Race Times - The Cherry Blossom Run

" I certify that I have acknowledged any code that I used from any other person in the class, from Piazza or any Web site or book or other source. Any other work is my own. "

Zhewen Hu

#9124090355

Stat 242 Assignment 1

April 13rd

Introduction

In this assignment, I analyze the data from the Cherry Blossom 10-mile running race, http://cherryblossom.org. This dataset involves 24 individual files and each file contains information, such as name, ranks, hometown etc. By using R, I preprocess the data so that I can analyze the data easily. By the R output, I find some interesting results about the winners(1999-2010) and participants(1999-2010).

Data Manipulation

For this part, because 24 files have different data formats. I use the functions to check the header of the table to have a brief idea about the data formats. Then I use a function to deal with most of the files. For exceptions, like women10Mile2009 which has no header or file with encoding problem, I will deal with them individually. In the end, I organize 24 files into a big list. Each element of the list is a data frame. So I can use this big list to explore data more easily. Also, in the explanatory data analysis part, I modify the big list to analyze further.

Explanatory Data Analysis

In this part, I analyze the data from two aspects. One is the general analysis and the other is winner analysis.

1. General Analysis
   (1) Number of Participants
   One can see from the plot below, from 1999-2010, the number of participants increase a lot both for men and women. For men, the number of participants increases by two times. For women, the number of participants increases by three times. For one side, this shows that people are aware of the importance of health and love exercising more than before. For the other side, the organization of Cherry Run has promoted itself a lot by different ways. I have checked the official website, and I notice that Cherry Run has its own blog and lottery. By various ways, Cherry Run attracts more and more participants. One interesting thing is that female participants increases faster than male participants. Probably, women are more interested in taking part in long-distance running and the route of Cherry Run suits women better.

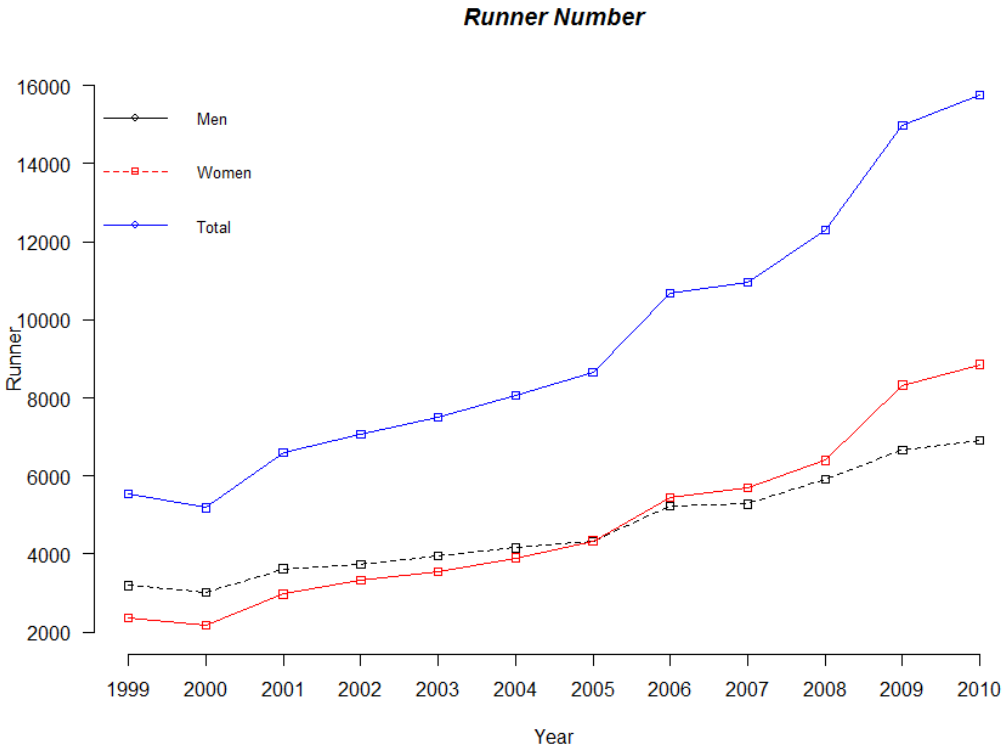

Fig.1

(2) Participants Age Distribution

One can see from the Fig.2 that the distribution is not normally distributed and it skewed a little bit. Mode occurs in the group of 25 – 30. And there are few participants in the group of 70-80 or 0-10. Apparently, the youth is the majority of participants. Also, the plot shows that Cherry Run does attract people from different age. Apparently, it is a big event for Washington D.C..

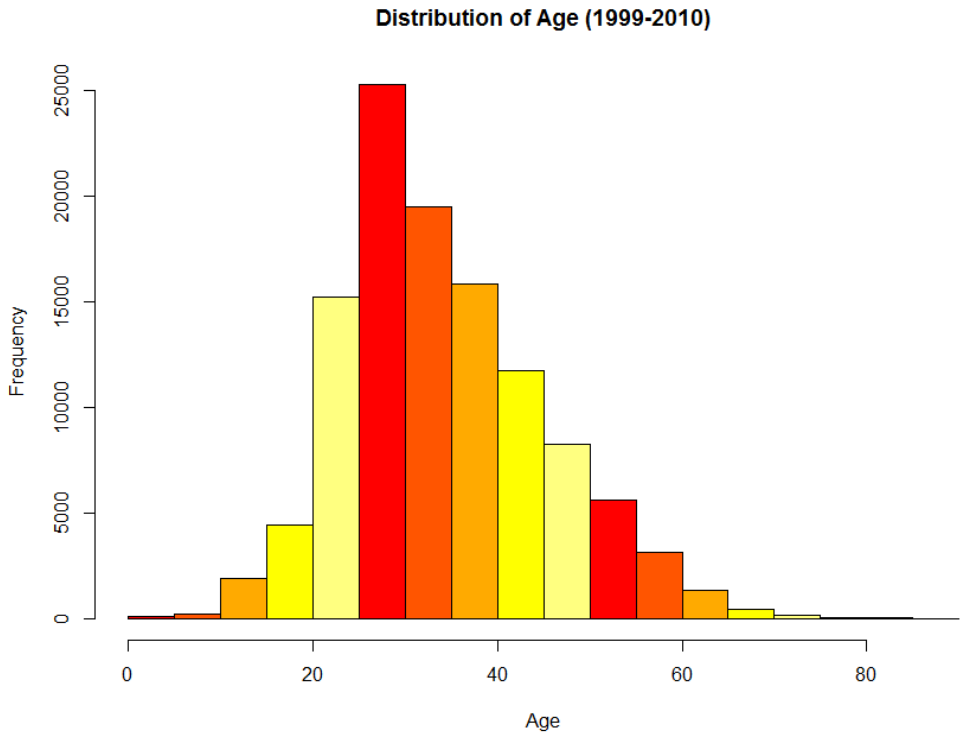

Fig.2

2. Winner Analysis

Because there are many participants in Cherry Run. I am more interested in exploring data about winners.

(1) Are winner the same?

One can see from the table 1 and table 2 that the winners are not the same. For men, Ridouane Harroufi won twice. For women, Lineth Chepkurui won three times. According to the official sites, though winners are not the same, the seeds people are similar.

Winner (Men)		
Worku Bikila	Reuben Cheruiyot	John KORIR
Rueben CHERUIYOT	John Korir	Nelson Kiplagat
John K Korir	Gilbert Okari	Tadesse Tola
Ridouane Harroufi	Ridouane Harroufi	Stephen Tum

Table 1

Winner(Women)		
Jane Omoro	Teresa Wanjiku	Elana MEYER
Luminita TALPOS	Olga Romanova	Isabella Ochichi
Nuta Olaru	Lidiya Grigoryeva	Teyba Erkesso
Lineth Chepkurui	Lineth Chepkurui	Lineth Chepkurui

Table 2

(2) Winner Age

One can clearly the age range of the winners is from 10 – 35. The mode is in the group of 20-25. From the biological side, people from age 20 to age 30 have the best physical condition and can achieve best score.

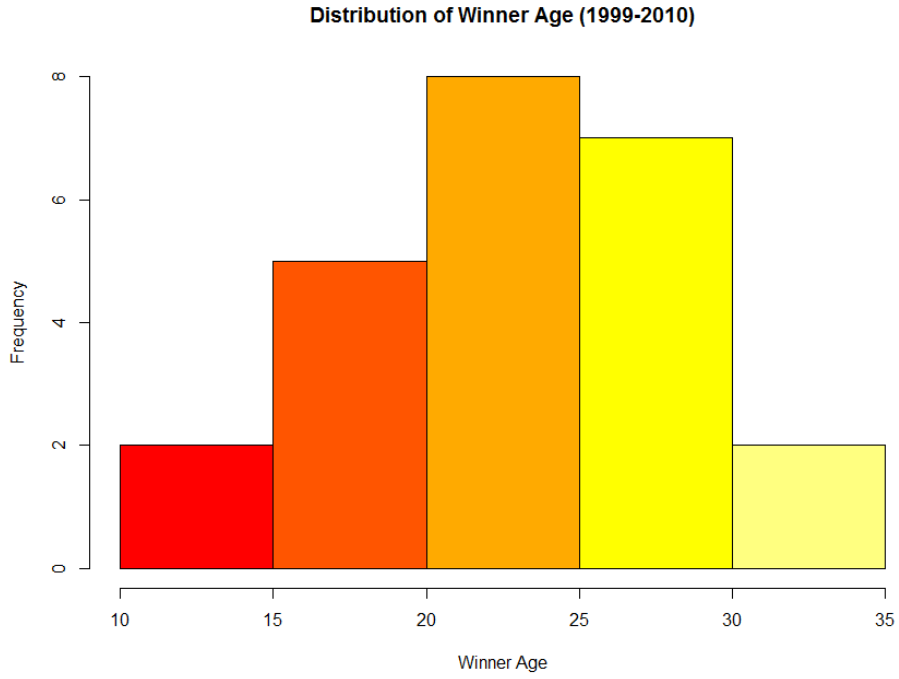

Fig.3

(3) Winner Hometown

From Fig. 4, one can see that most of the winners are from Kenya (As the matter of fact, I am not sure whether Ken also means Kenya). One can notice on the Olympic Games that African people are really good at long-distance running. Also, there is a science research paper reports that African people have an ideal body structure which suits running and can achieve good scores in the competition.

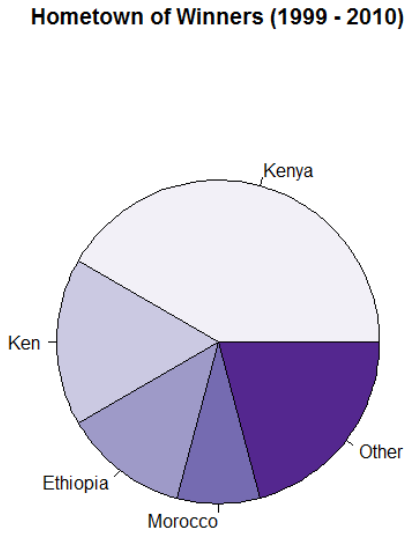

Fig. 4

(4) Winner Running Time

Winner Running Time fluctuates from 1999 – 2010. One can see that generally speaking, winners time decreases. Also, it is obvious that men run faster that women. Probably, the reason why winner time decrease is that people know how to train scientifically and people also gain some nutrition knowledge. Also, the shoes or running clothes they wear have improved. For example, the fabric is lighter and more comfortable. The reason why the data fluctuates is that the weather or some sudden occasions happened.

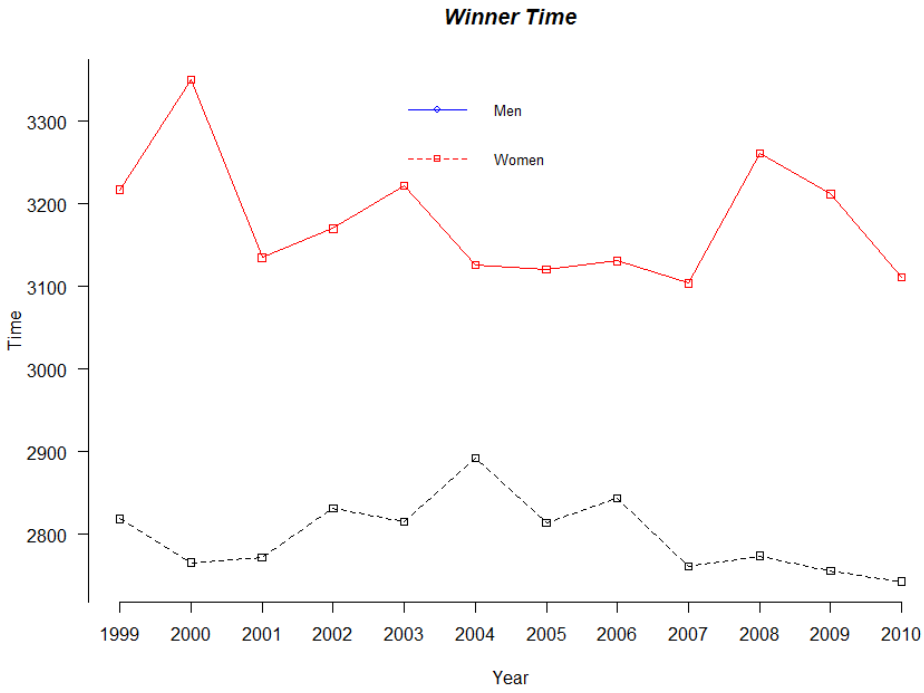

Fig.5

Discussion

Due to time limitation, I still did not analyze very thoroughly. For example, I have noticed that the organization of Cherry Run has set the system for seeds runners according to their historical data. I have spent too much time cleaning data. I should have analyze the people have the good performance from 1999 – 2010 instead of performance of winners. For example, the running time is faster than a certain number.

Also, I have noticed that Cherry Run has set up the lottery. I could do some predications about who will win in the following years. It is worth exploring as well.

Appendix

Appendix 1 (Data Manipulation)

```
library(stringr)

library(dplyr)

library(lubridate)

setwd("d:\\00 Davis\\03 2015 Spring Quarter\\Stat
242\\Data_Duncan\\stat242_2015\\Assignment1\\data")


## (1) Read all data into a big list

wholeList = lapply(list.files(), function(x)(readLines(x,encoding = "UTF-8")))


## (2) Detect "=" in the files

titleDetect = function (x){

  # x means the xth element in the wholeList(big list)

  wholeList = lapply(list.files(), readLines)

  wholeList[[15]] <- str_replace(wholeList[[15]], wholeList[[15]][18], wholeList[[3]][17])

  wholeList[[15]] <- str_replace(wholeList[[15]], wholeList[[15]][17], wholeList[[3]][16])

  grep("^=", wholeList[[x]])}

 # titleDect is assigned to check the =

   titleDect <- lapply(c(1:24), titleDetect)


## (3) try read.fwf to split the data according to fixed width

count = function (x) {

  # x means the xth files in the directory

  # use for counting the fixed width of the header

  # Add 1 to each width there is a blank after each cloumn

  nchar(x) +1

}


## (4) give the names
```

```
filenames = function(x) {list.files()[x]}

## (5) how many lines we should skip to get rid of the big title
skipLine = function (x) {titleDect[[x]]}

## (6) count width in the big data frame
widthC = function(x) {lapply(strsplit(wholeList[[x]][titleDect[[x]]]," "), count)}

## (7) colnames
cherryNames <- function (x) {
read.fwf(filenames(x), widthC(x), skip = skipLine(x)-2,
fill = TRUE,  na.strings = c("","NA"), comment.char = '',
 stringsAsFactors=FALSE, blank.lines.skip = TRUE,encoding = "UTF-8")
 }
cherryCol <- function (x) {tolower(cherryNames(x)[1,])}

## (8) read data by read.fwf
cherryBlossomT <- function (x) {
# x means the xth files in the directory
 read.fwf(filenames(x), widthC(x), skip = skipLine(x),
 col.names = cherryCol(x), check.names=FALSE, fill = TRUE,
 na.strings = c("","NA"), comment.char = '',
 blank.lines.skip = TRUE, encoding = "UTF-8")
 }

##(10) put into a big dataframe list
cherryRun <- lapply(c(1:24),cherryBlossomT)
cherryRun[[15]] <- w01 # exceptions
cherryRun[[11]] <- m09 # exceptions
```

```
## (11) exceptions


### Encoding solved

encoding09 <- readLines("men10Mile_2009", encoding = "UTF-8")

encoding09 <- gsub("[\u00A0]"," ", encoding09 )

widthEncoding <- lapply(strsplit(encoding09[grep("^=", encoding09)]," "),count)

m09 <- read.fwf(textConnection(encoding09),c( 6,12,7,23,3,21,8,7,2,6),

skip = skipLine(11),  col.names =  c("place","div/tot","num","name","ag","hometown","gun time","time","id","pace"),

check.names=FALSE, fill = TRUE,  na.strings = c("","NA"),

comment.char = '', blank.lines.skip = TRUE, encoding = "UTF-8")


### file without header name

w01 <- cherryBlossomT(15)

colnames(w01) <- cherryCol(3)


### some files have the incorrect "="

cherryNamesEX <- function (x) { read.fwf(filenames(x),

c(6,9,7,23,3,16,8,8,1,6,2), skip = skipLine(x)-2, fill = TRUE,

na.strings = c("","NA"), comment.char = '', stringsAsFactors=FALSE,

blank.lines.skip = TRUE,encoding = "UTF-8")}

cherryColEX <- function (x) {tolower(cherryNamesEX(x)[1,])}

cherryRun[[8]]  <- read.fwf(filenames(8), c(6,9,7,23,3,16,8,8,1,6,2),

skip = skipLine(8), col.names = cherryColEX(8), check.names=FALSE,

fill = TRUE,  na.strings = c("","NA"), comment.char = '',

blank.lines.skip = TRUE, encoding = "UTF-8")

cherryRun[[20]] <- read.fwf(filenames(20), c(6,9,7,23,3,16,8,8,1,6,2),

skip = skipLine(20), col.names = cherryColEX(20), check.names=FALSE,

fill = TRUE,  na.strings = c("","NA"), comment.char = '',

blank.lines.skip = TRUE, encoding = "UTF-8")
```

```
## remove the irrelevent symbols like # in time variable

cherryRun[[9]] <- read.fwf(filenames(9), c( 6,12,7,23,3,19,7,1,1,6,2,8),

 skip = skipLine(9),

 col.names =
c("place","div/tot","num","name","ag","hometown","time","id","na","pace","s","split"),

 check.names=FALSE, fill = TRUE,  na.strings = c("","NA"), comment.char = '', blank.lines.skip
= TRUE, encoding = "UTF-8")

cherryRun[[21]] <- read.fwf(filenames(21), c( 6,12,7,23,3,19,7,1,1,6,2,8),

skip = skipLine(21), col.names =
c("place","div/tot","num","name","ag","hometown","time","id","na","pace","s","split"),

check.names=FALSE, fill = TRUE,  na.strings = c("","NA"), comment.char = '', blank.lines.skip
= TRUE, encoding = "UTF-8")


cherryRun[[23]] <- read.fwf(filenames(23), c( 6,12,7,23,3,21,8,7,2,6),

skip = skipLine(23), col.names = c("place","div/tot","num","name","ag","hometown","gun
time","time","id","pace"),

check.names=FALSE, fill = TRUE,  na.strings = c("","NA"), comment.char = '', blank.lines.skip
= TRUE, encoding = "UTF-8")


cherryRun[[12]]<-  read.fwf(filenames(12),widthC(12),

skip = skipLine(12), col.names =  c("place","div/tot","num","name","ag","hometown","5
mile","time","net time","na","pace"),

check.names=FALSE, fill = TRUE,  na.strings = c("","NA"), comment.char = '', blank.lines.skip
= TRUE, encoding = "UTF-8")


cherryRun[[24]]<-  read.fwf(filenames(24),

c( 6,12,7,23,3,21,8,8,7,1,1,6,2), skip = skipLine(24),

col.names = c("place","div/tot","num","name","ag","hometown","5 mile","gun
time","time","id","na","pace","s"), check.names=FALSE,

 fill = TRUE,  na.strings = c("","NA"), comment.char = '', blank.lines.skip = TRUE, encoding =
"UTF-8")
```

Appendix 2

```
# (1)General Analysis

## Age

Ageall <- function(x) {select(cherryRun[[x]],contains("ag"))}

allAge <- as.numeric(unlist(sapply(c(1:24), Ageall)))

names(allAge) <- NULL

hist(allAge, main = "Distribution of Age (1999-2010)",

xlab = "Age", col = heat.colors(5))


## Number of Runner

runner <- function (x) {nrow(cherryRun[[x]])}

runnerTotal = runnerM + runnerF

runnerM = unlist(lapply(c(1:12), runner))

runnerF = unlist(lapply(c(13:24), runner))

plot(runnerM,type = "o", axes = FALSE, pch = 22, lty = 2,

ylim = c(2000,16000), ylab = 'Runner', xlab = "Year")

lines(runnerF,type = "o", pch =22, col = 'red')

lines(runnerTotal, type = "o", pch = 22, lty = 1, col = "blue")

title(main = "Runner Number", font.main = 4)

axis(1, at = 1:12, lab = as.character(c(1999:2010)))

axis(2,  las = 1, at = c(2000,4000,6000,8000,10000,12000,14000,16000))

legend("topleft", xjust = 0,c("Men","Women","Total"),

cex=0.8,col=c("black","red","blue"), pch = 21:22, lty = 1:2, bty = "n")

(runnerM[12]-runnerM[1])/runnerM[1]

(runnerF[12]-runnerF[1])/runnerF[1]

(runnerTotal[12]-runnerTotal[1])/runnerTotal[1]


# (2) Rank one problem

placeOne <- function(x){cherryRun[[x]][1,]}

rankOne <- lapply(c(1:24), placeOne)
```

```
## Person Name
wName<-sapply(c(1:24),function(x){select(rankOne[[x]],contains("name"))})
names(wName)<- NULL
## Hometown
Home <- function(x) {select(rankOne[[x]],contains("hometown"))}
winnerHome <- as.character(unlist(sapply(c(1:24), Home)))
Kenya <- sum(str_count(winnerHome, "Kenya\\s*"))/24
Ken <- sum(str_count(winnerHome,"[K[Ee][Nn]\\s]*"))/24
Ethiopia<- sum(str_count(winnerHome, winnerHome[1]))/24
Morocco <- sum(str_count(winnerHome, winnerHome[10]))/24
Others <- (24 - kenya - ken - ethiopia - morocco)/24
pie(c(Kenya,Ken,Ethiopia,Morocco,Others),
labels = c("Kenya","Ken","Ethiopia","Morocco","Other"),
col = brewer.pal(5,"Purples"), main = "Hometown of Winners (1999 - 2010)")
## Age
Age <- function(x) {select(rankOne[[x]],contains("ag"))}
winnerAge <- as.numeric(unlist(sapply(c(1:24), Age)))
names(winnerAge) <- NULL
hist(winnerAge, main = "Distribution of Winner Age (1999-2010)",
 xlab = "Winner Age", col = heat.colors(5))
## Time
Wtime <- function(x) {select(rankOne[[x]],matches("^time|^net|^ net tim\\s*$"))}
winnerTime <- lapply(c(1:24), Wtime)
timeConvert <- function(x){period_to_seconds(ms(unlist(winnerTime[[x]])))}
winnerTime[[12]] <- select(rankOne[[12]], matches("^time"))
winnerSecond <- as.numeric(sapply(c(1:24), timeConvert))
year <- as.numeric(1999:2010)
plot(winnerSecond[1:12],type = "o", axes = FALSE, pch = 22, lty = 2, ylim =
range(winnerSecond), ylab = 'Time', xlab = "Year")
lines(winnerSecond[13:24],type = "o", pch =22, col = 'red')
```

```
title(main = "Winner Time", font.main = 4)

axis(1, at = 1:12, lab = as.character(c(1999:2010)))

axis(2,  las = 1, at = c(2500,2600,2700,2800,2900,3000,3100,3200,3300,3400,3500,3600))

legend("top", xjust = 0,c("Men","Women") , cex=0.8,col=c("blue","red"), pch = 21:22, lty = 1:2,
bty = "n")
```