Intro to R: Lab 2-walkthrough

Simon Caton

Making the student.df data frame

```
name <- c("Amy", "Bill", "Carl")
DAD <- c(80, 65, 50)
BDA <- c(70, 50, 80)
gender <- as.factor(c("F", "M", "M"))
nationality <- as.factor(c("IRL", "UK", "IRL"))
age <- c(20, 21, 22)
student.df <- data.frame(name, age, gender, nationality, DAD, BDA)
student.df$average <- (student.df$BDA + student.df$DAD)/2
```

Run the functions: print, str, and summary on the student.df data.frame

```
print(student.df) # print the values of the data frame
```

```
##   name age gender nationality DAD BDA average
## 1  Amy  20      F         IRL  80  70    75.0
## 2 Bill  21      M          UK  65  50    57.5
## 3 Carl  22      M         IRL  50  80    65.0
```

```
str(student.df) # tells us the structure of the data frame
```

```
## 'data.frame':    3 obs. of  7 variables:
##  $ name       : Factor w/ 3 levels "Amy","Bill","Carl": 1 2 3
##  $ age        : num  20 21 22
##  $ gender     : Factor w/ 2 levels "F","M": 1 2 2
##  $ nationality: Factor w/ 2 levels "IRL","UK": 1 2 1
##  $ DAD        : num  80 65 50
##  $ BDA        : num  70 50 80
##  $ average    : num  75 57.5 65
```

```
summary(student.df) # provides some summary statistics of the values in the data frame
```

```
##    name        age       gender nationality      DAD            BDA
##  Amy :1   Min.   :20.0   F:1    IRL:2       Min.   :50.0   Min.   :50.00
##  Bill:1   1st Qu.:20.5   M:2    UK :1       1st Qu.:57.5   1st Qu.:60.00
##  Carl:1   Median :21.0                      Median :65.0   Median :70.00
##           Mean   :21.0                      Mean   :65.0   Mean   :66.67
##           3rd Qu.:21.5                      3rd Qu.:72.5   3rd Qu.:75.00
##           Max.   :22.0                      Max.   :80.0   Max.   :80.00
##     average
##  Min.   :57.50
##  1st Qu.:61.25
##  Median :65.00
##  Mean   :65.83
```

```
##  3rd Qu.:70.00
##  Max.   :75.00
```

Complete the following to cast student.df$name to a character vector

```
student.df$name <- as.character(student.df$name)
```

Cast nationality to a character vector

```
student.df$nationality <- as.character(student.df$nationality)
```

Add another 5-10 students

```
student.df <- rbind(student.df, c("Dennis", 23, "M", "UK", 55, 65))
student.df <- rbind(student.df, c("Emily", 23, "F", "FR", 50, 55))
student.df <- rbind(student.df, c("Fred", 23, "M", "US", 70, 75))
student.df <- rbind(student.df, c("George", 23, "M", "FR", 65, 70))
student.df <- rbind(student.df, c("Harriot", 23, "F", "IRL", 80, 80))
student.df <- rbind(student.df, c("Isabelle", 23, "F", "US", 57, 90))

#check nothing weird has happened:
head(student.df) #head prints the first 5 rows
```

```
##     name age gender nationality DAD BDA average
## 1    Amy  20      F         IRL  80  70      75
## 2   Bill  21      M          UK  65  50    57.5
## 3   Carl  22      M         IRL  50  80      65
## 4 Dennis  23      M          UK  55  65  Dennis
## 5  Emily  23      F          FR  50  55   Emily
## 6   Fred  23      M          US  70  75    Fred
```

So because we didn't provide sufficient values to fill all columns in our data frame, average now has some erroneous data in it.

Let's also ensure that nothing strange has happened to our data frame itself

```
str(student.df)
```

```
## 'data.frame':    9 obs. of  7 variables:
##  $ name       : chr  "Amy" "Bill" "Carl" "Dennis" ...
##  $ age        : chr  "20" "21" "22" "23" ...
##  $ gender     : Factor w/ 2 levels "F","M": 1 2 2 2 1 2 2 1 1
##  $ nationality: chr  "IRL" "UK" "IRL" "UK" ...
##  $ DAD        : chr  "80" "65" "50" "55" ...
##  $ BDA        : chr  "70" "50" "80" "65" ...
##  $ average    : chr  "75" "57.5" "65" "Dennis" ...
```

Oh dear, it seems that our two numerical columns DAD and BDA have become strings... I guess we weren't careful enough when we populated our data frame with new values. Let's fix up our data frame.

```
student.df <- student.df[ , -7] # remove the average column
student.df$DAD <- as.numeric(student.df$DAD)
student.df$BDA <- as.numeric(student.df$BDA)
# now recompute the average for each student
student.df$average <- (student.df$BDA + student.df$DAD)/2
head(student.df) # check again
```

```
##     name age gender nationality DAD BDA average
## 1    Amy  20      F         IRL  80  70    75.0
## 2   Bill  21      M          UK  65  50    57.5
## 3   Carl  22      M         IRL  50  80    65.0
## 4 Dennis  23      M          UK  55  65    60.0
## 5  Emily  23      F          FR  50  55    52.5
## 6   Fred  23      M          US  70  75    72.5
```

Looks fine now, we can move on.

Rebuild the nationality factor

```
student.df$nationality <- as.factor(student.df$nationality)
levels(student.df$nationality)
```

```
## [1] "FR"  "IRL" "UK"  "US"
```

So now we get a more interesting version of what was in the initial lab sheet:

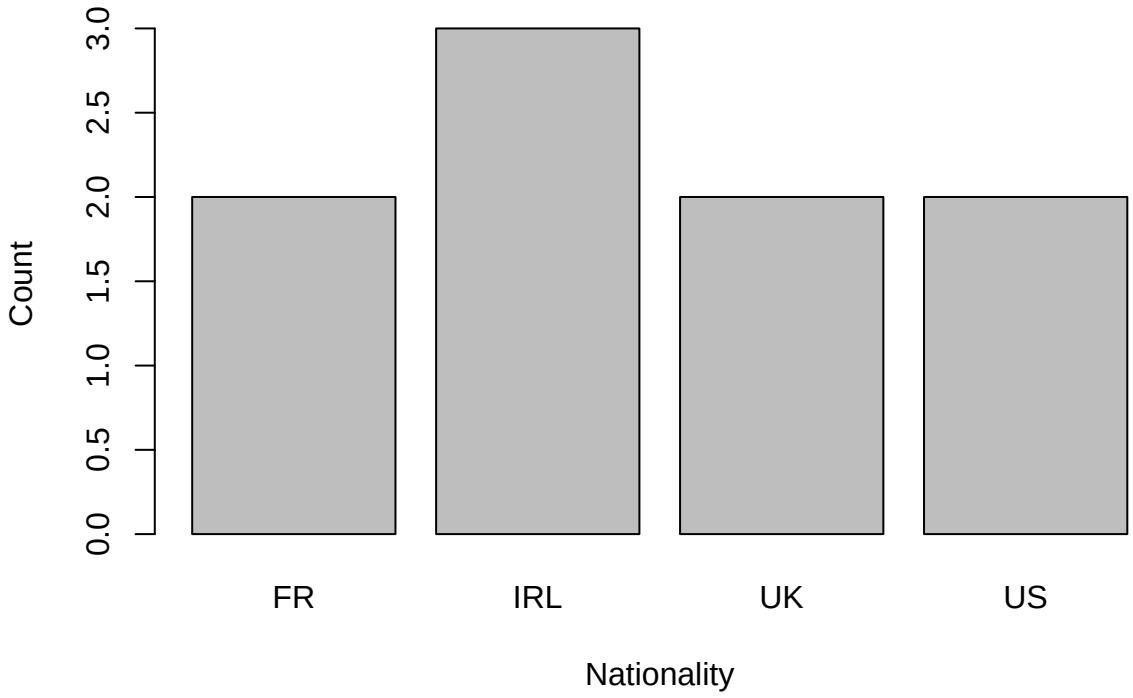
```
averages <- tapply(student.df$BDA, student.df$nationality, mean)
averages
```

```
##       FR      IRL       UK       US
## 62.50000 76.66667 57.50000 82.50000
```

```
table(student.df$nationality)
```

```
##
##  FR IRL  UK  US
##   2   3   2   2
```

```
barplot(table(student.df$nationality), xlab = "Nationality", ylab="Count")
```



compute the min, max, and standard deviation for BDA, and DAD

```
min(student.df$BDA)
```

```
## [1] 50
```

```
max(student.df$BDA)
```

```
## [1] 90
```

```
sd(student.df$BDA)
```

```
## [1] 12.61062
```

```
min(student.df$DAD)
```

```
## [1] 50
```

```
max(student.df$DAD)
```

```
## [1] 80
```

```
sd(student.df$DAD)
```

```
## [1] 11.56623
```

build a data.frame of the results from 1 + mean

A few different ways to do this, but this way is probably easiest (and most pragmatic)

```
mins <- c(min(student.df$BDA), min(student.df$DAD))
means <- c(mean(student.df$BDA), mean(student.df$DAD))
maxs <- c(max(student.df$BDA), max(student.df$DAD))
```

```
sds <- c(sd(student.df$BDA), sd(student.df$DAD))

subjects <- data.frame(mins, maxs, means, sds, row.names = c("BDA", "DAD"))
names(subjects) <- c("min", "max", "mean", "sd")
print(subjects)
```

```
##     min max     mean       sd
## BDA  50  90 70.55556 12.61062
## DAD  50  80 63.55556 11.56623
```

mtcars prep

```
data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$am <- factor(mtcars$am, labels=c("Automatic", "Manual"), levels=c(0,1))
```

Rerun str and summary to see the changes we made

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```