2018 서울시 빅데이터 캠퍼스 공모전

서울시 국공립 어린이집 최적 입지 찾기!

팀명: 빅캠아이(BigCami)

팀원: 신동걸, 임희진, 전현성, 최유리

사진출처: https://news.joins.com/article/21530904

CONTENTS

프로젝트 소개

프로젝트 배경 및 데이터 수집, 분석 방법 소개

1-1. 프로젝트 배경

WHY

국공립 어린이집인가?

선진국의 국공립 어린이집 비율은
얼마나 될까?

출산율과 가임기 여성 인구를
반영 해야할까?

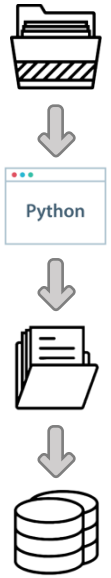
소외 지역 공공시설의
유휴공간을 활용하면 어떨까?


용산구립 샘물어린이집(성심여중·고 도서관 내)

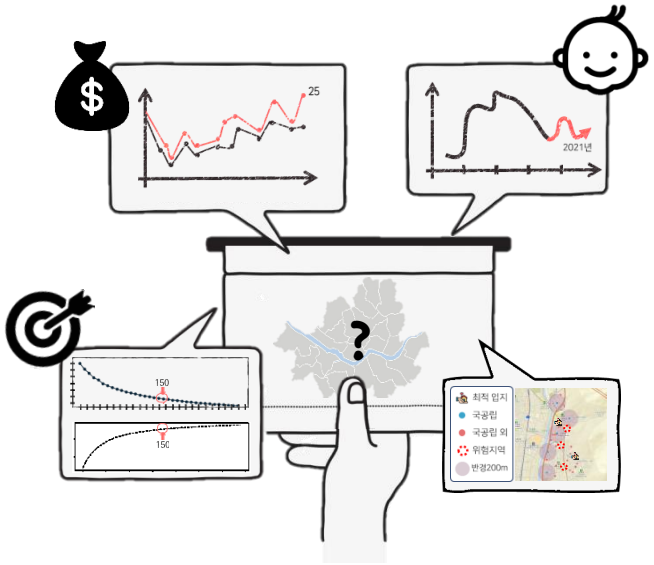
잠재 이용 아동 인구와 접근성 취약 지역을 분석하여
보다 시급한 지역의 유휴시설을 활용하여 국공립 어린이집을 확충할 필요가 있음

1-2. 분석 과정

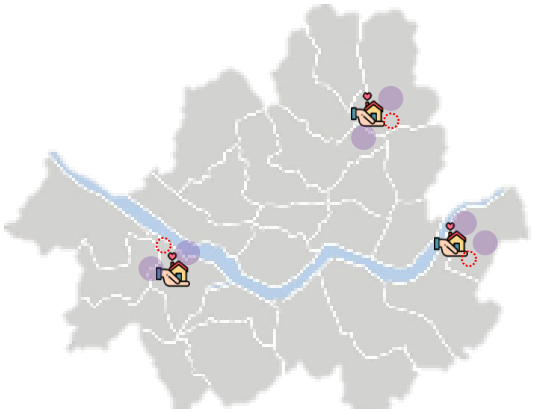
데이터 수집 및 저장



데이터 가공



머신러닝 및 딥러닝 활용
데이터 분석



GIS 분석을 통한
최적지 선정

1-3. 데이터 소개

	이용 데이터셋 목록	데이터 기간	활용 목적	출처
1	서울시 보육시설(어린이집) 현황	2018년 8월	GIS분석, 이용자 접근성, 시설 접근성 분석	서울시 빅데이터 캠퍼스
2	거주인구(행정동)	2010년 1월 - 2018년 9월	인구 추정	서울 열린 데이터 광장
3	거주인구(행정동 추가분)			행정안전부
4	주민등록 인구통계			
5	표준지 공시지가	2016년 1월 - 2018년 8월	부동산데이터 활용 소득 추정	국토교통부
6	주택 유형별 매매 실거래 자료(오픈API)			
7	주택 유형별 전월세 자료(오픈API)			
8	생활안전지도	2016년 1월- 2017년 6월	어린이집 최적지 선정 시 치안 환경 반영	국립재난안전연구원

※ 제외된 data set

1. 행정동별 소득
   - 개인정보보호법에 따라 데이터 접근이 제한됨
   - 실거래가 및 공시지가 데이터를 활용
   - 전월세 데이터 추후 보완(feature 추가)

2. 어린이집 대기 인원 수
   - Social Mining의 한계
   - 어린이집 정원데이터로 보완

3. 가임기 여성 및 기혼자 인구, 기혼자 교육 수준
   - 현재 사회 현상을 반영하는데 있어 한계가 존재

4. 주택 평수
   - 평당 가격을 구하고 표준화 스케일링(scaling)하여 사용

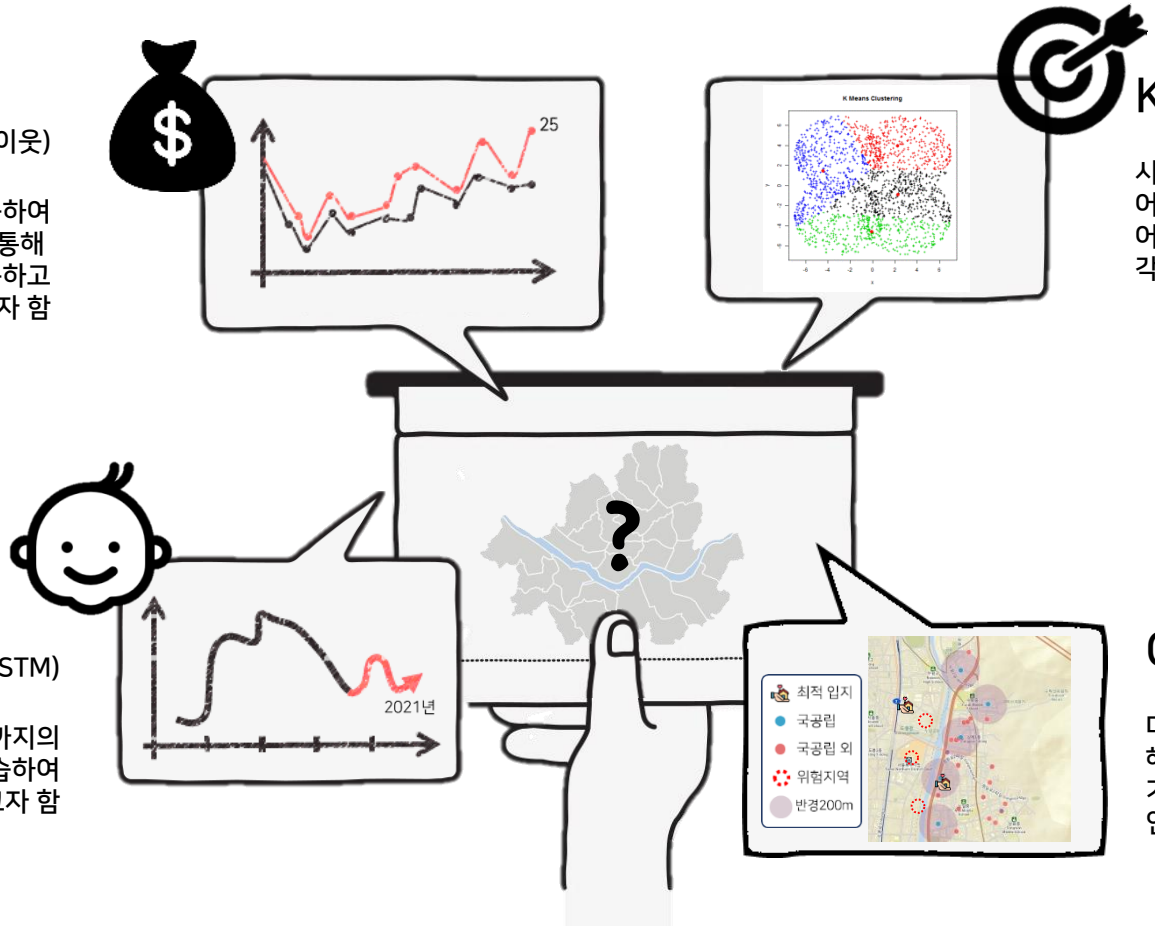
1-4. 주요 분석 기법 소개

KNN(K-최근접 이웃)

부동산의 위도 및 경도와 부동산 가격을 학습하여
어린이집 위도 및 경도 데이터를 통해
어린이집 인근 지역의 부동산 평균 가격을 예측하고
저소득 지역을 추출하고자 함

RNN LSTM(순환 신경망 LSTM)

2010년~2018년 9월까지의
영유아 인구 데이터를 학습하여
2021년의 행정동별 영유아 인구를 예측하고자 함



K-means clustering(K-평균 군집화)

시설 및 이용자 접근성 분석의 독립변수를 추출하기 위해
어린이집 위도 및 경도 데이터를 학습하여
어린이집 그룹(클러스터)의 centroid(중심점)과
각 어린이집의 거리를 구하고자 함

GIS 공간 분석

머신러닝과 딥러닝 모델을 통해 소외된 행정동을 파악하고
해당 행정동 내의 치안 위험 지역과
기존 국공립 어린이집 반경 200m를 고려하여
인근지역의 공공시설을 국공립 어린이집 최적지로 추천하고자 함

머신러닝 및 딥러닝, GIS 공간 분석 등을 활용하여 국공립 어린이집의 최적지를 발견하고자 함

머신러닝 기반 데이터 분석

소득 추정, 시설 접근성, 이용자 접근성, 인구 예측 분석

2-1. 머신러닝 기반 지역 소득 분석

실거래가, 공시지가 데이터를 활용한 지역 소득 분석(KNN)

- 아파트 및 연립주택의 부동산 데이터 표준화(약 40만 개)
- 부동산 데이터의 위도 및 경도를 KNN Regressor 알고리즘으로 클러스터 생성함.
- 어린이집을 해당 클러스터링에 보내 각 클러스터마다의 공시지가 평균값을 인근 지역 소득으로 추정하였음.

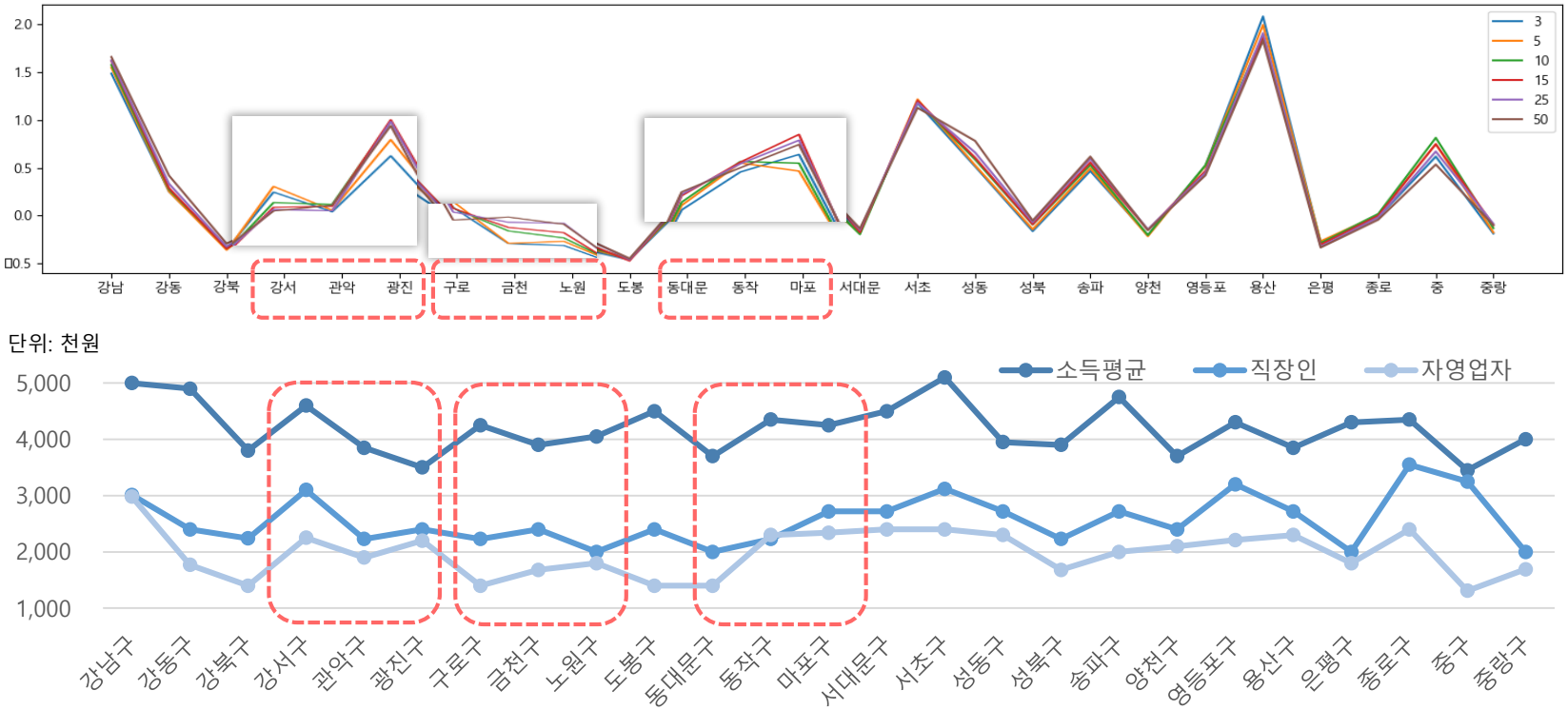


그림1. KNN 클러스터의 K값 비교 그래프(위: Python 분석 결과, 아래: 설명 참조)

[최적의 K값 찾기]

KNN's K = 각 클러스터 내부의 요소의 개수

최적의 K값을 찾기 위해 3, 5, ~ 50까지의 값을 적용하고 구 별 평균을 구하였음.

대부분의 지역 평균은 K값에 큰 영향을 받지 않았으나, 3개 구간에서 소득 평균의 순위가 크게 바뀌는 결과가 나왔음.

신한은행(2018년) 및 서울서베이(2016년)의 서울시 구별 소득자료를 참고하여 실제 값과 가장 근사한 K값을 선택하였음.

[K-NN(nearest neighbor)]

관측값마다의 거리를 기반으로 한 기계학습 클러스터링 알고리즘임 (K = 각 클러스터 내부의 구성원 수)

2-2. 딥러닝 기반 아동 인구 예측

미래의 수요와 개선 기간을 반영한 인구 추정 (RNN LSTM; RNN Long Short-Term Memory)

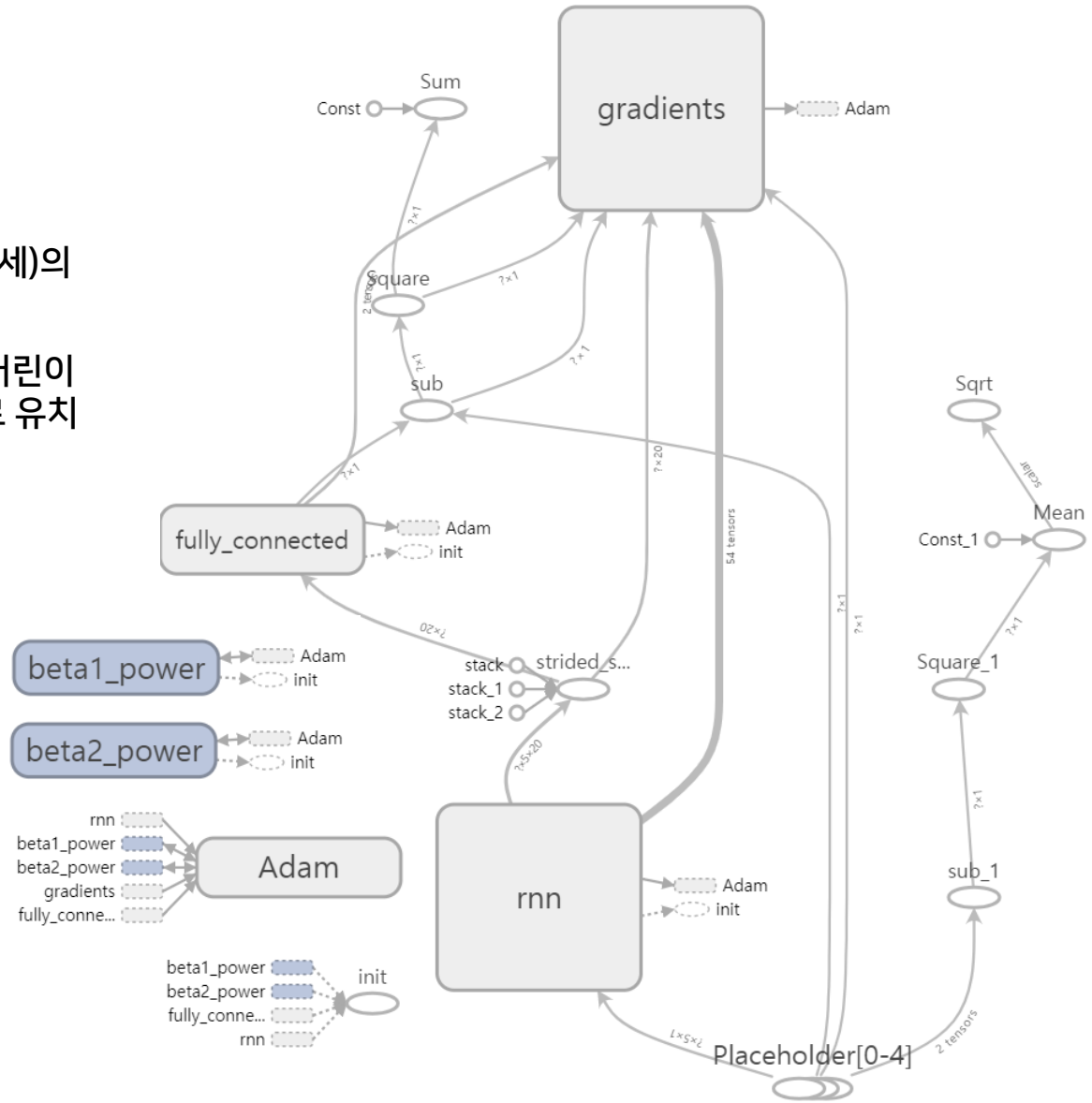
- 서울시 행정동의 2010년 1월 - 2018년 9월까지의 월별 아동 인구 수(만0-5세)의 시계열 패턴을 분석, 2021년 인구 추정하였음
- 어린이집을 이용하는 아동의 연령에 대한 범위는 선행연구를 참고하였으며, 어린이집 이용 아동들의 연령은 평균 생후 52개월 경으로 나타나, 만 5세부터는 주로 유치원을 많이 이용하는 것으로 봄
- 시계열 분석을 위해 딥러닝 알고리즘 중 RNN LSTM을 적용하여 모델 생성함
  - ✓ RNN LSTM 시퀀스 랭스(7): 연관 데이터 길이
  - ✓ 데이터 개수의 한계로 은닉 계층이 넓은 딥러닝 모델 구성

[RNN – LSTM]

과거의 데이터를 연결해 처리하는 순환 신경망

Memory cell이 존재해 시계열 분석에 유리한 딥러닝 알고리즘

그림2. RNN Tensor Board



2-2. 딥러닝 기반 아동 인구 예측

최종 후보지 중 한 곳인 오류 제2동의 2021년 인구 추이 예측

- LSTM 모델의 RMSE(Root Mean Squared Error)

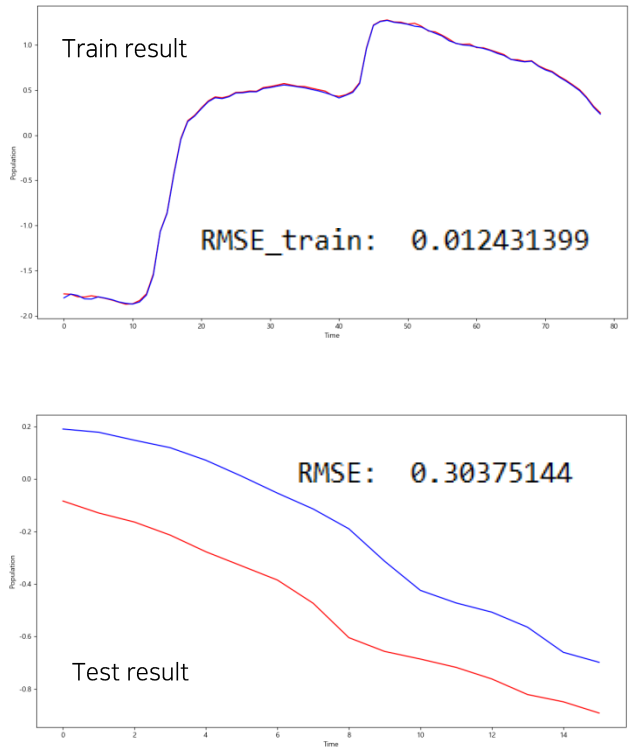


그림3. 오류 제2동 2021년 인구 예측 훈련 및 테스트 결과

그림4. 오류 제2동 2021년 인구 추이 예측

2-3. 머신러닝 기반 시설 접근성 분석

각 클러스터의 중심 위치와 어린이집 위치를 반영한 시설 접근성 분석(K-means)

- 위도, 경도의 위치기반으로 어린이집의 K-means 클러스터 생성하여 각 클러스터의 중심 위치와 어린이집 위치로 시설 접근성 분석함.
- K-means의 K값은 전체 클러스터의 개수를 의미하며, K값에 따라 각 클러스터 간의 교차 범위가 달라짐. [그림5]를 통해 K값이 150으로 귀결됨을 확인 할 수 있음.

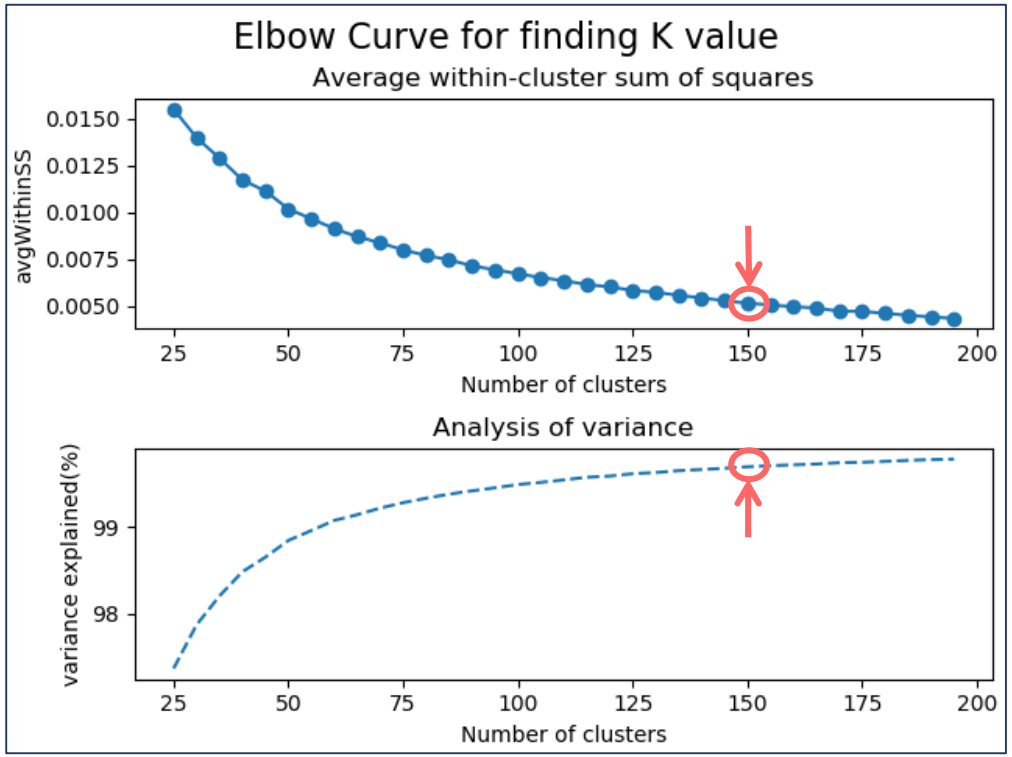


그림5. 최적 K값 찾기 - 엘보그래프

[최적의 K값 찾기- 엘보그래프]

최적의 K값 - K : 전체 클러스터의 개수

분산과 각 클러스터 간에 완만해지는 거리(클러스터 간의 교차 범위, k=150)

[K-means]

K(지정된 클러스터의 개수)에 따른 거리 기반의 클러스터링 비지도 학습으로 대표적인 기계학습 알고리즘임

2-3. 머신러닝 기반 시설 접근성 분석

각 클러스터의 중심 위치와 어린이집 위치를 반영한 시설 접근성 분석(K-means)

- 서울시 자치구의 개수(k=25)로 클러스터를 별도로 구성하여 비교 데이터로 활용

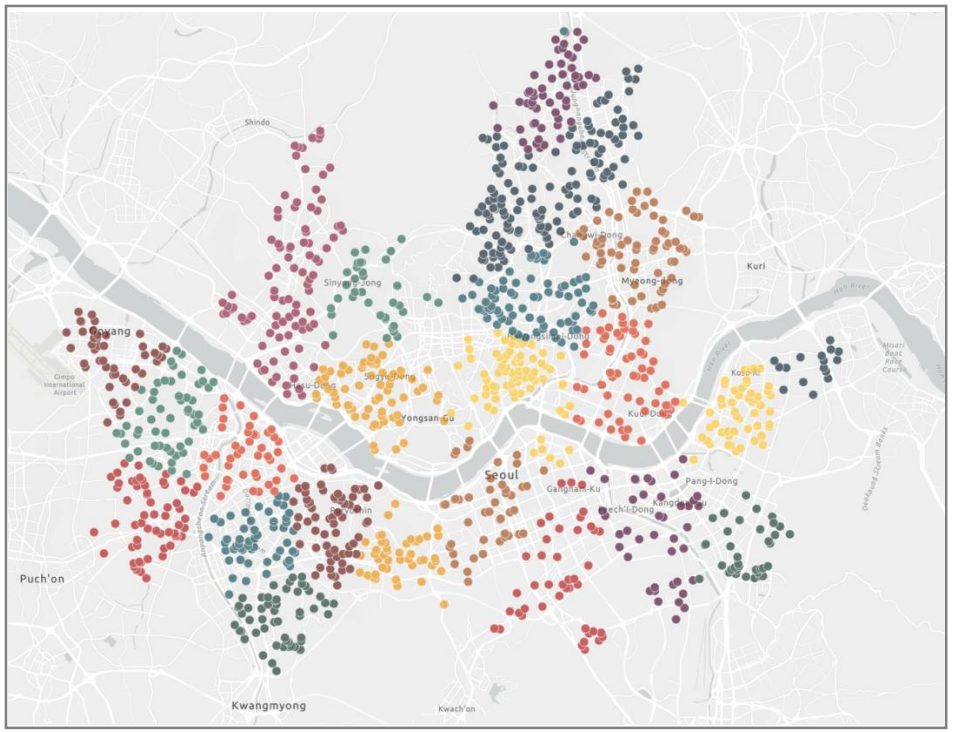


그림6. K값에 따른 클러스터 구성 (K=25)

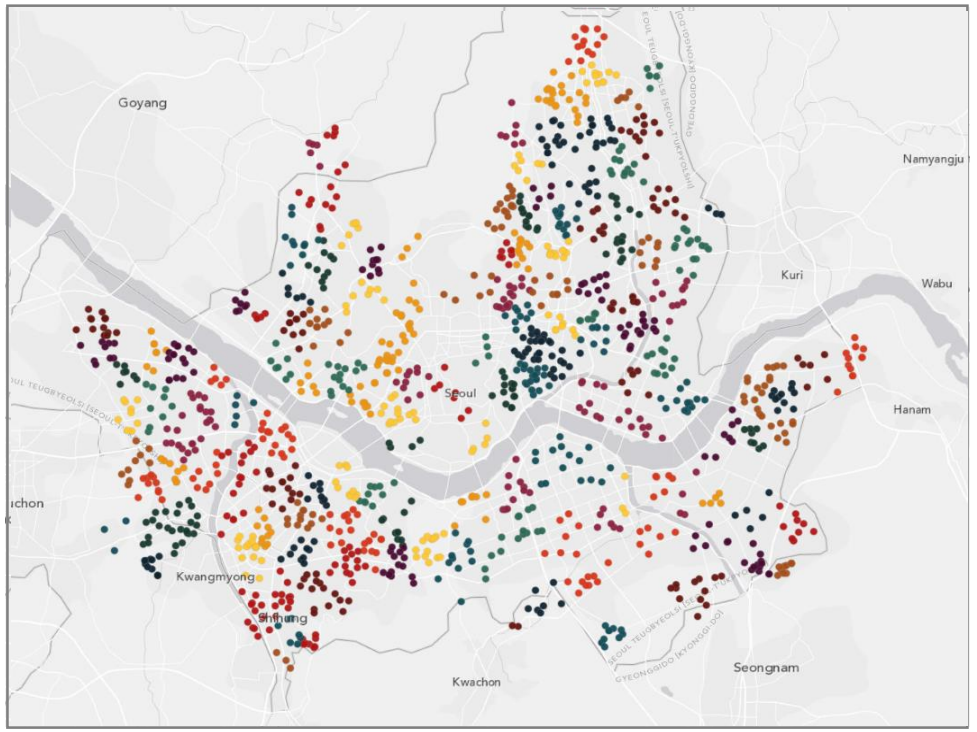


그림7. K값에 따른 클러스터 구성 (K=150)

2-3. 머신러닝 기반 시설 접근성 분석

각 클러스터의 중심 위치와 어린이집 위치를 반영한 시설 접근성 분석(K-means)

- 어린이집의 잠재적 이용 수요를 예측하기 위해 각 클러스터 간의 거리와 어린이집 간의 거리를 분석 (단, 위도 및 경도는 km로 변환)
- 행정동별 아동 추정 인구(만0-5세) 수를 추출하여 해당 인구가 거주하고 있는 인근 어린이집의 접근성이 평균값 이하인 지역을 추출

$$P_j = \sum_{i=1}^{m} E_i \, d_{ij}^{-\beta}$$

시설 접근성 분석 위한 수식

평균값 이상의 만 5세 이하 인구 + 평균값 이하의 접근성을 가지는 지역
P_j = 공립 보육시설 j의 접근성
E_i = i 클러스터에 거주하는 영유아 중 보육시설을 이용할 것으로 추정되는 인구
d_ij = i 클러스터의 도형 중심점과 보육시설 j 간의 거리
β = 거리마찰계수(1의 값 적용)
m = 총 클러스터 수
접근성 지수가 높으면 높을수록 보다 많은 인구가 해당 시설로 공간상에서 접근이 용이함을 의미함

2-4. 머신러닝 기반 이용자 접근성 분석

시설 접근성과 이용자 밀도의 비교 분석을 반영한 이용자 접근성 분석(K-means)

- 어린이집의 이용자 접근성을 예측하기 위해 각 어린이집의 이용 정원수와 각 클러스터 중심과의 거리(km)의 곱을 시설 접근성으로 나눔. (단, 위도 및 경도는 km로 변환)
- 특정 시설이 이용자에게 매우 가깝다고 하더라도 만약 그 시설에 극도로 많은 이용자가 몰린다면 그 시설은 이용자에게 접근성이 높은 시설이라고 말하기 어려움. 공급 용량은 이용자 거주지역에서 시설까지의 거리 값을 이용해 조정함.

$$A_i = \sum_{j=1}^{n} \frac{C_j d_{ij}^{-\beta}}{\sum_{k=1}^{m} E_k d_{kj}^{-\beta}}$$

이용자 접근성 분석 위한 수식

A_i = i 클러스터에 거주하는 만 5세 이하 인구의 공립 보육시설로의 평균 접근성
C_j = 공립 보육시설 j의 이용 정원 (시설의 규모 반영)
E_k = k 클러스터에 거주하는 만 5세 이하의 인구
m = 총 클러스터 수
n = 공립 보육시설의 총수

2-4. 머신러닝 기반 이용자 접근성 분석

각 클러스터의 중심 위치와 어린이집 위치를 반영한 이용자 접근성 분석(K클러스터)

- 접근성 지수가 낮을 수록 취약 지역임.

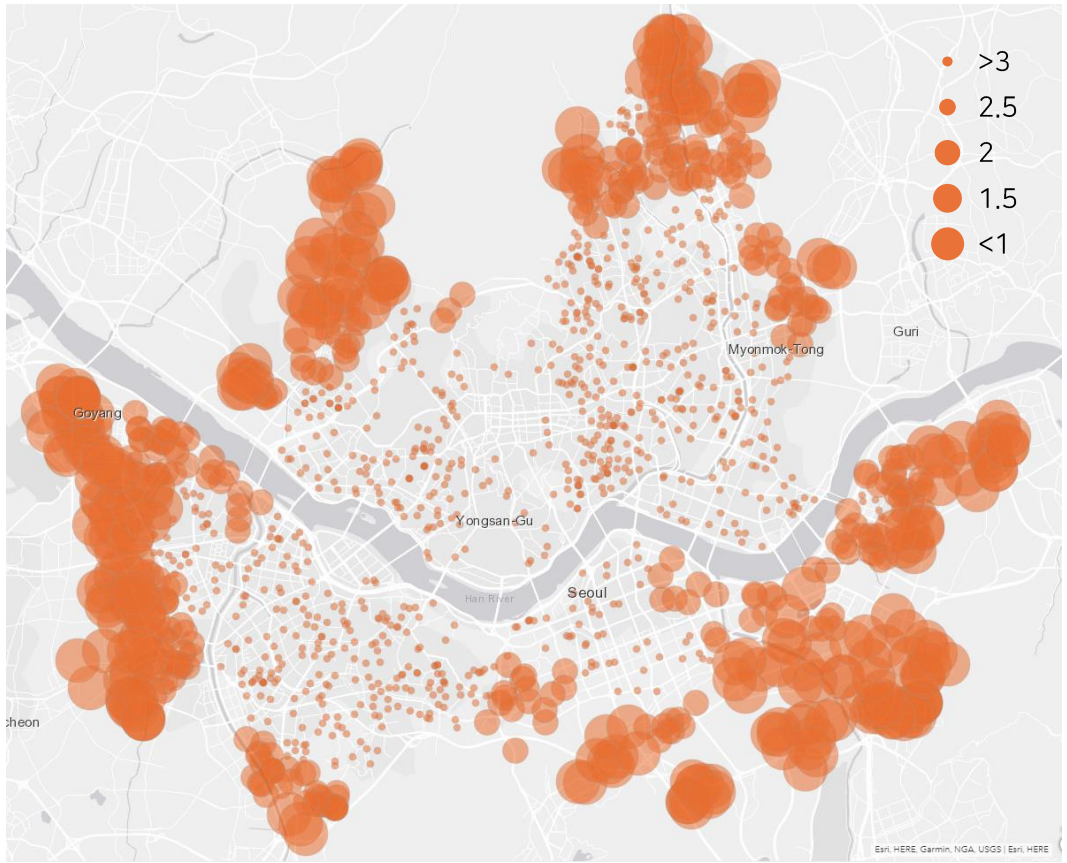


그림8. 이용자 접근성 지수 5분위 (K=150, reversed)

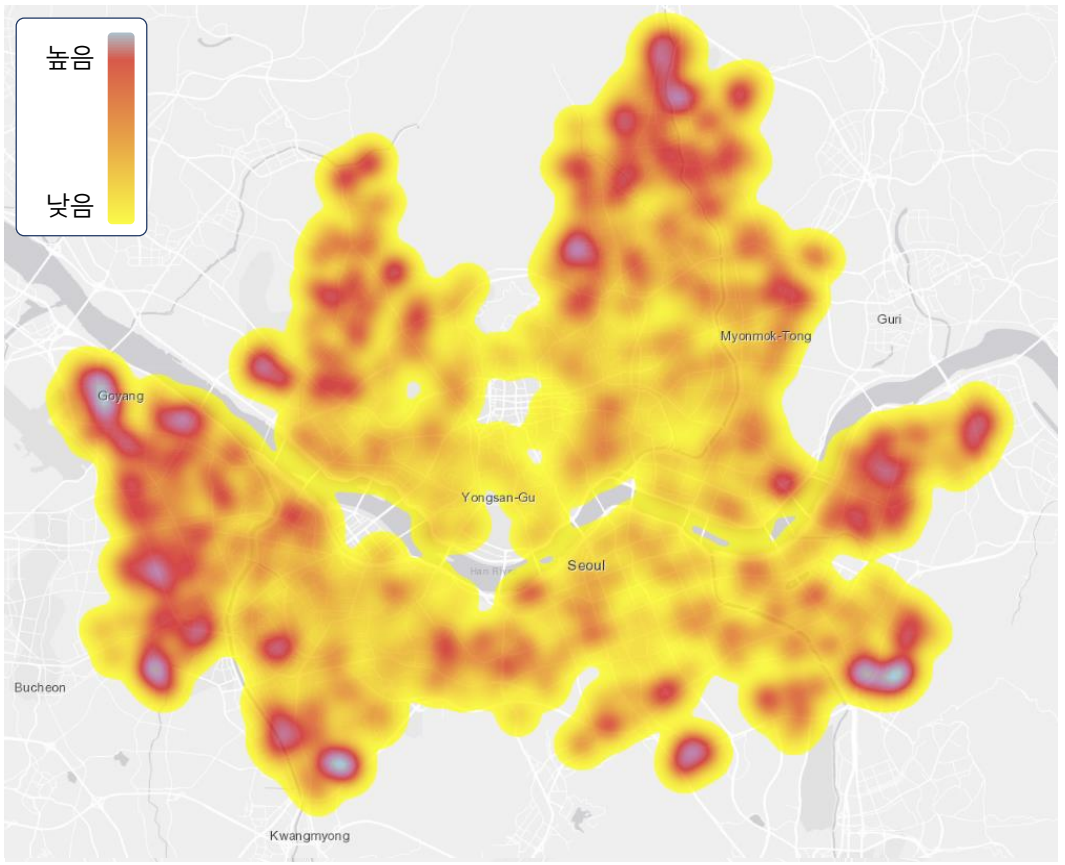


그림9. 이용자 접근성 지수 5분위 히트맵 (K=150, reversed)

2-5. 시설 접근성, 이용자 접근성, 추정 소득 종합 분석

관련 변수의 분석 결과 및 종합 분석

- 인구밀도가 높고 소득이 적은 지역(시설 접근성 및 이용자접근성 지수가 가장 낮은 지역)을 국공립 어린이집이 필요한 지역으로 해석함.
- 본 연구는 추정 소득 4분위, 현재 아동 인구 기준의 시설 접근성(행정동별 분위), 이용자 접근성(클러스터별 분위)를 분석함.

표1. 소외지역 어린이집 분위별 수치

구분	소득 분위	시설(현재)	이용자(현재)	시설(2021년)	이용자(2021년)
봄빛	1	0.007	0.020	0.002	0.020
상일	1	0.007	0.020	0.002	0.020
늘사랑	0	0.002	0.087	0.004	0.053
노원	0	0.002	0.087	0.004	0.053

* 0~1 사이의 값으로 정규화(MinMax Scaler)함. 낮을 수록 취약한 지역을 의미

표2. 관련 변수

관련 변수
추정 소득(공시지가, 실거래가)
추정 소득(전월세)
2018년 아동 인구(만0-5세)
2021년 추정 아동 인구(만0-5세)
국공립 어린이집 시설 접근성
국공립 어린이집 이용자 접근성

2-5. 전월세 데이터 보완 후 분석 결과

관련 변수의 분석 결과 및 종합 분석

- 소외 지역으로 선정한 5개 지역의 소득, 시설 접근성 및 이용자접근성(현재, 2021년 인구 예측) 등을 바탕으로 가장 취약한 어린이집을 추출함
- 주성분 분석(PCA)의 설명력이 가장 높은 지역 중 아래의 표와 같이 수치들을 비교한 결과 오류 제2동의 어린이집이 가장 소외지역으로 분석됨

표3. 소외 어린이집 후보군의 전월세 데이터를 포함한 변수 수치

	소득분위	전월세	시설(현재)	이용자(현재)	시설(2021년)	이용자(2021년)
항동어린이집	-0.424	-0.850	18.084	0.343	18.242	0.344
궁동어린이집	-0.451	-0.856	20.437	0.343	20.636	0.344

표4. 각 어린이집의 PCA 수치 비교

이름	PCA	Rank
도봉1동어린이집	0.068432	0.000692
궁동어린이집	0.081093	0.001383
항동어린이집	0.082988	0.002075
구립 한누리어린이집	0.085596	0.002766
도봉2동오감발달베이비마을어린이집	0.086008	0.003458
구립 꼬마대통령어린이집	0.086882	0.004149
구립 구파발어린이집	0.087983	0.004841
구립 진관어린이집	0.090093	0.005533
구립 기자촌어린이집	0.099242	0.006224
구립 큰솔어린이집	0.102182	0.006916

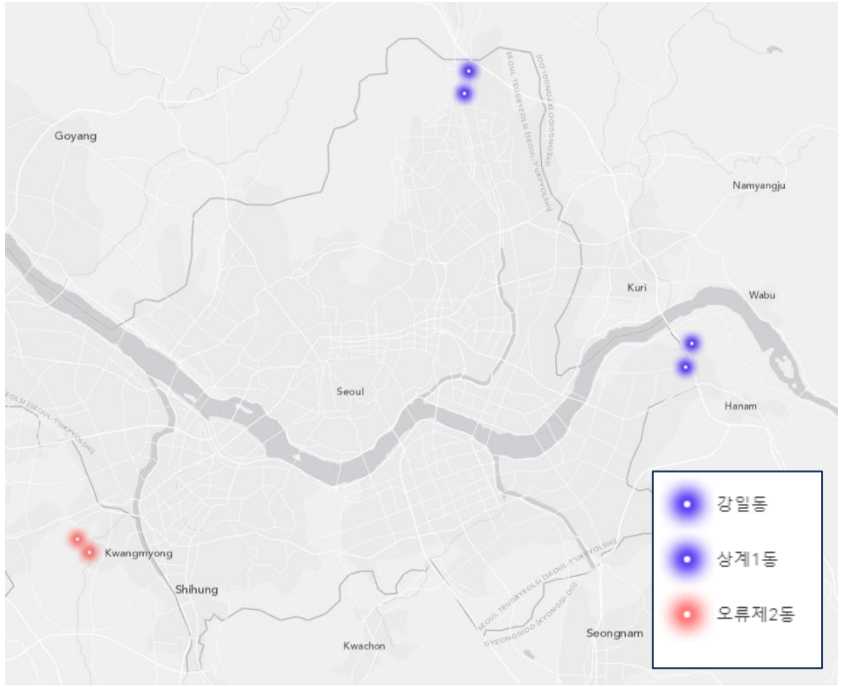


그림10. 서울시 국공립 어린이집 소외지역

2-5. 머신 러닝 및 딥러닝 기반 데이터 분석 결과

소외지역 후보(1): 강서구 강일동

	시설(현재)	이용자(현재)	시설(2021년)	이용자(2021년)
현재	0.677	0.187	0.606	0.181
개선 후	0.841	0.281	0.828	0.281

* 0~1 사이의 값으로 정규화(MinMax Scaler)함.

강서구 강일동 내 현재 어린이집 정원 각 30명씩 증원(330명),
국공립 어린이집 10개소(총 900명)를 증설할 때

전체 행정동의 하위 3.17%에서
하위 12.68% 수준까지 상승

소외지역 후보(2): 노원구 상계1동

	시설(현재)	이용자(현재)	시설(2021년)	이용자(2021년)
현재	0.002	0.080	0.004	0.050
개선 후	0.017	0.360	0.017	0.360

* 0~1 사이의 값으로 정규화(MinMax Scaler)함.

노원구 상계1동 내 현재 어린이집 정원 각 30명씩 증원(150명),
국공립 어린이집 7개소(총 700명)를 증설할 때

전체 행정동의 하위 0.97%에서
하위 11.46% 수준까지 상승

소외지역 후보(3): 구로구 오류 제2동

	PCA 분석 수치 결과
현재	0.082987894
개선 후	0.212323899

* 0~1 사이의 값으로 정규화(MinMax Scaler)함.

구로구 오류제2동 내 현재 어린이집 정원 각 30명씩 증원(300명),
국공립 어린이집 10개소(총 1600명)를 증설할 때

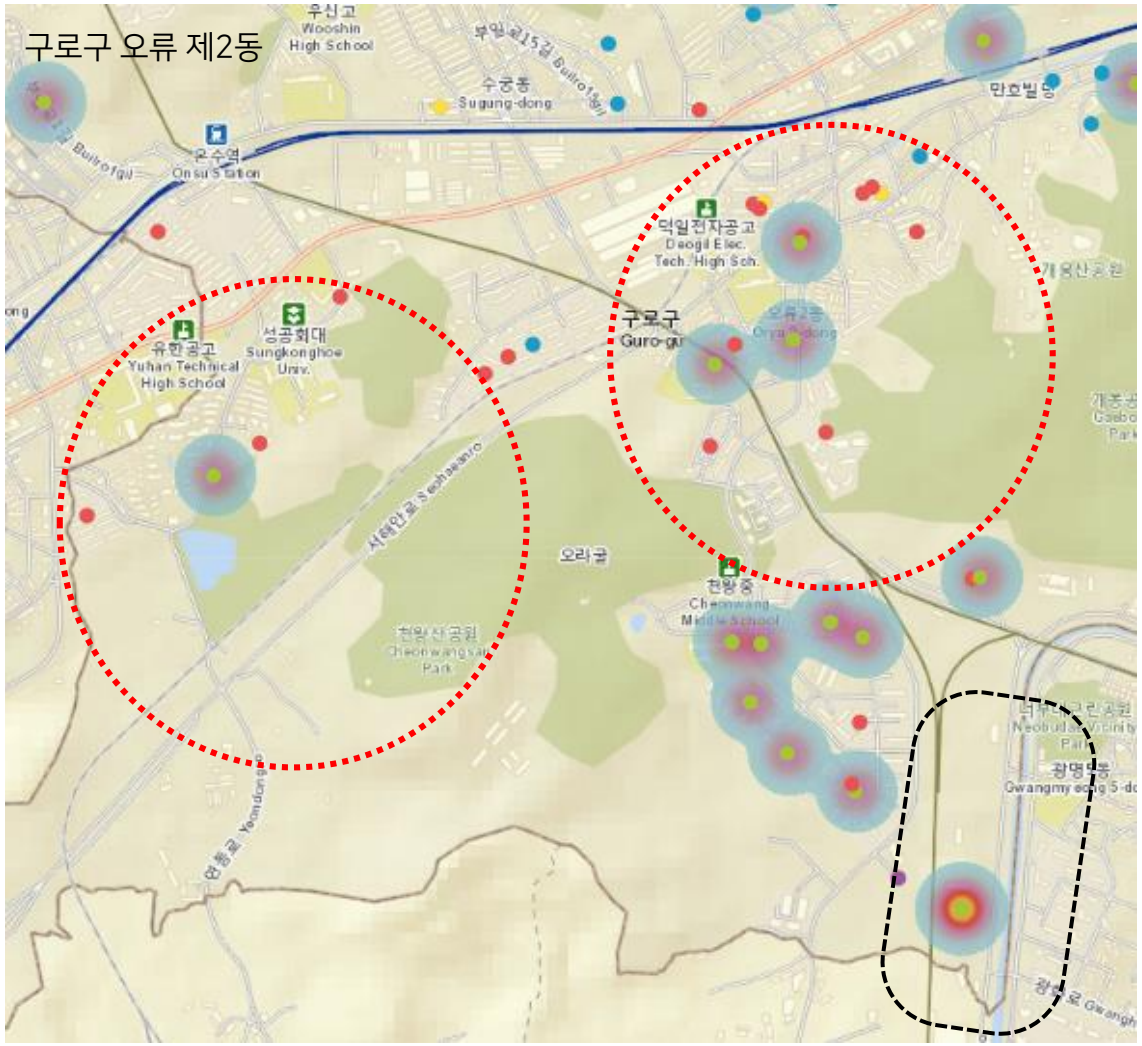
전체 행정동의 하위 0.3%에서
하위 19.6% 수준까지 상승

각 행정동에 국공립 어린이집을 증설하고 정원을 증원할 경우
이용자 접근성 지수가 크게 상승될 것으로 예측됨

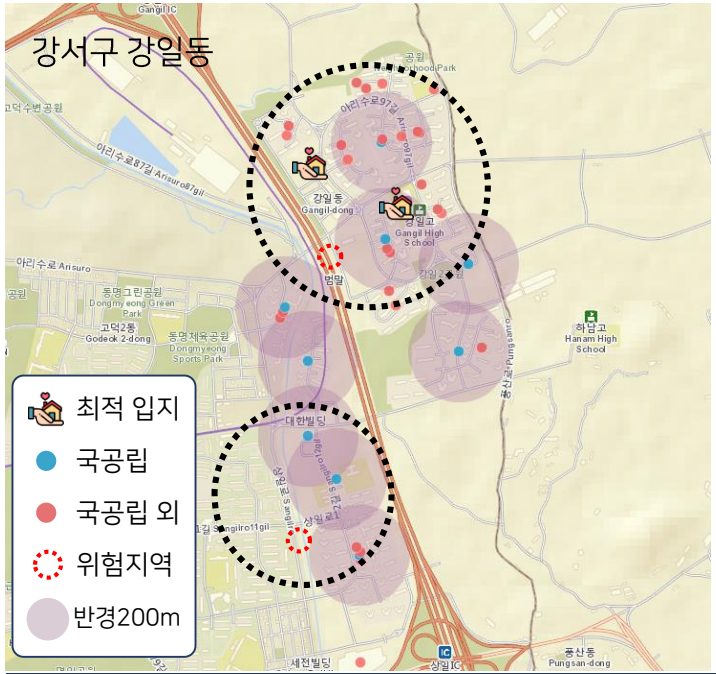
GIS 활용 종합 분석

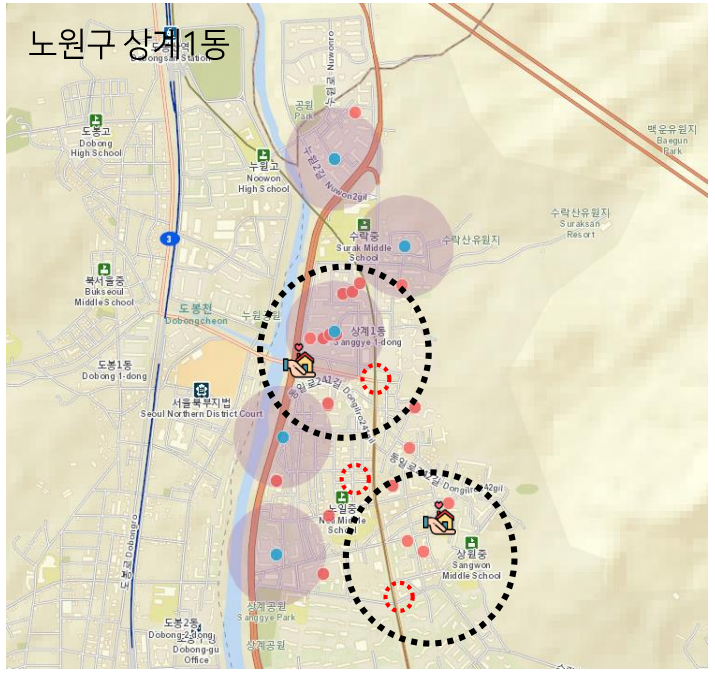
서울시 국공립 어린이집 최적지 분석 결과

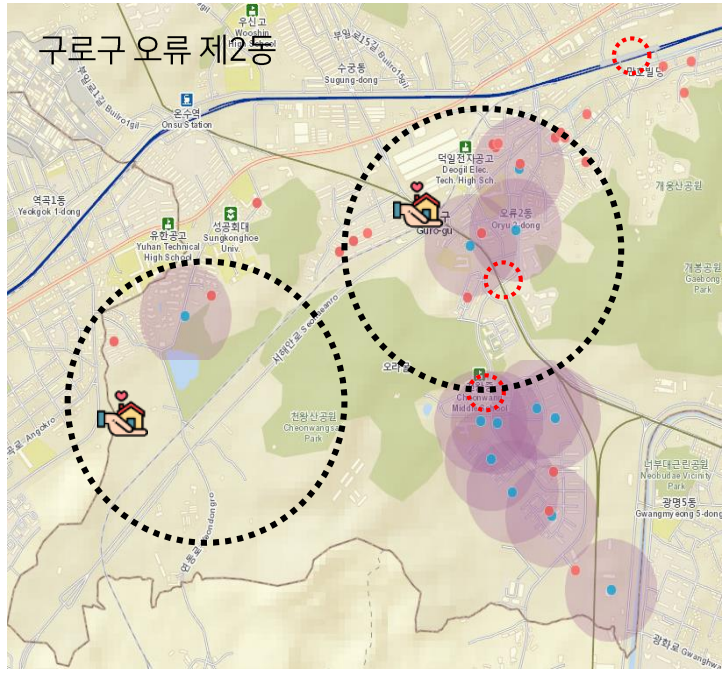
3-1. 후보 행정동 내 이용자 접근성 취약 지역 선별



3-2. 생활안전지도 기반 국공립 어린이집 최적지 선정







* 치안위험지역과 국공립 어린이집 반경 200m를 고려하여 최적 입지 선정

3-3. 연구의 한계점

1 부동산 데이터 기반의 소득 추정치는 실제 소득 수준과의 오차가 발생함

2 인구 데이터의 크기가 작아 모델의 학습량이 줄어드는 한계점이 있음

3 입지 선정에서 다른 알고리즘을 적용한다면 보다 폭넓은 분석이 가능할 것으로 기대됨

4. 참고 자료 및 분석 도구

참고 자료

- 김진영, 2014, 'GIS를 활용한 천안시 국공립 어린이집 최적지 분석에 관한 연구', 한국교원대학교 교육대학원
- 서문희·송신영, 2011, '우리나라의 보육 실태와 외국 사례-공립보육시설, 보육비용 지원, 양육수당을 중심으로', 기획재정부 육아정책연구소
- 손정렬·오수경, 2007, 'GIS 공간분석기법을 이용한 서울시 노인주간보호시설의 접근성 연구', 한국지역지리학회지, 제13-5호
- 신한은행, 2018, '서울시 생활금융지도 – 소득편'
- TAPAS, 2018, '서울 자치구별 월소득 비교해보니... 1위는 OO구', http://www.tapasnews.com (검색일:2018.10.29)

참고 사이트

- 서울시 빅데이터 캠퍼스 https://bigdata.seoul.go.kr
- 서울시 열린데이터 광장 http://data.seoul.go.kr/
- 행정안전부 https://www.mois.go.kr
- 국토교통부 https://www.realtyprice.kr
- 생활안전지도 http://www.safemap.go.kr/main/smap.do

분석 도구

- Python - Pandas, Numpy, Matplotlib, Tensorflow, Sklearn, Scipy
- PowerBI
- ArcGIS

감사합니다

서울시 국공립 어린이집 최적 입지를 찾았다!

팀명: 빅캠아이(BigCami)

팀원: 신동걸, 임희진, 전현성, 최유리