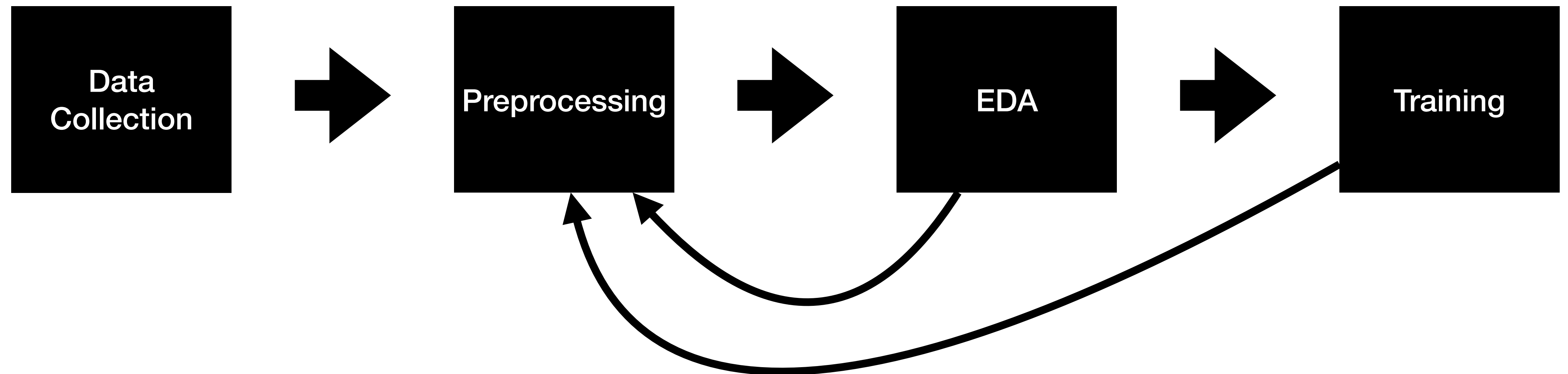


# **Big Data Analytics Programming**

**Week-06. Exploratory Data Analysis**

**Jungwon Seo, 2020-Fall**



# Exploratory Data Analysis

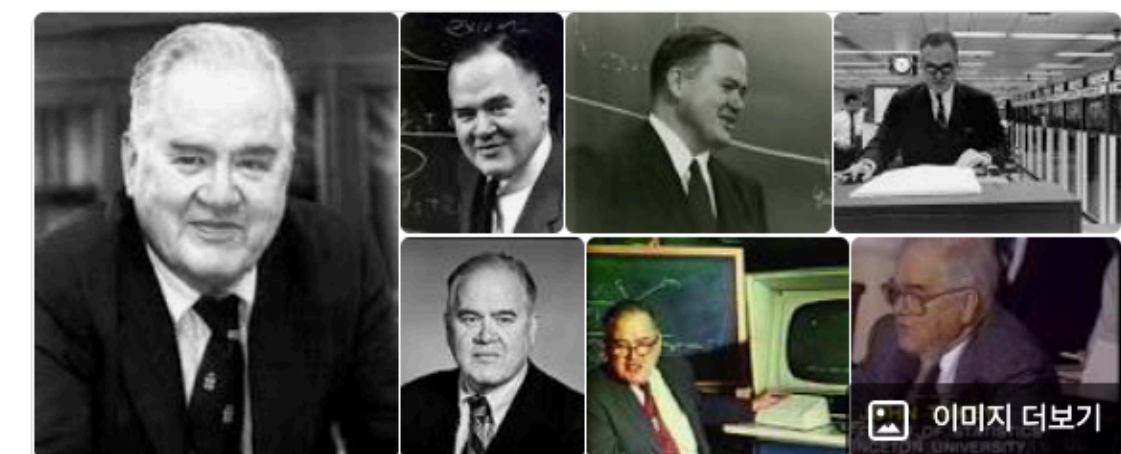
## EDA의 정의와 목적

- 탐색적 데이터 분석 (EDA)

- John Tukey가 창안한 자료분석 방법론
- 기존의 통계학이 정보의 추출에서 가설 검정등에만 너무 치우쳐져 자료 자체의 의미를 찾는데 어려움이 존재
  - 모델에만 집중을하는 머신러닝 프로세스도 마찬가지
- 시각적/수치적 요약 데이터로 데이터를 다양한 각도로 분석

- Goal\*

- Maximize insight into a data set;
- Uncover underlying structure;
- Extract important variables;
- Detect outliers and anomalies;
- Test underlying assumptions;
- Develop parsimonious models; and
- Determine optimal factor settings.



### 존 튜키

미국 수학자

영어에서 번역됨 - 존 와일더 터키 (John Wilder Tukey)는 Fast Fourier Transform 알고리즘과 박스 플롯 개발로 가장 잘 알려진 미국의 수학자입니다.  
[위키백과\(영어\)](#)

[원래 설명 보기](#) ▼

출생: 1915년 6월 16일, [미국 매사추세츠 뉴베드퍼드](#)

사망 정보: 2000년 7월 26일, [미국 뉴저지 뉴브런즈윅](#)

국적: 미국

\* <https://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm>

# Exploratory Data Analysis

## EDA의 정의와 목적

- How
  - EDA 접근 방식은 특정한 기술 집합이라기 보다, 데이터 분석에 대한 자세나 철학을 의미
  - 주로 시각적 방법론에 의존
    - 차트, 테이블
    - 사람은 시각적 정보를 해석하는 능력이 발달되어 있음
- Machine Learning without EDA?
  - 무의미한 속성에 대한 인코딩과 정규화
  - 결측값/이상치에 의한 예상 외의 결과
  - 불균형 데이터에 의한 모델 쓸림 현상
  - 결국 => 무의미한 모델 개선으로 이어짐

# Exploratory Data Analysis Techniques

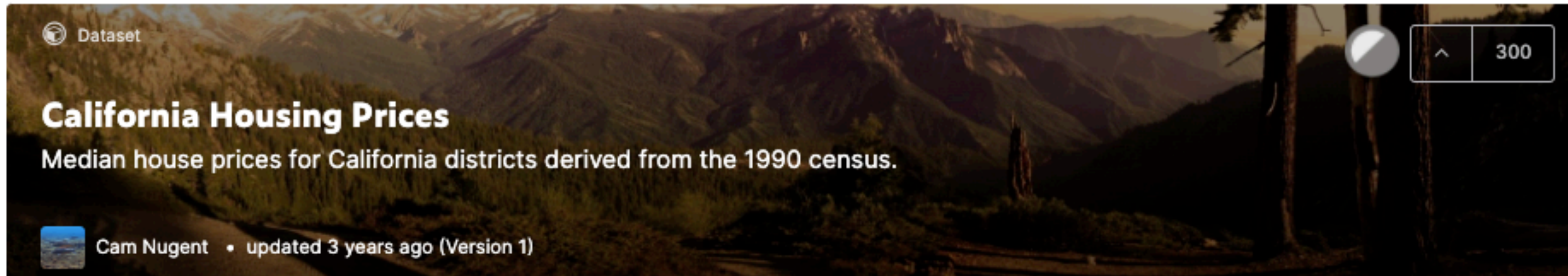
- 요약통계
  - Categorical 데이터
    - Frequency: 빈도수
    - Mode: 최빈값
  - Numerical 데이터
    - Mean, Median, Min, Max ..
- 시각화
  - Bar, Box, Pie, Scatter ..



# EDA 예제로 바로 알아보기




# 캘리포니아 주택 가격 예측하기





**California Housing Prices**  
Median house prices for California districts derived from the 1990 census.

Cam Nugent • updated 3 years ago (Version 1)

[Data](#) [Tasks \(1\)](#) [Notebooks \(163\)](#) [Discussion \(4\)](#) [Activity](#) [Metadata](#) [Download \(1 MB\)](#) [New Notebook](#)

 **Usability** 8.5

 **License** CC0: Public Domain

 **Tags** computer science, software, programming, social science, social issues and advocacy and 3 more

Description

### Context

This is the dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. It serves as an excellent introduction to implementing machine learning algorithms because it requires rudimentary data cleaning, has an easily understandable list of variables and sits at an optimal size between being too toyish and too cumbersome.

The data contains information from the 1990 California census. So although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.

# EDA 예제

## 데이터 구조 훑어보기

- Feature 확인

- longitude/latitude: 경도/위도
- housing\_median\_age: 주택 연식 중위값
- total\_rooms: 지역의 전체 방의 수
- total\_bedrooms: 지역의 전체 침실의 개수
- population: 지역의 인구
- households: 가구 수
- median\_income: 중위소득
- median\_house\_value: 주택가격의 중위값
- ocean\_proximity: 해안가와의 가까움 정도

(주의) 각 행(row)은 특정 집이 아닌 지역

|   | 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        |



# EDA 예제

## 데이터 구조 훑어보기

```
housing.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20640 entries, 0 to 20639
Data columns (total 10 columns):
longitude      20640 non-null float64
latitude       20640 non-null float64
housing_median_age  20640 non-null float64
total_rooms    20640 non-null float64
total_bedrooms 20433 non-null float64
population     20640 non-null float64
households     20640 non-null float64
median_income  20640 non-null float64
median_house_value 20640 non-null float64
ocean_proximity 20640 non-null object
dtypes: float64(9), object(1)
memory usage: 1.6+ MB
```

### 데이터 수

### 범주형 데이터에 대한 요약통계

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

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

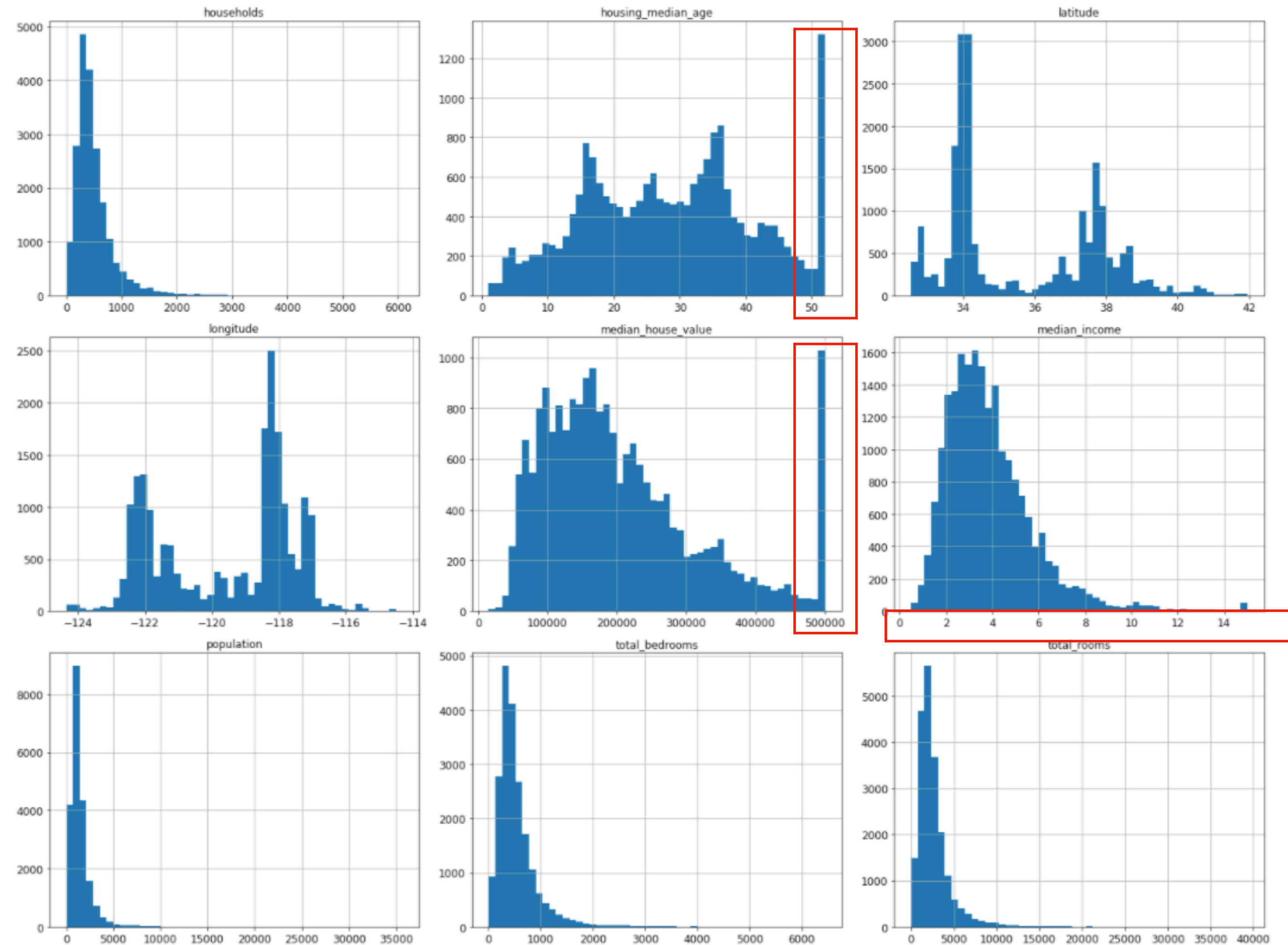
### 수치형 데이터에 대한 요약통계

| 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      |

# EDA 예제

## 데이터 구조 훑어보기

각 속성별 히스토그램



# EDA 예제

## 데이터 구조 훑어보기

- (주의) 데이터를 더 깊게 보기전에 미리 테스트 데이터셋을 분리해야함
- 중위소득의 단위 확인
- 침실 수의 결측값 인지
- 주택 연식과 가격의 최대값이 한정되어 있는데, 이 최대값을 넘어가는 예측이 필요한지 확인!
  - 데이터셋의 주택가격의 최대값이 \$500,000인데 실제 주택 가격이 몇 백만 달러일수도 있음
  - 만약 필요하다면, 실제 값을 확보
  - 확보 할 수 없다면 이 구간을 제거

# EDA 예제

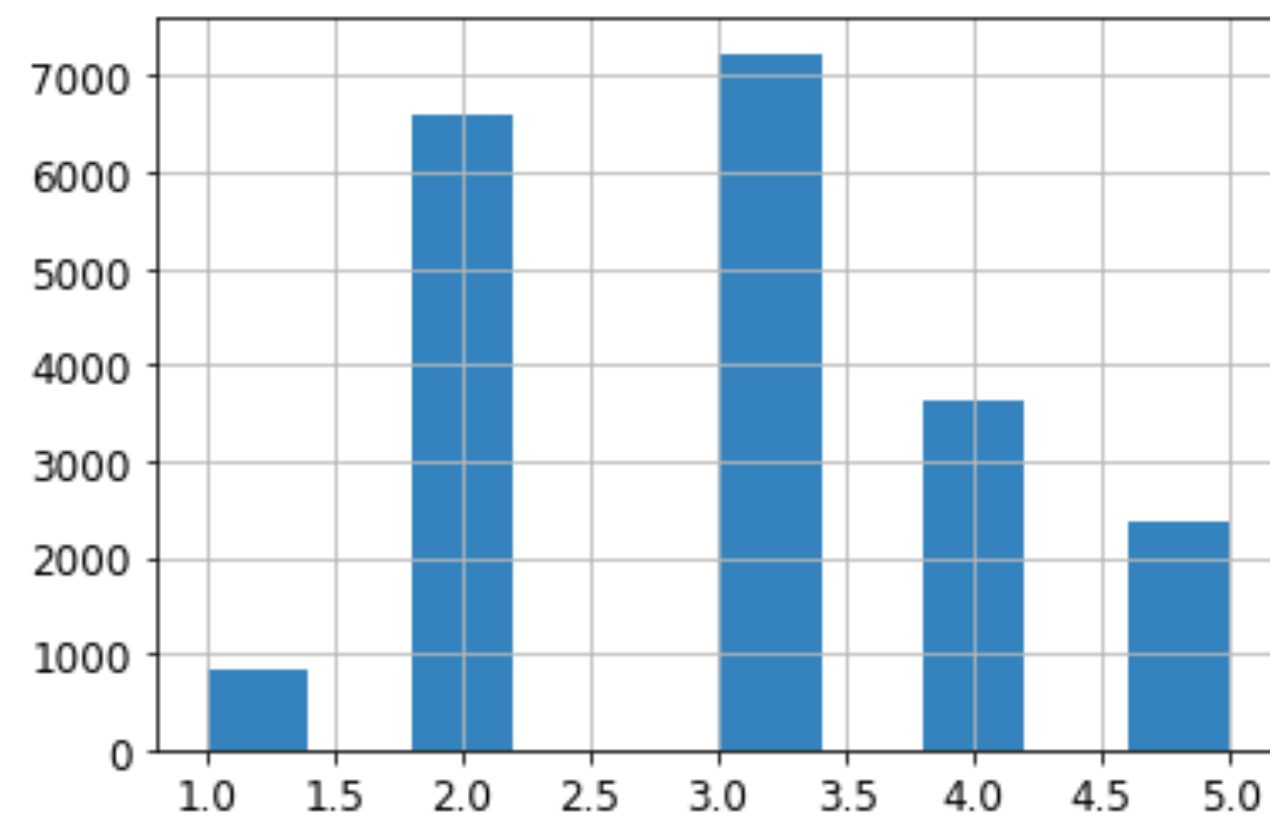
## 테스트 데이터 세트 분리하기

- 왜 테스트 세트를 만들어야 하나?
  - 일반화된 모델을 만들기 위해서는 최종 모델을 검증하기 전까지 사람도 모델도 테스트 세트에 대한 영향을 받으면 안 됨
  - 자칫 잘못하면 테스트 세트에 최적화된 모델을 개발 할 수가 있음
- 전통 적인 머신러닝에서는 일반적으로 전체 데이터 셋의 20%를 테스트 세트로 가져감
  - 물론 데이터가 1억개 이렇게 있다면, 2천만개를 굳이 테스트 세트로 가져가지 않아도 충분함
- 어떻게 20%를 선택할까?
  - 서울 부동산 가격을 예로 들때 강남구만 테스트 세트로 뽑는게 서울 부동산 가격을 대표한다고 볼 수 있을까?

# EDA 예제

## 테스트 데이터 세트 분리하기

- 단순 랜덤하게 추출을 한다면, 전체 표본을 대표하는 샘플을 얻는게 보장되지 않음 계층적 샘플링을 활용하여 표본 추출
- 도메인 지식을 활용하여, 표본 추출에 사용될 기준을 책정
  - 부동산 전문가에 따르면 해당 지역의 부동산가격은 해당 거주인의 소득과 크게 연관이 있다고함



소득분위(5분위)의 히스토그램

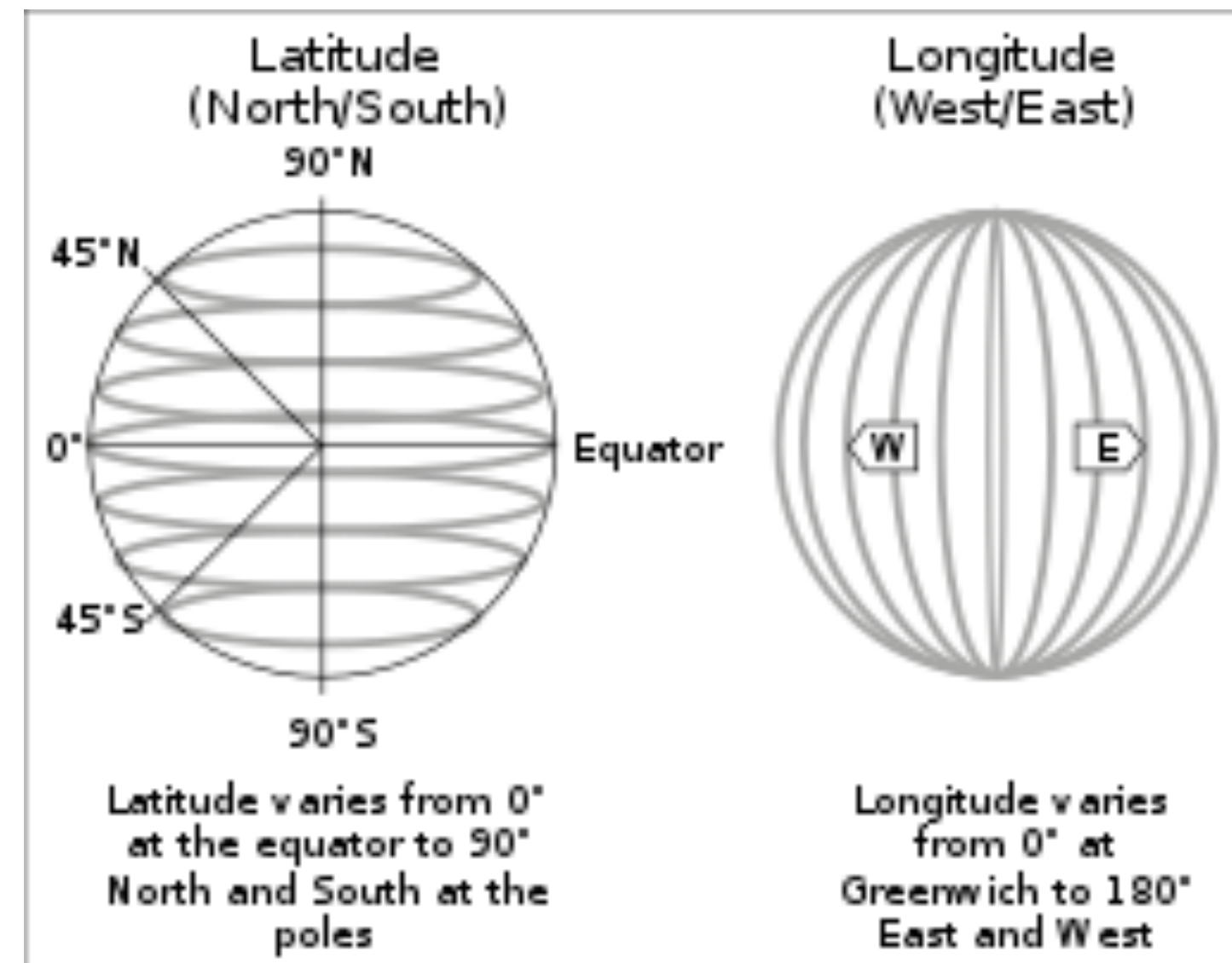
|   | Overall  | Stratified | Random   | Rand. %error | Strat. %error |
|---|----------|------------|----------|--------------|---------------|
| 1 | 0.039826 | 0.039729   | 0.040213 | 0.973236     | -0.243309     |
| 2 | 0.318847 | 0.318798   | 0.324370 | 1.732260     | -0.015195     |
| 3 | 0.350581 | 0.350533   | 0.358527 | 2.266446     | -0.013820     |
| 4 | 0.176308 | 0.176357   | 0.167393 | -5.056334    | 0.027480      |
| 5 | 0.114438 | 0.114583   | 0.109496 | -4.318374    | 0.127011      |

랜덤 샘플링 vs 계층 샘플링

# EDA 예제

## 지리 정보를 활용한 시각화

- 경도(longitude), 위도(latitude)란?
  - 예) <https://www.google.co.kr/maps/place/Yonsei+University/@37.5657882,126.9363833,17z>
- 지리 좌표계

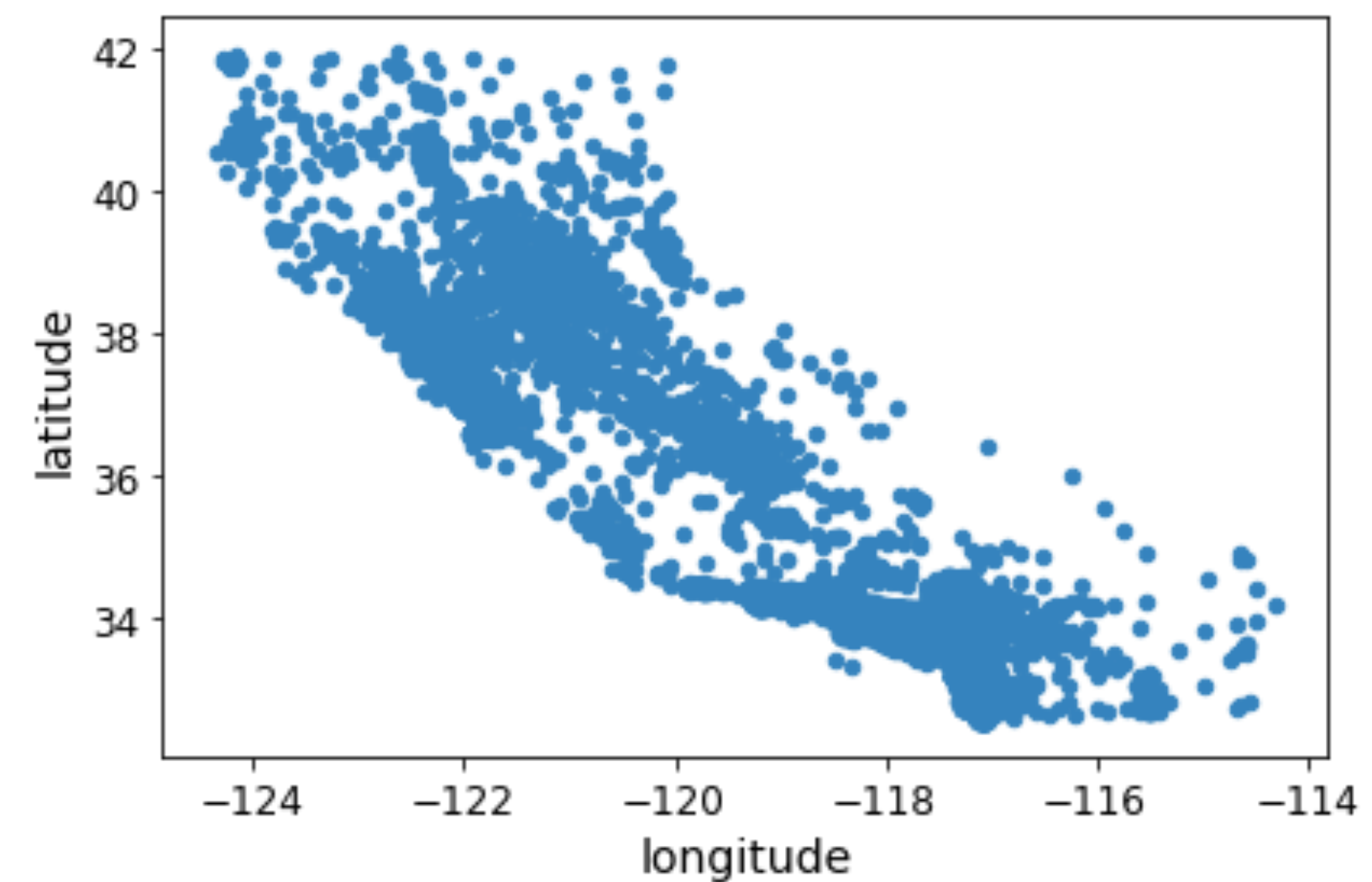




# EDA 예제

## 지리 정보를 활용한 시각화

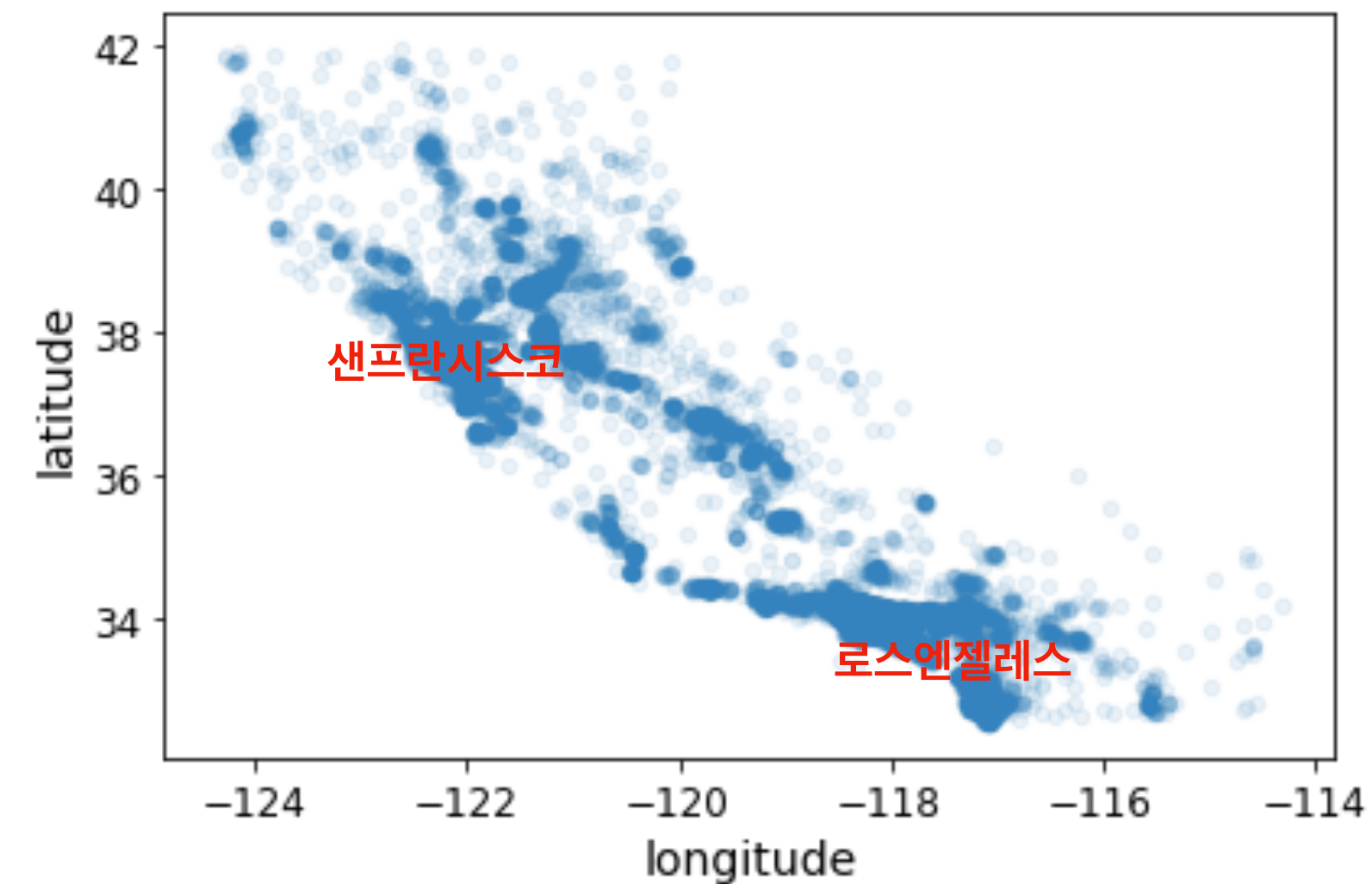
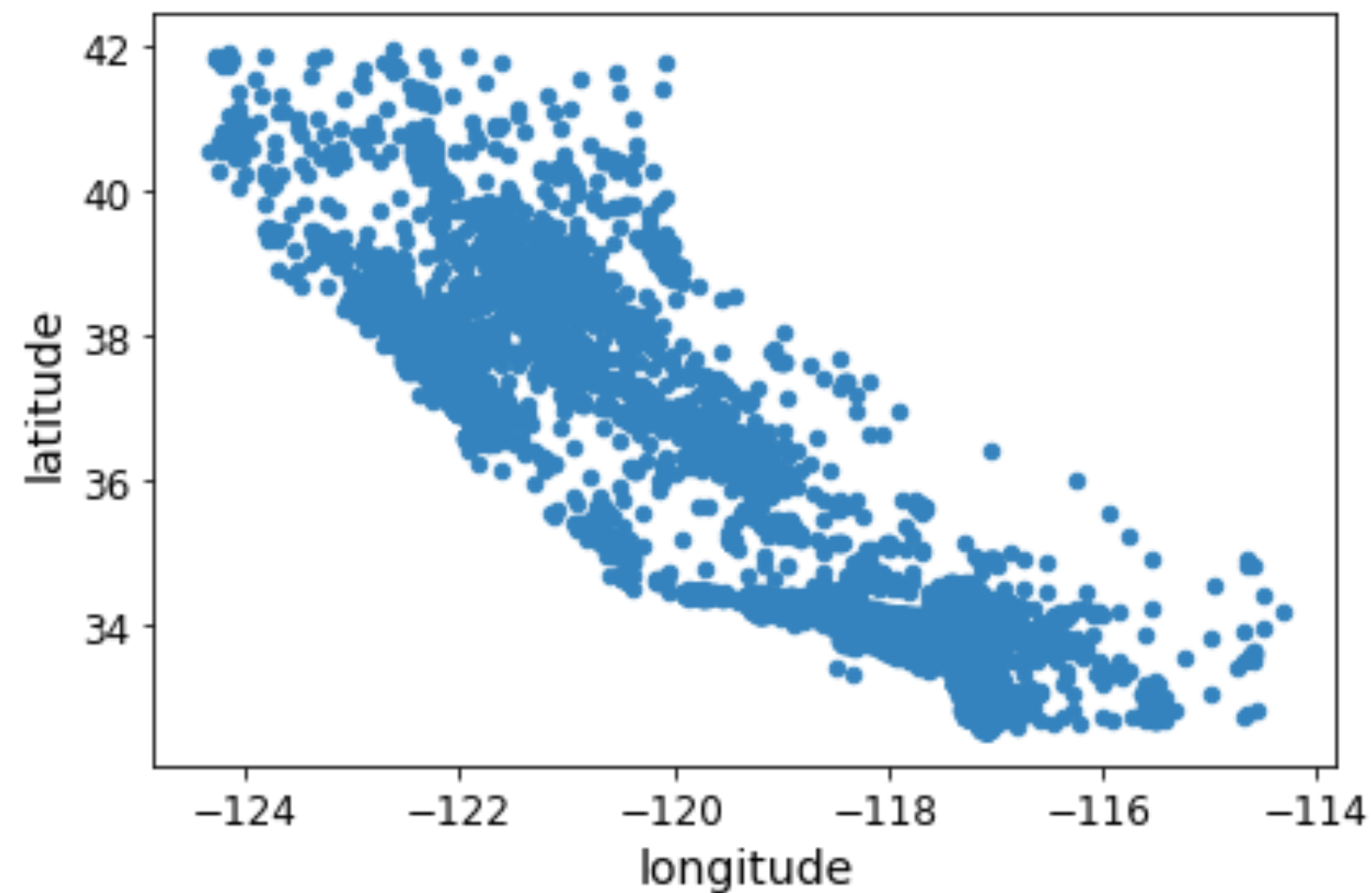
- 데이터셋의 경도/위도를 이용하여 Scatter plot 출력
- 오른쪽 차트는 캘리포니아 지역을 잘 나타내지만, 특별한 인사이트를 얻을 수 없음



# EDA 예제

## 지리 정보를 활용한 시각화

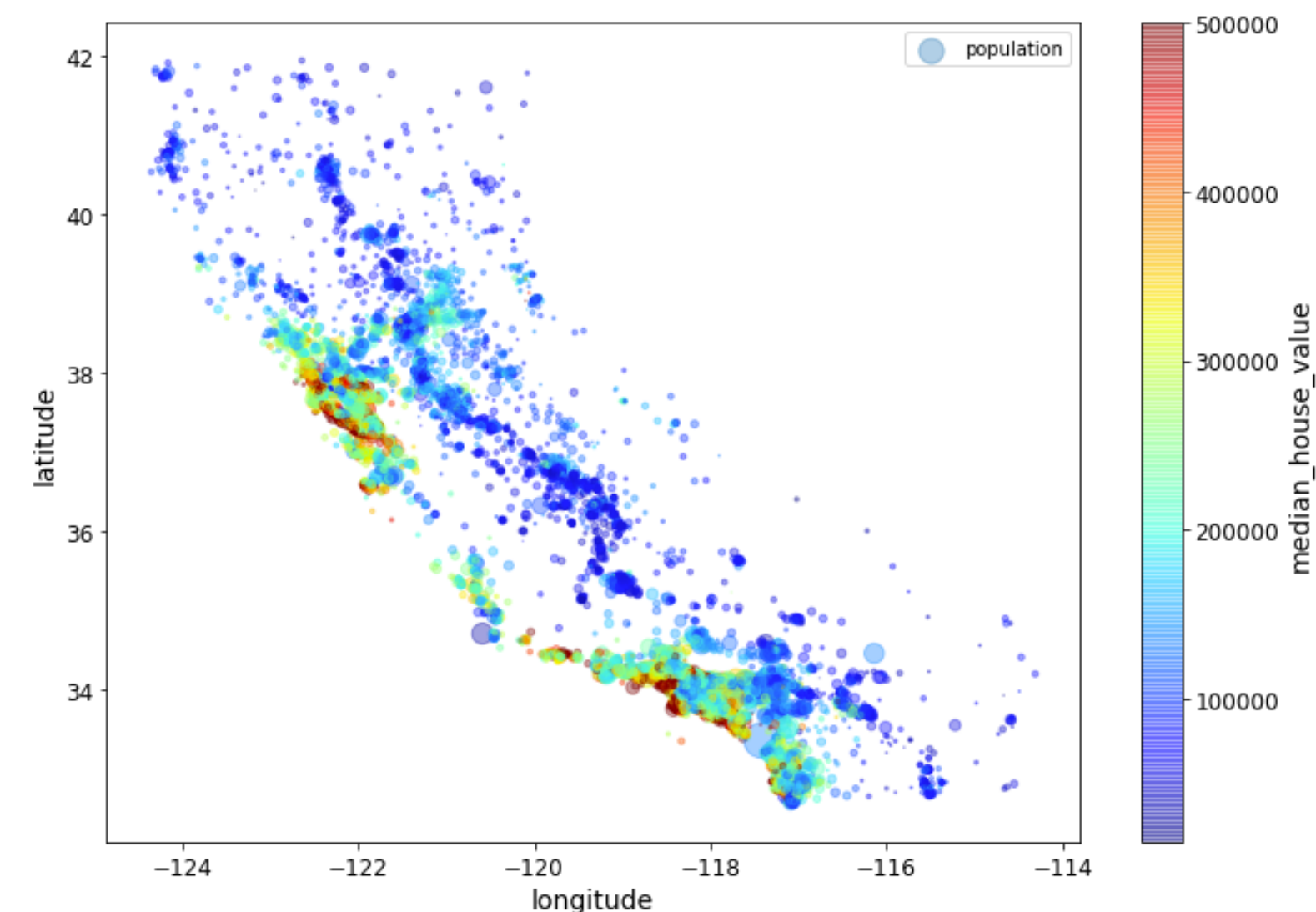
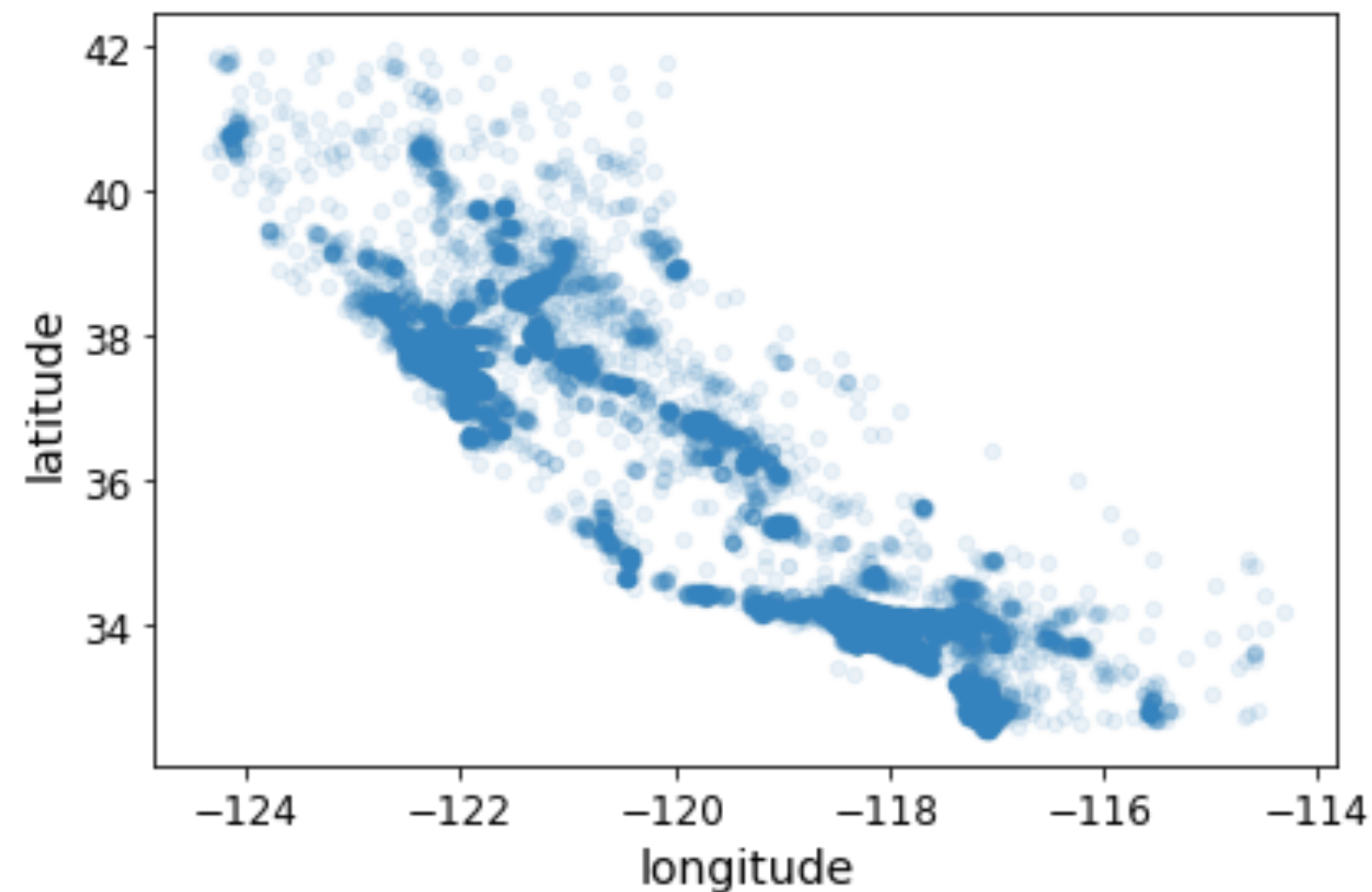
- 각 데이터셋의 투명도를 조절하여 출력
- 데이터셋에서 밀집된 지역을 찾을 수 있음
  - 연한 부분은 사실 주 거주 지역이 아님 (교외)
  - 또 해안가를 따라 주로 밀집한다는 것을 볼 수 있음



# EDA 예제

## 지리 정보를 활용한 시각화

- 각 포인트에 색과 크기를 넣어 추가적인 속성을 표현함
  - 밀집된 지역의 중심가가 주로 높은 가격을 보임(붉은색)
  - 외각 지역은 낮은 가격을 보임 (파란색)
  - 인구수가 많은 지역은 큰 동그라미 반대는 작은 동그라미
- 주택가격은 인구밀도와 연관이 있음



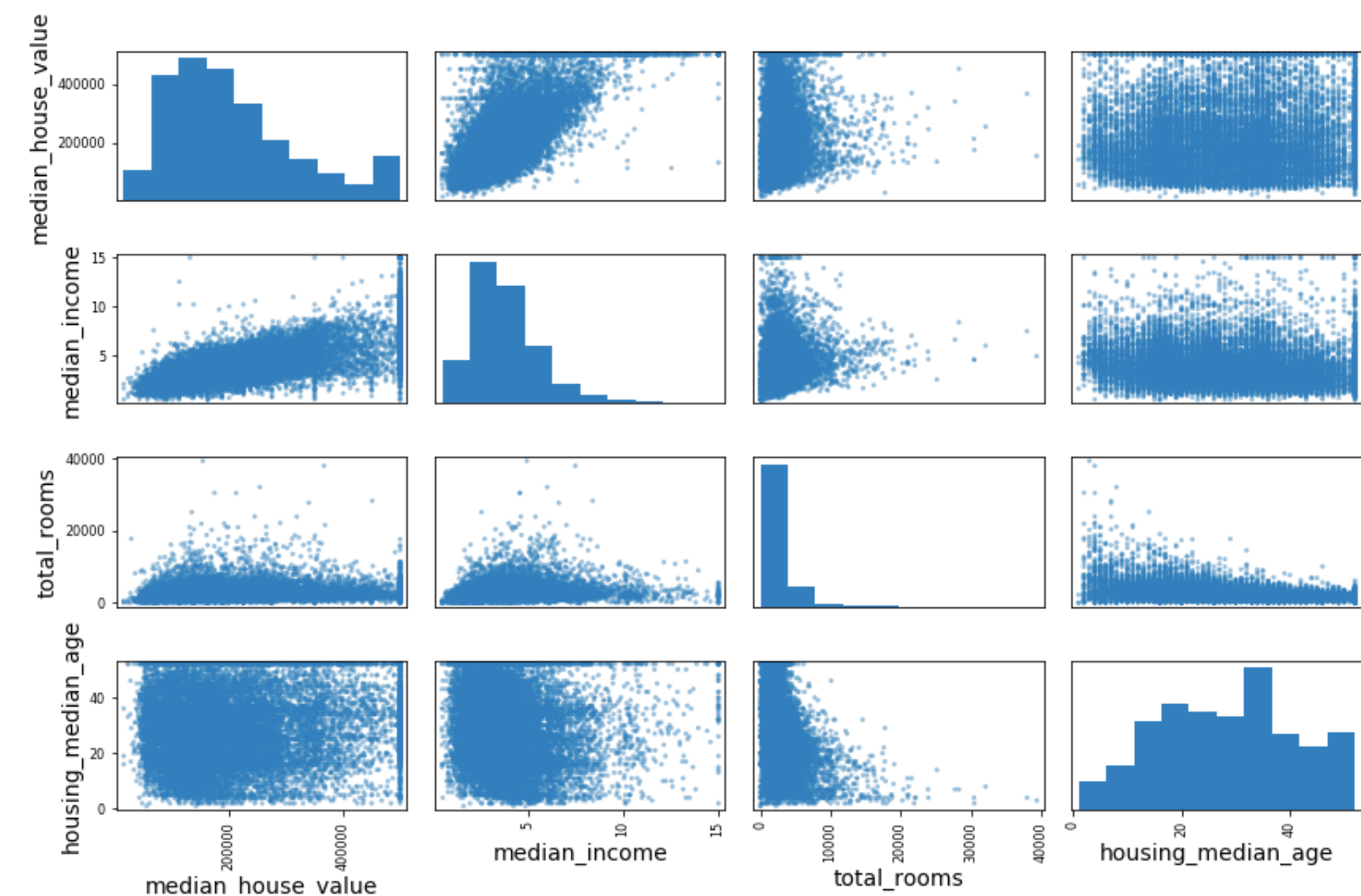


# EDA 예제

## 속성간 상관관계 분석

- 우리가 예측하고자 하는 주택 가격과 가장 상관관계가 큰 속성 확인
  - 상관관계는 -1부터 1사이의 값으로 절대값을 취했을 때 1에 가까울 수록 두 속성이 높은 선형적인 상관관계를 보인다고 할 수 있음
  - 1: 음의 상관관계, 1: 양의 상관관계, 0: 상관관계 없음

|                    |           |
|--------------------|-----------|
| median_house_value | 1.000000  |
| median_income      | 0.687160  |
| total_rooms        | 0.135097  |
| housing_median_age | 0.114110  |
| households         | 0.064506  |
| total_bedrooms     | 0.047689  |
| population         | -0.026920 |
| longitude          | -0.047432 |
| latitude           | -0.142724 |



# EDA 예제

## 속성간 상관관계 분석

- 속성들을 조합한 뒤에, 조합된 속성간의 상관관계 분석(feature 생성)
  - 예) 한 지역에 전체 방의 개수는 큰 의미가 없음, 중요한 것은 가구당 방 개수
- 조합의 예
  - rooms\_per\_household: 가구 당 방의 수 (total\_rooms/households)
  - bedrooms\_per\_room: 방 중 침실의 비율 (total\_bedrooms/total\_rooms)
  - population\_per\_household: 가구당 거주자 수 (population/households)

|                    |           |
|--------------------|-----------|
| median_house_value | 1.000000  |
| median_income      | 0.687160  |
| total_rooms        | 0.135097  |
| housing_median_age | 0.114110  |
| households         | 0.064506  |
| total_bedrooms     | 0.047689  |
| population         | -0.026920 |
| longitude          | -0.047432 |
| latitude           | -0.142724 |

|                          |           |
|--------------------------|-----------|
| median_house_value       | 1.000000  |
| median_income            | 0.687160  |
| rooms_per_household      | 0.146285  |
| total_rooms              | 0.135097  |
| housing_median_age       | 0.114110  |
| households               | 0.064506  |
| total_bedrooms           | 0.047689  |
| population_per_household | -0.021985 |
| population               | -0.026920 |
| longitude                | -0.047432 |
| latitude                 | -0.142724 |
| bedrooms_per_room        | -0.259984 |

# EDA를 통해 추가적인 인사이트를 얻은 후

- 추가적인 전처리
- 모델링

참고문헌: Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.



**E.O.D**