How Have Marathons Changed over the Past Few Decades

Haley Vemmer

IS 470: Business Intelligence and Big Data

Professor RJ Podeschi

December 15, 2020

Final Project

Outline

1. Project Proposal
   1. Background
   2. Data
   3. Expectations
2. Documentation
3. Findings
4. Works Cited
5. Project Proposal

Several weeks ago, I was reading a marathon training book, and there were several quotes about how the marathon has changed over the past few decades. Many of these dealt with the marathon attracting more recreational runners in addition to the elite runners who already participated in them and how this has affected the prestige of the event. According to Runner’s World, “Elite winning times are faster today, but the average runner back then was considerably quicker than modern mid-packers” (Cooper 2019). In my project, I would like to look at how the average finishing times for these races have changed as a whole, but also breaking them down into quartiles, gender, age range, and other variables.

I will be using a data set used in a study done by a professor at The University of Chicago in the Booth School of Business, which can be found at the following link <https://faculty.chicagobooth.edu/george.wu/research/marathon/data.htm>. I am going to use a sample of the full data set, as the sample data set on this website only has data from 2003 – 2013. I was planning to use Excel to clean up some of the data in Excel, but the data set contains more rows than Excel is able to handle. So, I will likely be doing all of the work on this project in R/R Studio.

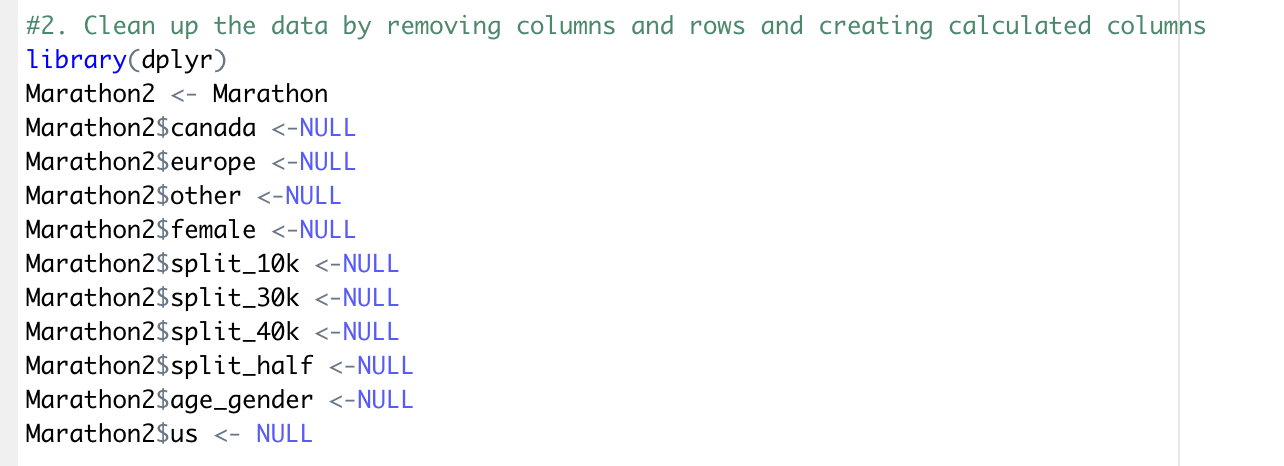
I’m expecting that my results will confirm that marathon races have seen a variety of changes over the past few decades. I believe that these changes will be seen in the finishing times as a whole, but specifically in the middle and lower quartiles. I don’t believe I will see a drastic change in the average time for the upper quartiles. I think there will be a drastic change in the average number of participants, and I also expect to see changes in the average ages and possibly gender of participants.

1. Documentation

The R Code for the following is available at <https://github.com/hvemmer/MarathonData>. My first step was to download the data from the link above and load it into R. I used both the full data file and a file containing information about each individual race.Graphical user interface, text, application

Description automatically generated

Next, I began cleaning up my data. The first step to clean the data was to remove the columns I knew wouldn’t be used.



After removing these columns, I then removed some outlier entries that would skew the data.

Text

Description automatically generated

After removing these entries, I created some calculated columns to be able to categorize the data later. Graphical user interface, text, application, table

Description automatically generated

I then removed a few more columns that were no longer necessary and created some small data sets broken up by decade, which were only used for getting some statistics. I also created one more calculated column.

Graphical user interface, application

Description automatically generated

Next, I pulled some statistics to look at.

Text

Description automatically generatedText

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

Finally, I began to create my visualizations. To do this, I first loaded in several packages. The first visualization I chose to make was a density plot of average times by decade. Chart, line chart

Description automatically generated

Figure 1: Mean Times by Decade

My second visualization was a density plot of average times by gender by decade.

A picture containing text, sky, flock

Description automatically generated

Figure 2: Mean Times by Gender by Decade

The third visualization I created was a box plot of chip times per decade, which I compared to my fourth visualization – a box plot of the average finishing times for each race.

Chart, box and whisker chart

Description automatically generated

Figure 3: Individual Chip Times by Decade

Chart, box and whisker chart

Description automatically generated

Figure 4: Average Race Time by Decade

My fifth visualization was a bar graph showing the average number of finishers per race in each decade. For comparison purposes, I also created bar graphs of the number of participants in each decade and the number of races in each decade.

Chart, bar chart

Description automatically generated

Figure 5: Average Number of Race Participants by Decade

Chart, bar chart

Description automatically generated

Figure 6: Number of Marathon Participants by Decade

Chart, histogram

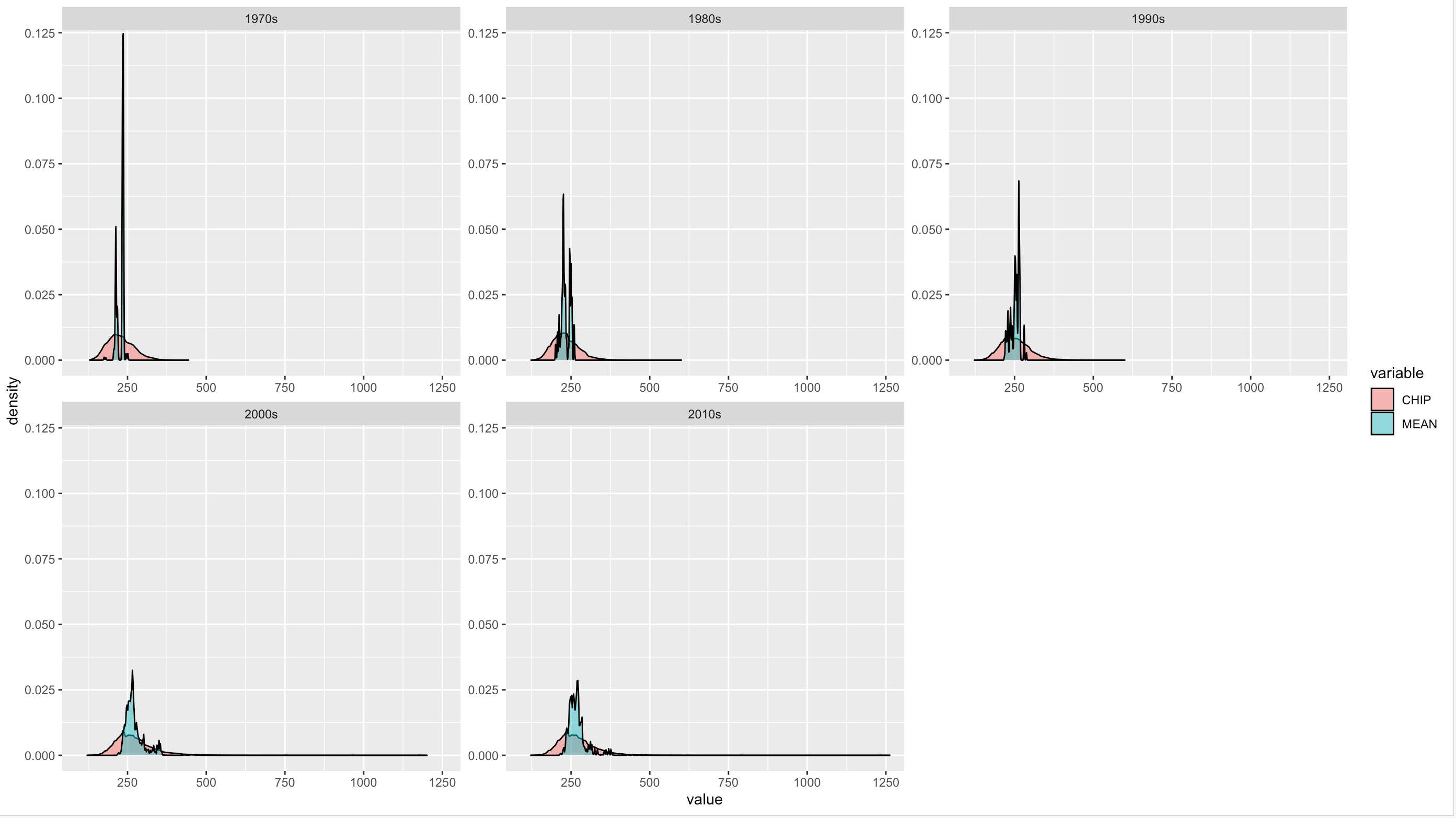
Description automatically generated

Figure 7: Number of Marathon Races in Each Decade

The sixth visualization I created shows the number of races by location in each decade.

Graphical user interface, application

Description automatically generated

The final visualization I created was a series of density plots illustrating the differences between the individual chip times and the average race times. 

1. Findings

Most of my findings from this analysis were what I expected to find. Marathons aren’t just for elite runners anymore. More people with an average pace are starting to run these races, thus altering the average times. The density plot of average times by decade shows that most races still have a concentration of participants finishing the race between 200 and 300 minutes. However, more recent races have many more outliers than ones in the past. The density plot breaking these times down by gender shows that this is occurring with both male and female participants. I also found it interesting to see how the increase in chip time outliers affected the average marathon finish time, as can be seen by my final visualization. These results are a combination of my first two visualizations, so they are what I expected, but I did find it interesting to see how they affected each other.

Similarly, the box plots I created to look at finishing times and average race times show that most marathoners are still elite runners; however, there is a wider range of outlying data points. The average chip time by decade ranges from about 229 minutes (1970s) to about 271 minutes (2000s), so the average chip time has increased about 42 minutes, which is a decrease of pace by about 96 seconds per mile. The average race finishing time by decade ranges from about 215 minutes (1970s) to about 277 minutes (2010s), so races in the 2010s were completed about an average of 62 minutes slower than races in the 1970s. This time difference equates to the average pace in the 2010s being about 2:22 slower than the average pace in the 1970s.

One thing that I learned during this analysis was about the large drop in number of average race participants by decade between the 1990s and 2000s. This was a bit unexpected, but not necessarily surprising, especially considering the increasing popularity of the race during this time would cause an increase in number of races being offered. It was interesting to see how fast the number of marathons being held increased between the decades.

One question that I was asked during class presentations was whether or not course time limits may have impacted who could participate in the older races. While I didn’t have an answer during class, I was curious, so I looked it up. While many races today impose a cut-off time of about six or six and a half hours (or at least the races I have looked into entering), most races didn’t really impose one at all fifty years ago, or really up until about ten years ago, as walkers and walk/joggers just didn’t enter them (McGillivray). People choosing (and being able to choose) to walk the distance or walk/jog the distance in intervals is likely the main factor behind the slower times seen in races today.

1. Works Cited

Cooper, Bob. “10 Ways Running Has Changed in the Last 50 Years.” *Runner's World*, 10 Jan. 2019, www.runnersworld.com/uk/training/a775570/10-ways-running-has-changed-in-the-last-50-years/.

McGillivray, Dave. “What's the Deal with Course Time Limits?” *Runner's World*, Runner's World, 17 Sept. 2020, www.runnersworld.com/races-places/a20820446/whats-the-deal-with-course-time-limits/.