

Assignment 1

Xiaoshu Gui

1/24/2019

```
# Load packages
library(dplyr)
library(purrr)
library(magrittr)
library(tidyr)
library(tibble)
library(stringr)

# Read data
jss <- read.csv("datjss.csv", header=T, sep=",")
sss <- read.csv("datsss.csv", header=T, sep=",")
stu <- read.csv("datstu.csv", header=T, sep=",", na.strings=c("", " ", "NA"))
# Note that there are empty cells in stu, we replace them with NAs first.
```

Exercise 1

- Number of students: 340823
- Number of schools: 6165
- Number of programs: 33

```
pgm <- stu %>%
  select(starts_with("choice")) %>% # get a subset of stu (choicepgm1 ~ 6)
  map(., function(x) unique(x)) %>% # find the number of programs in each "choicepgm"
  unlist %>%
  unique() %>% # join all the unique programs in 6 choices and find the unique value.
  # Note that one program does not have a name.
  length()
```

- Number of choices (school, program):

```
stu1 <- stu %>%
  subset(., is.na(.[5:16]))

# drop_na(., schoolcode1:choicepgm6) # tease out NAs in school and program choices

# map(., function(x) gsub("^$|^$", NA, x)) # replace all empty cells with NAs.
```

- Missing test score: 179887 (number of NAs in "score" in datstu)
- Apply to the same school (different programs)
- Apply to less than 6 choices: 20988

```
# rule out NAs in choicepgm, leaving us students who apply 6 choices
stu3 <- drop_na(stu, starts_with("choice"))
```

```
# then students who apply less than 6 choices should be total number of students minus those who apply 6 choices
340823 - count(stu3)
```

```
##          n
## 1 20988
```

```
# na.omit(stu[11:16])
```

```
summary(stu)
```

```
##          X          score          agey          male
## Min.      :    1    Min.    :158.0    Min.    : 9.00    Min.    :0.000
## 1st Qu.: 85206    1st Qu.:252.0    1st Qu.:16.00    1st Qu.:0.000
## Median :170412    Median :283.0    Median :17.00    Median :1.000
## Mean    :170412    Mean    :291.1    Mean    :17.13    Mean    :0.549
## 3rd Qu.:255618    3rd Qu.:324.0    3rd Qu.:18.00    3rd Qu.:1.000
## Max.    :340823    Max.    :469.0    Max.    :57.00    Max.    :1.000
##          NA's      :179887    NA's      :650
## schoolcode1    schoolcode2    schoolcode3    schoolcode4
## Min.      : 10101    Min.      : 10101    Min.      : 10101    Min.      : 10101
## 1st Qu.: 21502    1st Qu.: 21502    1st Qu.: 21502    1st Qu.: 21502
## Median : 50105    Median : 50107    Median : 50113    Median : 50202
## Mean     :239365    Mean     :244223    Mean     :264627    Mean     :315661
## 3rd Qu.: 61201    3rd Qu.: 61202    3rd Qu.: 61202    3rd Qu.: 61203
## Max.     :9100501    Max.     :9100501    Max.     :9100501    Max.     :9100501
## NA's     :102      NA's     :163      NA's     :195      NA's     :406
## schoolcode5    schoolcode6    choicepgm1
## Min.      : 10101    Min.      : 10101    General Arts :125850
## 1st Qu.: 21201    1st Qu.: 21203    Business     : 63167
## Median : 50204    Median : 50204    Home Economics : 51922
## Mean     : 47539    Mean     : 47354    General Science: 28777
## 3rd Qu.: 60801    3rd Qu.: 60704    Agriculture   : 26810
## Max.     :9100101    Max.     :9090401    (Other)       : 44194
## NA's     :17140    NA's     :17088    NA's         : 103
## choicepgm2    choicepgm3    choicepgm4
## General Arts :122728    General Arts :122794    General Arts :121461
## Business     : 65835    Business     : 62312    Business     : 58483
## Home Economics: 51044    Home Economics: 51702    Home Economics: 51500
## Agriculture   : 30313    Agriculture   : 32850    Agriculture   : 36925
## Visual Arts   : 26073    Visual Arts   : 27224    Visual Arts   : 27406
## (Other)       : 44667    (Other)       : 43746    (Other)       : 44642
## NA's         : 163      NA's         : 195      NA's         : 406
## choicepgm5    choicepgm6
## General Arts :122379    General Arts :124181
## Business     : 57820    Business     : 55411
## Home Economics: 51467    Home Economics: 50776
## Agriculture   : 34732    Agriculture   : 35734
## Visual Arts   : 21618    Visual Arts   : 23501
## (Other)       : 34174    (Other)       : 32266
## NA's         : 18633    NA's         : 18954
##          jssdistrict    rankplace
## Accra Metropolitan      : 33068    Min.      : 1.00
## Kumasi Metro             : 22640    1st Qu.: 1.00
## Tema                     : 12546    Median : 3.00
## Shama/Ahanta/East (Sekondi/Takoradi): 8464    Mean     :15.45
## Ga West (Amasaman)       : 7970    3rd Qu.: 4.00
## (Other)                  :256108    Max.     :99.00
## NA's                     : 27      NA's     :179888
```

```
summary(sss)
```

```
##           X                sssdistrict      ssslong
## Min.      :    1  Accra Metropolitan: 271  Min.      :-2.9267
## 1st Qu.:1542  Kumasi Metro      : 262  1st Qu.: -1.5972
## Median :3083                : 179  Median :-0.7990
## Mean    :3083  Accra Metro      : 170  Mean    :-0.8679
## 3rd Qu.:4624  Ho Municipal      : 104  3rd Qu.: -0.1971
## Max.    :6165  Tema              :  96  Max.    :  1.0327
##           (Other)              :5083  NA's     :3100
##           ssslat
## Min.      : 4.835
## 1st Qu.: 5.690
## Median : 6.383
## Mean    : 6.741
## 3rd Qu.: 7.031
## Max.    :11.036
## NA's     :3100
```

Excercise 2