

Lab 7

Fan Yang (fy2232)

November 28, 2017

Instructions

Before you leave lab today make sure that you upload a .pdf file to the canvas page (this should have a .pdf extension). This should be the PDF output after you have knitted the file, we don't need the .Rmd file (don't upload the one with the .Rmd extension). The file you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. Note, however, in the file you upload you should the above header to have the date, your name, and your UNI. Similarly, when you save the file you should replace **UNI** with your actualy UNI.

Introduction

In today's lab we will use data on the 2829 fastest men's and women's 100m sprint times saved as dataframes `sprint.m.csv` and `sprint.f.csv`.

1. Load the dataframe `sprint.m.csv` and save it as `sprint.m`. Append a column to the dataframe called `CityDate` that is defined by concatenating the string entries in the `City` and `Date` columns. For example, entries "Berlin" and "16.08.2009" in the `City` and `Date` columns, respectively, produce an entry of "Berlin 16.08.2009" in the `CityDate` column. Note that there will be some weird characters in the city names.

```
sprint.m <- read.csv("sprint.m.csv", header = T, as.is = T)
CityDate <- paste(sprint.m$City, sprint.m$Date)
sprint.m <- cbind(sprint.m, CityDate)
```

2. We assume that every unique combination of city and date in the `CityDate` column corresponds to a unique track meet. How many unique track meets occur? How many other sprint times were recorded in the same track meet as Usain Bolt's legendary time of 9.58 seconds?

```
length(table(sprint.m$CityDate))
```

```
## [1] 1181
```

```
sum(sprint.m$CityDate==sprint.m$CityDate[which(sprint.m$Time==9.58)])
```

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

3. Compute a reduced version of `sprint.m` that only keeps the fastest time from each track meet. For example, of all rows that correspond to sprint times recorded at the "Berlin 16.08.2009" track meet, we will only keep Usain Bolt's row since his time of 9.58 was fastest. Hint: There are many ways to do this, `tapply()` or `split()` might be helpful. You can do this without using a loop. Call the result `sprint.m.fastest` and check that the number of rows is the same as the number of unique men's track meets. Display the first five rows. Note that if there are ties for first place, choose only one runner to represent the race in the reduced data frame. How you chose doesn't matter.

```
whichmin <- function(data){
  return (data[which.min(data$Time),])
}
sprint.m.fastest <- lapply(split(sprint.m, sprint.m$CityDate),
```

```

      function(data) return (data[which.min(data$Time),]) )
sprint.m.fastest <- data.frame(t(sapply(sprint.m.fastest, rbind)))
nrow(sprint.m.fastest) == length(table(sprint.m$CityDate))

```

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

```
head(sprint.m.fastest,5)
```

```

##              Rank  Time Wind              Name Country
## &Eacute;vry-Bondoufle 11.07.1994 2276 10.08 0.5      Bruny Surin      CAN
## Abbotsford 19.07.1997      1202 10.03 -2.1      Donovan Bailey    CAN
## Abbotsford 23.05.1992      1581 10.05 1.2      Davidson Ezinwa      NGR
## Abilene 20.05.2004          2532 10.09 2      Christie van Wyk    NAM
## Abilene 29.05.1993          2276 10.08 1.2      Bryan Bridgewater    USA
##              Birthdate              City
## &Eacute;vry-Bondoufle 11.07.1994 12.07.67 &Eacute;vry-Bondoufle
## Abbotsford 19.07.1997      16.12.67      Abbotsford
## Abbotsford 23.05.1992      22.11.71      Abbotsford
## Abilene 20.05.2004          12.10.77      Abilene
## Abilene 29.05.1993          07.09.70      Abilene
##              Date CityDate
## &Eacute;vry-Bondoufle 11.07.1994 11.07.1994      1
## Abbotsford 19.07.1997      19.07.1997      2
## Abbotsford 23.05.1992      23.05.1992      3
## Abilene 20.05.2004          20.05.2004      4
## Abilene 29.05.1993          29.05.1993      5

```

4. Load the women's dataframe `sprint.w.csv` and repeat steps (1) - (3) on this dataset so that what remains is `sprint.w.fastest`. Display the first five rows.

```

#1)
sprint.w <- read.csv("sprint.w.csv", header = T, as.is = T)
CityDate <- paste(sprint.w$City, sprint.w$Date)
sprint.w <- cbind(sprint.w, CityDate)

```

```

#2)
length(table(sprint.w$CityDate))

```

```
## [1] 921
```

```

#3)
whichmin <- function(data){
  return (data[which.min(data$Time),])
}
sprint.w.fastest <- lapply(split(sprint.w, sprint.w$CityDate),
  function(data) return (data[which.min(data$Time),]) )
sprint.w.fastest <- data.frame(t(sapply(sprint.w.fastest, rbind)))
nrow(sprint.w.fastest) == length(table(sprint.w$CityDate))

```

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

```
head(sprint.w.fastest,5)
```

```

##              Rank  Time Wind              Name Country
## Abidjan 26.04.2014 1395 11.06 -0.6 Murielle Ahour&eacute; CIV
## Abuja 02.07.2008    1696 11.08 +0.2      Damola Osayomi    NGR
## Ad-Dawah 05.10.2000 1853 11.09 +1.6      Chryste Gaines    USA
## Ad-Dawah 06.05.2016 60 10.8 +0.7      Tori Bowie        USA

```

##	Ad-Dawhah	07.05.1998	722	10.99	+0.7	Beverly McDonald	JAM
##			Birthdate		City	Date	CityDate
##	Abidjan	26.04.2014	23.08.87		Abidjan	26.04.2014	1
##	Abuja	02.07.2008	26.07.86		Abuja	02.07.2008	2
##	Ad-Dawhah	05.10.2000	14.09.70		Ad-Dawhah	05.10.2000	3
##	Ad-Dawhah	06.05.2016	27.08.90		Ad-Dawhah	06.05.2016	4
##	Ad-Dawhah	07.05.1998	15.02.70		Ad-Dawhah	07.05.1998	5