빅데이터 프로젝트

201944015 박상원

계획

아파트 매매 분석 (분석하고자 하는 내용)

- 지역별 거래된 매매가 분석
- 층 수에 따른 매매가 관계 분석
- 출생률과 아파트 매매가 상관관계
- 아파트 종류별 매매가 분석
- 거래 횟수와 매매가 상관관계 분석

개발 환경

- 데이터는 공공 데이터 포털의 오픈 API를 통해 아파트 매매 실 거래 상세 자료를 통해 거래 내역들을 사용
- 출생률 데이터는 **출처 : 통계청,「인구동향조사」** 에서 행정구역(서울) 합계출산율 데이터 사용
- 모든 데이터는 2015~2021년의 데이터를 기준으로 수집
- Python 사용
    - lib: JSON, requests, XMLTODICT, pandas, matplotlib, seaborn이용

전 처리, 데이터 변환 과정

```python
# api 사용을 위한 인증키 입력
serviceKey = "내 키는 비밀"
url ="http://openapi.molit.go.kr:8081/내 키는 비밀"


1 usage
def get_df(lawd_cd, deal_ym): #lawd_cd= 법정동 코드 / deal_ym = 계
    global serviceKey
    base_url = "http://openapi.molit.go.kr:8081/OpenAPI_ToolInsta
    base_url += f'&LAWD_CD={lawd_cd}'
    base_url += f'&DEAL_YMD={deal_ym}'

    res = requests.get(base_url)
    data = json.loads(json.dumps(xmltodict.parse(res.text)))
    df = pd.DataFrame(data['response']['body']['items']['item'])
    drop_df(df)
    return df
```

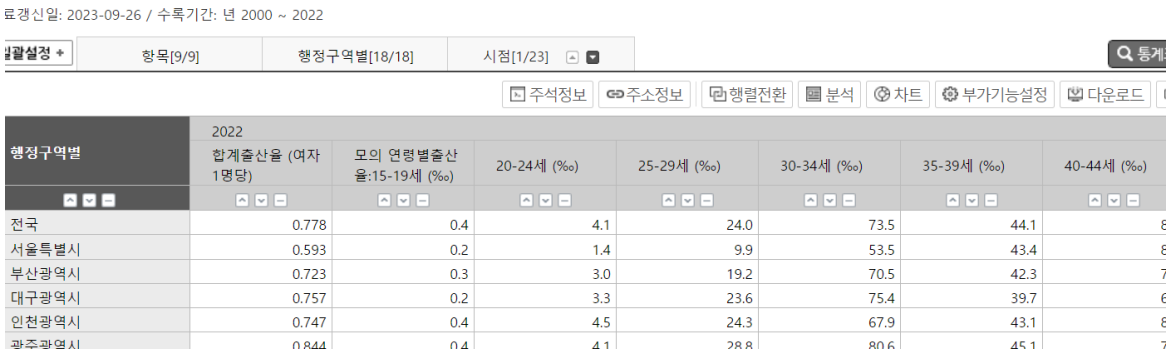
먼저 공공 데이터 포털의 **오픈 API**를 통해 데이터를 가져와서 내가 필요한 데이터

법정 동 코드, 매매가, 년, 월, 아파트명 등을 **data frame 형태로 변환**

이후 API 사용 **트래픽 제한**이 있어서

```python
for year in range(2015,2022):
    for month in range(1,13):
        ym = int(f'{year}{month:02d}')
        df = get_df( lawd_cd: 42750,ym)
        merge_df = pd.concat( objs: [merge_df,df],ignore_index = True)
        merge_df.to_csv( path_or_buf: '강릉시.csv', index=False)
```

필요한 **data들을 csv파일로 만들어서** 해당 csv 파일들로 프로젝트 진행



행정구역별	2022 합계출산율 (여자 1명당)	모의 연령별출산율:15-19세 (‰)	20-24세 (‰)	25-29세 (‰)	30-34세 (‰)	35-39세 (‰)	40-44세 (‰)
전국	0.778	0.4	4.1	24.0	73.5	44.1	8
서울특별시	0.593	0.2	1.4	9.9	53.5	43.4	8
부산광역시	0.723	0.3	3.0	19.2	70.5	42.3	7
대구광역시	0.757	0.2	3.3	23.6	75.4	39.7	6
인천광역시	0.747	0.4	4.5	24.3	67.9	43.1	8
광주광역시	0.844	0.4	4.1	28.8	80.6	45.1	7

←

출생률 데이터는 앞서 말했던 출처에서 해당 표처럼 생긴 **csv파일을 다운받아 data frame** 형태로 변환해서 사용

전 처리, 데이터 변환 과정

```
df = pd.read_csv('.csv')

# '거래금액'을 숫자로 변환
df['거래금액'] = df['거래금액'].str.replace(',', '').astype(int)
print(df)
# '년'과 '월'을 합쳐서 새로운 열 '년월' 생성
df['년월'] = df['년'].astype(str) + '-' + df['월'].astype(str)

# '년월'로 그룹화하여 거래금액의 평균 계산
monthly_avg = df.groupby('년월')['거래금액'].mean().reset_index()

# 시각화
plt.figure(figsize=(12, 6))
plt.plot( *args: monthly_avg['년월'], monthly_avg['거래금액'], marker='o')
plt.title('Average Transaction Amount by Month (2015-2021)')
plt.xlabel('Year-Month')
plt.ylabel('Transaction Amount (Average)')
plt.xticks(rotation=45)
plt.show()
```

월별로 15년도부터 21년도까지 매매가의 평균을 구해야 했기 때문에

년 열과 월 열을 그룹화한 뒤 매매가의 평균을 계산해서 시각화

간단한 시각화들은 **이처럼 모두 진행**

상관계수를 나타내야 할 때에는

```
# 'yearly_avg'와 'df'를 '년'을 기준으로 합침
result_table = pd.merge(yearly_avg, df, on='년')
```

공통으로 사용하는 데이터를 기준으로 merge를 시킨 뒤에

```
plt.subplot( *args: 2, 2, 4)
plt.scatter(result_table['거래금액'], result_table['값'])
plt.title('Scatter Plot: Transaction Amount vs. Birth Rate')
plt.xlabel('Transaction Amount')
plt.ylabel('Birth Rate')

plt.tight_layout()
plt.show()
```

Scatter를 사용해서 시각화

전 처리, 데이터 변환 과정

년도별로 층 별 매매가 데이터를 만들 때에는

```
files = [file for file in os.listdir() if file.endswith('*.csv')]

# 파일들을 읽어와서 데이터를 합치기
merged_data = pd.concat([pd.read_csv(file) for file in files], ignore_index=True)

merged_data['거래금액'] = merged_data['거래금액'].str.replace(',', '').astype(int)
```

일단 행정 동 별로 나눠진 데이터를 모두 하나로 합친 뒤에

```
# '층' 값을 15층 미만, 15~30층, 30층 초과로 나누기
merged_data['층 구간'] = pd.cut(merged_data['층'], bins=[-float('inf'), 15, 30, float('inf')],
                            labels=['15층 미만', '15~30층', '30층 초과'])
```

기준치 (~15, 15~30, 30~) 에 맞춰 데이터를 나눠서

```
# 년도별 층 구간별 평균 거래금액 계산
avg_price_by_year_floor_range = merged_data.groupby( by: ['년', '층 구간'], observed=False)['거래금액'].mean().reset_index()
```

층 구간 별로 그룹을 만들어 mean을 통해 평균 매매가를 계산한 뒤

```
plt.figure(figsize=(15, 10))
for i, year in enumerate(avg_price_by_year_floor_range['년'].unique(), 1):
    plt.subplot( *args: 3, 3, i)

    # 해당 년도의 데이터 선택
    year_data = avg_price_by_year_floor_range[avg_price_by_year_floor_range['년'] == year]

    # 막대 그래프 생성
    sns.barplot(x='층 구간', y='거래금액', data=year_data)

    # 서브플롯 제목 설정
    plt.title(f'년도: {year}')
```

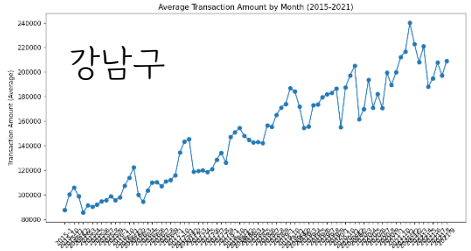
각 년도별로 subplot을 생성해 전체 년도의 데이터를 한번에 추출

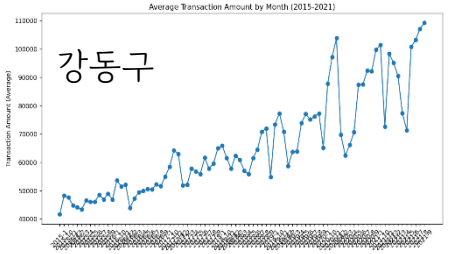
그 외 나머지 pie chart나 막대 그래프 등
간단한 시각화들은 matplotlib 를 통해 제작

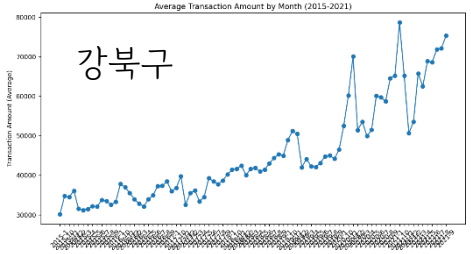
시각화

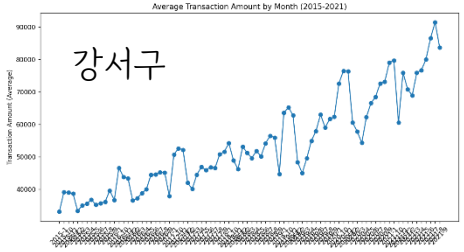
지역별 거래된 매매가 분석

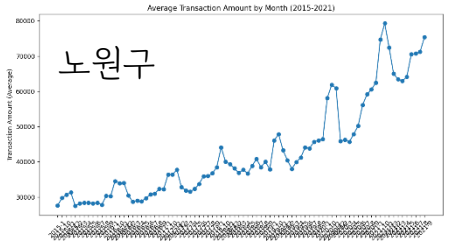
서울

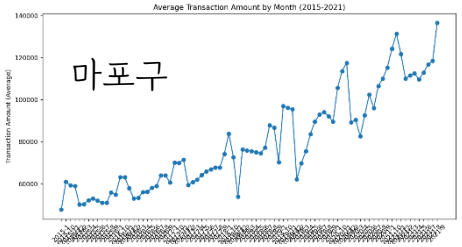










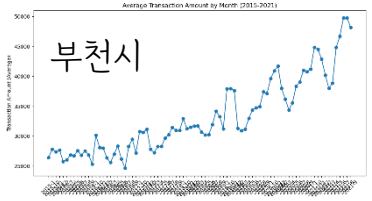


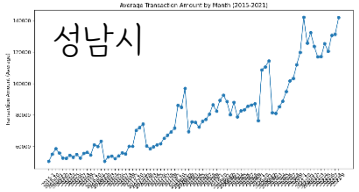
서울, **경기권**의 특정 지역 데이터를 토대로 **월별 거래 금액**의 **평균** 그래프

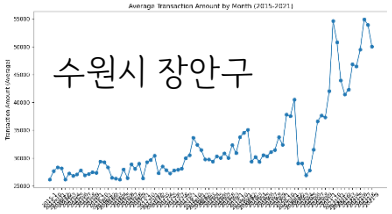
각 구마다 급상승 하는 년도가 다르지만, 전반적으로

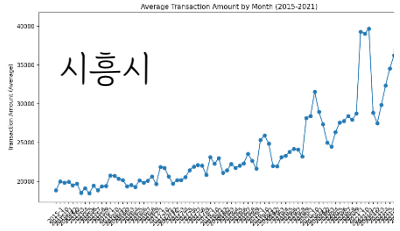
서울, 경기권은 모두 시간이 지날수록 가파르게 증가하는 **우 상향 그래프** 형태

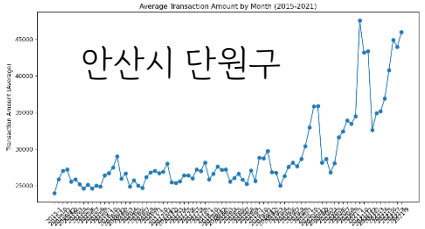
경기



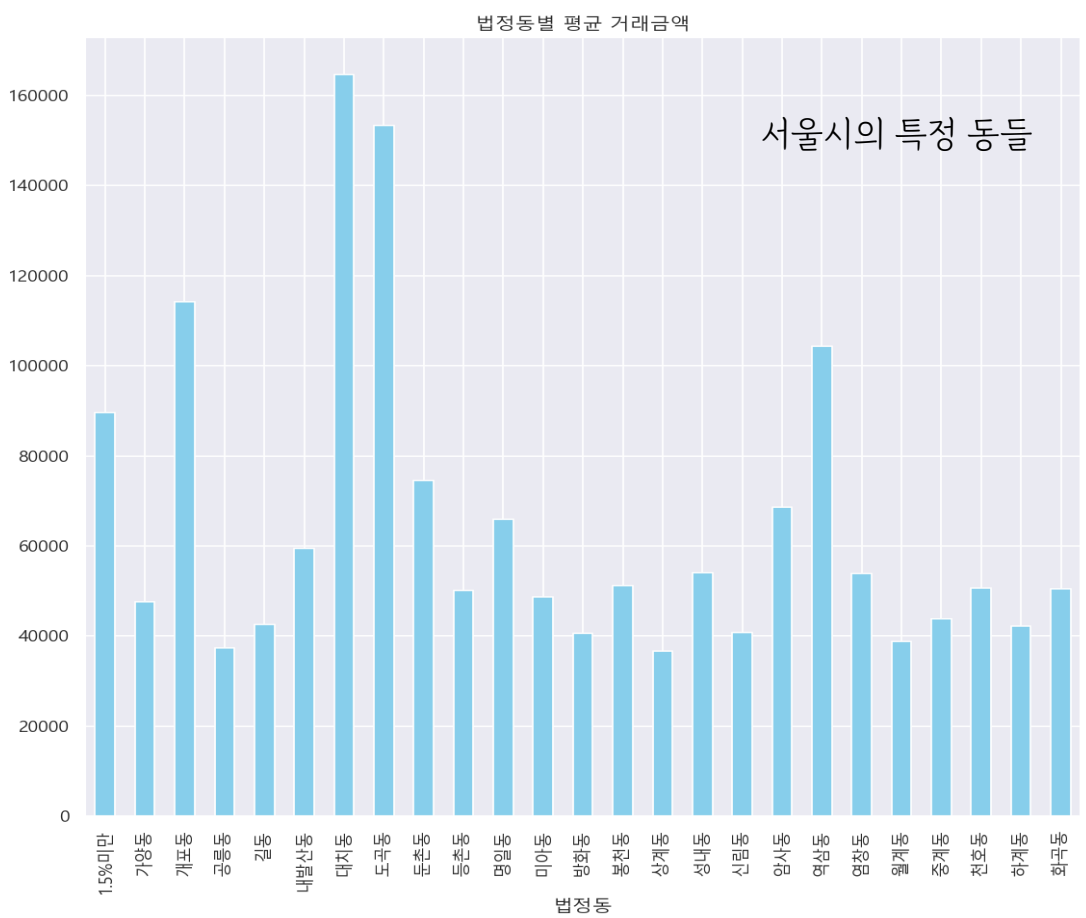


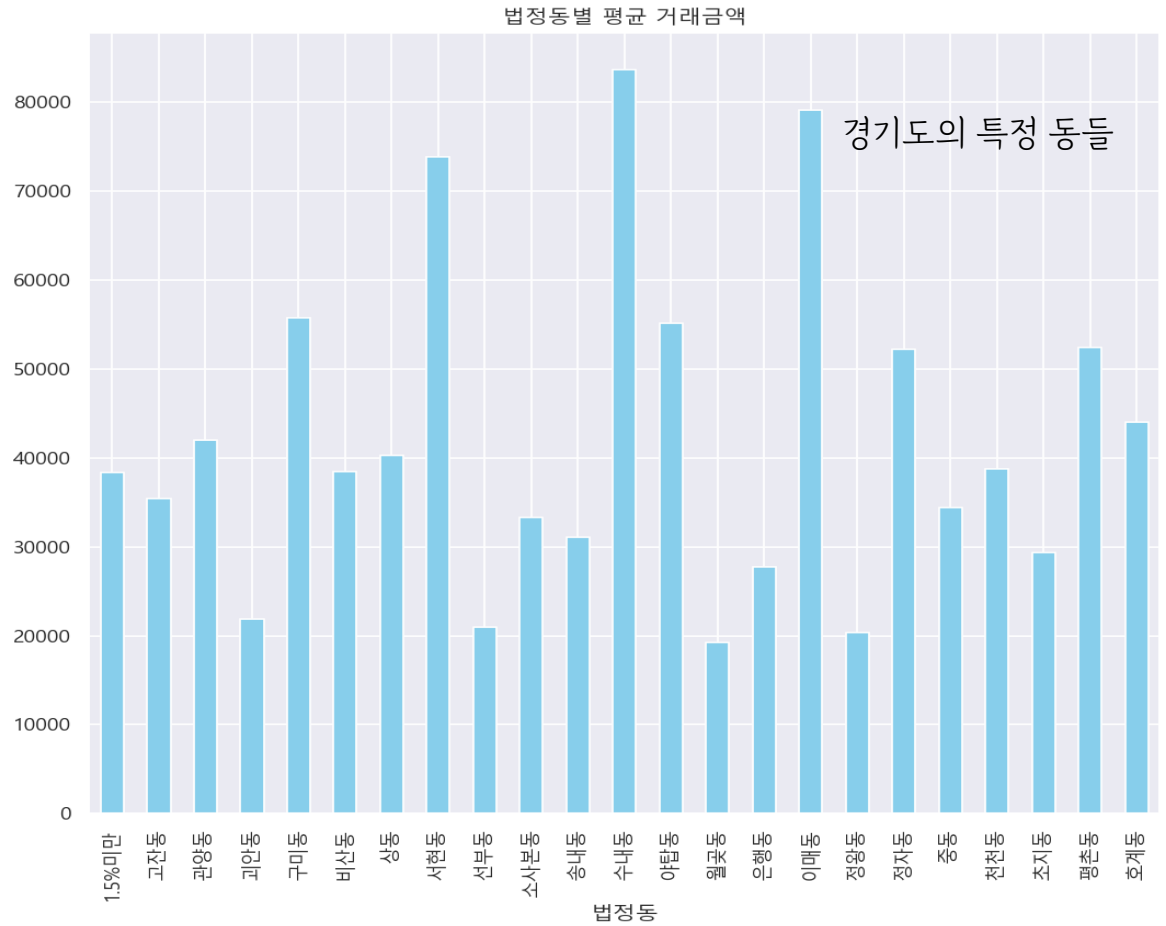






지역별 거래된 매매가 분석





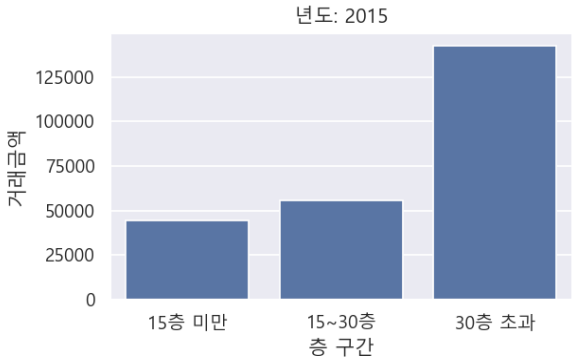
서울, **경기권**의 특정 동의 매매가에 대한 그래프

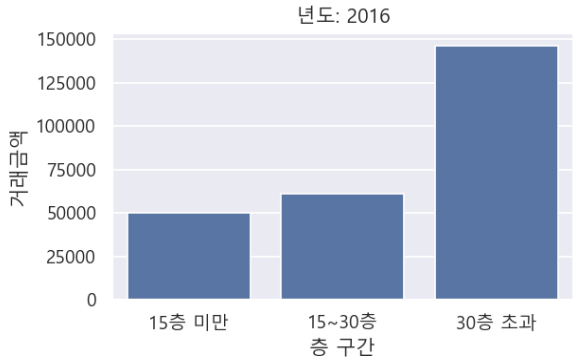
평균 매매가를 비교하면 기본적으로 서울권이 비싼 가격대에 형성돼 있으며 최대 **2배 이상 차이**나는 동도 존재

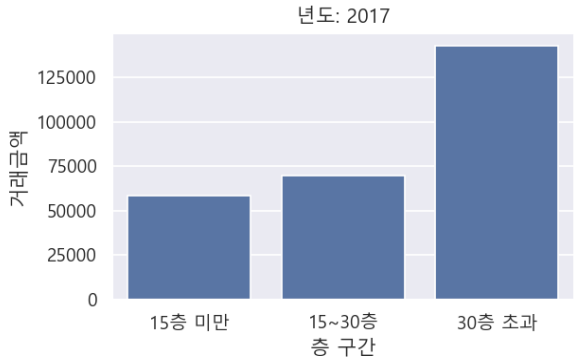
심지어 같은 시, 도 내에서도 **동에 따른 가격 차이의 폭이 크다**는 것을 알 수 있다.

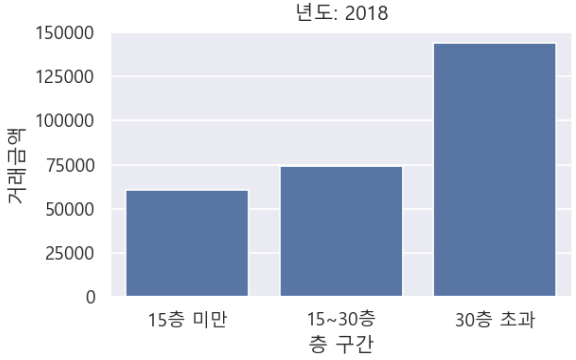
즉 특정 동의 **인프라에 따라** 가격이 **다르다**는 것을 알 수 있다.

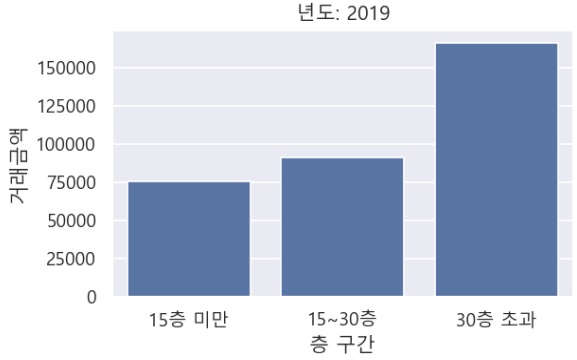
층 수에 따른 매매가 분석

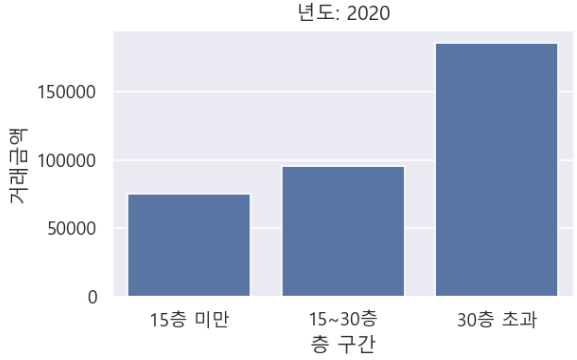


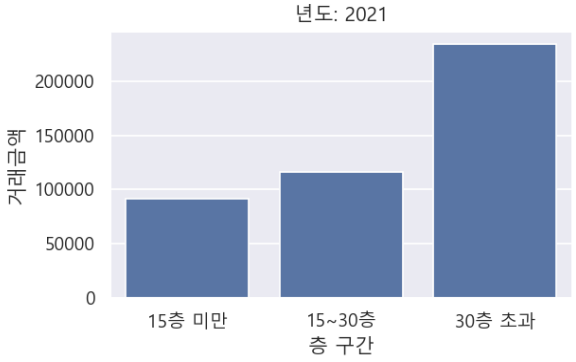












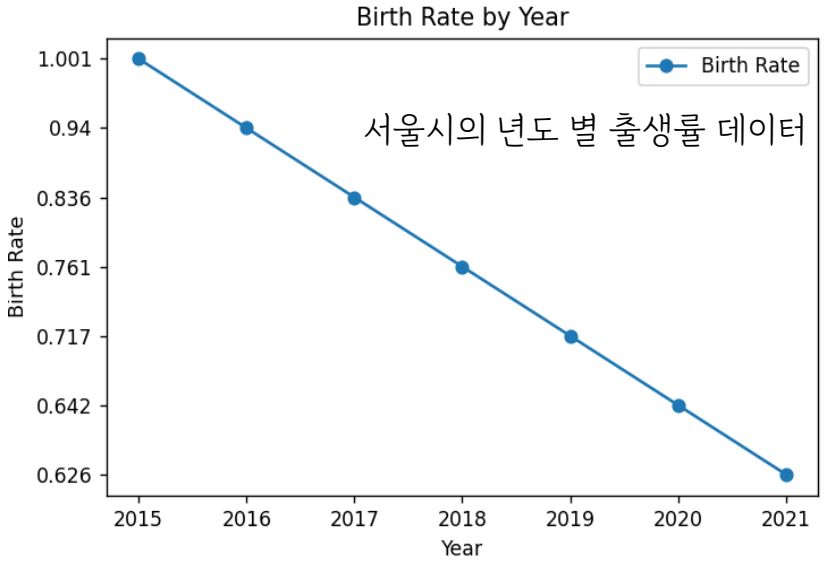
고층은 **흔들리는 요소**, **외출의 불편함** 때문에 가격대가 **낮게 형성될 것이라 예상**

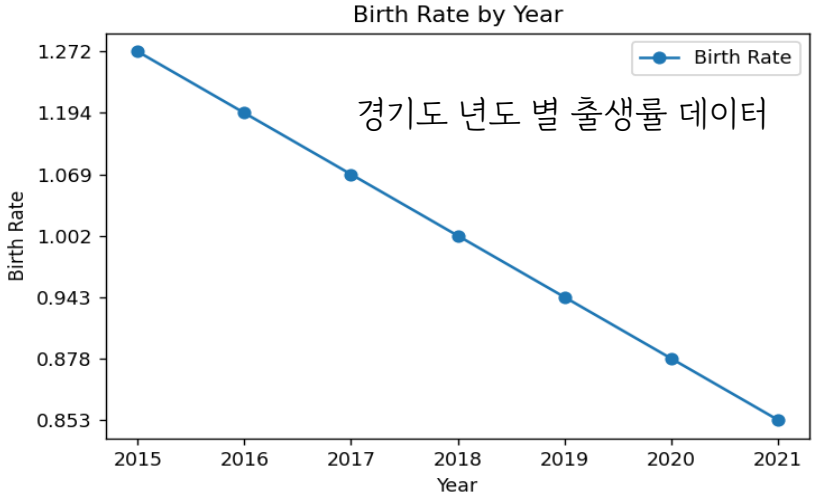
아파트 별로 층수가 모두 다르기 때문에 15~30층 구간을 기점으로
저층, 중간층, 고층으로 나누어서 평균 매매가를 분석

예상과 달리 년도 상관 없이 **모두 고층이 더 높은 매매가에 형성**

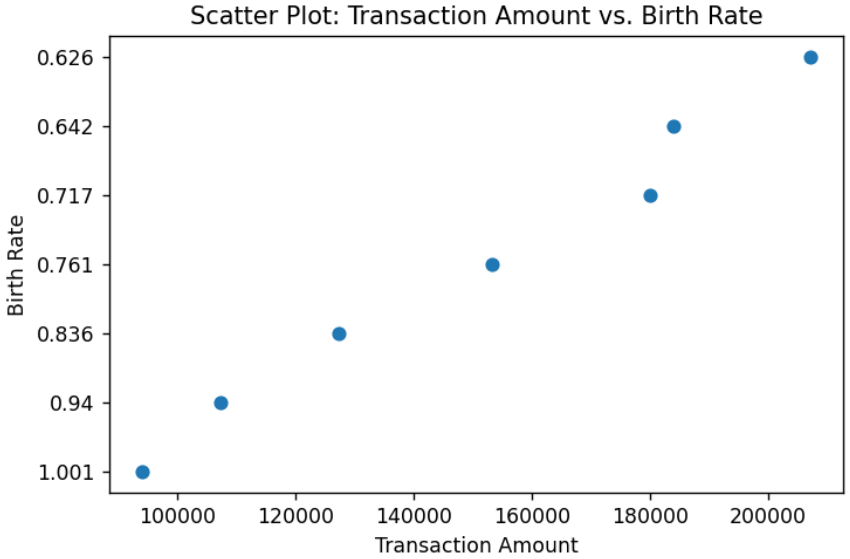
저층보다 고층이 선호되는 이유로는 여러가지가 있겠지만 해가 잘 드는 것이
가장 큰 이유로 예상된다.

출생률과 아파트 매매가 관계



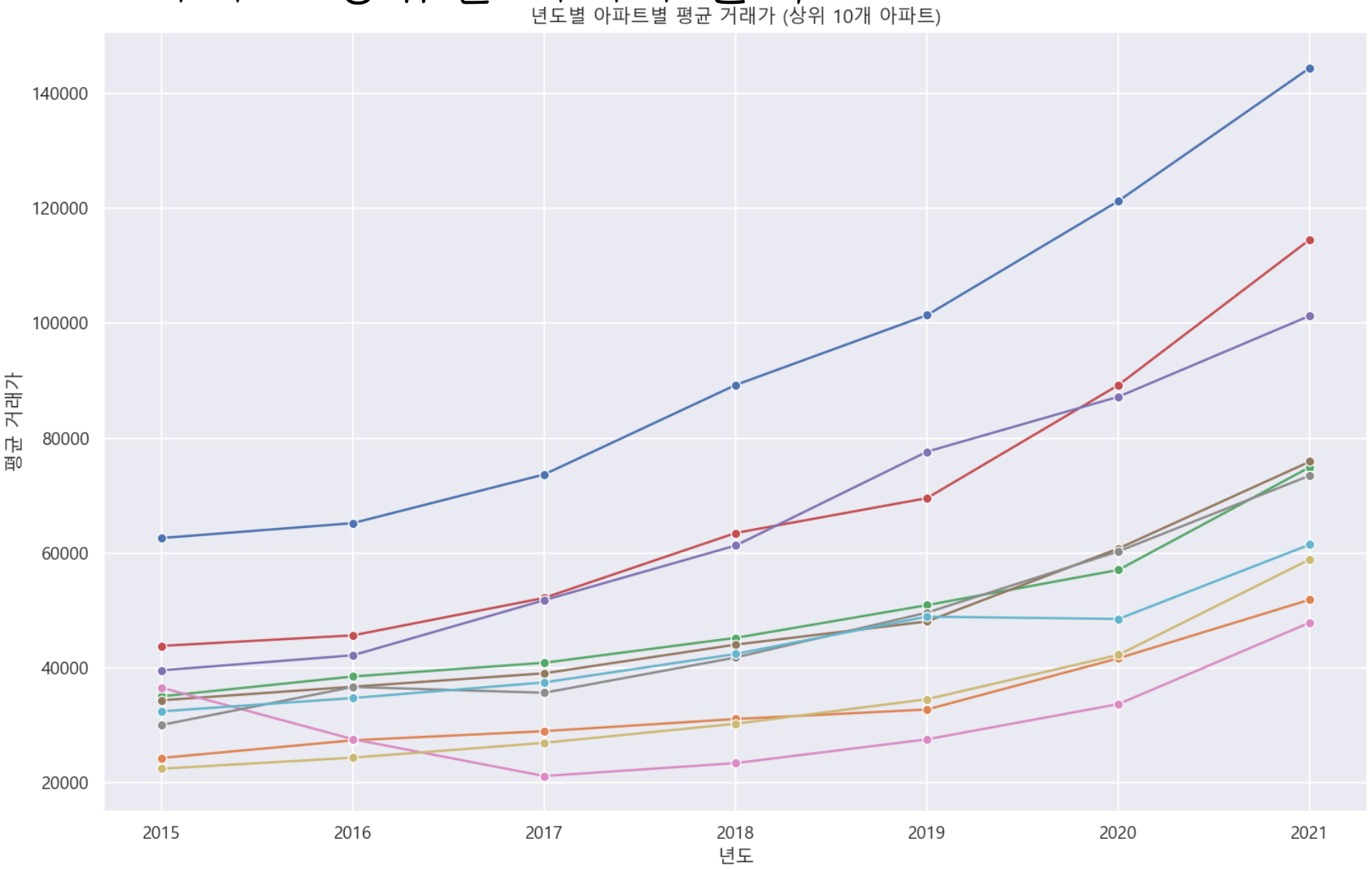


서울의 년도 별 출생률 데이터를 보면 점점 출생률이 줄어듦 **우 하향 그래프**
이를 서울의 년도 별 거래금액과 비교를 해보았을 때



거래금액은 증가하지만 출산율은 줄어드는 형태
출생률이 적어지면 매매가 즉 아파트 **가격대가 낮아질 것이라 예상**했지만
예상과는 다르게 출생률이 낮아짐과는 관계 없이 서울의 **매매 가격은 증가**
경기도 또한 출생률은 낮아지지만 앞의 월별 거래금액 평균치는 증가
앞으로 인구수는 적어지지만 서울, 경기의 아파트 매매가는 **떨어진다 할 수 없음**

아파트 종류별 매매가 분석

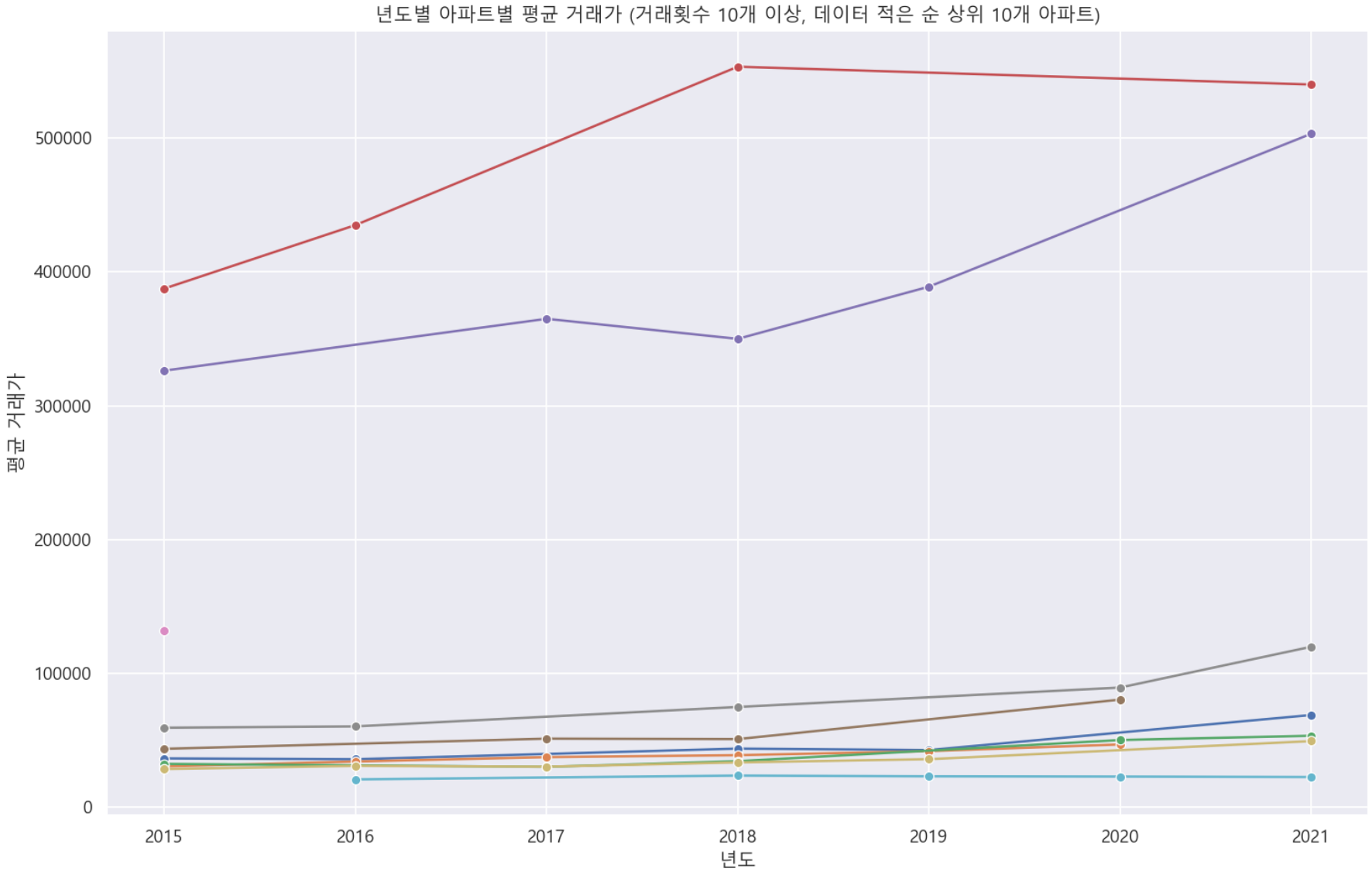


아파트 종류별 매매가 분석을 위해
거래내역이 많은 아파트들로 선별

평균 거래가를 축으로
년도 별 데이터를 시각화 했을 때

기본적으로 **아파트의 종류**마다 제각각
다른 가격대에 형성 돼 있다.

아파트 종류별 매매가 분석



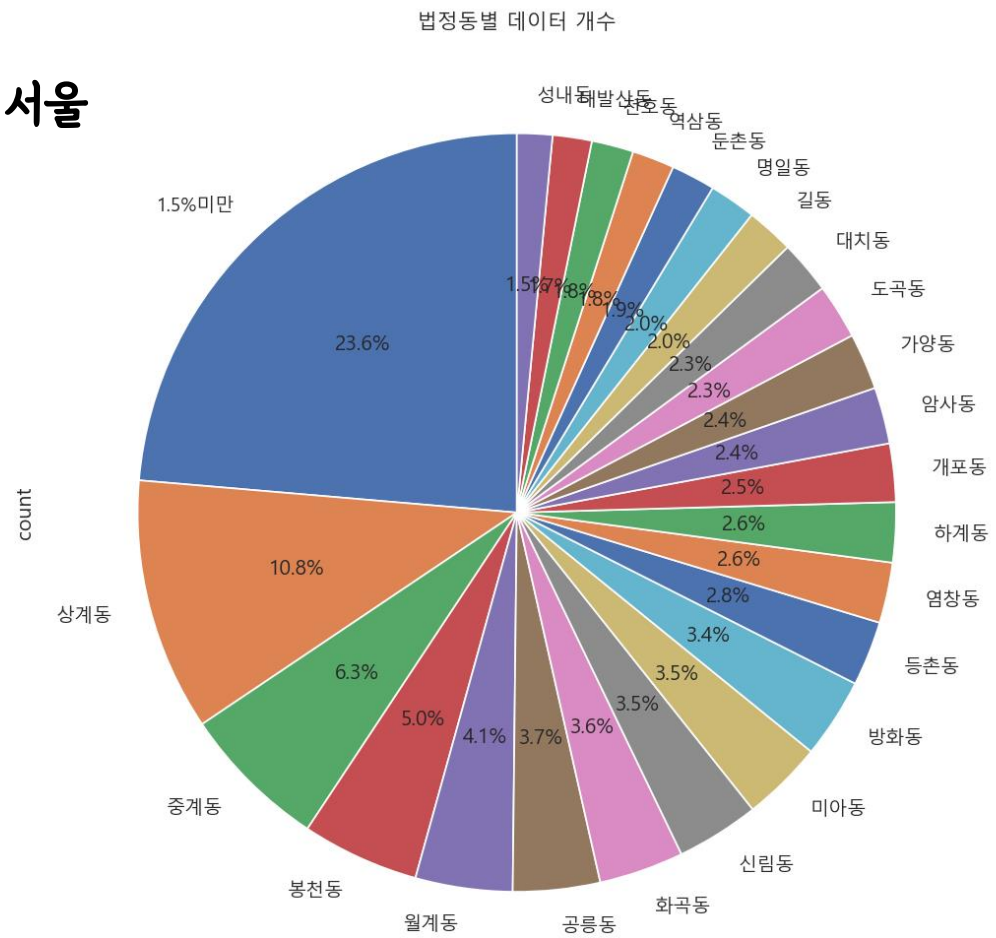
매매 데이터가 적은 아파트
거래량이 많지 않기 때문에 더욱 직관적이다.

거래량이 적은 아파트들은
비싼 가격대에 형성이 돼 있고
가격이 떨어지지 않는다.

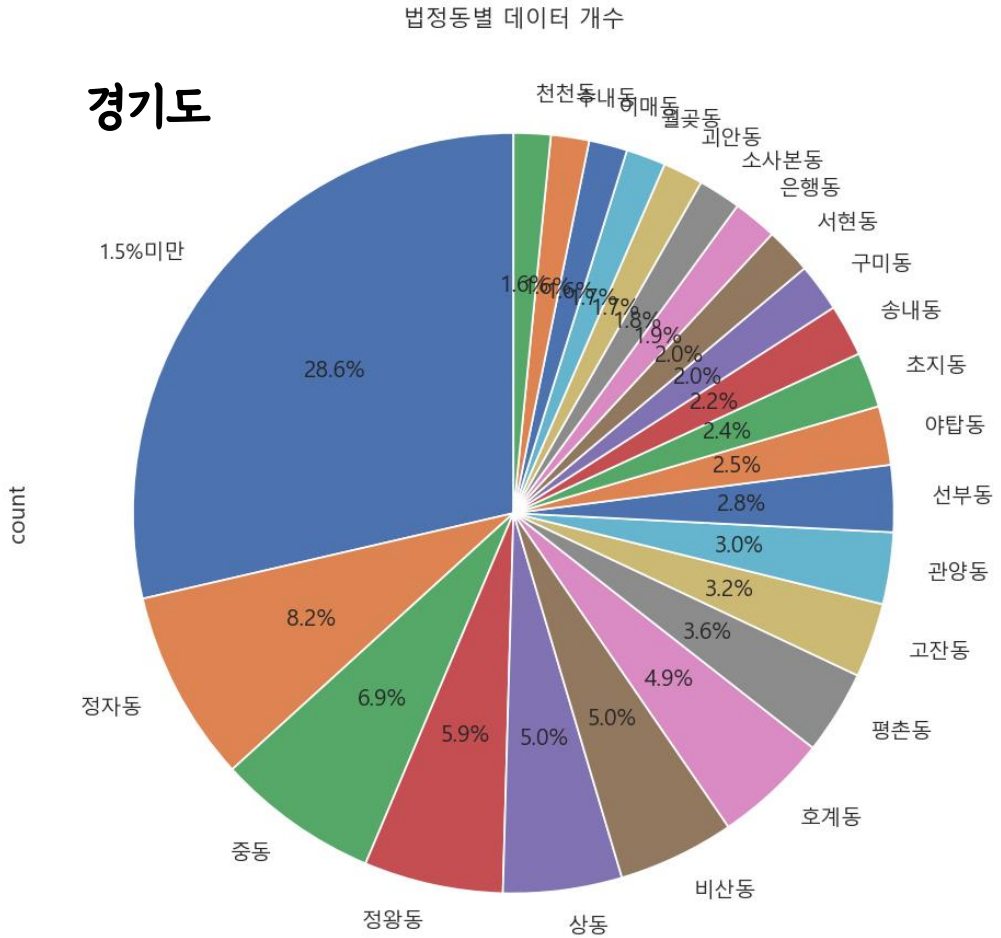
앞의 표와 현재 표를 비교해 보았을 때
다른 가격대에 형성돼 있다는 것은
아파트마다 **등급이 있는 것으로 보인다**.

거래횟수와 매매가 분석

서울



경기도

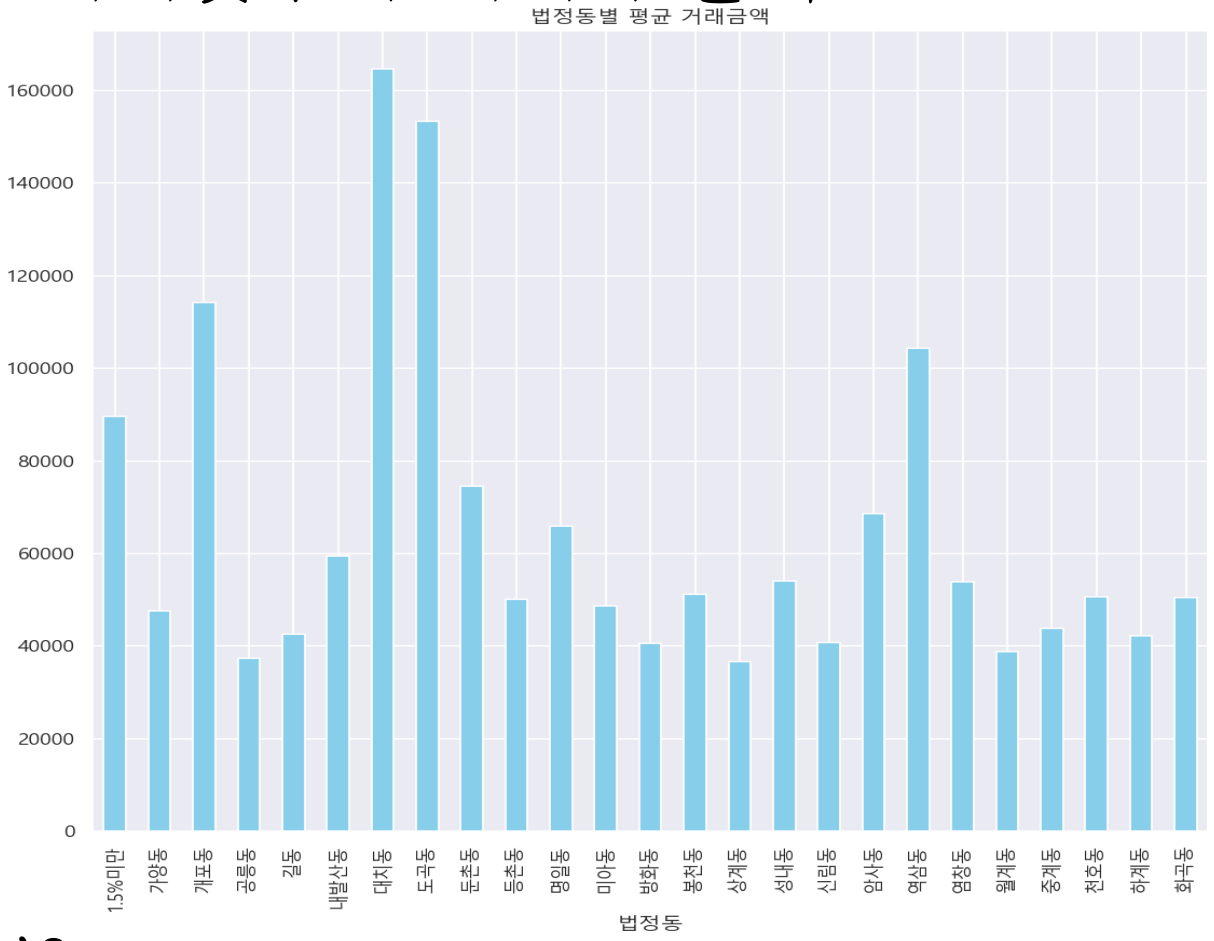


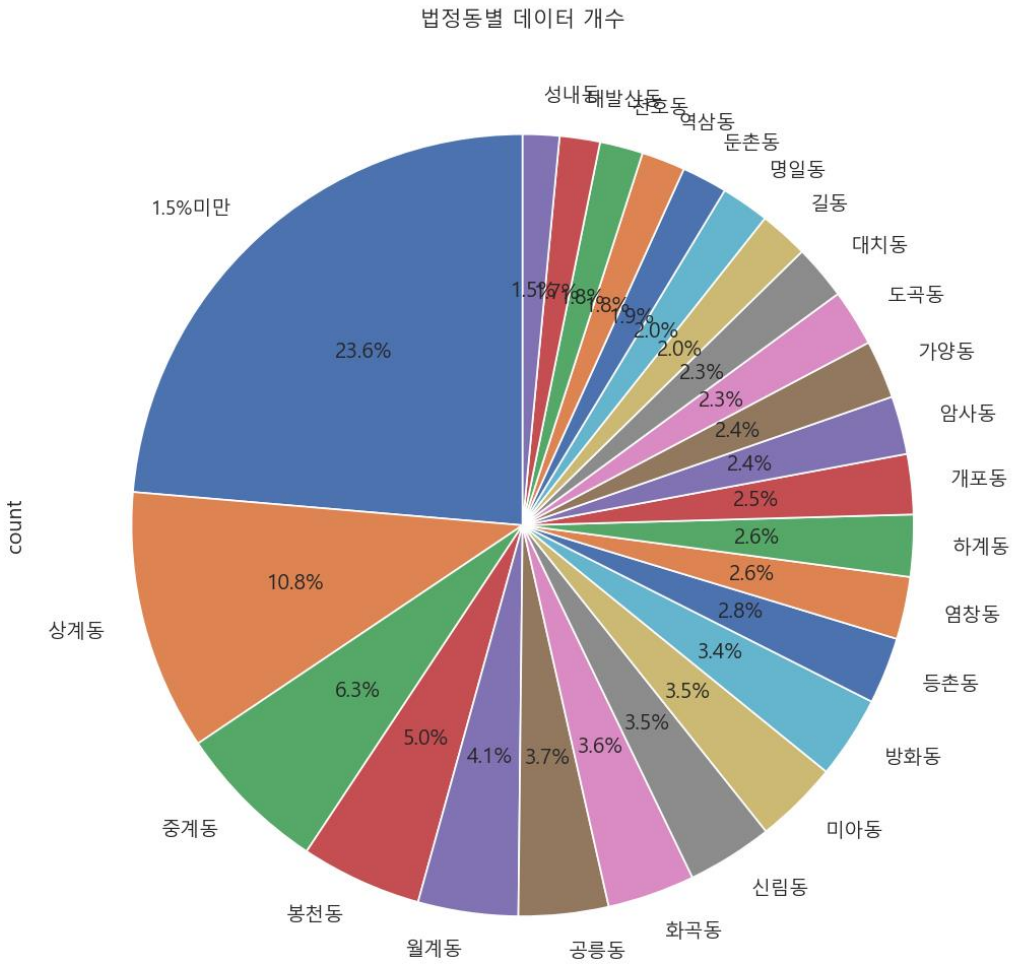
서울, **경기권**의 특정 동의 거래 횟수의 분포

거래 횟수가 1.5% 미만인 동들이 두 곳 모두 20%~30%대를 구성하고 있으며 1.5%~2.5%까지가 대부분을 이루고 있다.

그럼에도 특정 인기있는 동들은 거래 횟수가 전체의 5%를 넘기는 곳도 있다.

거래횟수와 매매가 분석





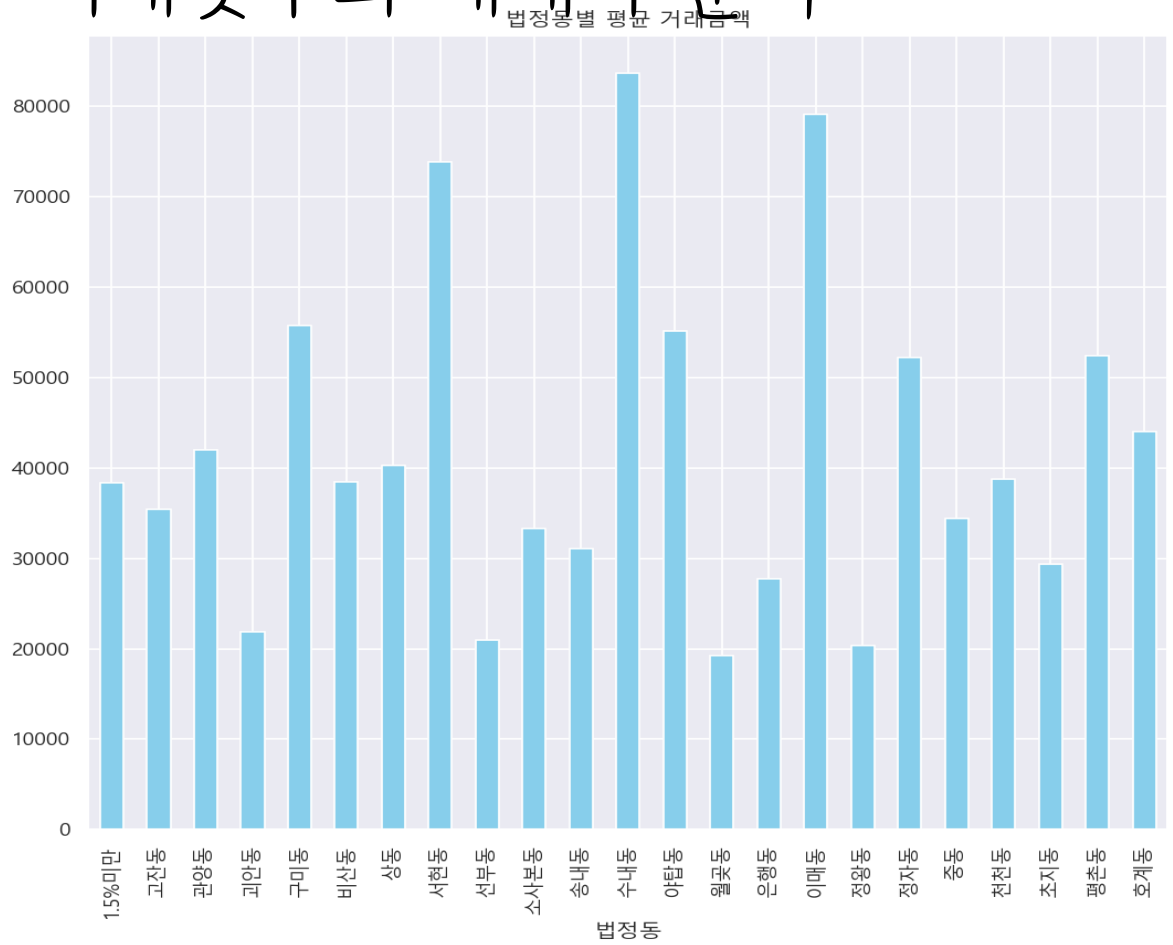
서울

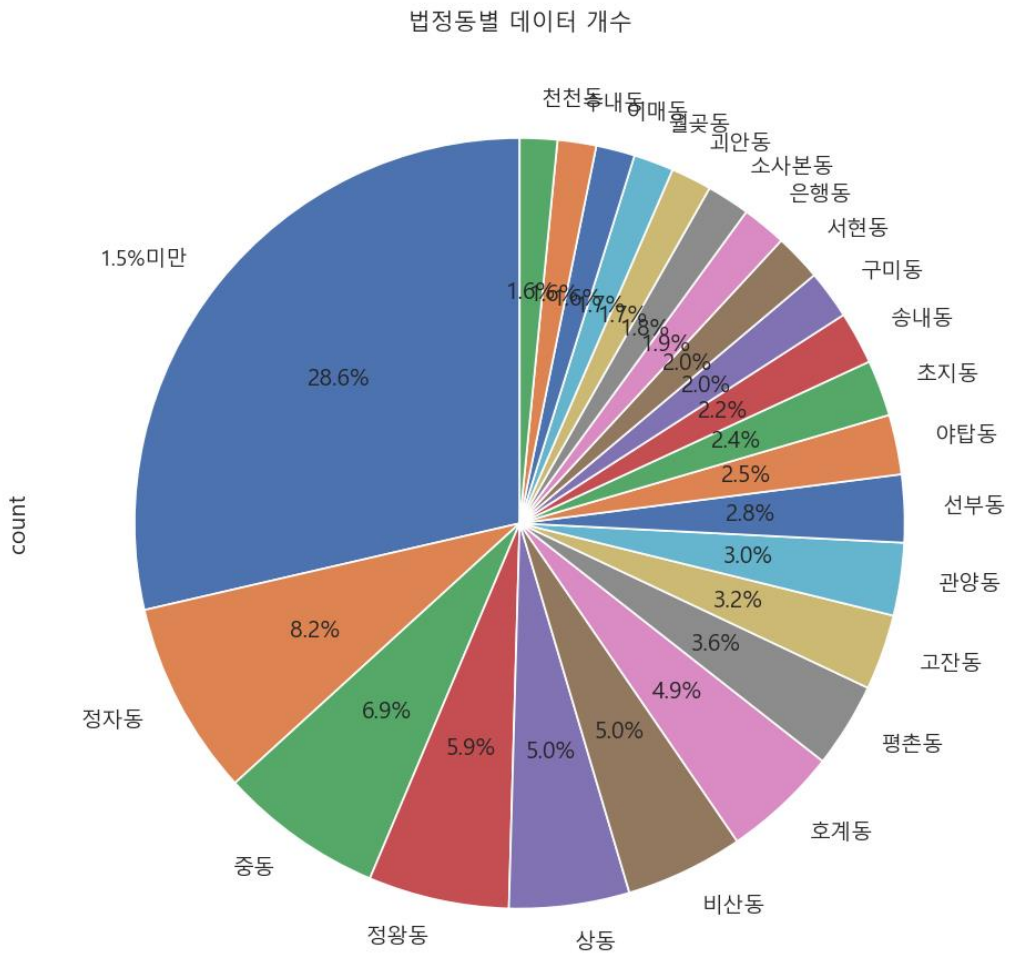
거래 횟수가 많다면 인기가 좋은 매물이고 **매매가도 높게 형성**될 것이라 **예상**

하지만 보이는 것과 같이 **거래 횟수가 전체의 10%이상을 차지하는 상계동**의 평균 거래금액을 보면 **중간보다도 낮은 가격대에 매매가가 형성**

중계동 또한 상계동 다음으로 횟수가 많지만 평균 매매가가 상위권에 차지하지 못함

거래횟수와 매매가 분석





경기도

경기도에서는 **가장 많은 거래횟수 분포를 가진 정자동이** 평균 매매가 중 **높은 가격대에** 형성

하지만 **두번째로 많은 중동은** 평균 매매가중 **높지도 낮지도 않은 가격대이며** 다음 동들 또한 **일정하지 않은 가격대에 형성**

거래 횟수에 따라 매매가가 형성되는 것이 일정하지 않으며 서로 비례적인 관계는 아니지만 연관되어 있을 수는 있다.

결론

아파트 매매가가 어떤 관계들을 통해 형성이 되는지 시각화 한 뒤 직접 분석해 보았는데,
본문에서 무엇을 어떻게 예상하며 데이터를 가공하고 또 그 값이 어떻게 나왔는지 모두 정리해 놓았다.

결론적으로는 아파트의 **매매가는 단순하게 어느 관점에 따라 비례 혹은 반비례적으로 형성되지 않는다.**
여기서 다루지 않았던 다른 요인들에 영향을 많이 받는 주제라고 생각한다.
그럼에도 시각화를 통한 데이터 분석으로 유의미한 결과도 분명 얻었다고 생각한다.

THANK YOU

박상원이었습니다.

https://github.com/qarksangwon/BigDataProject