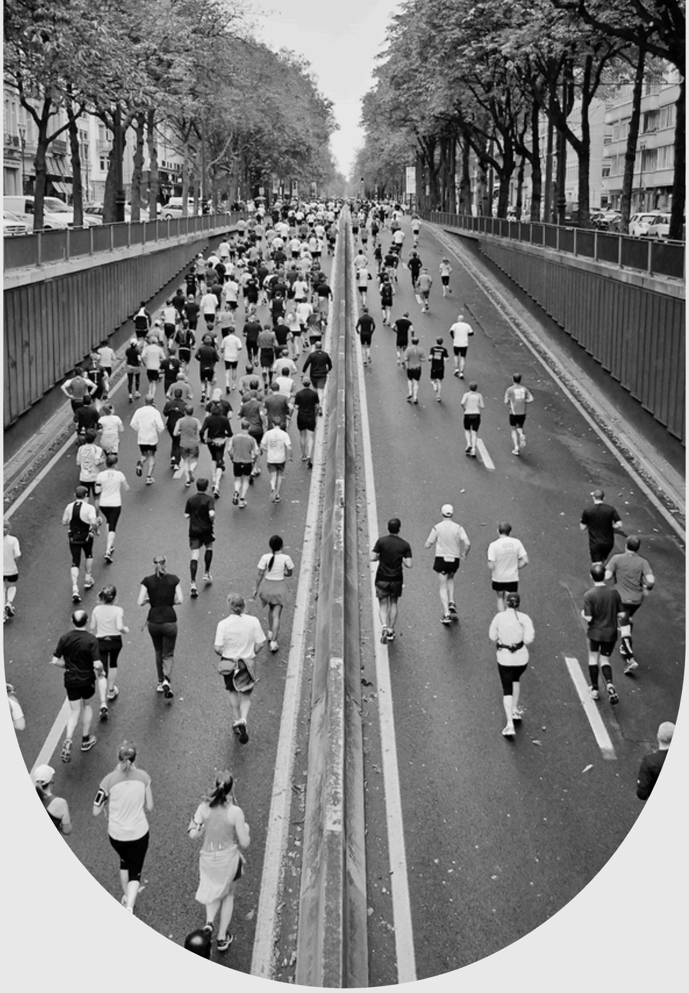
BOOTCAMP
DATA
SCIENCE



# EDA: MARATHON TIME PREDICTORS

## PREPARED BY KATIA BARSUK

# Index

# 01. Introduction

In this project we're going to analyze if there are any factors that can contribute to success when running a marathon. We'll be using a Kaggle dataset.

Source:
Andrea Girardi, Marathon time Predictions
Predict Marathon Results from Athletes Open Data Sources
https://www.kaggle.com/datasets/girardi69/marathon-time-predictions/data

## 01.a Formal context

The dataset consists of 10 columns and 87 entries. It includes information on training history, pacing strategies, and marathon results.

The dataset contains data from Prague'17 marathon, with the following key columns:

- MarathonTime: Total time to complete the marathon (in decimal hours).
- Wall21: Time taken to complete the first half of the marathon (in decimal hours). Indicates pacing consistency; higher values may suggest runners "hit the wall."
- km4week: Average kilometers run per week in the 4 weeks before the marathon (including the marathon itself).
- sp4week: Average speed during training in the 4 weeks before the marathon (in kilometers per hour).
- CrossTraining: Whether the runner engaged in cross-training (e.g., cycling or swimming).
- Category: Runner category (e.g., age group).

## 01.b Personal context

On December 6th, 2024 I ran my first marathon completely ignoring the popular recommendation that your first half of the marathon should be slower than the second one. Also, I got sick during the 2 last weeks of my Marathon preparation and was not able to do the required number of kilometers per week. I made a good time but I can get rid of the feeling that I could have run faster if I:

- Listened to the common sense and did the first 21kms at a slower pace than the second 21kms
- Didn't get sick and was able to maintain the good number of kms per week

I have subscribed to my second marathon on March 16th, 2025 and would like to have a better picture on what's important when preparing/running a marathon.

## 01.c Hypotheses

The project investigates the following hypotheses:

1. Running a marathon at a steady pace improves performance.
- Runners with Wall21 Ratio values close to 1.00 (indicating even pacing) tend to finish faster than those with higher values.
2. More training mileage in the last 4 weeks of marathon preparation improves performance
- Runners who run more kilometers during the last 4 weeks before the marathon usually have faster marathon times.
3. Training speed during the last 4 weeks before the marathon affects marathon performance.
- Runners with lower sp4week values (indicating faster training speeds) tend to perform better in marathons.

4. Elite runners are better at pacing (and at keeping up with common sense!)
- Elite runners are expected to have `Wall21 Ratio` values closer to 1.00 compared to amateurs.

(This last hypothesis has been added as an attempt to self-compassionate and cheer myself up with an old good idea that “Practice makes it perfect" 🙂)

# 02. Data Exploration

**Data Preparation and Cleaning**

- General info about the dataset, including data about null values (almost not present, we removed rows with empty cells in the Category column and replaces empty values with NaN in the Wall21 column)

- Checking if the dataset has unique values and duplicates (no duplicates found.

**Measures of central tendency**

- We analyzed the mean, mode, and median of the dataset. This gives us a general idea of the average runners profile and their Marathon results.

- We also observed that there are outliers in the dataset and removed some of them that were apparently a data entry error.

**Columns Understanding and Description**

- A short description of the columns and type assignment (Categorical or Numerical)

**Data Transformation & Validation**

- Creating a new column "MaxGroupAge" to translate the info in the category column into a more readable format.

- Wall21 and MarathonTime values have times in hours and fractions of hours which might be confusing sometimes. Also, the sp4week column appears to represent a measure of speed, but its values don’t match the common unit of minutes per kilometer. Instead, these values seem more aligned with kilometers per hour (km/h), where higher values indicate faster speeds. We decided to convert them to a more common format, i.e. hours and minutes.

- We introduce a new column Wall 21 Ratio (based on the Ideal Marathon Time). This new column is to represent how evenly a runner splits their first and second halves of a marathon.

# 03. Data Analysis

## 03.a Global Data Analysis

- Univariate Analysis:

Our target variable is Marathon Time as we are looking into factors that might influence it.

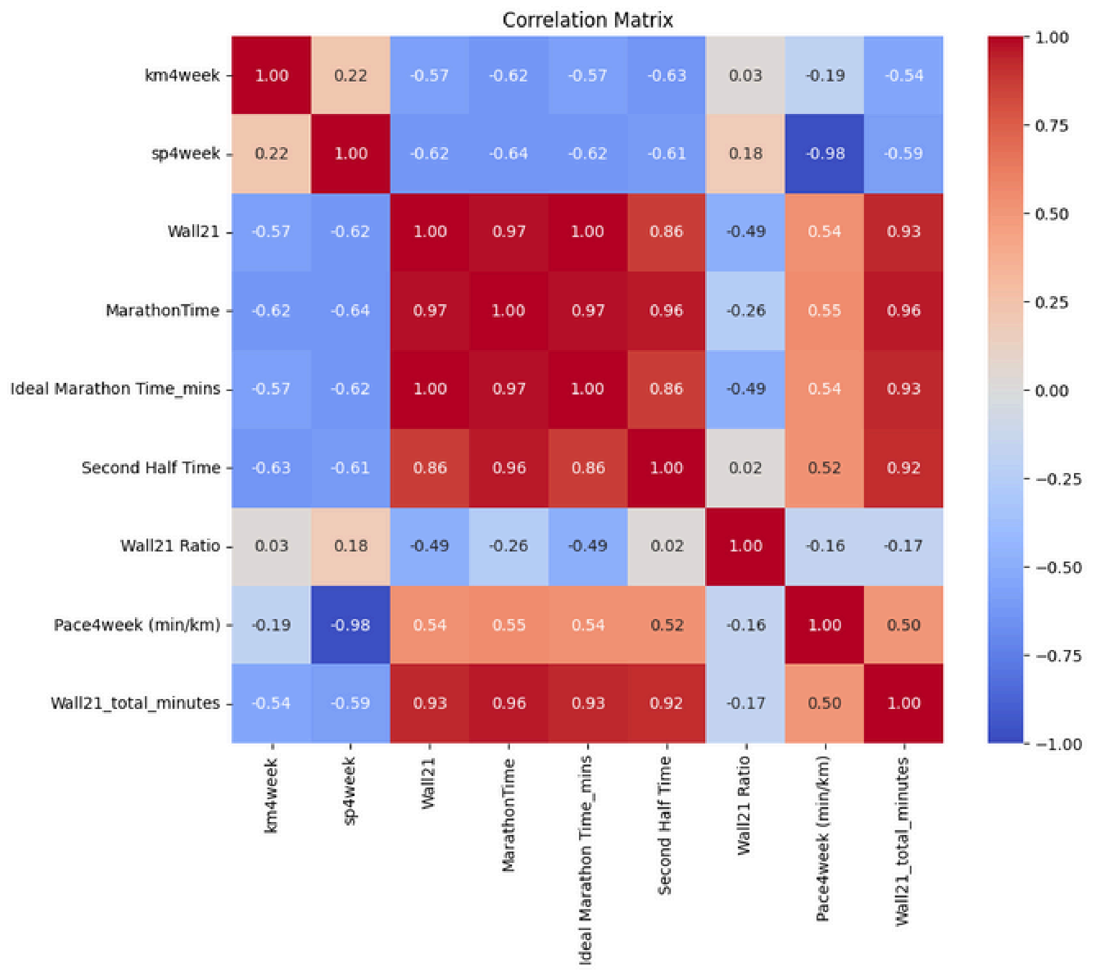
The distribution of the sp4week showed that the training speed is quite uniform across the runners. Also, the Wall21 distribution showed that some athletes may struggle to maintain pace beyond the first half.

The histogram of the Marathon time shows most participants finishing marathons between 3 to 3.5 hours. The MarathonTime and Ideal Marathon Time are close, suggesting participants train to align with their goals.

The Univariate Analysis of Categorical Variables demonstrates that the vast majority (68 runners have no cross-training and that the majority of runners fall into the 40 age group.

- Bivariate Analysis

We generate a heatmap to observe if there are any correlation between the variables:



| | km4week | sp4week | Wall21 | MarathonTime | Ideal Marathon Time_mins | Second Half Time | Wall21 Ratio | Pace4week (min/km) | Wall21_total_minutes |
|---|---|---|---|---|---|---|---|---|---|
| km4week | 1.00 | 0.22 | -0.57 | -0.62 | -0.57 | -0.63 | 0.03 | -0.19 | -0.54 |
| sp4week | 0.22 | 1.00 | -0.62 | -0.64 | -0.62 | -0.61 | 0.18 | -0.98 | -0.59 |
| Wall21 | -0.57 | -0.62 | 1.00 | 0.97 | 1.00 | 0.86 | -0.49 | 0.54 | 0.93 |
| MarathonTime | -0.62 | -0.64 | 0.97 | 1.00 | 0.97 | 0.96 | -0.26 | 0.55 | 0.96 |
| Ideal Marathon Time_mins | -0.57 | -0.62 | 1.00 | 0.97 | 1.00 | 0.86 | -0.49 | 0.54 | 0.93 |
| Second Half Time | -0.63 | -0.61 | 0.86 | 0.96 | 0.86 | 1.00 | 0.02 | 0.52 | 0.92 |
| Wall21 Ratio | 0.03 | 0.18 | -0.49 | -0.26 | -0.49 | 0.02 | 1.00 | -0.16 | -0.17 |
| Pace4week (min/km) | -0.19 | -0.98 | 0.54 | 0.55 | 0.54 | 0.52 | -0.16 | 1.00 | 0.50 |
| Wall21_total_minutes | -0.54 | -0.59 | 0.93 | 0.96 | 0.93 | 0.92 | -0.17 | 0.50 | 1.00 |

It allows us to think that:

**Higher weekly mileage (**km4week**)** and **faster speeds (**sp4week**)** are the **strongest predictors** of faster marathon times.

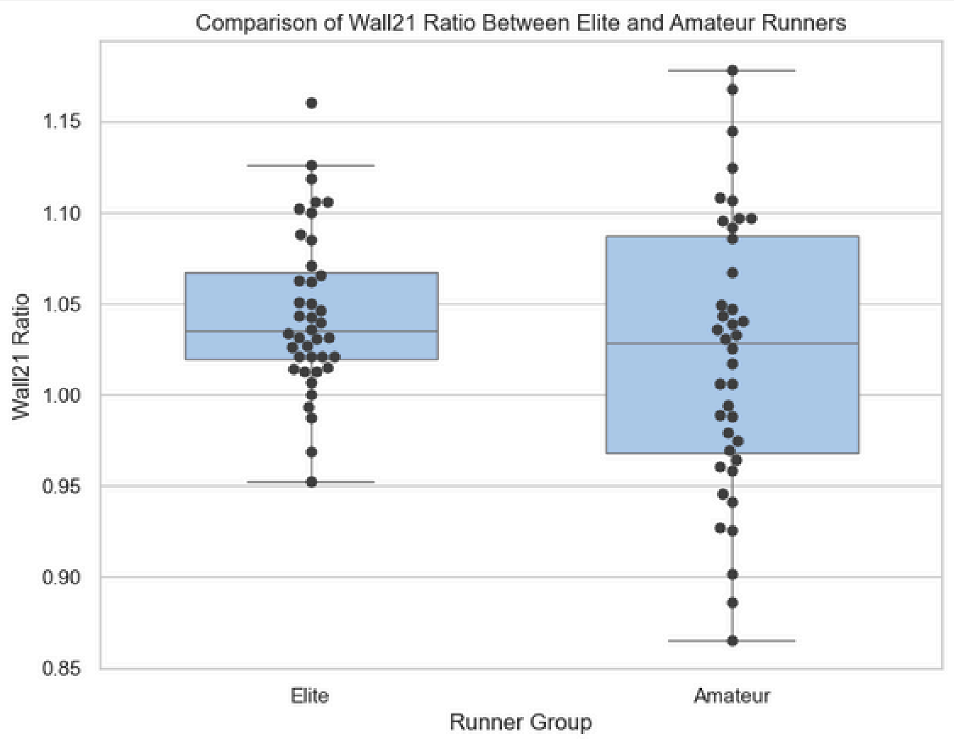
**Balanced pacing** (Wall21 Ratio close to 1.00) contributes positively but less strongly compared to mileage and speed.

Poor performance in the **first half of the marathon (**Wall21**)** is strongly associated with slower overall times and second-half struggles.

It seems interesting to investigate the relationship between Wall21 ration and Marathon Time further. We build a scatter plot to visualize it, as well as Wall21 Ration categories (bins) to range the values. We perform an ANOVA test to confirm that runners who maintain a balanced Wall21 Ratio (closer to 1.0) perform better overall.

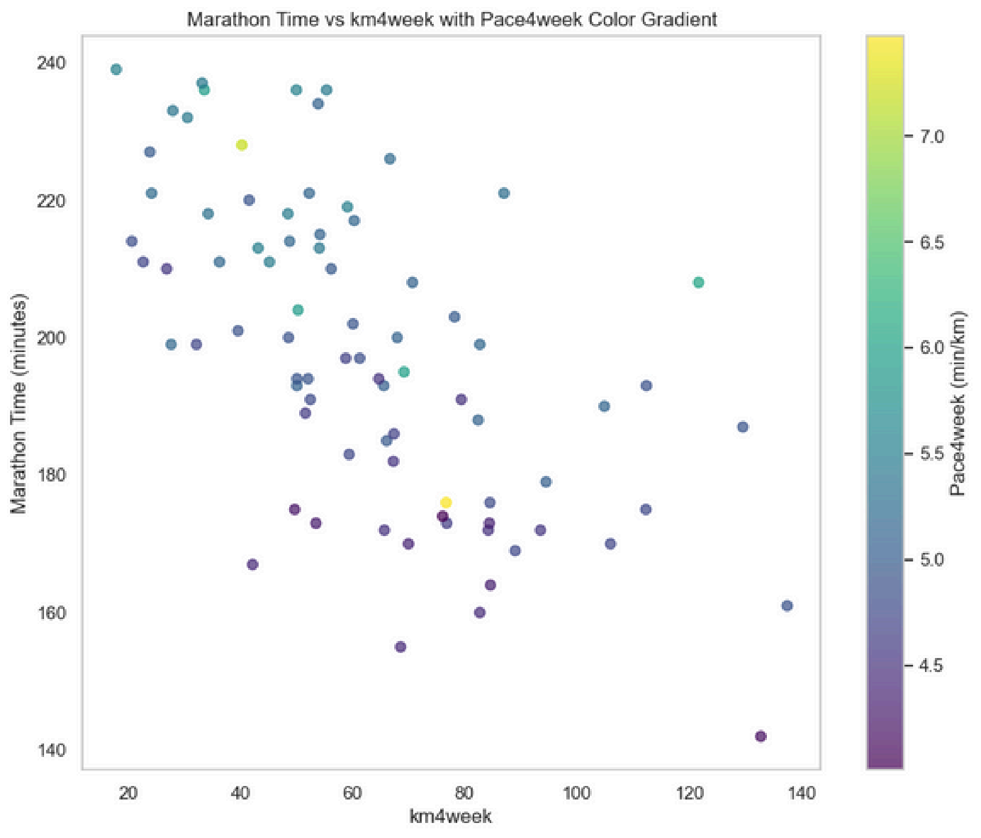
The analysis of categorical variables against our target variable (Marathon Time) showed that runners in younger or middle-aged male categories outperform others (expected). Although most runners do not perform cross-training, those who engage in 3–5 hours of cycling seem to have slightly better marathon times on average. Weekly distance and speed have a strong negative relationship with marathon time. (The more kilometers you run each week and the faster your running speed during training, the faster you are likely to finish the marathon.)

Then we divide all runners into to groups: Elite and Amateur to see if there any pattern within these to groups. The bloxplot shows that elite runners tend to have more even pace during the entire marathon:



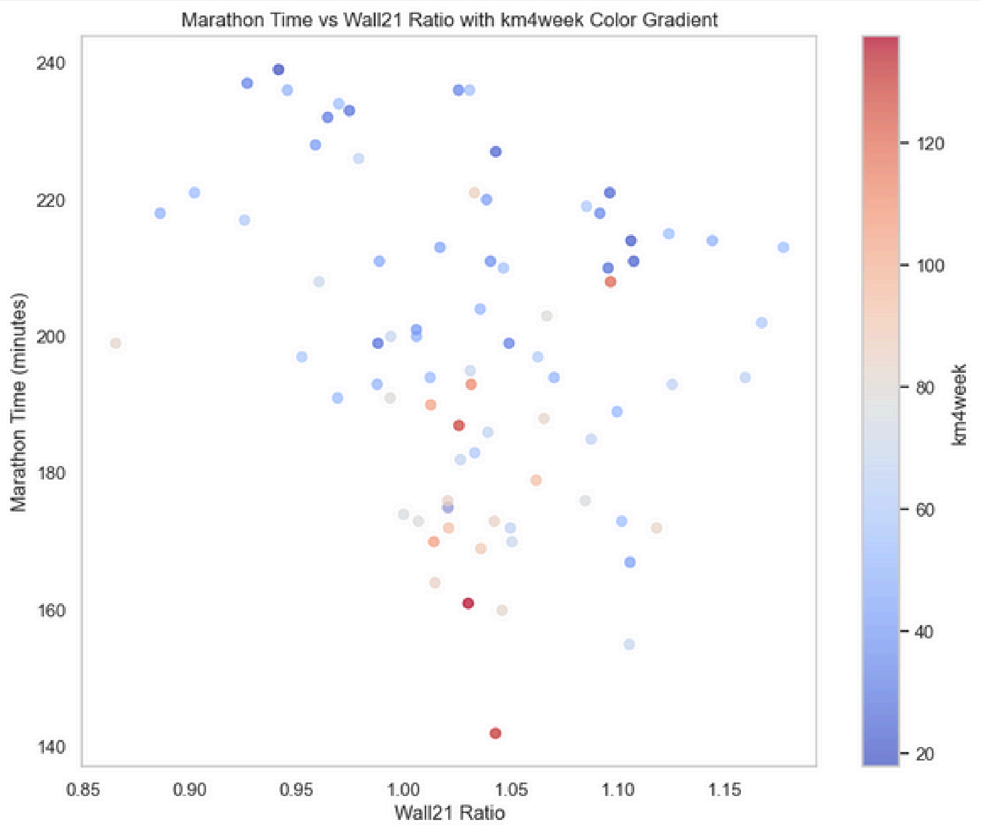
- Multivariale Analysis

We compared different pairs of variables against our target variable to see if any combination of factors influence Marathons success in a greater way.



## 03.b Data Analysis focused on responding to hypotheses

Our Global analysis already showed some correlation and in this second part we centered on visualizing the data that would allow to answer our 4 initial questions (hypothesis). We draw scatterplots and violin plot to illustrate the dependencies between the variables.



# 04. Conclusions

The analysis aimed to uncover factors contributing to marathon performance, particularly those that could inform better preparation strategies. Using a dataset from the Prague Marathon 2017, several key findings emerged:

**General Insights**

1. Training Volume and Speed: Weekly mileage (km4week) and training pace (sp4week) emerged as the most significant predictors of faster marathon times. Runners who consistently logged higher kilometers and maintained faster speeds during their training performed better.
2. Pacing Strategy: Balanced pacing, as reflected by a Wall21 Ratio close to 1.00, correlated with improved performance. Runners who experienced minimal slowdowns in the second half performed better overall. However, slight slowdowns (Wall21 Ratio between 1.0 and 1.05) appeared optimal, suggesting that perfectly even splits may not always be necessary.
3. Cross-Training: While most runners did not incorporate cross-training, those who did (e.g., cycling for 3–5 hours weekly) demonstrated slightly improved performance on average.
4. Age and Category: As expected, younger and middle-aged runners outperformed older participants.

**Responding to Hypotheses**

- Pacing and Performance: The hypothesis that steady pacing improves performance was supported. Elite runners, in particular, displayed more balanced pacing, as seen in their Wall21 Ratio values tightly clustered around 1.00.
- Training Volume: Higher weekly mileage in the 4 weeks leading up to the marathon strongly correlated with faster marathon times, validating the second hypothesis.
- Training Speed: Faster training speeds (lower sp4week values) were also linked to better outcomes, confirming the third hypothesis.
- Experience and Pacing: Elite runners showed superior pacing skills compared to amateurs, supporting the idea that experience enhances pacing ability.

**Practical Implications**

- Balanced Preparation: Marathoners should aim for a balance between higher training mileage and faster training speeds while targeting a Wall21 Ratio close to 1.05 for optimal performance.
- Incorporating Cross-Training: Although not decisive, integrating activities like cycling may offer additional benefits, particularly for injury prevention and performance gains.

This analysis highlights the importance of a structured and balanced training approach, emphasizing consistent mileage, pace, and pacing strategy.