Case Study 2, Unit 4 - Web Scraping
Daniel Davieau, Lei Jiang, Kristen Rollins
4 February 2020

Introduction

With the creation of the world wide web in the 1990s and its uncontested explosion since then, there is hardly a scarcity of data. In fact, there are more than 1.5 billion websites that exist today, containing more information than we can hope to digest in a lifetime. Nevertheless, there is great potential in retrieving relevant information to bring insights to personal as well as commercial applications. While some of this data is on private websites requiring authorization and authentication, there is still much data to be found on completely free and public webpages. Fortunately, last year a U.S. court ruled that scraping data from public websites is not illegal, even without prior approval. We must still take steps to respect and not harm the websites we would like to obtain data from, but we now have the freedom to collect public data from these sites.

Even with legal permission, web scraping is not trivial, as the data that floats on the web is often unstructured or inconsistently formatted. We experienced this first-hand, as we sought to collect data for the female runners of the Cherry Blossom race held annually in Washington D.C. This data is available on their free and public website, but is spread across many pages with varied formatting. With some effort, however, we were able to retrieve and shape this data into a data frame for our ultimate goal of analyzing the age distributions of the female runners across the years. From this analysis we found that there are very gradual but consistent changes in the age distributions, which decrease from 1999 through 2009, then slightly increase again from 2010 through 2012.

Background

The Cherry Blossom Ten Mile Run is held annually in Washington D.C. in early April, to coincide with the cherry blossom trees being in bloom. The road race began in 1973 as a training run for the Boston marathon, but its popularity has grown over the years, with nearly 17,000 participants in 2012. Each year, the official, individual-level results are published by the organizers to their public website at http://www.cherryblossom.org/, where the results are separated by men and women. We will only be looking at the historical results from 1999 through 2012, as the subsequent results were published in a very different format. This is part of the struggle of web scraping, as we have very little control over the data we want to collect.

The Nolan and Lang textbook, *Data Science in R*, describes in detail their process of retrieving data for the men's 10-mile races from 1999 to 2012. They also outline their analysis of comparing age and performance for these male runners. In their analysis, they found that the 1999 male runners were typically older than the 2012 male runners. With this framework provided, we sought to generalize the data extraction process to pull data for the female runners instead. Our goal was to scrape the appropriate pages from the Cherry Blossom website, parse the tables found, then clean and combine the data to create a full data frame of female runners

to perform analysis with. We could then use this data structure to visualize the age distributions of female runners across the years 1999 through 2012, to compare and expand the analysis from the Nolan and Lang text.

Methodology

To create a data frame for analysis with the 1999-2012 female Cherry Blossom runners, we first had to scrape this data from the Cherry Blossom website. We utilized the programming language R and the XML package to do so. Because web pages are subject to change arbitrarily and are not required to be in any consistent format, our web scraping task was iterative with lots of debugging. Our full process can be found in our code base (referenced in the Appendix), but here we will only outline some of our key findings and decisions to increase confidence in our final result.

We primarily followed the procedure provided in the Nolan and Lang text, only using the female result web pages rather than the male result pages. We had to make some tweaks to their `extractResTable` function since the women's pages did not follow the exact HTML structures of the men's pages, but this was accomplished through trial and error. In addition, we used similar functions to locate the header row and separator row (denoted by many "=" signs) for each year's result table. We did find that the women's table for 2001 did not have header or separator rows, but we were able to resolve this problem by substituting in those rows from the 2002 table.

Next we were able to extract the desired variables in a similar manner to the text. We used the `extractVariables` and `createDF` functions provided to separate the data into columns and make a data frame with appropriate data types. In the `createDF` function, we chose to only select the `name`, `home`, and `age` columns (as well as adding `year` and `sex` columns), because our ultimate analysis will only be concerned with comparing age distributions over the race years. Even though we will not specifically view individual-level results in our analysis, we retained the name and hometown columns as a method for error checking. If we found that some data point looked suspicious, we cross-referenced it with the original table on the Cherry Blossom website by searching for an individual's name in the appropriate year.

Lastly, we examined the NAs that were introduced into the data frame, to verify that they were not errors in parsing the data. We found that NAs only occurred where there were blank or invalid values in the age column, so we were satisfied there. We then chose to remove all rows with missing ages from our data frame. This eliminated 39 rows, with a maximum of 11 rows removed from any given year (2005). Some of these rows were not even records of runners, but rather trailing information that was picked up from the ends of the tables. With our full data frame having over 76,000 rows, we were confident that removing these 39 rows with missing ages would not significantly influence our distribution analysis. In this way we were satisfied that our web scraping and data cleaning procedures accurately obtained the female runner results, and we could proceed to analysis of the age distributions across the years.

Results

Objective 1: Web Scraping

After following the web scraping and data cleaning process described above, we stored our final data frame in an object called `cbWomenDF`. To confirm that our data frame could be used for analysis, we used the function `str` to display the structure of an R object. We obtained the following output describing our data frame.

```
> str(cbWomenDF)
'data.frame':     76038 obs. of  5 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           " "Eunice Sagero
" "Alla Zhilyayeva       " ...
 $ home: chr  "Kenya              " "Kenya              " "Kenya              "
"Russia             " ...
 $ age : num  26 29 20 29 24 38 27 30 30 37 ...
```

The two columns our analysis will be concerned with are `year` and `age`; from this output we verified that those two columns are in the appropriate formats of `int` and `num`, respectively. We also noted that the `name` and `home` values often had trailing whitespace. These columns could have been cleaned up further, but we chose to leave the values in their present state because they will not affect our analysis. We also confirmed that there were no NA values remaining in the data frame.

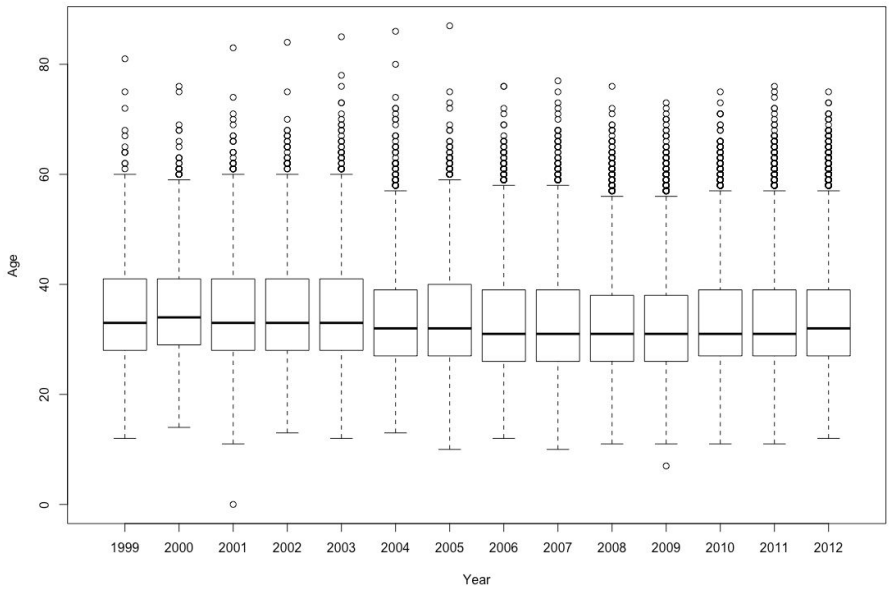

Figure 1. Age vs. Year Boxplots for Cherry Blossom Female Runners

To further confirm that we could appropriately use the `year` and `age` columns for our age analysis, we created a simple boxplot of ages versus years, seen in Figure 1. For the most part the distributions among years are relatively similar to each other, and the values are within a reasonable age range, leading us to believe that no errors occurred when we parsed the age column. We did investigate the obvious oddity, i.e. the point in 2001 that was recorded to have an age of 0. We found that this "runner", Loretta Cuce, was actually recorded in the original web page to have an age of 0, so this point was not a parsing error. We noticed that she was recorded immediately after a 31-year-old Carolann Cuce in the same year, and they had exactly the same running time; from this we made the connection that a mother probably completed the race while pushing her daughter in a baby carriage. Since this was a quirk, rather than an error, of the dataset, we chose to preserve this data point.

Objective 2: Comparing Age Distributions

The Nolan and Lang text found in their analysis of the male runner data that the 1999 runners were typically older than the 2012 runners. Our objective was to find out whether this same trend occurred in the female runner data, and extend our analysis to compare the age distributions from those and all years in between. First we compare 1999 and 2012 female runners in Figure 2. On the left we have a quantile-quantile (q-q) plot, which is a graphical technique for determining if two data sets come from the same distribution. If they are from the same distribution, we should see the points forming a line that is roughly straight. Here we see that the line is fairly but not exactly straight, indicating similar but not identical distributions. The density curves on the right confirm this, as we see that the ages of the 1999 female runners are spread out among slightly higher ages than the 2012 female runners. We also compared the mean and median values for these distributions -- the 1999 female runners had a mean age of 34.9 and median age of 33, while the 2012 women had a mean age of 33.88 and median age of 32. So, the age difference observed in the men's data is also observed in the women's data, but the difference is not as drastic for the female runners.

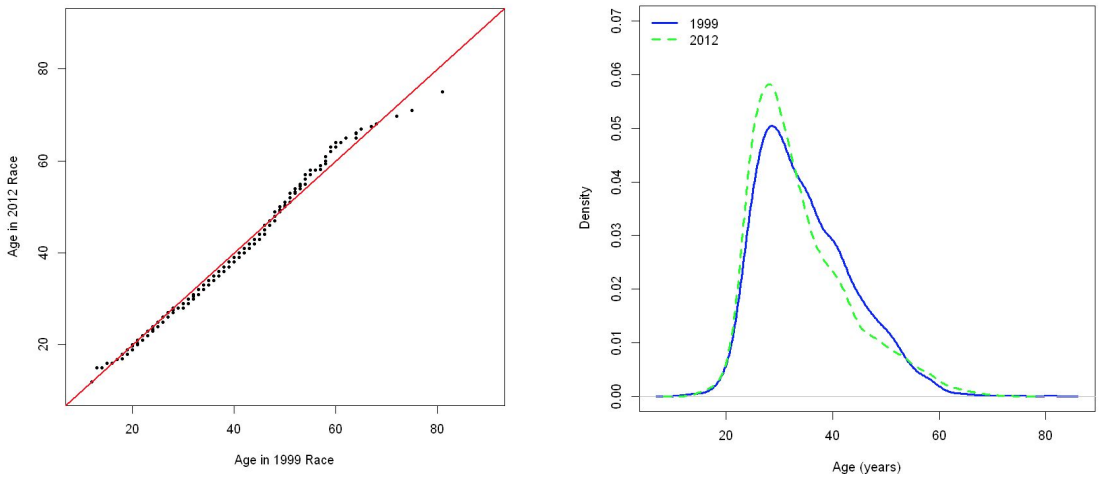

Figure 2. Quantile (left) and Density (right) Comparisons of 1999 and 2012 Female Runners

Since there was some evidence of decreasing female age based on the first and last years of our data, we want to explore how the distributions changed over all the years we have available. Figure 3 illustrates the age distributions for a selection of these years. Because the distributions regularly overlap, we could not effectively display all 14 years on this plot, but we selected a range of the years that portrays some differences in distributions. Here we pay most attention to where the density peaks occur; we observe that the peak age decreases from 1999 to 2005 to 2009, which peaks at the lowest age. Then there is a slight increase from 2010 through 2012. This plot confirms our suspicions that there is some variation in age distribution throughout all of the years, although these changes are rather slight and gradual.

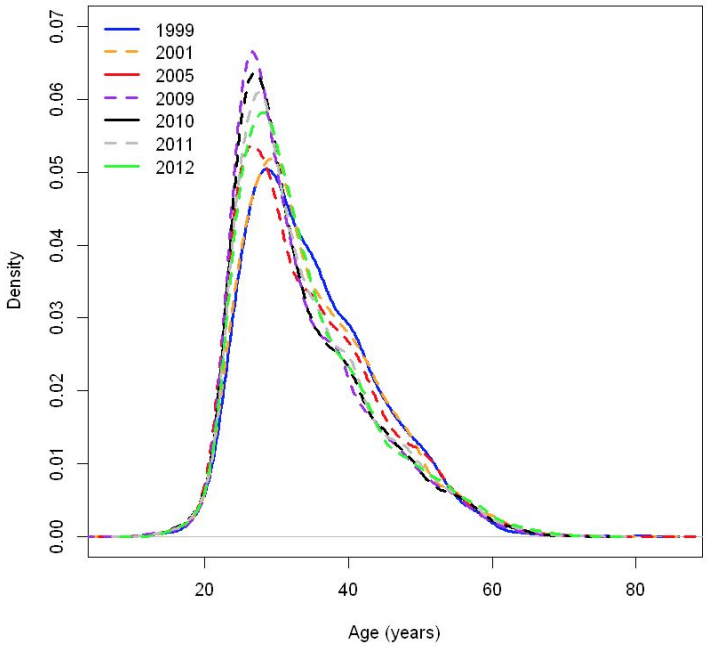

Figure 3. Age Density Curves for a Sample of Years

The violin plots in Figure 4 demonstrate both summary statistics (as we saw in the Figure 1 boxplots) as well as density curves for the female runner ages from 1999 through 2012. This extra visual information helps us to most clearly see the trends of the age distributions across the years. While we see that the maximum ages increase up through 2005, we found that these were due to the same woman running the Cherry Blossom every year from 2001 to 2005, where she was the oldest female runner at the age of 83 through 87. The white dots denoting the median value are more indicative of the trends of the distributions, since the median is not sensitive to such outliers. We observe that the median values appear to very gradually decrease from 1999/2000 all the way to 2009, then slightly increase again from 2010 to 2012.

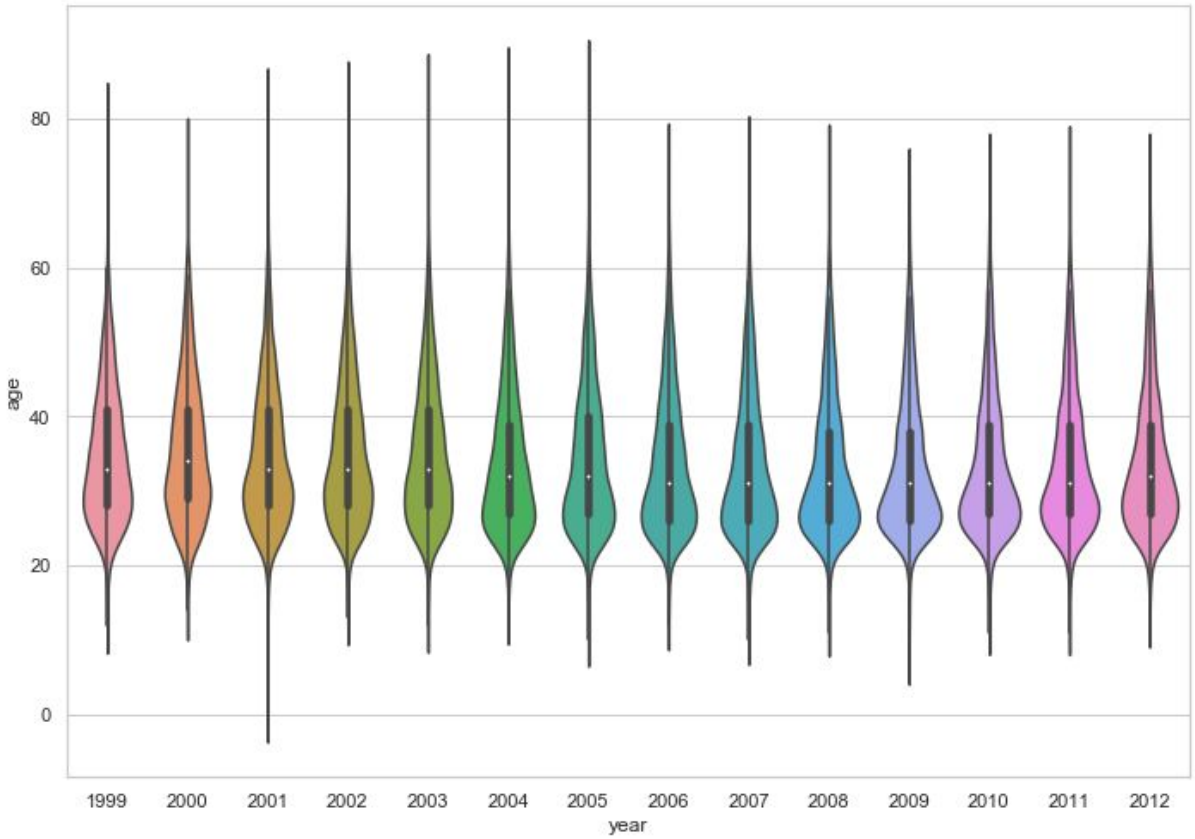

Figure 4. Age vs. Year Violin Plots for Cherry Blossom Female Runners

Conclusions

From our analysis of the Cherry Blossom female runners, we found that from 1999 to 2009 there was gradually a higher concentration of younger runners, then from 2010 to 2012 the age distributions gradually moved to be a bit older (though not as old as the earliest years). It is unclear exactly why these distributions changed so slowly but consistently. A possible explanation could be that U.S. women's health on average was declining over this time, leading to fewer older women being able to participate on average. Additional analysis could pull the Cherry Blossom data from 2013 through the most recent race, to determine whether the slight age increases in 2010-2012 was an indication of an upward trend or simply a minor departure from the actual downward trend. In addition, it seems that the difference in age distributions was more extreme for the male runners than the female runners. Further analysis could show comparisons between the two groups across the years as well.

References

- Nolan, D., and Temple Lang, D. (2015), Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving. Boca Raton, FL: CRC Press (NTL)
- R code provided by professor Blanchard, adapted from http://rdatasciencecases.org/

- https://www.internetlivestats.com/total-number-of-websites/
- https://techstartups.com/2019/09/13/web-scraping-legal-us-court-says-scraping-public-data-website-without-permission-not-illegal/

Appendix

Codebase

Our code has been included in a zipped folder called lJiang_kRollins_dDavieauCase2_Code.zip. The files in this folder include:

- lJiang_kRollins_dDavieauCaseStudy2.R - our primary code base, where we processed the data and made initial visualizations
- lJiang_kRollins_dDavieauUnit4CaseStudy.ipynb - copy of our data processing, with additional analysis performed
- SeabornViolinPlots.ipynb - data imported into Python to make additional visuals with the Seaborn package
- cbWomen.rda - data object that can be loaded directly into R, to skip data processing
- cbWomenDF.csv - csv file so that we could import the cleaned data into Python

Assignment

- **Q.7** Follow the approach developed in Section 2.2 to read the files for the **female** runners and then process them using the functions in Section 2.3 to create a data frame for analysis. You may need to generalize the createDF() and extractVariables() functions to handle additional oddities in the raw text files.
- **Q.10** We have seen that the 1999 runners were typically older than the 2012 runners. Compare the age distribution of the **[female]** runners across all 14 years of the races. Use quantile–quantile plots, boxplots, and density curves to make your comparisons. How do the distributions change over the years? Was it a gradual change?