《数学实践》作业二

许乐乐

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.

```
library(tidyverse)
library(ggplot2)
```

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
```

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
```

apply(ca_pa,c(1,2),is.na) 表示把 is.na 这个查找缺省值的函数应用到数据集 ca_pa 的行和列（c(1,2) 中,1 表示行，2 表示列）。colSums() 表示对列求和。因此结果显示该数据集中每列中有多少数据是 NA。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<-na.omit(ca_pa)
```

`e. How many rows did this eliminate?`

```
dim(read.csv("data./calif_penn_2011.csv"))[1]-dim(ca_pa)[1]
```

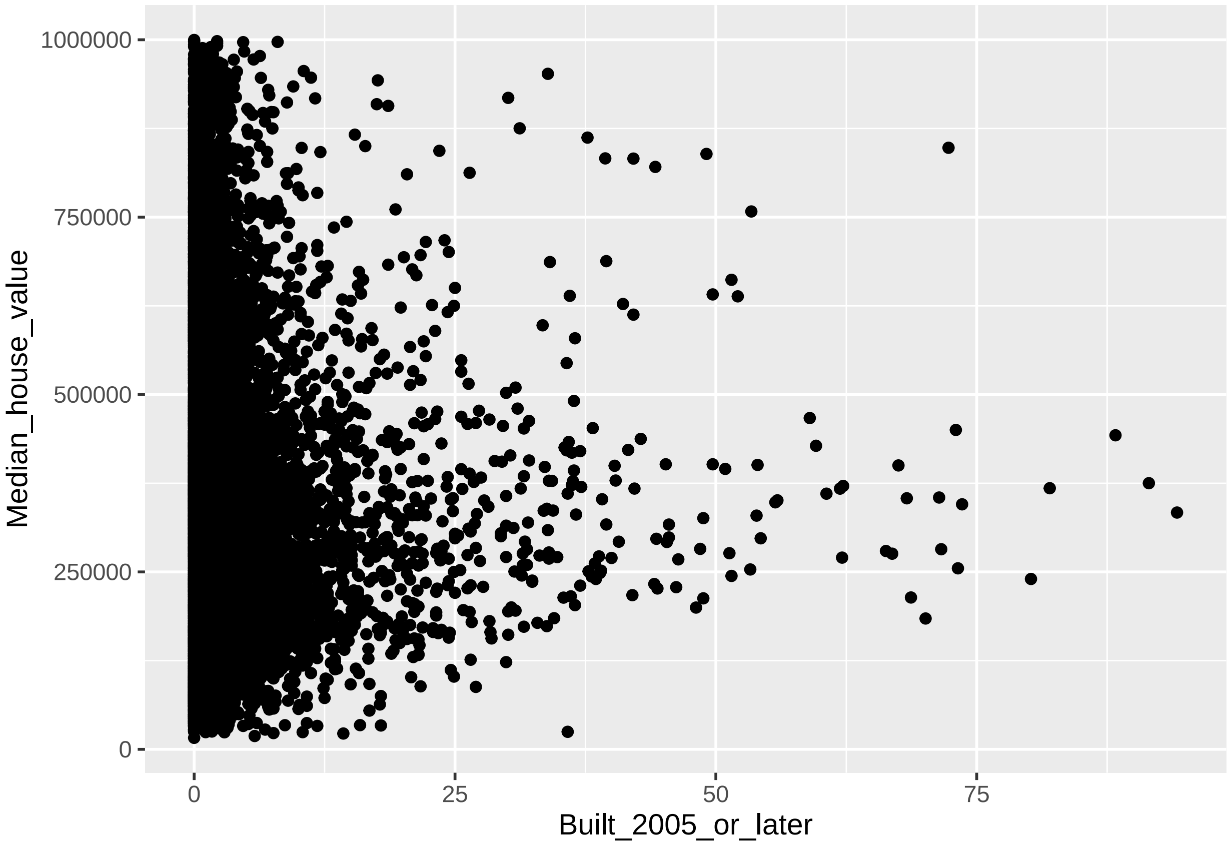
```
## [1] 670
```

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

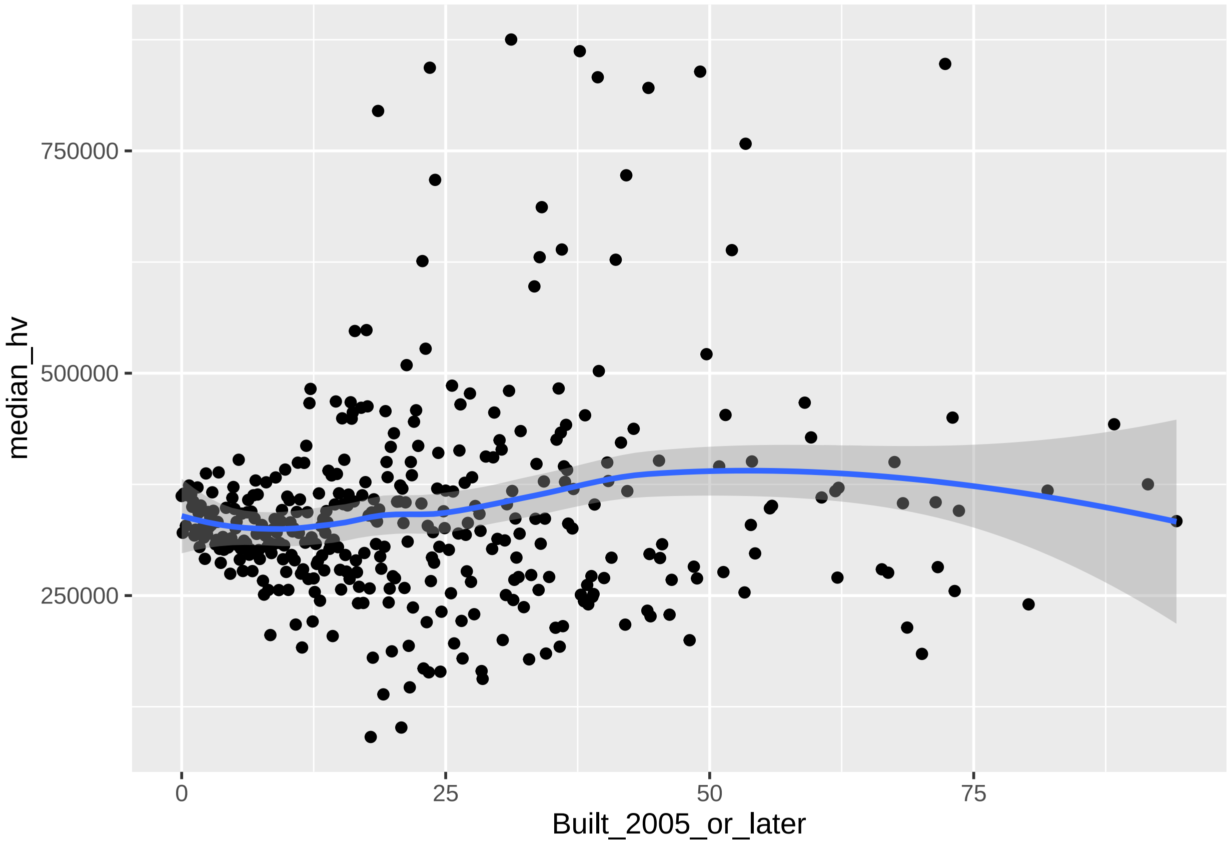
不一致。因为 (c) 中计算的是每列有几个 NA 数据，(e) 中计算的是至少含有一个 NA 数据的行数。

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.

```
attach(ca_pa)
ggplot(ca_pa) +
    geom_point(aes(x=Built_2005_or_later,y=Median_house_value))
```

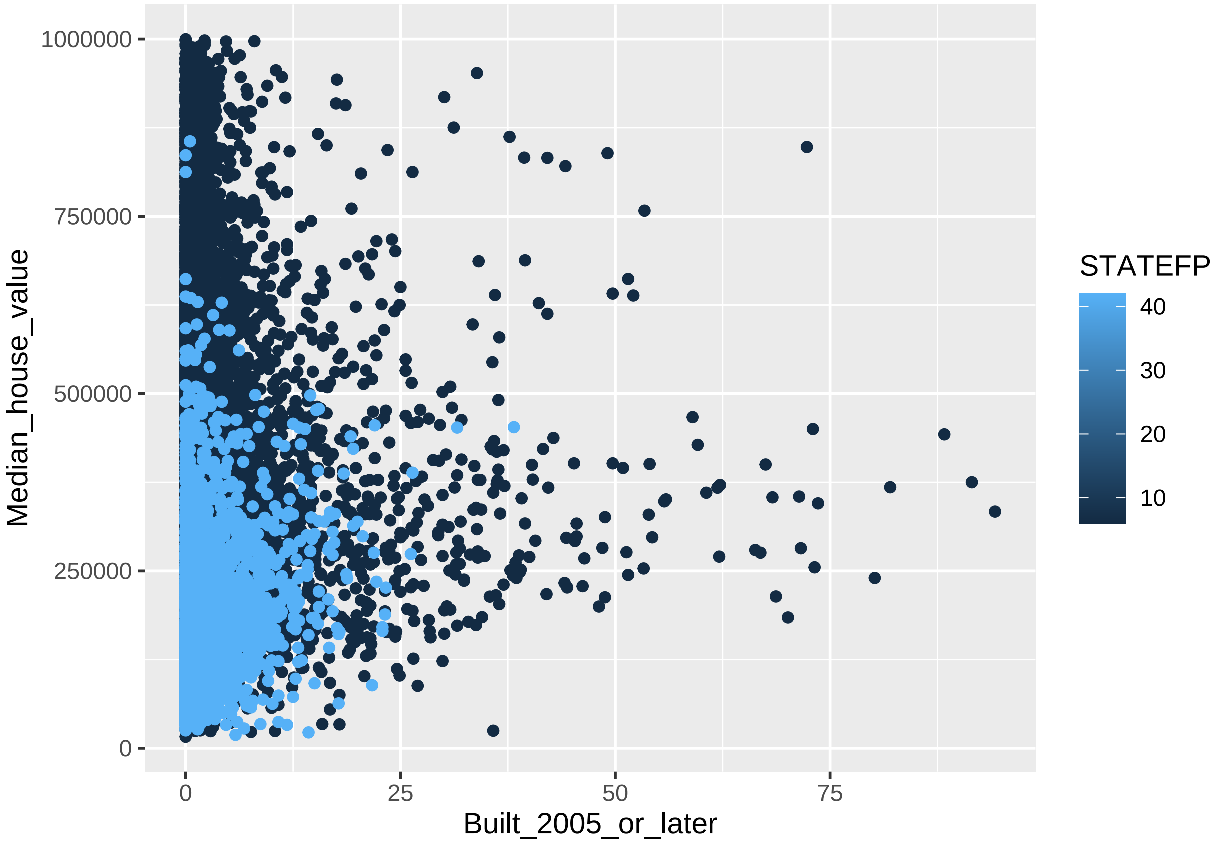


```
ca_pa%>%
  group_by(Built_2005_or_later)%>%
  summarize(median_hv=mean(Median_house_value))%>%
  ggplot(aes(Built_2005_or_later,median_hv))+
  geom_point()+
  geom_smooth()
```



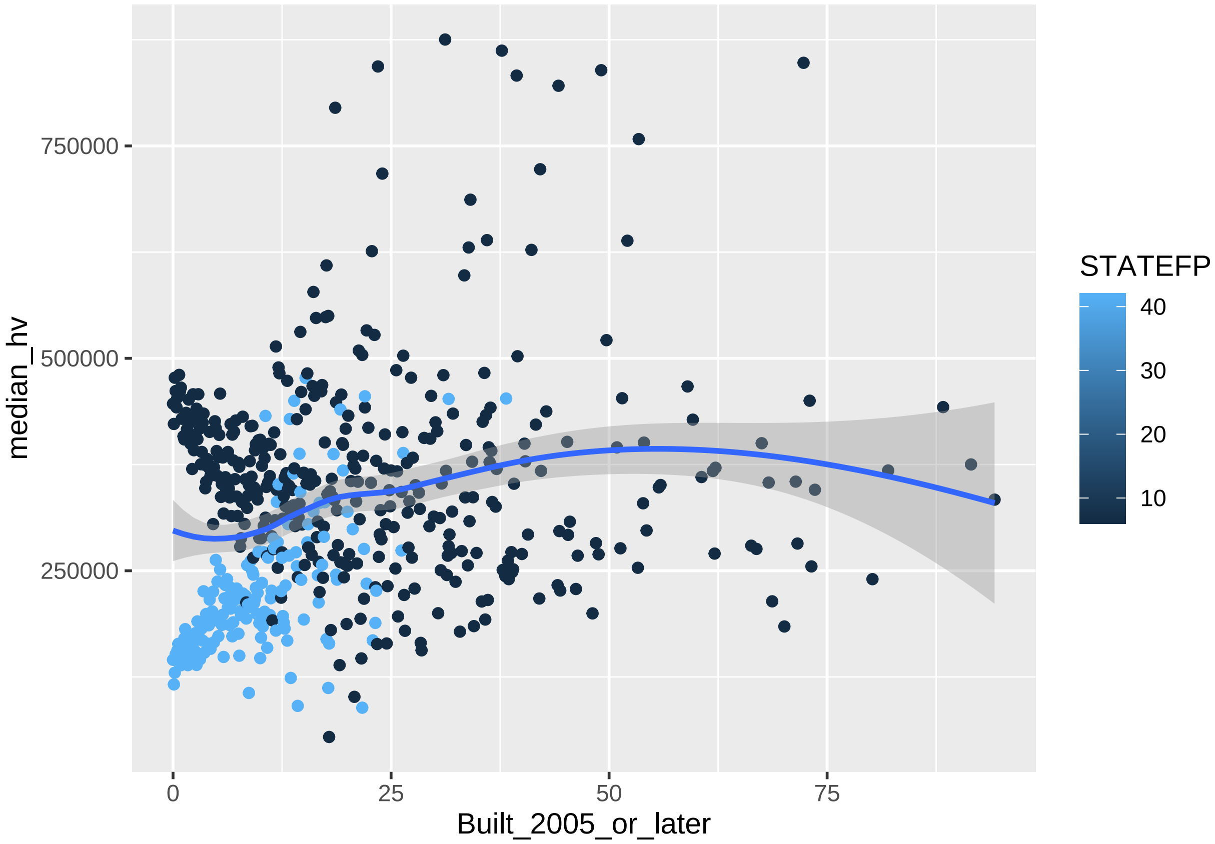
```
b. Make a new plot, or pair of plots, which breaks this out by state.  Note that the state is rec
```

```
ggplot(ca_pa)+
  geom_point(aes(x=Built_2005_or_later,y=Median_house_value,color=STATEFP))
```

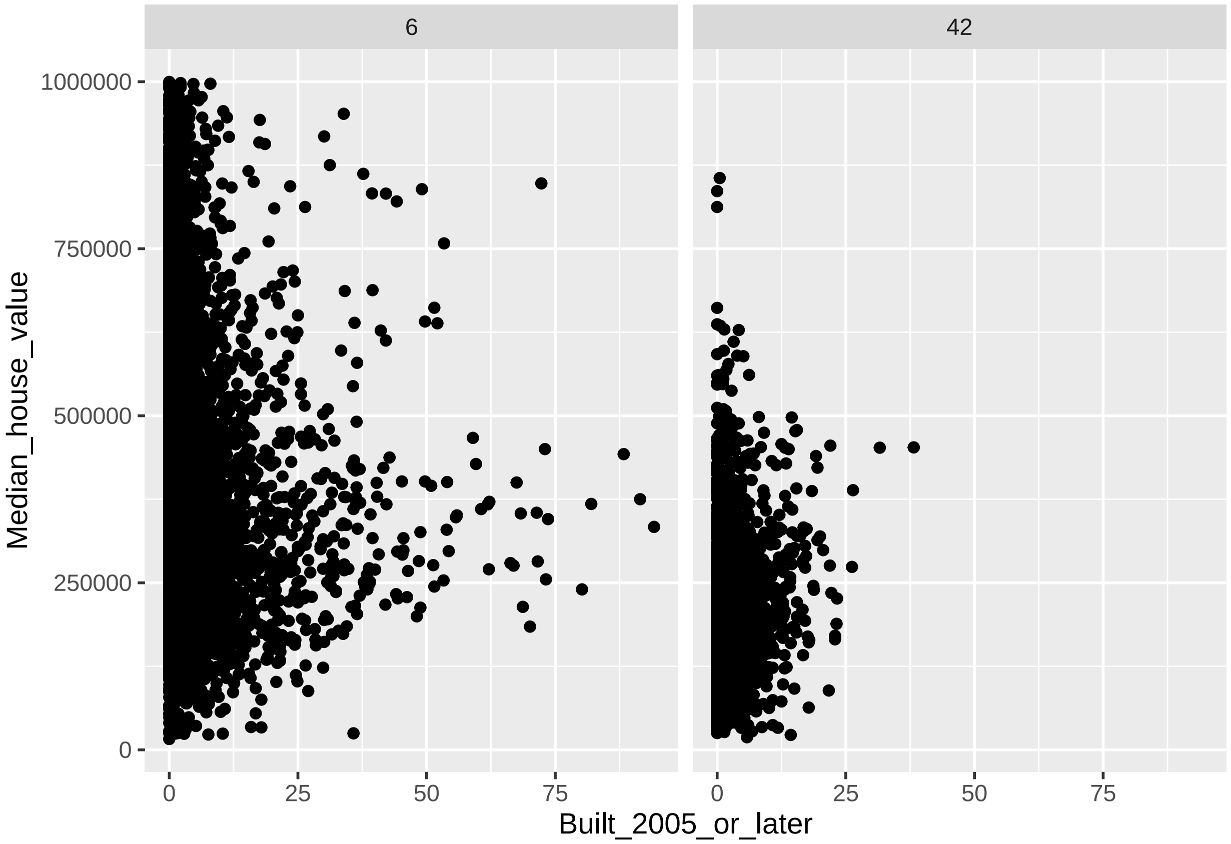


```
STATEFP<-as.factor(STATEFP)

ca_pa%>%
  group_by(Built_2005_or_later,STATEFP)%>%
  summarise(median_hv=mean(Median_house_value))%>%
  ggplot(aes(Built_2005_or_later,median_hv,color=STATEFP))+
  geom_point()+
  geom_smooth()
```



```
ggplot(ca_pa)+
  geom_point(aes(x=Built_2005_or_later,y=Median_house_value))+
  facet_wrap(~STATEFP)
```



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?

```
ca_pa <- ca_pa %>%
  mutate(vacancy_rate=Vacant_units/Total_units)
min(ca_pa[,"vacancy_rate"])
```

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

```
max(ca_pa[,"vacancy_rate"])
```

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

```
mean(ca_pa[,"vacancy_rate"])
```

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

```
median(ca_pa[,"vacancy_rate"])
```

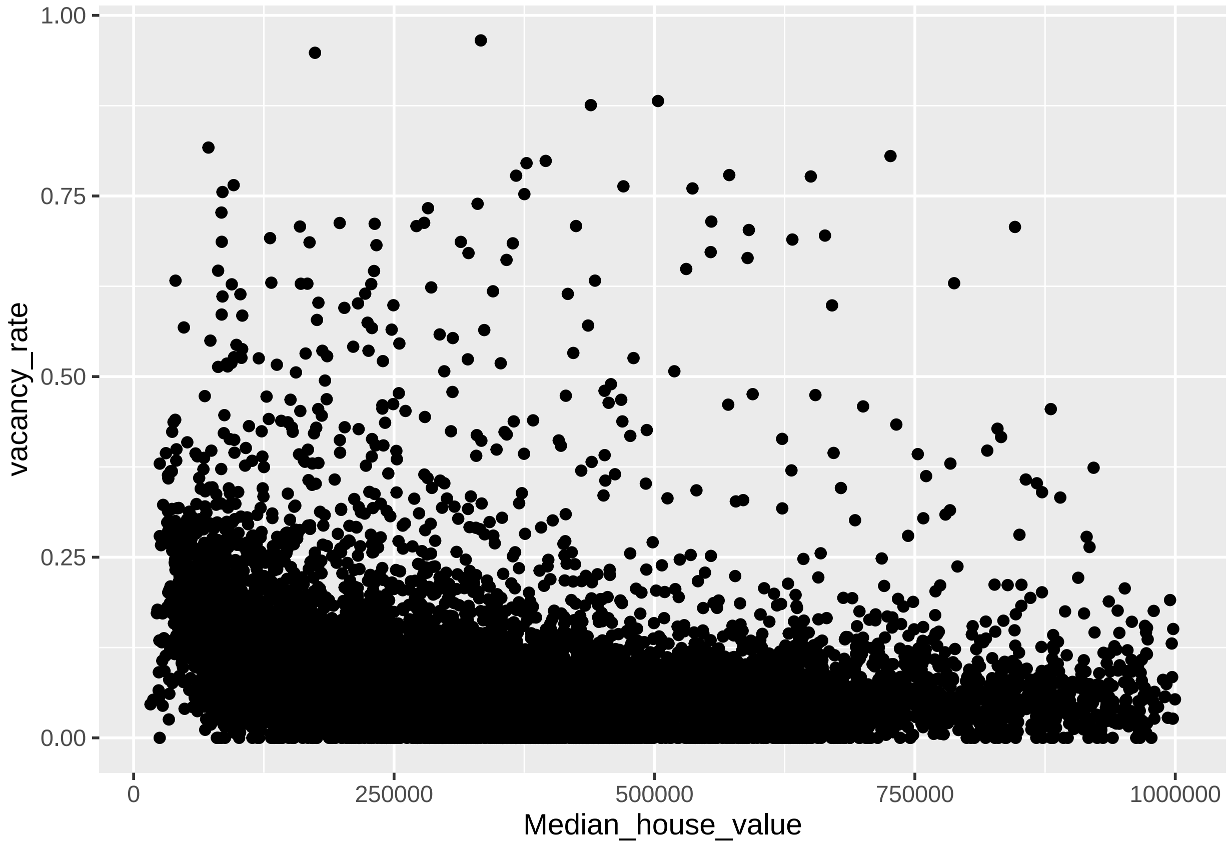
```
## [1] 0.06767283
```

```
summary(ca_pa$vacancy_rate)
```

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

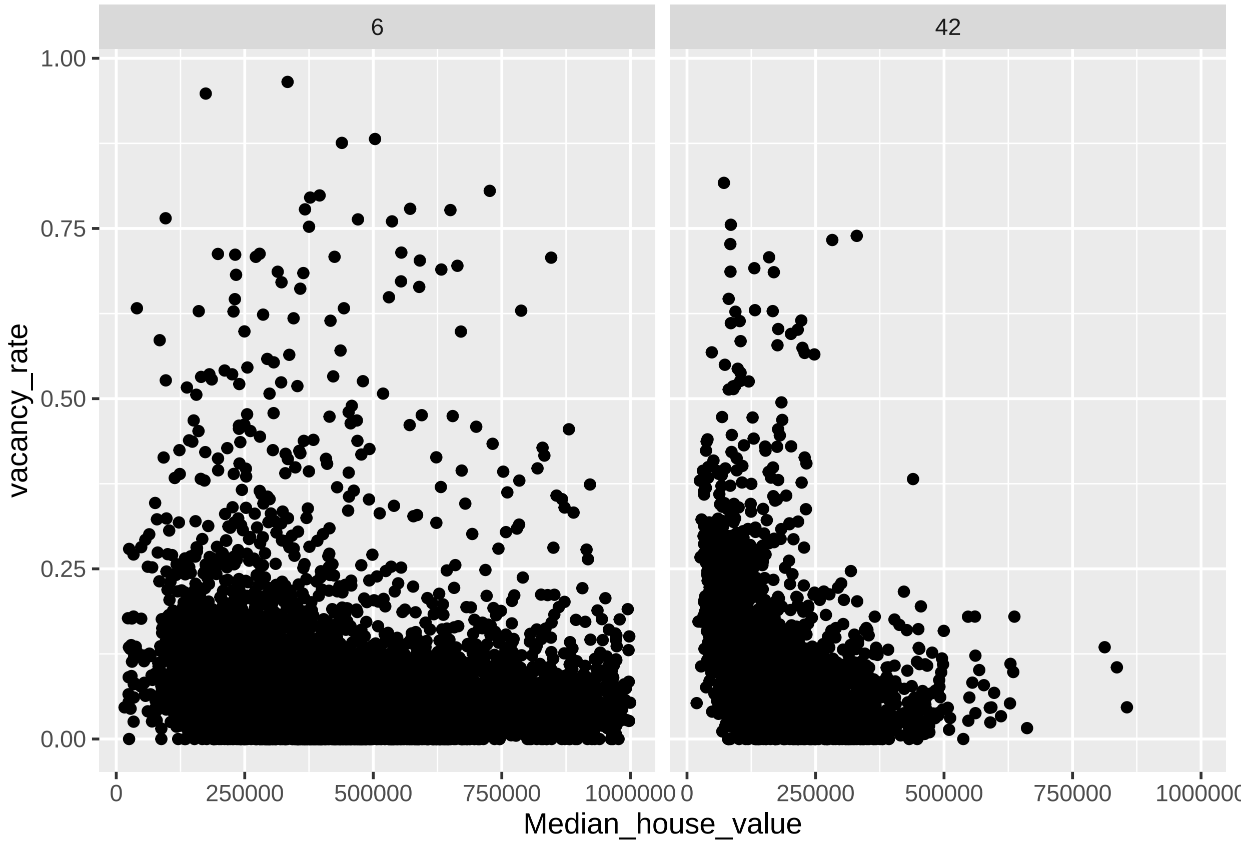
```
b. Plot the vacancy rate against median house value.
```

```
ggplot(ca_pa)+
  geom_point(aes(x=Median_house_value,y=vacancy_rate))
```



```
c. Plot vacancy rate against median house value separately for California and for Pennsylvania.
```

```
ggplot(ca_pa)+
  geom_point(aes(x=Median_house_value,y=vacancy_rate))+
  facet_wrap(~STATEFP)
```



YES.(with California being state 6 and Pennsylvania state 42.)Pennsylvania 的图在左下角集中，房价中位数较低，较高房价的空房率较低。而 California 的图对于房价中位数较为散布，房价中位数有高有低。

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)
```

先通过遍历 ca_pa 的行，通过查找 STATEFP 为 6 并且 COUNTYFP 为 1 的国家 Santa Clara，将这些观测值存储到向量 acca 中。再提取 Santa Clara 的每条数据的第十列 Median_house_value 房价中位数到向量 accamhv 中，并计算均值，从而得到 Santa Clara 该国的 Median_house_value 的中位数。

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.

```
ca_pa%>%
  filter(STATEFP==6,COUNTYFP==1)%>%
  dplyr::select(10)%>%
  unlist()%>%
  median()
```

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

```
c. For Alameda, Santa Clara and Allegheny Counties, what were the average percentages of housing
```

```
ca_pa%>%
  filter(STATEFP==6,COUNTYFP==1)%>%
  select("Built_2005_or_later")%>%
  unlist()%>%
  mean()
```

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

```
ca_pa%>%
  filter(STATEFP==6,COUNTYFP==85)%>%
  select("Built_2005_or_later")%>%
  unlist()%>%
  mean()
```

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

```
ca_pa%>%
  filter(STATEFP==42,COUNTYFP==3)%>%
  select("Built_2005_or_later")%>%
  unlist()%>%
  mean()
```

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

```
ca_2005<-
  ca_pa%>%
  filter((STATEFP==42&COUNTYFP==3)|(STATEFP==6&(COUNTYFP%in%c(1,85))))%>%
  group_by(COUNTYFP)%>%
  summarise(mean_2005=mean(Built_2005_or_later))
ca_2005
```

```
## # A tibble: 3 x 2
##   COUNTYFP mean_2005
##      <int>     <dbl>
## 1        1      2.82
## 2        3      1.47
## 3       85      3.20
```

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

```
mycor<-function(d){
  return(cor(d$Median_house_value,d$Built_2005_or_later))
}
mycor(ca_pa)
```

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

```
ca_pa%>%
  filter(STATEFP==6)%>%
  mycor()
```

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

```
ca_pa%>%
  filter(STATEFP==42)%>%
  mycor()
```

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

```
ca_pa%>%
  filter(STATEFP==6,COUNTYFP==1)%>%
  mycor()
```

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

```
ca_pa%>%
  filter(STATEFP==6,COUNTYFP==85)%>%
  mycor()
```

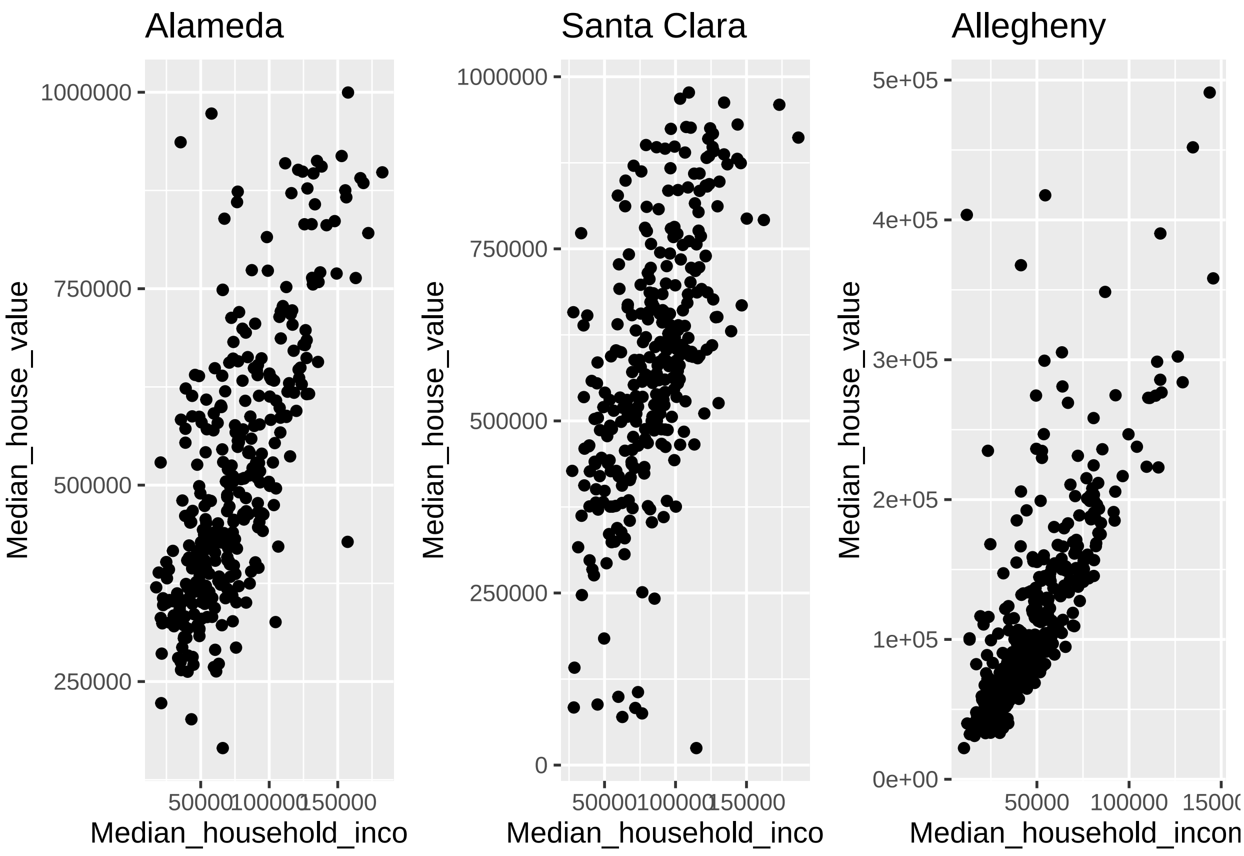
```
## [1] -0.1726203
```

```
ca_pa%>%
  filter(STATEFP==42,COUNTYFP==3)%>%
  mycor()
```

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

e. Make three plots, showing median house values against median income, for Alameda, Santa Clara,

```
library(ggpubr)
p1<-ca_pa%>%filter(STATEFP==6,COUNTYFP==1)%>%
  ggplot(aes(Median_household_income,Median_house_value))+
  geom_point()+ggtitle("Alameda")
p2<-ca_pa%>%filter(STATEFP==6,COUNTYFP==85)%>%
  ggplot(aes(Median_household_income,Median_house_value))+
  geom_point()+ggtitle("Santa Clara")
p3<-ca_pa%>%filter(STATEFP==42,COUNTYFP==3)%>%
  ggplot(aes(Median_household_income,Median_house_value))+
  geom_point()+ggtitle("Allegheny")
ggarrange(p1,p2,p3,ncol=3,nrow=1)
```



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() 使用交叉分类引自来建立每个因子水平组合的计数列联表。第一行：将性别作为一个因子 factor，具有两个水平 level:female 和 male。第二行：调换了两个水平的顺序。第三行：将 level 重定义为 Male 和 female，但是 Male 未赋值默认为 0。第四行：输出其他值 92，为原来 male 的计数。

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

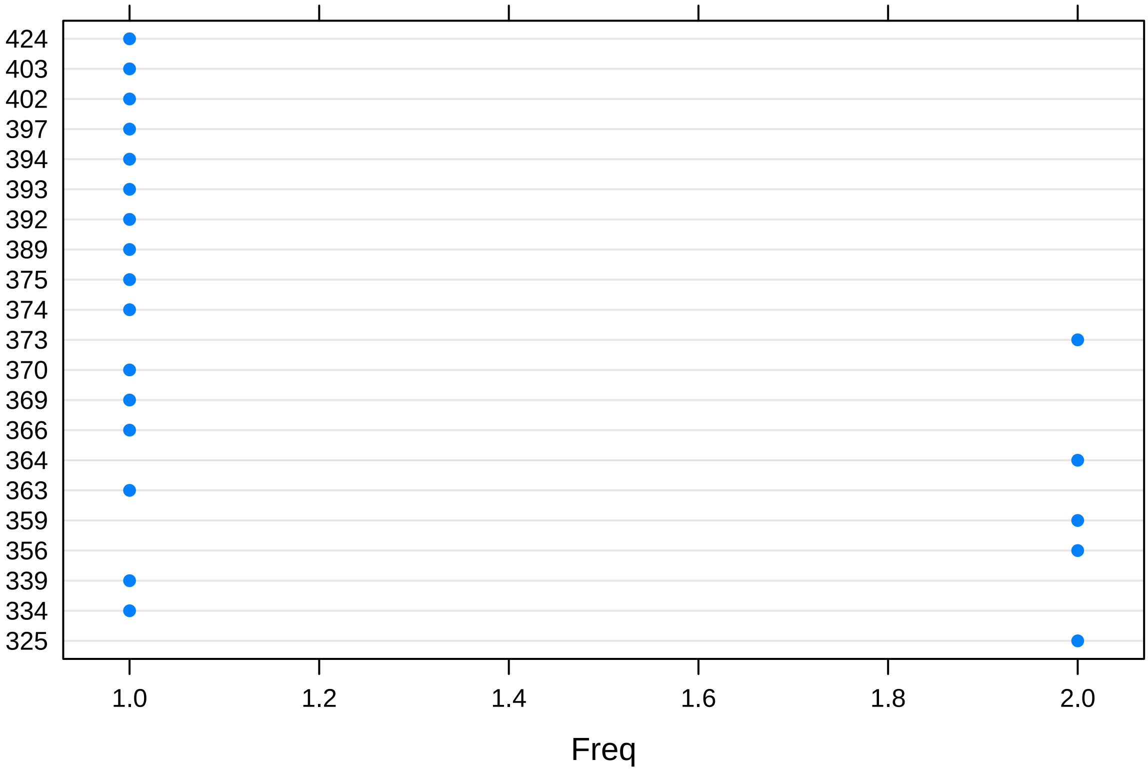
(a) Use the sequence of numbers 1, 2, . . . , 100 to check that this function gives the result that is expected.

```
cal_pro<-function(vec,cut){
  return(sum(vec>cut)/length(vec))
}
seq100<-seq(1,100)
cal_pro(seq100,50)
```

```
## [1] 0.5
```

(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)
dotplot(ex01.36)
```



```
cal_pro(unlist(ex01.36),7*60)
```

```
## [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

….

```
attach(Rabbit)
Dose <- unstack(Rabbit, Dose ~ Animal)[,1]
```

```
Treatment <- unstack(Rabbit, Treatment ~ Animal)[,1]
BPchange <- unstack(Rabbit, BPchange ~ Animal)
Rabbit.df <- data.frame(Treatment, Dose, BPchange)
Rabbit.df
```

```
##    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
## 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
```