아파트 가격 형성 요인에 관한 연구

-가격 급상승 지역을 중심으로-

통계적 데이터 마이닝 Term Project 최종보고

3조 김주형, 배승예, 정승민

2022.05.26

목차

1. 연구배경

2. 연구 설계

3. 데이터

4. 분석 및 결과

5. 결론

무주택자의 아파트 매매 가능성은 점점 희박해지고 있다.

- **아파트 매매 가격은 지속적으로 상승세에 있어 무주택자의 주택 구매 가능성이 더욱 낮아지고 있다.**
  - 아파트 가격을 결정하는 데에 있어 어떠한 요인이 영향을 미치는가를 파악함으로써 원인을 파악할 수 있다.
    정부의 부동산 정책 의사결정에 도움이 될 수 있다.

핵심질문

아파트 가격이 지속적으로 상승하는 원인은 무엇인가?

연구주제

아파트 가격을 결정하는 여러 요인 중 중요한 요소를 파악한다.

* 참고문헌: <"이럴줄 알았으면 5년전 무리해서라도 집 살 걸"...서울 아파트 '전세→매매' 6억 더 있어야>, 매일경제

1. 연구배경

2. 연구 설계

1) 연구대상

2) 독립변수

3. 데이터

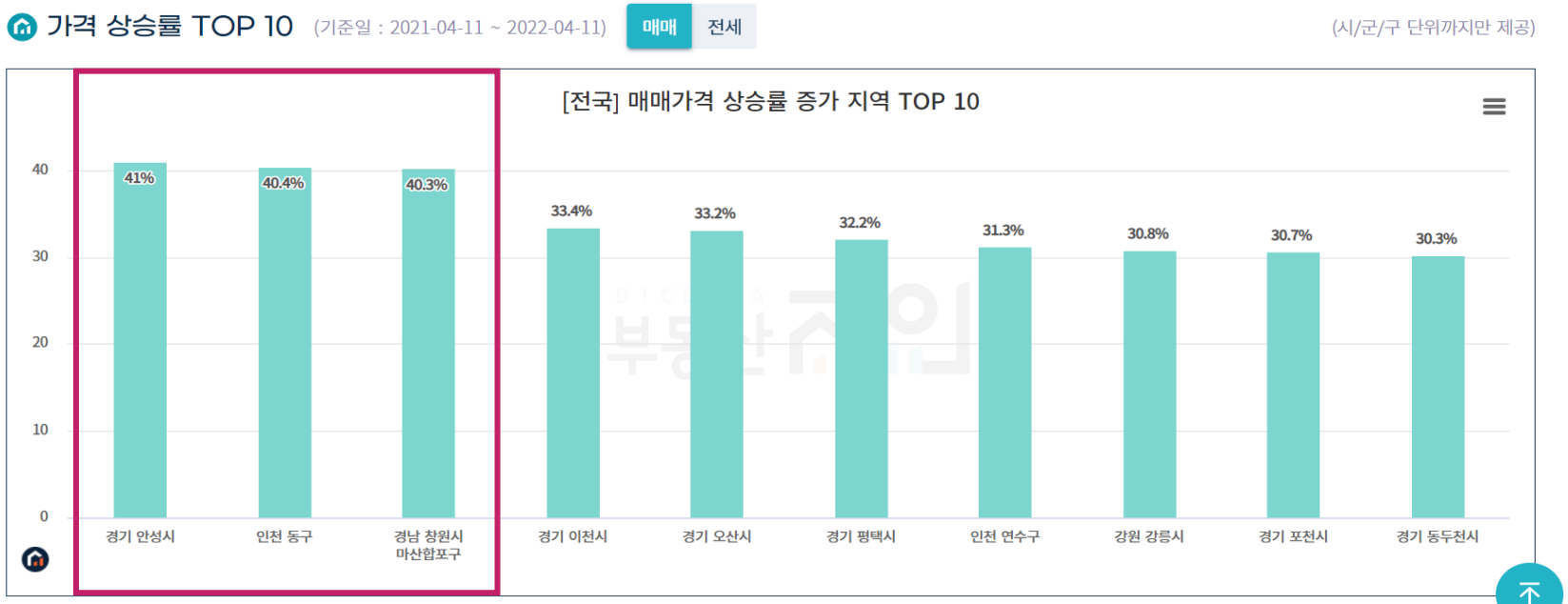
4. 분석 및 결과

5. 결론

최근 1년 간 아파트 가격이 가장 가파르게 증가한 지역은 어디인가?

- **최근 가격이 상승한 지역의 가격 결정요인이 현재의 아파트 결정요인을 반영한다고 가정**
  - 최근 1년(2021/04 ~ 2022/04) 기간동안 매매가격 상승 TOP3 지역을 분석 대상으로 선정
    경기도 안성시 / 인천광역시 동구 / 경상남도 창원시 마산합포구

전국 시/군/구 단위 아파트 매매가격 상승률 증가 TOP10 지역



* 출처: 부동산지인 (https://www.aptgin.com/root_main)

급상승 지역으로 훈련한 모델의 정확도는 어느 정도인가?

- **모델의 검증을 위한 검증 지역 선정**
  - 정량적 검증과 더불어 정성적 평가도 가능하도록 학교 근방 지역인 포항시 남구의 8개 지역 아파트에 모델 적용
    8개 지역: 대도동, 대잠동, 상도동, 송도동, 연일읍, 이동, 지곡동, 효자동

아파트 매매 가격 예측 및 요인 파악 모델링

요인 파악 모델

- **아파트 가격 상승 Top 3 지역**
  - 경기도 안성시
  - 경상남도 창원시 마산합포구
  - 인천광역시 동구
- **요인 파악을 위한 다중회귀분석**
  - OLS Regression
  - Ridge Regression
  - Lasso Regression
  - Elastic net Regression

검증

예측 모델

- **Dataset**
  - Train Data = 아파트 가격 상승 Top3 지역
  - Test Data = **포항시 남구 8개 지역**
- **Linear Models**
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
- **Ensemble Models**
  - RandomForest
  - XGBoost

아파트 가격을 결정하는 요인에는 어떤 것이 있는가?

- **Feature의 유형에 따라 크게 3가지 군집으로 나누고 이에 해당하는 변수를 수집 (총 21개 변수)**
  - 아파트 단지 특성 / 주변시설 (학군 및 편의시설) / 지역 특성 의 세 군집으로 분류

아파트 단지 특성 (9개)

Feature	Explanation
연식	년도 기준 아파트 연식 연식 = 2022 - 준공년도
세대수	아파트 단지 내 총 세대 수
브랜드	아파트 건설사의 21년 시공능력평가 상위 30개 브랜드 해당 여부
건폐율	대지면적에서 건축물이 차지하는 비율 (%)
용적률	대지면적에 대한 건축물의 연면적 비율 (%)
세대당 주차대수	세대 당 가능한 주차 대수
최저/최고층	단지 내 건물의 최저층/최고층
난방방식	중앙/개별 난방

주변시설 특성 (9개)

Feature	Explanation
학군 - 거리	아파트 단지부터 가장 가까운 초/중/고교 까지의 거리 (m)
학군 - 고등학교 순위	최단거리 고등학교의 학교 순위로 파악한 학군
보육시설	반경 1Km이내 국공립 어린이집/유치원 갯수
공원거리	최근접 공원까지의 거리
병원	반경 1Km이내 병원 갯수
약국	반경 1Km이내 약국 갯수
마트	반경 1Km이내 마트 갯수
편의점	반경 1Km이내 편의점 갯수
학원	반경 1Km이내 학원 갯수

지역 관련 특성 (3개)

Feature	Explanation
연평균 소득	$\frac{\text{해당 지역의 기업이 지급한 연간 급여액}}{\text{지역 근로자 수}}$
고령인구비율	$\frac{\text{만 60세 이상 인구}}{\text{해당 지역 전체인구}}$
유입인구비율	$\frac{\text{2021년 12월 인구} - \text{2020년 12월 인구}}{\text{2020년 12월 인구}}$

1. 연구배경

2. 연구 설계

3. 데이터

1) 데이터 수집

2) 데이터 전처리

3) EDA

4. 분석 및 결과

5. 결론

아파트 단지 특성 정보 수집

- **네이버 부동산, 호갱노노 사이트 및 2021 시공능력평가 자료를 활용하여 아파트 단지 특성의 데이터 수집**
  - 네이버 부동산, 호갱노노 – 건설사 정보 제외 아파트 특성 변수 크롤링 및 수집
  - 국토교통부 <21년도 건설업체 시공능력평가 공시> 결과 종합 상위 30개 건설사

아파트 특성 정보

- 네이버 부동산, 호갱노노 사이트를 크롤링하여 분석에 필요한 일부 독립변수와 종속변수 수집
- 독립변수
  - 연식, 세대수, 건설사, 건폐율, 용적률, 세대 당 주차 대 수, 최저층, 최고층, 난방 방식
- 종속변수
  - 최근 실거래 기준 1개월 평균 매매가, 공급면적 (단위면적($1m^2$) 당 가격으로 계산하여 정리)
  - **아파트 단지 당 세대가 가장 많은 평형**을 기준으로 단지의 전체 특성을 반영하고자 함

건설사 브랜드 정보

- 건설사 브랜드 측정을 위해 시공능력평가 해당 여부 수집
- 국토교통부 조사 <21년도 건설업체 시공능력평가 공시> 결과 종합 상위 30개 건설사

	업체 상호
1	삼성물산 주식회사
2	현대건설(주)
3	지에스건설(주)

주변 시설 특성 정보 수집

- **독립변수 중 학군과 주변 시설 정보에 관한 데이터 수집**
  - 학군 정보 – 거리 및 학교 순위의 두 가지 지표 활용
  - 아파트 단지 반경 1Km 이내에 위치하는 각종 편의시설의 수

거리로 측정한 학군정보

- 최단거리의 초/중/고교까지 거리정보
- 최단거리 고등학교 명
- 학구도안내서비스(GIS) 사이트 (https://schoolzone.emac.kr/gis/gis.do)

학교 평가로 측정한 학군정보

- '프람피 아카데미' 제공 고등학교 순위 정보
- 순위 산정방법
  - 순위 = 4년제 대학 진학률 + 서울대 합격자 수
  - 4년제 대학 진학률 = (4년제 대학 진학자 + 국외 대학교 진학자) / 전체 진학자
  - 서울대 합격자 수 = 2015~2018년 서울대 합격자 수 총합

주변 시설 정보

- 호갱노노 사이트에서 아파트 주변 편의시설 정보 수집
- 보육시설
  - 반경 1Km 내 국공립 어린이집/유치원 수
- 기타 주요 편의시설
  - 반경 1Km 내 병원/약국/편의점/마트/학원 수
- 최근접 공원까지의 거리

지역 관련 특성 정보 수집

- **각 지역의 연평균 소득, 고령인구 비율, 유입인구 비율 수집**
  - 연평균 소득 : 호갱노노 직장인 연봉 정보
  - 지역별 인구 정보 : 행정안전부 주민등록 인구통계 사이트

지역별 연평균 소득

- 호갱노노 사이트의 “직장인 연봉” 정보 수집
- 각 지역에 속한 기업이 지급한 연간 총 급여액을 해당 지역 근로자의 수로 나눈 값
- 급여 정보는 국민연금과 금융감독원의 정보를 활용하여 추산

지역 노령인구 비율

- 행정안전부 주민등록 인구통계 사이트 이용 (https://jumin.mois.go.kr/ageStatMonth.do)
- 2021년 인구 기준
- 해당 지역의 60세 이상 인구수를 해당 지역 전체 인구수로 나눈 값

지역별 유입인구 비율

- 행정안전부 주민등록 인구통계 사이트 이용 (https://jumin.mois.go.kr/ageStatMonth.do)
- $\frac{2021\text{년 }12\text{월 인구} - 2020\text{년 }12\text{월 인구}}{2020\text{년 }12\text{월 인구}}$
- 유입된 경우 양수, 유출된 경우 음수

Data Preprocessing

- **모델링에 적합한 형식으로 데이터 처리 및 변환 작업 수행**
  - 데이터별 결측치 파악 및 처리 / 문자형 데이터의 숫자 변환 / 범주형 변수에 대한 One-hot-encoding 진행

데이터별 결측치 처리

- **아파트 매매 최근 실거래가 데이터**
  - 총 47개 결측치 발생
  - KB부동산 사이트에서 31개 보완
  - 나머지 16개 row 데이터 삭제
- **건설사 데이터**
  - 총 53개 결측치 발생
  - 상위 30개 건설사에 해당하지 않는 경우로 처리
- **세대 당 주차 대 수 데이터**
  - 총 63개 결측치 발생
  - 세대 당 주차 대 수 평균값으로 대체
- **용적률/건폐율 데이터**
  - 총 59개 결측치 발생
  - 아파트를 지은 시기를 고려하기 위해 연식 데이터의 10년 단위별 평균 공급면적 당 용적률/건폐율 값을 계산 후, 이를 기준으로 아파트 공급면적에 대한 용적률/건폐율 대체

문자형 변수 처리

- **문자형 변수에 대해 숫자 변환**
  - 공급면적/용적률/건폐율에 대한 숫자 추출
  - 준공년도의 연도 추출 및 연식 계산
  - 아파트 총 주차대수에서 세대당 주차 수 추출
  - 주상복합 아파트에 대한 숫자 라벨링

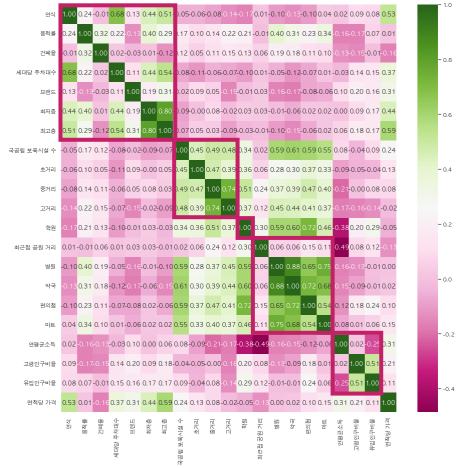
범주형 변수 처리

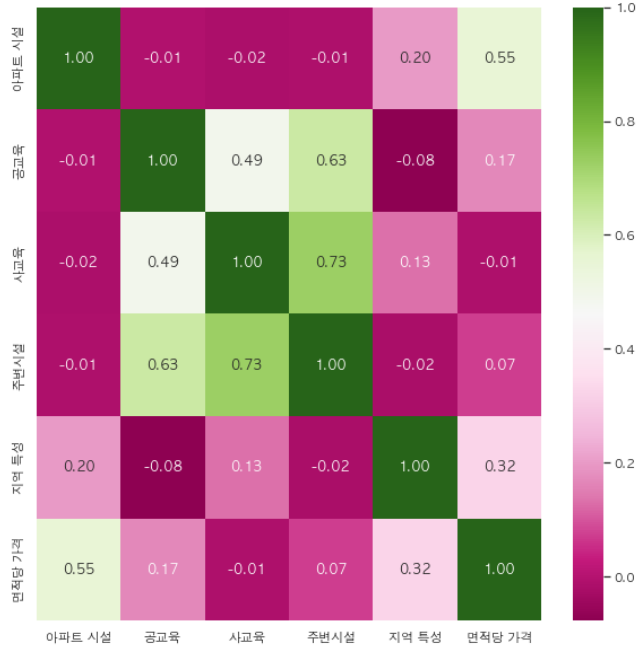
- **범주형 변수에 대해 One-hot-encoding 변환**
  - 난방 종류/아파트 구조 변수는 범주별 위계가 없는 데이터로, One-hot-encoding 방식으로 처리

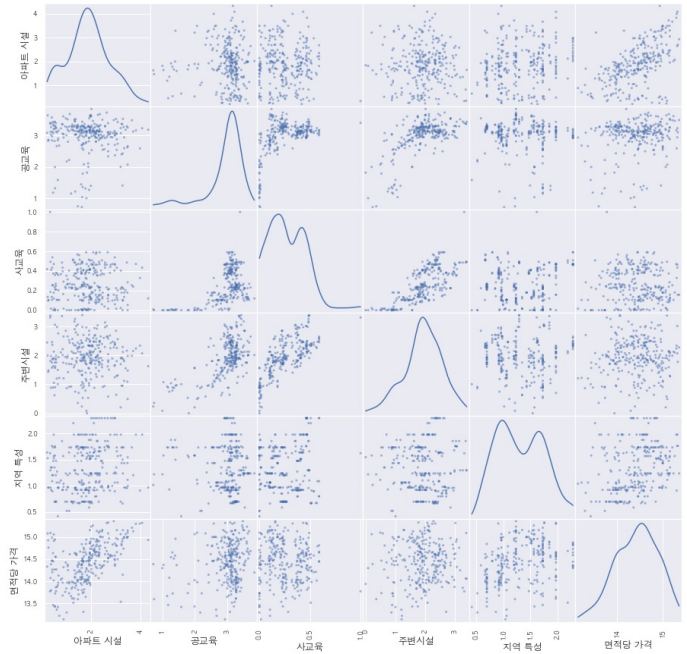
변수를 군집화하여 분석 시 변수 간 상관관계가 더 잘 드러남 (1/2)

- **20개 변수에 대해 아파트 시설 / 공교육 / 사교육 / 주변시설 / 지역 특성 5개로 군집화**
  - 아파트 시설 – 면적당 가격이 0.55로 높은 양의 상관관계를 가지고 있음 확인
  - 공교육 – 주변시설, 사교육 – 주변시설이 0.6 이상의 높은 양의 상관관계를 가지고 있음 확인

군집화한 변수별 상관관계



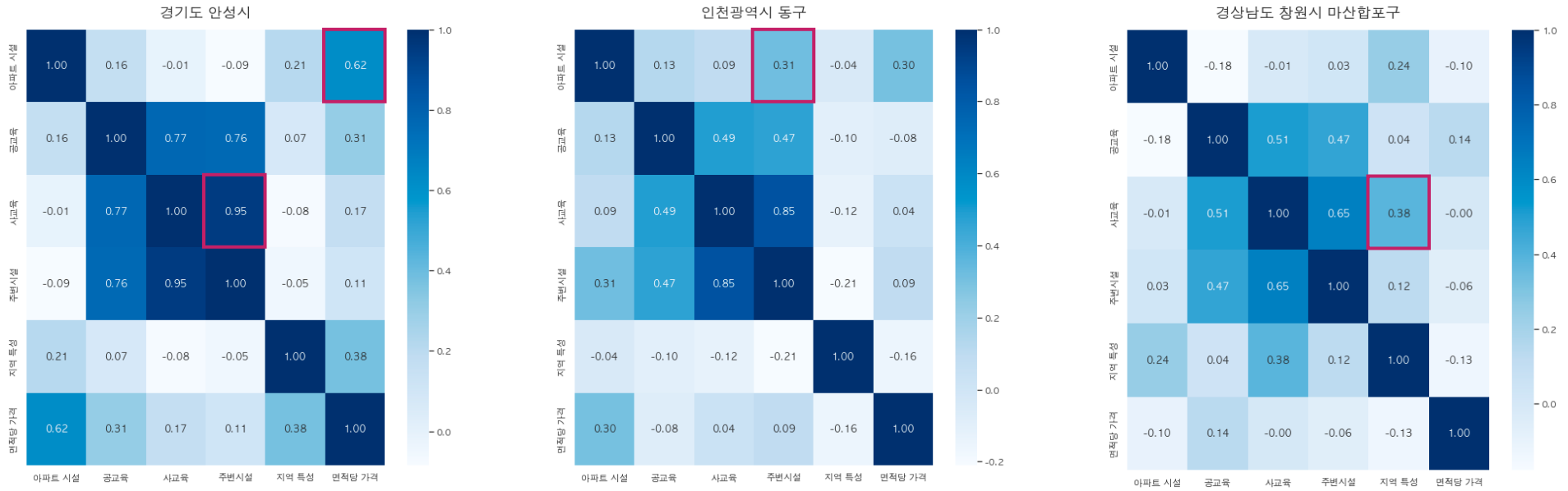




변수를 군집화하여 분석 시 변수 간 상관관계가 더 잘 드러남 (2/2)

- **지역별로 군집화한 변수 간의 상관관계가 다르게 나타남**
  - 경기도 안성시는 아파트 시설 – 면적당 가격이 0.62, 사교육 – 주변시설이 0.95로 높은 양의 상관관계를 가짐
  - 인천광역시 동구는 아파트 시설 – 주변 시설이 0.31로 다른 지역에 비해 높은 양의 상관관계를 가짐
  - 경상남도 창원시 마산합포구는 사교육 – 지역 특성이 0.38로 다른 지역에 비해 높은 양의 상관관계를 가짐

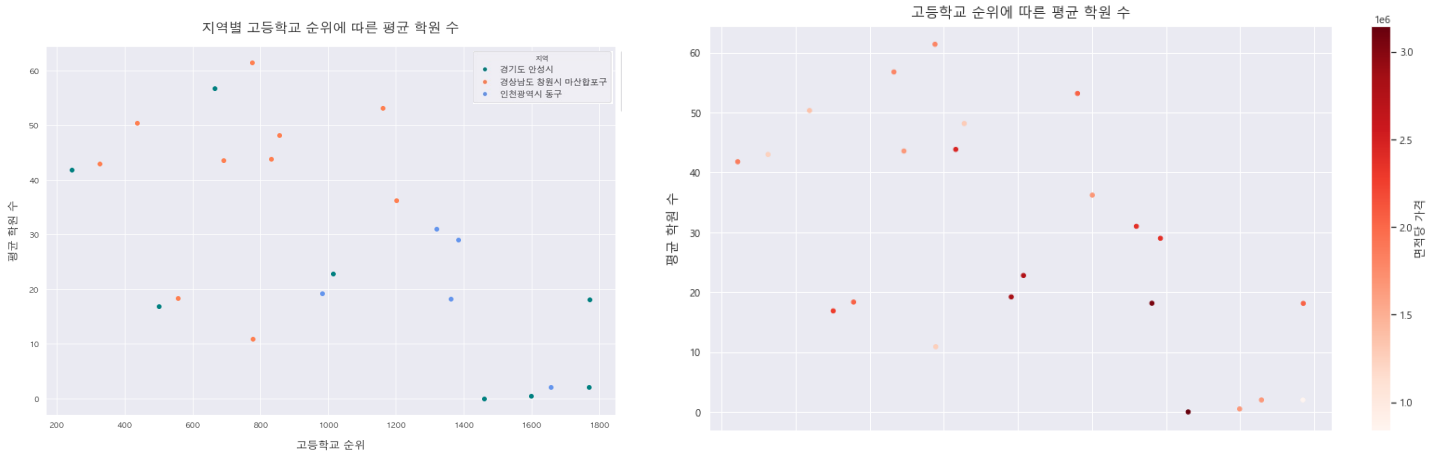
지역별 군집 변수 간 상관관계



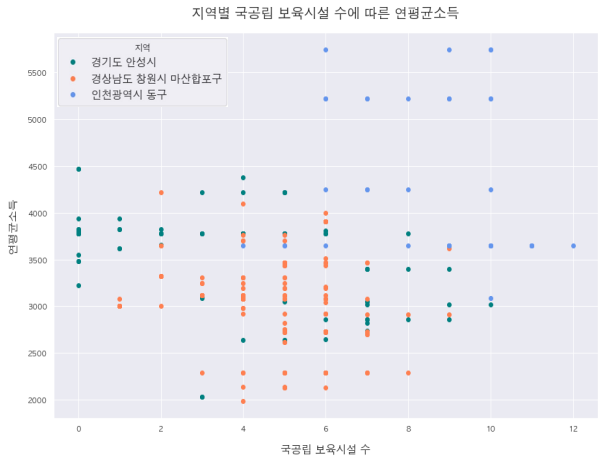
공교육과 아파트 가격 / 연평균소득 간 상관관계 나타나지 않음

- **지역별 고등학교 순위가 높을수록 아파트 주변 평균 학원 수가 많아지는 경향을 보임**
  - 그러나, 고등학교 순위가 높고 평균 학원 수가 많은 아파트가 꼭 높은 가격대를 형성하는 것은 아닌 것으로 나타남
- **연평균소득 별로 국공립 보육시설에 대한 선호도가 나뉘지 않는 것으로 확인됨**

공교육 학군과 학원 수의 관계



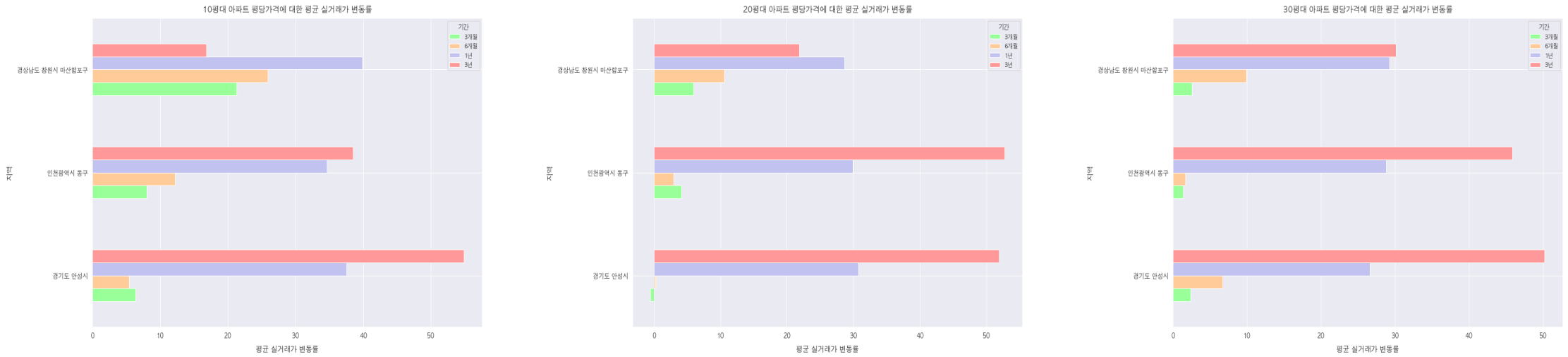
국공립 보육시설 수와 연평균소득의 관계



아파트 실거래가 변동률을 통해 지역별 차이 파악 가능

- **아파트 실거래가 변동률 데이터를 통해 가격 형성에 영향을 미치는 새로운 변수를 찾아보고자 함**
  - 변동률의 경우 결측치가 많아 분석 모델의 독립변수에서는 제외함
- **경기도 안성시 지역은 모든 평대 아파트의 실거래가 평균 변동률이 50%보다 높게 나타남**
  - 경남 창원시 마산합포구 지역은 타 지역에 비해 1년 이하의 기간 변동률이 높으며 특히 10평대 변동률이 높음

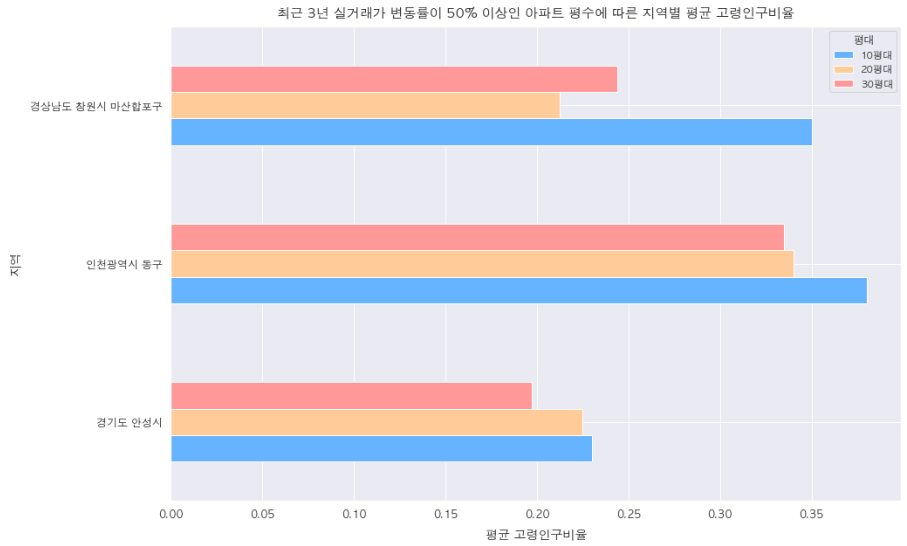
평대별 아파트 실거래가 변동률

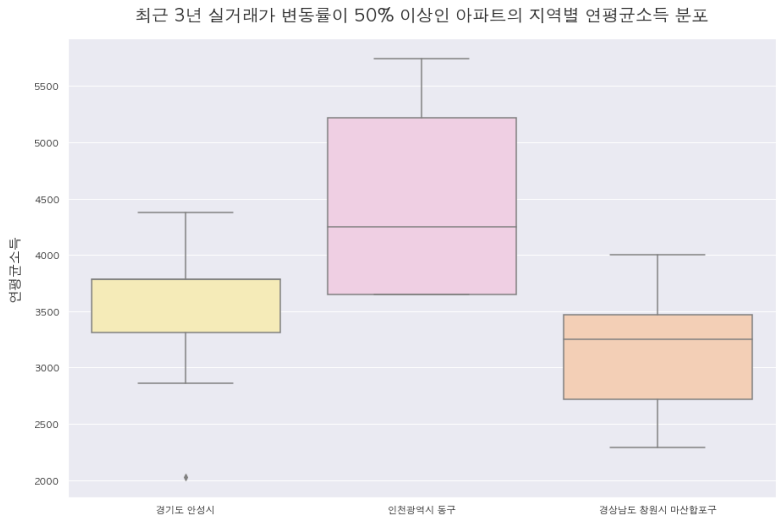


변동률이 높은 아파트는 고령인구비율과 연평균소득의 영향을 받음

- **최근 3년 실거래가 변동률이 50% 이상인 아파트 분석 시 변동률과 지역 특성의 관계가 나타남**
  - 변동률이 제일 높은 경기도 안성시의 평균 고령인구비율이 가장 낮게 나타남
  - 인천광역시 동구가 상대적으로 연평균소득이 높음

실거래가 변동률 50% 이상 아파트 지역별 특성

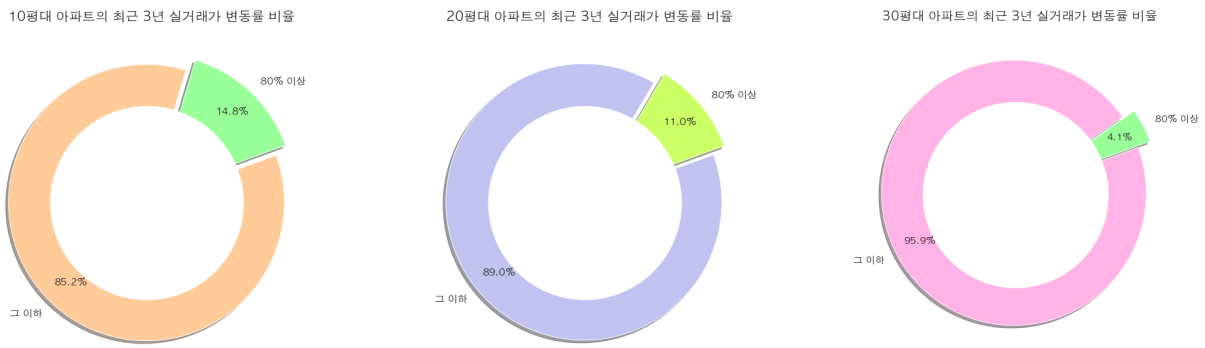




변동률 기준 Outlier의 특성 파악 (1/2)

- **모든 평대에서 80% 이상의 변동률을 가진 아파트가 최대 약 15%까지 나타남**
  - 18개의 아파트가 지역별로 분포되어 있으나 경기도 안성시에서 11개 아파트를 차지하며 가장 많이 나타남
    특별히 가격이 많이 오른 아파트들 사이에 공통된 특징이 있을 것으로 추정

평대별 실거래가 변동률 80% 이상 아파트 비율

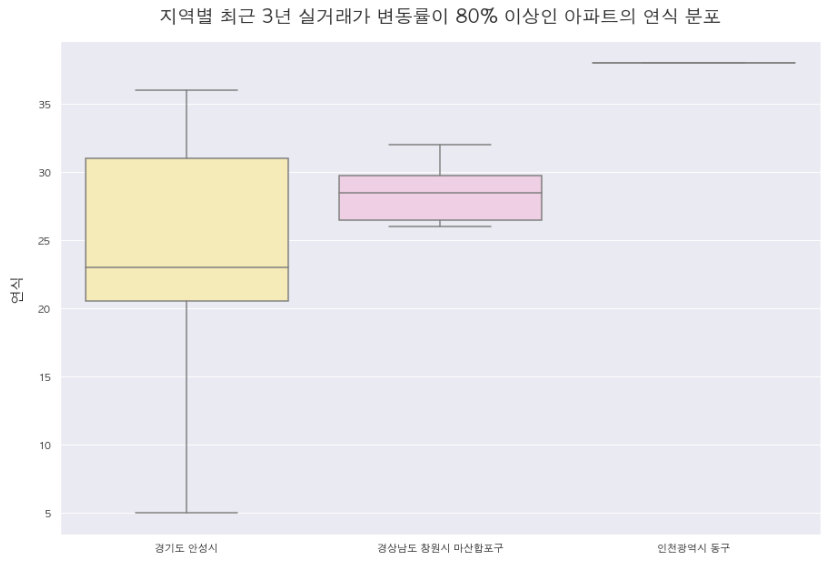


지역별 실거래가 변동률 80% 이상 아파트 분포

변동률 기준 Outlier의 특성 파악 (2/2)

- **80% 이상의 변동률을 가진 아파트의 연식이 20년 이상이거나 7년 이하로 양극화 분포가 확인됨**
  - 조사 결과, 연식이 높아 재개발 예정지가 될 수 있어 아파트 가격이 급상승한 것으로 추정
    특히 아파트 가격이 높지 않고 저층인 아파트에 많이 투자하는 것으로 보임
    연식이 오래될수록 가격이 하락하지만, 재개발 가능성이 존재하는 경우 예외가 발생할 수 있음

지역별 실거래가 변동률 80% 이상 아파트 연식



실거래가 변동률 80% 이상 아파트 정보

아파트 이름	3년 실거래가 변동률	최저/최고층	아파트 연식
안성롯데캐슬	80.92	11/20	7
대우1차	82.34	12/15	30
산수화753	83.18	6/20	20
⋮	⋮	⋮	⋮
주은풍림	107.06	6/20	21
⋮	⋮	⋮	⋮
아양주공1차	124.84	6/6	32
옥산주공	129.03	5/5	35
주은청설	131.23	11/20	23
삼부	135.85	5/5	38
다조맨션	141.67	6/10	26
두산1차	171.72	15/15	32
금산주공	213.93	5/5	36

1. 연구배경

2. 연구 설계

3. 데이터

4. 분석 및 결과

1) 요인 파악 모델

2) 검증 예측 모델

5. 결론

OLS 분석 결과 6개 변수가 유의한 영향을 미치는 것으로 나타남

- **Ordinary Least Square 모델의 설명력은 약 60%이며, 유의수준 0.05 기준 6개의 변수가 유의함**
  - 모델의 p-value 는 4.41 e-48 로 유의수준 0.05 기준 유의하다고 볼 수 있음
  - R-square = 0.656 / Adjusted R-square = 0.628 로 약 60% 수준으로 독립변수가 종속변수를 설명함
  - 최고층 / 세대수 / 연식 / 용적률 / 연평균소득 / 국공립보육시설 수
  - 유의한 변수만으로 OLS분석을 진행한 경우, Adjusted R-square = 0.629로 모델의 설명력이 상승함

OLS 분석 변수별 Coefficient

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2670	0.065	4.123	0.000	0.139	0.395
연식	-0.3526	0.048	-7.280	0.000	-0.448	-0.257
용적률	-0.2340	0.060	-3.914	0.000	-0.352	-0.116
건폐율	-0.0795	0.103	-0.773	0.440	-0.282	0.123
세대당 주차대수	-0.1072	0.060	-1.780	0.076	-0.226	0.011
국공립 보육시설 수	0.2294	0.065	3.510	0.001	0.101	0.358
세대수	0.3665	0.092	3.968	0.000	0.185	0.548
최저층	0.1189	0.103	1.151	0.251	-0.085	0.322
최고층	0.3882	0.086	4.514	0.000	0.219	0.558
초거리	-0.0317	0.072	-0.441	0.660	-0.173	0.110
중거리	-0.0292	0.089	-0.329	0.743	-0.204	0.145
고거리	-0.0320	0.082	-0.388	0.699	-0.194	0.130
학교순위	-0.0104	0.039	-0.263	0.793	-0.088	0.067
연평균소득	0.2249	0.060	3.735	0.000	0.106	0.343
고령인구비율	-0.0410	0.049	-0.834	0.405	-0.138	0.056
유입인구비율	0.0136	0.039	0.350	0.727	-0.063	0.090
최근접 공원 거리	0.0315	0.047	0.672	0.502	-0.061	0.124
병원	0.0175	0.126	0.138	0.890	-0.231	0.266
약국	0.0584	0.135	0.433	0.665	-0.207	0.324
편의점	0.0620	0.069	0.900	0.369	-0.074	0.198
마트	0.0636	0.104	0.613	0.541	-0.141	0.268
학원	-0.0571	0.089	-0.641	0.522	-0.233	0.118

유의한 변수의 OLS 분석 결과

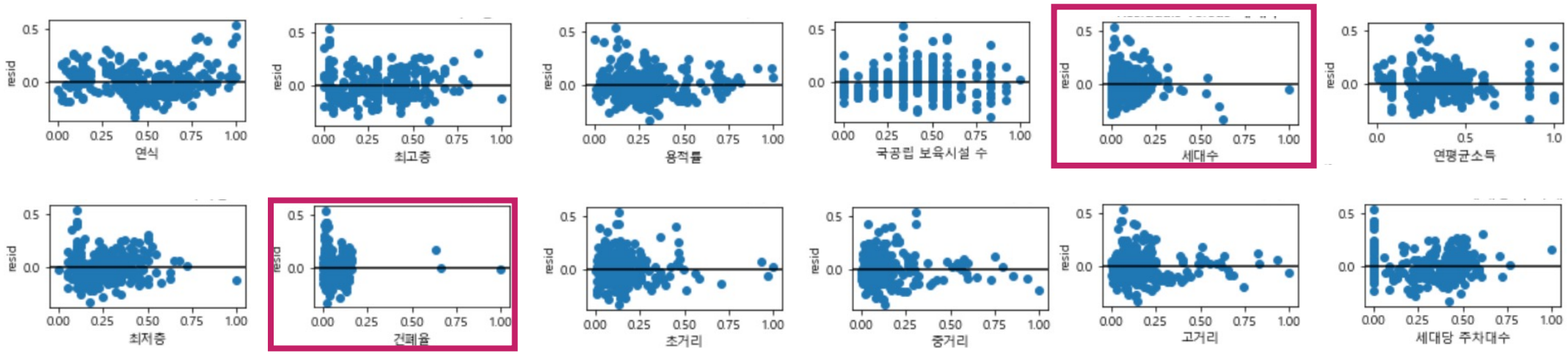
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1989	0.039	5.151	0.000	0.123	0.275
연식	-0.2993	0.038	-7.778	0.000	-0.375	-0.224
최고층	0.4369	0.056	7.805	0.000	0.327	0.547
용적률	-0.2034	0.050	-4.088	0.000	-0.301	-0.105
국공립 보육시설 수	0.3313	0.039	8.402	0.000	0.254	0.409
세대수	0.3696	0.088	4.195	0.000	0.196	0.543
연평균소득	0.2210	0.044	5.038	0.000	0.135	0.307

Residual Plot 확인 결과 세대수, 건폐율 변수의 변환 필요성 파악

- **회귀분석의 Error term에 대한 가정 확인을 위해 residual plot을 확인한 결과, 세대수와 건폐율 변수에서 일정한 패턴이 확인됨**
  - 현재의 linear model이 평당가격과 세대수, 건폐율 변수 간의 deterministic한 부분을 잡아내지 못함

    Variable Transformation의 필요성 파악
  - *회귀분석의 error term 은 {independent and normally distributed, mean=0, same variance}을 만족해야 한다.*
  - *따라서 좋은 residual plot은 0 근처의 밀도가 높고, 0을 기준으로 대칭적이며, 규칙성이 없어야 한다.*

OLS 분석 변수별 Residual Plot



세대수와 건폐율 변수의 로그변환 후 모델의 설명력이 상승

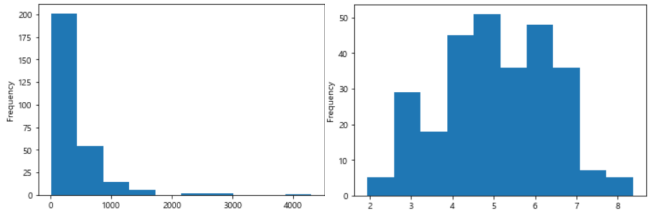
- **R-square : 0.656 → 0.690, Adjusted R-square : 0.629 → 0.665 상승하여 모델의 설명력이 높아짐**
  - 모델의 p-value는 1.08 e-53로 유의수준 0.05 기준 유의하다고 볼 수 있음
  - 세대수와 건폐율 변수의 log 변환한 값의 분포가 보다 symmetric하며 bell 모양을 따름

    회귀분석 모델의 성능을 높일 수 있음을 의미

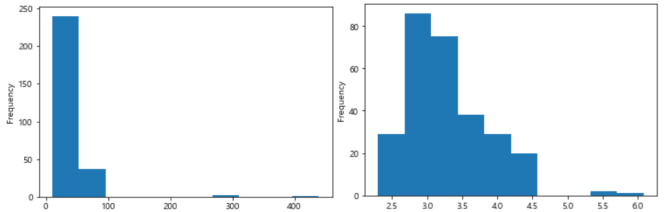
    세대수 → log(세대수)의 coefficient : 0.3665 → 0.3765

    건폐율 → log(건폐율)의 p-value : 0.440 → 0.065 (여전히 유의하지 않음)

세대수 log변환 후 분포 변화



건폐율 log변환 후 분포 변화



변수 변환 OLS 모델 변수별 Coefficient

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2239	0.067	3.354	0.001	0.092	0.355
연식	-0.4104	0.046	-8.843	0.000	-0.502	-0.319
용적률	-0.1451	0.068	-2.130	0.034	-0.279	-0.011
log(건폐율)	-0.1489	0.080	-1.854	0.065	-0.307	0.009
세대당 주차대수	-0.0595	0.058	-1.033	0.303	-0.173	0.054
국공립 보육시설 수	0.2476	0.062	3.998	0.000	0.126	0.370
log(세대수)	0.3765	0.063	6.007	0.000	0.253	0.500
최저층	0.1532	0.098	1.561	0.120	-0.040	0.347
최고층	0.1064	0.097	1.099	0.273	-0.084	0.297
초거리	-0.0731	0.068	-1.069	0.286	-0.208	0.062
중거리	-0.0566	0.084	-0.671	0.503	-0.223	0.110
고거리	-0.0341	0.078	-0.435	0.664	-0.188	0.120
학교순위	-0.0047	0.037	-0.125	0.901	-0.078	0.069
연평균소득	0.1664	0.058	2.875	0.004	0.052	0.280
고령인구비율	-0.0267	0.047	-0.572	0.568	-0.119	0.065
유입인구비율	0.0026	0.037	0.069	0.945	-0.071	0.076
최근접 공원 거리	0.0552	0.044	1.248	0.213	-0.032	0.142
병원	0.0938	0.121	0.776	0.438	-0.144	0.332
약국	0.0146	0.129	0.113	0.910	-0.239	0.268
편의점	0.0395	0.065	0.604	0.546	-0.089	0.168
마트	0.0482	0.098	0.490	0.624	-0.145	0.242
학원	-0.0462	0.085	-0.546	0.585	-0.213	0.120

Penalty 부과 모델 분석 결과 OLS 모델의 설명력이 가장 높게 나타남

- **세대수, 용적률 변수 변환 전/후 모두 OLS모델의 R-square값이 가장 높게 나타남**
  - 이는 OLS모델에서 Overfitting이 일어났을 수 있다는 점을 시사
  - Penalty가 부과된 모델 (Ridge, Lasso, ElasticNet) 모델에서는 overfitting이 방지되어 설명력이 낮게 측정될 수 있음
  - 어떤 모델이 더 적합한지 평가하기 위해선 Test set을 통한 evaluation이 필요

모델별 R-square 값

Model	R-square	변수 변환 R-square
Ordinary Least Square	**0.656**	**0.690**
Ridge Regression	0.651	0.685
Lasso Regression	0.638	0.679
Elastic Net	0.638	0.678

최고층, 연식이 가장 주요하게 아파트 가격 형성에 영향을 미침

- **변수의 로그 변환 전 모든 모델에서 최고층의 영향이 가장 높게 나타나며, 변환 후 연식의 영향이 높음**
  - 용적률이 낮을수록 최고층의 높이는 낮아지는데 어떤 의미일까?

    주어진 땅에 공급할 수 있는 집이 줄어든다.

    수요공급 법칙에서 공급이 적어 가격이 비싸지는 경우

    땅값(평당 가격)이 비싼 아파트에서 보이는 현상

모델별 주요 변수 Coefficient 값

	Ordinary Least Square		Ridge Regression		LASSO Regression		Elastic Net	
모델별 주요 변수	index	Coefficient	index	Coefficient	index	Coefficient	index	Coefficient
	최고층	**0.3882**	**최고층**	**0.3398**	**최고층**	**0.4117**	**최고층**	**0.4115**
	세대수	0.3665	연식	-0.3144	세대수	0.2996	세대수	0.2994
	연식	-0.3526	세대수	0.2958	연식	-0.2888	연식	-0.2887
	용적률	-0.234	국공립 보육시설 수	0.2122	국공립 보육시설 수	0.2545	국공립 보육시설 수	0.2544
	연평균소득	0.2249	연평균소득	0.2094	연평균소득	0.2152	연평균소득	0.2151
	국공립보육시설수	0.2294	용적률	-0.1932	용적률	-0.1582	용적률	-0.1581
변수 변환 모델별 주요 변수	index	Coefficient	index	Coefficient	index	Coefficient	index	Coefficient
	연식	**-0.4104**	**연식**	**-0.3495**	log(세대수)	0.3748	**연식**	**-0.3443**
	log(세대수)	0.3765	log(세대수)	0.3189	연식	-0.3645	log(세대수)	0.3385
	국공립보육시설수	0.2476	국공립 보육시설 수	0.2191	국공립 보육시설 수	0.2781	국공립 보육시설 수	0.2536
	연평균소득	0.1664	연평균소득	0.1614	연평균소득	0.1557	연평균소득	0.1562
	용적률	-0.1451	최고층	0.1555	log(건폐율)	-0.1287	최고층	0.1453

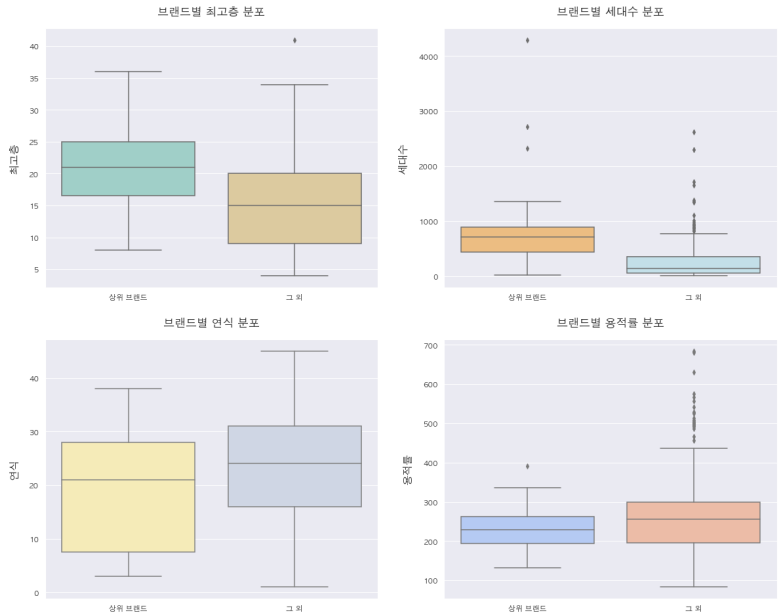
EDA 결과와 다르게 나타난 아파트 브랜드 변수에 대한 추가 분석

- **브랜드 변수에 따른 아파트 가격에 대한 분포가 regression 모델에 적합하지 않은 형태로 나타남**
  - 브랜드와 타 독립변수들 간의 관계에 분포 차이가 분명히 나타남

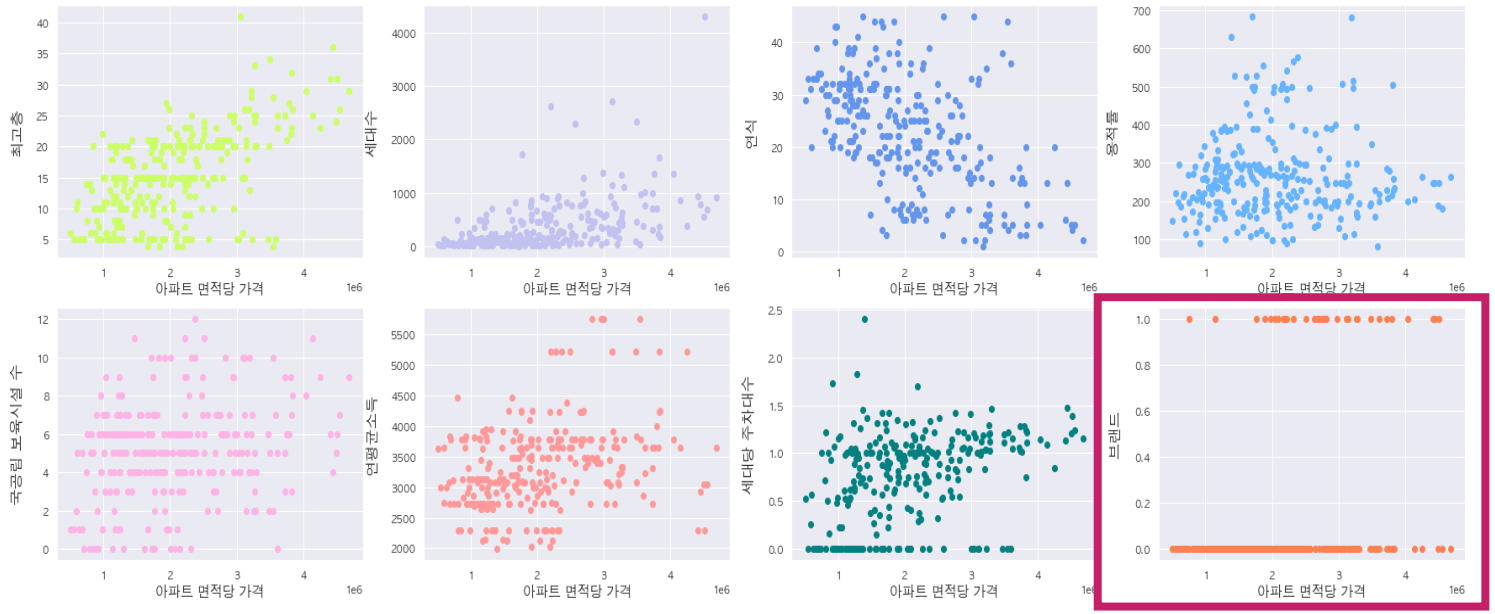
    브랜드 변수를 상위 30개 건설사 해당 여부로 분류하였기 때문으로 추정

    건설사를 상위부터 하위까지 그룹을 5개 나누어 5점부터 1점까지 브랜드를 점수화 했다면 중요도가 높아졌을 것으로 추정

브랜드에 대한 변수 분포



아파트 가격에 대한 각 변수 분포



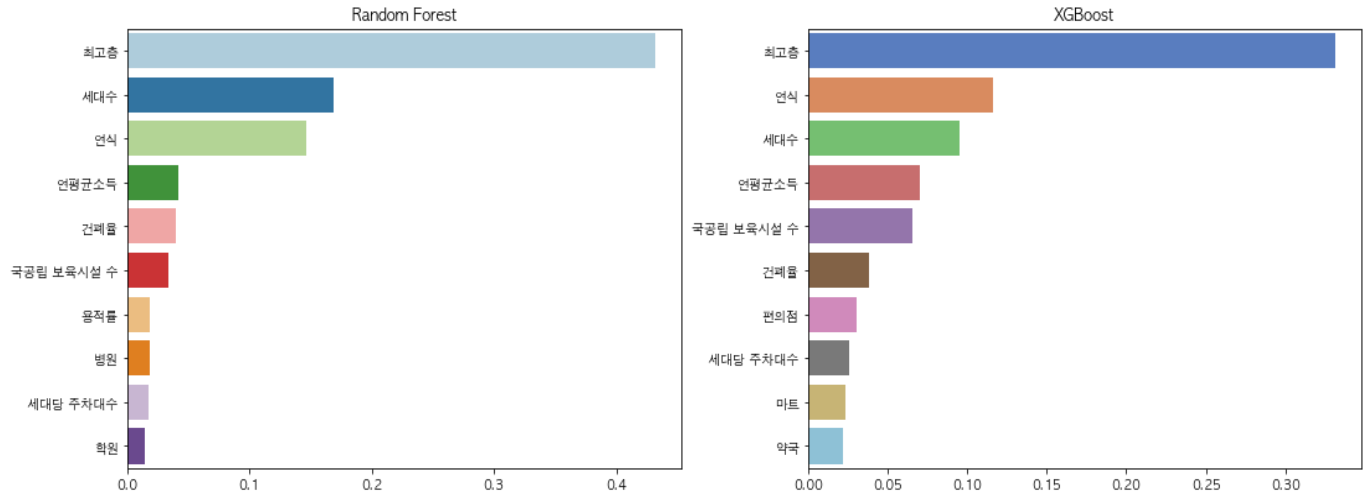
예측 모델 중 Ramdom Forest의 예측 정확도가 가장 높게 나타남

- **Random Forest 모델의 RMSE값이 0.258로 5개 모델 중 가장 낮게 나타남**
  - MAE값 기준 Ridge Regression 모델이 0.219로 가장 좋은 성능을 보임
  - RMSE 기준 RF > Ridge > XGBoost > Linear = Lasso
  - 앙상블 모델에서도 최고층 변수가 가장 많은 영향을 끼침

모델별 예측 성능 평가 결과

Model	MAE	MSE	RMSE
Linear Regression	0.220	0.075	0.275
Ridge Regression	**0.219**	0.069	0.262
LASSO Regression	0.222	0.076	0.275
Elastic Net	0.226	0.069	0.263
Random Forest	0.231	**0.066**	**0.258**
XGBoost	0.246	0.074	0.272

Ensemble Prediction Model 변수별 중요도



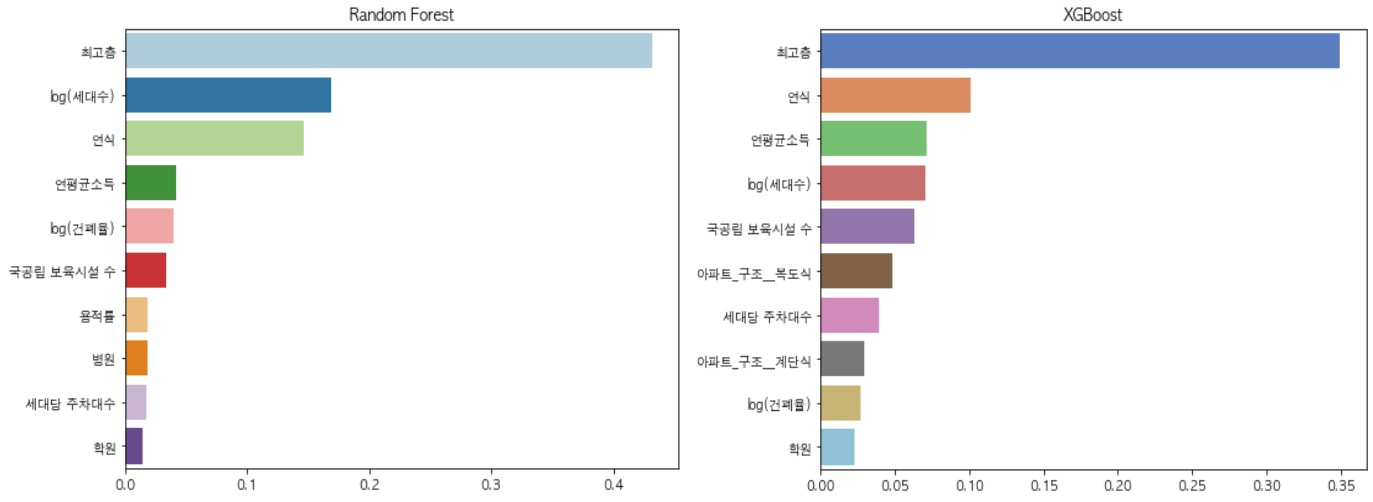
변수 로그변환 반영 시 예측 모델의 성능도 향상됨

- **RMSE값 기준 Elastic Net, MAE값 기준 LASSO 모델의 성능이 가장 좋음**
  - 대체적으로 Linear 방식이 Tree 기반 앙상블 모델에 비해 좋게 나타남
  - 앙상블 모델에서는 로그 변환 시에도 최고층 변수가 가장 중요한 영향을 미침

모델별 변수 변환 후 예측 성능 평가 결과

Model	MAE	MSE	RMSE
Linear Regression	0.174	0.067	0.259
Ridge Regression	0.175	0.063	0.252
LASSO Regression	**0.169**	0.063	0.250
Elastic Net	0.174	**0.062**	**0.249**
Random Forest	0.222	0.066	0.257
XGBoost	0.230	0.069	0.263

Ensemble Prediction Model 변수별 중요도



1. 연구배경

2. 연구 설계

3. 데이터

4. 분석 및 결과

5. 결론

분석 결과 아파트 가격에는 단지 자체 특성이 가장 많은 영향을 미침

- **최고층, 연식, 세대수의 아파트 단지 자체 특성이 많은 영향을 미치는 것으로 나타났다.**
  - 다만, EDA 결과와 달리 브랜드 정보의 영향이 적어 확인 결과 이진 분류로 인해 변수의 영향력이 줄었을 것으로 추정된다.
  - 단지가 크고 신축일수록 아파트 가격은 상승하지만, 연식이 매우 오래되어 재개발 가능성이 있는 경우 역시 높은 가격을 보였다.
- **지역 특성 중 지역별 연평균 소득이 가격 형성에 영향을 미치는 것으로 나타났다.**
  - 인구의 유입 정도나 연령 비율에 관한 사항은 큰 영향을 미치지 않는다.
- **학군 정보에서 국공립 보육시설 수가 주요하게 영향을 미치는 것으로 나타났다.**
  - 초/중/고교 보다는 미취학 아동을 위한 시설에 영향을 받음을 알 수 있다.
  - 사교육인 학원의 영향은 크게 나타나지 않았다.
- **이외 주변 시설의 분포와 편의성이 아파트 가격의 형성에 직접적으로 영향을 미치지는 않는 것으로 파악되었다.**

시사점 및 한계점

- **아파트 단지의 규모와 지역 평균 소득에 아파트 가격이 영향을 받는다는 데에서 정책 결정에 도움을 줄 수 있다.**
  - 평균 소득 기반으로 공공 주택 공급 결정하여 정부 수입 증가 가능
  - 부동산 규제 정책의 결정에 있어 지역 기반 및 단지 규모에 대한 특성 반영 가능
- **다만, 데이터 수집에 있어 절대적인 숫자가 적었다는 한계를 갖는다.**
  - 최근 1년간 집값이 급상승한 3개의 지역을 대상으로 하여 데이터가 280개로 다소 부족한 양이었다.
    이때, 지역을 늘려 데이터 확보 시, 지역별로 달리 나타나는 특성을 반영하기 위한 새로운 변수를 도입해볼 수 있다.
  - 아파트 가격이라는 매우 복합적인 현상을 설명하기에 21개의 변수는 부족하다.
    특성별로 나눈 카테고리 별 변수의 수 균형도 맞지 않아 추가적으로 고려해볼 만한 변수가 많다.
- **아파트 단지 단위가 아닌 매매 실거래 내역을 Sample로 삼아 연구 확장 시 더 나은 정확도를 추출할 수 있을 것으로 예상된다.**

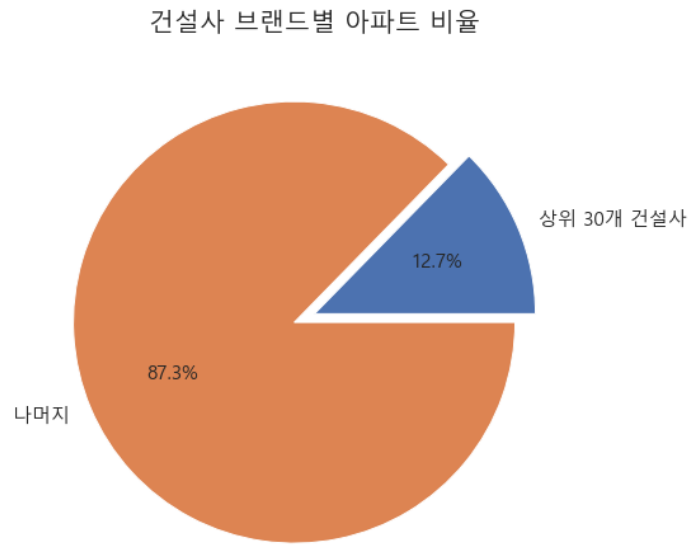
감 사 합 니 다

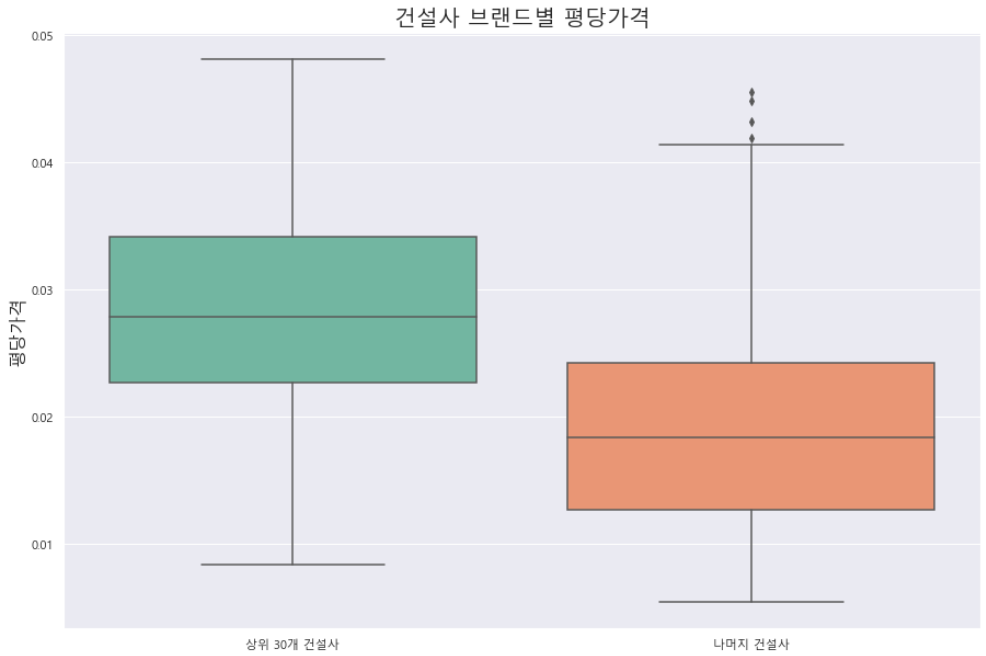
Appendix

- 중간 발표 시 진행한 EDA 사항 첨부

상위 30개 건설사에 해당하는 경우 기타 건설사보다 높은 가격을 보임

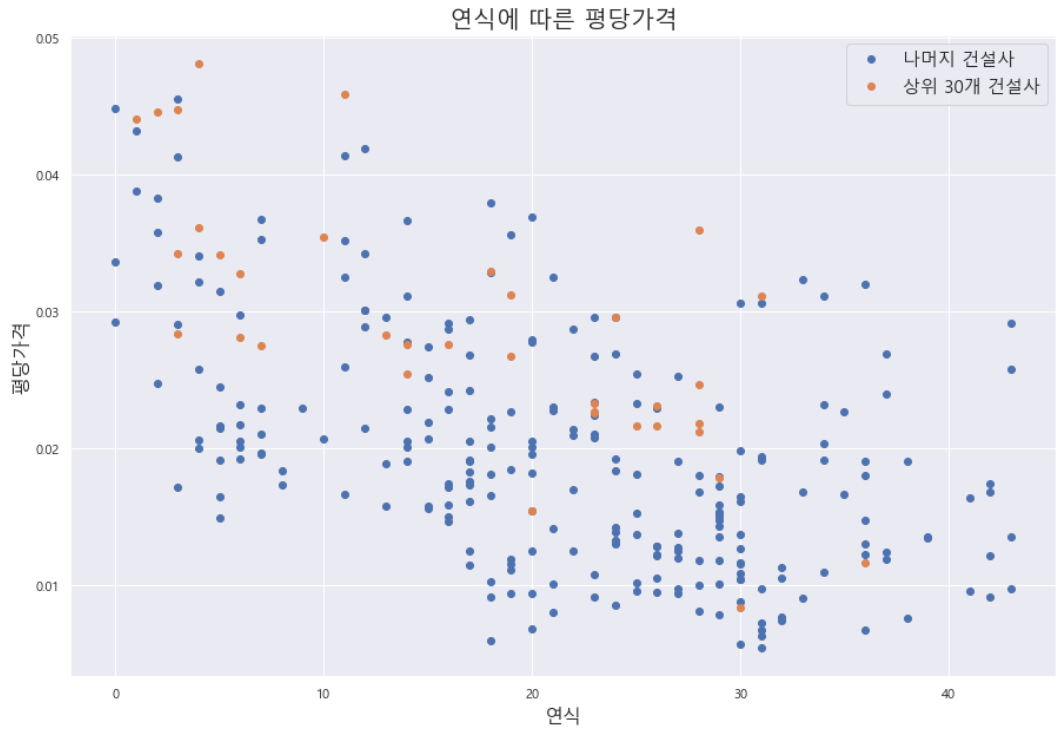
- **건설사 브랜드별 아파트 비율 및 평당가격**
  - 상위 30개 건설사 비율은 12.7%로 작게 차지
  - 상위 30개 건설사 평당가격 분포의 중앙값이 나머지 건설사의 상위 25% 평당가격보다 높음 확인





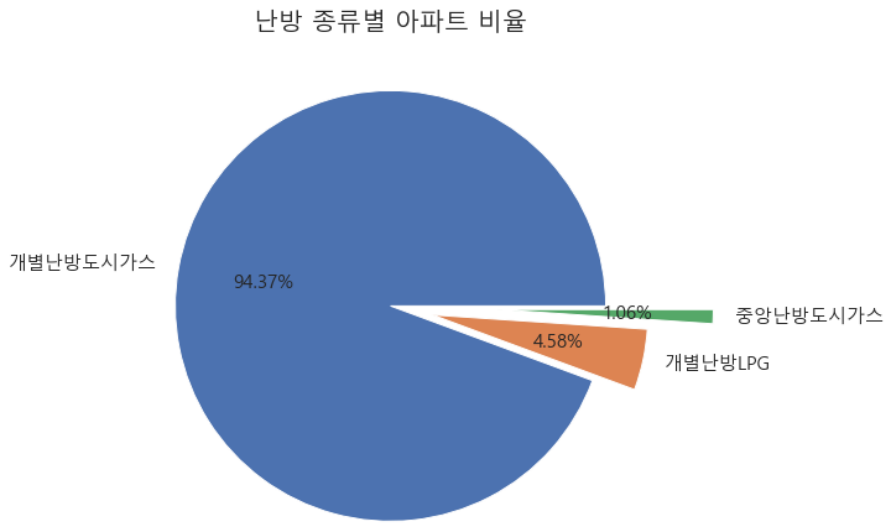
아파트 연식이 오래될 수록 가격이 낮게 나타남

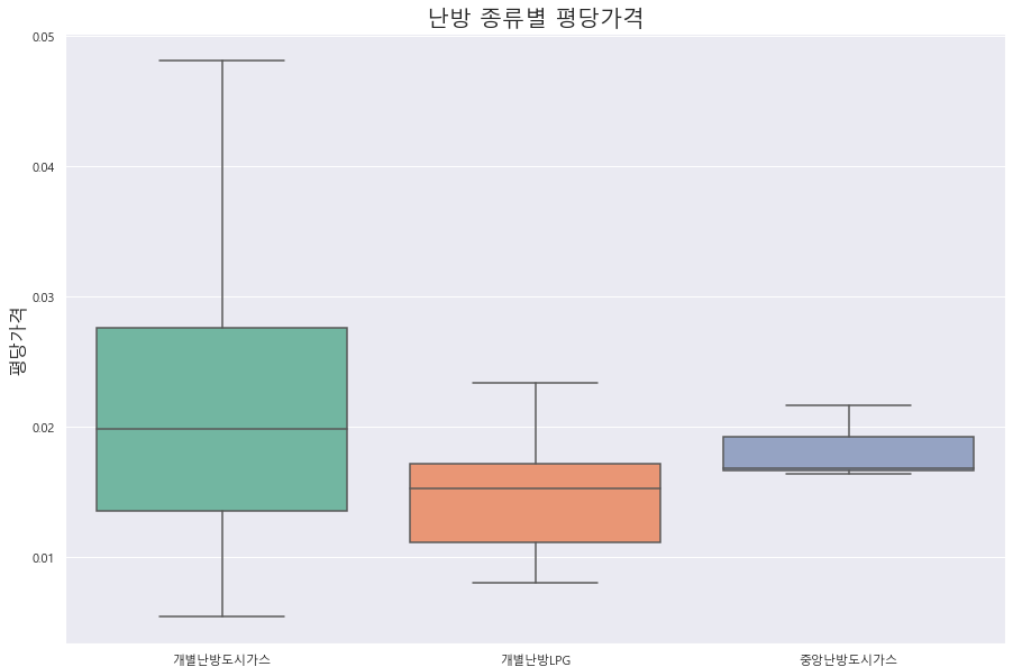
- **아파트 연식에 따른 평당가격 분포**
  - 연식이 높을수록 평당가격이 낮아지는 경향성 확인



LPG가스보다 도시가스 이용 난방방식의 아파트 가격이 더 높은 경향

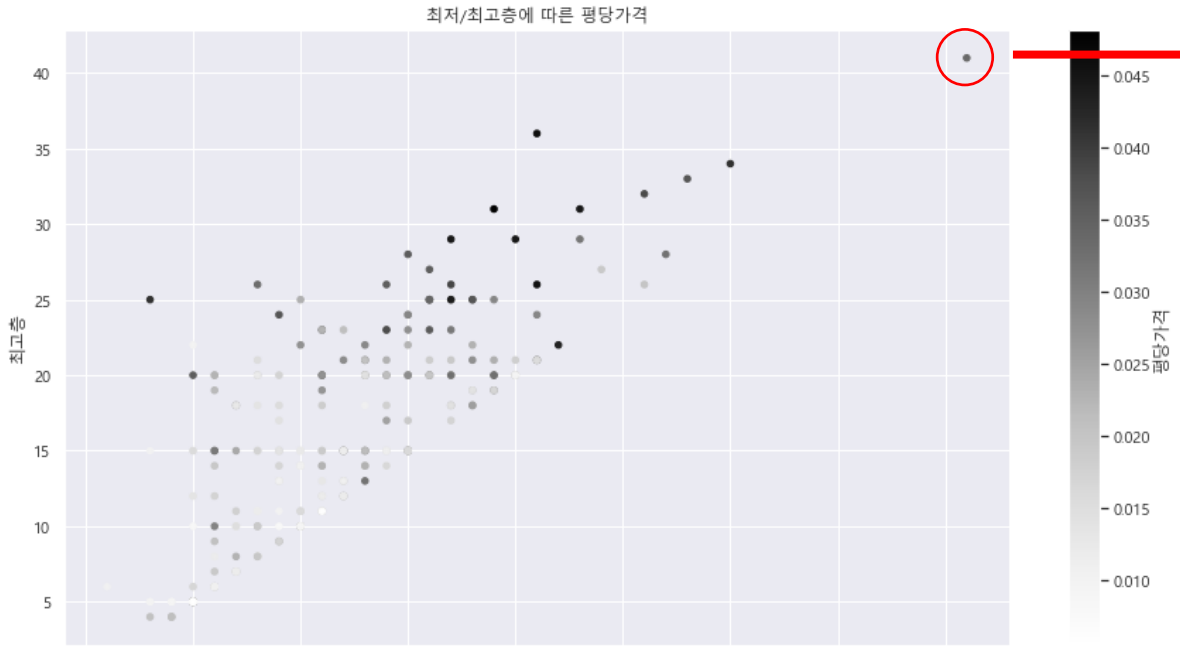
- **난방 종류별 아파트 비율 및 평당가격**
  - 개별난방도시가스 비율이 94.37%로 대부분 차지
  - 중앙난방도시가스 평당가격이 개별난방LPG 평당가격보다 높게 분포하고 있음 확인

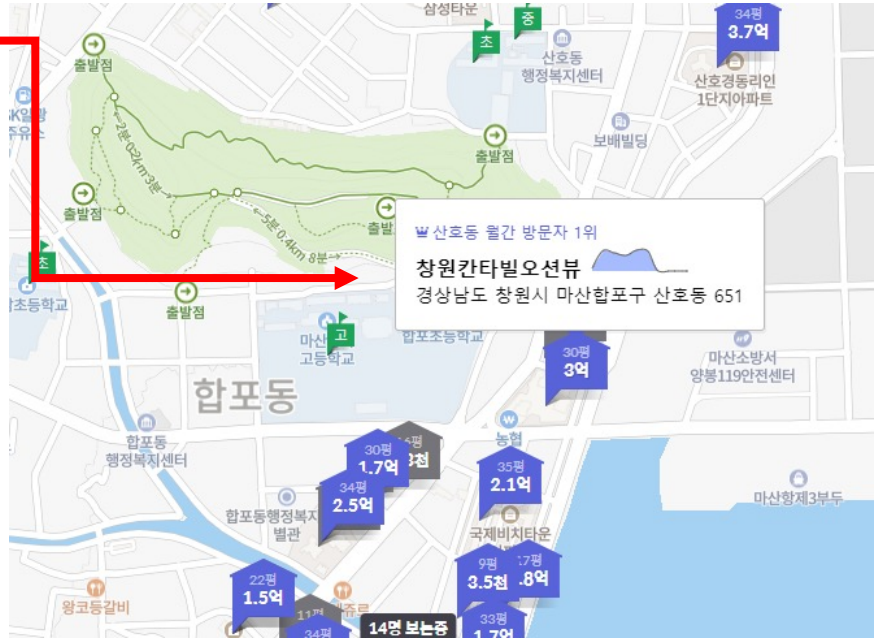




최저/최고층이 높은 아파트 단지일수록 가격이 높게 나타남

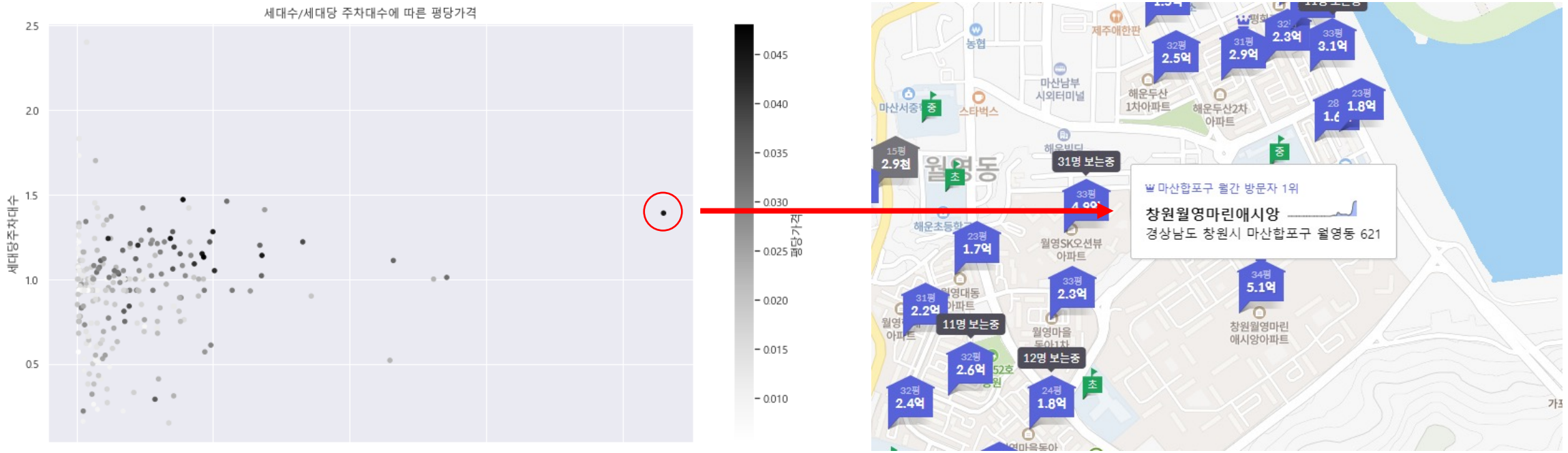
- **최저/최고층에 따른 평당가격**
  - 최저/최고층이 높을수록 평당가격이 높은 경향성 확인
  - 최저/최고층 이상치 조사 결과 지역에서 가장 인기있는 아파트임을 확인하여 제거하지 않기로 결정





세대수가 많을수록, 주차 가능 수가 많을수록 높은 가격이 형성됨

- **세대수/세대당 주차대수에 따른 평당가격**
  - 세대수/세대당 주차대수가 높을수록 평당가격이 높은 경향성 확인
  - 세대수/세대당 주차대수 이상치 조사 결과 지역에서 가장 인기있는 아파트임을 확인하여 제거하지 않기로 결정



변수 간 상관관계 분석

- **평당가격과의 관계분석에서 세대수와 최고층은 양(+)의 상관관계를, 연식은 음(-)의 상관관계를 보임**
  - 세대수 – 평당가격, 최고층 – 평당가격이 0.5 이상의 높은 양의 상관관계를 가지고 있음 확인
  - 연식 – 평당가격이 -0.55로 높은 음의 상관관계를 가지고 있음 확인