Homework 2

Di Fan

2020/07/08

The data set 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("data/calif_penn_2011.csv")
```

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

```
nrow(ca_pa)
```

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

```
ncol(ca_pa)
```

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

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

```
colSums(apply(ca_pa,c(1,2),is.na))#To Count the number of missing data in each column
```

```
##                          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
```

d. The function 'na.omit()' takes a dataframe and returns a new dataframe, omitting any row containing a

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

e. How many rows did this eliminate?

```
nrow(ca_panew)
```

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

```
eliminate_rownumber<-nrow(ca_pa)-nrow(ca_panew)
eliminate_rownumber
```

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

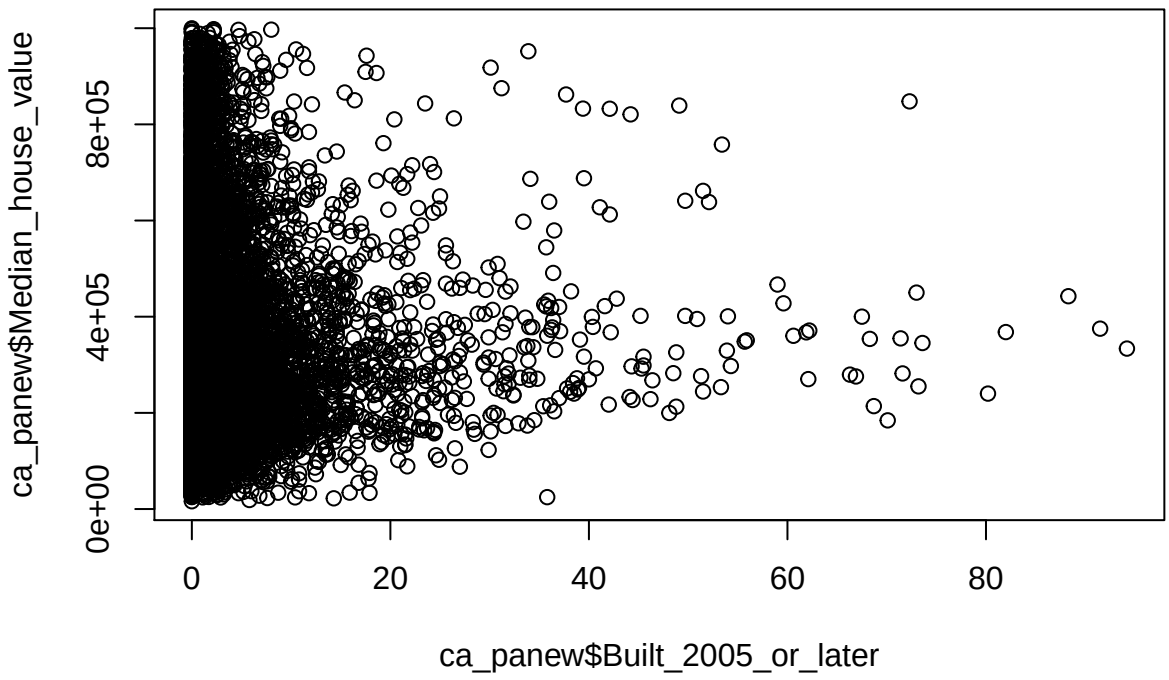
f. Are your answers in (c) and (e) compatible? Explain.

```
#It isn't contradict.Because in the (c),what we're counting is the number of missing columns in each co
```

2. *This Very New House*

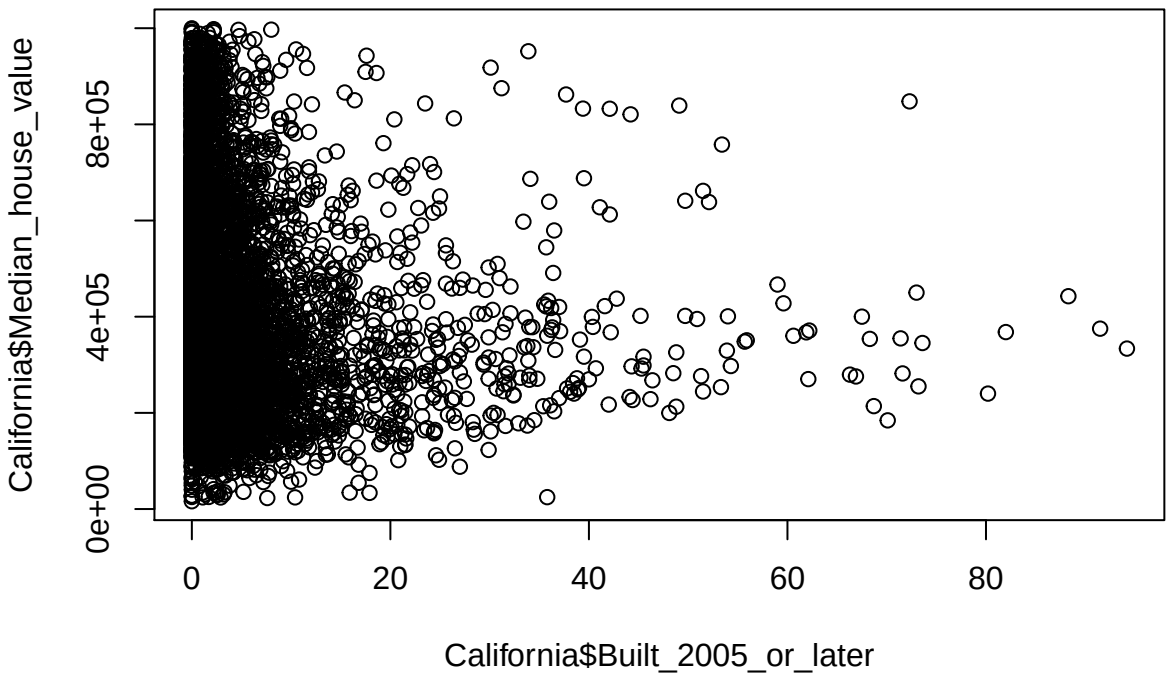
    a. The variable `Built_2005_or_later` indicates the percentage of houses in each Census tract built since 2005. Plot median house prices against this variable.

```
plot(ca_panew$Median_house_value~ca_panew$Built_2005_or_later)
```

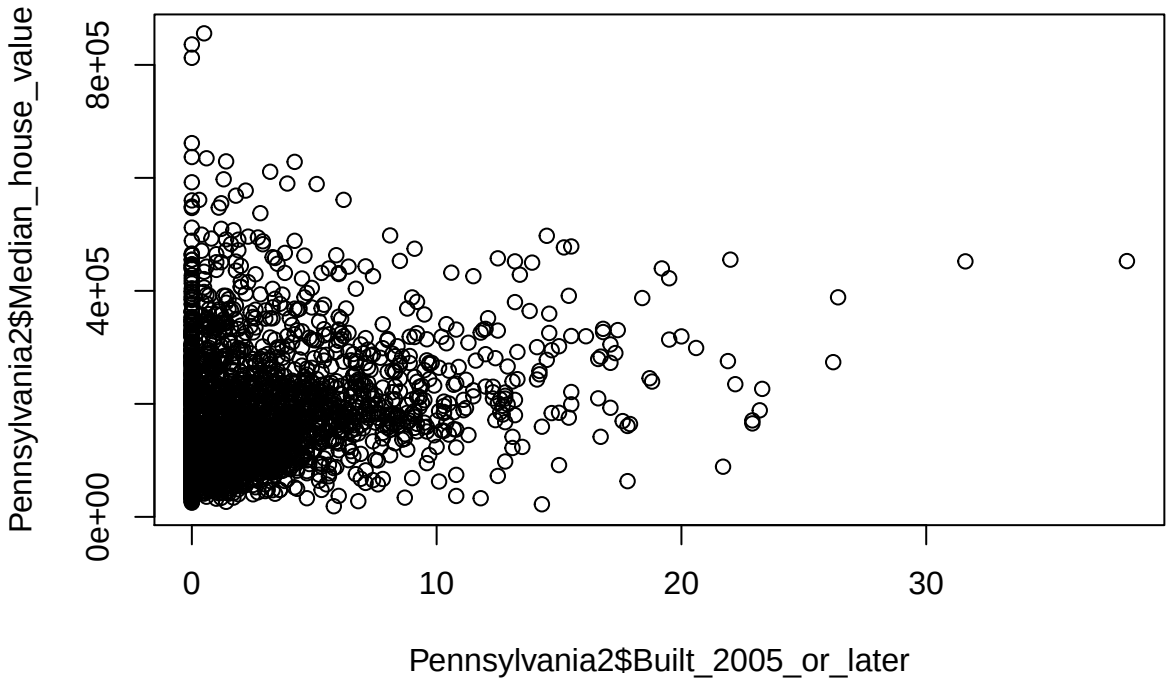


b. Make a new plot, or pair of plots, which breaks this out by state. Note that the state is recorded in the `STATEFP` variable, with California being state 6 and Pennsylvania state 42.

```
S6index<-which(ca_panew$STATEFP==6)
S42index<-which(ca_panew$STATEFP==42)
California<-ca_panew[S6index,]
Pennsylvania2<-ca_panew[S42index,]
plot(California$Median_house_value~California$Built_2005_or_later)
```




```
plot(Pennsylvania2$Median_house_value~Pennsylvania2$Built_2005_or_later)
```



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?

```
ca_panew$Vacany_rates<-ca_panew$Vacant_units/ca_panew$Total_units
```

```
min(ca_panew$Vacany_rates)
```

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

```
max(ca_panew$Vacany_rates)
```

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

```
mean(ca_panew$Vacany_rates)
```

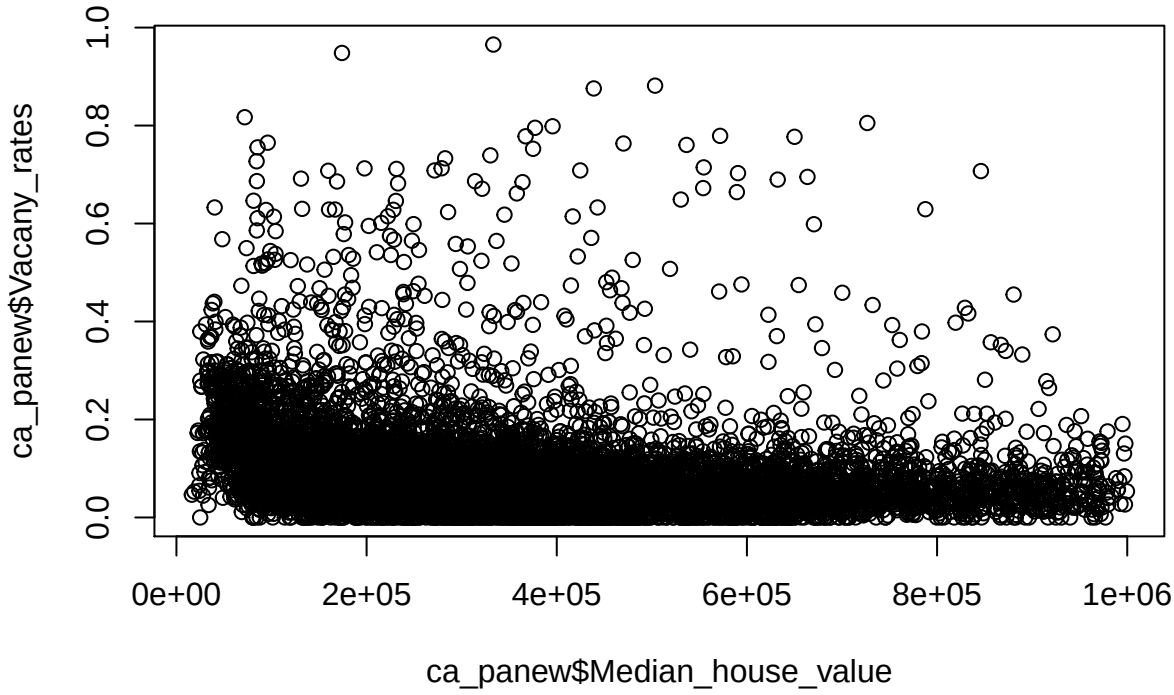
```
## [1] 0.08888789
```

```
median(ca_panew$Vacany_rates)
```

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

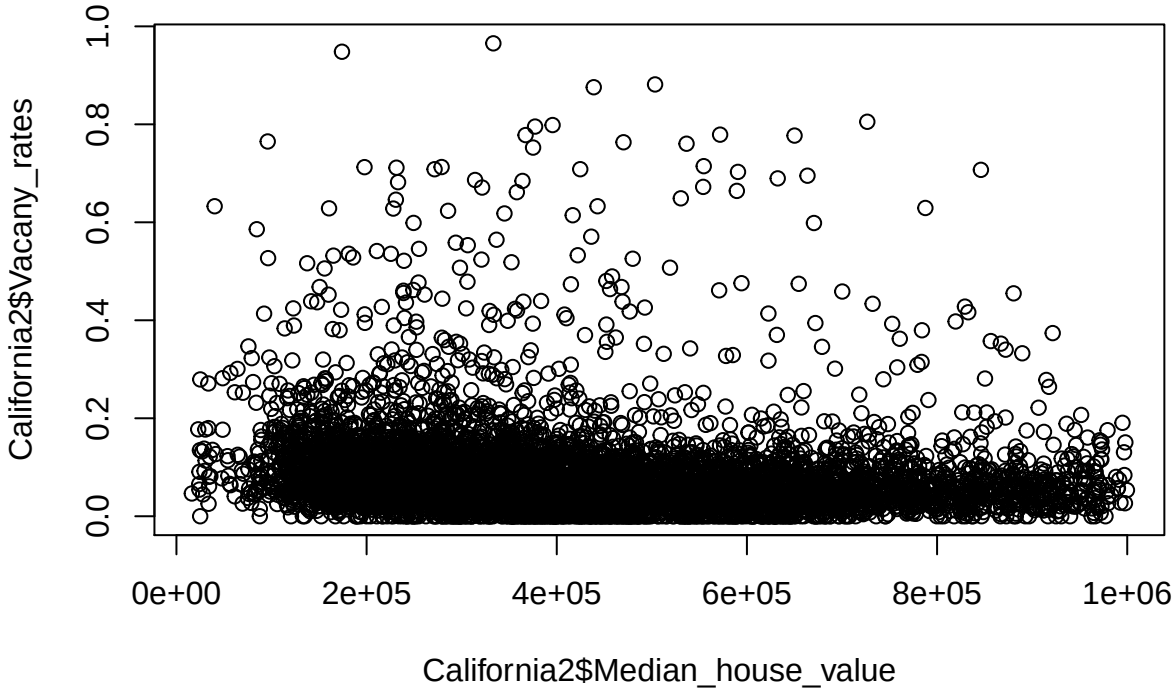
b. Plot the vacancy rate against median house value.

```
plot(ca_panew$Vacany_rates~ca_panew$Median_house_value)
```

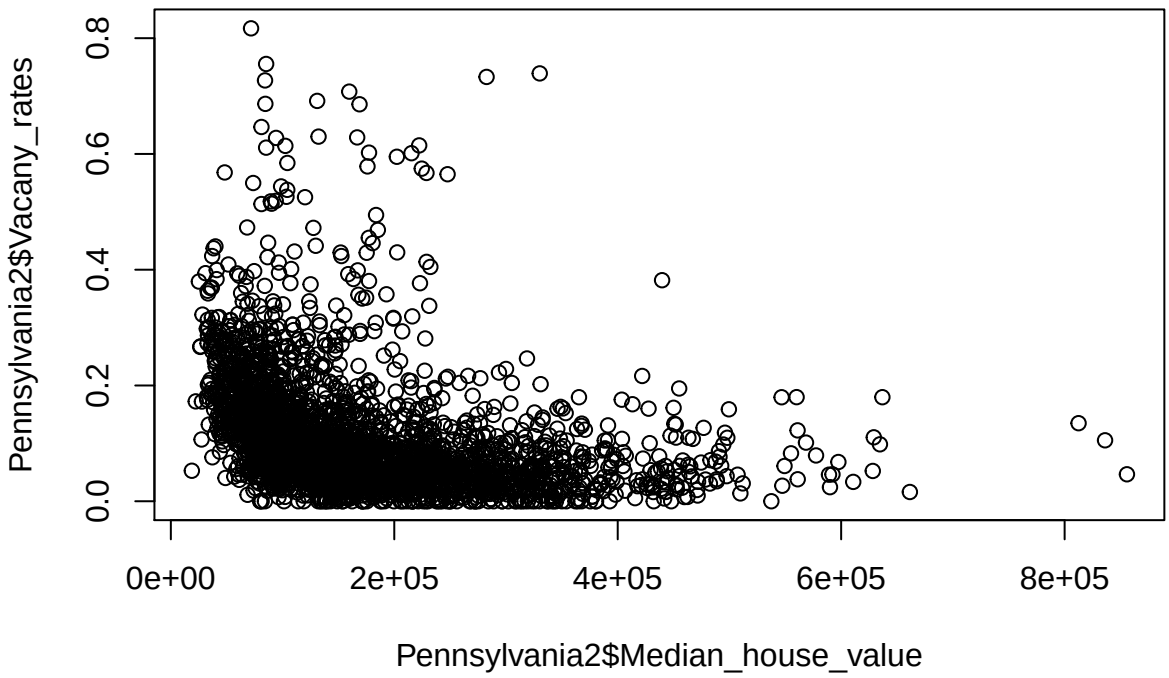


c. Plot vacancy rate against median house value separately for California and for Pennsylvania. Is there a difference?

```
S6index2<-which(ca_panew$STATEFP==6)
S42index2<-which(ca_panew$STATEFP==42)
California2<-ca_panew[S6index2,]
Pennsylvania2<-ca_panew[S42index2,]
plot(California2$Vacany_rates~California2$Median_house_value)
```




```
plot(Pennsylvania2$Vacany_rates~Pennsylvania2$Median_house_value)
```



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. Answer for a:the block of code intends to calculate the median of total units for three given counties.And the code for Alameda county is given below.It uses a loop to do that. 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.

```
median(ca_panew[which(ca_panew$STATEFP==6&ca_panew$COUNTYFP==1),10])
```

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

```
c. For Alameda, Santa Clara and Allegheny Counties, what were the average percentages of housing built s
```

```
 #ALameda County
mean(ca_panew[which(ca_panew$STATEFP==6&ca_panew$COUNTYFP==1),15])
```

```
## [1] 2.737018
```

```
#Santa Clara
mean(ca_panew[which(ca_panew$STATEFP==6&ca_panew$COUNTYFP==85),15])
```

```
## [1] 3.115112
```

```r
#Allegheny Counties
mean(ca_panew[which(ca_panew$STATEFP==42&ca_panew$COUNTYFP==3),15])
```

```
## [1] 2.012448
```

d. The ‘cor‘ function calculates the correlation coefficient between two variables. What is the correla

```r
#(i)the whole data
cor(ca_panew$Median_house_value,ca_panew$Built_2005_or_later)
```

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

```r
#(ii)all of California
cor(California2$Median_house_value,California2$Built_2005_or_later)
```

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

```r
#(iii) all of Pennsylvania
cor(Pennsylvania2$Median_house_value,Pennsylvania2$Built_2005_or_later)
```

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

```r
#(iv) Alameda County
Aindex1<-which(California2$COUNTYFP==1)
Sindex1<-which(Pennsylvania2$COUNTYFP==3)
Aindex2<-which(California2$COUNTYFP==85)
Alameda<-California2[Aindex1,]
cor(Alameda$Median_house_value,Alameda$Built_2005_or_later)
```

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

```r
#(v) Santa Clara County
Santa<-Pennsylvania2[Sindex1,]
cor(Santa$Median_house_value,Santa$Built_2005_or_later)
```

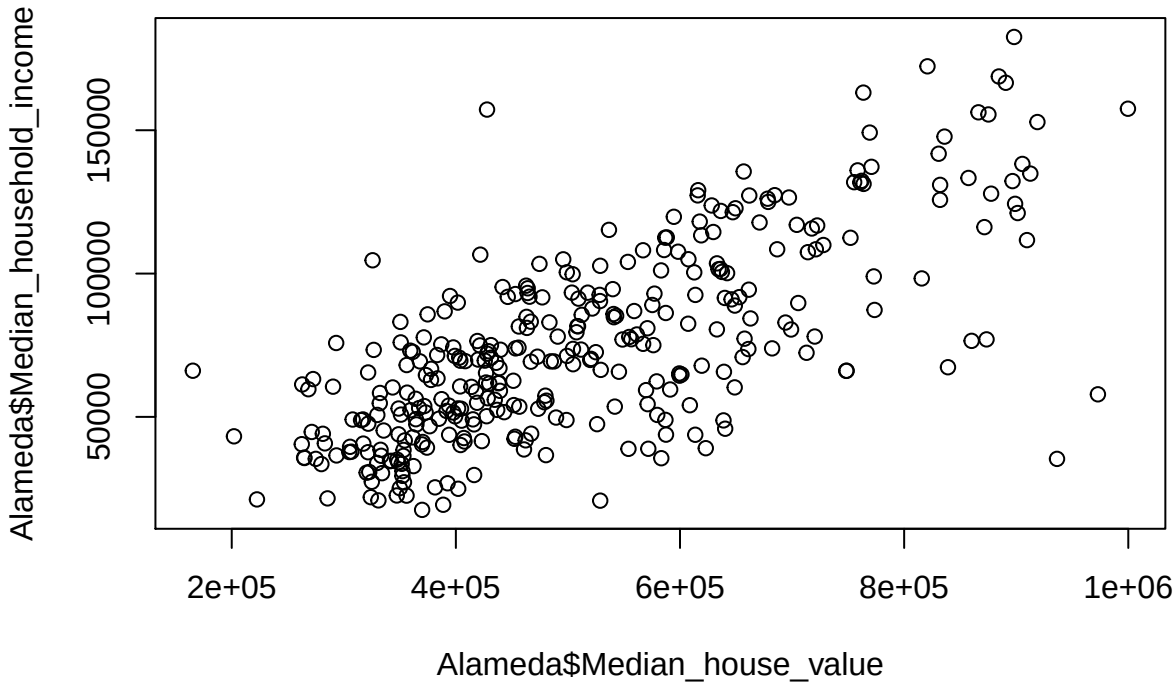
```
## [1] 0.1939652
```

```r
#(vi) Allegheny County
Allegheny<-California2[Aindex2,]
cor(Allegheny$Median_house_value,Allegheny$Built_2005_or_later)
```

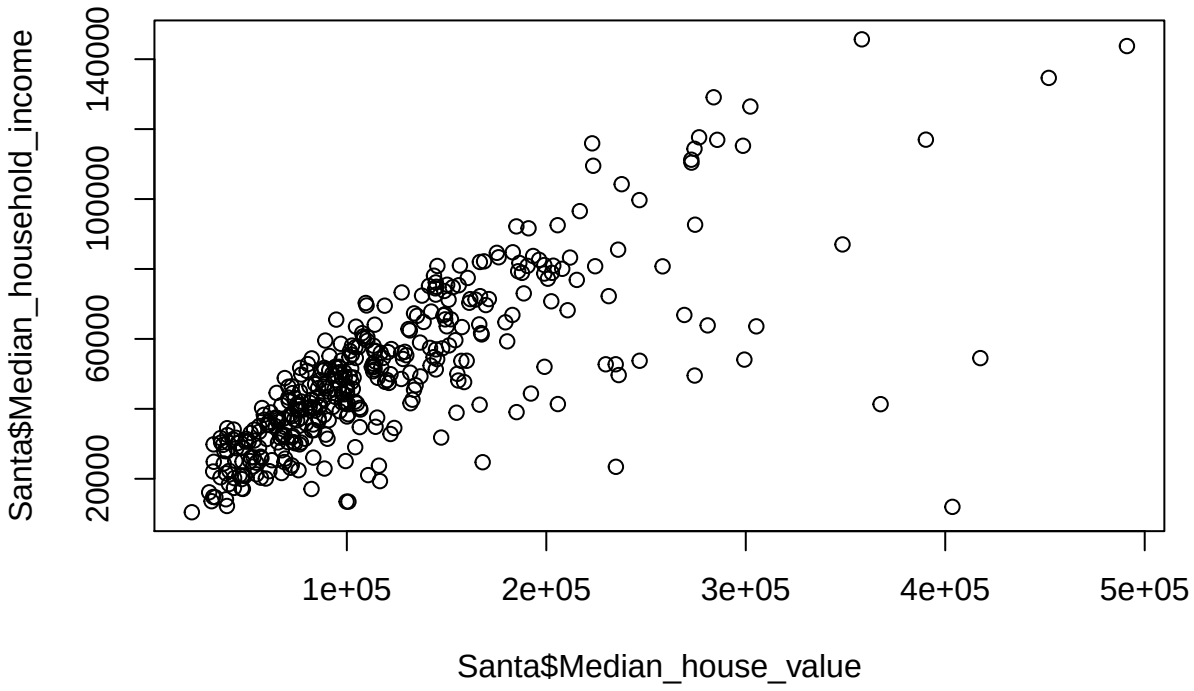
```
## [1] -0.1726203
```

e. Make three plots, showing median house values against median income, for Alameda, Santa Clara, and Al

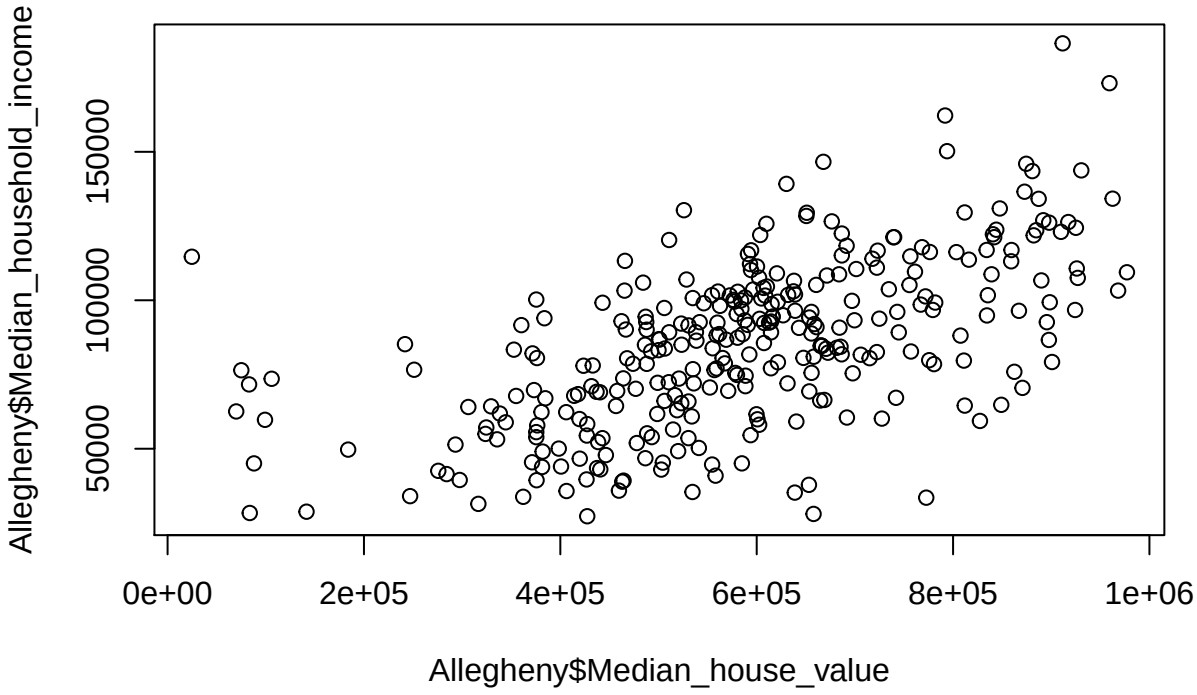
```
plot(Alameda$Median_house_value,Alameda$Median_household_income)
```




```
plot(Santa$Median_house_value,Santa$Median_household_income)
```




```
plot(Allegheny$Median_house_value,Allegheny$Median_household_income)
```




```
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)
    }
  }
}
accamhv <- c()
for (tract in acca) {
  accamhv <- c(accamhv, ca_pa[tract,10])
}
median(accamhv)
```

MB.Ch1.11. Run the following code:

```
gender <- factor(c(rep("female", 91), rep("male", 92)))
table(gender)
```

```
## gender
## female   male
##     91     92
```

```r
gender <- factor(gender, levels=c("male", "female"))
table(gender)
```

```
## gender
##   male female
##     92     91
```

```r
gender <- factor(gender, levels=c("Male", "female"))
# Note the mistake: "Male" should be "male"
table(gender)
```

```
## gender
##   Male female
##      0     91
```

```r
table(gender, exclude=NULL)
```

```
## gender
##   Male female   <NA>
##      0     91     92
```

```r
rm(gender)  # Remove gender
```

Explain the output from the successive uses of table().

MB.Ch1.12. Write a function that calculates the proportion of values in a vector x that exceed some value cutoff.

```r
exceedcutoff<-function(x,value){
  d<-length(x)
  k<-0
  for(i in 1:d){
    if(x[i]>value) k=k+1
  }
    ratio<-k/d
    ratio
}
```

(a) Use the sequence of numbers 1, 2, . . . , 100 to check that this function gives the result that is expected.

```r
#we set value to 25
exceedcutoff(c(1:100),25)
```

```
## [1] 0.75
```

```r
#we set value to 50
exceedcutoff(c(1:100),50)
```

```
## [1] 0.5
```

(b) Obtain the vector ex01.36 from the Devore6 (or Devore7) package. These data give the times required for individuals to escape from an oil platform during a drill. Use dotplot() to show the distribution of times. Calculate the proportion of escape times that exceed 7 minutes.

MB.Ch1.18. The Rabbit data frame in the MASS library contains blood pressure change measurements on five rabbits (labeled as R1, R2, . . . ,R5) under various control and treatment conditions. Read the help file for more information. Use the unstack() function (three times) to convert Rabbit to the following form:

Treatment Dose R1 R2 R3 R4 R5

1 Control 6.25 0.50 1.00 0.75 1.25 1.5

2 Control 12.50 4.50 1.25 3.00 1.50 1.5

....

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:DAAG':
##
##     hills
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
unstack(Rabbit,BPchange~Animal)
```

```
##       R1    R2    R3    R4   R5
## 1   0.50  1.00  0.75  1.25  1.5
## 2   4.50  1.25  3.00  1.50  1.5
## 3  10.00  4.00  3.00  6.00  5.0
## 4  26.00 12.00 14.00 19.00 16.0
## 5  37.00 27.00 22.00 33.00 20.0
## 6  32.00 29.00 24.00 33.00 18.0
## 7   1.25  1.40  0.75  2.60  2.4
## 8   0.75  1.70  2.30  1.20  2.5
## 9   4.00  1.00  3.00  2.00  1.5
## 10  9.00  2.00  5.00  3.00  2.0
## 11 25.00 15.00 26.00 11.00  9.0
## 12 37.00 28.00 25.00 22.00 19.0
```