접수번호	

「2022년 통계데이터 분석 · 활용대회」 데이터분석 보고서

제 목	주변 생활 시설에 따른 광역지방자치단체별 2, 30대 1인 가구 여성 인구의 비율 분석

신 청 자 명	소속/직위	고려대학교	성 명	윤민서
	휴대전화	010-3256-3401	전자우편	yms020615@gmail.com
제출일	2022-07-22			

주변 생활 시설에 따른 광역지방자치단체별 2, 30대 1인 가구 여성 인구의 비율 분석

1. 배경

□ 주제

대한민국 내 17개 광역지방자치단체별 2, 30대 1인 가구 여성 인구(이하 여성 인구)의 비율에 도서관, 치안 시설, 전기차 충전소 등 9개의 주변 시설이 어떠한 영향을 주는지 분석한다. 현재 상대적으로 여성 인구의 비율이 낮거나 증가 추세가 크지 않은 지역들이 어떠한 주변 생활 시설을 보충하면 좋을지 제시한다.

□ 분석 필요성(문제점) 및 전략

2016년에서 2020년까지 5년 동안 여성 인구가 빠르게 증가하고 있으며 증가 추세도 점점 커지고 있다. 2016년에서 2017년으로 넘어갈 당시에는 약 2.8%가 증가했으나 2020년에서 2021년 사이에는 약 12.73%로 증가 추세가 4년 사이에 약 5배나 증가하였다. 이런 사회적 흐름에 맞추어서 여성 인구의 비율이 빠르게 늘어나는 것을 광역시/도 차원에서 어떻게 대비해야 할지 미리 분석할 필요가 있다고 판단했다.

분석을 위해서 2020년 기준 전체 인구에 대한 여성 인구와 여성 인구의 증가율을 각각 x축, y축으로 하여 군집화를 진행한다. 이후 전체 인구에 대한 여성 인구의 비율을 종속변수로 설정하고 문화 시설, 도서관, 치안 시설, 30년 이상 노후 주택, 전기차 충전소, 치킨 전문점, 아파트, 연립 및 다세대 주택, 사설 학원 9개 변수에 대해서 산점도를 그리고 상관계수를 분석한다. 앞의 분석 결과를 바탕으로 최적화 다중 선형 회귀분석 알고리즘 중 하나인 단계별 선택 다중 선형 회귀분석을 통해 여성 인구에 영향을 유의미하게 주는 변수가 무엇인지 찾는다.

2. 데이터 분석

□ 데이터 선정

통계지리정보서비스에서 제공하는 데이터 중 2020년의 데이터를 제공하는 항목 중에서 여성 인구에 영향을 줄 수 있는 것을 우선으로 고려하였다. 여성 인구의 비율을 구하기 위해 인구 변화 카테고리에서 2020년 기준 전체 인구를 수집하였다. 차후에 상관계수 분석을 통해서 전체 인구에 대한 여성 인구와 상관관계가 강하게 나타나지 않는 변수를 제거를 하는 작업을 진행할 것이기 때문에 데이터

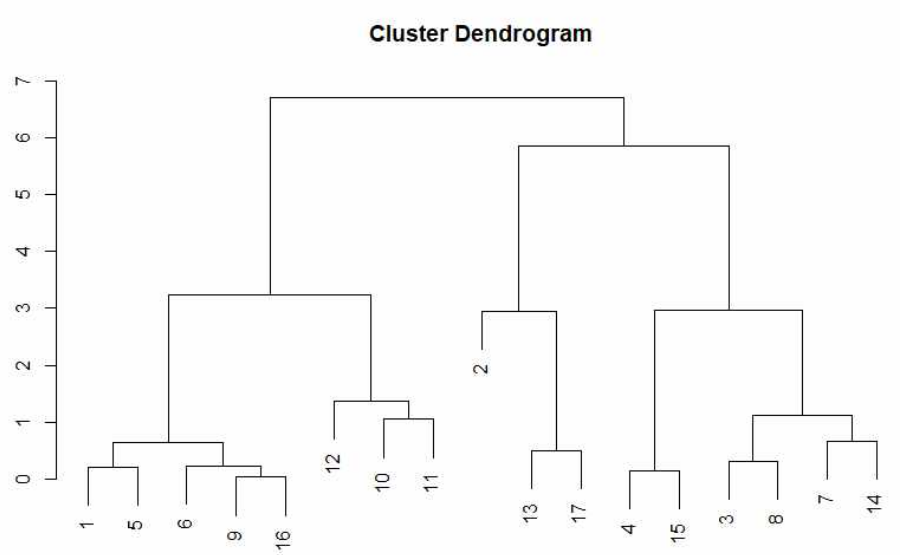
를 수집하는 초반 과정에서는 최대한 다양한 지표를 수집하고자 노력하였다.

□ 데이터 분석(분석 프로세스, 분석방법, 접근방법 등)

가장 먼저 광역지방자치단체별로 전체 인구에 대한 여성 인구의 비율을 구한다. 다음으로 전체 인구에 대한 여성 인구를 x축, 여성 인구 비율의 증가율을 y축으로 하여 군집화를 진행한다. 군집의 개수를 구하기 위해서 계층적 군집화 방법을 이용하여 계통도를 그린다. 군집의 개수가 명확히 나누어지지 않는다면 k-평균 알고리즘을 사용하여 군집의 개수인 k를 구한다.

분석 단계에서는 각각의 군집에 대하여 전체 인구에 대한 여성 인구의 비율을 종속변수로 설정하고 상단 전략 부분에서 서술한 9개의 변수에 대하여 산점도를 그리고 상관성을 시각적으로 분석할 수 있는 변수와 그 중에서 상관계수의 절댓값이 유의미한 정도로 크게 나타나는 변수를 추출한다. 이후 회귀분석을 통해 군집별로 어떤 변수가 얼마나 영향을 미치는지 찾는다. 회귀분석의 종속변수로 포함되는 변수는 산점도와 상관계수 분석으로 일차적으로 선별된 변수들이 된다. 이 과정에서 단계적 선택 다중 선형 회귀분석을 통하여 회귀방정식의 각각의 x 변수의 설명력, 회귀방정식 자체의 유의미성을 높게 만드는 변수를 선별한다.

□ 분석 결과 및 해석



[그림 1: 계통도]

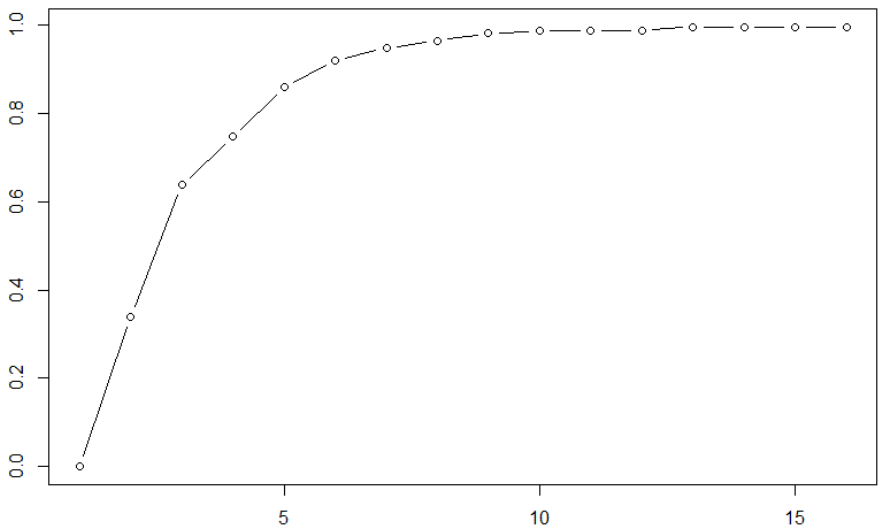
전체 인구에 대한 여성 인구의 비율, 여성 인구의 증가율 두 변수를 각각 x축, y축으로 하여 군집화를 하기 위해 계통도를 그렸다. 계통도를 그리기 전 x축 변수, y축 변수 모두 표준화를 진행하였다. 군집의 개수가 4, 5, 6개인 계층의 경계

가 명확하지 않다고 판단하여 계통도를 통해 군집화를 하는 것을 기각하였다. 잎 노드로 표현된 지역의 번호에 대응되는 지역은 다음과 같다.

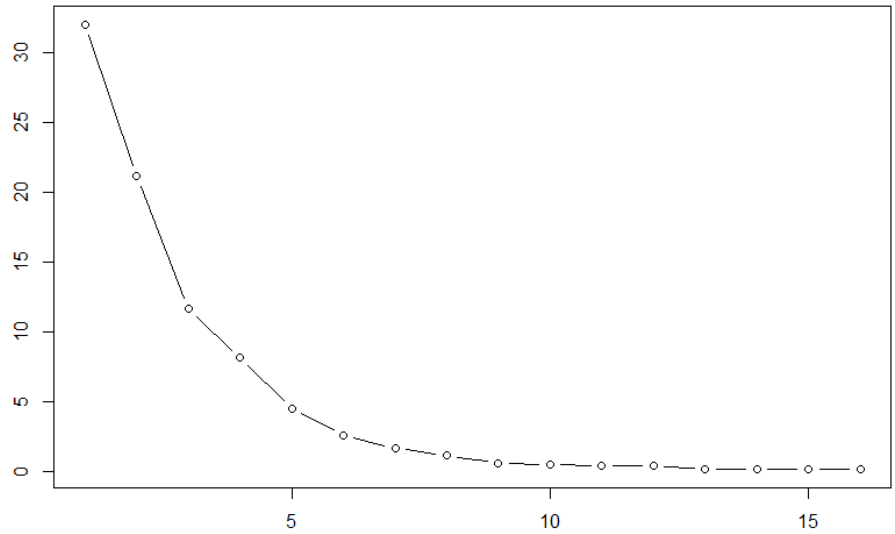
1	경기도
2	서울특별시
3	부산광역시
4	경상남도
5	인천광역시
6	경상북도
7	대구광역시
8	충청남도
9	전라북도
10	전라남도
11	충청북도
12	강원도
13	대전광역시
14	광주광역시
15	울산광역시
16	제주특별자치도
17	세종특별자치시

[표 1: 지역 번호]

다음으로는 k-평균 알고리즘을 이용하여 군집화를 진행하였다. 최적의 군집 개수를 구하기 위해 실루엣 방법, 팔꿈치 방법을 이용하였다.

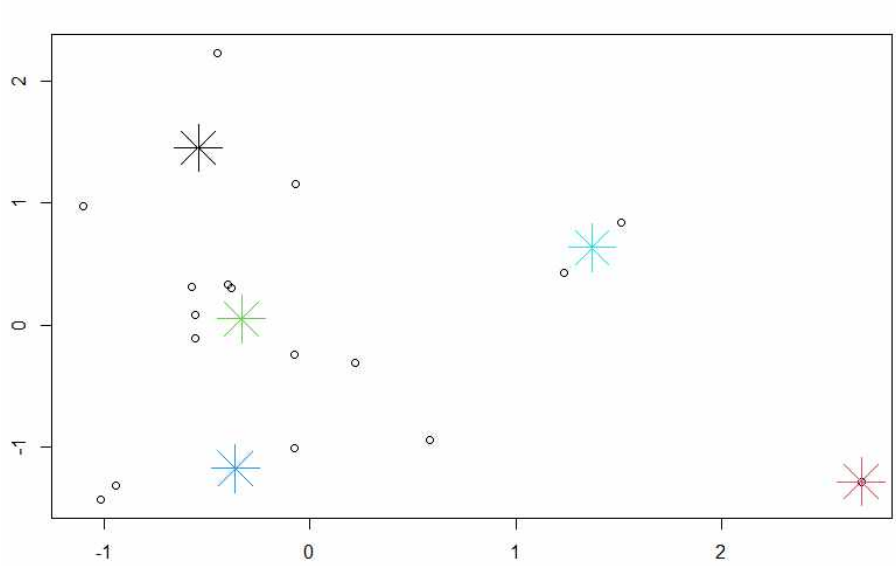


[그림 2: 실루엣 방법의 결과]



[그림 3: 팔꿈치 방법의 결과]

두 방법 각각의 결과 모두 k가 5일 때 군집이 가장 적절하게 형성된다고 판단하였다. 이러한 결과를 이용하여 17개의 광역시/도를 k-평균 알고리즘으로 5개의 군집으로 나누었다. 전체 인구에 대한 여성 인구의 비율, 여성 인구의 증가율 두 변수를 각각 x축, y축으로 하여 실행한 k-평균 알고리즘의 결과는 다음과 같다.

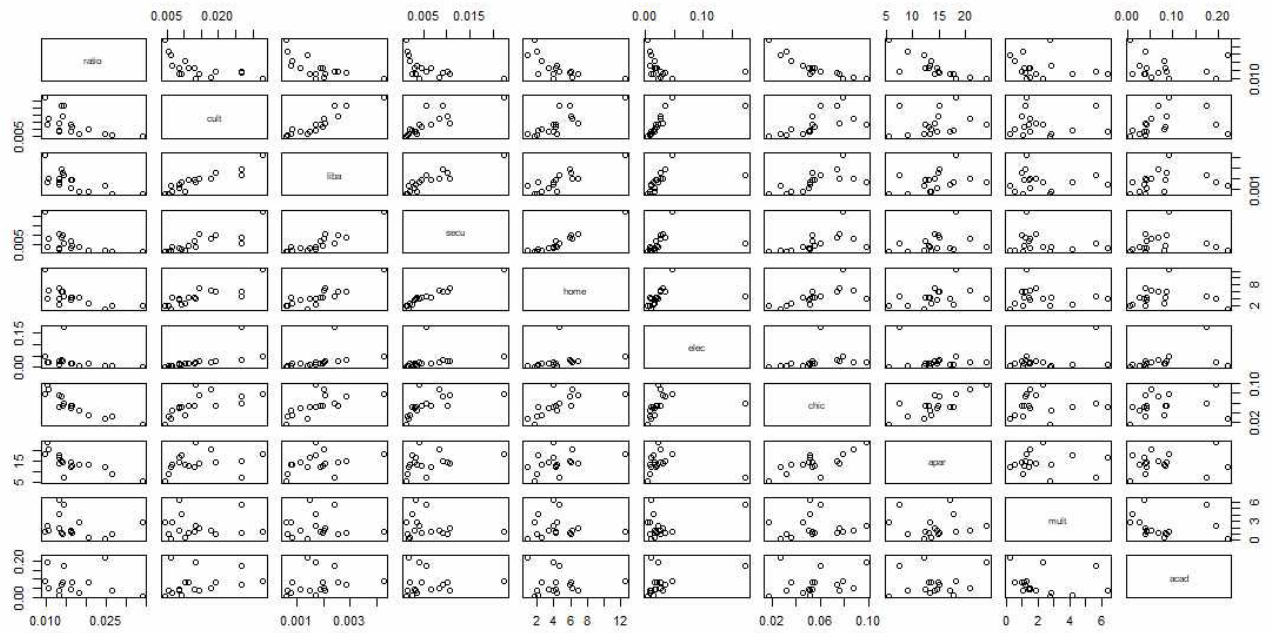


[그림 4: k-평균 알고리즘의 결과]

1 - 검은색	전라남도, 충청북도, 강원도
2 - 연두색	경기도, 부산광역시, 인천광역시, 경상북도, 충청남도, 전라북도, 제주특별자치도
3 - 파란색	경상남도, 대구광역시, 광주광역시, 울산광역시
4 - 하늘색	대전광역시, 세종특별자치시
5 - 빨간색	서울특별시

[표 2: 군집 별 지역]

군집 별 분석을 하기 전에 17개 광역시/도 전체에 대해 여성 인구의 비율과 9개의 주변 시설 간 산점도를 그렸다. 광역시/도 전체에 분포된 개수의 데이터인 도서관, 치안 시설, 30년 이상 노후주택, 전기차 충전소, 치킨 전문점, 아파트, 연립 및 다세대 주택은 여성 인구로 나누었다. 1개 당 인구수로 제공되는 데이터는 역수를 취하여 인구수를 곱한 뒤 여성 인구로 나누었고 1000명 당 사설 학원 개수 데이터는 1000을 곱한 뒤에 여성 인구로 나누었다. 산점도는 다음과 같다.



[그림 5: 변수 간 산점도]

1행(1열)부터 차례대로 전체 인구수에 대한 여성 인구의 비율, 문화 시설, 도서관, 치안 시설, 30년 이상 노후 주택, 전기차 충전소, 치킨 전문점, 아파트, 연립 및 다세대 주택, 사설 학원이다. 이 중 전체 인구수에 대한 여성 인구의 비율과 상관계수를 분석한 결과 절댓값이 0.6이 넘는 변수는 문화 시설(-0.613), 도서관(-0.671), 30년 이상 노후 주택(-0.647), 치킨 전문점(-0.868), 아파트(-0.766) 5개이다.

흔히 강한 음/양의 상관관계를 가진다고 할 때의 경곗값인 0.7을 기준으로 정하

려고 하였지만 표본의 수가 적은 것을 고려하여 적당한 음/양의 상관관계를 가진다고 할 때의 경곗값인 0.5와의 중간을 기준으로 결정하였다. 위 5개의 변수 모두 산점도 상으로도 상관관계가 있다고 판단하기 적절하다고 판정하고 5개의 변수에 대해 전체 인구에 대한 여성 인구의 비율과 단계적 선택 다중 선형 회귀분석을 실시하였다. 분석의 결과는 다음과 같다.

```
Call:
lm(formula = ratio ~ liba + chic + apar, data = c0)

Coefficients:
(Intercept)         liba         chic         apar
  0.0343404   -1.9024166   -0.1330578   -0.0004686
```

[그림 6: 단계적 선택 다중 선형 회귀분석 결과]

회귀분석 과정에서 문화 시설과 30년 이상 노후 주택이 제외되고 도서관, 치킨 전문점, 아파트 3개의 변수가 채택되었다.

전체 인구에 대한 여성 인구가 매우 높은 것으로 인해 군집에 혼자 분류된 서울특별시를 제외한 나머지 4개의 군집에 대하여 회귀방정식을 구한 결과는 다음과 같다. 군집의 번호는 표 2와 같다. 포함된 지역의 개수가 적은 군집은 세 개의 변수에 대하여 회귀분석을 완전히 시행하지 못하는 상황이 발생하기 때문에 각각의 종속변수에 대하여 회귀분석을 진행한 결과를 첨부하였다.

	도서관	치킨 전문점	아파트
전체	-4.7089	-0.2606	-0.0011
군집 1	-2.9236	-0.2372	-0.0013
군집 2	-2.3879	0.3747	-0.0006
군집 3	-8.8011	-0.1656	-0.0009
군집 4	-1.8038	-0.1012	-0.0002

[표 3: 군집 별 회귀분석 결과]

회귀분석 결과(표 3)에서 보이는 특이점은 다음과 같다.

1. 여성 인구의 증가율이 상대적으로 매우 낮은 군집 3에 대하여 도서관이 적을수록 여성 인구의 비율이 작아지는 결과가 두드러지게 관찰되었다.
2. 치킨 전문점의 경우에는 상관계수의 절댓값이 가장 크게 계산된 것에 반해 회귀분석 결과는 불규칙성을 보이고 있다.
3. 여성 인구의 비율이 커짐에 따라 아파트의 수가 적음이 여성 인구의 비율에 미치는 영향이 작아지고 있다.

3. 분석 활용 전략

□ 기대효과

위의 결과에 따라 현재 여성 인구의 비율이 증가하는 정도가 갈수록 커지고 있는 현재의 상황을 고려하였을 때 광역시/도 단위로 도서관과 아파트의 수를 늘리는 것이 앞으로 2, 30대 1인 가구 여성 인구가 빠르게 늘어나는 것에 대비하는 데에 적합한 방안이라고 결론을 내릴 수 있다.

사회적 흐름에 따라 젊은 여성이 혼자 살면 1인 가구를 노리는 범죄에 취약해질 수 있다는 시선에서 여성도 혼자 사는 경험을 만드는 것이 차후에도 좋은 영향을 미칠 것이라는 시선으로 변화하고 있다. 이러한 흐름에 맞추어 전체 인구에 대한 2, 30대 1인 가구 여성 인구의 비율의 증가 추세가 더욱 빨라질 것으로 예상되기 때문에 본 분석을 통하여 사회적 흐름, 사회적 추세에 대비를 할 수 있다고 예상한다.

□ 방향제시

광역지방자치단체 단위가 아닌 시군구 또는 읍면동 단위로 분석이 가능하다고 했을 때 표본의 개수가 자연스럽게 늘어나고 군집화, 회귀분석의 질이 더욱 높아질 것으로 예상이 된다. 통계지리정보서비스에서 제공하는 데이터는 시군구 단위의 데이터는 시군구의 경계가 경량화되어 있고 읍면동 단위의 데이터는 결측치가 잦기 때문에 전국 단위의 정확한 분석이 힘들다고 판단하여 광역시/도 단위로 분석을 할 수밖에 없었음이 한계점이라고 볼 수 있다.

통계지리정보서비스에서 제공하는 데이터의 시계열이 일정하지 않음을 또 다른 한계점으로 들 수 있을 것이다. 2020년을 기준으로 제공되는 데이터가 더욱 다양했으면 2, 30대 1인 가구 여성 인구에 영향을 미치는 변수를 현재보다 더 다양한 방면에서 분석할 수 있었을 것이다.

참고문헌

신동화, 이세희, 송현주, & 서진욱(2018). 산점도 진단분석과 분할 변수 선택 기법을 활용한 점진적인 시각적 분석. 정보과학회논문지, 45(8), 801-806.

이경화(2004). 상관관계의 교수학적 변환에 관한 연구. 학교수학, 6(3), 251-266.

임성운(2014). 차원축소를 통한 K-평균 군집분석의 비교.