Paper SAS2227-2018

Back to Basics: Get Better Insights from Data

Atrin Assa, SAS® Institute Inc.

ABSTRACT

Get the most out of your data, big or small. There is more data than ever, but more data doesn't always mean better insights. If you're not careful, more data can sometimes lead you down the wrong path. The worst thing that you can do is find patterns that aren't really there. SAS® Visual Analytics gives you the tools to avoid these types of problems. With simple drag-and-drop functionality, you can explore your data. You can see the shape of your data. You can see potential relationships in your data. You can get a grasp of the power and limits of your data. With that understanding, you can get better insights. It's time to move beyond just pretty visualizations, and this paper shows you how.

INTRODUCTION

Put down that average. Is that really what you want to use? Most of us have been through a basic statistics course at some point. Those of us who haven't are at least familiar with those basics. Well, hopefully, we have. So why do we still see fundamental mistakes in analysis? These mistakes cost organizations time and often huge sums of money. Sometimes, analysts pay for these errors with their jobs. The fundamentals aren't intuitive because we are not statistical creatures. So we learn them. Then, as time goes on, we start to forget as our attention moves on to other aspects of reporting and analysis. In this paper, we will look at a few fundamentals that will help you get better insights from your data. The topics we'll discuss in this paper aren't about advanced analytical methods. They're about the basics and how SAS Visual Analytics can help you get back to those basics. They apply to you whether you're building a simple sales dashboard made of bar charts and pie charts or trying to predict political instability. Getting these basics right will help you get better insights from your data.

LOOK AT YOUR DATA

It's mind-boggling and at the same time completely understandable how little time people spend visualizing their data. You've got data, you need to make something useful out of it, so you jump as quickly as you can into producing the final product, whether it's a report, dashboard, or model. Time is precious. You're under pressure to get to an answer fast. You might feel more like a producer than an analyst. Yet you might find yourself neglecting the most critical part of the analysis: graphing your data.

With the incredible array of analytical tools available to you, it's easy to get excited about the more advanced tools and forget the basics. The basics take time, and you rarely have much of it. SAS Visual Analytics not only gives you an incredible array of tools, but it also makes it easy to do the basics: like graphing your data.

Why is this important? Well, consider the following data set, put together by Francois Anscombe (Anscombe 1973).

DATASET 1		DATASET 2		DATASET 3		DATASET 4	
x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Table 1. Anscombe's Quartet

You have four sets of x and y data. Some of you might start by looking at summary statistics for these variables.

You can do this quickly with SAS Visual Analytics, using measure details. (See Figure 1.)

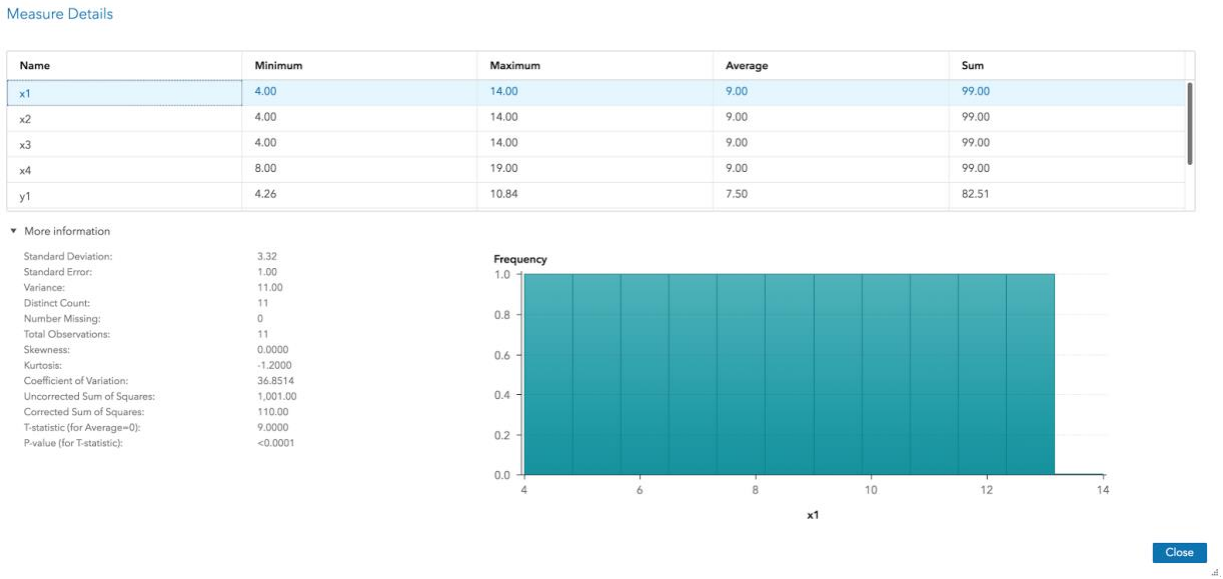

Figure 1. The Measure Details Dialog in SAS Visual Analytics

When you look at the summary statistics, you'll see that all the x variables and all the y variables tend to be remarkably similar—so similar in fact, that they are identical. x1, x2, x3, x4 all have the same standard deviation of 3.32. They also have an equal variance of 11.

You can even put these variables into a correlation matrix and see that all four x/y variables have the same strong correlation (around 0.82, as seen in Figure 2). In fact, if you ran a linear regression on each of the four data sets, you would get back the exact same regression line.

Figure 2. Correlation Matrix of Anscombe's Quartet

It's only when you just build a simple scatter plot of each data set that you start to realize how different these data sets are from one another, as seen in Figure 3.

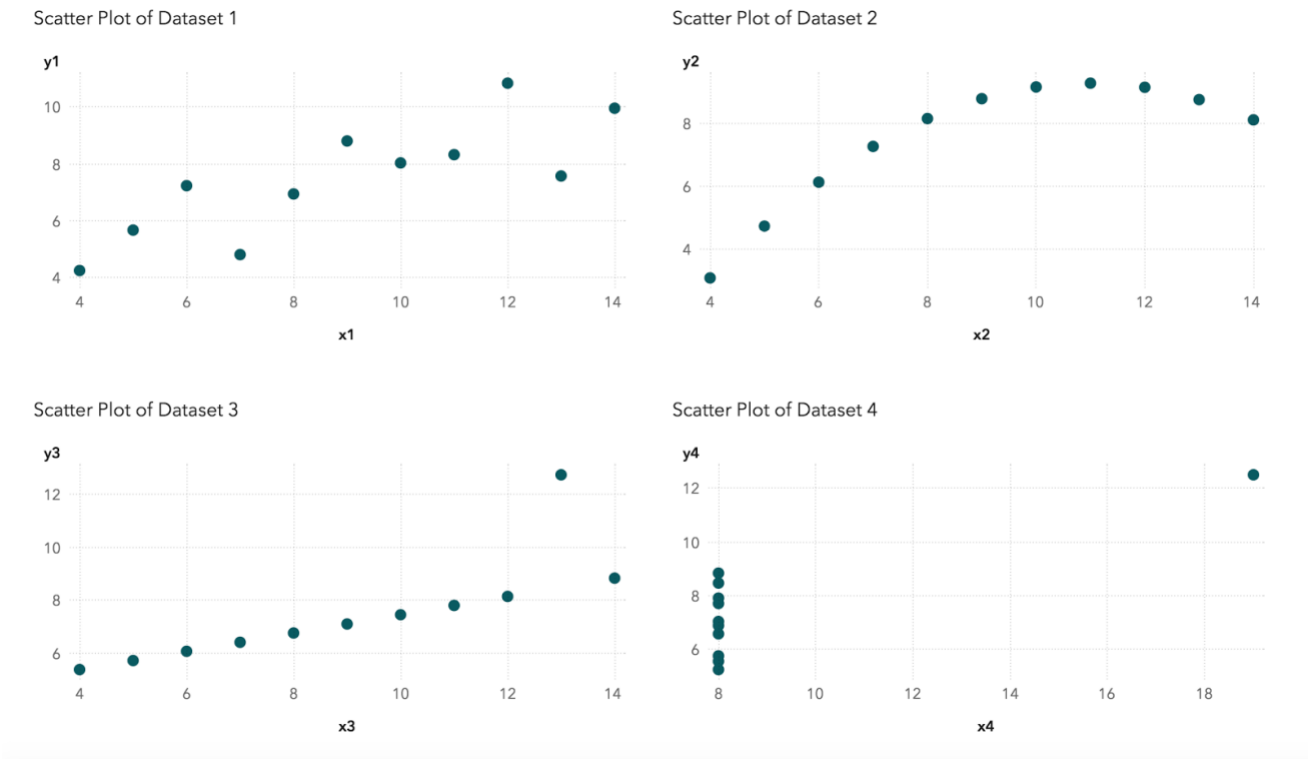

Figure 3. Scatterplots of Anscombe's Quartet

The single most powerful thing you can do is to plot your data and look at it. It is the fundamental tool of any activity with data. It's the best way to grasp the power and limitations of all the other means that you have available to you. If you don't look at your data, every tool you use beyond visualization can and will mislead you, your organization, and beyond.

RELY LESS ON AVERAGES

When you're pressed for time, it's better to take a broader approach to your data than a deeper one. Looking at select facets of your data is the fastest way to get misleading results. Averages, whether they are arithmetic means or medians, are commonly used summary facets of quantitative data. To take the most common example, if you've got sales figures for a year, you add them up and divide them by the number of observations, and that's your average.

Looking at data through averages causes a fundamental problem you need to consider. By their nature, averages distill hundreds, thousands, or billions of numbers (first-order information) and the variations between those numbers (second-order information) down to a single value. That means that you're likely to have potential information loss. You need to keep this in mind whenever you use an average. Is the information missing when you communicate an average valuable?

Imagine you're in charge of designing the thermal controls of a space station where astronauts will live. What information would be more useful for you? That the average temperature up there is 0ºF? Or that in direct sunlight the station heats up to 250ºF, and in the shade, the temperature is a frosty -250ºF (Price 2001)? In this case, the variation between the measurements is much more valuable to you than the average. If you design and engineer the thermal controls using the arithmetic mean, you'll have dead astronauts.

Think of averages less as a summary of the data and more as an obfuscation of your data. Whenever you use an average, you should stop and ask yourself: what information am I losing with this average? In general, it's better to communicate information another way. Even when the variations in your data aren't as extreme, these differences can have a disproportionate effect on your problem space.

One way to reduce the obfuscation is to take averages at lower levels in your data. The more local the average, the less you obfuscate information. In our space station example, an average for "in shadow" measurements and "under sunlight" measurements hides less information than an average of both shadow and sunlight measurements.

With a diverse set of visualizations, SAS Visual Analytics makes it easy for you to go beyond communicating information with averages. Your dashboards, your reports, or even your explorations and analyses can convey the richness of your data without obfuscating critical information.

UNDERSTAND THE SHAPE OF YOUR DATA

The distribution of your data is vital to identifying the power and limitations of the tools in front of you. Yet, it's easy to open up some data and to run analytical and statistical methods on it without ever checking or understanding the distribution of that data.

SAS Visual Analytics provides several tools to help overcome this instinct. Once again, the ability to view measure details (Figure 4) on any data set lets you quickly start to look at the shape and distribution of your data.

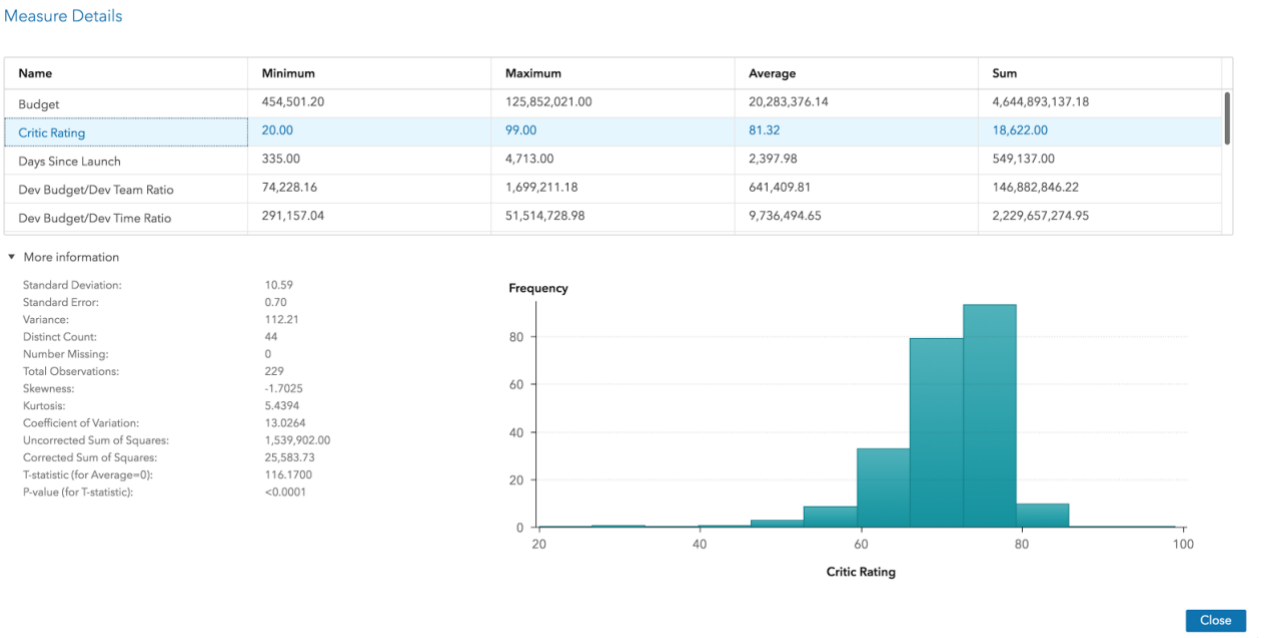

Figure 4. Measure Details in SAS Visual Analytics

The measure details dialog gives you one place to see a wealth of information about your quantitative variables or measures. You can see their minimum and maximum values, their averages, their sums, basic descriptive statistics, and best of all a histogram of the measure, as seen in Figure 5. This histogram can help you see how your measures are distributed. Are they normal? Skewed? Do you have outliers that might be of concern?

Information like kurtosis, found under more information (Figure 6), tells you how likely the data is to have outliers. So, for example, if we look at video game data revenue data, not only do we see a histogram with outliers, but also high kurtosis.

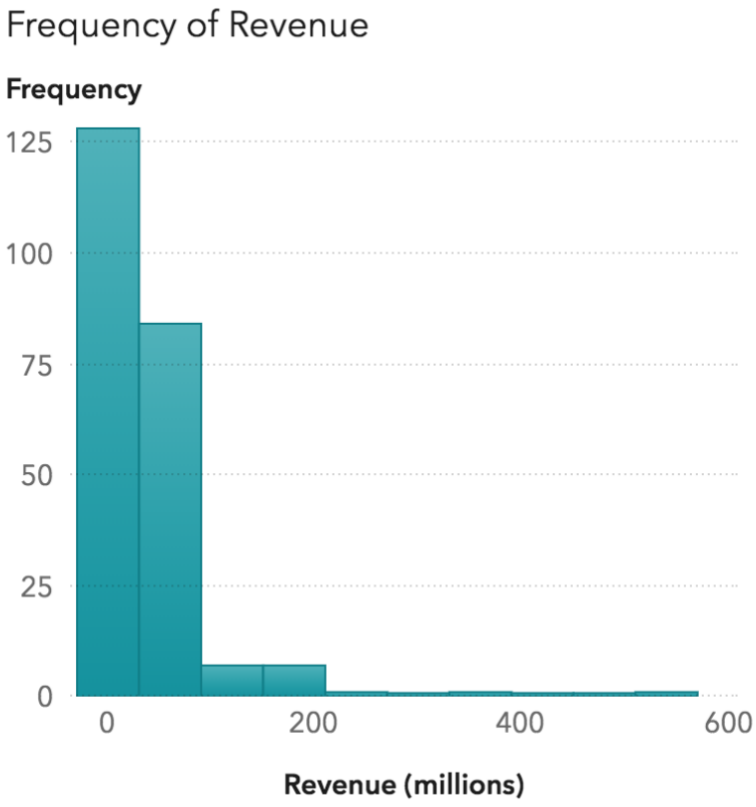

Figure 5. Histogram of Revenue

▼ More information

Standard Deviation:	55,048,581.94
Standard Error:	3,637,712.69
Variance:	3,030,346,373,928,910.00
Distinct Count:	229
Number Missing:	0
Total Observations:	229
Skewness:	4.9398
Kurtosis:	35.7871
Coefficient of Variation:	133.4897
Uncorrected Sum of Squares:	1,080,351,407,432,960,000.00
Corrected Sum of Squares:	690,918,973,255,791,000.00
T-statistic (for Average=0):	11.3363
P-value (for T-statistic):	<0.0001

Figure 6. More Information from Measure Details on Revenue

High kurtosis makes sense for the video game business. We could publish 100 video games, and most might not make much money. Every few years, one or two of our games could end up being big hits. These hits can be so big, that a single title could make more money than most or even all the other titles released combined.

SAS Visual Analytics also lets you create histograms as part of your report so that you can make them part of what you communicate to report consumers. While you're exploring, you even get fine-tuned control over your histograms. SAS Visual Analytics will try to find the best binning of the histogram. But if you want to tinker, you can adjust the size of the bins and see if that gives you a different perspective, all with a real-time, point-and-click interface.

The distribution of your data will determine what tools make sense for data. For example, when you have a normal distribution (the symmetric bell-curve we all know and love), it makes sense to use a measure of center like an average. But what if you have a distribution that looks like multiple peaks as seen in Figure 7?

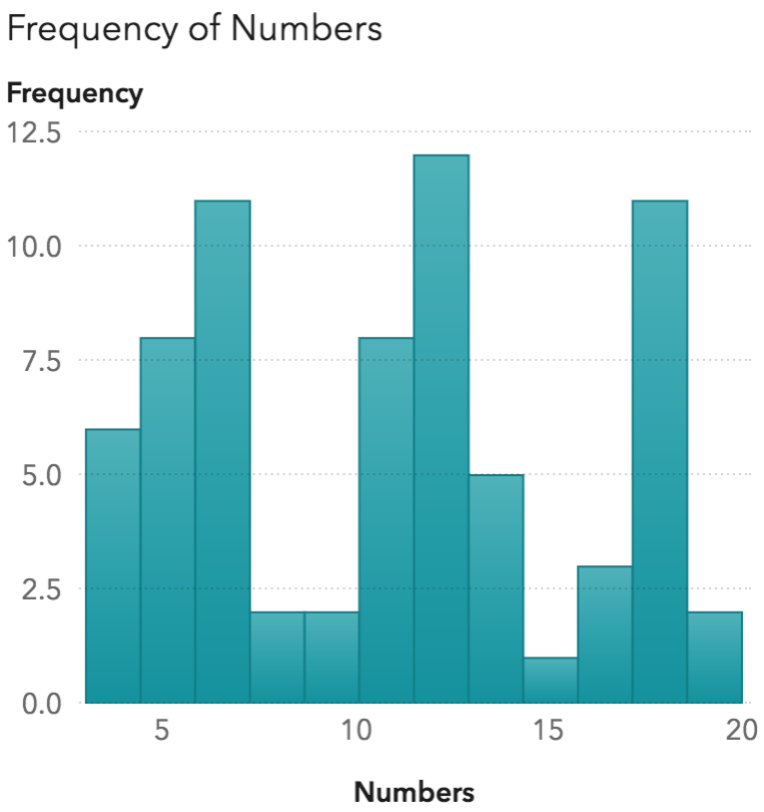

Figure 7. Multi-peaked Distribution

The center isn't as useful to us in this case, and if we used it, we could be getting and communicating the wrong "insight" from our data.

The shape of your data is essential to how you analyze and present insights about that data. With SAS Visual Analytics you have powerful tools to help you understand the shape of your data, so use them often and use them well.

MAKE HONEST GRAPHS

One of the simplest things you can do to ensure that your insights are useful and valuable is to provide accurate and honest graphs. In the realm of data visualization, it can be easy to mislead with minor tweaks to how you represent data.

SAS Visual Analytics helps you avoid these pitfalls in three ways. First, auto-charting enables you to get the best visualization based on your data. Second, visualizations are built by default to be as accurate as possible. Finally, a wealth of visualization options let you visualize the same data multiple ways so that you can get various perspectives on your data.

Whenever you drag columns of data to the canvas, SAS Visual Analytics assesses the columns and chooses a first visualization that makes sense for that data. So if you select revenues and time, you'll get a time series plot. And if you choose seven different measures, you'll get a scatterplot matrix. This can help you take the guesswork out of your visualization process.

Once you've made your visualizations, you'll notice a couple of things. SAS Visual Analytics ensures that every visualization is as representative as it can be of your actual data. For example, when you have a measure axis, you can always make sure that the axis will start at 0. This keeps your insights from being distorted by misleading axes scales as seen in Figure 8. When it comes to absolute versus relative frequencies, SAS Visual Analytics errs on frequencies rather than frequency percentages. It'll be up to you to decide how you want to show that information.

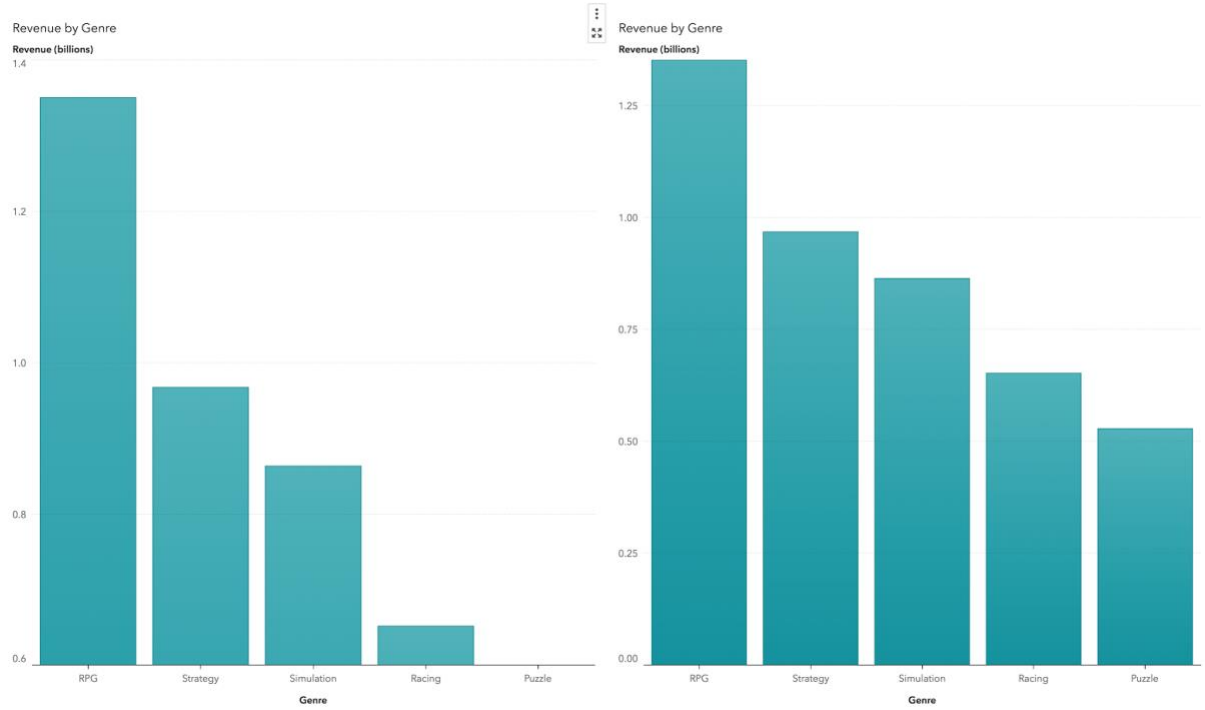

Figure 8. Same Data, Different y-axis Minimums

Finally, SAS Visual Analytics provides you multiple ways to visualize data. You can change a visualization to another one that might offer a different perspective at any time, like in Figure 9. You can duplicate the chart and change to another visualization with one click.

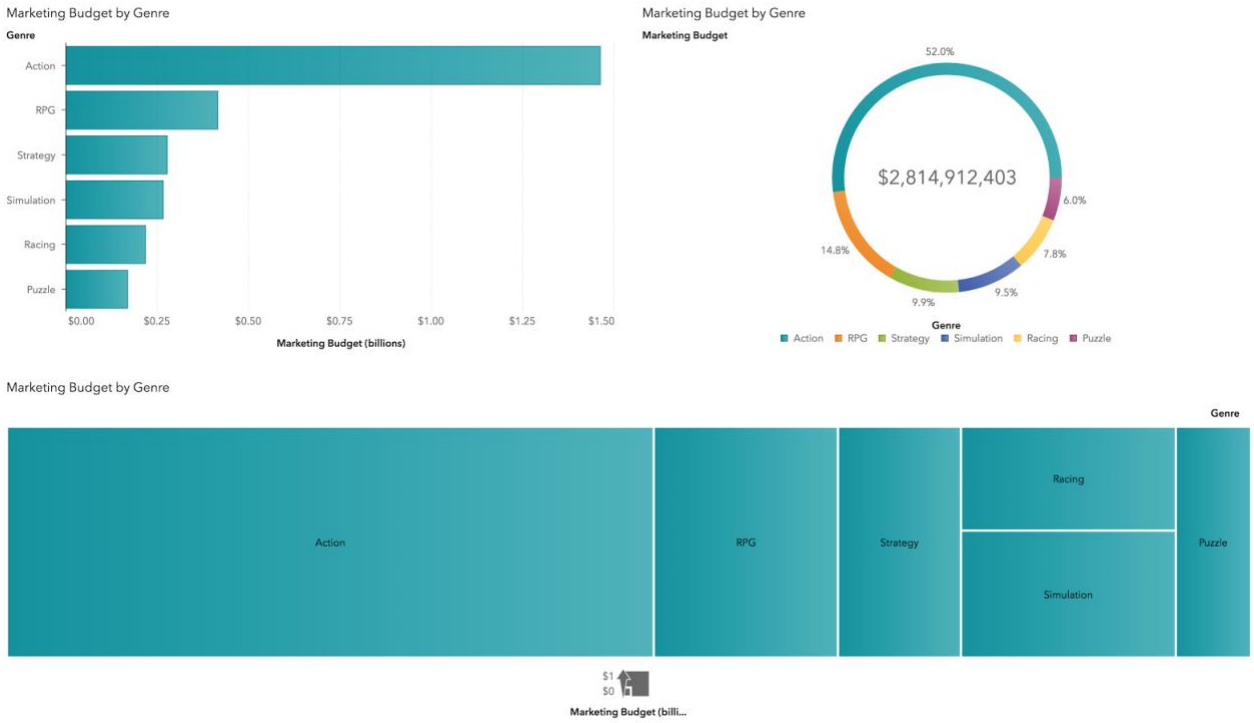

Figure 9. Multiple Representations of the Same Information

The multiple perspectives you can get from the same data with SAS Visual Analytics make it less likely that you misinterpret your data due to misleading graphs.

CONCLUSION

Getting insights from data isn't enough. We all want to get more reliable and accurate insights from our data. It's easy to think that more sophisticated methods and tools are the only way to get more accurate insights from your data. But you need to be careful. It's important to understand that just because you've used sophisticated tools and methods doesn't mean that your insights are accurate or even useful. The same is true for the simple tools. There is no absolute rule here: every tool at your disposal can lead you to misleading results if you forget the basics. The most sophisticated model is only as good as its underlying basic assumptions like validity and normality. A simple measure of center, like an average, can obfuscate crucial information. And even a bar chart can give you the wrong impression just by changing where the y-axis starts.

Luckily, SAS Visual Analytics gives you the power to get back to the basics. It puts you in control of your data and what you do with that data. It lets you see your data at the most basic level. It gives you the resources you need to decide when something like an average makes sense, by easily letting you understand the shape of your data before you get your insights. It gives you honest visualizations, from the most basic to the most sophisticated, so that you can get accurate and more reliable insights, while also being able to communicate them honestly with your organization.

REFERENCES

Anscombe, F. J. 1973. Graphs in Statistical Analysis. *The American Statistician*, 27(1), 17. doi:10.2307/2682899.

Price, S., T. Phillips, and G. Knier. 2001, March 21. Staying Cool on the ISS. Retrieved December 10, 2017, from https://science.nasa.gov/science-news/science-at-nasa/2001/ast21mar_1/.

RECOMMENDED READING

- *How to Lie with Statistics*
- *Statistical Models: Theory and Practice*

CONTACT INFORMATION <HEADING 1>

Your comments and questions are valued and encouraged. Contact the author at:

Atrin Assa
SAS Institute Inc.
Atrin.Assa@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.