Homework 2

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*
   1. Load the data into a dataframe called ca\_pa.

ca\_pa=read.csv("data/calif\_penn\_2011.csv")

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

dim(ca\_pa)

## [1] 11275 34

c. Run this command, and explain, in words, what this does:  
answer:求每个变量中，观测值为na的个数

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

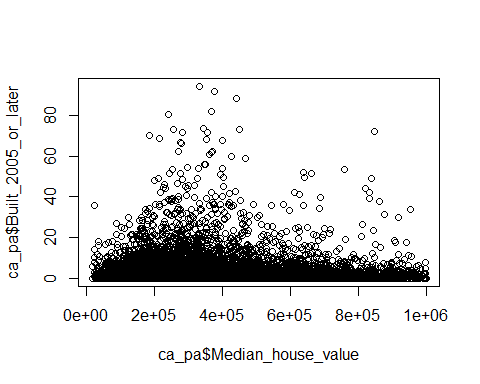
d. The function `na.omit()` takes a dataframe and returns a new dataframe, omitting any row containing an NA value. Use it to purge the data set of rows with incomplete data.

ca\_pa=na.omit(ca\_pa)

e. How many rows did this eliminate?  
answer:670  
f. Are your answers in (c) and (e) compatible? Explain.  
answer:不一样，因为（c)中同一列可能会有不止一个na，而（d)中只要一列中有一个na就要被删去

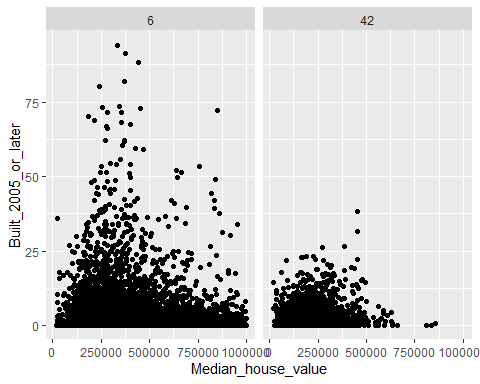
1. *This Very New House*
   1. The variable Built\_2005\_or\_later indicates the percentage of houses in each Census tract built since 2005. Plot median house prices against this variable.

plot(ca\_pa$Median\_house\_value,ca\_pa$Built\_2005\_or\_later)



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.

ggplot(data=ca\_pa)+  
 geom\_point(aes(x=Median\_house\_value,y=Built\_2005\_or\_later))+  
 facet\_wrap(~STATEFP)



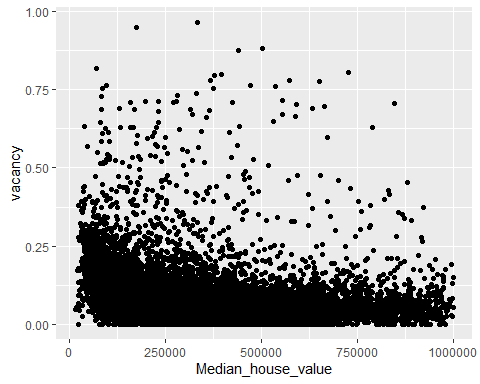
1. *Nobody Home*  
   The vacancy rate is the fraction of housing units which are not occupied. The dataframe contains columns giving the total number of housing units for each Census tract, and the number of vacant housing units.
   1. Add a new column to the dataframe which contains the vacancy rate. What are the minimum, maximum, mean, and median vacancy rates?

ca\_pa=ca\_pa %>% mutate(vacancy=Vacant\_units/Total\_units)  
summary(ca\_pa$vacancy)

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

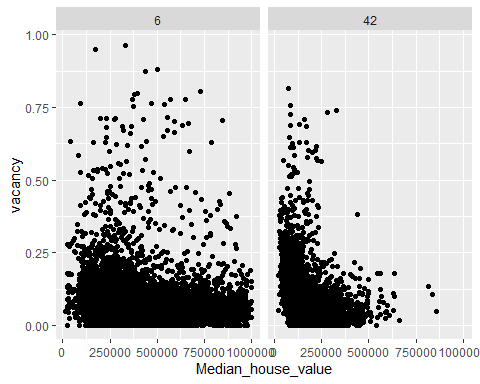
b. Plot the vacancy rate against median house value.

ggplot(data=ca\_pa)+  
 geom\_point(aes(x=Median\_house\_value,y=vacancy))



c. Plot vacancy rate against median house value separately for California and for Pennsylvania. Is there a difference?  
answer:在宾夕法尼亚中，住房空置率随着房价的上升而明显减小；而加州没有这么明显的趋势

ggplot(data=ca\_pa)+  
 geom\_point(aes(x=Median\_house\_value,y=vacancy))+  
 facet\_wrap(~STATEFP)



1. 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).
   1. Explain what the block of code at the end of this question is supposed to accomplish, and how it does it. answer:为了求加州编号为1的县中的median\_house\_value的中位数。方法：先将加州编号为1的县的列数选出，然后将这些列中的median\_house\_value构成一个新的数组，再求这个数组的中位数。
   2. 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.

ca\_pa %>% filter(COUNTYFP==1,STATEFP==6) %>% summarise(median(Median\_house\_value))

## median(Median\_house\_value)  
## 1 474050

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

ca\_pa %>% filter(COUNTYFP==1,STATEFP==6) %>% mutate(Alameda\_pr=Built\_2005\_or\_later/Total\_units) %>% summarise(mean(Alameda\_pr))

## mean(Alameda\_pr)  
## 1 0.002583202

ca\_pa %>% filter(COUNTYFP==85,STATEFP==6) %>% mutate(Santa\_Clara\_pr=Built\_2005\_or\_later/Total\_units) %>% summarise(mean(Santa\_Clara\_pr))

## mean(Santa\_Clara\_pr)  
## 1 0.001943991

ca\_pa %>% filter(COUNTYFP==3,STATEFP==42) %>% mutate(Allegheny\_pr=Built\_2005\_or\_later/Total\_units) %>% summarise(mean(Allegheny\_pr))

## mean(Allegheny\_pr)  
## 1 0.001185988

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?

cor\_whole=ca\_pa %>% mutate(pr=Built\_2005\_or\_later/Total\_units) %>% summarise(cor(pr,Median\_house\_value))  
cor\_whole

## cor(pr, Median\_house\_value)  
## 1 -0.008761551

cor\_ca=ca\_pa %>% mutate(pr=Built\_2005\_or\_later/Total\_units) %>% filter(STATEFP==6) %>% summarise(cor(pr,Median\_house\_value))  
cor\_ca

## cor(pr, Median\_house\_value)  
## 1 -0.06240432

cor\_pa=ca\_pa %>% mutate(pr=Built\_2005\_or\_later/Total\_units) %>% filter(STATEFP==42) %>% summarise(cor(pr,Median\_house\_value))  
cor\_pa

## cor(pr, Median\_house\_value)  
## 1 0.1980093

cor\_Alameda=ca\_pa %>% mutate(pr=Built\_2005\_or\_later/Total\_units) %>% filter(STATEFP==6,COUNTYFP==1) %>% summarise(cor(pr,Median\_house\_value))  
cor\_Alameda

## cor(pr, Median\_house\_value)  
## 1 -0.009840453

cor\_Santa\_Clara=ca\_pa %>% mutate(pr=Built\_2005\_or\_later/Total\_units) %>% filter(STATEFP==6,COUNTYFP==85) %>% summarise(cor(pr,Median\_house\_value))  
cor\_Santa\_Clara

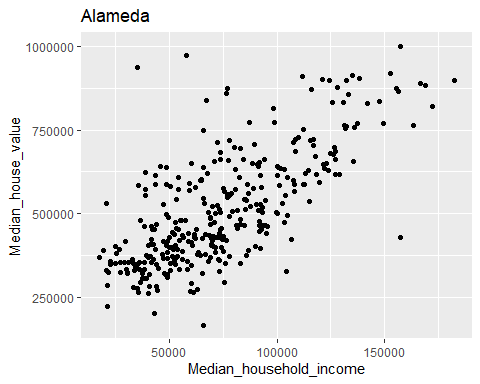
## cor(pr, Median\_house\_value)  
## 1 -0.1732909

cor\_Allegheny=ca\_pa %>% mutate(pr=Built\_2005\_or\_later/Total\_units) %>% filter(STATEFP==42,COUNTYFP==3) %>% summarise(cor(pr,Median\_house\_value))  
cor\_Allegheny

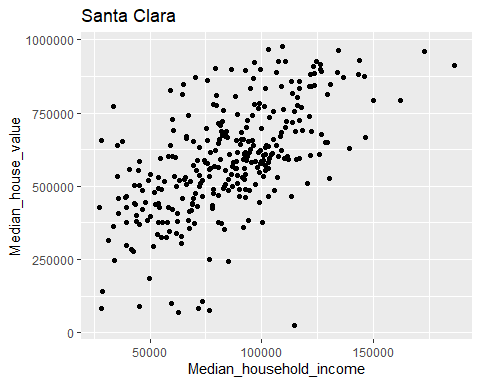
## cor(pr, Median\_house\_value)  
## 1 0.04878986

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.)

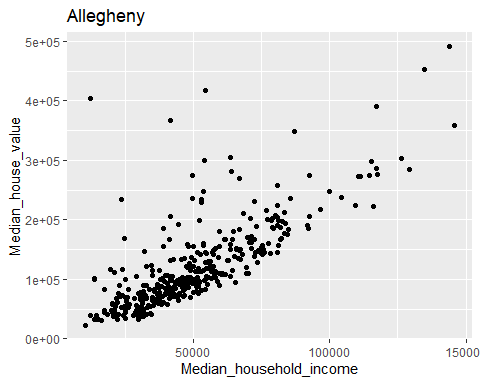
ggplot(data=ca\_pa %>%filter(COUNTYFP==1,STATEFP==6))+  
 geom\_point(aes(x=Median\_household\_income,y=Median\_house\_value))+  
 labs(title = "Alameda")



ggplot(data=ca\_pa %>%filter(COUNTYFP==85,STATEFP==6))+  
 geom\_point(aes(x=Median\_household\_income,y=Median\_house\_value))+  
 labs(title = "Santa Clara")



ggplot(data=ca\_pa %>%filter(COUNTYFP==3,STATEFP==42))+  
 geom\_point(aes(x=Median\_household\_income,y=Median\_house\_value))+  
 labs(title = "Allegheny")



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)

MB.Ch1.11. Run the following code:

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

Explain the output from the successive uses of table(). answer:当level的名称一致时，table的展现不会随着名称次序的改变而改变，但是当level名称改变时，结果就不一样了，原先的名称会出现在na中  
MB.Ch1.12. Write a function that calculates the proportion of values in a vector x that exceed some value cutoff.

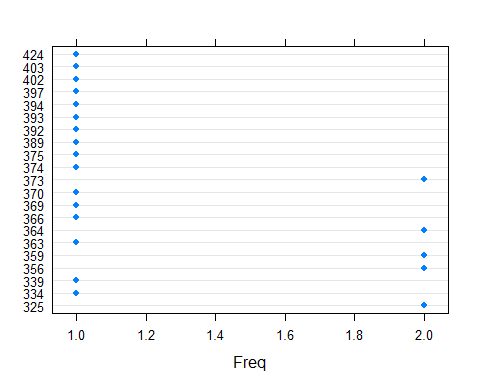
1. Use the sequence of numbers 1, 2, . . . , 100 to check that this function gives the result that is expected.

func=function(x,y){  
 pr=sum(x>y)/length(x)  
 return(pr)  
}  
func(seq(1,100),10)

## [1] 0.9

1. 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.

dotplot(ex01.36)



func(ex01.36$C1,420)

## [1] 0.03846154

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

….

A=Rabbit %>% filter(Treatment=='Control',Dose==6.25|Dose==12.5) %>% dplyr::select(-Run) %>% group\_by(Dose)  
B=unstack(A,BPchange~Animal)  
data.frame(Treatment=c("Control","Control"),Dose=c(6.25,12.5),B)

## Treatment Dose R1 R2 R3 R4 R5  
## 1 Control 6.25 0.5 1.00 0.75 1.25 1.5  
## 2 Control 12.50 4.5 1.25 3.00 1.50 1.5