Bring Analytics into Your Day Job

Atrin Assa, SAS Institute Inc., Cary, NC

ABSTRACT

If you rely on data but aren't analytics oriented, you're missing out on insights. Even worse, you could be sabotaging the decision-making that builds on insights from data. In this paper, we explore how you can bring analytics and analytical thinking into your decision-making by using SAS® Visual Analytics. We examine everything from smart, automated tools that can help you find insights within your data to simple charts and rules that can help you determine the usefulness of your insights.

INTRODUCTION

If you work in a modern organization, then you probably rely on data to complete some aspect of your job. You might depend on data about the number of bugs in software. You might rely on information about website visits and engagement. You might need reliability statistics about an engine. The list goes on. But can you get deeper, richer insights from that data? Can you unlock the secrets in the data that will help you make a difference? There are no guarantees. Some data can be taken at face value, while other data needs a bit of digging to find nuggets of insight. Yet the broader array of tools you apply, the more likely you are to discover hidden knowledge, should there be any. Analytics can give you that power. In this paper, we'll discuss some of the simplest yet most powerful analytic tools in SAS Visual Analytics that will help you see your data in new and deeper ways. If you know how to use these tools, you can quickly find areas to focus on in your data.

THE HISTOGRAM

The histogram (Display 1) is an under-appreciated visualization and analytical tool. It looks like a bar chart, but it gives you meta-insight about your data. It tells you the shape or distribution of your data. If you look at a measure like "Age" on a histogram, you can see the distribution of age for the people in your data. Each bar of the histogram shows you the count of data for that bin.

Understanding the shape of your data is important to how you'll make decisions based on that data. For example, if you've got normally distributed data, that is your histogram forms a neat bell curve with no extreme values, you can work with a variety of statistical techniques.

But if your histogram looks skewed (most of the bars fall on one side) or the ends of your bell curve seem to extend into the extremes, then it means you need to be careful how you analyze that data. Even a simple average of your measure might like "Age" might be misleading under these circumstances.
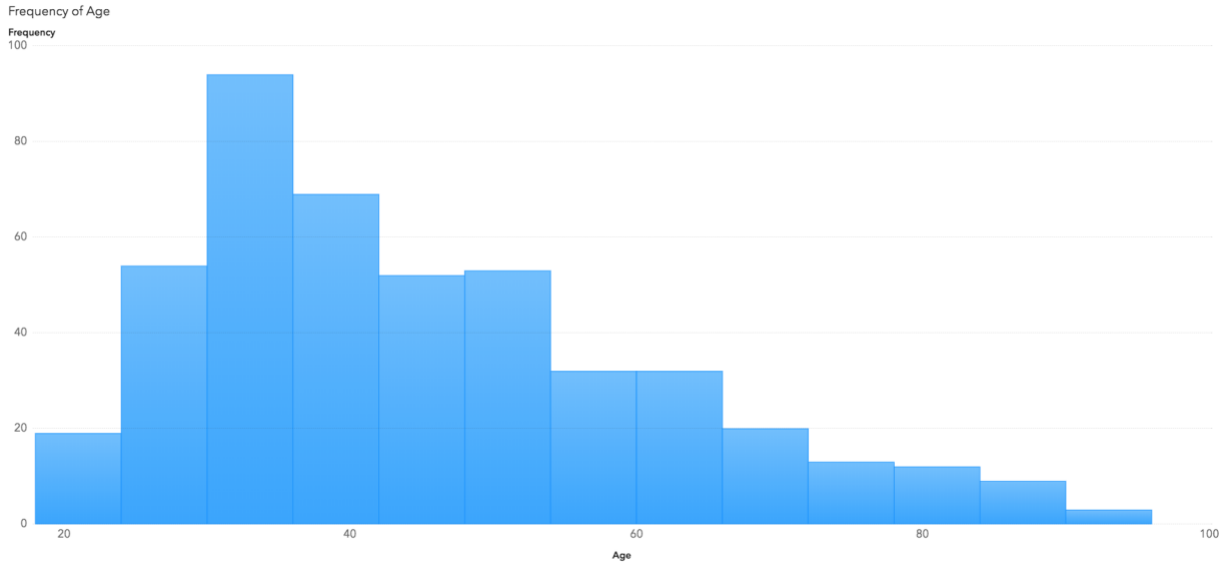
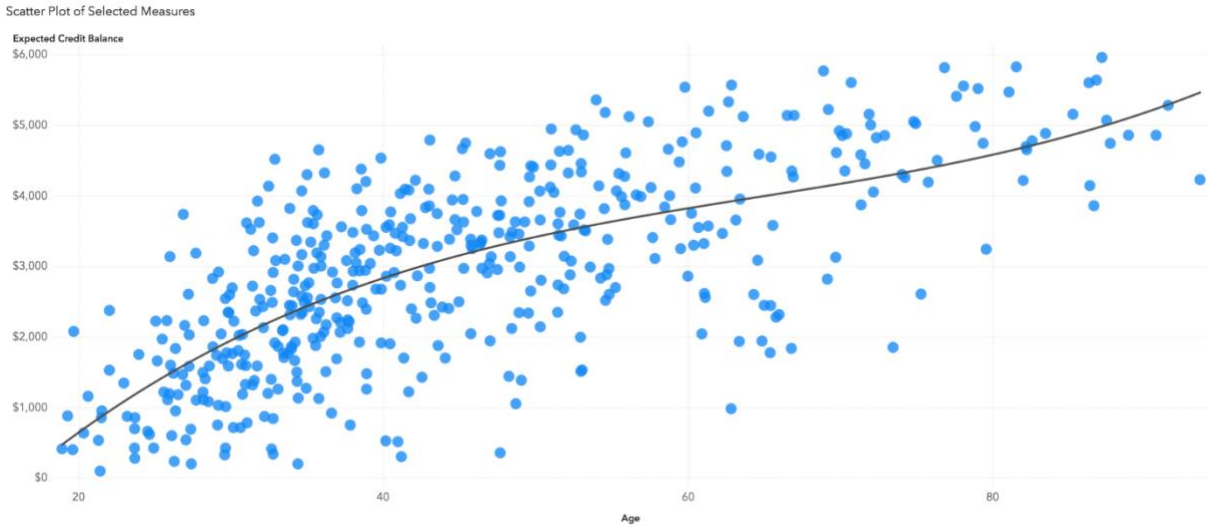**Display 1. A histogram in SAS Visual Analytics**

Consider this example: human height is usually a normally distributed variable. Most humans will have a height that falls somewhere near the middle of the distribution. The shortest humans are still fairly close to the center of that distribution. As are the tallest humans. That's why we don't see ant-sized humans or people the height of skyscrapers. You could take an average of people's heights and say most people would be somewhere near that average.

Suppose instead that you're a Lemonade Stand Conglomerate owner and trying to understand how much money each of your 50 lemonade stands earns. Most of your lemonade stands might only make a couple of hundred dollars daily. A couple of your lemonade stands, in very high traffic areas, could earn tens of thousands of dollars per day. These two lemonade stands will make your average revenue per lemonade stand look a lot bigger than it really is. But you can't remove them from your analysis, because these stands might also generate the bulk of your business revenue.

Understanding the distribution of your data helps you see the numbers and how they affect your organization in new ways. Histograms help you understand that distribution. That understanding can change your approach to making decisions.

THE SCATTER PLOT

A scatter plot (Display 2) is a quick way to see the shape of your data. Scatter plots visualize two measures against each other. In this way, you can see if the measures have any patterns together. For example, you might see that as age increases, write-off risk for credit cards decrease. Other times scatter plots might look completely random like someone spray painted points on the plot. That tells you that your two measures are unlikely to have a relationship. Keep in mind, however, that a relationship on a scatter plot doesn't mean there is an actual causal relationship (that is one measure causes the change in the other measure). It could be a fluke. The fewer rows of data you have, the more careful of this you need to be. Even two randomly generated columns can occasionally seem to have a relationship. But, we know for a fact that they were just randomly generated.

Display 2. A scatter plot in SAS Visual Analytics

If your data is too large to use a scatter plot, the auto-charting feature in SAS Visual Analytics will automatically select a heat map for you. The heat map bins your points together into a more readable format for larger data sets.

THE CORRELATION MATRIX

One quick way to find relationships in your data is through a correlation matrix (Display 3). This object in SAS Visual Analytics analyzes the measures you select and finds their correlation. Once again, as with a scatter plot, these correlations do not imply causation. Correlations can be negative or positive. If the correlation is negative, then as one measure increases, the other decreases. If the correlation is positive, then as one measure increases, the other also decreases.
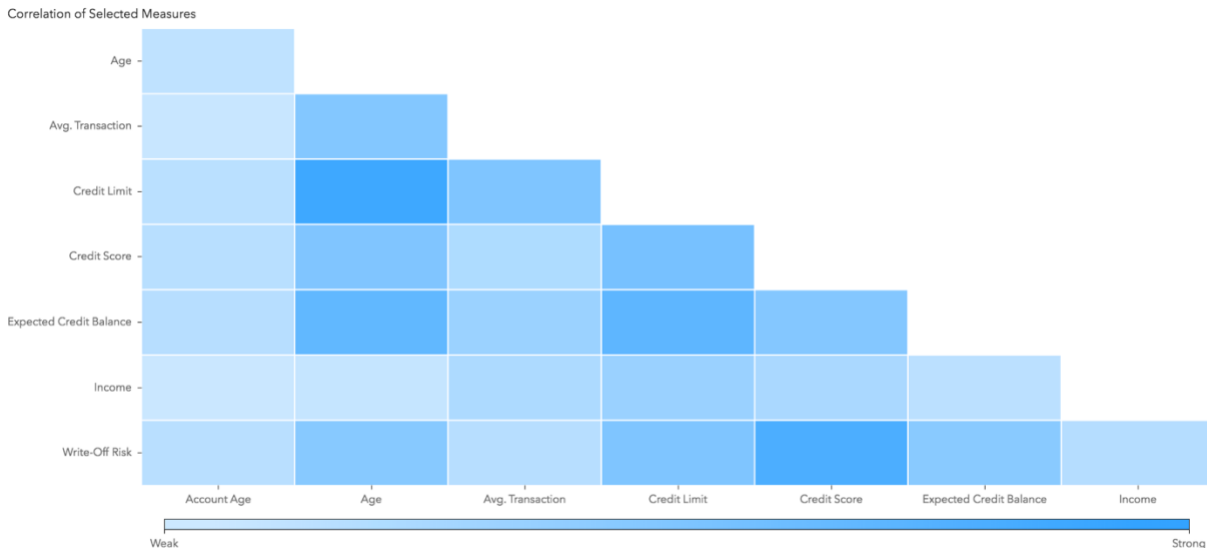
Display 3. A correlation matrix in SAS Visual Analytics

The correlation matrix quickly highlights these relationships with a heatmap style between all the measures you select. The darker the color you see, the stronger the correlation, whether negative or positive.

Once again, if you have a small data set or a lot of columns, you might see correlations where there really aren't any. You'll need to apply your domain knowledge and expertise to make sure the associations you see are genuine. For example, if you only have 20 rows of data and 20 columns, it's not unusual to see moderate correlations between even randomly generated variables, so be careful.

If you find intriguing correlations, use scatter plots to check if the relationship is really linear, roughly following a line, or if there is something else in the data, for example, an extreme value, that's causing a correlation.

AUTOMATED ANALYSIS

When you first start an analysis, it can be daunting. You might have a data set with hundreds of columns and a blank canvas to start with. Where do you begin? What should you look at? You might know which variables you're interested in, but how do you visualize them? What other variables should you look at?

Remember that we want to look at our data from as many perspectives as possible. That means being paralyzed by the blank canvas is only going to limit our viewpoint. Automated analysis (Displays 4 and 5) in SAS Visual Analytics 8.3 can help with that.

Suppose you're interested in a variable called "Revenue". You'd like to see how the other variables in your data set are driving revenues. You can go about manually creating charts. But if you've got hundreds of columns of data, it's going to be time-consuming. That's where automated analysis comes in.

Just open the data pane in SAS Visual Analytics, right-click the column you're interested in, and analyze it, either on the current page or on a new page. Behind the scenes, SAS Visual Analytics analyzes your data with techniques like decision trees and within seconds gives you results. The column you examine can be a measure or a category.

The resulting dashboard lets you quickly see some vital information about how your data relates to the variable you're interested in. If you've analyzed a measure, you'll get natural language overviews of things like the minimums, averages, maximums, as well as which predictors best explain your variable of interest. If you've analyzed a category, you'll get an explanation of how your data fits into that category, and which predictors best explain the column you're interested in.

7.7%   Average **Write-Off Risk** of 462 cases is 7.7%, with minimum 0.00% and maximum 15.00%. Credit Score, Credit Score Group, Credit Limit, Age, Expected Credit Balance, Age Group and twelve other predictors best explain **Write-Off Risk**.

| Credit Score | Credit Score Group | Credit Limit | Age | Expected Credit Balance | Age Group | State | Income | | | | | |

Groups by Average Value of Write-Off Risk

15%   If Credit Score Group is Poor or Fair, Account Age is 1, 2, 5, 6, 8, or 10 , Rewards is Travel, Avg. Transaction is less than $55, State is Alaska, Colorado, Illinois, New Hampshire, or Wisconsin , then the 6 cases have a high predicted Write-Off Risk of 15%.

15%   If Credit Score Group is Poor or Fair, Account Age is 3, 4, 7, or 9 , Avg. Transaction is less than $96, Rewards is None, State is Idaho, Michigan, New Hampshire, or Wisconsin , then the 5 cases have a high predicted Write-Off Risk of 15%.

14%   If Credit Score Group is Poor or Fair, Account Age is 1, 2, 5, 6, 8, or 10 , Rewards is Travel, Avg. Transaction is less than $55, State is Louisiana, Maryland, New York, or Oklahoma , then the 4 cases have a high predicted Write-Off Risk of 14%.
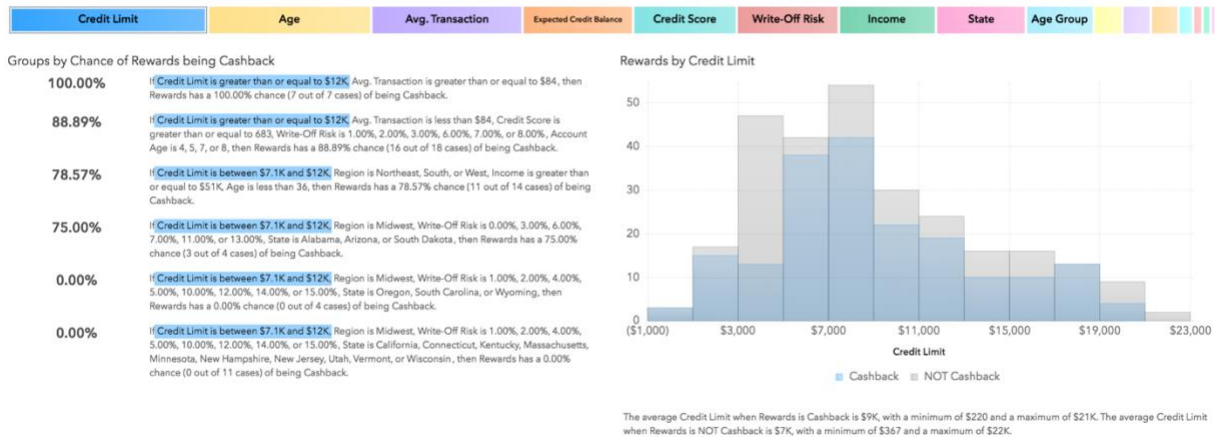
14%   If Credit Score Group is Poor or Fair, Account Age is 1, 2, 5, 6, 8, or 10 , Rewards is Cashback or None, Region is Midwest, Northeast, or West, Age is less than 30, then the 17 cases have a high predicted Write-Off Risk of 14%.

1.5%   If Credit Score Group is Good, Very Good, or Exceptional, Credit Score is greater than or equal to 773, Account Age is 1, 3, 4, 5, 6, 7, or 8 , Income is greater than or equal to $60K, then the 15 cases have a low predicted Write-Off Risk of 1.5%.

.79%   If Credit Score Group is Good, Very Good, or Exceptional, Credit Score is greater than or equal to 773, Account Age is 1, 3, 4, 5, 6, 7, or 8 , Income is less than $60K, then the 24 cases have a low predicted Write-Off Risk of .79%.

Write-Off Risk by Credit Score

Write-Off Risk and Credit Score have a strong negative relationship.

Display 4. Automated analysis of a measure in SAS Visual Analytics

| | Cashback | None | Travel |

40.91%   Overall, **Rewards** has a 40.91% chance (189 out of 462 cases) of being **Cashback**. Credit Limit, Age, Avg. Transaction, Expected Credit Balance, Credit Score, Write-Off Risk and twelve other predictors best explain **Rewards**.

| Credit Limit | Age | Avg. Transaction | Expected Credit Balance | Credit Score | Write-Off Risk | Income | State | Age Group | | | | |

Groups by Chance of Rewards being Cashback

100.00%   If Credit Limit is greater than or equal to $12K, Avg. Transaction is greater than or equal to $84, then Rewards has a 100.00% chance (7 out of 7 cases) of being Cashback.

88.89%   If Credit Limit is greater than or equal to $12K, Avg. Transaction is less than $84, Credit Score is greater than or equal to 683, Write-Off Risk is 1.00%, 2.00%, 3.00%, 6.00%, 7.00%, or 8.00%, Account Age is 4, 5, 7, or 8 , then Rewards has a 88.89% chance (16 out of 18 cases) of being Cashback.

78.57%   If Credit Limit is between $7.1K and $12K, Region is Northeast, South, or West, Income is greater than or equal to $51K, Age is less than 36, then Rewards has a 78.57% chance (11 out of 14 cases) of being Cashback.

75.00%   If Credit Limit is between $7.1K and $12K, Region is Midwest, Write-Off Risk is 0.00%, 3.00%, 6.00%, 7.00%, 11.00%, or 13.00%, State is Alabama, Arizona, or South Dakota, then Rewards has a 75.00% chance (3 out of 4 cases) of being Cashback.

0.00%   If Credit Limit is between $7.1K and $12K, Region is Midwest, Write-Off Risk is 1.00%, 2.00%, 4.00%, 5.00%, 10.00%, 12.00%, 14.00%, or 15.00%, State is Oregon, South Carolina, or Wyoming, then Rewards has a 0.00% chance (0 out of 4 cases) of being Cashback.

0.00%   If Credit Limit is between $7.1K and $12K, Region is Midwest, Write-Off Risk is 1.00%, 2.00%, 4.00%, 5.00%, 10.00%, 12.00%, 14.00%, or 15.00%, State is California, Connecticut, Kentucky, Massachusetts, Minnesota, New Hampshire, New Jersey, Utah, Vermont, or Wisconsin , then Rewards has a 0.00% chance (0 out of 11 cases) of being Cashback.

Rewards by Credit Limit

■ Cashback ■ NOT Cashback

The average Credit Limit when Rewards is Cashback is $9K, with a minimum of $220 and a maximum of $21K. The average Credit Limit when Rewards is NOT Cashback is $7K, with a minimum of $367 and a maximum of $22K.

Display 5. Automated analysis of a category in SAS Visual Analytics

Below the explanation, you'll see the variables you probably should investigate further displayed on a colorful strip. They are presented in order of importance. Below the strip, you'll see any groups identified within your data. When you select variables in the colorful strip, you'll see a visualization of how that particular variable ties into your variable of interest.

Automated analysis is a great, quick way to identify patterns in your data and decide what to focus on, especially when you have large amounts of data with a large number of variables. It's an easy way to quickly cover a lot of ground with the data and see multiple

variables and how they tie into your variable of interest, and that means you're less likely to miss insights within your data.

CONCLUSION

You don't need to be a statistician to bring some analytics to your day job. What you do need is some intuition and knowledge of the domain you're working in, and some simple, yet powerful tools that can help you along. SAS Visual Analytics provides some easy-to-use analytic tools like histograms, scatter plots, correlation matrices, and the new automated analysis feature that help you broaden your view of your data. With these tools you can quickly find patterns in your data and find insights that you might have otherwise missed.

RECOMMENDED READING

- Freedman, David. 2009. *Statistical Models: Theory and Practice*. 2d ed. Cambridge, United Kingdom: Cambridge University Press.

CONTACT INFORMATION

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

Atrin Assa
Atrin.Assa@sas.com