주택 가격은 사회문제를 내포하는가?

B2 윤민기

과제 정의

보스턴 주택 가격 인자 분석은 고전적인 통계 연습 모델이다. 그 과정에서 수많은 주제의 분석이 이루어졌겠지만 사회가 내포하는 문제와 주택 중위 가격 사이의 관계성 파악은 드물었으리라 판단했다.

해당 데이터는 1978년에 발표된 보스턴을 배경으로 한 데이터로서 목표 변수인 주택 중위 가격(MEDV)과 이를 분석하기 위한 13개의 설명 변수로 이루어져 있다.

이 중 미국 사회의 고질적인 문제인 CRIM(범죄율), 흑인 인권 문제로 2020년 지금까지 불타오르는 B(흑인 비율), 교육 양극화를 의미할 수 있는 PTRATIO(학생당 교사 비율), 소득 양극화를 의미할 수 있는 LSTAT(저소득층 비율), 환경 문제를 의미할 수 있는 NOX(산화질소 농도)가 H1 대립가설인 '사회문제를 내포한다' 를 지지하는 주요 변수로 가정했다.

이를 한 줄로 설명하면 다음과 같다.
범죄율, 학생당 교사 비율, 저소득층 비율, 산화질소 농도가 낮을수록 주택 중위 가격은 높을 것이다.

우선 가정하지 않은 설명 변수에서 유의미한 추론이 나올 수 있으며 설명 변수가 적기 때문에 모든 변수를 탐색적 분석을 거친 후 4개의 설명 변수를 중점적으로 모델링한다.

분석 계획

1. BOSTON_HOUSING.csv 데이터 소개

2. 그래프 분석

3. 상관관계 분석

4. two-samples t-test 검정

5. 파생변수 사용 유무 확인

6. 회귀분석 및 모델링 진행

7. 모델 평가

8. 모델 개선안

데이터 현황 - 통계

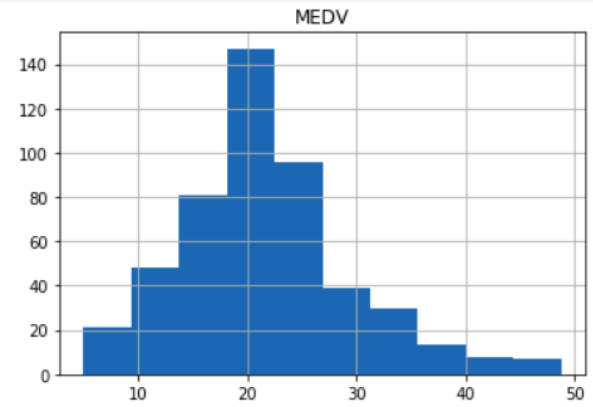
해당 데이터는 각 변수들이 무엇을 나타내는지, 단위는 무엇인지 등에 관한 설명이 불친절하다. 비율인지 개수인지 돈인지 구분은 데이터 분석에 핵심적으로 작용할 수 있기 때문에 이를 요약하면,

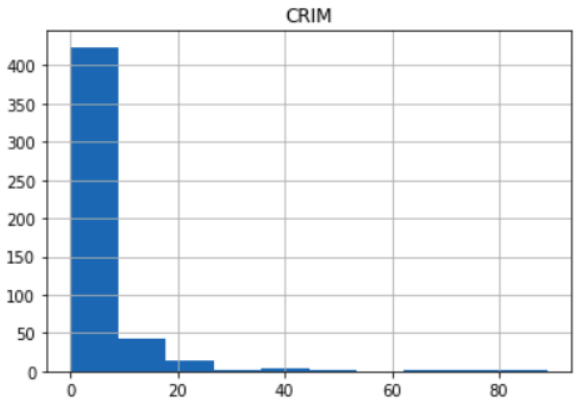
CRIM → 자치 시(town) 별 1인당 범죄율
ZN → 25,000 평방피트를 초과하는 거주지역의 비율
INDUS → 비소매상업지역이 점유하고 있는 토지의 비율
CHAS → 찰스강의 경계에 위치해 있으면 1, 그렇지 않으면 0
NOX → 10ppm당 농축 일산화질소
RM → 주택 1가구당 평균 방의 개수
AGE → 1940년 이전에 건축된 소유주택의 비율
DIS → 5개의 보스턴 직업센터까지의 접근성 지수
RAD → 방사형 도로까지의 접근성 지수
TAX → 10,000 달러 당 재산세율
PTRATIO → 자치 시(town)별 학생/교사 비율
B → 1000(Bk-0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함.
LSTAT → 모집단의 하위계층 비율(%)
MEDV → 본인 소유의 주택가격(중앙값) (단위: $1,000)

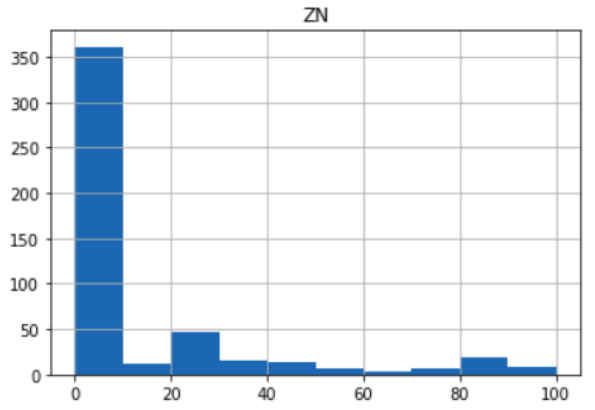
이를 참고해서 분석을 실행한다

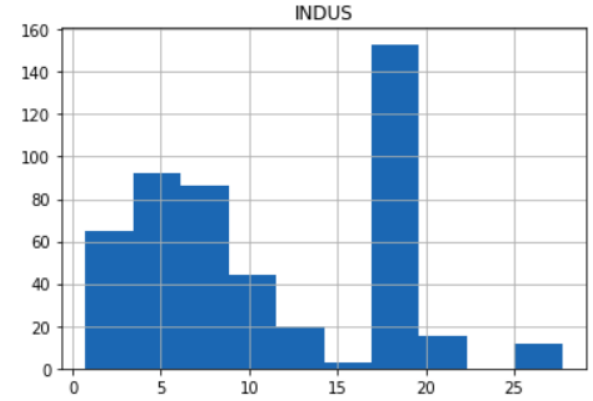
데이터 현황 - 통계

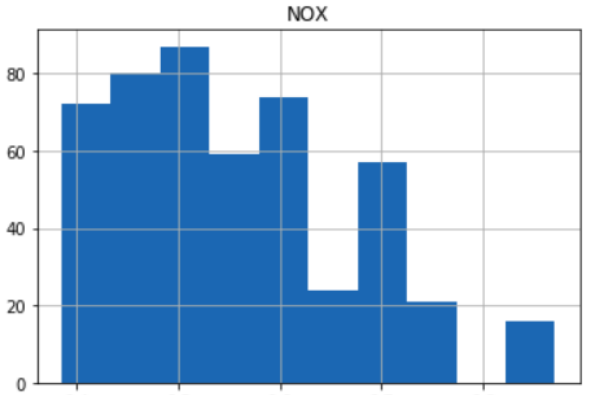
각 데이터의 분포는 다음과 같다.

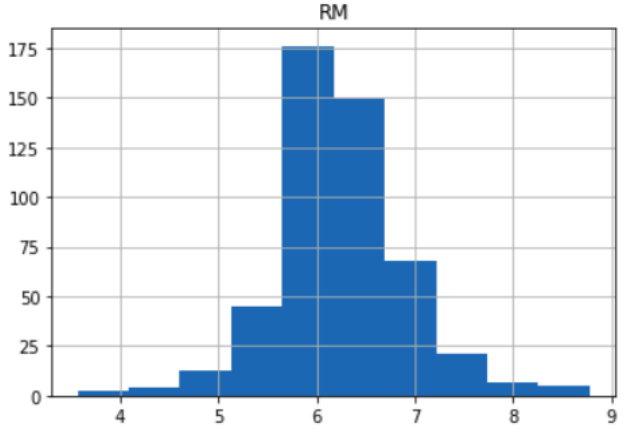


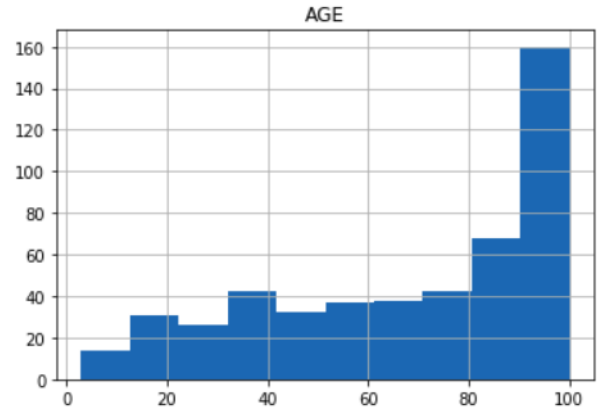


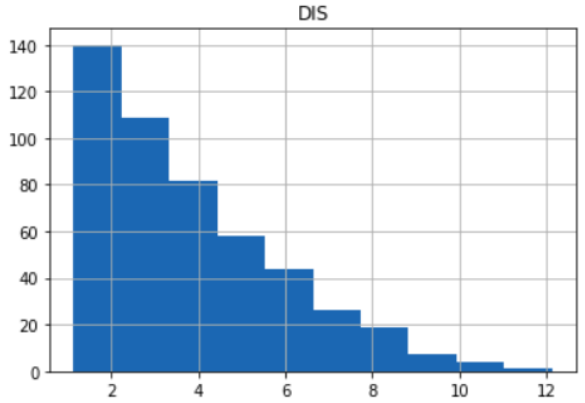


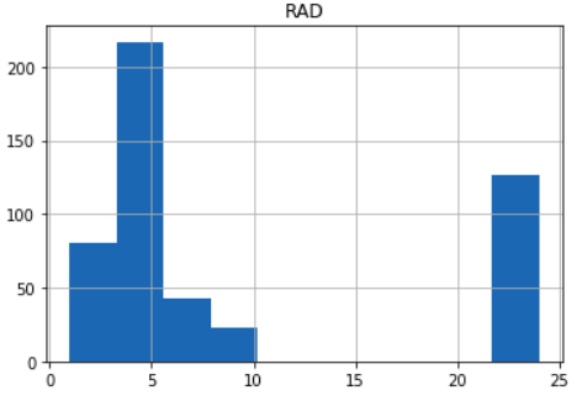






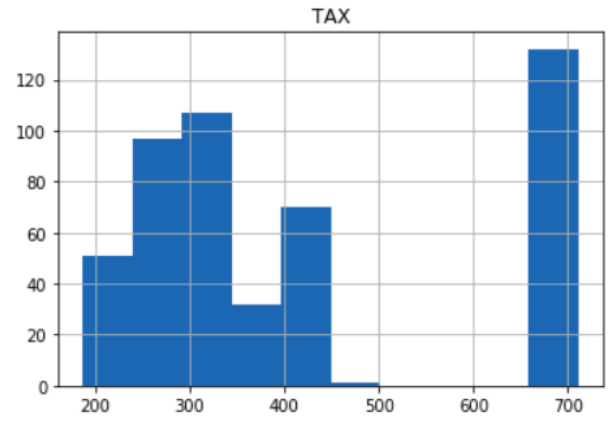


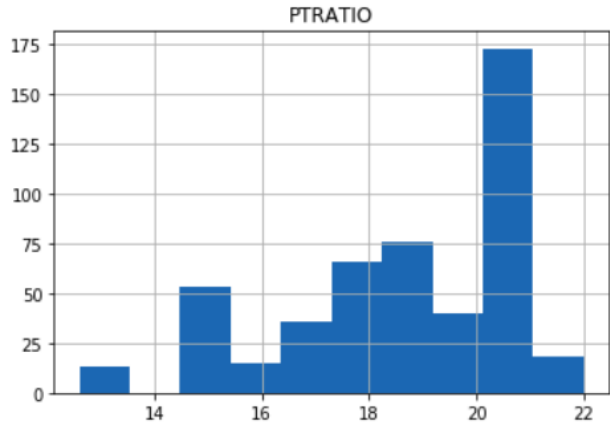


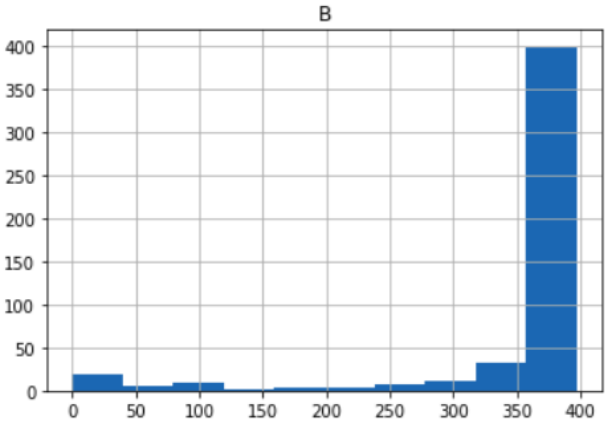


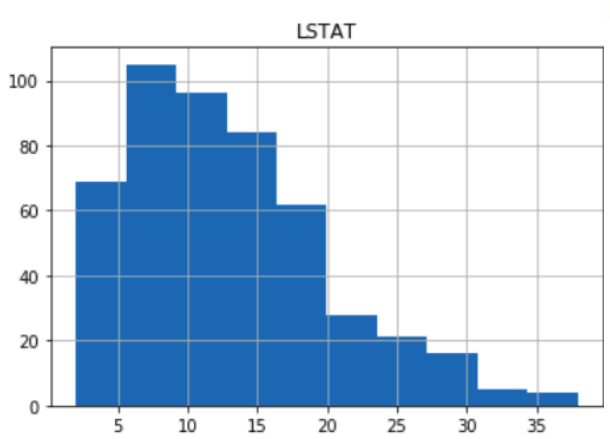
데이터 현황 - 통계

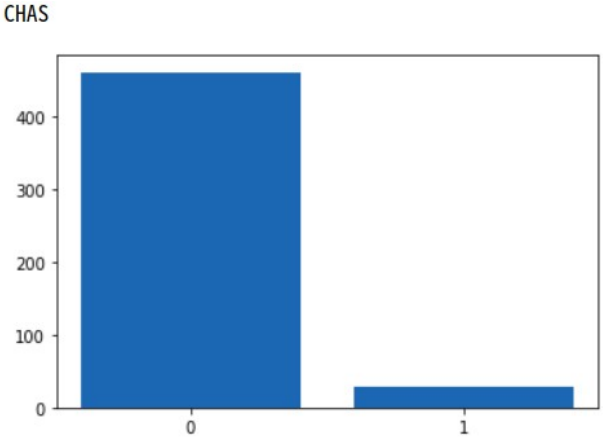
각 데이터의 분포는 다음과 같다.











데이터 현황

해당 데이터에서 수정해야 할 것이 몇 가지가 있다.

1. B(흑인 비율) 의 정확한 뜻을 알 수 없다. 이를 직관적으로 이해할 수 있는 비율로 변경해야 한다.
   (B 현재 값 = 1000(진짜비율-0.63)^2)

2. MEDV값 50 이상은 모두 50으로 처리되어 있다. 따라서 수가 16개 뿐이고 목표 변수로서의 정확도를 고려해 제거한다.

3. CRIM 은 1인당 범죄발생률을 의미한다. 해당 수치 25 이상은 평균 * 3시그마를 2배 이상 넘고, 수가 11개뿐임을 고려, 가정에서 중요한 변수로 설정한만큼 정확성을 위해 제거한다.

4. TAX 와 RAD 는 일정 수치 이상이 한 값에 몰리는 경우가 많지만, 그 수가 100개 이상으로 지웠을 때의 데이터 손실이 너무 커서 지우지 않고 해당 변수를 최소한으로 인용하는 방향으로 결정한다.

변수 B(수정된 지수) → B(비율), row 수 506 개 →479개

데이터 현황

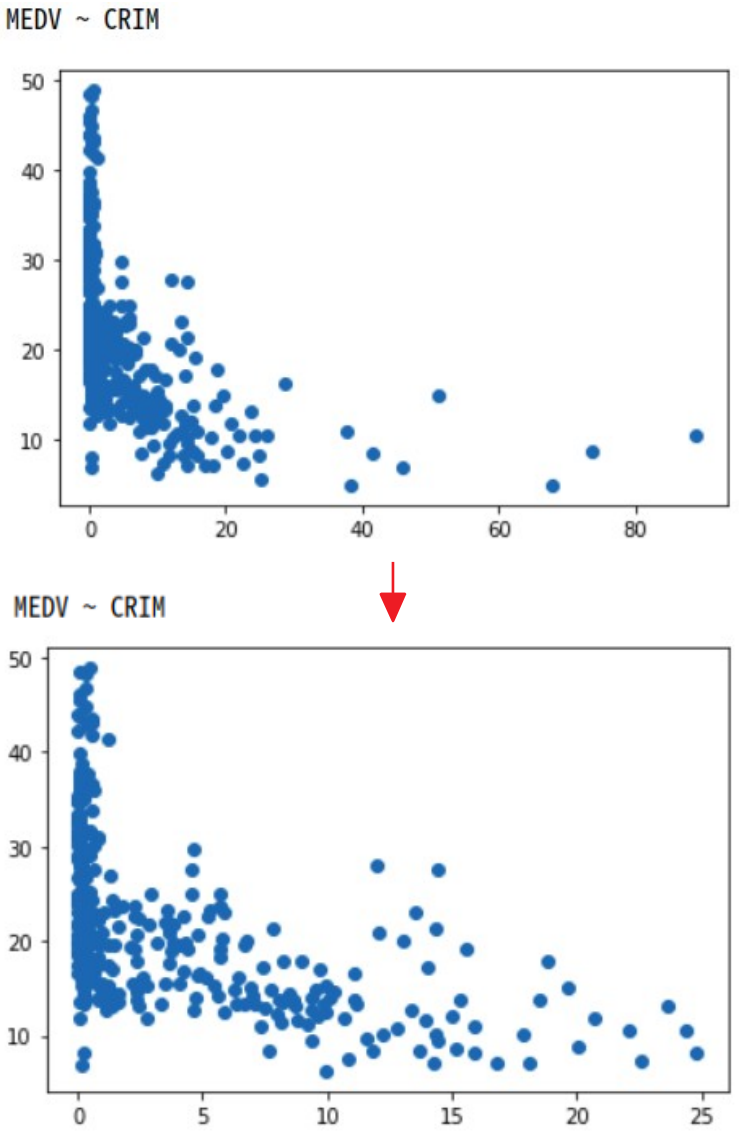
모든 변수들의 주택 중위 가격(이하 MEDV)을 각 설명 변수 별 y ~ x 꼴 그래프 분석을 진행 하였다.

그 결과 MEDV 가 50을 넘어가면 모두 이상치로 취급해 연속형 변수에서 50인 값이 16개가 등장했다.

전체 데이터셋에서 목표 변수가 가장 중요하므로 이상치를 모두 제거하는 방향으로 진행했다.

이에 더해 우측 상단의 그래프는 MEDV 이상치를 제거한 MEDV~ CRIM scatter graph 이다.

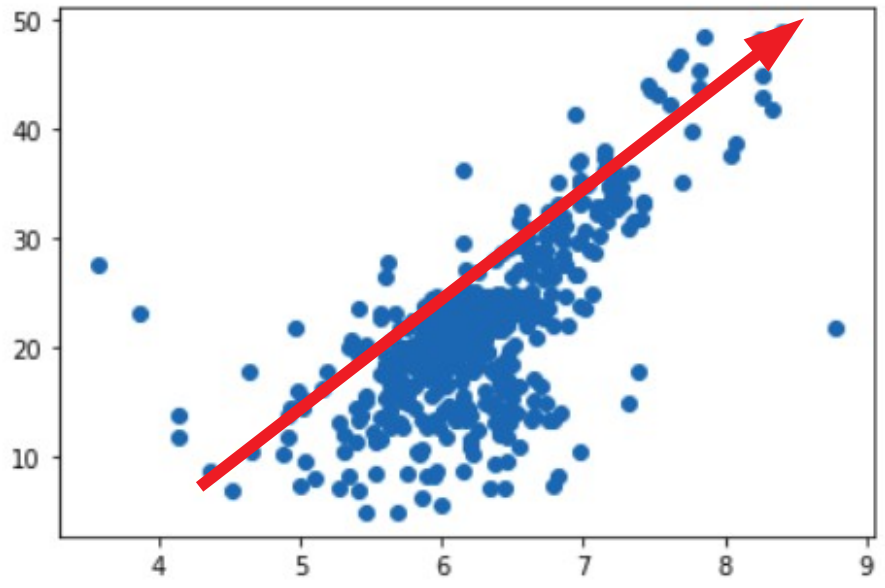
CRIM 은 1인당 범죄발생율을 뜻하는데 40 언저리를 넘어가는 숫자들은 지나치다고 판단, 이상치 처리를 진행하기로 했다. 보수적으로 CRIM 이 25 이상인 11 개의 데이터를 제거했다.



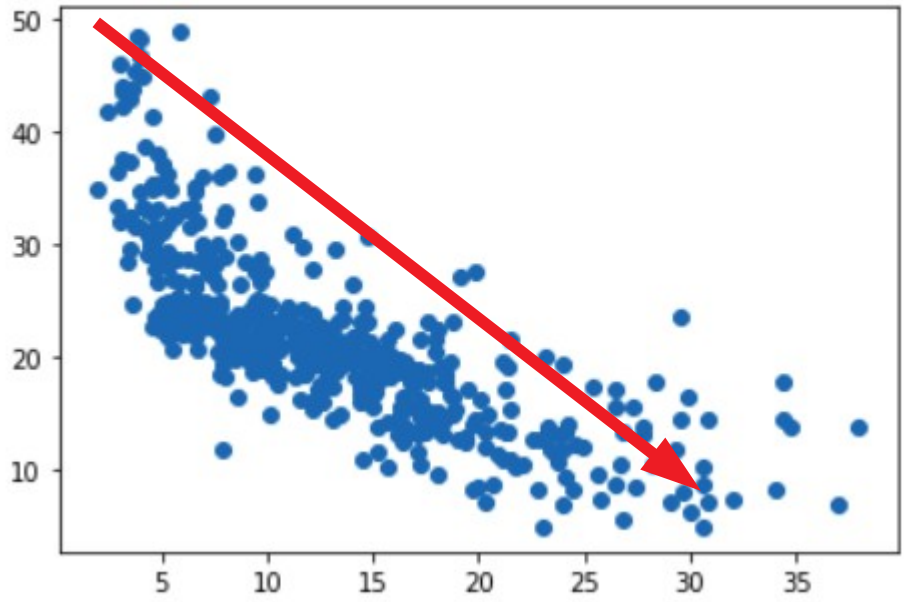
데이터 현황

목표변수 ~ 설명변수 그래프 분석 결과 뚜렷한 추세를 보이는 그래프는 MEDV ~ RM(주거당 평균 객실수) 와 MEDV ~ LSTAT(저소득층 비율) 뿐이었다. RM 은 우상향, LSTAT 는 우하향 경향을 보였다.

MEDV ~ RM



MEDV ~ LSTAT



이상치 처리가 된 데이터를 바탕으로 피어슨 상관계수를 확인하도록 한다

데이터 현황

	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
MEDV	1.0000	-0.5055	0.4006	-0.5883	0.0679	-0.5079	0.6887	-0.4784	0.3457	-0.4441	-0.5489	-0.5094	0.3397	-0.7509
CRIM	-0.5055	1.0000	-0.2628	0.5379	-0.0646	0.5663	-0.2717	0.4495	-0.4801	0.8243	0.7663	0.3604	-0.4128	0.5943
ZN	0.4006	-0.2628	1.0000	-0.5234	-0.0569	-0.5087	0.3073	-0.5603	0.6720	-0.2996	-0.2946	-0.3767	0.1572	-0.4225
INDUS	-0.5883	0.5379	-0.5234	1.0000	0.0423	0.7604	-0.4043	0.6302	-0.7029	0.5826	0.7094	0.3762	-0.3266	0.6331
CHAS	0.0679	-0.0646	-0.0569	0.0423	1.0000	0.0925	0.0408	0.0776	-0.0850	-0.0240	-0.0606	-0.1131	0.0396	0.0018
NOX	-0.5079	0.5663	-0.5087	0.7604	0.0925	1.0000	-0.3127	0.7234	-0.7634	0.6009	0.6584	0.1737	-0.3506	0.6045
RM	0.6887	-0.2717	0.3073	-0.4043	0.0408	-0.3127	1.0000	-0.2554	0.2321	-0.1713	-0.2634	-0.2847	0.0707	-0.6006
AGE	-0.4784	0.4495	-0.5603	0.6302	0.0776	0.7234	-0.2554	1.0000	-0.7375	0.4343	0.4847	0.2558	-0.2440	0.6351
DIS	0.3457	-0.4801	0.6720	-0.7029	-0.0850	-0.7634	0.2321	-0.7375	1.0000	-0.4725	-0.5152	-0.2320	0.2646	-0.5273
RAD	-0.4441	0.8243	-0.2996	0.5826	-0.0240	0.6009	-0.1713	0.4343	-0.4725	1.0000	0.9035	0.4430	-0.4270	0.4882
TAX	-0.5489	0.7663	-0.2946	0.7094	-0.0606	0.6584	-0.2634	0.4847	-0.5152	0.9035	1.0000	0.4393	-0.4179	0.5503
PTRATIO	-0.5094	0.3604	-0.3767	0.3762	-0.1131	0.1737	-0.2847	0.2558	-0.2320	0.4430	0.4393	1.0000	-0.1647	0.3468
B	0.3397	-0.4128	0.1572	-0.3266	0.0396	-0.3506	0.0707	-0.2440	0.2646	-0.4270	-0.4179	-0.1647	1.0000	-0.3445
LSTAT	-0.7509	0.5943	-0.4225	0.6331	0.0018	0.6045	-0.6006	0.6351	-0.5273	0.4882	0.5503	0.3468	-0.3445	1.0000

상관계수 분석 결과 INDUS, RM, NOX, TAX, PTRATIO, LSTAT 이 MEDV와 |0.5| 이상의 관련성이 있음을 확인할 수 있다.

추가로 CRIM, TAX, RAD 가 서로 높은 관련성을 가지고 있음을 확인할 수 있다.

처음에 해당 주제를 설명하기 위한 변수로서 B를 선정했다. 그러나 데이터 변환을 거쳐 B를 확인한 결과 1.26% 만 넘어도 모든 값들이 1.26%로 처리됨을 확인할 수 있었다.

무려 전체 479개 열 중 420개가 그러했다. 이는 데이터 분석에서 의미 없는 변수가 되었지만 이 자체로 해당 주제였던 '주택 가격의 사회 문제 내포'의 어느 정도의 반증이라고 볼 수도 있다. 겨우 주택 가격을 알기 위한 데이터임에도 흑인 비율은 미미한 수준 이상 관측 하지 않으며 타 인종은 언급조차 없다.

바로 앞 장에서 언급한 TAX, RAD, CRIM 의 상관성은 TAX 와 RAD 의 편향성 때문에 큰 의미가 없으리라 판단해서 진행하지 않는다.

다중선형회귀분석, DecisionTree, RandomForest, GradientBoosting 을 진행함에 있어서,
모든 설명 변수로 진행한 모델링 vs 4개의 설명 변수로 진행한 모델링, 형식으로 구성한다.

Score 나 설명력 모두 크게 감소할 것이 분명하나 해당 설명 변수로 얼마나 주택 가격을 설명할 수 있는지 보기 위함이다.

다중 선형 회귀 분석 전체 진행 후
p-value 유의값들로만 진행했으나 Adj R-squared가 거의 동일하였다.

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.772
Model:	OLS	**Adj. R-squared:**	0.765
Method:	Least Squares	**F-statistic:**	120.9
Date:	Wed, 16 Sep 2020	**Prob (F-statistic):**	6.46e-140
Time:	00:37:49	**Log-Likelihood:**	-1303.7
No. Observations:	479	**AIC:**	2635.
Df Residuals:	465	**BIC:**	2694.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]	
Intercept	23.1795	4.841	4.788	0.000	13.666	32.693	
C(CHAS)[T.1]	0.3579	0.739	0.484	0.629	-1.095	1.811	제거
CRIM	-0.1563	0.070	-2.224	0.027	-0.294	-0.018	
ZN	0.0355	0.011	3.142	0.002	0.013	0.058	
INDUS	-0.0452	0.050	-0.911	0.363	-0.143	0.052	제거
NOX	-11.6673	3.057	-3.817	0.000	-17.674	-5.661	
RM	3.9645	0.362	10.938	0.000	3.252	4.677	
AGE	-0.0263	0.011	-2.461	0.014	-0.047	-0.005	
DIS	-1.2137	0.161	-7.526	0.000	-1.531	-0.897	
RAD	0.2716	0.059	4.612	0.000	0.156	0.387	
TAX	-0.0138	0.003	-4.618	0.000	-0.020	-0.008	
PTRATIO	-0.8244	0.105	-7.869	0.000	-1.030	-0.619	
B	8.0473	1.797	4.478	0.000	4.516	11.579	
LSTAT	-0.3305	0.045	-7.317	0.000	-0.419	-0.242	

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.771
Model:	OLS	**Adj. R-squared:**	0.766
Method:	Least Squares	**F-statistic:**	143.1
Date:	Wed, 16 Sep 2020	**Prob (F-statistic):**	7.45e-142
Time:	00:55:13	**Log-Likelihood:**	-326.39
No. Observations:	479	**AIC:**	676.8
Df Residuals:	467	**BIC:**	726.8
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.258e-16	0.022	-5.68e-15	1.000	-0.043	0.043
CRIM	-0.0963	0.043	-2.220	0.027	-0.181	-0.011
ZN	0.1087	0.034	3.237	0.001	0.043	0.175
AGE	-0.0948	0.039	-2.433	0.015	-0.171	-0.018
DIS	-0.3234	0.043	-7.533	0.000	-0.408	-0.239
RAD	0.3148	0.063	4.986	0.000	0.191	0.439
B	0.1145	0.025	4.525	0.000	0.065	0.164
NOX	-0.1865	0.044	-4.219	0.000	-0.273	-0.100
RM	0.3366	0.030	11.109	0.000	0.277	0.396
TAX	-0.3221	0.058	-5.578	0.000	-0.436	-0.209
PTRATIO	-0.2314	0.028	-8.157	0.000	-0.287	-0.176
LSTAT	-0.2998	0.040	-7.476	0.000	-0.379	-0.221

MEDV ~ B + NOX + CRIM + PTRATIO 다중회귀분석 결과는 다음과 같다.

OLS Regression Results

Dep. Variable:	MEDV	**R-squared:**	0.468
Model:	OLS	**Adj. R-squared:**	0.463
Method:	Least Squares	**F-statistic:**	104.0
Date:	Wed, 16 Sep 2020	**Prob (F-statistic):**	1.51e-63
Time:	02:49:44	**Log-Likelihood:**	-528.73
No. Observations:	479	**AIC:**	1067.
Df Residuals:	474	**BIC:**	1088.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.258e-16	0.034	-3.75e-15	1.000	-0.066	0.066
CRIM	-0.1390	0.044	-3.128	0.002	-0.226	-0.052
B	0.1047	0.037	2.810	0.005	0.031	0.178
NOX	-0.3255	0.041	-7.899	0.000	-0.406	-0.245
PTRATIO	-0.3855	0.036	-10.718	0.000	-0.456	-0.315

Omnibus:	86.202	**Durbin-Watson:**	0.971
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	152.965
Skew:	1.047	**Prob(JB):**	6.08e-34
Kurtosis:	4.811	**Cond. No.**	2.29

Adj R-squared 가 0.463 으로 0.5가 되지 않는 것을 확인할 수 있다.

해당 분석에서 제시하는 coef 를 활용했을 때,

MEDV = -0.1390 * CRIM + 0.1047 * B + -0.3255 * NOX + -0.3855 * PTRATIO

식으로 MEDV 를 표현하는 것을 확인할 수 있다.

해당 분석은 목표 변수를 설명하는 것에 부족하다.

RandomForest → 모든 설명 변수 test_score: 81.38%
→ CRIM, NOX, PTRATIO, B 설명 변수 test_score: 52.66%

DecisionTree → 모든 설명 변수 test_score: 73.45%
→ CRIM, NOX, PTRATIO, B 설명 변수 test_score: 53.17%

Gradient Boosting → 모든 설명 변수 test_score: 85.97%
→ CRIM, NOX, PTRATIO, B 설명 변수 test_score: 61.86%

13개의 설명 변수와 비교했을 때, 많게는 30% 포인트, 적게는 24% 포인트 까지 차이가 발생하는 것을 확인할 수 있다.

데이터가 한정적임을 감안하더라도, 해당 설명변수로 목표변수를 설명하기는 한계가 분명하다.

초기 가설 설정 단계에서 MEDV ~ NOX + B + PTRATIO + CRIM 이 해당 데이터를 통해서는 유의미한 결론을 낼 수 없다는 것을 알 수 있었다.

그 이유를 분석하면
첫째, B 의 값의 87/28% 가 데이터 자체에서 이상치로 1.26% 지점에 몰려있다는 점

둘째, TAX + RAD + CRIM 으로 새로운 해석을 하고자 했으나 TAX 와 RAD 도 마찬가지로 왜곡된 데이터로 인해 파생변수 생성에 어떠한 인사이트도 줄 수 없었다는 점

셋째, 네 가지 변수에 너무 치우친 나머지 가장 상관계수가 높았던 LSTAT 과 RM 을 정작 외면했다는 점

넷째, 데이터가 왜곡되어 있다보니 regression 알고리즘 들도 충분한 설명력을 가지는 선형 방정식을 만들어내기 어려웠다는 점

등 여러 면에서 해당 과제의 목적은 달성되기 어려웠다.

개선방향 또는 결론

분석 결과가 주택 가격은 사회 문제를 반영한다 의 귀무가설 채택, 즉 주택 가격이 사회 문제를 반영한다고 할 수 없다, 기 때문에 생산적인 결과 문항을 창출하기가 어렵다.

그러나 해당 데이터만 놓고 보았을 때 1.26%가 최대지만 흑인 비율이 주택 가격에 미치는 영향력은 거의 없다는 점, NOX 역시 상관계수가 0.5 는 넘었지만 해당 모델의 설명계수로서 미치는 영향은 미미했다는 점,

오히려 거의 모든 설명력을 LSTAT 와 RM 이 담당했다는 점을 봤을 때 주택 중위 가격에 결정적으로 영향을 미치는 용인은 물리적인 집 공간의 수 및 크기와 원인과 결과가 뒤바뀌었다는 인상은 있지만 저소득층 비율이라고 할 수 있다.

해당 데이터를 가지고 이 두 설명 변수 없이 유의미한 결과를 창출하기란 어렵다는 결론이 나온다.

데이터는 시대를 반영하기 때문에 어떤 아이디어, 혹은 연구해보고 싶은 주제가 있더라도 그 시대성에서 해당 연구를 위한 자료가 부족하다면 주어진 데이터로 한계가 있다는 것을 깨달았다. 더 적극적으로 외부 정보를 활용했어야 하나, 하는 아쉬움이 있다.

혹여 초기의 가정이 결코 될 수 없는 방향으로 가고 있더라도 그쪽으로 계속 나간 후 결과를 보는 것이 맞다고 생각했었는데, 막상 실망스러운 결과를 마주하니 실험적인 주제 보다는 기존에 교재에 실려있듯이 안정적인 주제로 분석을 진행했어야 하는 아쉬움도 남는다.

애로사항은 본인의 실력 부족이 애로사항이었으며 데이터 분석 및 리포트 작성은 정말 많은 노력과 에너지, 그리고 그것이 잘 드러나지 않는다는 아쉬움이 필요한 작업임을 깨달았다.