Cherry Blossom

Hieu Nguyen, Nithya Devadoss, Ramesh Simhambhatla, Ramya Mandava Sep 22, 2018

Abstract

In this case study, we are interested in understanding how people's physical performance changes as they age. Cherry Blossom Ten Mile Run race results from 1999 to 2012 are used to study how the age distribution of the runner's change over the years. We have limited our study with women racers' during the period (1999-2012). We have acquired the publicly available data from the Cherry Blossom website, by year and gender, and different formats. We have downloaded the data to local computer, examined the data structure, completed parsing and cleaning of the data, and created a consistent data structure to start our analysis. We have learned that the data of 1999 runners were typically older than the 2012 runners. We will analyze the women runner's data across 14 years using quantile–quantile plots, boxplots, and density curves to make our comparisons. Our point of interest is to determine if the distributions change over the years; is the change gradual, if any (Question #10: Modeling Runners' Times in the Cherry Blossom Race [2]).

1 Introduction

Cherry Blossom Ten Mile Run is one of the annual road run races held in Washington D.C. in early April when the cherry trees are typically in bloom. The Cherry Blossom started in 1973 as a training run for elite runners who were planning to compete in the Boston Marathon. It has since grown in popularity and in 2012, around 17,000 runners ranging in age from 9 to 89 participated. The race has become so popular that entrants are chosen via a lottery or they guarantee a spot by raising $500 for an official race charity. After each year's race, the organizers publish the results at http://www.cherryblossom.org/. These data offer a tremendous resource for learning about the relationship between age and performance [2].

One source of data about this comes from Cherry Blossom road races. Hundreds of thousands of people participate in road races each year; the race organizers collect information about the runners' times and often publish individual-level data on the Web. These freely accessible data may provide us with insights to our question about performance and age [2].

Our approach in this case study is:

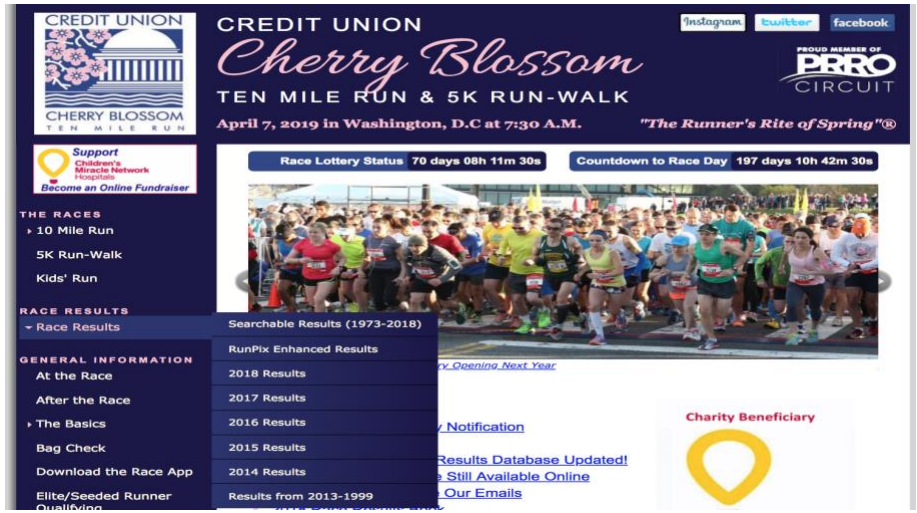
1) Capture Data from Cherry Blossom website

2) Examine Data and perform Data Cleaning
3) Create Tidy Data required for Analysis
4) Analyze Data using quantile–quantile plots, boxplots, and density curves
5) Compare the results

2 Methods

2.1 Data Acquisition:

We have captured the women runners' data from http://www.cherryblossom.org/. The picture illustrates the Cherry Blossom web page and source to the data by year.



Each year data published in a separate URL. We have visited each URL to download the data. We have observed that the data is published in different formats: XML, HTML, RSTL etc. We have attempted to read the data directly using R program (a code template given to us in the class by Prof. Slater). However, after thorough data exploration, we have found that the data parsing and creating a tidy data set is very complex and time consuming, which doesn't fit our schedule.

2.2. Create Tidy Data:

Since our objective is to perform analysis on the distribution of Women Runners; "Age" from 1999 to 2012 (14 years), and not on the coding of data capture, we have downloaded each file into a '.txt' file. We have then manually filtered, scrubbed and captured the 'Age' data into a '.csv' file. We have removed non-numeric data in this process to avoid unexpected results.

2.3. Data Analysis

Histogram of Ages:

As seen in figure 2.3.1, women between ages of 25 and 30 have participated the most in the Cherry blossom races, across years 1999:2012 together, with the range of ages being 10 – 81.

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
10.00   27.00   31.00   33.28   39.00   81.00
```

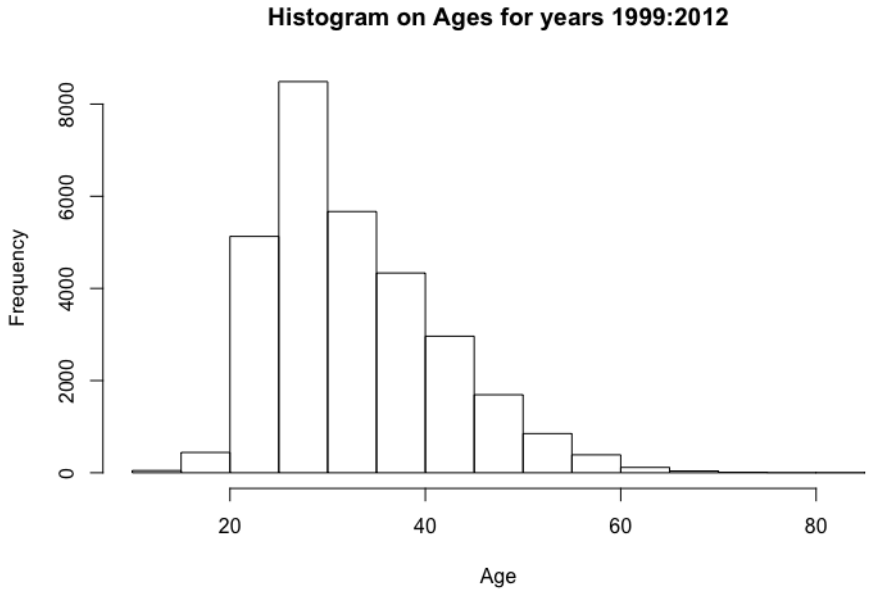


Figure 2.3.1 Histogram on Ages for years 1999:2012

Summary statistics by year:

As can be seen from the summary statistics (figure 2.3.2), the range(min:max) of ages has narrowed down from 1999 to 2012 by about 25% (range of 69 years in 1999 to 53years in 2012). We can also notice that the median age of female runners in 1999 is 33 years and it gradually decreased to 30 years by 2007 and it stabilized at that value.

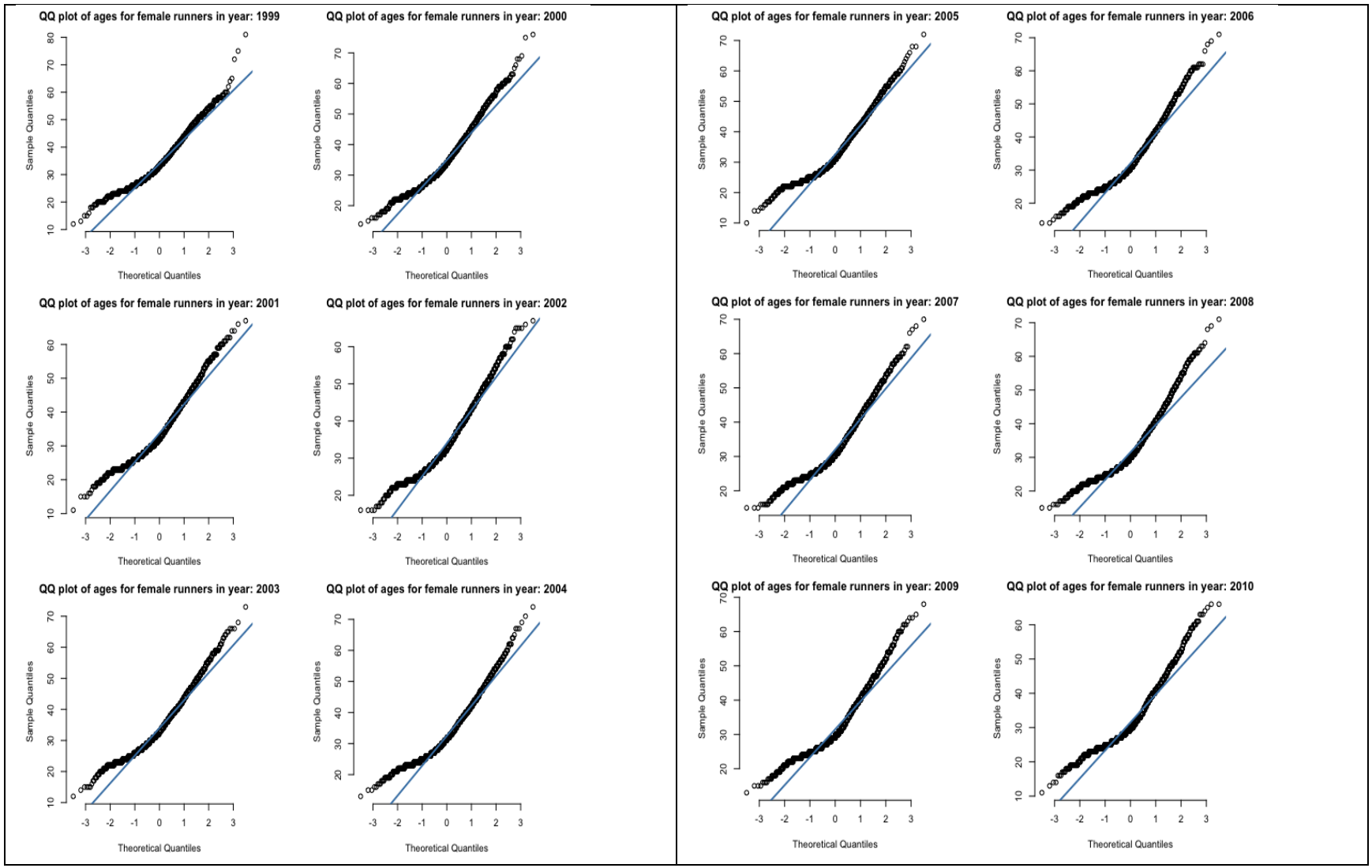
```
     X2012           X2011           X2010           X2009           X2008           X2007
 Min.   :15.0   Min.   :13.00   Min.   :11.00   Min.   :13.00   Min.   :15.00   Min.   :15.00
 1st Qu.:26.0   1st Qu.:26.00   1st Qu.:26.00   1st Qu.:26.00   1st Qu.:26.00   1st Qu.:26.00
 Median :30.0   Median :30.00   Median :30.00   Median :29.00   Median :30.00   Median :30.00
 Mean   :32.5   Mean   :32.55   Mean   :32.04   Mean   :31.87   Mean   :32.21   Mean   :32.54
 3rd Qu.:37.0   3rd Qu.:38.00   3rd Qu.:37.00   3rd Qu.:37.00   3rd Qu.:37.00   3rd Qu.:38.00
 Max.   :68.0   Max.   :72.00   Max.   :66.00   Max.   :68.00   Max.   :71.00   Max.   :70.00
     X2006           X2005           X2004           X2003           X2002           X2001
 Min.   :14.00   Min.   :10.00   Min.   :13.00   Min.   :12.00   Min.   :16.00   Min.   :11.00
 1st Qu.:26.00   1st Qu.:26.00   1st Qu.:26.00   1st Qu.:28.00   1st Qu.:28.00   1st Qu.:28.00
 Median :31.00   Median :31.00   Median :31.00   Median :33.00   Median :32.00   Median :32.00
 Mean   :32.79   Mean   :33.16   Mean   :33.22   Mean   :34.43   Mean   :34.32   Mean   :34.05
 3rd Qu.:38.00   3rd Qu.:39.00   3rd Qu.:39.00   3rd Qu.:40.00   3rd Qu.:40.00   3rd Qu.:39.50
 Max.   :71.00   Max.   :72.00   Max.   :74.00   Max.   :73.00   Max.   :67.00   Max.   :67.00
     X2000           X1999
 Min.   :14.00   Min.   :12.00
 1st Qu.:29.00   1st Qu.:28.00
 Median :34.00   Median :33.00
 Mean   :35.56   Mean   :34.61
 3rd Qu.:41.00   3rd Qu.:40.00
 Max.   :76.00   Max.   :81.00
```

Figure 2.3.2 Summary statistics on Ages of female runners for years 1999:2012

Quantile-Quantile plot:

Quantile-quantile plots are used to graphically observe if two data sets are coming from similar distributions. Figure 2.3.3 shows q-q plots of Theoretical quantiles vs. Data quantiles of ages of runners by year (from 1999 to 2012). It can be noticed that in the earlier years (around 1999) there are several older runners(outliers) with ages above 70 in the top 20% quantile, while this number gradually decreased as the years progressed. This could be because, the total number of runners increased over years and the probability of older

runners being picked in the lottery system is lower when compared to younger runners due to the uneven frequency distribution of ages (higher for ages 25-30).



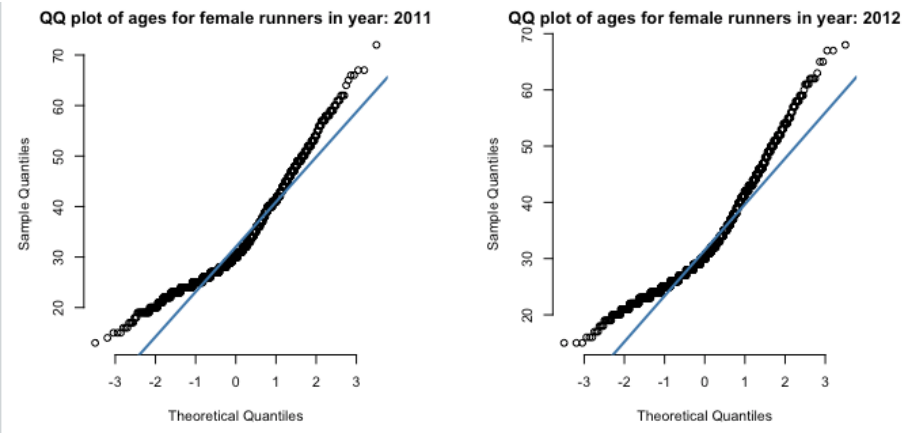


Figure 2.3.3 Quantile-Quantile plots of ages of female runners across years 1999:2012

Box Plots:

As can be seen from the boxplots (figure 2.3.4), there are more outliers (older runners) in the earlier years while the distribution tightened up slightly as years progressed. It can also be noticed that the inter-quartile range is about the same across all years. However, the median slightly decreased from 1999 to 2007 and it remained the same thereafter.

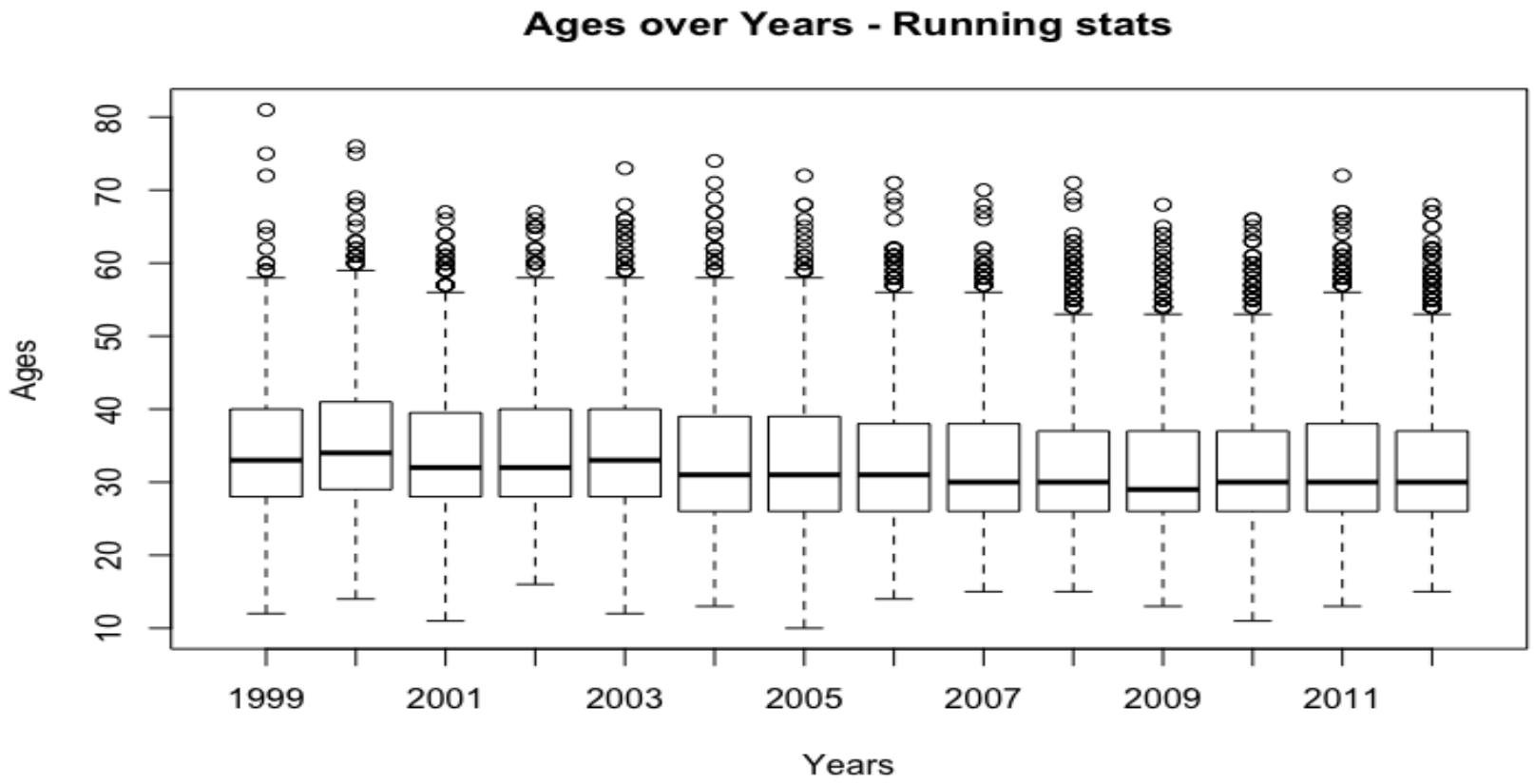


Figure 2.3.4 Box plot of Ages of female runners by Year from 1999 to 2012.

Density Curves:

Density plots have been used in order to look into the change of age demographics of the runners over years. Ggplot and plotly libraries have been used to achieve this. As seen in figure 2.3.5, the density plot is color coded by year (also shown in the key on our right-hand side of the plot).

Analyzing the density plot, it can be noted that the curves are shifting positions gradually from right to left, and the distribution is narrowed, while height of the plot is increasing as the years progressed from 1999 to 2012. These changes indicate that the mean age as of participants(women) in 2012 has decreased as compared to 1999.

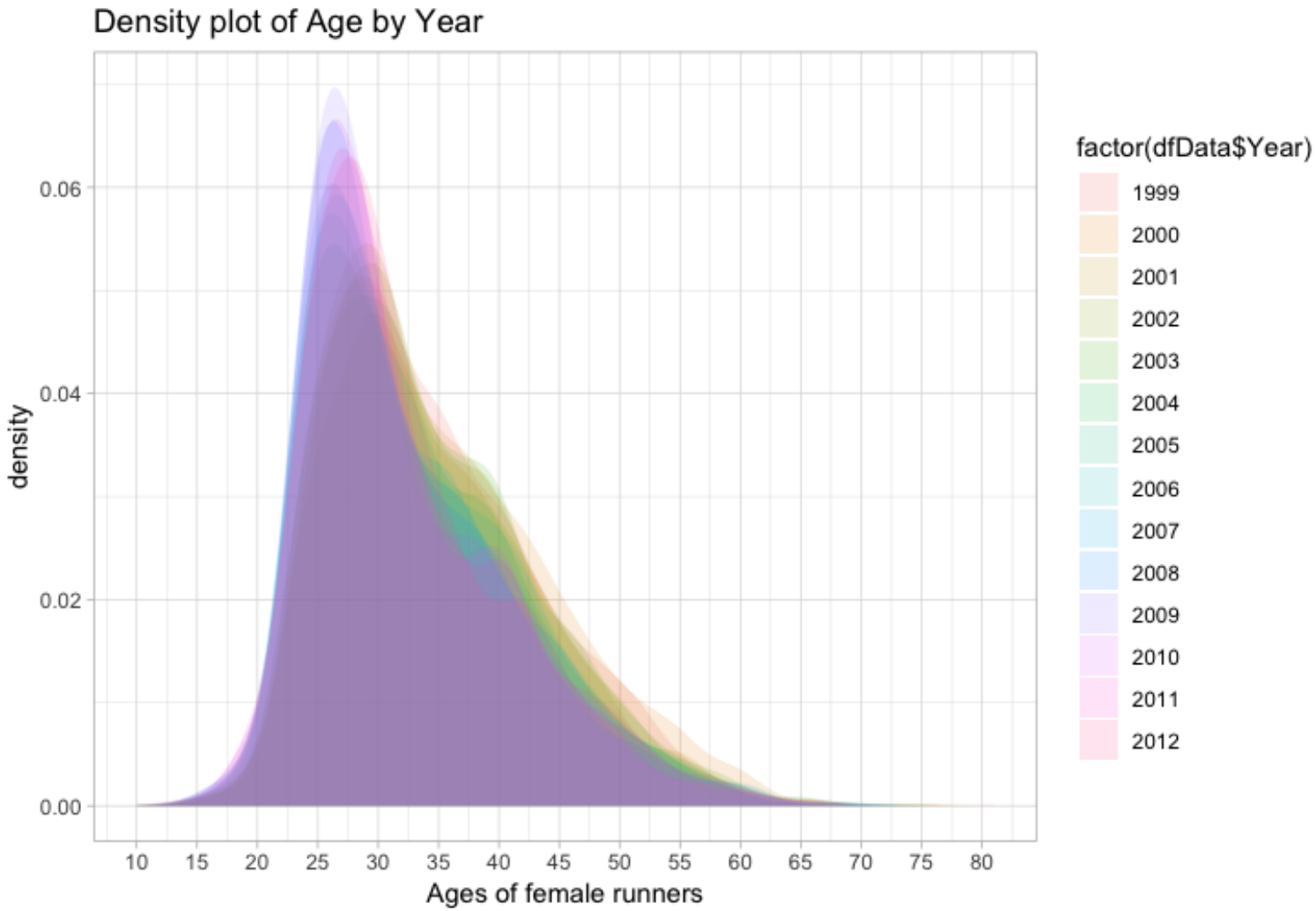


Figure 2.3.5 Density plot of Ages of female runners, color coded by Year (1999:2012)

3 Conclusion

In this case study, we acquired women runners' data from Cherry Blossom website, scrubbed the data manually to create a data structure for Age of the runners. We have further analyzed the Women Runners' Age data using various statistical techniques such as: QQ plots, histogram, Summary statistics, Density curves and Box plots.

From our analysis of age distribution across all years from 1999 to 2012, the general population female race runners display right-skewed distribution with the highest frequency of runners falling into the 25-30 years age bin for Women.

It can be noticed from the quantile-quantile plots that in the earlier years (around 1999) there are several older runners(outliers) with ages above 70 in the top 20% quantile. Stacked density plot with a single curve for each year, further shows a decrease in mean age of female runners from 1999 to 2012. This is further visually supported by the boxplots that shows the median line that decreases from ~35 to ~32 across years. However, the drop in this age has stopped in 2007 and remained stable for the rest of the years.

4 References

[1] Prof Slater's Jupyter Notebook sample and class material

[2] Nolan, Deborah; Lang, Duncan Temple. Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving (Chapman & Hall/CRC The R Series) (Pages 45, 101)

APPENDIX A

CODE:

```
library(XML)
library (plyr)
library(gdata)
library(ggplot2)
library(gridExtra)
ubase = "http://www.cherryblossom.org/"
womenURLs =
  c("results/1999/cb99f.html", "results/2000/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/2009/09cucb-F.htm",
    "results/2010/2010cucb10m-f.htm",
    "results/2011/2011cucb10m-f.htm",
    "results/2012/2012cucb10m-f.htm")

# http://www.cherryblossom.org/results/2000/Cb003f.htm
#http://www.cherryblossom.org/results/2009/09cucb-F.htm

urls = paste(ubase, womenURLs, sep = "")

urls[1:3]

urls

extractResTable =
  # takes a list of websites from the cherry blossom race
  # a list of years corresponding to the year the result is for
  # and the gender of the participant
```

```
# Retrieve data from web site,
# find the preformatted text,
# and write lines or return as a character vector.
# returns a list of strings corrsponding to lines in the web url
function(url = "http://www.cherryblossom.org/results/2009/09cucb-F.htm",
        year = 1999, sex = "female", file = NULL)
{
  doc = htmlParse(url)

  if (year == 2000) {
    # Get preformatted text from 4th font element
    # The top file is ill formed so the <pre> search doesn't work.
    ff = getNodeSet(doc, "//font")
    txt = xmlValue(ff[[4]])
    els = strsplit(txt, "\r\n")[[1]]
    print(year)
  }
  else if (year == 1999) {
    # Get preformatted text from <pre> elements
    pres = getNodeSet(doc, "//pre")
    txt = xmlValue(pres[[1]])
    els = strsplit(txt, "\n")[[1]]
    print(year)
  }
  else {
    # Get preformatted text from <pre> elements
    pres = getNodeSet(doc, "//pre")
    txt = xmlValue(pres[[1]])
    els = strsplit(txt, "\r\n")[[1]]
    print(year)
  }

  if (is.null(file)) return(els)
```

```
  # Write the lines as a text file.
    writeLines(els, con = file)
}

years = 1999:2012
womenTables = mapply(extractResTable, url = urls, year = years)
names(womenTables) = years
sapply(womenTables, length)


save(womenTables, file = "/Users/ramya/Desktop/CBMenTextTables.rda")

capture.output(womenTables, file = "CBWomenTextTables4.xls")


df <- ldply (womenTables, data.frame)
df1 <- df

df1 <- df1[-c(0:4), ]

colnames(df1) <- c("YEAR","PLACE","DIV /TOT","NAME", "AGE", "HOMETOWN" , "TIME", "PACE")


myData = read.csv("/Users/ramya/Documents/GitHub/QuantifyingTheWorld_CS1/Week2/data/DataFinal2.csv", header =
TRUE, sep = ",")

df3 <-myData
df3<- na.omit(myData)

dataFinal <- na.omit(myData)


dfData <- ldply (dataFinal, data.frame)
```

```
colnames(dfData) <- c("Year", "Age")

dfData$Year <- as.numeric(substring(dfData$Year, 2))


#Box Plots
boxplot(Age~Year, data=dfData, main="Ages over Years - Running stats",
        xlab="Years", ylab="Ages")


#simple qq plot
#require(gridExtra)

par(mfrow = c(3,2), mar=c(4,4,1,4) + 0.8)
#"QQ plot of ages for female runners in year:"
for (i in 2005:2012) {
  temp <- subset(dfData, dfData$Year == i)
  head(temp)
  qqnorm(temp$Age, pch = 1, frame = FALSE, main = paste("QQ plot of ages for female runners in year:",i),
xlab="Theoretical Quantiles", ylab="Sample Quantiles")
  qqline(temp$Age, col = "steelblue", lwd = 2)
}
dev.off()

#Density Plots
d <- density(dfData$Age) # returns the density data
plot(d) # plots the results


options(repr.plot.width=10, repr.plot.height=8)
age.d = ggplot(dfData, aes(dfData$Age, fill = factor(dfData$Year))) + geom_density(col=NA, alpha=0.15) + theme_light()+

  scale_x_continuous(breaks = pretty(dfData$Age, n = 20))+
```

```
  ggtitle("Density plot of Age by Year") + xlab("Ages of female runners")

age.d

#Summary Stats
summary(dfData)
summary(myData)

#Histogram
hist(dfData$Age, main="Histogram on Ages for years 1999:2012", xlab="Age")

par(mfrow = c(7,2), mar=c(4,4,1,4) + 0.8)
for (i in 1999:2012) {
  temp <- subset(dfData, dfData$Year == i)
  head(temp)
  hist(dfData$Age, main=paste("Histogram on Ages for year:",i), xlab="Age")
}

# to fix plot.new() margin too big error
dev.off()
```