



JMP Instructional data set 011

Marathon Runners' Performance

JMP platforms and statistical methods

Formula Editor:	Creating new variables
Distribution:	Histogram; box plots; summary statistics; bar chart
Graph Builder:	scatterplot
Multivariate:	Correlation; scatterplot matrix

Problem statement

The performance of marathon runners can be influenced by many factors. Two such factors are the age and sex of the runner.



A data set of 26,598 marathon runners who finished a recent Boston Marathon was gathered. In addition to the runners' age and sex, data on the time at which each runner reached five distances in the marathon was recorded at 5, 10, 15 kilometers, as well as 21.1 kilometers (for half marathons) and 42.2 kilometers (for full marathons).

The objectives of the analyses will be to:

1. Describe general patterns in marathon runners' performance.
2. Identify similarities and differences in running performance between male and female runners.
3. Describe any differences in performance across ages of the runners.
4. Compare and describe performance across different phases of the marathon.
5. Identify runners that are most unique in their performance.

Data

IDS_011.jmp

Placement	Order of completion
Age	Age of runner
Sex	Sex of runner (Male, Female)
5K Time	Time (h:mm:ss) of runner at 5 kilometer mark
10K Time	Time (h:mm:ss) of runner at 5 kilometer mark
15K Time	Time (h:mm:ss) of runner at 5 kilometer mark
21.1K Time	Time (h:mm:ss) of runner at 5 kilometer mark
42.2K Time	Time (h:mm:ss) of runner at 5 kilometer mark
Pace	Average run time per kilometer (h:mm:ss)

Exercises

1. Create the following new variables using the formula editor. 1. Difference between 10K and 5K Time ; 2. Difference between 15K and 10K Time ; 3. Difference between 21.1K and 15K ; 4. Difference between 42.2K and 21.1K Time ; 5. Difference between the time to complete 2nd half of the race from the 1st half.

Instructions:

Select Cols > New Columns. Name the column '10K-5K Time'. Choose Duration > hr:m:s for the Format. Select Formula under Column Properties drop down menu. In the Formula Editor, create a formula that subtracts the '5K Time' from the '10K Time'. $10K\ Time - 5K\ Time$

Click OK, then OK again in the New column dialog box.

Follow similar steps to create the '15K-10K Time', '21.1K-15K Time', and '42.2K-21.1K Time' variables.

For the '2nd Half – 1st Half Time' variable, the formula should be: $42.2K-21.1K\ Time - 21.1K\ Time$

2. To get familiar with the data, summarize the distribution of the variables 'Age', 'Sex', and all the Run Time variables. Provide a few brief sentences describing what the shape of the distributions for the run times reveals. Examine the age of runners for each sex. Are they distributed similarly for each sex or does one of the sexes tend to have more/less older runners? Examine the Run Time variables for each sex. Is there any indication that the Run Times are somewhat different for the two sexes.

Instructions:

Select Analyze > Distribution

Place the variables 'Age', 'Sex', '5K Time', ... , '2nd Half – 1st Time' in the Y, Columns. Click OK.

In the Distribution report, click on each sex in the bar chart and examine the part of the distribution highlighted in all the other variables.

3. Create a Scatterplot of all the Run Time variables for the runners to complete the marathon by their 'Placement' with 'Sex' as the overlay variable. Provide a few brief sentences describing what this graph reveals about the rate at which runners are completing the race throughout their placement. Example: Describe the general shape, variation, any outliers.

Instructions:

Graph > Graph Builder. Place '5K Time' on the Y axis, and 'Placement' on the X axis. Place 'Sex' in the Overlay role.

Choose Redo > Column Switcher under the red triangle. Select '5K Time' as the Initial Column to Switch. Select that variable and all the other Run Time variables as the Replacement Columns.

Select Local Data Filter under the red triangle. Choose the 'Sex' variable and click on "+" symbol.

Under the red triangle next to the Graph Builder label, select Lock Scales. This way, when using the Data Filter to switch between the two sexes, the Y axis will stay fixed.

4. Examine the following individual runners (2950, 11204, 16651, 13155, 17616, 19946, and 26598) in the Scatterplot created in Exercise 3. For each, Provide a brief sentence or two describing the performance of this runner throughout the marathon.

Instructions:

Highlight runner 2950 in the data table. Go to the scatterplot to see that runners '5K Time'. Using the Column Switcher, move through each Run Time variable to see how their time compares to the other runners and how it changes throughout the stages of the marathon.

Repeat for the other runners.

5. Create a Scatterplot of all the Run Time variables by their 'Age' with 'Sex' as the overlay variable. Provide a few brief sentences describing what these graphs reveal about the relationship between Run Times and Age, and comparisons between males and females. Is the relationship between the Run Times and Age, and differences between male and female similar for all distance ranges?

Instructions:

Graph > Graph Builder. Place '5K Time' on the Y axis, and 'Placement' on the X axis. Place 'Sex' in the Overlay role.

Choose Redo > Column Switcher under the red triangle. Select '5K Time' as the Initial Column to Switch. Select that variable and all the other Run Time variables as the Replacement Columns.

Because there are many runners (26,598), there is a lot of overlapping of the data points making it difficult to compare the two sexes. Choose to display the data points, smoother, and contour by selecting all three in the graph palette while holding down the shift key. 

Right-click on the graph and choose Graph > Marker Size > 0, Dot to make the smoother and contours more visible.

Select Local Data Filter under the red triangle. Choose the 'Sex' variable and click on "+" symbol. Under the red triangle next to the Graph Builder label, select Lock Scales.

6. Create a Scatterplot Matrix and Correlation table of the Run Time variables ('5K Time', '10K-5K Time', '15K-10K Time', '21.1K -15K Time', and '42.2K – 21.1K Time'). It of course is no surprise that there are strong correlations among these variables. Notice, however, the pairs that have the highest correlation and those that have the least. Interpret this.

Instructions:

Analyze > Multivariate Methods > Multivariate. Place the Run Time variables in the Y, Columns. Click OK.

7. Create a Scatterplot between the Run Time during the 2nd half of the race ('42.2K-21.1K Time') vs. 1st half ('21.K Time'). Provide a few brief sentences describing what this graph reveals. Identify a few runners where the difference in the two Runs times is greatest with regard to the second half being much longer than the first half. Describe the difference. Identify a few runners where the difference in the two Runs times is greatest with regard to the second half being less than the first half. Describe the difference.

Instructions:

Graph > Graph Builder. Place '42.2K – 21.1K' on the Y axis, and '21.1K Time' on the X axis. Place 'Sex' in the Overlay role.

Create a line representing equal time between the two variables:

Right-click on the graph and choose Customize. Click "+". Select the Templates drops down menu and choose 'Y Function'. Edit the script to read 'Y Function(y=x,x).

8. Create a Scatterplot between the difference in the Run Times between the 2nd half and 1st half of the marathon vs. Placement. Provide a few brief sentences describing what this graph reveals. Example: Does it appear that the difference in Run Times between the 2nd and 1st half of the race is much different across the 'Placement' of the runners?

Instructions:

Graph > Graph Builder. Place '2nd - 1st Half' on the Y axis, and 'Placement' on the X axis. Place 'Sex' in the Overlay role.

Add a reference line on the y axis representing no difference in Run Time between the two halves of the race:

Double-click on the Y axis. Type in the value 0:00:00, click "Add".

9. Create a Scatterplot between the difference in the Run Times between the 2nd half and 1st half of the marathon vs. Age. Provide a few brief sentences describing what this graph reveals. Example: Does it appear that the difference in Run Times between the 2nd and 1st half of the race is much different across the range in 'Age' of the runners?

Instructions:

Graph > Graph Builder. Place '2nd - 1st Half' on the Y axis, and 'Age' on the X axis. Place 'Sex' in the Overlay role.

Add a reference line on the y axis representing no difference in Run Time between the two halves of the race:

Double-click on the Y axis. Type in the value 0:00:00, click "Add".

10. What are some other factors that might impact runners' performance, and thus would be useful data to explore?