homework2 report

王梓

2024-07-01

1

a.

```
ca_pa <- read.csv("data/calif_penn_2011.csv", header = T)
```

b.

```
nrow(ca_pa)
```

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

```
ncol(ca_pa)
```

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

c.

```
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
```

这行代码计算出了每一列中失效值(missing value)的数量。 首先 `apply(ca_pa, c(1, 2), is.na)` 表示对 ca_pa 中每个值判断它是否是 NA，然后使用 `colSums` 把每列中是 NA 的数据的数量加起来。

d.

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

e.

```
nrow(ca_pa) - nrow(ca_pa_cleaned)
```

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

因此 `na.omit` 操作清除了670行数据。

f.

```
sum(colSums(apply(ca_pa,c(1,2),is.na)))
```

```
## [1] 3034
```

```
max(colSums(apply(ca_pa,c(1,2),is.na)))
```

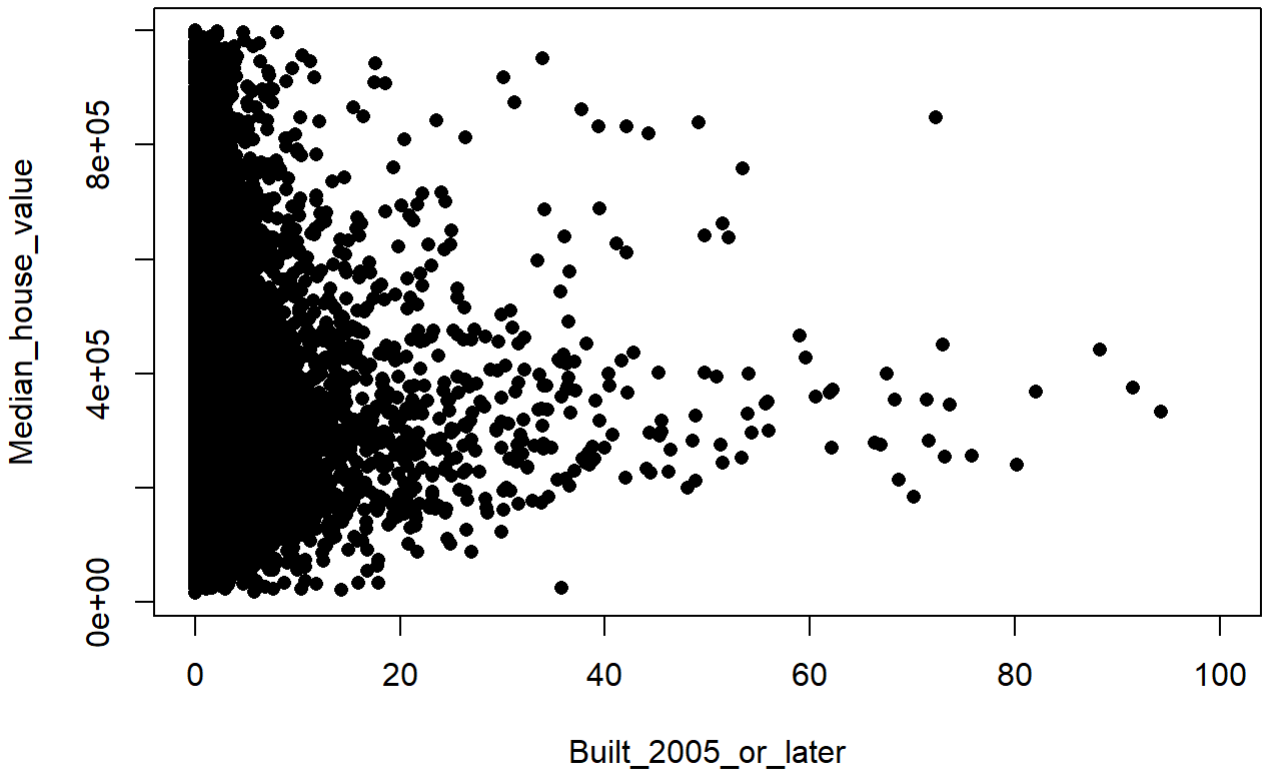
```
## [1] 599
```

c中和e中的结果是一致的，因为清除的行数大于等于列中无效数据的最大值，且小于所有列中无效数据的总和（每一行中可能有多列有无效数据）。

2.

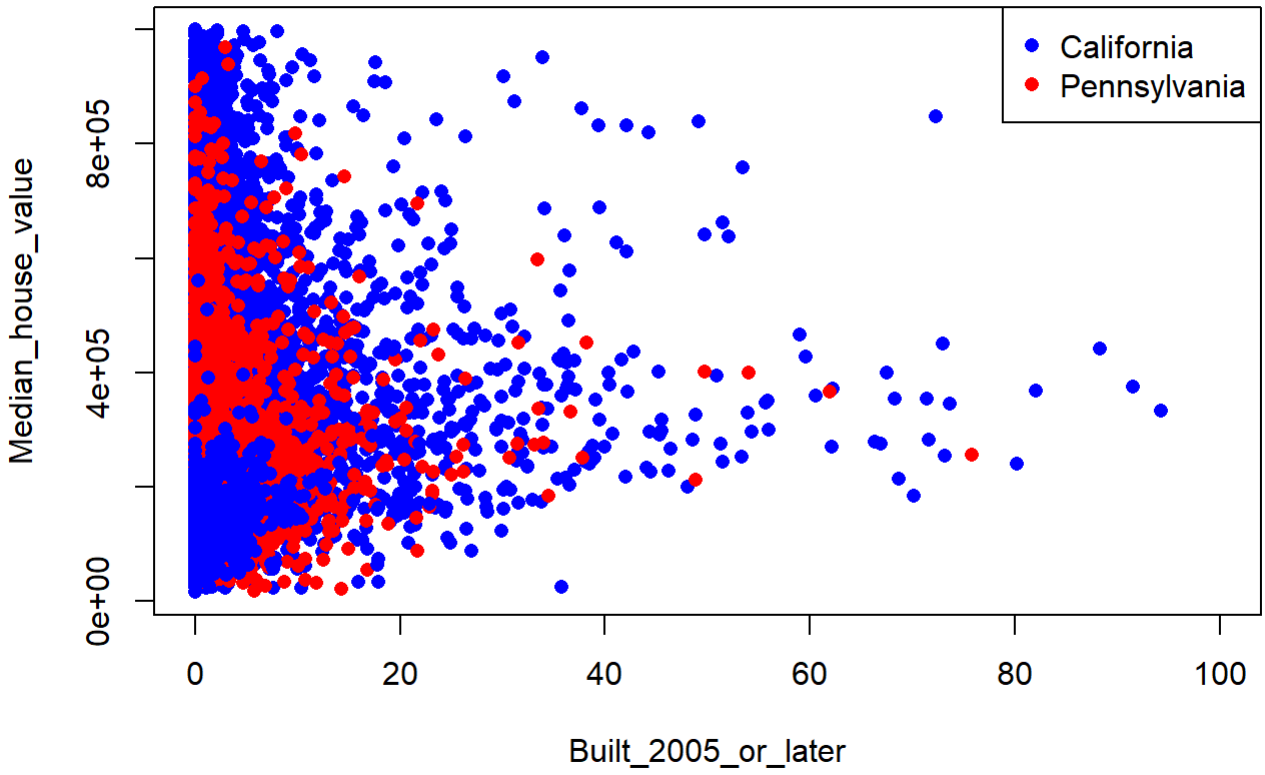
a.

```
plot(ca_pa$Median_house_value ~ ca_pa$Built_2005_or_later,
     pch = 16, xlab = "Built_2005_or_later", ylab = "Median_house_value")
```



b.

```
plot(ca_pa$Median_house_value ~ ca_pa$Built_2005_or_later,
     pch = 16,
     xlab = "Built_2005_or_later", ylab = "Median_house_value",
     col = ifelse(ca_pa_cleaned$STATEFP == 6, "blue", "red"))
legend("topright",
       legend = c("California", "Pennsylvania"),
       col = c("blue", "red"),
       pch = 16)
```



3.

a.

```
ca_pa$Vacant_rate <- ca_pa$Vacant_units / ca_pa$Total_units
min(ca_pa$Vacant_rate, na.rm = T)
```

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

```
max(ca_pa$Vacant_rate, na.rm = T)
```

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

```
mean(ca_pa$Vacant_rate, na.rm = T)
```

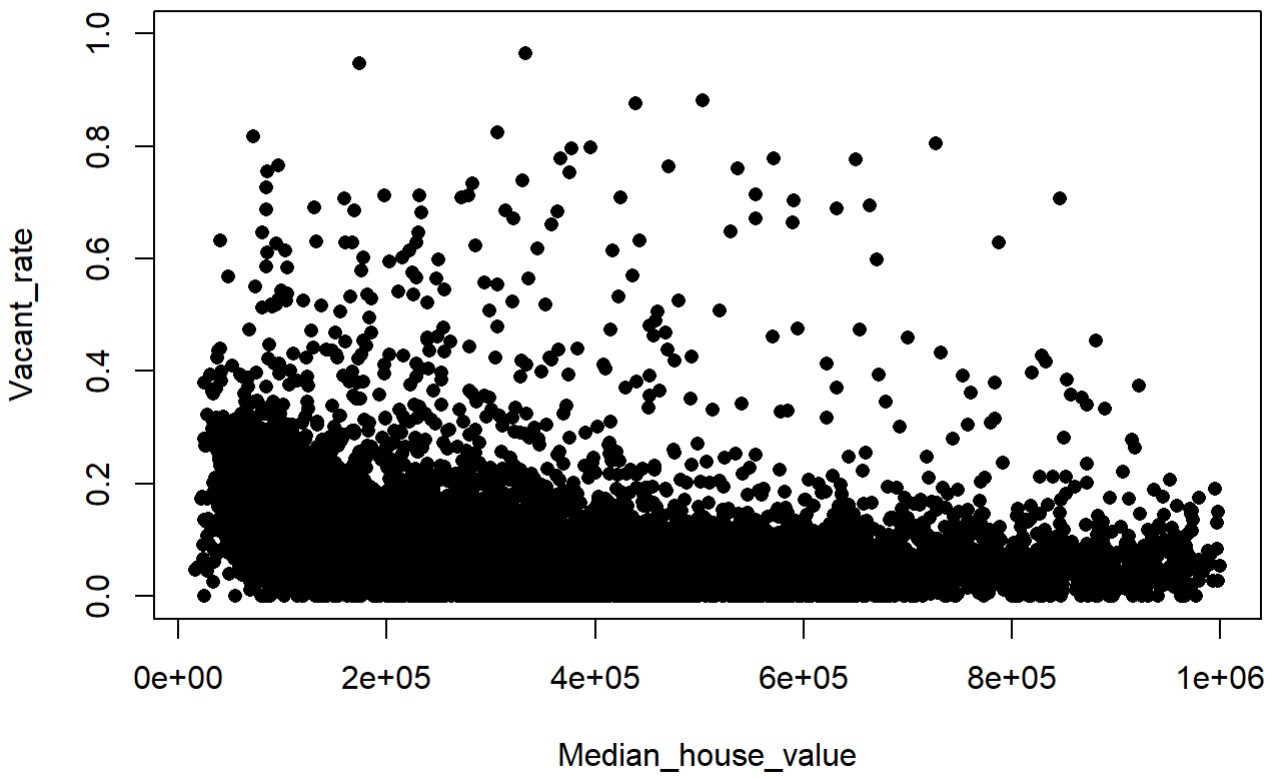
```
## [1] 0.08917878
```

```
median(ca_pa$Vacant_rate, na.rm = T)
```

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

b.

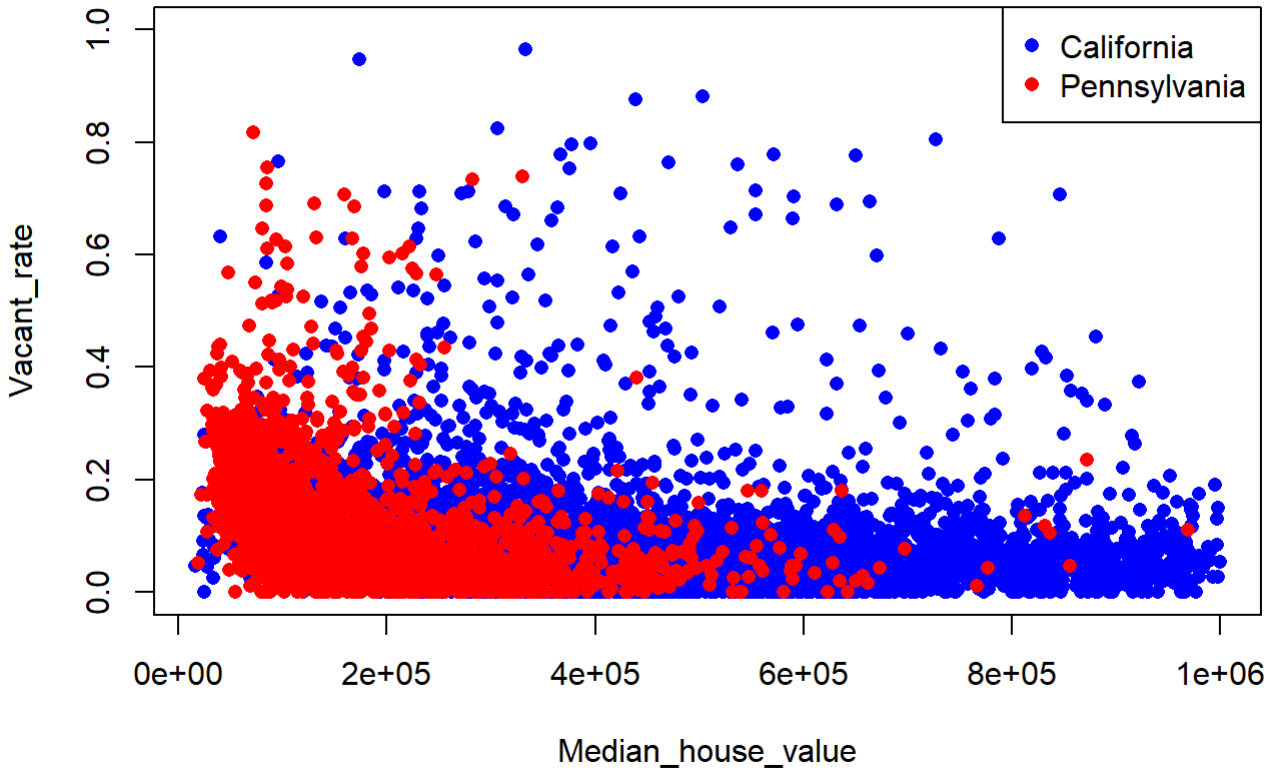
```
plot(ca_pa$Vacant_rate ~ ca_pa$Median_house_value,
     pch = 16,
     xlab = "Median_house_value", ylab = "Vacant_rate")
```



c.

```
plot(ca_pa$Vacant_rate ~ ca_pa$Median_house_value,
     pch = 16,
     xlab = "Median_house_value", ylab = "Vacant_rate",
     col = ifelse(ca_pa$STATEFP == 6, "blue", "red"))

legend("topright",
       legend = c("California", "Pennsylvania"),
       col = c("blue", "red"),
       pch = 16)
```



California 房价中位数较低的区域更多，空置率相较于 Pennsylvania 来说更少。

4

a.

这段代码在计算 Alameda, California 中每个 Census tract 的 Median_house_value 的中位数。

代码中使用 `acca` 来储存属于 Alameda, California 的数据行数，使用 for 循环来得到 `acca`。 `accamhv` 中储存的是 Alameda, California 的每个 Census tract 的 Median_house_value，通过访问 `acca` 中储存的对应行的数据来得到。 最后使用 `median()` 来得到其中位数。

b.

```
median(ca_pa[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1, "Median_house_value"], na.rm = T)
```

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

c.

```
percentages_of_housing_built_since_2005 <-
  ca_pa$Built_2005_or_later / ca_pa$Total_units
mean(percentages_of_housing_built_since_2005
     [(ca_pa$STATEFP == 6 & ca_pa$COUNTYFP %in% c(1, 85))
       | (ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3)],
     na.rm = T)
```

```
## [1] 0.0119391
```

d.

(i)

```
cor(ca_pa$Median_house_value, percentages_of_housing_built_since_2005, use = "complete.obs")
```

```
## [1] -0.00904388
```

(ii)

```
California <- ca_pa$STATEFP == 6
cor(ca_pa$Median_house_value[California], percentages_of_housing_built_since_2005[California],
use = "complete.obs")
```

```
## [1] -0.06328907
```

(iii)

```
Pennsylvania <- ca_pa$STATEFP == 42
cor(ca_pa$Median_house_value[Pennsylvania], percentages_of_housing_built_since_2005[Pennsylvani
a], use = "complete.obs")
```

```
## [1] 0.03593896
```

(iv)

```
Alameda <- ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1
cor(ca_pa$Median_house_value[Alameda], percentages_of_housing_built_since_2005[Alameda], use =
"complete.obs")
```

```
## [1] -0.00917091
```

(v)

```
Santa_Clara <- ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 85
cor(ca_pa$Median_house_value[Santa_Clara], percentages_of_housing_built_since_2005[Santa_Clar
a], use = "complete.obs")
```

```
## [1] -0.1732909
```

(vi)

```
Allegheny <- ca_pa$STATEFP == 42 & ca_pa$COUNTYFP == 3
cor(ca_pa$Median_house_value[Allegheny], percentages_of_housing_built_since_2005[Allegheny], us
e = "complete.obs")
```

```
## [1] 0.09210483
```

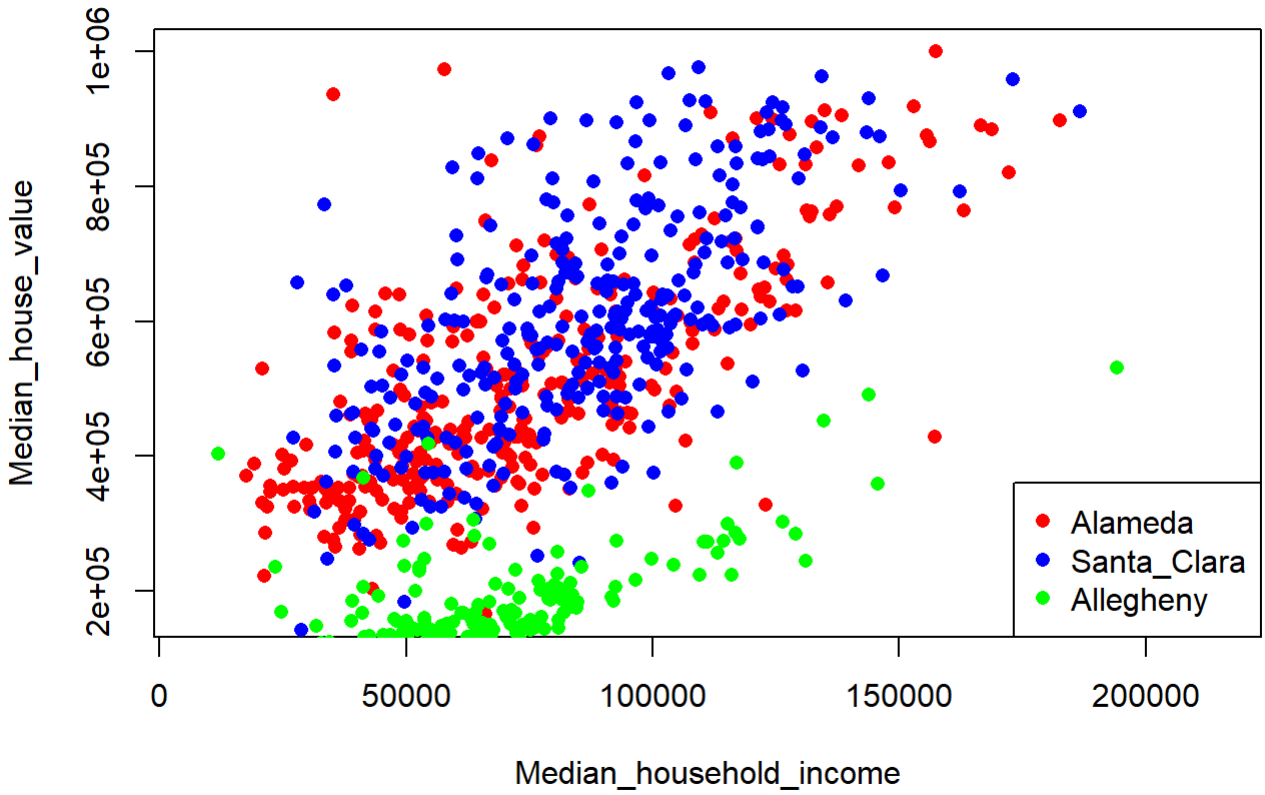
e.

```
plot(ca_pa$Median_house_value[Alameda] ~ ca_pa$Median_household_income[Alameda],
     pch = 16,
     xlab = "Median_household_income", ylab = "Median_house_value",
     col = "red")

points(ca_pa$Median_house_value[Santa_Clara] ~ ca_pa$Median_household_income[Santa_Clara],
     pch = 16,
     xlab = "Median_household_income", ylab = "Median_house_value",
     col = "blue")


points(ca_pa$Median_house_value[Allegheny] ~ ca_pa$Median_household_income[Allegheny],
     pch = 16,
     xlab = "Median_household_income", ylab = "Median_house_value",
     col = "green")

legend("bottomright",
       legend = c("Alameda", "Santa_Clara", "Allegheny"),
       col = c("red", "blue", "green"),
       pch = 16)
```



MB.Ch1.11

```
gender <- factor(c(rep("female", 91), rep("male", 92)))
table(gender)
```

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

使用 table() 统计 female 和 male 出现的频数。

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

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

先使用 factor() 将 female 和 male 的顺序互换，再调用 table()。

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

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

gender中没有 Male，因此 Male 的频数统计为0。

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

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

在上一步的操作中，`levels=c("Male", "female")` 中没有包含 male，因此 table 中原为 male 的数据变为 NA，在设置 `exclude=NULL` 后，可以看到 gender 中有 92 个 NA。

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

MB.Ch1.12.

```
exceeded_proportion <- function(x, cutoff) {
  return(sum(x > cutoff) / length(x))
}
```

(a)

```
exceeded_proportion(seq(1, 100), 40)
```

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

(b)

```
if(!require(Devore7)) install.packages("Devore7")
```

```
## 载入需要的程序包：Devore7
```

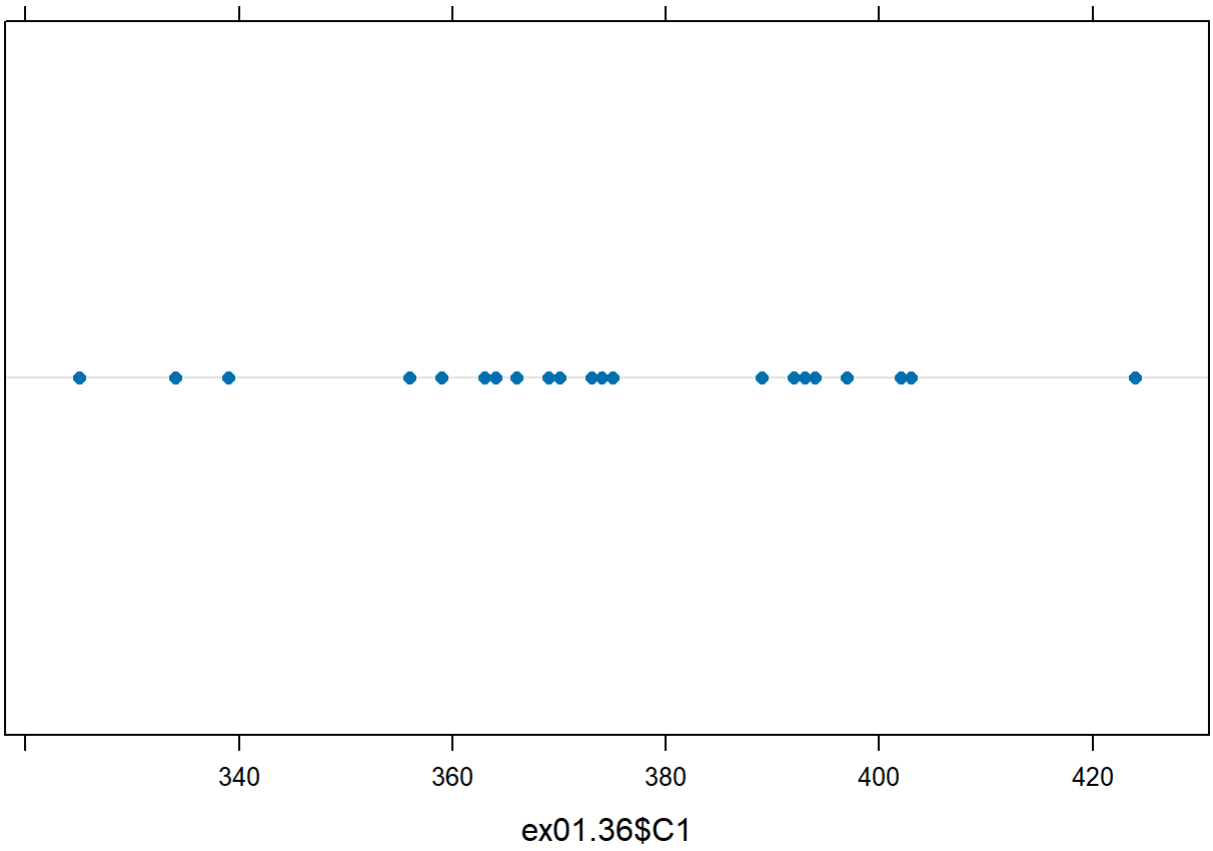
```
## 载入需要的程序包：MASS
```

```
## 载入需要的程序包：lattice
```

```
library(Devore7)

if (!require("lattice")) install.packages("lattice")
library(lattice)
```

```
dotplot(ex01.36$C1)
```




```
exceeded_proportion(ex01.36$C1, 7 * 60)
```

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

MB.Ch1.18

```
if(!require(MASS)) install.packages("MASS")
library(MASS)
```

```
df1 <- unstack(Rabbit, BPchange ~ Animal)
df2 <- unstack(Rabbit, Dose ~ Animal)
df3 <- unstack(Rabbit, Treatment ~ Animal)
new_df <- data.frame(row.names(df1), df3[1], df2[1], df1)
colnames(new_df) <- c("", "Treatment", "Dose", "R1", "R2", "R3", "R4", "R5")
new_df
```

```
##       Treatment   Dose    R1    R2    R3    R4   R5
## 1   1   Control   6.25  0.50  1.00  0.75  1.25  1.5
## 2   2   Control  12.50  4.50  1.25  3.00  1.50  1.5
## 3   3   Control  25.00 10.00  4.00  3.00  6.00  5.0
## 4   4   Control  50.00 26.00 12.00 14.00 19.00 16.0
## 5   5   Control 100.00 37.00 27.00 22.00 33.00 20.0
## 6   6   Control 200.00 32.00 29.00 24.00 33.00 18.0
## 7   7       MDL   6.25  1.25  1.40  0.75  2.60  2.4
## 8   8       MDL  12.50  0.75  1.70  2.30  1.20  2.5
## 9   9       MDL  25.00  4.00  1.00  3.00  2.00  1.5
## 10 10       MDL  50.00  9.00  2.00  5.00  3.00  2.0
## 11 11       MDL 100.00 25.00 15.00 26.00 11.00  9.0
## 12 12       MDL 200.00 37.00 28.00 25.00 22.00 19.0
```