Descriptive Stat

#使用AER包里提供的示例数据  
data("CPS1985")  
  
#查看数据结构和盖帽，先变成dplyr格式操作方便  
  
CPS1985 <- tbl\_df(CPS1985)  
CPS1985

## Source: local data frame [534 x 11]  
##   
## wage education experience age ethnicity region gender occupation  
## 1 5.10 8 21 35 hispanic other female worker  
## 2 4.95 9 42 57 cauc other female worker  
## 3 6.67 12 1 19 cauc other male worker  
## 4 4.00 12 4 22 cauc other male worker  
## 5 7.50 12 17 35 cauc other male worker  
## 6 13.07 13 9 28 cauc other male worker  
## 7 4.45 10 27 43 cauc south male worker  
## 8 19.47 12 9 27 cauc other male worker  
## 9 13.28 16 11 33 cauc other male worker  
## 10 8.75 12 9 27 cauc other male worker  
## .. ... ... ... ... ... ... ... ...  
## Variables not shown: sector (fctr), union (fctr), married (fctr)

str(CPS1985)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 534 obs. of 11 variables:  
## $ wage : num 5.1 4.95 6.67 4 7.5 ...  
## $ education : num 8 9 12 12 12 13 10 12 16 12 ...  
## $ experience: num 21 42 1 4 17 9 27 9 11 9 ...  
## $ age : num 35 57 19 22 35 28 43 27 33 27 ...  
## $ ethnicity : Factor w/ 3 levels "cauc","hispanic",..: 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 2 levels "south","other": 2 2 2 2 2 2 1 2 2 2 ...  
## $ gender : Factor w/ 2 levels "male","female": 2 2 1 1 1 1 1 1 1 1 ...  
## $ occupation: Factor w/ 6 levels "worker","technical",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ sector : Factor w/ 3 levels "manufacturing",..: 1 1 1 3 3 3 3 3 1 3 ...  
## $ union : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 1 ...  
## $ married : Factor w/ 2 levels "no","yes": 2 2 1 1 2 1 1 1 2 1 ...

#看工资的统计量  
summary(CPS1985$wage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 5.25 7.78 9.02 11.20 44.50

var(CPS1985$wage)

## [1] 26.41

sd(CPS1985$wage)

## [1] 5.139

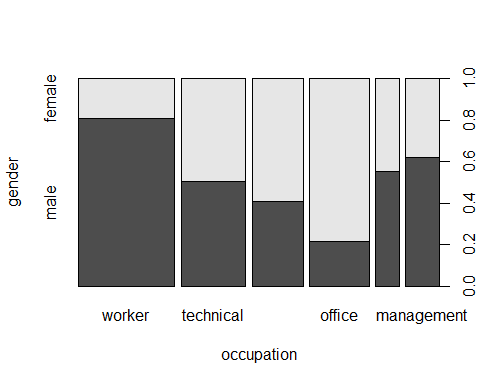
#统计类别变量  
  
table(CPS1985$occupation)

##   
## worker technical services office sales management   
## 156 105 83 97 38 55

table(CPS1985$occupation,CPS1985$gender)

##   
## male female  
## worker 126 30  
## technical 53 52  
## services 34 49  
## office 21 76  
## sales 21 17  
## management 34 21

#马赛克图  
plot(gender~occupation, data=CPS1985)



#先带上CPS1985数据集，可以直接用变量  
attach(CPS1985)  
#计算两列的Pearson相关系数  
cor(log(wage),education)

## [1] 0.3804

#计算两列的Spearman相关系数  
cor(log(wage), education, method = "spearman")

## [1] 0.3813

#计算数据集的相关系数矩阵，此时只需要一个含多列的dataframe即可  
cor(select(CPS1985,wage,education,age,experience),method = "spearman")

## wage education age experience  
## wage 1.0000 0.3813 0.2523 0.1690  
## education 0.3813 1.0000 -0.1075 -0.3061  
## age 0.2523 -0.1075 1.0000 0.9729  
## experience 0.1690 -0.3061 0.9729 1.0000

cor(select(CPS1985,wage,education,age,experience))

## wage education age experience  
## wage 1.00000 0.3819 0.177 0.08706  
## education 0.38192 1.0000 -0.150 -0.35268  
## age 0.17697 -0.1500 1.000 0.97796  
## experience 0.08706 -0.3527 0.978 1.00000

#按照民族统计平均工资，tapply(被统计的对象，分组的量，函数)  
tapply(wage,ethnicity,mean)

## cauc hispanic other   
## 9.278 7.283 8.058