MSDS 7333 Fall 2020: Case Study 2

Jayson Barker, Brandon Croom, Shane Winestock

Business Understanding

The Cherry Blossom Ten Mile Run is an annual 10-mile race in Washington, D.C. The race was originally founded in 1973 and has been run continuously ever since. The race is meant to coincide with the bloom of the cherry blossoms in the DC area and is part of the National Cherry Blossom Festival. The race started as a training race for professional runners planning to compete in the Boston Marathon. Over the years the race has evolved into a race that attracts both professional and local runners of all levels. Evaluating this change in runners is a critical component for the race planners to understand how to position and market the race to attract participants.

This report focuses on the 1999 through 2012 years of the race. The goal is to specifically analyze female runners over the years to determine the answers to a few key questions that the race planners need for marketing purposes. Specifically the following questions will be addressed:

- Have age distributions changed over time?
- Have race times increased or decreased over time?
- What trends should the planners be aware of that may occur in future years?

Data Acquisition

In order to perform the analysis the data for races from 1999 through 2012 must be acquired. The easiest way to gather this data is through web scraping of the Cherry Blossom race website. The website contains web pages listing all of the race results through the specified time period.

Web scraping has many aspects that must be taken into account when performed. First and foremost, the data must be allowed to be scraped from websites. The content of a website may be a revenue generation opportunity or built out content that the website owner has spent considerable time curating. When acquiring data from websites it must first be verified that the website allows scraping of the site's data. Scraping websites without site owner's consent may result in litigation if the data is presented without attribution and may be in violation of copyright laws. In this particular case, the data from the Cherry Blossom site is being presented to the race planners Since this data is for the race planners and is from the race website there are no legal implications present in scraping the website data for this analysis.

A second aspect that must be taken into account when scraping websites is understanding the data availability and working through the differences that may occur overtime with individual site webpages. In the Cherry Blossom case, the race results were housed on individual webpages for each race year. However, each year's webpages has subtle differences that required changes in the acquisition method to ensure the data was captured correctly. For example, some years of the Cherry Blossom race captured only race completion time, while others included completion time and gun time. The web scraping must be adapted to take all of these possible changes into account.

In building out the specific web scraping capability leveraged for this analysis a few key items in the individual webpages needed to be addressed. Each webpage for each year is analyzed for key items that must be taken into account to ensure the data can be scraped in a consistent fashion. Each webpage uses equal marks (‘=’) as column header delimiters. The specific spacing for each column for each year must be accounted for to ensure the correct fields were consistently mapped. In addition the necessary webpage tags (the layout information for each individual webpage) much be searched to find the correct tag that contains the race results. The years 1999, 2000 and 2009 had different formatting for these tags than other years, thus special analysis was required to scrape these years to ensure consistent results.

The results of scraping the Cherry Blossom site for the women’s results provides us the following record counts for the years in question. This information provides an initial check that the web scraping is working as expected and provides an individual record set by year for initial data cleansing.

```
##     Year Count
##  1: 1999  2356
##  2: 2000  2167
##  3: 2001  2973
##  4: 2002  3335
##  5: 2003  3544
##  6: 2004  3903
##  7: 2005  4338
##  8: 2006  5441
##  9: 2007  5696
## 10: 2008  6401
## 11: 2009  8329
## 12: 2010  8859
## 13: 2011  9034
## 14: 2012  9733
```

Data Cleaning

Now that the data has been broken down into individual record sets for each year, an initial look at the data structure for each individual record set is warranted. This will ensure that data mapped into the data sets represents the correct data types for the values that are expected to be pulled from the data. Since all individual record sets are pulled in the same format evaluating the 2000 record set will provide a quick indication that data was mapped correctly.

```
sapply(women_2000_df,class)
```

```
##       Place      DivTot         Num        Name          Ag    Hometown 
## "character" "character" "character" "character" "character" "character" 
##         Gun         Net 
## "character" "character"
```

From the above information it can be seen that every field is being read as a character field. Given that the data is being scraped from the web, this makes sense. Data conversion will be necessary to the respective data types as individual record sets are combined into a full record set.

The evaluation of missing data is another items that is best addressed at the individual record set level. Performing missing data checks at this level makes addressing them much easier prior to combining the data into a full record data set. The table below shows the results of missing data analysis for each of the individual record sets.

```
##     Year Count
##  1: 1999     0
##  2: 2000     0
##  3: 2001     0
##  4: 2002     0
##  5: 2003     0
##  6: 2004     0
##  7: 2005     0
##  8: 2006     0
##  9: 2007     0
## 10: 2008     0
## 11: 2009     0
## 12: 2010     0
## 13: 2011     0
## 14: 2012     0
```

From the table above, there is no data missing in the records. This shows that the data entered in to the web site is complete and that the data scraping methodology worked as expected.

As previously noted, the individual record sets have all data defined as characters. The time values (Gun, Net) and the age values need to be converted to numeric data types to ensure further analysis is correct. At the same time all of the data needs to be merged into a single record set containing all years. As part of this merging, a new data field for race time (Net_Conv) will be created. This data field will contain the race time converted to minutes. Just as with the individual record sets the evaluation of data types and missing values for the combined record set will occur to ensure data is ready for analysis.

Checking the combined record set data types, as noted below, all values are now more indicative of the respective data column. Age (Ag) is now numeric and the race time (Net_Conv) is also numeric.

```
##        Year       Place      DivTot        Name          Ag    Hometown
## "character" "character" "character" "character"   "numeric" "character"
##         Net        Pace    Net_Conv
## "character" "character"   "numeric"
```

Evaluating the data for missing values produces the following results:

```
##    Count Missing
## 1:             0
```

Based on the missing value check, it can be seen that the data has remained intact throughout the transformations. One additional check to provide summary statistics of the data will be performed prior to beginning analysis. This summary statistics check provides another sanity check for data evaluation.

```
##      Year              Place              DivTot              Name          
##  Length:75839       Length:75839       Length:75839       Length:75839      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##        Ag          Hometown             Net                Pace          
##  Min.   : 0.00   Length:75839       Length:75839       Length:75839      
##  1st Qu.:26.00   Class :character   Class :character   Class :character  
##  Median :31.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :32.73
##  3rd Qu.:39.00
##  Max.   :87.00
##     Net_Conv
##  Min.   : 51.73
##  1st Qu.: 89.00
##  Median : 98.18
##  Mean   : 99.31
##  3rd Qu.:108.40
##  Max.   :170.97
```

The summary statistics above reconfirm the information that was expected. The character fields should not have an statistical values. Additionally, the Age (Ag) and race time (Net_Conv) fields do show reasonable statistical values. It should be noted that the Age (Ag) field does show a minimum value of 0.00. This indicates that there may be incorrect data input into the age field for some runners or that the data was omitted. Evaluating the combined record set to pick out zero age record results in the data shown below.

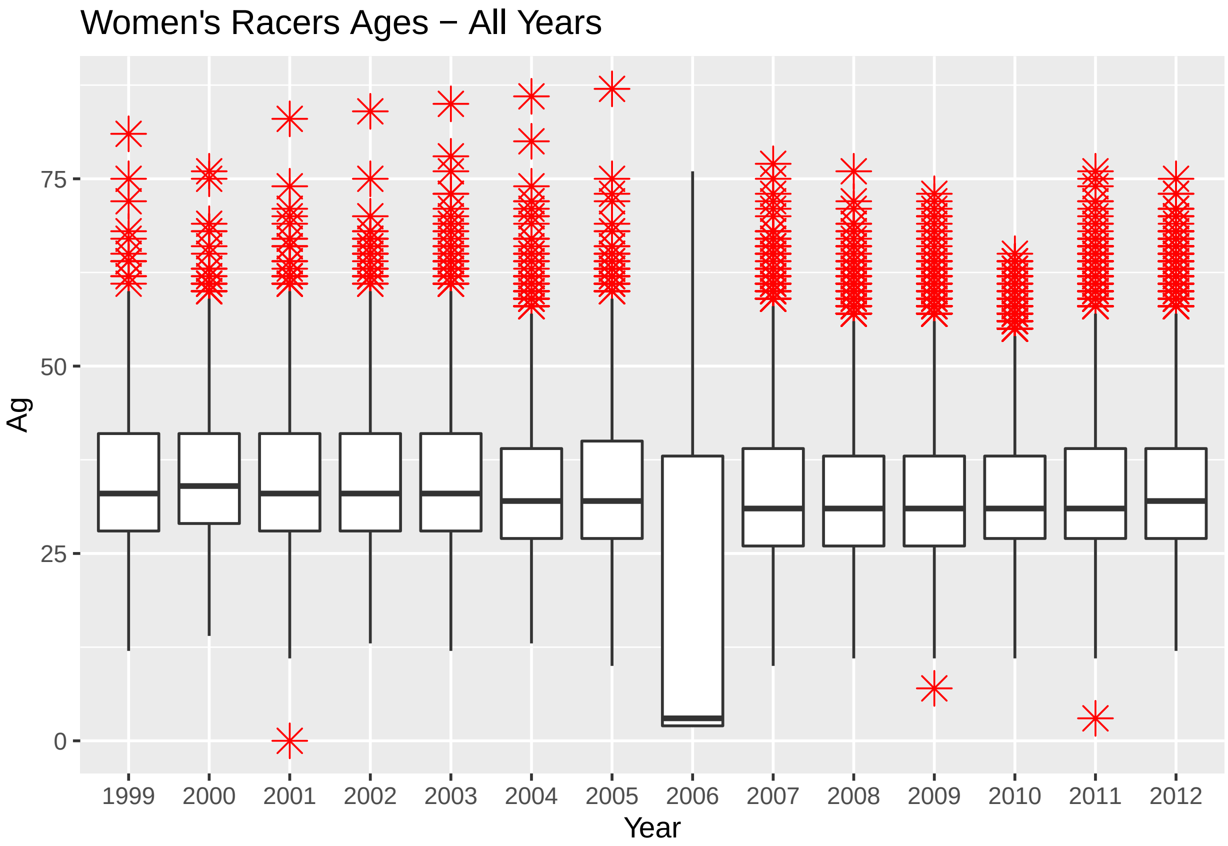
```
##      Year Place DivTot                  Name Ag           Hometown     Net Pace
## 7129 2001  2611      0 Loretta CUCE          0 Alexandria VA      1:53:38    0
##      Net_Conv
## 7129 113.6333
```

Evaluation of the race results web page for the 2001 race does confirm that the age entered for this individual was 0. As such this should be treated as an invalid data point, with respect to age.

Data Analysis

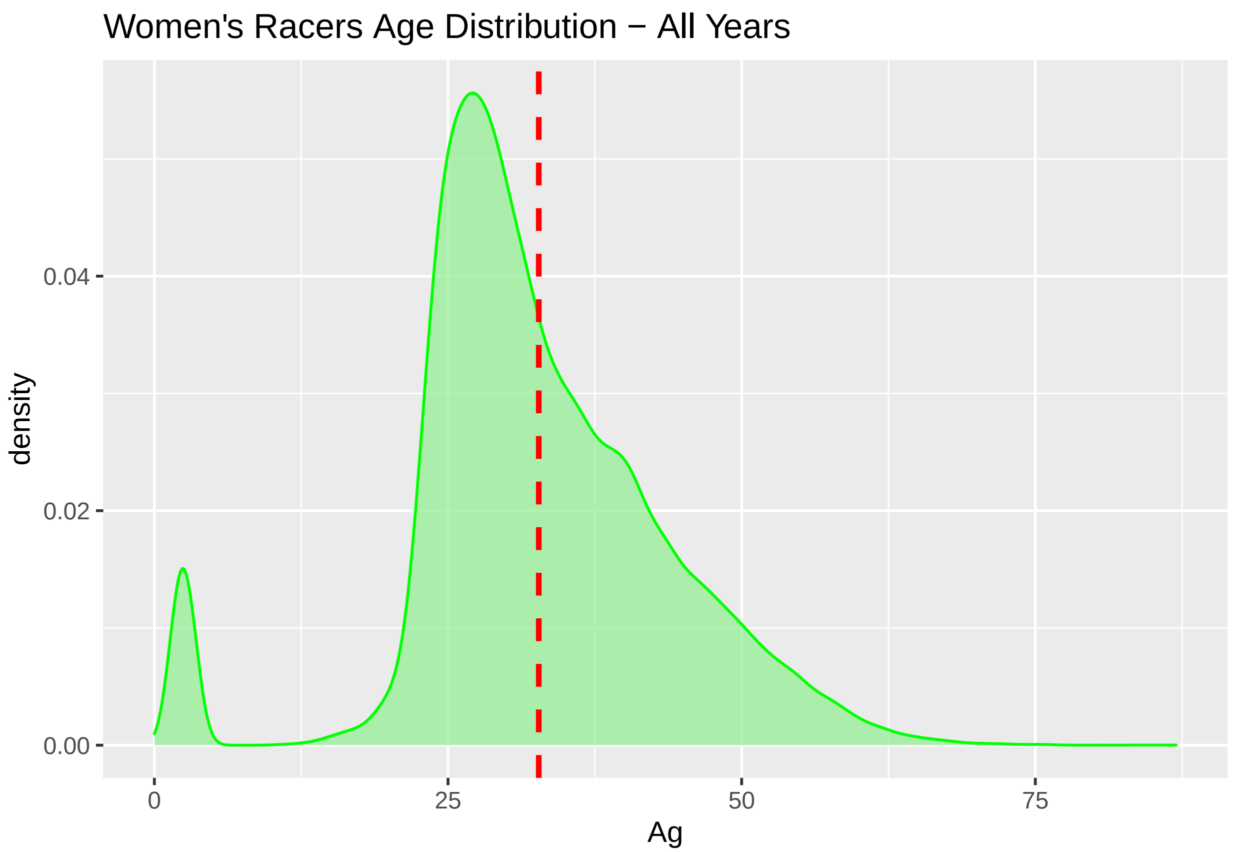
Now that the data has been scraped and cleaned for analysis, the process of answering the requested business questions can begin. The initial analysis will start with an analysis of racers ages over time. It will then move into the evaluation of race completion times over time.

Age Analysis



From the box plot it can be seen that 2006 had a lot of younger racers compared to the other years. Additionally, in 2006 many of these younger racers were at the tail end of the race times. This could indicate an influx of non-professional runners during the race for that year.

Diving into the distribution of ages across the race years the plot below shows a bimodal distribution of ages. This distribution reaffirms the large number of younger racers seen in 2006.



Continuing to evaluae runners ages the data provides an opportunity to evaluate mean age by race year as shown in the table below.

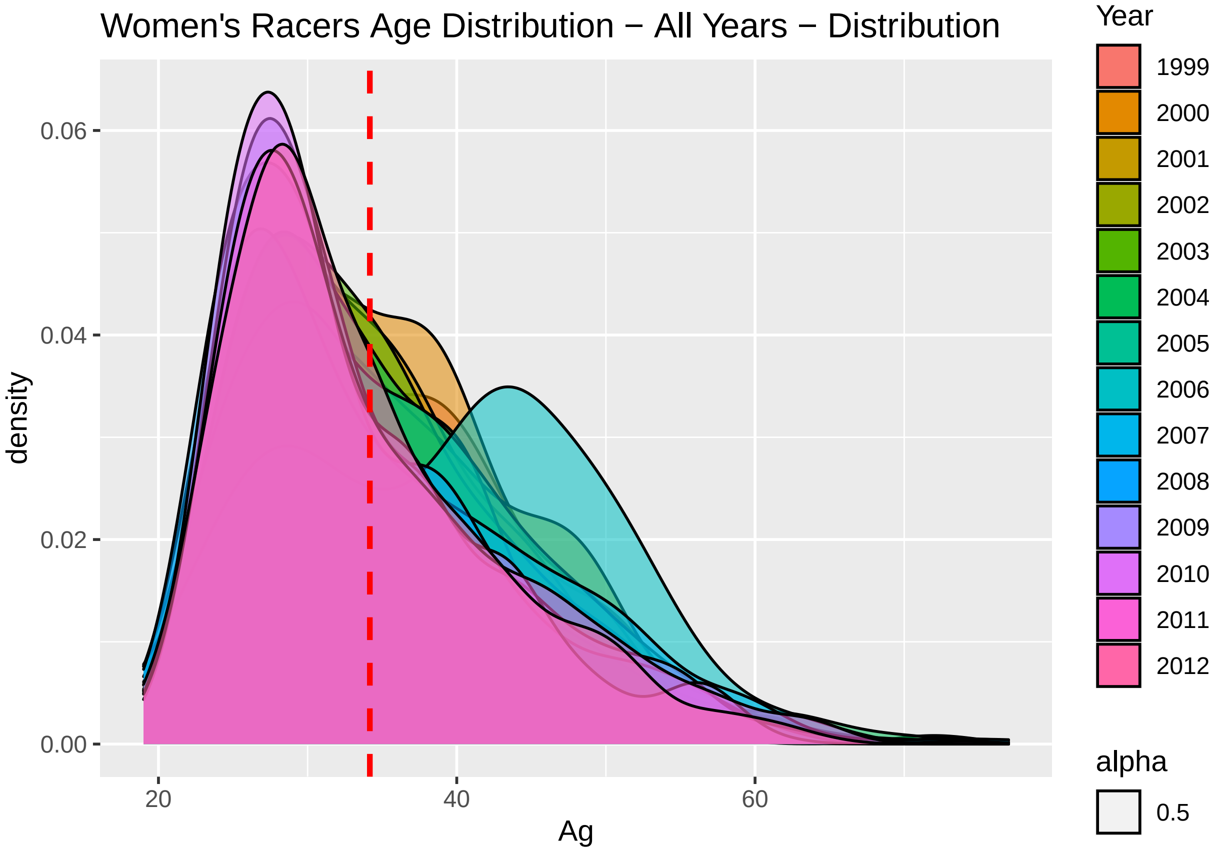
```
##       Year         Mean Age
##  1:   1999 34.9009353741497
##  2:   2000 35.5535549399815
##  3:   2001 34.8061911170929
##  4:   2002 35.1378378378378
##  5:   2003 35.0513833992095
##  6:   2004 33.9296534017972
##  7:   2005 34.1692485549133
##  8:   2006 18.3847994111152
##  9:   2007 33.4678571428571
## 10:   2008 33.2102548069408
## 11:   2009 33.0775147217882
## 12:   2010 33.0376171964327
## 13:   2011 33.7369878183832
## 14:   2012 33.8779936272998
## 15: Total 32.7250754888646
```

This table confirms the distribution plot and box plots above with respect to racers in 2006 skewing much younger for that year. Using the mean age of all races as 32, looking specifically at 2006 it can be seen that of the 5434 racers in 2006, 3784 of them were below the mean race age.

```
## [1] 5434
## [1] 3784
```

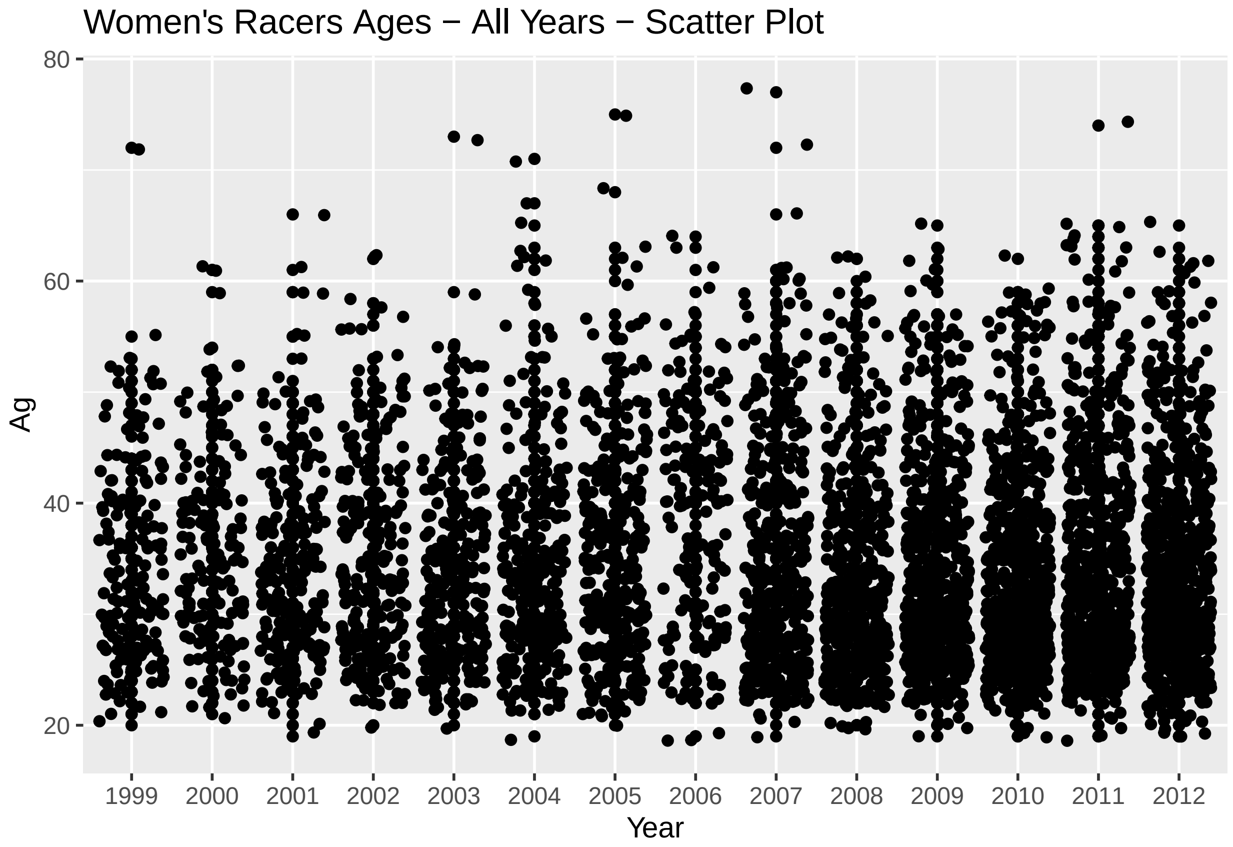
Reevaluating the age data to account for the 2006 outlier the following distribution can be seen:

```
## Warning in Year == Year_Sel: longer object length is not a multiple of shorter
## object length
```



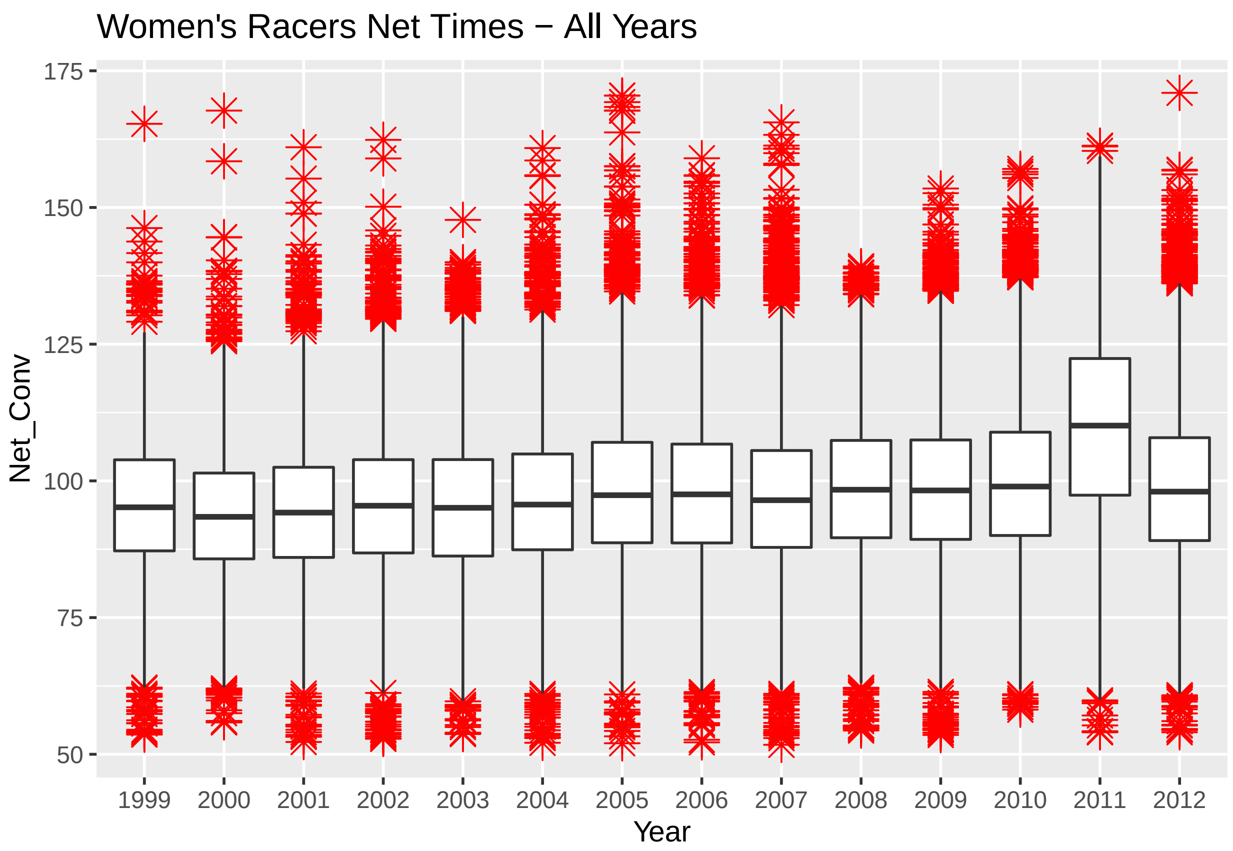
Accounting for the extremely young outliers in the distribution above it can be seen that there is still some skew towards younger ages. The outliers in this case are 2002 and 2007 where ages were actually a bit older. Leveraging the same data in the scatter plot below it can be seen that there is a stronger concentration of ages in the 2009-2012 year range compared to previous years. This could indicate that more racers are entering the race.

```
## Warning in Year == Year_Sel: longer object length is not a multiple of shorter
## object length
```



Time Analysis

Just as with the age data there are interesting trends the emerge when evaluating race time data. A simple box plot of race times begins to show some of these points of interest. In the box plot below of Net Race times over all years it can be seen that on average the race times were consistent in all years except for 2011.



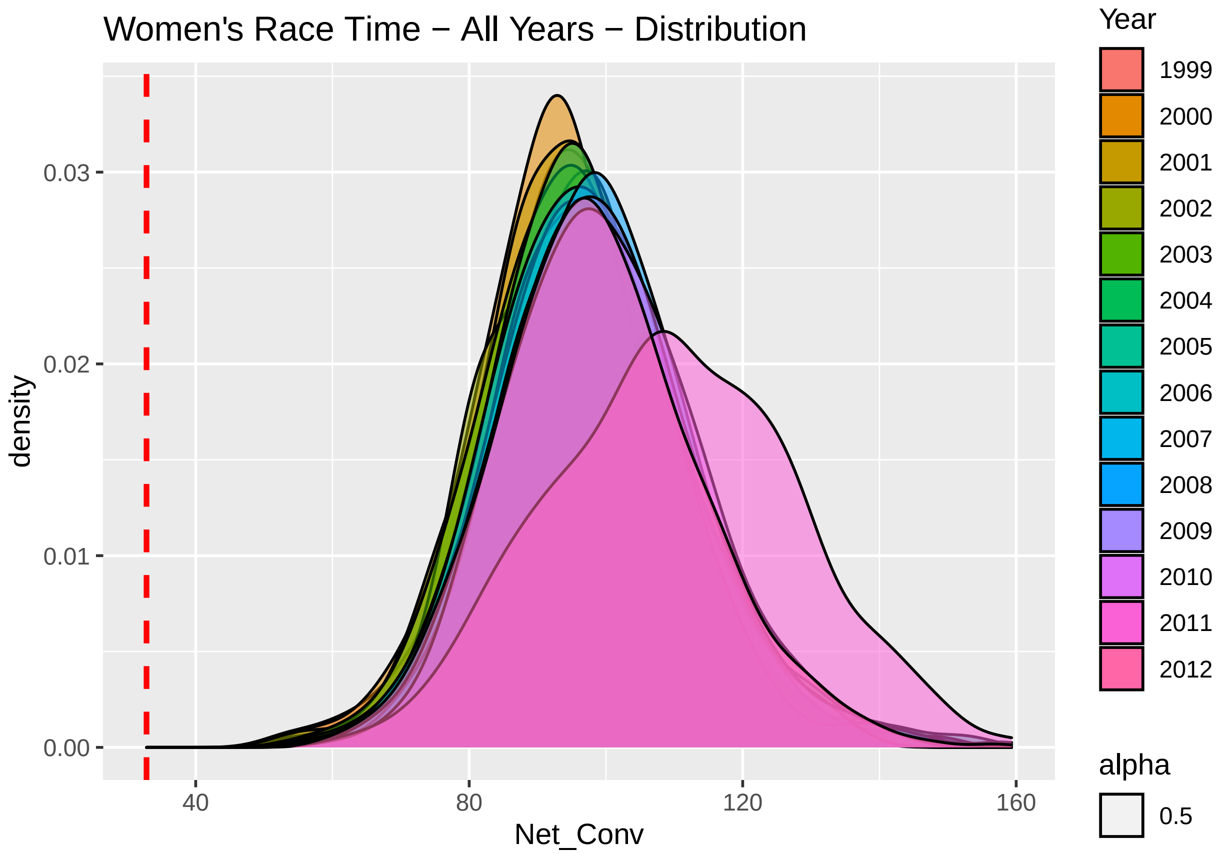
Digging deeper into mean runner time provides the following table:

```
##       Year         Mean Time
##  1:   1999  95.538102324263
##  2:   2000 93.9138504155125
##  3:   2001 94.6531123822342
##  4:   2002 95.9373323323323
##  5:   2003 95.5493459439112
##  6:   2004 96.7006718014549
##  7:   2005 98.6066319845857
##  8:   2006 98.1578241933505
##  9:   2007 97.1735476190476
## 10:   2008 98.7727710906154
## 11:   2009 98.8117333653808
## 12:   2010   99.92128401555
## 13:   2011 109.904671465485
## 14:   2012 99.0223027375201
## 15: Total  99.313335047491
```

From the mean time table above the 2011 outlier is reaffirmed. This outlier is almost 10 minutes higher than the mean of all races. Additional data would be needed to evaluate possible causes of this outlier discrepancy. Multiple factors could be at play to cause this outlier. Some examples of possible impacts could be: racers could have truly run the race at a slower pace, weather conditions could have impacted racers, or even the total number of racers could have impacted these times. Based on the data available, 2011 had the second highest total number of racers, with 9034, however this was only an increase of approximately 200 racers from 2010. In addition, the number of racers in 2012 went up to 9733 which was a much larger increase and race times for 2012 were back around the mean. Additional analysis would need to be performed with additional data sets to truly understand what caused 2011 to be such an outlier in race time.

The distrubution chart below also reaffirms that 2011 was an outlier year.

```
## Warning in Year == Year_Sel: longer object length is not a multiple of shorter
## object length
```



Conclusion

In conclusion the data on the Cherry Blossom race from 1999 through 2012 has provided some interesting insights. These insights should help race planners in evaluating any changes that may need to occur to maintain excitement about the race.

With respect to racers age, it was shown that the mean racers age of approximately 32 years old has remained consistent over time. There was a data anomaly in 2006 where many young racers performed in the race, however that anomaly was not seen in other years.

With respect to race completion times, like age, these too have remained fairly consistent over the 1999 through 2012 time period. The year 2011 was an outlier in this data and those results were not seen again in the data set. On average racers were finishing the the race in about 99 minutes.

The data also confirm that the number of racers entering the race continues to increase year over year. 1999 had 2365 runners while 2012 had 9733 runners. This is an extremely large increase. From a business perspective it could indicate that interest in the race still exists.

All in all, both the age and time data seem pretty consistent over the 13 year span of data that was evaluated. As the number of racers continue to increase over the years these values may shift as more racers begin to enter the race.

From a race planner perspective this data can assist in planning for future races on many levels. It can assist with planning out the race course. On average the racers are completing in about 99 minutes, in a congested city such as Washington, D.C., understanding how long streets may be blocked off and impact traffic around the city. In addition, this information may go into overall route planning by providing alternative race routes that fall into the length of the race but also can be completed in the average time. Understanding the average ages of the runners could largely help the planners in a marketing sense for the race. Knowing that the racer ages skew toward the low 30s provides the race planners with information about how best to reach those racers. It may be more beneficial for the race planners to market on websites and social media sites given the average racer age than to advertise in running magazines or newspapers.

Overall, the trends look to be relatively steady with respect to runners ages and completion times. There is a consistent increase in number of runners attending the races, however ages and finish times remain consistent. This trend looks to be consistent over the 13 period analyzed with only a few outliers in each case.