

# 서울시 1인 가구가 살기 좋은 동네 찾기

작성자: 전현아

## ▼ 목차

### 1인 가구 통계

[왜 1인 가구가 되었을까?](#)

[1인 가구 주거 유형은?](#)

[데이터 해석 및 프로젝트 계기](#)

### 서울시에서 1인 가구가 살기 좋은 동네는?

[서울 1인 가구의 안전 만족도는?](#)

[서울시 어느 구에 1인 가구 많이 살까?](#)

[서울시 5대 범죄 발생 수는?](#)

[서울시에 cctv가 몇 개 있을까?](#)

[서울시에 유흥업소는 몇 개 있을까?](#)

### 가설 검증

[인구수와 범죄 발생 간의 관계](#)

[범죄 발생과 CCTV 간의 관계](#)

[유흥업소와 범죄 발생 간의 관계](#)

[인사이트 도출](#)

### 예측모델

[데이터 설명 및 정제](#)

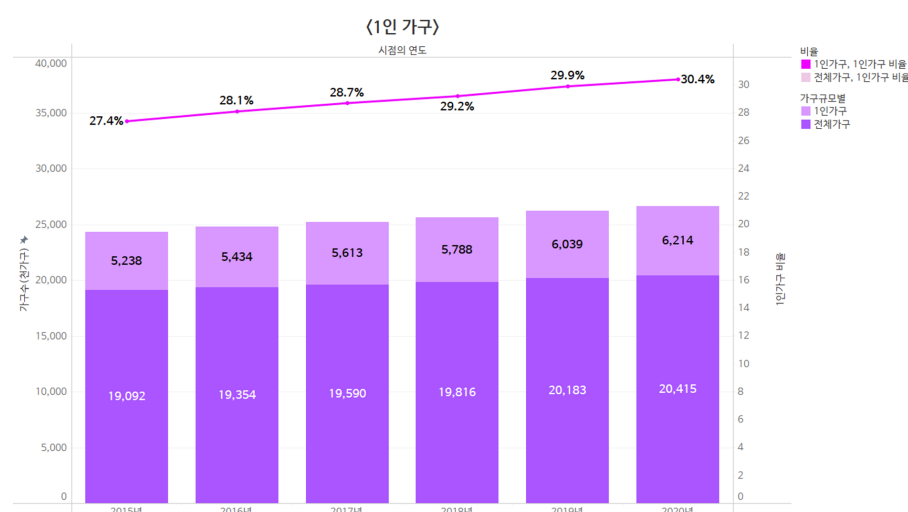
[예측모델](#)

[테스트 결과](#)

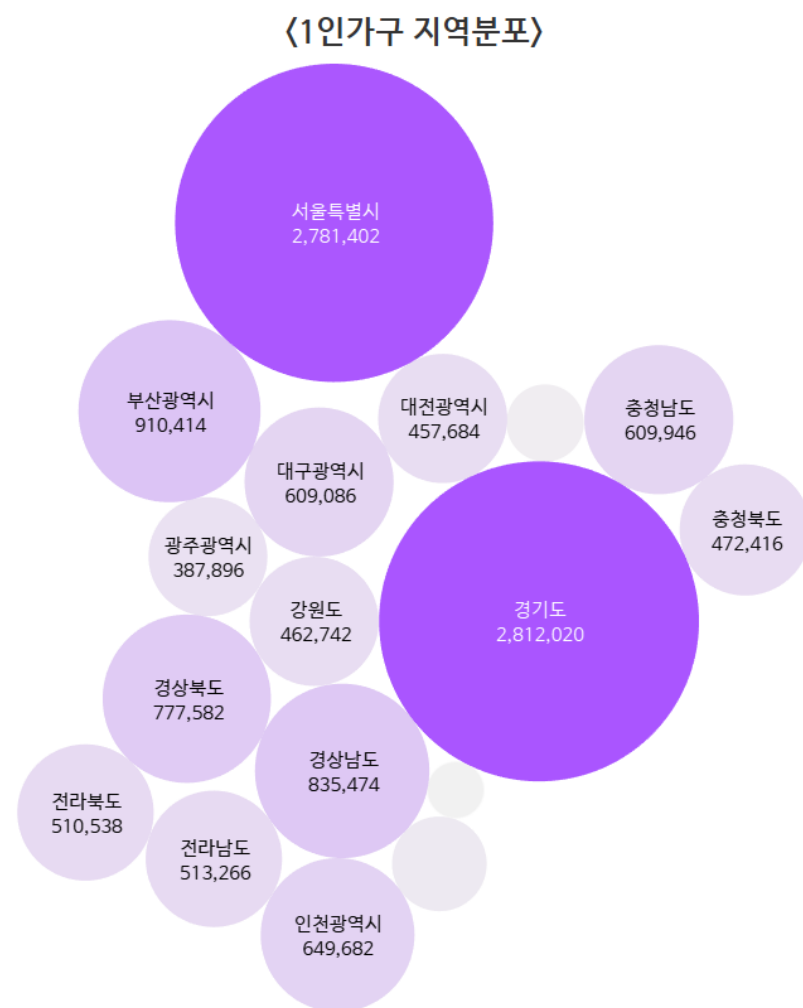
[최종 모델](#)

### 프로젝트 회고

## 1인 가구 통계

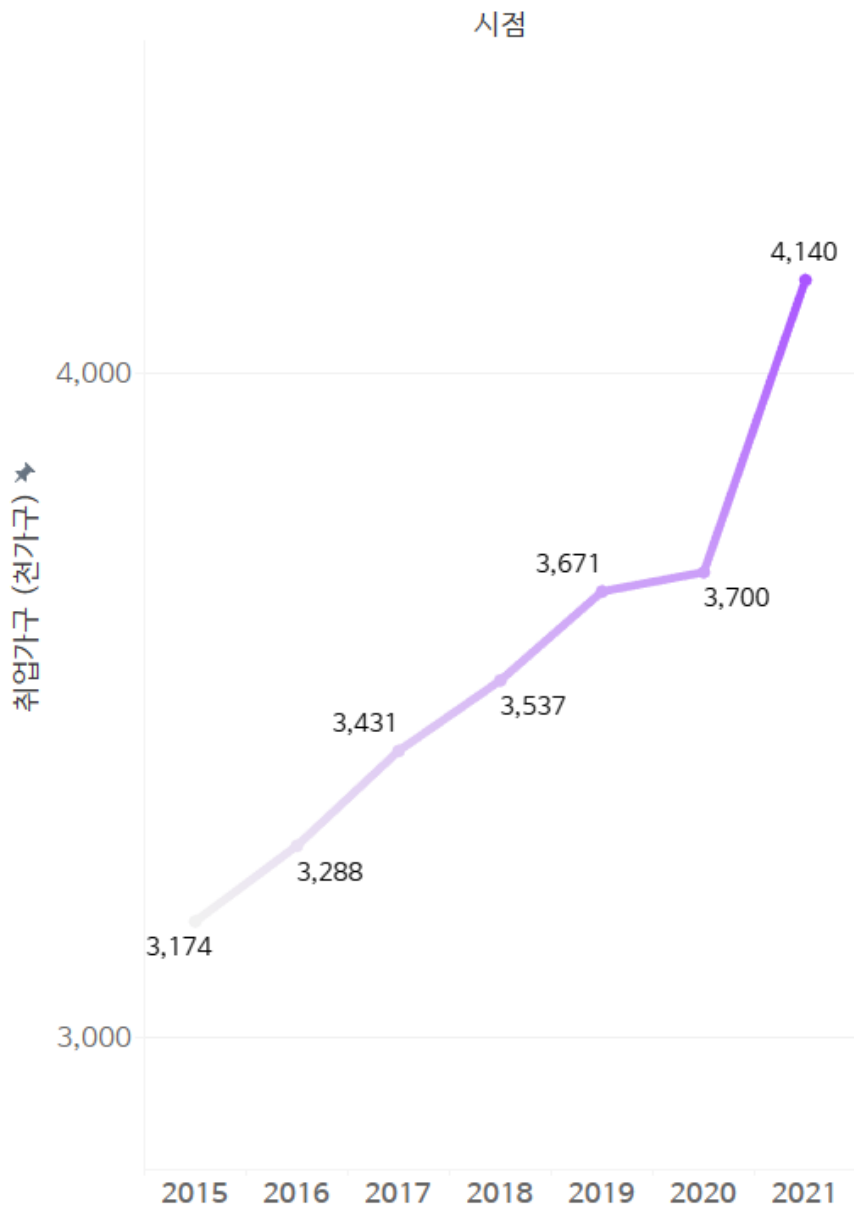


1인 가구가 2015년부터 점점 증가하고 있고 2020년에는 전체 인구의 **30.4%**를 차지하고 있다.



(2020년 기준) 1인 가구는 **경기도, 서울특별시, 부산광역시** 순으로 많았다. 특히 **수도권**에서 많이 몰려있는 것을 확인할 수 있다.

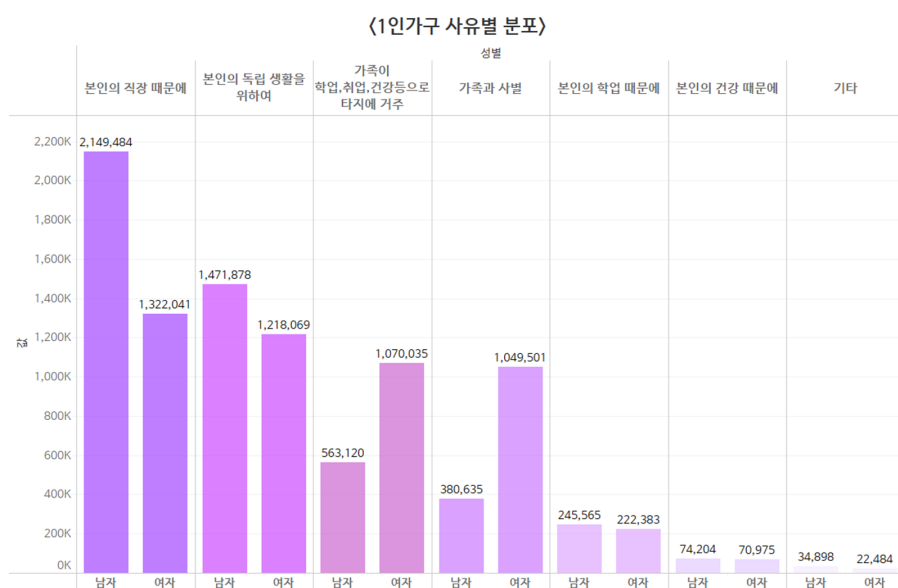
## 〈취업 1인가구〉



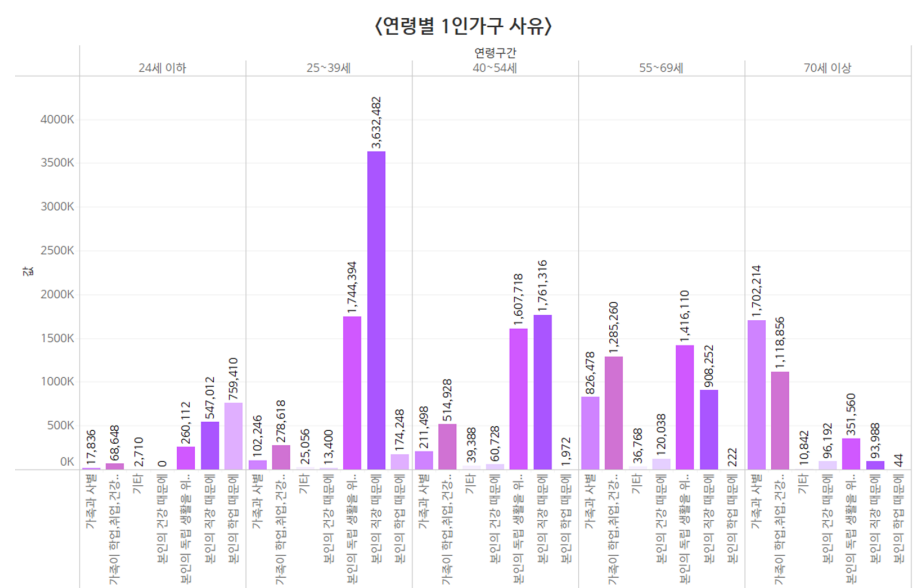
(취업 1인가구) 1인 가구가 많은 이유를 일단 취업 1인 가구를 보면 알 수 있다. 2015년부터 현재까지 꾸준히 상승하는 것을 볼 수 있다. 취업을 다른 지역으로 했기 때문에 1인 가구가 되었다는 것을 파악할 수 있다. **대부분의 일자리가 수도권에 분포해 있어** 다른 지역에서 수도권으로 많이 이동한다.

## 왜 1인 가구가 되었을까?

### • 성별



### • 연령별

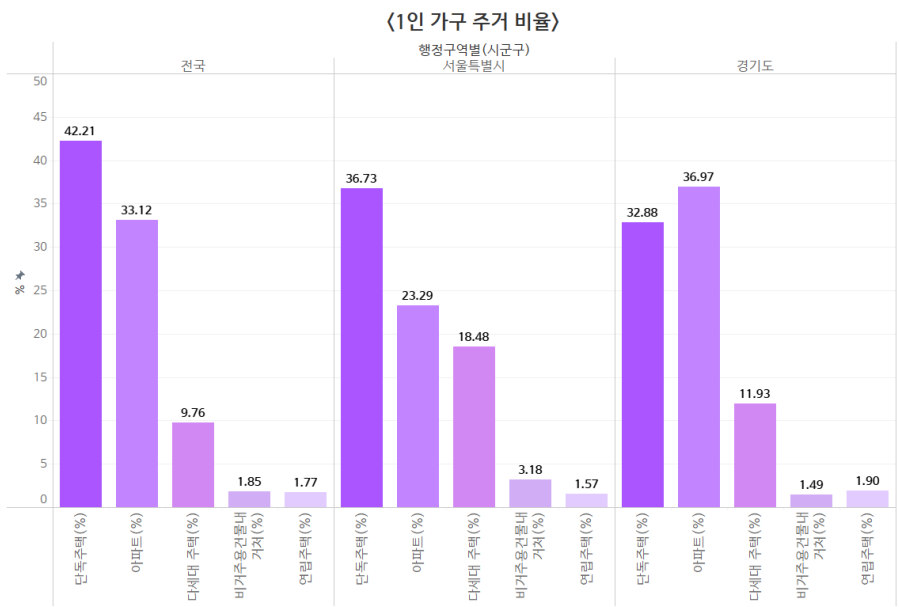


## (2020년 전국 1인 가구 통계)

- 2020년 기준 1인 가구의 주된 사유는 **본인 직장, 본인 독립** 순이다.
- 남자는 본인 사유**가 높게 나왔고 **여자는 본인, 가족 사유**가 비슷한 비율로 나온 것을 볼 수 있다.

- 연령별로 봤을 때, **25~54세는 본인 직장, 독립 사유**가 높게 나왔고, **60대 이후는 가족 관련 사유**가 높아지고 있다.
- 이것을 토대로 위의 취업 가구가 증가하는 이유를 뒷받침할 수 있다.

1인 가구 주거 유형은?



(2021년 기준)

- 전국 1인 가구 주거 비율은 **단독 주택이 42.21%, 아파트가 33.12%** 순으로 높다.
- 주로 1인 가구가 사는 **원룸, 빌라, 오피스텔** 등은 단독 주택에 포함된다.
- 1인 가구 비율이 높은 **수도권**을 봤을 때 **서울은 단독주택, 경기도는 아파트와 단독주택**에 많이 거주하는 것을 알 수 있다.

데이터 해석 및 프로젝트 계기

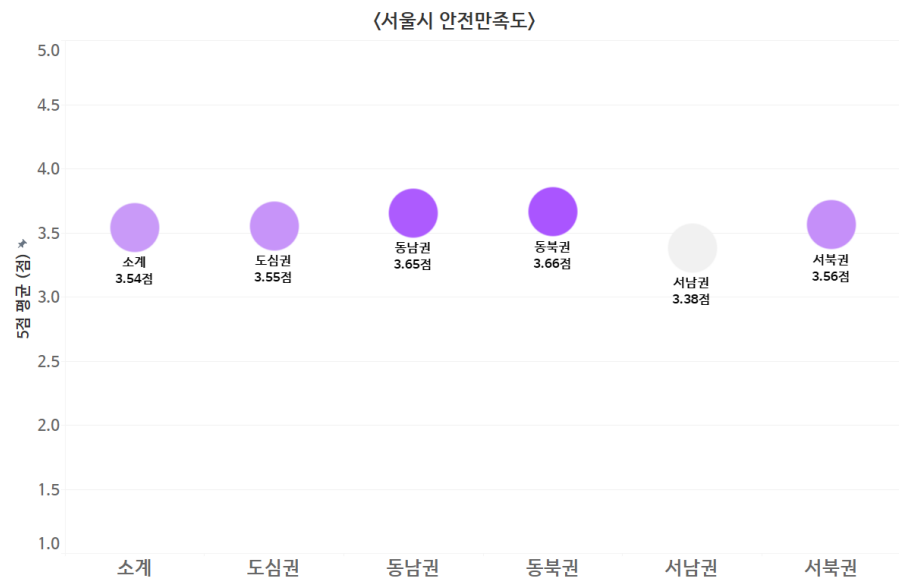
위의 데이터 분석을 통해 1인 가구는 **수도권**에 많이 분포해있고 1인 가구 사유는 **본인의 직장, 본인의 독립**이고 주로 **단독 주택**에 거주하고 있다.

수도권 중 특히 **서울**은 면적에 비해 1인 가구가 많이 분포해있다. **20-40대 1인 가구**는 첫 독립일 확률이 높기 때문에 어떤 곳이 좋고 월세가 어느 정도인지 알 수 없을 것이다. 그리고 1인 가구가 직장 주변에서 월세를 구하기에는 비싸기도 하고 치안이 어떤지 몰라 불안하다. 그래서 데이터 분석을 통해 **서울시에서 1인 가구가 살기 좋은 동네**를 확인하려고 한다.

또한 1인 가구가 가장 많이 주거하는 **원룸, 빌라, 오피스텔의 월세**를 예측하고 싶었다. 그래서 **직방 데이터를 크롤링해 서울시 행정구역별 월세를 예측하는 모델**을 만들었다.

서울시에서 1인 가구가 살기 좋은 동네는?

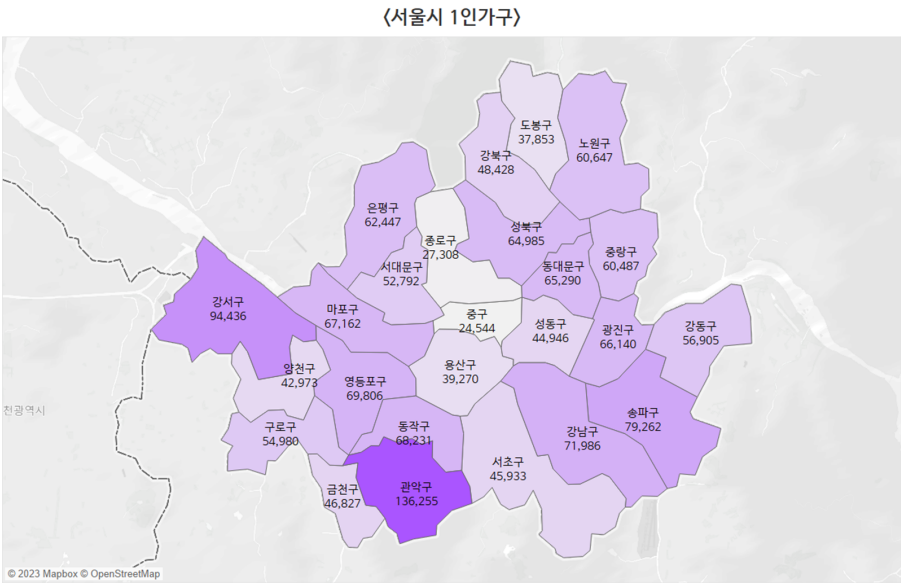
서울 1인 가구의 안전 만족도는?



(2020년 기준)

- 서울시 안전 만족도 평균은 3.54점이다.
- 동북권**(성동구, 광진구, 동대문구, 중랑구, 성북구, 강북구, 도봉구, 노원구)은 **3.66점**으로 가장 높고 **동남권(3.65점)** 순으로 높다.
- 안전만족도가 가장 낮은 곳은 **3.38점**을 얻은 **서남권**(성동구, 광진구, 동대문구, 중랑구, 성북구, 강북구, 도봉구, 노원구)이다.

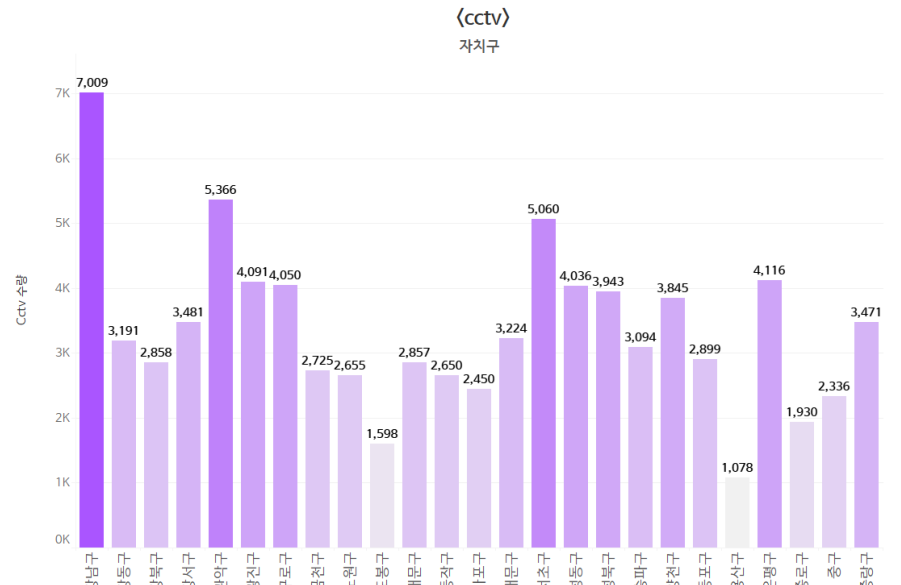
서울시 어느 구에 1인 가구 많이 살까?



(2021년 기준)

- 서울시에서 1인 가구가 **관악구, 강서구, 송파구, 강남구** 순으로 많이 거주하는 것으로 보인다.
- 서울시에 사는 1인 가구 **남녀 비율**은 남자가 약 47%, **여자 약 53%**로 더 많은 것을 알 수 있다.

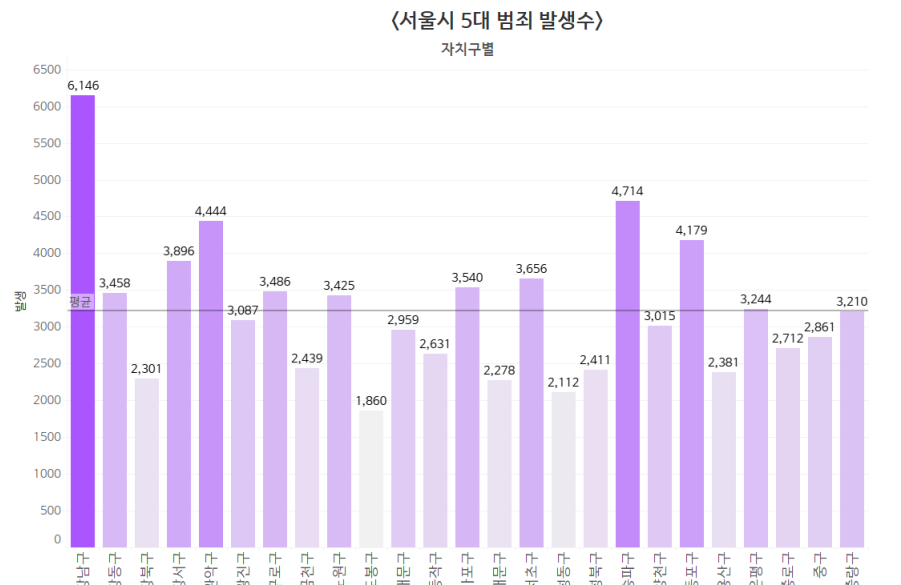
서울시에 cctv가 몇 개 있을까?



(2022년 기준)

- 총 서울시 cctv 갯수는 84,013개다.
- 그 중 **강남구**에 **7009개**, 그 다음으로 **관악구(5336개)**, **서초구(5060개)** 순으로 cctv가 많다. 인구가 많이 거주하거나, 범죄가 많이 일어나기 때문에 cctv가 많다고 해석할 수 있다.

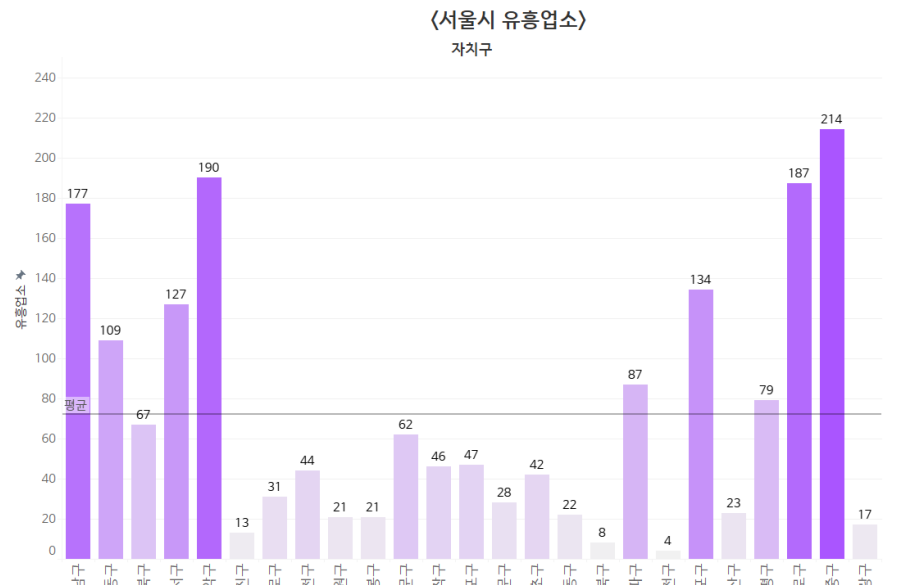
서울시 5대 범죄 발생 수는?



(2021년 기준)

- 5대 범죄는 **살인, 강도, 강간/강제추행, 절도, 폭력**이다.
- 범죄 발생 평균은 약 **3217.8건**이고 범죄가 많이 일어난 곳은 **강남구, 송파구, 관악구, 영등포구** 순이다.
- 적게 일어난 곳은 **도봉구, 성동구, 강북구** 순이다.

서울시에 유흥업소는 몇 개 있을까?

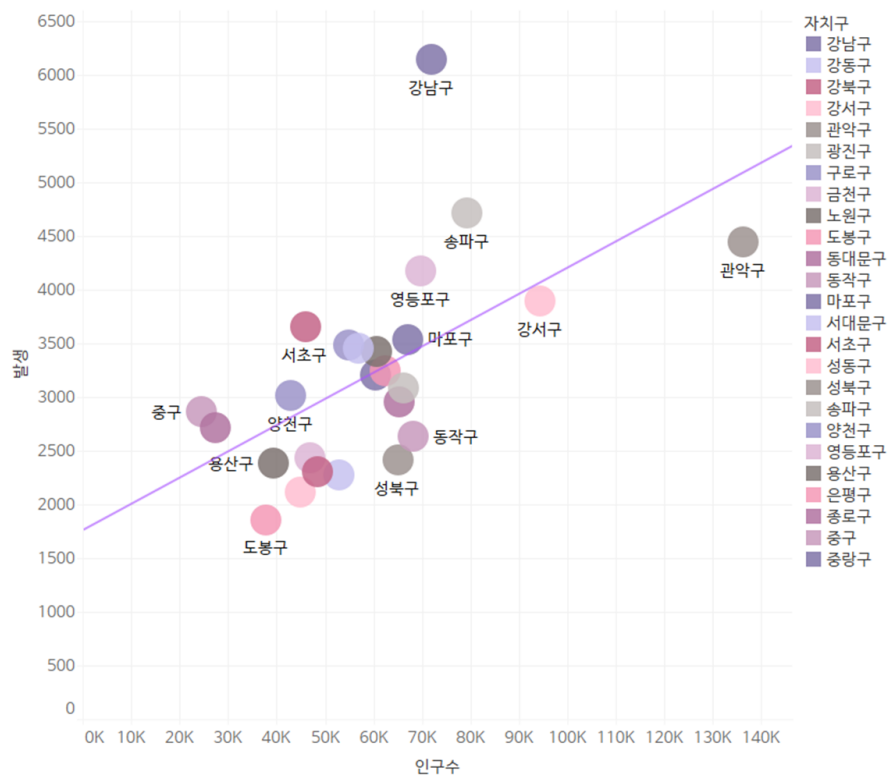


- 서울시에서 등록 되어있는 유흥업소 수는 총 **1800개**이다.
- 그중 유흥업소가 많은 곳은 **중구, 관악구, 종로구, 강남구** 순이다.
- 적은 곳은 **양천구, 성북구, 관진구, 중랑구** 순이다.

가설 검증

인구수와 범죄 발생 간의 관계

〈인구수와 범죄발생 간의 관계〉



선형 추세 모델

모델 수식: ( 인구수+절편 )  
모델링된 관측값 수: 25  
필터링된 관측값 수: 0  
모델 자유도: 2  
잔차 자유도(DF): 23  
SSE(오차제곱합): 1.44893e+07  
MSE(평균 제곱 오차): 629969  
R-제곱: 0.33444  
표준 오차: 793.706  
p-값(유의): **0.0024608**

• 범죄 발생 = 0.0244509\*인구수 + 1760.63

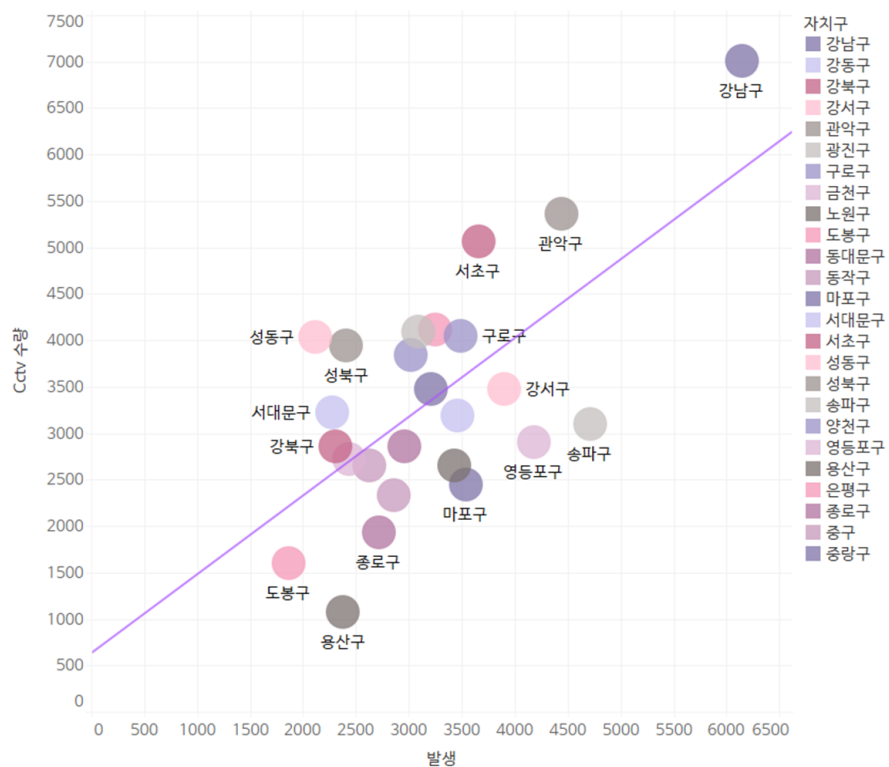
p-값이 0.00246으로 유의수준 0.05보다 작으므로 범죄 발생 수와 인구 수 간의 관계가 유의하다고 볼 수 있고 두 변수 간 상호 설명 능력을 나타내는 R-제곱이 0.33으로 중간의 설명력을 나타낸다.

또한 두 변수의 상관 관계를 봤을 때, 0.578308로 중간의 상관 관계가 존재하는 것을 알 수 있다.

즉, 인구수가 많을수록 범죄 발생도 많아진다.

범죄 발생과 CCTV 간의 관계

〈범죄발생과 CCTV 간의 관계〉



선형 추세 모델

모델 수식: ( 발생+절편 )  
모델링된 관측값 수: 25  
필터링된 관측값 수: 0  
모델 자유도: 2  
잔차 자유도(DF): 23  
SSE(오차제곱합): 2.16854e+07  
MSE(평균 제곱 오차): 942845  
R-제곱: 0.419302  
표준 오차: 971.002  
p-값(유의): **0.0004666**

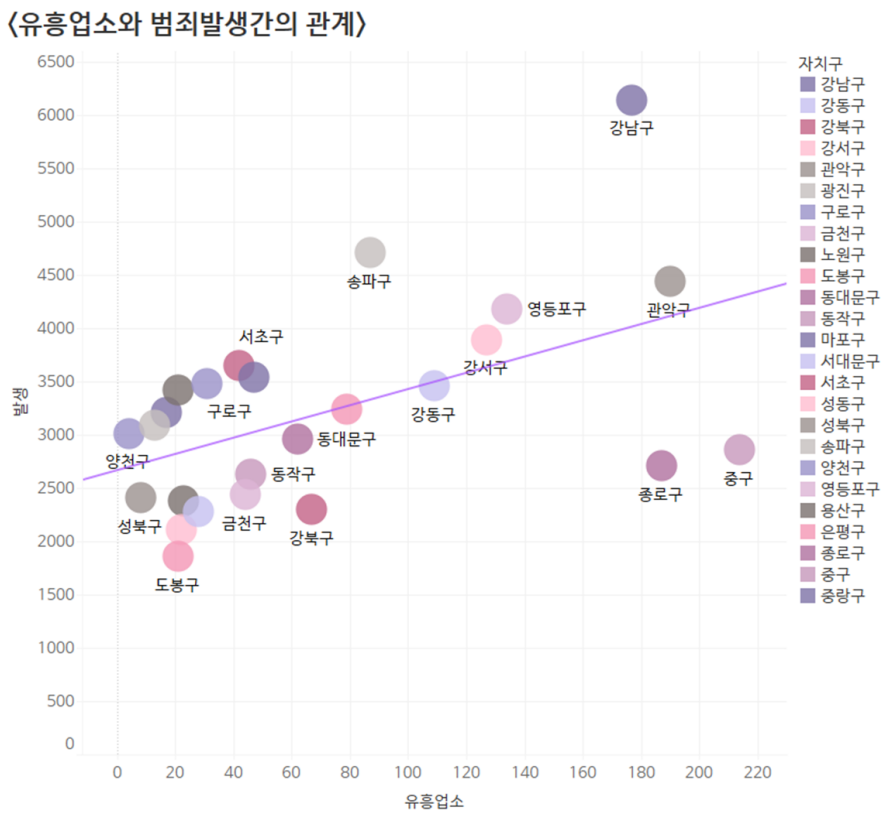
• CCTV수량 = 0.848091 \* 범죄 발생 + 631.534


p-값이 0.0004666으로 유의수준 0.05보다 작으므로 범죄 발생 수와 CCTV 수 간의 관계가 유의하다고 볼 수 있고 두 변수 간 상호 설명 능력을 나타내는 R-제곱이 0.42으로 중간의 설명력을 나타낸다.

또한 두 변수의 상관 관계를 봤을 때, 0.647535로 높은 상관 관계가 존재하는 것을 알 수 있다.

즉, 범죄 발생 수가 많을 수록 CCTV수량도 많아진다.

유흥업소와 범죄 발생 간의 관계



**선형 추세 모델**

모델 수식: ( 유흥업소+절편 )

모델링된 관측값 수: 25

필터링된 관측값 수: 0

모델 자유도: 2

잔차 자유도(DF): 23

SSE(오차제곱합): 1.60272e+07

MSE(평균 제곱 오차): 696833

R-제곱: 0.263799

표준 오차: 834.765

p-값(유의): **0.0086376**

• **범죄 발생 = 7.62 \* 유흥업소 + 2669.16**

p-값이 **0.0086376**으로 유의수준 0.05보다 작으므로 **범죄 발생 수와 유흥업소 수 간의 관계가 유의하다**고 볼 수 있고 두 변수 간 상호 설명 능력을 나타내는 R-제곱이 0.26으로 약한 설명력을 나타낸다.

또한 두 변수의 **상관 관계**를 봤을 때, **0.513614**로 중간의 상관 관계가 존재하는 것을 알 수 있다.

즉, 유흥업소가 많을 수록 범죄 발생 수가 많아진다.

인사이트 도출

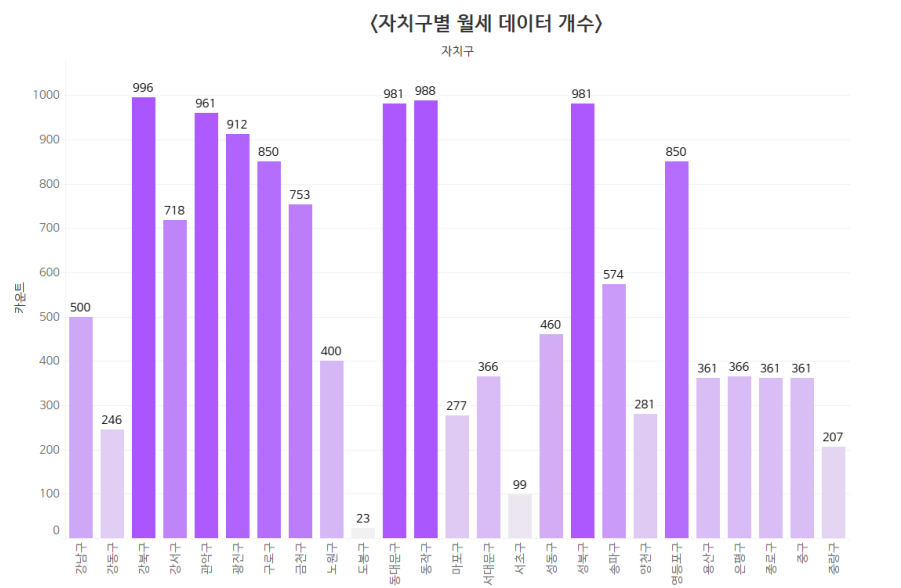
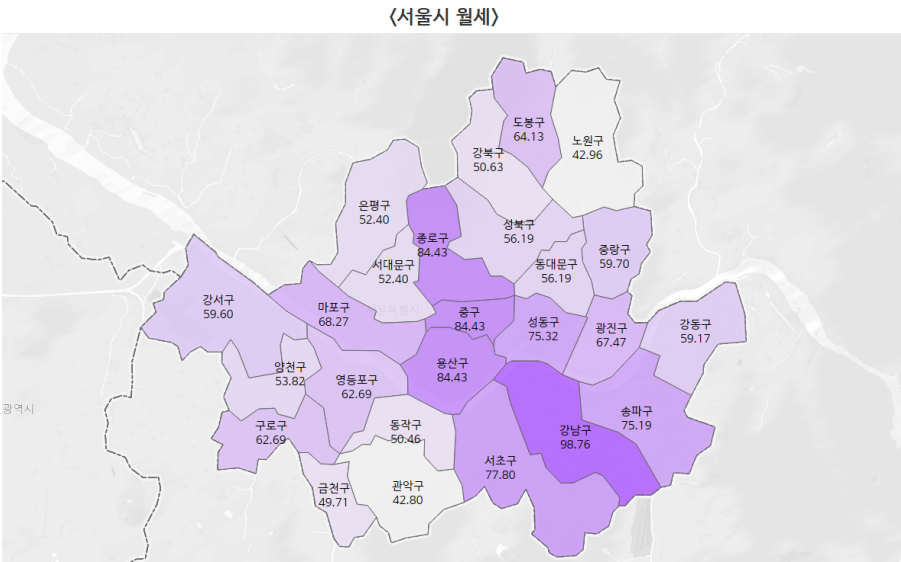
위의 결과를 토대로 서울시에서 살기 좋은 동네는 **동북권**을 추천한다.

동북권 중 **성동구, 강북구, 도봉구, 광진구**는 범죄 발생 수가 낮고 유흥업소도 적은 것으로 나타난다.

그리고 조심해야 하는 동네는 **강남구, 관악구, 양천구** 등이 있다.

예측모델

데이터 설명 및 정제



- **직방**에서 서울시 원룸, 빌라, 오피스텔 월세를 **크롤링**했다.



- 자치구마다 1000개 썩 월세를 모으고 싶었지만, 그만큼 존재하지 않는 동네가 있어 자치구 비율이 차이가 난다.
- 가장 월세가 비싼 동네는 **강남구, 종로구, 중구, 용산구**이다.

	자치구	보증금	월세	평수(㎡)	건물층	매물층	분류	평수(평)
0	종로구	1000	60	39.67	2	2	고	12.0
1	종로구	500	35	30.28	2	반지하	반지하	9.0
2	종로구	500	40	19.83	3	2	고	6.0
3	종로구	500	40	23.50	3	1	저	7.0
4	종로구	500	38	29.75	2	반지하	반지하	9.0
...	...	...	...	...	...	...	...	...
14341	송파구	15000	50	37.43	5	2	저	11.0
14342	송파구	5000	70	62.81	4	3	고	19.0
14343	송파구	4000	37	26.45	5	4	고	8.0
14344	송파구	500	30	23.14	3	반지하	반지하	7.0
14345	송파구	120	40	13.22	5	2	저	4.0

14346 rows × 8 columns

- 평수(㎡)을 3.3058로 나눠 round함수를 사용해 소수점은 반 올림해준 **평수(평)** 컬럼을 만들어줬다.
- 또한 건물층, 매물층을 기준으로 **반지하, 저, 중, 고, 옥탑방**으로 나눠줬다.

	보증금	월세	평수(㎡)	건물층	평수(평)
count	14346.000000	14346.000000	14346.000000	14346.000000	14346.000000
mean	3419.915517	63.174265	33.347456	7.200753	10.090966
std	5193.187627	45.911663	20.857093	5.233494	6.316671
min	1.000000	1.000000	4.000000	1.000000	1.000000
25%	500.000000	38.000000	19.830000	4.000000	6.000000
50%	1000.000000	50.000000	26.440000	5.000000	8.000000
75%	4500.000000	80.000000	42.900000	9.000000	13.000000
max	70000.000000	1250.000000	489.260000	46.000000	148.000000

- 데이터를 보면 보증금이 만원(단기임대)~7억원(반전세/전세), 월세는 만원~1250만원, 평수는 148평까지 존재한다.
- 그래서 **보증금은 50만원~2억, 월세는 250만원 이하, 평수는 30 평이하**로 기준을 정해두었다.

- 총 **13,872개**의 데이터를 사용했고 예측에 사용할 feature는 **자치구, 보증금, 분류, 평수(평)** 이다.
- 월세를 예측하려고 회귀분석을 사용하려 했지만 원핫인코딩으로 자치구가 25개로 나뉘어지기 때문에 컬럼이 너무 많아지기 때문에 적합하지 않다.
- 오디널인코딩은 숫자로 나뉘져 대소관계가 생기기 때문에 선형, 로지스틱 회귀모델에 사용하는 것은 적절하지 않다.
- 하지만 트리 기반 모델에서는 여러 번의 분기를 통해 양적 대소 관계가 점차 사라지기 때문에 오디널 인코딩을 사용할 수 있다.
- 따라서 **회귀트리**를 사용해 마지막 노드에 있는 타겟값들의 평균을 예측값으로 반환시켜준다.

## 예측모델

model	R2	rmse	mae
<b>base</b>		<b>34.621</b>	<b>25.990</b>
RandomForestRegressor	0.641	20.727	13.616
DecisionTreeRegressor	0.471	25.174	15.323
GradientBoostingRegressor	0.636	20.867	14.389
XGBRegressor	0.636	20.863	14.387

LGBMRegressor	0.652	20.421	14.018
randomcv (RandomForestRegressor)	0.651	20.445	13.525
<b>randomcv (LGBM)</b>	<b>0.714</b>	<b>19.907</b>	<b>13.625</b>

- 총 5개의 회귀트리모델을 돌렸고 과적합을 방지하기 위해 **K-Fold 교차검증**을 사용했다.
- 평가 지표인 R2가 높고 RMSE와 MAE가 작은 **RandomForest**와 **LightGBM**을 **RandomSearchCV**를 이용해 최적의 파라미터를 찾아내 다시 k-fold 교차검증을 사용해 결과를 확인해봤다.
- **LightGBM**이 **시간이 더 빠르고 가벼운 모델**이고 R2도 높게 나왔기 때문에 LightGBM을 선택했다.

## 테스트 결과

Weight	Feature
0.8106 ± 0.0263	평수(평)
0.3282 ± 0.0141	보증금
0.0293 ± 0.0089	자치구
-0.0001 ± 0.0146	분류

- **LightGBM Permutation Importance**

모델의 특성중요도를 확인해본 결과 **평수(평)**, **보증금**이 예측값에 영향을 많이 끼친다.

- ‘자치구의 수가 많아 특성중요도가 높지 않은 것인가’ 해서 자치구를 행정구역별로 5개로 나눠 다시 모델 예측을 해봤지만 비슷한 결과가 나와 원래의 예측모델을 선택했다.

- **test 결과**

<b>R2</b>	0.678
<b>RMSE</b>	20.582
<b>MAE</b>	14.035


test셋 결과를 봤을 때, train셋과 큰 차이가 없는 것을 보면 교차검증을 통해 과적합을 방지한 것을 볼 수 있다.

## 최종 모델

<b>R2</b>	0.709
<b>RMSE</b>	18.912
<b>MAE</b>	13.047

최종 모델을 만들기 위해 train, test 데이터를 합쳐 가장 성능이 좋았던 LGBM모델에 넣어 다시 학습을 진행했다.

- **자치구(오디널인코더)**



강북구(1) 동대문구(2) 서대문구(3) 광진구(4) 송파구(5) 양천구(6) 금천구(7)  
은평구(8) 동작구(9) 구로구(10) 성동구(11) 강서구(12) 노원구(13) 영등포구(14)  
성북구(15) 용산구(16) 중구(17) 중랑구(18) 종로구(19) 관악구(20)  
강남구(21) 마포구(22) 강동구(23) 도봉구(24) 서초구(25)

- **분류(오디널인코더)**





고(1) 반지하(2) 중(3) 옥탑방(4) 저(5)

- **실제로 예측해보기**

- **주의점:** 트리모델이기 때문에 **외삽이 불가능**하다.
- **보증금(만원) 50~20000, 월세(만원) 1~250, 평수 1~30** 안으로 설정해야한다.

- 만약, 성북구(15)에서 보증금이 500만원, 원하는 층은 고층(1), 8평을 원한다면

```
cv_lgb.predict([[{'자치구':15, '보증금':500, '분류':1, '평수(평)':8}])
```

- 모델은 **월세 45.55760199만원을** 예측했다.
- 이걸 토대로 원하는 위치, 평수, 층수로 월세를 예측할 수 있길 바란다.

## 프로젝트 회고

예측 모델의 성능이 0.7로 적당한 성능을 냈다. 하지만 나의 계획은 **자치구에 따라 월세 차이가 나길 바랬지만** 특성 중요도에서 낮은 점수를 얻어 아쉬웠다. 아마 월세가 비싼 몇 개의 구를 제외한 나머지 구들은 비슷한 월세를 가지고 있어서 그런 것 같다. 다음엔 각각의 자치구에 대한 월세 예측 모델을 만들면 좋을 것 같다.

이번 프로젝트를 하면서 까먹었던 section 1,2를 다시 공부해서 좋았고 당시 부족했던 지식들을 다시 돌아볼 수 있었고 조금은 더 발전했다. 그리고 처음 태블로를 써봤지만 시각화를 직접 하면서 더 많이 배울 수 있었고 더 배워서 멋진 대시보드를 만들고 싶다. 그리고 직접 사이트를 구현해보고 싶은 생각이 들어 section3에서 배운 페이지 만들기도 도전해볼 것이다.

많은 오류도 만나고 어떻게 해결해야 할지 막막했지만 그래도 포기하지 않고 끝까지 문제를 해결하려고 노력했다. 모르는 것이 있다면 동기들에게 물어보고 구글링도 열심히 했다. 예전이었다면 포기하고 다른 방법을 찾아 봤을텐데 이번 프로젝트를 진행하면서 포기하지 않는다면 방법은 있다는 것을 깨달았다.

### 노션 링크

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/2e2e1405-437f-4d04-9360-ff425b98d43d/cp1.ipynb>

#### ▼ 출처

- 통계청
- 서울시 범죄
- 서울시 유흥업소
- 서울시 1인 가구(연령별)
- 서울시 1인가구(거처종류별)
- 서울시 1인가구 지원정책

#### ▼ 통계용어

- 주택용어

- 서울시 주변환경 안전도
- 서울시 cctv



단독주택이란, 한 가구가 생활할 수 있도록 건축된 **일반 단독주택**과 여러 가구가 살 수 있도록 설계된 **다가구 단독주택**을 말한다.

- 서울시 권역명

권역명	행정구역
도심권	종로구, 중구, 용산구
동북권	성동구, 광진구, 동대문구, 중랑구, 성북구, 강북구, 도봉구, 노원구
서북권	은평구, 서대문구, 마포구
서남권	강서구, 양천구, 영등포구, 구로구, 금천구, 관악구, 동작구
동남권	강남구,강동구,서초구,송파구