**Case 2: Cherry Blossom 10 Miler**

**Introduction and Overview**

**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*

*Data Dictionary*

* *­*Gun time vs chip time vs comb time vs

*Assumptions*

* Runners with no time are dropped.
* Anybody not from the US will be lumped into non-US bucket
* 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+)
* 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.”

**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?

Overall Age Distribution

* Visual analysis of box plot
* QQ Plot
* Density Curves
* Violin Plots
* Change Point analysis

Overall Time by Age Distribution

* What is best way to do this? Series of box plots?

*MEN*

Overall Age Distribution

* Visual analysis of box plot
* QQ Plot
* Density Curves
* Violin Plots
* Change Point analysis

Overall Time by Age Distribution

*COMPARISON OF MEN VS WOMEN*

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

**Recommendation and Conclusion**

Conclusion

Recommendation:

**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