Exploring the Track Records Dataset

Jeremy Maslanko

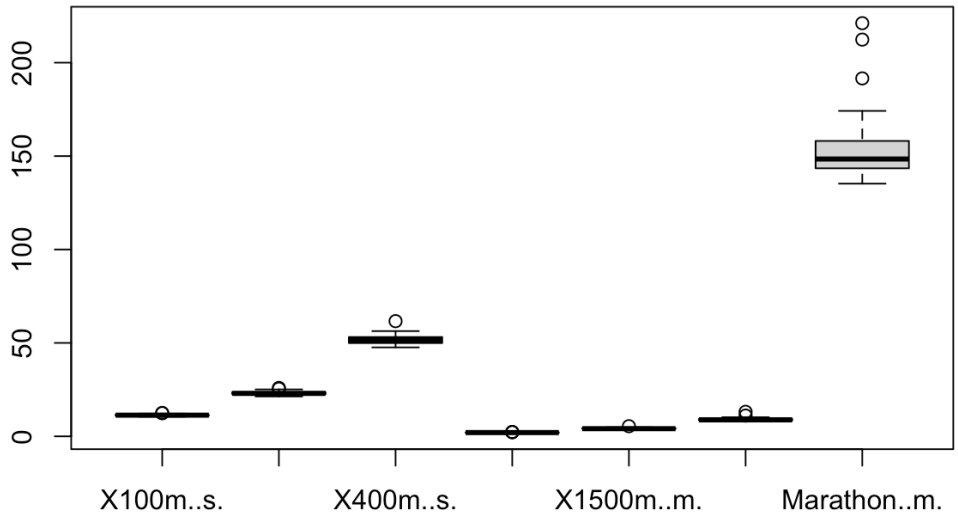
Introduction

Understanding track records by country gives us valuable insight into the sport that we would not be able to gain by simply looking at the overall records. We can see trends across countries as well as spot outliers. The data we have has the record for the following track events: 100m, 200m, 400m, 800m, 1500m, 3000m, and full marathon. Additionally, we have the records for 54 countries. Throughout this paper, we will complete simple exploratory data analysis to provide us with a clearer understanding of the data. This will help paint a picture of the track records data so that we can further analyze the data in the future.

Analysis

To begin, lets take a look at our data summary. We will take a look at the five number summary, but will display the data with box and whisker plots. This can be seen in Figure 1 below.

Figure 1: Box and Whisker

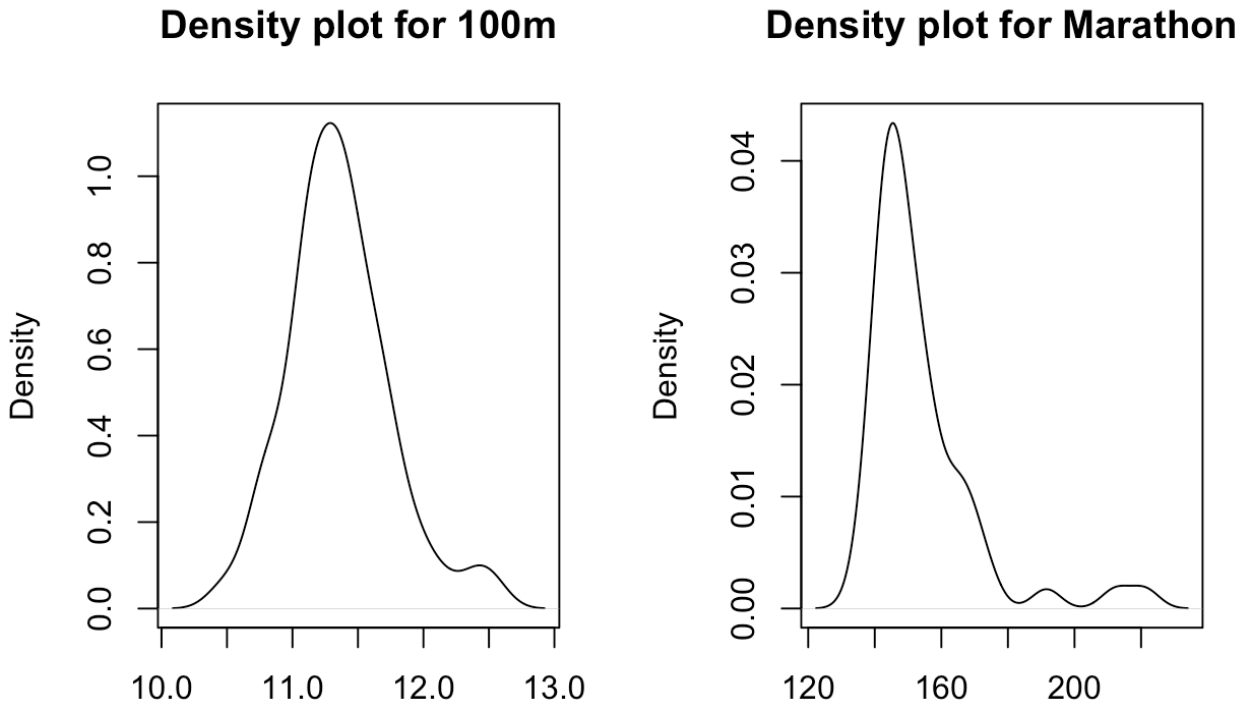


Initial findings from this plot show that as the distance gets longer, the spread in the data also gets larger. One thing to note, the 100,, 200m, and 400m and distances are measured in seconds, while

everything longer is measured in minutes. For the simplicity of this analysis, we will keep these units as is.

Not only do we see the spread of the data get larger as the distance increases, but it also appears that we have more outliers as the distance gets larger. This is most notable when looking at the Marathon distance plot.

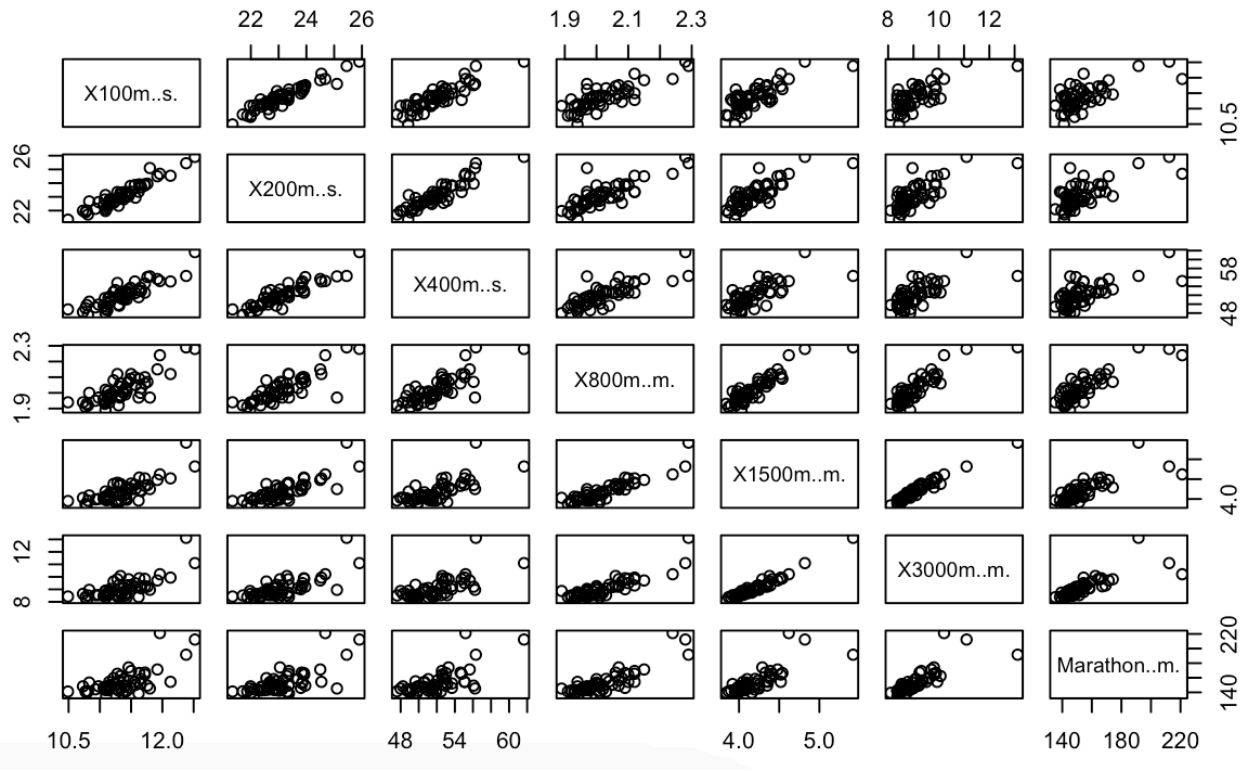
With a basic understanding of what the spread of data looks like, lets take a look at how the data is distributed. When we look at each distribution, we see a gradual change from a normal distribution to gamma distribution or simply a skewed normal distribution. Below we can see this change from the 100m to the marathon.



This aligns with what we saw earlier with the box and whisker plots, that as the race distance increases, so does the spread.

With an understanding of how the data looks, lets now look to see if there is any relationship between the different variables. Figure 2 shows us the correlation matrix, which provides the correlation plot between each of the variables in our dataset.

Figure 2: Correlation Matrix



Overall, most of our variables show positive linear correlation. For the shorter distances, there even appears to be strong positive linear correlation. But as we have seen with the other plots, the longer the distance, the more the data starts deviate. This is especially noticeable when comparing the 100m data with the marathon data. Not only are we seeing those outliers poke their head out again, but we also see less of a correlation. Below is a correlation table, which shows the correlation value between each variable.

	X100m..s.	X200m..s.	X400m..s.	X800m..m.	X1500m..m.	X3000m..m.	Marathon..m.
row names	1.0000000	0.9410886	0.8707802	0.8091758	0.7815510	0.7278784	0.6689597
X200m..s.	0.9410886	1.0000000	0.9088096	0.8198258	0.8013282	0.7318546	0.6799537
X400m..s.	0.8707802	0.9088096	1.0000000	0.8057904	0.7197996	0.6737991	0.6769384
X800m..m.	0.8091758	0.8198258	0.8057904	1.0000000	0.9050509	0.8665732	0.8539900
X1500m..m.	0.7815510	0.8013282	0.7197996	0.9050509	1.0000000	0.9733801	0.7905565
X3000m..m.	0.7278784	0.7318546	0.6737991	0.8665732	0.9733801	1.0000000	0.7987302
Marathon..m.	0.6689597	0.6799537	0.6769384	0.8539900	0.7905565	0.7987302	1.0000000

Just as we saw with the plots, we can see the correlation values gradually get smaller as we start to compare distances that are farther apart. From this, we can conclude that the track records for a shorter distance race are more predictable when knowing another distance track

record. As an example, the 100m and 200m races have a higher correlation than the 100m and Marathon distance.

Throughout this analysis, we have noticed a few outliers present, most noticeable in the marathon distance. These are data points that are noticeably different than the rest of the values. If we were going to complete more advanced modeling on this data (i.e. regression), it would be beneficial to think about removing the outliers. We would want to remove them due to their influence on any modeling that would be done. Put another way, they will skew any predictions we may make since they do not represent the majority of the data that we have.

Trying to explain these outliers can prove to be difficult just from looking at this data. Because there are so few and that they only for the longer distances, we may not want to look into other pieces of data that could explain this. One such way might be by looking at what the training programs look like for each country. Perhaps the counties that have outlier records for the marathon have different priorities for the training for that event. If they don't have as many resources available for training that discipline, then it would be reasonable to assume that that has an impact in track record for that event.

Future Improvements

In the future, there are a couple things I would recommend to improve this analysis. The first is to convert the variables to be in the same unit of measurement. This would provide a simpler understanding of the relationships of the variables and would lead to more accurate results.

The second recommendation would be to gather more data that could be used as additional features to explain the relationships. A couple recommendations are average hours per week spent training, average dollar amount spent on training materials, as well as climate of training facility. Additionally, it may be valuable to collect data on more countries to see if the outliers are real or just a consequence of the countries selected.

Conclusion

In conclusion, examining track records by country provides a nuanced understanding of sports beyond mere statistics, allowing us to identify trends and outliers across various events. With data spanning 100m sprints to full marathons from 54 countries, this paper's simple exploratory analysis lays the groundwork for deeper insights in the future. By painting a clearer

picture of track records, we gain valuable insights to inform decisions and foster appreciation for athletes' achievements globally.