Homework 2

□ □ □ 3220103172

2022 年 7 月 3 日

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?

```
dim(ca_pa)
```

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

It has 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))
```

```
##                          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
```

对每行每列的值是否是 NA 进行判断，FALSE=0,TRUE=1，并按列进行相加，得到 ca_pa 中每列中值为 NA 的个数。

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(ca_pa_new)
```

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

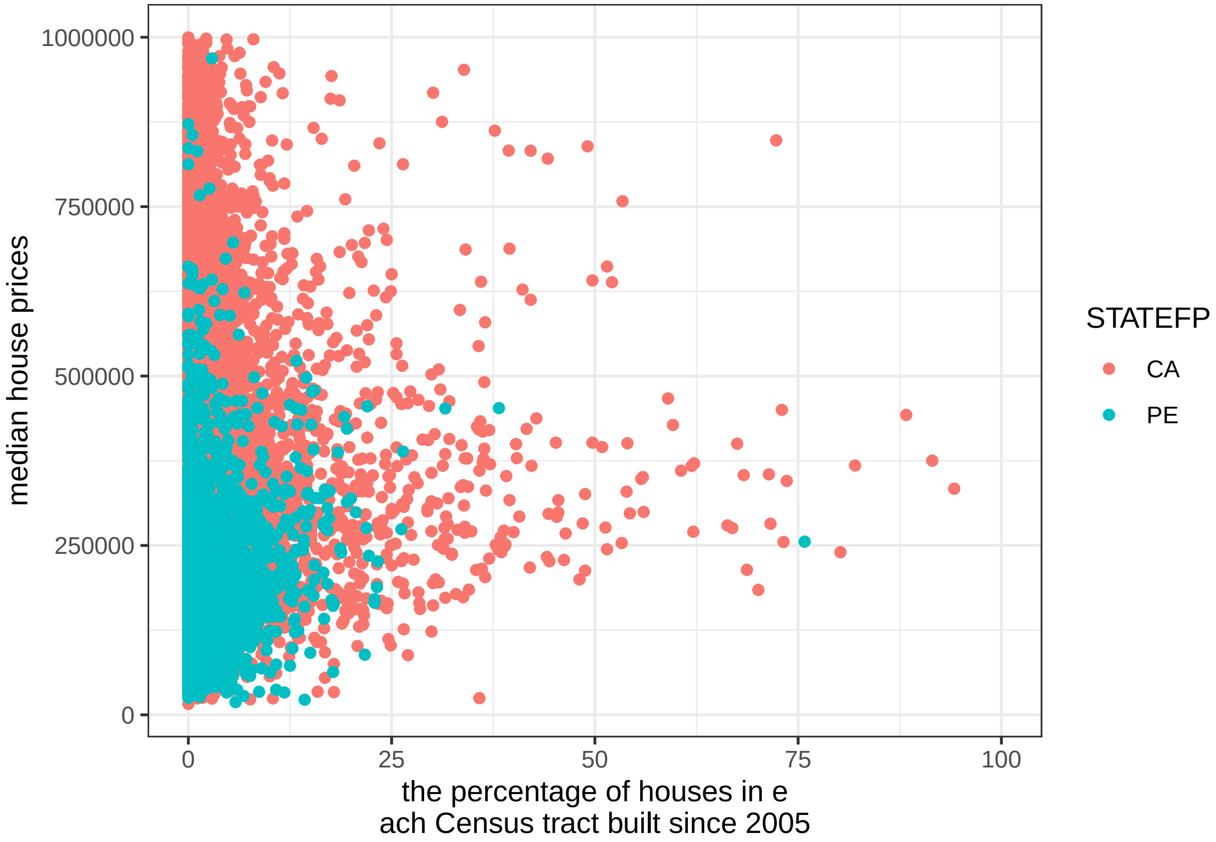
它有 10605 行和 34 列

f. Are your answers in (c) and (e) compatible? Explain. 结果是兼容的，因为 `na.omit()` 函数将表中 NA 的行去掉了，由 c 可知表中有 3034 个数据空缺，有可能某一行中有多个数据是空值，所以 c 和 e 的结果是兼容的。

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.

```
ggplot(data=ca_pa)+geom_point(aes(x = Built_2005_or_later,
                                  y = Median_house_value,color=factor(STATEFP)))+
labs(x = "the percentage of houses in e
ach Census tract built since 2005"
,y = "median house prices") +theme_bw()+ scale_colour_discrete(
    name="STATEFP",
    labels = c("CA","PE"))
```

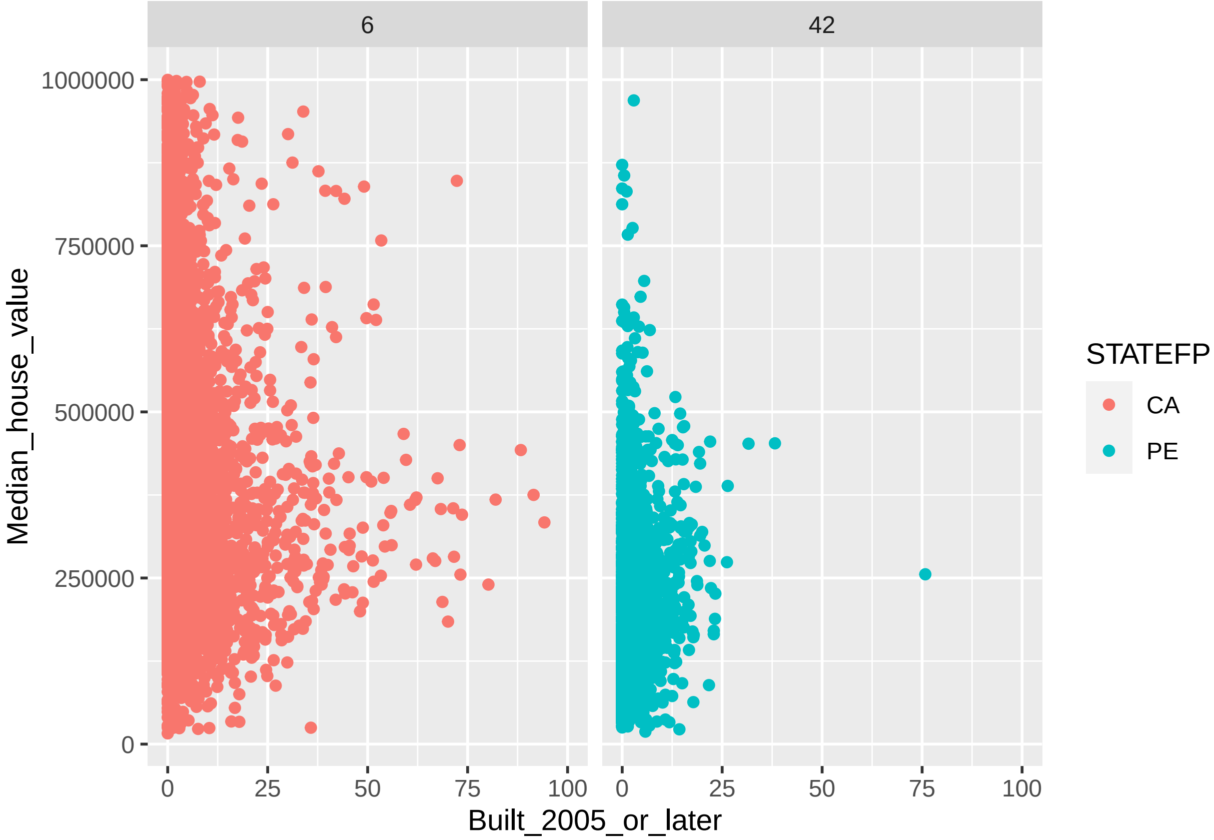
```
## Warning: Removed 599 rows containing missing values (geom_point).
```



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.

```
ggplot(data=ca_pa)+
  geom_point(aes(x = Built_2005_or_later,
                                  y = Median_house_value,color=factor(STATEFP)))+ scale_colour_discre
    name="STATEFP",
    labels = c("CA","PE"))+
  facet_wrap(~ STATEFP)
```

```
## Warning: Removed 599 rows containing missing values (geom_point).
```



其中

6 代表加利福尼亚州，42 代表宾夕法尼亚州

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?

```
library(dplyr)
ca_pa2 <- ca_pa_new %>% mutate(Vacancy = Vacant_units/Total_units)
vacancy_rate <- ca_pa2[,'Vacancy']
min(vacancy_rate)
```

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

```
max(vacancy_rate)
```

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

```
mean(vacancy_rate)
```

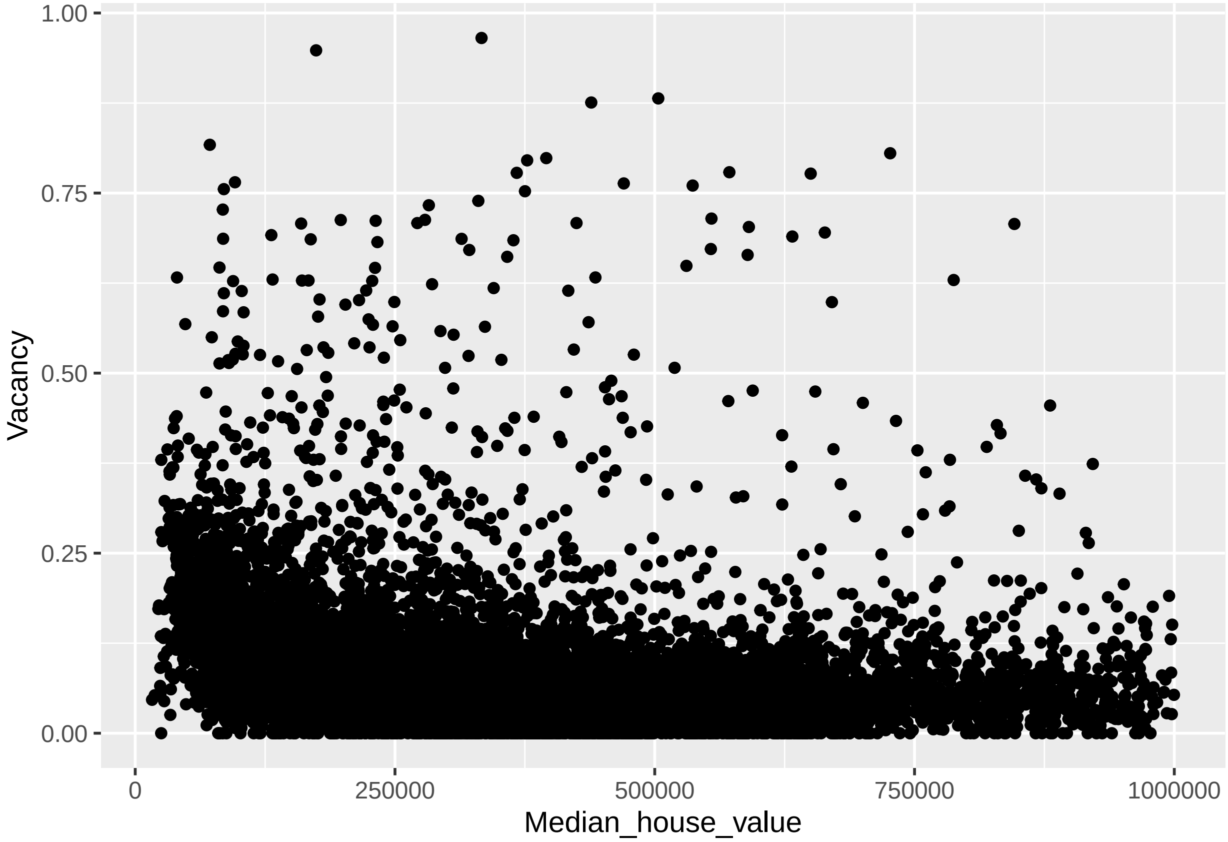
```
## [1] 0.08888789
```

```
median(vacancy_rate)
```

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

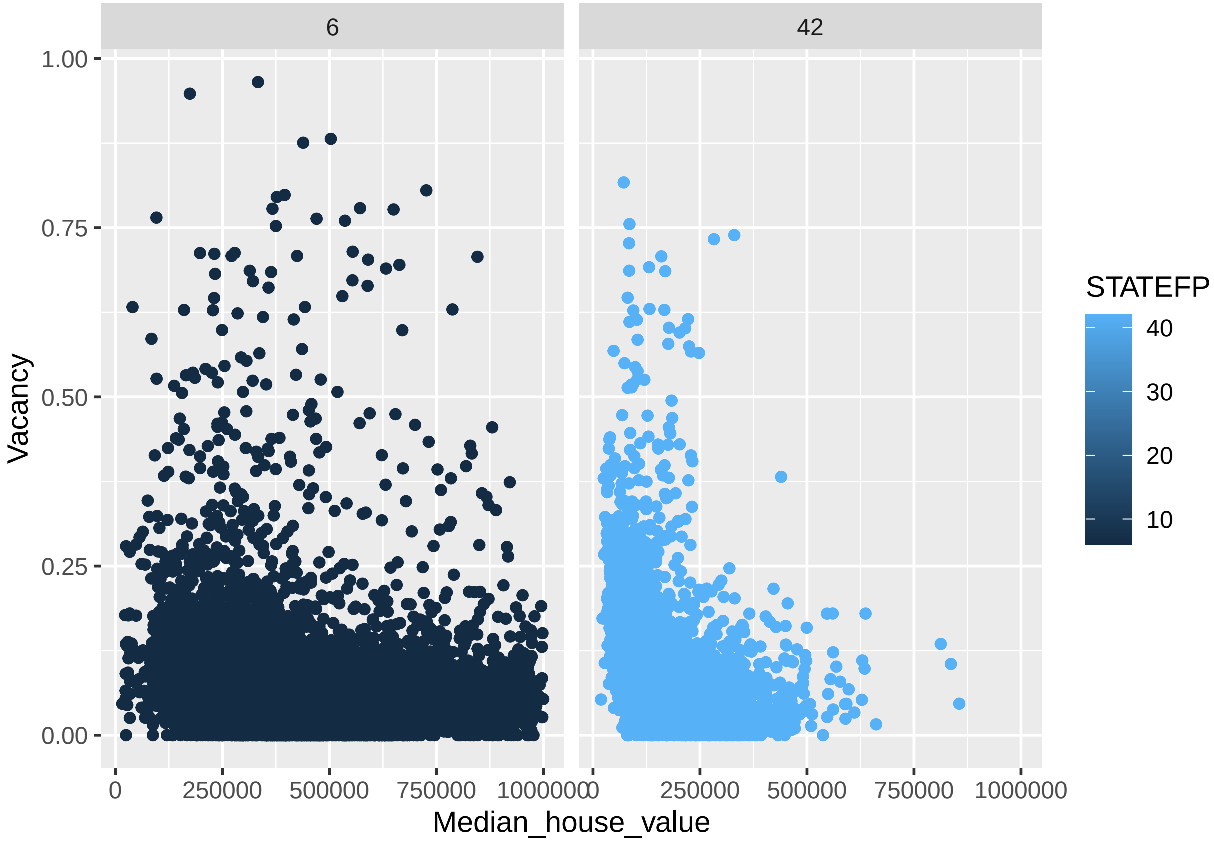
b. Plot the vacancy rate against median house value.

```
ggplot(data=ca_pa2)+geom_point(aes(x = Median_house_value,y=Vacancy))
```



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

```
ggplot(data=ca_pa2)+geom_point(aes(x = Median_house_value,y=Vacancy,color=STATEFP))+
  facet_wrap(~ STATEFP)
```



其中 6 代表加州，42 代表宾州。不同在于加州的空缺率相对更高。

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.

```
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)
```

```
## [1] NA
```

将 California 中的 Alameda County 的数据所在行的位置记录在向量 acca 中, 又将将 California 中的 Alameda County 的 Median_house_value 记录在向量 accamhv 中，并计算其中位数。由于 accamhv 中存在值 NA 的元素，因此使用 median() 函数得到的值一定是 NA。我们应该所以将 ca_pa 替换成 ca_pa_new。

```
acca <- c()
for (tract in 1:nrow(ca_pa_new)) {
 if (ca_pa_new$STATEFP[tract] == 6) {
 if (ca_pa_new$COUNTYFP[tract] == 1) {
 acca <- c(acca, tract)
 }
 }
}
accamhv <- c()
for (tract in acca) {
 accamhv <- c(accamhv, ca_pa_new[tract,10])
}
median(accamhv)
```

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

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.

```
library(dplyr)
median((ca_pa_new %>% filter(STATEFP == 6 & COUNTYFP == 1))[,10])
```

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

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

```
#For Alameda:
library(dplyr)
mean((ca_pa_new %>% filter(STATEFP == 6 &
                            COUNTYFP == 1))[,'Built_2005_or_later'])
```

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

```
#For Santa Clara:
library(dplyr)
mean((ca_pa_new %>% filter(STATEFP == 6 &
                            COUNTYFP == 85))[,'Built_2005_or_later'])
```

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

```
#For Allegheny Counties:
library(dplyr)
mean((ca_pa_new %>% filter(STATEFP == 42 &
                             COUNTYFP == 3))[,'Built_2005_or_later'])
```

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

Alameda, Santa Clara 和 Allegheny Counties 的 average percentages of housing built since 2005 分别是 2.820468，3.200319，1.474219。

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?

```
#(i) the whole data
cor(ca_pa_new[,'Median_house_value'],
    ca_pa_new[,'Built_2005_or_later'])
```

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

```
#(ii) all of California
cor((ca_pa_new %>% filter(STATEFP == 6))
    [,'Median_house_value'],
    (ca_pa_new %>%  filter(STATEFP == 6 ))
    [,'Built_2005_or_later'])
```

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

```
#(iii) all of Pennsylvania
cor((ca_pa_new %>% filter(STATEFP == 42 ))[,'Median_house_value'],
    (ca_pa_new %>%  filter(STATEFP == 42 ))
    [,'Built_2005_or_later'])
```

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

```
#(iv) Alameda County
cor((ca_pa_new %>% filter(STATEFP == 6 & COUNTYFP == 1))[,'Median_house_value'],(ca_pa_new %>% filter
    [,'Built_2005_or_later'])
```

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

```
#(v) Santa Clara County
cor((ca_pa_new %>% filter(STATEFP == 6 &
 COUNTYFP == 85))[,'Median_house_value'],(ca_pa_new %>%
filter(STATEFP == 6 &   COUNTYFP == 85))[,'Built_2005_or_later'])
```

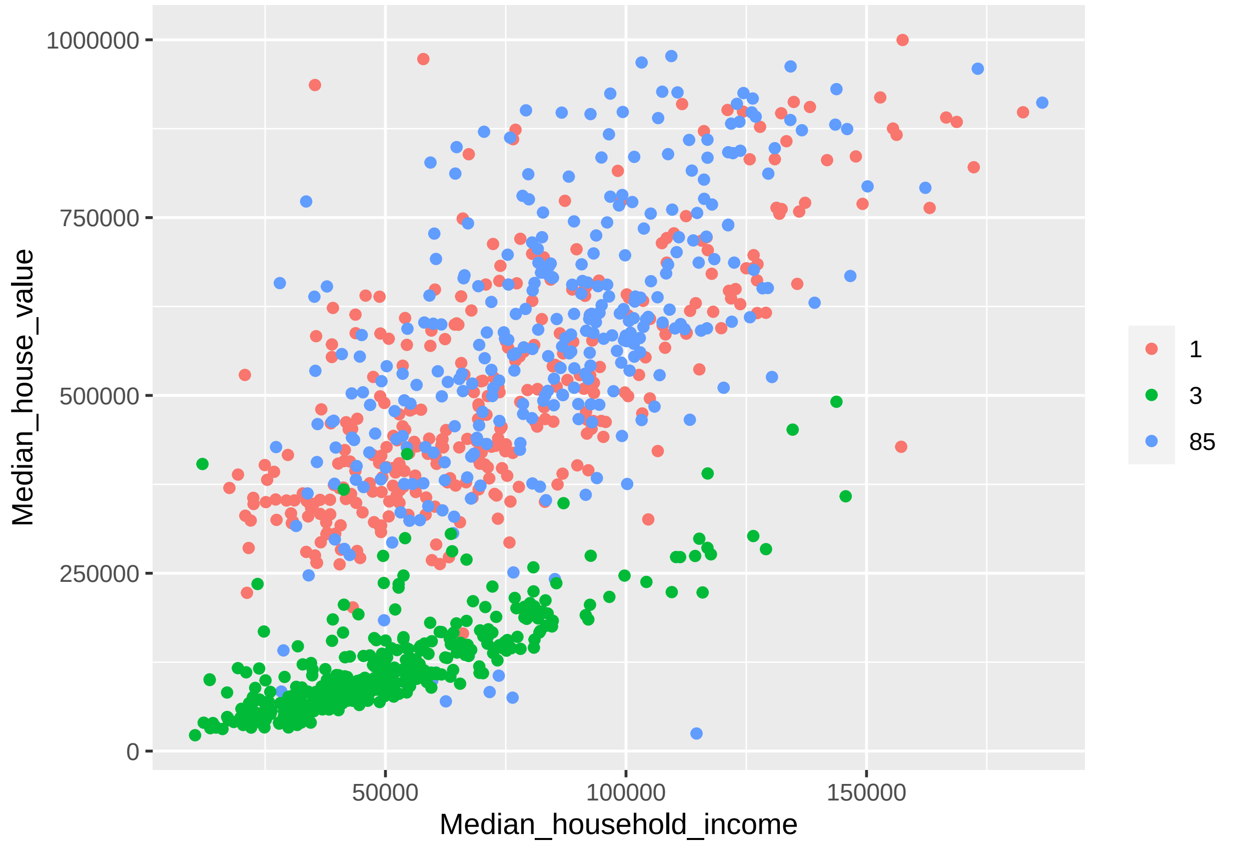
```
## [1] -0.1726203
```

```
#(vi) Allegheny County
cor((ca_pa_new %>% filter(STATEFP == 42 &
 COUNTYFP == 3))[,'Median_house_value'],(ca_pa_new %>% filter(STATEFP == 42 &COUNTYFP == 3))[,'Built_
```

```
## [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.)

```
tu<-ca_pa_new %>% filter((STATEFP == 6 &
                COUNTYFP == 1)|(STATEFP == 6 &
                  COUNTYFP == 85)|(STATEFP == 42 &
                                   COUNTYFP == 3))


ggplot(data=tu)+
  geom_point(aes(x = Median_household_income,y = Median_house_value,color=factor(COUNTYFP)))+
  theme(legend.title=element_blank())
```



其中 1,3,85 分别代表 Alameda, Santa Clara, Allegheny Counties。

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() 函数可以统计向量中各元素出现的次数, 并以 levels 为分类依据，计算出现次数（如果有 levels）, 若有 exclude=NULL，则显示 levels 以外的值的出现次数

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

```
fx <- function(x,a){sum(table(x[x>a]))/(length(x))}
```

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

```
fx(1:100,40)
```

```
## [1] 0.6
```

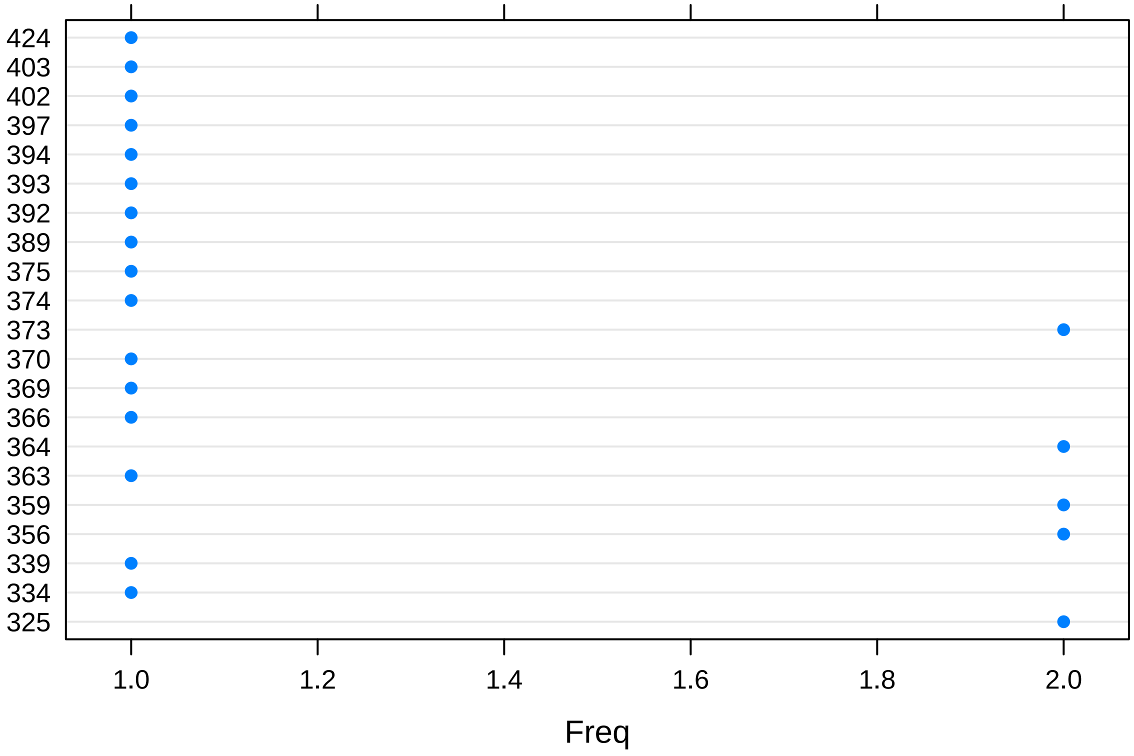
输出结果与预期相符合。

(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.

```
library('Devore7')
```

```
## 载入需要的程辑包：MASS

##
## 载入程辑包：'MASS'

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

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

## 载入需要的程辑包：lattice
```

```
dotplot(ex01.36)
```



```
fx(t(ex01.36),420)
```

```
## [1] 0.03846154
```

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')
library('dplyr')
a1 <- unstack(Rabbit,BPchange~Animal)
a2 <- unstack(Rabbit,Dose~Animal)
a3 <- unstack(Rabbit,Treatment~Animal)
result<-cbind(a3[,1],a2[,1],a1)
result
```

```
##    a3[, 1] a2[, 1]    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
## 7      MDL    6.25  1.25  1.40  0.75  2.60  2.4
## 8      MDL   12.50  0.75  1.70  2.30  1.20  2.5
## 9      MDL   25.00  4.00  1.00  3.00  2.00  1.5
## 10     MDL   50.00  9.00  2.00  5.00  3.00  2.0
## 11     MDL  100.00 25.00 15.00 26.00 11.00  9.0
## 12     MDL  200.00 37.00 28.00 25.00 22.00 19.0
```