SAS® Macros for Graphical Data Analysis
Doug Lassman, PRA International, Lenexa, KS

ABSTRACT
SAS/Graph has many useful procedures to aid in graphical data analysis including PROC GPLAY, PROC GCHART and PROC G3D. Combining these with PROC GREPLAY allows for even more powerful displays of data. Sometimes the exact display is not available with the previously mentioned procedures. This is where PROC GANNO and ANNOTATE datasets come to the rescue. Virtually any display imaginable can be created with a combination of dots, lines, bars, polygons and text. After a brief overview of ANNOTATE datasets, example SAS macros will be presented and discussed.

INTRODUCTION
SAS Graph has many useful procedures to aid in graphical data analysis but sometimes the exact display is not available with PROC GPLAY, PROC GCHART, or PROC G3D. Fortunately, SAS/Graph also allows the production of custom graphs using PROC GANNO and ANNOTATE datasets. First the ANNOTATE facility is discussed and then two example macros are presented and discussed.

ANNOTATE – AN OVERVIEW
Custom plots can be easily created with SAS annotate datasets. First a dataset is created and then a SAS/Graph procedure is called with an option of anno=dataset. The procedure can be PROC GPLAY or PROC GCHART if the annotate dataset contains desired labeling or special graphics such as error bars. If the entire presentation is coded in the annotate dataset, PROC GANNO can be used.

One of the most confusing (but versatile) aspects of the annotate facility is specifying the coordinate systems. These are the values for XSYS, YSYS, ZSYS and HSYS. There are 12 possible values but two are the most common. The absolute coordinate is predominantly used over the relative coordinate system. If labels or error bars are being added to a GPLAY or GCHART, the annotation would typically be confined to the data area (as opposed to the graphics output area and the procedure output areas) and the location of the annotation within the data area will be a value that is specified on the x and y axes. This would then be an XSYS, YSYS, ZSYS, and HSYS value of '2'. In situations where the entire presentation is contained in an annotate dataset, the graphics output area would most likely be used. The location of the annotation within the output area can be specified by either percentage of the area (0 to 100) or by the cell. Since the values of the cell are dependent upon GOPTIONS such as VPOS and HPOS but the percentage values are not, using percentage values would be preferred and the XSYS, YSYS, ZSYS and HSYS value would then be '3'.

After the coordinates are specified, then the graphics elements can be constructed. Although there are several elements, the most basic are the bar, the line, the label and the point. Others that will not be discussed include circles, pies, arcs and polygons. To create a bar, the x and y values of the lower-left corner is specified with a annotate function of 'move' and the observation is output. Then the x and y values of the opposite corner are specified with the annotate function of 'bar' and that observation is output. The color, line, and style can also be specified. To draw a line, the x and y value of the first point is specified with the annotate function of 'move' and the observation is output. The color, line, and style can also be specified. For a label, the x and y values and the function of 'label' are specified and the observation is output. The text, color, style, size, and position can also be given.

With these four elements, many graphs can be created.

EXAMPLES
Two macros have been developed to aid in the graphical analysis of data. The first creates a plot that is essentially the observation along the x-axis and a bar chart for the specified variables along the y-axis. The second is a Gabriel’s biplot which displays the correlational structure of specified variables.

The ALLGRAPH macro has four arguments: the dataset, the number of variables to plot, the number of variables to sort, and the list of the variables. The data are then sorted as specified (the first n variables in the list) and then a bar chart for each variable is created. These are then stacked so that each observation is vertically aligned. Outliers are easily detected with this display. They can be noted and removed for further analysis. Also with this display, other trends can be observed. The data can be sorted differently and redisplayed just by changing the macro arguments.

The BIPLOT macro has six arguments: the dataset, the number of variables, the list of variables, the principal component for the x and y axes, and optionally a class variable. This plot shows the relationship of the variables, the relative similarities of the individual data points, and the relative values of the observations for each variable. The plot is interpreted as follows: the length of the vector variable indicates how well the two dimensions represent that variable. The angle between two variable vectors represent their pairwise correlation. If the angle is near 0 degrees they are positively correlated. If the angle is near 180 degrees they are negatively correlated. If the angle is 90 degrees, they are not correlated. The spatial proximity of individual observations reflects their similarities in the two dimensions plotted. Finally, the relative values of the observations for a particular variable are seen by projecting the observation points onto the variable vector. Also, with this display, different combinations of the principal components can easily be produced.

CONCLUSION
Using the annotate facility, powerful custom graphs can easily be constructed. Two examples have been presented and further examples of tools for graphical data analysis will be posted to the web address below.

REFERENCES
CONTACT INFORMATION
Your comments and questions are valued and encouraged.
Contact the author at:
   Doug Lassman
   PRA International
   16400 College Blvd.
   Lenexa, KS 66219
   Work Phone: 913-438-7696
   Fax: 913-599-0344
   Email: lassmandoug@praintl.com
   Website : www.geocities.com/lassmandoug/sas.htm

TRADEMARK INFORMATION
SAS and SAS/GRAPH are registered trademarks of SAS Institute Inc. in the USA and other countries.
** Macro allgraph **

* Macro allgraph
*
* A macro to create a barchart for each 
* variable specified in the dataset.
*
* DSN - Dataset name
* NUMVARS - Number of variables 
* SORTVAR - Number of variables to sort 
* VARS - List of variables
*
******************************************************************************

%let one=1;

** Create macro variables for each variable**;

%do i=1 %to &numvars;
%let var&i=%scan(&vars,&i);
%end;

** Sort the data by the n sortvars **:

proc sort data=&dsn;
by %do i=1 %to &numsort; &var&i %end; ;
run;

data _null_;
set &dsn end=eof;
if eof then call symput('numobs',_n_);
run;

** Create the annotate dataset for each var**:

%do i=1 %to &numvars;
proc means data=&dsn noprint;
var &var&i;
output out=meanout min=min max=max;
run;

data _null_; 
set meanout;
call symput('min',min);
call symput('max',max);
run;

data ganno&i; 
set &dsn;
min=1*&min;
max=1*&max;
xlast=2+98*(_n_/&numobs+&one);
ylast=100*%eval(_n_/&numvars);
xnew=2+98*(_n_/&numvars+&one);
result=&var&i;
ynew=ylast+100*((result-min)/(max-

** Set all annotate datasets together **;

data anno;
length function $8;
set %do i=1 %to &numvars; ganno&i %end; xsys='3';
ysys='3';
color='Black';
function='Label';
%do i=1 %to &numvars; %end;

%allgraph(pharmsug.explore,10,4,tmt sex 
response baseline duration age bmi height 
wt_base wt_chg);
** BIPLOT MACRO
%macro biplot(dsn,numvars,vars,xaxis,yaxis,class);
**********************************************
* Macro biplot
* A macro to create a Gabriels Biplot for a *
given dataset.
* DSN  - Dataset name
* NUMVARS - Number of variables
* VARS  - List of variables
* XAXIS - Principal comp displayed on x-axis
* YAXIS - Principal comp displayed on y-axis
* PRIN1 to PRINx
* CLASS - Class variable
*********************************************;
** Proc PRINCOMP to get the principal **
components **;
proc princomp data=&dsn outstat=outstat
out=proc princomp data=88dsn outstat-outstat
out-out noprint;
  var &vars;
run;
quit;
** Create macro vars for each variable **;
%do i=1 %to &numvars;
  %let var&i=%scan(&vars,&i);
%end;
** Create the annotate dataset of the **
component vectors **;
proc sort data=outstat
out=proc princomp data=88dsn outstat-outstat
out-out(where=(/_type_='SCORE'));
  by _name_;
run;
%do i=1 %to &numvars;
  data prin&i;
    set sort;
    if &i=compress(_name_,'PRIN');
    %do j=1 %to &numvars;
      order=&j;
      var=upcase(symget("var&j"));
      prin&i=&&&var&j;
      output;
    %end;
    keep order var prin&i;
run;
proc sort; by order; run;
%end;
** Create annotate dataset of data points **;
proc means data=xp noprint;
  vari (&xaxis &yaxis);
run;
%if &class= %then %do;
  data out;
    set out;
    group=1;
run;
%end;
%else %do;
  proc sort data=out; by &class; run;
  data out;
    set out;
    by &class;
    if _N_=1 then group=1;
    else if first.&class then group=group+1;
    retain group;
run;
%end;
** Create annotate dataset of data points **;
data _null_;
  set minmax;
  xmax=max(abs(min1),max1);
  ymax=max(abs(min2),max2);
  call symput('xmax',xmax);
  call symput('ymax',ymax);
run;
%if &class= %then %do;
  data out;
    set out;
    group=1;
run;
%end;
%else %do;
  proc sort data=out; by &class; run;
  data out;
    set out;
    by &class;
    if _N_=1 then group=1;
    else if first.&class then group=group+1;
    retain group;
run;
%end;
data ano;
  set xp;
  length text $8 function $8;
xsys='3';
ysys='3';
color='Black';
x=50;
y=50;
function='Move';
output;
x=50+(&xaxis/&xmax)*45;
y=50+(&yaxis/&ymax)*45;
function='Draw';
output;
text=var;
function='Label';
  if &xaxis<0 and &yaxis<0 then position='D';
  if &xaxis>0 and &yaxis<0 then position='F';
  if &xaxis<0 and &yaxis>0 then position='A';
  if &xaxis>0 and &yaxis>0 then position='C';
  output;
run;

** Create annotate dataset for the axis and labels **;

data axis;
  length function $8;
  xsys='3';
  ysys='3';
  color='Black';
  x=5;  y=5;  function='Move';  output;
  x=95;  y=5;  function='Draw';  output;
  x=5;  y=95;  function='Move';  output;
  x=95;  y=95;  function='Draw';  output;
run;

data axislabl;
  length function $8 text $18;
  xsys='3';
  ysys='3';
  color='Black';
  x=50;  y=0;  position='B'; function='Label';
  text='&xaxis';  output;
  x=0;  y=50;  position='E'; function='Label';
  angle=90; text='&yaxis'; output;
run;

data anno;
  set axislabl anno anno1 axis ;
run;

title "Gabriel's Biplot";
proc ganno anno=anno;
run;
%mend;

** Example Call **;

%biplot(pharmsug.explore,7,baseline duration age bmi height wt_base wt_chg ,PRIN1,PRIN2,tmt);