**# Introduction**

In this Data Era, 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. The data has been already 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.

To answer the above questions, we are using the data for each gender and a Loess model to predict the smoothed run times after two rounds of normalization. Then we have done the distribution comparison for men runners and also for women runners in the year of 1999 and in the year of 2012.

Resources: Daniel Kaplan and Deborah Nolan in the Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving

**## Results**

1. **Distribution Comparison for Men Runner Times**

General speaking, distributions are different for men runners in 1999 and 2012. We also find an interesting point that runners are younger in 2012 but they 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: (two figures used, figure 11 and 15)

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

Let start off showing the density distribution of age for runners of these two years. In this visualization, 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 men runners are younger of 2012 that the peak appears around 30 years old while the peak appears around 40 of 1999. (figure 11).

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 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 men runners have longer times to finish the run of 2012 and in 1999 they have shorter times. (figure 15)

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

* QQ-plot: (Here I decided not to include the age qq plot, this section uses the qq-plot figure after figure 15, remember to change the name of it. It’s now called figure 12)

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 straight 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. (<https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot>)

* Summary Statistics. (I need a boxplot for this normalized data using time99Norm and time12Norm. Please create the plot and then put the description in.)

When looking at the boxplot, 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. (I need to see the skewness for the write up)

1. **Distribution Comparison for Women Runner Times**

Distributions are different for women runners in 1999 and 2012. We also find an interesting point that runners are younger in 2012 but they 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: (two figures used, one at cell 77 and one at cell 82. Title is needed, and figure dimension needs to change.)

In this part, we will also include the density distribution of age for runners in 1999 and 2012 here 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 (cell 77 result). 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 ages are similar, but their peak values are different which year of 2012 has a higher peak.

Then we look at the density distribution of the age normalized times for runners in the year of 1999 and the year of 2012(cell 87 result). 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 big 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 that women runners have a much longer times to finish the run in 2012 and they have shorter times in 1999.

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 are very different. People in 2012 are much slower than in 1999.

* QQ-plot: (I need a qq plot for this part. But I will describe it coz I know how it is supposed to look like)

Again, on the x-axis, it is the quantiles for the year of 1999 which we are treating as the expected quantiles. Just like the men data, at the very beginning, these two sets of quantiles are forming a straight line well and they do go along with the middle straight line. This indicates their distributions are very similar at the very left corner. After that, 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. (<https://stats.stackexchange.com/questions/101274/how-to-interpret-a-qq-plot>)

* Summary Statistics. (I need a boxplot for this normalized data using womentime99Norm and womentime12Norm. Please create the plot and then put the description in.)

When looking at the boxplot, 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. (I need to see the skewness of normalized female box plot for the write up)