6. 데이터 분석 실습

6.1 보스턴 주택 가격 데이터

이 데이터를 토대로 앞에서 배운 다중 선형 회귀분석을 이용해 주택가격을 예측하는 방법과 주성분분석을 이용해 변수를 선택하는 방법을 배워보자. 원래는 총 506개의 데이터가 404개의 훈련 데이터와 102개의 테스트 데이터로 분리돼 있어, 여기서는 전체를 다 한꺼번에 실습하고자 한다.

6.2 다중 선형 회귀 분석

먼저 보스턴 주택 가격 데이터를 이용해 주택가격을 회귀분석해보자.

독립변수의 개수는 전부 13개이며, 데이터의 개수는 506개이다. 이 데이터를 주택 가격을 예측하거나 주택가격에 영향을 미치는 요인을 찾는데 이용한다. 11개의 변수를 모두 사용해서 회귀분석을 바로 실행할 수도 있지만 머저 데이터가 어찌 구성돼 있는지 살펴보자.

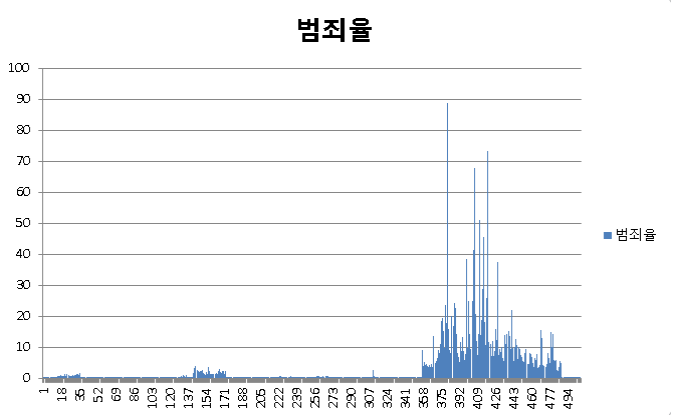
데이터를 통계수치, 그래프 등을 이용해 확인하고 탐색하는 과정을 EDA(Exploratory Data Analysis), 즉 택색적 데이터 분석이라고 한다. 이 과정은 1장에서 설명한 데이터 전처리 과정과 기초 통계 분석 과정에 포함돼 있는 과정으로서 데이터 분석을 하기 전에는 항상 데이터가 어떤 형태로 구성돼 있는지, 어떤 분포와 어떤 특징을 띠는지 파악하는 과정을 반드시 거쳐야 한다.

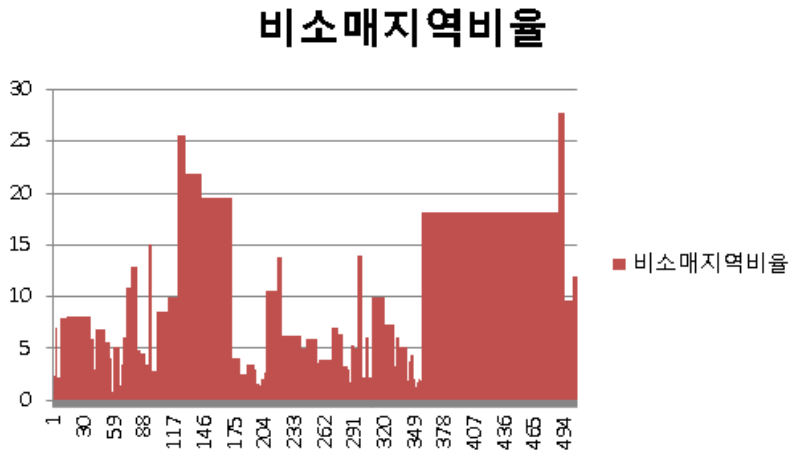
EDA는 다음과 같은 과정을 포함한다.

1. **각 변수의 정의와 데이터형 확인**
2. **데이터에 누락이 있는지 확인**
3. **변수의 데이터별로 이상값이 있는지 확인**
4. **히스토그램, 시계열차트, 박스플롯 등 그래프를 이용해 시각적으로 확인**
5. **기초 통곗값을 이용해 데이터의 특성을 확인**
6. **기타 등등**

바쁘다고 이 과정을 건너뛰었을 때는 최악의 경우 처음부터 다시 분석해야 하는 경우가 발생할 수 있으며, EDA를 통해 데이터에 대한 새로운 발견과 적용할 알고리즘 선택도 가능하므로 반드시 EDA를 진행하기 바란다.

먼저 데이터의 형태가 어떤지 다음 그림을 통해 살펴보자.





등등 이런 식으로 독립변수 11개에 대해 다 분석해줘야 한다.

그래프를 보면 고속도로 접근성과 재산세율이 비슷한 모양인 것을 볼 수 있다. 실제로 상관계수를 계산해 보면 0.91로 상당히 높다는 것을 확인할 수 있다. 이를 다중공선성으로 처리할 수도 있지만 여기서는 그냥 두기로 한다.

또한 범죄율과 거주지역비율, 찰스강더미의 값에 0이 많은 것을 확인할 수 있다. 이를 좀 더 자세히 살펴보면 범죄율은 실제로 0인 값은 별로 없고 0에 가까운 값이 많다는 사실을 확인 할 수 있다.

거주지역비율의 0 개수는 372개로, 약 74%에 해당하며, 주택가격과의 상관계수는 0.36으로 그다지 높지 않다. 찰스강더미의 0 개수는 471개로, 약 93%이며, 역시 주택가격과의 상관계수는 약 0.17임을 확인할 수 있다. 이로써 거주지역비율과 찰스강더미 변수는 변별력 없는 변수이므로, 11개를 고려하겠다.

물론 지금까지 설명한 과정이 정답은 아니다. 히스토그램, 상관행렬 등을 통해 여러 가지 방법으로 EDA를 진행해 보면 또 다른 아이디어가 생길 수 있으니 직접 확인해보자.

구해야할 변환 행렬 A는 다음과 같이 정의했다.

$$A = (X^T X)^{-1} X^T Y$$

Step 0. 상수항 추가

Step 1. X^T 계산

이건 범위가 넘 커서 엑셀 고유의 기능을 활용해보려 한다.

Step 2. $X^T X$ 계산

$$X^T \times X = XX^T$$
$$(12 \times 506) \times (506 \times 12) = (12 \times 12)$$

12×12 범위를 선택한 후 계산해라.

Step 3. $X^T X^{-1}$ 계산

Step 4. $X^T Y$ 계산

$$X^T \times Y = X^T Y$$
$$(12 \times 506) \times (506 \times 1) = (12 \times 1)$$

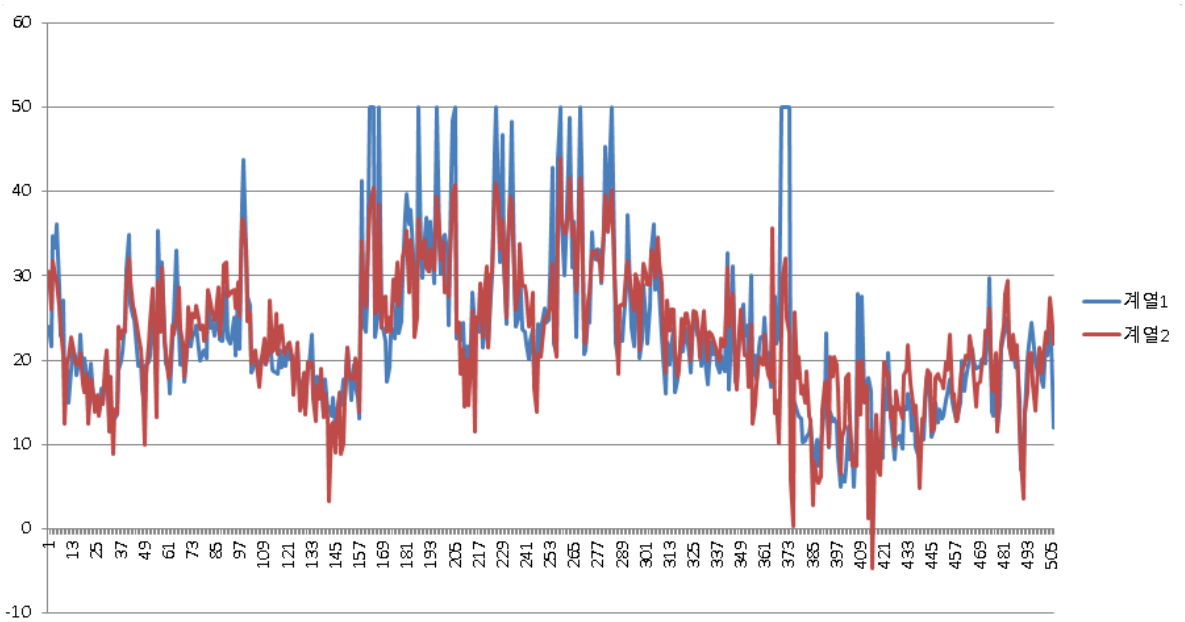
Step 5. $A = (X^T X)^{-1} X^T Y$ 계산

이 데이터는 표준화하지 않았기에 부호만 살펴보기로 한다. 회귀계수의 부호가 양수인 경우에는 해당 변수가 커질수록 주택가격이 상승한다는 것을 의미하며, 회귀계수의 부호가 음수인 경우에는 해당 변수가 커질수록 주택가격이 하락한다는 것을 의미한다.

살펴보면 일반적인 상식에 준하는 회귀계수도 있지만, 의외의 회귀계수도 있다는 사실을 알 수 있다. 여기에는 두 가지 가능성이 있다. 하나는 수집된 데이터에 오류가 섞여있을 가능성이고, 다른 하나는 지금까지 그러려니 생각했던 관계가 잘못된 경우다. 데이터 분석을 하다보면 후자의 경우도 의외로 많은 것을 경험할 수 있다. 일단 결과를 의심하고 보는 습관을 들이자.

이제 계산된 회귀계수를 통해 주택가격을 얼마나 잘 예측하는지 계산해보자.

실제값과 예측값 둘 간의 상관계수는 0.854이다.



위 그림은 실제 주택가격과 예측한 주택가격을 그래프로 그린것이다. 전체적인 트렌드는 잘 따라가고 있는 것으로 보이며 오차가 큰 부분이 군데 군데 보이는 것을 볼 수 있다.

여기서 더 할 수 있는 것은 예측 성능을 높이는 일과, 만약 예측 성능이 만족스럽다면 성능저하를 최소화하면서 좀 더 효율적으로 예측할 수 있게 하는 것이다.

좀 더 구체적으로 설명하면 먼저 예측 성능을 높이는 다양한 머신러닝 알고리즘을 사용해 문제에 많는 머신러닝 알고리즘을 찾고 개량화하는 것이다. 예를 들어 의사결정 나무, 서포트 벡터 머신, 딥러닝을 이용하거나 요즘 인기 있는 XGBOOST를 이용하면 좀 더 예측 성능이 좋은 모델을 구축할 수 있다.

다음으로 현재의 예측성능에 만족하는 경우로, 성능 저하를 최소화하는 변수나 특징을 찾는 것이다. 이를 특징 선택(feature selection)이나 특징 추출(feature extraction) 이라고 하며, 이를 위한 여러가지 방법론이 있지만 여기서는 주성분 분석을 이용한 방법을 진행해 보겠다.

6.3 주성분 분석을 이용한 특징 선택

3장에서는 주성분분석을 통해 데이터를 압축하는 방법과 새로운 정보를 만드는 방법을 배웠다. 여기서는 주성분 분석의 다른 활용 예로 변수를 선택하는 방법을 알아보겠다.

특징 선택이란 종속변수를 예측하기 위해 적당한 독립변수가 어떤 것인지 찾는 것을 말한다.

주성분 분석을 이용한 변수 선택은 일반적인 방법론은 아니지만, 주성분 분석의 개념을 응용한 방법으로 극히 일부에서 사용한다.

여기서 이 방법을 소개하는 이유는 일반적으로 알려진 알고리즘의 활용 예 외에도 다른 방법으로 응용 가능하다는 것을 소개하기 위한 것이므로 참고바란다.

Step1. 데이터 표준화

항상 해왔듯, 데이터 표준화 먼저한다.

$$z_i = \frac{(x_i - \bar{x})}{\sigma}$$

Step2. 상관행렬 계산

표준화 데이터를 이용해 상관행렬을 계산해보자. 표준화된 데이터의 행렬 Z 상관행렬 R은 다음과 같이 정의했다.

	범죄율	비소매지역비율	일산화질소농도	방개수	노후건물비율	고용센터접근성	고속도로접근성	재산세율	학생교사비율	흑인비율	하위계층비율	주택가격
범죄율	1	0.406583	0.420972	-0.21925	0.352734	-0.37967	0.625505	0.582764	0.289946	-0.38506	0.455621	-0.3883
비소매지역비율	0.406583	1	0.763651	-0.39168	0.644779	-0.70803	0.595129	0.72076	0.383248	-0.35698	0.6038	-0.48373
일산화질소농도	0.420972	0.763651	1	-0.30219	0.73147	-0.76923	0.611441	0.668023	0.188933	-0.38005	0.590879	-0.42732
방개수	-0.21925	-0.39168	-0.30219	1	-0.24026	0.205246	-0.20985	-0.29205	-0.3555	0.128069	-0.61381	0.69536
노후건물비율	0.352734	0.644779	0.73147	-0.24026	1	-0.74788	0.456022	0.506456	0.261515	-0.27353	0.602339	-0.37695
고용센터접근성	-0.37967	-0.70803	-0.76923	0.205246	-0.74788	1	-0.49459	-0.53443	-0.23247	0.291512	-0.497	0.249929
고속도로접근성	0.625505	0.595129	0.611441	-0.20985	0.456022	-0.49459	1	0.910228	0.464741	-0.44441	0.488676	-0.38163
재산세율	0.582764	0.72076	0.668023	-0.29205	0.506456	-0.53443	0.910228	1	0.460853	-0.44181	0.543993	-0.46854
학생교사비율	0.289946	0.383248	0.188933	-0.3555	0.261515	-0.23247	0.464741	0.460853	1	-0.17738	0.374044	-0.50779
흑인비율	-0.38506	-0.35698	-0.38005	0.128069	-0.27353	0.291512	-0.44441	-0.44181	-0.17738	1	-0.36609	0.333461
하위계층비율	0.455621	0.6038	0.590879	-0.61381	0.602339	-0.497	0.488676	0.543993	0.374044	-0.36609	1	-0.73766
주택가격	-0.3883	-0.48373	-0.42732	0.69536	-0.37695	0.249929	-0.38163	-0.46854	-0.50779	0.333461	-0.73766	1

위 표의 사관행렬을 살펴보면 데이터가 가진 변수 간의 관계를 살펴볼 수 있다. 범죄율과 방의 개수는 마이너스 상관관계로, 방의 개수가 늘어날수록 범죄율은 줄어든다는 것을 알 수 있다. 지극히 상식적인 이야기로, 방의 개수가 많은 동네는 부자 동네이므로 보안을 철저히 하기에 범죄율이 낮다는 이야기로 해석할 수 있다. 하지만 반대로도 생각할 수 있다. 도둑이라면 부잣집을 상대로 범죄를 저지르는 편이 한방에 일확천금을 노릴 수 있지 않을까? 라는 생각으로 데이터를 의심해 보는 것도 중요하다.

역시 상관행렬을 만든 후, 대칭행렬(대각선 성분의 값이 같은지)이 됐는지, 대각선의 값은 1인지 꼭 확인해야 한다.