15장 선형모델 연습

Sangkon Han(sangkon@pusan.ac.kr)

2023-03-29

캘리포니아 집 값 예측

데이터 불러오기

```
housing = read.csv("./data/housing.csv")
head(housing)
```

```
##   longitude latitude housing_median_age total_rooms total_bedrooms population
## 1   -122.23    37.88                41         880            129        322
## 2   -122.22    37.86                21        7099           1106       2401
## 3   -122.24    37.85                52        1467            190        496
## 4   -122.25    37.85                52        1274            235        558
## 5   -122.25    37.85                52        1627            280        565
## 6   -122.25    37.85                52         919            213        413
##   households median_income median_house_value ocean_proximity
## 1        126        8.3252             452600        NEAR BAY
## 2       1138        8.3014             358500        NEAR BAY
## 3        177        7.2574             352100        NEAR BAY
## 4        219        5.6431             341300        NEAR BAY
## 5        259        3.8462             342200        NEAR BAY
## 6        193        4.0368             269700        NEAR BAY
```

캘리포니아 집 값 예측 데이터 구조

- logitude, 경도
- latitude, 위도
- housing_median_age, 주변의 집을 그룹화 했기 때문에 중앙값 사용
- total_rooms, 정체 방 수
- total_bedrooms, 전체 침실 수
- population, 인구
- households, 세대수
- median_income, 소득(중앙값)
- median_house_value, 주택 가격(중앙값)
- ocean_proximity, 해안 근접도

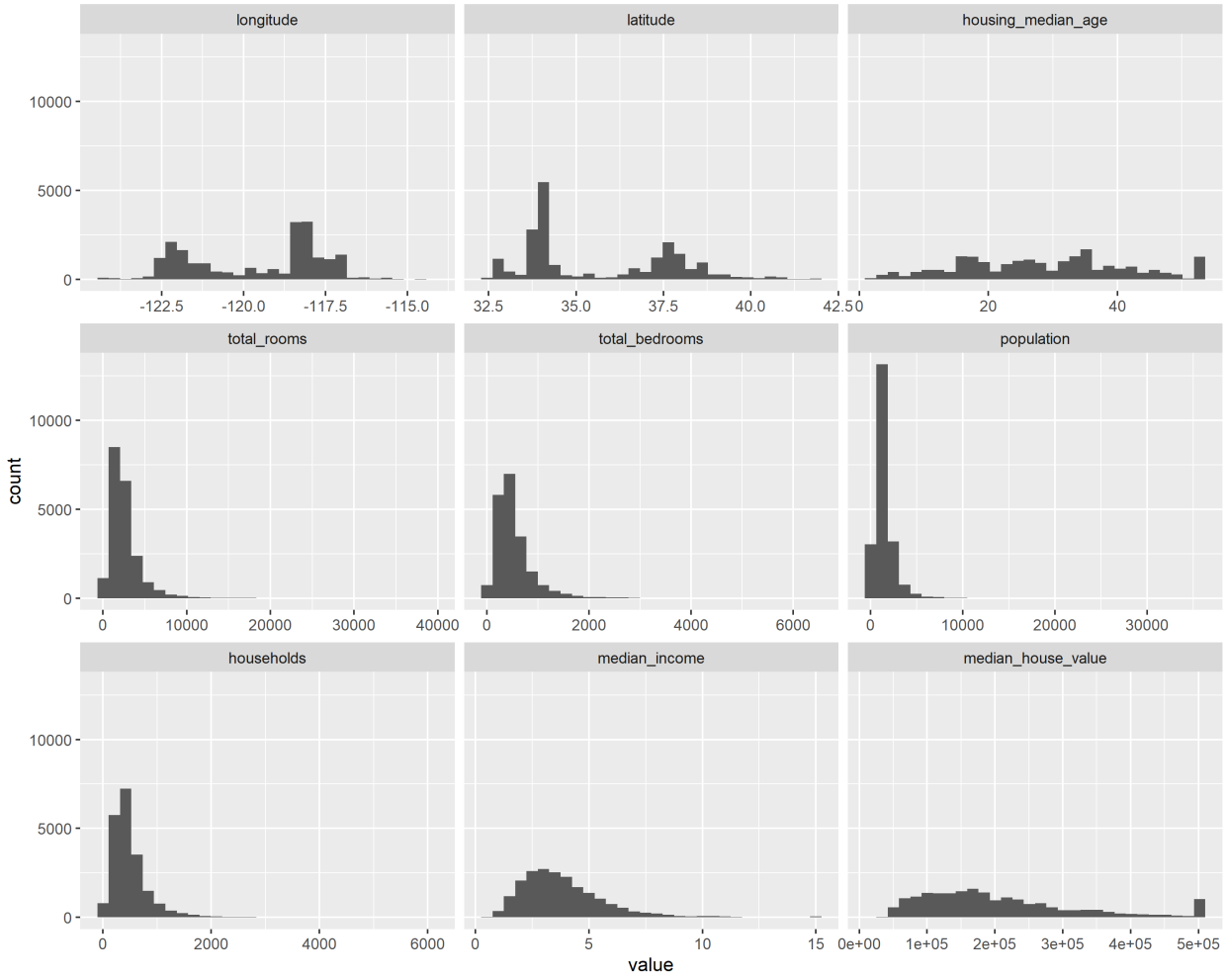
변수 요약 정보 확인

total_bedrooms에 NA값이 있음을 확인할 수 있습니다. 기술 통계 분석을 사용해서 출력된 정보를 토대로 (예를 들어 median_income과 median_house_value 등) 몇가지 가설을 세울 수 있습니다. ocean_proximity는 별도로 처리할 필요가 있습니다.

```
summary(housing)
```

```
##    longitude         latitude     housing_median_age  total_rooms
##  Min.   :-124.3   Min.   :32.54   Min.   : 1.00      Min.   :    2
##  1st Qu.:-121.8   1st Qu.:33.93   1st Qu.:18.00      1st Qu.: 1448
##  Median :-118.5   Median :34.26   Median :29.00      Median : 2127
##  Mean   :-119.6   Mean   :35.63   Mean   :28.64      Mean   : 2636
##  3rd Qu.:-118.0   3rd Qu.:37.71   3rd Qu.:37.00      3rd Qu.: 3148
##  Max.   :-114.3   Max.   :41.95   Max.   :52.00      Max.   :39320
##
##  total_bedrooms     population      households     median_income
##  Min.   :   1.0   Min.   :    3   Min.   :   1.0   Min.   : 0.4999
##  1st Qu.: 296.0   1st Qu.:  787   1st Qu.: 280.0   1st Qu.: 2.5634
##  Median : 435.0   Median : 1166   Median : 409.0   Median : 3.5348
##  Mean   : 537.9   Mean   : 1425   Mean   : 499.5   Mean   : 3.8707
##  3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0   3rd Qu.: 4.7432
##  Max.   :6445.0   Max.   :35682   Max.   :6082.0   Max.   :15.0001
##  NA's   :207
##  median_house_value ocean_proximity
##  Min.   : 14999     Length:20640
##  1st Qu.:119600     Class :character
##  Median :179700     Mode  :character
##  Mean   :206856
##  3rd Qu.:264725
##  Max.   :500001
##
```

```
ggplot(data = melt(housing), mapping = aes(x = value)) +
  geom_histogram(bins = 30) +
  facet_wrap(~variable, scales = 'free_x')
```



전처리

```
# 결측치 처리
housing$total_bedrooms[is.na(housing$total_bedrooms)] = median(housing$total_bedrooms , na.rm = TRUE)

# 파생 변수
housing$mean_bedrooms = housing$total_bedrooms/housing$households
housing$mean_rooms = housing$total_rooms/housing$households

# 기존 미사용 변수 삭제
drops = c('total_bedrooms', 'total_rooms')
housing = housing[ , !(names(housing) %in% drops)]
head(housing)
```

```
##   longitude latitude housing_median_age population households median_income
## 1   -122.23    37.88                41        322        126        8.3252
## 2   -122.22    37.86                21       2401       1138        8.3014
## 3   -122.24    37.85                52        496        177        7.2574
## 4   -122.25    37.85                52        558        219        5.6431
```

```
## 5   -122.25    37.85                52        565        259        3.8462
## 6   -122.25    37.85                52        413        193        4.0368
##   median_house_value ocean_proximity mean_bedrooms mean_rooms
## 1             452600        NEAR BAY     1.0238095   6.984127
## 2             358500        NEAR BAY     0.9718805   6.238137
## 3             352100        NEAR BAY     1.0734463   8.288136
## 4             341300        NEAR BAY     1.0730594   5.817352
## 5             342200        NEAR BAY     1.0810811   6.281853
## 6             269700        NEAR BAY     1.1036269   4.761658
```