Data cleaning

mahdi Babaloo

2022-10-04

{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)

Data Cleaning for year 98

First, we empty the entire memory. Then we upload the required packages and libraries.

“‘{r message=FALSE, warning=FALSE, paged.print=FALSE} #remove all rm(list=ls())

#Package_installation # install.packages(“dplyr”) library(“dplyr”) # install.packages(“tidyverse”)
library(“tidyverse”) # install.packages(“stargazer”)
library(“stargazer”) # install.packages(“haven”)
library(haven)

```
##  Data loading
This section we load  r data.
```

```
```{r , echo=TRUE}
#data_load
load("../data/98.RData")
```

Clear not necessary data

```
{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} rm("U98P3S01","U98P3S02","U98P3S04","U98
"U98P3S09","U98P3S10","U98P3S11","U98P3S14"    ,"U98P4S04","R98P1","R98P2","R98P3S01","R98P3S02",
"R98P3S04","R98P3S05","R98P3S06","R98P3S07","R98P3S08","R98P3S03",    "R98P3S09","R98P3S10","R98P3S11",
"R98P4S02","R98P4S03","R98P4S04","R98Data","U98P2") ## Data cleaning ### Division of provinces
```

“‘{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} Province <- c(Markazi=“00”, Gilan=“01”, Mazandaran=“02”, AzarbaijanSharghi=“03”, AzarbaijanGharbi=“04”, Kermanshah=“05”, Kouzestan=“06”, Fars=“07”, Kerman=“08”, KhorasanRazavi=“09”, Esfahan=“10”, SistanBalouchestan=“11”, Kordestan=“12”, Hamedan=“13”, CharmahalBakhtiari=“14”,Lorestan=“15”, Ilam=“16”, KohkilouyeBoyerahamad=“17”, Boushehr=“18”, Zanjan=“19”, Semnan=“20”, Yazd=“21”, Hormozgan=“22”, Tehran=“23”, Ardebil=“24”, Qom=“25”, Qazvin=“26”, Golestan=“27”, KhorasanShomali=“28”, KhorasanJonoubi=“29”, Alborz=“30”)

```
### Renaming characters
```{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE}
relation <- c(Head="1", Spouse="2", Child="3", SonDaughter_inLaw="4", GrandSonDaughter="5", Parent="6",
gender <- c(Male="1", Female="2")
literacy <- c(literate="1", illiterate="2")
yesno <- c(Yes="1", No="2")
education <- c(Elemantry="1", Secondary="2", HighSchool="3", Diploma="4", College="5", Bachelor="6", Mas
```

```
occupation <- c(employed="1", unemployed="2", IncomeWOJob="3", Student="4", Housewife="5", Other="6")
marital <- c(Married ="1", Widowed="2", Divorced="3", Single="4")
```

Census time

```
{r echo=FALSE, message=TRUE, warning=TRUE, paged.print=TRUE} U98Data <- U98Data %>%
rename(month = MahMorajeh) %>%  mutate(province = fct_recode(as.factor(substr(Address,
2, 3)), !!!Province),          town = as.integer(substr(Address, 4, 5)))%>%  select(Address,month,provi
### cleaning & rename {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} U98P1
<- U98P1 %>%   rename(     member = DYCOL01,  relation = DYCOL03,      gender = DYCOL04,
age = DYCOL05,     literacy = DYCOL06,      studying = DYCOL07,     degree = DYCOL08,
occupationalst = DYCOL09,maritalst = DYCOL10)%>%   mutate(across(where(is.character),
as.integer),          across(c(relation,gender,literacy,studying,degree,occupationalst,maritalst),
as.factor),          relation = fct_recode(relation, !!!relation),          gender =
fct_recode(gender, !!!gender),          literacy = fct_recode(literacy, !!!literacy),
studying = fct_recode(studying, !!!yesno),          degree = fct_recode(degree, !!!education),
occupationalst = fct_recode(occupationalst, !!!occupation),          maritalst = fct_recode(maritalst,
!!!marital))  ##  household_without_childeren  ###  part1  {r, echo=TRUE, message=FALSE,
warning=FALSE, paged.print=FALSE} U98P1 <- U98P1 %>%  mutate(Just_Married1 = case_when(
relation == "Child" ~ 0 ,          relation == "SonDaughter_inLaw" ~ 0,     relation
== "GrandSonDaughter" ~ 0, relation == "Parent" ~ 0 ,     relation == "Sibling" ~ 0,
relation == "OtherRelative" ~ 0 ,     relation == "NonRelative" ~ 0 ,     relation ==
"Head" ~ 1,     relation == "Spouse" ~ 1))%>%  select(Address,member,relation,Just_Married1,everything
### part2 "'{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE}
```

U98P1 <- U98P1%>% mutate(Just_Married2 = case_when(relation == "Child" ~ 1 , relation == "SonDaughter_inLaw" ~ 1, relation == "GrandSonDaughter" ~ 1,relation == "Parent" ~ 1 , relation == "Sibling" ~ 1, relation == "OtherRelative" ~ 1 , relation == "NonRelative" ~ 1 , relation == "Head" ~ 0, relation == "Spouse" ~ 0))%>% select(Address,member,relation,Just_Married2,everything())

```
### part3
```{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE}
U98P1 <- U98P1%>%
  mutate_at(vars(Address),as.character)%>%
  group_by(Address)%>%
  dplyr::mutate(Indicator1 = sum(Just_Married1,na.rm = T),
                Indicator2 = sum(Just_Married2,na.rm = T))%>%
  select(Address,member,relation,Indicator1,Indicator2,everything())
```

Final part

```
{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} U98P1<- U98P1%>%  filter(Indicator1
== 2 & Indicator2 == 0)
```

Income

define income_wage_eaerner

Change data type

```
{r, warning=FALSE} class(U98P4S01$Address) = "double"  First  Table  {r, warning=FALSE}
income_wage_earner <- U98P4S01%>%  rename(     member     =DYCOL01,     income     =DYCOL15
)%>%  select(Address,member,income)%>%  mutate(income = replace_na(income,0))
```
```

secound table

selfe_mployed: people who selfemployed {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} income_self_employed <-U98P4S02 %>% rename(member =DYCOL01, income =DYCOL15)%>% select(Address,member,income)%>% mutate(income = replace_na(income,0))

change double forcharacter

{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} income_wage_earner <-income_wage_earner %>% mutate(income =as.integer(income))

third table

other income {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} other<-U98P4S03 %>% rename(member =DYCOL01, income =DYCOL03,)%>% select(Address,member,income)% mutate(income = replace_na(income,0)) ### create a column by bindind columns of income_wage_earner and income_self_employed and other income {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} income_table<-bind_rows(income_wage_earner,income_self_employed,other) ### deleting income_self_emolyed and income_wage_earner and other {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} rm(income_wage_earner,income_self_employed,other) ### Change Data type {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} income_table <- income_table %>% mutate(income =as.integer(income)) ## calculate total income {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} income_table <- income_table%>% group_by(Address,member)%>% summarise(total_income = sum(income)) ### change data type {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} class(U98P1$Address) = "double" ### merging data {r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} Data<-left_join(x=U98P1, y=income_table, by=c("Address","member"))

clothes expendutre

third table

"'{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} U98P3S03<-U98P3S03%>% rename(code = DYCOL01, purchased = DYCOL02, value = DYCOL03)

U98P3S12<-U98P3S12%>% rename(code = DYCOL01, purchased = DYCOL02, value = DYCOL03) U98P3S13<-U98P3S13%>% rename(code = DYCOL01, purchased = DYCOL02, value = DYCOL03)

women clothes expendture & shoe

```{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE}
#subgroup1
ag1sp_1_1<-filter(U98P3S03,code==31232|code==31233|code==31234|code==31235|code==31415)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_1= sum(value))
#subgroup2
ag1sp_1_2<-filter(U98P3S03,code==31236|code==31237|code==31238|code==31239|code==31242)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_2= sum(value))
#subgroup3
ag1sp_1_3<-filter(U98P3S03,code==31415|code==31112|code==31244|code==31113|code==31114)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_3= sum(value))
#subgroup4
ag1sp_1_4<-filter(U98P3S03,code==31116|code==31117|code==31118|code==31119|code==31312)%>%
  group_by(Address) %>%
```

```
  summarise(ag1sp_1_4= sum(value))
#subgroup5
ag1sp_1_5<-filter(U98P3S03,code==31318|code==31319|code==31323|code==31316)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_5= sum(value))
#subgroup6
ag1sp_1_6 <- filter(U98P3S03,code==32121|code==32122)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_6= sum(value))
#subgroup7
ag1sp_1_7<- filter(U98P3S03,code==32123|code==32124|code==31211)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_7= sum(value))
#subgroup8
ag1sp_1_8<- filter(U98P3S12,code==121136|code==121316|code==121111|
                     code==121114|code==121112|code==121115|code==121336|
                     code==121353)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_8= sum(value))
#subgroup9
ag1sp_1_9<- filter(U98P3S12,
                   code==121316|code==123214|code==121341|code==123214)%>%
  group_by(Address) %>%
  summarise(ag1sp_1_9= sum(value))
```

men clothes expendture_subgroups

```
{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} #subgroup1 ag2sp_1_1 <-
filter(U98P3S03,code==31211|code==31212|code==31213|code==31216|code==31414)%>%   group_by(Address)
%>%    summarise(ag2sp_1_1= sum(value)) #subgroup2 ag2sp_1_2 <- filter(U98P3S03,code==31215|code==31214
group_by(Address) %>%    summarise(ag2sp_1_2= sum(value)) #subgroup3 ag2sp_1_3 <- filter(U98P3S03,code==
group_by(Address) %>%    summarise(ag2sp_1_3= sum(value)) #subgroup4 ag2sp_1_4 <- filter(U98P3S03,code==
group_by(Address) %>%    summarise(ag2sp_1_4= sum(value)) #subgroup5 ag2sp_1_5<- filter(U98P3S12,
code==121113|code==123216|code==123227)%>%   group_by(Address) %>%    summarise(ag2sp_1_5=
sum(value)) #subgroup6 ag2sp_1_6<- filter(U98P3S13,                    code==31415)%>%
group_by(Address)
```

merging_data

```
{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} ##### install.packages("plyr")
library(plyr) #merging_data data<-join_all(list(ag1sp_1_1,ag1sp_1_2,ag1sp_1_3,ag1sp_1_4,ag1sp_1_5,
ag1sp_1_6,ag1sp_1_7,ag1sp_1_8,ag1sp_1_9,ag2sp_1_1,                    ag2sp_1_2,ag2sp_1_3,ag2sp_1_4,ag
by='Address', type='left') #CLEARING_DATA remove(ag1sp_1_1,ag1sp_1_2,ag1sp_1_3,ag1sp_1_4,ag1sp_1_5,
ag1sp_1_6,ag1sp_1_7,ag1sp_1_8,ag2sp_1_1,ag1sp_1_9,ag2sp_1_2,ag2sp_1_3,        ag2sp_1_4,ag2sp_1_5,ag2sp
```

Combinig Data and save clean data The combing data set with expenditure and income family and add month in data.

“‘{r, echo=TRUE, message=FALSE, warning=FALSE, paged.print=FALSE} class(U98P1$Address) = “double” Data<-left_join(x=Data, y=data, by=c(“Address”)) Data<-left_join(x=Data, y=U98Data, by=c(“Address”)) # saving data in stata file

#write_dta(Data,“98.dta”) “‘ The same work on 1399 data.