**Case 2: Cherry Blossom 10 Miler**

**Introduction and Overview**

Cherry Blossom 10 Mile race in DC – one of most popular, part of Professional Road Running Organization circuit. In years from 1999 to 2012 the race has seen [WHAT SORT OF GROWTH] in participation, where women have [WHAT] and men have [WHAT]. The annual event has also grown in complexity, with seeding of elite runners, a fundraising opportunity that guarantees a lottery system to grant entry to runners, a team entry system, and a lottery that assigns additional entry.

The city of Washington DC has put pressure on Cherry Blossom race organizers to limit the time of the race. Given the other Cherry Blossom activity happening around the same time, it has become more difficult to keep the road closed for the many hours needed to set up and run the race. As a result, the race needs to be shorter in time with less participants.

The race committee has asked our team to assess the past 13 years of race results to understand how the race has grown, how the distribution of the age of runners has changed, how the pace of our runners has changed, to ultimately to suggest an option for how to shrink the race. The 2012 race had X participants, who had to finish within 2 hours and 20 minutes (2020 requirement), and we are looking to recommend a new set of qualifying times that reflects the age and pace of past participants while also reducing the overall number of runners.

**Data Preparation (Q7)**

The team collected data by web scrapping information from the Cherry Blossoms 10 Miler website. The results for female runners were obtained by accessing each year from 1999 to 2012 in the directory found at: <http://www.cherryblossom.org/results>. We found that several years of data were stored in slightly different formats (i.e. wider page headers, variable spaces between lines of data) and had different names (i.e. time vs net time vs gun time) for the same attribute. For this reason, we had to implement a function with different internal procedures for each year to import the data.

*Data Scraping and Parsing*

A comprehensive description of the steps we took to scrape and parse data from the Cherry Blossom website can be found in our respective R and Python notebooks. See CODE DIRECTORY at end of write up. The following are highlights of issues encountered by the team.

Women’s Data

* For most years the data is fairly consistent but there are anomalies that exist that required additional specialized treatment for some years.
* For the 1999 data the //pre node worked while for 2000 we had to use the //font node to obtain the data. Furthermore, for 1999 the linefeed carriage return character “\r” was missing so the newline character “\n” was used instead.
* Additionally, the header line “====” is not present for all years making it rather difficult to find the header and the spacer columns.  This required manual intervention to establish the variable lengths. This was the case for years 2001 and 2002.
* Remove leading and trailing whitespaces in data (using trimws in R)

Men’s Data

Similar issues existed for the men’s data as the women’s data with some exceptions. The biggest challenge in scraping the men’s data was year 2009. It required additional work to effectively scrape the data. Essentially, 2009 data was formatted in a Word-like document where the end of each data line was indicated by </pre><pre>. This required a separate function to scrape and parse the data.

*Data Cleaning*

Our initial review of our data once scraped and organized revealed additional inconsistencies and missing data. The team invested time to cross reference participants with missing ages against the “Searchable Results” page on the Cherry Blossom website; information was updated or marked as missing for more than 85 records. We also reviewed times for participants under the age of 12; while times for some of these young runners may be questionable, the information published was taken at face value, with the assumption that they would not significantly impact the analysis.

Additional data cleaning and preparation included:

* Runners with no time were dropped from the data set.
* Used “Time” when available; for runners with multiple time entries, used “Net Time” then “Gun Time” then “Comb Time” for time variable.
* Created age divisions in 5-year age brackets (with exception of under 19 and above 80); this is consistent with the Cherry Blossom race divisions in 2012; also created 10-year increments if needed.
* Created variables to assist in managing hours, minutes, seconds, and milliseconds of time variable; used this information to calculate pace mathematically (as pace from Cherry Blossom website had errors).

**Analysis (Q10)**

*WOMEN*

Box Plots

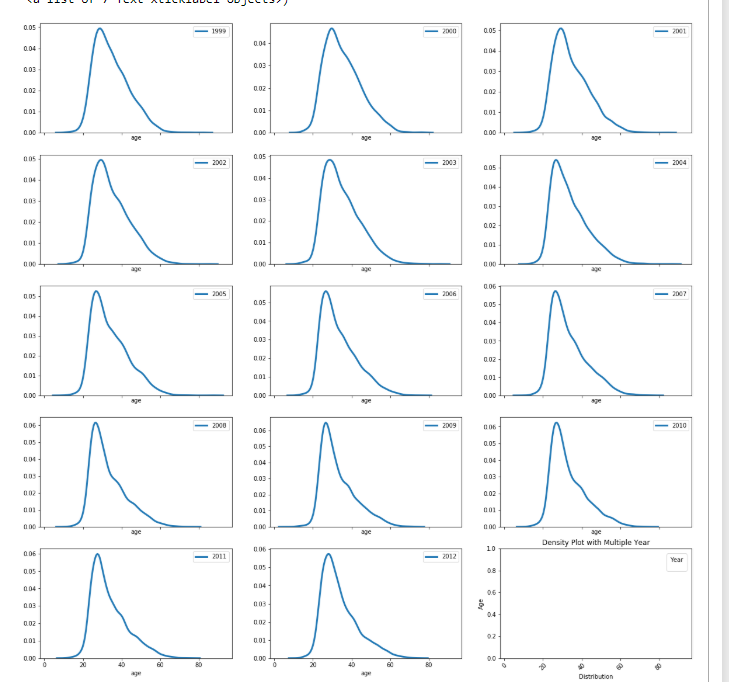
|  |  |
| --- | --- |
| AGE | TIME |
|  |  |

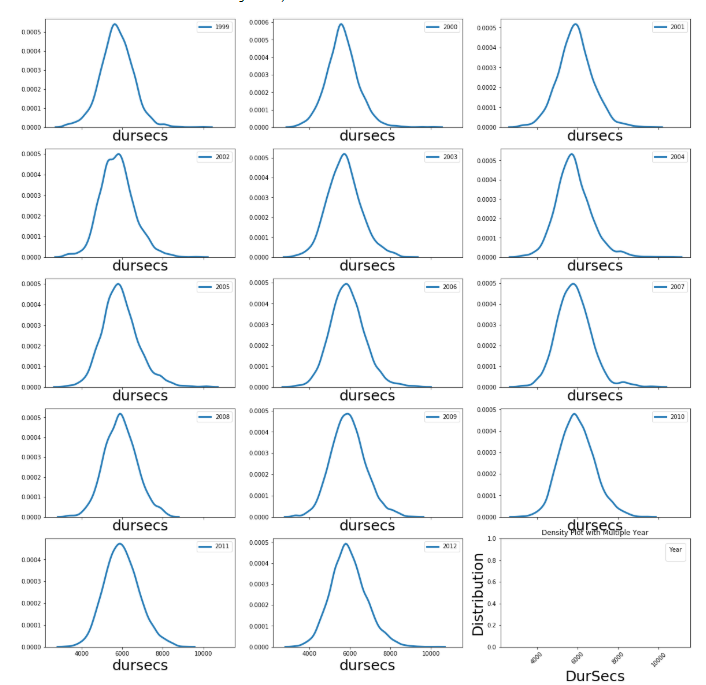
QQ Plots

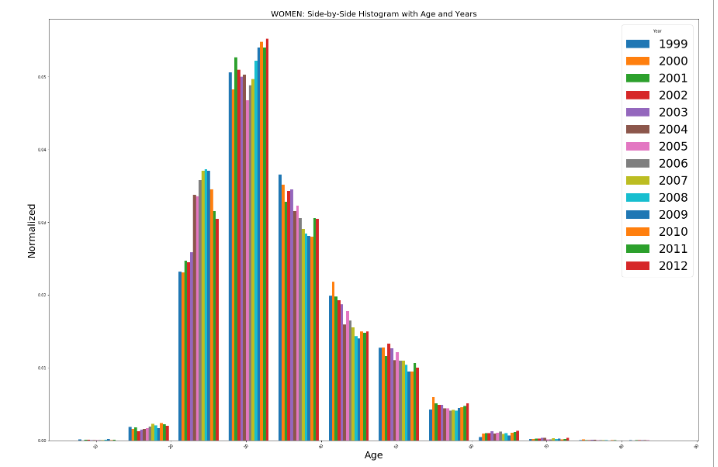
|  |  |
| --- | --- |
| AGE | TIME |
|  |  |

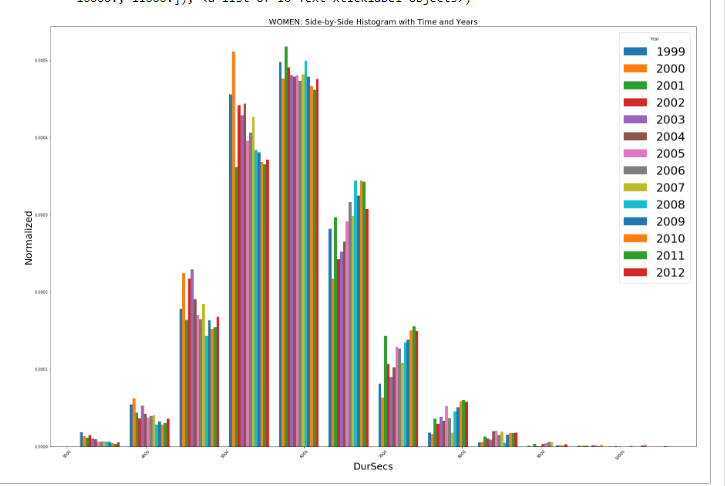
Density Plots

|  |  |
| --- | --- |
| AGE | TIME |
|  |  |

Density Plot – Age, Year by Year Detail

Density Plot – Time, Year by Year Detail

Histograms – Age and Years

Histograms – Time and Years

Overall Age Distribution

First, conducted a basic visual analysis of box plot series showing age distribution. [WHAT DID WE FIND]

Then, dug into age distribution more, [WHAT DID WE FIND].

* QQ plot here? … for each year, by age and gender

Next,

* Density or violin curves … Combo Violin Density

Pink Series of Charts:

Density is higher, and they show that there are differences. 1999 is higher, 2010 is lower, is this telling us they are not equal? The tails are much further apart…the spread is not as wide.

Lastly, conducted statistical assessment / change point analysis / what other method to use – t-test

CONCLUSION – the distributions [did/did not] change. The change was [gradual/sudden/non-existent].

Overall Time by Age Distribution

First, conducted a basic visual analysis of box plot series showing age distribution as related to overall race time. [WHAT DID WE FIND]

Then, dug into / age time distribution more, [WHAT DID WE FIND].

* QQ plot here?

Next,

* Density or violin curves

Lastly, conducted statistical assessment / change point analysis / what other method to use – t-test

CONCLUSION – the distributions [did/did not] change. The change was [gradual/sudden/non-existent].

*MEN*

First, conducted a basic visual analysis of box plot series showing age distribution. [WHAT DID WE FIND]

Then, dug into age distribution more, [WHAT DID WE FIND].

* QQ plot here?

Next,

* Density or violin curves

Lastly, conducted statistical assessment / change point analysis / what other method to use – t-test

CONCLUSION – the distributions [did/did not] change. The change was [gradual/sudden/non-existent].

Overall Time by Age Distribution

First, conducted a basic visual analysis of box plot series showing age time distribution. [WHAT DID WE FIND]

Then, dug into age distribution more, [WHAT DID WE FIND].

* QQ plot here?

Next,

* Density or violin curves

Lastly, conducted statistical assessment / change point analysis / what other method to use – t-test

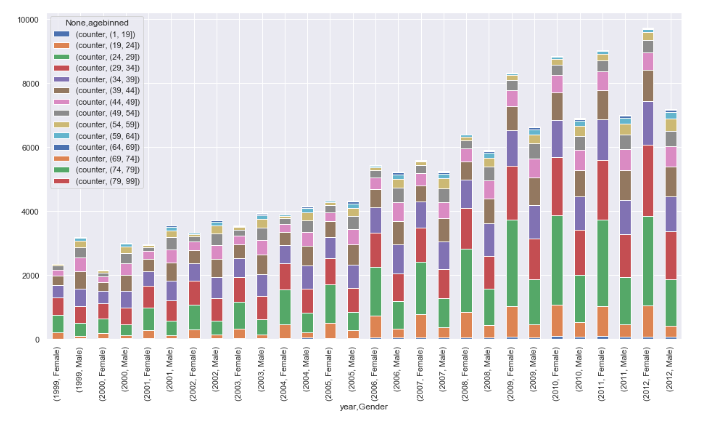
CONCLUSION – the distributions [did/did not] change. The change was [gradual/sudden/non-existent].

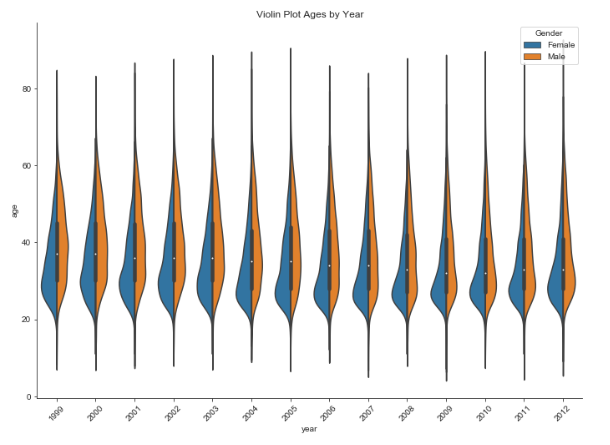
*COMPARISON OF MEN VS WOMEN*

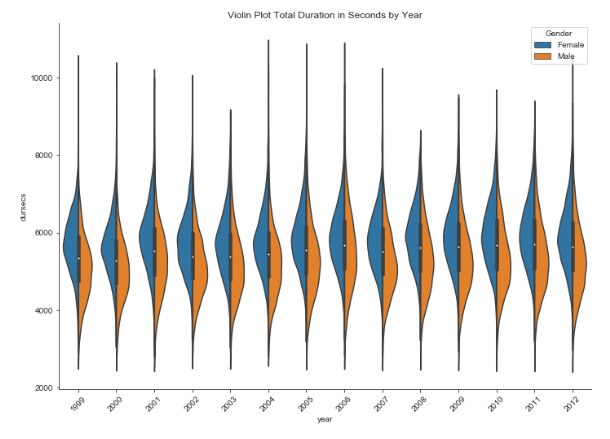
*Difference between duration ran for men and women*

*Difference in age between men and women*

*Do a test of medians for age and paces?? Simple regression – is it a positive or negative relationship*







**Recommendation and Conclusion**

Conclusion.

The age distribution of men and the age distribution of women [BLAH BLAH BLAH]. This indicates [WHAT]? What do we think about this? How can this help make the race shorter?

Recommendation

Propose new set of times for age groups to target

Take off what %

\*\* Green charts – regression fitted line, i.e. your time vs age, slopes of lines are mostly the same. From a straight regression perspective, people are just getting slightly older.

\*\* attempted Change Point, with Loess and Change Point

\*\* R change point model – converted from Python. Not super helpful

\*\* Python does not have elbow

\*\* Regression line from Age to Time, straight linear regression model – are there different slopes for different age groups?

\*\* LASSO Path – should be similar to piecewise, different slopes for different buckets.

\*\* Extra testing are t-test … Sklearn Lars Path, looks similar to the Change Point …. Is duration changing over time?

\*\* Six chart with the colors …

\*\* Violin plots – with Men and Women, men is slightly flatter than women, spread on men’s curve is wider, the variation on the women’s side is a little less, and men are more spread out. Looked at normalcy – the variance between the groups are different…

\*\* the t-test is too perfect. 0.001 are the average age of women equal to men;

\*\* overall, is there change from

\*\* in 2012 are younger people not

\*\* Side by Side Histogram…looks at age against the total.

**CODE DIRECTORY**

**APPENDIX**

|  |  |
| --- | --- |
| Women |  |
| Men |  |
| Both |  |

Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Action Taken** |
| Gender | Male or Female; gender of participants; scraped from individual Cherry Blossom website pages | Scraped, no cleaning necessary, data label |
| Year | 1999 to 2012; scraped from individual Cherry Blossom website pages | Scraped, no cleaning necessary, data label |
| Place | Place in each year’s race, separated by Gender and Year | Scraped, no cleaning necessary, did not use |
| Div\_Total | Place of Each Participant in Age Division; for each year | Scraped, no cleaning necessary, did not use |
| Name | Participant Name | Scraped, no cleaning necessary, did not use individual names |
| Age | Age of Each Participant | Scraped, cleaned up missing or outlier ages |
| Hometown | Home of Each Participants, either City ST or Country | Scraped, no cleaning, did not use |
| Time | One of: Time, Gun Time, Net Time; or Comb Time; Time – Overall Participant Time; Gun Time – Time of Participant from Gun to Individual Finish; Net Time – Time of Participant from Start Line to Finish Line. | Scraped, needed to assess which times provided; used in order – Net Time, Gun Time, or Comb Time; type depended on race year |
| Pace | Average Mile Per Hour for Each Participant | Scraped, discovered odd values; did not use |
| Num ID | Cherry Blossom Participant ID, not available for all years | Scraped, no cleaning necessary, did not use |
| Net Time | Net Time for Each Participant, where available | Scraped, where available |
| Comb Time | Combined Time of Each Participant, usually reflects Gun Time | Scraped, where available |
| Time Length | Created Variable to help calculate time | Created to assist in managing time format for calculations |
| F Combine | Created Variable | Created to assist in managing time format for calculations |
| Final Time | Created Variable | Created to assist in managing time format for calculations |
| Hour | Created Variable to separate time: hours | Created to assist in managing time format for calculations |
| Minutes | Created Variable to separate time: minutes | Created to assist in managing time format for calculations |
| Seconds | Created Variable to separate time: seconds | Created to assist in managing time format for calculations |
| Dur Secs | Calculated Variable; Total Race Time in Seconds | Created to assist in managing time format for calculations |
| Duration Min | Calculated Variable; Total Race Time in Minutes | Created to assist in managing time format for calculations |
| Calc Pace | Calculated Pace from Time | Created to assist in managing time format for calculations |
| Counter | Created Variable; tally of 1 for each participant | Created to assist in managing participant count |