



# LH X COMPAS 공공주택 공유차량 적정규모 산정

: 하남 교산 신도시 지구를 중심으로

LH내거



# Contents

1 프로젝트 소개 및 목적

2 분석 파이프라인

3 활용 데이터

4 EDA

5 데이터 전처리 및 모델링

6 클러스터링

7 적정 규모 산정

8 결론



# 1 프로젝트 소개 및 목적

# 공유차량 이용 기대



**공유차량 이용자 수와 공유차량 수 증가에 따라  
공공주택단지 내 입주민들의 공유차량 이용량 증가 예상**

하남 공유차량 수요



# 화성시 공공주택 단지 내 공유차량 이용현황 수요 분석을 통한 하남 교산 지구의 공공주택 단지의 공유차량 적정규모 산정

# 1 프로젝트 소개 및 목적

교산 신도시

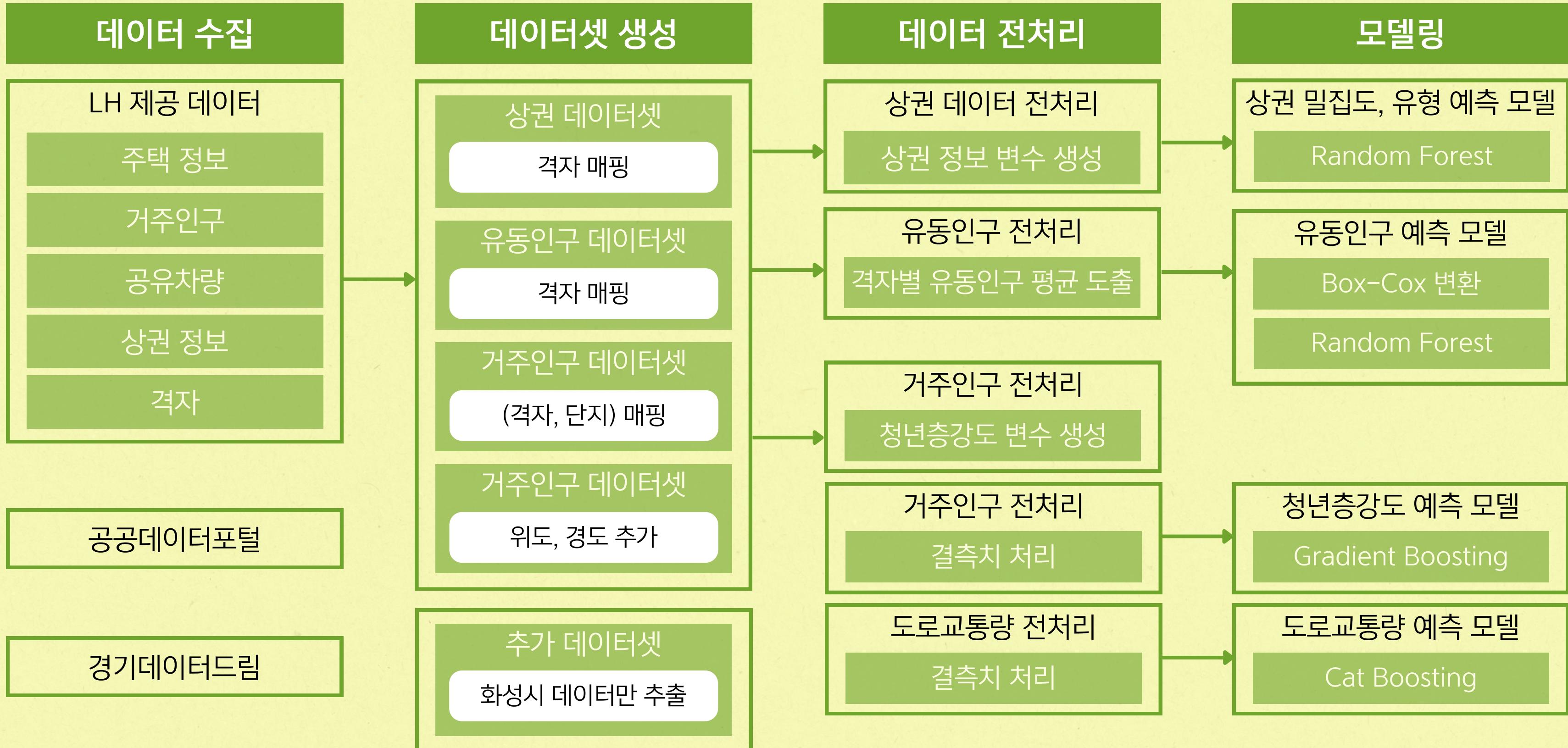
하남시 천현동, 교산동, 춘궁동, 상사창동, 하사창동 일원에 지어지는 신도시



## 계획 규모

- **인구 : 87,258인 (가구당 인구수는 2.4/호로 적용)**
  - 2040 하남도시 기본계획에서 설정한 2030년 기준
  - 주택건설 : 36,697호 (공동주택은 28,383호)
- **지하철 : 송파하남선 신설(3호선 연장)**
  - 3호선의 오금역과 5호선의 하남시청역을 연결
  - 총 3개역 신설 (교산 지구 내에서는 2개 신설 예정)
- **도로 교통 : 주요 상행 도로**
  - 교산지구를 횡단하는 고속도로, BRT 도로 등 신설

# 2 분석 파이프라인 1



# 2 분석 파이프라인 2



# 3 활용 데이터

	데이터명	컬럼 구성	출처
인구	화성시_거주인구.csv	격자ID, 성연령별 인구수	제공
	화성시_성연령별_유동인구.csv	기준년도, 성연령별 유동인구 수, 위/경도	제공
	화성시_시간대별_유동인구.csv	기준년도, 시간대별 유동인구 수, 위/경도	제공
	화성시_요일별_유동인구.csv	기준년도, 요일별 유동인구 수, 위/경도	제공
상권	화성시_상권정보.csv	상권분류, 위/경도	제공
	화성시_추정교통량.csv	도로 링크 ID, 도로 등급, 추정 교통량	제공
	화성시_상세도로망.geojson	도로 링크 ID, 도로 이름	제공
화성	화성시_버스정류장.csv	정류장명, 위/경도	제공
	화성시_지하철역.csv	역사명, 위/경도	제공
교통	화성시_공영주차장.csv	주차장명, 위/경도	제공
	화성시_공유차량_이용현황.csv	월별 이용건수, 월별 이용시간, 배치차종, 배치대수	제공
공유차량	공공주택임대_정보(화성시).csv	단지 코드, 지원 유형, 세대수, 위/경도	제공
	공공주택임대_임대료(화성시).csv	단지코드, 월/세금, 위/경도	제공
공공주택	화성시_격자.geojson	100MX100M 격자 ID	제공
격자			

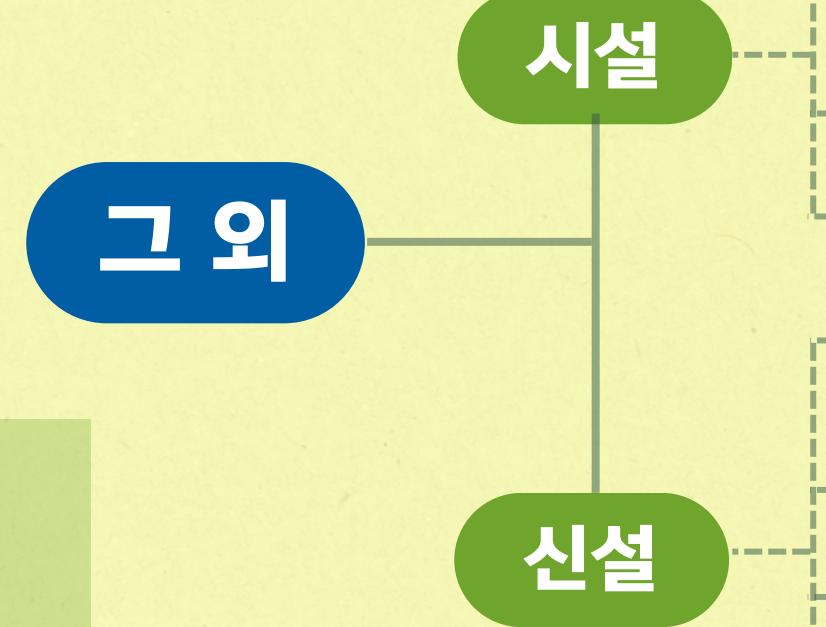
# 3 활용 데이터

데이터명	컬럼 구성	출처
하남시_거주인구.csv	격자ID, 성연령별 인구수	제공
하남시_성연령별_유동인구.csv	기준년도, 성연령별 유동인구 수, 위/경도	제공
하남시_시간대별_유동인구.csv	기준년도, 시간대별 유동인구 수, 위/경도	제공
하남시_요일별_유동인구.csv	기준년도, 요일별 유동인구 수, 위/경도	제공
하남시_상권정보.csv	상권분류, 위/경도	제공
하남시_추정교통량.csv	도로 링크 ID, 도로 등급, 추정 교통량	제공
하남시_상세도로망.geojson	도로 링크 ID, 도로 이름	제공
하남시_버스정류장.csv	정류장명, 위/경도	제공
하남시_지하철역.csv	역사명, 위/경도	제공
하남시_공영주차장.csv	주차장명, 위/경도	제공
하남시_공유차량_이용현황.csv	월별 이용건수, 월별 이용시간, 배치차종, 배치대수	제공
공공주택임대_정보(하남시).csv	단지 코드, 지원 유형, 세대수, 위/경도	제공
공공주택임대_임대료(하남시).csv	단지코드, 월/세금, 위/경도	제공
하남시_격자.geojson	100MX100M 격자 ID	제공

# 3 활용 데이터



데이터명	컬럼 구성	출처
대상구역도(하남교산).geojson	위치정보, 블록유형	공공데이터포털



데이터명	컬럼 구성	출처
전국초중등학교위치표준데이터.csv	학교명, 위/경도, 학교급구분	공공데이터포털
전국휴게소정보표표준데이터.csv	휴게소명, 위/경도	공공데이터포털
경기도 공장등록 현황.csv	공장명, 위/경도	경기데이터드림
교산 주택공급 계획.csv	공급 유형, 세대수, 위/경도	하남교산지구 단위계획결정조정서
교산 신설 학교.csv	학교명, 위/경도	자체 생성
교산 신설 지하철.csv	위/경도	자체 생성
교산 신설 주요도로.csv	도로 위치정보	자체 생성

# 4 EDA : 공유차량 데이터 관찰

## 사용 데이터



### 공유차량

화성시\_공유차량\_이용현황.csv  
공공주택임대\_정보(화성시).csv  
공공주택임대\_임대료(화성시).csv



## 관찰

- 블록에 따라 공유차량 1~4개 배치
- 동탄 신도시에 주로 배치
- car\_cnt에 NaN 14곳 존재 (총 30곳 중)
  - 서비스 종료된 블록



## 인사이트&활용



- 신도시 지역에 공유차량 수요가 큼
  - 교산 신도시와의 유사성 고려
- 행복주택이 위치한 곳에 우선적으로 배치되는 경향
- 신도시 외곽이 폐지되는 경향이 큼



- 통합공공임대주택 여부 고려
- 신도시 → 인구, 상권, 교통 등 고려

# 4 EDA : 거주인구&상권 관찰

## 사용 데이터

### 거주 인구



화성시\_거주인구 .csv

화성시\_공유차량\_이용현황.csv

df\_2022\_lh거주인구공유차량\_merge .csv

## 관찰



## 인사이트&활용

- 20~30대 **거주인구와 공유차량 간 상관관계**
  - 이용시간, 이용건수에 0.3 이상의 양의 상관관계
- **청년인구비율이 공유차량 수요예측의 주요 변수**

- 청년층 비율 파생변수 생성
- 청년층의 강도를 통해 수요 예측

### 상권

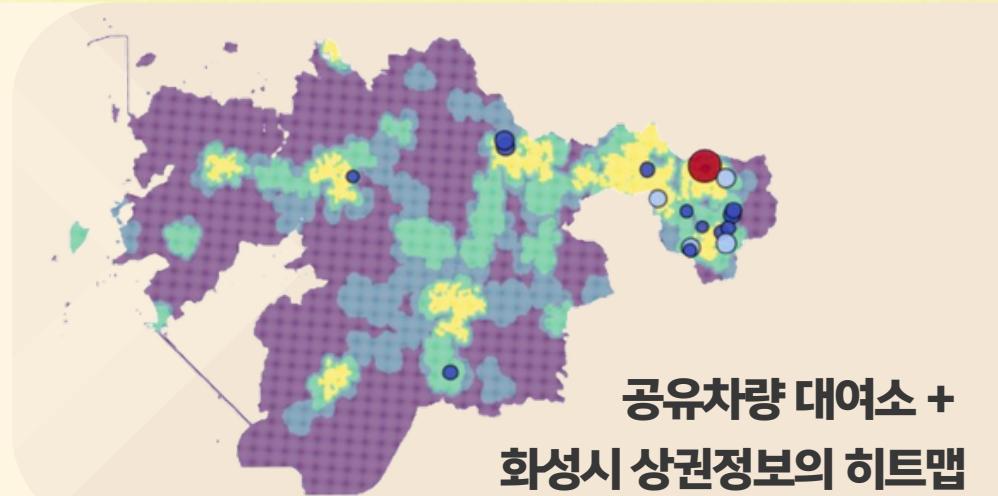


화성시\_상권정보\_2022.csv

(화성시\_상권정보 중 2022년의 데이터)

화성시\_격자.geojson

화성시\_거주인구.csv



- 상권 밀집도를 3단계로 구분했을 때 때
  - 2단계 이상인 영역에 **공유차량 대여소 존재**
- 상권-주거 밀집 구역 간에 관계 역시 밀접

- 군집 결과와 같이 상권 강도 추출
- 주거 구역과의 관계에 따라 유형 판별

# 4 EDA : 거주인구&상권 - 공유차량

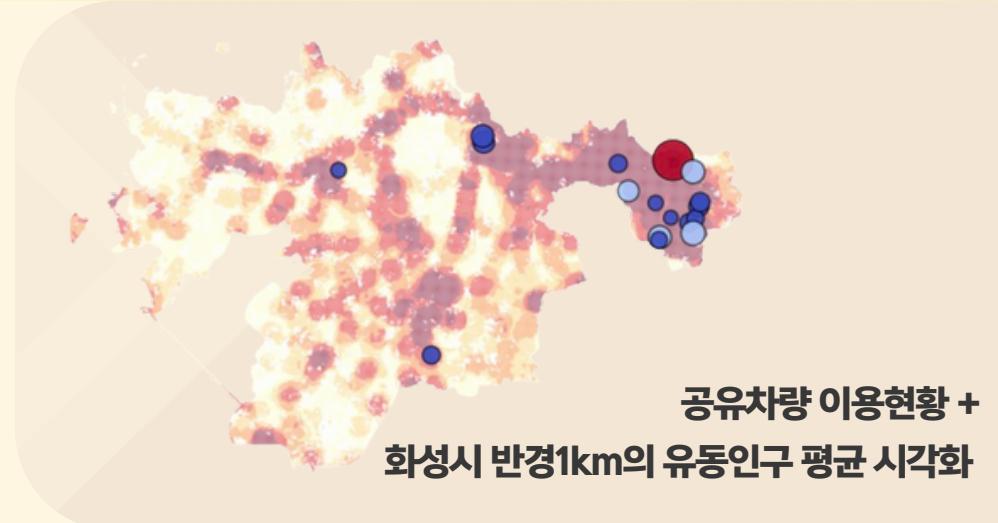


## 사용 데이터

### 유동 인구



1-8.화성시\_성연령별\_유동인구.csv



## 관찰

## 인사이트&활용

- 동탄 신도시와 일부 주거 지역에 밀집
    - 연결 지역따라 이동 경로 파악 가능
  - **공유 차량 대여 블록 역시 이동 경로 사이에 존재**
    - 유동인구 밀집한 동탄에서 공유차량 이용 건수 높음
- ▶ • 블록 기준 반경 1km 유동인구 계산  
• 교통량 예측의 변수로 활용 가능



### 교통량

화성시\_교통추정량\_2022.csv

(화성시 교통추정량 중 2022년의 데이터)

화성시\_상세도로망.geojson

화성시\_공유차량\_이용현황.csv



- 동탄 신도시 지역에 도로 밀집
  - **상행하는 고속도로와 공유차량 대여소 인접**
    - 교통량이 가장 높은 경부고속도로와 화성 동탄2 LH2 인접
- ▶ • 주택 기준 주요 도로와의 접근성 파악  
• 인접한 도로의 교통량 활용

# 4 EDA : QGIS 사각화\_ 화성

## 주요 도로 + 인구

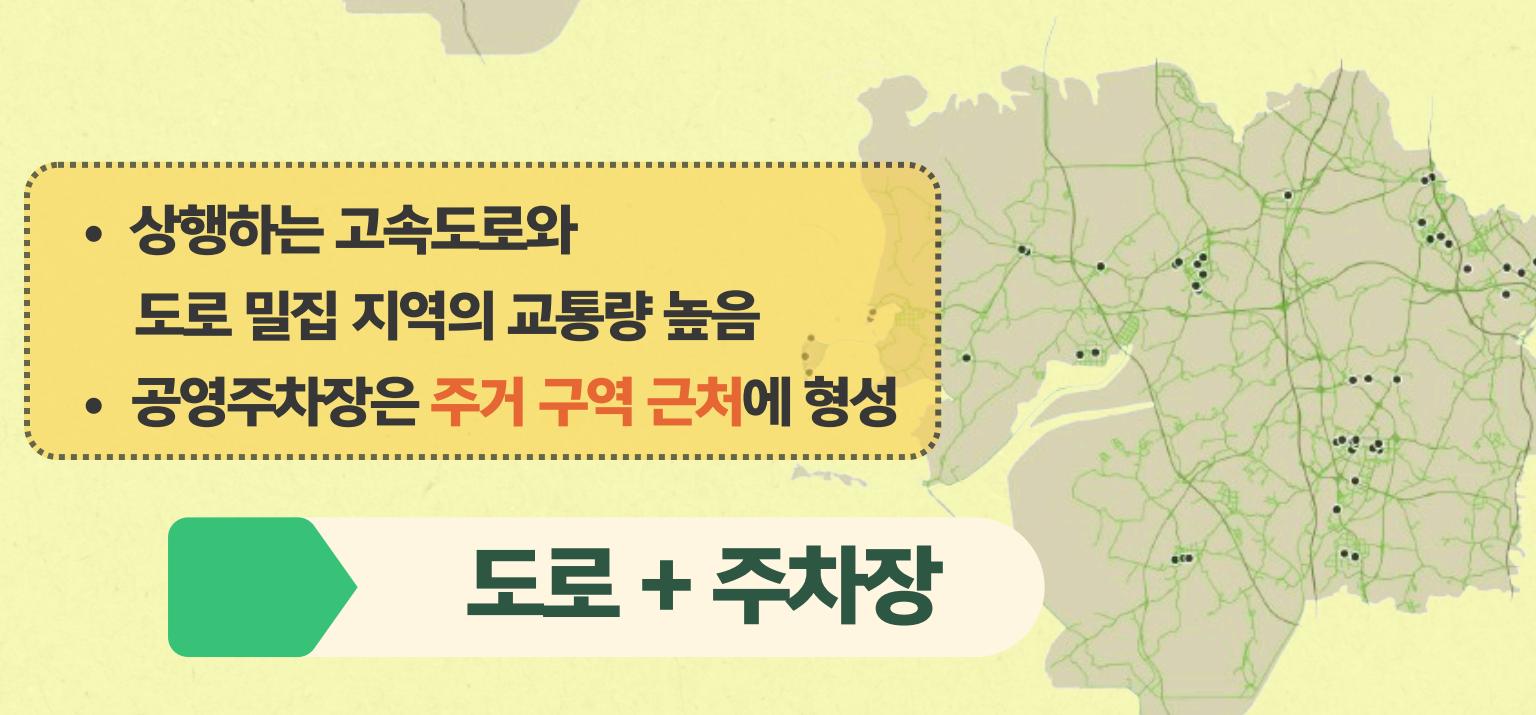


## 상권 3단계 구분



- 상행하는 고속도로와  
도로 밀집 지역의 교통량 높음
- 공영주차장은 주거 구역 근처에 형성

## 도로 + 주차장

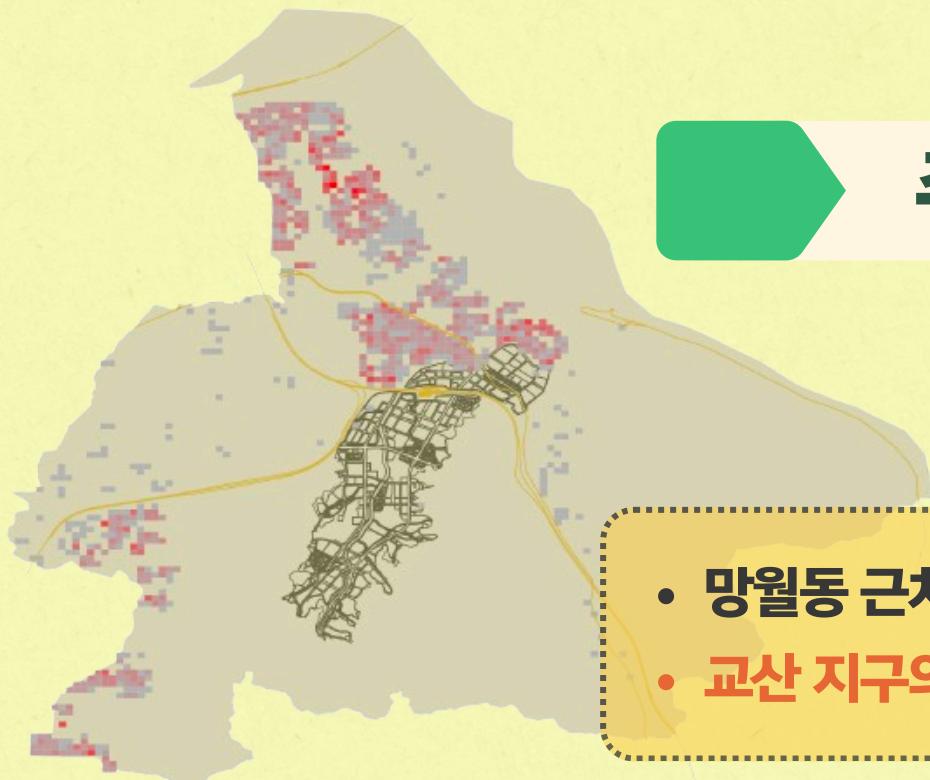


- 주거 구역과 비례하게 정류장 형성
- 지하철은 동탄 제외 골고루 분배

## 버스 + 지하철



# 4 EDA : QGIS 사각화\_ 하남



주요 도로 + 인구



상권 3단계 구분



도로 + 주차장



버스 + 지하철

# 5 데이터 전처리 및 모델링

## 예측의 필요성

교산 신도시는 새롭게 개발되는 지역이므로 존재하는 데이터가 한정적



- 허나, EDA를 통해 상권이나 인구 데이터가 공유차량과 밀접한 관계를 가짐을 확인
- 분석을 통해 상권 정보/유동인구 데이터/청년층 강도/교통량 예측



## 가용한 변수

교산신도시에 대해 가용한 변수는 공동주택 정보 / 주차장 / 지하철 / 버스정류장



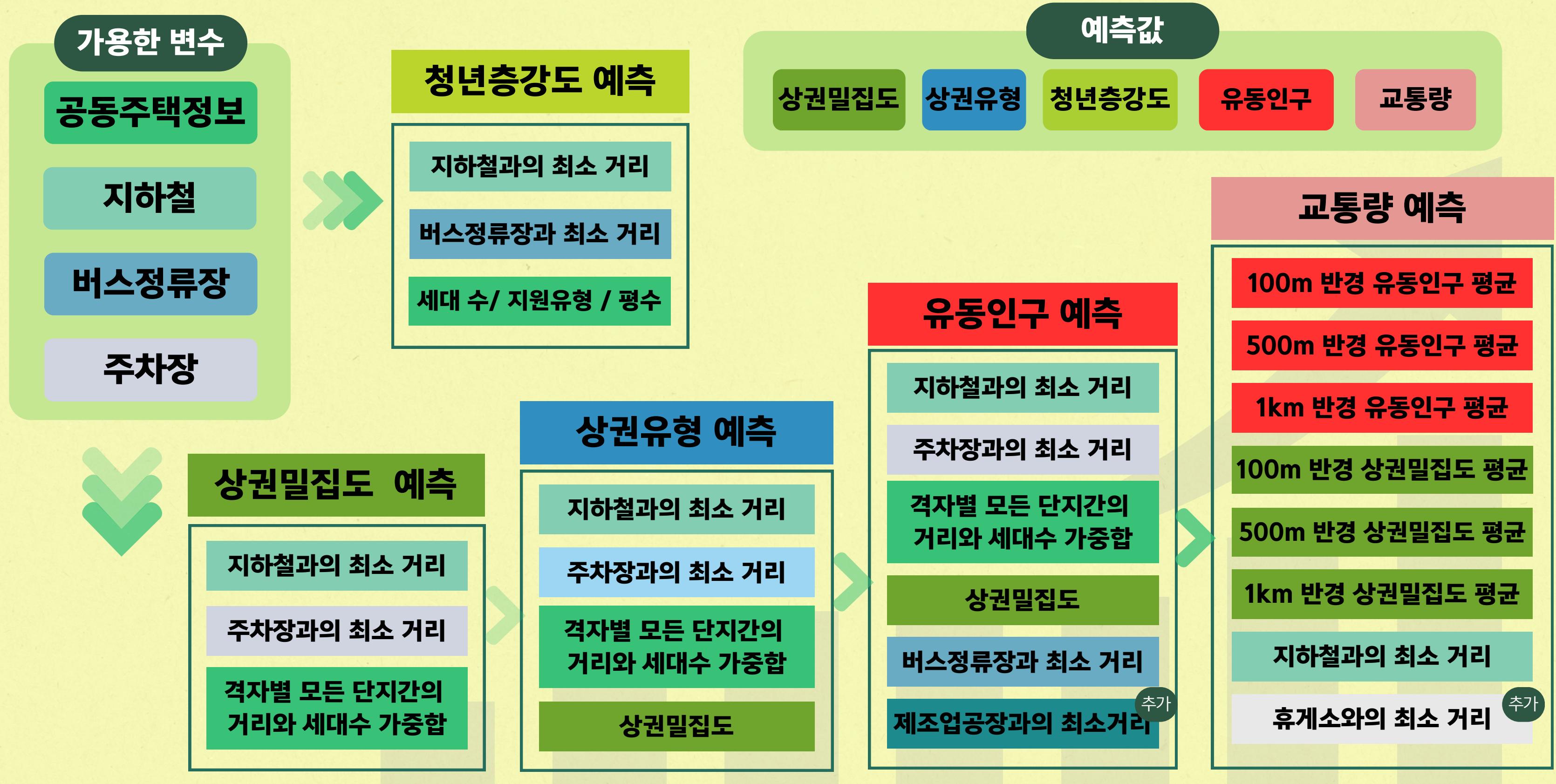
교산 신도시에 대한 지구계획 서류 참고하여 생성



각종 시설의 위치는 지구계획 도면을 참고하여 생성



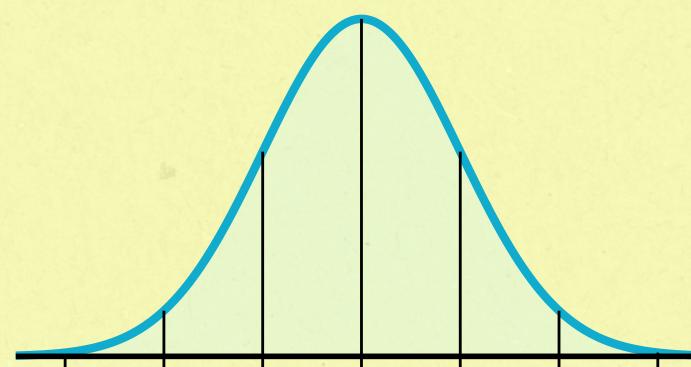
# 5 데이터 전처리 및 모델링



# 5 데이터 전처리 및 모델링

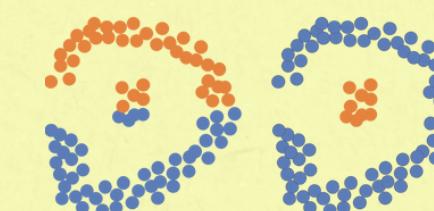
## 전처리 기법

- Standard Scaling
- MinMax Scaling
- Boxcox transform
- Yeojohnson transform
- Factor Analysis



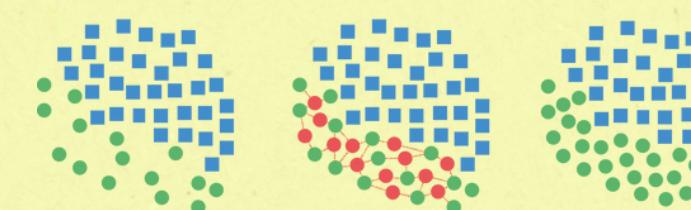
## 군집화 기법

- spectral clustering
- DBSCAN



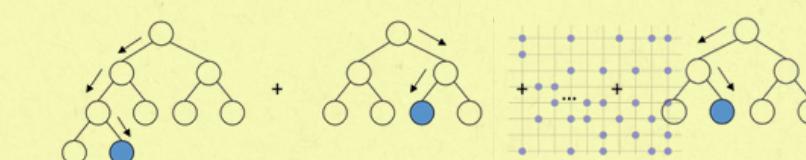
## 증강 기법

- SMOTE



## 모델링 기법

- RandomForest
- Gradient Boosting
- SVM
- GridSearch
- CatBoost



# 5 데이터 전처리 및 모델링 : 상권 분석

## 분석목표

- 화성의 상권 위치 정보로부터 밀집한 상권을 포착하고 강도에 따라 구분
- 주거구역과의 관계에 따라 상권의 유형을 적절하게 분류

## 사용 데이터

화성시\_상권정보.csv / 화성시\_공영주차장.csv  
화성시\_지하철역.csv / 화성시\_공공주택정보.csv

## 상권밀집도 / 상권유형 예측

격자	가장 가까운 지하철역과의 거리	단지와의 거리 및 세대수 가중합	가장 가까운 주차장과의 거리



# 5 데이터 전처리 및 모델링 : 상권 분석

## 데이터 전처리

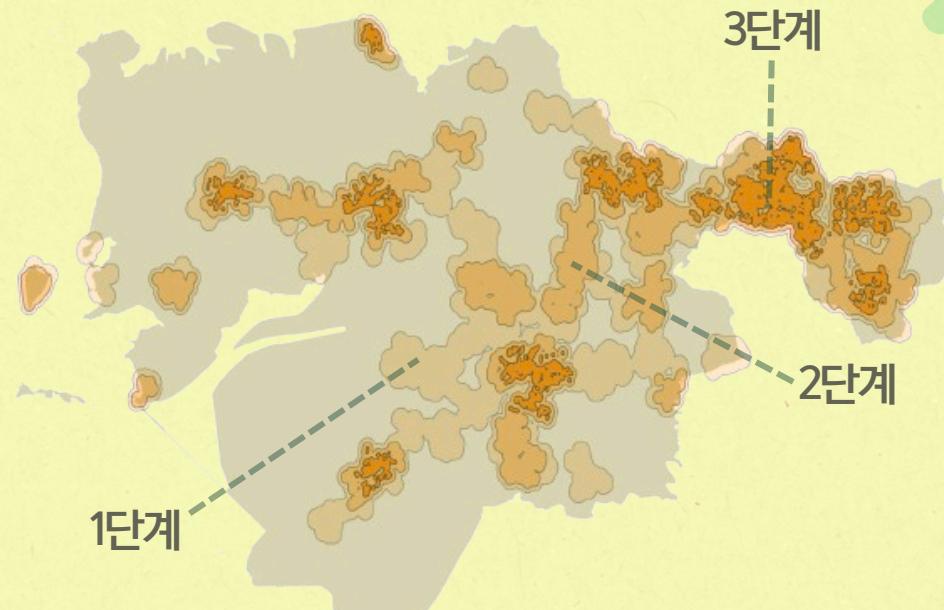
### 1. DBSCAN 군집화

위/경도를 기준으로 상권이 밀집한  
클러스터 생성

1단계 : eps 100m / min\_samples 50

2단계 : eps 100m / min\_samples 100

3단계 : eps 100m / min\_samples 300



### 2. 3단계 군집 강도 매핑

상권 클러스터가 지나는 격자에 지정한  
상권 강도를 새로운 컬럼으로 매핑

3	2		2
3		1	
	2	1	2
1		2	

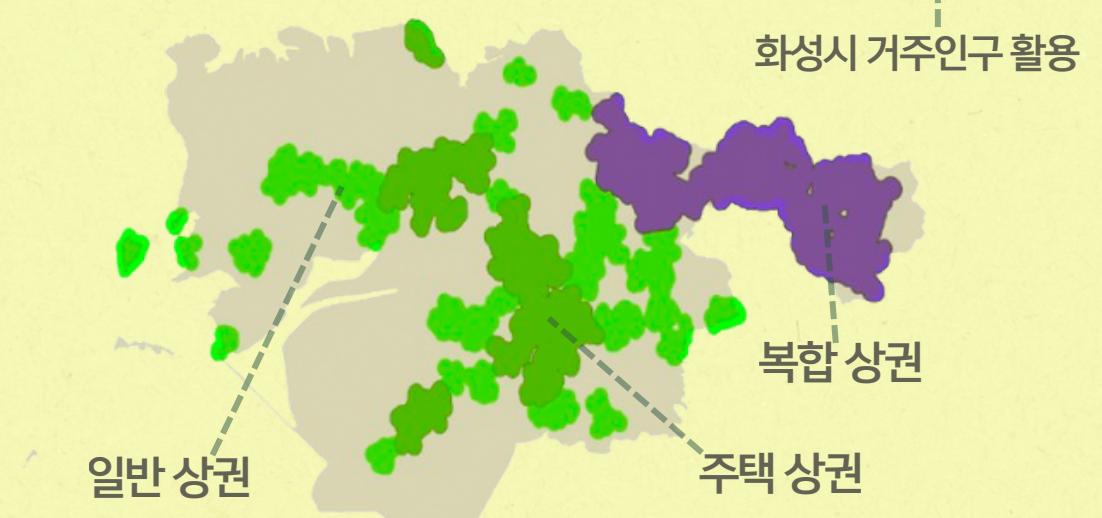
1단계 : 13391개  
2단계 : 9824개  
3단계 : 9401개

### 3. 상권 유형 매핑

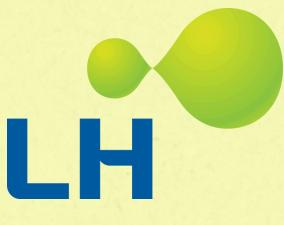
상권은 주변 환경에 따라 번화 상권, 주택 상권, 교외상권,  
대학상권, 역세권 상권, 사무실 상권 등이 존재

여기선 3가지로 분류하여 격자에 컬럼으로 생성

- 일반 상권 : 주택 상권 X + 번화 상권 X
- 주택 상권 : 주택 상권 O + 번화 상권 X
- 복합 상권 : 주택 상권 O + 번화 상권 O



# 5 데이터 전처리 및 모델링 : 상권 밀집도 및 유형 예측



## 피처 셀렉션 / 모델링

### 지하철역



- 격자 별 가장 가까운 지하철역과의 거리를 구함

### 주차장



- 격자 별 가장 가까운 주차장과의 거리

### 임대주택



- 거리의 역수와 세대수를 기반으로 가중치를 계산
- 가까운 거리에 세대수가 많은 단지가 결과에 더 큰 영향을 줌

$$\text{가중치}_j = \frac{\text{단지 } j\text{의 세대수}}{\text{격자 중심과 단지 } j\text{의 거리}}$$

$$\sum_{j=1}^n (\text{가중치}_j \times \text{단지 } j\text{의 세대수})$$
$$\sum_{j=1}^n \text{가중치}_j$$

### 상권 밀집도 예측 모델

#### Random Forest

- max\_depth = 15
- min\_samples\_leaf = 1
- min\_samples\_split = 2

ACC 0.85

F1 0.79

과적합 방지

위한 소극적인  
하이퍼 파라미터  
조정

### 상권 유형 예측 모델

#### Random Forest

- max\_depth = 15
- min\_samples\_leaf = 1
- min\_samples\_split = 2

ACC 0.98

F1 0.96

# 5 데이터 전처리 및 모델링 : 격자별 유동인구 분석



## 분석목표

격자 기준으로  
상권, 버스정류장과 제조업 공장의 거리를 통해 유동인구 예측



## 사용 데이터

화성시\_성연령별\_유동인구.csv / 화성시\_격자.geojson

상권정보예측.csv / 화성시\_버스정류장.csv

화성시\_2021년, 2022년 격자별 유동인구의 평균.csv

화성시\_공장등록현황.csv



## 유동인구 평균 계산

### 1. 위경도별 유동인구 평균

2021년 ~ 2022년의 기준일자 및 위경도별  
전체 유동인구 수 도출

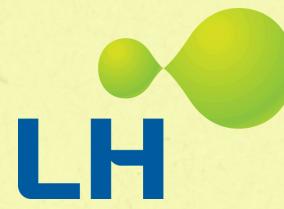
	STD_YM	lat	lon	pop_sum
0	202101	37.198232	126.971339	66.85
1	202101	37.198683	126.971335	92.62
2	202101	37.198235	126.971902	91.37
3	202101	37.199136	126.971896	25.29
4	202101	37.198237	126.972465	59.63
...	...	...	...	...
1968108	202212	37.203731	127.111036	354.82
1968109	202212	37.204182	127.111034	64.00
1968110	202212	37.204633	127.111031	20.11
1968111	202212	37.205083	127.111029	52.22
1968112	202212	37.205534	127.111027	74.07

	lat	lon	pop_avg
0	37.000817	126.724329	1.49375
1	37.003201	126.744535	0.35500
2	37.004027	126.732726	0.86500
3	37.004030	126.733288	0.85500
4	37.004084	126.741717	0.68500
...	...	...	...
187567	37.305562	126.668739	6.85000
187568	37.305574	126.670431	10.40800
187569	37.305578	126.670996	10.22600
187570	37.306464	126.668729	1.73500
187571	37.308195	126.658554	0.16500

gid	pop_avg	lon	lat
0	다바145999	0.170000	126.538832
1	다바154953	1.159996	126.549198
2	다바154954	1.654074	126.549279
3	다바154955	3.482000	126.549459
4	다바155954	2.120791	126.550311
...	...	...	...
58248	다사698102	3.711139	127.160397
58249	다사698103	82.010139	127.160486
58250	다사698104	9.990000	127.160108
58251	다사699102	60.077917	127.161241
58252	다사699103	37.148750	127.161239

58253 rows × 4 columns

# 5 데이터 전처리 및 모델링 : 유동인구 예측



## 데이터 전처리

### 1. Box-Cox 변환

데이터가 비정상성을 보일 때,  
데이터를 정규분포로 변환하는 기법

- Box-Cox 변환은 입력 값이 양수이어야 함
- y에 아주 작은 상수를 더하여 최소값이 0보다 크게 함( $y = y + 1e-9$ )
- ‘scipy.stats.boxcox’를 사용하여 변환 적용

### 3. RandomForestRegressor

결정 트리(Decision Tree)를 기반으로 한  
양상블 학습(Ensemble Learning) 방법

- ‘RandomForestRegressor’ 클래스를 사용하여 모델 생성
- 입력 데이터 X\_train\_scaled와 타겟 데이터 y\_train을 학습에 사용
- 학습된 모델로 테스트 데이터에 대한 예측 수행

## 평가지표(R-square, MSE)

Mean Squared Error = 2.69

R - square = 0.68

### 2. StandardScaler

각 특성을 평균이 0이고 분산이 1인  
정규분포에 가깝게 변환하는 기법

- train dataset의 각 특성별 평균과 표준편차를 계산한 후 데이터 변환
- 일관성을 유지하기 위해 같은 방식으로 test dataset에 적용하여 데이터 변환

### 4. Hyperparameter Tuning

머신러닝 모델의 학습 과정이나 구조를 결정하는  
설정 값을 조정하는 기법

- n\_estimators
- max\_depth
- min\_samples\_split
- bootstrap
- min\_samples\_leaf
- max\_features

→ GridSearch를 통해 최적의 Hyperparameter 탐색

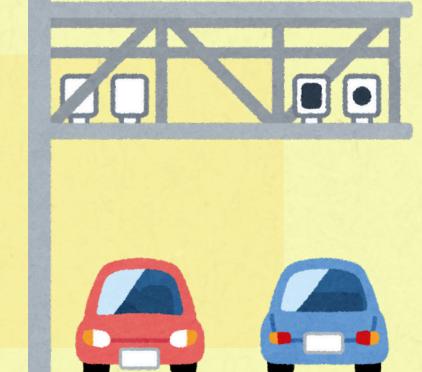
## 하이퍼 파라미터

n\_estimators = 200  
max\_depth = 20  
max\_features = ‘sqrt’  
min\_samples\_split = 2  
min\_samples\_leaf = 1  
bootstrap = True

# 5 데이터 전처리 및 모델링 : 격자별 교통량 추정

## 분석목표

- 도로의 접근성을 고려하여 인접한 도로의 교통량 계산
- 주택 단지별로 교통량 계산



## 사용 데이터

화성시\_추정교통량.csv / 화성시\_성연령별\_유동인구.csv

화성시\_상세도로망.csv / 화성2022.csv

- road\_rank가 101, 102, 103인 경우만 필터링



## 피처셀렉션

링크ID	반경 100m 상권밀집도 평균	반경 500m 상권밀집도 평균	반경 1000m 상권밀집도 평균	반경 100m 유동인구 평균	반경 500m 유동인구 평균	반경 1000m 유동인구 평균	가장 가까운 지하철과의 거리	가까운 휴게소와의 거리	추정교통량

특정 도로 반경에 위치한  
모든 상권밀집도의 평균

특정 도로 반경에 위치한  
모든 유동평균의 평균

타겟

# 5 데이터 전처리 및 모델링 : 격자별 교통량 추정

## CatBoost

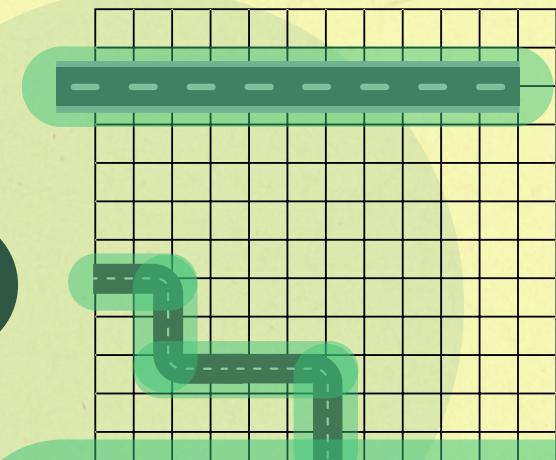
### 하이퍼 파라미터

- depth : 12 / iterations : 400 / learning\_rate : 0.2

### 평가지표(R-square, MSE)

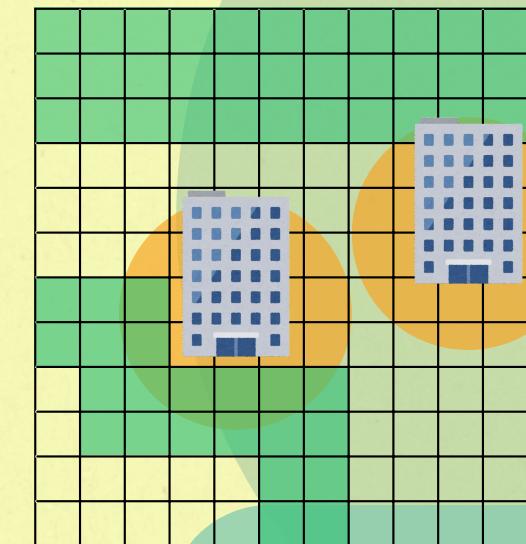
- Mean Squared Error = 7131.83
- R - square = 0.78

### 격자 매핑



격자를 거치는  
도로의 교통량의 합계 매핑

### 블록 매핑



블록 기준 지정한 반경만큼의  
격자의 교통량 합계 계산하여 매핑

화성	1.5km
하남	0.5km

### 반경 결정 기준

지역에 따라 통상적으로 블록이  
상행 고속도로와 만나는데 필요한 최소 거리

# 5 데이터 전처리 및 모델링 : 청년층 강도 예측

## 분석목표

- 거주 인구로부터 청년층이 밀집한 지역을 탐색
- 공공주택 단지별로 청년층의 강도 여부를 파악 하는 모델 생성



## 사용 데이터

공공주택임대\_정보(화성시).csv  
공공주택임대\_임대료(화성시).csv  
화성시\_공유차량\_이용현황.csv  
화성시\_상권유형.csv  
화성시\_격자.geojson

전국초중등학교위치표준데이터.csv  
경기도 공장등록 현황.csv  
화성시\_버스정류장.csv  
화성시\_지하철역.csv  
화성시\_거주인구.csv

b) 잔여 결측치는에 대해  
→ 80 ~ 100 대: 0으로 처리  
→ 20~70대: 코사인 유사도 기반 대체

## 데이터 전처리

### 1. 병합

단지코드, 위도, 경도를  
기준으로 거주인구 데이터 병합

단지코드	단지명	세대수	경도	위도	gid	m_20g_pop	w_20g_pop	m_30g_pop	w_30g_pop	m_40g_pop	w_40g_pop	m_50g_pop	w_50g_pop	m_60g_pop	w_60g_pop	m_70g_pop	w_70g_pop	m_80g_pop	w_80g_pop	m_90g_pop	w_90g_pop	m_100g_pop	w_100g_pop	total_pop	youth_pop	
0 000414	화성시_단지_12	1178	127.041121	37.217795	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN		
1 000415	화성시_단지_13	990	127.049830	37.212454	601127	122.0	108.0	171.0	... 62.0	146.0	35.0	100.0	597126	597127	597128	597129	597130	597131	597132	597133	597134	597135	597136	597137	597138	597139
2 000416	화성시_단지_14	836	127.047309	37.215054	598129	222.0	214.0	143.0	... 86.0	105.0	22.0	45.0	597140	597141	597142	597143	597144	597145	597146	597147	597148	597149	597150	597151	597152	597153
3 000447	화성시_단지_15	649	126.880070	37.295522	450187	50.0	33.0	59.0	... 74.0	50.0	0.0	44.0	451188	451189	451190	451191	451192	451193	451194	451195	451196	451197	451198	451199	451100	451101

### 2. 결측치처리

a) 2023년 기준으로  
거주인구 2022년 결측치 대체

단지코드	단지명	지원유형	세대수	경도	위도	gid	m_20g_pop	w_20g_pop	m_30g_pop	w_30g_pop	m_40g_pop	w_40g_pop	m_50g_pop	w_50g_pop	m_60g_pop	w_60g_pop	m_70g_pop	w_70g_pop	m_80g_pop	w_80g_pop	m_90g_pop	w_90g_pop	m_100g_pop	w_100g_pop	total_pop	youth_pop
0			0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

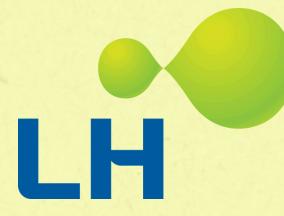
단지코드	단지명	지원유형	세대수	경도	위도	gid	m_20g_pop	w_20g_pop	m_30g_pop	w_30g_pop	m_40g_pop	w_40g_pop	m_50g_pop	w_50g_pop	m_60g_pop	w_60g_pop	m_70g_pop	w_70g_pop	m_80g_pop	w_80g_pop	m_90g_pop	w_90g_pop	m_100g_pop	w_100g_pop	total_pop	youth_pop
0			0	0	0	0	27	28	28	28	28	29	32	35	35	39	39	39	39	56	78	79	79	1	1	

대체 전

대체 후  
단지코드 별로 병합

df\_2022\_임대주택거주인구.csv

# 5 데이터 전처리 및 모델링 : 청년층 강도 예측



## 데이터 전처리2

파생변수 생성  $\Rightarrow$  X / y 구분  $\Rightarrow$  훈련/테스트 분할  $\Rightarrow$  데이터증강(SMOTE)

### 1. 변수 정리

- 지원유형통합:  
[공공임대, 통합공공임대]
- 평수 인코딩:  
[10평이하, 20평이하, 30평이하]
- 거주비율 인구 추가:  
20~30대 거주인구/전체 거주인구

### 2. 경도/위도 기반 거리 생성

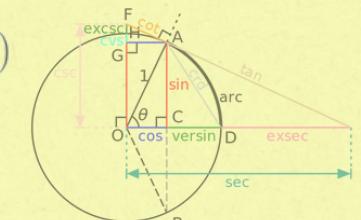
- 가장 가까운 대중교통까지와의 거리
- 가장 가까운 학교까지의 거리
- 가장 가까운 공장까지의 거리

#### ① 'haversine' 방식으로 거리 계산

$$a = \sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta\text{lon}}{2}\right)$$

$$c = 2 \cdot \text{atan}2\left(\sqrt{a}, \sqrt{1-a}\right)$$

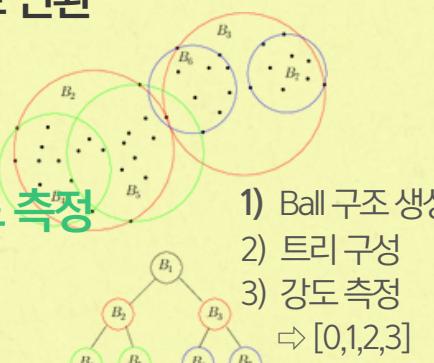
$$d = R \cdot c$$



### 3. 강도 기반 변수 생성

- 3km 반경의 상권밀집도에 기반한 변수 생성
- 청년층비율을 청년층 강도로 변환  
청년층비율  $\leq 0.3 \Rightarrow '0'$   
청년층비율  $> 0.3 \Rightarrow '1'$

#### ② 'BALLTREE'를 활용해 강도 측정



### 4. 데이터 분리

- X(피처), y(타겟) 구분
  - X(피처) : 주택정보, 대중교통 접근성, 주차장 접근성, 상권 강도, 공장 접근성, 학교 접근성
  - y(타겟) : 청년층 강도
- 훈련/테스트 분할



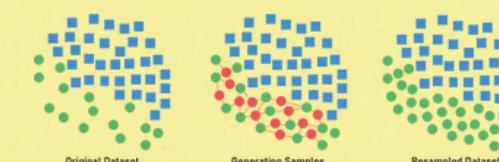
### 5. 데이터 증강

- 전체 데이터 수가 79행으로 매우 적고, 분류 모델을 학습시키기 위해 train set이 63행으로 적음.
- train set에 대한 데이터 증강(\*\*SMOTE 기법)을 통해 데이터 불균형을 제어하고 소수 클래스에 대한 학습력을 높임.

73	182	1	1	0	1	165.880033
61	450	1	1	0	1	142.091777
55	359	1	1	0	1	189.312526
40	1456	0	1	1	0	135.853891
9	2342	0	1	0	1	156.909011
—	—	—	—	—	—	—
20	1547	1	1	0	1	102.189552
60	440	1	1	0	1	131.959404
71	149	0	1	0	1	116.060009
14	583	0	1	0	1	74.323639
51	538	1	1	0	1	275.754472
—	—	—	—	—	—	—
63 rows × 17 columns						

0	182	1	1	0	1	165.880033
1	450	1	1	0	1	142.091777
2	359	1	1	0	1	189.312526
3	1456	0	1	1	0	135.853891
4	2342	0	1	0	1	156.909011
—	—	—	—	—	—	—
65	809	0	0	0	1	154.198848
66	804	0	1	0	1	113.811208
67	442	0	0	0	0	118.587749
68	828	0	1	0	0	157.335741
69	731	1	0	0	1	503.797348
70 rows × 17 columns						

#### \*\* SMOTE 기법



선형 보간법을 사용하여 소수 클래스 데이터의 새로운 샘플을 생성

# 5 데이터 전처리 및 모델링 : 청년층 강도 예측

## 모델 생성

임대주택정보, 공영주차장, 상권, 대중교통 접근성, 공장 접근성, 학교 접근성을 토대로  
“청년층 강도”  $\Rightarrow$  [0,1]를 예측하는 분류 모델을 생성함



<모든 변수(feature) 활용>

	Random Forest	Gradient Boosting	SVM (RBF)
정확도	0.50	0.44	0.62
정밀도	0.50	0.43	0.79
재현율	0.50	0.44	0.62
F1-score	0.47	0.42	0.56



Feature selection

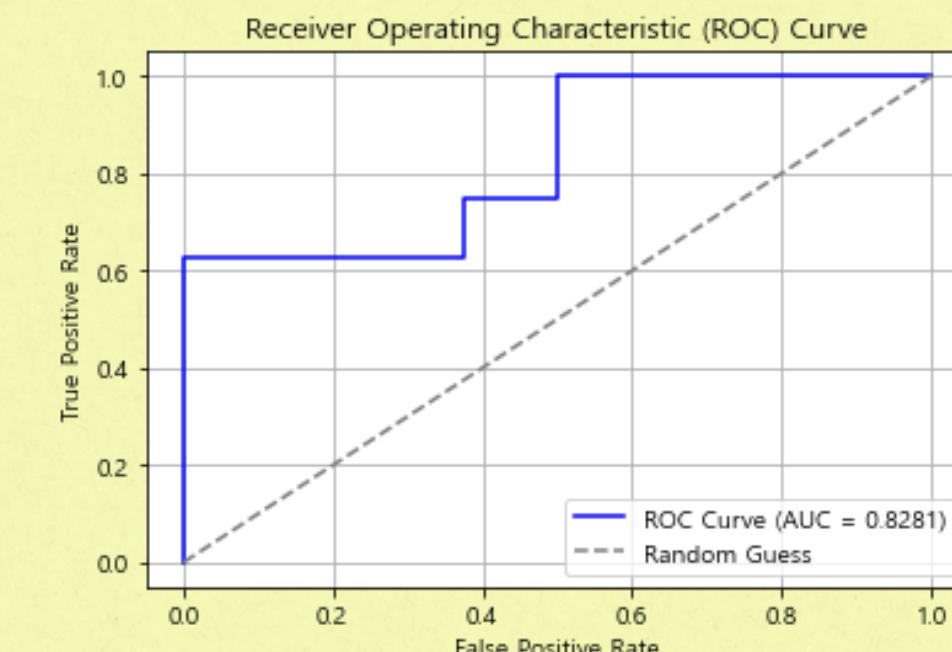


<특정 변수[주택정보, 대중 교통 접근성]만 활용>

## 그리드 서치

learning\_rate(학습률) : 0.2  
max\_depth(최대 깊이) : 7  
n\_estimators(트리 개수) : 200  
subsample(샘플링 비율) : 0.8

## 최종 모델

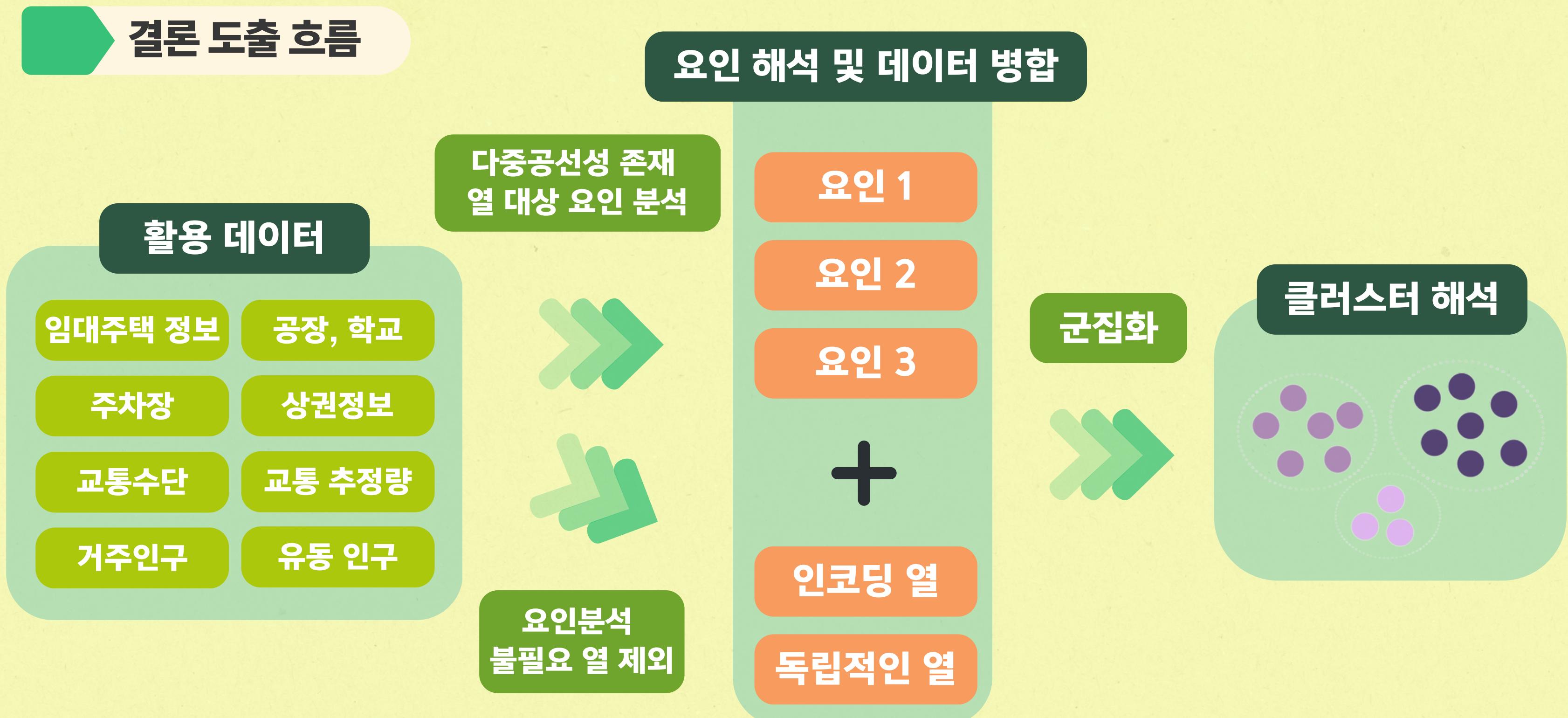


◆ 최종모델  
 $\Rightarrow$  83% 확률로 양성과 음성을 올바르게 구분 가능

## 하이퍼 파라미터

정확도: 0.75  
정밀도: 0.83  
재현율: 0.75  
f1-score: 0.73

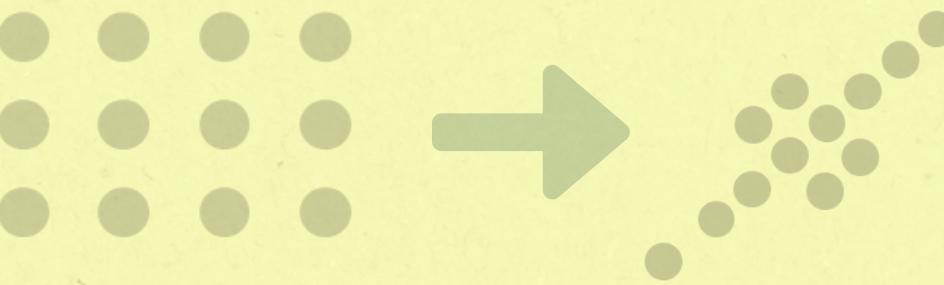
# 6 요인분석 및 클러스터링



# 6 요인분석 및 클러스터링

## 요인 분석 과정

### 요인분석 실행 근거



- **다중공선성 제거: 상관관계가 존재하는 열 대상 요인분석 실행**
  - One-Hot encoding 된 열은 요인분석에 제외
  - 다중공선성이 존재하지 않는 독립적인 열은 요인분석에 제외
- **차원의 저주 예방: 추후 군집화에서 효율적**
  - 요인분석으로 인해 차원이 줄어들어 효과적인 거리 계산 가능

### 다중공선성 존재 열

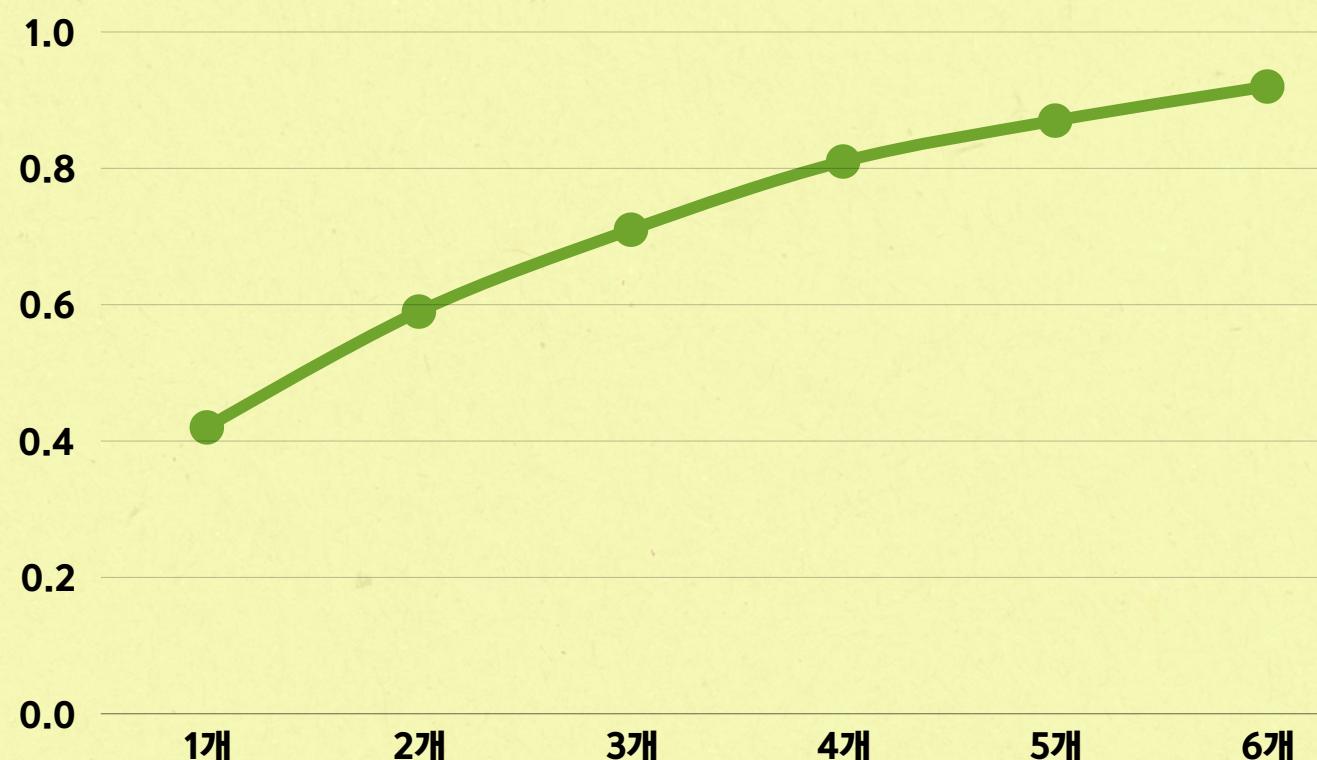
데이터열
가장 가까운 주차장까지의 거리
가장 가까운 지하철역까지의 거리
반경 3km 내의 상권 유형 최빈값
반경 3km 내의 상권 강도 평균값
반경 1.5km 내의 교통 추정량 총합
가장 가까운 공장까지의 거리
가장 가까운 초,중,고등학교까지의 거리
반경 1km 유동인구 총합

# 6 요인분석 및 클러스터링

## 요인 분석 과정 \_ 화성시

### 요인분석 실행 결과

- **요인 개수 선택**
  - 누적 설명력이 0.8을 넘어가는 지점
  - 4개의 요인으로 80.2%를 설명하고 있는 것으로 확인



### 요인 해석

#### • 요인 1: 주차장 및 교육 접근성 요인

- 가장 가까운 주차장 거리 ↑
- 초,중,고등학교까지의 거리 ↑

Factor 1

#### • 요인2: 상권 및 유동 인구 요인

- 반경 3km 상권강도 ↑
- 반경 3km 상권유형 ↑

Factor 2

#### • 요인3: 교통량 및 산업 요인

- 근방 교통 추정량의 합 ↑
- 가까운공장까지의 거리 ↓

Factor 3

#### • 요인4: 역세권 요인

- 가장 가까운 지하철역 거리 ↑

Factor 4

# 6 요인분석 및 클러스터링

## 군집화 결과 \_ 화성시

### 데이터 병합

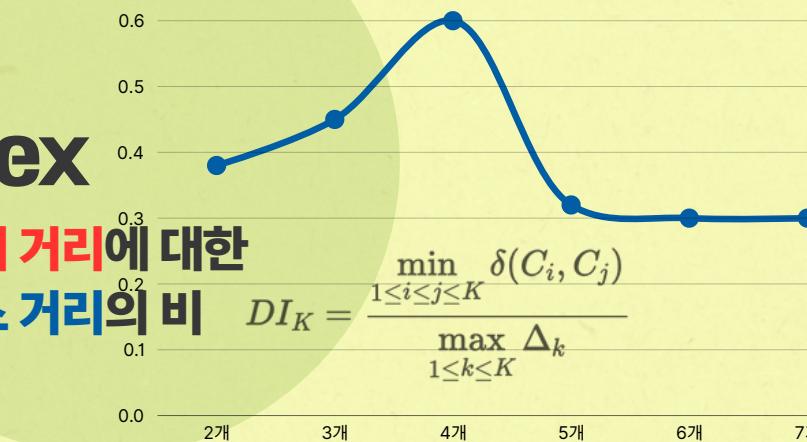
축소된 요인 4개와 요인분석하지 않은 열을 병합한 후 MinMaxScaler

- 0과 1 사이의 값으로 변환하여 군집화 효율
- 요인들 간의 상관관계가 없음을 확인
- 3단계 : eps 100m / min\_samples 300

### Dun-index

클러스터 내 최대 거리에 대한  
클러스터 간 최소 거리의 비

$$DI_K = \frac{\min_{1 \leq i \leq j \leq K} \delta(C_i, C_j)}{\max_{1 \leq k \leq K} \Delta_k}$$



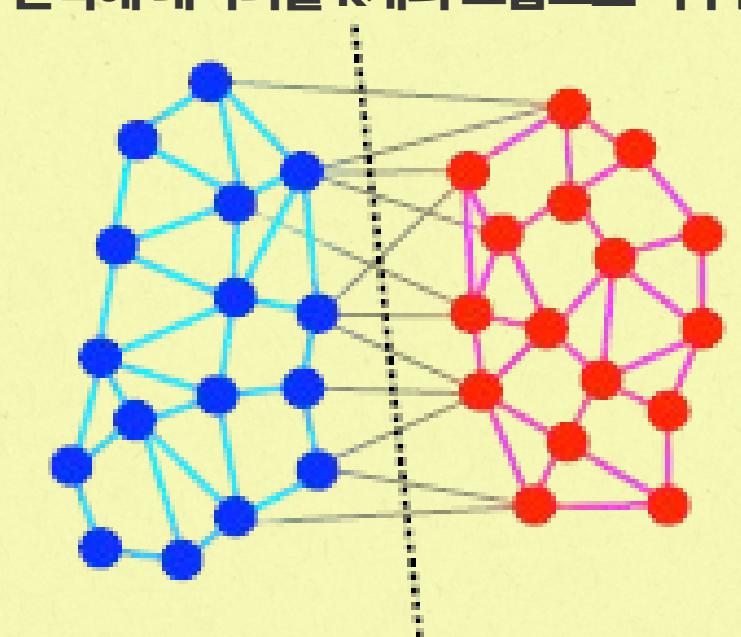
### Silhouette score

할당된 군집 내 데이터와 얼마나 가까우며,  
타군집의 데이터와 얼마나 멀리 있는지

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

### Spectral-Clustering

- 데이터를 그래프 형식으로 변환
- 거리 기반으로 유사성을 계산하여 유사한 데이터끼리 연결
- 그래프 구조를 분석해 데이터를 k개의 그룹으로 나누는 방식

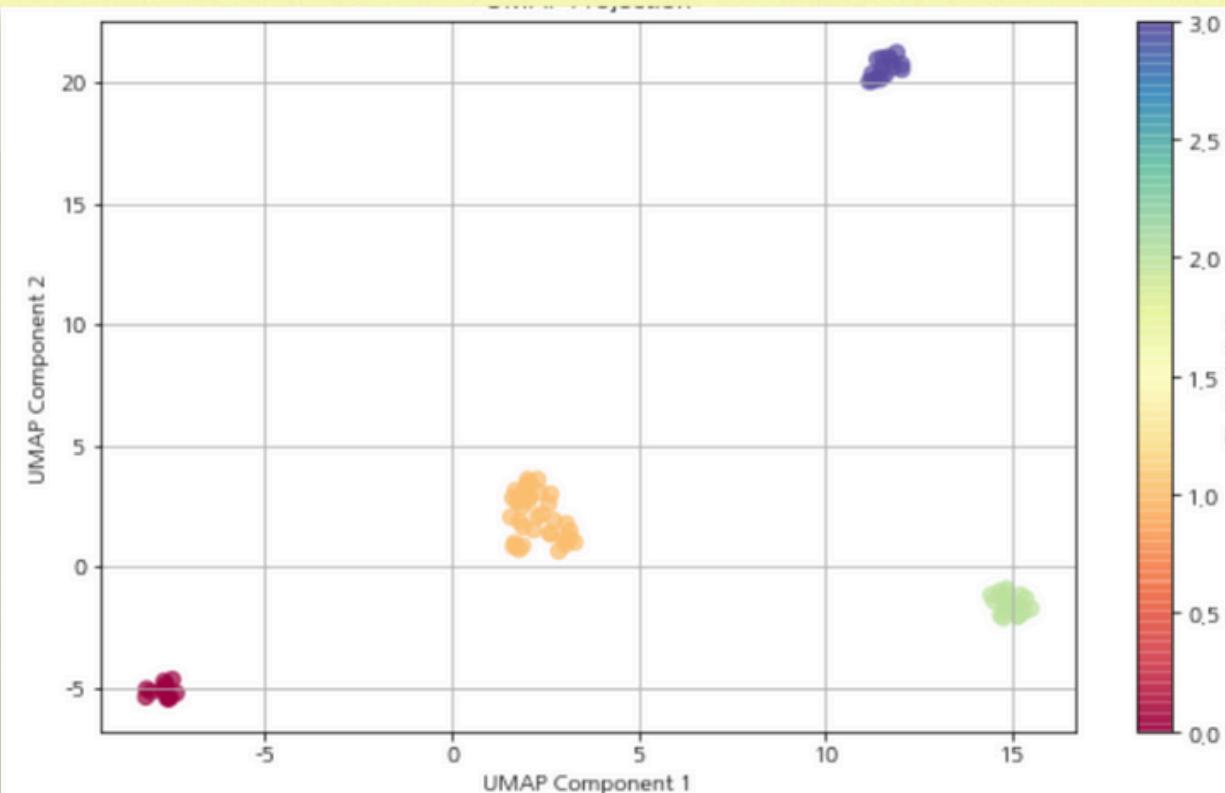


# 6 요인분석 및 클러스터링

\*\*20~30대 연령의 인구와 대중교통 승·하차 인원이 많은 지역에서 카셰어링 이용이 활발했음

근린환경 요인을 고려한 카셰어링 대여소 배치 방안 연구 - 서지민, 석종수 (2017).

## 클러스터링 해석

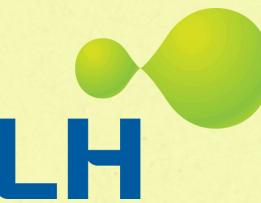


화성시에 새롭게 공유차량을 선정함으로써, 기존에  
공유차량이 적절하게 산정되지 않고 있었음을 확인할 수 있음

## 클러스터

- **클러스터 0 : (3대 배치가 적정해 보임)**
  - \*\*공공임대 + 청년층 + 높은 유동인구/상권강도 + 가까운 버스정류장  
⇒ 수요량에 긍정적 영향
  - 0%의 공유차량 폐쇄됨.
- **클러스터 1 :(1대 배치가 적당해 보임)**
  - 통합공공임대 + 청년층 + 가까운 버스정류장  
⇒ 적당한 수요 예상
  - 25%의 공유차량이 폐쇄됨.
- **클러스터 2 :(0대 배치가 적당해 보임)**
  - 통합공공임대 + 청년층 X + 낮은 유동인구 + 먼 버스정류장  
⇒ 수요에 부정적 영향
  - 92%의 공유차량이 폐쇄됨.
- **클러스터 3 : (2대 배치가 적당해 보임)**
  - 통합공공임대 + 높은 유동인구/상권강도 + 가까운 버스정류장  
⇒ 수요에 긍정적 영향
  - 25%의 공유차량이 폐쇄됨.

# 7 적정 규모 산정 : 하남시 데이터 생성



# 하남교산지구단위계획결정조사



1

교산\_신설공동주택정보.csv

블록	세대수	통합공임 여부	10평 이하	20평 이하	30평 이하
A1	...	...	...	...	...
...	...	...	...	...	...
C3	...	...	...	...	...

토지이용계획도\_지구계획변경(2차)-모형



학교



### 3 지하철



1

- 신설 학교 14곳의 좌표 생성

하남 내 신설하는 3개의 역사 좌표 생성  
기존 하남 지하철 데이터와 concat

모형에 따라 QGIS 상에서 주요 도로의 shp 생성  
일부 교산 일대의 도로 데이터와 concat



주차장

- 신설 공영주차장 정보가 없으므로 자체적으로 입지선정

# 7 적정 규모 산정 : 하남시 주차장 입지 선정

\*\*상권 : 대상구역도의 blockType 중 '공동주택', '단독주택', '주상복합'은 제외

## 사용 데이터

하남시\_격자.geojson / 3.대상구역도(하남교산).geojson  
하남시\_버스정류장.csv / 하남\_지하철.csv / 하남시\_공영주차장.csv

### 1. 교산지구에 해당하는 격자 추출

대상구역도(하남교산)과  
하남시 격자 데이터 병합

⇒ 'result\_grid' 생성

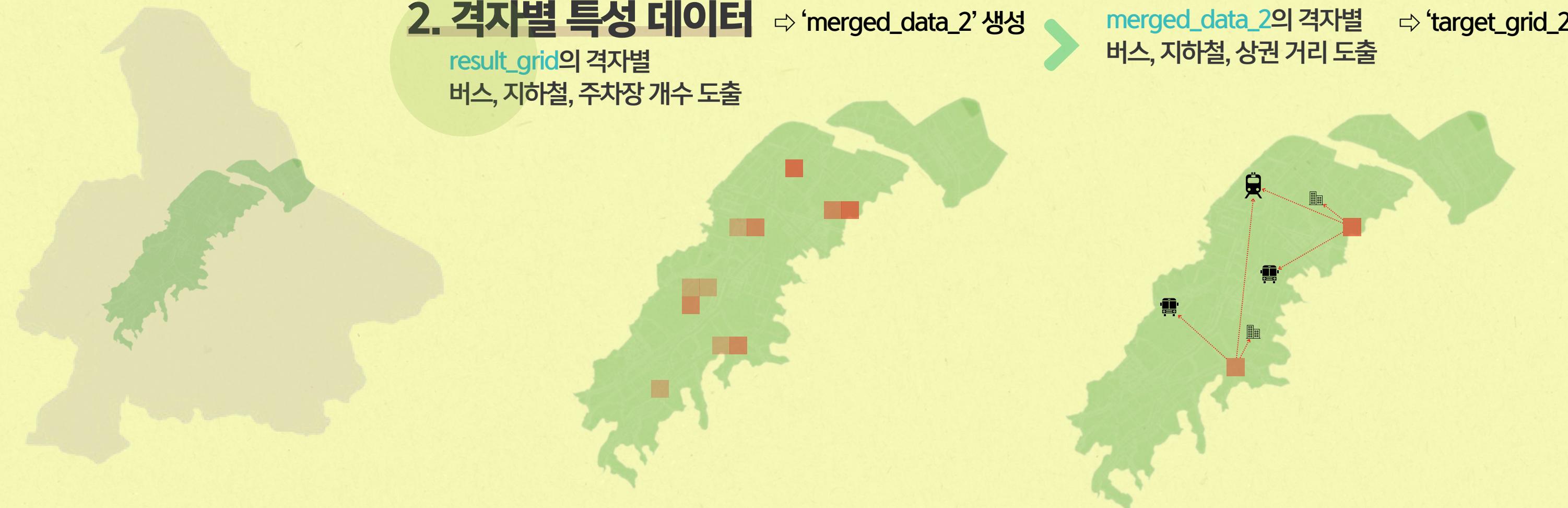
대상구역도의 blockType이 '공동주택', '단독주택', '주상복합'이 아닌 데이터 필터링  
격자별 버스 정류장, 지하철 정거장, \*\*상권 과의 최소 거리 도출

### 2. 격자별 특성 데이터

result\_grid의 격자별  
버스, 지하철, 주차장 개수 도출

⇒ 'merged\_data\_2' 생성

merged\_data\_2의 격자별  
버스, 지하철, 상권 거리 도출  
⇒ 'target\_grid\_2' 생성



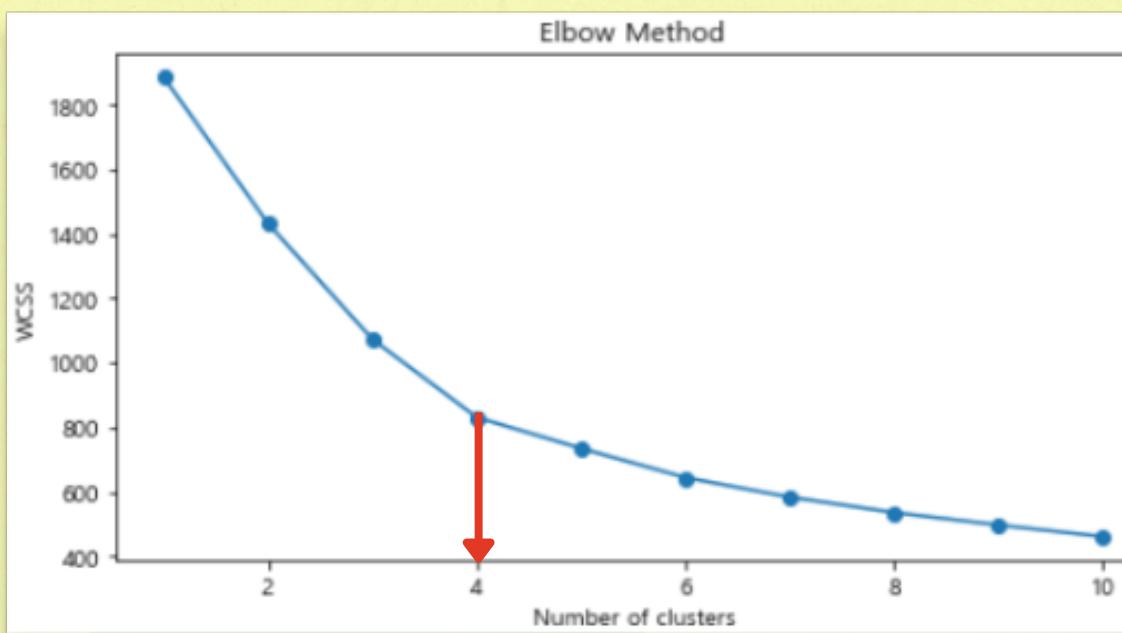
# 7 적정 규모 산정 : 하남시 주차장 입지 선정

## 클러스터링

target\_grid\_2의 버스 정류장, 지하철 정거장, 상권의 거리 데이터로 클러스터링

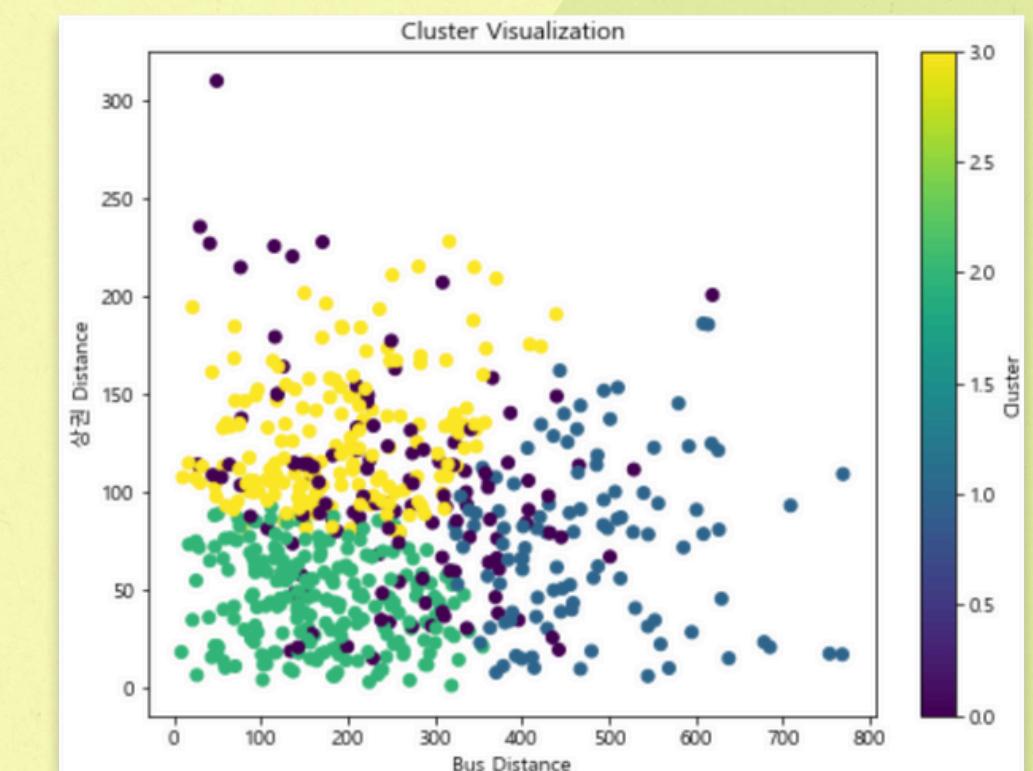
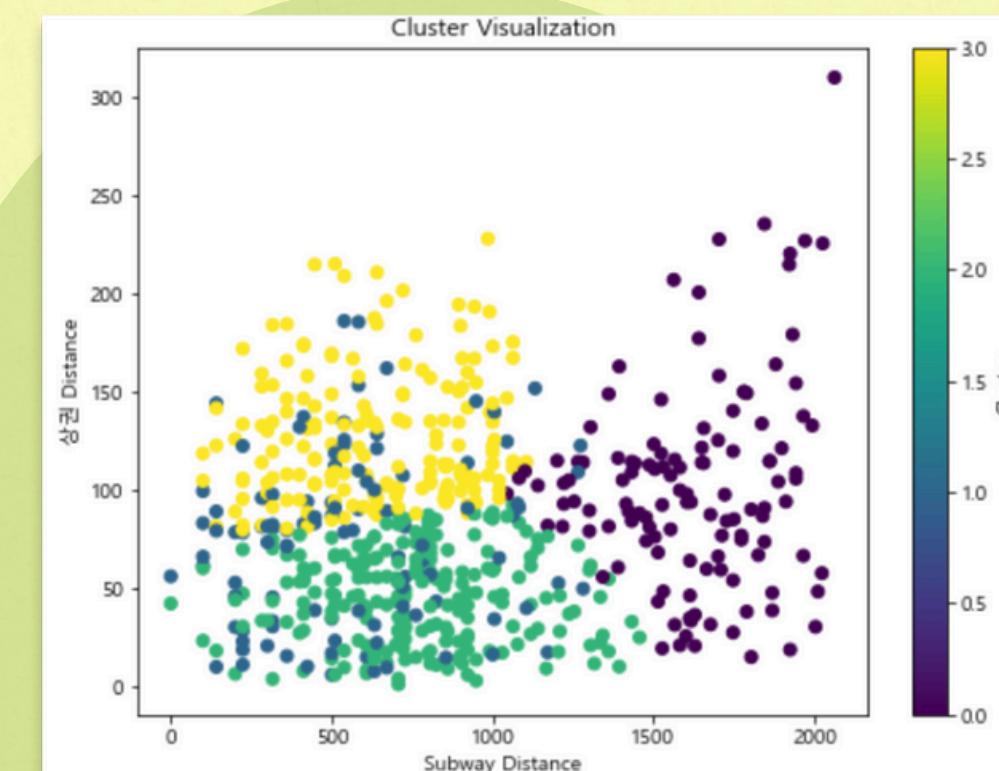
### 1. Elbow Method $\Rightarrow$ 4개로 설정

- Elbow Method로 적정 클러스터 개수 설정



### 2. 입지 선정 후보 cluster 선택

- 버스 정류장과 지하철 정거장과의 거리가 멀고, 상권과의 거리는 짧은 cluster 1로 선택



# 7 적정 규모 산정 : 하남시 주차장 입지 선정

## 최종 입지 선정

### 1. 입지 선정 후보 격자

클러스터가 1인 격자

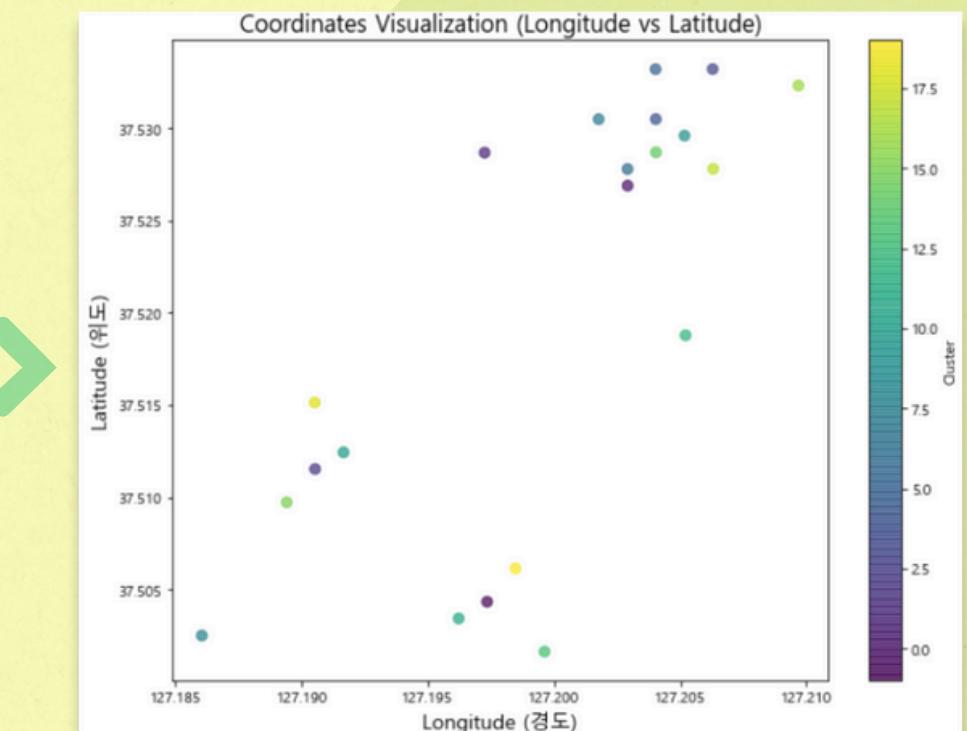
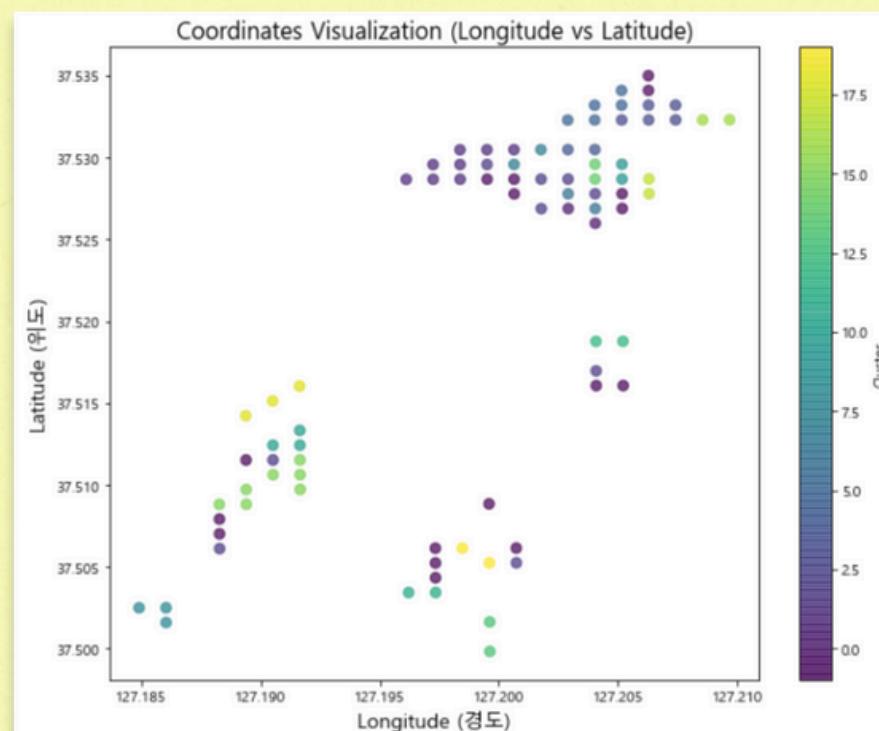
gid	nearest_subway_distance	nearest_bus_distance	nearest_fac_distance	cluster	
0	다사737475	100.039167	440.146793	61.652351	1
1	다사732450	728.295840	417.271102	81.456037	1
2	다사732477	510.101580	354.791265	112.523778	1
3	다사733479	500.195768	544.895314	6.095312	1
4	다사732478	538.727312	452.581971	125.615930	1
...	...	...	...	...	
100	다사734451	721.392454	530.396719	40.819657	1
101	다사735476	200.078320	326.066649	52.990849	1
102	다사740484	854.735097	400.078157	14.865199	1
103	다사738478	223.694384	595.247550	28.538802	1
104	다사738463	1000.391470	393.339180	16.425877	1

105 rows × 5 columns

blockType이 ‘주차장’에 해당한 격자 제외



### 2. 최종 입지 선정



DBSCAN으로 클러스터링을 진행하여  
클러스터별 대표 격자 추출

# 7 적정 규모 산정 : 청년층강도 및 상권 예측

## 하남 데이터 적용

## 교산 지구에 대한 청년층의 강도와 상권 예측

### Gradient Boosting Classification

#### 청년층강도 예측

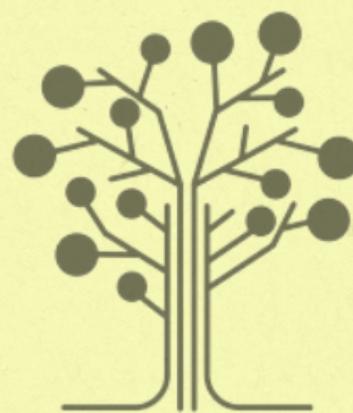
GridSearch로 학습한 최적의 parameter 기준으로  
교산지구의 데이터를 적용하여 교산지구 청년층강도 예측

- learning\_rate = 0.2
- max\_depth = 7
- n\_estimators = 200
- subsample = 0.8



#### 청년층강도 예측값

array([0, 0, 0, 0, 1, ..., 1, 0, 0])



### Random forest Classification

#### 상권밀집도 예측

GridSearch로 학습한 최적의 parameter 기준으로  
하남시의 데이터를 적용하여 하남시의 상권 밀집도 예측

- max\_depth = 15
- min\_samples\_leaf = 1
- min\_samples\_split = 2
- n\_estimators = 100



#### 상권밀집도 예측값

array([3., 3., 1., ..., 2., 0., 0.])

#### 상권유형 예측값

array([2, 2, 1, ..., 2, 0, 0])

#### 상권유형 예측

GridSearch로 학습한 최적의 parameter 기준으로  
하남시의 데이터를 적용하여 하남시의 상권 유형 예측

- max\_depth = 30



# 7 적정 규모 산정 : 유동인구 및 교통 추정량 예측

## 하남 데이터 적용

## 교산 지구에 대한 유동인구와 교통 추정량 예측

### Random forest Regression

#### 유동인구 예측

GridSearch로 학습한 최적의 parameter 기준으로  
하남시의 데이터를 적용하여 하남시의 유동인구 예측

- max\_depth = 20
- min\_samples\_leaf = 1
- min\_samples\_split = 2
- n\_estimators = 200
- bootstrap = True
- max\_features = 'sqrt'



#### 유동인구 예측값

```
array([[3.13195059 1.79761927 3.13179804  
... 1.85282169 1.98655641 2.0796592 ]])
```



### CatBoosting Regression

#### 교통추정량 예측

GridSearch로 학습한 최적의 parameter 기준으로  
하남시의 데이터를 적용하여 하남시의 상권 밀집도 예측

- depth = 11
- iterations = 400
- learning\_rate = 0.2



#### 교통추정량 예측값

```
array([441134.32, 922094.43, ...  
980388.02, 2499372.24])
```



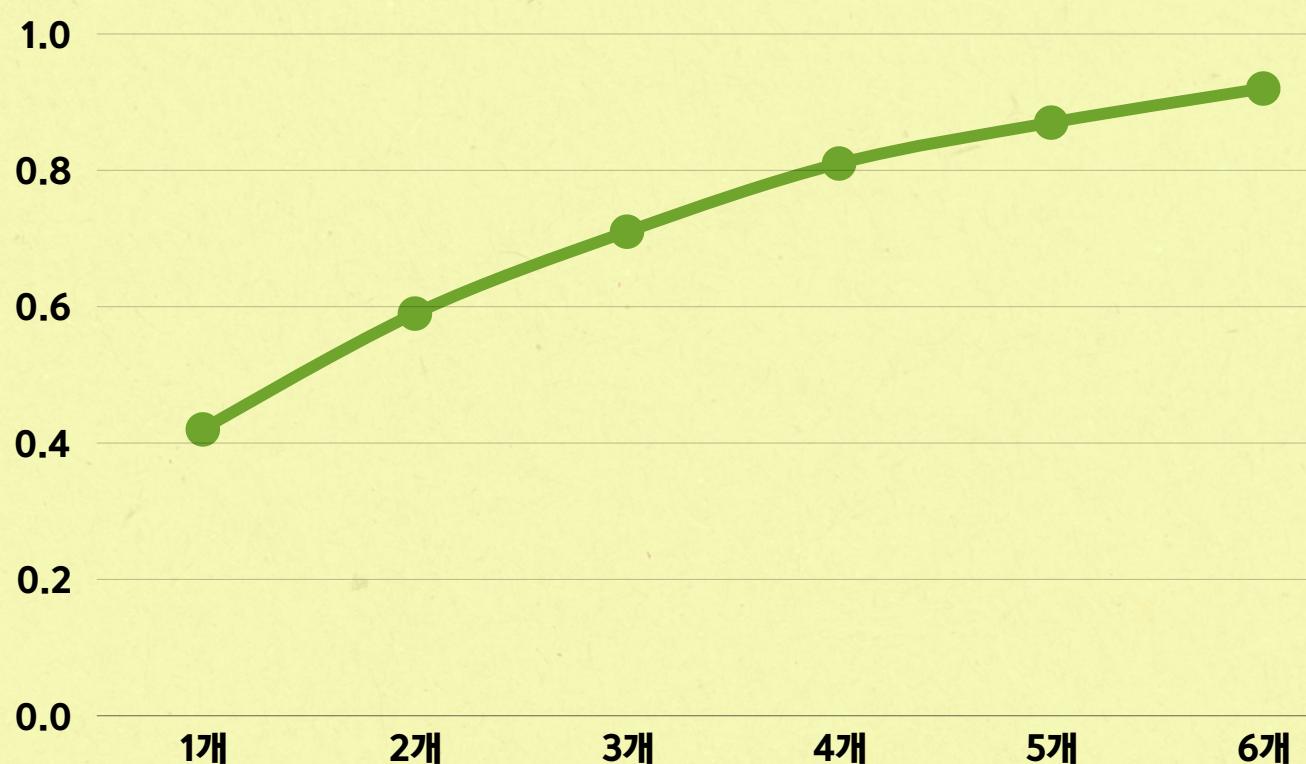
# 7 적정 규모 산정 - 요인분석

## 요인 분석 과정 \_ 교산지구

### 요인분석 실행 결과

#### • 요인 개수 선택

- 누적 설명력이 0.8을 넘어가는 지점
- 4개의 요인으로 80%를 설명하고 있는 것으로 확인



### 요인 해석

#### • 요인 1: 교육 접근성 및 유동인구, 상권 요인

- 초, 중, 고등학교까지의 거리 ↑
- 반경 1km 유동인구 ↓
- 반경 3km 상권강도 ↓

Factor 1

#### • 요인 2: 교통량 및 주차, 역세권 요인

- 공장으로 인한 근방 교통 추정량의 합 ↑
- 가장 가까운 지하철역 거리 ↑

Factor 2

#### • 요인 3: 청년층과 세대수 및 버스정류장 요인

- 청년층 강도, 세대수 ↑
- 가까운 버스정류장 거리 ↓

Factor 3

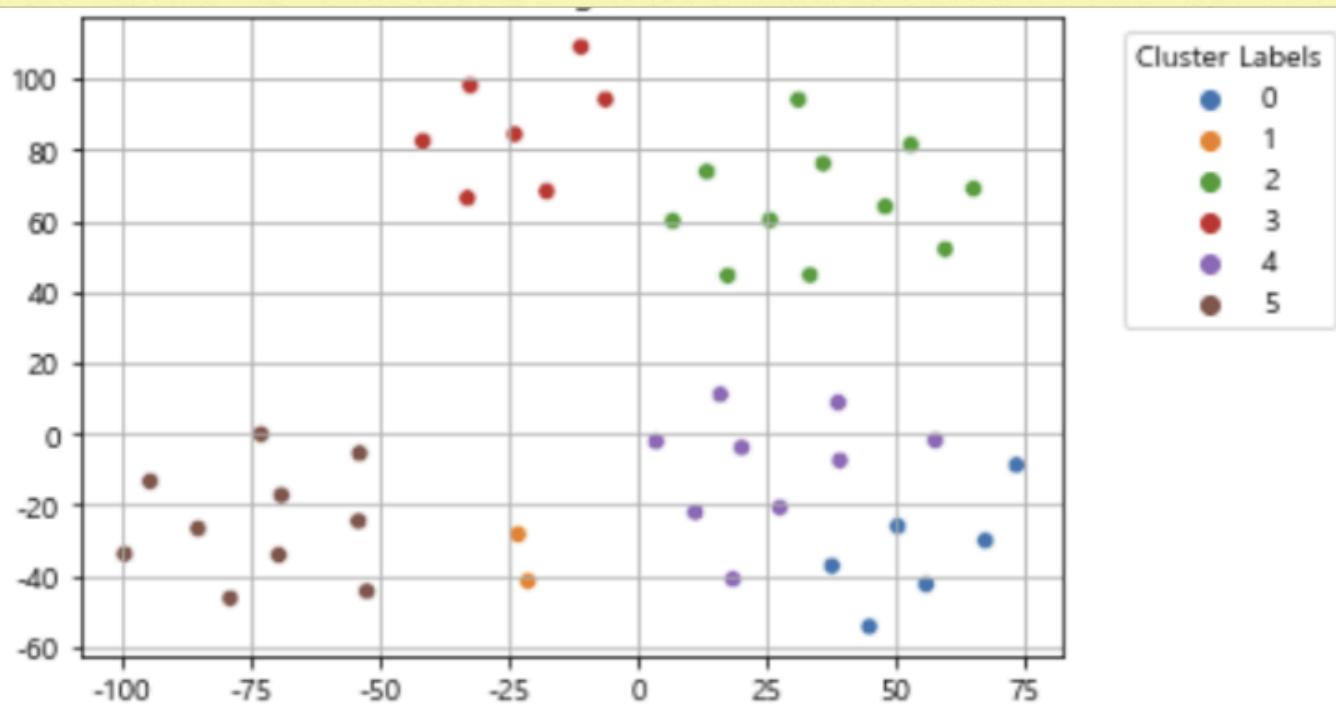
#### • 요인 4: 산업단지 요인

- 가장 가까운 산업단지 거리 ↓

Factor 4

# 7 적정 규모 산정 - 클러스터링

## 클러스터링 해석



화성시에 존재하지 않았던 30평 이상의 교산단지에 적절한 공유차량을 선정해주기 위해 정교한 클러스터링이 필요함

- **클러스터 0 :** (1대 배치가 적정해 보임)
  - 높은 교통량 + 청년층 + 버스정류장 가까움  $\Rightarrow$  수요량에 긍정적 영향
  - 30평대(공공임대)  $\Rightarrow$  아이가 없는 20-40대 신혼부부
  - [화성시의 클러스터 0과 유사하지만](#) / 유동인구  $\downarrow$   $\Rightarrow$  장거리 운행위주  $\uparrow$
- **클러스터 1 :** (3대 배치가 적당해 보임)
  - 교육 접근성  $\uparrow$  + 청년층 + 유동인구  $\uparrow$   $\Rightarrow$  수요 대폭 증가
  - 10평대(공공임대)  $\Rightarrow$  아이가 있는 젊은 신혼부부
  - [화성시의 클러스터 0과 가장 유사함](#)
- **클러스터 2 :** (1대 배치가 적당해 보임)
  - 대중교통 접근성  $\downarrow$  + 청년층 X + 교육접근성  $\downarrow$   $\Rightarrow$  수요에 부정적 영향
  - 20평대(공공임대) / 중간수준의 유동인구
  - [화성시의 클러스터 3과 유사하지만](#) / 낮은 교육접근성  $\Rightarrow$  소극적 대여
- **클러스터 3 :** (0대 배치가 적당해 보임)
  - 청년층 X + 먼 버스정류장 + 낮은 유동인구  $\Rightarrow$  수요에 부정적 영향
  - 20평대(통합공공임대)  $\Rightarrow$  저소득 고령층
  - [화성시의 클러스터 2와 가장 유사](#)
- **클러스터 4 :** (2대 배치가 적당해 보임)
  - 교육 접근성  $\uparrow$  + 유동인구  $\uparrow$  + 산업단지 가까움  $\Rightarrow$  수요에 긍정적 영향
  - 30평대(공공임대) / 혼합된 세대  $\Rightarrow$  안정적 생활
  - [화성시의 클러스터 0과 유사하지만](#) / 적은 청년층  $\Rightarrow$  자차 보유 가능성  $\uparrow$
- **클러스터 5 :** (2대 배치가 적당해 보임)
  - 높은 교통량(가까운 공장) + 버스정류장 가까움 + 청년층  $\uparrow$   $\Rightarrow$  수요에 긍정적 영향
  - 10평대(통합공공임대)  $\Rightarrow$  저소득 청년층  $\Rightarrow$  자차 보유 부담  $\uparrow$
  - [화성시의 클러스터 1과 가장 유사함](#)

# 7 적정 규모 산정

## 최종 데이터 조정



블록코드	차량대수
A1	3
A1	3
A3	1
A3	0
A6	1
A6	0
	...

최종적으로 공유차량을 산정한 이후, 같은 블록코드  
를 갖는 행 16개에 대해서 확인해본 결과

블록코드	차량대수
A1	3
A3	1
A6	1
	...

같은 블록 코드(동일 단지) 내에  
1개의 블록 코드에 3 대가 배정되어 있었으며,  
6개의 블록 코드에 1대와 0대가 배정되어 있었음

“  
블록 코드를 고려하지 않았음에도 동일 단지 내에서  
자연스럽게 균형잡힌 차량배정이 이루어짐

클러스터링 모델이 **지역적·인구적 특성을 효과적으로 학습했음**을 간접적으로 입증함.  
데이터의 내재적 특성을 기반으로 해석 가능한 결과를 도출하며,  
**차량 수요를 반영한 성공적인 클러스터링 모델**을 구축했음을 보여줌.

# 7 적정 규모 산정 조정

## 최종 선정 결과



차량 1대



차량 2대



차량 3대



## 기대효과

- 입주민 이동 편의성 향상
  - 공유차량 가용성을 높여 입주민이 편리하게 이동 가능
- 교통 불편 지역 해소
  - 대중교통이 부족한 지역에 공유 차량을 배치하여 교통 소외 문제 완화



## 활용 방안

- 차량 배치 최적화
  - 분석 결과를 바탕으로 지역별 공유차량의 수요를 예측
  - 적정 개수를 배치하여 차량 부족 또는 과잉 문제를 해소
- 유지비용 절감
  - 적정 개수만 운영함으로써 차량의 유지 및 관리 비용 절감

## 한계점

- 교산지구 신도시 데이터가 충분하지 않음
  - 화성시 및 교산을 제외한 하남시에는 존재하는 데이터가 미개발 도시인 교산시에 존재하지 않았기 때문에 정확한 공유차량 산정에 어려움을 겪었음



감사합니다.

Thank You

