Data Science

Correlation & Regression Analysis

노기섭 교수

(kafa46@hongik.ac.kr)

Lecture Goals

상관 분석 (Correlation Analysis)

- 상관의 개념, 공분산, 피어슨 상관계수
- 코사인 유사도
- 상관행렬과 히트맵 실습 (Boston Housing 예제)

단순 회귀 분석 (Simple Regression)

- 회귀의 개념과 최소제곱법 (MSE)
- np.polyfit 실습 (자동차 연비 예제)
- R^2(결정계수)의 의미와 해석

다중 회귀 분석 (Multiple Regression)

- Car Price 데이터 실습
- 회귀계수 해석, 모델 평가 (R^2, RMSE)
- 잔차 분석

상관 분석 (Correlation Analysis)

상관의 기본 개념

정의

- 두 변수가 얼마나 함께 움직이는지를 나타내는 통계적 척도

목적

- 두 변수가 함께 변하는 정도를 수치로 표현한 것
- 한 변수가 변할 때, 다른 변수의 변화 방향과 정도를 파악
- 상관계수(Correlation Coefficient) 로 수치화 (범위: -1 ~ 1)

상관의 종류

구분	설명
양의 상관	두 변수가 함께 증가 (공부시간 ↑ → 점수 ↑)
음의 상관	한 변수가 증가할 때 다른 변수는 감소 (운동시간 ↑ → 체중 ↓)
상관 없음	두 변수의 변화가 무관 (신발 사이즈 ↔ 수학점수)

상관계수의 해석

상관계수 (r)	관계 유형	해석
(r = +1)	완전한 양의 상관	두 변수 완벽히 같은 방향으로 이동
(0 < r < 1)	양의 상관	함께 증가하는 경향
(r = 0)	상관 없음	선형 관계 거의 없음
(-1 < r < 0)	음의 상관	한쪽 증가 시 다른쪽 감소
(r = -1)	완전한 음의 상관	완벽히 반대 방향으로 이동

피어슨 상관계수 (Pearson Correlation Coefficient)

개념

- 두 변수의 선형적 관계(Linear Relationship) 를 수치로 표현하는 지표
- 공분산(Covariance)을 표준화(Standardization) 한 형태
- 단위에 관계없이 -1 ≤ r ≤ 1 범위로 표현됨

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

공분산(Covariance)의 개념

- 두 변수가 함께 변하는 정도를 나타내는 값
- 방향성 해석
  - 양수 → 두 변수가 같은 방향으로 움직임
  - 음수 → 두 변수가 반대 방향으로 움직임
  - 0에 가까움 → 관계 없음

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

공분산 계산 예시

공부시간과 점수의 분포

학생	X (공부시간)	Y (점수)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
홍길동	1	2	-1	-1.67	1.67
이순신	2	3	0	-0.67	0.00
유관순	3	6	1	2.33	2.33

$$Cov(X,Y) = \frac{1.67 + 0 + 2.33}{3} = 1.33$$

공분산이 양수 → 같은 방향으로 함께 증가

$$r = \frac{Cov(X,Y)}{STD_X \times STD_Y}$$

공분산의 한계

- 단위(scale)에 따라 값이 달라짐(예: '시간' vs '분' 단위로 계산 시 값이 커지거나 작아짐)
- 서로 다른 단위의 변수 간 비교 어려움 ➔ 정규화된 지표인 피어슨 상관계수 사용

공분산 요약

■ 주의할 점

- 상관 ≠ 인과관계(Causation)
  - 상관계수가 높다고 해서 한 변수가 다른 변수를 '원인'으로 만든다는 뜻은 아님.
- 예시:아이스크림 판매량 ↑ 익사 사고 ↑

  → 둘 다 여름(계절)의 영향Slide

■ 핵심 요약

- 공분산: 방향성만 알려줌
- 피어슨 상관계수: 단위 제거 후 비교 가능
- r 값으로 선형 관계의 방향과 강도 파악
- 해석 시 반드시 맥락(Context) 고려

코사인 유사도 (Cosine Similarity)

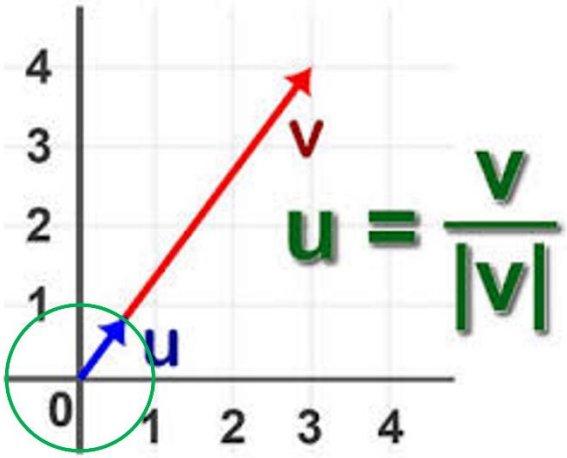
개념

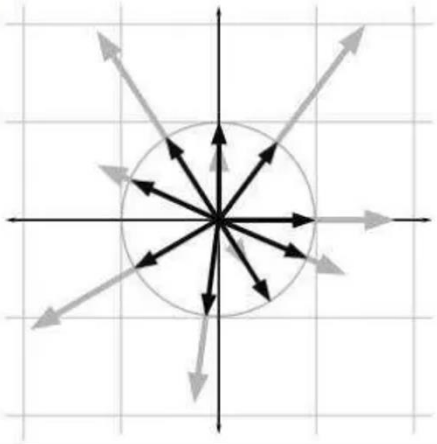
- 두 벡터의 방향(각도) 기반 유사도 측정 지표
- 벡터의 크기나 단위에 영향을 받지 않음
- 방향이 같을수록 유사도가 높음

$$Cosine\ Similarity(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

수학적 해석

- $x \cdot y$: 두 벡터의 내적
- $\|x\|, \|y\|$: 각 벡터의 크기 (L2 Norm)
- 값의 범위:
  - 1: 완전히 같은 방향
  - 0: 직각, 관계 없음
  - -1: 완전히 반대 방향





코사인 유사도 – 예시

■ **같은 방향**

$$x = [1, 2, 3], \qquad y = [2, 4, 6]$$

$$x \cdot y = 28$$

$$\|x\| = \sqrt{14}$$

$$\|y\| = \sqrt{56}$$

$$Cosine\ Similarity(x, y) = \frac{28}{\sqrt{14} \cdot \sqrt{56}} = 1$$

■ **벡터의 방향이 약간 달라진다면?**

$$x = [1, 2, 3], \quad y = [101, 102, 103]$$ ← **평행 이동된 벡터**

$$Cosine\ Similarity(x, y) < 1$$

- 코사인 유사도는 1보다 작음.
- 그러나 피어슨 상관계수는 여전히 1

피어슨 상관계수 vs. 코사인 유사도

구분	피어슨 상관계수	코사인 유사도
초점	평균 제거 후 선형 관계 측정	벡터의 방향적 유사성 측정
정규화 기준	평균 및 표준편차	L2 노름 (벡터 길이)
단위 영향	평균 제거로 일부 제거됨	완전히 제거됨
활용 분야	통계, 상관 분석	텍스트, 문서, 추천 시스템

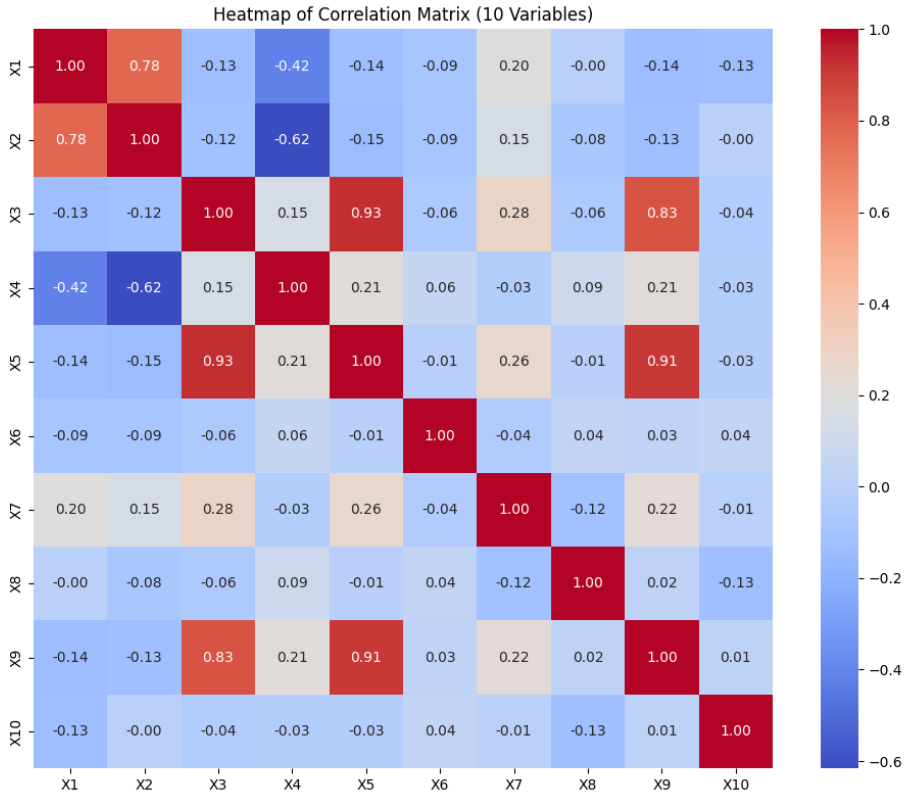
- 코사인 유사도는 **데이터의 이동(shift)** 에 민감
- 즉, 모든 값에 일정한 상수를 더하면 벡터의 방향이 바뀌어 유사도가 달라질 수 있음
- 반면 피어슨 상관계수는 평균을 제거하기 때문에 이동에 영향을 받지 않음
- 코사인 유사도는 고차원 벡터(예: 문장 임베딩, 사용자-아이템 벡터)에서

  두 벡터의 **방향적 유사성**을 측정할 때 유용

상관행렬 및 히트맵 시각화

실습 코드

https://www.deepshark.org/courses/data_science/w/06_correlation_regression#correlation_matrix



양의 상관(positive correlation) 값은 붉은색 계열,

음의 상관(negative correlation) 값은 파란색 계열,

0에 가까운 값은 흰색 또는 중간색으로 표시

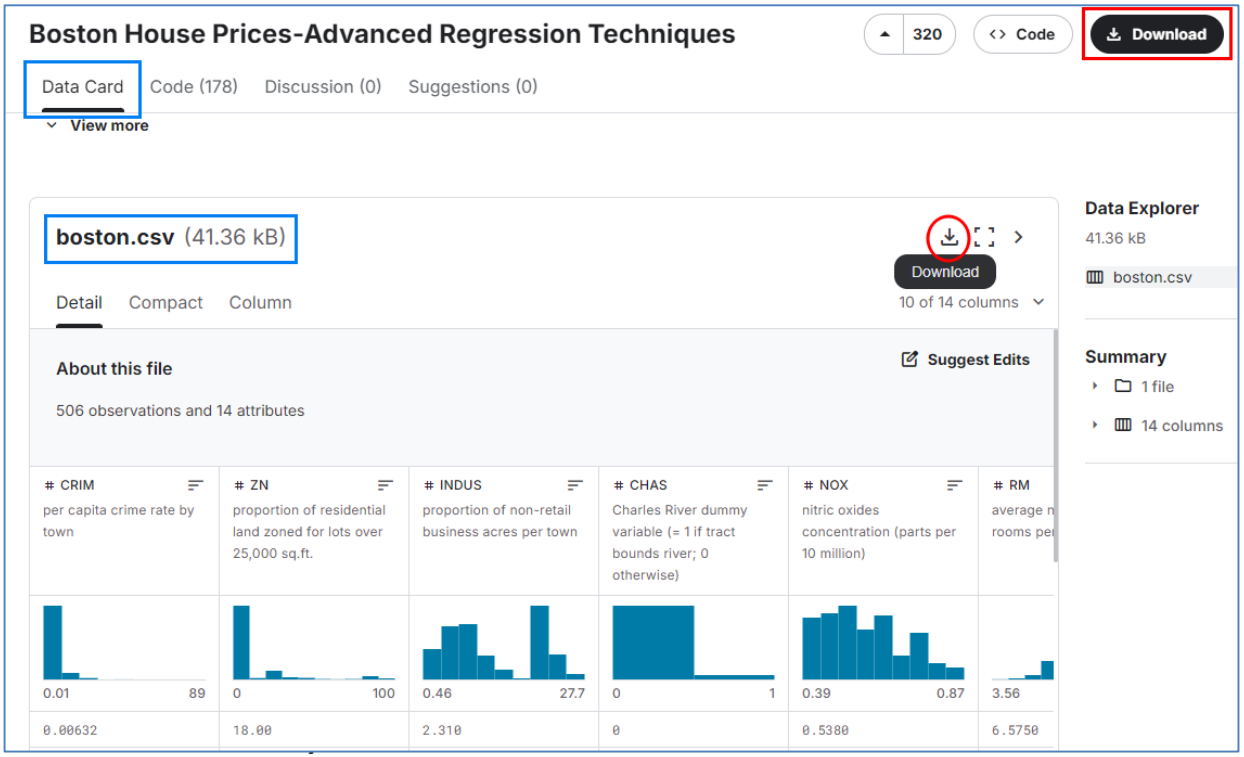
Kaggle 데이터 분석 (주택 가격에 영향을 주는 요인 분석)

도시의 주택 가격은 여러 요인(방 수, 인근 범죄율, 학교 수준, 공기 질 등)에 영향을 받는다.

실습을 통해 Kaggle의 Boston House Price 데이터셋을 사용하여 다음을 분석한다.

- 변수 간의 상관관계 분석
- 상관행렬 및 히트맵 시각화
- 주요 요인 탐색

■ 데이터셋 다운로드



보스톤 주택 변수 설명

변수	설명
CRIM	마을별 1인당 범죄 발생률
ZN	25,000 평방피트(약 2323㎡) 이상의 대형 주택용 토지로 지정된 주거용 토지 비율
INDUS	마을별 비소매(non-retail) 상업 지역 비율
CHAS	찰스강과 접해 있는지 여부를 나타내는 더미 변수 (1: 접함, 0: 접하지 않음)
NOX	대기 중 질소 산화물 농도 (1000만분의 1 단위)
RM	주택당 평균 방 개수
AGE	1940년 이전에 지어진 자가주택의 비율
DIS	보스턴의 5개 주요 고용 중심지까지의 가중 거리
RAD	고속도로 접근성 지수
TAX	재산세율 (10,000달러당 세금액)
PTRATIO	학생-교사 비율
B	흑인 인구 비율을 기반으로 계산된 지표. 식: $1000(Bk - 0.63)^2$, 여기서 Bk는 흑인 인구 비율
LSTAT	하위 계층(저소득층) 인구의 비율(%)
MEDV	자가주택의 **중앙 가격** (단위: 1,000달러)

보스톤 집값 변수별 상관관계 분석 절차

■ 데이터 불러오기 및 기본 확인

- pandas로 CSV 불러오기
- df.info(), df.describe(), df.isnull().sum() 등으로 결측치/변수 유형 확인

■ 수치형 변수만 추출하여 상관행렬 계산

- df.corr() 사용
- MEDV (자가주택의 중앙 가격)와의 상관계수를 내림차순 정렬하여 주요 요인 탐색

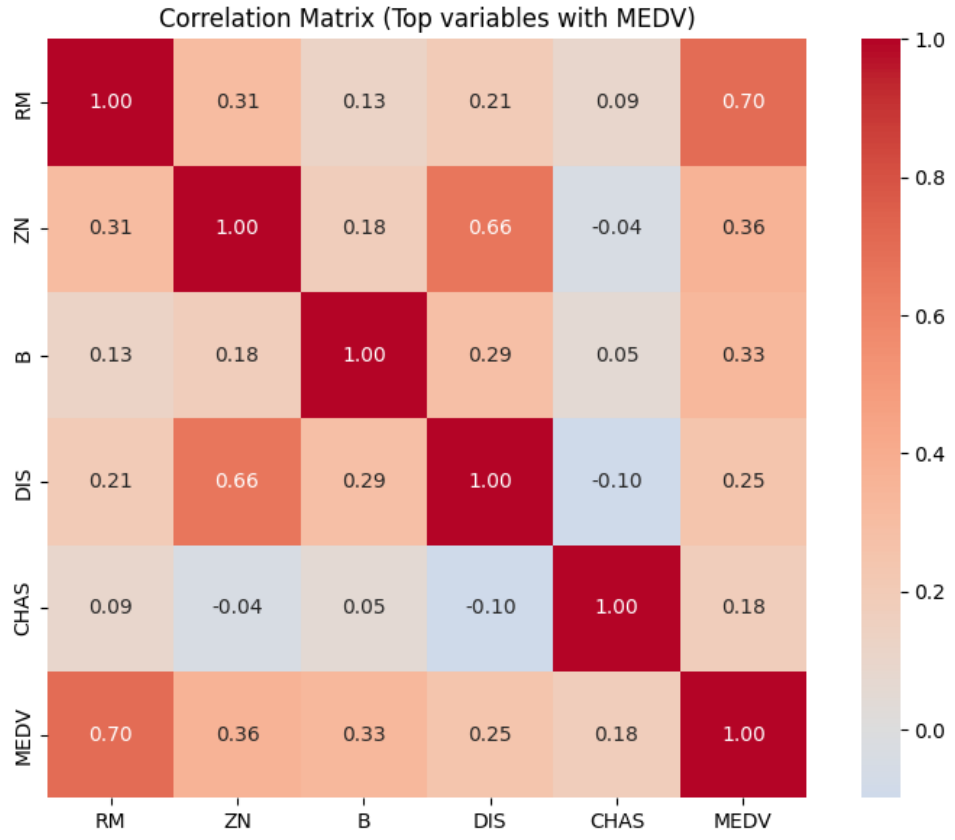
■ 히트맵 시각화

- sns.heatmap으로 전체 상관행렬 (correlation matrix) 시각화
- MEDV (자가주택의 중앙 가격) 상관이 높은 상위 5개 변수 만 뽑아 별도 히트맵 생성

구현 및 분석

소스 코드

https://www.deepshark.org/courses/data_science/w/06_correlation_regression#boston_house_code



순위	변수	상관계수	변수 설명
1	**RM**	0.695	주택당 평균 방 개수 방이 많을수록 집값이 높아지는 경향
2	**ZN**	0.360	25,000 평방피트(약 2,323㎡) 이상의 대형 주택용 토지 비율 고급 주거지일수록 집값이 높음
3	**B**	0.333	흑인 인구 비율 기반 지표: $1000(Bk - 0.63)^2$ 인종 관련 지표로 당시 집값과 통계적 상관
4	**DIS**	0.250	보스턴 주요 고용 중심지까지의 거리 거리가 멀수록 선호도가 낮아질 수 있음
5	**CHAS**	0.175	찰스강 인접 여부 (1: 인접) 강 인접 지역의 집값이 더 높음

단순 회귀 분석

단순 회귀분석(Simple Linear Regression)

■ 하나의 **독립 변수(Independent Variable)** x 와 **종속 변수(Dependent Variable)** y 사이의 **선형 관계(Linear Relationship)** 를 모델링하는 통계적 방법

- 즉, x가 변할 때 y가 어떻게 변하는지를 하나의 직선(line) 으로 근사하여 예측하는 것이다.
- 예를 들어, 공부 시간(x)과 시험 점수(y) 간의 관계

  공부 시간이 많을수록 점수가 높아진다면, 두 변수 사이에는 양의 선형 관계가 존재한다.

$$y = \beta_0 + \beta_1 x + \epsilon$$

- y: 종속 변수 (예측하려는 값)
- x: 독립 변수 (입력 또는 설명 변수)
- β_0: 절편 (intercept), $x = 0$ 일 때의 값
- β_1 : 기울기 (slope), x 가 1 단위 증가할 때 의 변화량
- ϵ: 오차항 (error term), 실제 값과 예측 값의 차이, 실제값 y가 회귀선에서 얼마나 벗어나는지를 설명
  - ϵ 값은 우리가 직접 알 수 없는 확률적 잡음(random noise)에 해당

최소제곱법(OLS; Ordinary Least Squares)

- 회귀선은 모든 데이터 점들로부터의 거리(잔차, residual)가 최소가 되도록 설정
  - 즉, 각 점의 오차 제곱을 합한 값을 SSE (Sum of Squared Error) 최소화하는 방법을 최소제곱법 (OLS) 이라고 한다.

$$Minimize \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

MSE (Mean Squared Error)

- 잔차 제곱합을 데이터 개수 n으로 나눈 평균값, 즉 평균제곱오차
- OLS의 평균을 구하는 점만 다를 뿐 최소화 되는 지점은 동일함

$$Minimize\ MSE$$

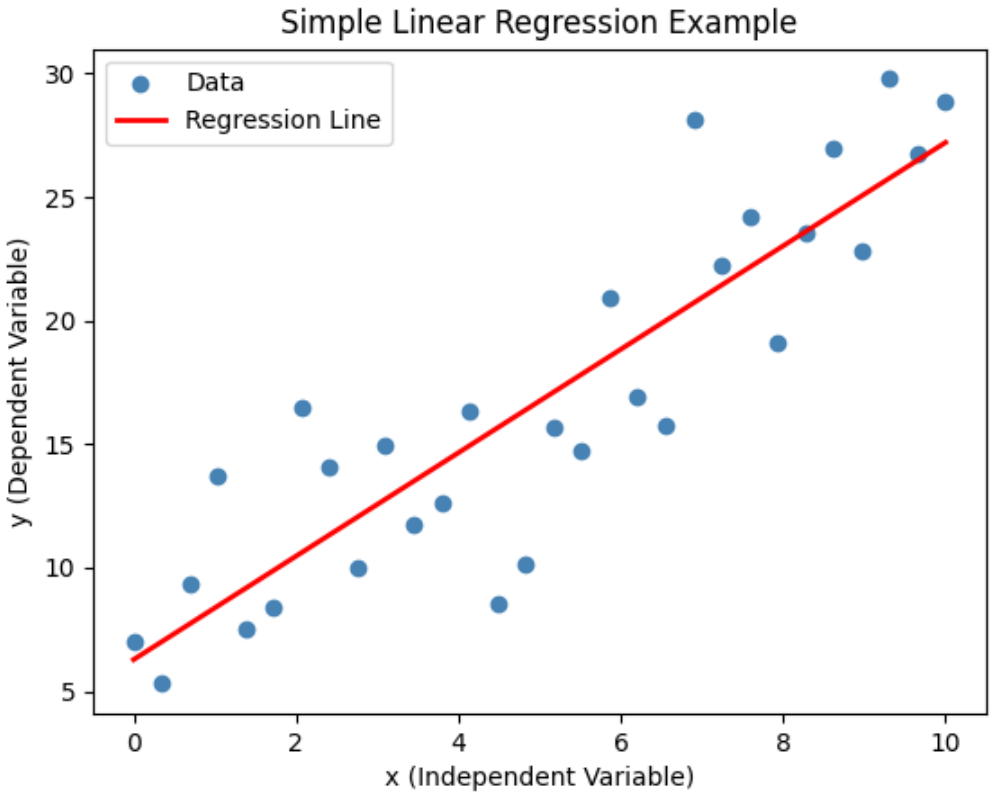
$$= Minimize \frac{SSE}{n}$$

$$= Minimize \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

시각적 이해

소스 코드

https://www.deepshark.org/courses/data_science/w/06_correlation_regression#simple_regression



회귀 계수: [2.0923798 6.28551344]

$y = 2.0923798x + 6.28551344$

Kaggle 데이터 분석: 자동차 연비 예측

- 자동차의 연비(MPG, Miles Per Gallon)는 엔진 성능과 차량의 물리적 특성에 따라 달라진다.
- Kaggle의 Auto MPG 데이터셋을 사용하여 엔진의 마력(horsepower)이 연비(mpg, mile per gallon) 에 어떤 영향을 미치는지를 분석하라.

데이터 다운로드

- https://www.kaggle.com/datasets/uciml/autompg-dataset

변수 설명

변수명	설명
mpg	자동차의 연비 (Miles per gallon, 종속 변수)
cylinders	실린더 수
displacement	배기량 (cubic inches)
horsepower	마력 (horse power)
weight	차량 무게 (lbs)
acceleration	0→60mph 가속 시간 (seconds)
model year	생산 연도 (model year)
origin	생산 지역 코드 (1: 미국, 2: 유럽, 3: 일본)
car name	차량 이름

구현 및 분석

데이터 분석 코드

https://www.deepshark.org/courses/data_science/w/06_correlation_regression#auto_mpg_code1

단순 회귀 분석 코드: Horsepower → MPG

https://www.deepshark.org/courses/data_science/w/06_correlation_regression#auto_mpg_code2

- 출력 결과
  - 절편 (intercept): 39.93586102117047
  - 기울기 (slope): -0.15784473335365365
  - R^2: 0.606
  - RMSE: 4.893

R^2 (결정계수, Coefficient of Determination)는 회귀 모델이 실제 데이터를 얼마나 잘 설명하는지를 나타내는 지표

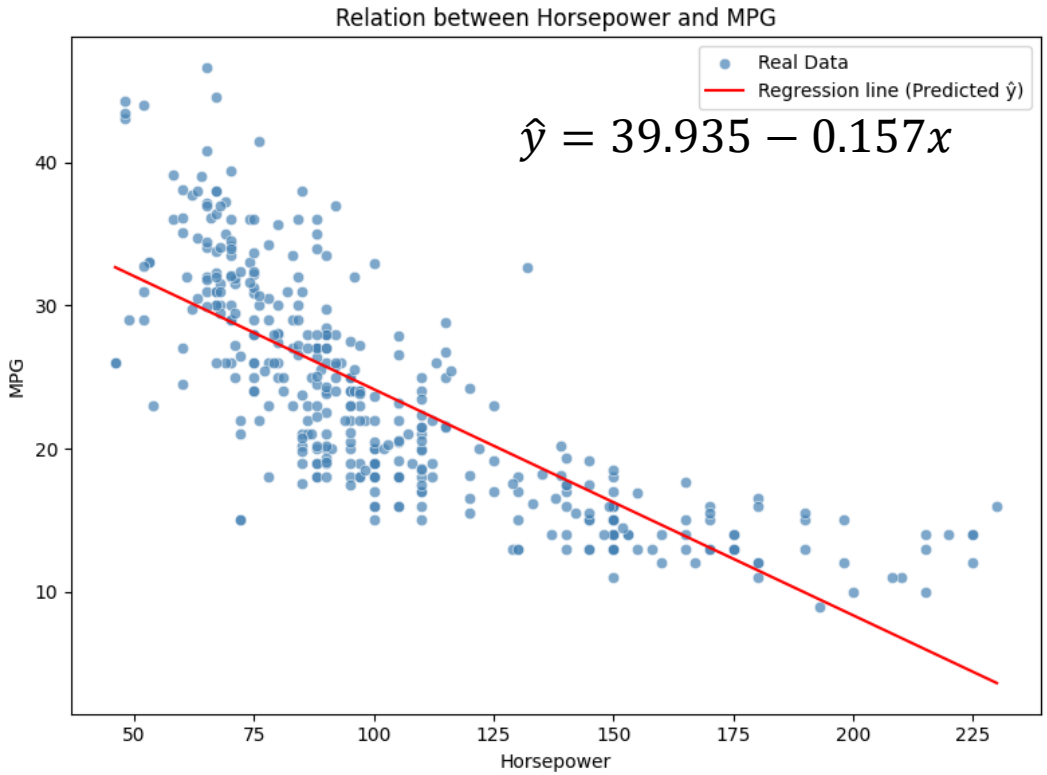
$$R^2 = 1 - \frac{SSE\ (sum\ of\ squared\ errors)}{SST(total\ sum\ of\ squares)}$$

- "차량의 마력(horsepower)만으로 연비(MPG) 변동의 약 60.6%를 설명할 수 있다."
- 나머지 40%는 차량 무게, 배기량, 운전 습관 등의 다른 요인에 의해 발생

시각화: 회귀선과 데이터 분포

시각화 코드

- https://www.deepshark.org/courses/data_science/w/06_correlation_regression#auto_mpg_code3



- 마력이 1 증가할 때마다 평균적으로 연비는 약 0.157 MPG 감소한다.
- 즉, 고성능(고마력) 차량일수록 연비 효율은 낮아지는 경향을 보인다.
- 결정계수 R^2 = 0.606는, 연비의 약 60%가 마력 변수로 설명될 수 있음을 의미한다.

다중 회귀 분석

다중 회귀분석(Multiple Linear Regression)

■ **다중 회귀분석(Multiple Linear Regression) 은 두 개 이상의 독립 변수가 동시에 종속 변수에 어떤 영향을 미치는지를 분석**

- 여러 요인을 함께 고려해 종속 변수의 변동을 설명하고 예측 정확도를 높이는 모델이다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- y : 종속 변수 (예측 대상)
- x_i : 독립 변수들
- β_0 : 절편 (Intercept)
- β_i : 각 독립 변수의 회귀계수. 다른 변수들이 일정할 때, 해당 변수의 변화가 y에 미치는 영향
- ϵ : 오차항, 모델이 설명하지 못하는 부분(잡음)

Kaggle 데이터 분석: 자동차 가격 예측 (다중 회귀)

- 자동차의 판매 가격은 엔진 크기, 마력, 연식, 브랜드 등 여러 요인의 영향을 받는다.
- Kaggle의 Car Price Prediction 데이터셋을 이용해 자동차 가격(price)이 어떤 변수들에 의해 설명되는지를 분석하라.

데이터셋 다운로드

- https://www.kaggle.com/datasets/hellbuoy/car-price-prediction

변수 설명

변수명	설명
symboling	보험 위험 등급
wheelbase	축간 거리 (inch)
carlength, carwidth, carheight	차량 크기
curbweight	공차중량 (lbs)
enginesize	엔진 배기량 (cc)
horsepower	마력 (hp)
peakrpm	최대 회전수 (rpm)
citympg, highwaympg	도심/고속도로 연비
price	자동차 판매가격 (종속변수)

실습 단계

실습 코드 흐름

- 데이터 불러오기 및 확인
- 특징 변수 선택 (다중 독립변수 구성)
- 학습/검증 데이터 분리
- 모델 학습 및 회귀계수 해석
- 모델 성능 평가 (R^2, RMSE)
- 시각화 (실제 vs 예측 비교)

소스 코드

https://www.deepshark.org/courses/data_science/w/06_correlation_regression#car_price

실행 결과

데이터 크기: (205, 26)

	price	enginesize	horsepower	curbweight	citympg
0	13495.0	130	111	2548	21
1	16500.0	130	111	2548	21
2	16500.0	152	154	2823	19
3	13950.0	109	102	2337	24
4	17450.0	136	115	2824	18

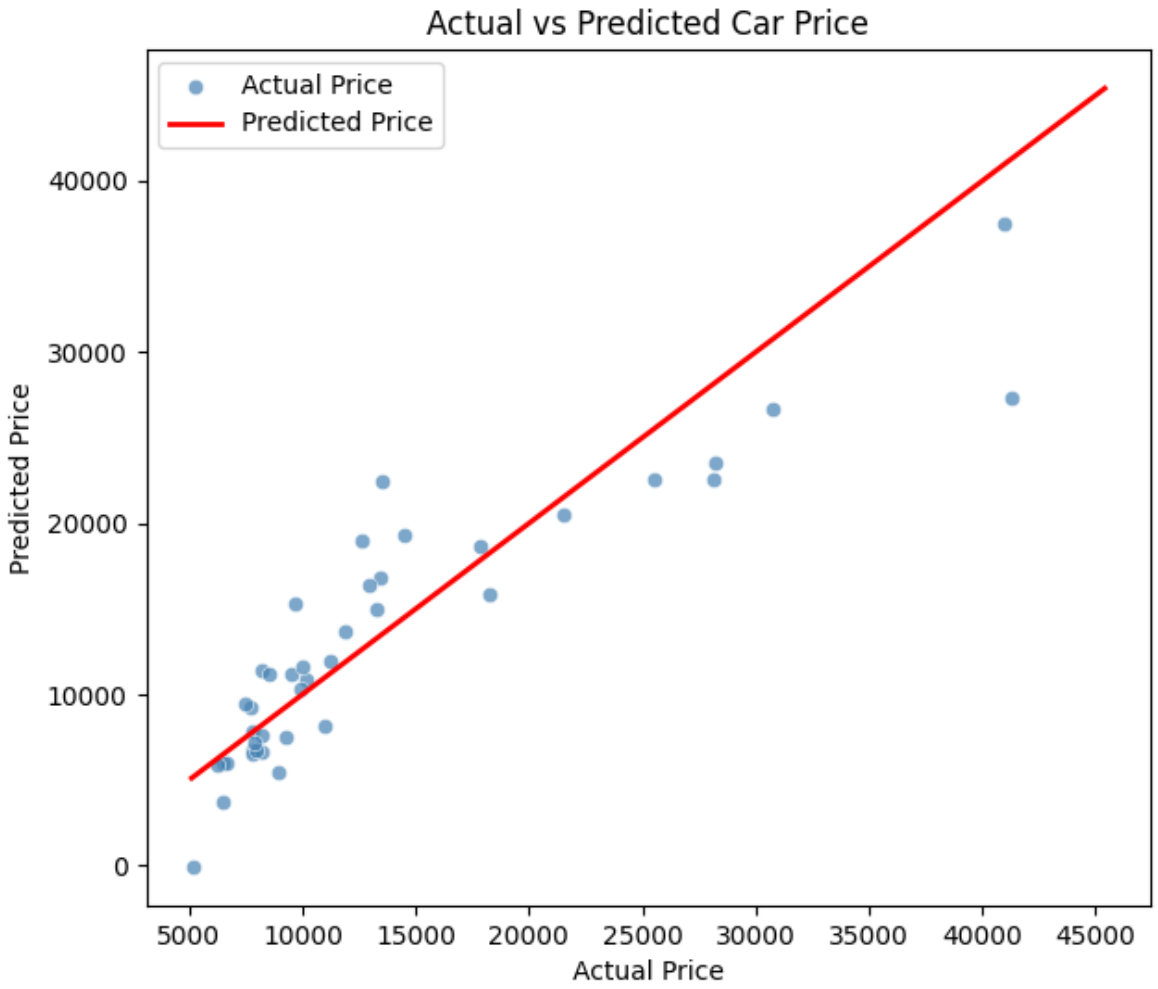
회귀계수:

	Variable	Coefficient
0	enginesize	80.397
1	horsepower	48.838
2	curbweight	3.936
3	citympg	-47.919

절편 (Intercept): -10913.718

R^2: 0.816, RMSE: 3814.81

시각화: Actual vs Predicted Car Price



결과 해석

■ **모델식**

$$\hat{y} = -10913 + 80.4 \times enginesize + 48.84 \times horsepower + 3.94 \times curbweight - 47.92 \times citympg$$

■ **해석**

- 엔진 크기(enginesize)와 마력(horsepower)이 클수록 차량의 가격이 상승한다.
- 차량 무게(curbweight)가 높을수록 가격이 다소 상승하는 경향이 있다.
- 연비(citympg)가 높을수록(연비가 좋은 차일수록) 가격이 낮은 경향 — 고성능 차량일수록 연비가 낮기 때문이다.

■ **R^2 = 0.816**

- 모델이 자동차 가격 변동의 약 81.6%를 설명한다.
- 나머지 18.4%는 브랜드, 디자인, 옵션 등 비수치적 요인으로 설명 가능하다.

■ **RMSE = 3814.81**

- 예측 오차의 표준편차 수준이 약 3,800달러임을 의미한다.

잔차(Residual)의 개념

■ **잔차(Residual):** 실제값과 예측값의 차이

$$e_i = y_i - \hat{y}_i$$

- 모델이 각 데이터 포인트를 얼마나 잘 예측했는지를 보여주는 지표
- 회귀모델의 성능이 높더라도, 잔차의 분포를 분석해야 모델이 선형성을 가정한 것이 타당한지 확인할 수 있다.

■ 이상적인 잔차의 특징

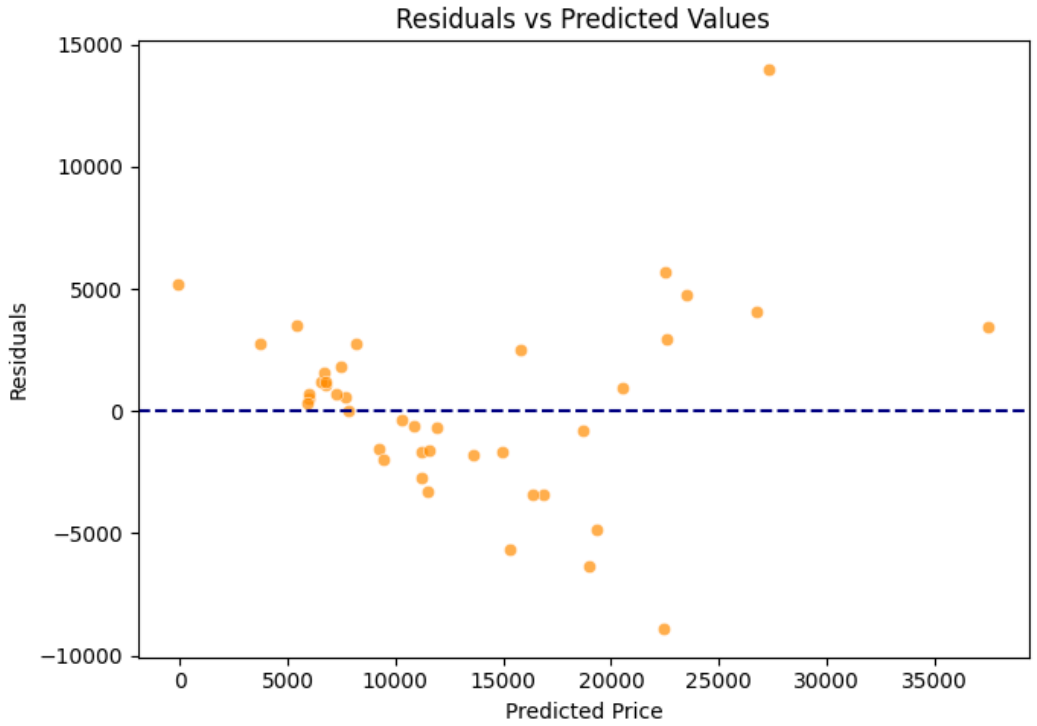
조건	설명
평균이 0에 가까움	예측값과 실제값의 평균적인 차이가 거의 없음
등분산성 (Homoscedasticity)	예측값의 크기에 따라 잔차의 분산이 일정해야 함
독립성 (Independence)	잔차들이 서로 독립적이어야 함 (시간 순서나 패턴이 없어야 함)
정규성 (Normality)	잔차의 분포가 정규분포에 가까워야 함

잔차 시각화 실습

- **실습 코드**

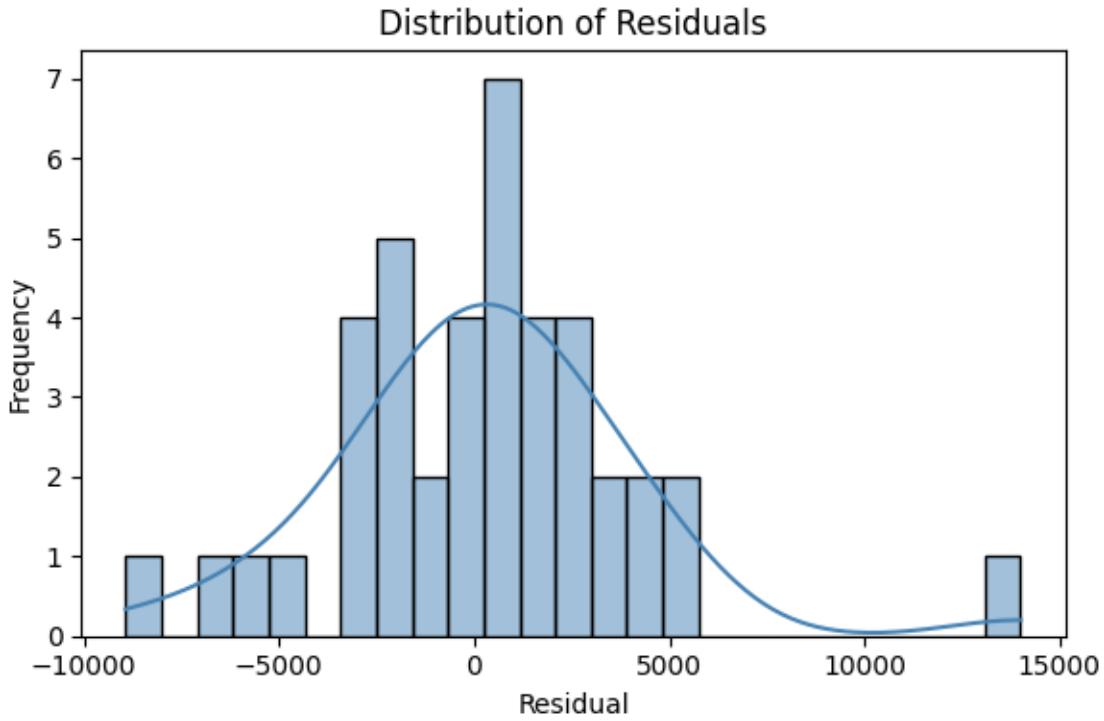
  https://www.deepshark.org/courses/data_science/w/06_correlation_regression#residual_plot

Residual vs. Predicted Plot



평가 항목	관찰 결과	해석
평균 중심성	잔차가 0을 중심으로 대체로 대칭적	편향 없는 모델
등분산성	고가 구간에서 분산 증가	이분산성 일부 존재
선형성	곡선 패턴 없음	선형 가정 타당
이상치	일부 극단값 존재	고가 차량 데이터 영향 가능성

Residual Distribution Plot

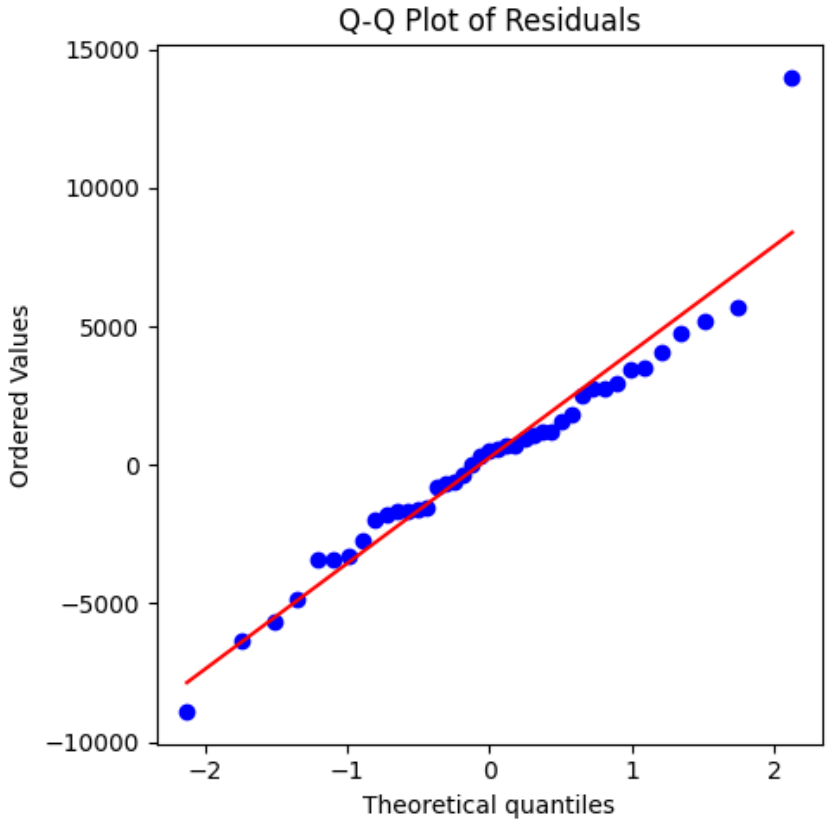


평가 항목	관찰 결과	해석
평균 중심성	잔차의 평균이 0 근처	예측에 편향 없음
분포 형태	약간의 비대칭, 종 모양 유지	정규성 대체로 만족
극단값 존재	±10,000 근처 일부 존재	고가 차량의 예측 오차 영향 가능성

QQ plot of Residuals

Q-Q(Quantile–Quantile) Plot

- 회귀모델의 잔차(residuals)가 정규분포(normal distribution)를 따르는지 시각적으로 검증
- 가로축: 정규분포의 이론적 분위 값 (Theoretical Quantiles)
- 세로축: 실제 잔차의 분위값(Ordered Values)
- 빨간 선: 완벽한 정규분포를 따를 경우 점들이 위치해야 할 기준선



평가 항목	관찰 결과	해석
중앙부 패턴	직선에 근접	잔차 대부분이 정규분포 따름
꼬리 부분	약간 벗어남	일부 이상치 존재
전체 형태	전반적으로 직선형	정규성 가정 대체로 만족

수고하셨습니다 ..^^..