

2021 SDC

각 시도별 도시 경쟁력 지표 측정 및 이를 통한 시도간 인구이동의 분석과 이해

21.6.20 – 21.8.10
2021년 통계데이터 분석 및 활용대회

Contents



1 분석 배경

문제 현황, 이전 연구, 본 연구의 목적

2 데이터 수집

도시 경쟁력 지표 설립, 자료 수집, 도로네트워크, 데이터 전처리

3 데이터 분석

도시경쟁력 지표 측정, 모델생성

4 분석 결과 및 해석

시도별 도시경쟁력 측정, 회귀모델 분석

5 결론 및 한계점

결론, 한계점 및 향후 연구 방향

분석 배경

문제 현황 , 이전 연구와의 차별성, GPCI와 도시 경쟁력 지표, 본 연구의 목적



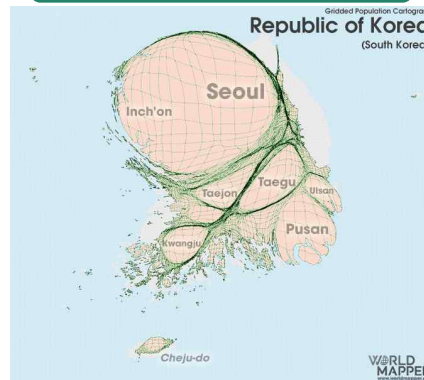
01°

01. 분석배경_문제현황

수도권 인구 집중도의 변화



격자형 인구분포 지도



수도권 인구집중 완화 정책

수도권 규제

(과밀억제권역·성장관리권역·자연보전권역으로 나누어 공장이나 학교, 대형 업무 시설 등 인구 집중 유발 시설의 설치를 제한)

지역 혁신성장 거점 조성

(혁신도시, 행복도시)

광역거점 중심의 교통망 연계강화 등
(광역철도, 고속도로 건설)

- ❖ 우리나라는 서울, 인천, 경기 등 3개 시도를 포함한 수도권으로의 인구이동 현상이 계속 발생하고 있다.
- ❖ 수도권은 전체 국토면적의 12%이지만 통계청에 따르면 2020년 수도권 인구는 전체 인구의 50.1%로 수도권의 인구집중이 심화되었다는 것을 확인할 수 있다.
- ❖ 지역 간 불균형한 성장 문제를 해결하기 위해 인구이동에 영향을 끼치는 요소에 대한 분석이 필요하다.

WorldMapper 2012

<http://www.kado.net/news/articleView.html?idxno=1004127>

01. 분석배경_이전 연구와의 차별성

최열, 김형준.

“ 수도권 및 비수도권의 주거이동 결정요인 비교 분석 ”

국토계획 47.4 (2012): 219-231.

수도권과 비수도권의 주거이동자를 가구특성, 주거환경특성, 경제적 특성(성별, 연령, 주택, 문화지출 등을 포함)에 대해 분석

최진호

“ 한국 지역 간 인구이동의 선별성과 이동 이유 ”

한국인구학 31.3 (2008): 159-178.

수도권 이동자를 수도권 전입이동, 전출이동, 그리고 수도권내 이동으로 구분하여 비이동자의 특성과 비교

최은영

“ 선택적 인구이동과 공간적 불평등의 심화 수도권을 중심으로 ” 7.2 (2004): 57-70

젊은 연령층의 수도권 집중에 대해 살펴보기 위해 연령을 기준으로 인구이동의 특징을 분석

이성우

“ 지역특성이 인구이동에 미치는 영향 : 독립이동과 연계이동 ”

지역연구 18.1 (2002): 49-82.

이주하는 지역에 연고가 있는 연결형 이주자와 아닌 독립형 이주자에 대하여, 개인특성과 지역수준을 이용한 지역간 이주 모델 설립 및 분석

이전 연구는 지역 간 인구이동에 영향을 끼치는 요소를 이동 인구의 개인적 특성 수준 (전입인구의 나이, 성별, 직업, 소득 등) 에 대해 분석하는 연구는 많이 진행되었으나, 도시의 특징에 따라 인구이동이 어떻게 이루어지는지에 대해서는 연구가 상대적으로 부족하였다. **우리는 수도권 인구집중 완화 정책이 수도권의 과도한 인프라 집중을 제한하거나 타 지역에 대하여 도시 발전을 꾀한다는 점에 주목하여** 각 도시의 경쟁력이 인구이동에 미치는 영향을 분석하고자 한다. 따라서 본 연구에서는 시도간 이동 요인을 도시 별 특성 수준에서 살펴보고자 한다. 즉,

1. 데이터를 수집하여 시도별 도시 경쟁력 지표를 측정하고
2. 도시 경쟁력 지표를 이용해 인구이동 현상 및 수도권 쏠림 현상을 분석한다.

01. 분석배경 _ GPCI와 도시 경쟁력 지표

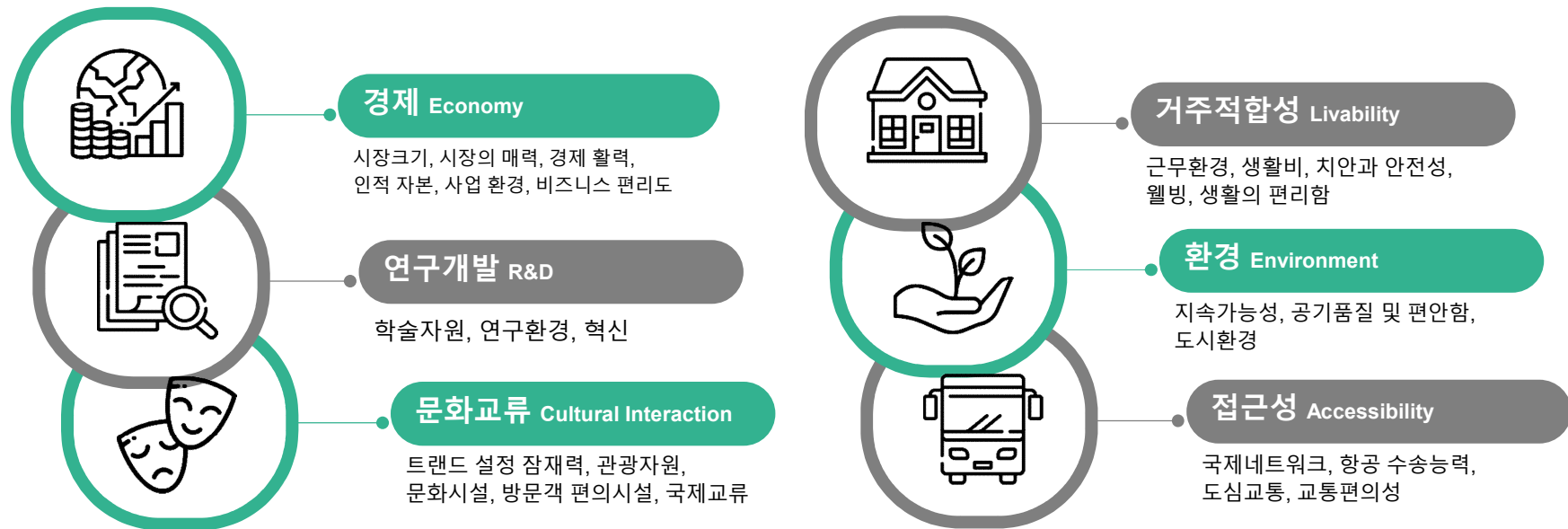
도시경쟁력 지표를 어떻게 설립할 수 있을까?

우리는 **글로벌파워 도시지수(Global Power City index, GPCI)**를 이용하였다.

일본 도시개발 조사기관인 모리기념재단(MMF)의 산하 도시전략연구소(IUS)에서 만든 세계 도시의 경쟁력을 종합적으로 판단하고자 만든 도시 지수이다. 도시전략연구소는 이를 이용하여 세계 48개 도시를 분석하고 2008년부터 매년 레포트를 발간하고 있다.

GPCI는 아래와 같은 6가지 항목을 26가지 소분류로 나눈 후, 이를 총 70여개의 지표를 통해 측정한다.

이 GPCI를 국내 도시 경쟁력을 더 잘 나타내도록 수정 하여 이를 본 연구에서의 도시 경쟁력 지표로 둔다.



01. 분석배경_연구 목적

시도별 도시 경쟁력 측정

국내에 적합한 도시 경쟁력 지표를 설립하고, 이를 이용하여 17개 시도별로 도시경쟁력을 측정한다. 그후 각 도시의 년도에 따른 변화와 최근 년도인 2019년에 대하여 시각적으로 이동, 전입인구와의 관계성을 살펴본다.

도시경쟁력과 전입 인구와의 연관성 분석

앞서 시각적으로 분석한 도시 경쟁력 지표가 시도별 전입인구에 대하여 유효한 영향을 끼치는지를 머신 러닝을 활용해 자세히 확인한다.

시도간 도시 경쟁력차이와 시도간 이동관의 연관성 분석

만약 도시경쟁력과 전입 인구와의 관계가 유효하다면 시도간 이동에 대해서도 유효하게 설립하는지를 시도간 도시경쟁력 지표 차이 값을 이용해 분석한다.



데이터 수집

도시 경쟁력 지표 설립, 자료 수집, 도로네트워크, 데이터 전처리



02°

02. 데이터 수집_ 도시 경쟁력 지표 설립

GPCI 를 기준으로 수정하여 도시 경쟁력 지표를 설립하였다

시도별 공통 지표를 제외

경제자유지수, 자연재해로 인한 경제 손실 등 → 삭제

국가 단위 지표를 지역 단위 지표로 변경

GDP(국내총생산) → GRDP(지역내총생산)
외국인 방문객 수 → 시도별 관광지 외국인 방문 수의 합
세계문화유산 접근성 → 시도별 문화재수 등

데이터 수집이 어려울 경우 유사한 지표 추가

교통 - 교통 접근성을 측정하기 위해 고속도로 네트워크
제작, BC,CC값 추가



지표를 지역 특성이 더 잘 나타내도록 수정

도심 50km 이내 (인구, 인프라 밀집도가 높은 한국엔 맞지 않음)
→ 해당 도시 내 (필요한 경우 시도 면적으로 나눠서 계산)



각 항목의 모든 소분류에 대하여 자료 조사



GPCI의 6개의 항목의 각 항목에는 소분류가 존재 (항목-소분류-세부지표순. 예를 들어 환경의 경우 지속가능성, 공기품질 및 편안함 등이 소분류에 해당) (GPCI의 모든 소분류에 대해 적어도 하나 이상의 세부지표 수집)

긍정적, 부정적 요소 모두 측정



경제 → 인당지방세, 거주 적합성 → 근무환경 등 청각
항목에 대해 최대한 부정적 요소를 포함시켜 지표 측정

02. 데이터 수집_ 자료 수집

분류	소분류	세부 지표 (변수명)	데이터 설명
경제 	시장크기	GRDP	지역내총생산
		1인당GRDP	지역내총생산액 / 지역별연앙추계인구
	시장의매력	경제성장률	지역내 총생산의 전년대비 증감률
	경제활력	매출top기업수	시도별 포함하는 전국 매출금액 상위 500개 기업 수
	인적자본	취업인구	시도별 취업자수
		서비스직비율	기업분류코드로 계산한 서비스업 종사자 수 / 전체 종사자 수
	사업환경	임금	지역별 5인 이상 기업의 인당 임금총액 평균 (정액, 초과, 특별급여 포함)
	비즈니스 수행의 감편성	인당지방세	지방세 1인당 부담현황
연구 	학술자원	연구원비율	시도별 취업자수 대비 연구원 비율
		top대학수	THE에서 매년 평가하는 세계 대학 1000위 이내의 대학 수
	연구환경	인당연구개발비	시도별 연구개발비를 취업자수로 나눈 값
		유학생수	THE에서 평가한 세계 대학 1000위 이내 각 대학의 학생수와 유학생 비율로 추정된 유학생 수의 합
	혁신	특허등록수	각 지역별 특허 등록 수
		스타트업수	창업 3년 이내의 중소기업 수

문화 교류 	트렌드 설정 잠재력	문화예산비율	공연예술예산 지역별 현황/지역별 총예산
		문화재수	시도지정문화재수(무형, 유형, 기념물, 시도민속문화재 포함)
	문화시설	극장수	공연장 시설 수
		문화기반시설수	문화기반시설(도서관, 박물관, 미술관, 문예회관 등) 수
	방문객 편의시설	호텔수	100석 이상을 가지고 있는 호텔 수
		영업점포	자기소유점포 + 임차점포의 수
		식당수	음식점업 전체 수
	국제 교류	외국인거주자수	시도별 거주 외국인 (한국국적 미취득 포함) 수, 매년 11월 1일 기준
		외국인 방문객 수	한국 여행시 방문지 (중복포함)
거주 	근무환경	청년실업률	년도별 15~29세 경제활동인구 중 실업자가 차지하는 비율
		근로시간	지역별 총 상용근로시간
	치안과안정 성	범죄발생률	인구 천명 당 범죄발생 건수
	웰빙	자살률	인구 십만명당 자살률(사망원인 중 자살로 인한 사망률)
	생활의 편리함	의사비율	각 시도별 인구 천명 당 의료기관 종사자 수
		소매업수	자동차업을 제외한 소매업 사업체 수

02. 데이터 수집_ 자료 수집

환경 	지속가능성	재생에너지비율	신재생에너지 지역별 공급비중
		폐기물재활용비율	재활용량/생활폐기물량(사업장, 일반가정)
	공기품질	CO2배출	지역별 co2배출량/ 1인당 GRDP
		미세먼지	도시별 미세먼지(PM2.5) 대기오염도 1년 평균
	도시환경	인당도시공원	인구천명당 도시공원조성면적 (국가도시공원+생활권공원+주제공원+조례가정하는공원)
교통 	교통편의성	자전거도로수	지역별 자전거 도로 노선 수
		도로혼잡비용	2016년 이후 혼잡비용 산출방식에 따른 교통혼잡비용 (차량운행 비용 + 시간가치 비용)
	도심교통	철도역수	각 시도별 철도 항목 역 수
		대중교통이용	요일별 대중교통 총이용인원의 평균
	교통접근성	BC*	전국 고속도로 네트워크를 통해 계산한 각 노드의 Betweenness Centrality 값의 평균
		CC*	전국 고속도로 네트워크를 통해 계산한 각 노드의 Closeness Centrality 값의 평균

* BC, CC의 경우 본 연구에서 교통접근성을 측정하기 위해 새롭게 추가한 지표, p12에서 설명

❖ 6가지 부분을 41개의 세부지표로 측정

❖ 2014년 – 2019년, 17개 시도별

❖ 부정적 요소 (회색 표시) :

인당 지방세, 청년실업률, 근로시간, 범죄발생률, 자살률, CO2배출량, 미세먼지, 도로혼잡비용

❖ 시도별 전체 연령의 데이터가 없을 경우 경제활동인구를 기준으로 자료 수집

❖ KOSIS 데이터를 중점으로 수집하였으며 부가적으로 MDIS 및 SDC의 데이터를 사용

해당 지표의 자료 출처 및 참고한 GPCI의 세부지표 설명은 PPT 뒷장에 수록

❖ 값이 존재하지 않을 경우 Nan값 처리

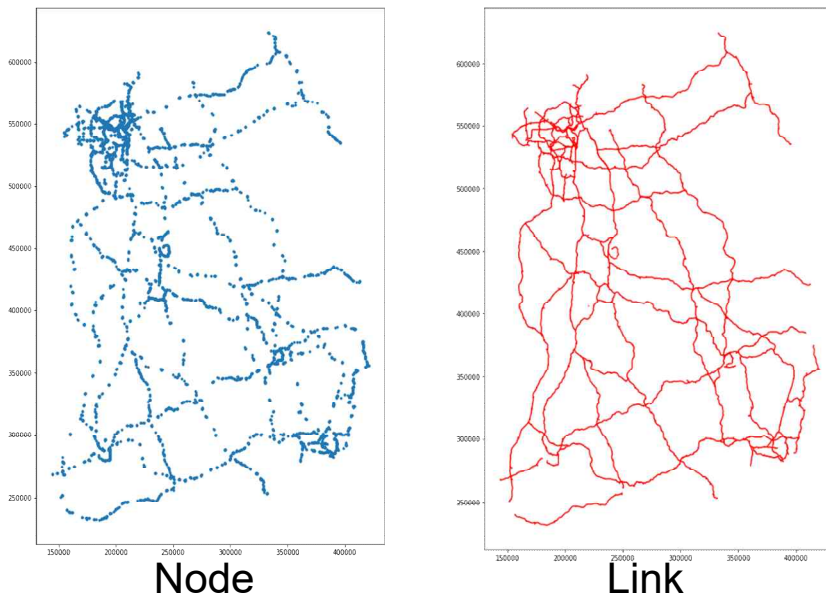
특히 세종특별자치시의 경우 한 항목의 전체년도 데이터 값이 없는 경우가 있어 주의 필요

02. 데이터 수집_ 도로네트워크

시도 간의 교통 접근성을 측정하기 위해 전국 도로 네트워크를 생성

- ❖ 국토교통부에서 제공하는 표준노드링크 정보에서 도로 등급 101과 102에 해당하는 고속국도와 도시고속국도를 추출, 이를 연결하는 노드를 추출하여 다음과 같은 고속도로 네트워크를 생성
- ❖ 노드 : 교차로, IC/JC 링크 : 도로
- ❖ 모든 노드의 매개중심성(BC)과 근접중심성(CC)을 구한 후 각 시도에 포함되는 노드들의 평균 값을 해당 지역을 대표하는 도심 접근성 지표로 사용

한국 도시 고속화도로 네트워크



매개 중심성 (Betweenness Centrality)

특정 노드가 네트워크에 존재하는 모든 노드들 간의 최단 경로에서 존재하는 빈도수

각 노드의 네트워크 상의 중요성을 의미. 매개중심성이 높은 노드(교차로)가 폐쇄된다면 도로 흐름에 큰 영향을 미침

접근 중심성 (Closeness Centrality)

특정 노드와 네트워크에 존재하는 다른 노드들 사이의 최단 거리를 평균 낸 값

접근 중심성이 큰 노드일 수록 이 노드에서 다른 노드로 상대적으로 빠르게 이동 가능하다는 것이며 허브 교차로일 확률이 크다.

이정원, 이강원 "네트워크 중심성 지표를 이용한 서울 수도권 지하철망 특성 분석",
한국철도학회논문집(2017.06) 20-3, P413-422을 참고하여 도시 접근성 지표를 BC,CC로 측정하였음

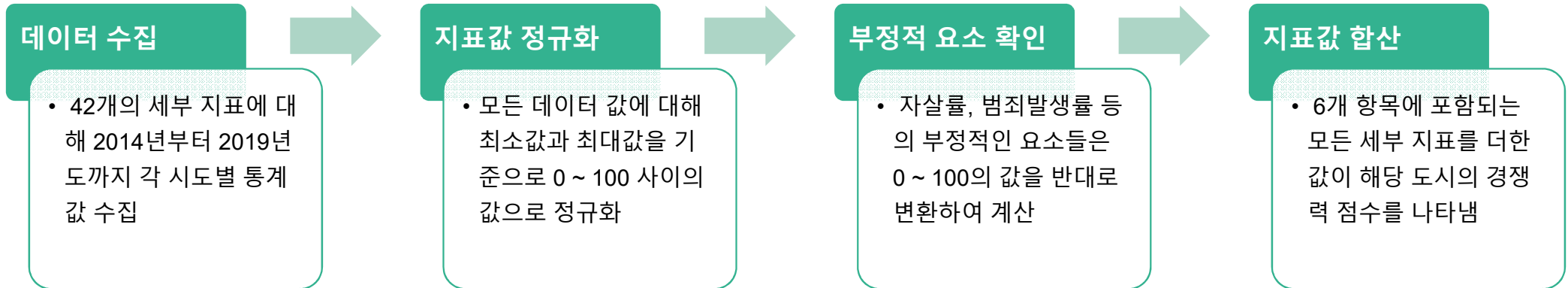
데이터 분석

도시경쟁력 지표 측정, 모델생성



03

03. 데이터 분석_도시경쟁력 지표 측정



	2014	2015	2016	2017	2018	2019
서울특별시	370.940941	381.473449	377.981070	383.863918	383.802443	368.918975
부산광역시	175.628774	179.780965	192.621626	190.804758	187.358542	155.109915
대구광역시	146.469118	148.488833	157.320011	157.332518	144.889618	124.230073
인천광역시	153.534722	159.584716	170.094353	172.360720	154.196428	140.504087
광주광역시	161.382513	164.300257	168.068573	178.669266	163.195448	139.565429
대전광역시	217.286580	213.759123	224.616776	203.630129	182.985015	167.311911
울산광역시	189.021038	173.340210	183.069278	189.014896	163.291626	145.619995
세종특별자치시	161.027645	153.679786	144.422028	143.955645	122.659793	136.769451
경기도	332.188296	349.625595	336.875043	326.355769	355.395266	347.141609
⋮	⋮					⋮

분석 내용

1. 전국 시도별 도시경쟁력 점수 확인
2. 2014 ~ 2019년도 도시 경쟁력 지표의 변화
3. 6개 항목별 TOP 5 도시 분석
4. 도시경쟁력과 전입 인구 사이의 연관성 분석

03. 데이터 분석_모델 생성

회귀분석을 위한 데이터처리

결측 값의 주변 데이터 분포에 따라 평균값,
근접데이터 사용하여 채우기
Nan값이 포함된 해당 도시 삭제

다중공선성 해결

과적합을 예방하고자 VIF(Variance
Inflation Factor) 및 피어슨 상수를
사용하여 독립인 변수만을 선택)

머신 러닝 - 회귀분석



데이터 분포 조정

이상치 데이터의 영향을 최소화 하고
독립변수의 데이터 분포를 비슷하게 하기
위해 Robust Scaling를 진행

성능이 제일 높은 모델 선택

Python 패키지 중 하나인 pycaret의
Compare Model 기능을 이용하여 MAE
기준으로 성능이 가장 높은 모델을 선택



a) 각 시도별 전입 인구와 도시 지표 6개 항목의 선형분석

b) 각 시도별 전입 인구와 6개의 항목의 Rank 선형분석

03. 데이터 분석_모델 생성

회귀분석을 위한 데이터처리

❖ 모든 년도에 대해서 데이터 합치기 __ 분석(a)

머신 러닝을 통한 모델의 훈련을 위해 6개년도 데이터를 모두 합쳐 데이터 수를 늘렸다.
102 rows (시도명_년도 17x6) x 7 columns (종속변수인 전입인구 + 6가지 항목 값)

❖ 시도간 도시 경쟁력 지표 차이 계산 __ 분석(b)

각 년도에 대해서 17개의 시도 간의 도시 경쟁력 지표 값의 차이를 계산하여 17 x 17 x 43의 3차원 행렬로 나타낸 후, 같은 도시에서의 차이를 제외하여 사용하였다.

272 rows (A시도에서B시도로) x 43 columns (종속변수인 A에서의 B로의 이동 인구수 + 6가지 항목 값)

❖ NaN (결측 값) 처리 __ 공통

각 시도의 6개년도 데이터를 사용하여, 해당 년도에 값이 없을 땐 3년도 미만일 경우 나머지 년도의 평균값 데이터를 사용하였고, 3개년도 이상일 경우 근접 년도의 데이터 값을 사용하였다.

데이터 분포 조정

❖ Robust Scaling __ 분석(b)

중앙 값이 0, IQR (제 3사분위수 - 제 1사분위수) 값이 1이 되도록 변환
이상치 데이터의 영향을 최소화 시키고 특히 시도간 도시 경쟁력 지표 차이 계산으로 인한
독립변수의 고르지 못한 데이터 분포를 조정하기 위해 사용하였다.
후에 기술할 학습에 쓰일 데이터와 성능 분석에 쓰일 예측 데이터를 따로 스케일링 하여
모델이 미리 예측 데이터를 학습하는 것을 방지하였다.

```
from sklearn.preprocessing import RobustScaler
robustScaler = RobustScaler()
print(robustScaler.fit(dataset))
data_use_robustScaled = robustScaler.transform(data_use)
data_use_robustScaled = pd.DataFrame(data_use_robustScaled, columns=data_use.columns)
data_use_robustScaled2 = data_use_robustScaled[data_use_robustScaled.iloc[:, :-1] != 0]
data_use_robustScaled2["순이동"] = data_use_robustScaled["순이동"]
data_use_robustScaled2.dropna(axis=0)
```

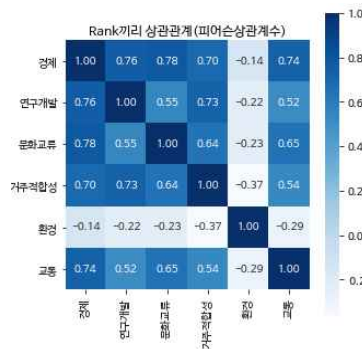

03. 데이터 분석_모델 생성

다중 공선성 해결

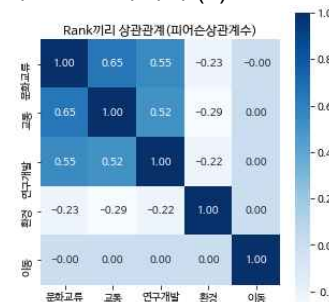
❖ VIF(Variance Inflation Factor)을 사용하여 독립인 변수만을 선택

회귀분석시 독립변수 간에 높은 선형관계가 존재할 경우(다중 공선성 문제) 분석 결과의 회귀 계수 (모델에서 해당 변수가 미치는 영향 정도를 의미)가 제대로 설명 되지 못한다. 이를 해결하기 위해 우리는 독립변수인 항목 간의 피어슨 상관계수를 시각화 하고, 다중공선성 문제가 있을 시에 VIF Factor 값이 가장 큰 항목을 제외하였다. 피어슨 상수가 0.7보다 작고, VIF 값이 2보다 작아 질때까지 이를 반복하여 다중공선성 문제를 해결하였다.

각 시도간 이동 인구와 6개의 항목의 지표 값 차이 선형분석에서 (b)



VIF Factor	features
0	5.852906 경제
1	3.141019 연구개발
2	2.917430 문화교류
3	2.885035 거주적합성
5	2.460271 교통
4	1.305744 환경



VIF Factor	features
1	1.938835 문화교류
3	1.888909 교통
0	1.541339 연구개발
2	1.099478 환경

```
#다중공선성값 측정 후 제거 VIF(Variance inflation Factors)
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor(valuedata.values, i) for i in range(valuedata.shape[1])]
vif["features"] = valuedata.columns
vif = vif.sort_values(by="VIF Factor", ascending=False)
```

```
#VIF 계수가 높은 feature 제거
v = valuedata
for i in range(3):
    li = list(vif.head(1)["features"])
    v = v.drop(li, axis=1)
    vif = pd.DataFrame()
    vif["VIF Factor"] = [variance_inflation_factor(v.values, i) for i in range(v.shape[1])]
    vif["features"] = v.columns
    vif = vif.sort_values(by="VIF Factor", ascending=False)
    vif
```

03. 데이터 분석_모델 생성

성능이 제일 높은 모델 선택

❖ Pycaret 사용하여 여러 모델 비교 __ 공통

AutoML을 지원하는 파이썬 라이브러리인 pycaret은 scikit-learn패키지를 기반으로 여러 모델을 생성 및 비교해준다.

연구 재현을 위해 session_id 는 42로 설정하였고, 데이터의 약 90%을 모델 훈련에 사용하였고 10%는 모델 성능 분석 및 예측에 사용하였다.

머신 러닝을 돌리기전 setup 기능을 이용하여 추후에 필요한 환경을 initialize 시켰다.

그 후 데이터 값이 크지 않아 머신 러닝의 Fold 값은 10으로 설정하여 pycaret에서 지원하는 선형 회귀분석의 여러 모델에 대해서 성능을 측정하고, MAE가 제일 낮은 모델을 선택하였다.

시도별 전입 인구와 6개의 항목의 지표 값의 선형분석에서 (a)

```
#data setup
clf = setup(data=data_use_robustScaled2, target = '순이동', session_id = 42)
```

	Description	Value
0	session_id	42
1	Target	순이동
2	Original Data	(92, 6)
3	Missing Values	False
4	Numeric Features	5
5	Categorical Features	0
6	Ordinal Features	False
7	High Cardinality Features	False
8	High Cardinality Method	None
9	Transformed Train Set	(64, 5)

```
bestmodel = compare_models(10)
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	0.7683	1.5161	1.1324	0.6632	0.2951	2.3912	0.409
gbr	Gradient Boosting Regressor	0.7753	1.7574	1.1817	0.6423	0.3274	2.8019	0.041
et	Extra Trees Regressor	0.7895	1.7604	1.1852	0.5907	0.3096	1.8945	0.354
ada	AdaBoost Regressor	0.8057	1.3491	1.1319	0.4685	0.2960	2.6422	0.062
dt	Decision Tree Regressor	0.8380	2.1511	1.2130	0.4015	0.3627	2.9953	0.013
llar	Lasso Least Angle Regression	1.6427	9.0899	2.7687	-0.4453	0.7572	2.5811	0.012
lasso	Lasso Regression	1.6612	9.6870	2.8541	-0.5186	0.7110	2.2405	0.013
knn	K Neighbors Regressor	1.5146	9.3700	2.6501	-0.5581	0.5780	2.3246	0.062
lightgbm	Light Gradient Boosting Machine	1.8241	10.5870	2.9769	-0.6897	0.6385	5.3681	0.063

분석 결과 및 해석

시도별 도시경쟁력 측정, 회귀모델 분석

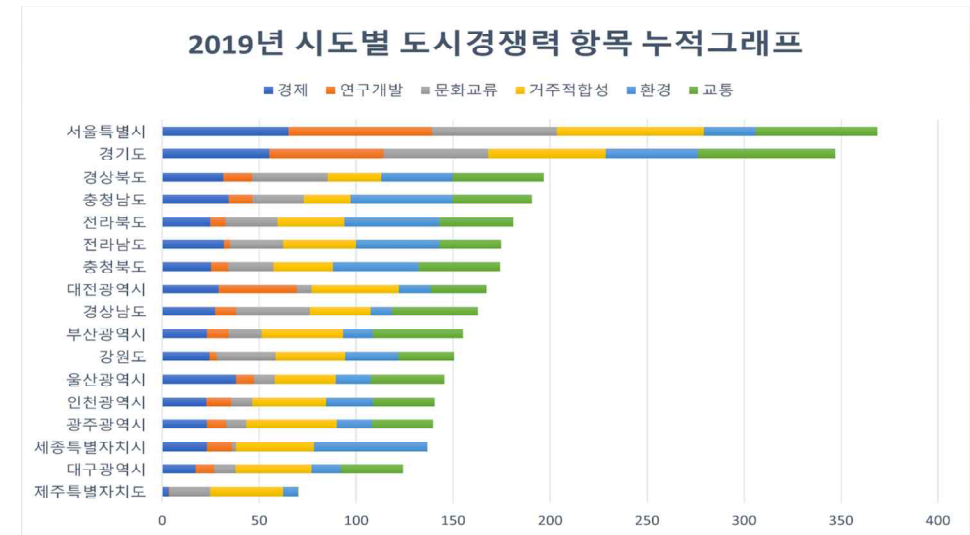


04

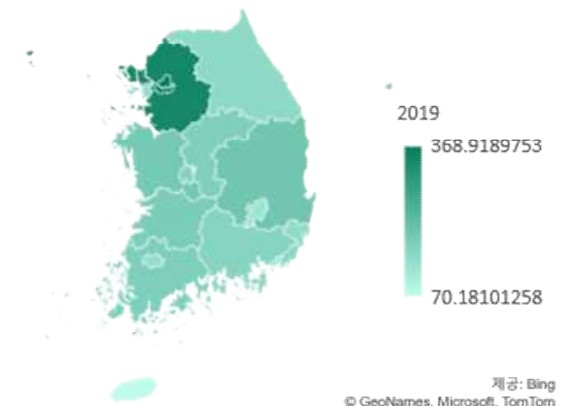
04. 분석 결과 및 해석_ 시도별 도시경쟁력 측정

2019년 기준 17개 시도의 도시 경쟁력 측정 및 분석

- ❖ 서울특별시와 경기도가 각각 369점과 347점으로 다른 도시들에 비해 압도적으로 높은 점수를 보인다. 이를 통해 수도권에 인구집중의 원인이 인구 특성을 떠나 도시 자체의 경쟁력 (도시의 매력도)가 무척이나 높다는 이유도 있다는 걸 알 수 있다.
- ❖ 반면, 제주특별자치도는 약 70점으로 제일 낮은 점수를 기록했으며, 특히 섬이라는 지리적 특성 때문에 연구개발과 교통 측면 점수는 거의 0점에 가깝다.
- ❖ 광역시와 수도권을 제외한 지역들의 경우 환경 점수가 커 전체적인 도시경쟁력 지표가 높아졌다.
- ❖ 세종의 경우에는 거주적합성과 환경의 항목 값이 높다. 이는 세종이 도시를 계획할 때 인구이동을 유인하기 위해 환경과 거주환경을 고려하였다고 추측할 수 있다



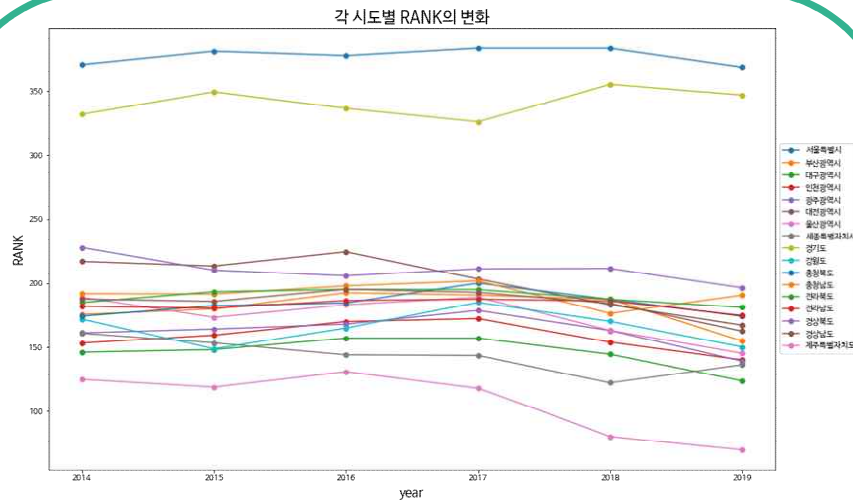
2019년 17개 시도별 도시경쟁력 지도



04. 분석 결과 및 해석_ 시도별 도시경쟁력 측정

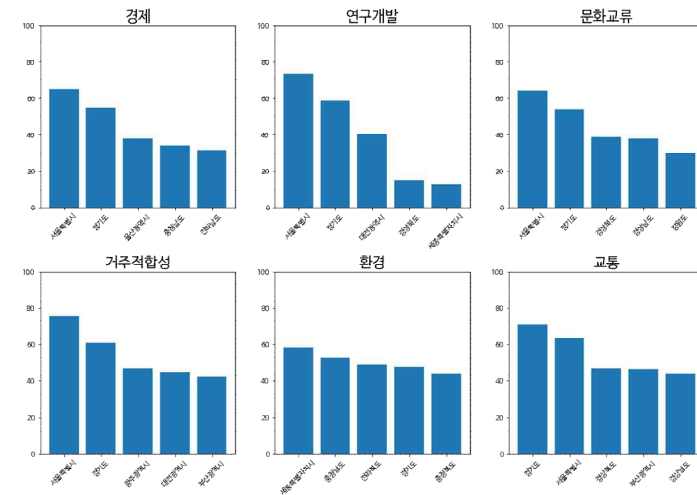
2019년 기준 17개 시도의 도시 경쟁력 측정 및 분석

도시 경쟁력 지표의 변화



- ❖ 2014년부터 2019년까지 지역별로 모든 세부지표의 값을 합산한 경쟁력 지표의 변화를 그래프로 나타내었다
- ❖ 서울과 경기도는 높은 순위를 유지하고 있으며, 이외 지역은 150 ~ 200 사이에서 유사한 값을 가진다
- ❖ 과거 데이터를 기반으로 다양한 시계열 분석 알고리즘을 통해 2021년 이후 도시 지표의 변화를 예측할 수 있다

항목별 Top 5 도시 분석

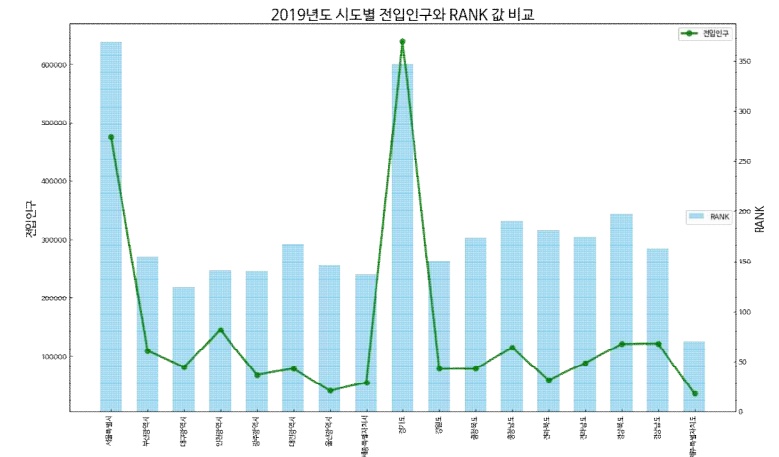


- ❖ 경제는 수도권을 제외하고 울산광역시와 충청남도가 높음
- ❖ 연구 개발 항목에서는 대전광역시가 상대적으로 높은 점수를 얻음
- ❖ 거주 적합성은 서울과 경기도 다음으로 광주, 대전, 부산 등의 대표 광역시가 높은 값을 가짐
- ❖ 환경은 2012년 설립된 세종이 가장 높은 점수를 얻었으며, 서울을 제외한 지방도시가 높은 순위를 보임

04. 분석 결과 및 해석_ 시도별 도시경쟁력 측정

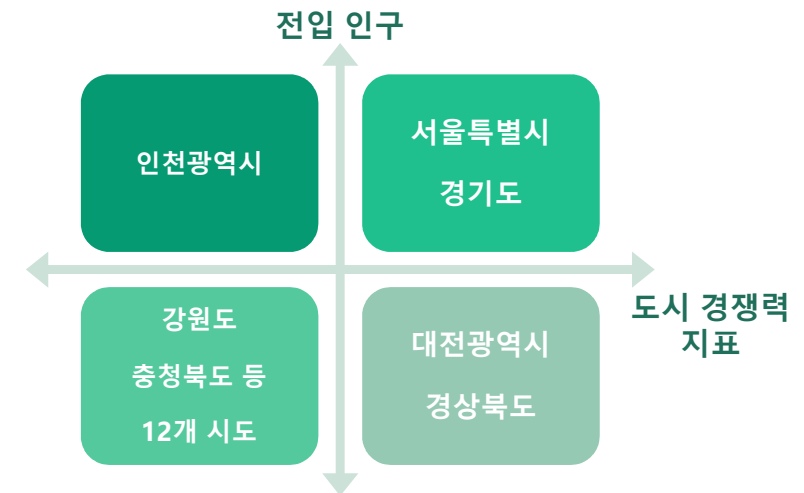
도시경쟁력 지표와 전입 인구와의 상관관계 분석

- ❖ 오른쪽 그래프는 2019년도의 시도별 전입인구와 도시경쟁력 지표를 나타낸다
- ❖ 이와 같은 시각적 분석을 통해 전입인구와 도시경쟁력 지표 사이의 연관성이 존재하는 것을 알 수 있다
⇒ 머신 러닝을 통한 전입인구 예측 등 심층적 분석을 진행하였다



도시 경쟁력 지표 기반 17개 시도 분류

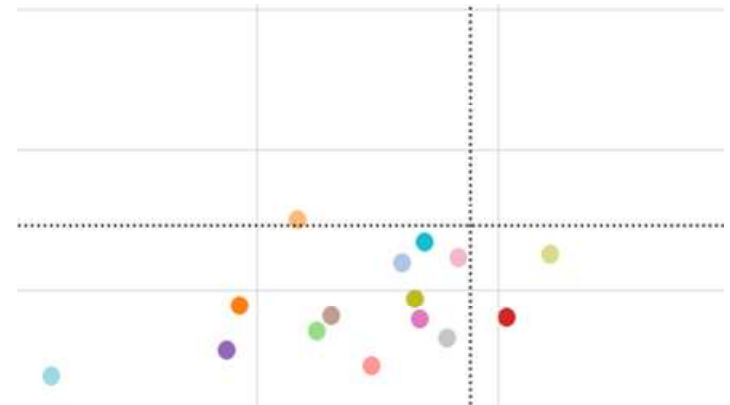
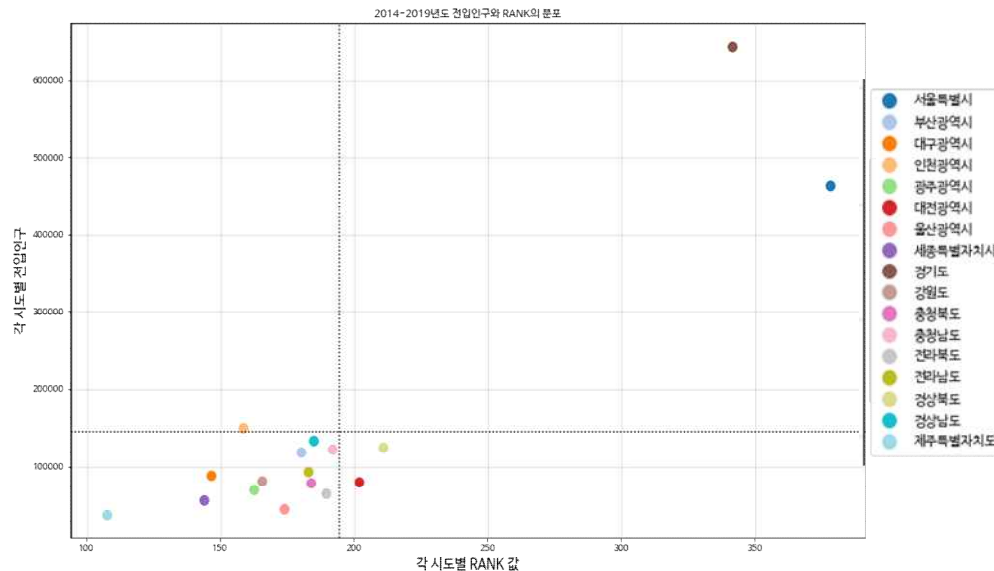
- ❖ 17개 시도를 전입인구와 도시 경쟁력 지표를 토대로 아래 그림과 같이 4개의 클러스터로 분류하였다
- ❖ 2014년부터 2019년까지 모든 도시의 전입 인구 수 평균과 도시경쟁력 지표 평균을 기준으로, 전입인구와 도시 경쟁력 지표를 분류하였다
- ❖ 인천광역시는 수도권에 속하는 지역이기 때문에 전입인구는 많지만 경쟁력 지표는 상대적으로 작다
- ❖ 대전과 경북지역은 전입인구는 평균 이하이지만 도시의 발전 수준은 평균 이상으로 높다



04. 분석 결과 및 해석_ 시도별 도시경쟁력 측정

도시 경쟁력 지표 기반 17개 시도 분류

- ❖ 우리는 여기서 낮은 Rank를 지니고 있지만 전입 인구가 많은 도시와 높은 Rank를 지니고 있지만 전입 인구가 적은 도시들의 특징을 살펴볼 필요가 있다. 다음 그림에서 볼 수 있듯이 전자의 경우인 인천광역시는 높은 Rank를 지닌 서울특별시와 경기도와 인접해 있어 자체의 도시 Rank는 낮게 측정되어도 거주하고 있는 주민이 느끼는 Rank의 값은 더 높을 것이다. 반면 후자의 경우, 대전광역시와 경상북도는 인접한 Rank가 큰 도시가 없기에 도시의 Rank값이 평균 이상임에도 불구하고 주민이 느끼는 실질적 Rank값은 낮을 것이다.



04. 분석 결과 및 해석_모델 분석

a) 각 시도별 전입 인구와 도시 지표 6개 항목의 선형분석

- ❖ 시도별 전입인구 ~ 문화교류, 교통, 연구개발, 환경
- ❖ Random Forest Regressor Model 사용
- ❖ 경제, 거주적합성을 제외하고 최종적으로 Regrssion model 학습에 사용된 데이터는 (102,5)이며, 이 중에서 (92,5)을 학습에 (10,5)을 만든 모델의 성능 테스트 (예측)에 사용

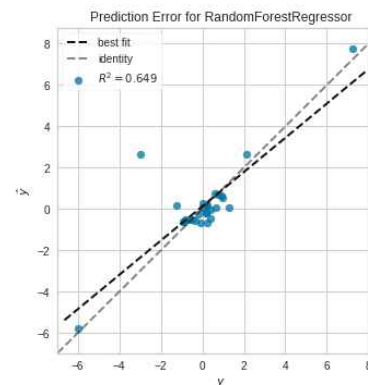
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	0.5761	1.5469	1.2437	0.6165	0.2332	1.219

모델링 결과

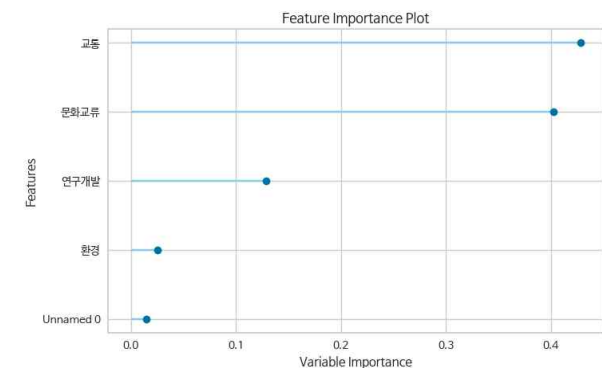
RF 모델은 MAPE 약 1.2%의 오차로 전입 인구를 예측하며, Prediction Error 그래프에서 $R^2 = 0.61$ 로 다소 높은 선형성을 보이고 있다. 또한, 도시로 들어오는 전입 인구에 가장 큰 영향을 미치는 요소는 교통과 문화교류이다.

Unnamed: 0	교통	문화교류	환경	연구개발	이동인구	Label
0	0.821782	-2.383140	-1.105825	1.274260	0.158786	1.622062
1	0.841584	2.354439	1.885704	0.448484	5.410729	8.497389
2	0.861386	-0.458519	0.521978	-1.090984	-0.908017	0.254768
3	0.881188	0.415557	0.121939	0.192278	-0.322196	0.310791
4	0.900990	0.362212	0.296416	0.823901	0.095778	0.102626
5	0.920792	0.159389	0.303089	0.568088	-0.414339	-0.638190
6	0.940594	-0.259762	0.354611	0.123330	-0.971110	-0.392780
7	0.960396	0.738537	1.005545	-0.350609	0.382452	-0.131443
8	0.980198	0.558665	0.961812	-2.318309	-0.083420	-0.425130
9	1.000000	-2.368103	-0.029602	-2.575140	-1.256479	0.333783

실제 전입인구 수(이동인구)와
모델의 예측값(Label)



Prediction Error



Feature Importance Plot

04. 분석 결과 및 해석_모델 분석

b) 각 시도 간의 이동 인구와 도시 지표 6개 항목의 선형분석

- ❖ 시도간 이동인구 ~ 문화교류, 교통, 연구개발, 환경
- ❖ Extra Tree Regressor 모델 사용
- ❖ 경제, 거주적합성을 제외하고 최종적으로 regression model 학습에 사용된 데이터는 (1734,5) 이며, 이 중에서 (1561,5)을 학습에 (173,5)을 만든 모델의 성능 테스트 (예측)에 사용

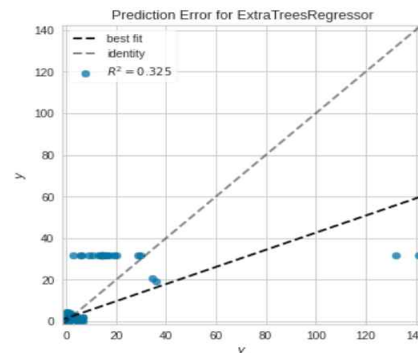
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Extra Trees Regressor	1.8448	63.9674	7.998	0.3251	0.4377	3.9958

모델링 결과

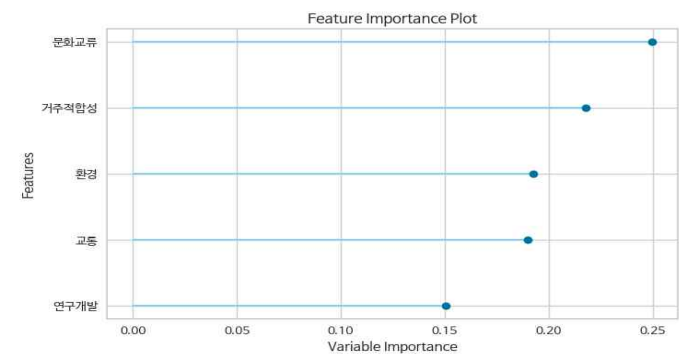
ET 모델로 도시 간의 연간 이동 인구를 모델링 하였을 때, MAPE 약 4.0%의 오차를 가지며, 결정계수는 $R^2 = 0.325$ 로 다소 낮은 설명력을 가지고 있다. 이는 앞서 언급한 대로 인접 도시간의 영향성을 고려하지 못하였으며, 년도에 따라 주가지역을 선택하는 요인이 달라지는데 이를 고려하지 못하였기 때문이라 사유 된다.

	거주적합성	연구개발	문화교류	교통	환경	이동	Label
0	-1.025738	1.668067	-0.416207	-1.439902	2.246275	-0.255684	-0.225263
1	-0.912480	-0.607347	-0.041714	-1.970571	-1.186729	-0.050737	0.277617
2	0.400353	-0.276248	-0.256954	-0.318907	-0.422345	0.549684	0.200558
3	-2.266411	-3.766822	-2.096160	-1.562557	-0.371989	0.753368	1.256786
4	-2.512763	-2.901573	-2.485301	-1.284424	0.375305	4.379684	0.799006
...
464	-0.352137	0.276045	-0.013479	1.015572	0.563117	-0.275158	0.285155
465	1.893237	4.211370	1.883823	0.182913	-0.429741	1.700211	1.160551
466	0.920876	-1.952033	0.818326	1.548947	-1.268458	-0.321789	-0.298672
467	0.000000	0.000000	0.000000	0.000000	0.000000	15.523052	31.687737
468	-2.249039	-4.156759	-2.086991	-1.331676	-0.521055	0.054632	0.940859

실제 도시 간 이동인구(이동)와
모델의 예측값(Label)



Prediction Error



Feature Importance Plot

결론 및 한계점

결론, 한계점 및 향후 연구 방향



05°

결론

Conclusion

- 약 40여개의 지표를 이용하여 2014-2019년까지 시간의 흐름에 따라 우리나라의 시도별 도시 경쟁력이 어떻게 변화하는지 6가지 항목에 대해서 파악할 수 있었다. 이를 이용하여 각 시도별로 도시 발전 정도의 특징을 세부적으로 분석하고, 수도권 인구 집중 현상을 설명하였다.
- 인구 이동에 대한 원인을 주로 설문조사를 통해 파악한 통계자료들이 다수 존재, 본 연구에서 더 나아가 제안한 도시경쟁력 지표와 각 시도별 전입 인구, 이동인구와의 상관관계를 분석함으로써 정량적인 방법으로 이동 인구의 원인을 분석한 통계자료를 제시하여 수도권 쏠림 현상을 해결하기 위한 지방 발전 방향 설립에 필요한 정보를 제공한다.
- 시도별 전입인구와 항목별사이에는 선형 관계가 있었으며 Feature Importance 분석을 통해 6가지의 항목 중에 교통, 문화교류 (경제, 거주적합성은 제외) 가 도시로 인구가 유입되는 데 중요한 영향을 미치는 것을 알 수 있다.
- 시도간 이동인구와 시도간 경쟁력 지표 차이에서는 선형 관계가 없는 것을 봤을 때, 좀 더 낮은 단계인 시도간 이동에서는 두 도시의 경쟁력 차이 외에 다른 요인이 있음을 추측할 수 있다.
- 수도권 쏠림 현상을 완화하기 위해서는 경제적인 측면에서의 발전 뿐만 아니라 문화 관광 자원의 확대, 도심 내 교통시설의 확충 등의 도시의 종합적인 경쟁력을 끌어올리는 정책이 필요하다.

한계점 및 향후 연구 방향

Limitation

- 자료 수집의 한계로 인하여 GPCI의 모든 세부 지표 내용을 사용하지 못하였고 우리나라 사람들이 선호하는 거주환경에 맞는 다른 세부 지표의 추가가 필요하다.
- 동일 도시의 동일 항목일지여도 개인의 특성에 따라 (특히 연령) 인식하는 경쟁력 지표 값의 차이가 크다. 하지만 우리는 개인의 특성을 본 연구에서 고려하지 않았다.
- 머신 러닝을 통한 모델 학습 시에 시간대가 서로 다른 데이터를 한번에 사용하였으며 다중선공성 문제가 발생하여 6가지 모든 항목에 대하여 인구이동에 영향을 끼치는 중요도를 분석하지 못하였다
- 도시 경쟁력 지표와 인구이동 간의 관계를 단순 선형 관계에서만 분석을 진행하였다.

Future work

- 도시 경쟁력 지표 측정시에 사람들이 이주 도시 선택 시 중요하게 생각하는 항목의 순서로 가중치를 부여한다면, 인구이동 현상을 분석하는데 더 유용한 도시 경쟁력 지표가 될 것이다.
- 이동 인구가 아닌 해당 도시의 거주 인구와의 관계를 분석해 볼 수도 있다
- 각 세부 지표 점수의 시계열 예측을 통해 미래 시점에서의 도시 경쟁력 지표를 예측하고, 해당 모델을 발전시켜 미래 시점에서의 도시 경쟁력 지표를 이용한 미래 유입 인구 예측 모델을 제작한다. 이를 이용하면 수도권 인구 집중 현상을 완화하기 위한 여러 정책에 대해 해당 모델을 이용한 시뮬레이션이 가능하다.

(부록) 데이터 출처

데이터 출처	변수 이름
KOSIS - 시도별 경제활동별 지역내총생산	GRDP
KOSIS - 시도별 1인당 지역내총생산, 지역총소득, 개인소득	1인당GRDP
KOSIS-경제성장률(시도)	경제성장률
통계데이터센터 - 기업통계등록부	매출top기업수
KOSIS - 시군(9개도)/연령별 취업자 및 고용률 (전체)	취업인구
통계데이터센터 - 기업통계등록부	서비스직비율
KOSIS - 행정구역(시도)/ 산업/규모별 임금 및 근로시간	임금
KOSIS - 지방세 1인당, 세대당 부담현황	인당지방세
KOSIS - 연구수행주체별 연구원(지역)	연구원비율
konema - THE World University Ranking : Full List	top대학수
KOSIS - 연구수행주체별 연구개발비(지역)	인당연구개발비
konema - THE World University Ranking : Full List	유학생수
지식재산통계서비스 - 시도별 등록	특허등록수
통계데이터센터 - 기업통계등록부	스타트업수
KOSIS - 지방자치단체 : 총예산/문화예산/문화예술예산/공연예술예산 지역별 현황	문화예산비율
KOSIS - 시도지정문화재현황	문화재수
KOSIS - 공연시설 : 공연시설 내 보유 공연장 현황	극장수
KOSIS - 인구 십만명당 문화기반시설수(시도/시/군/구)	문화기반시설수
KOSIS - 시도/객석수 규모별 음식점업 현황	호텔수
KOSIS - 시도/객석수 규모별 음식점업 현황	영업점포
KOSIS - 시도/객석수 규모별 음식점업 현황	식당수
KOSIS - 시도별 외국인주민 현황	외국인거주자수
KOSIS - 한국 여행 시 방문지(17개 시도)	외국인방문객 수

데이터 출처	변수 이름
KOSIS - 청년실업률(시도)	청년실업률
KOSIS - 행정구역(시도)/ 산업/규모별 임금 및 근로시간	근로시간
KOSIS - 인구 천명당 범죄발생건수(시도)	범죄발생률
KOSIS - 시군구/사망원인(50항목)/성/ 사망자수, 사망률, 연령표준화 사망률(1998~)	자살률
KOSIS - 인구 천명당 의료기관 종사 의사수(시도/시/군/구)	의사비율
KOSIS - 전국, 산업별, 성별, 규모별 사업체수 및 종사자수(종사상지위별)	소매업수
KOSIS - 전국, 산업별, 성별, 규모별 사업체수 및 종사자수(종사상지위별)	음식주점업수
KOSIS-지역별 신·재생에너지 생산량(비재생폐기물 제외, 2019년 4/4분기~)	재생에너지비율
KOSIS-폐기물 처리현황_생활계폐기물(생활폐기물_사업장생활계폐기물)	폐기물재활용비율
국가미세먼지정보센터 - 시도별 배출량	CO2배출
KOSIS-미세먼지(PM2.5) 월별 대기오염도(측정망별, 시도별, 도시별, 측정지점별)	미세먼지
KOSIS - 인구 천명당 도시공원조성면적(시도)	인당도시공원
국가교통DB - 자전거도로 및 주차시설 현황	자전거도로수
한국교통연구원 - 도로교통혼잡비용	도로혼잡비용
KOSIS - 교통시설 : 도로 등	철도역수
국가대중교통DB - 지역내 대중교통	대중교통이용
국토교통부_국가교통정보 표준노드링크 정보	BC
국토교통부_국가교통정보 표준노드링크 정보	CC
MDIS - 국내이동인구통계	전입 인구

(부록) GPCI

GPCI 참고	변수 이름
GDP	GRDP
GDP per capita	1인당GRDP
GDP의 연평균 성장률	경제성장률
Fortune의 "Fortune Global 500"에 포함된 대상 도시에 위치한 회사의 총 점수	매출top기업수
해당 도시의 취업자 수	취업인구
대상 도시의 서비스 직원 비율	서비스직비율
UBS의 "가격 및 수입"에 제공된 대상 도시의 임금 수준	임금
KPMG의 "기업 세율표"에서 대상 도시 또는 대상 도시 국가의 법인 세율.	인당지방세
유네스코 통계 연구소의 "UIS 통계"에서 국가 및 대상 도시의 직원 수에 비례하여 추정	연구원비율
THE 세계 대학 순위에서 상위 1000위 대학 중 도심 50km 이내에 위치한 대학의 랭킹 점수	top대학수
UIS 통계에 나열된 해당 국가 및 대상 도시의 직원 수에 비례하여 추정	인당연구개발비
THE의 세계 대학 순위에서 상위 1000위 안에 드는 각 대학의 학생 수와 유학생 비율로 추정되는 대학에 다니는 유학생 수	유학생수
지난 10년 동안의 특허 등록 평균을 기반으로 대상 도시의 직원 수에 비례하여 추정.	특허등록수
지난 3년 동안 대상 도시에서 설립된 평균 스타트업 수	스타트업수
문화 이벤트 수	문화예산비율
유네스코 세계 문화 유산의 규모와 유형에 따라 부여되는 총 포인트	문화재수
트립 어드바이저에 나열된 극장 및 콘서트 홀	극장수
De Gruyter Saur에 등록된 박물관 수	문화기반시설수
Hotels.com에 표시된 도심에서 10km 이내에 위치한 총 5성급 호텔 객실 수	호텔수
트립 어드바이저에 나열된 대상 도시에 위치한 모든 상업 및 쇼핑물에 대한 평점의 총 값	영업점포

GPCI 참고	변수 이름
La Liste의 세계 Top 1000 레스토랑 중 도심에서 10km 이내에 위치한 레스토랑 수	식당수
대상 도시의 국가에 등록 된 외국인 또는 시민권이 없는 거주자 수	외국인거주자수
해당도시를 방문하는 외국인 방문객 수	외국인방문객 수
대상 도시의 총 실업률	청년실업률
UBS의 "가격 및 수입"에 제공된 대상 도시의 근무 시간	근로시간
대상 도시의 인구 100만 명당 연간 살인 수	범죄발생률
장에 보정 수명에 대한 점수와 인구 십 만명당 자살률의 평균	자살률
OECD의 보건통계와 WHO의 세계보건관측소에 등재된 각 도시 직원 수 대비 평균 의사 수	의사비율
도시 중심에서 10km 이내에 위치한 OpenStreetMap에 등록된 소매점 수	소매업수
도시 중심에서 10km 이내에 위치한 OpenStreetMap에 등록된 음식점 수	음식주점업수
총 1차 에너지 공급 대비 재생에너지 공급 비율	재생에너지비율
재활용된 도시 폐기물의 평균 비율	폐기물재활용비율
연료 연소로 인한 CO2 배출량을 인당GDP로 나눈 값	CO2배출
IQAir에서 대상 도시의 연간 평균 PM2.5 농도	미세먼지
Numbeo의 "Pollution"에서 대상 도시에 대한 "녹색 및 공원의 질" 점수	인당도시공원
택시 또는 자전거의 의한 이동 편의성 평균	자전거도로수
TomTom의 "Traffic Index"에 따라 교통 혼잡으로 인해 발생한 평균 추가 이동 시간의 백분율	도로혼잡비용
도심 10km 이내에 OpenStreetMap에 나열된 기차 및 트램 정류장의 밀도.	철도역수
Numbeo의 Traffic에 따른 도시의 대중 교통 사용 비율.	대중교통이용
-	BC
-	CC

(부록) 참고 문헌

The Mori Memorial Foundatoin(2021). **Global Power City Index(GPCI)-2020 YEARBOOK.**

정면진,강동규(2016). 수도권 주거 이동 가구의 주거입지 선택 요인 분석 . *지역인구*, 제32권 제1호

최열, & 김형준. (2012). 수도권 및 비수도권의 주거이동 결정요인 비교 분석. *국토계획*, 47(4), 219-231.

최진호. (2008). 한국 지역 간 인구이동의 선별성과 이동 이유: 수도권을 중심으로. *한국인구학*, 31(3), 159-178.

최은영. (2004). 선택적 인구이동과 공간적 불평등의 심화: 수도권을 중심으로. *한국도시지리학회지*, 7(2), 57-70.

이성우. (2002). 지역특성이 인구이동에 미치는 영향: 독립이동과 연계이동. *지역연구*, 18(1), 49-82.

이정원, & 이강원. (2017). 네트워크 중심성 지표를 이용한 서울 수도권 지하철망 특성 분석. *한국철도학회 논문집*, 20(3), 413-422.