Rearranging Data Frame

Boyie Chen

10/8/2019

I. Setting Working Directory

In this section, we are going to deal with data frame. The example data is from: https://udb.moe.edu.tw/StatCardList/University/000012CE619F/0003/ We can download the data from the website and put the csv file in the diretory.

```
#set work directory
path = "/Users/Andy 1/Google 雲端硬碟/108-1/1 三 345 統計 TA/3 TA 課內容/0 formal/Data"
setwd(path)
```

Then `read.csv()` is a function that helps us get access to the data frame.

```
data = read.csv('107NTU Student.csv')
head(data) #1st row is not the real label...
```

```
##   學1.1.正式學籍在學學生人數.以.系.所..統計       X       X.1             X.2
## 1                                   學年度  設立別  學校類別  學校統計處代碼
## 2                                      107   公立  一般大學            0003
## 3                                      107   公立  一般大學            0003
## 4                                      107   公立  一般大學            0003
## 5                                      107   公立  一般大學            0003
## 6                                      107   公立  一般大學            0003
##            X.3      X.4                X.5          X.6             X.7
## 1     學校名稱  系所代碼              系所名稱      學制班別  在學學生數小計
## 2 國立臺灣大學  01131001 華語教學碩士學位學程  碩士班(日間)              54
## 3 國立臺灣大學  02151002          音樂學研究所  碩士班(日間)              27
## 4 國立臺灣大學  02151002          音樂學研究所        博士班              10
## 5 國立臺灣大學  02152003              戲劇學系  學士班(日間)             178
## 6 國立臺灣大學  02152003              戲劇學系  碩士班(日間)              33
##            X.8          X.9
## 1 在學學生數男  在學學生數女
## 2            4           50
## 3            9           18
## 4            5            5
## 5           79           99
## 6           12           21
```

```
data[1,]
```

```
##   學1.1.正式學籍在學學生人數.以.系.所..統計       X       X.1             X.2
## 1                                   學年度  設立別  學校類別  學校統計處代碼
##        X.3      X.4      X.5      X.6            X.7          X.8
## 1 學校名稱  系所代碼  系所名稱  學制班別  在學學生數小計  在學學生數男
##            X.9
## 1 在學學生數女
```

```
#Let's drop the 1st row, and assign a new DataFrame
data = data[2:nrow(data),] #what is the disadvantage of doing this? Your 'data' is overlapped
head(data)
```

```
##   學1.1.正式學籍在學學生人數.以.系.所..統計     X       X.1  X.2
## 2                                        107 公立 一般大學 0003
## 3                                        107 公立 一般大學 0003
## 4                                        107 公立 一般大學 0003
## 5                                        107 公立 一般大學 0003
## 6                                        107 公立 一般大學 0003
## 7                                        107 公立 一般大學 0003
##            X.3      X.4                   X.5          X.6 X.7 X.8 X.9
## 2 國立臺灣大學 01131001 華語教學碩士學位學程 碩士班(日間)  54   4  50
## 3 國立臺灣大學 02151002         音樂學研究所 碩士班(日間)  27   9  18
## 4 國立臺灣大學 02151002         音樂學研究所       博士班  10   5   5
## 5 國立臺灣大學 02152003             戲劇學系 學士班(日間) 178  79  99
## 6 國立臺灣大學 02152003             戲劇學系 碩士班(日間)  33  12  21
## 7 國立臺灣大學 02152003             戲劇學系       博士班   3   2   1
```

```
data = data[1:302,] #further drop out the last several rows
colnames(data) #Then we have another problem...the name of col is incorrect
```

```
##  [1] "學1.1.正式學籍在學學生人數.以.系.所..統計"
##  [2] "X"
##  [3] "X.1"
##  [4] "X.2"
##  [5] "X.3"
##  [6] "X.4"
##  [7] "X.5"
##  [8] "X.6"
##  [9] "X.7"
## [10] "X.8"
## [11] "X.9"
```

```
#Modify the col names
colnames(data) = c('year', 'x1', 'x2', 'x3', 'x4', 'x5', '系所','BachMD','在學生數','Male','Female')
colnames(data)
```

```
##  [1] "year"     "x1"       "x2"       "x3"       "x4"       "x5"
##  [7] "系所"     "BachMD"   "在學生數" "Male"     "Female"
```

```
#we further drop out other cols that are not interesting
data = data[,c('year','系所','BachMD','在學生數','Male','Female')]

#drop out the rows that we are not interested at
BachMD = data$BachMD # 因為每個 col 是 factor，所以我不能寫：
data2 = data['BachMD' == '學士班 (日間)',] # 這是 char 對 char 的邏輯判斷
# 但我的 col 裡沒有 char，只有 factor，所以要這樣寫才能比：
data = data[BachMD == '學士班 (日間)',] # 才能篩選出學士班的 row

head(data['在學生數'])
```

```
##     在學生數
## 5        178
## 10       259
## 13       169
## 16       178
## 19       555
## 22       255
```

```r
numStudent = data$在學生數 #We can use chinese as a variable's name, cool!
numStudent = data$'在學生數' #or write in this way
head(numStudent)
```

```
## [1] 178 259 169 178 555 255
## 153 Levels:  1 10 101 104 106 109 11 111 112 113 114 115 118 119 12 ... 在學學生數小計
```

```r
mean(numStudent) #error, because of "factor"
```

```
## [1] NA
```

```r
is.numeric(numStudent) #only when data type == numeric, then you can calculate
```

```
## [1] FALSE
```

```r
is.factor(numStudent) #What is 'factor'? 類別變數
```

```
## [1] TRUE
```

II. factor 類別變數

the data type "factor" helps us to deal with ranks, levels or just to distinguish what category something is. For example, we may have a set of data that is the ranking of happiness. 1 may represents the least happy situation, and 5 represent the happiest situation. In this case, the number 1 to 5 no longer have the property of real numbers. They become orders.

Another example is that factors can be labels. We can assign `DogLovers` as a variable to represent whether someone is a dog lover or not. So the value contained in `DogLovers` may be `True` or `False`(`0` or `1`)

In order to recover the property of real numbers, we have to transform `factor` into `numeric`

The following is a *WRONG* way.

```r
#change factor to numeric, a WRONG way
numStudent2 = as.numeric(numStudent) #as.numeric() : only do the ranking
is.numeric(numStudent2)
```

```
## [1] TRUE
```

```r
head(numStudent2)
```

```
## [1]  42  67  39  42 116  66
```

```r
mean(numStudent2)
```

```
## [1] 68.8
```

The above method gives you the order among the vector. The true value is missing. Thus, this is the right way:

```r
#The right way
numStudent3 = as.numeric(levels(numStudent))[numStudent]
is.numeric(numStudent3)
```

```
## [1] TRUE
```

```r
head(numStudent3)
```

```
## [1] 178 259 169 178 555 255
```

```r
mean(numStudent3) #finally we get the average number of student in NTU among all departments.
```

```
## [1] 276.7333
```

We can save the file. It shows up in your working directory.

```
#Store the dataFrame that you've clean up
write.table(data, "data.txt", row.names=F) #write txt file
write.csv(data, "data.csv", row.names = F) #write csv file
```

III.Sorting 排序 & Barplot

```
#sort
# 將整個 dataFrame 的特定 col 轉成 numeric
data$在學生數 = as.numeric(levels(data$在學生數))[data$在學生數]
mean(data$在學生數) #same result as above
```

```
## [1] 276.7333
```

```
data = data[order(data$在學生數),] #order() 會回傳排序的 index
write.csv(data, "data2.csv", row.names = F) #save a sorted ver.

#plot
data2 = read.csv("data2.csv")

max(data$在學生數); min(data$在學生數)
```

```
## [1] 858
```

```
## [1] 3
```

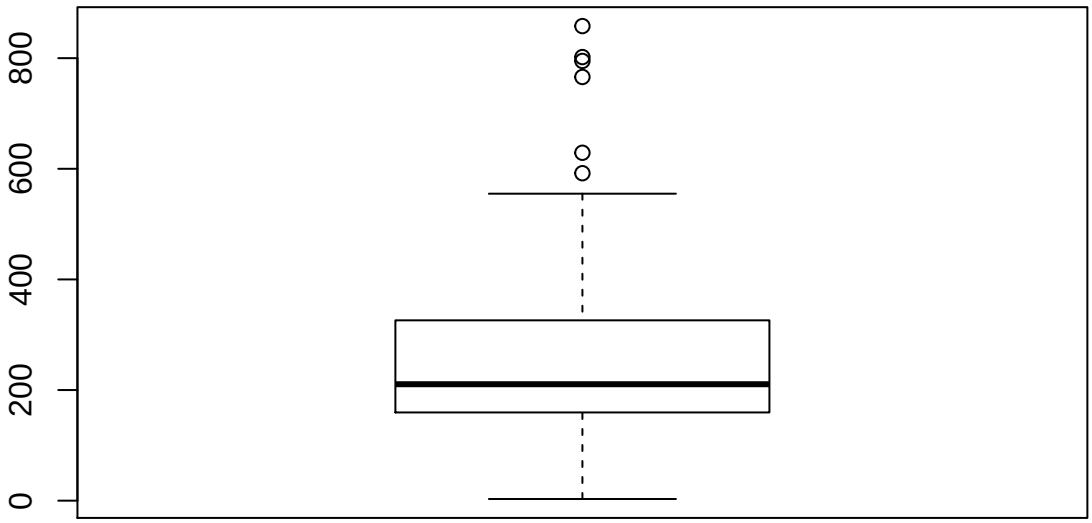
```
median(data$在學生數)
```

```
## [1] 210.5
```

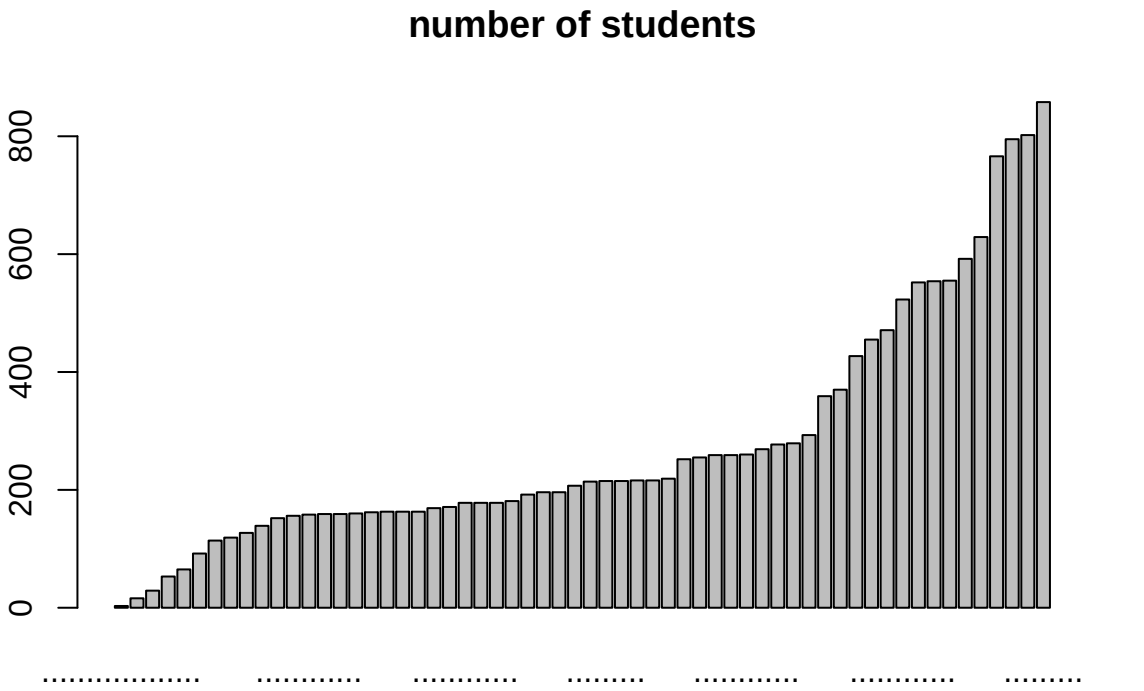
```
summary(data$在學生數)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0   159.8   210.5   276.7   309.5   858.0
```

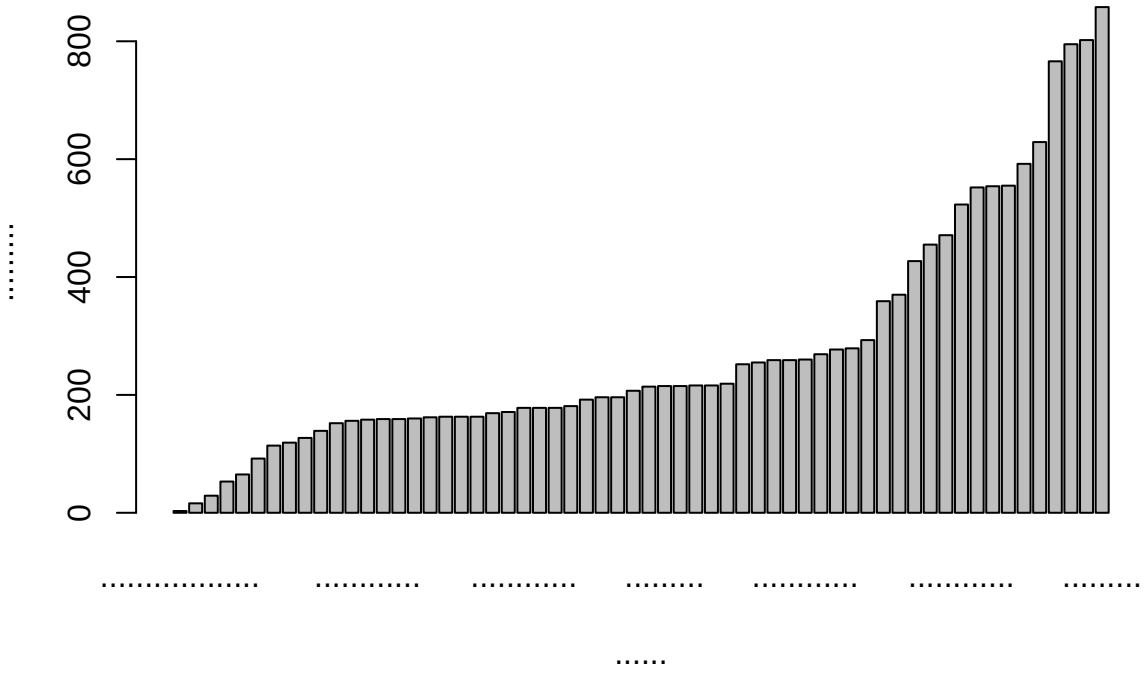
```
#boxplot
boxplot(data2$在學生數)
```



```r
#RStudio does not fully support chinese...QQ
barplot(data2$在學生數, names.arg = data2$系所, xlab = 'department', main = 'number of students')
```




```r
barplot(data2$在學生數, names.arg = data2$系所, xlab = '系所', ylab = '學生數', main = '各系所學生人數')
```



引入中文字型來解決繪圖問題

Reference: https://blog.gtwang.org/r/how-to-use-your-favorite-fonts-in-r-charts/

```
#install.packages("showtext")
#let us import chinese fonts
library(showtext)
```

```
## Loading required package: sysfonts
```

```
## Loading required package: showtextdb
```

```
showtext.auto(enable = TRUE)
```

```
## 'showtext.auto()' is now renamed to 'showtext_auto()'
## The old version still works, but consider using the new function in future code
```

```
font_add("PingFang", "/System/Library/Fonts/PingFang.ttc") #load your own font

#save the bar plot
png("output.png", width = 1920, height = 1080)
barplot(data2$在學生數, names.arg = data2$系所, xlab = '系所', ylab = '學生數', main = '各系所學生人數',family = "P
dev.off()
```

```
## pdf
##   2
```

如果是乾淨得多的 data frame，就好處理得多

```
##Another way to read csv
#if you first make the table cleaner...
data = read.csv('107NTU Student-2.csv', header = F) #F if the first row is not the name of variables
head(data)
```

```
##        V1     V2       V3              V4           V5       V6
## 1 學年度 設立別 學校類別 學校統計處代碼     學校名稱 系所代碼
## 2    107   公立 一般大學               3 國立臺灣大學  1131001
## 3    107   公立 一般大學               3 國立臺灣大學  2151002
## 4    107   公立 一般大學               3 國立臺灣大學  2151002
## 5    107   公立 一般大學               3 國立臺灣大學  2152003
## 6    107   公立 一般大學               3 國立臺灣大學  2152003
##                      V7           V8             V9          V10
## 1              系所名稱     學制班別 在學學生數小計 在學學生數男
## 2 華語教學碩士學位學程 碩士班(日間)             54            4
## 3          音樂學研究所 碩士班(日間)             27            9
## 4          音樂學研究所       博士班             10            5
## 5              戲劇學系 學士班(日間)            178           79
## 6              戲劇學系 碩士班(日間)             33           12
##            V11
## 1 在學學生數女
## 2           50
## 3           18
## 4            5
## 5           99
## 6           21
```

```
data = read.csv('107NTU Student-2.csv', header = T) #F if the first row is not the name of variables
head(data)
```

```
##   學年度 設立別 學校類別 學校統計處代碼     學校名稱 系所代碼
```

```
## 1    107    公立  一般大學              3 國立臺灣大學  1131001
## 2    107    公立  一般大學              3 國立臺灣大學  2151002
## 3    107    公立  一般大學              3 國立臺灣大學  2151002
## 4    107    公立  一般大學              3 國立臺灣大學  2152003
## 5    107    公立  一般大學              3 國立臺灣大學  2152003
## 6    107    公立  一般大學              3 國立臺灣大學  2152003
##                 系所名稱      學制班別 在學學生數小計 在學學生數男
## 1 華語教學碩士學位學程 碩士班(日間)           54           4
## 2          音樂學研究所 碩士班(日間)           27           9
## 3          音樂學研究所       博士班           10           5
## 4              戲劇學系 學士班(日間)          178          79
## 5              戲劇學系 碩士班(日間)           33          12
## 6              戲劇學系       博士班            3           2
##   在學學生數女
## 1           50
## 2           18
## 3            5
## 4           99
## 5           21
## 6            1
```

Note that... 請盡量將要處理的變數（行向量）bind 回 data frame，而非單獨抓出來處理

```
#easy to assign vlbs and do the operation
#A WRONG way to operate by vectors
male = data$'在學學生數男'
female = data$'在學學生數女'
ratio = male/female
tail(ratio) #there are a lot of NAs
```

```
## [1] NA NA NA NA NA NA
```

```
sum(is.na(ratio)) #I have 16 NA values
```

```
## [1] 16
```

```
max(ratio)
```

```
## [1] NA
```

```
max(ratio, na.rm=TRUE) #see the max value without NA
```

```
## [1] Inf
```

```
#remove NA from vectors
ratio = ratio[!is.na(ratio)]
sum(is.na(ratio)) #no NAs anymore
```

```
## [1] 0
```

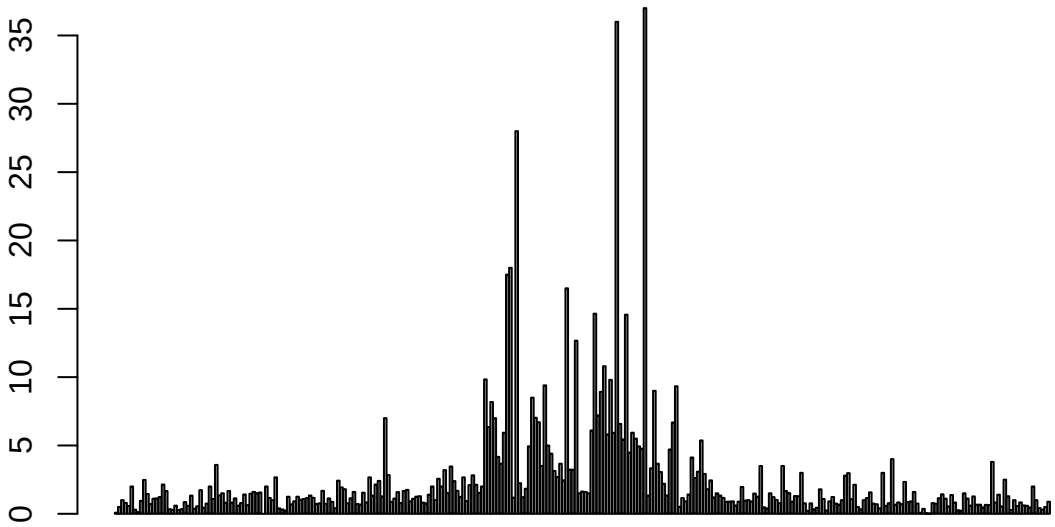
```
#remove Inf values from vectors
sum(ratio == Inf) #I have 3 Inf
```

```
## [1] 3
```

```
ratio = ratio[!is.infinite(ratio)]
sum(ratio == Inf)
```

```
## [1] 0
```

```r
#alternative way
ratio = ratio[is.finite(ratio)]
barplot(ratio) #no tags, not good enough, and hard to recover the tags
```




```r
#I'll show in a more proper way

#We just look at the undergraduate students
male = data2$Male
female = data2$Female
ratio = male/female
data2 = cbind.data.frame(data2, ratio) #add a new col to the dataFrame
#plot without handeling the Inf, so we have to set limit
png("output2.png", width = 3840, height = 2160)
barplot(data2$ratio, names.arg = data2$系所, ylim = c(0,10), family = "PingFang")
dev.off()
```

```
## pdf
##   2
```

也可以從網路引入 data frame

```r
#Use Data from Internet
#we can also import data from the internet
library(foreign)
gpa1 = read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/gpa1.dta")
```