2020 서울서베이 가구원_데이터.xlsx')
raw2021 = pd.read_excel(f'{src}/2021 서울서베이 가구원_데이터.xlsx')
```



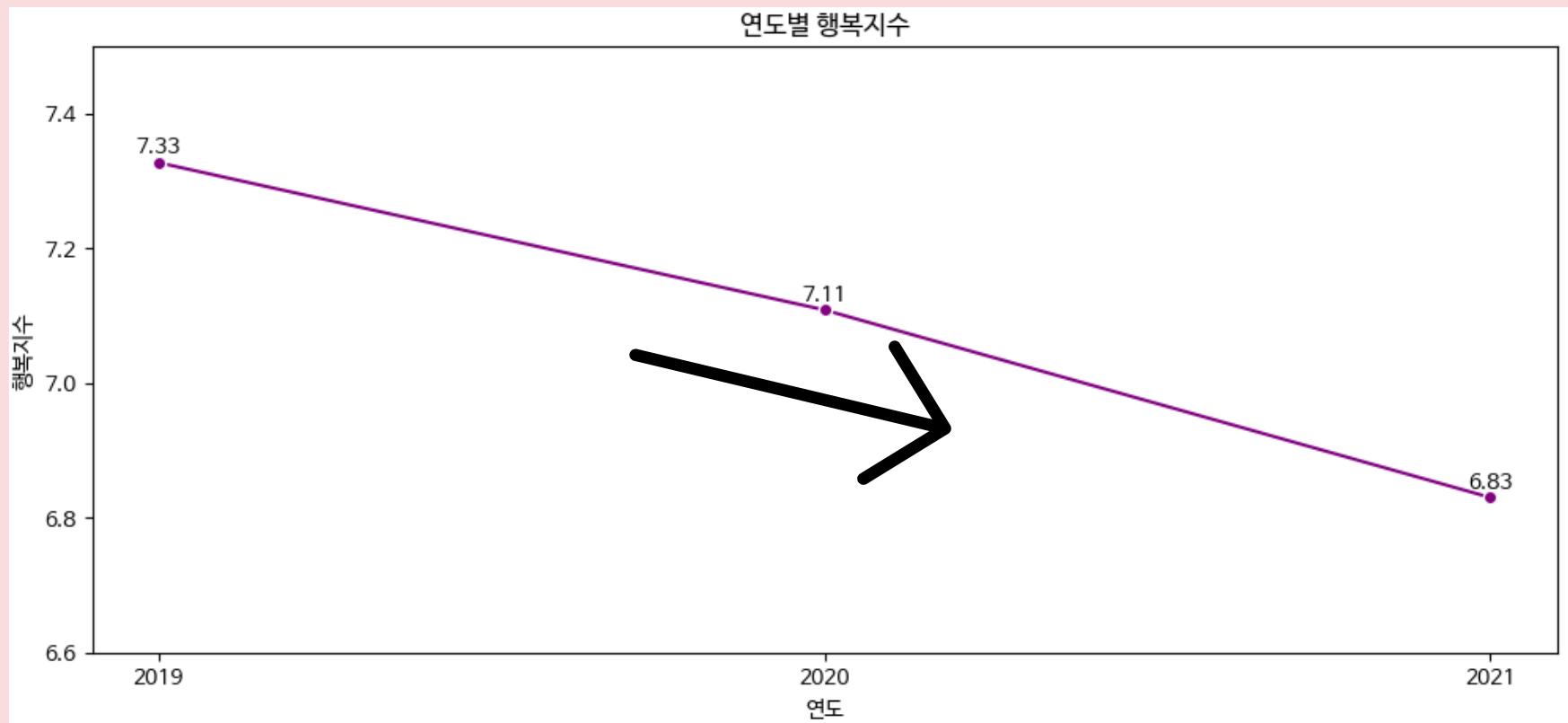
```
# 중복 행 제거, 부호 변환, 공백 제거
buho = pd.read_excel(f'{src}/공통부호.xlsx')
buho.drop([0,7,20,21,33], axis=0, inplace=True)
buho.replace({'AQ21': 'SQ1_1'}, inplace=True)
buho = buho.apply(lambda x: x.str.replace(' ', ''))

raws = [raw2018, raw2019, raw2020, raw2021]
years = range(2018, 2022)
for year, data in zip(years, raws):
    cols = buho[year].tolist()
    data[cols].to_csv(f'processed{year}.csv', index=False, encoding='utf-8-sig')
```

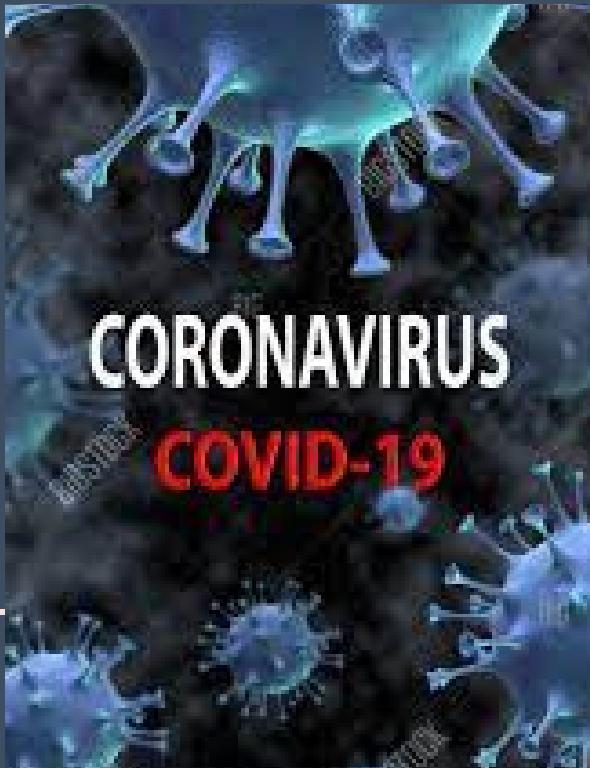
1-10 중,  
당신의 건강 행복지수는  
어느 정도 입니까?

# Seoul Survey

서울시 평균 건강 행복지수



# Happiness § Health with Covid 19



# Coding

```
● ● ●

def crawling(page, keyword, sort_type, start_date, end_date):
    url = f"https://search.naver.com/search.naver?where=news&sm=tab_pge&query={keyword}&sort={sort_type}&photo=0&field=0&pd=3&ds={start_date}&de={end_date}&cluster_rank=23&mynews=0&office_type=0&office_section_code=0&start={page}1"
    try:
        response = urllib.request.urlopen(url)
        rescode = response.getcode()
        if rescode == 200:
            response_body = response.read()
            soup = BeautifulSoup(response_body, 'html.parser')
            news_items = soup.select("#main_pack > section > div > div.group_news > ul")
            news_titles = [item.text for item in news_items]
            return news_titles
        else:
            raise urllib.error.HTTPError(f"HTTP Error {rescode}")
    except:
        return print('error')

def crawl_news(keyword, start_date, sort_type=0, pages=100, month=0, week=0, day=0, iters=10):
    if month != 0:
        filename = f"{keyword}_{start_date.strftime('%Y-%m-%d')}_{{month*iters}}월_{{month}}월간격_{{pages}}p.csv"
    elif week != 0:
        filename = f"{keyword}_{start_date.strftime('%Y-%m-%d')}_{{week*iters}}주_{{week}}주간격_{{pages}}p.csv"
    elif day != 0:
        filename = f"{keyword}_{start_date.strftime('%Y-%m-%d')}_{{day*iters}}일_{{day}}간격_{{pages}}p.csv"
    else:
        filename = f"{keyword}_{start_date.strftime('%Y-%m-%d')}_{{pages}}p.csv"

    news_list = []
    keyword_encoded = urllib.parse.quote(keyword)
    interval = relativedelta(months=month, weeks=week, days=day)
    for iter in tqdm(range(iters), desc="Progress:"):
        start = (start_date + interval * i).strftime("%Y.%m.%d")
        end = (start_date + interval * (i+1)).strftime("%Y.%m.%d")
        args_list = [(page, keyword_encoded, sort_type, start, end) for page in range(pages)]
        with Pool() as pool:
            page_results = pool.starmap(crawl_news, args_list)
        for page_result in page_results:
            news_list.extend(page_result)
    news_df = pd.DataFrame(news_list)
    news_df.to_csv()
    return news_df(filename, encoding='utf-8-sig')
```



# Coding

---



```
def outlier(data, value):
    data = pd.Series(data)
    level_1q = data.quantile(0.25)
    level_3q = data.quantile(0.75)
    IQR = level_3q - level_1q
    rev_range = value # 제거 범위 조절 변수
    df = data[(data <= level_3q + (rev_range * IQR)) & (data >= level_1q - (rev_range *
IQR))]
    if df.empty:
        return None
    else:
        df = df.reset_index(drop=True)
        return df.mean()

df = pd.read_csv('./19-21_Year&Gu&Hp.csv', index_col=0)
df_ol_15 = df.groupby(['구', 'year'])['행복지수(건강)'].agg(lambda x: outlier(x, 1.5)).unstack()
df_ol_30 = df.groupby(['구', 'year'])['행복지수(건강)'].agg(lambda x: outlier(x, 3)).unstack()
df_ol_15.to_csv('./19-21_hp_mean_byOutlier15.csv', encoding='utf-8-sig')
df_ol_30.to_csv('./19-21_hp_mean_byOutlier30.csv', encoding='utf-8-sig')
```

# Coding



```
## wordcloud 생성 및 이미지 mask 처리
mask=np.array(Image.open(f'{src}/happiness.png'))
text = word_counter

wc = WordCloud(
    font_path="NanumBarunGothic",
    background_color="white",
    max_words=200000,
    mask=mask,
    min_font_size=10,
    max_font_size=100
).generate_from_frequencies(text)

image_colors = ImageColorGenerator(mask)
wc = wc.recolor(color_func=image_colors)

## 그래프 설정하고 출력
plt.figure(figsize=(90,76))
plt.imshow(wc)
plt.axis("off")
plt.tight_layout(pad=0)
plt.show()
```

# Coding

---



```
kkma = Kkma()
data = pd.read_csv(f"{src}/건강 만족도_2020-01-01_52주_1주간격_10p.csv", index_col = 0)
nouns_data = data['0'].apply(lambda x: kkma.nouns(x))
```



```
word_counter = Counter()
[word_counter.update(i) for i in nouns_data]
word_counter = {i:word_counter[i] for i in word_counter if (len(i)>1) & ~(i.isdigit())}
word_counter = dict(sorted(word_counter.items(), key = lambda x: x[1], reverse=True))

word_counter.pop('만족도')
word_counter.pop('건강')

word_freq = pd.Series(word_counter)
pd.DataFrame(word_freq).to_csv('word_frequency.csv', encoding='utf-8-sig')
```

2019년

# Wordcloud



2020년

# Wordcloud

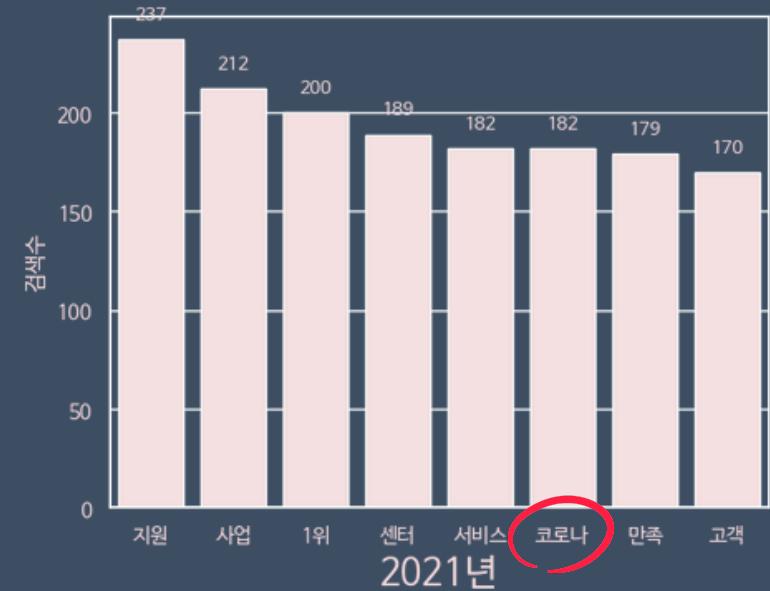


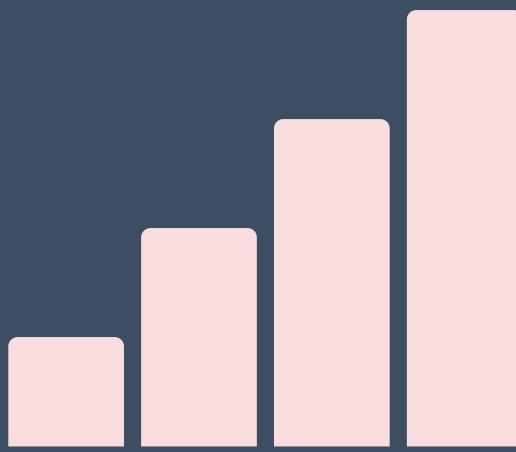
2021년

# Wordcloud



# Keywords





Pre-  
investigation  
- 사전조사 -

# 사전조사

- 건강관련 시설
  - 병원, 약국
- 환경
  - 농지 면적 / 공원, 체육 시설
- 개인의 재정상태
  - 평균 소득
- 연령별 분포도



# Coding



```
gu_lat_long = pd.read_excel(f'{src}/구별위경도_25.xlsx',header = None)
df_seoul = pd.read_csv('./merged_survey.csv')

geo_path = f'{src}/geo_state.json'
try:
    geo_data = json.load(open(geo_path, encoding='utf-8'))
except:
    geo_data = json.load(open(geo_path, encoding='utf-8-sig'))
```



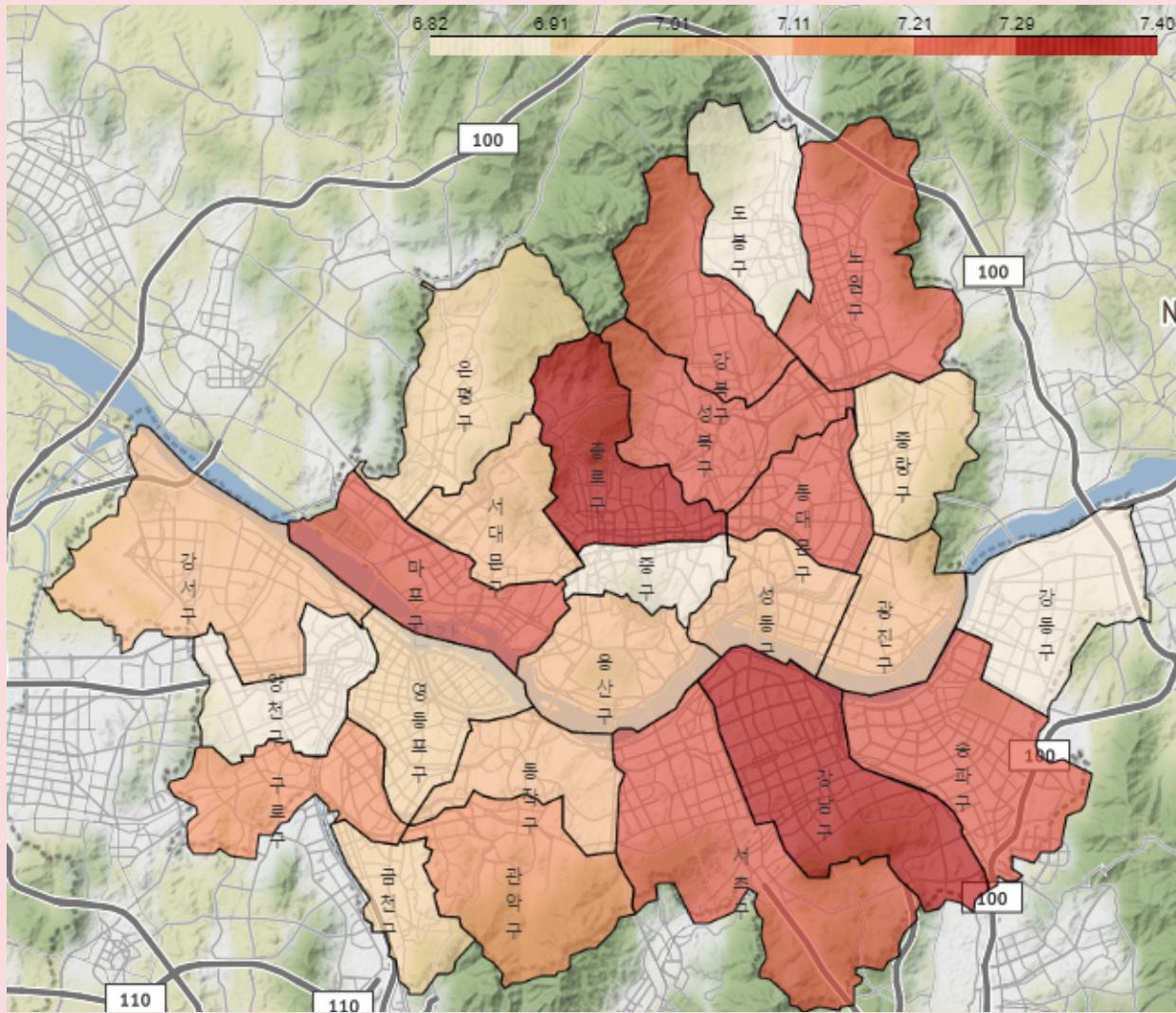
```
seoul_map = folium.Map(
    location=[37.55,126.98],
    tiles ='Stamen Terrain',
    zoom_start=12
)
gu_lat_long.columns = ['경도','위도','구']
happiness = df_seoul.groupby(['구'])[['행복지수(건강)']].mean()
```

# Coding



```
folium.Choropleth(  
    geo_data = geo_data, # 지도 경계  
    data = happiness['행복지수(건강)'], # 표시하려는 데이터  
    columns=[happiness.index, happiness['행복지수(건강)']], # 열 기준  
    threshold_scale = [6.5,6.7,6.9,7.1,7.4],  
    key_on = 'feature.properties.name',  
    fill_color = 'OrRd'  
).add_to(seoul_map)  
  
for i, row in gu_lat_long.iterrows():  
    temp = row["구"]  
    s = "Info"  
    folium.Marker(  
        [row["위도"], row["경도"]],  
        icon = folium.DivIcon(  
            html = f"""  
                <div style = "font-family: fantasy; color: black; font-size: 10px; font-weight: 900; ">  
                    {"{}".format(temp)}</div>  
                """  
        )  
    ).add_to(seoul_map)  
seoul_map.save('~/seoul.html') #png 형태로도 저장 가능
```

# 구별 평균 건강 행복지수



# The Happiness Experiment

---



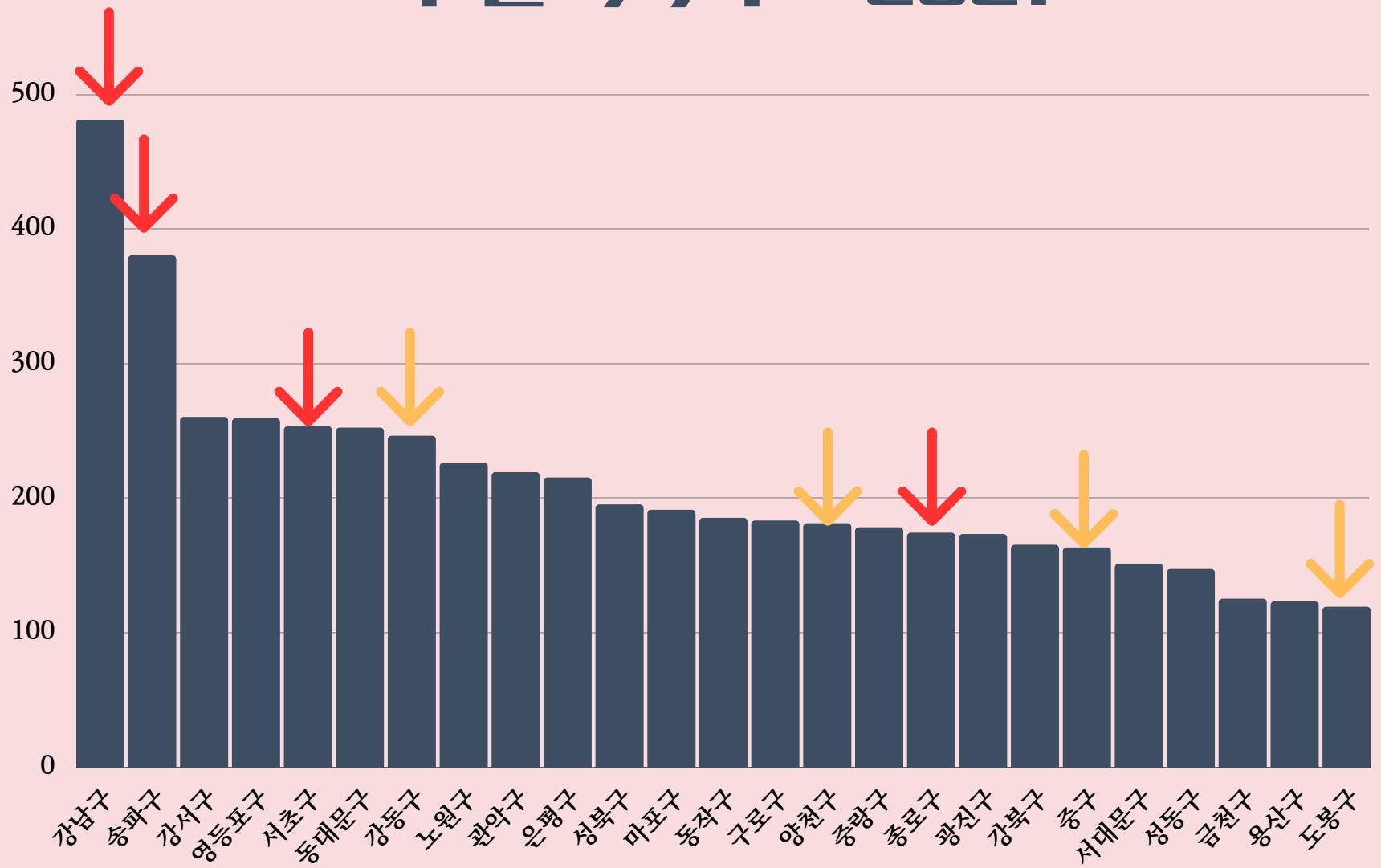
# Coding



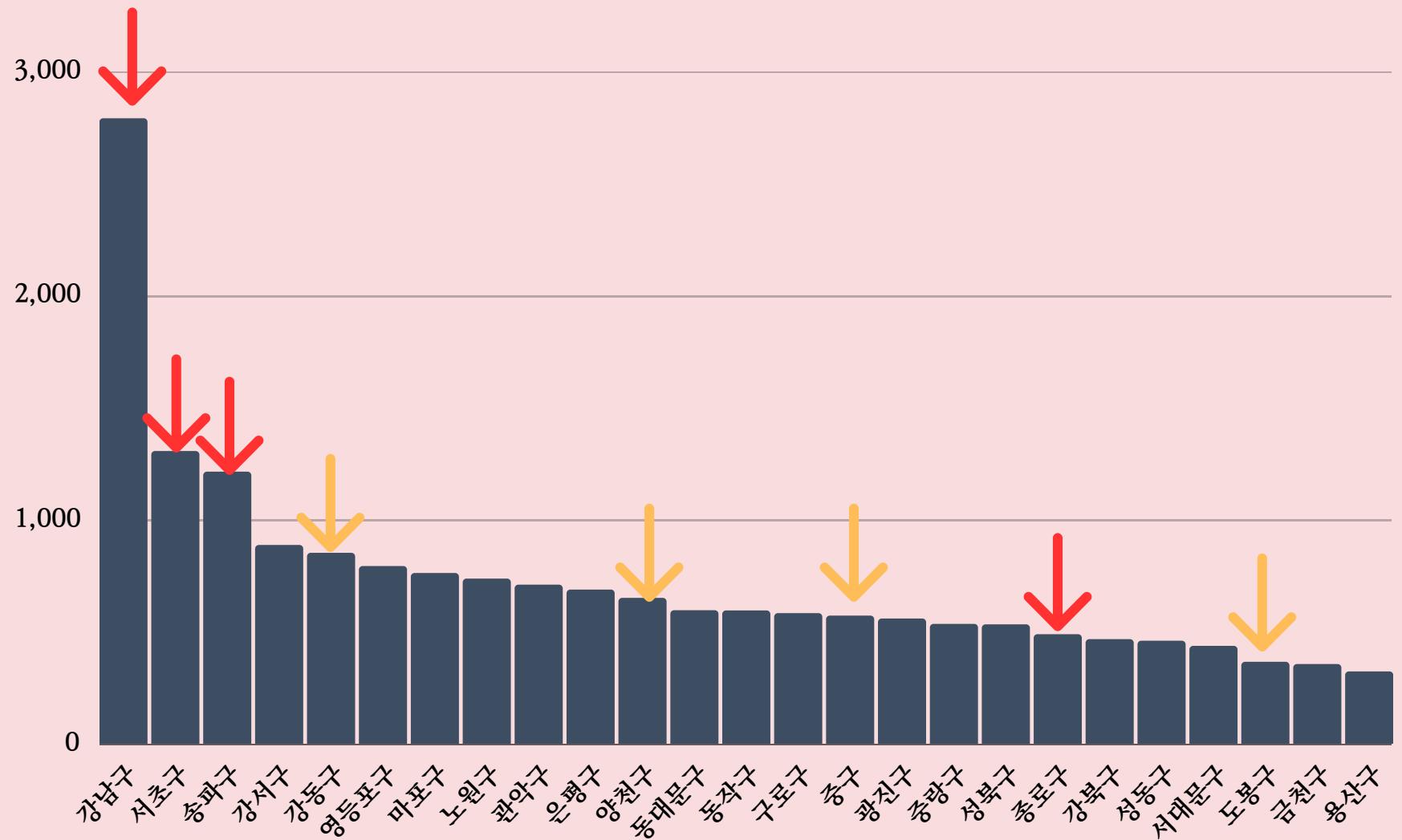
```
url = f'{src}/processing_data/processing_data(병원수).csv'
df_hospital = pd.read_csv(url)

df_hospital_sorted = df_hospital.sort_values(by="병원수", ascending = False)
plt.figure(figsize=(15,10))
ax_hospital = sns.barplot(
    data = df_hospital_sorted,
    x = '지역구',
    y = '병원수'
)
for p in ax_hospital.patches:
    ax_hospital.text(
        p.get_x() + p.get_width() / 2,
        p.get_height() + 10,
        int(p.get_height()),
        ha='center'
    )
plt.show()
```

# 구별 약국 수 - 2021

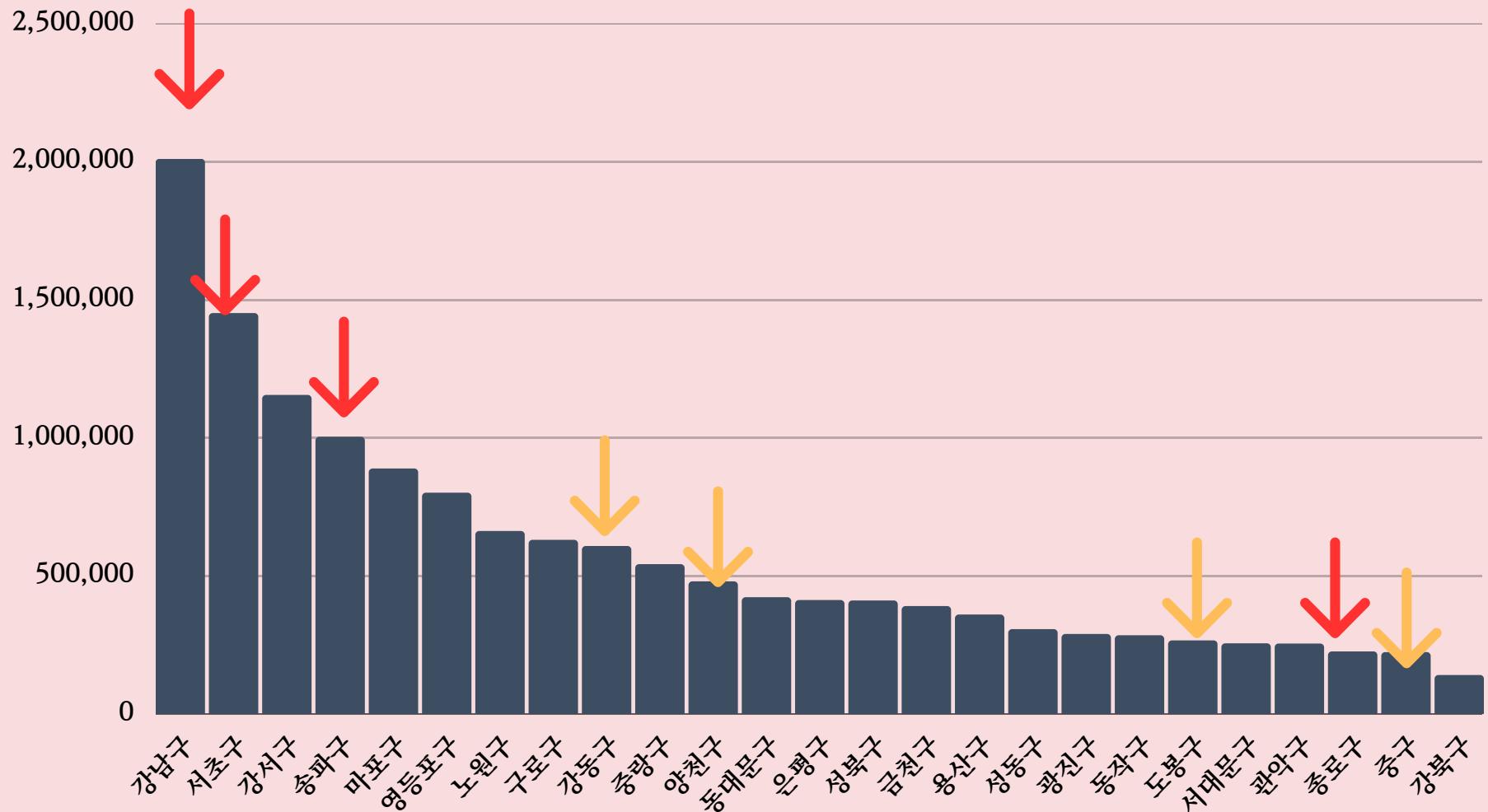


# 구별 병원 수 - 2021



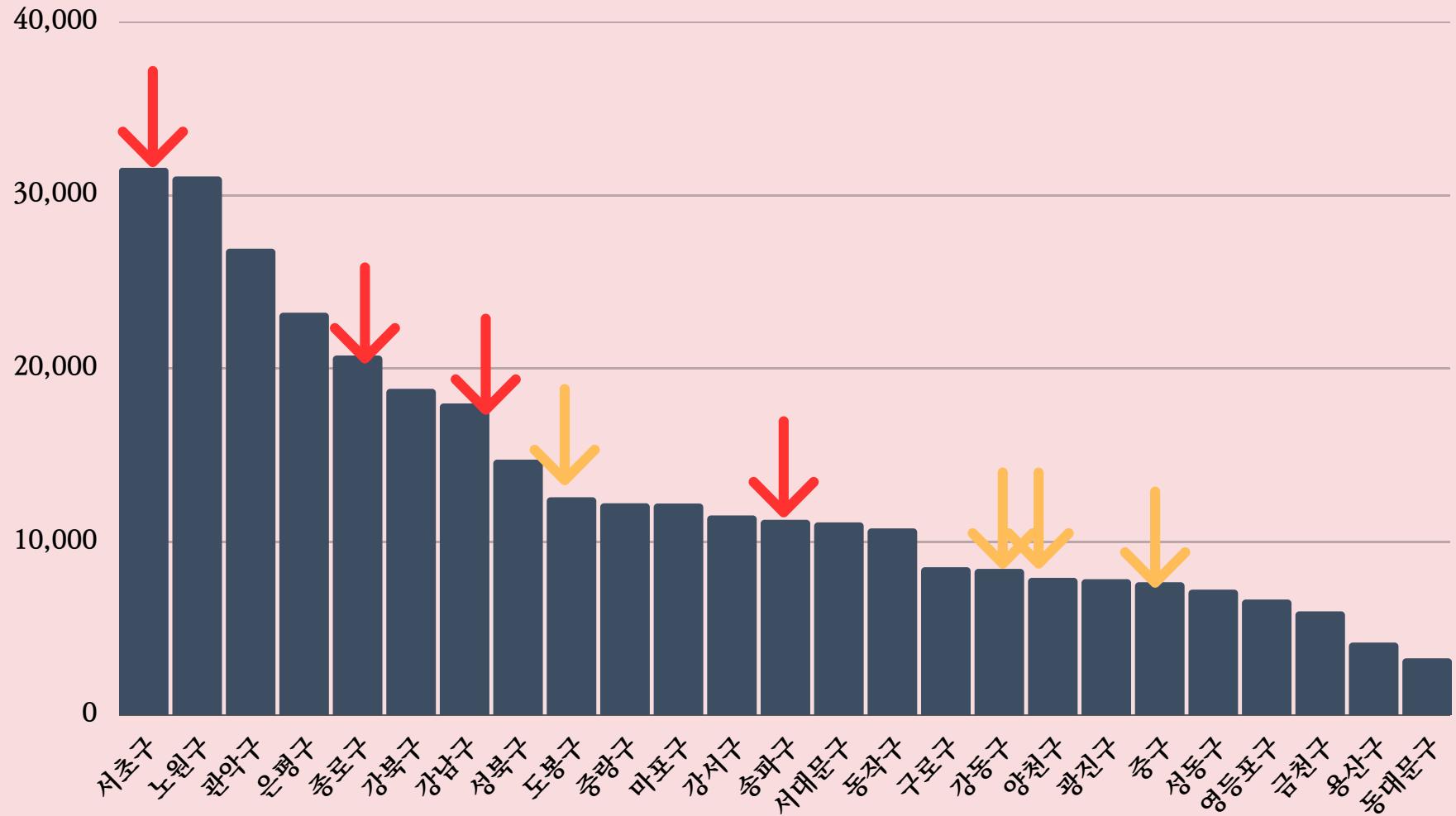
# 구별 녹지 면적 - 2021

■ 단위: m<sup>2</sup>



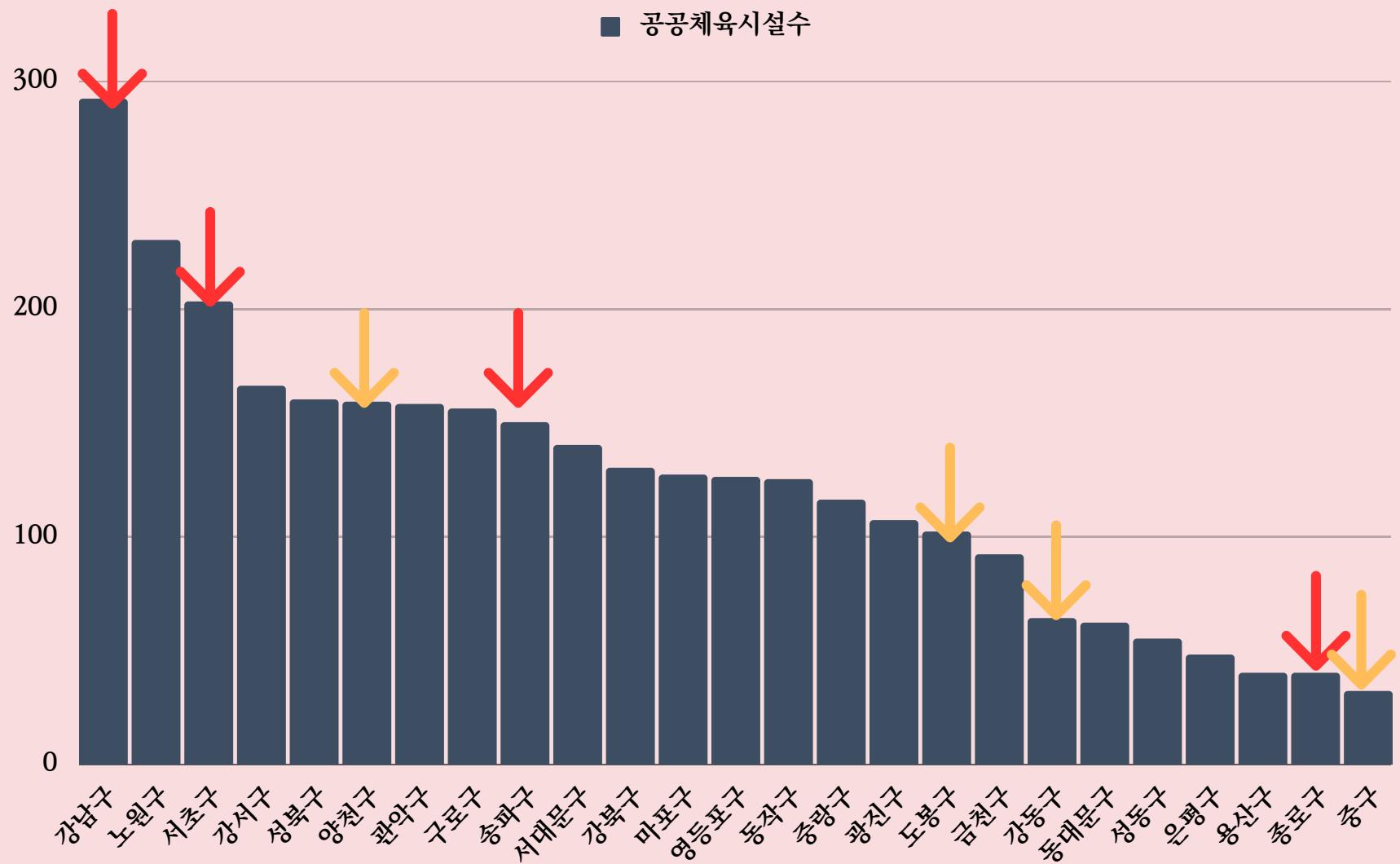
# 구별 공원 면적 - 2021

■ 단위: 천m<sup>2</sup>

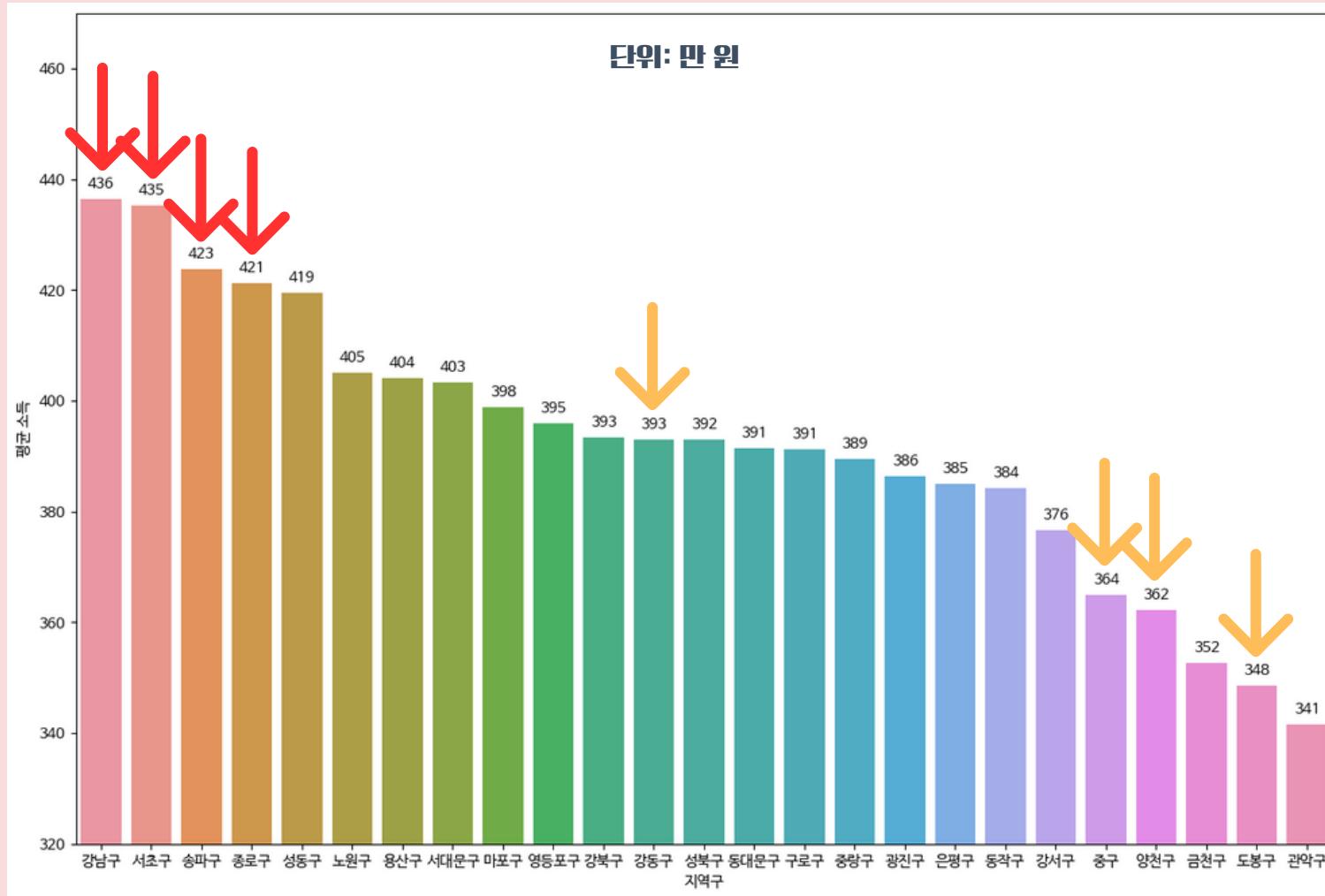


# 구별 공공 체육시설 수 - 2021

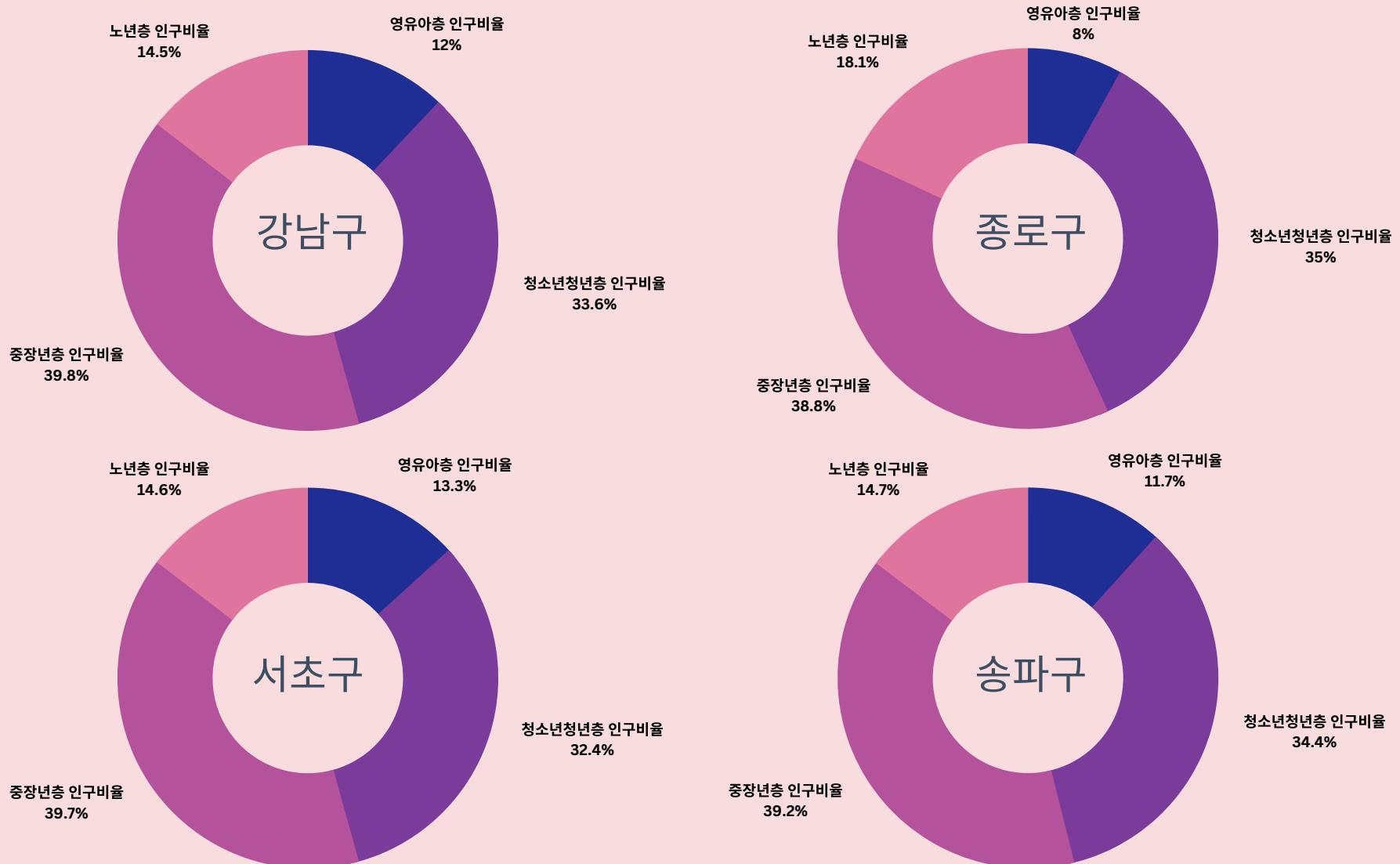
■ 공공체육시설수



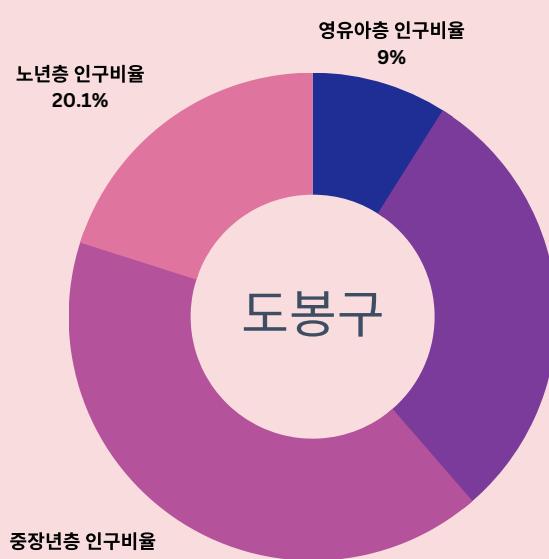
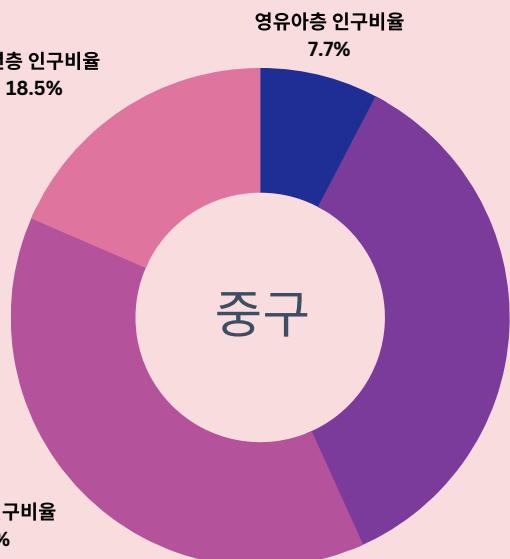
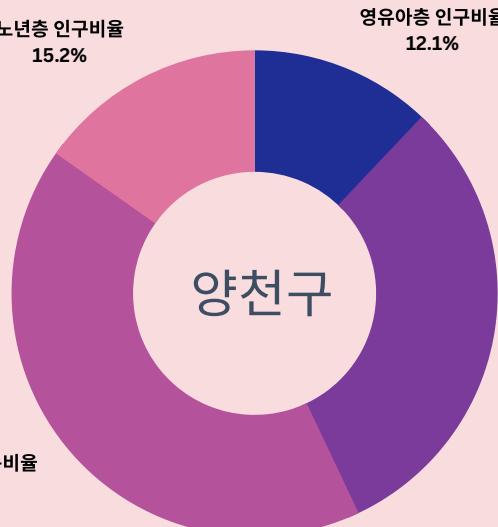
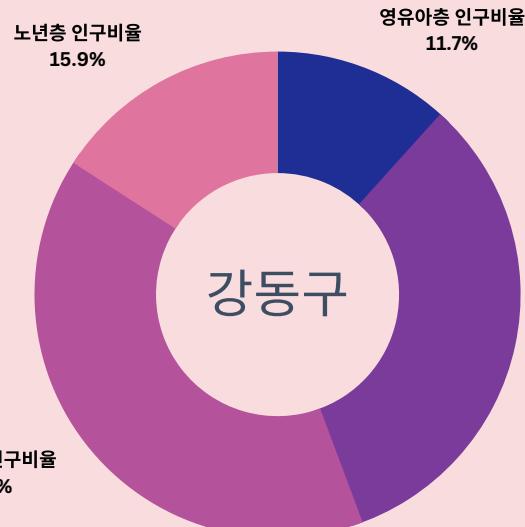
# 구별 평균 소득 - 2021



# 구별 연령대별 인구 비율



# 구별 연령대별 인구 비율



# Coding

```
● ● ●

df = pd.read_excel(f'{src}/outlier_0_1_제거 총합.xlsx')

df['mean'] = (df[2019] + df[2020] + df[2021])/3
data_x = [
    '공공체육시설수',
    '공원 총 면적',
    '녹지(총면적)',
    '약국수',
    '평균 소득',
    '병원수',
    '노년층 인구수'
]
fig = plt.figure(figsize = (15, 10))

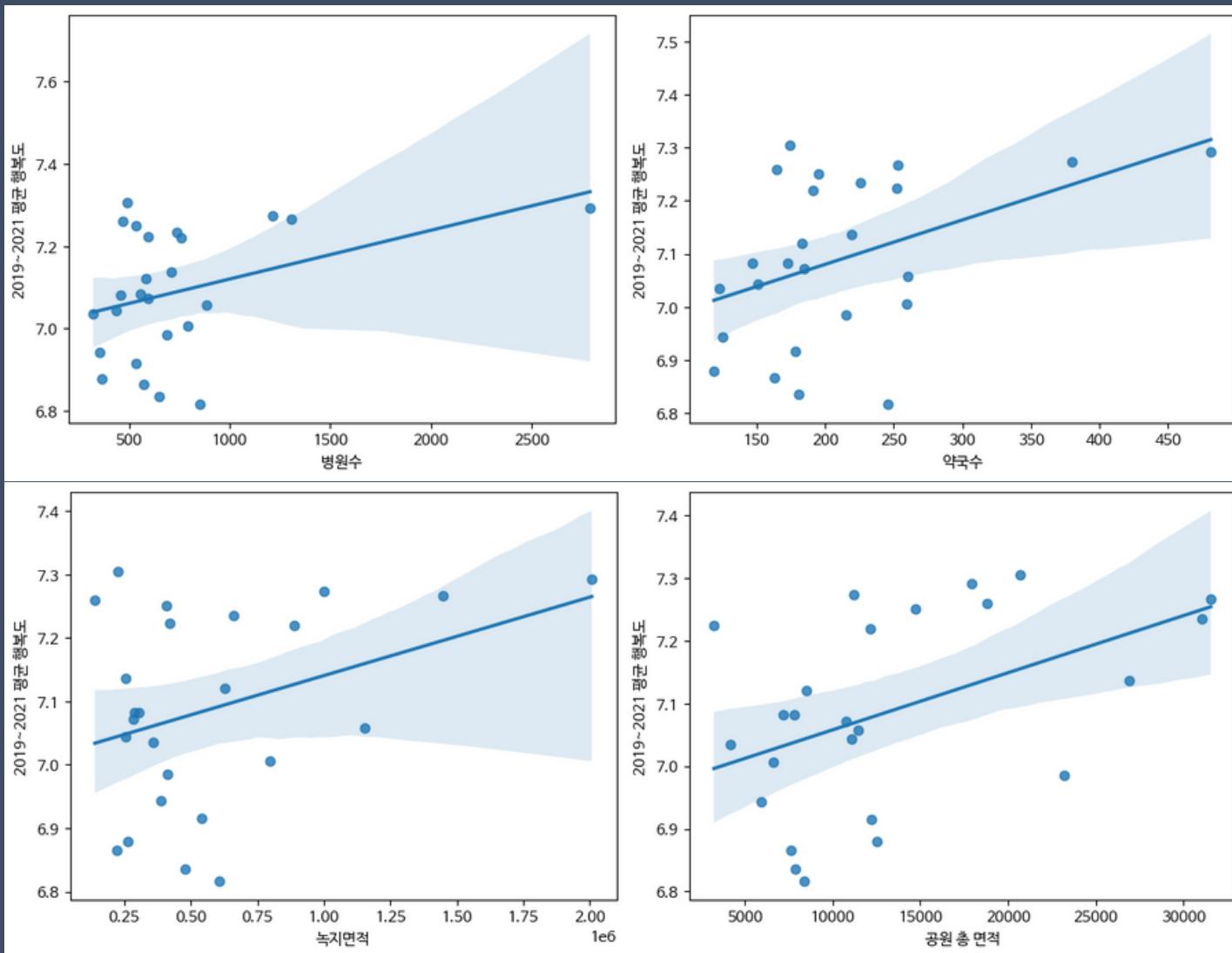
for i in range(len(data_x)):
    temp_ax = fig.add_subplot(2,4,i)
    sns.regplot(data=df, x=data_x[i], y='mean', ax=temp_ax)
plt.show()
```

## 4가지 요인과 건강 행복도 의 상관성

THE RESULTS

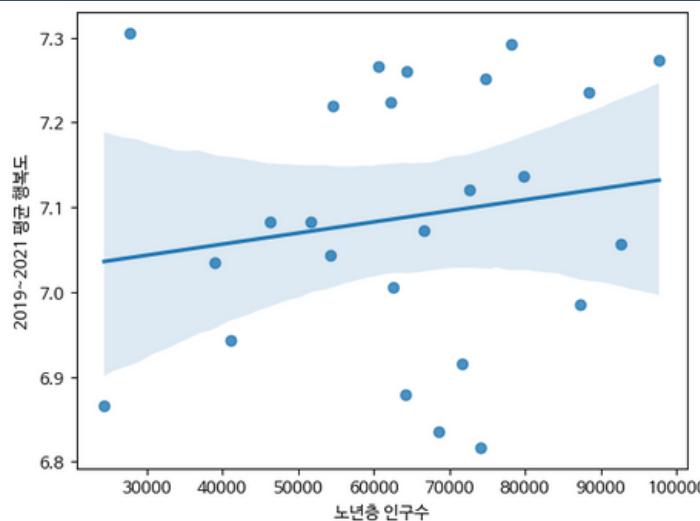
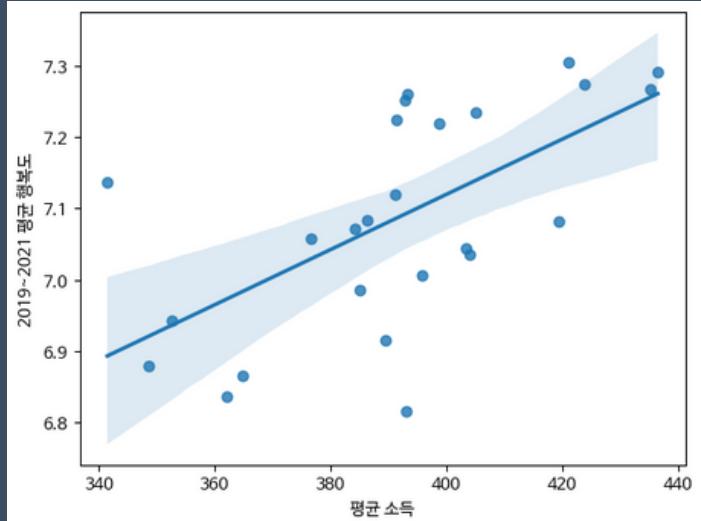
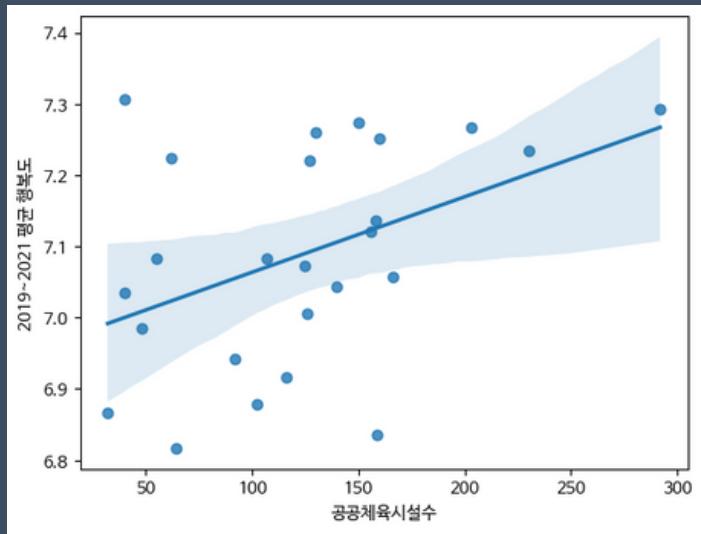
# 4가지 요인과 건강 행복도 의 상관성

## THE RESULTS



# 4가지 요인과 건강 행복도 의 상관성

## THE RESULTS



# 4가지 요인과 건강 행복도 의 상관성



```
dfs = [
    df_hospital,
    df_drugstore,
    df_greener,
    df_gym,
    df_park,
    df_population,
    df_income
]
df_total = pd.concat(dfs, axis = 1, join = 'outer')
```



```
df_total = df_total[
    df_total.columns[0],
    df_total.columns[1],
    df_total.columns[3],
    df_total.columns[5],
    df_total.columns[7],
    df_total.columns[9],
    df_total.columns[20],
    df_total.columns[22],
]
df_total['평균 행복지수'] = df_happiness['평균 행복지수']
```

THE RESULTS

# 4가지 요인과 건강 행복도의 상관성

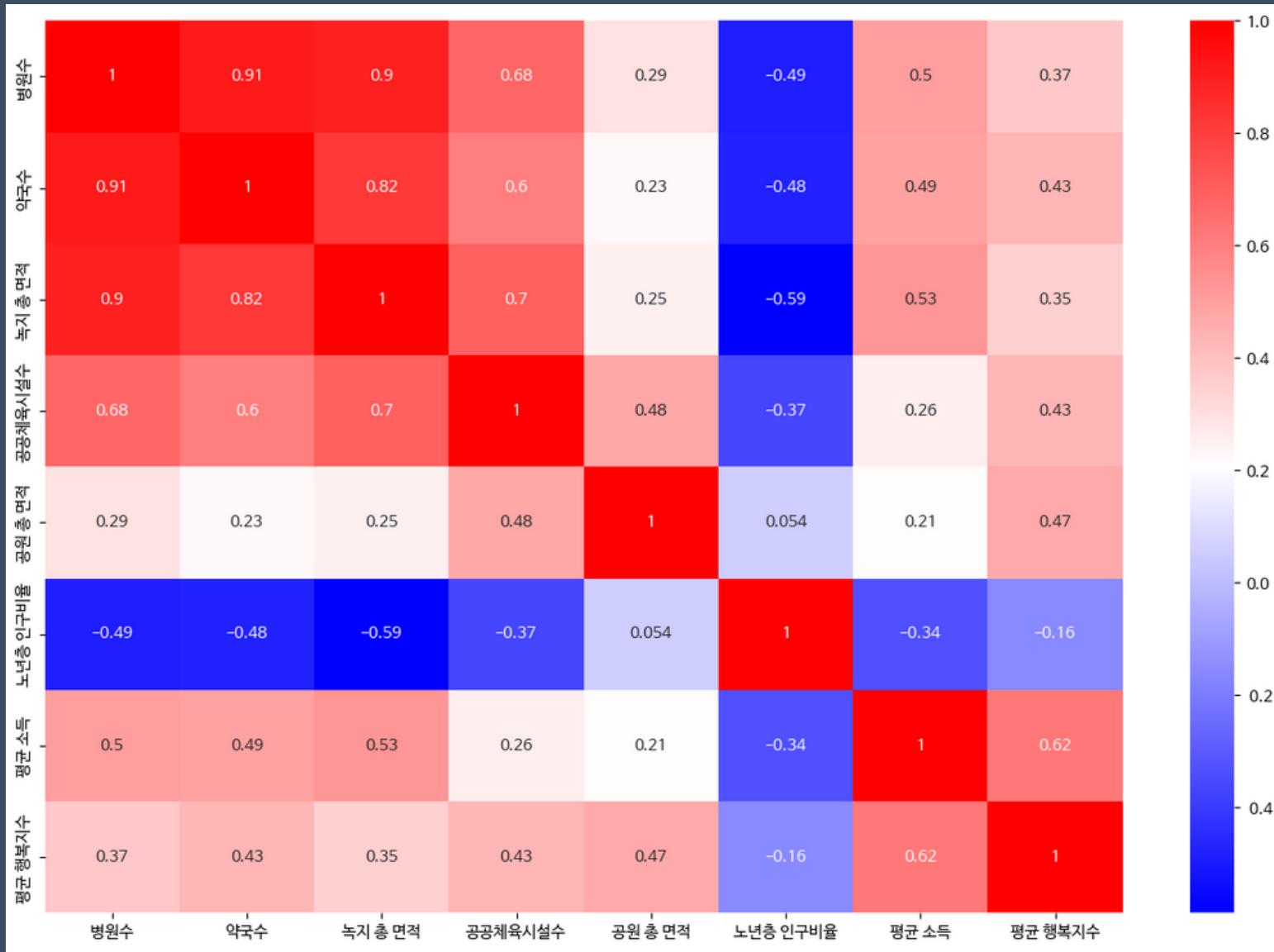
---



```
corr_matrix = df_total.corr(method='pearson')
plt.figure(figsize=(15,10))
sns.heatmap(corr_matrix, cmap = 'bwr', annot=True)
```

THE RESULTS

# THE RESULTS



# Analysis of Results

-



## 서울 SURVEY

행복지수 감소  
건강과 행복지수의 관계

## 뉴스 크롤링

코로나 언급 증가  
행복에 영향

## 상관관계

4가지의 요인 확인

CONCLUSION

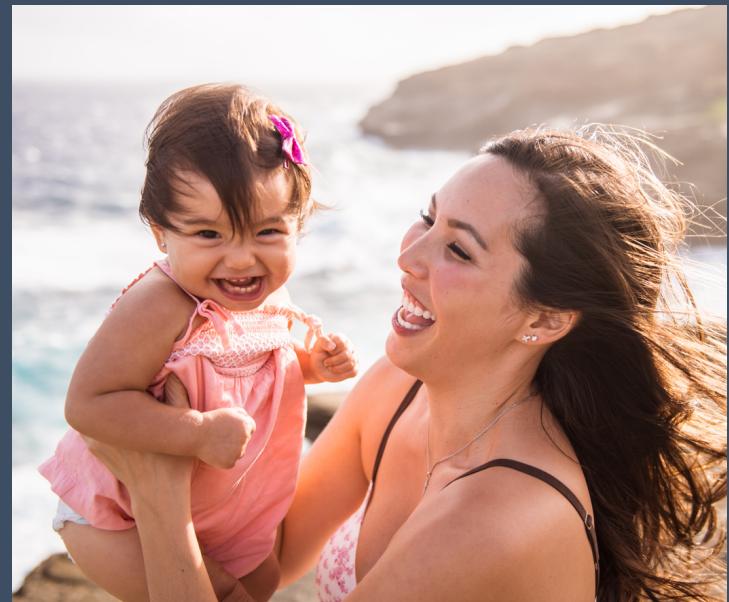
## 정부의 역할

- 자연환경시설 보완
- 의료 서비스 공급 증가
- 건강에 대한 만족도 증가 기대

## 보완이 필요한 사항

- 다양한 원인이 복합적으로 영향
- 이상치 처리
- 단순선형모델 - 인과관계 설명 필요

# Analysis of Results



# REFERENCES - 참고문헌

- 김동진, 윤흔미, 이정아, & 채희란. (2014). 의료패널자료를 활용한 우리나라의 의료이용 불평등 측정. *Health and Social Welfare Review*, 34(3), 33-58.
- 김동환. (2020). 지리공간적 접근성과 지역의료이용규모. *심사평가연구소*, 45-57.
- 김상곤, & 김성중. (2008). 지역주민의 행복도 결정요인에 관한 연구. *Journal of Community Welfare*, 25, 325-351.
- 김창엽. (2004). 사회계층과 전반적 건강수준 및 건강행태. *보건복지포럼*, 2004(6), 18-25.
- 문정화, & 강민아. (2017). 노인의 종사상 지위가 삶의 만족도에 미치는 영향과 소득수준, 건강수준의 조절효과. *Social Welfare Policy*, 44(1), 79-103.
- 장인수, & 김홍석. (2016). 서울시민들의 주관적 건강 행복지수 결정요인 분석. *Health and Social Welfare Review*, 36(3), 85-118.
- 허성호, 김종대, & 정태연. (2011). 취업이 노인의 삶에 미치는 영향 분석. *Journal of the Korean Gerontological Society*, 31(4), 1103-1118.

# THANK YOU.

감사합니다.



# Q & A

