Sensitivity Training for Building Better Bar Charts with SAS/GRAPH® Software

Perry Watts, SAS® User, Elkins Park, PA

ABSTRACT

This workshop combines instructions for chart building with principles defined by the statistical graphics experts, Edward Tufte and William Cleveland. Step-by-step instructions are provided for building midpoint, group and subgroup charts that use the graphics principles to generate publication-quality output. Gratitude is extended to SAS Institute Inc. for granting permission to use the DEPT data set from *The How-To Book for SAS/GRAPH® Software* by Thomas Miron **[4]**.

Besides learning the GCHART procedure, time is spent on looking at how the PATTERN, AXIS, FORMAT and LEGEND statements are incorporated into chart building. Instructions are also provided in the workshop for creating a high-resolution graphics file and showing how it is transferred to PowerPoint for display.

Even if you have minimal experience with SAS/GRAPH software, you should come away from this presentation knowing how to create a polished bar chart with GCHART.

THE BAR CHART AS A GRAPHIC CONSTRUCT

Definition:

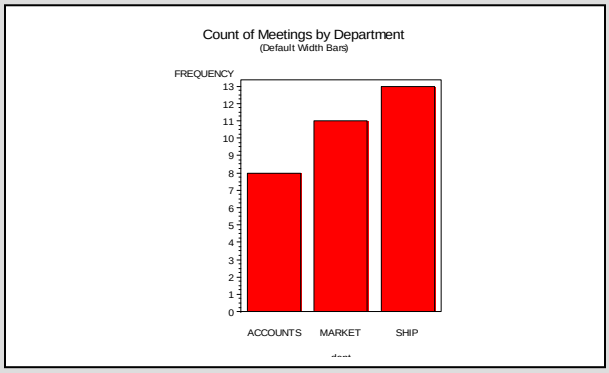
From Wikipedia:

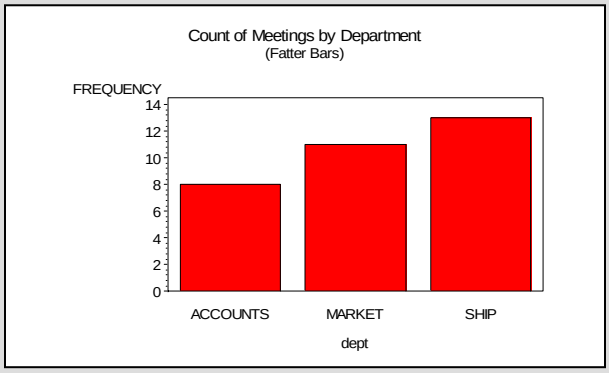
> A **bar chart**, also known as a **bar graph**, is a chart with rectangular bars of lengths usually proportional to the magnitudes or frequencies of ***what*** they represent. Bar charts are used for ***comparing*** two or more values. The bars can be horizontally or vertically oriented **[6]** (italics added).

Missing from the definition is a description of the "what" that is being represented. For the bar chart that "what" is discrete nominal or ordinal data. Histograms are better suited for displaying ranges of continuous data.

It can also be deduced from the Wikipedia definition that information in a bar chart is conveyed solely by bar length. As Figure 1 demonstrates, bar widths can be changed without altering the graph's message.

Figure 1. Two bar charts convey exactly the same information, because the width of the bars has no intrinsic meaning.

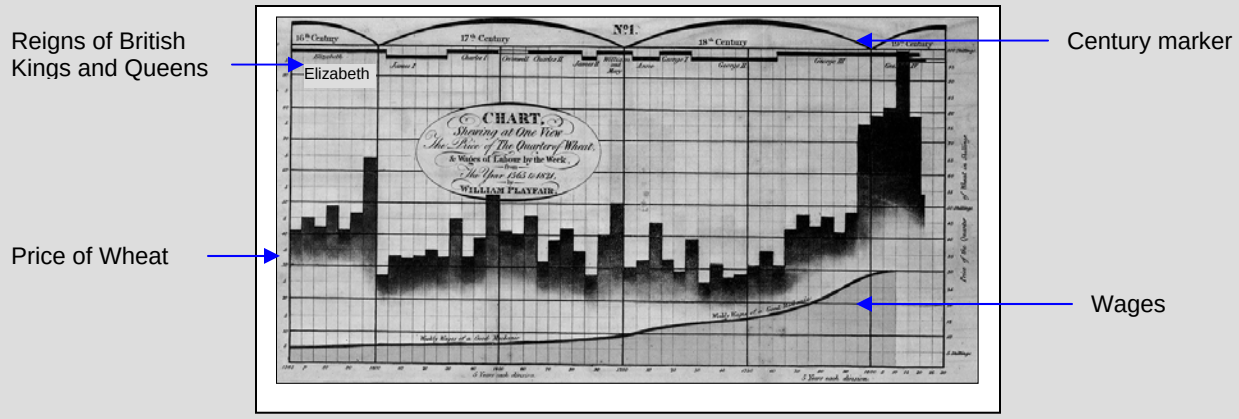




History:

Possibly the first bar charts appeared in the *Commercial and Political Atlas* (London, 1786) by William Playfair **[6]**. Displayed in Figure 2 is a hazy reproduction of Playfair's chart entitled *CHART, Shewing at One View The Price of The Quarter of Wheat, & Wages of Labour by the Week --from-- The Year 1565 to 1824*. A more readable version of this graph appears in *The Visual Display of Quantitative Information* by Edward Tufte **[5, p. 34]**.

Figure 2. An early bar chart by William Playfair compares the price of wheat to the wages for skilled laborers.



Playfair enhanced his chart by adding line plots for wages, reigns of monarchs and centuries. With these enhancements he anticipated the arrival of PROC GBARLINE in version 9 that produces an augmented vertical bar chart with a (line) plot overlay **[15, p. 739].** Nevertheless, the stand-alone bar chart with its simple structure has managed to endure over the centuries. Possibly its longevity can be attributed to its transparency. Bar charts use summary data to convey brief messages quickly!

The Bar Chart Gets Bad Press

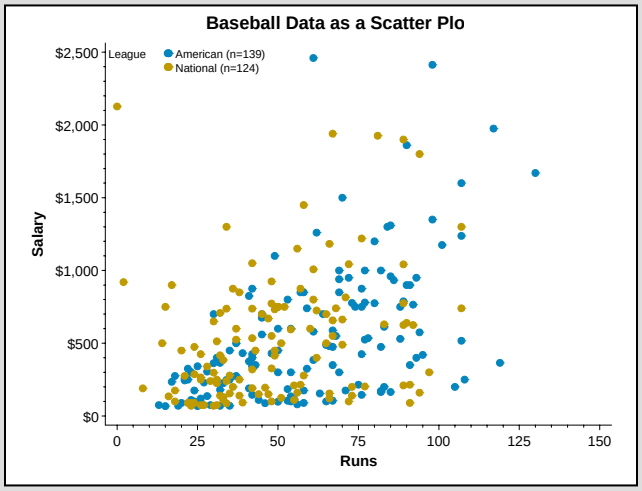
Despite their popularity, graphics experts do not have a high regard for bar charts. Cleveland makes no reference to them in *The Elements of Graphing Data* **[2]**, and Tufte points to a low *data-ink ratio* defined below as a reason for not taking them seriously.

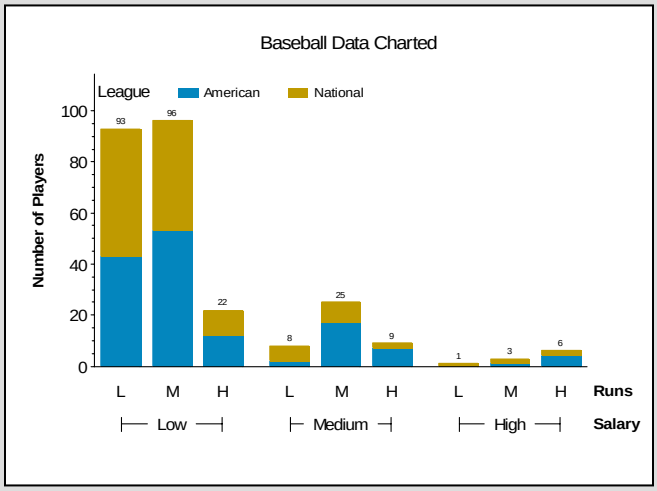
$$\text{Data-ink ratio} = \frac{\text{data-ink}}{\text{total ink used to print the graphic}}$$

$$= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information } \textbf{[5, p. 93]}$$

For Tufte, the only data-ink in a bar chart is the height or altitude of the bars. Furthermore, he shows that redundancy increases when a single labeled shaded bar conveys the same "altitude in six separate ways (any five of the six can be erased and the sixth will still indicate the height)" **[5, p. 96].** However, the data-ink ratio is not an appropriate metric to apply when the stated goal of a graphic is to *compare* results (see Wikipedia definition). In Figure 3, for example, the scatter plot fails to convey any difference between the major baseball leagues in terms of runs vs. salary even though the data-ink score is much higher. Part of the failure can be attributed to the overlapping plotting symbols that represent ties within and between leagues.

Figure3. Tufte's data ink ratio collides with Cleveland's maxim that "overlapping plotting symbols must be visually distinguishable" **[2, p. 50]**. Overlap is not an issue with a bar chart, because the data are summarized. Data source in **[7]**.





BAR CHARTS IN SAS

The Environment

The code that supports the concepts described in this paper has been fully tested in Version 9.1.3 SAS software. However, the version 8 manuals referenced in this paper accurately document PROC GCHART and all supporting SAS/GRAPH statements. In addition, graphics output has been written to EMF files. To find out more about the EMF (**E**nhanced Windows **M**eta**f**ile) device driver, see *TS-DOC:TS-674* fully cited in the reference section **[10]**.

SAS/GRAPH Statements

SAS/GRAPH procedures are supported by *statements* that affect how a graph is rendered. These statements are external to the PROCs and the ones used in this paper are global in scope. See Table 1 below.

Table 1. SAS/GRAPH statement definitions from **[8], pp.161-162** that are used in this paper.

Statement	Description
AXIS	Modifies the appearance, position, and range of values of axes in charts and plots
GOPTIONS	Submits graphics options that control the appearance of graphics elements by specifying characteristics such as … fonts or text height. Graphics options can also temporarily change device settings.
LEGEND	Modifies the appearance and position of legends generated by procedures that produce charts, plots, and maps.
PATTERN	Controls the color and fill of patterns assigned to areas in charts, maps, and plots
TITLE	Add titles to graphics output.

SYMBOL is omitted from the list, because it cannot be used in PROC GCHART.

More about GOPTIONS

To get graphics output, a GOPTIONS statement must be included in the code. Like SAS system OPTIONs, scope is global **[4], p. 39**. For a more comprehensive discussion about GOPTIONS see references **[1], [3], [4], [8], [9],** and **[10]**. Here is the complete GOPTIONS statement for the chart in Example #1:

```
GOPTIONS device=emf gunit=pct rotate=landscape ftext="Arial"
 ftitle="Arial/Bold";
TITLE1 "Count of Meetings by Department";
FILENAME chrt1 "c:\N08\HOWGCHART\Grf\chrt1.emf";
GOPTIONS htext=5pct htitle=6pct gsfname=chrt1;
```

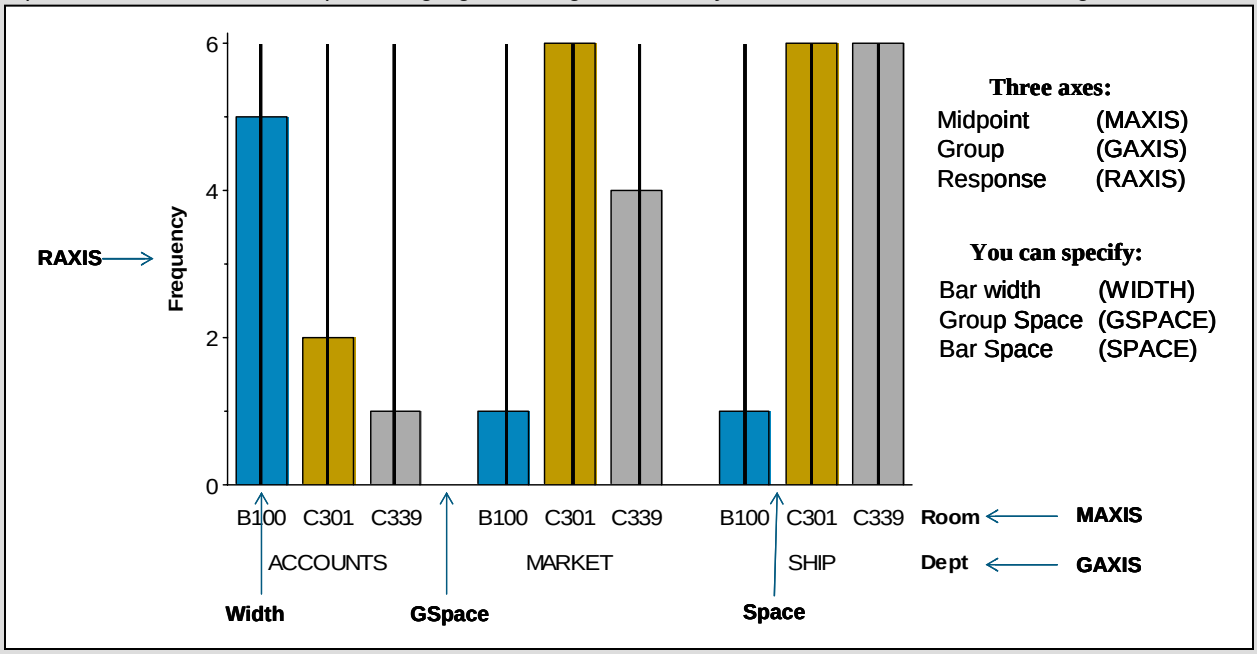
- **device=emf** for **E**nhanced Windows **M**eta**f**ile is a device driver that is supported in SAS. EMF uses a vector format that produces a high-resolution graphic that can be resized without loss in quality **[10], p.4**. Raster or bitmap graphs such as *Playfair.jpg* displayed in Figure 2 are pixel-based which means the picture will degrade when it is enlarged.
- **gunit=pct** specifies the default unit of measure used with height specifications. Choices include CELLS (default), CM (centimeters), IN (inches), PCT (percentage of graphics output area), and PT (points) **[8], p.350.**
- **rotate=landscape** means the graph's width is greater than its height.
- **ftext="Arial"** sets the font for all text in the graph. The argument is device dependent. Double quotes are required. **fitle="Arial/Bold"** emboldens title text**.**
- A second **GOPTIONS** statement includes the HTEXT, HTITLE and GSFNAME clauses so that they can be altered for each graph that is generated in a multi-graph program.
- **gsfname=chrt1** The argument points to the emf file referenced by FILENAME that can be inserted into PowerPoint or WORD. The same graph can be viewed in the Graphics Window of the enhanced editor by double-clicking on GCHART output listed in the RESULTS window.

Complete GOPTIONS statements do not appear elsewhere in the paper, but they can be found in the unabridged program in the NESUG proceedings. Code lines of lesser significance are also diminished in size in the examples that follow.

Bar Chart Structure: No X-coordinates

The horizontal axis for x-coordinates in the GPLOT procedure is replaced by the midpoint (MAXIS) axis in GCHART. If needed, a group axis (GAXIS) can be added. The only addressable coordinates in a SAS bar chart lie along the black vertical lines in Figure 4. Even the coordinates along the group axis are not accessible to the programmer. This arrangement limits access to the graphic structure, but it makes it possible to easily change the values for WIDTH, SPACE and GSPACE that control the appearance of the bars.

Figure 4. GCHART replaces the horizontal axis in GPLOT with a midpoint axis. Default locations for the joint midpoint and group axes labels, *Room* and *Dept*, are highlighted. Programmers only have access to coordinates along the black lines.



Procedure Syntax

Only the options that are used in the paper are included here. For a complete listing, see the SAS/GRAPH Version 8 and Version 9 reference manuals **[8], pp. 541-557 and [9], pp. 796-814.**

PROC GCHART<DATA=*input-data-set*>
VBAR | HBAR | VBAR3D *chart-variable(s)* </*option(s)*>
Where options following the forward slash (/) can include the following:

■ appearance options
CERROR=*error-bar-color*
COUTLINE=*bar-outline-color* | SAME
GSPACE=*group-spacing* (in cells)
LEGEND=LEGEND<1…99>
NOLEGEND
PATTERNID=GROUP | MIDPOINT | SUBGROUP
SHAPE=*3D-bar-shape* (HBAR3D and VBAR3D only)
SPACE=*bar-spacing* (in cells)
WIDTH=*bar-width* (in cells)
WOUTLINE=*bar-outline-width* (in pixels)

■ statistic options
CLM=*confidence-level*
ERRORBAR=BARS | BOTH | TOP
FREQ
FREQLABEL='*column=label*' (HBAR only)
INSIDE=*statistic* (VBAR only)
OUTSIDE=*statistic* (VBAR only)
SUMLABEL='*column=label*' (HBAR only)
SUMVAR=*summary-variable*
TYPE=*statistic*

■ axes options
DESCENDING
GAXIS=AXIS<1…99>
MAXIS=AXIS<1…99>
RAXIS=AXIS<1…99>

■ midpoint options
DISCRETE
GROUP=*group-variable*
MIDPOINTS=*value-list*
RANGE
SUBGROUP=*subgroup-variable*

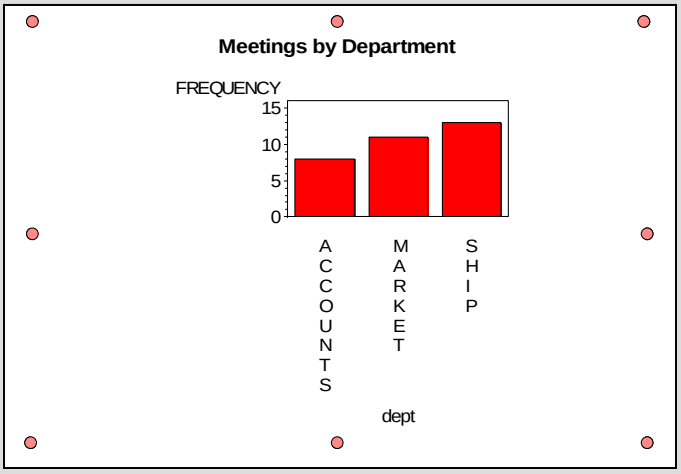
MIDPOINT BAR CHARTS

The basic bar chart is a midpoint chart where the group axis is not defined. As in all types of charts, the variable assigned to VBAR (vertical bar) or HBAR (horizontal bar) always references a midpoint axis value. The term *midpoint* is apt, since the value is center-justified under the bar for VBAR and half-way up a bar for HBAR. In Figure 4 the vertical black lines coincide with the midpoint values and mark the only areas in the chart that can be referenced by the GCHART procedure. This positioning is ideal for charting nominal or ordinal discrete data.

Example #1: Taking the Defaults

Example #1: A midpoint chart where default settings are used for the AXES, bar WIDTHS and PATTERNS. The frame around the data rectangle qualifies as *chartjunk*, a term coined by Tufte to reference anything on a graph "that does not tell the viewer anything new" [5, p. 107]. SAS frames by default, and without a PATTERN statement all bars are colored red.

```
%let outpath = c:\N08\HOWGChart\GRF;
title1 "Meetings by Department";
filename chrt "&outpath\chrt1.emf";
goptions htext=5 htitle=6 gsfname=chrt;

proc gchart data=N08.meetings;
  vbar dept;
run;
quit;
```



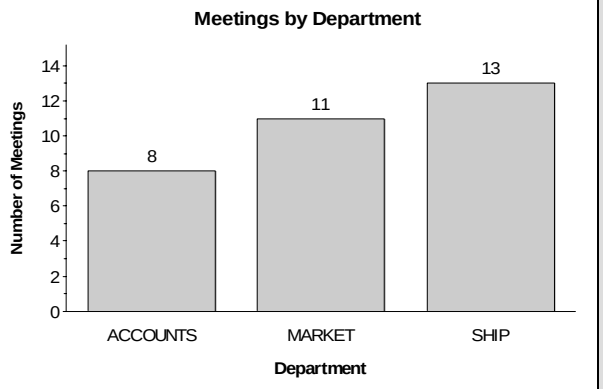
- **%let outpath =** Assigning a path to a macro variable makes it easy to reassign subdirectories.
- **goptions htext=5** text height is enlarged to show that it does not affect bar width.
- **proc gchart vbar dept** DEPT is the midpoint variable because it comes right after the VBAR keyword.

Unfortunately bar widths do not automatically accommodate text size in PROC GCHART. Therefore the result is an unappealing graph with a lot of white space surrounding a very small data rectangle. The selection handles are highlighted in POWERPOINT to demonstrate just how much space is wasted because of the vertical midpoint axis values and the horizontal response axis label. Possibly the excessive amount of white space could qualify as *chartjunk*, since it obstructs what is being communicated in the graph.

Example #2: A Better Midpoint Chart

Example #2: Fix the default bar chart by adding SAS/GRAPH statements and GCHART options.

```
title1 move=(+7,+0)  "Meeting by Department";
filename chrt "&outpath\chrt2.emf";
goptions htext=5pct htitle=6pct gsfname=chrt;

pattern1 color=grayCC;

axis1
 label=(a=90 f="Arial/Bold"
 "Number of Meetings")
 order=(0 to 14 by 2) minor=(n=1);

axis2 label=(f="Arial/Bold" "Department");

proc gchart data=N08.meetings;
 vbar dept / noframe
  width=50 outside=freq
  raxis=axis1 maxis=axis2
  coutline=black woutline=1;
run; quit;
```



- **title1 move=(+7, +0)** While the relative move issues a WARNING to the log, it centers the text directly over the data rectangle.
- **pattern1** The pattern statement sets the color and type of fill for all areas in a graph. The default type for a pattern fill is SOLID. Other options such as EMPTY and slanted hatch marks are also available. Because PATTERNID is not specified as a GCHART option in the example, all the bars are colored the same.
- **axis1** The label for the response axis is rotated (a=90), and emboldened (f="Arial/bold"), provided with 1 minor tick mark that references an integer (n=1), and given a more expansive range (0 to 14 by 2). For the label, formatting clauses (a= f=) ***must*** precede text assignments or they won't take effect.
- **vbar dept/ noframe** To remove the *chartjunky* frame from the chart use NOFRAME.

- **vbar dept/ width=50** WIDTH is always defined in cells which vary in size by graphics device. Setting WIDTH to an arbitrarily high number generates an acceptable bar chart that expands the data rectangle.
- **vbar dept/ outside=freq** displays the specified statistic (in this case FREQ) above the bar. While numbers pasted to the interior of a chart can add to graph clutter, their presence can be helpful in pinpointing values that are furthest from the origin.
- **vbar dept/ raxis=axis1 maxis=axis2** How the axes statements are linked to RAXIS and MAXIS in GCHART.
- **vbar dept/ coutline=black woutline=1** the bars are outlined in black. Although defined as *chartjunk* by Tufte, the outlines don't take up extra space and are especially helpful when bar background colors are very light.

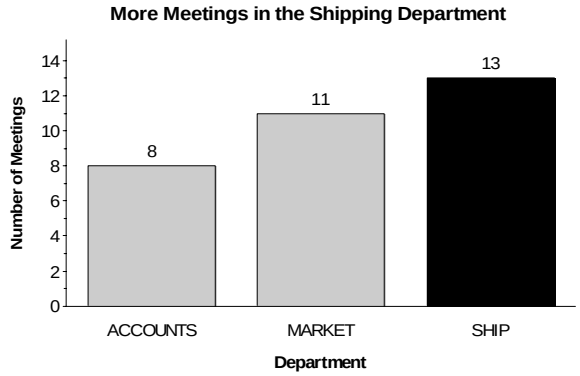
With the OUTSIDE= COUTLINE= and WOUTLINE= (in pixels) defined in this example, it is now possible to show how Tufte defines the six different ways that altitude is displayed in a bar chart:

> (1) height of the left line, (2) height of shading, (3) height of right line, (4) position of top horizontal line, (5) position (not content) of number at bar's top, and (6) the number itself. **[5. p.96].**

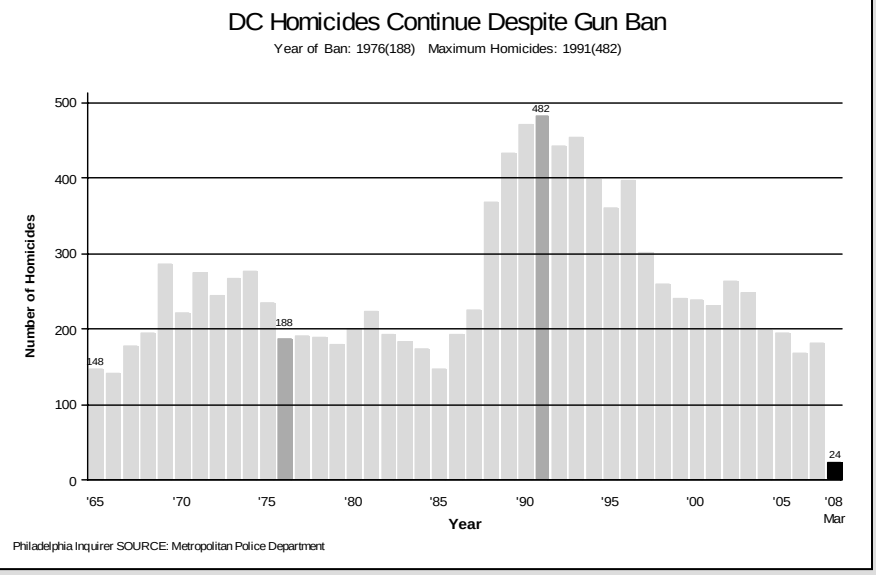
Example #3: Assigning Colors to a Midpoint Chart

Example #3: Take advantage of the REPEAT=option to assign the same color to multiple, unrelated bars.

```
title1 move=(+7,+0)
 "More Meetings in the Shipping Department";
filename chrt "&outpath\chrt3.emf";
goptions htext=5pct htitle=6pct gsfname=chrt;

pattern1 color=grayCC repeat=2;
pattern2 color=black  repeat=1;

axis1
 label=(a=90 f="Arial/Bold"
 "Number of Meetings")
 order=(0 to 14 by 2) minor=(n=1);
axis2 label=(f="Arial/Bold" "Department");

proc gchart data=N08.meetings;
 vbar dept / noframe patternID=midpoint
  width=50 outside=freq
  raxis=axis1 maxis=axis2
  coutline=black woutline=1;
run; quit;
```



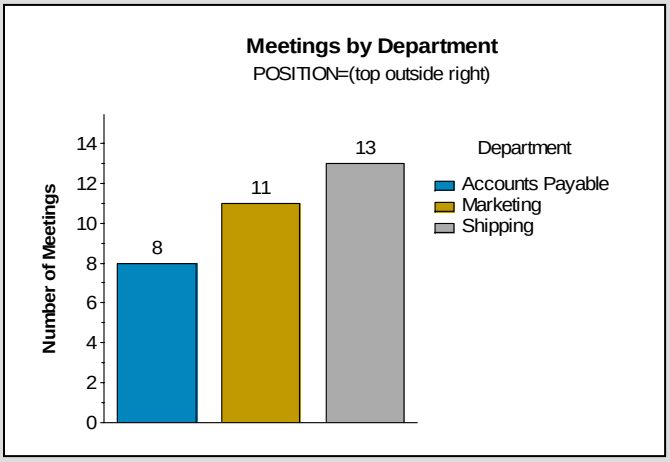

```
pattern1 c=CXdadada repeat=11;
pattern2 c=CXababab repeat=1;
pattern3 c=CXdadada repeat=14;
pattern4 c=CXababab repeat=1;
pattern5 c=CXdadada repeat=16;
pattern6 c=black repeat=1;
```

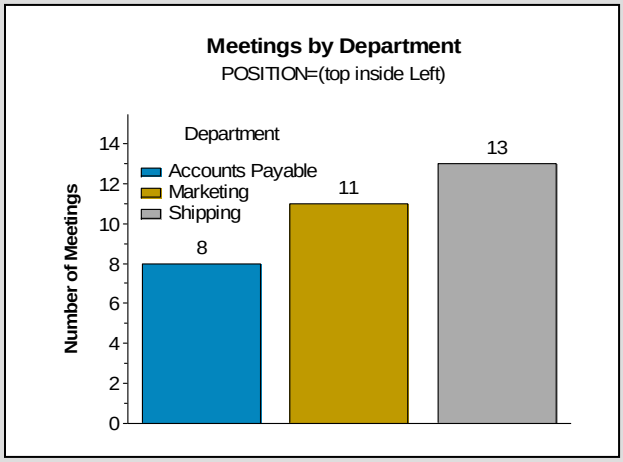


Color can be used to emphasize the high number of meetings held in the shipping department. Additions to the code include PATTERN statements and the PATTERNID= option. The REPEAT=option in the PATTERN statement was especially helpful when the number of homicides over a 44 month period was translated into SAS from a graph that appeared in *The Philadelphia Inquirer* on March 14, 2008.

Example #4: Assigning a Legend to a Midpoint Chart

Example #4: *n* patterns are assigned to an *n*-midpoint bar chart to form a legend that replaces axis annotation. Use the SUBGROUP=option and set the legend POSITION to INSIDE to maximize the size of the data rectangle.






```
proc format;
value $dptLfm
  'ACCOUNTS'='Accounts Payable' 'MARKET'='Marketing' 'SHIP'='Shipping';
run;
filename chrt "&outpath\chrt4B.emf";
goptions htext=5 htitle=6 gsfname=chrt;

pattern1 color=CX0386BE; pattern2 color=CXBF9900; pattern3 color=CXABABAB;

legend1
  across=1 shape=bar(3,2) label=( position=(top center)"Department")
  position=(top inside left) mode=share;

axis1 label=(a=90 f="Arial/Bold" "Number of Meetings") order=(0 to 14 by 2)
  minor=(n=1);
axis2 label=none value=none;

title1 move=(+7,+0) 'Meetings by Department';
title2 move=(+7,+0pct) 'POSITION=(top inside Left)';

proc gchart data=N08.meetings;
  vbar dept /noframe width=15 type=FREQ
    outside=FREQ subgroup=dept
    legend=legend1 raxis=axis1 maxis=axis2
    coutline=black woutline=1;
  format dept $dptLfm.;
run; quit;
```

- **proc format...$dptLfm** While the VALUE= option in LEGEND statement could be hard-coded to modify DEPT descriptions, a format forges a more reliable link between data and display. The format is applied inside PROC GCHART.
- **pattern1... pattern2... pattern3...** Three pattern statements for three bars.
- **legend1** The legend statement contains two POSITIONS. The first is a sub-option in the LABEL=option and the second is a full-fledged option for legend placement inside the data rectangle. MODE=share prevents graphics elements from being overwritten. LEGEND1 is linked to PROC GCHART via the LEGEND= option.
- **axis2** label=none value=none shows how to turn off all annotation along the midpoint axis. Remember minor=none and major=none are NOT required, since ticks are not represented along the midpoint axis in GCHART.
- **vbar dept/ width=15** A specific width has been defined so that the EMF file can be enlarged in WORD .
- **vbar dept/ subgroup=dept** Using this option generates a midpoint chart with a legend. The syntax is counter-intuitive, because there is no subgroup in the chart, and DEPT is a midpoint not a subgroup variable. Also missing from the code is PATTERNID=MIDPOINT that would assign colors by DEPT. SAS will not insert a legend unless PATTERNID=SUBGROUP which automatically happens when SUBGROUP=DEPT **[14, p. 551]**.

While setting the legend POSITION= option to INSIDE increases the size of the data rectangle, Cleveland warns that clutter will increase unless the key or legend is fully separate from the data **[2, pp.36-37, 46-47].** However, his warning comes with reservations:

> One disadvantage, compared with data labels inside the scale-line rectangle, is that identification is slightly harder because we must look back and forth between the key and the data **[2, p.46]**.

Ways to cope with INSIDE legend placement include:

1)Place the legend over low-frequency bars.

2)Always set MODE=SHARE so that the data are never hidden.

3)Increase the interval in the response axis range for added space. This maneuver will generate a justifiable increase in the maximum. For example, if <0 to 14 by 2> is changed to <0 to 16 by 4>, then 13 still falls within the last interval. Merely increasing the maximum and keeping the interval fixed at 2 will violate Cleveland's maxim to choose scales "so that the data rectangle fills up as much of the scale-line rectangle as possible"**[2, p.81]**. With an interval of 2 and a maximum of 16, the data and scale rectangles are no longer in synchrony, because there are no data points between 14 and 16.

4)Reduce the size of the text in the legend. (See Example #5).

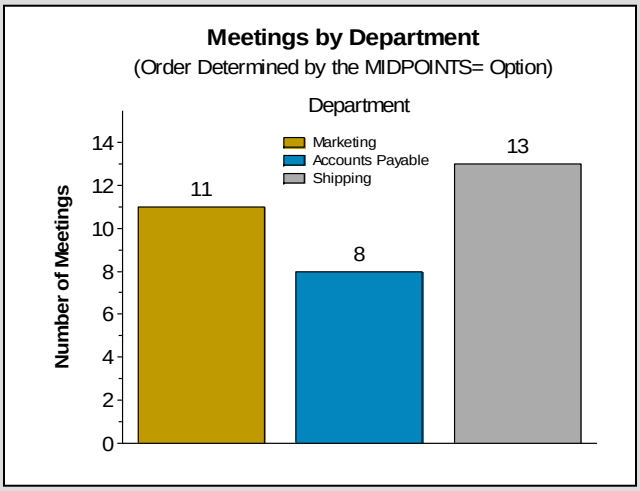
5)Use a relative move with OFFSET= to get the legend out of the way. (See Example #5).

6)Change the direction of the legend from vertical to horizontal with ACROSS= (See example #8).

Example #5: Changing the Bar Order in a Midpoint Chart

Example #5. Bar order is changed with the MIDPOINTS= option in PROC GCHART. The VALUE option in LEGEND1 is used instead of a format that no longer works. The legend with a reduced text size is moved upwards a little from the center of the graph with the OFFSET option. ORDER= and VALUE= in the legend are set to reflect the new bar order.

```
filename chrt "&outpath\chrt5.emf";
goptions htext=5 htitle=6 gsfname=chrt;
pattern1 color=CX0386BE; pattern2 color=CXBF9900;
pattern3 color=CXababab;
legend1
 across=1 shape=bar(3,2)
 label=(position=(top center) "Department")
 position=(top inside center) mode=share
 offset=(+0,+4) order=("MARKET" "ACCOUNTS" "SHIP")
 value=(h=4 'Marketing' 'Accounts Payable'
            'Shipping');
axis1
 label=(a=90 f="Arial/Bold" "Number of Meetings")
 order=(0 to 14 by 2) minor=(n=1);
axis2
 label=none value=none;
title1 move=(+7,+0) 'Meetings by Department';
title2 move=(+7,+0)
 '(Order Determined by the MIDPOINTS= Option)';
proc gchart data=N08.meetings;
  vbar dept /noframe width=15 type=FREQ
    outside=FREQ subgroup=dept
    midpoints="MARKET" "ACCOUNTS" "SHIP"
    legend=legend1 raxis=axis1 maxis=axis2
    coutline=black woutline=1;
run; quit;
```



Bar order can changed from the default alpha-numeric order by adding a MIDPOINTS= option to PROC GCHART. The ORDER and VALUE options in the legend reflect the new order. Additional GCHART options for changing bar order include ASCENDING= and DESCENDING= where bars are ordered by their heights.

SUBGROUP OR STACKED BAR CHARTS

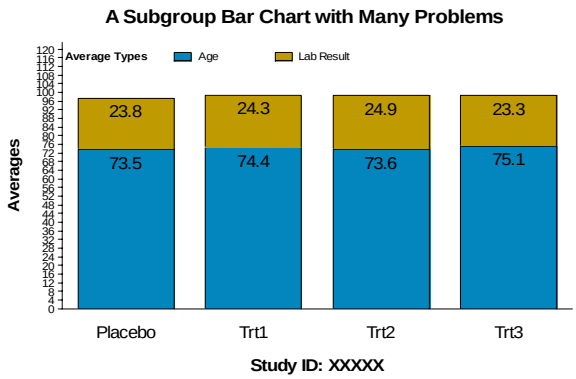
A subgroup bar chart has the same axis configuration as the midpoint bar chart; namely only two axes are involved: the midpoint axis (MAXIS) and the response axis (RAXIS). While the midpoint axis was featured in the previous section, attention is now being paid the response axis where summary statistics are displayed. Like PROC REPORT and PROC TABULATE; summary statistics are calculated *within* PROC GCHART, but sometimes their display gets muddled when bars are stacked on top of each other. An example of how bad things can get is shown in the first two panels of Example #6.

Example #6: Displaying Means in a Subgroup Bar Chart

Example#6: Problems with the stacked bar chart in panel 2 are listed in the first panel. Solutions applied to the subgroup bar chart of the MEETINGS data set that appears in the last panel require the use of pre-processed data.

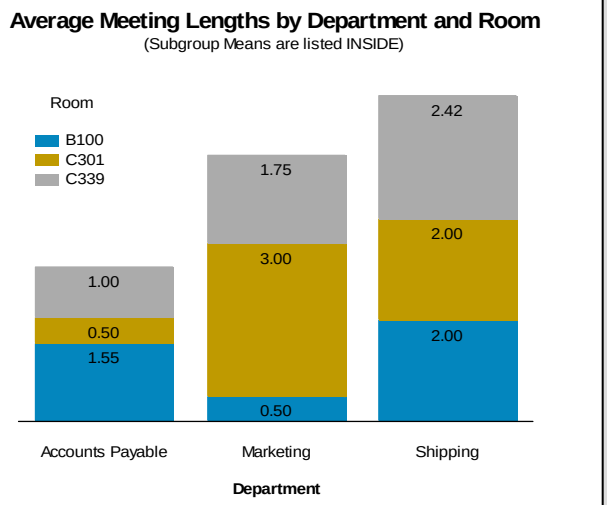
Problems

1) The graph is **cluttered** because of the large number of major tick marks along the response axis. Cleveland recommends anywhere from 3 to 10 **[2, p.39].**
2) Since longest bar is less than 100, an axis maximum of 120 is **misleading**. In other words, data and scale rectangles don't coincide. However, there would be no place to put an INSIDE legend with an axis range of <0 to 100 by 20>.
3) The display of the response axis counts as **chartjunk**, since all subgroups in the graph are explicitly labeled. It can be removed without any loss in information.
4) **Bar outlines** for subgroups should also be **removed**, since they don't add information and they don't translate properly from SAS to Microsoft® WORD or PowerPoint. (See TRT3).
5) The response axis is **misleading**, since the highest average is 75.1, not 98.5. It makes no sense to sum averages in a subgroup bar chart. Again, removing the axis would solve the problem.
6) Actually, this chart should be **replaced** with **two midpoint charts**; one for lab results and another for age. Concluding that the average lab result is approximately one third the average age for every arm in the study is meaningless.



Problem Solving Requires Pre-Processed Data

```
filename chrt "&outpath\chrt6.emf";
goptions htext=4 htitle=5 gsfname=chrt;
pattern1 ...; pattern2...; pattern3 ...;

legend1
  across=1 shape=bar(3,2)
  label=( position=(top center) "Room")
  position=(top inside left) mode=share;

axis1 label=none minor=none major=none
 value=none style=0 order=(0 to 6.5 by 0.5);
axis2 label=(f="Arial/Bold" "Department");

title1 'Average Meeting Lengths by Department and Room';
title2 '(Subgroup Means are listed INSIDE)';

proc gchart data=work.Avg;
  vbar dept /noframe
    width=15 type=SUM sumVar=AvgHours
    subgroup=room inside=SUM
    legend=legend1  raxis=axis1 maxis=axis2;
  format dept $dptLfm.;
run; quit;
```



- **goptions htext=4 htitle=5** HTEXT is reduced to 4 percent so that the averages appear on the chart. When the text is too large, it simply disappears without a warning being issued to the LOG. HTITLE is also reduced to accommodate a longer title that incorporates text usually reserved for the response axis label.
- **axis1 label=none ... style=0** The response axis is removed from the display. STYLE=0 removes the axis line as well.
- **proc gchart data=work.Avg** Using the INSIDE= option to insert means from raw data into a subgroup chart generates:

  ```
  WARNING: The statistic MEAN is not supported inside subgrouped bars on the VBAR statement.
  ```

  Therefore a summary data set is created as a work-around with the following SQL command:

  ```
  proc sql noprint;
    create table AVG as
    select distinct dept, room, mean(hours) as AvgHours
    from n08.meetings
    group by dept, room;
  quit;
  ```

  The output data set has nine records, one with an average for each combination of ROOM and DEPT.

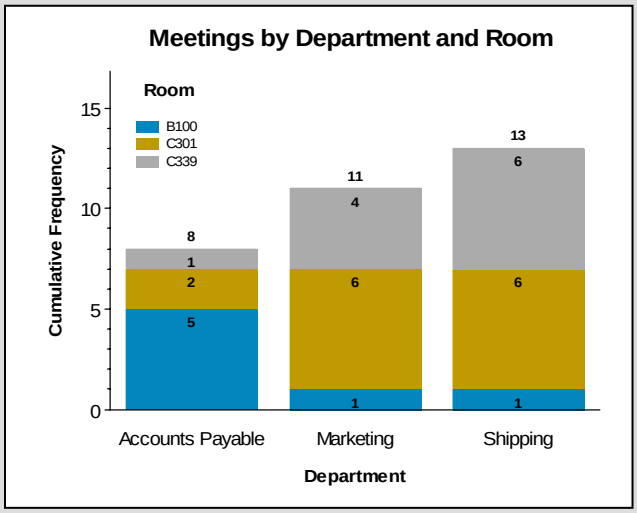
- **vbar dept / type=SUM sumVar=AvgHours ...inside=SUM** Since the sum of one observation within a set of classification variables is itself, type=SUM and inside=SUM works as intended. SUMVAR=, a new option, is used for identifying the variable that supplies the statistic.

The subgroup bar chart is best-suited for displaying frequencies, since no data pre-processing is required. Also the response axis can be meaningfully labeled as "Cumulative Frequency".

Example #7: Displaying Frequencies in a Subgroup Bar Chart

Example #7. Frequencies work well in a subgroup bar chart where the response axis is appropriately labeled *Cumulative Frequency.*

```
filename chrt "&outpath\chrt7.emf";
goptions htext=3.75 ftext="Arial/Bold"
  htitle=6 gsfname=chrt;
pattern1 ...; pattern2 ...; pattern3 ...;
legend1
  across=1 shape=bar(3,2)
  label=(h=4.5 position=(top center) "Room")
  value=(f="Arial") position=(top inside left)
  mode=share;
axis1 minor=(n=4)
  label=(a=90 h=4.5 "Cumulative Frequency")
  order=(0 to 15 by 5) value=(f="Arial" h=4.5);
axis2 label=(h=4.5 "Department")
  value=(f="Arial" h=4.5);
title1 move=(+9pct,+0pct) 'Meetings by Department and Room';
proc gchart data=n08.meetings;
  vbar dept /noframe
    width=15
    subgroup=room outside=FREQ inside=FREQ
    legend=legend1 raxis=axis1 maxis=axis2;
  format dept $dptLfm.;
run;quit;
```



- **goptions htext=3.75 ftext="Arial/Bold"** defines a smaller, bolder text for the OUTSIDE= INSIDE= options. If too large in either the vertical or horizontal direction, the numbers won't print. This means larger text sizes and regular fonts for labels and values in the LEGEND and AXES statements must be explicitly defined. They are highlighted in the source code above.
- **vbar dept/ subgroup=room** automatically generates a legend unless NOLEGEND is specified.
- **vbar dept/ outside=FREQ inside=FREQ** Three INSIDE frequencies add up to a single OUTSIDE frequency.

GROUP BAR CHARTS

When groups are accommodated in a bar chart the number of bars typically increases from *m* (#*m*idpoints) to *mXg* (*mX* #groups). Bar widths decrease and midpoint axis text becomes even more constricted. Unless a legend is used, a vertical arrangement of midpoint axis values can hardly be prevented.

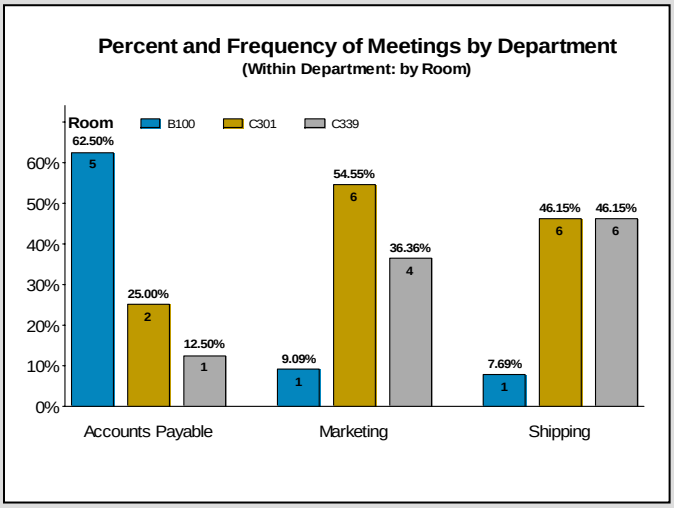
The SUBGROUP= option required for inserting legends into midpoint charts also works for group bar charts (see Example #4). To prevent text from going vertical, set SUBGROUP= to the midpoint variable *and* hide both the midpoint axis *values* and *label*. The group label should also be hidden, because SAS automatically positions all MAXIS and GAXIS labels to the right of the data rectangle. If an axis label is long, white space is increased with a corresponding reduction in size of the data rectangle. Centering the group axis label below the axis values saves space, but requires the use of ANNOTATE which is beyond the scope of this paper.

If possible, use a group rather than a subgroup bar chart to display statistics. All translate easily and accurately in PROC GCHART to a group chart. Available statistics include: FREQ, CFREQ, PERCENT (PCT), CPERCENT (CPCT), SUM, and MEAN **[14, p. 556].** If the selected statistic TYPE is a SUM or MEAN, then the SUMVAR= option for *summary variable* must also be used **[14, p. 579]**.

Example #8: Assigning a Legend and Displaying Percents in a Group Bar Chart

Example #8. A group bar chart with a legend. Group percents are easily calculated with the G100 (group 100) option.

```
filename chrt "&outpath\chrt8.emf";
goptions htext=3.75pct ftext="Arial/Bold" ...;
pattern1 ...; pattern2 ...; pattern3 ...;
legend1
  across=3 shape=bar(3,2)
  label=(h=4.5 position=(left) "Room")
  value=(f="Arial") position=(top inside left)
  mode=share;
axis1 label=none order=(0 to 70 by 10)
  value=(f="Arial" h=4.5
  '0%' '10%' '20%''30%' '40%' '50%' '60%' ' ')
  minor=NONE major=(h=0.25) offset=(,4pct);
axis2 label=none value=none;
axis3 label=none value=(f="Arial" h=4.5);
title1 ...; title2 ...;
proc gchart data=N08.meetings;
 vbar room / type=PCT g100 inside=freq noframe
  outside=pct group=dept subgroup=room
  width=10 gspace=5 space=3 legend=legend1
  raxis=axis1 maxis=axis2 gaxis=axis3
  coutline=black woutline=1;
  format dept $dptLfm.;
run; quit;
```

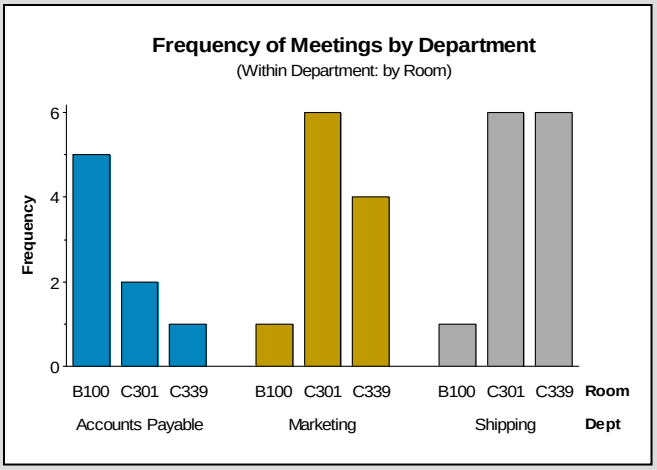


- **legend1 across=3** To display a horizontal legend, increase the number assigned to ACROSS from one to three to accommodate each meeting room in the data.
- **axis1 label=none ... value=(... '0%' ... '60%' ' ') minor=NONE major=(h=0.25) offset=(,4pct);** Since the response axis is references a *statistic* rather than a *variable*, a format cannot be used to affix a percent sign on to the end of each value. Instead use tick labeling in the VALUE= option of the AXIS statement. All highlighted options in this AXIS statement serve to reduce *chartjunk* by eliminating the need for an axis label. The value '70%' and minor tick marks are also not needed. In addition, the height for major tick marks is also reduced to disguise the fact that blank 70% is parallel to "Room" on the chart. With this "trick", OFFSET in the vertical direction can be reduced from 7 to 4 percent.
- **axis2** references the midpoint axis. The axis must be declared in order to be hidden. Defaults generate values!
- **axis3** references the group axis. Since "Department" appears in the title, the axis label now related to *chartjunk* can be removed.
- **vbar room/ group=DEPT subgroup=room ... legend=legend1** Using these three options generates a group bar chart with a legend that references midpoint values. Again, SAS will not insert a legend unless PATTERNID=SUBGROUP which happens when SUBGROUP=DEPT **[8], p. 551.**

Example #9: Hiding the Legend and Displaying Frequencies in a Group Bar Chart

Example #9. .Example #8 is simplified. No special processing is required for labeling vertical axis values or positioning a legend. However, the PATTERNID=option must be set to GROUP or MIDPOINT, so that colors are correctly assigned.

```
filename chrt "&outpath\chrt9.emf";
goptions htext=4.5 ftext="Arial" ftitle="Arial/Bold"
  htitle=6 gsfname=chrt;
pattern1 ...; pattern2 ...; pattern3 ...;
axis1 label=(f="Arial/Bold" a=90 "Frequency")
  order=(0 to 6 by 2) minor=(n=1);
axis2 label=(f="Arial/Bold" "Room");
axis3 label=(f="Arial/Bold" "Dept");

title1 move=(+7,+0) 'Frequency of Meetings by
Department';
proc gchart data=N08.meetings;
  vbar room / noframe
    Group=dept patternID=Group
    width=10 gspace=5 space=3
    raxis=axis1 maxis=axis2 gaxis=axis3
    coutline=black woutline=1;
    format dept $dptLfm.;
run; quit;
```



- **Axes Labels "frequency" "Room" "Dept"** Diminish the size of the data rectangle even though A(ngle) is set to 90 for "Frequency" and Department is abbreviated as "Dept".
- A legend is not produced, because the SUBGROUP= option is not filled in.
- **vbar room/ patternID=Group** Causes bar assignments to be made by DEPT affiliation. To assign color by ROOM, PATTERNID would have to be set to MIDPOINT.

ADDITIONAL BAR CHARTS (EXTRA CREDIT)

Non-Hierarchical Group Bar Charts

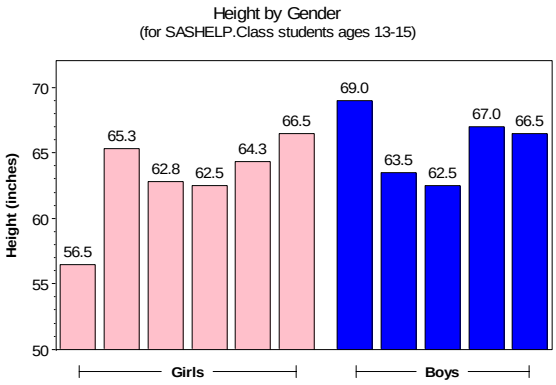
Data structures are not hierarchical when group affiliations are being charted for individuals. For example, William from SASHELP.Class can't be both male and female!

Obs	Name	Sex	Age	Height
1	Alfred	M	14	69.0
2	Alice	F	13	56.5
3	Barbara	F	13	65.3
4	Carol	F	14	62.8
5	Henry	M	14	63.5
6	Janet	F	15	62.5
7	Jeffrey	M	13	62.5
8	Judy	F	14	64.3
9	Mary	F	15	66.5
10	Ronald	M	15	67.0
11	William	M	15	66.5

The NOZERO option is used to generate the non-hierarchical bar chart in Example #10.

Example #10. The SASHELP.class data set is used to plot heights for students ranging in age from 13 to 15 years.

```
filename chrt "&outpath\chrt9.emf";
goptions htext=4.5 ftext="Arial" ftitle="Arial/Bold"
  htitle=6 gsfname=chrt;
pattern1 ...; pattern2 ...; pattern3 ...;
pattern1 color=pink; pattern2 color=blue;
axis1
  label=(a=90 f="Arial/Bold" "Height (inches)")
  order=(50 to 70 by 5) minor=(n=4);
axis2 label=NONE value=NONE;
axis3 label=none
 value=(f="Arial/Bold" "Girls" "Boys");
title1 'Height by Gender';
title2 '(for SASHELP.Class students ages 13-15)';
proc gchart data=N08.sasClass;
 vbar Name / width=5 type=SUM sumvar=Height
   outside=sum Group=sex patternID=group NOZERO
   raxis=axis1 maxis=axis2 gaxis=axis3
   coutline=black woutline=1;
run; quit;
```



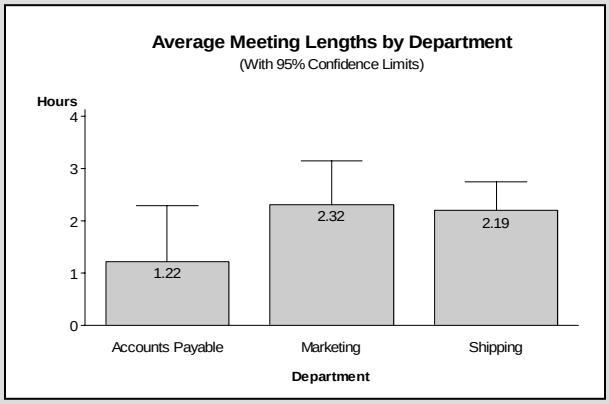
- **pattern1 pattern2** reference 'F' and 'M' in alphabetical order from variable, SEX. While group order can be changed for *hierarchical* group bar charts re-ordering is not allowed with NOZERO. So girls must come first.
- **axis1 order=(50 to 70 by 5)** The data/ink ratio is increased when the axis minimum is set to 50. With minor differences emphasized, outliers can be identified as the first children in both groups: a short girl and a tall boy.
- **axis2** The midpoint axis hidden, because there is no need for knowing *who* is being measured.
- **axis3** Group axis *values* are being emboldened to match the *label* for the response axis.
- **vBar Name/** declares (hidden) NAME as the midpoint variable.
- **vBar Name/ width=5** easily accommodates OUTSIDE text at 4.5PCT. When WIDTH is reduced to 2 cells, the OUTSIDE values disappear.
- **vBar Name/ type=SUM sumvar=Height** Since the data are in summary format (one record per midpoint value), type becomes SUM. If FREQ were charted, all the bars would have a height of 1.
- **vBar Name/ group=sex patternID=group** The group variable is defined and its values are linked to the pattern statement via PATTERNID. Since PATTERNID is not set to SUBGROUP, a legend is not produced.
- **vBar Name/ NOZERO** suppresses the display of a bar when its frequency is zero (i.e. nonexistent). Thus the 'William' bar is excluded from the group labeled GIRLS. If NOZERO were omitted, space for 22 bars would have to be allocated in the chart.

Generating Error Bars

Error bars can be easily added to a bar chart when SUMVAR is set to MEAN or PCT. Algorithms for calculating error bar heights from raw data are described in the SAS/GRAPH manual **[8], p.547.** In Example #11, error bars are displayed for average meeting times per department.

Example #11. Generating Error Bars in PROC GCHART.

```
filename chrt "&outpath\chrt11.emf";
goptions htext=4.5 ftext="Arial" ftitle="Arial/Bold" hti-
tle=6 gsfname=chrt;

pattern1 color=grayCC;
axis1 label=(f="Arial/Bold" "Hours") minor=NONE;
axis2 label=(f="Arial/Bold" "Department");
title1 move=(+10,+0) 'Average Meeting Lengths by Depart-
ment';
title2 move=(+10,+0) '(With 95% Confidence Limits)';
proc gchart data=N08.meetings;
  vbar dept /noframe
    width=50 type=MEAN inside=MEAN
    errorbar=TOP clm=95 sumvar=Hours
    raxis=axis1 maxis=axis2
    coutline=black woutline=1;
run; quit;
```



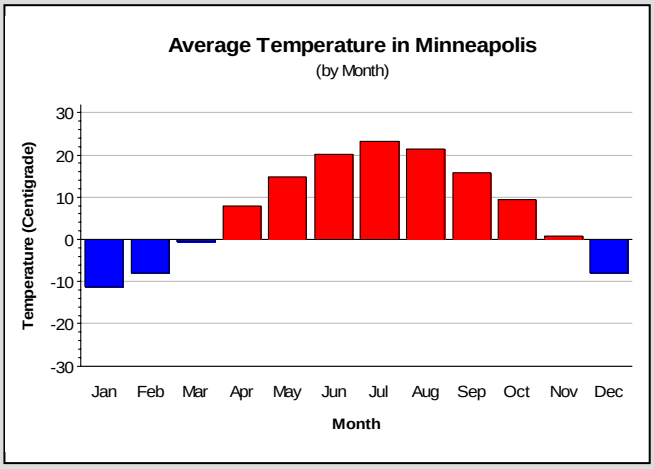
- **vBar Name/ type=MEAN inside=MEAN sumvar=Hours** Error bars are being generated for the means of a continuous variable (HOURS). TOP error bars and INSIDE statistics can be displayed together.
- **vBar Name/ errorbar=TOP clm=95** Besides TOP, BOTH (tops and bottoms) and BARS shown on page 595 in **[8]** are available for selection. In this example, error bars with 95% confidence limits are being plotted.

Generating a Min/Max Chart by Manipulating the SUBGROUP= Option

The method for assigning whole bars single colors in a subgroup chart comes from an unpublished CD entitled *Robert Allison's SAS/Graph Examples!* In the example from the CD, SUBGROUP is set to a mutually exclusive binary variable. Similarly in Example #12, ABOVEZEROYVN is identified as the subgroup variable, because an average monthly temperature cannot be both above and below zero.

Example #12. Generating a Min/Max chart in PROC GCHART.

```
filename chrt "&outpath\chrt12.emf";
goptions htext=4.5 ftext="Arial" ftitle="Arial/Bold" htitle=6
gsfname=chrt;

pattern1 color=Blue; pattern2 color=Red; pattern3;
axis1 label=(a=90 f="Arial/Bold" 'Temperature (Centigrade)')
  minor=(n=4) order=(-30 to 30 by 10);
axis2 label=(f="Arial/Bold" 'Month');
title1 move=(+10,+0) 'Average Temperature in Minneapolis';
title2 move=(+10,+0) '(by Month)';

proc gchart data=n08.MplsWeather;
 vbar month / discrete noframe
   width=5 type=sum sumvar=TC
   subgroup=AboveZeroYvN nolegend
   raxis=axis1 maxis=axis2
   autoref clipref cref=graybb
   coutline=black woutline=1;
 format month monthfmt.;
run; quit;
```



- **vbar month / discrete** The unformatted values for month are numbers ranging from 1 to 12. Up to now all midpoint values have been character. For numbers, GCHART internally calculates midpoints from the data range. What DISCRETE does is to instruct GCHART to omit the calculation and use the raw values instead.
- **vbar month / subgroup=AboveZeroYvN nolegend** The variable AboveZeroYvN contains an 'N' when the temperature is below zero and a 'Y' when it is above zero. Thus PATTERN1 is set to blue (cold) and PATTERN2 is red (hot). Since a temperature cannot be both above and below zero, all the bars are assigned single colors. NOLEGEND also has to be specified, since a variable has been assigned to SUBGROUP.
- **vbar month / autoref clipref cref=graybb** Reference lines are used to make it easier to determine the temperature at a given month. CLIPREF places the lines behind the bars, CREF colors them light gray so that they aren't too prominent, and NOFRAME gets rid of the default frame around the plotting region.

SUMMARY AND CONCLUSIONS

Step-by-step instructions for building bar charts that conform at least in spirit to principles defined by statistical graphics experts have been presented in this tutorial. Frames, redundant labels and overly busy axes have been removed from charts, and an effort has been made to reduce the amount of white space caused by text placement in SAS software. Since the recommended OUTSIDE legend also increases white space in charts, it has been moved INSIDE. However, a conscious effort has been made to keep legend and chart data completely separate.

Besides learning what impact GCHART procedure options have on graphics output, an effort has been made to show the GOPTIONS, PATTERN, AXIS, FORMAT and LEGEND statements are incorporated into chart building.

COPYRIGHT STATEMENT

REFERENCES

[1] Cassidy, Deb. *An Introduction to SAS/Graph®*. Proceedings of the SAS Global Forum, Cary, NC: SAS Institute Inc., 2007, paper #112.

[2] Cleveland, William S. *The Elements of Graphing Data: Revised Edition.* Summit, NJ: Hobart Press, 1994.

[3] Cochran, Ben. *A Gentle Introduction to SAS/GRAPH® Software*. Proceedings of the Twenty-Eighth SAS User Group International Conference, Cary, NC: SAS Institute Inc., 2003, paper #200.

[4] Miron, Thomas. *The How-To Book for SAS/GRAPH Software.* Cary, NC: SAS Institute Inc., 1995. Copyright 1995, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC

[5] Tufte, Edward R. *The Visual Display of Quantitative Information: Second Edition*. Cheshire, CT: Graphics Press, 2001.

Web Citations:

[6] http://en.wikipedia.org/wiki/Bar_chart. *Bar chart: From Wikipedia, the free encyclopedia*. A definition and history of the bar chart is provided.

[7] http://lib.stat.cmu.edu/datasets/baseball.data. *From StatLib --- DataSets Archive.* This was the 1988 ASA Graphics Section Poster Session dataset, organized by Lorraine Denby.

SAS Institute References:

[8] SAS Institute Inc. *SAS/GRAPH® Reference, Version 8,* Cary NC: SAS Institute Inc., 1999.

[9] SAS Institute Inc. *SAS/GRAPH® Reference, Volumes 1, 2, and 3,* Cary NC: SAS Institute Inc., 2004.

[10] TS-DOC:TS-674. *An Introduction to Exporting SAS/GRAPH Output to Microsoft Office SAS Release 8.2 and higher.* http://support.sas.com/techsup/technote/ts674/ts674.html.

RELATED PAPERS BY THE AUTHOR:

Watts, Perry. *Building a Better Bar Chart with SAS/Graph® Software*. Proceedings of the 20th Annual Northeast SAS Users Group Conference. Baltimore, MD, 2007, paper #NP16.

Watts, Perry. *Charting the Basics with PROC GCHART*. Proceedings of the 20th Annual Northeast SAS Users Group Conference. Baltimore, MD, 2007, paper #FF17.

WHAT'S IN THE NESUG PROCEEDINGS:

1) Data Sets
   - **bb_bchart.sas7bdat** baseball data in Figure 3.
   - **meetings.sas7bdat** supports Examples #1-9, #11.
   - **sasclass.sas7bdat** supports Example #10
   - **mplsweather.sas7bdat** supports Example #12
2) SAS Programs
   - **ChartBBData.sas** uses the base ball data to chart a group|midpoint|subgroup chart.
   - **HOWDemo.sas** Source code that generates Examples #1-12.

LISTINGS OF THE *MEETINGS* AND *MPLSWEATHER* DATA SETS

Meetings **Data Set**

Obs	Dept	Room	Room Number	Date	Hours
1	ACCOUNTS	C339	339	01/10/1995	1.00
2	ACCOUNTS	B100	100	01/24/1995	0.50
3	SHIP	C339	339	01/30/1995	2.00
4	MARKET	C301	301	02/24/1995	3.50
5	SHIP	C339	339	02/28/1995	4.00
6	MARKET	C301	301	03/01/1995	4.00
7	ACCOUNTS	B100	100	03/03/1995	3.50
8	ACCOUNTS	B100	100	03/08/1995	0.50
9	ACCOUNTS	B100	100	03/21/1995	0.25
10	SHIP	C301	301	03/27/1995	1.50
11	ACCOUNTS	C301	301	03/29/1995	0.50
12	SHIP	C339	339	04/12/1995	0.50
13	MARKET	C301	301	04/25/1995	1.50
14	MARKET	B100	100	05/02/1995	0.50
15	SHIP	C301	301	05/12/1995	2.50
16	MARKET	C301	301	05/25/1995	3.50
17	SHIP	C301	301	06/01/1995	2.00
18	SHIP	C339	339	06/07/1995	3.00
19	SHIP	C301	301	06/14/1995	2.00
20	MARKET	C339	339	07/12/1995	0.50
21	ACCOUNTS	B100	100	08/03/1995	3.00
22	SHIP	C301	301	08/14/1995	2.50
23	SHIP	C339	339	08/28/1995	1.50
24	MARKET	C339	339	09/15/1995	3.00
25	ACCOUNTS	C301	301	09/28/1995	0.50
26	MARKET	C301	301	10/11/1995	3.00
27	SHIP	C301	301	11/01/1995	1.50
28	MARKET	C339	339	11/15/1995	1.00
29	SHIP	B100	100	11/21/1995	2.00
30	SHIP	C339	339	11/22/1995	3.50
31	MARKET	C339	339	12/05/1995	2.50
32	MARKET	C301	301	12/21/1995	2.50

MplsWeather **Data Set**

Month	TF	TC	Above ZeroYvN
1	11.8	-11.22	N
2	17.9	-7.83	N
3	31.0	-0.56	N
4	46.4	8.00	Y
5	58.5	14.72	Y
6	68.2	20.11	Y
7	73.6	23.11	Y
8	70.5	21.39	Y
9	60.5	15.83	Y
10	48.8	9.33	Y
11	33.2	0.67	Y
12	17.9	-7.83	N

TRADEMARK CITATION

CONTACT INFORMATION

The author welcomes feedback via email at perryWatts@comcast.net