MSDS 7333 – 406

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

Modeling Runners’ Times in the Cherry Blossom Race

Case Study

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

Martin Garcia, Michael Catalano, Jeremy Otsap, and Christian Nava

June 3, 2020

Introduction

The Cherry Blossom Ten Mile Run is a race held in Washington, D.C. each spring as the cherry blossom trees are in full bloom. The results of this race are published publicly on a website (http://www.cherryblossom.org). These race data provide a challenging candidate for web scraping and analysis due to the formatting of the text from year to year. A web scraper is faced with a host of issues when attempting to manipulate this data, including data discrepancies, input errors, formatting issues, missing headers, and inconsistent values across columns.

The bulk of this analysis is focused on gathering clean, tidy, and tabular data through web scraping and preprocessing. Next, we dive into the race results, conduct statistical analysis to understand distributions in age and race times over the years 1999-2012, and finish with a change-point analysis. The goal of this change-point analysis is to determine at what ages the changes to average run time are occurring.

Summary of Experiment

The primary tool used in this case study is R, with a host of packages to assist in the web scraping and follow-on analysis. Helper functions are created with the goal of cycling through the yearly results and keeping the reformatting process dynamic, although this poses challenges when certain race years require a tailored approach to preprocessing.

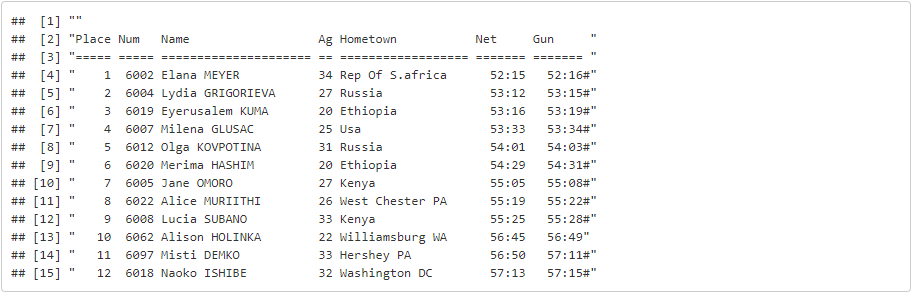
For the statistical analysis portion of the case, we conduct basic EDA and check out the distributions within the data. One specific use case requires averaging run performance across each age group and looking at the relationship between age and run times. Our change-point analysis consists of two tests, one using piecewise linear modeling for detecting a single change-point, and a non-parametric test (PELT) for detecting multiple change-points in ages by average run time.

Data Cleaning & Preparation

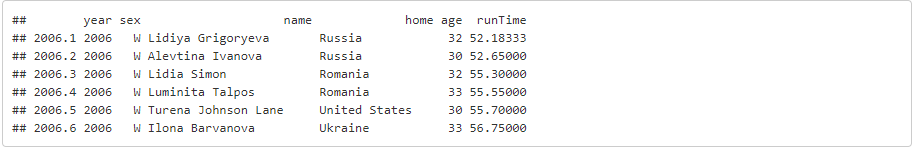
As was prevalent throughout most of this case, race year data comes in all sorts of formats. The year 1999 does not perform line breaks in a standardized fashion, which requires brute reformatting to correct. 2001 race data suffers from a lack of headers (Table 1). The fix is to copy headers from subsequent years and impute (Table 2). We notice ages with missing values, zeros, and values under seven, which are removed prior to conducting statistical summaries or applying the change-point analyses. Race time formats are sporadic across the 14 races and require conversion and remapping.

2006 poses a peculiar challenge: the extractVariables() function does not aggregate hometown and time columns due to a misaligned spacer row in the header. Fixing this issue allows the data to import properly, else all 5,432 lines are missed downstream. Table 3 shows our final data frame once all cleaning and preprocessing was finished.

*Table 1. 2001 race data as brought into R for preprocessing*



*Table 2. 2001 race data once cleaned*

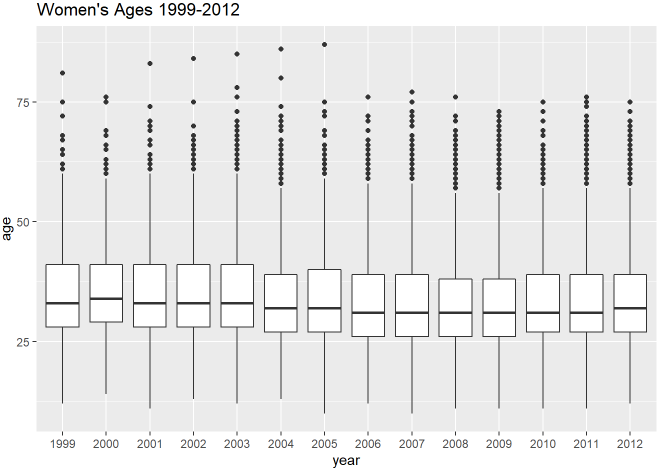


*Table 3. Finalized race result data-frame*

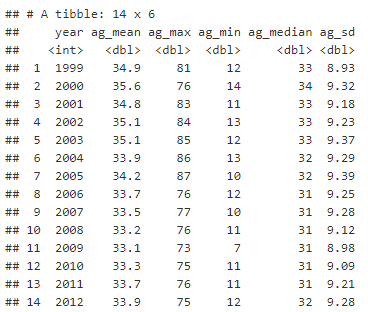
Race Analysis

*Data Distributions*

Overall distribution of ages for women from 1999-2012 using a box and whisker plot are displayed in Figure 1. Box and whisker plot are a useful way to visualize differences among samples or groups. The results of this plot point to consistent dispersion throughout the years. The median does not show drastic differences, although one could argue the race has gotten slight younger due to the lack of women over the age of 77 from 2008-2012; narrower distributions at the end of the “whiskers” during these years indicate this. The summary statistics in highlight this hypothesis by showing the average and median ages for the 14-year span of women’s races.



*Figure 1. Age distribution of women runners from 1999-2012*



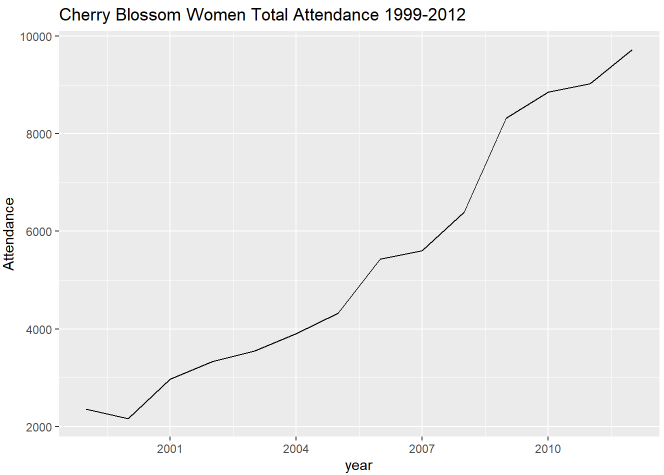
*Table 4. Summary statistics for women runners from 1999-2012*

Figure 2 below shows two views of age distribution: the graph on the left shows the overall age distribution for all available years (1999-2012). It appears to portray a right skewed distribution, with a mean age of 33.8. The graph on the right shows only two years: 1999 and 2012. Relatively similar results which reiterate, graphically, what we surmise from the summary statistics.

|  |  |
| --- | --- |
|  |  |

*Figure 2. Age distribution density plots*

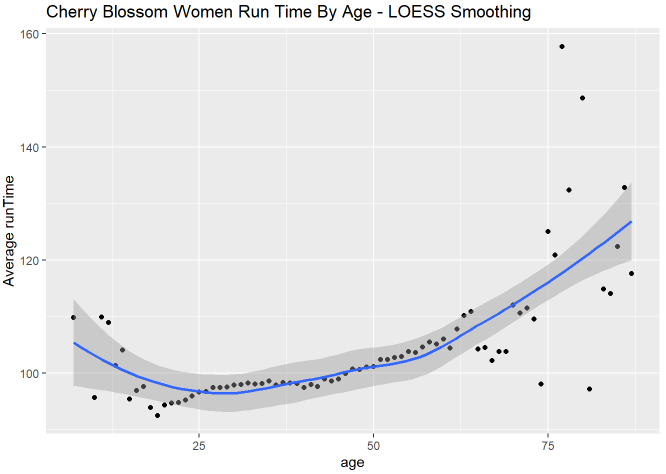
The following sets of graphs highlight attendance and average run time trends, all showing an increase from 1999 to 2012 (Figure 3 and Figure 4). In Figure 5, we pair average run time with age in a scatter plot, with the intent of visualizing the non-linear relationship. For this graph, we utilize the LOESS curve fitting method. We see some erratic behavior at the lower and higher ends of the age spectrum as denoted by the quadratic nature of the plot.



*Figure 3. Attendance by year*



*Figure 4. Annual run times by year*

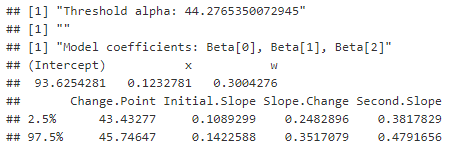


*Figure 5. Annual tun times and age scatter plot*

*Change-point Analysis*

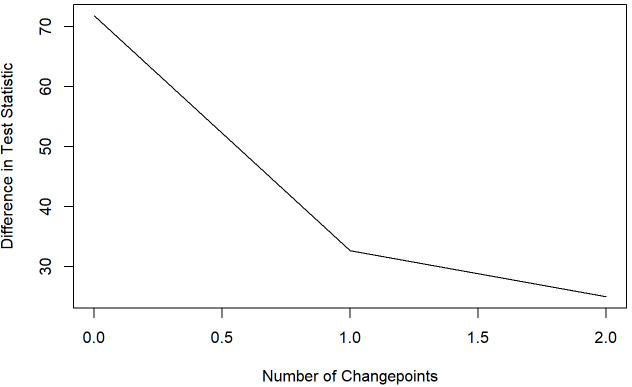
The variance in Figure 5 spurs the idea of conducting a change-point analysis. We can surmise where we think a change in variance has occurred, however we do not know for sure. The purpose of conducting this analysis is to answer the question of changing age distributions as it relates to average run time over the course of the 14-year race study. The analysis is run twice: once with the intent of using a piecewise linear function, and a second run which uses the PELT method for identifying change-points in each set of summary statistics for a specified cost function and penalty.

Figure 6 highlights the R output for the single change-point analysis. Note a possible change-point (threshold alpha) at age 44. Our model coefficients are also shown, as well as a confidence interval of our change-point (43.4 - 45.7). Important to note here: zero is not included in any of our confidence intervals; there appears to be evidence of a change-point in average run time at age 44.

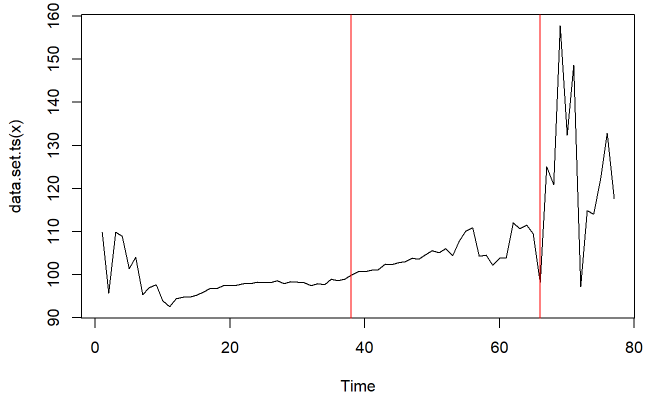


*Figure 6. Single change-point analysis results*

In reality, there may be multiple change-points in this race dataset. To that end, we utilize the Pruned Exact Linear Time (PELT) method to search the solution space in the most efficient manner. We use the CROPS penalty in conjunction with the PELT method for this scenario. This method utilizes a non-parametric cost function based on the empirical distribution of the data. The diagnostic plot (Figure 7) gives us an idea of how many change-points are evident (the point on the elbow corresponding to the lowest difference in test statistic). Finally, a new plot (Figure 8) shows the same group age and average run time data as well as the locations of our estimated change-points (ages 38 and 66).



*Figure 7. Diagnostic plot for number of change-points*



*Figure 8. Multiple change-point results using PELT method*

Conclusion

While the crux of the case relies heavily on web scraping and reformatting of disparate race data from the annual Cherry Blossom Race, we can glean some insights from the distribution of data. Yearly age ranges have remained relatively consistent from the boxplot visualization (Figure 1). The number of attendants and average run times of the race have steadily increased in the 14-year span of available data. The relationship between age and average run time is quadratic, with higher times at both ends of the age spectrum, and we have derived change-points for age as it relates to average run times, which happen to be at ages 38 and 66.

*Drawbacks of methods employed and recommendations for improvements*

**Web scraping methodology**

The R method for web scraping and parsing, while painstakingly manual, proved sufficient. An additional technique would have been to use the beautiful soup package in Python.

**Additional analysis**

Deeper dives into the data could be undertaken; constructing a record for each individual runner across all race years might prove useful, especially for modelling the change in running time on an individual basis as each runner ages. This might have required additional cleaning on the town variable for detecting the same runners from different towns, or duplicated runner names.