

Paper 200-2016

Make the Jump from Business User to Data Analyst in SAS® Visual Analytics

Ryan Kumpfmiller, Zencos Consulting

ABSTRACT

SAS® Visual Analytics is effective in empowering the business user with the skills to build reports and dashboards. The tool is easy to use and navigate, but it also has capabilities that go beyond just presenting data. There are additional data analysis features, such as forecasting, fit lines, and correlations, which can give those business users better insight into their data. This paper is going to go into what each of those features are, how to interpret them, and what objects they are used with in SAS® Visual Analytics.

INTRODUCTION

Analytics comes from the intersection of business, technology, and statistics. Finding people talented in all three areas is rare, and more often than not, users come from one of these areas with an interest in the others. Self-service analytics tools try to bridge that gap by providing user-friendly software that helps overcome any lack of technical knowledge. Therefore, users coming from the business or statistical areas will not have to learn as much technical code before diving into their analysis. However, what about the business or technical users that may not be well versed on the statistical side?

Everyone knows how to set up a bar chart and line graph, but when you start to go beyond measuring a single category, things may not be as clear. Knowing what types of objects and analysis to use to display your data can be the difference between finding those key takeaways, which is what analytics is all about. On top of all of the objects that SAS Visual Analytics has, there are also data analysis tools such as forecasting, fit lines, and correlations. Each of these can be used within one or more of the objects and can add insight to the takeaways that users are looking to get when using this tool. While they all have underlying statistical calculations, SAS Visual Analytics makes them very easy to apply to the objects. In the following sections, we're going to explore those underlying calculations so that anyone from the tech or business side of analytics can better understand what these features do and then be able to apply the methods themselves.

FORECASTING

SAS Visual Analytics users can apply the forecasting feature to predict how their data trends into the future. Using data that contains a time frame, users can use the Forecasting option in the Explorer Line chart object that models the data to some upcoming time frame. However, using and understanding how it works is a little easier said than done. In this section, you will learn how the forecasting is done and how to use the Scenario Analysis option.

How does it work in SAS Visual Analytics?

Forecasting can only be done with a line chart in the Explorer section of SAS Visual Analytics. In the Roles tab of the line chart, there is an option for forecasting. The option is grayed out until a date item is added in the Category section. Once that is populated, then you can select the Forecasting option. When selected, a vertical line appears in the line chart dividing the ending date of the user's data and the beginning of the forecasting results.

As long as you have a date field and a measure, anything can be forecasted. Popular examples include sales, weather, and company performance. For this example, we are going to stick with the finance industry and look at the most analyzed company in the stock market, Apple. In Figure 1 is an example of the forecast using Apple's closing stock price at quarters end for the past ten years. (Data Source: <https://ycharts.com/companies/AAPL/>)

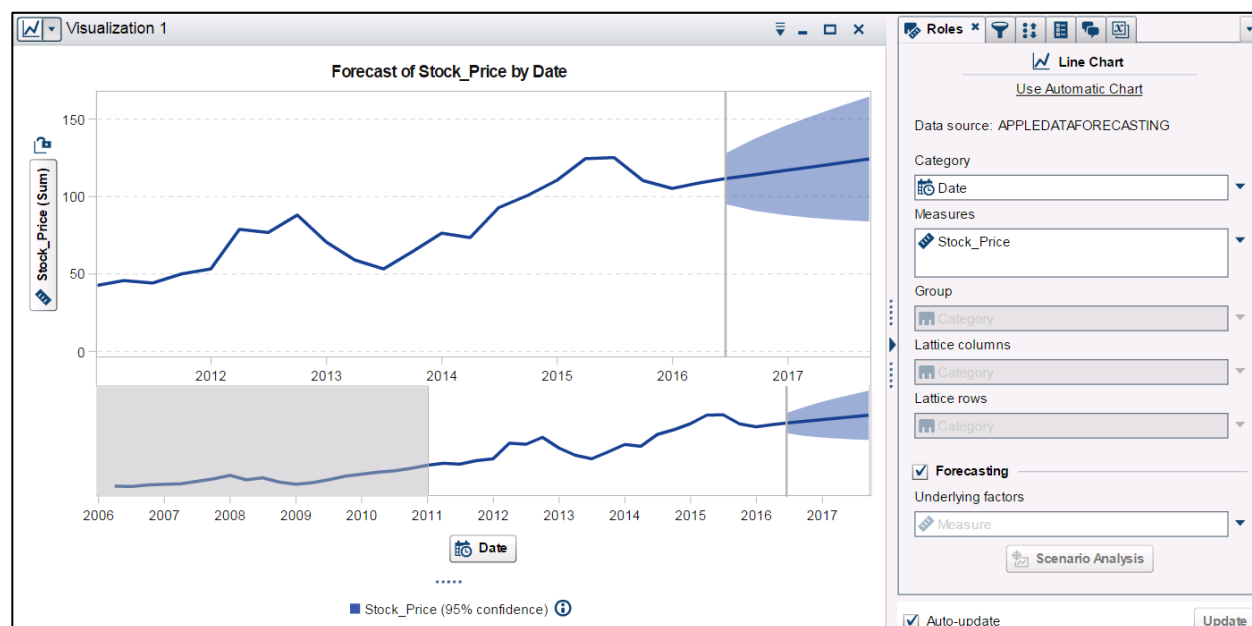


Figure 1 Forecast of Apple Stock Price

The data ends after quarter 1 of 2016 so the forecast starts at the end of quarter 2 which is the end of June where the gray vertical line is placed. The dark blue line in the forecast shows the most likely trajectory of the stock price and the blue shaded area is the confidence interval. By looking at the legend at the bottom, you can see that we are working with a 95% confidence interval. This means that the model projects a 95% chance that the future stock price will be somewhere in the blue shaded area.

For this example, the forecast is only going out to the next six quarters. This is called the forecast duration and can be changed with the confidence interval by going to the Properties tab. At the bottom of the tab, there is an option to change those values, shown in Figure 2.

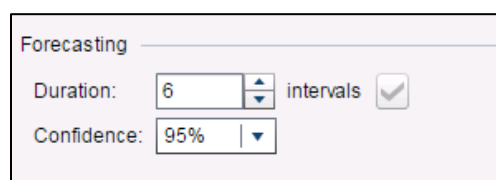


Figure 2 Forecasting Options

As you increase the forecast duration the confidence band typically expands since the further into the future you go there is more uncertainty. It's important to note here that models like these work better with as much data as you can give them. If you only have a few points, then the model is going to have a hard time coming up with accurate results.

How is the data modeled?

One of the best aspects of SAS Visual Analytics is that it enables a business user to harness the power of analytics. The forecast is an example of that since it is able to run your data through six different models and picks the one with the best fit. Here is a list of the different models available [4]:

- Damped-trend exponential smoothing
- Linear exponential smoothing
- Seasonal exponential smoothing
- Simple exponential smoothing
- Winters method (additive)
- Winters method (multiplicative)

As the data is modeled, the Root-Mean-Square-Error (RMSE) is calculated for each model behind the scenes. [1] The RMSE is a measure of how close the predicted values are to the real data. The lower the RMSE, the more accurate the model is. SAS Visual Analytics then selects the model with the lowest RMSE to use in the forecast.

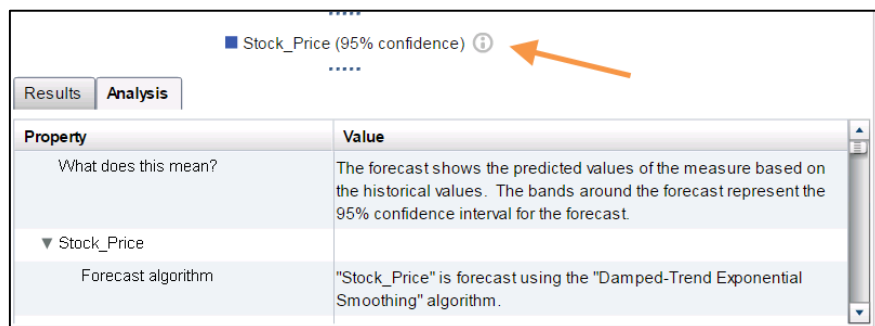


Figure 3 Forecast Details

After selecting the forecasting option, you can see which model was used as well as a table of the results by clicking on the (i) at the bottom of the line chart. Shown in Figure 3, the Damped-Trend Exponential Smoothing algorithm was selected for the forecast used in the first example.

Look for Underlying Factors

In order to improve our analysis, we don't just want to look at one historical measure and base the forecast on those values. There could also be other data points that might have an influence on that measure, and if they are incorporated then our model can become even stronger since it will have multiple variables incorporated.

The models that SAS Visual Analytics runs to build our forecast can also include other measures into the analysis. By going to the Underlying factors section in the Roles tab. By clicking the drop down, you can add one or more measures from your data set into the analysis. As with the original forecasting, SAS runs the data through the models, adding autoregressive integrated moving average models (ARIMA) to go with the original six, to determine the best fit. If the added measure does not have an influence on the model, then it will be grayed out. When the new measure does influence the model, the chart is updated with the results as shown in Figure 4.

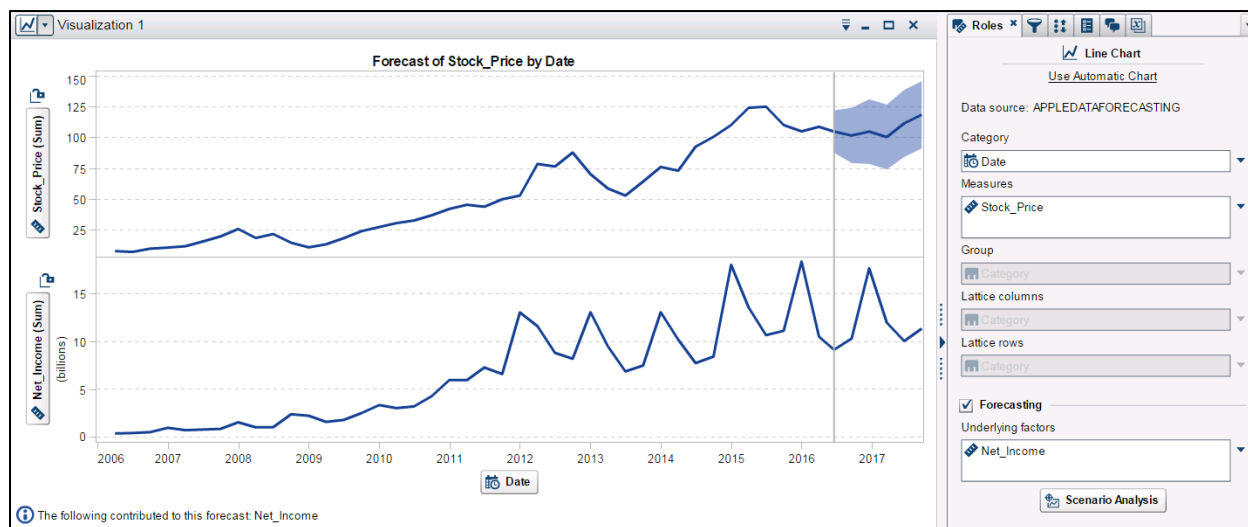


Figure 4 Forecast with Underlying Factors

Continuing our forecast example, adding Net_Income as a possible underlying factor, the forecasting has been updated with the results. The top chart is similar to our original forecast of Apple's stock price except now the forecasted section has improved. In our first run, Quarter 1 of 2017 had a 95% confident predicted stock price in the range of \$86.23-\$152.10. When using Net_Income as a factor, that confidence band is now narrowed to \$74.37-\$126.91, which is a notable reduction in the range.

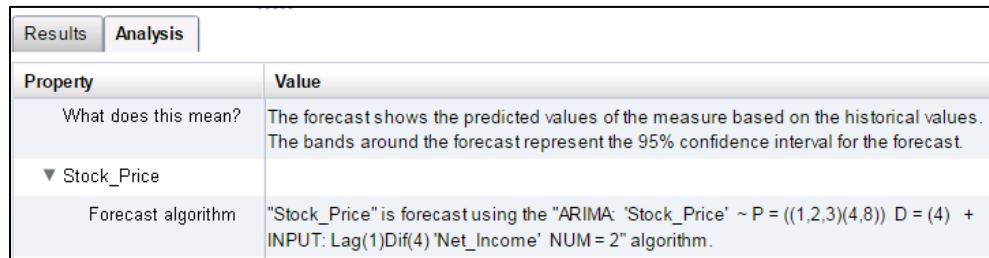


Figure 5 Forecast Details with Underlying Factors

A closer look at the bottom analysis section in Figure 5 and you can also see that the forecast used an ARIMA model as opposed to the Damped-trend that was used originally.

Using Scenario Analysis and Goal Seeking

Once you have found an underlying factor that influences the forecast, the Scenario Analysis button at the bottom of the Roles tab becomes available to use. After clicking on it, a window shows the forecasted data field and the underlying factor. There are two options for users to change, Goal Seeking and Scenario Analysis.

Figure 6 Advanced Forecasting Options

With Scenario Analysis, you can go in and manipulate the underlying factors and see how the forecast would change based on those new values. In our example, we envision that Apple is introducing a line of products this quarter and that those products are planned to drive net income up 50% for the foreseeable future compared to following the normal path. We can set this expectation by clicking on the Net_Income button on the left side of the screen and selecting “Set Series Values”. A window like the one shown in Figure 6 will pop up and this is where the values can be set with a fixed number, a numeric increment, or a percentage increase.

After selecting ‘OK’ the forecasted numbers for the Net_Income are updated with the 50% increase. There is a gray line in the underlying factor’s forecast section that indicates the original data points. Since the underlying factor has been altered, only Scenario Analysis is available to use and is the only option available in the right menu. When Apply is selected, the forecast is then updated with the new results.

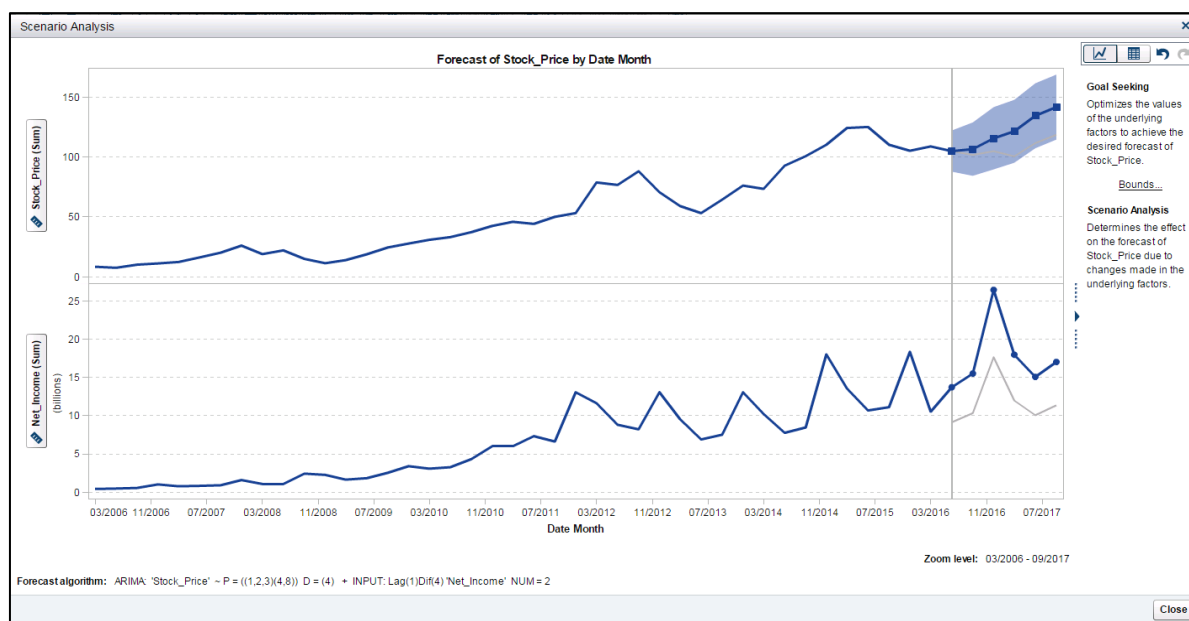


Figure 7 Forecast with Scenario Analysis

In Figure 7, the data points and the confidence band have now started to trend higher. The gap is not that far off from the original with the first forecasted quarter, but over the next 3-5 quarters, the new forecast really starts to move away from the original. You could take away from this model that the stock price is expected to rise as the net income grows over time.

Goal seeking works in a similar way except that you are changing the forecasted values and then seeing how the underlying factors would have to change to get those results. Since the underlying factors can have just a small influence on the forecast and they also do not have a confidence range, you only get an accurate result with something that is heavily correlated. So for this example, let's use Apple's revenue per quarter as the forecast and the number of iPhones sold as the underlying factor since iPhones are one of Apple's primary products.

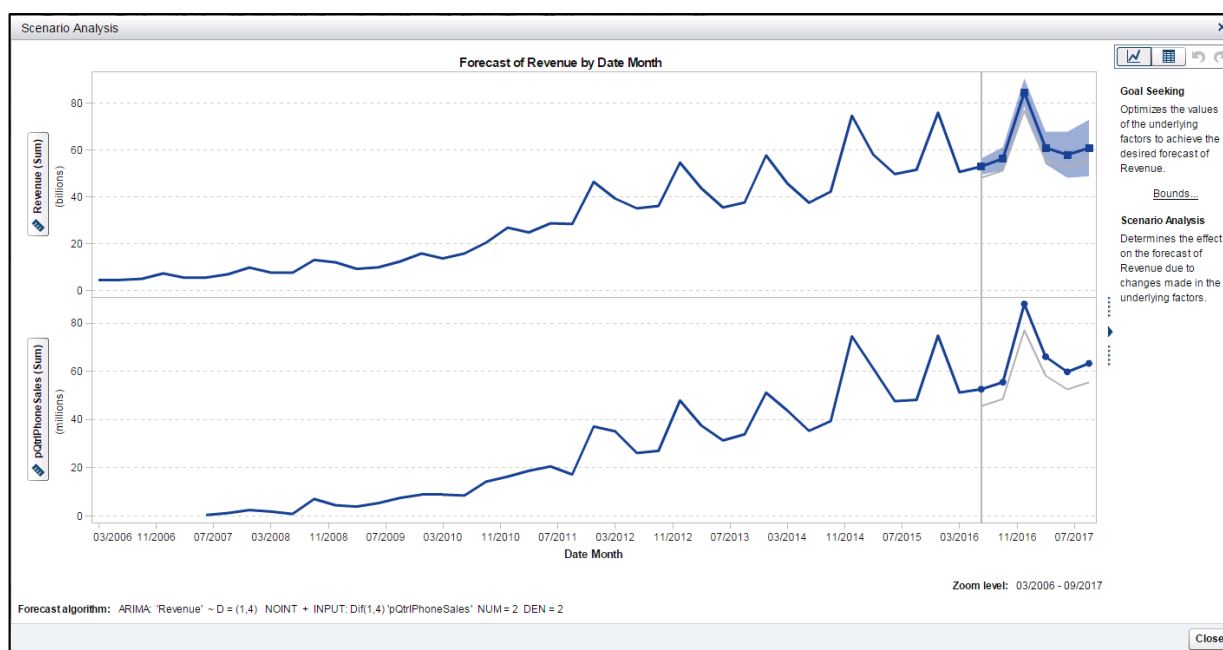


Figure 8 Forecast with Goal Seeking

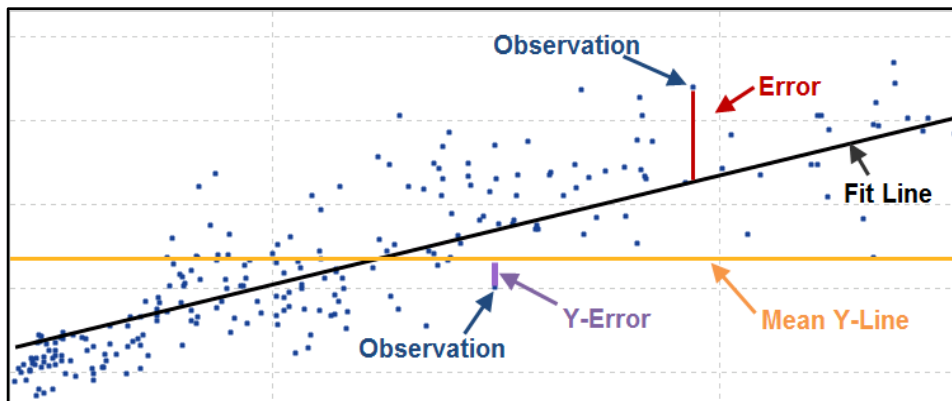
For this analysis in Figure 8, we increased the forecasted revenue by 10% in the same way that we increased the net income by 50% in the last example. You can see that two line graphs are very similar. Since the iPhone was released in Quarter 2 of 2007, the sales of the iPhone can be closely tied to the revenue of Apple since it is one of their premier products. Consequently, when we increase the revenue by 10%, there is similar change into what the percent increase in iPhone sales would need to be. For the first forecasted quarter the iPhone sales have increased from 45.44 million to 52.54 million, an increase of 15.6%. The percentage increases are similar across the next 5 quarters and end with an average 14.25%. So this goal seeking analysis is telling us that pending any other factors, this is what iPhone sales would have to be in order to hit the increased revenue target.

USING FIT LINES

Along with foresight, another key objective of data analysis is to find relationships between variables that might not be so obvious when looking from afar. When a user discovers a relationship, such as the net income and stock price shown in the forecasting example, that becomes critical information with which a business or organization can then take action. However, being able to track down these relationships is no easy task. Using lines of best fit is one way to determine if a relationship exists between two variables.

What are Lines of Best Fit?

Lines of best fit are a way to model the relationship between variables. This is done in SAS Visual Analytics with two measures. The fit line is formed between the two measures by taking in all of the data points and calculating a line that best represents the relationship for your data. The calculation is done by evaluating each of the data points and finding the line that yields the highest R-squared value. An example with random data put into a scatter plot is shown in Figure 9.



The Mean Y-Line is just the average of your Y values. This line represents a fit line that takes no X values into consideration at all. The line of best fit is the line through the data that relates the X and Y values and has the highest R-squared value compared to any other possible line using that X measure. [3]

Figure 9 Understanding R-Square Calculation

The R-squared value is calculated using the distances of error in the fit line (Error line in the figure) and the Y-Line (Y-Error in the figure). The calculation is shown below:

$$1 - (\text{Total Error Squared}) / (\text{Total Y Error Squared})$$

Each of the error and y-error values are squared and then aggregated into totals. The quotient of those totals is then subtracted from one and you get your R-squared value. The modeling process minimizes the Total Error Squared, which then results in the line with the highest R-squared value. Since the calculation divides the fit line error by the Y-Error, the higher the value signifies how much the line of best fit captures a relationship between the data. In other words, with the addition of the X values, this line of best fit shows a greater relationship between the two variables the closer the R-square number is to one than zero.

There is also more than one type of line of best fit. What is shown in Figure 9 is an example of a linear best fit line, which is just a straight line through the data. Aside from linear, SAS Visual Analytics also has the options of Quadratic, Cubic, and PSpline. Quadratic and Cubic can be used if your data is curved or has multiple points where a trend takes the data in a new direction. Quadratic lines have one curve where Cubic lines have two, similar to an S shape.

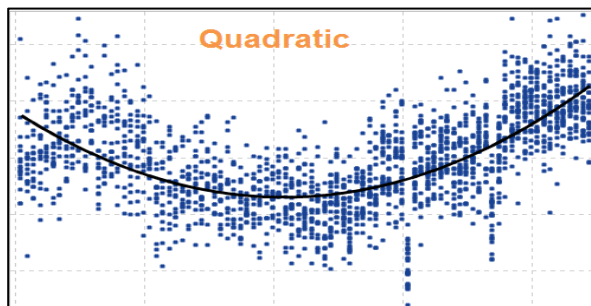


Figure 11 Quadratic Fit Line

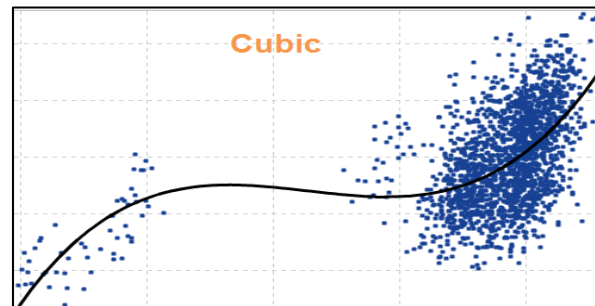


Figure 10 Cubic Fit Line

The PSpline line on the other hand fits the line in pieces, which can have multiple curves and breaks across the data.

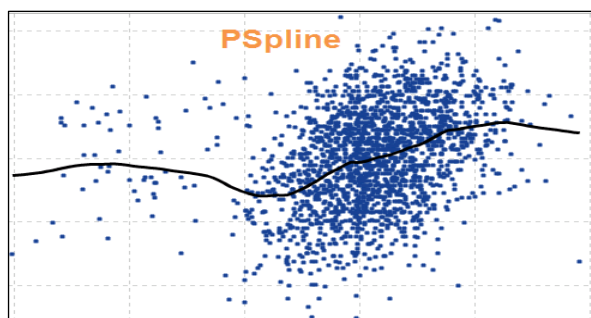


Figure 12 PSpline Fit Line

How do they 'Fit' in with SAS Visual Analytics Objects?

Fit lines are available with two objects in SAS Visual Analytics, the scatter plot and the heat map.

Scatter Plot

All of the examples above used a Scatter Plot object in SAS Visual Analytics to display the fit lines. A scatter plot is a graph that plots individual points for each row of data based on where they land according to the X-axis and Y-axis variables. The scatter plot variables must be defined as measures in the source data the option to select a fit line is in the properties tab of the Scatter Plot.

The default is none, but the other options are all of the different lines that were mentioned in the previous section as well as best fit. The best fit option selects the highest R-Square value from linear, quadratic, and cubic. PSpline is not considered for best fit.

In Figure 14, the scatter plot shows student math versus reading scores in grades 6, 7, and 8 from the VA_SAMPLE_K12_STUDENT data that comes with SAS Visual Analytics. The Best Fit option was selected as the Fit Line.

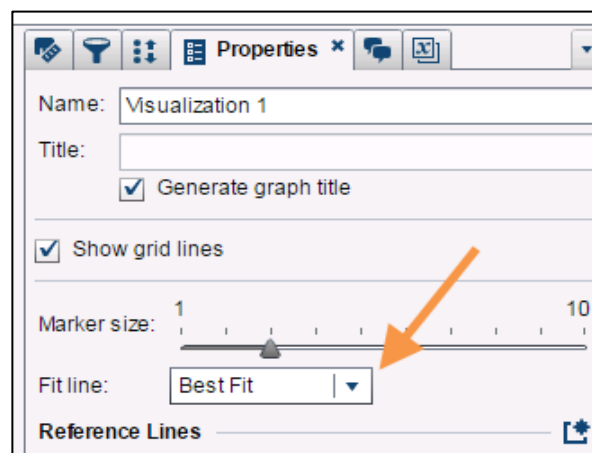


Figure 13 Best Fit Options

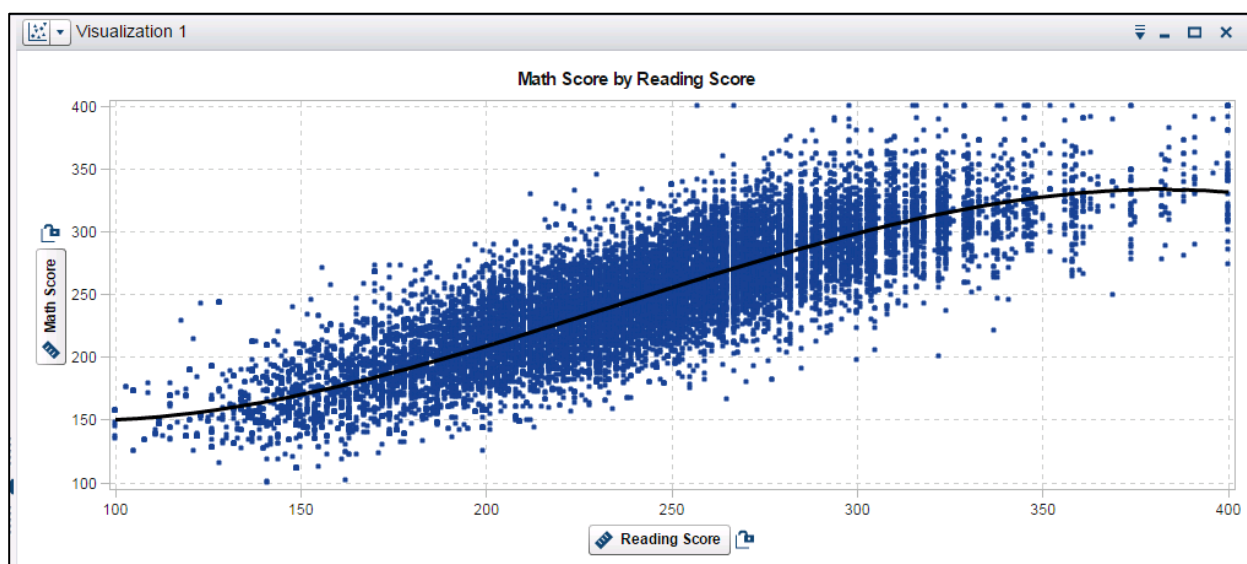


Figure 14 Fit Line Scatter Plot Example

You can see how the fit line runs through the data and has a few curves to it. This looks like a cubic line but to be sure we can check the analysis tab at the bottom shown in Figure 15.

Results Analysis	
Property	Value
Model type	Best Fit (Cubic)
"Best fit"	Best fit evaluates linear, quadratic, and cubic fit lines and picks the one which represents the relationship between two variables most appropriately.
Model description	The cubic fit line is used to describe the relationship between two variables when that relationship exhibits a curve with two inflection points, or changes in curvature (an S-shaped curve, for example). If the points on the scatter plot are tightly clustered around the line, then it likely provides a good approximation for the relationship. If not, another fit line should be considered to represent the relationship.
R-square value	0.73

Figure 13 Analysis Tab for Fit Line in Scatter Plot

The analysis tab gives the full breakdown on selection, description, and the R-square value. After selecting just the linear and quadratic lines, those R-square values were both 0.72. At 0.73, the cubic line was our best fit for this model.

Understanding Heat Maps

Heat maps are similar to scatter plots in that each data point has a specific spot on graph with respect to the X-axis and Y-axis. Heat maps are different in that you can bin the measures so that instead of an individual point, you now have a range bucket that counts the frequency or aggregates any other measure for all the points within that range. If you did choose an aggregate, it would have no effect on the fit line since the fit line is shown based on the two measures on the X-axis and Y-axis. You can also use a category as one of the axis if you would like, but fit lines do not work with a category since R-square has to be calculated between two measures.

As they relate to fit lines, heat maps and scatter plots are the same in how they calculate the line and display it on the graph. Heat maps are better from a visual aspect in that if you have too many data points on a scatter plot, the heat map categorizes them into areas and shows the intensity of the frequency through color in the blocks. In Figure 16, we use the same student score data that was in the scatter plot example.

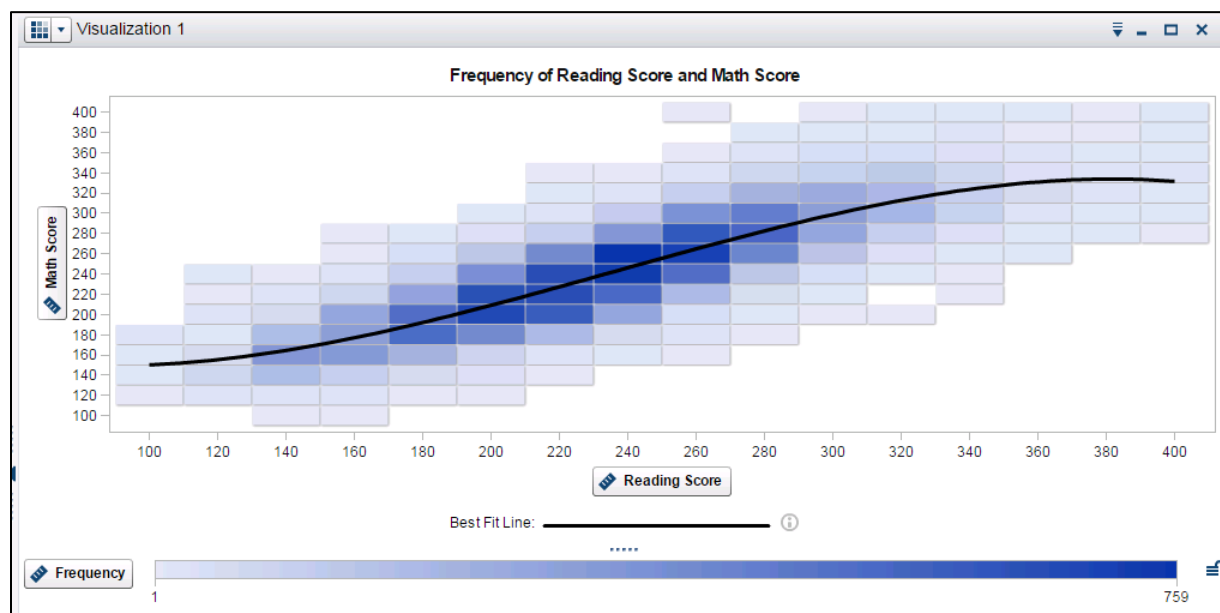


Figure 14 Fit lines with Heat maps

You can see that it is definitely a lot easier on the eye to look at since the data points have been replaced with blocks of color. The legend at the bottom shows the level of frequency based on the color so that the user can grasp an understanding of how many points are in each block. The options in the Properties tab and the results in the bottom Analysis tab remain the same as the scatter plot.

How to interpret the line?

Now that we have gotten our line modeled and understand how SAS Visual Analytics does the modeling, it is time for the analysis part. As mentioned before, lines of best fit are a way to show relationships between measures. The higher a line gets on the R-square value, the more variability in your Y is captured by your model, indicating better model fit. There are a few other things to consider about the line for analysis. [2]

- **Direction** – Does the slope go up or down? If the slope is going up then you have a positive relationship which means as one measure increases, so does the other. When the slope is going down then you have a negative relationship and as one measure increases, the other decreases.
- **Strength** – How condensed are the data points to the line? If most of the data points follow the line then the relationship is going to be stronger. However, if they are scattered all over or they are all compressed into one small area, then a relationship might not be as obvious.
- **Shape** – Is it a straight line or curved? A straight line signifies a simple relationship, when one measure goes one way, the other measure goes follows suite. A curve means that there could be a changing point. This means that as your data is following the line, there becomes a point where the relationship changes. These points of curvature can be very important to understand more about your data.
- **Outliers** - Are there any outliers? Outliers can be good to find examples of what doesn't follow the relationship.

Back to our example, we know that we have a well fit model based on our R-square value (0.73). The line flows in an upward direction which tells us that a student with a high math score should score relatively as high on reading and *vice versa*. Nearly all of our data points are in-between the 200-300 range for both scores and that is where our line stays in an upward slope which indicates a strong association between the two measures. The shape is where things aren't as straightforward. Since this is a cubic line, the ends of the line start to straighten out and we no longer have our slope. This indicates a non-linear relationship between reading and math scores where in the extreme ends of the data, the lower (100-160) and higher (340-400), we would expect a smaller increase in math score for each additional point on reading score than we would in the middle range with greater slope.

UNDERSTANDING CORRELATIONS

As our lines of best fit in the previous example, correlations are another way to determine relationships between measures.

How does SAS Visual Analytics Calculate Correlations?

In the previous section, we reviewed the calculation and meaning of the R-Square Value. Correlations in SAS Visual Analytics are calculated in a similar manner except they use Pearson's product-moment correlation coefficient calculation. [4] This calculation takes in two measures and determines how much they are related in a linear manner.

The range of the Correlation value can be anywhere from -1 to 1. Anything from -1 to 0 indicates a negative relationship, which means that as one of the measures increases the other decreases. A correlation of 0 shows no relationship at all. Positive numbers from 0 to 1 indicate a positive relationship, which means that as one measure increases so does the other. SAS identifies these ranges of ratings for correlations as being Weak, Moderate, or Strong.

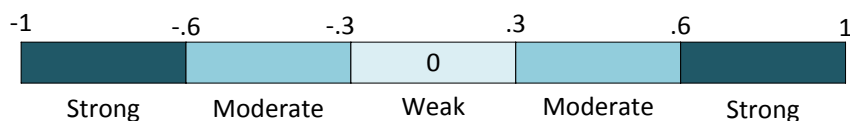


Figure 15 SAS Visual Analytics Ratings for Correlations

Where Can You Find Them in Data Objects

Correlations between two measures can be calculated in the correlation matrix, or through a linear fit line in the heat map and scatter plot.

Can a Correlation Matrix Get Us to the Playoffs?

In a Correlation Matrix, there are two options in the Roles tab under Show Correlations to display the correlations between the measures that you want. The option **within one set of measures** takes a set of measures and displays them in a matrix against themselves so that, in a triangle format, you will see each measures correlation against one another. In Figure 18, we measure seasonal team baseball statistics against one another. This dataset (Data Source:

<http://www2.stetson.edu/~jrasp/data.htm>) combines all team seasons from 1921-2009 and totals up team statistics such as Hits, Home Runs, ERA, and so on.

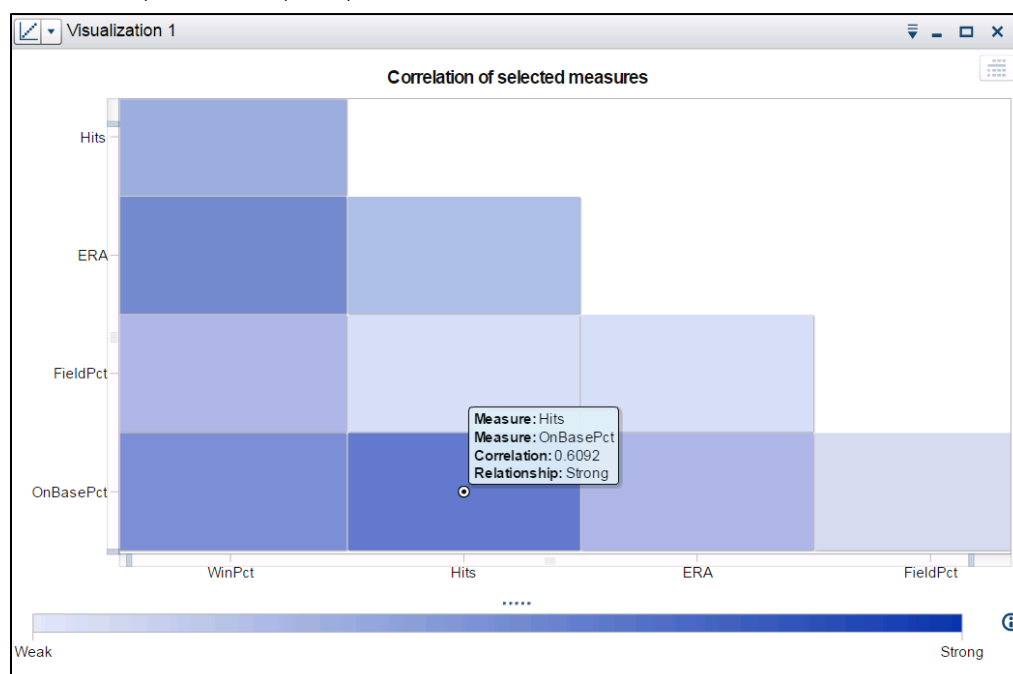


Figure 16 Correlation Matrix with one set of measures

After adding in WinPct (Win Percentage), Hits, ERA (Earned Run Average = Measure of earned runs given up per 9 innings), FieldPct (Fielding Percentage = Measure of successful defensive plays), and OnBasePct (On Base Percentage = Measure of times a batter gets on base per plate appearance) to the measures in the Roles tab, we get our matrix of correlations. The bar at the bottom shows that the color displays how strong the correlations are. If you hover over any of the boxes, then you see the data point box which gives you the measures that were calculated, the correlation, and how SAS categorizes that correlation. In this example, Hits and OnBasePct have a strong correlation which makes sense because every hit that a batter gets directly influences their on base percentage (OnBasePct).

Now let's look at something that might be useful for our analysis. Win Percentage (WinPct) is the goal of all baseball teams, since you need to have one of the top win percentages to make the playoffs each year. In the next figure, **between two sets of measures** is chosen and Win Percentage is put on the X-axis. Then the Y-Axis is filled in with all of the measures that we want to compare against one another to see which statistic is most heavily correlated with WinPct.

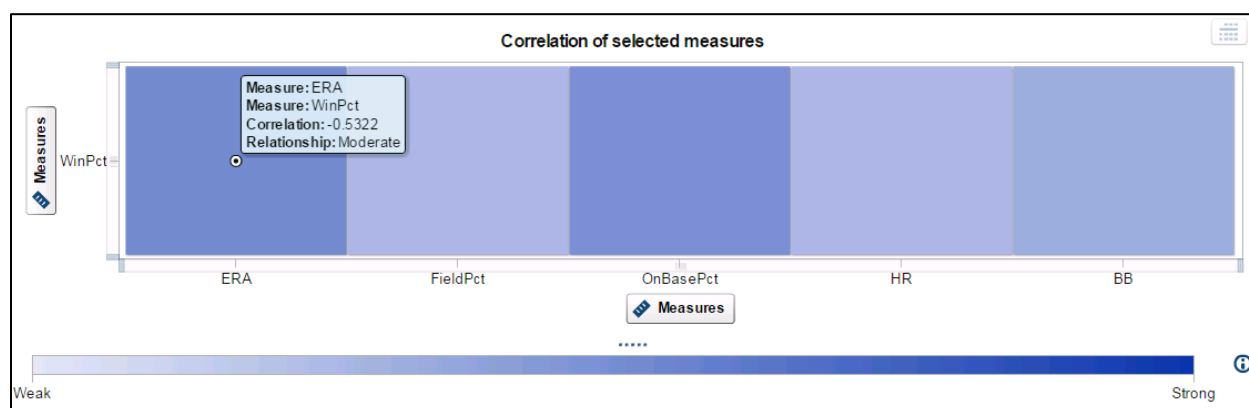


Figure 17 Correlation Matrix with two sets of measures

Using this option helps cut down on the matrix and allows the user to see just the set of correlations that they want to compare. You can add more measures to the X-axis but the point is that it cuts out the full matrix that you get with the one set of measures option.

Linear Fit Lines

From the previous section, fit lines were covered in scatter plots and heat maps. In each of those examples if the user selects linear fit line or selects best fit and the best fit is linear, then the correlation value will be calculated between the two measures in the Analysis section at the bottom of each object. In Figure 20 is the WinPct and ERA correlation shown in a Scatter Plot.

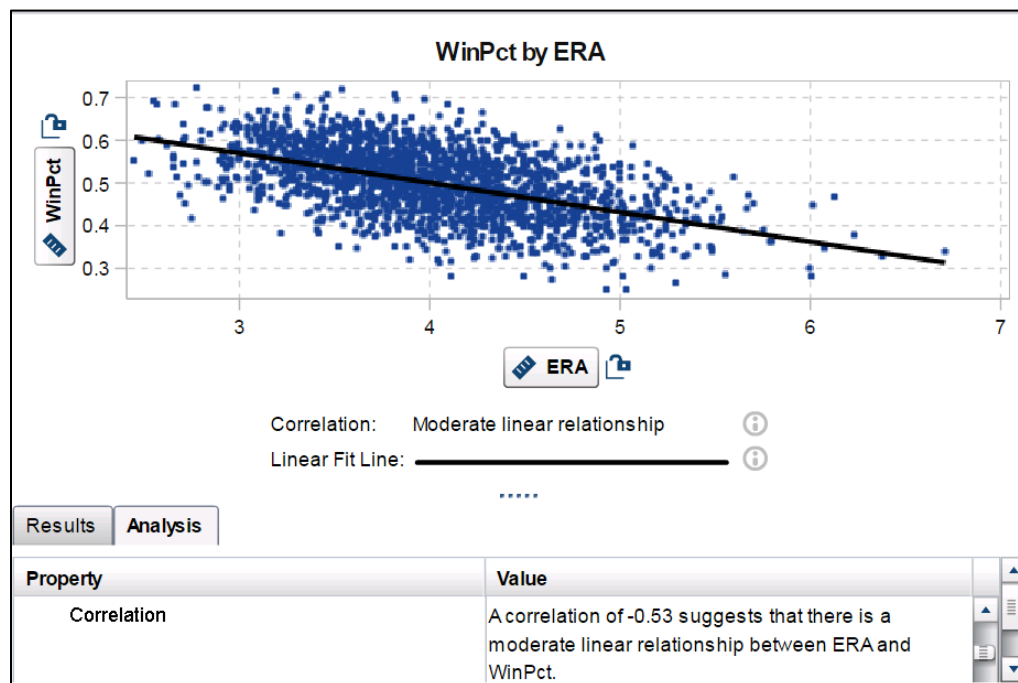


Figure 18 Correlations in Linear Fit Lines

Interpreting the Correlation Value

With a lot of data and a strong correlation between two measures you might assume that they have found a relationship between measures. Sometimes that is not always the case. The phrase *correlation does not equal causation* is common in the field of statistics and means that just because two measures have values that are related—which is measured by correlation—it does not mean that the concepts behind the measures have a direct relationship. There are many different forms of an apparent relationship between data items. In Steven Few's book *Now You See It*, he breaks down correlations to meaning one of four possibilities [2]:

- One measure causes the others behavior
- Neither causes the other's behavior, both are caused by other variables
- Neither causes the other's behavior, another variable connects them
- Correlation is erroneous due to insufficient or bad data

So in the previous two figures we were looking at win percentage against other measures to see which ones were the most correlated. In Figure 19, each of the five measures has a moderate relationship with win percentage. This makes sense because all of those measures have an influence on the outcome of the game. ERA had the strongest correlation at -.53. This means that as a team's pitchers gives up fewer runs on average, we would expect them to have a higher win percentage. The correlation indicates that a lower ERA causes a higher win percentage, which we know to be true based on the rules of baseball. In this example then, the correlation of the values of the measures was indicative of a conceptual relationship between those two measures.

Now when dealing with measures that aren't so directly linked you may find out that there is a hidden connector between them or that they just happen to be correlated by chance. In the below figure, per capita beef consumption in pounds (Data Source: <http://www.disastercenter.com/crime/uscrime.htm>) is compared to burglaries per 100,000 people in the United States from 1960-2014 (Data Source: <http://www.nationalchickencouncil.org/about-the-industry/statistics/per-capita-consumption-of-poultry-and-livestock-1965-to-estimated-2012-in-pounds/>)

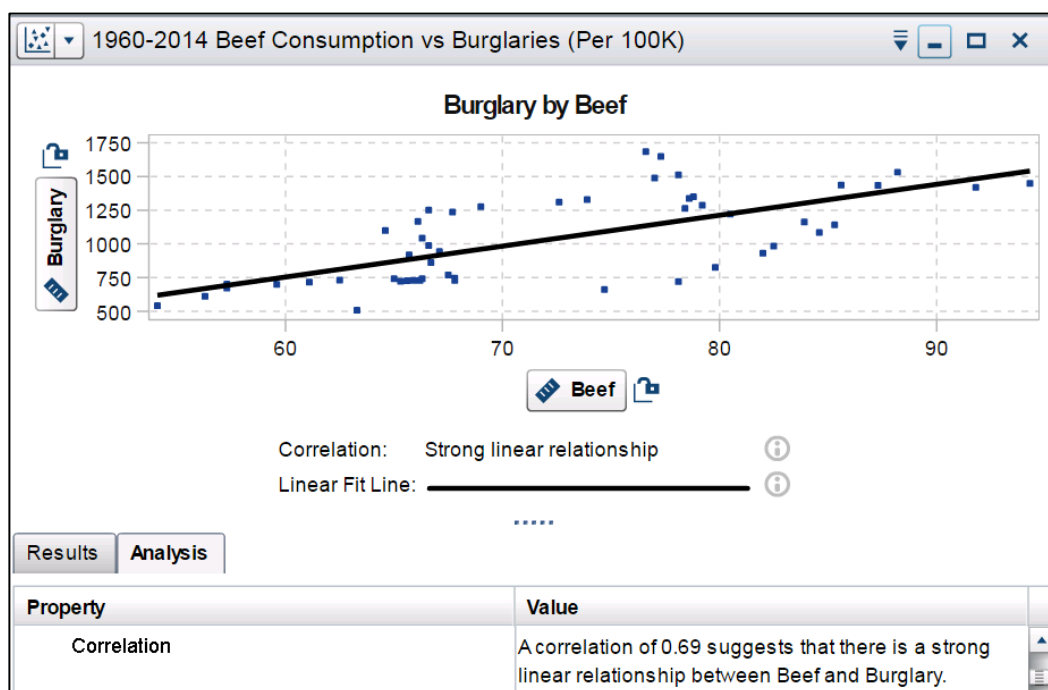


Figure 19 Example of a correlation with no relationship

Well it turns out that there is a strong correlation between the two. Does this mean that one causes the other? There's no logical reason to expect that the more burglaries in the United States that there are, then more beef will be consumed or *vice versa*. Correlations show possible relationships between data items, it's up to the user to then do further investigation onto where the connection lies.

CONCLUSION

Throughout using all of these features, we have been able to learn more about the data at hand. In forecasting, we were able to see what measures had an influence on Apple's stock price and how it would react if some conditions changed. Using the test score data with lines of best fit, it could be seen that there was a direct relationship between subject scores but only for certain sections in the data. In looking at correlations, it was determined that amongst the measures reviewed, ERA had the most influence on a team's winning percentage throughout MLB history. These datasets all came from vastly different areas but they all had many data fields and these features of SAS Visual Analytics enabled us to learn more about the relationships between those data fields. Hopefully after reading this, you can take these concepts back to your organization and be able to apply them to other scenarios.

SOURCES

- [1] Chawla, V. "Correlations, forecasts, and making sense of it all with visualization" SAS, May 2016. Available at: <http://blogs.sas.com/content/sascom/2016/05/27/correlations-forecasts-and-making-sense-of-it-all/>
- [2] Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis. Oakland, CA: Analytics Press.
- [3] Frost, J. "Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?". The Minitab Blog, August 2013. Available at: <http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>
- [4] SAS Institute SAS Visual Analytics 7.3 User's Guide. Available at: <http://support.sas.com/documentation/cdl/en/vaug/68648/PDF/default/vaug.pdf>

RECOMMENDED READING

- SAS® *Visual Analytics User Guide*, latest version
- *Now you see it: Simple visualization techniques for quantitative analysis*, Steven Few

CONTACT INFORMATION

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

Ryan Kumpfmiller

Zencos Consulting

Cary, NC

rkumpfmiller@zencos.com

www.zencos.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.