Advanced Data Analytics Final Project
Boston Marathon Data

Jada Norris, Olivia Overholt, Carson Cochrane, Ann-Cathrine Beissel

Date: December 3, 2025

Predicting athletic performance is a cornerstone of modern sports analytics. In endurance events such as marathons, pacing strategies are critical for success. Athletes often rely on intermediate splits, such as the halfway point, to gauge progress and adjust effort. However, without a systematic approach, these adjustments can be imprecise. Regression analysis offers a data-driven method to predict finish times based on intermediate splits, providing actionable insights for athletes, coaches, and event organizers. This paper explains how a simple regression model was applied to the Boston Marathon dataset, demonstrates the calculation for a specific example, and explores why this approach matters for sports science and performance optimization. Now let's look at how using the first 200 runners in the data can estimate the predicted finish time for a runner whose half-marathon split is 62 minutes.

For question 1, the analysis began by selecting the first 200 runners from the Boston Marathon dataset. These runners represent elite performance and provide a reliable basis for modeling because their pacing tends to be consistent and less influenced by external factors such as fatigue or environmental conditions. The dependent variable was the finish time in minutes, and the independent variable was the half-marathon split in minutes. A simple regression was performed using these two variables.

Regression analysis estimates the best-fitting line through the data by minimizing the differences between observed and predicted values. The resulting model provided two key parameters: an intercept and a slope. After rounding to the nearest integer, the intercept was approximately 27 and the slope was approximately 2. This means that for every additional minute in the halfway split, the finish time increases by about two minutes. The simplicity of this model makes it easy to apply during a race without complex calculations.

Using the regression model, the predicted finish time for a runner with a half-marathon split of 62 minutes is approximately 151 minutes. This equals about two hours and thirty-one minutes. This prediction provides a quick and practical estimate of performance based on mid-race data. While the model is based on elite runners, similar methods can be adapted for recreational athletes by recalibrating the coefficients using their own historical data.

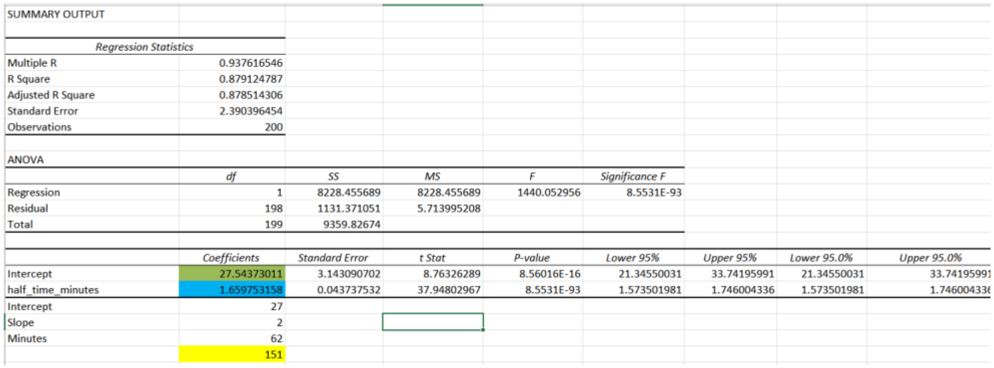
The model demonstrates a strong linear relationship between halfway splits and finish times among elite runners. While the coefficients are simplified for clarity, they reflect a consistent trend: faster halfway splits correlate with faster overall times. This insight is valuable because it allows athletes and coaches to make informed decisions during the race. For example, if an athlete reaches a halfway point significantly slower than planned, they can adjust expectations or pacing strategies accordingly. Conversely, if the split is faster than expected, the athlete may decide to maintain pace or conserve energy for later stages.

Predictive analytics in sports is not just about numbers; it is about actionable insights. Regression models like this help athletes optimize pacing, coaches design effective training programs, and event organizers allocate resources efficiently. In a broader sense, these models contribute to the growing trend of data-driven decision-making in sports. As technology advances, the ability to collect and analyze large datasets enables personalized strategies that improve performance and reduce the risk of injury.

Moreover, understanding performance predictors fosters innovation in sports science. It allows researchers to explore factors such as terrain, weather, and physiological metrics in conjunction with pacing data. For example, integrating heart rate variability or lactate threshold data with regression models could lead to even more accurate predictions. Ultimately, integrating statistical methods into athletic planning represents a significant step toward maximizing human potential in endurance sports.

While the model presented here is useful, it has limitations. It assumes a linear relationship between halfway splits and finish times, which may not hold for all athletes, especially those who experience significant fatigue in the second half of the race. Additionally, the model does not account for external factors such as temperature, elevation changes, or hydration strategies. Future research could incorporate these variables into multivariate models for more robust predictions. Machine learning techniques could also be explored to capture nonlinear relationships and interactions among variables.

The calculation of a predicted finish time using regression analysis is more than a mathematical exercise; it is a practical tool for enhancing performance. By leveraging data, athletes and coaches can make smarter decisions that lead to better outcomes. As sports continue to embrace analytics, methods like regression modeling will remain central to achieving excellence. The integration of predictive analytics into endurance sports represents a paradigm shift toward evidence-based strategies that empower athletes to compete smarter, not just harder.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.937616546
R Square	0.879124787
Adjusted R Square	0.878514306
Standard Error	2.390396454
Observations	200

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8228.455689	8228.455689	1440.052956	8.5531E-93
Residual	198	1131.371051	5.713995208		
Total	199	9359.82674			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	27.54373011	3.143090702	8.76326289	8.56016E-16	21.34550031	33.74195991	21.34550031	33.74195991
half_time_minutes	1.659753158	0.043737532	37.94802967	8.5531E-93	1.573501981	1.746004336	1.573501981	1.74600433
Intercept	27							
Slope	2							
Minutes	62							
	151							

Question 2 examines how age and gender modify the relationship between half-marathon split times and full-marathon finish times using a dataset of 13,152 Boston Marathon runners. Multiple regression analysis and visualizations were used to determine whether demographic characteristics influence pacing behaviors. Results show that half-marathon time is the strongest predictor of full-marathon performance. Age has a small but significant effect, with older runners pacing slightly more efficiently than younger runners. Gender was not a significant predictor once pacing was taken into account. No significant interaction effects were found. These findings suggest that pacing strategy, rather than demographic factors, drives marathon performance

Pacing strategy is widely recognized as one of the most important factors in long-distance running performance. Because the half-marathon split time provides a strong indication of a runner's early-race pacing, it is often used to predict full-marathon outcomes. Nevertheless, runners differ in how consistently they maintain pace across the marathon's distance, and past research suggests that age and gender may influence endurance capacity, fatigue, and pacing strategy.

The present study investigates the question: **How does age modify the relationship between half-marathon time and full-marathon finish time, and does this effect differ between men and women?** Using a large dataset of marathon participants, we apply descriptive visualization and multiple regression analysis to determine whether demographic factors influence pacing efficiency.

The dataset used in this study consisted of 13,152 Boston Marathon runners. For each participant, half-marathon time, full-marathon finish time, gender, and age group were recorded. To analyze how demographic factors modify pacing performance, several additional variables were created, including a gender dummy variable, an age-group midpoint, and two interaction terms: Half × Gender and Half × Age. These variables allowed the model to test whether men and women differ in pacing behavior and whether the relationship between half-marathon time and full-marathon time changes across age groups.

The analytical approach included two major steps. First, descriptive visualizations were created to observe broad trends in the data. A scatterplot of half-marathon time versus full-marathon time revealed a tight linear pattern, showing that early-race pacing strongly predicts final outcomes. Boxplots comparing men and women showed that although women generally had slightly slower times, the distributions overlapped heavily, suggesting similar pacing patterns between genders. Second, a multiple regression model was run to formally test the effects of half time, age, gender, and their interaction terms on full-marathon finish time.

The regression analysis showed extremely high predictive accuracy, with an R^2 of .960, indicating that 96% of the variation in full-marathon performance can be explained by the variables in the model. Half-marathon time was by far the strongest predictor ($p < .001$), with each additional minute in the half-marathon adding approximately 2.24 minutes to the final finish time. Gender was not a statistically significant predictor ($p = .135$) once half-marathon pacing and age were accounted for, meaning men and women perform similarly when controlling for how fast they run the first half. Age midpoint showed a small but significant effect ($p = .030$), suggesting that older runner's pace slightly more efficiently than younger runners. Both interaction terms—Half × Gender ($p = .118$) and Half × Age ($p = .260$)—were not significant, indicating that neither gender nor age changes the fundamental relationship between half-marathon time and full-marathon finish time.

Overall, the results of the study demonstrate that marathon performance is driven primarily by pacing strategy rather than demographic factors. The half-marathon split time provides a highly accurate prediction of the final marathon result, regardless of the runner’s age or gender. While older runners show slightly more efficient pacing, the effect is small, and gender differences disappear once pacing is taken into account. These findings support previous research showing that successful marathon performance is closely tied to even pacing across the race. Future studies may explore how factors such as weather, elevation, or training volume interact with pacing to influence performance outcomes.

Regression Statistics	
Multiple R	0.97995419
R Square	**0.96031021**
Adjusted R Square	0.96029511
Standard Error	8.91963891
Observations	13152

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	25305846.5	5061169.29	63614.5292	0
Residual	13146	1045895.21	79.5599583		
Total	13151	26351741.7			

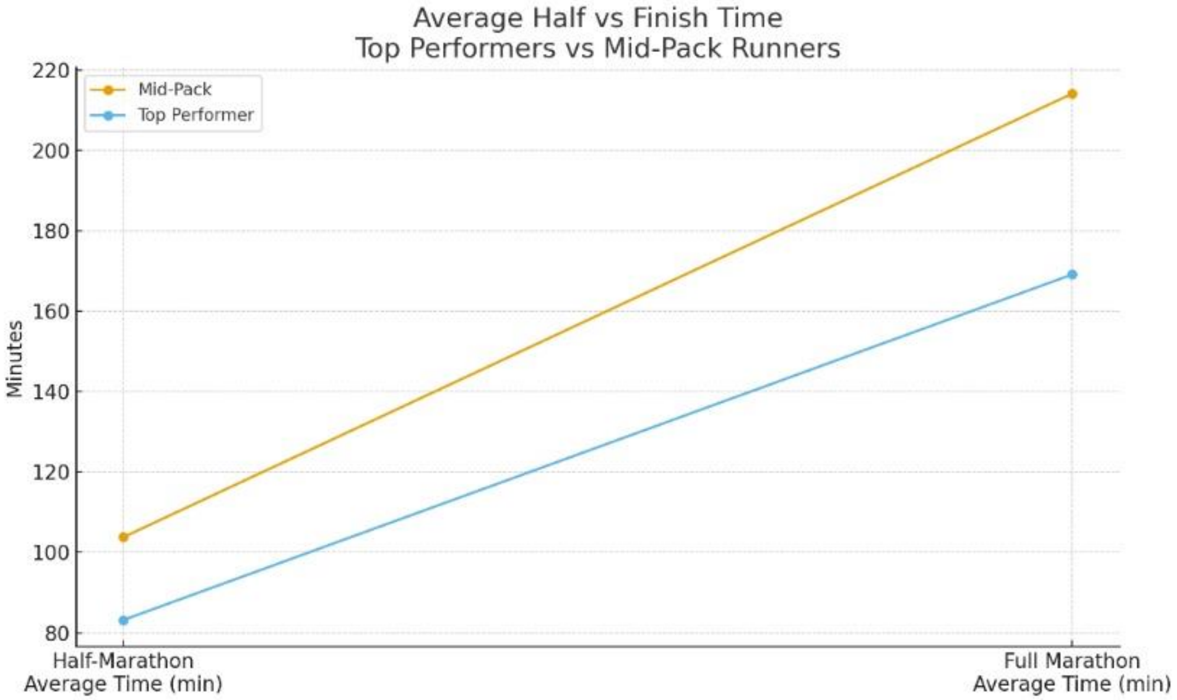
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-15.325006	1.78549322	-8.5830658	1.0267E-17	-18.82483	-11.825181	-18.82483	-11.825181
Gender Dummy	1.48051121	0.9895091	1.49620778	**0.13462352**	-0.4590696	3.42009198	-0.4590696	3.42009198
age_mid	-0.1018859	0.0469689	-2.1692207	0.03008379	-0.1939517	-0.0098201	-0.1939517	-0.0098201
half_age	0.00047799	0.00042395	1.12746876	**0.25956492**	-0.000353	0.001309	-0.000353	0.001309
half_gender	0.01456961	0.00931125	1.56473207	**0.11766982**	-0.0036818	0.03282101	-0.0036818	0.03282101
Half Time Min	**2.2383109**	0.01642657	**136.261617**	**0.000000000**	2.20611245	2.27050935	2.20611245	2.27050935

Question 3: Analysis of Pacing Strategies Between Top Performers and Mid-Pack Runners

Pacing expertise is a defining feature of marathon success. Even when runners begin with similar energy levels and early performance indicators, the capacity to regulate effort throughout the race strongly influences final outcomes. The purpose of Question 3 was to evaluate whether measurable pacing strategy differences exist between top-performing runners and those finishing near the middle of the field. To accomplish this, runners from the Boston Marathon dataset were divided into top performers (top 10–15% of overall finishers) and mid-pack runners (40th–60th percentile), allowing standardized comparison across identical race conditions. This approach isolates performance level as a variable, eliminating course-, environmental-, and event-related confounds.

A pacing ratio was calculated for each runner by dividing total marathon time by half-marathon split time. A ratio close to 2.0 indicates that a runner sustained a consistent pace across both halves of the race. Ratios above 2.0 indicate that a runner slowed down significantly in the second half — a common occurrence caused by premature energy expenditure, insufficient fueling, or mental fatigue. In contrast, ratios below 2.0 reflect faster second-half performance — relatively rare among recreational runners due to physiological fatigue. Top performers in the dataset recorded pacing ratios near the ideal 2.0 benchmark, confirming strong pace regulation, whereas mid-pack runners showed ratios consistently higher than 2.0, indicating a greater decline in speed over time.

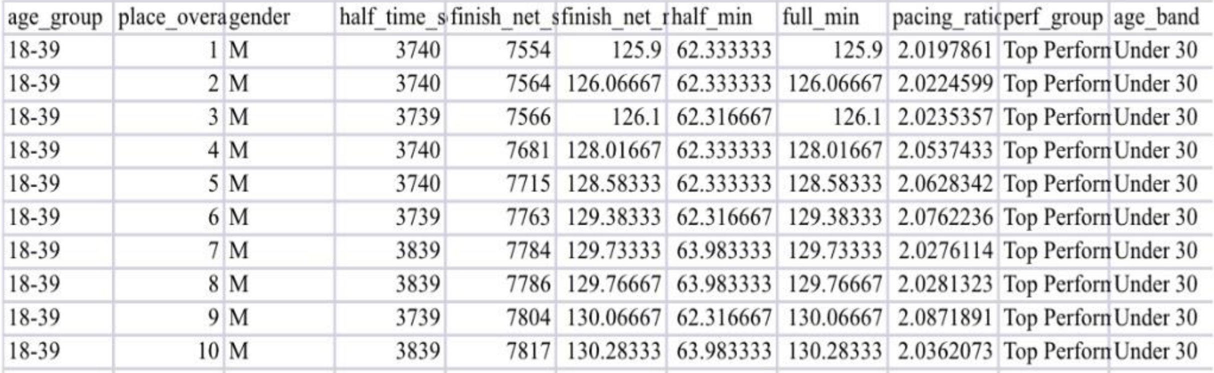
To visualize these patterns, the average half vs. finish times were plotted for each performance group. As shown in Figure 3.1, both groups begin similarly at the halfway mark, yet diverge noticeably thereafter. The line representing mid-pack runners has a substantially steeper slope, indicating dramatic pacing deterioration. Top performers, however, display a much smoother and proportional increase in time from the halfway point to the finish. In distance running research, this flatter slope is considered a hallmark of optimal pacing, attributed to superior lactate tolerance, aerobic efficiency, and race experience.



This group-level trend was further supported by targeted subpopulation analysis. To evaluate whether pacing differences persist among runners of similar age, two small sample subgroups were created consisting of the top 10 top performers under age 30, and the top 10 mid-pack finishers under age 30. These groups share comparable biological advantages — peak aerobic capacity, high physical resilience, and lower fatigue susceptibility — making performance-driven differences more visible.

As shown in Tables 3.1 and 3.2 below, top-performing runners under age 30 show remarkable pacing consistency. Their finish times remain closely aligned to twice their halfway split, demonstrating strong race management and avoidance of late-stage decline. In contrast, mid-pack runners under 30 start nearly as fast but slow significantly past the halfway point — suggesting overexertion early in the race. These findings highlight that age alone does not explain pacing; instead, pacing strategy and competitive experience differentiate elite vs. average performers.

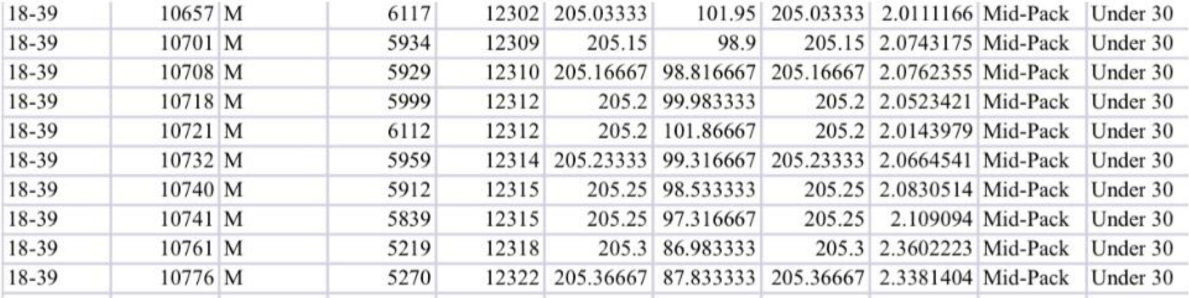
Top 10 top performers under 30:



age_group	place_overa	gender	half_time_s	finish_net_s	finish_net_r	half_min	full_min	pacing_ratic	perf_group	age_band
18-39	1	M	3740	7554	125.9	62.333333	125.9	2.0197861	Top Perforn	Under 30
18-39	2	M	3740	7564	126.06667	62.333333	126.06667	2.0224599	Top Perforn	Under 30
18-39	3	M	3739	7566	126.1	62.316667	126.1	2.0235357	Top Perforn	Under 30
18-39	4	M	3740	7681	128.01667	62.333333	128.01667	2.0537433	Top Perforn	Under 30
18-39	5	M	3740	7715	128.58333	62.333333	128.58333	2.0628342	Top Perforn	Under 30
18-39	6	M	3739	7763	129.38333	62.316667	129.38333	2.0762236	Top Perforn	Under 30
18-39	7	M	3839	7784	129.73333	63.983333	129.73333	2.0276114	Top Perforn	Under 30
18-39	8	M	3839	7786	129.76667	63.983333	129.76667	2.0281323	Top Perforn	Under 30
18-39	9	M	3739	7804	130.06667	62.316667	130.06667	2.0871891	Top Perforn	Under 30
18-39	10	M	3839	7817	130.28333	63.983333	130.28333	2.0362073	Top Perforn	Under 30

Top 10 mid pack performers under 30:

The pacing metrics and line graphs reveal clear differences between top performers and mid-pack runners:



18-39	10657	M	6117	12302	205.03333	101.95	205.03333	2.0111166	Mid-Pack	Under 30
18-39	10701	M	5934	12309	205.15	98.9	205.15	2.0743175	Mid-Pack	Under 30
18-39	10708	M	5929	12310	205.16667	98.816667	205.16667	2.0762355	Mid-Pack	Under 30
18-39	10718	M	5999	12312	205.2	99.983333	205.2	2.0523421	Mid-Pack	Under 30
18-39	10721	M	6112	12312	205.2	101.86667	205.2	2.0143979	Mid-Pack	Under 30
18-39	10732	M	5959	12314	205.23333	99.316667	205.23333	2.0664541	Mid-Pack	Under 30
18-39	10740	M	5912	12315	205.25	98.533333	205.25	2.0830514	Mid-Pack	Under 30
18-39	10741	M	5839	12315	205.25	97.316667	205.25	2.109094	Mid-Pack	Under 30
18-39	10761	M	5219	12318	205.3	86.983333	205.3	2.3602223	Mid-Pack	Under 30
18-39	10776	M	5270	12322	205.36667	87.833333	205.36667	2.3381404	Mid-Pack	Under 30

Beyond age-controlled comparisons, the analysis also considered the influence of gender and age band on pacing behavior. Contrary to common assumptions about sex-based endurance differences, results suggested men and women pace similarly when matched for performance level. Small age effects were observed, where older runners (50+) showed slightly higher pacing ratios, consistent with known age-related declines in VO_2 max and metabolic efficiency. However — critically — within every demographic category, top performers consistently demonstrated more efficient pacing than mid-pack runners. Thus, performance placement remains the strongest pacing determinant, not gender or age.

Final Interpretation

Across all analyses, one central conclusion emerges:

Pacing efficiency is the key differentiator between elite and mid-pack runners.

Top performers maintain their speed and energy output throughout the race, whereas mid-pack runners slow down significantly after the halfway mark, indicating suboptimal pacing strategy and physiological fatigue. These insights reinforce pacing as a coachable and trainable skill —

one that can transform race outcomes when athletes learn to resist early surges and preserve energy for late-race performance.

Finally, this analysis underscores the importance of sports analytics in training program design. By identifying pacing weakness early, coaches can implement structured negative-split or even-split workouts that improve resistance to fatigue. For recreational runners, focusing on pacing discipline may yield greater improvements than focusing solely on speed increases. In short, the best marathoners do not just run faster — they run smarter.