Homework 2 Solutions

3190300985 LUIS LUZERN YUVEN

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Exercise 1 *Loading and Cleaning*

(a),(b)

```
ca_pa <- read.csv("data/calif_penn_2011.csv")
dim(ca_pa)
```

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

The dataframe has 11275 rows and 34 columns

(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
```

The command counts the number of NA values in each variable (column)

(d),(e),(f)

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

```
## [1] 10605    34
```

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

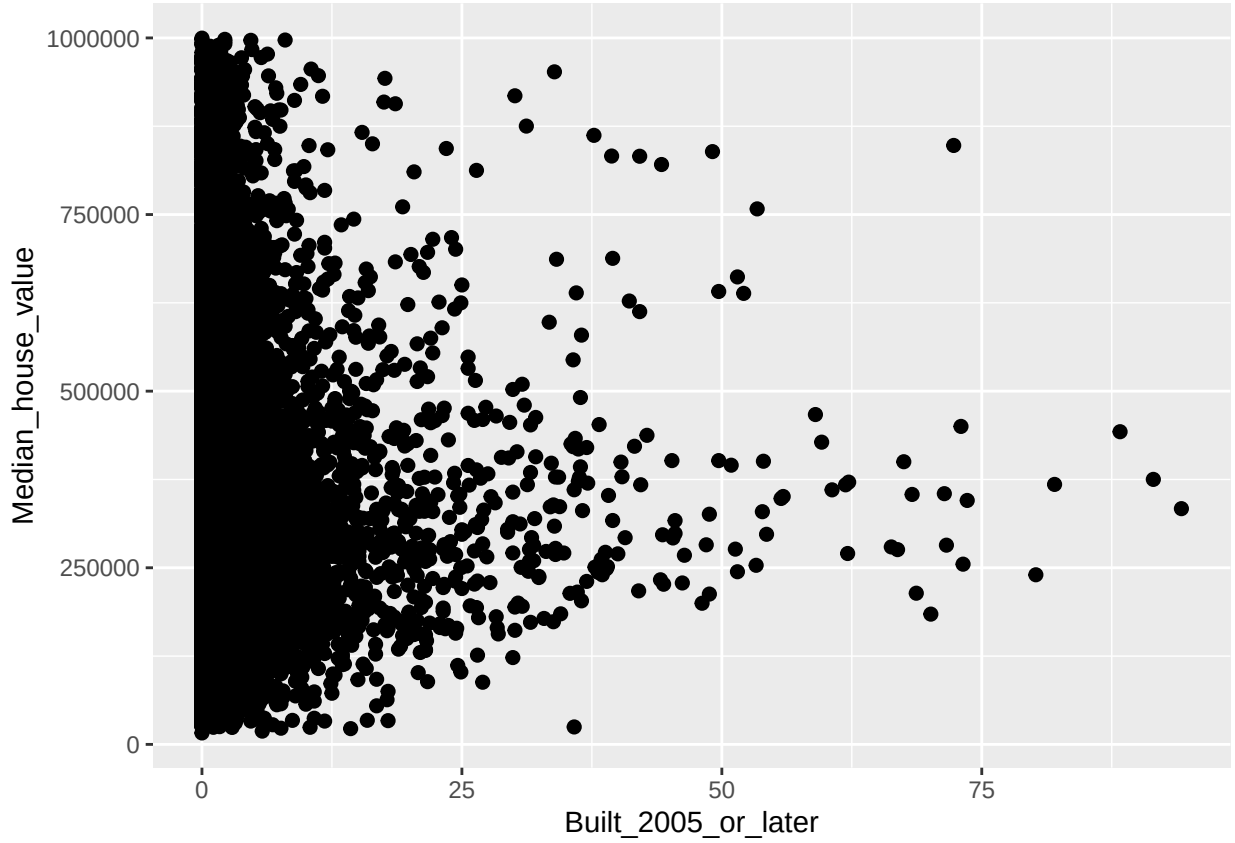
```
##                           X                      GEO.id2
##                           0                            0
##                     STATEFP                     COUNTYFP
##                           0                            0
##                     TRACTCE                   POPULATION
##                           0                            0
##                    LATITUDE                    LONGITUDE
##                           0                            0
##           GEO.display.label          Median_house_value
##                           0                            0
##                 Total_units                Vacant_units
##                           0                            0
##                Median_rooms  Mean_household_size_owners
##                           0                            0
## Mean_household_size_renters         Built_2005_or_later
##                           0                            0
##          Built_2000_to_2004                 Built_1990s
##                           0                            0
##                 Built_1980s                 Built_1970s
##                           0                            0
##                 Built_1960s                 Built_1950s
##                           0                            0
##                 Built_1940s       Built_1939_or_earlier
##                           0                            0
##                  Bedrooms_0                  Bedrooms_1
##                           0                            0
##                  Bedrooms_2                  Bedrooms_3
```

```
##                         0                         0
##                 Bedrooms_4          Bedrooms_5_or_more
##                         0                         0
##                    Owners                    Renters
##                         0                         0
##     Median_household_income       Mean_household_income
##                         0                         0
```

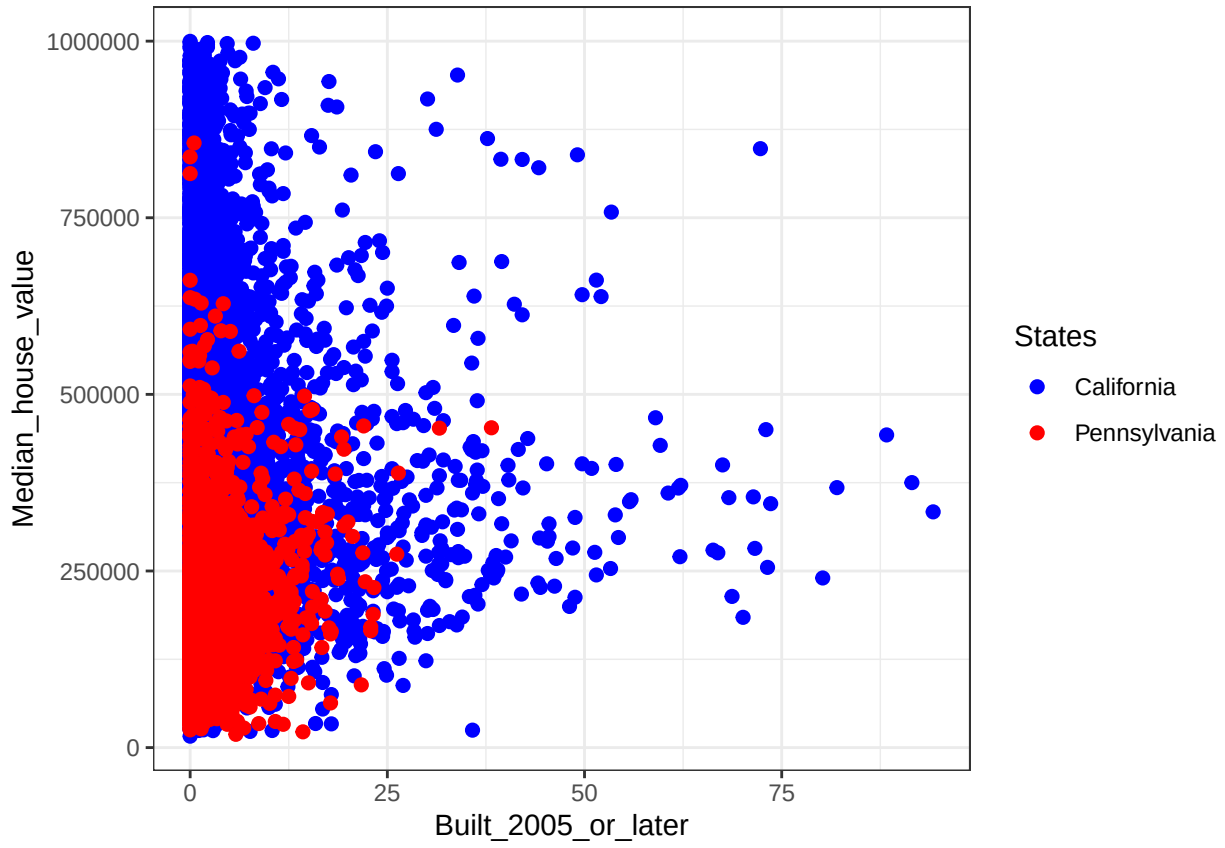
A total of 670 rows are eliminated. The answers in (c) and (e) are compatible, since from the results of the last command we can see that there is no NA values anymore.

Exercise 2 *This Very New House*

```
ca_pa %>% ggplot(aes(x = Built_2005_or_later, y = Median_house_value)) +
  geom_point(size=2) + labs(x = "Built_2005_or_later", y = "Median_house_value")
```




```
ca_pa %>% ggplot(aes(x = Built_2005_or_later, y = Median_house_value,
                     color = factor(STATEFP))) + geom_point(size=2) +
  labs(x = "Built_2005_or_later", y = "Median_house_value", color = "States") +
  theme_bw() + scale_color_manual(values = c("blue","red"),
                                  labels = c("California", "Pennsylvania"))
```



Exercise 3 *Nobody Home*

(a)

```
ca_pa <- ca_pa %>% mutate(Vacancy_rate = Vacant_units/Total_units*100)
min(ca_pa$Vacancy_rate)
```

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

```
max(ca_pa$Vacancy_rate)
```

```
## [1] 96.5311
```

```
mean(ca_pa$Vacancy_rate)
```

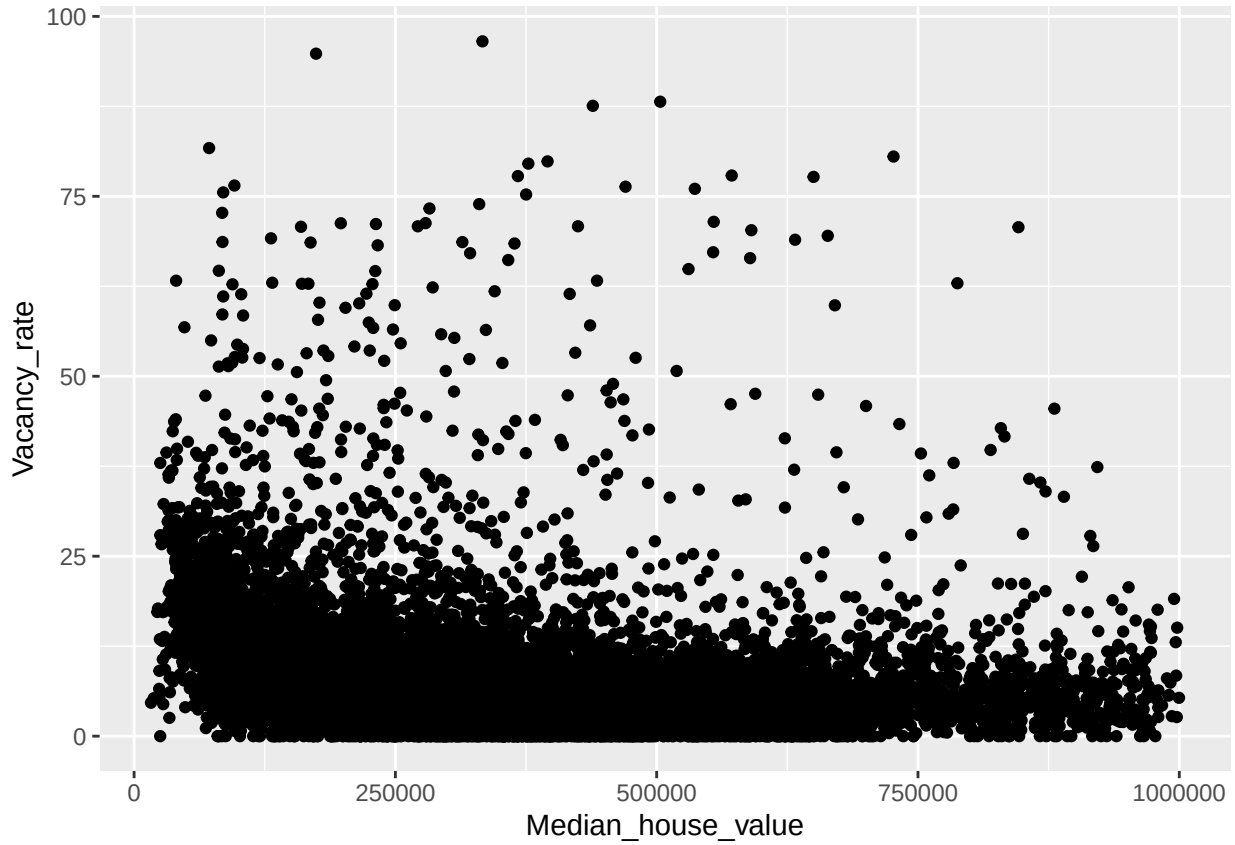
```
## [1] 8.888789
```

```
median(ca_pa$Vacancy_rate)
```

```
## [1] 6.767283
```

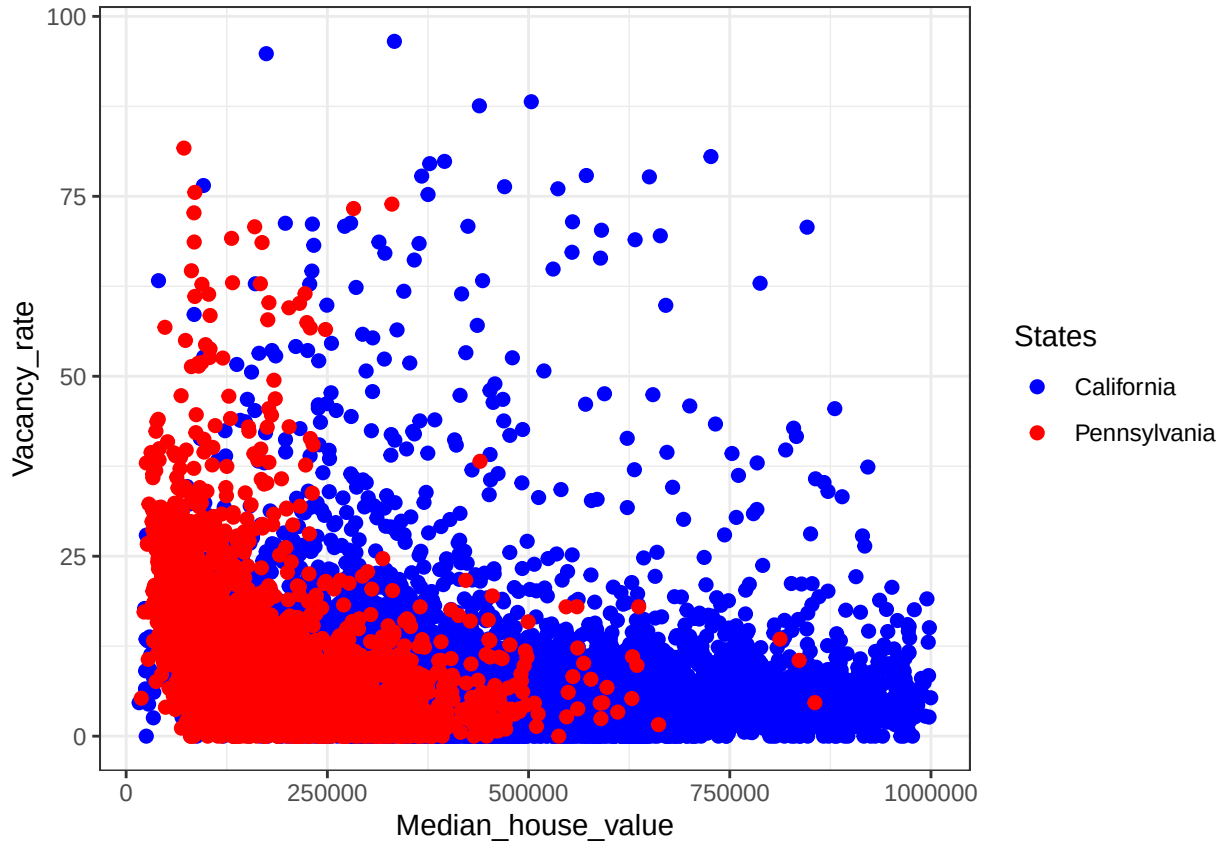
(b)

```
ggplot(data = ca_pa) + geom_point(aes(x = Median_house_value, y = Vacancy_rate)) +
  labs(x = "Median_house_value", y = "Vacancy_rate")
```



(c)

```
ca_pa %>% ggplot(aes(x = Median_house_value, y = Vacancy_rate,
                     color = factor(STATEFP))) + geom_point(size=2) +
  labs(x = "Median_house_value", y = "Vacancy_rate", color = "States") + theme_bw() +
  scale_color_manual(values = c("blue","red"), labels = c("California", "Pennsylvania"))
```



Exercise 4

```
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)
```

(a)

`acca` stores the index of rows in `ca_pa` which belongs to Alameda County, while `accamhv` stores the median house value of the corresponding rows stored in `acca`, and then `median(accamhv)` finds the median value of `accamhv`

(b)

```
which(ca_pa$STATEFP == 6 & ca_pa$COUNTYFP == 1)
```

(c)

```
# Alameda
alameda <- ca_pa %>% dplyr::filter(STATEFP == 6, COUNTYFP == 1)
sum(alameda$Built_2005_or_later) / length(alameda)
```

```
## [1] 27.56
```

```
# Santa Clara
santa_clara <- ca_pa %>% dplyr::filter(STATEFP == 6 & COUNTYFP == 85)
sum(alameda$Built_2005_or_later) / length(santa_clara)
```

```
## [1] 27.56
```

```
# Allegheny
allegheny <- ca_pa %>% dplyr::filter(STATEFP == 42 & COUNTYFP == 3)
sum(allegheny$Built_2005_or_later) / length(allegheny)
```

```
## [1] 16.17429
```

(d)

```
# (i) Whole Data
cor(ca_pa$Median_house_value, ca_pa$Built_2005_or_later)
```

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

```
# (ii) California
cor(ca_pa$Median_house_value[which(ca_pa$STATEFP == 6)],
    ca_pa$Built_2005_or_later[which(ca_pa$STATEFP == 6)])
```

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

```
# (iii) Pennsylvania
cor(ca_pa$Median_house_value[which(ca_pa$STATEFP == 42)],
    ca_pa$Built_2005_or_later[which(ca_pa$STATEFP == 42)])
```

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

```
# (iv) Alameda County
cor(alameda$Median_house_value, alameda$Built_2005_or_later)
```

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

```
# (v) Santa Clara County
cor(santa_clara$Median_house_value, santa_clara$Built_2005_or_later)
```

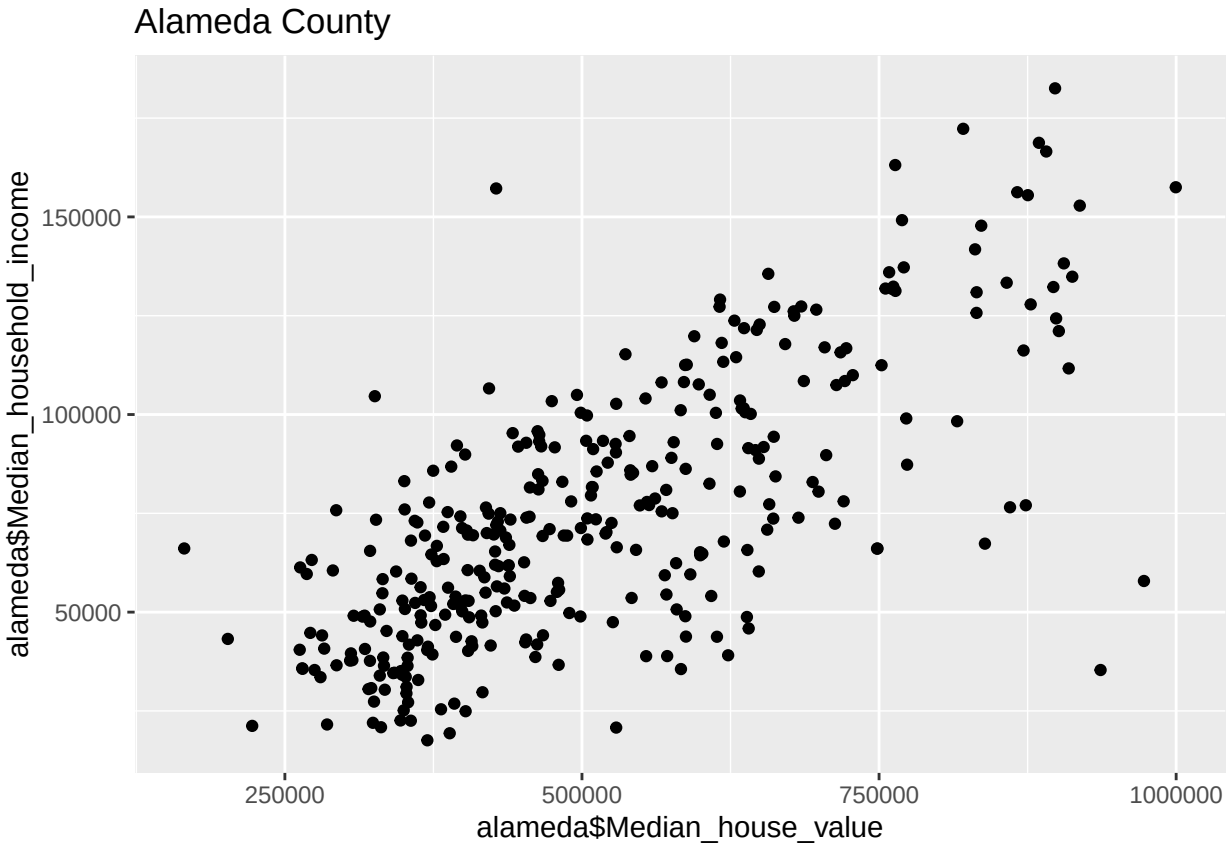
```
## [1] -0.1726203
```

```
# (vi) Allegheny County
cor(allegheny$Median_house_value, allegheny$Built_2005_or_later)
```

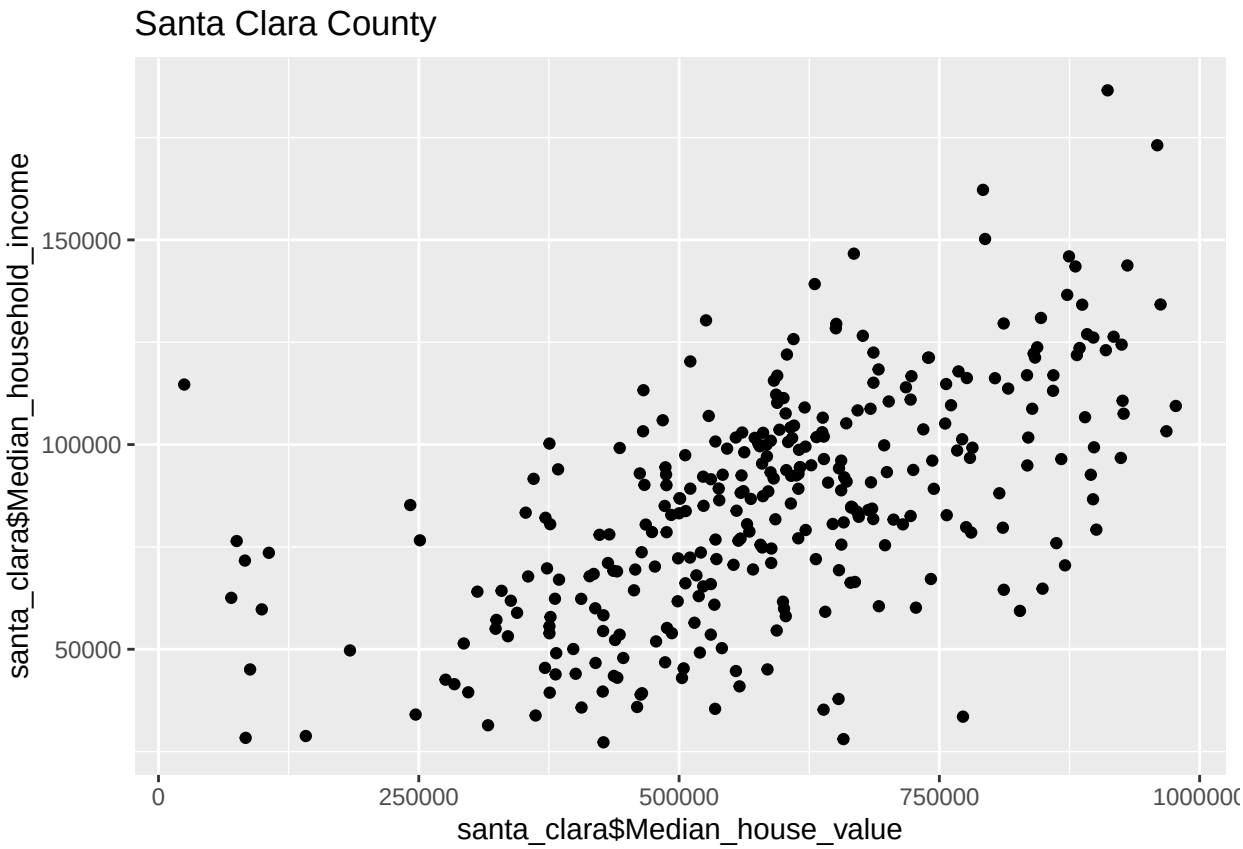
```
## [1] 0.1939652
```

(e)

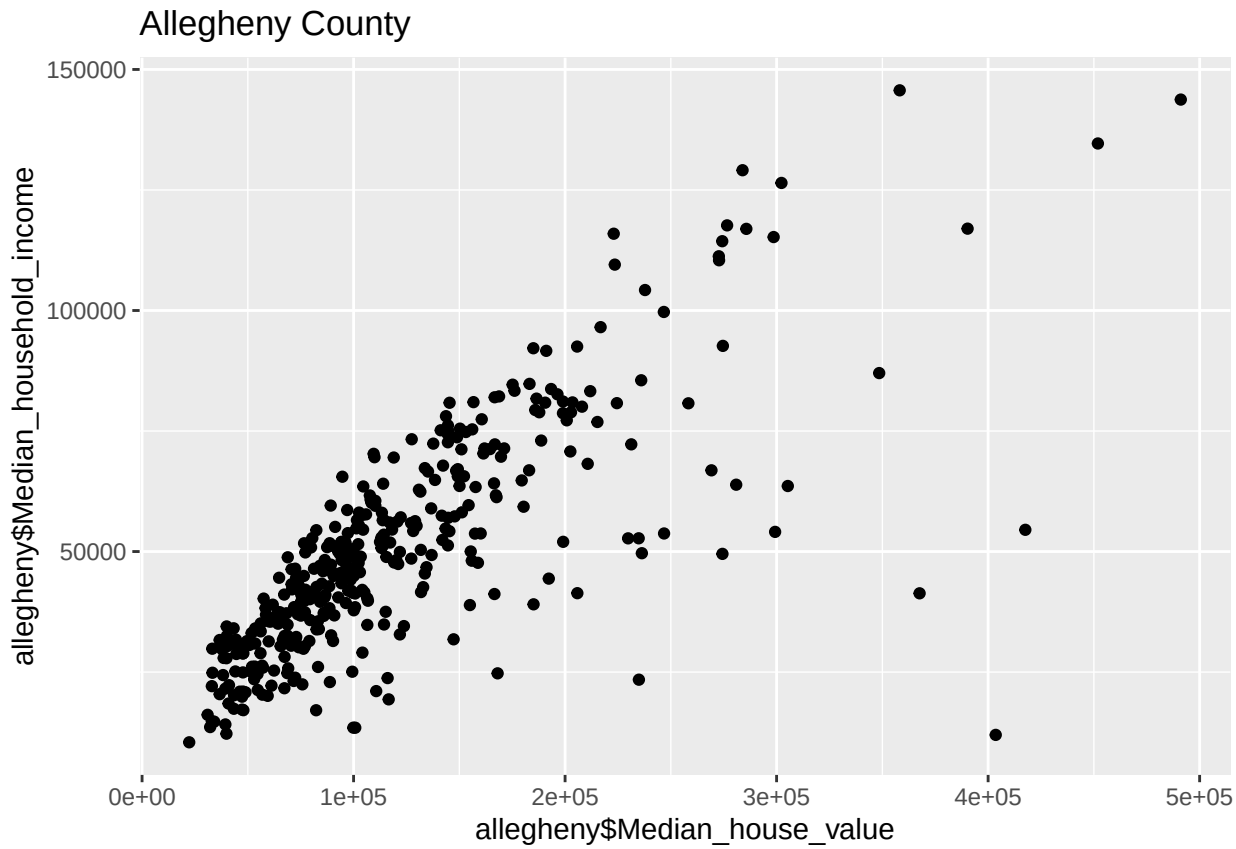
```
ggplot() + geom_point(aes(x = alameda$Median_house_value, y = alameda$Median_household_income)) + labs(
```




```
ggplot() + geom_point(aes(x = santa_clara$Median_house_value, y = santa_clara$Median_household_income))
```




```
ggplot() + geom_point(aes(x = allegheny$Median_house_value, y = allegheny$Median_household_income)) + la
```



MB.Ch1.11

```
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
```

From the first use of `table()`, we see that the function automatically counts the number of `female` and `male` in the vector. The second use shows us that the function counts the frequency of the variables specified in `levels`, but in the third use we know that the name of the variable is case-sensitive. When using the `table()`, we can also count the number of `NA` values.

MB.Ch1.12

```
cutoff <- function(x,a) {
  return (x[which(x > a)])
}
```

(a)

```
x <- seq(1,100)
cutoff(x,90)
```

```
##  [1]  91  92  93  94  95  96  97  98  99 100
```

(b)

```
library(Devore7)
```

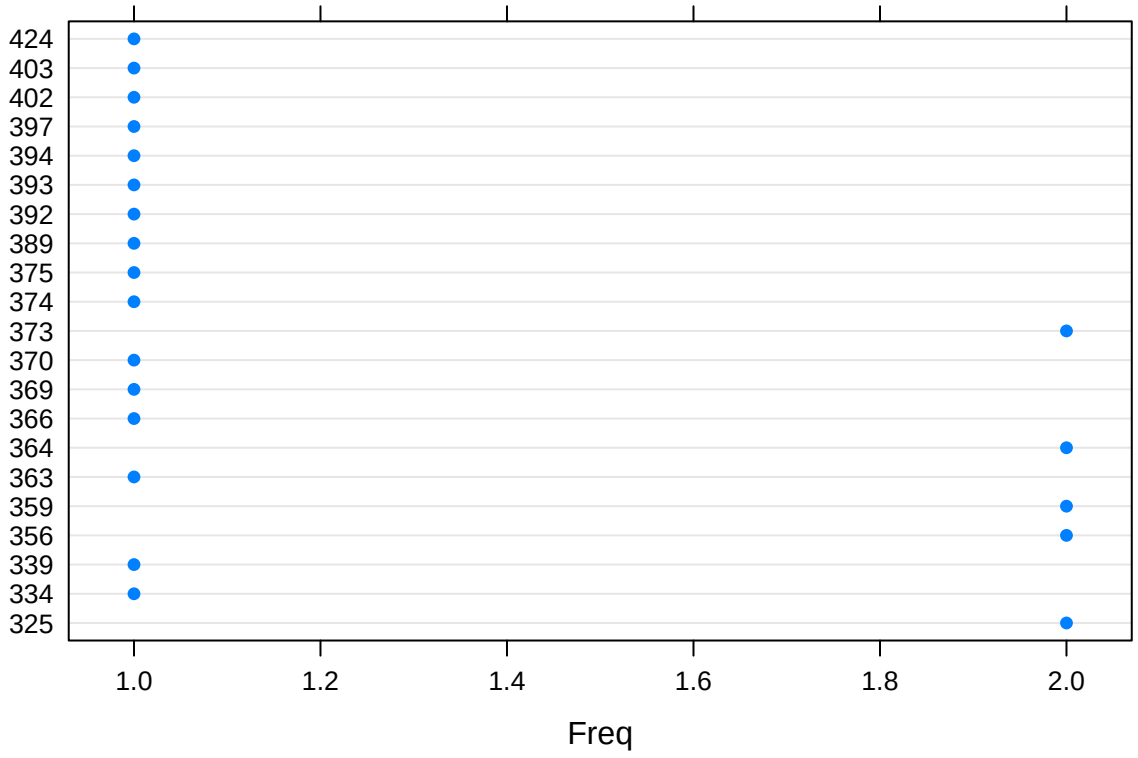
```
## Loading required package: MASS
```

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

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

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

```
dotplot(ex01.36)
```




```
length(cutoff(ex01.36$C1,420))/dim(ex01.36)[1]
```

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

MB.Ch1.18

```
library(MASS)
rabbit1 <- unstack(Rabbit,BPchange~Animal)
rabbit2 <- unstack(Rabbit,Dose~Animal)
rabbit3 <- unstack(Rabbit,Treatment~Animal)
rabbit <- cbind(rabbit3[5],rabbit2[5],rabbit1)
colnames(rabbit) <- c("Treatment","Dose","R1","R2","R3","R4","R5")
rabbit
```

```
##    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
```