혼자 사는 청년들을 위한 서울시 살 곳 추천

I·SEOUL·U

4팀 데이터마이닝
박정현 김지민 노정화 염예빈 전규리

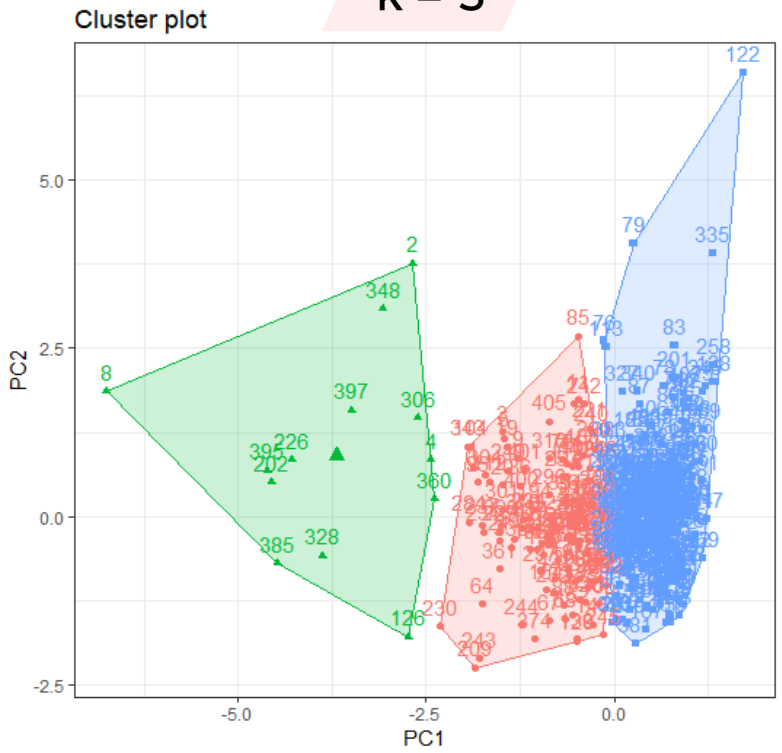
02 PCA(주성분 분석)

PCA란?

클러스터링 결과

k = 2

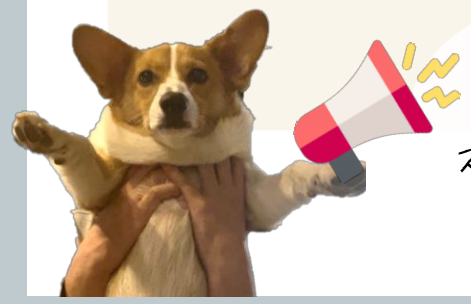
k = 3



02 Feature Engineering : 내부적 요소 (1) 시세

2주차 복습 …

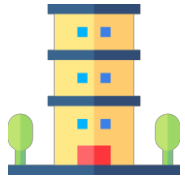
- ✓ 전세와 매매는 거래가 월세보다 많이 일어나지 않아 데이터가 아예 없는 행정동이 있음
- ✓ 1인 가구의 주된 거주 형태는 월세




아파트


연립다세대


단독다가구


오피스텔

mean_mdeposit : 월세 보증금

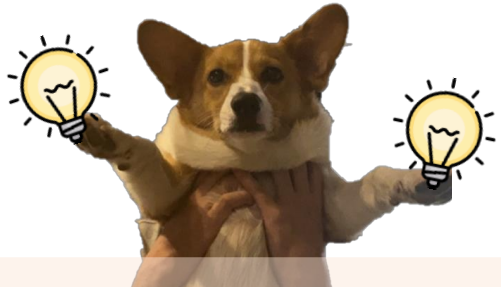
mean_month : 월세액

각 주택 종류별로 2개의 변수가 월세 데이터 반영

02 Feature Engineering : 내부적 요소 (1) 시세

$$\text{월세 이율} = \frac{\text{월세 가격}}{\text{전세금 - 월세 보증금}} \times 100$$

월 단위의
전월세 전환율



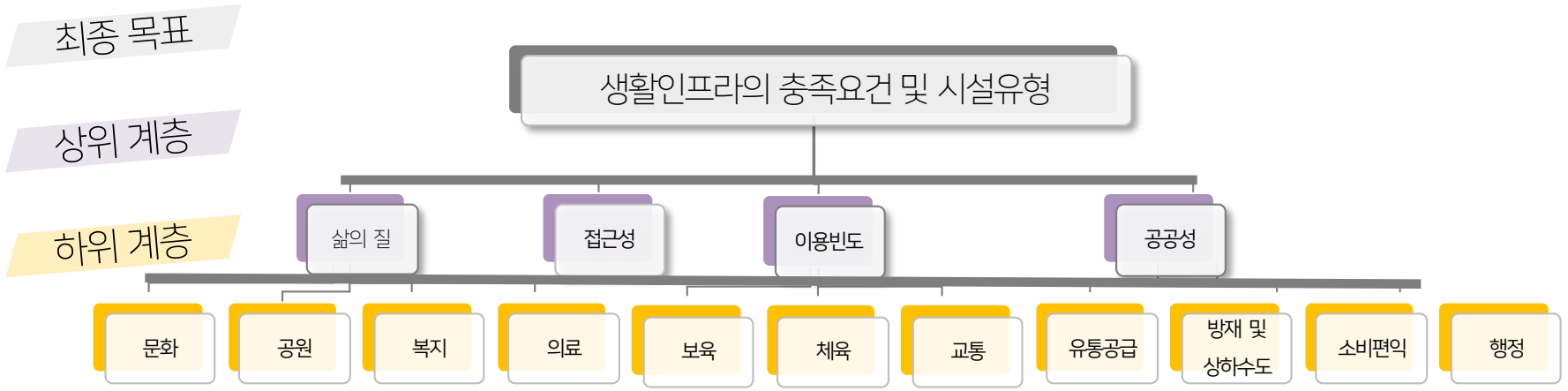
월세 이율

전월세 전환율 ÷ 12

02 Feature Engineering : 외부적 요소(1) 생활인프라 지수

종합적 지표 제안 : ' 생활인프라 지수 '

생활인프라 충족요건 및 시설유형 판단요인 계층구조



국토연구원, 『생활인프라 실태의 도시간 비교분석 및 정비방안』

02 Feature Engineering : 외부적 요소(1) 생활인프라 지수

종합적 지표 제안 : ' 생활인프라 지수 '

AHP분석기법을 적용해 산출된 가중치

0.427 삶의 질 (QoL)

0.263 접근성

0.173 이용빈도

0.137 공공성

02 Feature Engineering : 외부적 요소(2) 치안

치안 관련 파생변수 정제: 2. 범죄 관련 구단위 변수들

치안	CCTV개수
	죄종별 범죄발생건수(구단위)
	도시위험도지표(구단위)

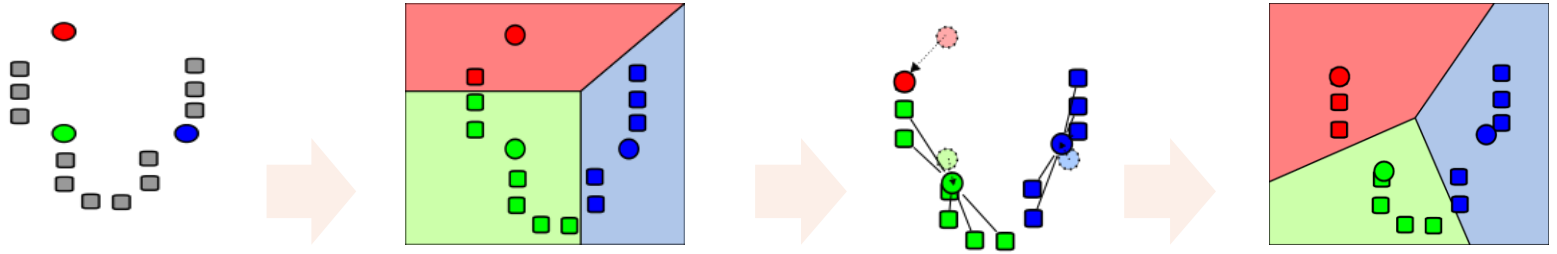
자료 특성상 구단위로밖에 제공하지 않음,
행정동별로 변환하더라도 정확도 상실

→과감히 삭제!

03 클러스터링

1 " K-Means Clustering "

주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로,
각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작



무작위로 데이터 중 K개의 중심점 추출

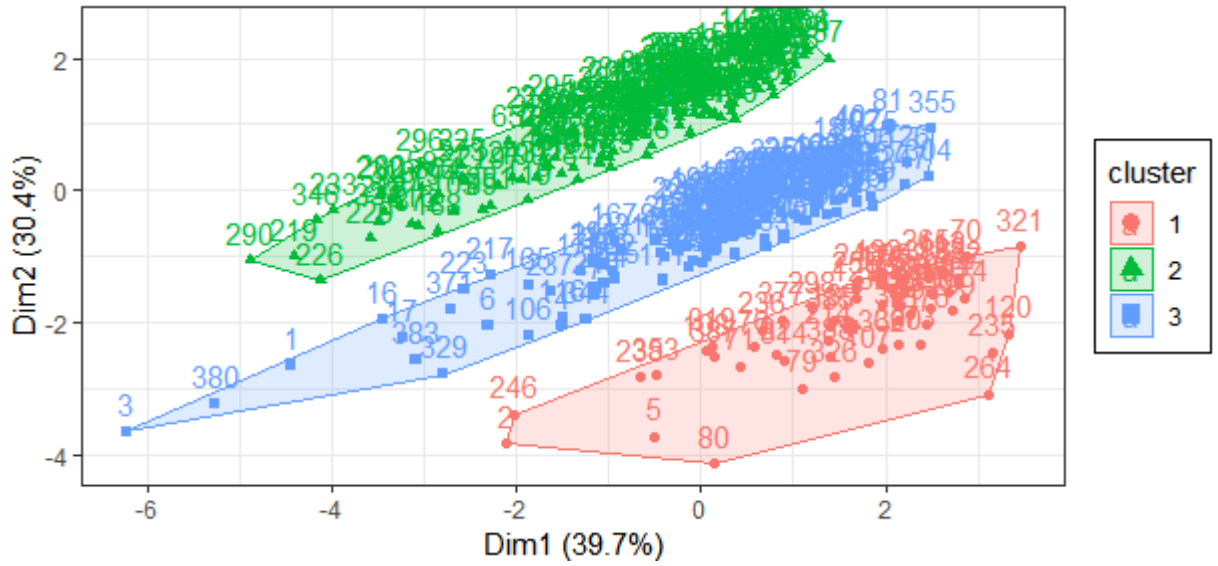
각 오브젝트들이 가장 가까운 중심점을 기준으로 묶임

각 클러스터에서 중심점 이동

고정된 클러스터로 수렴할 때까지 반복

03 클러스터링 : K-MEANS

Cluster plot

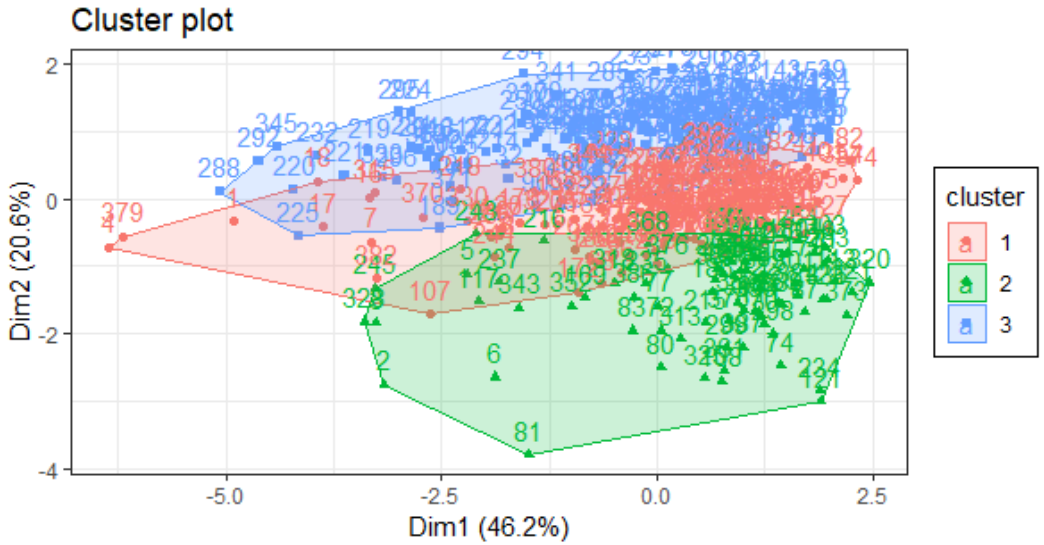


대체적으로 고르게 잘 묶인 모습!

03 클러스터링 : K-medoids

2 K-Medoids Clustering

실루엣 계수가 K-means clustering의 것보다 작음



K-means

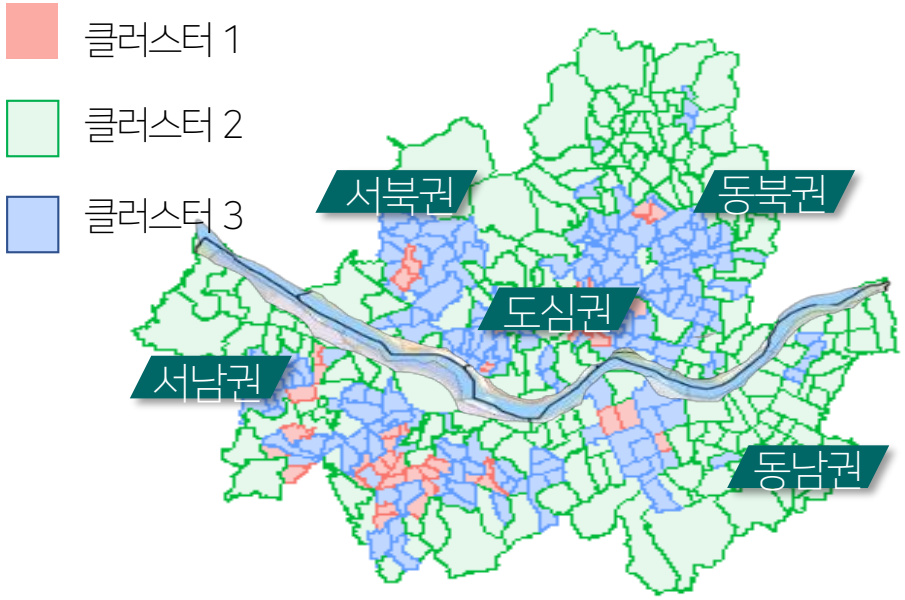
```
> sil_num
[1] 0.5902643
```

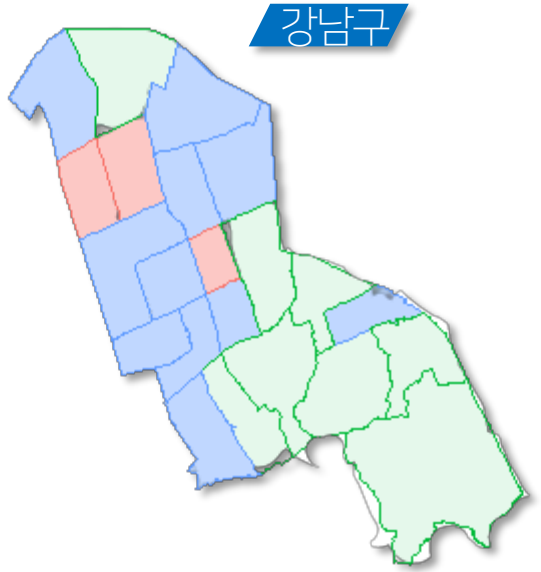
∨

```
> sil_num
[1] 0.5754666
```

sil무룩..

03 최종 모델 선택 및 결과 해석 : K-MEANS

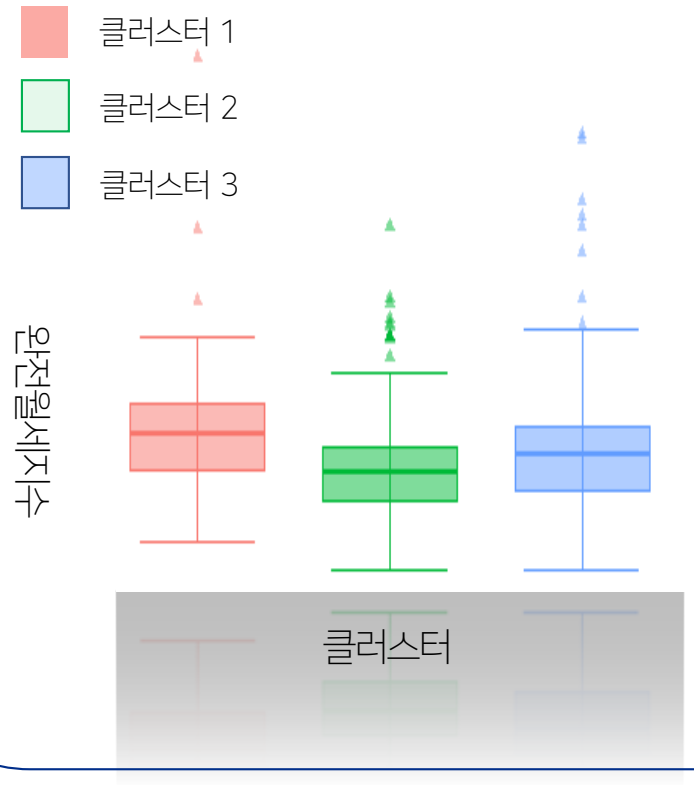


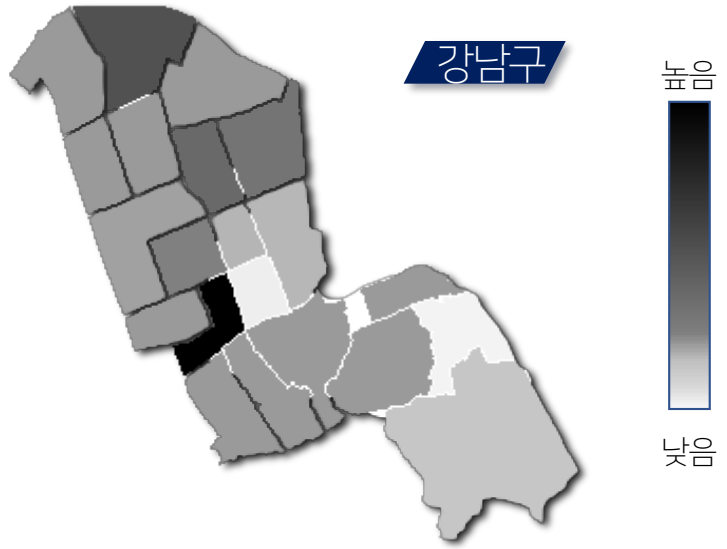


- 클러스터 2는 도시 외곽 지역,
  클러스터 1,3 은 주로 도시 중심에 분포
- 북쪽과 동남권의 중심은 클러스터 1,3
  서남권의 중심에는 클러스터 2가 주로 분포
- 도시 중심과 가까운 곳에 클러스터 1,3

03 최종 모델 선택 및 결과 해석 : K-MEANS

내부적 요인(1) : 시세

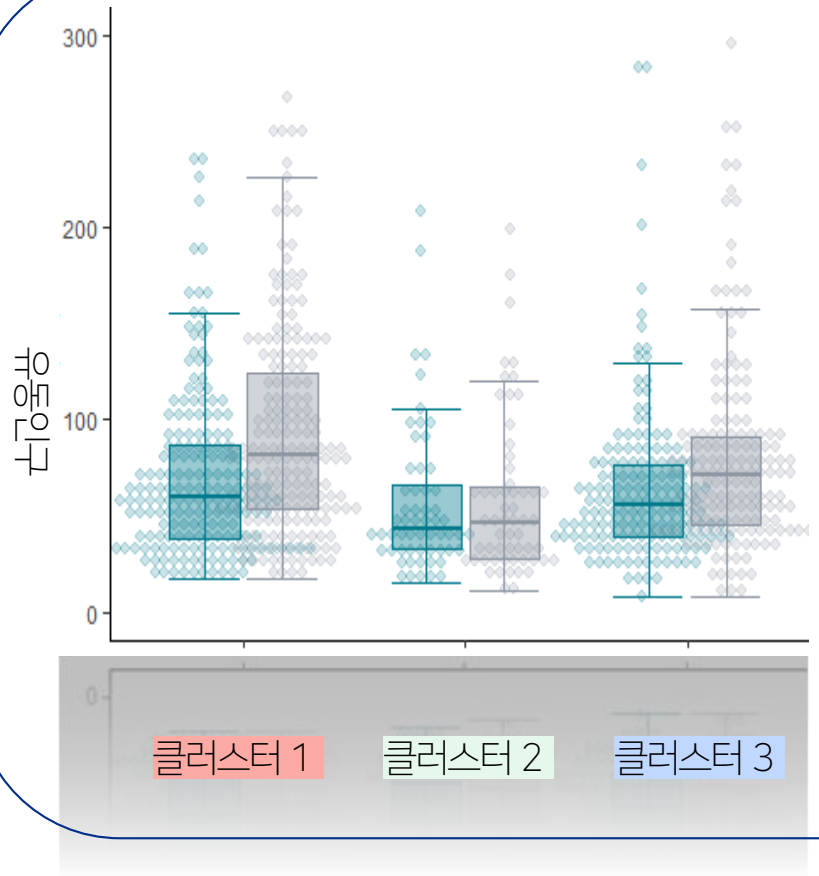


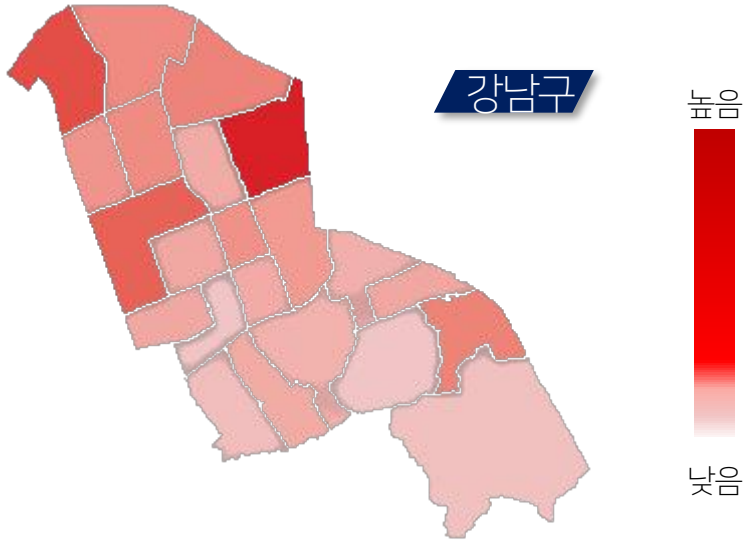


- 클러스터 3 =1→2순으로 높은 시세 형성
- 중심부와 가까울수록 시세가 높은 경향

03 최종 모델 선택 및 결과 해석 : K-MEANS

내부적 요인(2) : 라이프스타일

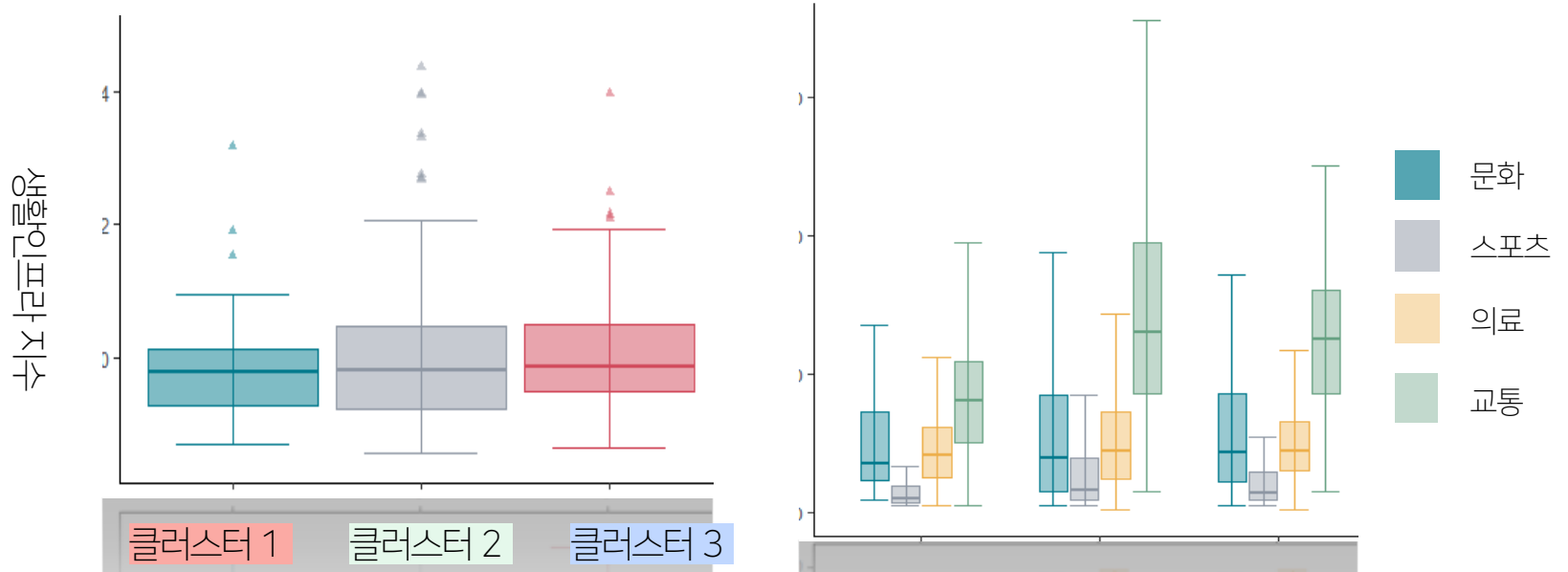




- 클러스터 1 →3→2순으로 유동인구 많음
- 주민등록인구 대비 유동인구 평균과 분산은 비례
  유동인구 절대량이 많을수록 분산도 크다.
  즉, 밤에도 유동인구가 많다!

03 최종 모델 선택 및 결과 해석 : K-MEANS

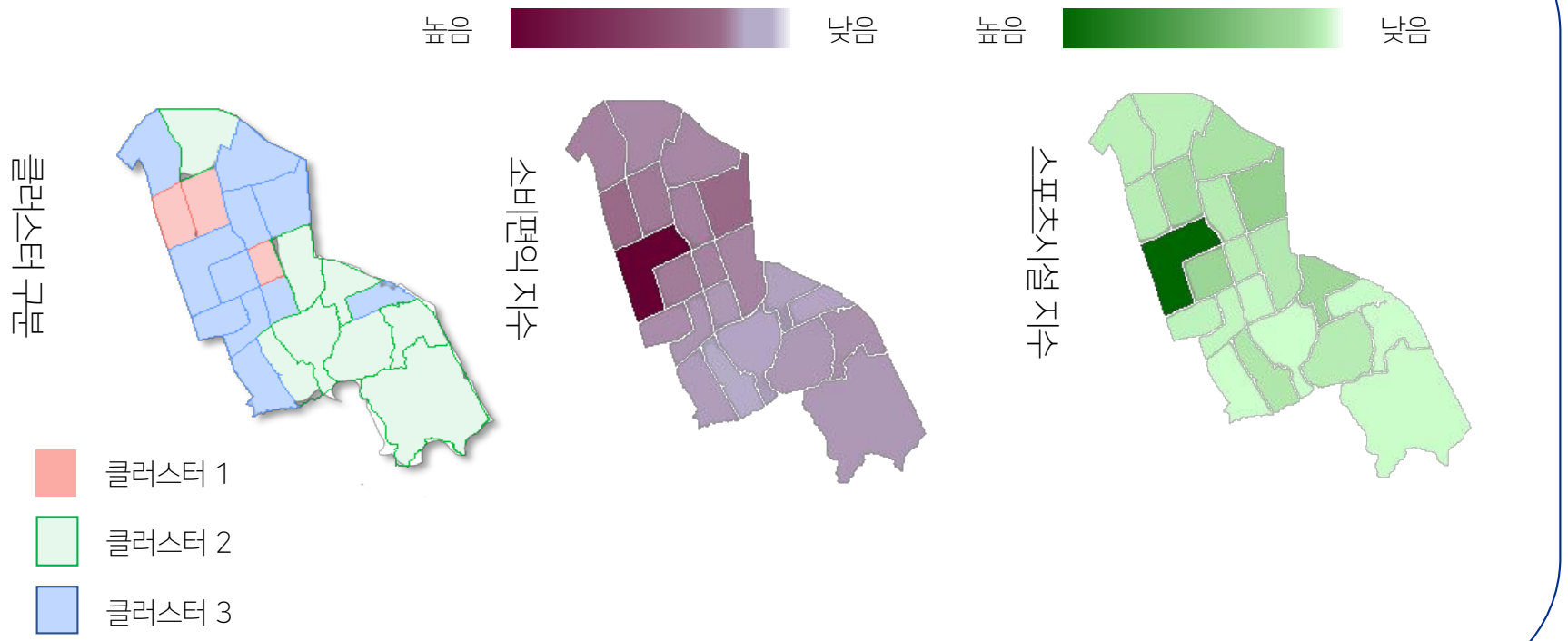
외부적 요인(1) : 생활인프라 지수



- 클러스터 2→3 →1 순으로 생활인프라지수가 높음
- 클러스터 2: 도시 외곽의 생활인프라가 좋은 주거 지역

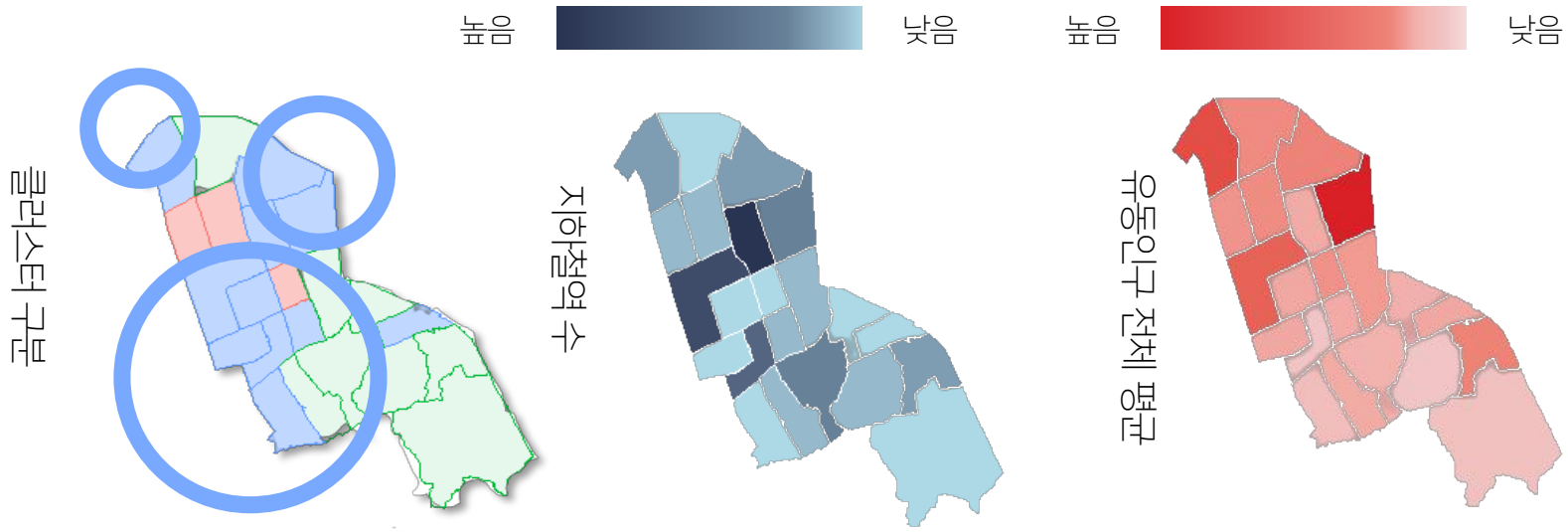
03 최종 모델 선택 및 결과 해석 : K-MEANS

외부적 요인(1) : 생활인프라 지수 – 소비, 스포츠



03 최종 모델 선택 및 결과 해석 : K-MEANS

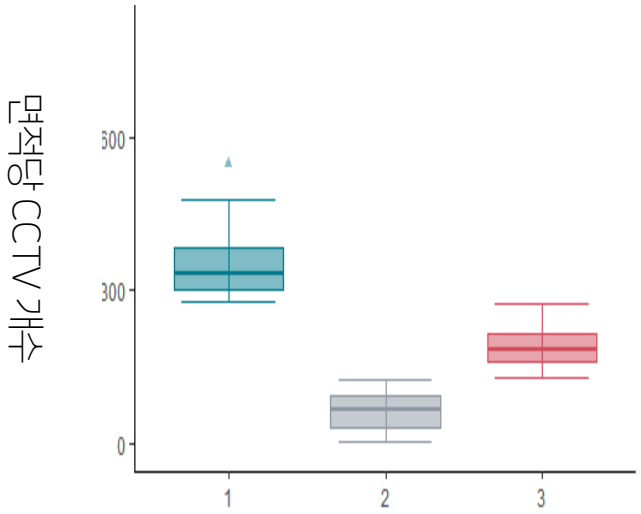
외부적 요인(1) : 생활인프라 지수 – 교통

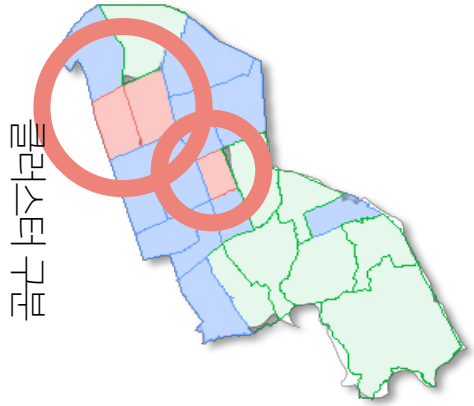


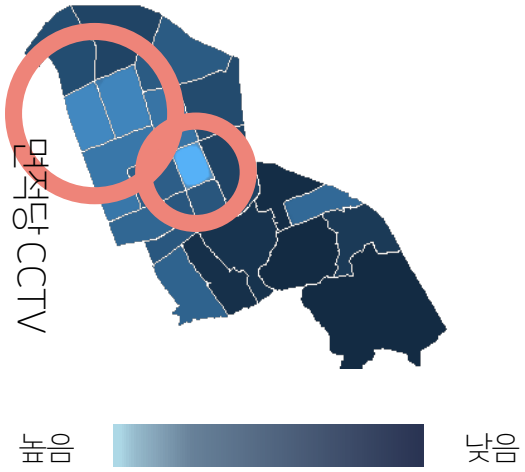
클러스터 3: 주민등록인구 대비 유동인구가 많고,
지하철역, 상가와 스포츠시설 등이 모여 있는 번화가, 교통 요충지

03 최종 모델 선택 및 결과 해석 : K-MEANS

외부적 요인(2) : 치안



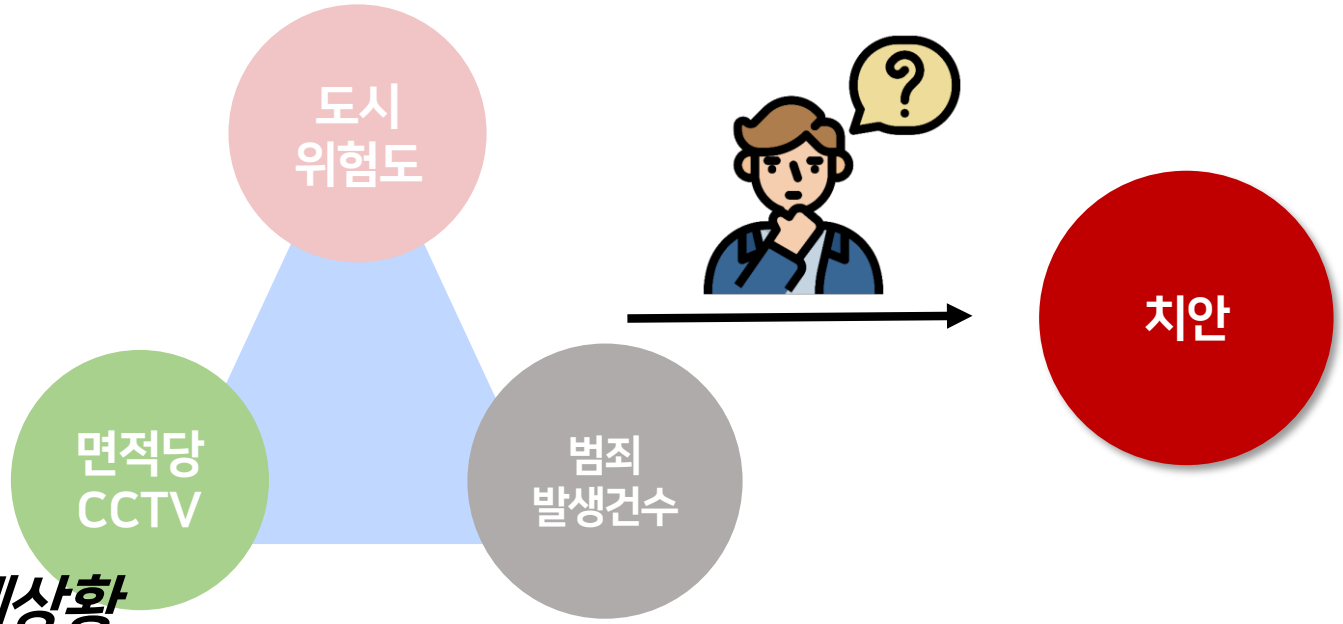




클러스터 1: 치안이 좋고, 교통 요충지와 조금 떨어진 주거 지역

03 최종 모델 선택 및 결과 해석 : K-MEANS

잠깐 ! 면적당 CCTV 개수와 치안의 관계는?

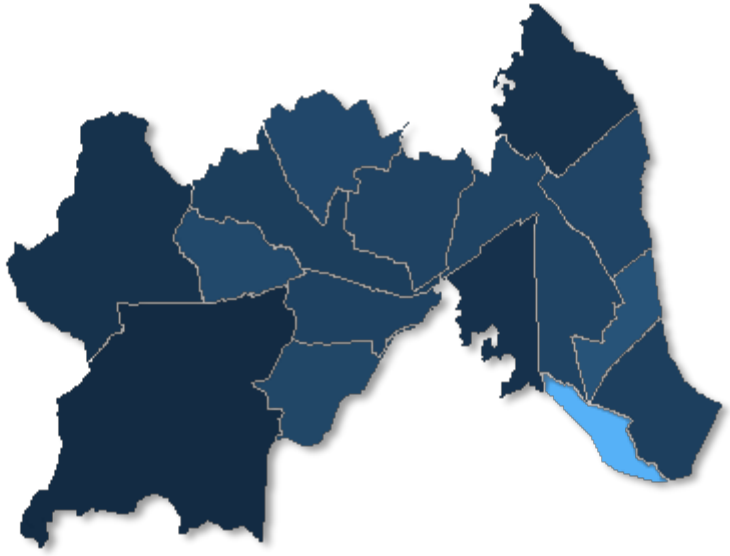


문제상황

1. 면적당 CCTV, 범죄발생건수, 도시위험도 간의 관계가 유의하지 않음
2. 치안을 직접적으로 설명할 수 있는 도시위험도, 범죄발생건수는 자치구별 데이터로, 클러스터링에 사용되지 않음

03 최종 모델 선택 및 결과 해석 : K-MEANS

잠깐 ! 면적당 CCTV 개수와 치안의 관계는?

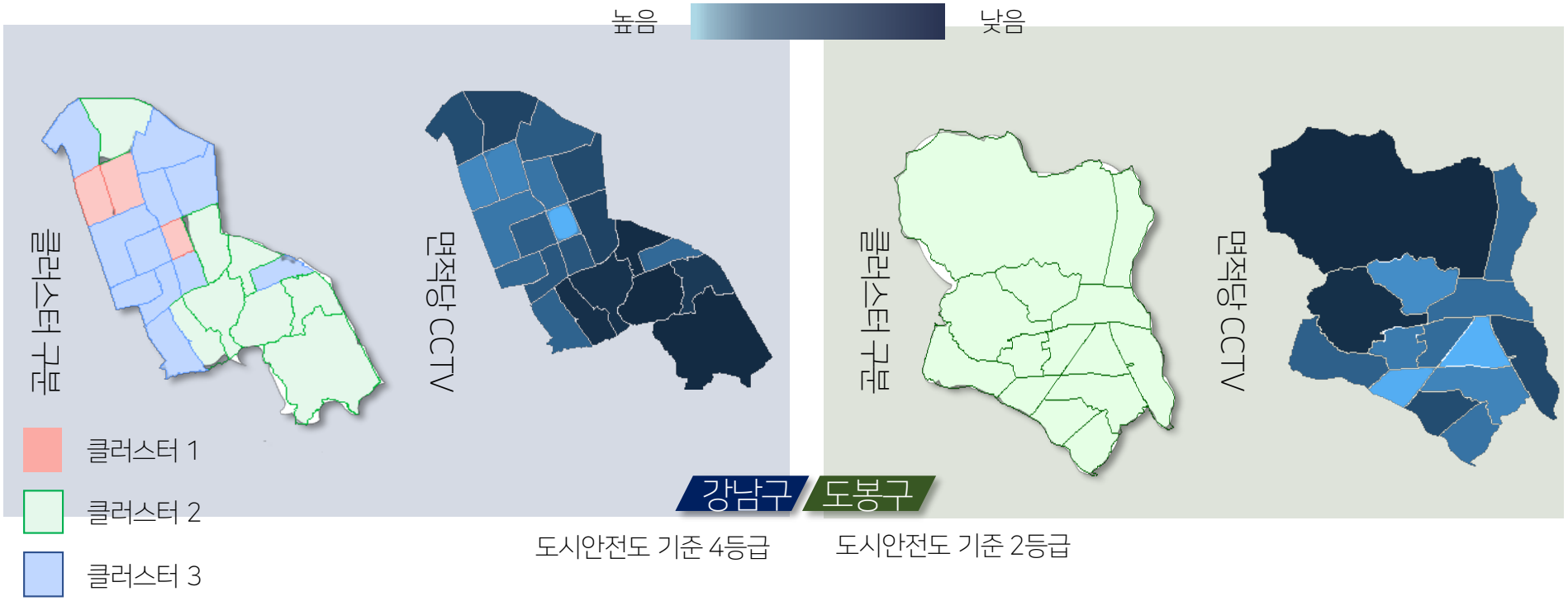


문제상황

3. 해당 데이터는 정부에서 생활방범, 교통단속, 어린이보호 등을
목적으로 설치한 공공CCTV 데이터
→사설 CCTV는 반영되지 않음

03 최종 모델 선택 및 결과 해석 : K-MEANS

잠깐 ! 면적당 CCTV 개수와 치안의 관계는?



실제 치안에 대한 지표인 도시안전도가 낮은 강남구/ 도봉구 비교

→ 도시 외곽쪽으로 갈수록 CCTV 밀도가 낮아지는 동일한 추세!

03 최종 모델 선택 및 결과 해석 : K-MEANS

요약

	변수 구분	클러스터 1	클러스터 2	클러스터 3
내부적 요소	시세	높음	상대적으로 낮은편	높음
	유동인구 평균 및 분산	가장 높음	상대적으로 낮은편	높음
외부적 요소	생활인프라	상대적으로 낮은편	가장 높음	높음
	면적당 CCTV	가장 높음	낮음	낮은 편
	종합	교통 요충지 주변의 주거 지역	도시 외곽의 주거지역	교통 요충지 및 번화가

04 추천알고리즘 함수 구현

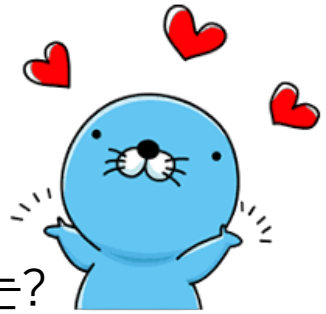
실제 구현

화려한 조명이 날 감싸네님의 취향 고려!

- 월세액은 **60만원에서 40만원** 사이로!
- 보증금은 **5천만원에서 1천만원** 사이로!
- 오피스텔 평수는 **25평에서 15평** 사이로!

04 추천알고리즘 함수 구현

실제 구현



화려한 조명이 날 감싸네님을 위한 데마팀의 추천 결과는?

<서울특별시 마포구 합정동 월드오피스텔>

- ✓ 평수 20.52평
- ✓ 보증금 5천만원
- ✓ 월세 45만원