**Abstract**

In this case study we perform comparative analysis on age-normalized times for the runners in 1999 and 2012 for both the male and female runners. The analysis required scraping the race data for the years 1999 to 2012 to get the data for the male and female runners from the Cherry Blossom Race website provided. The data was also cleaned prior to conducting analysis.

**Introduction**

The Cherry Blossom 10-Mile Race is held in April, during the time of the Cherry Blossom Festival in Washington DC. According to the National Cherry Blossom Festival official website, “The National Cherry Blossom Festival commemorates the 1912 gift of 3,000 cherry trees from Mayor Yukio Ozaki of Tokyo to the city of Washington, DC, and celebrates the enduring friendship between the people of the United States and Japan.T.” [2] The race was first held in 1973 and at that time was considered training grounds for start athletes to compete in more prominent races like the Boston Marathon. [1] Over time the Cherry Blossom race has evolved and is now considered a local activity where people of all abilities, and not just start athletes, participate. The race results are available at the official website of the Cherry Blossom Race, <http://www.cherryblossom.org/aboutus/results.php>

We are interested in conducting an analysis to observe if a runner’s age has any bearing on their run-time. In other words, we want to compare running performances for the years 1999 and 2012 to know if there is any correlation between a runner’s age and their running performance. We would like to conduct this survey for both the male and female runners.

**Background**

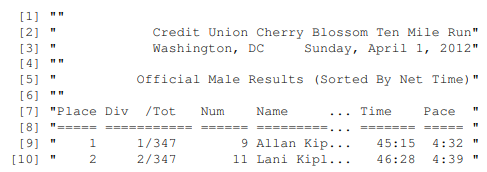
The analysis that we plan on conducting has been performed and published in the book Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving by Daniel Kaplan and Nolan. We utilized the code provided for our analysis.

For conducting our analysis of the possible relationship between age and run-rate for both the men and women runners at the Cherry Blossom 10-miles race, 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 thy are ready to be used for analysis.

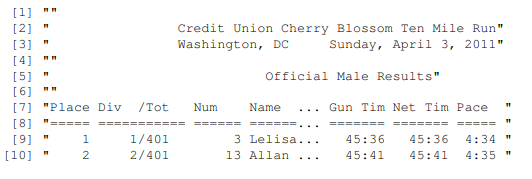
**Methods**

As mentioned earlier, we scraped the results of the races for both men and women participants from 1999-2012 from the Cherry Blossom 10-Miles race’s official website. We began by scraping men’s files and then the women’s race results files. It was observed that all files did not have the same formatting or even the same columns (that is the headers differ between years). Figures 1 and 2 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

**Figure 1: First 10 rows of Men’s Results for 2012**



**Figure 2: First 10 Rows of Men’s Results for 2011**



The women’s race results file for the years 2001, 2003, and 2006 had header missing. Headers were also missing for the men’s race results files for the years 2003 and 2006. All header issues were resolved before.

Next, men’s and women’s race results files were extracted and saved as csv files individually for separate years.

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

**Table 1: Number of Rows of Results per File per Year for Men’s Results**

|  |  |
| --- | --- |
| **Year** | **Number of Rows** |
| 1999 | 3190 |
| 2000 | 3016 |
| 2001 | 3561 |
| 2002 | 3723 |
| 2003 | 3946 |
| 2004 | 4156 |
| 2005 | 4324 |
| 2006 | 5235 |
| 2007 | 5274 |
| 2008 | 5905 |
| 2009 | 6649 |
| 2010 | 6909 |
| 2011 | 7011 |
| 2012 | 7193 |

**Table 2: Number of Rows of Results per File per Year for Women’s Results**

|  |  |
| --- | --- |
| **Year** | **Number of Rows** |
| 1999 | 2356 |
| 2000 | 2166 |
| 2001 | 2972 |
| 2002 | 3334 |
| 2003 | 3542 |
| 2004 | 3899 |
| 2005 | 4333 |
| 2006 | 5435 |
| 2007 | 5690 |
| 2008 | 6397 |
| 2009 | 8323 |
| 2010 | 8853 |
| 2011 | 9030 |
| 2012 | 9730 |

Add info on columns here after touching base with Bin. Also what was the issue with line 46 in men’s 2006 body and women’s 2006 body

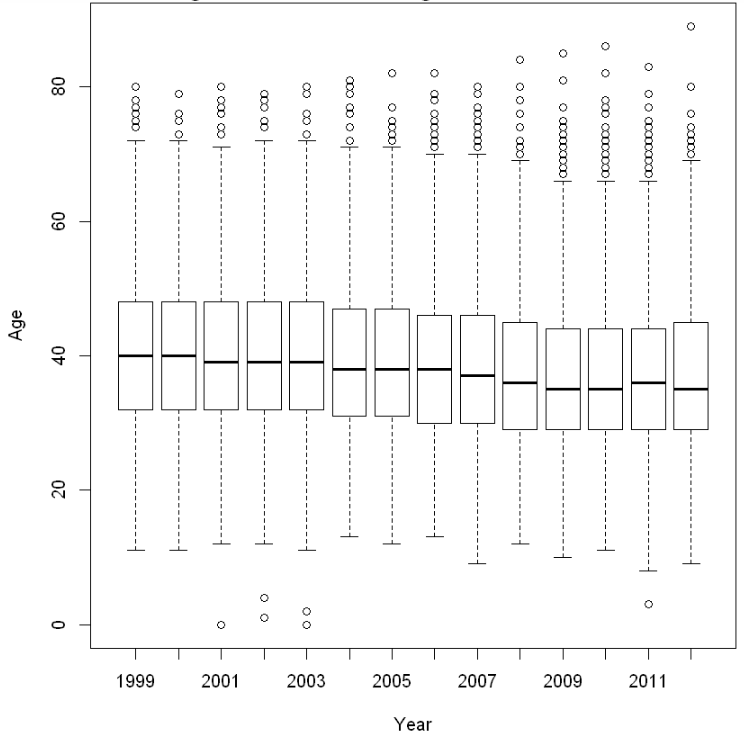
**Results**

We begin our analysis by running code to get basic statistical data. Below is a box plot for the age distribution of male runners who participated in the Cherry Blossom 10-Mile race from years 1999 to 2012.

It can be seen from the box plots that the average age of male runners for the years 1999 and 2000 is 40 years. The average age for male runners for the years 2001 through 2003 looks to be around 38 years. For the years 2004 to 2006, the average age drops to 38. For 2007, the average age is 37.5 or so. It goes down a bit in 2008 to around 36 years. The average age dips to around 35 years for 2009 and 2010. It edges up to 36 years for 2011 and is at 35 years for the year 2012.

Overall the average age of male runners is in the 35-40 year age bracket from 1999 to 2012. We see an outlier in 2012, a single runner at around 89 years. We also observe outliers denoting very young ale runners. In 2011 a runner is around 10 years. And for year 2001 and 2003 a male runner is in the 5 years or less bracket. These are

**Figure 1: Male Race Runner’s Age Distribution by Year**



References:

1. <https://en.wikipedia.org/wiki/Cherry_Blossom_Ten_Mile_Run>
2. <https://nationalcherryblossomfestival.org/about-us/>