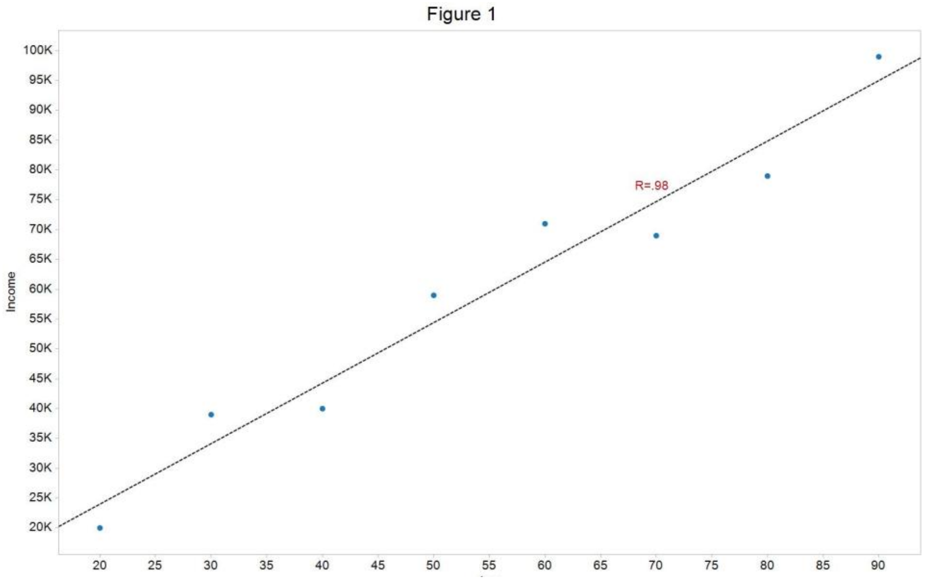
Nuts and Bolts of Data Mining: Correlation & Scatter Plots

By Tim Graettinger

In this article, I continue the "Nuts and Bolts of Data Mining" series. We will tackle two, intertwined tools/topics this time: correlation and scatter plots. These tools are fundamental for gauging the relationship (if any) between pairs of data elements. For instance, you might want to view the relationship between the age and income of your customers as a scatter plot. Or, you might compute a number that is the correlation between these two customer demographics. As we'll soon see, there are good, bad, and ugly things that can happen when you apply a purely computational method like correlation. My goal is to help you avoid the usual pitfalls, so that you can use correlation and scatter plots effectively in your own work.

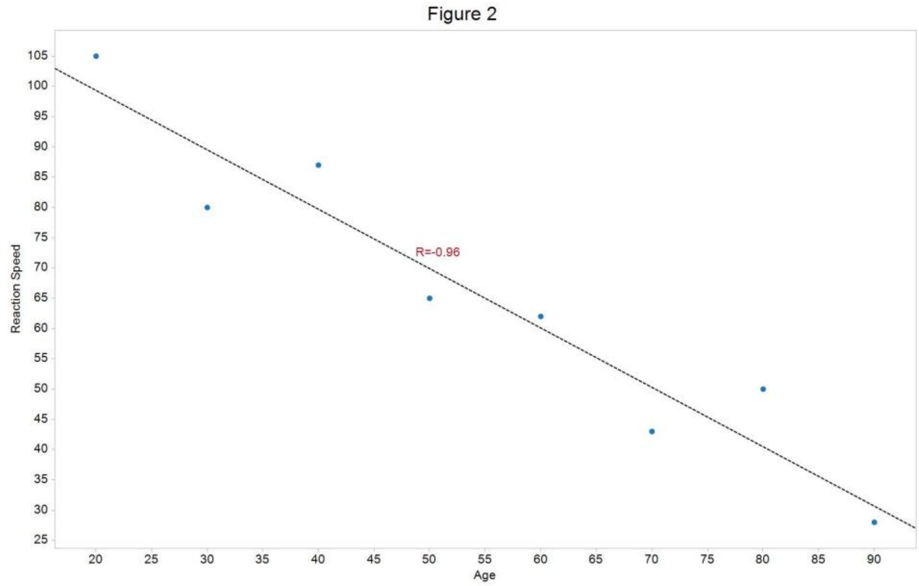
The Good

Consider the graph of income versus age that is shown in Figure 1. In this scatter plot, each dot represents a different customer. Note also the dotted line in the plot that is the "best fit" of a <u>straight</u> line to the age-income data. Finally, the legend on the plot reports the correlation measure, r=0.98. ("R" or "r" is the traditional letter designation for correlation).



Age vs. Income.

Figure 1 - Example of positive correlation



Age vs. Reaction Speed.

Figure 2 - Example of negative correlation

Correlation is a statistical measure, and it indicates how well, or poorly, a straight line conforms to a pair of data elements. By design, the correlation value can range from -1 to +1. A positive correlation is associated with a best-fit line that slants upward to the right, like that in Figure 1. A best-fit line slanting downward to the right, depicted in Figure 2, indicates a negative correlation. A line of best fit that is flat, or nearly so, has a correlation near zero, as shown in Figure 3. The more tightly the data hugs the best-fit line, the larger the magnitude of the correlation – that is, the closer it will be to either -1 or +1.

Sounds ideal, doesn't it? With a single number, we get a quantitative measure of whether or not two data elements, such as age and income, are related, and to what degree. In addition, the sign (+ or -) of the correlation indicates the "sense" of the relationship. For instance, referring again to Figure 1, the positive correlation value tells us that income increases with age – and the scatter plot confirms it. What could possibly go wrong? As it turns out, plenty can go wrong.

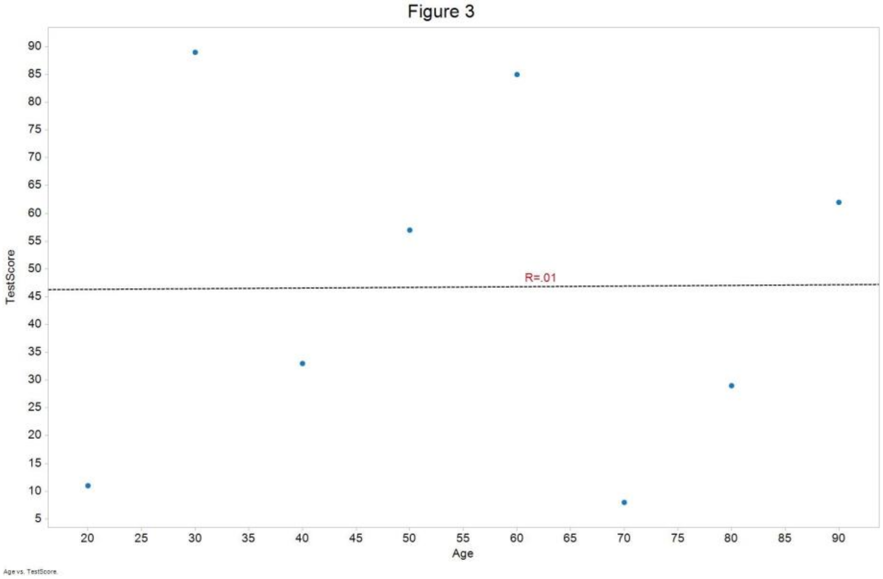


Figure 4 - Example with no correlation

The Bad

As a first example of what can go wrong, suppose you have two data elements that are uncorrelated. The scatter plot of these elements looks like the "shotgun blast" depicted in Figure 3. Suppose further that, in loading the data set which includes these two elements, a column alignment problem occurs[1]. The misalignment occurs in only a single row of data, and large values from adjacent columns are inserted into the wrong slots. The result is shown in Figure 4, with the outlying value clearly marked in the scatter plot.

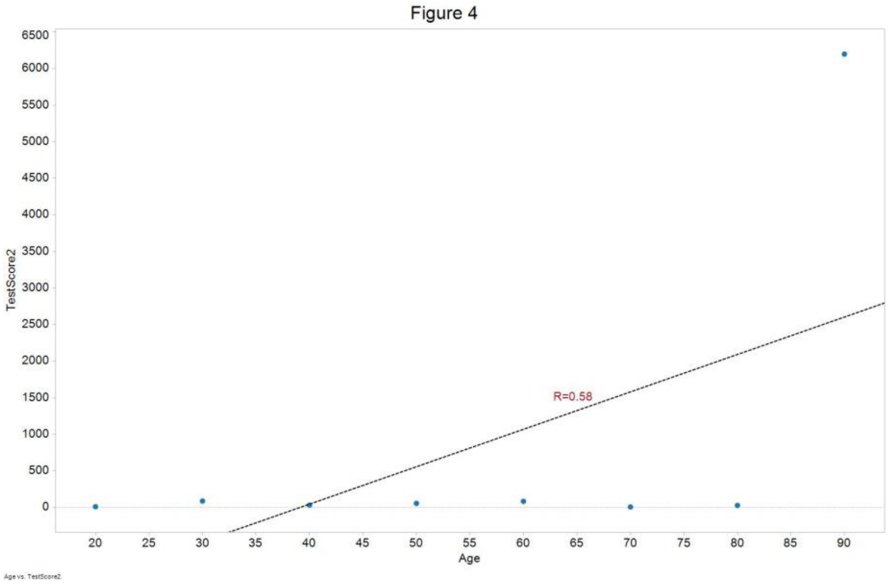


Figure 3 - Outlier induces a correlation where there is none

In Figure 3, we saw that the correlation was approximately zero for the properly aligned data. Now, the single outlying value induces a correlation value of 0.58 – certainly high enough to suggest a strong relationship where, in fact, there really is none.

The Ugly

Let's turn now to a second example of correlation gone awry. Suppose that a single data entry error occurs[2] (let's say a shifting of the decimal point), and the data elements originally shown in Figure 1 appear now as in Figure 5. Like the previous example, we are challenged by a single outlying value.

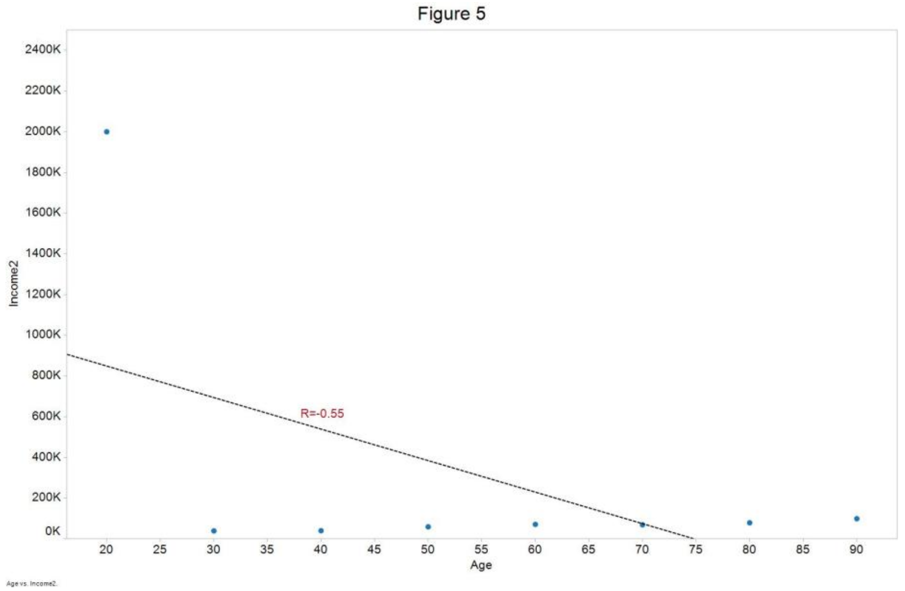


Figure 5 - Outlier induces sense & strength mistakes by the correlation measure

There are important differences in this case, however, compared to the previous. Here, a strong relationship does exist between the two data elements, and it has a positive sense. That is, the true, best-fit line slants upward to the right, as we see in Figure 1. The introduction of one outlier, per Figure 5, not only reduces the strength of the correlation, it also destroys the sense – note how the best-fit line in Figure 5 slants down to the right. Blindly using only the correlation value, we would infer that income decreases with age.

The Remedy

How can we remedy these situations? Seriously, I'm asking – stop reading for a minute and think …

Did you say, "Look at the data"? The outlying values are obvious in the scatter plots of Figures 4 and 5. (They are also obvious in histogram plots of the individual data elements – see "Nuts and Bolts of Data Mining: The Histogram" for more information).

Once you identify the outlying values, what is the next step? Think it over, I'll wait…

Did you say, "Delete the offending data point(s)"? Hopefully not. The response I was looking for was, "Ask if the outlying value is a real, legitimate value." If you can't answer this question yourself, ask someone who can. My philosophy is always, "diagnose before prescribing." You need to understand why the outlying value is present so that you can plan how to handle it.

If you can confirm that any outlying values are NOT legitimate, the next step is to create a "handler". A handler, as the name implies, is a piece of software[3] that can detect an anomalous value, and then handle it appropriately. When I'm exploring data and relationships between elements, I will create a simple range-checking filter to handle outliers like the ones we've seen in this article. The filter will check each value in the column versus the range (lower and upper limits on the value) that I specify. If a value is out-of-range, then the outlying element is filtered – that is, it is ignored for any subsequent calculations, such as correlation. It is also excluded from any scatter plots or other charts.

Why Bother?

You might wonder, "Why bother with all that? Why not just delete the record?" Here's why: By simply deleting an anomalous record, or several, you have not created any capability to repeat the appropriate action on a future data file. In my own work, it is common to receive multiple iterations of a data set. We need a reproducible means of processing, exploring, analyzing, and modeling these variants of a data set.

For another reason, there may be a problem with only one element in the record. The other elements may be good and useful for analysis. You don't want to throw away any data that you don't have to.

If all that still isn't enough to convince you, please allow me to pile on one more. One day, your boss or client will ask you to do a little "quick-and-dirty" analysis. She may give you sincere, heartfelt assurances to the contrary, but chances are very good that you will work with that same data (or a variant) again for another "little project" – three, or six, or twelve months later. By then, you'll have long since forgotten what rows or elements you excluded from the prior analysis, and why. If you're very fastidious, and lucky, you might find some handwritten notes jotted on a legal pad that may help you remember the steps you took previously. Unfortunately, reconstructions will take at least as much time as doing the work from scratch – again. (Ask me how I know!) Or, you can encapsulate/document your original "quick-and-dirty" work in a handler, allowing you to breeze through any subsequent analysis[4].

Anything Else?

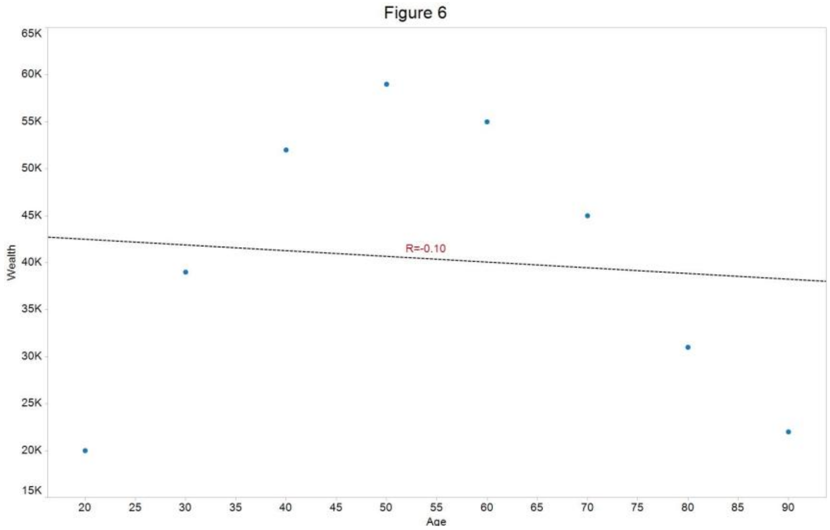


Figure 6 - Nonlinear relationship is not detected by correlation

There is one more common situation where the correlation measure can fail. Figure 6 provides an example where the data elements are clearly related, but the correlation between them is nearly zero. Why zero? Recall that correlation measures how well, or poorly, a straight line conforms to a pair of data elements. In Figure 6, the elements have a nonlinear relationship that is fit best by a curved line.

In contrast to the previous problematic examples, the issue here is not outlying values. Rather, it is the wrong assumption that there is a straight-line relationship between the elements. For my money, the best way to detect and handle nonlinear patterns is one we noted earlier … join in any time now … yes, to LOOK AT THE DATA. Using a scatter plot, you can spot a nonlinear curve in a heartbeat.

Wrap Up

In this "Nuts & Bolts" article, we focused on two mainstays of data mining and exploratory data analysis: correlation and scatter plots. We saw that straight-line relationships can be quantified nicely by the correlation measure.

Outliers, however, can severely impact the correlation calculation. In certain instances, a single outlier can suggest a strong relationship where none exists. In other instances, an outlier can destroy both the sense and quality of a truly strong relationship. In the article, I stressed the importance of building software "handlers" to detect and limit the impact of such outliers.

Lastly, we saw that correlation is ineffective for discovering nonlinear relationships. In contrast, the scatter plot really shines in this area. With it, you can actually SEE the nonlinear relationship. So remember, when you're exploring, analyzing, and modeling - LOOK AT THE DATA.

Tim Graettinger, Ph.D., is the President of Discovery Corps, Inc. (http://www.discoverycorpsinc.com), a Pittsburgh-area company specializing in data mining, visualization, and predictive analytics.

Your comments and questions about this article are welcome. Please contact Tim at (724)-743-3642 or tgraettinger@discoverycorpsinc.com

[1] Exactly this problem occurred in my own work about two months ago.
[2] The data entry error is a misplaced decimal point. In my work with survey data, such errors occur regularly.
[3] The piece of software might be a function written in code, or it might be an element with user-specified parameters in a data flow diagramming tool.
[4] Do what you think is best. You know where I stand.