Cherry Blossom Race Results Analysis

Laura Lazarescou, John Rodgers, Maysam Mansor and Mel Schwan

February 1, 2021

1 Introduction

Websites provide a tremendous amount of data on any topic that would interest you. Web scraping is an automated method used to extract large amounts of data from websites. The data on the websites are unstructured. Web scraping helps collect these unstructured data and store it in a structured form. One example of the many annual road races is the Cherry Blossom Ten Mile Run held in Washington D.C. in early April, when the cherry trees are typically in bloom. The Cherry Blossom Race started in 1973 as a training run for elite runners planning to compete in the Boston Marathon. Nearly 17,000 runners ranging in age from 9 to 89 have participated. Our team objectives are to scrape the race results for men and women from 1999 to 2012. We will do exploratory data analysis to discover any unique dynamics that exist in the race results.

2 Methods

2.1 Data Collection and Cleaning

In this case study, we could not scrape race data from the Cherry Blossom website since it recently changed from pure HTML format to a form query of a backend database. Fortunately, we were able to discover an archive of the website that contained the original HTML results pages. The "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org" site includes the race results, including the HTML parsing errors. We use several approaches to extracting this data into text files. We have extracted the men and female runners data for the years 1999 to 2012

We developed the following approach for web scraping the race results for the men and women. This approach downloads the HTML of each URL race result location into a temporary file. This file is then loaded into an array containing each HTML line of code. The scraping function then looks for the HTML tag for the beginning of the table ("

"), which indicates the end of the table of race results. The function then extracts the content between the table start and end array index. This extracted content is stripped of any incidental HTML tags and saved to the text file with names representing the year within the sex labeled directory.

This process worked with every year of racing except for 2009 results. These HTML pages were malformed and could not be scraped. For the 2009 dataset, we manually extracting that year's results.

In addition to managing the 2009 exceptions, we found that the 2001 txt file lacked a header so we added one using another file as a template. We also edited both the men's and the women's 2006 files to correct the header which lacked a separator between two columns.

2.2 Data Visualization

Using boxplots, Q-Q plots and scatterplots with linear regression curves, we were able to visualize the different dimensions of the data.

Once the data was in a clean structure, we found that Cherry Blossom Race participants are a broad distribution of runners between the ages of 9 and 89. The average age of runners in the ten-mile race between 1999 and 2012 was 38.6 years for men, compared with 33.9 years for women. As we saw in the boxplots, the average age of both men and women is decreasing over time. Younger runners are being drawn to the race as time is passing. This may also be reflective of the current demographics as millenials become a greater percent of the active population.

Comparing the average run times across the 14-year period between men and women, we see that men complete the race in 87.54 minutes vs. women's average rate of 98.16 minutes. We did not visualize this value, we calculated it to provide an average across all boxplots for all years.

We evaluated the normality of the women's age and run time data with Q-Q plots to see if there was excessive variance. In most years the Q-Q plots were close to linear, which indicates normality. We also know that a significant sample size allows us to assume normality of the sample, and we have more than 3000 runners in each race so these tests are more symbolic than necessary.

We also looked at the linearity between age and run time for women, and we found that there is a slightly positive relationship - younger women tend to have lower run times, while older women may have higher run times. However, there was so much variance across the age span, that one would be wise to seek additional factors if they were trying to predict run time based on gender and age.

```
##  [1] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/1999/cb99f.html
##  [2] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/Cb003f.htm"
##  [3] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2001/oof_f.html
##  [4] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2002/ooff.htm"
##  [5] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2003/CB03-f.HTM
##  [6] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2004/women.htm"
##  [7] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2005/CB05-f.htm
##  [8] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2006/women.htm"
##  [9] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2007/women.htm"
## [10] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2008/women.htm"
## [11] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2009/2009cucb-F
## [12] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2010/2010cucb10
## [13] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2011/2011cucb10
## [14] "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/results/2012/2012cucb10
```

2.2.1 Comparing Age Between Women and Men by Year

As we see in the plots below, men are older than women on average each year, and the mean age of both men and women decreased from 1999 to 2012.

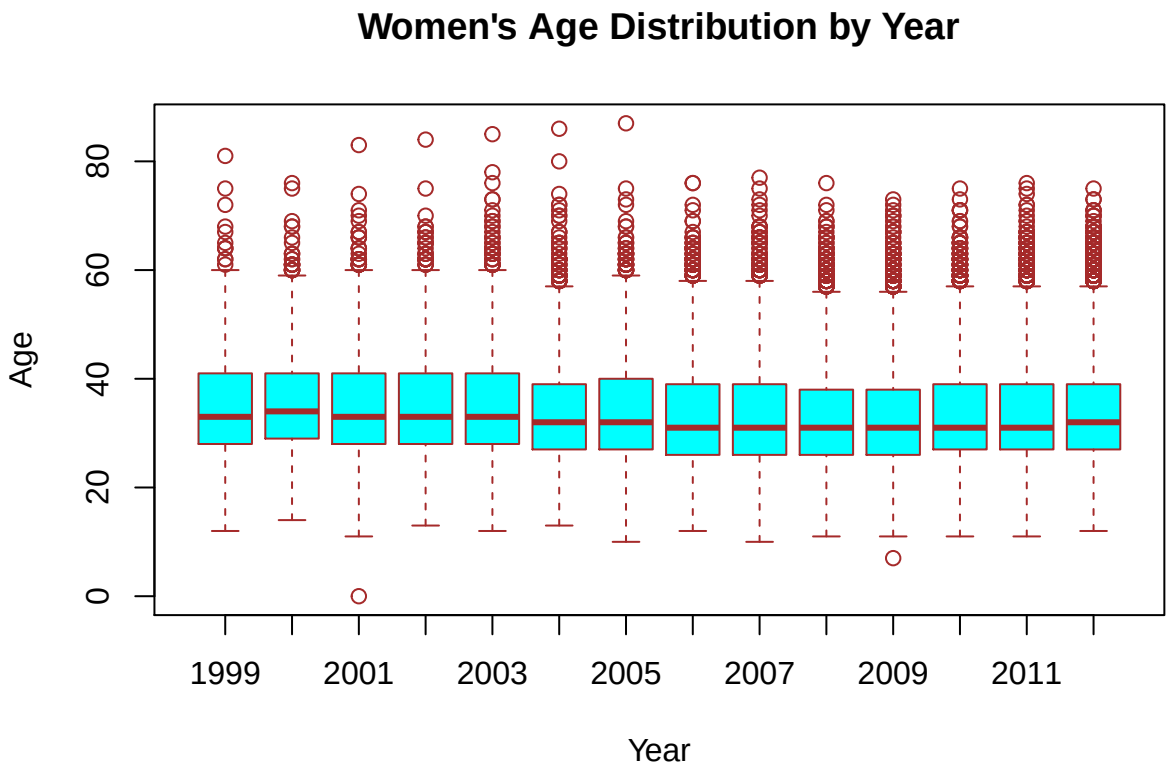


Fig.1−Women's Age Distribution by Year

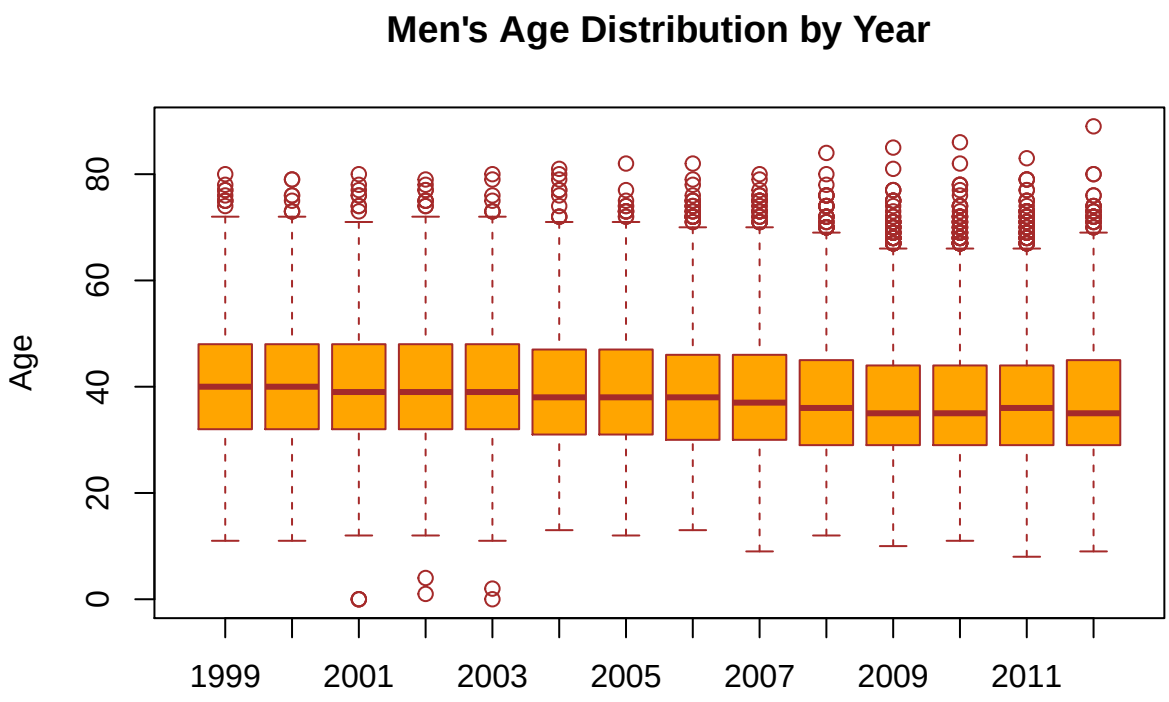
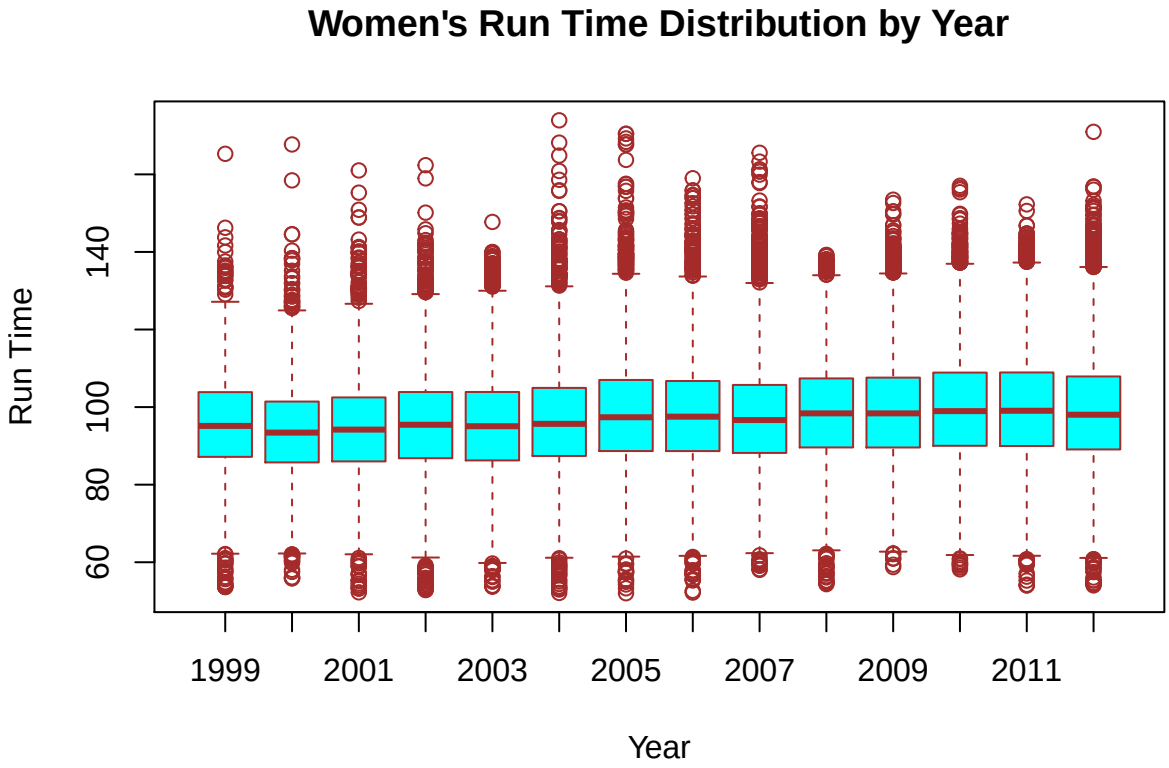


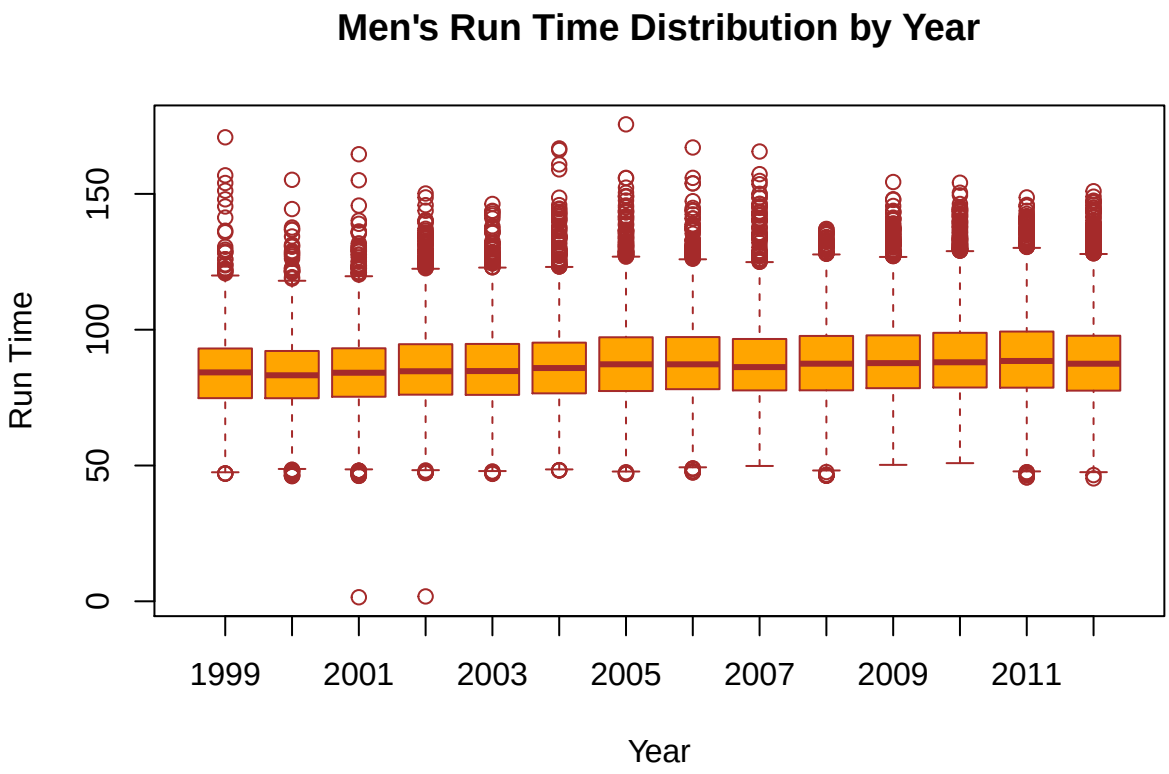
Fig.2−Men's Age Distribution by Year

2.2.2 Comparing Run Time Between Women and Men by Year

Men are faster and average run time is decreasing year over year for both groups. This may indicate that modern fitness routines are more effective. It would be worthwhile to look at trends by age group before we assume that all participants are running faster as time progresses.



Fig−3−Women's Run Time Distribution by Year.



Fig−4−Men's Run Time Distribution by Year.

2.2.3 Mean Ages and Run Times Across all Years

```
## [1] "Average Age of Women Runners = "
```

```
## [1] 33.85235
```

```
## [1] "Average Age of Men Runners = "
```

```
## [1] 38.61034
```

```
## [1] "Average Run Time of Women Runners = "
```

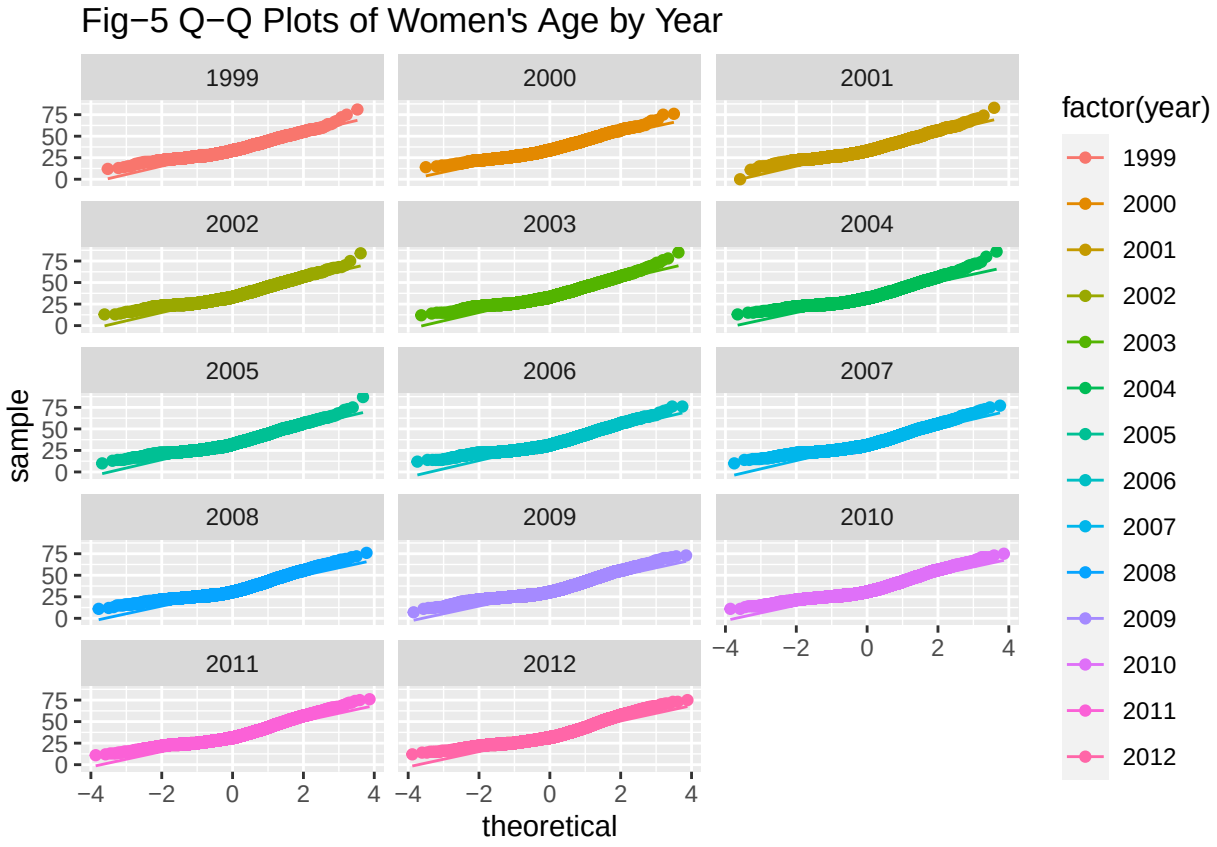
```
## [1] 98.16419
```

```
## [1] "Average Run Time of Men Runners = "
```

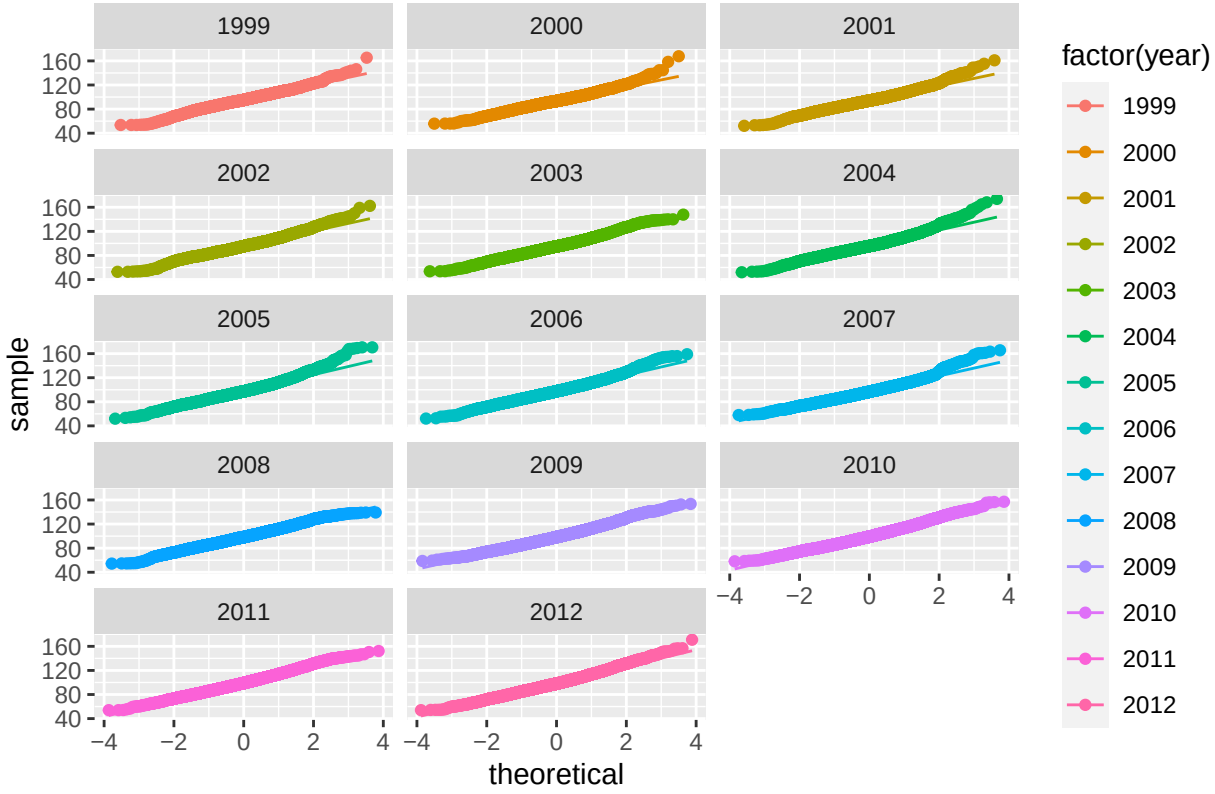
```
## [1] 87.54202
```

2.2.4 Evaluating Normality and Correlation Between Age and Run Time

The Q-Q plots for Women's Age are close to normal each year. Run time Q-Q plots are even closer to a linear pattern which confirms normality. Given the sample sizes of more than 3000 per year, normality may be assumed.

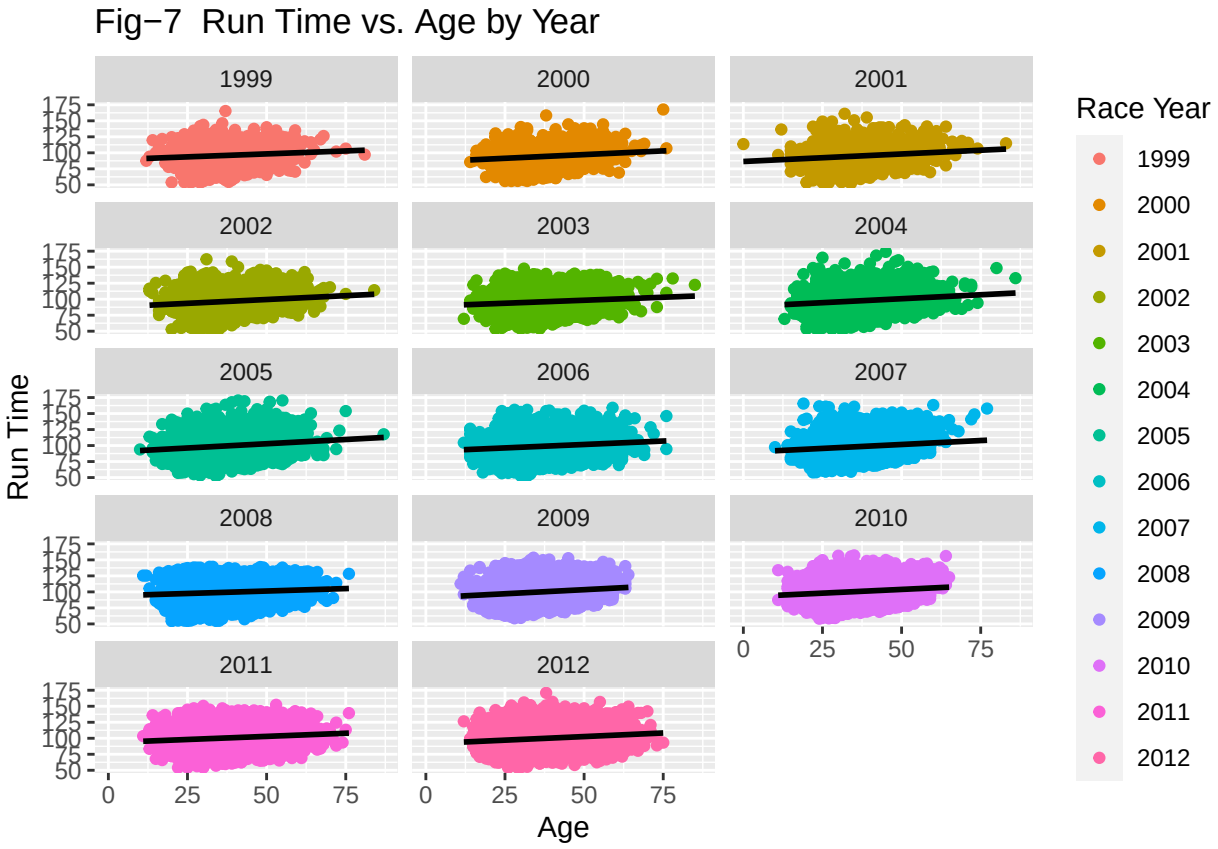


Fig−6 Q−Q Plots of Women's Run Time by Year



2.2.5 Linearity or Correlation Between Age and Run Time for Women

We do see a slightly positive correlation for each year of women's results. We also see a broad range of run time values for each age value, so this factor is important but it is not enough to form the basis of a prediction.



3 Conclusion

This project helped us realize the complexity of data gathering and data wrangling. Despite changes to the Cherry Blossom Race website, we were able to locate an archive dataset and complete our study. However, there were issues with specific years, and in the case of the year 2000, the archive was missing but we were able to locate an alternative file. The Cherry Blossom Race participants may continue to grow younger as the baby boomers stop running and more millenials reach their thirties. It would be interesting to see how the data has evolved since 2012. Overall we found that men are older and faster, but women showed more variation in the boxplots which means that more women are attempting the race at different levels of fitness. This may be encouraging for others who have not dared to enter this race. Now that our web scraping and data gathering skills are sharpened, we can take on any question that may be answered by unstructured data.

4 Appendix: All code for this report

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)

## LL Environment
#setwd("C:/SMU_Local/SMU_T5_QTW_CapA/QTW_7333/Unit 3 and 4/CS4")

library(pacman, htm2txt)
p_load(XML, tidyverse,ggplot2,  sjmisc)
ubase = "http://web.archive.org/web/20180721140041/http://www.cherryblossom.org/"
menURLs =
  c("results/1999/cb99m.html", "Cb003m.htm", "results/2001/oof_m.html",
    "results/2002/oofm.htm", "results/2003/CB03-M.HTM",
    "results/2004/men.htm", "results/2005/CB05-M.htm",
    "results/2006/men.htm", "results/2007/men.htm",
    "results/2008/men.htm", "results/2009/2009cucb-M.htm",
    "results/2010/2010cucb10m-m.htm",
    "results/2011/2011cucb10m-m.htm",
    "results/2012/2012cucb10m-m.htm")
urlsmen = paste(ubase, menURLs, sep = "")
womenURLs =
   c("results/1999/cb99f.html", "Cb003f.htm", "results/2001/oof_f.html",
    "results/2002/ooff.htm", "results/2003/CB03-f.HTM",
    "results/2004/women.htm", "results/2005/CB05-f.htm",
    "results/2006/women.htm", "results/2007/women.htm",
    "results/2008/women.htm", "results/2009/2009cucb-F.htm",
    "results/2010/2010cucb10m-f.htm",
    "results/2011/2011cucb10m-f.htm",
    "results/2012/2012cucb10m-f.htm")
urlswomen = paste(ubase, womenURLs, sep = "")
print(urlswomen)
# Clean html tags from text file lines
cleanFun <- function(htmlString) {
  return(gsub("<.*?>", "", htmlString))
}
  # Uses the archive list of websites for the cherry blossom race
  # and the years corresponding to the year and the gender
  # of the racer. Then retrieves the html into a text document
  # Finds the <pre> html tag which starts the table
  # and the </pre> which ends the table to extract only the racing data
```

```
  # Writes array to file labeled with year.

extractResults =
  function(url, year , sex )
  {
    if (year == 2009) { return()}
    download.file(url,"temp_page.txt")
    doc = readLines("temp_page.txt")
    # loop through HTML lines looking for the beginning and end of table of results
    for (i in 1:length(doc)) {
       if (str_contains(doc[i], "<pre>")) {
         start = i + 1
         break
       }
    }
    for (i in 1:length(doc)) {
       if (str_contains(doc[i], "</pre>")) {
         end = i -1
         break
       }
    }
    # Extract out only race results lines
    docline =  doc[start:end]
    # Clean any html tags out
    docline = cleanFun(docline)
    if (sex == "male") {
      outfile = paste("MenTxt/", year, sep = '', ".txt")
      }
    else {
      outfile = paste("WomenTxt/", year, sep = '', ".txt")
      }
    write_lines(docline, outfile)
  }
#Verify download locations exist for files
if (!file.exists("MenTxt/")) dir.create("MenTxt/")
if (!file.exists("WomenTxt/"))dir.create("WomenTxt/")


# Web scrape both men's and women's race results
years = 1999:2012
results = mapply(extractResults, url = urlsmen, year = years, "male")
results = mapply(extractResults, url = urlswomen, year = years, "women")
```

```
### Begin converting .txt files into data values
# Data cleaning

findColLocs = function(spacerRow) {

  spaceLocs = gregexpr(" ", spacerRow)[[1]]
  rowLength = nchar(spacerRow)

  if (substring(spacerRow, rowLength, rowLength) != " ")
    return( c(0, spaceLocs, rowLength + 1))
  else return(c(0, spaceLocs))
}
selectCols = function(shortColNames, headerRow, searchLocs) {
  sapply(shortColNames, function(shortName, headerRow, searchLocs){
    startPos = regexpr(shortName, headerRow)[[1]]
    if (startPos == -1) return( c(NA, NA) )
    index = sum(startPos >= searchLocs)
    c(searchLocs[index] + 1, searchLocs[index + 1])
  }, headerRow = headerRow, searchLocs = searchLocs )
}
extractVariables =
function(file, varNames =c("name", "home", "ag", "gun",
                           "net", "time"))
{

  # Find the index of the row with =s
  eqIndex = grep("^===", file)
  # Extract the two key rows and the data
  spacerRow = file[eqIndex]
  headerRow = tolower(file[ eqIndex - 1 ])
  body = file[ -(1 : eqIndex) ]
       # Remove footnotes and blank rows
  footnotes = grep("^[[:blank:]]*(\\*|\\#)", body)
  if ( length(footnotes) > 0 ) body = body[ -footnotes ]
  blanks = grep("^[[:blank:]]*$", body)
  if (length(blanks) > 0 ) body = body[ -blanks ]


  # Obtain the starting and ending positions of variables
  searchLocs = findColLocs(spacerRow)
  locCols = selectCols(varNames, headerRow, searchLocs)
```

```
  Values = mapply(substr, list(body), start = locCols[1, ],
                  stop = locCols[2, ])
  colnames(Values) = varNames

  #Values['net'] =
  return(Values)
}
convertTime = function(time) {
  timePieces = strsplit(time, ":")
  timePieces = sapply(timePieces, as.numeric)
  sapply(timePieces, function(x) {
                      if (length(x) == 2) x[1] + x[2]/60
                      else 60*x[1] + x[2] + x[3]/60
                      })
}
createDF =
function(Res, year, sex)
{
       # Determine which time to use
  useTime = if( !is.na(Res[1, 'net']) )
              Res[ , 'net']
            else if( !is.na(Res[1, 'gun']) )
               Res[ , 'gun']
            else
               Res[ , 'time']

  runTime = convertTime(useTime)

  Results = data.frame(year = rep(year, nrow(Res)),
                       sex = rep(sex, nrow(Res)),
                       name = Res[ , 'name'],
                       home = Res[ , 'home'],
                       age = as.numeric(Res[, 'ag']),
                       runTime = runTime,
                       stringsAsFactors = FALSE)
  invisible(Results)
}
mfilenames = paste("MenTxt/", 1999:2012, ".txt", sep = "")

menFiles = lapply(mfilenames, readLines)

names(menFiles) = 1999:2012
# fix spacer row on 2006 due to no gap between Hometown and Net Time
```

```
menFiles[['2006']][7] = "===== ======== ====== ==================== == ============== ======= =======
menResMat = sapply(menFiles, extractVariables)
menDF = mapply(createDF, menResMat, year = 1999:2012,
               sex = rep("M", 14), SIMPLIFY = FALSE)
cbMen = do.call(rbind, menDF)
wfilenames = paste("WomenTxt/", 1999:2012, ".txt", sep = "")
womenFiles = lapply(wfilenames, readLines)
names(womenFiles) = 1999:2012

# fix spacer row on 2006 due to no gap between Hometown and Net Time
womenFiles[['2006']][7] = "===== ======== ====== ==================== == ============== ======= ====

library("R.utils")

# Insert the header row from the men's file (rows 13-14) just above results on women's file
womenFiles[['2001']] = insert(womenFiles[['2001']],ats=3,values=menFiles[['2001']][3:4])
womenResMat = sapply(womenFiles, extractVariables)
# Create a dataframe of all data for women

womenDF = mapply(createDF, womenResMat, year = 1999:2012,
               sex = rep("F", 14), SIMPLIFY = FALSE)

cbWomen = do.call(rbind, womenDF)
# Plots and Analyses


# PLot Women's Age by Year
boxplot(age ~ year,
data=cbWomen,
main="Women's Age Distribution by Year",
xlab="Year",
ylab="Age",
col="cyan",
border="brown"
,sub=("Fig.1-Women's Age Distribution by Year"))


# Recall the PLot of Men's Age by Year
boxplot(age ~ year,
data=cbMen,
main="Men's Age Distribution by Year",
xlab="Year",
ylab="Age",
```

```
col="orange",
border="brown"
,sub="Fig.2-Men's Age Distribution by Year")

?boxplot
# PLot Women's Run Time by Year
boxplot(runTime ~ year,
data=cbWomen,
main="Women's Run Time Distribution by Year",
xlab="Year",
ylab="Run Time",
col="cyan",
border="brown",
sub="Fig-3-Women's Run Time Distribution by Year.")

# PLot Men's Run Time by Year
boxplot(runTime ~ year,
data=cbMen,
main="Men's Run Time Distribution by Year",
xlab="Year",
ylab="Run Time",
col="orange",
border="brown"
,sub="Fig-4-Men's Run Time Distribution by Year."
)


# Mean values across all years for Men and Women

womenAvgAge=mean(cbWomen$age,na.rm = TRUE)
paste("Average Age of Women Runners = ")
print(womenAvgAge)

menAvgAge=mean(cbMen$age, na.rm=TRUE)
paste("Average Age of Men Runners = ")
print(menAvgAge)

womenAvgRT=mean(cbWomen$runTime,na.rm = TRUE)
paste("Average Run Time of Women Runners = ")
print(womenAvgRT)

menAvgRT=mean(cbMen$runTime, na.rm=TRUE)
paste("Average Run Time of Men Runners = ")
```

```r
print(menAvgRT)


# Q-Q Plots for Women's Age by Year show normal distribution of data
# Thank you to the helpful tutorial at http://r-statistics.co/ggplot2-Tutorial-With-R.html

library(ggplot2)
ggplot(cbWomen, aes(sample = age, colour = factor(year))) +
  stat_qq() +
  stat_qq_line()+
  labs(title="Fig-5 Q-Q Plots of Women's Age by Year", x="theoretical", y="sample") +
  facet_wrap( ~ year, ncol=3)  # columns defined by 'year'

# Q-Q Plots of runTime for Women

ggplot(cbWomen, aes(sample = runTime, colour = factor(year))) +
  stat_qq() +
  stat_qq_line()+
  labs(title="Fig-6  Q-Q Plots of Women's Run Time by Year", x="theoretical", y="sample") +
  facet_wrap( ~ year, ncol=3)
# columns defined by 'year

# Plot Run Time vs. Age by Year

ggplot(cbWomen, aes(x=age, y=runTime, colour= factor(year))) + geom_point() + geom_smooth(method = "lm"
```