Age-Dist Case Study II

James Vasquez, Javier Saldana, Sara Zaheri

Table of Contents

Table of Figures

Age-Dist Case StudyII

James Vasquez, Javier Saldana, Sara Zaheri

1. Introduction

The Cherry Blossom Ten Mile Run & 5K Run-Walk is a race that takes place in Washington DC during the spring. The race has been a staple for the city and is in its 21st year of running. The race is sponsored by multiple credit unions in the US. The organization is referenced as America's Credit Union, this sponsored event is used to support the Children's Hospitals that belong to the Children's Miracle Network Hospitals. The organization is used to treat millions of children in both the US and Canada. (1)

The race takes place 0.5 mile south of the White House and runs through a scenic route along the Potomac River and finishes at the same point. The race has seen a constant increase of runners from the inaugural year up to 2019. The 2020 race was done with virtual submissions due to the COVID-19 Pandemic. From 1999 the race had 2,352 runners where in 2012 the runners had increased to 9,729 runners. In recent years, the runners have come worldwide from the US, Kenya, Poland, and Ethiopia.

The analysis done on the data web scraps the historical race results from 1999 to 2012. In recent years the organization has moved their results from web data tables into a historical search enabled database. Though the database is user friendly it does not gather all data from the previous race results.

In recent years of the race, family and friends have been able to track their runners through GPS and receive status updates. Not only does GPS tracking allow friends and family to track their runners but they can position themselves to cheer the runners on at multiple locations. It also helps with meet up times when runners finish and find their friends. The Cherry Blossom Race has grown in popularity but also has grown in the methods race results and technology for tracking runners.

The analysis takes on the challenge comparing age distribution between the year's web scrapped for the data. Statistical testing is also done to quantify any significance difference between years.

2. Methods

The objective for this study was to determine if there was a statistically significant difference between the racers each given year. While there are multiple ways to address this question, we reviewed the data set for assumptions and addressed the question of interest using a Tukey-Kramer test.

The first step was to ensure all of the required assumptions for a Tukey-Kramer test had been met. To utilize the Tukey-Kramer test, the dataset must have a normal distribution, homogeneity of variance, and be independent. First and foremost, we can see that based on the given age distribution, the age distribution is right skewed. There is a higher number of runners that are younger, and the number slowly decreases as the age increases. Some ways to address this violation would be to consider a log transformation.

To determine if the assumption of equal variance is met, we utilized the Levene Test for Equality of Variances. The Levene test was the preferred test of choice because of the normality violation addressed earlier. Unlike the Bartlett test, which also measures variance equality, the Levene test is robust to moderate violations of normal distribution. The Levene test is defined as follows:

$$H_0: \qquad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$

$$H_a: \qquad \sigma_i^2 \neq \sigma_j^2 \quad \text{For at least one pair } (i,j)$$

The null hypothesis states that the variance for all groups is equal. The alternative hypothesis states that the variance from at least one group is different. Where N is the sample size and k is the number of subgroups divided within, the test statistic to measure the variance is calculated as follows:

$$W = \frac{(N-k)}{(k-1)} \times \frac{\sum_{i=1}^{k} N_i \left(\underline{Z}_{i.} - \underline{Z}_{..}\right)^2}{\Sigma_{i=1}^{k} \sum_{j=1}^{N_i} \left(\underline{Z}_{ij} - \underline{Z}_{i.}\right)^2}$$

We did not use the Brown and Forsythe enhancements of substituting the mean with the median or a trimmed mean because to increase power since the test is robust as is considering the data sets are largely similar. As a result, where $\underline{Z}_i$ is the group means of Z_{ij} and $\underline{Z}_{..}$ is the overall mean of Z_{ij},we calculate Z_{ij} as the following:

$$Z_{ij} = \left|Y_{ij} - \underline{Y}_{i.}\right| \quad \text{Where } Y \text{ is the mean of the } i\text{-th subgroup}$$

The final assumption to be addressed is independence. For entries to be independent, the entries from one group must not be related or dependent on the prior group. This causes concern in this instance since the dataset does not provide any insight into the dependency of the entries. None of the data available provides a unique identifier which could be used across different years. As a result, there is no true manner to test this assumption based on the limited available data. However, considering marathons are likely to have repeat runners, it is highly likely there may be some dependency issues that may arise in this data set. Since dependency is not something we can control after data collection, we will proceed with the assumption that the data is independent.

Once all assumptions have been met, we used the Tukey-Kramer test to determine if there was in fact a statistical difference between the years. The Tukey-Kramer method is appropriate for this study because of the unequal populations in each group. Using this methodology, we were able to determine if there is a statistical difference between the means of the groups. The Tukey-Kramer method is applied to all group and therefore, produces results for all combinations available with group comparisons.

2.1. Data Source

The data used for analysis is web scrapped from their historical repository of race results.

http://www.cherryblossom.org/aboutus/results_list.php\

The team choice of language to scrap and analyze the data is Python. Within Python the Beautiful Soup library is used to access web tags and the data table which holds the data needed. A nested approach was ideally used to locate the proper tag for the data table, due to improperly built sites several tags were not opened and closed correctly. However, this is expected with data from 1999 where websites were still being developed in many ways. Rather than relying on the tag, a function was built to access the content and search for the word 'Place' in almost every table this word is found and is the start of the data table needed.

Throughout the years different features were captured by races, for instance the feature time is captured by 'time', 'gun time', 'net time', 'net tim', etc…. It was agreed that the data frame to be used for analysis would concentrate only on fields that were in common between all the years. These columns are:

- Place
- Name
- Age
- Hometown
- Time
- Year

All years were able to parse easily except for 2001 & 2006. Race results from 2001 did not have a header row to allow the function to find headers it was looking for. Race results from 2006 combine the columns of hometown and time together which had to be split into their proper columns. It was found to be more time efficient to handle these years slightly different rather than try to accommodate the functions for this one offs.

2.2. Exploratory Data Analysis (EDA)

This section serves as an exploratory data analysis on the age distribution of recares from year to year. We present various plots to investigate the assumption of having the same distribution. Before diving into a bunch of plots we illustrate the distribution of one year. I fairly choose 2005.

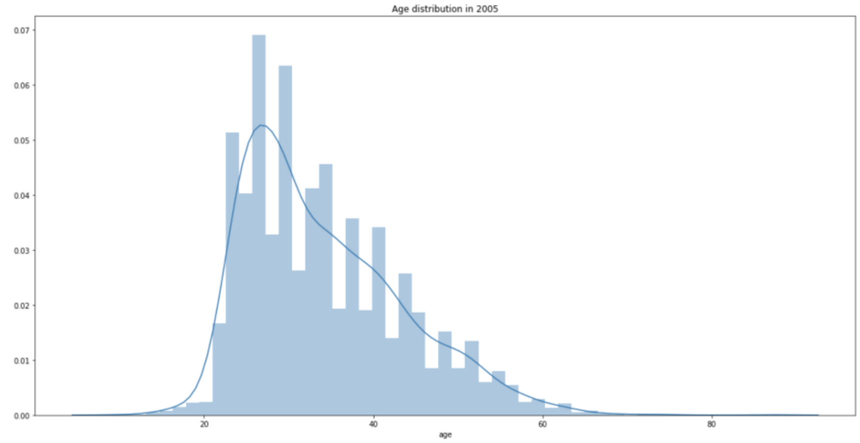


Figure 1. Age distribution in 2005

One might infer from the plot above that there is a sharp increase (say exponentially) after the age of 18 which consistently holds until the age of 26-27. Then the diagram falls linearly until 60. Plotting the same diagram for different ages confirms this observation. I present the density curve for all years from 1999 to 2012 as follows.

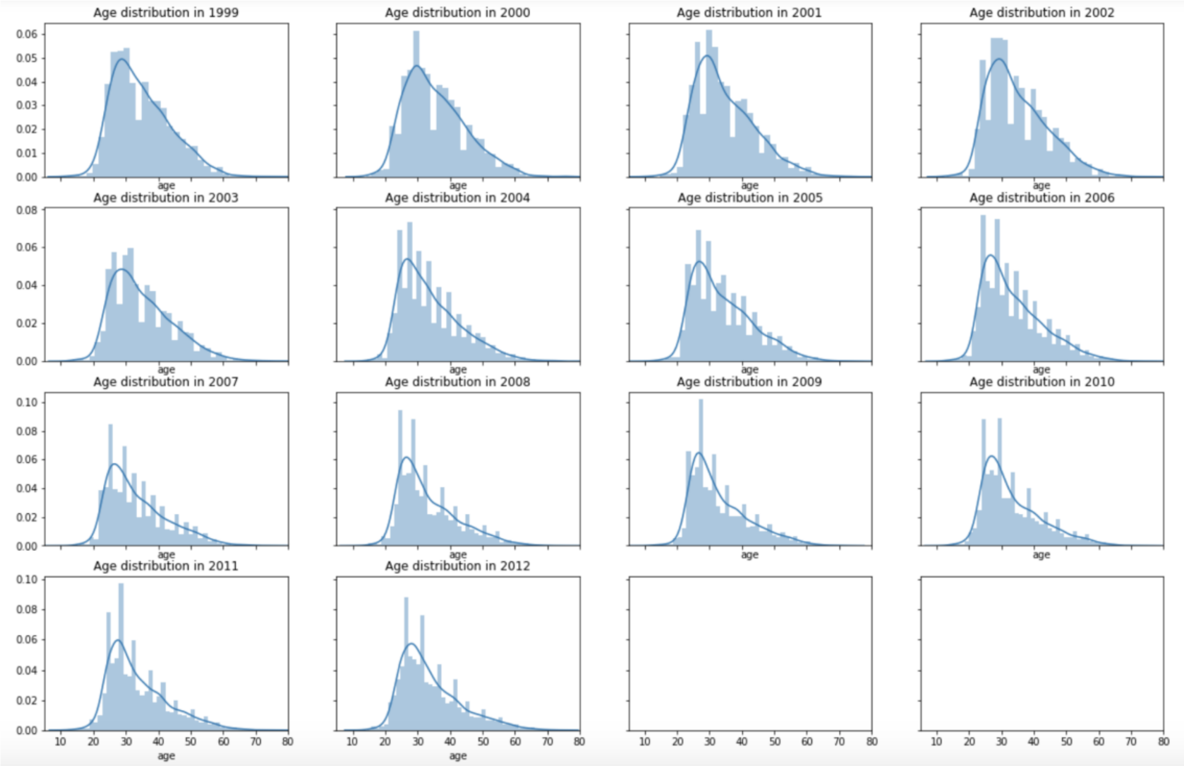


Figure 2. Age distribution for all years from 1999 to 2012

From what we have plotted so far, it is not certain the equality of distribution over the years. To call this equality into question we plot all density curves in figure, plotting with a color map called magma to distinguish between years.

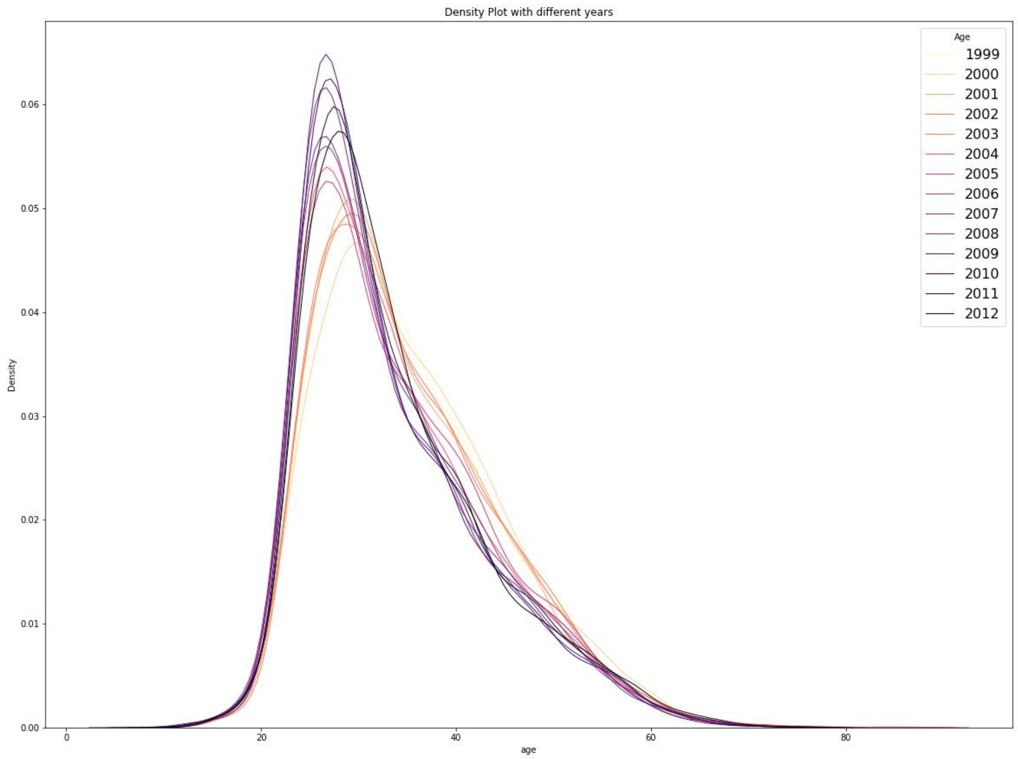


Figure 3. Age distribution over 14 years using color map

The former figure along with the following which is cumulative distribution function support the argument that in the early years the population was older compared to the last years. It is visible from these figures that until 30 year-old the purple curves (i.e. last years) are above the yellow and red curves (i.e. first years).

The other method to compare two different distributions is Q-Q plot which draws a comparison between two sets of quintiles from two datasets. 14 years of dataset arises 91 pairs to compare which is quite a lot. First we present an example in the following page to clarify our analysis. The examples demonstrate the Q-Q plot between 2000 and 2010. The closer the scatter plotted quantiles to the 45 degree line are the more similar the distribution of two datasets is. For first years the points are above the 45 line, then they become below. This observation is noticeable for almost all pairs of years as we moving forward from 1999 to 2012.

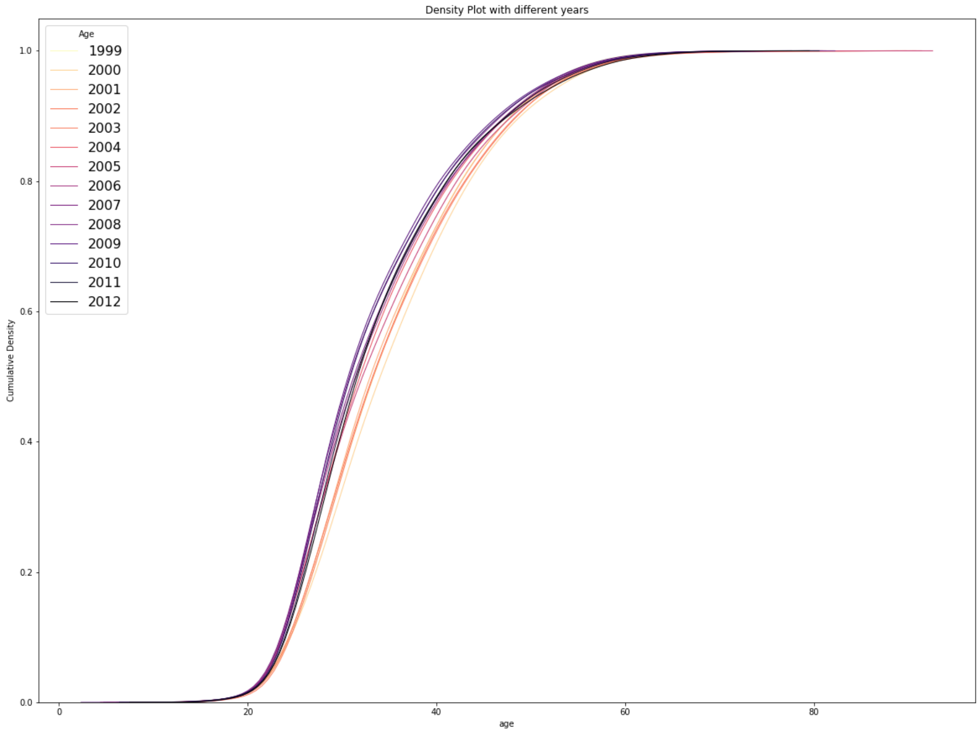


Figure 4. Commulative distribution curves using color map

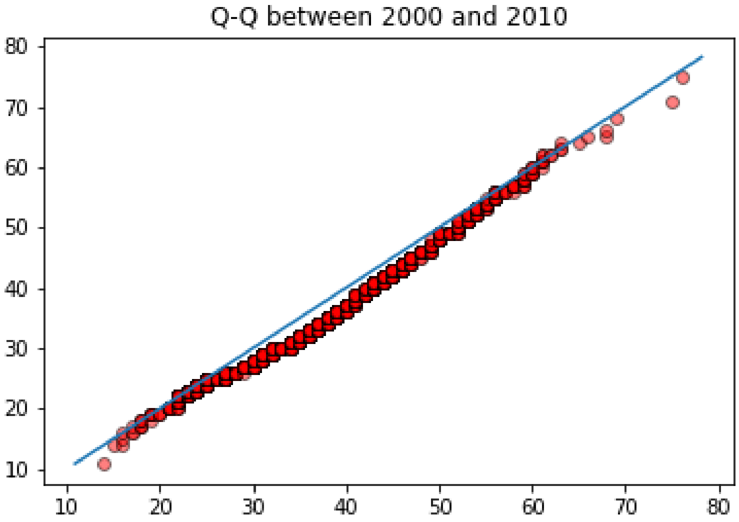


Figure 5. Q-Q plot between 2000 and 2010

We elaborate on Q-Q plot by having each year against the next and previous year, as well as 1999 and 2012.

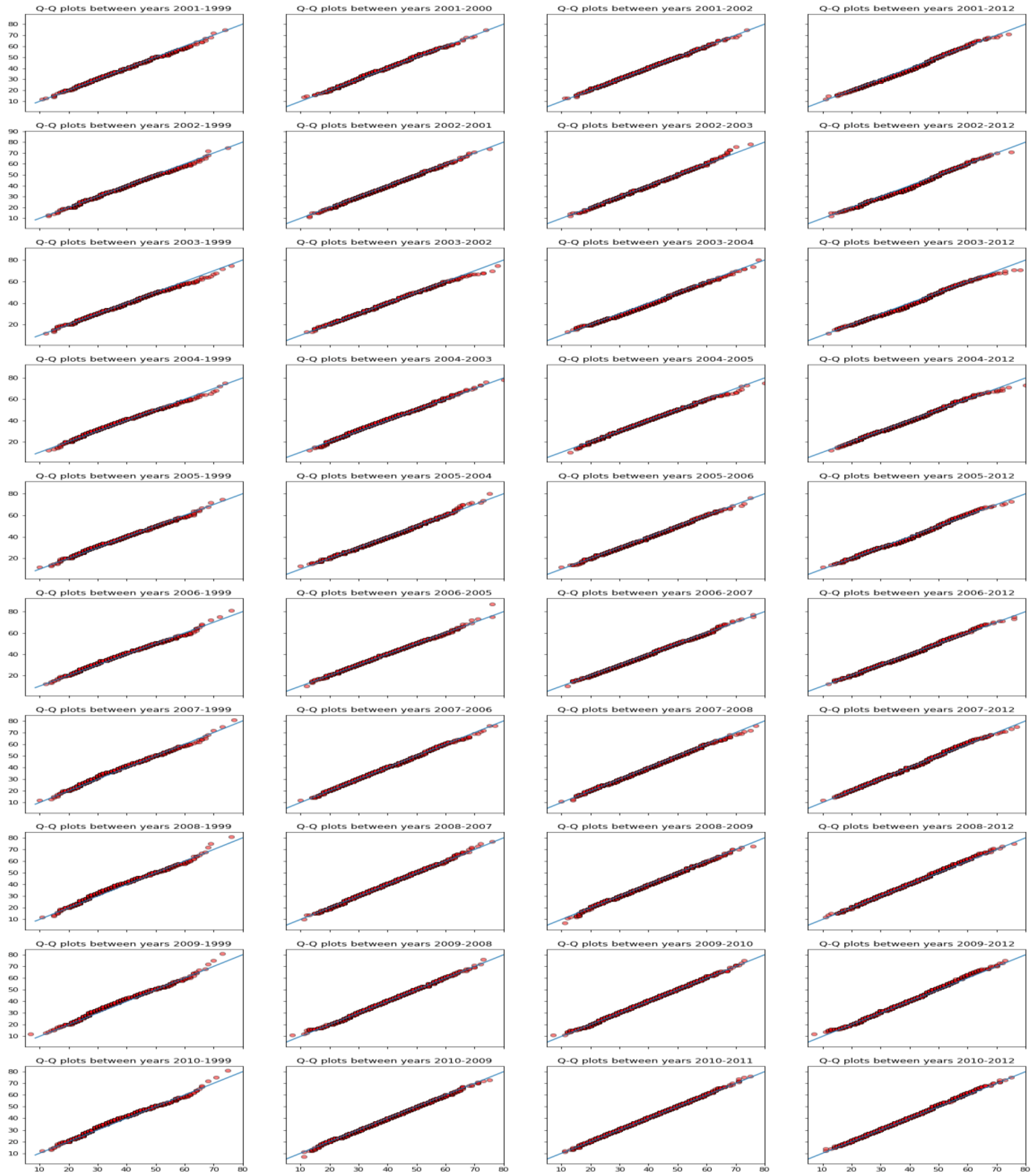


Figure 6. Q-Q plot for years from 2001 to 2010. Each year has been plotted against the previous, next, first, and last year.

Our last but not the least plot tool to compare two distributions is box plot. Box plots illustrate the quartiles, max, min, and median in the form of a box. In this way, we can see a gradual decrease in age for those important quantiles over the years which concludes the distribution slightly changed along those years.

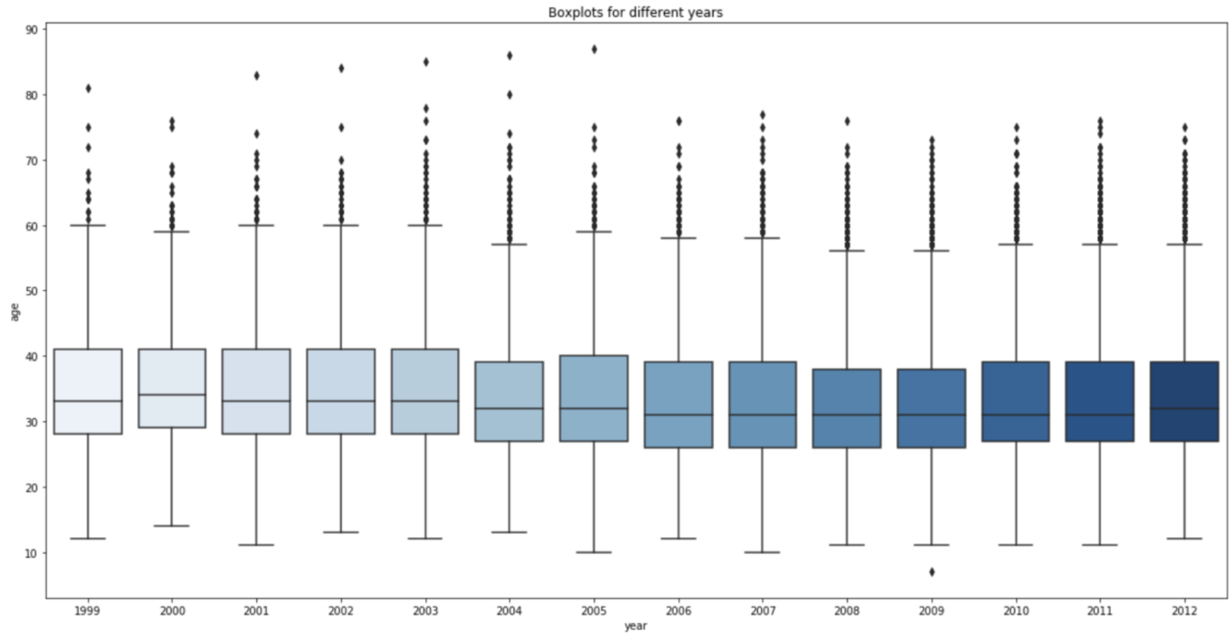
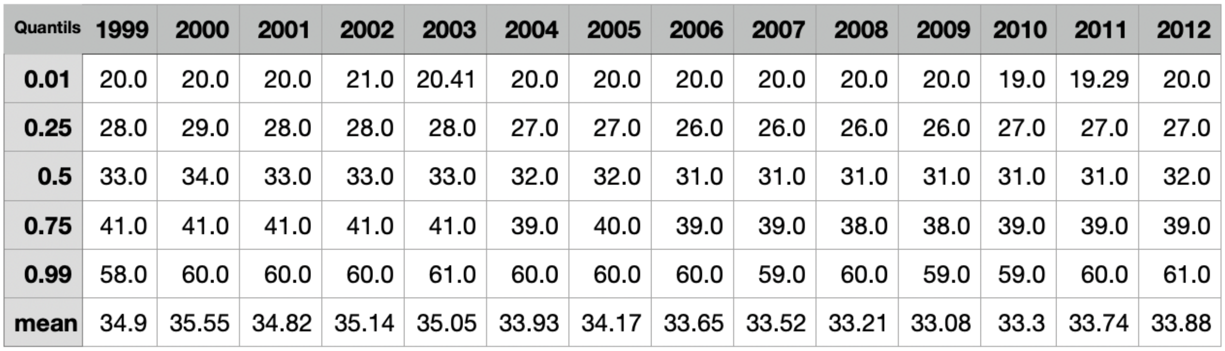


Figure 7. Box plot of quartiles over the years

Finally, I place a table of quartiles and 0.99 and 0.01 (which are almost max and min after removing the outliers) for different years in the following. One can see that the first and third quartiles, as well as the median are increasing and has the difference of 1 or 2 years comparing first and last years.



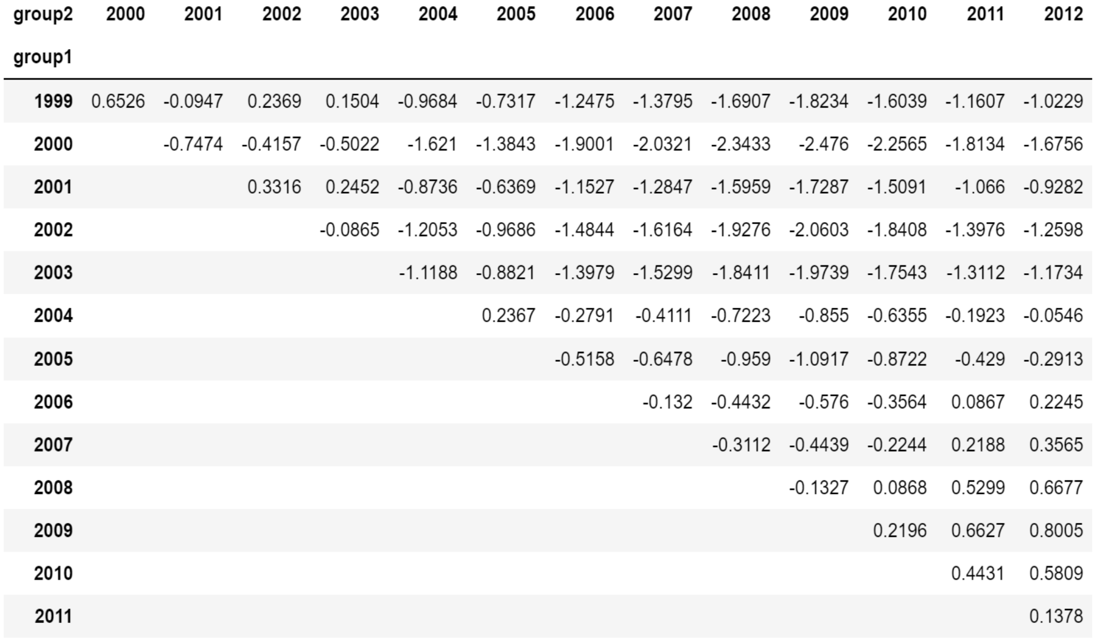
Quantils	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
0.01	20.0	20.0	20.0	21.0	20.41	20.0	20.0	20.0	20.0	20.0	20.0	19.0	19.29	20.0
0.25	28.0	29.0	28.0	28.0	28.0	27.0	27.0	26.0	26.0	26.0	26.0	27.0	27.0	27.0
0.5	33.0	34.0	33.0	33.0	33.0	32.0	32.0	31.0	31.0	31.0	31.0	31.0	31.0	32.0
0.75	41.0	41.0	41.0	41.0	41.0	39.0	40.0	39.0	39.0	38.0	38.0	39.0	39.0	39.0
0.99	58.0	60.0	60.0	60.0	61.0	60.0	60.0	60.0	59.0	60.0	59.0	59.0	60.0	61.0
mean	34.9	35.55	34.82	35.14	35.05	33.93	34.17	33.65	33.52	33.21	33.08	33.3	33.74	33.88

Figure 8. Table of quartiles

2.3. Analysis

Based on the results from Levene’s test, we would fail to reject the null hypothesis (p-value = 0.001) that the variance from the groups are equal with 95% confidence. This result, along with the other two conditions put us in prime position to conduct the Tukey-Kramer test with confidence.

The Tukey-Kramer test produces p-values and confidence intervals for each comparison. Instead of illustrating a table of p-values, we produced this table that demonstrates which comparisons have a statistically significant difference than others. The Tukey-Kramer test provides the mean difference between the years and also whether the comparison rejects or fails to reject the null hypothesis. By comparing the years to all of them, we were able to gain insight into the change of the racers' age. While the number of racers increased substantially, the mean age changed slightly throughout the years.



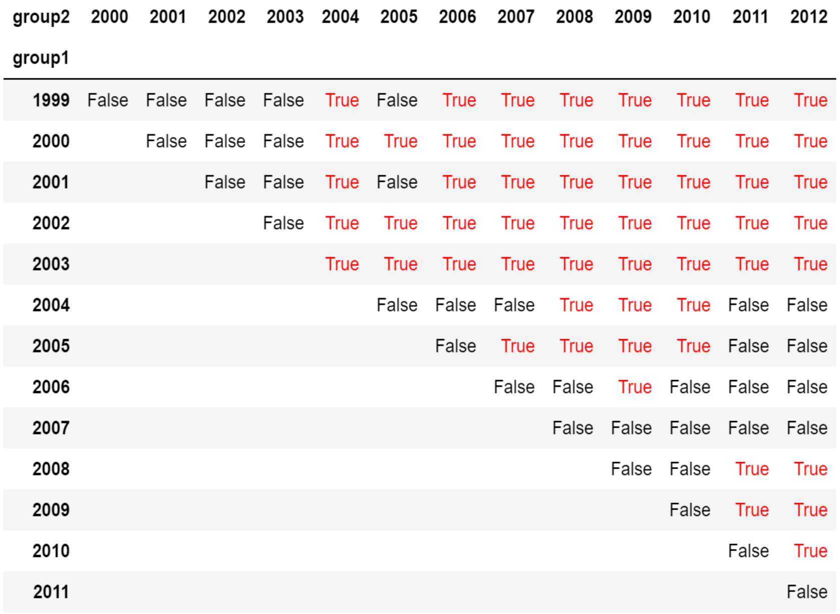
group2 group1	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
1999	0.6526	-0.0947	0.2369	0.1504	-0.9684	-0.7317	-1.2475	-1.3795	-1.6907	-1.8234	-1.6039	-1.1607	-1.0229
2000		-0.7474	-0.4157	-0.5022	-1.621	-1.3843	-1.9001	-2.0321	-2.3433	-2.476	-2.2565	-1.8134	-1.6756
2001			0.3316	0.2452	-0.8736	-0.6369	-1.1527	-1.2847	-1.5959	-1.7287	-1.5091	-1.066	-0.9282
2002				-0.0865	-1.2053	-0.9686	-1.4844	-1.6164	-1.9276	-2.0603	-1.8408	-1.3976	-1.2598
2003					-1.1188	-0.8821	-1.3979	-1.5299	-1.8411	-1.9739	-1.7543	-1.3112	-1.1734
2004						0.2367	-0.2791	-0.4111	-0.7223	-0.855	-0.6355	-0.1923	-0.0546
2005							-0.5158	-0.6478	-0.959	-1.0917	-0.8722	-0.429	-0.2913
2006								-0.132	-0.4432	-0.576	-0.3564	0.0867	0.2245
2007									-0.3112	-0.4439	-0.2244	0.2188	0.3565
2008										-0.1327	0.0868	0.5299	0.6677
2009											0.2196	0.6627	0.8005
2010												0.4431	0.5809
2011													0.1378

Figure 9. Turkey-Kramer mean difference. Shows the difference in mean age across all years.

3. Results

The results from the Tukey-Kramer test indicate the mean age did change over time as the years progressed. From a consecutive standpoint, the year 2004 saw the greatest decrease in mean age as the mean difference between 2003 and 2004 was -1.1188. Prior to 2004, the mean age for all the years was equal. After 2004, we begin to see a slight shift in the mean age that only lasts up to 2 years. We can see which comparisons reject the null hypothesis by looking at the red "True" values in the table. Based on this analysis, it is clear something considerably different occurred in 2004 that warrants additional research to identify the root cause of the drastic shift.

The implications of these findings have the capability to reduce costs by optimizing advertising dollars. Even though the average age is slowly decreasing over the years, the cost of acquisition for a 25-year old runner is much lower in 2012 than it was in 1999. The primary reason for that would be the revolution of advertising on the internet and the rise of social media. The race has been able to increase its turnout substantially over the years, which started with just over 2,300 racers in 1999 and ended with over 9,700 in 2012. While the current marketing expenses are readily available, the organization is able to reduce its cost of acquisition and retention with the implementation of a reliable customer resource management system. Maintaining a database of prior runners and their information (such as age) will give the organization insight into when a runner might be less likely to enter the race. It can shift those efforts to novel marketing ideas such as trendy social media campaigns or sporting event sponsorships in order gain new racers.



group2	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
group1													
1999	False	False	False	False	True	False	True	True	True	True	True	True	True
2000		False	False	False	True	True	True	True	True	True	True	True	True
2001			False	False	True	False	True	True	True	True	True	True	True
2002				False	True	True	True	True	True	True	True	True	True
2003					True	True	True	True	True	True	True	True	True
2004						False	False	False	True	True	True	False	False
2005							False	True	True	True	True	False	False
2006								False	False	True	False	False	False
2007									False	False	False	False	False
2008										False	False	True	True
2009											False	True	True
2010												False	True
2011													False

Figure 10. Turley-Kramer null hypthosis rejection. Illustrates which age groups reject the null hypothesis and are different from the control group.

Table 1. Number of racers by year

Year	Racers
2012	9729
2011	9030
2010	8853
2009	8321
2008	6397
2007	5688
2006	5434
2005	4325
2004	3899
2003	3542
2002	3330
2001	2972
2000	2166
1999	2352

4. Conclusion

In this study we draw conclusion from applying different methods in data analysis that a gradual difference in distribution appears over the years. This presumption is based on using various plots such as Q-Q, box plots, and density curve, as well as Tukey-Kramer test that confirms the plots. These evidences demonstrate there is a moderate decrease in age of racers over 14 years. Future analysis should concentrate on participants by state and age group. This work could help the organizers concentrate on where and who to market their race to for a greater support and a monetary increase in the amount of funds raised.

5. References

1. http://www.cherryblossom.org/index.php
2. Tukey, John (1949). "Comparing Individual Means in the Analysis of Variance". *Biometrics*. **5** (2): 99–114. JSTOR 3001913
3. Thode, Henry C. (2002), *Testing for normality*, New York: Marcel Dekker, ISBN 0-8247-9613-6, Section 2.2.2, Quantile-Quantile Plots, p. 21.
4. Spear, Mary Eleanor (1952). *Charting Statistics*. McGraw Hill. p. 166.

6. Appendix (Code)

The code is implemented in Jupyter notebook. We leave this part blank and submit the .ipython file together with the project.