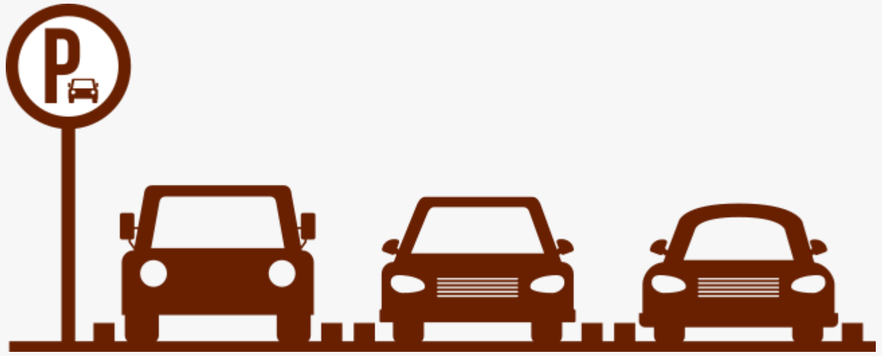
주차수요예측 프로젝트

마주연, 신해빈, 이서현, 안도현

CONTENTS



01 데이터 소개 및 EDA

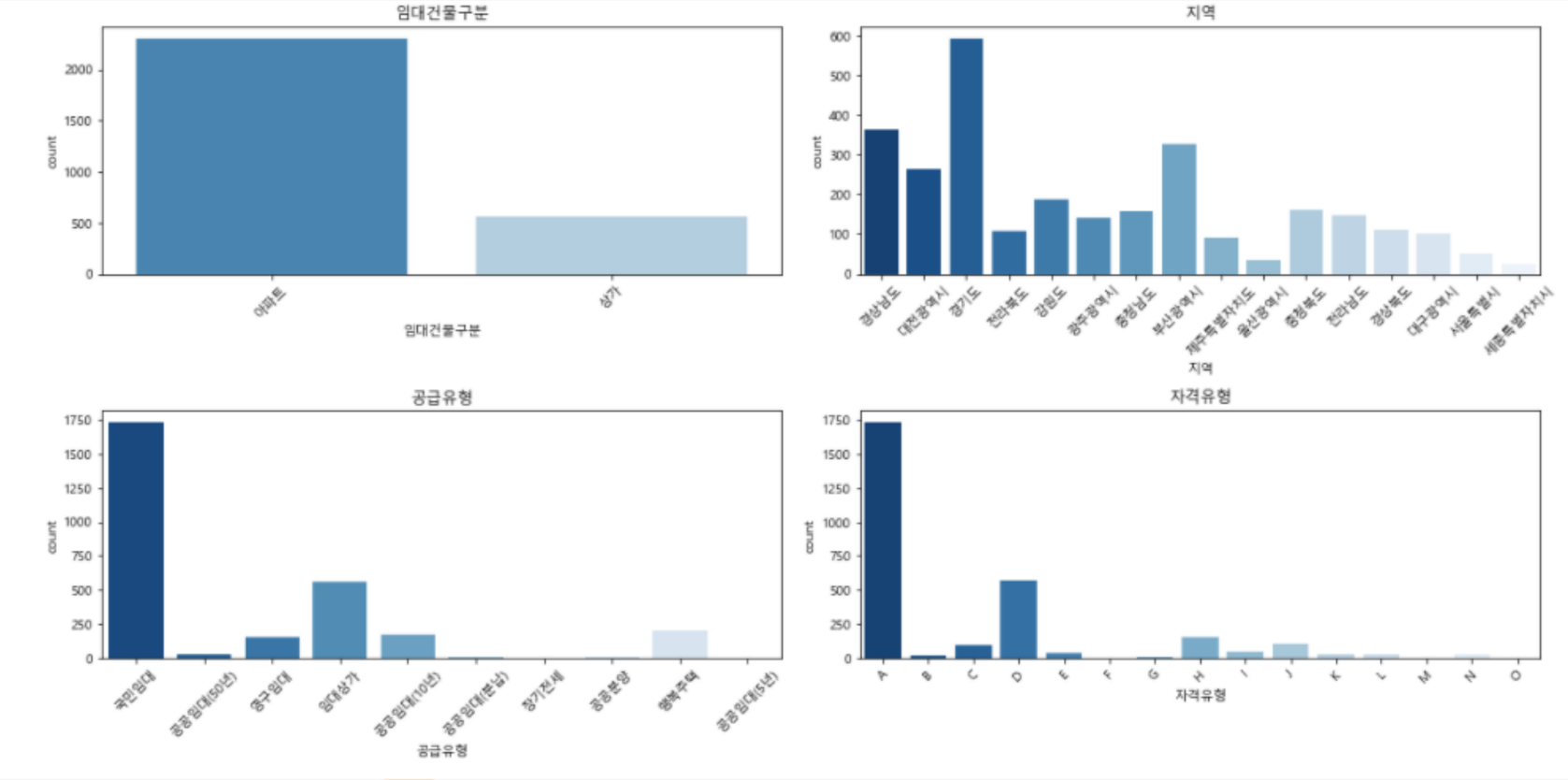
단지코드	자격유형
총세대수	임대보증금
임대건물구분	임대료
지역	지하철역수
공급유형	버스정류장수
전용면적	단지내주차면수
전용면적별세대수	등록차량수
공가수	

2549 x 15

유형별 임대주택 설계 시,
단지 내 적정 주차 수요를 예측

01 데이터 소개 및 EDA

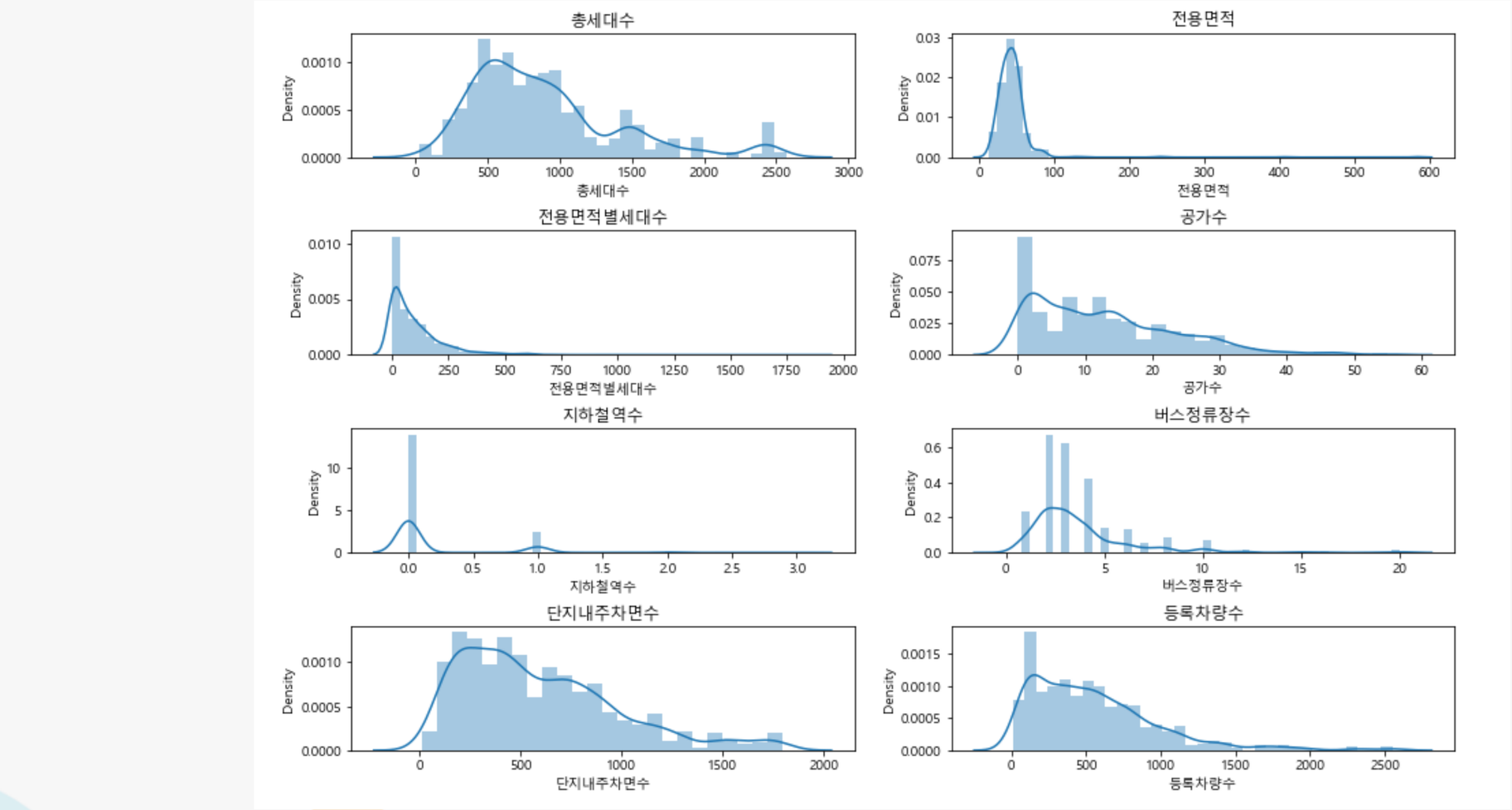
범주형변수확인



- 임대건물구분
- 지역
- 공급유형
- 자격유형

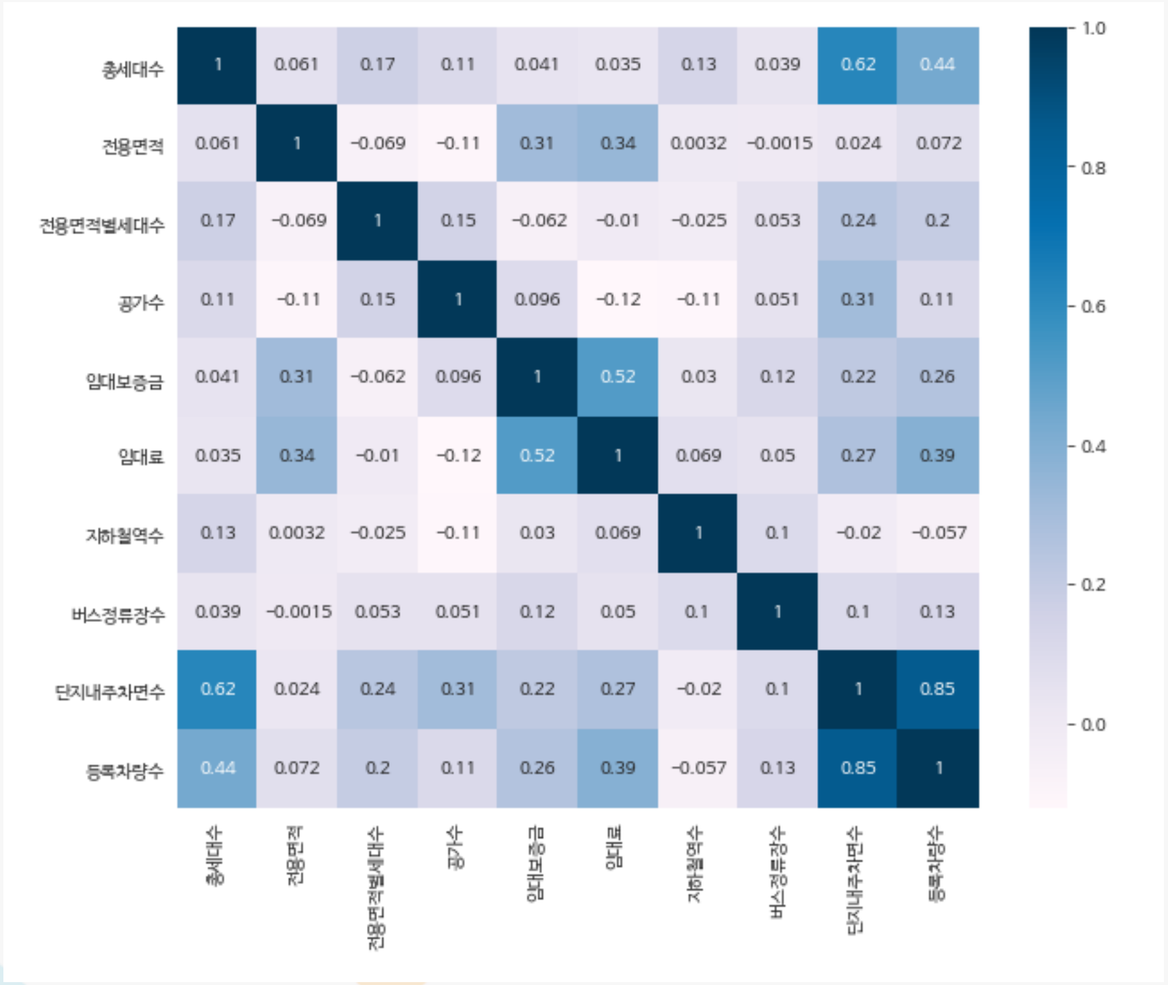
01 데이터 소개 및 EDA

연속형변수확인



01 데이터 소개 및 EDA

타겟변수와의 상관관계

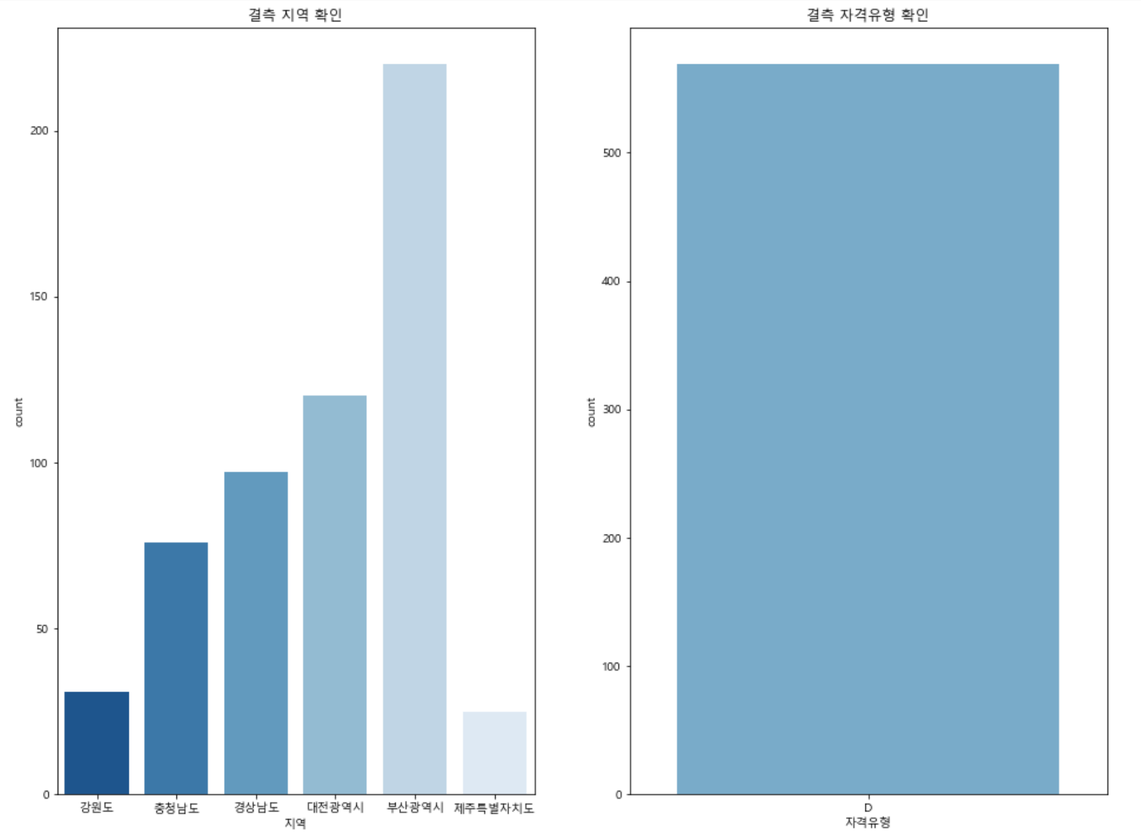


- 등록차량수 - 단지내 주차면수
- 등록차량수 - 총세대수
- 단지내 주차면수 - 총세대수
- 임대료 - 임대보증금

양의 상관관계를 보임

02 데이터 전처리

결측치의 지역과 자격유형 확인



결측치 지역확인

부산광역시 > 대전광역시 > 경상남도 > 충청남도 > 제주특별자치도

결측치 자격유형 확인

D (공공분양,임대상가)

02 데이터 전처리

결측치처리

☑ 임대보증금, 임대료 → LH /주택관리공단 홈페이지 이용하여 직접처리

(1) (Train) 부산 임대주택 결측치

(2) (Train) 강원도 행복주택 결측치

(3) (Train) 대구광역시 임대주택 결측치

(4) (Test) 부산 행복주택 결측치

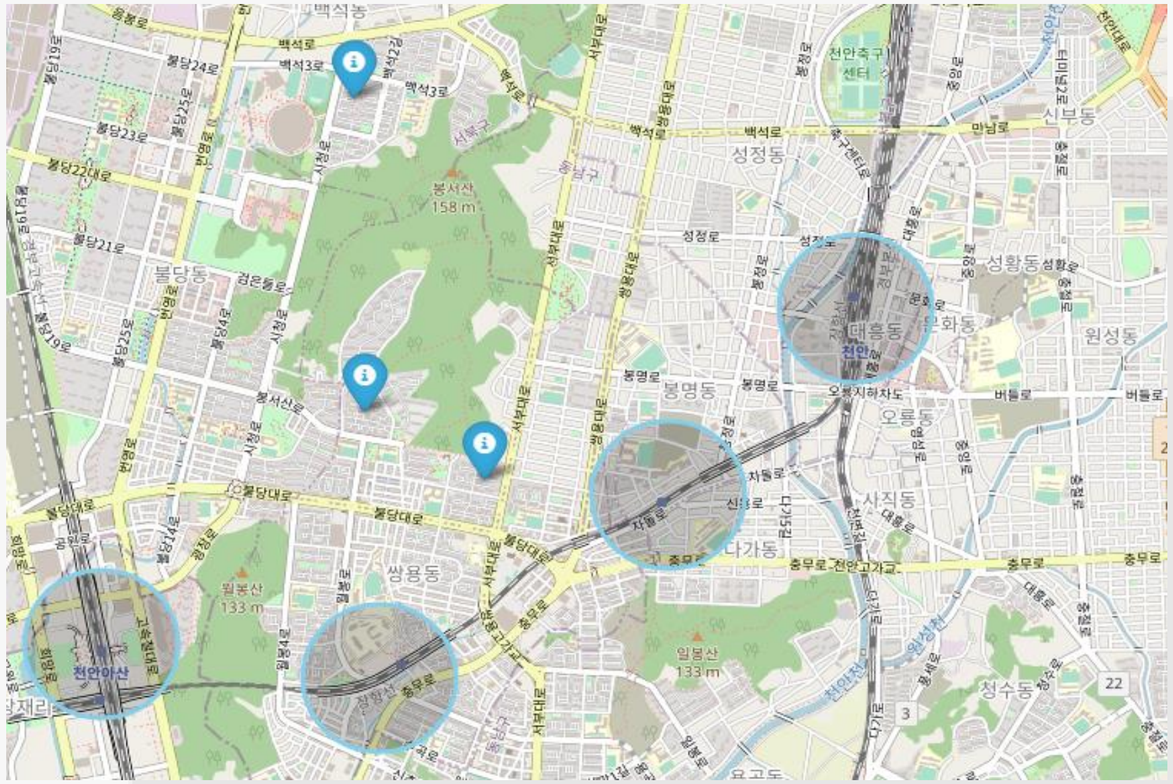
(5) (Test) 강원도 영구임대 결측치

(6) (Test) 대전광역시 영구임대 결측치

평형	임대보증금	임대료
24A,24B	7505500	114740
26	8148000	152250
37	15868000	212290
46	24768000	277690

02 데이터 전처리

결측치의 처리



지하철 결측치가 있는 지역: **충청남도, 대전광역시**

위도, 경도 데이터를 활용하여
지도상으로 인접한 지하철 역 존재여부 확인

결측치 모두 **0**으로 처리

02 데이터 전처리

결측치의 처리

	단지코드	총세대수	임대건물구분	지역	공급유형	전용면적	전용면적별세대수	공가수	자격유형	임대보증금	임대료	지하철역수	버스정류장수	단지내주차면수	등록차량수	평수
2284	C1350	1401	아파트	대전광역시	공공분양	74.94	317	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	32.0
2285	C1350	1401	아파트	대전광역시	공공분양	74.94	137	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	32.0
2286	C1350	1401	아파트	대전광역시	공공분양	74.94	22	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	32.0
2287	C1350	1401	아파트	대전광역시	공공분양	84.94	164	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	36.0
2288	C1350	1401	아파트	대전광역시	공공분양	84.94	19	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	36.0
2289	C1350	1401	아파트	대전광역시	공공분양	84.96	26	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	36.0
2290	C1350	1401	아파트	대전광역시	공공분양	84.97	26	2.0	D	NaN	NaN	NaN	6.0	1636.0	2315.0	36.0

결측치 중 임대건물구분이 **아파트**인
지역은 모두 **대전광역시**

공공분양은 임대가 아니므로
임대보증금과 임대료가 있을 필요X
0으로 처리

02 데이터 전처리

결측치의 처리

☑ 공급유형이 **공공분양**인 경우

☑ 공급유형이 **장기전세**인 경우

→ 결측치 모두 0으로 처리

02 데이터 전처리

결측치의 처리

임대상가 결측치처리 → **Random forest** 이용

	단지코드	총세대수	임대건물구분	지역	공급유형	전용면적	전용면적별세대수	공가수	자격유형	임대보증금	임대료	지하철역수	버스정류장수	단지내주차면수	등록차량수
80	C1925	601	상가	강원도	임대상가	32.10	1	9.0	D	NaN	NaN	0.0	4.0	117.0	75.0
83	C1925	601	상가	강원도	임대상가	72.16	1	9.0	D	NaN	NaN	0.0	4.0	117.0	75.0
93	C1874	619	상가	충청남도	임대상가	12.62	1	2.0	D	NaN	NaN	0.0	2.0	97.0	62.0
94	C1874	619	상가	충청남도	임대상가	17.40	1	2.0	D	NaN	NaN	0.0	2.0	97.0	62.0
96	C1874	619	상가	충청남도	임대상가	22.89	1	2.0	D	NaN	NaN	0.0	2.0	97.0	62.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

임대료

train_score: 0.985

test_score: 0.925

임대보증금

train_score: 0.983

test_score: 0.801

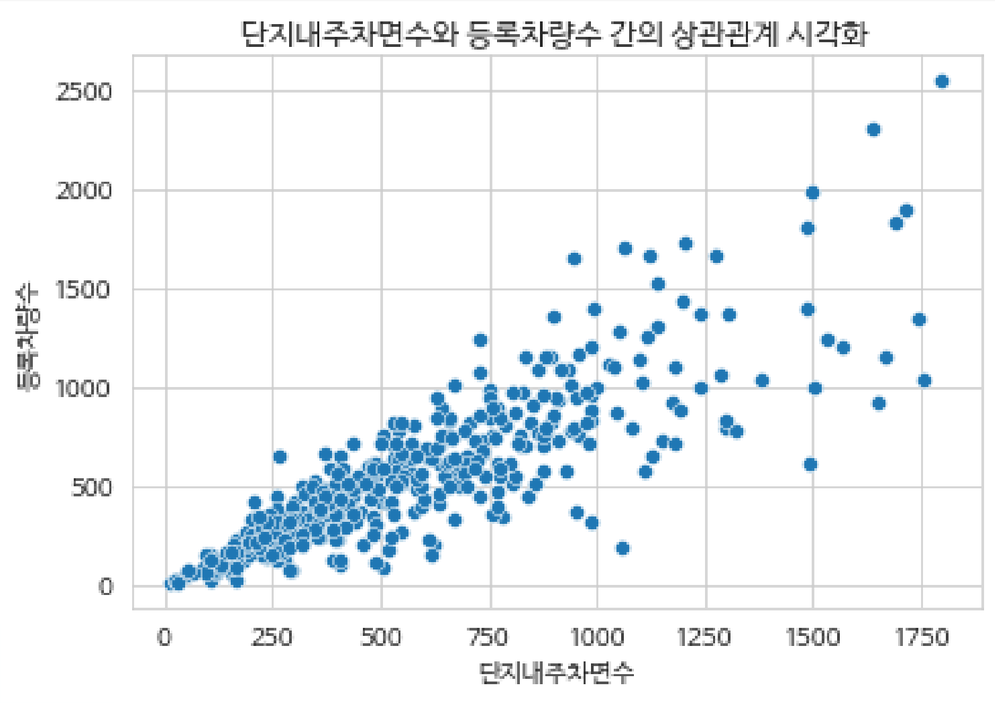
02 데이터 전처리

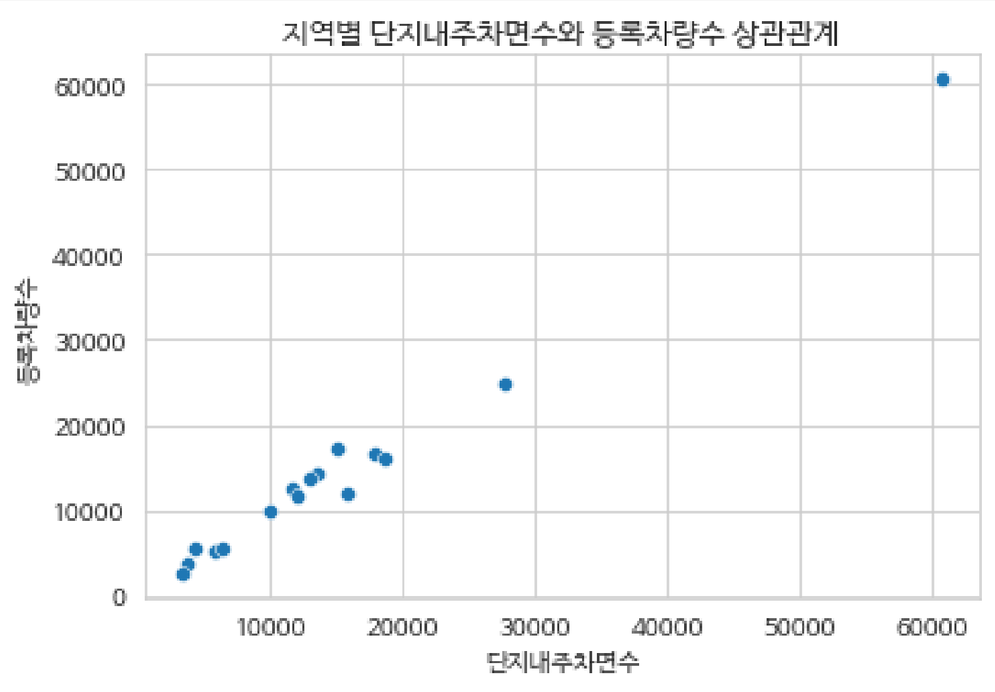
파생변수 생성

① 지역변수 - Clustering

- 지역 피쳐수가 많으므로 그룹화

Target 변수와 상관관계가 높은 변수 기준



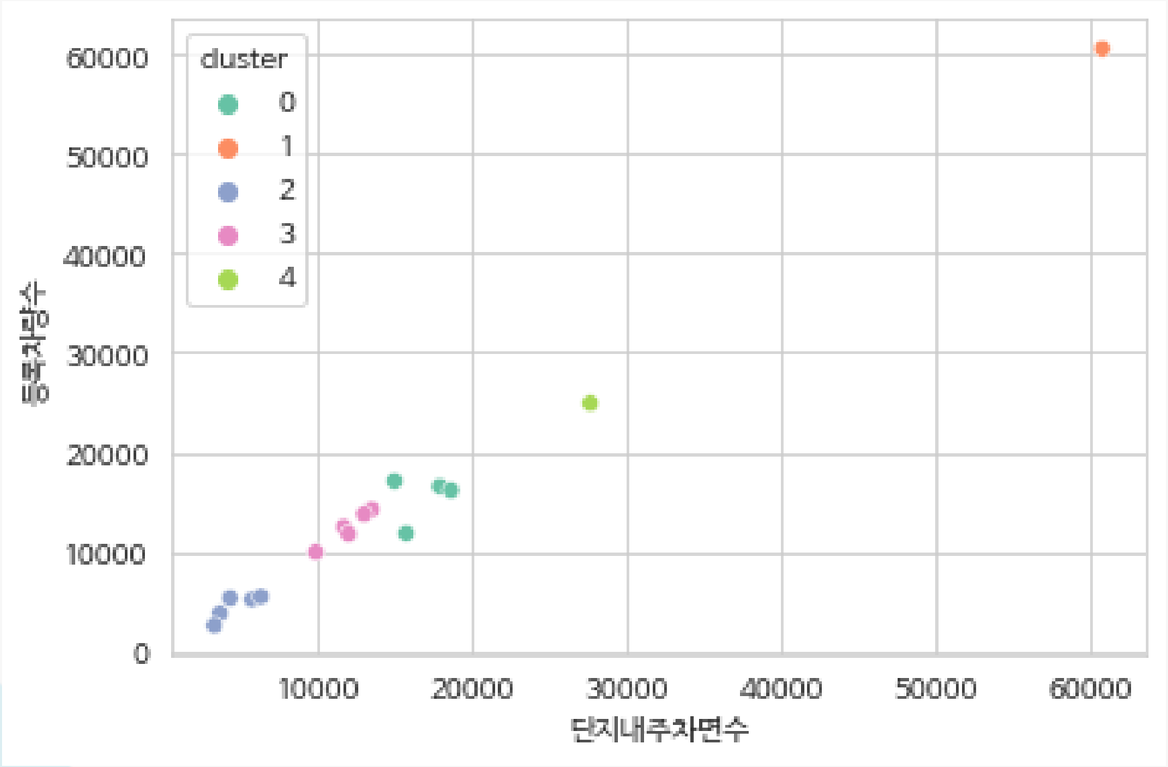


02 데이터 전처리

파생변수 생성

① 지역변수 - Clustering

- K-means Clustering : 각 클러스터와 거리 차이의 **분산을 최소화**하여 데이터를 k개로 그룹화 하는 방식



Group 0 강원도, 경상북도, 대구광역시,전라남도,전라북도

Group 1 경기도

Group 2 광주광역시, 대전광역시, 부산광역시,충청북도

Group 3 서울특별시,세종특별자치시,울산광역시,
제주특별자치도,충청남도

Group 4 경상남도

02 데이터 전처리

파생변수 생성

① 지역변수 - 수도권/비수도권 및 광역시/비광역시 변수 생성

- 수도권 및 광역시 지역 특성상 인구밀집도가 높아 자료의 **편향**이 있음

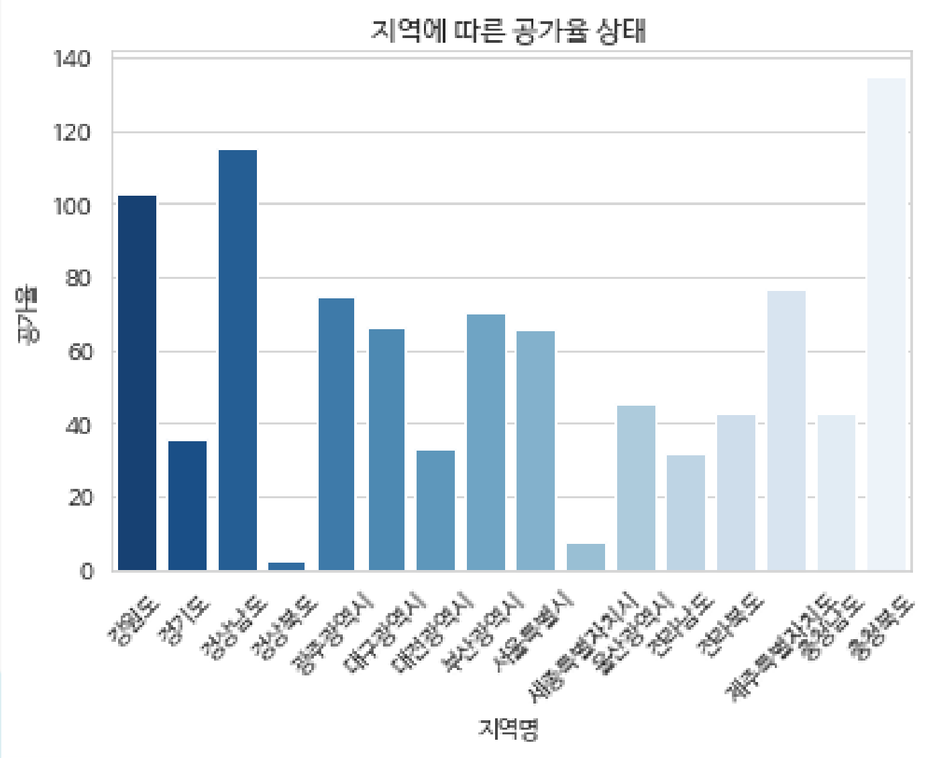
수도권	비수도권	광역시	비광역시
서울특별시	그 외 지역	부산광역시 대구광역시	그 외 지역
경기도		대전광역시 광주광역시	
		울산광역시 세종특별자치시	
		제주특별자치도	

02 데이터 전처리

파생변수 생성

② 공가수 - '공가율' 변수 생성

- **총세대수 대비 공가수 비율**이 자동차 등록대수 예측에 더 적합하다고 판단



$$\text{공가율} = \frac{\text{공가수}}{\text{총세대수}}$$

02 데이터 전처리

파생변수 생성

③ 공급유형 재범주화

- 자격유형의 경우, 한쪽에 치우쳐져 있어 **유형 특성**에 따라 재범주화 필요

국민임대

30년간 임대기간이며
분양전환 불가능

공공임대

모집 공고시 5,10년 등 기간 결정
임대기간 후 입주자 우선 분양

영구임대

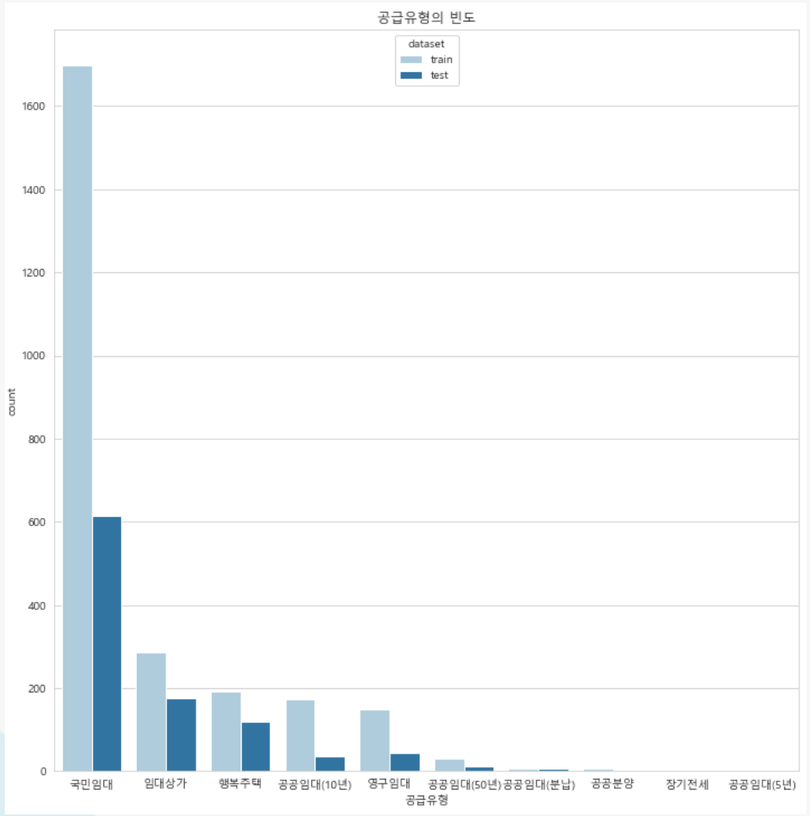
영구적 임대만 가능
해당 지자체에 영구임대가 있는 경우
신규공급 X

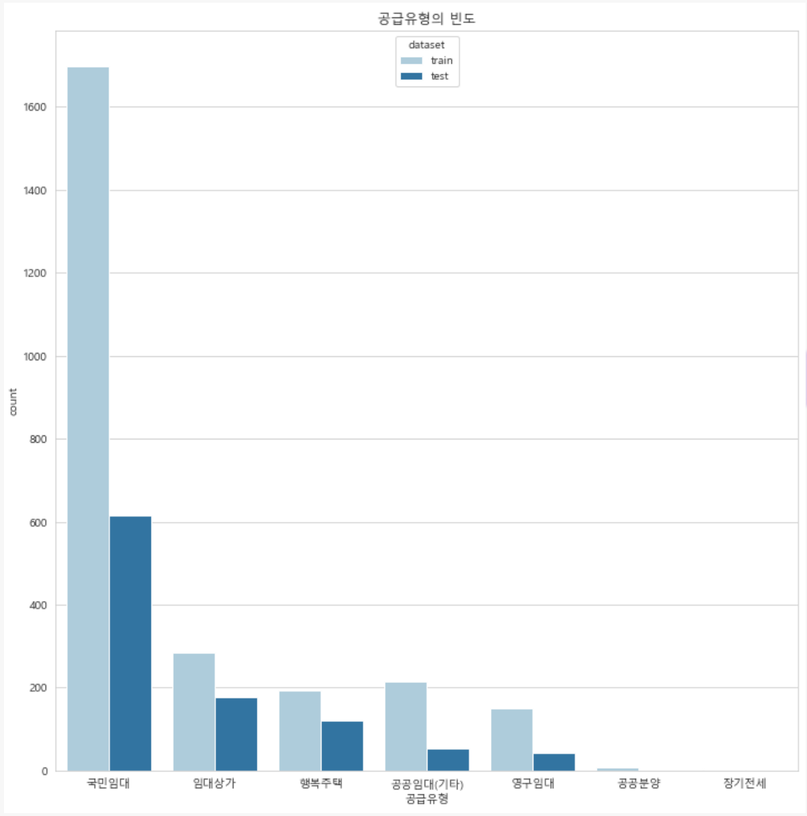
공공임대 기준으로 그룹화

02 데이터 전처리

파생변수 생성

③ 공급유형 재범주화





02 데이터 전처리

파생변수 생성

④ Age 데이터 이용

- 성별에 따른 유의미한 차이가 없으므로 **통합**하여 연령대 변수 생성

	지역	0	10	20	30	40	50	60	70	80	90	2060	3050
0	경상북도	0.063353	0.117706	0.127956	0.103005	0.156273	0.159295	0.146573	0.077889	0.040573	0.007378	0.770990	0.418573
1	경상남도	0.054303	0.108825	0.135538	0.113964	0.144691	0.159442	0.156763	0.081633	0.038655	0.006185	0.792032	0.418098
2	대전광역시	0.057289	0.083282	0.125081	0.135501	0.142592	0.154570	0.158728	0.088153	0.046205	0.008598	0.804626	0.432663
3	경기도	0.077537	0.106536	0.118378	0.141036	0.161562	0.142203	0.132656	0.075148	0.037490	0.007455	0.770983	0.444801
4	전라북도	0.057154	0.119765	0.136585	0.098623	0.143650	0.155790	0.144678	0.086773	0.047690	0.009293	0.766098	0.398062
5	강원도	0.059569	0.103999	0.123531	0.114142	0.144203	0.159443	0.158437	0.080493	0.046542	0.009640	0.780250	0.417789
6	광주광역시	0.066560	0.100105	0.131960	0.131304	0.159212	0.147345	0.135516	0.080306	0.040090	0.007604	0.785641	0.437860
7	충청남도	0.063080	0.121499	0.140597	0.132398	0.152278	0.148467	0.127970	0.070977	0.035413	0.007321	0.772687	0.433143
8	부산광역시	0.044950	0.068193	0.111030	0.095915	0.122722	0.151754	0.194591	0.132131	0.068136	0.010579	0.808142	0.370390
9	제주특별자치도	0.071385	0.123174	0.138801	0.098834	0.153078	0.157333	0.129965	0.080998	0.038167	0.008263	0.759010	0.409245
10	울산광역시	0.049987	0.110842	0.138928	0.105241	0.140767	0.152554	0.165510	0.092326	0.038452	0.005392	0.795327	0.398562
11	충청북도	0.070017	0.121378	0.125760	0.117694	0.154967	0.157023	0.137817	0.069380	0.038974	0.006989	0.762642	0.429684
12	전라남도	0.066577	0.122533	0.135855	0.111448	0.154082	0.157760	0.133817	0.073238	0.036652	0.008036	0.766201	0.423290
13	대구광역시	0.049390	0.071995	0.130810	0.133032	0.138067	0.165291	0.167645	0.092486	0.044442	0.006842	0.827331	0.436390
14	서울특별시	0.030950	0.051330	0.109495	0.099532	0.122889	0.141963	0.198579	0.158593	0.071316	0.015354	0.831050	0.364385
15	세종특별자치시	0.073760	0.103219	0.144116	0.210567	0.150086	0.123696	0.116443	0.052279	0.020979	0.004854	0.797188	0.484350

02 데이터 전처리

파생변수 생성

⑤ 외부변수 - 연령별 운전면허 소지자

- 통계청 자료 이용 및 연령 범주화

	연령대	총계	1종_보통	2종_합	2종_보통	2종_소형
0	16	1807	0	1807	0	0
1	17	4078	0	4078	0	0
2	18	62697	33393	29304	25507	152

재범주화

16세 이상 20세 미만 → 10대

20세 이상 30세 미만 → 20대

30세 이상 40세 미만 → 30대

40세 이상 50세 미만 → 40대

50세 이상 60세 미만 → 50대

60세 이상 70세 미만 → 60대

70세 이상 80세 미만 → 70대

80세 이상 90세 미만→ 80대

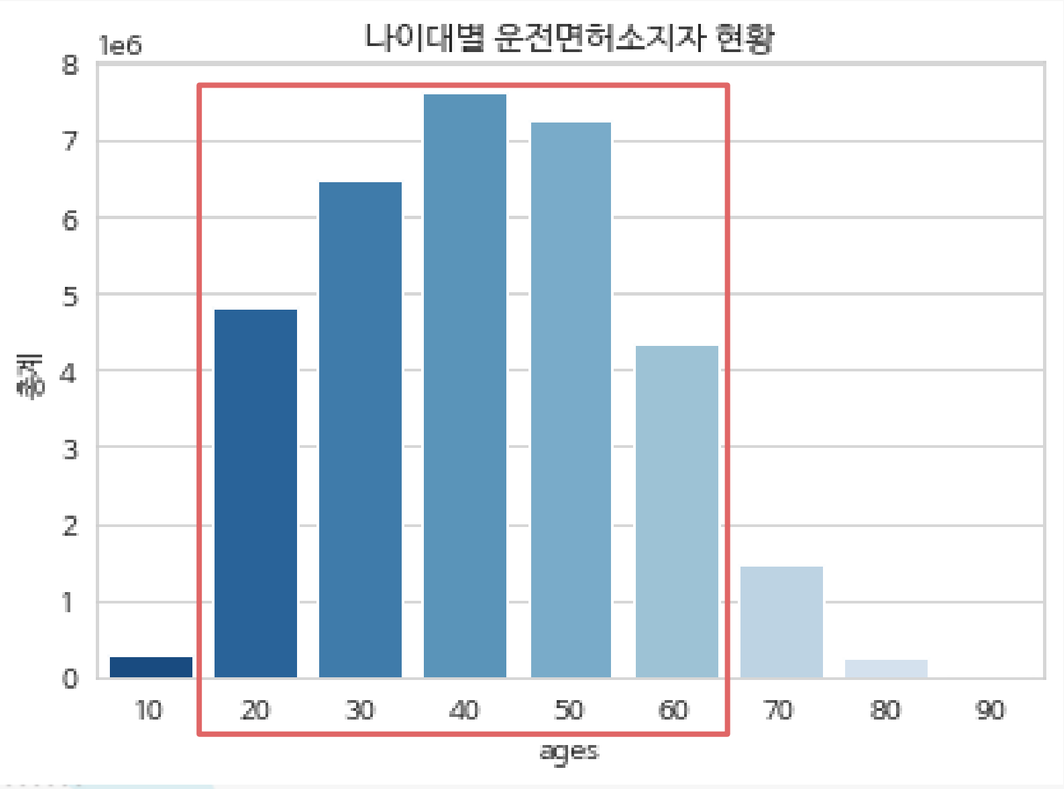
90세 이상 → 90대 이상

02 데이터 전처리

파생변수 생성

⑤ 외부변수 - 연령별 운전면허 소지자

- 통계청 자료 이용 및 연령 범주화



운전면허소지 비율이 높은 연령대를
기준으로 변수 생성

age 2060 : 20~60대

age 3050 : 30~50대

age 2060 , age 3050 각각 평균을 기준

높은지역 **1**

낮은지역 **0**

02 데이터 전처리

파생변수 생성

⑤ 외부변수 - 지역별 1인당 자동차등록대수 및 주민등록인구

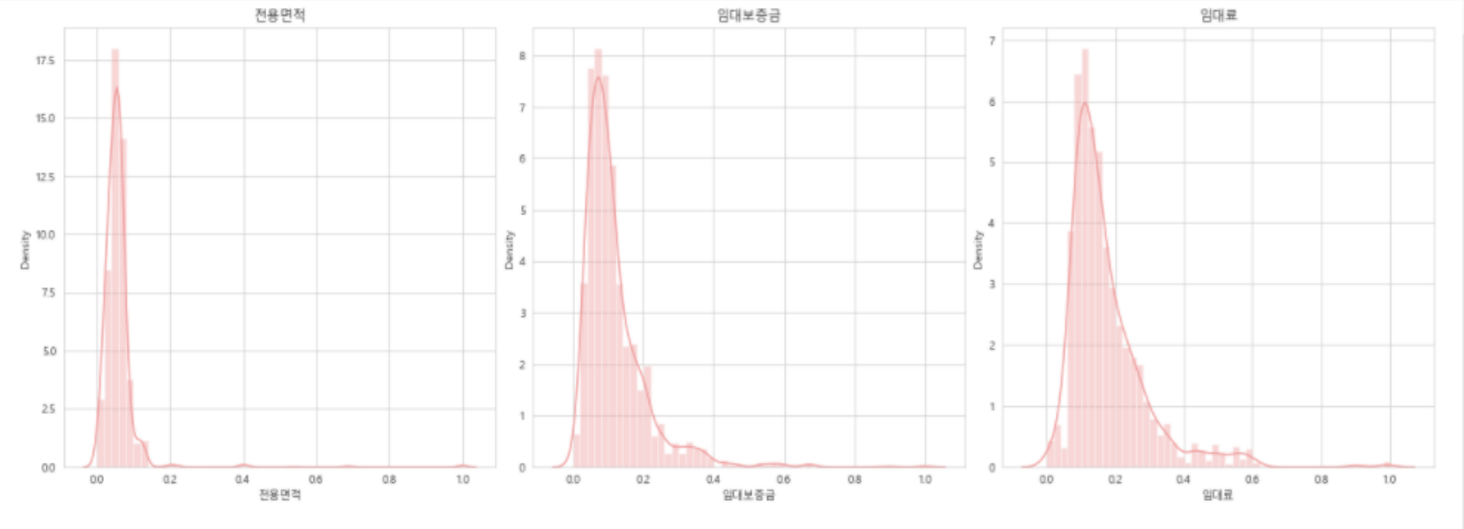
- 통계청 자료 이용
- 지역별 주민등록 인구 및 자동차등록대수가 단지별 등록차량수에 영향을 미칠것이라 판단

	지역명	1인당자동차등록대수	자동차등록대수	주민등록인구
0	서울특별시	0.3	3157361	9668465
1	부산광역시	0.4	1429040	3391946

03 모델링

연속형 및 범주형 변수 처리

연속형 변수



전용면적 , 임대보증금, 임대료 외 다른 연속형 변수 또한 **MIN-MAX Scaling** 진행

Min-Max

최소값을 0 , 최대값을 1로 하여 정규화 하는 방법

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

03 모델링

연속형 및 범주형 변수 처리

범주형 변수

	강원도	경기도	경상남도	경상북도	광주광역시	대구광역시	대전광역시	부산광역시	세종특별자치시	울산광역시	...
0	0	0	0	0	0	0	0	0	0	0	...

지역, 자격유형, 수도권, 광역시 등 범주형 변수 **Dummy 변수화**

Dummy

질적 변수를 0과 1로 변환한 변수

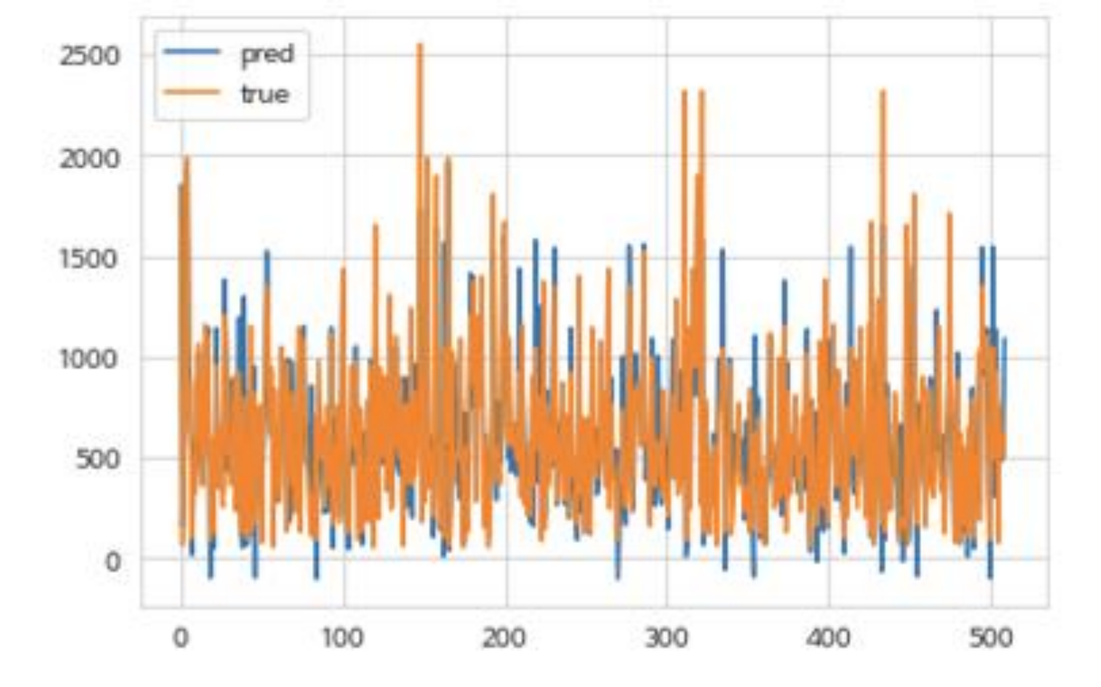
회귀식에서 해당 변수의 효과를 **0** 또는 **상수값**으로 만들어 유의미한 효과 파악 가능

03 모델링 (1)선형 회귀 모델

- Adjusted R^2

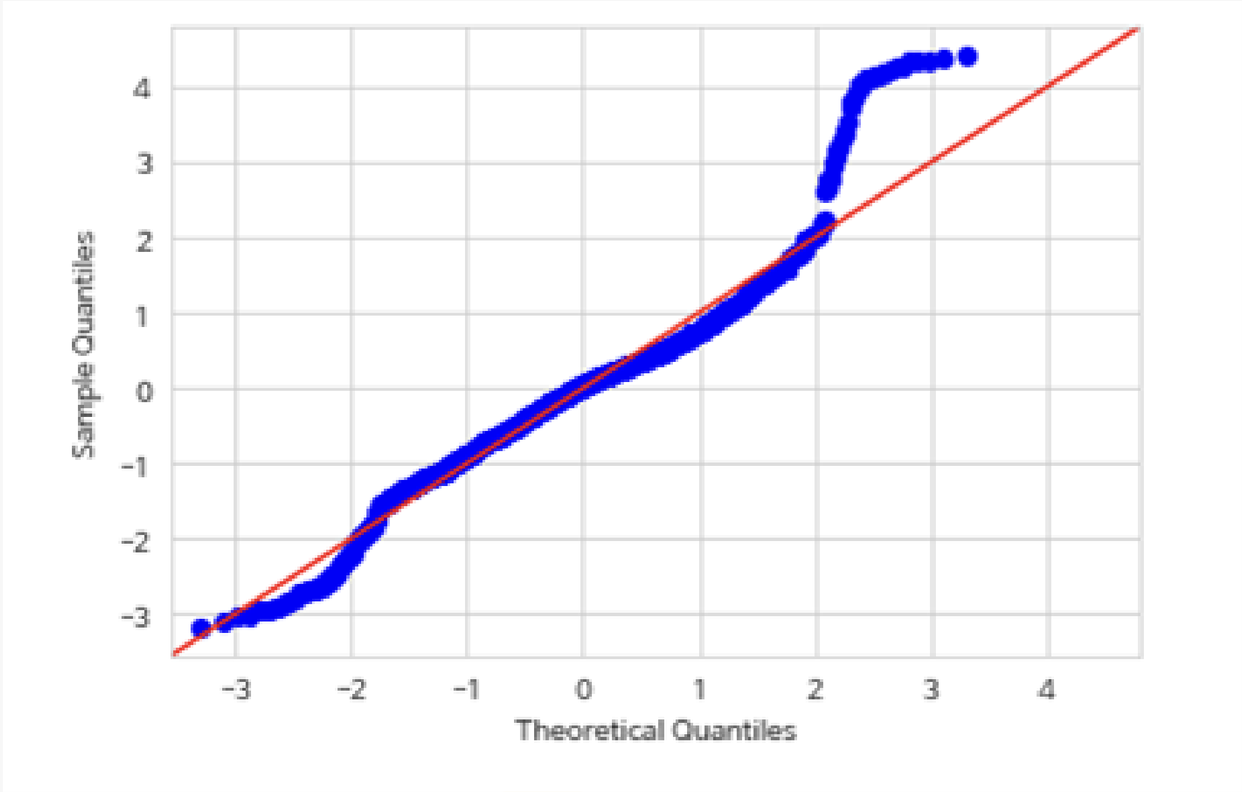
  0.806

- 성능 확인



03 모델링 (1)선형 회귀 모델

잔차 분석



잔차가 정규성을 따름

03 모델링 (1)선형 회귀 모델

Cook's distance

- 이상치 판별

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)\hat{\sigma}^2}$$

i번째 관측값을 제거하고 추정한 회귀 계수와
모든 관측값으로 추정한 값의 차이

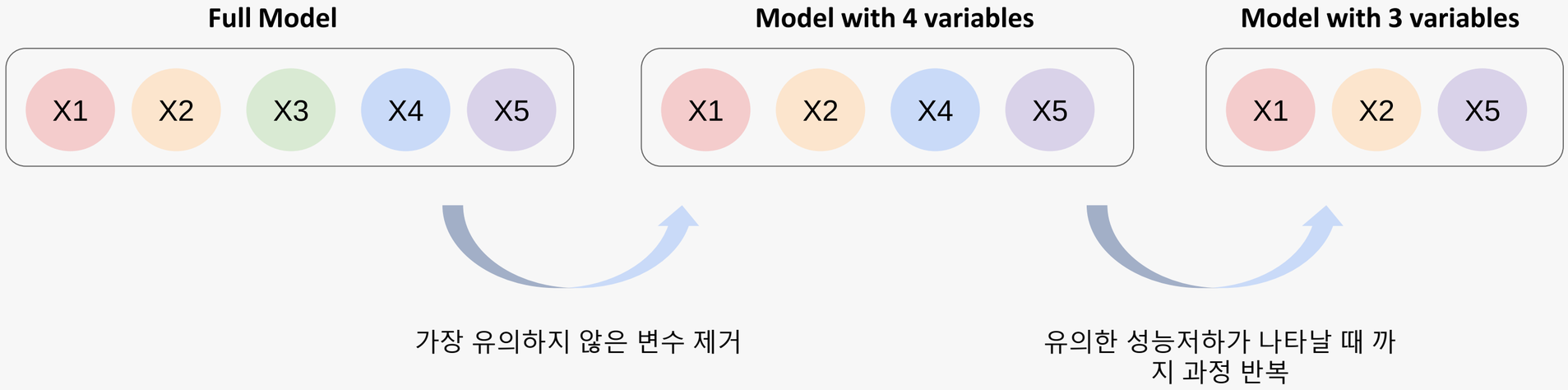
1856	0.015604
1858	0.015003
1857	0.014904
127	0.011532
1695	0.011186

유의한 극단값 없음

03 모델링 (1)선형 회귀 모델

변수선택 - 후진소거법

- 모든 변수가 포함된 모형에서 설명력이 가장 낮은 변수를 제거하며 성능지표를 비교하는 방법



03 모델링 (1)선형 회귀 모델

후진 소거법으로 선택된 변수
(34개 변수 중 20개)

'경상북도', '광주광역시', '대구광역시', '울산광역시', '전라북도', '제주특별자치도', '충청남도', '충청북도', 8, '단지코드', '임대건물구분', '지역', '공가수', '임대보증금', '임대료', '지하철역수', '버스정류장수', '단지내주차면수', '공가율', 'age2_3050'

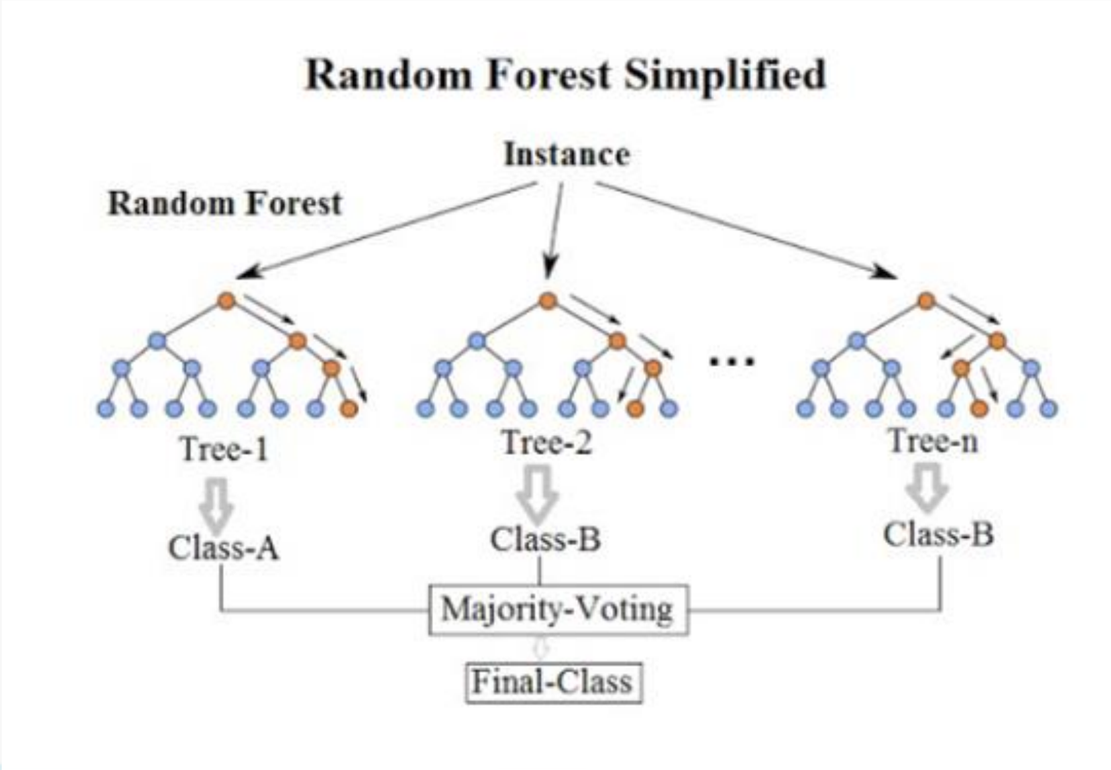
변수 제거 후 Adjusted R^2

0.932

변수 제거 후 성능 개선

03 모델링 (2) 랜덤 포레스트 (Random Forest)

- 앙상블 학습 방법의 일종
- 다수의 결정 트리를 만들고 이들의 결과를 모아 다수결로 결과를 도출



오버피팅(overfitting) 가능성 저하

간편하고 빠르며 정확도가 높음

03 모델링 (2) 랜덤 포레스트 (Random Forest)

모델 성능

Train	Test
MSE : 28603.3562	MSE : 31467.6623
RMSE : 169.13	RMSE : 177.39

Train과 Test의 RMSE 차이가 크지 않음

03 모델링 (2) 랜덤 포레스트 (Random Forest)

하이퍼 파라미터 튜닝

- 모델링을 위해 설정해주는 값

• 최선의 하이퍼 파라미터

max_depth : **16** # 결정 트리의 최대 깊이

n_estimators : **1000** # 결정 트리의 개수

min_sample_leaf : **2**

리프 노드의 최소한의 샘플 데이터 수

min_sample_split : **2**

노드를 분할하기 위한 최소한의 샘플 데이터수

• 최선의 예측 정확도

0.957

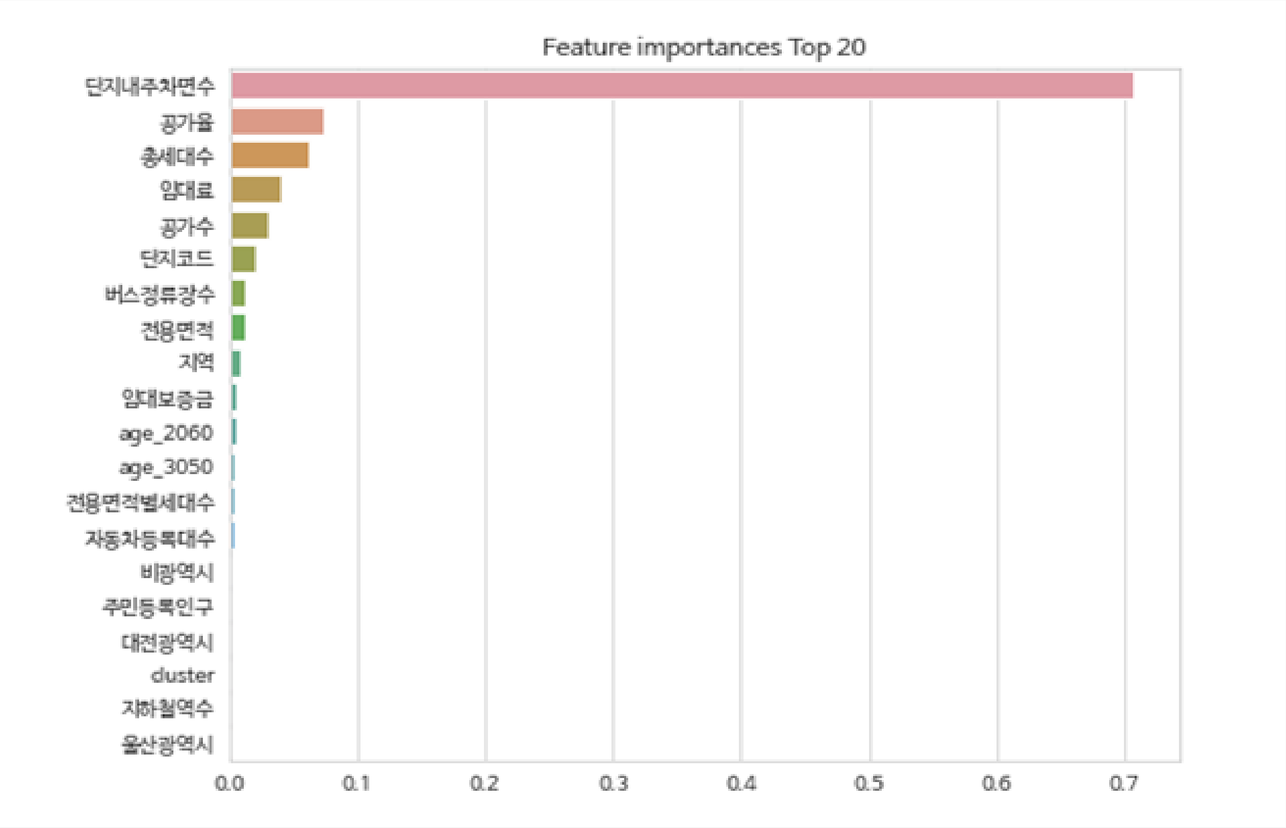
03 모델링 (2) 랜덤 포레스트 (Random Forest)

최선의 하이퍼 파라미터로 학습 및 예측

R^2 : 0.957

MSE : 2951.26

변수 중요도



04 결론

선형 회귀 모델

Accuracy Score

(private score) 412.16

(public score) 314.70

R^2

0.932

랜덤 포레스트

Accuracy Score

(private score) 144.61

(public score) 129.75

R^2

0.957

최종 모델 : **랜덤 포레스트**

04 결론

적용방안 및 한계점

1. 단지내 주차면수, 공가율, 총세대수, 임대료가 등록차량수에 유의한 영향을 끼침
- 공가율 같은 경우는 미리 예측하기 어려우므로 주민등록인구 또는 입주자격에 맞는 인구 분포등을 확인하여 산정한다.
1. 지역을 그룹화 한 것보다 개별 지역이 등록차량수에 더 영향을 끼침
- 각 지역별 특성을 우선적으로 고려할 필요가 있다.
1. 차량 소유 비율이 높은 30 ~ 50대가 유의미한 영향을 미침
- 읍/면/동같은 세부적인 지역별 연령대를 고려하거나 가정을 이루었을 가능성이 있으므로 상권 분석이나 학교 유무와 같은 주변 특징을 파악한다.

04 결론

적용방안 및 한계점

1. 변수 20개 중 단지내 주차면수와 타겟변수간의 상관계수가 0.8 이상인 반면, 다른 변수들은 0.4 이하로 낮았음
2. 각 단지의 준공년도, 위도/경도, 구 등 세부 사항이 없어 외부데이터 이용 및 특성 파악에 어려움이 있었음
3. 모든 상가 데이터의 임대보증금과 임대료가 결측되어 있었음

- LH 외부데이터를 분석해본 결과, 같은 단지의 상가여도 계약 년도나 층수에 따라 임대 보증금과 임대료의 차이가 커 결측치를 처리하는데 어려움이 있었음

ThankYou