Paper 243-2008

Butterflies, Heat Maps, and More: Explore the New Power of SAS/GRAPH®

Susan Schwartz, SAS Institute Inc., Cary, NC

ABSTRACT

In SAS® 9.2, SAS/GRAPH® introduces the statistical graphics (SG) procedures. The SG procedures provide an easy way to produce commonly used analytical graphs. This presentation will demonstrate how to use these new tools to create butterfly plots, heat maps, risk maps, stacked plots, and other unique charts. Sample code will be presented so you can start building your own library of new charts. You will learn how to create paneled plots to compare results across multiple classification variables and overlay multiple plots onto one chart.

INTRODUCTION

An effective plot can reveal trends or patterns in your data that might be difficult to see in tabular form. The new SAS/GRAPH procedures SGPLOT and SGPANEL provide additional tools to help you view your data in interesting and effective ways. Because you can create overlaid plots, the output you can produce is limited only by your imagination. This paper will highlight the power of the new SAS/GRAPH SG procedures.

STACKED BAND PLOTS

A band plot is a graph that applies a color scale to two variables. Each color band represents some commonality between the x and y values. Bands are useful for displaying ranges, whether from the performance of an investment portfolio, the distribution of behaviors, or the progression of an infectious disease.

Figure 1 shows a stacked band plot that displays the four categories of the Body Mass Index (underweight, normal weight, overweight, and obese) as a function of height and weight. It was generated with an SGPLOT procedure and the BAND statement.

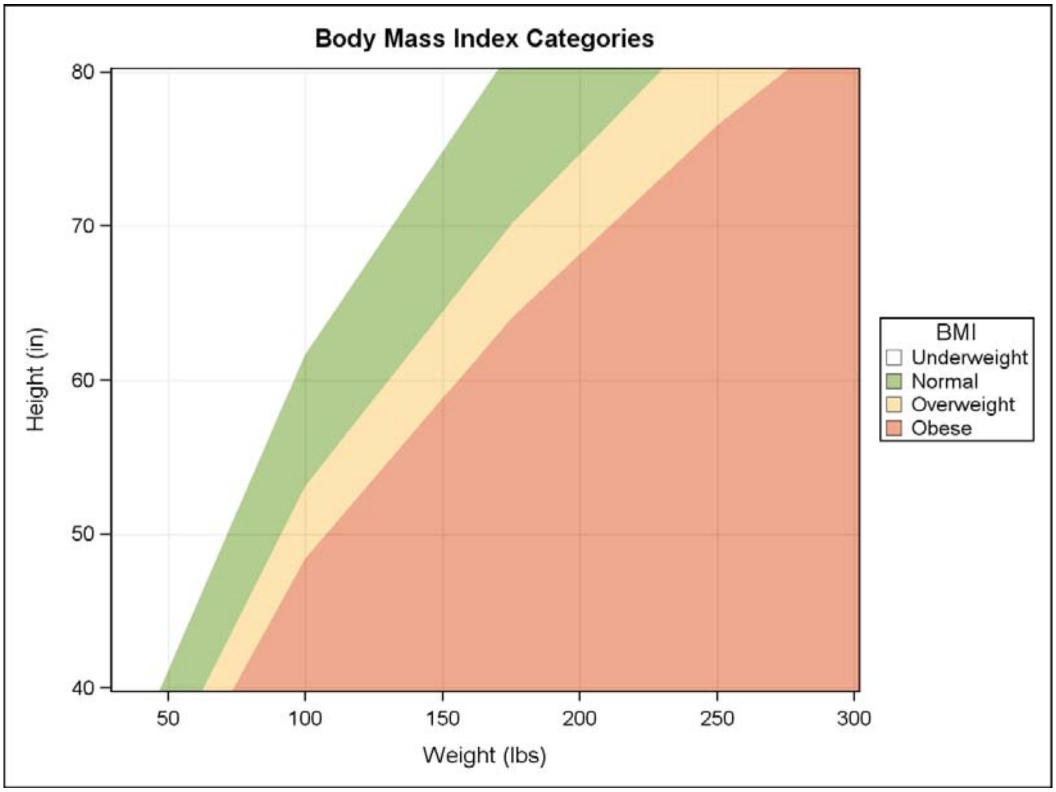


Figure 1: Stacked Band Plot using the SGPLOT procedure

CODE FOR STACKED BAND PLOT:

The Body Mass Index (BMI) and the BMI categories are defined by a formula. Our example will apply the BMI formula to five different weights as shown in Figure 2:

Body Mass Index (BMI):	BMI categories:	Weights (in lbs):
$BMI = 703 \times \frac{weight\,(lb)}{height^2\,(in^2)}$	Underweight: BMI < 18.5 Normal: 18.5 ≤ BMI < 25.0 Overweight: 25.0 ≤ BMI < 30.0 Obese: BMI ≥` 30.0	25.0 100.0 175.0 250.0 325.0

Figure 2: Body Mass Index (BMI) categories and body weights for Stacked Band Plot

First, define the BMI categories and generate the BMI categories for five different weight values:

```
data bmi;
   infile datalines;
   input category $ 1-11 upper_bmi lower_bmi weight_bmi;
   lower_hgt=(703*weight_bmi/lower_bmi)**.5;
   upper_hgt=(703*weight_bmi/upper_bmi)**.5;
   nn=_n_;
   datalines;
Underweight 1    18.5    25.0 4
Underweight 1    18.5   100.0 4
Underweight 1    18.5   175.0 4
Underweight 1    18.5   250.0 4
Underweight 1    18.5   325.0 4
Normal      18.5 25.0    25.0 5
Normal      18.5 25.0   100.0 5
Normal      18.5 25.0   175.0 5
Normal      18.5 25.0   250.0 5
Normal      18.5 25.0   325.0 5
Overweight  25.0 30.0    25.0 6
Overweight  25.0 30.0   100.0 6
Overweight  25.0 30.0   175.0 6
Overweight  25.0 30.0   250.0 6
Overweight  25.0 30.0   325.0 6
Obese       30.0 150.0  25.0   7
Obese       30.0 150.0  100.0 7
Obese       30.0 150.0  175.0 7
Obese       30.0 150.0  250.0 7
Obese       30.0 150.0  325.0 7
;
run;
```

Second, define a style that has a green to red color ramp:

```
proc template;
define style styles.bmi;
  parent=styles.listing;
  style graphcolors from graphcolors /
   'gcdata1'=CX31035E
   'gdata1'=CXFFFFFF
   'gdata3'=CXFDC861
   'gdata4'=CXDC531F
   'gdata2'=CX679920;
end;
run;
```

Third, generate the band plot specifying that X is WEIGHT_BMI, the upper and lower limits of the Y band are UPPER_HGT and LOWER_HGT respectively, and the GROUP is CATEGORY. The bands automatically change color for each value of the group variable.

```
ODS listing style=bmi;
title 'Body Mass Index Categories';

proc sgplot data=bmi;
band x=weight_bmi
  upper=upper_hgt lower=lower_hgt /
  transparency=.5
  group=category
  name="bmi";
xaxis min=30 max=300 label='Weight (lbs)' grid;
yaxis min=40 max=80  label='Height (in)'  grid;
keylegend "bmi" /
  position=right across=1 title='BMI';
run;
```

The XAXIS and YAXIS statements define the range of the axis values, the axis labels, and enable the display of vertical and horizontal grid lines. By default, each band has a color transparency set to 0 (opaque) which would hide the grid lines; the transparency is set to .5 to let the grid lines appear. The KEYLEGEND statement generates a legend on the right side of the plot.

USING STACKED BAND PLOTS AS A RISK MAP

Now that we have a visual of the four BMI categories we can determine a person's BMI category by plotting the height and weight directly on the band plot. Consider that we have the growth records of three patients taken over 20 years and we plot all of the height and weight combinations onto the BMI map. By adding a SERIES plot to our BAND plot we can produce the risk map in Figure 3.

CODE FOR RISK MAP:

First, define the patient data:

```
data records;
   infile datalines;
   nn=_n_;
   input wgt hgt year name $ ;
datalines;
45 52 1975 Bob
80 56 1980 Bob
125 66 1986 Bob
175 67 1990 Bob
200 68 1995 Bob
215 69 2000 Bob
250 69 2005 Bob
80  43 1975 Fred
85  58 1980 Fred
135 70 1986 Fred
180 70 1990 Fred
195 70 2005 Fred
50  42 1975 Jack
135 54 1980 Jack
140 64 1985 Jack
160 73 1990 Jack
170 76 2000 Jack
180 76 2005 Jack
;
run;
```

Second, merge the bmi data set with the records data set. Note that both data sets contain a variable *nn* which is used only to construct the merged observations.

```
proc sort data= records;
  by nn;
run;

proc sort data= bmi;
  by nn;
run;

data merged;
  merge records bmi;
  by nn;
  if name='' then name='Jack';
run;
```

The bmi data set contains 20 observations and the records data set contains only 18. By default, the NAME variable will be a missing value in the last two observations. Replace the missing value with the valid name ‘Jack’ so the NAME legend will be correct.

Lastly, construct the final plot:

```
ODS listing style=bmi;
title '3 Patient Growth Charts with their BMI Categories';

proc sgplot data=merged;
band x=weight_bmi
  upper=upper_hgt lower=lower_hgt /
  transparency=.5 group=category name="bmi";
xaxis min=30 max=300 label='Weight (lbs)' grid;
yaxis min=40 max=80  label='Height (in)'  grid;
keylegend "bmi" / position=right across=1 title='BMI';
series x=wgt y=hgt /
  datalabel=year
  lineattrs=(pattern=solid thickness=2px)
  markers markerattrs=(symbol=circlefilled )
  group=name name="pts";
keylegend "pts" / position=bottom title='Patient: ';
inset "BMI=703 x Weight / Height (*ESC*){sup '2'}"  " "/
  position=bottomright textattrs=graphfootnotetext;
run;
```

The BAND, XAXIS, YAXIS, and the first KEYLEGEND statements are the same as before. The SERIES, second KEYLEGEND, and INSET statements are overlaid onto the previous plots. The result is a composite plot of all the elements.

The SERIES statement specifies that X=WGT and Y=HGT (patient height and weight) and GROUP=NAME. Options specify that solid circle markers are labeled with the year of the measurement, and these are connected by solid lines that are two pixels wide. Just as in the band plot, the color of the markers and lines will change with each value of the group.

The second KEYLEGEND statement generates a legend of the patient names in the SERIES. The INSET displays the formula for the BMI computation. The string **(*ESC*){sup '2'}** invokes the INSET’s special formatting controls to display the 2 as a superscript.

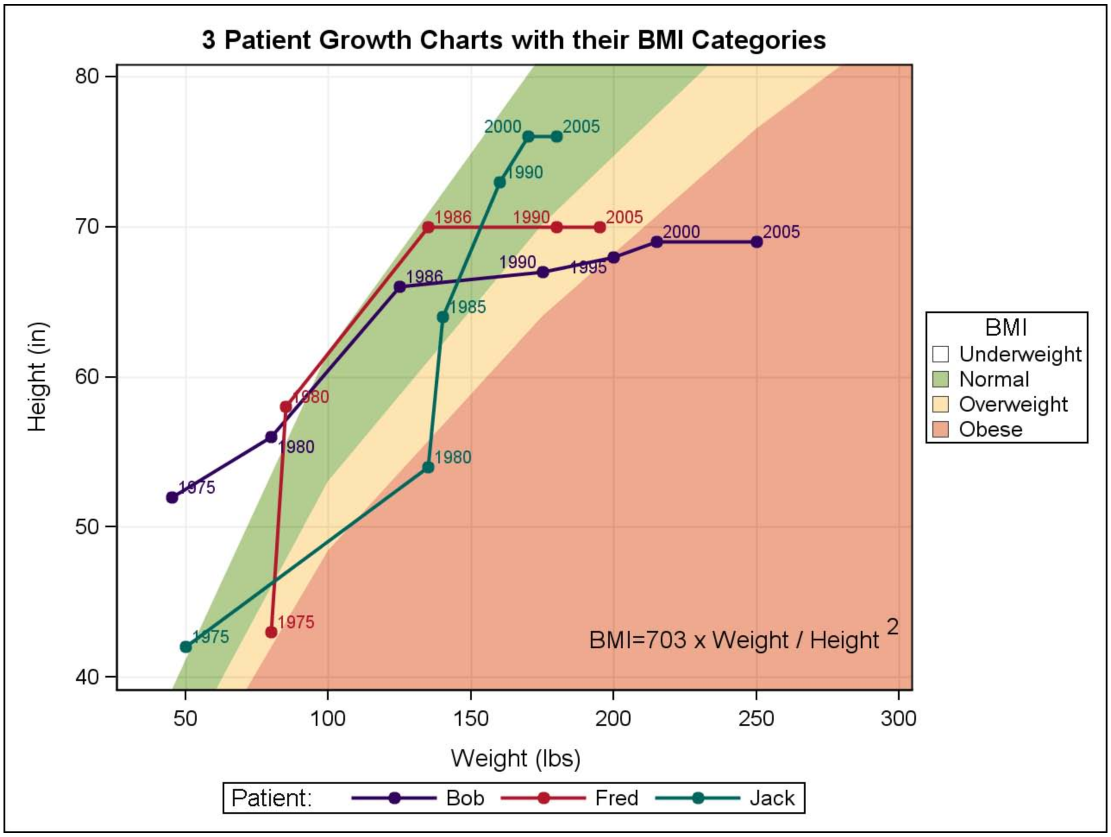


Figure 3: Stacked Band and Series Plots using the SGPLOT procedure

PANELED GRAPHS

A paneled graph is a grid of cells generated across multiple classification variables. The classification variables can be arranged in rows and columns (lattice) or in a series of multiple classifier values (panel). The cell area can be as simple or complex as desired. Bar charts, series plots, regression plots, and overlaid plots are just some of the graphs available for paneling.

LATTICE LAYOUT

With a lattice layout you specify two classification variables. The first classification variable determines the total number of columns of the panel, and the second classification variable determines the total number of rows of the panel.

Figure 4a shows a horizontal bar chart within a lattice panel. It was generated with the SGPANEL procedure and HBAR statement.

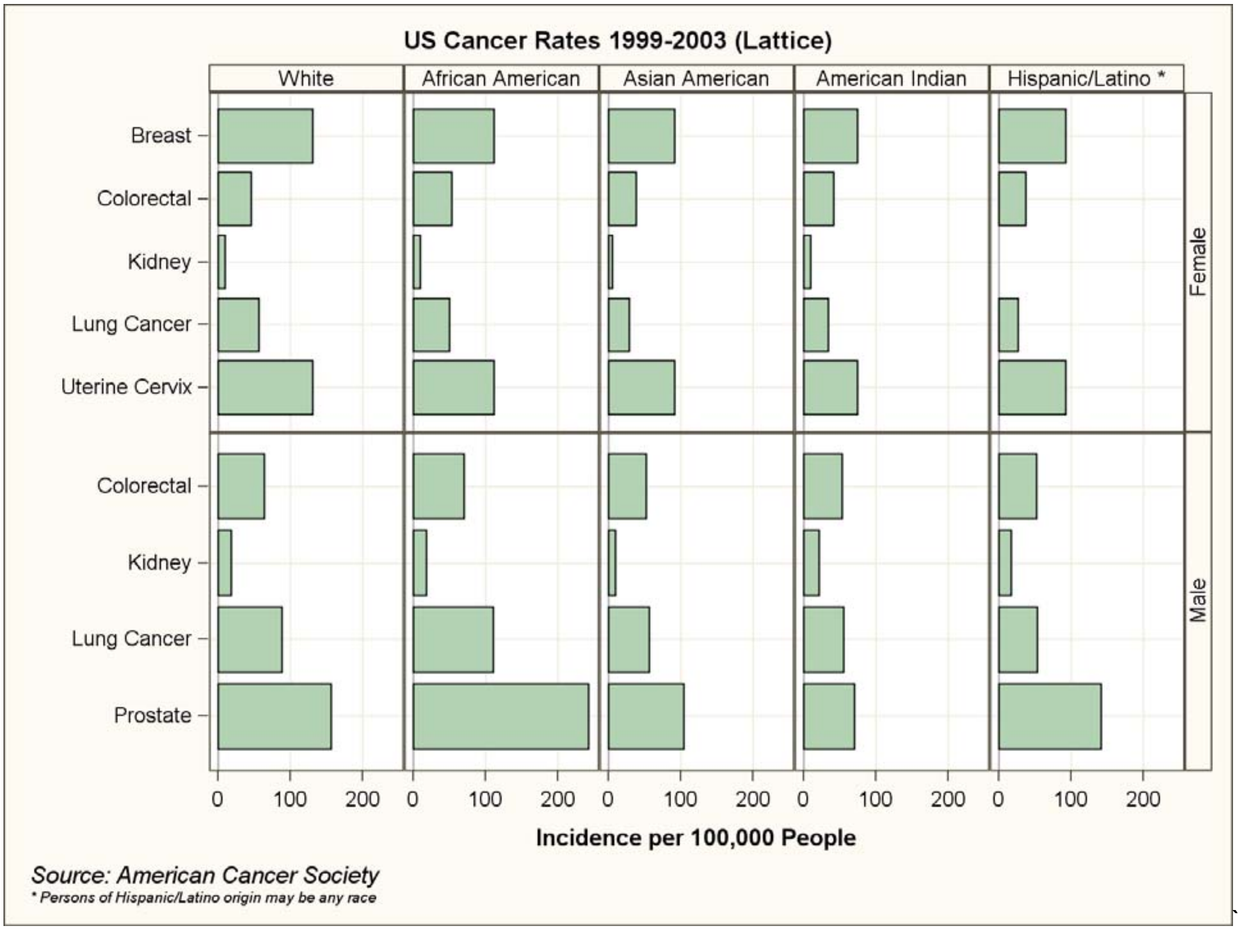


Figure 4a: Paneled Lattice Plot using the SGPANEL procedure (*race x cause*)

Figure 4a indicates some interesting information about cancer incidence in the U.S.

- In the male population, prostate cancer has the highest incidence, and it occurs more in the African American population than in other ethnic groups.
- In the female population, breast and uterine/cervical cancers appear to have the same incidence rate, and the rate is the same across all ethnic groups.
- These indications warrant further analysis for confirmation.

CODE FOR LATTICE LAYOUT:`

First, generate the data set and user-defined format:

```
data work.cancer;
   infile datalines;
   input cause $ 1-20 sex $ race incidence;
   datalines;
Lung Cancer         Male    1   88.8
Lung Cancer         Male    2  110.6
Lung Cancer         Male    3   56.6
Lung Cancer         Male    4   55.5
Lung Cancer         Male    5   52.7
Lung Cancer         Female  1   56.2
Lung Cancer         Female  2   50.3
Lung Cancer         Female  3   28.7
Lung Cancer         Female  4   33.8
Lung Cancer         Female  5   26.7
Colorectal          Male    1   63.7
Colorectal          Male    2   70.2
Colorectal          Male    3   52.6
Colorectal          Male    4   52.7
Colorectal          Male    5   52.4
Colorectal          Female  1   45.9
Colorectal          Female  2   53.5
Colorectal          Female  3   38.0
Colorectal          Female  4   41.9
Colorectal          Female  5   37.3
Breast              Female  1  130.8
Breast              Female  2  111.5
Breast              Female  3   91.2
Breast              Female  4   74.4
Breast              Female  5   92.6
Prostate            Male    1  156.0
Prostate            Male    2  243.0
Prostate            Male    3  104.2
Prostate            Male    4   70.7
Prostate            Male    5  141.1
Kidney              Male    1   18.0
Kidney              Male    2   18.5
Kidney              Male    3    9.8
Kidney              Male    4   20.9
Kidney              Male    5   16.9
Kidney              Female  1    9.3
Kidney              Female  2    9.5
Kidney              Female  3    4.9
Kidney              Female  4   10.0
Uterine Cervix      Female  1  130.8
Uterine Cervix      Female  2  111.5
Uterine Cervix      Female  3   91.2
Uterine Cervix      Female  4   74.4
Uterine Cervix      Female  5   92.6
;
run;
proc format;
value ethnicity
1 = "White"
2 = "African American"
3 = "Asian American"
4 = "American Indian"
5 = "Hispanic/Latino *";
run;
```

Then use the SGPANEL procedure to generate the lattice:

```
ODS listing style=analysis;
title 'US Cancer Rates 1999-2003 (Lattice)';

footnote justify=left italic 'Source: American Cancer Society';
footnote2 justify=left italic h=6pt
   '* Persons of Hispanic/Latino origin may be any race';

proc sgpanel data=cancer;
format race ethnicity.;
panelby race sex /
  columns=5
  layout=lattice
   novarname
  uniscale=column;
hbar cause / response=incidence;
colaxis grid label='Incidence per 100,000 People';
rowaxis grid label=' ';
run;
```

The PANELBY statement requests a lattice LAYOUT with five columns and the NOVARNAME option. By default the classifier variable name and value are included in the cell header. The NOVARNAME option changes the header to the value only.

The PANELBY statement also specifies UNISCALE=COLUMN. UNISCALE=ALL is the default which does a full union of all row values and all column values. This would force all values of cancer CAUSE to display for each SEX. Because several of the cancer types are gender specific, we want to eliminate the empty rows.

The HBAR statement specifies that each cell should contain a horizontal bar chart with Y=CAUSE and X=INCIDENCE. The HBAR statement does not contain the STAT= option, so it defaults to SUM.

The COLAXIS and ROWAXIS statements refer to the panel's X and Y axes (not each cell's X and Y axes). They define the column and row axis labels and turn on column and row grid lines in each cell.

The lattice chart in Figure 4a enables you to compare the incidence of cancer within one sex or within one race. What is more difficult to examine is the incidence of any one type of cancer across sex and race.

Figure 4b shows the same lattice with RACE and CAUSE switched on the PANELBY and HBAR statements:

```
proc sgpanel data=cancer;
format race ethnicity.;
panelby cause sex /
  columns=5
  layout=lattice
   novarname
  uniscale=column;
hbar race / response=incidence;
colaxis grid label='Incidence per 100,000 People';
rowaxis grid label=' ';
run;
```

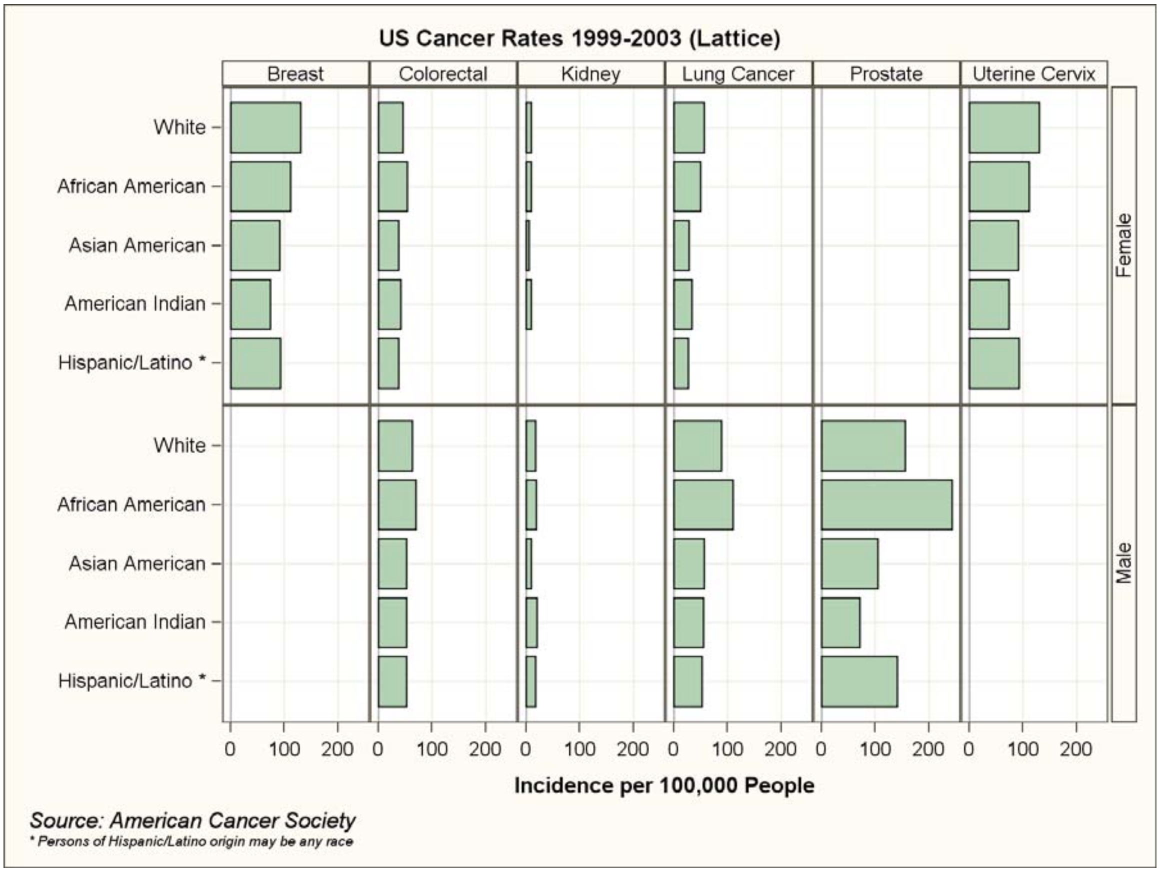


Figure 4b: Paneled Lattice Plot using the SGPANEL procedure (*cause x race*)

Figure 4b permits a different examination of U.S. cancer incidence:

- In the male population, we still see that prostate cancer has the highest incidence and occurs more in the African American population. But now we see that lung cancer is also higher in the male African American population than in other male populations.
- In the female population, breast and uterine/cervical cancers now appear to have a higher incidence rate in White females than in other female ethnic groups.
- Lung cancer has the second highest incidence rate In the male population, while it is third in the female population.
- Again, these indications warrant further analysis for confirmation.

PANEL LAYOUT:

With a panel layout cell headers are added to each cell indicating the current values of the classifiers.

Figure 5 shows a vertical bar chart within a panel layout. It was generated with the SGPANEL procedure and VBAR and REFLINE statements. Note how the cell headers cycle from left to right and top to bottom.

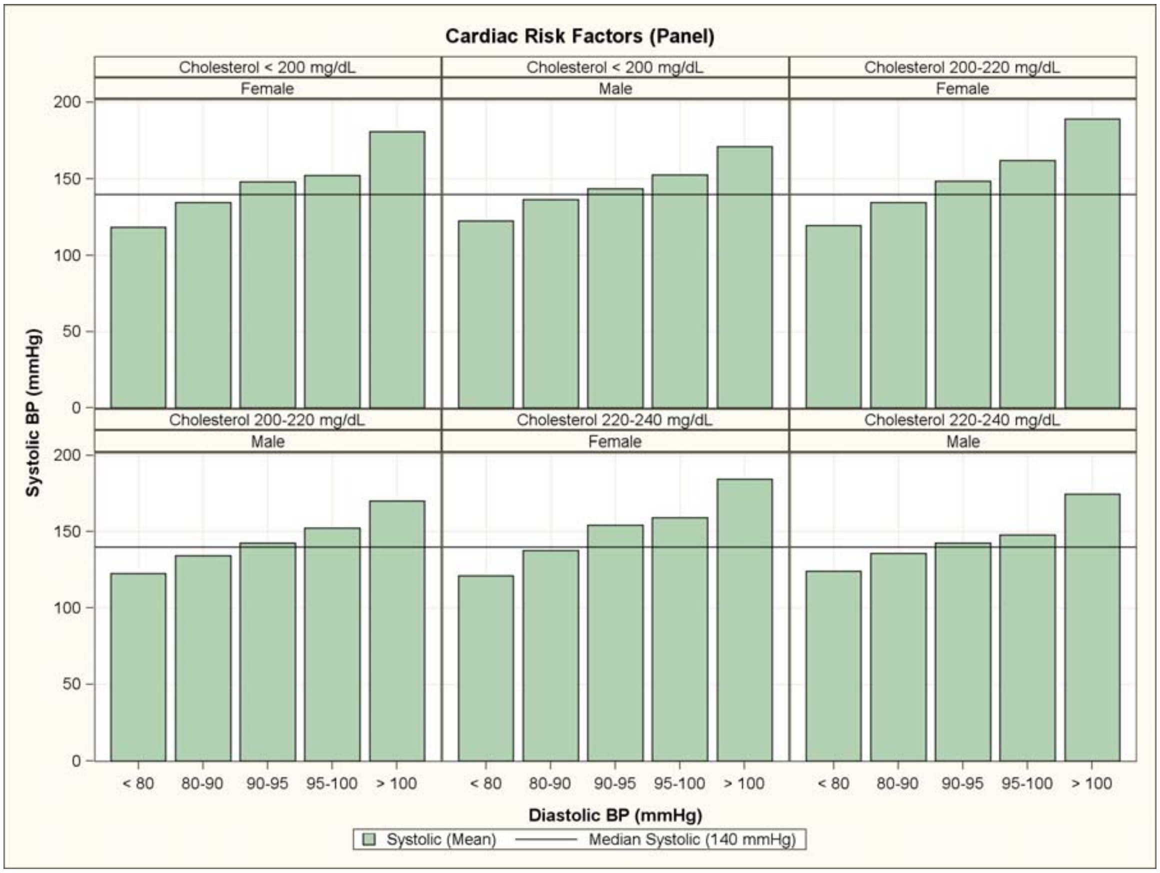


Figure 5: Paneled Layout Plot using the SGPANEL procedure

CODE FOR PANEL LAYOUT:

Define user-defined formats, and use the SGPANEL procedure to generate the panel layout:

```
proc format;
value cholfmt
    0-200 = "Cholesterol < 200 mg/dL"
  201-220 = "Cholesterol 200-220 mg/dL"
  221-240 = "Cholesterol 220-240 mg/dL"
  241-999 = "Cholesterol > 240 mg/dL";

value diasfmt
    0-80 = "< 80"
   81-90 = "80-90"
   91-95 = "90-95"
  96-100 = "95-100"
 101-999 = "> 100";
run;

ODS listing style=analysis;
title 'Cardiac Risk Factors (Panel)';

proc sgpanel data=sashelp.heart;
format cholesterol cholfmt.
       diastolic   diasfmt.;
panelby cholesterol sex /
        columns=3 rows=2
        novarname;
vbar diastolic /
     response=systolic
     stat=mean;
refline 140 / axis=y
     lineattrs=(color=black)
     legendlabel='Median Systolic (140 mmHg)'
     name=”refline;
colaxis grid label='Diastolic BP (mmHg)';
rowaxis grid label='Systolic BP (mmHg)';
run;
```

The PANELBY statement sets the classifiers to be CHOLESTEROL and SEX, requests a 3 x 2 panel LAYOUT, and sets the NOVARNAME option. By default the classifier variable name and value are included in the cell header. The NOVARNAME option changes the header to the value only.

CHOLESTEROL and DIASTOLIC are aggregated values because there formats contain “many to one” mappings.

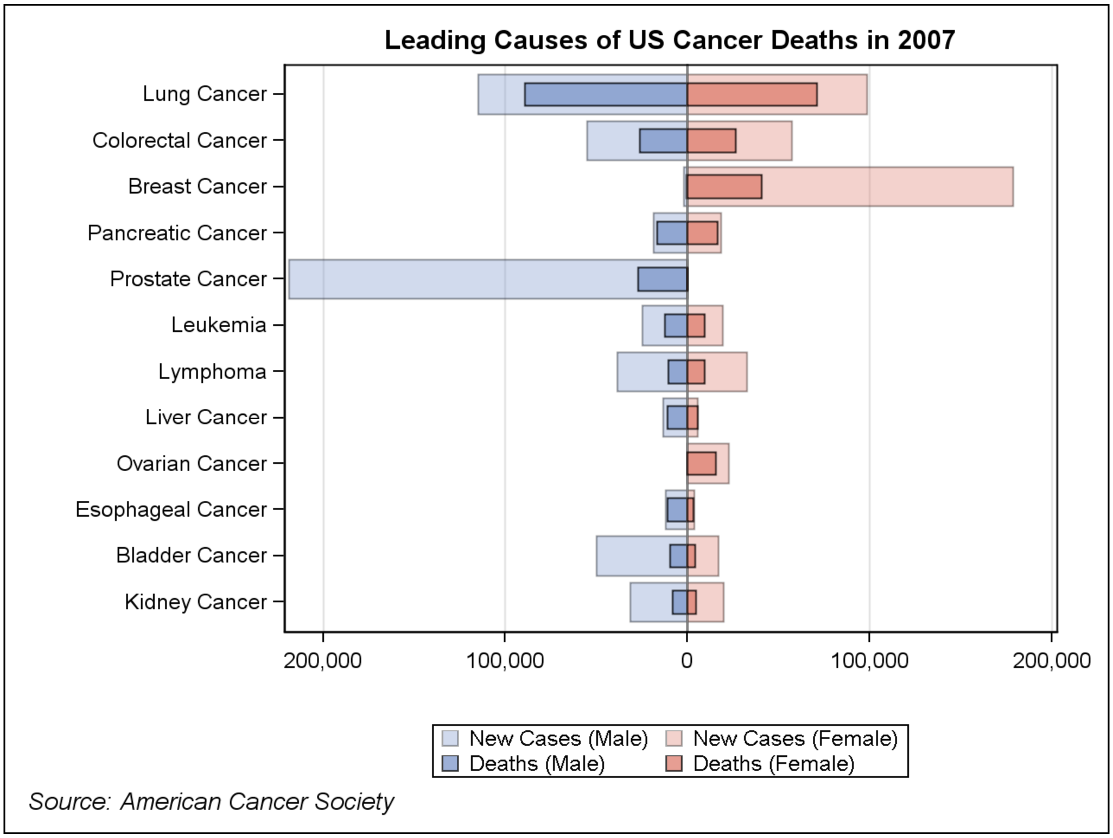
The VBAR statement specifies that each cell should contain a vertical bar chart with X=DIASTOLIC and Y=SYSTOLIC. The mean SYSTOLIC value is computed because STAT=MEAN is specified.

The REFLINE statement draws a horizontal line at 140 on the Y axis, manually sets the line color to black, and defines the text for the legend. The COLAXIS and ROWAXIS statements refer to the panel’s X and Y axes (not the cell’s X and Y axes). The statements define the column and row axis labels and turn on column and row grid lines in each cell.

The data set SASHELP.HEART contains eight crossings of SEX and the formatted classifier CHOLESTEROL. (CHOLESTEROL has four and SEX has two). That would normally produce a panel with eight cells. The PANELBY statement explicitly asks for three rows and two columns, so the SGPANEL procedure produces two images, the first with six cells, and the second with two cells. Figure 5 contains only the first panel.

BUTTERFLY PLOTS

A butterfly plot is a plot drawn across a centered axis, and is particularly useful when comparing two halves of a whole. Figure 6 shows a butterfly plot generated with the SGPLOT procedure.



Source: American Cancer Society

Figure 6: Butterfly Plot using the SGPLOT procedure

This butterfly plot presents a great deal of information in one compact picture:

- The larger, more transparent bars show the number of new diagnoses by cancer type in 2007.
- The smaller, more opaque bars show the number of deaths by cancer type in 2007.
- The size of a small bar relative to its larger bar shows the ratio of deaths to diagnoses. Pancreatic cancer has the highest deaths to diagnoses ratio.
- The blue bars show diagnoses and deaths for men, the pink bars for women.
- The data are ordered from top to bottom by the total number of cancer deaths. Of the 12 most deadly forms of cancer in 2007, lung cancer was the most deadly and kidney cancer the least.

CODE FOR BUTTERFLY PLOT:

First, generate the data set. The butterfly plot is really displaying negative data (for males) and positive data (for females). The DATA step multiples the values of MCASES and MDEATHS by -1. The negative (left) side of the butterfly will display as positive values via a picture format.

```
data work.cancer;
  infile datalines;
  input cause $ 1-20 mcases fcases mdeaths fdeaths;
  deaths=mdeaths + fdeaths;
  mcases= -1 * mcases;
  mdeaths= -1 * mdeaths;
  datalines;
Lung Cancer         114760  98620  89510  70880
Colorectal Cancer    55290  57050  26000  26180
Breast Cancer         2030 178480    450  40460
Pancreatic Cancer    18830  18340  16840  16530
Prostate Cancer     218890      0  27050      0
Leukemia             24800  19440  12320   9470
Lymphoma             38670  32710  10370   9360
Liver Cancer         13650   5510  11280   5500
Ovarian Cancer           0  22430      0  15280
Esophageal Cancer    12130   3430  10900   3040
Bladder Cancer       50040  17120   9630   4120
Kidney Cancer        31590  19600   8080   4810
;
run;
```

Second, sort the data set in descending order by deaths, so the cancer with the most deaths is first:

```
proc sort data=cancer out=cancer_sorted;
by descending deaths;
run;
```

Third, define the picture format that will be used to display the negative (male) values as positive values on the axis. Both negative and positive values display with a comma format. The value “0” is intentionally excluded from the picture format to force it to display on the axis.

```
proc format;
picture positive low-<0='000,000'
                0<-high='000,000';
run;
```

Lastly, use the SGPLOT procedure with four HBAR statements to generate the butterfly plot:

```
ODS listing style=listing;
title 'Leading Causes of US Cancer Deaths in 2007';
footnote justify=left italic 'Source: American Cancer Society';
proc sgplot data=cancer_sorted;
format mcases mdeaths fcases fdeaths positive.;
hbar cause / response=mcases
  fillattrs=graphdata1 transparency=.65
  legendlabel="New Cases (Male)" name="mcases";
hbar cause / response=mdeaths
  barwidth=.5 fillattrs=graphdata1 transparency=.25
  legendlabel="Deaths (Male)" name="mdeaths";
hbar cause / response=fcases
  fillattrs=graphdata2 transparency=.65
  legendlabel="New Cases (Female)" name="fcases";
hbar cause / response=fdeaths
  barwidth=.5 fillattrs=graphdata2 transparency=.25 name="fdeaths"
  legendlabel="Deaths (Female)" name="fdeaths";
keylegend "mcases" "fcases" "mdeaths" "fdeaths" / across=2;
xaxis label=" " grid;
yaxis label=" " discreteorder=data;
run;
```

The first HBAR statement draws the bars for MCASES. The color is set to the graphdata1 style element with the transparency set to .65, the text for the legend is set via the LEGENDLABEL option, and the statement is named for reference in the KEYLEGEND statement to follow.

The second HBAR draws the bars for MDEATHS. The color is again set to the graphdata1 style element but with the transparency set to .25, setting this bar to a more opaque version of the color used for MCASES. The BARWIDTH is set to .5 so the MDEATHS bar overlays only half of the MCASES bar. The text for the legend is set via the LEGENDLABEL option, and the statement is named for later reference.

The third HBAR statement draws the bars for FCASES. The color is set to the graphdata2 style element with the transparency set to .65, the text for the legend is set via the LEGENDLABEL option, and the statement is named for later reference.

The fourth HBAR statement draws the bars for FDEATHS. The color is set to the graphdata2 style element with the transparency set to .25, setting this bar to a more opaque version of the color used for FCASES. The BARWIDTH is set to .5 , text for the legend is set via the LEGENDLABEL option, and the statement is named for later reference.

The KEYLEGEND statement generates legend entries for the four HBAR statements in the desired order, two to a line.

The XAXIS statement draws a blank X axis label and vertical grid lines. The YAXIS statement draws a blank Y axis label and tells SGPLOT to preserve the incoming data order. Because the data have already been sorted in descending order (by total number of deaths), the bars are displayed in that order without total number of deaths appearing anywhere on the plot.

PANELED BUTTERFLY PLOT

By making a small number of changes to our butterfly plot code, we can compare US Cancer Deaths from 2007 to US Cancer Deaths in 1997 in a paneled display of the butterfly plot (Figure 7).

First, add data from 1997 along with a variable YEAR to the data set:

```
data work.cancer;
   infile datalines;
   input cause $ 1-20 Year $ mcases fcases mdeaths fdeaths;
   deaths=mdeaths + fdeaths;
   mcases= -1 * mcases;
   mdeaths= -1 * mdeaths;
   datalines;
Lung Cancer         2007 114760  98620  89510  70880
Colorectal Cancer   2007  55290  57050  26000  26180
Breast Cancer       2007   2030 178480    450  40460
Pancreatic Cancer   2007  18830  18340  16840  16530
Prostate Cancer     2007 218890      0  27050      0
Leukemia            2007  24800  19440  12320   9470
Lymphoma            2007  38670  32710  10370   9360
Liver Cancer        2007  13650   5510  11280   5500
Ovarian Cancer      2007      0  22430      0  15280
Esophageal Cancer   2007  12130   3430  10900   3040
Bladder Cancer      2007  50040  17120   9630   4120
Kidney Cancer       2007  31590  19600   8080   4810
Lung Cancer         1997  98300  79800  94400  66000
Colorectal Cancer   1997  45500  48600  22600  24000
Breast Cancer       1997   1400 180200   290   43900
Pancreatic Cancer   1997  13400  14200  13500  14600
Prostate Cancer     1997 334500      0  41800      0
Leukemia            1997  15900  12400  11770   9540
Lymphoma            1997  34200  26900  13220  12060
Liver Cancer        1997   9100   4500   7500   4900
Ovarian Cancer      1997      0  26800      0  14200
Esophageal Cancer   1997   9400   3100   8700   2800
Bladder Cancer      1997  39500  15000   7800   3900
Kidney Cancer       1997  17100  11700   7000   4300
;
```

```
  run;
```

Second, change the sort to include YEAR as a sort key:

```
  proc sort data=cancer out=cancer_sorted;
    by descending year descending deaths;
  run;
```

Use the same user-defined format POSITIVE:

```
  proc format;
    picture positive low-<0='000,000'
                    0<-high='000,000';
  run;
```

Lastly, we make several changes to the SGPLOT code:

- Change the SGPLOT procedure to the SGPANEL procedure.
- Add a PANELBY statement to generate one cell for each YEAR.
- Change the XAXIS and YAXIS statements to COLAXIS and ROWAXIS statements.

```
  ODS listing style=listing;
  title 'Leading Causes of US Cancer Deaths 2007 vs. 1997';
  footnote justify=left italic 'Source: American Cancer Society';

  proc sgpanel data=cancer_sorted;
  panelby year / columns=1 uniscale=all;
  format mcases mdeaths fcases fdeaths positive.;
  hbar cause / response=mcases
    fillattrs=graphdata1 transparency=.65
    legendlabel="New Cases (Male)" name="mcases";
  hbar cause / response=mdeaths
    barwidth=.5 fillattrs=graphdata1 transparency=.25
    legendlabel="Deaths (Male)" name="mdeaths";
  hbar cause / response=fcases
    fillattrs=graphdata2 transparency=.65
    legendlabel="New Cases (Female)" name="fcases";
  hbar cause / response=fdeaths
    barwidth=.5 fillattrs=graphdata2 transparency=.25
    legendlabel="Deaths (Female)" name="fdeaths";
  keylegend "mcases" "fcases" "mdeaths" "fdeaths" / across=2;
  colaxis label=" " grid;
  rowaxis label=" " discreteorder=data;
  run;
```

The paneled butterfly plot is displayed in Figure 7.

All the attributes of the original butterfly plot are still true:

- The larger, more transparent bars show the number of new diagnoses by cancer type in 2007 and 1997.
- The smaller, more opaque bars show the number of deaths by cancer type in 2007 and 1997.
- The size of a small bar relative to its large bar shows the ratio of deaths to diagnoses. Pancreatic cancer has the highest deaths to diagnoses ratio in both 2007 and 1997.
- The blue bars show diagnoses and deaths for men, the pink bars for women.
- The data are ordered from top to bottom by the total number of cancer deaths. Of the 12 most deadly forms of cancer, lung cancer was the most deadly in both 2007 and 1997.

The paneled butterfly plot reveals some additional information:

- There were more diagnoses of lung cancer in 2007 than in 1997.
- The death to diagnosis ratio for lung cancer was slightly better in 2007 than in 1997.
- There were many fewer diagnoses of prostate cancer in 2007 compared to 1997, but only slightly fewer deaths.

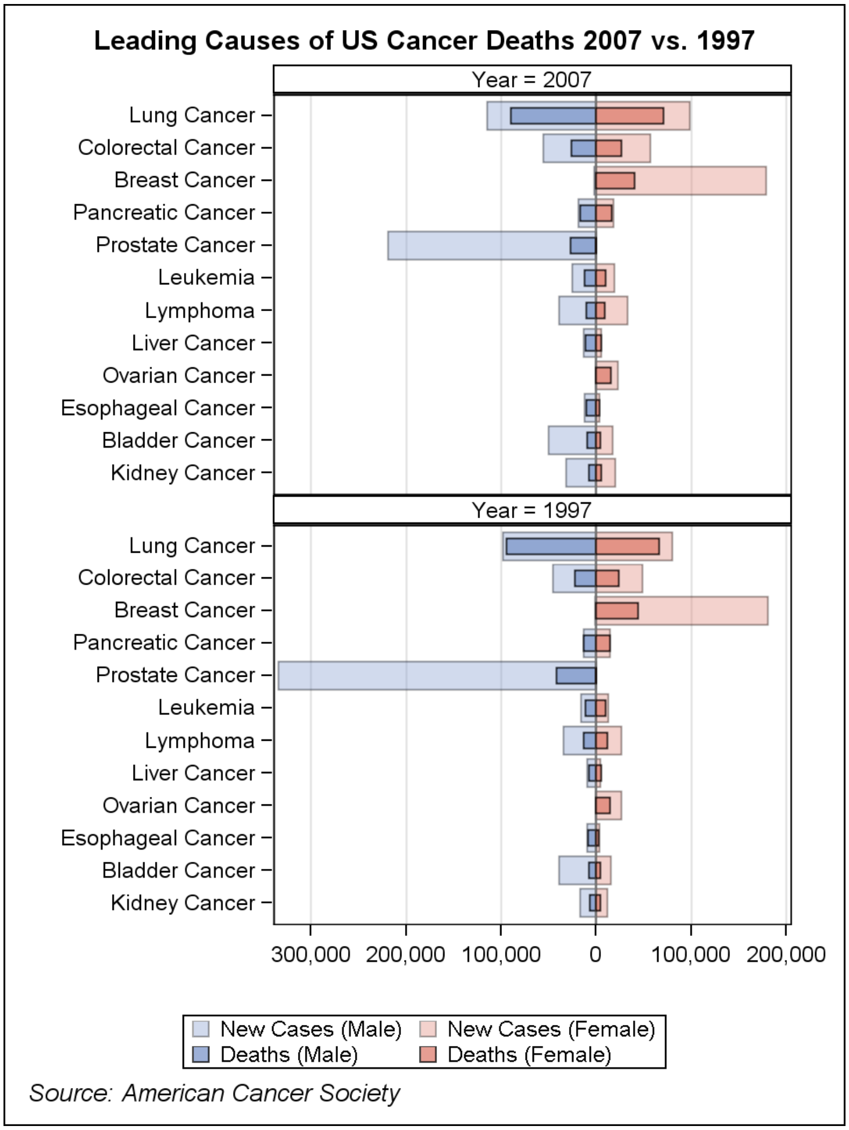


Figure 7: Paneled Butterfly Plot using the SGPANEL procedure

HEATMAP

A heatmap is a graph that displays combinations of two categorical variables via the use of tiles, with each tile colored according to a color scale. The color scale represents some third characteristic common to the categorical variables.

Figure 8 shows a heatmap generated by the SGPLOT procedure and a SCATTER statement.

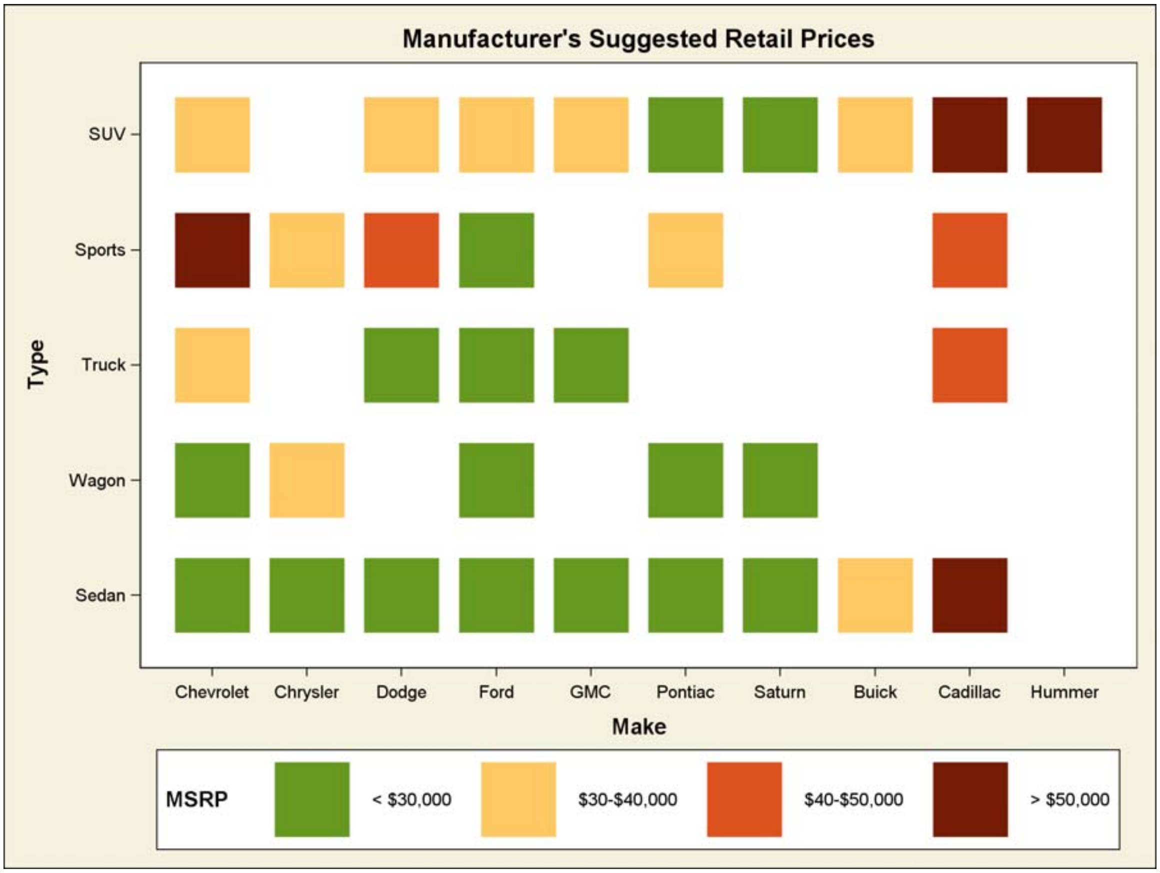


Figure 8: Heatmap using the SGPLOT procedure

CODE FOR HEATMAP:

First, define the color scale and legend format:

```
/*  define green to red color ramp  */
proc template;
define style styles.heatmap;
  parent=styles.harvest;
  style graphcolors from graphcolors /
   'gcdata1'=CX679920
   'gcdata2'=CXFDC861
   'gcdata3'=CXDC531F
   'gcdata4'=CX751C07;
end;
run;

proc format;
  value price
    1='< $30,000'
    2='$30-$40,000'
    3='$40-$50,000'
    4='> $50,000';
run;
```

Next, subset the data set, compute the mean MSRP by make and type, and assign the group category variable:

```
proc sort data=sashelp.cars (where=(make in
  ('Buick' 'Cadillac' 'Chevrolet' 'Chrysler' 'Dodge'
   'GMC' 'Ford' 'Saturn' 'Pontiac' 'Hummer')))
  out=cars_sorted;
  by make type;
run;

proc means data=cars_sorted noprint;
  by make type;
  output out=cars_average mean=avgmsrp;
run;

data cars_final;
  set cars_average;
  format group price.;
  label group='MSRP';
  if avgmsrp le 30000 then group=1;
    else if avgmsrp le 40000 then group=2;
    else if avgmsrp ge 50000 then group=3;
    else group=4;
run;
```

Then sort by the group category to ensure that the legend values are in order:

```
proc sort data=cars_final out=cars_heatmap;
  by group;
run;
```

Lastly, generate the heatmap with MAKE and TYPE as the classifiers (X and Y values), and the formatted MSRP as the GROUP:

```
ODS listing style=heatmap;
title "Manufacturer's Suggested Retail Prices";

proc sgplot data=cars_heatmap;
scatter x=make y=type /
  markerattrs=(size=.28in symbol=squarefilled)
  group=group;
run;
```

The SCATTER statement uses an enlarged filled square as the marker. Normally, when a group variable is used, the scatter point markers change appearance (symbol and color) for each group value. Because the marker symbol is being set explicitly, only the color will change from group to group.

CONCLUSION

As you have seen in these examples the new SGPLOT and SGPANEL procedures are quite powerful. The ability to overlay plots gives you the flexibility to create plots specific to your needs. ODS styles enable you to simply and effectively use colors. The paneling features provide a concise way to view data by multiple levels of classification. And best of all, the procedure syntax is straightforward and easy to use. We look forward to seeing the creative plots the SAS community will develop using these new procedures.

REFERENCES

The Body Mass Index (BMI) formula used to create the stacked band plots is available from http://en.wikipedia.org/wiki/Heat_map.

The data for the butterfly plots is available from the American Cancer Society website: www.cancer.org.

ACKNOWLEDGMENTS

The author would like to thank Dan Heath and Sanjay Matange for their invaluable input.

RECOMMENDED READING

Heath, Dan. 2007. "New SAS/GRAPH Procedures for Creating Statistical Graphics in Data Analysis."
Proceedings of the 2007 SAS Global Forum. April, 2007.
http://www2.sas.com/proceedings/forum2007/193-2007.pdf

Heath, Dan. 2008. "Effective Graphics Made Simple using SAS/GRAPH 'SG' Procedures."
Proceedings of the 2008 SAS Global Forum. April, 2008

Matange, Sanjay. 2008. "Statistical Graphics Editor." *Proceedings of the 2008 SAS Global Forum.* April, 2008

Matange, Sanjay. 2008. "Introduction to Graph Template Language." *Proceedings of the 2008 SAS Global Forum.* April, 2008

O'Connor, Daniel. 2008. "SAS® Graphics on ODS 9.2 Performance-Enhancing Steroids." *Proceedings of the 2008 SAS Global Forum.* April, 2008

CONTACT INFORMATION

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

Susan Schwartz
SAS Institute
SAS Campus Drive
Cary, NC 27513
susan.schwartz@sas.com