实 验 报 告

学 院：

课程名称：

专 业：

姓 名：

学 号：

1. 实验目的

掌握数据新变量的生成；

掌握k-means聚类，变量压缩（WOE）；

掌握函数的编写。

1. 实验内容
   1. 使用k-means聚类对数据集“各省人口基本信息.csv”中的变量“15岁以上人口受教育年限”进行离散化，k取4；并在原数据中生成一新变量“year”用于表示4种不同类别（年限短、年限一般、年限较长、年限长）。
   2. 首先根据变量“地区总人口”生成一列响应变量“pop”，若人口数量大于等于8千万，“pop”取1，否则取0；然后根据“pop”和“year”，使用WOE的方法把变量“year”连续化，生成新变量“WOE\_year”。
   3. 讲上述两步的过程进行函数封装，要求函数最终的返回结果为数据框，包括的变量有：地区、pop、15岁以上人口受教育年限、year、WOE\_year
2. 实验要求

完成实验内容中的每一步，请附上代码与每一步数据结果的前15行。

# 代码

library(tidyverse)

#3.2.1

na<-which(is.na(total\_info$X15岁以上人口受教育年限))

kc<-kmeans(na.omit(total\_info$X15岁以上人口受教育年限),4)

cluster<-(kc$cluster)

cluster<-append(cluster,'',after = na-1)

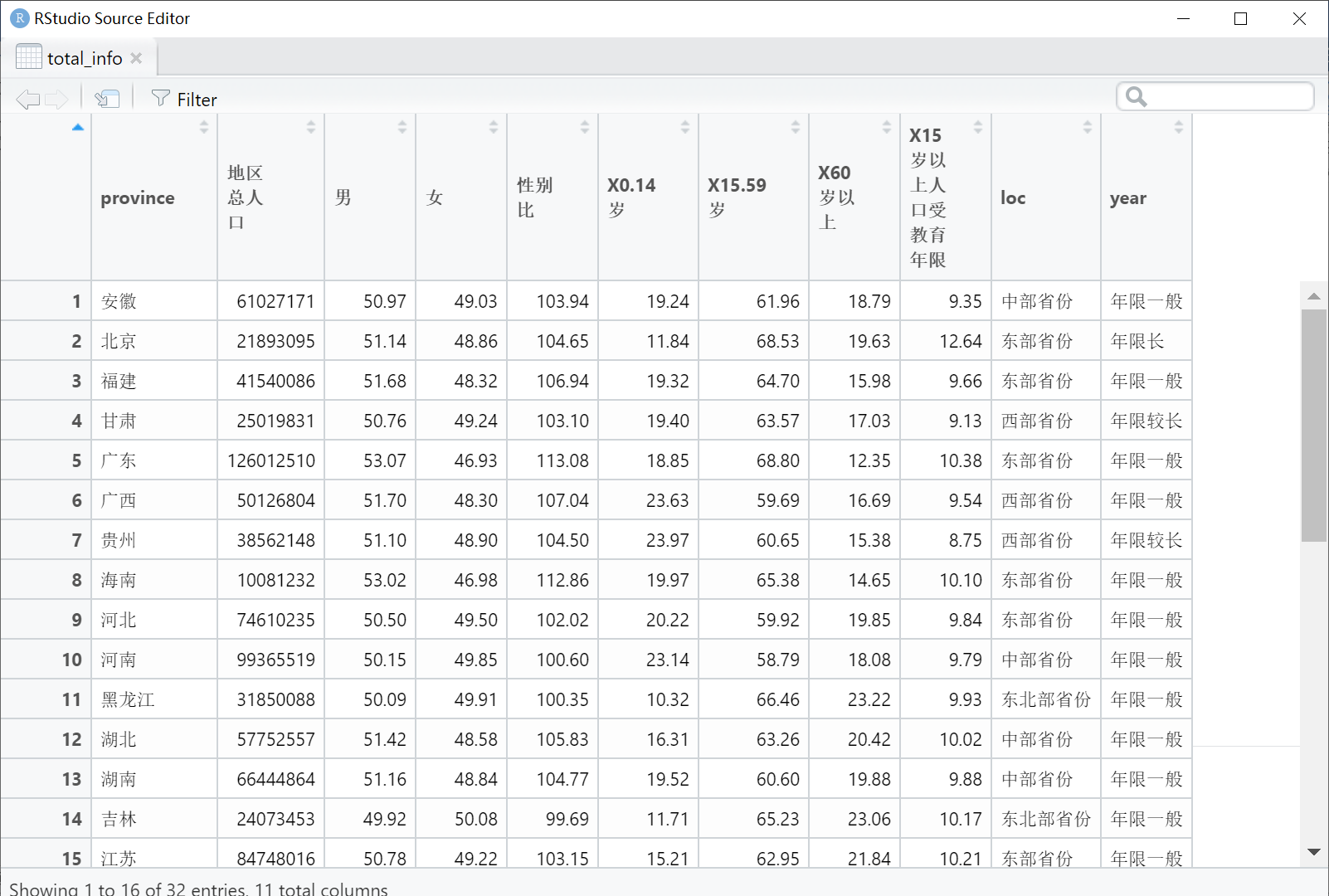
total\_info$year<-cluster

total\_info$year[which(total\_info$year==1)]<-'年限短'

total\_info$year[which(total\_info$year==2)]<-'年限一般'

total\_info$year[which(total\_info$year==3)]<-'年限较长'

total\_info$year[which(total\_info$year==4)]<-'年限长'



#3.2.2

total\_info$pop[total\_info$地区总人口>=8e7]=1

total\_info$pop[total\_info$地区总人口<8e7]=0

noarmy<-total\_info[-28,]

woe\_f<-function(df,flag\_loc,var\_loc){

n\_1t<-table(df[,flag\_loc])[2]

n\_0t<-table(df[,flag\_loc])[1]

varname<-names(df)[var\_loc]

fact\_n<-as.vector(unique(df[,var\_loc]))

fact<-c()

woe<-c()

iv<-c()

for (i in fact\_n){

subsetdf<-paste("subset(df,",varname,"=='",i,"')",sep="")

df\_i<-eval(parse(text=subsetdf))

n\_1i<-table(df\_i[,flag\_loc])[2]

n\_0i<-table(df\_i[,flag\_loc])[1]

p\_1i<-n\_1i/n\_1t

p\_0i<-n\_0i/n\_0t

woe\_i<-log(p\_1i/p\_0i)

iv\_i<-(p\_1i-p\_0i)\*woe\_i

fact<-c(fact,i)

woe<-c(woe,woe\_i)

iv<-c(iv,iv\_i)

}

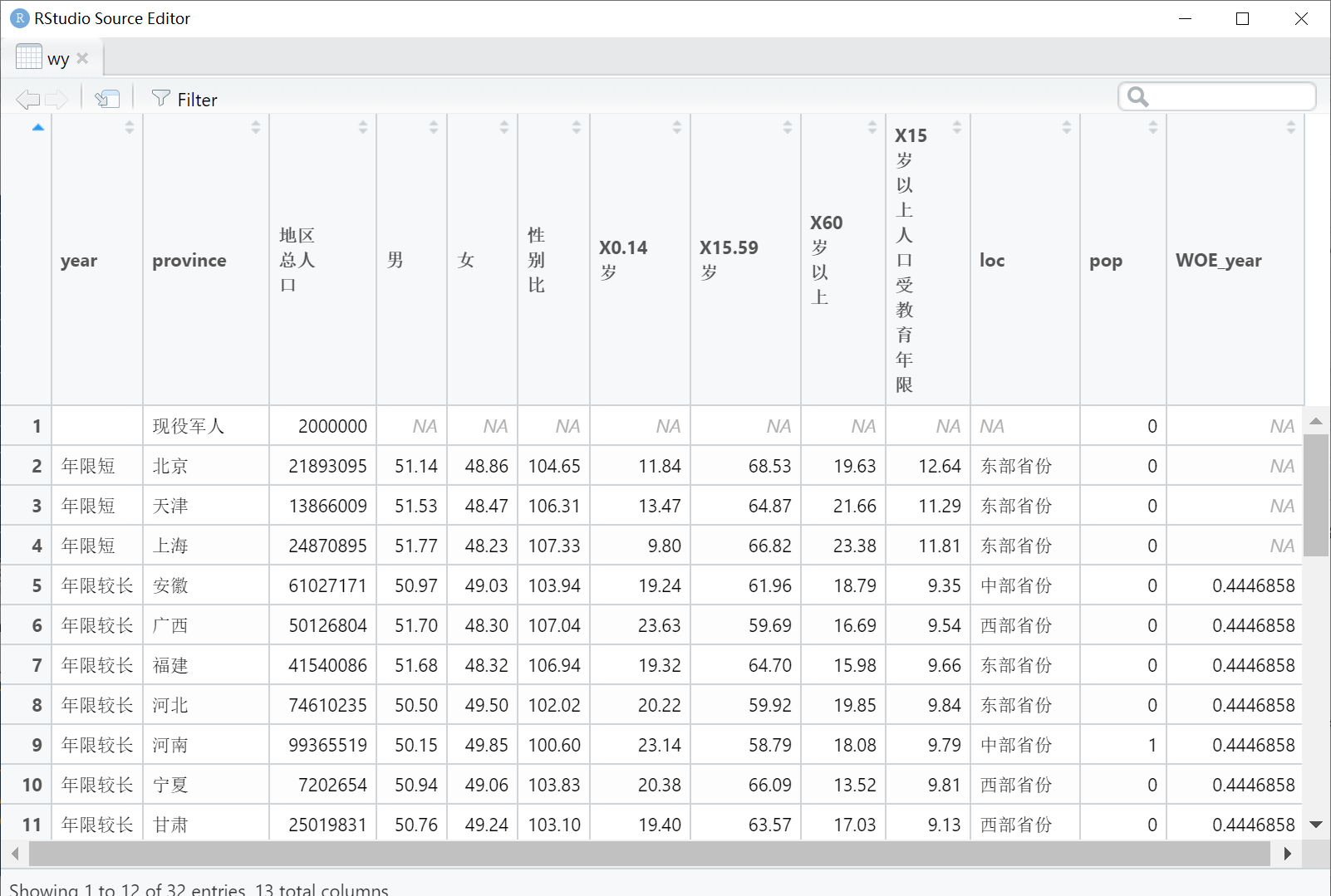
woeres<-paste("woeres<-data.frame(",varname,"=fact,woe\_",varname,"=woe)",sep='')

eval(parse(text=woeres))

ivres<-sum(iv)

return(list(woe=woeres,iv=ivres))

}



WOE\_year<-woe\_f(noarmy,which(names(noarmy)=='pop'),which(names(noarmy)=='year'))

#3.2.3

wy<-function(){

na<-which(is.na(total\_info$X15岁以上人口受教育年限))

kc<-kmeans(na.omit(total\_info$X15岁以上人口受教育年限),4)

cluster<-(kc$cluster)

cluster<-append(cluster,'',after = na-1)

total\_info$year<-cluster

total\_info$year[which(total\_info$year==1)]<-'年限短'

total\_info$year[which(total\_info$year==2)]<-'年限一般'

total\_info$year[which(total\_info$year==3)]<-'年限较长'

total\_info$year[which(total\_info$year==4)]<-'年限长'

total\_info$pop[total\_info$地区总人口>=8e7]=1

total\_info$pop[total\_info$地区总人口<8e7]=0

noarmy<-total\_info[-28,]

woe\_f<-function(df,flag\_loc,var\_loc){

n\_1t<-table(df[,flag\_loc])[2]

n\_0t<-table(df[,flag\_loc])[1]

varname<-names(df)[var\_loc]

fact\_n<-as.vector(unique(df[,var\_loc]))

fact<-c()

woe<-c()

iv<-c()

for (i in fact\_n){

subsetdf<-paste("subset(df,",varname,"=='",i,"')",sep="")

df\_i<-eval(parse(text=subsetdf))

n\_1i<-table(df\_i[,flag\_loc])[2]

n\_0i<-table(df\_i[,flag\_loc])[1]

p\_1i<-n\_1i/n\_1t

p\_0i<-n\_0i/n\_0t

woe\_i<-log(p\_1i/p\_0i)

iv\_i<-(p\_1i-p\_0i)\*woe\_i

fact<-c(fact,i)

woe<-c(woe,woe\_i)

iv<-c(iv,iv\_i)

}

woeres<-paste("woeres<-data.frame(",varname,"=fact,woe\_",varname,"=woe)",sep='')

eval(parse(text=woeres))

ivres<-sum(iv)

return(list(woe=woeres,iv=ivres))

}

a<-woe\_f(noarmy,which(names(noarmy)=='pop'),which(names(noarmy)=='year'))

a<-data.frame(a)

names(a)[1:3]<-c('year','WOE\_year','iv')

return(merge(total\_info,a[,1:2],by='year',all.x=T))

}

wy<-wy()

