Shahab Geravesh, Statistical Computing 206 (002), Homework 2

a. Load the data into a dataframe calledca_pa Loading the dataset.

```
ca_pa<-read.csv("http://www.stat.cmu.edu/~cshalizi/uADA/13/hw/01/calif_penn_2011.csv", header=TRUE)
```

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

```
nrow(ca_pa)
```

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

The dataset has 11275 rows

```
ncol(ca_pa)
```

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

The dataset has 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
```

By using apply function, it loops through ca_pa and returns the true nulls and it shows the number of nulls in each columns

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

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

Eliminating the rows with null values

e. How many rows did this eliminate?

```
nrow(ca_pa) -nrow(na_Omit)
```

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

It eliminated 670 rows

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

```
colSums(apply(na_Omit,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                           0
##                Total_units                Vacant_units
##                          0                           0
##               Median_rooms  Mean_household_size_owners
##                          0                           0
## Mean_household_size_renters         Built_2005_or_later
##                          0                           0
##          Built_2000_to_2004                 Built_1990s
##                          0                           0
##                 Built_1980s                 Built_1970s
##                          0                           0
##                 Built_1960s                 Built_1950s
##                          0                           0
##                 Built_1940s       Built_1939_or_earlier
##                          0                           0
##                  Bedrooms_0                  Bedrooms_1
##                          0                           0
##                  Bedrooms_2                  Bedrooms_3
##                          0                           0
##                  Bedrooms_4          Bedrooms_5_or_more
##                          0                           0
##                      Owners                     Renters
```

```
##                             0                             0
##     Median_household_income       Mean_household_income
##                             0                             0
```

By running this command for na_Omit, it shows the missing values are now zero and it means that we were able to eliminate the null values from the dataset successfully.

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.

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

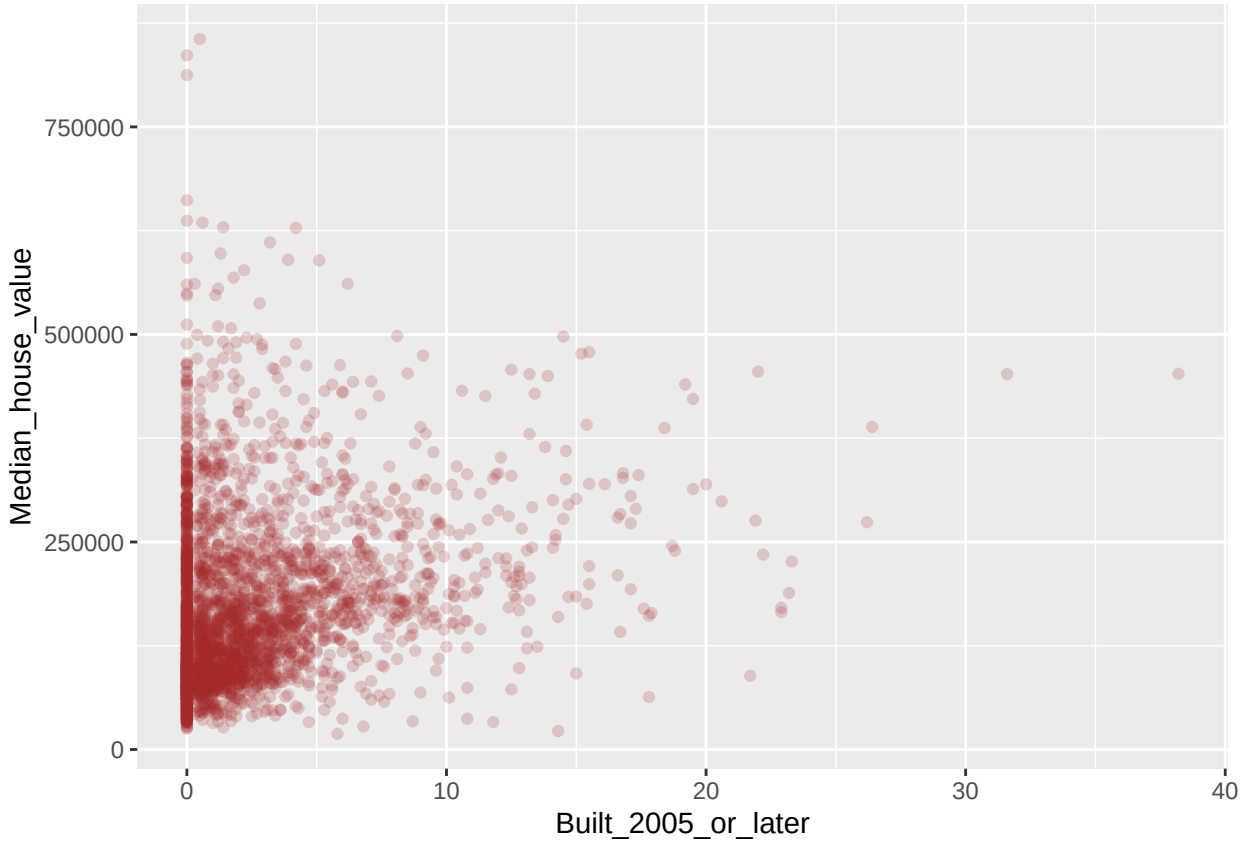
```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

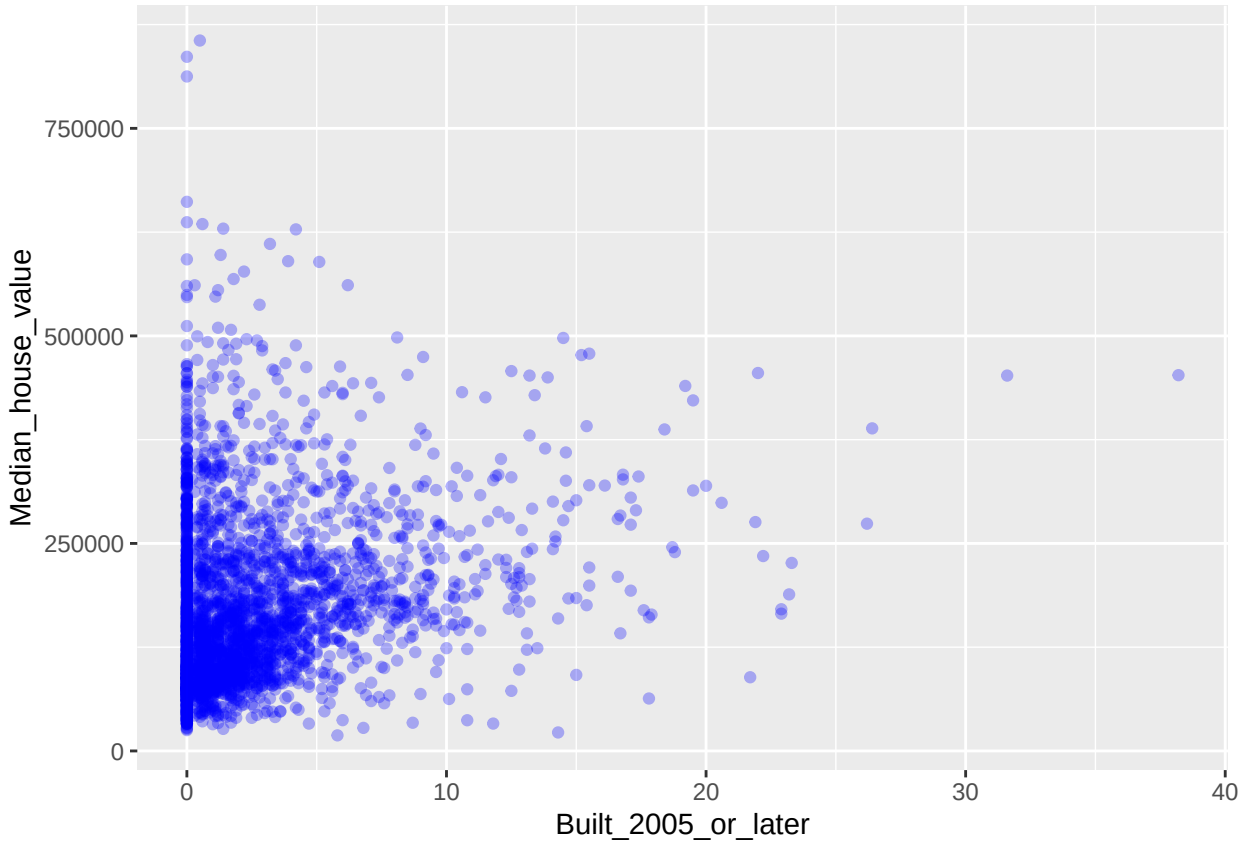
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
na_Omit%>% filter(STATEFP != 6)  %>% ggplot(aes(Built_2005_or_later, Median_house_value)) + geom_point(a
```



```
na_Omit %>% filter(STATEFP != 6)  %>% ggplot(aes(Built_2005_or_later, Median_house_value)) + geom_point
```



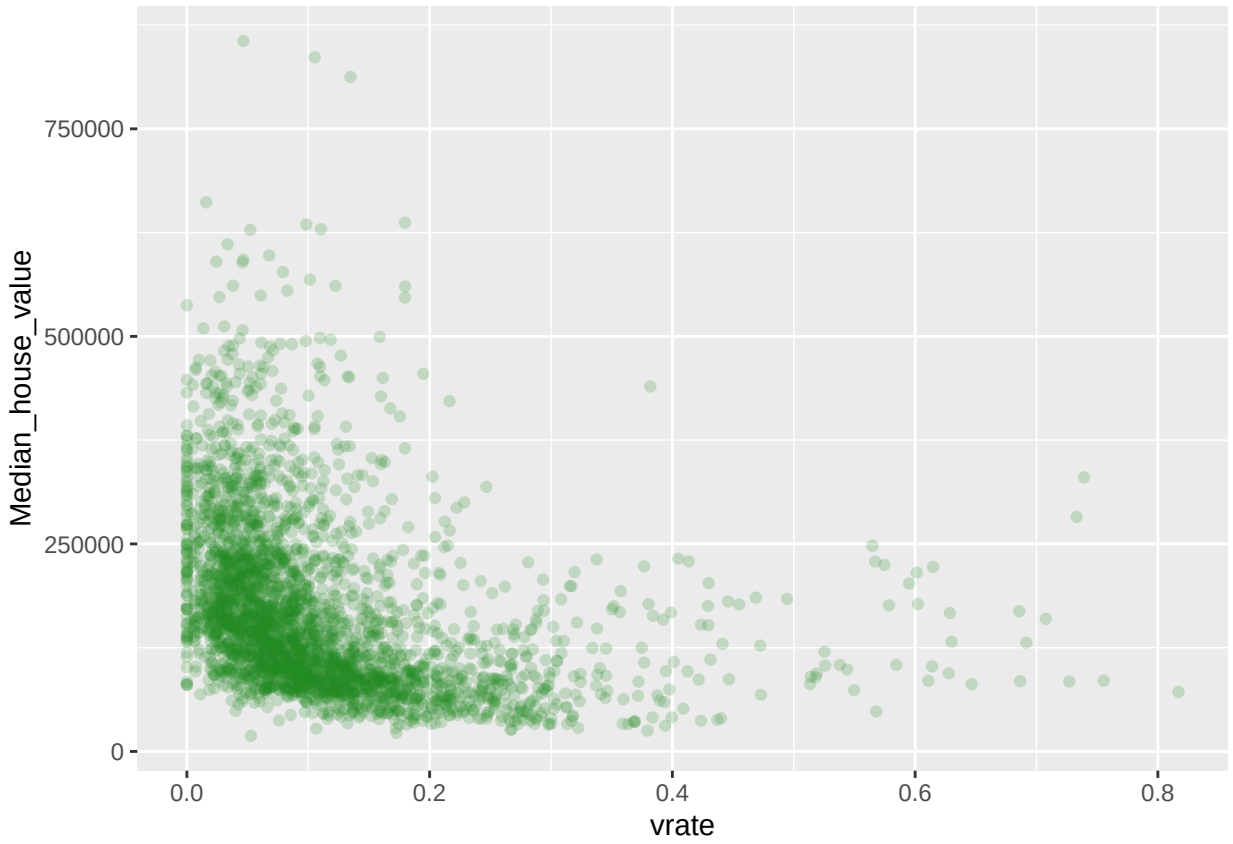
3. Nobody Home The vacancy rate is the fraction of housing units which are not occupied. The dataframe contains columns giving the total number of housing units for each Census tract, and the number of vacant housing units. a. Add a new column to the dataframe which contains the vacancy rate. What are the minimum, maximum, mean, and median vacancy rates?

```
na_Omit$vrate <- na_Omit$Vacant_units/na_Omit$Total_units
summary(na_Omit$vrate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03846 0.06767 0.08889 0.10921 0.96531
```

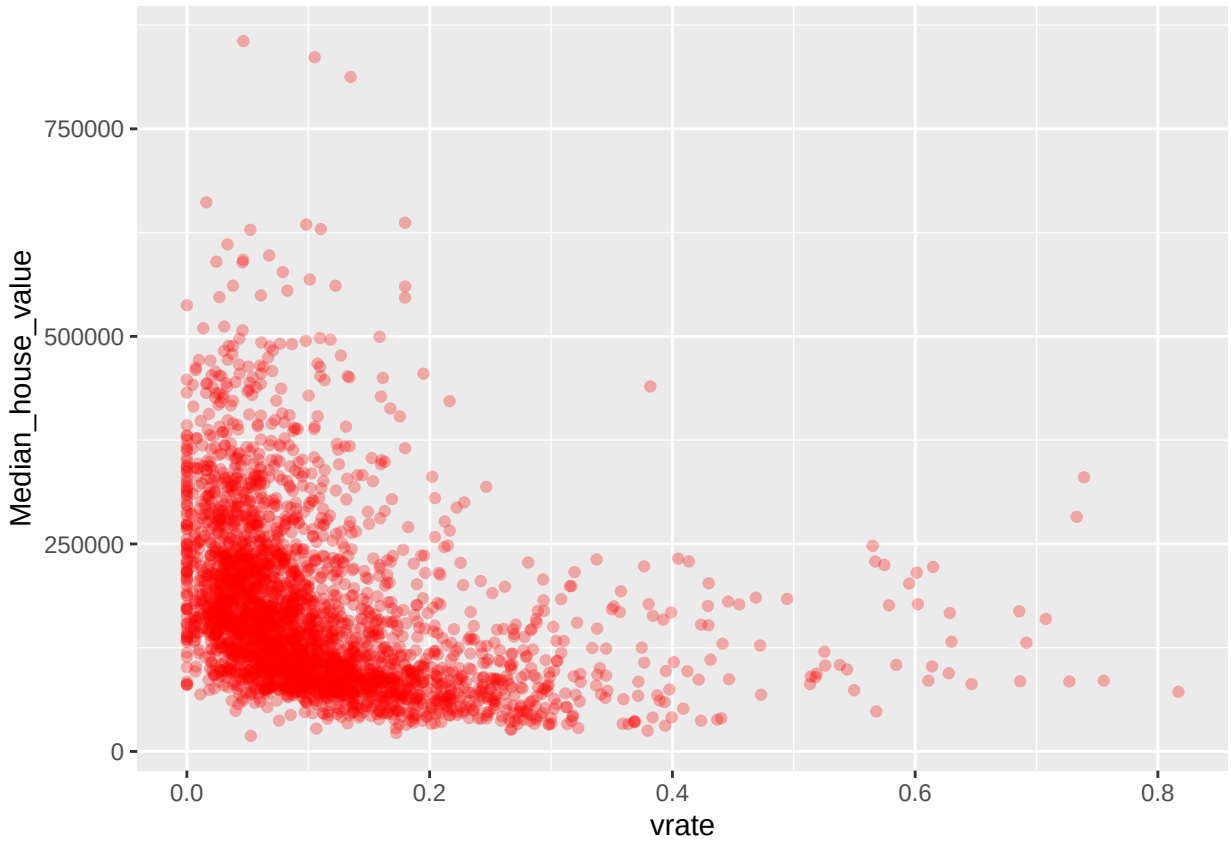
b. Plot the vacancy rate against median house value.

```
na_Omit %>% filter(STATEFP != 6)  %>% ggplot(aes(vrate, Median_house_value)) + geom_point(alpha=.2, col
```



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

```
na_Omit%>% filter(STATEFP != 6)  %>% ggplot(aes(vrate, Median_house_value)) + geom_point(alpha=.3, col=
```



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.

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(na_Omit[na_Omit$STATEFP == 6 && na_Omit$COUNTYFP == 1,]$Median_house_value)
```

```
## [1] 311100
```

It is supposed to calculate the median of track median house in Alameda county.

```
alameda.avg <- na_Omit$Built_2005_or_later[na_Omit$STATEFP == 6 &na_Omit$COUNTYFP == 1]
mean(alameda.avg)
```

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

```
santaclara.avg <- na_Omit$Built_2005_or_later[na_Omit$STATEFP == 6 &na_Omit$COUNTYFP == 85]
mean(santaclara.avg)
```

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

```
allegheny.avg <- na_Omit$Built_2005_or_later[na_Omit$STATEFP == 42 &na_Omit$COUNTYFP == 3]
mean(allegheny.avg)
```

```
## [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?

```
cor.wholedata <-cor(na_Omit$Median_house_value, na_Omit$Built_2005_or_later)
cor.wholedata
```

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

```
cor.california <-cor(na_Omit$Median_house_value[na_Omit$STATEFP == 6],na_Omit$Built_2005_or_later[na_Omi
cor.california
```

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

```
cor.pennsylvania <-cor(na_Omit$Median_house_value[na_Omit$STATEFP == 42],na_Omit$Built_2005_or_later[na_
cor.pennsylvania
```

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

```
cor.alameda <-cor(na_Omit$Median_house_value[na_Omit$STATEFP == 6 &na_Omit$COUNTYFP == 1],na_Omit$Built_
cor.alameda
```

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

```
cor.santaclara <-cor(na_Omit$Median_house_value[na_Omit$STATEFP == 6 &na_Omit$COUNTYFP == 85],na_Omit$Bu
cor.santaclara
```

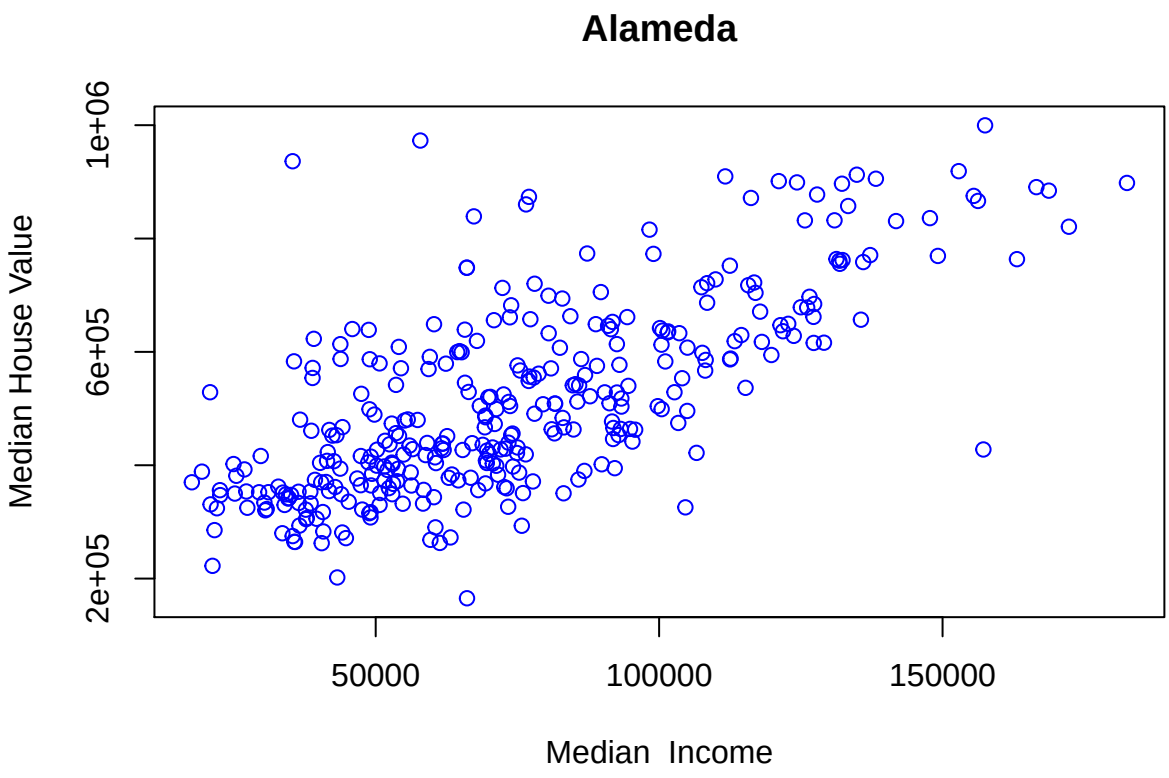
```
## [1] -0.1726203
```

```
cor.allegheny <-cor(na_Omit$Median_house_value[na_Omit$STATEFP == 42 &na_Omit$COUNTYFP == 3],na_Omit$Bui
cor.allegheny
```

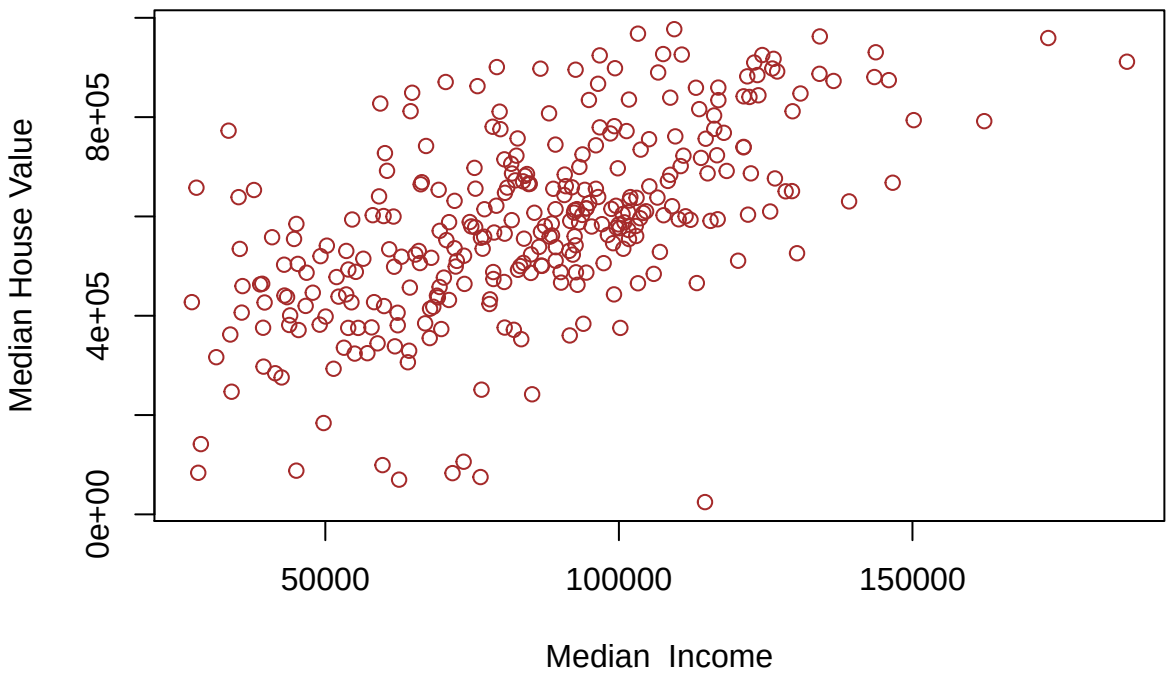
```
## [1] 0.1939652
```

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.)

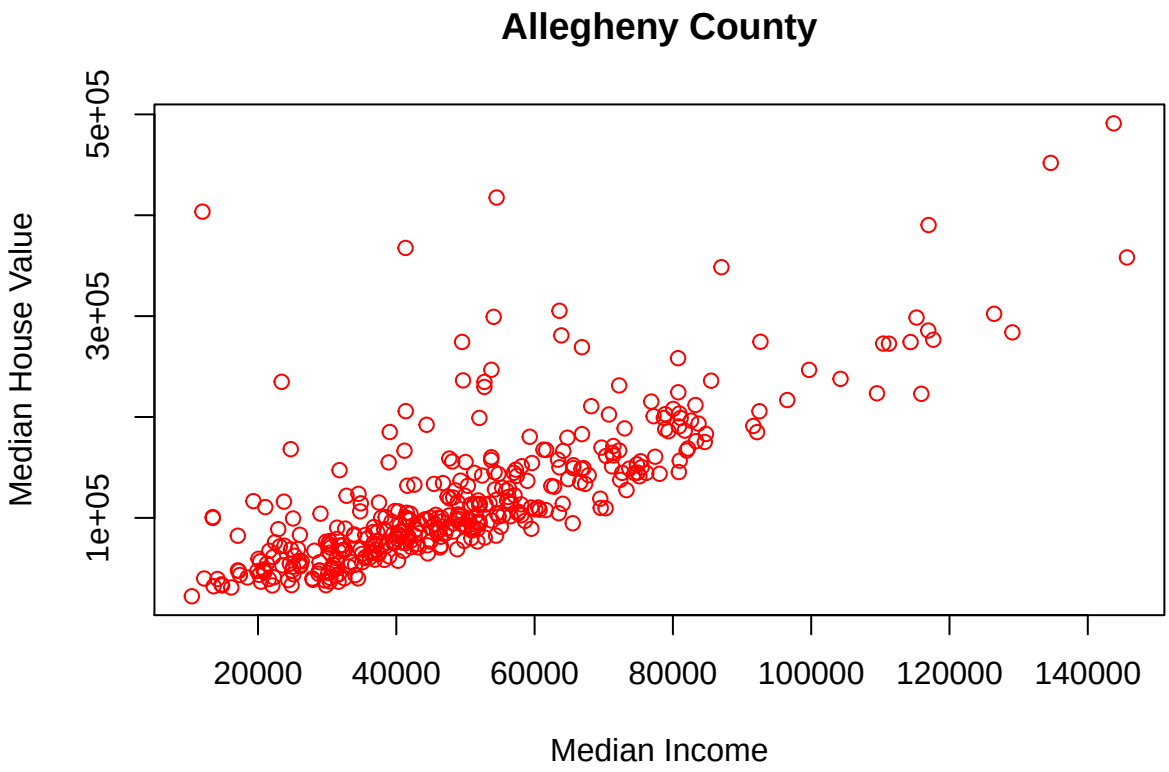
```
plot(na_Omit$Median_household_income[na_Omit$STATEFP == 6 &na_Omit$COUNTYFP == 1],na_Omit$Median_house_v
```




```
plot(na_Omit$Median_household_income[na_Omit$STATEFP == 6 &na_Omit$COUNTYFP == 85],na_Omit$Median_house
```




```
plot(na_Omit$Median_household_income[na_Omit$STATEFP == 42 &na_Omit$COUNTYFP == 3],na_Omit$Median_house
```


Allegheny County
Median House Value
1e+05
3e+05
5e+05
20000
40000
60000
80000
100000
120000
140000
Median Income