MWSUG 2017 - Paper DV01

An Introduction to ODS Statistical Graphics

Kirk Paul Lafler, Software Intelligence Corporation, Spring Valley, California

Abstract

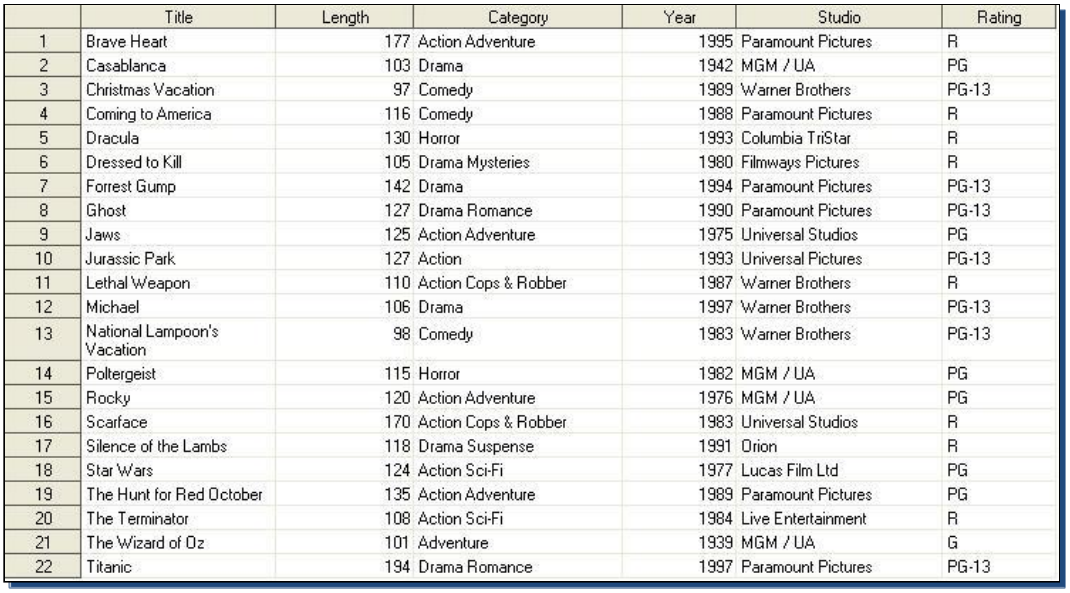
Delivering timely and quality looking reports, graphs and information to management, end users, and customers is essential. This presentation provides SAS® users with an introduction to ODS Statistical Graphics found in the Base-SAS software. Attendees learn basic concepts, features and applications of ODS statistical graphic procedures to create high-quality, production-ready output; an introduction to the statistical graphic SGPLOT procedure, SGPANEL procedure, and SGSCATTER procedure; and an illustration of plots and graphs including histograms, vertical and horizontal bar charts, scatter plots, bubble plots, vector plots, and waterfall charts.

Introduction

It's commonly understood that data patterns and differences are not always obvious from tables and reports. This is where graphical output and their ability to display data in a meaningful way have a clear advantage over tables, reports and other traditional communication mediums. Through the use of effective visual techniques the ability to better understand data is often achieved – a need that appeals to data analysts, statisticians and others. With the availability of ODS Statistical Graphics the SAS user community has a powerful set of tools to produce high-quality graphical output for data exploration, data analysis, and statistical analysis. All that is required to leverage the power available in ODS Statistical Graphics is a SAS® Base license. Once that is secured, ODS Statistical Graphics and its many capabilities are available to users of SAS® Base, SAS/STAT®, SAS/ETS® and SAS/QC® software. The purpose of this paper is to introduce SAS users to the world of ODS Statistical Graphics, its features and capabilities, and basic syntax associated with using the SGPLOT, SGSCATTER and SGPANEL procedures found in SAS Base software.

Data Set Used in Examples

The examples used throughout this paper utilize a data set called, MOVIES. The Movies table consists of six columns: title, length, category, year, studio, and rating. Title, category, studio, and rating are defined as character columns with length and year being defined as numeric columns, shown below.



	Title	Length	Category	Year	Studio	Rating
1	Brave Heart	177	Action Adventure	1995	Paramount Pictures	R
2	Casablanca	103	Drama	1942	MGM / UA	PG
3	Christmas Vacation	97	Comedy	1989	Warner Brothers	PG-13
4	Coming to America	116	Comedy	1988	Paramount Pictures	R
5	Dracula	130	Horror	1993	Columbia TriStar	R
6	Dressed to Kill	105	Drama Mysteries	1980	Filmways Pictures	R
7	Forrest Gump	142	Drama	1994	Paramount Pictures	PG-13
8	Ghost	127	Drama Romance	1990	Paramount Pictures	PG-13
9	Jaws	125	Action Adventure	1975	Universal Studios	PG
10	Jurassic Park	127	Action	1993	Universal Pictures	PG-13
11	Lethal Weapon	110	Action Cops & Robber	1987	Warner Brothers	R
12	Michael	106	Drama	1997	Warner Brothers	PG-13
13	National Lampoon's Vacation	98	Comedy	1983	Warner Brothers	PG-13
14	Poltergeist	115	Horror	1982	MGM / UA	PG
15	Rocky	120	Action Adventure	1976	MGM / UA	PG
16	Scarface	170	Action Cops & Robber	1983	Universal Studios	R
17	Silence of the Lambs	118	Drama Suspense	1991	Orion	R
18	Star Wars	124	Action Sci-Fi	1977	Lucas Film Ltd	PG
19	The Hunt for Red October	135	Action Adventure	1989	Paramount Pictures	PG
20	The Terminator	108	Action Sci-Fi	1984	Live Entertainment	R
21	The Wizard of Oz	101	Adventure	1939	MGM / UA	G
22	Titanic	194	Drama Romance	1997	Paramount Pictures	PG-13

Graphical Design Principles

Good graphical design begins with displaying data clearly and accurately. Data (and information) should be conveyed effectively and without ambiguity. Unnecessary information often distracts from the message, therefore it should be excluded. In their highly acclaimed book, Statistical Graphics Procedures by Example (2011), Mantange and Heath share this message about graphical output, *"A graph is considered effective if it conveys the intended information in a way that can be understood quickly and without ambiguity by most consumers."*

Best Practice Graphical Techniques

Best practice graphical techniques emphasize essential visual elements in a graph or chart for maximum effectiveness. Although the SAS software effectively adheres to many, if not most, of these techniques, users are cautioned to be aware of these techniques. The following best practice graphical techniques are valuable considerations as you design and develop visual output.

- ✓ Group similar elements together to differentiate them from other elements
- ✓ Use distinct colors to set elements apart
- ✓ Use different shapes to distinguish points on a line graph
- ✓ Varying 2-dimensional elements in a line graph offers an effective way to indicate differences, as well as how much difference exists
- ✓ Length serves as an excellent way to communicate differences between elements
- ✓ Size can show that differences exist but doesn't show the amount of difference between elements
- ✓ Width is good for showing that differences exist but fails to show the actual difference between elements
- ✓ Color saturation is useful for showing distinctions and differences between elements

Bar graphs, line graphs and pie charts are popular graphical choices for many SAS users. The following tips serve as guidelines to consider.

Tips for Designing Effective Bar Graphs

- ✓ Avoid misleading scales by starting quantities at zero
- ✓ Verify that the vertical axis is about 25% shorter than the horizontal axis
- ✓ Verify that bars are all the same width
- ✓ Verify that bars are arranged in a logical sequence
- ✓ Use tick marks to display amounts
- ✓ Verify that the space between bars is about half as wide as the bar itself

Tips for Designing Effective Line Graphs

- ✓ Avoid misleading scales by starting quantities at zero
- ✓ Verify that the vertical axis is about 25% shorter than the horizontal axis
- ✓ Use grid lines to display quantities and amounts

Tips for Designing Effective Pie Charts

- ✓ Limit the number of pie slices to no more than six
- ✓ Specify the largest pie slice at the top and work clockwise in decreasing order
- ✓ Combine “small” slices into a “Miscellaneous” or “Other” slice
- ✓ Label each pie slice inside the pie slice
- ✓ Avoid, if possible, the use of legends
- ✓ Emphasize a particular pie slice by color or by extracting it from the pie itself
- ✓ Avoid fill patterns and use color instead

Gestalt Principles of Visual Perception

Gestalt, borrowed from the world of psychology, means “unified whole.” The theory of visual perception was developed by German psychologists in the 1920s. For those interested, more information can be found at, http://www.scholarpedia.org/article/Gestalt_principles.

The SGPLOT Procedure by Example

Whether the data you work with is large or small, or some size in between, the SGPLOT procedure is a powerful tool for handling many of your graphical needs. PROC SGPLOT creates single-cell scatter plots, line plots, histograms, bar charts, box plots and an assortment of other plot types quickly and easily. The SGPLOT procedure supports the following plot statements.

Plot Statement	Plot Statement
Scatter	HBAR / VBAR
Histogram	VLINE
HBOX / VBOX	HighLow
Step	Bubble
Dot	Vector

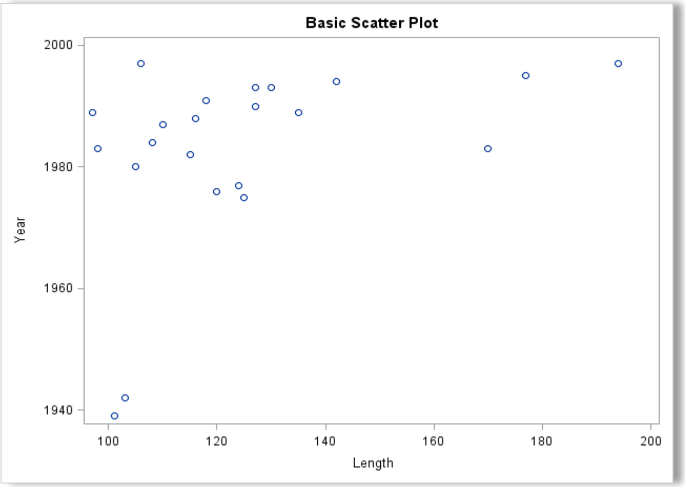
The generalized syntax for the SGPLOT procedure is displayed below.

```
PROC SGPLOT < DATA=data-set-name > < options > ;
  Plot-statement(s) plot-request-parameters < / options > ;
RUN ;
```

Creating a Scatter Plot

Scatter plots are typically produced to display data points on a horizontal and vertical axis for the purpose of determining whether a possible relationship exists between one variable and another. The relationship between two variables is often referred to as their correlation. The SCATTER plot output is produced with the SGPLOT procedure and a SCATTER statement. The resulting scatter plot contains the data points for the LENGTH variable on the horizontal (x-axis) and the YEAR variable on the vertical (y-axis) from the MOVIES data set. As can be seen, a TITLE statement is also specified to display additional information at the top of the scatter plot.

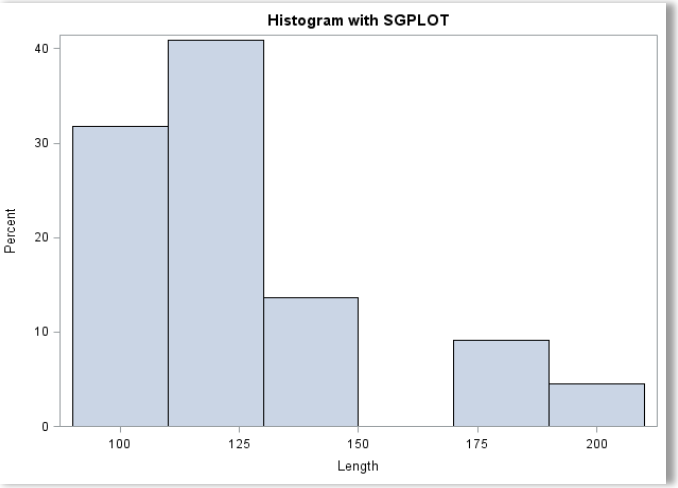
```
TITLE ‘Basic Scatter Plot’ ;
PROC SGPLOT DATA=MOVIES ;
  SCATTER X=Length Y=Year ;
RUN ;
```



Creating a Histogram

Histograms are vertical bar charts that display the distribution of a set of numeric data. They are typically used to help organize and display data to gain a better understanding about how much variation the data has. Much can be learned from viewing the shape of a histogram. The various histogram shapes include: **Bell-shape** – displays most of the data clustered around the center of the x-axis, **Bimodal** – displays two data values occurring more frequently than any other, **Skewed Right** or **Skewed Left** – displays values that tend to be occurring around the high or low points of the x-axis, **Uniform** – displays data equally (no peaks) across a range of values. The next SGPLOT shows the LENGTH variable specified in a HISTOGRAM statement with the MOVIES data set as input. The LENGTH variable is displayed on the horizontal (x-axis) of the histogram output.

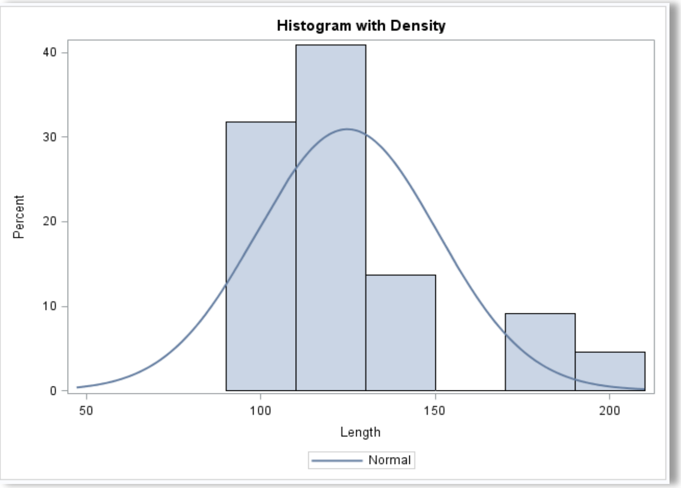
```
TITLE ‘Histogram with SGPLOT’ ;
PROC SGPLOT DATA=MOVIES ;
  HISTOGRAM Length ;
RUN ;
```



Combining a Histogram with a Density Plot

Histograms can also have a density curve applied to display the distribution of values for numeric data. The next SGPLOT shows the LENGTH variable specified in a HISTOGRAM statement combined with the LENGTH variable displayed in a density plot with the DENSITY statement. As before, the LENGTH variable is displayed on the horizontal (x-axis) of the histogram output.

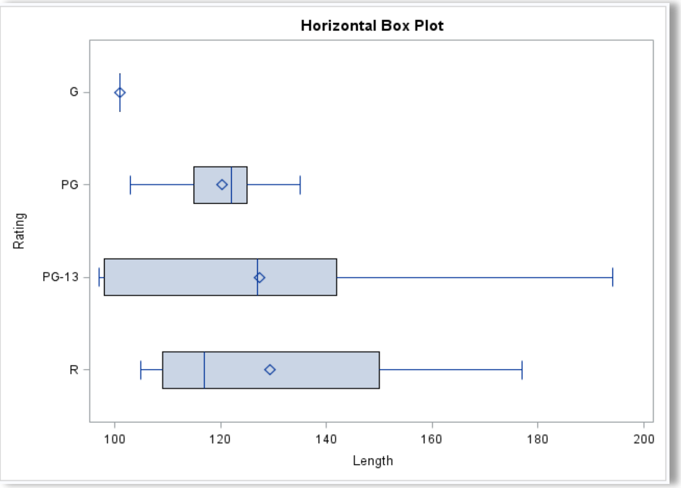
```
TITLE ‘Histogram with Density’ ;
PROC SGPLOT DATA=MOVIES ;
  HISTOGRAM Length ;
  DENSITY Length ;
RUN ;
```



Creating a Horizontal Box Plot

Box plots are useful for displaying data outliers and for comparing distributions for continuous variables. Box plots split the data into quartiles where the range of values traverses the first quartile (Q1) through the third quartile (Q3). The median of the data is represented by a vertical line drawn in the box at the Q2 quartile. Box plots also display the range (or spread) of the data indicating the distance between the smallest and largest value. The next SGPLOT shows the LENGTH variable specified in a HBOX statement with the RATING variable specified in the CATEGORY= option. The LENGTH variable is displayed on the horizontal (x-axis) and the categorical variable, RATING, is displayed on the vertical (y-axis) of the horizontal box plot output. It should also be noted that vertical box plots can be created with the SGPLOT procedure.

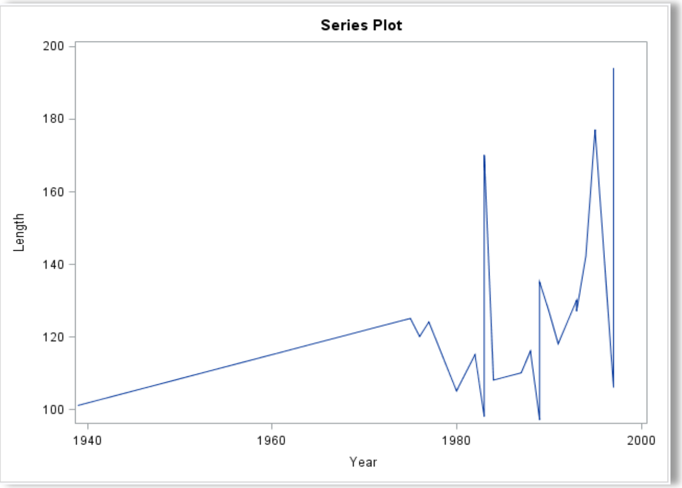
```
TITLE ‘Horizontal Box Plot’ ;
PROC SGPLOT DATA=MOVIES ;
  HBOX Length /
         CATEGORY=Rating ;
RUN ;
```



Creating a Series Plot

Series (or Time Series) plots are useful for displaying trends on paired data at different points in time. Series plots display the time variable along the horizontal (x-axis) and data values along the vertical (y-axis). The next SGPLOT selects the YEAR (or time) variable along the horizontal axis and LENGTH along the vertical axis in a SERIES statement.

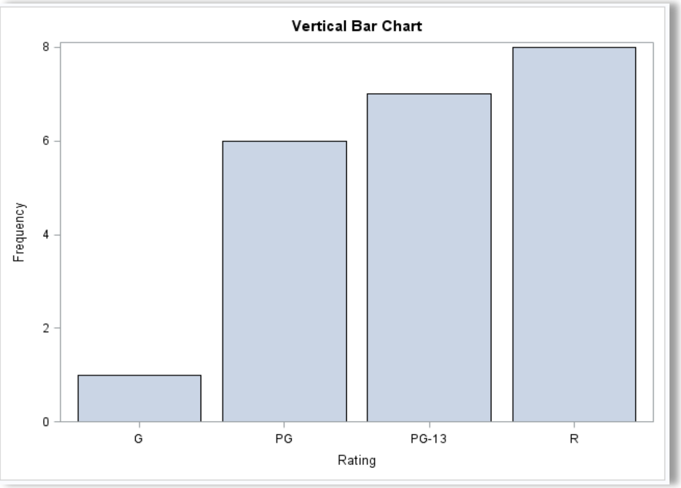
```
PROC SORT DATA=Movies
           OUT=Sorted_Movies ;
  BY Year ;
RUN ;

TITLE ‘Series Plot’ ;
PROC SGPLOT DATA=MOVIES ;
  SERIES X=Year Y=Length ;
RUN ;
```



Creating a Bar Chart

Bar charts are a commonly used graph type to display and compare the quantity, frequency or other measurement for discrete categories or groups. Bar charts display the selected variable along the horizontal (x-axis) and the frequency along the vertical (y-axis). The next SGPLOT displays the distinct values for the RATING variable along the horizontal axis with the frequency value along the vertical axis in a VBAR statement.

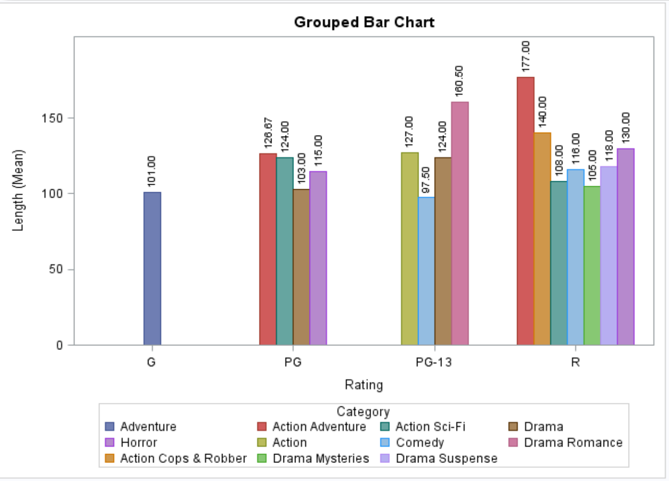
```
TITLE ‘Vertical Bar Chart’ ;
PROC SGPLOT DATA=MOVIES ;
  VBAR RATING ;
RUN ;
```



Creating a Grouped Vertical Bar Chart

Like the bar chart in the previous example, a grouped bar chart is a commonly used graph to display and compare the quantity, frequency or other measurements for discrete categories or groups. A grouped bar chart displays the selected variable along the horizontal (x-axis) and uses a grouping variable to display the various sub-category values for each discrete value of the VBAR variable. The next SGPLOT displays the distinct value(s) for the RATING variable and the distinct value(s) for the grouped variable, CATEGORY, along the horizontal axis, and the frequency value along the vertical axis in a VBAR statement. Since the mean value is computed and displayed at the top of each grouped bar for the CATEGORY variable, a FORMAT statement is specified to display values in the desired formatting.

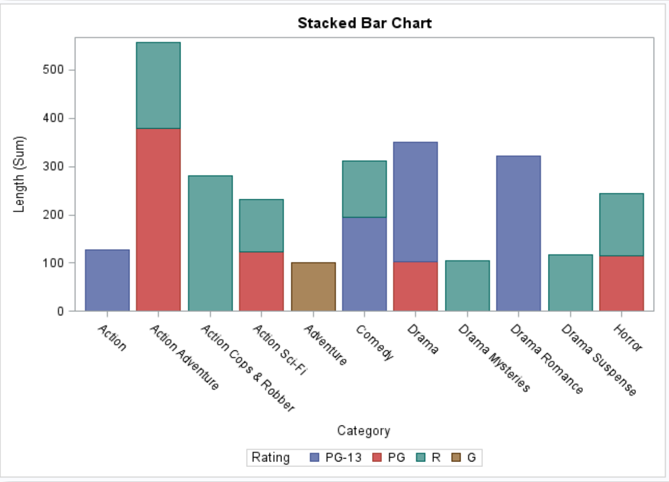
```
TITLE ‘Grouped Bar Chart’ ;
PROC SGPLOT DATA=MOVIES ;
  FORMAT Length 5.2 ;
  VBAR RATING  /
         RESPONSE=Length
         STAT=Mean
         GROUP=CATEGORY
         GROUPDISPLAY=Cluster
         DATALABEL ;
RUN ;
```



Creating a Stacked Vertical Bar Chart

Like the bar charts in the previous two examples, a stacked bar chart displays and compares the quantity, frequency or some other measurement for discrete categories or groups. A stacked bar chart displays the distinct values for the selected variable of the VBAR statement and stacks the grouping variable, designated with the GROUP= option) along the horizontal (x-axis). The next SGPLOT displays the distinct value(s) for the CATEGORY variable and stacks the distinct value(s) for the grouped variable, RATING, along the horizontal axis, and the frequency value along the vertical axis in a VBAR statement.

```
TITLE ‘Stacked Bar Chart’ ;
PROC SGPLOT DATA=MOVIES ;
  VBAR CATEGORY /
         RESPONSE=Length
         GROUP=RATING ;
RUN ;
```



The SGSCATTER Procedure by Example

Whether the data you work with is large or small, or some size in between, the SGSCATTER procedure is a powerful tool for handling many of your graphical needs. PROC SGSCATTER provides single-statement control to a number of scatter plot panels and matrices. The SGSCATTER procedure supports the following plot statements.

Plot Statement
Plot
Compare
Matrix

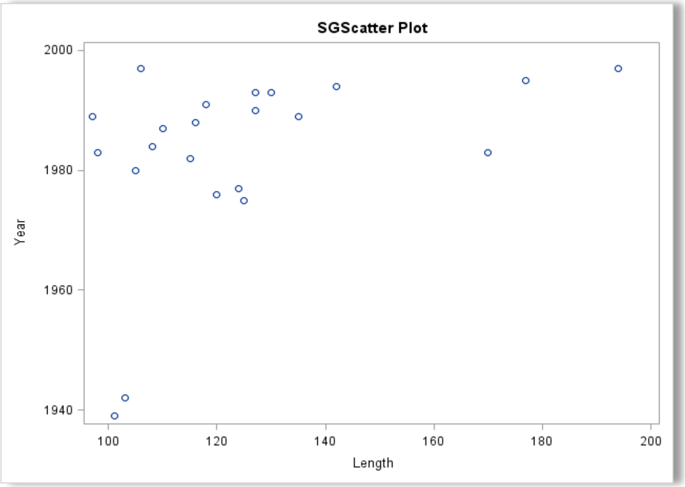
The generalized syntax for the SGSCATTER procedure is displayed below.

```
PROC SGSCATTER < DATA=data-set-name > < options > ;
  Plot variable1 * variable2 / < options > ;
  Compare X=(variable(s)) Y=(variable(s)) < / options > ;
  Matrix variable(s) < / option(s) > ;
RUN ;
```

Creating a Scatter Plot with SGSCATTER

SGScatter plots display data points on a horizontal and vertical axis for the purpose of determining whether a possible relationship exists between one variable and another. The relationship between two variables is often referred to as their correlation. The SCATTER plot output is produced with the SGSCATTER procedure and a PLOT statement. The resulting scatter plot contains the data points for the LENGTH variable on the horizontal (x-axis) and YEAR variable on the vertical (y-axis) from the MOVIES data set.

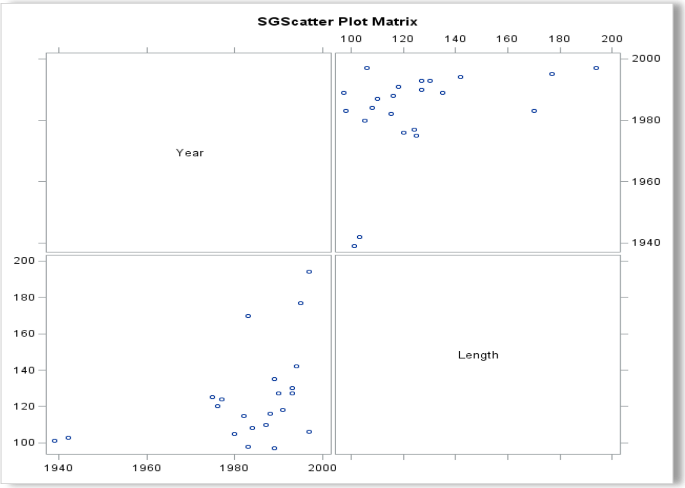
```
TITLE ‘SGScatter Plot’ ;
PROC SGSCATTER DATA=MOVIES ;
  PLOT Year * Length ;
RUN ;
```



Creating a Scatter Matrix Plot with SGSCATTER

Scatter matrix plots display data points on a horizontal and vertical axis for the purpose of determining whether a possible relationship exists between one variable and another. The SCATTER plot output is produced with the SGSCATTER procedure and a MATRIX statement. The resulting scatter plot contains a two-pair scatter matrix plot from the MOVIES data set.

```
TITLE ‘SGScatter Plot Matrix’ ;
PROC SGSCATTER DATA=MOVIES ;
  MATRIX Year Length ;
RUN ;
```



The SGPANEL Procedure by Example

The SGPANEL procedure provides the ability to create a single-panel of plots or charts using one or more classification variables. Unlike a single-page plot or chart as is produced with the SGPLOT procedure, the SGPANEL procedure is able to produce a panel of plots or charts in a single image, or multiple plots or charts displayed in multiple panels. The result is a panel of plots or charts that display relationships among variables. The SGPANEL supports a number of plot types including histograms, horizontal bar charts, vertical bar charts, horizontal box plots, vertical box plots, and plot and chart types. The SGPANEL procedure supports the following statements.

SGPanel Statement
Panelby
Other Plot / Chart Statements

The generalized syntax for the SGPANEL procedure is displayed below.

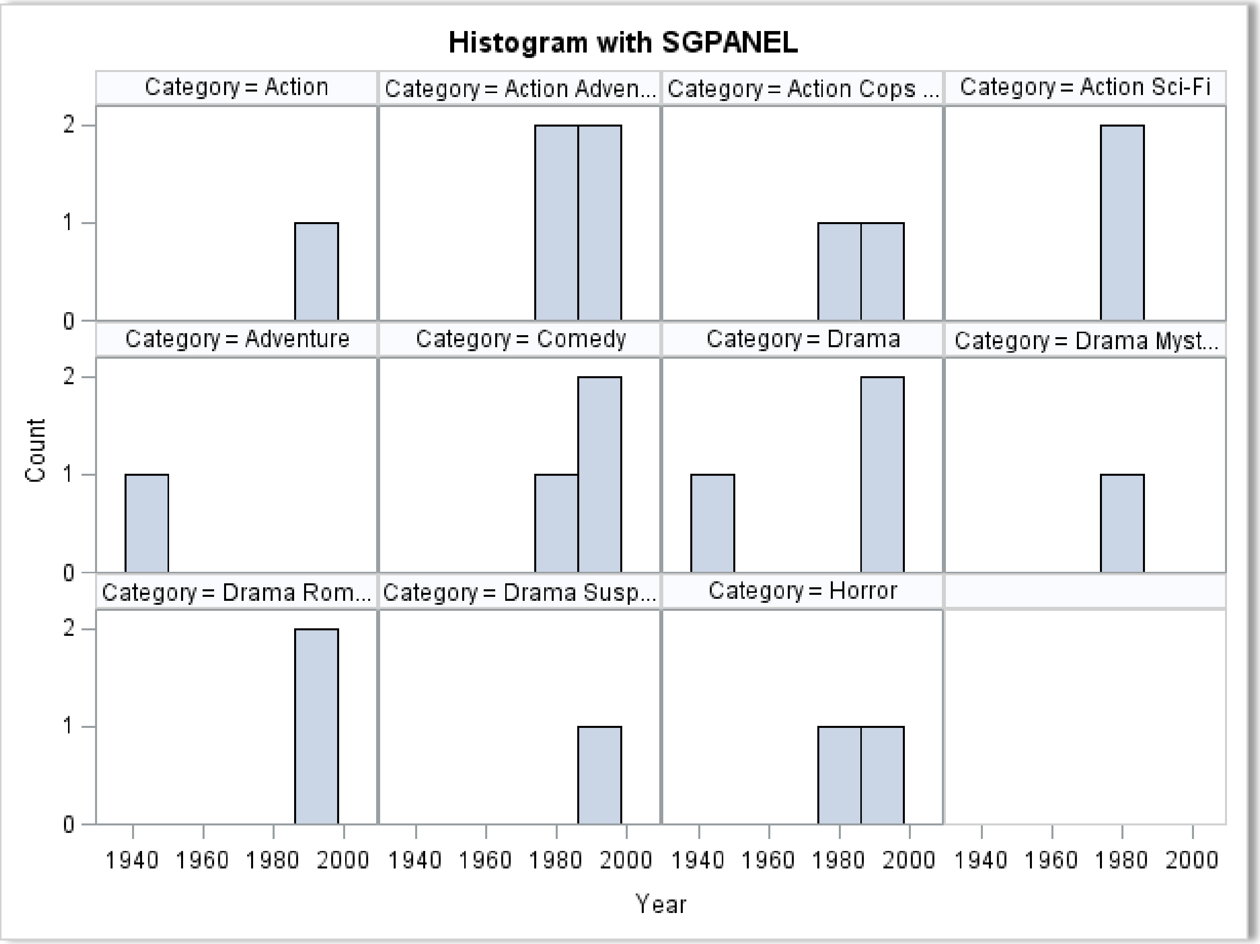
```
PROC SGPANEL < DATA=data-set >  < options > ;
  PANELBY classvar1 < classvar2 . . . < classvarn > /
              options > ;
  < plot / chart statement > ;
RUN ;
```

Creating a Panel of Histograms with SGPANEL

A panel of plots and charts created with the SGPANEL procedure displays values of one or more classification variables. The purpose of a panel of plots is to be able to compare one or more variables with one another. The next example shows a panel of histograms created with the SGPANEL procedure and a HISTOGRAM statement. The resulting panel of histograms contains the number of movies (SCALE=Count) by category corresponding to the year the movie was produced from the MOVIES data set.

```
TITLE ‘Histogram with SGPANEL’ ;
PROC SGPANEL DATA=MOVIES ;
  PANELBY CATEGORY / ROWS=3 COLUMNS=4 ;
  HISTOGRAM YEAR   / SCALE=COUNT ;
RUN ;
```

SGPANEL Output

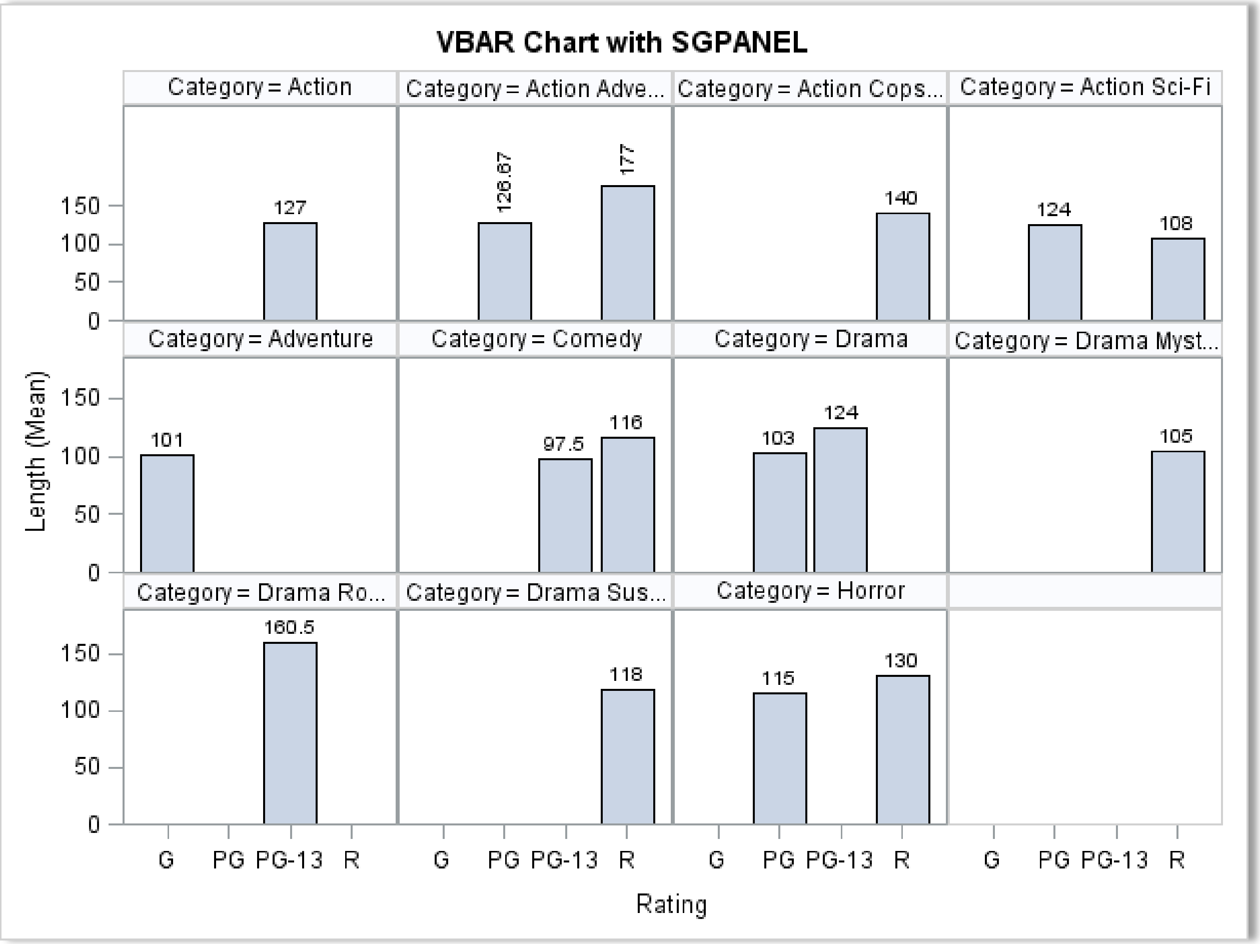


Creating a Panel of Bar Charts with SGPANEL

A panel of bar charts can be created with the SGPANEL procedure using one or more classification variables. As before, the purpose of a panel of bar charts could be for comparing one or more variables values against another. The next example shows a panel of vertical bar charts created with the SGPANEL procedure and a VBAR statement. The resulting panel of vertical bar charts contains the average movie lengths (STAT=Mean) for each category of movie corresponding to the movie rating (i.e., G, PG, PG-13, and R) from the MOVIES data set.

```
TITLE ‘VBAR Chart with SGPANEL’ ;
PROC SGPANEL DATA=MOVIES ;
  PANELBY CATEGORY / ROWS=3 COLUMNS=4 ;
  VBAR RATING      / RESPONSE=Length STAT=Mean DATALABEL ;
RUN ;
```

SGPANEL Output



Conclusion

It's commonly understood that data patterns and differences are not always obvious from tables and reports. This is where graphical output and their ability to display data in a meaningful way have a clear advantage over tables, reports and other traditional communication mediums. Through the use of effective visual techniques the ability to better understand data is often achieved. Good graphical design begins with displaying data clearly and accurately. Data (and information) should be conveyed effectively and without ambiguity. Because unnecessary information often distracts from the message, it should be excluded. This paper introduced SAS users to the world of ODS Statistical Graphics, its features and capabilities, and basic syntax associated with using the SGPLOT, SGPANEL and SGSCATTER procedures found in SAS Base software.

References

Kincaid, Chuck (2010), *"SGPANEL: Telling the Story Better,"* Proceedings of the 2010 SAS Global Forum (SGF) Conference, COMSYS, Portage, MI, USA.

Lafler, Kirk Paul; Joshua M. Horstman and Roger D. Muller (2017), *"Building a Better Dashboard Using SAS® Base Software,"* Proceedings of the 2017 Pharmaceutical SAS Users Group (PharmaSUG) Conference, The Trinomium Group, USA.

Lafler, Kirk Paul; Joshua M. Horstman and Roger D. Muller (2016), *"Building a Better Dashboard Using SAS® Base Software,"* Proceedings of the 2016 Pharmaceutical SAS Users Group (PharmaSUG) Conference, The Trinomium Group, USA.

Lafler, Kirk Paul (2016), *"Dynamic Dashboards Using SAS®,"* Proceedings of the 2016 SAS Global Forum (SGF) Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2015), *"Dynamic Dashboards Using Base SAS® Software,"* Proceedings of the 2015 South Central SAS Users Group (SCSUG) Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Lafler, Kirk Paul (2015), *"Dynamic Dashboards Using SAS®,"* Proceedings of the 2015 SAS Global Forum (SGF) Conference, Software Intelligence Corporation, Spring Valley, CA, USA.

Matange, Sanjay and Dan Heath (2011), Statistical Graphics Procedures by Example, SAS Institute Inc., Cary, NC, USA. Click to view the book at the SAS Book store.

Sams, Scott (2013), *"SAS® BI Dashboard: Interactive, Data-Driven Dashboard Applications Made Easy,"* Proceedings of the 2013 SAS Global Forum (SGF) Conference, SAS Institute Inc, Cary, NC, USA.

Slaughter, Susan J. and Lora D. Delwiche (2010), *"Using PROC SGPLOT for Quick High-Quality Graphs,"* Proceedings of the 2010 SAS Global Forum (SGF) Conference, SAS Institute Inc, Cary, NC, USA.

Zdeb, Mike (2004), *"Pop-Ups, Drill-Downs, and Animation,"* Proceedings of the 2004 SAS Users Group International (SUGI) Conference, University at Albany School of Public Health, Rensselaer, NY, USA.

Acknowledgments

The author thanks Choo Hyung Lee, MWSUG 2017 Data Visualization and Graphics Section Chair, for accepting my abstract and paper; Misty Johnson, MWSUG 2017 Academic Chair; John Xu, MWSUG 2017 Operations Chair; the MidWest SAS Users Group (MWSUG) Executive Board; and SAS Institute for organizing and supporting a great conference!

Trademark Citations

About the Author

Kirk Paul Lafler is an entrepreneur, consultant and founder of Software Intelligence Corporation, and has been using SAS since 1979. Kirk is a SAS application developer, programmer, certified professional, provider of IT consulting services, mentor, advisor, professor at UC San Diego Extension and educator to SAS users around the world, and emeritus sasCommunity.org Advisory Board member. As the author of six books including Google® Search Complete (Odyssey Press. 2014) and PROC SQL: Beyond the Basics Using SAS, Second Edition (SAS Press. 2013); Kirk has written hundreds of papers and articles; been an Invited speaker and trainer at hundreds of SAS International, regional, special-interest, local, and in-house user group conferences and meetings; and is the recipient of 25 "Best" contributed paper, hands-on workshop (HOW), and poster awards.

Comments and suggestions can be sent to:

Kirk Paul Lafler
SAS® Consultant, Application Developer, Programmer, Data Analyst, Educator and Author
Software Intelligence Corporation
E-mail: KirkLafler@cs.com
LinkedIn: http://www.linkedin.com/in/KirkPaulLafler
Twitter: @sasNerd