Hope Muller

Professor Mijanovich

Statistical Consulting

20 December 2020

An Analysis of New York City Marathon Finishers

Introduction & Motivation

The first Sunday in November in New York City is a day where streets in all boroughs are lined with crowds cheering for friends, family, and strangers as they attempt the feat of running 26.2 miles. It is an inspiring event for anyone that happens upon the crowds.

Fifty-five years ago, the New York City Marathon had 55 finishers, and it has now grown to the world's largest marathon, with 53,639 finishers in the 2019 race. People travel from all over the world to race, some running the entire 26.2 miles at a pace faster than I can run one mile.

As a runner, I am intrigued by running trends and conclusions that I read about online. Some say that runners are getting faster over the years (Minsberg, 2020). But are they? I have read that men and females peak in their running times at specific ages (Zavorsky 2017). However, is that true? And what about cold and hot days? Journalists give definitive numbers on how someone's mile-pace is affected by temperature changes on race day (Hosier, 2019). I looked more into these three topics.

Research Questions

I considered three questions about New York City Marathon finishers.

1. Has median race pace, subsetted by age, sex, and speed quartiles, changed over the years?
2. At what age do runners peak in their race pace?
3. Do cold and hot race days impact runners' race pace compared to average temperature race days?

Data

The New York Road Runners reports individual-level runner data on their public website. I used the RSelenium package in R to scrape runner data from several years of the race. The data set includes most runners (99% - 100%). The missing 0-1% is due to glitches in the scraping tool, server, or internet, during the scraping. In total, there were 461,163 runners in the dataset, with 68,874 of those runners repeating the race more than once. Of the repeat runners, 143 ran the race all thirteen years that I collected, 2001-2003 and 2009-2019. Figure 1 offers more detail about the demographics of each year's race. There is an increase of runners each year. There are more males than females each year, and on average, males are faster and older than females. Figure 2 visualizes race participant growth over the years within specific age groups. The age group with the most runners is 40 to 44, while minimal runners under age 20 or over age 70 run the race.

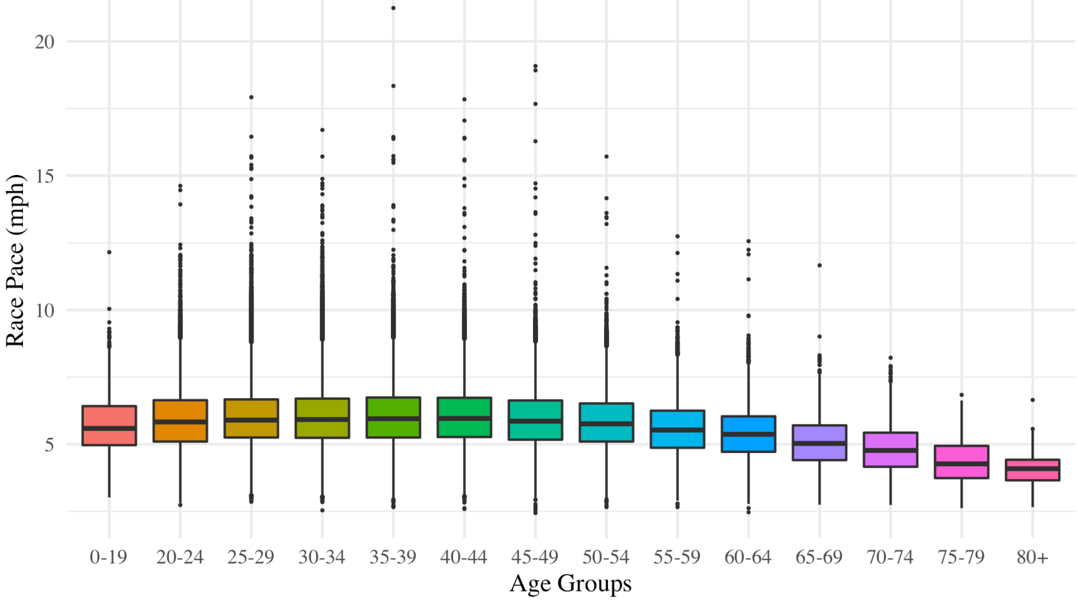
In addition to runners specific data, I collected the minimum, mean, and maximum temperature, in Fahrenheit, for each marathon race day since 1970 from the Farmer's Almanac website. Inches of precipitation is also included. There was no marathon in 2012 due to Hurricane Sandy.

Figure 1: NYC Marathon Finisher Summary (2001-2003, 2009 - 2019)

Race Year	Sex	Total Runners Inc.	Mean Fin. Time (h:m:s)	Mean Age	Mean Speed (mph)	Mean Temp (℉)	Rain (in.)
2001	F	6,851	4:46:09	37	5.7	54.9	0
	M	16,669	4:21:57	40	6.3		
2002	F	10,194	4:50:33	37	5.7	41.3	0
	M	21,610	4:23:48	40	6.2		
2003	F	11,718	4:58:33	37	5.5	63.8	0
	M	23,018	4:32:54	40	6		
2009	F	15,168	4:42:12	38	5.7	56.1	0.14
	M	28,441	4:15:23	42	6.4		
2010	F	16,156	4:44:43	39	5.7	44.3	0
	M	28,950	4:18:15	42	6.3		
2011	F	15,335	4:42:40	39	5.8	47.9	0
	M	28,149	4:15:51	42	6.4		
2013	F	19,568	4:45:00	39	5.7	50.9	0
	M	30,695	4:18:35	42	6.3		
2014	F	20,420	4:50:36	39	5.6	44.3	0
	M	30,101	4:23:33	42	6.2		
2015	F	20,696	4:53:33	39	5.5	56.6	0
	M	28,891	4:27:39	42	6.1		
2016	F	21,302	4:53:33	39	5.5	54.4	0
	M	29,558	4:26:09	42	6.1		
2017	F	21,049	4:53:09	39	5.5	55	0.01
	M	29,631	4:25:09	42	6.2		
2018	F	22,117	4:53:51	39	5.5	48.9	0.37
	M	30,618	4:26:51	43	6.2		
2019	F	22,715	4:52:18	40	5.6	47.9	0
	M	30,859	4:23:39	43	6.2		

Figure 2 shows a summary of the NYC Marathon race paces divided by age groups. The majority of runners between ages 25 and 44 have very similar time ranges, shown by the nearly identical box plots. However, there are numerous outliers for the majority of ages. Some marathoners run much faster than the majority of people. It is also noticeable there are some lower outliers as well. Some people walk the entire race, which results in 8 - 10 hour finishing times.

Figure 2: NYC Marathon Runner's Race Paces Compared by Age

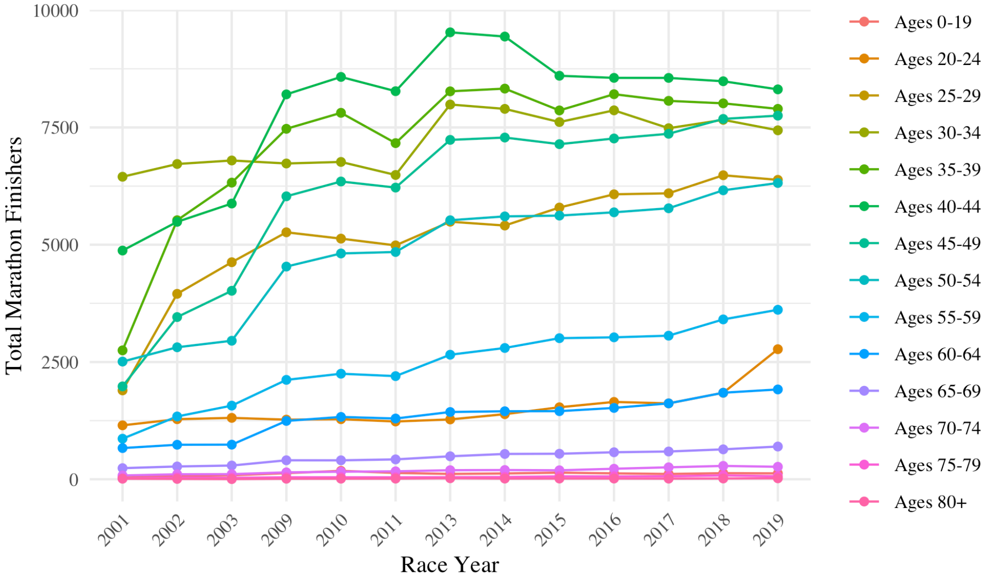


Measures & Methods

The analysis was completed in RStudio with the following R packages: data.table, dplyr, ggplot2, lubridate, MatchIt, h2o, lme4, colbalt, and rsq.

For the first question around if runners are getting faster, I found the 12.5 percentile time, the median, and the 87% percentile time, for each year and age group, subsetted by sex. I explored slope trends between years to identify if runners' pace is changing over time. I hypothesized that fast runners (the top 25%) are getting faster, on average, average runners (middle 25% - 75%) are remaining the same, and slower runners (bottom 25%) are getting slower. As I mentioned in my introduction, it is thought that fast runners are faster than ever. My hypothesis around slower runners getting slower, on average, is because running is getting more popular, and marathons have a growing number of participants. More runners between ages 20 and 69 years old have been running the NYC Marathon over the years, as you can see in Figure 3.

Figure 3: NYC Marathon Finisher Trends by Age Group



For the second question around runners' peak age for pacing, I started with a quadratic model for each sex with age squared, age, and race pace using all runners included once. For repeat runners, I randomly sampled one race for each person. I then looked into a best-fit model with race pace as the outcome, for each sex with all runners included once, considering minimal variables. Finally, I produced a repeat runner model for those running the marathon five or more times in my data set. There were 5,979 participants. I explored models with various variables and tried subsetting by sex, using a median-centered race pace as the outcome.

As you can see in Figure 4, there are vastly different patterns between individuals in their pacing from year to year. Figure 4 includes all female individuals in the data set that ran the NYC Marathon for every included year. The pace shown is median-adjusted with the individual's race year, sex, and age group. See the individual runners shown below. It is challenging to recognize a consistent trend across all race participants as some increase in speed, others are decreasing in speed, and several have a clear arch in their pace trend. As the last part of this question, I used a quadratic model with age and age squared, and random effects for each individual in an attempt to get a more precise model.

Figure 4: Individual Female Runners' Race Paces During 13 NYC Marathons

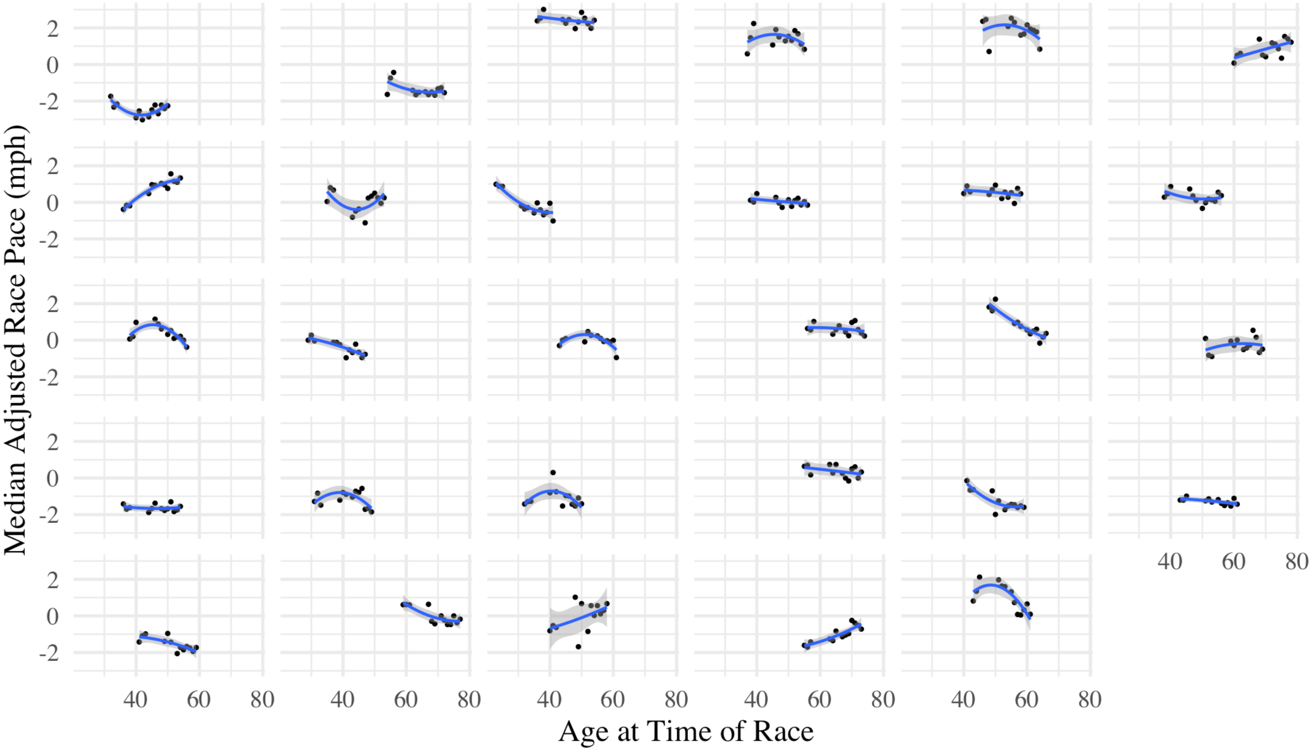
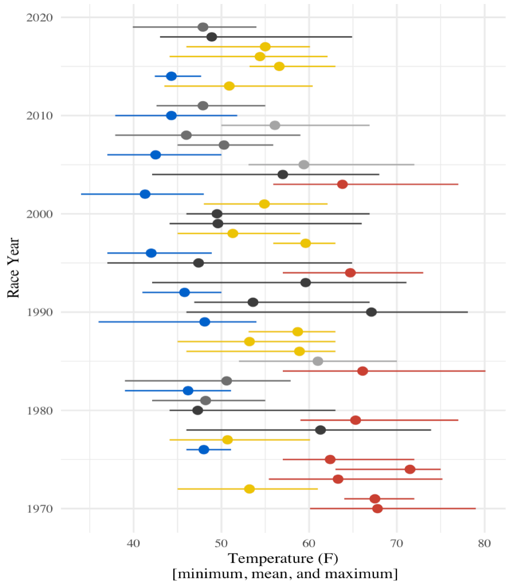


Figure 5:
NYC Marathon Race Day Temperature History

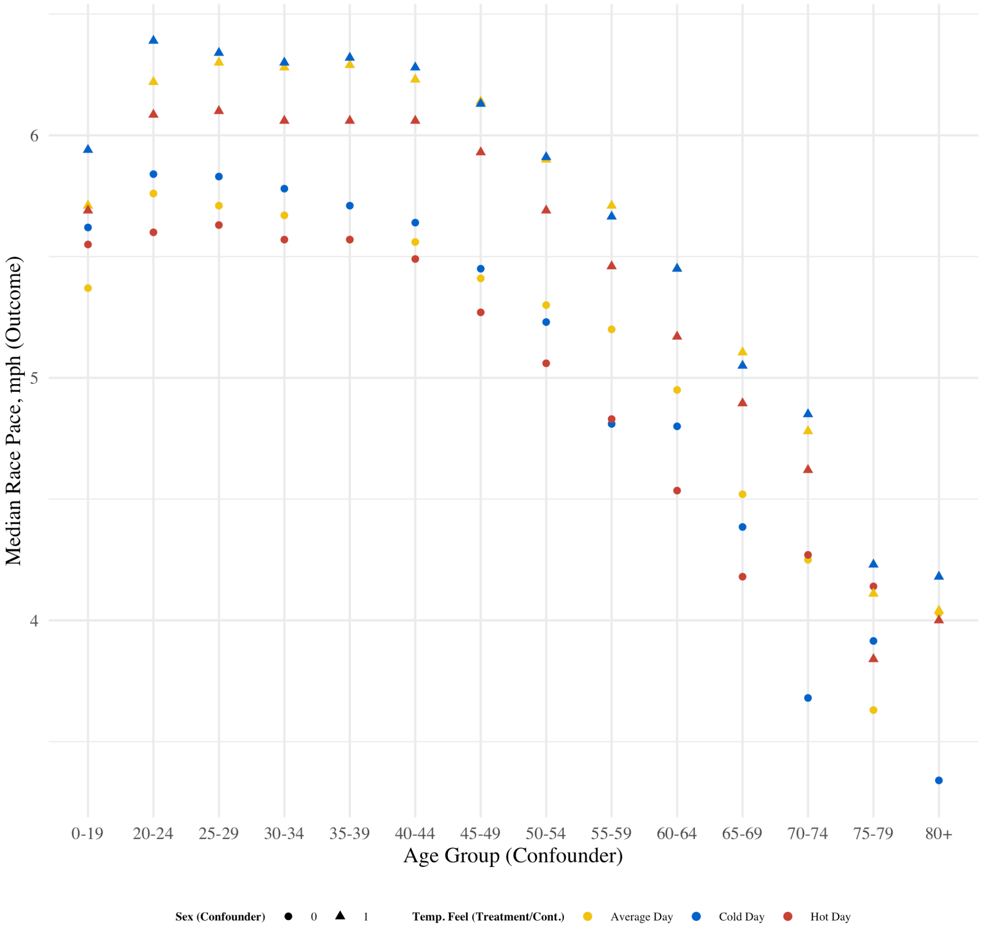


For question three, since I have been learning about propensity score matching, I used 1-1 nearest neighbor and full-model propensity score matching for two different situations. The outcome variable was the runner's speed in miles per hour; the treatment variables were whether it was a cold day versus an average temperature day or a hot day versus an average temperature day.

The cold day was defined as race days that temperatures did not rise above 54°F. Hot days were defined as race days whose mean temperatures were above 60°F, but whose minimum temperatures did not fall below 55°F. Average days were defined as race days whose mean temperature was between 50°F and 60°F, and a minimum temperature that did not fall below 44°F and maximum did not rise above 63°F. I looked specifically at finishers from 2001, 2002, and 2003 to have consistent running trends in the same timeframe. I also added 2017 runners to the control group since 2001 may have biased finishing times. After the tragedies in New York City on September 11th, 2001, the marathon had a reduced number of runners that year. Also, runners' finishing times may have been impacted by the camaraderie that existed after the September 11th attacks in New York City and in the United States. After matching, I ran linear models with the treatment variable, corral number, age, and sex.

There were apparent pace differences between hot, cold, and average days that motivated me to look into potentials causal effects of hot or cold days on runners' pace. See Figure 6.

Figure 6: Median NYC Marathon Speeds by Age, Sex, and



Challenges & Limits

The initial challenge with this project was scraping the data from the New York Road Runner website. There are around 75,000 or fewer runner numbers for each race year. Each runner's data is on a separate webpage. The scraping process took a lot of diligence and intentionality. The organization also routinely closes its results website for maintenance.

The primary limit to the project was knowing minimal information about each runner. It would have been interesting to know when each runner started their running career, their birthdate, height, weight, and a personal best marathon time.

Other constraints surround the method accuracy of connecting the repeat runners. Some runners may have gotten married and changed their names between races. Their birthday may have been in the first week of November, so their birth year would have been miscalculated. Additionally, runners with the same birth year and same name may have been connected, overlooked in the editing, and are not the same person.

Also, there are much more precise methods for detecting impacts of temperature on runners' pace. It would be possible to use models that take sex and age into account. There are likely differences in how much temperature changes impact various kinds of runners.

If I had a few more months to work on this project, I would scrape the NYC Marathon data for all 50 years of the race. I would spend more time combing through the connected repeat runners to verify that they are the same person. Moreover, I would dig deeper into the questions I first posed and others around mile pacing and placement.

Results

For question one, I hypothesized that fast NYC Marathoners were getting faster, on average, average people were staying the same speeds, on average, and slower runners were getting slower. However, as you can see in Figures 5 and 6, most median marathon times stayed consistent from year to year within their sex, age groups, and speed quartiles. For the colors in the plots, green represents the top 25% of times for each age group within sex for each year. Yellow represents the middle 50% of finishers for each each group, within sex. And red represents the slowest 25% of runner. Some significant trends (p

< .05) are fast males between 55 - 74 increased in their median race pace over the years. Males over 70 in the slowest quartile decreased in their median race pace over the years. Meanwhile, the fastest quartile of females under 20 and over 80 were both faster in their median times over the years. Nearly all females between 55 - 69 also showed increasing speed. As far as females slowing down, the slowest quartile of females between ages 20 to 44 and 35 to 44 showed decreasing median times. I thought the decline in median race times for females might be due to the rising popularity of marathon running (see Figure 3 for growth numbers). However, when you see Figure 3, most race participant growth is happening in other age groups. Figures 7 and 8 depict the median pace trends over the years. The solid and dotted lines represent significant relationships (p < .05).

Figure 7: Male NYC Marathoners Median Pace Trends Over the Years

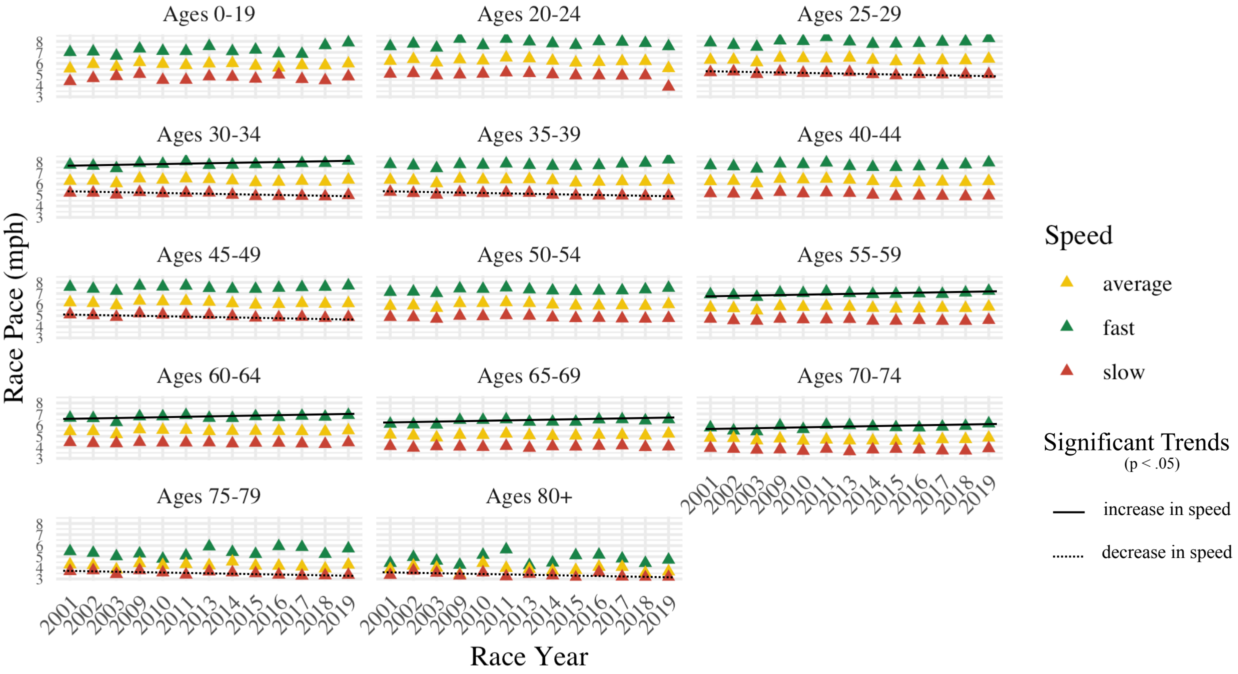
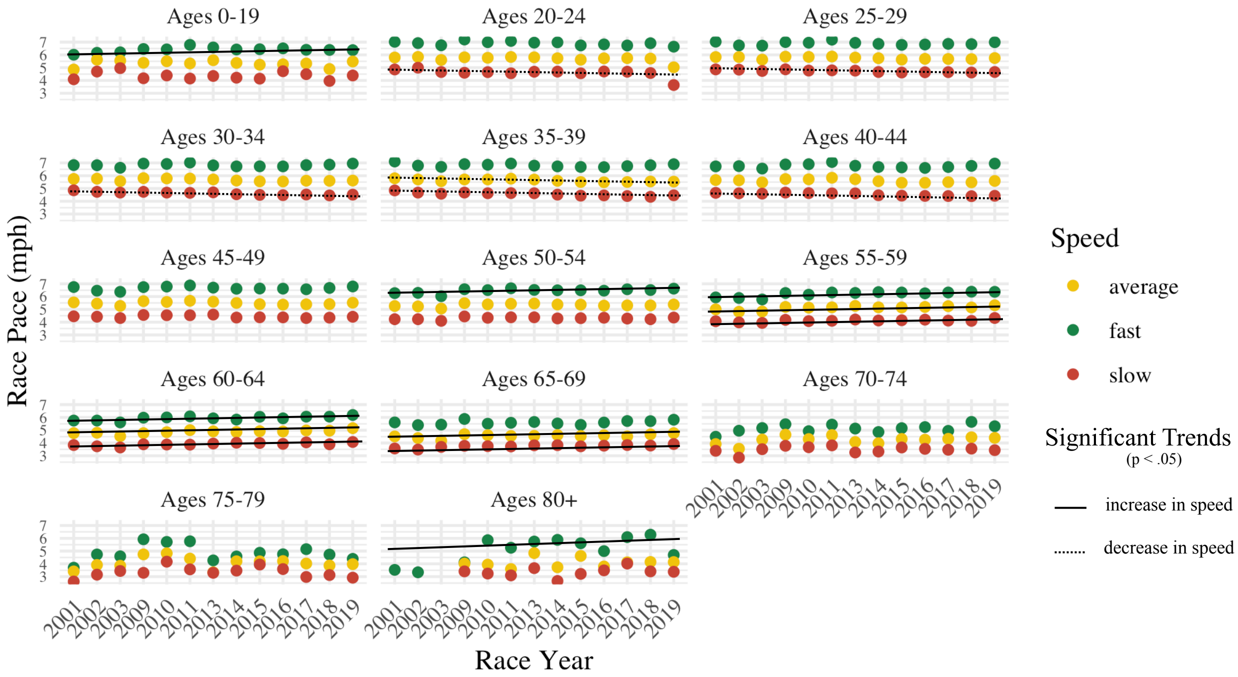
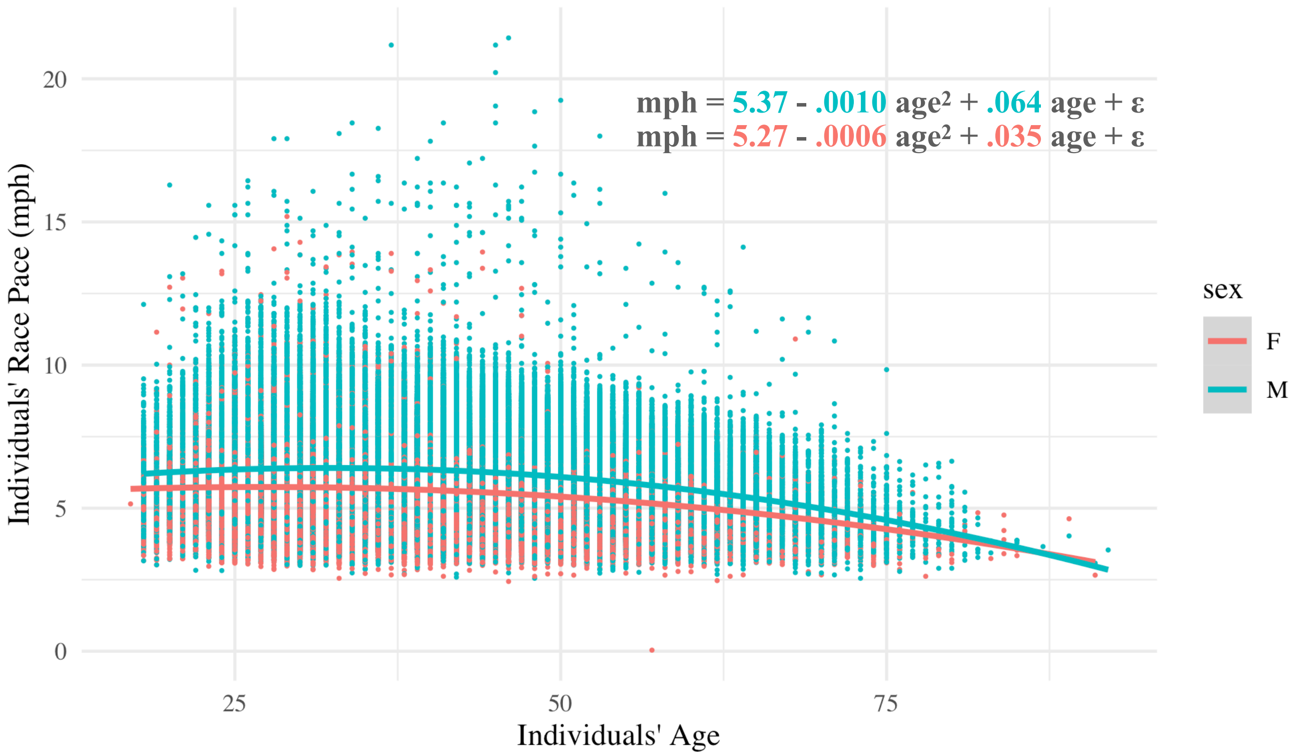


Figure 8: Female NYC Marathoners Median Pace Trends Over the Years



Regarding question two, to find the peak age, I started with a quadratic model using age squared and age with all of the runners in the data set, including one year for each person. Both variables were significant (p < .05). The simple models, noted in Figure 9, demonstrate the peak age for a male's fastest pace is 32 years old, and the peak age for a female's fastest pace is 29 years old. However, as you can tell by the spread of the runners' paces in Figure 9, the R^2 is low. The male and female models have an R^2 of .04.

Figure 9: NYC Marathoners Age Compared with Speed



Next, I ran stepwise functions and experimented with numerous models with all the runners in the data set, included one time. I hoped to find a simple model using sex, age, age squared, and one other predictor variable. This model was chosen after exploration and AIC and BIC tests. It includes sex, age, age squared, New York Road Runners' age group, and daily mean temperature on race day. All coefficients were significant ($p < .05$). Age group was equal to the sum of two and age minus 20, divided by five. The R^2 is better than the previously reported model since it included temperature. The R^2 was .1.

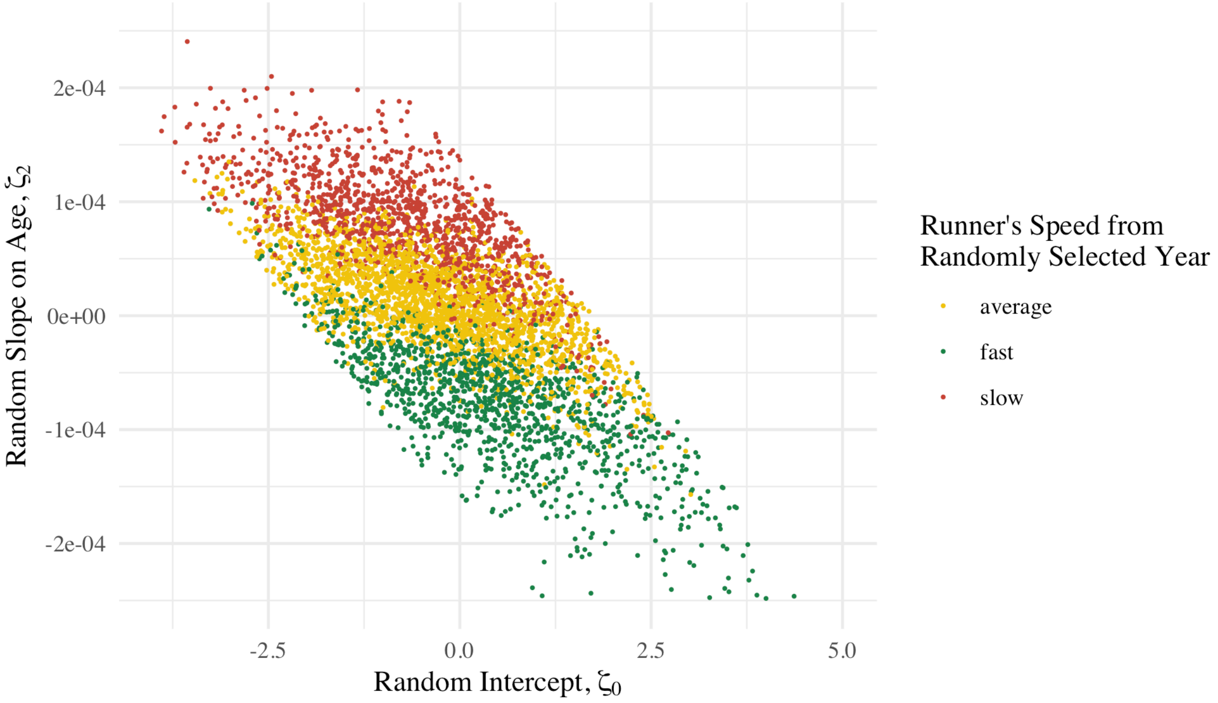
$$\text{mph} = 5.585 - .00086\ \text{age}^2 + .0301\ \text{age} + .1151\ \text{age group} + .6734\ \text{sex} - .0079\ \text{daily mean temperature} + \varepsilon$$

In search of a better model, lastly, I produced a random-effects model on the 5,979 repeat runners who ran five or more marathons within these years. By exploring different variables and anova tests, this was the simplest, most effective model that I found. Surprisingly, the model with sex included was less accurate.

$$\text{mean adjusted mph} = -\ 1.437 + \zeta_0 + (-\ .0005463 + \zeta_1)\ \text{age}^2 + (.05947 + \zeta_2)\ \text{age} + \varepsilon$$

The random-effects unsurprisingly correlated to each other, coinciding with the individual's speed category during the marathon. I chose a race speed at random for each repeat runner for the plot and colored the points with that speed. For example, they may have run slowly one year and at an average pace the next year, but the color in this plot shows one of those values. Random intercepts were smaller for faster runners, and the random slope on age was smaller for slower runners. See Figure 10.

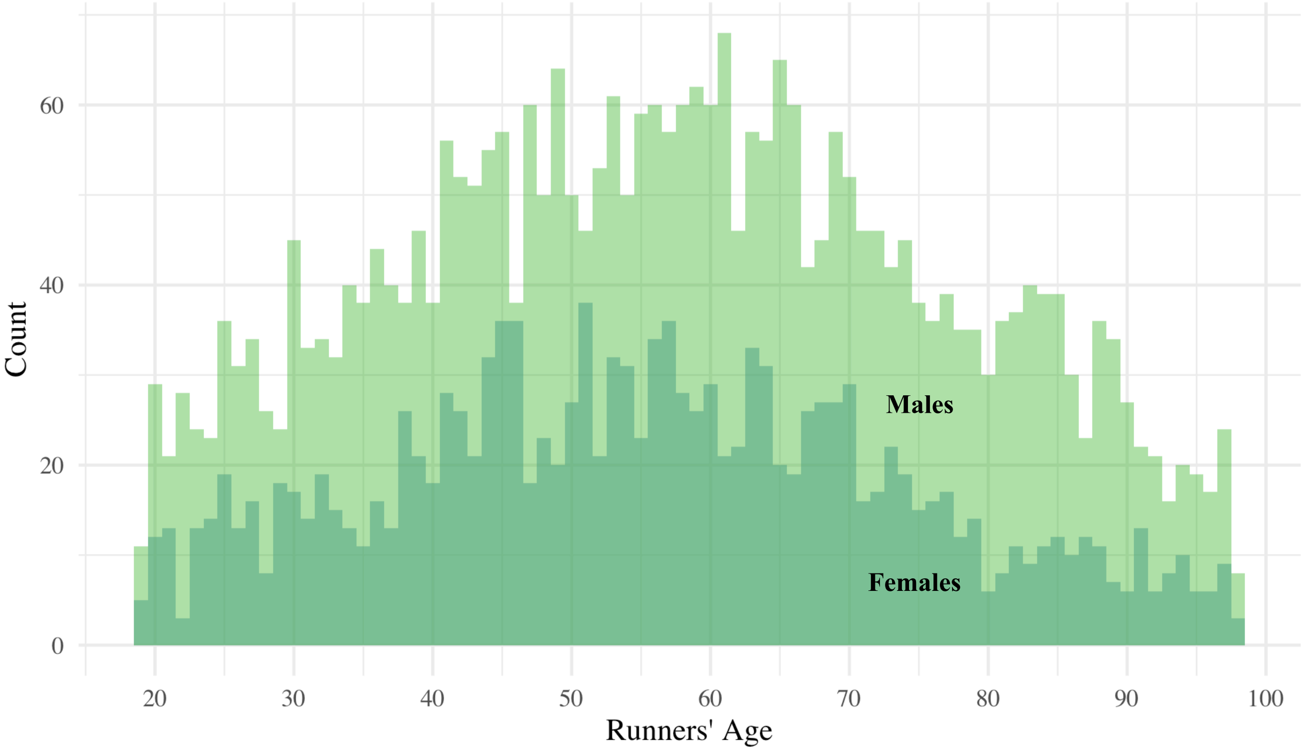
Figure 10: Repeat NYC Marathon Repeat Runners' Individual BLUPS



Lastly, with the random-effects model I found, it was nearly impossible to predict the runner's likely peak age for pace. There are a variety of ages that runners peak in their NYC Marathon pace. People start racing at different times in their life and have various approaches to training throughout the years. As you can see in Figure 11, for some runners, the model predicts they will peak at age 98. Many people are shown as peaking at age 84 and older. However, there were only seven people age 84 and over in the entire repeat runner data, the oldest being 89. So, the model made inferences about people's capabilities without understanding the realities of human aging. It also estimated that 20% of runners would peak below age 18 or above age 100. These values were not included in the graph.

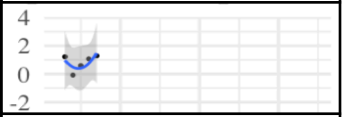
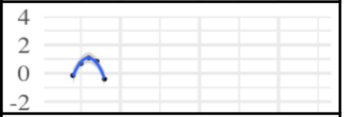
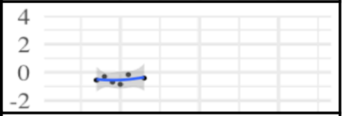
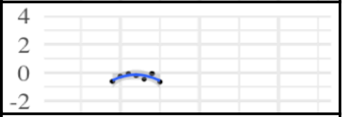
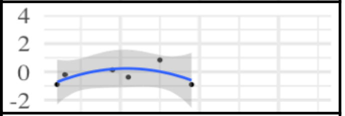
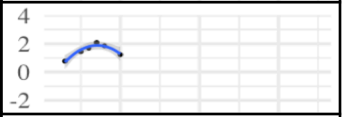
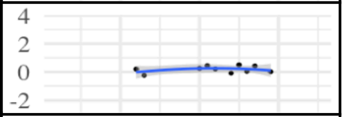
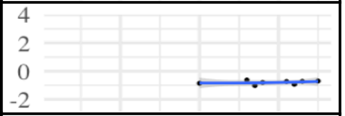
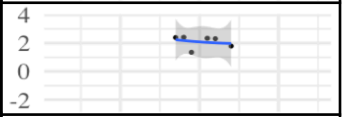
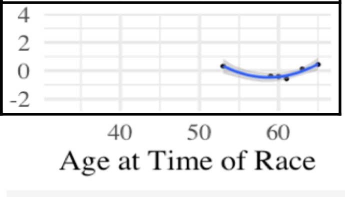
The model estimated that the mode for female peak-pace age was around age 51, for participants running the NYC Marathon. The median was age 54 for females. The model for males, who are repeat marathoners, was age 61, for participants running the NYC Marathon. The median was age 56 for males. Nevertheless, the model is undoubtedly limited since there are so many predictions outside of reasonable age values.

Figure 11: Repeat NYC Marathon Male and Female Peak Age for Fastest Pace



In Figure 12, you can see details about ten specific repeat runners and the age that the model predicted they would peak in their race pace. As you can see in the last column, there are noticeably different paces the runners ran over their NYC Marathons years. The model is making a prediction based on that data.

Figure 12: Ten Repeat NYC Marathoner's & Predicted Age for Peak Racing Time

Name	Sex	City	Country	Birth Year	Speed	Peak Age	Median-Adjusted Race Paces
Antonio Ventimiglia	M	Milano	ITA	1970	slow	56.417	
Aron Kressel	M	New York, NY	USA	1959	average	58.755	
Brian Halbreich	M	Woodside, NY	USA	1976	average	58.583	
Curtis Arluck	M	New York, NY	USA	1953	slow	52.29	
David Stewart	M	New York, NY	USA	1962	fast	59.525	
Isabelle Bricout	F	Elizabeth, NJ	USA	1974	average	50.254	
Melodie Fratt	F	New York, NY	USA	1978	fast	58.656	
Michaelyn Fesler	F	New York, NY	USA	1981	fast	45.342	
Stephanie Tang	F	Brooklyn, NY	USA	1982	fast	55.851	
Ted Ahn	M	Eastchester, NY	USA	1950	average	54.51	

For the question around weather's impact on marathon speeds, the distribution of the confounders and propensity scores overlapped in the matched data. The variables had difference of means under .01 and variance ratios near one. See Figures 13 and 14.

Figure 13: Balance Table for Propensity Score Matching on NYC Marathoners (Cold with Average Temperatures)

	Cold Day, 1-1 Nearest Neighbor				**Cold Day, Full-Matching on 5,000 Runners**			
Variables	*Unmatched Difference in Means*	*Unmatched Variance Ratio*	*Matched Difference in Means*	*Matched Variance Ratio*	*Unmatched Difference in Means*	*Unmatched Variance Ratio*	*Matched Difference in Means*	*Matched Variance Ratio*
sex	0.0555	NA	0.0014	NA	.-0701	NA	0.0074	NA
age	-0.1360	0.9199	0.0025	1.0177	-0.1074	0.9659	0.0227	1.0386
corral	-0.0003	1.0004	-0.0017	1.0014	-0.0059	1.0160	0.0076	1.0417

Figure 14: Balance Table for Propensity Score Matching on NYC Marathoners (Hot with Average Temperatures)

	Hot Day, 1-1 Nearest Neighbor				**Hot Day, Full-Matching on 5,000 Runners**			
Variables	*Unmatched Difference in Means*	*Unmatched Variance Ratio*	*Matched Difference in Means*	*Matched Variance Ratio*	*Unmatched Difference in Means*	*Unmatched Variance Ratio*	*Matched Difference in Means*	*Matched Variance Ratio*
sex	0.0387	NA	0	NA	0.0213	NA	-0.001	NA
age	-0.1367	0.9043	0	1.0051	-0.1672	0.9006	0.0053	1.0095
corral	0.0011	0.9988	0.0001	1.0008	0.0222	0.9727	0.0260	0.9813

The conclusion for a cold day was the estimated average treatment effect on the treated using 1-1 nearest neighbor matching was .027, with all p-values for each variable less than .05. The average treatment effect on the treated using full matching with a sample of 6,000 runners with the cold and average days was .038. This implies, holding all other variables constant, as a result of a cold first Sunday in November, runners finish the New York City marathon on average of 30 seconds or 55 seconds faster in their overall time than those running on an average temperature marathon day.

For a hot day, the estimated average treatment effect on the treated using 1-1 nearest neighbor matching was -.148. The average treatment effect on the treated using full matching with a sample of 6,000 runners with the hot and average days was -.150. This implies, holding all other variables constant, as a result of a hot first Sunday in November, runners finish the New York City marathon on average 9.3 minutes slower in their overall time than those running on an average temperature marathon day. The ATT was the same minute value for both the 1-1 nearest neighbor and full matching with finishing time and race pace as the outcome. With all of these models, all variables had p-values less than .05.

Conclusion

After researching the New York City Marathon, three things have become more apparent. First, the thought that people are getting faster every year was confirmed regarding fast people who run the NYC Marathon within certain age groups, specifically the top 25% of females between 50 and 64 and the top 25% of males between 55 and 74. Second, the statistical models could not conclude the marathon pacing peak age for all runners. However, the models' medians for peak age were age 54 for females and 56 for males. Overarchingly, peak age may also be closely related to other variables like the number of years as a competitive runner and training rigor, which were not in the data. Thirdly, race day temperature impacts running pace during the NYC Marathon, holding all other variables constant. I love being a runner and diving into numbers to understand more about this sport that I'm passionate about. I look forward to doing additional research on this topic in the future.

Works Cited

Hosier, G. “How Much Does Heat Slow Your Race Pace?” *Podium Runner.* 2 Aug. 2019, https://www.podiumrunner.com/training/how-much-does-heat-slow-your-race-pace/. Accessed 20 Dec. 2020.

Minsberg, T, Quealy, K. “Why are American Women Running Faster Than Ever? We Asked Them - Hundreds of Them”. New York Times, 28 Feb. 2020, https://www.nytimes.com/interactive/2020/02/28/sports/womens-olympic-marathon-trials.html. Accessed 20 Dec. 2020.

New York Road Runners, 2020. Retrieved Nov. 2020 (https://results.nyrr.org/home).

Zavorsky GS, Tomko KA, Smoliga JM. “Declines in marathon performance: Sex differences in elite and recreational athletes”. PLoS ONE, 12(2): e0172121, 10 Feb. 2017, https://doi.org/10.1371/journal.pone.0172121. Accessed 20 Dec. 2020.