

2장 머신러닝 프로젝트 처음부터 끝까지 (1부)

개요



2.1. 실전 데이터 활용

실전 데이터 저장소

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

캘리포니아 주택가격 데이터

- 1990년 미국 캘리포니아 주에서 수집한 주택가격 데이터

	A	B	C	D	E	F	G	H	I	J
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
2	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
3	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
4	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
5	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
6	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
7	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
8	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
9	-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY
10	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
11	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY
12	-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY
13	-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY
14	-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY
15	-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY
16	-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY
17	-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY
18	-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY
19	-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY
20	-122.26	37.84	50	2239	455	990	419	1.9911	158700	NEAR BAY

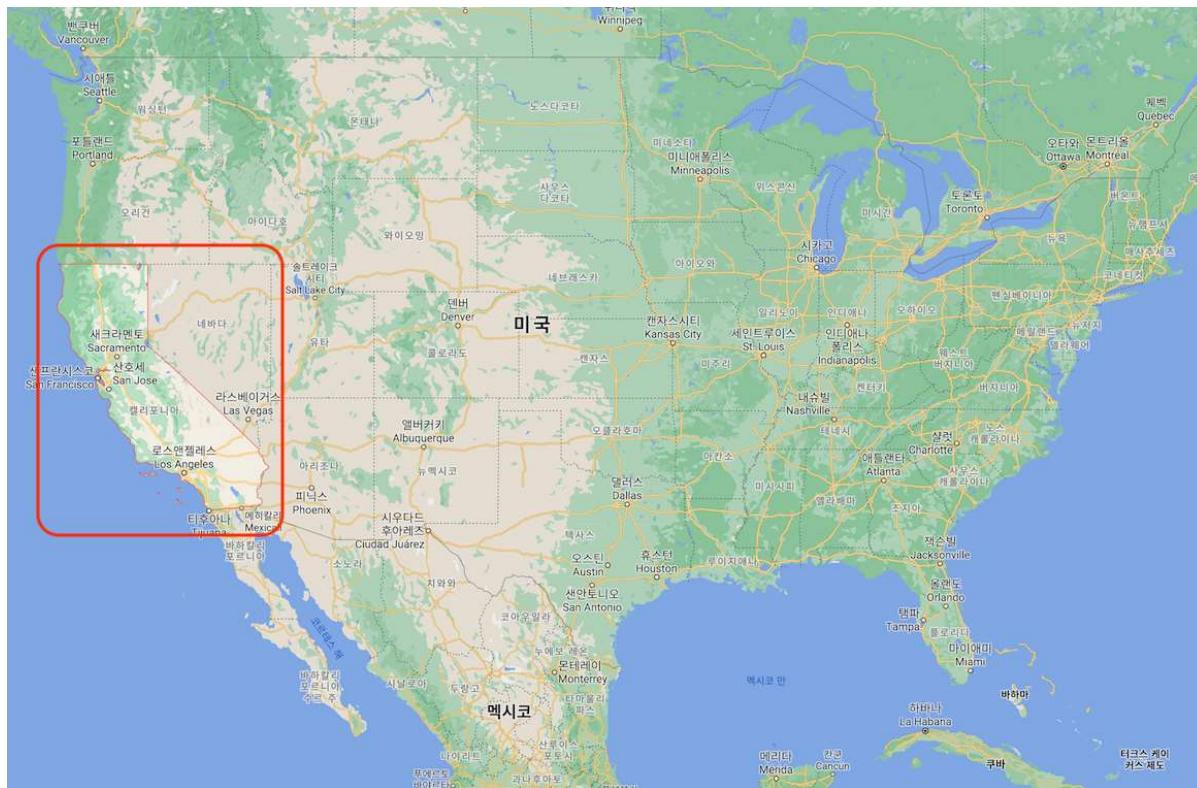
2.2. 큰 그림 그리기

- 데이터 기초 정보 확인
- 훈련 모델 확인

2.2.1. 데이터 기초 정보 확인

- 미국 캘리포니아 주의 20,640개 지역별 주택가격 데이터
- 특성 10개: 경도, 위도, 주택 건물 중위연령, 총 방 수, 총 침실 수, 인구, 가구 수, 중위소득, 주택 중위가격, 해안 근접도
- 목표: 구역별 주택 중위가격 예측 시스템 구현

미국 캘리포니아 지도



2.2.2. 훈련 모델 확인

- 지도 학습: 타깃은 구역별 주택 중위가격.
- 회귀: 주택 중위가격 예측, 즉 이산형 값이 아닌 연속형 값 예측.
 - 다중 회귀: 구역별로 여러 특성을 주택 가격 예측에 사용
 - 단변량 회귀: 구역별로 한 종류의 값만 예측
- 배치 학습: 빠르게 변하는 데이터에 적응할 필요가 없으며, 데이터셋의 크기도 충분히 작기에 데이터셋 전체를 대상으로 훈련 진행

이산형 데이터 vs 연속형 데이터

- 이산형 데이터: 1, 2, 3, 등 값과 값 사이를 명확하게 구분할 수 있는 데이터
- 연속형 데이터: 유리수, 실수 처럼 두 개의 값 사이에 항상 새로운 값이 존재하는 데이터

2.3. 데이터 구하기

- 캘리포니아 주택가격 데이터: 많은 공개 저장소에서 다운로드 가능
- 여기서는 개인 깃허브 리포지토리에 압축파일로 저장한 파일을 다운로드
- `load_housing_data()` 함수: 캘리포니아 주택가격 데이터를 다운로드한 후에 Pandas 데이터프레임으로 반환

```
housing = load_housing_data()
```

2.4. 데이터 탐색과 시각화

- 데이터프레임과 데이터 탐색
- 훈련셋과 테스트셋
- 데이터 시각화

2.4.1. 데이터프레임과 데이터 탐색

- pandas의 데이터프레임 활용
- `head()`, `info()`, `describe()`, `hist()` 등 데이터프레임 메서드를 사용하여 데이터 기초 정보 확인
- 범부형/수치형 특성 탐색

head() 메서드

```
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

- 데이터 일부 확인 용도

info() 메서드

```
housing.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
 #   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  
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

- 20,640개 구역별로 10개의 특성 조사.
- 해안 근접도를 뜻하는 `ocean_proximity` 특성은 범주형 categorical이고 나머지는 수치형 numerical 특성임.
- 총 방 수를 뜻하는 `total_bedrooms` 특성은 207개의 null 값, 즉 결측치 포함.

범주형 특성 탐색: value_counts() 메서드

```
housing[ "ocean_proximity" ].value_counts()
```

```
<1H OCEAN      9136
INLAND         6551
NEAR OCEAN     2658
NEAR BAY        2290
ISLAND            5
Name: ocean_proximity, dtype: int64
```

- '해안 근접도'는 5개의 범주로 구분
- value_counts() 메서드: 사용된 특성값과 각각의 특성값이 사용된 횟수 확인

해안 근접도 특성값

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

수치형 특성 탐색: `describe()` 메서드

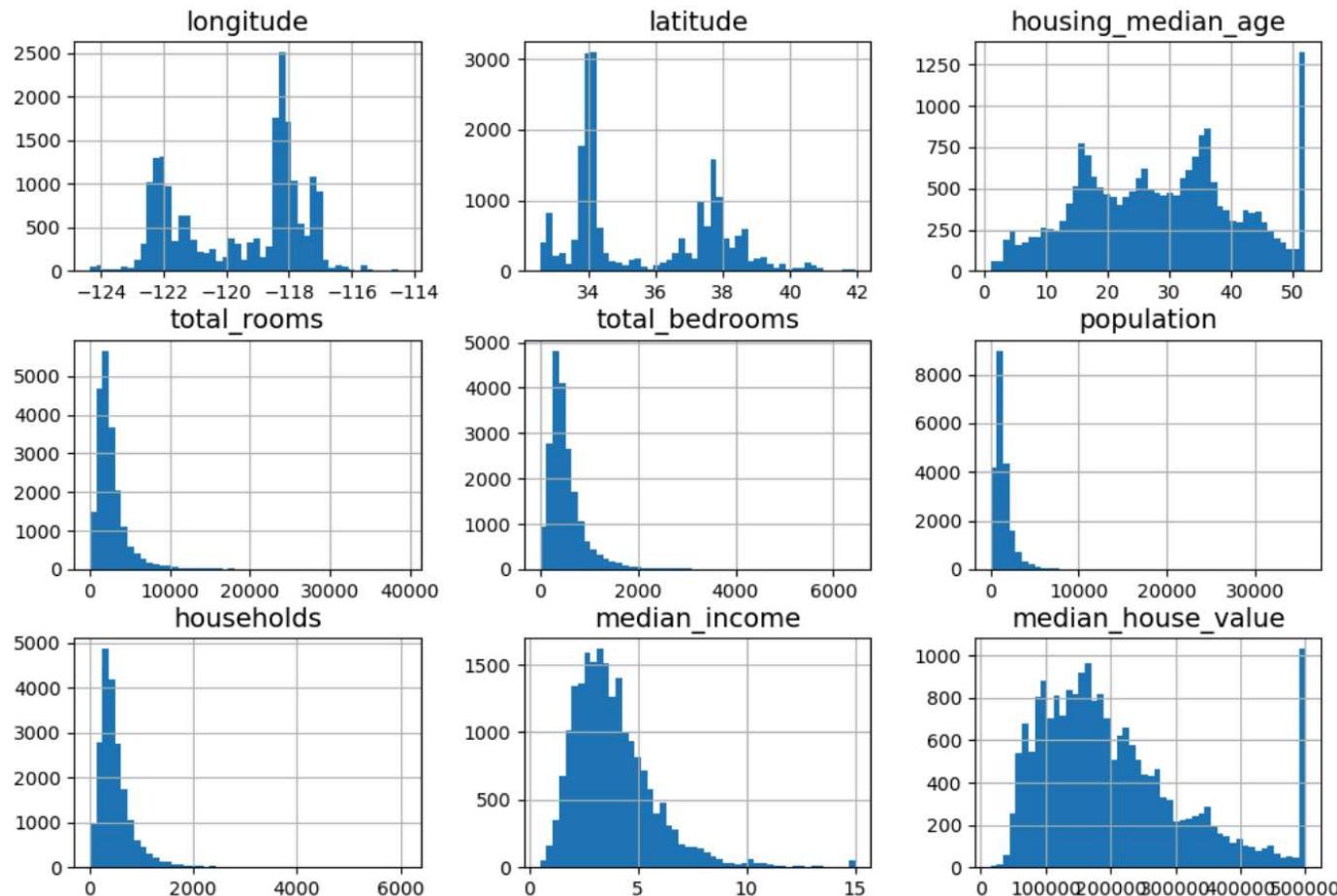
```
housing.describe()
```

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

- 수치형 특성들의 정보 요약: 평균값, 표준편차, 사분범위

수치형 특성별 히스토그램

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

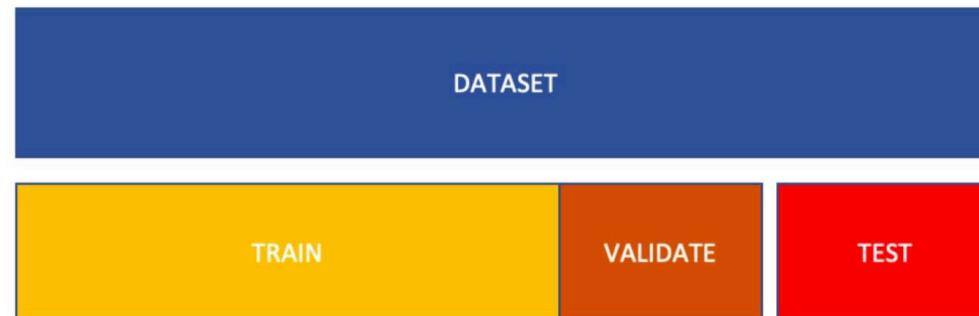


수치형 특성별 데이터 분포

- 각 특성마다 사용되는 단위와 스케일(척도)이 다르다. 1 단위부터 만 단위까지 다양하다.
- 일부 특성은 한쪽으로 치우쳐져 있다. 예를 들어 `total_rooms`, `total_bedrooms`, `population`, `households` 등의 특성값들이 오른쪽 꼬리를 길게 갖는다.
- 일부 특성은 값을 제한한 것으로 보인다. 예를 들어 `housing_median_age`, `median_house_value` 등의 특성값 상한값이 임의로 지정되어 잘린 것처럼 보인다.

2.4.2. 훈련셋과 테스트셋

- 모델 학습 시작 이전에 준비된 데이터셋을 훈련셋과 테스트셋으로 구분.
- 테스트셋 크기: 전체 데이터 셋의 20% 이하. 너무 크지 않게.
- 훈련셋의 일부는 훈련 중에 훈련의 진척 정도를 측정하는 검증 용도로 활용.
- 훈련셋과 데이터 세트를 구분하는 방식에 따라 결과가 조금씩 달라짐
 - 무작위 샘플링 vs. 계층 샘플링
 - 여기서는 계층 샘플링 활용

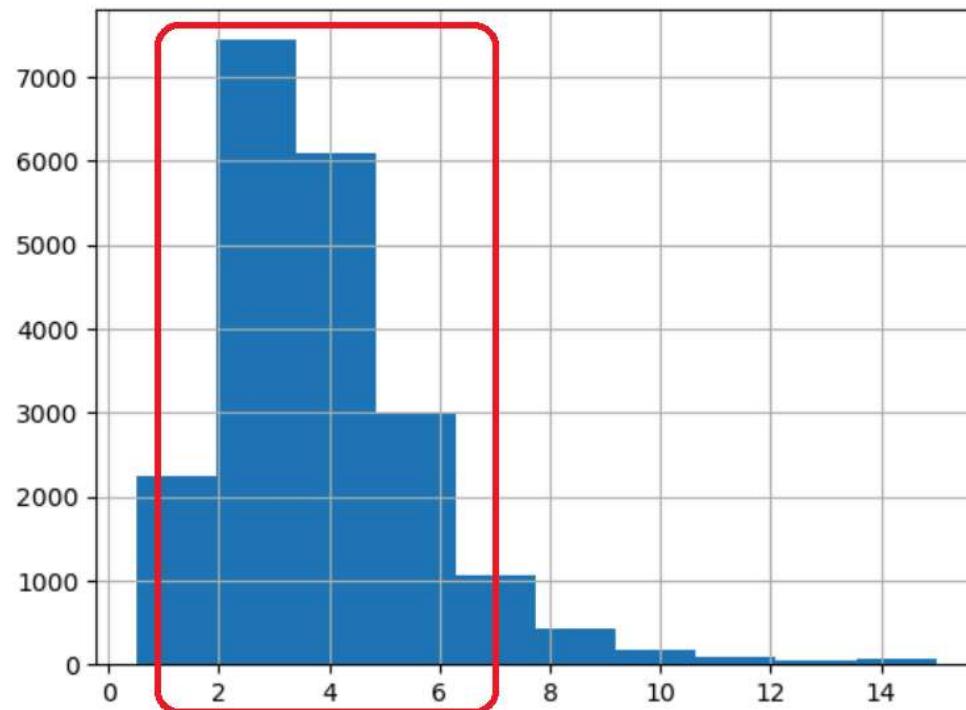


계층 샘플링

- 계층: 유사한 성질의 데이터로 구성된 그룹.
- 계층 샘플링: 계층별로 적절한 양의 샘플을 추출하는 기법
- 계층별로 충분한 크기의 샘플이 훈련셋으로 추출되어야 훈련 과정에서 편향이 발생하지 않음.
- 여기서는 소득 구간으로 구분된 계층 활용. 주택가격에 가장 큰 영향을 주는 특성이 기 때문임.
- 특정 소득 구간에 포함된 샘플이 과하게 적거나 많으면 해당 계층의 중요도가 과대 혹은 과소 평가됨

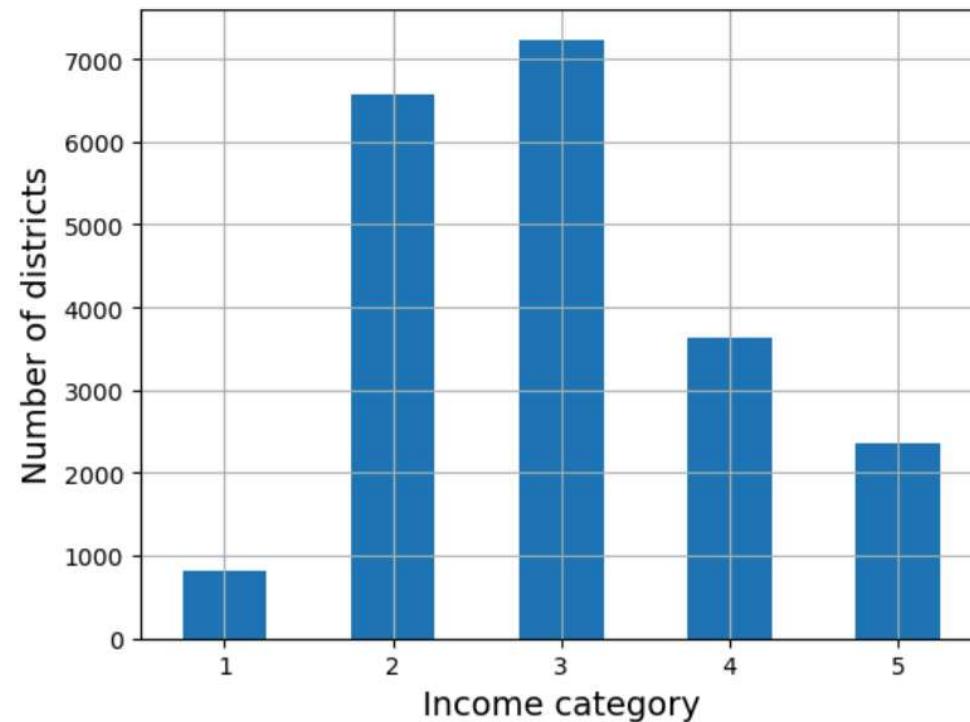
중위소득 히스토그램

- 대부분 구역의 중위소득이 **1.5~6.0**, 즉 15,000에서 60,000 달러 사이임을 확인



소득 구간별 데이터 분포

```
housing["income_cat"].value_counts().sort_index().plot.bar(rot=0, grid=True)
```



계층 샘플링과 무작위 샘플링 비교

- 계층 샘플링을 활용한 훈련셋과 테스트셋 구분

```
from sklearn.model_selection import train_test_split  
  
strat_train_set, strat_test_set = train_test_split(housing,  
                                                 test_size=0.2,  
                                                 stratify=housing[ "income_cat" ],  
                                                 random_state=42)
```

- 무작위 샘플링을 활용한 구분과의 비교

소득 구간	전체(%)	계층 샘플링(%)	무작위 샘플링(%)
1	3.98	4.00	4.24
2	31.88	31.88	30.74
3	35.06	35.05	34.52
4	17.63	17.64	18.41
5	11.44	11.43	12.09

2.4.3. 데이터 시각화

- 테스트셋을 제외한 훈련셋에 대해서만 시각화를 이용하여 탐색. 아래 코드는 `housing` 변수를 훈련셋을 가리키도록 업데이트.

```
housing = strat_train_set.copy()
```

- 테스트셋에 대해서 훈련 전에 너무 많은 정보를 알게 되면 이를 이용하여 훈련을 달리할 수 있음.
- 그러면 제대로 된 실전 테스트를 진행할 수 없게 되어 훈련된 모델의 실전 성능을 정확히 파악하기 어려움.

지리적 데이터 시각화

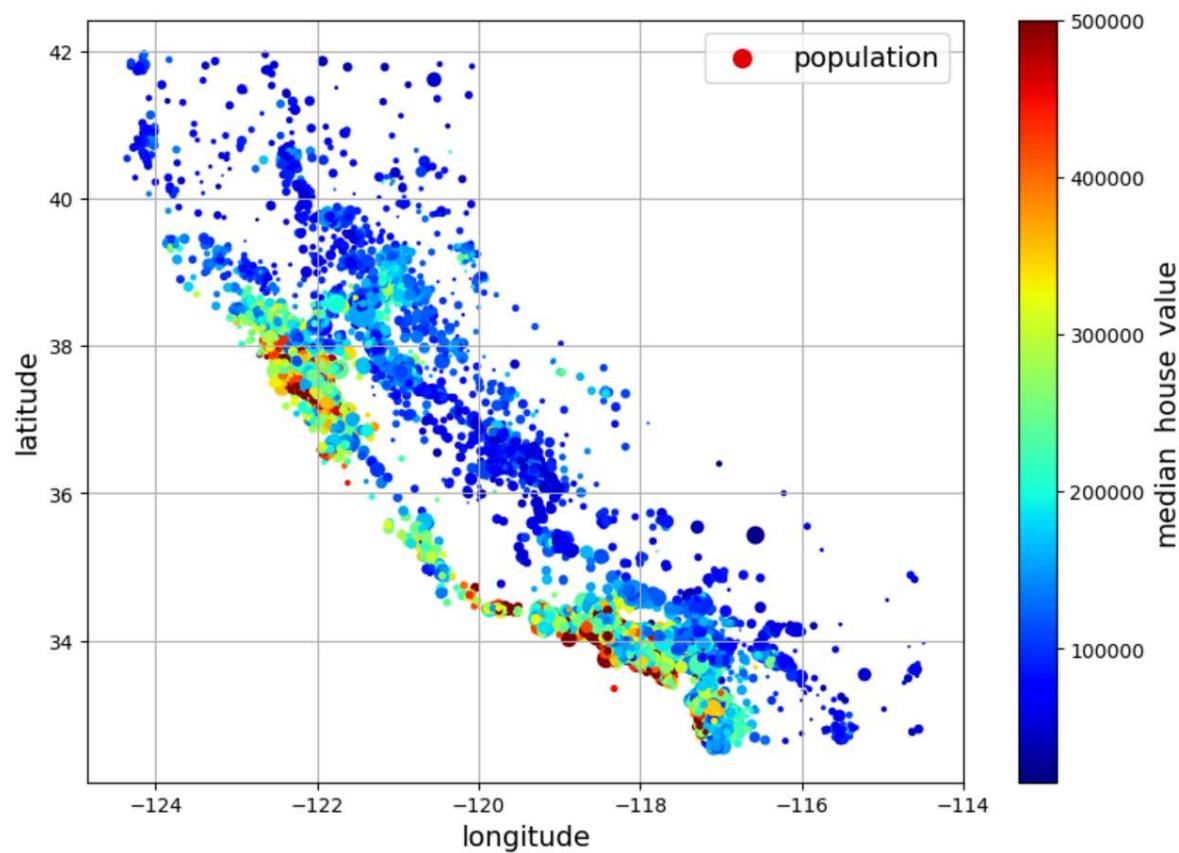
- 경도와 위도 정보를 이용하여 구역 정보를 산포도로 표현하여 인구밀도 확인 가능.
- 샌프란시스코의 Bay Area, LA, 샌디에고 등 유명 대도시의 특정 구역이 높은 인구 밀도를 가짐.

데이터프레임의 plot() 메서드

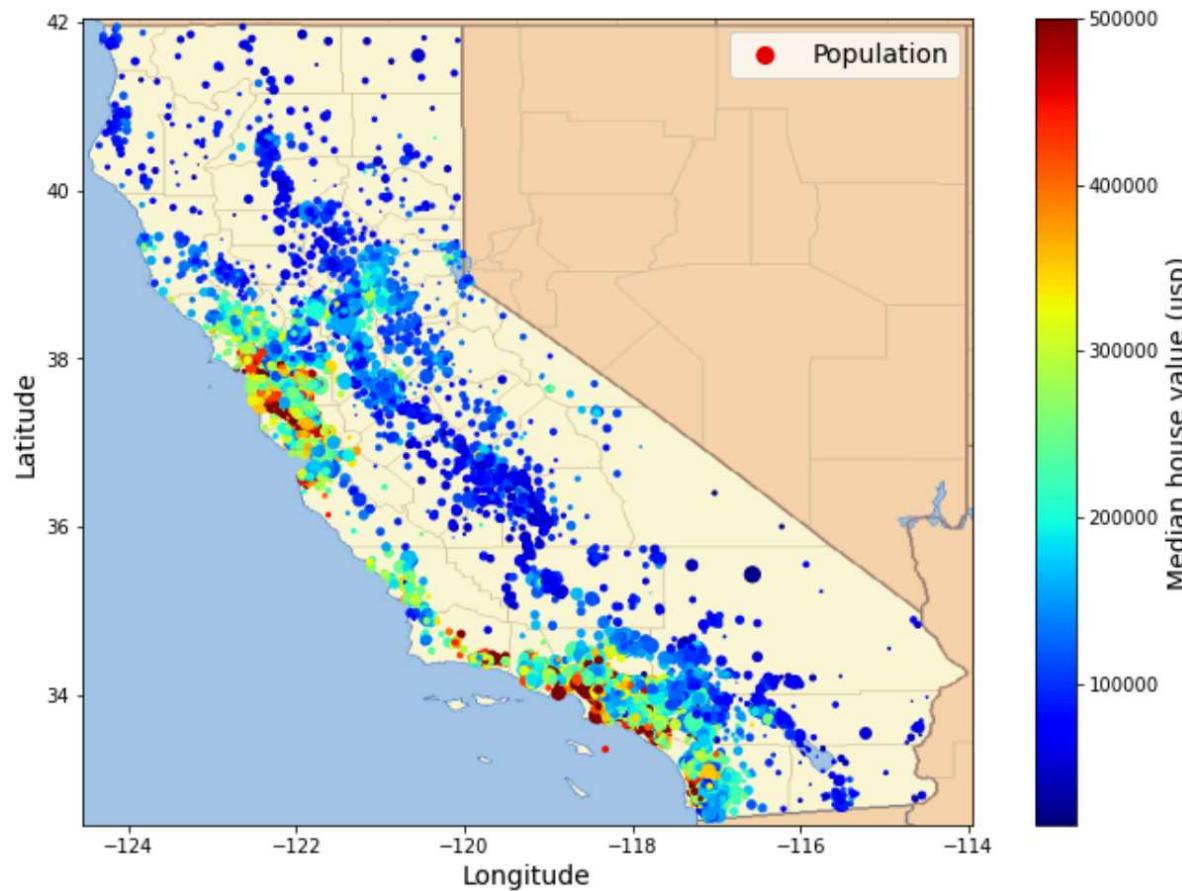
- 산점도 그리기

```
housing.plot(kind="scatter",
              x="longitude",
              y="latitude",
              grid=True,
              s=housing["population"] / 100,
              label="population",
              c="median_house_value",
              cmap="jet",
              colorbar=True,
              legend=True,
              figsize=(10, 7))
```

산점도



실제 지도 활용



상관관계

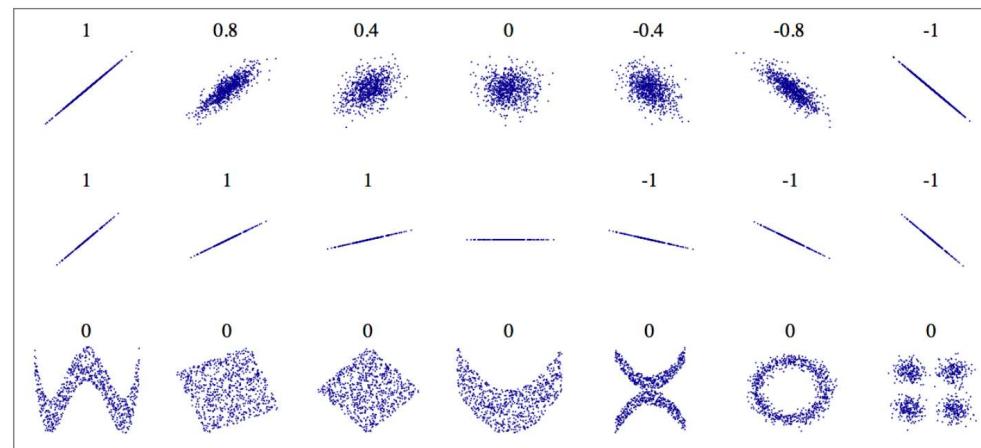
- 데이터프레임의 `corr()` 메서드는 수치형 특성들 사이의 선형 상관계수를 계산

```
corr_matrix = housing.corr()
```

- 주택 중위가격 특성과 다른 특성들 사이의 상관계수

```
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
```

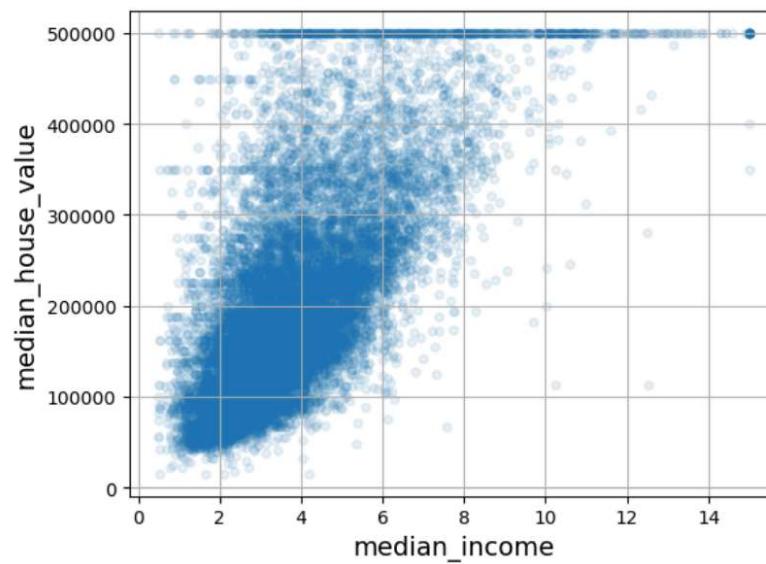
상관계수와 상관관계



- 상관계수가 1에 가까울 수록: 강한 양의 선형 상관관계
- 상관계수가 -1에 가까울 수록: 강한 음의 선형 상관관계
- 상관계수가 0에 가까울 수록: 매우 약한 선형 상관관계

주택 중위가격과 중위소득의 상관관계

- 상관계수가 0.68로 중위소득에 따라 주택 중위가격 함께 변하는 경향 감함.
- 점들이 너무 넓게 퍼져 있음. 완벽한 선형관계와 거리 멎.
- 50만 달러 수평선: 가격 제한 결과로 보임
- 이상한 형태를 학습하지 않도록 해당 구역을 제거하는 것이 좋음. (여기서는 그대로 두고 사용)



특성 조합 활용

- 구역별 총 방 수와 총 침실 수 대신 아래 특성이 보다 유용함
 - 가구당 방 개수(rooms_per_household)
 - 방 하나당 침실 개수(bedrooms_per_room)
 - 가구당 인원(population_per_household)
- 주택 중위가격과 방 하나당 침실 개수의 상관관계가 중위소득을 제외한 기존의 다른 특성들에 비해 높음.