MARATHON RUNNERS PERFORMANCE

RELEVANT JMP PLATFORMS AND STATISTICAL TECHNIQUES

Distribution :	Histograms, Boxplots, Summary Statistics, Bar Chart
Graph Builder :	Scatterplots
Multivariate :	Correlations, Scatterplot Matrix

PROBLEM STATEMENT

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

A dataset of 26,598 marathan 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 5 distances in the marathon was recorded (5, 10, 15 kilometers as well as 21.1 kilometers (1/2 marathon) and 42.2 kilometers (full marathon)). A few 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; 4. Identify runners that are most unique in their performance.

DATA SET

MARATHON_RUNNERS_PERFORMANCE.JMP

Placement	Order of finishers
Age	Age of runner
Sex	Sex of the runner (male or female)
5K Time	Time (h:mm:ss) of runner at 5 kilometer mark
10K Time	Time (h:mm:ss) of runner at 10 kilometer mark
15K Time	Time (h:mm:ss) of runner at 15 kilometer mark
21.1K Time	Time (h:mm:ss) of runner at ½ marathon mark (21.1 kilometers)
42.2K Time	Time (h:mm:ss) of runner at completion of marathon (42.2 kilometers)
Pace	Average run time per kilometer

EXCERCISES

Tips: *This dataset contains run times for 26,598 runners. As you'll see in the scatterplots that will be made, data points will have a tremendous amount of overlap and in certain cases can make it hard to see patterns and make comparisons. As a result, it can be advantageous to view scatterplots using different displays: data points*  *, density contours*  *, and a smoother.*  *Note: The number of contour levels and amount of smoothing can be adjusted. In addition, choosing 'Sex' as an overlay variable in the Scatterplots while also setting it up in the Local Data Filter can help with comparisons. It's best if you choose to Lock Scales under the Graph Builder red triangle so the axes don't change as you switch between male and female runners. Lastly, for each of the scatterplots created below, it's best to set up the Column Switcher using all of the time variables so as to provide a way to choose any variable to display in one graph view.*

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.

2. To get familiar with the data, summarize the distribution of the variables Age, Sex, and all the Run Time variables. *Tip: Choose to use 1 minute bin increments. Double-click or right-click on the Y axis to open the axis settings.*
   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 similarily for each sex of does one of the sex tend have more/less older runners?
   Hint: Click on the two different sexes in the bar chart and examine where the data is being highlighted in the Age distribution.

3. Create a Scatterplot of the time 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.

Per the tips listed at the beginning: 1. add Sex as the overlay variable, 2. create a local data filter for Sex, 3. add a Column Switcher including all the Run Time variables, and 4. choose to Lock Scales.

a. Highlight the runner who placed 2,950 in the data table. Now examine all the scatterplots (i.e., switching the time variable on the y axis) with that runner highlighted. Provide a brief sentence or two describing the performance of this runner thoughout the marathon.

b. Highlight the runner who placed 11,204 in the data table. Now examine all the scatterplots with that runner highlighted. Provide a brief sentence or two describing the performance of this runner thoughout the marathon.

c. Highlight the runner who placed 16,651 in the data table. Now examine all the scatterplots with that runner highlighted. Provide a brief sentence or two describing the performance of this runner thoughout the marathon.

d. Highlight the runner who placed 17,616 in the data table. Now examine all the scatterplots with that runner highlighted. Provide a brief sentence or two describing the performance of this runner thoughout the marathon.

e. Highlight the runner who placed 19,946 in the data table. Now examine all the scatterplots with that runner highlighted. Provide a brief sentence or two describing the performance of this runner thoughout the marathon.

4. Create a Scatterplot of the time for the runners to complete the marathon by their Age.
Per the tips listed at the beginning: 1. add Sex as the overlay variable, 2. create a local data filter for Sex, 3. add a Column Switcher including all the Run Time variables, and 4. choose to Lock Scales.
Provide a few brief sentences describing what these graphs reveals about the relationship between run times and age, and comparions between males and females.
Is this relationship between run time and age, and differences between male and female similar for all distance ranges? Where is the least amount of variation in Run Times (young or old aged runners)?

5. Calculate the correlations and create scatterplots between all the Run Times interval variables (5K Time, 10K-5K Time, 15K-10K Time, 21.1K-15K Time, and 42.2K-21.1K Time)
Tip: Use Analyze > Multivariate.
It if of course no surprise that there is strong correlations among these vairables. Notice, however, the pairs that have the highest correlation and those that have the least. Interpret this.

6. 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).
Per the tips listed at the beginning: 1. add Sex as the overlay variable, 2. create a local data filter for Sex, and 3. choose to Lock Scales. Also, add a line representing y=x. Right-click on the graph and choose 'Customize'. Click "+". Select the Templates drop down and choose 'Y Function'. Edit the script to read 'Y Function(y=x,x);'.
Provide a description of features that this graph reveals.

7. Create a Scatterplot between the difference in the run times between the 2nd half and 1st half of the marathon vs. Placement.
   Per the tips listed at the beginning: 1. add Sex as the overlay variable, 2. create a local data filter for Sex, and 3. choose to Lock Scales. Also, add a reference line at 0:00:00. Double-click or right-click on the Y axis to open the axis settings.
   Provide a description of features that this graph reveals. Where, in terms of placement, tend to be the runners with the least/most difference in run times between 1st and 2nd half of the marathon? Are there any differences observed between males and females?

8. Create a Scatterplot between the difference in the run times between the 2nd half and 1st half of the marathon vs. Age.
   Per the tips listed at the beginning: 1. add Sex as the overlay variable, 2. create a local data filter for Sex, and 3. choose to Lock Scales. Also, add a reference line at 0:00:00. Double-click or right-click on the Y axis to open the axis settings.
   Does it appear that the difference in Run Times between the 2nd and 1st half of the race is much different across the age range of the runners?

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