

|  |
| --- |
| 빅데이터 분석  결과 보고서 |
| MZ세대 1인 가구를 위한  거주지 추천 서비스 |
| 12월 21일  WhereHouse-2 (이재서, 정범진)  작성자 : 이재서 |



목차

1. 분석 배경 및 기획 의도
   1. 분석 배경
   2. 기획 의도
2. 분석 목표
3. 분석 데이터
   1. 데이터 수집
   2. 데이터 정합성 검정
   3. 데이터 전처리
4. 분석 과정 및 결과
   1. 검거 율 분석 과정
   2. 검거 율 분석 결과 및 시각화
5. 활용 방안
6. 참고문헌

1 분석 배경 및 기획 의도

* 1. 분석 배경

2022년 1인 가구는 전체 가구의 34.5% 인 750만 2천 가구로 가장 큰 비중을 차지했다. 2000년대 들어 크게 늘어나기 시작한 1인 가구의 수는 2000년도 15.5% 였던 것에 비하면 상당히 증가함을 알 수 있다.

통계청은 2023년 12월 12일 이러한 내용을 담은 ‘2023 통계로 보는 1인가구’ 에 대한 내용을 발표했는데, 이 보고서에 따르면 전체 3가구 중 1가구 이상이 1인 가구이며, 해가 지날수록 1인 가구가 증가하는 모양새를 보이고 있다.

또한 시도 간 이동한 인구 중, 서울의 유입이 유출보다 45,098가구가 많아 그 다음 유입이 더 많은 충남(3,888가구)보다 압도적인 수치를 자랑했다.

1인 가구가 늘어남에 따라 MZ세대들의 1인 가구 비율 또한 함께 증가했는데.

서울시 여성 가족 재단은 16년, 6월 한달 간 여성 1인가구 1천 800여명의 생활실태를 세대별로 조사하였고, 그 결과 2030세대 (MZ세대) 여성의 약 46%가 일상 생활이 안전하지 않다고 느꼈다고 한다. 이 수치는 2명중 한 명 꼴이며, 위급 상황에 도와줄 사람이 없다는 답변이 절반이었다.

정부에선 현재까지 여성 안심 귀가 서비스, 안심 비상벨, 안심 홈 세트 지원 등 증가하고 있는 1인 가구의 안전정책수요에 대응해 1 인 가구의 불안감을 해소할 수 있는 사업을 지속적으로 강화하고 있지만 아직까지 불안한 건 마찬가지다.

이 문제는 비단 여성만의 문제가 아니다.

경찰청은 23년 1월부터 8월까지 발생한 이상 동기 범죄 (일명 묻지마 범죄)는 23건으로 파악됐다며, 특히 신림역과 서현역 사건 등이 발생한 8월에서 10월 사이에는 5건이나 증가 했다고 보고했다.

* 1. 기획 의도

MZ세대를 위한 주거지 추천 서비스를 개발해 서울에 거주 혹은 이주 계획중인 MZ세대 들에게 주거 안전성 보장을 위한 정보를 제공한다.

사용자가 지정한 구에 대한 범죄 발생 횟수 및 검거 횟수, 시내 주요 기관 (경찰 · 소방관서) 의 수, 인구 수와 밀집 정도를 수집하여 분석하고 각 데이터마다의 상관 관계를 분석해 안전성 요소로써 사용한다.

안전성 점수는 사용자가 지정한 포인트의 좌표를 토대로 가장 가까운 파출소까지의 거리, 반경 500M내 CCTV의 수, 그리고 분석한 검거 율을 토대로 각각의 가중치를 적용해 제공한다. 이로 인해 사용자는 거주 혹은 이주 계획 시 더욱 안전한 장소를 비교, 고려할 수 있도록 도움을 제공한다.

2 분석 목표

- 검거 율을 분석하게 되면 안전성에 유의미한 영향을 줄 수 있다고 생각.

각 구의 범죄 발생 수와 검거 수를 가져 와 검거 율을 산출한다.

- 각 구의 안전과 관련된 요소들을 가져와 산출해 낸 검거 율과 분석한다.

1. 분석 데이터
   1. 데이터 수집

모든 데이터들은 일관된 분석을 위해 2022년을 기준으로 수집했으며, 모든 데이터들의 출처는 서울 열린 데이터 광장에서 수집하였다.

- 검거 율 데이터

검거 율 데이터로 각 구의 5대 범죄 (살인, 강도, 강간 · 강제추행, 절도, 폭력 ) 발생 및 검거 데이터를 사용 하였다. 각 범죄들을 합해 범죄 발생 소계와 검거 소계를 계산한 뒤

검거 소계를 범죄 발생 소계로 나누어 검거 율을 산출 하였다.

* 서울시 5대 범죄 발생 현황 통계 \_ 서울 열린 데이터 광장

- 인구 수와 밀집도 데이터

검거 율에 영향을 미치는 요인 중에는 시민들의 협조도 있을 것이라고 판단하여 조사 하였다. 인구 수만이 검거 율에 영향을 미칠 것이라 생각했지만 인구의 밀도 또한 영향을 미친다는 것을 알게 되었고, 두 개 모두 데이터 수집을 하였다.

* 서울시 인구밀도 (구별) 통계 \_ 서울 열린 데이터 광장

- 시내 주요 기관 (경찰 · 소방관서) 데이터

시내 주요 기관 데이터는 각 구의 경찰청 수, 경찰서 수, 치안센터 (지구대 · 파출소)의 수, 소방본부 수, 소방서 수, 119 안전 센터의 수를 사용 하였다. 각각의 데이터들에 비중을 따로 두지 않고 시내 주요 기관으로 묶어 소계로 사용 하였다.

* 서울시 시내주요기관 (경찰 · 소방관서) 통계 \_ 서울 열린 데이터 광장

-서울시 CCTV 데이터

분석 전 계획을 수립할 때 CCTV가 많으면 많을 수록 검거 율에 영향을 미치는 정도가 클 것이라 판단 했다.

* 서울시 자치구 연도별 CCTV 설치 현황 \_ 서울 열린 데이터 광장
  1. 데이터 정합성 검정

향후 분석을 위해 각 데이터의 정합성을 검사한 결과, 몇 가지 문제점들이 발견 되었다.

- 모든 데이터들에 , (콤마) 가 붙어 데이터를 불러올 때, 숫자가 아닌 문자열로 불러와졌다.

- 검거 율과 관련된 데이터에서는 몇몇 구에 특정 범죄에 대해 일부 결측 치가 발견 되었다.

- 시내 주요 기관과 관련된 데이터에서는 경찰청과 소방 본부에 관해 거의 대부분의 구가 결측 치를 가지고 있었다.

* 1. 데이터 전처리

조사한 데이터를 분석에 사용하기 위해서는 데이터 전 처리와 정리 과정을 거쳐야 했다.

조사한 모든 데이터들을 먼저 DataFrame 형태로 정리하였다. 모든 데이터들에 대해 적용되어 총 3개의 DataFrame이 만들어 졌다.

필요한 데이터는 소계를 사용하므로 각 Column 들을 더할 필요가 있었고

문자열로 되어 있는 데이터들을 모두 as.numeric을 통해 숫자로 변환하였다.

그 과정 중 결측 치를 가지고 있던 데이터들이 NA로 변환 되었고

각각의 벡터를 결측 값이 있다면 0으로 치환 해 데이터 프레임 전체를 전처리 하였다.

서울시CCTV 데이터의 23년의 데이터가 일부 결측 치가 발견 되었지만 우리가 필요한 것은 22년의 CCTV 수이고 그 이전의 연도까지의 설치 현황에선 결측 치가 발견되지 않아 23년도의 데이터는 삭제하고 진행하였다.

1. 분석 과정 및 결과
   1. 검거 율 분석 과정

본 프로젝트는 검거 율과 시내 주요 기관, 인구수, 그리고 인구 밀집 정도의 상관 관계를 분석 하기 위해 다중 선형 회귀 분석 (Multiple Linear Regression)을 사용한다.

회귀 분석 모델을 수식 화 한 후, 가장 우수한 회귀 모델을 판별한다.

종속 변수로 둘 검거 율을 수집하기 위해 서울 열린 데이터 광장에서 수집한 서울시 5대 범죄 발생 현황 통계를 불러와 5대 범죄의 발생 건 수 소계와 5대 범죄의 검거 수 의 소계를 계산했다. 5대 범죄의 검거 수 소계를 5대 범죄의 발생 건 수 소계로 나눠 검거 율을 구했다.

독립 변수로 둘 시내 주요 기관 또한 서울 열린 데이터 광장에서 데이터를 수집해 각 자치구 마다 소계를 구했다.

계산해 논 각 자치구의 검거 율을 분석하기 위해 여러 독립 변수 (시내 주요 기관의 수, 인구수, 인구 밀집 정도, CCTV 수)들을 고려하여 모델을 수식 화 하였다.

* 1. 검거 율 분석 결과 및 시각화

|  |  |  |  |
| --- | --- | --- | --- |
| 설명 변수 | B | t | p |
| Constant | 8.246e-01 | 10.296 | 3.28e-09 |
| 시내 주요 기관 수 | 7.659e-03 | 2.730 | 0.0133 |
| 인구수 | -2.916e-07 | -2.403 | 0.0267 |
| 인구 밀집 정도 | -6.886e-06 | -2.584 | 0.0182 |
| CCTV수 | -1.190e-05 | -1.075 | 0.2958 |
| Adjusted R² | 0.5135 | | |
| p - value | 0.001159 | | |

Note : \*p<0.05, \*\*p<0.01, \*\*\*p<0.001

B : 가중치, t : 가중치/표준 오차, p : 유의 수준, p – value : 유의 수준,

Adjusted R² : 조정된 결정 계수

[표 1-1] 검거 율 상관 관계 분석

[표 1-1]의 분석 결과를 단계적 선택법 StepAIC(model, direction = “both”) 을 사용 하여 변수를 다시 선택 하였다.

|  |  |  |  |
| --- | --- | --- | --- |
| 설명 변수 | B | t | p |
| Constant | 8.104e-01 | 10.219 | 2.19e-09 |
| 시내 주요 기관 수 | 6.977e-03 | 2.543 | 0.01936 |
| 인구 수 | -3.374e-07 | -2.958 | 0.00777 |
| 인구 밀집 정도 | -6.682e-06 | -2.504 | 0.02107 |
| Adjusted R² | 0.5097 | | |
| p - value | 0.0005745 | | |

[표 1-2] 단계적 선택법을 사용한 검거 율 상관 관계 분석

단계적 선택법을 사용하여 변수 선택 후 결과를 확인해보니 시내 주요 기관 수, 인구 수, 인구 밀집 정도가 독립 변수로 선택 되었다.

위 모델에서 얻은 회귀 식 결과는 다음과 같다.

# 검거 율 = (8.104e-01)+(6.977e-03\* 시내 주요 기관 수)

+(-3.374e-07\*인구 수)+(-6.682e-06\*인구 밀집 정도)

p – value : 0.0005745

Adjusted R² : 0.5097 -> 50.97%

회귀 분석에서 p – value 값은 구한 회귀 모델이 의미 있는 모델인지를 나타내는 것으로, 이 값이 작을수록 의미 있는 모델인 것을 나타낸다. 평균적으로 p-value 값이 유의 수준인 0.05보다 작으면 해당 독립 변수는 종속 변수에 통계적으로 유의적인 영향을 미치는 것으로 간주된다.

Adjusted R² 값은 모델의 설명력을 나타내며, 0 ~ 1 사이의 값을 가진다. 이 값이 클수록 회귀 모델이 오차가 적다는 의미를 나타낸다.

결론적으로 p – value 의 값은 유의미하게 도출 되었지만 Adjusted R² 의 값에 의해 약간의 오차가 존재한다고 보여진다.

분석 결과, 독립 변수로 선정한 시내 주요 기관은 수가 많아질수록 검거 율에 유의미한 영향을 미치는 것으로 보인다.

하지만 내 생각과는 다르게 CCTV와 인구 수의 경우 반대로 흘렀는데,

CCTV의 경우, 범죄가 많이 발생한 곳을 기준으로 설치되는 후행 성 지표이다 보니 검거 율 증가에는 영향이 없었고, 회귀 식을 설명하는데 오히려 방해만 될 뿐 이었다.

또한 인구 수의 경우 사람이 많은 지역에서 범죄 발생 건수도 함께 높게 나오다 보니

사람이 많다는 건 검거에 유의미한 영향을 미치지만 그보다 더 범죄 발생이 많이 늘어나므로 단순히 검거 율에 미치는 영향만 계산하기에는 복잡한 문제가 발생했다.

우리가 만드는 프로젝트에서 상관 관계 분석을 통한 검거 율 회귀 식은 각 지역의 검거 율을 100% 완벽하게 예측 하진 못하지만, 참고 용으로 사용이 가능한 수준이며,

우리의 프로젝트는 사용자가 지정한 핀 포인트를 기반으로 동작하므로 안전성 점수의 한 요소로 사용할 것이다.

1. 활용 방안

현재 서울로 거주 혹은 이주 계획중인 MZ세대 들에게 특정 지역이 왜 안심이 되는 가를 상세히 설명해줄 수 있다. 그로 인해 선호 하는 장소 물색에 유의미한 도움을 줄 수 있을 것이다. 또한 자치구 혹은 서울시 전체에서 어느 장소가 안전 하거나 그렇지 못하거나 참고하는 용도로도 사용할 수 있을 것이다. 특정 좌표를 통해 근처 CCTV와 파출소까지의 거리를 고려하고 검거 율과 비교한다면 시민들을 위한 정책 마련에 도움이 될 것이다.

1. 참고 문헌

- 통계청 - 2023 통계로 보는 1인가구

- 서울 열린 데이터 광장

- e-나라 지표

- 서울시 여성 가족 재단 – 세대별 여성 생활 만족도 조사