**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**

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:

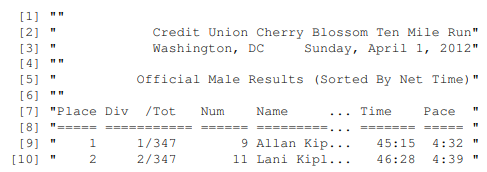
**Web Scraping Race Run Time Results**

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](http://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.

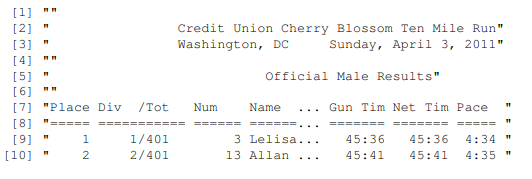
Not sure if we want to keep the below in yellow & the 2 figures

It is 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**



**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. And to apply consistency and uniformity and ease of data extraction, we convert them all to lowercase. (I don’t think we used tolower anywhere? We can delete). 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. (Understand from Bin what the issues are to write here)

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

**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 |

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.

|  |  |
| --- | --- |
| **Year** | **Number of NA** |
| 1999 | 0 |
| 2000 | 0 |
| 2001 | 0 |
| 2002 | 0 |
| 2003 | 0 |
| 2004 | 0 |
| 2005 | 0 |
| 2006 | 0 |
| 2007 | 0 |
| 2008 | 0 |
| 2009 | 0 |
| 2010 | 0 |
| 2011 | 0 |
| 2012 | 0 |

References:

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