

homework2

3170102187 孙晨旭

1 Loading and cleaning

a.

```
ca_pa <- read_csv("~/github/Rcourse_1/data/calif_penn_2011.csv")
```

b.rows and columns

```
dim(ca_pa)
```

```
## [1] 11275      34
```

c.

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

```
##           Bedrooms_4      Bedrooms_5_or_more      Owners
##                98                98                100
##           Renters Median_household_income Mean_household_income
##                100                115                126
```

apply 函数将 is.na() 作用于 ca_pa 的每一个元素，colSums() 统计每一列中为 na 类型的元素个数。

d.e.

```
ca_pa_1<-na.omit(ca_pa)
(delete_nrows = nrow(ca_pa)-nrow(ca_pa_1))
```

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

f.

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

```
## [1] 3034
```

```
sum(rowSums(apply(ca_pa,c(1,2),is.na))!=F)
```

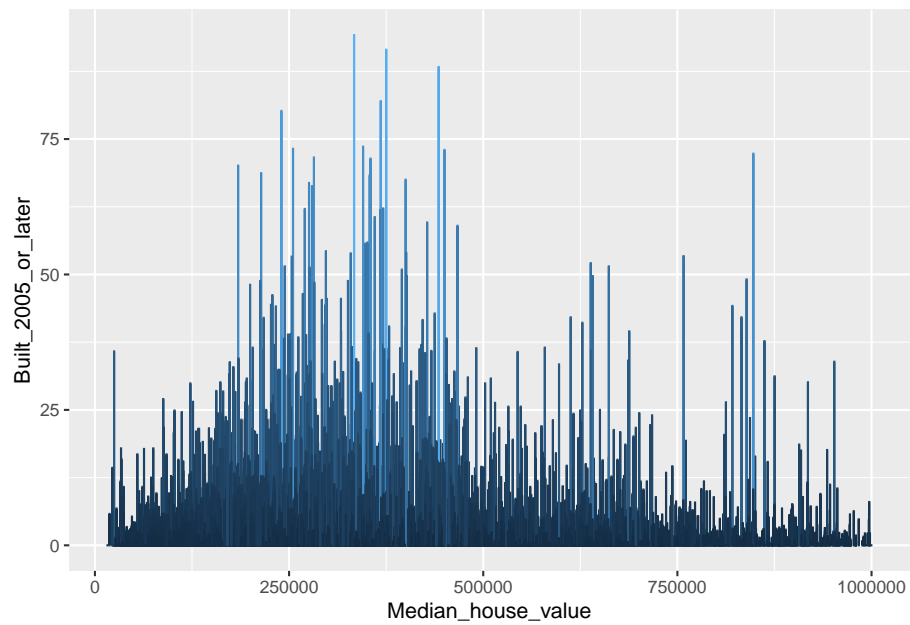
```
## [1] 670
```

c 中统计为每一列中 na 的数量，累加后为 na 元素的数量。由于每一行可以包含多个 na，所以，最终统计至少包含一个 na 的行数为 670，所以 (c),(e) 相符。

2 This Very New House

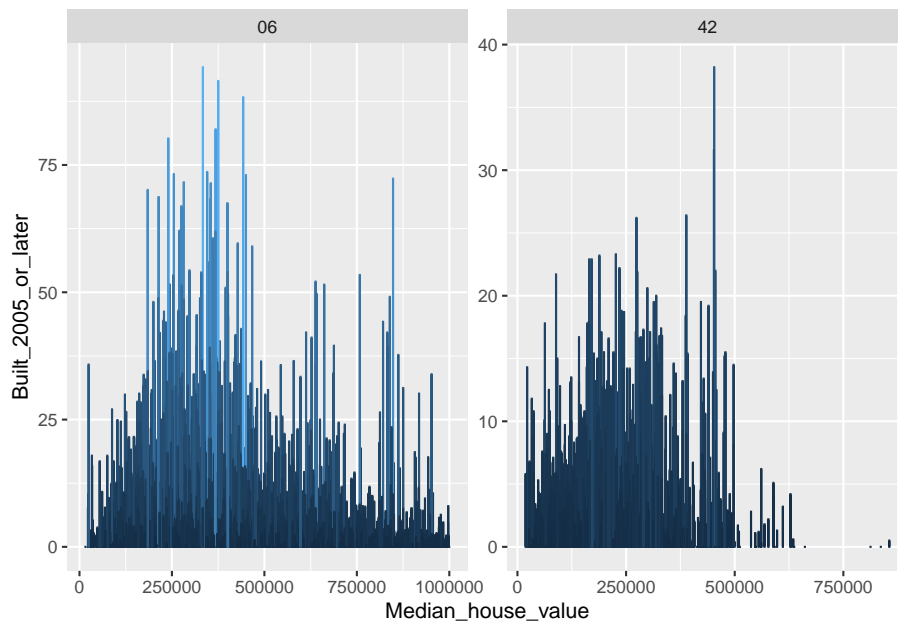
a.

```
ca_pa_1 %>% ggplot(aes(x = Median_house_value,y = Built_2005_or_later,  
                      col = Built_2005_or_later)) +  
  geom_col(position = "dodge") +  
  theme(legend.position = "none")
```



b.

```
ca_pa_1 %>% ggplot(aes(x = Median_house_value, y = Built_2005_or_later,
                        col = Built_2005_or_later)) +
  geom_col(position = "dodge") +
  facet_wrap(~STATEFP, scales = "free") +
  theme(legend.position = "none")
```



3 Nobody Home

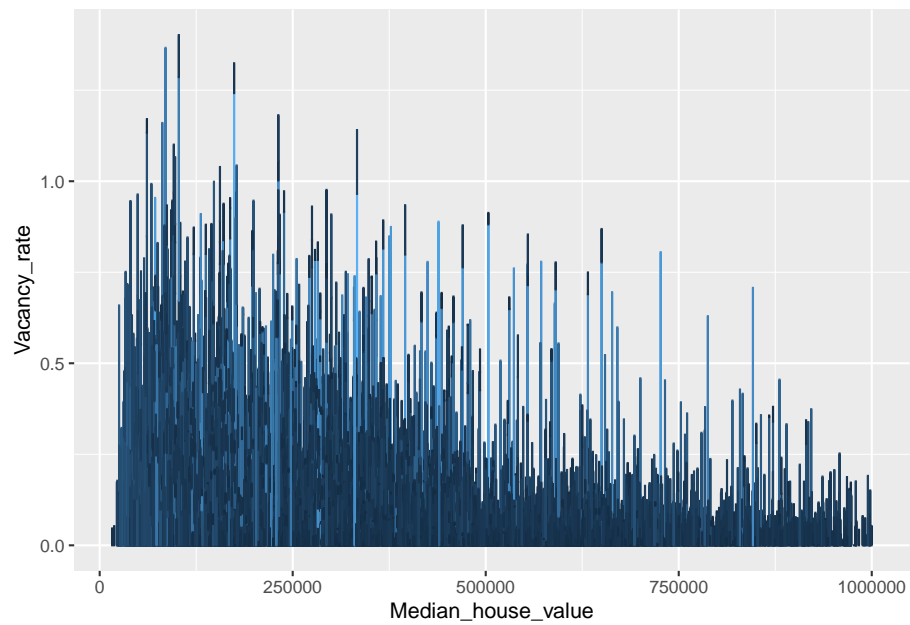
a.

```
ca_pa_2 <- ca_pa_1 %>% mutate(Vacancy_rate = Vacant_units/Total_units)
summary(ca_pa_2$Vacancy_rate)
```

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

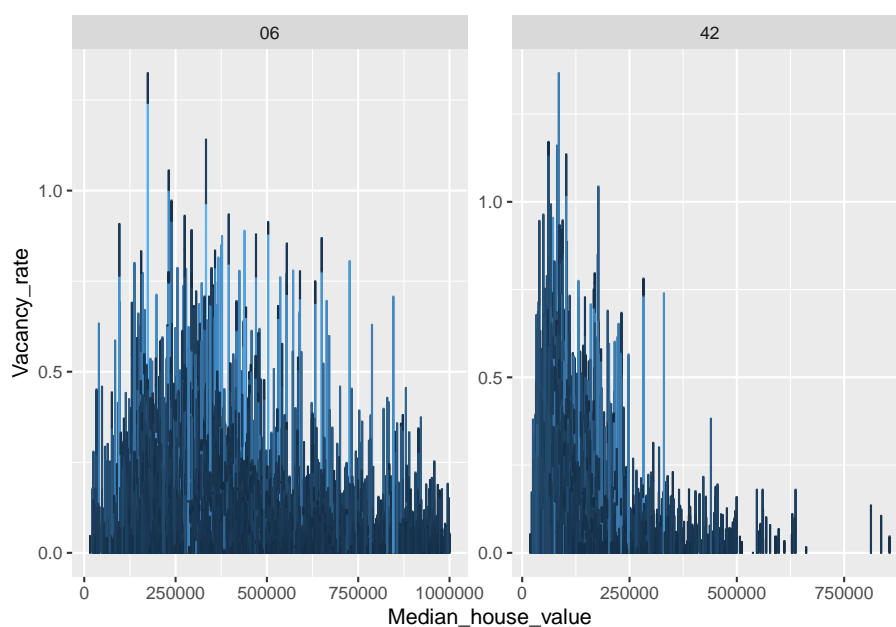
b.

```
ca_pa_2 %>% ggplot(aes(x = Median_house_value, y = Vacancy_rate
                        , col = Vacancy_rate)) +
  geom_col() +
  theme(legend.position = "none")
```



c.

```
ca_pa_2 %>% ggplot(aes(x = Median_house_value, y = Vacancy_rate,  
                        col = Vacancy_rate)) +  
  geom_col() +  
  theme(legend.position = "none") +  
  facet_wrap(~STATEFP, scales = "free")
```



可以看出，加州各个价位的房子都有空闲，而宾州的空房子主要集中在较低价位。此外，结合房屋建造的时间，说明可能宾州新建的房屋空闲概率较大。

4 COUNTYFP

- 前一个循环用于找出 `ca_pa` 中为于加州 Alameda County 的数据下标，后一个循环根据以上下标提取出相应 `Median_house_value` 这一列的数据，最后求中位数。
-

```
median((ca_pa_1 %>% filter(STATEFP == "06",COUNTYFP=="001"))$Median_house_value)
```

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

5 MB.Ch 1.11.

```
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)
```

首先建立一个 factor 类型的数据 gender，其中前 91 个为 “female” 后 92 个为 “male”。之后重新设置 levels，若能在 gender 内找到，就正常返回，

若有设定外的 levels 则将其返回为 NA。默认下 table 不显示 NA 的值，有关 NA 的输出由参数 useNA 控制。最后一个操作中 exclude=NULL 参数表示不排除任何数据，但却将 useNA 默认为 “ifany”，若 NA 个数为正，此时就会输出一个 levels。

6 MB.Ch 1.12.

```
exceed <- function(x,cutoff){
  sum(x>cutoff)/length(x)
}
exceed(1:100,60)
```

```
## [1] 0.4
```

```
exceed(1:100,0)
```

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

```
exceed(1:100,19.5)
```

```
## [1] 0.81
```

7 MB.Ch.1.18.

```
cbind(Treatment = unstack(Rabbit, Treatment ~ Animal)[,1],
      Dose = unstack(Rabbit, Dose ~ Animal)[,1],
      unstack(Rabbit, BPchange ~ Animal))
```

```
##      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