Exploratory Data Analysis
-Boston Housing-

소프트웨어학부 2018044566 박서연

Overview of the dataset

- 보스턴 시의 주택 가격에 대한 데이터

- **주택 가격에 영향을 미치는 요소에는 무엇이 있는가?**

  ** 재산세율(TAX)는 재산의 보유 사실에 대해 과세하는 세금이다. 그러므로 주택 가격에 영향을 미치는 요소가 아닌 주택 가격에 영향을 받는 요소라 판단해 재산세율은 데이터 분석에서 제외하였다.

  가설 1) 범죄율이 높은 지역은 주택 가격이 낮을 것이다.
  가설 2) 비소매상업지역의 토지 점유율이 높을 수록 주택 가격이 낮을 것이다.
  가설 3) 질소 산화물 농도가 0.7ppm 이상일 경우 인체 기도 저항이 증가된다. 그러므로 질소 산화물 농도가 높을 수록 주택 분포가 적고, 주택 가격도 낮을 것이다.
  ⇒ 가설은 주택 가격에 영향을 미치는 요소인가?
  ⇒ 이 외에도 주택 가격에 영향을 미치는 요소가 있는가?

General Description : Data Set

- 총 506개의 데이터
- 14개의 속성이 존재

CRIM	town별 1인당 범죄율
ZN	25,000평방 피트가 넘는 토지에 대한 거주지의 비율
INDUS	비소매상업지역이 점유하고 있는 토지의 비율
CHAS	찰스강 인접 변수(강의 경계에 위치한 경우 1, 아니면 0)
NOX	10ppm 당 일산화질소(질소산화물) 농도
RM	주택당 평균 방의 개수
AGE	1940년 이전에 건설된 본인 소유의 집 사용자 비율
DIS	5개의 보스턴 직업센터의 가중거리
RAD	방사형 고속도로까지의 접근성지수
TAX	10000달러 당 재산세율
PTRATIO	town별 학생/교사 비율
B	town별 흑인 비율
LSTAT	모집단의 하위계층 비율
MEDV	본인 소유의 주택가격(중앙값) (단위: $1,000)

** CAT.MEDV = if MEDV > $30,000 then 1, else 0

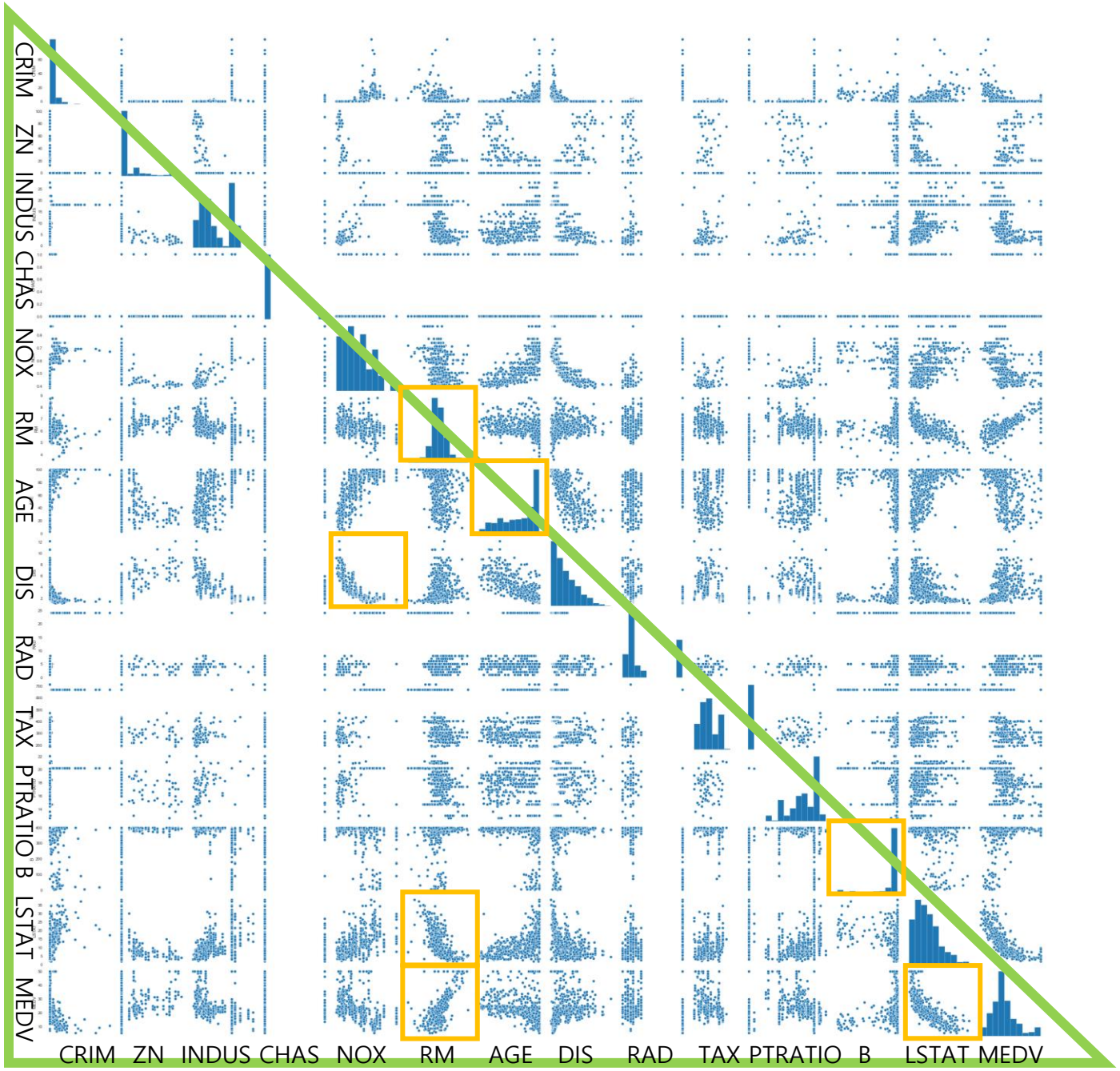
Descriptive Statistics

ID	속성	Min	Max	Avg	Std
CRIM	수치형	0.00632	88.99762	3.614	8.602
ZN	수치형	0.0	100.0	11.364	23.322
INDUS	수치형	0.46	27.74	11.1368	6.860
CHAS	명목형	0	1		
NOX	수치형	0.385	0.871	0.555	0.116
RM	수치형	3.561	8.78	6.285	0.703
AGE	수치형	2.9	100	68.575	28.149
DIS	수치형	1.1296	12.1265	3.795	2.106
RAD	정수형	1	24	9.549	8.707
TAX	정수형	187	711	408.237	168.537
PTRATIO	수치형	12.6	22.0	18.456	2.165
B	수치형	0.32	396.9	356.674	91.295
LSTAT	수치형	1.73	37.97	12.653	7.141
MEDV	수치형	5.0	50.0	22.533	9.197

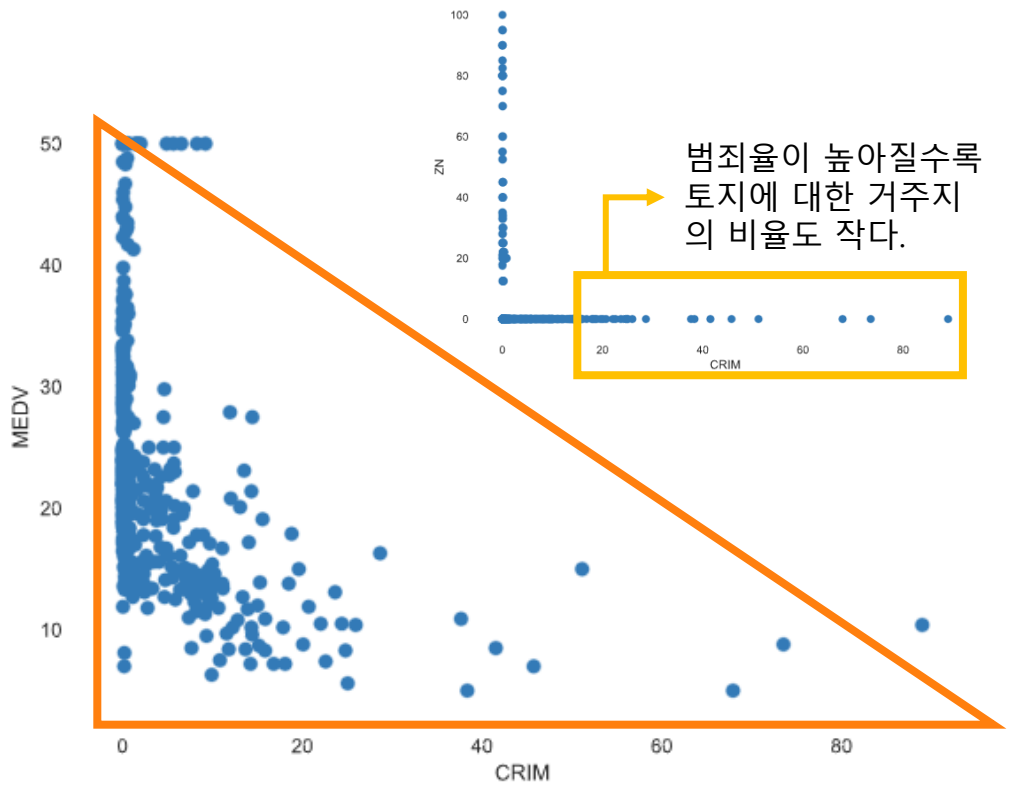
- 수치형 및 정수형 속성은 13개 명목형 속성은 1개이다.
- 최저 주택 가격(중앙값)은 $5,000이고 최고 주택 가격은 $50,000이다.
- 주택 가격(중앙값)이 $30,000이 넘는 주택의 수는 84채이고, 그렇지 않은 주택의 수는 442채이다.

Data Visualization

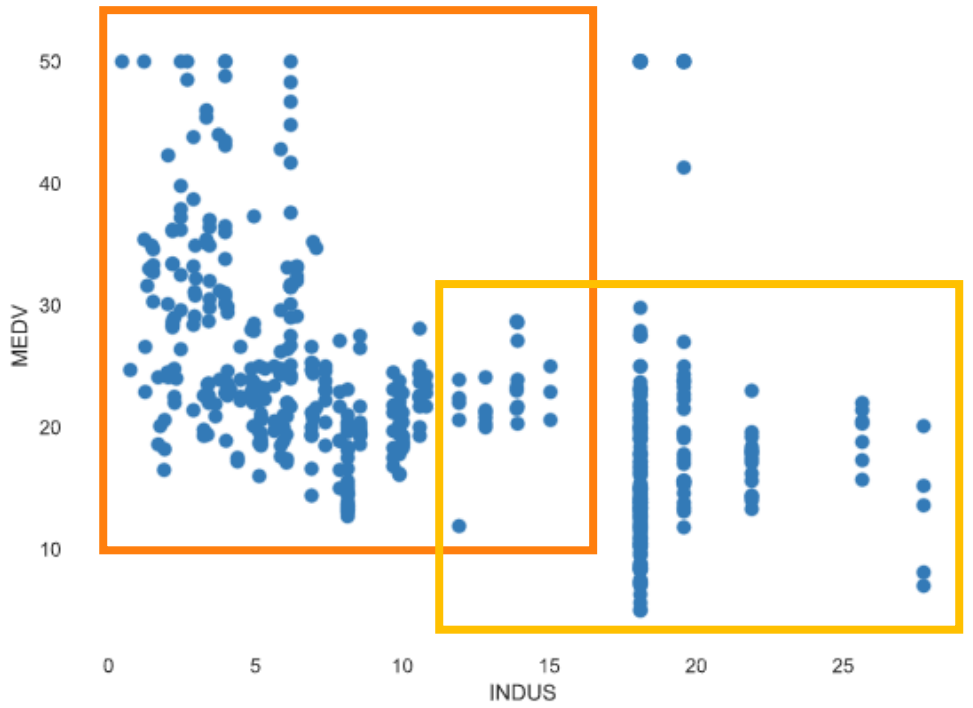
- 각 속성 간의 관계를 파악하기 위해 단변량 분석 및 다변량 분석을 시행하였다.
- 하위계층 비율과 방의 개수는 상관관계를 보인다.
- 직업센터의 가중거리와 일산화질소 농도는 음의 상관관계를 보인다.
- 보스턴시에는 오래된 주택이 많다.
- town별 흑인 비율은 크게 차이가 없다.
- 방의 개수는 비대칭 분포이지만, 정규분포에 가깝다.(Kurtosis=1.9)



Data Visualization

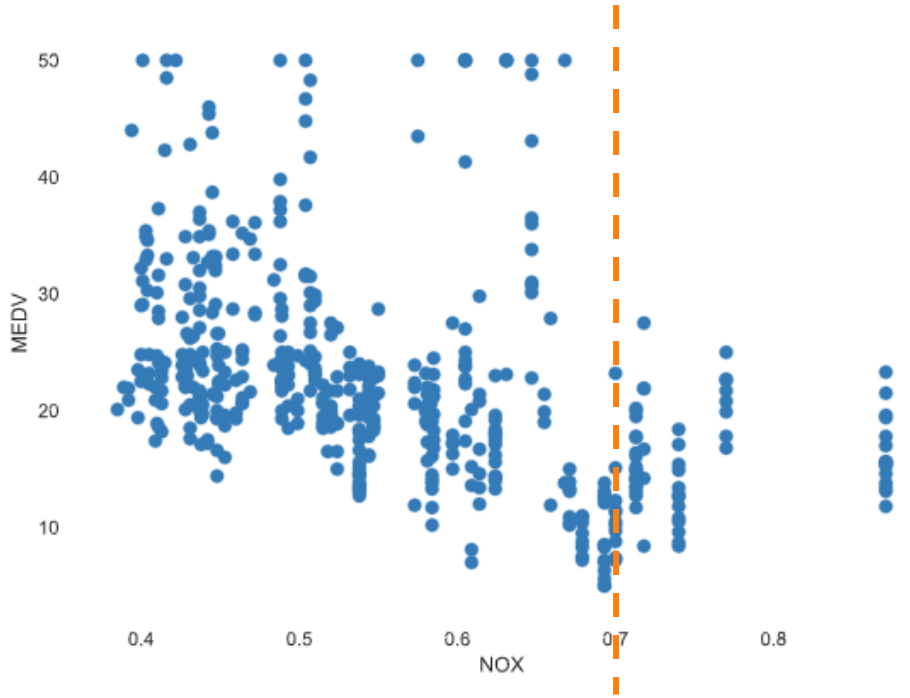


- 범죄율이 높을 수록 집값은 낮아진다. 그러나, 주택의 분포에 있어서 범죄율이 높은 곳은 거의 존재하지 않기 때문에, 상관관계가 있다고 할 수는 없다.



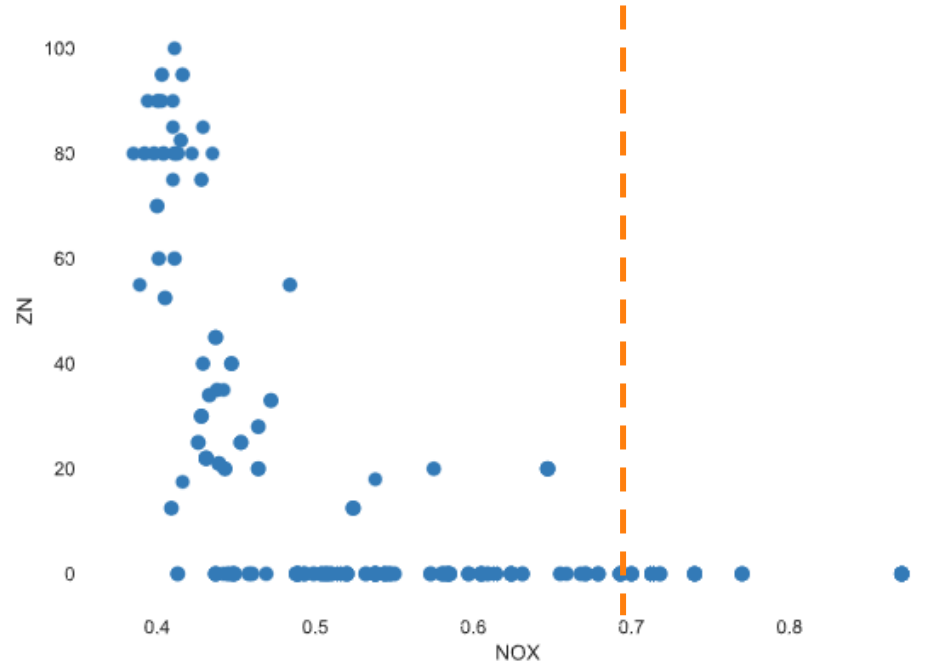
- 비소매상업지역의 토지 점유율이 높을 수록 집값은 낮은 쪽에 분포하고 있고, 토지 점유율이 낮을 수록 집값은 상대적으로 높은 쪽에 분포하고 있음을 확인할 수 있다. 그러나, 차이가 크지 않기 때문에, 집값에 영향을 주는 요소라 할 수 없다.

Data Visualization



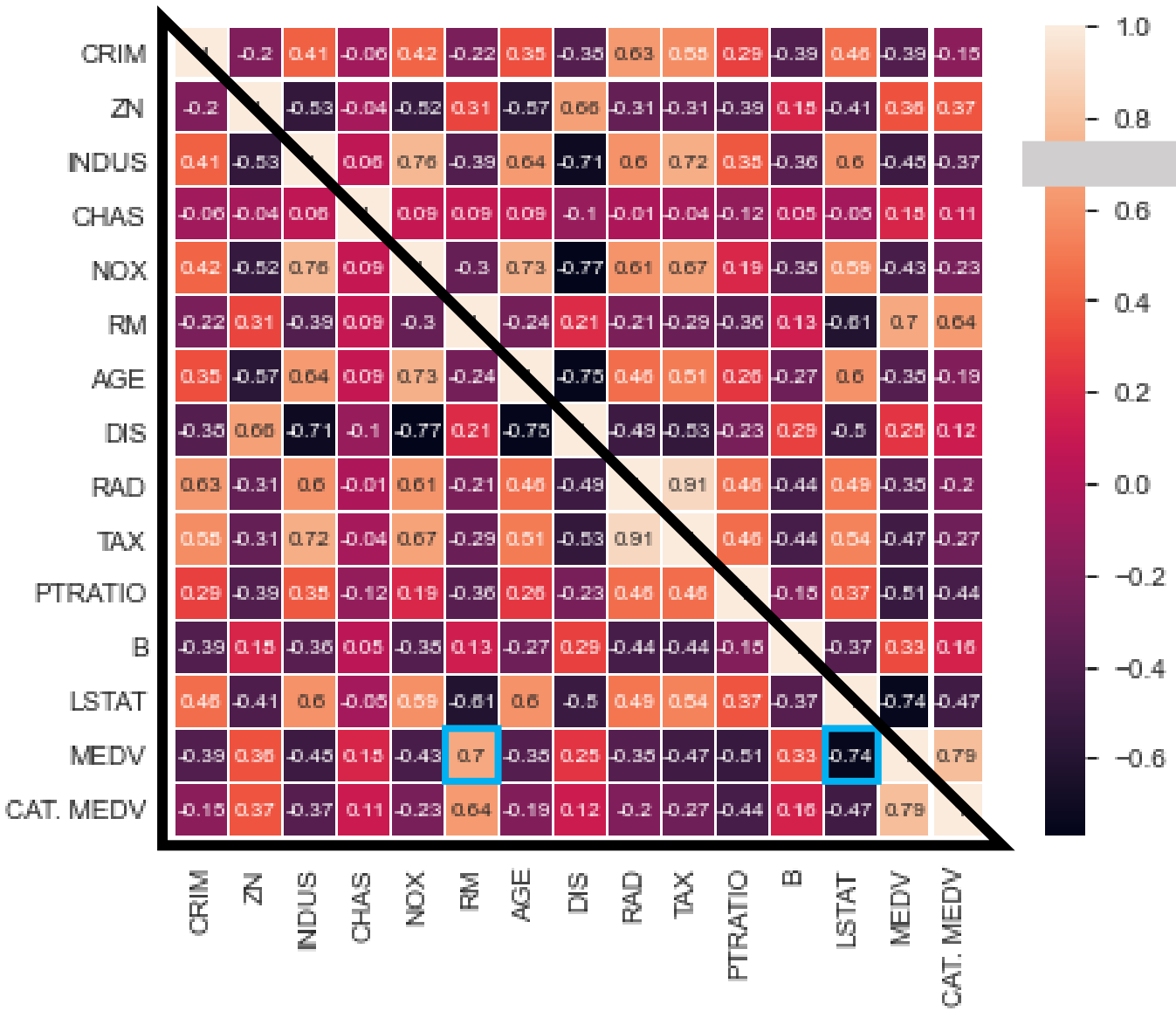
** 0.7ppm이상일 경우 인체 기도 저항이 증가

NOX가 약 0.7이상일 경우 주택의 분포가 적다. 집값이 싼 곳은 NOX 농도가 낮은 곳에서도 나타나기때문에, 주택의 수에는 영향을 주지만, 주택 가격에 영향을 준다고 할 수는 없다.

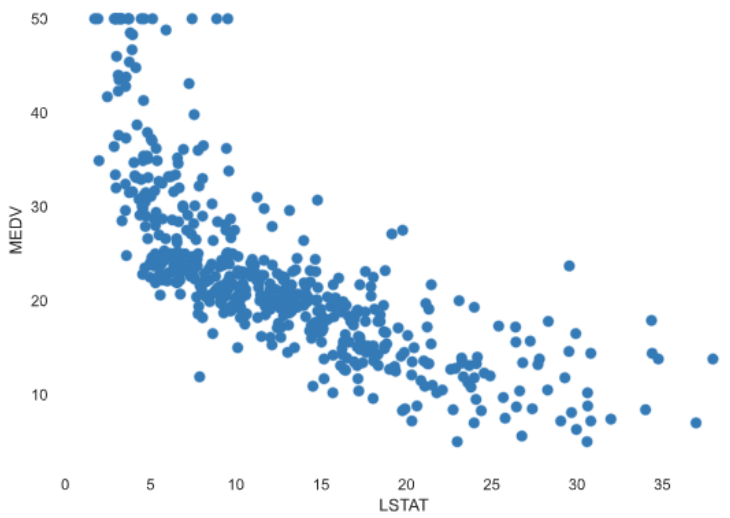


NOX가 높아질수록 토지에 대한 거주지의 비율이 낮아짐을 확인 할 수 있다.

Data Visualization

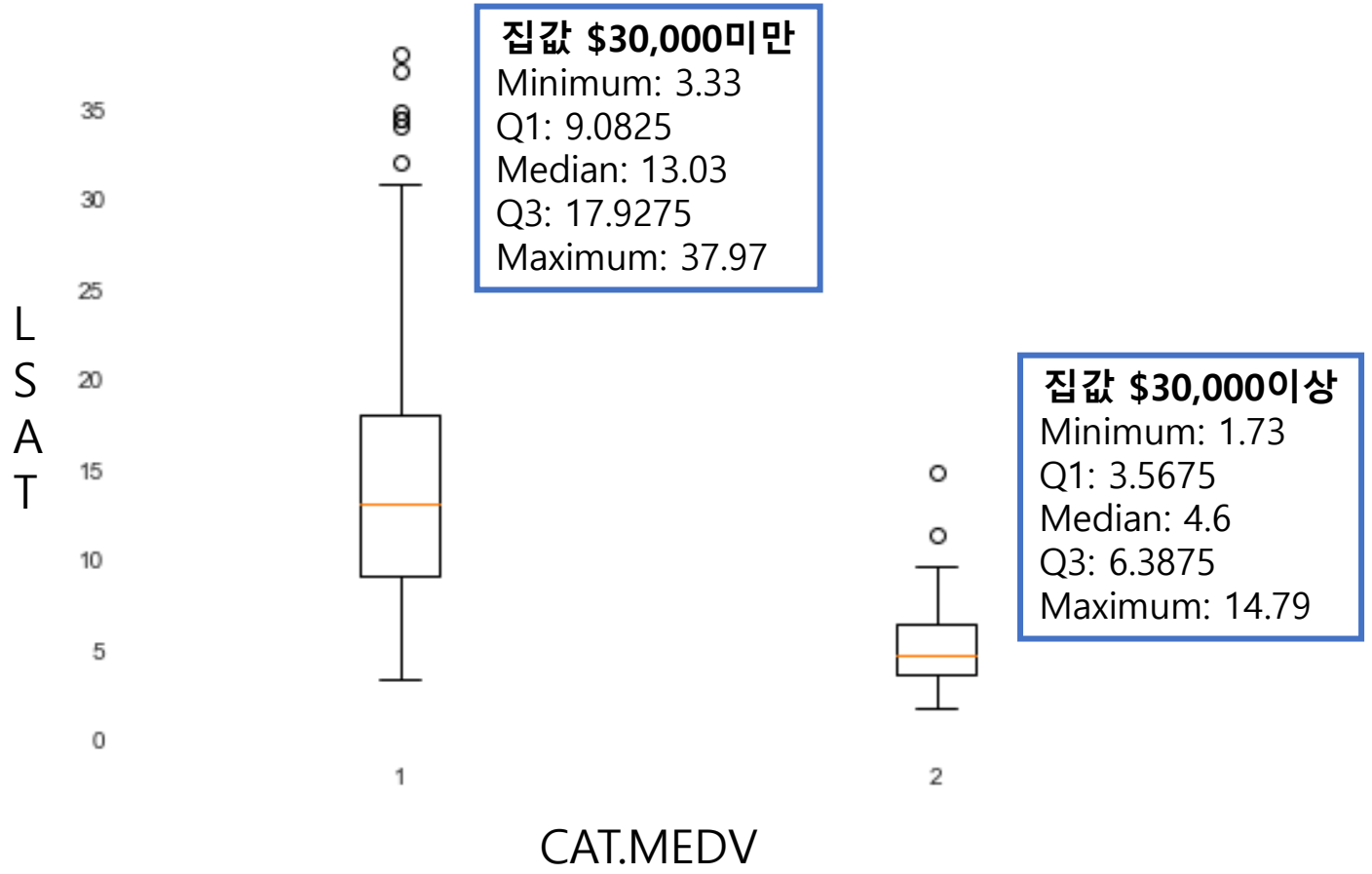


- MEDV-LSTAT



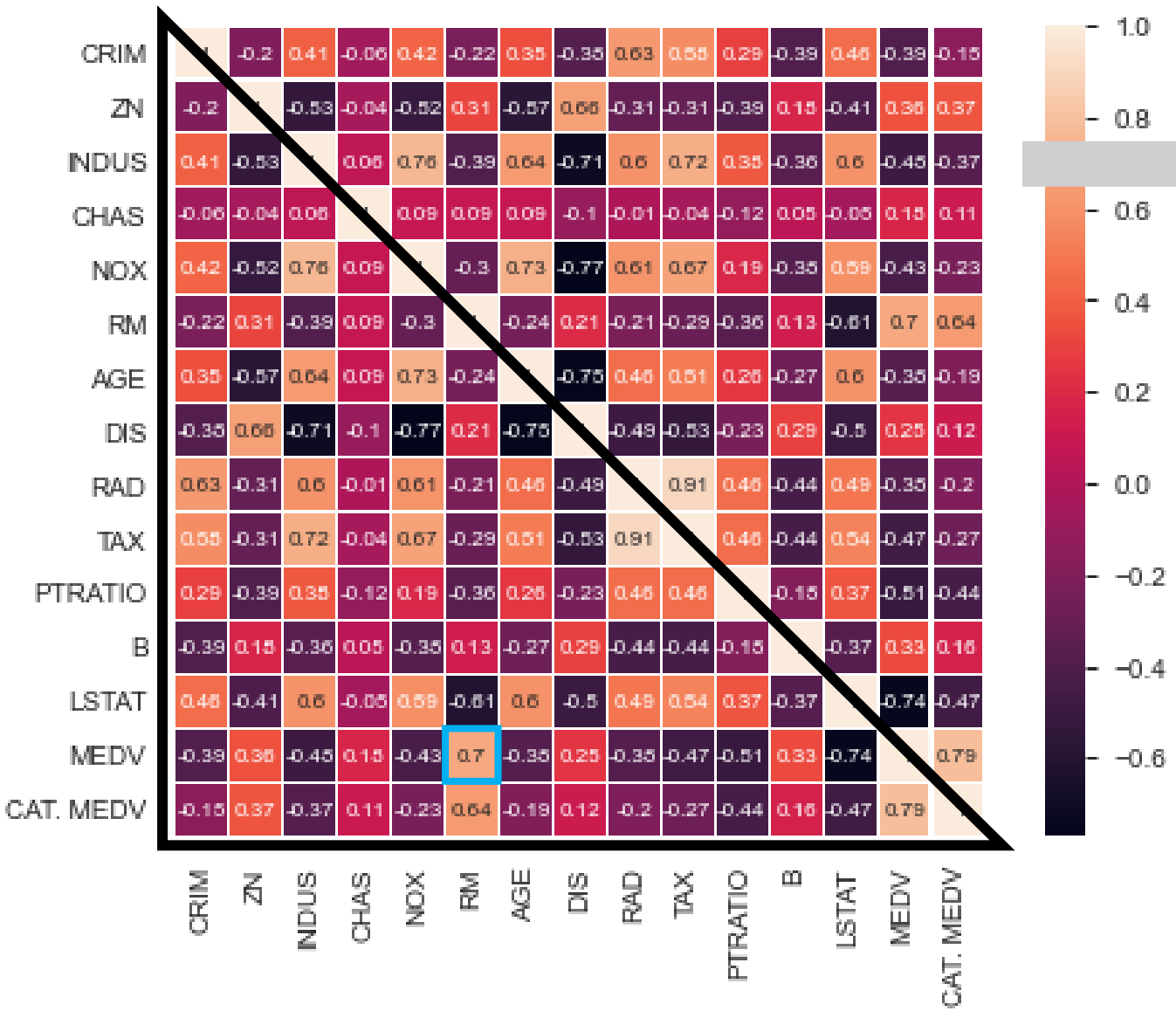
하위계층 비율이 높을 수록
집값이 낮다.

Data Visualization

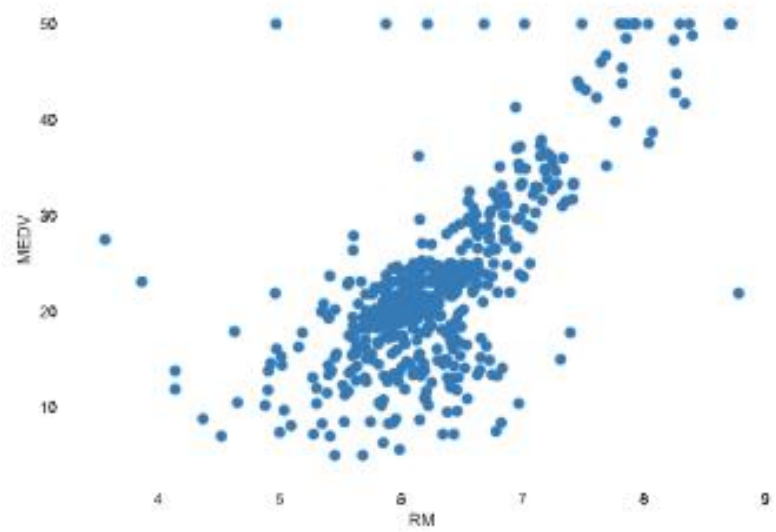


- $30,000를 기준으로 하위계층 비율의 편차를 확인하였다.
- 집값이 $30,000미만인 경우 하위계층 비율의 편차가 크다.

Data Visualization

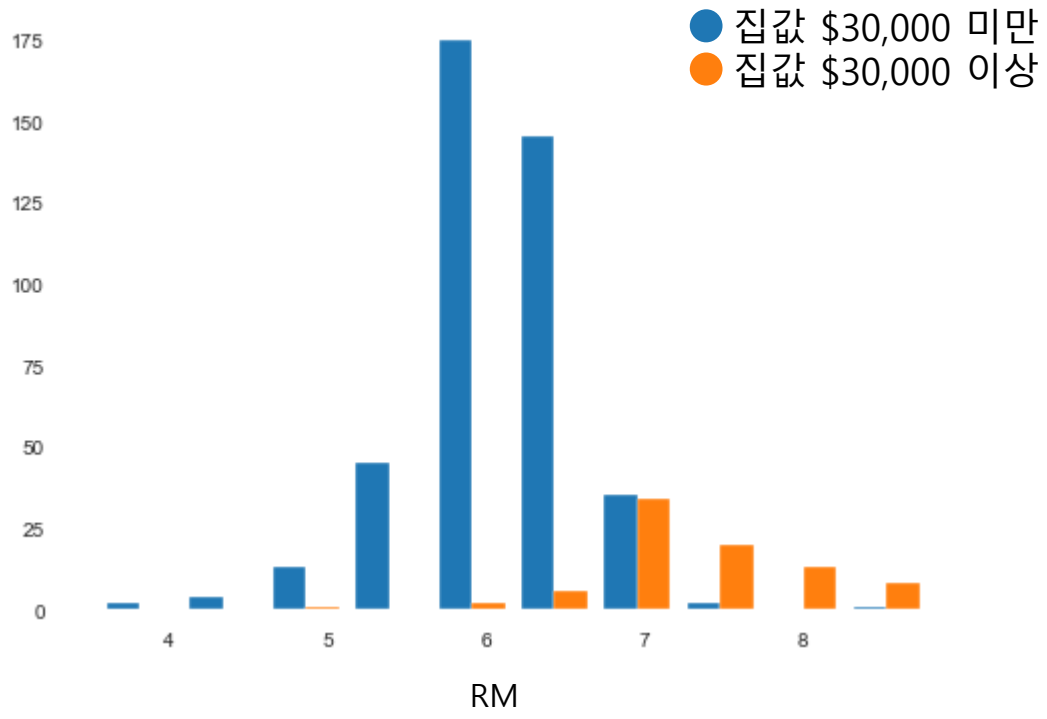


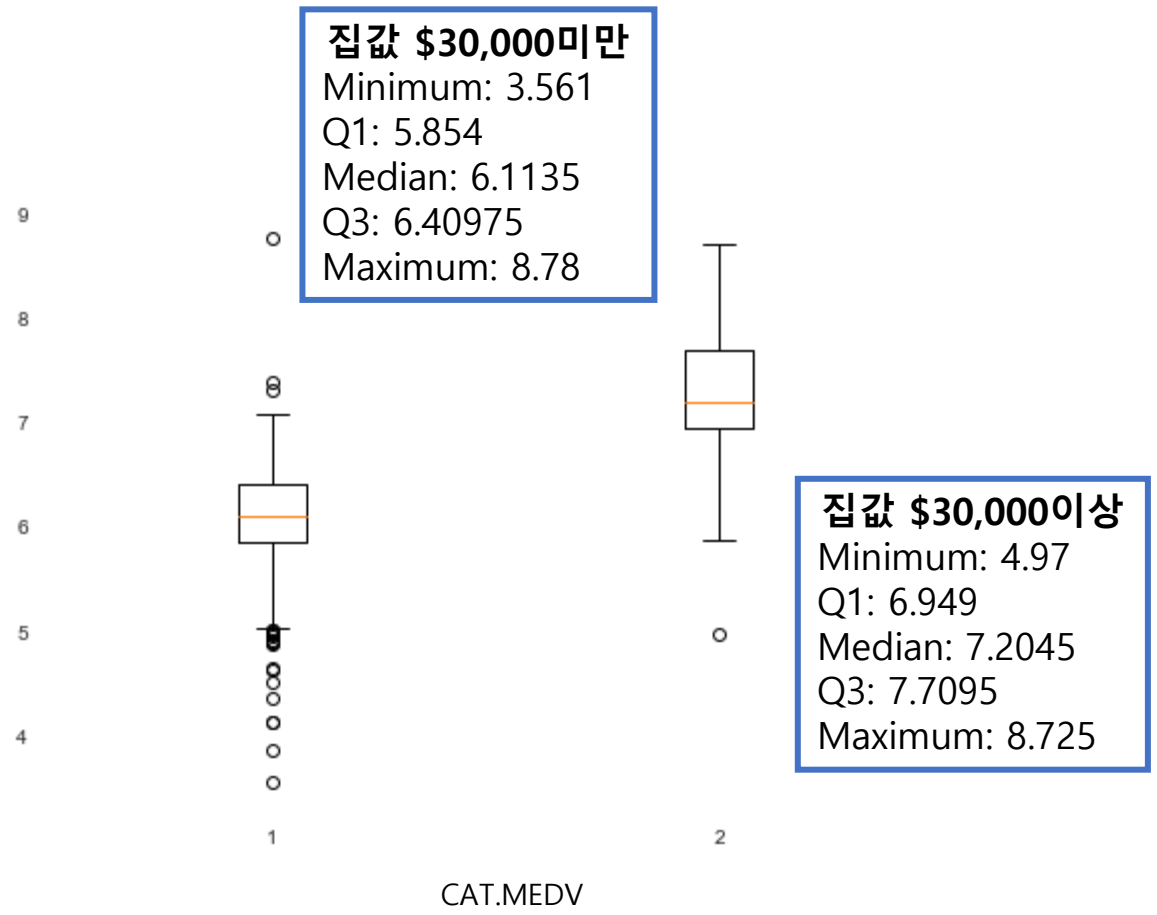
- MEDV-RM



방의 개수가 많을 수록
집값이 높다.

Data Visualization

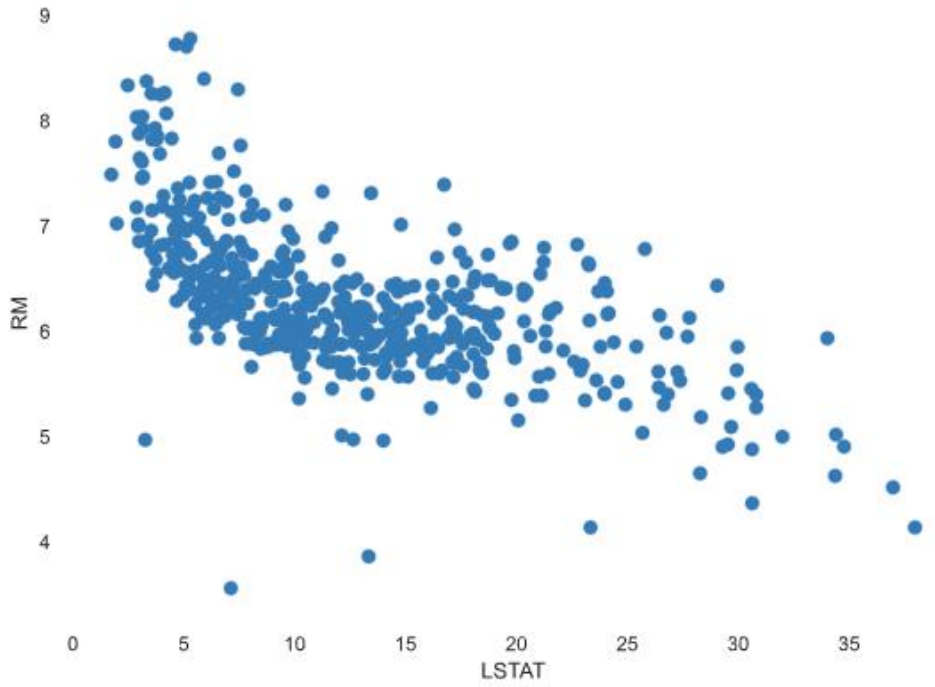




- $30,000를 기준으로 방의 개수의 분포를 확인하였다.
- 집값이 높을 수록 방의 개수의 분포가 더 높은 곳에 위치한다.

Data Visualization

- 앞서, 하위계층 비율이 적을수록, 방의 개수가 많을 수록 집값이 높음을 확인하였다.
- 하위계층 비율과 방의 개수는 음의 상관관계를 가진다. 즉, 하위계층 비율이 높을 수록 방의 개수는 줄어든다.



Result

- 주택 가격에 영향을 미치는 요소는 다음과 같다.
  ① 방의 개수
  ② 하위계층 비율

- 질소 산화물 농도, 범죄율은 주택 가격에 주는 영향보다, 주택의 분포에 영향을 주었다.
  농도가 높을 수록, 범죄율이 높을 수록 주택의 수는 현저히 적었다.

- 데이터셋 분석은 다음과 같은 프로그램을 사용하였다.
  ① python
    I. pandas-profiling: descriptive statistics
    II. matplotlib: boxplot, histogram,
    III. seaborn: heatmap, scatter matrix