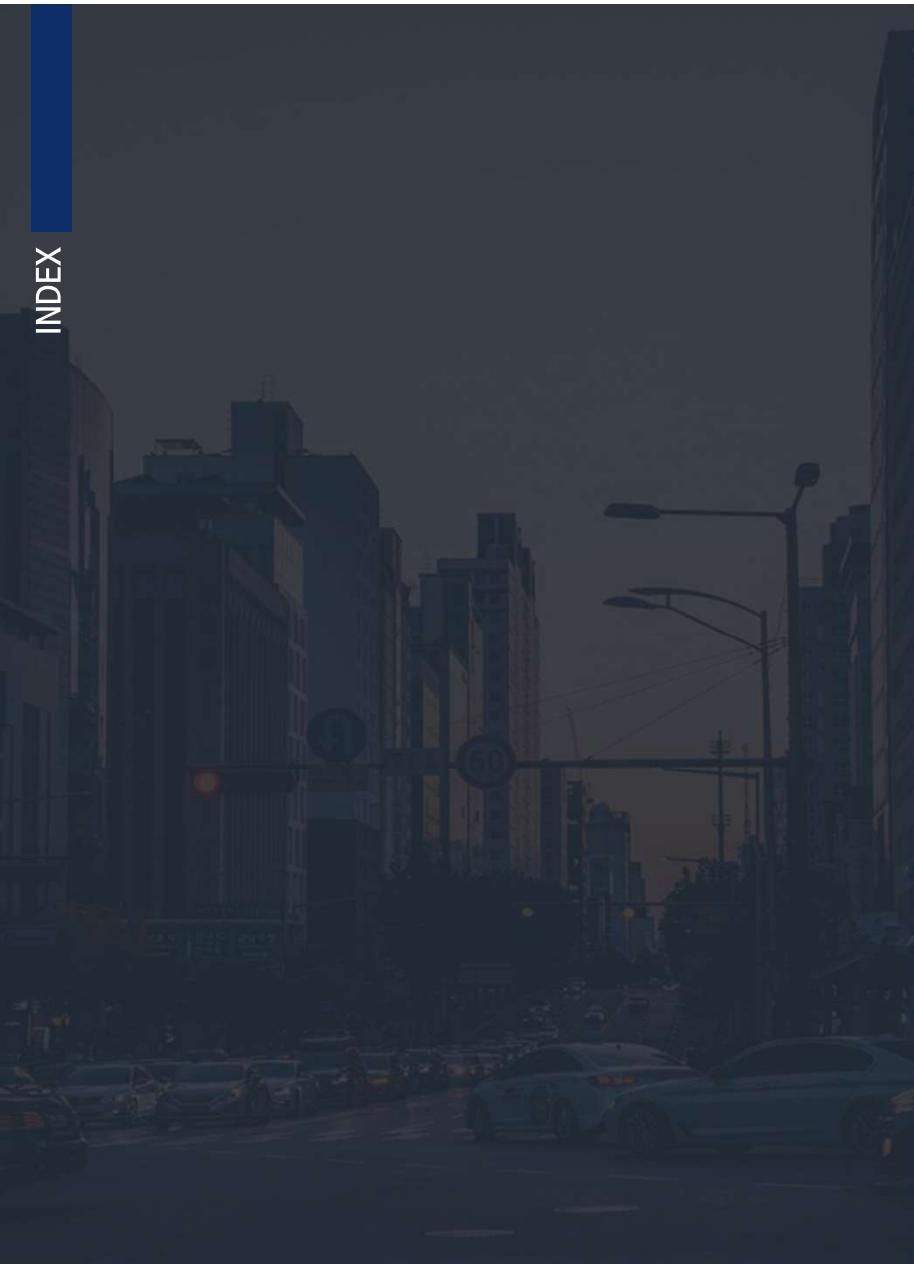


사회 초년생 특화 부동산 매물 검색 서비스

집구하기 힘들조

박현준, 김지윤, 김은비, 김지연, 박성준, 이서현



01. 서론

- 01. 분석배경
- 02. 분석목적

02. 데이터

- 01. 데이터 수집 및 변수 선정
- 02. EDA
- 03. 데이터 전처리

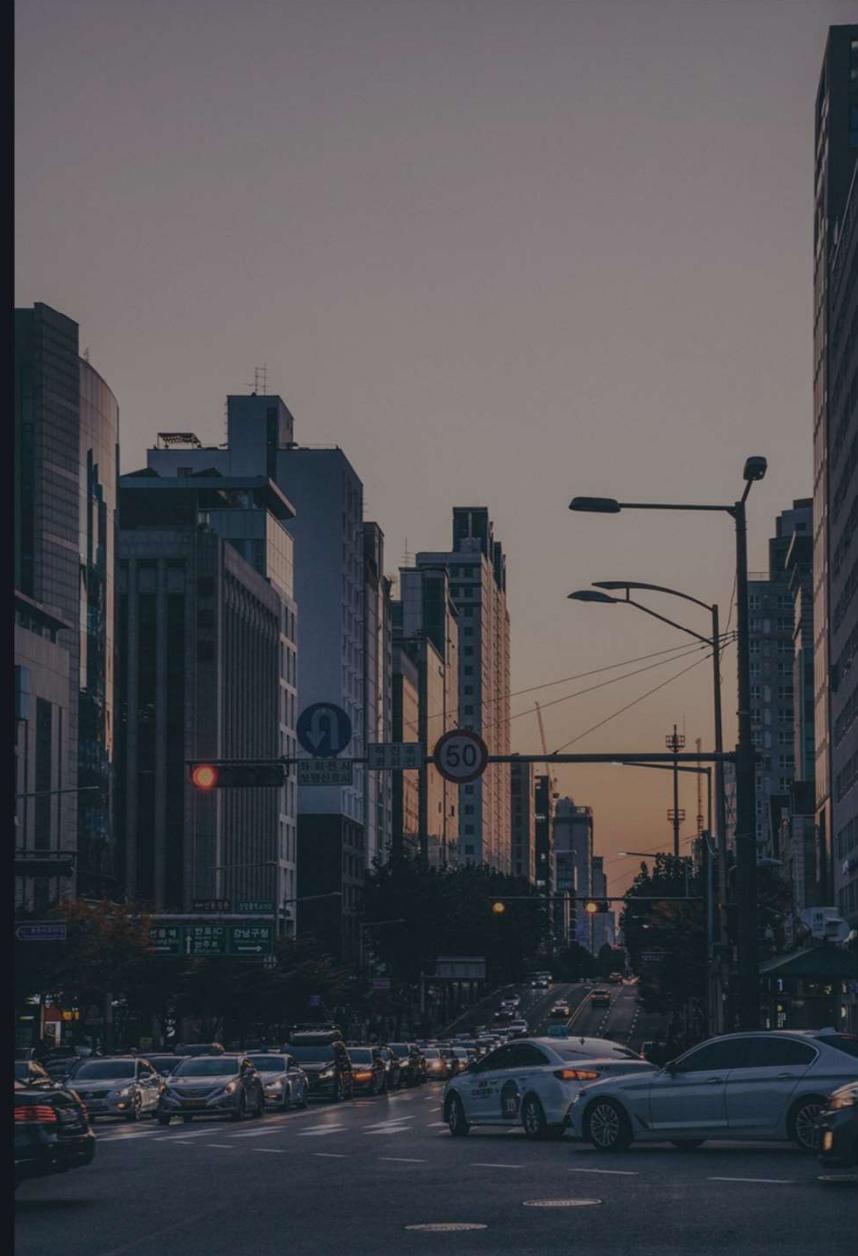
03. 모델링

- 01. 군집 분석
 - (1) DBCSAN
 - (2) K-MEANS
- 02. 회귀 분석
 - (1) 선형회귀모델
 - (2) Tree 기반 모델
- 03. 텍스트 분석
 - (1) LDA 토픽 분석
 - (2) Word Cloud

04. 결론

- 01. 서비스 가상 시나리오
- 02. 분석 시사점 및 기대효과

1. 서론



01. 프로젝트 배경

- 문제 현황

- 학업과 취업을 이유로 타 지역에서 수도권으로 유입되는 청년들의 경우, 첫 정착 지역 탐색에서 정보가 부족
- 부동산 지식과 경험이 부족한 청년들은 해당 매물이 조건에 맞게 가격이 잘 책정되었는지 알기 힘듦
- **기존 매물 정보 서비스의 경우 매물 자체의 특징은 제공하지만, 지역의 거주, 연령대, 교통의 편의성, 문화, 편의시설 등 개인의 라이프스타일에 관련한 특징은 부족하거나 확인하기 어려운 실정**
- 한정된 예산 안에서 청년들의 선호요소를 고려하여 가격과 주거 만족도를 동시에 상승시킬 매물의 탐색 필요성 대두

- 주제 선정

“ 청년들의 예산 뿐만 아니라 *라이프 스타일까지 고려한 매물을 찾도록 도와주자 ”

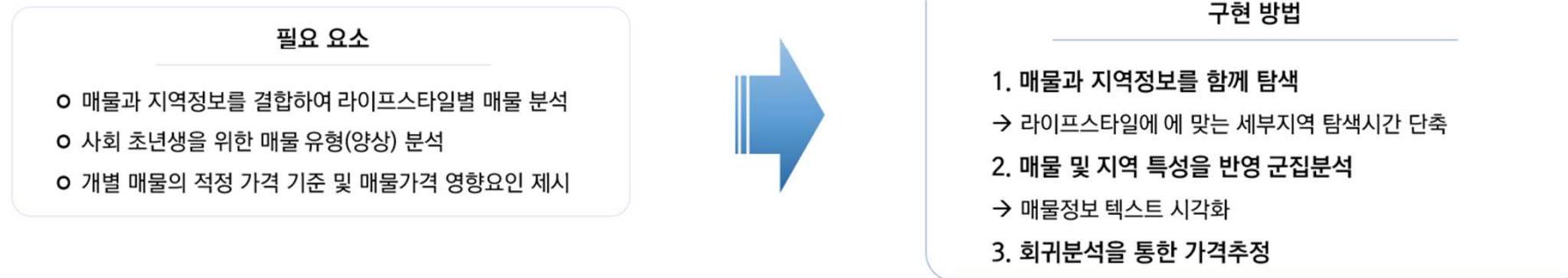
*편의시설 수, 교통, 안전, 지하철까지 거리 등 지역, 생활영향 요인

→ 매물과 지역정보를 결합하여 라이프스타일에 맞는 매물 탐색

[군집 및 텍스트 분석] 부동산 지식과 경험이 없는 사회초년생을 위한 매물양상 분석

[회귀분석] 가격적절성 평가 및 매물가격 영향요인 제시

02. 분석 목적



1. 매물 자체에 대한 정보 뿐 아니라 매물이 속한 지역의 정보를 함께 분석

사용자의 선호 주거환경(매물특징 + 지역특징)에 맞게 지역과 매물 조회 가능

사용자가 미처 생각하지 못한 지역 및 매물을 발견, 사용자가 직접 세부지역의 매물을 탐색하는 시간을 줄여줄 수 있음

2. 군집분석 전체 매물들을 지역과 매물데이터를 바탕으로 군집분석을 실시하여 매물들의 양상 파악

3. 회귀분석 가격에 대한 회귀분석을 바탕으로 매물의 가격에 영향을 미치는 요인들 분석

1) 좋은 주택의 요건도출 : 일반적으로 가격에 사람들이 선호하는 요건이 반영

→ 가격영향력이 큰 요소들을 추출하면 주택선택의 주요 요인들을 찾는데 도움이 될 것이라고 예상

2) 매물가격 적절성 평가 : 매물 가격이 조건(매물, 지역)을 고려했을 때 가격이 적절하게 형성되어 있는가?

→ 회귀분석으로 집값 추정한 뒤에 실제 매물의 가격과 비교(과대, 과소평가된 매물 파악)

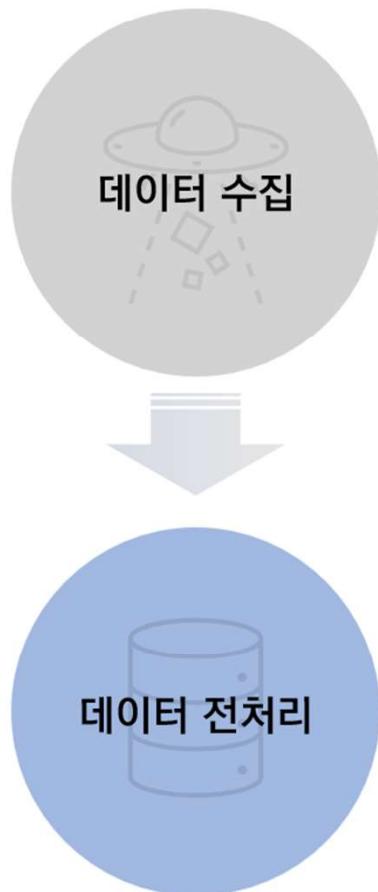
2. 데이터



01. 데이터 수집 및 적재

매물 정보와 지역 정보를 종합적으로 제시하기 위한 데이터 수집 및 전처리

매물 정보만 제공되는 기존 서비스와 차별화



매물 데이터

: 매물 종류, 가격, 옵션, 설명 텍스트 등

수집처: 직방, Kakao API

지역 정보 데이터

: 안전, 교통, 시설(편의시설 개수), 주거(1인당 연면적), 주거형태(단독주택, 아파트 비율)

수집처: 통계청 API (SGIS), 공공 데이터 포털

pandas & SQL 쿼리 활용

전처리 후 pymysql 활용 MySQL 데이터베이스 적재

01. 데이터 수집 및 적재

- API 통한 데이터 수집

: 통계청 공공데이터, 통계청 SGIS API, 직방 API 수집 (지역 코드 기준으로 Join)



- GeoHash 지도에서 지역 선택
- 직방 API를 통해서 매물 정보 크롤링 (서울, 경기도, 인천 한정)

```
GeoHash Explorer
wydm
Go

geo_list = ['wydr', 'wydn', 'wydq', 'wydw', 'wydy', 'wydj', 'wydm', 'wydt',
            'wydh', 'wydk', 'wyds', 'wydu', 'wyd5', 'wyd7', 'wyd3', 'wydg',
            'wyd4', 'wyd6', 'wydd', 'wygv']

apt_type_list = ['ws', 'js']

for geo in geo_list:
    for apt_type in apt_type_list:
        csv_file_name = f'{geo}_{apt_type}.csv'
        filepath = os.path.join(csv_dir, csv_file_name)

        if os.path.isfile(filepath):
            print(f'{csv_file_name} already exists')
        else:
            crawl(geo, apt_type)
```



- 기존 데이터의 위도 · 경도 데이터를 활용하여 카카오 API 데이터 수집
- 해당 매물로부터 가장 가까운 시설까지의 거리 데이터 수집

```
url = "https://dapi.kakao.com/v2/local/search/keyword.json"
queryString = {"query": "대형마트",
               "x": df['x_w84'][0],
               "y": df['y_w84'][0],
               "category_group_code": "MT1"}
header = {'Authorization': f'KakaoAK {API_KEY}'}

response = requests.get(url, headers=header, params=queryString)
tokens = response.json()
tokens['documents'][0]['distance']
```

supermarket_dist	convenience_store_dist	school_dist	subway_dist	cultural_venue_dist	public_institution_dist	hospital_dist
3312	305	254	645	562	692	906
2752	49	620	506	392	680	204
2774	31	613	521	371	682	218
4285	305	636	562	1222	270	1162
2808	18	442	477	267	538	216
2016	120	452	378	1057	758	578
1578	13	576	266	1514	843	338
2103	175	529	286	968	673	492
2808	18	442	477	267	538	216

01. 데이터 수집 및 적재

• 데이터 피처명 & 예시

: 통계청 공공데이터, 통계청 SGIS API, 직방 API, Kakao API 수집 (지역 코드 기준으로 Join)

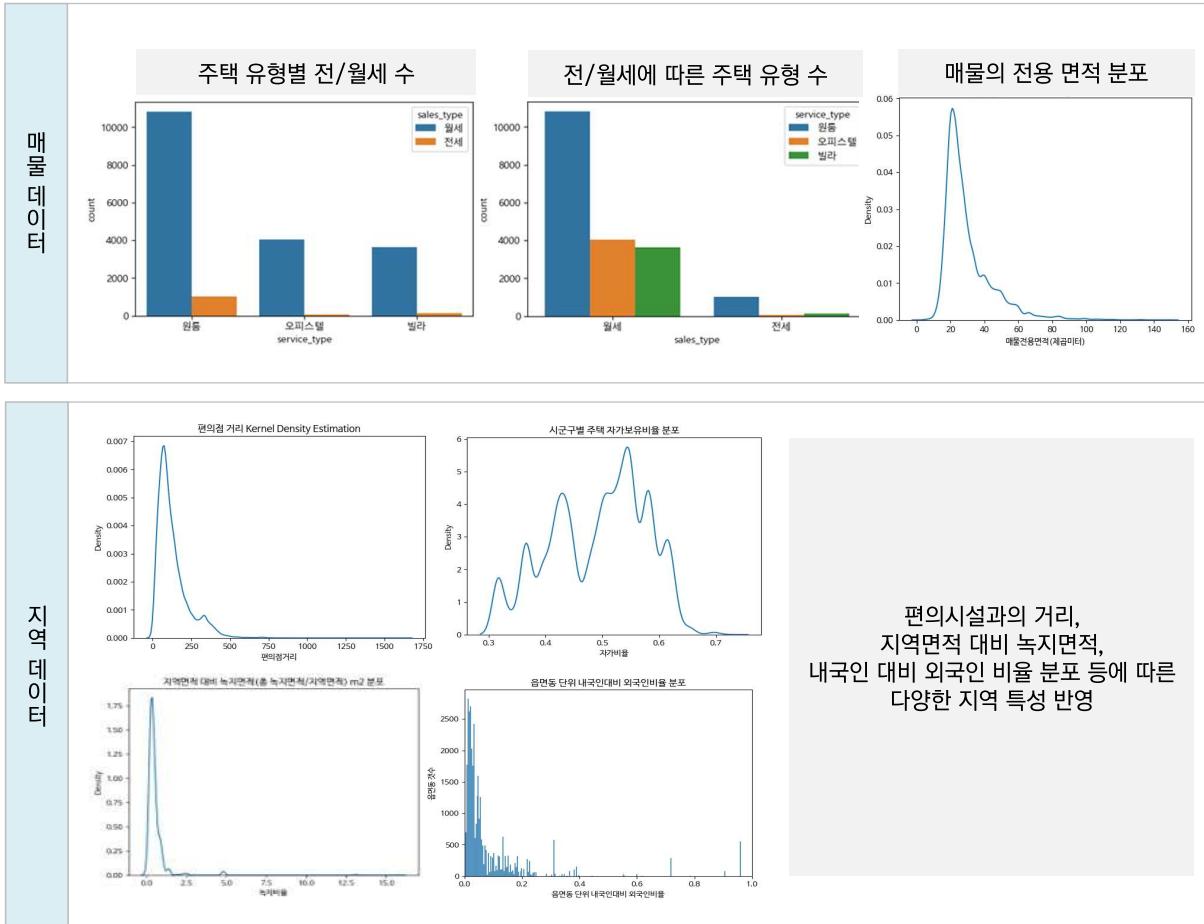
매물 정보	
고유번호	매물 id
주소 코드	시군구 코드, 읍면동 코드
주소	세부주소, 시도, 시군구, 읍면리동, X좌표, Y좌표,
매물 종류	빌라/오피스텔/원룸, 월세/전세
가격	임대료, 보증금, 관리비
옵션	관리비에 포함된 옵션, 옵션 개수
시설	엘리베이터
인근 시설	대형마트, 편의점, 학교, 문화시설, 공공기관, 병원
교통	근처 지하철 역
매물 설명	매물 상세설명, 매물 제목



지역 정보	
주소 코드	시군구 코드, 읍면동 코드
주소	시도, 시군구, 읍면동
면적	다세대 1인당 연면적, 단독주택 1인당 연면적, 비거주건물 거주 1인당 연면적, 아파트 1인당 연면적, 연립주택 1인당 연면적, 오피스텔 1인당 연면적
인구	총인구, 인구밀도, 성비, 내국인대비 외국인비, 65세 이상 인구수 비율, 노령화지수, 인구 수 대비 총 사업체 수, 청장년 인구 수 비율, 순 이동률
가구	평균 가구원 수, 총 가구, 집합가구 비율, 일반가구 비율, 친족가구 수 비율, 1인가구 수 비율
주택 유형	단독주택 비율, 아파트 비율, 다세대 비율, 비거주용 건물내 주택 비율, 주택 이외의 거처 비율
환경	지역면적 대비 공원 면적, 지역면적 대비 녹지 면적, 미세먼지
안전	지역 안전지수
시설	인구 10만명 당 편의시설 수, 쇼핑시설 수, 잡화점 수, 음식점 수, 의료시설 수, 문화시설 수, 체육시설 수

02. EDA

- 데이터 특징 파악 통한 분석 방향, 방법 설정



매물 데이터

- 원룸, 월세 매물이 압도적으로 많음
- 분포가 집중되어 있는 데이터가 많음

지역 데이터

- 시군구, 읍면동에 따른 다양한 지역별 특징

매물+지역 데이터 특징

- 전체 데이터의 분포 고르지 X, 치우친 데이터 O
- 왜도, 첨도 탐색
- 왜도, 첨도 개선을 위해 로그변환, 로버스트 스케일링 등 고려

분석방향

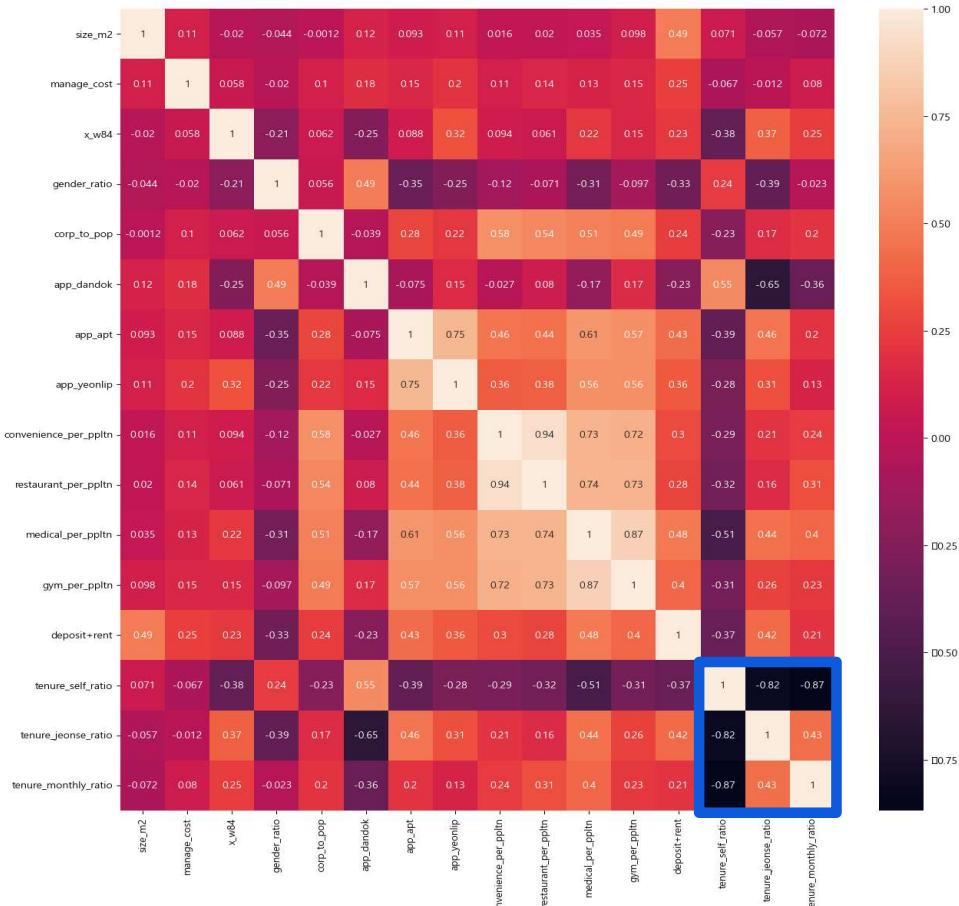
- 매물 데이터와 지역 데이터를 연결시키면
지역의 다양한 특징에 따라 매물을 분류하고,
그 특징에 따른 가격 추정 가능

분석 방법

- 군집 : 매물의 특성에 따른 군집
- 회귀 : 지역별 특성에 따른 매물 가격 추정
- 텍스트 분석: 토픽 모델링

02. EDA

- 데이터 상관관계 확인 (Heatmap)



전세비율과 자가비율 상관관계가 높음
→ 하나가 증가하면 다른 하나 감소 (Trade-off)
→ 변수 선택 고려

03. 데이터 전처리

매물 데이터

1. 매물-지역 데이터 간 지역 코드 통일

- 매물 데이터 : 법정동 코드
- 지역데이터 : 행정동 코드
- 변환 방식: 매물 데이터의 주소 변수를 좌표로 변환하여
매물 데이터의 법정동 코드를 행정동 코드로 변환

시군구	행정구역명	법정동	행정기관코드	법정동코드
서울특별시	서울특별시	서울특별시	1100000000	1100000000
종로구	종로구	종로구	1111000000	1111000000
종로구	청운효자동	청운동	1111051500	1111010100
종로구	청운효자동	신교동	1111051500	1111010200
종로구	청운효자동	궁정동	1111051500	1111010300
종로구	청운효자동	효자동	1111051500	1111010400
종로구	청운효자동	창성동	1111051500	1111010500
종로구	청운효자동	통인동	1111051500	1111010800
종로구	청운효자동	누상동	1111051500	1111010900
...

지역 정보 데이터

1. 인구, 가구 수 등 절대 수치형 데이터의 상대적 비교 위한 비율화

예)

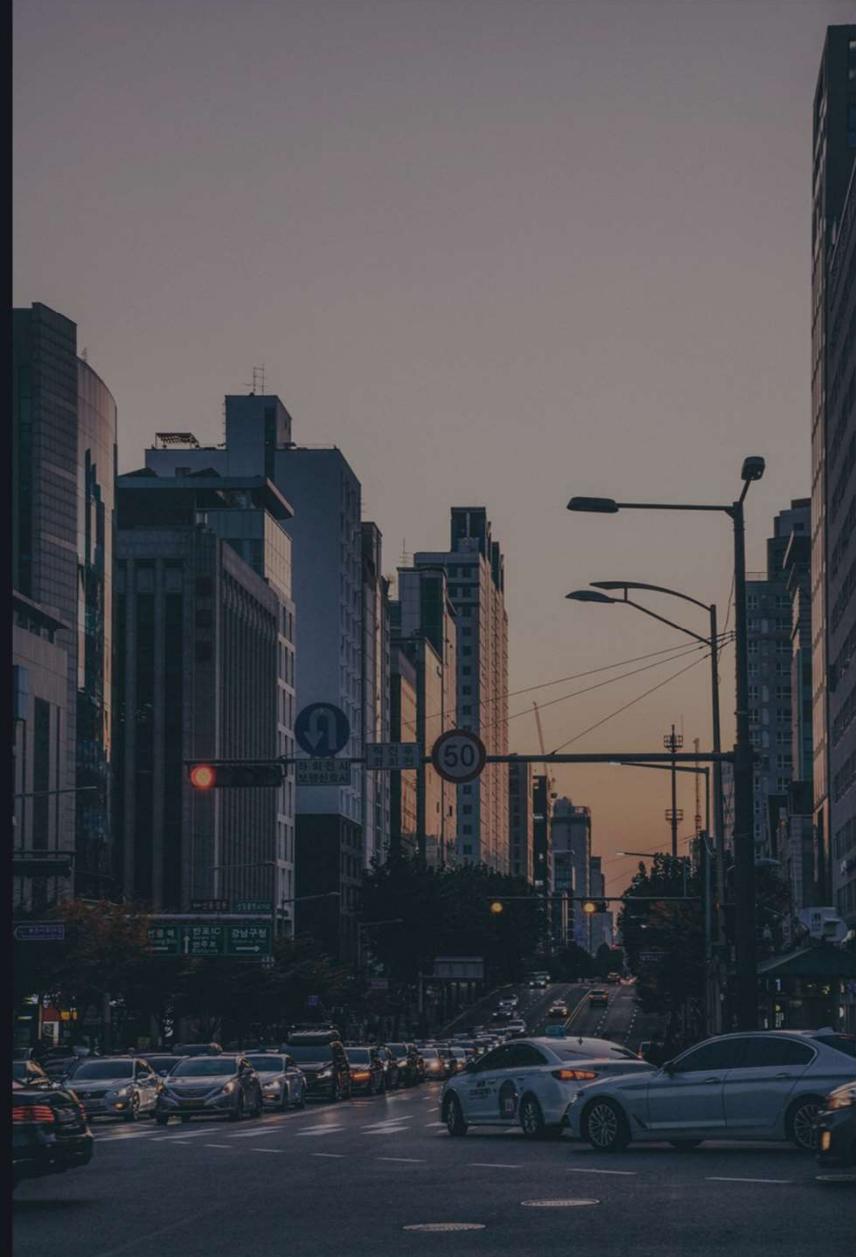
오리지널 데이터	변환된 데이터
점유유형 중 자가 총합	총 점유유형 중 자가 비율 (자가 / 점유유형 총계)
점유유형 총계	
총 사업체 수	총 인구수 대비 사업체 수 (총 사업체 수 / 총 인구수)
총 인구수	
편의시설 수	1인당 편의시설 수 (편의시설 수 / 총 인구수)
총 인구수	
...	...

2. 지하철, 옵션, 관리비 포함 등 텍스트 데이터의 수치 변환

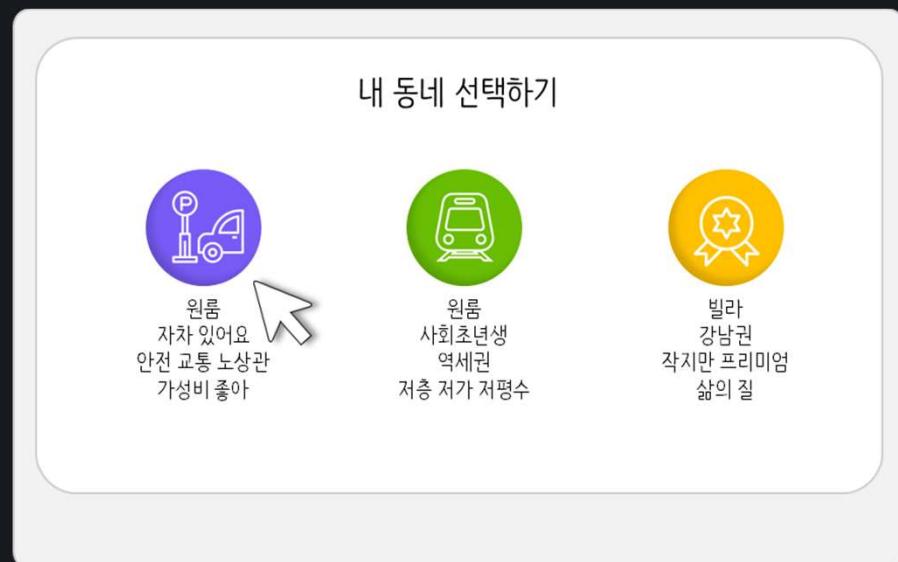
예)

오리지널 데이터	데이터	변환된 데이터
지하철	지행역(1호선), 동두천중앙역(1호선)	2
옵션	에어컨, 냉장고, 세탁기	3
관리비 포함 요소	전기세, 수도, 인터넷	3

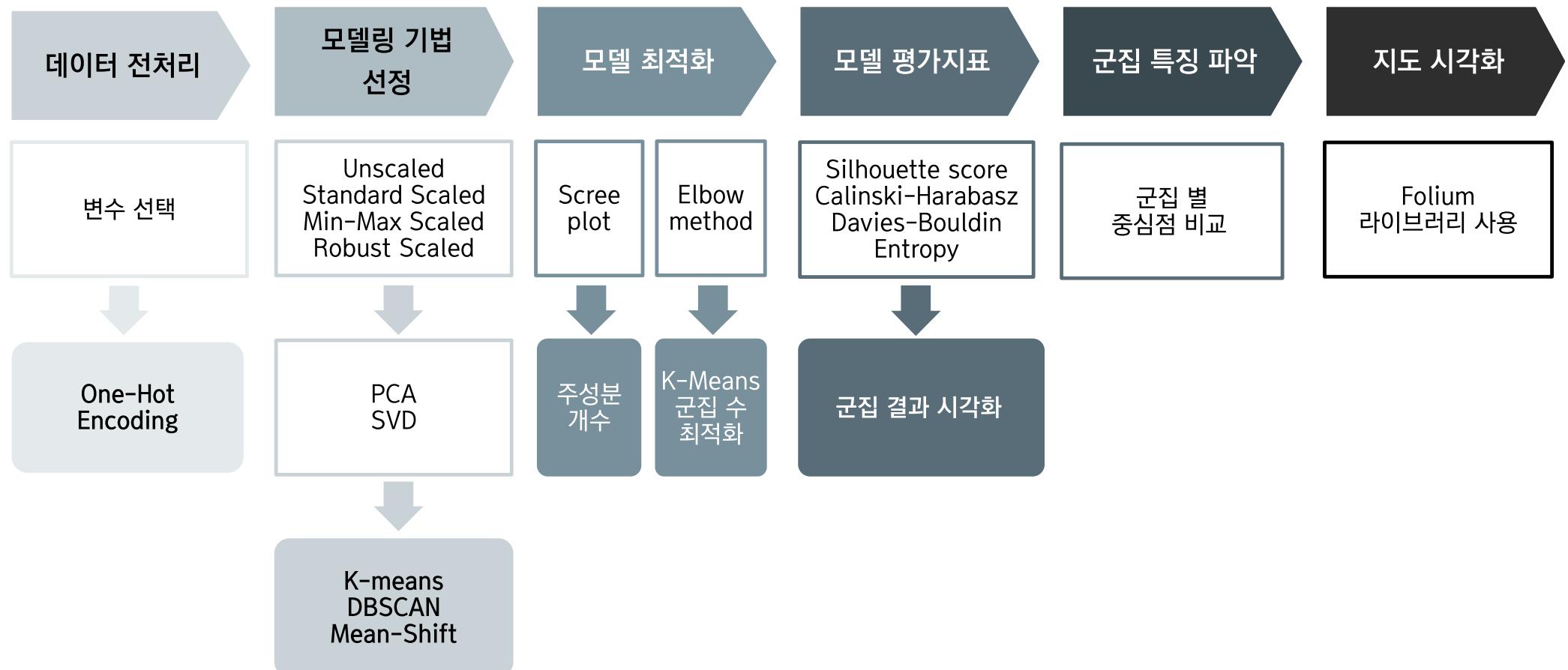
3. 모델링



03-01. 군집



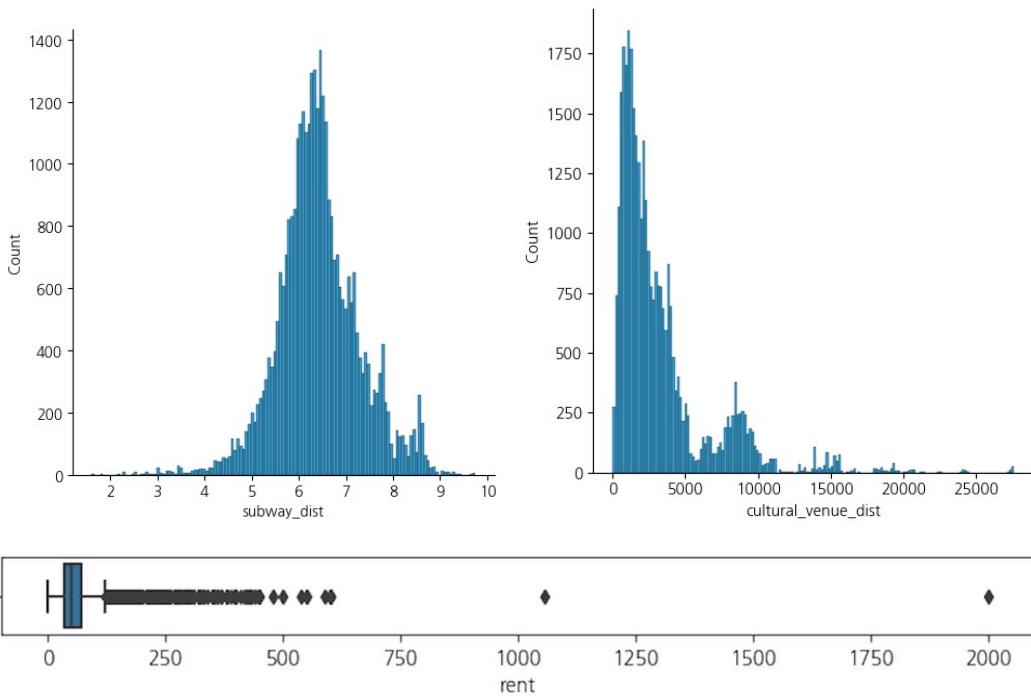
01. 군집 분석 프로세스



02. 군집 데이터 전처리

- 데이터의 특징 EDA

school_dist	4037
public_institution_dist	4278
hospital_dist	4706
supermarket_dist	14268
subway_dist	17040
cultural_venue_dist	27603
ppltn_dnsty	49681



- 데이터의 피처 중 값이 수만 단위로 늘어나는 경우 존재
→ 스케일링 필요: KMeans 같은 알고리즘은 스케일에 민감해서
→ 모든 특성이 동일한 범위 (0~1) 내에 위치하도록 변환, 군집
알고리즘에서 각 특성이 동등한 가중치를 가질 수 있도록 함
- 첫번째 그래프: 데이터가 정규분포에 가까움
→ 이 경우, StandardScaler 사용이 이상적
- 두번째 그래프: 데이터가 정규분포에서 멀어짐
→ 이 경우, Min-Max 스케일링 사용이 적절, 기존 분포를 유지
- 박스플롯: 이상치가 존재함을 확인
→ 이 경우, 이상치에 강한 Robust Scaling 사용 고려

스케일링

Standard Scaler
Min-Max Scaler
Robust Scaler

02. 군집 데이터 전처리

• 지역 안전 지수 재산정

: 행정안전부 지역안전지수 등급 산정 기준을 바탕으로 동일 행정구역 내 안전등급 → 수도권 내 안전등급으로 재산정

□ 지역안전지수 산출식

위해	취약	경감	의식
결과 지표, 감축 필요 분야별 사망자수 및 발생건수 등	위해 발생의 인적, 물적요인이 되는 사회환경 지표로 관리 필요	위해 발생 예방 및 대응하기 위한 지자체 노력 지표	위해 발생 예방 및 대응하기 위한 주민 노력 지표

= 100 - (위해지표 × 50%) - (취약지표 × 10%) + (경감지표 × 20%) ± (의식지표 × 20%)

※ 등급은 광역 시/도, 기초 시/군/구 5개 그룹별로 1등급 10%, 2등급 25%, 3등급 30%, 4등급 25%, 5등급 10% 비율로 산정

변환 방식 : 등급 환산

```
safety_range = list(range(7, 28))

safety_df = pd.DataFrame(safety_range, columns=['safety_value'])

percentiles = [0.10, 0.35, 0.65, 0.90, 1]

safety_list = []

safety_list = safety_df['safety_value'].quantile(percentiles).tolist()

for i in range(len(df)):
    for j in range(len(safety_list)):
        if df['safety_idx'][i] <= safety_list[j]:
            df['safety_idx'][i] = j
            break
```

변환 후

16.0	5479
12.0	4062
20.0	3389
13.0	3334
14.0	2935
18.0	2704
15.0	2011
17.0	1915
21.0	1716
19.0	1642
9.0	1612
11.0	1314
10.0	1214
27.0	241
26.0	229
8.0	177
24.0	120
22.0	60
23.0	57
7.0	36

Name: safety_idx, dtype: int64



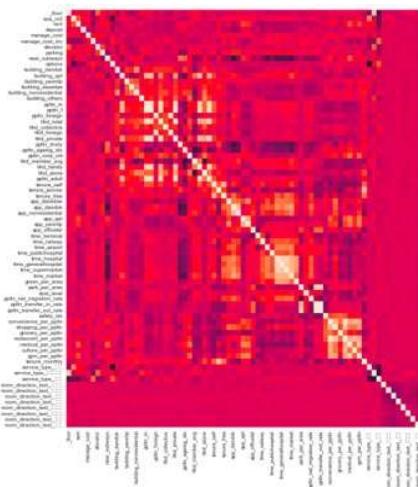
2.0	17140
1.0	12859
3.0	1953
0.0	1825
4.0	470

Name: safety_idx, dtype: int64

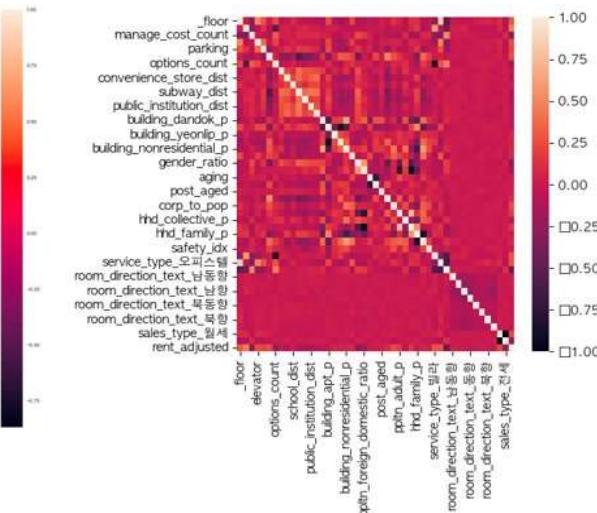
02. 군집 데이터 전처리

• 변수 선택

히트맵 분석을 통해 군집화 모델에 넣을 변수들 취사선택



변수 선택 전 히트맵



변수 선택 후 히트맵

유사도 높은 변수, 군집에 영향이 적다고 여겨지는 변수 제외

→ 시군구 기준으로 수집한 데이터는 매물 정보 데이터 기준에는 적합하지 않다고 판단되어 제외하고 대체 가능한 데이터를 카카오 REST API를 활용해 새로 수집함

• 인코딩

One Hot Encoding:

범주형 변수들(원룸, 빌라 등 주거 형태, 집의 방향, 월세, 전세)

→ 군집 분석에 적용할 수 있는 형태로 변환

원룸	19028
오피스텔	8272
빌라	6947
Name: service_type,	

월세	32291
전세	1956
Name: sales_type,	

One Hot Encoding 적용

room_direction_text_북향	room_direction_text_서향	service_type_빌라	service_type_오피스텔	service_type_원룸	sales_type_월세	sales_type_전세
0	0	0	0	1	1	0
0	1	0	0	1	1	0
0	0	0	0	1	1	0
0	0	1	0	0	1	0
0	0	0	0	1	1	0

03. 군집 모델링 기법 선정 과정

	사용 기법
알고리즘	K-Means DBSCAN MeanShift
스케일링	Standard Scaler Min-Max Scaler Robust Scaler
차원축소	PCA SVD
평가 기법	실루엣 계수 엔트로피 Calinski-Harabasz Score Davies-Bouldin Score

Example

- 군집화 기법:
K-Means / DBSCAN / Mean Shift
- 주성분 개수:
7 / 8 / 9
- 군집 개수:
3 / 4 / 5 / 6
- 스케일링:
Unscaled / Standard / Minmax / Robust

03. 군집 모델링 기법 선정 과정

3-1) 주성분 개수 선정

Data Scaling → 주성분 개수 후보 선정

Unscaled			Min-Max Scaler		
주성분 개수	고윳값	누적기여율	주성분 개수	고윳값	누적기여율
PC2	1.010003e+07	0.9837	PC9	0.1205	0.7341
			PC10	0.1154	0.7710
			PC11	0.1042	0.8043
Standard Scaler			Robust Scaler		
주성분 개수	고윳값	누적기여율	주성분 개수	고윳값	누적기여율
PC16	1.0738	0.7180	PC7	1.4502	0.7457
PC17	1.0569	0.7405	PC8	1.1568	0.7710
PC18	0.9971	0.7617	PC9	0.9928	0.7956
PC19	0.9329	0.7816	PC10	0.8184	0.8104

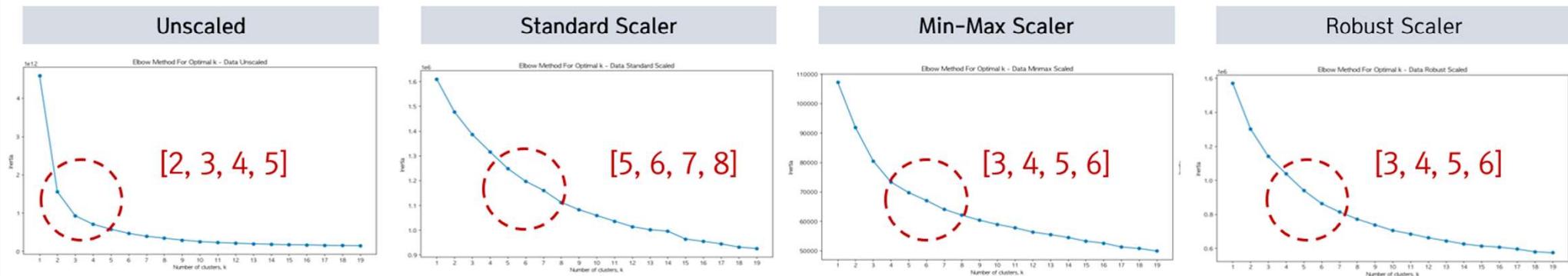
주성분 개수를 결정하기 위해 서로 기법이 다른 스케일링을 적용한 데이터에 PCA를 적용하여
누적기여율 70%~80% 의 주성분을 후보 주성분 개수로 결정

03. 군집 모델링 기법 선정 과정

3-2) 군집 개수 후보 선정 : Elbow Method → 클러스터링 모델에 적용할 n_clusters의 최적값 선정 기준

[K-Means 예시] : 군집 알고리즘에서 사용할 군집의 개수를 설정하기 위해 Elbow Method 확인

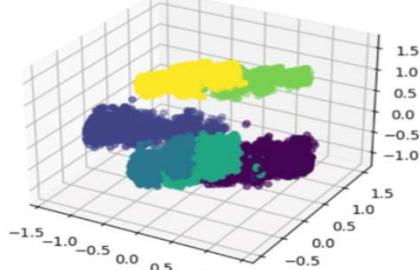
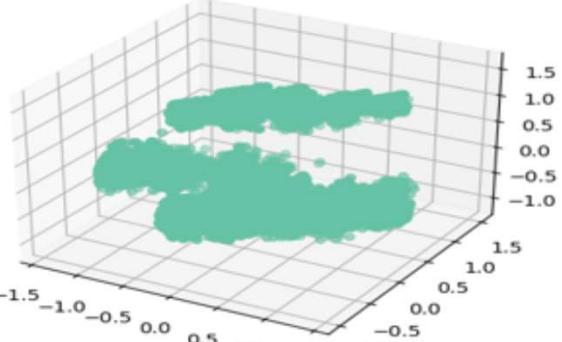
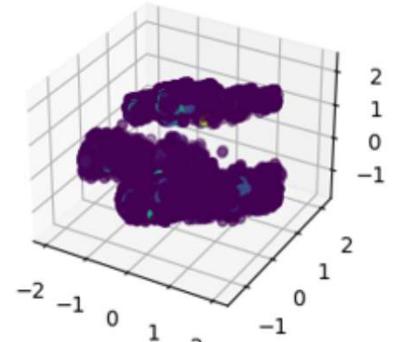
그래프가 완만해지는 점을 기준으로 하여 군집 개수 후보 정함



03. 군집 모델링 기법 선정 과정

3-3) 군집 알고리즘 선정

- 후보 주성분 개수, 군집 개수 후보, 차원축소 기법, 스케일링 기법을 조합 → 각 모델링 기법에 적용
- 각 알고리즘 별 실루엣 계수가 높은 모델 시각화 후 각 군집별로 데이터의 수가 고르게 분포되었는지 확인

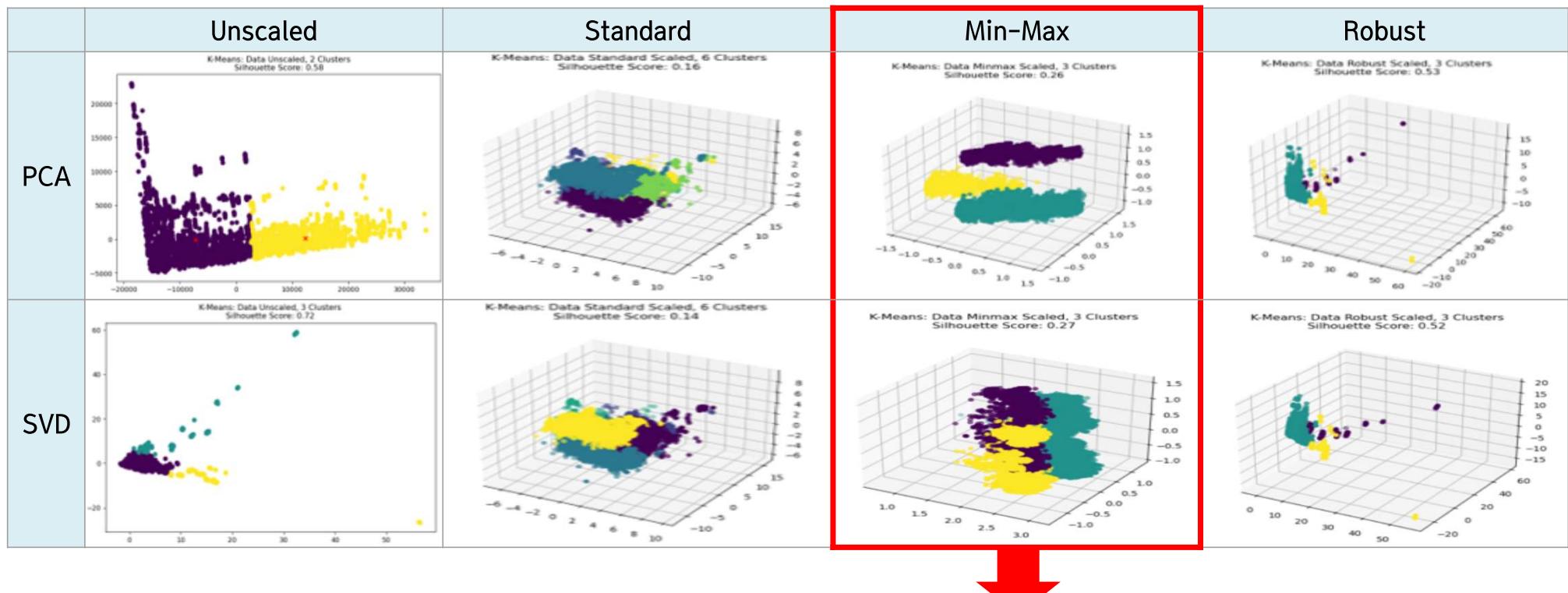
	K-Means	DBSCAN	MeanShift
시각화	<p>K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.24 Calinski Harabasz Score: 7067.18 Davies Bouldin Score: 1.65 Entropy: 10.09</p> 	<p>Silhouette Score Undefined</p> 	<p>bandwidth: 2.537, silhouette score: -0.448</p> 
실루엣 계수	0.24 (상대적으로 높진 않음)	없음	-0.448 (낮음)
군집	군집내는 밀집되어 있고 군집간 거리가 멀어보임	여러 군집으로 형성되지 않고 하나의 군집으로만 표현됨	다른 군집의 색이 희미하게 보이지만 대부분의 값들이 하나의 군집에 몰림

K-Means 선택

04. K-Means 파라미터 최적화

4-1) 최종 스케일링 기법 선정

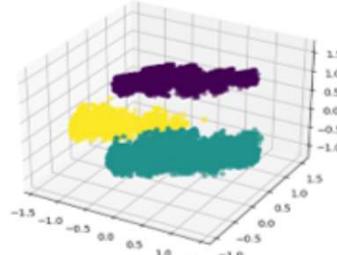
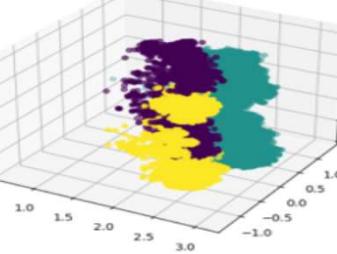
- 스케일링 하지 않은 데이터, Robust Scaler를 적용한 데이터 → 실루엣 계수가 더 높음
- 시각화 결과 Min-Max Scaler를 통해 데이터의 특성을 잘 나타낸 군집을 형성할 수 있다고 판단함



Min-Max Scaler 최종 결정

04. K-Means 파라미터 최적화

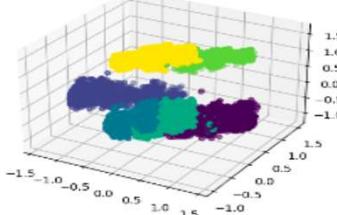
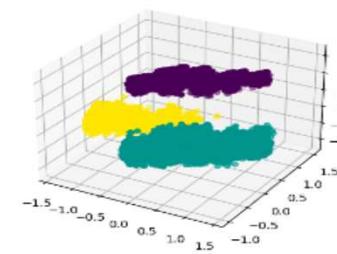
4-2) 최종 차원축소 기법 선정

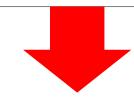
	Min-Max 결과	실루엣 계수
PCA	<p>K-Means: Data Minmax Scaled, 3 Clusters Silhouette Score: 0.26</p> 	0.26
SVD	<p>K-Means: Data Minmax Scaled, 3 Clusters Silhouette Score: 0.27</p> 	0.27



근소한 차이로 **PCA** 의 성능이 좋음

4-3) 최종 군집 개수 선정

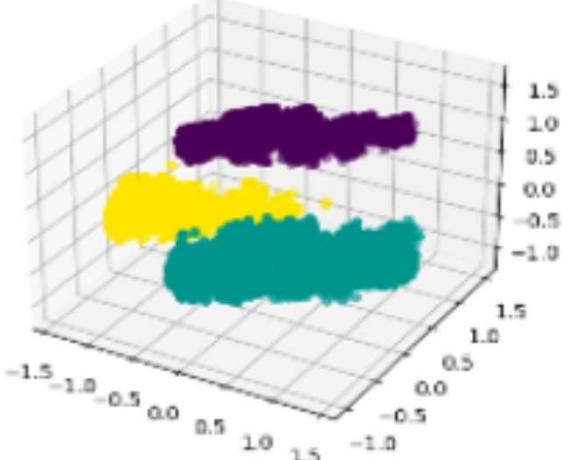
	Min-Max 결과	실루엣 계수	군집 수
	<p>K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.24 Calinski Harabasz Score: 7067.18 Davies Bouldin Score: 1.65 Entropy: 10.09</p> 	0.24	6
	<p>K-Means: Data Minmax Scaled, 3 Clusters Silhouette Score: 0.26 Calinski Harabasz Score: 8750.86 Davies Bouldin Score: 1.55 Entropy: 10.16</p> 	0.26	3



모델 파라미터의 경우 **n_clusters = 3** 으로 설정했을 때 가장 성능이 좋음

05. 군집 모델링 결과

- 최종 군집 모델

	최종 선택	시각화
알고리즘	K-Means	<p>K-Means: Data Minmax Scaled, 3 Clusters Silhouette Score: 0.26 Calinski Harabasz Score: 8750.86 Davies Bouldin Score: 1.55 Entropy: 10.16</p>
스케일링	Min-Max scaler	
차원축소	PCA	
군집 개수	3	 A 3D scatter plot visualizing the clustered data. The x-axis ranges from -1.5 to 1.5, the y-axis from -1.0 to 1.5, and the z-axis from -1.0 to 1.5. Three distinct clusters of points are shown: one cluster in purple at the top, one in yellow in the middle, and one in teal at the bottom. The axes are labeled with numerical values.

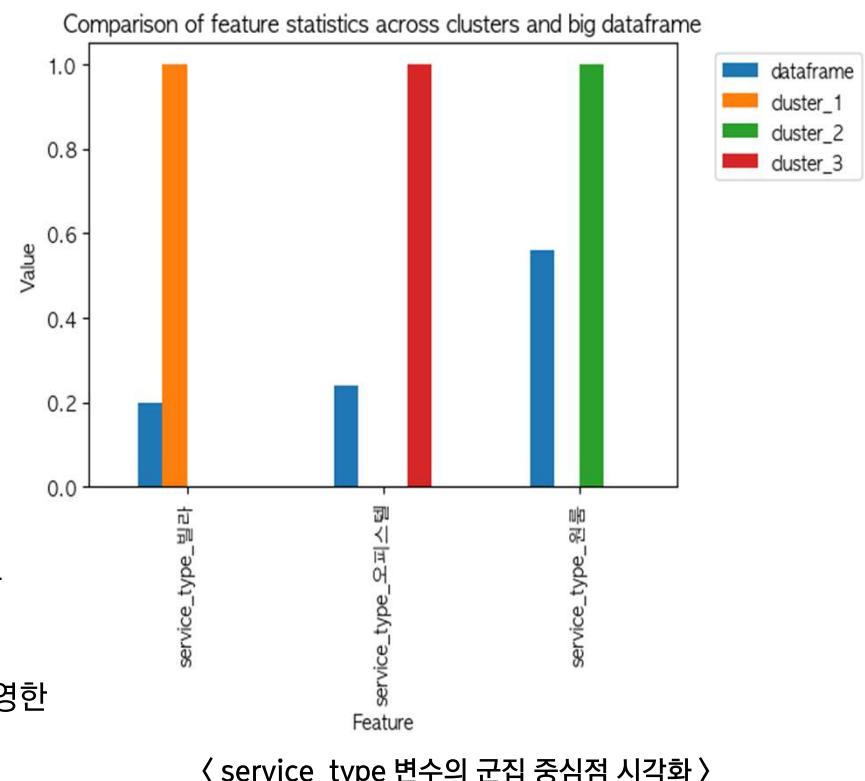
05. 군집 모델링 결과

5-1) 군집 중심점 분석

- 최종 군집 결정 후 군집의 대표적인 특징을 파악하기 위해 각 군집의 중심점을 변수별로 시각화
- 주택 유형별(service_type 변수)로 3개의 군집이 형성됨 (원룸, 빌라, 오피스텔)
- 군집에 사용된 원본 데이터가 주택 유형별 특징에 영향을 많이 받았기 때문에 이러한 특징을 토대로 3가지 군집이 형성된 것으로 판단

5-2) 군집 결과 개선 방향

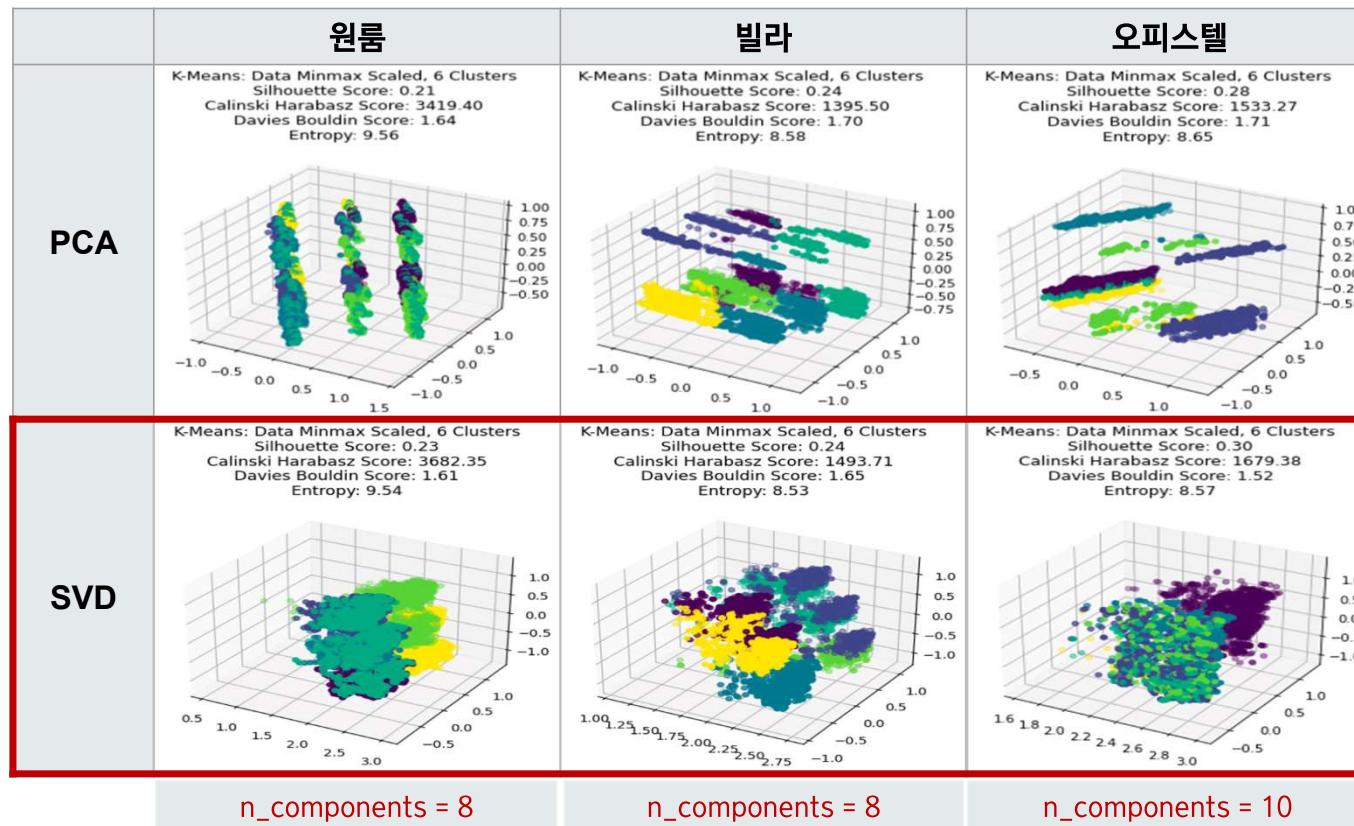
- 군집분석 목적
- 매물의 특성과 매물이 위치하는 지역의 특성을 반영한 군집을 형성하여 사용자에게 제공
→ 사용자가 자신의 성향과 맞는 군집을 선택하여 매물 정보 조회
- 사용자에게 좀더 자세한 정보를 제공하고 service_type이 아닌 다른 변수의 특성을 반영한 군집을 보여주기 위해 service_type별로 클러스터링 진행



06. 군집 모델링 발전

6-1) 차원 축소 기법 비교

- 차원 축소 기법을 적용한 군집화 모델 실행 결과를 시각화로 확인



기존 모델과 다르게 새로운 모델에서는
SVD를 사용하기로 결정

06. 군집 모델링 발전

6-1) 주성분 특성 파악

Component 1:
sales_type_월세: 0.3916
hhd_private_p: 0.3861
ppltn_adult_p: 0.3417
parking: 0.2693
options_count: 0.257

Component 2:
aged: 0.6691
aging: 0.6008
parking: 0.2563
ppltn_dnsty: 0.1355
near_subways_count: 0.1337

Component 3:
parking: 0.587
elevator: 0.5301
aging: 0.3945
near_subways_count: 0.1811
hhd_alone_p: 0.1724

Component 4:
room_direction_text_남향: 0.8616
room_direction_text_동향: 0.4221
room_direction_text_남서향: 0.1708
room_direction_text_서향: 0.0981
room_direction_text_남동향: 0.0893

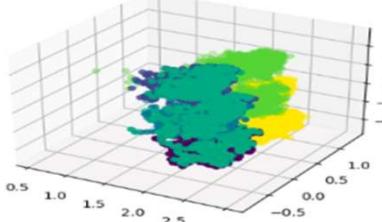
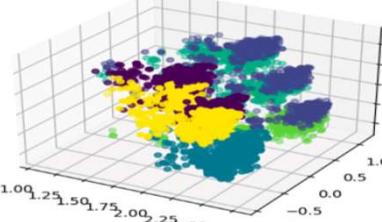
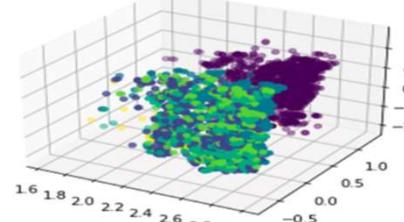
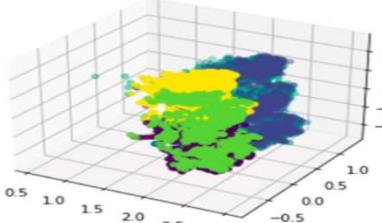
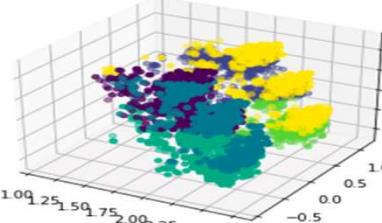
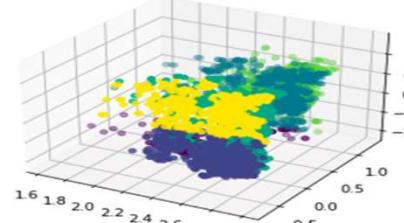
원룸			
components1	components2	components3	components4
sales_type_월세	aged	parking	room_direction_text_남향
hhd_private_p	aging	elevator	room_direction_text_동향
ppltn_adult_p			room_direction_text_남서향
			room_direction_text_서향
			room_direction_text_남동향
인구, 가구 특성 변수	고령 인구 비율 변수	건물 설비 변수	주택 방향 변수

- 실루엣 계수를 비교하며 찾은 최적의 주성분 개수 8, 10개
- 변수 별 주성분 요소 값(components_)을 계산하여 각 주성분을 설명하는 변수를 파악하고자 함
- 확인 결과 같은 변수가 여러 주성분을 설명하고 있어 각 주성분의 특성을 파악하지 못하여 군집 성능이 저하된다고 예상함
→ 각 주성분을 설명하는 변수가 다른 주성분과 차별점이 있도록 차원의 개수를 줄임

06. 군집 모델링 발전

6-2) 주택 유형별 최종 군집 모델

: 차원의 개수를 4로 줄인 결과 모델의 성능이 개선됨

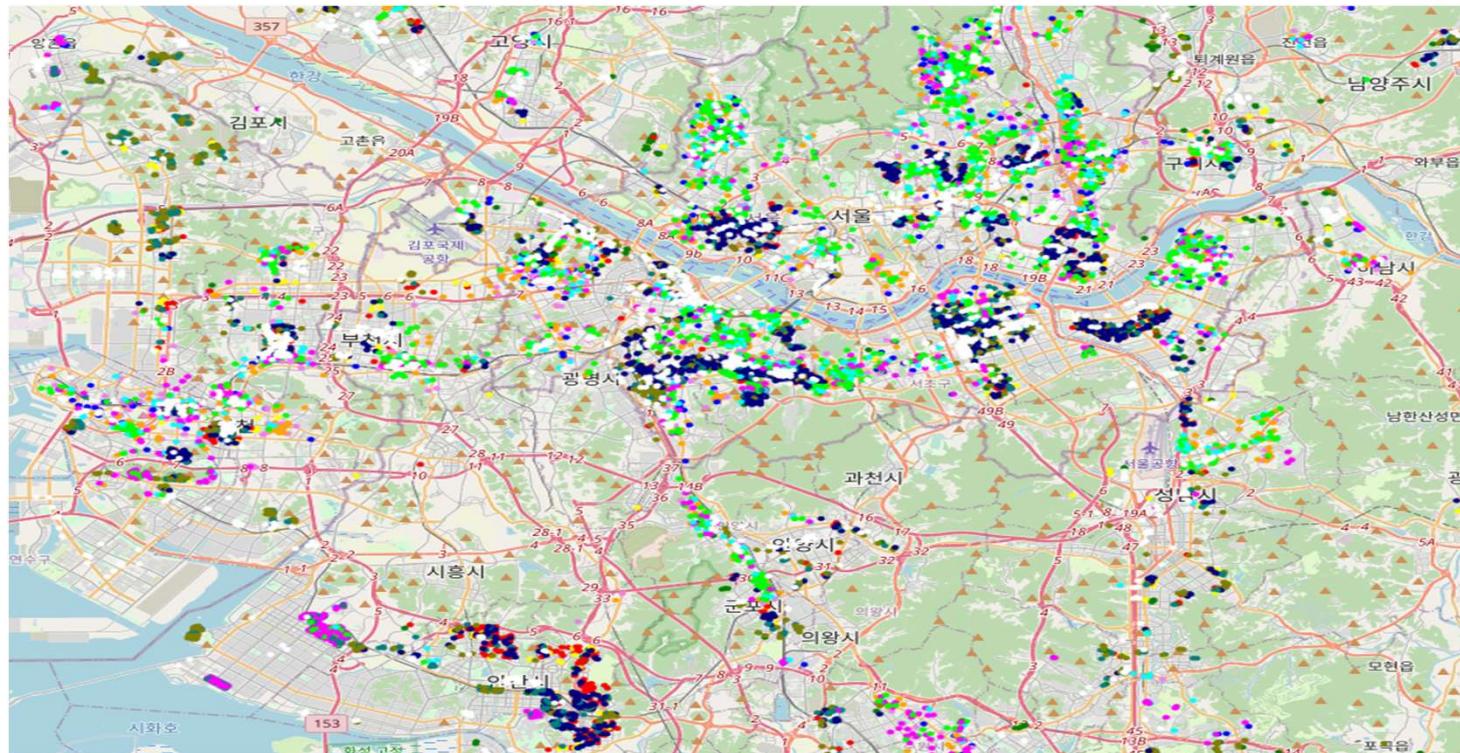
	원룸	빌라	오피스텔
차원 수 8, 10	K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.23 Calinski Harabasz Score: 3682.35 Davies Bouldin Score: 1.61 Entropy: 9.54 	K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.24 Calinski Harabasz Score: 1493.71 Davies Bouldin Score: 1.65 Entropy: 8.53 	K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.30 Calinski Harabasz Score: 1679.38 Davies Bouldin Score: 1.52 Entropy: 8.57 
차원 수 4	K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.41 Calinski Harabasz Score: 11452.99 Davies Bouldin Score: 0.87 Entropy: 9.50 	K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.44 Calinski Harabasz Score: 4869.26 Davies Bouldin Score: 0.82 Entropy: 8.51 	K-Means: Data Minmax Scaled, 6 Clusters Silhouette Score: 0.51 Calinski Harabasz Score: 8001.97 Davies Bouldin Score: 0.67 Entropy: 8.70 
평가 지표 변화	Silhouette : 0.23 → 0.41 Calinski-Harabasz : 3682 → 11452 Davies-Bouldin : 1.61 → 0.87 Entropy : 9.54 → 9.50	Silhouette : 0.24 → 0.44 Calinski-Harabasz : 1493 → 4869 Davies-Bouldin : 1.65 → 0.82 Entropy : 8.53 → 8.51	Silhouette : 0.30 → 0.51 Calinski-Harabasz : 1679 → 8001 Davies-Bouldin : 1.52 → 0.67 Entropy : 8.57 → 8.70

차원의 개수를 4로 줄인 결과
모델의 성능이 개선됨

07. 군집 특징 파악

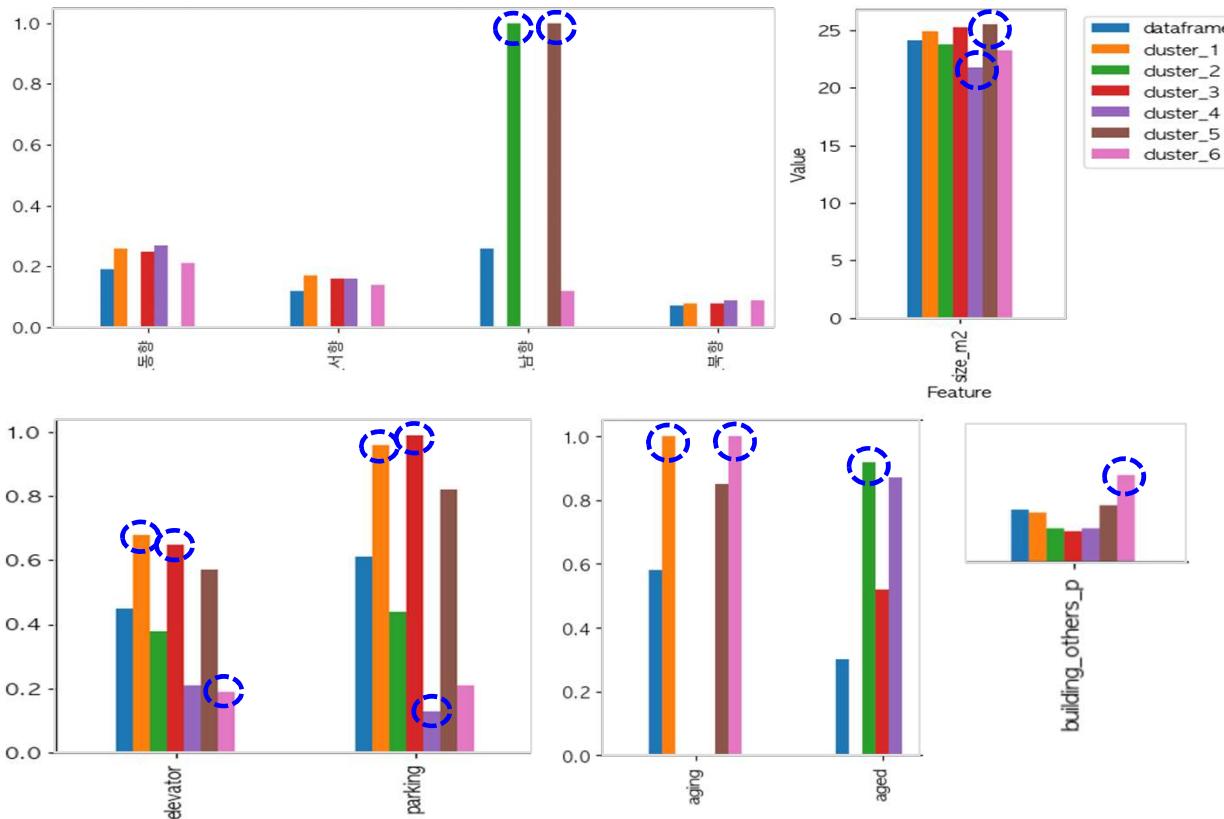
총 18개의 군집 지도 시각화

예) 빌라: 6개 군집, 원룸: 6개 군집, 오피스텔: 6개 군집



07. 군집 특징 파악

7-1) 군집 중심점 분석

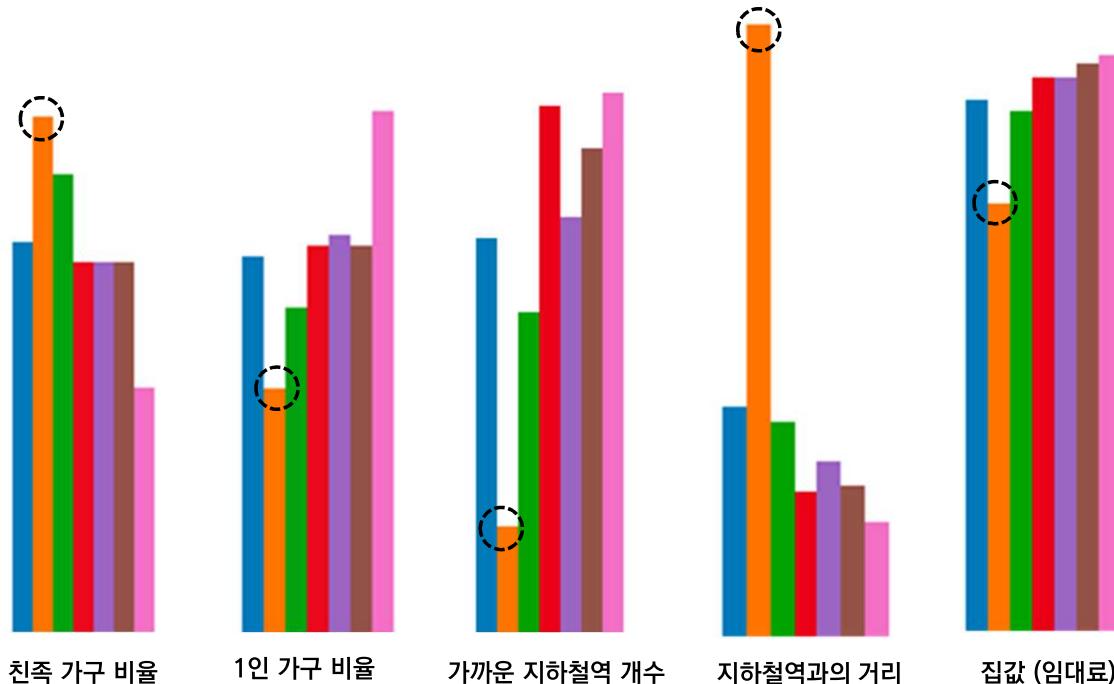


군집 별 중심점 시각화

- 군집의 특징을 파악하기 위해 44개의 변수에 대해 각 군집 중심점을 계산하여 시각화
- 시각화 결과를 바탕으로 다른 군집에 비해 높거나 낮은 변수들을 찾아냄
- 특징이 눈에 띠는 변수를 통해 군집의 특징을 정의

07. 군집 특징 파악

7-1) 군집 중심점 분석



오피스텔

Cluster 1

친족 가구 비율 높은 군집

1인 가구 비율 낮은 군집

아파트가 많은 군집

가까운 지하철 역이 적은 군집

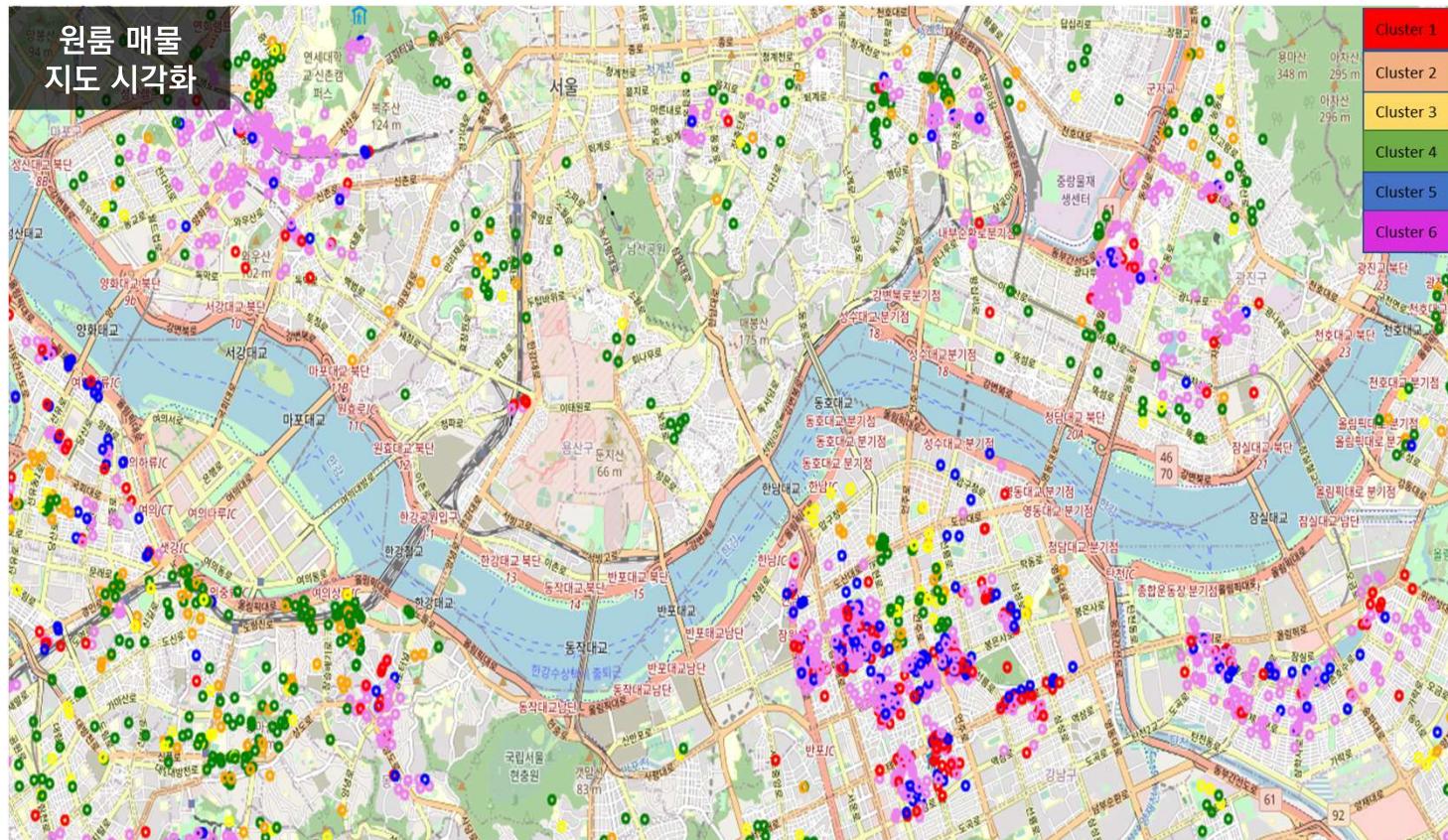
높은 총 수의 매물이 많은 군집

지하철역과의 거리가 먼 군집

집값(임대료)이 저렴한 군집

집값이 비교적 저렴하지만 교통이 좋지 않은 군집
아파트가 많아 친족 가구가 많이 사는 군집

07. 군집 특징 파악



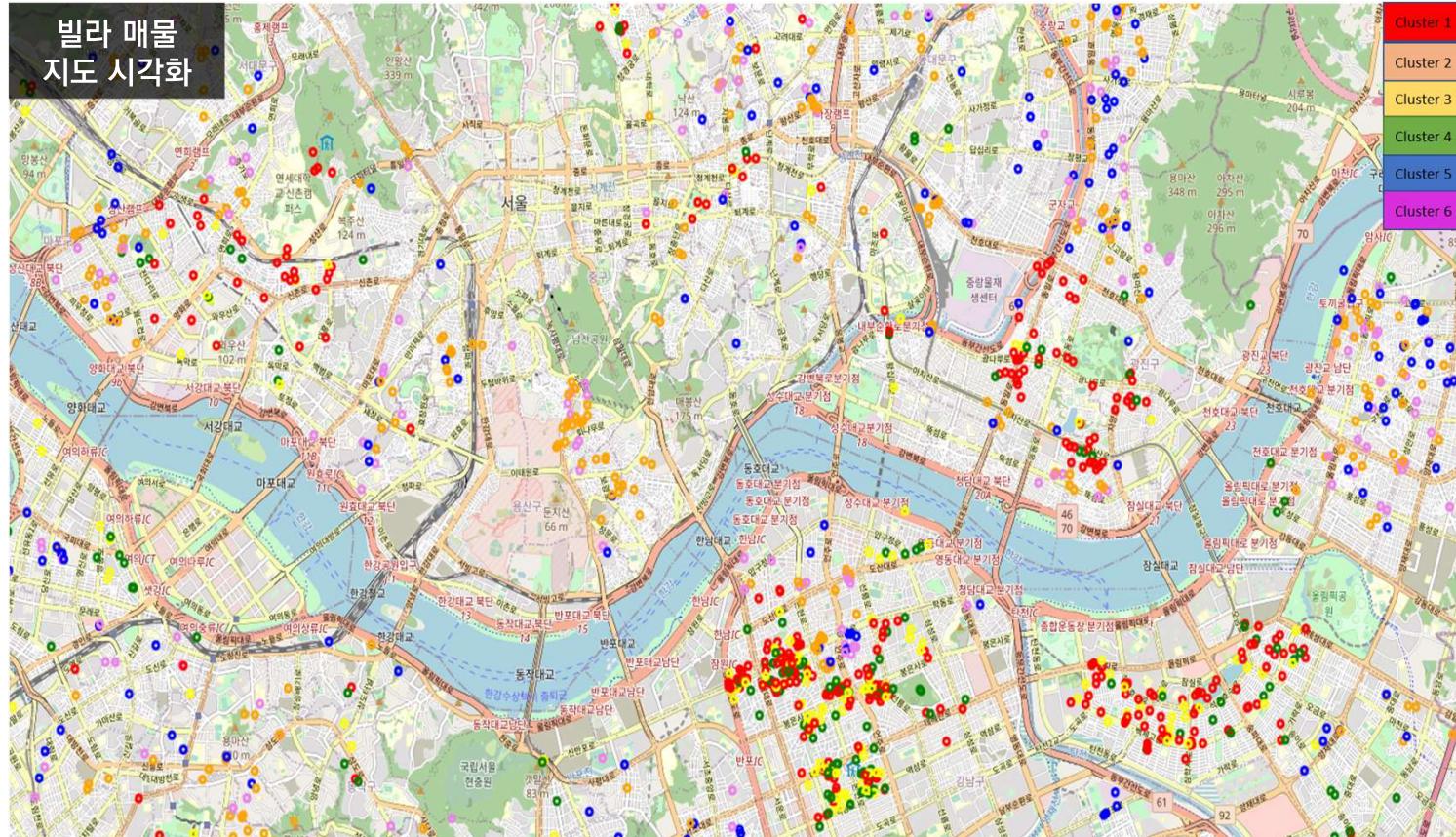
군집별 특징

Cluster 1	Cluster 4
엘리베이터 있는 매물이 많음	가까운 지하철역이 많음
주차공간이 있는 매물이 많음	증수가 낮은 매물이 많음
안전지수가 낮음	매물의 크기가 작음
가까운 지하철역이 적음	임대료가 저렴한 군집
건물 설비가 좋지만 주거 환경은 안 좋고	저층의 작은 평수의 매물이 많지만
교통이 좋지 않은 군집	고동이 편리하고 집값이 저렴한 군집

Cluster 2	Cluster 5
1인 가구 비율 높은 군집	매물의 크기가 작음
남녀성비가 낮음	지하철역 그리고 다른 시설들이랑 가까움
안전지수가 높음	인구 밀도가 높은 군집
매물의 크기가 작음	인구 밀도가 높고 작은 평수의 매물이지만 편의시설이랑 가까운 군집
자취비율이 높고 안전지수가 높지만	
작은 평수의 매물인 군집	

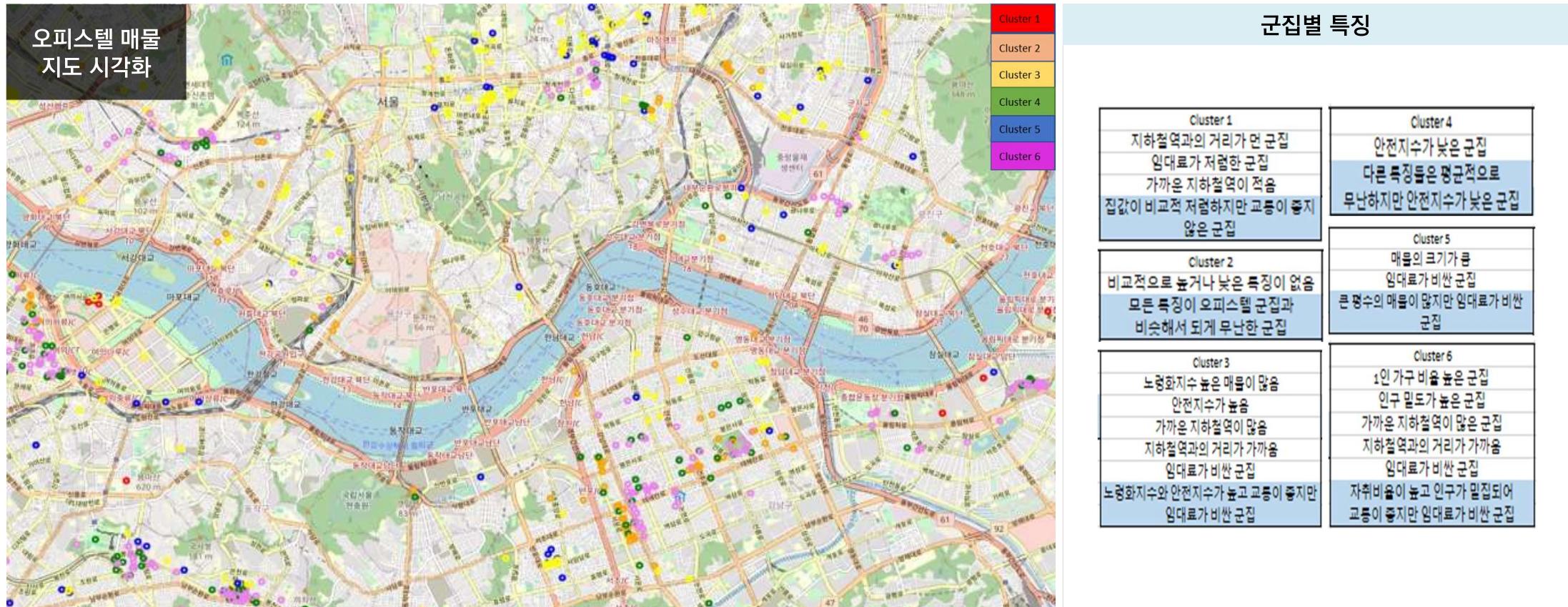
Cluster 3	Cluster 6
엘리베이터 있는 매물이 많음	안전지수가 높음
주차공간이 있는 매물이 많음	임대료가 비싼 군집
가까운 지하철역이 많음	가까운 지하철역이 가까움
남성 인구 비율이 높은 군집	지하철역이 가까움
교통이 좋지 않고 안전하지 않지만 좋은 건물	안전지수가 높고 교통이 좋지만
설비에 비해 집값이 저렴한 군집	임대료가 비싼 군집

07. 군집 특징 파악

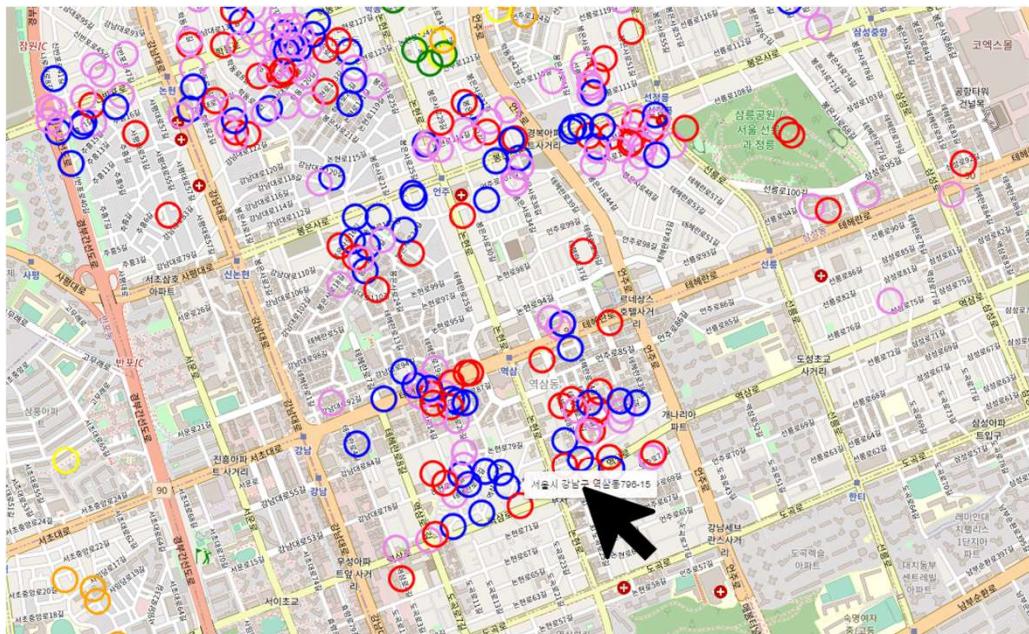


군집별 특징	
Cluster 1	노령화지수 높은 매물이 많음 엘리베이터 없는 매물이 많음 주차공간이 없는 매물이 많음 임대료가 저렴한 군집
Cluster 2	외국인 비율이 높고 안전지수가 높은 군집
Cluster 3	자취비율이 높음 안전지수가 높음 외국인 비율이 높음 외국인 비율 자취 비율이 높고 안전지수가 높은 군집
Cluster 4	노령화지수 높은 매물이 많음 엘리베이터 없는 매물이 많음 주차공간이 없는 매물이 많음 임대료가 낮은 군집
Cluster 5	엘리베이터 있는 매물이 많음 주차공간이 있는 매물이 많음 낮은 중 수의 매물이 많음 임대료가 저렴한 군집 건물 설비가 좋지 않고 낮은 중수지만 임대료가 저렴한 군집
Cluster 6	가까운 지하철역이 적음 지하철역과의 거리가 멀음 인구 밀도가 낮은 군집 엘리베이터 있는 매물이 많음 주차공간이 있는 매물이 많음 건물 설비가 좋지만 교통이 좋지 않은 군집

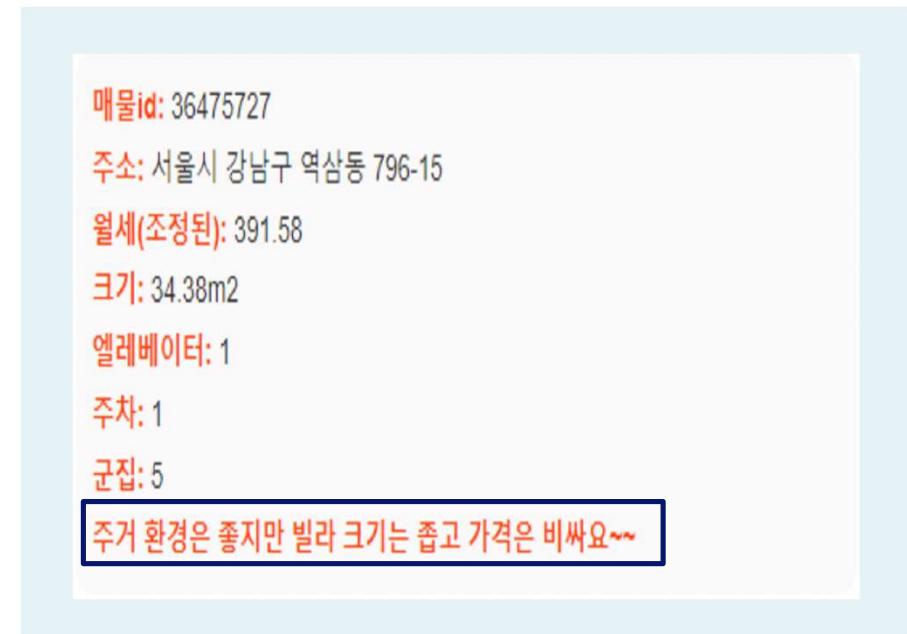
07. 군집 특징 파악



08. 빌라 매물 정보 및 군집 특성 예시

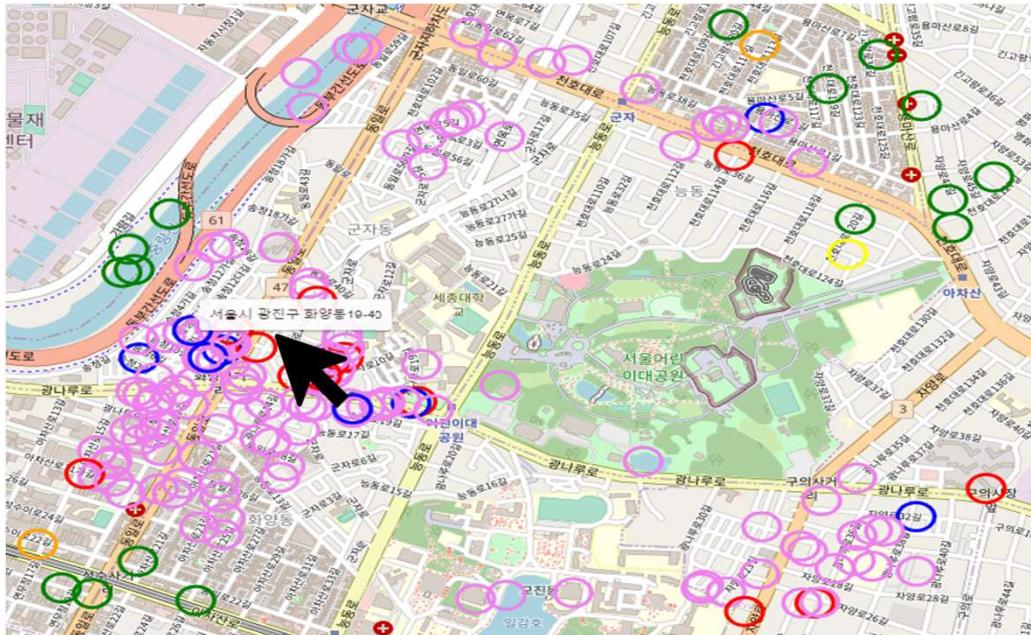


빌라 매물 지도에서 매물 선택하여 클릭

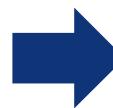


해당 매물의 정보와 군집 특징이 말풍선에 표시됨

08. 원룸 매물 정보 및 군집 특성 예시



원룸 매물 지도에서 매물 선택하여 클릭



매물id: 36362509

주소: 서울시 광진구 화양동 19-40

월세(조정된): 89.17

크기: 29.75m²

엘레베이터: 1

주차: 1

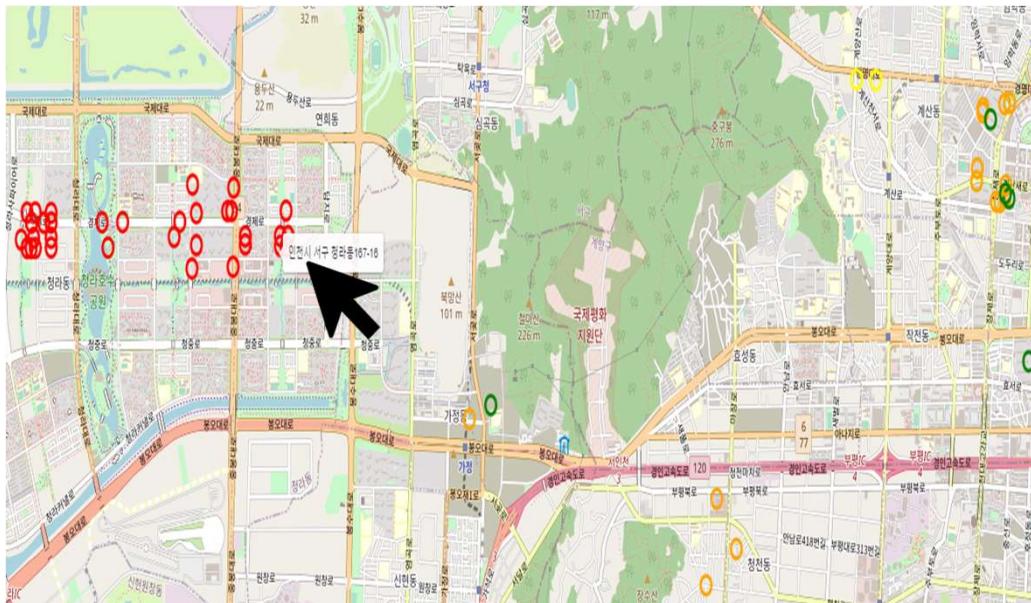
군집: 5

남향: 1

원룸 크기도 크고 햇빛 잘 들어오는 남향집~~~

해당 매물의 정보와 군집 특징이 말풍선에 표시됨

08. 오피스텔 매물 정보 및 군집 특성 예시



오피스텔 매물 지도에서 매물 선택하여 클릭



매물id: 36584528
주소: 인천시 서구 청라동 167-16
월세(조정된): 38.08
크기: 27.64m²
엘레베이터: 1
주차: 1
군집: 1
남향: 0
근처 지하철역 개수: 0
가까운 지하철역 까지: 2144.0 미터
저렴한 오피스텔~~ 대중교통은 안 좋아요

해당 매물의 정보와 군집 특징이 말풍선에 표시됨

03-02. 회귀

매물+지역정보 → 매물가격

14 매물 둘러보기

매물 정보

- 원룸 40m² 4층 서향
- 월세 **45/5500**
- 관리비 15

5만원
비싸요

- 옵션 세탁기, 전자레인지, 책상
- 엘리베이터 유
- 지하철역 영등포역 ⑤ ① 3km
- 주차공간 유

01. 회귀 모델 선정

1-1) 월 주거비용 추정을 위한 회귀 모델

후보 모델

선형 회귀모델	Linear Regression
	Polynomial Linear Regression
선형 회귀 규제 모델	Ridge
	Lasso
	ElasticNet
트리 기반 모델	RandomForest
	XGB
	LGBM
	Gradient Boost

목표 변수

2023 전월세 전환율			
기준금리 3.5%	주택	민특법	상가
전세 → 월세	5.5%	5.5%	12%
월세 → 전세	합의	5.5%	합의

- 목표 변수 산출식

- 사용 변수: deposit, rent, manage cost
최근 1년간 3.75~5.5% 사이에서 변동→ 5%로 변환

- 부동산 전월세 전환율:
최근 1년간 3.75~5.5% 사이에서 변동→ 5%로 변환

- 월 주거비용
= 보증금(deposit)*0.05/12 + 월세(rent) + 관리비(manage cost)

01. 회귀 모델 선정

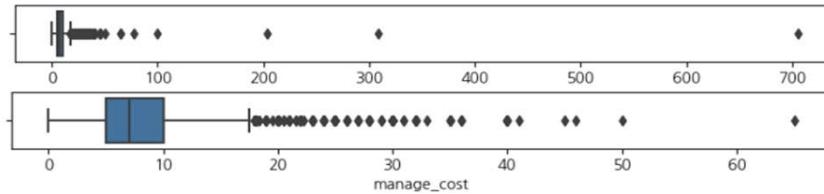
1-2) 데이터 전처리

매물 데이터 : 입력 오류 등으로 인한 이상치 데이터 제거

→ 지역을 고려했을 때 관리비, 월세, 평수 등이 합당하지 않으면 제거

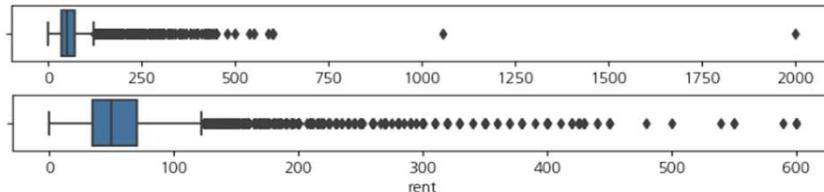
- 관리비 > 100만원

id	address1	service_type	size_m2	sales_type	rent	deposit	manage_cost
36598408	인천시 계양구 계산동	오피스텔	30.48	월세	40	500	705.0
36551565	경기도 의왕시 내순동	원룸	25.88	월세	55	1000	308.0
36250177	경기도 이천시 갈산동	원룸	26.45	월세	55	500	204.0



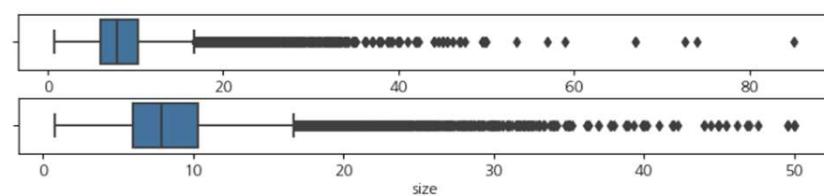
- 월세 > 1000만원 이상

id	address1	service_type	size_m2	sales_type	rent	deposit	manage_cost
36383385	서울시 강서구 화곡동	빌라	48.66	월세	2000	10	10.0
36485147	인천시 미추홀구 용현동	오피스텔	84.17	월세	1056	2000	20.0



- 매물 사이즈 > 150 등

id	address1	service_type	size_m2	sales_type	rent	deposit	manage_cost
36504233	경기도 파주시 탄현면 법흥리	원룸	221.49	월세	27	300	3.0
36517921	경기도 용인시 기흥구 상갈동	원룸	148.76	전세	0	8000	5.0



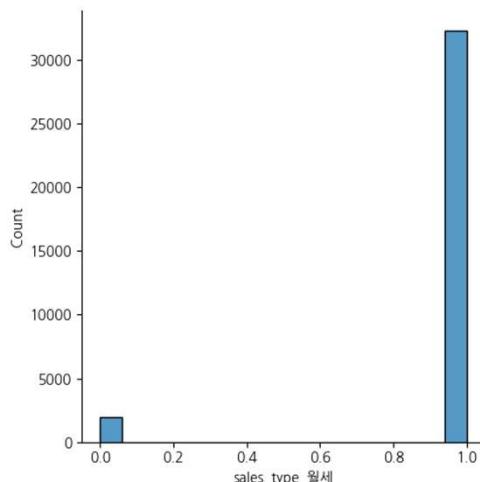
01. 회귀 모델 선정

1-2) 데이터 전처리

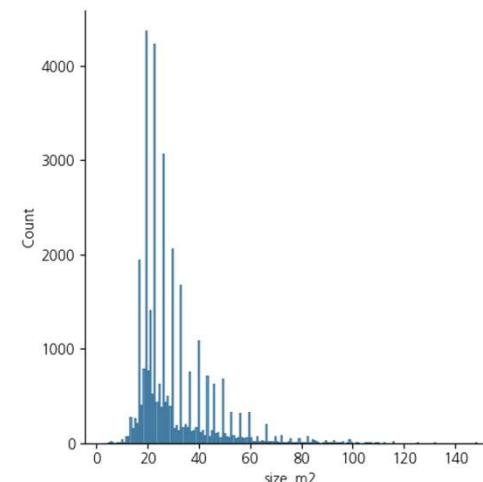
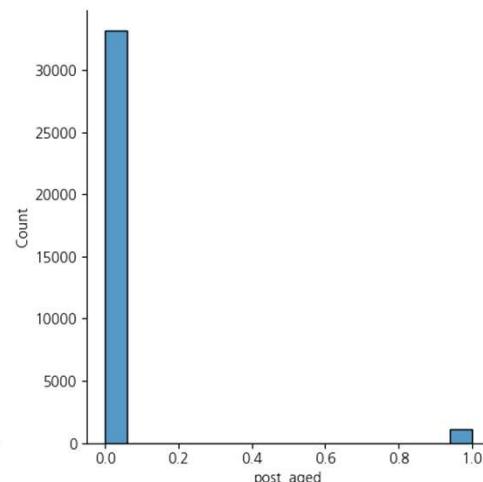
분포 확인 목적: 회귀모델을 실시하기 전 변수들의 분포를 확인 → 향후 적용할 스케일링, 모델 기법 선정방향 설정

분포 확인 결과:

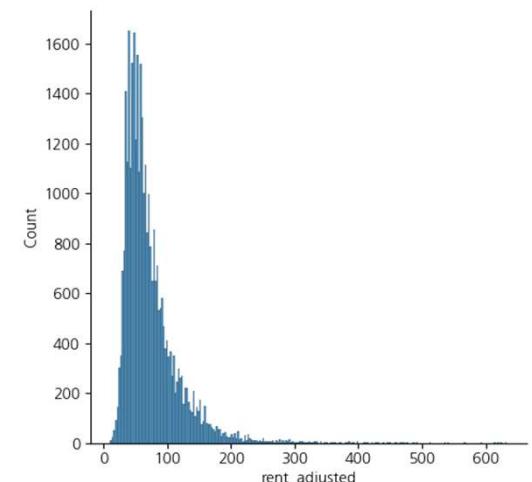
- 범주형 데이터 포함 → one-hot-encoding, 트리기반 모델
- 각자 데이터의 범위가 달라 스케일링 고려
- 분포가 균일하지 않은 경우 불균형한 분포를 조정해주는 로그변환 등을 고려



범주형 변수



왜도와 첨도가 높은 분포

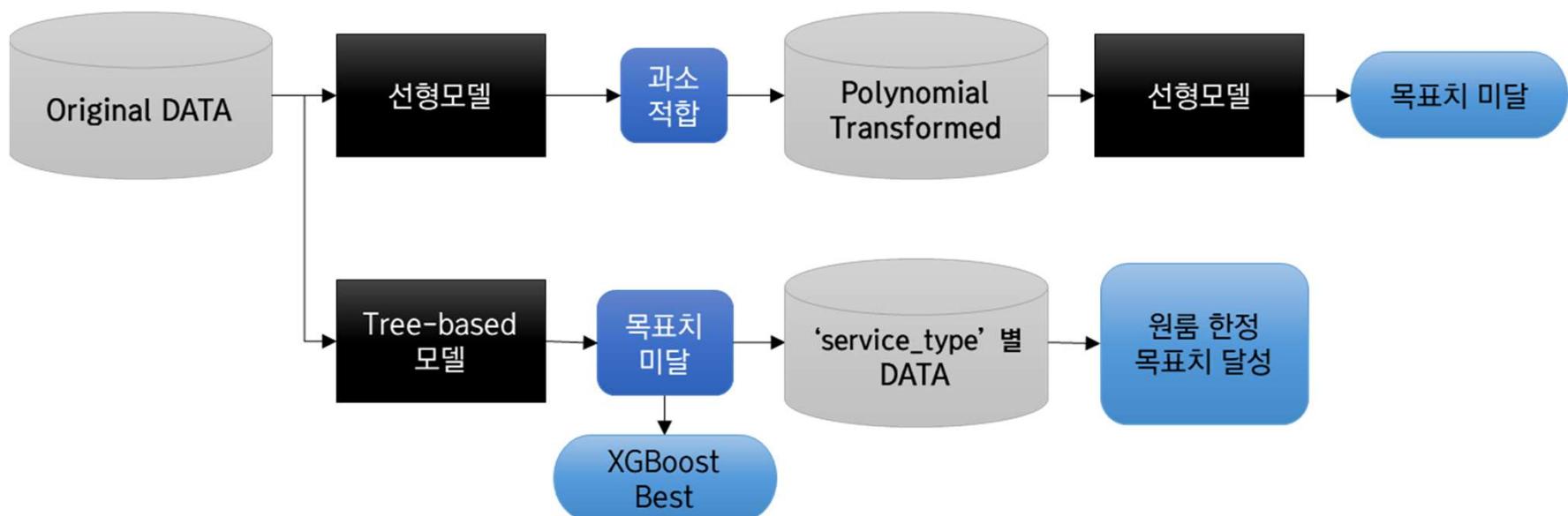


01. 회귀 모델 선정

1-3) 모델링 과정

- 성능 목표: RMSE 10 이내 (추정 가격 오차 10만원 이내)

후보모델	선형 회귀 모델	Tree-based 모델
	Linear Regression Lasso Ridge ElasticNet	Random Forest Gradient Boost XGB LGBM



02. 회귀 모델링 결과

2-1) Linear Regression

- 성능 목표: RMSE 10 이내 (추정 가격 오차 10만원 이내)

선형회귀 모델	Lasso	Ridge	Elastic Net
Train RMSE: 25.424 Test RMSE: 24.638	Train RMSE: 27.066 Test RMSE: 26.145	Train RMSE: 25.484 Test RMSE: 24.669	Train RMSE: 26.280 Test RMSE: 26.3247

- 데이터 복잡도 증가 필요 → **다항 변환 실시**
- L1 규제 보다 L2 규제의 성능 개선 효과 → 변수선택의 필요성 X, **ElasticNet** 모델에 약한 L1 규제 적용



Best

Train RMSE: 18.114 Test RMSE: 20.311	Train RMSE: 19.725 Test RMSE: 19.800	Train RMSE: 18.182 Test RMSE: 19.766	Train RMSE: 18.802 Test RMSE: 19.497
---	---	---	---

여전히 목표 성능에 미치지 못함
선형 모델보다 복잡한 모델 사용의 필요성

02. 회귀 모델링 결과

2-2) Tree model

◆ 트리모델의 선택이유

- 1) 불연속성이 높은 데이터에 강함 : 데이터 중 불연속성이 높은 피쳐 존재
- 2) 가격 결정 요인의 설명할 때 모델의 해석력이 좋음

	Random Forest	Gradient Boost	XGB	LightGBM
원본 데이터 Baseline model 결과	<ul style="list-style-type: none">• Train RMSE: 6.227• Test RMSE: 16.250	<ul style="list-style-type: none">• Train RMSE: 19.425• TEST RMSE: 20.678	<ul style="list-style-type: none">• Train RMSE: 10.489• Test RMSE: 15.808	<ul style="list-style-type: none">• Train RMSE: 14.376• Test RMSE: 17.599
하이파라미터 조정 (Grid Search, Bayesian ops)	Bayesian optimization 사용하여 아래의 범위에서 파라미터 탐색 n_estimators: (100~5000) max_features: (30~69) max_depth: (5~40) min_samples_leaf:(1~50) min_samples_split': (2~50)	learning_rate: (0.01, 0.1), n_estimators : (100, 1000), subsample: (0.5,1)}	n_estimators: [200, 500, 1000] max_depth:[2, 3, 6, 8, 10], colsample_bytree: [0.5, 0.75, 1] subsample:[0.5, 0.75, 1]	learning_rate: (0.01, 0.15), num_leaves: (10, 50), max_depth: (3, 15), n_estimators : (100, 1000), min_child_samples: (20,40)}
최종 선택 파라미터	min_samples_leaf=1 min_samples_split= 5 n_estimators= 621	1 'learning_rate': 0.09, 2 'n_estimators': 954, 3 'subsample': 0.8	colsample_bytree=0.75, learning_rate=0.1, n_estimators=2000, max_depth=6 subsample= 0.75, gamma = 10	learning_rate: 0.1, max_depth: 15, min_child_samples: 40, n_estimators: 448, num_leaves: 23
최종 결과 (best_params_)	<ul style="list-style-type: none">• Train RMSE: 7.2• Test RMSE: 16.9• 파라미터로 개선 어려움	<ul style="list-style-type: none">• Train RMSE: 13.810• Test RMSE: 17.480	<ul style="list-style-type: none">• Train RMSE: 1.726• Test RMSE: 10.116	<ul style="list-style-type: none">• Train RMSE: 11.490• Test RMSE: 15.698

02. 회귀 모델링 결과

2-3) Best Model - XGB

XGBoost model 개선방향

- 트리모델 중 베스트 모델이었던 XGB로 파라미터 조정해봐도 여전히 오차가 한 자리 수(만원 대) 이내로 오차가 줄지 않음
- 주거형태 별로 특징 상이함 → 주거형태 별로 모델을 돌려 가격 추정을 다르게 해보면 오차가 개선되지 않을까?
- 오차 한 자리수의 벽을 넘고자 주거형태(원룸, 오피스텔, 빌라) 별로 트리 모델 돌려봄
- 주거형태(원룸, 오피스텔, 빌라)에 따라 가격에 영향을 미치는 요소가 다를 것이라고 생각하여 분리하여 회귀분석을 진행

결과

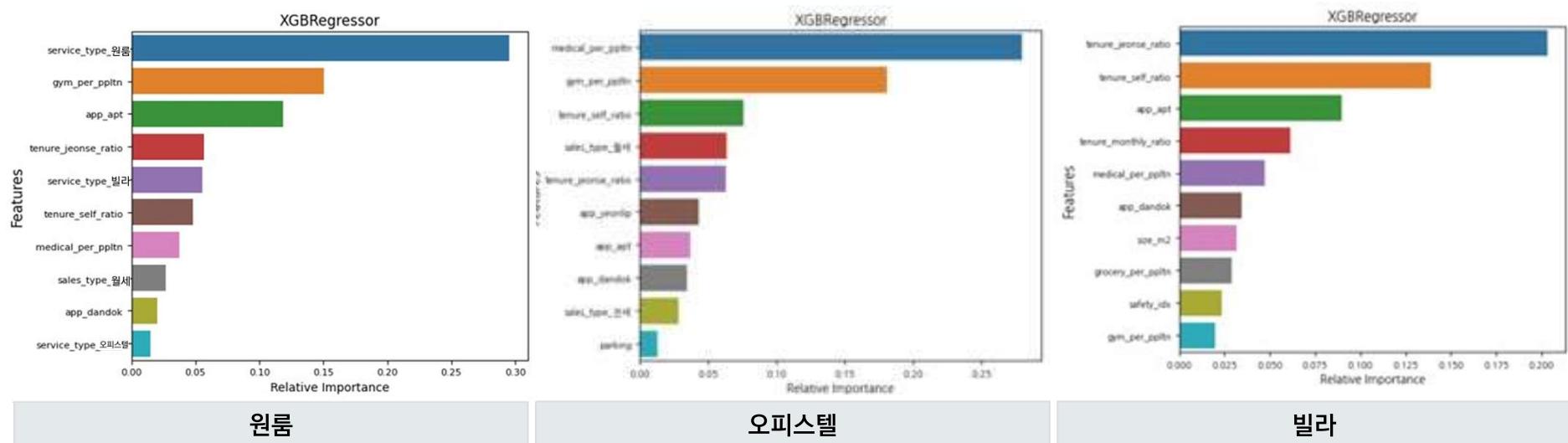
	원룸	오피스텔	빌라
Train RMSE	1.528	5.966	5.55
Test RMSE	9.935	18.680	22.208

- 주거형태 중 원룸의 오차는 한자리 수로 개선
- 그.러.나... 오피스텔, 빌라에서는 데이터 샘플이 적었기 때문에 모델의 결과가 개선되지 않았음(오피스텔 8,259개 / 빌라 6,936개(원룸 19,013개))
- 오피스텔 및 빌라에 대한 매물 데이터 추가 확보 필요
- 전체 데이터셋에 대한 회귀모델의 정확도가 개선되지 않은 이유가 오피스텔, 빌라 매물데이터에 대한 예측 성능이 낮기 때문이라고 추정 가능

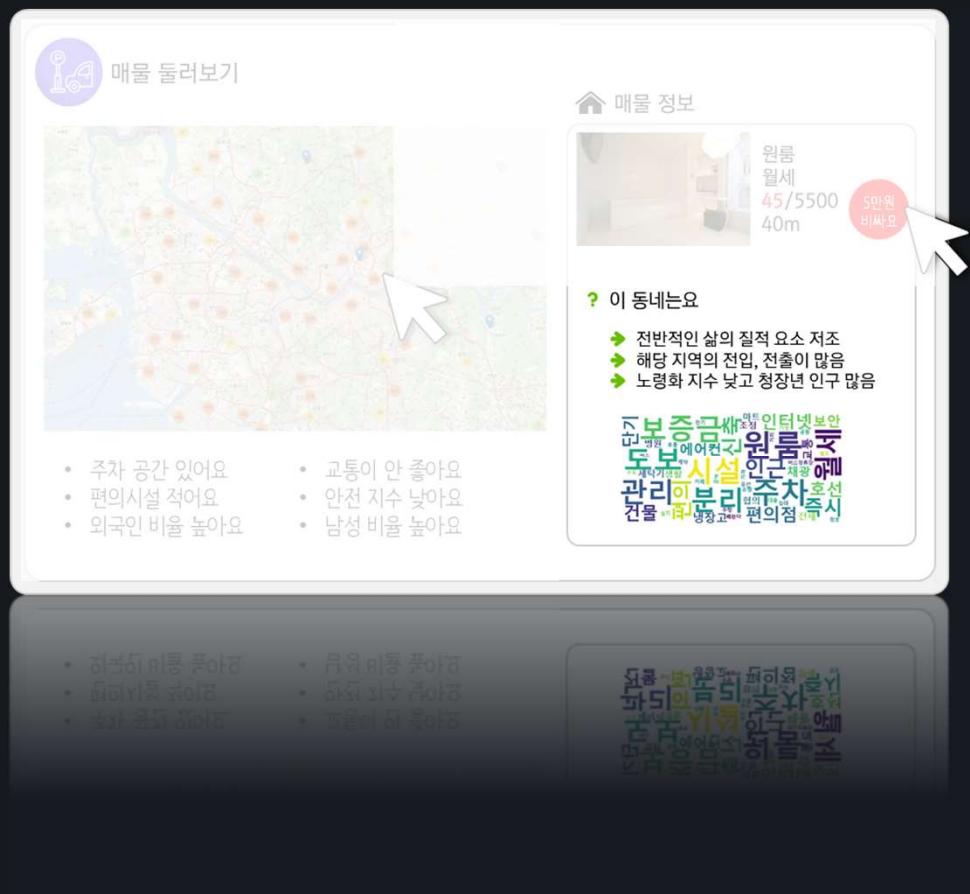
02. 회귀 모델링 결과

2-3) Best Model - XGB

Feature Importance



03-03. 텍스트

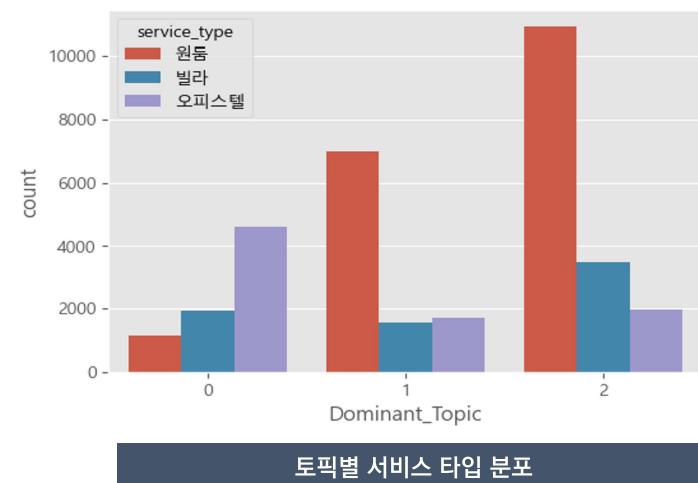
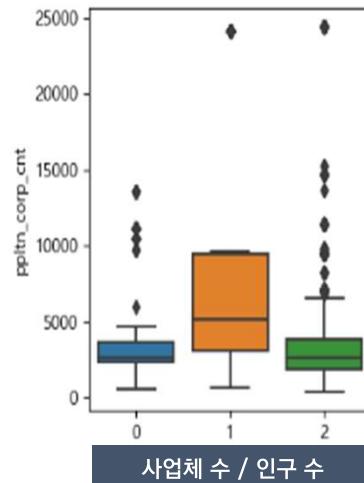
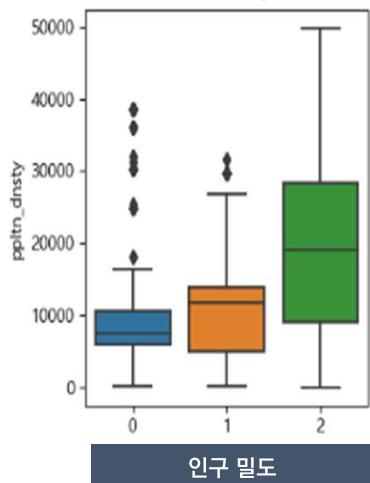
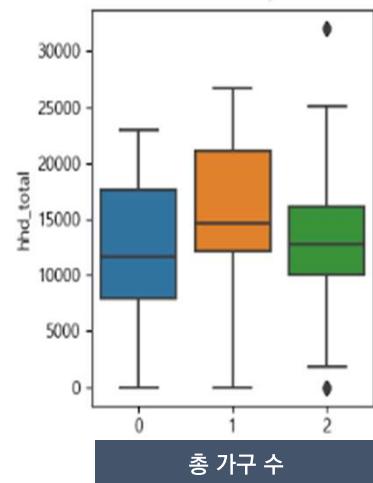


03. 텍스트_LDA 토픽분석

03. 텍스트_LDA 토픽분석

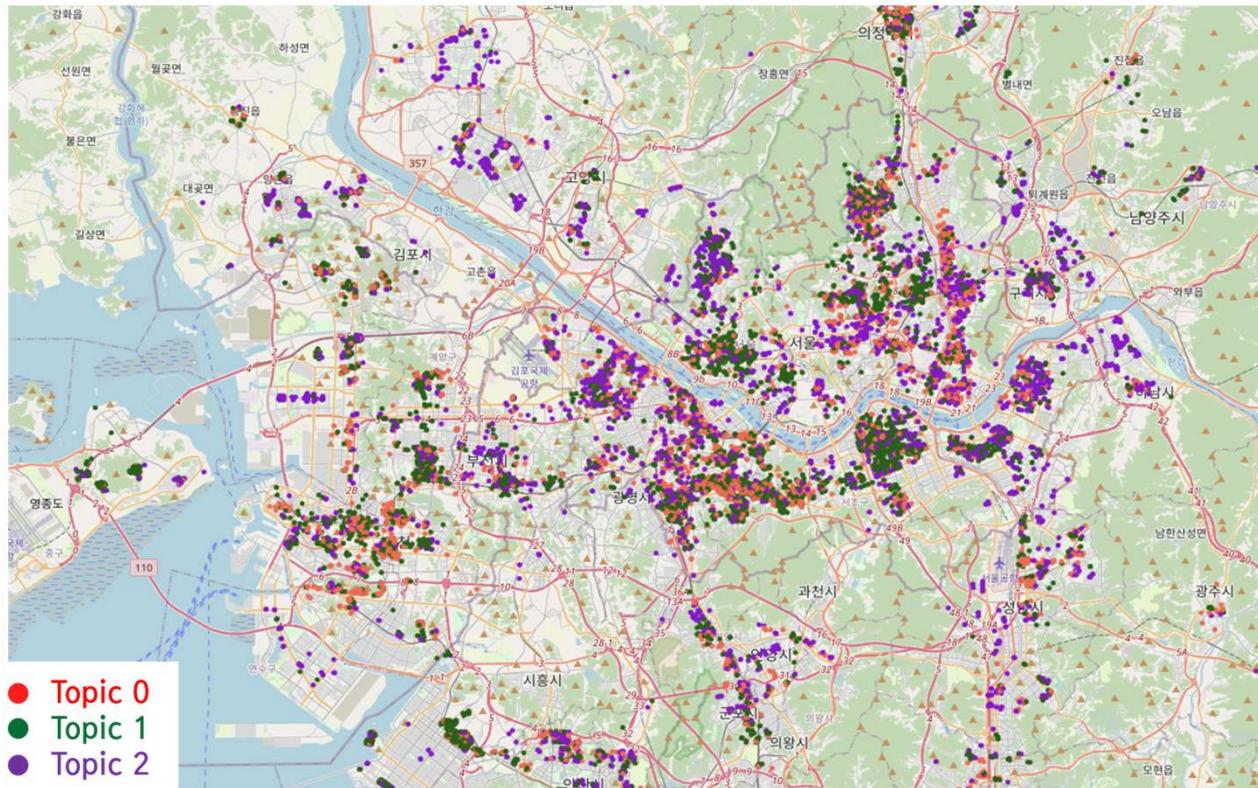
토픽 별 non-text feature 분석

- boxplot 통해 topic별 특징 비교(매물 및 지역특징)
- 각 토픽에 포함될 확률이 50% 이상 매물들로 한정하여 확인



03. 텍스트_LDA 토픽분석

Folium 지도 시각화



Topic 0

주요단어: '보증금, 원룸, 월세, 관리'
원룸, 자취생 → 관악구

Topic 1

주요단어: 신축, 시설, 편의 채광
직장인 → 강남구 집중분포

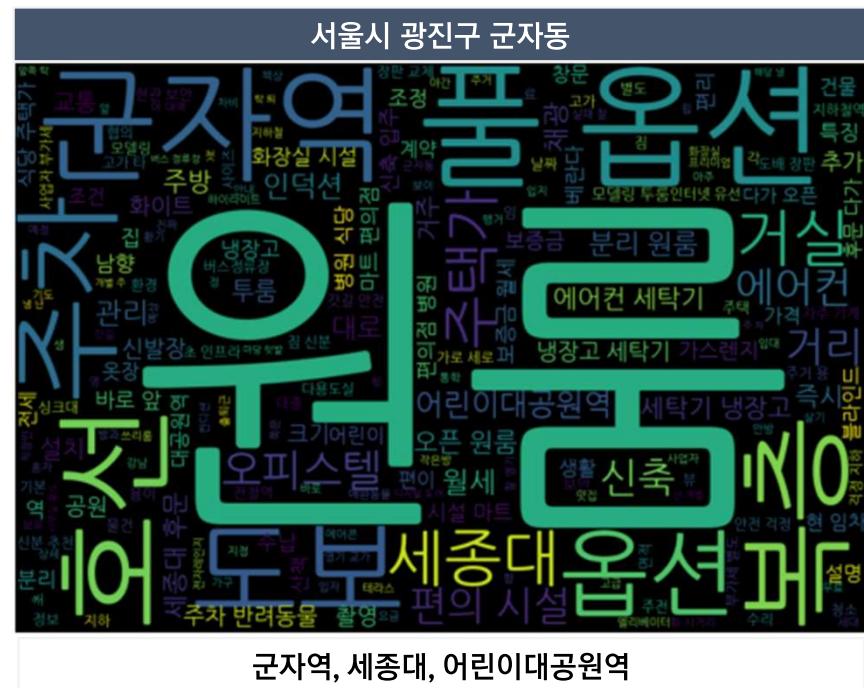
Topic 2

주요단어: 오피스텔, 주차
가족단위 → 하남시, 외곽

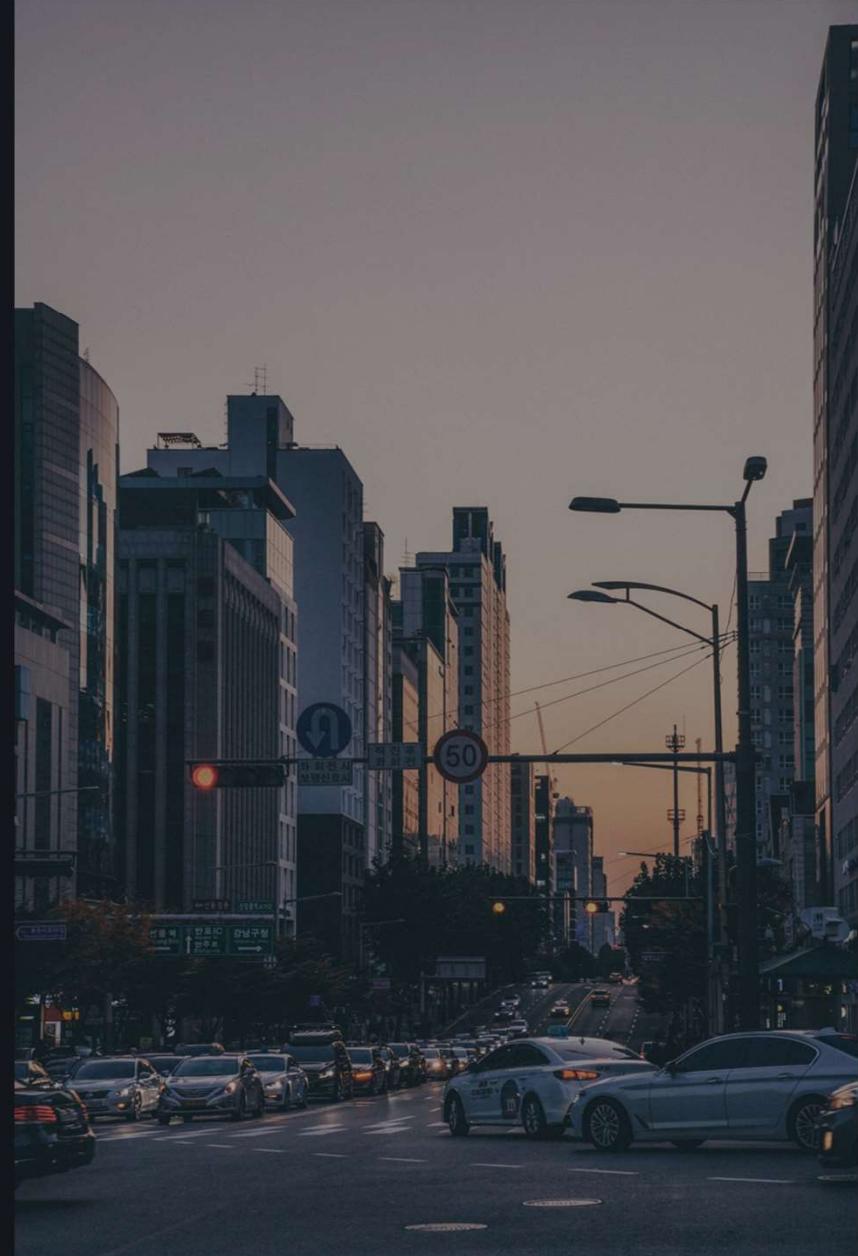
03. 텍스트 WordCloud

시군구, 읍면동 단위 매물 워드 클라우드

매물이 속한 지역에 대해 파악할 수 있도록
시군구, 읍면동별 매물 키워드 집계



4. 결론



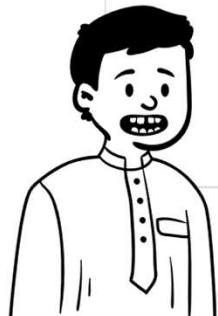
04. 결론

서비스 시나리오

취업을 하게 되어 상경한 A씨

인터넷 상 정보는 많지만

어느 동네가 생활 패턴에 맞을지,
관심 가는 매물의 가격이 맞는 가격인지,
그래서 어디서 자취를 할지
모르는 상황!



내 동네 선택하기



원룸
자차 있어요
안전 교통 노상관
가성비 좋아



원룸
사회초년생
역세권
저층 저가 저평수



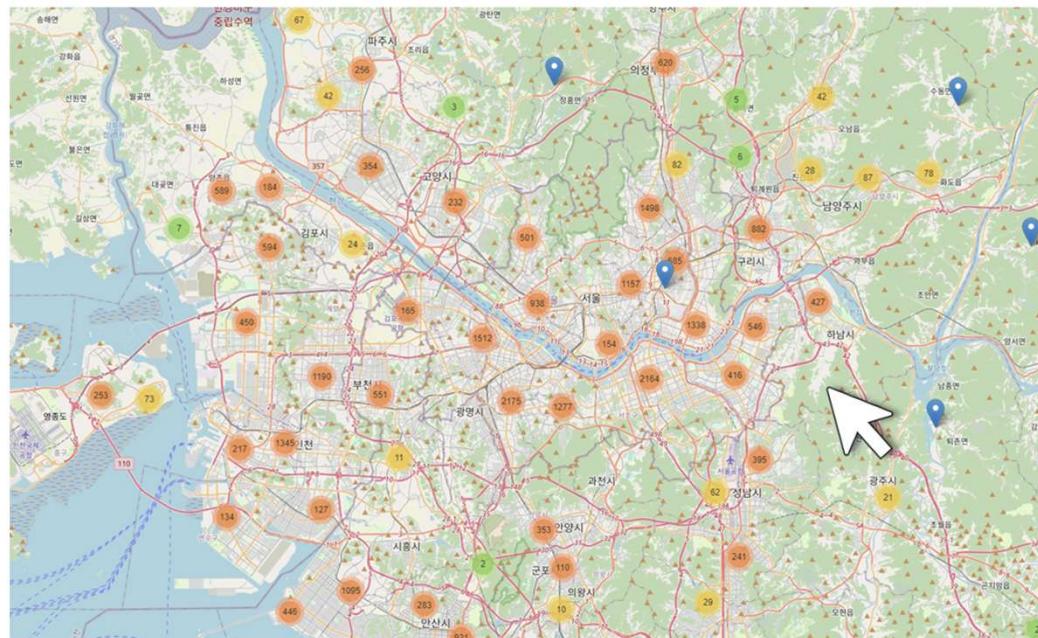
빌라
강남권
작지만 프리미엄
삶의 질

04. 결론

서비스 시나리오



매물 둘러보기



- 주차 공간 있어요
- 편의시설 적어요
- 외국인 비율 높아요

- 교통이 안 좋아요
- 안전 지수 낮아요
- 남성 비율 높아요

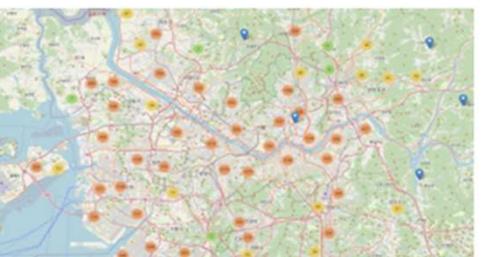
04. 결론

서비스 시나리오





매물 둘러보기



? 이동네는요

- ▶ 전반적인 삶의 질적 요소 저조
- ▶ 해당 지역의 전입, 전출이 많음
- ▶ 노령화 지수 낮고 청장년 인구 많음



매물 정보



- 원룸 40m² 4층 서향
- 월세 45/5500
- 관리비 15

5만원
비싸요


• 옵션	세탁기, 전자레인지, 책상
• 엘리베이터	유
• 지하철역	영등포역 ⑤ ① 3km
• 주차공간	유

04. 결론

서비스 시나리오





매물 둘러보기



〈 가격 평가 상세 분석 〉

가격 분석에 중요하게 작용한 요인들은 아래와 같습니다!



요인	작용 범위
service_type_new	전반적인 서비스 유형
gmt_per_ppltn	인구 밀도
emd_cd_2022	행정 구역
app_apt	아파트 타입
size_m2	면적

? 이 동네는요

- ▶ 전반적인 서비스 유형
- ▶ 해당 지역의 인구 밀도
- ▶ 노령화 지수

본인의 취향에 맞게 이런 요인들을 고려해서 매물을 탐색하시기를 추천합니다.

- 주차공간 유



05. XGB_실제 원룸매물 예측 결과

매물 id	실제가격	예측가격	면적(m2)	관리비포함 옵션수	주차가능	지하철역 수	옵션 수	슈퍼마켓 거리	원룸여부	월세여부
36447553	42.250	42.017	26.45	3	1	2	7	1208	1	1
36528800	36.417	38.232	16.53	3	0	3	6	1898	1	1
36401555	47.083	48.854	26.45	3	1	1	7	1167	1	1
36544669	36.250	42.476	16.53	3	1	3	9	958	1	1
36453311	34.250	35.612	19.83	1	1	0	6	864	1	1
36516168	37.417	32.666	26.45	1	0	2	5	674	1	1
36592305	63.083	65.722	17.49	0	1	2	8	734	1	1
36378653	59.167	68.549	26.45	3	0	2	7	529	1	1
36595845	34.417	34.562	26.45	3	1	2	5	463	1	1
36353593	60.833	50.173	13.22	0	1	1	10	1542	1	1

04. 결론

분석 시사점

프로젝트 진행 방향

1. 목적: 매물 가격에 큰 영향을 주는 요인을 사용자에게 보여주어 청년들의 집 구하기에 도움이 되고자 함

2. 분석방법

1) 회귀

- 사용자가 실제 매물의 정보들을 입력
- 우리가 추측한 매물 가격 vs 실제 가격 비교
- 가격 적정성 평가

2) 군집

- 자신이 어떤 주거환경을 선호하는지, 주택의 특징이 어떠한지 잘 모르는 사용자에게 각기 다른 특징을 가진 ‘주택 군집’ 추천
- 사용자가 원하는 주택 군집 선택 시, 해당 군집에 속한 매물 조회(사용자 취향 반영)

서비스 구현 시나리오

구현 방안

- Landing page = 폴리엄을 통해 구현한 매물 지도 + 매물특징(군집분석 결과 + 토픽별 결과)
- 주거형태(원룸/빌라/오피스텔) 별로 회귀분석 진행한 후 각 주거형태의 중요 변수(feature importance)를 사용자에게 보여줌
- 특정 주거형태에서 특정한 요인의 중요도가 높다는 것 = 해당 요인에 의해 매물 가격이 크게 좌우된다는 뜻

04. 결론

분석 시사점

한계점

- XGB 모델의 정확도가 떨어졌던 이유 = 데이터 양의 부족 (회귀분석, 군집분석)
- 매물 개별로 세분화된 데이터가 많으면 많을수록 좋다는 것을 깨달음
- 오피스텔과 빌라의 경우 데이터의 수가 원룸에 비해 충분하지 않아 예측 정확도가 원룸에 비해 떨어짐

군집 분석

- 이상치라고 판단되는 매물에 대해 사용자에게 “이 매물의 정보를 신뢰할 수 없습니다.” 와 같은 경고 메세지를 보여주는 서비스를 제공 하고자 함
- 매물 데이터의 이상치에 대해 주관적으로 해석하기 어려워 이상치에 대한 처리를 하지 않음

추후 발전 방향

[분석 측면]

- 오피스텔 및 빌라 매물 데이터 추가 → 오피스텔과 빌라 매물에 대한 분석 정확도 향상 도모
- 유의미한 변수 추가 확보 통한 분석 및 모델링 성능 향상 가능성
- 군집 분석에서 이상치에 대한 처리에 집중하여 스케일링 기법, 모델 성능 향상 도모

[서비스 측면]

- 웹 구현 및 서비스화
- 웹 상에 선호하는 지역조건을 범위로 조정하면 해당되는 매물을 지도 위에 보여주는 기능

End of Documents