HousingPricesCAvsPA

Muhammed Khan

Thursday, May 28, 2015

```
#1
#a
ca_pa <- read.csv("http://people.csail.mit.edu/sylvain/houseprices2011.csv")
# Loading the file into R
#b
dim(ca_pa)  # display the dimension of the dataframe (nrow, ncol)
```

```
## [1] 11275    34
```

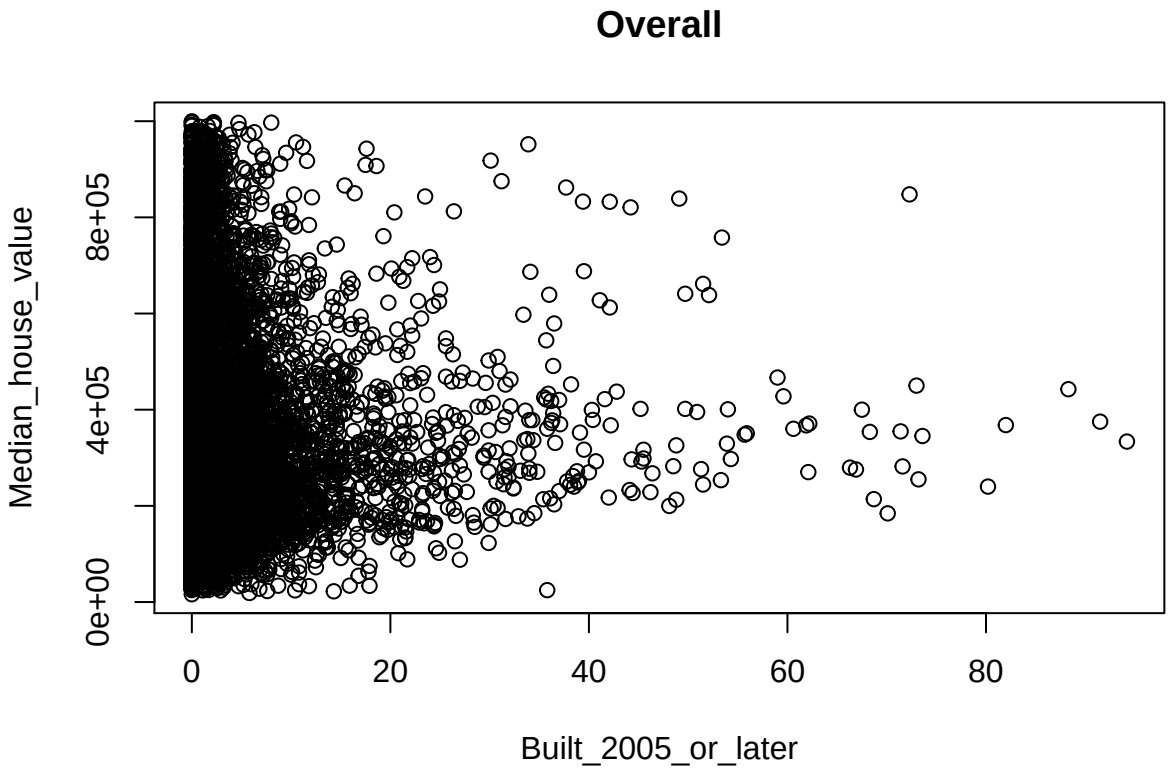
```
#c
colSums(apply(ca_pa,c(1,2),is.na)) # this apply function checks for null values in the
```

```
##                          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
```

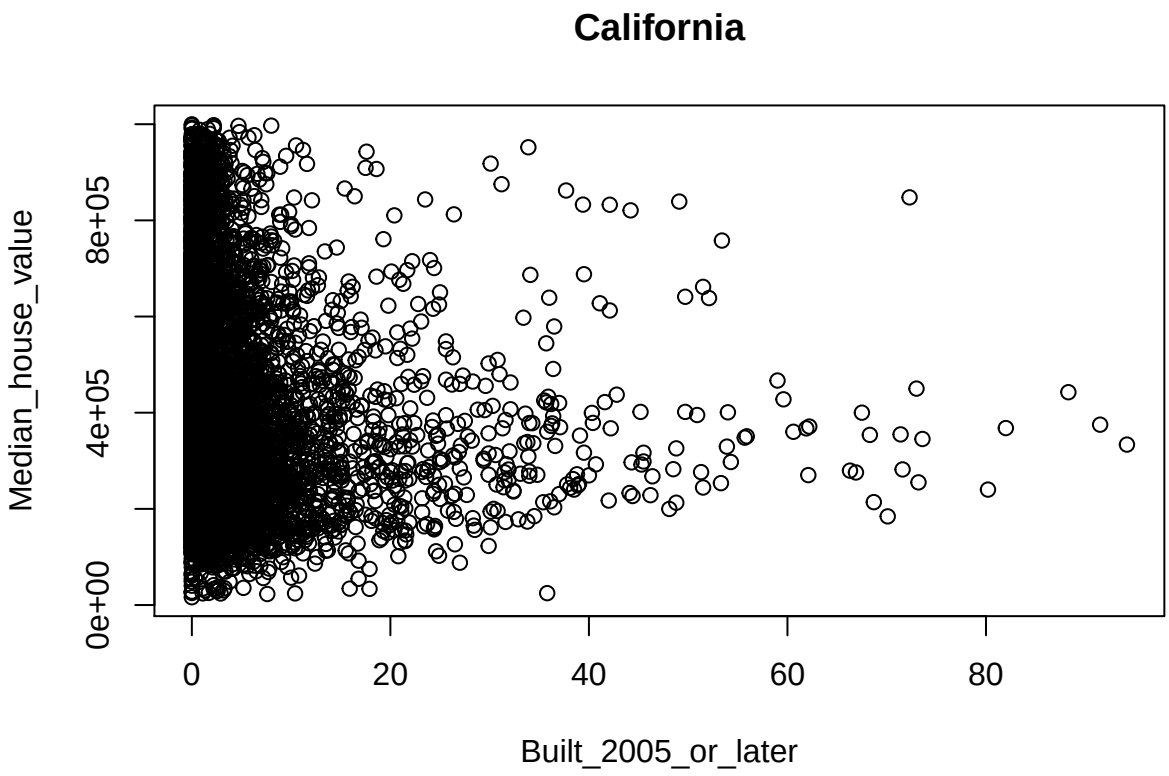
```r
#given dataframe and replaces null with 1.
#ThecolSums displays how many null values are there in each column.
# a vector giving the subscripts which the function will be applied over. E.g.,
# for a matrix 1 indicates rows, 2 indicates columns, c(1, 2) indicates rows and columns.
# Where X has named dimnames, it can be a character vector selecting dimension names.
x <- nrow(ca_pa)
#d
ca_pa <- na.omit(ca_pa)
# eliminates the null records from the dataset
#e
row_omit <- x - nrow(ca_pa)
# no of rows eliminated
print(row_omit)
```

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

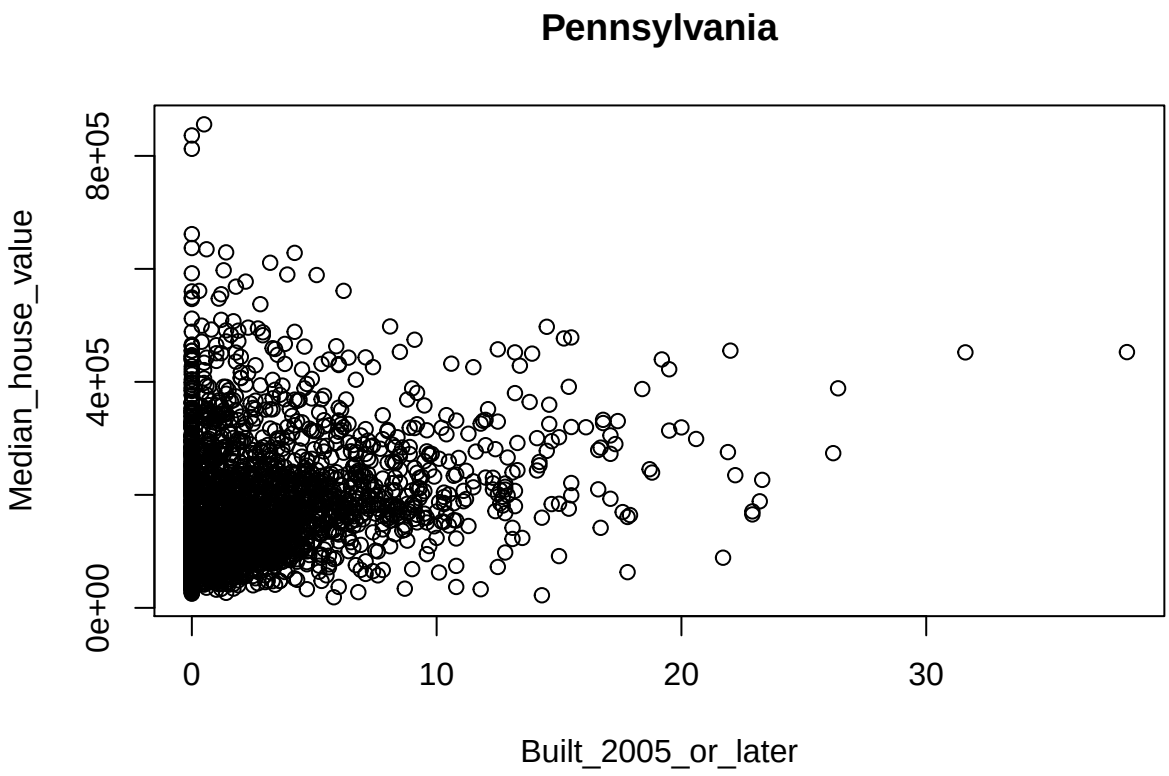
```r
#670
#f

#******
#2)

#a)
# Attach is a function that allows one to search variable in R shell path
attach(ca_pa)
plot(Built_2005_or_later, Median_house_value, main= "Overall", xlab = "Built_2005_or_later", ylab = "Me
```




```
#b)
ca <- subset(ca_pa, ca_pa$STATEFP == 6)
#Filtering out the data for California
pa <- subset(ca_pa, ca_pa$STATEFP == 42)
#Filtering out the data for Pennsylvania
plot(ca$Built_2005_or_later, ca$Median_house_value, main= "California", xlab = "Built_2005_or_later", yl
```



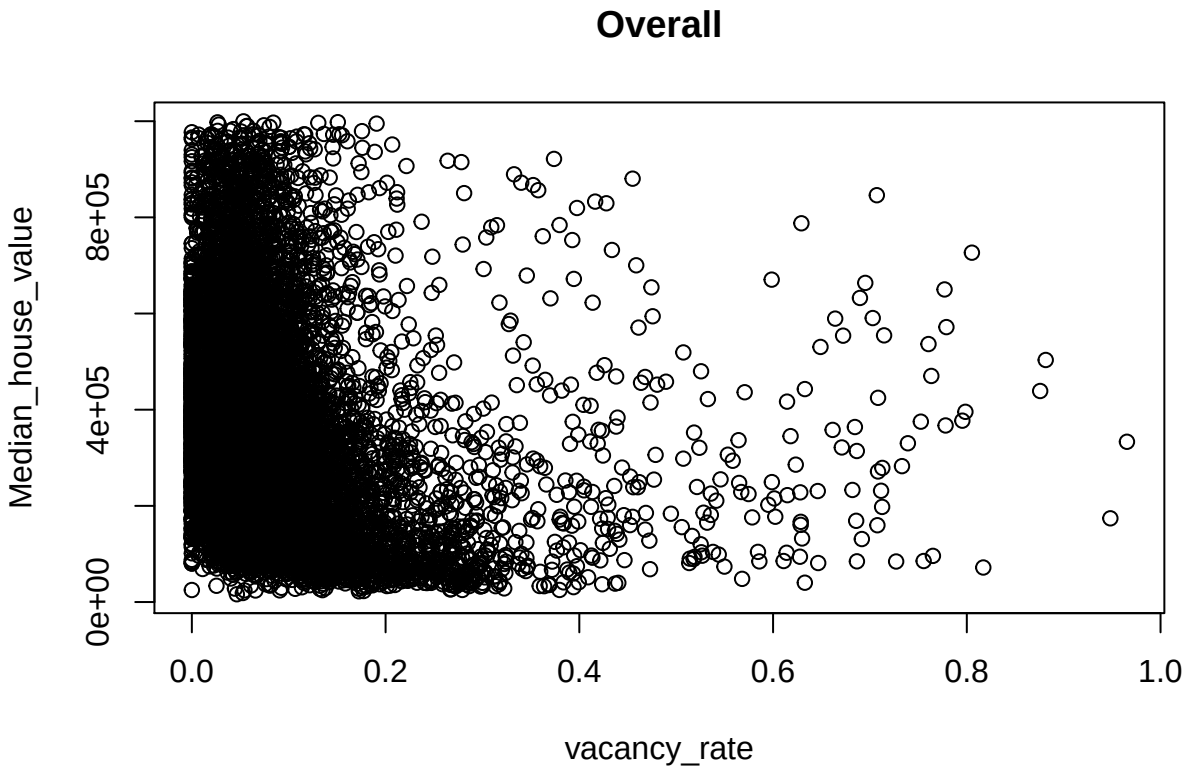

```
plot(pa$Built_2005_or_later, pa$Median_house_value, main= "Pennsylvania", xlab = "Built_2005_or_later",
```



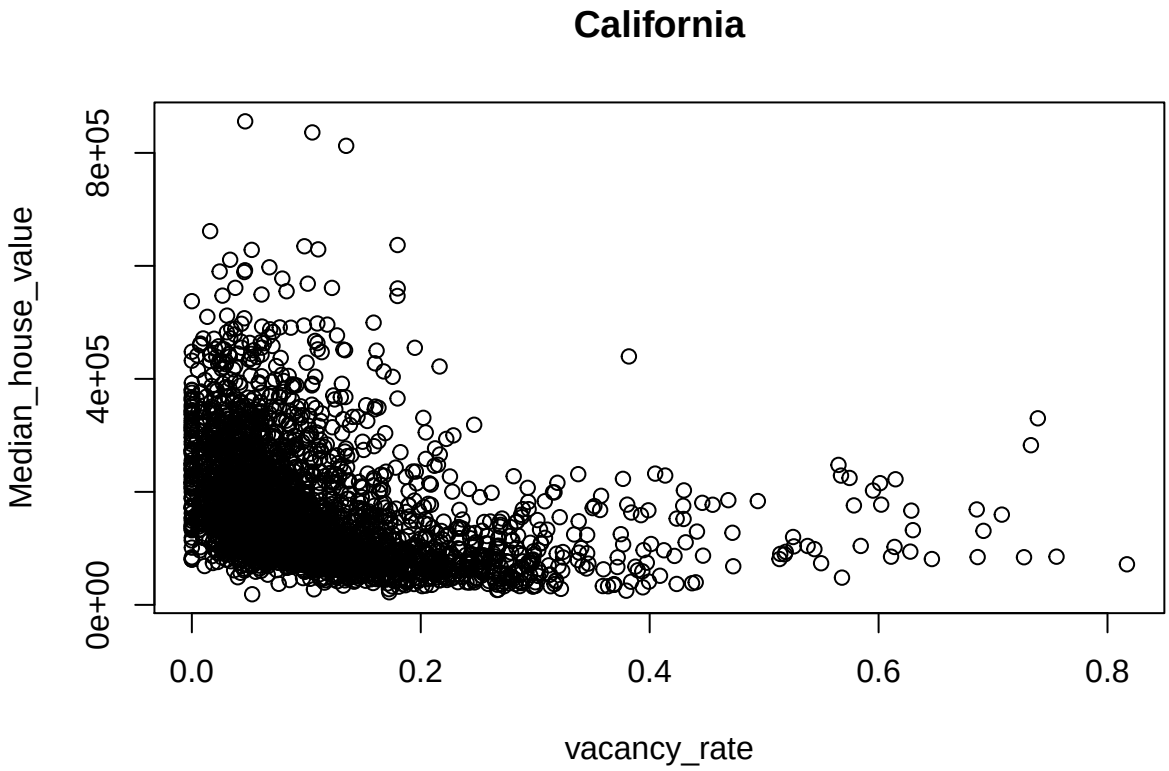

```
#3)
#a)

ca_pa$vacancy_rate <- ca_pa$Vacant_units/ca_pa$Total_units
# Create new column vacancy rate
summary(ca_pa$vacancy_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03846 0.06767 0.08889 0.10920 0.96530
```

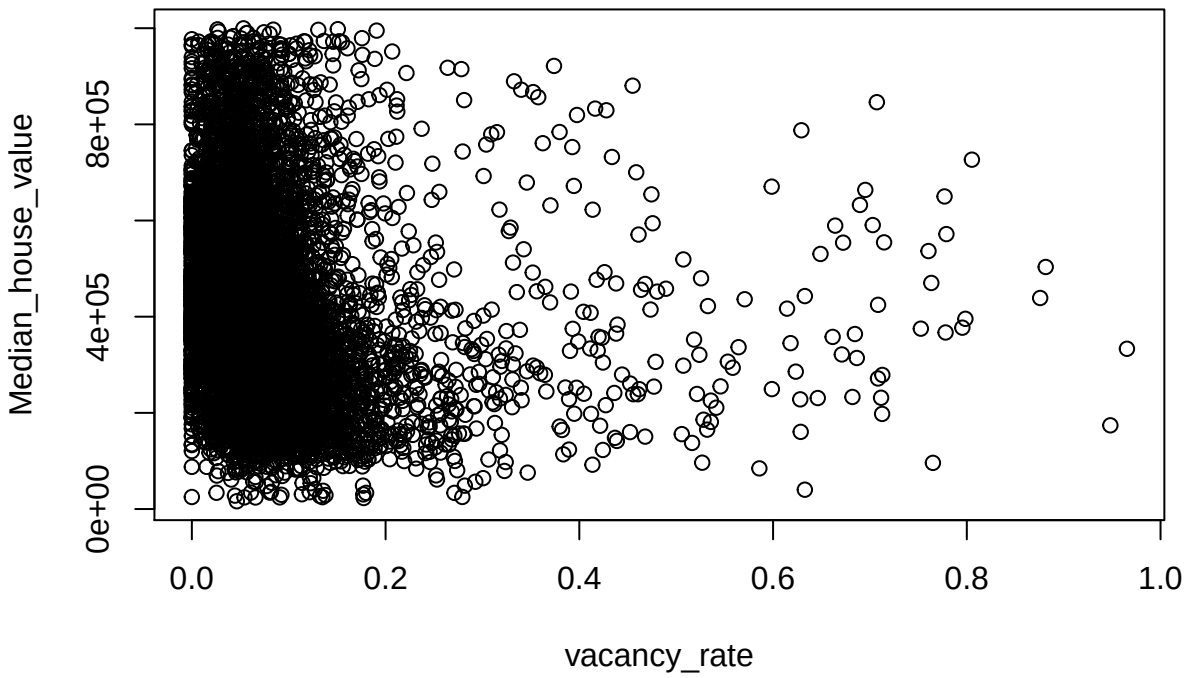
```
# function summary gives the min max, mean and median of the particular column.
#to visualize before allocating computing resources to plot
#b)
plot(ca_pa$vacancy_rate, ca_pa$Median_house_value, main= "Overall", xlab = "vacancy_rate", ylab = "Medi
```




```
#c)
ca <- subset(ca_pa, ca_pa$STATEFP == 6)
pa <- subset(ca_pa, ca_pa$STATEFP == 42)
plot(pa$vacancy_rate, pa$Median_house_value, main= "California", xlab = "vacancy_rate", ylab = "Median_
```




```
plot(ca$vacancy_rate, ca$Median_house_value, main= "Pennsylvania", xlab = "vacancy_rate", ylab = "Median
```




```
#is there a difference between ca and pa vacancy rate plots?

#4)
#a)
acca <- c()
for (tract in 1:nrow(ca_pa)) {
  if (ca_pa$STATEFP[tract] == 6) {
    if (ca_pa$COUNTYFP[tract] == 1) { #county embeded in state
      acca <- c(acca, tract)
    }
  }
}
# this code extracts the row names of the data where state = 6 AND county = 1 which is #Alameda County,
accamhv <- c()
for (tract in acca) {
  accamhv <- c(accamhv, ca_pa[tract,10])
}
#extracts the median house value for the county Alameda
#median(accamhv) median of the median house value column

#b)
bracket_code = median((subset(ca_pa, (ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1)))[,10])
#subset of ca_pa where StateFP equivalent to 6 && County =1,

#c)
ala <- subset(ca_pa, (ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1))
```

```r
#create subsets for counties
santa <- subset(ca_pa, (ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85))
allegh <- subset(ca_pa, (ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3))

mean(ala$Built_2005_or_later)
```

```
## [1] 2.820468
```

```r
mean(santa$Built_2005_or_later)
```

```
## [1] 3.200319
```

```r
mean(allegh$Built_2005_or_later)
```

```
## [1] 1.474219
```

```r
#d)
cor(ca_pa$Median_house_value, ca_pa$Built_2005_or_later) # Calculating correlation between #median hous
```

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

```r
cor(ca$Median_house_value, ca$Built_2005_or_later)
```

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

```r
cor(pa$Median_house_value, pa$Built_2005_or_later)
```

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

```r
cor(ala$Median_house_value, ala$Built_2005_or_later)
```

```
## [1] 0.01303543
```

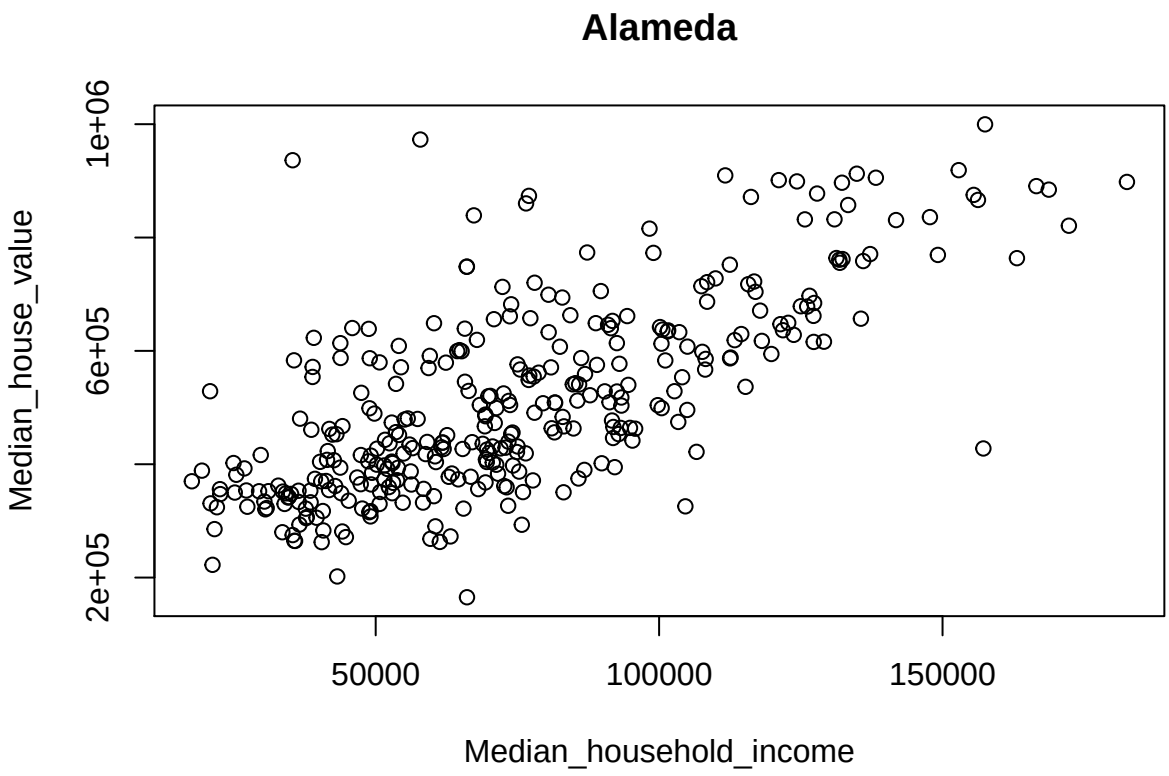
```r
cor(santa$Median_house_value, santa$Built_2005_or_later)
```

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

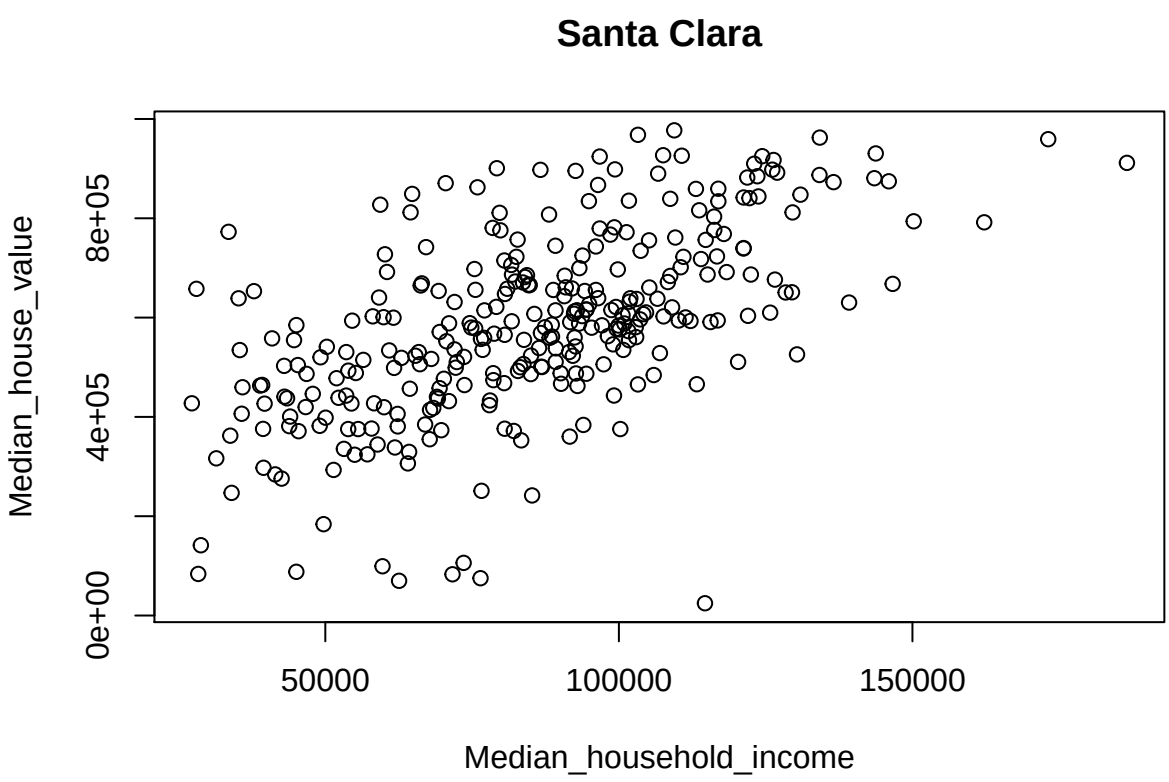
```r
cor(allegh$Median_house_value, allegh$Built_2005_or_later)
```

```
## [1] 0.1939652
```

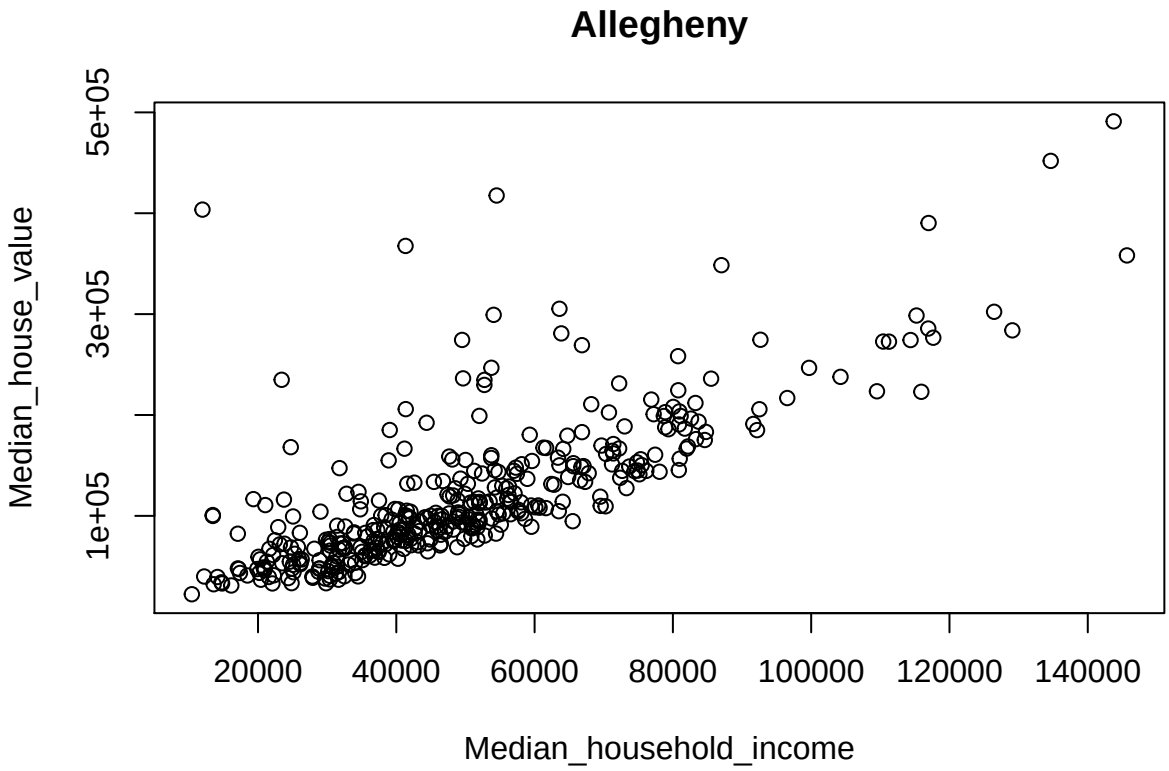
```r
#e)
plot(ala$Median_household_income, ala$Median_house_value, main= "Alameda", ylab = "Median_house_value",
```




```
plot(santa$Median_household_income, santa$Median_house_value, main= "Santa Clara", ylab = "Median_house
```




```
plot(allegh$Median_household_income, allegh$Median_house_value, main= "Allegheny", ylab = "Median_house
```


Allegheny
Median_house_value
1e+05
3e+05
5e+05
20000
40000
60000
80000
100000
120000
140000
Median_household_income