

Visualizing Data Techniques, Including Auto-Charting and Big Data for SAS® Global Forum 2012

Justin Choy, SAS Institute Inc., Cary, NC, USA

ABSTRACT

Visualizing data of varying sizes can be challenging. This paper discusses the issues concerning visualizing data and provides suggestions on how to address these issues. The paper assists users who don't know which visualization to use for their data

INTRODUCTION

"A picture is worth a thousand words." This is especially true when you are visualizing data. In order to effectively create meaningful visuals on your data there are some tips and techniques you should consider. This paper is by no means all-encompassing; it covers some general visualization techniques. Data size and column composition plays an important role when selecting which visualizations to use to represent the data. "Big data", the latest buzz, brings its own unique set of challenges when creating visualizations and this paper will cover some of the challenges and potential solutions.

Big Data is usually described as Data of such volume, variety or velocity that an organization must move beyond its comfort zone technologically to derive intelligence for effective decisions. Volume refers to the size of the data, where variety describes whether the data is structured and/or unstructured. And velocity is the speed in which you can access the data.

Auto-charting is a method to assist users who may not know which visual to use by automatically creating the best possible visual for them based on the data selected. SAS Visual Analytics is a new SAS product that utilizes intelligent auto charting to help business users visualize their data. The auto-charting in SAS Visual Analytics utilizes many of the tips, techniques and suggestions covered in this paper.

SAS Visual Analytics is a business intelligence solution that leverages SAS high performance analytic technologies, to empower organizations to explore huge volumes of data very quickly in order to see patterns, trends and identify opportunities for further analysis, and surface results to information consumers.

SAS' in-memory server, SAS LASR Analytic Server, is the heart and soul of our high performance analytic technologies, and provides a game changing ability to accelerate analytic computations. The combination of high performance analytics and the highly visual, easy to use, data exploration interface --- enables organizations to derive value from massive amounts of data ('big data') faster than ever before. This creates an unprecedented ability to solve difficult problems, improve business performance and mitigate risk – rapidly and confidently.

Here are some tips to assist you in generating the best visuals for your data.

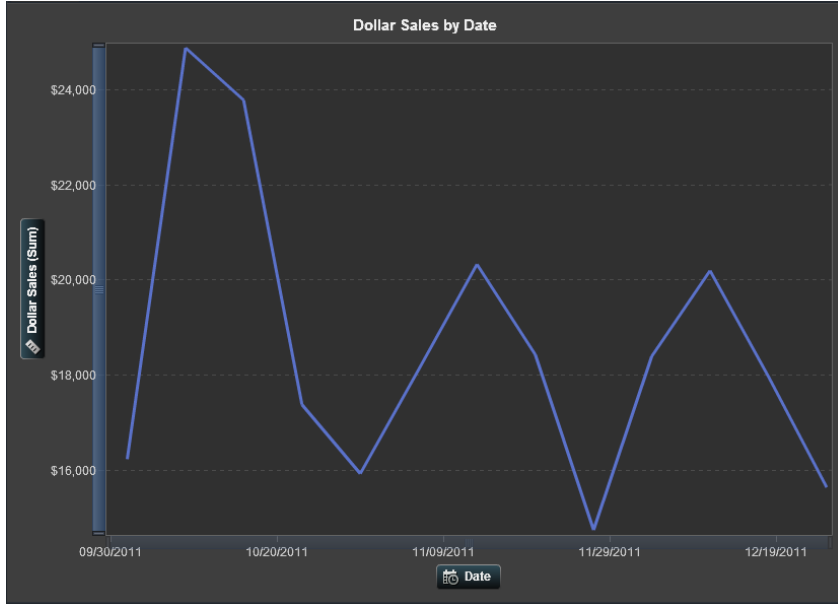
- Understand the data you are trying to visualize, including its size and cardinality.
- What you are trying to visualize and the information you wish to convey.
- Know your audience and understand how they process visual information.
- Use a visual that conveys the information in the best and simplest form for your audience.

CHARTING 101

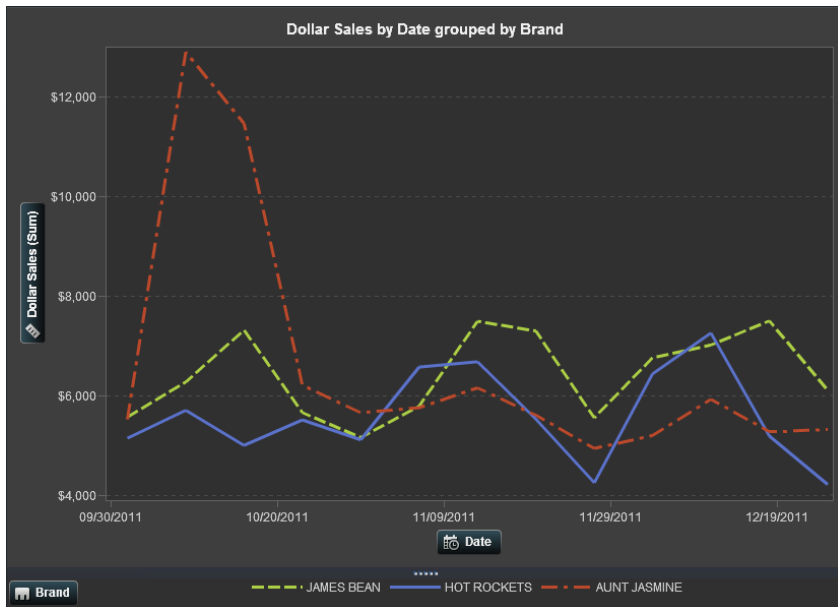
For users that don't know or unsure of which graph type to use when here is a quick guide to use when determining which chart type to use.

LINE GRAPHS

Line graphs are good to use when you are trying to track changes or trends over time. (See Display 1.) Line charts are also good when you are trying to compare multiple items over the same time period for example. (See Display 2.) You can only use line charts when the values on the x axis are continuous.



Display 1. Line Graph



Display 2. Multiple Category Line Graph

BAR GRAPHS

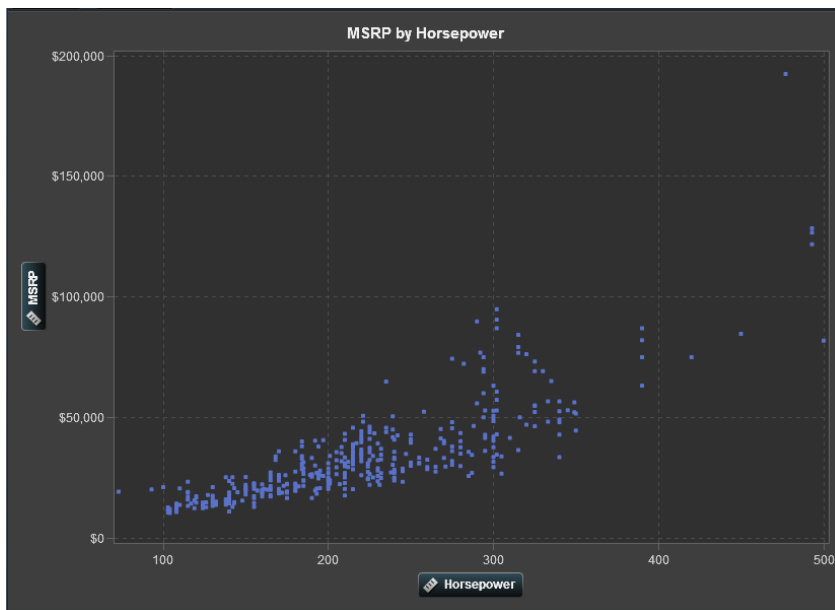
Bar graphs are good to use when you want to compare quantities of different categories or groups. (See Display 3.)



Display 3. Bar Graph

SCATTER PLOTS

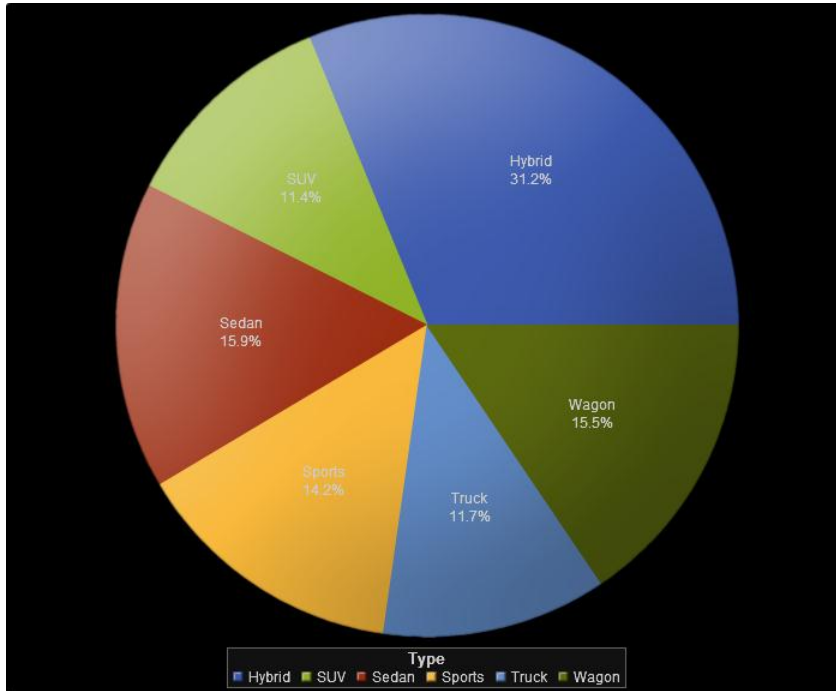
Scatter or X-Y plots are good to use when you want to compare or understand if there is a relationship between the X and Y variables. (See Display 4.)



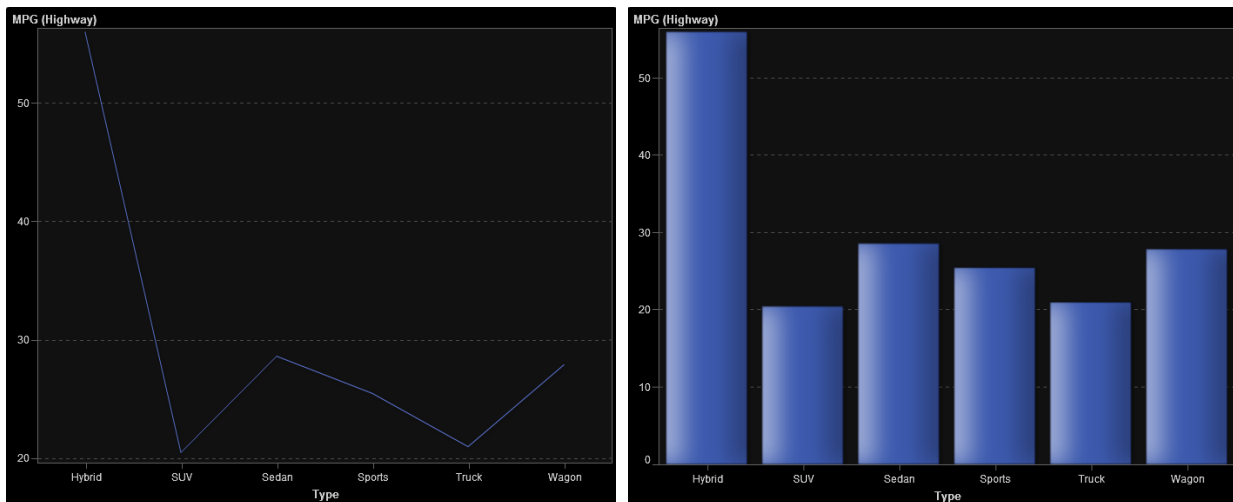
Display 4. Scatter Plot

PIE CHARTS

Pie Charts, some people love them and some people hate them, these can be used to compare parts of a whole. If you choose to use a pie chart then the slices should be a percentage of the whole. (See Display 5.) These charts aren't useful if there are too many slices or each slice is very close in size. There are alternate charts such as bar or line that can show the same information better. (See Display 6.)



Display 5. Pie Chart



Display 6. Alternates for Pie Charts

Of course there are many other chart types that you can choose from such as bubble plots, heat maps, tile charts and many more. The selection of these chart type will usually depend upon the number of categories and measures (or dimensions) you have to visualize. By following the tips outlined in the introduction and understanding the use cases

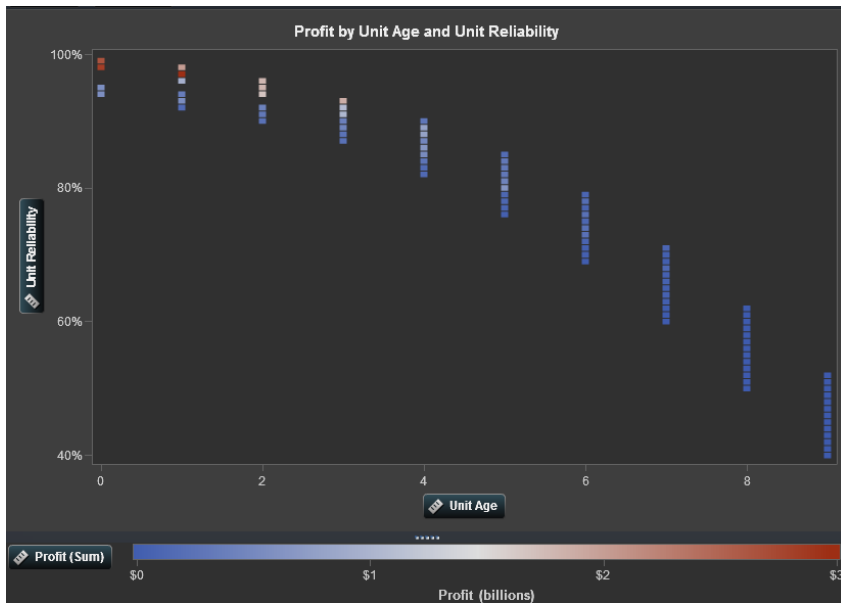
you may need to try different to visuals and test these with your audience to make sure the correct information is being conveyed.

BIG DATA

Big data brings new challenges to visualization there is the volume, the variety and velocity that needs to be taken into account. As well as the cardinality of the columns you are trying to visualize.

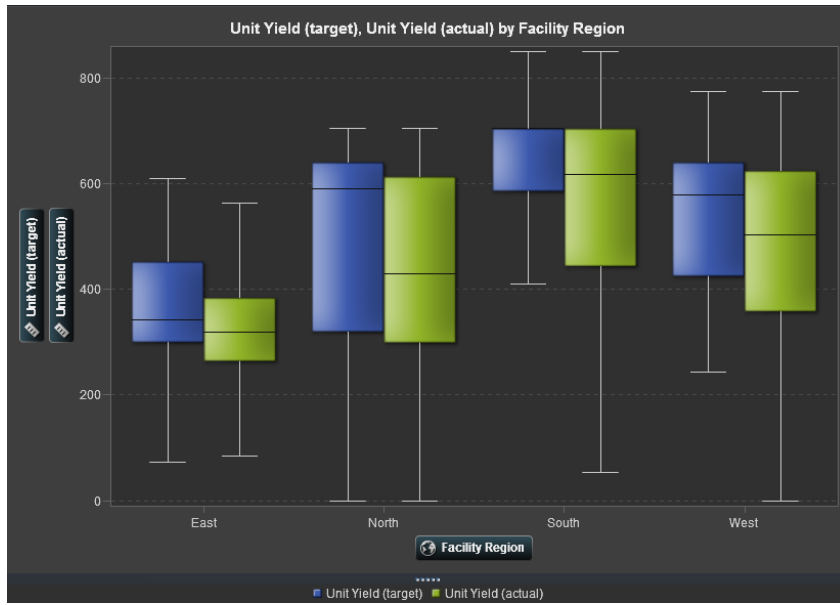
Volume can become an issue because you may not be able to process the sheer amount of data in a timely manner, thus requiring you to have to make compromises and aggregate the details to visualize.

For example, what if you have a billion rows dataset and you want to create a scatter plot on two measures? The user trying to viewing a billion points in a scatter plot will have a hard time trying to see that many points. And, the application creating the visual may not be able to plot a billion points in a timely or effective manner. One potential solution is to use binning on both axes so that you can effectively visualize the big data. (See Display 7).



Display 7. Using Binning to Visualize Big Data.

Box plots are another example of how the volume of data can impact how a visual is shown. Regularly, these will show fine when using big data. (See Display 8.) With this specific visual, sometimes users want to show the outliers. With outliers being one to five percent of the data, one percent of a billion rows is 10 million outlier points. Looking at Display 8, we have one category and two measures and to visualize outliers could mean plotting 20 million outlier points. How could you effectively show the outliers for box plots in a meaningful way?

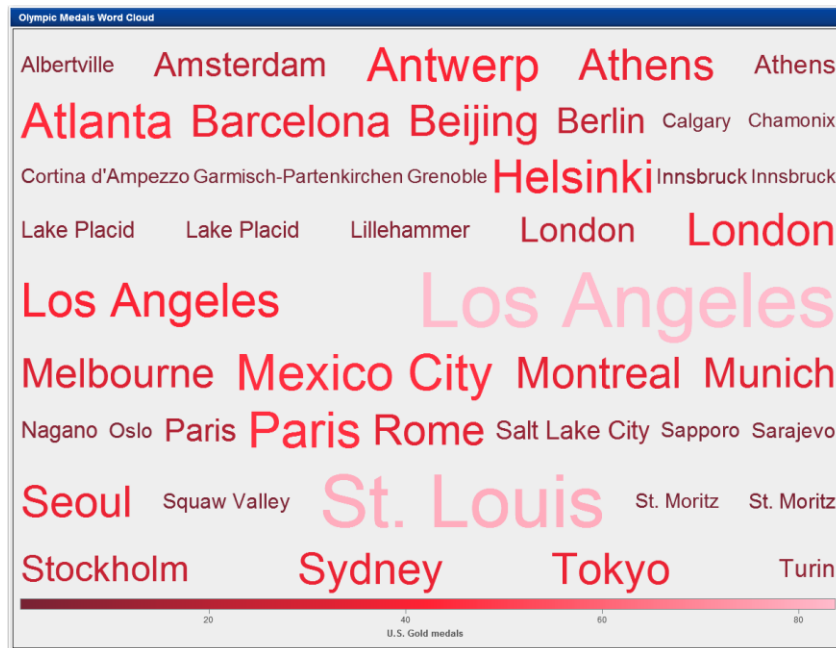


Display 8. Box Plot on Big Data

UNSTRUCTURED DATA

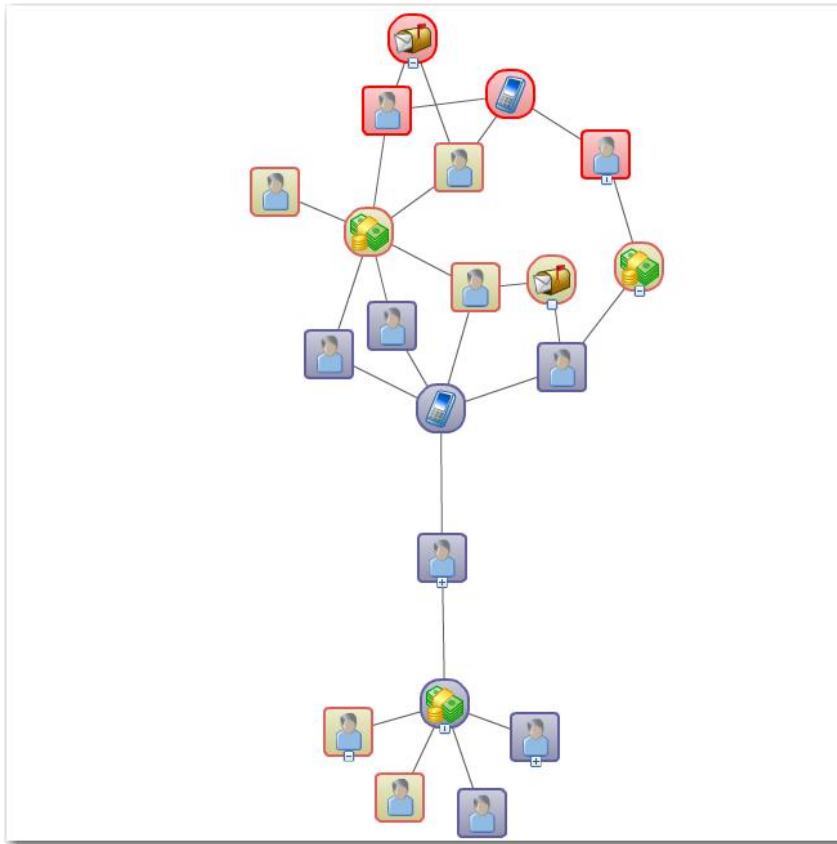
Variety brings challenges in that visualizing structured data is fairly simple, though unstructured data requires new visualization techniques.

A word cloud visual can be used on unstructured data to visualize high or low frequent word occurrences. (See Display 9.)



Display 9. Word Cloud

Another example is using a network diagram to show the relationship of someone tweeting and their followers. (See Display 11.)

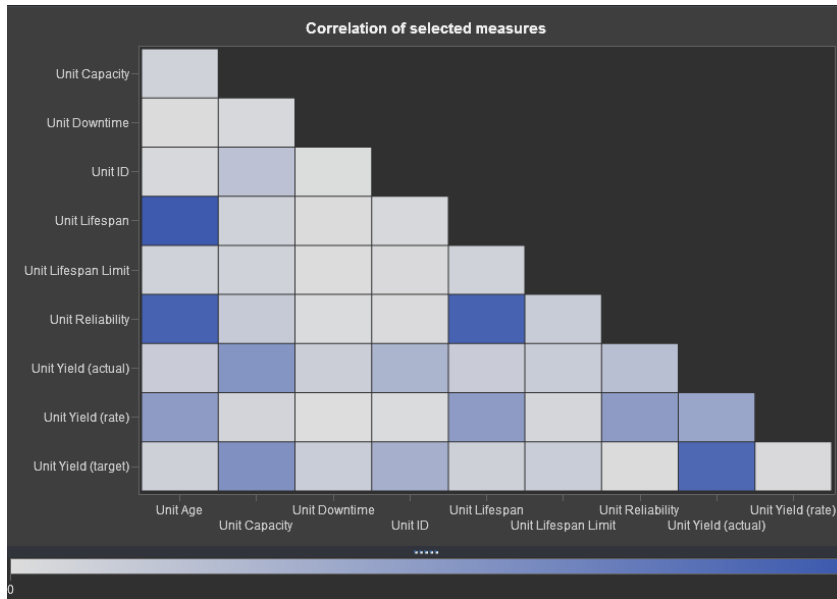


Display 10. Network Diagram

VISUALIZATION VELOCITY

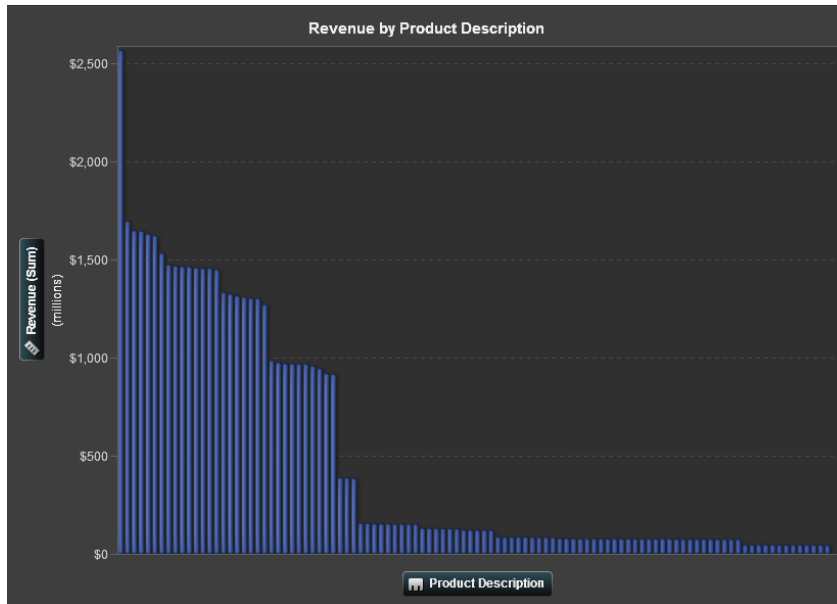
Velocity for big data is all about the speed you can access and process it. For visualizations this translates to the speed in which you can visualize it.

The example below visualizes forty five correlation calculations on just over 1.1 billion rows of data. This graph shows the correlation values and returns in two to six seconds using the SAS LASR Analytic Server. If you were to run this using SAS Enterprise Guide, for example, it would take you around six hours for the visual to render. (See Display 11.)



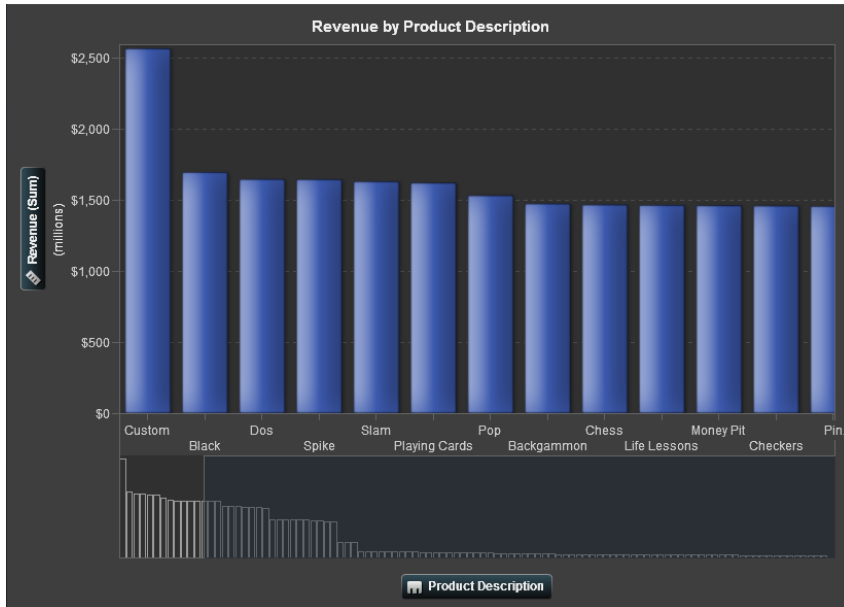
Display 11. Correlation Matrix

Cardinality becomes a concern in big data as the data may have many more unique values per column. The example in Display 12 shows only 105 unique Product Descriptions. With big data we may have over a million unique values. As you can see from the example even viewing 105 bars the viewer cannot see the labels for each of the bars and now the graph is becoming less meaningful. Imagine if we had 1,000,000 bars could we even see them?



Display 12. High Cardinality Bar Chart and Big Data

One method to deal with high cardinality SAS has adopted in SAS Visual Analytics for bar charts is to provide an overview bar which zooms into the bar chart and provides a way for the user to scroll through the whole chart, as well as the amount of zoom. It also shows the whole graph in the overview bar. Comparing Display 13 to Display 12, it's easy to see that Display 13 show the information more clearly and easily.

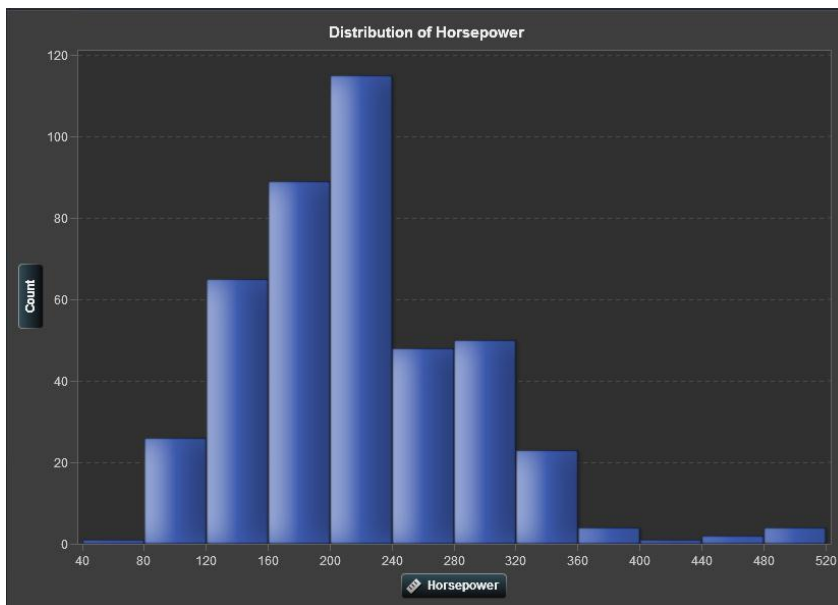


Display 13. Overview Axis Bar Visual of High Cardinality big data

AUTO-CHARTING

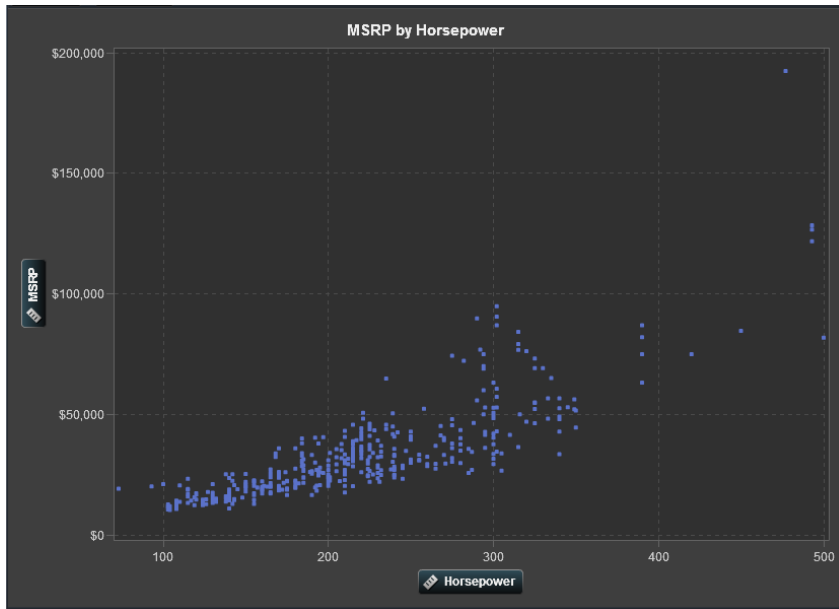
In SAS' newest product SAS Visual Analytics, the exploration capability provides intelligent auto-charting that will show the best possible visual based on what they user drags and drops on to the visual palette. It's important to note that auto-charting may not always create the visual the user had in mind. Therefore we allow the user to select the specific visual they wish build. In some instances, they can use auto-charting to start and then switch to the specific visual.

For example, when a single measure is selected, auto-charting shows the distribution of that measure. (See Display 14)



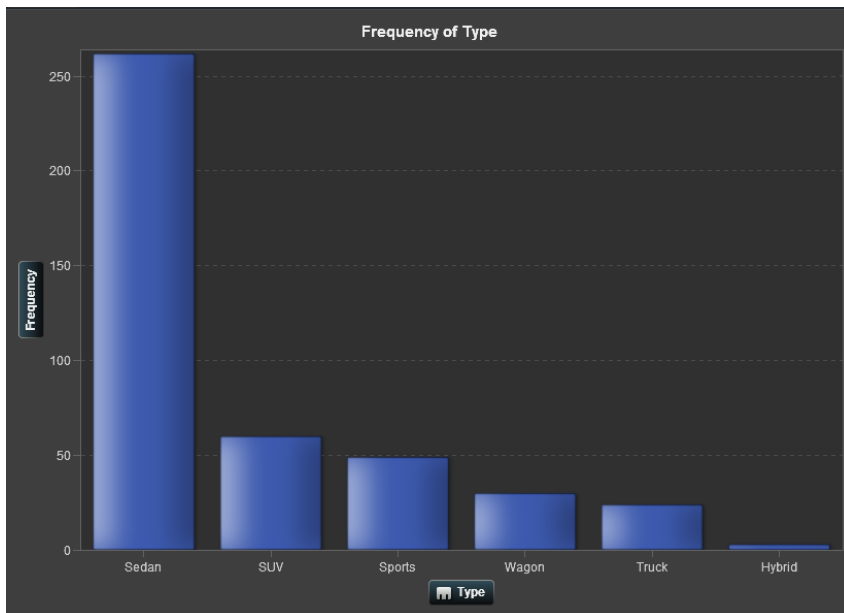
Display 14. Distribution of Data

The addition of a second measure results in a scatter plot. (See Display 15)



Display 15. Two measures results in a Scatter Plot

A category can be one of three types, standard, date and geographic. When the categorical type is standard, the visual will show a frequency count of data. (See Display 16) If the category is a date, then a measure is also required and the visual will be a line graph. (See Displays 1, 2 and 6)



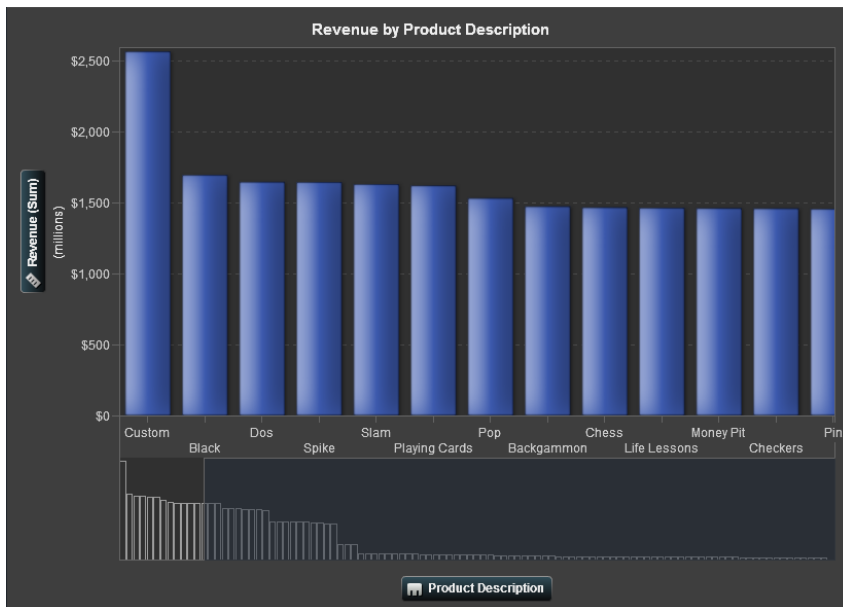
Display 16. Frequency of Data

Whereas, if the category is geographic, then a geographic map will be shown. (See Display 16.)



Display 17. Geographic Map Frequency

The auto-charting in SAS Visual Analytics takes into account the cardinality of the data and adjusts the visuals accordingly. Using the visual in Display 18 as an example, the cardinality of the column Product Description was 105 auto-charting checked the cardinality of the selected column and automatically provided the overview bar axis as the cardinality was deemed high. The overview axis is an option that can be turned on and off as required.



Display 18. Auto-charting with Overview Axis

CONCLUSION

Visualizing your data is both fun and can be challenging “a picture is worth a thousand words” could not be truer. It’s definitely easier to understand information in a visual as compared to a large table with lots of rows and columns. Though sometimes the visual creator can get swept up with all the cool and unique visual choices they can choose from and end up presenting the information using an inferior visual. In some cases there are specific visuals you should use, while in others your audience may dictate the visual. For the latter scenario showing your audience an alternative visual that can convey the information easier and clearer may sway them over. Understanding your data and its composition, what information you are trying to convey visually to your audience and knowing how they process visual information will enable you to choose the appropriate visualization. Or there are products like SAS Visual Analytics that will help users best visualize there data.

RECOMMENDED READING

- Chawla, Varsha and Whitman, Lisa. [Finding the Best Display Type for Your Data](#), SAS Global Forum 2011.

CONTACT INFORMATION

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

Justin Choy
SAS Institute, Inc.
SAS Campus Drive
Cary, NC 27513
E-mail: Justin.Choy@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.