# Abstract

This case study is intended to familiarize with data ingestion via web scraping technique, clean, and parse data for statistical analysis and modeling. In addition to extension of work performed by Nolan and Lang utilizing Credit Union Cherry Blossom Ten Mile Run, we performed out analysis to compare make and female age-normalized time distribution for the runners in 1999 and 2012. The outcome is then represented in Normal Q-Q plots, boxplots and density curves.

Distributions are different for men runners in 1999 and 2012. We find an interesting point that runners are younger in 2012 but they have longer runner times.

In this Data Era [1], there are potentials in analyzing data and bringing insights to a problem we have interests in. For instance, to find out whether there is a relationship between physical performance and aging and what the relationship is, road races data can be the data we could do analysis on. This kind of data is collected by the race organizers and it is at individual level. Data is often published on the Web. We can bring some insights to the question we have interests in.

Cherry Blossom Ten Mile Run is one of the annual races held in Washington D.C. early April which the cherry blossom is blooming around this time. This race run started in 1973 as a training run for runners planning to compete in Boston Marathon. Its participants range from age 9 to 89. The organizer publishes the results at http://www.cherryblossom.org/. The data published provides a very helpful resource including name, sex, hometown and final time of participants for us to learn the relationship between physical performance and aging.

The original analysis was provided by the book of Daniel Kaplan and Deborah Nolan in the Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. In our case study, the data has been web-scraped and it has results from all years from 1999 to 2012. The data is divided by genders meaning MenTxt for men and WomenTxt for women.

# Background

For conducting our analysis, we scraped the results of the races from 1999 to 2012 data from the race’s official website as text files. The results were contained in individual files for each year. There were separate files for men and women. It was noticed that all files did not have the same formatting. Also, the features/columns contained in the file were not consistent across all. This warranted cleaning of the scraped files so that they are ready to be used for analysis.

The cleaned data contains 6 features including runner name, gender, hometown and the results. There are 70,070 male runners and 75,972 female runners attended Cherry Blossom Run across 14 years. There are two rounds of normalization involved. First normalization is each male runner's time divided by the fastest time for the runner of the same age. Then the normalized results are used in a Loess model to get the predictions. Second normalization is each male runner's time divided by these predictions. Then we compare the distribution of runner times in 1999 and 2012

# Methods

The methods used to get the results of the Cherry Blossom races for the years 1999-2012 for both men and women participants and get the data ready to be used for our analysis, included the below:

#### Web Scraping Race Run Time Results

We begin the acquisition of data by extracting the files for the years 1999 through 2012 from the official website for the Cherry Blossom race (www.cherryblossom.org). A separate link is provided for the race results for each year for women and men races. We did notice that the URLs for the individual years’ results were not consistent. Because of this inconsistency, we had to extract data by visiting each year’s results web page. In order to read data into R, each years’ race result file is extracted, and the web data is converted to a text document. We created a function to accomplish this and also added some conditional statements to resolve differences between the files.

It is observed that all files did not have the same formatting or even the same columns (that is the headers differ between years). Figure 1 taken from the book Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving by Daniel Kaplan and Nolan show clearly the difference in column headings

<center><img src="Men's 2011 xolumns.png"> <img src="Men's 2012 xolumns.png"> <center><b>Figure 1. Men's Race Results for 2011 and 2012</b> </center></center>

#### Extracting Men and Women’s Race Results Files

We can now extract the results files for each year from the URLs and save them to the local directories. We apply the function to the men’s results file to extract the data and save it to the /MenTxt/ directory. We take the same steps for the women’s results and save the extracted files to the /WomenTxt/ directory.

We notice that the women’s race results file for the year 2001 had the header missing. This issue was resolved by adding the header to the Women’s 2001 race results file in the file directly.

#### Data Import and Cleanup

For the purposes of our analysis, we create a function that import the text files and only keep the variables for name of the runner (“name), their hometown (“home”), age (“ag”), gun time (“gun”), net time (“net”), and time (“time”). Only the first few characters that are unique to the variables are used, since the full column names differ between the result files. Blank rows and footnotes are also removed to clean the data with this function.

We also notice that the columns can have different widths and it’s not necessary that the data for the column is placed at the same position in all the files. To resolve this issue, cleaning the data also includes creating function (“findColLocs”) to determine the locations of the columns so that values of the columns can be extracted. It does this by searching for spaces in the rows. This is done to remove the dependency on variable names to extract the data. We extract all the columns for the variables we have determined we need, using the starting and ending positions of the columns. This function is usable with both the men’s and women’s results files.

Next, we also create a function (“selectCols”) that would let us select columns by name. We need, as inputs to this function, the names of the desired columns, the header row that contains the column names, and the locations of the blanks in the separator row.[1] This function is usable with both the men’s and women’s results files.

We wrap up these two newly created functions in another function (“extractVariables”) that can be applied to each year’s data for either the men’s or women’s race results files. We create the Men’s results files first applying our function. We see that that the 2003 and 2006 files require further cleaning. Once the men’s results files are created, we proceed with creating the results files for the races by year for the women. We apply the same fixes to the women’s 2003 and 2006 race results files.

After all the cleaning was performed and the data was deemed ready to be used in our analysis, we had individual files by year for both men and women racing results. Below tables summarize the number of rows per file available for us to use:

<center><img src="Number of rows all mens files.png"> <center><b>Table 1. Number of rows in Men's Race Results by year</b> </center>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

<center> <img src="Number of rows all womens files.png"> <center> <b>Table 2. Number of rows in Women's Race Results by year</b> </center>

We also need to clean up the time variable. The data for time needs to be formatted into hh:mm:ss and the data type should be numeric, not character. Our function convertTime helps resolves the formatting issues for the time variable. The use of strsplit, helps in splitting the time variable in the desired format of hours, minutes, and seconds. After the time variable is split, it is converted to type numeric and aggregated into one value with time being reported in minutes. [1]

We need to have data frames created for all the results for men’s and women’s races. We create a function to handle this requirement. Included in the function is the ability to determine which time is to be used in the data frame from amongst the three provided. Also, two new variables namely “year” and “sex” are also created and the data for year and gender will be used as inputs. Other functionalities included are dropping rows with no time and removing # and \* and blanks from the time variable. Applying the function to create a data frame for men, we see that it does not include any “NA” values for the time variable.

# Results

#### 1. Distribution Comparison for Men Runners Times

Generally speaking, distributions are different for men runners in 1999 and 2012. We also find an interesting point that most runners were younger in 2012 but they have had longer runner times. Here we break out the distribution comparison into three parts density plot comparison, qq-plot comparison and the summary statistics comparison.

\* \*\*Density Plot\*\*:

In this part, we included the density distribution of age for runners in 1999 and 2012 to tell our interesting findings.

Let us start off by showing the density distribution of age for runners of these two years. We find that they have different distributions. The density distribution shows evidence of age shifting when comparing men runners from these two years. When you look at where the peaks of two curves appear, you can see that in 2012 men runners have a mode around 29 and the mode is around 40 in 1999 curve. This tells us most men runners were younger in 2012 than 1999(Figure 2).

<center><img src="CB\_AgeDensity99vs12.png"><center><b>Figure 2. Density Curves for the Age of Male Runners in 1999 and 2012.</b> These two density curves have quite different shapes and different modes(where the peaks at). In 2012 men runners have a mode around 29 and the mode is around 40 in 1999 curve. Most men runners were younger in 2012 than 1999</center>

Then we look at the density distribution of the age normalized times for runners in the year of 1999 and the year of 2012. We find that they have different distributions. Density curve of 1999 runners has a slimmer shape and it has a good bell shape while 2012 density curve has a slightly left skewed at top, broader shape in the middle and it is not as perfectly bell-shaped. The density distribution shows evidence of runner time shifting when comparing men runners from these two years. When you look at where the peaks of two curves appear, you can see that most men runners take longer times to finish the race of 2012 and in 1999 they have had shorter times(Figure 3).

<center><img src="CB\_TimeDistributionDensity99vs12FIXED.png"></center><b>Figure 3. Density Curves for the Runner Times of Male Runners in 1999 and 2012.</b> Density curve of 1999 runners has a slimmer shape and it has a good bell shape while 2012 density curve has a slightly left skewed at top, broader shape in the middle and it is not as perfectly bell-shaped. The density distribution shows evidence of runner time shifting when comparing men runners from these two years. Most men runners have longer times to finish the run of 2012 and in 1999 they have shorter times.</center>

This is very interesting. We would assume that 2012 runners would have a shorter run time because most males who joined 2012 run were younger, but the fact is 1999 people were faster.

\* \*\*QQ-plot\*\*:

Now let us look into QQ-plot where we can compare two sets of quantiles against each other and we can also find out whether they have the same distribution. On the x-axis, it is the quantiles for the year of 1999 which we are treating as the expected quantiles. At the very beginning, these two sets of quantiles are forming a straight line well which they do go along with the middle reference line. This indicates their distributions are very similar at the very left corner. After that, the line is still roughly forming a straight line, but this straight line is going off the reference line in the middle. By the trend of this QQ-plot we can see that 2012 runner times are more spread-out than 1999 runner times. We can also see that 2012 runner times have higher values(Figure 4). [2]

<center><img src="CB\_TimeDistributionQQPlot99vs12.png"></center><b>Figure 4. Quantile-quantile Plot of Male Runner Normalized Time.</b>Their distributions are very similar at the very left corner. After that, the line is still roughly forming a straight line, but this straight line is going off the reference line in the middle. By the trend of this QQ-plot we can see that 2012 runner times are more spread out than 1999 runner times. We can also see that 2012 runner times have higher values.</center>

\* \*\*Summary Statistics\*\*

When looking at the five number summary and boxplot(1999 is on the left), we can say they are close but runner times in 2012 are still higher than that of 1999 in terms of median values, 1st and 3rd quantiles and the max value. What we can also see is that 2012 data are more spread-out than 1999 data which is another evidence of what we have talked about in QQ-plot above(Table 3 and Figure 5).

<center><img src="men5numbersummary.png"><center> <b>Table 3. Summary Stats for Male Runner Normalized Time 1999 vs. 2012. </b> Five number summary of the normalized runner time shows that runner times in 2012 still have higher vales except for the minimum. </center>

<center><img src="CB\_TimeBoxPlot99vs12.png"><center> <b>Figure 5. Box Plot for Male Runner Normalized Time 1999 vs. 2012. </b> Box plot shows how their data are skewed and means are different. 2012 data have a broader range than them in 1999. Also 2012 runner times have a higher mean.</center>

#### 2. Distribution Comparison for Women Runner Times

Distributions are also different for women runners in 1999 and 2012. We also find an interesting point that most runners in these years have similar age group but in 2012 they have had longer runner times. Here we break out the distribution comparison into three parts density plot comparison, qq-plot comparison and the summary statistics comparison.

• \*\*Density Plot\*\*:

In this part, we also include the density distribution of age for runners in 1999 and 2012 tell our interesting findings.

Again, let start off showing the density distribution of age for runners of these two years. This time we find that they have similar shapes. Unlike for men, the density distribution shows very little evidence of age shifting when comparing women runners from these two years meaning when you look at where the peaks of two curves appear, you can see that most females runners have a similar age in 2012 to participants in 1999, but their peak values are different which year of 2012 has a higher peak(Figure 6).

<center><img src="CB\_AgeDensityWomen99vs12.png"><center><b>Figure 6. Density Age Distribution for Female runner Normalized Time 1999 vs. 2012. </b>The density distribution shows very little evidence of age shifting when comparing women runners from these two years meaning when you look at where the peaks of two curves appear, most people had similar age, but their peak values are different which year of 2012 has a higher peak.</center></img>

Then we look at the density distribution of the age normalized times for runners in the year of 1999 and the year of 2012. We find that they have similar but slightly different distributions. Density curve of 2012 runners has a slimmer shape in the middle and it has a good bell shape while 1999 density curve has an almost perfect bell shape. The density distribution shows evidence of a bigger runner time shifting when comparing women runners from these two years. When you look at where the peaks of two curves appear, you can see two modes are quite different meaning most women runners have had a much longer times to finish the run in 2012 and they have had shorter times in 1999(Figure 7).

<center><img src="CB\_TimeDistributionDensityWomen99vs12FIXED.png"><center><b>Figure 7. Density Distribution Plot for Female runner Normalized Time 1999 vs. 2012. </b> The density distribution shows evidence of a bigger runner time shifting when comparing women runners from these two years. When looking at where the peaks of two curves appear, most women runners have had a much longer times to finish the run in 2012 and they have shorter times in 1999.</center></img>

This is different from what we have observed when examining the men data. With not that much age differences in these two years, the runner times were quite different. Most people in 2012 were much slower than in 1999.

• \*\*QQ-plot\*\*:

Again, on the x-axis, it is the quantiles for the year of 1999 which we are treating as the expected quantiles. The line is roughly forming a straight line, but this straight line is steeply going off the line in the middle. By the trend on this qq-plot we can see that 2012 runner times are much more spread out than 1999 runner times. Based on the line on this plot, we can also see that 2012 runner times have much higher values(Figure 8).[2]

<center><img src="CB\_TimeDistributionQQPlotwomen99vs12.png"><center><b>Figure 8. QQ Plot for Female runner Normalized Time 1999 vs. 2012. </b>The line is roughly forming a straight line, but this straight line is steeply going off the line in the middle. By the trend on this qq-plot we can see that 2012 runner times are much more spread out than 1999 runner times. We can see that 2012 runner times have much higher values.</center></img>

• \*\*Summary Statistics\*\*

When looking at the boxplot(1999 is on the left) and the 5 number summary(Table 4 and Figure 9), we can tell there is an obvious difference. The means are not as close as them of men data. Based on the five-number summary, runner times in 2012 are much higher than that of 1999. What we can also see is that 2012 data are more spread out than 1999 data.

<center><img src="women5numbersummary.png"><center> <b>Table 4. Summary stats for female runner Normalized Time 1999 vs. 2012. </b>The five number summary shows data in 2012 are larger than those in 1999.</center>

<center><img src="CB\_TimeBoxPlotwomen99vs12.png"><center> <b>Figure 9. Box Plot for Female runner Normalized Time 1999 vs. 2012. </b>Box plot proves what we see from the 5 number summary. 2012 data have a broader range than them in 1999. Also 2012 runner times have a higher mean.</center> </img>

#### 2. Distribution Comparison for Women Runner Times

Distributions are also different for women runners in 1999 and 2012. We also find an interesting point that most runners in these years have similar age group but in 2012 they have had longer runner times. Here we break out the distribution comparison into three parts density plot comparison, qq-plot comparison and the summary statistics comparison.

• \*\*Density Plot\*\*:

In this part, we also include the density distribution of age for runners in 1999 and 2012 tell our interesting findings.

Again, let start off showing the density distribution of age for runners of these two years. This time we find that they have similar shapes. Unlike for men, the density distribution shows very little evidence of age shifting when comparing women runners from these two years meaning when you look at where the peaks of two curves appear, you can see that most females runners have a similar age in 2012 to participants in 1999, but their peak values are different which year of 2012 has a higher peak(Figure 6).

<center><img src="CB\_AgeDensityWomen99vs12.png"><center><b>Figure 6. Density Age Distribution for Female runner Normalized Time 1999 vs. 2012. </b>The density distribution shows very little evidence of age shifting when comparing women runners from these two years meaning when you look at where the peaks of two curves appear, most people had similar age, but their peak values are different which year of 2012 has a higher peak.</center></img>

Then we look at the density distribution of the age normalized times for runners in the year of 1999 and the year of 2012. We find that they have similar but slightly different distributions. Density curve of 2012 runners has a slimmer shape in the middle and it has a good bell shape while 1999 density curve has an almost perfect bell shape. The density distribution shows evidence of a bigger runner time shifting when comparing women runners from these two years. When you look at where the peaks of two curves appear, you can see two modes are quite different meaning most women runners have had a much longer times to finish the run in 2012 and they have had shorter times in 1999(Figure 7).

<center><img src="CB\_TimeDistributionDensityWomen99vs12FIXED.png"><center><b>Figure 7. Density Distribution Plot for Female runner Normalized Time 1999 vs. 2012. </b> The density distribution shows evidence of a bigger runner time shifting when comparing women runners from these two years. When looking at where the peaks of two curves appear, most women runners have had a much longer times to finish the run in 2012 and they have shorter times in 1999.</center></img>

This is different from what we have observed when examining the men data. With not that much age differences in these two years, the runner times were quite different. Most people in 2012 were much slower than in 1999.

• \*\*QQ-plot\*\*:

Again, on the x-axis, it is the quantiles for the year of 1999 which we are treating as the expected quantiles. The line is roughly forming a straight line, but this straight line is steeply going off the line in the middle. By the trend on this qq-plot we can see that 2012 runner times are much more spread out than 1999 runner times. Based on the line on this plot, we can also see that 2012 runner times have much higher values(Figure 8).[2]

<center><img src="CB\_TimeDistributionQQPlotwomen99vs12.png"><center><b>Figure 8. QQ Plot for Female runner Normalized Time 1999 vs. 2012. </b>The line is roughly forming a straight line, but this straight line is steeply going off the line in the middle. By the trend on this qq-plot we can see that 2012 runner times are much more spread out than 1999 runner times. We can see that 2012 runner times have much higher values.</center></img>

• \*\*Summary Statistics\*\*

When looking at the boxplot(1999 is on the left) and the 5 number summary(Table 4 and Figure 9), we can tell there is an obvious difference. The means are not as close as them of men data. Based on the five-number summary, runner times in 2012 are much higher than that of 1999. What we can also see is that 2012 data are more spread out than 1999 data.

<center><img src="women5numbersummary.png"><center> <b>Table 4. Summary stats for female runner Normalized Time 1999 vs. 2012. </b>The five number summary shows data in 2012 are larger than those in 1999.</center>

<center><img src="CB\_TimeBoxPlotwomen99vs12.png"><center> <b>Figure 9. Box Plot for Female runner Normalized Time 1999 vs. 2012. </b>Box plot proves what we see from the 5 number summary. 2012 data have a broader range than them in 1999. Also 2012 runner times have a higher mean.</center> </img>

# Conclusion

In this project, we are to compare the distribution of the age-normalized times for male and female runners in 1999 and 2012, then describe our findings. After Loess modelling and two rounds of normalizations, we have compared the density distribution, QQ-plot, boxplots and the five number summary.

From the runner times density distribution and the age density distribution outputs, we can see that in 2012 most male runners were younger but they were slower. We have also confirmed this interesting finding through the following five number summary and QQ plot outputs. Unlike male runners, most female runners for these two years had similar age but again they were slower in 2012.

From what we have observed, for both men and women, the age normalized times distribution in 1999 is different from 2012.

[Back to Contents](#Contents)