**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.

**\*\* cut off times …**

**\*\* also**

**Joe – 2009 data**

**Steven – flesh out the t-test, confidence internal, try one sided**

**Data Preparation (Q7)**

**Q.7** Follow the approach developed in Section 2.2 to read the files for the **female** runners and then process them using the functions in Section 2.3 to create a data frame for analysis. You may need to generalize the createDF() and extractVariables() functions to handle additional oddities in the raw text files.

The team collected data by web scrapping information from the Cheery 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. [NOTE A FEW] and had different names (i.e. [NOTE A FEW] 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 Parsing*

* Cross referenced missing ages with Searchable results on Cherry Blossom site to enter any ages found (43 looked up, some found/some NA) 33 under age of 12 – reviewed those as well;
* Needed to spend a good amount of time on 2009 Men’s data, required much more cleaning and parsing; appeared that original layout was in Word. End of line </pre><pre>

*Data Dictionary*

* *­*Gun time vs chip time vs comb time vs
* Gender, Year, Name, Div\_Total, Age, hometown, time, pace, numId, guntime, nettime, combtime.

*Assumptions*

* Runners with no time are dropped.
* Use the 5 year intervals, lumped see the Box plots.
* To accommodate changes in age divisions over the years, team used age divisions from 2012, which corresponded to 5-year age brackets (with exception of under 19 and 80+) Also have age by 10 year increments as well.
* What time to use: USTAF only uses Gun Time / From Cherry Blossom: “With the wave start all runners except those finishing among the top 25 men and women will be scored and placed based on net times. The top 25 men and women will be scored using gun times.”
* Ages 14 and under for 10 Miler may be questionable, but took data at face value; recognize that the younger ages might not make difference in race

We used what time – nettime, guntime, or combin time.

We calculated the pace of runners with their time and distance, did not use the Pace from Cherry Blossom site as there were too many error.

**Analysis**

*WOMEN*

**Q.10** We have seen that the 1999 runners were typically older than the 2012 runners. Compare the age distribution of the **[female]** runners across all 14 years of the races. Use quantile–quantile plots, boxplots, and density curves to make your comparisons. How do the distributions change over the years? Was it a gradual change?

\*\* Slight change, confirm with test stastic

\*\* is age getting longer or is time getting longer

\*\* participants by year bar chart with linear growth, and the age distribution,

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*

* Men vs Women Violin Plots
* Men vs Women Density Plots

**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.

**Goal of project: Can we set qualifying time to limit capacity for men and women**

* Question to ask
  + **Has there been a change in age distribution?**
  + How many age groups
  + Avg time per Age group
    - Avg time per mile
  + ~~Avg time per country~~
  + Compare Men vs Women
  + ~~Participation % that did not finish~~
* ~~“Group by” to find people who have ran all races.~~
* Group by age and get the mean time
* How many people are getting to the hundred mile club

**MISC NOTES ABOUT CSV FILE**

1999:

* Duplicate Records, NEED TO ADDRESS
* NEED TO EVALUATE times that are Min : Sec : Milisecond vs Hour : Min : Sec, did those work correctly?
* 4 ages are blank, DO WE WANT TO IMPUTE? We can look at the participant database to see if we can get/calculate them
* City and State together, Country abbrev

2000:

No data, NEED TO PULL

2001:

* No pace, only net time, could calculate pace if we need it/want it
* Gun time = Comb time < net time THIS IS WEIRD, NO OTHER ONE IS LIKE THIS, I’m researching it more
* Min : Sec : Milisecond vs Hour : Min : Sec … IS THIS CORRECT?
* One “0” age, we can fix that.

2002:

* Gun time = comb time > net time
* 4 ages to impute – DO THIS
* City State together, country abbrev

2003:

* Gum time = combTime > net time
* No age blanks
* CHECK Min : Sec : Milisec vs Hour : Min : Sec
* City State combined, Country Abbrev

2004:

* Gun time = combtime > Net Time
* No age blanks
* City St combined, Country abbrev

2005:

* Gun time = combin time > net time
* 8 ages to clean up; 8 missing
* City weirdness where we lost the first character
* Select times have # or \*
* City state together, Country spelled out

2006:

* First year with pace
* Gun time = combtime > net time
* Select times with \* or #
* Lost first character of the City, no state
* 1 missing age

2007:

* Gun time = combtime, no net time
* City State together, Country spelled out
* Min : Sec : Milisecond vs Hour : Min : Sec
* 2 ages missing

2008:

* Gun time = combtim, no net time
* City state together, Country spelled out
* Min : Sec : Milisecond vs Hour : Min : Sec
* Paces are weird
* No ages missing

2009:

* Age 7? Blank ages = 2
* Gun time = combtime > netTime
* Check Min : Sec : Milisecond vs Hour : Min : Sec

2010:

* Gun time = Combtime > net Time
* No age missing
* City State

2011:

* One row off: Marie-Laure Poir
* Paces are weird
* Gum time = CombTime > netTime
* No missing ages

2012:

* Gun time = Comb Time, no net Time
* Paces are weird