STAT 206 Homework 2_Lihua_Xu

Due Monday, October 16, 5:00 PM

General instructions for homework: Homework must be completed as an R Markdown file. Be sure to include your name in the file. Give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. (Examining your various objects in the "Environment" section of RStudio is insufficient – you must use scripted commands.)

The data set at [http://www.stat.cmu.edu/~cshalizi/uADA/13/hw/01/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`.

```
loading_data<-read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/13/hw/01/calif_penn_2011.csv")
ca_pa <- data.frame(loading_data)
```

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

```
dim.data.frame((ca_pa))
```

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

```
#There are 11275 rows and 34 columns.
```

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

```
colSums(apply(ca_pa,c(1,2),is.na))
```

```
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
```

```
#Find the number of value "NA" for each clomns.
```

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

e. How many rows did this eliminate?

```
dim.data.frame(ca_pa_new)
```

```
## [1] 10605    34
```

```
#There are 10605 rows.
```

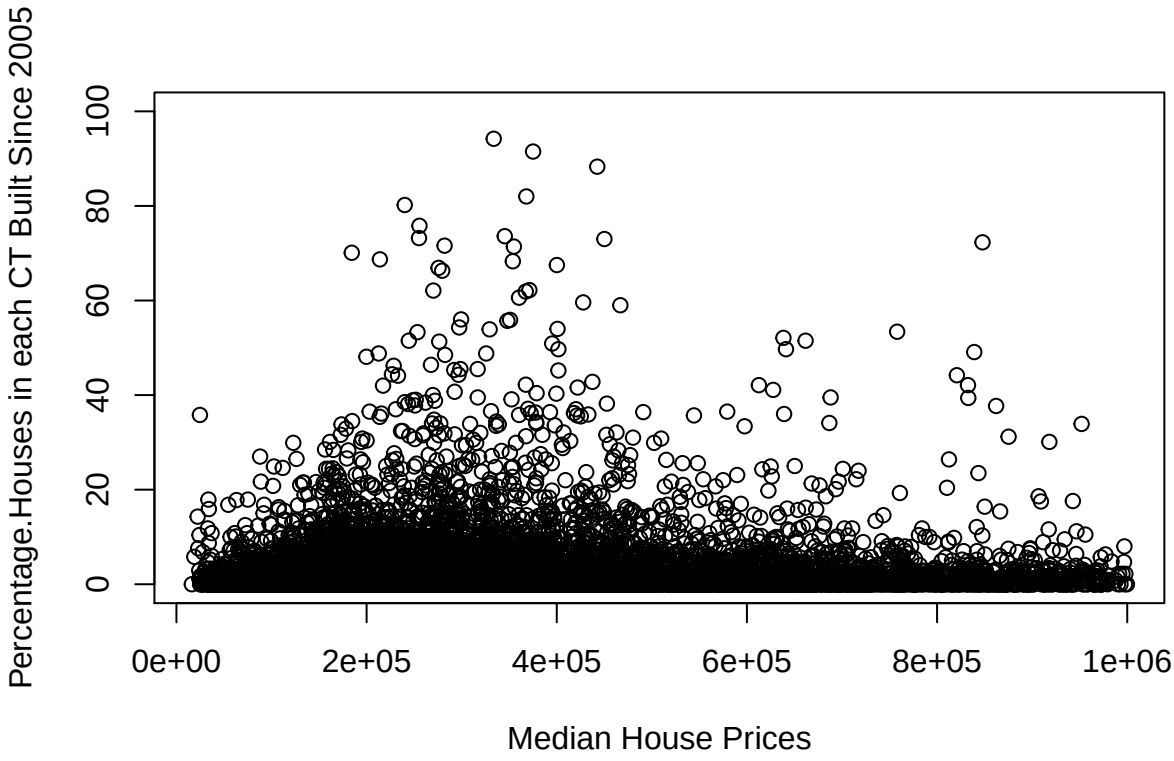
f. Are your answers in (c) and (e) compatible? Explain.

```
#Yes,they are compatible.
#From the problem (c), the number of rows which contain
#at leat 1 NA is larger than the value 599.
#From the problem (e), omitting any row containing an NA value,
#the row number become 10605. Compared with the original 11275 rows,
#the differences is 670. 670 is larger than 599.
#This means the data produced by (c) and (e) are both resonable.
```

2. *This Very New House*

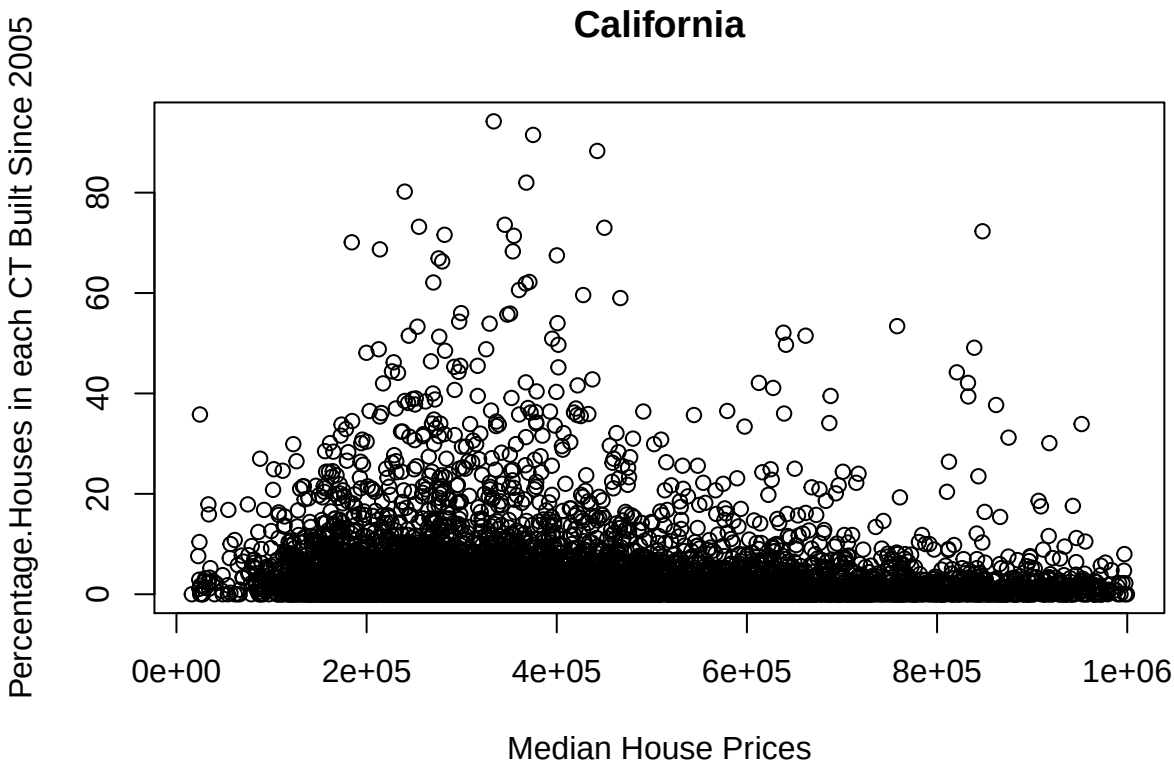
a. The variable `Built_2005_or_later` indicates the percentage of houses in each Census tract built since 2005. Plot median house prices against this variable.

```
plot(ca_pa$Median_house_value,ca_pa$Built_2005_or_later,
     xlab="Median House Prices",
     ylab="Percentage.Houses in each CT Built Since 2005")
```

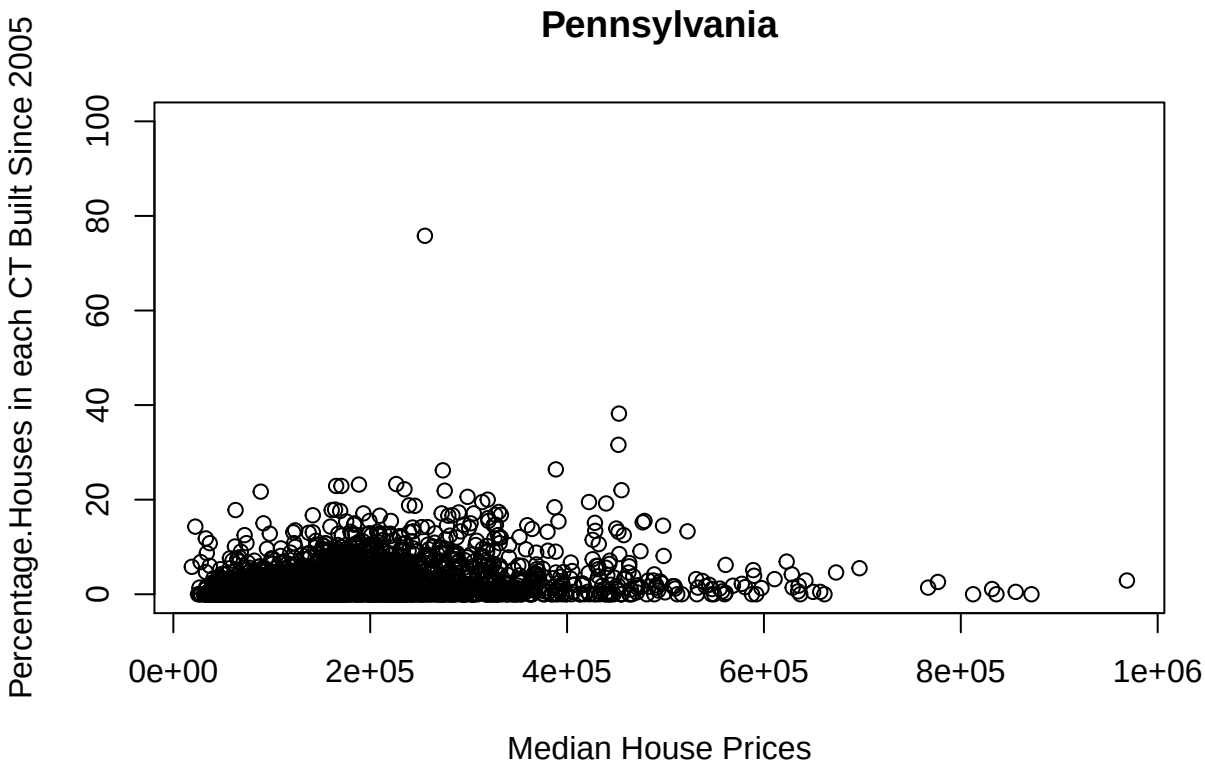


b. Make a new plot, or pair of plots, which breaks this out by state. Note that the state is recorded in the `STATEFP` variable, with California being state 6 and Pennsylvania state 42.

```
plot(ca_pa[ca_pa$STATEFP==6,]$Median_house_value,
     ca_pa[ca_pa$STATEFP==6,]$Built_2005_or_later,
     xlab="Median House Prices",
     ylab="Percentage.Houses in each CT Built Since 2005",
     main="California")
```




```
plot(ca_pa[ca_pa$STATEFP==42,]$Median_house_value,
     ca_pa[ca_pa$STATEFP==42,]$Built_2005_or_later,
     xlab="Median House Prices",
     ylab="Percentage.Houses in each CT Built Since 2005",
     main="Pennsylvania")
```



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?

```
a <- ca_pa$Vacant_units
b <- ca_pa[ca_pa$Total_units!=0,]$Total_units
vacancy_rate_included<-transform(ca_pa,vacancy_rate=a/b)
```

```
## Warning in a/b: longer object length is not a multiple of shorter object
## length
```

```
minimum_vacancy_rate <- min(vacancy_rate_included$vacancy_rate)
minimum_vacancy_rate
```

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

```
#The minimum vacancy rates is 0.
maximum_vacancy_rate <- max(vacancy_rate_included$vacancy_rate)
maximum_vacancy_rate
```

```
## [1] 62.75
```

```
#The maximum vacancy rates is 62.75.
mean_vacancy_rate <- mean(vacancy_rate_included$vacancy_rate)
```

```
mean_vacancy_rate
```

```
## [1] 0.1482118
```

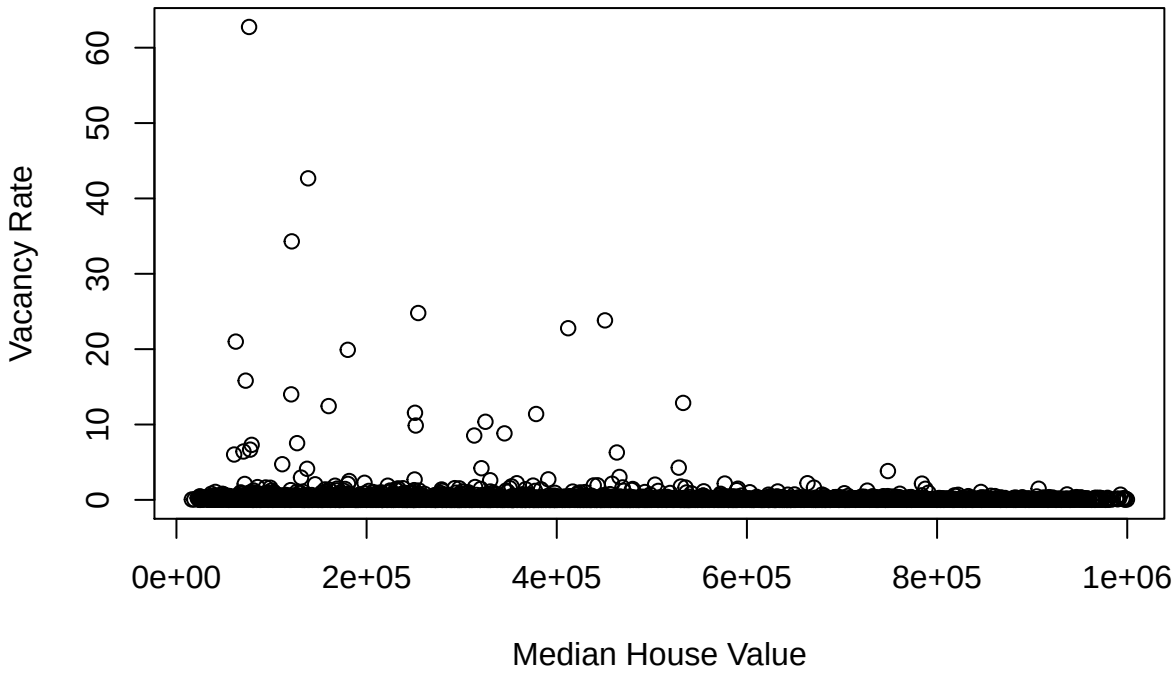
```
#The mean vacancy rates is 0.1482118.
median_vacancy_rate <- median(vacancy_rate_included$vacancy_rate)
median_vacancy_rate
```

```
## [1] 0.06588855
```

```
#The median vacancy rates is 0.06588855.
```

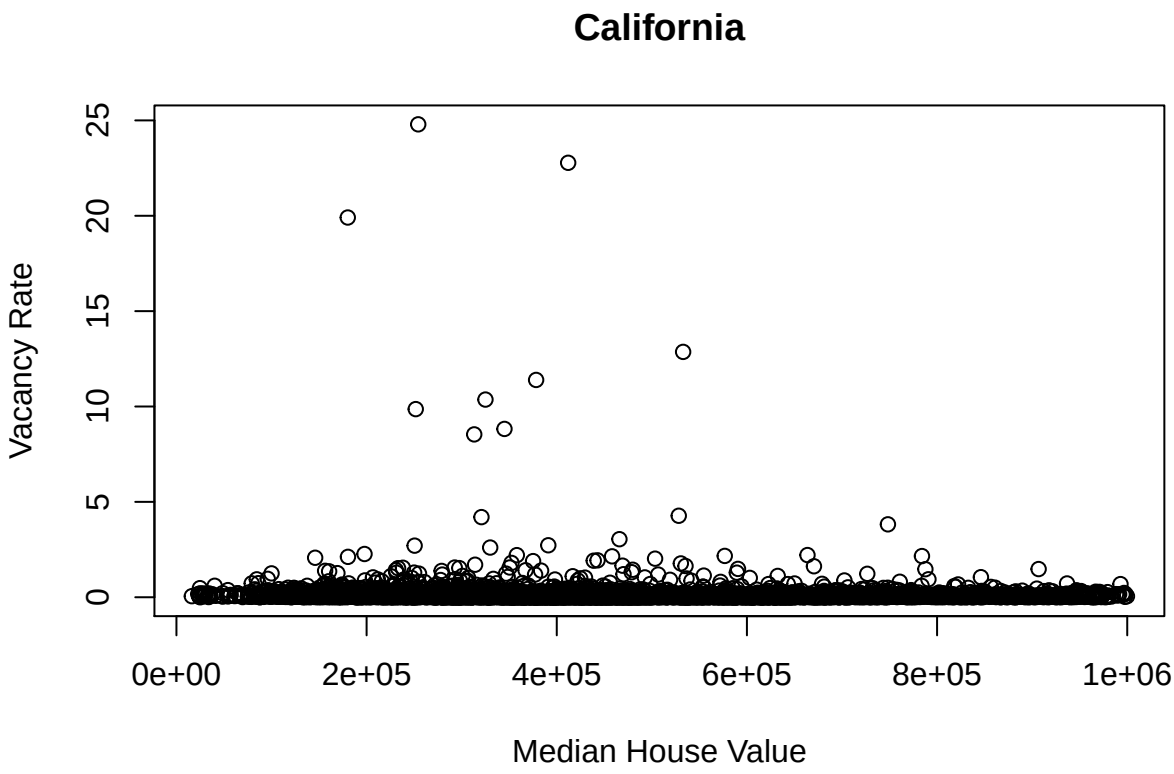
b. Plot the vacancy rate against median house value.

```
plot(vacancy_rate_included$Median_house_value,vacancy_rate_included$vacancy_rate,
     xlab="Median House Value",
     ylab="Vacancy Rate")
```

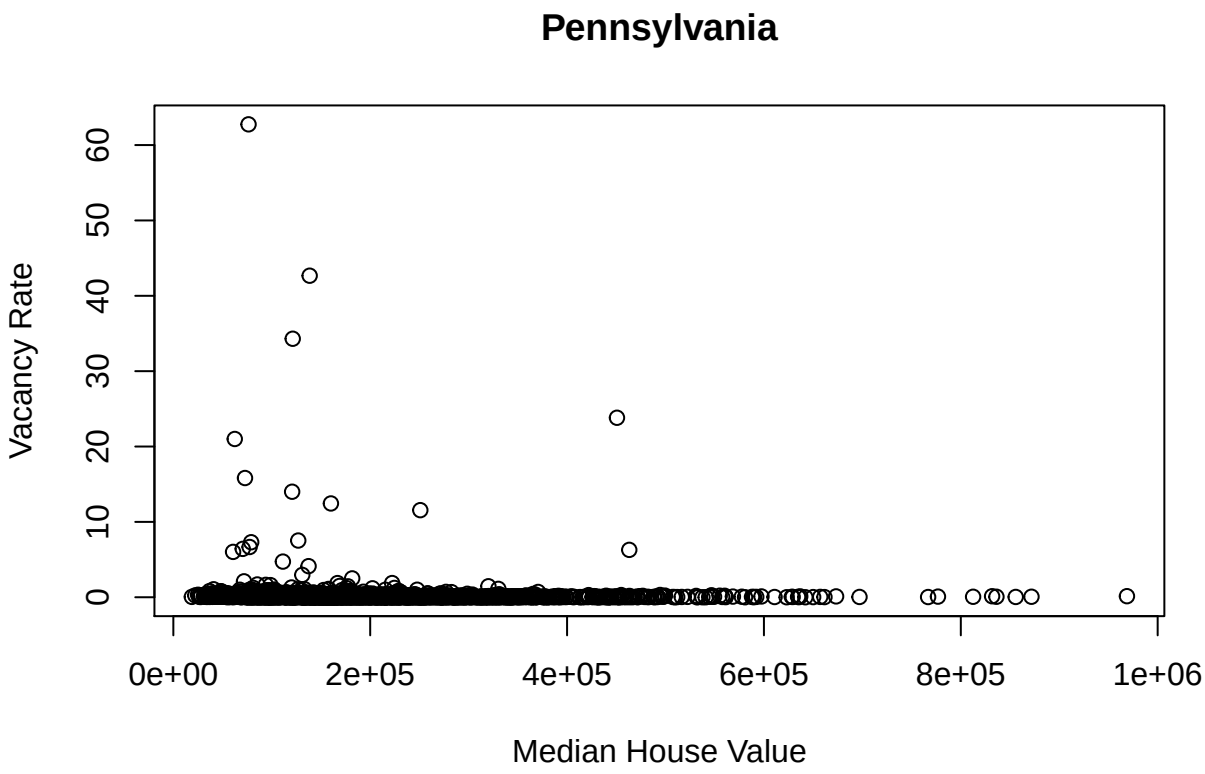


c. Plot vacancy rate against median house value separately for California and for Pennsylvania. Is there a difference?

```
plot(vacancy_rate_included[vacancy_rate_included$STATEFP==6,]$Median_house_value,
     vacancy_rate_included[vacancy_rate_included$STATEFP==6,]$vacancy_rate,
     xlab="Median House Value",
     ylab="Vacancy Rate",
     main="California")
```




```
plot(vacancy_rate_included[vacancy_rate_included$STATEFP==42,]$Median_house_value,
     vacancy_rate_included[vacancy_rate_included$STATEFP==42,]$vacancy_rate,
     xlab="Median House Value",
     ylab="Vacancy Rate",
     main="Pennsylvania")
```




```
#For Calfornia, the distribution of the vacancy rate is homogeneous
#among all different Median House Value.
#For Pennsylvania, the distribution of the vacancy rate is not that homogeneous.
#Especially for the higher Median House Value, there is almost no vacancy rate.
```

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.

```
#The block of code at the end of this question is supposed to obtain the
#each row numbers that comtain the information for the Alameda County in California.

#First, acca was consigned with a NULL value.
#Second, doing a "for" loop (which go through every row in the dataframe)
#in order to find the rows with information for California (STATEFP[tract]==6).
#If it doesn't find the information for California, it will directly go to the next loop.
#Once it find the California information, using "if" to judge
#whether this row contain the information for Alameda County (COUNTYFP[tract]==1).
#If it contain no information regarding Alameda County, a new cycle will start.
#But if this row contain the information regarding Alameda County,
#the number for this row would be recorded in the vector "acca".
#At the end first section, we will get a vector "acca",
#which contains all the row numbers comtaining the information related
#to the Alameda County in California.
```

```
#For the nect section, using the row numbers containing in the vector "acca" to
#extract the median house value for each row.
#And at the end, we will get a median number for this series of the median house values.
```

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.

```
#The code at the end of the section gives the result "NA".
#So I think we need to substitute the missing value as 0
ca_pa[is.na(ca_pa)]<-0
#So the one line code should be as following.
median(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==1,]$Median_house_value)
```

```
## [1] 262700
```

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

```
mean(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==1,]$Built_2005_or_later)
```

```
## [1] 2.386565
```

```
#The average percentages of housing built since 2005 in Alameda, California is 2.386565.
mean(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==85,]$Built_2005_or_later)
```

```
## [1] 3.160215
```

```
#The average percentages of housing built since 2005 in Santa Clara, California is 3.160215.
mean(ca_pa[ca_pa[ca_pa$STATEFP==42,]$COUNTYFP==3,]$Built_2005_or_later)
```

```
## [1] 3.326057
```

```
#The average percentages of housing built since 2005 in Allegheny, Pennsylvania is 3.326057.
```

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?

```
cor(ca_pa$Median_house_value,ca_pa$Built_2005_or_later)
```

```
## [1] -0.01183982
```

```
#The correlation between median house value and the percent of housing built
#since 2005 in the whole data is -0.01183982.
cor(ca_pa[ca_pa$STATEFP==6,]$Median_house_value,
    ca_pa[ca_pa$STATEFP==6,]$Built_2005_or_later)
```

```
## [1] -0.08323476
```

```
#The correlation between median house value and the percent of housing built
#since 2005 in all of California is -0.08323476.
cor(ca_pa[ca_pa$STATEFP==42,]$Median_house_value,
    ca_pa[ca_pa$STATEFP==42,]$Built_2005_or_later)
```

```
## [1] 0.2023056
```

```
#The correlation between median house value and the percent of housing built
#since 2005 in all of Pennsylvania is 0.2023056.
cor(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==1,]$Median_house_value,
    ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==1,]$Built_2005_or_later)
```

```
## [1] 0.08096986
```

```
#The correlation between median house value and the percent of housing built
#since 2005 in Alameda County is  0.08096986.
cor(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==85,]$Median_house_value,
    ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==85,]$Built_2005_or_later)
```

```
## [1] -0.08851101
```

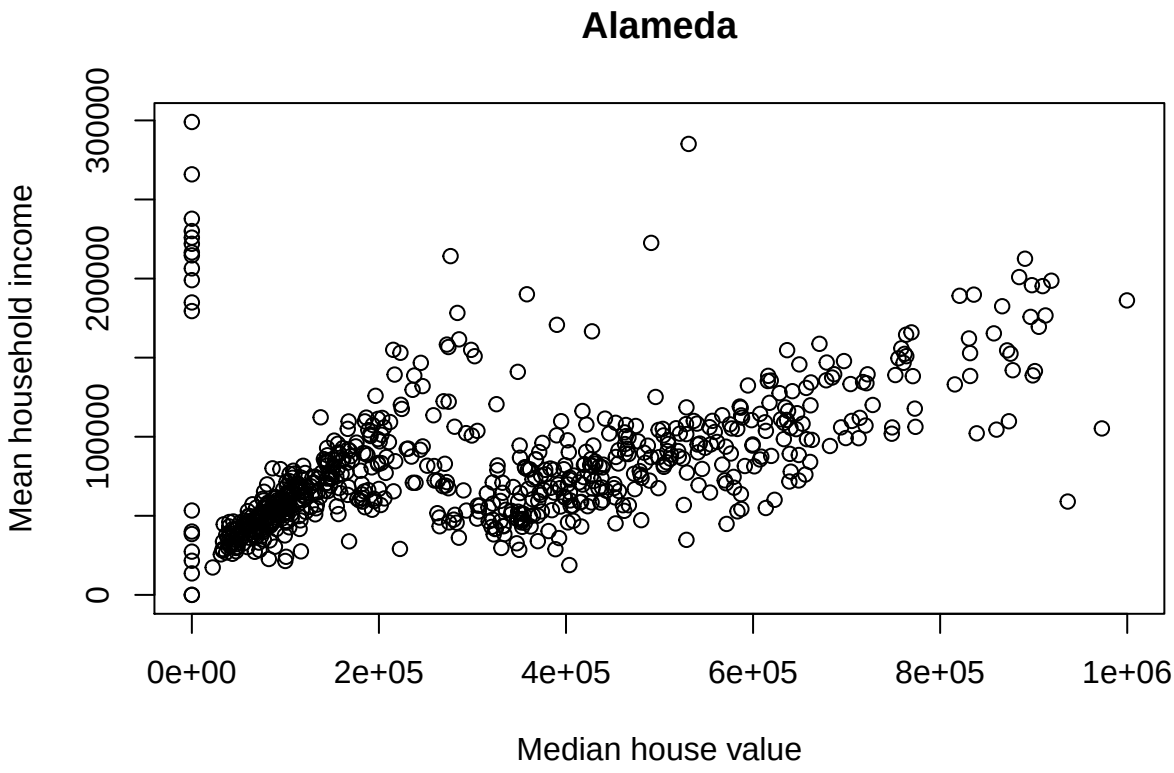
```
#The correlation between median house value and the percent of housing built
#since 2005 in Santa Clara County is -0.08851101.
cor(ca_pa[ca_pa[ca_pa$STATEFP==42,]$COUNTYFP==3,]$Median_house_value,
    ca_pa[ca_pa[ca_pa$STATEFP==42,]$COUNTYFP==3,]$Built_2005_or_later)
```

```
## [1] -0.01480825
```

```
#The correlation between median house value and the percent of housing built
#since 2005 in Allegheny County is -0.01480825.
```

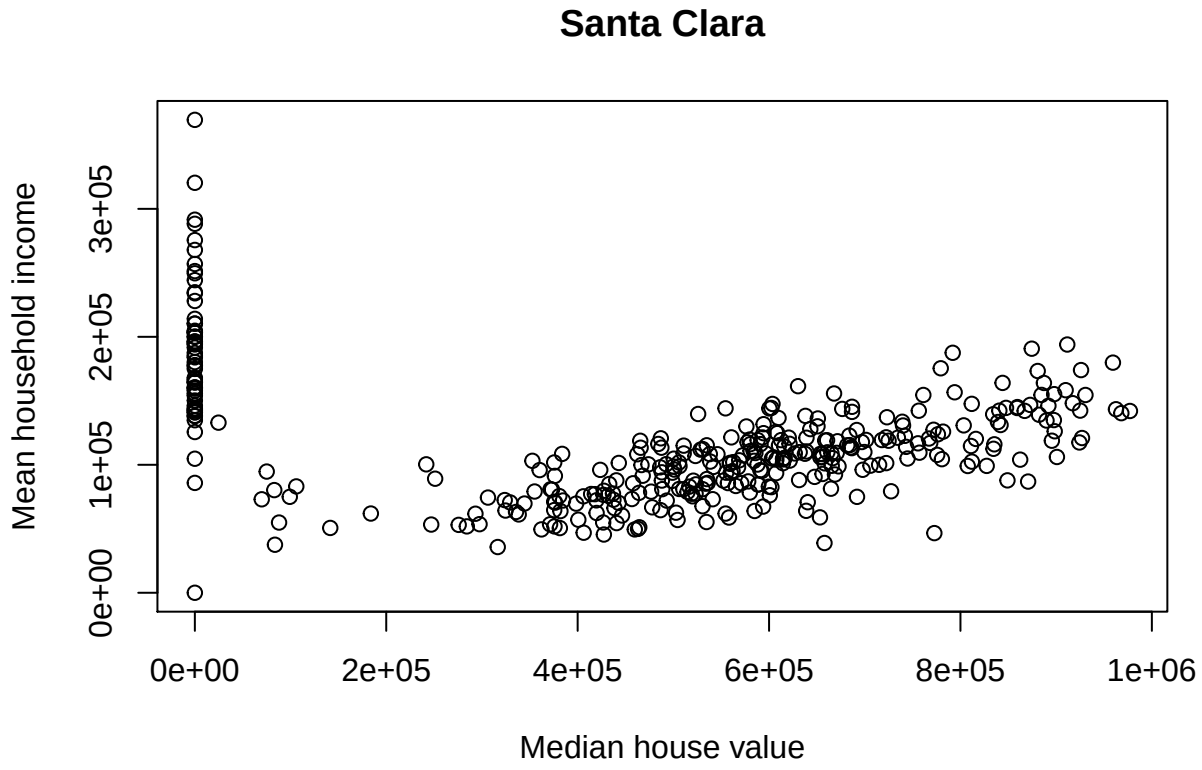
e. Make three plots, showing median house values against median income, for Alameda, Santa Clara, and Allegheny Counties. (If you can fit the information into one plot, clearly distinguishing the three counties, that's OK too.)

```
plot(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==1,]$Median_house_value,
     ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==1,]$Mean_household_income,
     xlab = "Median house value",ylab = "Mean household income",main="Alameda")
```

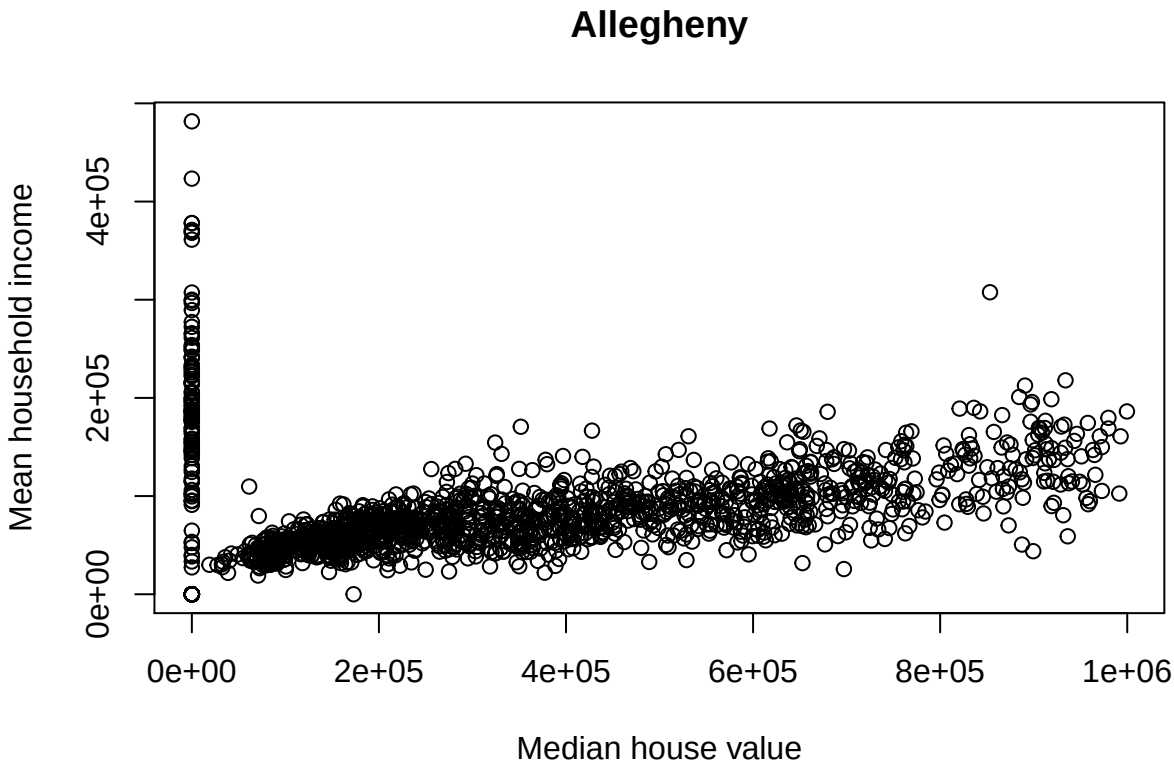



```
plot(ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==85,]$Median_house_value,
     ca_pa[ca_pa[ca_pa$STATEFP==6,]$COUNTYFP==85,]$Mean_household_income,
```

```
     xlab = "Median house value",ylab = "Mean household income",main="Santa Clara")
```




```
plot(ca_pa[ca_pa[ca_pa$STATEFP==42,]$COUNTYFP==3,]$Median_house_value,
     ca_pa[ca_pa[ca_pa$STATEFP==42,]$COUNTYFP==3,]$Mean_household_income,
     xlab = "Median house value",ylab = "Mean household income",main="Allegheny")
```




```
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)
```