Cherry Blossom Analysis

Matt Chinchilla, Rikel Djoko, Drew Larsen

9/20/2020

Business Case:

The Credit Union Cherry Blossom Ten Mile Race is an annual road race that takes place in Washington D.C. . The race has been held since 1973 and during that time valuable race result data has been collected and is available on the race's official website. The objective of this analysis is to extract race result data for female runners between the years of 1999 to 2012, a total of 14 years. Using this data, we hope to be able to help race planners gain new insight into patterns and trends as they relate to female runners of the race.

A few questions explored in this analysis include: Have age distributions changed over the years? Have race times increased or decreased both in total and by age groups for female runners? We will also look for trends and other insights within the data that may be valuable to bring to the attention of race planners. With a better understanding of female participants, race planners can make adjustments in routes, sponsorship outreach, and marketing that could help increase female participation and improve the overall race experience.

Data Extraction(Prep):

To collect the necessary race data our team will be using software to extract the race results data published directly from the Cherry Blossom Ten Mile Race website. This technique is known as web scrapping and is widely used to collect data from the internet. The race results data published on the Cherry Blossom website is freely available to the public and there are no known restriction to either collecting (scrapping) this data or analyzing it.

The web scrapping process is a fairly straightforward one as it relates to this project. The website itself is constructed using Hypertext Markup Language or HTML. HTML is what is known as a markup language, but put simply, there are tags within HTML that give the web page its structure. In viewing the underlying HTML code, we can find the tags that encapsulate the data we are interested in. Once we know the relevant tags related to the content, we instruct our software to search through the websites HTML code find the tags we are interested in and scrape the data contained in the tags. This process is repeated for each web page of race results from 1999 to 2012 until all the raw race results data has been extracted. The next section will review how the raw data is then transformed to a usable state for analysis.

Below are the character lengths scrapped from the website for each year.

```
## 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012
## 2356 2166 2972 3335 3543 3899 4334 5436 5691 6397 8324 8854 9030 9729
```

Below is an example of the scrapped raw data.

```
##
## Place Num   Name                 Ag Hometown           Net     Gun
## ===== ===== ==================== == ================== ======= =======
##     1  6002 Elana MEYER          34 Rep Of S.africa      52:15   52:16#
##     2  6004 Lydia GRIGORIEVA     27 Russia               53:12   53:15#
##     3  6019 Eyerusalem KUMA      20 Ethiopia             53:16   53:19#
##     4  6007 Milena GLUSAC        25 Usa                  53:33   53:34#
##     5  6012 Olga KOVPOTINA       31 Russia               54:01   54:03#
##     6  6020 Merima HASHIM        20 Ethiopia             54:29   54:31#
##     7  6005 Jane OMORO           27 Kenya                55:05   55:08#
```

Data Extraction (Execution):

Raw data extraction:

After separating the data into individual rows, we had to extract the variables from the raw data. We use the headers at the top and the row with the equal signs to guide our data extraction. For most years, the breaks in the row of equal signs correspond to a new variable. We use the extract variables function below to create matrices from the raw data. The function takes the raw data that is separated by row, finds the locations of the breaks in the row with the equal signs, and uses that location as a guide where to split every line in the raw data into a new variable. These variables are then named using the header row above the equal sign line. We found that year 2001 did not have a header row but did have the equal sign separator. 2002 and 2001 had the same header row structure, so we copy the header row from 2002 and use it for 2001. 2011 had a problem initially with parsing using UTF-8. Changing the htmlParse function in the extractResTable3 function to have encoding = 'latin1' solved the problem. It was also found that 2006 did not have a separation in the equal sign row between location and time, so we manually changed an equal sign to a space in the row so our function would automatically separate these variables. Our extractVariables function was successfully run on all raw data objects and outputted a list of individual matrices for each year.

Clean up:

After extracting raw results, the data needed significant amounts of cleaning. There were NA's and outliers all over the place. We had to assign header names and change data types. Header names were addressed in our extractVariables function, and the remaining variables are ("name", "home", "ag", "gun", "net", "time"). Name is the name of the runner, home is their home location, ag is their age, gun is their gun time (if applicable), net is their net time (if applicable) and time is their race time (if applicable). If we have a net time variable, we use that to calculate time in the final dataframe. If we have a gun time variable but no net time variable, we use that to calculate time in the final dataframe. If we don't have a gun or net time but we have a time variable, we use that to calculate time in the final dataframe. After addressing header names, we changed the age variable to numeric. We had to do some data parsing to change the time variable to numeric as well: the data came in as hours:minutes:seconds. We were interested in analyzing the time variable in terms of minutes, so we multiplied hours by 60, divided seconds by 60 and kept the minutes, and added those 3 variables together to get total minutes. This result was then turned into a numeric variable

Missing values and outliers:

There were outliers and NA values in both of our numeric columns: age and time. These needed to be dealt with. Looking at box plots, it was clear that 2003 was definitely an issue, and 2009, 2001 and 2011 could potentially be issues with young runners. To deal with 2003, we found that the age variable fluctuated by a column in either direction as we went down the raw data. Increasing the search by 1 column for the age variable solved the issue with 2003 and 2011. In 2001, there was a racer with an age of 0. When we went back and looked at the raw data, it was found that this racer was listed as 0, so we just removed this racer from the data. In 2011, there was a racer that was 7 years old. This was confirmed in the raw data. While the racer was young, it was determined that it was possible for a 7-year-old to run a 10-mile race, so the data point was kept. There were still a few outliers in the data frame, but those were kept in for analysis in our EDA. Now that we dealt with NA's and outliers in the age variable, we decided to look at the time variable. Stars and hashtags in the time variable were messing up our multiplication initially. Removing those in the initial time calculation function removed a significant amount of NA's in many years. There were still NA's in 2002 and 2006. 2006 had times that were in the location variable. It was determined that the data wasn't being parsed correctly due to a missing break in the equal sign row. The break was manually inputted, which fixed the NA's in 2006. The one remaining NA in 2002 was a footer line, which was removed. Looking at box plots of the run times by year, nothing stood out as problematic in terms of outliers, so we proceed with our analysis.

Data Description

dataframe with 75971 observations on the following variables:

- home - is the coutry of residency of the runner
- runTime - is the is the official time from starting gun to finish line.
- sex - is the gender of the participant F for woman and M for man in this case we only have all woman runner
- age - is the age of the runner
- name - is the name of runner

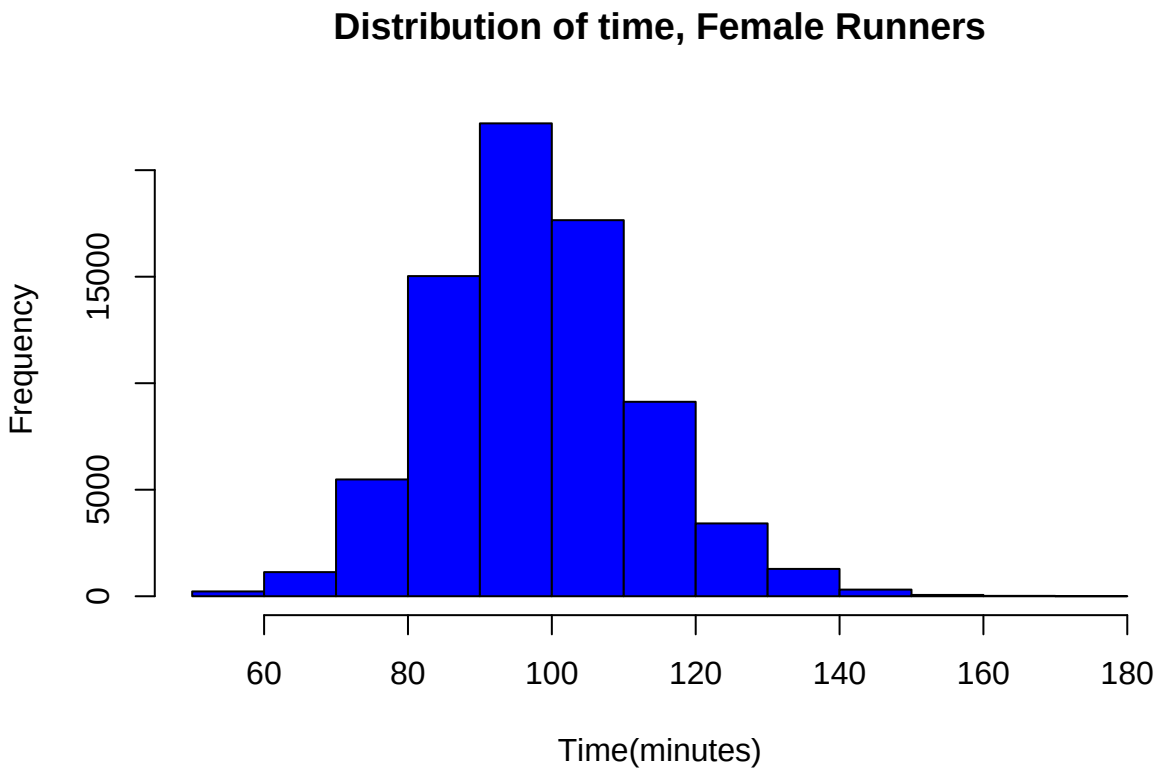
Below are high level summary statistics of our overall data

```
## 'data.frame':    75970 obs. of  6 variables:
##  $ year   : int  1999 1999 1999 1999 1999 1999 1999 1999 1999 1999 ...
##  $ sex    : chr  "F" "F" "F" "F" ...
##  $ name   : chr  "Jane Omoro            " "Jane Ngotho           " "Lidiya Grigoryeva     " "Eunice 
##  $ home   : chr  "Kenya              " "Kenya              " "Russia             " "Kenya
##  $ age    : num  26 29 NA 20 29 24 38 NA 27 30 ...
##  $ runTime: num  53.6 53.6 53.7 53.9 54.1 ...
```

```
##       year          sex                name               home
##  Min.   :1999   Length:75970       Length:75970       Length:75970
##  1st Qu.:2005   Class :character   Class :character   Class :character
##  Median :2008   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2007
##  3rd Qu.:2010
##  Max.   :2012
##
##       age            runTime
##  Min.   : 7.00   Min.   : 51.73
##  1st Qu.:27.00   1st Qu.: 88.53
##  Median :32.00   Median : 97.33
##  Mean   :33.85   Mean   : 98.09
##  3rd Qu.:39.00   3rd Qu.:106.78
##  Max.   :87.00   Max.   :177.52
##  NA's   :21
```

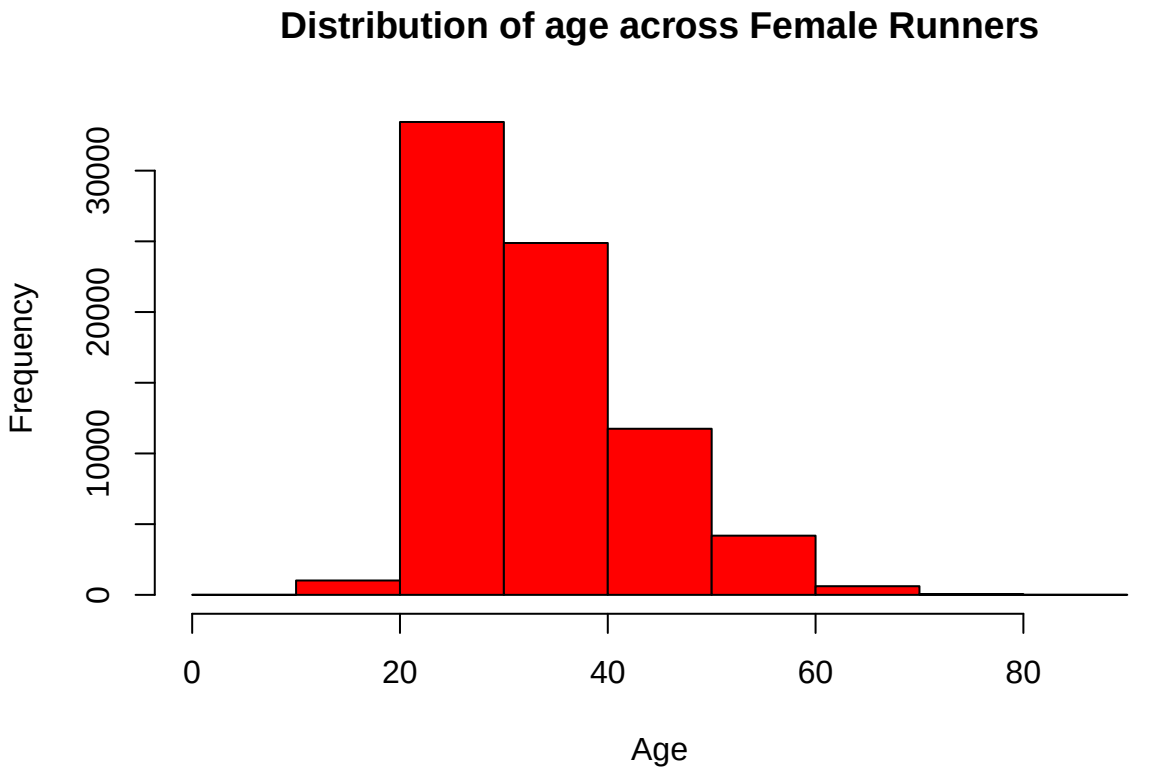
Distribution of times:

From the run time distribution below we can see that the run time is normally distributed from 1999-2012. This means that the expected female run time is around 98 Min. The histogram below shows the overall distribution of female runner times.



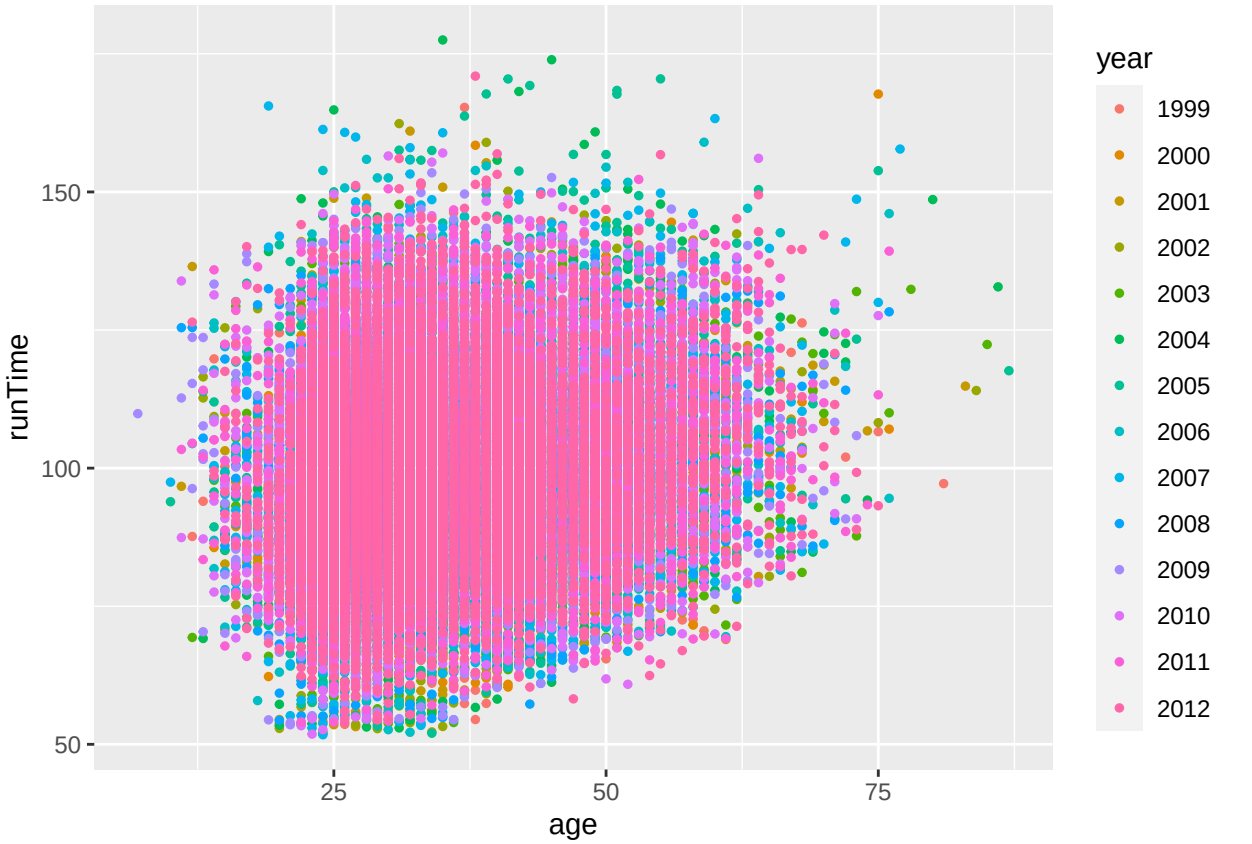
Age Distribution:

From the age distribution below we can see the participant's age goes from 10 to 70 and most of the participant are between 20-40. The data is skewed to the right so we have some other participant form 40 to 60.



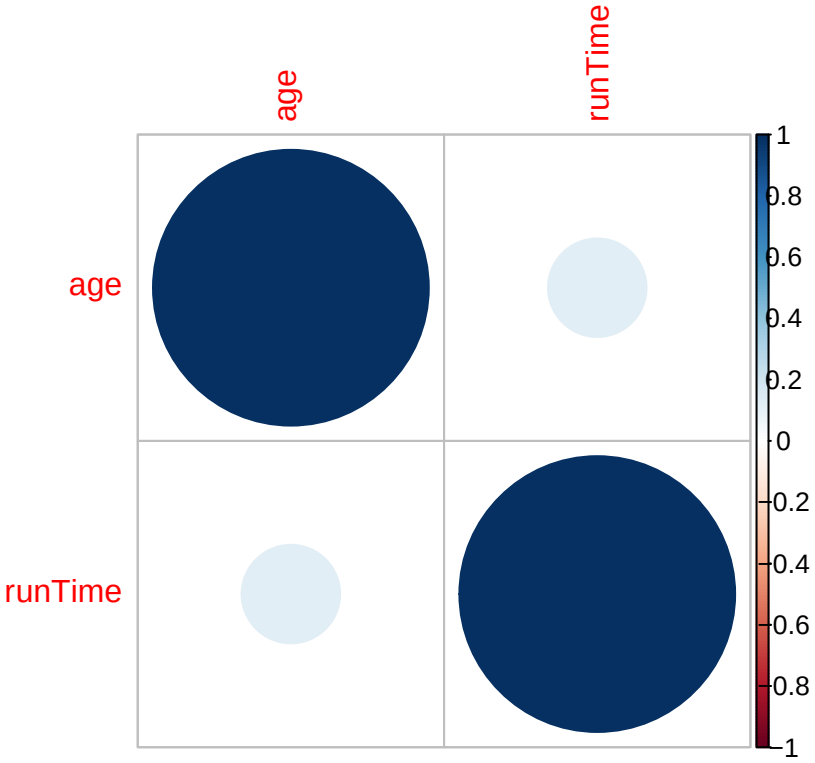
Scatterplot:

The scatter plot below across years doesn't show any particular linear trend but shows a big cluster which tells us that year by year most of the participants are between age 25 and 50 and runtimes are between 75 and 125 min.



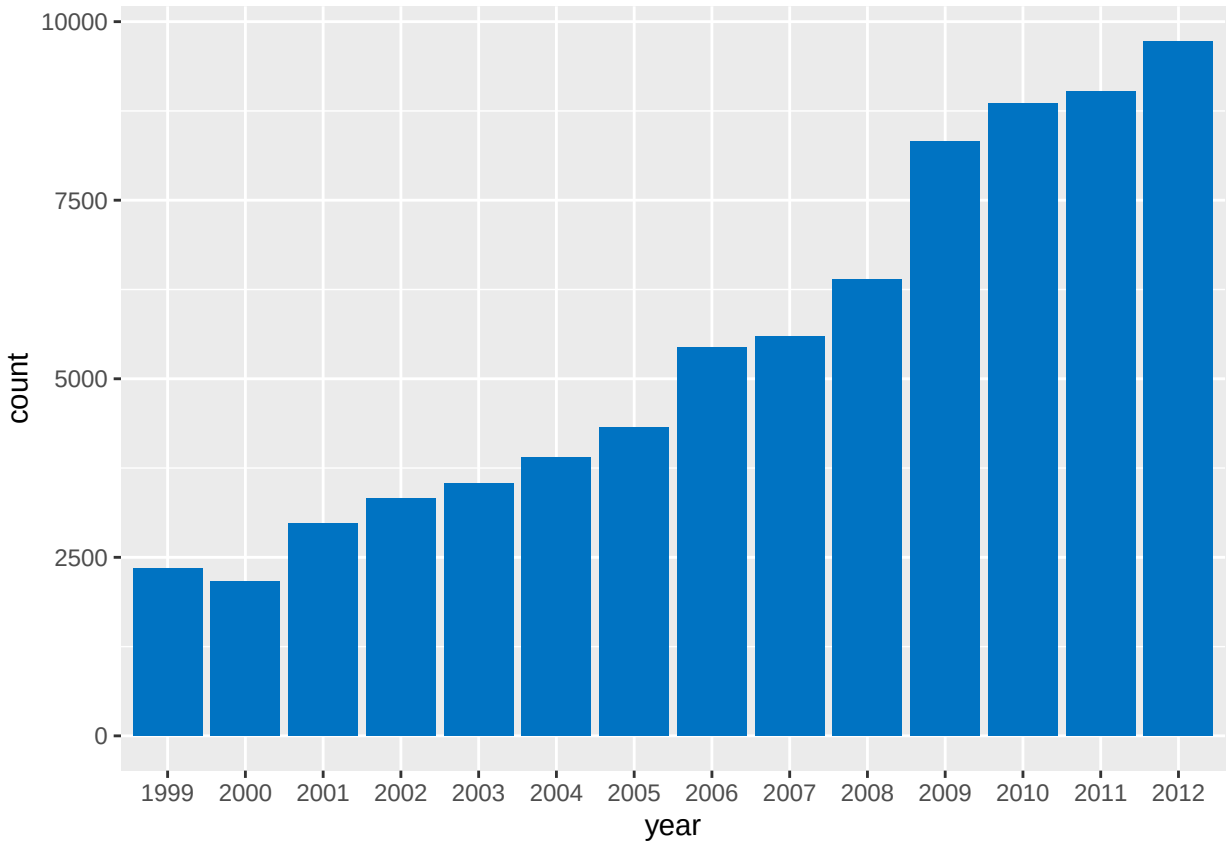
Correlation Plot:

The correlation plot below is between the age and run time, confirmed on the idea of weak correlation between the two variables.



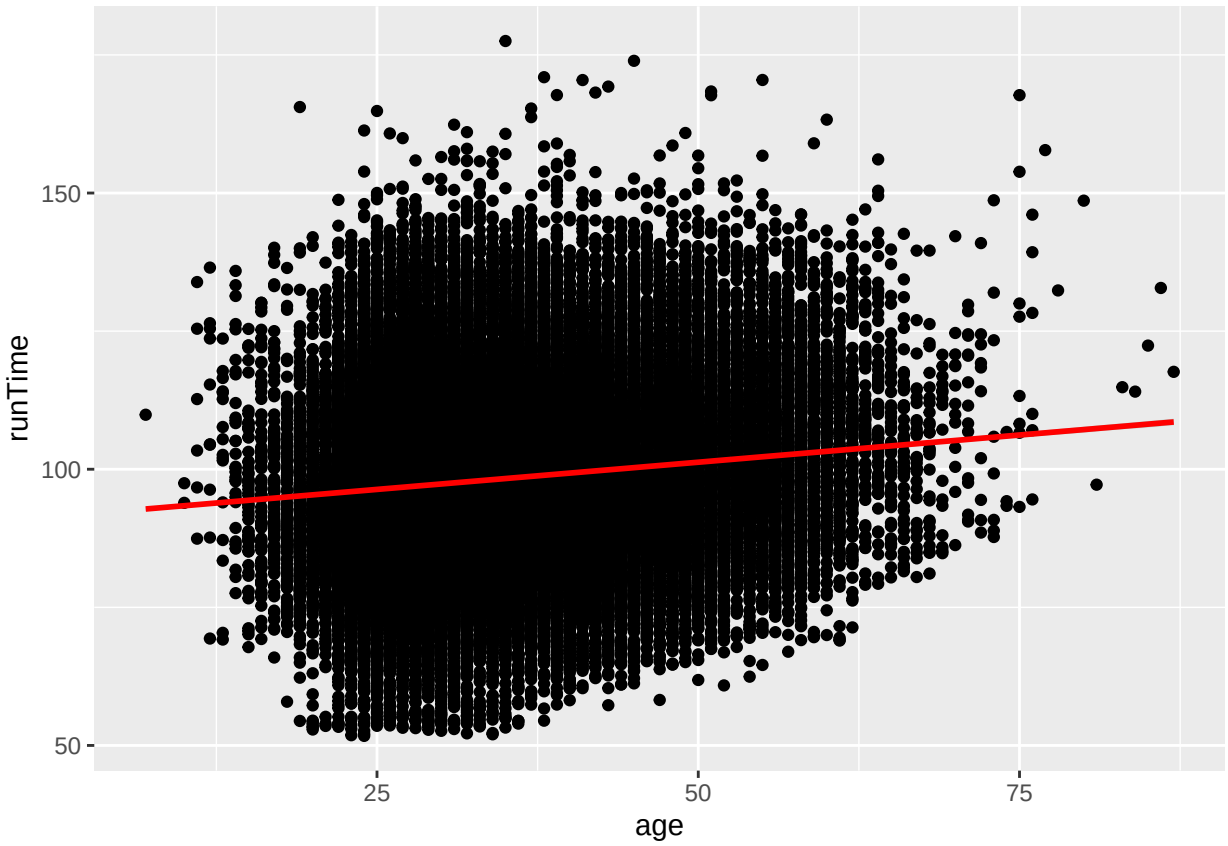
Bar Plot of # of participants per race:

As you can see from the plot below the number of participants is increasing every year, from 1999 to 2012 there are almost 4X the number of participants, growing from 2500 to 10000.



Scatter plot of run time vs. age, not colored:

From the scatter plot below we can see there's not a clear linear trend or correlation between the age and the run time.

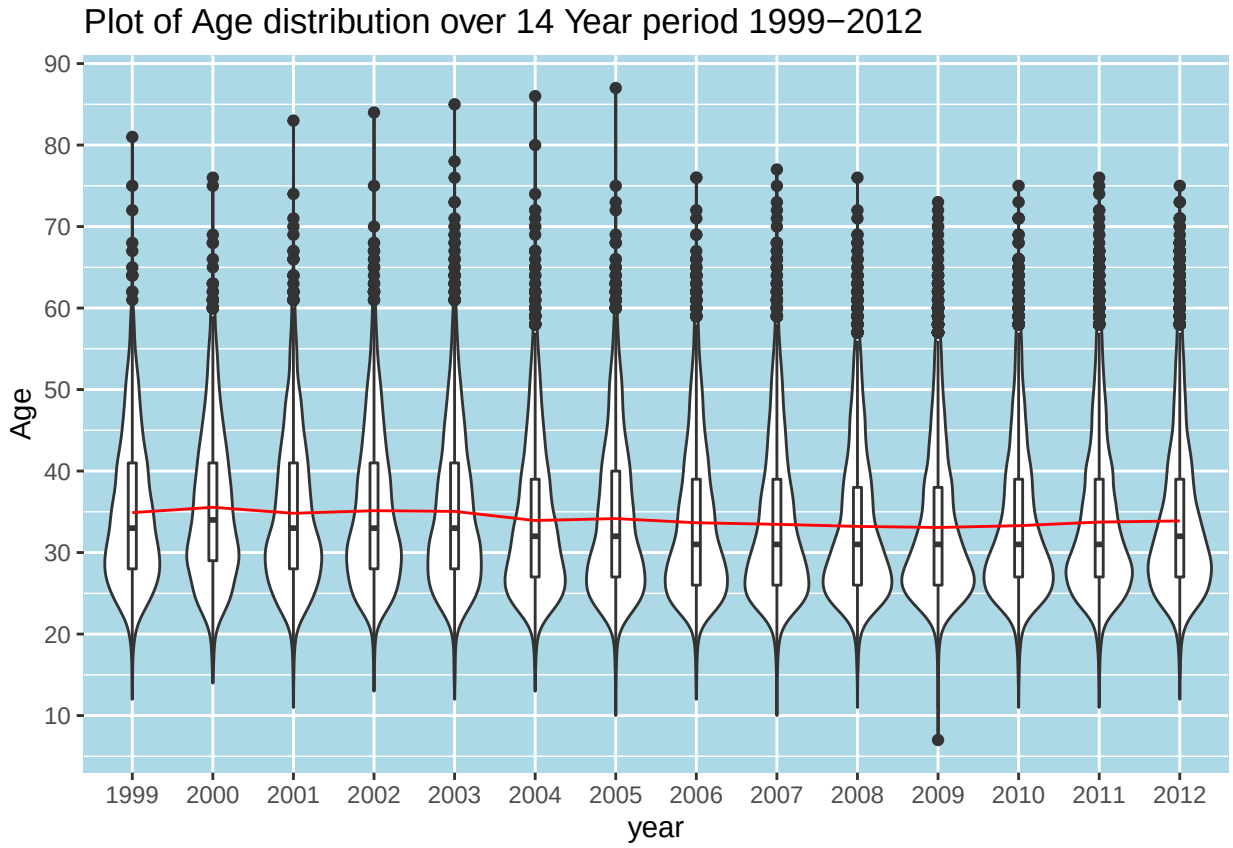


Business Analysis

Analysis: Have the ages of runners changed over the years?

Through the analysis below we have come to the conclusion that there is a statistically significant decrese in age over time. The mean age seemed to float around 35 for 1999 - 2002, then drop down to 34ish for 2004 and 2005, then drop down below 34 for 2006 - 2010, bottoming out at 33.07 in 2009. The age then begins to come back up in 2010, 2011 and 2012, where the mean age is 33.88. Is a difference of a year or two different enough to make differences to the race structure? We expect 2013 to have a mean age between 33.5 - 34.

The violin plot helps us to visualize how female age has been distributed over the 14 year period.The red line running through the plots represents the mean age each year. We can see that overall the age distributions has not varied considerably. Consistently we can see that the majority of female racers are between the ages of 25 to 35 years old with the overall mean around 34 and the overall median just a bit lower at around 32. There isn't a lot of evidence that the mean age has changed year over year. We will run a more formal analysis to see if this is true or not.




```
## [1] "The mean female age is"
```

```
## [1] 33.84923
```

```
## [1] "The median female age is"
```

```
## [1] 32
```

We are interested in running an ANOVA test to see if there is a statistical difference between the mean age in any year. First, we look at assumptions. The 3 assumptions are that the observations are independently and randomly drawn from the population, the data is normally distributed in each year, and the populations have a common variance. Looking at the boxplot above, the data visually look to have a common variance, but there does seem to be a skew toward higher values in these distributions. I don't think ht's enough of a right skew to be a problem, so we continue as if the populations are normal. As far as independence goes, we know that the populations have gotten larger as time has gone on, but I don't think that the age distribution of one race would depend on the age distribution of another. Also, these data were not randomly chosen from a population, but due to the nature of the data (the population of the ages of people running the race each year), it wouldn't make sense to do a random selection technique. Therefore, I think we can run an ANOVA test in order to compare these year's distribution of age.

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## year            1   18025   18025   212.6 <2e-16 ***
## Residuals   75947 6439701      85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 21 observations deleted due to missingness
```

Using our ANOVA test, it appears that there is a statistical significant difference in the mean age per year. Which years? We use the Bonferroni correction for multiple comparisons. Bonferroni was chosen due to it being a fairly conservative test: we want to make sure that the ages are different and the difference between

ages in each year isn't due to us running many tests. It seems that the further the year is away, the more likely that the year has a significantly different age.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fullwomendf$age and fullwomendf$year
##
##      1999    2000    2001    2002    2003    2004    2005    2006
## 2000 1.00000 -       -       -       -       -       -       -
## 2001 1.00000 0.42246 -       -       -       -       -       -
## 2002 1.00000 1.00000 1.00000 -       -       -       -       -
## 2003 1.00000 1.00000 1.00000 1.00000 -       -       -       -
## 2004 0.00502 4.4e-09 0.00704 2.6e-06 1.5e-05 -       -       -
## 2005 0.17314 9.9e-07 0.28028 0.00045 0.00211 1.00000 -       -
## 2006 3.6e-06 4.0e-14 2.6e-06 2.1e-11 1.8e-10 1.00000 0.53922 -
## 2007 2.1e-08 < 2e-16 9.1e-09 9.8e-15 9.8e-14 1.00000 0.01504 1.00000
## 2008 2.3e-12 < 2e-16 3.2e-13 < 2e-16 < 2e-16 0.01011 1.1e-05 0.81899
## 2009 1.9e-15 < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.00015 2.2e-08 0.03006
## 2010 5.1e-12 < 2e-16 5.8e-13 < 2e-16 < 2e-16 0.02957 2.9e-05 1.00000
## 2011 4.6e-06 1.6e-14 2.8e-06 6.1e-12 5.9e-11 1.00000 1.00000 1.00000
## 2012 0.00012 1.6e-12 9.9e-05 8.2e-10 7.3e-09 1.00000 1.00000 1.00000
##      2007    2008    2009    2010    2011
## 2000 -       -       -       -       -
## 2001 -       -       -       -       -
## 2002 -       -       -       -       -
## 2003 -       -       -       -       -
## 2004 -       -       -       -       -
## 2005 -       -       -       -       -
## 2006 -       -       -       -       -
## 2007 -       -       -       -       -
## 2008 1.00000 -       -       -       -
## 2009 1.00000 1.00000 -       -       -
## 2010 1.00000 1.00000 1.00000 -       -
## 2011 1.00000 0.03845 0.00019 0.11618 -
## 2012 0.71444 0.00059 5.1e-07 0.00156 1.00000
##
## P value adjustment method: bonferroni
```

We ran the first analysis to find where differences were. This analysis looks at which years have higher ages. It looks like the ages aren't getting higher over time. The only statistically significant times where age got higher as year got higher is comparing 2008 to 2011 and 2012, 2009 to 2011 and 2012, and 2010 to 2012.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fullwomendf$age and fullwomendf$year
##
##      1999    2000    2001    2002    2003    2004    2005    2006
## 2000 0.78208 -       -       -       -       -       -       -
## 2001 1.00000 1.00000 -       -       -       -       -       -
## 2002 1.00000 1.00000 1.00000 -       -       -       -       -
## 2003 1.00000 1.00000 1.00000 1.00000 -       -       -       -
## 2004 1.00000 1.00000 1.00000 1.00000 1.00000 -       -       -
## 2005 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 -       -
```

```
## 2006 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 -
## 2007 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
## 2008 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
## 2009 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
## 2010 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
## 2011 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
## 2012 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000 1.00000
##      2007    2008    2009    2010    2011
## 2000 -       -       -       -       -
## 2001 -       -       -       -       -
## 2002 -       -       -       -       -
## 2003 -       -       -       -       -
## 2004 -       -       -       -       -
## 2005 -       -       -       -       -
## 2006 -       -       -       -       -
## 2007 -       -       -       -       -
## 2008 1.00000 -       -       -       -
## 2009 1.00000 1.00000 -       -       -
## 2010 1.00000 1.00000 1.00000 -       -
## 2011 1.00000 0.01922 9.7e-05 0.05809 -
## 2012 0.35722 0.00030 2.6e-07 0.00078 1.00000
##
## P value adjustment method: bonferroni
```

Typically, it looks like age gets younger or stay the same as we move into the future. The outliers here are 2011 and 2012, which seem to have gotten older compared to 2008, 2009 and 2010.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fullwomendf$age and fullwomendf$year
##
##      1999    2000    2001    2002    2003    2004    2005    2006
## 2000 1.00000 -       -       -       -       -       -       -
## 2001 1.00000 0.21123 -       -       -       -       -       -
## 2002 1.00000 1.00000 1.00000 -       -       -       -       -
## 2003 1.00000 1.00000 1.00000 1.00000 -       -       -       -
## 2004 0.00251 2.2e-09 0.00352 1.3e-06 7.3e-06 -       -       -
## 2005 0.08657 4.9e-07 0.14014 0.00022 0.00105 1.00000 -       -
## 2006 1.8e-06 2.0e-14 1.3e-06 1.0e-11 8.9e-11 1.00000 0.26961 -
## 2007 1.0e-08 < 2e-16 4.6e-09 4.9e-15 4.9e-14 0.70182 0.00752 1.00000
## 2008 1.1e-12 < 2e-16 1.6e-13 < 2e-16 < 2e-16 0.00505 5.4e-06 0.40949
## 2009 9.6e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 7.6e-05 1.1e-08 0.01503
## 2010 2.6e-12 < 2e-16 2.9e-13 < 2e-16 < 2e-16 0.01479 1.5e-05 1.00000
## 2011 2.3e-06 7.9e-15 1.4e-06 3.0e-12 3.0e-11 1.00000 0.53005 1.00000
## 2012 5.9e-05 8.0e-13 5.0e-05 4.1e-10 3.7e-09 1.00000 1.00000 1.00000
##      2007    2008    2009    2010    2011
## 2000 -       -       -       -       -
## 2001 -       -       -       -       -
## 2002 -       -       -       -       -
## 2003 -       -       -       -       -
## 2004 -       -       -       -       -
## 2005 -       -       -       -       -
## 2006 -       -       -       -       -
## 2007 -       -       -       -       -
```

```
## 2008 1.00000 -       -       -       -
## 2009 0.64037 1.00000 -       -       -
## 2010 1.00000 1.00000 1.00000 -       -
## 2011 1.00000 1.00000 1.00000 1.00000 -
## 2012 1.00000 1.00000 1.00000 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

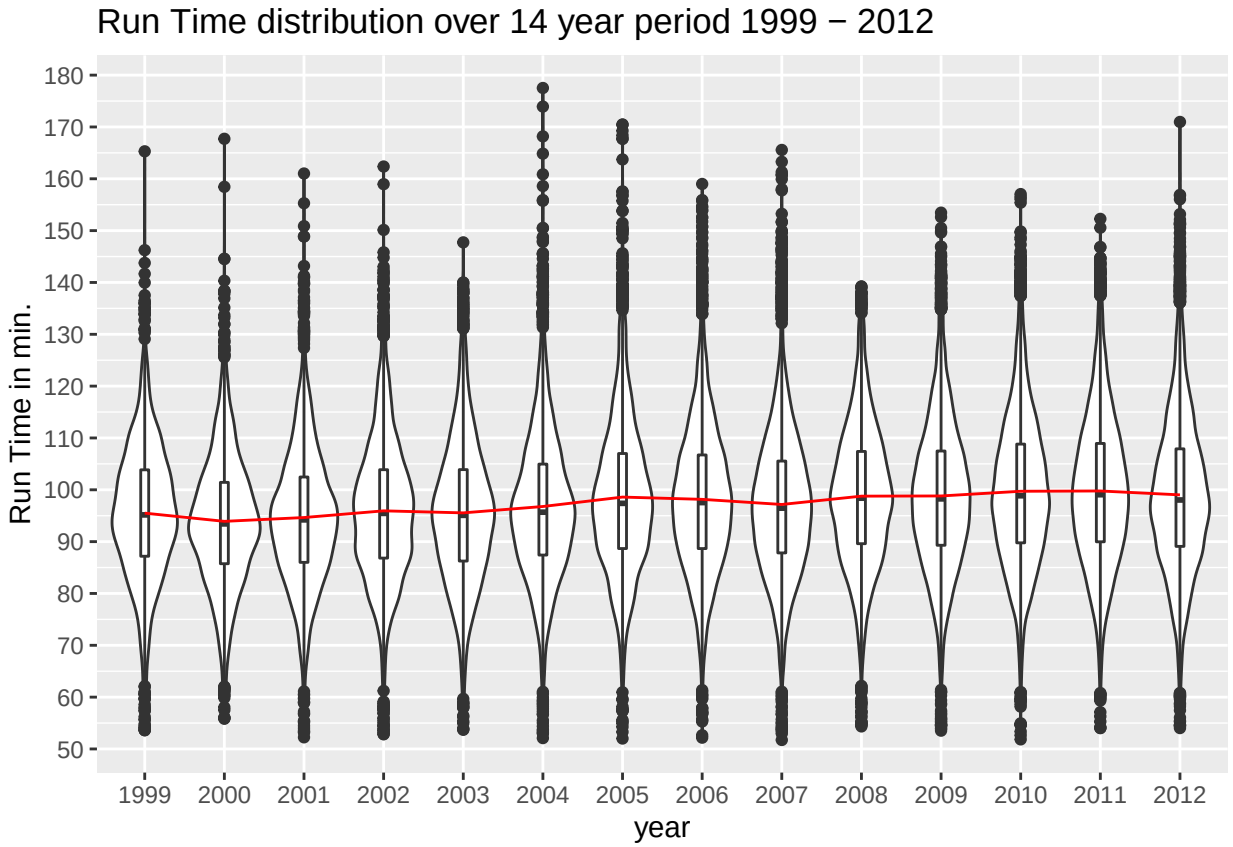
We have statistically significant differences, but are the differences practical? The mean age seemed to float around 35 for 1999 - 2002, then drop down to 34ish for 2004 and 2005, then drop down below 34 for 2006 - 2010, bottoming out at 33.07 in 2009. The age then begins to come back up in 2010, 2011 and 2012, where the mean age is 33.88. Is a difference of a year or two different enough to make differences to the race structure? I expect 2013 to have a mean age between 33.5 - 34.

```
## # A tibble: 14 x 2
##     year   age
##    <int> <dbl>
##  1  1999  34.9
##  2  2000  35.6
##  3  2001  34.8
##  4  2002  35.1
##  5  2003  35.1
##  6  2004  33.9
##  7  2005  34.2
##  8  2006  33.7
##  9  2007  33.5
## 10  2008  33.2
## 11  2009  33.1
## 12  2010  33.3
## 13  2011  33.7
## 14  2012  33.9
```

Analysis: Have the times of runners changed over the years?

Through the analysis below we have come to the conclusion that there is a statistically significant increase in time as years have increased but given this increase is only 42 seconds per-mile across the largest variances we do not think that in a practical sense the change is significant enough to warrant any structural changes to the race.

Looking at the violin plots below we are able to see the distributions of run time over the years along with the mean which is represented by the red line running through each plot. It doesn't look as if run time has changed much over time. We will run a more formal statistical analysis to check if times have changed statistically.




```
## [1] "The mean female run times"
```

```
## [1] 98.08692
```

```
## [1] "The median female run times"
```

```
## [1] 97.33333
```

We are interested in running an ANOVA test to see if there is a statistical difference between the mean age in any year. First, we look at assumptions. The 3 assumptions are that the observations are independently and randomly drawn from the population, the data is normally distributed in each year, and the populations have a common variance. Looking at the boxplot above, the data visually look to have a common variance. There may have been a slightly smaller variance in early years, but I think that has more to do with the 5-10 longest times in later years being fairly high. Looking at the size of the whiskers in the box plot, they look to be fairly constant over the years. Next we look at normality. The distributions appear to be normal, aside from some rather large outliers. The 5-10 longest times wouldn't have a huge effect on the shape of the distribution, since our sample sizes per year is in the thousands. As far as independence goes, we know that the populations have gotten larger as time has gone on, but I don't think that the time distribution of one race would depend on the time distribution of another. Also, these data were not randomly chosen from a population, but due to the nature of the data (the population of the times of people running the race each year), it wouldn't make sense to do a random selection technique. Therefore, I think we can run an ANOVA test in order to compare these year's distribution of times.

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## year            1   167884  167884   838.2 <2e-16 ***
## Residuals   75968 15215007     200
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a statistically significant difference in the mean run time of at least one of the years. Which years?

It looks like many years have statistically different times. This is a two sided test, so we are not sure if times have gotten larger or smaller, just that they're different. Now that we know they have changed in many years, we want to know how they have changed.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fullwomendf$runTime and fullwomendf$year
##
##      1999    2000    2001    2002    2003    2004    2005    2006
## 2000 0.01560 -       -       -       -       -       -       -
## 2001 1.00000 1.00000 -       -       -       -       -       -
## 2002 1.00000 1.8e-05 0.02555 -       -       -       -       -
## 2003 1.00000 0.00202 0.93473 1.00000 -       -       -       -
## 2004 0.04654 3.8e-12 5.6e-08 1.00000 0.01666 -       -       -
## 2005 1.0e-15 < 2e-16 < 2e-16 3.6e-14 < 2e-16 5.3e-07 -       -
## 2006 2.0e-12 < 2e-16 < 2e-16 9.3e-11 1.1e-15 0.00029 1.00000 -
## 2007 0.00013 < 2e-16 3.6e-13 0.00671 8.8e-06 1.00000 5.6e-05 0.02103
## 2008 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 3.4e-10 1.00000 1.00000
## 2009 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 1.2e-11 1.00000 0.76543
## 2010 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.00164 1.4e-08
## 2011 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.00057 2.6e-09
## 2012 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 4.9e-15 1.00000 0.02833
##      2007    2008    2009    2010    2011
## 2000 -       -       -       -       -
## 2001 -       -       -       -       -
## 2002 -       -       -       -       -
## 2003 -       -       -       -       -
## 2004 -       -       -       -       -
## 2005 -       -       -       -       -
## 2006 -       -       -       -       -
## 2007 -       -       -       -       -
## 2008 5.0e-08 -       -       -       -
## 2009 1.7e-09 1.00000 -       -       -
## 2010 < 2e-16 0.00420 0.00230 -       -
## 2011 < 2e-16 0.00133 0.00064 1.00000 -
## 2012 4.8e-13 1.00000 1.00000 0.07353 0.02486
##
## P value adjustment method: bonferroni
```

Looking into the future, it seems that in general, times have gotten larger. Comparing 1999 to future years, all years from 2004 - 2012 have a statistically significant increase in run time.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fullwomendf$runTime and fullwomendf$year
##
##      1999    2000    2001    2002    2003    2004    2005    2006
## 2000 1.00000 -       -       -       -       -       -       -
## 2001 1.00000 1.00000 -       -       -       -       -       -
## 2002 1.00000 9.1e-06 0.01278 -       -       -       -       -
## 2003 1.00000 0.00101 0.46737 1.00000 -       -       -       -
## 2004 0.02327 1.9e-12 2.8e-08 0.55985 0.00833 -       -       -
## 2005 5.1e-16 < 2e-16 < 2e-16 1.8e-14 < 2e-16 2.6e-07 -       -
```

```
## 2006 1.0e-12 < 2e-16 < 2e-16 4.7e-11 5.7e-16 0.00015 1.00000 -
## 2007 6.6e-05 < 2e-16 1.8e-13 0.00336 4.4e-06 1.00000 1.00000 1.00000
## 2008 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 1.7e-10 1.00000 0.84695
## 2009 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 6.0e-12 1.00000 0.38271
## 2010 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.00082 7.2e-09
## 2011 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.00028 1.3e-09
## 2012 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 2.5e-15 1.00000 0.01417
##      2007    2008    2009    2010    2011
## 2000 -       -       -       -       -
## 2001 -       -       -       -       -
## 2002 -       -       -       -       -
## 2003 -       -       -       -       -
## 2004 -       -       -       -       -
## 2005 -       -       -       -       -
## 2006 -       -       -       -       -
## 2007 -       -       -       -       -
## 2008 2.5e-08 -       -       -       -
## 2009 8.5e-10 1.00000 -       -       -
## 2010 < 2e-16 0.00210 0.00115 -       -
## 2011 < 2e-16 0.00067 0.00032 1.00000 -
## 2012 2.4e-13 1.00000 1.00000 1.00000 1.00000
##
## P value adjustment method: bonferroni
```

Looking at years where run time has decreased as we look into the future, only 2000 is less than 1999, 2007 is less than 2005 and 2006, and 2012 is less than 2010 and 2011. The trend seems to be an increasing mean time as we move into the future, aside from 2000, 2007 and 2012 being slight outliers.

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  fullwomendf$runTime and fullwomendf$year
##
##      1999   2000   2001   2002   2003   2004   2005    2006   2007
## 2000 0.0078 -      -      -      -      -      -       -      -
## 2001 1.0000 1.0000 -      -      -      -      -       -      -
## 2002 1.0000 1.0000 1.0000 -      -      -      -       -      -
## 2003 1.0000 1.0000 1.0000 1.0000 -      -      -       -      -
## 2004 1.0000 1.0000 1.0000 1.0000 1.0000 -      -       -      -
## 2005 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 -       -      -
## 2006 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  -      -
## 2007 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 2.8e-05 0.0105 -
## 2008 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  1.0000 1.0000
## 2009 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  1.0000 1.0000
## 2010 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  1.0000 1.0000
## 2011 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  1.0000 1.0000
## 2012 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000  1.0000 1.0000
##      2008   2009   2010   2011
## 2000 -      -      -      -
## 2001 -      -      -      -
## 2002 -      -      -      -
## 2003 -      -      -      -
## 2004 -      -      -      -
## 2005 -      -      -      -
## 2006 -      -      -      -
```

```
## 2007 -      -      -      -
## 2008 -      -      -      -
## 2009 1.0000 -      -      -
## 2010 1.0000 1.0000 -      -
## 2011 1.0000 1.0000 1.0000 -
## 2012 1.0000 1.0000 0.0368 0.0124
##
## P value adjustment method: bonferroni
```

We know that there is a statistically significant increase in mean run time as years increase. Is there a practical difference? The lowest mean run time is about 94 minutes in 2000, while the longest mean run time is about 100 minutes in 2011. In a ten mile race, that is a mean increase of 42 seconds per mile. Is that a big enough difference to make changes in your race, considering that change occurred over 11 years? I expect 2013 to have a mean run time between 99 and 100 minutes for the 10 mile race.

```
## # A tibble: 14 x 2
##     year runTime
##    <int>   <dbl>
##  1  1999    95.5
##  2  2000    93.9
##  3  2001    94.6
##  4  2002    95.9
##  5  2003    95.5
##  6  2004    96.8
##  7  2005    98.6
##  8  2006    98.2
##  9  2007    97.2
## 10  2008    98.8
## 11  2009    98.8
## 12  2010    99.7
## 13  2011    99.8
## 14  2012    99.0
```

Conclusion

Based on our analysis of age and runner time we see statistically significant differences in both time and age but we do not believe that the differences are large enough to recommend or warrant any major changes to the Cherry Blossom 10 mile Race. We would recommend that given more time further analysis of runner home town/location could highlight new insights into questions such as where are most of the runners from, what countries are represented, and how popular is the race internationally. As we have seen in the summary data that the race itself is increasing in participation and has grown by more than 500 participants a year over the 14 year period from 1999 to 2012. We would recommend that based on the growth in popularity of the race considerations should be taken to add more facilities, water and water stations, gear/giveaways supplies, and make sure that any concessions are appropriately stocked for the increase in runners. We would also recommend a forecasting analysis of future race growth as a next step to help with budgeting, staffing, and purchase of supplies.