hw02.R

Zhenyok Nazedwox

Sun Dec 18 02:37:45 2016

```
# 01 data
data <- read.csv("../data/calif_penn_2011.csv")
nrow(data)
```

```
## [1] 11275
```

```
ncol(data)
```

```
## [1] 34
```

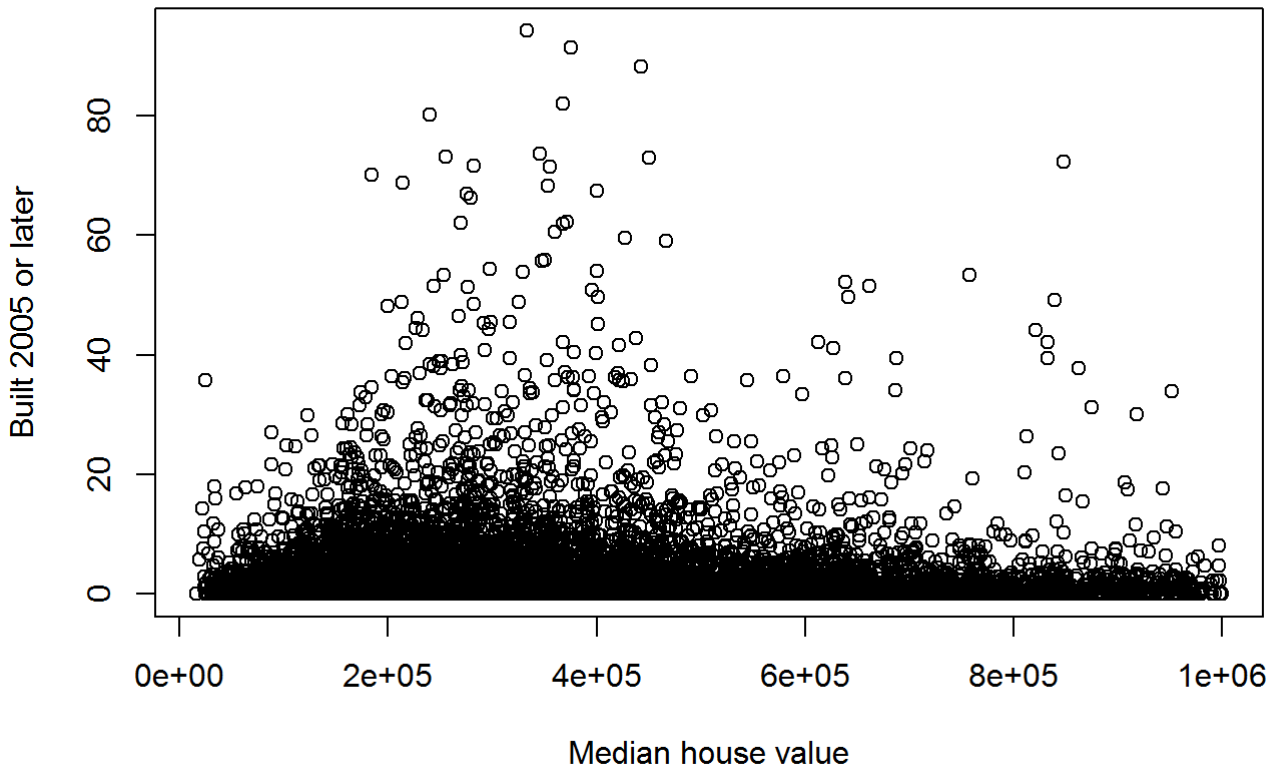
```
# количество NA по каждому столбцу для каждой строки
colSums(apply(data,c(1,2), is.na))
```

```
##                           X                      GEO.id2
##                           0                            0
##                     STATEFP                     COUNTYFP
##                           0                            0
##                     TRACTCE                   POPULATION
##                           0                            0
##                    LATITUDE                    LONGITUDE
##                           0                            0
##           GEO.display.label          Median_house_value
##                           0                          599
##                 Total_units                Vacant_units
##                           0                            0
##                Median_rooms  Mean_household_size_owners
##                         157                          215
## Mean_household_size_renters         Built_2005_or_later
##                         152                           98
##          Built_2000_to_2004                 Built_1990s
##                          98                           98
##                 Built_1980s                 Built_1970s
##                          98                           98
##                 Built_1960s                 Built_1950s
##                          98                           98
##                 Built_1940s       Built_1939_or_earlier
##                          98                           98
##                  Bedrooms_0                  Bedrooms_1
##                          98                           98
##                  Bedrooms_2                  Bedrooms_3
##                          98                           98
##                  Bedrooms_4          Bedrooms_5_or_more
##                          98                           98
##                      Owners                      Renters
##                         100                          100
##     Median_household_income       Mean_household_income
##                         115                          126
```

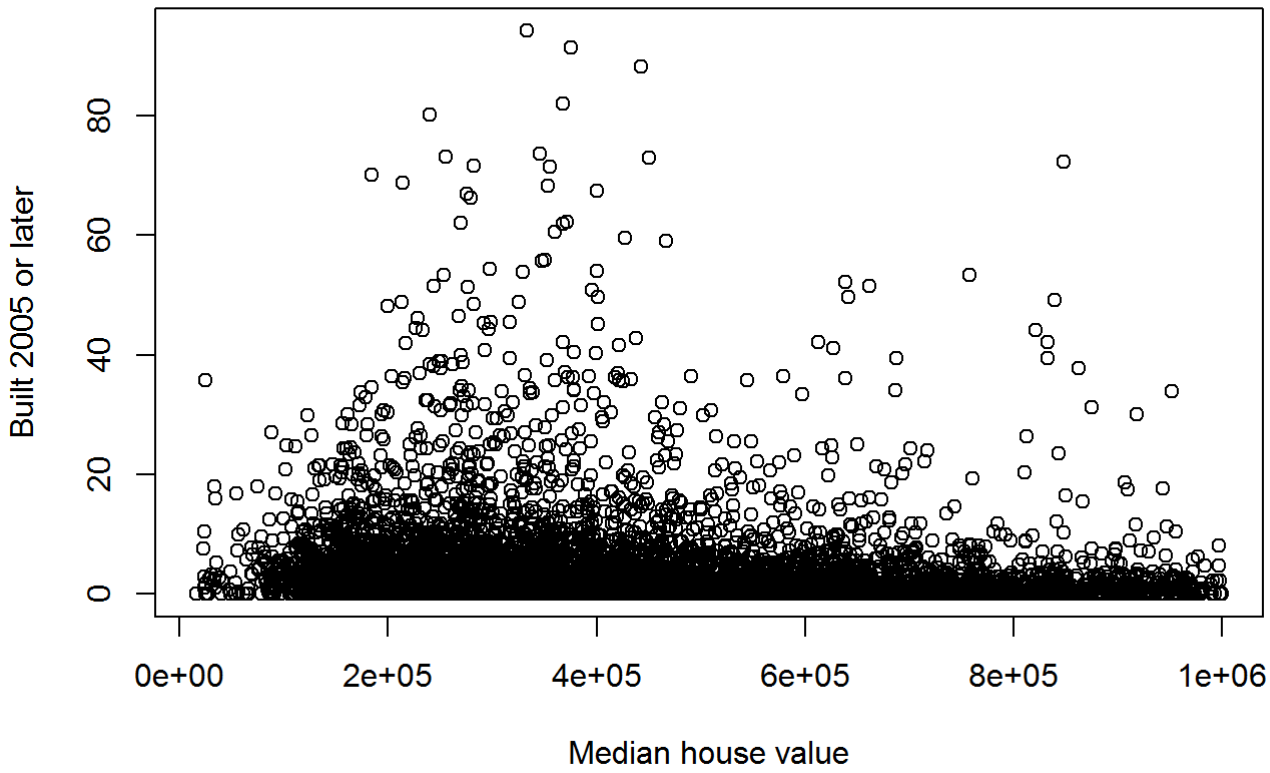
```
omdata <- na.omit(data)
# сколько строк было удалено
nrow(data) - nrow(omdata)
```

```
## [1] 670
```

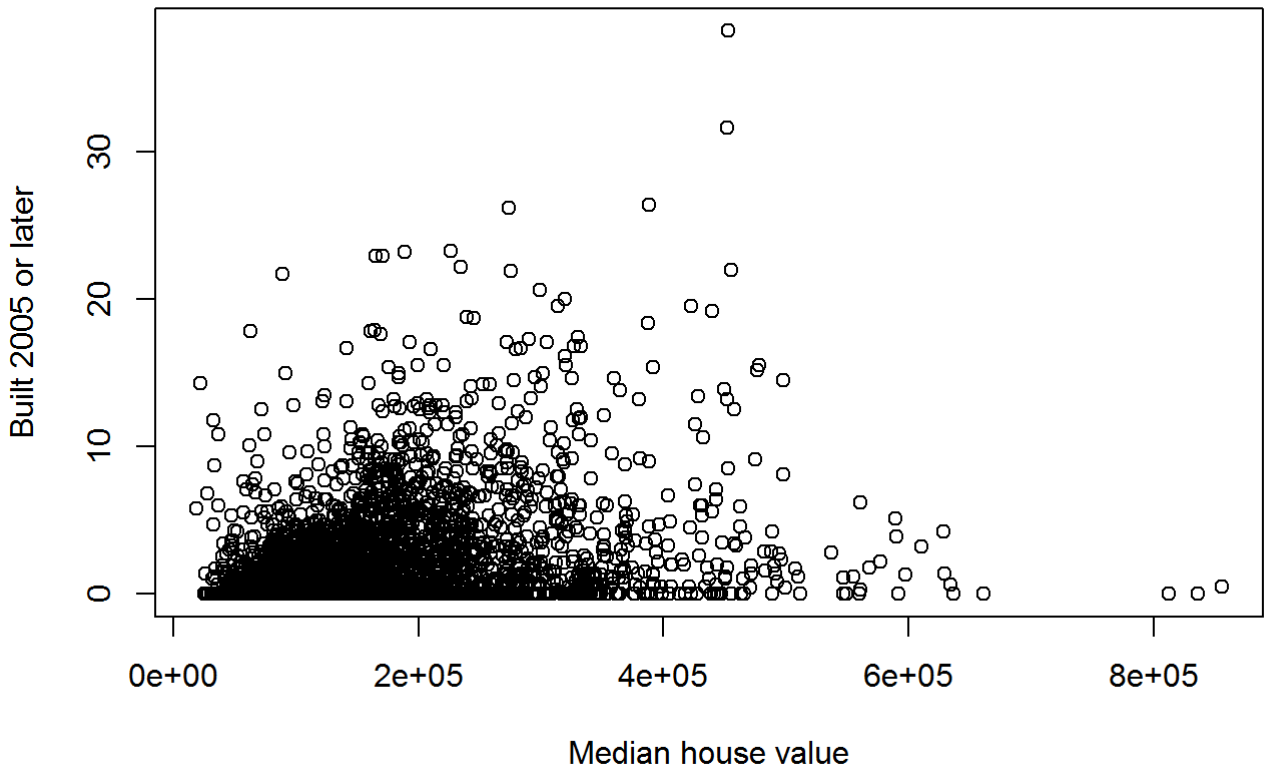
```
# 02 new houses
plot(omdata$Median_house_value,
     omdata$Built_2005_or_later,
     xlab = "Median house value", ylab = "Built 2005 or later")
```




```
# California 6
plot(omdata$Median_house_value[omdata$STATEFP == 6],
     omdata$Built_2005_or_later[omdata$STATEFP == 6],
     xlab = "Median house value", ylab = "Built 2005 or later")
```




```
# Pennsylvania 42
plot(omdata$Median_house_value[omdata$STATEFP == 42],
     omdata$Built_2005_or_later[omdata$STATEFP == 42],
     xlab = "Median house value", ylab = "Built 2005 or later")
```




```
# 03 vacancy
omdata$vacancy_rate = omdata$Vacant_units / omdata$Total_units
min(omdata$vacancy_rate)
```

```
## [1] 0
```

```
max(omdata$vacancy_rate)
```

```
## [1] 0.965311
```

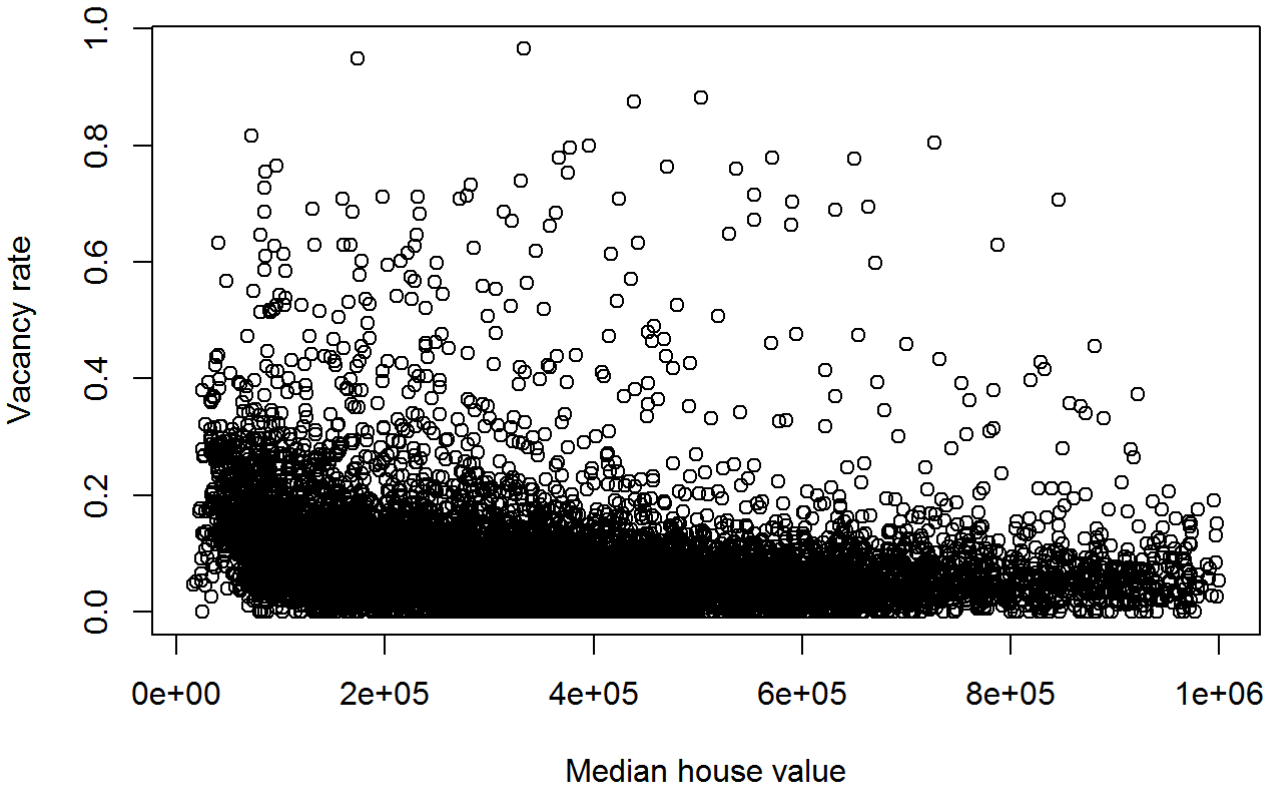
```
mean(omdata$vacancy_rate)
```

```
## [1] 0.08888789
```

```
median(omdata$vacancy_rate)
```

```
## [1] 0.06767283
```

```
plot(omdata$Median_house_value,
     omdata$vacancy_rate,
     xlab = "Median house value", ylab = "Vacancy rate")
```


Vacancy rate
0.0
0.2
0.4
0.6
0.8
1.0
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
Median house value

```
# чем больше средняя стоимость дома, тем их лучше покупают, так как доля незанятых до
мов меньше

# 04 correlation
# в acc записываются индексы для строк штата 6 и округа 1
# в accmv записываются значения Median house value для отобранных строк
# считается медиана по полученным значениям
acc <- c()
for (tract in 1:nrow(omdata)) {
  if (omdata$STATEFP[tract] == 6) {
    if (omdata$COUNTYFP[tract] == 1) {
      acc <- c(acc, tract)
    }
  }
}
accmv <- c()
for (tract in acc) {
  accmv <- c(accmv, omdata[tract,10])
}
fw = median(accmv)

# second way
sw = median(as.numeric(unlist(subset(omdata, STATEFP == 6 & COUNTYFP == 1, select = 1
0))))

# average percent of builds
# Butte County
bc = mean(as.numeric(unlist(subset(omdata, STATEFP == 6 & COUNTYFP == 7, select = c(1
6:24)))))
# Santa Clara
sc = mean(as.numeric(unlist(subset(omdata, STATEFP == 6 & COUNTYFP == 85, select =
c(16:24)))))
# York County
yc = mean(as.numeric(unlist(subset(omdata, STATEFP == 42 & COUNTYFP == 133, select =
c(16:24)))))

# cor function
# all dataset
cor(omdata[[10]], omdata[[16]])
```

```
## [1] -0.01893186
```

```
# California
cor(omdata[which(omdata$STATEFP == 6), 10], omdata[which(omdata$STATEFP == 6), 16])
```

```
## [1] -0.1153604
```

```
# Pennsylvania
cor(omdata[which(omdata$STATEFP == 42), 10], omdata[which(omdata$STATEFP == 42), 16])
```

```
## [1] 0.2681654
```

```
# Butte County
bc = omdata[which(omdata$STATEFP == 6 & omdata$COUNTYFP == 7),]
cor(bc[[10]], bc[[16]])
```

```
## [1] 0.04203267
```

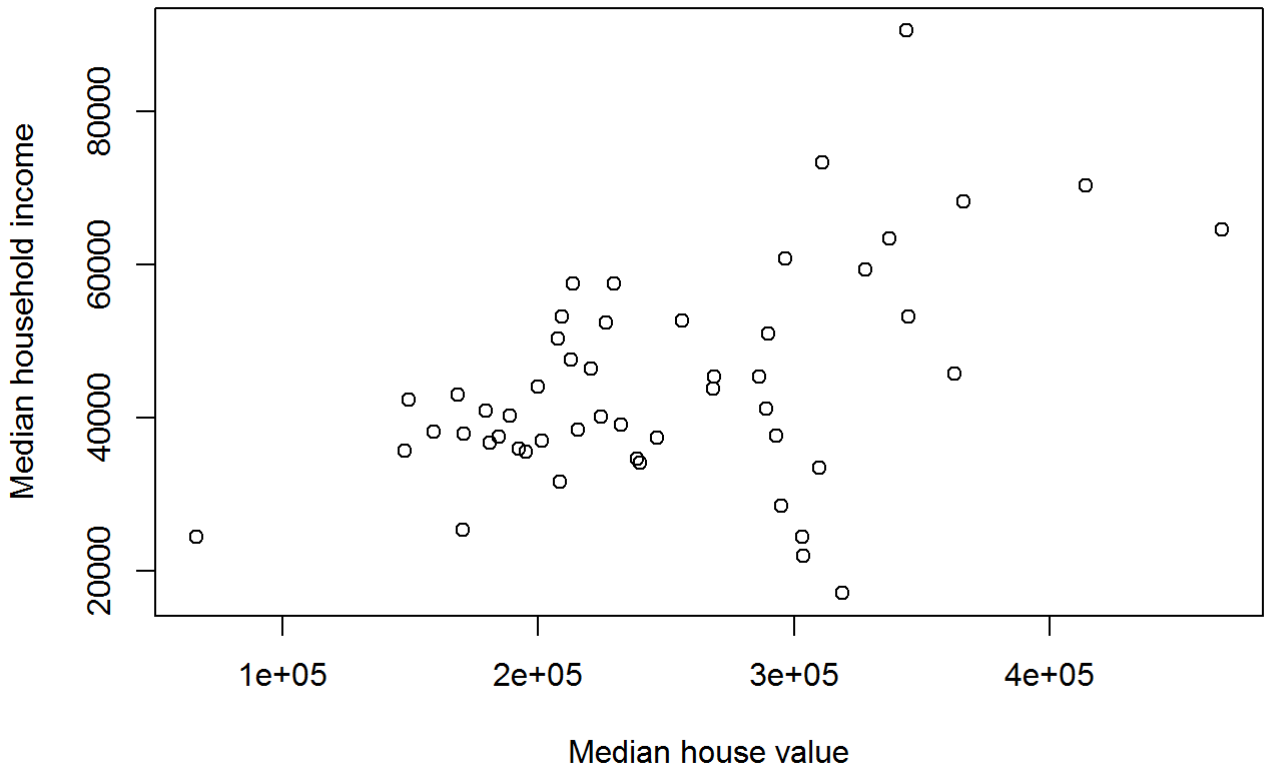
```
# Santa Clara
sc = omdata[which(omdata$STATEFP == 6 & omdata$COUNTYFP == 85),]
cor(sc[[10]], sc[[16]])
```

```
## [1] -0.1726203
```

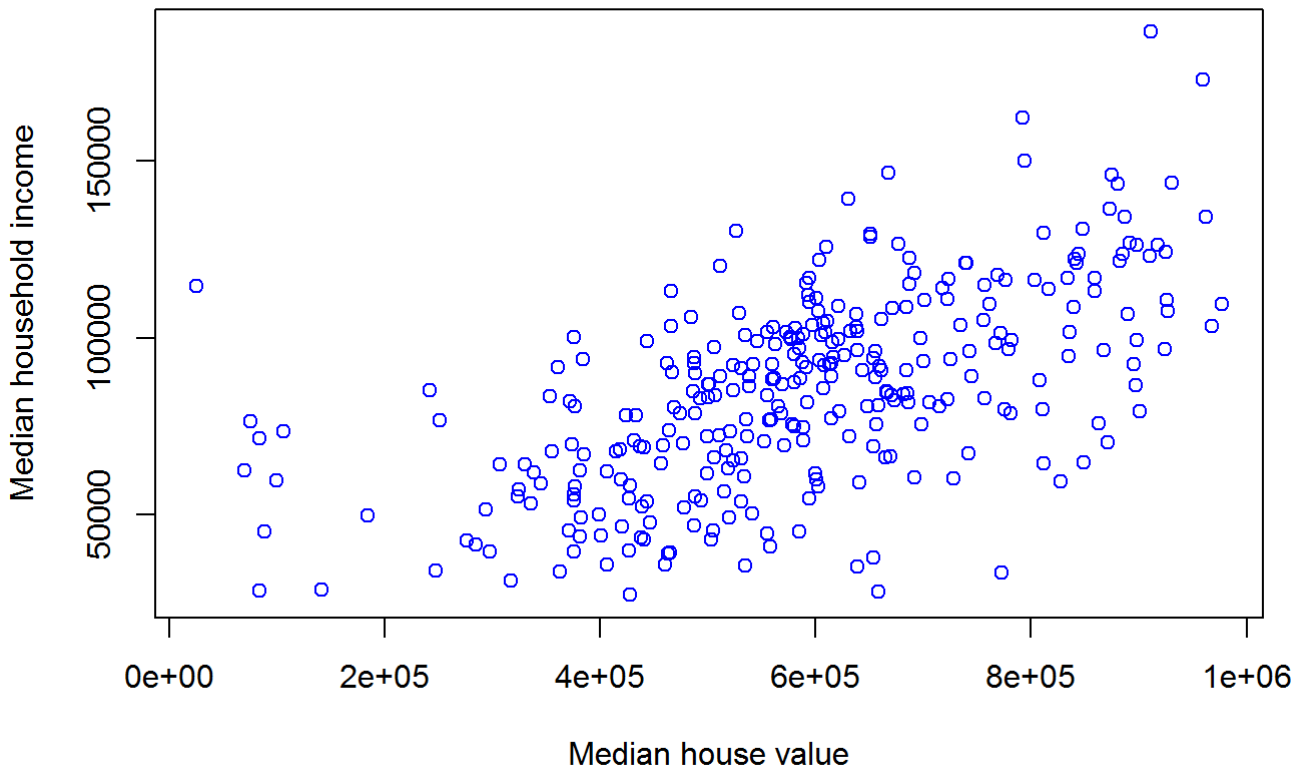
```
# York County
yc = omdata[which(omdata$STATEFP == 42 & omdata$COUNTYFP == 133),]
cor(yc[[10]], yc[[16]])
```

```
## [1] 0.3860773
```

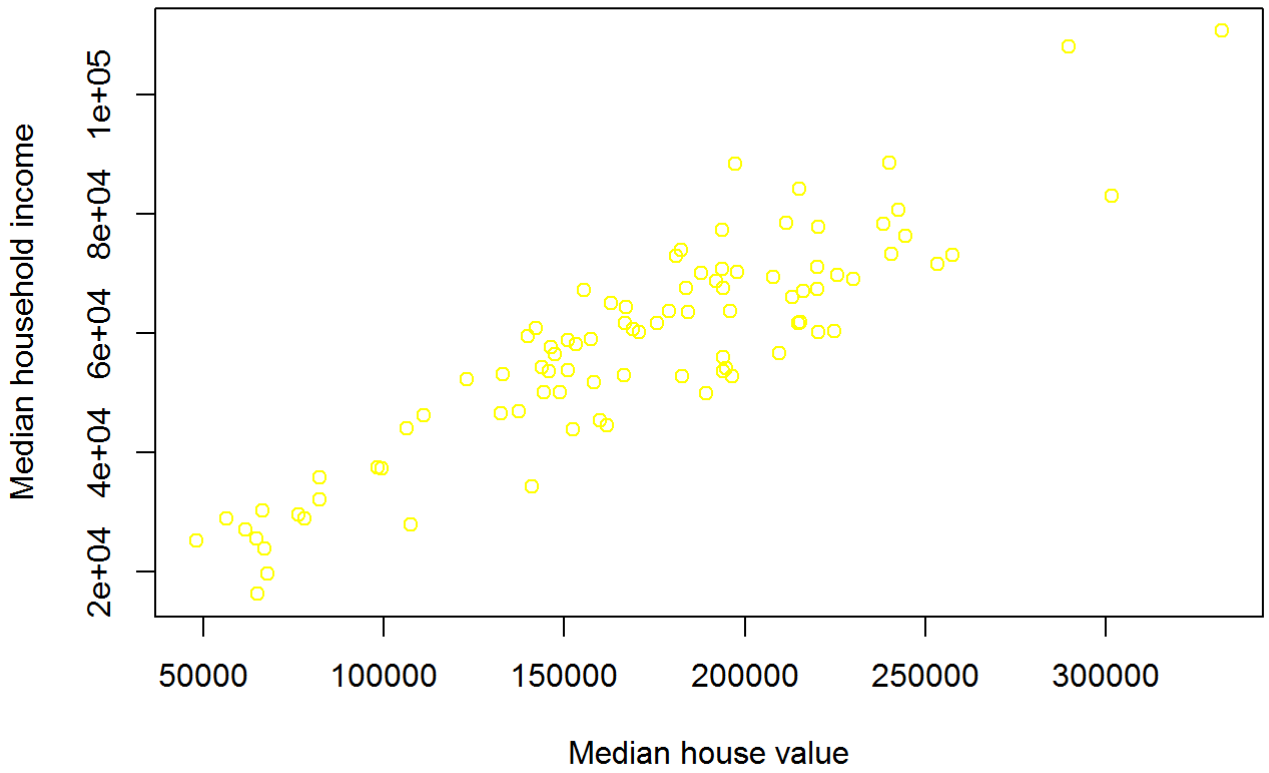
```
# plots
plot(bc$Median_house_value, bc$Median_household_income,
     xlab = "Median house value",
     ylab = "Median household income")
```



```
plot(sc$Median_house_value, sc$Median_household_income,
     col = "blue",
     xlab = "Median house value",
     ylab = "Median household income")
```




```
plot(yc$Median_house_value, yc$Median_household_income,
     col = "yellow",
     xlab = "Median house value",
     ylab = "Median household income")
```


Median household income
2e+04
4e+04
6e+04
8e+04
1e+05
50000
100000
150000
200000
250000
300000
Median house value