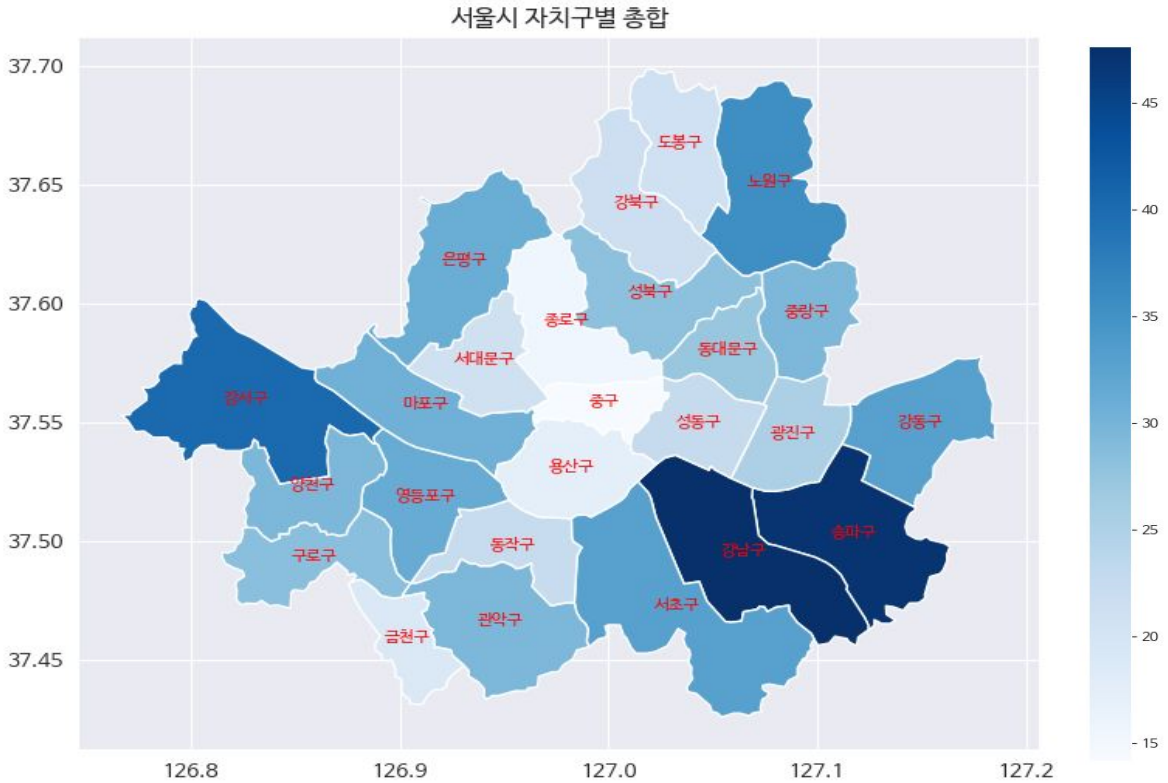
삼정 KPMG

[도시의 재구성: 사람들은 무엇에 이끌리는가]

TEAM Urban Vision: 김다은, 김민지, 박서형, 지준희



어떤 곳에 살고싶은가?

색깔이 의미하는 것은?

색의 분포가 왜 다를까?

단순한 불균형인가?

불균형의 해소방법은?

목차

1. 프로젝트 기안

1-1. 주제 선정 배경

지방 분리 시대 (1995 ~ 2025)

- 자치 제도를 통한 지역 경제 발전 도모
- 지방 광역시·도별 발달 불균형 심화
- 서울특별시 자치구별 불균형 심화
- 지역 경제 활성화 방안의 필요성 부상

지방 통합 시대 (2025 ~)

- 지방 광역시·도 메가시티 권고안 제기
- 서울특별시 지역균형발전계획 수립
- 자치구별 도시 구성 요인 비교분석
- 도시 발전 구조 예측 및 개발 전략 수립

1. 프로젝트 기안

1-1. 주제 선정 배경

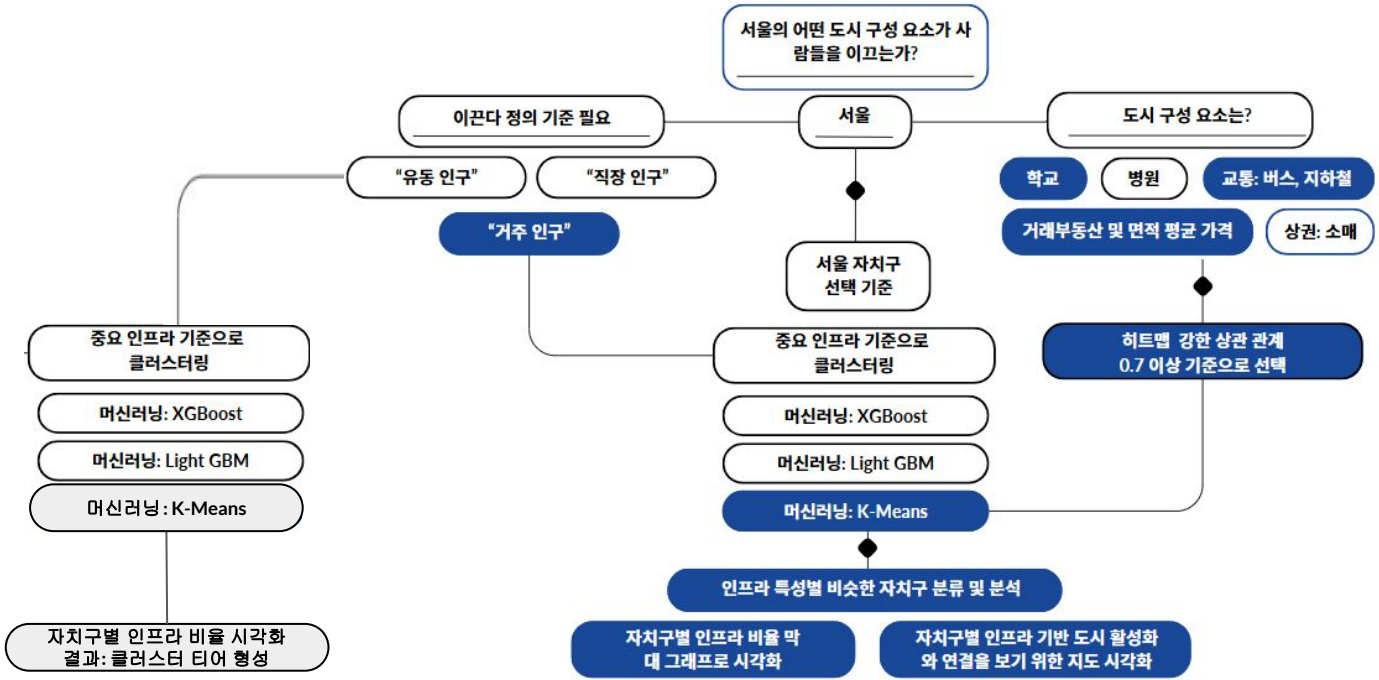
서울특별시도 동남권(강남3구)에 집중된 **불균형을 해소**하기 위한 지역균형발전계획의 수립

어떤 자치구끼리 역할을 통합해야 할지 결정하기 위해서는
도시를 구성하는 요인들이 무엇인지 분석이 선행되어야 함

수립

1. 프로젝트 기안

1.1 마스터 플랜: 팀 Urban Vision은 다음과 같은 기획으로 프로젝트를 진행함



2. 프로젝트 개발환경

2-1. 작업 라이브러리

클라우드 개발환경 : Google Colaboratory

- **데이터 전처리 (Data Preprocessing)**
  - Excel, Pandas(Python), Numpy
- **데이터 시각화 (Data Visualization)**
  - Matplotlib, Seaborn, Plotly, Folium, Geopandas
- **지리공간적 분석 (Geospatial Analysis)**
  - Geopandas, Shapely, PySAL, Rasterio
- **머신러닝 분석 및 예측 (Machine Learning & Prediction)**
  - XGBoost, LightGBM, Scikit-learn

2. 프로젝트 개발환경

2-2. 플랫폼 아키텍처

운영체제 (OS) : Windows 10 Pro (64bit)

- **데이터 분석 (Data Analysis)**
  - Python 3.11.11, Pandas(Python)(2.2.2), Numpy(1.26.4)
- **데이터 시각화 (Data Visualization)**
  - Matplotlib(3.10.0), Seaborn(0.13.2), Plotly(5.24.1), Folium(0.19.4), Geopandas(1.0.1)
- **머신러닝 (Machine Learning)**
  - XGBoost(2.1.3), LightGBM(4.5.0), Scikit-learn(1.6.0)

3. 프로젝트 추진체계

3-1. 팀 소개

TEAM Urban Vision

- **김다은 (Daeun, Kim) : Team Leader**
  - 교통 및 인프라 데이터 분석
  - PCA(주성분 분석) Clustering
- **김민지 (Minji, Kim)**
  - 상권 및 부동산 데이터 분석
  - PCA(주성분 분석) Clustering
- **박서형 (Seohyung, Park) : Vice Team Leader**
  - 인구 데이터 분석
  - Heat-map & Bar-plot
- **지준희 (Joonhee, Ji)**
  - 직장 데이터 분석
  - Heat-map & Bar-plot

3. 프로젝트 추진체계

3-2. 일정 계획

프로젝트 기간 : 2025.01.17(금) - 2025.01.24(금)

Task	Time Schedule						
	1/17(Fri)	1/19(Sun)	1/20(Mon)	1/21(Tue)	1/22(Wed)	1/23(Thu)	1/24(Fri)
Data Preprocessing							Presentation Due Date
Data Cleaning							
Data Analysis							
Data Visualization							
Machine Learning							
Final Adjustments							

삼정 KPMG

프로젝트 1단계:
전처리 및 선택 기준의 상관계수

4. 프로젝트 데이터 셋

4-1. 원본 데이터

서울 열린데이터 광장

- 등록인구_20250116170835
- 서울시상권분석서비스(직장인구-자치구)_2024.12.11
- 서울시 지하철 호선별 역별 유_무임 승하차 인원 정보
- 서울시 시내버스 정류소 현황 2024
- 서울시_상권분석서비스(추정매출-상권)_2023년
- 서울시 부동산 실거래가 정보_2025.01.15.
- 서울시 학교 기본정보 (2024)
- 서울시 병의원 위치 정보 (2024)

인구
(등록인구)

직장
(직장인구, 유동인구)

교통
(지하철 역, 버스 정류장)

상권
(소매, 외식, 직장인, 기타)

부동산
(가격, 거래량)

인프라
(학교, 학원, 병원)

4. 프로젝트 데이터 셋

4-2. 데이터 전처리

자치구명으로 분류된 통합 csv 데이터 파일 생성

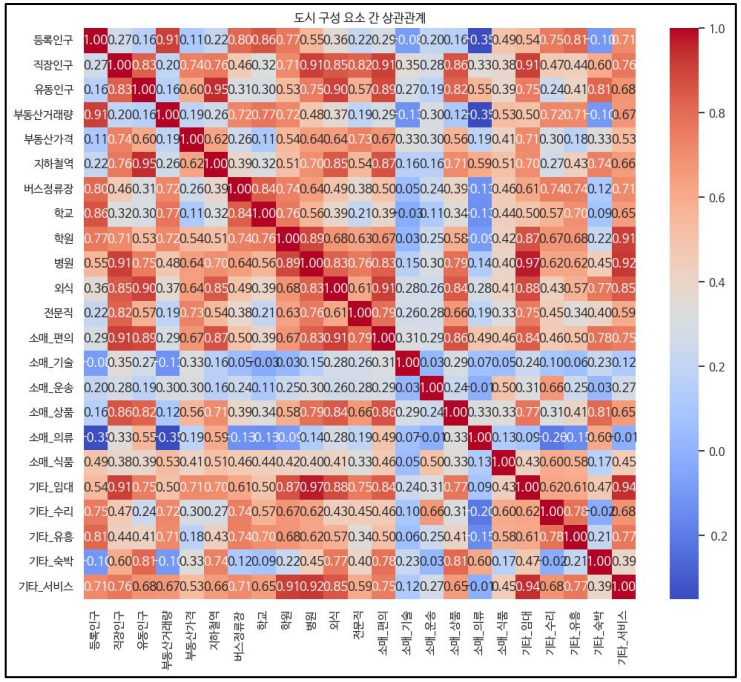
자치구명	등록인구	직장인구	유동인구	부동산거래량	부동산가격	지하철역	버스정류장	학교	학원	병원	외식	전문직
강남구	564280	1121201	503302811	4961	2337.322965	43	424	85	4624	3451	13967	3026
강동구	469464	75000	123064048	5277	1290.043635	18	302	64	1879	981	5317	170
강북구	289678	28246	86536158	2571	683.6846052	12	195	38	687	523	3839	118
강서구	563515	149719	177323432	6058	876.2260366	19	335	82	2120	1033	7284	412
관악구	496469	72879	136579127	3220	828.5015155	10	259	59	1406	784	5867	217
광진구	349307	81443	126469385	3273	1236.278512	11	174	46	1147	642	5343	210
구로구	412441	139820	159862041	3597	814.0908479	12	254	60	1191	640	4968	402
금천구	239577	126426	64103637	2340	826.6017607	4	168	35	592	404	3803	405
노원구	498358	56701	159311071	4620	917.5642511	16	390	100	2152	837	4586	120
도봉구	306926	36388	69991699	2771	649.0097005	8	232	46	943	410	2998	192
동대문구	359219	80350	119079063	3887	1092.453337	11	258	50	1077	684	5042	189
동작구	387792	78066	193600630	3963	1299.738743	20	141	51	1386	668	4284	149
마포구	373874	142802	288018206	4833	1414.107629	29	252	52	1859	860	10020	531
서대문구	319749	66165	77063767	3486	1112.833098	9	211	43	1143	495	4536	82
서초구	412611	576938	212438876	4199	2304.641884	23	325	57	2726	1700	7420	4444
성동구	282385	80690	150399028	3570	1744.107387	17	187	39	996	540	4789	359
성북구	435492	110756	98814397	4209	980.0545308	13	282	62	1482	614	4842	83
송파구	657991	399728	260133120	6277	1651.265657	32	439	96	2814	1426	8743	993
양천구	435867	139602	54510743	4257	1165.708525	6	257	64	2634	730	4024	345
영등포구	397514	397723	260703845	4233	1317.820248	26	291	47	1342	903	8018	865
용산구	218370	176381	120116395	2215	1991.513481	14	231	38	624	365	5470	183
은평구	466809	48907	103631922	4859	840.3519531	13	323	68	1529	772	4779	122
종로구	150011	234085	230827930	1274	1062.744647	17	189	47	636	559	7015	447
중구	131589	334297	355879391	1246	1300.94439	38	164	36	416	665	6756	754
중랑구	386131	39323	81478637	3295	863.4355569	14	357	48	911	633	4392	83

5. 프로젝트 진행 과정

5-1. 상관관계 히트맵

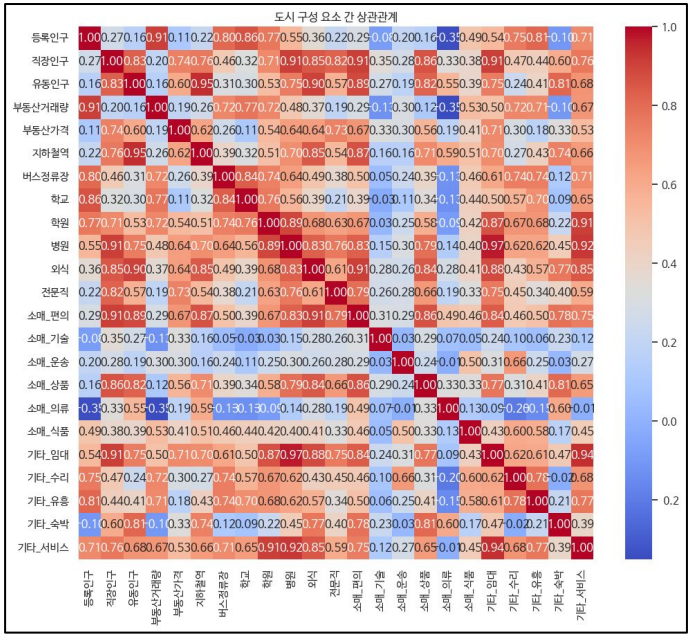
1차 상관관계 분석 결과

- **상관관계가 높을수록 도시 발달에 필수적인 요소**
- **다중 공선성 발생**
  - 1에 가까운 상관관계로 변수 간 영향 중복
  - 상위 개념에 해당하는 변수 채택
- **스케일 에러 발생**
  - 명수, 가격, 개수 등 여러 가지 단위 혼재
  - 표준화 및 정규화 작업 진행



5. 프로젝트 진행 과정

5-1. 상관관계 히트맵



도시 발달 기준을 Work vs. Life로 구분

Work

유동인구
지하철 역
부동산 가격
병원
소매_편의
소매_상품
소매_외식

Life

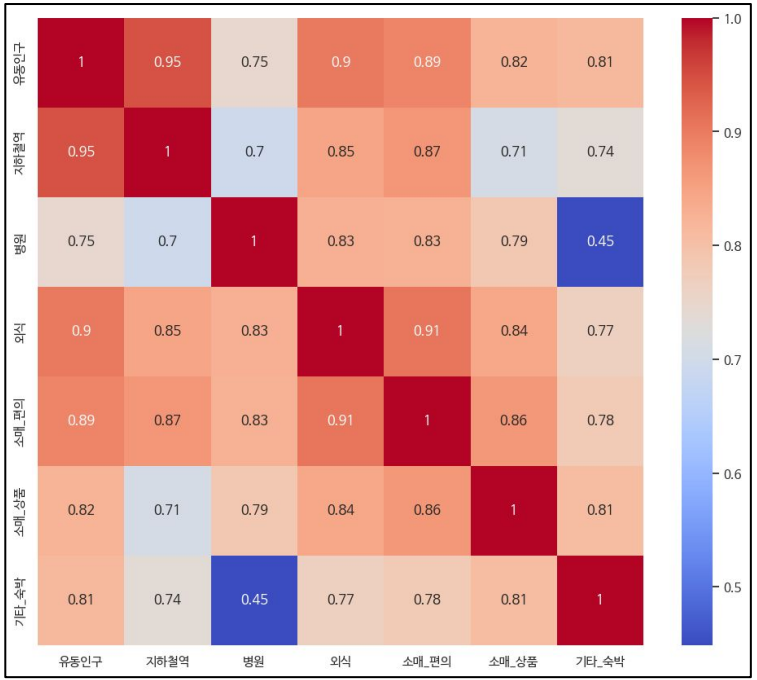
등록인구
버스 정류장
부동산 거래량
학교
기타_수리업
기타_유흥업
기타_서비스업

5. 프로젝트 진행과정

5-1. 상관관계 히트맵

유동인구 중심
상관관계 (Correlation) 히트맵 (Heat-map)

- 유동인구가 많은 자치구는 업무 단지로 발달됨
- 유동인구와 지하철 역의 상관관계 (0.95)
- 유동인구와 외식의 상관관계 (0.90)

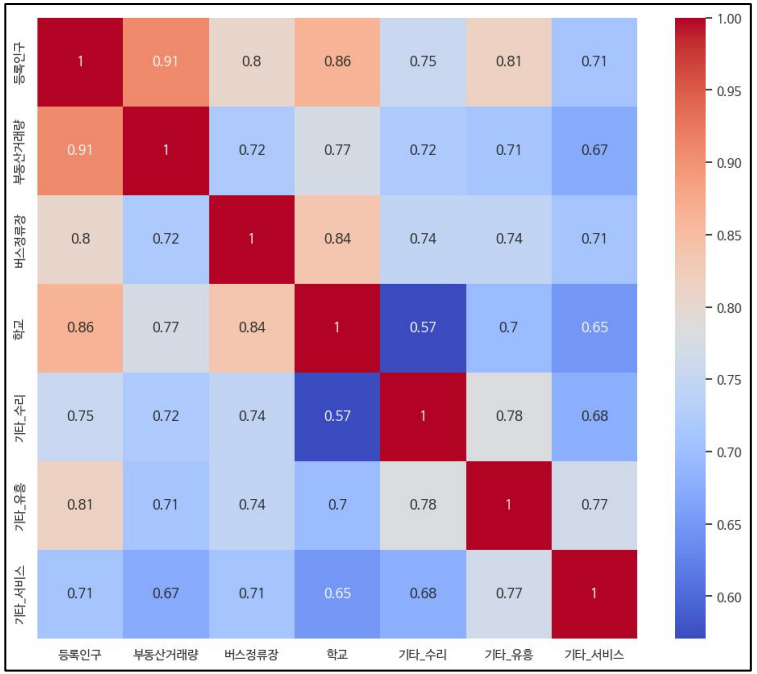


5. 프로젝트 진행과정

5-1. 상관관계 히트맵

등록인구 중심
상관관계 (Correlation) 히트맵 (Heat-map)

- 등록인구가 많은 자치구는 거주 단지로 발달됨
- 등록인구와 부동산거래량의 상관관계 (0.91)
- 등록인구와 학교의 상관관계 (0.86)



삼정 KPMG

클러스터링 : K-MEANS

자치구별 인프라 기준으로 분리

클러스터링 : K-Means 선택 기준

총 6가지의 인프라 데이터 기반으로 정확한 자치구 분류 혹은 군집화를 위해 라벨링을 하지 않는 비지도 학습 K-means 알고리즘 채택함

	XGBoost	Light GBM	K-Means
인프라 데이터 기반된 자치구 분리	새로운 수치 데이터 필요	새로운 수치 데이터 필요	✓
인프라 구조의 비슷한 정도로 자치구 분리	새로운 수치 데이터 필요	새로운 수치 데이터 필요	✓
라벨링 사용하지 않음	✗	✗	✓

클러스터링 - 전처리

클러스터링 결과의 정확도를 높이기 위해 Min Max Scaler를 사용한 정규화를 진행하였으며, 서울지 자치구의 경계 정보와 데이터를 병합함

표준화 : Standard Scaler

서울시 자치구 경계 정보 데이터와 병합

code	name	name_eng	base_year	geometry	자치구명	cluster	Cluster	등록인구	거래된부동산	버스정류장	수리업	유흥업	서비스업	학교
11250	강동구	Gangdong-gu	2013	POLYGON ((127.1152 37.55753, 127.1188 37.55722...	강동구	1	0	0.641857	0.801232	0.540268	0.619687	0.480994	0.240458	0.446154
11240	송파구	Songpa-gu	2013	POLYGON ((127.06907 37.52228, 127.07496 37.520...	송파구	2	0	1.000000	1.000000	1.000000	0.948546	1.000000	0.507888	0.938462
11230	강남구	Gangnam-gu	2013	POLYGON ((127.05867 37.5263, 127.06907 37.5222...	강남구	2	0	0.821978	0.738422	0.949664	1.000000	0.808480	1.000000	0.769231
11220	서초구	Seocho-gu	2013	POLYGON ((127.01397 37.52504, 127.01918 37.520...	서초구	1	0	0.533854	0.586961	0.617450	0.798658	0.483918	0.355725	0.338462
11210	관악구	Gwanak-gu	2013	POLYGON ((126.98368 37.47386, 126.98464 37.469...	관악구	1	0	0.693158	0.392367	0.395973	0.503356	0.535088	0.224173	0.369231

- 특성들을 표준화하여 변수 간 스케일을 조정
- K-Means 알고리즘의 성능 향상을 통해 정확한 클러스터 파악

- 서울시 자치구 경계 정보가 담긴 데이터를 geopandas를 사용해 로드
- 정규화된 데이터프레임과 병합하여 지도를 만들기 위한 준비 진행

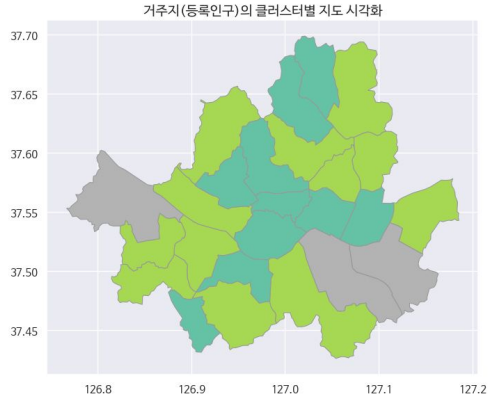
클러스터링 - K-mean

Elbow method를 통해 최적의 클러스터 개수를 파악하고, 정규화된 데이터를 기반으로 K-Means 클러스터링 진행

Elbow Method으로 k 값 지정

- 급격하게 꺾이는 지점에서 최적 k 값 파악
  → **K-Means 알고리즘 클러스터링** 에 적용
- K = 3으로 설정

K = 3 기반으로 자치구 군집화



강남구 송파구 강서구

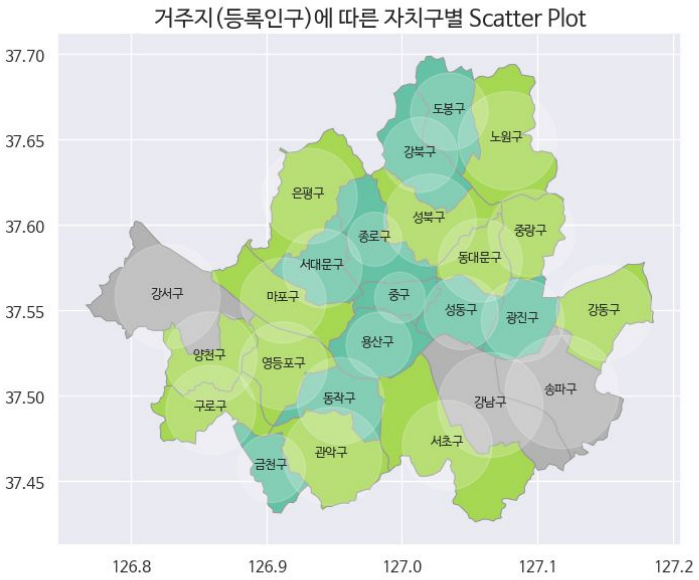
노원구 은평구 성북구 양천구
강동구 구로구 관악구 마포구
동대문구 중랑구 서초구
영등포구

동작구 서대문구 종로구 용산구
도봉구 중구 금천구 강북구
성동구 광진구

거래된 부동산 수, 버스정류장 수, 수리업 소매 상권 수, 유흥업소 수, 서비스업소 수, 그리고 학교 수의 기반되어 자치구를 3가지 클러스터로 군집화 했음

클러스터링 - 지도 시각화

Geopandas의 plot, centroid와 matplotlib의 scatter과 annotate를 사용하여 자치구별 지도를 시각화함



Geopandas의 plot()

클러스터별 색상을 구분하여 지도 생성

Geopandas의 centroid

각 자치구의 중심점을 계산하여 지도에 표기할 마커의 좌표를 형성

Matplotlib의 scatter

자치구 중심점에 인구 규모에 비례하는원형 마커를 표시

도출 인사이트

군집화된 자치구들의 인프라 구성 비슷함은 그 자치구들의 거주 인구 수와 비례하지 않음

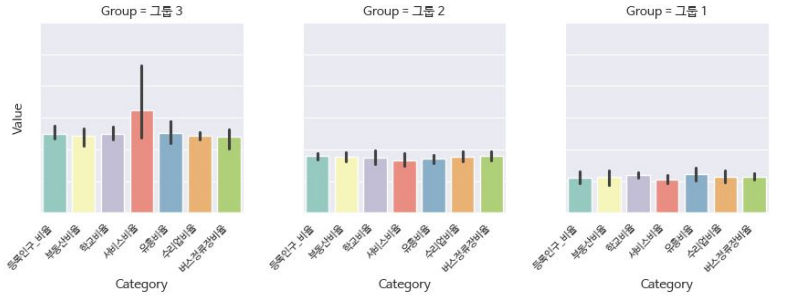
하지만 **1번 클러스터** 에 해당하는 자치구 (중구, 용산구, 동작구, 등등)들은 **서울 자치구 중 하위 순위에 해당함** 으로 거주 인구 지역에 적합하지 않다고 판단

3번 2번 클러스터에 해당하는 자치구와 1번 자치구의 구조 차이를 분석할 필요

6. 결과: 클러스터 자치구별 인프라 비교: 막대 그래프

프로젝트의 도출 데이터 분석 결과의 토대로, 인프라 구성을 비교하여, 어떤 인프라를 향상해야하는지 효율적으로 분석 할 수 있음

자치구 클러스터별 평균 인프라 보유 비율

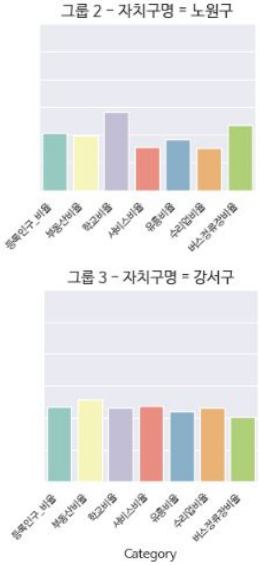


계산법: (자치구의 해당 인프라 수/ 서울시에 있는 해당 인프라 수) * 100

인사이트:

- 평균적으로 클러스터 3, 2, 1 순서로 인프라 비율이 나뉜다
- 그러므로 **3번 클러스터의 비율과 유사해지는 것을 도시 구조 벤치마킹** 사례로 설정할 수 있다

노원구->강서구로 보는 벤치마킹 인사이트



클러스터 비율 평균

	1	2	3
부동산거래 수 비율	20%	30-40%	50-60%
학교 수 비율	10%	20-30%	40-50%
서비스업소 수 비율	10%	20-30%	60%
유흥업소 수 비율	10%	30-40%	40%
수리업소 수 비율	10%	30-40%	40%
버스정류장 수 비율	15%	20-30%	30-40%

6. 결론: 개선점과 한계점

프로젝트 진행 결과, 다음과 같은 한계점과 개선점을 찾아볼 수 있었다

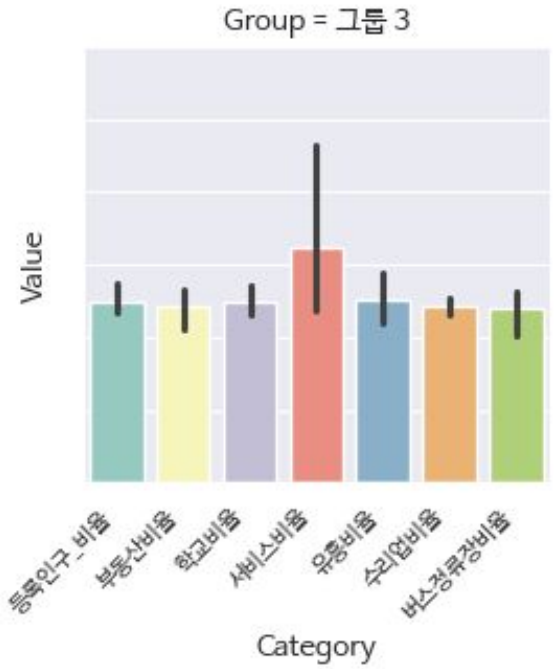
개선점

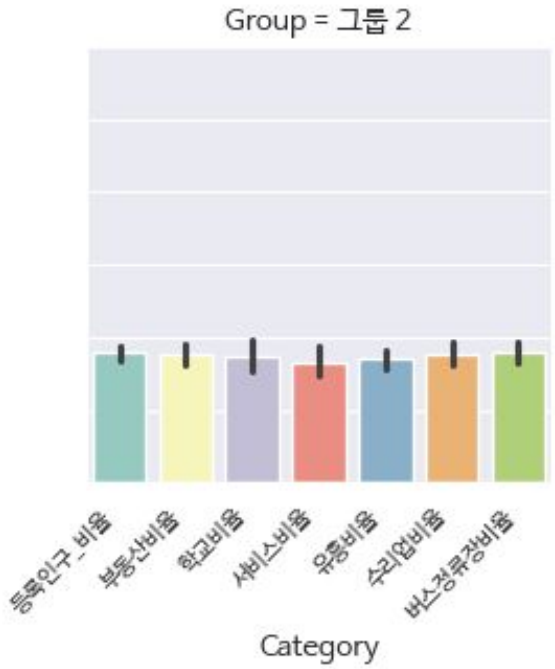
- 변수가 가지는 요소들을 고려했다면?(ex_학교별 진학률)
- 지역만이 가지는 요소들(공공청사, 문화재 등)을 고려했다면?
- 더 많은 데이터를 참고했다면?

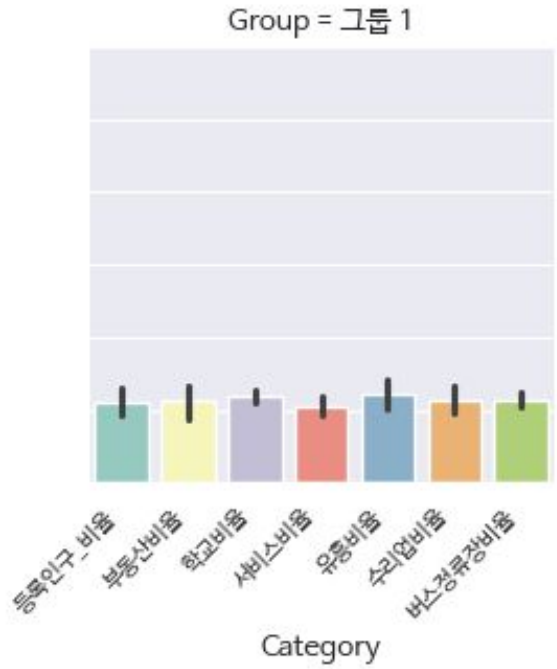
한계점

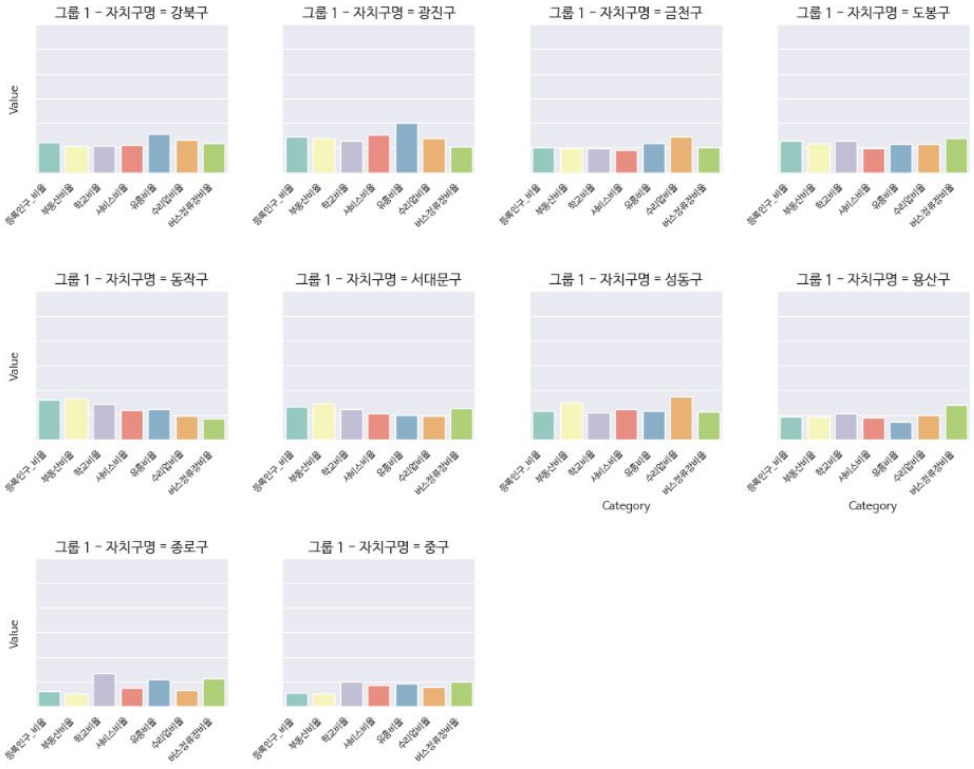
- 서울의 요소들이 지방에도 똑같이 적용된다고 볼 수는 없다.
- 도시 발전의 핵심이 되는 key값을 찾을 수 없었다.(주거지역, 상업지역이 다름)
- 요소들에 가중치를 부여하기 어렵다.

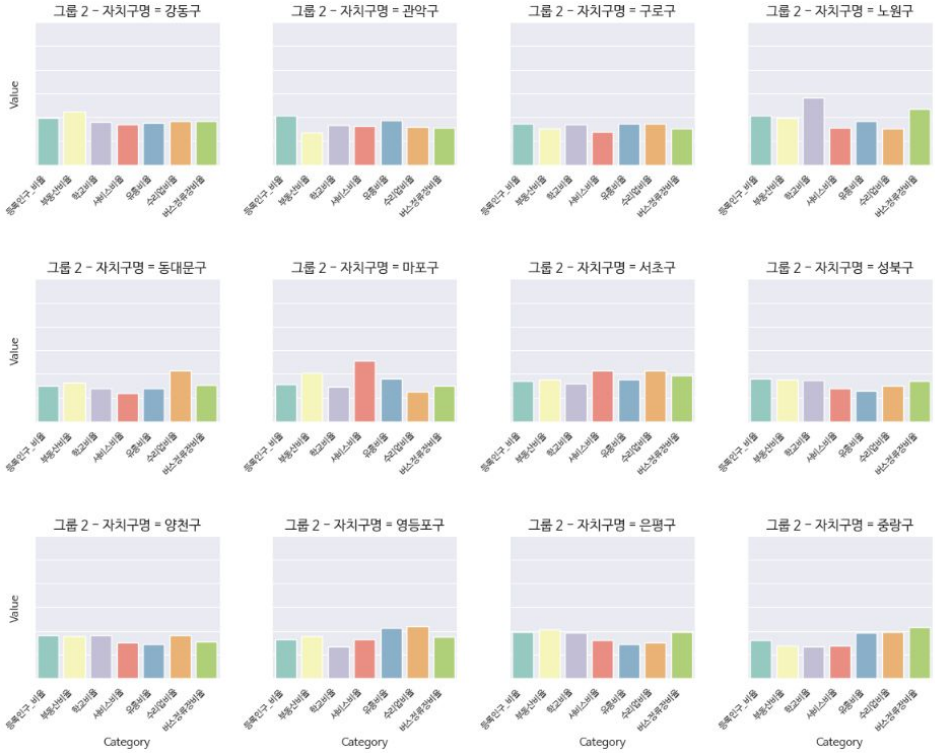
APPENDIX

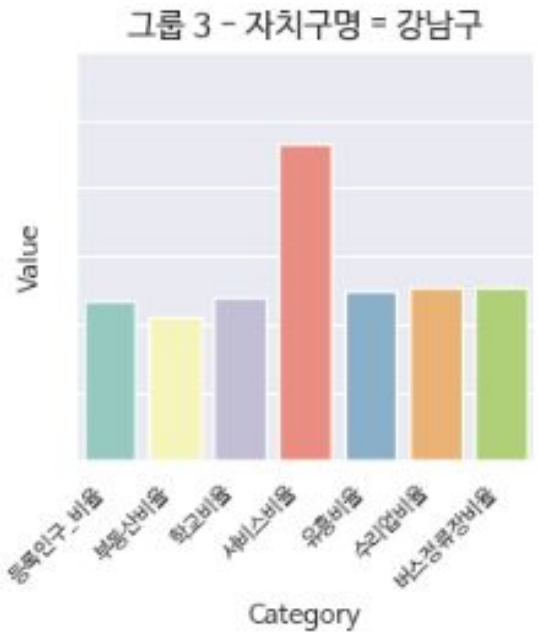

Group = 그룹 3
Value
등록인구_비율
부동산비율
학교비율
서비스비율
유통비율
수리업비율
버스정류장비율
Category

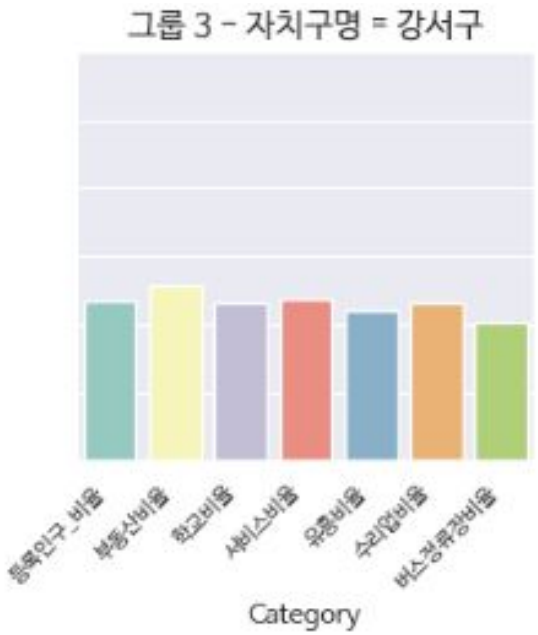

Group = 그룹 2
등록인구_비율
부동산비율
학교비율
서비스비율
유통비율
수리업비율
버스정류장비율
Category

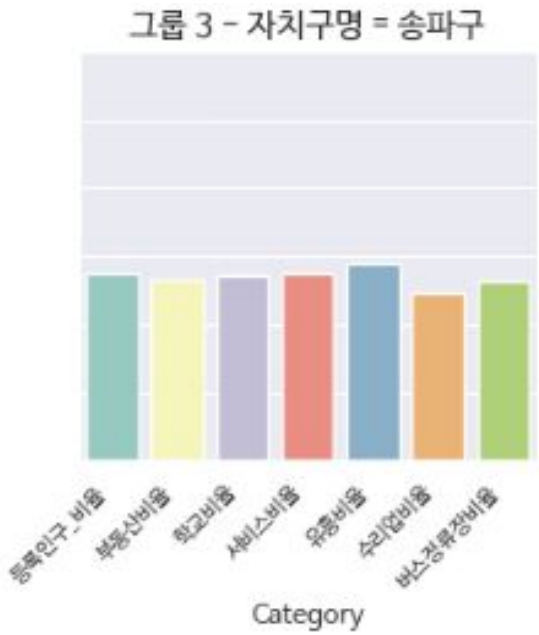

Group = 그룹 1
등록인구_비율
부동산비율
학교비율
서비스비율
유통비율
수리업비율
버스정류장비율
Category


그룹 1 - 자치구명 = 강북구
그룹 1 - 자치구명 = 광진구
그룹 1 - 자치구명 = 금천구
그룹 1 - 자치구명 = 도봉구
그룹 1 - 자치구명 = 동작구
그룹 1 - 자치구명 = 서대문구
그룹 1 - 자치구명 = 성동구
그룹 1 - 자치구명 = 용산구
그룹 1 - 자치구명 = 종로구
그룹 1 - 자치구명 = 중구
Value
Category
등록인구 비율
보행신비율
학교비율
서비스비율
유흥비율
숙박업비율
버스정류장비율

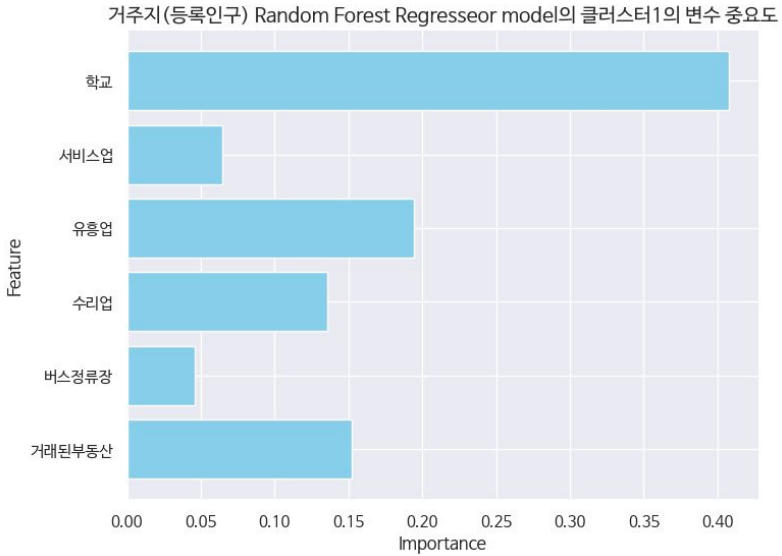

그룹 2 - 자치구명 = 강동구
그룹 2 - 자치구명 = 관악구
그룹 2 - 자치구명 = 구로구
그룹 2 - 자치구명 = 노원구
그룹 2 - 자치구명 = 동대문구
그룹 2 - 자치구명 = 마포구
그룹 2 - 자치구명 = 서초구
그룹 2 - 자치구명 = 성북구
그룹 2 - 자치구명 = 양천구
그룹 2 - 자치구명 = 영등포구
그룹 2 - 자치구명 = 은평구
그룹 2 - 자치구명 = 중랑구
Value
Category
음폐인구 비율
부동산비율
학교비율
서비스비율
유흥비율
숙박업비율
버스정류장비율

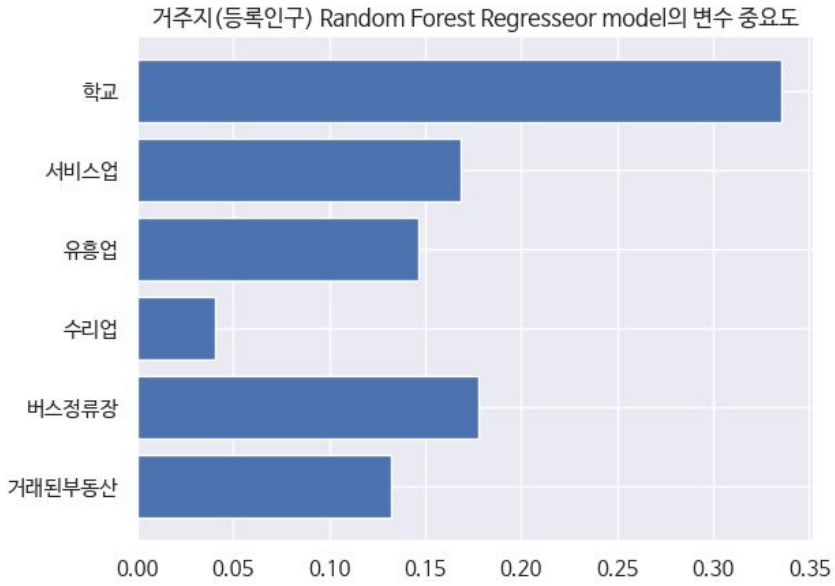

그룹 3 - 자치구명 = 강남구
Value
등록인구_비율
부동산비율
학교비율
서비스비율
유흥비율
수리업비율
버스정류장비율
Category

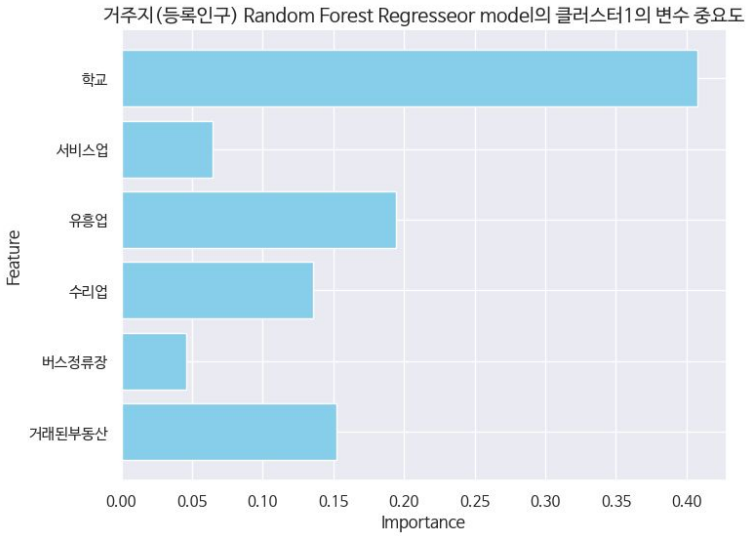

그룹 3 - 자치구명 = 강서구
등록인구_비율
부동산비율
학교비율
서비스비율
유흥비율
수리업비율
버스정류장비율
Category

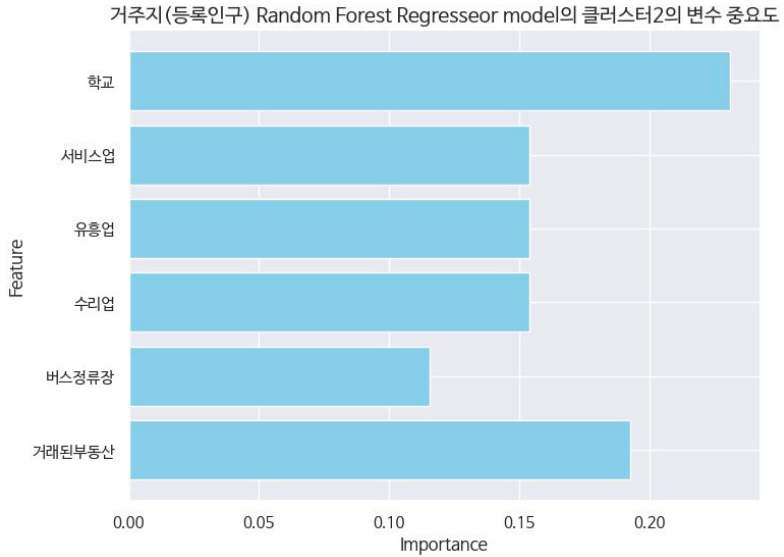

그룹 3 - 자치구명 = 송파구
등록인구_비율
부동산비율
학교비율
서비스비율
유흥비율
수리업비율
버스정류장비율
Category

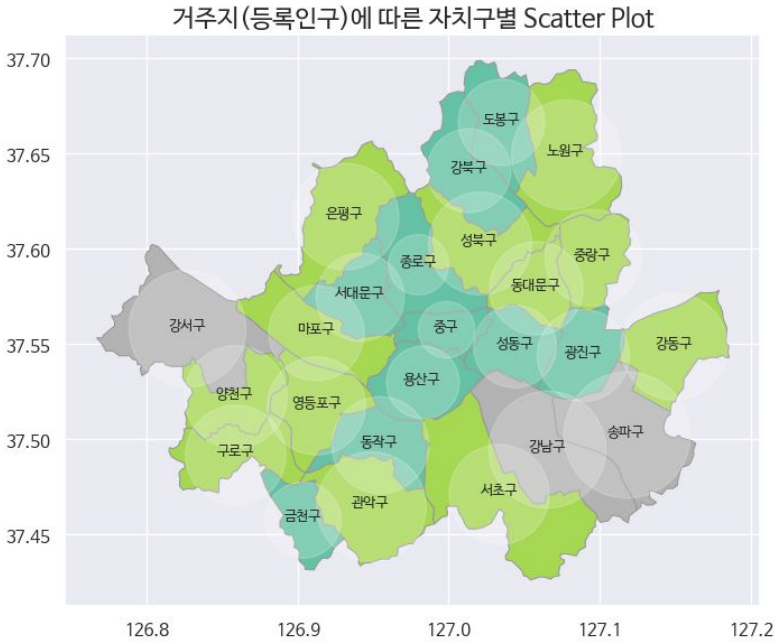
클러스터링 - 다른 알고리즘 시도

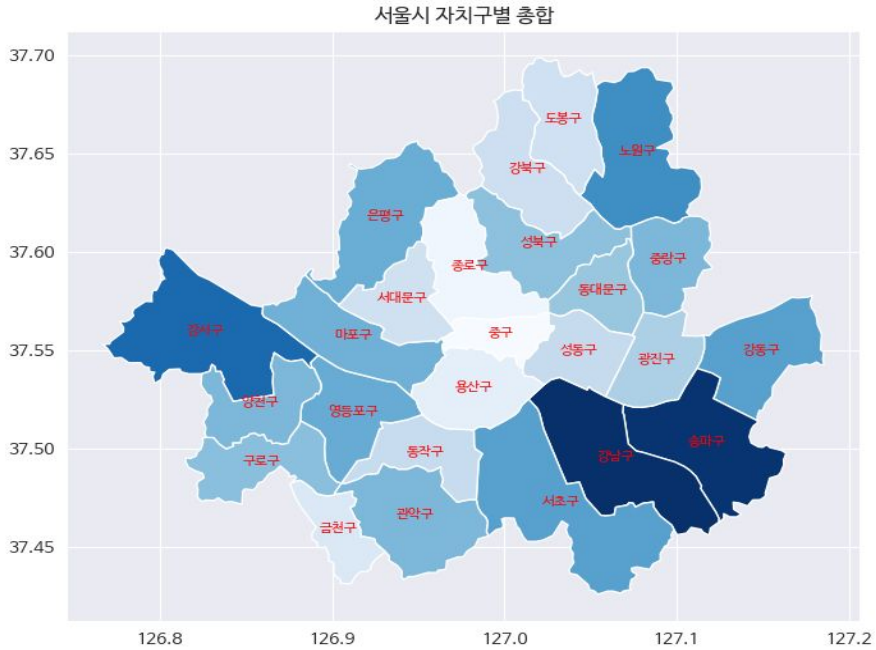





거주지(등록인구) Random Forest Regresseor model의 클러스터1의 변수 중요도
학교
서비스업
유흥업
수리업
버스정류장
거래된부동산
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance


거주지(등록인구) Random Forest Regresseor model의 클러스터2의 변수 중요도
학교
서비스업
유흥업
수리업
버스정류장
거래된부동산
Feature
0.00
0.05
0.10
0.15
0.20
Importance


거주지(등록인구)에 따른 자치구별 Scatter Plot
37.70
37.65
37.60
37.55
37.50
37.45
126.8
126.9
127.0
127.1
127.2
도봉구
노원구
강북구
은평구
성북구
중랑구
종로구
동대문구
서대문구
강서구
마포구
중구
성동구
광진구
강동구
용산구
양천구
영등포구
동작구
송파구
구로구
강남구
서초구
관악구
금천구


서울시 자치구별 총합
37.70
37.65
37.60
37.55
37.50
37.45
126.8
126.9
127.0
127.1
127.2
도봉구
노원구
강북구
은평구
성북구
중랑구
종로구
동대문구
서대문구
강서구
마포구
중구
성동구
광진구
강동구
용산구
양천구
영등포구
동작구
송파구
구로구
강남구
서초구
관악구
금천구

Thank You