Data\_Cleaning

离散数值变量转化为类别变量，并加上标签,需要使用car包

school <- c(2, 1, 1, 2, 0, 1, 1,3,3,5,4,35,6,5,8)  
#recode接收的参数:第一个是一个数值变量，第二个参数是归类参数  
#lo表示最小值,hi标示最大值，但是必须用在lo:5,7:hi这样的仿序列语法里  
#c(5,8)也是允许的，但是c(5,8,1:3)这样的语法是不行的，作者没定义  
#else表示所有没匹配的  
#最后的参数as.factor.result如果不设置，默认结果是一个string向量  
  
f1\_school<-recode(school,"lo:1='law';2:3='business';4='trade';c(5,8)='finance';10:hi='art';else=NA",  
 as.factor.result=TRUE,  
 levels=c("law","business","finance","trade","art"))  
  
f2\_school<-recode(school,"lo:1='law';2:3='business';4='trade';c(5,8)='finance';10:hi='art';else=NA",  
 as.factor.result=TRUE)  
f1\_school

## [1] business law law business law law law   
## [8] business business finance trade art <NA> finance   
## [15] finance   
## Levels: law business finance trade art

f2\_school

## [1] business law law business law law law   
## [8] business business finance trade art <NA> finance   
## [15] finance   
## Levels: art business finance law trade

连续数值变量编码为factor变量，并加上标签

#模拟的数据distance是路段到A点的距离，gas\_stations是路段上加油站数  
#要做的事是先根据距离把distance划分为近、中、远，然后分别汇总统计加油站数  
  
distance<-rnorm(20)\*100  
gas\_stations<-sample(20)  
df<-data\_frame(distance,gas\_stations)  
summary(distance)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -232.0 -3.7 34.4 54.2 158.0 271.0

df$abs\_dis<-abs(df$distance)  
df$dist\_class<-cut(df$abs\_dis,breaks=c(0,20,50,100,1000),  
 right=TRUE,  
 ordered\_result=TRUE)  
df$dist\_class

## [1] (20,50] (100,1e+03] (0,20] (100,1e+03] (0,20]   
## [6] (0,20] (100,1e+03] (50,100] (100,1e+03] (100,1e+03]  
## [11] (50,100] (100,1e+03] (0,20] (50,100] (20,50]   
## [16] (100,1e+03] (20,50] (100,1e+03] (100,1e+03] (0,20]   
## Levels: (0,20] < (20,50] < (50,100] < (100,1e+03]

aggregate(df$gas\_stations,list(df$dist\_class),sum)

## Group.1 x  
## 1 (0,20] 37  
## 2 (20,50] 28  
## 3 (50,100] 40  
## 4 (100,1e+03] 105

数据缩尾处理Winsorize，去除数据里的异常值，一般默认去掉1%双侧 这个有些包里有此函数，但是专门调用个包不值得，自己参考网络写了个

#winsor函数定义  
winsor<-function(x,p=0.01) {  
 if(length(p) != 1 || p < 0 || p > 0.5) {  
 stop("bad p-value for winsorization")  
 }  
 lim <- quantile(x, probs=c(p, 1-p))  
 x[ x < lim[1] ] <- lim[1]  
 x[ x > lim[2] ] <- lim[2]  
 return(x)  
}  
#用AER包里地CPS1985演示对wage，age，education进行0.05缩尾  
data(CPS1985)  
summary(CPS1985$wage)

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

n\_cps<-as.data.frame(sapply(select(CPS1985,wage,age,education),winsor,p=0.05))  
summary(n\_cps$wage)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.50 5.25 7.78 8.86 11.20 20.00