Effectively Grappling With Graphics
Chris Potter
Syntex Laboratories, Inc.

Introduction - This tutorial is a collection of examples taken from my last two SUGI presentations.

A picture may be worth a thousand words, but it is a rare picture that can adequately describe a thousand word set.

We often assume that a graph or plot of data will automatically convey our message. However an inappropriate or poorly designed plot may raise more questions than it answers. Effective graphics must be designed to communicate. I hope to show some of the ways that this communication can be assured.

In this paper I will be discussing:

1. The Purpose of Graphics.
2. Steps for Creating a Chart.
3. Rules for Graphical Integrity.
4. Effective Chart Design.
5. Using PROC GPLOT and GCHART.

Before continuing, I should mention an excellent reference book on this subject: The Visual Display of Quantitative Information, by Edward R. Tufte, Graphics Press. Several examples and definitions were obtained from this source. In addition examples of typical chart errors were taken from an IBM reference manual "Pointers on Effective Chart Design". Unfortunately I have no other information on this manual.

THE PURPOSE OF GRAPHICS

Graphics has been defined as the art of drawing. Recently, the use of computers has changed graphics to a "push a button - get a graph" simplicity that has compromised the art portion of graphics.

Graphics should be thought of as the communication of ideas through visual means.

Graphics may be used to:

• Show the structure in the data
• Summarize large amounts of data
• Demonstrate how things are connected
• Show organizational relationships
• Provide advertising
• Illustrate training
• Communicate ideas
• Display humor
• Set up a situation or feeling

Good graphics should:

• Show the data clearly
• Display your conclusions
• Avoid distorting the truth
• Avoid ornamentation
• Be interesting and aesthetic

Some Don'ts:

* Don't let software defaults determine the final output.
* Don't get carried away with fonts, patterns, colors or 3d effects.
* Don't forget what you're trying to show.
* Don't omit or distort information.
* Don't use the wrong type of chart.
Steps for Creating a Chart

The following ten point checklist has been quite useful to me for preparing graphics for presentation:

1. **Know your audience** - Effective graphics for one audience may be confusing to another group. Tailor your presentation to match the level of your audience.

2. **Decide what you want to show** - For graphics to be a form of communication you need to know what you want to say. Make sure your data supports what you want to say.

3. **Make a sketch** - Quick sketches can help to visualize the data and determine the proper chart type.

4. **Start simple** - When you first plot the data don't bother with triplex fonts, headings, labels, patterns, etc. This will save time and money. Plot to the screen instead of the plotter as your start if possible.

5. **Add titles and symbols** - Titles should be sized and positioned for legibility and appropriate balance and use of space. Type fonts should be simple but bold enough to stand out. Symbols should be used to mark the data if necessary, and should be sized to allow differentiation between different data. Avoid large symbols if possible.

6. **Check colors and shading for impact** - Impact means an aesthetic chart with the data presented clearly. Impact does not mean multicolored op-art masquerading as a plot.

7. **Does the plot show what you want** - This is the reality check of any plot. Does the plot demonstrate your conclusion without distorting the data.

8. **Check for Chartjunk** - Is all the ink necessary and useful.

9. **Define and simplify if possible** - Remove any chartjunk or patterns that don't help the plot.

10. **Try the plot out on someone** - Does another person see what you are trying to show.

**Rules for Graphical Integrity:**

- The graph should represent proportional to the changes in the data represented.
- The graph should be clearly labeled.
- Show data variation, not design variation.
- Use deflated or standardized units for money when shown to a time series plot.
- Data should not be shown out of context.
- The number of information carrying dimensions should not exceed the number of dimensions in the graph.

**Effective Chart Design**

The following points are adapted from the IBM manual "Pointers on Effective Chart Design". They cover many of the common graphical errors.

1. **Use Well-Designed Charts for all Your Graphics.**

   Often charts used for analysis find their way into a presentation or a report. Poor charts will confuse the audience and reflect poorly on the presenter.

2. **Determine What Relationships You Want to Show.**

   Decide what point you want to make before you choose a chart, then design the chart to illustrate your point. You may want to write down a statement of what the data shows and later see if the chart shows that relationship effectively.

3. **Make curves thicker than your grids.**

4. **Label curves if there is room.**

5. **Write axis labels horizontally.**

6. **Make axis numbers large enough to read.**

7. **Use the same scale when comparing trends.**

8. **Include zero when comparing levels or totals.**

9. **When you omit zero let your reader know.**

10. **Use scales that make interpolation easy.**
11. Use different line thickness for data - use dashed or dotted lines for projections or extensions.

12. Limit pie charts to five or fewer segments.

13. Make bars of columns wider than the spaces between them.

14. Order shade patterns from darkest to lightest.

15. Place labels in shaded areas of surface chart if possible.

16. Order data in surface charts so irregular layers are on top.

17. Avoid semilog scales for audiences unfamiliar with them.

18. Order shade patterns on maps to facilitate remembering scale ends.

19. Don't allow grid lines to pass through columns or bars.

20. Avoid garish shade patterns.

21. Use a single type font from chart to chart to maintain consistency.

Following these guidelines will result in a clear presentation of the data.

As a final tip don't be afraid of drawing or typing on your plots. You may want to add labels, reference lines, or other information. Color, patterns, shading, or pictures may be added manually to increase the impact of the graphics and allow you to create effects not possible or convenient with your computer system. Combine media for more interesting presentations. Use of a slide bureau or graphic arts department may be necessary for certain types of graphics and are resources that should not be overlooked.

You should always be proud of your graphics.

Now, let's talk about how we can get SAS/Graph to make these great graphics.

Using PROC GPLOT

We'll use PROC GPLOT to make some plots and slowly add options to get the plot we want.

PROC GPLOT

Features:
- Automatically scales the values on axes or axes and tick marks can be explicitly defined.
- Provides horizontal and vertical reference lines.
- Plots can be superimposed.
- A variety of symbols can be used to represent points.
- Colors, symbols, line styles and interpolation method can be selected.

Input and output options allow you to designate the data set to be plotted or the output graphics catalog where the plot will be sent. Annotate data sets can be specified.

- DATA=SASdataset
- GOUT=graphicscatalog
- ANNOTATE=SASdataset
- ANNO=SASdataset

The general plot statement looks like this

- PLOT yvariable=xvariable/.../options
- OVERLAY
- AREAS=n
- SKIPMISS
- LEGEND=LEGENDn
- NOLEGEND

Some of the options provide: overlaid plots, filled in areas, gaps in lines at missing values, and control of legend statements.

PROC GPLOT

PLOT yvariable=xvariable/.../options
- OVERLAY
- AREAS=n
- SKIPMISS
- LEGEND=LEGENDn
- NOLEGEND

195
Plots can be enhanced with symbols, patterns, titles and footnotes and can be plotted using a "BY" variable.

**PROC G PLOT options**

- **SYMBOLn options**
- **PATTERNn options**
- **TITLEn options 'text'
- **FOOTNOTEn options 'text'
- **NOTE options 'text'
- **BY variables**

Options allow for the control of axes or frames and color attributes for text.

**PROC G PLOT**

```
PROC GPLOT;
PLOT variable+variable.../options;
```

A variety of color and line type options are provided.

**PROC G PLOT**

```
PROC GPLOT;
PLOT variable+variable.../options;
```

This is some fictitious sales data for an even more fictitious pet store.

Now that we have our SAS data set let's see what it looks like. Using the simplest form of the GPLOT procedure:

**PROC GPLOT:**

```
PROC GPLOT;
PLOT AMT * MONTH;
```

196
We get the following graph.

It needs some work doesn't it.

By changing the PLOT Statement to:

```
PROC GPLOT;
PLOT AMT * MONTH = PETS;
```

We get colored points and a legend telling us what color goes with what pet type.

What we want are lines to show the sales of each pet. In addition, we want markers at each point and thicker lines. The SYMBOL statement provides these options:

Symbol Statement

Example:

```
SYMBOL Options
  C=color (depends on output device)
  V=symbol (A~W and special symbols)
  F=font (generally I.ft out)
  H=height (height of symbol=PCT IN CM or Cell)
  W=width (width of line)
  JoinMethod (JOIN HILO RLINE)
  line_type (for dashed lines)
```

By using the following symbol statements we can specify the height of the symbol, line width, color of the line, symbol to be used and the type of line drawn (a simple line joining the points).

```
PROC GPLOT;
PLOT AMT * MONTH = PETS;
SYMBOL1 H=2 W=2 C=BLACK V=SQUARE I=JOIN;
SYMBOL2 H=2 W=2 C=RED  V=HASH   I=JOIN;
SYMBOL3 H=2 W=2 C=GREEN V=TRIANGLE I=JOIN;
SYMBOL4 H=2 W=2 C=BLUE  V=STAR   I=JOIN;
```

The result is:

We now notice the Y axis starts at 20 and ends at 90. Remembering our rules for effective chart design we want the graph to start at 0 to allow us easier comparisons. The axis options provide the solution by providing a VAXIS statement.
VAXIS Examples

Specification

10 20
10, 20, 30
10, 20, 30, 40
10, 20, 30, 40, 50
10, 20, 30, 40, 50, 60
10, 20, 30, 40, 50, 60, 70
10, 20, 30, 40, 50, 60, 70, 80
10, 20, 30, 40, 50, 60, 70, 80, 90
10, 20, 30, 40, 50, 60, 70, 80, 90, 100
10, 20, 30, 40, 50, 60, 70, 80, 90, 100

We include the VAXIS option in our program:

PROC GPLOT;
        PLOT AMT * MONTH = PETS/VAXIS=0 TO 100 BY 10;
        SYMBOL1 H=2 W=2 C=BLACK V=SQUARE I=JOIN;
        SYMBOL2 H=2 W=2 C=RED V=HASH I=JOIN;
        SYMBOL3 H=2 W=2 C=BLUE V=TRIANGLE I=JOIN;
        SYMBOL4 H=2 W=2 C=GREEN V=STAR I=JOIN;

The result looks like this.

Now, let's really fix this plot. A PROC FORMAT is used to replace the month numbers with month names:

PROC FORMAT:
        VALUEMFMT 1= 'JAN'
        VALUEMFMT 2= 'FEB'
        VALUEMFMT 3= 'MAR'
        VALUEMFMT 4= 'APR'
        VALUEMFMT 5= 'MAY'
        VALUEMFMT 6= 'JUN'
        VALUEMFMT 7= 'JUL'
        VALUEMFMT 8= 'AUG'
        VALUEMFMT 9= 'SEP'
        VALUEMFMT 10= 'OCT'
        VALUEMFMT 11= 'NOV'
        VALUEMFMT 12= 'DEC'

A TITLE statement is used to provide a title.

Example:

 Title options "Title" (depends on output device)
 Width (see list)
 Height (height of characters in Cm or Cell)
 Alignment (LEFT RIGHT CENTER)
 Angle (left -90 to +90)
 Rotation (character rotation)
 B=0,1,2,3 (draw box)
 U=0,1,2,3 (underlines)

The resulting program is shown. Notice the use of the HPOS and VPOS statements in the GOPTIONS statement. This tells the plotter or other output device how many character positions there are (in this case 80x40). By changing the parameters the size of the lettering is changed.
The final result:

This plot can be read even without the colors present.

Now let's try something different. We want a graph of each pet and a regression line showing the projected sales in December.

SAS makes this simple. We use the "I" parameter in the SYMBOL statement. The following interpolation options are available.

In addition, the R series (for regression) option allows us to choose Linear Quadratic or Cubic regression with confidence limits on the Mean or Individual points with 20%, 95% or 99%.

We'll sort by PET, PROC GPLOT by PET, and use I=RLCLM95 for regression line w/confidence limits on the mean at 95%.
The program:

```
DATA SALES:
PROC FORMAT:
VALUEMFMT 1='JAN'
VALUEMFMT 2='FEB'
VALUEMFMT 3='MAR'
VALUEMFMT 4='APR'
VALUEMFMT 5='MAY'
VALUEMFMT 6='JUN'
VALUEMFMT 7='JUL'
VALUEMFMT 8='AUG'
VALUEMFMT 9='SEP'
VALUEMFMT 10='OCT'
VALUEMFMT 11='NOV'
VALUEMFMT 12='DEC'
PROC SORT; BY MONTH:
PROC GCHART; BY PETE:
PLOT AMT * MONTH /
    VAR=MONTH;
    HSLAB=1 TO 12;
    VHSLAB=1 TO 12;
    VAXIS=NONE;
    HAXIS=NONE;
    SYMBOL=H
    SYMBOLS=BLACK
    VSYMBOL=BLACK
    PATTERN=BLACK
    NOCLONE;
    TITLE=(H, 'PETE'S PETS');
    LABEL AMT=PROFIT IN $;
    FORMAT MONTH MFMT.;
```

The output is four plots. I'm just showing the first and last:

```
PETE'S PETS
```

Looks like lizards are the way to go.

Using PROC GCHART

PROC GCHART provides the following chart types:

- Horizontal bar
- Vertical bar
- Block
- Pie
- Star

The basic form is:

```
PROC GCHART
```

Option allow the use of annotate data sets, the ability to sort the chart in ascending or descending sequence (good for Pareto charts) and axis options for labeling the axes.

```
```

200
Suppressing options allow you to remove axes, stats symbols and legends if desired:

**PROC GCHART**

```plaintext
HBAR variables... / options
or YBAR variables... / options
Options:
  * NOAXIS
  * NOAXES
  * NOSTATS
  * NOLOG
  * NOLEGEND
  * NOZEROS
```

Color options provide control over the frame and text colors.

**PROC GCHART**

```plaintext
HBAR variables... / options
or YBAR variables... / options
Options:
  * CAXIS=color
  * CFRAME=color
  * COUTLINE=color
  * CTEXT=color
  * FRAME
  * PATTERN=method
```

The grouping options allow you to determine what the major group and subgroup variables will be. If you group a bar chart by month you will get one set of bars for each month. Then you can use SUBGROUP's variable to determine how each month will be divided.

**PROC GCHART**

```plaintext
HBAR variables... / options
or YBAR variables... / options
Options:
  * GROUP=variable
  * CSSPACE=number
  * SUBGROUP
  * SPACE=number
  * LEGEND=legend number
  * LEVELS=number
  * DISCRETE
  * MINOR=n
  * MISSING
```

Bar representation options are used to specify the type of statistic each bar represents. Frequency is the default unless SUMVAR= variable is used. When SUMVAR is specified TYPE=SUM.

**PROC GCHART**

```plaintext
HBAR variables... / options
or YBAR variables... / options
Options:
  * TYPE=FREQ
  * TYPE=CFRQ
  * TYPE=PERCENT
  * TYPE=CPERCENT
  * TYPE=SUM
  * TYPE=MEAN
```

In horizontal bar charts the following options specify which statistics are printed on the chart.

**PROC GCHART**

```plaintext
HBAR variables... / options
or YBAR variables... / options
Options:
  * FREQ
  * CFRQ
  * PERCENT
  * CPERCENT
  * SUM
  * MEAN
  * SUMVAR
```

(SUMVAR= must be specified for SUM + MEAN)
Let's try out a PROC GCHART on our pet data. Our first program looks like this:

```
DATA SALES;
PROC GCHART;
VBAR MONTH/SUMVAR=AMT
   GROUP=PET
   DISCRETE;
```

The result of this may surprise you. We've asked for a vertical bar chart grouped by PET by month. The result is:

A more terse representation would be to do the following:

```
PROC GCHART;
VBAR MONTH/SUMVAR=AMT
   SUBGROUP=PET
   DISCRETE;
```

How we use SUBGROUP=PET to split each bar into the contributing variable.

In both cases SUMVAR=AMT is used to create a plot of total sales and the DISCRETE option is included to get one bar for each month.

A BLOCK chart of this data is hard to read so it will not be included. Remember, although 3D charts are impressive they are often a poor way to present data.

I hope this brief tour of these procedures encourages you to try them out. Take a simple data set and experiment with the options. You'll be an expert in no time.