**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 dada by web scrapping data from the Cheery Blossoms organization’s website. These results were obtained by accessing each years from <http://www.cherryblossom.org/results> URL directory. Several years were stored in different formats and had different names for the same attribute. For that reason, we had to implement a function that had different sub procedures for each year to import the data.

**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?

**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

Assumptions:

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