종합실습 1

# BOSTON
# 집값 예측

집 값 에  영 향 을  주 는  요 인  분 석

A반 3조 김준오

# INDEX

0 과제 정의

# 분석 배경

미국 인구 조사국에서 수집한 보스턴 시의 주택 가격에 대한 데이터를 통해,
주택의 가격에 영향을 미치는 인자를 분석하고자 한다.

다양한 예측모델을 이용하여 집값에 영향을 주는 영향인자를 객관적으로 도출하고,
선정한 영향인자를 활용하여 예측한다.

# 1 가설 설정

1. 범죄율(CRIM)이 높을수록 주택가격은 낮아질 것이다. (-)
2. 주거당 평균 객실 수가 클수록 주택가격은 올라갈 것이다. (+)
3. 중심지(노동센터) 접근 거리가 작을수록 주택가격은 상승할 것이다. (-)
4. 학생당 교사 비율이 높을수록 주택 가격은 올라갈 것이다. (+)
5. 저소득층 비율이 높을수록 주택 가격은 내려갈 것이다. (-)

# 2 데이터 처리 – 결측치, 데이터 타입 확인

| | MEDV | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 24.000000 | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.199997 | 4.0900 | 1 | 296 | 15.300000 | 396.899994 | 4.98 |
| 1 | 21.600000 | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.900002 | 4.9671 | 2 | 242 | 17.799999 | 396.899994 | 9.14 |
| 2 | 34.700001 | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.099998 | 4.9671 | 2 | 242 | 17.799999 | 392.829987 | 4.03 |
| 3 | 33.400002 | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.799999 | 6.0622 | 3 | 222 | 18.700001 | 394.630005 | 2.94 |
| 4 | 36.200001 | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.200001 | 6.0622 | 3 | 222 | 18.700001 | 396.899994 | 5.33 |

데이터 확인 : df.head()

```
1 df_raw.isnull().sum(axis=0)
```

```
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
```

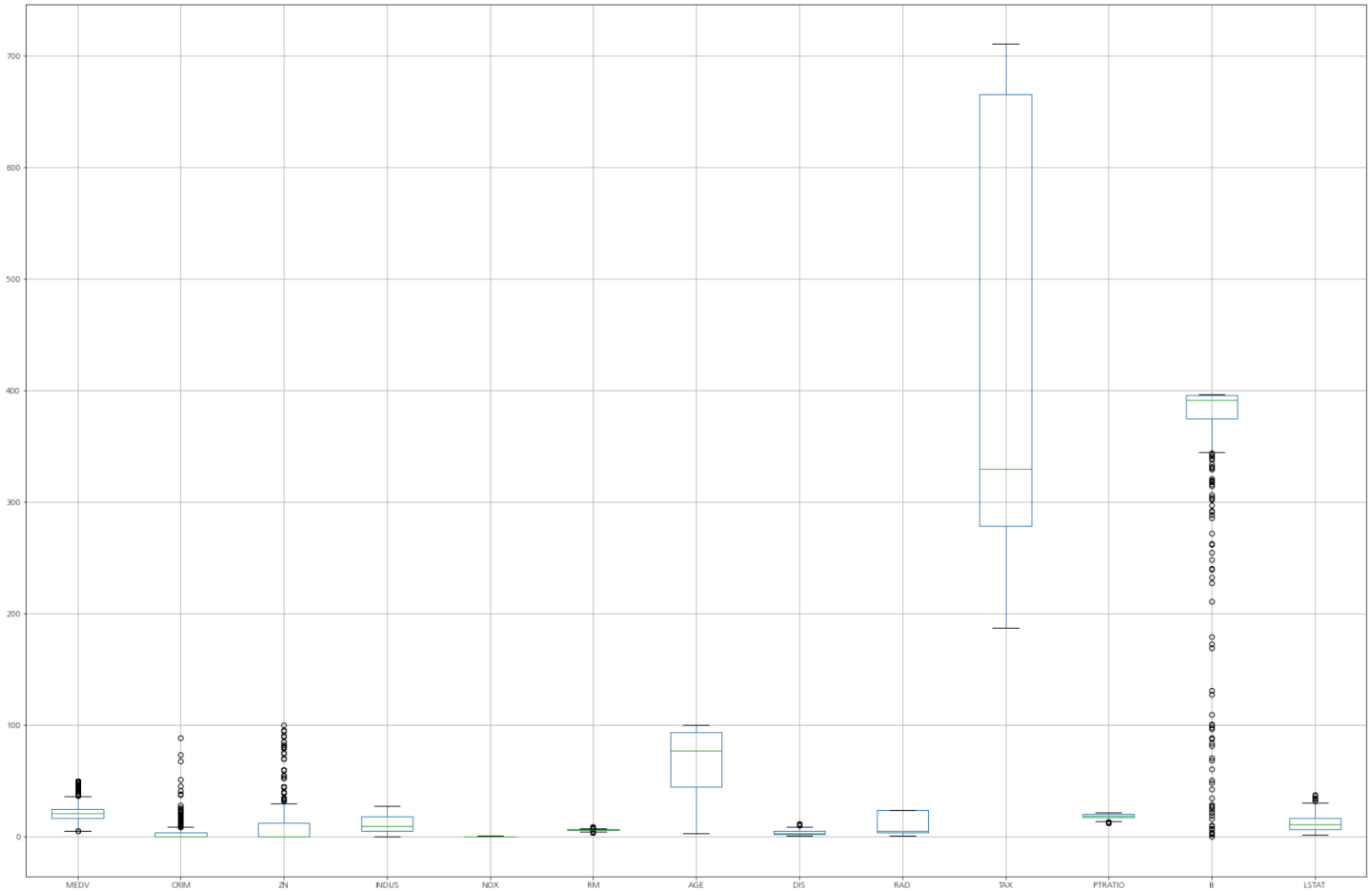
결측치 확인

```
1 df_raw.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   MEDV     506 non-null    float64
 1   CRIM     506 non-null    float64
 2   ZN       506 non-null    float64
 3   INDUS    506 non-null    float64
 4   CHAS     506 non-null    int64
 5   NOX      506 non-null    float64
 6   RM       506 non-null    float64
 7   AGE      506 non-null    float64
 8   DIS      506 non-null    float64
 9   RAD      506 non-null    int64
 10  TAX      506 non-null    int64
 11  PTRATIO  506 non-null    float64
 12  B        506 non-null    float64
 13  LSTAT    506 non-null    float64
dtypes: float64(11), int64(3)
memory usage: 55.5 KB
```

데이터 타입 확인

# 2 데이터 처리 – 이상치 확인




Boxplot 을 활용하여
데이터의 이상치를
확인한 결과,
특별한 이상치를
확인할 수 없었다.

이상치 확인

# 2 데이터 처리

| | MEDV | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 24.000000 | 0.00632 | 18.0 | 2.31 | 0 | 0.538 | 6.575 | 65.199997 | 4.0900 | 1 | 296 | 15.300000 | 396.899994 | 4.98 |
| 1 | 21.600000 | 0.02731 | 0.0 | 7.07 | 0 | 0.469 | 6.421 | 78.900002 | 4.9671 | 2 | 242 | 17.799999 | 396.899994 | 9.14 |
| 2 | 34.700001 | 0.02729 | 0.0 | 7.07 | 0 | 0.469 | 7.185 | 61.099998 | 4.9671 | 2 | 242 | 17.799999 | 392.829987 | 4.03 |
| 3 | 33.400002 | 0.03237 | 0.0 | 2.18 | 0 | 0.458 | 6.998 | 45.799999 | 6.0622 | 3 | 222 | 18.700001 | 394.630005 | 2.94 |
| 4 | 36.200001 | 0.06905 | 0.0 | 2.18 | 0 | 0.458 | 7.147 | 54.200001 | 6.0622 | 3 | 222 | 18.700001 | 396.899994 | 5.33 |

데이터 확인 : df.head()

```
1 df_raw.isnull().sum(axis=0)
```

```
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

```
1 df_raw.info()
```

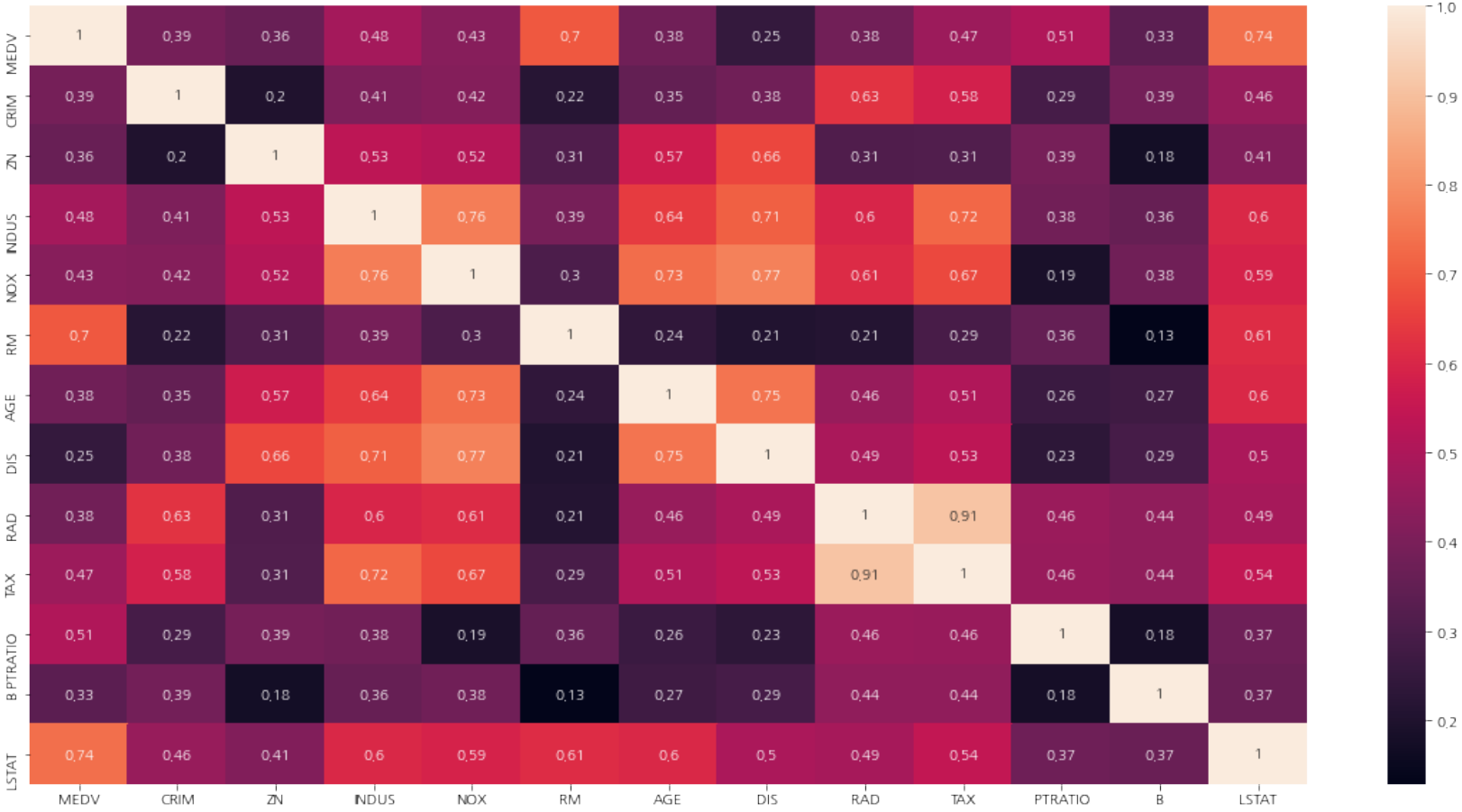
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 506 entries, 0 to 505
Data columns (total 14 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   MEDV     506 non-null    float64
 1   CRIM     506 non-null    float64
 2   ZN       506 non-null    float64
 3   INDUS    506 non-null    float64
 4   CHAS     506 non-null    int64
 5   NOX      506 non-null    float64
 6   RM       506 non-null    float64
 7   AGE      506 non-null    float64
 8   DIS      506 non-null    float64
 9   RAD      506 non-null    int64
 10  TAX      506 non-null    int64
 11  PTRATIO  506 non-null    float64
 12  B        506 non-null    float64
 13  LSTAT    506 non-null    float64
dtypes: float64(11), int64(3)
memory usage: 55.5 KB
```

데이터 타입 확인

데이터 확인 결과,

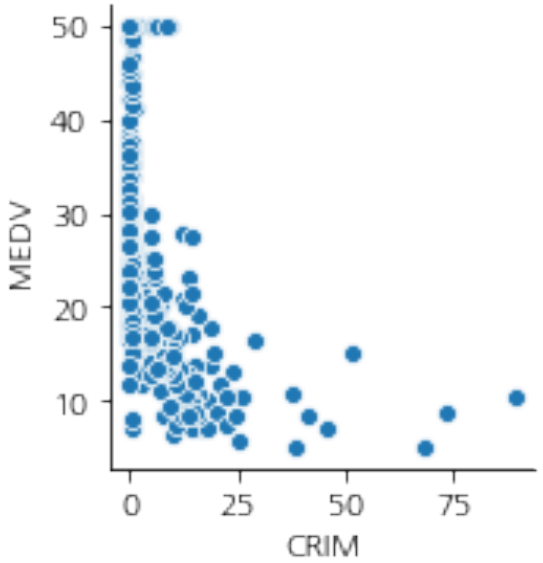
1. 결측치, 이상치 모두 존재하지 않았다.
2. “CHAS” 데이터 타입을 object 로 변경해야한다.

# 3 탐색적 분석

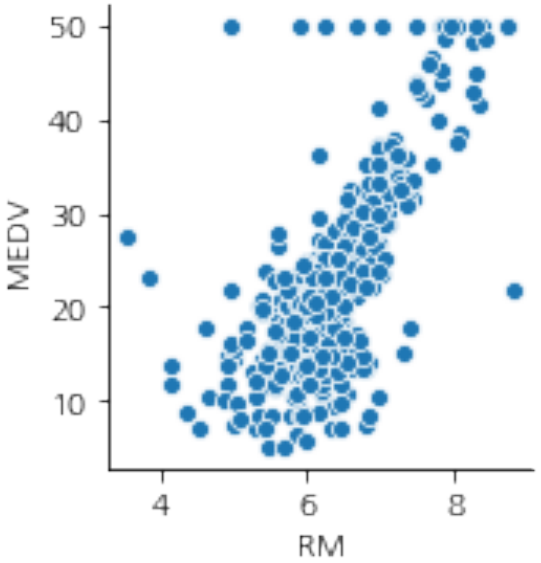



히트맵을 통해 전체적인 상관관계 분석

# 3 탐색적 분석



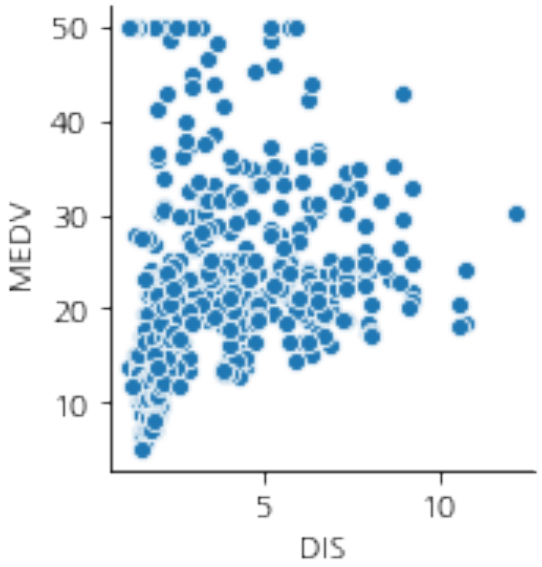

**가설1**
범죄율과 주택가격



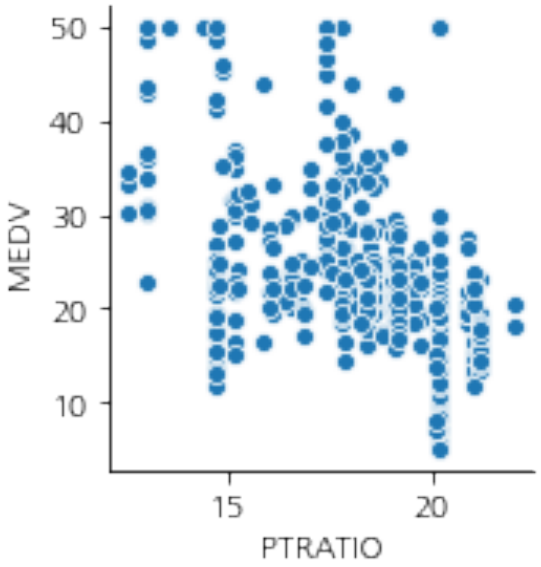

**가설2**
주거당 평균 객실 수와 주택가격

가설1 : 범죄율이 올라갈수록 주택가격이 내려가는 경향을 확인할 수 있다.
가설2 : 평균 객실 수가 커지면, 주택가격이 올라가는 경향을 확인할 수 있다.

# 3 탐색적 분석



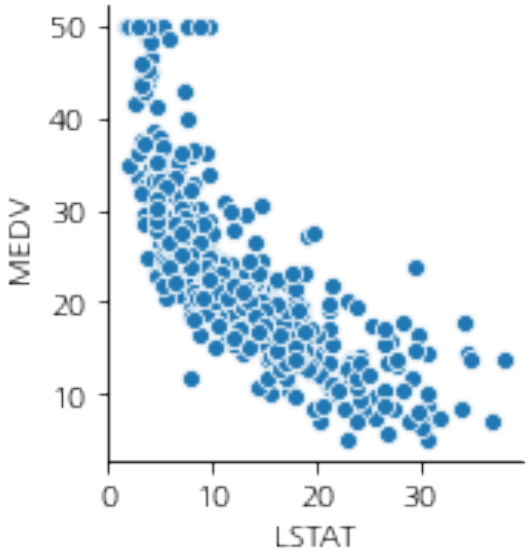

**가설3**
중심지 접근 거리와 주택가격




**가설4**
학생당 교사 비율과 주택가격

가설3 : 중심지 접근 거리와 주택가격 사이의 확실한 상관관계를 파악하기 어려웠다.
가설4 : 학생당 교사 비율과 주택가격 사이의 확실한 상관관계를 파악하기 어려웠다.

# 3 탐색적 분석




**가설5**
저소득층 비율과 주택가격

가설5 : 저소득층 비율이 커질수록 주택가격이 내려가는 경향성을 확인할 수 있었다.

# 4 모델링

# 4 모델링 – 01. 다중 회귀분석

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | MEDV | R-squared: | 0.741 |
| Model: | OLS | Adj. R-squared: | 0.734 |
| Method: | Least Squares | F-statistic: | 108.1 |
| Date: | Wed, 25 Nov 2020 | Prob (F-statistic): | 6.72e-135 |
| Time: | 02:38:47 | Log-Likelihood: | -1498.8 |
| No. Observations: | 506 | AIC: | 3026. |
| Df Residuals: | 492 | BIC: | 3085. |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 36.4595 | 5.103 | 7.144 | 0.000 | 26.432 | 46.487 |
| C(CHAS)[T.1] | 2.6867 | 0.862 | 3.118 | 0.002 | 0.994 | 4.380 |
| CRIM | -0.1080 | 0.033 | -3.287 | 0.001 | -0.173 | -0.043 |
| ZN | 0.0464 | 0.014 | 3.382 | 0.001 | 0.019 | 0.073 |
| INDUS | 0.0206 | 0.061 | 0.334 | 0.738 | -0.100 | 0.141 |
| NOX | -17.7666 | 3.820 | -4.651 | 0.000 | -25.272 | -10.262 |
| RM | 3.8099 | 0.418 | 9.116 | 0.000 | 2.989 | 4.631 |
| AGE | 0.0007 | 0.013 | 0.052 | 0.958 | -0.025 | 0.027 |
| DIS | -1.4756 | 0.199 | -7.398 | 0.000 | -1.867 | -1.084 |
| RAD | 0.3060 | 0.066 | 4.613 | 0.000 | 0.176 | 0.436 |
| TAX | -0.0123 | 0.004 | -3.280 | 0.001 | -0.020 | -0.005 |
| PTRATIO | -0.9527 | 0.131 | -7.283 | 0.000 | -1.210 | -0.696 |
| B | 0.0093 | 0.003 | 3.467 | 0.001 | 0.004 | 0.015 |
| LSTAT | -0.5248 | 0.051 | -10.347 | 0.000 | -0.624 | -0.425 |

| | | | |
|---|---|---|---|
| Omnibus: | 178.041 | Durbin-Watson: | 1.078 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 783.126 |
| Skew: | 1.521 | Prob(JB): | 8.84e-171 |
| Kurtosis: | 8.281 | Cond. No. | 1.51e+04 |

## 다중 회귀분석 결과,

1. 모델의 설명력은 0.741(74.1%)이다.
2. F-검정 결과, P-value는 0.05보다 매우 작으므로 통계값은 유효하다.
3. INDUS, AGE 값은 P-value가 0.05보다 매우 크므로 의미가 없는 값이다.

# 4 모델링 – 01. 다중 회귀분석 : 가설 검증

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | MEDV | R-squared: | 0.741 |
| Model: | OLS | Adj. R-squared: | 0.734 |
| Method: | Least Squares | F-statistic: | 108.1 |
| Date: | Wed, 25 Nov 2020 | Prob (F-statistic): | 6.72e-135 |
| Time: | 02:38:47 | Log-Likelihood: | -1498.8 |
| No. Observations: | 506 | AIC: | 3026. |
| Df Residuals: | 492 | BIC: | 3085. |
| Df Model: | 13 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 36.4595 | 5.103 | 7.144 | 0.000 | 26.432 | 46.487 |
| C(CHAS)[T.1] | 2.6867 | 0.862 | 3.118 | 0.002 | 0.994 | 4.380 |
| CRIM | -0.1080 | 0.033 | -3.287 | 0.001 | -0.173 | -0.043 |
| ZN | 0.0464 | 0.014 | 3.382 | 0.001 | 0.019 | 0.073 |
| INDUS | 0.0206 | 0.061 | 0.334 | 0.738 | -0.100 | 0.141 |
| NOX | -17.7666 | 3.820 | -4.651 | 0.000 | -25.272 | -10.262 |
| RM | 3.8099 | 0.418 | 9.116 | 0.000 | 2.989 | 4.631 |
| AGE | 0.0007 | 0.013 | 0.052 | 0.958 | -0.025 | 0.027 |
| DIS | -1.4756 | 0.199 | -7.398 | 0.000 | -1.867 | -1.084 |
| RAD | 0.3060 | 0.066 | 4.613 | 0.000 | 0.176 | 0.436 |
| TAX | -0.0123 | 0.004 | -3.280 | 0.001 | -0.020 | -0.005 |
| PTRATIO | -0.9527 | 0.131 | -7.283 | 0.000 | -1.210 | -0.696 |
| B | 0.0093 | 0.003 | 3.467 | 0.001 | 0.004 | 0.015 |
| LSTAT | -0.5248 | 0.051 | -10.347 | 0.000 | -0.624 | -0.425 |

| | | | |
|---|---|---|---|
| Omnibus: | 178.041 | Durbin-Watson: | 1.078 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 783.126 |
| Skew: | 1.521 | Prob(JB): | 8.84e-171 |
| Kurtosis: | 8.281 | Cond. No. | 1.51e+04 |

1 범죄율(CRIM)이 높을수록 주택가격은 낮아질 것이다. **(-)**

→ CRIM의 coef는 음수로 **음의 상관관계**를 보인다. (가설 O)

2 주거당 평균 객실 수가 클수록 주택가격은 올라갈 것이다. **(+)**

→ RM의 coef는 큰 양수로 **양의 상관관계**를 보인다. (가설 O)

3 중심지(노동센터) 접근 거리가 작을수록 주택가격은 상승할 것이다. **(-)**

→ DIS의 coef는 큰 음수로 음의 상관관계를 보이는 듯 하나, P-value**가 0.05보다 크므로** 성립하지 않는다. (X)

4 학생당 교사 비율이 높을수록 주택 가격은 올라갈 것이다. **(+)**

→ PTRATIO의 coef는 음수로 **음의 상관관계**를 보인다. (가설X)

5 저소득층 비율이 높을수록 주택 가격은 내려갈 것이다. **(-)**

→ LSTAT의 coef는 음수로 **음의 상관관계**를 보인다. (가설O)

# 4 모델링 – 01. 다중 회귀분석 : VIF

| | 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 |

VIF 결과

→ 다중공선성도 확인한 결과, 모든 설명변수들의 VIF 값이 10 이하로, 다중공선성을 보이는 설명변수는 없는 것으로 판단할 수 있다.

# 4 모델링 – 01. 다중 회귀분석 : 후진제거법

후진제거법 적용 후 선택된 변수들 : 'NOX', 'RM', 'DIS', 'PTRATIO', 'LSTAT' (5개)
후진제거법 적용 후 제거된 변수들 : 'CRIM', 'ZN', 'INDUS', 'AGE', 'RAD', 'TAX', 'B' (7개)

후진제거법을 적용해 선택된 5개의 변수 NOX(산화질소 농도), RM(주거당 평균 객실 수), DIS(중심지 접근 거리), PTRATIO(학생당 교사 비율), LSTAT(저소득층 비율)이 다중 회귀분석에서 주요 변수들이다.

# 4 모델링 – 01. 다중 회귀분석

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | MEDV | R-squared: | 0.708 |
| Model: | OLS | Adj. R-squared: | 0.705 |
| Method: | Least Squares | F-statistic: | 242.6 |
| Date: | Wed, 25 Nov 2020 | Prob (F-statistic): | 3.67e-131 |
| Time: | 02:41:20 | Log-Likelihood: | -1528.7 |
| No. Observations: | 506 | AIC: | 3069. |
| Df Residuals: | 500 | BIC: | 3095. |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 37.4992 | 4.613 | 8.129 | 0.000 | 28.436 | 46.562 |
| NOX | -17.9966 | 3.261 | -5.519 | 0.000 | -24.403 | -11.590 |
| RM | 4.1633 | 0.412 | 10.104 | 0.000 | 3.354 | 4.973 |
| DIS | -1.1847 | 0.168 | -7.034 | 0.000 | -1.516 | -0.854 |
| PTRATIO | -1.0458 | 0.114 | -9.212 | 0.000 | -1.269 | -0.823 |
| LSTAT | -0.5811 | 0.048 | -12.122 | 0.000 | -0.675 | -0.487 |

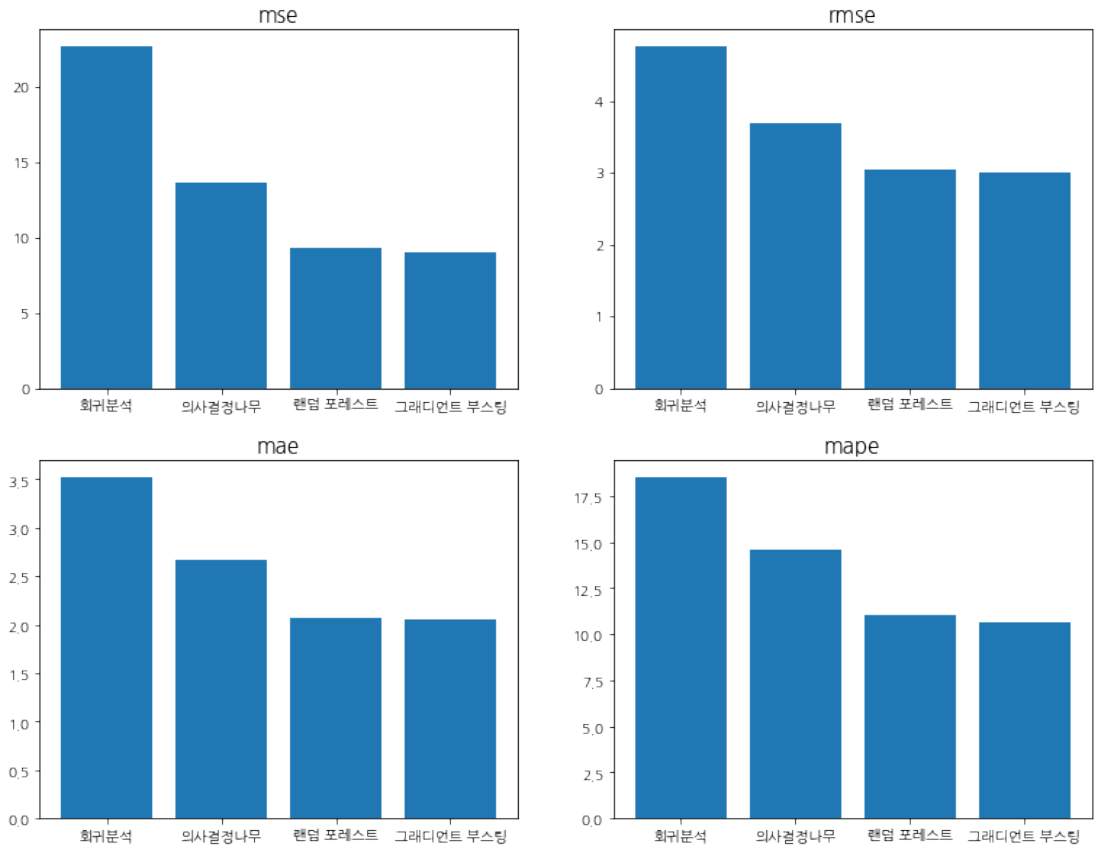
| | | | |
|---|---|---|---|
| Omnibus: | 187.456 | Durbin-Watson: | 0.971 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 885.498 |
| Skew: | 1.584 | Prob(JB): | 5.21e-193 |
| Kurtosis: | 8.654 | Cond. No. | 545. |

## 다중 회귀분석 결과,

1. 모델의 설명력은 0.708(70.8%)이다.
2. F-검정 결과, P-value는 0.05보다 매우 작으므로 통계값은 유효하다.
3. 모든 설명변수들의 P-value가 0.05보다 작으므로 의미가 있다.
4. 식으로 표현하자면 다음과 같다.

MEDV = 37.4992 + (-17.9966)*NOX + (4.1633)*RM + (-1.1847)*DIS + (-1.0458)*PTRATIO + (-0.5811)*LSTAT

# 5 분석결과 – 다양한 모델 테스트 결과 비교




에러값들을 모델별로 산출해보았다. MSE, RMSE, MAE, MAPE를 계산했으며,
그 결과, 에러는 그래디언트 부스팅 < 랜덤 포레스트 < 의사결정나무 < 회귀분석 순으로 산출되었다.

6 소감

# 데이터 분석을 하며,

집값 분석이라는, 와 닿는 주제를 활용하여 공부해서,
더 흥미를 가지고 분석 실습에 임할 수 있었습니다.

통계에 대한 기본을 더 확실히 하고, 이를 전공 영역에 접목시킬 수 있는
역량을 가진다면, 정말 큰 도움이 될 것이라 생각했습니다.