Assignment 1

보건환경융합과학부

2019250311

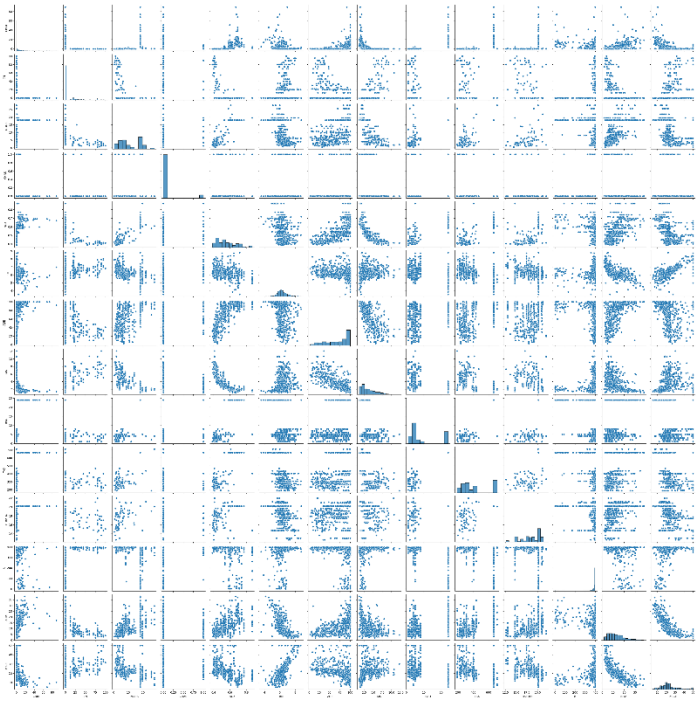
김모세

Task 1

```
for i, col in enumerate(boston_df.columns):
  plt.figure(figsize=(8,4))
  plt.plot(boston_df[col])
  plt.title(col)
  plt.xlabel('Town')
  plt.tight_layout()
for i, col in enumerate(boston_df.columns):
  plt.figure(figsize=(8,4))
  plt.scatter(boston_df[col], boston_df['PRICE'])
  plt.ylabel('PRICE', size=12)
  plt.xlabel(col, size=12)
  plt.tight_layout()
```

```
import seaborn as sns

sns.pairplot(boston_df);
```



Task 2

1. 변수 분석
   A. CRIM : 높은 집은 가격이 낮다.
   B. ZN : PRICE와 비례관계이다.
   C. INDUS : 크게 관계가 없다.
   D. CHAS : 높으면 PRICE 최저치가 높다.
   E. NOX : 높으면 PRICE가 낮다.
   F. RM : PRICE와 비례 관계를 가진다.
   G. AGE : 높으면 PRICE가 낮은 경향이 있다.
   H. DIS : 낮으면 PRICE가 낮은 경향이 있다.
   I. RAD : 높으면 PRICE가 낮다.
   J. TAX: 높으면 PRICE가 낮다
   K. PTRATIO : 높으면 PRICE가 낮다.
   L. B : 상관이 크게 없다.
   M. LSTAT : PRICE와 반비례한다.

변수들의 분류를 위생요인과 동기요인으로 나눠보면, 선형적인 관계를 가지는 요인을 동기요인으로, 그렇지 않고 특정 조건일 때 결과에 영향을 미치는 요인을 위생요인으로 나눌 수 있을 것 같다.

선형적인 관계를 가지는 ZN, RM, LSTAT은 동기요인이고, 특정 범위에서 PRICE에 큰 영향을 미치는 CRIM, CHAS, NOX, AGE, DIS, RAD, TAX, PTRATIO는 위생요인이다. 동기요인 중 RM과 LSTAT은 반비례 관계를 가진다. 따라서 multicollinearity를 고려하여 LSTAT을 RM에 포함시키는 것이 적절해보인다.

Task 3

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

y = boston_df['PRICE']
X = boston_df.drop(labels=['PRICE'], axis=1)

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3 )
print(X_train.shape, y_train.shape)

model = LinearRegression()
model.fit(X_train, y_train)
print("학습 데이터 점수: {}".format(model.score(X_train, y_train)))
print("평가 데이터 점수: {}".format(model.score(X_test, y_test)))
```

```
from sklearn.metrics import mean_squared_error, r2_score

predict = model.predict(X_train)
rmse = (np.sqrt(mean_squared_error(y_train, predict)))
r2 = r2_score(y_train, predict)

print('RMSE: {}'.format(rmse))
print('R2 Score: {}'.format(r2))
```

| 모델 | 평가 데이터 점수 | RMSE | R2 Score for test |
|---|---|---|---|
| 'INDUS', 'B', 'LSTAT' columns를 제거한 경우 | 0.5-0.7 | 4.67-5.52 | 0.71 - 0.65 |
| 'INDUS', 'B' columns를 제거한 경우 | 0.70 – 0.73 | 4.7 | 0.74 |
| 아무 column도 제거하지 않은 경우 | 0.67-0.78 | 4.55 – 4.82 | 0.73 - 0.76 |

세 가지 경우로 나눠서 여러 번 점수를 측정해보았다.

Task2에서 예상한대로 'INDUS', 'B', 'LSTAT' columns를 제거한 경우를 측정한 결과 평가 데이터 점수가 데이터 셋에 따라서 0.5-0.7 정도로 변동이 심하였다. 하지만 'LSTAT'을 포

함하여 측정하자 0.70이상으로 결과가 나와 multicollinearity를 잘못 평가했다고 알 수 있었다. 아무 column도 제거하지 않은 경우 데이터셋에 따라서 0.67-0.76정도의 정확성을 가진 것으로 나타났다.

따라서 데이터셋에 따른 모델의 안정성은 'INDUS', 'B' columns만 제거한 경우 더 높았고, 모델의 정확도는 아무 컬럼도 제거하지 않은 경우가 제일 높다고 결론내렸다.

변수들의 계수는

```
print('labels\n',X.columns)
print('Coefficients: \n', model.coef_)
print('Intercept: \n', model.intercept_)
print('R2 for Train)', model.score( X_train, y_train ))
print('R2 for Test (cross validation)', model.score(X_test, y_test))
```

```
labels
 Index(['CRIM', 'ZN', 'INDUS', 'CHAS', 'NOX', 'RM', 'AGE', 'DIS',
'RAD', 'TAX',
       'PTRATIO', 'B', 'LSTAT'],
      dtype='object')
Coefficients:
 [-1.19378921e-01  4.11125026e-02  4.08459356e-02  3.01302362e+00
 -1.98650082e+01  3.38602559e+00  3.92024316e-03 -1.55679024e+00
  3.66765928e-01 -1.25857614e-02 -9.94433671e-01  1.15650279e-02
 -5.60017122e-01]
```

로 나타났다.

따라서 Task2에서의 'INDUS', 'B'가 PRICE와 큰 관계가 없다는 예측은 맞았다고 할 수 있다. 하지만 관계가 있을거라 생각했던 'ZN', 'RM', 'LSTAT'에서 'RM'을 제외하고는 PRICE에 큰 영향을 미치는 변수가 없었다. 다른 변수인 'NOX', 'DIS', 'CHAS'가 더 큰 영향을 미치고 있었다.

Task 4

계수들을 분석한 결과, PRICE를 결정하는데 있어서 동기요인보다 위생요인이 더 큰 영향을 미치는 것으로 알 수 있다. 즉, 사람들은 집값을 평가하는데 있어서 여러 특정 조건들이 만족되는 것이 더 중요하다. 예를 들어, 주요 변수들인 'NOX', 'DIS', 'CHAS', 'RM'에서 위생요인인 'NOX', 'DIS', 'CHAS'를 모두 다 만족하는 주변에 공원 같은 것이 있어 공기가 좋고, Boston employment centres와 가깝고, Charles River가 근처에 있는 조건인 집이 집값이 높은 것이다. 집값과 선형적인 관계이 있는 'ZN', 'LSTAT'은 위생요인에 비하여 영향이 적고, 방의 개수인 'RM'정도만이 동기요인과 비슷한 영향을 미친다.

해당 모델은 데이터의 개수가 적어 데이터셋에 따라서 정확도의 변동이 심한 것이 한계라 생각한다. 이를 보완하기 위해 더 많은 데이터를 통하여 모델을 학습시킬 수 있다면 좋은 결과를 도출할 수 있다고 생각한다.