보스턴 외곽지역에 거주하는 저소득층 흑인 비율과 보스턴 주택 가격의 관계

KAGGLE BOSTON HOUSING

B반 4조 이정하

과제 정의

1970년대 중반 보스턴 외곽지역의 주택 가격에 영향을 미치는 인자가
다양한 예측 모델을 이용하여 주택 가격에 영향을 주는지 알아내고
선정한 영향 인자를 활용하여 집값을 예측하는 과제.

분석 배경

1970년대 중반까지 미국은 **저소득층이 입주할 수 있는**
공영 주택을 짓는 프로젝트를 국가적으로 진행하고 있었다.

해당 데이터가 미국 내 인종차별이 심했던 1970년대 중반에 수집되었다는 점,
그리고 공영 주택 프로젝트가 중앙정부 주도적으로 저소득층의
주택을 건설해주는 방향으로 진행되었다는 점에 중점을 두었다.

이를 통해 아래와 같은 예측을 할 수 있었다.

- **1970년대에 인종차별이 심했기 때문에 흑인의 재산(수입)이 적을 것이다.**
- **미국 내 저소득층을 인종을 기준으로 나눈다면 흑인이 많을 것이다.**
- **그렇기 때문에 공영 주택에 저소득층 흑인이 많이 입주했을 것이다.**
- **보스턴 외곽에 위치해 있을수록 주택 가격이 낮을 것이다.**
- **따라서 저소득층 흑인은 보스턴 외곽에 거주할 것이며, 주택 가격은 낮을 것이다.**

위와 같은 생각을 바탕으로
주거지 비율, 중심지 접근 거리, 저소득층 비율, 흑인 인구 비율을 중심으로
각 인자가 보스턴 주택 가격과 어떤 관련성을 가지고 있는지 알아보기로 했다.

가설 설정

보스턴 외곽지역에 거주하는 저소득층 흑인 비율이 높을수록
보스턴 주택 가격이 낮을 것이다.

가설에 따른 분석 변수 선정

'보스턴 외곽지역에 거주하는 저소득층 흑인 비율'을 분석해야 하기 때문에
다음과 같은 4개의 변수를 분석 대상으로 선정했다.

1. **주거지 비율 = ZN:** 2500 평방미터를 초과하는 거주지역의 비율
2. **중심지(노동 센터) 접근 거리 = DIS:** 보스턴 직업센터까지의 접근성 지수
3. **흑인 인구 비율 = B:** 자치시별 흑인의 비율
4. **저소득층 비율 = LSTAT:** 모집단의 하위계층 비율

분석 계획

1. 데이터 현황
2. 탐색적 분석

   boxplot으로 분포 파악, 사용변수 확정
3. 모델링 적용

   Gradient Boosting으로 설명변수 중요도 분석
4. 분석 결론
5. 분석을 통해 깨달은 점

1. 데이터 현황

[기술통계량 확인]

	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
count	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000	506.000
mean	22.533	3.614	11.364	11.137	0.069	0.555	6.285	68.575	3.795	9.549	408.237	18.456	356.674	12.653
std	9.197	8.602	23.322	6.860	0.254	0.116	0.703	28.149	2.106	8.707	168.537	2.165	91.295	7.141
min	5.000	0.006	0.000	0.460	0.000	0.385	3.561	2.900	1.130	1.000	187.000	12.600	0.320	1.730
25%	17.025	0.082	0.000	5.190	0.000	0.449	5.886	45.025	2.100	4.000	279.000	17.400	375.377	6.950
50%	21.200	0.257	0.000	9.690	0.000	0.538	6.209	77.500	3.207	5.000	330.000	19.050	391.440	11.360
75%	25.000	3.677	12.500	18.100	0.000	0.624	6.623	94.075	5.188	24.000	666.000	20.200	396.225	16.955
max	50.000	88.976	100.000	27.740	1.000	0.871	8.780	100.000	12.127	24.000	711.000	22.000	396.900	37.970

보스턴의 주택 가격에 영향을 미치는 영향 인자 데이터를 불러온다.
해당 과제에서는 목표변수가 보스턴 주택 가격(MEDV)이다.

그리고 강 조망 여부(CHAS)를 제외한 모든 변수는 연속형 데이터이다.
CHAS 변수는 강 조망이 있으면 1, 없으면 0으로 나타난다.

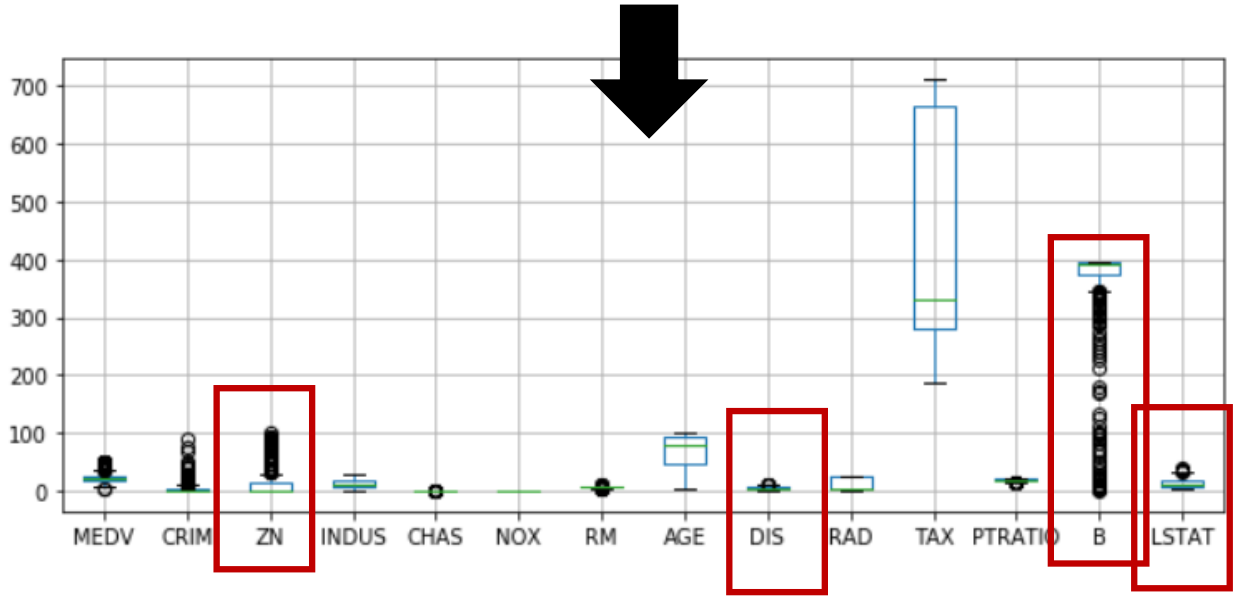
```
MEDV        0
CRIM        0
ZN          0
INDUS       0
CHAS        0
NOX         0
RM          0
AGE         0
DIS         0
RAD         0
TAX         0
PTRATIO     0
B           0
LSTAT       0
dtype: int64
```

[결측치 확인]

해당 데이터에 결측치가 있는지 판단한다.
결측치가 없으므로 데이터 변환 없이 그대로 진행한다.

[이상치 확인]

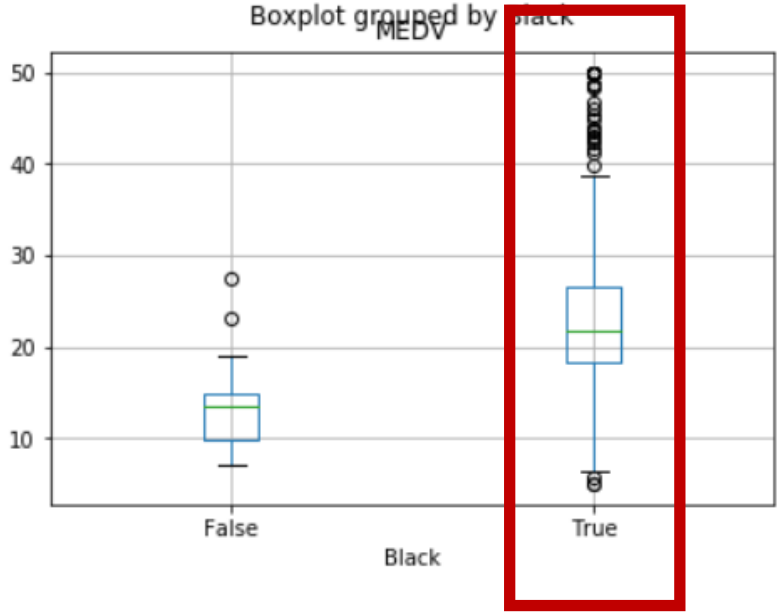
BOX PLOT 을 통해서 해당 데이터에 이상치가 있는지 판단한 결과,
ZN, DIS, B, LSTAT 변수에서 이상치가 확연히 존재함을 확인했다.
이상치가 나온 이유에 대한 원인 파악이 중요하다고 판단했다.



2. 탐색적 분석

(1) Boxplot을 통해 데이터의 분포, 관계 파악

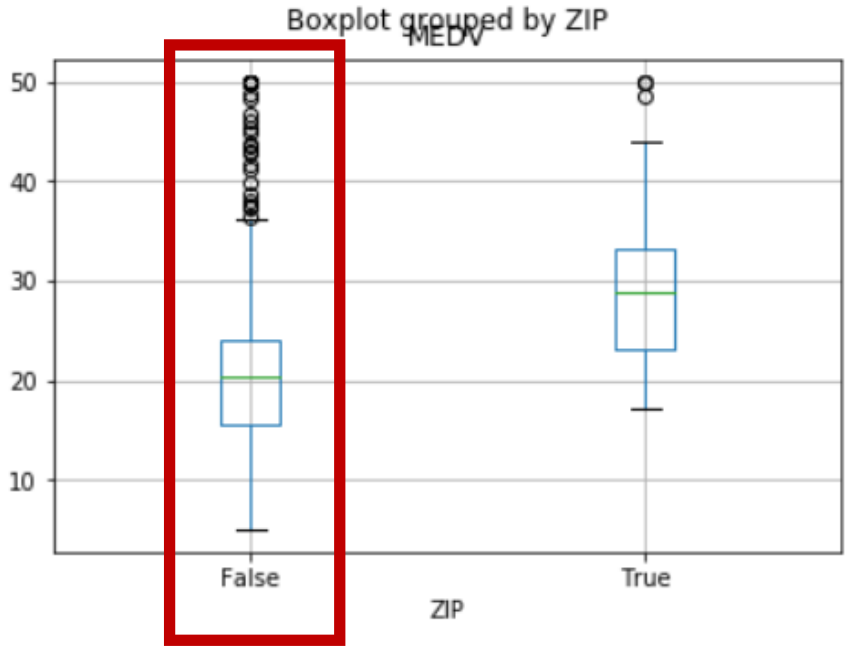
* 흑인 인구 비율과 보스턴 주택 가격의 관계



흑인 비율이 350 이상인 지역에서는
주택 가격이 월등히 높게 나타난다.

이를 통해 기존의 편견과 다르게
흑인 인구 비율이 높을수록 보스턴 주택 가격은 높게 나타난다는 사실을 알 수 있었다.

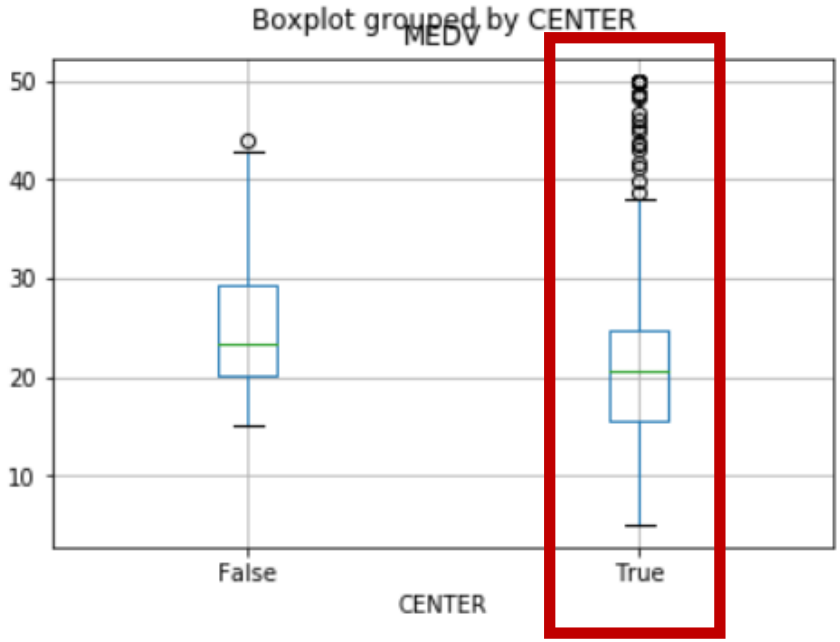
* 주거지 비율과 보스턴 주택 가격의 관계



주거지 비율이 30 이상인 지역에서 주택 가격 평균이 조금 더 높으나,
주거지 비율이 30 이하인 지역에서는 이상치가 의미 있게 많이 나타났다.

따라서 주거지 비율이 낮은 지역에서 주택 가격의 격차가 큰 것으로 판단했으며,
주거지 비율이 낮은 지역에서도 부동산 가격이 높은 경우가 많다고 판단했다.

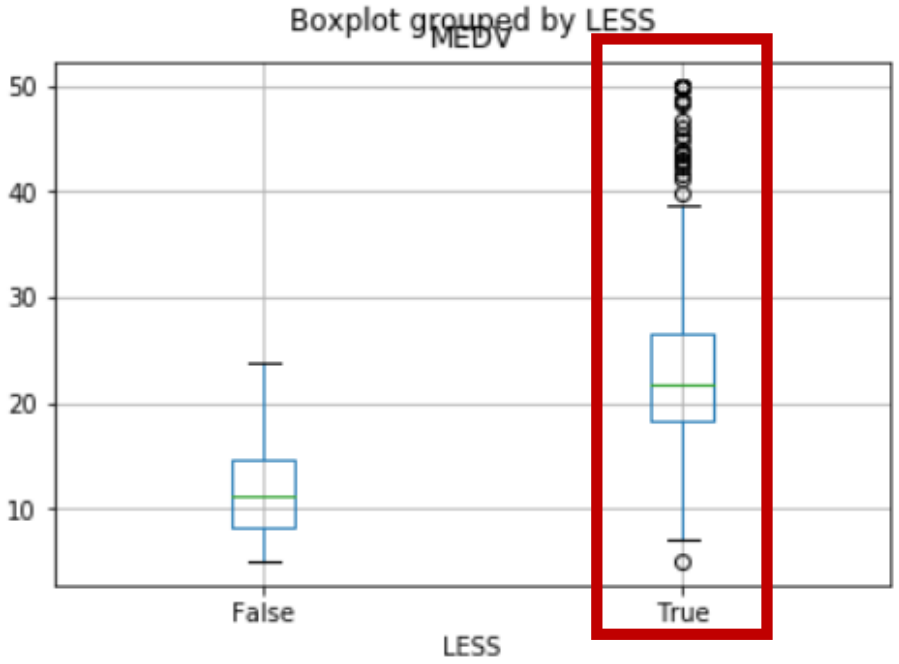
* 중심지(노동센터) 접근 거리와 보스턴 주택 가격의 관계



중심지(노동센터)와의 접근거리가 6 이하일수록 주택 가격 평균이 조금 더 낮으나,
이상치가 의미 있게 많이 나와 중심지와의 접근거리가 가까울수록
주택 가격의 편차가 심하다고 판단했다.

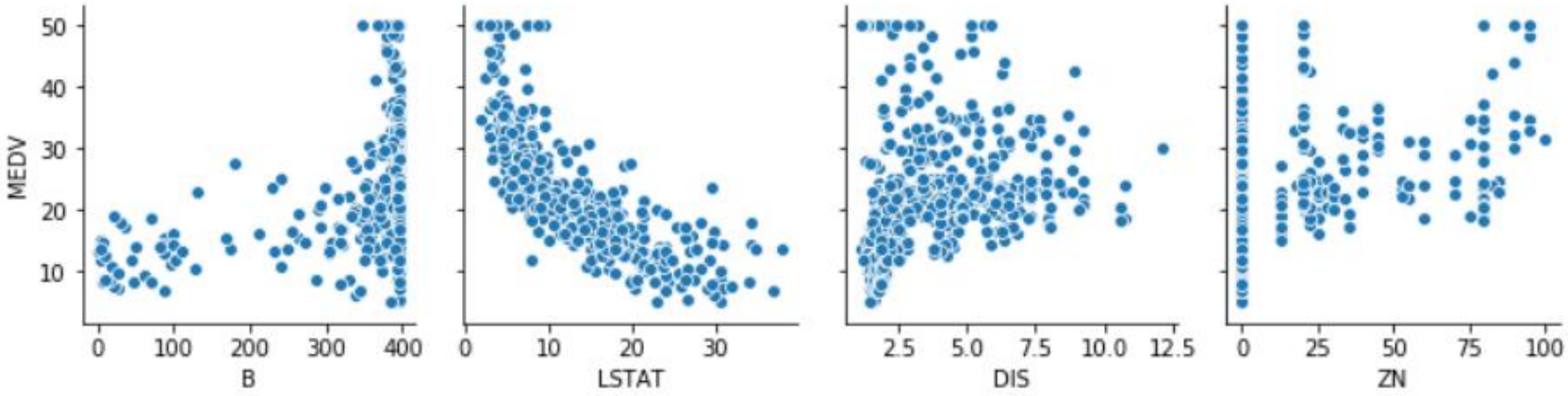
그러나 접근거리가 가까운 경우에도 주택 가격이 높은 경우가 많이 나와
이 부분에 대한 추가적인 검토가 필요하다고 생각했다.

* 저소득층 비율과 보스턴 주택 가격의 관계



저소득층 비율이 25보다 낮을수록 부동산 가격이 월등히 높게 나왔다.
따라서 저소득층 비율이 높으면 부동산 가격이 낮다는 결론을 내릴 수 있었다.

(2) Scatter Plot을 통해 변수 간 관계 파악



중심지 접근거리(DIS) 변수를 제외하고
흑인 인구 비율(B), 저소득층 비율(LSTAT),
중심지 접근거리(DIS), 주거지 비율(ZN) 변수로
Scatter Plot 을 그린 결과, 각 변수가 목표변수 **보스턴 주택 가격(MEDV)**과
다음과 같은 관계를 가지고 있음을 알 수 있었다.

1) 흑인 인구 비율이 높을수록 주택 가격이 높다.
2) 저소득층 비율이 높을수록 주택 가격이 낮다.
3) 중심지와의 거리가 가까울수록 주택 가격의 편차가 크다.
4) 주거지 비율이 낮을수록 주택 가격의 편차가 크다.

3. 모델링 적용

선택한 모델: Gradient Boosting

잔여 오차를 지속적으로 개선하기 때문에 점진적으로 오차를 보완할 수 있다.

	Feature	Importance
12	LSTAT	0.535
5	RM	0.266
7	DIS	0.100
4	NOX	0.023
11	B	0.016
10	PTRATIO	0.015
0	CRIM	0.013
6	AGE	0.011
9	TAX	0.010
2	INDUS	0.007
8	RAD	0.003
3	CHAS	0.001
1	ZN	0.000

학습 데이터와 테스트 데이터의 **설명력**

Score on train set: 0.987

Score on test set: 0.907

```
Score on training set: 0.987
Score on test set: 0.907
```

설명변수 **중요도**

1위: LSTAT(저소득층 비율)
2위: RM(주거당 평균 객실 수)
3위: DIS(중심지와의 접근거리)순으로 크게 나왔다.

따라서 처음에 세운 가설에 필요한 변수였던
B(흑인 인구 비율), ZN(주거지 비율) 변수의 경우,
추가 검토가 더 필요해 보였다.

	Feature	Importance
12	LSTAT	0.535
5	RM	0.266
7	DIS	0.100
4	NOX	0.023
11	B	0.016
10	PTRATIO	0.015
0	CRIM	0.013
6	AGE	0.011
9	TAX	0.010
2	INDUS	0.007
8	RAD	0.003
3	CHAS	0.001
1	ZN	0.000

설명변수 중요도 확인을 통한 변수 재선정

1) 설명변수 중요도가 0.000 이 나온 ZN(주거지 비율)은 DIS(중심지와의 접근거리)로 대체 가능하므로, 선정대상에서 삭제한다.

2) LSTAT(저소득층 비율), DIS(중심지와의 접근거리)는 높은 설명변수 중요도를 가지고 있으므로 채택한다.

3) B(흑인 인구 비율)의 설명변수 중요도는 낮으나 가설에서 없어서는 안 되는 특징이므로 채택한다.

모델링을 통해 얻은 통찰(Insights)

1) 기존 가설에서 중요하게 생각했던
흑인 비율의 설명변수 중요도가 생각보다 낮게 나왔다.

2) 이와 함께 처음했던 예상과 다르게
흑인과 저소득층의 상관관계가 낮다는 것을 알 수 있었다.

3) 주거지 비율 변수의 경우, 중요도가 0.000으로 나왔기 때문에
주거지 비율 변수는 보스턴 주택 가격에
영향을 미치지 않는다는 통찰을 얻었다.

4) 처음에 세운 가설을 설명할
'보스턴 외곽지역 거주 여부'는
DIS(중심지와의 접근거리) 변수를 통해
나타내야 한다는 것을 깨달았다.

5) 최종적으로 가설을 설정한 배경에 대한 검토가
다시 이루어져야 한다고 판단했다.

*****통찰을 통한 배경 검토*****

처음 생각) 저소득층 비율 중
흑인이 상대적으로 많을 것이다!

분석 후 판단) 저소득층 비율 중
흑인이 많다고 할 수 없다.

4. 분석 결론

(1) 가설과 변수 수정 필요성 확인

가설 보스턴 외곽지역에 거주하는 저소득층 흑인 비율이 높을수록 보스턴 주택 가격이 낮을 것이다.

➡ 가설 검토

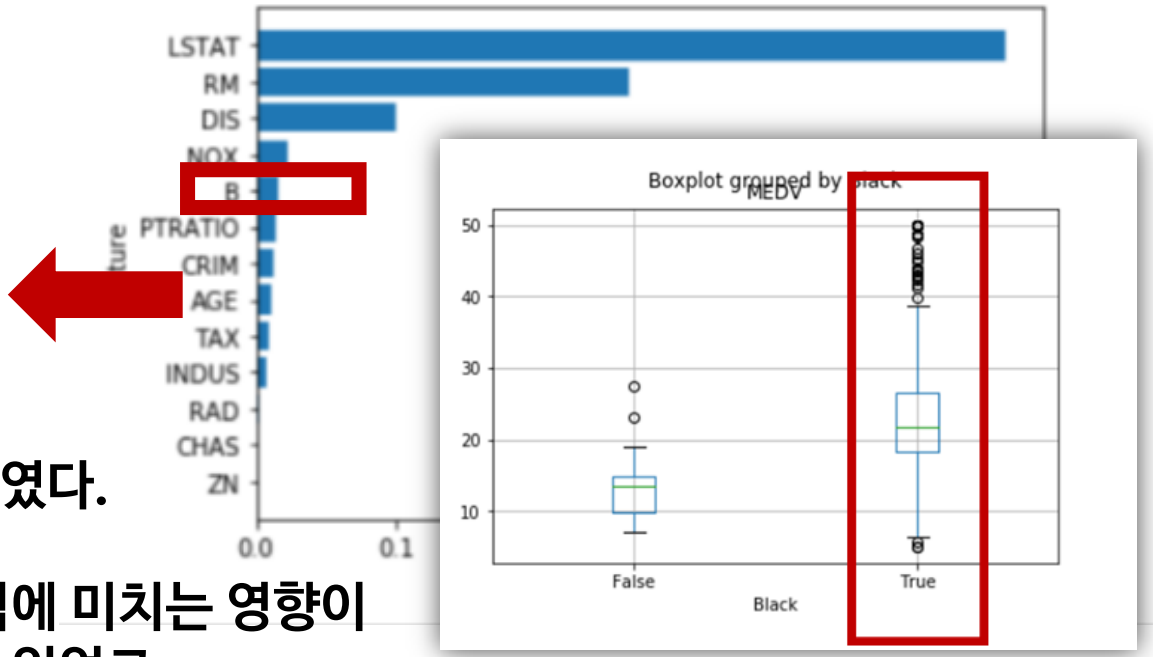
저소득층과 흑인의 상관관계가 낮으므로,
가설을 증명하기 위해서는 해당 데이터에 집계된 **저소득층 중 흑인이 얼마나 되는지 파악이 이루어져야 한다.**

➡ 변수 검토

'보스턴 외곽지역' = **DIS(중심지와의 접근거리)**
'저소득층 비율' = **LSTAT(저소득층 비율)**
'흑인 비율' = **B(흑인 인구 비율)**

가설을 증명하기 위해서는 설명변수 간의 연관성과 당시 보스턴의 주거 형태에 대한 조사가 더 이루어져야 한다.

(2) 설명변수 중요도를 통한 분석 배경 재검토



처음에 세운 가설을 증명할 때
필요했던 설명변수 중,
B(흑인 인구 비율)는
설명변수 중요도가 낮았으며,
처음 가설과 반대되는 경향을 보였다.

흑인 인구 비율 변수가 주택 가격에 미치는 영향이
생각보다 낮다는 결론을 내릴 수 있었고,
분석의 배경이 된 공영주택 프로젝트와
1970년대 중반 보스턴의 인종차별 현황에 대한 조사가
더 이루어져야 한다고 판단했다.
또, 데이터의 수집주체와 신뢰도에 대한 판단도 필요하다고 생각했다.

5. 분석을 통해 깨달은 점

1) 뚜렷한 주관을 가져야 하는 데이터 분석가

분석 업무를 하며 변수 선택과 가설 수용의 기준이 애매함을 느꼈다.
이를 통해 분석가가 스스로 중요하게 여기는 점이 무엇인지,
비즈니스 관점에서 기업이 중요하게 생각하는 것이
무엇인지 분석가 스스로 알고 있는 것이 중요하다고 생각했다.

2) 도메인 지식의 부족함을 통해 깨달은 내용전문가의 중요성

세운 가설이 부분적으로 증명되었을 때, 분석 배경을 검토하는 것은
해당 부분에 대한 도메인 지식이 있어야 가능하다고 생각했다.
따라서 내용 전문가의 역할이 데이터 분석의 8할이라는 것을 깨달았다.

3) 수학적 지식의 중요성과 앞으로의 공부 다짐

개인적으로 과제를 진행하며 통계와 수학 관련 지식이 아직도 부족함을 느꼈다.
앞으로 공부를 통해 부족한 부분을 채워나가야겠다는 다짐을 했다.

감 사 합 니 다