Data Analysis on Cherry Blossom 10 Mile Run

Hong Fan

1. Introduction

Cherry Blossom Run is an annual event organized on April in Washington, D.C. The races include 10 Mile Run, 5K Run-Walk and 1/2 Mile Kids' Run. Starting from 1973, the cherry blossom races have attracted people from all over the world. The data are results from the Cherry Blossom 10-mile running race collected from 1999 to 2010 for both men and women participants.

2. Data reading and transformations

There are 24 files in total. The first 12 files contain male participants' racing records separated by years from 1999 to 2010. The second 12 files include female participants' racing information separated the same way. Since categories vary from file to file, some names are adjusted to make them consistent between files. In order to analyze them as a whole and make comparisons, place, name, age, hometown, gun time (official time) are selected as common columns and year, gender, birth and identifier are added to the final combined data frame resulting in 9 columns in total. In addition, there are 239 observations which are incomplete (information in certain columns are not given). These incomplete observations are removed from the whole data. Time is transformed into seconds.

3. Data analysis

3.1 Summary of Time spent on 10 Mile Run

After cleaning the data, there are 112951 observations in total from 1999 to 2010. In Figure 1, the left boxplot shows that the distribution of participants' running time from 1999 to 2010 is approximately normal with mean value 5807.

The plot in the right shows the total number of participants in each year and corresponding gender proportions. Starting from 1999, proportion of male participants is greater than female participants, but the gap is vanishing gradually. From 2009, the total number of women runners in Cherry Blossom 10 Mile Run exceeds men runners.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2743	5058	5758	5807	6490	10650

Table 1 Summary of Running Time

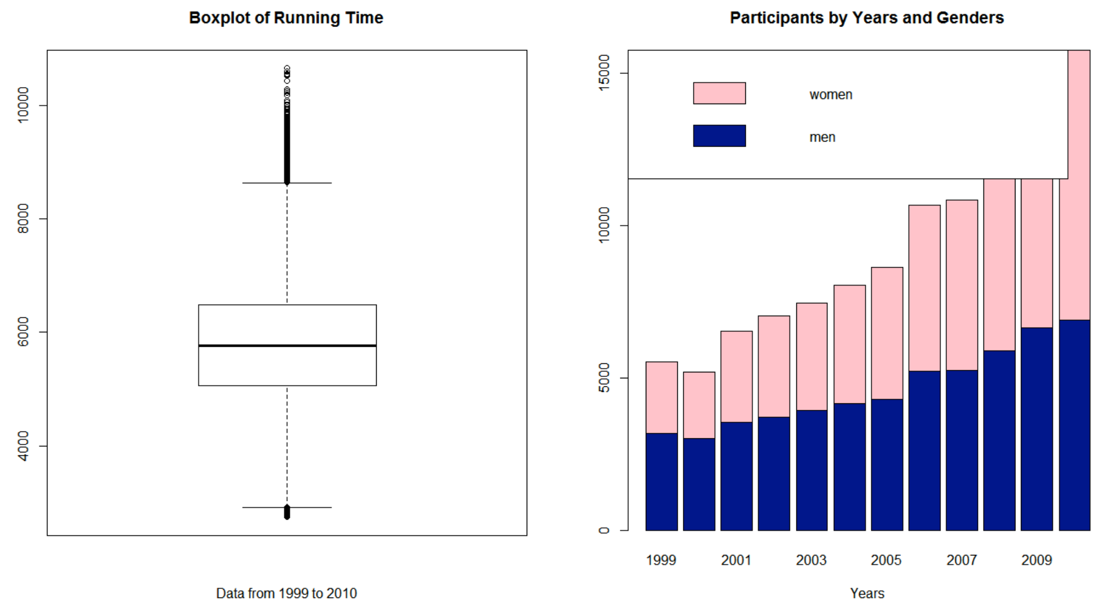

Figure 1 Left: Boxplot of Runing time. Right: Stacked plot of Participants by Years and Genders

3.2 Comparison between genders

In this section, we make comparisons between male and female participants and density plot is used to show the overall distribution of total time spent on racing in these two groups. In Figure 2 below, time distributions for both groups are close to normal but slightly right-skewed. Interestingly, these two distributions are very similar and are consistent with the distribution indicated by the boxplot above. Table 2 shows the corresponding summary for each group.

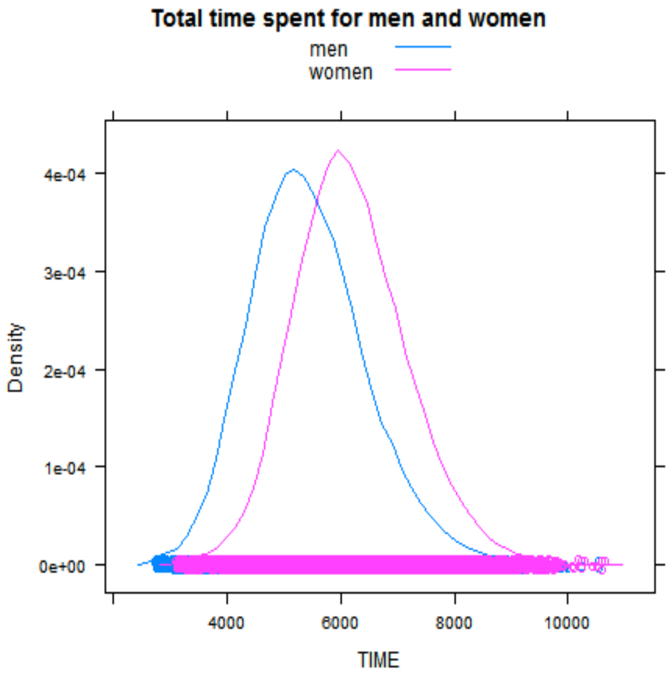

Figure 2 Density Plot of Time Spent on 10 Mile Racing: Men and Women

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Male	2743	4719	5354	5433	6062	10550
Female	3104	5488	6108	6171	6803	10650

Table 2 Summary of Time for Male and Female Groups

In order to make understanding the data, we make comparisons between these groups year by year. Figure 3 below shows the comparisons of overall and average ages for men and women runners. The average age for men runners in each year is about 5 years older than that for women runners, but the average age gap is slightly decreasing. Interestingly, the top of Figure 3 shows most of runners whose ages are above 60 are males, which contribute to greater average age for male runners.

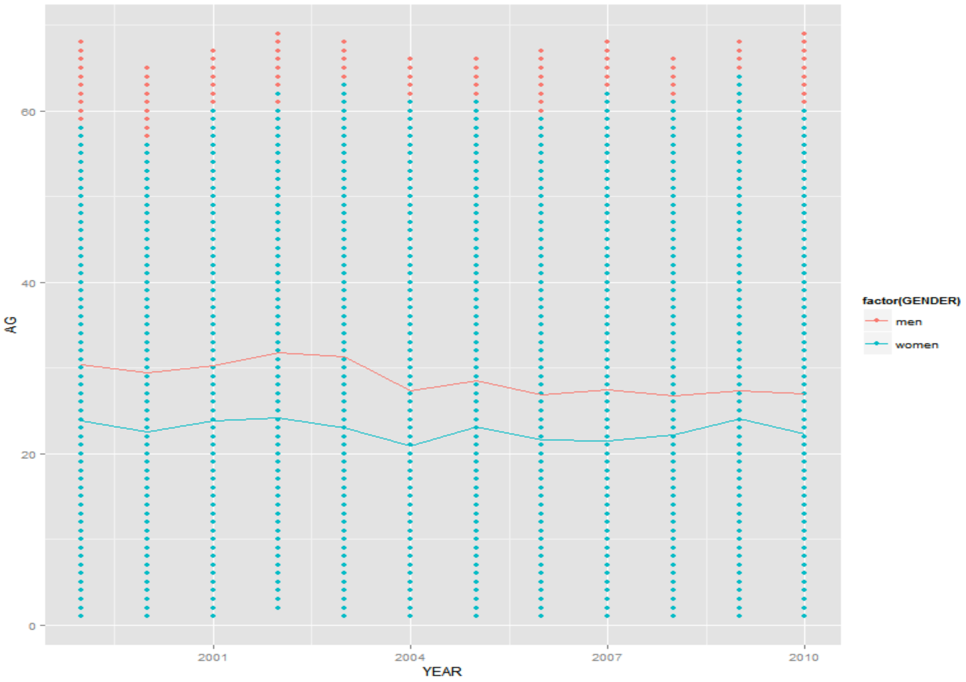

Figure 3 Age against Year for Men and Women runners

The bottom of Figure 4 shows that most of the top participants in each year are dominated by men players. It is clear to see the average racing time for men group (denoted by pink line) is approximately 750 seconds (12.5 min) faster than the average racing time for women group (denoted by blue line). Interestingly, the two lines almost coincide with each other if the blue line is shifted downward.

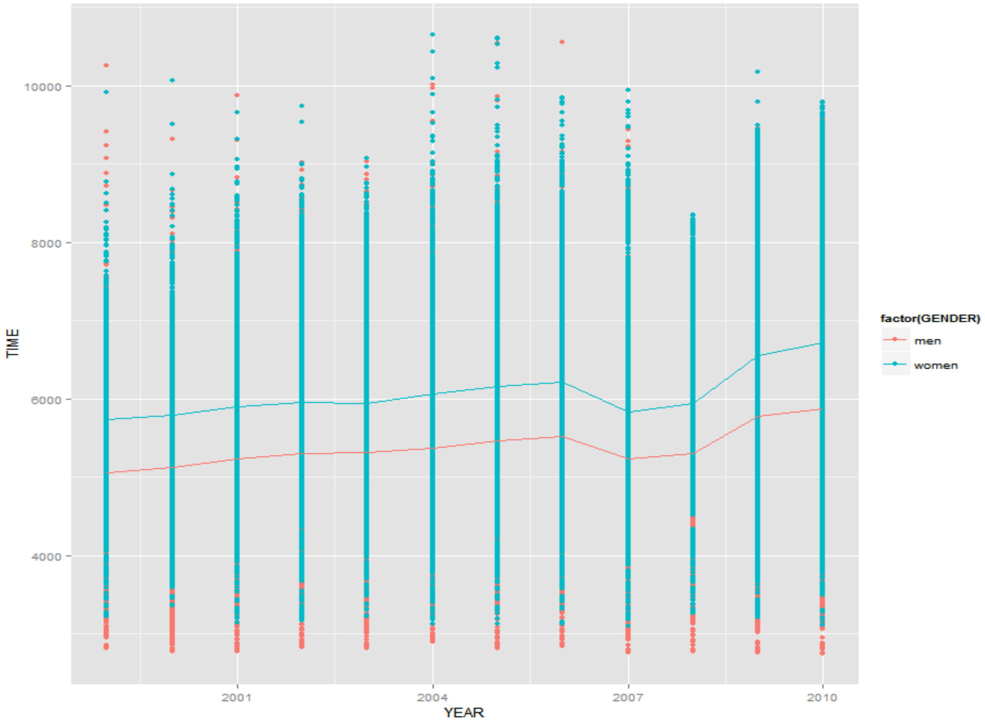

Figure 4 Racing Time against Years

In general, both lines have increasing trends with time going by, but decreased at year 2007 and slightly increased in year 2008, indicating better overall performance for the participants in 2007 and 2008. However, the reason behind this is not clear. When we zoom in at the top of this figure, it is found that larger time values in 2007 are not as scattered as previous years. After checking the 239 incomplete cases cleaned from the whole data, it is found the last 22 men participants and last 88 women participants (ranked by running time) in 2007 are removed since corresponding time information are missing. This fact strongly supports Figure 3 in which the average time for both men and women went down. In 2008, 34 observations which has relatively large time values are removed which may result in the decrease of the average time.

3.3 Comparison between US participants and Non-US participants:

In this section, we analyze the data by dividing them into two groups – US runners and non-US runners. Due to the fact that there is no column clearly specifying whether each participant is domestic. We use the U.S. state name obtained in the HOMETOWN column to separate people from other non-US countries. However, this approach does not work for data in 2006. 2006 data, for both men and women, rarely contain U.S. states (most of the information included in HOMETOWN only recorded city names.), which makes itself difficult to be identified.

The following table shows the number of US participants and Non-US participants in each year which are identified by the approach mentioned above.

	1999	2000	2001	2002	2003	2004
Non-US	25	49	52	45	36	64
US	5518	5133	6479	7002	7434	7985
Total	5543	5182	6531	7047	7470	8049
Proportion*	0.45%	0.95%	0.80%	0.60%	0.48%	0.80%
	2005	2006	2007	2008	2009	2010
Non-US	59	8213	60	56	79	87
US	8579	2449	10784	12212	14888	15663
Total	8638	10662	10844	12268	14967	15750
Proportion*	0.68%	Ignore	0.55%	0.46%	0.53%	0.55%

Table 3 Non-US and US Participants over Years

Since 2006 data do not contain state names, the results (in grey) are quite different from other years (shown in Table 3), thus 2006 is ignored in this section. Although the number of Non-US participants has the increasing trend which increases to 87 in 2010, their proportions still within the 1% of the total population.

Figure 5 is a stack plot used to show comparisons of proportion for US and Non-US participants. It is clear to see that total participants are increasing over years and in 2009, it becomes more than 15000, but the proportion of Non-US runners remains slim.

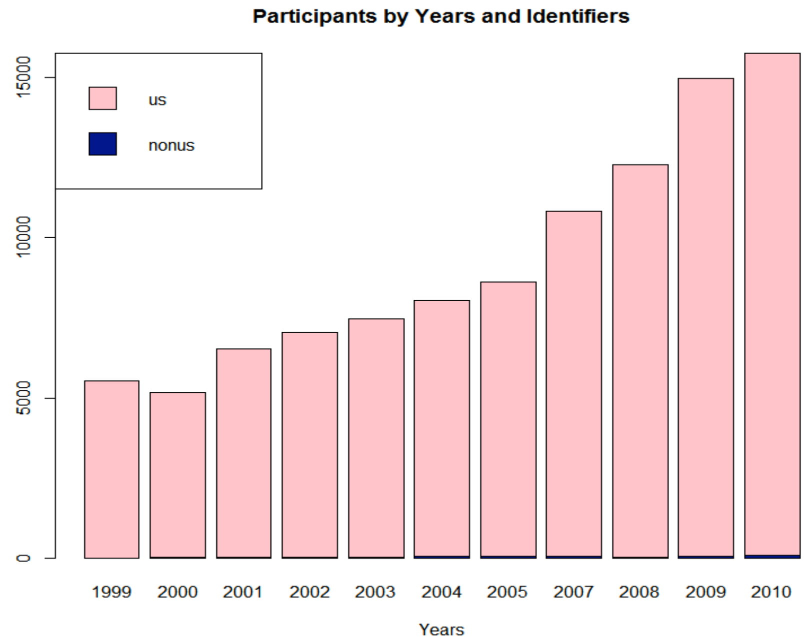

Figure 5 Stacked Plot of Non-US and US Participants over Years

3.3 Regarding Matching Participants Who Attended 10 Mile Run in Different Years

In this section, we want to extract participants who attended this running event more than once. Based on the given information in the data, we use name, hometown and birth to identify participants. However, there are many problems. For example, a women participant whose name is Lineth Chepkurui attended the running event from 2008 to 2010 — three times in total, but her age is 21 at 2009 and 23 at 2010, resulting in different values in birth — 1988 and 1987 respectively.

In addition, the way that recorded data in hometown vary from year to year which makes the matching process difficult. In order to have relatively accurate matching, uniform way to measure each variable over years is necessary.

4. Discussion

In this report, data from The Cherry Blossom 10 Mile Run is analyzed in different ways. We first check the distribution of the overall running time and make comparisons regarding population proportions between female and male runners. In order to better understand the data, we compare overall ages and average ages against years between male and female participants. After these, we compare overall running time and average running time in each year between these two groups. In this section removed data are used to explain the decrease of average time for both groups in 2007 and 2008. After making above comparisons, we divide data into two groups- US and Non-US participants and draw stacked plot to show the proportion of both groups in each year's total number of runners and find out the total number of non-US runners are increasing but not as fast as domestic runners and the total number of non-US participants is within 1% of the total runners. In the last section, we try to match runners who attended the event more than once. However, since some variables are recorded in different ways and some involves errors, it is hard to matching people accurately. Therefore, uniform way to measure each variable over years is suggested.

In the data, there are 239 incomplete cases out of 113190, which may result in different or counter-fact inferences since some missing information may direct the inferences in different ways.

Appendix1:

```r
# Set the working directory
setwd("C:/Users/Administrator/stat242_2015/Assignment1/data")
filenames = list.files()
n = length(filenames)

header.m = function(filename){            #headerline manipulation function :
the total number of characters in this line remains the same
  lines = readLines(filename, encoding = "UTF-8")
  lines.new = gsub("[\u00A0]", " ", lines)
  header = grep("^place", lines.new, value = TRUE, ignore.case = TRUE)
  header = toupper(header)
  header = gsub("([A-Z]+)\\sTIM", replacement = "\\1-TIM", header) ##change GUN
TIM into GUN-TIME and NET TIM into NET-TIME but still keeps its previous number
of characters. will deal with DIV  /TOT and DIV /TOT in the following functions!!
(specify this when finishing)
  header = gsub('([0-9]+)\\s([A-Z]+)', '\\1-\\2', header) ##deal with 5 Mi and
10 Km
  header = gsub("MILE", "MI  ", header)
  return(header)
}

headerline = sapply(filenames, header.m)
```

```r
for (i in 1: n) {                       ###deal with files which do not have h
eaders using the fact that the first half of the files include men's records
 from 1999 t0 2010 and the second half includes women's data in correspondin
g years
  if (identical (headerline[[i]], character(0)) && i > n/2) {
    headerline[[i]] = headerline[[i-n/2]]
  } else if (identical (headerline[[i]], character(0)) && i<= n/2){
    headerline[[i]] = headerline[[i+n/2]]
  } else {
    headerline[[i]] = headerline[[i]]}
}

timeline = function (filename){     #index for lines that contain time
  line <- readLines(filename, encoding = "UTF-8")
  line.new = gsub("[\u00A0]", " ", line)
  index.t = grep("[0-9]+:[0-9]+", line.new)
  return(index.t)
}
```

```r
width.f = function (filename){
  ind = match(filename, filenames)
  headerstring = headerline[[ind]]
  headersplit = strsplit(headerstring, split= "\\s([0-9]|[A-Z])") #split the header to get width
  width = as.numeric(sapply(headersplit, nchar))
  l = length(width)
  width.new = width + c(1, rep(2, l-2), 6) #last string's width does not affect other strings' widths; for files women10Mile_2001 and men10Mile_2001, there is no space after last column name "GUN", so 4 is added in case # or * are contained.
  return(width.new)
}

width = lapply(filenames, width.f)

header.f = function (filename){        # function of getting names by using headerline
  ind = match(filename, filenames)
  headerline[[ind]] = gsub("\\s+/", replacement = "/", headerline[[ind]]) ##get rid of the space in DIV /TOT
  header = gsub("GUN|GUN-TIM", "TIME", headerline[[ind]])  #TIME equals to GUN TIME  (reference in the website)
  header = strsplit(header, split = "\\s+")
  header.n = unlist(header)
  return(header.n)
}

header = sapply(filenames, header.f)

readdat = function (filename, indext = timeline(filename)){  #data manipulation function
    ind = match(filename, filenames)
    dat = read.fwf(filename, widths = width[[ind]], encoding = "UTF-8", skip=indext[1] - 1, comment.char="", stringsAsFactors = FALSE, strip.white = TRUE)
    dat.new = data.frame (gsub("[\u00A0]|\\#|\\*", "", as.matrix(dat)))
    time = indext - (indext[1]-1)
    dat.new = dat.new[time, ] #obtain rows that only contain time
    names(dat.new) = header[[ind]]
    names(dat.new)[names(dat.new)=="GUN"] <- "TIME"
    return(dat.new)
}
```

```
Appendix 2
data = lapply(filenames, readdat)

time.c = function (x) {              #Time conversion function quoted from TA Nick's postings on piazza.
  time = strsplit(as.character('x'), ':')[[1]]
  conv = 60 ^ seq.int(length(time) - 1, 0)
  time = sum(conv * as.integer(time))
  return(time)
}

usaidentifier = function(x){   #identify whether the participant is from us
  pos = gregexpr('[A-Z]{2}', x)[[1]]
  string = substring(x, pos, pos+1)
  if (grepl("usa|united states|washington DC", x, ignore.case = TRUE)) { #deal with several special cases
    "us"
  } else if (!is.na(match(string, state.abb))| !is.na(match(x, state.name))) {
    "us"
  } else {
    "nonus"
  }
}   #this function cannot accurately identify us and non us players in 2006 since most Hometowns only contain city names

data.m = function (filename, time = time.c(x), id = usaidentifier(x),dat = readdat(filename, indext = timeline(filename))){
    ind = match(filename, filenames)
    dat[dat==""]  <- NA     #replace blank entries by NAs
    pattern = "([A-z]+)10Mile_([0-9]+)"
    gender = gsub(pattern, "\\1", filename)
    year = gsub(pattern, "\\2", filename)
    dat$YEAR = rep(year, nrow(dat))
    dat$GENDER = rep(gender, nrow(dat))
    dat[, c("YEAR", "AG")] = sapply(dat[, c("YEAR", "AG")], as.numeric)
    dat[, c("HOMETOWN", "NAME", "PLACE")] = sapply(dat[, c("HOMETOWN", "NAME", "PLACE")], as.character)
    dat$BIRTH = dat[, "YEAR"] - dat[, "AG"]
    dat$TIME = sapply(dat$TIME, time)
    tmp=lapply(dat$HOMETOWN, usaidentifier)
    tmp=unlist(tmp)
    dat$IDENTIFIER = tmp
    dat = dat[, c("PLACE", "NAME", "AG", "HOMETOWN", "TIME", "YEAR", "GENDER", "BIRTH", "IDENTIFIER")]  ###Obtain common columns in the whole data frame
s.
    return(dat)
}

dat = lapply(filenames, data.m)

data_w = do.call("rbind", dat)   #combine 24 data frames into 1

tt= complete.cases(data_w)
table(tt) #frequency table

data_t = data_w [which (tt == TRUE),] # remove 239 incomplete cases
data_f = data_w [which (tt == FALSE),] # incomplete cases
data_f [which(data_f$YEAR== 2008), ] # check removed cases in 2007

par(mfrow = c(1,2))

library("ggplot2")

boxplot(data_t$TIME, main = "Boxplot of Running Time", xlab = "Data from 199
9 to 2010")
counts <- table(data_t$GENDER, data_t$YEAR)
barplot(counts, main="Participants by Years and Genders",
        xlab="Years", col=c("darkblue","pink"),
        legend = rownames(counts), args.legend = list(x="topleft"))



#Men and Women group comparison
plot1 <- qplot(YEAR, TIME, data = data_t) + aes(colour = factor(GENDER)) + s
tat_summary(fun.y = mean, geom="line")
plot2 <- qplot(YEAR, AG, data = data_t) + aes(colour = factor(GENDER)) + sta
t_summary(fun.y = mean, geom="line")

#Comparison between mean time for both genders from 1999 to 2010
by(data_t[, "TIME"], data_t[, c("GENDER", "YEAR")], mean, na.rm = TRUE)


dat.x = data_t[-which(data_t$YEAR==2006), ]
counts <- table(dat.x$IDENTIFIER, dat.x$YEAR)
barplot(counts, main="Participants by Years and Identifiers",
        xlab="Years", col=c("darkblue","pink"),
        legend = rownames(counts), args.legend = list(x="topleft"))
```

```
#participant matching over years
nm <- c("NAME", "HOMETOWN")
res <- do.call(rbind, lapply(split(data_t, as.character(interaction(data_t[,
nm]))),function(x) {
      x[duplicated(x[, nm]) | duplicated(x[, nm], fromLast = TRUE), ]}))
nm1 <- c("NAME", "HOMETOWN", "BIRTH")
res1 <- do.call(rbind, lapply(split(data_t, as.character(interaction(data_t[,
nm1]))),function(x) {
  x[duplicated(x[, nm1]) | duplicated(x[, nm1], fromLast = TRUE), ]}))
```