

# 2장 머신러닝 프로젝트 처음부터 끝까지

## 2.1 실전 데이터 활용

## 실전 데이터 저장소

다양한 실전 데이터를 모아놓은 데이터 저장소를 머신러닝 공부에 잘 활용할 수 있어야 한다.

- [OpenML](#)
- [캐글\(Kaggle\) 데이터셋](#)
- [페이퍼스 위드 코드](#)
- [UC 얼바인\(UC Irvine\) 대학교 머신러닝 저장소](#)
- [아마존 AWS 데이터셋](#)
- [텐서플로우 데이터셋](#)

## 분석 대상 데이터셋

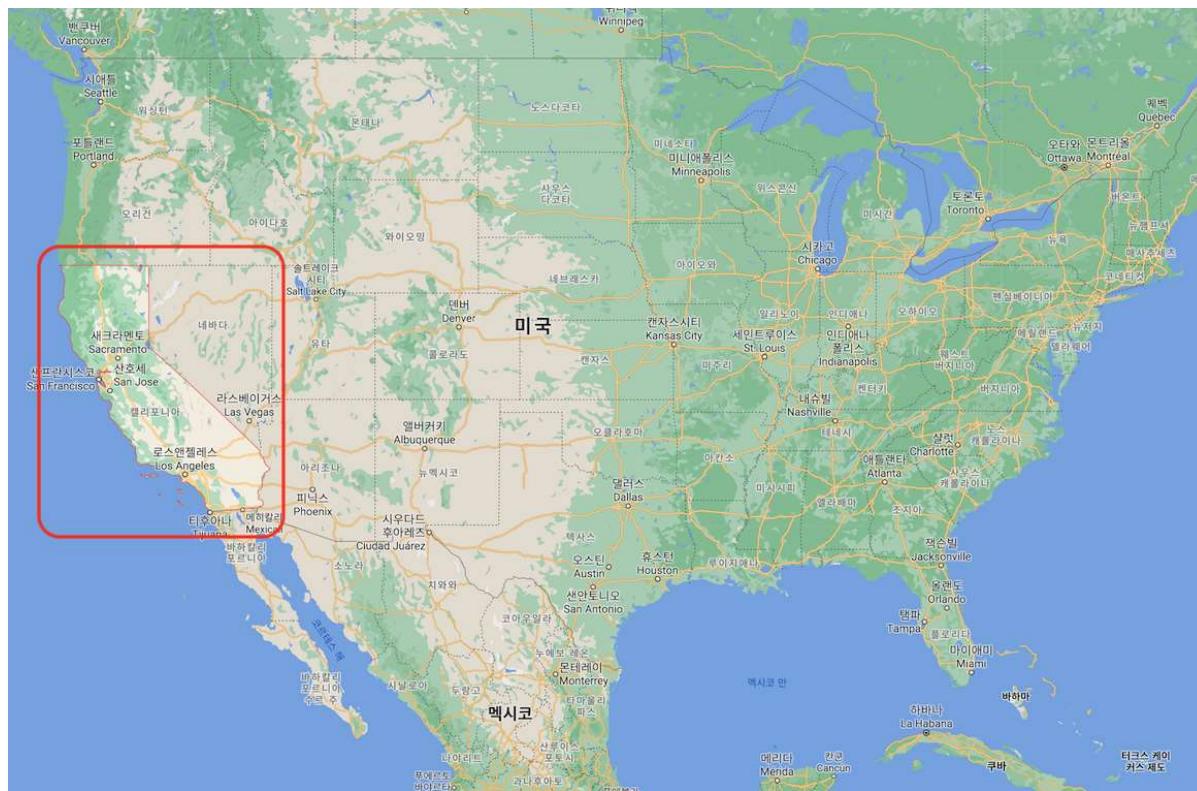
- 1990년 미국 캘리포니아 주에서 수집한 인구조사 데이터

## 2.2 큰 그림 그리기

## 데이터 정보 확인

- 미국 캘리포니아 주의 20,640개 지역별 인구조사 데이터
- 특성 10개: 경도, 위도, 중간 주택 연도, 방의 총 개수, 침실 총 개수, 인구, 가구 수, 중간 소득, 중간 주택 가격, 해안 근접도
- 목표: 구역별 중간 주택 가격 예측 시스템(모델) 구현하기

- 미국 캘리포니아 지도



## 훈련 모델 확인

- 지도 학습(supervised learning)
  - 타깃: 구역별 중간 주택 가격
- 회귀: 중간 주택 가격 예측
  - 다중 회귀: 여러 특성을 활용한 예측
  - 단변량 회귀: 구역마다 하나의 가격만 예측
- 배치 학습: 빠르게 변하는 데이터에 적응할 필요가 없음

## 훈련 모델 성능 측정 지표

선형 회귀 모델의 경우 일반적으로 아래 두 기준 중 하나를 사용한다.

- 평균 제곱근 오차(RMSE)
- 평균 절대 오차(MAE)

## 평균 제곱근 오차(root mean square error, RMSE)

- 유클리디안 노름(Euclidean norm) 또는  $\ell_2$  노름(norm)으로도 불림
- 참고: 노름(norm)은 거리 측정 기준을 나타냄.

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - y^{(i)})^2}$$

- 기호 설명
  - $\mathbf{X}$ : 훈련용 입력 데이터셋 전체 샘플들의 특성값으로 구성된 행렬, 타깃 제외.
  - $\mathbf{x}^{(i)}$ :  $i$  번째 입력 샘플의 전체 특성값 벡터.
  - $y^{(i)}$ :  $i$  번째 샘플의 타깃
  - $h$ : 예측 함수
  - $\hat{y}^{(i)} = h(\mathbf{x}^{(i)})$ :  $i$  번째 샘플에 대한 예측 값

## 훈련셋과 2D 어레이

- 훈련 입력 데이터셋에  $m$  개의 샘플이 포함되어 있고 각각의 샘플이  $n$  개의 특성을 갖는 경우의 훈련 입력 데이터셋:

( $m, n$ ) 모양의 numpy의 2D 어레이

- 예제:  $m = 5, n = 4$  인 경우의 훈련 입력 데이터셋  $\mathbf{X}$ :

```
array([[-118.29, 33.91, 1416, 38372],  
..... [-114.30, 34.92, 2316, 41442],  
..... [-120.38, 35.21, 3444, 29303],  
..... [-122.33, 32.95, 2433, 24639],  
..... [-139.31, 33.33, 1873, 50736]])
```

- 각각의  $\mathbf{x}^{(i)}$ 는  $i$  번째 행에 해당. 예를 들어  $\mathbf{x}^{(1)}$ 은 첫째 행의 1D 어레이:

```
array([-118.29, 33.91, 1416, 38372])
```

## 평균 절대 오차(mean absolute error, MAE)

- MAE는 맨해튼 노름 또는  $\ell_1$  노름으로도 불림

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^m |h(\mathbf{x}^{(i)}) - y^{(i)}|$$

- 이상치가 많은 경우 활용
- RMSE가 MAE보다 이상치에 더 민감하지만, 이상치가 많지 않을 경우 일반적으로 RMSE 사용

## 2.3 데이터 훑어보기

## 전제 조건

- `housing` 변수에 데이터프레임으로 적재된 캘리포니아 인구조사 데이터셋 할당

## 데이터 기본 정보 확인

- pandas의 데이터프레임 활용
- `head()`, `info()`, `describe()`, `hist()` 등을 사용하여 데이터 구조 훑어보기

```
housing.head()
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

## housing.info()

#	Column	Non-Null Count	Dtype
0	longitude	20640 non-null	float64
1	latitude	20640 non-null	float64
2	housing_median_age	20640 non-null	float64
3	total_rooms	20640 non-null	float64
4	total_bedrooms	20433 non-null	float64
5	population	20640 non-null	float64
6	households	20640 non-null	float64
7	median_income	20640 non-null	float64
8	median_house_value	20640 non-null	float64
9	ocean_proximity	20640 non-null	object

- 구역 수: 20,640개
- 구역별로 경도, 위도, 중간 주택 연도, 해안 근접도 등 총 10개의 조사 항목
  - '해안 근접도'는 범주형 특성이고 나머지는 수치형 특성.
- '방의 총 개수'의 경우 누락된 데이터인 207개의 null 값 존재

## 범주형 특성 탐색

- '해안 근접도'는 5개의 범주로 구분

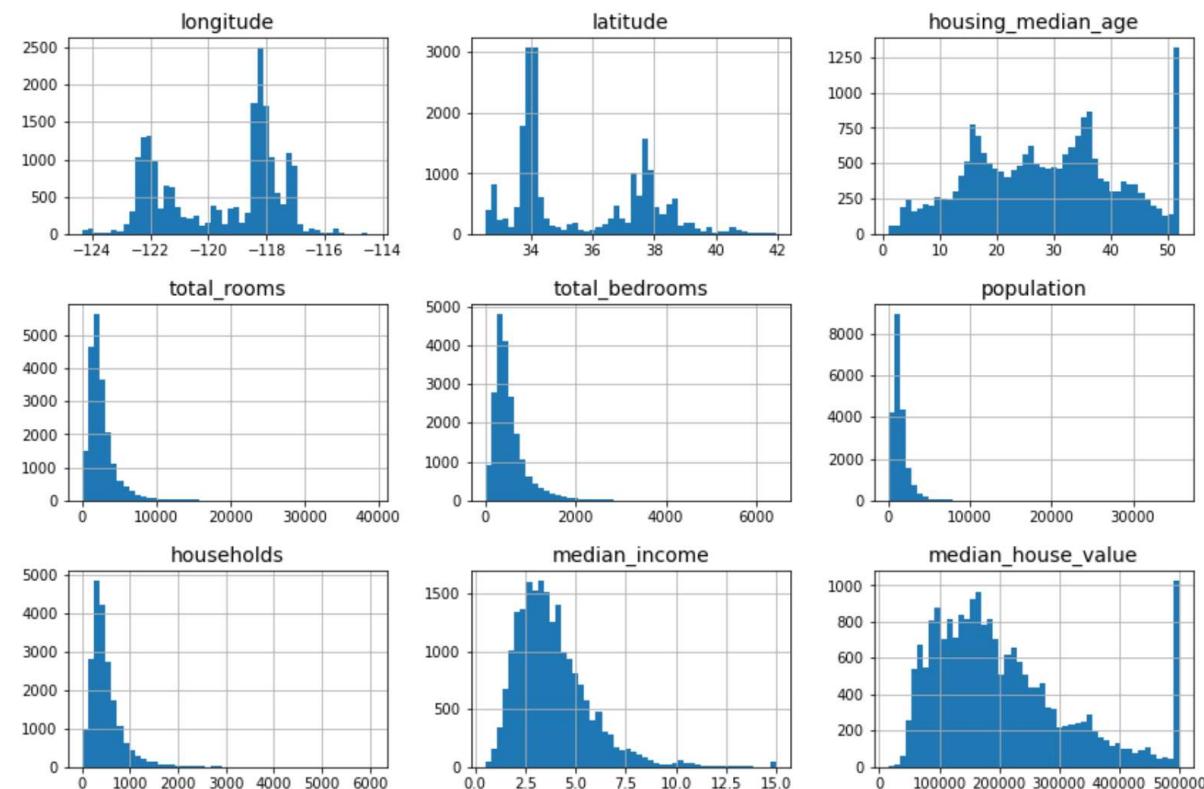
특성값	설명
<1H OCEAN	해안에서 1시간 이내
INLAND	내륙
NEAR OCEAN	해안 근처
NEAR BAY	샌프란시스코의 Bay Area 지역
ISLAND	섬

# 수치형 특성 탐색

```
housing.hist(bins=50, figsize=(12, 8))
```

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
<b>count</b>	20640.000000	20640.000000	20640.000000	20640.000000	20433.000000	20640.000000	20640.000000	20640.000000	20640.000000
<b>mean</b>	-119.569704	35.631861	28.639486	2635.763081	537.870553	1425.476744	499.539680	3.870671	206855.816909
<b>std</b>	2.003532	2.135952	12.585558	2181.615252	421.385070	1132.462122	382.329753	1.899822	115395.615874
<b>min</b>	-124.350000	32.540000	1.000000	2.000000	1.000000	3.000000	1.000000	0.499900	14999.000000
<b>25%</b>	-121.800000	33.930000	18.000000	1447.750000	296.000000	787.000000	280.000000	2.563400	119600.000000
<b>50%</b>	-118.490000	34.260000	29.000000	2127.000000	435.000000	1166.000000	409.000000	3.534800	179700.000000
<b>75%</b>	-118.010000	37.710000	37.000000	3148.000000	647.000000	1725.000000	605.000000	4.743250	264725.000000
<b>max</b>	-114.310000	41.950000	52.000000	39320.000000	6445.000000	35682.000000	6082.000000	15.000100	500001.000000

## 수치형 특성별 히스토그램



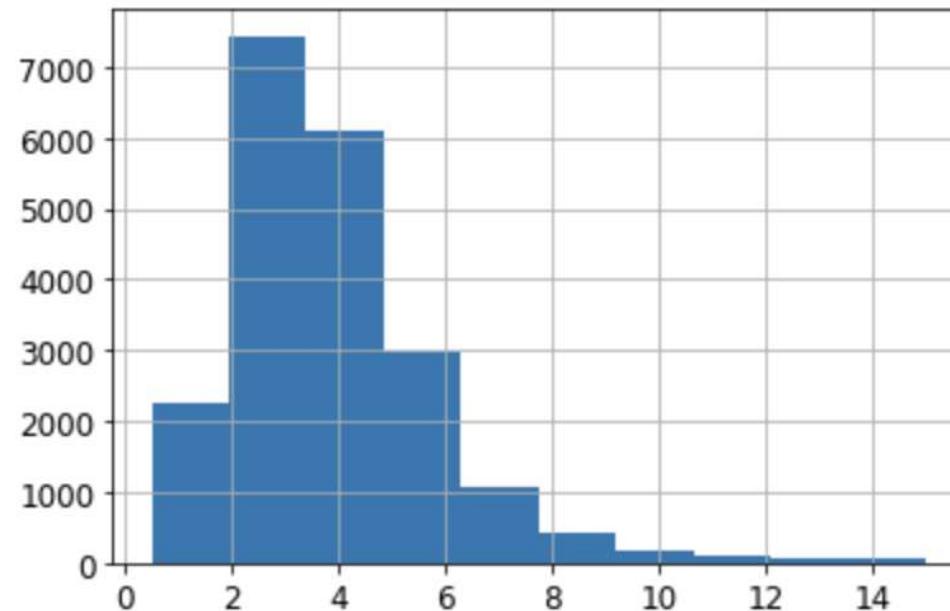
## 훈련셋과 테스트셋

- 모델 학습 시작 이전에 준비된 데이터셋을 훈련셋과 테스트셋로 구분
  - 테스트셋 크기: 전체 데이터 셋의 20%
- 테스트셋에 포함된 데이터는 미리 분석하지 말 것.
  - 테스트셋을 미리 분석하면 미래의 데이터 특징을 미리 가정하는 **데이터 스누핑 편향** 오류를 범할 가능성이 높아짐
  - 미리 보면서 알아낸 직관이 학습 모델 설정에 영향을 미칠 수 있음
- 훈련셋과 데이터 세트를 구분하는 방식에 따라 결과가 조금씩 달라짐
  - 무작위 샘플링 vs. 계층 샘플링
- 여기서는 계층 샘플링 활용

## 계층 샘플링

- 계층: 동질 그룹. 예를 들어, 소득별 계층 등 사용.
- 테스트셋: 전체 계층을 대표하도록 각 계층별로 적절한 샘플 추출
- 예제: 소득 범주
  - 계층별로 충분한 크기의 샘플이 포함되도록 지정해야 학습 과정에서 편향이 발생하지 않음
  - 특정 소득 구간에 포함된 샘플이 과하게 적거나 많으면 해당 계층의 중요도가 과대 혹은 과소 평가됨

- 전체 데이터셋의 중간 소득 히스토그램 활용



- 대부분 구역의 중간 소득이 **1.5~6.0**, 즉 15,000에서 60,000 달러 사이

- 소득 구간을 아래 숫자를 기준으로 5개로 구분

구간	범위
1	0 ~ 1.5
2	1.5 ~ 3.0
3	3.0 ~ 4.5
4	4.5 ~ 6.0
5	6.0 ~

## 계층 샘플링과 무작위 샘플링 비교

소득 구간	전체(%)	계층 샘플링(%)	무작위 샘플링(%)	계층 샘플링 오류율	무작위 샘플링 오류율(%)
1	3.98	4.00	4.24	0.36	6.45
2	31.88	31.88	30.74	-0.02	-3.59
3	35.06	35.05	34.52	-0.01	-1.53
4	17.63	17.64	18.41	0.03	4.42
5	11.44	11.43	12.09	-0.08	5.63

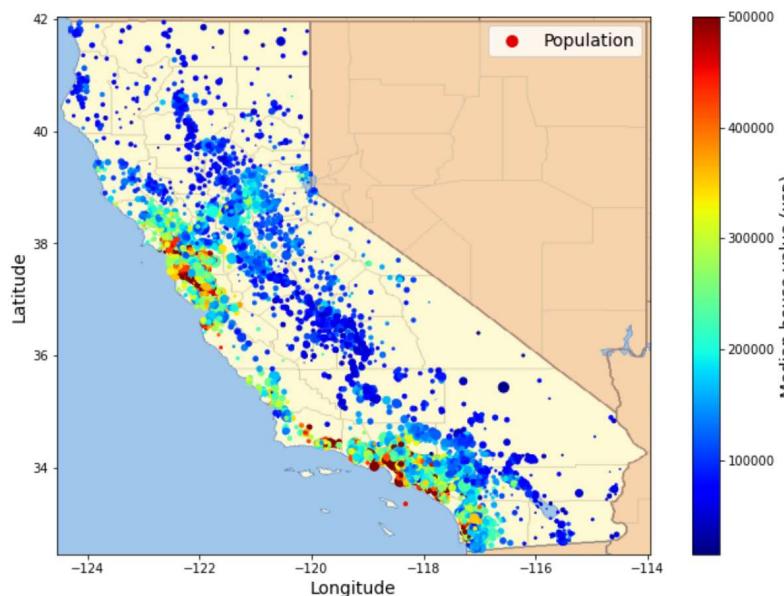
## 2.4 데이터 탐색과 시각화

## 주의 사항

- 테스트셋을 제외한 훈련셋에 대해서만 시각화를 이용하여 탐색
- 데이터 스누핑 편향 방지 용도

## 지리적 데이터 시각화

- 구역이 집결된 지역과 그렇지 않은 지역 구분 가능
- 주택 가격이 해안 근접도 또는 인구 밀도와 관련이 큼. 샌프란시스코의 베이 에어리어, LA, 샌디에고 등 밀집된 지역 확인 가능
- 해안 근접도: 위치에 따라 다르게 작용
  - 대도시 근처에선 해안 근처 주택 가격이 상대적 높지만 북부 캘리포니아 지역은 다름

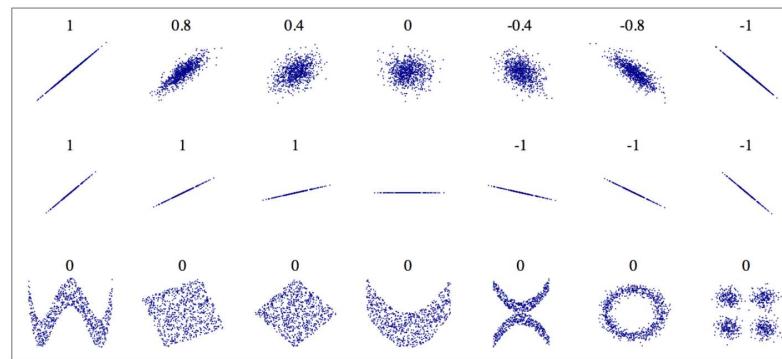


## 상관관계 조사

- 중간 주택 가격 특성과 다른 특성 사이의 상관관계: 상관계수 활용

```
median_house_value    1.000000
median_income       0.688380
total_rooms        0.137455
housing_median_age 0.102175
households         0.071426
total_bedrooms     0.054635
population        -0.020153
longitude          -0.050859
latitude           -0.139584
```

## 상관계수의 특징

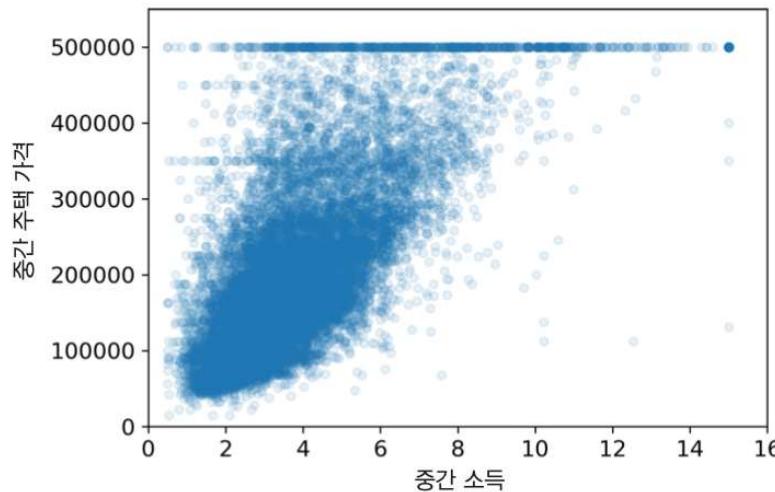


<그림 출처: [위키백과](#)>

- 상관계수:  $[-1, 1]$  구간의 값
- 1에 가까울 수록: 강한 양의 선형 상관관계
- -1에 가까울 수록: 강한 음의 선형 상관관계
- 0에 가까울 수록: 매우 약한 선형 상관관계. 하지만 다른 관계 존재 가능.
- 기울기 정도와 아무 연관 없음

## 중간 주택 가격과 중간 소득의 상관관계

- 상관계수가 0.68로 가장 높음
- 중간 소득이 올라가면 중간 주택 가격도 상승하는 경향이 있음
- 점들이 너무 넓게 퍼져 있음. 완벽한 선형관계와 거리 멎.
- 50만 달러 수평선: 가격 제한 결과로 보임
  - 45만, 35만, 28만, 그 아래 정도에서도 수평선 존재. 이유는 알려지지 않음.
  - 이상한 형태를 학습하지 않도록 해당 구역을 제거하는 것이 좋음. (여기서는 그대로 두고 사용)



## 특성 조합 활용

- 구역별 방의 총 개수와 침실의 총 개수 대신 아래 특성이 보다 유용함
  - 가구당 방 개수(rooms for household)
  - 방 하나당 침실 개수(bedrooms for room)
  - 가구당 인원(population per household)
- 중간 주택 가격과 방 하나당 침실 개수의 연관성 다소 있음