호흡기 전담클리닉 최적 입지 선정

회귀모델을 통한 군집분석과 입지 선정 공간분석을 중심으로

E215

구아영 김민채 김창현 김현솔 김형순 문희소 박은우



목차

O1-WHY

- 분석 배경

- 문제 인식

02 WHAT

- 목적

03 How

- 프로젝트 개요

- 구별 개소수 선정

CONCLUSION

- 활용방안

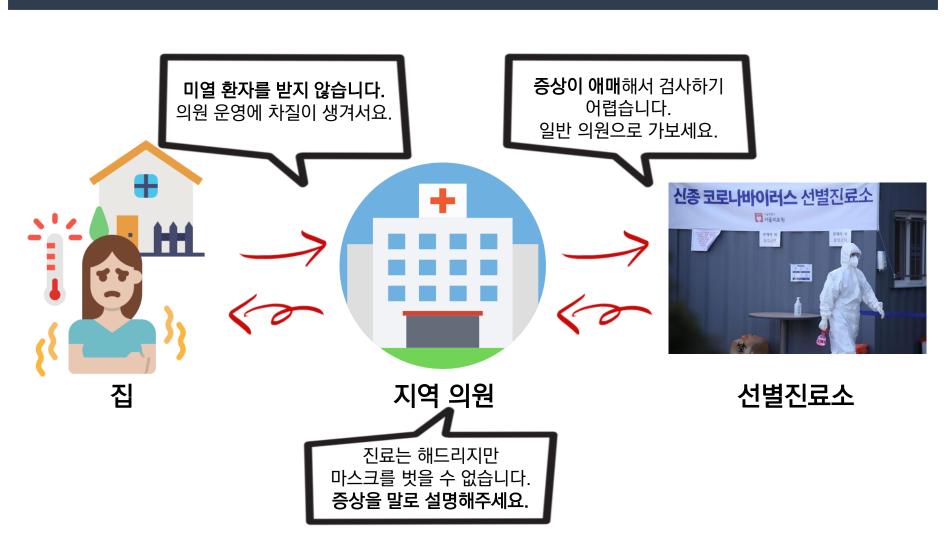
- 기대 효과 및 개선방향

- 위치 선정 요인 데이터 분석

- 분석 결과

WHY 1. 분석 배경

현재의 코로나19 진료 및 예방 체계는 증상만으로 여타 호흡기감염 구분이 어려워 일선 의료기관에서 호흡기·발열 환자에 대한 적극적인 진료의 제한이 존재



WHY 1. 분석 배경



ॱ❤️ 현 코로나19 대응 체계의 문제점



1. 의원&병원

- 코로나 확진자 방문으로 인한 의료기관 운영 중단 우려
- 의료진&환자 2차감염



2. 선별진료소

- 증상만으로 코로나19와 호흡기감염 구분이 어려워 진료의 제한 존재



3. 국민안심병원

- 입원치료를 담당하는 종합병원급 병원에 확진자 방문으로 인한 운영 차질 우려

현재의 제한적인 코로나19 대응 체계는 호흡기&발열 환자의 진료 공백과 의료기관 집단 감염에 취약함

이에 정부는 코로나19 장기화 및 동절기 호흡기감염(감기, 독감 등) 동시 유행을 대비하여 호흡기 전담 클리닉을 선정함으로써 호흡기·발열 환자의 안전한 의료이용체계 확립 계획 수립



·**쌀**· 호흡기 전담 클리닉이란?

호흡기 증상이 나타나는 환자에게 독감 등 여부를 진단·처방하고. 코로나 환자로 의심될 경우 직접 진단 검사를 하거나 선별진료소에서 검사하도록 하는 진료시스템으로 호흡기·발열 화자의 일차의료를 담당

호흡기전담클리닉 운영 방안



- 21년까지 **전국 1.000개소의** 호흡기전담클리닉 설치 발표
- 서울 내에서 보건소 25개를 제외한 **약 120개소** 의료기관형 호흡기전담클리닉 설치 예정 * 올해 서울 75개소 설치 계획 확정

구분	개방형 클리닉	의료기관형 클리닉
운영	보건소, 공공시설을 활용	일선 의·병원급 의료기관이 호흡기를 전담

WHY 2. 문제 인식

의료기관형 ▋클리닉▮

시군구별 클리닉 설치 개소수 확정

보건복지부가 공개한 호흡기전담클리닉 선정 절차는

'개소수 선정'과 '입지 평가 방식'의 한계가 존재하여 효과적 운영의 제한이 예상됨



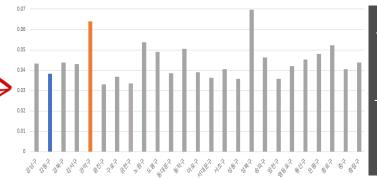
지정 프로세스

개방형 클리닉

* 호흡기전담클리닉 설치 수 기준



Problem1: 단순 인구수 계산에 근거한 선정방식



인구수 대비 코로나 확진자 확률분포

Problem2: 다양한 요건을 고려하지 않은 선정 평가 방식

- 현 '10만명당 1개소' 기준은 코로나 예측 환자 및 고위험군 등 관련 위험요인을 고려하지 않음
- 자치구별 확진자 밀도의 차이가 존재하여 현 클리닉 설치 개소수 기준의 개선 필요

지자체-지역의료계

10만명당 1개소

협의

클리닉 공간 확보 시설. 설비 구비

의료 인력 모집

지정

현장 방문 및 평가

지정

민간 의원&병원

사업 참여 신청

시설. 설비 구비

강남구 보건소 4.3km 강남구 강남구 개포1동



- 정량적 평가 방식의 부재 (거리, 주거인구, 밀도 등 다양한 요건을 고려한 평가 필요)
- 설치가 확정된 구 보건소의 경우 보건소와 주거지 간 접근성 문제 존재

WHAT \Address

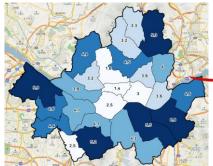
프로젝트 목적: 서울시의 자치구별 호흡기 전담 클리닉 설치 개소 수 선정 및 최적 입지 추천

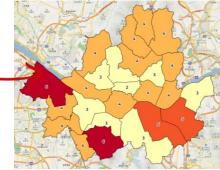
Problem 1:단순 인구수 계산에 근거한 개소 수 선정방식

- Solution: 시군구별 코로나 확진자 수에 영향을 **미치는 요인**들을 고려하여 개소 수 선정

AS-IS

TO-BE





단순 구별 인구수만 고려 진할수록 많은 수요 예상

(강서구 6개, 용산구 1개)

- 유의미한 변수를 고려하여 개소 수 책정



- (예시1) 호흡기 질환 평가 등급이

- (예시2) 주거인구가 높고, 유동인구가 낮은 행정동에 우선 배치해야 함

"우수"인 병원

Problem 2: 다양한 요건을 고려하지 않은 선정 평가 방식

- Solution: 고려요건, 필수요건을 고려하여 호흡기 전담 클리닉 최적 입지를 도출



- (예시1) 특정 지역에 설치가 집중되지 않아야 함
- (예시2) 보건소가 설치된 시군구에는 1개소 이상 설치하도록 함



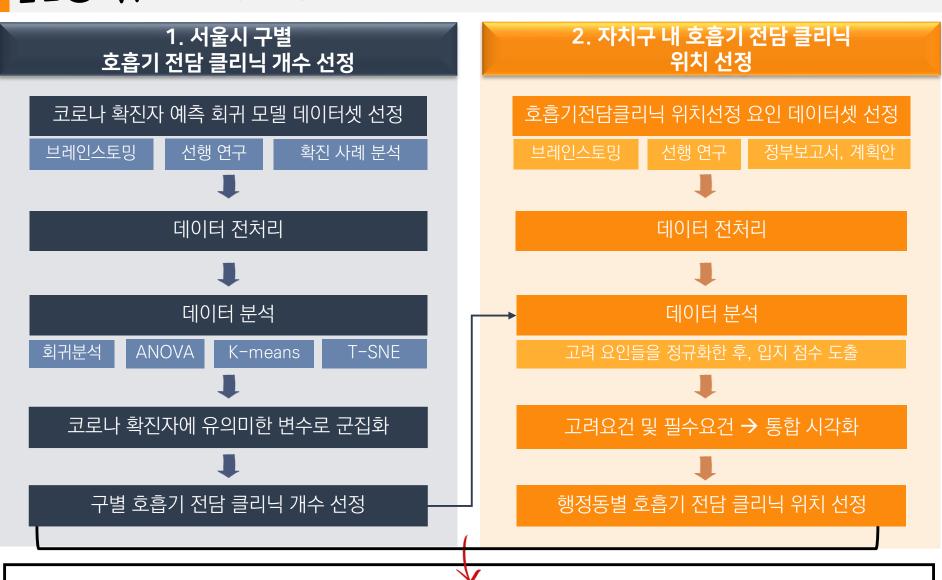
코로나19 취약 지역 도출



우선 방문 평가 지역 선정

"신속하고 효과적인 전염성 호흡기질병 진료를 통한 선제적 확산 대응"

HOW 프로젝트 개요



3. 입지 선정

코로나 확진에 영향을 미치는 변수 선정

1. 인구

N인 가구수

전염병의 감염은 주로 가족간 밀접한 접촉에 의한 군집감염 사례들이 보고되어 있다.

가구수



시군구 별 N인 가구수의 차이가 존재하며. 그에 따른 코로나 발생에 대한 영향이 존재할 것이다.



서울시 가구 수별 비중

2. 확산위험시설

물류센터, 종교, 노래방, 다단계판매업체, 주점업, 목욕탕

병원

° 방역당국 발표 12종 고위험시설

유흥주점, 노래연습장, 실내공연장, 헌팅포차, PC방, 실내 집단 운동 시설, 단란주점, 감성주점, <mark>방문판매</mark> 등 직접판매 홍보관, 대형학원, 콜라텍, 뷔페

○ 종교시설, 물류센터

교회, 절등 종교시설에서 집담감염 사례 다수 발생 경기도 중심 <mark>물류센터</mark>에서 집담감염 사례 다수 발생

○병원

전염병 감염은 주로 <mark>의료기관내 감염</mark>의 양상을 보인다.

3. 집단시설

집단시설: 남남인 사람이 6인 이상의 집단으로 거주하는 시설

요양시설

아동(고아원)

대구, 경북 지역 중심으로 장기요양시설 거주자 사이에서 대규모 환자가 발생 요양시설 환자들은 코로나19 취약계층이며, 대부분 연령이 높고 만성질환을 가지고 있음.

출처: Surveillance of COVID-19 at long-term care facilities

아동의 경우 무증상 사례가 많다. 이는 아동의 경우 감염 상태가 감지되지 않은 채 방치될 수 있음을 의미한다.

출처: 경기도청, 경기도 일일상황분석 보고서 180호, p.10

4. 질병

당뇨병. 고혈압

진해거담제

고혈압, 당뇨 질환자의 경우, 코로나19에 감염되기 쉬우며 사망률도 높다.

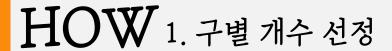
- 출처: Dong Wook Kim, Kyeon Hyang Byeon, Jaiyong Kim, an Observational Study Based on the Korean National
- : 심한 기침이나 가래, 급성 상기도 감염 등의 호흡기질환의 경우에 주로 처방받는 내복약의

진해거담제

→ 진해거담제 판매량이 높을수록 코로나19 확진 발생에 영향을 줄 수 있을 것이다.

출처: 심평원 요양급여 적용기준 및 세부사항

대 모



다양한 데이터셋의 수집 및 통합(join)



데이터 수집

시군구 당뇨병 환자 수

시군구 종교단체 수

다단계판매업체 인허가 정보



수도권 데이터만 추출

서울특별시 - 구별 인천광역시 - 구별

경기도 – 시군구별

→ 총 77개(raw)

데이터 전처리





▼ 파생변수(평균, 합계, count)

회귀분석 최종 데이터셋: 수도권(서울특별시, 인천광역시, 경기도)별 데이터

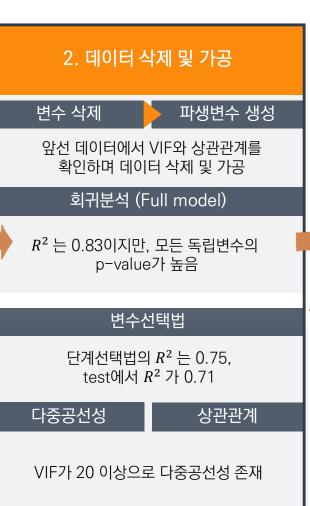
독립변	수명	
총가구_가구원_계_명	방문판매_종사자_비율	종교단체_수
일반가구_가구원_계_명	 당뇨병소계	대규모점포수
집단가구_비혈연_6인이상_가구_가구원_계_명	 고혈압소계	목욕장업수
집단가구_집단시설가구_가구원_계_명	유입인구	유흥주점영업수
가구원_모두_65세이상_가구수	r l 유출인구 l	노래연습장업수
N1인_가구수	 PC방_총사업체수	병원수
N2인_가구수	 전화권유업체수	병원_의료인수
N3인_가구수	 다단계판매업체수	병원_입원실수
N4인_이상가구	 단란주점영업수	병원_병상수
	 음식점_수 	의원수

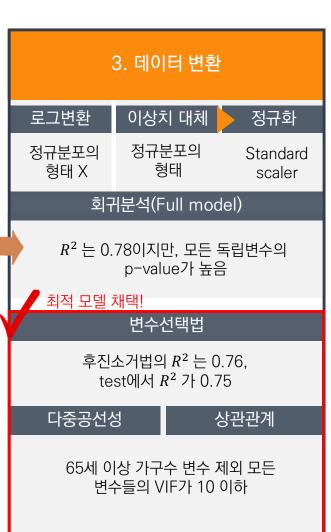
	SIG_CD	도 시	자 치 구	총가구_ 가구원_ 계_명	일반가구 _가구원_ 계_명	집단가구_비혈연 _6인이상_가구_ 가구원_계_명	집단가구_집 단시설가구_ 가구원_계_명	가구원_모 두_65세이 상_가구수	8 명 소 계	고혈 압소 계		규 모 점 포 수	독욕 장업수	유흥 주점 영업 수	노래 연습 장업 수	병 원 수	병원 _의 료인 수	병원 _입 원실 수	병원 _병 상수	의 원 수	확 진 자 수
0	11110	서 울 특 별 시	종로구	154969	136649	640	4854	8079	374	797		2	30	434	102	6	4769	835	2890	56	98
1	11140	서울특별시	중 구	130957	115665	690	2328	6990	317	712		15	45	654	77	5	1325	258	927	176	66
2	11170	서 울 특 별 시	용 산 구	227181	204196	428	2387	11019	555	1112	Γ.	77	(ro	DWC	·) X	2	<u> </u>	CO	lur	nn)



회귀분석 Process









최종 선택 회귀모델 review

회귀모델 해석

OLS Regression Results Dep. Variable: 확진자수 R-squared: 0.758 Adj. R-squared: 0.741 Model: OLS F-statistic: Method: Least Squares 43.91 Date: Mon, 09 Nov 2020 Prob (F-statistic): 1.21e-16 Time: 01:25:08 Log-Likelihood: -327.26 664.5 No. Observations: 61 AIC: Df Residuals: BIC: 675.1

 Covariance Type: nonrobust

 coef
 std er
 t
 P>It
 [0.025]
 0.975]

 N1인_가구수
 22.0136
 13.211
 1.666
 0.100
 4.452
 48.479

 가구원_모두_65세이상_가구수
 76.7934
 14.197
 5.409
 0.000
 8.353
 105.234

 병원_병상수
 -28.0608
 11.667
 -2.405
 0.019
 51.433
 -4.688

 병원_의료인수
 20.6774
 10.415
 1.985
 0.052
 0.187
 41.542

 const
 157.9127
 6.925
 22.805
 0.000
 44.041
 171.784

- Test에서의 R²

Df Model:

Backward_best_model.summary()

print(r2_score(test_y3, pred_y_backward))

0.7459918415378652

- 1. p-value가 유의수준 0.1내에서 유의
- 2. train에서 R^2 가 0.758, test에서 R^2 가 0.746로 높음

사관가수 1.00 0.86 0.64 0.50 가구원_모두_65세이상_가구수 0.86 1.00 0.66 0.43 병원_병상수 0.64 0.66 1.00 0.73 병원_의료인수 0.50 0.43 0.73 1.00

다중공선성(VIF)

features	VIF Factor
가구원_모두_65세이상_가구수	10.203407
병원_의료인수	9.454477
병원_병상수	2.721537
N1인_가구수	1.000000

"65세 이상 가구수" 변수가 **변수간 상관관계가** 높음

→ <u>ANOVA</u>를 통해 해당변수를 제거한 모델과 비교

ANOVA

> anova(lm1, lm2)

Analysis of Variance Table

Model 1: 확진자수 ~ 가구원 모두 65세이상_가구수 + N1인_가구수 + 병원_병상수 + 병원_의료인수

Model 2: 확진자수 ~ N1인_가구수 + 병원_병상수 + 병원_의료인수

Res.Dt RSS Dt Sum ot Sq 1 72 92219 2 73 147493 -1 -55274 43.1

of Sq F Pr(>F)
-55274 43.155 6.814e-09 ***

p-value가 낮으므로, "65세 이상 가구수" 변수가 회귀모형에 대한 **기여도가 높음**

→ 따라서 <u>해당 변수를 제거하지 않고</u> 기존 모델을 최종 모델로 선정

HOW 1. 구별 개수 선정

데이터전처리 변수선정

회귀분석

군집분석



최종 변수 해석

: 1인 가구 수, 가구원 모두 65세 이상 가구 수, 병원 병상 수, 병원 의료인 수

1인 가구 수



인사이트: 기혼자 가구 이 예방행동 1인 가구 예방행동

회귀분석: 회귀계수 22.0136

P-value 0.100*

(가구원 모두) **65세 이상 가구수**



인사이트: 65세 이상 감염 발생 多 가구수 多

76.7934 회귀분석: 회귀계수

> P-value 0.000***

병원 병상 수



낮은 감염 인사이트: 병상수 환자 통제 多 용이 발생

회귀분석: 회귀계수 -28.0608

P-value 0.019**

병원 의료인 수



인사이트: 의료인 감염 확진자 발생

회귀분석: 회귀계수 20.6774

P-value 0.052*



최종 회귀 모델의 변수

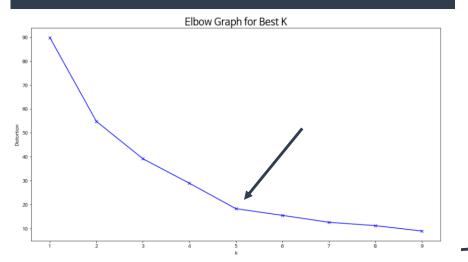
정규화

단위가 큰 변수에 의해 군집이 결정되는 현상 방지
→ Standard 정규화 수행

자치구 N1인_가구수 가구원_모두_65세이상_가구수 병원_병상수 병원_의료인수

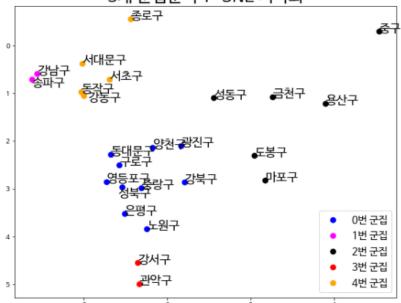
종로구	-0.627366	-0.913460	0.207114	2.040004
중구	-0.754035	-1.086548	-1.030646	-0.120702
용산구	-0.115559	-0.446171	-1.087395	-0.324601

K-means



5개 군집 분석 T-SNE 시각화

5개 군집분석 T-SNE 시각화



최종 회귀 모델의 변수들로 군집분석 실행

> 총 5개의 군집으로 호흡기전담클리닉 설치 개수 지정

변수선정



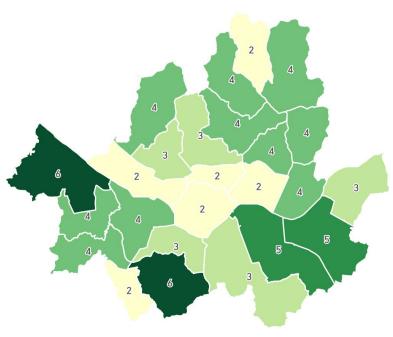
군집 해석



1인 가구수, 65세 이상 가구수가 가장 많은 군집



병원 의료인 수가 높은 군집



C군집(4개소) 평균



1인 가구수, 65세 이상 가구수가 가장 적은 군집



C군집보다 병상 수가 많은 군집

접근성

호흡기 전담 클리닉 선정 고려요건 정리, 관련 데이터 수집 및 전처리 진행



고려 요건 : 의료 인프라, 접근성, 주거 대 유동비, 경제적 취약계층

의료 인프라



행정동별 의료인 수



주차장 구획 수



도로폭 위치 정보

건강보험심사평가원 발표 호흡기질환 우수병원 중심 선정

호흡기질환 우수 병원

내과 의원, 종합병원 의료인 수는 지역내 **환자 수용력**과 관련

- 자차나 도보 이용 권장
- **주차 구획이 충분**한 지역에 선정

- 접근성을 위해 **대로변에 위치**한 병원 선정
- 도로폭이 12m 이상

주거 대 유동비

주거인구 / 유동인구

주거 대 유동비가 높은 지역에 선정

- ^{높은} 지역 : 주거인구 ▲ 유동인구 ▼ **유동인구 수 유동인구 수**

- 낮은 지역: 주거인구 ▼ 유동인구 ▲

경제적 취약 계층



기초생활수급자 수

- 경제적 취약 계층의 사람들이 감염병에 노출되기 쉬움
- **기초생활수급자 수**가 많은 지역 고려

2020, "경기도 코로나바이러스 일일상황분석 보고서"

호흡기 전담 클리닉 고려 요건 변수 점수화 및 행정동별 순위 시각화 진행



고려 변수 데이터 점수화 및 시각화

데이터 정규화

점수화(score)

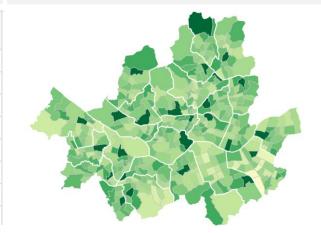
구 내 행정동별 입지 점수(score) 순위 산정

시각화

구 내 행정동 별 순위 시각화

자치구	[‡] 행정동	의료 [‡] 인수	주차구 [‡] 획수	총수급 [‡] 자_인원	주거대 [‡] 유동	score
강남구	수서동	0.0045	0.3315	0.5843	0.2316	1.1519
강동구	천호2동	0.0321	1.0000	0.3145	0.1711	1.5176
강북구	번2동	0.0077	0.2102	0.2905	0.5470	1.0554
강서구	등촌3동	0.0117	0.0553	1.0000	0.3866	1.4537
관악구	난곡동	0.0070	0.0964	0.2964	0.4188	0.8187
광진구	자양3동	0.2119	0.0000	0.0526	0.4877	0.7522
구로구	구로4동	0.0023	0.2685	0.1269	0.4888	0.8865
금천구	시흥4동	0.0002	0.3132	0.1983	0.4974	1.0090
노원구	중계2.3동	0.0010	0.1274	0.8350	0.3594	1.3228
도봉구	도봉1동	0.0010	0.8223	0.2267	0.3908	1.4409
동대문구	제기동	0.0039	0.2421	0.2321	0.2939	0.7719

자치 [‡] 구	행정 [‡] 동	ranks
강남구	수서동	1
강남구	일원본동	2
강남구	개포1동	3
강남구	일원2동	4
강남구	삼성1동	5
강남구	일원1동	6
강남구	개포4동	7
강남구	도곡1동	8



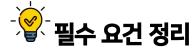
- 데이터 단위 통일을 위한 정규화
- *min-max scaler

$$x_{scaled} = rac{x - x_{min}}{x_{max} - x_{min}}$$

· 정규화 데이터 합으로 입지점수 계산

- · 구 내 행정동별 입지점수 순위 산정
- · 구 내 행정동별 입지점수 순위 시각화
- · 구별 호흡기전담클리닉 할당 수를 고려하여 순위가 높은 행정동에 병원 선정

호흡기전담클리닉 설치·운영 계획을 확인한 후, 위치 선정 필수 요건 도출



호흡기전담클리닉 설치·운영 계획

"병원급 국민안심병원은 호흡기전담클리닉으로 전환"

"구 보건소에 1개소 이상 설치"

"특정 지역에 설치가 집중되지 않도록 유의"

필수 요건

서울 내 병원급 <u>국민안심병원</u>은 **선정 배제**

서울 구 보건소는 선정 배제

선정 호흡기전담클리닉 <u>반경</u> 500m이내 선정 배제

보건복지부

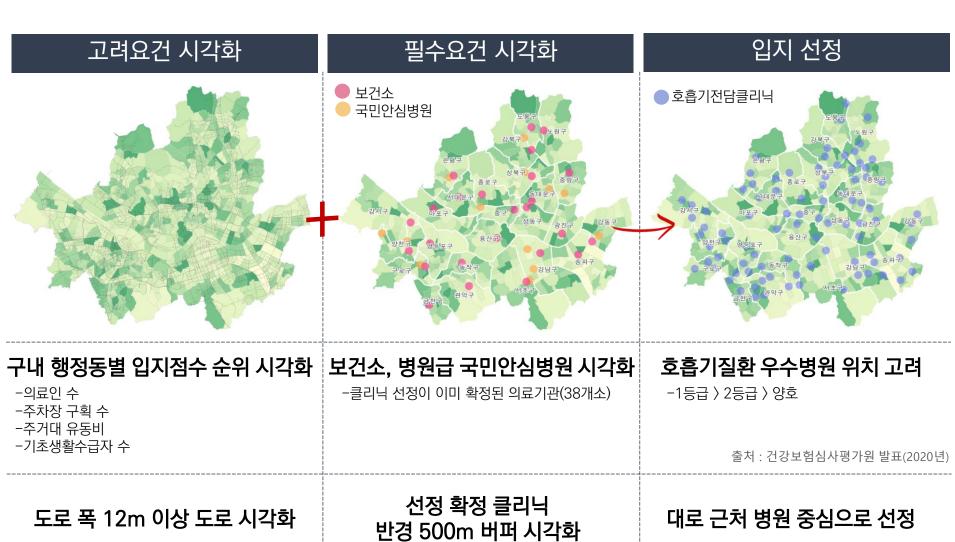
코로나19 장기화 대비

호흡기전담클리닉 설치·운영 계획

2020. 7. 13.



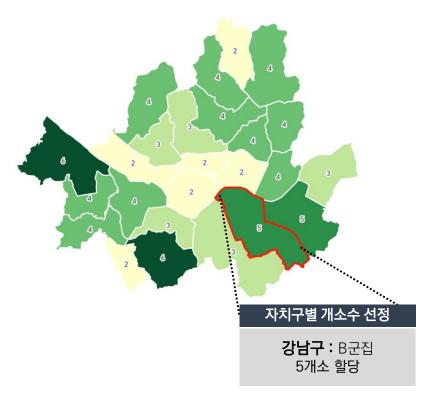
고려요건, 필수요건 통합 시각화



입지 선정 과정 (예시 – 강남구)

Step 1

호흡기전담클리닉의 서울 자치구별 개소 수 선정



자치구	label	real_value	predict_value	호흡기전담클리닉_설치수
강남구	1	367	263.764843	5

Step 2

행정동별 코로나 입지선정 표준점수 순위 시각화



자치 [‡] 구	행정 [‡] 동	ranks	÷
강남구	수서동		1
강남구	일원본동		2
강남구	개포1동		3
강남구	일원2동		4
강남구	삼성1동		5
강남구	일원1동		6
강남구	개포4동		7

행정동별 순위 시각화

의료인프라, 접근성, 주거인구, **차상위계층** 점수 순위

강남구에서는 '<mark>수서동'</mark>이 의료인프라, 접근성, 주거인구, 경제적 취약 계층 점수가 가장 높은 행정동이다.

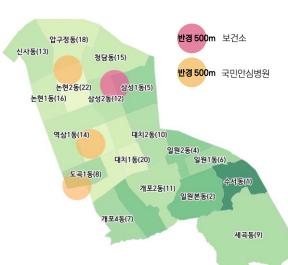
1순위로 선정!

19

-일지 선정 과정 (예시 – 강남구)

Step 3

보건소와 국민안심병원 위치 시각화



필수요건 고려

이미 호흡기전담클리닉 선정이 확정된 **보건소와 병원급 국민안심병원** 위치 시각화

특정 지역에 설치가 집중되지 않도록 **반경 500m** 버퍼 시각화

Step 4

폭 12m 이상 도로, 호흡기질환우수병원 시각화

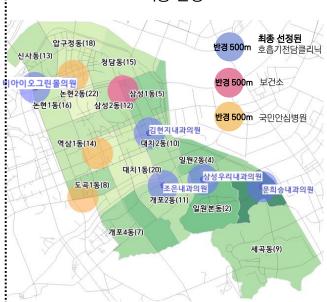


도로 폭, 우수병원 시각화

호흡기질환 우수병원 후보에서 선정하되, 접근성을 고려하여 폭 12m 이상인 도로 인접 병원 우선 선정

Step 5

최종 선정



최종 입지선정

행정동 순위별로 Step 3 ~ 4을 반복하여 호흡기전담클리닉 최종 입지 선정

반경 500m 버퍼 시각화를 통해 **인접 지역**에는 선정 배제

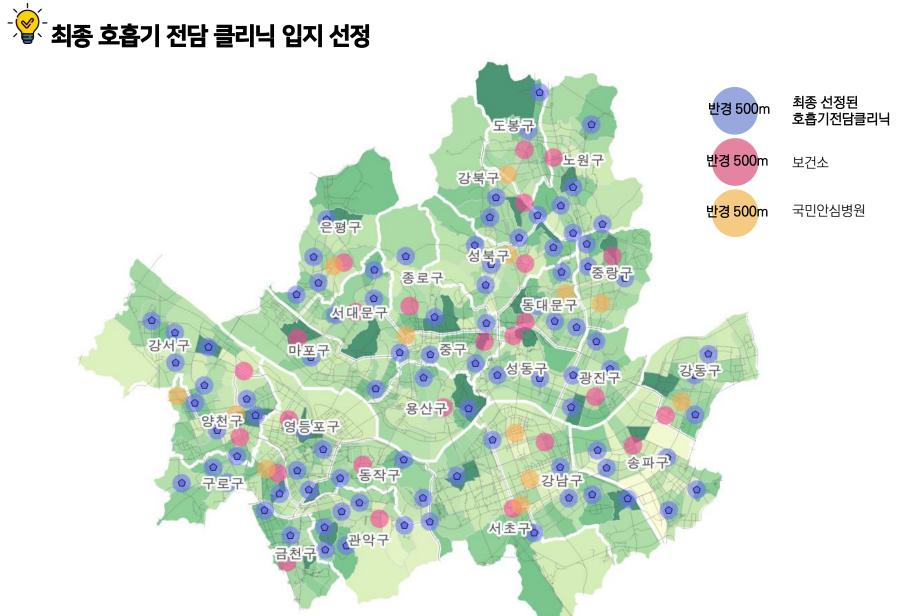
군집별 입지 선정 결과



화곡2동(13)







CONCLUSION 1. 활용방안



서울시



자치구 호흡기전담클리닉 개소수 선정

- 표준화된 모델 기반의 분석 프로세스 운영으로 신속하고 효율적인 정책반영 가능
- 호흡기전담클리닉 관련 정책 수립 및 적용에 활용 가능

서울시 자치구



자치구 내 호흡기전담클리닉 입지 선정

- 주거인구, 인구밀도, 교통 인프라, 의료인 수 등을 종합적으로 고려하여 입지별 추천 순위 제공
- 입지 선정 모델을 기반으로 자치구별 변수
 가중치를 조정하여 해당 자치구 맞춤형 호흡기
 전담 클리닉 입지 선정 가능

CONCLUSION 2. 기대효과 및 개선방향



기대효과

객관적인 선정 기준 마련

의사결정권자 및 평가자의 주관이 반영되는 기존 호흡기전담클리닉 선정 방식과 달리 **객관적인 선정 기준 마련**

취약계층 호흡기 질환 의료 접근성 개선

자치구별 고령인구, 전염병 취약질환 보유여부 등을 고려한 클리닉 개소 수 산출을 통해 전염병 **취약계층의 의료 접근성 개선**

효과적인 호흡기질환클리닉 운영

효과적인 클리닉 운영을 통해 병원 내 전파와 의료 마비 방지하고 의 · 병원의 정상적인 기능을 보장

개선방향



다양한 질병 관련 보건의료 빅데이터 이용

인플루엔자, 메르스 등 <u>다양한</u> 전염성 호흡기질환 데이터를 추가하여 **더욱 정교한 모델을 설계할 수 있을 것으로 예상**

보건의료 빅데이터의 개별데이터 확보를 통한 모델 정교화 가능

CONCLUSION



참고문헌

사용도구



2015, "메르스, 그리고 한국에서의 발생과 그 확산", 채윤태

2020.07. "코로나바이러스감염증-19 국내 발생 현황", 중앙방역대책본부

2020.09. "코로나바이러스감염증-19 국내 발생 현황", 중앙방역대책본부

2020, "경기도 코로나바이러스 일일상황분석 보고서", 경기도감염병관리지원단

2020.08, "요양급여의 적용기준 및 방법에 관한 세부사항", 보건복지부

2020.07, "코로나19 장기화 대비 호흡기전담클리닉 설치·운영 계획", 보건복지부

2020.07, "2020년도 호흡기전담클리닉 설치·운영 지원사업 예산집행지침" 보건복지부

2020.05" Surveillance of COVID-19 at long-term care facilities in the EU/EEA", ECDC











기사

2020.06. "교회 고위험사업장 학원 정은경이 꼽은 코로나 확산 장소", 한국경제, 이지현, 배태웅

2020.08, "오늘부터 수도권 '12종 고위험시설' 닫는다", 머니투데이, 최태범

의원 인허가정보 목욕장업 인허가정보

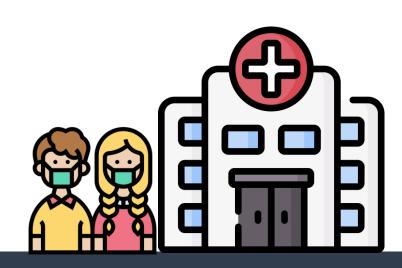


구분	활용 데이터	생성 주기	주요속성	데이터 소스
인구조사	시군구별 흡연율 현황	매년	흡연율	KOSIS 국가 통계 포 털
	고령자(65세이상) 가구(시군구)		가구원 모두 65세 이상 가 구수, 가구원 중 1명 이상 65세 이상 가구	
	가구형태별 가구		총가구, 1~5인 및 6인 이 상 가구수, 남자가구원, 여 자가구원	
	가구원수별가구		1~6인 및 7인 이상 가구수	
의료정보	당뇨병 진단 경험률		당뇨병 환자수	
	고혈압 진단 경험률		고혈압 환자수	
	진해거담제 청구건수	매달	진해거담제 청구건수	보건의료빅데이터개방 시스템
업종	물류창고업체 인허가정보	수시	좌표정보, 창고면적	지방행정 인허가 데이 터 개방
	유흥주점영업 인허가정보		좌표정보, 시군구 별 수	
	단란주점영업 인허가정보			
	노래연습장업 인허가정보			
	다단계판매업체 인허가정보			
	전화권유판매업 인허가정보			
	병원 인허가정보		좌표정보, 의료인 수, 병상 수, 진료과목내용명, 입원	

좌표. 소재지 면적

구분	활용 데이터	생성 주기	주요속성	데이터 소스
인구조사	서울시 시군구 단위 성연령별 1 인가구 수 정보(2019,2016)	매년	시군구 단위 1인가구 성연 령별 수	서울시 빅데이터 캠퍼 스
	서울시 주거인구	수시	동별 위치 좌표, 연령대별 주거인구	국가공간정보포털 오 픈마켓
	행정동별 서울생활인구(내국인)	매일	행정동별, 연령별, 성별 생 활인구	서울 열린데이터 광장
	서울시 국민기초생활보장수급자 (동별) 통계	매년	동별 국민기초생활보장수 급자 수	
위치정보	검진기관정보(2013)	수시	보건소 위치	공공데이터포털
	전국 주차장 정보 표준데이터	정기	주차장 주소, 주차장 위치, 주차장 구획수	
	서울시 도로구간 위치 정보	수시	도로 폭 정보, 도로위치	서울 열린 데이터 광장
	병원 인허가 정보, 의원 인허가 정보		병원 의원수, 의료인 수, 동 별 병원수, 병원 주소 및 이 름	지방행정 인허가 데이 터개방
	국민안심병원(병원급) 위치		국민안심병원 위치	보건복지부
	호흡기질환 우수병원 리스트	정기	우수 병원 등급, 우수 병원 주소 및 이름	
보건	서울, 인천, 경기 시군구 별 코로 나 확진자 수	매일	서울, 인천, 경기 시군구 별 코로나 확진자 수	자치구 행정 사이트

감사합니다:)



Appendix

1. 데이터 가공 이전(전체 데이터)

_																				
							더	0 E	付섯	Į										
	SIG_CI) 도 시	자 치 구	총가구_ 가구원_ 계_명	일반가구 _가구원_ 계_명	집단가구_비혈연 _6인이상_가구_ 가구원_계_명	집단가구_집 단시설가구 가구원_계_명	두 65세이	당 뇨 병 소 계	고혈 압소 계	대 규 모 점 포 수	업	주점	노래 연습 장업 수	병 원 수	의	병원 _입 원실 수	병원 _병 상수	의 원 수	확 진 자 수
0	1111	서 울 등 별 시	로	154969	136649	640	485	4 8079	374	797	2	30	434	102	6 4	4769	835	2890	56	98
1	1114	서 울 열 년 시	_	130957	115665	690	232	8 6990	317	712	15	45	654	77	5	1325	258	927	176	66
2	1117	서 울 열 시		227181	204196	428	238	7 11019	555	1112	5	32	88	82	3	1000	215	837	90	150
								<u> </u>		7	7(ı	O\	(v	Χ	29	9(co	lu	m	n)
						회귀	분석	– F	ull	n	าด	d€	el							
	D	ep.	Vai	riable		확진지	 	R-sq	uared	i:	0	.860								
			N	/lodel:		0	LS A	dj. R-sq	uared	i:	0	.684								
			Ме	ethod	: 1	Least Squar	res	F-sta	atistic	::	4	.883								
				Date	Mon	, 09 Nov 20	20 Pro	ob (F-sta	tistic):	0.000	115								
				Time	:	11:19:	42 L	og-Likeli	ihood	i:	-27	0.14								
N	lo. O	bse	rva	tions	:		53		AIC	:	6	00.3								
	0	f R	esi	duals			23		BIC	:	6	59.4								
		[Of N	/lodel:	:		29													
	Cova	riar	ice	Type		nonrob	ust													
									coef	st	d en		t	F)> t	t	[0.0	025	0.	975]
						N19	인_가구속	è -0.	0081	(0.030	-(.270	0.	790	0	-0.0		0	.054
							· 인 가구=		0327	(0.071	-(.458	0.	65	1	-0.1	181	0).115
							· · 인_가구=		0069		0.091		0.076		940		-0.1			.195
							이상가-		0629		0.134		.470		643		-0.3			.214
						PC방 총			0341		0.171		.199		844		-0.3			.389
						FC0_8													_	

1.1572 35.404 0.033 0.974 -72.082 74.397

상관관계, VIF 확인 VIF Factor features 0 1.577685e+04 N1인_가구수 N2인_가구수 1 6.284581e+04 2 7.375177e+04 N3인_가구수 3 1.942908e+05 N4인_이상가구 4 5.865408e+00 PC방_총사업체수 가구원_모두_65세이상_가구수 5 1.677044e+02 +12-84



1. 데이터 가공 이전(전체 데이터)

							E-	0 E	l셋											
	SIG_CD	도시	자 치 구	충가구_ 가구원_ 계_명	일반가구 _가구원_ 계_명	집단가구_비혈연 _6인이상_가구_ 가구원_계_명	집단가구_ 단시설가구 가구원_계_'	L_ 두_65세이		고혈 압소 계	대 규모 점 포 수	목 약 장 접 수	유흥 주점 영업 수	노래 연습 장업 수	8	_의	병원 _입 원실 수	병원 _병 상수	의 원 수	확 진 자 수
0	11110	서울특별시	종로구	154969	136649	640	485	54 8079	374	797	. 2	30	434	102	6 4	1769	835	2890	56	98
1	11140	서 울특 별 시	중구	130957	115665	690	232	28 6990	317	712	. 15	45	654	77	5 1	1325	258	927	176	66
2	11170	서울 특별시	용 산 구	227181	204196	428	238	37 11019	555	1112	. 5	32	88	82	3 1	1000	215	837	90	150
								\downarrow		7	7(r	OV	v)	Χ	29	9(c	col	lur	m	n)
						회귀	분석	– F	ull	m	100	dε	əl							
	De	p. \	Vai	iable:		확진자	·수	R-sq	uared	:	0.	860	Г							
			N	lodel:		0	LS A	Adj. R-sq	uared	l:	0.	684								
			Me	thod:		Least Squar	es	F-sta	atistic	::	4.	883								
				Date:	Mon	, 09 Nov 20	20 Pr	ob (F-sta	tistic): 0	.000	115								
				Time:		11:19:	42 L	.og-Likeli	ihood	l:	-270).14								
N	lo. Ob	sei	va	tions:			53		AIC	:	60	0.3								
	Di			duals:			23		BIC	:	65	9.4								
			f N	lodel:			29													
	_			-																
(Covar	ian	ce	Type:		nonrobi	ust													
(Covar	ian	ce	Type:		nonrob	ust		coef	sto	l err		1		P> t	ı	[0.0]	25	0.	975]
(Covar	ian	ce	Type:			ust 인_가구:		coef 0081		l err .030	-0	.270		P> t .790		[0.0 -0.0			9 75] .054
(Covar	ian	ce	Type:		N19		수 -0.		0				0)	•	70	0	
(Covar	ian	ce	Туре:		N19 N29	인_가구:	수 -0.0 수 -0.0	0081	0	.030	-0	.270	0	.790) 	-0.0	70 81	0	.054
(Covar	ian	ce	Туре:		N19 N29 N39	인_가구: 인_가구:	수 -0.6 수 -0.6 수 0.6	0081 0327	0.	.030	-0 0	.270 .458	0	.790 .651) 	-0.0 -0.1	70 81 81	0	.054

1.1572 35.404 0.033 0.974 -72.082 74.397



상관관계와 VIF를 확인하며, 가공할 데이터 혹은 삭제한 데이터 선정

1. 다른 변수들과 상관관계가 높고, VIF가 높은 변수

- 일반가구_가구원_계 삭제
- 병원_입원실수 삭제
- 총가구_가구원_계_명 삭제

2. 상관계수 plot에서 특정 변수와 패턴이 비슷한 변수

- 당뇨병, 고혈압 변수와 가구원_모두_65세이상_가구수 변수의 패턴이 비슷함
 - → 65세 이상의 사람들이 당뇨병, 고혈압이 많을 것
 - → 당뇨병, 고혈압 변수 삭제
- 유입인구와 유출인구의 패턴이 비슷함
 - → 유입유출인구: 유입+유출
- N2인_가구수, N3인_가구수, N4인이상_가구수 변수의 패턴이 비슷함
 - → N2인이상_가구: N2인 + N3인 + N4인이상



2. 데이터 삭제 및 가공

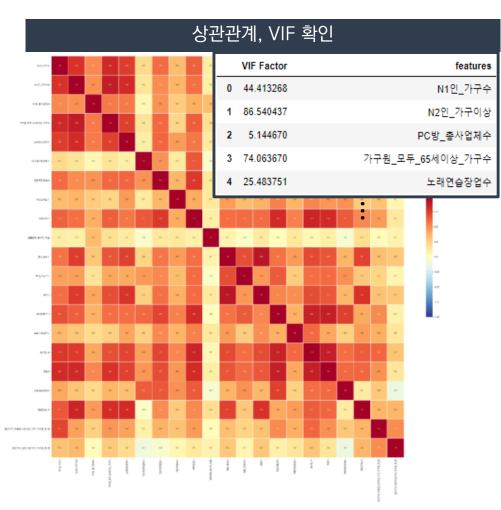
	데이터셋																				
	SIG_CD	도시	자 치 구	집단가구_비혈연 _6인이상_가구_ 가구원_계_명	집단가구_집단 시설가구_가구 원_계_명	가구원_모 두_65세이 상_가구수	방문판 매_종사 자_비율	N1인 _가구 수	N2인 _가구 이상	종 교 단 체_ 수		단란 주점 영업 수	대 규모 점포 수	목 욕 장 업 수	유흥 주점 영업 수	노래 연습 장업 수	병 원 수	병원 _의 료인 수	병원 _병 상수	의 원 수	확 진 자 수
0	11110	서울특별시	종로구	640	4854	8079	3.14	24671	37981	407		117	2	30	434	102	6	4769	2890	56	98
1	11140	서 울특별 시	중구	690	2328	6990	32.56	22078	32418	163		84	15	45	654	77	5	1325	927	176	66
2	11170	서울특별시	용산구	428	2387	11019	3.96	35148	58029	180		101	5	32	88	82	3	1000	837	90	150

77(row) X **21(column)**

회귀분석 - Full model

Dep. Variable:	5	학진자수		R-90	uared:	0.88	8	
Model:		OLS	A	dj. R-eq	uared:	0.76	7	
Method:	Least	Squares		F-st	atietic:	7.35	7	
Date:	Wed, 23 S	ep 2020	Pro	b (F-eta	itletic):	1.73e-0	6	
Time:	(09:55:33	Lo	g-Likel	lhood:	-249.7	2	
No. Observations:		53			AIC:	555.	4	
Of Residuals:		25			BIC:	610.	6	
Df Model:		27						
Covariance Type:	no	onrobust						
		00	of a	etd err	t	P>Iti	[0.025	0.975]
N14	인 가구수	0.001	-	0.001	0.703	0.489	-0.002	0.004
	_							
N2인	이상_가구	-0.001	12	0.003	-0.365	0.718	-0.008	0.005
PC방_é	· 흥종사자수	0.092	23	0.206	0.447	0.658	-0.333	0.517
	const	-22.180	04	17.443	-1.272	0.215	-58.104	13.743
가구원_모두_65세이	상_가구수	0.002	26	0.005	0.552	0.586	-0.007	0.012
3	고혈압소계	-0.008	33	0.047	-0.176	0.861	-0.106	0.089

 R^2 는 높지만, 변수 모두 p-value가 높음

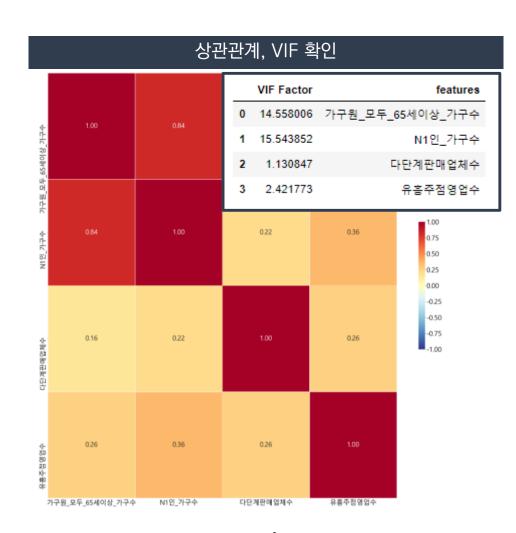


- 상관관계: 강한 양의 상관관계의 변수가 많음
- VIF: 높음
- 독립변수간 상관관계가 높음



	변수신	선택법	- 단계	선택			
Dep. Variable:	į	박진자수	R-s	quared:	0.7	756	
Model:	Model: OLS Adj. R-squared:		0.7	736			
Method:	Least	Squares	F-s	statistic:	37	.17	
Date:	Mon, 09 N	ov 2020	Prob (F-s	tatistic):	3.81e	-14	
Time:		11:20:13	Log-Lik	elihood:	-285	.83	
No. Observations:		53		AIC:	58	1.7	
Df Residuals:		48		BIC:	59	1.5	
Df Model:		4					
Covariance Type:	no	onrobust					
		coef	std err	t	P> t	[0.025	0.975
가구원_모두_65세이	상_가구수	0.0083	0.002	3.675	0.001	0.004	0.01
N1	인_가구수	0.0017	0.001	2.544	0.014	0.000	0.00
다단계	판매업체수	0.9113	0.541	1.686	0.098	-0.176	1.99
유흥	주점영업수	0.0689	0.042	1.650	0.105	-0.015	0.15
	const	-41.5863	19.883	-2.092	0.042	-81.563	-1.60

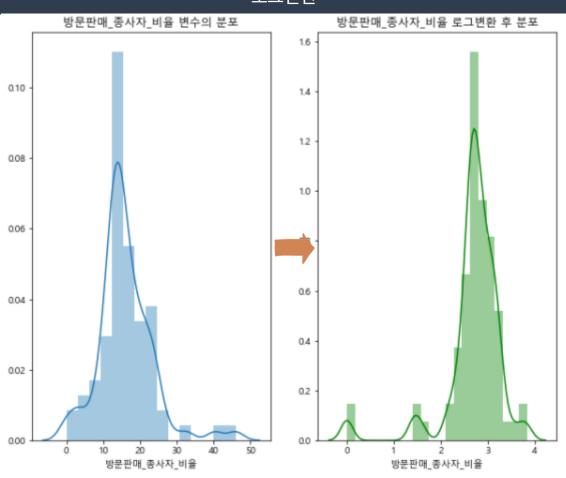
- Full model보다 R² 는 낮지만, p-value가 낮음
- Test에서의 R²
 - stepwise model r2: 0.7136888525617864



- 상관관계: 인구 관련 변수끼리 상관관계가 0.84로 높음
- VIF: 10 이상인 변수 2개



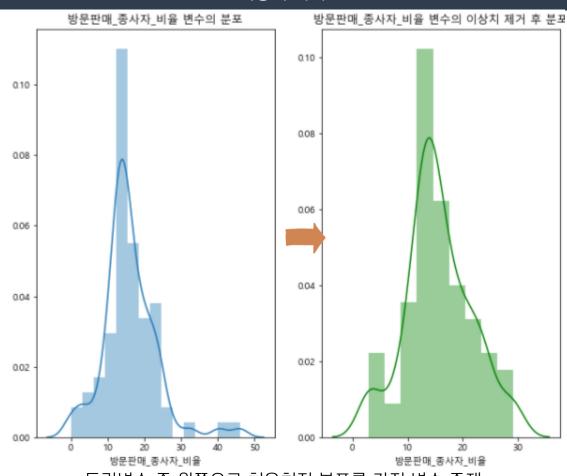




로그변환을 수행해도 정규분포의 형태를 가지지 않음 → 따라서 로그변환한 데이터로 회귀모델을 진행하지 않고 이상치 대체 수행



이상치 대체



독립변수 중 왼쪽으로 치우쳐진 분포를 가진 변수 존재

→ 이상치 대체만으로도 정규분포의 형태를 보임

→ 이후 정규화 단계에서 이상치 대체한 데이터로 분석 진행



대규모점포수

병원수

의원수

목욕장업수 유흥주점영업수 노래연습장업수

병원_의료인수

병원_병상수

이상치 대체

이상치 대체 전 변수 분포의 왜도, 첨도

집단가구_비혈연_6인이상_가구_가구원_계_명 Skewness: 01.61 Kurtosis: 03.87 집단가구_집단시설가구_가구원_계_명 Skewness: 01.36 가구원_모두_65세이상_가구수 Skewness: 00.43 Kurtosis: -0.41 방문판매_종사자_비율 Skewness: 01.17 Kurtosis: 04.09

M1인_가구수 Skewness: 00.97 Kurtosis: 02.25 N2인_가구이상 Skewness: 00.48 Kurtosis: 00,46 종교단체 _수 Skewness: 01.27 Kurtosis: 03.76 PC방_총사업체수 Skewness: 02.61 Kurtosis: 09.19 유입유출인구 Skewness: 01.23 Kurtosis: 04.03 음식점_수 Skewness: 00.83 Kurtosis: 01.12 전화권유업체수 Skewness: 02.85 Kurtosis: 09.18 다단계판매업체수 Skewness: 07.93 Kurtosis: 65.91 단한주점영업수 Skewness: 01.36 Kurtosis: 02.32

> Skewness: 01.80 Kurtosis: 07.59 Skewness: 01.88 Kurtosis: 04.20 Skewness: 00.64 Kurtosis: 00.41 Kurtosis: 03.11 Skewness: 01.81

Skewness: 01.32 Kurtosis: 02.21 Skewness: 01.37 Kurtosis: 03.86 Kurtosis: 02.54 Skewness: 01.17 Skewness: 00.88 Kurtosis: 01.27

이상치 대체 후 변수 분포의 왜도. 첨도

집단가구_비혈면_6인이상_가구_가구원_계_명 Skewness: 00.81 Kurtosis: 00.06 집단가구_집단시설가구_가구원_계_명 Skewness: 00.88 Kurtosis: -0.05 가구원_모두_65세미상_가구수 Skewness: 00.43 Kurtosis: -0.41

방문판매_종사자_비율 Skewness: 00.15 Kurtosis: 00.16

N1인_가구수 N2인_가구이상 종교단체 _수 PC방_총사업체수 유입유출인구 음식점_수 전화권유업체수 다단계판매업체수 단한주점영업수 대규모점포수 목욕장업수 유흥주점영업수 노래연습장업수

Skewness: 00.45 Kurtosis: -0.08 Skewness: 00.37 Kurtosis: 00.10 Skewness: 00.44 Kurtosis: -0.20 Skewness: 00.98 Kurtosis: 00.23 Skewness: 00.28 Kurtosis: -0.05 Skewness: 00.59 Kurtosis: 00.28 Skewness: 01.17 Kurtosis: 00.25 Skewness: 01.12 Kurtosis: -0.35 Skewness: 00.73 Kurtosis: -0.19 Skewness: 00.93 Kurtosis: 00.09 Skewness: 00,41 Kurtosis: 00.02 Skewness: 01.02 Kurtosis: 00.11 Kurtosis: -0.09 Skewness: 00.51 Skewness: 00.59 Kurtosis: 00.07 Skewness: 01.24 Kurtosis: 00.61 Skewness: 00.64 Kurtosis: -0.06 Skewness: 00.65 Kurtosis: 00.45

이상치 대체 후 왜도, 첨도의 절댓값이 줄어들었음 * 왜도, 첨도가 절댓값 2를 넘으면 정규분포가 아니라고 봄

병원수

의원수

병원_의료인수

병원_병상수



정규화

Standard 정규화 후 데이터 * 이상치 제거한 데이터로 정규화 수행

N1인_가 구수	N2인_가 구이상	PC방_총 사업체수	가구원_ 모두_65 세이상_ 가구수	노래연습 장업수	다단계 판매업 체수	단란주점 영업수	대규모점 포수	목욕장업 수	방문판매 _종사자_ 비율
-0.651283	-1.155693	-0.909361	-0.913460	-0.853037	0.360619	1.274706	-0.327192	0.479187	-2.095273
-0.787091	-1.276813	2.258481	-1.086548	-1.064280	1.932375	0.480437	2.866390	1.808025	2.275856
-0.102551	-0.719202	-1.128210	-0.446171	-1.022032	0.360619	0.889606	0.737335	0.656365	-1.957758
0.229505	-0.212930	-0.482605	-0.162141	-0.236208	0.360619	-0.554519	-0.327192	-0.229527	-1.122606
1.131715	0.002812	0.600698	0.048616	1.039699	1.932375	0.360094	0.027650	0.124830	-0.466895

77(row) X **21(column)**

회귀분석 - Full model

OLS Regression Results

Dep. Variable: 확진자수 R-squared: 0.782 Model: OLS Adi. R-squared: 0.634 Method: F-statistic: 5.284 Least Squares Date: Mon, 09 Nov 2020 Prob (F-statistic): 1.72e-05 00:55:41 Log-Likelihood: -282.89 Time: No. Observations: 53 AIC: 609.8 Df Residuals: 31 BIC: 653.1

Df Model: Covariance Type: nonrobust

> coef std err t P>|t| [0.025 0.975] const 153.3863 9.727 15.770 0.000 133.549 173.224 N1인 가구수 35.0186 35.201 0.995 0.328 - 36.775 106.812 N2인_가구이상 46.0581 48.461 0.950 0.349 -52.778 144.894 PC방_총사업체수 -9.5438 17.344 -0.550 0.586 -44.917 25.829 가구원_모두_65세이상_가구수 13.5137 37.273 0.363 0.719 -62.505 89.532 노래연습장업수 6.5875 27.266 0.242 0.811 -49.023 62.198

변수선택법 성능 평가

Test set에서의 R2

```
from sklearn.metrics import r2_score
print('forward model r2:',r2_score(test_y3, pred_y_forward))
print('backward model r2:',r2_score(test_y3, pred_y_backward))
print('stepwise model r2:',r2_score(test_y3, pred_y_stepwise))
print('for model r2:',r2_score(test_y3, pred_y_for))
```

forward model r2: 0.7309266781063978 backward model r2: 0.7459918415378652 stepwise model r2: 0.7309266781063978 for model r2: 0.7441071869920552



군집분석 결과

	SIG_CD	도시	자치구	label
0	11110	서울특별시	종로구	4.0
1	11140	서울특별시	중구	2.0
2	11170	서울특별시	용산구	2.0
3	11200	서울특별시	성동구	2.0
4	11215	서울특별시	광진구	0.0
5	11230	서울특별시	동대문구	0.0
6	11260	서울특별시	중랑구	0.0
7	11290	서울특별시	성북구	0.0
8	11305	서울특별시	강북구	0.0
9	11320	서울특별시	도봉구	2.0
10	11350	서울특별시	노원구	0.0
11	11380	서울특별시	은평구	0.0
12	11410	서울특별시	서대문구	4.0
13	11440	서울특별시	마포구	2.0
14	11470	서울특별시	양천구	0.0
15	11500	서울특별시	강서구	3.0
16	11530	서울특별시	구로구	0.0
17	11545	서울특별시	금천구	2.0
18	11560	서울특별시	영등포구	0.0
19	11590	서울특별시	동작구	4.0
20	11620	서울특별시	관악구	3.0
21	11650	서울특별시	서초구	4.0
22	11680	서울특별시	강남구	1.0
23	11710	서울특별시	송파구	1.0
24	11740	서울특별시	강동구	4.0

확진자수 예측값

자치구 real_value predict_value

	사지구	real_value	predict_value
0	종로구	98	115.234902
1	중구	66	86.210305
2	용산구	150	147.095327
3	성동구	145	146.211707
4	광진구	144	196.020764
5	동대문구	190	235.544398
6	중랑구	211	260.945494
7	성북구	357	264.560194
8	강북구	159	238.078911
9	도봉구	231	210.164013
10	노원구	337	356.563967
11	은평구	278	281.719488
12	서대문구	150	219.209164
13	마포구	189	212.164743
14	양천구	199	219.579106
15	강서구	313	362.516344
16	구로구	210	231.314615
17	금천구	98	134.031342
18	영등포구	220	202.282192
19	동작구	270	263.798461
20	관악구	483	339.540146
21	서초구	242	206.575501
22	강남구	367	263.764843
23	송파구	432	305.110873
24	강동구	207	224.692670

군집별 호흡기전담클리닉 설치개수 선정

	group_label	mean_predict_values	size_count	rank	호흡기전담클리닉_설치개수
0	3	351.028245	2	1	6
1	1	284.437858	2	2	5
2	0	248.660913	10	3	4
3	4	205.902139	5	4	3
4	2	155.979573	6	5	2

- 1. 구별 군집분석 결과와 구별 확진자수 예측값 join
- 2. 군집별 확진자수 예측값 평균(mean_predict_values)
- 3. mean_predict_values 순위별 개수 차등 선정
- 4. 서울시에 추가로 설치되는 개수는 약 120개
- 5. 설치 개소수는 2~6개가 적정