homework2

xbw

2025-09-19

1. a.

```
ca_pa <- read.csv("E:/postgraduate/data science/mynotes/data/calif_penn_2011.csv")
```

b.

```
nrow(ca_pa) #[1] 11275
```

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

```
ncol(ca_pa) #[1] 34
```

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

c.统计每一列中元素是缺失值(NA)的数量

d.

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

e.omitted 10576 rows
f.(c)和(e)的答案不一致，因为(c)统计的是每一列有多少个元素是na，而(e)删除的是有na元素的行

2. a.

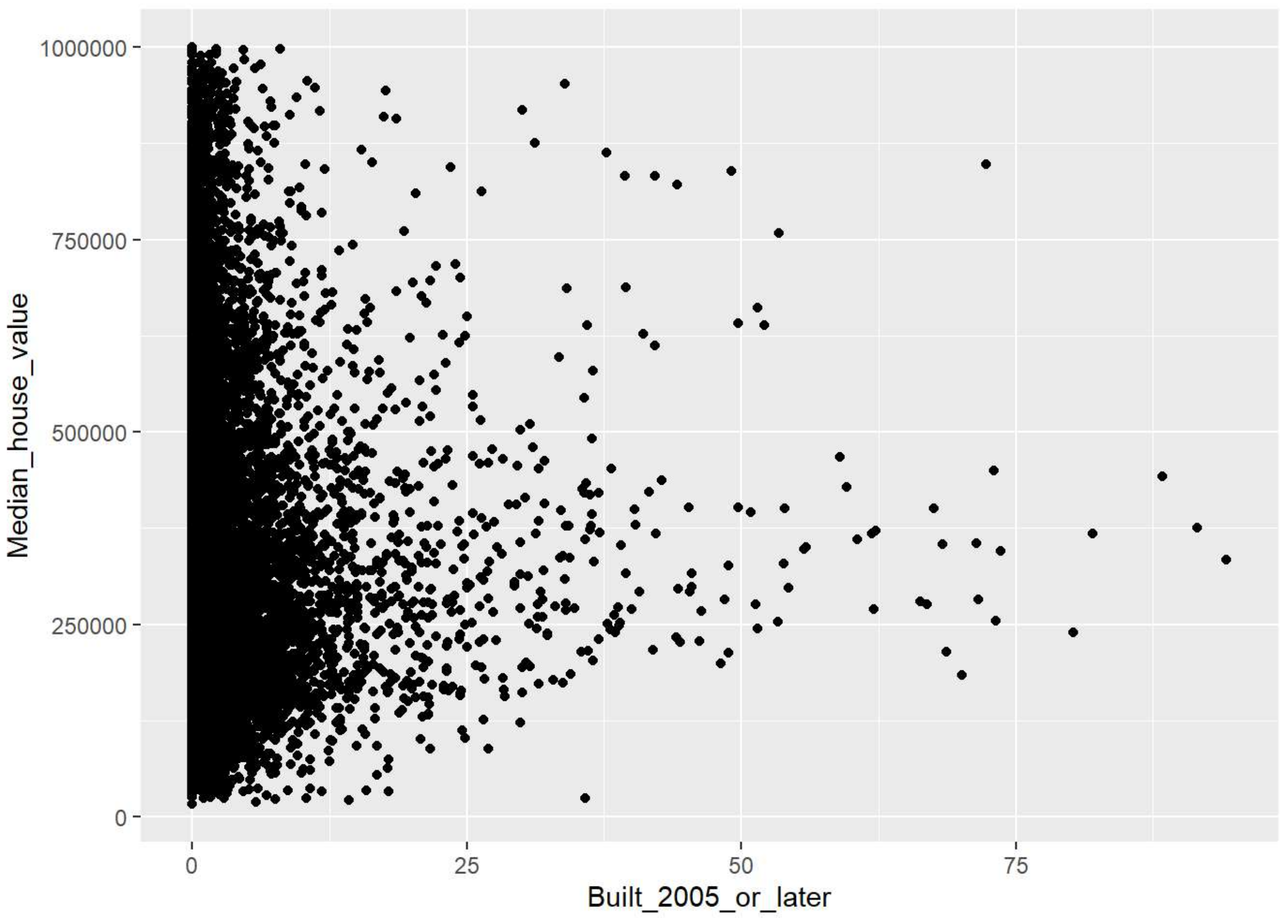
```
install.packages("tidyverse")
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\78471\AppData\Local\Temp\RtmpKa7YJP\downloaded_packages
```

```
library(tidyverse)
```

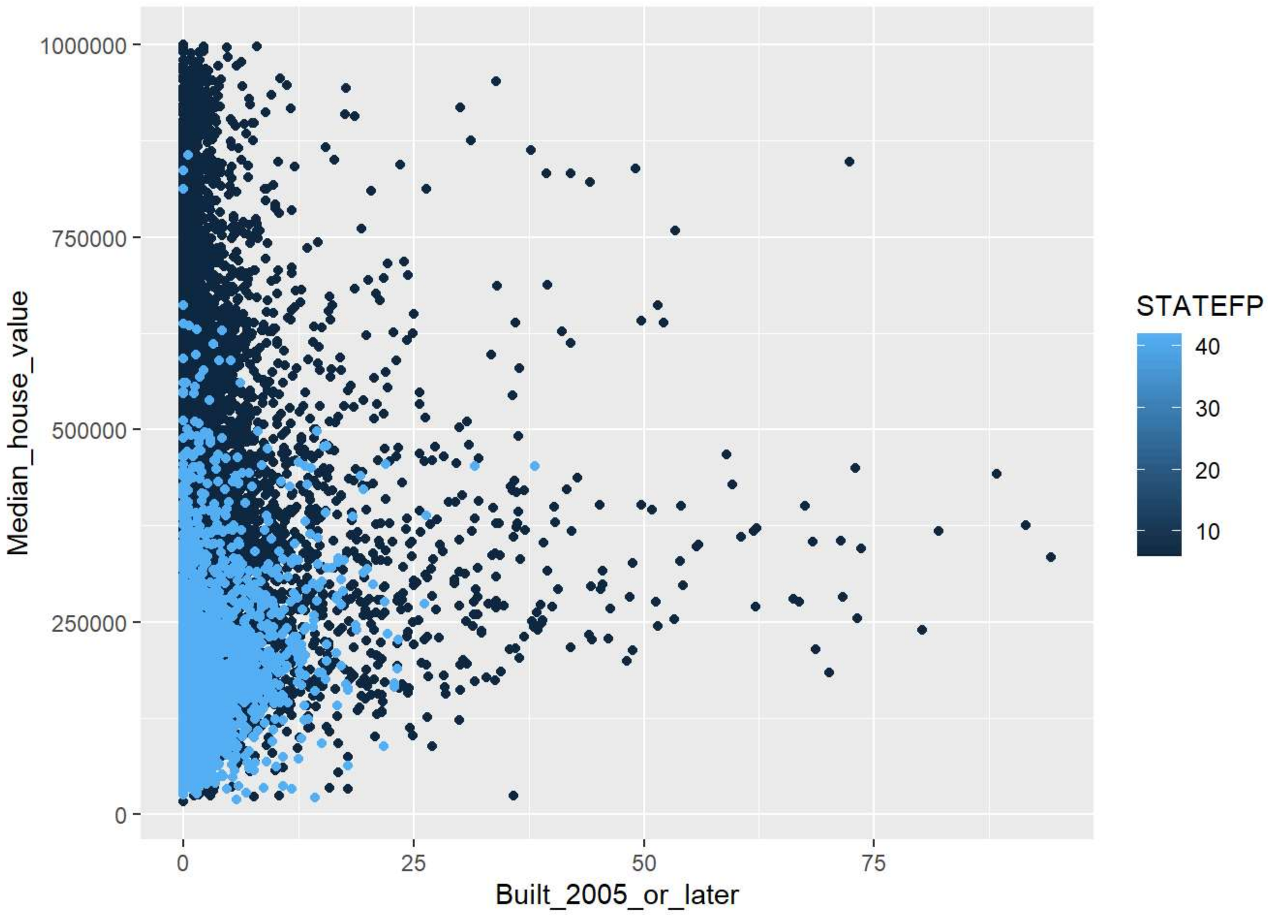
```
## ── Attaching core tidyverse packages ──────────────────────────────
── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.1     ✔ stringr   1.5.2
## ✔ ggplot2   4.0.0     ✔ tibble    3.3.0
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.1.0
## ── Conflicts ──────────────────────────────────────────────────────
────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

```
# Plot median house prices 与 Built_2005_or_later的关系
ca_pa_omitna |> ggplot()+aes(x=Built_2005_or_later,y=Median_house_value)+geom_point()+ la
bs( x = "Built_2005_or_later", y = "Median_house_value")
```



b.

```
#根据所在州对数据进行区分
ca_pa_omitna |> ggplot()+aes(x=Built_2005_or_later,y=Median_house_value,color=STATEFP)+ge
om_point()+ labs( x = "Built_2005_or_later", y = "Median_house_value")
```



c.

```
ca_pa$vacancy_rate <- ca_pa$Vacant_units/ca_pa$Total_units
```

3. a.

```
min(ca_pa$vacancy_rate,na.rm=TRUE) #0
```

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

```
max(ca_pa$vacancy_rate,na.rm=TRUE) #1
```

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

```
mean(ca_pa$vacancy_rate,na.rm=TRUE) #0.08917878
```

```
## [1] 0.08917878
```

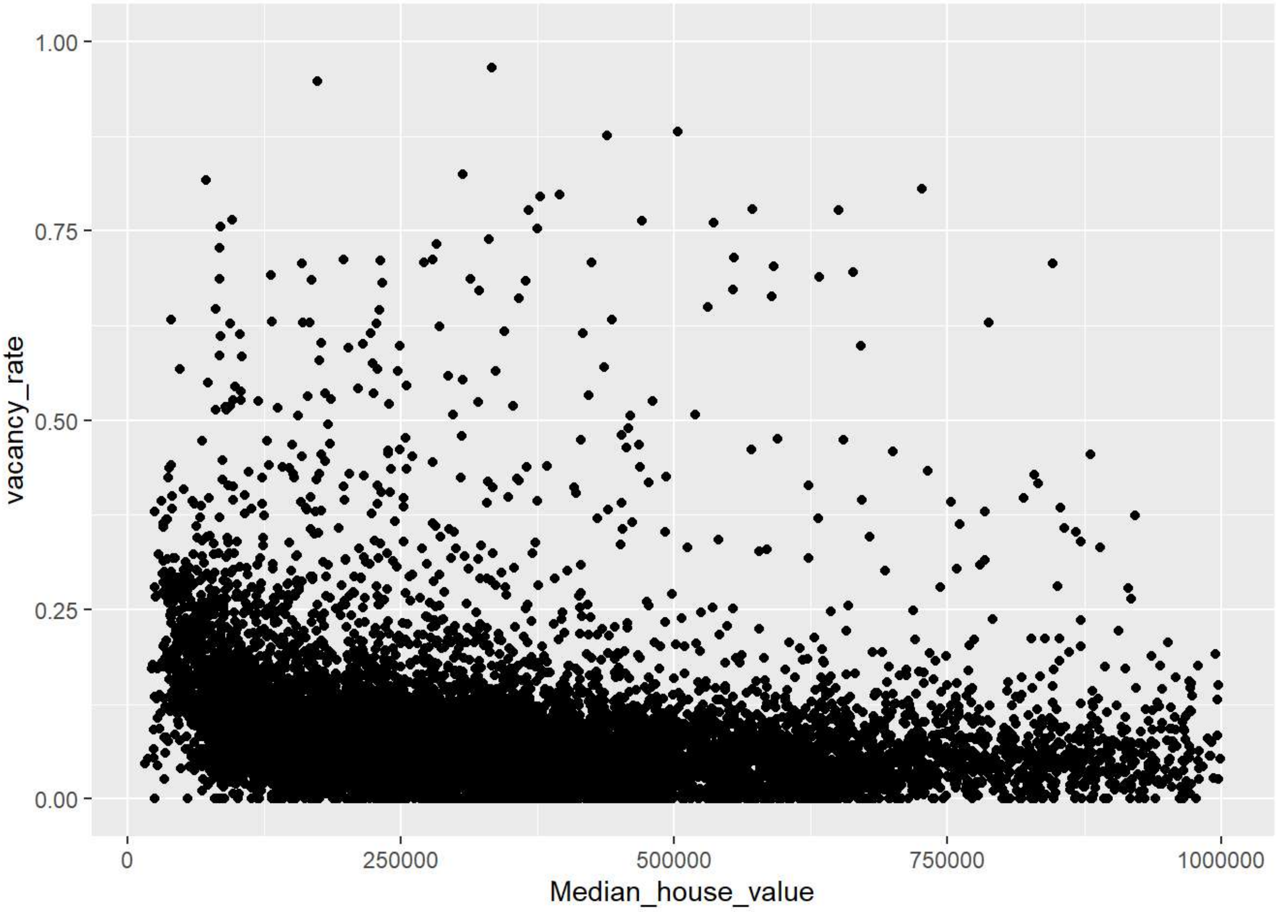
```
median(ca_pa$vacancy_rate,na.rm=TRUE) #0.06766326
```

```
## [1] 0.06766326
```

b.

```
ca_pa |> ggplot()+aes(x=Median_house_value,y=vacancy_rate)+geom_point()+labs(x="Median_ho
use_value",y="vacancy_rate")
```

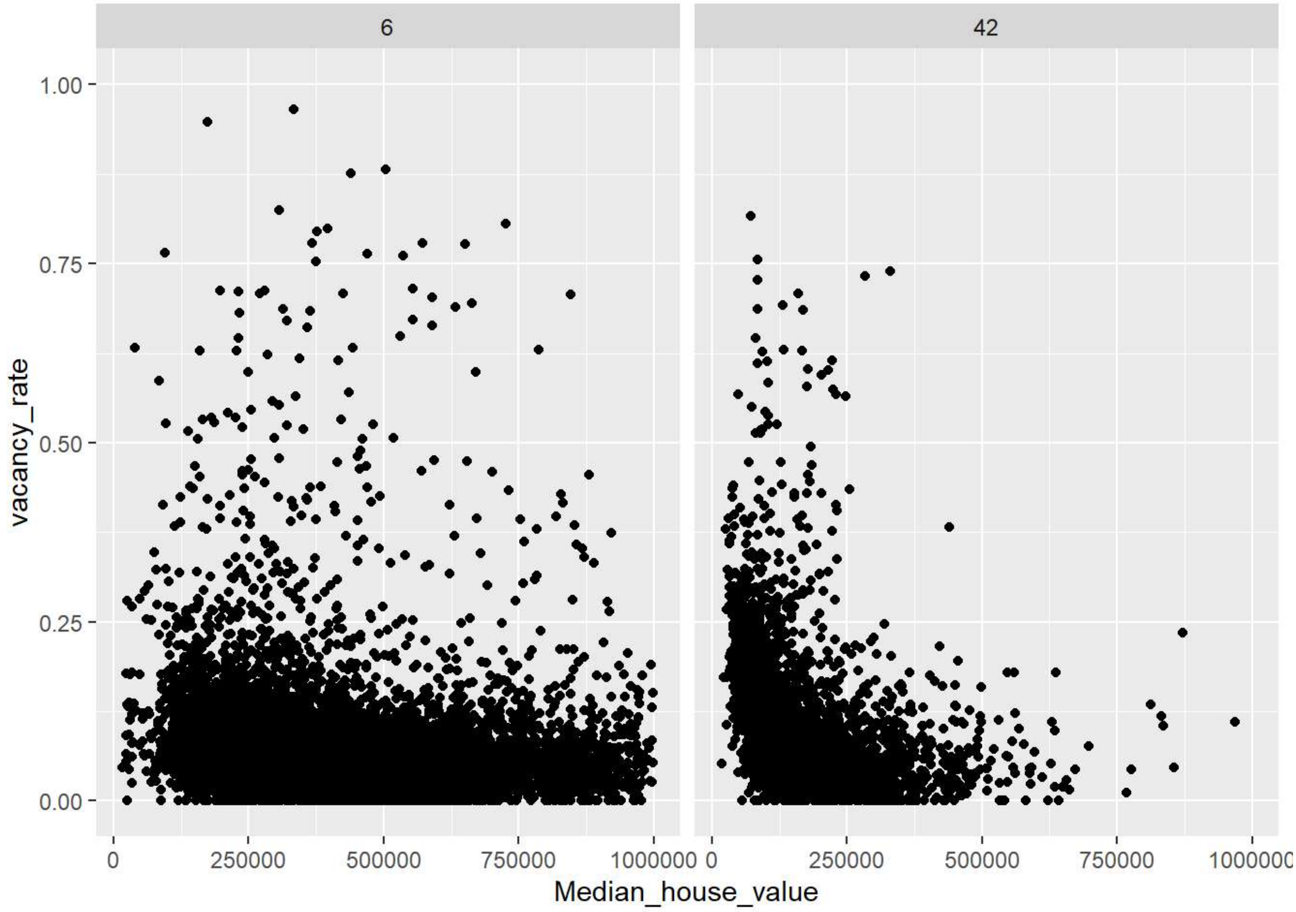
```
## Warning: Removed 599 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



c.

```
ca_pa |> ggplot()+aes(x=Median_house_value,y=vacancy_rate)+geom_point()+labs(x="Median_ho
use_value",y="vacancy_rate")+facet_wrap(~STATEFP)
```

```
## Warning: Removed 599 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
#宾夕法尼亚州中位房屋价值越高的区域空置率越低，而加利福尼亚州的房屋空置率分布比较均匀
```

4. a.

```
#这段代码要实现将Alameda County，California的median_house_value列成一个vector，并求中间值
acca <- c() #先创建一个空vector，用于存储行号
 for (tract in 1:nrow(ca_pa)) {#从第一行遍历到末尾
 if (ca_pa$STATEFP[tract] == 6) {
 if (ca_pa$COUNTYFP[tract] == 1) {
 acca <- c(acca, tract) #将STATEFP==6且COUNTYFP==1的行号添加到acca中
 }
 }
 }
 accamhv <- c() #创建空vector,
 for (tract in acca) { #遍历acca
 accamhv <- c(accamhv, ca_pa[tract,10]) #将每一行的第十列即median_house_value添加到accamh
v中
 }
 median(accamhv)#求Alameda County，California的median_house_value的中间值
```

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

b.

```
median(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1, 10], na.rm = TRUE)
```

```
## [1] 473500
```

c.

```
#Alameda County average percentages of housing built since 2005 is 2.932778
mean(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1,"Built_2005_or_later"],na.rm = TRUE)
```

```
## [1] 2.932778
```

```
# Santa Clara average percentages of housing built since 2005 is 3.160215
mean(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85,"Built_2005_or_later"],na.rm = TRUE)
```

```
## [1] 3.160215
```

```
#Allegheny County average percentages of housing built since 2005 is 1.883375
mean(ca_pa[ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3,"Built_2005_or_later"],na.rm = TRUE)
```

```
## [1] 1.883375
```

d.(i).the whole data

```
# the whole data -0.02052684
cor(ca_pa$Median_house_value,ca_pa$Built_2005_or_later,use = "complete.obs")
```

```
## [1] -0.02052684
```

(ii). all of California

```
# 0.2339447
 cor(ca_pa[ca_pa$STATEFP == 42,c("Median_house_value","Built_2005_or_later")],use = "comp
lete.obs")
```

```
##                       Median_house_value Built_2005_or_later
## Median_house_value             1.0000000           0.2339447
## Built_2005_or_later            0.2339447           1.0000000
```

(iii). all of Pennsylvania

```
# -0.1160322
cor(ca_pa[ca_pa$STATEFP == 6,c("Median_house_value","Built_2005_or_later")],use = "comple
te.obs")
```

```
##                       Median_house_value Built_2005_or_later
## Median_house_value             1.0000000          -0.1160322
## Built_2005_or_later           -0.1160322           1.0000000
```

(iv). Alameda County

```
#  0.01432789
cor(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1,c("Median_house_value","Built_2005_or_
later")],use = "complete.obs")
```

```
##                       Median_house_value Built_2005_or_later
## Median_house_value          1.00000000          0.01432789
## Built_2005_or_later         0.01432789          1.00000000
```

(v). Santa Clara County

```
# -0.1726203
cor(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85,c("Median_house_value","Built_2005_or_later")],use = "complete.obs")
```

```
##                       Median_house_value Built_2005_or_later
## Median_house_value           1.0000000          -0.1726203
## Built_2005_or_later         -0.1726203           1.0000000
```

(vi). Allegheny County

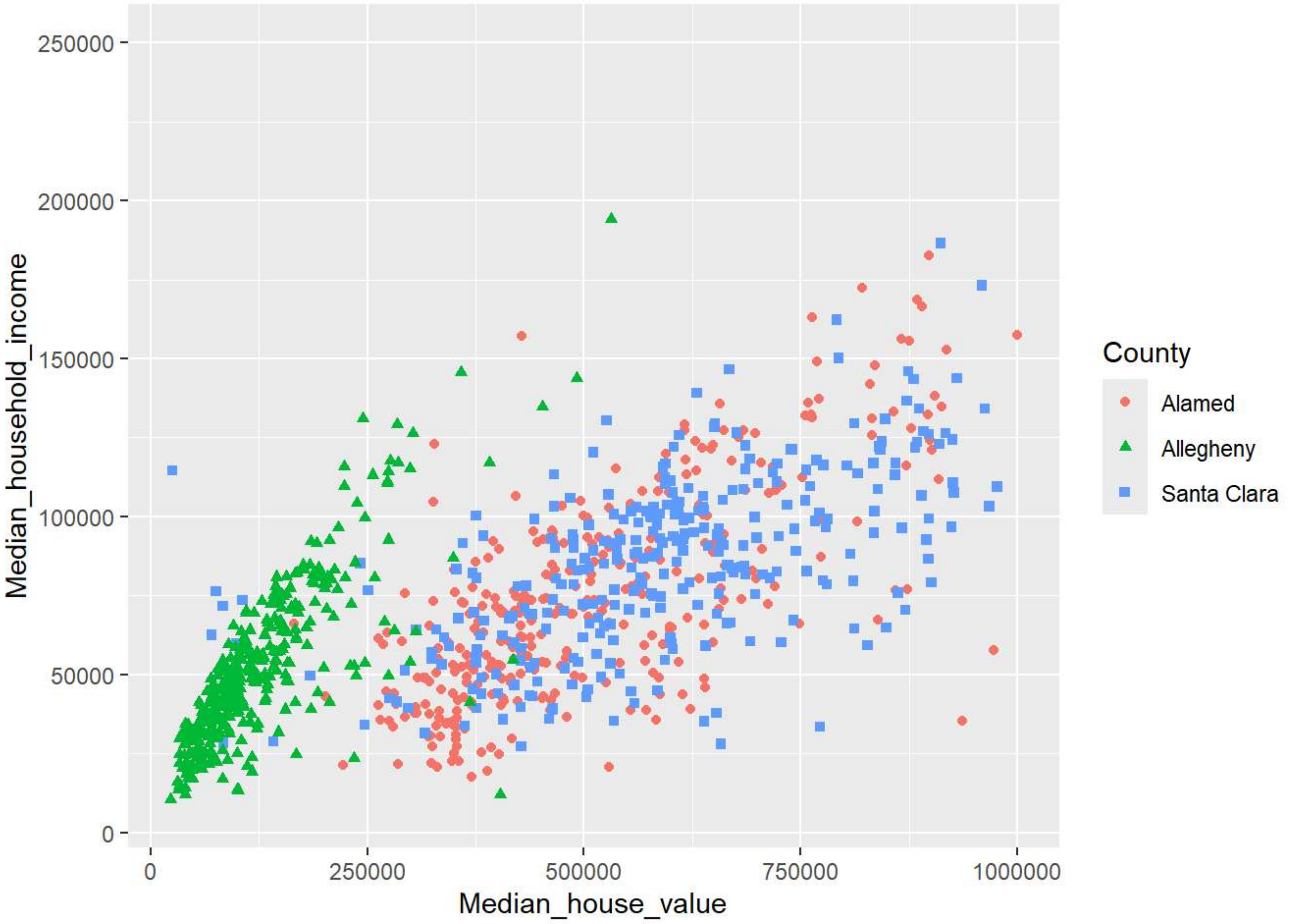
```
# 0.1868602
cor(ca_pa[ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3,c("Median_house_value","Built_2005_or_later")],use = "complete.obs")
```

```
##                       Median_house_value Built_2005_or_later
## Median_house_value           1.0000000           0.1868602
## Built_2005_or_later          0.1868602           1.0000000
```

e.

```
#先将三个城市的信息提取出来
ca_pa_sub <- ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1 |ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85 | ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3,c("STATEFP","COUNTYFP","Median_house_value","Median_household_income")]
#添加城市列名
ca_pa_sub$County <- ifelse(ca_pa_sub$STATEFP == 6 & ca_pa_sub$COUNTYFP == 1, "Alamed", ifelse(ca_pa_sub$STATEFP == 6 & ca_pa_sub$COUNTYFP == 85, "Santa Clara","Allegheny"))
#画出三个城市的图，用County区分
ca_pa_sub |> ggplot()+aes(x=Median_house_value,y=Median_household_income,color=County,shape = County)+geom_point() + labs(x="Median_house_value",y="Median_household_income",color="County",shape = "County")
```

```
## Warning: Removed 91 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



MB.Ch1.11.

```
#将"female"重复91次，"male"重复92次连接成一个字符串向量，将字符串向量转化为一个因子
gender <- factor(c(rep("female", 91), rep("male", 92)))
 table(gender)#统计gender因子中的每个水平的频次
```

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

```
 gender <- factor(gender, levels=c("male", "female")) #将因子gender的水平顺序从默认顺序改为指定的c("male", "female")顺序

 gender <- factor(gender, levels=c("Male", "female"))
 #"male"无法匹配"Male"，所以被转为NA
 table(gender, exclude=NULL)
```

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

```
 #exclude = NULL：强制显示所有水平，包括NA的计数，上面92个"male"被转换为了NA，所以显示92个NA
```

MB.Ch1.12.

```
proportion_fun <- function(x,cutoff){
  #统计x中大于cutoff的数量
  above_cutoff <- sum(x>cutoff,na.rm=TRUE)
  #统计x的总数
  total <- sum(!is.na(x))
  return (above_cutoff/total)#返回超过cutoff的比例
}
```

(a).

```
test1 <- 1:100
proportion_fun(test1,25) #0.75
```

```
## [1] 0.75
```

```
proportion_fun(test1,76) #0.24
```

```
## [1] 0.24
```

(b).

```
install.packages("Devore7")
```

```
## package 'Devore7' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\78471\AppData\Local\Temp\RtmpKa7YJP\downloaded_packages
```

```
library(Devore7)
```

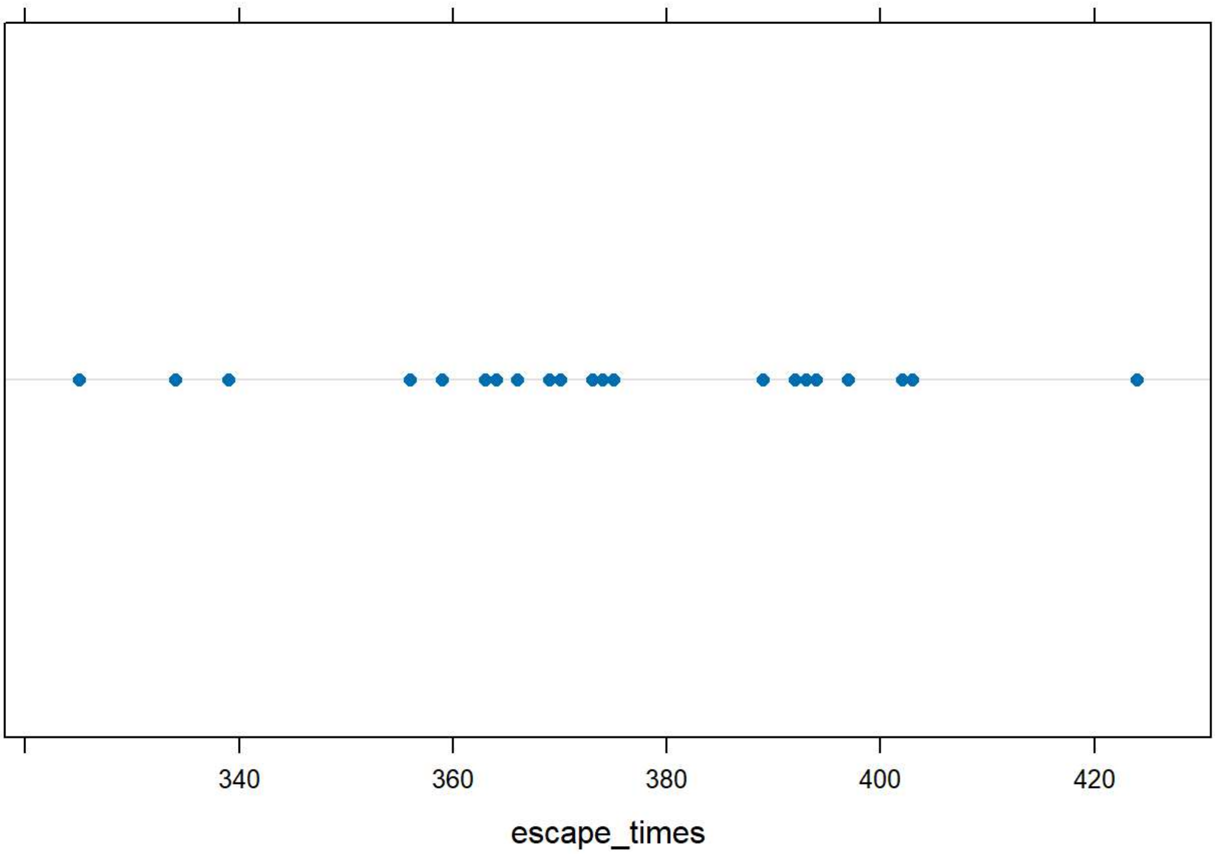
```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

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

```
## Loading required package: lattice
```

```
data(ex01.36)
escape_times <- ex01.36
dotplot(~escape_times)
```




```
proportion_fun(escape_times,420) #0.03846154
```

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

MB.Ch1.18.

```
install.packages("MASS")
```

```
## Warning: package 'MASS' is in use and will not be installed
```

```
library(MASS)
#将Rabbit根据Treatment、Dose、Animal的优先级排序
rabbit_sorted <- Rabbit[order(Rabbit$Treatment, Rabbit$Dose, Rabbit$Animal), ]
# 使用unstack将Animal展开为列
rabbit_unstack1 <- unstack(rabbit_sorted, BPchange ~ Animal)
# 提取唯一的Treatment和Dose组合
treatment_dose_combos <- unique(rabbit_sorted[, c("Treatment", "Dose")])
# 将treatment_dose_combos 和 rabbit_unstack1 组合成为最终结果
rabbit_unstack <- cbind(treatment_dose_combos,rabbit_unstack1)
print(rabbit_unstack)
```

```
##    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
```