Subject: Inquiry regarding data merging with DHS 2018 using R (2)
Posted by woojae1995 on Thu, 09 Feb 2023 03:49:12 GMT
View Forum Message <> Reply to Message

Following to my other question, I have another issue with data merging using the Nigeria DHS 2018 household data (HR) and household members data (PR)

HRdata <- read_dta('NGHR7BFL.DTA')
PRdata <- read_dta('NGPR7BFL.DTA')

I have done the following :

1.
HRtemp <- HRdata %>%
  select(hv001, hv002, starts_with("hml10"))
HRmerge <- merge(HRtemp, PRdata, by = c("hv001","hv002"))
HRmerge <- filter(HRmerge, hc1>=6 & hc1<=59)

This gives me 11590 observations for HRmerge

2.
Because I had to restrict the HR data to children of 6-59 months before I did the above merging as the following; I have tried this:

PRtemp =subset(PRdata, select=c(hv001, hv002, hc1), 'NA'= TRUE)
HRdata <- merge(HRdata,PRtemp,by=c("hv001", "hv002"))
HRdata <- filter(HRdata, hc1 >=6, hc1 <=59)
rm(PRtemp)

This will give me 11590 observations for HRdata

but then after I do the following merging to analyze another indicator as the folloinwg

HRmerge <- merge(HRtemp, PRdata, by = c("hv001","hv002"))
HRmerge <- filter(HRmerge, hc1>=6 & hc1<=59)

Now I have 21802 observations for HRmerge.

From my understanding whether I merge with a data that I have filtered already, or I filter it after, I don't think it should be different.

How come when I merge differently, I get different numbers? Is it because I merged with the PR data twice and somehow there were duplicates included? How can I manage it then?