AIVLE

AI 개발자 트랙 미니프로젝트 2차

신규 아파트 주차 수요 예측

AI 1반 1조

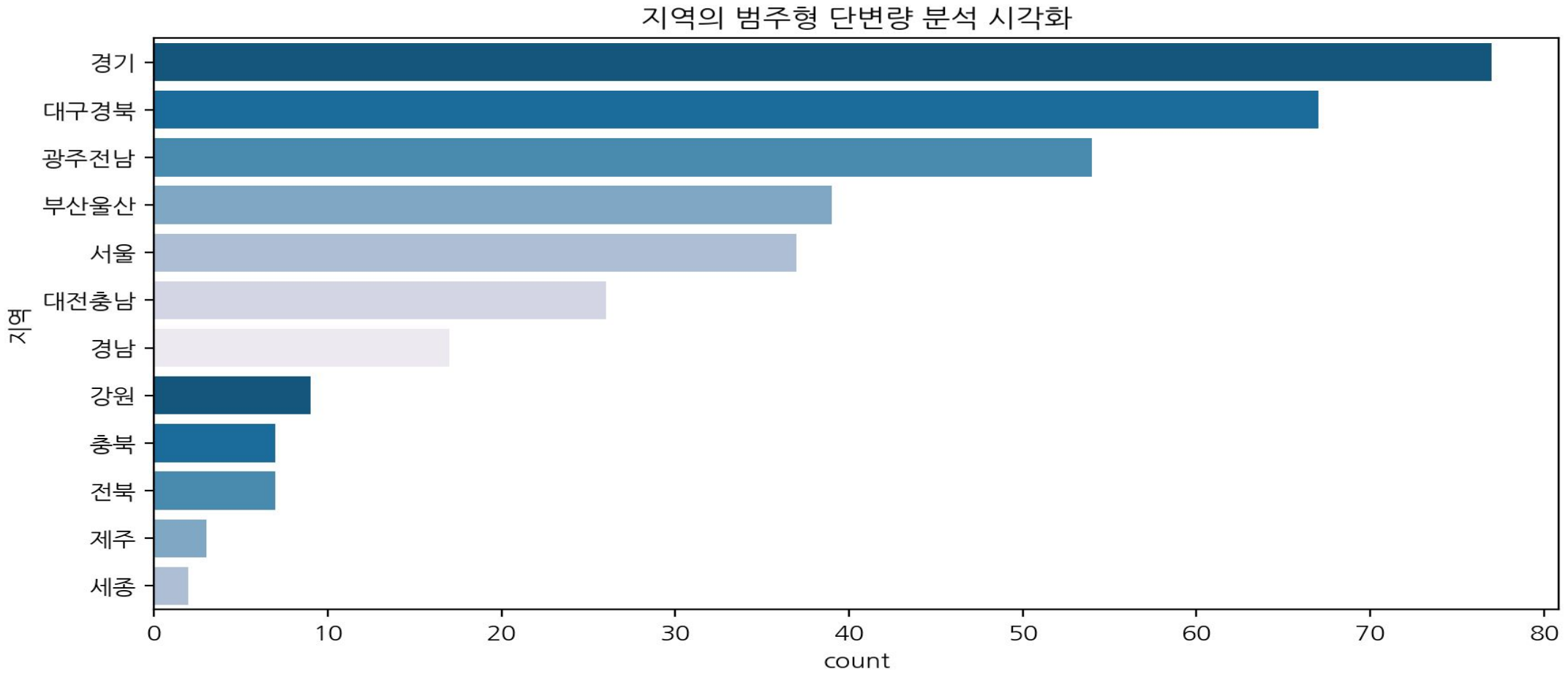
김예은, 김태헌, 박정은, 박지성, 정재원, 조강윤, 황유성

AIVLE

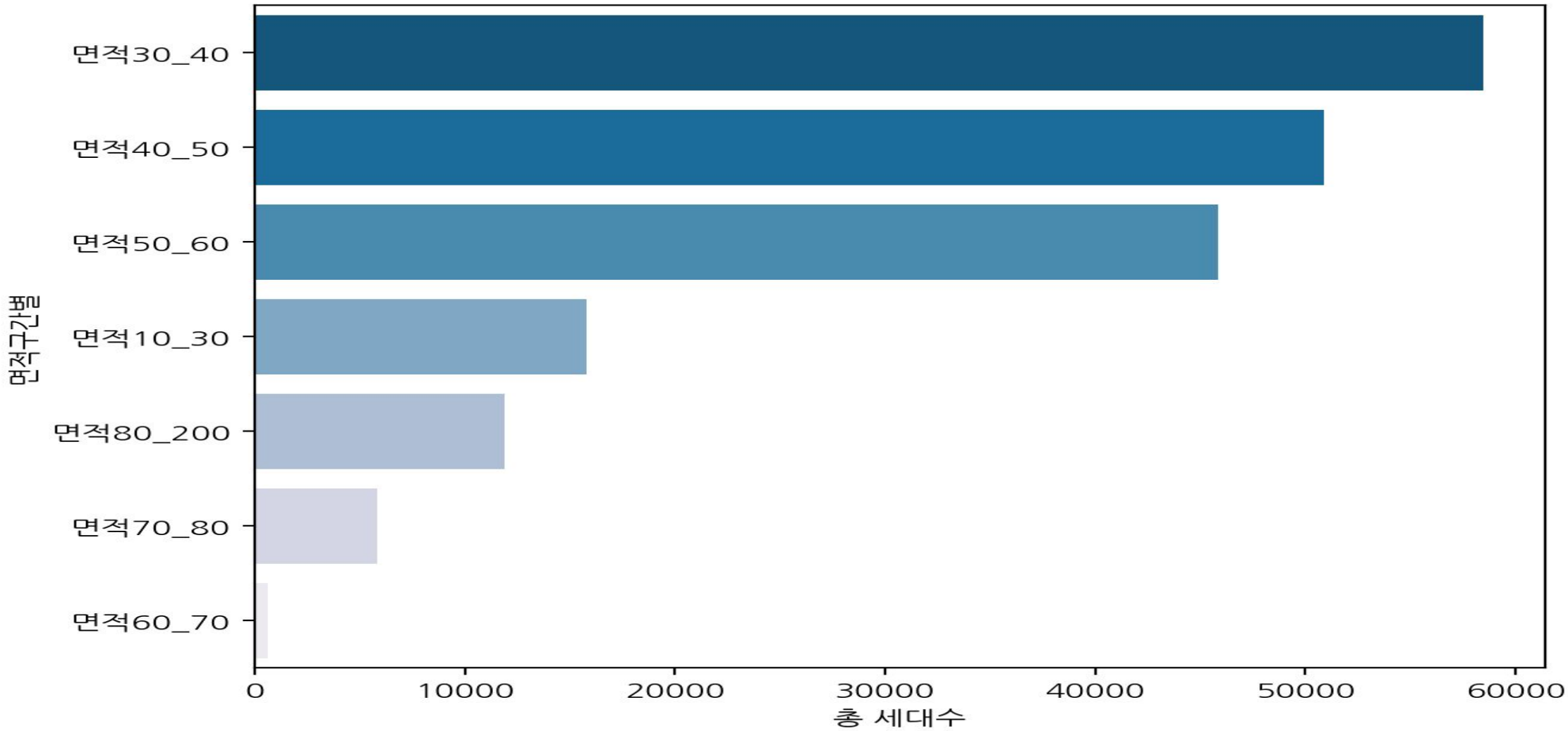
Make it possible

목차

데이터 분석 : 단변량 분석

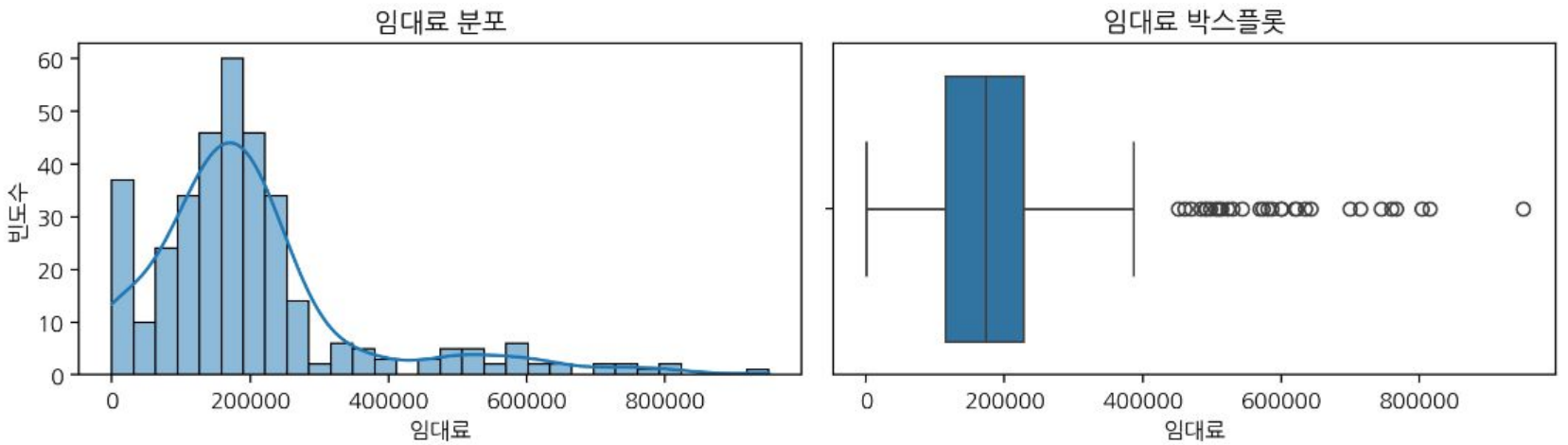

지역의 범주형 단변량 분석 시각화
지역
경기
대구경북
광주전남
부산울산
서울
대전충남
경남
강원
충북
전북
제주
세종
0
10
20
30
40
50
60
70
80
count

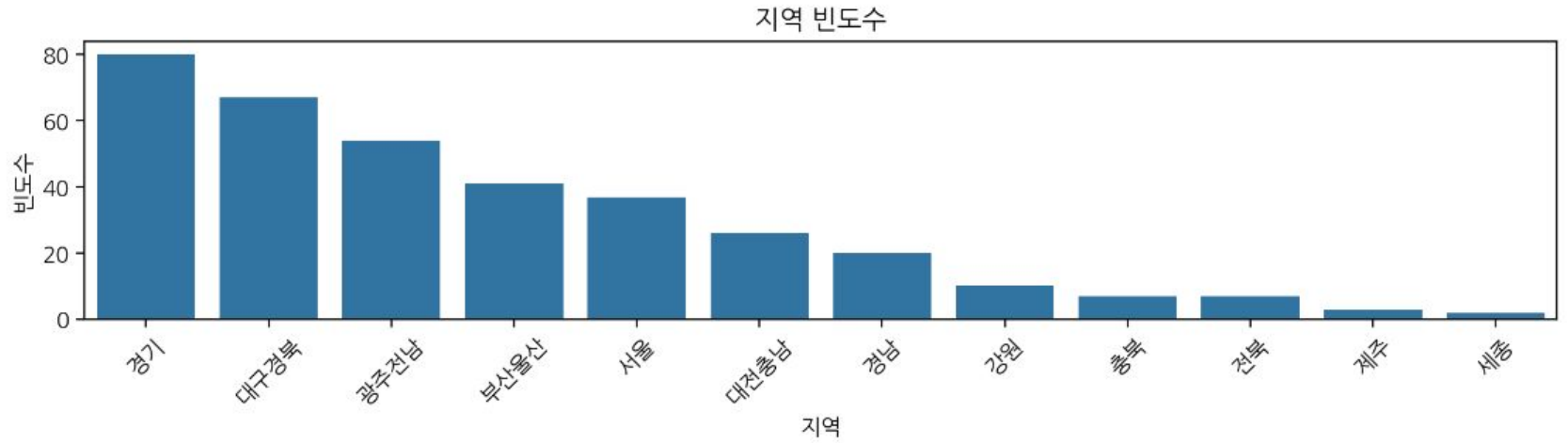
데이터 분석 : 단변량 분석


면적30_40
면적40_50
면적50_60
면적10_30
면적80_200
면적70_80
면적60_70
면적구간별
0
10000
20000
30000
40000
50000
60000
총 세대수

데이터 분석 : 단변량 분석

- 연속형과 범주형의 차이에 따라 단변량 분석 진행.





- **연속형**: '실차량수', '총세대수', '총면적', '준공연도', '임대료', '임대보증금'

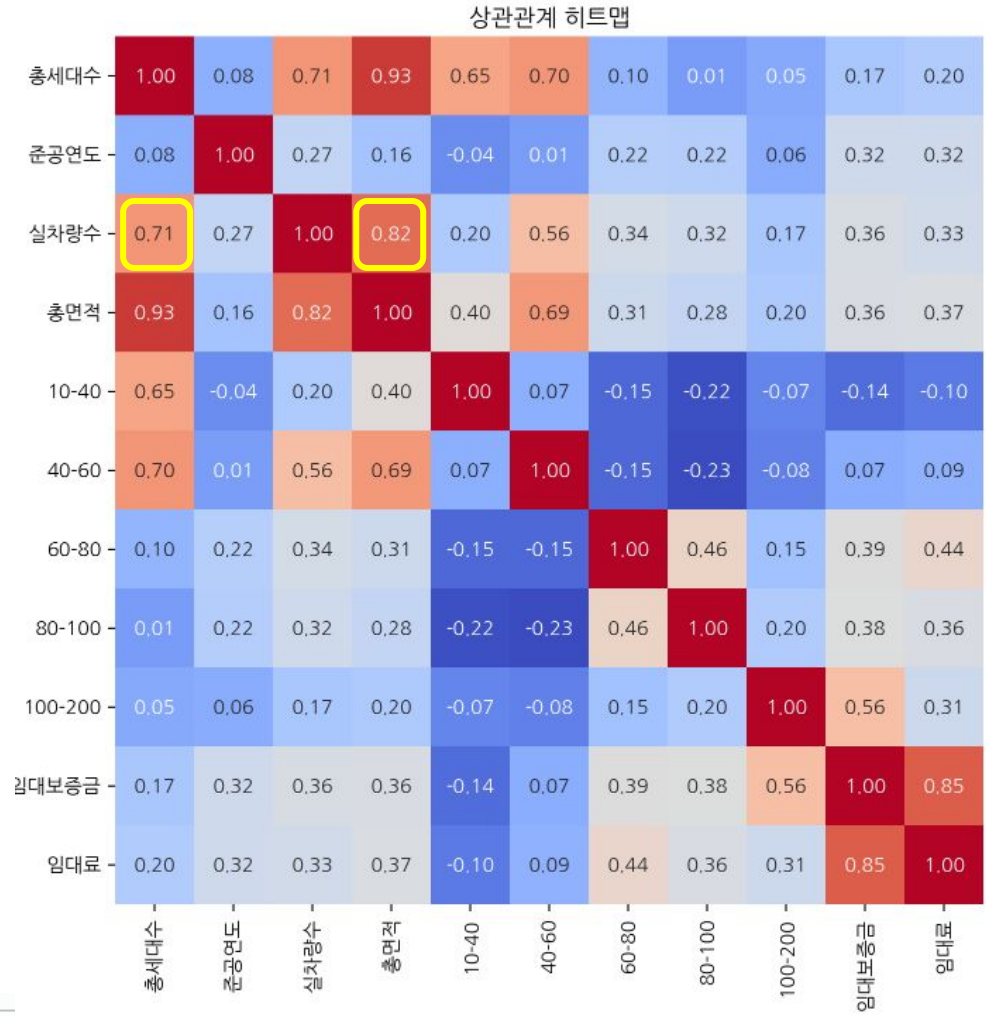
  →기술통계, Histogram, Boxplot 이용해 단변량 분석

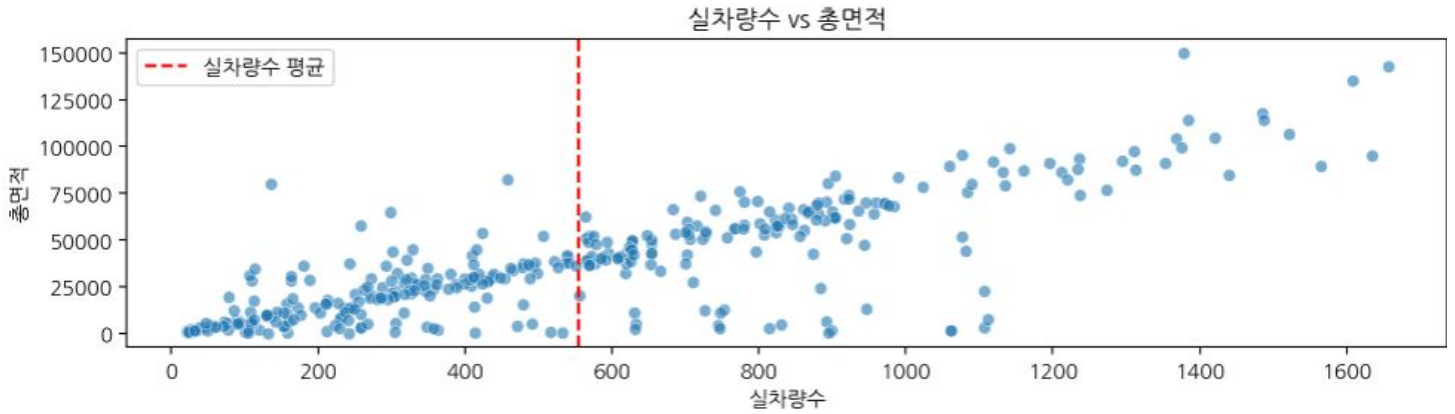
- **범주형**: '지역', '건물형태', '난방방식','승강기설치여부'

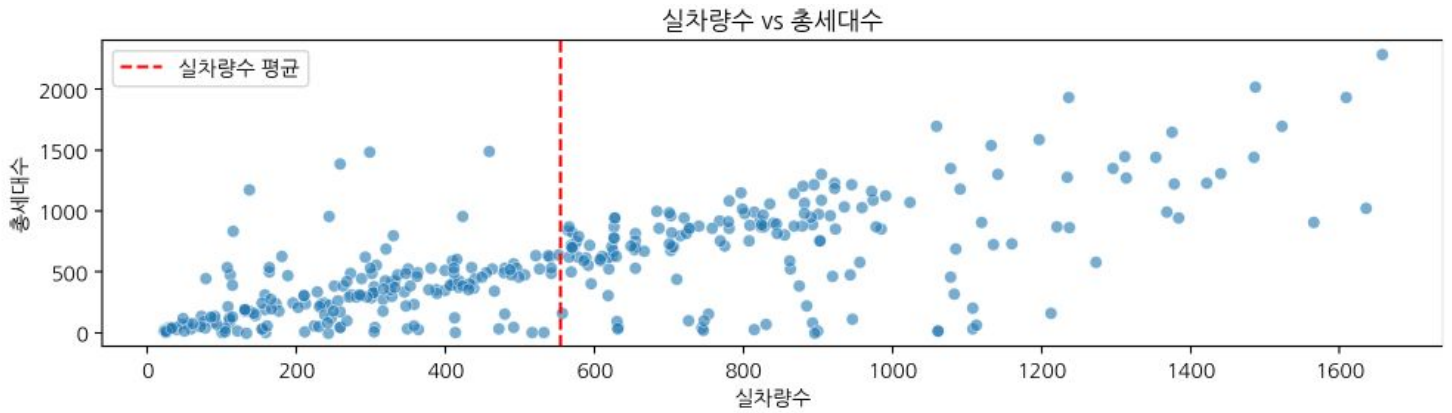
  → arplot 이용해 단변량 분석

데이터 분석 : 이변량 분석

실차량수와 총세대수, 총면적의 상관계수 값이 높은 것을 확인

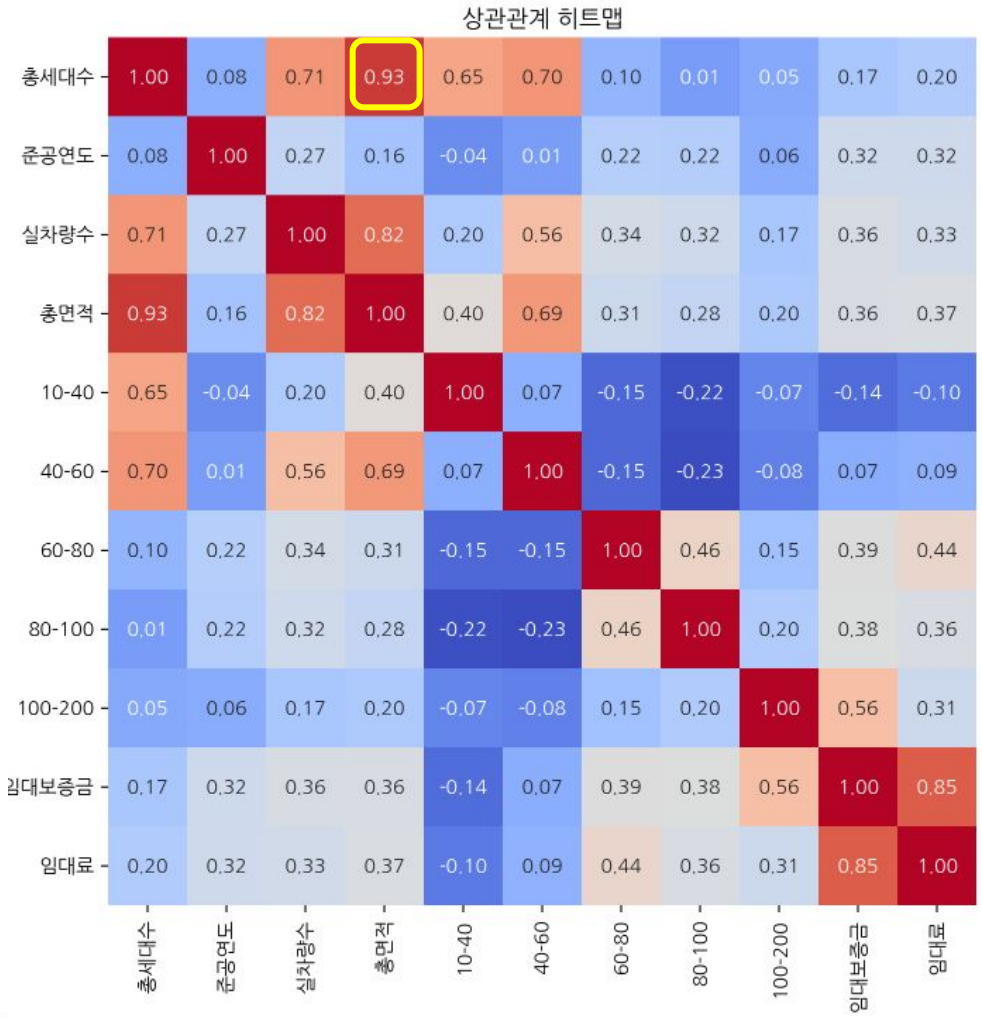






데이터 분석 : 이변량 분석

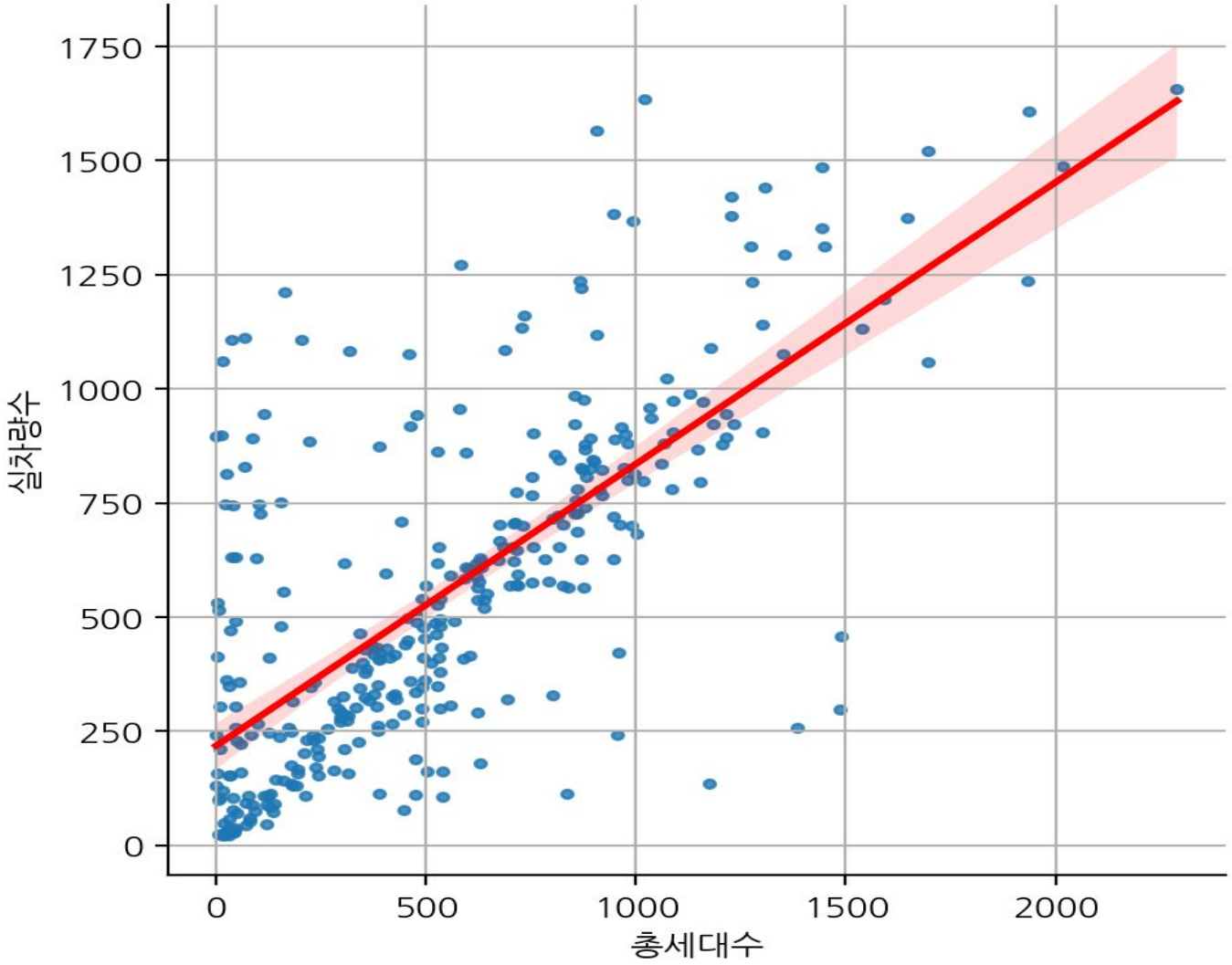
실차량수와 총세대수, 총면적의 상관계수 값이 높은 것을 확인



하지만, 총세대수와 총면적과의 상관관계가 높게 나옴

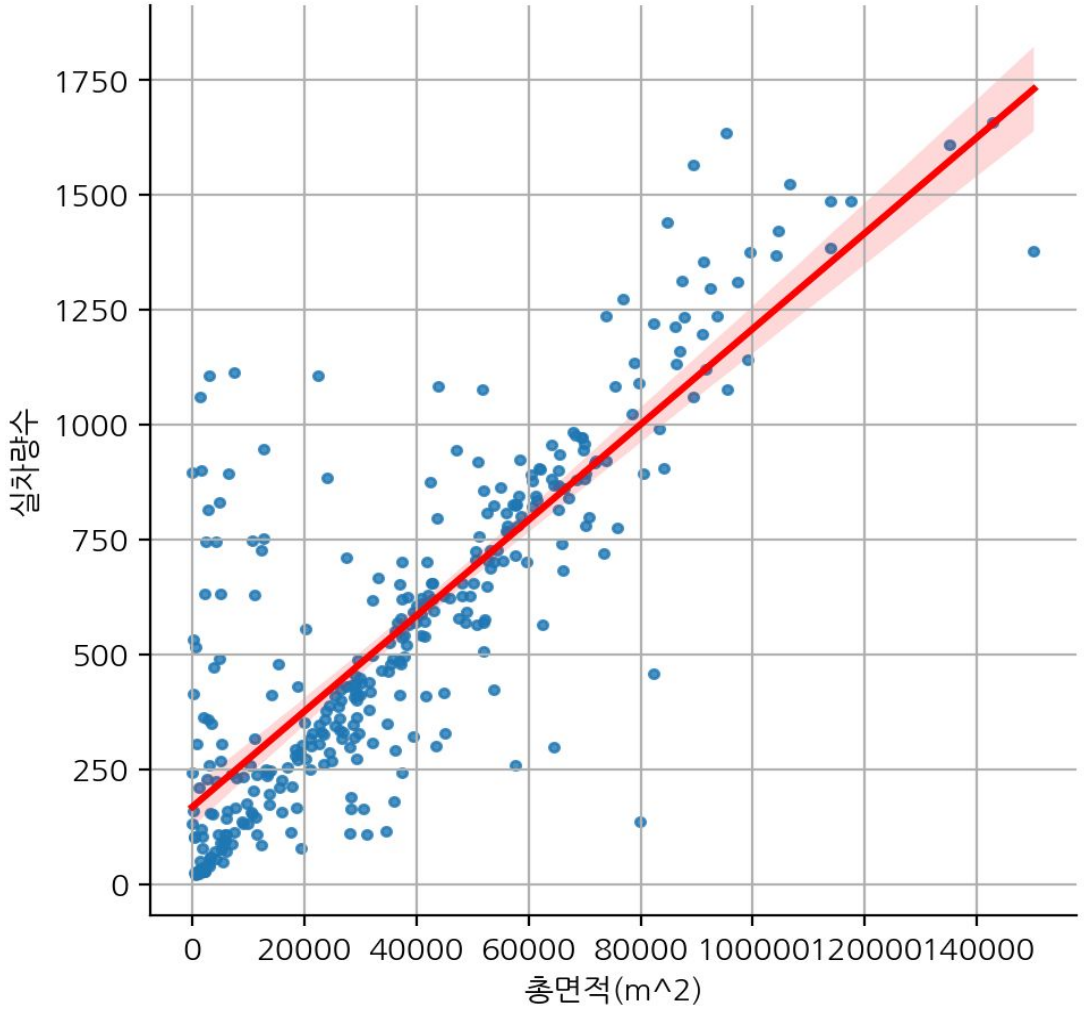
다중 공선성을 주의할 필요성을 느꼈습니다.

데이터 분석 : 이변량 분석



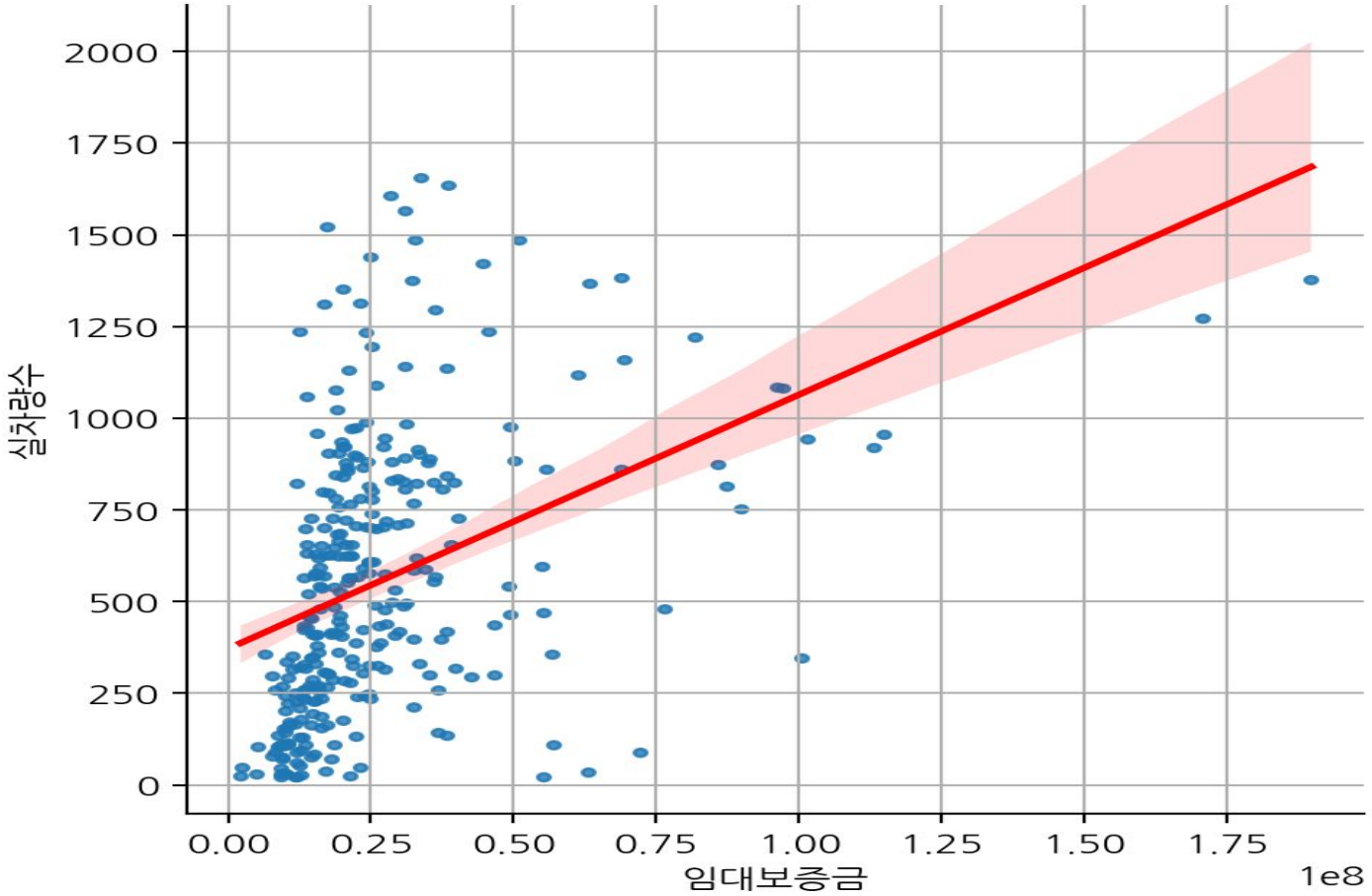
귀무가설을 기각합니다. (즉, 대립가설을 채택 : 두 변수와의 상관관계가 존재합니다.)
P-value: 1.0221045732952228e-54입니다.
두 변수의 상관관계는 0.7124746462088557 입니다.

데이터 분석 : 이변량 분석



귀무가설을 기각합니다. (즉, 대립가설을 채택 : 두 변수와의 상관관계가 존재합니다.)
P-value: 6.032798759323943e-86입니다.
두 변수의 상관관계는 0.8221825549027681 입니다.

데이터 분석 : 이변량 분석

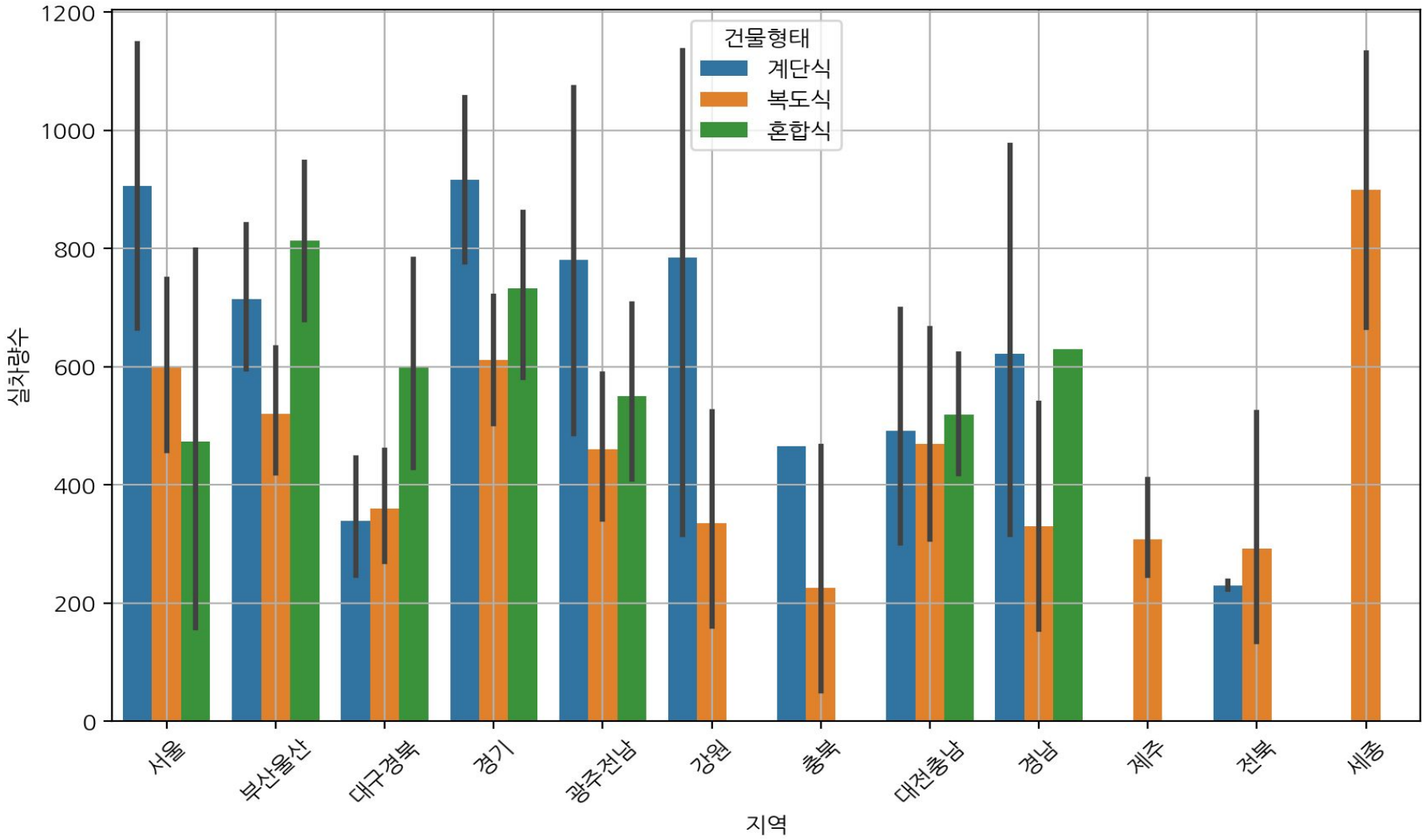


임대보증금은 등분산성을 만족하지 않는다.

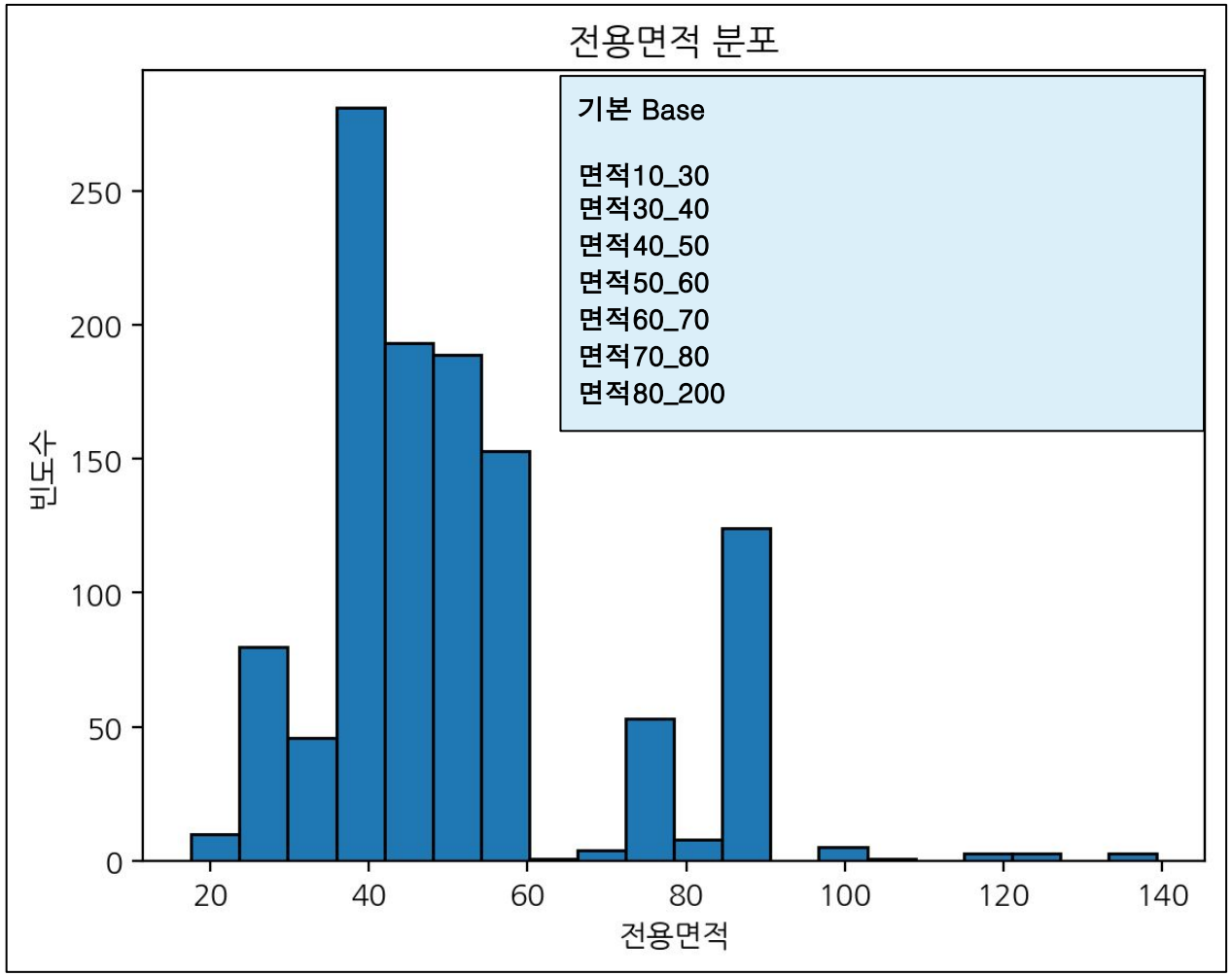
등분산성이란 예측변수의 모든 값에서 회귀선과 예측값의 분산이 일정해야함을 의미함

임대 보증금이 커질수록 실차량수(내가 차를 선택하지 않을수도 있고 차를 선택해서 구매할 수도 있는 그러한 선택지가 커짐을 알 수 있다.)

데이터 분석 : 이변량 분석


건물형태
계단식
복도식
혼합식
실차량수
1200
1000
800
600
400
200
0
서울
부산울산
대구경북
경기
광주전남
강원
충북
대전충남
경남
제주
전북
세종
지역

데이터 전처리 : 전용면적구간 나누기



1. **분포모양**
   - 대체로 40-60에 분포

2. **비대칭성**
   - 분포가 왼쪽으로 치우쳐져 있으며, 대부분의 데이터가 상대적으로 낮은 전용면적을 가짐

3. **이상치**
   - 100 이상의 이상치가 있으며, 이러한 데이터 포인트는 대형 주택이나 특수한 경우를 뜻함

데이터에 기반한 전용면적구간 나눔
10-40, 40-60, 60-80, 80-100, 100-200

일관된 간격으로 나눌 경우 데이터의 실제 분포를 반영하지 못 할 수 있음

따라서 모델 성능 향상 효과를 기대하며,
전용면적구간을 나누었다.

데이터 전처리 : 결측치 처리

KNN 알고리즘을 사용한 이유?

```
data['난방방식'].value_counts()

난방방식
개별가스난방     210
지역난방         84
지역가스난방      30
중앙가스난방      14
중앙난방          3
중앙유류난방      2
개별유류난방      1
지역유류난방      1
Name: count, dtype: int64
```

```
data['건물형태'].value_counts()

건물형태
복도식     203
계단식     104
혼합식      47
Name: count, dtype: int64
```

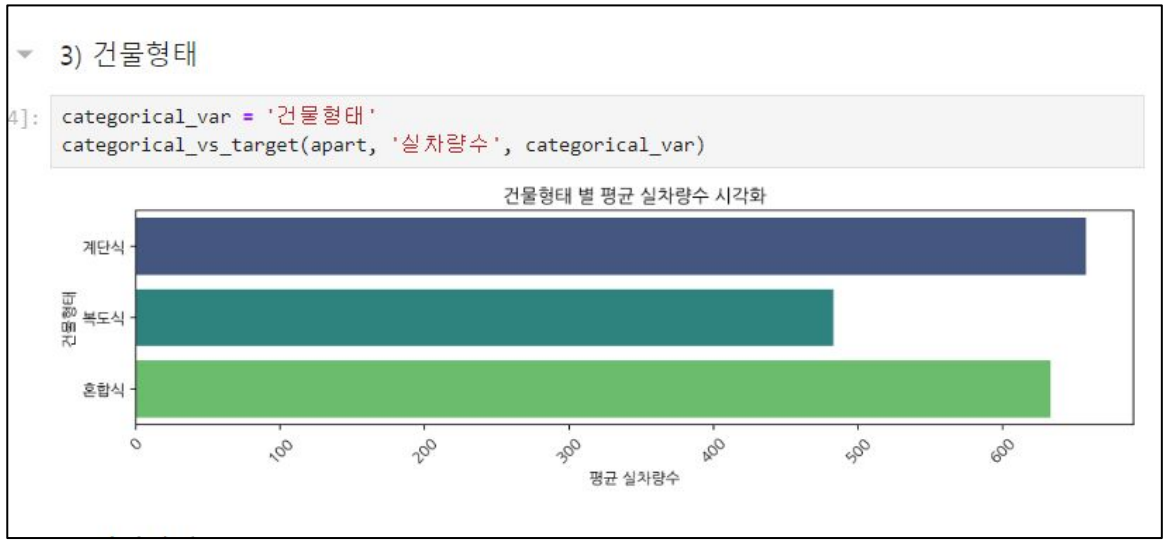
최빈값을 사용해서 결측치를 처리하기 보다 분류 알고리즘을 사용해서 결측치 처리하자는 아이디어

데이터 전처리 : 결측치 처리

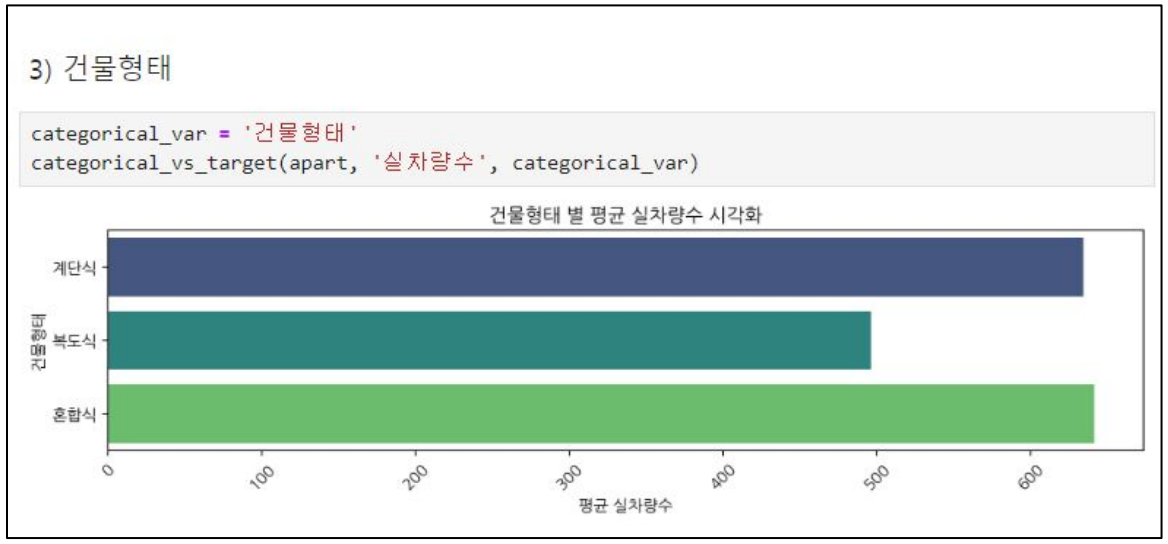
결측치 발생 칼럼 : 건물형태, 난방방식, 승강기설치여부

승강기설치여부 : 최빈값 활용
건물형태, 난방방식 : KNN 알고리즘 활용 결측치 처리

KNN을 이용한 건물형태 결측치 처리

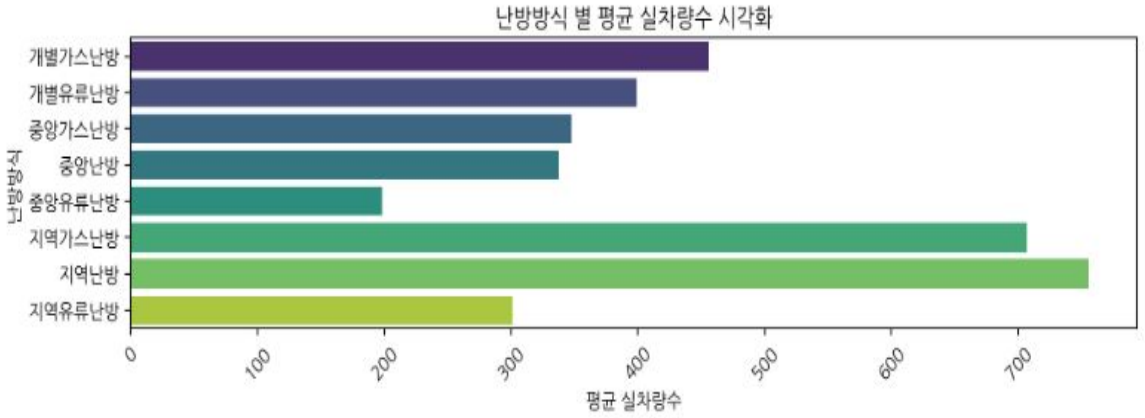


최빈값을 활용한 건물형태 결측치 처리



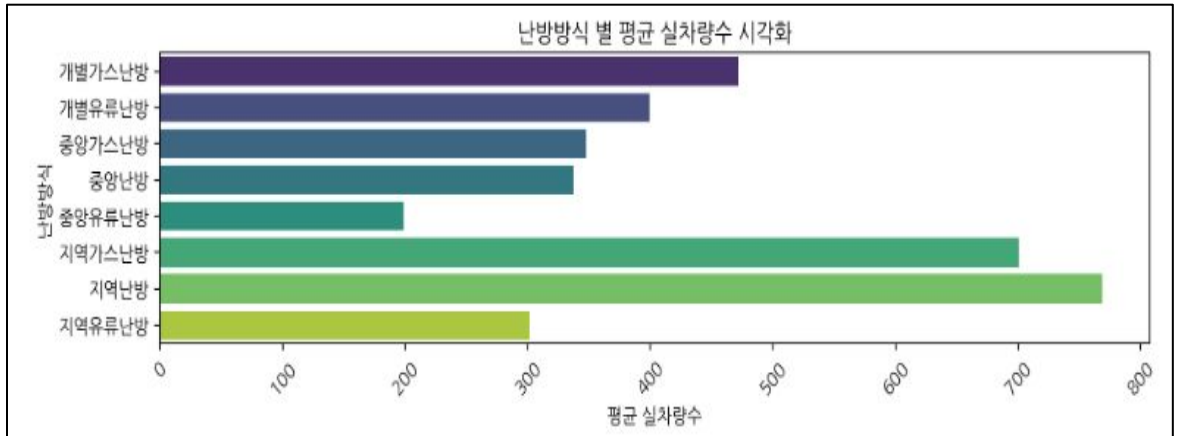
데이터 전처리 : 결측치 처리

KNN을 이용한 난방방식 결측치 처리




```
난방방식
지역난방       755.787879
지역가스난방    706.838710
개별가스난방    456.162562
개별유류난방    399.000000
중앙가스난방    347.428571
중앙난방       337.666667
지역유류난방    301.000000
중앙유류난방    198.500000
Name: 실차량수, dtype: float64
```

최빈값을 활용한 난방방식 결측치 처리




```
난방방식
지역난방       768.821429
지역가스난방    700.333333
개별가스난방    471.709524
개별유류난방    399.000000
중앙가스난방    347.428571
중앙난방       337.666667
지역유류난방    301.000000
중앙유류난방    198.500000
Name: 실차량수, dtype: float64
```

데이터 전처리 : 결측치 처리

KNN 알고리즘을 이용한 결측치 처리

2) 기본 정보 조회

```
apart.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 354 entries, 0 to 353
Data columns (total 18 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   단지코드     354 non-null    object
 1   총세대수     354 non-null    int64
 2   지역       354 non-null    object
 3   준공연도     354 non-null    int32
 4   건물형태     354 non-null    object
 5   난방방식     354 non-null    object
 6   승강기설치여부  354 non-null    object
 7   실차량수     354 non-null    int64
 8   총면적      354 non-null    float64
 9   10-30    354 non-null    int64
 10  30-40    354 non-null    int64
 11  40-50    354 non-null    int64
 12  50-60    354 non-null    int64
 13  60-70    354 non-null    int64
 14  70-80    354 non-null    int64
 15  80-200   354 non-null    int64
 16  임대보증금    354 non-null    float64
 17  임대료      354 non-null    float64
dtypes: float64(3), int32(1), int64(9), object(5)
memory usage: 48.5+ KB
```

"9개"

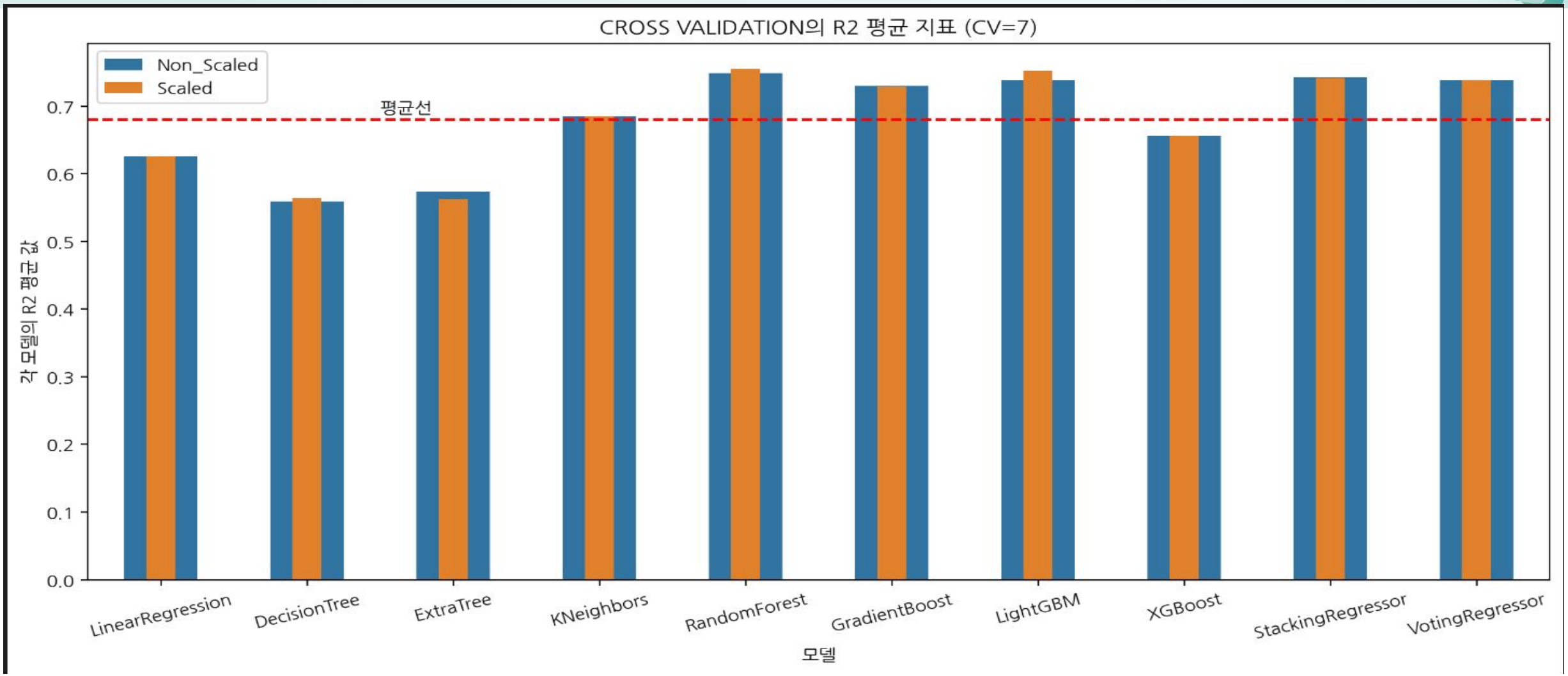
최빈값을 활용한 결측치 처리

2) 기본 정보 조회

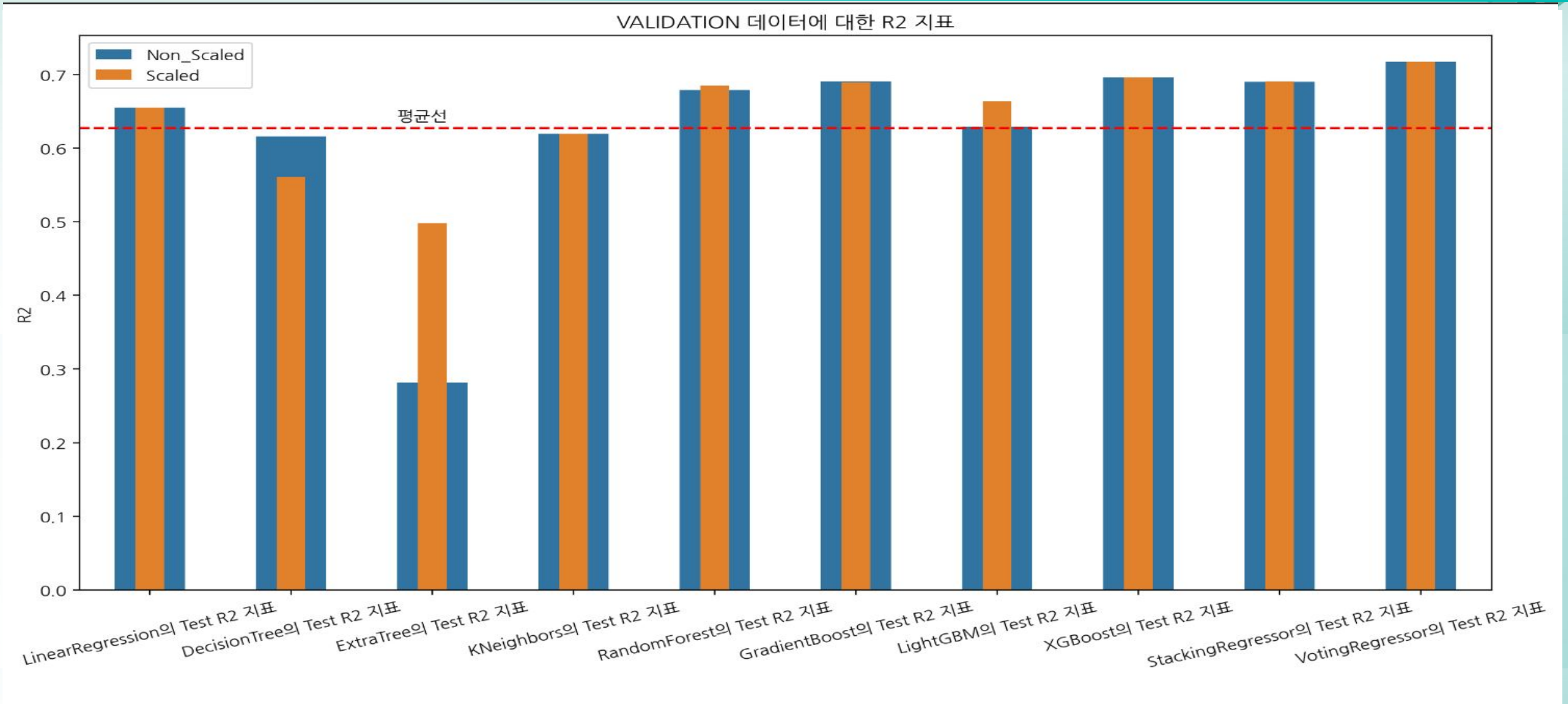
```
apart.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 345 entries, 0 to 344
Data columns (total 18 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   단지코드     345 non-null    object
 1   총세대수     345 non-null    int64
 2   지역       345 non-null    object
 3   준공연도     345 non-null    int32
 4   건물형태     345 non-null    object
 5   난방방식     345 non-null    object
 6   승강기설치여부  345 non-null    object
 7   실차량수     345 non-null    int64
 8   총면적      345 non-null    float64
 9   10-30    345 non-null    int64
 10  30-40    345 non-null    int64
 11  40-50    345 non-null    int64
 12  50-60    345 non-null    int64
 13  60-70    345 non-null    int64
 14  70-80    345 non-null    int64
 15  80-200   345 non-null    int64
 16  임대보증금    345 non-null    float64
 17  임대료      345 non-null    float64
dtypes: float64(3), int32(1), int64(9), object(5)
memory usage: 47.3+ KB
```

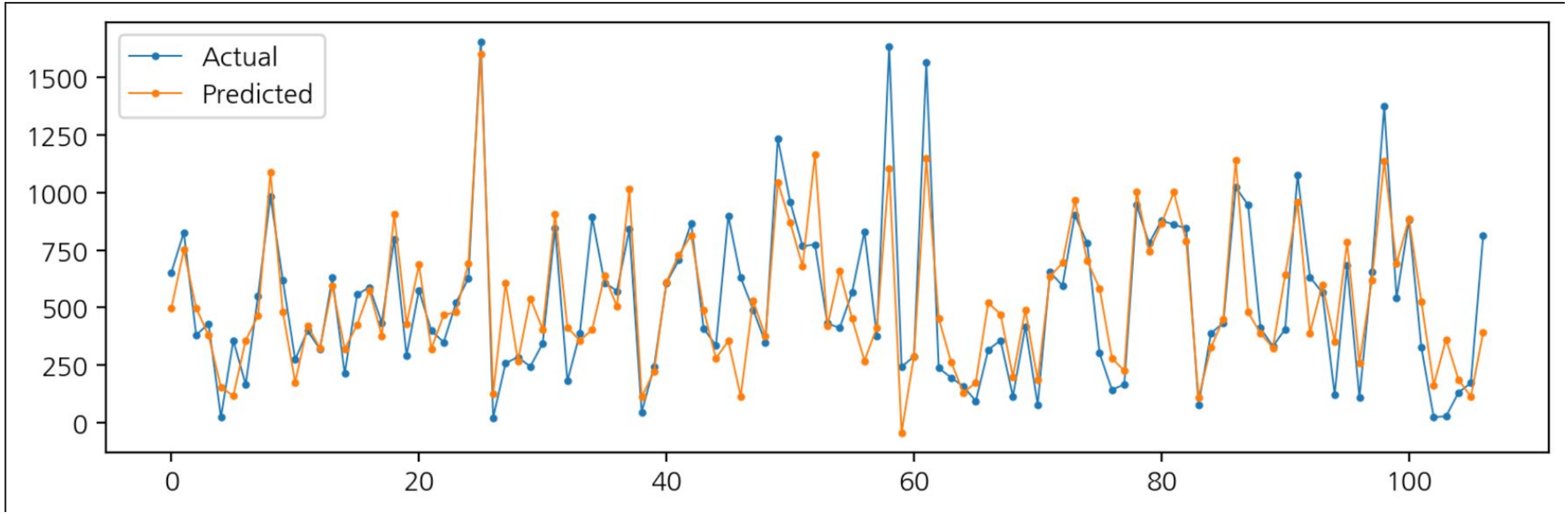
머신러닝 모델링 : 기존 BaseModel



머신러닝 모델링 : 기존 BaseModel

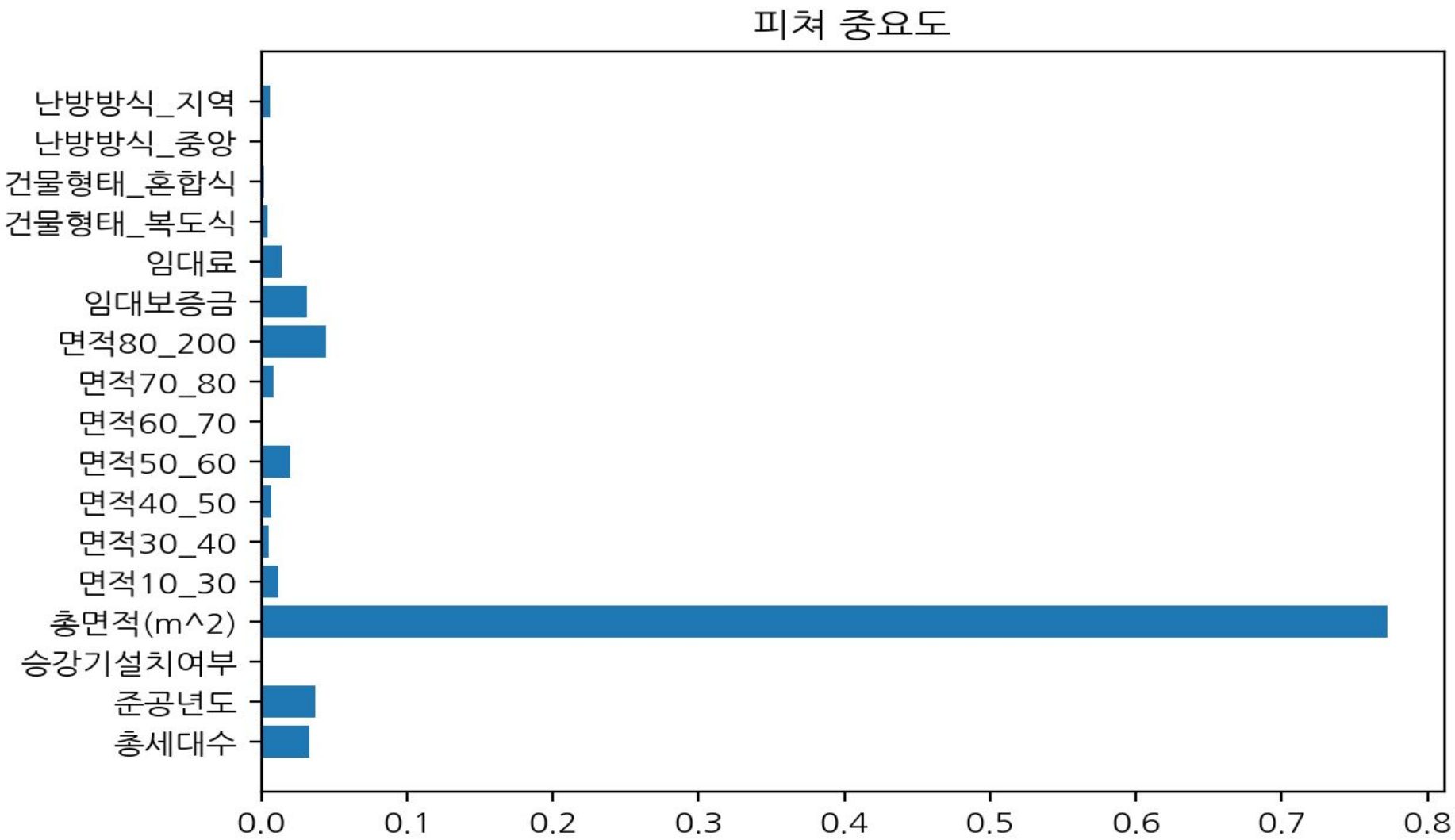


머신러닝 모델링 : Testing Model




```
Stacking Mean Squared Error: 28622.52
Stacking Root Mean Squared Error: 169.18
Stacking R-squared: 0.76
```

결과: 기존 BaseModel



결론

- 전용면적구간 10-40, 40-60, 60-80, 80-100, 100-200 으로 나누고, 결측치를 KNN으로 처리한 후 모델링함.

- LinearRegression, XGBoost, RandomForest 앙상블해서 나온 결과가 가장 성능이 좋았음. (StackingRegressor)

- Stacking RMSE : 169.18, Stacking R-squared : 0.76

	단지코드	단지명	총세대수	지역	예상차량수
0	C0005	서울석촌 도시형주택(공임10년)	20	서울	199
1	C0017	대구혁신센텀힐즈	822	대구경북	498
2	C0034	광교 호반마을 22단지 아파트	112	경기	518
3	C0046	죽미마을 휴먼시아12단지	122	경기	543
4	C0055	파주운정 한울마을 6단지	262	경기	614
5	C0072	벌교제석휴먼시아	35	광주전남	501
6	C0073	무안남악오룡마을	47	광주전남	300
7	C0084	진해 석동우림필유 아파트	152	경남	651
8	C0085	김해북부 두산위브 아파트	73	경남	576
9	C0112	운암주공6단지아파트	571	경기	497
10	C0114	가평읍내주공아파트	355	서울	292
11	C1149	부산장안A-1BL	96	부산울산	151
12	C0154	정관신도시휴먼시아1단지	1533	부산울산	1527
13	C0159	율하휴먼시아 7단지	712	대구경북	641
14	C0174	대구서재휴먼시아	757	대구경북	507
15	C0177	인천소래 휴먼시아 1단지	882	경기	730
16	C0182	동양주공아파트2단지	668	경기	604
17	C0200	매곡부림2차아파트	62	광주전남	152
18	C0225	신완마을 휴먼시아 6단지	1116	광주전남	1013
19	C0258	수원호매실 휴먼시아 8단지 아파트	1270	경기	896
20	C0267	신곡주공3단지	322	서울	307
21	C0272	용암마을13단지	996	서울	835
22	C0276	동편마을1단지	1017	경기	747
23	C0285	소사벌휴먼시아2단지	992	경기	837
24	C0286	소사벌 이곡마을3단지	1191	경기	893
25	C2306	양주옥정8	344	서울	264
26	C0309	진천이월	82	충북	192
27	C0352	대전판암4	2389	대전충남	1187
28	C0353	대전판암3	768	대전충남	383
29	C0360	하남미사 A26BL	588	서울	896

AIVLE