감염병 대응 빅데이터 활용 아이디어 공모전

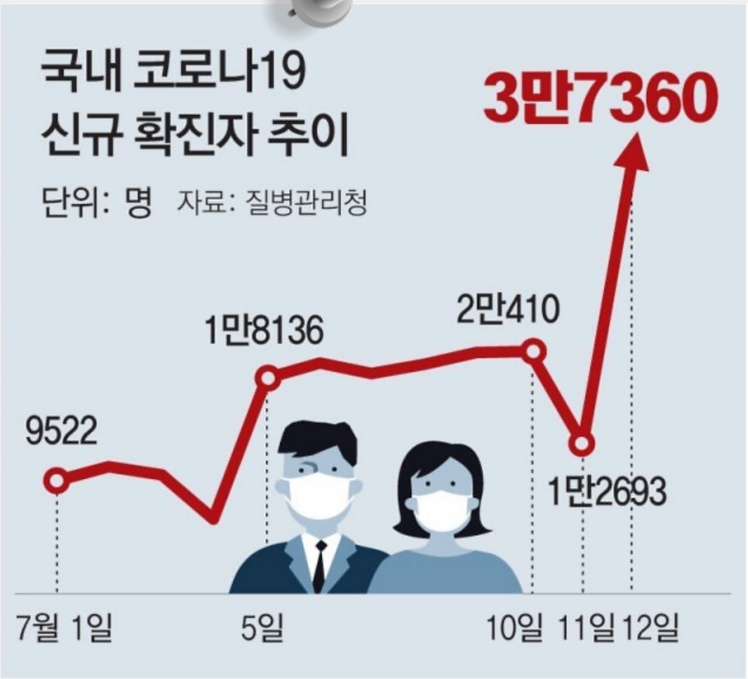
김대양
태정수

인구데이터 및 감염병 데이터 분석을 통한 독거노인 취약지역 판별 및 모델 제안

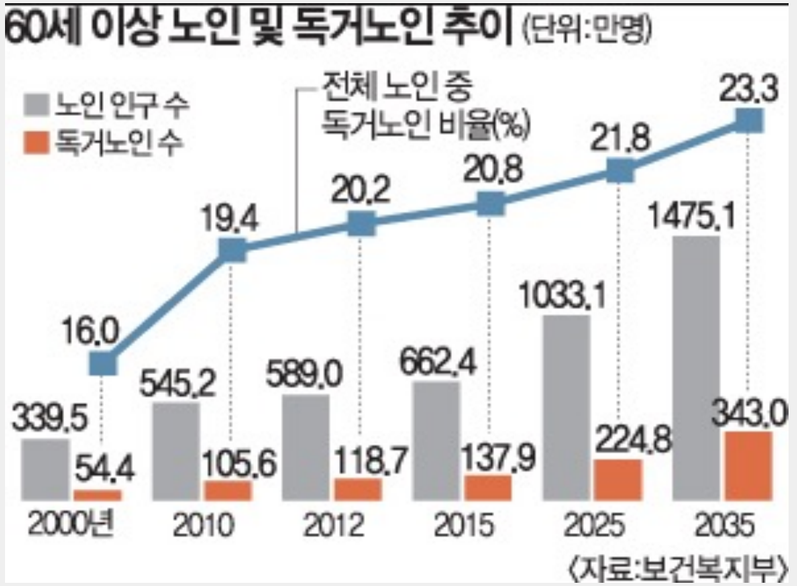
목차

01 추진 배경

주거 취약지역에 감염병 위험 높음

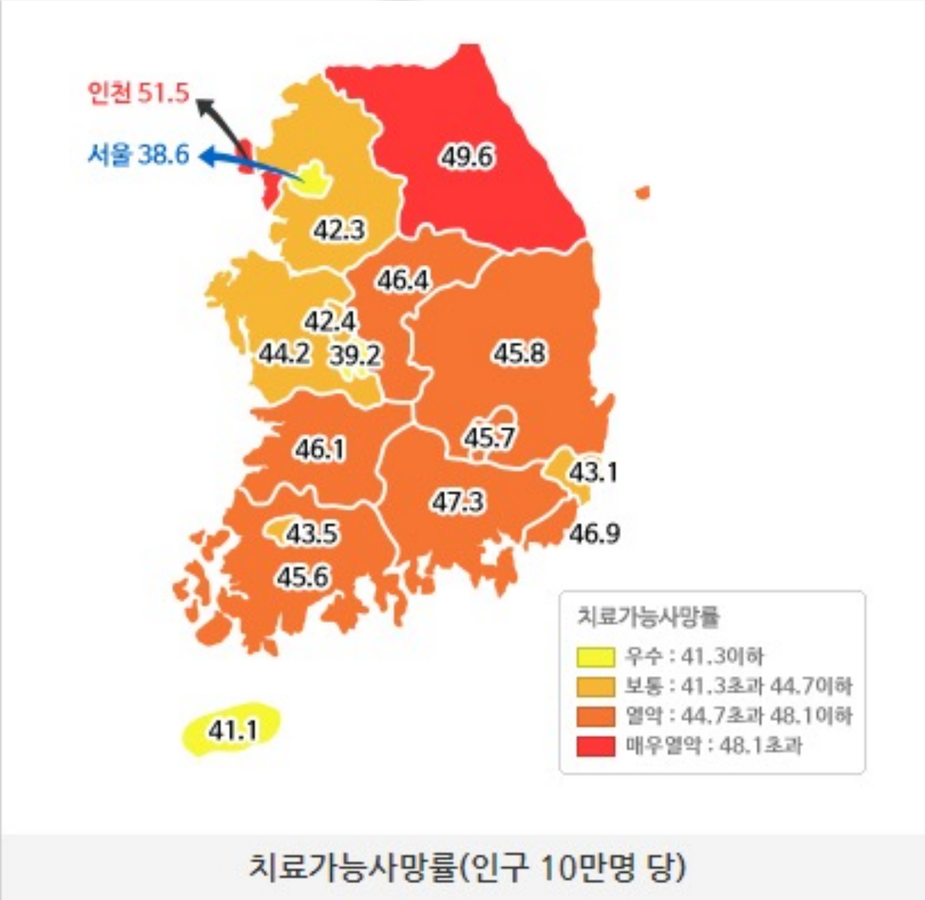


신종코로나 증가



최근 독거노인 수 증가

01 필요성



치료가능사망률(인구 10만명 당)

자료:공공의료연계망

2022년 인구데이터와 독거노인 데이터를 토대로 상대적으로 취약한 지역을 찾아내서 , 미리 예측할 수 있는 모델을 만듬으로서

독거노인에 대한 의료 위기를 미리 예방할 수 있게 하는 시스템이 필요하다고 생각합니다.

그래서 현재 지역별 시설 현황과 인구데이터를 분석하여 어느 지역이 개선이 필요한지 분류 할 수 있는 모델을 만들고자 합니다.

02 데이터 분석

활용 데이터 소개

연령별 감염병 환자 정보

1. 수집 플랫폼 : 감염병 빅데이터 거래소
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석

연령별 감염병 환자 발생 비율

1. 수집 플랫폼 : 감염병 빅데이터 거래소
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석

지역별 감염병 환자 정보

1. 수집 플랫폼 : 감염병 빅데이터 거래소
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집 분석 및 모델링

연령별 감염병 환자 발생 비율

1. 수집 플랫폼 : 감염병 빅데이터 거래소
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집 분석 및 모델링

02 데이터 분석

행정안전부
주민등록
인구통계

1. 수집 플랫폼 : 행정안전부
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집 분석 및 모델링

KOSIS
국가통계
포털

1. 수집 플랫폼 : 국가통계포털
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집 분석 및 모델링

02 데이터 분석

시도별 응급의료기관 및 응급의료시설

1. 수집 플랫폼 : 응급의료통계포털 MEDIS
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집 분석 및 모델링

내원요일별 응급실 이용(성별 + 연령별)

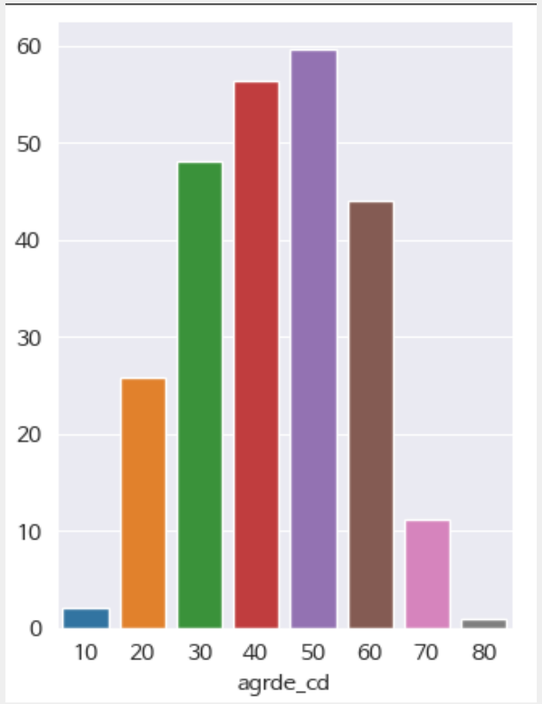
1. 수집 플랫폼 : 응급의료통계포털 MEDIS
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집 분석 및 모델링

내원요일별 응급실 이용(시도별)

1. 수집 플랫폼 : 응급의료통계포털 MEDIS
2. 관측 연도 : 2022년
3. 활용 단계 : 감염병 데이터 분석 , 군집

02 데이터 분석

감염병 빅데이터 거래소



연령별 환자 수

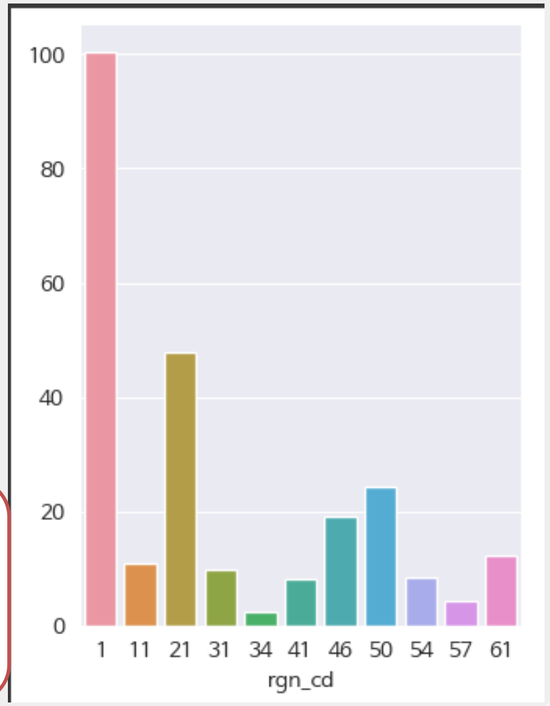
연령 데이터가 들어있는 2022년 환자 정보 데이터를 살펴 볼때 , 연령별로 그룹핑을 해서 환자 수를 살펴보면 , 30 ~ 60 부근에서 가장 많은 환자 발생 수를 볼 수 있었습니다.

지역 데이터가 들어있는 2022년 환자 정보 데이터를 살펴 볼때 , 지역별로 그룹핑을 해서 환자 수를 살펴보면 , 서울 , 인천 부근에서 가장 많은 환자 발생 수를 살펴 볼 수 있었습니다.

< 지역 라벨링 >

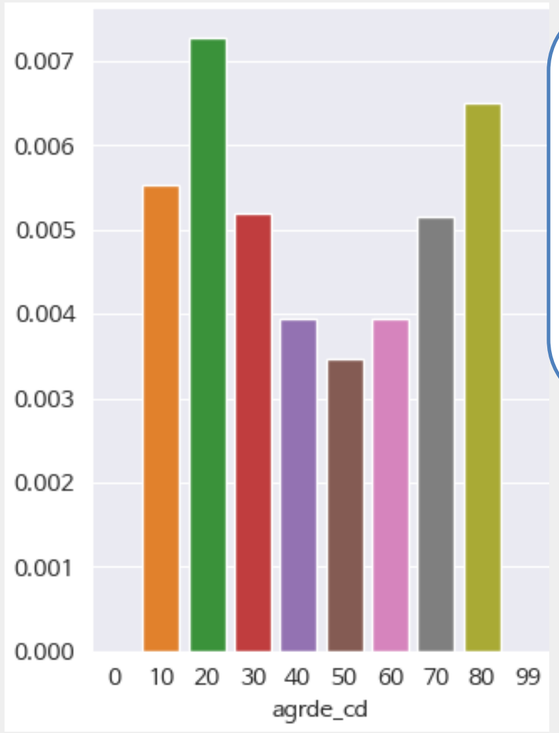
01 : 서울 11 : 경기 21 : 인천 31 : 충남 34 : 대전
41 : 대구 46 : 부산 50 : 경남 54 : 전북 57 : 전남
61 : 광주

지역별 환자 수



02 데이터 분석

감염병 빅데이터 거래소

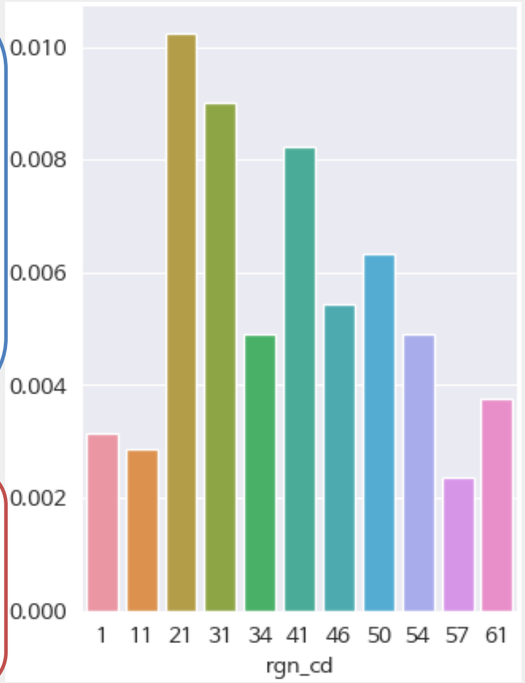


연령별 감염병 환자 비율

전체 환자수와 감염병 환자수가 들어있는 데이터에서 감염병 환자 비율을 , 연령별로 살펴보면 , 20대가 가장 많은 걸로 보인다 , 하지만 전체 환자 수를 보면 40 ~ 50대가 전체적인 환자수가 많이 있기 때문에 ,
결국에는 40 ~ 50 대에서 병에 걸릴 위험이 높다는 것을 파악할 수 있었습니다.

전체 환자수와 감염병 환자수가 들어있는 데이터에서 감염병 환자 비율을 , 지역별로 살펴보면 , 인천 , 충남 , 대구 등에서 감염병 환자 비율이 많은 것으로 보이지만 , 전체데이터에서 살펴 볼때는 전체 환자수 자체는 서울이 가장 많았기 때문에 ,
수도에서 감염병 환자와 그 외의 환자가 양쪽에서 많다는 것을 파악할 수 있었습니다.

지역별 감염병 환자 비율

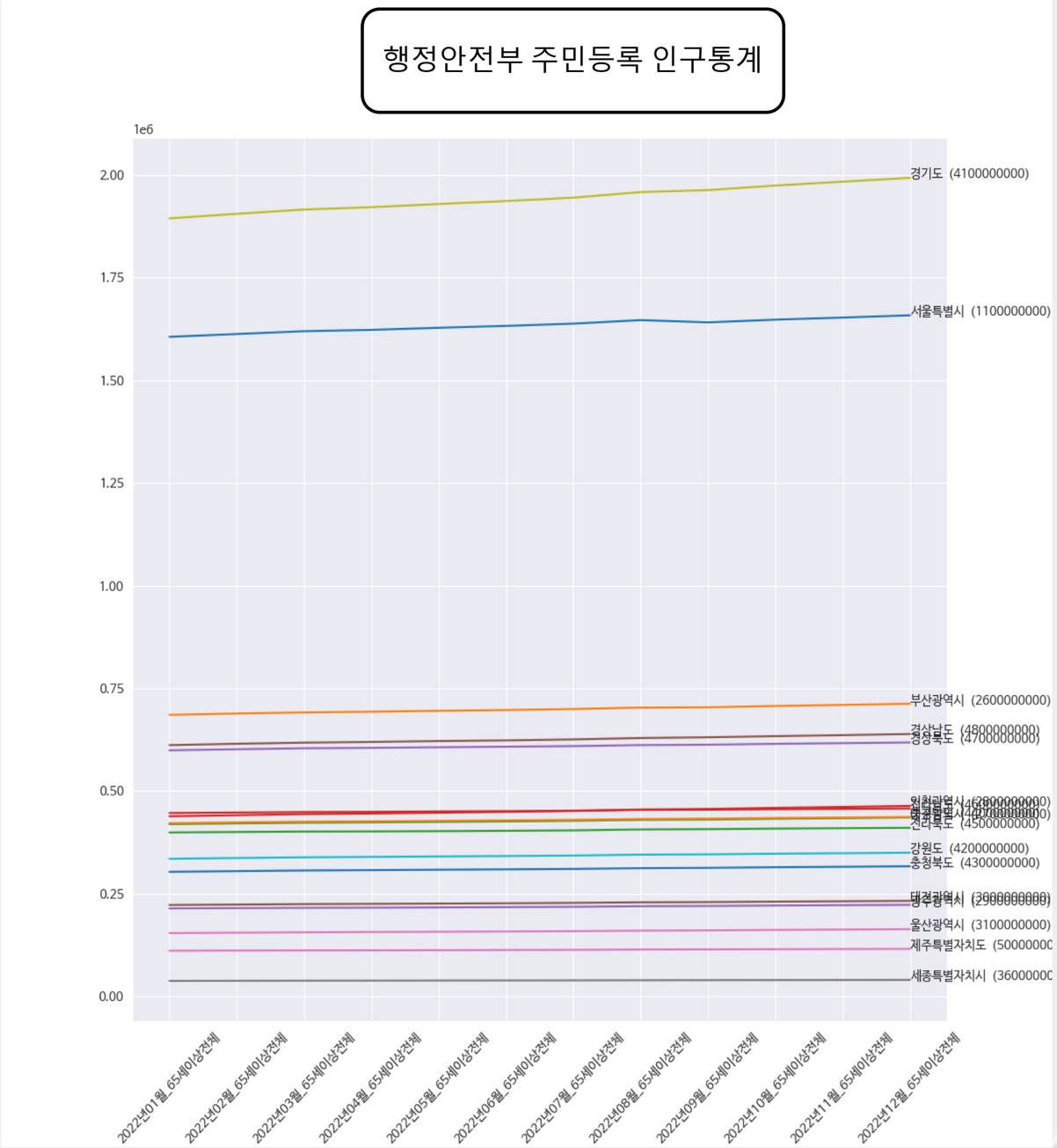


< 지역 라벨링 >

01 : 서울 11 : 경기 21 : 인천 31 : 충남 34 : 대전
41 : 대구 46 : 부산 50 : 경남 54 : 전북 57 : 전남
61 : 광주

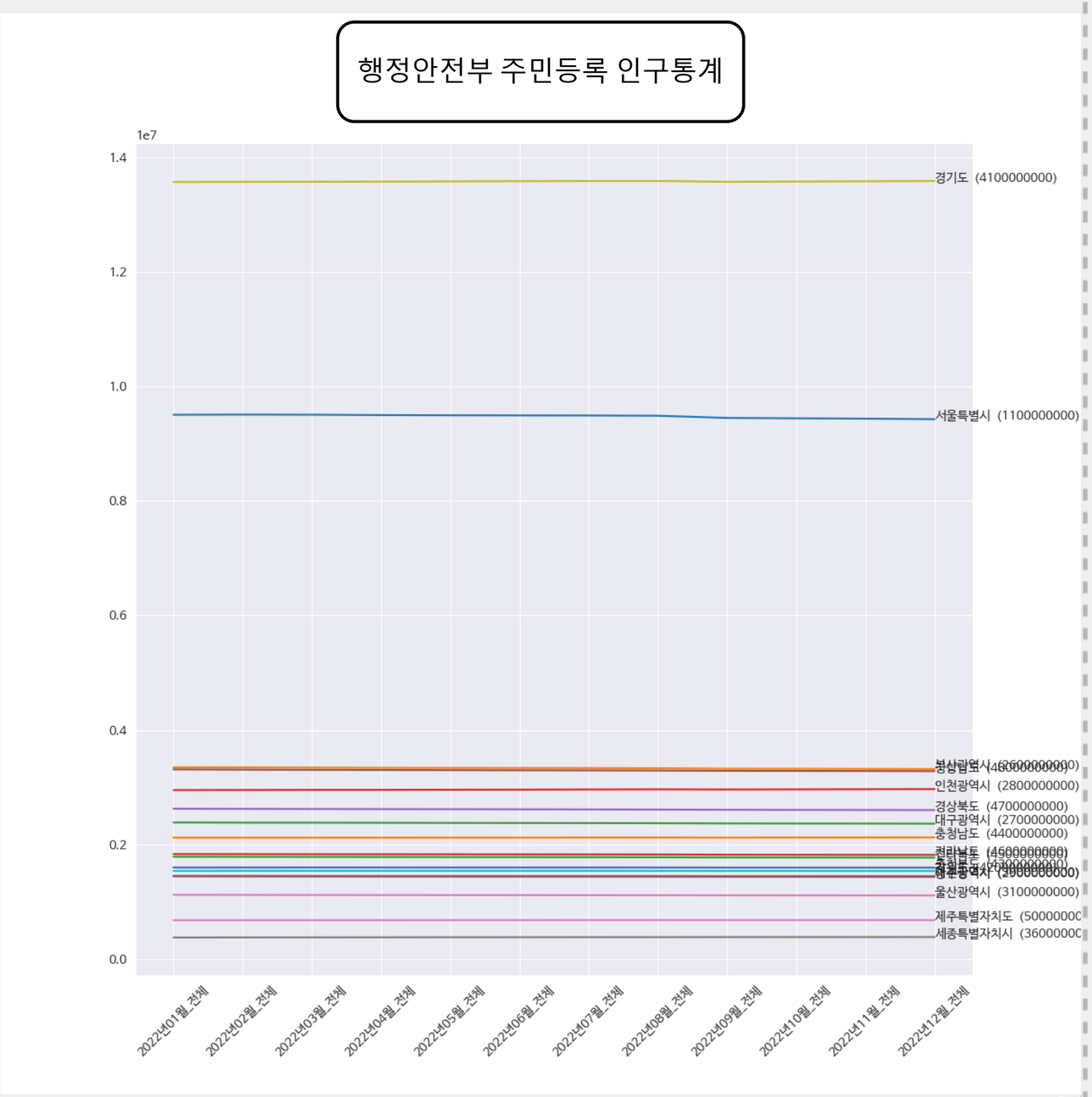
02 데이터 분석

행정안전부 주민등록 인구통계 데이터에서 보면 , 2022년 데이터에서 65세 이상의 인구 통계 데이터를 지역별로 보면 , 점점 올라가는 추세를 보이고 있으며 , 지역별로 보이면 경기도 , 서울특별시가 전체적인 인구가 가장 많습니다.



02 데이터 분석

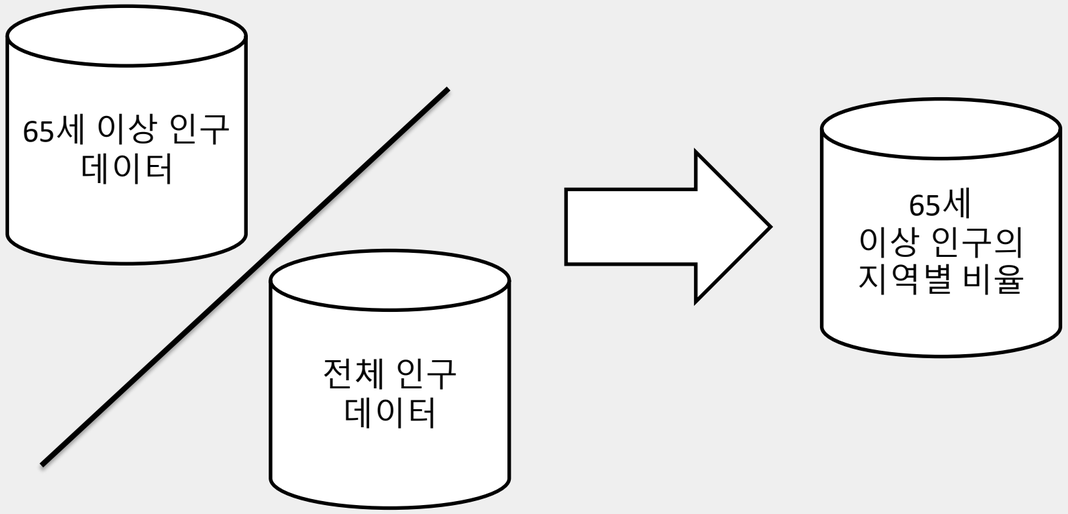
행정안전부 주민등록 인구통계 데이터에서 보면,
2022년 데이터에서 전체적인 데이터를 살펴보면, 65세 이상
인구에서처럼 확실하게 올라가는 추세는 보이지 않으며, 기울기가
거의 없는 것처럼 보입니다. 그리고 지역별로는 여전히 경기도,
서울특별시에서 많은 인구를 확인 할 수 있습니다.

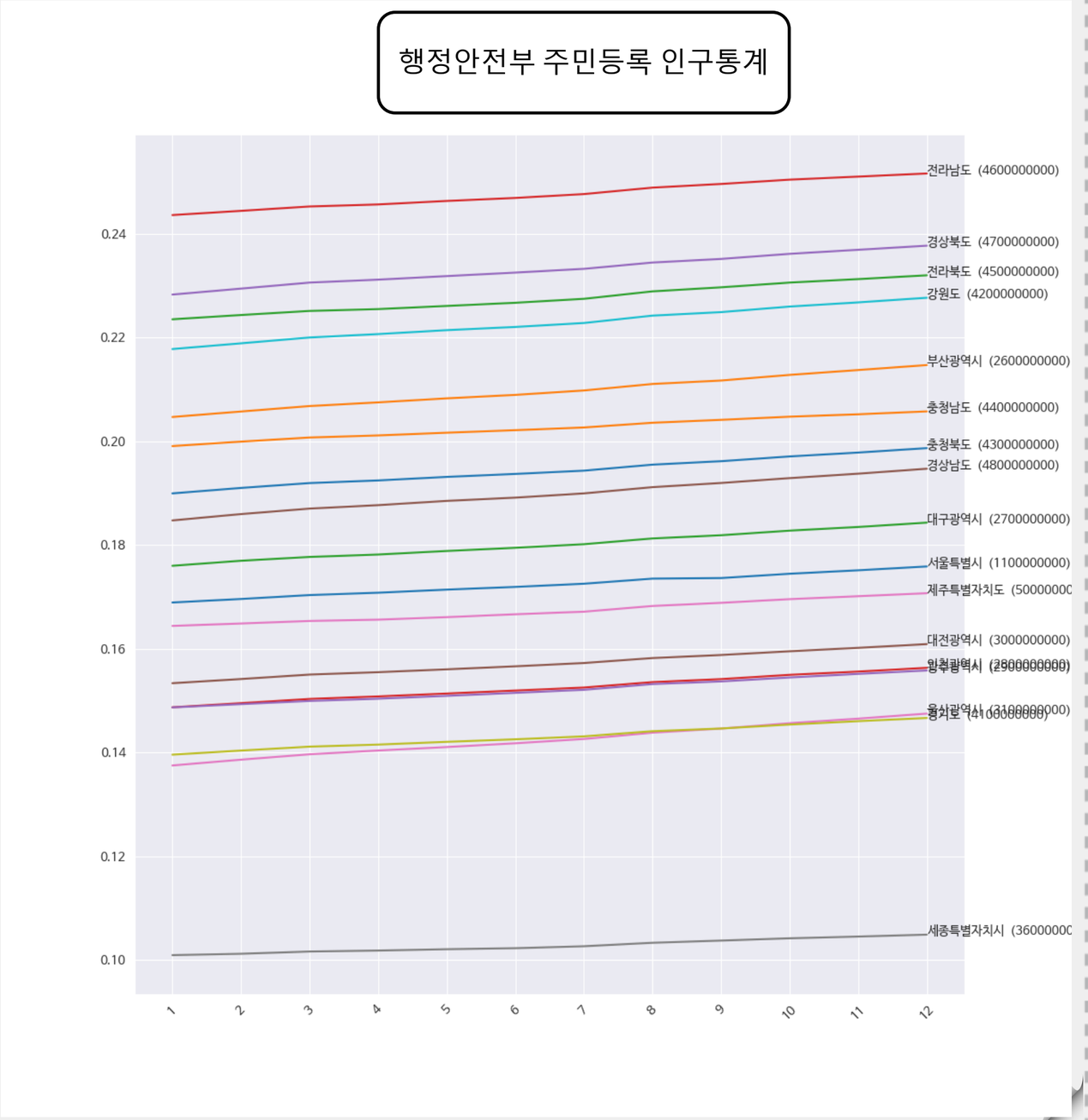


02 데이터 분석

2022년 1월부터 12월까지의 전체 인구 데이터를 2022년 1월부터 12월까지의 65세 이상 인구 데이터로 나눌 경우 , 비율을 파악할 수 있기 때문에 , 새로운 비율 데이터를 제작하고 시각화를 할 경우 ,

위처럼 65세 이상 인구는 확실하게 점점 늘어나는 추세로 볼 수 있으며 , 인구 자체는 서울특별시 , 경기도에서 많은 인구가 관측되었습니다만 , 비율로는 전라남도가 가장 많고 그 다음으로 경산북도 , 전라북도 , 강원도 순으로 많다는 것을 파악 할 수 있었습니다.



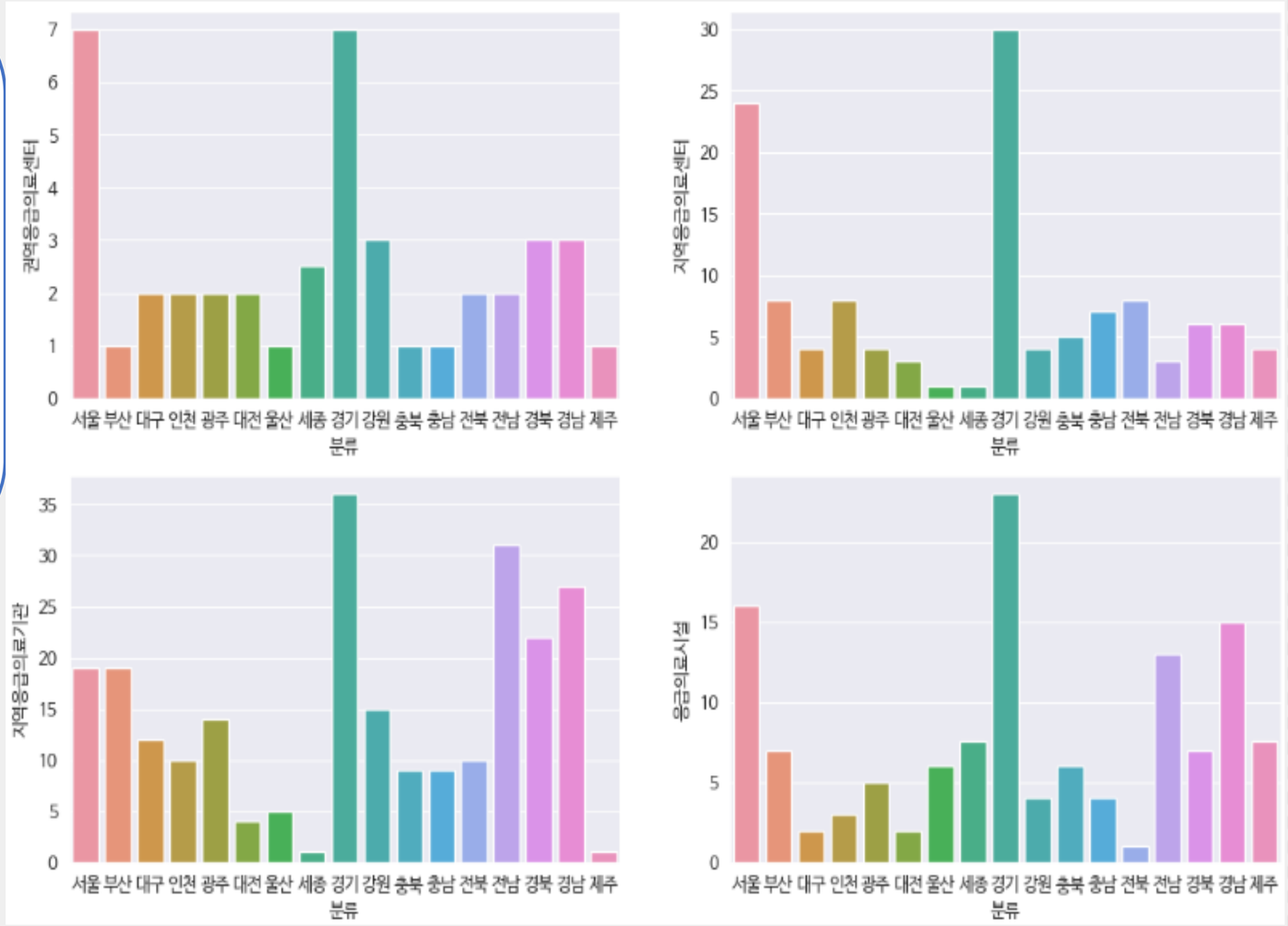


02 데이터 분석

응급의료통계포털 MEDIS

시도별 응급의료기관 및 응급의료시설 데이터를 통해서 , 통찰을 얻기 위해서 데이터에서 시각화를 해보았는데 , 여기에서 지역별로 각 의료기관 빈도를 볼 경우 ,

전체적으로 서울 , 경기에서 많은 수의 응급시설을 가지고 있음을 알 수 있고 , 반대로 지역응급의료기관의 경우 전남 , 경북 , 경남 부근에서 시설이 많은 것을 볼 수 있다.



권역응급 의료센터 → 지역응급 의료센터 → 지역응급 의료기관

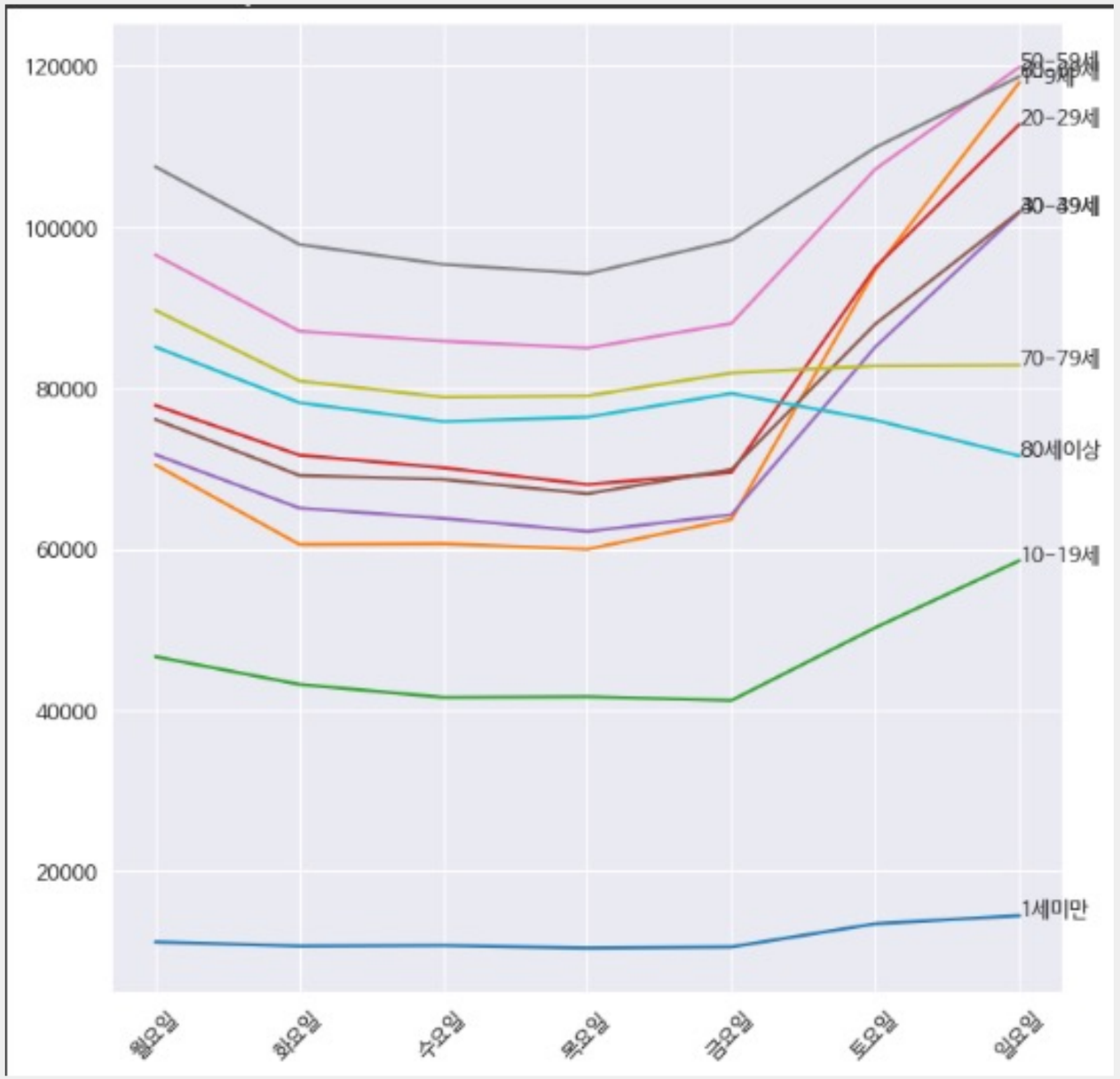
응급의료기관 체제 (규모)

02 데이터 분석

응급의료통계포털 MEDIS

내원요일별 응급실 이용데이터에서 연령이 기록되어 있는 데이터를 분석해보면 , 20 ~ 69세 까지 넓게 응급실을 이용하고 70세 이상이 되면 오히려 낮아지는 것을 볼 수 있습니다.

그리고 요일별로 확인이 되는 것은 , 월 , 토 , 일 이렇게 점점 상승을 하며 , 화 ~ 금의 경우 이용하는 사람들이 줄어드는 경향이 전체적으로 존재하는 것을 볼 수 있습니다.

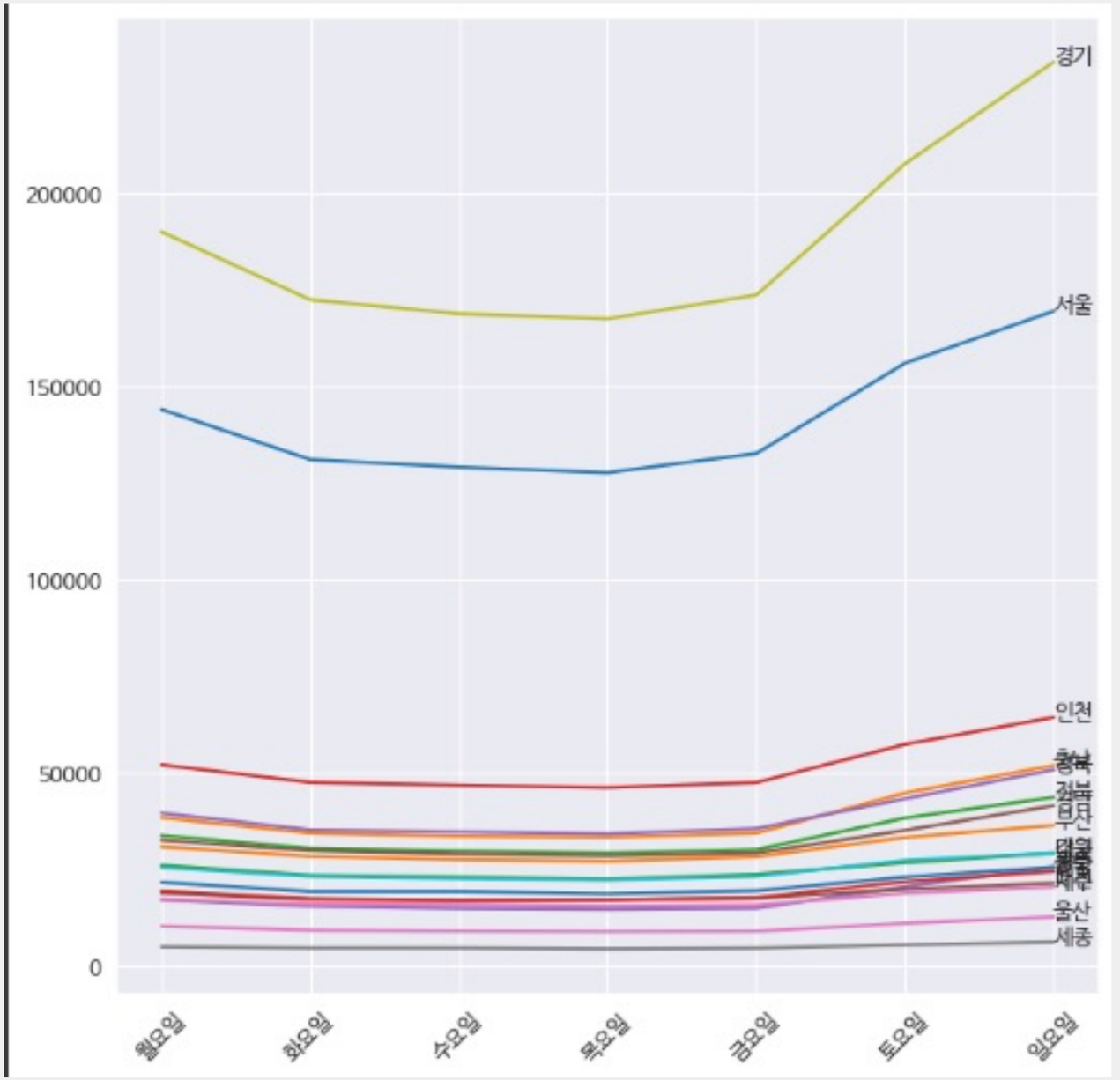


02 데이터 분석

응급의료통계포털 MEDIS

내원요일별 응급실 이용데이터에서 지역별로 데이터가 기록이 된 데이터를 분석해보면 , 여기서도 똑같이 월 , 토 , 일에서 이용객이 늘어나는 경향이 있다는 것을 볼 수 있습니다.

그리고 인구데이터에서 인구가 많아서 그만큼 환자 수도 많은 것으로 확인이 된 경기도 , 서울특별시에서 많은 내원 환자수를 볼 수 있습니다.

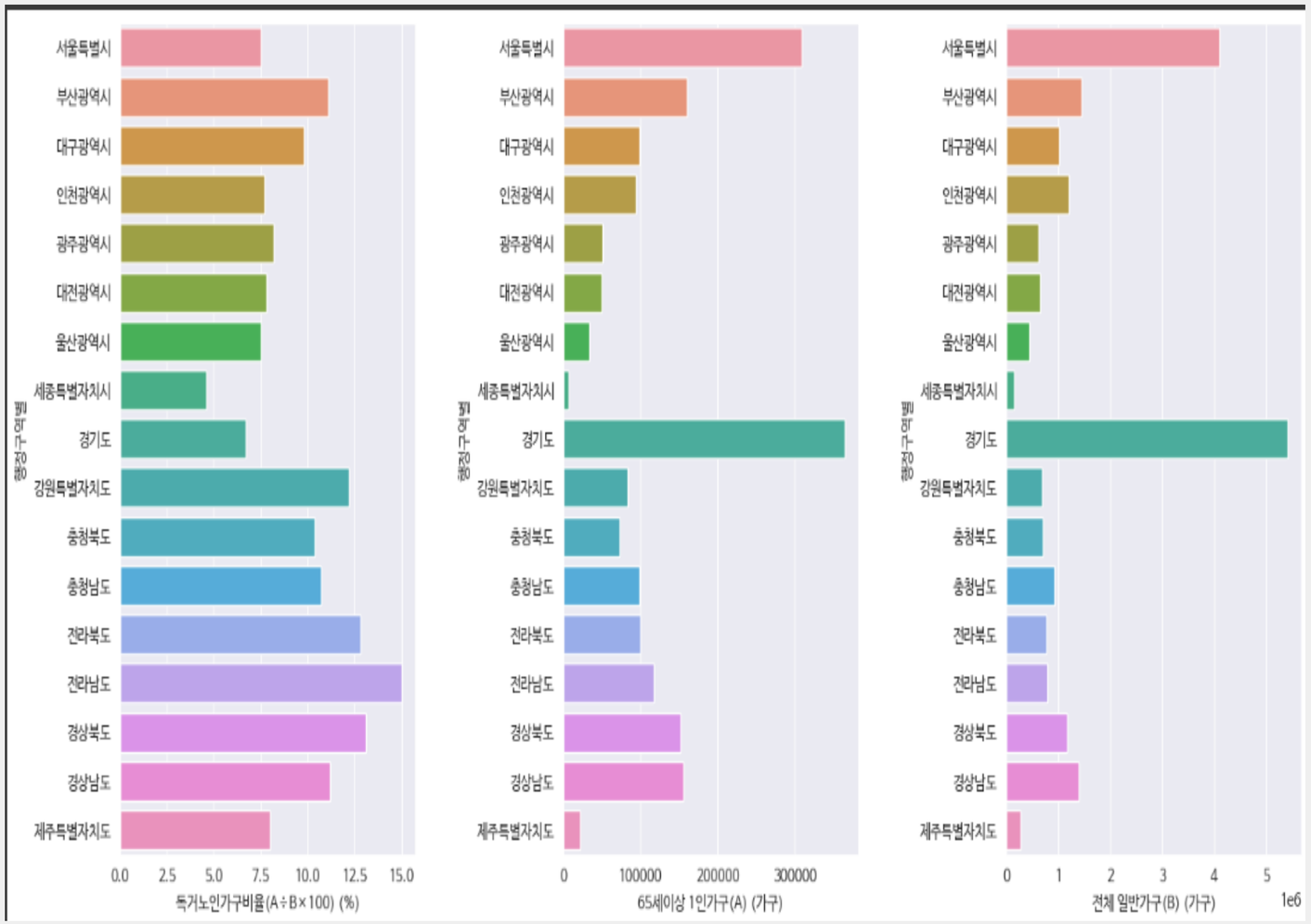


02 데이터 분석

KOSIS 국가통계포털

2022년에 기록된 , 전체 일반가구와 65세 이상 1인가구데이터 그리고 전체 일반가구와 65세 이상 1인가구데이터의 비율을 토대로 계산을 해낸 독거노인가구비율 데이터를 분석해보면 ,

가구수 자체는 서울특별시 , 경기도에서 많은 수를 보이면 그만큼 65세 이상 1인가구수도 많이 보이지만 전체 일반가구 중에서 65세 이상 1인가구의 비율을 살펴보면 , 전라도 , 경상도 부근에서 가장 많은 가구 수가 관측이 됩니다.



02 데이터 분석

파생변수 만들기

정리

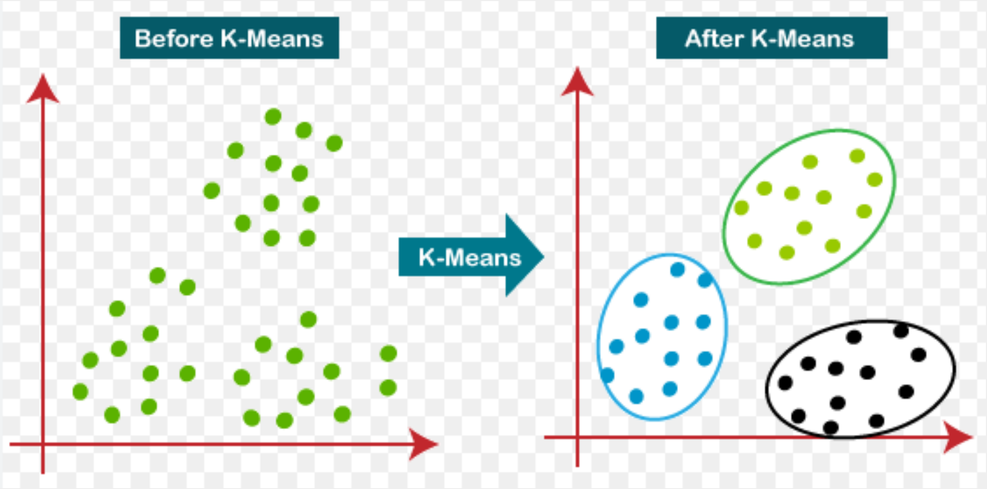
1. 지역에 따라서 응급시설의 비율 차이가 상당히 많이 있다. 그렇기 때문에 해당 지역의 인구에 따라서 어느정도 수의 시설이 마련이 되어 있는지 비율에 따른 비교 가능
2. 응급시설의 전체인구에 따른 비율과 함께 65세 이상 인구를 사용한 비율로도 비교가 가능하다
3. 인구 수 데이터에서 지역에 따라서 전체 인구중 65세 인구 수가 어느 정도의 비율이 있는지를 계산해서 지역 간 65세 인구 비율 비교가 가능
4. 지역에 따라서 독거노인 수에 비해 어느 정도의 시설이 준비가 되어있는지에 대해서 비교 가능
5. 지역에 따라서 65세 인구 중 65세 이상 1인가구인 독거노인 수가 어느정도의 비율로 존재하는지에 따라서 지역간 비교 가능

시도별 응급의료기관 및 응급의료시설 데이터 / 전체 인구 데이터	**인구 대비 응급 시설수를 지역별로 비교**
시도별 응급의료기관 및 응급의료시설 데이터 / 65세 이상 인구 데이터	**65세 이상 인구 대비 응급 시설수를 지역별로 비교**
65세 이상 인구 데이터 / 전체 인구 데이터	**지역별로 어느 정도의 비율로 65세 이상이 차지하는지 비교**
시도별 응급의료기관 및 응급의료시설 데이터 / 독거노인 인구 데이터	**독거노인 인구 데이터 대비 응급 시설수를 지역별로 비교**
독거노인 인구 데이터 / 65세 이상 인구 데이터	**지역별로 65세 이상 중 1인 가구가 어느 정도로 비율을 차지하는지 비교**

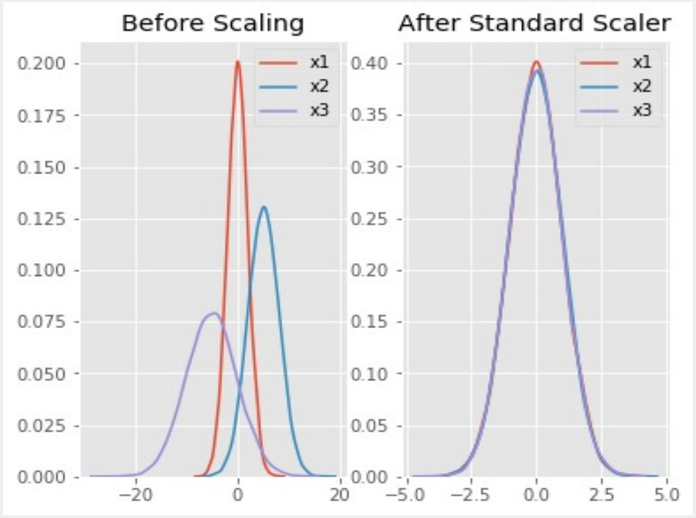
02 데이터 분석

K means를 통해서 할 경우 거리를 통해서 군집을 생성하기 때문에 데이터 간의 거리가 중요

전처리 및 군집화



[사진 출처]
https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/



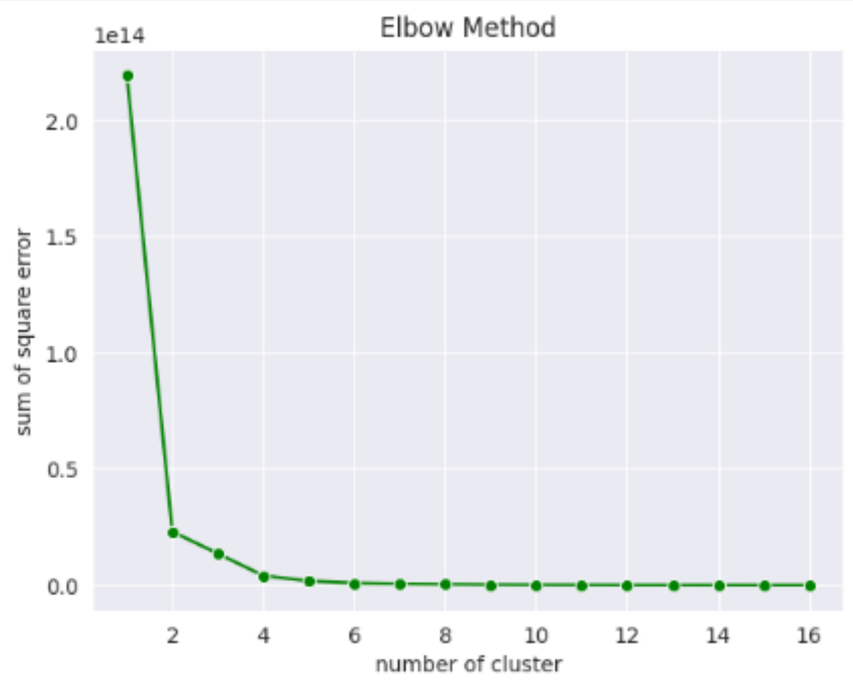
[사진 출처]
https://datascience.stackexchange.com/questions/43972/when-should-i-use-standardscaler-and-when-minmaxscaler

데이터의 단위가 다르기때문에 k means의 성능이 떨어질 위험이 있어서 , 각 데이터들의 분포를 평균 0 표준편차 1로 맞추는 표준화를 실시합니다.

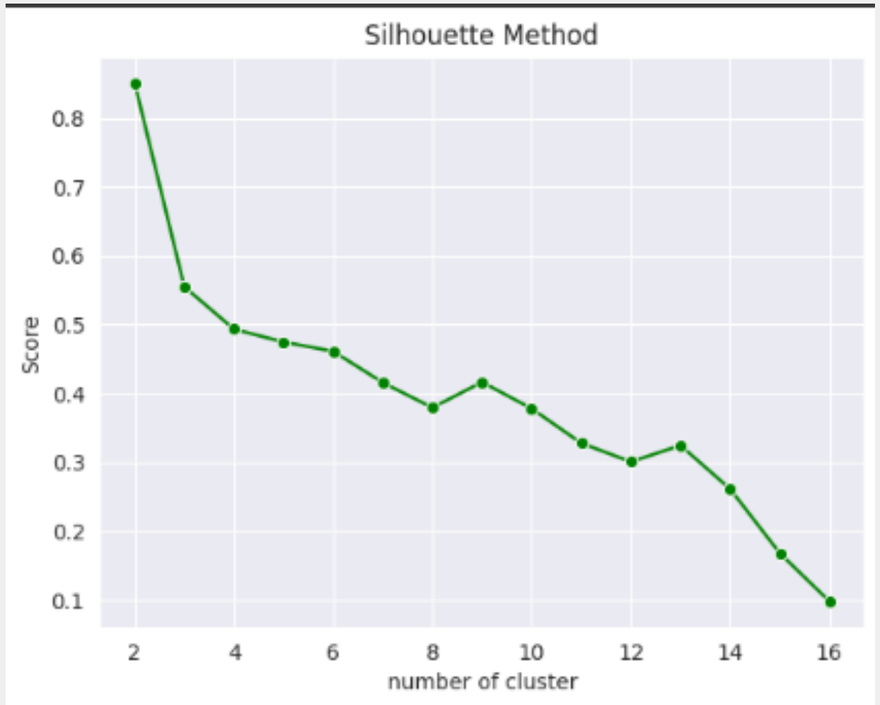
02 데이터 분석

적절한 K 찾기

전처리 및 군집화



엘보우 방법에서 군집의 영향력을 보면
2 ~ 3정도가 적당한 것 같습니다

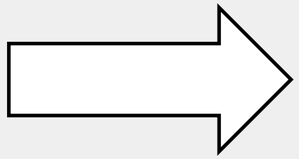


실루엣 분석에서 각 군집에서
데이터들의 떨어진 거리를 보면 2 ~ 3
정도에서 군집끼리 데이터가 잘 모여
있습니다

02 데이터 분석

전처리 및 군집화

K = 3 으로 Kmean 클러스터링을 이용하여 , 군집 분석을 실시한 결과



군집 0	대구 , 인천 , 광주 , 대전 , 울산 , 세종 , 제주
군집 1	서울 , 경기
군집 2	부산 , 강원 , 충북 , 충남 , 전북 , 전남 , 경북 , 경남

02 데이터 분석

군집분석 결론

군집 0

군집 0의 경우 군집 2의 지역들보다는 더 독거 노인 비율이 적으며 , 시설 수도 독거 노인 비율에 비해서 괜찮은 쪽에 속합니다.

적당

군집 1

군집 1의 경우에는 인구 수도 다른 지역에 비해 많으며 , 그만큼 독거노인의 비율도 많지만 그만큼 시설 수도 많은 편에 속합니다.

양호

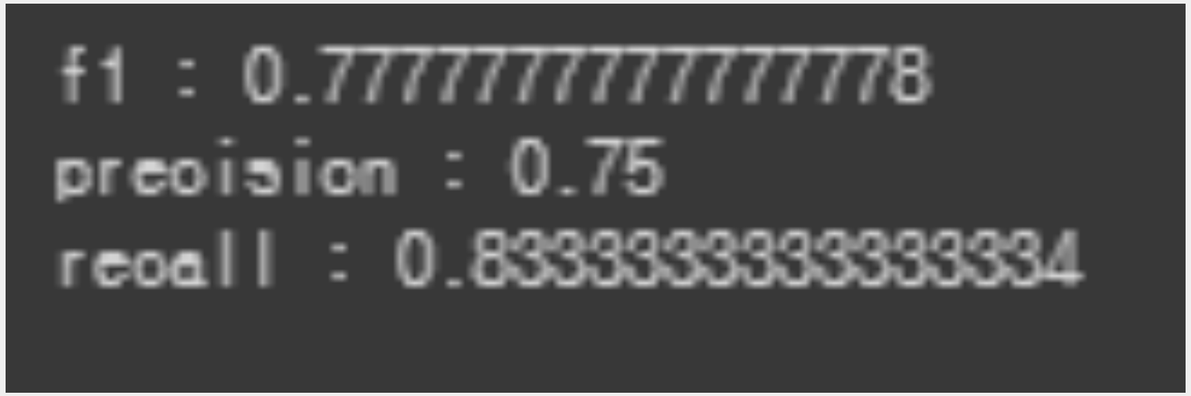
군집 2

독거노인의 비율이 군집0 보다는 많으며 군집 1보다는 적지만 시설 수가 독거노인의 비율에 비해서 조금 부족한 경향이 있는 지역들입니다.

부족

03 개선 방안

RandomForestClassifier



모델의 성능을 보면 ,

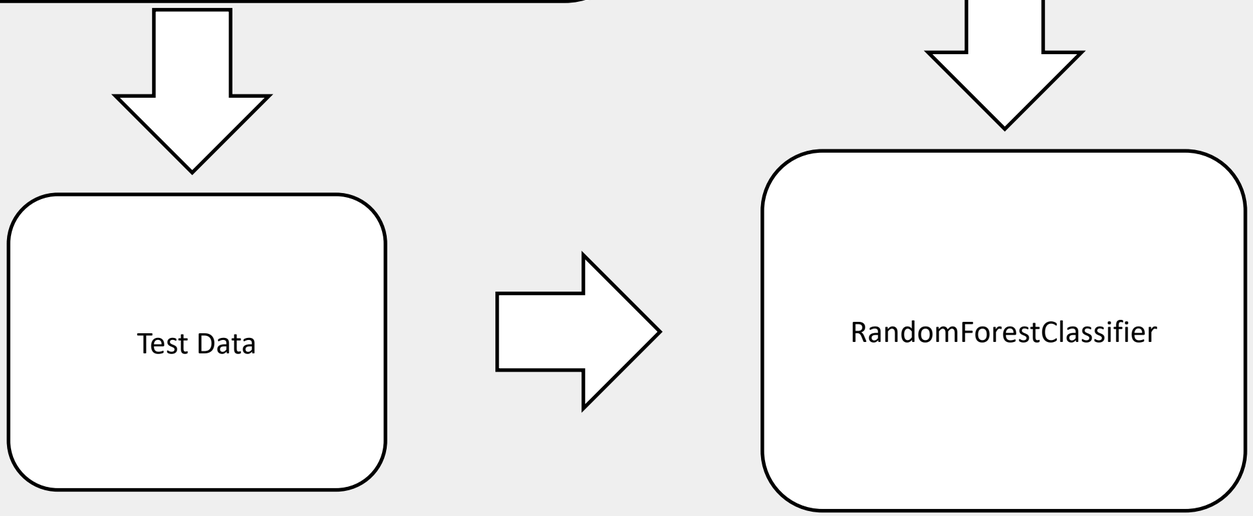
1. F1 : 튜닝하지 않은 모델로 괜찮은 값이 나왔습니다.

2. Precison , recall : F1은 precision과 recall의 조화평균이기 때문에 정말로 성능이 괜찮은지 볼 경우 둘다 확인해 볼 필요가 있는데 엄청난 차이가 있지 않으므로 괜찮은 성능을 보인다고 생각할 수 있습니다.

03 개선 방안

군집화를 토대로 3가지의 클래스를 가지는 종속변수 데이터를 만들었습니다.

여기에서 분류 모델을 만들기 위해서 데이터를 학습데이터 , 검증 데이터로 나눕니다.



03 개선 방안

모델 제안

현재 저희가 만든 모델은 프로토타입으로 여기에 더 데이터를 추가하거나 , 패러미터를 튜닝 하는 등으로 성능을 끌어올릴 수 있기 때문에

저는 이렇게 선보인 여러 데이터들을 토대로 매년 바뀌는 인구 수와 독거노인 인구 수에 근거하여 의료적으로 시설이 개선이 필요한 지역을 찾아내는 분류모델을 통해 미리 예측을 하면서 필요한 대책을 미리 하는 것으로 , 새로운 감염병이 나타난 시점에서도 유연하게 대처가 가능할 것 같다고 생각합니다.

구급차와 같은 의료용 운송 수단 개선

전국에 의료 시설이 많기 때문에 , 모든 시설에 대해서 운송 수단을 강화하는 것은 불가능 하기에 이러한 분류 모델을 통해서 필요한 지역을 찾아내서 미리 대처를 하는 것으로 비용과 시간 면에서 적절한 대처가 가능합니다.

의료 시설의 병상 수 개선

운송 수단 개선과 같이 전국의 의료 시설에 대해서 모두 개선하려면 비용적으로 불가능 하기 때문에 , 모델을 통해서 필요한 지역을 찾아 , 병상 수를 늘리는 것으로 의료붕괴 사태를 막을 수 있습니다.

04 기대 효과

시간과 비용 면에서 적절한 판단이 가능

모든 지역에 대해서 개선을 하는 것은 현실적으로 불가능 하기 때문에 , 개선이 필요한 지역에 집중하여 개선에 필요한 조치를 취할 수 있고 , 그것으로 시간과 비용을 절약할 수 있는 적절한 판단이 가능하게 됩니다.

시간 분류와 빠른 업데이트

이러한 모델을 토대로 개선할 경우 , 빠르게 변하는 인구 데이터를 실시간으로 적용시켜서 빠르게 모니터링이 가능하며 , 또 모델의 패러미터를 갱신하면서 상황에 맞게 적절한 업데이트를 하여 운용을 간편하게 할 수 있습니다.