Data cleaning

mahdi Babaloo

2022-10-04

Data Cleaning for year 98

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

```
#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)
#install.packages("remotes")
library(remotes)
#install_github("hemken/Statamarkdown")
library(Statamarkdown)
```

Data loading

This section we load r data.

```
#data_load
load("../data/98.RData")
```

Clear not necessary data

```
rm("U98P3S01","U98P3S02","U98P3S04","U98P3S05","U98P3S06","U98P3S07","U98P3S08",
   "U98P3S09","U98P3S10","U98P3S11","U98P3S14"
   ,"U98P4S04","R98P1","R98P2","R98P3S01","R98P3S02",
   "R98P3S04","R98P3S05","R98P3S06","R98P3S07","R98P3S08","R98P3S03",
   "R98P3S09","R98P3S10","R98P3S11","R98P3S12","R98P3S14","R98P4S01",
   "R98P4S02","R98P4S03","R98P4S04","R98Data","U98P2")
```

Data cleaning

Division of provinces

```
Province <- c(Markazi="00", Gilan="01", Mazandaran="02", AzarbaijanSharghi="03", AzarbaijanGharbi="04",
Esfahan="10", SistanBalouchestan="11", Kordestan="12", Hamedan="13", CharmahalBakhtiari="14",Lorestan="
```

```
Golestan="27", KhorasanShomali="28", KhorasanJonoubi="29", Alborz="30")
```

Renaming characters

```
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", Ma
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

cleaning & rename

```
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

```
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

```
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

```
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

```
U98P1<- U98P1%>%
  filter(Indicator1 == 2 & Indicator2 == 0)
```

Income

define income_wage_eaerner

Change data type

```
class(U98P4S01$Address) = "double"
```

First Table

```
income_wage_earner <- U98P4S01%>%
  rename(
    member     =DYCOL01,
    income     =DYCOL15
  )%>%
  select(Address,member,income)%>%
  mutate(income = replace_na(income,0))
```

secound table

```
income_self_employed <-U98P4S02 %>%
  rename(
    member     =DYCOL01,
    income     =DYCOL15 )%>%
  select(Address,member,income)%>%
  mutate(income = replace_na(income,0))
```

selfe_mployed: people who selfemployed

change double forcharacter

```
income_wage_earner <-income_wage_earner %>% mutate(income =as.integer(income))
```

third table

```
other<-U98P4S03 %>%
  rename(
    member     =DYCOL01,
    income     =DYCOL03,
  )%>%
  select(Address,member,income)%>%
  mutate(income = replace_na(income,0))
```

other income

create a column by bindind columns of income__wage__earner and income__self__employed and other income

```
income_table<-bind_rows(income_wage_earner,income_self_employed,other )
```

deleting income__self__emolyed and income__wage__earner and other

```
rm(income_wage_earner,income_self_employed,other)
```

Change Data type

```
income_table <- income_table %>% mutate(income =as.integer(income))
```

calculate total income

```
income_table <- income_table%>%
  group_by(Address,member)%>%
  summarise(total_income = sum(income))
```

change data type

```
class(U98P1$Address) = "double"
```

merging data

```
Data<-left_join(
  x=U98P1,
  y=income_table,
  by=c("Address","member")
)
```

clothes expendutre

third table

```
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

```
#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

```
#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|code==31217|code==31218|code==31219)%>%
  group_by(Address) %>%
  summarise(ag2sp_1_2= sum(value))
#subgroup3
ag2sp_1_3 <- filter(U98P3S03,code==31221|code==31213|code==31313|code==31111|code==31115)%>%
  group_by(Address) %>%
  summarise(ag2sp_1_3= sum(value))
#subgroup4
ag2sp_1_4 <- filter(U98P3S03,code==32111|code==32112|code==32113|code==32114)%>%
  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

```
##### 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,ag2sp_1_5,ag2sp_1_6), 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_1_6)
```

Combinig Data and save clean data

The combing data set with expenditure and income family and add month in data.

```
 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.

Clean R data

```
#remove all
rm(list=ls())
```

Read StATA data

```
use"C:\Users\Mahdi Babaloo\Dropbox\theis\R_markdown\Data\98.dta"
su
```

```
Error in running comma
```