Exploring the Factors Influencing the 10 Mile Run Race Times

--- Data Analysis based on Cherry Blossom 10 Mile Run 1999-2010 data

The Cherry Blossom 10 Mile Run is an annual 10 Mile Race happening in Washington D.C. From this event's official page, people can download the Race results starting from 1999 both for women and men. Although the documents' formats are slightly different among the years, after several trials, it is still manageable to read the documents, for years 1999 through 2010, into a structured data frame in R. Such data set contains more than 100,000 observations and helps in providing some insights upon which factors may influence one's 10 mile run race times. Since the variables that being provided vary in different years, only the ones that each documents have in common and being considered as informative have been used for analysis. What follows is the list of those variables and the questions related.

- **Race Time** This variable takes the names as "Time", "Gun Tim" or "Gun" in the documents, which is the official time that used for determining one's place in the race. Although net time has been suggested to be a better indicator for the real time that one spends for the race[1], in some old files there is no record for this variable and makes it impossible for comparison. The Race Time has been transformed from its old hh:mm:ss format into measuring by minutes for the sake of analysis. In this study the Race Time is the major response variable and possible transformation may apply for stabilize its variance. In total, there are 113,078 cases which having the Race Time in the data.

- **Year** As indicated in the data resources, we have twelve years' data from 1999 to 2001. The first question that raised is whether the race performance becomes better over the years. A surprising results will be posted in the first part of the data analysis.

- **Gender** For each year, we have the data both for women and men race runners. Does gender really matter for the race pace? From the following first analysis, one can see the gender's influence on overall yearly race performance.

- **Age** One may assume that people have a golden age for running. Is it true? The second part of the analysis will focus on whether it is a good idea to fit a linear model for Age and Race Time or not.

- **Frequency** By applying an identifying criteria for each runner by using their name, birth year, gender and hometown, one can create an ID for each runner and count how many times they participate The Cherry Blossom 10 Mile Run. This frequency variable indicates one's passion for the race. The third part of the analysis will focus on the influence of frequency on one's performance by also combing the effects of age.

- **Experience** After getting the overall participation times, one can also count how many times the runner has participated The Cherry Blossom 10 Mile Run for each year. This variable is different with frequency since it's more concerned with one's running experience. The third part of the analysis will also do an analysis of the influence of experience and compare the results with the influence of frequency.

[1] In a large race, it may take up to ten minutes before a participant actually crosses the starting line. Resource: http://blog.atlasrfidstore.com/race-chip-timing-vs-gun-timing#sthash.P5imFHYf.dpuf

- **State** Since the hometown data are inputted in different manners among the years, which makes it really hard for a sufficient analysis. In the last part of the data analysis, an exploratory analysis based on different state's performance will be provided.

- **Place** It is a variable assisting in limiting the scope of the data analysis. For example, to analyze each state's race performance, one can focus on the top 500 runners.

Transformation of Response Variable: Race Time

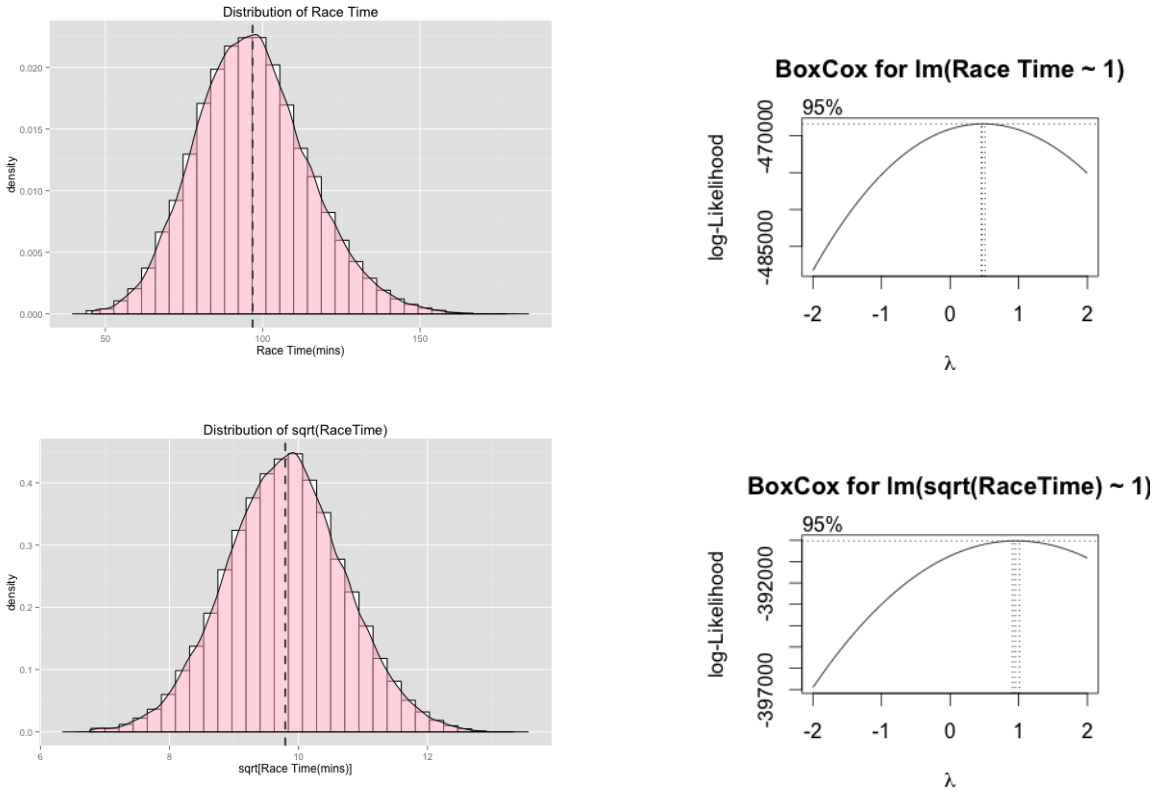


Figure 1: Transformation of Response Variable

The distribution of Race Time has been provided at the top left of Figure 1, which is slightly positive skewed. This indicates there exists some extreme long Race Time. The BoxCox function suggests that taking the square root of Race Time could stabilize the variance and make the distribution more normal. This transformation is confirmed by the distribution plot of Sqrt(Race Time), which located at the bottom left of Figure 1. So in the following data analysis, the response variable will take the form of Sqrt(Race Time).

Data Analysis I: Race Performance over Years

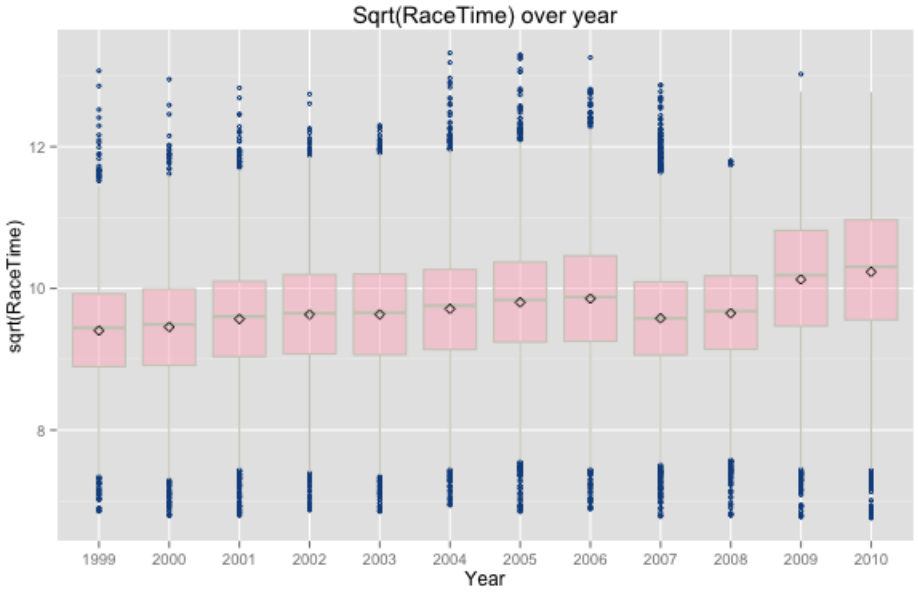


Figure 2: Race Performance (Sqrt(Race Time)) over Years

It was at first assumed that the race performance becomes better over the years. However, this assumption is truncated by the Box-plots for each year's overall race performance shown in Figure 2. It looks that there is actually an upward trend for the race time spent by the runners, except for the case of year 2007 and 2008. Three hypotheses have been suggested for explaining this phenomenon due to the increasing popularity of the Cherry Blossom 10 Mile Run.

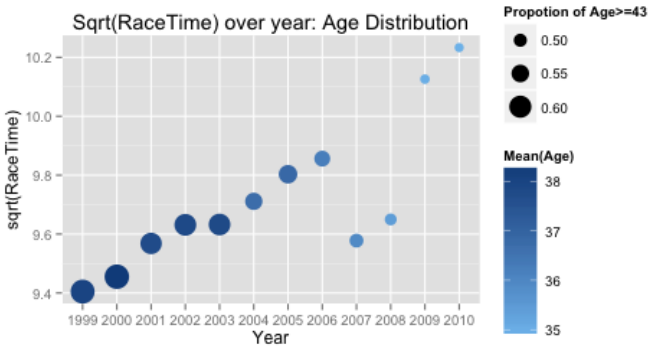
(1) There are more elder people involved in the race;

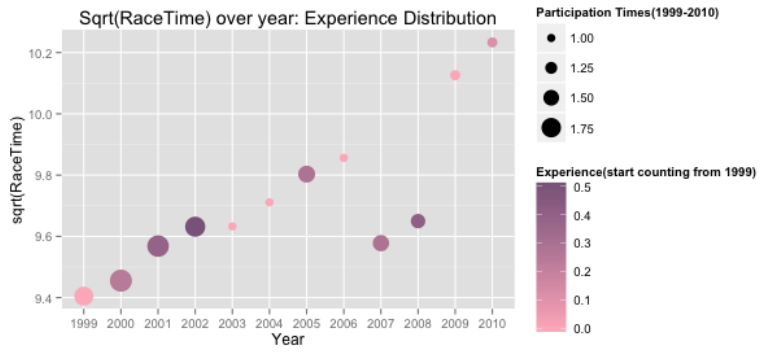
(2) There are more new race runners (with no race experience);

(3) There are more female involved in the race.

In order to test those three hypotheses, the plot of average race performance has been visualized by applying different colors and sizes.

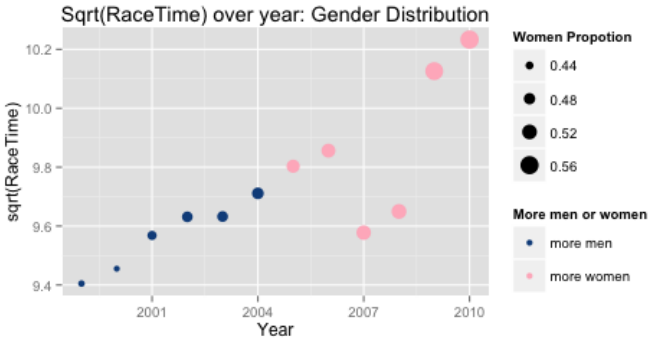
Hypothesis (1) is easily to be seen as rejected from the plot in the right. 43 is the upper quartile of the overall age distribution and there are actually less and less people over 43 participate the race along the years. Also the average mean of age for each year decreases along the years. Does this indicate age actually has a negative relationship with the Race Time? Further analysis will be provided later.





The left plot is a little difficult to explain for that there is no clear trend for either the colors or the sizes of the dots. However, some extreme dots may provide the insights. Since the participation times and experience started counting from 1999, it is nature to think that the dots should become smaller and darker over the time if lots of race runners participate the race several times. Based on this logic, 2006 and 2007 have more frequent runners than 2006, 2009 and 2010, which might be the reason why the overall race performance for those two years didn't follow the upward trend. However, just based on this plot, we don't have enough evidence for supporting hypothesis (2). For example, it's hard to explain why 2005 year still follow the upward trend since the darker color of the dot indicates more player with experience participated in 2005 race.

Hypothesis (3) is easily to be found being supported by the right plot. The proportion of women race runners increases along the years. Starting with 2005, there are even more female runners. The support for this hypothesis also suggests that the analysis of Race Performance should be separated among women and men. In this case, the above three plots has been plotted again by dividing with women and men.



However, even after divided by Gender, the upward trend of Race Time still exists in both plots for men and women. From Figure 3, we can see such trends are quite similar with the one in Figure 2. Since Age and Experience don't provide sufficient explanations for this upward trend, there might be some outside factors underlying. Those factors include climate, the difficulty of the race, as so on, which are outside the reach of our current database.

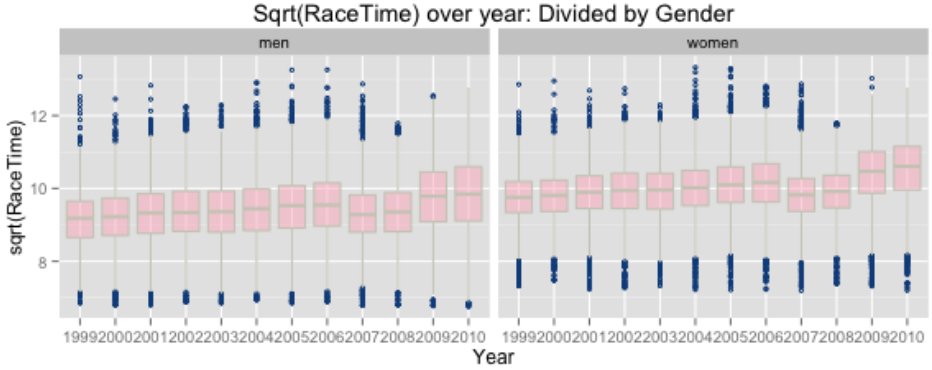
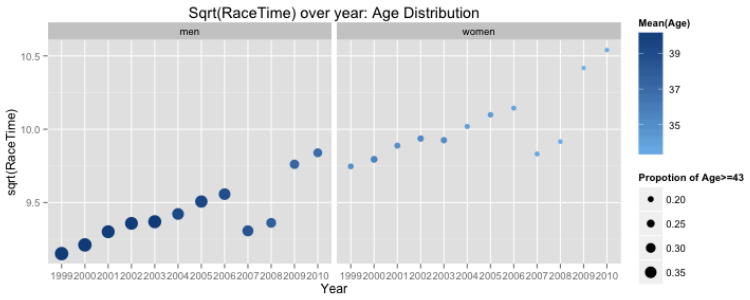
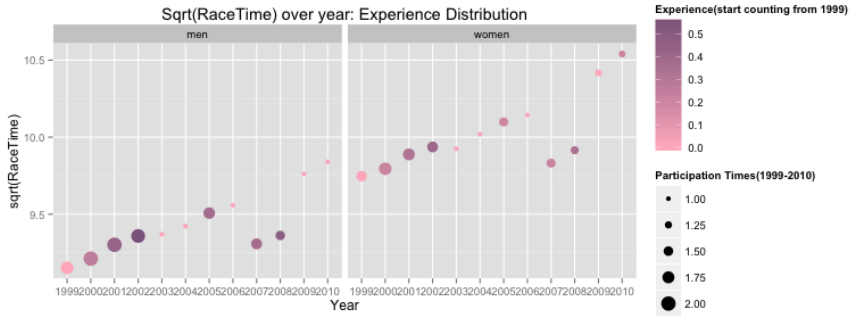


Figure 3: Race Performance (Sqrt(Race Time)) over Years---Divided by Gender



It should be emphasized though that Gender really matters by influencing the Age's performance on Race Performance over year. The right plot demonstrates that not only the female runners run slowly than the male runners overall, the female runners also tend to be younger. In other words the "negative" effect of Age on Race Time might be exaggerated by the Gender. In data analysis II, it will focus on the pure effect of Age on the Race Performance.



The left plot shows that the Participation Times and Experience for men and women share a very similar trend. It might not be very useful to combining the effects of gender for analyzing the Experience's effect on Race Performance. However, Experience and Participation Times are highly related with age, since people will become older after participating several times. This type of analysis will be shown in the third part of the data analysis.

Data Analysis II: Race Performance vs Age: A Linear Relationship?

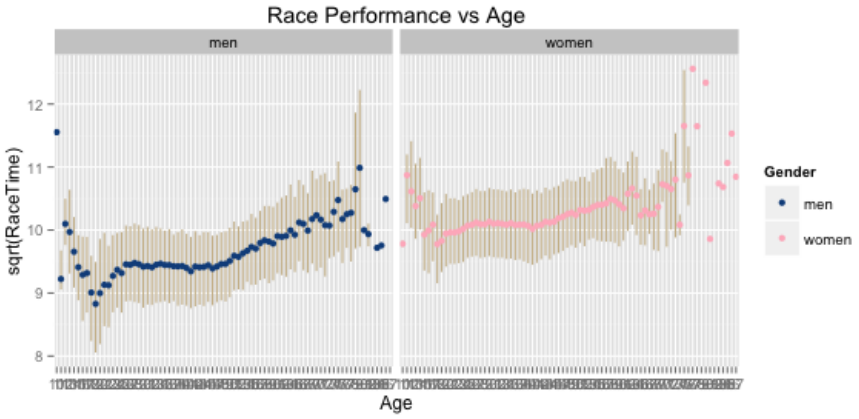


Figure 4: Race Performance vs Age --- Box-plot

To analyze the effect of Age on the Race Performance, a Box-plot is first given for the purpose to show the overall relationship. There still seems to be some positive linear trend between Age and Race Time, and such linear trend is more evident in the male runners' case. A closer investigation of the trend shows that

the best ages for running is around one's early 20th. While after one's 40th, the race speed slowed down even quicker.

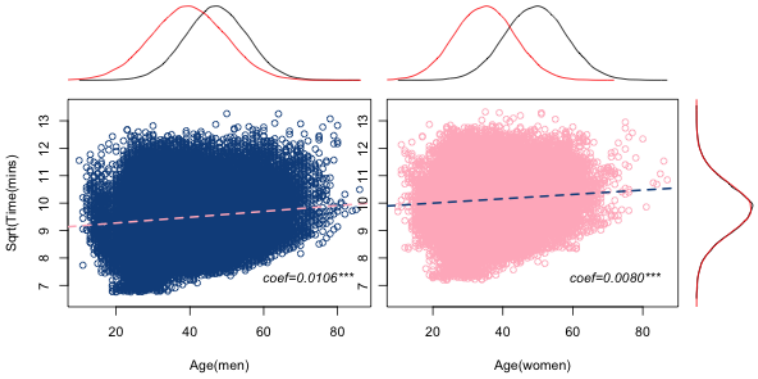
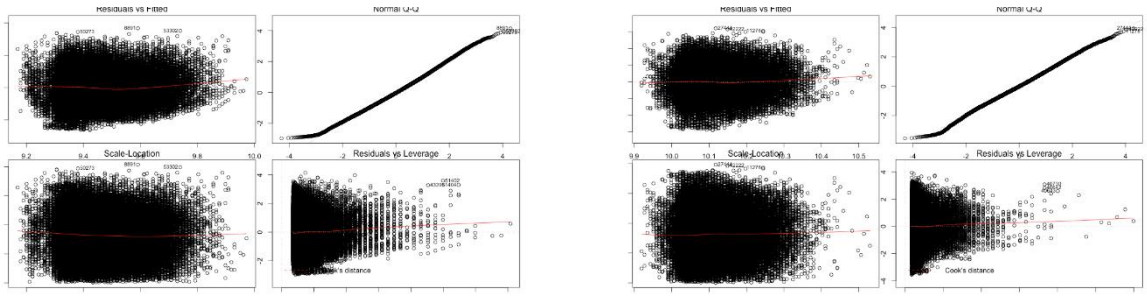


Figure 5: Race Performance vs Age --- Fit Linear Models

Linear Regression Models have also been tried in this case and Figure 5 is the summary of the findings. Both for women and men the Age distribution is negatively skewed. Against with the author's previous hypothesis, there are actually more very young runners and elder runners. Although the coefficients in both models are still significant and being positive, the absolute value (magnification) is quite small, which is around 1%. Also the R-squares for both models are no more than 2%. Figure 6 is the test for diagnosing whether the assumptions hold for the two linear models. It is shown except for some outsiders (see from Figure 4: some extreme young and old cases), the assumptions are satisfied. In sum, although there are some kind of linear trend between Age and Race Time, such effect is considerably small.



Men's Age Model **Women's Age Model**

Figure 6: Tests of Linear Models

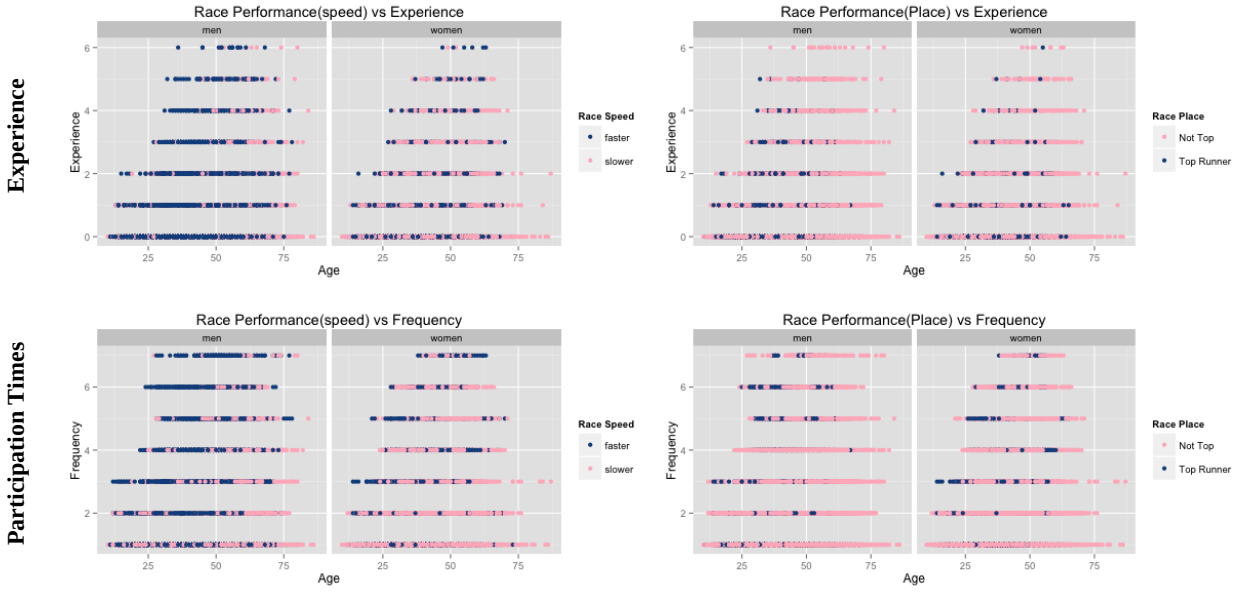
Data Analysis III: Race Performance vs Experience: after considering the effects of Age

The following four plots are intended to show the influence of experience and participation times (passion) not only differs between genders, but also mixed with the effect of age. There are typically two conclusions that can be drawn:

(1) Experience and passion have a stronger and longer influence the female runners.

(2) A typical quick or top runner is a frequent runner not older than 50 years old.

Race Performance measured by Speed | **Race Performance measured by Place**

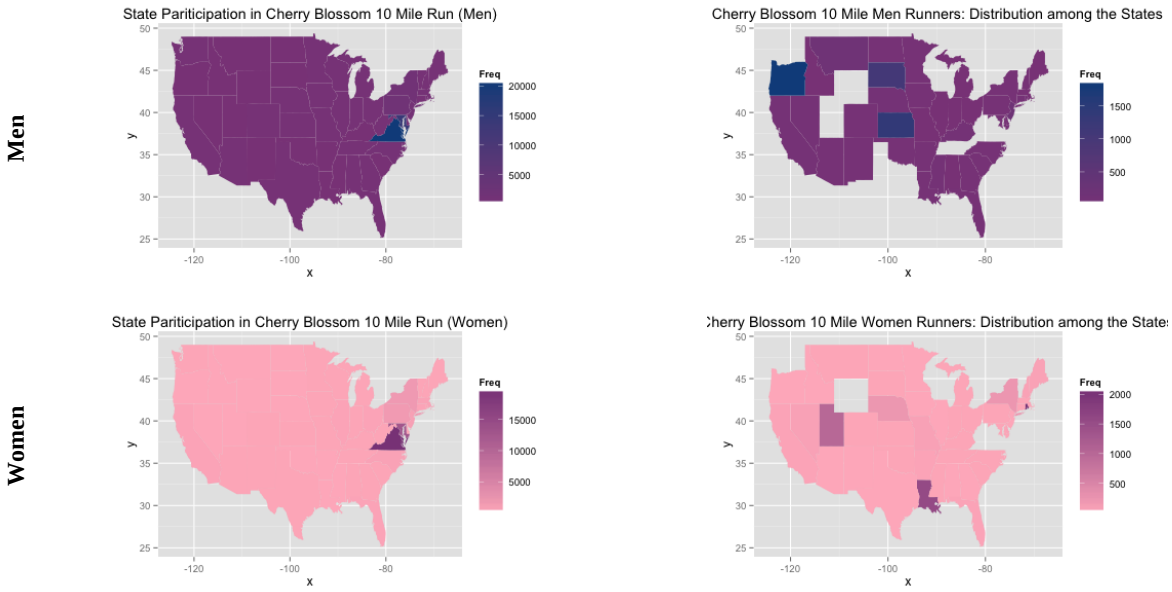


Data Analysis IV: Exploring State's Influence

Since the hometown data are inputted in different manners among the years, which makes it really hard for a sufficient analysis. The last data analysis is just to visualize the states performance. The Cherry Blossom 10 Mile Run is hold in Washington DC. So there is no wonder that more people living around the D.C area participate in the event (showing in the two left plots). Even more people in the east coast participate in the competition, people from other states are actually performs better.

Top 500 Men Runners mostly come from Oregon, Kansas and South Dakota. While Top 500 Women Runners mostly come from Rhode Island, Louisiana and Utah. Those findings are interesting, which worth further analysis of the states.

Overall Participation | **Top 500 Runners Distribution**



Resources

[1] Max Woolf. An Introduction on How to Make Beautiful Charts With R and ggplot2
http://minimaxir.com/2015/02/ggplot-tutorial

[2] Jie Zhou. How to draw good looking maps in R

https://uchicagoconsulting.wordpress.com

[3] Jiahui (Kavi) Tan. Use read.fwf() Piazza @39

https://piazza.com/class/i73w08p749y6xt?cid=39

[4] Duncan Temple Lang. xyDensityPlot.R

http://eeyore.ucdavis.edu/stat242/xyDensityPlot.R

[5] SAPE. ggplot2 Quick Reference

http://sape.inf.usi.ch/quick-reference/ggplot2/linetype