Homework 2 - Report

3170105743 李政达

2020/7/7

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`.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

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

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

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   STATEFP = col_character(),
##   COUNTYFP = col_character(),
##   TRACTCE = col_character(),
##   GEO.display.label = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

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

```
rows <- dim(ca_pa)[1]
columns <- dim(ca_pa)[2]
```

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

Ans: this command can figure out the number of missing values in every column.

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

```
##                            X1                     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
```

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?

```
rows - dim(ca_pa)[1]
```

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

f. Are your answers in (c) and (e) compatible? Explain.

Ans: They are compatible. The command in (c) check the number of missing values in every column, and the command in (e) check the number of rows with incomplete data. We can infer that after purging, the number of missing values in every column will be zero. And we can use the command below to check out the truth.

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

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

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.