Marathon Time Prediction

Garrett Marshall

2023-02-24

### Introduction

Their are a myriad of factors that an athlete must consider when training for a marathon. In this project I will explore those factors and see if they can be used to predict an athlete’s marathon finish time. I chose this topic to research because I am currently training for a marathon. My goal is to learn about the important factors that go into marathon running and apply them to my training in hopes of me being able to train and race more efficiently. The data set I will use in this project is called “Marathon Time Predictions Data” and was found on Kaggle.com. The data set consists of athlete’s training data for the 2017 Prague Marathon. The data was collected from the Strava app, which is a social network service used for tracking the physical exercise of it’s users. The data set contains many variables about the athletes and their training history. I am hoping to use these variables to find trends and make predictions about the athlete’s finish time.

### Data Description

The marathon data set consists of 81 observations and 8 variables. The following table lists and describes each variable in the data set.

| **Variables** | **Description** |
| --- | --- |
| Category | The sex and age group of the runner as a factor. The factors are MAM: males under 40 years, WAM: women under 40 years, M40: males between 40 and 45 years, M45: males between 45 and 50 years, M50: males between 50 and 55 years, and M55: males between 55 and 60 years. |
| km4week | The total number of kilometers an athete ran in the 4 weeks prior to the marathon. |
| sp4week | The average speed in kilometers/hour of an athlete in the 4 weeks prior to the marathon. |
| Wall | The amount of time in hours it took the athlete to run the first half of the marathon. |
| MarathonTime | The athlete’s marathon finish time in hours. |
| FinishCategory | The marathon finish times grouped into 4 factors. The factors are A: results under 3 hours, B: results between 3 hours and 3 hours 20 minutes, C: results between 3 hours 20 minutes and 3 hours 40 minutes, and D: results between 3 hours 40 minutes and 4 hours. |
| first\_percent\_split | Related to the ‘Wall’ variable. The percentage of the finish time allocated to the first half of the marathon. |
| cross\_train\_y\_n | Was the athlete a cross training athlete? Did they train cycling and swimming as a triathlete? This is a yes or no factor. |

I analyzed each variable above in hopes of discovering import insights about predicting marathon finish times. The variables I decided to explore in more detail were ‘km4week’, ‘first\_split\_percent’, ‘cross\_train\_y\_n’, and ‘MarathonTime’.

### Data Evaluation

The variable I was most interested in analyzing was ‘cross\_train\_y\_n’. This is because I am incorporating a form of cross-training into my marathon training. Not only does my training include running but I am doing weight training as well. Since I am not solely running, I was curious how this may affect my marathon time. In figure.1, I compare cross-training athletes to non-cross-training athletes’ finish times. An athlete who cross-trains would be someone who also trains in another sport like cycling or swimming. The main difference we can see in these two groups is their means. The athletes who cross-train have a higher average finish time than those who do not. I suspect we see this discrepancy because cross-training athletes would have a bulkier physical build than athletes who solely do marathon training, which would result in a slower pace. Figure.1 is an important visualization because it provides information to athletes who cross-train and what they can expect when running a marathon.

A relationship I wanted to explore was between the percentage of the first half split and marathon finish time. For an athlete to run an efficient marathon it is very important to pay attention to the split between the first half compared to the second half. Ideally, it should take an athlete the same amount of time to run the second half as it did to run the first half. If the first half took significantly less time then we can say that the runner burnt their self out or hit the ‘wall’. Since my data was collected only from the 2017 Prague Marathon, which is an invitational race, all of the athletes are either very experienced or professionals. This caused me to not observe a runner that had a first-half split time that was significantly less than the second half. All athletes ran very efficient marathons. I explore each athlete’s percent of the first half split in figure.2. In the left dot plot in the figure.2 we see a positive trend where athletes with a lower percentage of first-half split ran a faster marathon than those who had a higher first-half split percent.

Another relationship I wanted to explore was between the distance ran in the four weeks leading up to the race and the marathon finish time. In the right dot plot in the figure.2 we so a fairly strong negative relationship between the two variables. The more an athlete ran before the race resulted in a faster marathon time. This is an important metric to consider when training for a marathon as you need to get your body used to that volume of running.

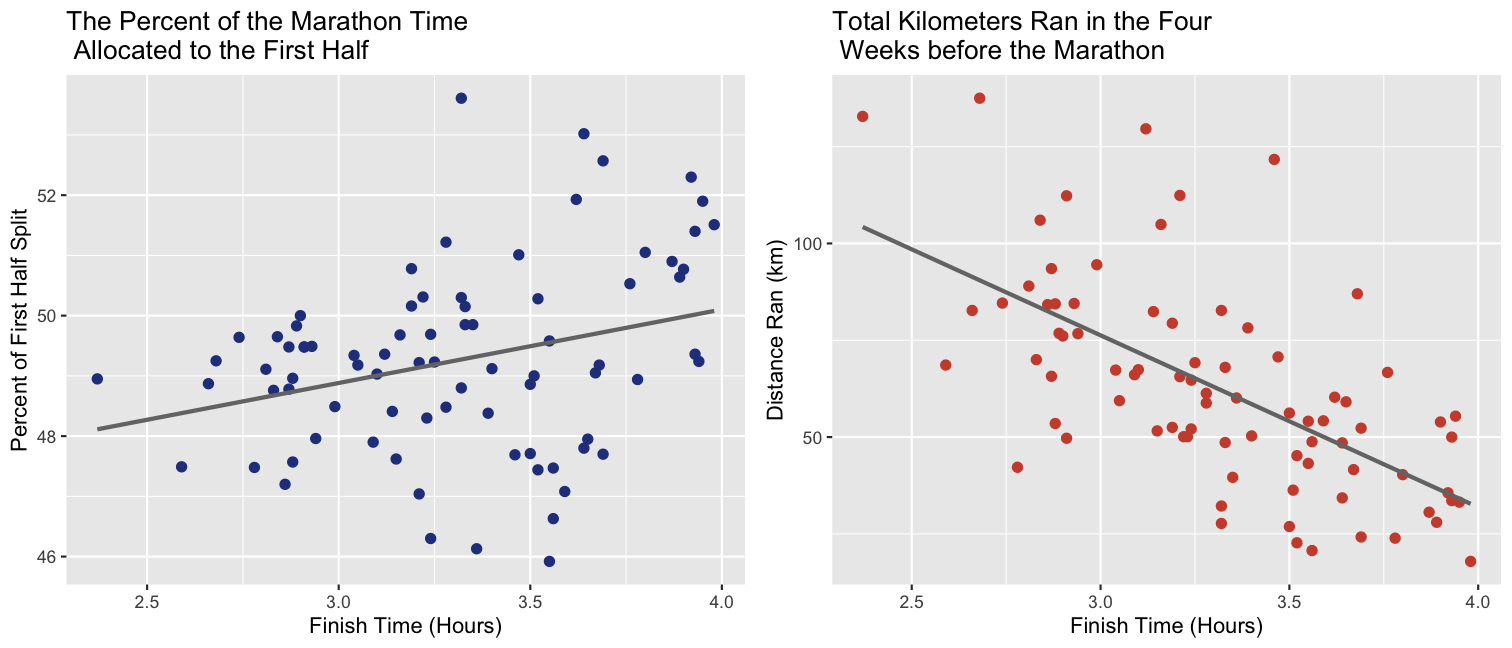
**Figure 1**

Chart, box and whisker chart

Description automatically generated

The box plots above show the average marathon finish time between athletes who cross-train and athletes who do not cross-train. An athlete who cross-trains would be someone who also trains another sport like cycling or swimming. The main difference we can see in these two groups is their means. The non-cross-training athletes have a faster marathon time on average compared to the cross-training athletes.

**Figure 2**



The graphic above shows two scatter plots, the left shows the percent of the finish time allocated to the first half of the race compared to the athlete's finish time, and the right shows the total distance each athlete ran in the four weeks leading up to the race compared to their finish time. On the left, we see a positive trend where athletes with a lower percentage of their first half split finished the race faster than those whose first half split took up a higher percentage. In the plot on the right, we see a negative trend where athletes who ran more leading up to the marathon achieved a lower finish time. The percent of first half split and the total distance ran before the race are two important metrics that contribute to an athlete running an efficient marathon.

### Results and Analysis

For my statistical analysis tests, I decided to perform two simple linear regressions. The first is between the percent of the first half split and marathon finish time, which I will refer to as ‘test one’. The second test will be between the distance ran in the four weeks leading up to the race and the marathon finish time, which I will refer to as ‘test two’. For these tests, I first need to verify if they meet the assumptions of simple linear regression. These assumptions are the independence of observations, constant variance (homoscedasticity), and normality of the data.

For test one, the observations are independent as each runner is independent of one another. The homoscedasticity can be examined in the residual plot in figure 3. In the plot, we can see that the variance is not as constant as we would like to see but it is not horrible. We can use the Q-Q plot in figure 3 to examine the normality of the data. Similarly to the residual plot, we do not see what we would want with the two ends flaring outwards. Although our data does not meet the assumptions perfectly, I still plan on completing the simple linear regression.

For test two the observations are independent for the same reason as test one, each runner is independent of one another. The homoscedasticity can be examined in the residual plot in figure 4. We can see in this residual plot that the variance is somewhat constant. We can use the Q-Q plot in figure 4 to examine the normality of the data. In this plot, we see that the data is normally distributed. Overall, test two meets all of the assumptions of simple linear regression. The table below shows important outputs from our first test.

| **Test One Results** |  |
| --- | --- |
| Formula: |  |
| P-Value: |  |

The results of this first test suggests a significant relationship between the percent of first half split and marathon finish time, with a p-value of at a % significance level. Regardless of the results form this test I can not be confident that there is a significant relationship because the assumptions of simple linear regression was not met. Moving on to our second test, the table below shows the results.

| **Test Two Results** |  |
| --- | --- |
| Formula: |  |
| P-Value: |  |

The results of the second test suggests s significant relationship between the distance ran in the four weeks leading up to the race and the marathon finish time, with a p-value of at a % significance level. Unlike test one, I can be very confident in the result of test two as all assumptions of simple linear regression were met.

**Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generatedFigure 3**

The Q-Q plot (left) can be used to address normality of the data. In this case, we see points flaring outwards at either end which would not suggest normality. Ideally, we would want to see the points much more in-line with the 45-degree dotted line.

The residual plot (right) can be used to address the spread of the variance or homoscedasticity of the data. In this case, we do not see the constant variance we want. Ideally, we would want to see no patterns but rather a random scatter of points spread along the dotted horizontal line.

### Figure 4

### Chart, scatter chart Description automatically generatedChart, scatter chart Description automatically generated

The Q-Q plot (left) can be used to address normality of the data. In this case, we see points that closely follow the 45-degree dotted line, which would suggest normality.

The residual plot (right) can be used to address the spread of the variance or homoscedasticity of the data. In this case, we do see the constant variance we want. The random spread of points is not perfect but does follow the dotted horizontal line better than we saw in figure 3.

### Conclusion

In this project I explored the many factors that goes into training for a marathon. I am able to conclude with confidence that the amount of running someone does in the weeks leading up to the marathon can have a significant impact on the results of the race. This is one of many metrics a runner should keep in mind when preparing for a marathon.

Although I enjoyed this topic for my project, I think it would be very interesting to recreate my tests with different data. The data set I used only contained information about professional runners, as mentioned in the ‘Data Evaluation’ section. I think that I could find interesting results from a data set that contains information of all types of runners, from amateurs to professionals to people would simply walk the marathon. I believe that this kind of data set could provide more applicable information to myself and runners alike.

### Appendix and Citations

ANDREA GIRARDI. (2017). Marathon Time Predictions, Version 2. Retrieved February 1, 2023 from <https://www.kaggle.com/datasets/girardi69/marathon-time-predictions>.

All code regarding data cleaning/wrangling and statistical tests can be found on my Github. <https://github.com/Garrett456/MarathonProject>