

2022년 통계데이터 분석 · 활용대회 데이터 분석

서울특별시 행복주택 최적 입지 선정



Contents



1. 분석 배경

- 주제 선정 배경
- 분석 필요성
- 선행 연구 및 분석 전략

2. 데이터 분석

- 활용 데이터
- 분석 절차
- 데이터 전처리
- 클러스터링
- 다목적 최적화

3. 결론

- 최종 입지 선정
- 기대효과
- 방향 제시

2022 통계 데이터 분석 활용대회

01 분석 배경

주제 선정 배경

‘행복주택’이란?



- 정부의 보편적 주거복지 정책의 일환인 임대주택
- 청년계층(사회초년생, 신혼부부, 대학생)
- 저렴 & 양질의 주거 환경 제공이 목표



제공에 있어 현실적인 문제 발생

분석 필요성

행복주택의 부실 현황



문제점 1: 경제성 부실

- 행복주택의 보증금은 보통 4000~8000만 원, 임대료는 주변 시세의 60~80% 수준으로 비교적 저렴하긴하나 사회 초년생들에게는 여전히 부담스러운 가격
- 여기에 관리비는 월세에 비해 터무니없이 높음.
- 이 때문에 100:1의 경쟁률을 뚫고도 청년들의 계약률은 70%대에 그침

문제점 2: 접근성 부실

- 많은 행복주택들이 경제성만을 생각
→ 접근성이 좋지 않은 도시 외곽에 건설되어 있음.
- 상대적으로 학교 및 사업체와 떨어져 있는 도심 외곽에 행복주택 건설
→ 대중교통을 사용하는 청년층 입주민의 수요를 제대로 반영 X

도심 밖으로 밀린 행복주택, 관심 밖으로 밀려 입주 미달

입력: 2021-08-17 19:35:02 | 수정: 2021-08-17 19:36:12



[KNN주간시정]-외곽지역 몰린 행복주택 효과 있을까?

작성자 김 성기 등록일 2021년 9월 1일 조회수 1585 의견 0

[사설] 기피시설로 도심 외곽 밀려난 행복주택 현실

선행 연구 및 분석 전략

선행연구 소개



행복주택 최적 입지 선정에 관한 연구:
다목적 공간 최적화 접근
(진찬우, 이건학, 2015)

본 연구와의 차별성

선행연구의 분석 방식

- 대중교통 접근성이 우수하면서 저렴한 가격으로 공급할 수 있는 행복주택의 최적 입지를 제안
- 다목적 공간 최적화 모델에서 접근성과 경제적 비용을 고려

본연구의 선행연구의 차별성

- 경제성과 접근성에 지역 인프라 상황까지 고려를 한 다목적 최적화 모델을 만들어 차별성을 두고자 함.
- 2015년의 상황과 현 2022년의 상황은 다를 것이라 판단.



■ 최근 자료를 바탕으로 다양한 분석 툴(R, Python)을 사용하여 분석

선행 연구 및 분석 전략

프로젝트 분석 전략과 목표



“ 접근성, 경제성, 인프라를 최대한 골고루 만족하는
서울특별시 행복주택의 최적 입지를 분석 ”

→ 총 3개의 최종 입지 선정

2022 통계 데이터 분석 활용대회

02 데이터분석

활용 데이터 - 5개 영역에 대해 데이터 수집

	데이터명	출처	사용용도		데이터명	출처	사용용도
경제	(1) 서울시 주택가격지수(매매) 통계(txt)	서울 열린데이터 광장	서울시 자치구별 주택매매가격	교통	(1) 서울시 통근통학 및 주간인구 통계(txt)	서울 열린데이터 광장	통근통학 인구 비율
	(2) 서울시 자치구별 공시지가(csv)	서울 열린데이터 광장	서울시 자치구별 공시지가		(2) 서울시 자치구 별 정보(csv)	서울교통공사	서울시 자치구별 지하철역 개수
수요	(1) 서울시 대학교별 기숙사 수용 현황(xlsx)	서울 열린데이터 광장	서울시 대학교 재학생 수 및 기숙사 수용률		(3) 서울시 지하철 운행 현황 통계(txt)	서울 열린데이터 광장	서울시 지하철 호선별 평일 운행회수 및 수송 인원
	(2) 서울시 사업체수 데이터	MDIS(통계데이터센터 공공용 마이크로데이터)	서울시 내 자치구별 사업체수 및 합계종사자 수	인프라	(1) 서울시 시내주요기관(경찰, 소방관서) 통계(txt)	서울 열린데이터 광장	서울시 자치구별 경찰서, 지구대 개수
기타	(1) 전국 기존임대주택(행복주택) (csv)	한국 주택공사	서울시 자치구별 기존 행복주택 개수		(2) 서울시 자치구별 유통업체현황(xlsx)	kosis 국가통계포털	서울시 자치구별 대형 마트, 백화점 등 개수
	(2) 서울시 재정비 촉진구역 목록	서울시 누리집 공식 사이트	서울시 행복주택 입지 후보지		(3) 서울시 문화공간 정보(csv)	서울 열린데이터 광장	서울시 자치구별 문화공간 개수 분포
	(3) 행복주택 공급계획(csv)	서울주택도시공사	서울 특별시에 공급될 행복주택의 면적 및 세대수				

→ 모든 데이터는 각각의 기관의 데이터베이스 활용 지침을 준수하여 추출하였으며, 검증된 공식 통계 사이트에서 최신의 데이터를 사용하였다.

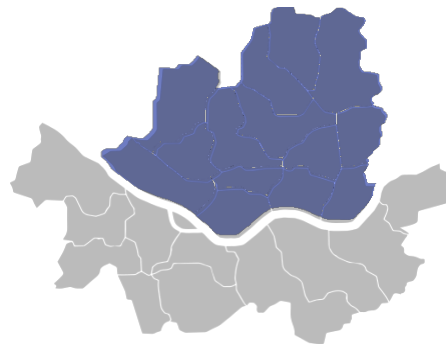
분석 절차

전처리

- 1) 결측치 & 이상치
- 2) Min/max Scaling
- 3) PCA
- 4) 지수생성



클러스터링



접근지수와 경제지수로 클러스터링



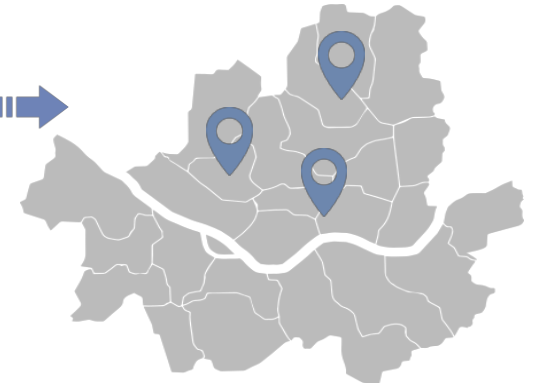
다목적 최적화

Maximize Z

3개의 자치구 추출

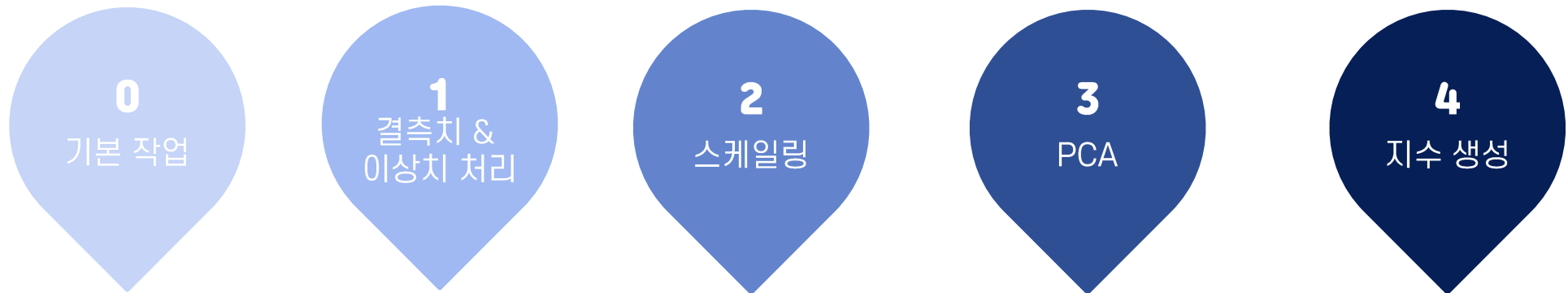


최종 입지 선정



전처리

전처리 과정



데이터에서 필요한 변수만을 추출

서울시 데이터 추출

변수명 변경(가독성)

중복 행 제거

변수 형태 변환 등

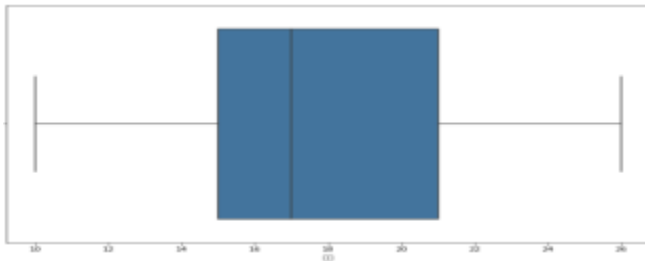
전처리

결측치 처리

→ 추가적인 데이터 조사를 통해 결측치를 제거하거나 보간

이상치 처리

→ **Boxplot** 을 이용하여 이상치를 확인하고 제거하거나 보간함.



ex) 경찰서 개수 데이터에서 극단적인 이상치가 보이지 않는다

파생변수 생성

대학교별 통학하는 대학생 수의 정도를 '수요량'이라는 새로운 변수 생성

$\text{수요량} = (\text{재학생수} \times (1 - \text{기숙사수용률}))$

최적 입지의 자치구를 선정하기 위하여 데이터의 단위를 **자치구로 변환**

전처리

최종 변수

	변수명	내용	자료형
1	자치구	서울 특별시 내 자치구명	object
2	재학생수	해당 자치구 소재 대학 합계재학생 수	int64
3	수용률	$((\text{기숙사수용인원수})/(\text{재학생수})) \times 100(\%)$	float64
4	수요량	$(\text{재학생수} \times (1 - \text{기숙사수용률}))$	int64
5	공공기관수	해당 자치구 내 도서관, 기타 공공기관 합계	int64
6	경찰서	해당 자치구 내 경찰서 합계	int64
7	쇼핑	해당 자치구 내 쇼핑공간 합계	int64
8	역개수	해당 자치구 내 역개수 합계	int64
9	평일운행횟수	해당 자치구 내 지하철역 평균 평일운행횟수	float64
10	수송인원	해당 자치구 내 지하철역 평균 하루 수송인원	float64
11	공지시가	단위 면적()당 토지 가격	float64
12	주택매매가격지수	2021.6월 기준 주택매매가격 변동률	float64
13	세대수	해당 자치구 소재 행복주택 세대수	int64
14	주택공급물량	해당 자치구 내 공급될 행복주택 물량	int64
15	사업체수	해당 자치구 소재 사업체 총개수	int64
16	종사자수	해당 자치구 소재 사업체 합계종사자 수	int64

전처리

MinMax Scaling

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

전체 변수의 척도가 다르기 때문에 스케일링 진행
변수 값을 0 ~ 1로 변환 (정규화)

PCA

고차원 데이터를 저차원의 데이터로 변환
차원의 저주 문제를 해소

결과 →

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
재학생수	0.194	0.509	0.059	0.109	0.141
수용률	0.177	0.443	0.086	-0.237	0.138
수요량	0.188	0.508	0.072	0.119	0.152
공공기관수	-0.251	0.149	0.238	0.248	-0.219
경찰서	-0.223	0.304	0.050	-0.081	0.200
쇼핑	-0.330	0.095	-0.171	0.282	0.275
역개수	-0.339	0.151	0.127	-0.124	0.091
평일운행횟수	-0.014	0.097	-0.623	-0.064	0.004
수송인원	-0.006	0.096	-0.626	-0.040	-0.020
공지시가	-0.383	0.153	0.005	0.190	-0.193
주택매매가격지수	-0.205	0.028	0.246	-0.558	-0.163
세대수	-0.153	-0.101	-0.068	-0.543	0.494
주택공급물량	-0.030	0.291	-0.171	-0.318	-0.668
사업체수	-0.424	0.034	-0.057	0.062	0.098
종사자수	-0.408	0.026	-0.044	0.037	-0.050

» 본 연구에서는 자체 생성한 지수로 충분하다 판단
따라서 PCA 차원 축소한 변수는 사용 X

전처리

지수 생성

접근성

수요량
종사자수
역개수
수송인원

경제성

공지시가
주택매매가격지수

인프라

공공기관수
경찰서
쇼핑

→ 각 **지수**에 해당하는 변수 설정

랜덤 포레스트

해당 지수들의 변수들을 독립변수로, 주택공급물량을 종속변수로 둔 랜덤 포레스트 모델 구축

랜덤포레스트의 **변수 중요도**를 사용 → 변수 가중치 설정의 근거로 사용

최종 지수식

접근지수 = $0.3 \times (\text{수요량}) + 0.3 \times (\text{종사자수}) + 0.1 \times (\text{역 개수}) + 0.3 \times (\text{수송인원})$

경제지수 = $0.6 \times (\text{공지시가}) + 0.4 \times (\text{주택매매가격지수})$

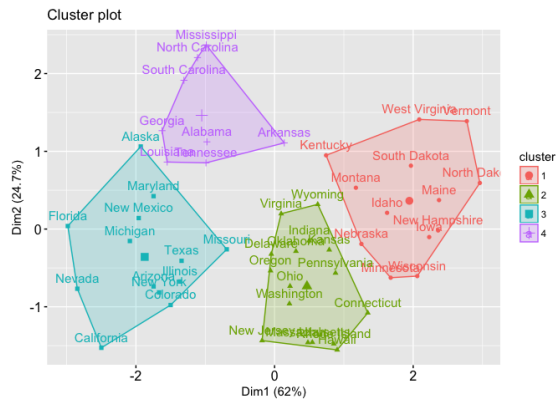
인프라지수 = $0.3 \times (\text{공공기관수}) + 0.3 \times (\text{경찰서}) + 0.4 \times (\text{쇼핑})$

클러스터링 - 분석 소개

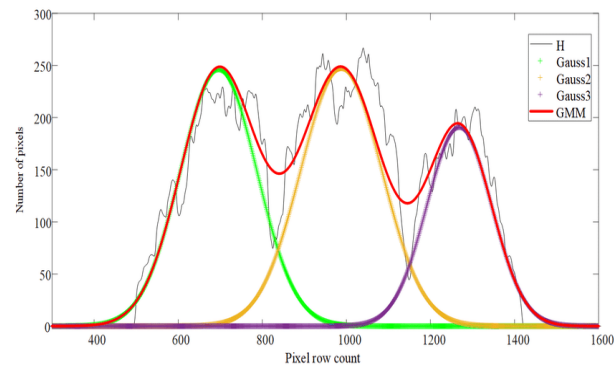
》》 각 데이터의 유사성을 측정하여 집단으로 분류하는 비지도 학습

x축을 **접근지수**로 y축을 **경제지수**로 두어 클러스터링 진행
접근지수가 크고 경제지수가 작은 군집을 찾는 것이 목표

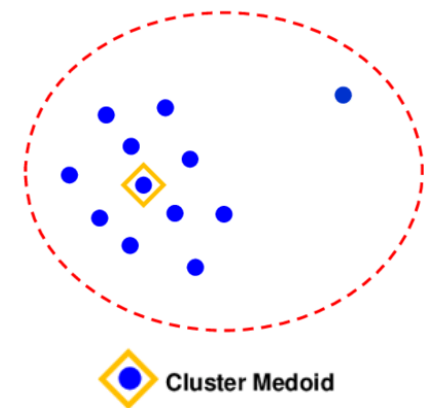
K - means



GMM



K - medoids

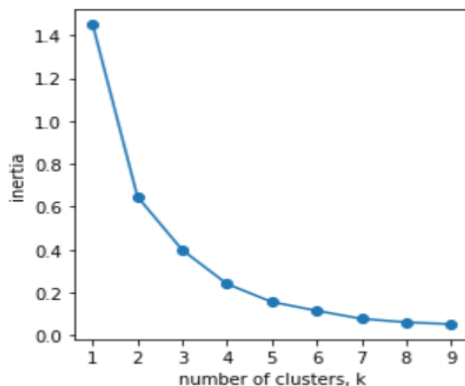


K-means 클러스터링

» K개의 중심점을 정하여 각 데이터로부터 그 데이터가 속한 군집의 중심까지의 평균거리를 최소화시키는 알고리즘

1. K개의 중심점을 정하기
2. 각 데이터들을 가장 가까운 중심점과 같은 그룹에 할당
3. 그룹의 중심점을 다시 업데이트
4. 중심점이 업데이트 되지 않을 때까지 반복

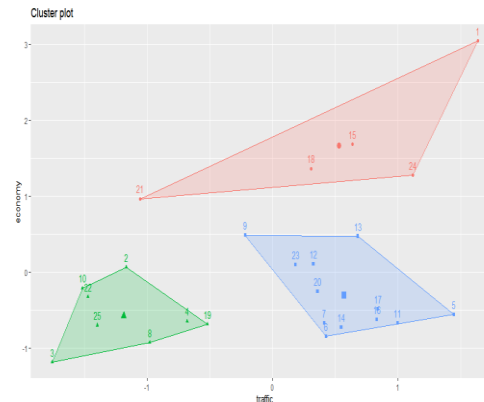
군집 수 k 정하기



→ Elbow Method 사용
K값을 3으로 결정

Elbow method란 군집 수에 따라 SSE의 변동을 비교하여 최적의 군집 수를 정하는 방법

결과



접근지수가 높고 경제지수가 낮은 클러스터는 생성되었으나
각 클러스터에 속한 데이터의 개수가 12개, 5개, 8개로
클래스 불균형이 심하다고 판단
→ 따라서 K-means 방법을 기각

GMM 클러스터링

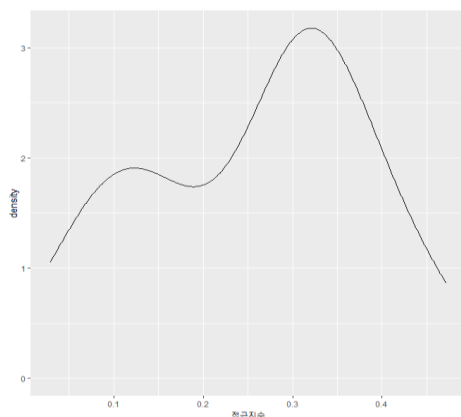
» 데이터가 여러 개의 **가우시안 분포**를 가진 데이터 집합들이 섞여서 생성되었다는 가정하에 군집화를 수행하는 방식

1. 각 정규분포에 데이터 포인트들을 임의로 할당

2. 개별 데이터 포인트들의 가능도(likelihood)를 계산하여 가장 높은 확률을 가진 정규분포에 할당

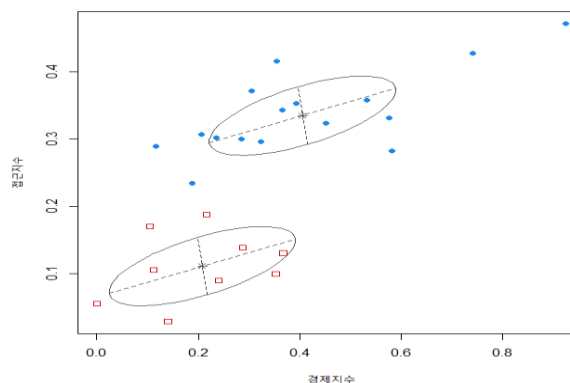
3. Maximum Likelihood Estimation으로 모평균과 모분산을 추정

군집 수 k 정하기



→ 봉우리가 2개이므로 k를 2로 결정

결과



접근지수가 높고 경제지수가 낮은 클러스터가 뚜렷하게 생성되지 않음

→ 따라서 **GMM** 방법을 기각

K-medoids 클러스터링

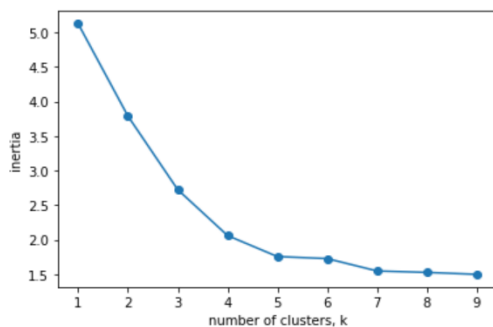
» K-means 방법을 변형한 것으로 군집의 중심점을 데이터의 평균이 아닌 중앙점을 사용하여 클러스터링을 진행

1. K-means와 작동원리는 동일

2. 중앙점을 사용하기 때문에 k-means보다 이상치로부터 덜 민감함

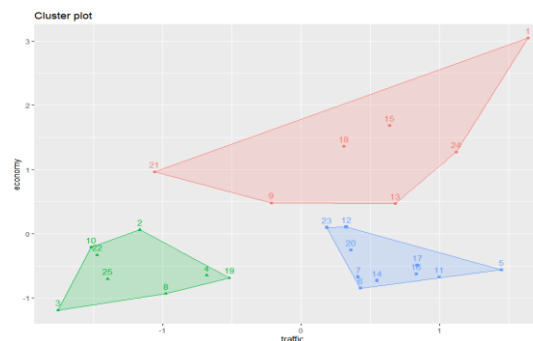
3. 데이터의 크기가 작을 때 적용하기 유리

군집 수 k 정하기



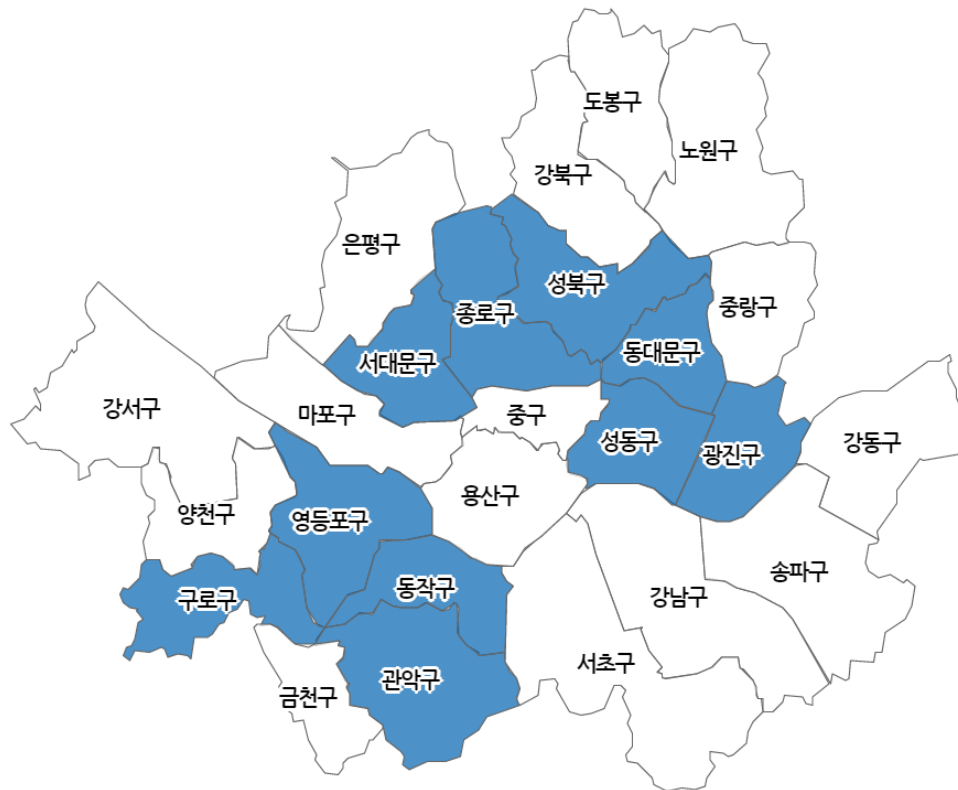
→ Elbow Method 사용
K값을 3으로 결정

결과



각 군집에 속한 데이터의 개수가 7개, 8개, 10개로 클래스 불균형이 심하지 않고 접근지수가 높고 경제지수가 낮은 클러스터가 만들어짐
→ 따라서 K-medoids 채택

클러스터링 결과



선정된 자치구 : 관악구, 광진구, 구로구, 동대문구, 동작구,
서대문구, 성동구, 성북구, 영등포구, 종로구

» K-medoids 클러스터링 결과로
접근지수가 높고 경제지수가 낮은
10개의 자치구가 선택되었음

다목적 최적화

목적함수

$Maximize Z_1 = 0.3 \times (\text{수요량}) + 0.3 \times (\text{종사자수}) + 0.1 \times (\text{역 개수}) + 0.3 \times (\text{수송인원})$ 접근성 지수

$Minimize Z_2 = 0.6 \times (\text{공지시가}) + 0.4 \times (\text{주택매매가격지수})$ ⇒ 경제성 지수

$Maximize Z_3 = 0.3 \times (\text{공공기관수}) + 0.3 \times (\text{경찰서}) + 0.4 \times (\text{쇼핑})$ 인프라 지수

최종 식

$$Maximize Z = a \times Z_1 + b \times (1 - Z_2) + c \times Z_3$$

- 1) a, b, c는 각 목적함수의 가중치로 단위는 0.1이다.
- 2) $a + b + c = 1$
- 3) $a, b > c$

행복주택의 주요 목적인 접근성(a)과 경제성(b)은 주변 인프라(c)보다 더 중요한 요소임

랜덤 포레스트

가중치 a, b, c의 값을 설정하기 위해 독립변수로 접근성, 경제성, 인프라 지수를,
종속변수로 주택공급물량을 두어 변수중요도 탐색

변수중요도 : 접근성 > 경제성 > 인프라

다목적 최적화

가능한 경우

(1) $a = 0.5, b = 0.3, c = 0.2$

채택

(2) $a = 0.4, b = 0.4, c = 0.2$

→ 다목적 최적화 결과, 두 경우 모두 동일하게

동대문구, 관악구, 성북구가 최종 자치구 후보지로 선정

→ 따라서 다목적 최적화 결괏값이 더 큰 경우인

(2) $a = 0.4, b = 0.4, c = 0.2$ 를 선택함

최종 다목적 최적화 식

$$\text{Maximize } Z = 0.4 \times Z_1 + 0.4 \times (1 - Z_2) + 0.2 \times Z_3$$

자치구	(1)	(2)
관악구	0.522	0.561
광진구	0.442	0.449
구로구	0.481	0.534
동대문구	0.530	0.577
동작구	0.393	0.432
서대문구	0.418	0.471
성동구	0.458	0.505
성북구	0.512	0.557
영등포구	0.428	0.474
종로구	0.458	0.499

다목적 최적화

서울특별시 주거지형 재정비촉진지구 목록에서
동대문구, 관악구, 성북구에 해당하는 재정비 촉진지구를 각각 하나씩 선정

장위 재정비촉진지구

- 성북구 장위동 일원

신림 재정비촉진지구

- 관악구 신림동 일원



전농·답십리 재정비촉진지구

- 동대문구 전농 1, 4동
- 답십리 1, 3, 5동

2022 통계 데이터 분석 활용대회

03

결론

최종 입지 선정

각 재정비촉진지구의 지역 특성과 주변 환경 그리고 구역별 재개발 진행 단계를 조사하여 최종 입지 선택

전농·답십리 재정비촉진지구 : 전농8 구역



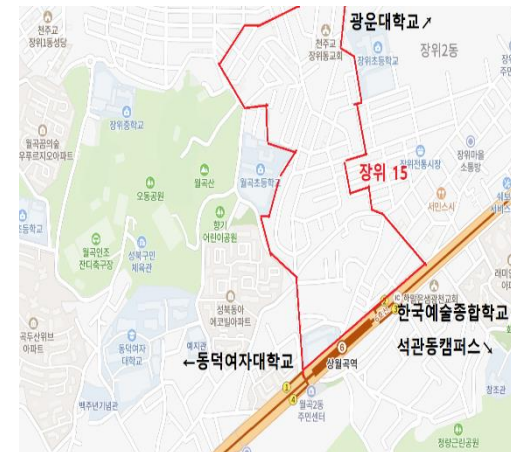
- [교통] 서울 동부의 주요 교통 거점인 청량리역과 가까움
- [환경] 인근에 3곳의 공원이 있음
- [학교] 서울시립대학교와 가까움
- [기타] 전농로타리시장이 있고 주변 상권이 관참음

신림 재정비촉진지구 : 신림4 구역



- [교통] 도보로 10분 거리에 서울대벤처타운역(신림선)이 있음
- [환경] 하측에 돌산이 붙어있고, 근처에 어린이공원이 2곳 있음
- [학교] 미림여고와 신우초 사이에 위치함

장위 재정비촉진지구 : 장위15 구역



- [교통] 북부간선도로가 있고 상월곡역(6호선)과 가까움
- [환경] 월곡산이 있음
- [학교] 광운대, 동덕여대, 한국예술종합학교 (석관동 캠퍼스)와 가까움
- [기타] 장위전통시장과 가까움

기대효과



저렴한 가격

월 임대료 6만원대 ~ 16만원대

**편리한
대중교통**

평균 63%의 출근 & 통학 인구
-KOSIS 국가통계포털-

삶의 질 향상

주변 편의시설, 치안, 환경, 학교까지 고려

- 행복주택의 주요 수요자인 대학생과 사회초년생들의 경제적인 부담을 덜 수 있음
- 접근성의 한계에서도 벗어나 출근과 통학에도 어려움이 없음
- 주변 환경까지 고려한 입지 선정으로 삶의 질 향상까지 노릴 수 있음

방향제시



접근성, 경제성, 주변 인프라까지 고려한 최적화된 입지대안



공간만 따지는 비효율적인 방법 X
실제 행복주택이 가지고 있던 문제점 해결 가능한 현실적 대안

다양한 현실적 상황을 고려해야 할 때 적절히 활용 가능

참고문헌

김지현, 강수진(2020), 「행복주택 성과지표 개선방안에 대한 연구 = A Study on the Improvement of Performance Indicators for Happy Housing Project」, 한국주택학회

민지현, 박하연, 정하은, 정희진, Zheng Haihua(2020), 「청년 공공임대주택의 문제점과 개선방안에 대한 연구」,
梨花女子大學校 法政大學 行政學科

진찬우, 이건학(2015), 「행복주택 최적 입지 선정에 관한 연구: 다목적 공간 최적화 접근」, 한국도시지리학회지 제18권
2호(81~95), 한국도시지리학회

이경운 기자, <대학생·청년·신혼부부 내 집 마련 '파주법원 행복주택' 추가공급>, [대학생·청년·신혼부부 내 집 마련 '파주법원
행복주택' 추가공급 - 국토일보 \(ikld.kr\)](http://ikld.kr)

2022년 통계데이터 분석·활용대회 데이터 분석

감사합니다

서울특별시 행복주택 **최적 입지 선정**