Paper RI-08

Do SAS® users read books? Using SAS graphics to enhance survey research

Barbara B. Okerson, WellPoint, Richmond, VA

ABSTRACT

In survey research, graphics play two important but distinctly different roles. *Visualization graphics* enable analysts to view respondent segments, trends and outliers that may not be readily obvious from a simple examination of the data. *Presentation graphics* are designed to quickly illustrate key points or conclusions to a defined audience from the analysis of the survey responses. SAS provides the tools for both these graphics roles through SAS/Graph and ODS graphics procedures. Using a survey of the Virginia SAS Users Group (VASUG) as the data source, this paper answers the above question and more while illustrating several SAS techniques for survey response visualization and presentation. The techniques presented here include correspondence analysis, spatial analysis, heat maps and others.

Results included in this paper were created with version 9.2 of SAS on a Windows 64-bit server platform and use Base SAS, SAS/STAT and SAS/GRAPH. SAS Version 9.1 or later and a SAS/GRAPH license are required for ODS graphics extensions. The techniques represented in this paper are not platform-specific and can be adapted by beginning through advanced SAS users.

INTRODUCTION

Conducting surveys is one of the easiest and most accessible methods for studying populations. SAS provides several tools that assist with survey design and analysis. Proc SURVEYSELECT can be used to identify a sample. SAS procedures used for standard analysis include MEANS, FREQ, TABULATE, and REPORT. For more advanced analysis SAS includes the SURVEYMEANS, SURVEYFREQ, SURVEYLOGISTIC and SURVEYREG procedures. Many papers have been written about the use and abuse of these SAS procedures. Much less has been written about how to display the results during analysis and after the survey has been conducted and analyzed. Often the results are just thrown into a table or chart without thought as how to best display. The focus of this paper is not on the survey process, but rather on using SAS graphics for meaningful display of the data and results.

Members of VASUG were surveyed to provide a convenience sample as a source of data to illustrate survey graphics techniques. The survey was conducted through Survey Monkey and was available to all those on the VASUG e-mail list. Participants were not identified and remain anonymous. Survey questions were designed solely for the purpose of graphics illustration and no attempt to validate was made. A copy of the survey is available at the end of this paper as Appendix 1.

VASUG is a non-profit organization serving SAS users throughout the commonwealth of Virginia. VASUG states its purpose as "educating and rendering assistance to fellow SAS Users in the Commonwealth of Virginia, providing an accessible source of information pertaining to the use of SAS software."

VISUALIZATION GRAPHICS

Visualization graphics differ from presentation graphics in that the focus is to understand the data rather than clearly depict results. With that in mind, these graphics likely incorporate large numbers of data points, enabling the viewing of segmentation and trends that may not be obvious from the viewpoint of tabular reports or summary display graphics. Additionally these graphics are a great way to see the impact of outliers.

Visualization graphics are noisy by design, containing lots of information and capturing several dimensions across multiple axes and quadrants and often including variations in colors and sizes. Because of the complexities of the displays, these graphics typically require meticulous study and careful interpretation to identify their meaning. Examples of visualization graphics include correspondence analysis and heat maps.

EXAMPLE ONE – CORRESPONDENCE ANALYSIS

Correspondence analysis is an exploratory data analytic technique that allows rows and columns of a cross-tabulation (two-way and multi-way) to be displayed as points in two-dimensional space. Correspondence analysis is a multivariate technique developed by Jean-Paul Benzécri in the 1970s. It is in effect a principal component analysis that is applied to categorical rather than continuous data. It provides a means of displaying a set of data in two-dimensional graphical form and has also been referred to as reciprocal averaging, optimum scaling, homogeneity analysis, and scalogram analysis.

SAS provides a procedure for correspondence analysis in SAS/STAT, Proc CORRESP. The CORRESP procedure performs simple and multiple correspondence analysis and can read two kinds of input: categorical responses on two or more classification variables and two-way contingency tables. For this example, raw categorical responses will be used. When using categorical data, the information from the cross-tab, as seen in the example below, is usually collected using simple multi-coded questions or semantic rating scales.

Here is the syntax for Proc CORRESP:

PROC CORRESP < options > **;**
TABLES < row-variables, > column-variables **;**
VAR variables **;**
BY variables **;**
ID variable **;**
SUPPLEMENTARY variables **;**
WEIGHT variable **;**

Until the introduction ODS statistical graphics, the %PLOTIT macro was the method of choice to display these results as high-resolution scatter plots of labeled points. This macro positions labels, draws curves, vectors, and circles, and shade to show density or a third variable. The %PLOTIT macro also can control the colors, sizes, fonts, and general appearance of the output plots and is a part of the SAS autocall library. Using the VASUG survey, the following SAS code is used for the %plotit macro. The plot is also shown below.

```
Proc corresp data=vasug2 outc=corr;
   tables gender sas_level,grocery;
   supvar gender;
run;
%plotit(data=Corr, datatype=corresp)
```

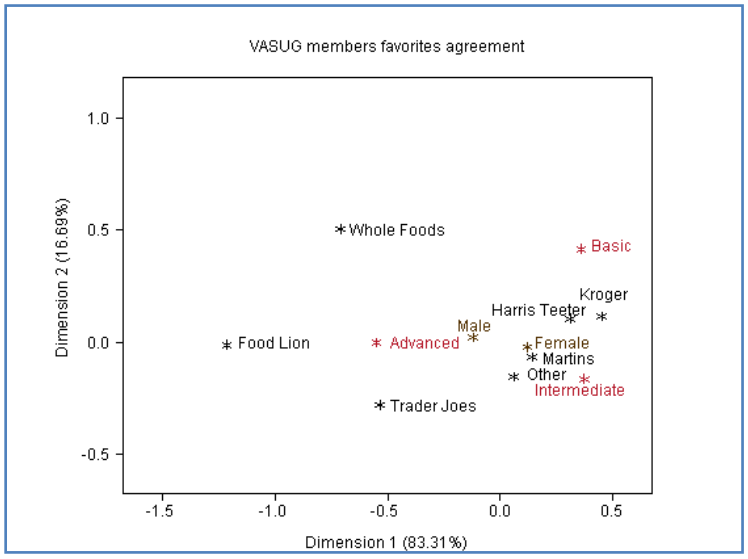


Figure 1a. Plotit Macro Correspondence Plot.

This shows SAS code and display for the same information using ODS Graphhics.

```
ods html;
ods graphics on;
Proc corresp data=vasug2 outc=corr;
   tables gender sas_level,grocery;
   supvar gender;
run;
ods graphics off;
ods html close;
```

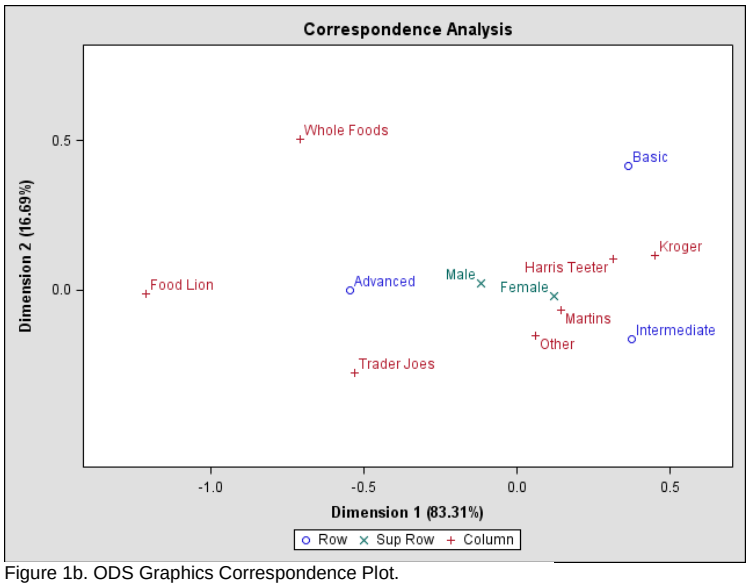


Figure 1b. ODS Graphics Correspondence Plot.

Interpreting a correspondence analysis graphic requires understanding that it is in actuality a depiction of a crosstabs. Rows with similar patterns of counts produce points that are close together, and columns with similar patterns of counts also produce points close together. The two closest points on the graphics above are between females and Martins, suggesting that female SAS users are more likely to prefer Martins as their grocery store while male SAS users had no distinct preferences. The level of SAS user does not appear to make any difference.

EXAMPLE TWO – HEAT MAP

A heat map is a nothing more than a graphical representation of data where the individual values are represented as a matrix of colors. While software designer Cormac Kinney originally coined and trademarked the term "heatmap" in 1991 to describe a 2D grayscale display of financial data, a heat map is any visual display that encodes data values as variations in color. While geographic maps can be used as the underlying structure for heat maps, this is not necessary for this technique and more often this is not the case. The advantage of displaying data as a heat map is the ability to convey a large amount of information in a relatively small space for speedy visual identification of any existing pattern or patterns.

In the following basic heat map example, SAS user levels (as self reported on the VASUG survey) are compared with the number of pages in their reported favorite book. We can now say that SAS Users read books.

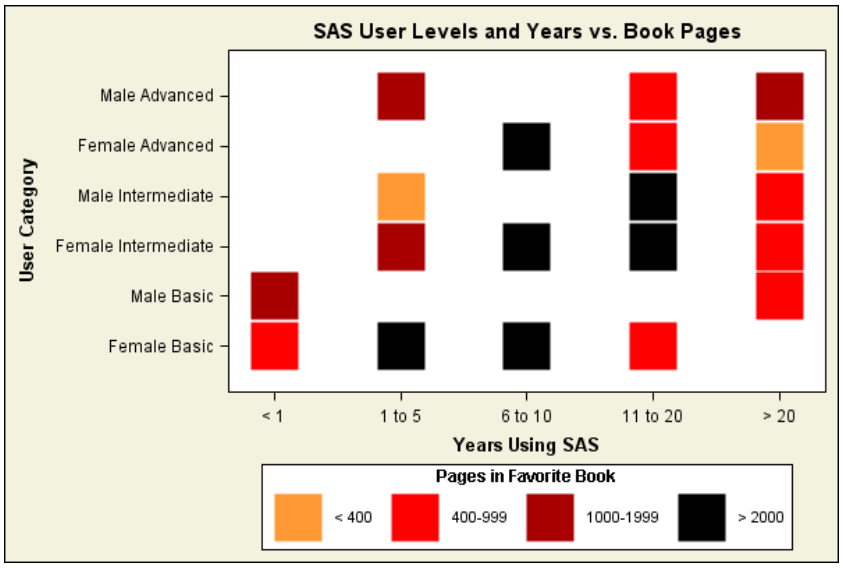


Figure 2. Heat Map with Proc SGPLOT.

In the above example, Proc SGPLOT was used to create the heat map. Proc Template was used to define the colors used in the heat map. The graphic shows that regardless of gender, SAS user level or years using SAS, SAS users read some voluminous tomes. It does appear that those middle range users-- both in terms of self-defined SAS level and years using SAS--spend more time reading trilogies, the entire Harry Potter series, and other lengthy works. The heat map SAS code is below.

```
    /*Set work directory as template directory */
ODS PATH(PREPEND) work.templat(update);
    /*Set the colors for the heat map */
    proc template;
      define style styles.heatmap;
      parent=styles.harvest;
      style graphcolors from graphcolors /
       'gcdata1'=CXFF9933
       'gcdata2'=CXFF0000
       'gcdata3'=CXA80000
       'gcdata4'=CX000000;
      end;
 run;
    /*Sort the group variable - needs to be sorted to display in order*/
 proc sort data=vasug;
  by group;
    run;
    /*Open ODS html, set style, turn on ODS graphics, set dimensions */
    ods html style=heatmap;
    ods graphics on;
    ods graphics / reset width=600px height=400px;
    /*Set title */
    title h=12pt "SAS User Levels and Years vs. Book Pages";
    /*Create the heat map. */
    proc sgplot data=vasug;
      scatter x=SY y=uc/ group=group
      markerattrs=(size=.35in symbol=squarefilled);
    label uc="User Category" SY="Years Using SAS"
      group="Pages in Favorite Book";
    yaxis discreteorder=formatted;
    xaxis discreteorder=formatted;
    run;quit;
    /*Close ODS Graphics and ODS HTML */
    ods graphics off;
    ods html close;
```

While this example uses Proc SGPLOT, SAS does include other procedures that can produce heat maps. Proc SGRENDER combined with Proc TEMPLATE can produce lattice and scatter heat maps. The SAS Visual Analytics product includes a heat map option. Heat maps can also be produced with Proc GMAP. When using GMAP, data values can be linked to geographic coordinates.

EXAMPLE THREE – SPATIAL ANALYSIS

Traditionally, maps have been used as a static display to relay known results to an intended audience after the research has been completed. More recently, maps are increasingly used as geo-visualization or exploratory interfaces to the data being analyzed. More and more frequently, both static and interactive maps are used to support thought processes, next steps, and decisions. Viewing geographic points on a map allows:

- Visual examination of the distribution
- Identification of global and local outliers
- Identification of potential trends
- View of local variation, and
- View of possible spatial autocorrelations

The map below takes an initial look at the location of the VASUG members who responded to the survey. Zip code information was collected as part of the survey.

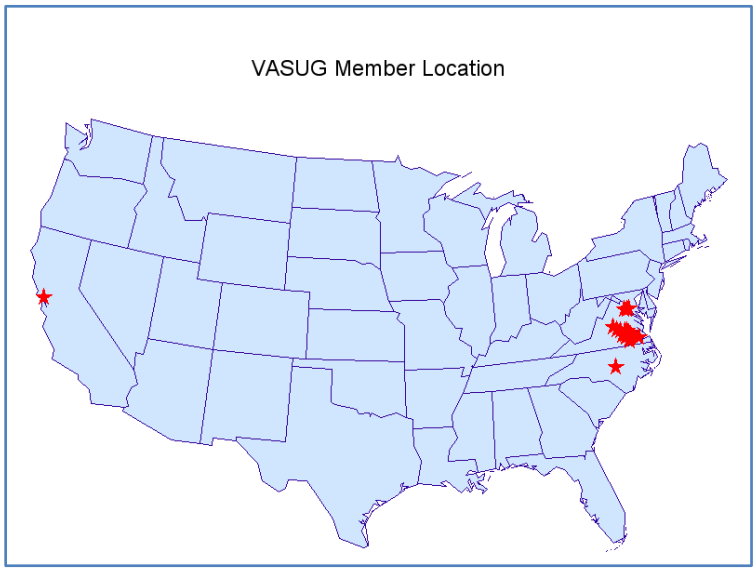


Figure 3a. VASUG Member US GMAP.

The map is created using the maps.states SAS map data set with Alaska, Hawaii, and Puerto Rico not included. Default map code often also removes the District of Columbia but that was necessary for this map. The SAS Code for the map is below.

```
/* Read in and sort ZIP codes. */
data myzip;
  set h.vasugm(keep=zip);
run;
proc sort data=myzip;
  by zip;
run;
/*Get coordinates for zips and convert to radians*/
data longlat;
  merge myzip(in=vasug)
  sashelp.zipcode(rename=(x=long y=lat)keep=x y zip);
  by zip; if vasug;
  x=atan(1)/45*long;
  y=atan(1)/45*lat;
  x=-x;
  keep zip x y;
run;
 /* Create an annotate data set to mark zip code locations.*/
data anno;
  set longlat;
  retain xsys ysys '2' function 'label' size 2.5 flag 1 when 'a';
  style='special';
  text='M'; /*star*/
  color='red';
output;
run;
 /* Combine the map data set with annotate data and remove */
 /*Alaska, Hawaii, and Puerto Rico. Project and separate.*/
data all;
  set maps.states(where=(state not in(2 15 72))) anno;
run;
proc gproject data=all out=allp;
  id state;
run; quit;
data map star;
  set allp;
```

```
  if flag=1 then output star;
  else output map;
run;
 /* Define pattern and title for the map. */
pattern1 v=ms c=c=cxD0E6FF r=50;
title1 ' ';
title3 h=1.5 'VASUG Member Location';
 /* Generate the map. */
proc gmap data=map map=map;
  id state;
  choro state / anno=star nolegend coutline=cx330099;
run; quit;
```

From this map, we can see that almost all of the survey respondents were from Virginia and tended to be in the eastern or northern part of the state. At least one respondent was from California and another from North Carolina, suggesting that our VASUG contact list includes SAS users outside of Virginia. The SAS log also reported two invalid zip codes. Here is another view showing distribution in Virginia.

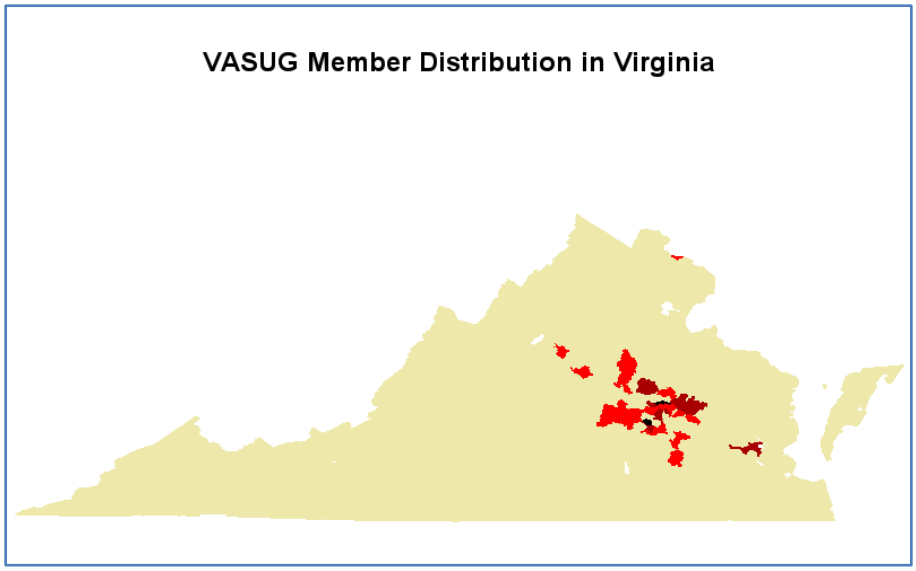


Figure 3b. VASUG Member State Zip Code GMAP.

Darker colors represent larger clusters of users. COUTLINE=same is used to remove boundary lines. Data is at zip code level. A map like this can be further examined by isolating the center section. For zip code boundaries, 2010 census ZCTA files were used. These are available for download by state free of charge. More information can be found at: http://www.census.gov/geo/ZCTA/zcta.html. The SAS code for the above map is below.

```
/* Import Zip Code boundary file for Virginia. */
proc mapimport datafile="r:\bokerson\sesug\sesug 2012\zip
VA\tl_2010_51_ZCTA510.shp"
  out=VAZIP92;
  id zcta5CE10;
run;
/* Match name and format of zip variable in data file for map ID */
data zipva;
  length zcta5CE10 $ 5;
  set vasug; zcta5CE10=zip;
run;
/* Open html and ods graphics, set patterns and title */
ods html;
ods graphics on;
goptions cback=white;
pattern1 c=CXEEE8AA; pattern2 c=CXFF0000;
```

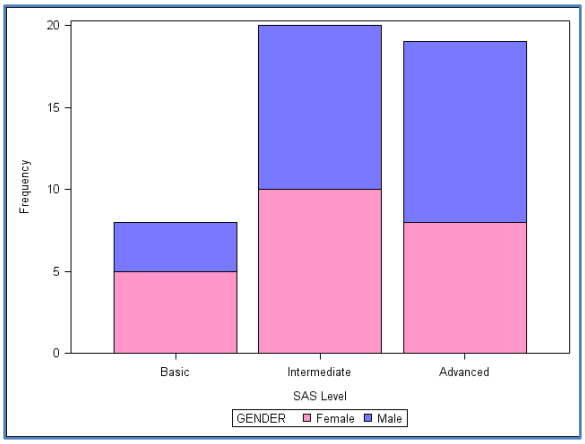
```
pattern3 c=CXA80000; pattern4 c=CX000000;
title4 h=2 "VASUG Member Distribution in Virginia";
/* Create the map. */
proc gmap data=Zipva map=vazip92;
  choro count/nolegend coutline=same levels=4;
  id zcta5CE10;
run;quit;
ods graphics off;
ods html close;
```

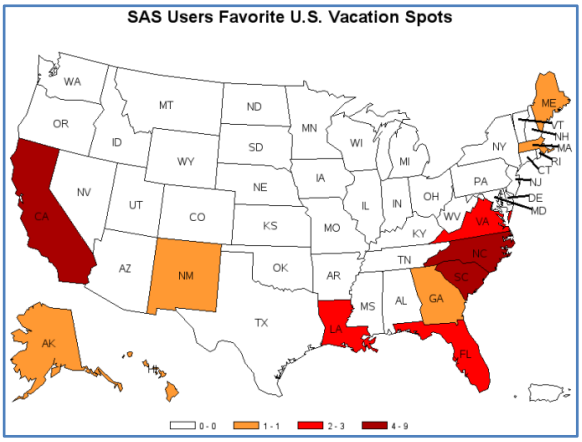
While showing examples of correspondence analysis, heat map and spatial mapping as visualization techniques, these only scratch the surface of graphics that can be used as an initial tool when examining survey. Other visualization graphic techniques include constellation charts, tree charts, scatter plots, radar charts, mosaic plots and cross graphs. Some of these graphics require special training for effective interpretation and all require an ability to focus in on the aspects of the view provided in order to interpret and understand these graphics.

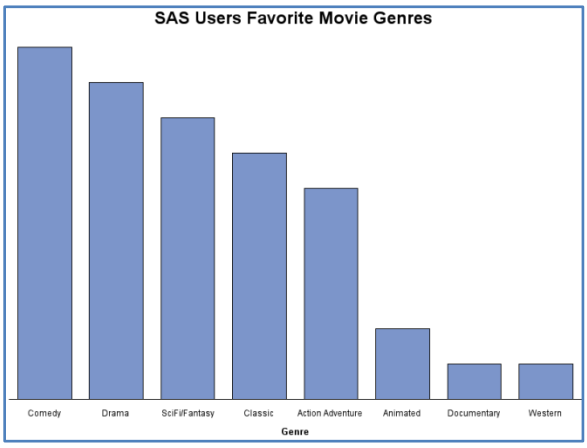
PRESENTATION GRAPHICS

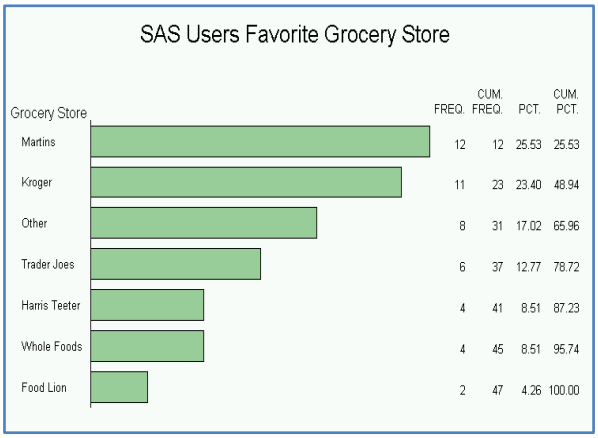
Presentation graphics differ from visualization graphics in that presentation graphics need to immediately convey the point that the graphic is making. They are intentionally focused and avoid any unnecessary dimensions. Good presentation graphics subscribe to the "picture is worth a thousand words" philosophy and, where possible, preclude the need for large amounts of supporting explanatory text. Because the ultimate goal of a survey data analyst is to communicate the results to defined audiences in a manner that they can comprehend, presentation graphics are often developed from the understanding of the visualization graphics.

The graphics that follow all present results from the VASUG survey. Each graphic is intended to be self-explanatory and is intentionally not complex. Enjoy them to learn more about a cross-section of your fellow SAS users.









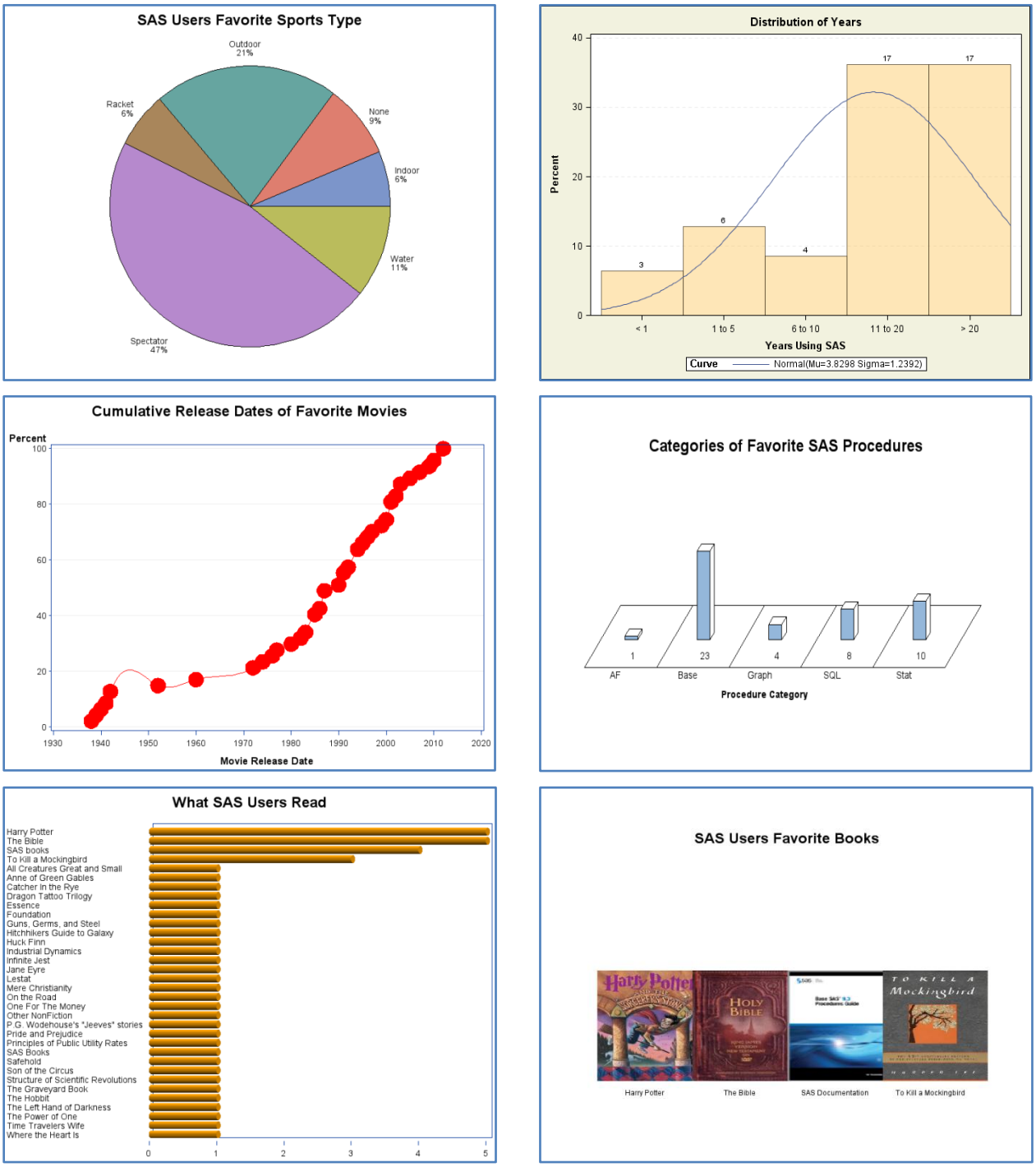


Figure 4. Presentation Graphics Sampler.

The graphics above use a variety of techniques and procedures, including Proc GMAP, Proc GCHART, Proc GPLOT, Proc SGPLOT and Proc UNIVARIATE. Some of the graphics are run with ODS Graphics, including ODS Statistical Graphics, while others are output directly from the SAS procedure.

SAS Code for any of the sampler graphics is available by request from the author.

CONCLUSION

While previous papers on SAS and survey data focus on the use of a variety of features in the SAS survey procedures, this paper has illustrated that SAS also provides the tools for displaying the results of these surveys after and during analysis, providing both visualization and display procedures and techniques. And we can now say that SAS users read books.

REFERENCES

Benzécri, J.P. *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. 1973. Paris, France: Dunod.

Bergland, P. "Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2." Proceedings of the SAS Global Forum 2009 Conference. Available at: http://support.sas.com/resources/papers/proceedings09/246-2009.pdf

Dickinson W. and Hall B., "Proc Corresp for Categorical Data: Correspondence Analysis for Discovery, Display, and Decision-Making, SAS Global Forum, 2008. Available at: http://www2.sas.com/proceedings/forum2008/227-2008.pdf

Hontz, B. "Visualization techniques - graphic visualization vs. presentation," PAI Blog. Available at: http://info.paiwhq.com/bid/53073/Visualization-techniques-graphic-visualization-vs-presentation.

Kemp. K. *Encyclopedia of Geographic Information Science*. Sage Publications, 2007.

SAS Institute. *SAS/Stat 9.1 Users Guide*, SAS Institute, 2004.

Suhr, D. "Selecting a Stratified Sample with PROC SURVEYSELECT," Proceedings of the SAS Global Forum 2009 Conference. Available at: http://support.sas.com/resources/papers/proceedings09/058-2009.pdf.

U.S. Census Bureau. http://www.census.gov/geo/ZCTA/zcta.html

Villacorte, Renato G. "SAS® Fundamentals for Survey Data Processing." Proceedings of the SAS Global Forum 2009 Conference. Available at: http://support.sas.com/resources/papers/proceedings09/150-2009.pdf.

Virginia SAS Users Group. http://www.vasug.org.

Wilkinson, L and Friendly, M. "The History of the Cluster Heat Map," *The American Statistician*, 63:2, 2009, 179-184.

Yeh, S. "SAS® Constellation Diagram Has Many Faces," Proceedings of the SAS Global Forum 2007 Conference, Available at: http://www2.sas.com/proceedings/forum2007/164-2007.pdf.

ACKNOWLEDGMENTS

I would like to acknowledge the members of the Virginia SAS User Group for taking time to respond to the survey used in this paper. Additionally I would like to thank them and members of the WellPoint Client Advisory & Reporting Services division for their suggestions and support in producing this paper.

CONTACT INFORMATION

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

Barbara B. Okerson, Ph.D., CPHQ, FAHM
Senior Health Information Consultant,
West Region Client Advisory & Reporting Services
WellPoint Health & Wellness Solutions
Phone: 804-662-5287
Email: barbara.okerson@wellpoint.com

Appendix 1 – The Survey

This survey was designed solely for the purpose of illustrating graphics techniques in survey research and has not been validated. Survey respondents represent a convenience sample. The survey was administered through Survey Monkey and survey respondents were anonymous.

1. In what ZIP code do you live?
    a. 5 digit zip code
2. What is your gender?
    a. Male
    b. Female
3. How many years have you used SAS?
    a. 1 to 5
    b. 6 to 10
    c. 11 to 20
    d. >20
4. How do you classify your expertise with SAS?
    a. Basic
    b. Intermediate
    c. Advanced
5. What is your favorite book?
6. What is your favorite movie?
7. Where is your favorite vacation destination? (geographic place)
8. What is your favorite sport?
9. What is your favorite grocery store?
10. What is your favorite SAS procedure?