Homework 2

ywx 3220101739

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`.
   b. How many rows and columns does the dataframe have?
   c. Run this command, and explain, in words, what this does:

   ```
   colSums(apply(ca_pa,c(1,2),is.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.
   e. How many rows did this eliminate?
   f. Are your answers in (c) and (e) compatible? Explain.

```r
# 1a. 加载数据
ca_pa <- read.csv("data/calif_penn_2011.csv")

# 1b. 查看行列数
dim(ca_pa)
```

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

```r
# 1c. apply(ca_pa, c(1,2), is.na) 对数据框的每个元素检查是否为 NA，返回一个逻辑值矩阵；
# colSums() - 对每列的 TRUE 值 (即 NA 值) 进行求和。
# 所以会计算数据框中每一列中缺失值 (NA) 的数量
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
```

```
# 1d. 使用 na.omit() 删除含 NA 的行
ca_pa_clean <- na.omit(ca_pa)

# 1e. 计算被删除的行数
rows_eliminated <- nrow(ca_pa) - nrow(ca_pa_clean)
rows_eliminated
```

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

```r
# 1f. 检查 c,e 是否兼容
na_per_col <- colSums(is.na(ca_pa))
max_na <- max(na_per_col)
total_na <- sum(na_per_col)
rows_eliminated >= max_na && rows_eliminated <= total_na
```

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

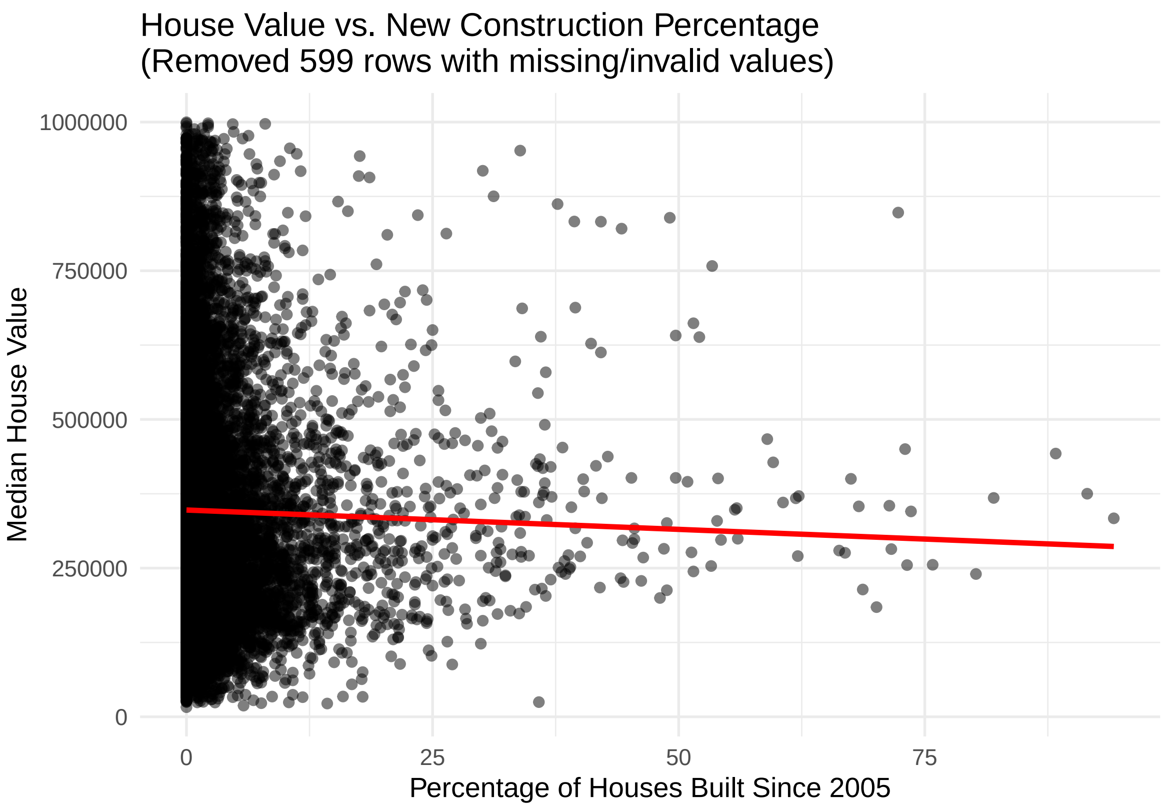
解释：c 中计算的是各列中 NA 的数量，e 的结果为删去含有 NA 的行数。含有 NA 最多的列的所有 NA 所在的行必须被删除，且可能有多列共享同一行的 NA，因此有 max_na <= rows_eliminated <= total_na 成立。

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.
   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.

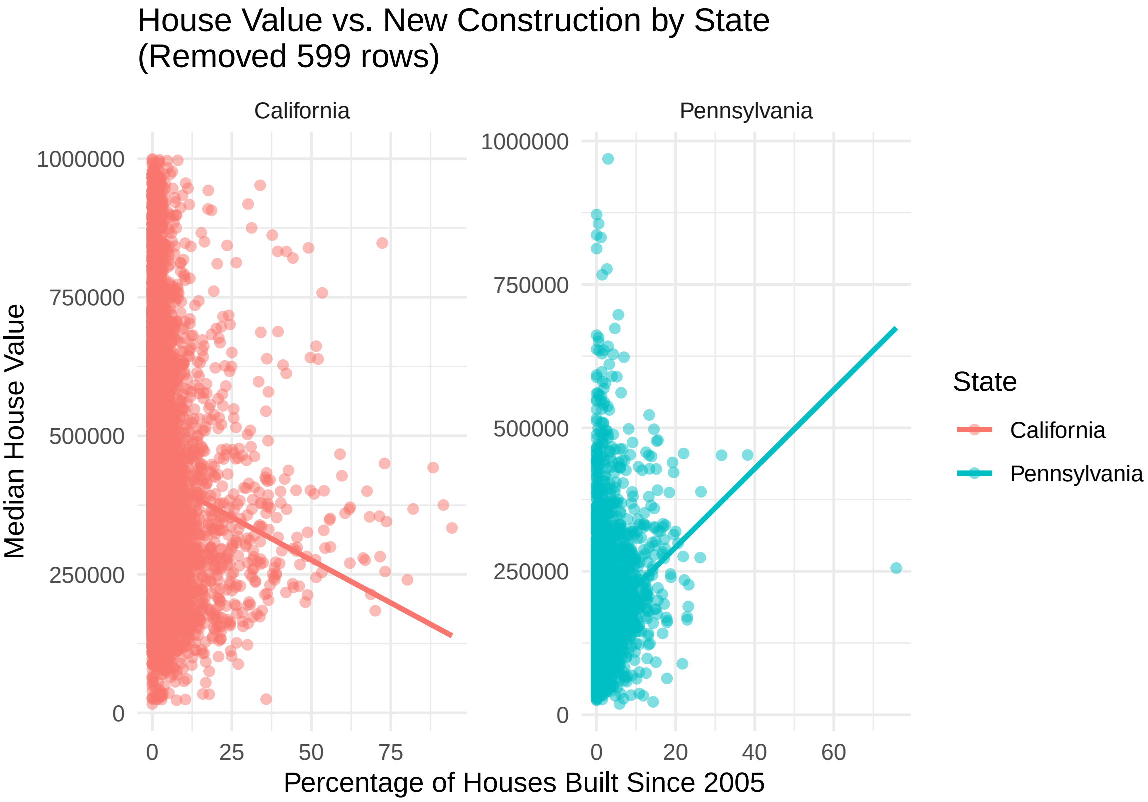
```r
# 移除有缺失值的行
ca_pa_clean <- ca_pa %>%
  filter(!is.na(Median_house_value),
         !is.na(Built_2005_or_later),
         is.finite(Median_house_value),
         is.finite(Built_2005_or_later))
# 使用清理后的数据
ggplot(ca_pa_clean, aes(x = Built_2005_or_later, y = Median_house_value)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Percentage of Houses Built Since 2005",
       y = "Median House Value",
       title = paste("House Value vs. New Construction Percentage",
                     "\n(Removed", nrow(ca_pa) - nrow(ca_pa_clean), "rows with missing/invalid
  theme_minimal()
```

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



```
# 创建州名变量并清理数据
ca_pa_clean <- ca_pa %>%
  mutate(State = case_when(
    STATEFP == "6" ~ "California",
    STATEFP == "42" ~ "Pennsylvania",
    TRUE ~ "Other"
  )) %>%
  filter(!is.na(Median_house_value),
         !is.na(Built_2005_or_later),
         is.finite(Median_house_value),
         is.finite(Built_2005_or_later),
         Built_2005_or_later >= 0 & Built_2005_or_later <= 100,
         Median_house_value > 0)

# 绘制分面图
ggplot(ca_pa_clean, aes(x = Built_2005_or_later, y = Median_house_value, color = State)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  facet_wrap(~ State, scales = "free") +
  labs(x = "Percentage of Houses Built Since 2005",
       y = "Median House Value",
       title = paste("House Value vs. New Construction by State",
                     "\n(Removed", nrow(ca_pa) - nrow(ca_pa_clean), "rows)")) +
  theme_minimal()
```

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



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

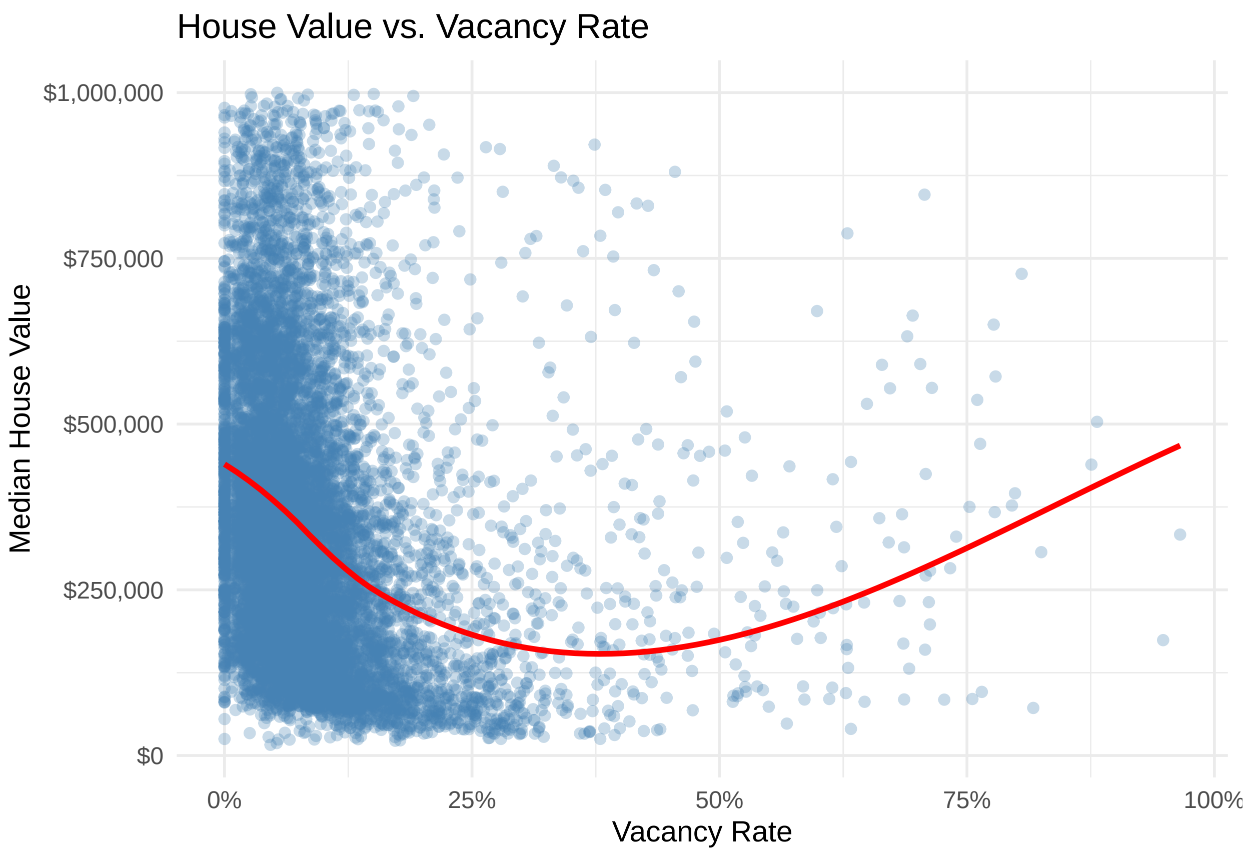
```
# 添加空置率列
ca_pa <- ca_pa %>%
  mutate(Vacancy_rate = Vacant_units / Total_units)
```

```r
# 计算统计量
vacancy_stats <- ca_pa %>%
  summarise(
    Min = min(Vacancy_rate, na.rm = TRUE),
    Max = max(Vacancy_rate, na.rm = TRUE),
    Mean = mean(Vacancy_rate, na.rm = TRUE),
    Median = median(Vacancy_rate, na.rm = TRUE)
  )
print(vacancy_stats)
```

```
##   Min Max       Mean     Median
## 1   0   1 0.08917878 0.06766326
```

```r
# 3b. 绘制空置率与房价中位数的关系图 (使用清理后的数据)
ca_pa_clean <- ca_pa_clean %>%
  mutate(Vacancy_rate = Vacant_units / Total_units)
ggplot(ca_pa_clean, aes(x = Vacancy_rate, y = Median_house_value)) +
  geom_point(alpha = 0.3, color = "steelblue") +
  geom_smooth(method = "loess", color = "red", se = FALSE) +
  labs(x = "Vacancy Rate",
       y = "Median House Value",
       title = "House Value vs. Vacancy Rate") +
  theme_minimal() +
  scale_y_continuous(labels = scales::dollar) +  # 房价用美元格式
  scale_x_continuous(labels = scales::percent)   # 空置率用百分比格式
```

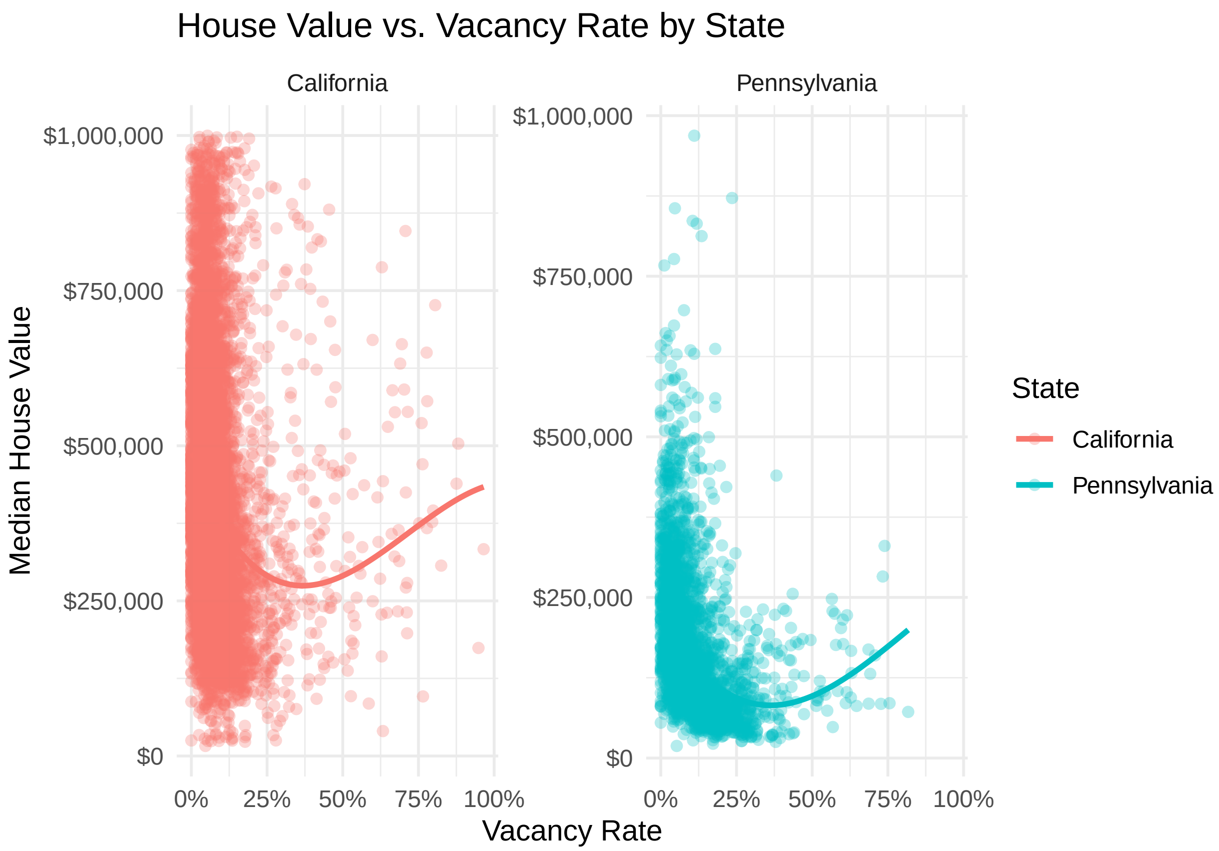
```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
# 3c
ca_pa_clean <- ca_pa %>%
  mutate(State = case_when(
    STATEFP == "6" ~ "California",
    STATEFP == "42" ~ "Pennsylvania",
    TRUE ~ "Other"
  )) %>%
  # 计算空置率
  mutate(Vacancy_rate = Vacant_units / Total_units) %>%  # 替换为实际列名
  # 数据清理
  filter(
    State != "Other",
    !is.na(Median_house_value),
    !is.na(Vacancy_rate),
    is.finite(Median_house_value),
    is.finite(Vacancy_rate),
    Vacancy_rate >= 0 & Vacancy_rate <= 1,  # 空置率应在 0-1 之间
    Median_house_value > 0
  )
```

```r
# 2. 绘制分面图（使用清理后的数据 ca_pa_clean）
ggplot(ca_pa_clean, aes(x = Vacancy_rate, y = Median_house_value)) +
  geom_point(alpha = 0.3, aes(color = State)) +  # 颜色映射到 State
  geom_smooth(method = "loess", se = FALSE, aes(color = State)) +  # 颜色映射
  facet_wrap(~ State, scales = "free_y") +  # 按 State 分面
  labs(x = "Vacancy Rate",
       y = "Median House Value",
       title = "House Value vs. Vacancy Rate by State") +
  theme_minimal() +
  scale_color_manual(values = c("California" = "#F8766D",
                                "Pennsylvania" = "#00BFC4")) +
  scale_x_continuous(labels = scales::percent) +  # x 轴显示为百分比
  scale_y_continuous(labels = scales::dollar)    # y 轴显示为美元
```

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



```
# 3. 差异检验（同样使用清理后的数据）
ca_vacancy <- ca_pa_clean %>%
  filter(State == "California") %>%
  pull(Vacancy_rate)
pa_vacancy <- ca_pa_clean %>%
  filter(State == "Pennsylvania") %>%
  pull(Vacancy_rate)
t_test_result <- t.test(ca_vacancy, pa_vacancy)
print(paste("加州和宾州空置率差异p值：", round(t_test_result$p.value, 4)))
```

```
## [1] "加州和宾州空置率差异p值： 0"
```

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

```
# 4a. 代码块的功能及实现原理
  # 遍历数据框的每一行，筛选出同时满足以下条件的行索引：州代码为6（加州）县代码为1（Alameda县）
  # 结果：acca是一个向量，存储所有Alameda县普查区的行号（索引）
    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)
        }
      }
    }
  # 根据第一部分的行索引，从数据框的第 10 列（假设是房价列）提取 Alameda 县所有普查区的房价值。
  # 结果：accamhv 是一个向量，存储 Alameda 县的房价数据。
    accamhv <- c()
    for (tract in acca) {
      accamhv <- c(accamhv, ca_pa[tract,10])
    }
  # 计算 accamhv 的中位数，即 Alameda 县的房价中位数。
    median(accamhv)
```

```
#4b.
median(ca_pa$Median_house_value[ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1], na.rm = TRUE)
```

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

```
#4c. 计算三县 2005 年后新建房屋的平均百分比
ca_pa %>%
  filter(
    (STATEFP == 6 & COUNTYFP == 1) |    # Alameda, CA
    (STATEFP == 6 & COUNTYFP == 85) |   # Santa Clara, CA
    (STATEFP == 42 & COUNTYFP == 3)     # Allegheny, PA
  ) %>%
  group_by(STATEFP, COUNTYFP) %>%
  summarise(
    Avg_Percent_New_Housing = mean(Built_2005_or_later, na.rm = TRUE),
    .groups = "drop"
  )
```

```
## # A tibble: 3 x 3
##   STATEFP COUNTYFP Avg_Percent_New_Housing
##     <int>    <int>                   <dbl>
## 1       6        1                    2.93
## 2       6       85                    3.16
## 3      42        3                    1.88
```

```
#4d
library(dplyr)
# (i) 全部数据
cor_all <- cor(ca_pa$Median_house_value, ca_pa$Built_2005_or_later, use = "complete.obs")
```

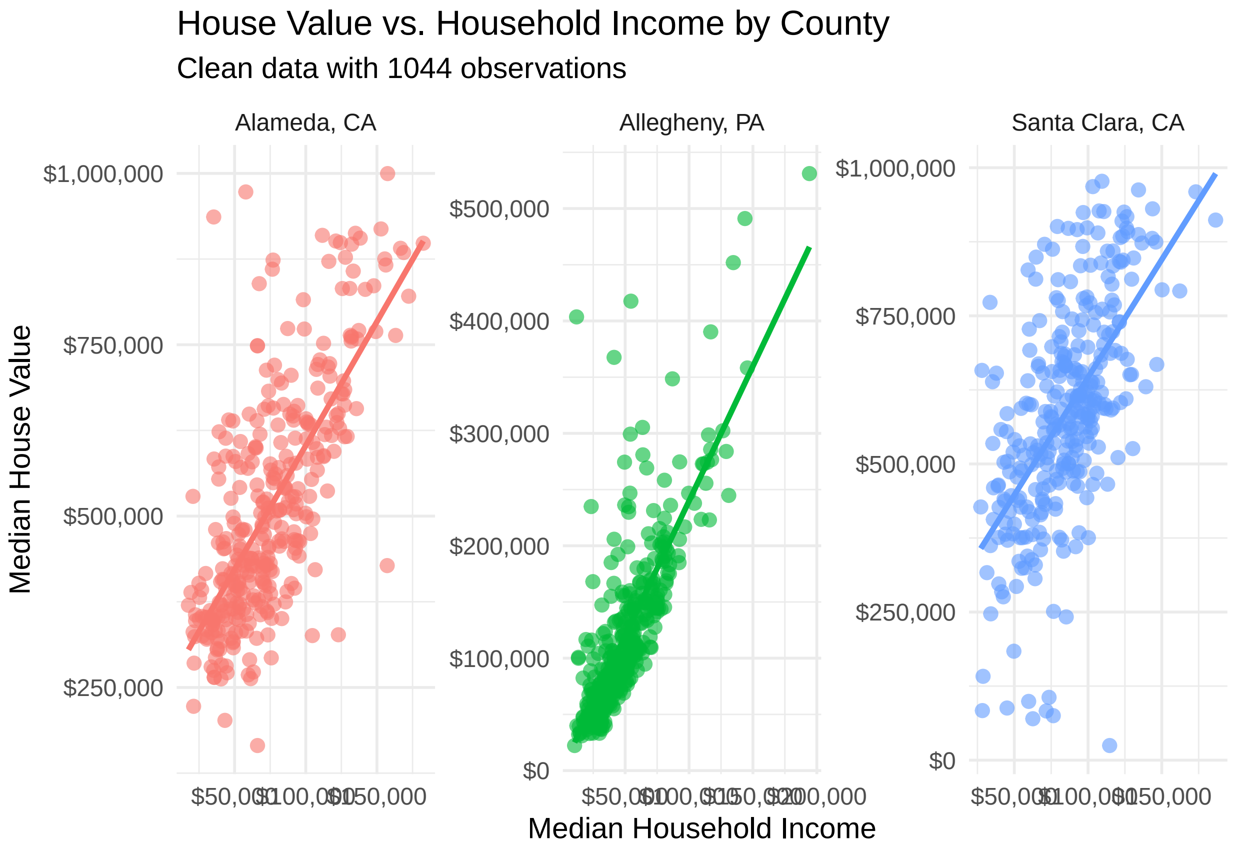
```r
# (ii) 加州全部
cor_ca <- ca_pa %>%
  filter(STATEFP == 6) %>%
  summarise(cor = cor(Median_house_value, Built_2005_or_later, use = "complete.obs")) %>%
  pull(cor)

# (iii) 宾州全部
cor_pa <- ca_pa %>%
  filter(STATEFP == 42) %>%
  summarise(cor = cor(Median_house_value, Built_2005_or_later, use = "complete.obs")) %>%
  pull(cor)

# (iv) Alameda 县 (CA-1)
cor_alameda <- ca_pa %>%
  filter(STATEFP == 6, COUNTYFP == 1) %>%
  summarise(cor = cor(Median_house_value, Built_2005_or_later, use = "complete.obs")) %>%
  pull(cor)

# (v) Santa Clara 县 (CA-85)
cor_santa_clara <- ca_pa %>%
  filter(STATEFP == 6, COUNTYFP == 85) %>%
  summarise(cor = cor(Median_house_value, Built_2005_or_later, use = "complete.obs")) %>%
  pull(cor)

# (vi) Allegheny 县 (PA-3)
cor_allegheny <- ca_pa %>%
  filter(STATEFP == 42, COUNTYFP == 3) %>%
  summarise(cor = cor(Median_house_value, Built_2005_or_later, use = "complete.obs")) %>%
  pull(cor)

# 汇总结果
results <- data.frame(
  Region = c("All Data", "California", "Pennsylvania",
             "Alameda County", "Santa Clara County", "Allegheny County"),
  Correlation = round(c(cor_all, cor_ca, cor_pa,
                        cor_alameda, cor_santa_clara, cor_allegheny), 3)
)
```

```r
print(results)
```

```
##                Region Correlation
## 1            All Data      -0.021
## 2          California      -0.116
## 3        Pennsylvania       0.234
## 4      Alameda County       0.014
## 5 Santa Clara County      -0.173
## 6    Allegheny County       0.187
```

```r
#4e
  # 创建目标县数据集
target_counties <- ca_pa %>%
  filter(
    (STATEFP == 6 & COUNTYFP == 1) |
    (STATEFP == 6 & COUNTYFP == 85) |
    (STATEFP == 42 & COUNTYFP == 3)
  ) %>%
  mutate(
    County = case_when(
      COUNTYFP == 1 ~ "Alameda, CA",
      COUNTYFP == 85 ~ "Santa Clara, CA",
      COUNTYFP == 3 ~ "Allegheny, PA"
    )
  )
  # 清理数据
target_counties_clean <- target_counties %>%
  filter(
    is.finite(Median_household_income),
    is.finite(Median_house_value),
    Median_household_income > 0,
    Median_house_value > 0
  )
  # 绘图
ggplot(target_counties_clean, aes(x = Median_household_income, y = Median_house_value)) +
  geom_point(alpha = 0.6, size = 2, aes(color = County)) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1, aes(color = County)) +
  facet_wrap(~ County, scales = "free") +
  scale_y_continuous(labels = scales::dollar) +
  scale_x_continuous(labels = scales::dollar) +
  labs(
    x = "Median Household Income",
    y = "Median House Value",
    title = "House Value vs. Household Income by County",
    subtitle = paste("Clean data with", nrow(target_counties_clean), "observations")
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

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



MB.Ch1.11. Run the following code: Explain the output from the successive uses of table().

```
#1. 初始创建 gender 因子
gender <- factor(c(rep("female", 91), rep("male", 92)))
table(gender)
```

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

```r
#2. 指定因子水平顺序
gender <- factor(gender, levels=c("male", "female"))
table(gender)
```

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

```r
# 通过 levels 参数将水平顺序强制设为 male 在前、female 在后
#3. 修改水平标签（首字母大小写）
gender <- factor(gender, levels=c("Male", "female"))  # "Male"   "male"
table(gender)
```

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

```r
#4. 显示包含 NA 的统计
table(gender, exclude=NULL)
```

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

```r
#5. 清理环境
rm(gender)  # 删除 gender 变量
```

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

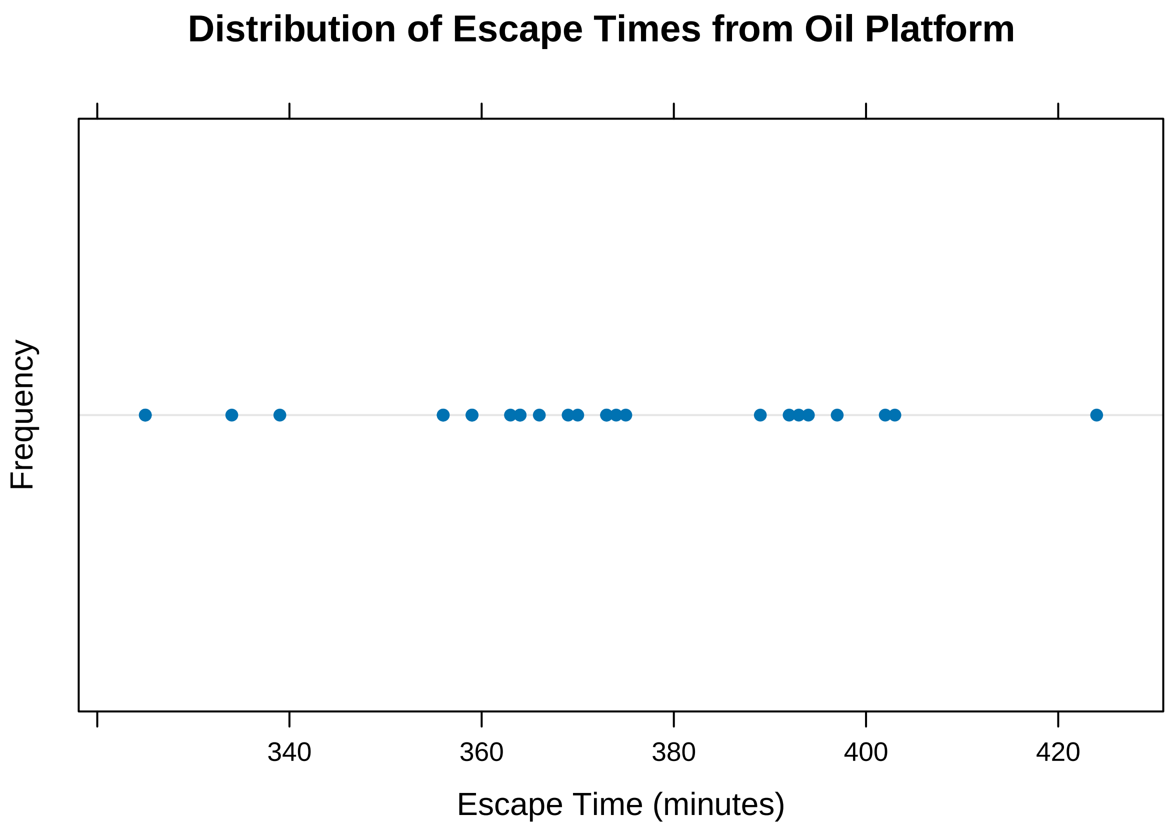
(a) Use the sequence of numbers 1, 2, . . . , 100 to check that this function gives the result that is expected.

```r
# 函数实现
prop_above_cutoff <- function(x, cutoff) {
  mean(x > cutoff, na.rm = TRUE)
}
# 生成测试序列
test_vec <- 1:100
# 验证不同阈值
cutoffs <- c(50, 75, 90, 101)
results <- sapply(cutoffs, function(c) {
  prop_above_cutoff(test_vec, c)
})
# 输出结果
data.frame(
  Cutoff = cutoffs,
  Expected_Prop = (100 - cutoffs) / 100,  # 理论比例
  Actual_Prop = results
)
```

```
##   Cutoff Expected_Prop Actual_Prop
## 1     50          0.50        0.50
## 2     75          0.25        0.25
## 3     90          0.10        0.10
## 4    101         -0.01        0.00
```

(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.

```
my_data <- dplyr::select(iris, Species)  # 确保使用 dplyr 的 select
hills_data <- MASS::hills                # 确保使用 MASS 的 hills
# 加载数据
data(ex01.36)
escape_times <- ex01.36$C1
# 绘制点图
dotplot(escape_times,
        xlab = "Escape Time (minutes)",
        ylab = "Frequency",
        main = "Distribution of Escape Times from Oil Platform")
```



```
# 计算比例
prop_over_7min <- prop_above_cutoff(escape_times, 420)
print(paste("Proportion exceeding 7 minutes:", round(prop_over_7min, 3)))
```

```
## [1] "Proportion exceeding 7 minutes: 0.038"
```

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

….

```
data(Rabbit)
# 第一次 unstack(): 将血压变化按兔子和处理条件分组
```

```r
unstacked1 <- unstack(Rabbit, BPchange ~ Animal)
# 第二次 unstack(): 处理剂量信息
unstacked2 <- unstack(Rabbit, Dose ~ Treatment)
# 第三次 unstack(): 可能需要进一步整理数据格式
# 最终整理
result <- data.frame(
  Treatment = rep(unique(Rabbit$Treatment), each = length(unique(Rabbit$Dose))/length(unique(Rabbi
  Dose = unique(Rabbit$Dose),
  R1 = unstacked1$R1,
  R2 = unstacked1$R2,
  R3 = unstacked1$R3,
  R4 = unstacked1$R4,
  R5 = unstacked1$R5
)
# 查看结果
result
```

```
##    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        MDL  50.00 26.00 12.00 14.00 19.00 16.0
## 5        MDL 100.00 37.00 27.00 22.00 33.00 20.0
## 6        MDL 200.00 32.00 29.00 24.00 33.00 18.0
## 7    Control   6.25  1.25  1.40  0.75  2.60  2.4
## 8    Control  12.50  0.75  1.70  2.30  1.20  2.5
## 9    Control  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
```