Steven Hayden, Kevin Mendonsa,
Joe Schueder, Nicole Wittlin
7333 QTW Case 2

Case 2: Cherry Blossom 10 Miler

Introduction and Overview

The Cherry Blossom 10 Mile race in DC is one of the most popular races in the area, and it is part of the Professional Road Running Organization race circuit. In the years from 1999 to 2012, the race has seen growth in the number of both female and male runners. While there was a slight dip in participation between 2000 and 2001, the year over year growth has been as high as 25-30%, as seen in **Figure 1**. Overall, the Cherry Blossom 10 miler has seen a greater increase in women runners (314% from 1999 to 2012), compared to a smaller increase of male runners (126% over the same time period). The bottom graph of **Figure 1** shows a larger proportion of women over the years, and by 2012, the race had growth to nearly 17,000 participants.

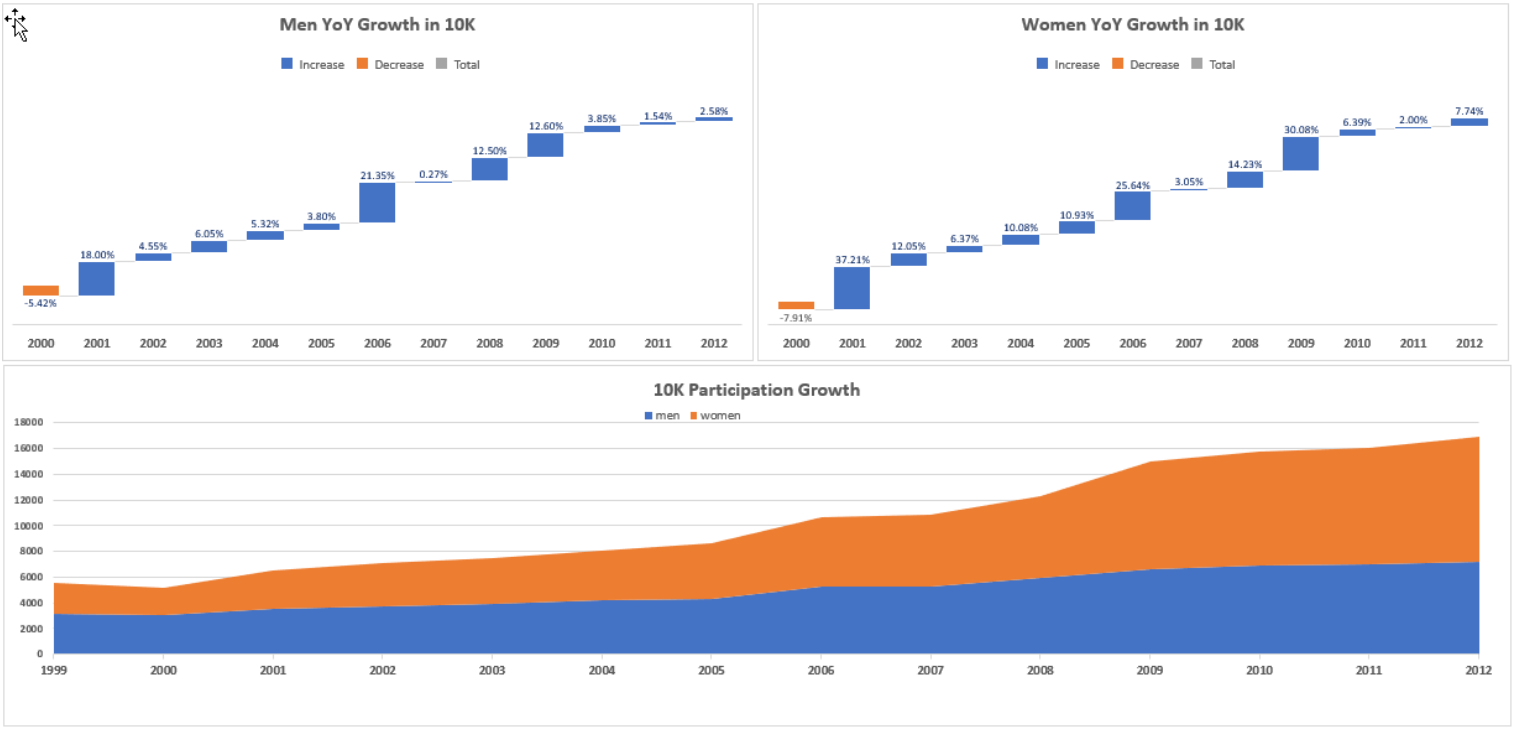

Figure 1: Year Over Year Participation by Gender

The annual event has also grown in complexity, with the seeding of elite runners for guaranteed entry, fundraising opportunities to guarantee a race spot for non-seeded runners, a team entry system and competition, and a lottery system to award entry to additional runners. In 2012, the slowest race times were well past two and a half hours, plus the many hours needed for set up and tear down. The city of Washington DC has put pressure on Cherry Blossom race organizers to limit the time of the race. Given the other Cherry Blossom activities happening around the same time, it has become more difficult to keep roads closed that length of time. Therefore, the race needs to take less time in the years ahead.

The race committee has asked our team to assess the past 13 years of race results to understand how the race has grown, how the age distribution of runners has changed, and how the pace of runners has changed. This information will become the foundation of changes to the Cherry Blossom 10 Mile race.

Data Preparation (Q7)

The team collected data by web scraping information from the Cherry Blossoms 10 Miler website. The results from 1999 to 2012 for both female and male runners were obtained by accessing each individual year's website in the directory found at: http://www.cherryblossom.org/results. We found that several years of data were stored in slightly different formats (i.e. wider page headers, variable spaces between lines of data) and had different names for the same attribute (i.e. time vs net time vs gun time to represent the total race time for each runner).

Data Scraping and Parsing

Our team designed and implemented a function that accommodated these format variations to effectively scrape the data from the website. The comprehensive methods we employed can be found in the respective R code and Python files. See **Code Directory** at the conclusion of this write up. The following are highlights of issues encountered by the team.

Women's Data

- For most years, the data was fairly consistent but there were anomalies that required additional specialized treatment in some years.
- For the 1999 data, the **"//pre"** node worked to obtain data, while for 2000, the **"//font"** node was needed. Furthermore, in 1999, the linefeed carriage return character **"\r"** was missing, and the newline character **"\n"** was used instead.
- Additionally, the header line **"===="** was not present in all years, making it rather difficult to find the header and the spacer columns. This required manual intervention to establish the variable lengths in 2001 and 2002.
- Various functions were used to trim leading and trailing whitespaces and replace extraneous characters and spaces (similar techniques but different syntax used in both).

Men's Data

- Similar issues existed for the men's data as the women's data with some exceptions.
- The biggest challenge in scraping the men's data was for year 2009. It required additional work to effectively scrape the data and then to structure, clean, and transform as needed.
- Lastly, 2009 data was formatted in a Word-like document on the website, where the end of each data line was indicated by **</pre><pre>.** This required a separate function to scrape, parse, and clean the data.

Data Cleaning

After compiling all the data, our initial review revealed additional inconsistencies and missing data. The team invested time to find missing ages by cross referencing against the "Searchable Results" page on the website; information was updated or marked as missing (NA) for more than 85 records. We also specifically reviewed times for participants under the age of 12; while times for some of these young runners may be questionable, the information published was taken at face value, with the assumption that there were too few to significantly impact the analysis.

Additional data cleaning and preparation included:

- Runners without race time were dropped from the data set.
- Race time was presented as "Time," "Net Time," or "Gun Time," depending on race year. For consistency in analysis, we used "Net Time" when available, followed by "Gun Time" or "Time." Details of these definitions can be found in **Appendix Figure A**, our data dictionary.

- Participations were binned into 10-year age brackets (with the exception of under 19 and above 80) to assist in understanding age trends.
- We created variables to analyze race accurately, given the variation in hours, minutes, seconds, and milliseconds of the time variable. Race pace was calculated mathematically (as the Cherry Blossom website had errors).

Analysis

With the objective to understand how participants' age and race time have changed over the years, the team first looked at women and men separately.

WOMEN

The first assessment by the team was to look at how the women in each of the 10-year age bins changed over time. As seen in **Figure 2**, the growth of women participants is most notable in the 19-29 and 29-39 categories age groups, while the other age categories grew more proportionally. Starting in 2009, this 20-year age span reflected a significant percentage of female participants (nearly 78% in 2009).

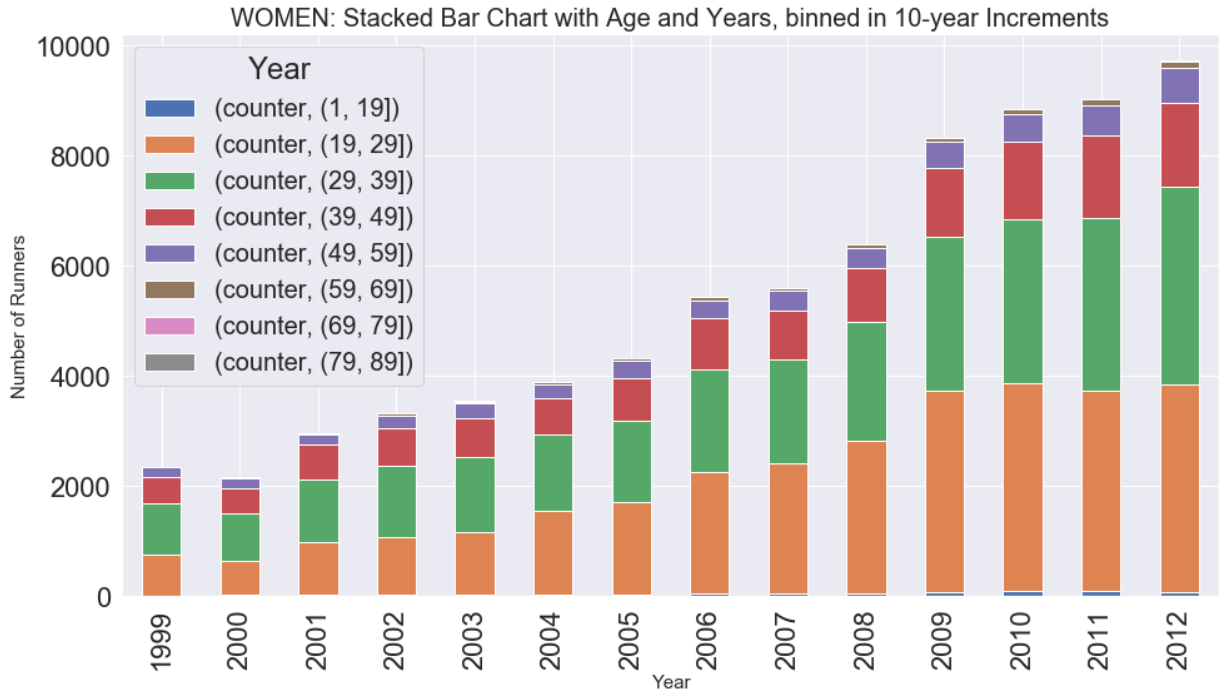

Figure 2: 10-Year Age Bins by Year

We looked at the same information presented another way: the number of Runners Group by Age for each year. **Figure 3** illustrates the significant increase in runners roughly age 20-40.

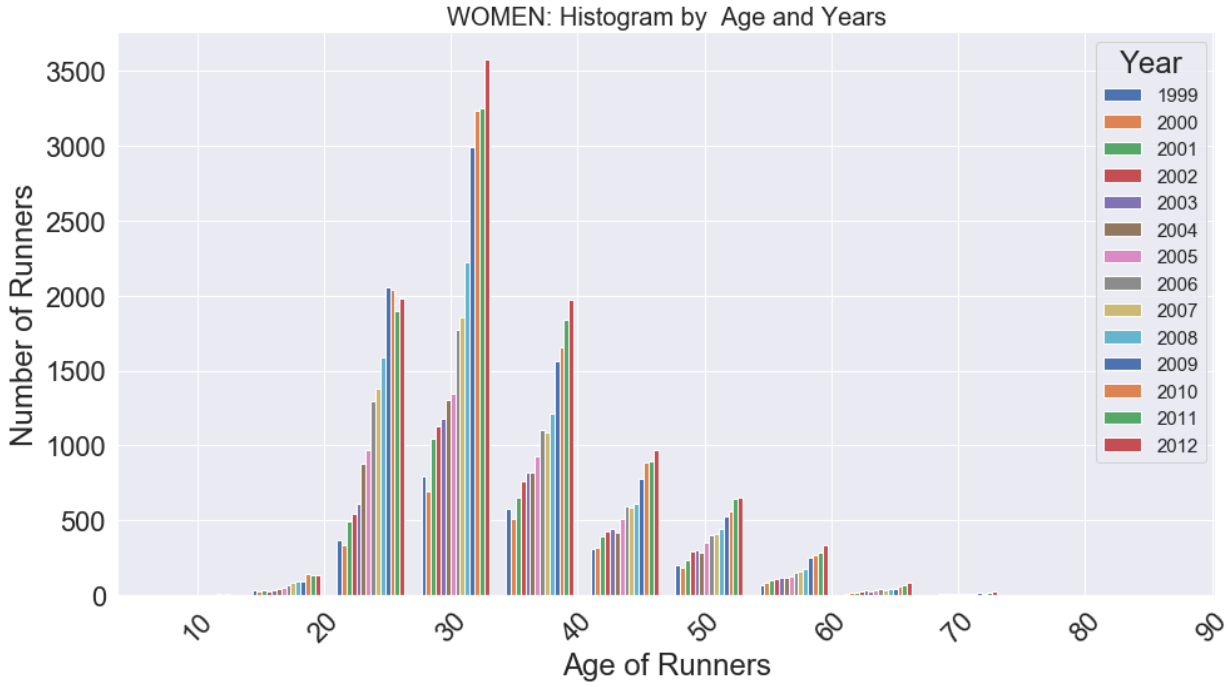

Figure 3: Runner Count by Age Groups Each Year

Then, we looked at two boxplots that compared women's age by year (**Figure 4**) and race time by year (**Figure 5**). The boxplots provided better insights into the trends for age and time. The red dotted line is equal to the median age in 2000, which saw the highest average of the oldest runners. The median age decreases slightly year over year, with a noticeable drop in 2006. There were no late 70 or 80+ aged runners in 2006, which also might contribute to a change in average age. Overall, there are not dramatic shifts in the distribution of the plots but certainly a younger trend at all quartiles.

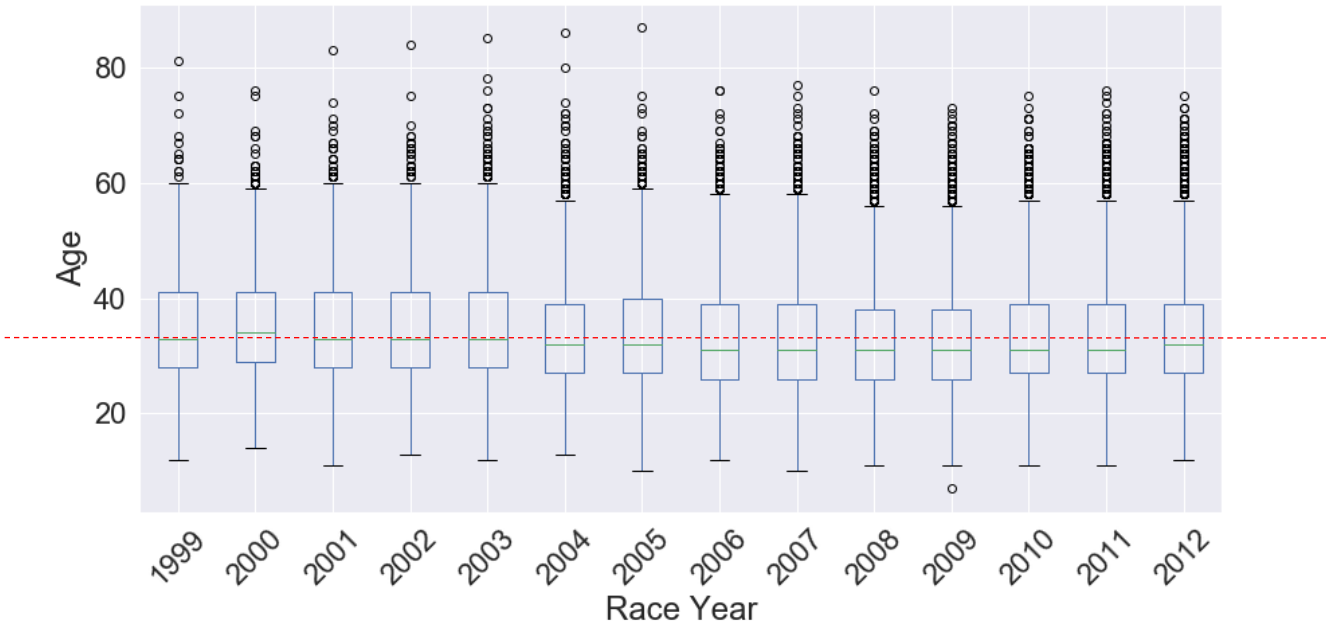

Figure 4: Boxplot of Year and Age

The second boxplot the team analyzed was a comparison of race time in minutes over the years to see if changes are occurring with the speed of female runners. The red dotted line reflects the median 2001 – a slow race from the early years – against the average race times for all years. As the years progress, the average race time does continue to lengthen by one or two minutes; upper quartiles show even larger variations in time. It also does not appear that the women are getting faster over the years, as seen in **Figure 5**.

To assess speed year over year, the team also looked at a density plot of female race times. **Figure 6** shows a better comparison of the distribution of race times over the years. While 2000 does show a marked difference from the others – it was the fastest year in the data set – the other years are roughly consistent when the distributions are plotted in this way.

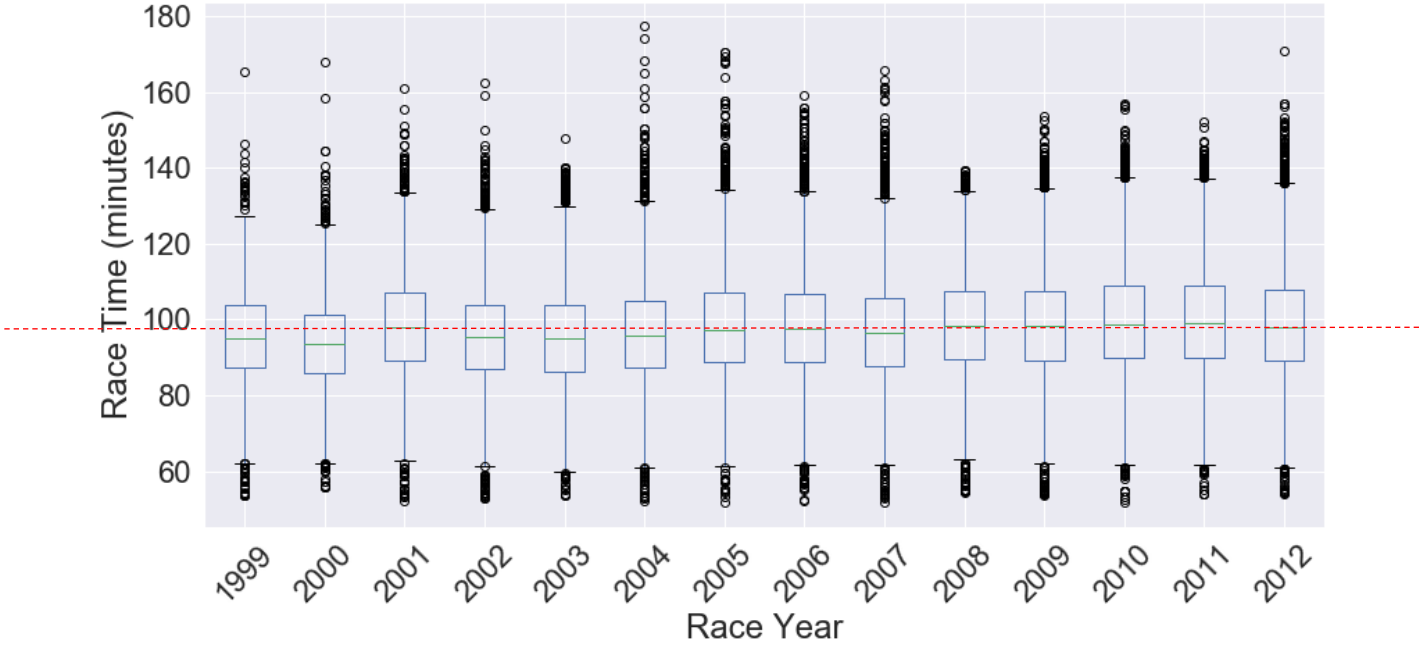

Figure 5: Boxplot of Year and Race Time

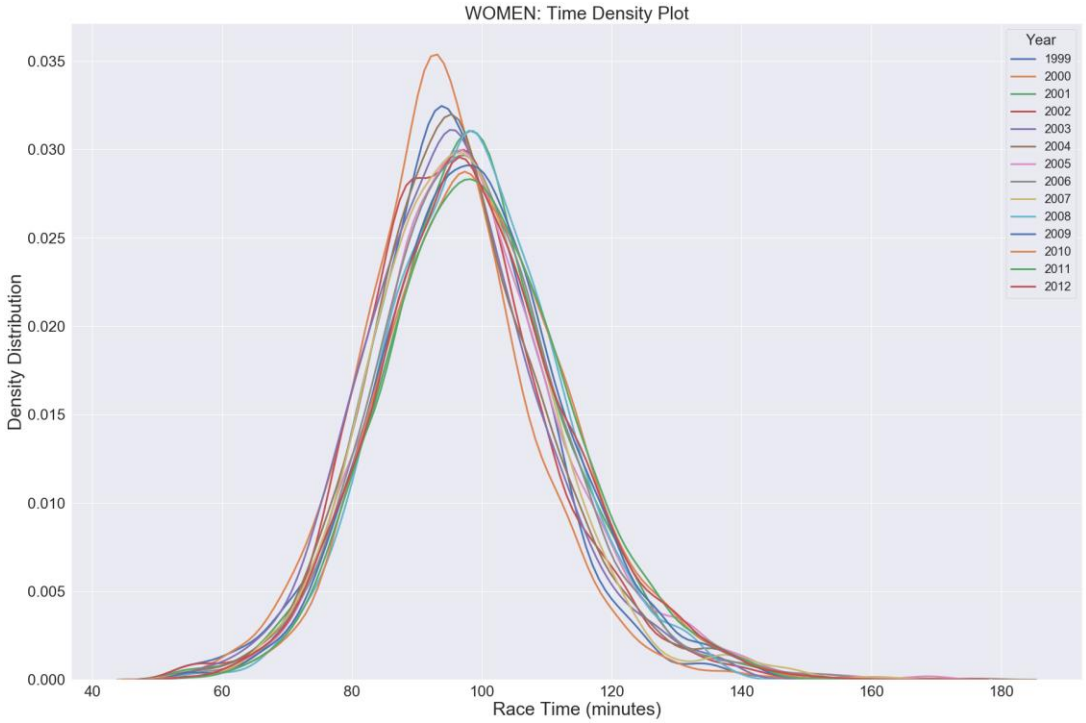

Figure 6: Density Plot of Time

MEN

In parallel to the analysis of age and time for female runners, the team looked at the same set charts for men. The first assessment was to look at how the men in each of the 10-year age bins changed over time. As seen in **Figure 7**, the growth of men participants is most prominent in the 19-29 and 29-39 age groups, but the changes are not nearly as dramatic as the female changes. The proportions of age categories over 40 as a reflection of the whole population remain relatively consistent.

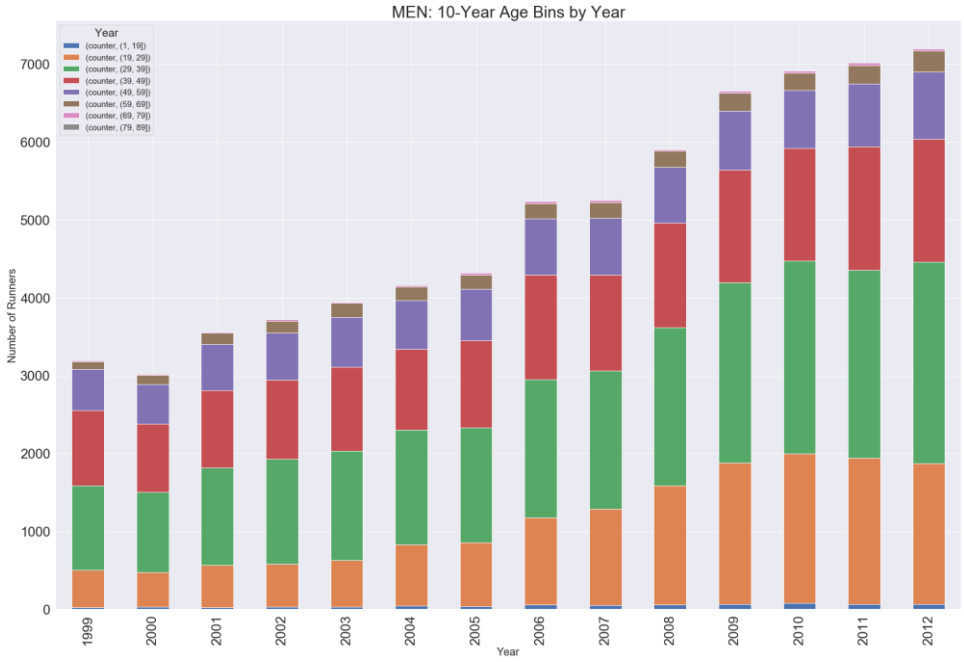

Figure 7: 10-Year Age Bins by Year

Figure 8 highlights the growth in runners. Comparing each bar to those around it reveals the changes from year to year. The biggest peaks are in runners age 20-40, especially in recent years.

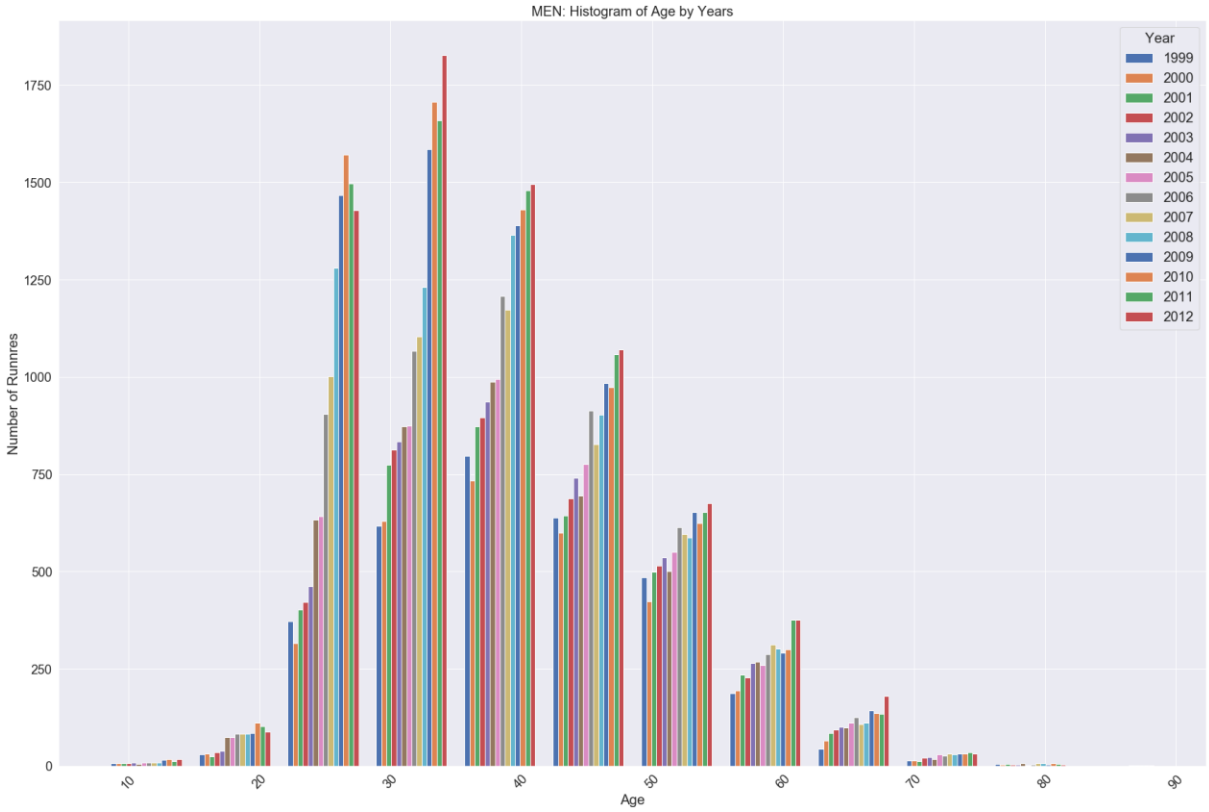

Figure 8: Runner Count by Age Groups Each Year

The two boxplots (**Figure 9** and **Figure 10**) compared men's age by year and race time by year. The boxplots provided better insights into the trends for age and time. In **Figure 9**, the red dotted line is equal to the median age in 1999 and 2000, which saw the oldest average runner age. The median age decreases slightly year over year. Overall, there are not dramatic shifts in the distribution of the plots but certainly a younger trend at all quartiles.

Figure 10 analyzes race time in minutes over the years to see if changes are occurring in the race time of male runners. The red dotted line reflects the median of the slowest year's race – 2006. It also does not appear that the men are getting faster over the years. While 2004 to 2007 had some slow outliers, the general trend overall shown in **Figure 11** reflects roughly consistent pace among male runners. As before, the distribution of race time is mostly consistent. However, 2006 appears to be an odd year; this year's curve is slightly flatter and shifted to the right, indicating a slower distribution of time. The weather looked OK for running, with no precipitation or extreme temperatures. However, there was some wind, which could have slowed race times. It does not seem to be significant to warrant discount that year, but important to note as a trend.

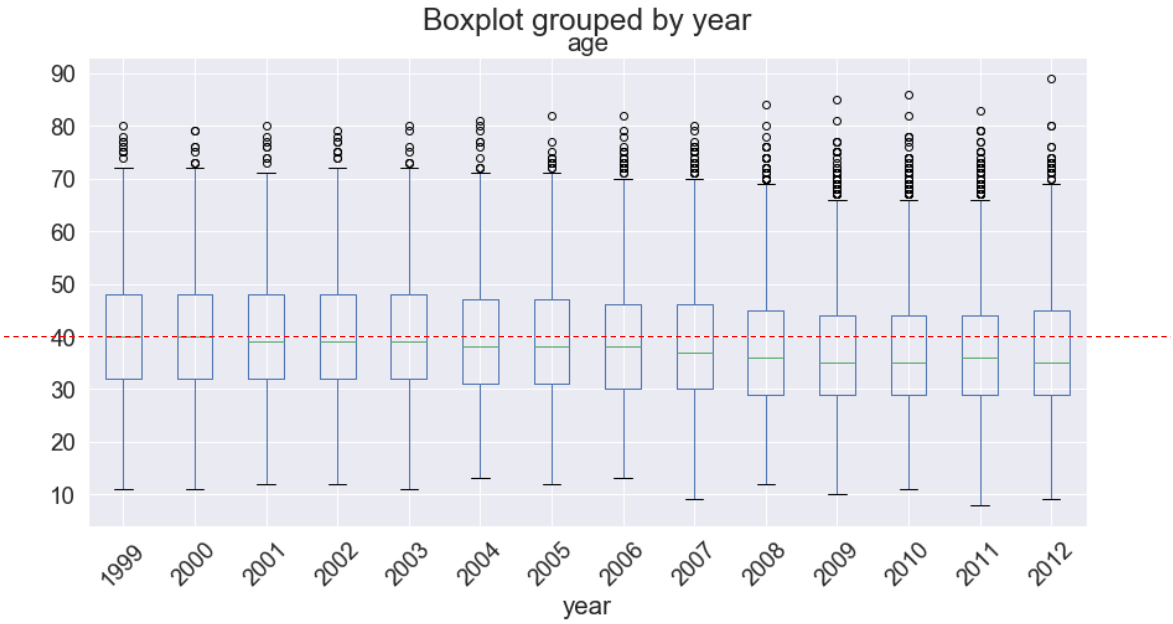

Figure 9: Boxplot of Year and Age

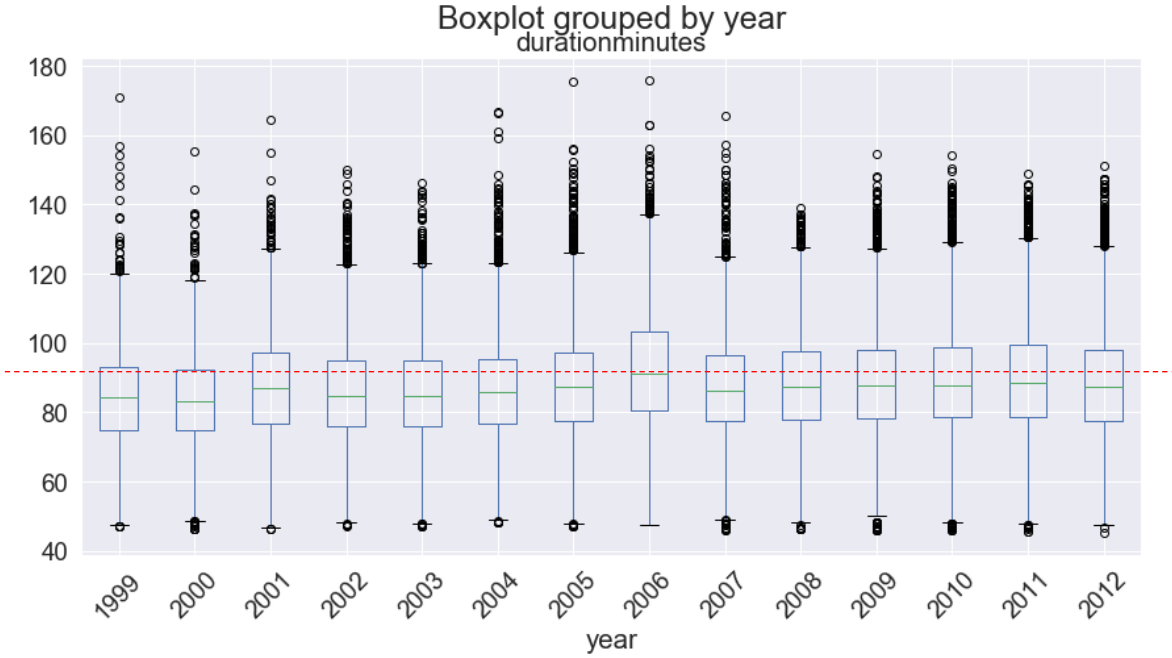

Figure 10: Boxplot of Year and Race Time

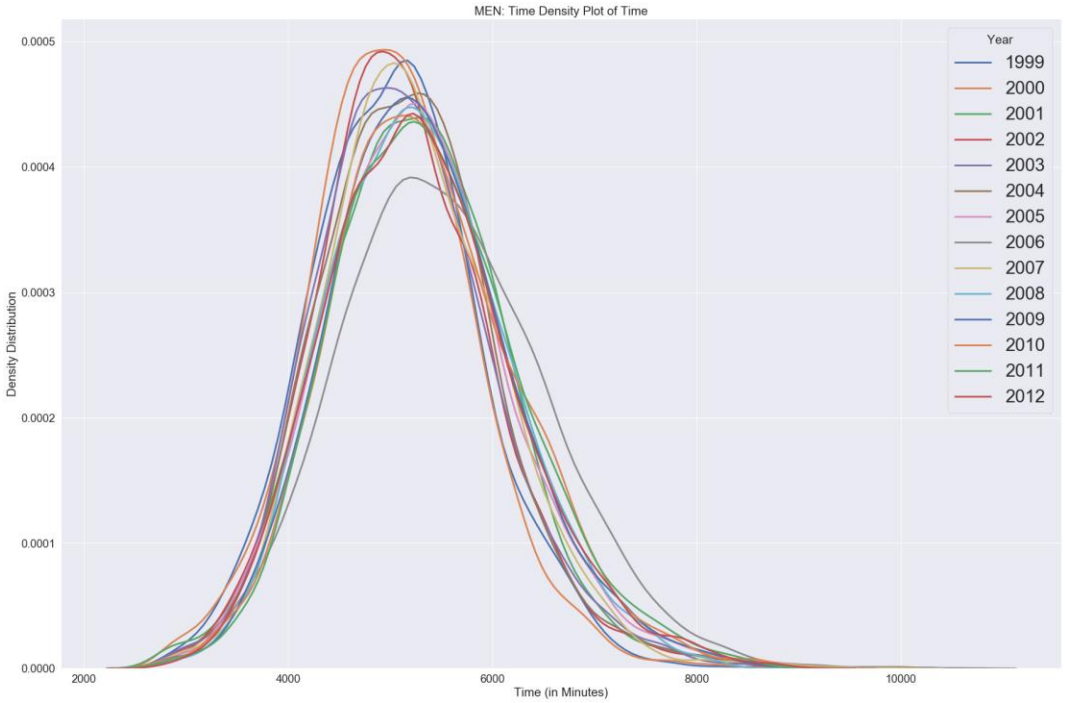

Figure 11: Density Plot of Time

COMPARISON OF WOMEN AND MEN

To compare men and women, we created violin plots help to demonstrate the differences. The age distribution of men compared to woman appears to be wider throughout the years, indicating a larger age spread. The distribution for women is skewed younger than men. More recent trends show that the female age distribution is less skewed in later years, as compared to earlier years. This skewness indicates that competition for the slots in the 30s age group for females could be high, given the increase in size of that age group.

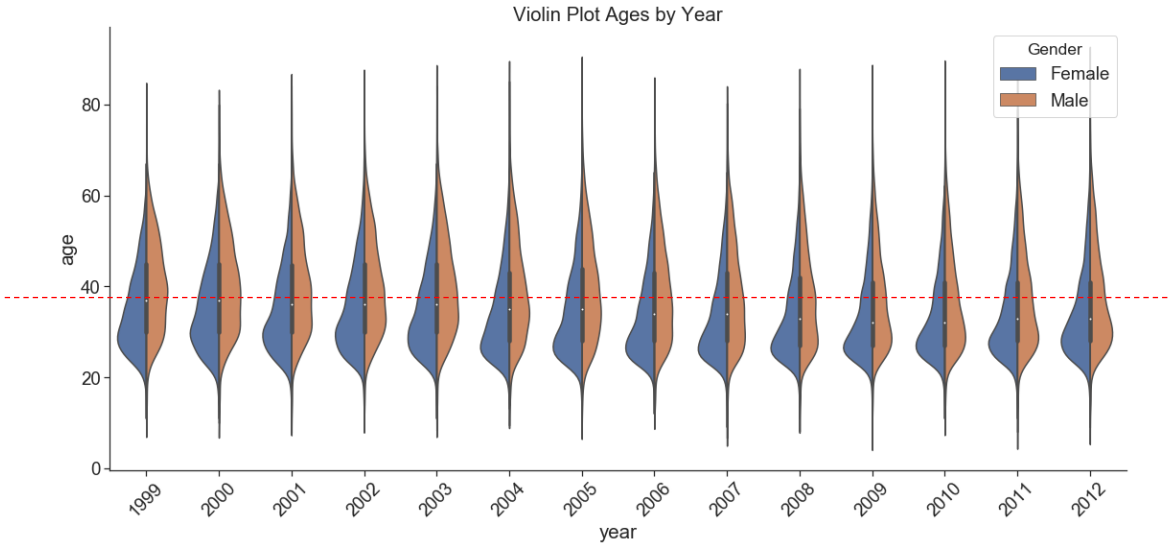

Figure 12: Violin Plot of Age

To compare men and women race times (in secs), we created violin plots to help demonstrate the differences. The race time distribution of men compared to woman appears to be wider throughout the years, indicating a larger race time spread. The distribution for the women displays less spread in race times. The race times appear to be fairly normal for both men and women. It is also evident from the plot that the men have a faster race completion time than women.

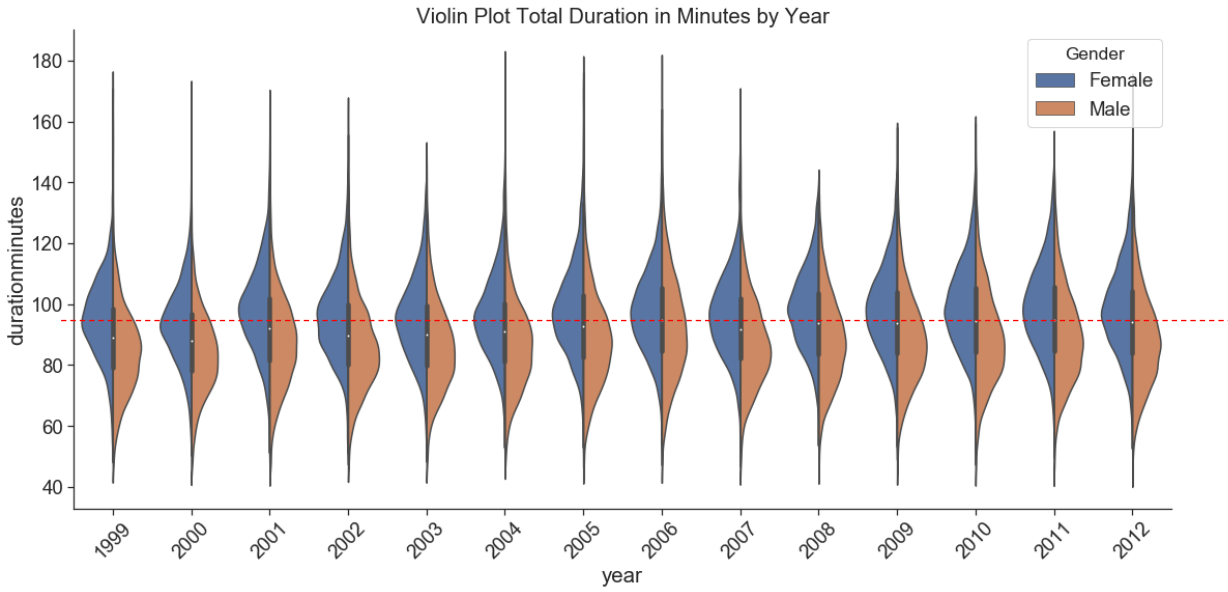

Figure 13: Violin Plot of Time

Statistical Analysis

We also conducted statistical analysis to understand if the age difference between women and men is significant. First, we ran a preliminary QQ Plot for both age and time for the entire data sets to visually check for normality. A QQ Plot illustrates the distribution of the data against a normal distribution; a straight line at 45 degrees indicates the data is normally distributed. **Figure 14** is a QQ Plot for age data and shows the data does not exhibit normality. This will help us select appropriate statistical tests.

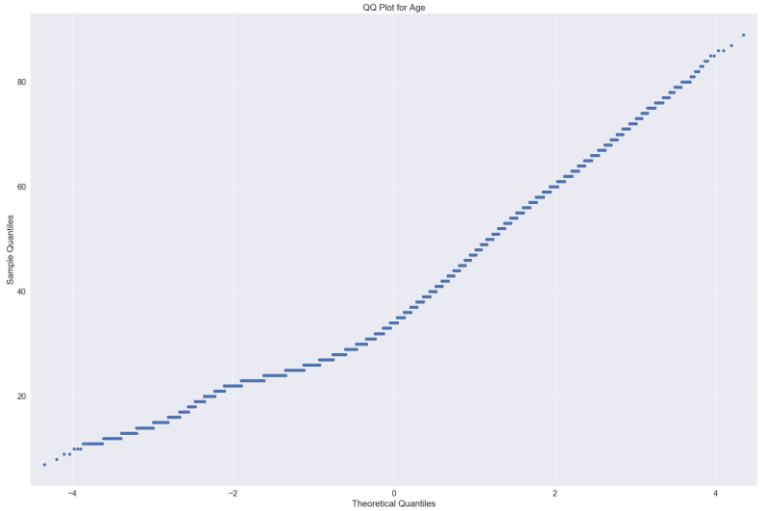

Figure 14: QQ Plot for Age

We further investigated normality with the Shapiro-Wilk test. The null hypothesis (H_o) is that ages are normally distributed for each gender. The test shows strong evidence that the distribution is not normal. Results shown in **Figure 15**. This backs the findings of both QQ plots.

	W	pval	normal
Female	0.932958	0.0	False
Male	0.970396	0.0	False

Figure 15: Shapiro-Wilk Test for Age

To complete our analysis, we tested the difference in age means between male and female runners using a t-test. The evidence suggests that the data is not normally disturbed and has unequal sample sizes between males and females. The Welch–Satterthwaite equation is used to approximate the adjusted degrees of freedom to help correct for this.

```
#########################################################################
t-test for difference in mean between Males and Females(Age)
#########################################################################
               T            dof       tail  p-val         CI95%   cohen-d  \
T-test  89.760711  137762.829028  two-sided    0.0  [4.66, 4.86]  0.473348

       BF10  power
T-test  inf    1.0
```

Figure 16: Welch-Satterthwaite Test for Age

Based on the results of the Welch–Satterthwaite we reject the null hypothesis (H_o). There is strong evidence (p-value = 0.0) that the means are different. Results shown in **Figure 16.**

An ANOVA test was run to test the difference in mean between years and age to see if a change in year

	sum_sq	df	F	PR(>F)
C(year)	1.805059e+05	13.0	131.525065	0.0
Residual	1.541255e+07	145994.0	NaN	NaN

is significantly related to age. The null hypothesis (H_o) is that there is no differences in mean age between years. The test yields a p-value = 0.0 indicating strong evidence to reject the null hypotheses **Figure 17**. We can extrapolate by looking back at **Figure 12** that the mean age is decreasing over the years. To help support this, a pair wise pairwise tukey was ran across the years. Comparing the early years to the later years we see significant P-values. This indicates the results from Tukey HSD suggests shows the null hypothesis should be rejected and that there is statistically significant differences. The result of this test can be found in **Appendix Figure B.**

A simple OLS regression model helps us further understand the relationship of age, time, and gender. The results of the regression test indicate that age, gender, and year are all significant (small p-values less than 0.000) for each dependent variable. The R^2 value of 0.947 indicates a high "goodness of fit" by the model. However, the results seem too perfect and over-fitting is a concern, so further investigation with other is highly recommended.

Model:	OLS	Adj. R-squared (uncentered):	0.947
Dependent Variable:	durationminutes	AIC:	1313567.1296
Date:	2020-06-01 21:36	BIC:	1313596.8039
No. Observations:	146008	Log-Likelihood:	-6.5678e+05
Df Model:	3	F-statistic:	8.721e+05
Df Residuals:	146005	Prob (F-statistic):	0.00
R-squared (uncentered):	0.947	Scale:	472.79

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
age	1.6004	0.0030	538.3436	0.0000	1.5946	1.6062
Is_Female	23.3931	0.1091	214.3619	0.0000	23.1792	23.6069
year_encoded	2.4775	0.0130	190.5169	0.0000	2.4520	2.5029

Omnibus:	113.700	Durbin-Watson:	1.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	124.501
Skew:	-0.034	Prob(JB):	0.000
Kurtosis:	3.126	Condition No.:	74

Figure 18: Results of Regression Model

Conclusion (Q10)

Through statistical analysis and visual analysis of supporting charts, our team assessed how the race has grown from 1999 to 2012, and how the age and time distributions have changed during that timeframe. Our analysis further shows that gender does have a significant impact on race time, as evident from the regression model (**Figure 18**) and the violin chart (**Figure 14**). As expected, the race times do slightly increase as age increases. It is taking runners, on average, approximately 2.5 minutes longer each year. This is more evident in the results of female runners than males. Given the larger proportion of female runners in recent years, their slower times may be overshadowing any gains in time by male runners. These changes are all very gradual, which could be driven by several factors such as how many elite runners participate, the performances by runners in the largest age groups, or the weather on any given race day.

Across women runners, the average age decreased between 1999 and 2012. This can be attributed both to the growth of participants in the 20-40 age group and fewer older participants in recent years. The change was gradual, except between 2005 and 2006 when there was a drop in older runners.

Men runners, on average, are also younger, likely due in part to the growth in participants age 20-40 as well. This change was also gradual. Interestingly, recent years have more older male runners that are outliers than women.

As a starting place for the Race Committee to think about decreasing the number of runners, we recommend next doing a cost comparison year over year for the 1999 to 2012 timeframe, as a cost

increase may dissuade participation. The committee could also use the information in this report to explore the creation of more strict qualifying times for the 10 mile and 5K races.

CODE DIRECTORY

Web Scraping Code in R (both genders)	**R Studio File**
Web Scraping Code in Python	**Men:** Python File **Women:** Python File
Analysis	**Master File:** Python Jupyter File **Additional Files:** Python Jupyter FIle 2, Python Jupyter File 3 *Note, there is overlap between the Jupyter files

APPENDIX

Appendix Figure A:
Data Dictionary

Variable	Description	Action Taken
Gender	Male or Female; gender of participants; scraped from individual Cherry Blossom website pages	Scraped, no cleaning necessary, data label
Year	1999 to 2012; scraped from individual Cherry Blossom website pages	Scraped, no cleaning necessary, data label
Place	Place in each year's race, separated by Gender and Year	Scraped, no cleaning necessary, did not use
Div_Total	Place of Each Participant in Age Division; for each year	Scraped, no cleaning necessary, did not use
Name	Participant Name	Scraped, no cleaning necessary, did not use individual names
Age	Age of Each Participant	Scraped, cleaned up missing or outlier ages
Hometown	Home of Each Participants, either City ST or Country	Scraped, no cleaning, did not use
Time	One of: Time, Gun Time, Net Time; or Comb Time; Time – Overall Participant Time; Gun Time – Time of Participant from Gun to Individual Finish; Net Time – Time of Participant from Start Line to Finish Line.	Scraped, needed to assess which times provided; used in order – Net Time, Gun Time, or Comb Time; type depended on race year
Pace	Average Mile Per Hour for Each Participant	Scraped, discovered odd values; did not use
Num ID	Cherry Blossom Participant ID, not available for all years	Scraped, no cleaning necessary, did not use
Net Time	Net Time for Each Participant, where available	Scraped, where available

Comb Time	Combined Time of Each Participant, usually reflects Gun Time	Scraped, where available
Time Length	Created Variable to help calculate time	Created to assist in managing time format for calculations
F Combine	Created Variable	Created to assist in managing time format for calculations
Final Time	Created Variable	Created to assist in managing time format for calculations
Hour	Created Variable to separate time: hours	Created to assist in managing time format for calculations
Minutes	Created Variable to separate time: minutes	Created to assist in managing time format for calculations
Seconds	Created Variable to separate time: seconds	Created to assist in managing time format for calculations
Dur Secs	Calculated Variable; Total Race Time in Seconds	Created to assist in managing time format for calculations
Duration Min	Calculated Variable; Total Race Time in Minutes	Created to assist in managing time format for calculations
Calc Pace	Calculated Pace from Time	Created to assist in managing time format for calculations
Counter	Created Variable; tally of 1 for each participant	Created to assist in managing participant count

Appendix Figure B:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1999	2000	0.3527	0.8816	-0.3132	1.0185	FALSE
1999	2001	-0.2171	0.9	-0.8464	0.4122	FALSE
1999	2002	-0.1604	0.9	-0.779	0.4581	FALSE
1999	2003	-0.175	0.9	-0.7855	0.4356	FALSE
1999	2004	-1.3177	0.001	-1.919	-0.7163	TRUE
1999	2005	-1.1656	0.001	-1.7585	-0.5726	TRUE
1999	2006	-1.7929	0.001	-2.3634	-1.2224	TRUE
1999	2007	-2.1205	0.001	-2.6894	-1.5516	TRUE
1999	2008	-2.6207	0.001	-3.1781	-2.0633	TRUE
1999	2009	-3.0431	0.001	-3.5848	-2.5013	TRUE
1999	2010	-3.1175	0.001	-3.6556	-2.5794	TRUE
1999	2011	-2.6317	0.001	-3.1686	-2.0949	TRUE
1999	2012	-2.5029	0.001	-3.0362	-1.9696	TRUE

2000	2001	-0.5698	0.1448	-1.2109	0.0713	FALSE
2000	2002	-0.5131	0.2621	-1.1436	0.1174	FALSE
2000	2003	-0.5276	0.2039	-1.1503	0.0951	FALSE
2000	2004	-1.6703	0.001	-2.284	-1.0567	TRUE
2000	2005	-1.5182	0.001	-2.1237	-0.9128	TRUE
2000	2006	-2.1456	0.001	-2.729	-1.5621	TRUE
2000	2007	-2.4732	0.001	-3.0551	-1.8913	TRUE
2000	2008	-2.9733	0.001	-3.544	-2.4027	TRUE
2000	2009	-3.3957	0.001	-3.9511	-2.8403	TRUE
2000	2010	-3.4702	0.001	-4.022	-2.9184	TRUE
2000	2011	-2.9844	0.001	-3.535	-2.4338	TRUE
2000	2012	-2.8556	0.001	-3.4027	-2.3084	TRUE
2001	2002	0.0567	0.9	-0.5351	0.6485	FALSE
2001	2003	0.0422	0.9	-0.5413	0.6257	FALSE
2001	2004	-1.1005	0.001	-1.6743	-0.5267	TRUE
2001	2005	-0.9485	0.001	-1.5135	-0.3834	TRUE
2001	2006	-1.5758	0.001	-2.1172	-1.0344	TRUE
2001	2007	-1.9034	0.001	-2.4431	-1.3637	TRUE
2001	2008	-2.4036	0.001	-2.9312	-1.876	TRUE
2001	2009	-2.8259	0.001	-3.337	-2.3149	TRUE
2001	2010	-2.9004	0.001	-3.4076	-2.3933	TRUE
2001	2011	-2.4146	0.001	-2.9204	-1.9088	TRUE
2001	2012	-2.2858	0.001	-2.7878	-1.7838	TRUE
2002	2003	-0.0145	0.9	-0.5864	0.5573	FALSE
2002	2004	-1.1572	0.001	-1.7192	-0.5952	TRUE
2002	2005	-1.0051	0.001	-1.5581	-0.4522	TRUE
2002	2006	-1.6325	0.001	-2.1613	-1.1036	TRUE
2002	2007	-1.9601	0.001	-2.4872	-1.433	TRUE
2002	2008	-2.4602	0.001	-2.9749	-1.9456	TRUE
2002	2009	-2.8826	0.001	-3.3803	-2.3849	TRUE
2002	2010	-2.9571	0.001	-3.4508	-2.4634	TRUE
2002	2011	-2.4713	0.001	-2.9636	-1.9789	TRUE
2002	2012	-2.3425	0.001	-2.8309	-1.854	TRUE
2003	2004	-1.1427	0.001	-1.6959	-0.5895	TRUE
2003	2005	-0.9906	0.001	-1.5347	-0.4465	TRUE
2003	2006	-1.6179	0.001	-2.1375	-1.0984	TRUE
2003	2007	-1.9456	0.001	-2.4633	-1.4278	TRUE
2003	2008	-2.4457	0.001	-2.9508	-1.9406	TRUE
2003	2009	-2.8681	0.001	-3.3559	-2.3803	TRUE
2003	2010	-2.9426	0.001	-3.4263	-2.4589	TRUE
2003	2011	-2.4568	0.001	-2.9391	-1.9744	TRUE

2003	2012	-2.3279	0.001	-2.8063	-1.8496	TRUE
2004	2005	0.1521	0.9	-0.3816	0.6858	FALSE
2004	2006	-0.4753	0.096	-0.9839	0.0334	FALSE
2004	2007	-0.8029	0.001	-1.3097	-0.2961	TRUE
2004	2008	-1.303	0.001	-1.7969	-0.8091	TRUE
2004	2009	-1.7254	0.001	-2.2016	-1.2492	TRUE
2004	2010	-1.7999	0.001	-2.2719	-1.3279	TRUE
2004	2011	-1.3141	0.001	-1.7847	-0.8435	TRUE
2004	2012	-1.1852	0.001	-1.6517	-0.7188	TRUE
2005	2006	-0.6273	0.002	-1.126	-0.1287	TRUE
2005	2007	-0.9549	0.001	-1.4518	-0.4581	TRUE
2005	2008	-1.4551	0.001	-1.9388	-0.9715	TRUE
2005	2009	-1.8775	0.001	-2.343	-1.412	TRUE
2005	2010	-1.952	0.001	-2.4132	-1.4907	TRUE
2005	2011	-1.4662	0.001	-1.926	-1.0064	TRUE
2005	2012	-1.3373	0.001	-1.7929	-0.8817	TRUE
2006	2007	-0.3276	0.5222	-0.7974	0.1422	FALSE
2006	2008	-0.8278	0.001	-1.2836	-0.3719	TRUE
2006	2009	-1.2501	0.001	-1.6867	-0.8136	TRUE
2006	2010	-1.3246	0.001	-1.7566	-0.8926	TRUE
2006	2011	-0.8388	0.001	-1.2693	-0.4084	TRUE
2006	2012	-0.71	0.001	-1.136	-0.284	TRUE
2007	2008	-0.5002	0.0157	-0.954	-0.0463	TRUE
2007	2009	-0.9225	0.001	-1.357	-0.4881	TRUE
2007	2010	-0.997	0.001	-1.4269	-0.5671	TRUE
2007	2011	-0.5112	0.0049	-0.9395	-0.0829	TRUE
2007	2012	-0.3824	0.1283	-0.8062	0.0414	FALSE
2008	2009	-0.4224	0.0464	-0.8417	-0.003	TRUE
2008	2010	-0.4968	0.0046	-0.9114	-0.0823	TRUE
2008	2011	-0.011	0.9	-0.424	0.4019	FALSE
2008	2012	0.1178	0.9	-0.2905	0.5261	FALSE
2009	2010	-0.0745	0.9	-0.4678	0.3188	FALSE
2009	2011	0.4113	0.0288	0.0197	0.8029	TRUE
2009	2012	0.5402	0.001	0.1535	0.9268	TRUE
2010	2011	0.4858	0.002	0.0993	0.8723	TRUE
2010	2012	0.6146	0.001	0.2332	0.9961	TRUE
2011	2012	0.1288	0.9	-0.2509	0.5086	FALSE