

## 직방 아파트 매매 데이터베이스 Exploratory Data Analysis (EDA)

Zigbang Apartment Sale Information Database Exploratory Data Analysis (EDA)

By 강호중 (Ho Jong Kang)

탐색적 데이터 분석(EDA: Exploratory Data Analysis)은 본격적인 모델링에 들어가기 앞서 선형되어야 하는 과정입니다. 데이터의 분포, 변수간 관계를 파악하기 위해 히스토그램, 산점도 등 다양한 시각화 방법이 동원됩니다.

이에 직방 데이터 베이스에 있는 국토 교통부의 아파트 매매 실거래가 데이터를 바탕으로 탐색적 데이터 분석을 아래와 같이 실행했습니다.

### 데이터베이스 연결 및 필요한 패키지 불러오기

Connecting to Database and Importing Packages Needed

```
In [1]: %run include/header.ipynb
```

Number of CPU Count = 8

```
In [2]: import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set_style("darkgrid", {"font.sans-serif": ['AppleMyungjo', 'Arial']})
```

### 데이터베이스에 있는 4개의 테이블을 Join한 후 필요한 Column들만 Select

Joining 4 Tables in Database and Selecting Columns Needed

직방 데이터 베이스에 있는 4개의 테이블을 합쳐 필요한 정보만 선택한 것입니다.

```
In [3]: all_join = """
SELECT
A.id,
A.area_danji_id,
A.총,
A.거래년,
A.거래월,
A.deposit AS 거래가,
(A.deposit/D.전용면적 * 3.3) AS 평당가격,
B.local1 AS 시도,
B.local2 AS 시군구,
B.local3 AS 읍면동,
B.category,
B.category2,
C.난방방식,
C.난방연료,
C.총주차대수,
C.가구당주차대수,
C.준공년월,
C.총세대수,
C.총동수,
C.최고층수,
C.최저층수,
D.공급면적,
D.전용면적,
D.방수,
D.욕실수,
D.현관구조
FROM jikbang.danji_molit_item AS A
JOIN jikbang.area_danji AS B
ON A.area_danji_id = B.id
JOIN jikbang.danji_detail AS C
ON A.area_danji_id = C.area_danji_id
JOIN jikbang.danji_room_type AS D
ON A.room_type_id = D.id
WHERE A.거래유형 = 1 AND B.enabled2 = 4 AND now() BETWEEN D.begin AND D.end
"""
meta_table = pd_execute.query(all_join)
```

### 만들어진 테이블 (10개의 Example Row)

Finished Table (10 Example Rows Shown)

이 과정을 통해 아래와 같은 테이블이 만들어집니다. (일단 예시로 10개만 표시하였습니다.)

```
In [5]: meta_table.head(10)
```

Out[5]:

	id	area_danji_id	총	거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방방식	난방연료	총주차대수	가구당주차대수	준공년월	총세대수	총동수	최고층수	최저층수	공급면적	전용면적	방수	욕실수	현관구조
0	4671331	1	12	2006	1	7480	589.820810	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
1	4671332	1	3	2006	1	7300	575.627261	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
2	4671333	1	15	2006	1	7250	571.684609	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
3	4671334	1	5	2006	2	7500	591.397871	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
4	4671335	1	6	2006	2	7900	622.939091	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
5	4671336	1	8	2006	2	7450	587.455219	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
6	4671337	1	14	2006	2	8000	630.824396	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
7	4671338	1	3	2006	2	7500	591.397871	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
8	4671339	1	14	2006	3	7400	583.512566	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식
9	4671340	1	12	2006	3	7500	591.397871	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식

# 이상점(Outlier) 제거

## Removing Outliers

정확한 분석을 위해 데이터에 있는 아파트 평당 매매가를 기준으로 하여 너무 높거나 너무 낮은, 즉 상위와 하위 각 0.001% 값을 가진 데이터를 제거 했습니다.

이것들은 실수로 적힌 가능성이 높기 때문에 분석에서 제외해야 합니다.

```
In [24]: np.percentile(meta_table['평당가격'], [0.001, 99.999])
```

```
Out[24]: array([ 62.7540612 , 9537.57214918])
```

```
In [4]: upper = meta_table['평당가격'].quantile(0.99999)
```

```
lower = meta_table['평당가격'].quantile(0.00001)
```

```
low_outlier_removed = meta_table[lower < meta_table['평당가격']]
```

```
outlier_removed = low_outlier_removed[low_outlier_removed['평당가격'] < upper]
```

```
outlier_removed.head()
```

```
Out[4]:
```

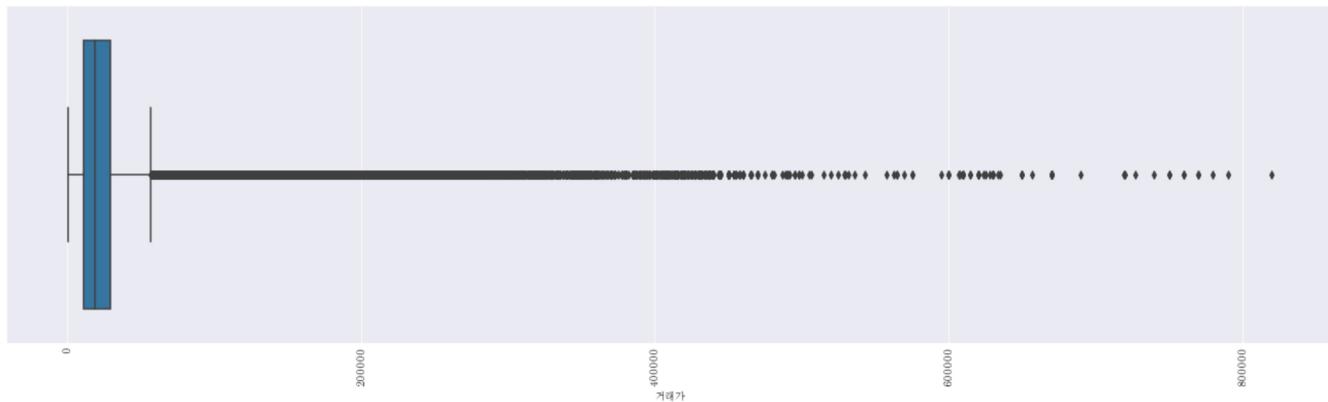
	id	area_danji_id	총	거래	년	거래	월	거래	가	평당가격	시도	시군구	읍면동	category	category2	난방방식	난방연료	총주차대수	가구당주차대수	준공년월	총세대수	총동수	최고층수	최저층수	공급면적	전용면적	방수	욕실수	현관구조
0	4671331	1	12	2006	1	7480	589.820810	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식			
1	4671332	1	3	2006	1	7300	575.627261	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식			
2	4671333	1	15	2006	1	7250	571.684609	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식			
3	4671334	1	5	2006	2	7500	591.397871	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식			
4	4671335	1	6	2006	2	7900	622.939091	경기도	군포시	산본동	아파트	아파트	지역난방	열병합	1210.0	0.47	199306	2550	26	25.0	15.0	57.4	41.85	2.0	1.0	복도식			

## 데이터베이스의 아파트 매매 거래가 Boxplot

### Apartment Sale Price Boxplot

아래는 상자 그림(Box Plot)으로 아파트 매매가 데이터를 표시한 것입니다. 상자 그림은 어디에 많은 데이터가 집중적으로 모여있는지를 시각화한 것입니다.

```
In [7]: plt.subplots(figsize=(23, 6))
sns.boxplot(x='거래가', data=meta_table)
plt.xticks(rotation=90);
```



- 평균 ≈ 23620 ≈ 2억 3620만원 (Mean Sale Price is around 236 million 200 thousand Won)

```
In [8]: meta_table['거래가'].mean()
```

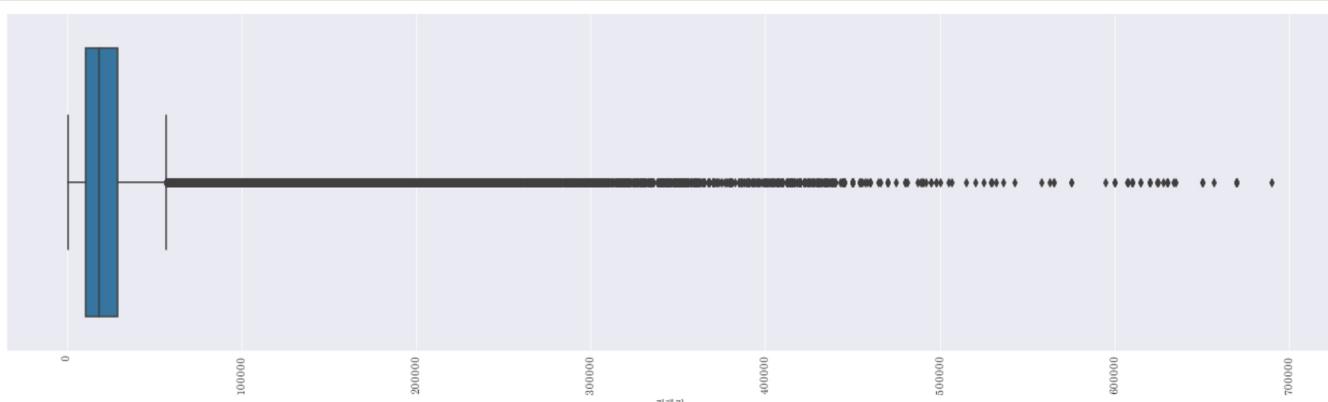
```
Out[8]: 23620.560618294516
```

## 데이터베이스의 아파트 매매 거래가 (Outlier 제거) Boxplot

### Apartment Sale Price Boxplot (Outlier Removed)

아래는 상자 그림(Box Plot)으로 이상점을 제거한 아파트 매매가 데이터를 표시한 것입니다. 이상점 제거로 평균이 감소한 걸 볼 수 있습니다.

```
In [26]: plt.subplots(figsize=(23, 6))
sns.boxplot(x='거래가', data=outlier_removed)
plt.xticks(rotation=90);
```



- 평균 ≈ 23617 ≈ 2억 3617만원 (Mean Sale Price is around 236 million 170 thousand Won)

```
In [27]: outlier_removed['거래가'].mean()
```

```
Out[27]: 23617.767794045365
```

## 평당가격 내림차순

Sale Price Per 3.3m<sup>2</sup> Descending Order

테이블을 평당 매매가의 내림차순으로 나열한 것입니다.(5개만 예시로 표시)

이 데이터 베이스에서 제일 높은 평당 가격은 1억 6947만원이었습니다. 이것은 서울 특별시 강동구 명일동에 위치한 약 25평의 삼익 그린 11차 아파트의 평당 가격이었고 거래가가 43억 4천만원이었습니다.

에러값으로 추측이 되어 다른 자료 (<http://www.aptbong.com/content.php?bt=apt&ct1=110000000&ct2=117400000&ct3=117401010&dn=226&y=2006>)를 비교해 보니 실거래가가 4억 3400만원이었습니다.

0을 하나 더 표기하여 생긴 실수로 판단되어집니다.

In [11]: `meta_table.sort_values(by='평당가격', ascending=False).head()`

Out[11]:

id	area_danji_id	총 거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방 방식	난방 연료	총주차 대수	기구당주차 대수	준공년월	총세대수	총동수	최고 총수	최저 총수	공급면적	전용면적	방수	유실수	현관구조	
3460367	944475	3752	9 2006	6 434000	16947.106423	서울특별시	강동구	명일동	아파트	아파트	중앙 난방	도시 가스	170.0	1.12	198606	152	1	15.0	12.0	100.89	84.51	3.0	2.0	계단식	
399163	12514835	1027	3 2018	2 100000	12981.903974	서울특별시	강동구	둔촌동	아파트	재건축	중앙 난방	도시 가스	959.0	0.70	198001	1372	47	5.0	5.0	25.42	25.42	1.0	1.0	계단식	
1029477	12664292	3850	11 2006	4 301900	11747.082050	경기도	용인시	기흥구	마북동	아파트	아파트	지역 난방	열병합	2018.0	1.28	200009	1576	27	20.0	11.0	107.83	84.81	3.0	2.0	계단식
2484550	15229576	22655	3 2016	12 820000	11138.094398	서울특별시	용산구	한남동	아파트	아파트	개별 난방	도시 가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식	
2484547	15228743	22655	3 2016	1 790000	10730.603140	서울특별시	용산구	한남동	아파트	아파트	개별 난방	도시 가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식	

## 평당가격 내림차순 (Outlier 제거)

Sale Price Per 3.3m<sup>2</sup> Descending Order (Outliers Removed)

이상점을 제거한 테이블을 평당 매매가의 내림차순으로 나열한 것입니다.(5개만 예시로 표시)

이 데이터 세트에서 제일 높은 평당 가격은 9515만원이었습니다. 이상점을 제거 안 했을 때 있던 에러값들이 제거된 걸 볼 수 있습니다.

In [28]: `outlier_removed.sort_values(by='평당가격', ascending=False).head()`

Out[28]:

id	area_danji_id	총 거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방 방식	난방 연료	총주차 대수	기구당주차 대수	준공년월	총세대수	총동수	최고 총수	최저 총수	공급면적	전용면적	방수	유실수	현관구조
7138214	12262488	10777	2 2017	12 340000	9515.732043	서울특별시	강남구	암구정동	아파트	아파트	지역 난방	열병합	420.0	2.47	197707	170	6	5.0	5.0	140.26	117.91	4.0	2.0	계단식
1484163	12902587	6251	28 2009	12 563000	9509.161396	서울특별시	강남구	삼성동	아파트	아파트	지역 난방	열병합	1253.0	2.79	200403	449	3	46.0	23.0	243.76	195.38	4.0	2.0	복합식
2130662	12411346	17303	15 2018	1 242500	9422.465526	서울특별시	서초구	반포동	아파트	아파트	지역 난방	열병합	3893.0	1.59	200907	2444	28	32.0	23.0	113.71	84.93	3.0	2.0	계단식
2130307	12411345	17303	15 2018	1 242500	9422.465526	서울특별시	서초구	반포동	아파트	아파트	지역 난방	열병합	3893.0	1.59	200907	2444	28	32.0	23.0	113.70	84.93	3.0	2.0	계단식
2131038	12411347	17303	15 2018	1 242500	9422.465526	서울특별시	서초구	반포동	아파트	아파트	지역 난방	열병합	3893.0	1.59	200907	2444	28	32.0	23.0	114.62	84.93	3.0	2.0	계단식

## 평당가격 오름차순

Sale Price Per 3.3m<sup>2</sup> Ascending Order

테이블을 평당 매매가의 오름차순으로 나열한 것입니다.(5개만 예시로 표시)

이 데이터 베이스에서 제일 낮은 평당 가격은 10만원이었습니다. 이것은 경기도 안양시 동안구 호계동에 위치한 약 35평의 삼마을임광 아파트의 평당 가격이었고 거래가가 380만원이었습니다.

에러값으로 추측이 되어 다른 자료 (<http://www.aptbong.com/content.php?bt=apt&ct1=410000000&ct2=411730000&ct3=411731040&dn=5415&y=2012>)를 비교해 보니 당시 실거래가가 4~5억이었습니다.

In [13]: `meta_table.sort_values(by='평당가격', ascending=True).head()`

Out[13]:

id	area_danji_id	총 거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방 방식	난방 연료	총주차 대수	기구당주차 대수	준공년월	총세대수	총동수	최고 총수	최저 총수	공급면적	전용면적	방수	유실수	현관구조
6997939	3730123	605	1 2012	6 380	10.544904	경기도	안양시 동안구	호계동	아파트	아파트	지역 난방	열병합	549.0	1.41	199212	390	8	24.0	15.0	138.47	118.92	4.0	2.0	계단식
6083633	28100003	16067	15 2006	11 500	19.414049	충청남도	홍성군	홍성읍	아파트	아파트	지역 난방	도시 가스	711.0	1.25	200610	569	13	15.0	14.0	110.03	84.99	3.0	2.0	계단식
6381418	28175302	18653	14 2006	6 500	19.462138	경상남도	김해시	관동동	아파트	아파트	지역 난방	열병합	824.0	1.04	200506	794	12	18.0	14.0	105.34	84.78	3.0	2.0	계단식
6381431	28175315	18653	18 2006	6 500	19.462138	경상남도	김해시	관동동	아파트	아파트	지역 난방	열병합	824.0	1.04	200506	794	12	18.0	14.0	105.34	84.78	3.0	2.0	계단식
6381423	28175307	18653	16 2006	6 500	19.462138	경상남도	김해시	관동동	아파트	아파트	지역 난방	열병합	824.0	1.04	200506	794	12	18.0	14.0	105.34	84.78	3.0	2.0	계단식

## 평당가격 오름차순 (Outlier 제거)

Sale Price Per 3.3m<sup>2</sup> Ascending Order (Outliers Removed)

이상점을 제거한 테이블을 평당 매매가의 오름차순으로 나열한 것입니다. (5개만 예시로 표시)

이 데이터 세트에서 제일 낮은 평당 가격은 62만원이었습니다. 이상점을 제거 안 했을 때 있던 예외값들이 제거된 걸 볼 수 있습니다.

아직 예외값들이 존재하지만 오차가 큰 값들은 제거되었습니다.

In [29]: `outlier_removed.sort_values(by='평당가격', ascending=True).head()`

Out[29]:

id	area_danji_id	총 거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방	난방	총주차대수	기구당주차대수	준공년월	총세대수	총동수	최고총수	최저총수	공급면적	전용면적	방수	용실수	현관구조	
											방식	연료													
3965792	1066025	9306	10	2015	3 1614	62.868271	부산광역시	동래구	온천동	아파트	주상복합2	개별난방	도시가스	348.0	1.27	199903	274	1	28.0	28.0	115.89	84.72	3.0	2.0	복도식
6838018	6985435	26969	12	2007	1 1001	63.064145	경상북도	청도군	청도읍	아파트	아파트	개별난방	-	103.0	0.35	199512	293	2	15.0	8.0	69.10	52.38	3.0	1.0	복도식
6827423	6596189	26878	4	2006	6 1000	63.157895	전라남도	광양시	금호동	아파트	아파트	지역난방	열병합	Nan	0.00	198801	365	14	5.0	5.0	58.73	52.25	2.0	1.0	계단식
6828430	6596301	26881	5	2006	8 1000	63.230506	전라남도	광양시	금호동	아파트	아파트	지역난방	열병합	Nan	0.00	198610	661	26	5.0	5.0	58.67	52.19	2.0	1.0	계단식
6827333	6596093	26878	4	2006	6 1000	63.242621	전라남도	광양시	금호동	아파트	아파트	지역난방	열병합	Nan	0.00	198801	365	14	5.0	5.0	58.73	52.18	3.0	1.0	계단식

## 거래가 내림차순

Sale Price Descending Order

테이블을 매매가의 내림차순으로 나열한 것입니다.(5개만 예시로 표시)

이 데이터 베이스에서 제일 높은 거래가는 82억원이었습니다. 이것은 서울특별시 용산구 한남동에 위치한 약 72평의 한남더힐 아파트의 매매가였고 평당 가격은 약 1억 1138만원이었습니다.

다른 자료 (<http://bizn.donga.com/realestate/East/3/all/20180322/89219551/1>)를 찾아보니 이것은 실수가 없는 실제 거래였습니다.

In [15]: `meta_table.sort_values(by='거래가', ascending=False).head()`

Out[15]:

id	area_danji_id	총 거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방	난방	총주차대수	기구당주차대수	준공년월	총세대수	총동수	최고총수	최저총수	공급면적	전용면적	방수	용실수	현관구조	
											방식	연료													
2484550	15229576	22655	3	2016	12 820000	11138.094398	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484547	15228743	22655	3	2016	1 790000	10730.603140	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484551	15229653	22655	3	2017	6 780000	10594.772720	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484543	15228719	22655	3	2015	7 770000	10458.942301	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484542	15228711	22655	3	2015	2 770000	10458.942301	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식

## 거래가 내림차순 (Outlier 제거)

Sale Price Descending Order (Outliers Removed)

이상점을 제거한 테이블을 매매가의 내림차순으로 나열한 것입니다. (5개만 예시로 표시)

이 데이터 세트에서 제일 높은 매매가는 위와 같은 아파트로 69억원이었습니다. 이상점을 제거로 실수가 없는 실거래가 몇개 제거된 걸 볼 수 있습니다.

In [30]: `outlier_removed.sort_values(by='거래가', ascending=False).head()`

Out[30]:

id	area_danji_id	총 거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	category2	난방	난방	총주차대수	기구당주차대수	준공년월	총세대수	총동수	최고총수	최저총수	공급면적	전용면적	방수	용실수	현관구조	
											방식	연료													
2484544	15228724	22655	1	2015	9 690000	9372.298945	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484560	15229972	22655	1	2018	1 670000	9100.638106	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484556	15229739	22655	1	2017	11 670000	9100.638106	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484558	15229747	22655	1	2017	11 670000	9100.638106	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식
2484557	15229742	22655	1	2017	11 670000	9100.638106	서울특별시	용산구	한남동	아파트	아파트	개별난방	도시가스	1732.0	2.89	201101	600	32	13.0	3.0	331.11	242.95	5.0	2.0	계단식

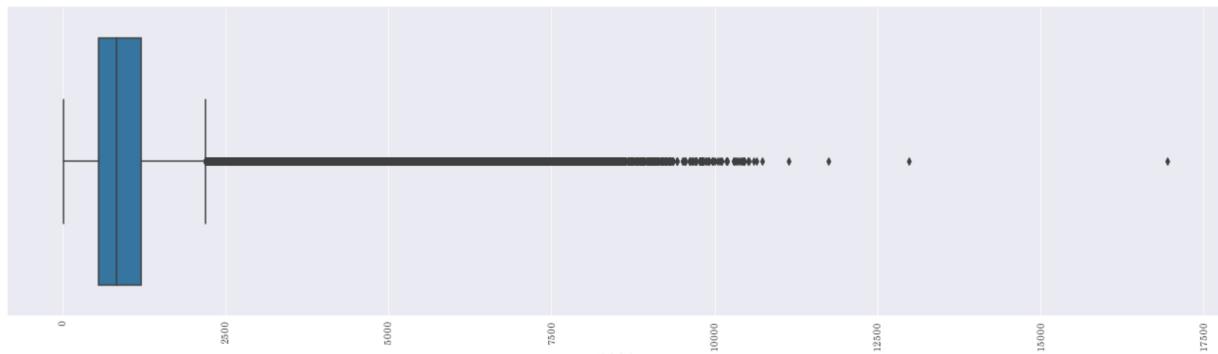
## 데이터베이스의 아파트 매매 평당가격 Boxplot

Apartment Sale Price Per 3.3m<sup>2</sup> Boxplot (3.3m<sup>2</sup> ≈ 1 Pyung)

아래는 상자 그림(Box Plot)으로 아파트 평당 매매가 데이터를 표시한 것입니다.

평균 평당 매매가는 약 987만원입니다.

```
In [31]: plt.subplots(figsize=(23, 6))
sns.boxplot(x='평당가격', data=meta_table)
plt.xticks(rotation=90);
```



```
In [156]: meta_table['평당가격'].mean()
```

```
Out[156]: 987.2867854787039
```

## 데이터베이스의 아파트 매매 평당가격 (Outlier 제거) Boxplot

Apartment Sale Price Per 3.3m<sup>2</sup> (Outlier Removed) Boxplot (3.3m<sup>2</sup> ≈ 1 Pyung)

아래는 상자 그림(Box Plot)으로 이상점을 제거한 아파트 평당 매매가 데이터를 표시한 것입니다.

평균 평당 매매가는 약 989만원으로 이상점 제거 전보다 조금 더 높은걸 볼 수 있습니다.

```
In [32]: plt.subplots(figsize=(23, 6))
sns.boxplot(x='평당가격', data=outlier_removed)
plt.xticks(rotation=90);
```



```
In [33]: outlier_removed['평당가격'].mean()
```

```
Out[33]: 989.0893030548791
```

## 현관구조별 평당가격 Boxplot

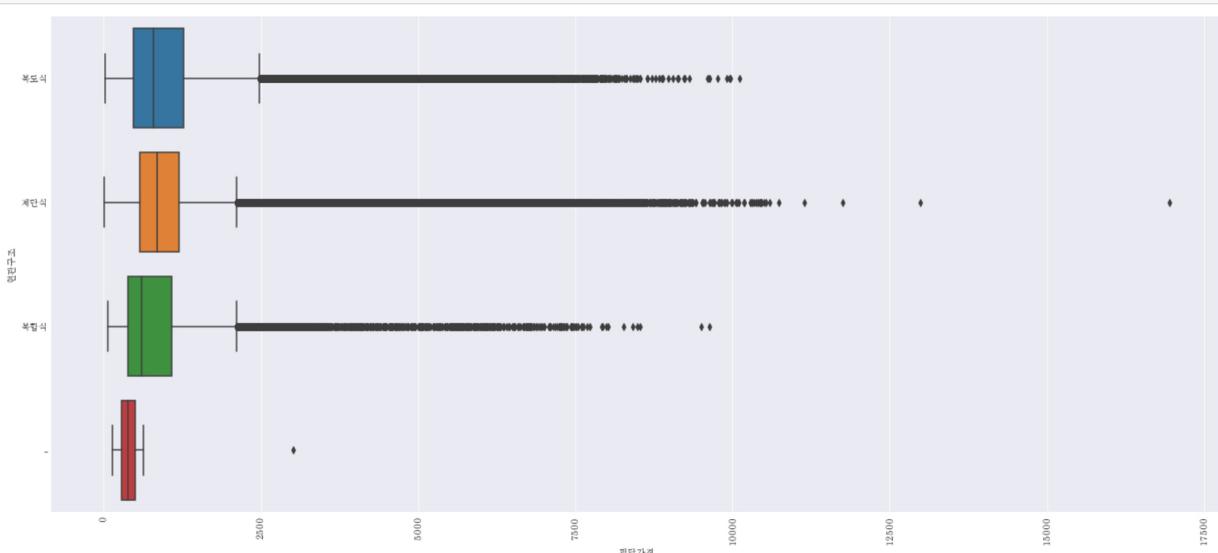
Different Entrance Structure's Sale Price Per 3.3m<sup>2</sup> Boxplot

각 현관구조 용어의 정의는 다음과 같습니다:

- 계단식: 같은 층의 2개구가 하나의 승강기를 사용하여 마주 보고 있는 형태의 아파트
- 복도식: 같은 층의 모든 세대가 복도를 한쪽이나 중앙으로 길게 설치하여 공동으로 사용하는 아파트
- 복합식: 같은 층의 3개구 이상이 하나의 승강기를 중심으로 배치되어 있는 형태의 아파트 (타워형 아파트에서 볼 수 있는 구조)

상자 그림(Boxplot)으로 데이터를 시각화를 하고 평균을 비교한 결과 평당가격은 계단식 아파트가 가장 높았고, 그 다음으로는 복도식, 복합식 순으로 나타났습니다.

```
In [159]: plt.subplots(figsize=(23, 10))
sns.boxplot(x='평당가격', y='현관구조', data=meta_table,
plt.xticks(rotation=90);
```



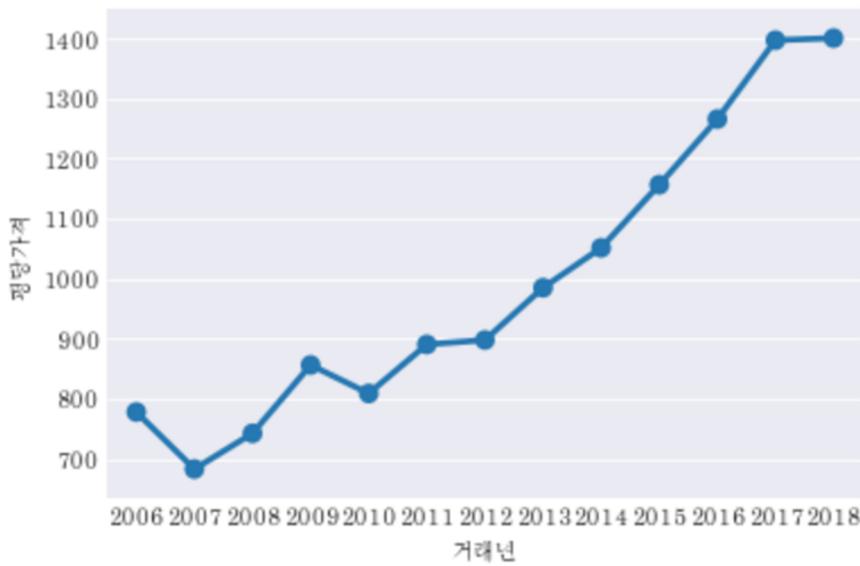
## 거래년 vs 평당가격 Pointplot

Year Sold vs Sale Price Per 3.3m<sup>2</sup> Pointplot

연별 평균 아파트 평당가격을 Pointplot으로 시각화하였습니다.

그 결과, 전국 평균 2010년 거래된 아파트 평당가격이 갑자기 하락한걸 볼 수 있었고 그 이후에는 상승세를 유지하는걸 볼 수 있습니다.

```
In [36]: sns.pointplot(x='거래년', y='평당가격', data=meta_table);
```



## 연별 각 시도의 평균 평당가격 변화 피벗 테이블

Average Sale Price Per 3.3m<sup>2</sup> Change Over Time In Different Cities and Provinces Pivot Table

연별 각 시도의 평균 평당가격 변화를 보여주는 피벗 테이블(Pivot Table)을 만들었습니다.

시도마다 평균 평당가격 변화 추세가 다른 걸 알 수 있고 특히 제일 흥미로운 점은 2010년 평균 평당가격이었습니다.

위에 Pointplot에도 나왔듯이 전국 평균적으로는 평당 가격이 많이 내려갔습니다. 하지만 피벗 테이블을 본 결과 오히려 지방은 가격이 오른 걸 볼 수 있습니다.

```
In [37]: meta_table.pivot_table(values='평당가격', index='시도',
...                           columns=['거래년'], aggfunc=np.mean)
```

Out[37]:

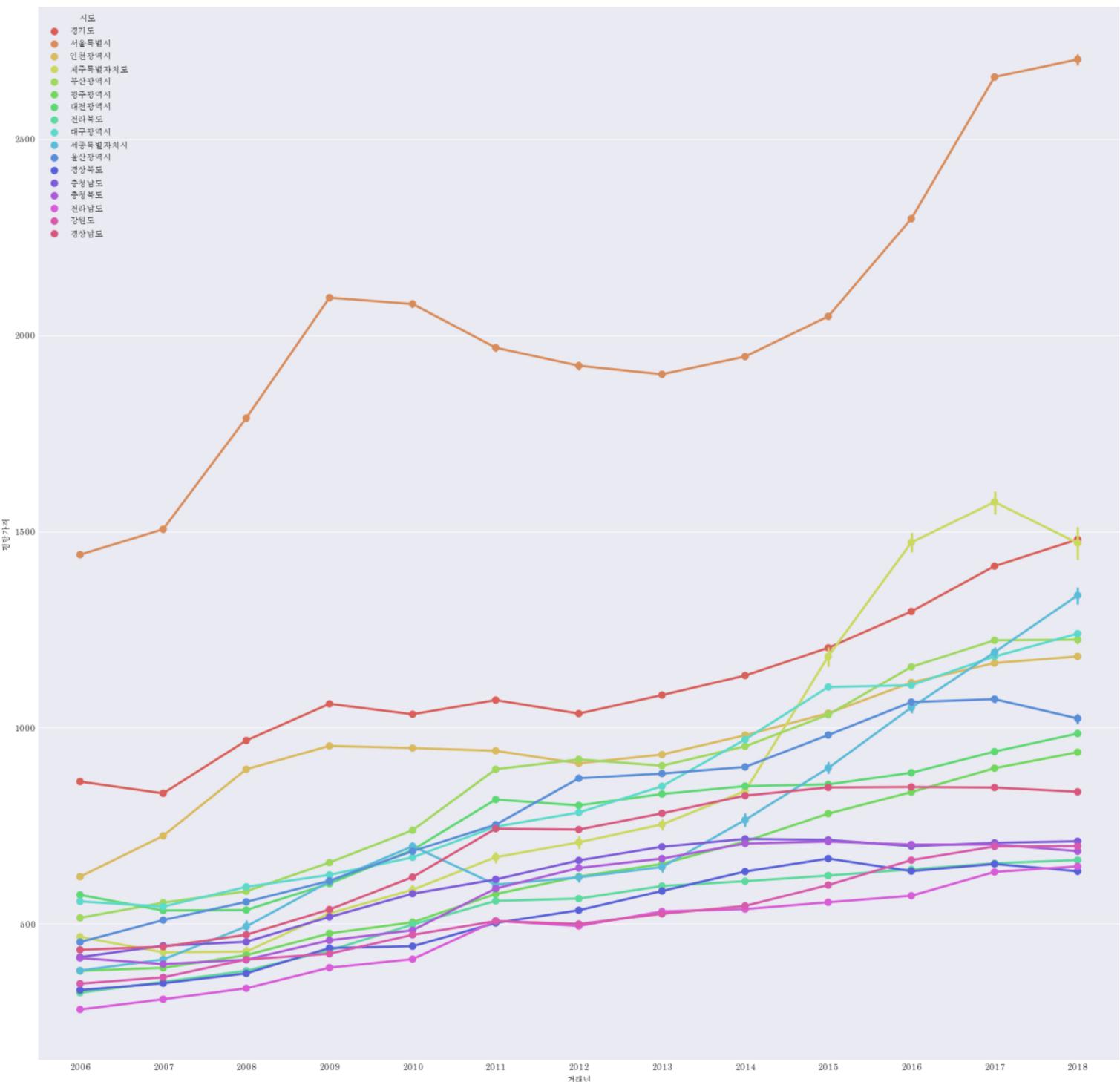
거래년 시도	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
강원도	347.449376	363.654461	408.978485	423.680329	471.811820	506.837008	499.451806	525.420069	545.579171	598.685029	662.214518	696.896425	698.049409
경기도	862.320189	832.358937	906.948916	1060.397355	1033.815324	1069.910988	1035.479267	1082.670021	1132.419989	1203.040929	1295.735283	1411.155217	1479.261333
경상남도	433.544419	441.680861	472.080058	536.341625	619.005872	742.283930	740.063416	781.355342	826.606226	847.404684	848.621482	847.167683	836.313505
경상북도	331.055959	348.561490	373.641693	438.053509	442.775098	502.024469	534.568468	583.593709	633.146960	666.147234	634.194346	652.444273	633.867414
광주광역시	380.297498	387.642689	420.173037	475.455002	503.782137	575.852774	620.313168	652.440061	709.935820	780.609910	835.595691	896.387507	937.015710
대구광역시	557.072048	543.760565	594.179551	624.578224	669.204298	747.356877	783.530322	850.413911	969.368930	1103.154858	1107.865357	1180.783916	1239.138056
대전광역시	573.700006	533.854166	535.129945	602.380223	688.133185	816.582321	801.655833	830.689439	850.731755	855.242213	884.802834	938.624771	984.598586
부산광역시	515.060724	553.705521	583.017269	656.044250	738.303611	893.724516	918.803830	902.760665	952.070007	1032.638658	1154.508990	1222.176423	1223.912295
서울특별시	1440.305461	1505.178656	1788.488026	2095.165468	2079.153235	1967.746769	1921.796552	1900.043953	1945.104785	2047.421054	2296.385656	2657.349498	2702.036669
세종특별자치시	380.728870	409.102505	493.168038	607.753607	697.579977	599.830110	618.402322	644.511520	764.494445	896.884296	1050.907230	1191.727790	1336.747921
울산광역시	453.490734	509.377860	555.846727	609.402162	685.051655	752.346904	870.736344	882.681235	899.585953	980.821538	1064.867455	1072.380741	1023.132720
인천광역시	620.048746	723.978746	893.739465	953.169497	947.710177	940.616775	908.868985	931.020436	980.298401	1036.348882	1114.613534	1164.365305	1181.189873
전라남도	281.724231	307.757449	335.703995	388.110611	409.901591	507.487676	494.410005	531.406354	537.400613	554.828799	571.470359	632.040295	646.390984
전라북도	324.424553	351.967753	380.286481	432.033923	497.855446	558.311715	564.184402	596.189314	608.480759	622.912693	638.637436	654.128405	662.399259
제주특별자치도	466.681803	427.137847	429.015113	526.405679	586.646528	669.611120	707.550207	753.049788	838.132764	1181.208367	1471.506219	1574.605228	1470.060044
충청남도	415.044653	444.083063	454.079523	517.253223	576.629271	613.055072	661.501971	696.269919	716.213323	713.718317	697.664343	706.166084	710.240866
충청북도	412.910576	397.159242	408.399061	457.957977	483.416073	589.969154	642.177618	666.121610	704.486482	709.663803	701.794758	702.174510	685.048566

## 연별 각 시도의 평균 평당가격 변화 Pointplot

Average Sale Price Per 3.3m<sup>2</sup> Change Over Time In Different Cities and Provinces Pointplot

Pointplot으로 연별 각 시도의 평균 평당가격 변화를 시각화 하였습니다.

```
In [40]: plt.subplots(figsize=(23, 23))
sns.pointplot(x='거래년', y='평당가격', hue='시도', data=meta_table, palette=sns.color_palette("hls", 17));
```

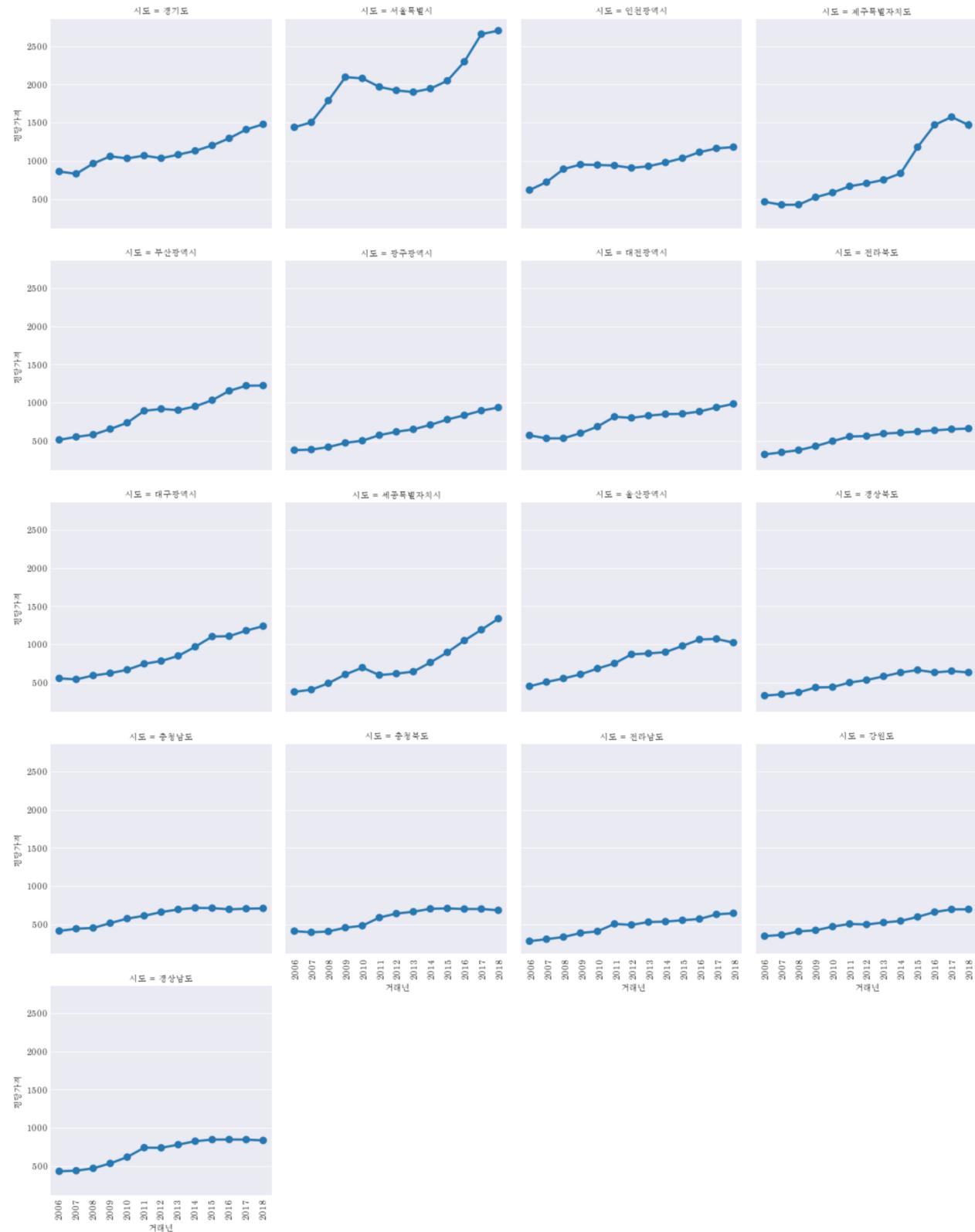


## 연별 각 시도의 평균 평당가격 변화 Point Factorplot

Average Sale Price Per 3.3m<sup>2</sup> Change Over Time In Different Cities and Provinces Point Factorplot

Point Factorplot을 사용하여 연별 각 시도의 평균 평당가격 변화를 개별 그래프로 시각화 하였습니다.

```
In [41]: sns.factorplot(x='거래년', y='평당가격', col='시도', col_wrap=4, data=meta_table, kind='point')
plt.xticks(rotation=90);
plt.xticks(rotation=90);
plt.xticks(rotation=90);
plt.xticks(rotation=90);
```



## 시도별 거래수

Number of Records Per Cities and Provinces

사용한 데이터 베이스에서 시도별 거래수를 구했습니다.

경기도와 서울특별시 거래 수가 다른 시도보다 월등히 높기 때문에 전국평균값을 구할 때 더욱 많은 영향을 끼치는 걸로 해석됩니다.

```
In [42]: meta_table.groupby('시도').count()['id']
```

```
Out[42]: 시도
강원도      230448
경기도      1964741
경상남도    474642
경상북도    308024
광주광역시  315903
대구광역시  422702
대전광역시  289787
부산광역시  552248
서울특별시   960264
세종특별자치시 27923
울산광역시  189383
인천광역시  448415
전라남도    199406
전라북도    262159
제주특별자치도 19559
충청남도    333828
충청북도    239513
Name: id, dtype: int64
```

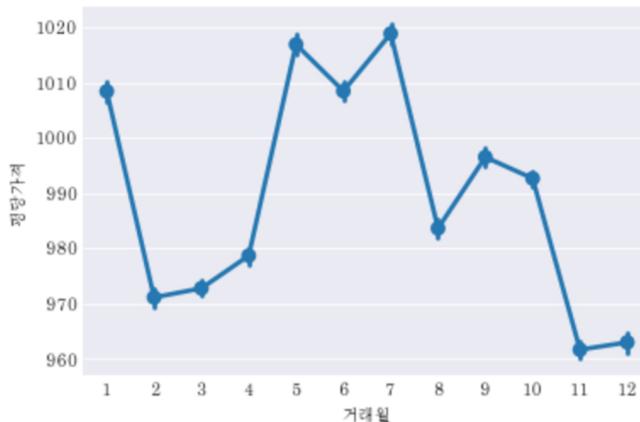
## 거래월별 평균 평당가격 Pointplot

Average Sale Price Per 3.3m<sup>2</sup> in Different Months Pointplot

거래월별 평균 평당가격을 Pointplot으로 나타냈습니다.

비교한 결과, 7월이 평균 평당가격이 가장 높았고, 그 다음으로는 5월, 1월, 6월, 9월, 10월, 8월, 4월, 3월, 2월, 12월, 11월 순으로 나타났습니다.

```
In [17]: sns.pointplot(x='거래월', y='평당가격', data=meta_table);
```



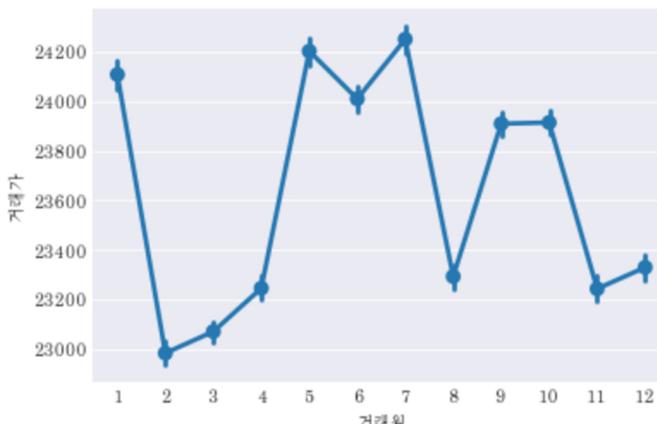
## 거래월별 평균 거래가 Pointplot

Average Sale Price in Different Months Pointplot

거래월별 평균 거래가를 Pointplot으로 나타냈습니다.

비교한 결과, 7월이 평균 거래가가 가장 높았고, 그 다음으로는 5월, 1월, 6월, 10월, 9월, 12월, 8월, 11월, 4월, 3월, 2월 순으로 나타났습니다.

```
In [18]: sns.pointplot(x='거래월', y='거래가', data=meta_table);
```



## 층별 평균 평당가격 Barplot

Average Sale Price Per 3.3m<sup>2</sup> in Different Floors Barplot

층별 평균 평당가격을 바 플롯(Bar Plot)으로 나타냈습니다.

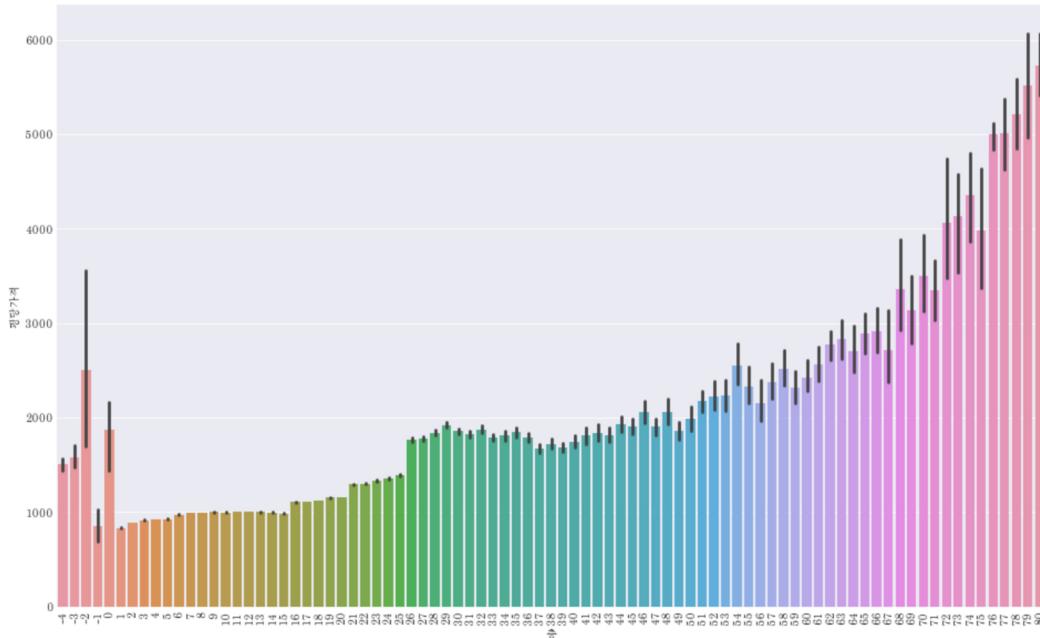
층수가 높아질수록 평균 평당가격도 올라가는 걸 볼 수 있습니다.

40층이 넘는 고층 아파트들은 대부분 고급아파트이기 때문에 평균 평당가격이 더 가파르게 올라가는 걸로 해석됩니다.

또한, 사용 데이터 베이스에 0층과 층수가 마이너스인 값을 가진 데이터가 존재하는걸 볼 수 있습니다.

이런 값을 가진 데이터에 대한 조취가 필요합니다.

```
In [177]: plt.subplots(figsize=(16, 10))
sns.barplot(x='층', y='평당가격', data=meta_table);
plt.xticks(rotation=90);
```



## 방수별 평균 평당가격 Barplot

Average Sale Price Per 3.3m<sup>2</sup> Depending on the Number of Rooms Barplot

방수별 평균 평당가격을 바 플롯(Bar Plot)으로 나타냈습니다.

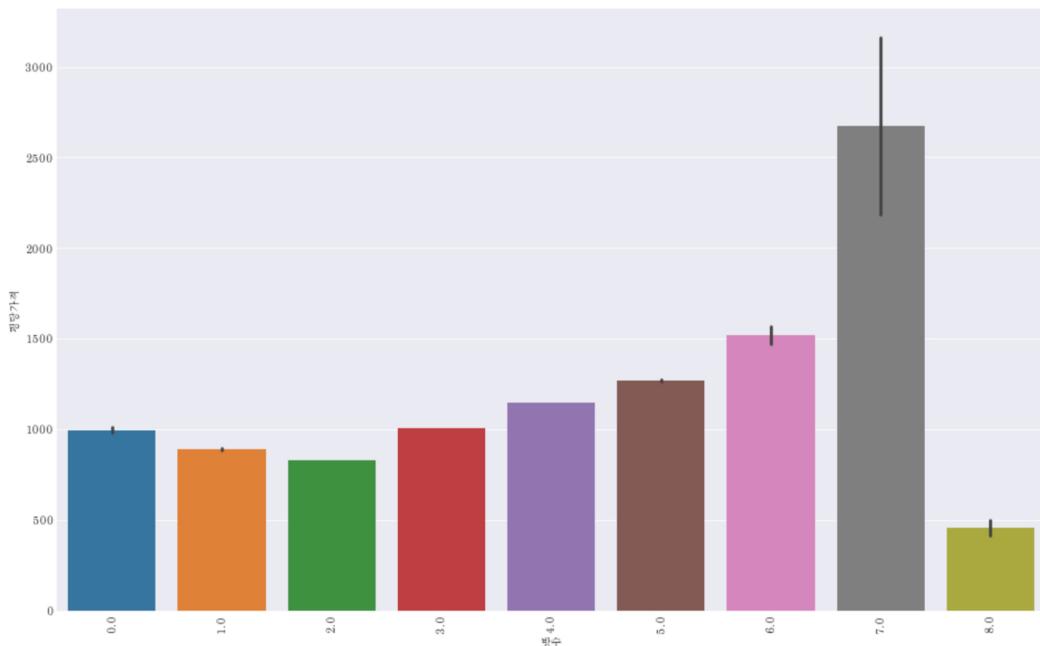
방수가 높아질수록 평균 평당가격도 올라가는 걸 볼 수 있습니다.

하지만 8방 같은 경우 평균가격이 금하락하는 걸로 보입니다.

또한, 방수가 0인 값을 가진 데이터가 존재합니다.

이것이 스튜디오 타입을 기재한 것인지 아니면 실수로 기재한 것인지 확인이 필요합니다.

```
In [43]: plt.subplots(figsize=(16, 10))
sns.barplot(x='방수', y='평당가격', data=meta_table);
plt.xticks(rotation=90);
```



## 욕실수별 평균 평당가격 Barplot

Average Sale Price Per 3.3m<sup>2</sup> Depending on the Number of Bathrooms Barplot

욕실수별 평균 평당가격을 바 플롯(Bar Plot)으로 나타냈습니다.

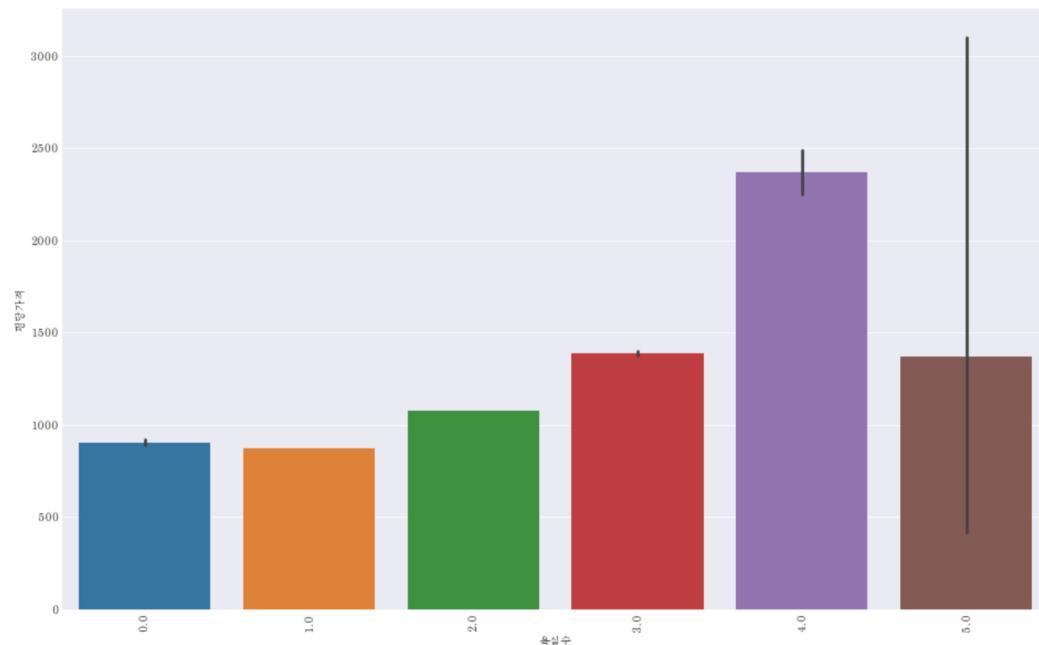
욕실수가 높아질수록 평균 평당가격도 올라가는 걸 볼 수 있습니다.

하지만 욕실 5개 같은 경우 평균가격이 급하락하고 에러 바(Error Bar)도 넓어보입니다.

또한, 욕실수가 0인 값은 가진 데이터가 존재합니다.

이것이 실수로 기재한 것인지 확인이 필요합니다.

```
In [19]: plt.subplots(figsize=(16, 10))
sns.barplot(x='욕실수', y='평당가격', data=meta_table);
plt.xticks(rotation=90);
```



## 난방방식별 평당가격 Boxplot

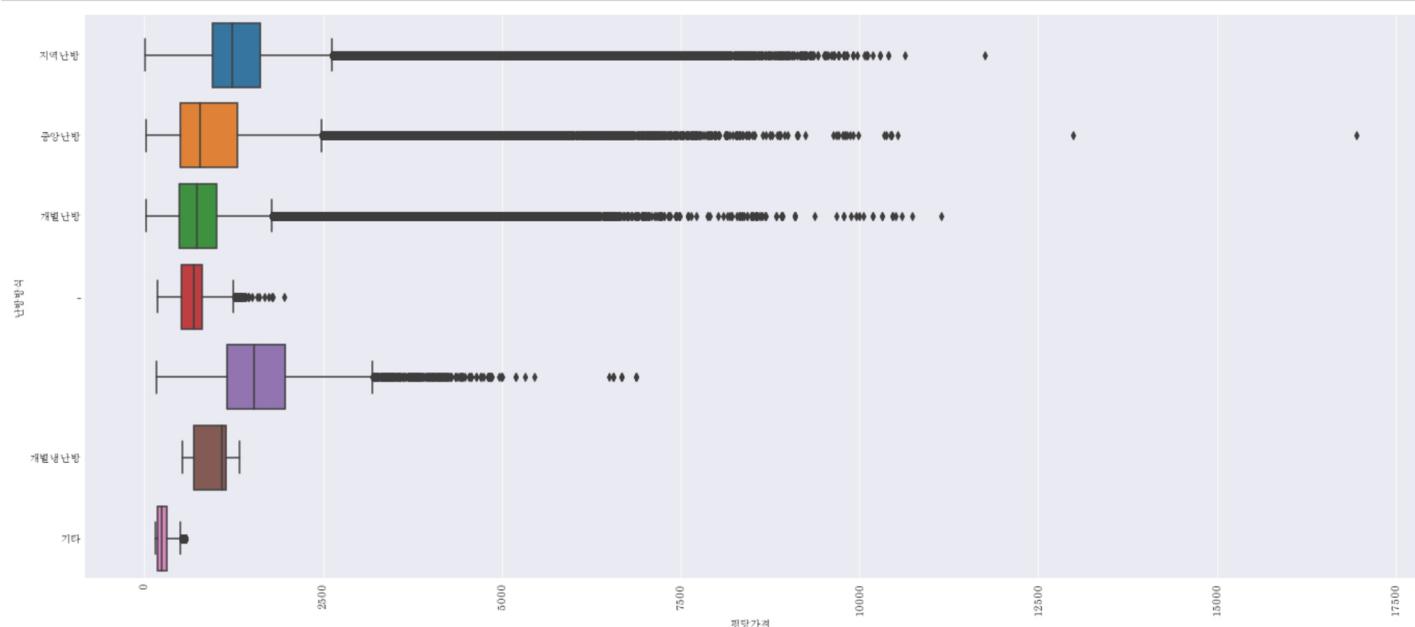
Sale Price Per 3.3m<sup>2</sup> Depending on Type of Heating Boxplot

난방방식별 평당가격 데이터가 어떻게 분포되어있는지 나타내는 그림 상자 (Box Plot)입니다.

평균을 비교하면 지역난방이 제일 높고, 그 다음으로는 중앙난방, 개별냉난방, 개별난방 순으로 나타납니다.

"-", "기타", "" 이런 값을 가진 데이터도 존재하는걸 알 수 있습니다.

```
In [20]: plt.subplots(figsize=(23, 10))
sns.boxplot(x='평당가격', y='난방방식', data=meta_table);
plt.xticks(rotation=90);
```



## 'category' 칼럼에 있는 Value Count

Value Count of column 'category'

'category' 칼럼에는 "아파트"라는 값만 존재하는 걸 볼 수 있습니다.

In [44]: `meta_table.groupby('category').count()`

Out[44]:

거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category2	난방방식	난방연료	총주차대수	기구당주차대수	준공년월	총세대수	총동수	최고층수	최저층수	공급면적	전용면적	방수	욕실수	현관구조
238945	7238945	7238945	7238945	7238945	7238945	7238945	7238945	7238945	7238945	6545596	6996977	7238945	7238945	7238945	7237979	7231557	7238945	7238945	7238681	7177765	

## 'category2' 칼럼에 있는 Value Count

Value Count of column 'category2'

하지만 'category2' 칼럼에는 6가지의 값이 존재합니다.

필터링이 필요한지 논의가 필요합니다.

In [45]: `meta_table.groupby('category2').count()`

Out[45]:

거래년	거래월	거래가	평당가격	시도	시군구	읍면동	category	난방방식	난방연료	총주차대수	기구당주차대수	준공년월	총세대수	총동수	최고층수	최저층수	공급면적	전용면적	방수	욕실수	현관구조
794	794	794	794	794	794	794	794	794	794	794	794	730	794	794	794	794	794	794	794	794	
26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	26	
7132860	7132860	7132860	7132860	7132860	7132860	7132860	7132860	7132860	7132860	6441306	6892127	7132860	7132860	7132860	7131929	7125507	7132860	7132860	7132654	7132654	
7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7188	7055	
83186	83186	83186	83186	83186	83186	83186	83186	83186	83186	81640	82264	83186	83186	83186	83184	83184	83186	83186	83186	83186	
14891	14891	14891	14891	14891	14891	14891	14891	14891	14891	14642	14642	14891	14891	14891	14858	14858	14891	14891	14833	14833	

## 거래가 vs 전용면적 LMplot

Sale Price vs Area for Exclusive Use LMplot

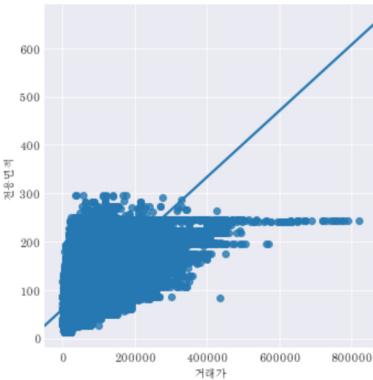
거래가와 전용면적의 상관관계를 보기 위해 LMplot을 사용하여 시각화하였습니다.

서로 양의 상관관계(Positive Correlation)를 갖고 있습니다.

전체적으로 전용면적이 넓을수록 거래가도 높은 걸 볼 수 있습니다.

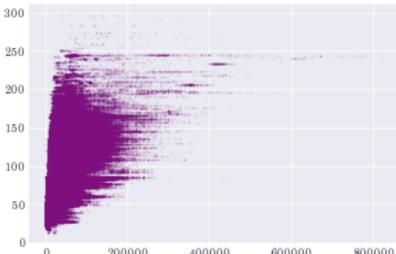
In [186]: `sns.lmplot(x='거래가', y='전용면적', data=meta_table)`

Out[186]: <seaborn.axisgrid.FacetGrid at 0x1ba06be1d0>



어떻게 데이터가 분포되어 있는지 보기 위해 마커사이즈는 감소를 하고 투명도 증가하였습니다.

In [222]: `plt.plot('거래가', '전용면적', data=meta_table, linestyle='', marker='o', markersize=1, alpha=0.05, color="purple");`



## 거래가 vs 전용면적 (Outlier 제거) LMplot

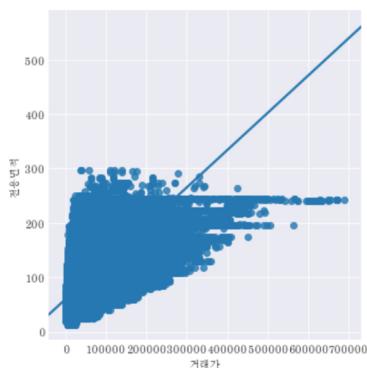
Sale Price vs Area for Exclusive Use (Outliers Removed) LMplot

이상점이 제거된 데이터의 거래가와 전용면적의 상관관계를 보기 위해 LMplot을 사용하여 시각화하였습니다.

위 그래프와 비교했을 때 다른 데이터와 멀리 떨어져 있는 아랫부분 데이터들이 제거된 걸 볼 수 있습니다.

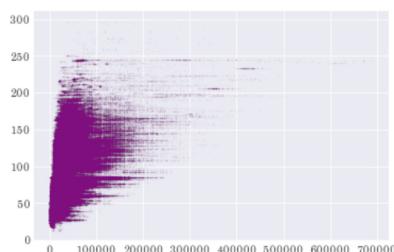
```
In [46]: sns.lmplot(x='거래가', y='전용면적', data=outlier_removed)
```

```
Out[46]: <seaborn.axisgrid.FacetGrid at 0x1b459d0c50>
```



어떻게 데이터가 분포되어 있는지 보기 위해 마커사이즈는 감소를 하고 투명도 증가하였습니다.

```
In [47]: plt.plot('거래가', '전용면적', data=outlier_removed, linestyle=' ', marker='o', markersize=0.3, alpha=0.05, color="purple");
```



## 거래가 vs 전용면적 Correlation

Sale Price vs Area for Exclusive Use Correlation

거래가와 전용면적의 상관계수(Correlation Coefficient)를 산출했습니다.

약 0.536으로 거래가와 전용면적은 꽤 많은 상관관계를 갖고 있다고 해석할 수 있습니다.

```
In [48]: meta_table['거래가'].corr(meta_table['전용면적'])
```

```
Out[48]: 0.5366069891686992
```

## 거래가 vs 전용면적 (Outlier 제거) Correlation

Sale Price vs Area for Exclusive Use (Outliers Removed) Correlation

이상점을 제거한 데이터의 거래가와 전용면적의 상관계수(Correlation Coefficient)를 산출했습니다.

약 0.537로 오차점 제거 전보다 많이 조금 더 높았습니다.

```
In [49]: outlier_removed['거래가'].corr(outlier_removed['전용면적'])
```

```
Out[49]: 0.5370864411627309
```

## 총세대 수 vs 평당가격 LMplot

Number of Households vs Sale Price Per 3.3m<sup>2</sup> LMplot

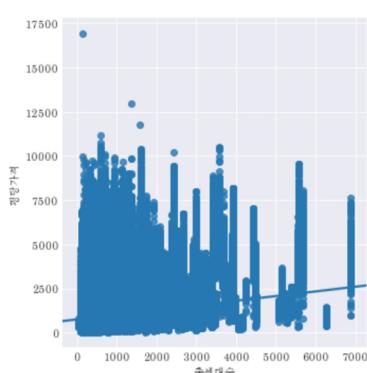
총세대 수와 평당가격의 상관관계를 보기 위해 LMplot을 사용하여 시각화하였습니다.

위처럼 강하지는 않지만 그래도 서로 양의 상관관계(Positive Correlation)를 갖고 있습니다.

전체적으로 총세대 수가 클수록 평당가격도 높은 걸 볼 수 있습니다.

```
In [190]: sns.lmplot(x='총세대수', y='평당가격', data=meta_table)
```

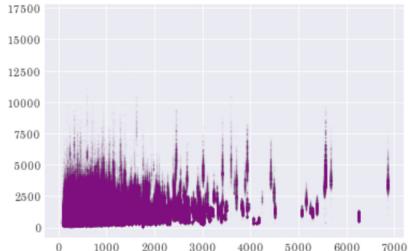
```
Out[190]: <seaborn.axisgrid.FacetGrid at 0x1abf863c88>
```



어떻게 데이터가 분포되어 있는지 보기 위해 마커사이즈는 감소를 하고 투명도 증가하였습니다.

```
In [226]: plt.plot('총세대수', '평당가격', data=meta_table, linestyle=' ', marker='o', markersize=0.2, alpha=0.05, color="purple")
```

```
Out[226]: <matplotlib.lines.Line2D at 0x1acdcc2278>
```



## 가구당 주차대수 vs 평당가격 LMplot

*Number of Parking Lots per Household vs Sale Price Per 3.3m<sup>2</sup> LMplot*

가구당 주차대수와 평당가격의 상관관계를 보기 위해 LMplot을 사용하여 시각화하였습니다.

서로 양의 상관관계(Positive Correlation)를 갖고 있는걸 볼 수 있습니다.

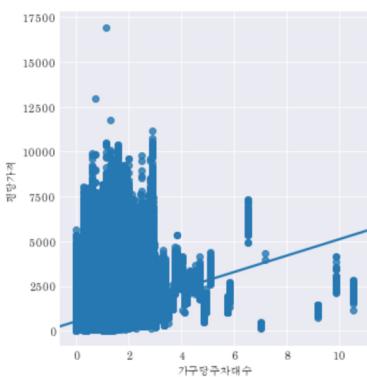
전체적으로 가구당 주차대수가 클수록 평당가격도 높은 걸 볼 수 있습니다.

하지만 그래프 오른쪽에 있는 데이터를 보면 주차대수가 많지만 최적합 직선(Best Fit Line) 아래 있습니다.

이건 어느 정도 가구당 주차대수를 넘으면 오히려 평당가격이 높지 않다는걸 의미합니다.

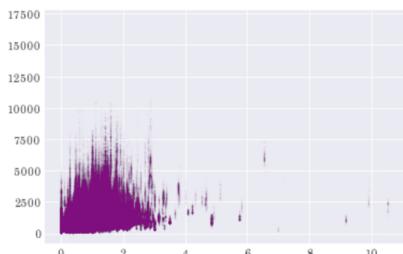
인구 밀집 지역을 벗어나기 위해 땅 가격이 낮기 때문인 거로 해석됩니다.

```
In [192]: sns.lmplot(x='가구당주차대수', y='평당가격', data=meta_table);
```



어떻게 데이터가 분포되어 있는지 보기 위해 마커사이즈는 감소를 하고 투명도 증가하였습니다.

```
In [231]: plt.plot('가구당주차대수', '평당가격', data=meta_table, linestyle=' ', marker='o', markersize=0.2, alpha=0.05, color="purple");
```



## 전용면적 vs 공급면적 LMplot

*Area for Exclusive Use vs Area for Common Use LMplot*

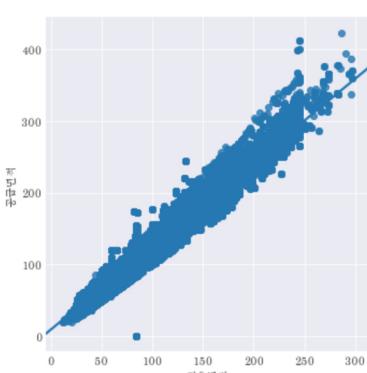
전용면적과 공급면적을 LMplot을 사용하여 시각화하였습니다.

그래프를 보면 하단에 전용면적은 약 70인데 공급면적은 0인 값이 있는 게 보여집니다.

이건 실수로 기재된 데이터인데 시각화를 통해 이런 데이터의 존재를 쉽게 파악할 수 있습니다.

```
In [194]: sns.lmplot(x='전용면적', y='공급면적', data=meta_table)
```

```
Out[194]: <seaborn.axisgrid.FacetGrid at 0x1bc8ba9b38>
```



## 사례 연구(Case Study): 한남동

Case Study = Hannam-dong

이상점 제거가 실제 데이터에 끼치는 영향을 알아보기 위한 사례 연구(Case Study)를 진행하였습니다.

한남동 선정 이유는 높은 거래가로 에러가 없는 실제 거래기록이 이상점 제거로 함께 지워진 거로 보였기 때문입니다.

```
In [9]: hannam = meta_table[meta_table['읍면동'] == '한남동']
```

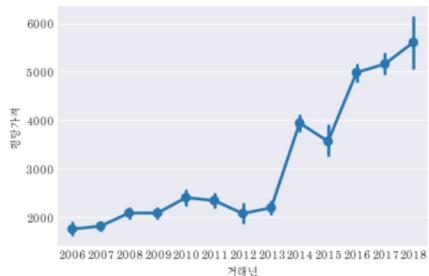
```
In [10]: hannam_outliergone = outlier_removed[outlier_removed['읍면동'] == '한남동']
```

### 한남동 거래년 vs 평당가격 Pointplot

Hannam-dong Year Sold vs Sale Price Per 3.3m<sup>2</sup> Pointplot

한남동 시세 변화를 보기위해 Pointplot을 이용해 시각화하였습니다.

```
In [11]: sns.pointplot(x='거래년', y='평당가격', data=hannam);
```



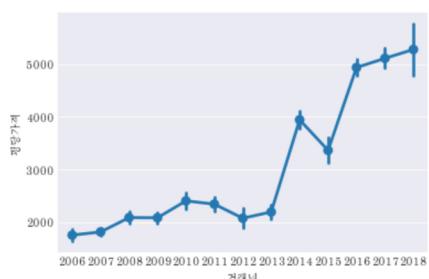
### 한남동 거래년 vs 평당가격 (Outlier 제거) Pointplot

Hannam-dong Year Sold vs Sale Price Per 3.3m<sup>2</sup> (Outliers Removed) Pointplot

이상점을 제거한 데이터로 같은 Pointplot을 만들었습니다.

이상점 제거로 평균 평당가격이 감소하였지만 큰 추세는 변화가 없는걸 볼 수 있습니다.

```
In [12]: sns.pointplot(x='거래년', y='평당가격', data=hannam_outliergone);
```

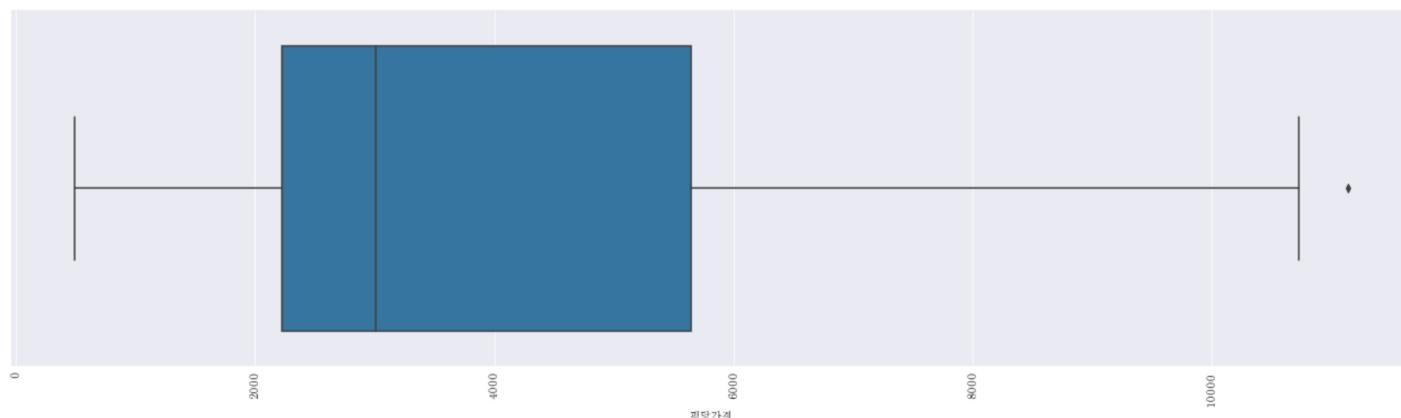


### 한남동 평당가격 Boxplot

Hannam-dong Sale Price Per 3.3m<sup>2</sup> Boxplot

아래는 상자 그림(Box Plot)으로 한남동 평당가격 데이터를 표시한 것입니다.

```
In [13]: plt.subplots(figsize=(23, 6))
sns.boxplot(x='평당가격', data=hannam)
plt.xticks(rotation=90);
```



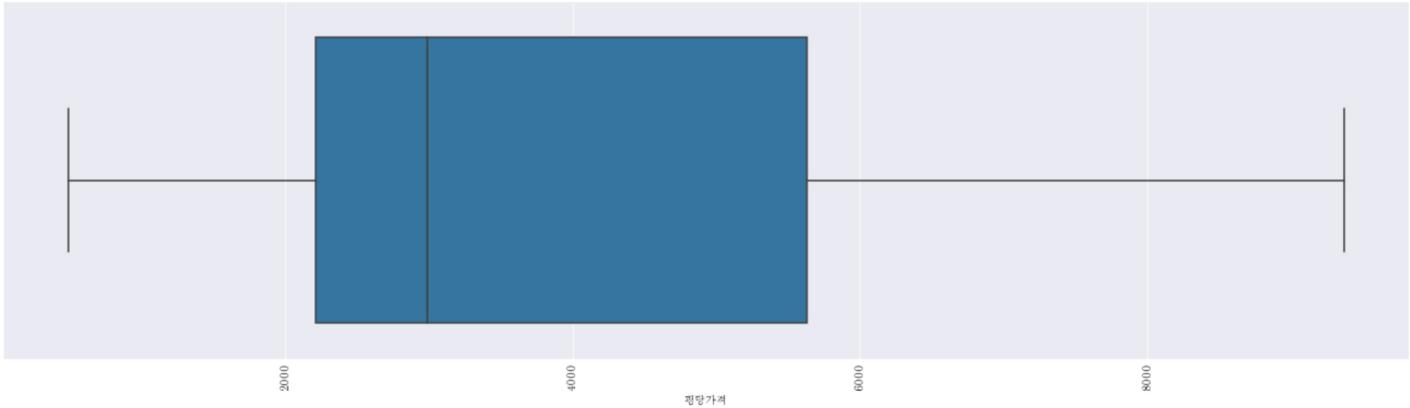
## 한남동 평당가격 (Outlier 제거) Boxplot

Hannam-dong Sale Price Per 3.3m<sup>2</sup> (Outliers Removed) Boxplot

이상점 제거 데이터로 상자 그림(Box Plot)을 만들었습니다.

최댓값이 많이 감소하고 극한값이 제거된 걸 볼 수 있습니다.

```
In [14]: plt.subplots(figsize=(23, 6))
sns.boxplot(x='평당가격', data=hannam_outliergone)
plt.xticks(rotation=90);
```



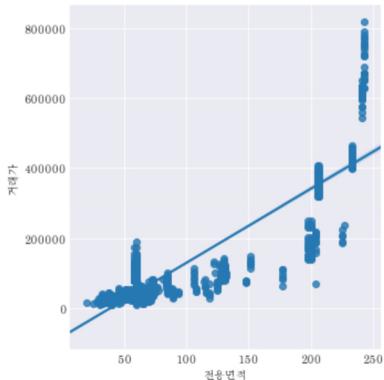
## 한남동 전용면적 vs 거래가 Pointplot

Hannam-dong Area for Exclusive Use vs Sale Price Pointplot

한남동 전용면적과 거래가의 상관관계를 나타내는 Pointplot입니다.

```
In [15]: sns.lmplot(x='전용면적', y='거래가', data=hannam)
```

```
Out[15]: <seaborn.axisgrid.FacetGrid at 0x1b0ba5ccf8>
```



## 한남동 전용면적 vs 거래가 (Outlier 제거) Pointplot

Hannam-dong Area for Exclusive Use vs Sale Price (Outliers Gone) Pointplot

위와 비교했을 때 오른쪽 상단에 있는 데이터 포인트들이 이상점으로 계산되어 제거되었습니다.

이상점 제거로 한남동의 가격이 높은 실거래가 제거된 사실을 확인하였습니다.

전체 데이터를 갖고 더 자세한 분석을 할 때나 기계학습(Machine Learning) 모델을 만들 때 데이터의 이상점 제거가 필요한 경우가 있습니다.

이상점 제거를 진행할 때 이 사례연구를 참고하면 도움이 될 것으로 보입니다.

```
In [16]: sns.lmplot(x='전용면적', y='거래가', data=hannam_outliergone)
```

```
Out[16]: <seaborn.axisgrid.FacetGrid at 0x1b0e7cad30>
```

