

서울시 아동보호전문기관 우선 입지 선정



Contents



01

1주차 피드백

02

클러스터링

03

입지 선정

04

한계와 의의



01

1주차 피드백

1) 데이터 보완

2) 최종 데이터

01. 데이터 보완



- 이혼 가정 데이터

1주차 피드백

이혼 후, 거주지를 이동하는 가정이 많은데
이 부분을 고려해야하지 않을까요?



이혼 가정의 거주지 움직임을 반영하는 데이터를 찾지 못해
이혼가정 데이터를 **삭제하기로 결정!**



01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

- 소득 수준 데이터

1주차 피드백

소득 수준과 관련된 데이터가
추가되었으면 좋겠어요!

행정동별 **소득 수준 데이터**를 위해 희망을 갖고
빅데이터 캠퍼스 방문!



01.
데이터 보완

최종 데이터

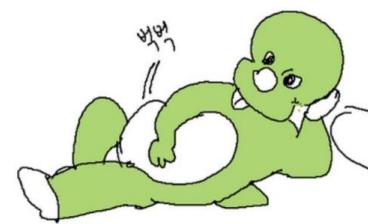
02.
클러스터링

03.
입지 선정

04.
한계와 의의

- 서울시 행정동별 소득수준 통계.csv

기준연도	행정동코드	구분	평균소득(만원)	소득금액구간	구간별 인원수 비중
2018	11110515	전체	9999	2천만원 미만	100
2018	11110515	전체	9999	3천만원 미만	100
2018	11110515	전체	9999	4천만원 미만	100
2018	11110515	전체	9999	5천만원 미만	100
2018	11110515	전체	9999	5천만원 이상	100



빅데이터 캠퍼스

“원본데이터로 역변환 할 수 없게 알아서 통계처리 하라고”



01. 데이터 보완



- 서울시 행정동별 소득수준 통계.csv

기준연도	행정동코드	구분	평균소득(만원)	소득금액구간	구간별 인원수 비중
2018	11110515	전체	9999	2천만원 미만	100
2018	11110515	전체	9999	3천만원 미만	100
2018	11110515	전체	9999	4천만원 미만	100
2018	11110515	전체	9999	5천만원 미만	100
2018	11110515	전체	9999	5천만원 이상	100



1. 구분이 '전체' 인 행만 필터링



2. 소득금액구간에 대해 Spread한 후,
3. 행의 합으로 나누어 구간별 인원 비율 구함

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 데이터 보완



- 서울시 행정동별 소득수준 통계.csv

행정동코드	2천만원 미만	3천만원 미만	4천만원 미만	5천만원 미만	5천만원 이상
11110515	0.44	0.09	0.15	0.14	0.18
11110520	0.25	0.05	0.11	0.13	0.46
11215710	0.09	0.31	0.43	0.11	0.05
11260590	0.08	0.31	0.42	0.14	0.06
11650540	0.24	0.06	0.11	0.13	0.47

경제요인 중 저소득비율 뿐만 아니라, 고소득비율도 고려하기 위해
행정동별 '5천만원이상 비율' 변수 추가

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 데이터 보완



- 장애아동, 치안기관 수 데이터

1주차 피드백

변수를 만들 때

행정동별 인구나 면적에 대한 고려가 필요하지 않을까요?

$$\frac{\text{장애아동 수}}{\text{아동 인구 수}}$$

$$\frac{\text{치안기관 수}}{\text{행정동 인구}}$$

01.
데이터 보완

최종 데이터

02.
클러스터링

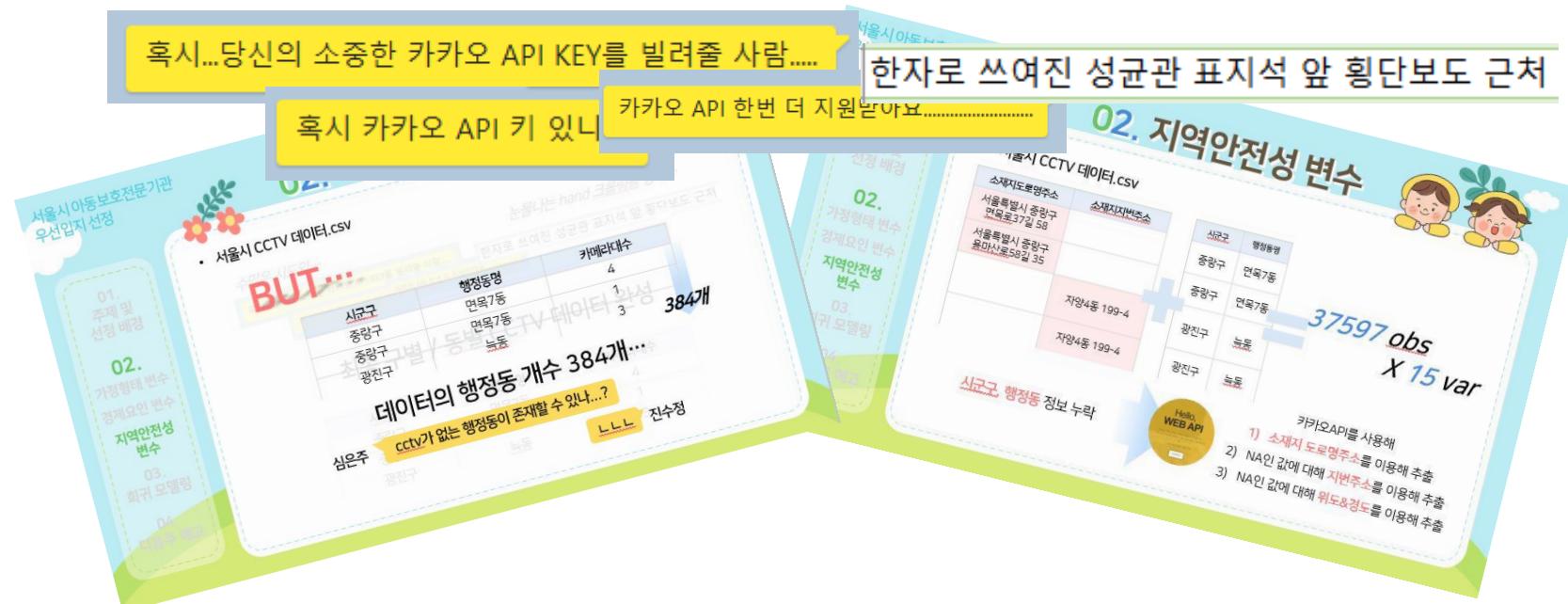
03.
입지 선정

04.
한계와 의의

01. 데이터 보완



- CCTV 변수



실패로 끝났던, 행정동 CCTV 데이터

01. 데이터 보완



- CCTV 변수

수많은 시도와...

눈물나는 hand 크롤링을 통해...

구	API 키	행정동명	카메라대수	횡단보도 근처
시군구	카카오 API 키	면목7동	4	
중랑구	있니	면목7동	1	
중랑구		면목7동	3	
광진구		늑동		

384개

나머지 행정동의 CCTV 수를 채우기 위해...

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 데이터 보완

- CCTV 변수

서울특별시_강동구_CCTV현황.csv
876 obs x 13 variables

서울특별시_도봉구_CCTV현황.csv
834 obs x 13 variables

서울특별시_동대문구_CCTV현황.csv
615 obs x 10 variables

누락된 행정동의 CCTV 데이터를
따로 가져와 채워 줌



카카오API를 사용해

- 1) 소재지 도로명주소를 이용해 추출
- 2) NA인 값에 대해 지번주소를 이용해 추출
- 3) NA인 값에 대해 위도&경도를 이용해 추출

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
환경과 의의

01. 데이터 보완



- CCTV 변수

CCTV 변수 완성...!

행정구	행정동	CCTV수	면적	CCTV비율
강남구	개포1동	15	1.27	11.81102
강남구	개포2동	57	2.51	22.70916
강남구	개포4동	105	1.49	70.4698
강남구	논현1동	130	1.25	104
강남구	논현2동	133	1.47	90.47619

⋮

425obs x 5 variables

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 데이터 보완



- 위험도 지수 산출 데이터셋

행정동	장애인아동 비율	부동산 단위가격	지구대파출소 비율	5천만원이상 비율	CCTV 비율	아동 수	기초수급가구 비율
사직동	0.0377	1016.3605	0.0305	0.2704	43.0894	1139	0.0135
삼청동	0.0154	0	0.0343	0.1468	14.7651	259	0.0155
사직동	0.0377	1016.3605	0.0305	0.2704	43.0894	1139	0.0135

⋮

425obs x 8 variables

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 위험도 지수 산출

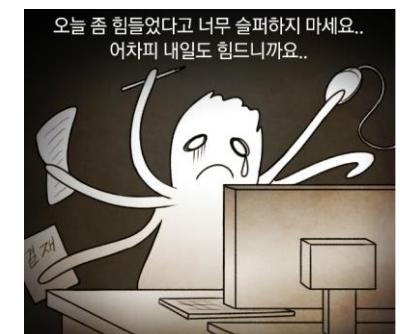
1주차 피드백

위험도 지수를 산출할 때

회귀 모델링 이외의 다른 방법도 사용해보는 건 어떤가요?



상관계수분석, 구조방정식, PCA, FA 등
다양한 방법 도전!



01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

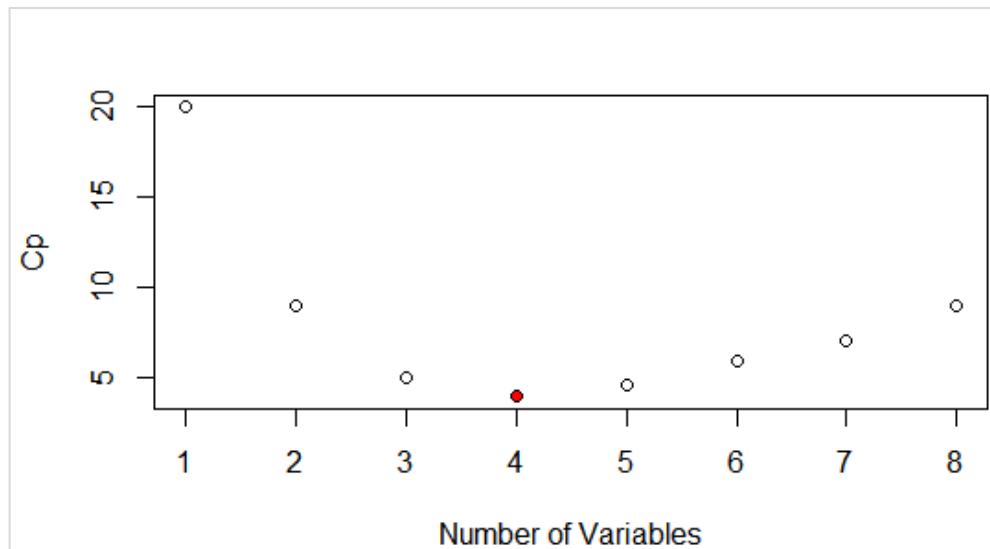
01. 최종 데이터



- Best Subset Selection

유의한 변수만 추출하기 위해 Best Subset Selection 진행

가능한 모든 변수들의 조합을 다 고려하는 방법



Mallows's Cp를 기준으로
최적의 변수 조합 선택

5천만원이상비율, 아동 수,
기초수급가구비율, CCTV 비율

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- Best Subset Selection

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

변수
5천만원 이상 비율
부동산 단위가격
기초수급 가구 비율
아동 인구
면적당 CCTV 비율
지구대 파출소 비율

경제적
요인

가정형태

지역
안전성

선택된 변수
5천만원 이상 비율
기초수급가구비율
아동 인구
면적당 CCTV 비율

위험도와 관련이 있는 각 요인을
대표할 수 있는 변수들로 잘 선택됨!

01. 최종 데이터



• 가중치 산출 - 상관분석

피어슨 상관계수를 이용해 변수의 가중치를 계산

$$\text{각 변수의 가중치} = \frac{\text{각 변수의 상관계수}}{\text{모든 상관계수 절댓값의 합}}$$

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

	5천만원이상 비율	CCTV 비율	아동 수	기초수급가구 비율
가중치	0.1071428	-0.2602040	0.1836734	-0.2397959

01. 최종 데이터



• 가중치 산출 - 상관분석

피어슨 상관계수를 이용해 변수의 가중치를 계산

각 변수의 상관계수는
각 변수의 상관계수

상관계수는 변수들 간의 상관관계를
표현하는 수치로, 상관계수는 상관관계가 있는 경우에만 계산된다.

개략적으로 파악하도록 도와주는 수치로,

가중치를 만들어 사용하기엔 적절하지 않음

	5천만원이상 비율	CCTV 비율	아동 수	기초수급가구 비율
가중치	0.1071428	-0.2602040	0.1836734	-0.2397959

사용 X

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



• 가중치 산출 - 경로분석

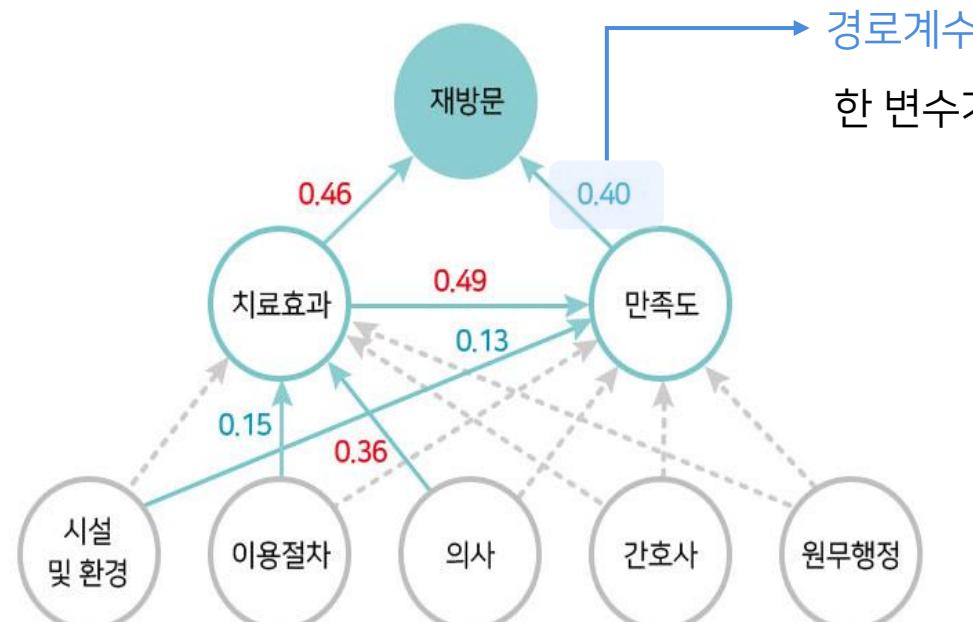
인과관계를 가진 여러 관측변수들 간의 선형관계를 **분해**하고, 이를 **해석**하는 방법

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의



경로계수

한 변수가 다른 한 변수에 미치는 직접적인 인과효과

각 변수의 가중치

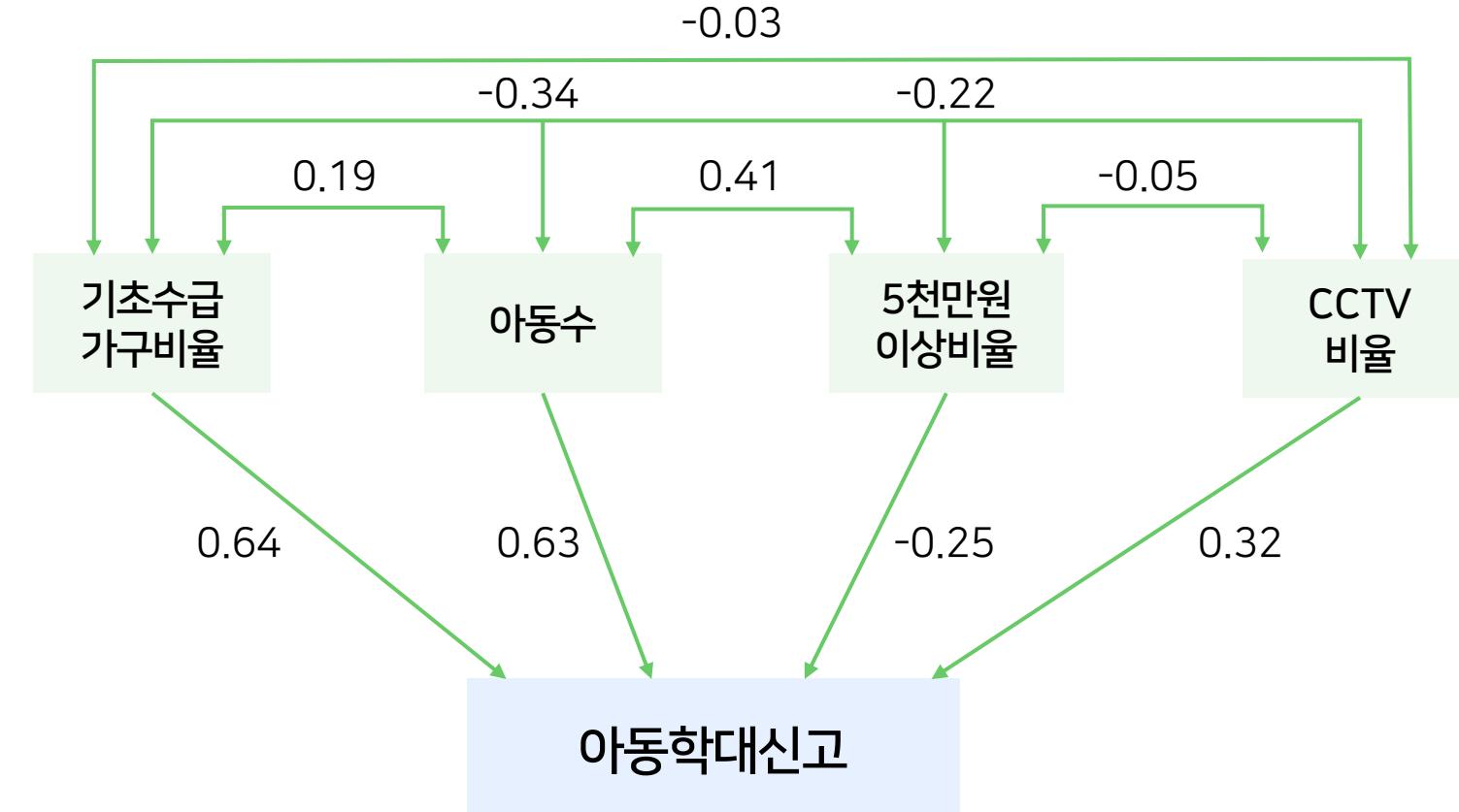
II

$$\frac{\text{각 변수의 경로계수}}{\text{모든 경로계수 절댓값의 합}}$$

01. 최종 데이터



- 가중치 산출 - 경로분석



01.
데이터 보완
최종 데이터

02.
클러스터링

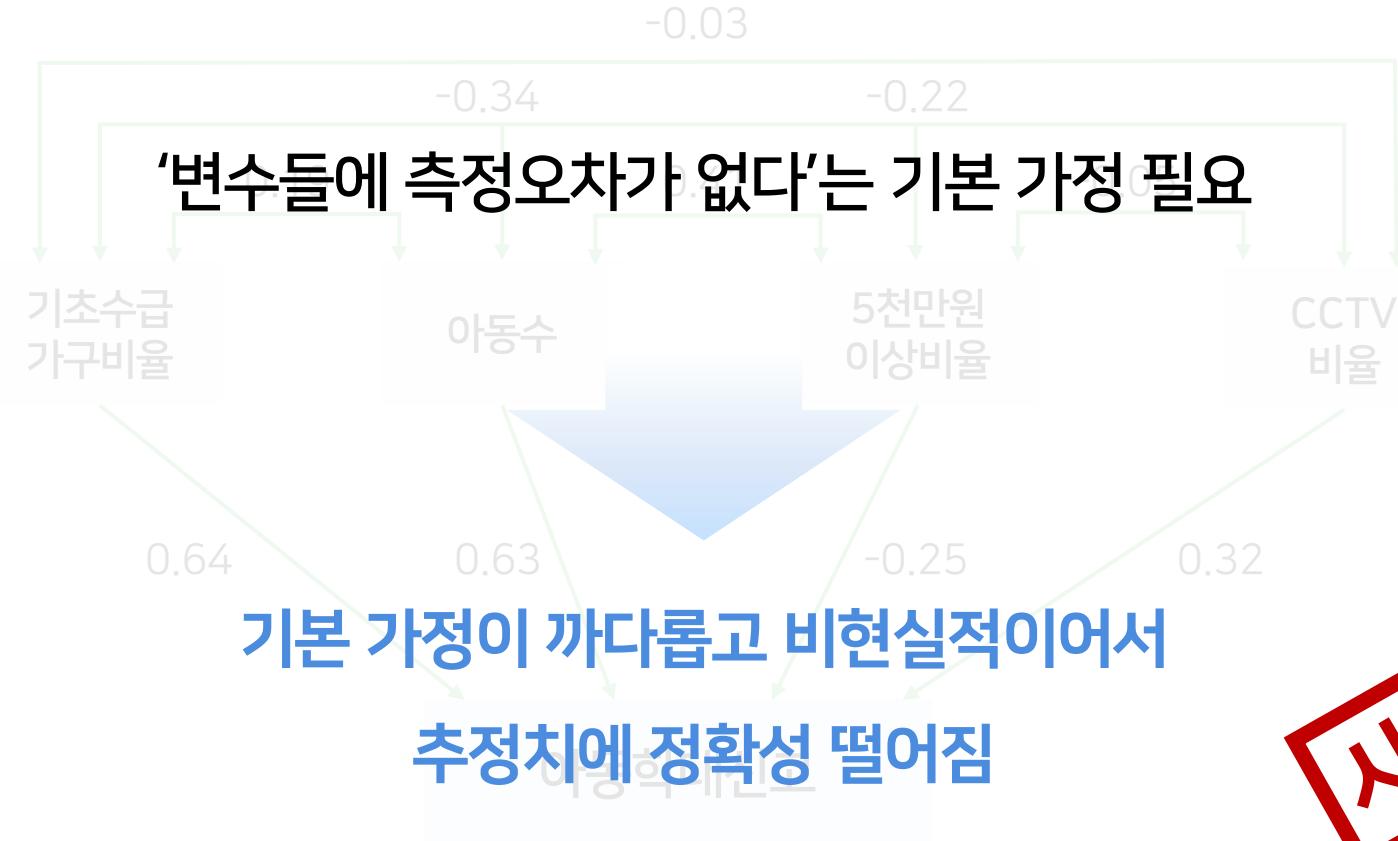
03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 가중치 산출 - 경로분석



01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

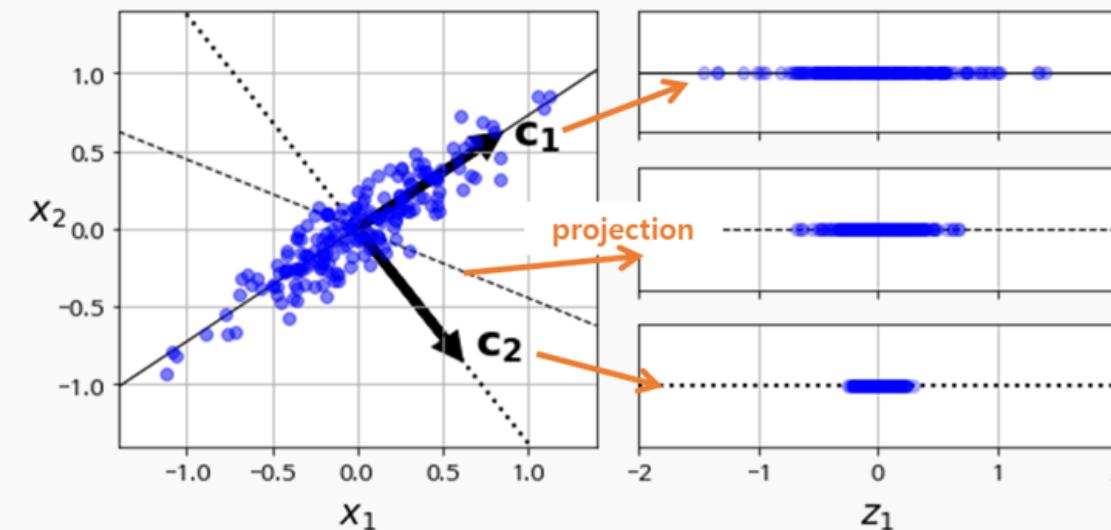
04.
한계와 의의

01. 최종 데이터



- 가중치 산출 - 주성분 분석

데이터의 분산이 **최대**가 되는 **초평면에 투영**시키는 방법



즉, 원본의 데이터 셋과 투영된 데이터 셋 간의 평균제곱거리를 **최소화** 하는 방법

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 가중치 산출 - 주성분 분석

데이터의 분산이 **최대**가 되는 **초평면에 투영**시키는 방법

1. 공분산 행렬

x1	x2	x3
2	1	3
8	4	7
1	4	4

- 변수의 **절대적 수치**가 중요할 때 사용

2. 상관행렬

나이	키	무게
13	143.4	40.5
25	178.0	72.3
32	169.5	57.2

- 변수의 **단위**가 다를 때 사용
- 변수의 **상대적인 변화**에 집중할 때 사용

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 가중치 산출 - 주성분 분석

데이터의 분산이 **최대**가 되는 **초평면에 투영**시키는 방법

1. 공분산 행렬

x1	x2	x3
2	1	3
8	4	7
1	4	4

- 변수의 **절대적 수치**가 중요할 때 사용

2. 상관행렬

나이	키	무게
13	143.4	40.5
25	178.0	72.3
32	169.5	57.2

- 변수의 **단위**가 다를 때 사용
- 변수의 **상대적인 변화**에 집중할 때 사용

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



• 가중치 산출 - 주성분 분석

데이터의 분산이 **최대**가 되는 **초평면에 투영**시키는 방법

	PC1	PC2
5천만원이상비율	-0.2755071	-0.22457696
CCTV 비율	0.1648474	-0.17644183
기초수급가구비율	0.2585080	0.14367463
아동 수	-0.3011375	-0.63094691

가중치 산출

변수	가중치
5천만원이상비율	-0.2755071
CCTV 비율	0.1648474
기초수급가구비율	0.2585080
아동 수	-0.3011375

한 개의 PC를 이루는 각 변수의 영향력을 가중치로 둠

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 가중치 산출 - 주성분 분석

데이터의 분산이 최대가 되는 초평면에 투영시키는 방법

PC1이 설명하는 분산 : 0.404

	PC1	PC2	변수	가중치
5천만원이상비율	-0.2755071	-0.22457696	5천만원이상비율	-0.2755071
CCTV 비율	0.1648474	-0.17644183	CCTV 비율	0.1648474
기초수급가구비율	0.2585080	0.14367463	기초수급가구비율	0.2585080
아동 수	-0.3011375	-0.63094691	아동 수	-0.3011375

4개의 변수를 충분히 설명하지 못한다고 판단

한 개의 PC를 이루는 각 변수의 영향력을 가중치로 둠

사용 X

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 가중치 산출 - 요인 분석

다수의 변수들에 잠재되어 있는 **공통인자**를 찾아내는 방법

기존 변수

변수
수학
과학
영어
중국어

Factor Analysis

수리적
능력

변수
수학
과학

외국어
능력

변수
영어
중국어

본래의 변수보다 더 적절한 **변수**를 생성할 수 있음

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

- 가중치 산출 - 요인 분석

5천만원이상 비율

기초수급가구 비율

아동 인구

CCTV 비율



*Factor
Analysis*

“ Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 9.23 on 2 degrees of freedom.
The p-value is 0.00992 ”

요인의 개수는 1개면 충분하며, 요인 분석 결과는 통계적으로 유의하다!

01. 최종 데이터

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

- 가중치 산출 - 요인 분석

5천만원이상 비율

기초수급가구 비율

아동 인구

면적당 CCTV 비율

*Factor
Analysis*

행정동	위험도지수
사직동	9.499004781
아현동	7.923649984
소공동	9.096211928
무악동	9.518269197

첫 번째 Factor의 계수를 이용하여
위험도지수를 산출



01. 최종 데이터



- 대체기관 변수의 문제점

1.

유치원, 학교 등이 포함되어 '아동 수' 변수와 상관관계가 매우 큼

행정동	대체기관 수
용답동	11
왕십리도선동	28
금호2.3가동	20
옥수동	25



행정동	대체기관 수
용답동	1
왕십리도선동	2
금호2.3가동	2
옥수동	1

어린이집, 유치원, 학교 제외

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터



- 대체기관 변수의 문제점

2.

행정동 별 크기 차이를 고려하지 못함

01.
데이터 보완
최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

대체기관 비율

II

행정동 별 대체기관 수
행정동 면적

면적 대비 대체기관 수 고려



행정동	대체기관 비율
용답동	0.4310
왕십리도선동	2.7777
금호2.3가동	3.125
옥수동	0.5128

01. 최종 데이터



- 최종 데이터

행정구	행정동	위험도 지수	대체기관 수 (학교 제외/면적)
종로구	사직동	9.499005	0
종로구	삼청동	9.84396	0
종로구	부암동	9.730225	1.321586
종로구	평창동	9.971461	0
종로구	무악동	9.518269	2.777778

⋮
425obs x 4 variables

“ 클러스터링을 위한 최종 데이터 완성 ”

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

01. 최종 데이터

01.
데이터 보완

최종 데이터

02.
클러스터링

03.
입지 선정

04.
한계와 의의

- 전체적인 플로우

데이터 수집 및 전처리

자치구, 행정동별 데이터

- 아동/장애 아동
- 소득수준
- 기초생활보장 수급자
- CCTV / 경찰서파출소
- 이혼 가정
- 부동산 가격

Factor Analysis

위험도 지수 산출

+ 대체기관 수

클러스터링

- K Means
- K Medoids
- DBSCAN
- HDBSCAN
- 계층적 클러스터링
- GMM clustering

Target Cluster

우선입지 선정

1 LSCP

설치할 아동보호전문기관 개수 확정

2 반경 내 수요 최대화 & P-Median

우선 입지 행정동 선정

3 최종 우선 입지 위치

입지선정 기준에 충족하는 건물 선택



02

클러스터링

- 1) 이상치 제거
- 2) 사용 기법
- 3) 결과 해석

02. 이상치 제거



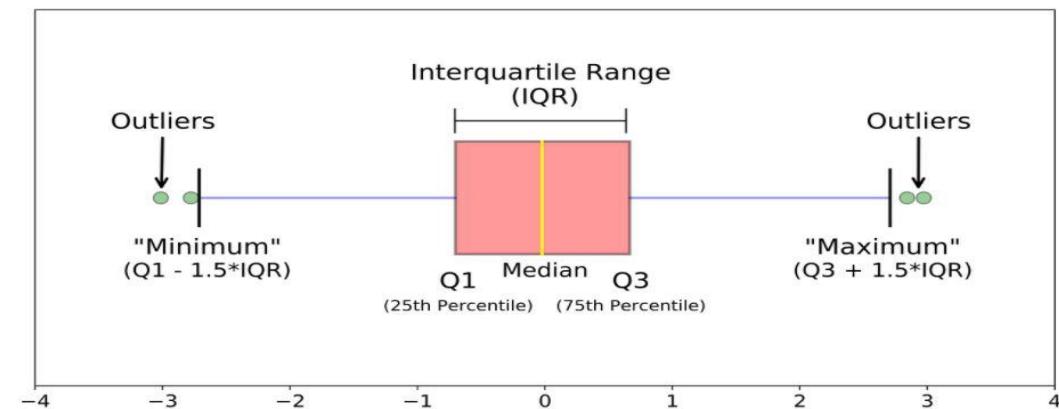
01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

- IQR 이상치 제거 방식



- $IQR = Q3 - Q1$
- $\text{Max} = Q3 + 1.5 \times IQR$
- $\text{Min} = Q1 - 1.5 \times IQR$

Max보다 크거나 Min보다 작으면 이상치

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 이상치 제거

- IQR 이상치 제거 방식

행정구	행정동
종로구	교남동
종로구	창신2동
중구	다산동
용산구	후암동
용산구	효창동
성동구	왕십리2동
광진구	자양1동

IQR 이상치 제거 방식을 통해
위험도지수가 지나치게 낮거나,
대체기관 비율이 지나치게 높은
47개의 행정동 제거



Max=Q3+1.5*IQR Min=Q1-1.5*IQR
Max보다 크거나 Min보다 작으면 이상치

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

- 클러스터링

GOAL

위험도 지수, 대체기관 **비율**을 이용한 **클러스터링**을 진행하여

아동학대 발생 위험율이 높고, 보호기관이 적은 행정동을 파악!

K-means

K-medoids

Hierarchical

DBSCAN

HDBSCAN

SOM

GMM

02. 사용 기법



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

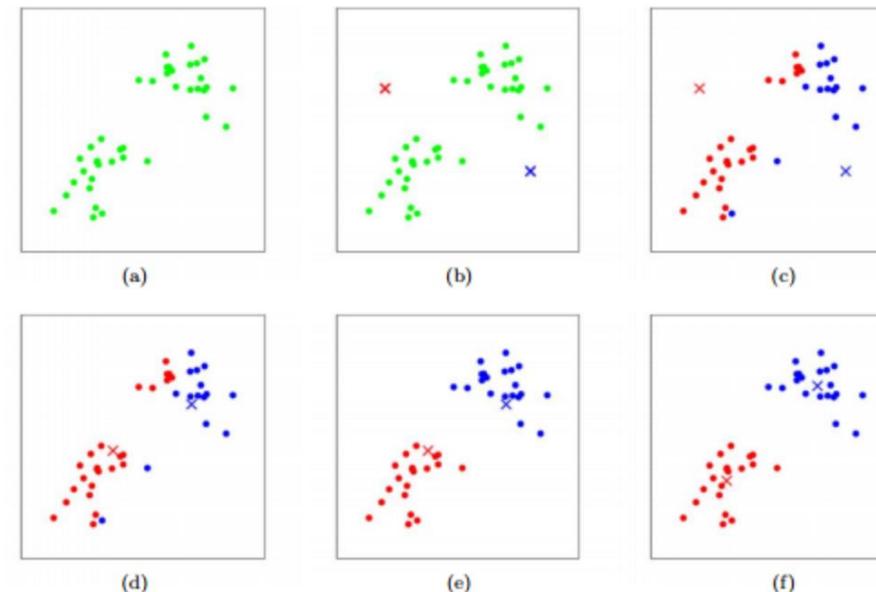
03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-means

각 데이터로부터 그 데이터가 속한 클러스터의 **중심까지의 평균을 최소화**하는 방법



1. K개의 임의의 중심점을 배치
2. 각 데이터들을 가장 가까운 중심점으로 할당
3. 군집으로 지정된 데이터들을 기반으로 해당 군집의 중심점 업데이트
4. 중심점이 수렴할 때까지 위의 2-3번 과정을 반복

01.
1주차 피드백

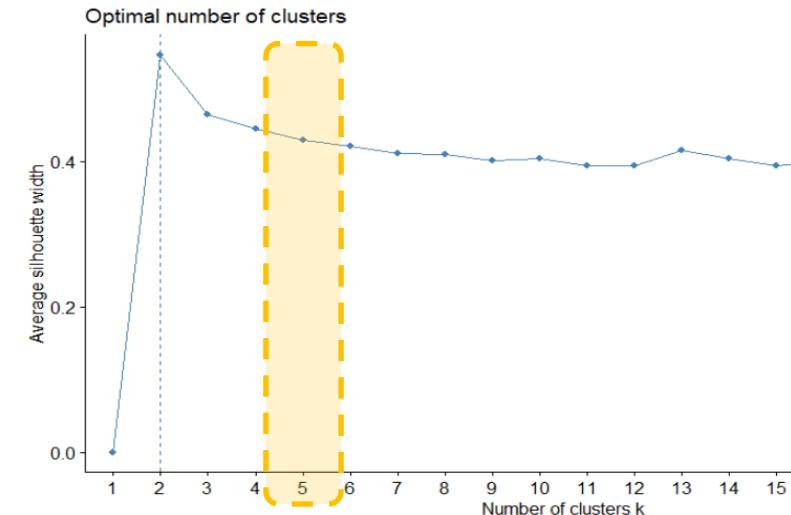
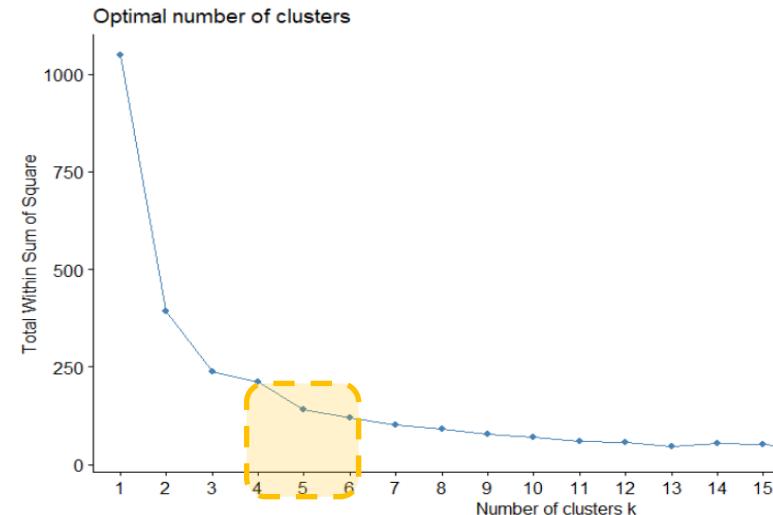
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-means



*Elbow point*와 *Silhouette*값을 고려해
클러스터 개수 5개 설정

01.
1주차 피드백

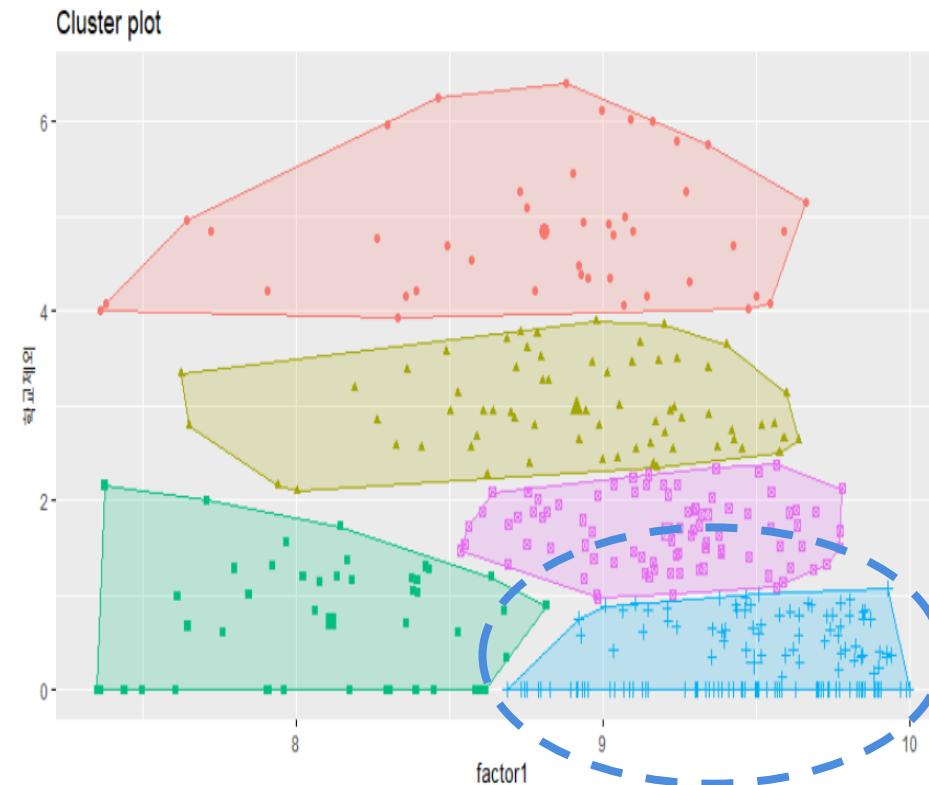
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-means



- Silhouette: 0.4299

- 클러스터별 행정동 개수

cluster	1	2	3	4	5
행정동	42	66	44	140	85

위험도지수가 높으면서, 보호기관수가 적은

행정동 140개

01.
1주차 피드백

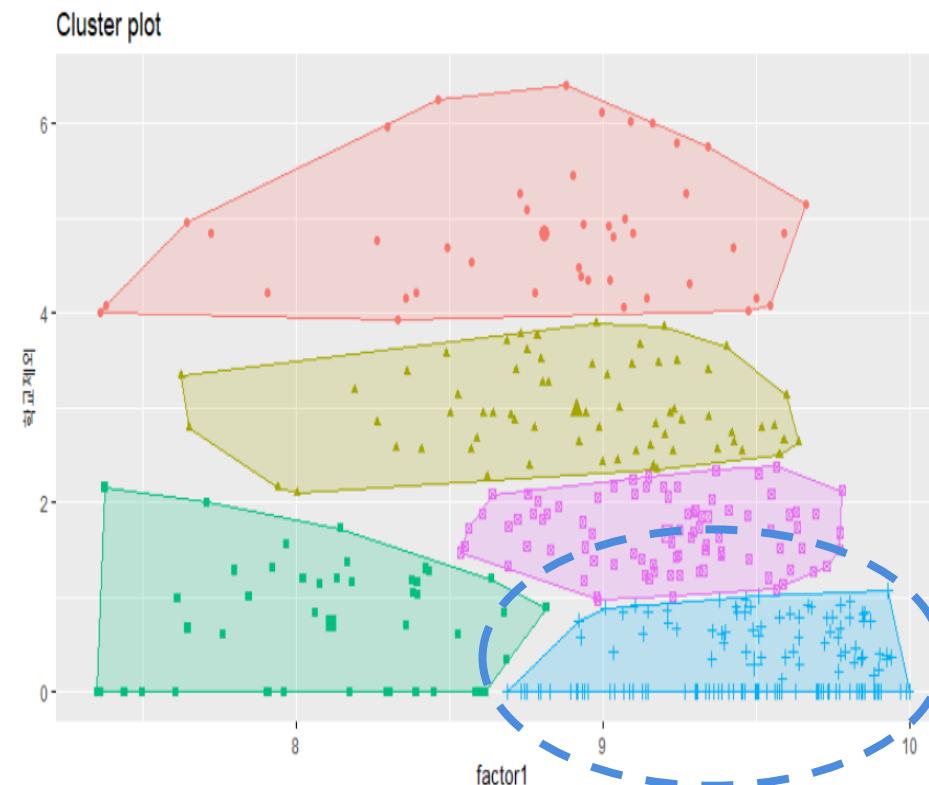
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-means



- Silhouette: 0.4299

- 클러스터별 행정동 개수

cluster	1	2	3	4	5
행정동	42	66	44	140	85

위험도지수거나 오염수 방지 기관 소재는
클러스터간 불균형이 심해 기각!
행정동 140개



01.
1주차 피드백

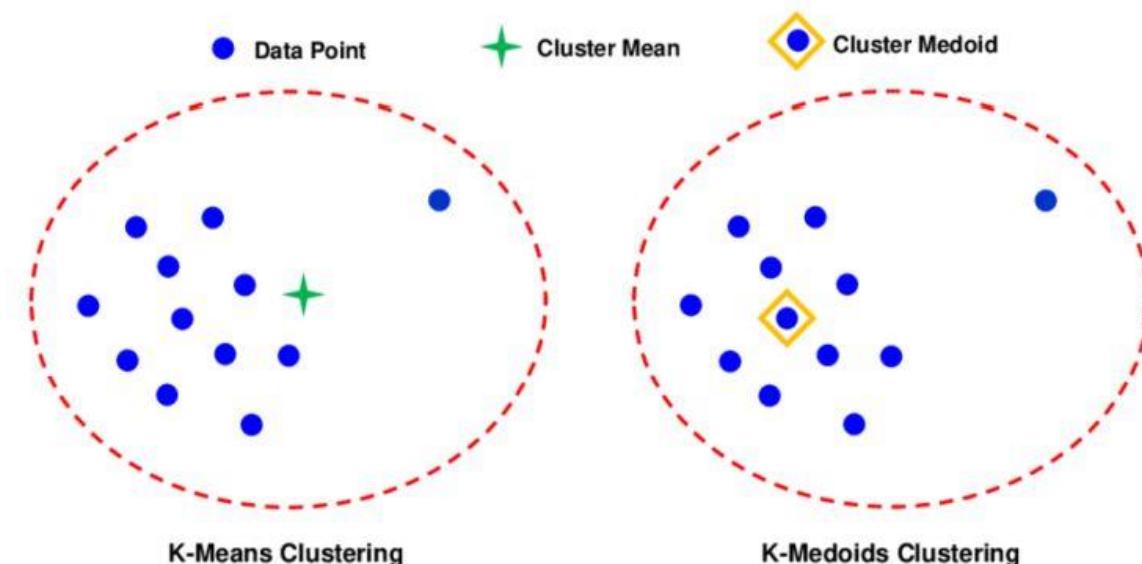
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

• K-medoids

'PAM' 기법으로도 불리며, 해당 군집의 **중앙값**을 중심점으로 보는 방법



중앙값을 취할 경우 이상치로부터 받는 영향이 적음

01.
1주차 피드백

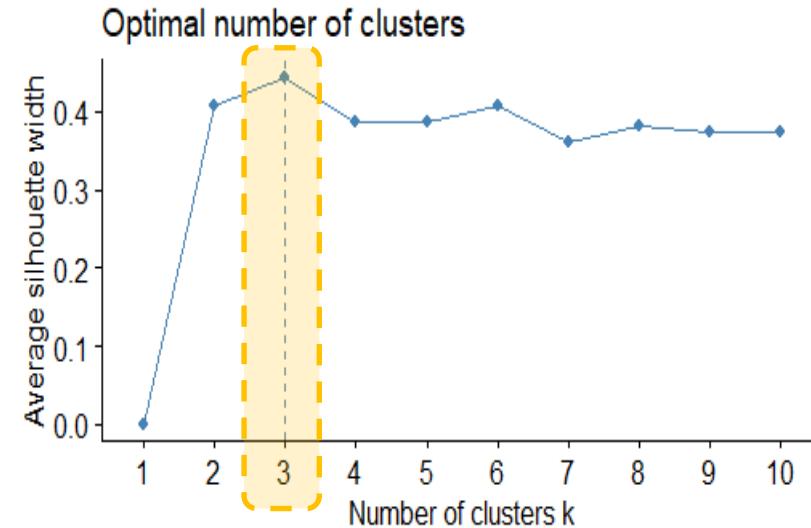
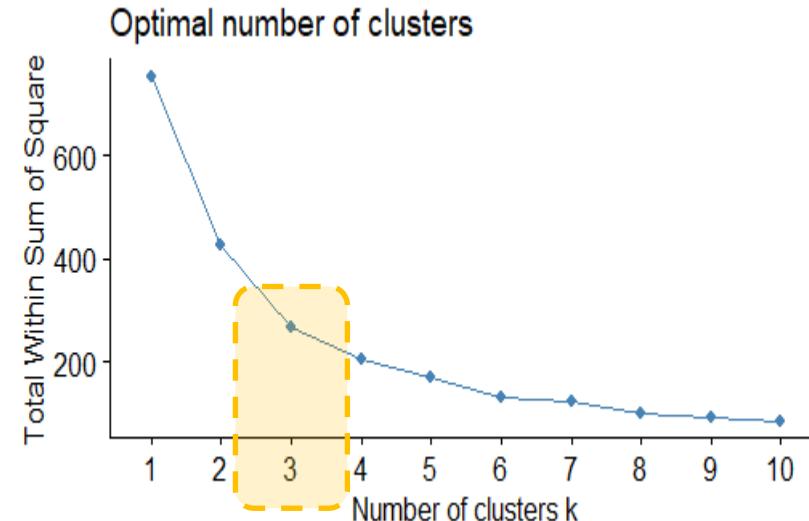
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-medoids



*Elbow point*와 *Silhouette*값을 고려해
클러스터 개수 3개 설정

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

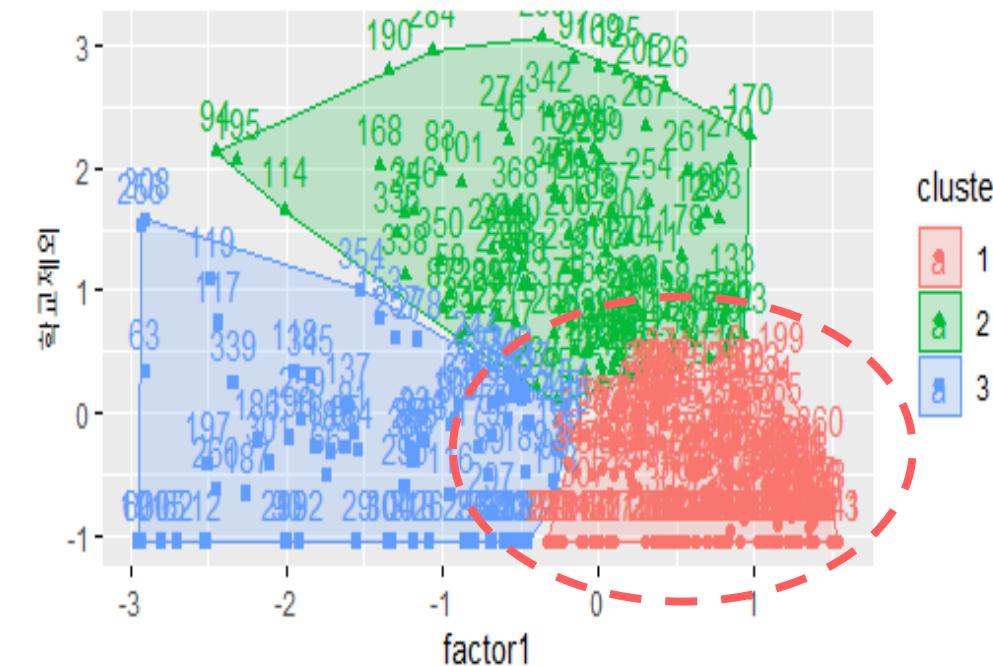
03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-medoids

Cluster plot



- Silhouette: 0.443279

- 클러스터별 행정동 개수

cluster	1	2	3
행정동	182	111	84

위험도지수가 높으면서, 보호기관수가 적은
행정동 182개

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

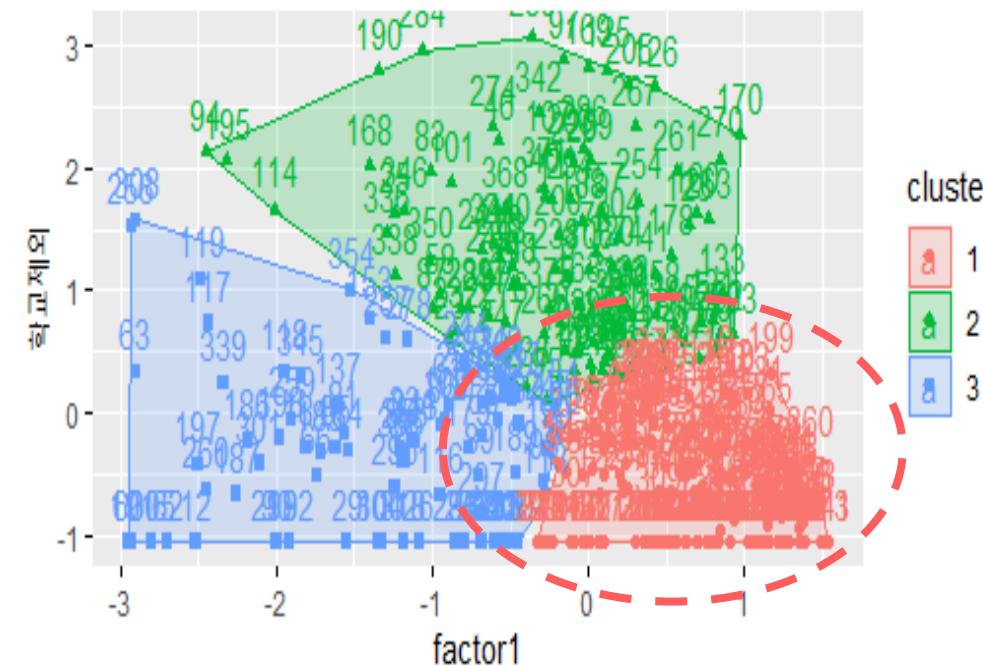
03.
입지 선정

04.
한계와 의의

02. 사용 기법

- K-medoids

Cluster plot



- Silhouette: 0.443279

- 클러스터별 행정동 개수

cluster	1	2	3
행정동	182	111	84

위험도 진실성이 0명이 생후 기관 소재은
클러스터간 불균형이 심해 기각!
행정동 182개



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

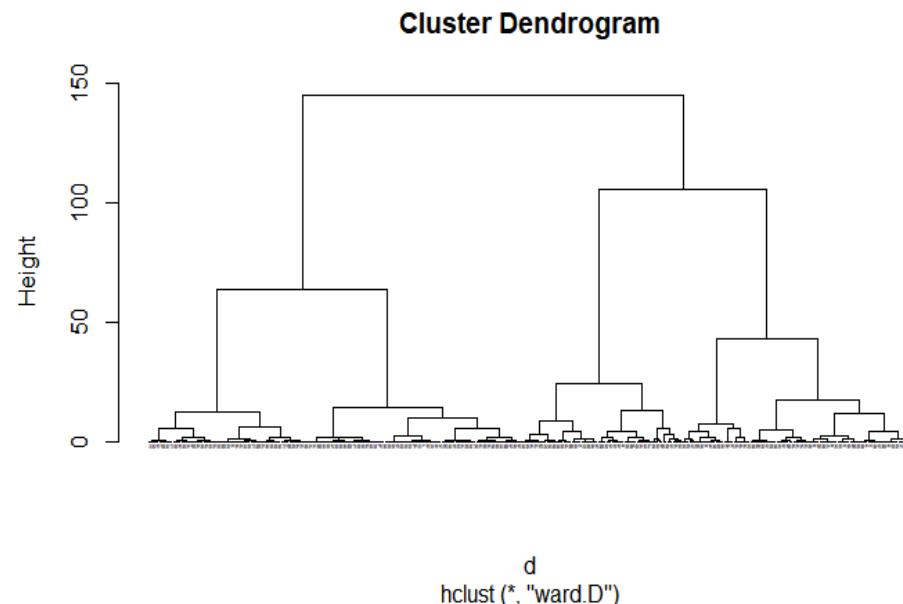
03.
입지 선정

04.
한계와 의의

02. 사용 기법

• 계층적 클러스터링 (Hierarchical Clustering)

개체들을 가까운 집단부터 순차적이고 계층적으로 차근차근 뭉어 나가는 방식



1. 모든 개체들 사이의 거리에 대한 유사도 행렬 계산
2. 거리가 인접한 관측치끼리 클러스터 형성
3. 유사도 행렬 업데이트



01.
1주차 피드백

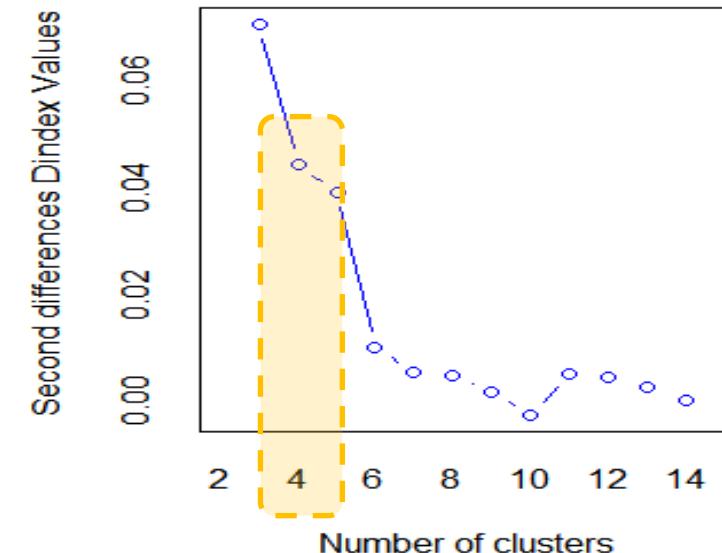
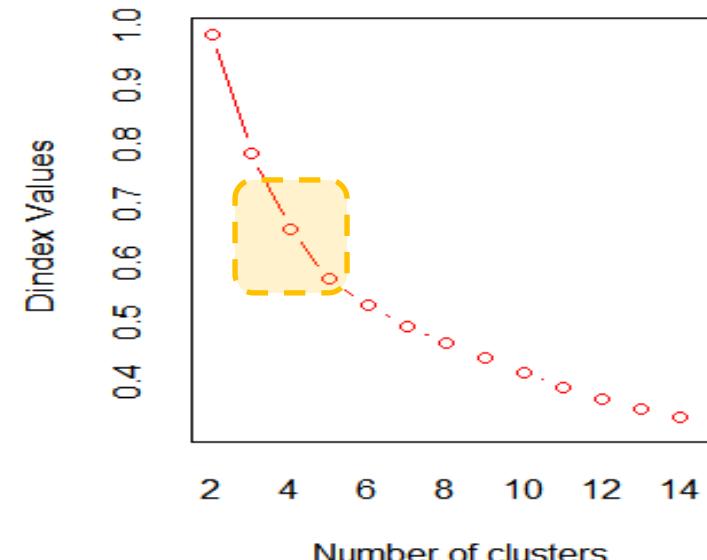
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- 계층적 클러스터링 (Hierarchical Clustering)



*Elbow point*와 *Silhouette*값을 고려해
클러스터 개수 4개 설정

01.
1주차 피드백

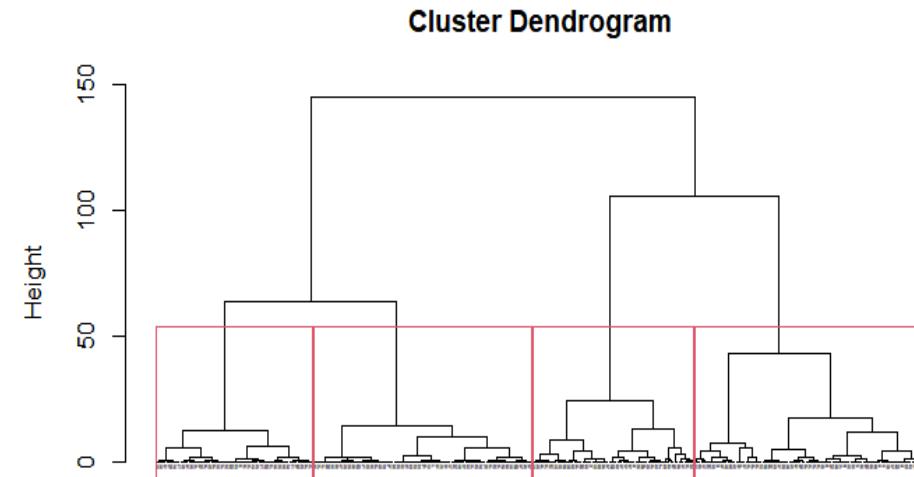
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- 계층적 클러스터링 (Hierarchical Clustering)



d
hclust (*, "ward.D")

상대적으로 분산이 작고,
노이즈나 이상치에 덜 민감한 방법

- Silhouette: 0.3385737
- 클러스터별 행정동 개수

Cluster	1	2	3	4
행정동	107	77	79	114

실루엣 값이 낮고,
타겟 클러스터 존재 X



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

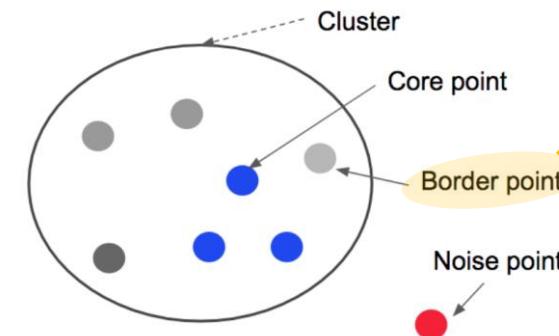
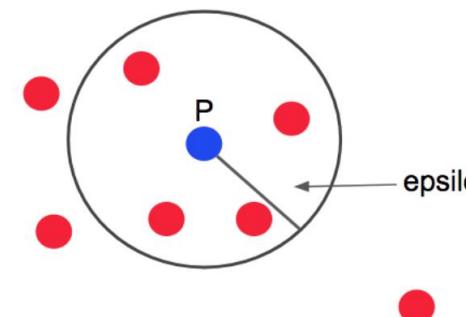
02. 사용 기법



- DBSCAN (Density-based spatial clustering of applications with noise)

밀도 기반 군집분석

1. 충분히 가까운(epsilon) 두 개의 핵심 점(Core point)이 동일한 클러스터에 배치
2. 핵심 점에 충분히 가까운 모든 경계 점(Border point)은 핵심 점과 동일한 클러스터에 배치
3. Noise points는 버림



Border point
최소한의 점의 개수를 만족하지
못하는 클러스터의 점들

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

- DBSCAN (Density-based spatial clustering of applications with noise)

파라미터 결정 방법

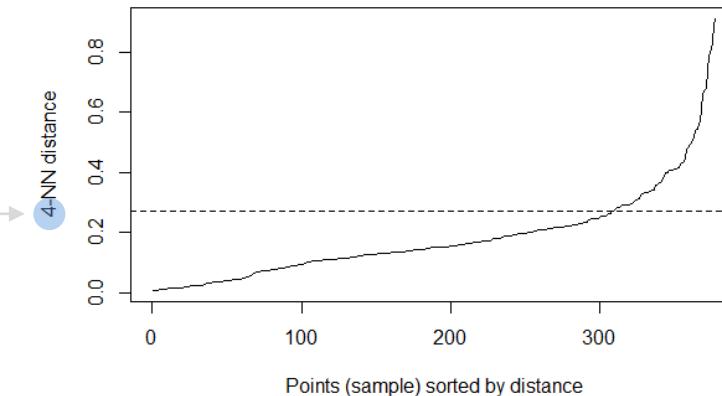
1. Eps (*epsilon*)

sorted k-dist graph에서
elbow point가 되는 k-dist

2. MinPts (*min points*)

데이터셋 별 객체 개수를 고려한
Heuristic 방법으로, $\ln(\text{객체 개수})$

2차원 데이터에서는
 $\text{MinPts} = 4$ 로 하는 것을 권장



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

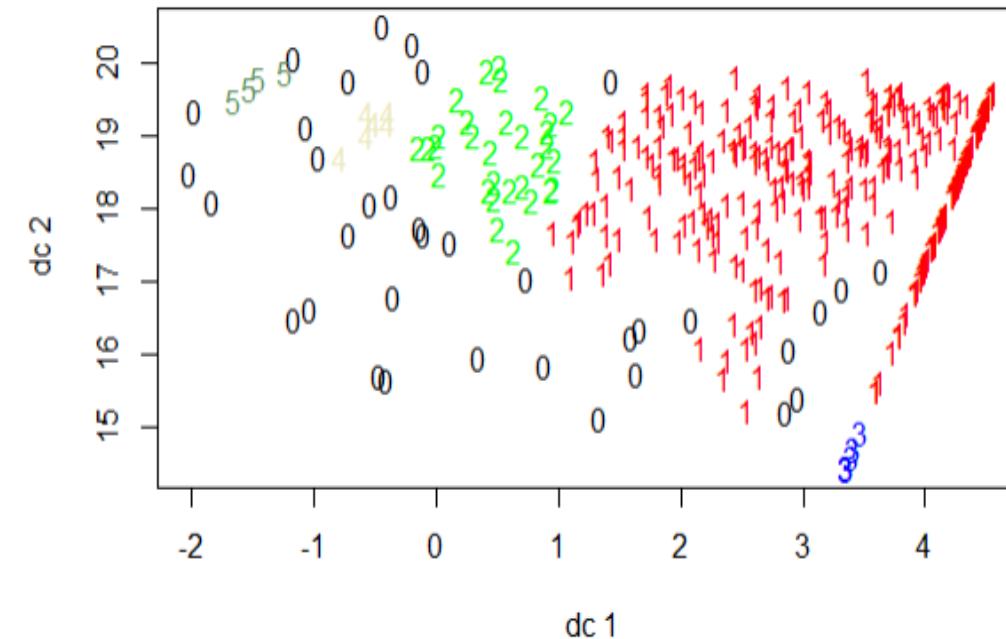
03.
입지 선정

04.
한계와 의의

02. 사용 기법



- DBSCAN (Density-based spatial clustering of applications with noise)



- Silhouette: 0.20612
- 클러스터별 행정동 개수

Cluster	0	1	2	3	4	5
행정동	36	295	31	5	6	4



하나의 cluster가 아니라 noise points임

01.
1주차 피드백

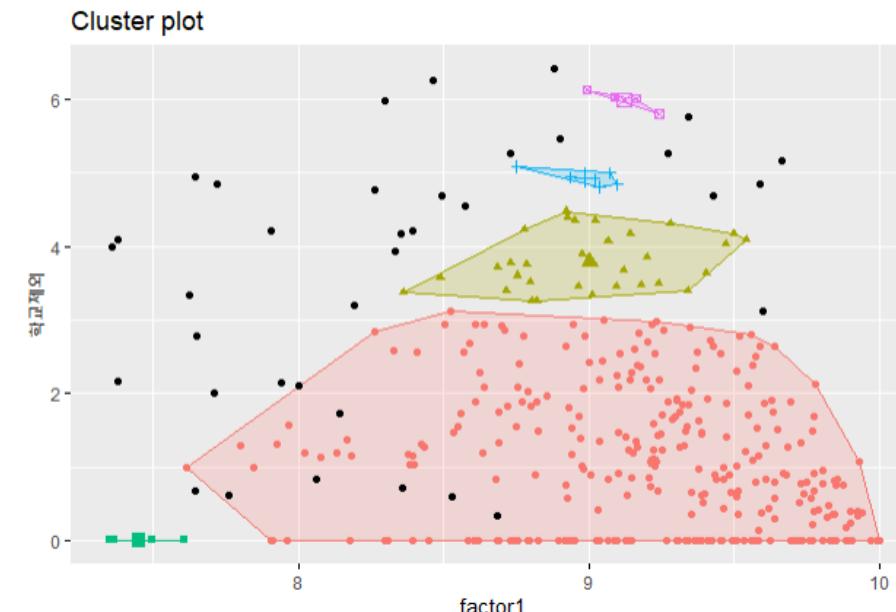
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- DBSCAN (Density-based spatial clustering of applications with noise)



- Silhouette: 0.20612
- 클러스터별 행정동 개수

Cluster	0	1	2	3	4	5
행정동	36	295	31	5	6	4

클러스터간 불균형이 심하고,
타겟 클러스터 존재 X



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

• HDBSCAN (Hierarchical DBSCAN)

계층적 밀도 기반 군집분석

DBSCAN

데이터들의 계층적 구조를 반영한 clustering이 불가능

HDBSCAN

- ϵ 파라미터는 더 이상 필요하지 않으며 MinPts만 존재
- 따라서, hyper-parameter tuning 비용이 상당히 줄게 됨

02. 사용 기법



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

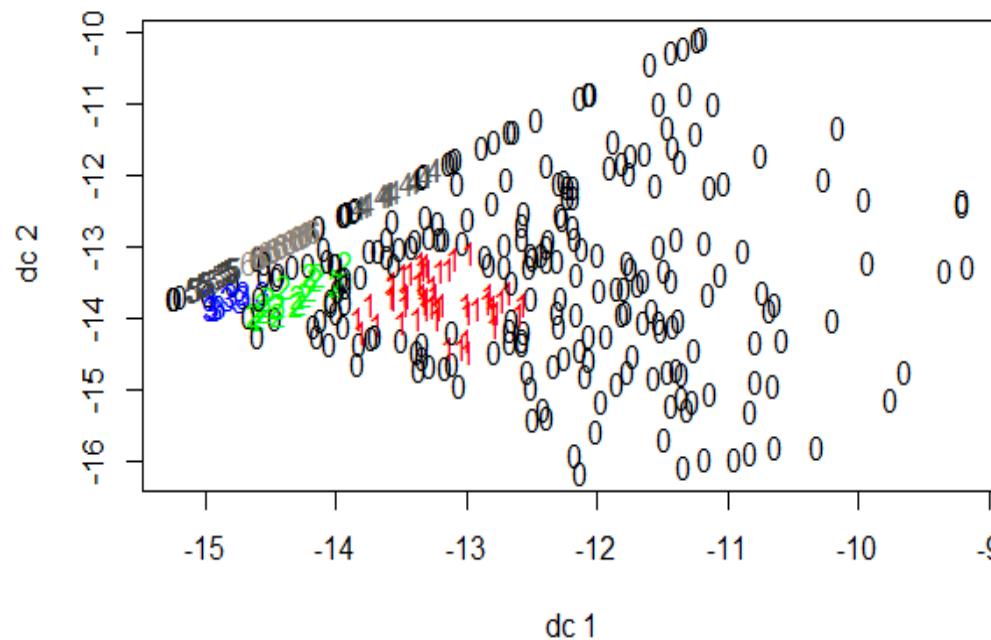
03.
입지 선정

04.
한계와 의의

02. 사용 기법

- HDBSCAN (Hierarchical DBSCAN)

계층적 밀도 기반 군집분석



- Silhouette: -0.03002
- 클러스터별 행정동 개수

Cluster	0	1	2	3	4	5	6
행정동	229	53	21	12	15	23	24

하나의 cluster가 아니라 noise points

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

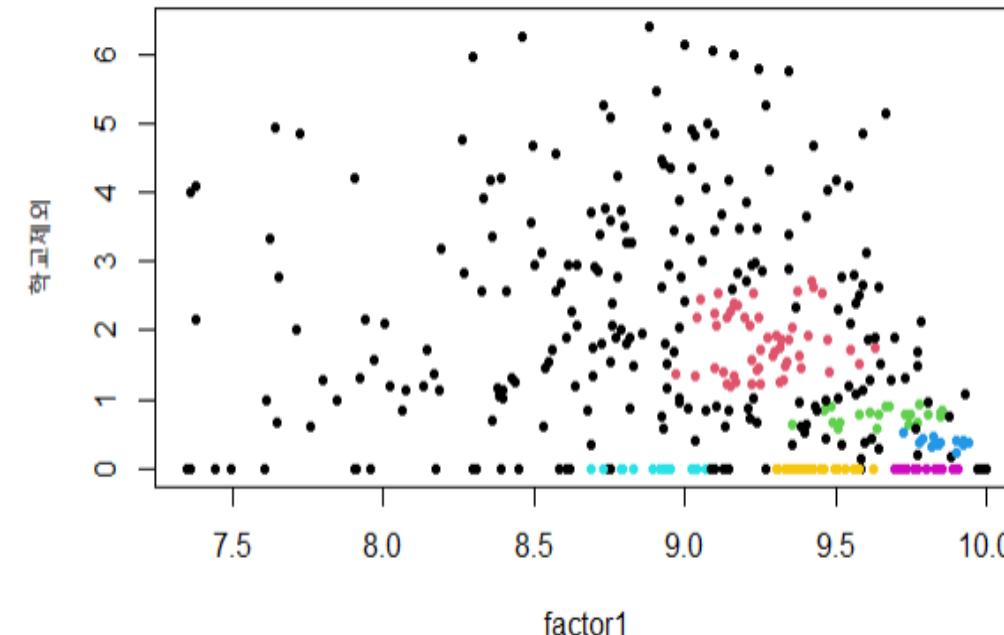
03.
입지 선정

04.
한계와 의의

02. 사용 기법

- HDBSCAN (Hierarchical DBSCAN)

계층적 밀도 기반 군집분석



- Silhouette: -0.03002
- 클러스터별 행정동 개수

Cluster	0	1	2	3	4	5	6
행정동	229	53	21	12	15	23	24

클러스터간 불균형이 심하고,
타겟 클러스터 존재 X



01.
1주차 피드백

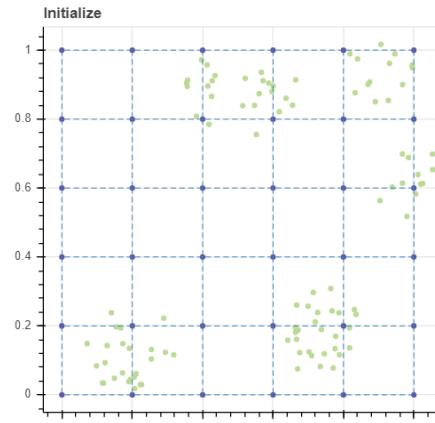
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

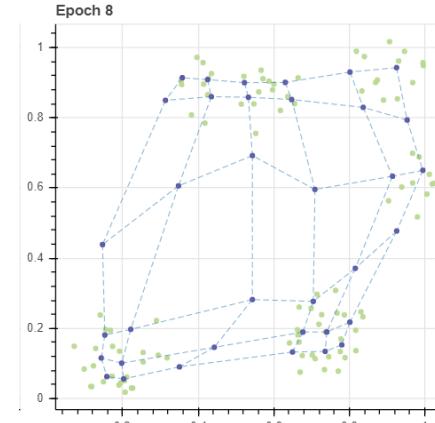
04.
한계와 의의

• SOM (Self Organized Map)

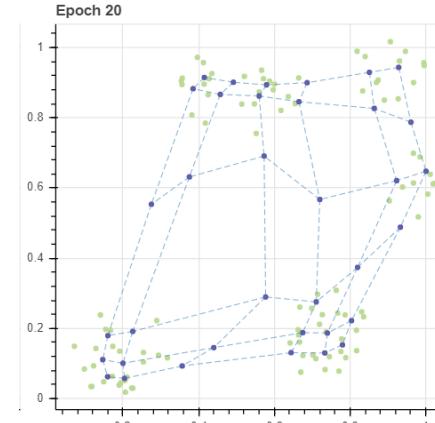
클러스터의 중심점을 관측치와 가깝게 업데이트해가는 DNN 기반의 클러스터링 방법



Random한 값을 가진
격자 벡터로 학습 시작



격자 벡터의 각 뉴런과의
거리를 계산하여 벡터의
가중치가 업데이트



입력된 모든 Input data의
가중치가 업데이트
될 때까지 반복

02. 사용 기법



01.
1주차 피드백

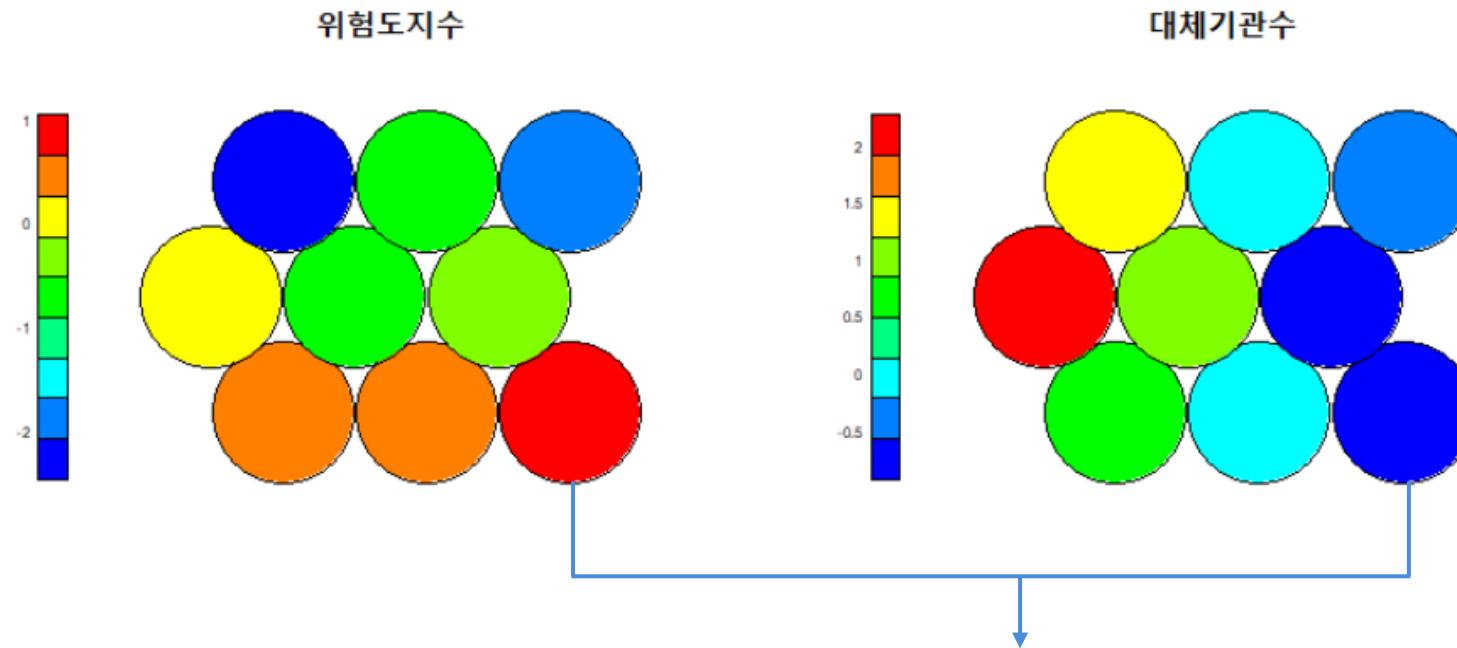
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- SOM (Self Organized Map)



위험도 지수가 높고, 대체기관 수는 적은 타겟 클러스터

01.
1주차 피드백

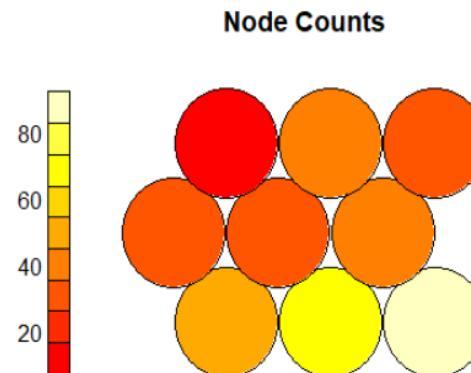
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

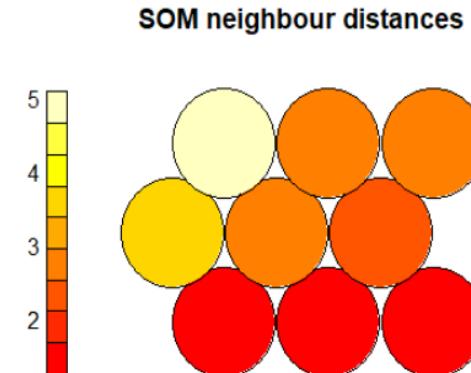
04.
한계와 의의

02. 사용 기법

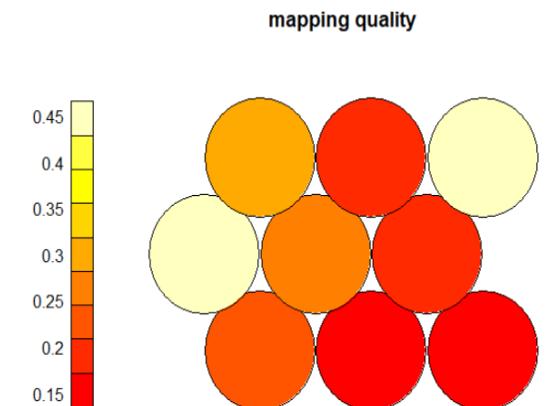
- SOM (Self Organized Map)



클러스터 내 obs. 개수



클러스터 간의 거리



클러스터 내 데이터들 간의 유사도

클러스터 간 **사이즈 불균형**이 존재하며,
클러스터 내의 **유사도** 또한 작음



01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

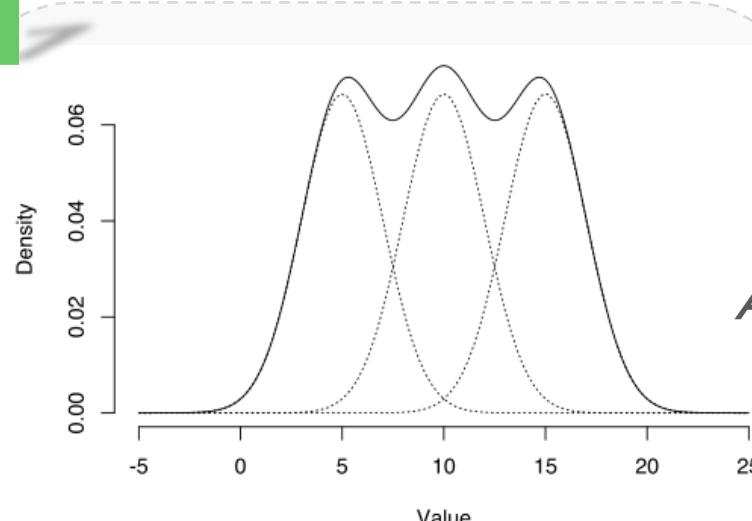
04.
한계와 의의

02. 사용 기법

• GMM (Gaussian Mixture Model) Clustering

데이터를 K개의 가우시안 분포의 **선형결합**으로 생각하고, 각 데이터를 적절한 분포에 배정하는 방법

1



K개의 가우시안 분포를 선형결합하여
만들어진 분포

EM
Algorithm

2

μ_k : 각 가우시안 분포의 mean

Σ_k : 각 가우시안 분포의 covariance

π_k : 각 가우시안 분포의 크기

EM 알고리즘을 이용하여
파라미터 추정



02. 사용 기법

01.
1주차 피드백

02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

• GMM (Gaussian Mixture Model) Clustering

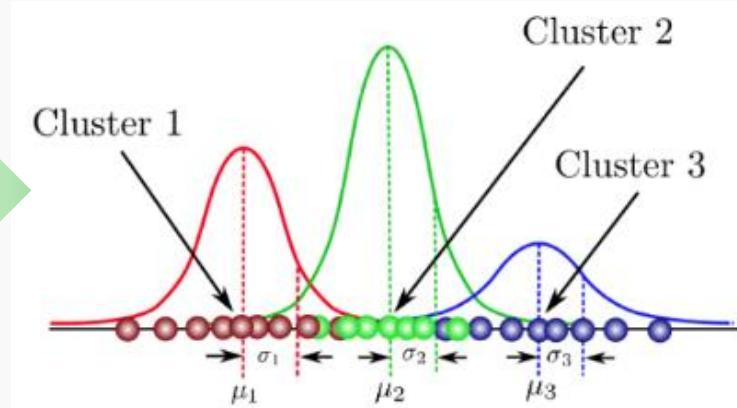
데이터를 K개의 가우시안 분포의 **선형결합**으로 생각하고, 각 데이터를 적절한 분포에 배정하는 방법

3

$$P(z_k = 1|x_n) = \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_1^K \pi_j N(x_n|\mu_j, \Sigma_j)}$$

각각의 **데이터**가 K개의 가우시안 분포 중
몇 번째 가우시안 분포로부터 왔는지
확률(책임값) 계산

4



가장 **책임값이 큰** K로
클러스터링

01.
1주차 피드백

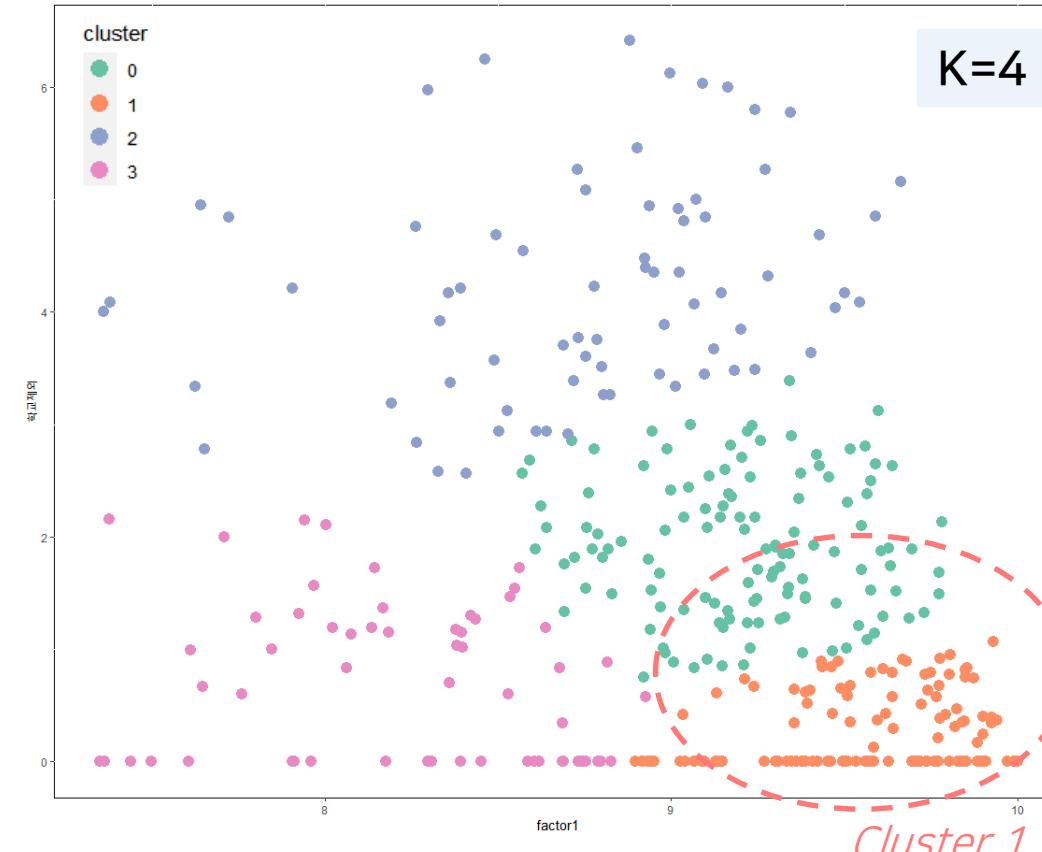
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

- GMM (Gaussian Mixture Model) Clustering



- Silhouette: 0.4291546
- 클러스터별 행정동 개수

Cluster	0	1	2	3
행정동	125	122	72	58

위험도지수가 높으면서 보호기관수가 적은
122개의 행정동

01.
1주차 피드백

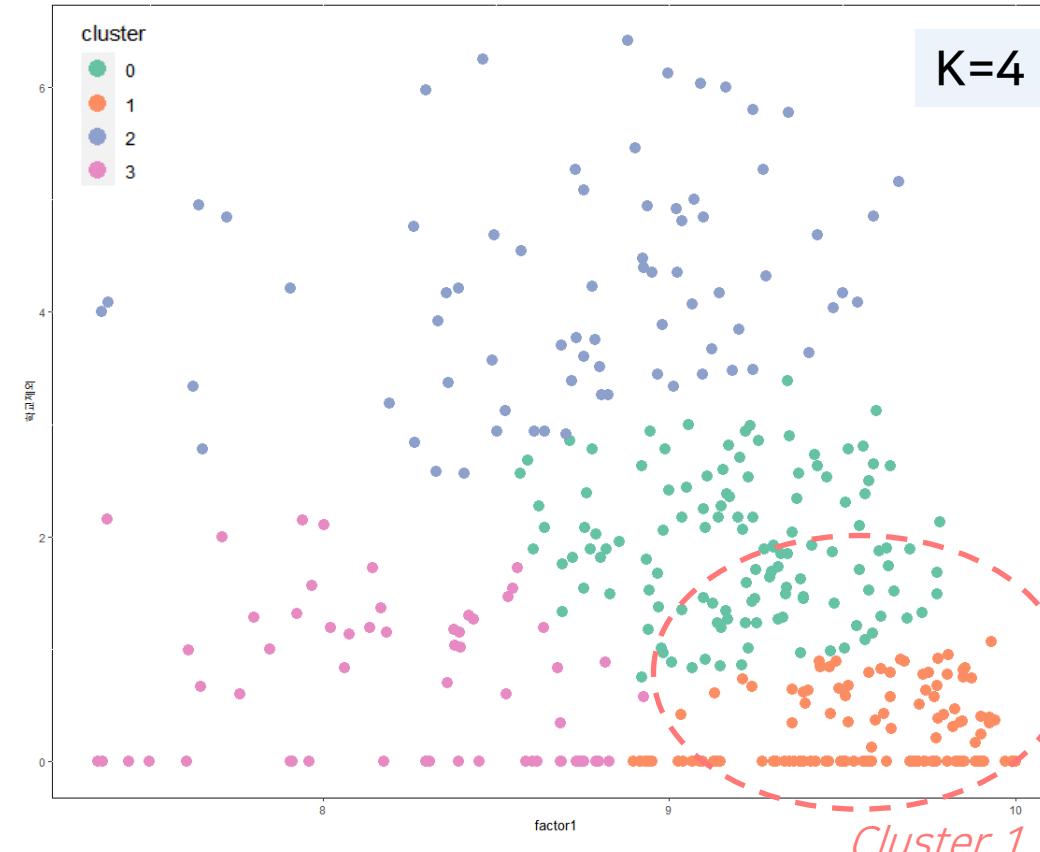
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 사용 기법

GMM (Gaussian Mixture Model) Clustering



- Silhouette: 0.4291546
- 클러스터별 행정동 개수

Cluster	0	1	2	3
행정동	125	122	72	58

실루엣 값과 클러스터 사이즈를 고려해 GMM 클러스터링 채택!



01.
1주차 피드백

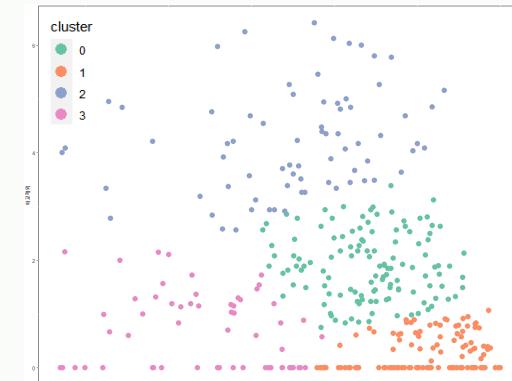
02.
이상치 제거
사용 기법
결과 해석

03.
입지 선정

04.
한계와 의의

02. 결과 해석

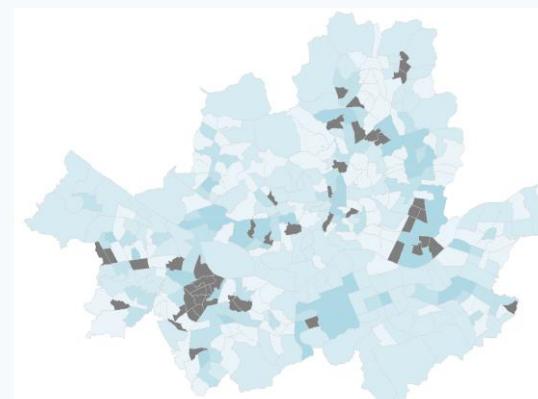
- GMM (Gaussian Mixture Model) Clustering



클러스터별 평균적인 위험도지수, 대체기관비율 값

Cluster	0	1	2	3
평균 위험도지수	9.21	9.57	8.76	8.23
평균 대체기관비율	1.90	0.268	4.21	0.711

Target!



클러스터링 결과

행정동	Cluster
사직동	1
삼청동	1
회현동	3
청파동	2



03

입지 선정

- 1) 행정동 선정
- 2) 최종 입지

03. 행정동 선정

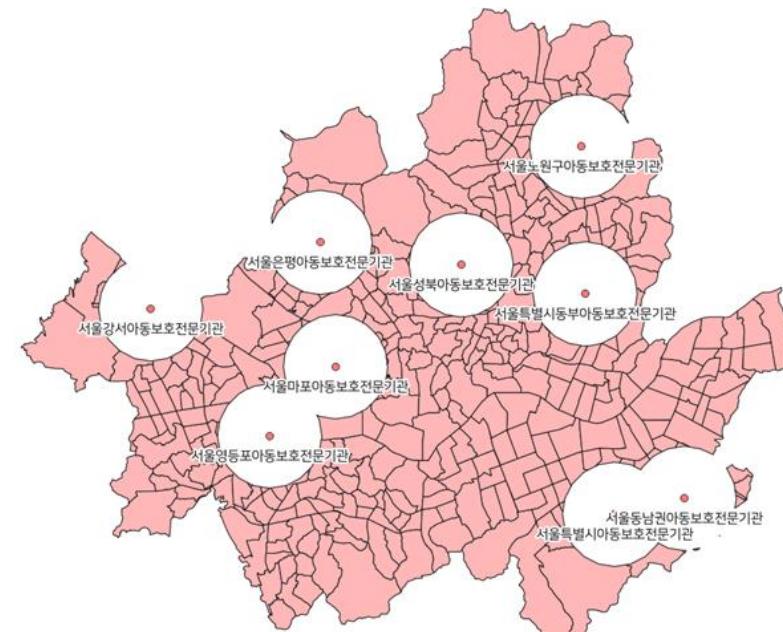
01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

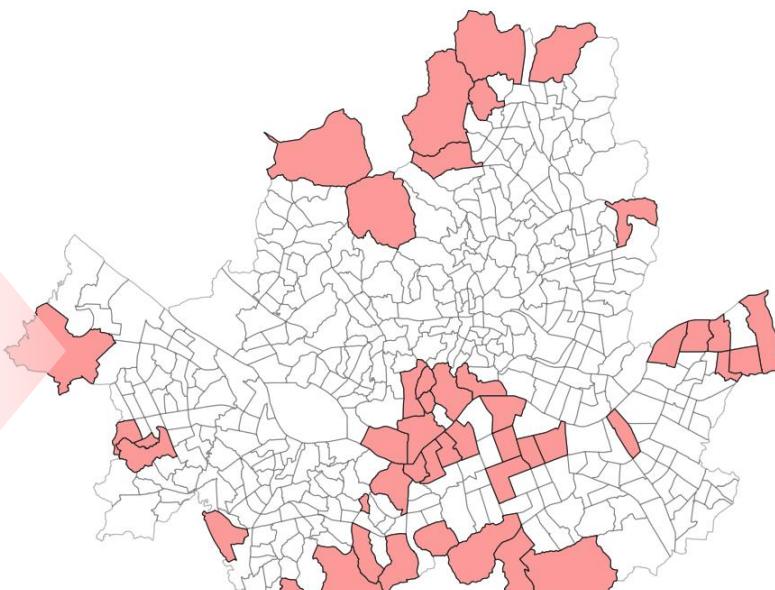
- 입지 후보 행정동 선출



현재 존재하는 아동보호전문기관의 2.5km
반경 내에 존재하는 행정동 제외

보건복지부 「사회보장제도 「아동의 안전한 성장을 위한 제도·전달체계 심층분석」 핵심평가」 참고

GMM 클러스터링에서 채택한 클러스터와
겹치는 행정동만 추출



01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

- 입지 선정 기법

LSCP Location Set Covering Problem

특정한 지리적 범위 내에 모든 수요가 커버될 수 있도록 시설물의 수를 최소화하는 모델.

특정 영역을 포함하는 시설물의 개수와 그 지점을 도출하기 위한 방법.

MCLP Maximal Covering Location Problem

주어진 시설물의 개수로 각 좌표가 가지는 수요를 최대로 커버하며, 그 합을 최대화하도록 입지를 선정하는 방법.



03. 행정동 선정



- LSCP (Location Set Covering Problem)

GAASS 알고리즘을 적용해 모든 수요를 커버할 수 있는 최소의 행정동 개수 확인

→ **GAASS** (*Greedy Adding Algorithm with Substitution Procedure*)

Greedy 하게 순간마다 최적이라고 생각되는 후보지를 결정하는 방식.

1. 임의적으로 하나의 입지 선택
2. 주변에 존재하는 입지를 후보에서 삭제
3. 남은 입지 중 score가 가장 높은 입지 선택 반복

→ 선택된 입지가 커버할 수 있는 수요

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 행정동 선정

- LSCP (Location Set Covering Problem)

1. 거리계산 함수를 이용하여 행정동 별 거리를 계산

행정동	Center_lat	Center_long
평창동	37.61551	126.9679
용산2가동	37.53884	126.9858
이촌1동	37.51467	126.9700
이태원1동	37.53354	126.9929
이태원2동	37.54156	126.9923
서빙고동	37.51813	126.9899

각 행정동의 중심점을
기준으로 1:1 대응을
통해 행정동 간의
거리를 구함



01.
1주차 피드백

02.
클러스터링

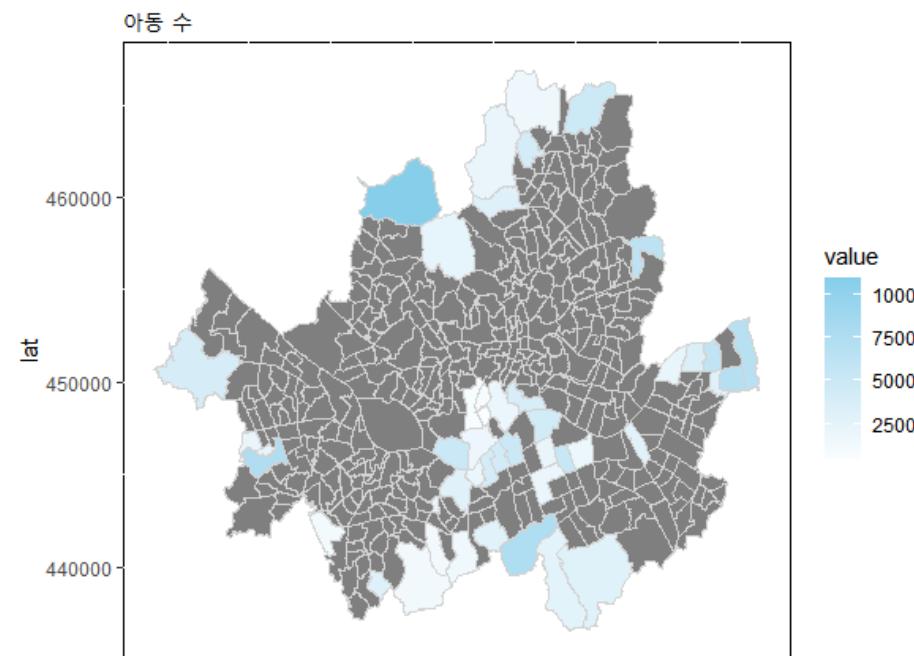
03.
행정동 선정
최종 입지

04.
한계와 의의

03. 행정동 선정

- LSCP (Location Set Covering Problem)

2. 행정동 별 2.5km 근방 아동 수를 합산하여 계산



행정동 별 2.5km 근방
아동 수를 합산하여
시각화한 결과,
다음과 같은 모양을 띨

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 행정동 선정

- LSCP (Location Set Covering Problem)

3. 커버되지 않은 아동이 없을 때까지 설치 위치를 선정

설치 위치:
남은 아동수: 160372
설치 위치: 43
남은 아동수: 133752

설치 위치: 43 29
남은 아동수: 107968

설치 위치: 43 29 39
남은 아동수: 92221

설치 위치: 43 29 39 31
남은 아동수: 80196

⋮

설치 위치: 43 29 39 31 16 15 12 9 33 14 24 2 19 21 11 40 1 45
남은 아동수: 5071

설치 위치: 43 29 39 31 16 15 12 9 33 14 24 2 19 21 11 40 1 45 38
남은 아동수: 3024

설치 위치: 43 29 39 31 16 15 12 9 33 14 24 2 19 21 11 40 1 45 38 36
남은 아동수: 1190

설치 위치: 43 29 39 31 16 15 12 9 33 14 24 2 19 21 11 40 1 45 38 36 20
남은 아동수: 0



입지를 선정할 때마다
커버되는 아동 수를
삭제해가며, 커버되지 않은
아동 수가 0명이 될 때까지
진행

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

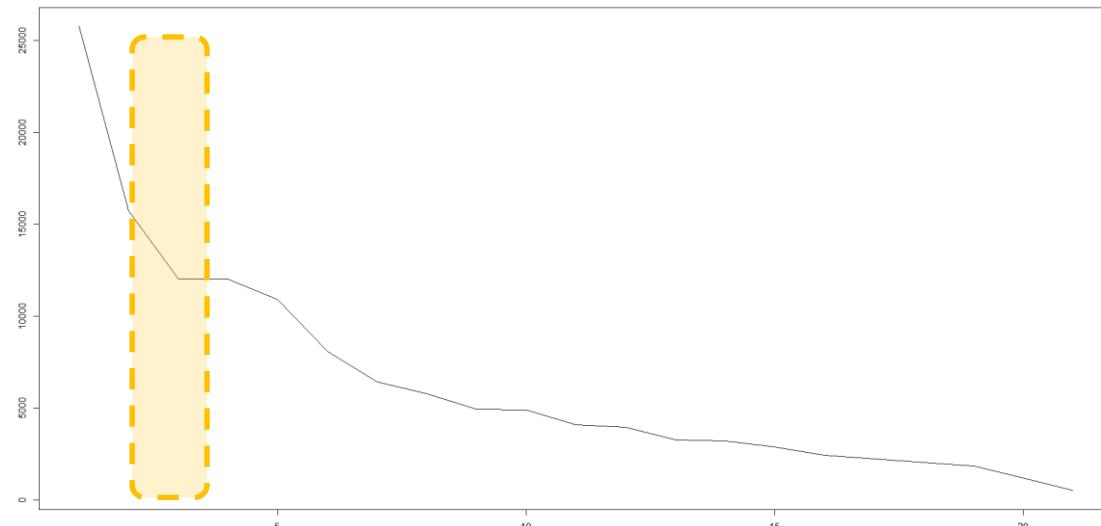
04.
한계와 의의

03. 행정동 선정

- LSCP (Location Set Covering Problem)

총 21개의 행정동 필요

설치 위치: 43 29 39 31 16 15 12 9 33 14 24 2 19 21 11 40 1 45 38 36 20
남은 아동수: 0



시설을 설치할 때마다
남은 아동 수를
시각화 했을 때,
Elbow point인 3으로
시설 개수 결정!

03. 행정동 선정

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

- MCLP : 반경 내 수요 최대화

설치할 시설물의 개수가 고정되어 있을 때,

지역 수요를 최대한 커버할 수 있도록 입지를 선정하는 방법

아동인구가 많으면 아동학대가 존재할 가능성이 높기 때문에,

아동인구를 고려해서 선정

1주차 피드백



03. 행정동 선정

01.
1주차 피드백

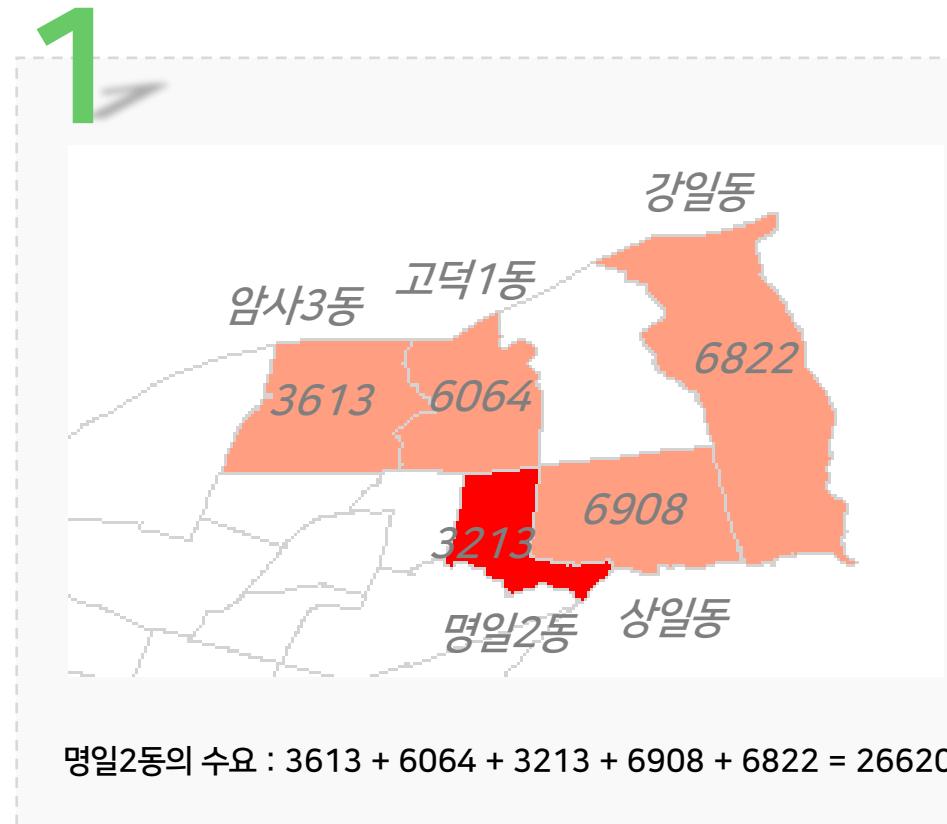
02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

- MCLP : 반경 내 수요 최대화



후보 행정동의 중심점을 기준으로
반경 2.5km 이내의 아동 수를 합산하여
각 행정동의 수요를 계산

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

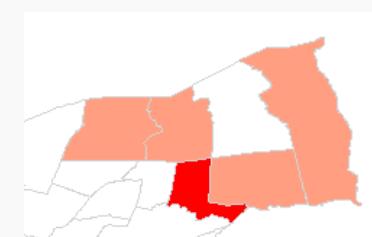
03. 행정동 선정

- MCLP : 반경 내 수요 최대화

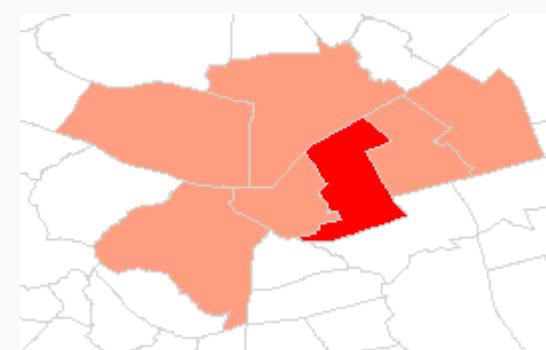
2



신월6동 (12022명)



명일2동 (26620명)



반포2동 (25784명)

반경 내 가장 아동 수가 많은
행정동을 우선 입지 지역에 포함



이미 선정된 행정동과
커버리지가 겹치지 않도록
수요를 최대화하는 3개의 행정동 선정

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

03. 행정동 선정

- MCLP : P-median

수요와 거리의 곱을 최소화하는 입지를 찾는 알고리즘

$$\text{Minimize } \sum_j h_i d_{ij} y_{ij} \quad \dots \quad (1)$$

$$\sum_j y_{ij} = 1 \quad \dots \quad (2)$$

$$\text{Subject to } \sum_j x_j = p \quad \dots \quad (3)$$

$$y_{ij} \leq x_j \quad \dots \quad (4)$$

$$x_j, y_{ij} \in \{0,1\} \quad \dots \quad (5)$$



h_i : 수요지의 수요량 (아동 인구)

d_{ij} : 수요지와 입지점의 거리

p : 시설 개수

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 행정동 선정

- MCLP : P-median

수요와 거리의 곱을 최소화하는 입지를 찾는 알고리즘

$$\text{Minimize } \sum_j h_i d_{ij} y_{ij} \quad \dots \quad (1)$$

$$\sum_j y_{ij} = 1 \quad \dots \quad (2)$$

$$\text{Subject to} \quad \sum_j x_j = p \quad \dots \quad (3)$$

$$y_{ij} \leq x_j \quad \dots \quad (4)$$

$$x_j, y_{ij} \in \{0,1\} \quad \dots \quad (5)$$



i = 수요지, j = 입지

x_j : $\begin{cases} \text{노드 } j \text{에 시설물이 설치되면, 1} \\ \text{그렇지 않으면, 0} \end{cases}$

y_{ij} : $\begin{cases} \text{노드 } j \text{의 시설물이 노드 } i \text{의 총수요를 충족시키면, 1} \\ \text{그렇지 않으면, 0} \end{cases}$

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

03. 행정동 선정

- MCLP : P-median

수요와 거리의 곱을 최소화하는 입지를 찾는 알고리즘

$$\text{Minimize } \sum_j h_i d_{ij} y_{ij} \quad \dots \quad (1)$$

$$\sum_j y_{ij} = 1 \quad \dots \quad (2)$$

$$\text{Subject to } \sum_j x_j = p \quad \dots \quad (3)$$

$$y_{ij} \leq x_j \quad \dots \quad (4)$$

$$x_j, y_{ij} \in \{0,1\} \quad \dots \quad (5)$$



i = 수요지, j = 입지

(1) 수요와 거리의 곱을 최소화 시킨다.

(2) 각 수요지를 커버하는 기관은 1개이다.

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

03. 행정동 선정

- MCLP : P-median

수요와 거리의 곱을 최소화하는 입지를 찾는 알고리즘

$$\text{Minimize } \sum_j h_i d_{ij} y_{ij} \quad \dots \quad (1)$$

$$\sum_j y_{ij} = 1 \quad \dots \quad (2)$$

$$\text{Subject to } \sum_j x_j = p \quad \dots \quad (3)$$

$$y_{ij} \leq x_j \quad \dots \quad (4)$$

$$x_j, y_{ij} \in \{0,1\} \quad \dots \quad (5)$$



i = 수요지, j = 입지

(3) 기관 설치 개수는 p개로 제한한다.

(4) 수요지 i를 입지 j에서 커버한다면, 반드시 노드 j에 기관이 존재해야 한다.

01.
1주차 피드백

02.
클러스터링

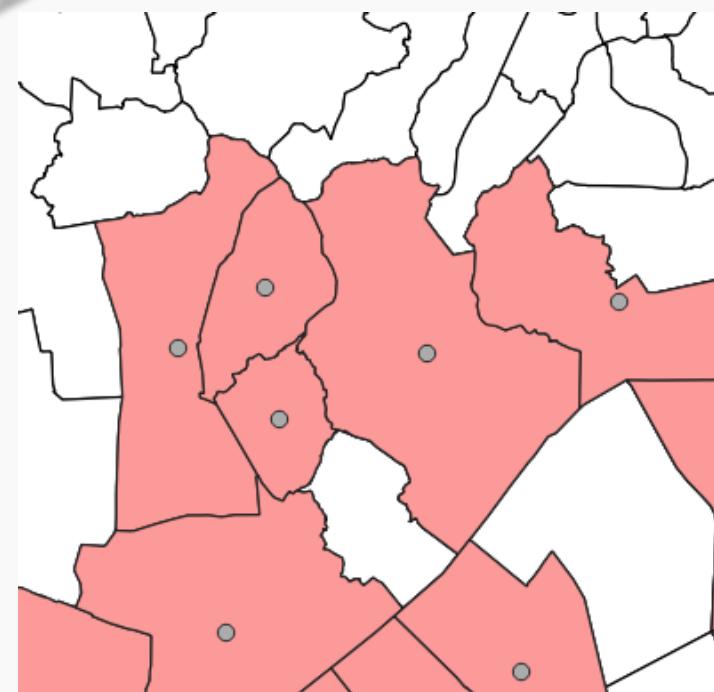
03.
행정동 선정
최종 입지

04.
한계와 의의

03. 행정동 선정

- MCLP : P-median

1



1사분위	2사분위	3사분위	4사분위
1배	2배	3배	4배

아동 수와 비례하도록

각 행정동마다 가중치 부여

03. 행정동 선정

01.
1주차 피드백

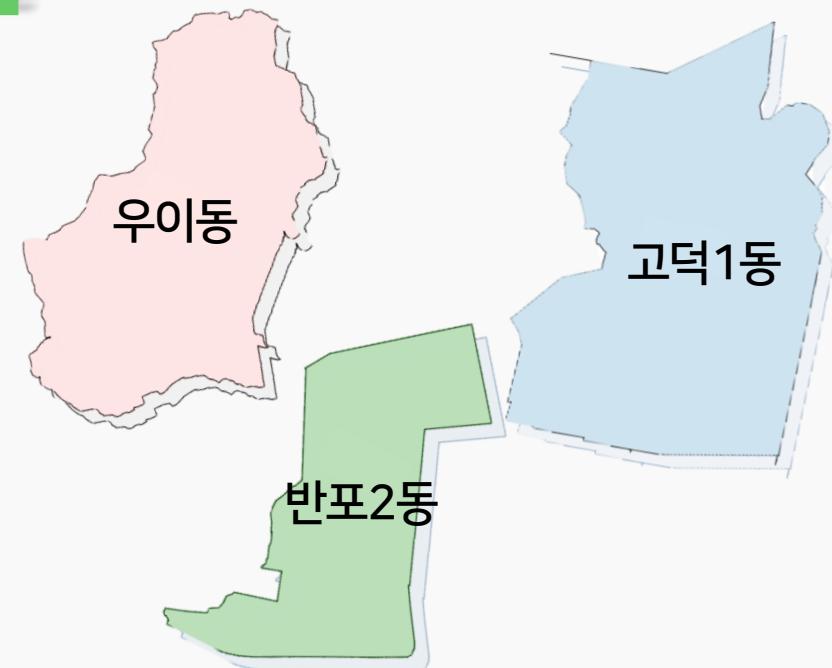
02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

- MCLP : P-median

2



후보 행정동의 **중심점**을 기준으로
반경 2.5km 이내의 **아동 수**와
거리의 곱을 최소화 하는
3개의 행정동 선정



01.
1주차 피드백

02.
클러스터링

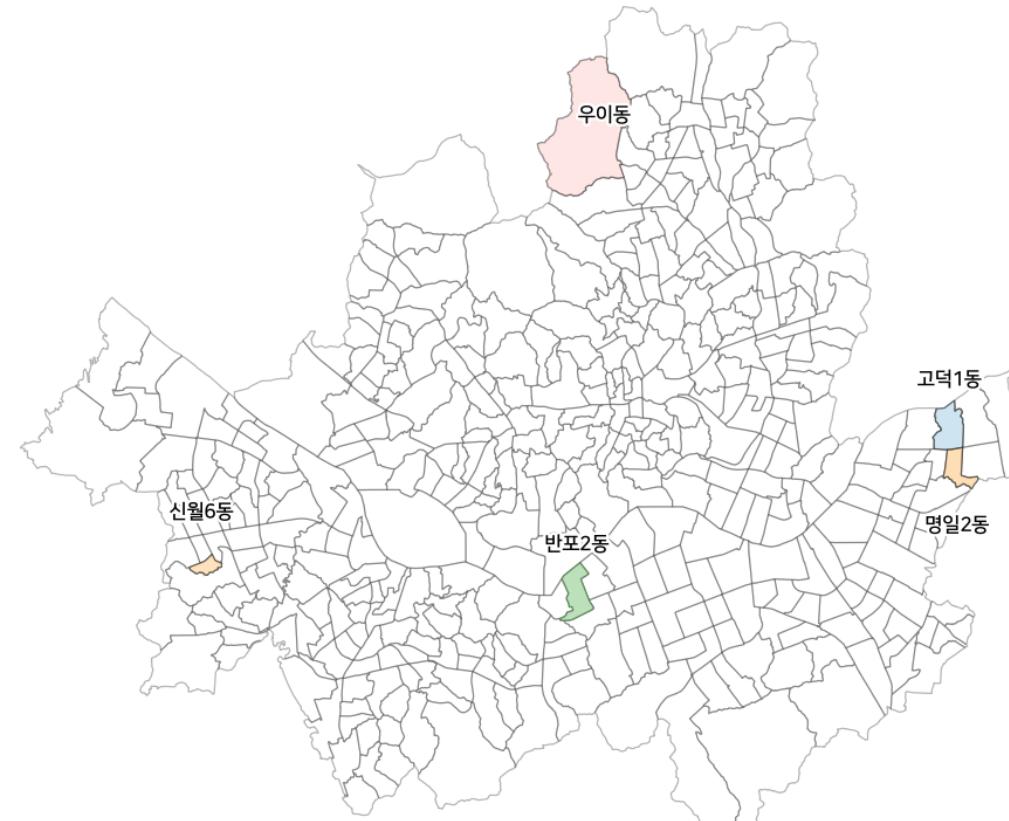
03.
행정동 선정

최종 입지

04.
한계와 의의

03. 행정동 선정

- 최종 행정동 선정



'반포2동'이 중복되기 때문에

총 5개의 행정동에 대해

최종 입지 선정 진행

01.
1주차 피드백

02.
클러스터링

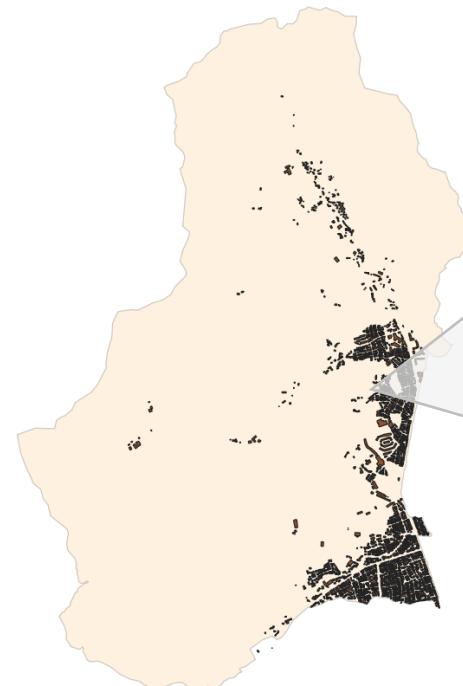
03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동

1. 청소년 유해업소 반경 50m 내 건물 제거



전체 건물 중
청소년 유해업소 반경 50m 버퍼 설치해
버퍼 내 건물 제거



01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

• 청소년 유해업소

청소년의 출입과 고용이 청소년에게 유해한 것

청소년 고용금지업소

청소년 출입·고용금지 업소

유해업종명	유해업종코드
호텔	11101
여관	11102
단란주점	12001
유흥주점	12002
주점영업	12007

업종 코드를 이용해
건물 제거

외 41개 유해업종 코드

01.
1주차 피드백

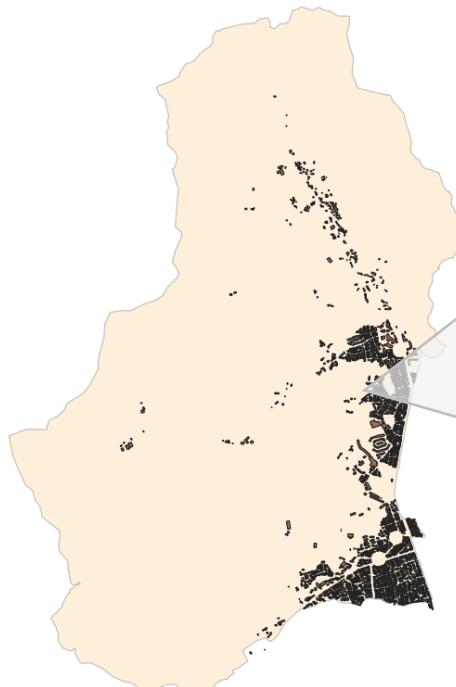
02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동
- 2. 연면적 $150m^2$ 이상 건물 추출



연면적이 $150m^2$ 미만인 건물 제거

- 연면적 $150m^2$ 이상
- 청소년 유해업소 제거 후 건물

01.
1주차 피드백

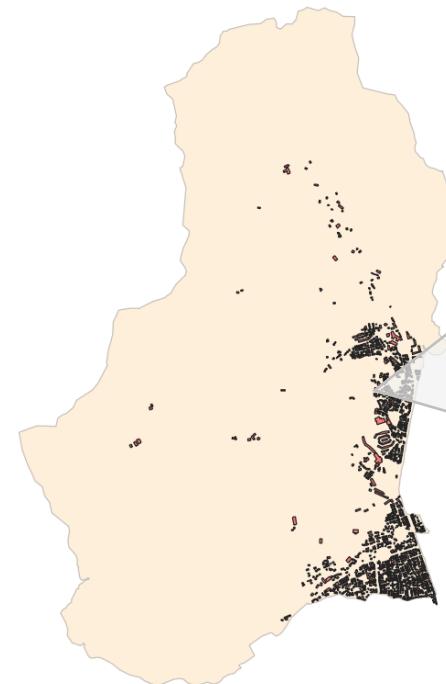
02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동
- 3. 실질적 주거 시설 건물 제거



실질적 주거 시설 (아파트, 빌라 등) 제거

- 주거 시설
- 연면적 150m² 이상 추출 후 건물

01.
1주차 피드백

02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동
3. 실질적 주거 시설 건물 제거

거주시설명	거주시설코드
단독주택	01000
다중주택	01002
다가구주택	01003
공관	01004
공동주택	02000

외 6개 거주시설 코드

업종 코드를 이용해
건물 제거

실질적 주거 시설 (아파트, 빌라 등) 제거

주거 시설

연면적 150m² 이상 추출 후 건물



01.
1주차 피드백

02.
클러스터링

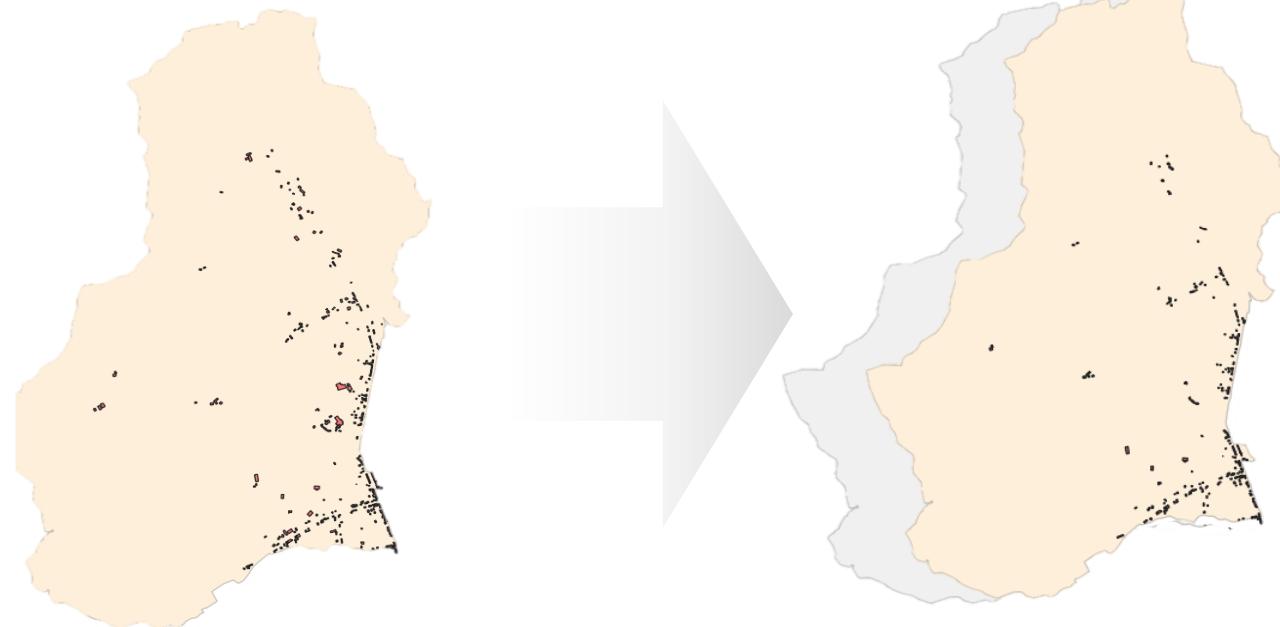
03.
행정동 선정

최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동
- 3. 명확한 목적성이 없는 건물 추출



우체국, 학교와 같이 확실한 목적을 가진 건물을 제외하기 위해
건물명을 기준으로 'NULL' 또는 '_빌딩' 추출



01.
1주차 피드백

02.
클러스터링

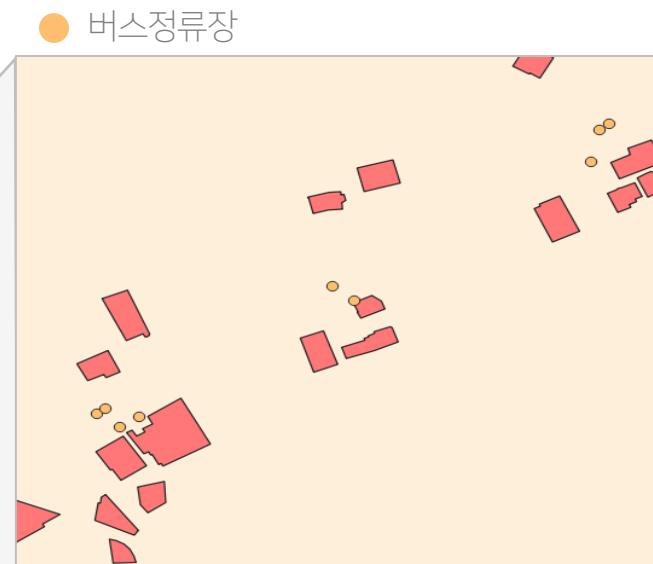
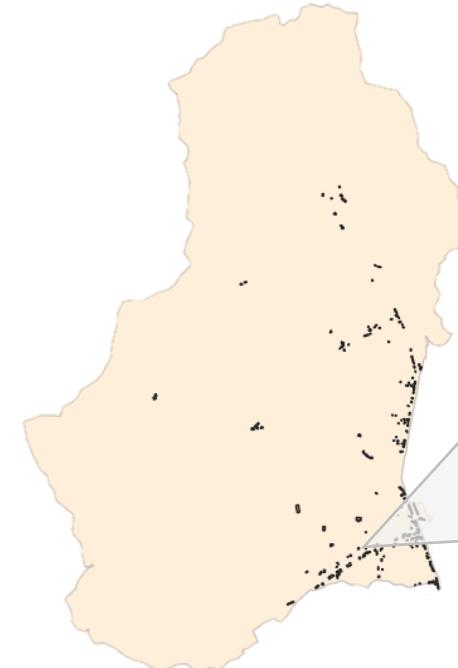
03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동

4. 시설까지의 접근성 고려 - 버스정류장



모든 후보지 별
최근점 버스정류장 3개까지의 평균 거리 계산해
가장 짧은 후보지 5개 선정



01.
1주차 피드백

02.
클러스터링

03.
행정동 선정

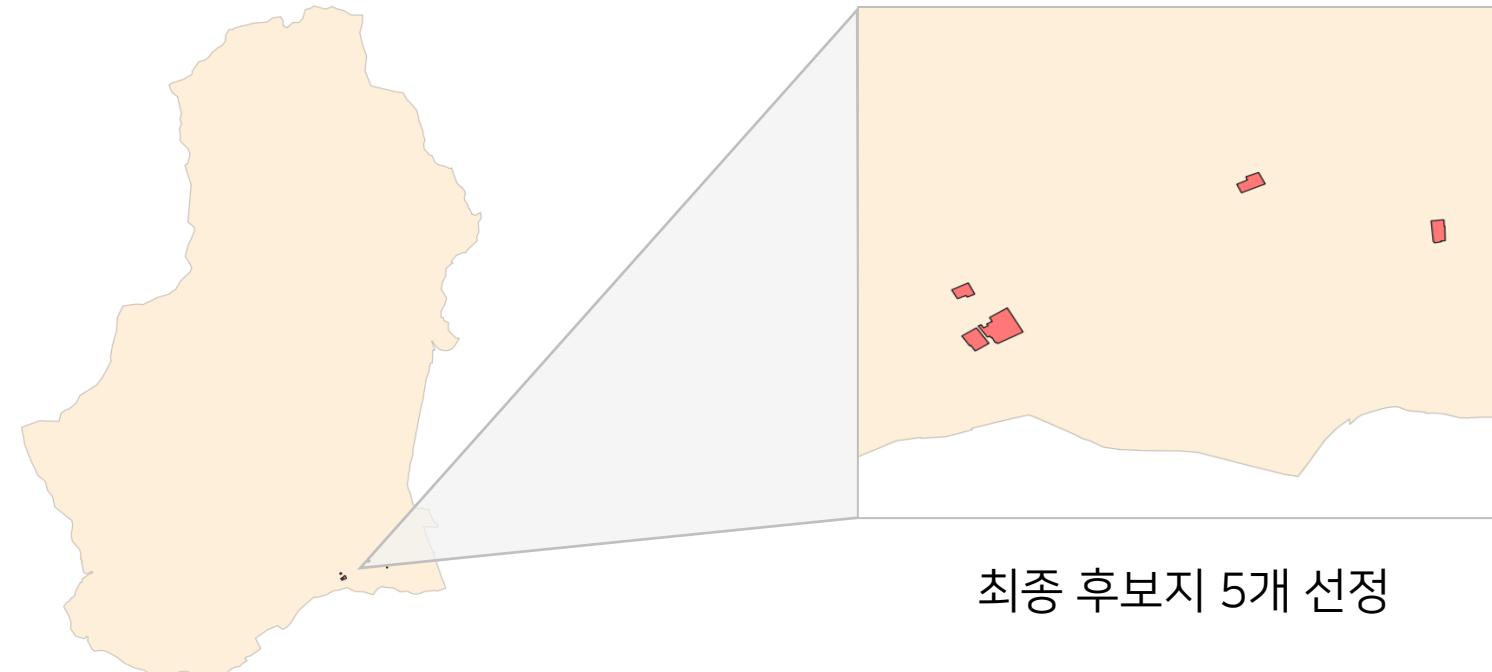
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 우이동

5. 최종 후보지 선정



03. 최종 입지

01.
1주차 피드백

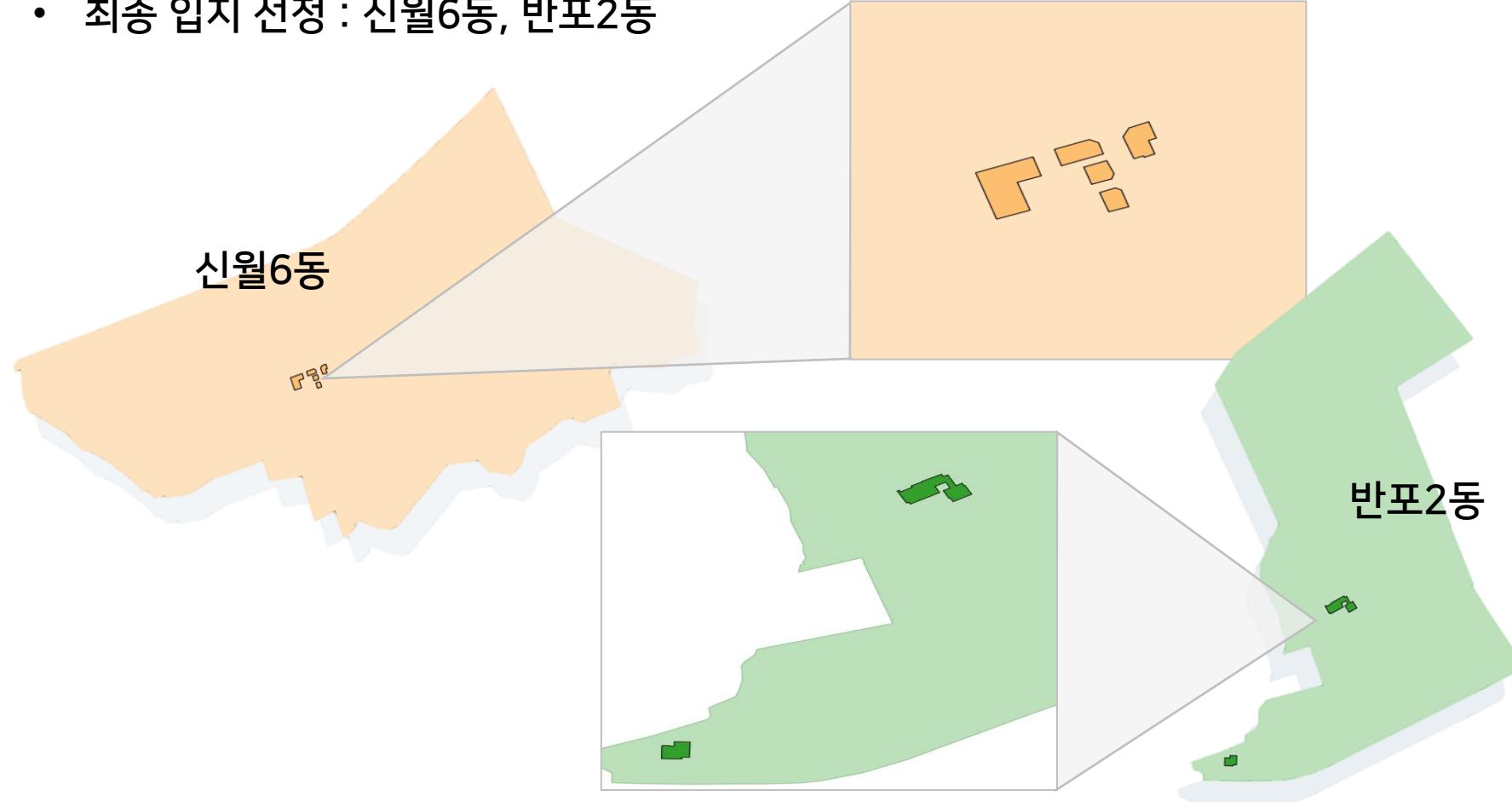
02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

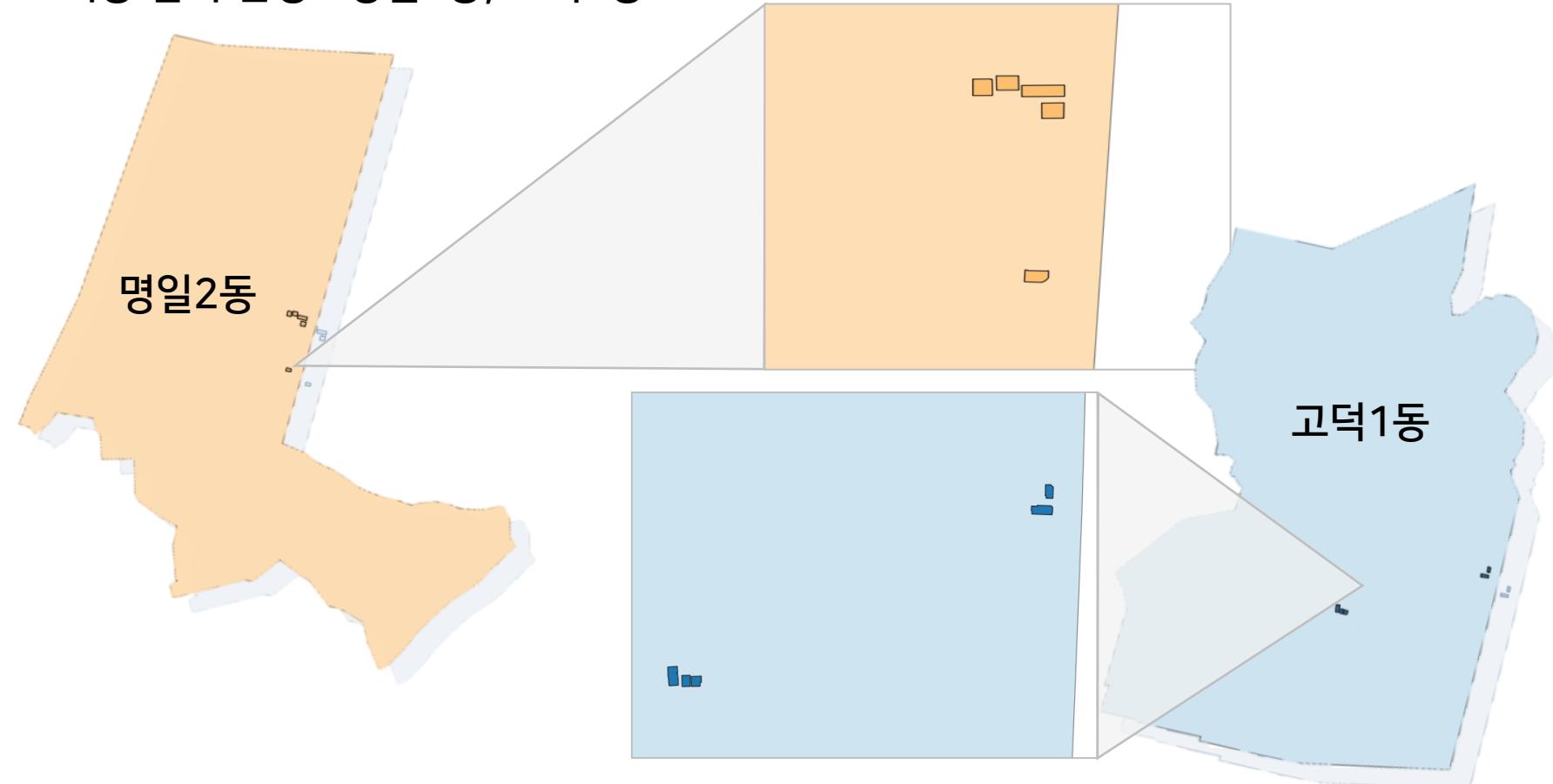
- 최종 입지 선정 : 신월6동, 반포2동



03. 최종 입지



- 최종 입지 선정 : 명일2동, 고덕1동



01.
1주차 피드백

02.
클러스터링

03.
행정동 선정

최종 입지

04.
한계와 의의

01.
1주차 피드백

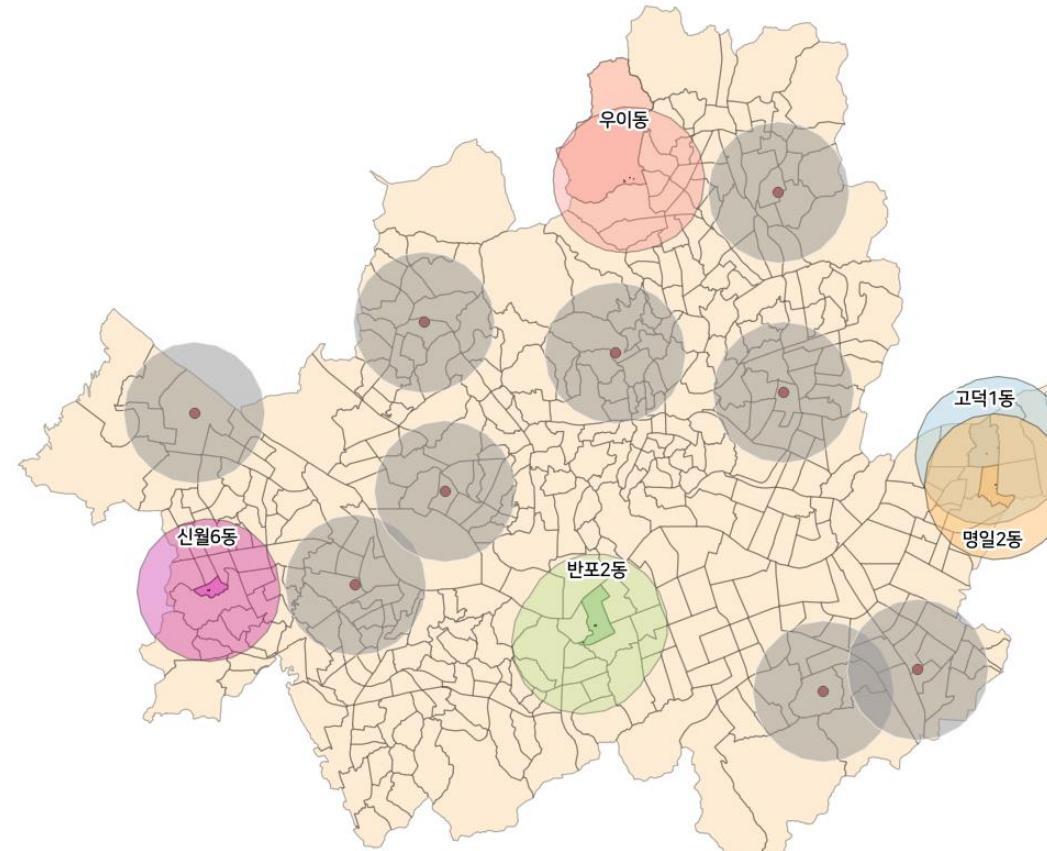
02.
클러스터링

03.
행정동 선정
최종 입지

04.
한계와 의의

03. 최종 입지

- 최종 입지 선정 : 전체 커버리지



현재 존재하는
아동보호전문기관의 커버리지와
겹치지 않으며,
부족한 부분을 잘 채워주는
입지가 선정되었음

● 현재 존재하는
아동보호전문기관의 커버리지



04

한계와 의의

주분끝!

04. 한계와 의의



- **한계**



행정동 별 데이터의 부족으로 다양한 변수 고려 부족



예산, 건물 상황 등 현실적인 문제를 고려하지 못함



가중치 산출, 클러스터링 등에서 주관적 해석이 반영됨

01.
1주차 피드백

02.
클러스터링

03.
입지 선정

04.
**한계와
의의**

01.
1주차 피드백

02.
클러스터링

03.
입지 선정

04.
한계와
의의

04. 한계와 의의

- 의의

카카오 API, QGIS 등 다양한 툴을 사용함



R, QGIS로 지도 시각화 뿐만 아니라



다양한 클러스터링 기법 사용

여러 입지선정 알고리즘 고려



가장 귀여운 PPT로 피셋 귀여움 담당

결론: 회귀가 짱먹었다



감사합니다

