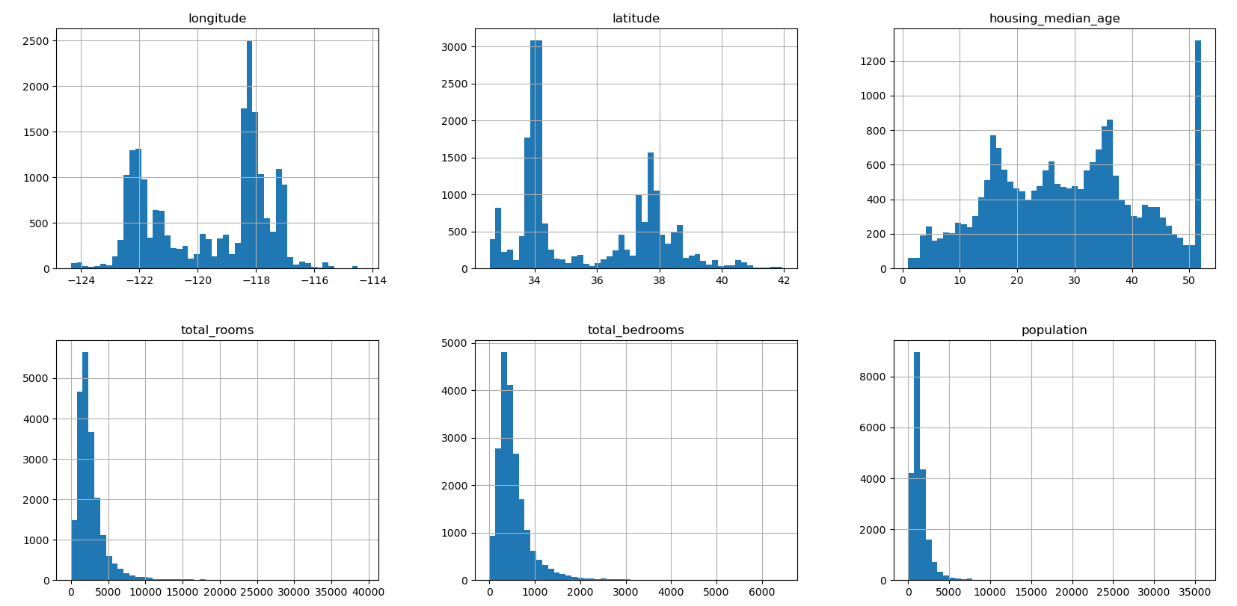
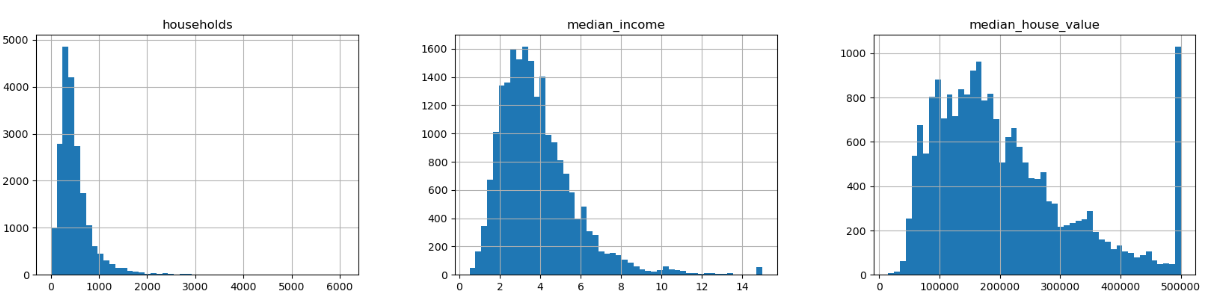
2.3.3 데이터 구조 훑터보기 : padas DataFrame





1. 중간소득: 0.5 ~ 15 사이로 표현이 됨 ( 의미 파악 필요: 3 => 약 3만 달러 의미 )
2. 중간주택연도와 중간주택가격은 최대값, 최소값이 한정이 된 듯. 이런 경우 선택방법 두가지
   1. 한계값 밖의 구역에 대한 정확한 레이블을 구한다.
   2. 훈련 셋트에서 이런 구역을 제거합니다.
3. 특성들의 스케일이 서로 많이 다릅니다.
4. 히스토그램의 꼬리가 두껍습니다. 가운데에서 왼쪽과 오른쪽 비교

2.3.4 테스트 세트 만들기

-. 테스트 세트로 일반화 오차 추정하면 매우 낙관적인 추정이 됨. 이를 데이터 스누핑(data snooping)편향이라고 함

-. 테스트 세트를 생성하는 것은 이론적으로는 간다. 무작위로 20% 샘플링

-. 데이터 셋이 바뀌어도 테스트 세트가 동일하기 유지할 수 있어야 한다. ???

-. 사이킷런은 데이터셋을 여러 서브셋으로 나누는 다양한 방법을 제공합니다. ( train\_test\_split )

* 데이터셋이 충분히 크다면 일반적으로 괜찮지만, 그렇지 않다면 샘플링 편향이 생길 가능성이 큽니다.
* 설문조사 기관에서 설문을 위해서 1000명에게 질문을 한다.
  + 전체 인구를 대변할 수 있는 1000명을 선택
  + 미국인구 비율: 여성 51.3%, 남성 48.7 %
  + 계층적 샘플링( stratified sampling ): 전체 인구를 계층이라는 동질의 그룹으로 나누고 테스트 세트가 전체 인구를 대표하도록 각 계층에서 올바른 수의 샘플을 추출
* 전문가가 중간 소득이 중간 주택가격을 예측하는 데 매우 중요하다고 이야기를 해주었다고 가정
  + 중간 소득이 연속적인 숫자형 특성이므로 소득에 대한 카테고리 특성을 만들어야 함

housing["income\_cat"] = pd.cut(housing["median\_income"],

bins=[0., 1.5, 3.0, 4.5, 6., np.inf],

labels=[1, 2, 3, 4, 5])

* + 소득 카테고리 기반으로 계층 샘플링을 할 준비가 됨.

사이킷 런의 StratifiedShuffleSplit를 사용

from sklearn.model\_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n\_splits=1, test\_size=0.2, random\_state=42)

for train\_index, test\_index in split.split(housing, housing["income\_cat"]):

strat\_train\_set = housing.loc[train\_index]

strat\_test\_set = housing.loc[test\_index]