homework2

zza

2025-06-29

1 Lodinng and cleaning

a

```r
ca_pa <- read.csv("calif_penn_2011.csv", stringsAsFactors = FALSE)
```

b

```
## 行数: 11275
```

```
## 列数:  34
```

c

```r
temp<-colSums(apply(ca_pa,c(1,2),is.na))
```

对 ca_pa 里的所有元素进行 is.na 判断，并对 apply 返回的逻辑值矩阵进行按列求和，得出每一列缺失值的数目

d

```r
  ca_pa_clean<-na.omit(ca_pa) ## 清除后的 dataframe
```

e

```
## 清除的行数:  670
```

f

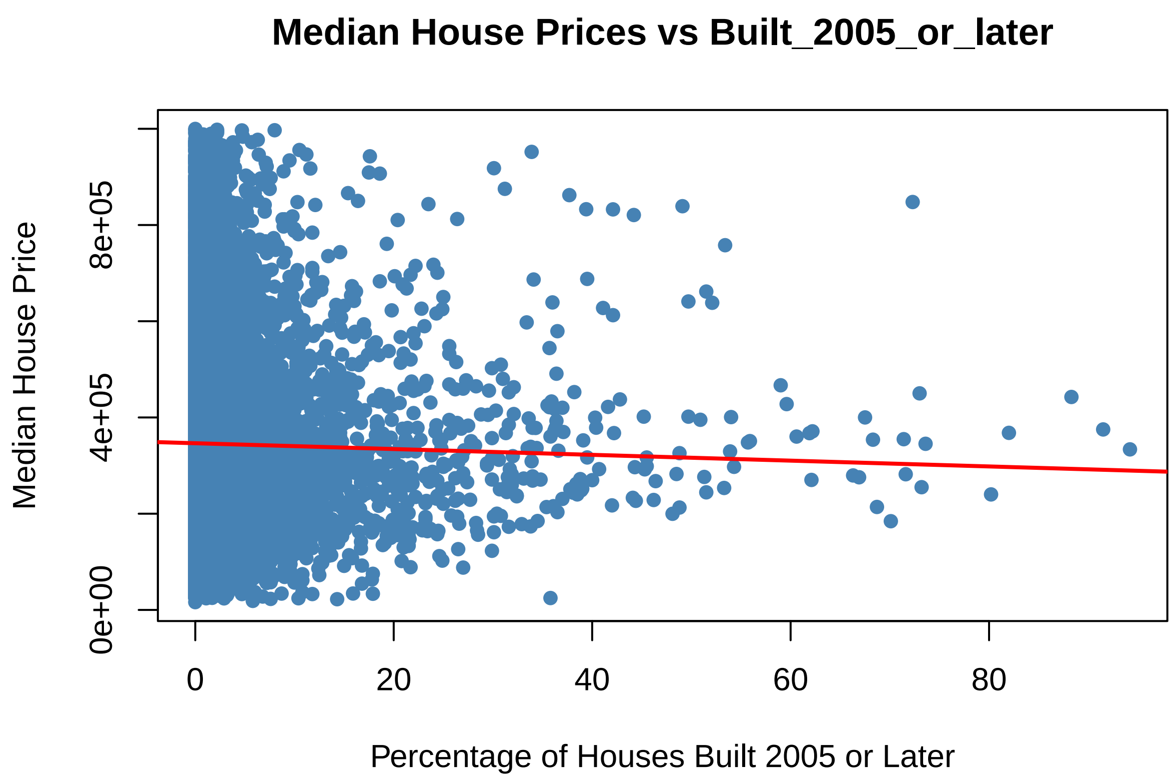
```
## c中各列na数量的总和： 3034
```

其中 c 中的结果和大于清除的总行数，这正说明了它们的一致性，因为每行可能含有多个 na 值

2 This very new House

a. 绘制房价中位数与 Built_2005_or_later 的关系图

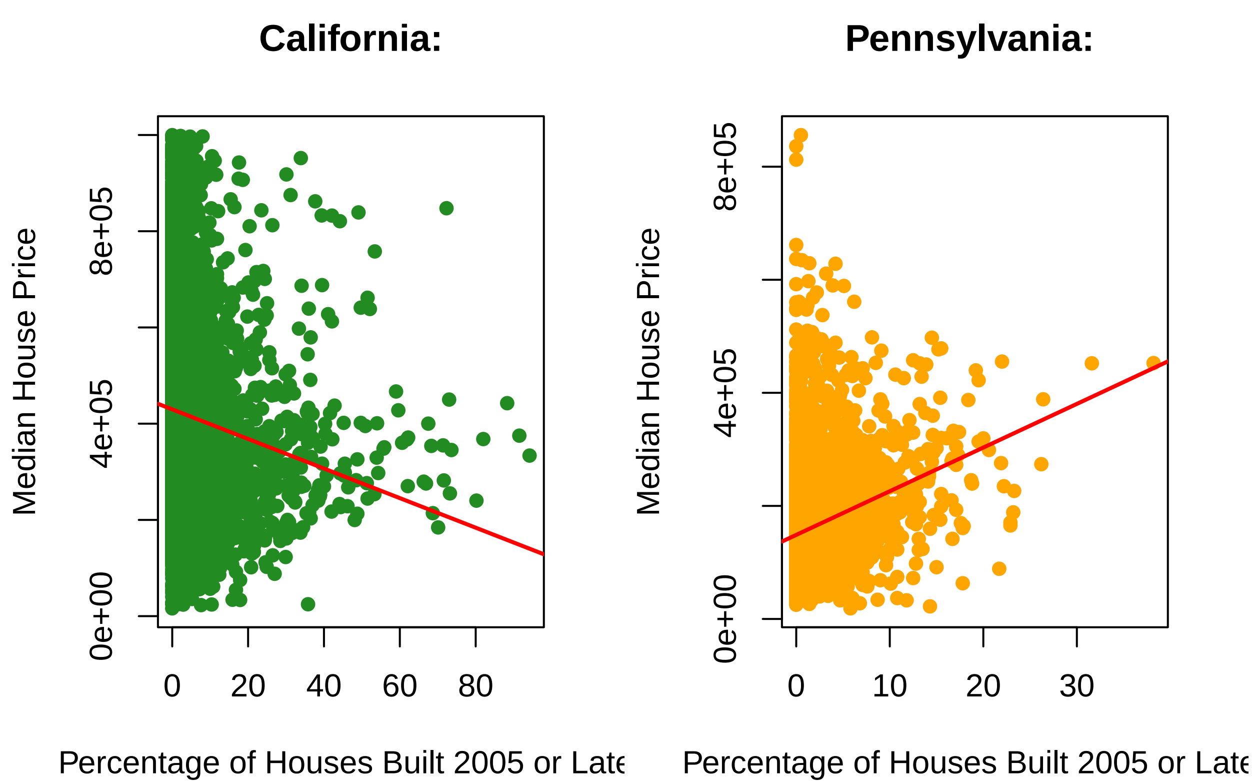
```r
# 绘制散点图，房价中位数列是 Median_house_value，Built_2005_or_later 是对应列
plot(
  x = ca_pa_clean$Built_2005_or_later,
  y = ca_pa_clean$Median_house_value,
  main = "Median House Prices vs Built_2005_or_later",
  xlab = "Percentage of Houses Built 2005 or Later",
  ylab = "Median House Price",
  pch = 16,
  col = "steelblue"
)

abline(lm(Median_house_value ~ Built_2005_or_later, data = ca_pa_clean),
       col = "red",
       lwd = 2)
```



b

```
# 筛选加利福尼亚州（STATEFP == 6）和宾夕法尼亚州（STATEFP == 42）的数据
ca_data <- ca_pa_clean[ca_pa_clean$STATEFP == 6, ]
pa_data <- ca_pa_clean[ca_pa_clean$STATEFP == 42, ]

# 绘制分组散点图，使用 par(mfrow = c(1, 2)) 将图形排列在一行两列
par(mfrow = c(1, 2))

# 加利福尼亚州的散点图
plot(
  x = ca_data$Built_2005_or_later,
  y = ca_data$Median_house_value,
  main = "California:",
  xlab = "Percentage of Houses Built 2005 or Later",
  ylab = "Median House Price",
  pch = 16,
  col = "forestgreen"
```

```
)
abline(lm(Median_house_value ~ Built_2005_or_later, data = ca_data),
       col = "red",
       lwd = 2)

# 宾夕法尼亚州的散点图
plot(
  x = pa_data$Built_2005_or_later,
  y = pa_data$Median_house_value,
  main = "Pennsylvania: ",
  xlab = "Percentage of Houses Built 2005 or Later",
  ylab = "Median House Price",
  pch = 16,
  col = "orange"
)
abline(lm(Median_house_value ~ Built_2005_or_later, data = pa_data),
       col = "red",
       lwd = 2)
```



```r
par(mfrow = c(1, 1))
```

3 Nobody Home

a

```
## 空置率最小值： 0
```

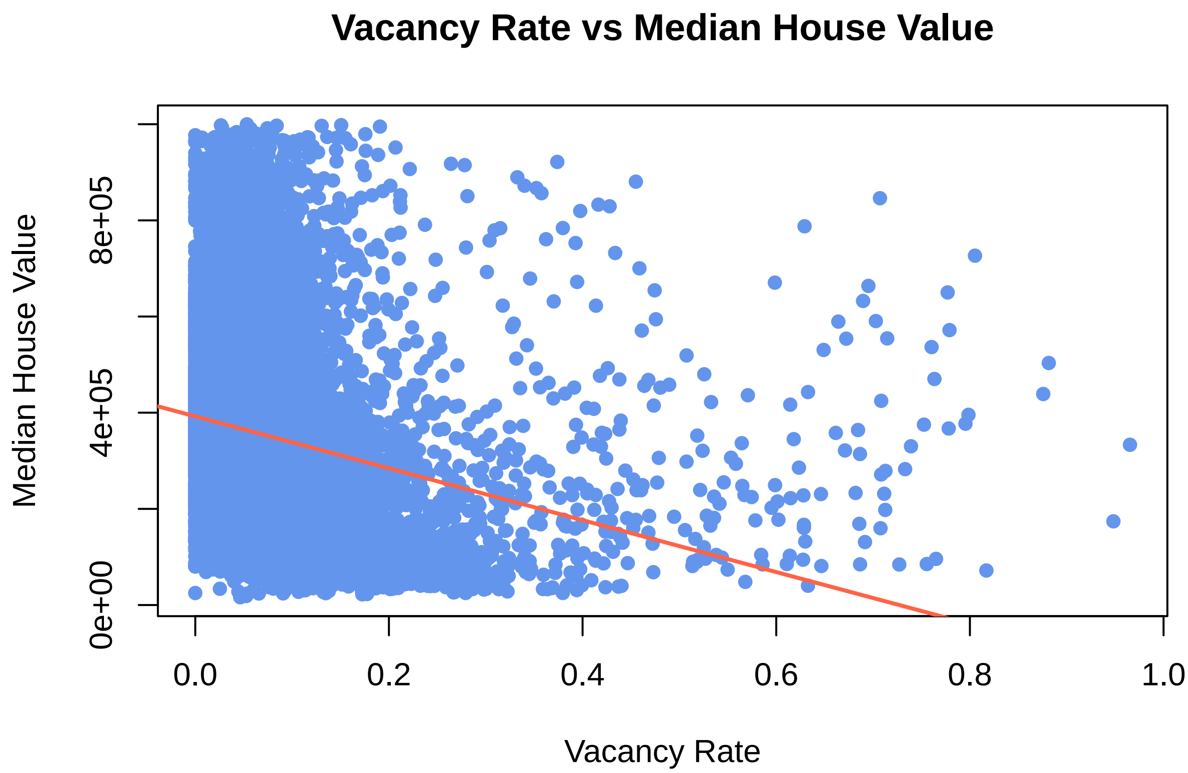
```
## 空置率最大值： 0.965311
```

```
## 空置率均值： 0.08888789
```

```
## 空置率中位数： 0.06767283
```

b

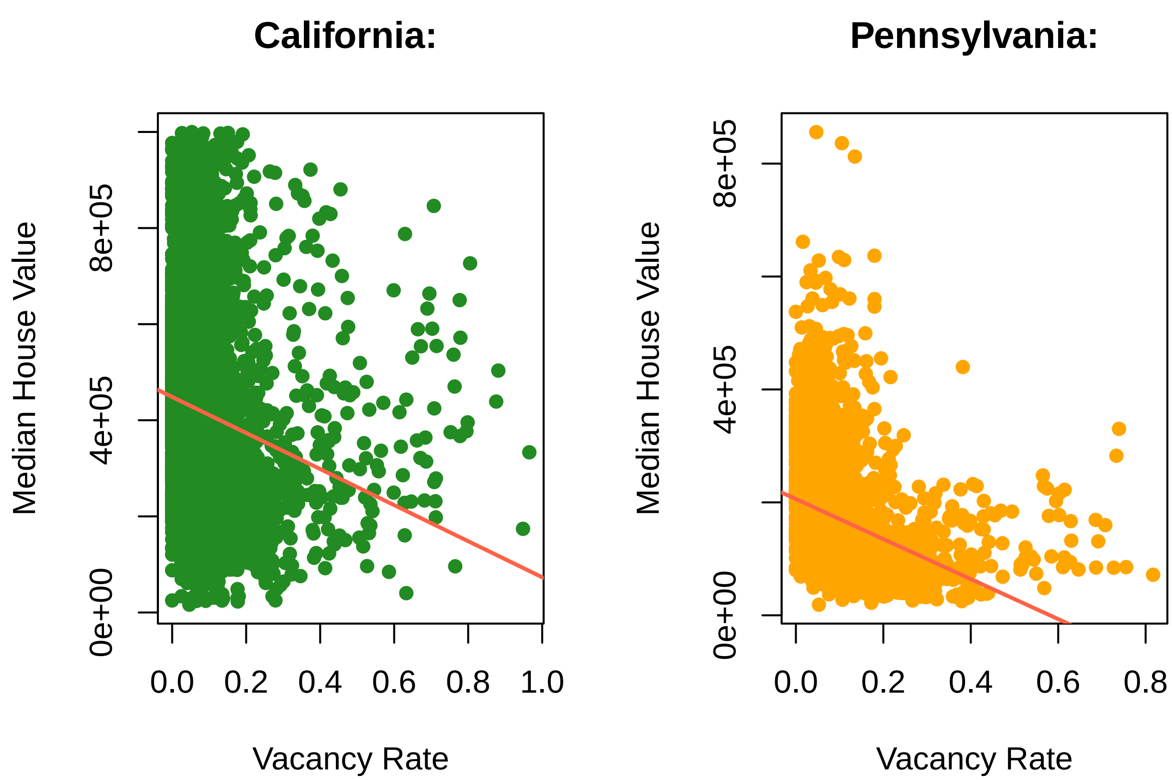
```r
plot(
  x = ca_pa_clean$vacancy_rate,
  y = ca_pa_clean$Median_house_value,
  main = "Vacancy Rate vs Median House Value",
  xlab = "Vacancy Rate",
  ylab = "Median House Value",
  pch = 16,
  col = "cornflowerblue"
)

abline(lm(Median_house_value ~ vacancy_rate, data = ca_pa_clean),
       col = "tomato",
       lwd = 2)
```



c

```
ca_data <- ca_pa_clean[ca_pa_clean$STATEFP == 6, ]
pa_data <- ca_pa_clean[ca_pa_clean$STATEFP == 42, ]

par(mfrow = c(1, 2))

# 加利福尼亚州散点图
plot(
  x = ca_data$vacancy_rate,
  y = ca_data$Median_house_value,
  main = "California: ",
  xlab = "Vacancy Rate",
  ylab = "Median House Value",
  pch = 16,
  col = "forestgreen"
)
abline(lm(Median_house_value ~ vacancy_rate, data = ca_data),
```

```r
       col = "tomato",
       lwd = 2)

# 宾夕法尼亚州散点图
plot(
  x = pa_data$vacancy_rate,
  y = pa_data$Median_house_value,
  main = "Pennsylvania:",
  xlab = "Vacancy Rate",
  ylab = "Median House Value",
  pch = 16,
  col = "orange"
)
abline(lm(Median_house_value ~ vacancy_rate, data = pa_data),
       col = "tomato",
       lwd = 2)
```



```r
par(mfrow = c(1, 1))
```

两个州图像存在较大差异

4

a

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

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

遍历 ca_pa 中每一行，将为加利福尼亚州 Alameda 县的行索引加入到 acca 中（我这里 ca_pa_clean 是清理过的 ca_pa）

遍历 acca 中的行索引，提取这些行的第 10 列数据房价中位数，存储到 accanhv。然后取其中位数

b

```r
median(ca_pa_clean[ca_pa_clean$STATEFP == 6 & ca_pa_clean$COUNTYFP == 1, 10])
```

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

c

```
## Alameda 县平均新建住房比例： 2.820468
```

```
## Santa Clara 县平均新建住房比例： 3.200319
```

```
## Allegheny 县平均新建住房比例： 1.474219
```

d

```
## (i) 整个数据的相关性： -0.01893186
```

```
## (ii) 加利福尼亚州的相关性： -0.1153604
```

```
## (iii) 宾夕法尼亚州的相关性： 0.2681654
```

```
## (iv) Alameda 县的相关性： 0.01303543
```

```
## (v) Santa Clara 县的相关性： -0.1726203
```

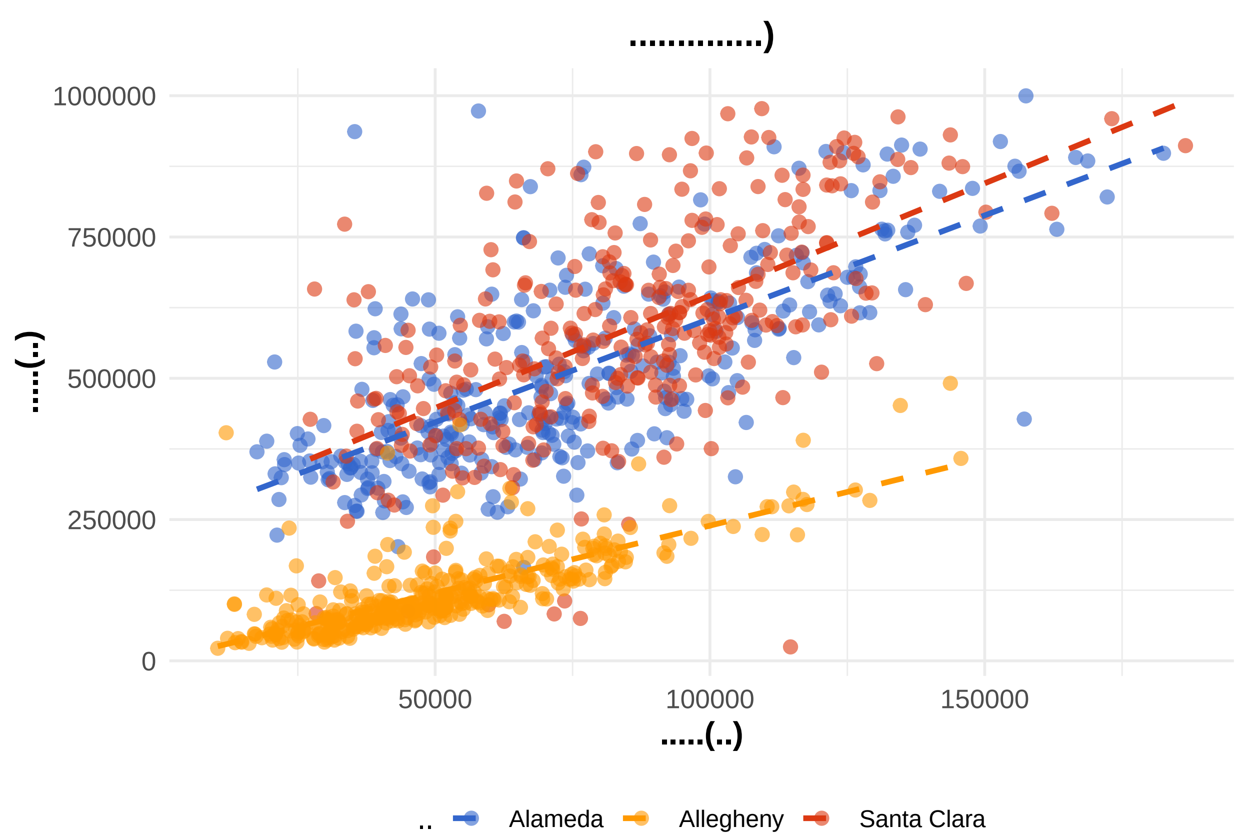
```
## (vi) Allegheny 县的相关性： 0.1939652
```

e. 绘制房价中位数与收入中位数的关系图（按县分组）

```r
# 加载 ggplot2 包（如果未加载）
library(ggplot2)

alameda <- ca_pa_clean[ca_pa_clean$STATEFP == 6 & ca_pa_clean$COUNTYFP == 1, ]
santa_clara <- ca_pa_clean[ca_pa_clean$STATEFP == 6 & ca_pa_clean$COUNTYFP == 85, ]
allegheny <- ca_pa_clean[ca_pa_clean$STATEFP == 42 & ca_pa_clean$COUNTYFP == 3, ]

# 为每个县数据添加标识列，再合并，让数据框直接包含要引用的列
alameda$County <- "Alameda"
santa_clara$County <- "Santa Clara"
allegheny$County <- "Allegheny"

county_data <- rbind(alameda, santa_clara, allegheny)

# 使用 ggplot2 绘制分组散点图，直接引用列名
ggplot(county_data, aes(x = Median_household_income, y = Median_house_value, color = County)) +
  geom_point(alpha = 0.6, size = 2) +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed") +
  labs(
```

```r
    title = " 房价中位数与收入中位数的关系）",
    x = " 收入中位数（美元）",
    y = " 房价中位数（美元）",
    color = " 县名"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "bottom",
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12, face = "bold")
  ) +
  scale_color_manual(values = c(
    "Alameda" = "#3366CC",       # 蓝色，区分 Alameda 县
    "Santa Clara" = "#DC3912",  # 红色，区分 Santa Clara 县
    "Allegheny" = "#FF9900"     # 橙色，区分 Allegheny 县
  ))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



MB.Ch1

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

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

```r
# 初始 gender 因子有 female（91 个）和 male（92 个），table 按默认水平统计，输出对应数量
```

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

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

```r
 # 重新指定因子水平，table 按新水平顺序（male 在前、female 在后）统计
```

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

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

```r
 # 指定水平为 c("Male", "female")，原数据无 Male 水平，仅 female 匹配，故 Male 计数为 0，female 为 91
```

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

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

```r
# 加入 exclude=NULL 后，未匹配的 male 被归为 NA 显示，输出 Male（0）、female（91）、<NA>
```

MB.Ch1.2

a

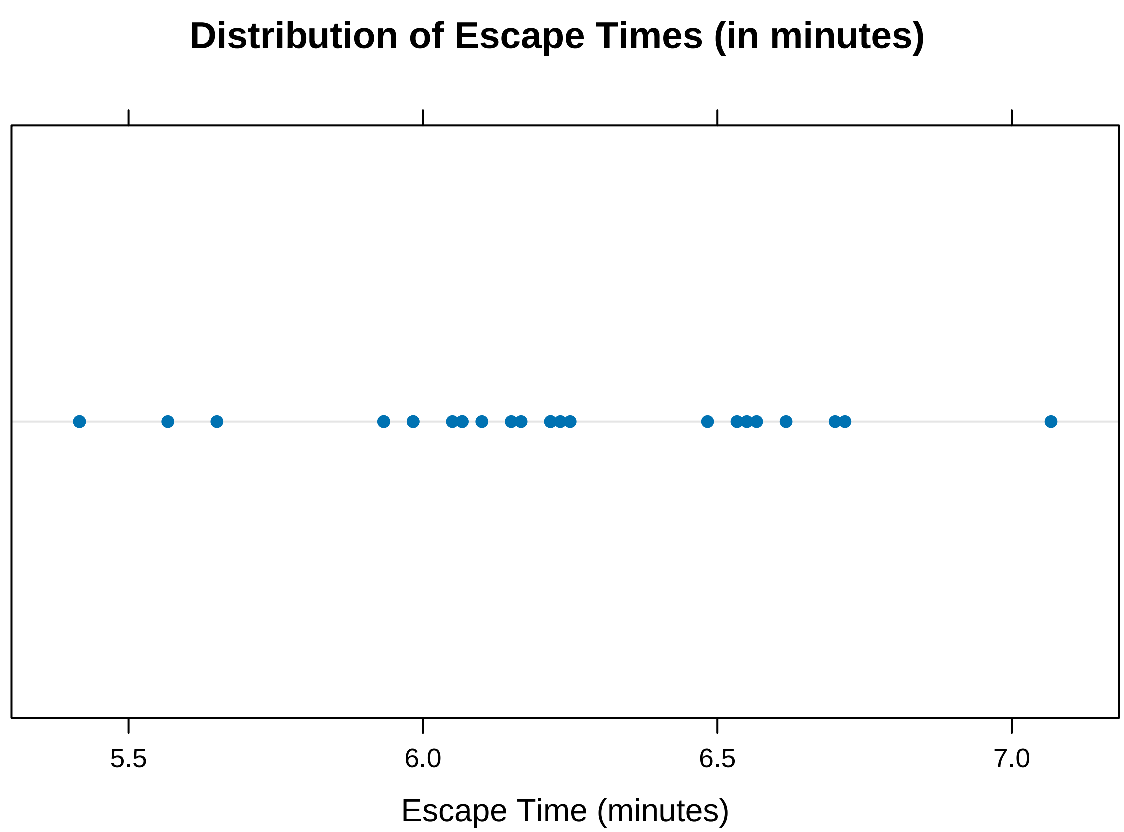
```r
prop_exceed <- function(x, cutoff) {
  mean(x > cutoff)
}
x <- 1:100
# 用 1 到 100 计算超过 50 的比例
result_a <- prop_exceed(x, 50)
```

b

```r
prop_exceed <- function(x, cutoff) {
  mean(x > cutoff)
}

library(Devore7)
```

```
## Loading required package: MASS

## Loading required package: lattice
```

```r
data("ex01.36")

# 将秒换算为分钟
escape_times <- ex01.36$C1 / 60

dotplot(escape_times,
        main = "Distribution of Escape Times (in minutes)",
        xlab = "Escape Time (minutes)")
```




```
# 计算超过 7 分钟的比例
result_b <- prop_exceed(escape_times, 7)
result_b
```

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

MB.Ch 1.18

```
library(MASS)
data(Rabbit)



rabbit_unstacked <- unstack(Rabbit, BPchange ~ Animal)



Rabbit$id <- with(Rabbit, paste(Dose, Treatment, sep = "_"))


dose_treatment <- unique(Rabbit[c("Dose", "Treatment", "id")])
```

```r
final_result <- cbind(dose_treatment, rabbit_unstacked)

final_result$id <- NULL

final_result <- final_result[, c("Treatment", "Dose", "R1", "R2", "R3", "R4", "R5")]

final_result <- final_result[order(final_result$Dose), ]

rownames(final_result) <- NULL

final_result
```

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