数据清洗

吴沛豪

目录

```
# set chunk options
knitr::opts_chunk$set(echo = TRUE,
                      fig.align = "center",
                      message = FALSE,
                      warning = FALSE)
rm(list = ls())
cat("\014") # Clear Workspace and Console
```

```
library(tidyverse)
library(survey)
library(foreign)
library(nhanesA)
library(VIM)
```

```
path_new <- paste0(getwd(),'/data_clean/')
```

```
p.na <- function(x){
  sum(is.na(x)/length(x))
}

f.detec <- function(data){
  data = data
  miss_col <- which(colSums(is.na(data))>0)
  miss_data <- data[,miss_col]
  aggr(miss_data,labels=names(miss_data),cex.axis=0.7)
}
```

DEMO 数据

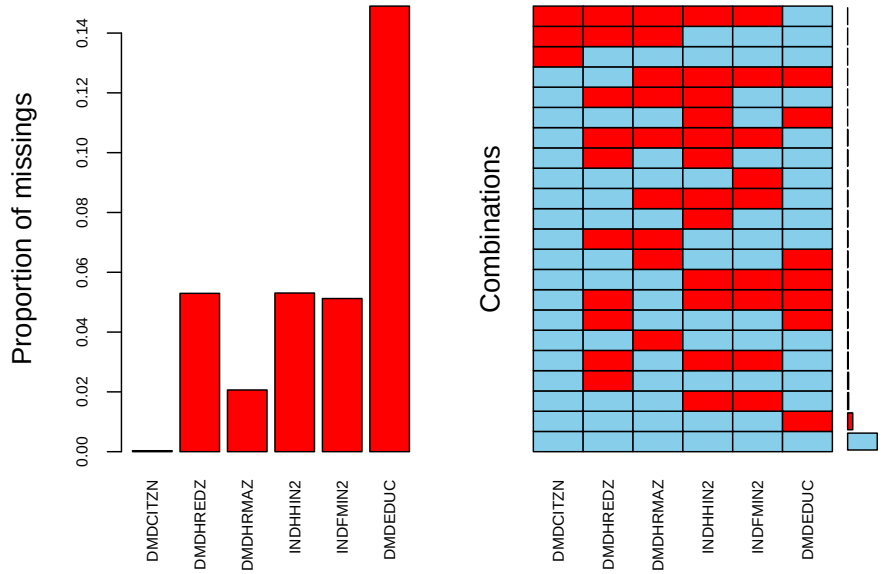
```
demo_data <- read.csv('demo_data.csv')[-1]
demo_vars <- names(demo_data)
demo_data <- nhanesTranslate('DEMO_J',demo_vars,data = demo_data)
```

```
demo_data_clean <- demo_data %>%
  dplyr::select(1,4,5,6,8,13,14,16,17,30,31,32,33,34,35,36,37,38,44,45) %>%
  mutate(AGE = if_else(!is.na(RIDAGEMN),round(RIDAGEMN/12,2),RIDAGEYR),
         DMDEDUC = if_else(!is.na(DMDEDUC3),DMDEDUC3,DMDEDUC2)) %>%
  dplyr::select(-c(RIDAGEMN,RIDAGEYR,DMDEDUC3,DMDEDUC2))

remove(demo_data)
remove(demo_vars)

write.csv(demo_data_clean,paste0(path_new,'demo_data_clean.csv'))
```

```
f.detec(demo_data_clean)
```



DIETARY 数据

```
DR1TOT <- read.csv('DR1TOT_J.csv')[-1]
DR1TOT2 <- DR1TOT %>%
  dplyr::select(1,2,3,13,15,18,32:101,125)
DR1TOT2 <- DR1TOT2 %>%
  na.omit()
DR1TOT_vars <- names(DR1TOT2[,c(4:6,77)])
DR1TOT2 <- nhanesTranslate('DR1TOT_J',DR1TOT_vars,data = DR1TOT2)
DR1TOT_clean <- DR1TOT2 %>%
  mutate(WTDR2D = if_else(WTDR2D==0,WTDRD1,WTDR2D))

remove(DR1TOT)
remove(DR1TOT2)
remove(DR1TOT_vars)

write.csv(DR1TOT_clean,paste0(path_new,'DR1TOT_clean','.csv'))
```

EXAMINATION 数据

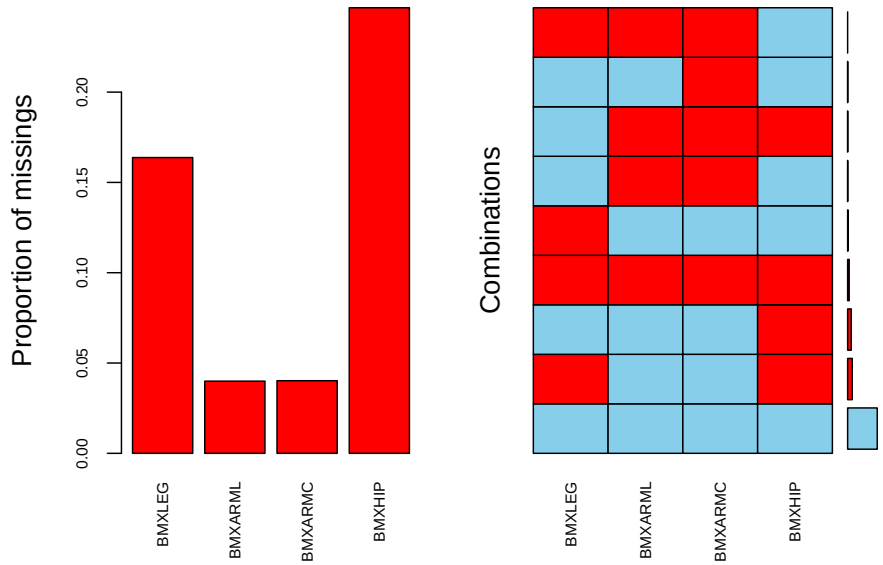
Body Measures

```
BMX <- read.csv('BMX_J.csv')[-1]
BMX_vars <- names(BMX)
BMX_clean <- BMX %>%
  dplyr::select(1,3,9,11,12,14,16,20) %>%
  dplyr::filter(!is.na(BMXBMI))

remove(BMX)
remove(BMX_vars)

write.csv(BMX_clean,paste0(path_new,'BMX_clean','.csv'))
```

```
f.detec(BMX_clean)
```



Blood Pressure

```
BPX <- read.csv('BPX_J.csv')[-1]
BPX_vars <- names(BPX)
BPX_clean <- BPX %>%
  dplyr::select(1,6,7,8,9,10,11,13,14,16,17,19,20) %>%
  dplyr::filter(!is.na(BPXPLS))
BPX_vars <- names(BPX_clean[,c(3,4)])
BPX_clean <- nhanesTranslate('BPX_J',BPX_vars,data = BPX_clean)
BPX_clean <- BPX_clean %>%
  mutate(BPXSY1 = if_else(BPXSY1 == 0,NA,BPXSY1),
         BPXSY2 = if_else(BPXSY2 == 0,NA,BPXSY2),
         BPXSY3 = if_else(BPXSY3 == 0,NA,BPXSY3),
         BPXSY4 = if_else(BPXSY4 == 0,NA,BPXSY4),
         BPXDI1 = if_else(BPXDI1 == 0,NA,BPXDI1),
         BPXDI2 = if_else(BPXDI2 == 0,NA,BPXDI2),
         BPXDI3 = if_else(BPXDI3 == 0,NA,BPXDI3),
         BPXDI4 = if_else(BPXDI4 == 0,NA,BPXDI4))
BPX_clean <- BPX_clean %>%
  mutate(BPXSY = apply(BPX_clean[,c(6,8,10,12)], 1, mean,na.rm = T) %>% round(2),
         BPXDI = apply(BPX_clean[,c(7,9,11,13)], 1, mean,na.rm = T) %>% round(2)) %>%
  dplyr::select(-c(5:13)) %>%
  dplyr::filter(BPXSY > 0 &  BPXDI > 0)

remove(BPX)
remove(BPX_vars)

write.csv(BPX_clean,paste0(path_new,'BPX_clean','.csv'))
```

Liver Ultrasound Transient Elastography

```
LUX <- read.csv('LUX_J.csv')[-1]
vars <- names(LUX)
LUX_clean <- LUX %>%
  dplyr::select(1,10,11,12,13,14) %>%
  na.omit()
```

```
remove(LUX)
remove(vars)

write.csv(LUX_clean,paste0(path_new,'LUX_clean','.csv'))
```

LABORATORY 数据

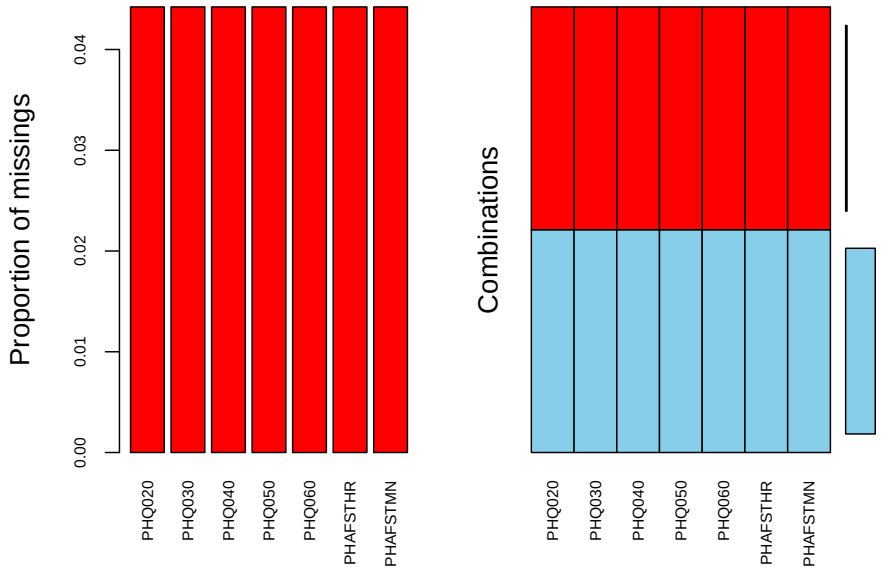
Fasting Questionnaire——用于校正

```
FASTQX <- read.csv('FASTQX_J.csv')[-1]
vars <- names(FASTQX)
FASTQX_clean <- FASTQX %>%
  dplyr::select(1,2,5,8,11,14,17,18,19)
vars <- names(FASTQX_clean)
FASTQX_clean <-  nhanesTranslate('FASTQX_J',vars,data = FASTQX_clean)

remove(FASTQX)
remove(vars)

write.csv(FASTQX_clean,paste0(path_new,'FASTQX_clean','.csv'))
```

```
f.detec(FASTQX_clean)
```



Cholesterol - Total

```
TCHOL <- read.csv('TCHOL_J.csv')[-1]
TCHOL_clean <- TCHOL %>%
  na.omit()

remove(TCHOL)

write.csv(TCHOL_clean,paste0(path_new,'TCHOL_clean','.csv'))
```

Cholesterol - High - Density Lipoprotein (HDL)

```
HDL <- read.csv('HDL_J.csv')[-1]
HDL_clean <- HDL %>%
  na.omit()

remove(HDL)

write.csv(HDL_clean,paste0(path_new,'HDL_clean','.csv'))
```

QUESTIONNAIRE 数据

Part A: outcomes

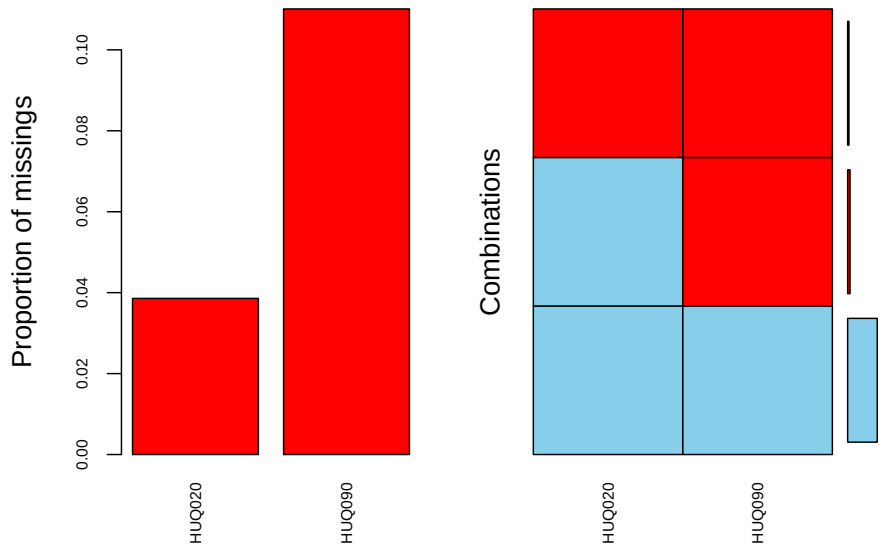
Hospital Utilization & Access to Care

```
HUQ <- read.csv('HUQ_J.csv')[-1]
vars <- names(HUQ)
HUQ_clean <- HUQ %>%
  dplyr::select(1,2,3,6,10)
vars <- names(HUQ_clean)
HUQ_clean <-  nhanesTranslate('HUQ_J',vars[2:5],data = HUQ_clean)

remove(HUQ)
remove(vars)

write.csv(HUQ_clean,paste0(path_new,'HUQ_clean','.csv'))
```

```
f.detec(HUQ_clean)
```



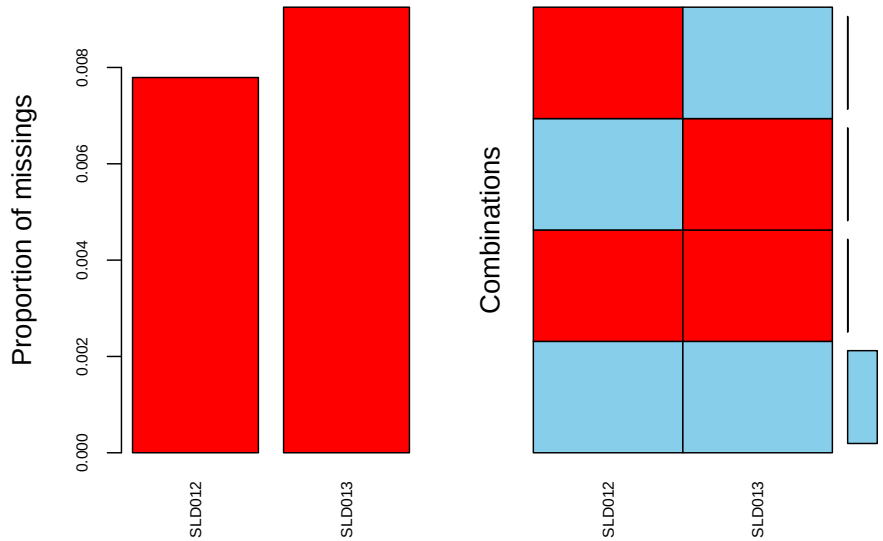
Sleep Disorders

```
SLQ <- read.csv('SLQ_J.csv')[-1]
vars <- names(SLQ)
SLQ_clean <- nhanesTranslate('SLQ_J',vars[2:length(vars)],data = SLQ)

remove(SLQ)
remove(vars)

write.csv(SLQ_clean,paste0(path_new,'SLQ_clean','.csv'))
```

```
f.detec(SLQ_clean)
```



Mental Health - Depression Screener

```
DPQ <- read.csv('DPQ_J.csv')[-1]
DPQ_clean <- DPQ %>%
  dplyr::select(1:10) %>%
  na.omit()
vars <- names(DPQ_clean)
DPQ_clean <- nhanesTranslate('DPQ_J',vars[2:length(vars)],data = DPQ_clean)
```

```
remove(DPQ)
remove(vars)

write.csv(DPQ_clean,paste0(path_new,'DPQ_clean','.csv'))
```

Part B: factors

Physical Activity

```
PAQ <- read.csv('PAQ_J.csv')[-1]
vars <- names(PAQ)
PAQ_clean <- PAQ %>%
  dplyr::select(1,2,5,8,11,14)
vars <- names(PAQ_clean)
PAQ_clean <- nhanesTranslate('PAQ_J',vars[2:length(vars)],data = PAQ_clean)

remove(PAQ)
remove(vars)

write.csv(PAQ_clean,paste0(path_new,'PAQ_clean','.csv'))
```

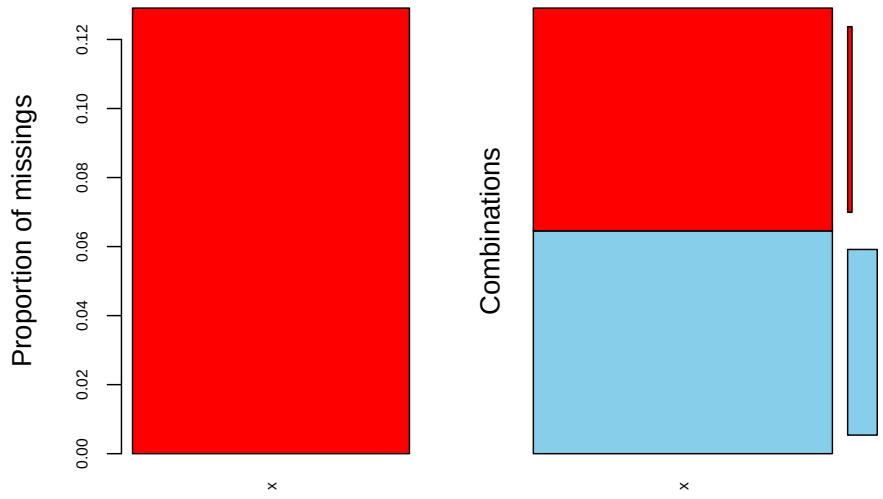
Smoking - Cigarette Use

```
SMQ <- read.csv('SMQ_J.csv')[-1]
vars <- names(SMQ)
SMQ_clean <- SMQ %>%
  dplyr::select(1,2)
vars <- names(SMQ_clean)
SMQ_clean <- nhanesTranslate('SMQ_J',vars[2],data = SMQ_clean)

remove(SMQ)
remove(vars)

write.csv(SMQ_clean,paste0(path_new,'SMQ_clean','.csv'))
```

```
f.detec(SMQ_clean)
```



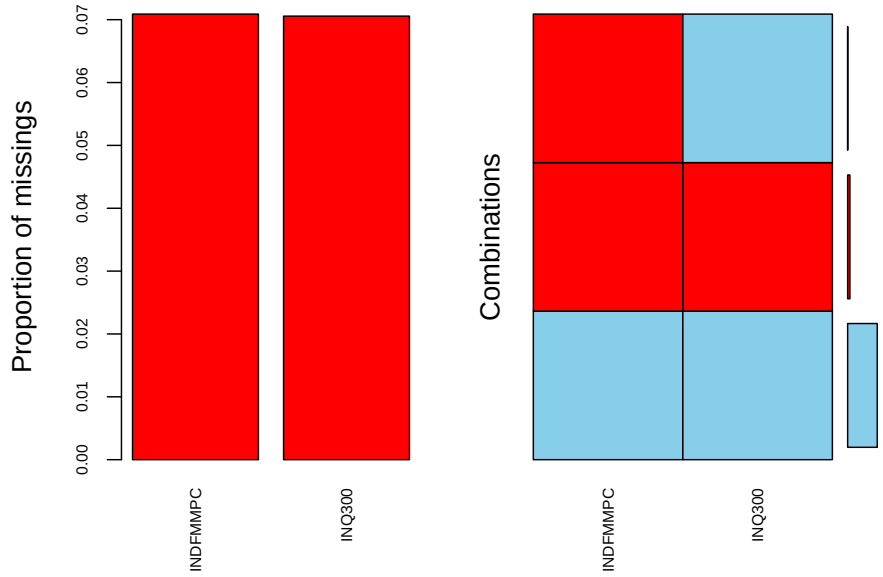
Income

```
INQ <- read.csv('INQ_J.csv')[-1]
vars <- names(INQ)
INQ_clean <- INQ %>%
  dplyr::select(1,13,14)
vars <- names(INQ_clean)
INQ_clean <- nhanesTranslate('INQ_J',vars[2:length(vars)],data = INQ_clean)


remove(INQ)
remove(vars)

write.csv(INQ_clean,paste0(path_new,'INQ_clean','.csv'))
```

```
f.detec(INQ_clean)
```



Occupation

```
OCQ <- read.csv('OCQ_J.csv')[-1]
vars <- names(OCQ)
OCQ_clean <- OCQ %>%
  dplyr::select(1,2,9)
vars <- names(OCQ_clean)
OCQ_clean <- nhanesTranslate('OCQ_J',vars[2:length(vars)],data = OCQ_clean)

remove(OCQ)
remove(vars)

write.csv(OCQ_clean,paste0(path_new,'OCQ_clean','.csv'))
```

Alcohol Use

```
ALQ <- read.csv('ALQ_J.csv')[-1]
vars <- names(ALQ)
ALQ_clean <- ALQ %>%
  dplyr::select(1,2,3)
vars <- names(ALQ_clean)
ALQ_clean <- nhanesTranslate('ALQ_J',vars[2:length(vars)],data = ALQ_clean)

remove(ALQ)
remove(vars)

write.csv(ALQ_clean,paste0(path_new,'ALQ_clean','.csv'))
```

```
f.detec(ALQ_clean)
```