Imputation(add new variables)

Xinyi Lin

9/29/2019

## Import data

Import data “R1\_R18\_partcleaned\_20190913.dta”.

## Creat the variable

Notation:

Variable name: student\_r

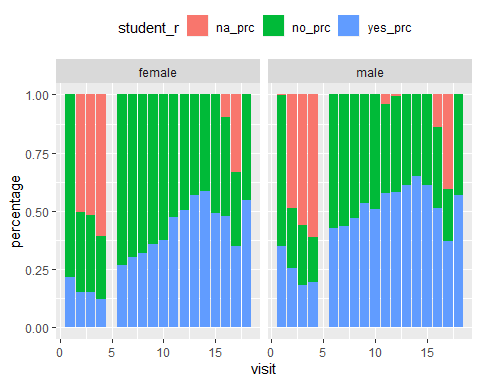
Created from: occup1\_r, occup2\_r

Label: student\_r — yes-is a student, no-not a student, NA-occup1\_r and occup2\_r are NAs

## Explore student\_r

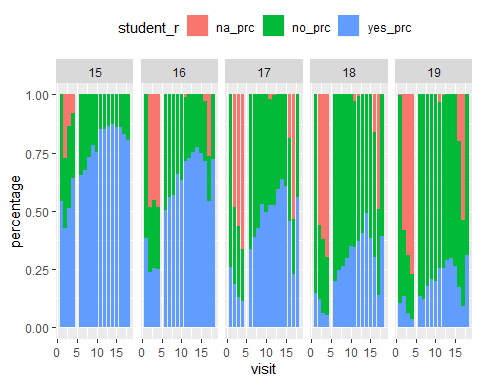
For ‘gender’

## Warning: Factor `student\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

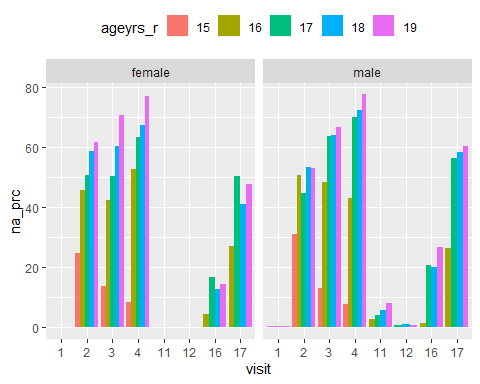


For ‘age’

## Warning: Factor `student\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`



## Warning: Factor `student\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| gender | visit | 15 | 16 | 17 | 18 | 19 |
| female | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| female | 2 | 24.77 | 45.86 | 50.57 | 58.63 | 61.80 |
| female | 3 | 13.88 | 42.38 | 50.51 | 60.41 | 70.69 |
| female | 4 | 8.57 | 52.80 | 63.36 | 67.27 | 76.92 |
| female | 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| female | 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| female | 16 | 0.00 | 4.32 | 16.61 | 12.71 | 14.49 |
| female | 17 | 0.00 | 26.92 | 50.34 | 41.18 | 47.75 |
| male | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.43 |
| male | 2 | 31.01 | 50.75 | 44.90 | 53.26 | 53.12 |
| male | 3 | 13.14 | 48.39 | 63.79 | 64.15 | 66.56 |
| male | 4 | 7.69 | 43.02 | 70.05 | 72.43 | 77.57 |
| male | 11 | 0.00 | 2.74 | 4.19 | 5.86 | 8.16 |
| male | 12 | 0.00 | 0.00 | 0.79 | 1.06 | 0.79 |
| male | 16 | 0.00 | 1.30 | 20.90 | 19.94 | 26.88 |
| male | 17 | 0.00 | 26.34 | 56.41 | 58.31 | 60.54 |

## Analysis missing data

### Select variables and tidy data

Predictors including: visit, ageyrs\_r, gender, area, educate\_r, currmarr\_r, sexp1yr, SEScat, study\_id, pregnow\_r, numchild\_r, eversex\_r, currrltn.

First, check each variable.

## student\_r visit ageyrs\_r gender   
## no :21816 18 : 3596 Min. :15.00 Length:47690   
## yes :19669 17 : 3407 1st Qu.:16.00 Class :character   
## NA's: 6205 16 : 3187 Median :17.00 Mode :character   
## 14 : 3126 Mean :17.17   
## 6 : 2969 3rd Qu.:18.00   
## 3 : 2939 Max. :19.00   
## (Other):28466   
## area educate\_r currmarr\_r sexp1yr   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. : 0.0   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 1.0   
## Median :0.0000 Median :1.0000 Median :0.0000 Median : 2.0   
## Mean :0.2956 Mean :0.9775 Mean :0.1953 Mean :42.7   
## 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:98.0   
## Max. :2.0000 Max. :1.0000 Max. :1.0000 Max. :99.0   
## NA's :8 NA's :4 NA's :1390   
## SEScat study\_id pregnow\_r numchild\_r   
## Min. :0.000 Length:47690 Min. :0.000 Min. :0.00   
## 1st Qu.:1.000 Class :character 1st Qu.:0.000 1st Qu.:0.00   
## Median :2.000 Mode :character Median :0.000 Median :1.00   
## Mean :1.645 Mean :0.126 Mean :0.78   
## 3rd Qu.:3.000 3rd Qu.:0.000 3rd Qu.:1.00   
## Max. :3.000 Max. :1.000 Max. :4.00   
## NA's :208 NA's :22584 NA's :43762   
## eversex\_r currrltn   
## Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:2.000   
## Median :1.0000 Median :2.000   
## Mean :0.5848 Mean :1.783   
## 3rd Qu.:1.0000 3rd Qu.:2.000   
## Max. :1.0000 Max. :9.000   
## NA's :347

For sexp1yr:

## 0 1 2 3 4 5 6 7 8 9 10 11   
## 3840 17944 3161 863 276 118 79 54 29 12 25 2   
## 12 13 15 16 17 18 20 21 22 24 30 93   
## 6 2 12 2 2 2 5 1 1 1 2 33   
## 97 98 99 NA's   
## 1 19826 1 1390

93 means a lot, 97 means don’t know, 98 means not applicable and 99 means not response. 97, 98, 99 can be marked as NA.

For pregnow\_r:

## # A tibble: 4 x 3  
## # Groups: gender [2]  
## gender pregnow\_r n  
## <chr> <dbl+lbl> <int>  
## 1 female 0 [no] 21945  
## 2 female 1 [yes] 3161  
## 3 female NA 661  
## 4 male NA 21923

For male, pregnow\_r can be marked as 98. And pregnow\_r should be catagorical variable.

For currrltn:

## 0 1 2 7 8 9   
## 54 10415 37193 12 5 11

7 means don’t know. *0, 8, 9 can be marked as missing value(not sure).*

As there are too much missing value in numchild\_r, and too much categories in study\_id it is not suitable to be a predictor.

## student\_r visit ageyrs\_r gender area   
## no :21816 18 : 3596 Min. :15.00 female:25767 0:36211   
## yes :19669 17 : 3407 1st Qu.:16.00 male :21923 1: 8861   
## NA's: 6205 16 : 3187 Median :17.00 2: 2618   
## 14 : 3126 Mean :17.17   
## 6 : 2969 3rd Qu.:18.00   
## 3 : 2939 Max. :19.00   
## (Other):28466   
## educate\_r currmarr\_r sexp1yr SEScat pregnow\_r   
## 0 : 1073 0 :38373 Min. : 0.000 0 : 9888 0 :21945   
## 1 :46609 1 : 9313 1st Qu.: 1.000 1 :10995 1 : 3161   
## NA's: 8 NA's: 4 Median : 1.000 2 :12703 98 :21923   
## Mean : 1.273 3 :13896 NA's: 661   
## 3rd Qu.: 1.000 NA's: 208   
## Max. :93.000   
## NA's :21218   
## eversex\_r currrltn   
## 0 :19657 1 :10415   
## 1 :27686 2 :37193   
## NA's: 347 7 : 12   
## NA's: 70   
##   
##   
##

### Get test data

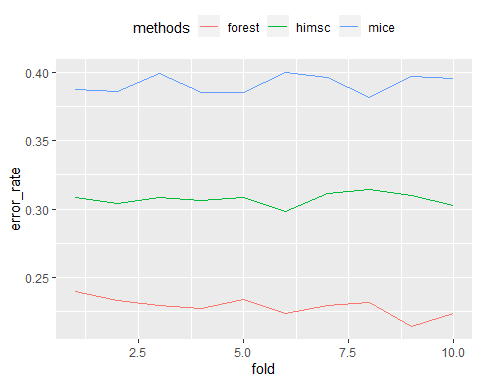
First, choose observations with known student\_r as test dataset. There are 41485 observations in total.

### Test three packages

## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150  
  
## Warning: Number of logged events: 150

We can find that error rates of Himsc package is around 33% and error rates of missForest is around 24%. Following is the plot of error rates.

## $himsc  
## [1] 0.3085824 0.3044107 0.3082671 0.3064127 0.3083414 0.2982160 0.3110843  
## [8] 0.3143684 0.3099542 0.3027965  
##   
## $forest  
## [1] 0.2396336 0.2328272 0.2296939 0.2270974 0.2336066 0.2234812 0.2293976  
## [8] 0.2311958 0.2137865 0.2232401  
##   
## $mice  
## [1] 0.3878978 0.3863582 0.3993733 0.3857281 0.3854870 0.4001929 0.3963855  
## [8] 0.3816297 0.3969631 0.3953713



### Permutation

Resample each variables separately to test their contribution to imputation. According to results, we can find ageyrs\_r have most contribution to correct imputation.

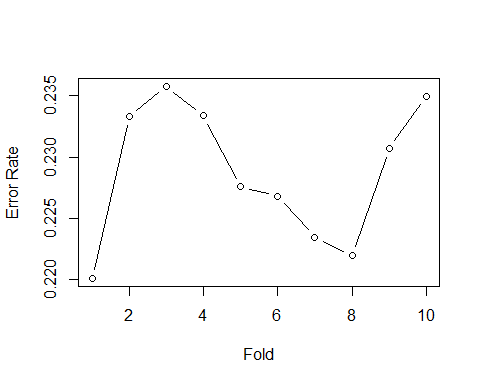
## currrltn gender area sexp1yr educate\_r pregnow\_r   
## 0.2275518 0.2283232 0.2298661 0.2307576 0.2314573 0.2316743   
## currmarr\_r SEScat eversex\_r ageyrs\_r   
## 0.2346151 0.2390504 0.2396531 0.2457038

### Test imputation performance

Then, randomly split test dataset into ten subsets and set them as missing values to test error rates of each fold when using different methods.

For misclassification rate:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2201 0.2243 0.2291 0.2288 0.2334 0.2358



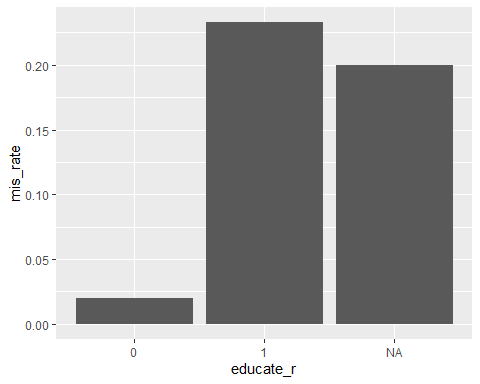
For the misclassification dataset:

There are 41485 observations are misclassified. Marginal distribution of misclassification rate in different gourps are significant among following variables: educate\_r， currmarr\_r, sexp1yr， pregnow\_r.

For educate\_r: **significant difference**

## Warning: Factor `educate\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

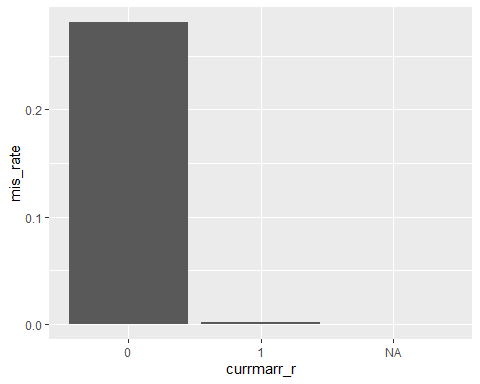
|  |  |  |
| --- | --- | --- |
| educate\_r | mis\_rate | sum |
| 0 | 0.0201183 | 845 |
| 1 | 0.2331488 | 40635 |
| NA | 0.2000000 | 5 |



For currmarr\_r: **significant difference**

## Warning: Factor `currmarr\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

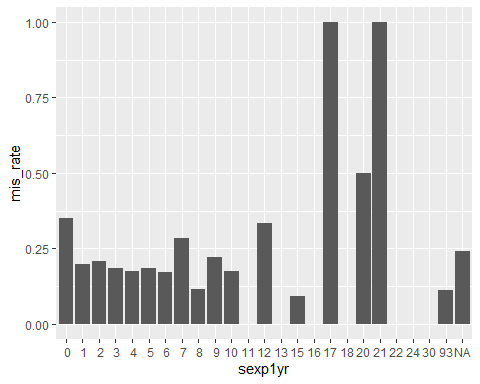
|  |  |  |
| --- | --- | --- |
| currmarr\_r | mis\_rate | sum |
| 0 | 0.2813214 | 33691 |
| 1 | 0.0017972 | 7790 |
| NA | 0.0000000 | 4 |



For sexp1yr: **significant difference**

## Warning: Factor `sexp1yr` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

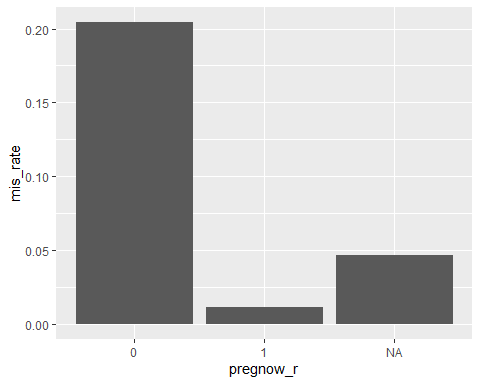
|  |  |  |
| --- | --- | --- |
| sexp1yr | mis\_rate | sum |
| 0 | 0.3499678 | 3106 |
| 1 | 0.1982607 | 15868 |
| 2 | 0.2095305 | 2854 |
| 3 | 0.1841432 | 782 |
| 4 | 0.1740891 | 247 |
| 5 | 0.1851852 | 108 |
| 6 | 0.1714286 | 70 |
| 7 | 0.2857143 | 49 |
| 8 | 0.1153846 | 26 |
| 9 | 0.2222222 | 9 |
| 10 | 0.1739130 | 23 |
| 11 | 0.0000000 | 1 |
| 12 | 0.3333333 | 6 |
| 13 | 0.0000000 | 2 |
| 15 | 0.0909091 | 11 |
| 16 | 0.0000000 | 1 |
| 17 | 1.0000000 | 2 |
| 18 | 0.0000000 | 2 |
| 20 | 0.5000000 | 4 |
| 21 | 1.0000000 | 1 |
| 22 | 0.0000000 | 1 |
| 24 | 0.0000000 | 1 |
| 30 | 0.0000000 | 2 |
| 93 | 0.1111111 | 27 |
| NA | 0.2411115 | 18282 |



For pregnow\_r: **significant difference**

## Warning: Factor `pregnow\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

|  |  |  |
| --- | --- | --- |
| pregnow\_r | mis\_rate | sum |
| 0 | 0.2045148 | 19270 |
| 1 | 0.0117302 | 2728 |
| NA | 0.0467938 | 577 |



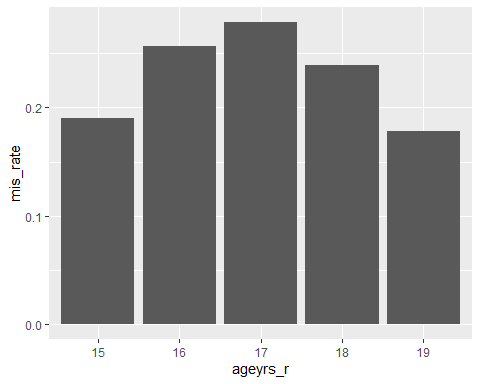
For visit:

|  |  |  |
| --- | --- | --- |
| visit | mis\_rate | sum |
| 1 | 0.2326072 | 2846 |
| 2 | 0.2512456 | 1405 |
| 3 | 0.2544118 | 1360 |
| 4 | 0.2477157 | 985 |
| 6 | 0.2300438 | 2969 |
| 7 | 0.2350863 | 2548 |
| 8 | 0.2529741 | 2858 |
| 9 | 0.2308064 | 2071 |
| 10 | 0.2625198 | 1897 |
| 11 | 0.2317909 | 2334 |
| 12 | 0.2256318 | 2770 |
| 13 | 0.2311771 | 2829 |
| 14 | 0.2114523 | 3126 |
| 15 | 0.2201772 | 2934 |
| 16 | 0.2102455 | 2811 |
| 17 | 0.1877912 | 2146 |
| 18 | 0.2182981 | 3596 |



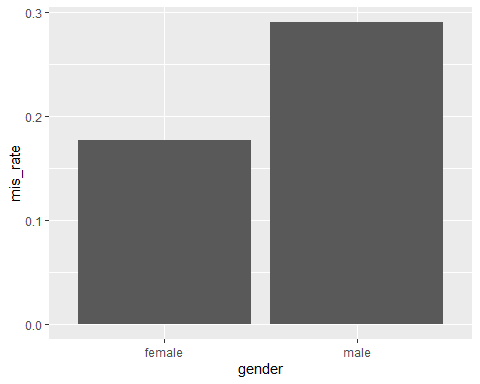
For ageyrs\_r:

|  |  |  |
| --- | --- | --- |
| ageyrs\_r | mis\_rate | sum |
| 15 | 0.1905854 | 7414 |
| 16 | 0.2566554 | 8264 |
| 17 | 0.2790518 | 7762 |
| 18 | 0.2389831 | 9440 |
| 19 | 0.1785009 | 8605 |



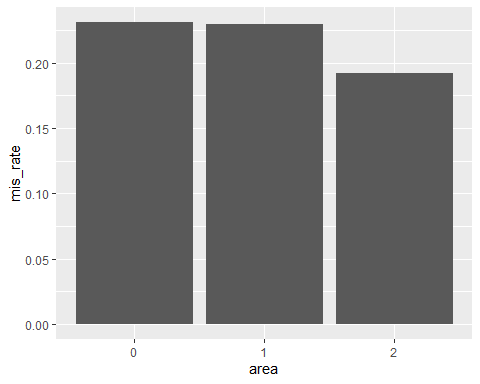
For gender:

|  |  |  |
| --- | --- | --- |
| gender | mis\_rate | sum |
| female | 0.1771872 | 22575 |
| male | 0.2904283 | 18910 |



For area:

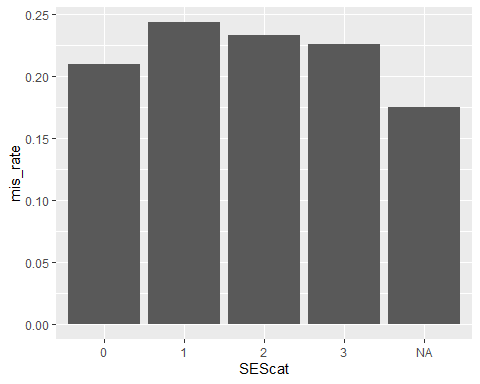
|  |  |  |
| --- | --- | --- |
| area | mis\_rate | sum |
| 0 | 0.2313695 | 31266 |
| 1 | 0.2298160 | 7828 |
| 2 | 0.1919699 | 2391 |



For SEScat:

## Warning: Factor `SEScat` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

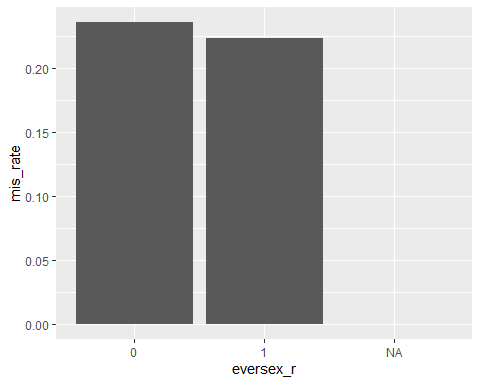
|  |  |  |
| --- | --- | --- |
| SEScat | mis\_rate | sum |
| 0 | 0.2101110 | 8110 |
| 1 | 0.2439437 | 9453 |
| 2 | 0.2336167 | 11292 |
| 3 | 0.2258220 | 12470 |
| NA | 0.1750000 | 160 |



For eversex:

## Warning: Factor `eversex\_r` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

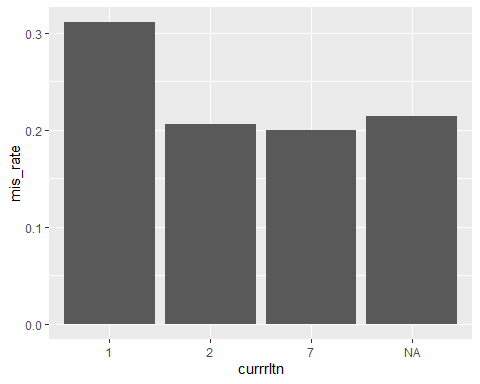
|  |  |  |
| --- | --- | --- |
| eversex\_r | mis\_rate | sum |
| 0 | 0.2361626 | 17814 |
| 1 | 0.2232878 | 23669 |
| NA | 0.0000000 | 2 |



For currrltn:

## Warning: Factor `currrltn` contains implicit NA, consider using  
## `forcats::fct\_explicit\_na`

|  |  |  |
| --- | --- | --- |
| currrltn | mis\_rate | sum |
| 1 | 0.3111453 | 9089 |
| 2 | 0.2057025 | 32372 |
| 7 | 0.2000000 | 10 |
| NA | 0.2142857 | 14 |



knitr::opts\_chunk$set(echo = FALSE)  
library(tidyverse)  
library(missForest)  
library(ggplot2)  
library(caret)  
library(Hmisc)  
library(mice)  
library(haven)  
new\_data <- read\_dta("R1\_R18\_partcleaned\_20190913.dta")  
student\_data = new\_data %>%  
 as.data.frame() %>%   
 mutate(student\_r = ifelse(occup1\_r == 8 | occup2\_r == 8, "yes", "no"),  
 student\_r = ifelse(is.na(student\_r) & occup1\_r == 20, "no", student\_r))  
#head(student\_data)  
plot\_data = student\_data %>%   
 filter(ageyrs\_r <= 19 & ageyrs\_r>=15) %>%   
 mutate(student\_r = as.factor(student\_r),  
 ageyrs\_r = as.factor(ageyrs\_r),  
 gender = ifelse(female==1, "female", "male")) %>%   
 select(visit, ageyrs\_r, gender, student\_r)  
plot\_data %>%   
 group\_by(visit, gender, student\_r) %>%   
 dplyr::summarize(count = n()) %>%   
 spread(key = student\_r, value = count) %>%   
 mutate(na = ifelse(is.na(`<NA>`), 0, `<NA>`),  
 sum = no + yes + na) %>%   
 select(-`<NA>`) %>%  
 mutate(no\_prc = no/sum,  
 yes\_prc = yes/sum,  
 na\_prc = na/sum) %>%   
 select(-c(no,yes,na,sum)) %>%   
 gather(key = student\_r, value = percentage, no\_prc:na\_prc) %>%   
 ggplot(aes(x=visit, y=percentage, fill=student\_r))+  
 geom\_bar(stat='identity') +  
 facet\_grid(. ~ gender) +  
 theme(legend.position="top")  
plot\_data %>%   
 group\_by(visit, ageyrs\_r, student\_r) %>%   
 dplyr::summarize(count = n()) %>%   
 spread(key = student\_r, value = count) %>%   
 mutate(na = ifelse(is.na(`<NA>`), 0, `<NA>`),  
 sum = no + yes + na) %>%   
 select(-`<NA>`) %>%  
 mutate(no\_prc = no/sum,  
 yes\_prc = yes/sum,  
 na\_prc = na/sum) %>%   
 select(-c(no,yes,na,sum)) %>%   
 gather(key = student\_r, value = percentage, no\_prc:na\_prc) %>%   
 ggplot(aes(x=visit, y=percentage, fill=student\_r))+  
 geom\_bar(stat='identity') +  
 facet\_grid(. ~ ageyrs\_r) +  
 theme(legend.position="top")  
table = plot\_data %>%  
 group\_by(visit, student\_r, gender, ageyrs\_r) %>%  
 dplyr::summarize(count = n()) %>%  
 spread(key = student\_r, value = count) %>%   
 mutate(na = ifelse(is.na(`<NA>`), 0, `<NA>`),  
 sum = no + yes + na) %>%   
 select(-`<NA>`) %>%  
 mutate(no\_prc = round(no/sum, 4)\*100,   
 yes\_prc = round(yes/sum, 4)\*100,   
 na\_prc = round(na/sum, 4)\*100) %>%  
 select(gender, visit, ageyrs\_r, no, no\_prc, yes, yes\_prc, na, na\_prc, sum) %>%  
 ungroup()  
table %>%   
 filter(na\_prc > 0) %>%   
 select(gender, visit, ageyrs\_r, na\_prc) %>%   
 mutate(visit = as.factor(visit)) %>%   
 mutate(ageyrs\_r = as.factor(ageyrs\_r)) %>%   
 ggplot(aes(x=visit, y=na\_prc, fill=ageyrs\_r))+  
 geom\_bar(stat='identity', position="dodge") +  
 facet\_grid(. ~ gender) +  
 theme(legend.position="top")  
table %>%  
 filter(visit %in% c(1,2,3,4,11,12,16,17)) %>%   
 select(gender, visit, ageyrs\_r, na\_prc) %>%   
 spread(key = ageyrs\_r, value = na\_prc) %>%   
 knitr::kable(digits = 3)  
target\_data = student\_data %>%   
 filter(ageyrs\_r <= 19 & ageyrs\_r>=15) %>%   
 mutate(student\_r = as.factor(student\_r),  
 visit = as.factor(visit),  
 gender = ifelse(female==1, "female", "male")) %>%   
 select(student\_r, visit, ageyrs\_r, gender, area, educate\_r, currmarr\_r, sexp1yr, SEScat, study\_id, pregnow\_r, numchild\_r, eversex\_r, currrltn)  
summary(target\_data)  
summary(as.factor(target\_data$sexp1yr))  
target\_data %>%   
 group\_by(gender, pregnow\_r) %>%   
 dplyr::summarize(n = n())  
summary(as.factor(target\_data$currrltn))  
impu\_data = target\_data %>%   
 mutate(sexp1yr = ifelse(sexp1yr %in% c(97, 98, 99), NA, sexp1yr)) %>%   
 mutate(pregnow\_r = ifelse(gender=="male", 98, pregnow\_r),  
 pregnow\_r = as.factor(pregnow\_r)) %>%   
 mutate(currrltn = ifelse(currrltn %in% c(0, 8,9), NA, currrltn),  
 currrltn = as.factor(currrltn)) %>%   
 select(-c(numchild\_r, study\_id)) %>%   
 mutate(visit = as.factor(visit),  
 area = as.factor(area),  
 educate\_r = as.factor(educate\_r),  
 currmarr\_r = as.factor(currmarr\_r),  
 SEScat = as.factor(SEScat),  
 eversex\_r = as.factor(eversex\_r),  
 gender = as.factor(gender))  
summary(impu\_data)  
test\_data = impu\_data[!is.na(impu\_data$student\_r),]  
set.seed(123)  
flds <- createFolds(1:nrow(test\_data), k = 10, list = TRUE, returnTrain = FALSE)  
#flds  
# error\_rate = vector("list", 10)  
set.seed(123)  
test\_data = as.data.frame(as.matrix(test\_data))  
error\_himsc = rep(NA, 10)  
error\_forest = rep(NA, 10)  
error\_mice = rep(NA, 10)  
for (n in 1:10){  
 na\_data = test\_data  
 na\_data[flds[[n]], 1] = NA  
 # missForest  
 impu\_forest = missForest(na\_data)  
 impu\_forest\_df = impu\_forest$ximp  
 # Himsc  
 impu\_himsc = aregImpute(~ student\_r + visit + ageyrs\_r + gender + area + educate\_r + currmarr\_r + SEScat + pregnow\_r + eversex\_r + currrltn, data = na\_data)  
 impu\_himsc\_l = impute.transcan(impu\_himsc, data=na\_data, imputation=1, list.out=TRUE, pr=FALSE, check=FALSE)  
 impu\_himsc\_df = as.data.frame(impu\_himsc\_l)  
 # mice  
 mice\_data = mice(impu\_data,seed = 123)  
 impu\_mice\_df = complete(mice\_data, 1)  
 # calculate error rates  
 # yes = 2, no = 1  
 error\_himsc[n] = sum(abs(as.numeric(impu\_himsc\_df[flds[[n]], 1]) - as.numeric(test\_data[flds[[n]], 1])))/length(test\_data[flds[[n]], 1])  
 error\_forest[n] = sum(abs(as.numeric(impu\_forest\_df[flds[[n]], 1]) - as.numeric(test\_data[flds[[n]], 1])))/length(test\_data[flds[[n]], 1])  
 error\_mice[n] = sum(abs(as.numeric(impu\_mice\_df[flds[[n]], 1]) - as.numeric(test\_data[flds[[n]], 1])))/length(test\_data[flds[[n]], 1])  
}  
res\_error = list(himsc = error\_himsc, forest = error\_forest, mice = error\_mice)  
res\_error  
res\_error %>%   
 as.data.frame() %>%   
 mutate(fold = 1:10) %>%   
 gather(key = methods, value = error\_rate, himsc:mice) %>%   
 ggplot(aes(x = fold, y = error\_rate, color = methods)) + geom\_line() +  
 theme(legend.position="top")  
set.seed(123)  
var\_mis <- matrix(data=NA,nrow=10,ncol=10) # 10 variables and 10 folds for each variables  
for (j in 3:12) {  
 permu\_data = test\_data  
 flds <- createFolds(1:nrow(permu\_data), k = 10, list = TRUE, returnTrain = FALSE)  
 target\_variable = permu\_data[,j]  
 new\_variable = sample(target\_variable, nrow(permu\_data), replace = T)  
 permu\_data[,j] = new\_variable  
 permu\_data = as.data.frame(as.matrix(permu\_data))  
 for (n in 1:10) {  
 na\_data = permu\_data  
 na\_data[flds[[n]], 1] = NA  
 # missForest  
 impu\_forest = missForest(na\_data)  
 impu\_forest\_df = impu\_forest$ximp  
 # calculate error rates  
 # yes = 2, no = 1  
 var\_mis[j-2,n] = sum(abs(as.numeric(impu\_forest\_df[flds[[n]], 1]) - as.numeric(permu\_data[flds[[n]], 1])))/length(permu\_data[flds[[n]], 1])  
 }  
}  
colnames(var\_mis) = c("fold1", "fold2", "fold3", "fold4", "fold5", "fold6", "fold7", "fold8", "fold9", "fold10")  
rownames(var\_mis) = c("ageyrs\_r","gender","area","educate\_r","currmarr\_r","sexp1yr","SEScat","pregnow\_r","eversex\_r","currrltn")  
apply(var\_mis,1,mean) %>%   
 sort()  
flds <- createFolds(1:nrow(test\_data), k = 10, list = TRUE, returnTrain = FALSE)  
#flds  
# error\_rate = vector("list", 10)  
set.seed(123)  
test\_data = as.data.frame(as.matrix(test\_data))  
error\_forest = rep(NA, 10) # indicate misclassification rate  
error\_index = rep(0,nrow(test\_data)) # indcate misclassification rows  
for (n in 1:10){  
 na\_data = test\_data  
 na\_data[flds[[n]], 1] = NA  
 # missForest  
 impu\_forest = missForest(na\_data)  
 impu\_forest\_df = impu\_forest$ximp  
 # calculate error rates  
 # yes = 2, no = 1  
 error\_forest[n] = sum(abs(as.numeric(impu\_forest\_df[flds[[n]], 1]) - as.numeric(test\_data[flds[[n]], 1])))/length(test\_data[flds[[n]], 1])  
 error\_vector = abs(as.numeric(impu\_forest\_df[, 1]) - as.numeric(test\_data[, 1]))  
 error\_index = error\_index + error\_vector  
}  
summary(error\_forest)  
plot(1:10,error\_forest,xlab="Fold",ylab="Error Rate",type="b",lty=1)  
error\_data = test\_data %>%   
 mutate(miscla = error\_index,  
 miscla = as.factor(miscla))  
educate\_error = error\_data %>%   
 group\_by(miscla, educate\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(educate\_r, mis\_rate, sum)   
educate\_error %>% knitr::kable()  
educate\_error %>%   
 ggplot(aes(educate\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
currmarr\_error = error\_data %>%   
 group\_by(miscla, currmarr\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(currmarr\_r, mis\_rate, sum)   
currmarr\_error %>% knitr::kable()  
currmarr\_error %>%   
 ggplot(aes(currmarr\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
sexp1yr\_error = error\_data %>%   
 group\_by(miscla, sexp1yr) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1),  
 miscla\_0 = ifelse(is.na(miscla\_0), 0, miscla\_0)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(sexp1yr, mis\_rate, sum)   
sexp1yr\_error %>% knitr::kable()  
sexp1yr\_error %>%   
 ggplot(aes(sexp1yr, mis\_rate)) +  
 geom\_bar(stat = "identity")  
pregnow\_error = error\_data %>%   
 group\_by(miscla, pregnow\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(pregnow\_r, mis\_rate, sum)   
pregnow\_error[c(1,2,4),] %>% knitr::kable()  
pregnow\_error[c(1,2,4),] %>%   
 ggplot(aes(pregnow\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
visit\_error = error\_data %>%   
 #mutate(visit = as.numeric(visit)) %>% # if as.numeric directly, it will be based on level  
 mutate(visit = as.numeric(as.character(visit))) %>%   
 group\_by(miscla, visit) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(visit, mis\_rate, sum)   
visit\_error %>% knitr::kable()  
visit\_error %>%   
 ggplot(aes(visit, mis\_rate)) +  
 geom\_bar(stat = "identity")  
age\_error = error\_data %>%   
 group\_by(miscla, ageyrs\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(ageyrs\_r, mis\_rate, sum)   
age\_error %>% knitr::kable()  
age\_error %>%   
 ggplot(aes(ageyrs\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
gender\_error = error\_data %>%   
 group\_by(miscla, gender) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(gender, mis\_rate, sum)   
gender\_error %>% knitr::kable()  
gender\_error %>%   
 ggplot(aes(gender, mis\_rate)) +  
 geom\_bar(stat = "identity")  
area\_error = error\_data %>%   
 group\_by(miscla, area) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(area, mis\_rate, sum)   
area\_error %>% knitr::kable()  
area\_error %>%   
 ggplot(aes(area, mis\_rate)) +  
 geom\_bar(stat = "identity")  
SEScat\_error = error\_data %>%   
 group\_by(miscla, SEScat) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(SEScat, mis\_rate, sum)   
SEScat\_error %>% knitr::kable()  
SEScat\_error %>%   
 ggplot(aes(SEScat, mis\_rate)) +  
 geom\_bar(stat = "identity")  
eversex\_error = error\_data %>%   
 group\_by(miscla, eversex\_r) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(miscla\_1 = ifelse(is.na(miscla\_1), 0, miscla\_1)) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(eversex\_r, mis\_rate, sum)   
eversex\_error %>% knitr::kable()  
eversex\_error %>%   
 ggplot(aes(eversex\_r, mis\_rate)) +  
 geom\_bar(stat = "identity")  
currrltn\_error = error\_data %>%   
 group\_by(miscla, currrltn) %>%   
 dplyr::summarize(n = n()) %>%   
 ungroup() %>%   
 mutate(miscla = str\_c("miscla\_", miscla)) %>%   
 spread(key = miscla, value = n) %>%   
 mutate(sum = miscla\_0 + miscla\_1,  
 mis\_rate = miscla\_1/sum) %>%   
 select(currrltn, mis\_rate, sum)   
currrltn\_error %>% knitr::kable()  
currrltn\_error %>%   
 ggplot(aes(currrltn, mis\_rate)) +  
 geom\_bar(stat = "identity")