Stat 597A Homework Problem Set2

Thananya Saksuriyongse

Sept 25, 2015

The data set at [http://www.math.umass.edu/~jstauden/calif_penn_2011.csv] contains information about the housing stock of California and Pennsylvania, as of 2011. Information as aggregated into "Census tracts", geographic regions of a few thousand people which are supposed to be fairly homogeneous economically and socially.

1. *Loading and cleaning*

a. Load the data into a dataframe called `ca_pa`.

```
ca_pa <- read.csv("http://www.math.umass.edu/~jstauden/calif_penn_2011.csv")
#summary(ca_pa)
```

b. How many rows and columns does the dataframe have?

```
nrow(ca_pa)                 #number of rows
```

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

```
ncol(ca_pa)                 #number of columns
```

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

```
oldrow <- nrow(ca_pa)
```

c. Run this command, and explain, in words, what this does:

```
colSums(apply(ca_pa,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
```

This command gives the total number that NA appears in each column. First, 'apply(ca_pa,c(1,2),is.na)' check through both row and column if there is any NA present if yes give 1 and 0 for no. Then the 'colsums' just sum those numbers for each column.

d. The function `na.omit()` takes a dataframe and returns a new dataframe, omitting any row containing an NA value. Use it to purge the data set of rows with incomplete data.

```
ca_pa <- na.omit(ca_pa)
nrow(ca_pa)
```

```
## [1] 10605
```

```
newrow <- nrow(ca_pa)                    # number of row after eliminate NA
```

e. How many rows did this eliminate?

```
oldrow - nrow(ca_pa)      # number of rows eliminated
```

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

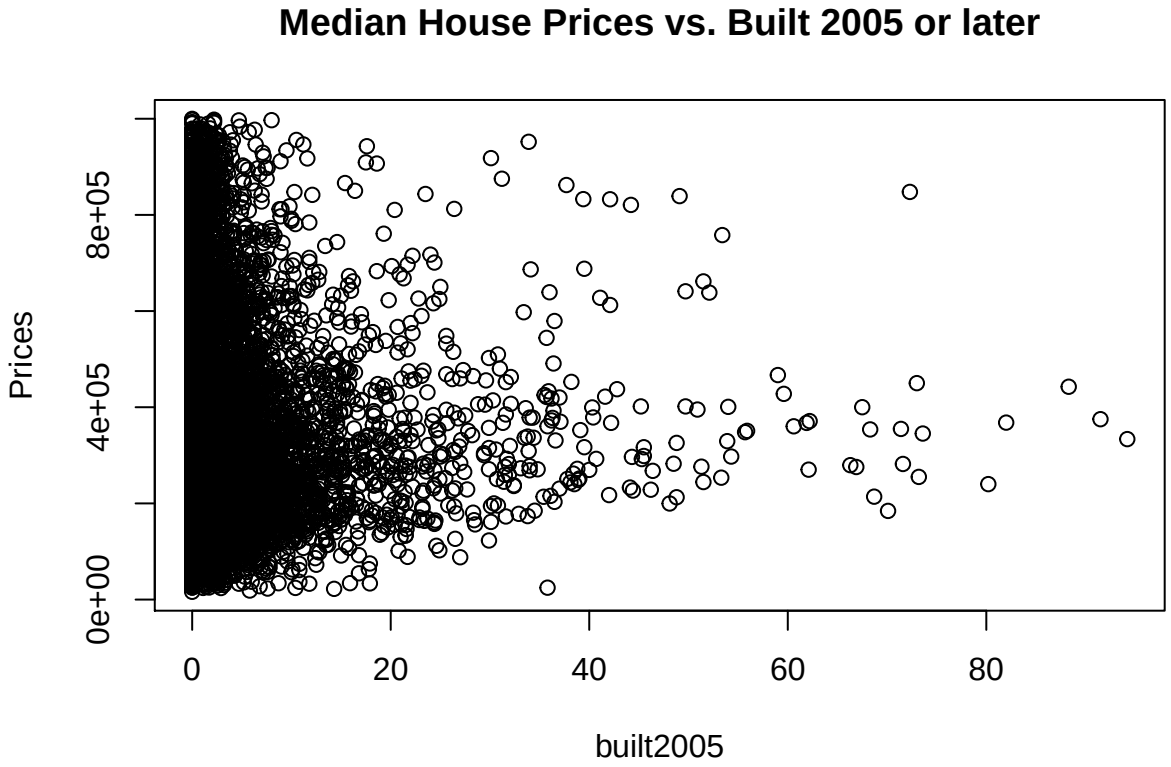
f. Are your answers in (c) and (e) compatible? Explain.

No, I don't think the answers are compatible. In part (c), it goes through row and column to check if there is any NA present and count number of time seeing NA (count twice for each entry) in each column. For part(e), it tells us the total number of line that has NA.

2. *This Very New House*

a. The variable `Built_2005_or_later` indicates the percentage of houses in each Census tract built since 2005. Plot median house prices against this variable.

```
price <- ca_pa$Median_house_value
built2005 <- ca_pa$Built_2005_or_later
plot(built2005, price , ylab = "Prices" , main = "Median House Prices vs. Built 2005 or later")
```

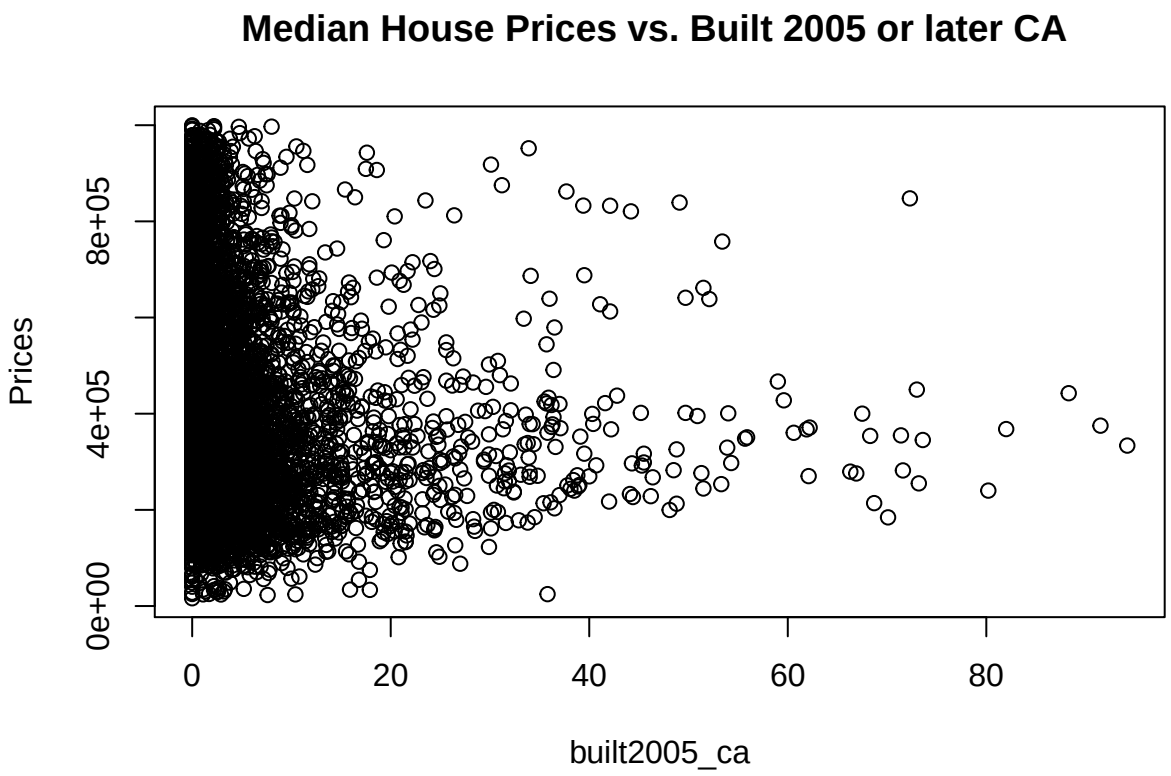


b. Make a new plot, or pair of plots, which breaks this out by state. Note that the state is recorded in the `STATEFP` variable, with California being state 6 and Pennsylvania state 42.

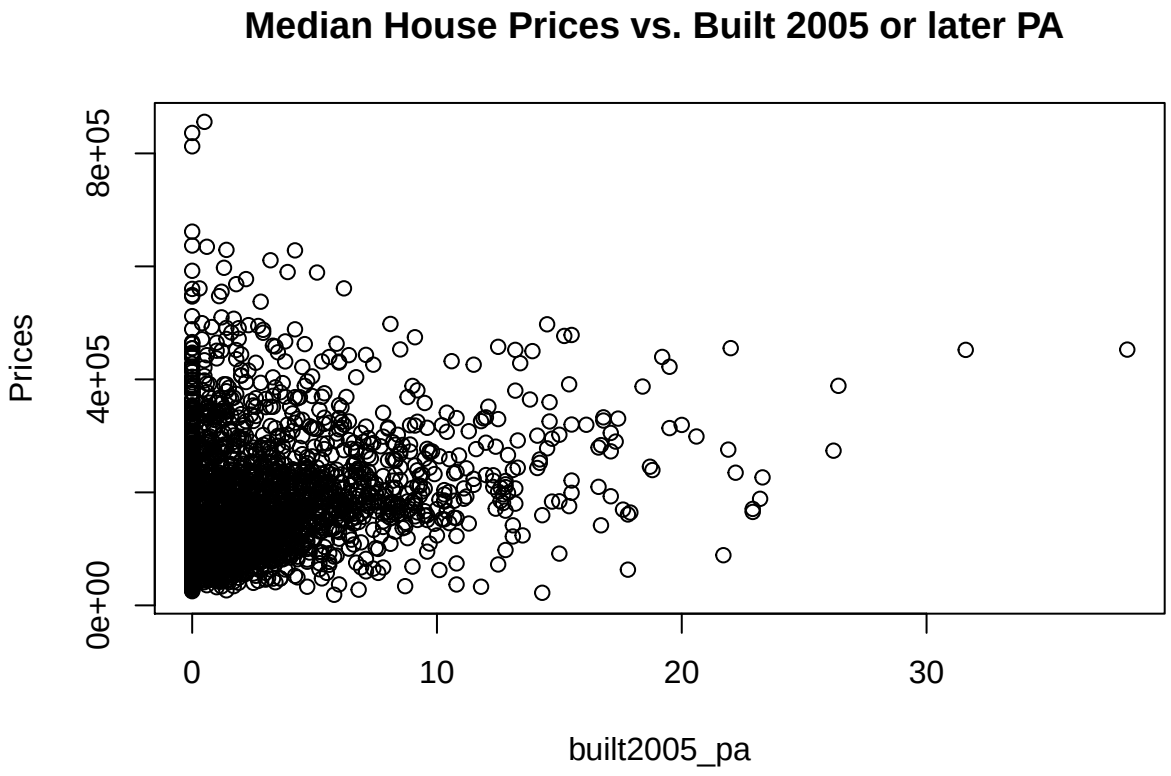
```
index <- ca_pa$STATEFP == 6
ca <- ca_pa[index, ]                    #get all info for the row that is ca
price_ca <- ca$Median_house_value
built2005_ca <- ca$Built_2005_or_later

pa <- ca_pa[ca_pa$STATEFP == 42, ]
price_pa <- pa$Median_house_value
built2005_pa <- pa$Built_2005_or_later

plot(built2005_ca, price_ca, ylab = "Prices" , main = "Median House Prices vs. Built 2005 or later CA "
```




```
plot(built2005_pa, price_pa, ylab = "Prices" , main = "Median House Prices vs. Built 2005 or later PA "
```



3. *Nobody Home*
   The vacancy rate is the fraction of housing units which are not occupied. The dataframe contains

columns giving the total number of housing units for each Census tract, and the number of vacant housing units.

a. Add a new column to the dataframe which contains the vacancy rate. What are the minimum, maximum, mean, and median vacancy rates?

```
total <- ca_pa$Total_units
vacant <- ca_pa$Vacant_units
vacancy_rate <- vacant/total

ca_pa$Vacancy_rate <- vacancy_rate          #adding new column to ca_pa1
min(vacancy_rate)
```

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

```
max(vacancy_rate)
```

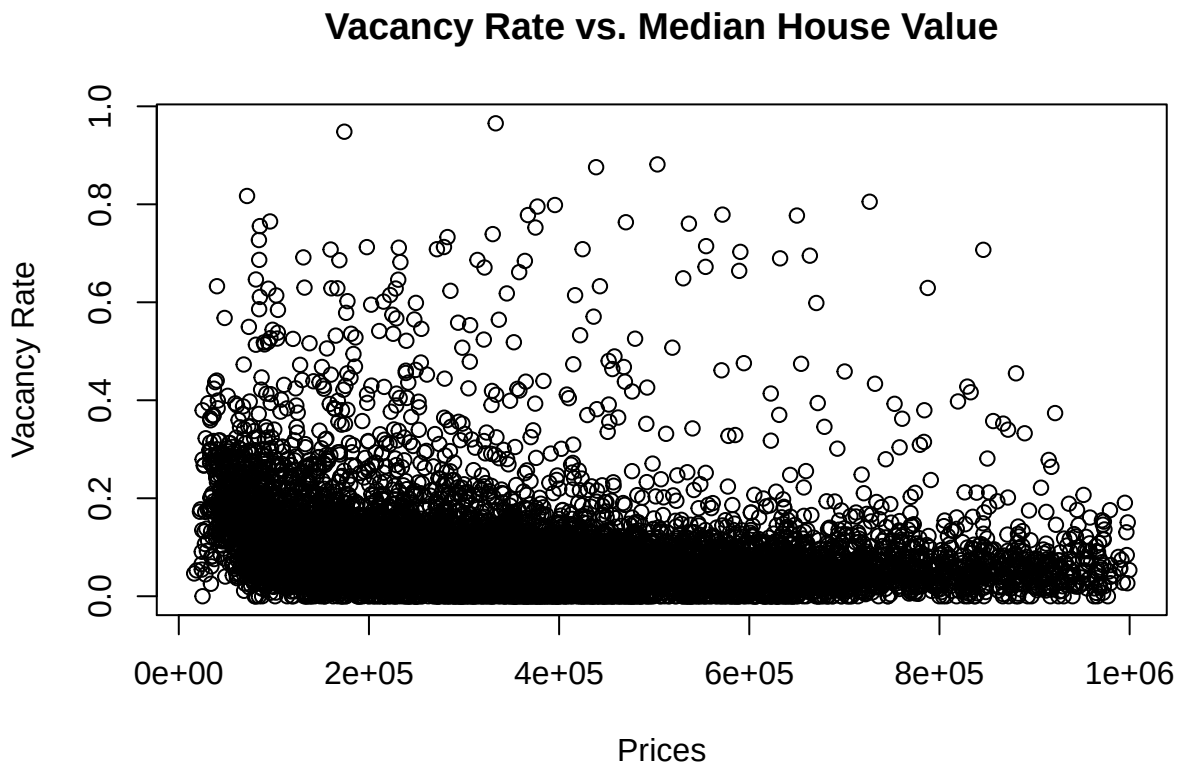
```
## [1] 0.965311
```

```
median(vacancy_rate)
```

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

b. Plot the vacancy rate against median house value.

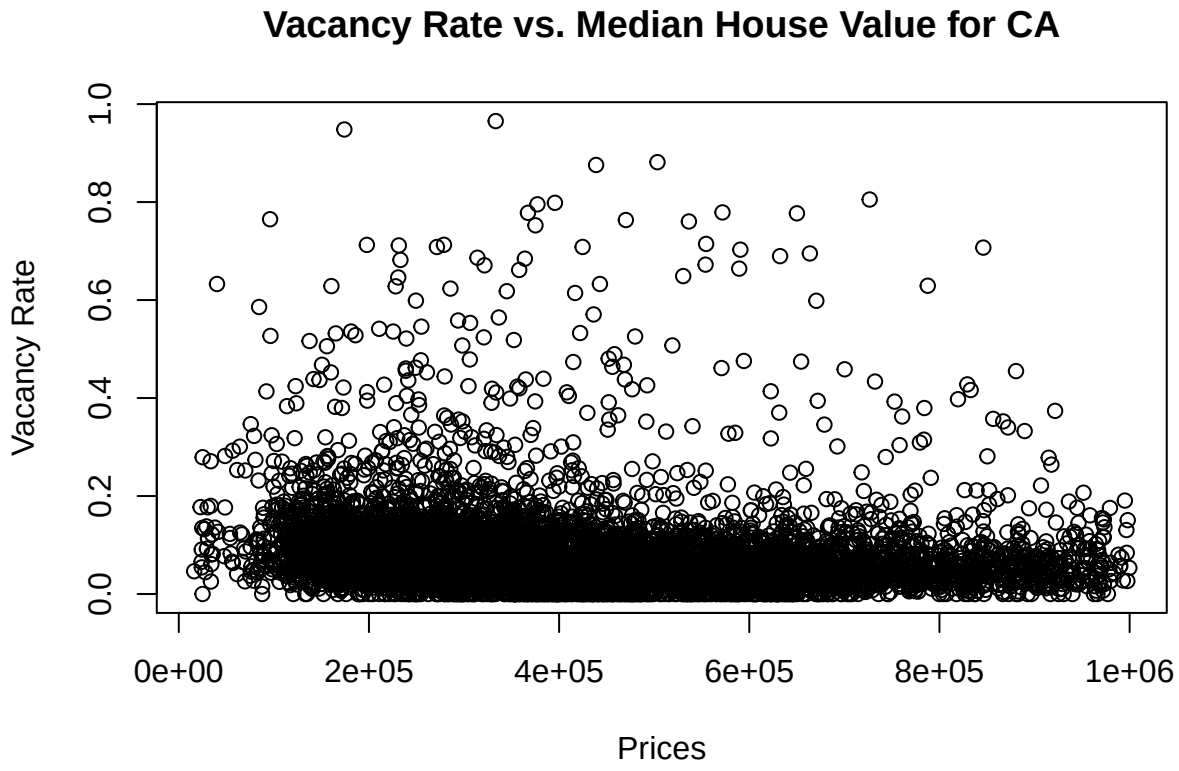
```
plot( price, vacancy_rate, xlab = "Prices" , ylab = "Vacancy Rate", main = "Vacancy Rate vs. Median Hous
```



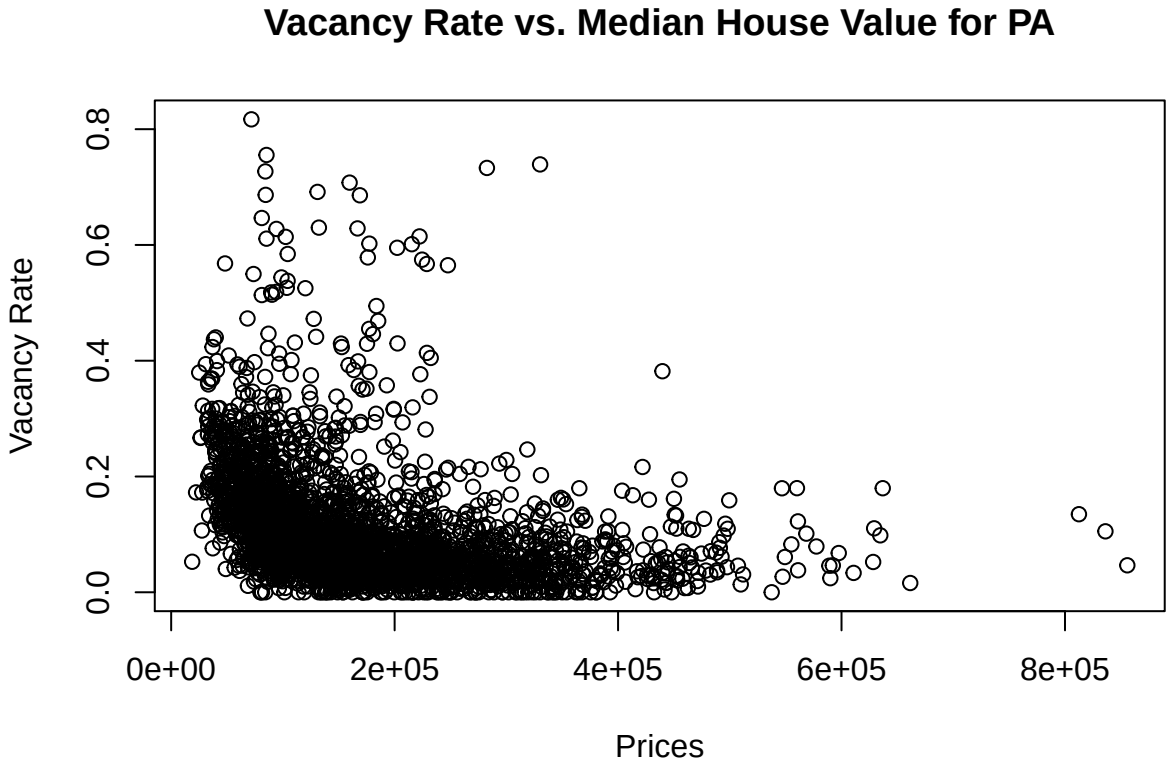
c. Plot vacancy rate against median house value separately for California and for Pennsylvania. Is there a difference?

```
rate_ca <- vacancy_rate[ca_pa$STATEFP == 6]
rate_pa <- vacancy_rate[ca_pa$STATEFP == 42]

plot( price_ca, rate_ca, xlab = "Prices" , ylab = "Vacancy Rate", main = "Vacancy Rate vs. Median House
```




```
plot( price_pa, rate_pa, xlab = "Prices" , ylab = "Vacancy Rate", main = "Vacancy Rate vs. Median House
```



4. The column `COUNTYFP` contains a numerical code for counties within each state. We are interested in Alameda County (county 1 in California), Santa Clara (county 85 in California), and Allegheny County (county 3 in Pennsylvania).

a. Explain what the block of code at the end of this question is supposed to accomplish, and how it does it.

```
acca <- c()
for (tract in 1:nrow(ca_pa)) {
  if (ca_pa$STATEFP[tract] == 6) {
    if (ca_pa$COUNTYFP[tract] == 1) {
      acca <- c(acca, tract)                                    # get a vector [1,...,342]
      }
    }
  }
```

The for loop will go through every row of dataframe, each time it will check if it is in state CA. If yes, then it checks if it is in Alameda country and store the row indices in a vector called acca. To get Santa Clara, we can change the second if statement to be if (ca_pa$COUNTYFP[tract] == 85). Similar idea to get Allegheny country (need to change both if statements == 42 and == 3).

```
accamhv <- c()
for (tract in acca) {
  accamhv <- c(accamhv, ca_pa[tract,10])                    # get a value oh mhv for the first 342
  }
median(accamhv)
```

b. Give a single line of R which gives the same final answer as the block of code. Note: there are at least two ways to do this; you just have to find one.

```r
# cali <- ca_pa[ca_pa$STATEFP == 6 , ]
# alemeda <- cali[ca_pa$COUNTYFP == 1, ]
# alemeda <- ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1 , 10]
# median(alemeda)
median(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1 , 10])
```

```
## [1] 474050
```

c. For Alameda, Santa Clara and Allegheny Counties, what were the average percentages of housing built since 2005?

```r
alemeda_house <- ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1 , ]$Built_2005_or_later
mean(alemeda_house)
```

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

```r
santa_clara_house <- ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85 , ]$Built_2005_or_later
mean(santa_clara_house)
```

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

```r
allegheny_house <- ca_pa[ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3 , ]$Built_2005_or_later
mean(allegheny_house)
```

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

d. The `cor` function calculates the correlation coefficient between two variables. What is the correlation between median house value and the percent of housing built since 2005 in (i) the whole data, (ii) all of California, (iii) all of Pennsylvania, (iv) Alameda County, (v) Santa Clara County and (vi) Allegheny County?

```r
cor(ca_pa$Median_house_value, ca_pa$Built_2005_or_later)              # correlation for the whole data
```

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

```r
cor(ca$Median_house_value , ca$Built_2005_or_later)                   # correlation for all of Californi
```

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

```r
cor(pa$Median_house_value , pa$Built_2005_or_later)                   # correlation for all of Pennsylva
```

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

```r
alemeda <- ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1 , ]
santa_clara <- ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85 , ]
allegheny <- ca_pa[ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3 , ]

cor(alemeda$Median_house_value, alemeda$Built_2005_or_later)            # correlation for Alameda
```

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

```
cor(santa_clara$Median_house_value, santa_clara$Built_2005_or_later)    # correlation for Santa Clara C
```

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

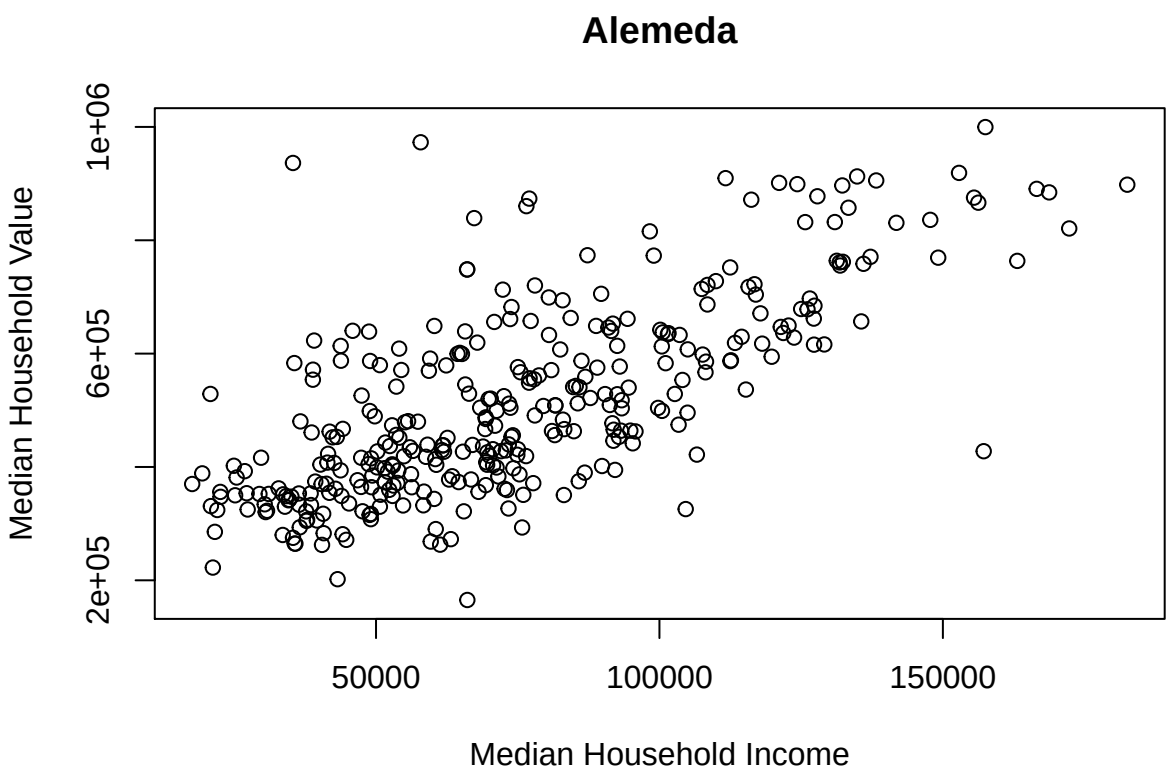
```
cor(allegheny$Median_house_value, allegheny$Built_2005_or_later)        # correlation for Allegheny
```

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

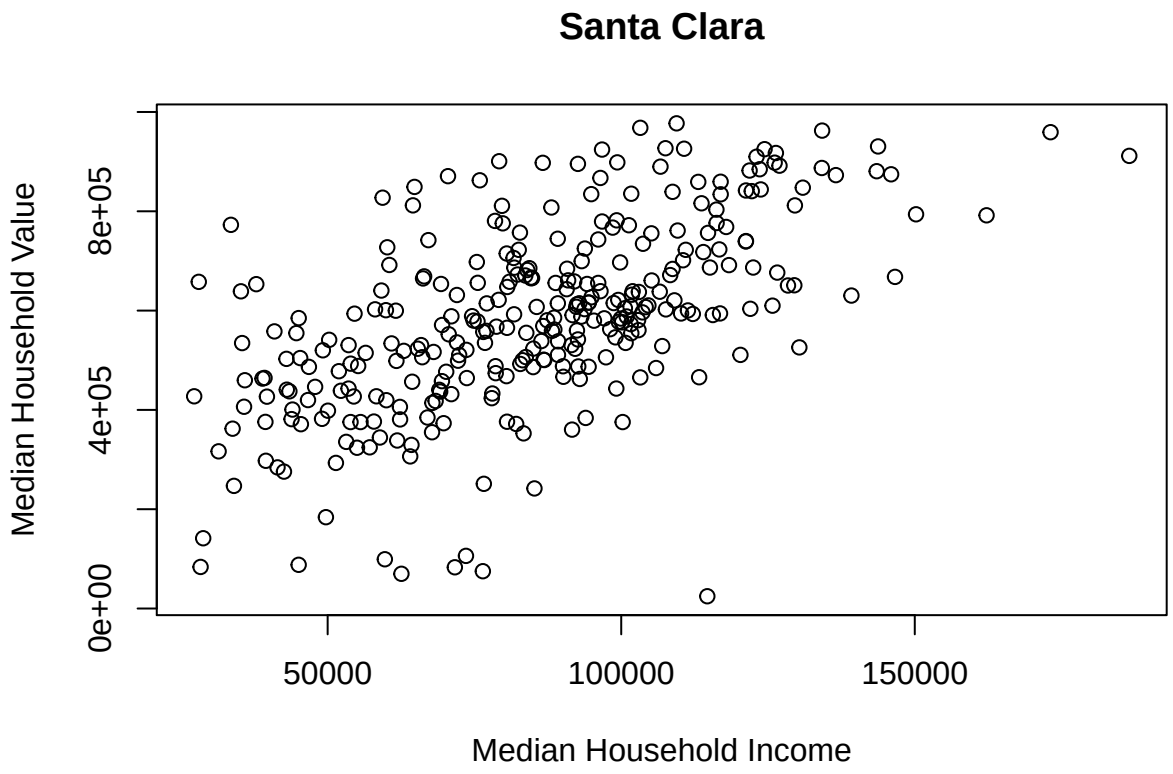
```
# cor(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1 , "Median_house_value"], ca_pa[ca_pa$STATEFP == 6 &
# cor(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85 , "Median_house_value"], ca_pa[ca_pa$STATEFP == 6
# cor(ca_pa[ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3 , "Median_house_value"], ca_pa[ca_pa$STATEFP == 4
```

e. Make three plots, showing median house values against median income, for Alameda, Santa Clara, and Allegheny Counties.
(If you can fit the information into one plot, clearly distinguishing the three counties, that's OK too.)

```
plot(alemeda$Median_household_income , alemeda$Median_house_value ,
     xlab = "Median Household Income" , ylab = "Median Household Value", main = "Alemeda")
```




```
plot(santa_clara$Median_household_income , santa_clara$Median_house_value ,
     xlab = "Median Household Income" , ylab = "Median Household Value", main = "Santa Clara")
```




```
plot(allegheny$Median_household_income , allegheny$Median_house_value ,
     xlab = "Median Household Income" , ylab = "Median Household Value", main = "Allegheny")
```