Analysis about Cherry Blossom Race Results

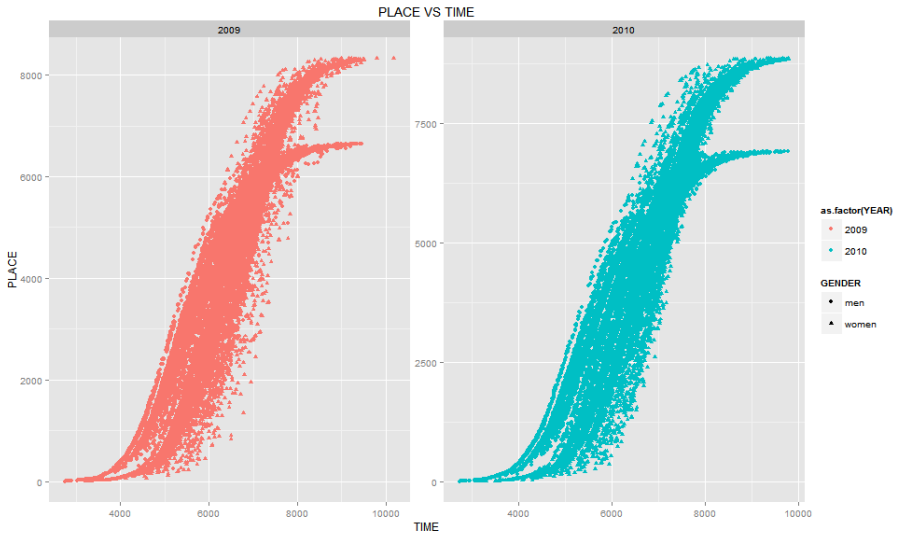
Yibing Luo

912491862

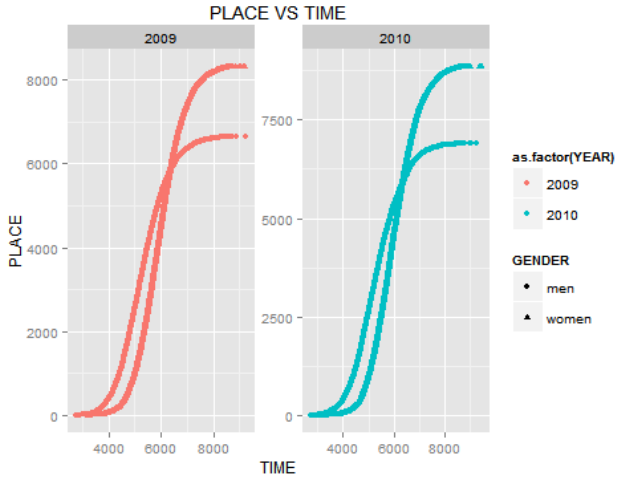
2015/4/13

Overview Of The Whole Data

In analysis part, a dataset called ‘wholeData’ is created. There are originally 15 variables. 4 more variables are created there:
‘age’: This is a category variable indicts the groups of the age.
‘GENDER’: This is a category variable indicts the gender.
‘YEAR’: This is a numeric variable shows the year.
‘FOREIGN’: This is a category variable indicts whether runner is foreign.
‘INDICT’: This is the numeric variable aims as finding people who took part in the running several years (Also, a new dataset called **‘forName’** is created for this purpose)
‘TIME’: For those years which don’t have ‘TIME’, gun time is used to replace it if year is between 1999~2008. Otherwise, net time is used to do so. Because PLACE is decided based on net time since 2009(It can be find through the dataset).

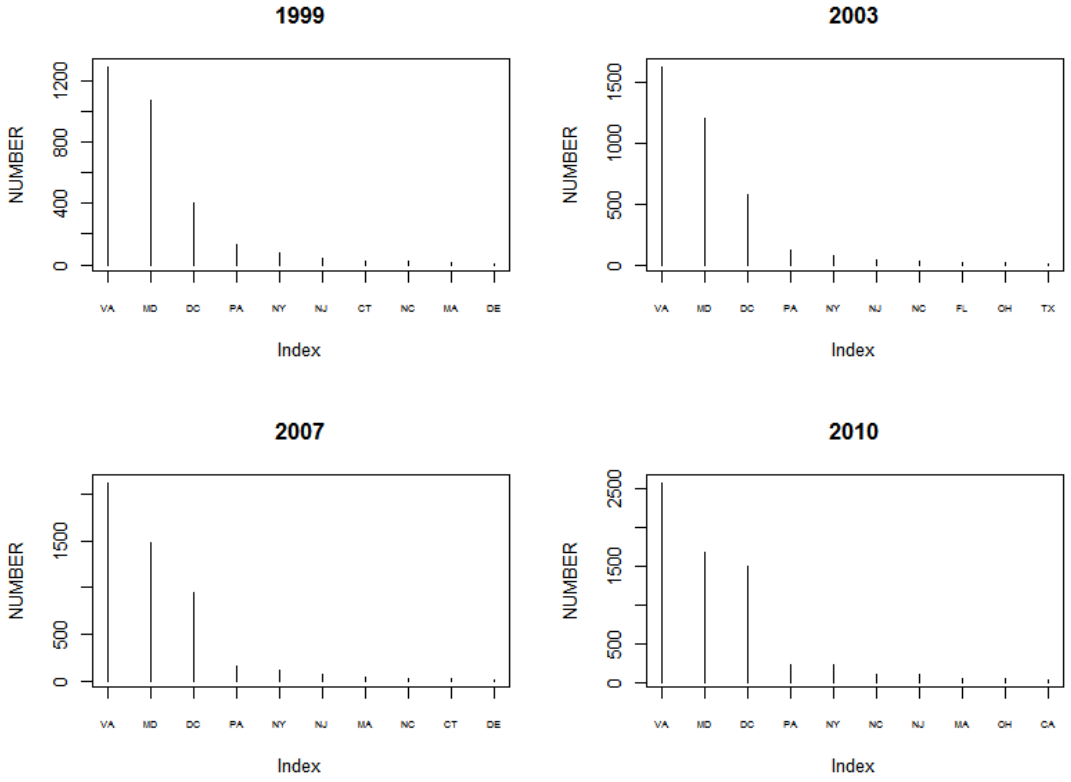


The plot above is PLACE vs GUNTIME in 2009, 2010. It give the evidence that this race no longer uses gun time since 2009. The fix plot is below as we use NETTIME instead of GUNTIME.



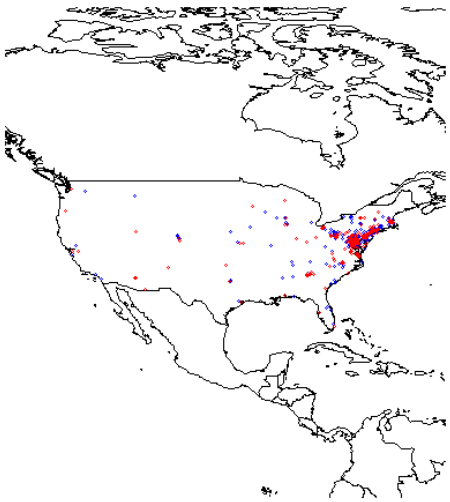
State And City

A variable in the data named ‘HOMETOWN’ normally can provide information about the states and cities. I want to use it to see the change about the areas runners came from across the year. It may show some patterns and tendency which interested me a lot.
Because of the limited space, only 1999, 2003, 2007, and 2010(2006 is not supposed to be analyzed because ‘HOMEWTOWN’ in 2006 is not easy to process) are chosen to analyze.

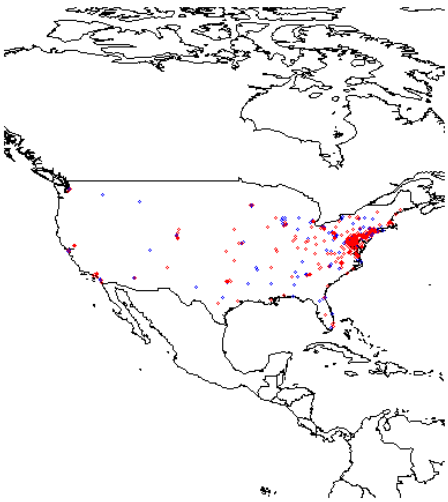


As shown above, states ‘VA’,’MD’ and ‘DC’ have the largest number of participators these four year. It makes sense for the reason the Cherry Blossom 10-mile running race is held in Washington DC. People in closed states are convenient to take part in this competition.

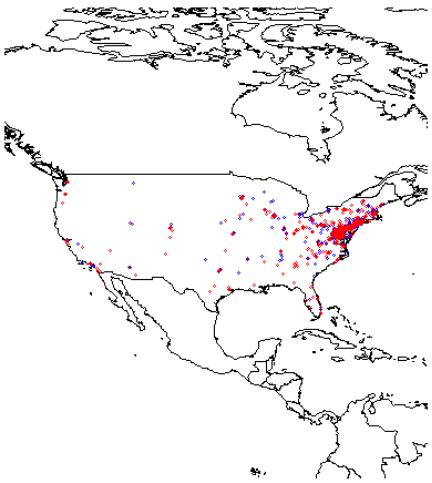
YEAR=1999



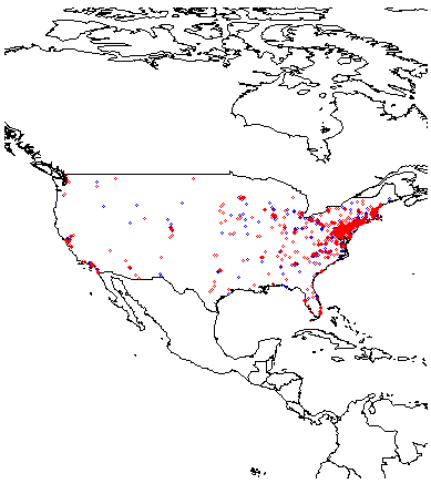
YEAR=2003



YEAR=2007



YEAR=2010



(Blue points represent male and red ones for female)

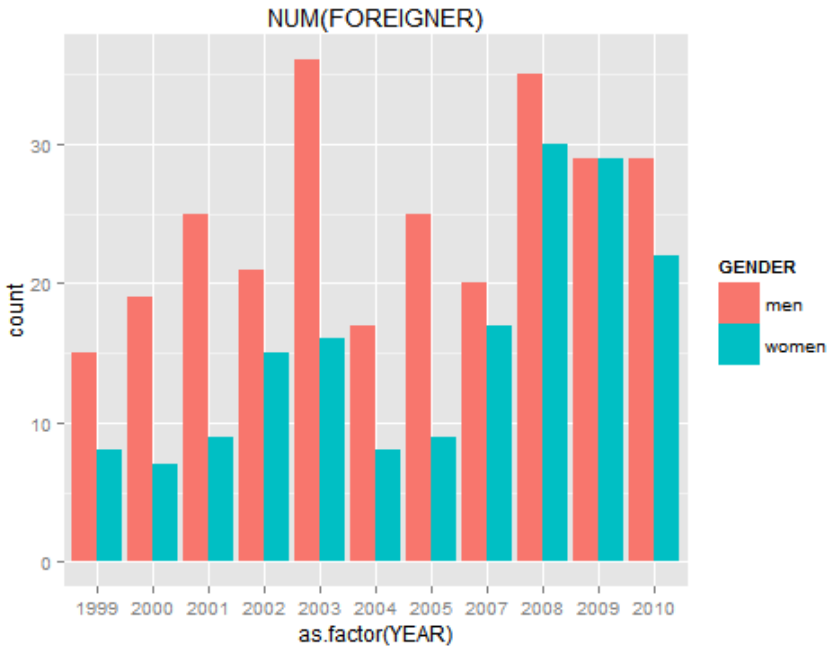
The plots above display the change of cities where athletes came. There are some findings:

1. More cities are in northern America, and most of which concentrate around DC area.
2. More and more cities involved in this game, especially in east of America, but cities in west and middle of America change slowly. It can conclude that the closer the area to the DC, the faster change in it. Such result may reflect the development of ways to transport as well as the popularity of this game.
3. Even in 2010, there are quite few cities in western American involved in this game. Quite a few cities concentrate around the sea coast. (Don't know why.

Maybe the organization of this game pays more attention there.)

Athletes From Foreign Countries

There are some foreign runners there. Observing there grades and change can be interesting. We can know people from which countries are better in long-distance running or like this race best. Did these foreign athletes' performances correspond to the performances in other games, such as Olympic?

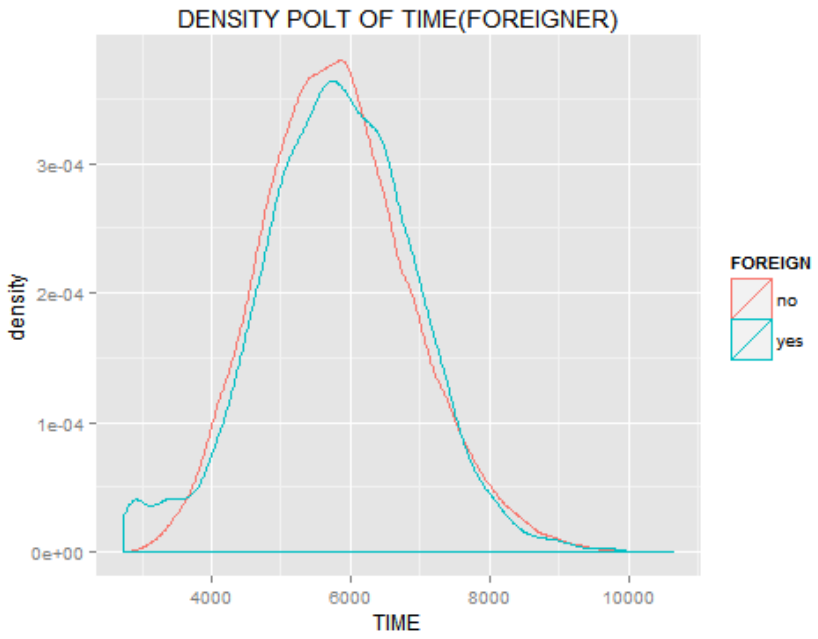


The above shows the number of foreign runners remains around 20. There is no obvious pattern for the change.

First take the number of athletes in different countries into consideration. The table gives top1~6 countries:

	Kenya	Canada	Ethiopia	Russian	Japan	Romania
Number	155	45	36	14	8	7

Then take a look at the density plot about variable TIME, which can measure performance very well.



From this plot, we may conclude that foreigners have high probability to get a high rank.

Then the PLACE1~4 runners are chosen from each year. We summary their hometowns and select the five largest number ones. Below is the table:

	Kenya	Ethiopia	Romania	Russian	Morocco
number	52	12	6	4	3

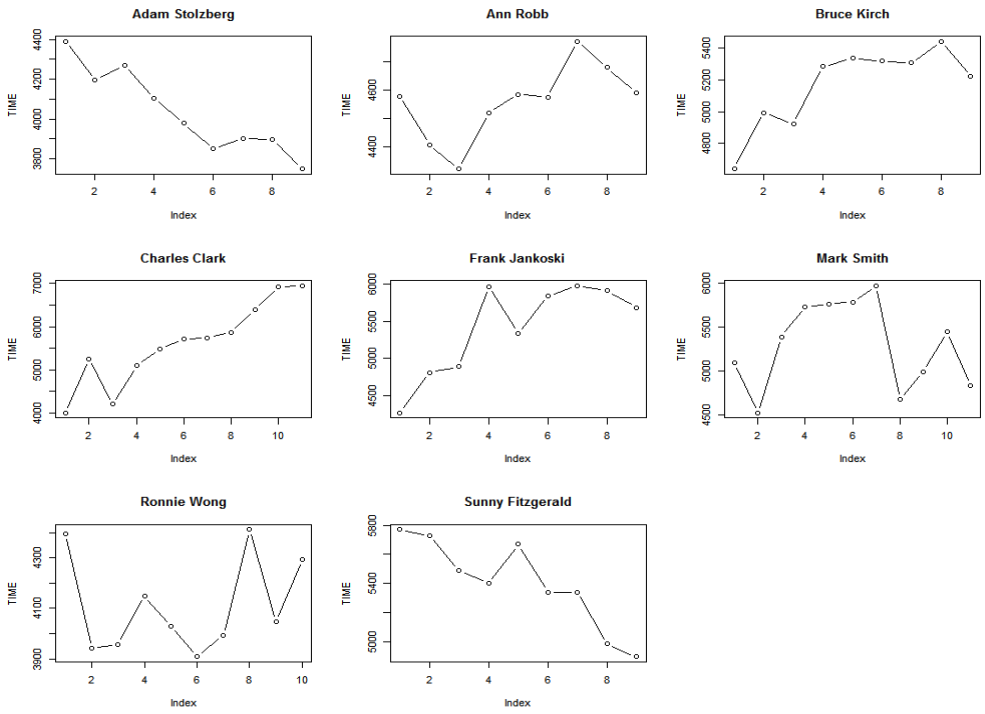
Kenyan athletes are absolutely the kings of long-distance race! And it's corresponding to their performances in Olympic. Also, Romanian athletes are excellent on the point of their small number of participators. Ethiopia is also great.

Canada has a large number of athletes, however the grade is not that good. In my opinion, I think the reason is it would be convenient to go to the states in Canada. So people would like to go their only for their interest.

However, it shows a high relationship between the number of participators and their performance among foreign athlete.

Athlete Who Took The Games Several Times

In this part, we want to track those athletes who join this game for several times. I wrote a function to do so. However, there is an error '`protect(): protection stack overflow`'. I cannot solve it. So I can just deal with a small part of the dataset and take a look on it.

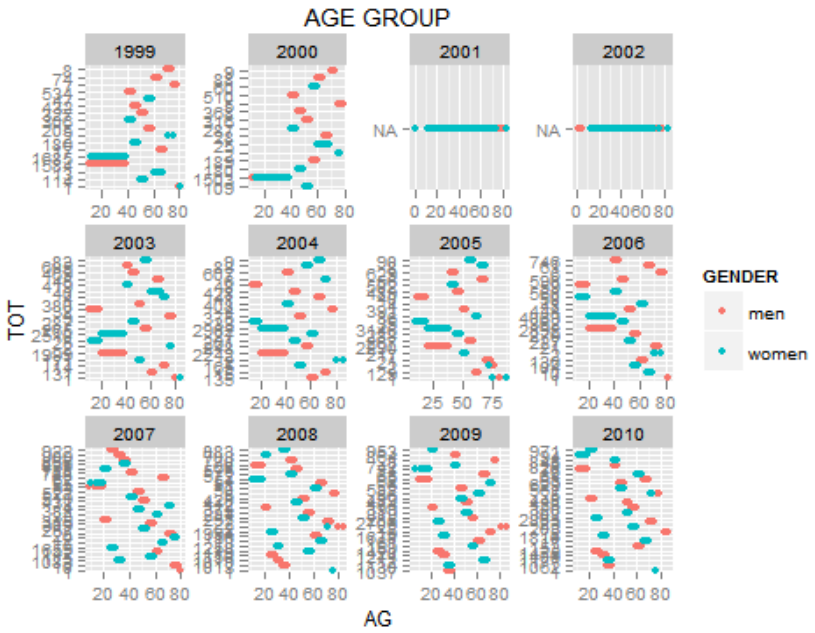


The plots below are random selected. These people are all ‘old men’ who took the game more than 8 times. However, the plots show no obvious pattern
As far as I am concerned, I think there is no obvious pattern for the ‘old men’. The people who just want to play or exercise in the game, their grades may be stable or decrease because of the age. Those who aim at getting good grades, can gain better results after practiced through years.

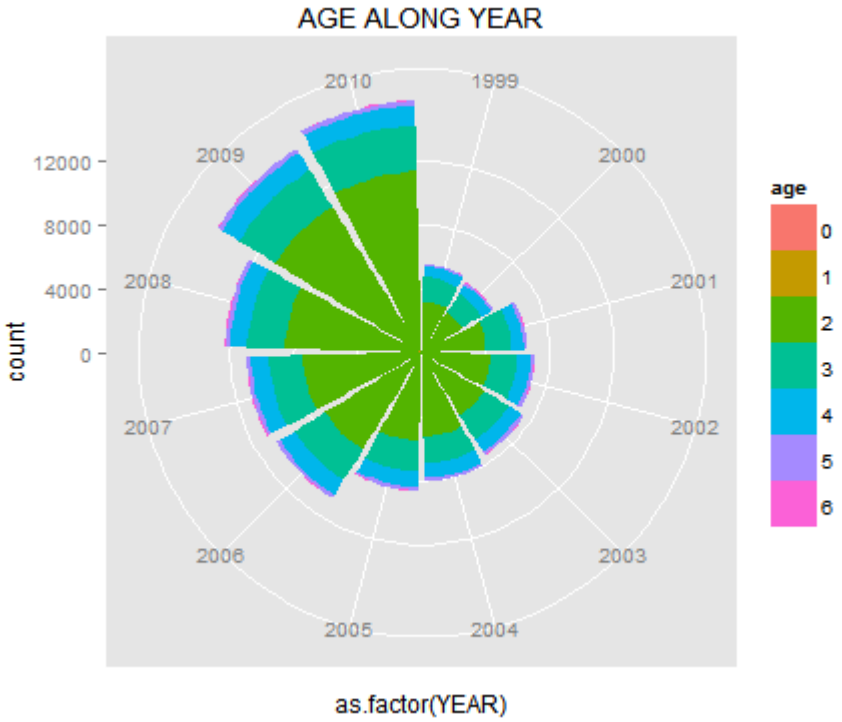
Age, Year, Gender and Time

In this part, some basic but important variables are taken into consideration. I am interested in them, because they usually give us a big picture of the dataset.

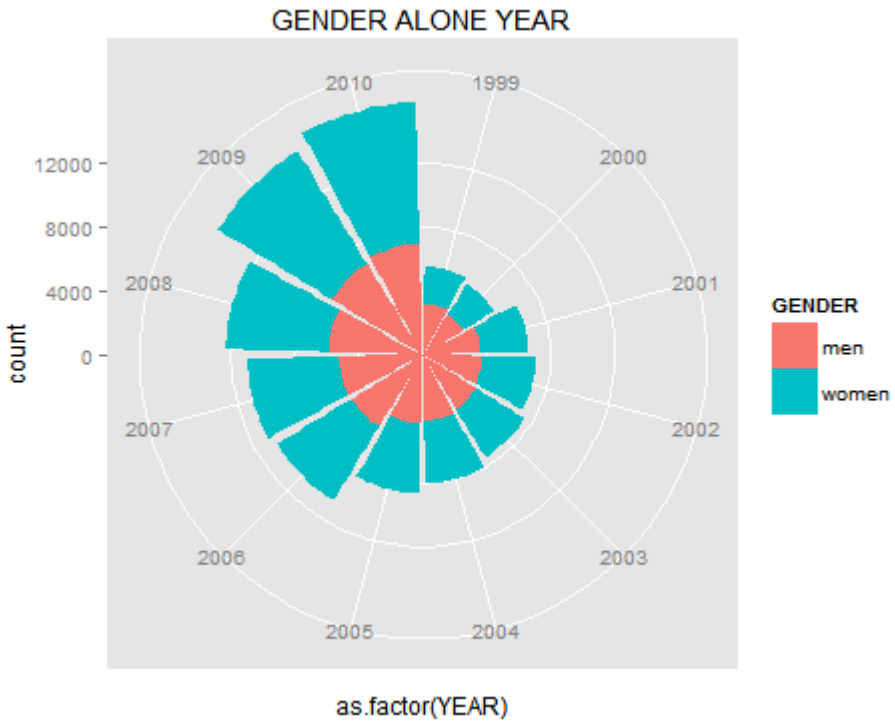
Variable ‘TOT’ gives us concept about grouping. The plot below shows that the interval for each age group is narrower. I think the increasing number of runners leads to this.



The plot above tells us the change about proportion of age group. It obvious that the number of participators is increasing. All age groups' proportions remain almost same except group2 (20<age<40). Group2 increase more fast.



The plot below is about variable 'GENDER'. We can obtain that the number of female athletes changes way fast than male athletes alone the year. This may reflect the change of attitude about exercise for women.



The plot below support that the younger people are more likely to get better grades. However, older athletes, e.g. group 6, improved in speed through all these years. The well-developed live and medical environment may interpret it. People today have a better physical condition.