# Problem set 2

*Ankita Shankhdhar*

*25 September 2015*

Question 1 - Loading the data

Part a:

Loading the dataframe and calling it ca_pa

```
ca_pa=read.csv('http://www.math.umass.edu/~jstauden/calif_penn_2011.csv')
```

Part b:

This data has:

```
dim(ca_pa)
```

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

Part c:

This command goes through each variable and checks if there are any missing values. Then the colSums command sums up all the times there is a missing value found.

```
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
```

Part d:

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

Part e:

This omits 670 rows

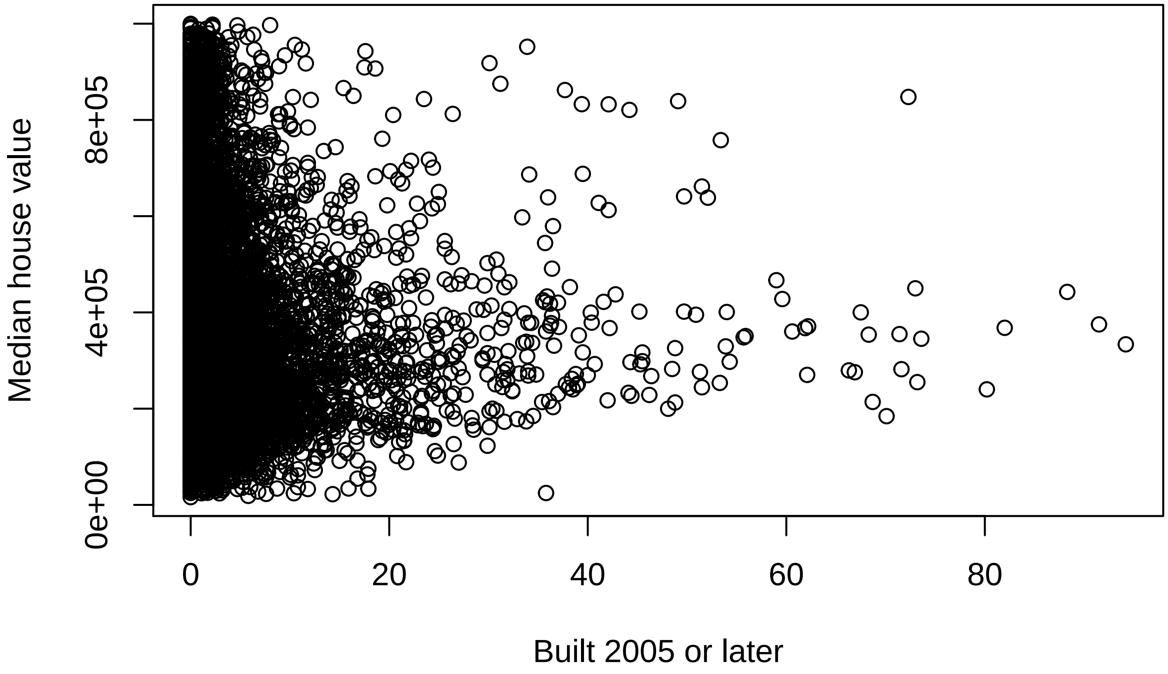
Part f:

Yes they are compatible! Part c gives us the amount of na in each column. However, part e just gives the sum of all of them. So part e is the union of part c.

Question 2 - This Very New House

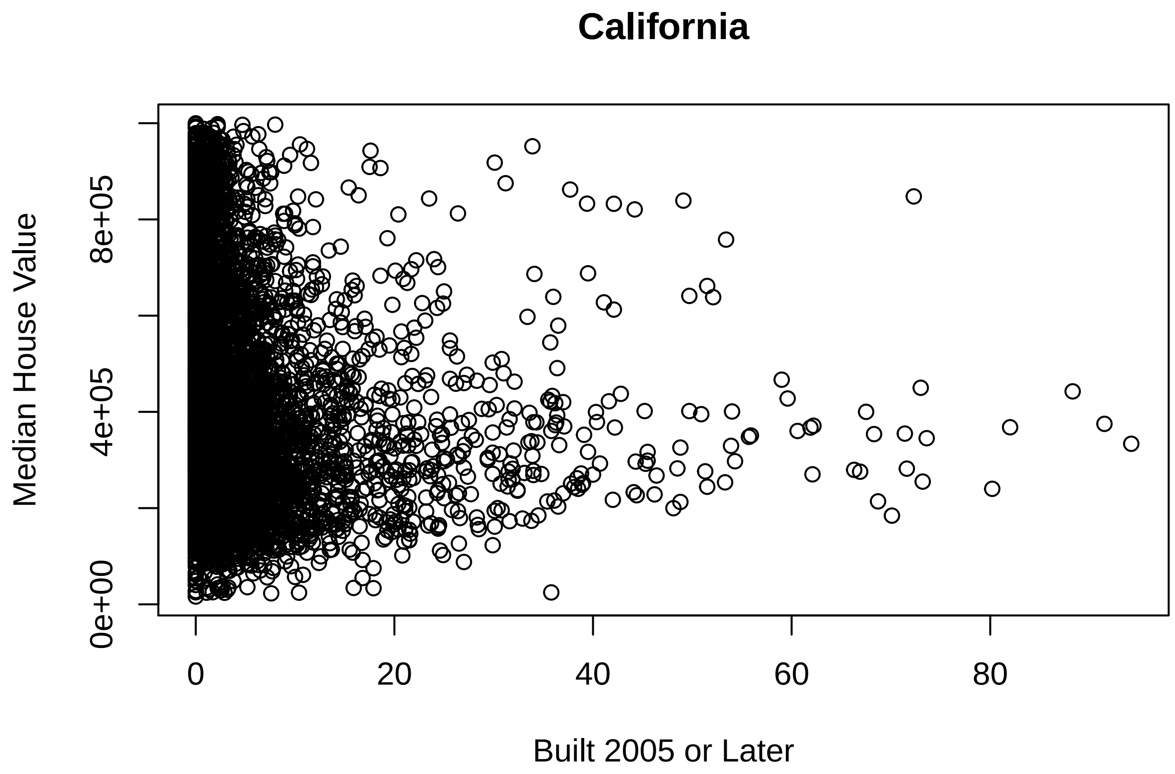
Part a:

```r
plot(ca_pa$Built_2005_or_later, ca_pa$Median_house_value,
     xlab="Built 2005 or later",ylab="Median house value")
```

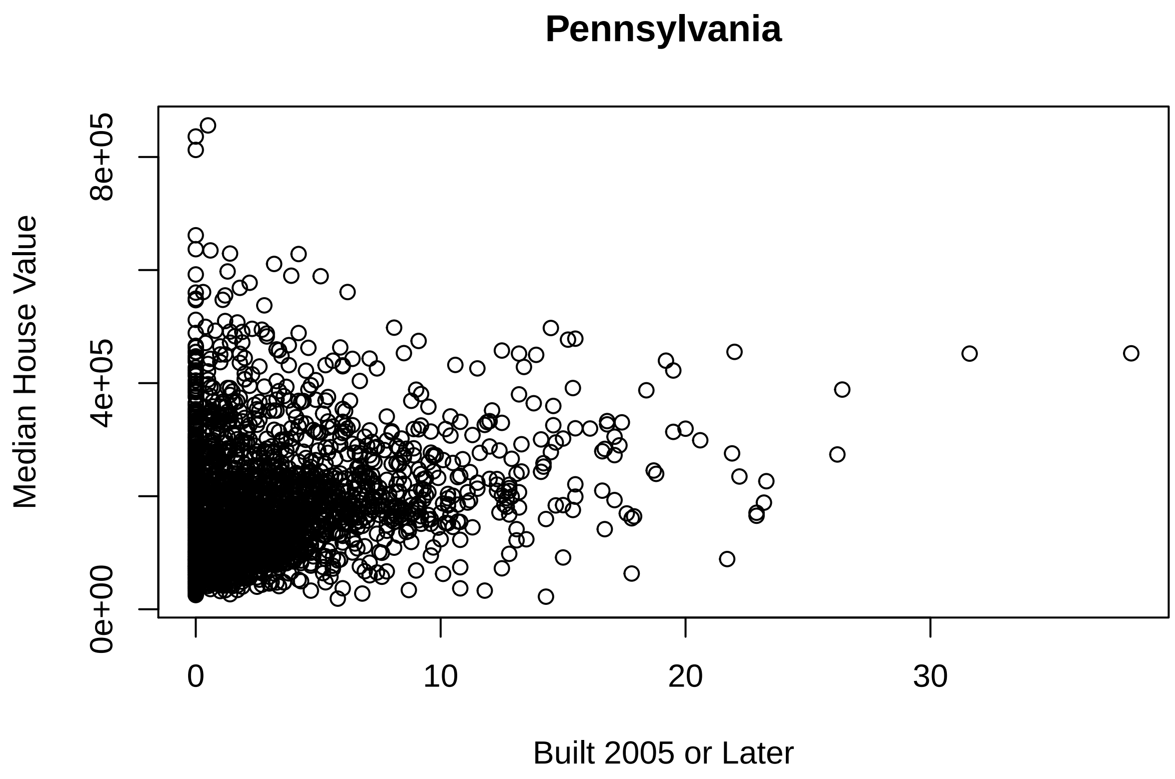


Part b:

```
cali<-which(ca_pa$STATEFP==6)
penn<-which(ca_pa$STATEFP==42)
plot(ca_pa$Built_2005_or_later[cali], ca_pa$Median_house_value[cali]
     , xlab='Built 2005 or Later',ylab='Median House Value',main='California')
```



```
plot(ca_pa$Built_2005_or_later[penn], ca_pa$Median_house_value[penn]
     , xlab='Built 2005 or Later',ylab='Median House Value',main='Pennsylvania')
```



Question 3 - Nobody Home

Part a:

```
ca_pa<-cbind(ca_pa,c(ca_pa$Vacant_units/ca_pa$Total_units))
names(ca_pa)[35] <- "Vacancy_rate"
min(ca_pa$Vacancy_rate)
```

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

```
max(ca_pa$Vacancy_rate)
```

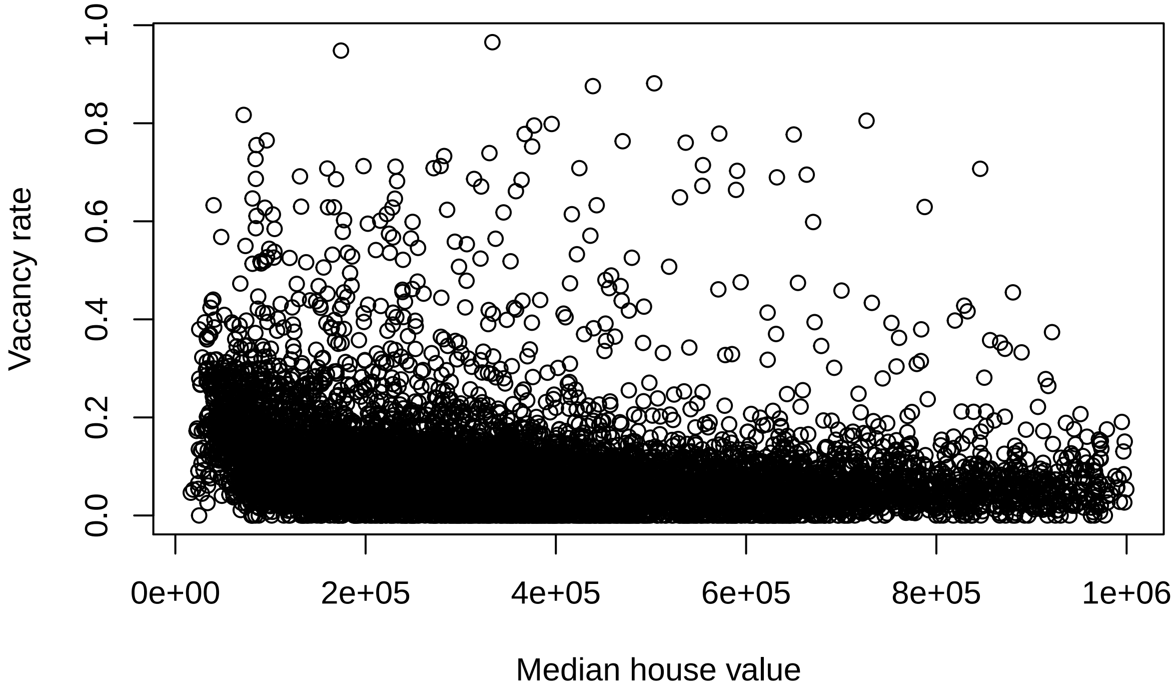
```
## [1] 0.965311
```

```
median(ca_pa$Vacancy_rate)
```

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

Part b:

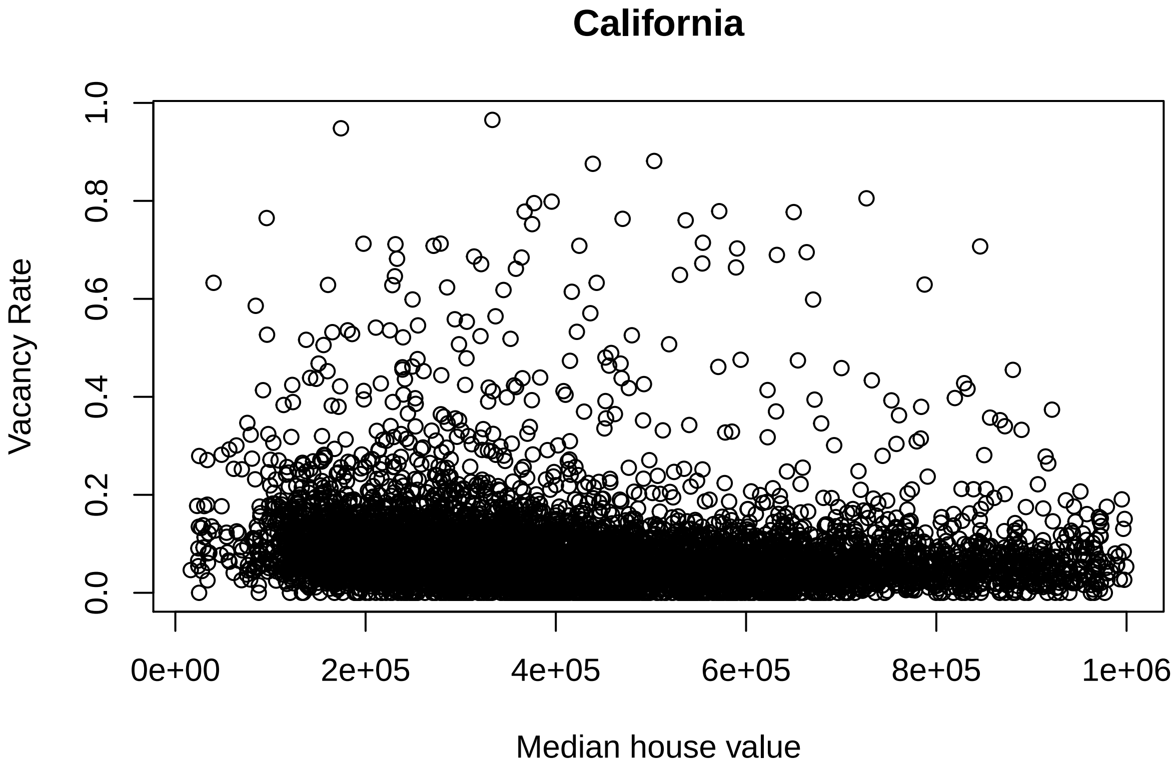
```
plot(ca_pa$Median_house_value,ca_pa$Vacancy_rate,
     xlab="Median house value",ylab="Vacancy rate")
```



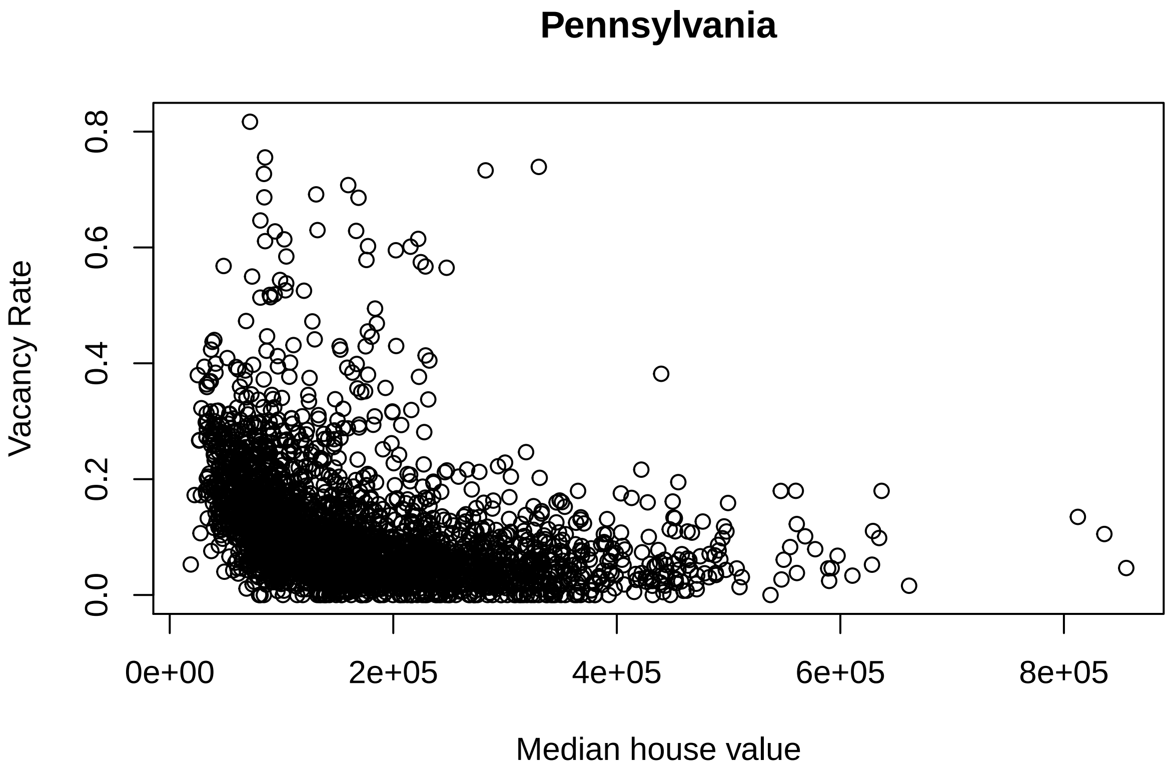
Part c:

There does seem to be a difference

```r
plot(ca_pa$Median_house_value[cali],ca_pa$Vacancy_rate[cali],
     xlab='Median house value',ylab='Vacancy Rate',main='California')
```



```
plot(ca_pa$Median_house_value[penn],ca_pa$Vacancy_rate[penn],
     xlab='Median house value',ylab='Vacancy Rate',main='Pennsylvania')
```



Question 4:

Part 1:

This code is creating a empty vector (acca) and walks through our dataset using the iterator tract. Whenever it finds the state to be 6 (California), if the county is is 1 (Alameda County). It is is then it takes index and saves it in the vector acca.

```
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)
      }
    }
  }
```

This code is creating a empty vector (accamhv). Then we loop through acca (vector of indices for Alameda county) and set accamhv to be all the median house value as

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

Part b:

```r
median(ca_pa$Median_house_value[ca_pa$STATEFP==6 & ca_pa$COUNTYFP==1])
```

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

Part c:

The average percentage of houses built since 2005 in Alameda is:

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

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

The average percentage of houses built since 2005 in Santa Clara is:

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

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

The average percentage of houses built since 2005 in Allegheny is:

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

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

Part d:

i.

```r
all_pct <-ca_pa$Built_2005_or_later/sum(ca_pa$Built_2005_or_later)
cor(ca_pa$Median_house_value,all_pct )
```

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

ii.

```r
cali_pct<-ca_pa$Built_2005_or_later[ca_pa$STATEFP==6]/
  sum(ca_pa$Built_2005_or_later[ca_pa$STATEFP==6])
cor(ca_pa$Median_house_value[ca_pa$STATEFP==6],cali_pct)
```

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

iii.

```r
penn_pct<-ca_pa$Built_2005_or_later[ca_pa$STATEFP==42]/
  sum(ca_pa$Built_2005_or_later[ca_pa$STATEFP==42])
cor(ca_pa$Median_house_value[ca_pa$STATEFP==42],penn_pct)
```

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

iv.

```r
alam_house <- ca_pa$Median_house_value[ca_pa$STATEFP==6 & ca_pa$COUNTYFP==1]
cor(alam_house,alameda)
```

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

v.

```r
sc_house <- ca_pa$Median_house_value[ca_pa$STATEFP==6 & ca_pa$COUNTYFP==85]
cor(sc_house,sc)
```

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

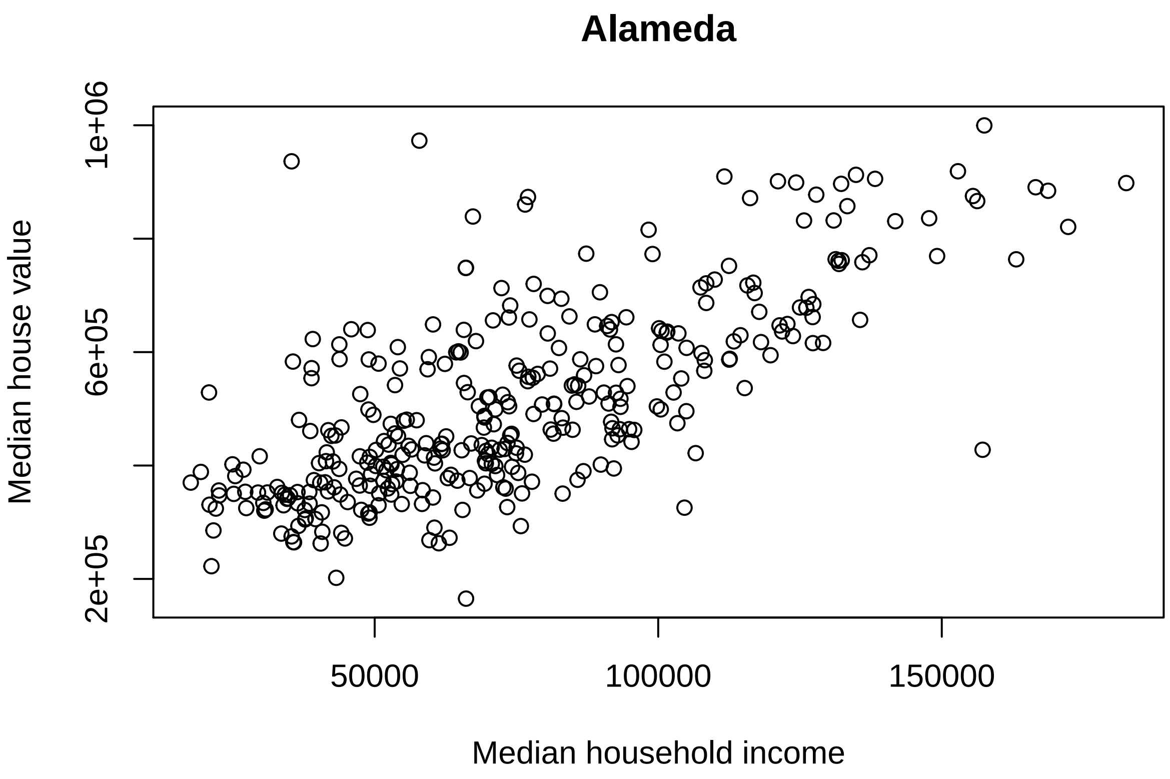
vi.

```r
alleg_house <- ca_pa$Median_house_value[ca_pa$STATEFP==42 & ca_pa$COUNTYFP==3]
cor(alleg_house,alleg)
```

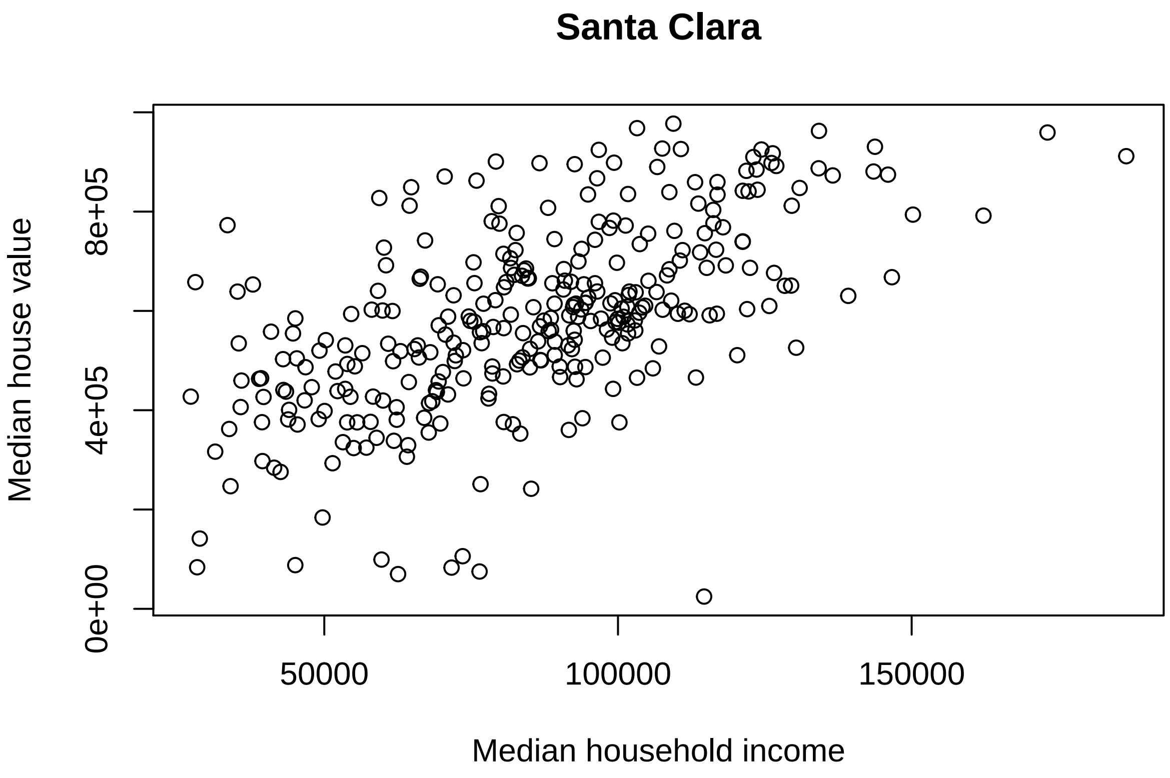
```
## [1] 0.1939652
```

Part e:

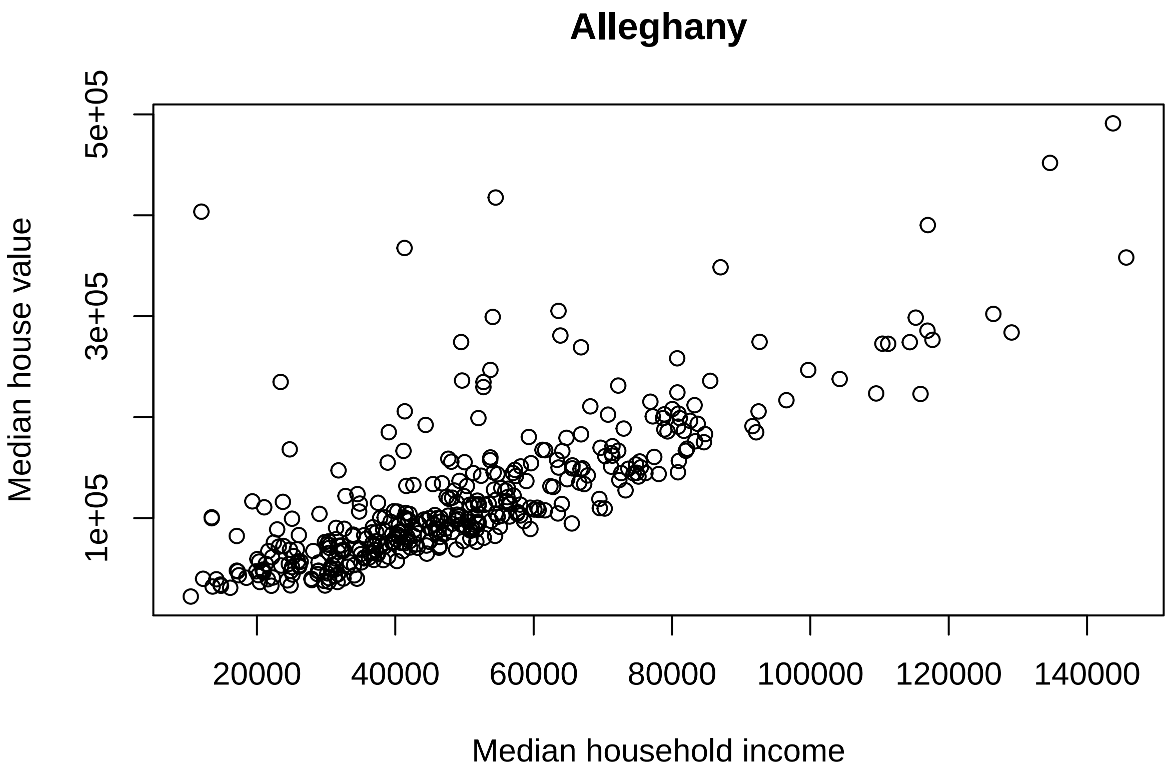
```r
alam_income <-ca_pa$Median_household_income[ca_pa$STATEFP==6 &
                                            ca_pa$COUNTYFP==1]
plot(alam_income,alam_house,xlab="Median household income",
     ylab="Median house value",main="Alameda")
```



```
sc_income <-ca_pa$Median_household_income[ca_pa$STATEFP==6 &
                                            ca_pa$COUNTYFP==85]
plot(sc_income,sc_house,xlab="Median household income",
     ylab="Median house value",main="Santa Clara")
```



```
alleg_income<- ca_pa$Median_household_income[ca_pa$STATEFP==42 &
                                              ca_pa$COUNTYFP==3]
plot(alleg_income,alleg_house,xlab="Median household income",
     ylab="Median house value",main="Alleghany")
```


Alleghany
Median house value
1e+05
3e+05
5e+05
Median household income