Homework 2

zhang zhuohan

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`.

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

- b. How many rows and columns does the dataframe have?

```
nrow(ca_pa)
```

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

```
ncol(ca_pa)
```

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

- c. Run this command, and explain, in words, what this does:

```
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
```

It's used to count the number of missing data in each column.

- 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.

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

- e. How many rows did this eliminate?

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

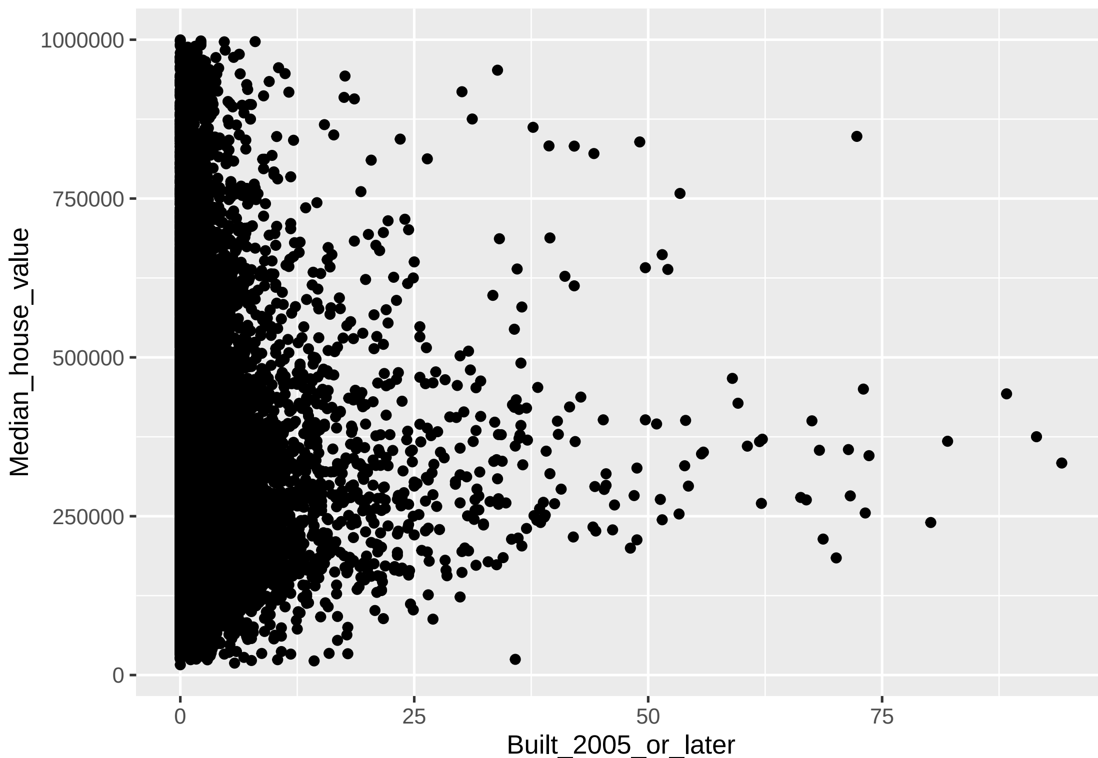
```
## [1] 670
```

- f. Are your answers in (c) and (e) compatible? Explain.
  It's compatible, because some rows may have more than one missing data.

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.

```
library(ggplot2)
ggplot(ca_pa, aes(Built_2005_or_later, Median_house_value)) +
    geom_point() # + geom_smooth()
```



- 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.

```
ca_pa$STATEFP <- factor(ca_pa$STATEFP)
ggplot(ca_pa, aes(Built_2005_or_later, Median_house_value,
       color=STATEFP)) + geom_point()
```

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.

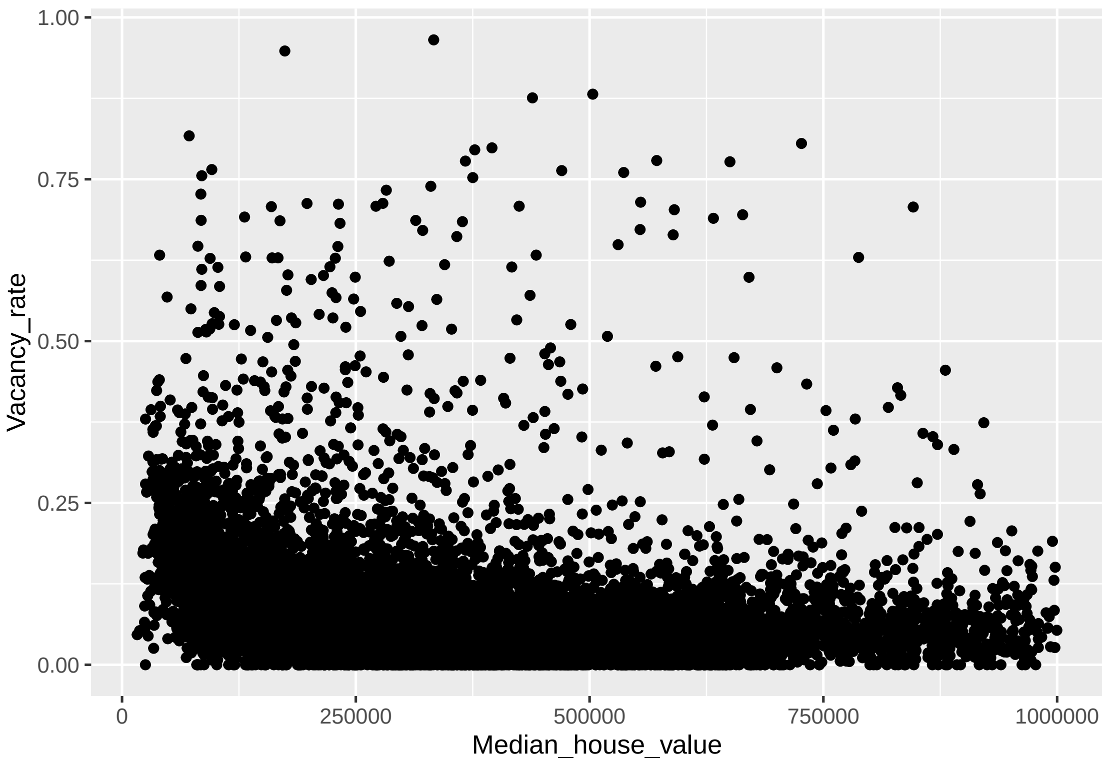
- a. Add a new column to the dataframe which contains the vacancy rate. What are the minimum, maximum, mean, and median vacancy rates?

```
library(dplyr)
ca_pa <- ca_pa |> mutate(Vacancy_rate = Vacant_units/Total_units)
summary(ca_pa$Vacancy_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03846 0.06767 0.08889 0.10921 0.96531
```

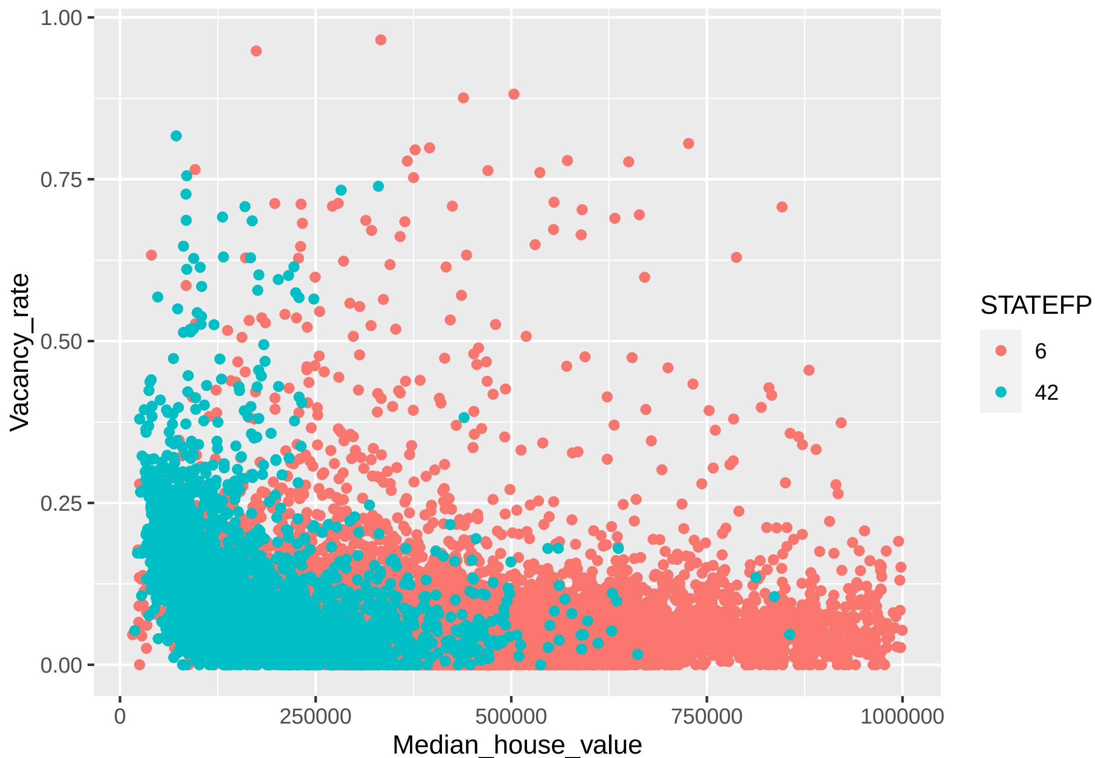
- b. Plot the vacancy rate against median house value.

```
ggplot(ca_pa, aes(Median_house_value, Vacancy_rate))+
    geom_point()
```



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

```
ggplot(ca_pa, aes(Median_house_value, Vacancy_rate,
       color=STATEFP)) + geom_point()
```



The prices of the houses are lower in Pennysylvania, so the points in the iamge are concentrated in the lower left portion. And the distributions of vacancy rate are generally similar.

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.

  Objective: To calculate the median number of total units in Alameda County, California.

- 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.

```
library(dplyr)
ca_pa %>% filter(STATEFP==6, COUNTYFP==1) %>%
      select(Total_units) %>% unlist() %>% median()
```

```
## [1] 1606
```

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

```
result <- ca_pa %>%
    filter(((STATEFP==6) & (COUNTYFP %in% c(1,85))) |
           ((STATEFP==42) & (COUNTYFP)==3)) %>%
    group_by(COUNTYFP) %>%
    summarise(rate_2005_later=mean(Built_2005_or_later/Total_units))
result[,1] <- c("Alameda", "Allegheny", "Santa Clara")
result
```

```
## # A tibble: 3 x 2
##   COUNTYFP    rate_2005_later
##   <chr>                 <dbl>
## 1 Alameda             0.00258
## 2 Allegheny           0.00119
## 3 Santa Clara         0.00194
```

- 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 the following situations?
  - (i) the whole data

```
unq_cor <- function(fea){
    return(cor(fea$Median_house_value, fea$Built_2005_or_later))
}
unq_cor(ca_pa)
```

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

  - (ii) all of Pennsylvania

```
ca_pa %>% filter(STATEFP==42) %>% unq_cor
```

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

  - (iii) Alameda County

```
ca_pa %>% filter(STATEFP==6 & COUNTYFP==1) %>% unq_cor
```

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

  - (iv) Santa Clara County

```
ca_pa %>% filter(STATEFP==6 & COUNTYFP==85) %>% unq_cor
```

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

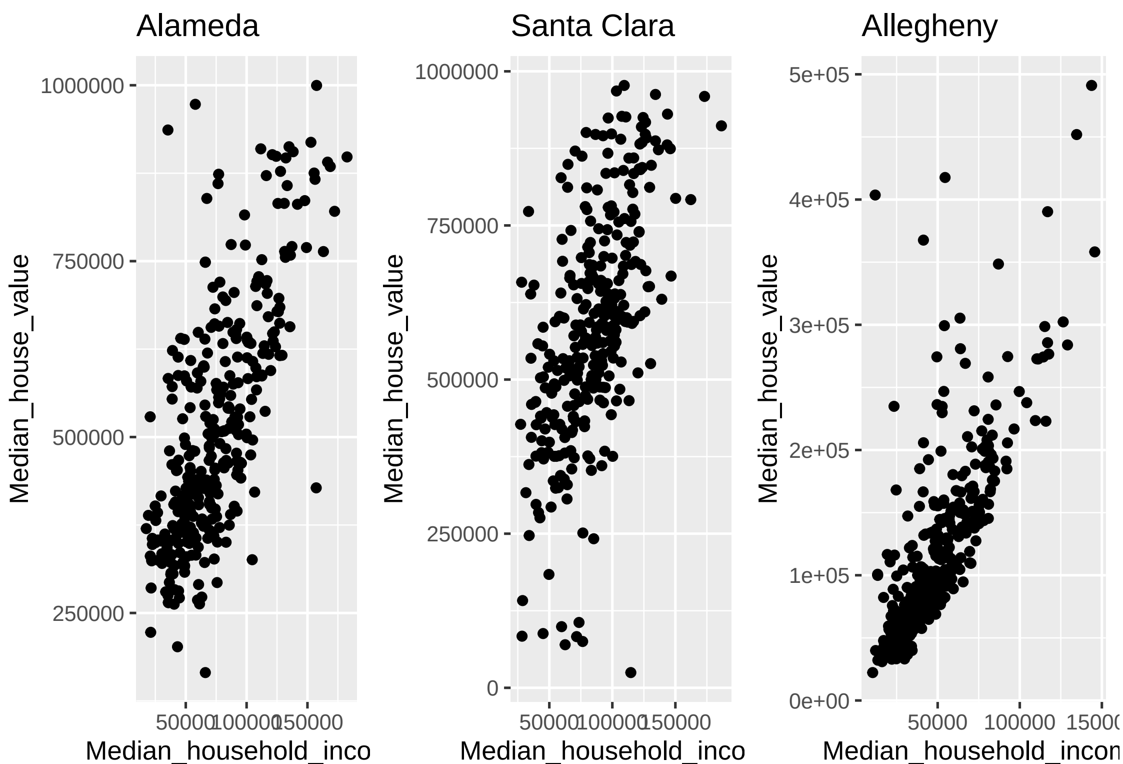
  - (vi) Allegheny County

```
ca_pa %>% filter(STATEFP==42 & COUNTYFP==3) %>% unq_cor
```

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

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

```
library(ggpubr)
p1 <- ca_pa %>% filter(STATEFP==6 & COUNTYFP==1) %>%
        ggplot(aes(Median_household_income, Median_house_value))+
        geom_point() + ggtitle("Alameda")
p2 <- ca_pa %>% filter(STATEFP==6 & COUNTYFP==85) %>%
        ggplot(aes(Median_household_income, Median_house_value))+
        geom_point() + ggtitle("Santa Clara")
p3 <- ca_pa %>% filter(STATEFP==42 & COUNTYFP==3) %>%
        ggplot(aes(Median_household_income, Median_house_value))+
        geom_point() + ggtitle("Allegheny")
ggarrange(p1,p2,p3, ncol=3, nrow=1)
```



5. (MB.Ch1.11.) Run the following code, and explain the output from the successive uses of table().

```
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
```

there are four things it does.

- First, it makes `gender` a factor and generate a table.
- Second, it changes the order in the table `gender`.
- Third, it makes the level `"Male", "female"`, but there is no "Male" in `gender`.
- Last, it shows the count number of NA, which is the number of "female".

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

```
exceed_some_val <- function(x, val){
    n = length(x); count = 0
    for(i in x)
        if(i > val)
            count <- count+1
    return (count/n)
}
```

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

  ```
  x <- seq(1,100); val <- 40
  exceed_some_val(x, val)
  ```

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

- 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.

  ```
  # There is no such package in recent versions of R.
  ```

7. (MB.Ch1.18.) The Rabbit data frame in the MASS library contains blood pressure change measurements onfive rabbits (labeled as R1, R2, . . . ,R5) under various control and treatment conditions. Read the help filefor more information. Use the `unstack()` function (three times) to convert Rabbit to the special form.

```
library(MASS)
Dose <- unstack(Rabbit, Dose~Animal)[,1]
Treatment <- unstack(Rabbit, Treatment~Animal)[,1]
BPchange <- unstack(Rabbit, BPchange~Animal)
data.frame(Treatment, Dose, BPchange)
```

```
##    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
```