1970'S 어느 집을 사는 것이 현명할까?

부제: 집 값에 영향을 많이 주는 변수는 무엇일까?

우영빈

과제 정의

주제 : **1970년대 보스턴에서 어느 집을 사는 것이 현명할까?**

배경지식:

보스턴의 경제에 중요한 산업

관광 산업

하버드 대학 등 유명 대학과 최초의 공립학교 설립

교육 도시

1970년대

집 값 폭등

1970년대 들어 보스턴 지역은

경기침체와 구조변화

다수의 고급 주택

미국에서 가장 높은 생계비를 요구하는 도시

보스턴 시의 가장 심각한 문제

도심의 교통체증

과제 정의

분석하고자 하는 방향: EDA를 통해 각 변수에 대한 인사이트를 얻은 후, 지금까지 배운 각 모델링에 모든 변수를 대입하여 가장 설명력이 높은 모델링과 가장 큰 중요도를 가지는 변수를 택한다. 이를 통해1970년대 보스턴에서 집을 구매할 때, 가장 중요하게 고려해야하는 요소와 가장 좋은 집을 택한다.

각 변수에 대한 가정

1) 범죄율: 범죄율이 낮을수록 집 값이 높아질까?

2) 주거지 비율: 주거지 비율이 높을수록 집 값이 높을까?

3) 비소매업 비율: 비소매업 비율이 낮을수록 집 값이 높을까?

4) 조망권: 조망권이 집 값이 높을까?

5) 산화질소: 산화질소 농도는 집 값과 관련이 없을 것 같은데 진짜일까?
산화질소 농도를 매번 체크하고 민감하게 반응하는 사람은 흔하지 않기 때문에 관련이 없을 것이라는 추측을 한다

6) 주거당 평균 객실 수 : 주거당 평균 객실 수가 많을수록 집 값이 높을까?

7) 노후 건물 비율: 노후건물 비율이 낮을수록 집 값이 높을 것이라 예상된다.

과제 정의

8) 중심지 접근거리: 중심지 접근거리가 가까울수록 집 값이 높을까? 회사가 집에서 가까운 것이 최고이기때문에 높을 것이라 예상한다.

9) 고속도록 접근 편이성 지수: 고속도로 접근 편이성 지수가 낮을수록 집 값이 낮을까? 고속도록 접근 편이성이 높으면 매연 등 좋지 않은 공기와 소음이 발생하기 때문에 집 값이 낮을것이라고 예상된다.

10) 재산세율: 재산세율이 높을수록 집 값이 높을까? 집 값이 비싸야 세금이 높아지기 때문이다.

11) 학생당 교수비율: 학생당 교수 비율은 영향이 없을 것 같다.

12) 흑인 인구 비율: 흑인 인구 비율이 낮을수록 집 값이 높을까? 흑인이 많은 곳은 할렘가 같은 곳이 연상되기때문에 낮을 것 같다.

13) 저소득층 비율: 저소득증 비율이 낮을수록 집 값이 높을까?

분석계획

1) 주제 선정 및 각 변수에 대한 가정 세우기

2) 데이터 현황 파악 1 - 갯수, 평균, 표준편차, 최소값, 최대값, IQR

3) 데이터 현황 파악 2 - 이상치, 결측치 확인, 산점도 ,히스토그램

4) 탐색적 분석 : 각 변수에 대한 가설 검정과 인사이트 도출

5) 변수간의 상관관계 파악

6) 모델링과 요약 : 회귀분석, 의사결정나무,랜덤포레스트, 그레디언트부스팅

7) 변수 중요도 파악

8) 개선 방향과 결론 도출

10) 실습 과정을 통해 배운점 또는 느낀점, 애로사항 정리

데이터 현황

데이터 현황 확인: **count, mean, std, min, 25% 50% 75%, max**

	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM
count	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
mean	22.532806	3.613524	11.363636	11.136779	0.069170	0.554695	6.284634
std	9.197104	8.601545	23.322453	6.860353	0.253994	0.115878	0.702617
min	5.000000	0.006320	0.000000	0.460000	0.000000	0.385000	3.561000
25%	17.025000	0.082045	0.000000	5.190000	0.000000	0.449000	5.885500
50%	21.200001	0.256510	0.000000	9.690000	0.000000	0.538000	6.208500
75%	25.000000	3.677083	12.500000	18.100000	0.000000	0.624000	6.623500
max	50.000000	88.976196	100.000000	27.740000	1.000000	0.871000	8.780000

AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
506.000000	506.000000	506.000000	506.000000	506.000000	506.000000	506.000000
68.574901	3.795043	9.549407	408.237154	18.455534	356.674030	12.653063
28.148862	2.105710	8.707259	168.537116	2.164946	91.294863	7.141062
2.900000	1.129600	1.000000	187.000000	12.600000	0.320000	1.730000
45.025000	2.100175	4.000000	279.000000	17.400000	375.377487	6.950000
77.500000	3.207450	5.000000	330.000000	19.050000	391.440002	11.360000
94.074999	5.188425	24.000000	666.000000	20.200001	396.225006	16.954999
100.000000	12.126500	24.000000	711.000000	22.000000	396.899994	37.970001

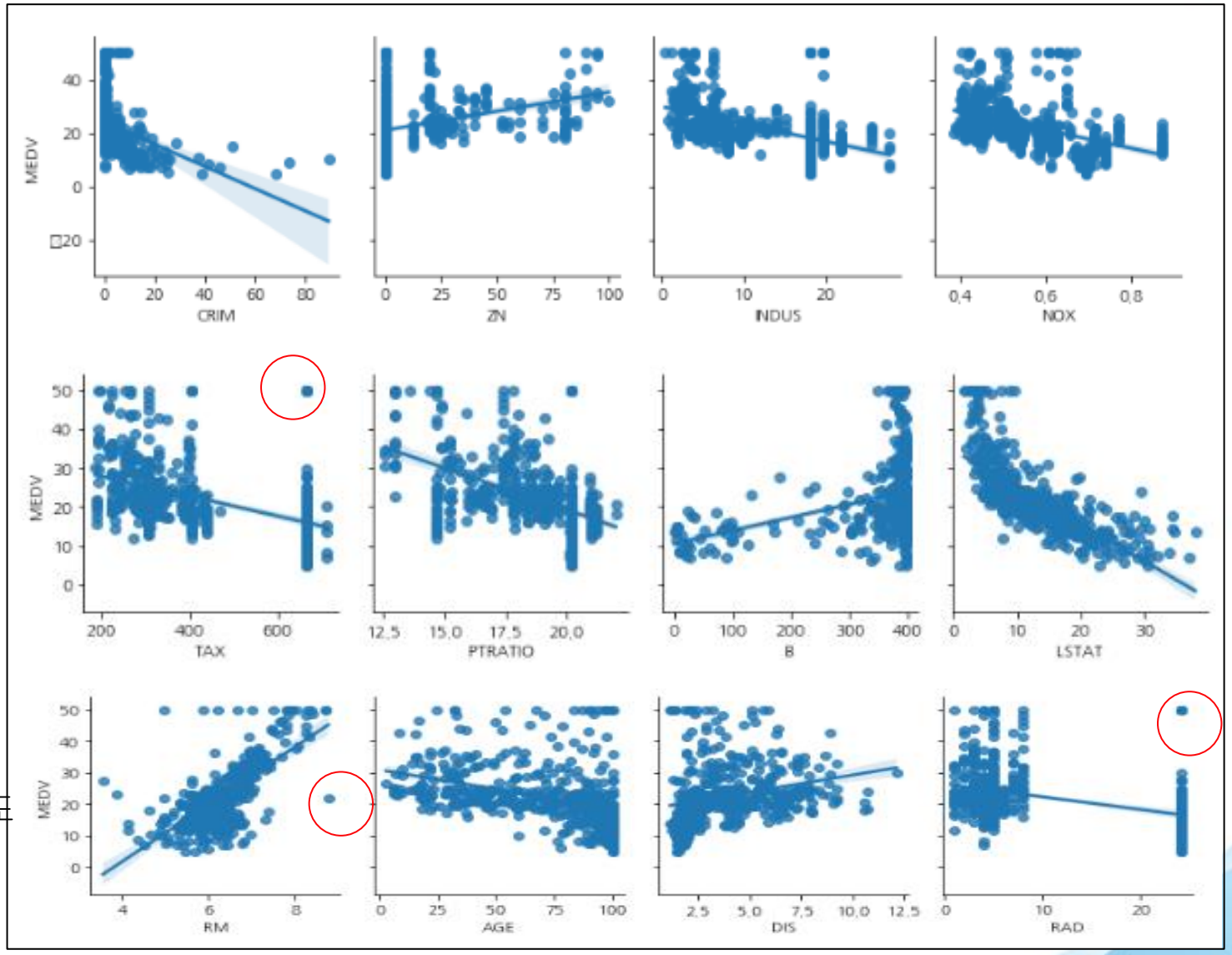
데이터 현황

결측치

```
MEDV       0
CRIM       0
ZN         0
INDUS      0
CHAS       0
NOX        0
RM         0
AGE        0
DIS        0
RAD        0
TAX        0
PTRATIO    0
B          0
LSTAT      0
dtype: int64
```

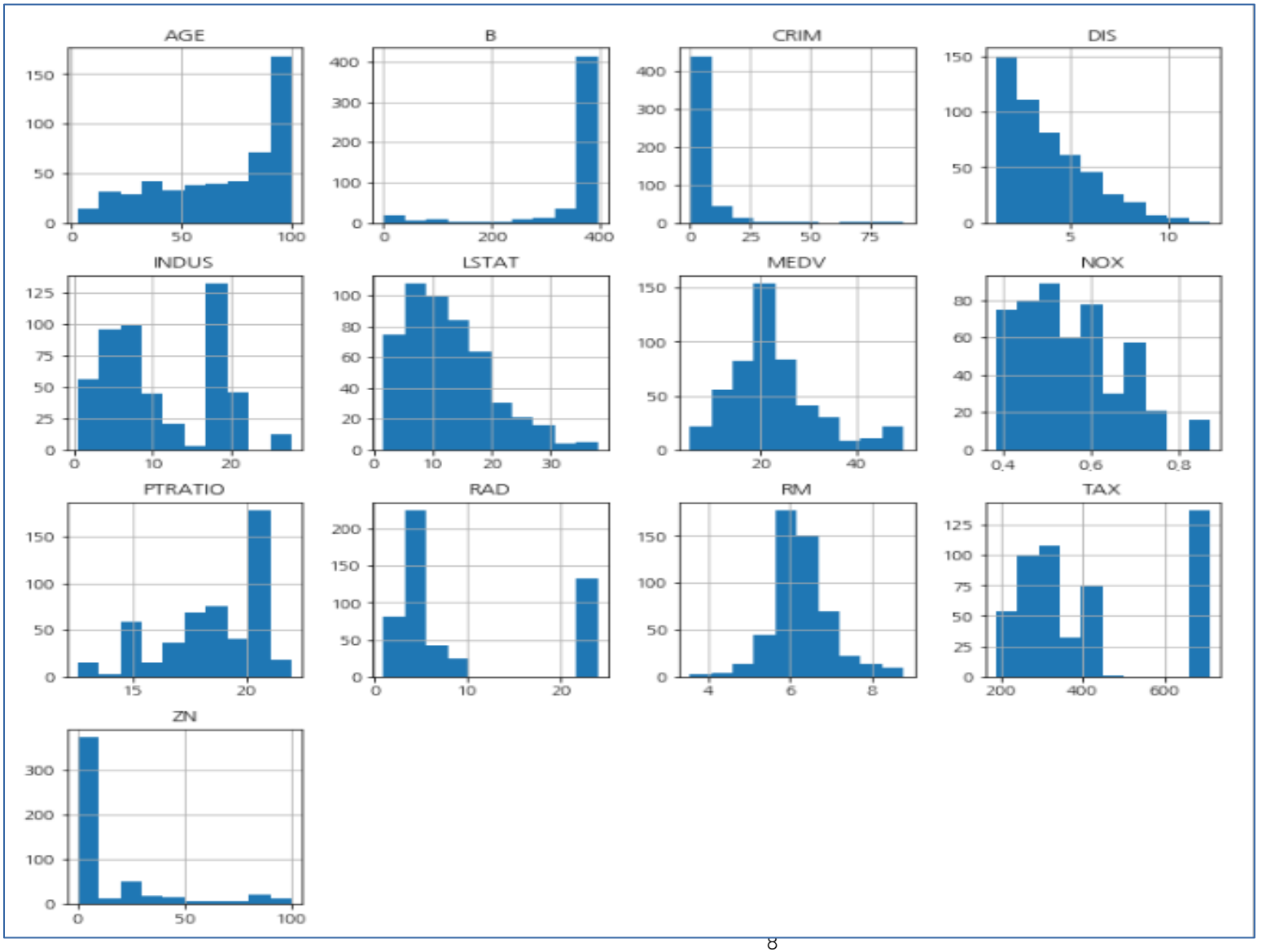
이상치가 같은 것이 보이나
교수님께서 의미있는 수일 수도
있다고 하셔서
제거하지 않기로 했다.

산점도 그래프



데이터 현황

히스토그램

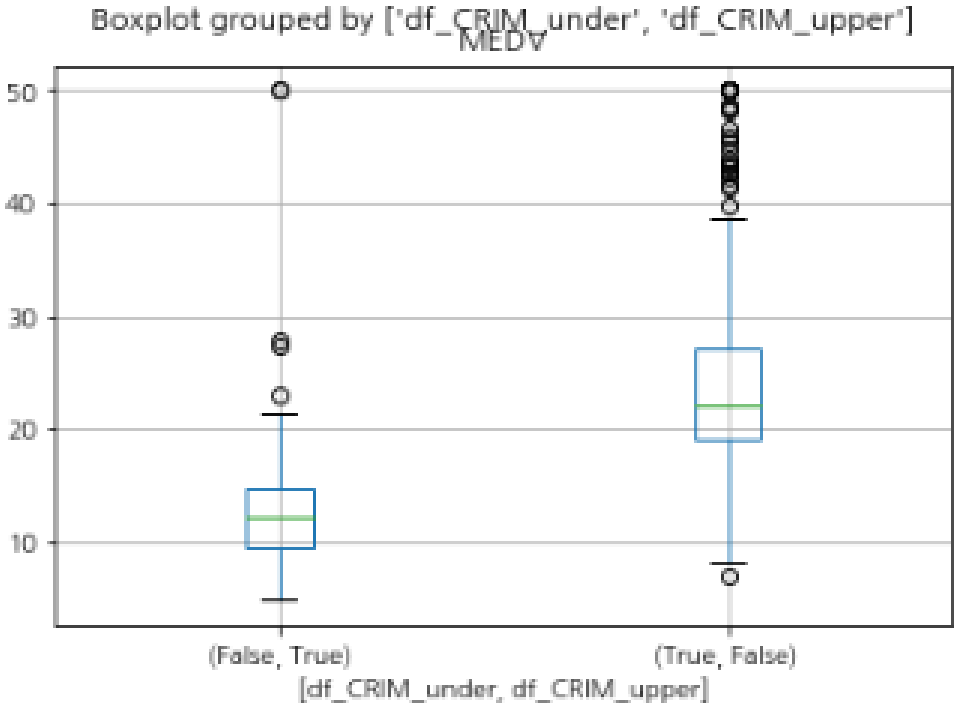


B,CRIM,INDUS
MEDV,RAD,CRIM

양극화된 경향 보임

각 변수에서 인사이트를 얻기 위해 T-test 검정 시행

범죄율(CRIM)이 낮을수록 집 값이 비싸질까?

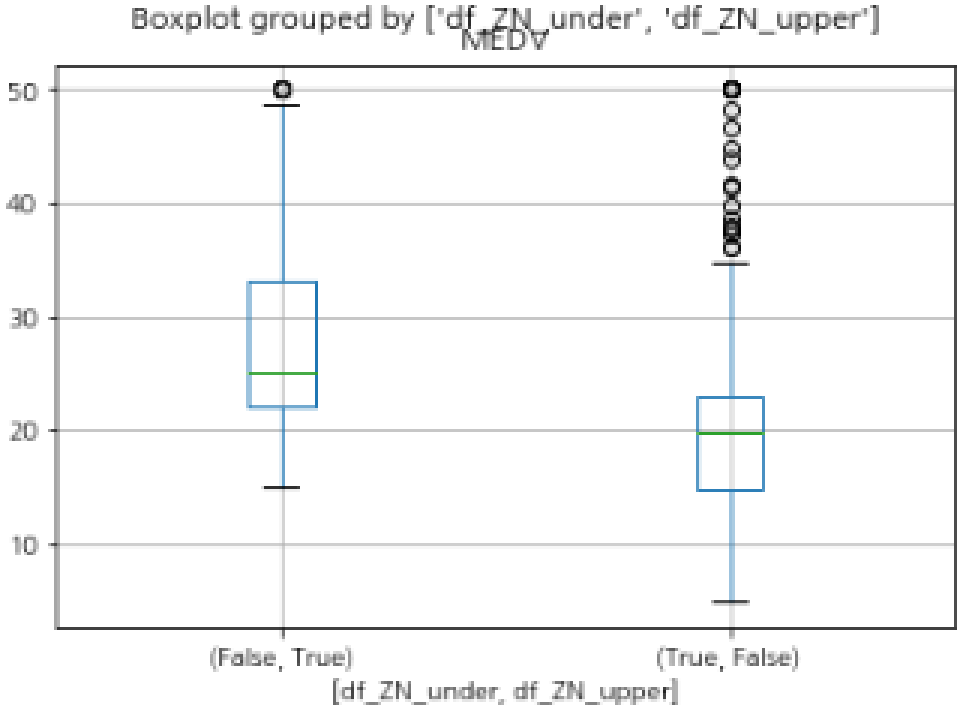



```
2-Sample t-test
Ttest_indResult(statistic=9.319889480008591, pvalue=3.6333822874473695e-19)
t:9.32
p:0.0
```

(범죄율7.5를 기준으로 컴죄율이 높은 지역과 낮은 지역 두 부류로 나눔)

귀무가설: 범죄율이 높은 지역과 범죄율이 낮은 지역의 주택 가격 평균이 같다.
대립가설: 범죄율이 높은 지역과 범죄율이 낮은 지역의 주택 가격 평균이 다르다
→**T-test**와 **BOX-Plot**을 통해, 범죄율이 낮은 지역의 주택 가격 평균이 높다는 것을 알 수 있다.

주거지 비율(ZN)이 높을수록 집 값이 비싸질까?

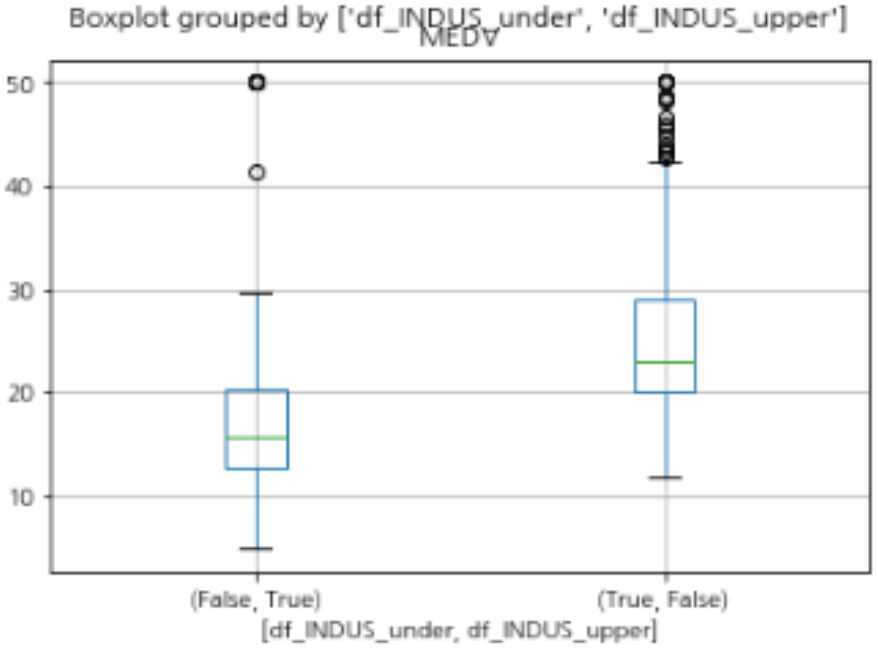



```
2-Sample t-test
Ttest_indResult(statistic=-9.038052730
869948, pvalue=3.41982474583652 73e-18)
t:-9.038
p:0.0
```

(주거지비율율 10 기준으로 두 부류로 나눈다.)

귀무가설: 주거지 비율이 높은 지역과 주거지 비율이 낮은 지역의 주택 가격 평균이 같다.
대립가설: 주거지 비율이 높은 지역과 주거지 비율이 낮은 지역의 주택 가격 평균이 다르다
.
→**T-test**와 **BOX-Plot**을 통해, 주거지 비율이 높은 지역일수록 주택 가격이 높은 것을 알 수 있다.

비소매업 비율(INDUS)이 높수록 집 값이 비싸질까?



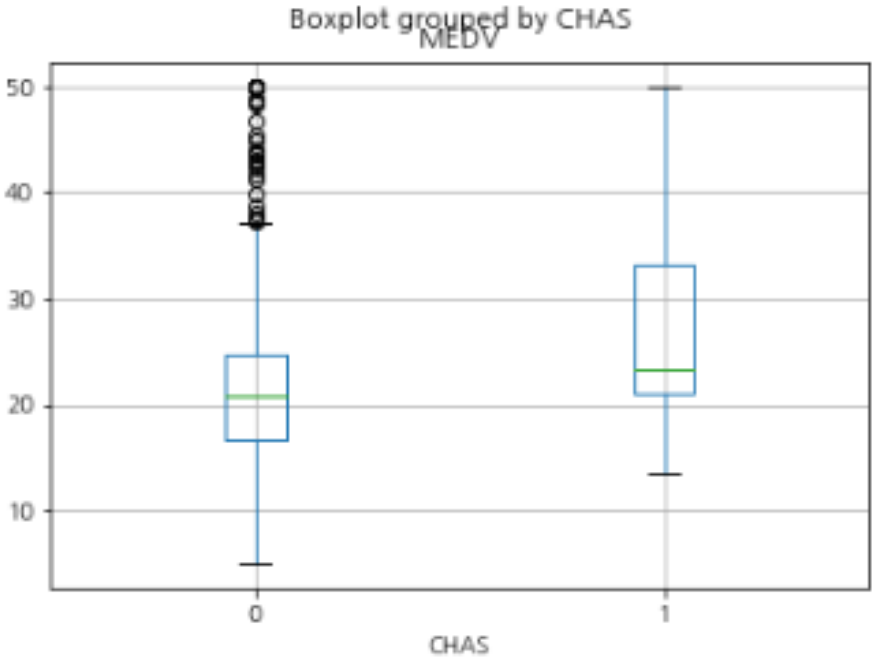

```
Sample t-test
est_indResult(statistic=10.163196066
023, pvalue=3.3702023559466714e-22)
10.163
0.0
```

(비소매업 비율 16을 기준으로 두 부류로 나눈다.
)

귀무가설: 비소매율이 높은 지역과 비소매율이 낮은 지역의 주택 가격 평균이 같다.
대립가설: 비소매율이 높은 지역과 비소매율이 낮은 지역의 주택 가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 비소매율이 낮은 지역이 주택 가격이 높다는 것을 예상할 수 있다.

강조망 여부(CHAS) 조망일수록 집 값이 비싸질까?



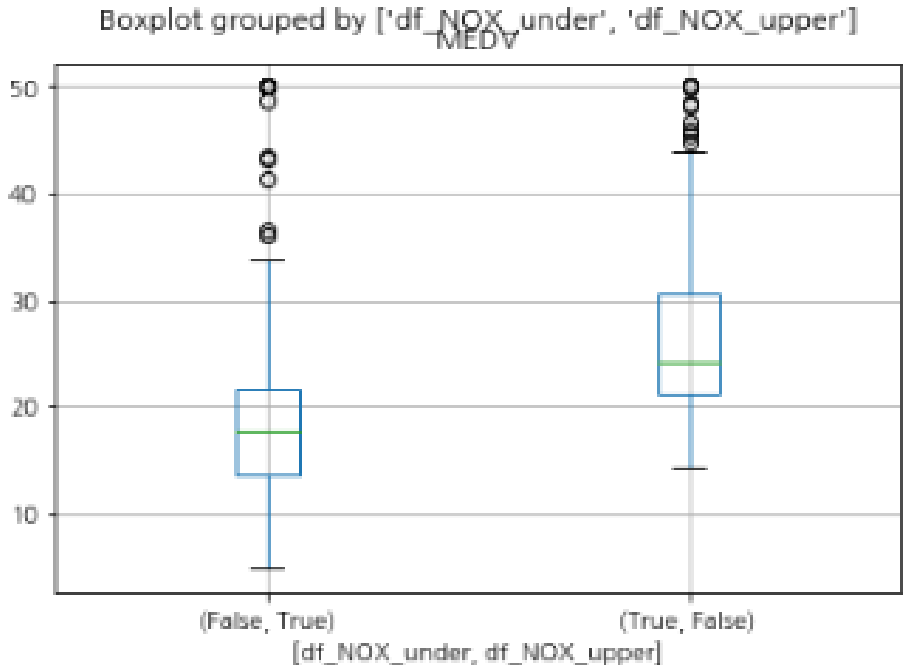

```
2-Sample t-test
Ttest_indResult(statistic=53.960094894
90948, pvalue=8.757522855106094e-300)
t:53.96
p:0.0
```

(조망과 비조망 두 부류로 나눈다
.)

귀무가설: 비조망과 조망의 주택가격 평균이 같다.
대립가설: 비조망과 조망의 주택가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 조망권과 비조망권의 차이가 유의미한다고 나왔지만 데이터셋 자체의 비조망권의 개수는 **471**개로 무수히 많고 조망권의 개수는 **35**개 적다. 불균형한 데이터셋이라고 판단하여 결과는 믿을 수 없다는 결론을 냈다.

산화질소 농도(NOX) 집 값이 영향이 있을까?

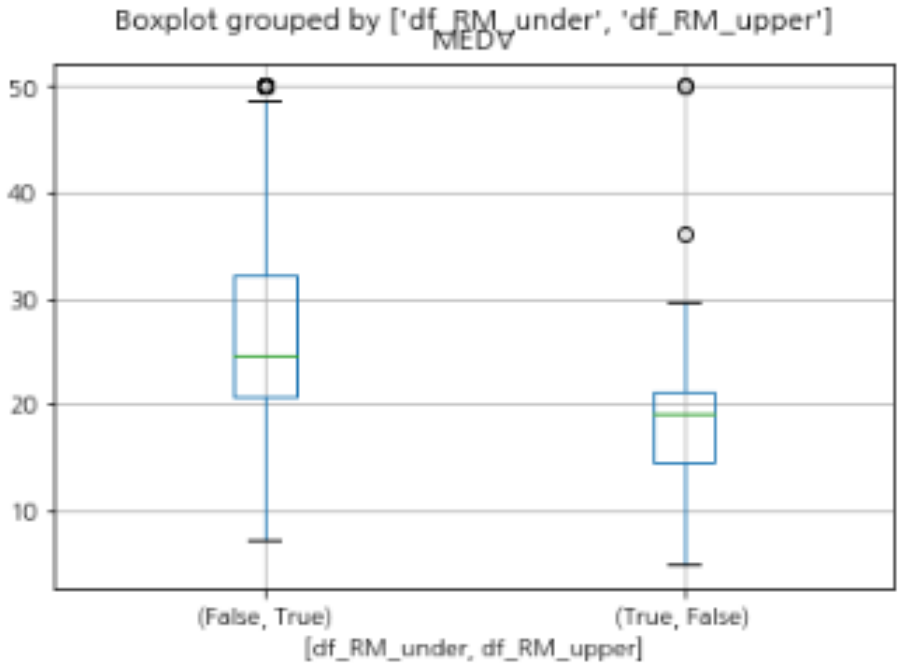



```
2-Sample t-test
Ttest_indResult(statistic=5.5430811266
6087, pvalue=4.2352484884514224e-08)
t:5.543
p:4.24e-08
```

(산화질소가 0.52보다 높을 때와 낮을 때 두 부류로 나눈다.)

귀무가설: 산화질소 농도가 높은 집단의 주택 가격 평균과 산화질소 농도가 낮은 집단의 주택가격 평균이 같다.
대립가설: 산화질소 농도가 높은 집단의 주택 가격 평균과 산화질소 농도가 낮은 집단의 주택가격 평균이 다르다.
→ **T-test**와 **BOX-Plot**을 통해, 산화질소 농도가 낮을수록 집값이 높다는 것을 알 수 있다.

주거당 평균 객실 수(RM)이 많으면 집 값이 높아질까?



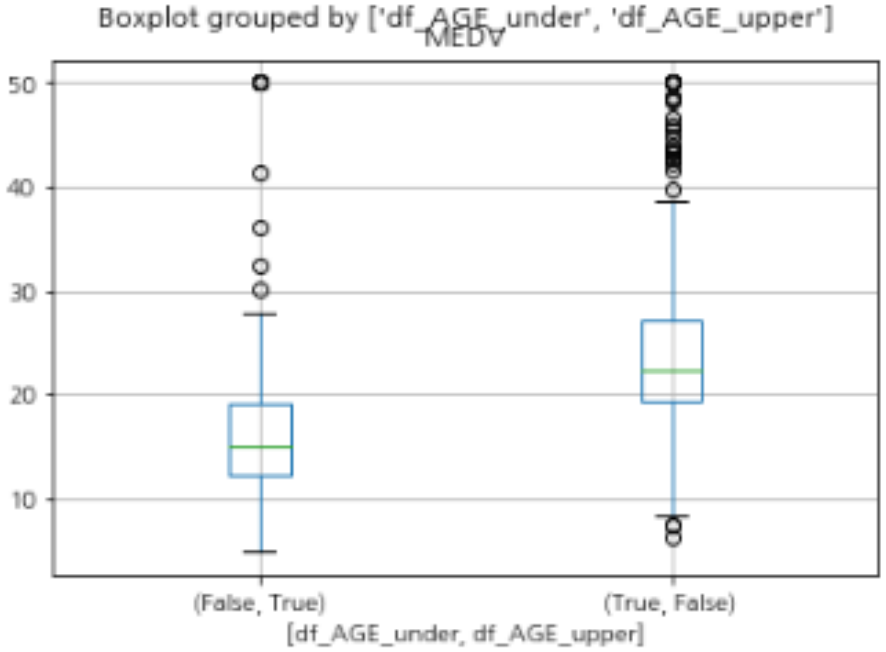

```
2-Sample t-test
Ttest_indResult(statistic=-11.732459745145503, pvalue=2.850535845980495e-28)
t:-11.732
p:0.0
```

(주거당 평균 객실 수를 6.2를 기준으로 두 부류로 나눈다.
)

귀무가설: 주거당 평균 객실 수가 많은 집단의 주택 가격 평균과 주거당 평균 객실 수가 적은 집단의 주택가격 평균이 같다.
대립가설: 주거당 평균 객실 수가 많은 집단의 주택 가격 평균과 주거당 평균 객실 수가 적은 집단의 주택가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 주거당 평균 객실 수가 많을수록 집 값이 높다는 것을 알 수 있다.

노후 건물 비율(RM)이 많으면 집 값이 높아질까?



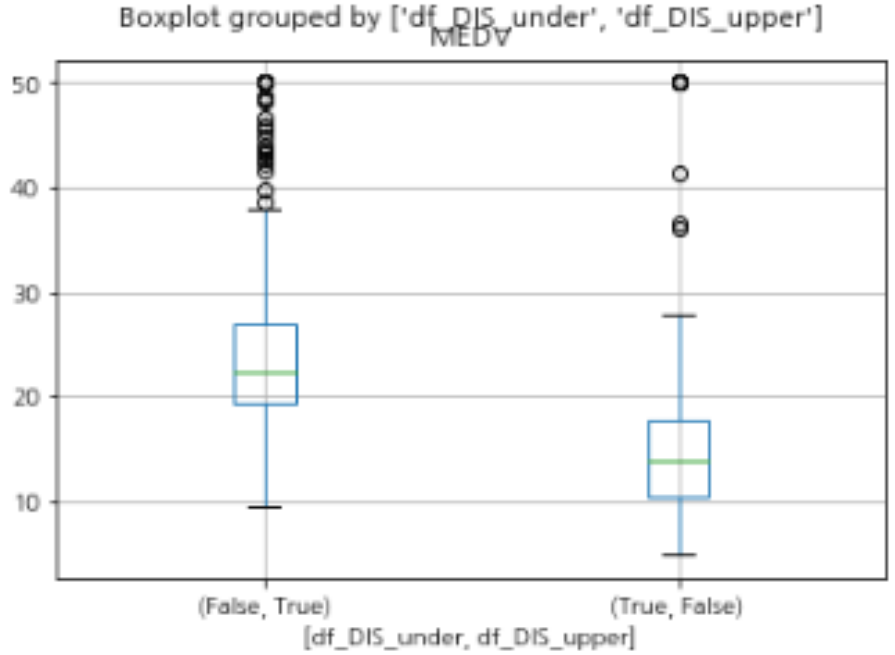

```
2-Sample t-test
Ttest_indResult(statistic=8.093983771830333, pvalue=4.359012441225016e-15)
t:8.094
p:0.0
```

(노후 건물 비율을 95 기준으로 두 부류로 나눈다.)

귀무가설: 노후 건물 비율이 높은 집단의 주택 가격 평균과 노후 건물 비율이 낮은 집단의 주택가격 평균이 같다.
대립가설: 노후 건물 비율이 높은 집단의 주택 가격 평균과 노후 건물 비율이 낮은 집단의 주택가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 노후 건물 비율이 낮을수록 집단의 주택가격 평균이 높다.

중심지(노동센터)거리(DIS) 가까울수 집 값이 높아질까?

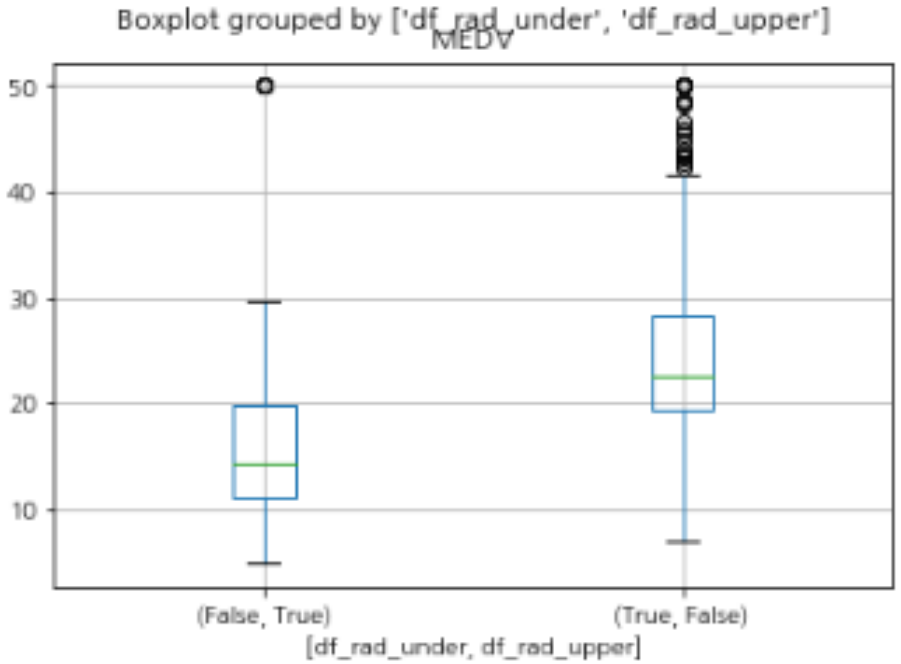



```
2-Sample t-test
Ttest_indResult(statistic=-7.785196453
373674, pvalue=3.98329830151541e-14)
t:-7.785
p:0.0
```

(중심지 거리를 2 기준으로 두 부류로 나눈다.
)

귀무가설: 중심지(노동센터) 접근 거리가 먼 집단의 주택 가격 평균과 중심지(노동센터) 접근 거리가 가까운 집단의 주택가격 평균이 같다.
대립가설: 중심지(노동센터) 접근 거리가 먼 집단의 주택 가격 평균과 중심지(노동센터) 접근 거리가 가까운 집단의 주택가격 평균이 다르다.
→ **T-test**와 **BOX-Plot**을 통해,중심지 거리가 멀수록 주택가격 평균이 높다.

고속도로 접근 편이성 지수(RAD)가 멀수록 집 값이 높아질까?



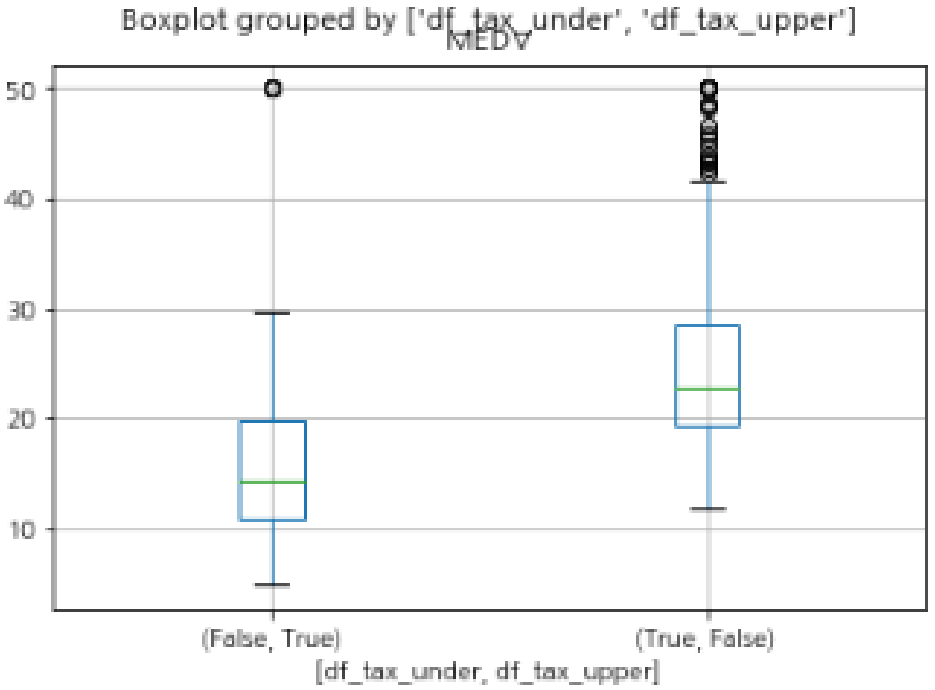

```
2-Sample t-test
Ttest_indResult(statistic=9.690255282930414, pvalue=1.7778936162464715e-20)
t:9.69
p:0.0
```

(고속도로 접근 편이성 지수를 24 기준으로 두 부류로 나눈다.)

귀무가설: 고속도로 접근 편이성 지수가 높은 집단의 주택 가격 평균과 고속도로 접근 편이성 지수이 낮은 집단의 주택가격 평균이 같다.
대립가설: 고속도로 접근 편이성 지수가 높은 집단의 주택 가격 평균과 고속도로 접근 편이성 지수이 낮은 집단의 주택가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 고속도로 접근성이 낮을수록 주택가격 평균이 높다.

재산세율(TAX)가 높을수 집 값이 높아질까?

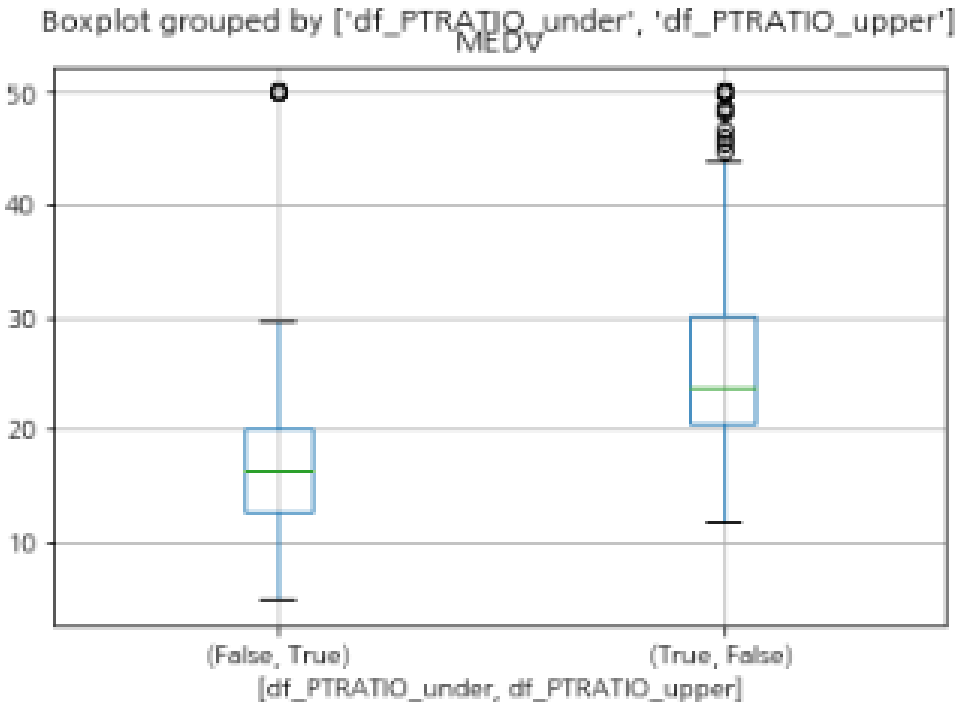



```
2-Sample t-test
Ttest_indResult(statistic=10.245571183
482726, pvalue=1.6682534285380135e-22)
t:10.246
p:0.0
```

(재산세율을 600 기준으로 두 부류로 나눈다.
)

귀무가설: 재산세율이 낮은 집단의 주택 가격 평균과 높은 집단의 주택가격 평균이 같다.
대립가설: 재산세율이 낮은 집단의 주택 가격 평균과 높은 집단의 주택가격 평균이 다르다.
→ **T-test**와 **BOX-Plot**을 통해, 재산세율이 높을수록 주택 가격 평균이 낮다..
→ 결과가 유의미한다고 나왔지만, 데이터셋에 **600** 이상은 약 **150**개로 적고 미만은 **350**개로 많다. 불균형한 데이터셋이라고 판단하기 위해서는 더 확연한 차이가 나야한다.

학생당 교수 비율(PTRATIO)이 집 값에 영향을 미칠까?



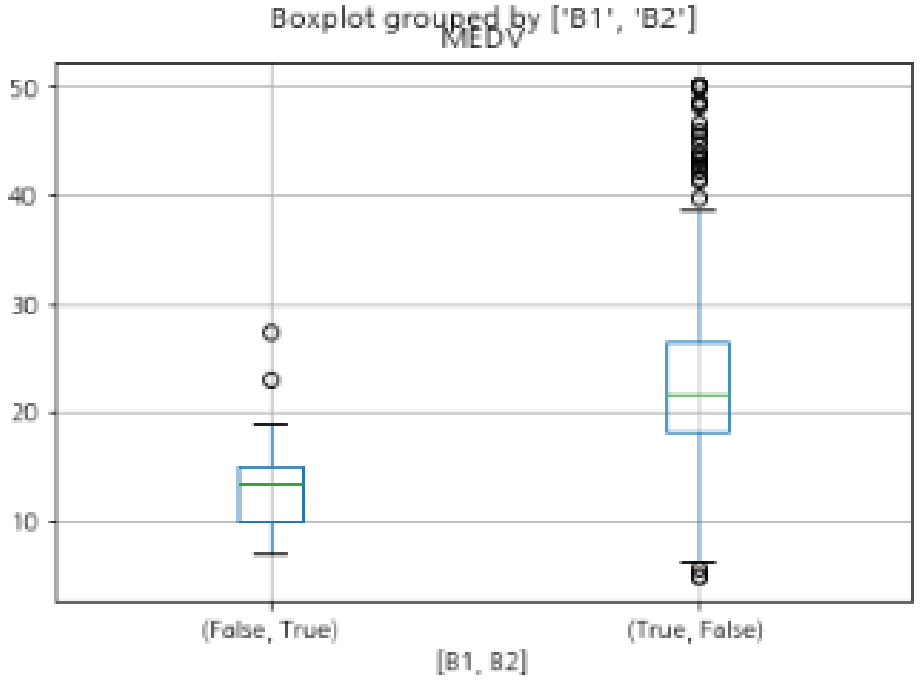

```
2-Sample t-test
Ttest_indResult(statistic=12.766967231
954494, pvalue=1.5395704416719087e-32)
t:12.767
p:0.0
```

(학생당 교수 비율을 20 기준으로 두 부류로 나눈다.)

귀무가설: 학생당 교수 비율 적은 집단의 주택 가격 평균과 많은 집단의 주택가격 평균이 같다.
대립가설: 학생당 교수 비율이 많은 집단의 주택 가격 평균과 많은 집단의 주택가격 평균이 다르다

→ **T-test**와 **BOX-Plot**을 통해, 학생당 교수 비율이 작을수록 집 값이 높다는 것을 예상할 수 있다.상식적으로는 학생당 교수 비율이 높아야 좋은 학군이고, 집값이 높을 것이라 예측되는데 다시 생각해보니 좋은 학교일수록 학생수가 무수히 많다. 그러므로 학생당 교수 비율이 적을수록 집 값이 높다는 인사이트를 얻었다.

흑인 인구 비율(B)이 낮을수록 집 값이 높을까?




```
2-Sample t-test
Ttest_indResult(statistic=6.953768186960278
5, pvalue=1.107587290213691e-11)
t:6.954
p:0.0
```

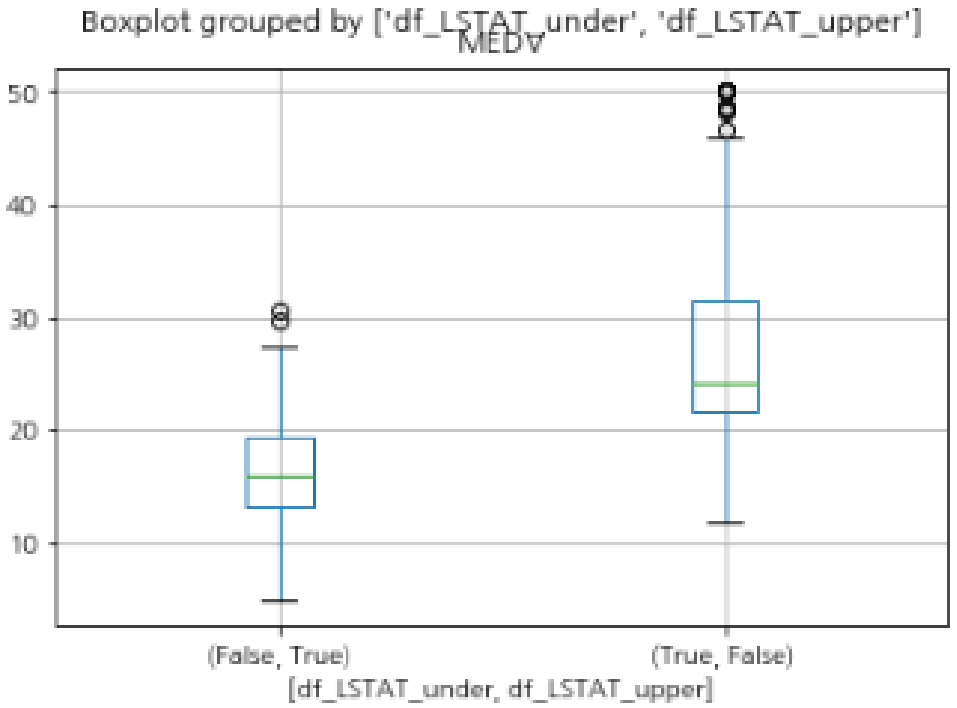
(흑인비율을 200 기준으로 두 부류로 나눈다.
)

귀무가설: 흑인이 많은 집단의 주택 가격 평균과 백인이 많은 집단의 주택가격 평균이 같다.

대립가설: 흑인이 많은 집단의 주택 가격 평균과 백인이 많은 집단의 주택가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 흑인이 많은 집단의 집 값이 더 비싸다는 편견을 깨는 인사이트를 얻었다.

저소득층 비율(LSTAT)이 낮을수록 집 값이 높을까?



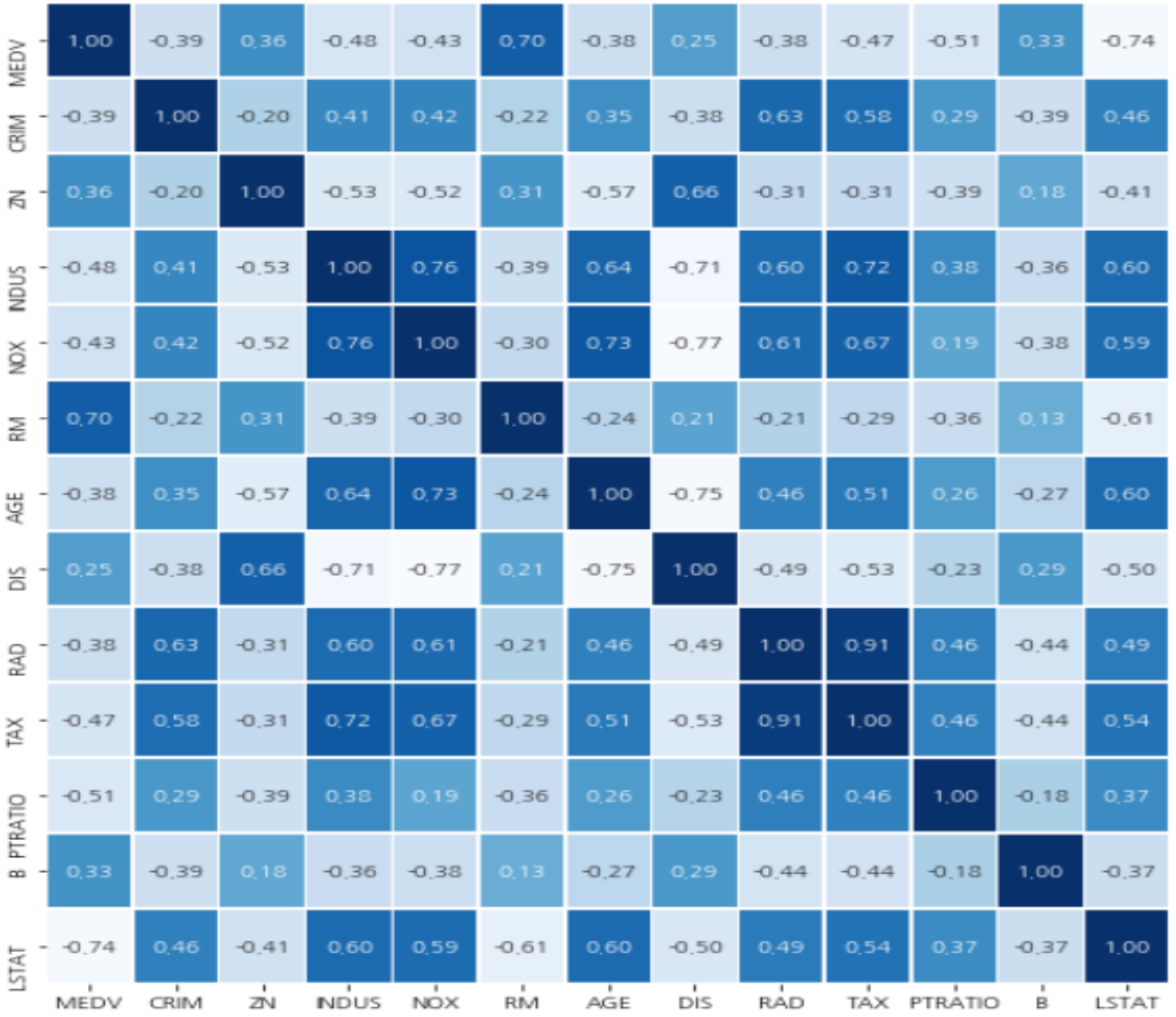

```
2-Sample t-test
Ttest_indResult(statistic=6.9537681869
602785, pvalue=1.107587290213691e-11)
t:6.954
p:0.0
```

(저소득층 비율을 13을 기준으로 두 부류로 나눈다.
)

귀무가설: 저소득층이 낮은 집단의 주택 가격 평균과 높은 집단의 주택가격 평균이 같다.
대립가설: 저소득층이 높은 집단의 주택 가격 평균과 낮은 집단의 주택가격 평균이 다르다.

→ **T-test**와 **BOX-Plot**을 통해, 저소득층 비율이 낮을수록 집 값이 더 비싸다는 인사이트를 얻었다.

각 변수 간의 상관관계는?



각 변수간 상관관계가 대체로 낮음

회귀분석

	variable	VIF
11	B	1.345
10	PTRATIO	1.783
1	CRIM	1.788
5	RM	1.932
2	ZN	2.298
12	LSTAT	2.931
6	AGE	3.093
3	INDUS	3.949
7	DIS	3.955
4	NOX	4.389
8	RAD	7.398
9	TAX	8.876
0	const	584.833

```
Durbin-Watson:                   1.011
Jarque-Bera (JB):              920.733
Prob(JB):                   1.16e-200
Cond. No.                    1.47e+04

Omnibus:                       192.780
Prob(Omnibus):                   0.000
Skew:                            1.632
Kurtosis:                        8.746
```

→ 다중공선성이 10이상인 것은 존재하지 않는다.

MEDV=36.6203+-0.1141CRIM+0.0457ZN
-16.4692NOX+3.8446RM-1.5261DIS+0.3155RAD
-0.0127TAX-0.9784PTRATIO+0.0097B-0.5281LSTAT

회귀분석을 통해 얻은 결론 요약

1) p 값이 0.05이상인 INDUS, AGE을 제거했으나 더빈왓슨이 1.011이기 때문에 잔차들은 양의 상관관계가 있다고 할 수 있다.

2)상관분석에서 값이 낮게 나온 변수인 "ZN"(0.36),"DIS"(0.25),"CRIM"(-0.388),"NOX"(-0.427),"INDUS"(-0.484)를 제거하고 다시 생각해본다.

3) "ZN"(0.36),"DIS"(0.25),"CRIM"(-0.388),"NOX"(-0.427),"INDUS"(-0.484)를 제거하여도 수정된 R-squaired 값은 늘어나지 않고, 줄어든다.

4) 가장 R-squaired가 가장 크도록 하는 변수는 CRIM, ZN, NOX, RM, DIS, RAD, TAX, PTRATIO, B, LSTAT이다.

5) 하지만, Durbin-Watson 값이 1.011이기 때문에 잔차가 독립적이지 않고, Prob(JB) 0.05 보다 작으므로 잔차의 정규성이 성립하지 않는다.

6) 변수를 줄이거나 제거하거나 scaling을 해도 잔차의 독립성, 정규성, 등분산성이 성립하지 않는다.

모델링 & 요약

7) 이 데이터는 잔차의 등분산성, 독립성, 정규성이 성립하지 않으므로 다중선형회귀 모델을 사용할 수 없다.

8) 이론적으로는 사용할 수 없다고 배우나 현업에서 종사하신 분들을 인터뷰한 결과, 회귀모델은 coefficient 값으로 각 변수가 끼치는 영향을 숫자로 알 수 있기 때문에 사용하기 좋은 모델링이라고 한다.

9) 이러한 이유로 잔차의 독립성, 정규성, 등분산성이 성립하지 않아도 현업에서는 사용하기도 한다.

변수 설명력 순서

NOX > RM > DIS > PTRATIO > LSTAT > RAD

모델링 & 요약

의사결정나무

max_depth=6, min_samples_leaf=1, min_samples_split=14를 선택하면

83.2%의 정확도를 보인다

랜덤포레스트

min_samples_leaf=1,max_depth=10 min_samples_split=4 를 선택하면

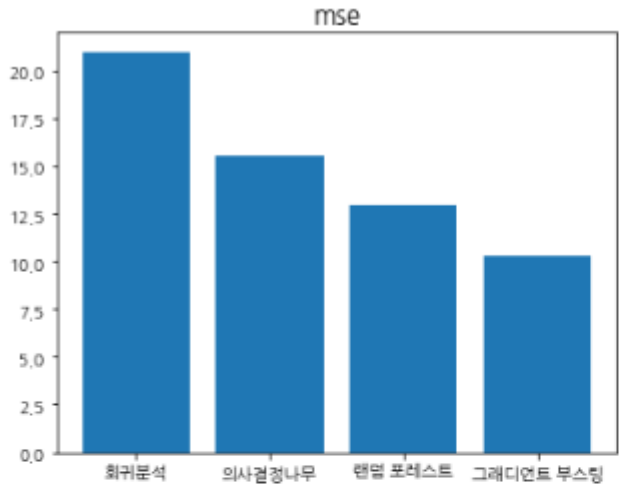
89.8 %의 정확도를 보인다

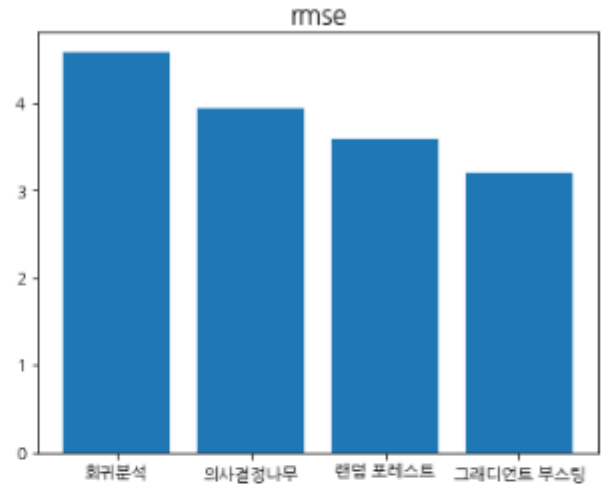
그레디언트부스팅

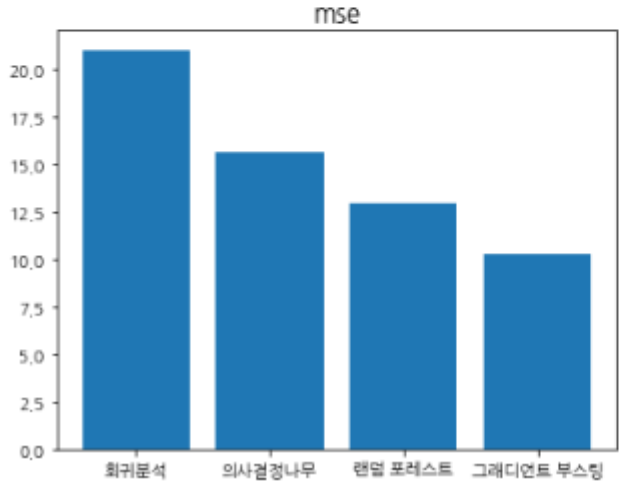
min_samples_leaf=2,max_depth=6 learning_rate=0.1을 선택하면

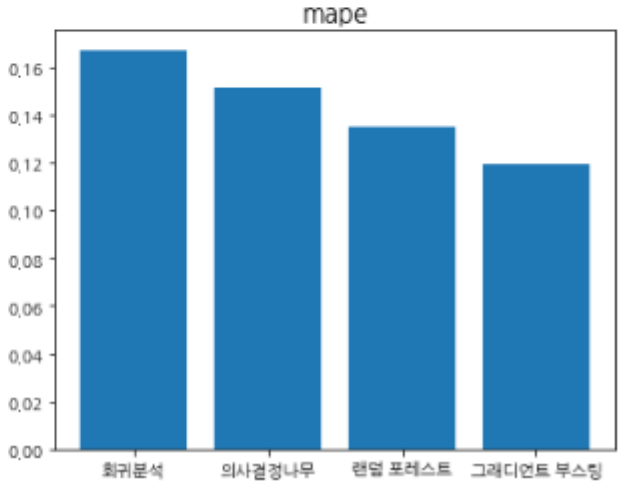
90%의 정확도를 보인다.

모델평가





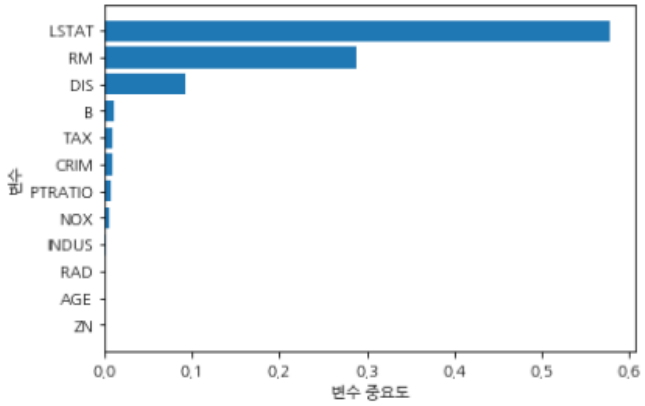




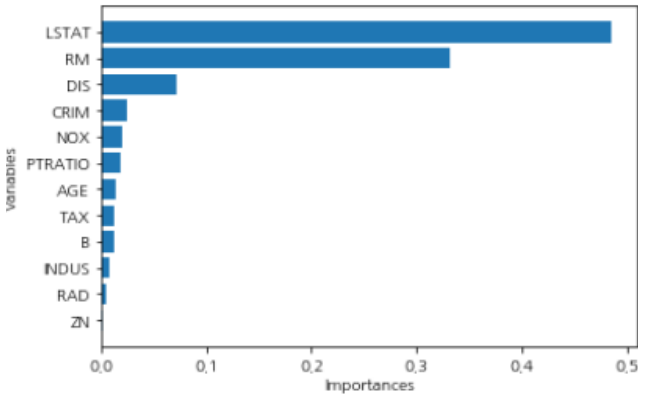
회귀분석 모델링
결과가 가장 좋음

모델링 & 요약

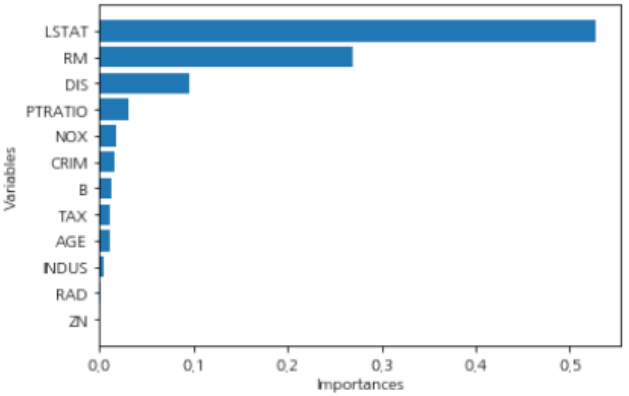
변수의 중요도



의사결정나무



랜덤포레스트



그레디언트부스팅

변수의 중요도 순서(세 모델링 모두 같은 순서가 3위권인 변수)
LSTAT -> RM -> DIS

개선방향 또는 결론

1, 보스턴 집 값에 대한 배경지식이 존재하지 않은 채로 과제 진행을 시작했다. 이 후 인사이트를 얻을 때, 나만의 의사결정을 하지 못한다는 것을 깨닳았다. 이 후, 다시 배경지식 공부부터 과제 진행을 시작했다. 다음부터는 시작할 때 배경지식을 알아보고 과제 수행을 해야겠다
.

2, 회귀분석은 잔차의 독립성, 등분산성, 정규성이 성립하지 않기에 사용할 수 없고, 나머지 모델링에서 중요 변수는 LSTAT, RM, DIS 순이다.
-> 따라서 1970년대 보스턴에서 집을 살 때, 다른 어떤 변수보다도 LSTAT, RM, DIS 에 유의하는 것이 좋을 것이다.

3,T-test 가설 검정을 모든 변수에 대해 두 그룹으로 나누고 진행했다. 교수님께서 말씀해주신 방법을 모든 변수에 적용했는데, 다음부터는 한 변수에 대해 양극화되는 경우에만 T-test 가설 검정을 사용할 것이다. 왜냐하면 정규분포, 양의 상관관계 등 양극화되지 않은 경우에 하는 T-test 검정은 무의미하기 때문이다.

실습 과정을 통해 배운점과 통찰, 아이디어, 애로사항

1) 처음 실습 과제가 주어졌을 때는 어떻게 가설 검정을 하고 가설을 세우는지 전혀 알지 못해서 문제가 생겼었다. 과제를 해결하며 이를 배웠다.

2) 상관관계가 높은 변수끼리 결합하여 파생변수를 만들면 더욱 좋은 모델이 될 수 있을 것이라는 생각이 들었다. 이를 하기 위해 부족한 점은 지금까지 BMI로만 파생변수를 구해봤는데, 이는 명확한 식이 있어서 어렵지 않았다. 그럼 명확한 식이 존재하지 않는 것은 어떻게 할 수 있을까?
→ 도출한 아이디어: 각 변수를 scale 한 후 각각 0.5를 곱해서 비율을 같게 더한다.

3) 회귀분석 모델링을 사용할 때 잔차의 독립성, 정규성, 등분산성이 성립하지 않아서 곤란했다 . 변수를 조정하며 재모델링을 했고 scale을 조정했음에도 불구하고 잔차의 독립성, 정규성, 등분산성이 성립하는 조합을 찾을 수 없었다.
→ 이를 통해, 데이터셋에 따라 사용할 수 없는 모델링도 존재한다는 것을 경험했다.

4) 모든 변수를 모델링에 대입하는 것이 맞는지에 대한 의문이 들었다.
→ 교수님께 여쭤봤더니 머신러닝 시대가 오기 전 데이터 분석 시대엔 EDA를 통해 얻은 인사이트를 모델링에 적용해서 필요하지 않다고 판단되는 변수는 제거하였다고 하셨다. 하지만, 머신러닝이 도래한 시대 이후엔 모든 변수를 모델링에 대입하여 결과를 도출한다고 한다. 그래서 이러한 조언에 따라 모델링을 진행했다.
'모델링에 모든 변수를 사용하는것일까?'라는 의문을 통해 또 한 번 성장을 한 계기였다.

5) 이 과제를 수행하며 CHAS 데이터가 불균형한 데이터이기에 모델링 전에 제거하는 것이 맞다고 생각했다. 이후, 불균형한 데이터를 처리하는 방법을 배웠다. 이 방법은 데이터 수가 극심히 적을 때 사용한다.
→ 숫자값으로 대체, 평균으로 대체

6) 데이터 분석가는 배경지식이 중요하다는 것을 알았다. 배경지식이 없으면 변수에 대한 정확한 결론을 짓기 어렵다.