Homework 2

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`.
   b. How many rows and columns does the dataframe have? 11275 行 34 列
   c. Run this command, and explain, in words, what this does: 这个程序的作用是计算每一列中的空缺值的个数

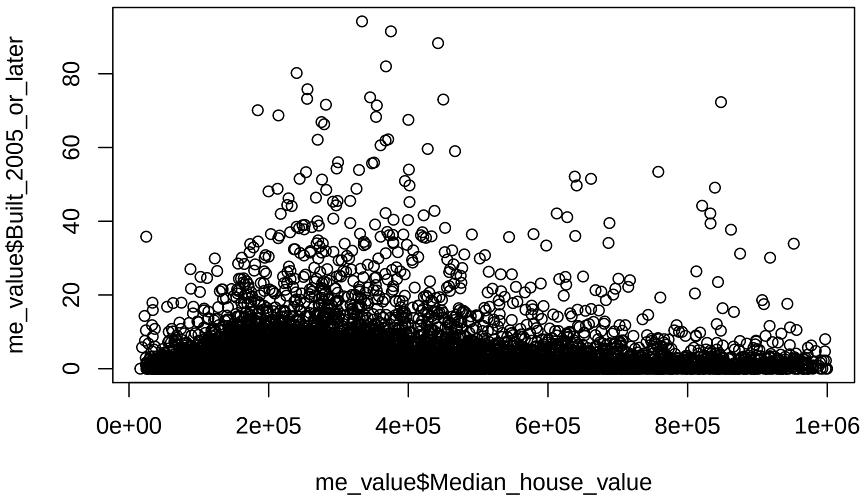
```
ca_pa <- read.csv("data/calif_penn_2011.csv",header=T)
a <- dim(ca_pa)
b <- colSums(apply(ca_pa,c(1,2),is.na))
```

```
d. The function `na.omit()` takes a dataframe and returns a new dataframe, omitting any
e. How many rows did this eliminate? 消除了670行
f. Are your answers in (c) and (e) compatible? Explain.
```

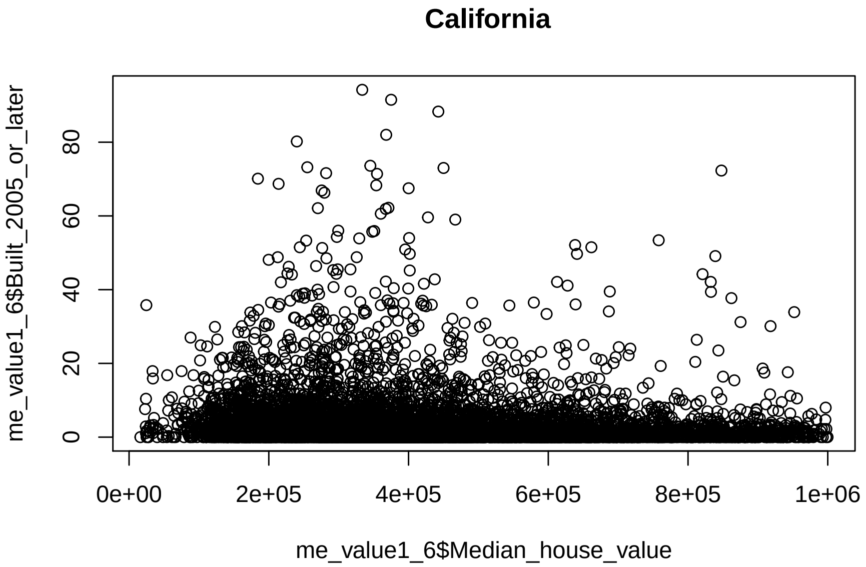
兼容。虽然 sum(colSums(apply(ca_pa,c(1,2),is.na))) 的结果是 3034，比 670 多，但是有可能有一行中出现多个 NA 的情况，所以不矛盾。2. *This Very New House*

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.

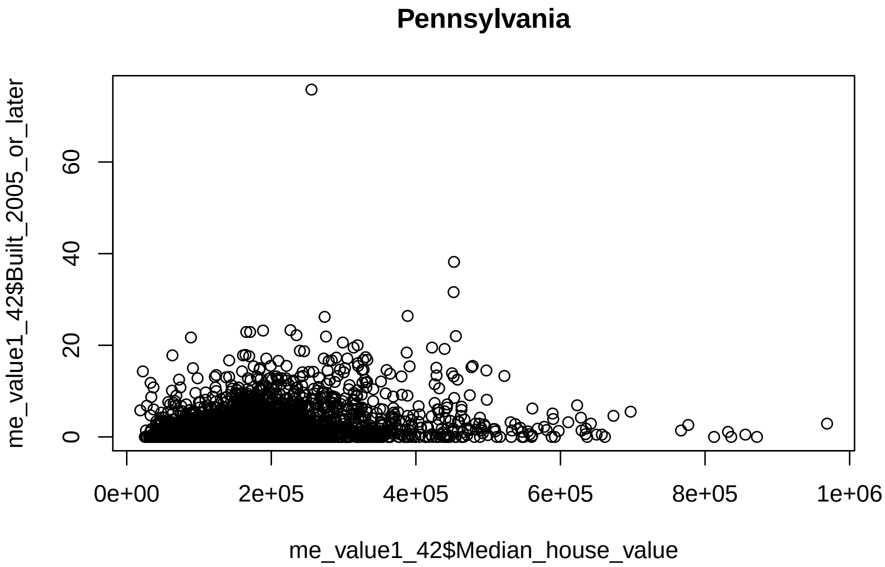
```
#a.中位价格与2005年以后建的房子所占的比例之间的关系
me_value <- na.omit(data.frame( ca_pa["Built_2005_or_later"], ca_pa["Median_house_value
plot(y = me_value$Built_2005_or_later,x = me_value$Median_house_value)
```



```
#b.按州划分
me_value1 <- na.omit(data.frame(ca_pa["STATEFP"], ca_pa["Built_2005_or_later"], ca_pa["
for (i in 1:length(me_value1$STATEFP)){
   if(me_value1$STATEFP[i] == 42){
     break
   }
}
num1 <- seq(from = 1, to = i)
num2 <- seq(from = i+1, to = length(me_value1$STATEFP))
me_value1_6 <- data.frame( Built_2005_or_later = me_value1$Built_2005_or_later[num1],Me
me_value1_42 <- data.frame(Built_2005_or_later = me_value1$Built_2005_or_later[num2],Me
plot(y = me_value1_6$Built_2005_or_later,x = me_value1_6$Median_house_value,main = " Ca
```



```
plot(y = me_value1_42$Built_2005_or_later,x = me_value1_42$Median_house_value, main = "
```



3. *Nobody Home*

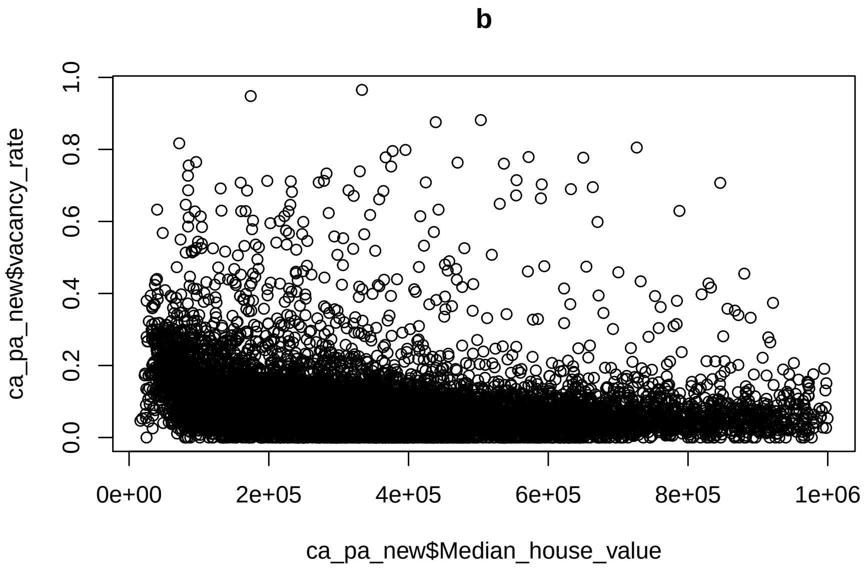
The vacancy rate is the fraction of housing units which are not occupied.

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

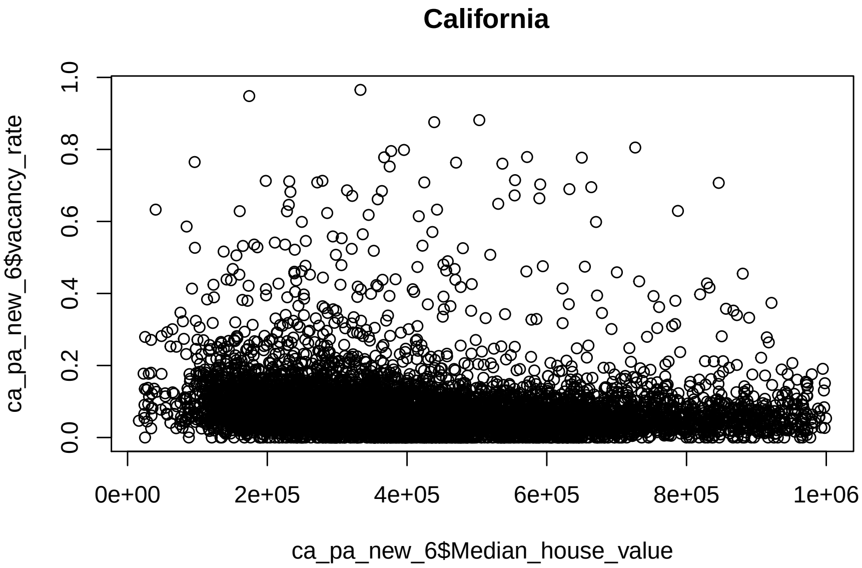
a. Add a new column to the dataframe which contains the vacancy rate. What are the minimum, maximum, mean, and median vacancy rates? 最小值是0，最大值是0.965311，平均数是0.08888789，中位数是0.06767283

b. Plot the vacancy rate against median house value. c. Plot vacancy rate against median house value separately for California and for Pennsylvania. Is there a difference? 宾夕法尼亚州的房屋价格和空置率呈现明显的房价越高，空置率越低。加利福利亚州则表现得比较平均。

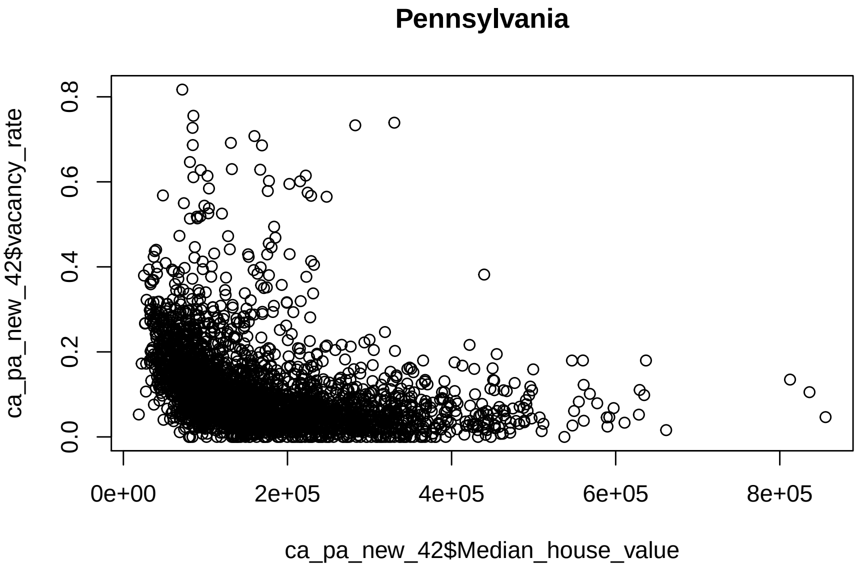
```r
#a
ca_pa_1 <- na.omit(ca_pa)
vacancy_rate <- ca_pa_1$Vacant_units/ca_pa_1$Total_units
ca_pa_new <- data.frame(ca_pa_1,vacancy_rate = vacancy_rate)
min_vacancy_rate <- min(vacancy_rate)
max_vacancy_rate <- max(vacancy_rate)
mean_vacancy_rate <- mean(vacancy_rate)
median_vacancy_rate <- median(vacancy_rate)
#b
plot(ca_pa_new$Median_house_value,ca_pa_new$vacancy_rate,main = "b")
```



```
#c
for (i in 1:length(ca_pa_new$STATEFP)){
   if(ca_pa_new$STATEFP[i] == 42){
     break
   }
}
num1 <- seq(from = 1, to = i)
num2 <- seq(from = i+1, to = length(ca_pa_new$STATEFP))
ca_pa_new_6 <- data.frame( vacancy_rate = ca_pa_new$vacancy_rate[num1],Median_house_val
ca_pa_new_42 <- data.frame(vacancy_rate = ca_pa_new$vacancy_rate[num2],Median_house_val
plot(y = ca_pa_new_6$vacancy_rate,x = ca_pa_new_6$Median_house_value,main = " Californi
```



```r
plot(y = ca_pa_new_42$vacancy_rate,x = ca_pa_new_42$Median_house_value, main = "Pennsyl
```



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. 计算 AlamedaCounty 的平均房价。是先找出有哪些单元是是属于 Alameda County，然后计算平均房价。

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
median_value = median(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 1),"Median
```

c. For Alameda, Santa Clara and Allegheny Counties, what were the average percentages o

Alameda自2005年以来建造房屋的平均百分比是2.820468%。

Santa Clara自2005年以来建造房屋的平均百分比是3.200319%

Allegheny Counties自2005年以来建造房屋的平均百分比是1.474219%

```r
 mean_rate_Alameda = mean(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 1),"Bu
mean_rate_Santa_Clara = mean(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 85)
mean_rate_Allegheny_Counties = mean(ca_pa_1[which(ca_pa_1$STATEFP == 42&ca_pa_1$COUNTYF
```

d. The `cor` function calculates the correlation coefficient between two variables. Wh

(i)-0.01893186

(ii)-0.1153604

(iii)0.2681654

(iv)0.01303543

(v)-0.1726203

(vi)0.1939652

```r
R_total = cor(ca_pa_1$Median_house_value,ca_pa_1$Built_2005_or_later)
R_California = cor(ca_pa_1[which(ca_pa_1$STATEFP == 6),"Median_house_value"],ca_pa_1[wh
R_Pennsylvania = cor(ca_pa_1[which(ca_pa_1$STATEFP == 42),"Median_house_value"],ca_pa_1
```

~~~~
R_Alameda_County = cor(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 1),"Built
R_Santa_Clara_County = cor(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 85),"
R_Allegheny_County = cor(ca_pa_1[which(ca_pa_1$STATEFP == 42&ca_pa_1$COUNTYFP == 3),"Bu
```
e. Make three plots, showing median house values against median income, for Alameda, Sa

```r
#Alameda
plot(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 1),"Median_house_value"],ca
```

![](Homework-02-1-_files/figure-latex/unnamed-chunk-7-1.pdf)<!-- -->

```r
#Santa Clara
plot(ca_pa_1[which(ca_pa_1$STATEFP == 6&ca_pa_1$COUNTYFP == 85),"Median_house_value"],c
```

![](Homework-02-1-_files/figure-latex/unnamed-chunk-7-2.pdf)<!-- -->

```r
#Allegheny Counties
plot(ca_pa_1[which(ca_pa_1$STATEFP == 42&ca_pa_1$COUNTYFP == 3),"Median_house_value"],c
```

![](Homework-02-1-_files/figure-latex/unnamed-chunk-7-3.pdf)<!-- -->
~~~~

```r
acca <- c()
for (tract in 1:nrow(ca_pa_1)) {
  if (ca_pa_1$STATEFP[tract] == 6) {
    if (ca_pa_1$COUNTYFP[tract] == 1) {
      acca <- c(acca, tract)
    }
  }
```

```
}
accamhv <- c()
for (tract in acca) {
  accamhv <- c(accamhv, ca_pa_1[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
```

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

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

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

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

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

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

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

Explain the output from the successive uses of table().
table() 的作用是统计频次，如果没有出现输出 0，也可以统计除了给出的剩下还有多少。

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

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

```
#定义函数
f_exceed <- function(vector,value){
  k = 0
  for (i in vector){
    if (i > value){
      k <- k + 1
    }
  }
  return(k)
}
#检验
vector_1 <- 1:100
value_exceed <- f_exceed(vector_1,1)
```

(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
```

```
rabbit1 <- data.frame(BPchange = Rabbit$BPchange,Animal =  Rabbit$Animal)
rabbit2 <- data.frame(Treatment = Rabbit$Treatment,Dose =  Rabbit$Dose)
rabbit_new <- data.frame(unique(rabbit2),unstack(rabbit1))
print(rabbit_new)
```

```
##    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
## 3    Control  25.00 10.00  4.00  3.00  6.00  5.0
## 4    Control  50.00 26.00 12.00 14.00 19.00 16.0
## 5    Control 100.00 37.00 27.00 22.00 33.00 20.0
## 6    Control 200.00 32.00 29.00 24.00 33.00 18.0
## 31       MDL   6.25  1.25  1.40  0.75  2.60  2.4
## 32       MDL  12.50  0.75  1.70  2.30  1.20  2.5
## 33       MDL  25.00  4.00  1.00  3.00  2.00  1.5
## 34       MDL  50.00  9.00  2.00  5.00  3.00  2.0
## 35       MDL 100.00 25.00 15.00 26.00 11.00  9.0
## 36       MDL 200.00 37.00 28.00 25.00 22.00 19.0
```