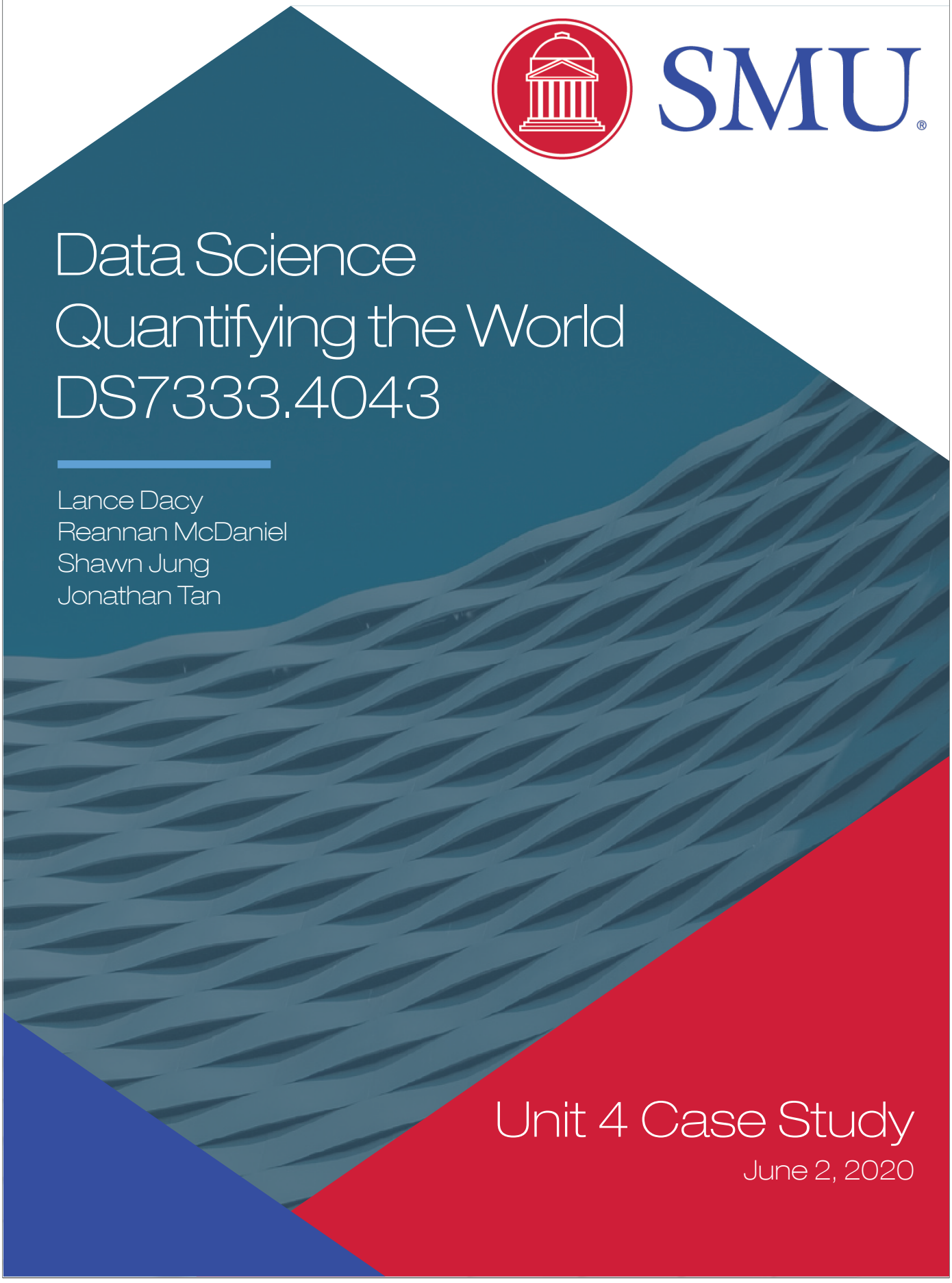


SMU

Data Science Quantifying the World DS7333.4043

Lance Dacy
Reannan McDaniel
Shawn Jung
Jonathan Tan

Unit 4 Case Study

June 2, 2020

Table of Contents

Introduction

Technology today has provided a culture of having everything we want/need available within a moments notice. You want groceries? No problem, go online and order them; they will be at your door step in 2 hours. Need data on airlines and their rates for the past year? Download it at any agency that is tracking that data. It is usually free. Want personal data about how someone purchases items or tracks their fitness results in a day? Well…data privacy laws might prevent you from having access to that data. In fact, you will usually have to pay for a service where people opt-in to provide you that data.

While data surrounds us, having access to it or permission to use it can be a completely different problem. Legislation today protects people from having personal information shared freely unless it is provided through a platform or result set where permission is implied. Take for instance the problem statement about understanding how people's physical performance changes as they age. This data is not always readily available or if it is; it must be supplied by a person or purchased. One source of data comes to mind though; myriad road races, triathlons, or health performance events that happen across the country each year.

Out of the hundreds of thousands of people that participate in road races each year; the race organizers collect information about the runners' times and often publish individual-level data on the Web. These freely accessible data may provide us with insights to our question about performance and age. One example of the many annual road races is the Cherry Blossom Ten Mile Run held in Washington D.C. in early April when the cherry trees are typically in bloom. This is such a popular event; lotteries have to be installed to ensure participation is limited to a safe amount. Once the race is finished; the organizers publish the data online for all to see. It is plausible that this data (over 14 years of history) could supply the data to inspect the question regarding change in a person's performance year over year.

Method

The stakeholders have a particular question of interest that is posed. We have seen that the 1999 runners were typically older than the 2012 runners. They want to compare the age distribution of the runners across all females runners through the 14 years of the races.

- *"Visualize how the distributions change over the years? Was it gradual?*

The age of runners range through to about 80 years, with some outliers at either extreme. Age distribution is expected to be close to normal as runners are chosen by lottery, with exceptions for sponsored teams, elite athletes, or charity fundraisers. In order to reach the question of interest at hand, multiple methods must be used to get the data in a consumable state (mentioned in the Data section above) and various tactics to discern the data will be described in this section.

Given that each year is in a single file and be extracted individually then uploaded to a repository; some years will have varying formats from year to year and will require manipulation using in Python using "BeautifulSoup".

Example URL for Python:

```
http://www.cballtimeresults.org/performances?
division=Overall+Women&page={PAGE_NUMBER}
&section=10M&sex=W&utf8=%E2%9C%93&year={YEAR_NUMBER}
```

The {} sections were iterated with simple counters for the ranges of number of pages per year, and years from 1999 to 2012 respectively. Using the XML tags in hierarchical order, locations for desired data were located using:

- (th) for table headers
- (tr) for rows
- (td) for the cells within that row

Calling for the (td) in each page results in a single dimensional array that contains every cell in the table. In the interest of speed, each row was populated via loop for the desired 9 cells per row. Iterating with counters allows the results of each row to be stored in a single array per year. The array was then written to a csv file for easy retrieval without running all the functions again; providing for re-usability.

All these tasks were wrapped into a single function and called once for each year for bug testing, though in future iterations, this will be plausible for years as well. The resulting thirteen (13) csv files were then imported into R to take advantage of ggplot2.

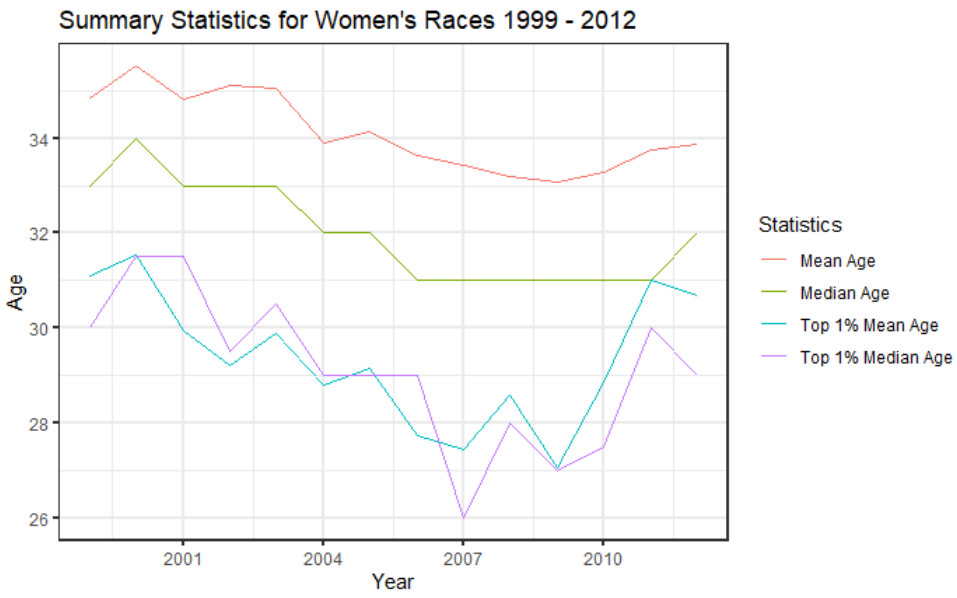
As the data was checked for inconsistencies; the program found that the state name was generating an extra column next to hometown for some of the years and subsequently dropped as the state name of the runners was not pertinent information for the scope of this case study). In addition the data was converted into data-frames, vertically merged with rbind, then exported again as the final csv with all the race results data from 1999 to 2012.

The final csv contained all female runner's results from races in years 1999 - 2012, and served as the baseline for further analysis presented in the conclusion section.

Figure 5: Sample of Final (Combined) csv File of Data

	Race	Name	Age	Time	Pace	PiS.TiS	Division	PiD.TiD	Hometown
1	1999 10M	Jane Omoro	26	0:53:37	5:22	Jan-58	W2529	1/559	Kenya
2	1999 10M	Jane Ngotho	29	0:53:38	5:22	Feb-58	W2529	2/559	Kenya
3	1999 10M	Lidiya Grigor	NR	0:53:40	5:22	Mar-58	NR	NR	Russia
4	1999 10M	Eunice Sager	20	0:53:55	5:24	Apr-58	W2024	1/196	Kenya
5	1999 10M	Alla Zhilyaye	29	0:54:08	5:25	May-58	W2529	3/559	Russia
6	1999 10M	Teresa Wanj	24	0:54:10	5:25	Jun-58	W2024	2/196	Kenya
7	1999 10M	Elana Viazov	38	0:54:29	5:27	Jul-58	W3539	1/387	Ukraine
8	1999 10M	Gladys Asiba	NR	0:54:50	5:29	Aug-58	NR	NR	Kenya
9	1999 10M	Nnenna Lync	27	0:55:39	5:34	Sep-58	W2529	4/559	Concord
10	1999 10M	Margaret Ka	30	0:55:43	5:34	Oct-58	W3034	1/529	Kenya
11	1999 10M	Susannah Be	30	0:56:13	5:37	Nov-58	W3034	2/529	Eugene
12	1999 10M	Kelly Keeler (	37	0:57:23	5:44	Dec-58	W3539	2/387	Bloomingtor
13	1999 10M	Marie Boyd (	39	0:57:24	5:44	13/2358	W3539	3/387	Albuquerque
14	1999 10M	Betsy Kempt	32	0:57:51	5:47	14/2358	W3034	3/529	Chapel Hill
15	1999 10M	Naoko Ishibe	30	0:58:05	5:49	15/2358	W3034	4/529	Washington
16	1999 10M	Bea Marie A	31	0:58:36	5:52	16/2358	W3034	5/529	Columbia
17	1999 10M	Connie Buck	NR	0:59:36	5:58	17/2358	NR	NR	Lancaster
18	1999 10M	Sharon Servi	25	0:59:42	5:58	18/2358	W2529	5/559	Alexandria
19	1999 10M	Donna Moor	38	0:59:48	5:59	19/2358	W3539	4/387	Silver Spring

Figure 6: Summary Statistics for Combined Race Information



All participants in each year's race mostly decreases from 1999 to 2007 for both the average participant and the fastest runners. The sharp downward spike in mean age for elite runners in 2007 seems to have been an unusually young group of elite. This trend is not at all shown in the mean age of all participants, indicating that enough ordinary runners are beginning to attend each race that they balance out the lower mean ages of the top finalists.

Furthermore, age appears to bottom out in 2007, either indicating that relative age group cohort has continued to be the fastest, or that enough middle aged runners are performing at a high level to balance out the youngest runners.

The rise in average age past 2008 is reflected in the increased total race participation, with 2009 reporting 8,000 runners; 2,000 more than the previous year. Enough of these new runners are older and performing high enough to drive up the average age of the top finalists significantly, though for both elite and normal runners, the median age increases less than the mean, indicating that a few older than average runners skew the mean age distribution upwards.

Data

Data Source

After each year's race, the organizers publish the results at http://www.cherryblossom.org/. This data offers a tremendous resource for learning about the relationship between age and performance.

Data collecting can be managed by various methods. Web Scraping is a technique used to retrieve information from websites and then saved in a database or repository for data analytics. There are enormous amounts of data available across many interests. Information is collected on road races in the running community that includes several metrics such as time, year of race, gender and distance.

These data are consistent enough that repeatable programming methods can be utilized to allow the stakeholders to vary their questions related to the data provided.

Exploratory Data Analysis

The per year results for the race does not have standardized formats. For instance, 2008 and 2007 have completely different row titles and orders. Initial attempts to parse this data from an XML format in R with rvest xml2 was arduous and time consuming, as each new format required a different set of rules to govern the separation of characters into the appropriate strings and columns.

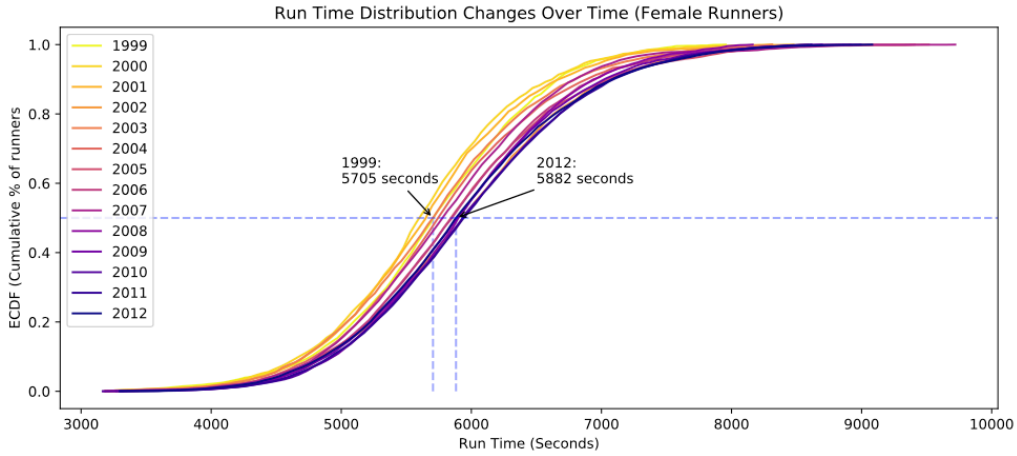
The non-standard nature of each year's results page also meant that XML structure locations for tables and rows changed each year. The problem of different race formats was solved by looking at the paginated results and switching to Python, which has a more novice friendly package; BeautifulSoup. The searchable results page effectively contained all the results from all years, where the columns and formats are all standardized.

By iterating on the page numbers of the results for each race, BeautifulSoup can aggregate the table from each page of the race results with consistent columns and formats. Given that all the paginated results are accessed through the same format URL, Python can iterate through the page number and year to access all the results in this category from 1999 to 2012.

Figure 1: Small Sample of Resulting Scraped csv File

Race	Name	Age	Time	Pace	PiS/TiS	Division	PiD/TiD	Hometown
1999 10M	Jane Omoro	26	0:53:37	5:22	Jan-58	W2529	1/559	Kenya
1999 10M	Jane Ngotho	29	0:53:38	5:22	Feb-58	W2529	2/559	Kenya
1999 10M	Lidiya Grigor	NR	0:53:40	5:22	Mar-58	NR	NR	Russia
1999 10M	Eunice Sager	20	0:53:55	5:24	Apr-58	W2024	1/196	Kenya
1999 10M	Alla Zhilyaye	29	0:54:08	5:25	May-58	W2529	3/559	Russia
1999 10M	Teresa Wanj	24	0:54:10	5:25	Jun-58	W2024	2/196	Kenya
1999 10M	Elana Viazov	38	0:54:29	5:27	Jul-58	W3539	1/387	Ukraine
1999 10M	Gladys Asiba	NR	0:54:50	5:29	Aug-58	NR	NR	Kenya

Figure 2: Run Time Distribution Changes Over Time



In the quest to answer the question of interest, the data would need to be inspected to see if the distribution of running time has been changing over the years. The visual plot of empirical CDF shows the running time of mid-range runners are increasing over time. For instance, the running time of 50th female participants rose from 5,705 to 5,882 during the 1999 to 2012 period of inspection.

Figure 3: KL Divergence Reference

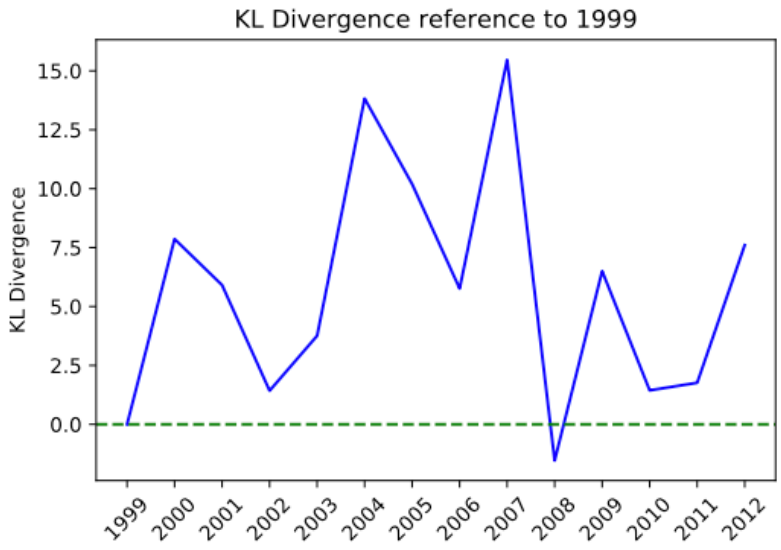
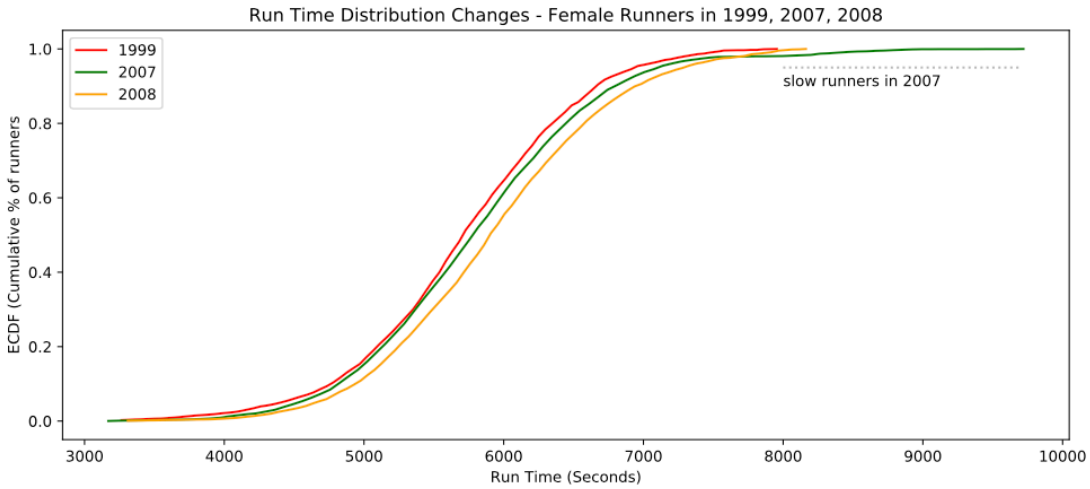


Figure 4: Run Time Distribution Changes for 1999, 2007, and 2008



In an effort to conclude if there is distinct variation in 1999 to other years; the KL (Kullback-Leibler) divergence was used to quantify the difference and provide visualizations to inspect the divergence.

Figure 3 displays the KL Divergence of the year 2007 as the highest (15.4706), and the next year; 2008 as the lowest (-1.5351). The interpretation of this data is that there were numerous slow runners in 2007 compared to 2008. Figure 4 indicates that the distribution in 1999 and 2008 were similar given the short tails.

Conclusion

The work to tidy the data into a repeatable/consumable format was significant work. Upon inspecting the various files produced by the programming routines, confidence was restored in the cleanliness of the data.

Figure 7: Box-plot for Age Distribution by Year

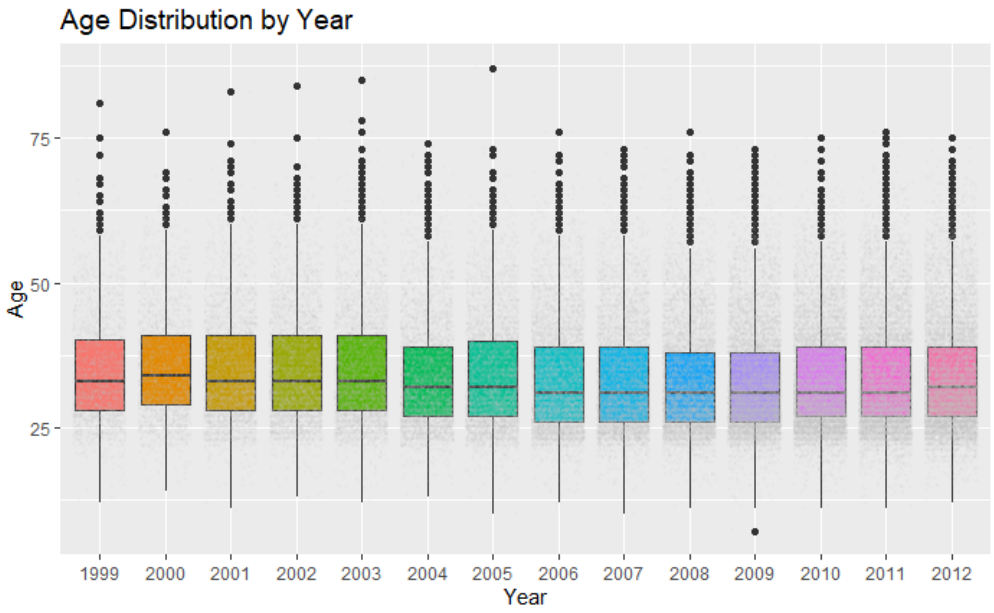


Figure 4 allows for the visualization of the descriptive statistics obtained through the combined data set. For each year from 1999-2012 there are outliers near the max tails, as age increases with more dispersion in the first seven years.

The min values remains consistent with only one year having an outlier for young age in 2009. The quartiles have very little variation across age year over year. Years 1999 and 2001 appear to have the most spread and this could be due to the amount of data collected and how it was collected and logged.

The year with the least amount deviation is 2012, which could simply be improvements in the data collection process as technology improves. To answer the question of interest:

- *There is minimal variation of the female runners year over year.*

Figure 8: Age Distribution by Year

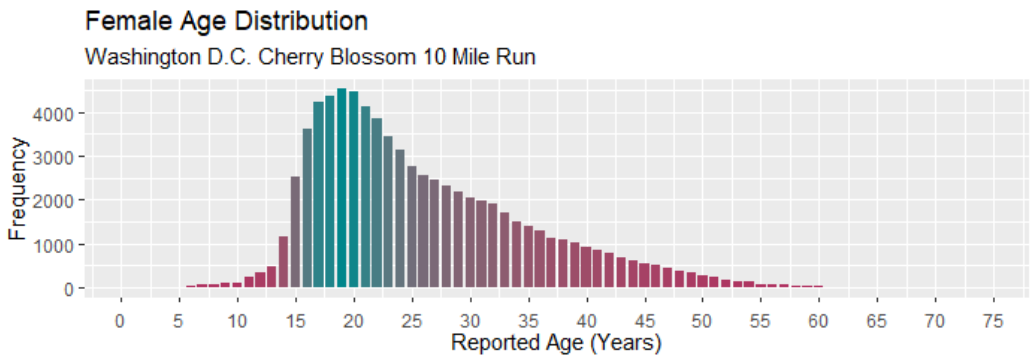
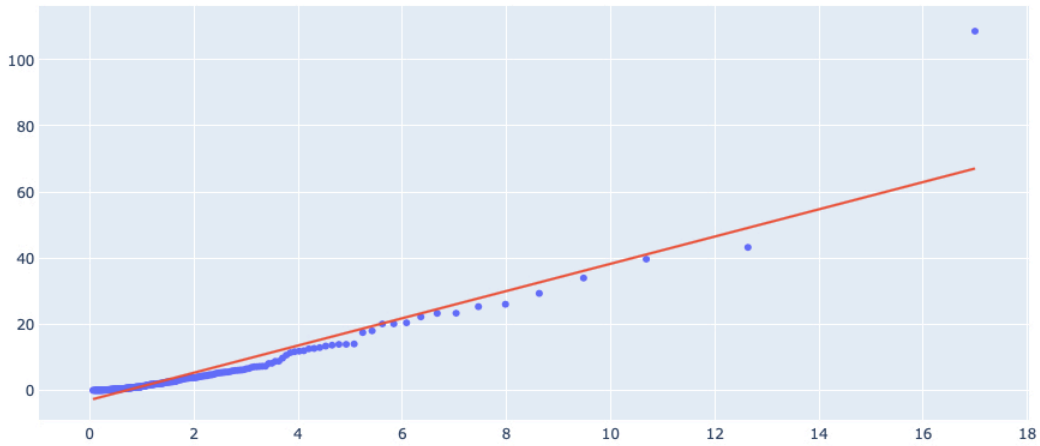


Figure 9: Age Distribution by Year



The majority of runners from all 14 years of races were between 18 and 25 years old, though the highest performing runners averaged around 28 years old. Given the random nature of the lottery entry, the relatively high number of runners under 30 could indicate the level of engagement with running as an activity. Further studies on data from different running events in other locations could determine if this is a wider trend or specific to this event/location.

Recommendation

Given the risk of consuming data that has been scraped from a screen as well as the legal implications of using this data for studies beyond novel; it is recommended in the future that the stakeholder purchase this data in a clean consumable format for distinct analysis. Myriad health information exists from data providers or this type of data as well as events that track performance of its attendees beyond simple race times.

Information such as speed, course statistics, and weather conditions could be combined to assist in determine the runners actual rate and performance conclusions based on age year over year. Scraping this information from a website proves primitive at best. Given the proper resources (time, money, and skill), a much more predictable model could be created.

Appendix

As with any technology; the system required to generate meaningful data, cleaning of the data, as well as the models that were implemented require various software packages. The raw "code" used to explore the methods detailed above can be found in the following libraries:

- GitHub Repository for R and Python