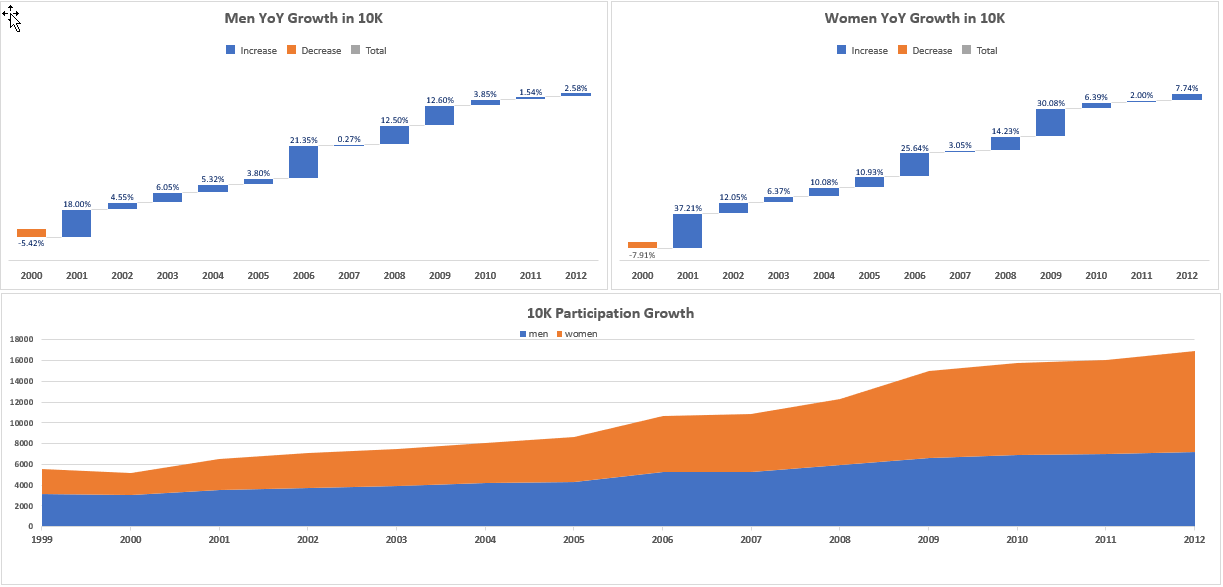
**Case 2: Cherry Blossom 10 Miler**

**Introduction and Overview**

The Cherry Blossom 10 Mile race in DC is one of the most popular races in the areas, and it is part of Professional Road Running Organization circuit. In the years from 1999 to 2012, the race has seen growth in the number of both female and male runners. While there was a slight dip in participation between in 2000, the year over year growth has been as high as 25-30%, as seen in **Figure 1**. Overall, the Cherry Blossom 10 miler has seen a greater increase in women runners (314% from 1999 to 2012), compared to a smaller increase of male runners (126% over the same time period). The bottom graph of **Figure 1** shows a larger proportion of women over the years, and by 2012, the race had growth to nearly 17,000 participants.

**Figure 1: Year over Year Participation by Gender**

The annual event has also grown in complexity, with the seeding of elite runners for guaranteed entry, fundraising opportunities to guarantee a race spot for non-seeded runners, a team entry system and competition, and a lottery system to award entry to additional runners. In 2012, race times were well past two hours, plus the many hours needed for set up and tear down. 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 activities happening around the same time, it has become more difficult to keep the road closed that length of time. Therefore, the race needs to take less time in the years ahead.

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 age distribution of runners has changed, how the pace of runners has changed, and ultimately, to [RECOMMEND WHAT].

**Data Preparation (Q7)**

The team collected data by web scrapping information from the Cherry Blossoms 10 Miler website. The results from 1999 to 2012 for both female and male runners were obtained by accessing each individual year’s website 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 for the same attribute (i.e. time vs net time vs gun time to represent the total race time for each runner).

*Data Scraping and Parsing*

Our team designed and implemented a function that accommodated these format variations to effectively scrape the data from the website. The comprehensive methods we employed can be found in the respective R code and Python notebooks. See **Code Directory** at the conclusion of this 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 and substituting unwanted characters using **gsub** to replace extraneous characters and spaces.

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 for year 2009. It required additional work to effectively scrape the data and then to structure, clean and transform the data as appropriate.
* Essentially, 2009 data was formatted in a Word-like document on the website where the end of each data line was indicated by **</pre><pre>.** This required a separate function to scrape, parse, clean and transform the data.

*Data Cleaning*

After compiling all the data, our initial review revealed additional inconsistencies and missing data. The team invested time to find missing ages by cross referencing against the “Searchable Results” page on the website; information was updated or marked as missing (NA) for more than 85 records. We also specifically 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:

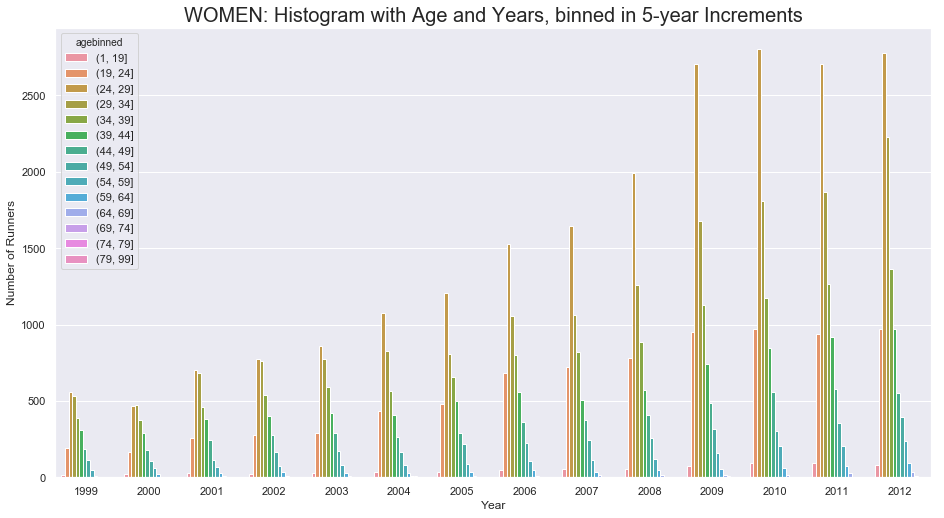
* Runners without race time were dropped from the data set.
* Race time was presented as “Time,” “Net Time,” or “Gun Time,” depending on race year. For consistency in analysis, we used “Net Time” when available, followed by “Gun Time” or “Time.” Details of these definitions can be found in **Appendix Figure A**.
* Participations were binned into 5-year age brackets (with the exception of under 19 and above 80); this is consistent with the Cherry Blossom race divisions in 2012. We also created 10-year increments if needed.
* We created variables to assist in managing race time, given the variation in hours, minutes, seconds, and milliseconds of the time variable. Race pace was calculated mathematically (as the Cherry Blossom website had errors).

**Analysis (Q10)**

With the objective to understand how participants’ age and race time have changed over the years, the team first looked at women and men separately.

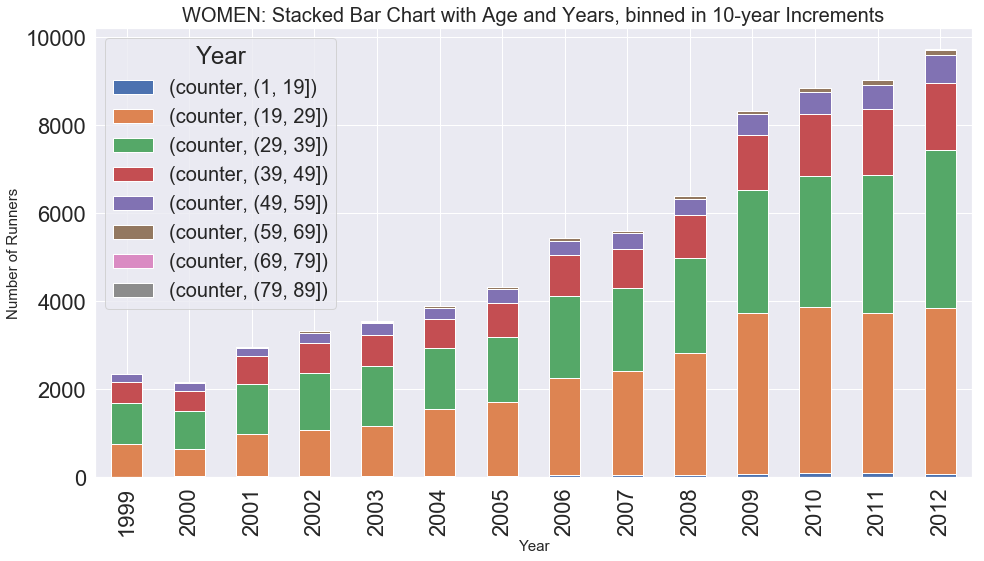
*WOMEN*

The first assessment by the team was to look at how the women in each of the 5-year age bins changed over time. As seen in **Figure 2**, the growth of women participants is most notable in the 24-29 and 30-34 age groups, while the other age categories grew more proportionally.

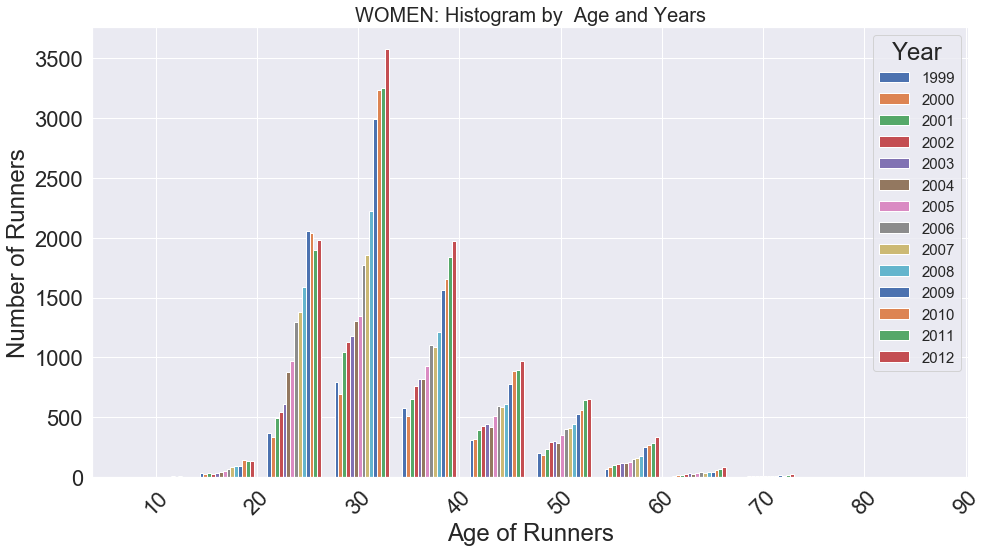


**Figure 2: 5-year Increment Age Bins by Year**

Next, we conducted a visual analysis of the 10-year age bins against year to see how the proportions of each changed over time. In **Figure 3**, the largest changes are seen in the 19-29 and 29-39 categories. Starting in 2009, this 20-year age span reflected a significant percentage of female participants. [INCLUDE HOW MUCH?]

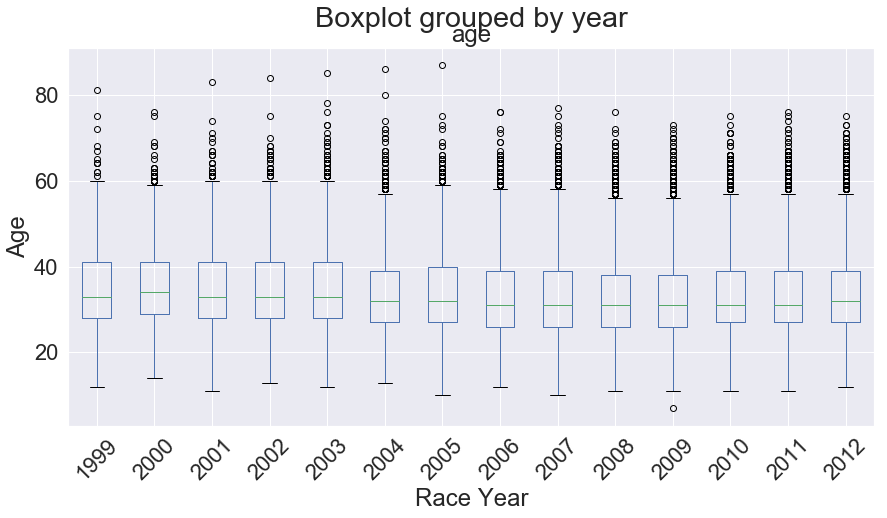
**Figure 3: 10-Year Age Bins by Year**

We looked at the same information presented another way: the number of Runners Group by Age for each year. **Figure 4** illustrates the significant increase in runners roughly age 20-40.



**Figure 4: Runner Count by Age Groups Each Year**

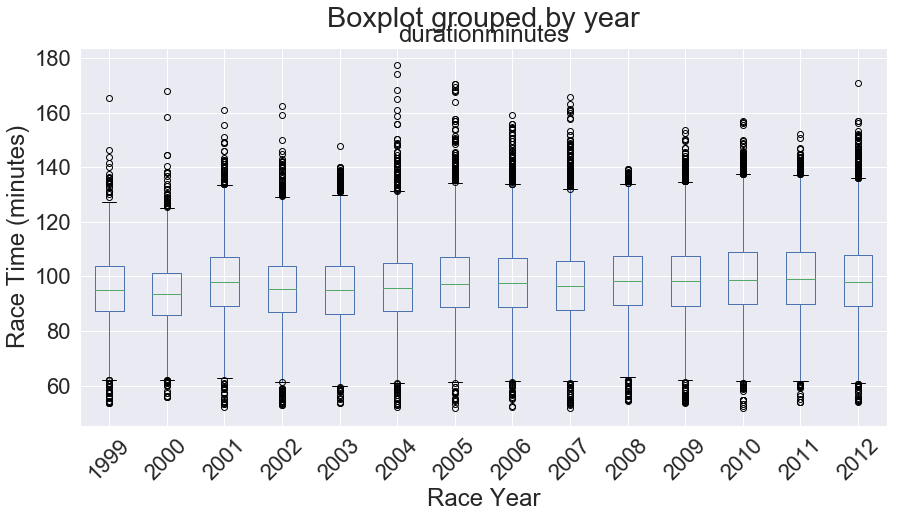
Then, we looked at two boxplots that compared women’s age by year (**Figure 5**) and race time by year (**Figure 6**). The boxplots provided better insights into the trends for age and time. The red dotted line is equal to the median age in 2000, which saw on average the oldest runners. The median age decreases slightly year over year, with a noticeable drop in 2006. There were no late 70 or 80+ aged runners in 2006, which also might contribute to a change in average age. Overall, there are not dramatic shifts in the distribution of the plots but certainly a younger trend at all quartiles.

****

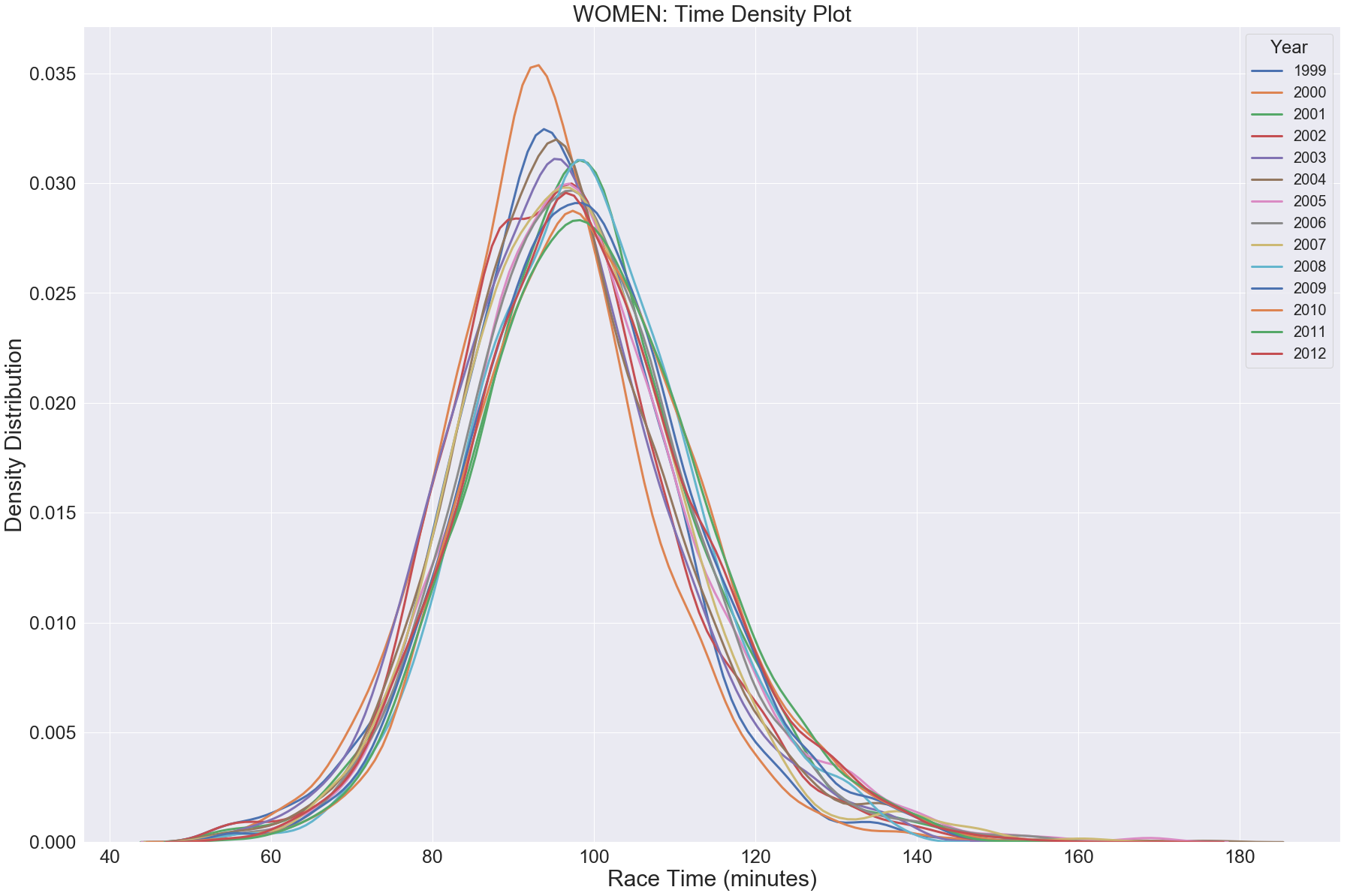
**Figure 5: Boxplot of Year and Age**

The second boxplot the team analyzed was a comparison of race time in minutes over the years to see if changes are occurring with the speed of female runners. The red dotted line reflects the median of the slowest year’s race – 2001 [WAS THIS ACTUALY SLOWEST YEAR]– although 2008 to 2012 are right in line with the median in 2001. It also does not appear that the women are getting faster over the years; in contrast, there is some slight increase in race time among the upper quartiles. However, the general trend seems to reflect roughly consistent pace among female runners (with the exception of a few slow outliers) as seen in **Figure 6**.

To assess speed year over year, the team also looked at a density plot of female race times. **Figure 7** [SHOWS WHAT]. [THIS IS WHERE I STOPPED]



**Figure 6: Boxplot of Year and Race Time**



**Figure 7: Density Plot of Time**

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.

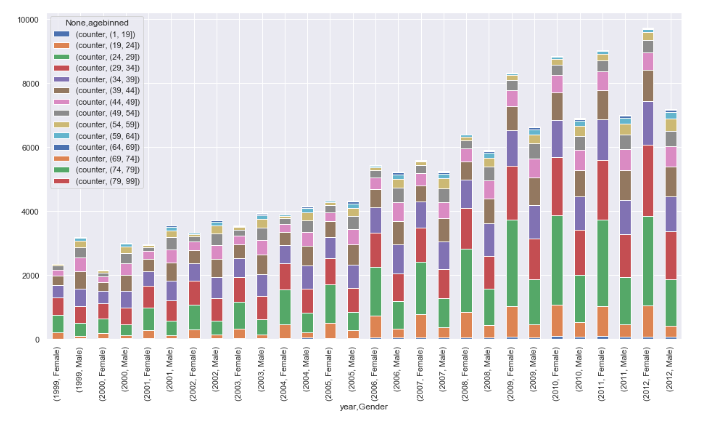
Lastly, conducted statistical assessment / change point analysis / what other method to use – t-test

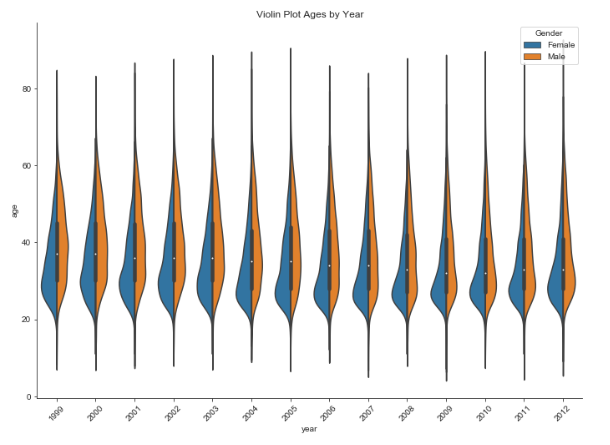
*COMPARISON OF MEN VS WOMEN*

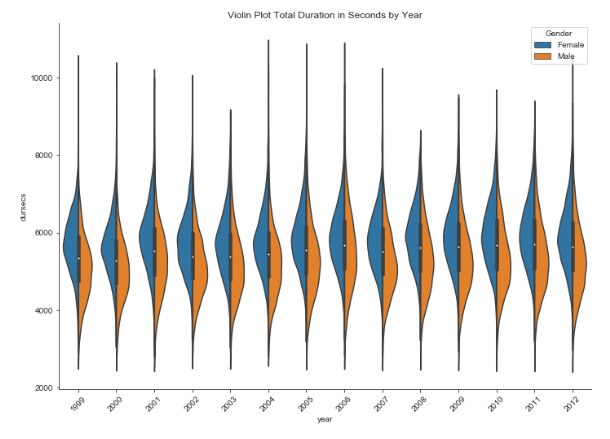
*Difference between duration ran for men and women*

*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**

**Appendix Figure A:**

**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 |