PharmaSUG 2016 - Paper DG15

Elevate your Graphics Game: Violin Plots

Spencer Childress, Rho, Inc., Chapel Hill, NC

ABSTRACT

If you've ever seen a box-and-whisker plot you were probably unimpressed. It lives up to its name, providing a basic visualization of the distribution of an outcome: the interquartile range (the box), the minimum and maximum (the whiskers), the median, and maybe a few outliers if you're (un)lucky. Enter the violin plot. This data visualization technique harnesses density estimates to describe the outcome's distribution. In other words the violin plot widens around larger clusters of values (the upper and lower bouts of a violin) and narrows around smaller clusters (the waist of the violin), delivering a nuanced visualization of an outcome. With the power of SAS/GRAPH®, the savvy SAS® programmer can reproduce the statistics of the box-and-whisker plot while offering improved data visualization through the addition of the probability density 'violin' curve. This paper covers various SAS techniques required to produce violin plots.

INTRODUCTION

The box-and-whisker plot is a basic data visualization which with a little SAS magic can be improved drastically. The SAS programmer needs a few tools to round those hard corners. All data visualizations begin with the underlying data. Throughout this paper the dataset in reference is SASHELP.CARS, which contains qualitative and quantitative data on a number of vehicles. The idea for this paper comes from Sanjay Matange's blog on violin plots.

The primary purpose of this paper is to illustrate similarities and differences between the box-and-whisker plot and the violin plot. Secondly I will discuss augmentations to the violin plot which provide additional information about the data. Producing a violin plot in SAS requires kernel density estimates, descriptive statistics, some data manipulation, and PROC SGPANEL, and each will be thoroughly explained.

THE BOX-AND-WHISKER PLOT

The box-and-whisker plot gives a quick outline of the distribution of continuous data. It's a visualization of the five-number summary, i.e. the sample minimum, first quartile, median, third quartile, and sample maximum. Figure 1 displays an example of the box-and-whisker plot with horsepower as the outcome, continent of origin as the group comparison, and number of cylinders as the panel comparison.

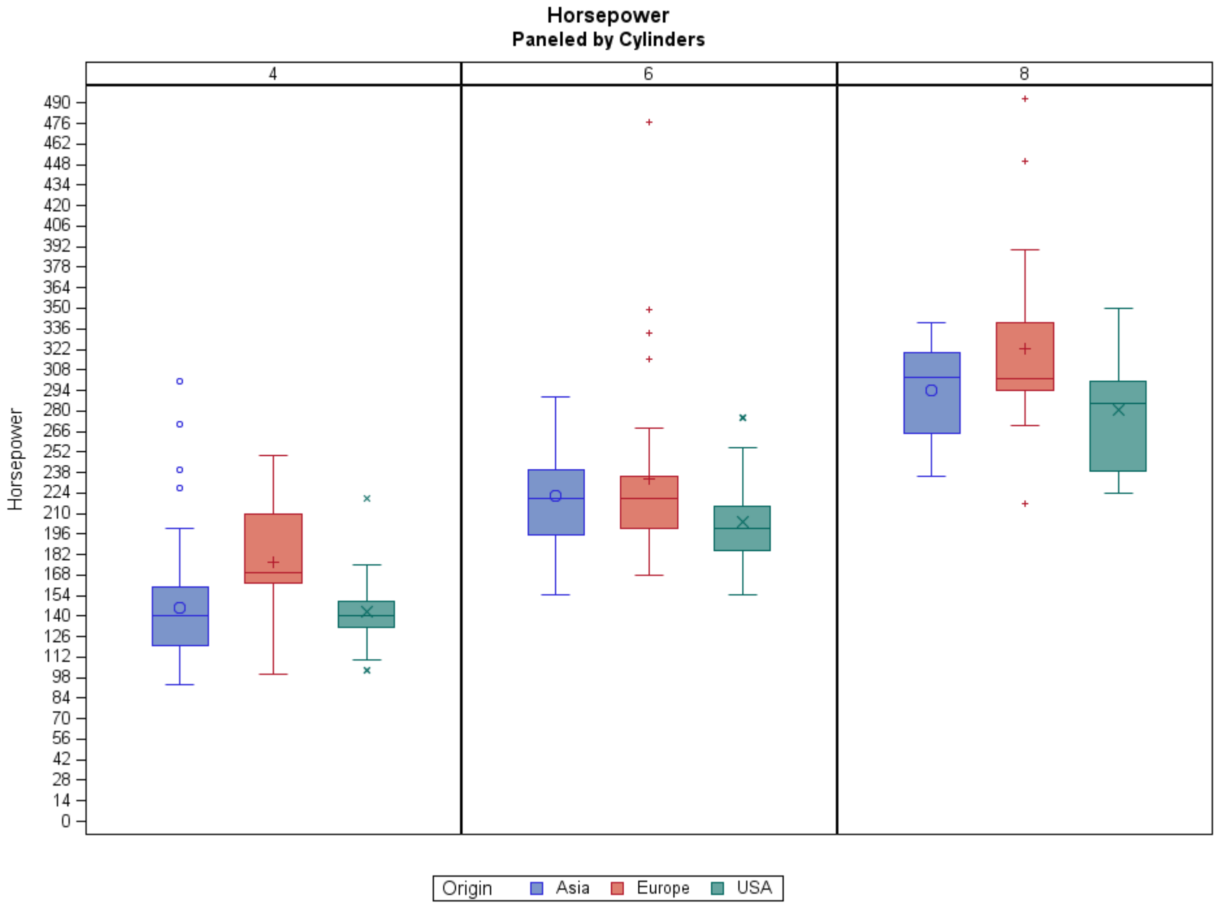

Figure 1. Box-and-Whisker Plots

This plot visualizes the distribution of horsepower by number of cylinders and continent of origin. Each solid box encompasses all points between the first and third quartiles, otherwise known as the interquartile range. The "whiskers" encompass all points inside 1.5x the interquartile range. Points outside the interquartile range are considered outliers for the purposes of this plot. The following code will produce this plot:

```
proc sort
    data = sashelp.cars (where = (cylinders in (4 6 8)))
    out = cars;
    by Cylinders Origin Horsepower;
run;

proc sgpanel
    data = cars;
    panelby Cylinders / novarname
        rows = 1;
    vbox Horsepower /
        group = Origin;
run;
```

It's pretty basic and I know we can do better.

THE VIOLIN PLOT

The violin plot is a box plot with a kernel density plot instead of a box. You might call it an outside-the-box plot. "In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable" ("Kernel Density Estimation," 2016, para. 1). Don't let that scare you away. SAS makes KDE easy and besides, KDE is beyond the scope of this paper.

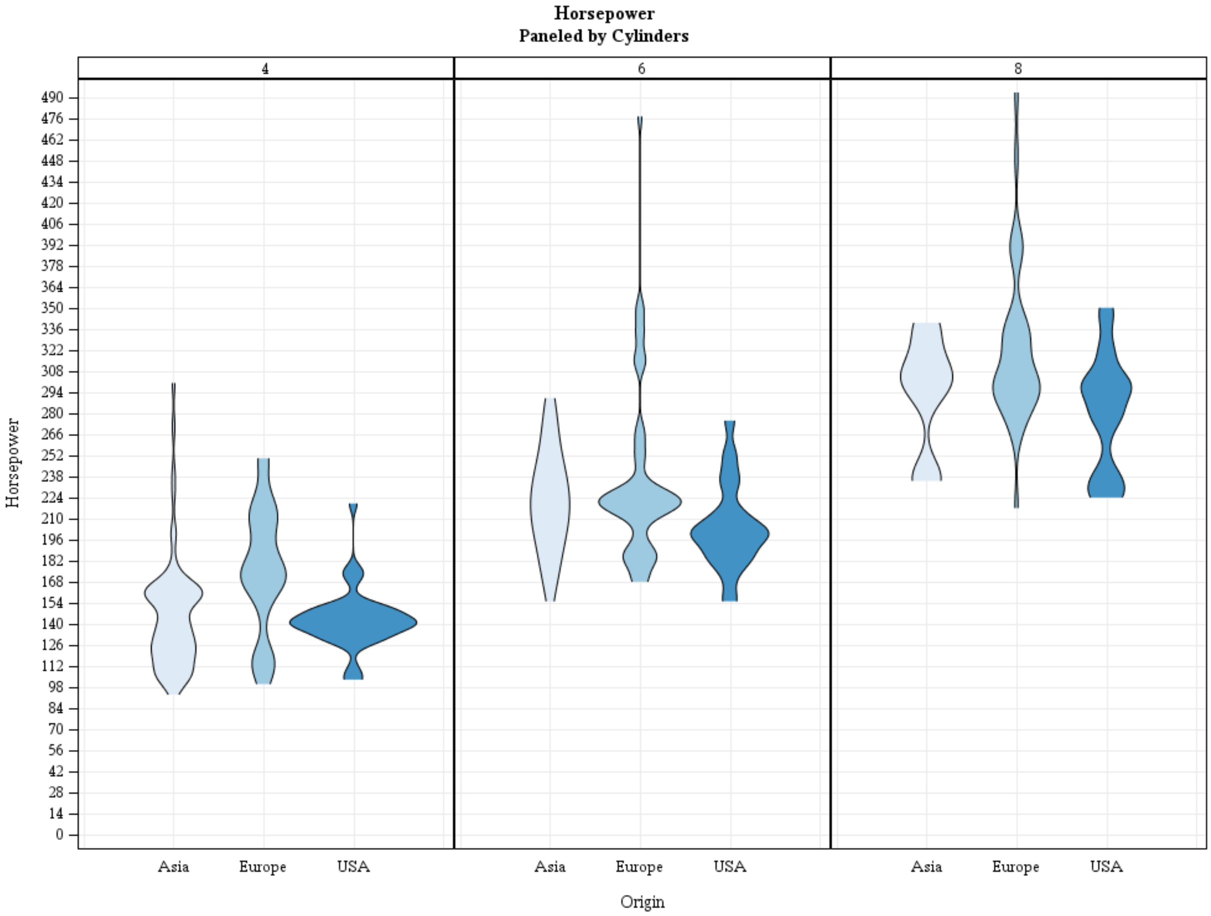

Figure 2. Violin Plots

These violin plots should look pretty familiar. They represent the same range on the y-axis as the box-and-whisker plots, which makes sense because the underlying data are the same. The similarities begin to fade upon further examination. Where the box-and-whisker plot represents a rigid interquartile range the violin plot provides a more nuanced visualization of the heart of the distribution. In some cases the "center" of the distribution is immediately apparent, as in American 4-cylinder vehicles. In other cases it's not so apparent, as in the somewhat bimodal American 8-cylinder vehicles.

At this stage the box-and-whisker plot affords more insight into the data: the quartiles and the mean. Let's add in these statistics:

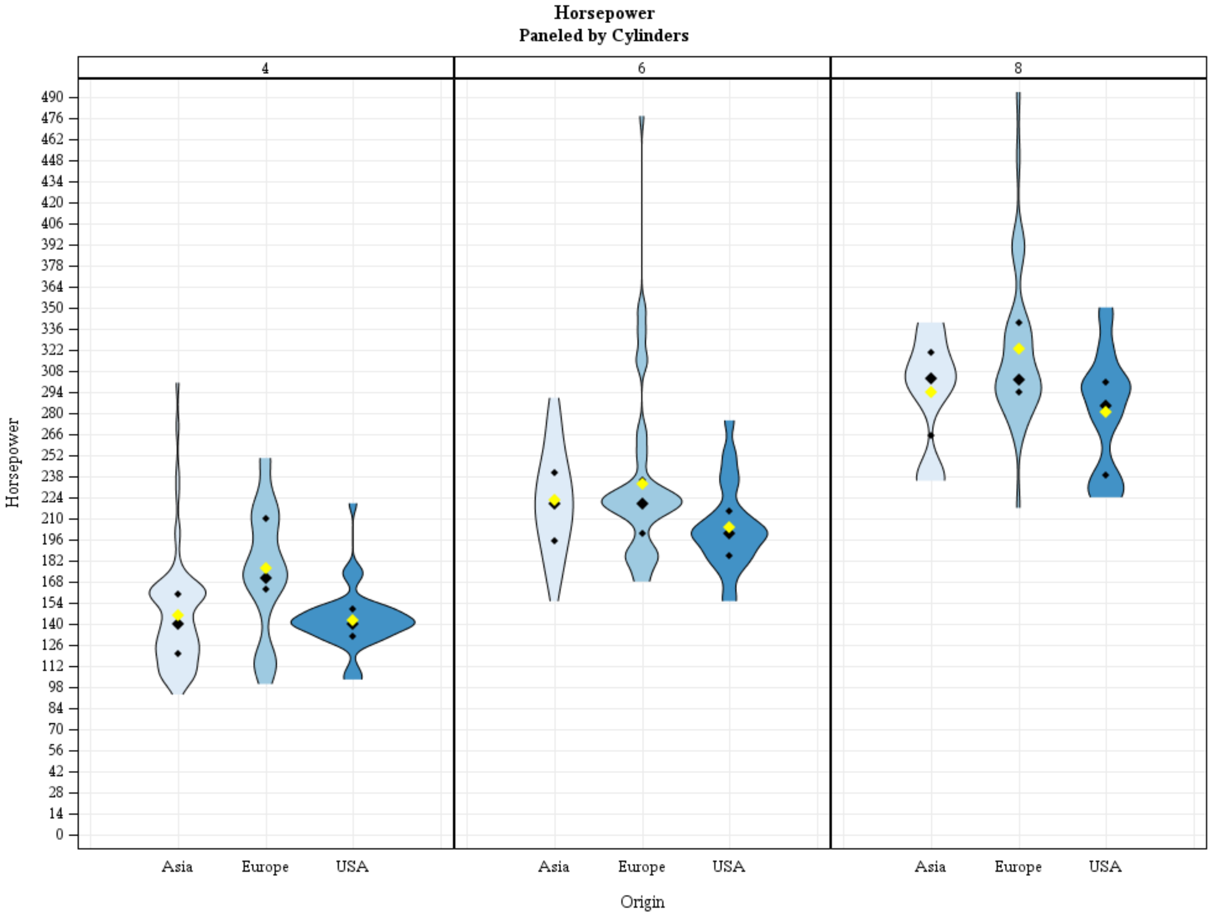

Figure 3. Violin Plots with Quartiles and Mean

The violin plot is quickly catching up with the tried-and-true box-and-whisker plot. The small diamonds represent the first and third quartiles, the large black diamond represents the median, and the large yellow diamond represents the mean. This plot would be a bit more interesting and accessible if we color-coded the quartiles rather than placed symbols.

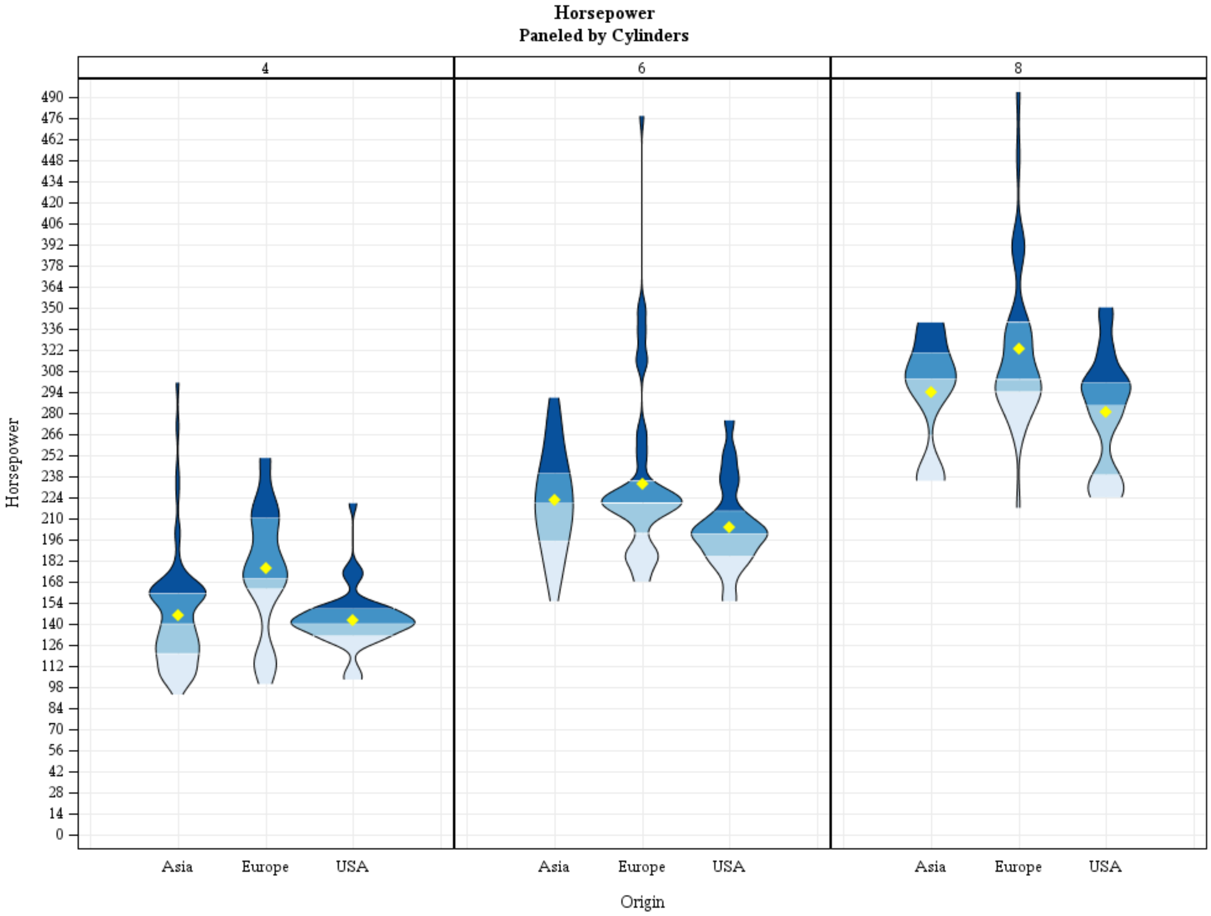

Figure 4. Violin Plots with Color-coded Quartiles and Mean

Not too shabby. We've reproduced the statistics represented in the box-and-whisker plot, namely the quartiles and the mean. With the exception of the "outliers" outside 1.5x the interquartile range these violin plots have got it all. However, perhaps the data analyst would like to see the data points which make up the distribution and further throw in some trend lines between the medians. In Figure 5 we restructure the plot and add jittered data points and trend lines.

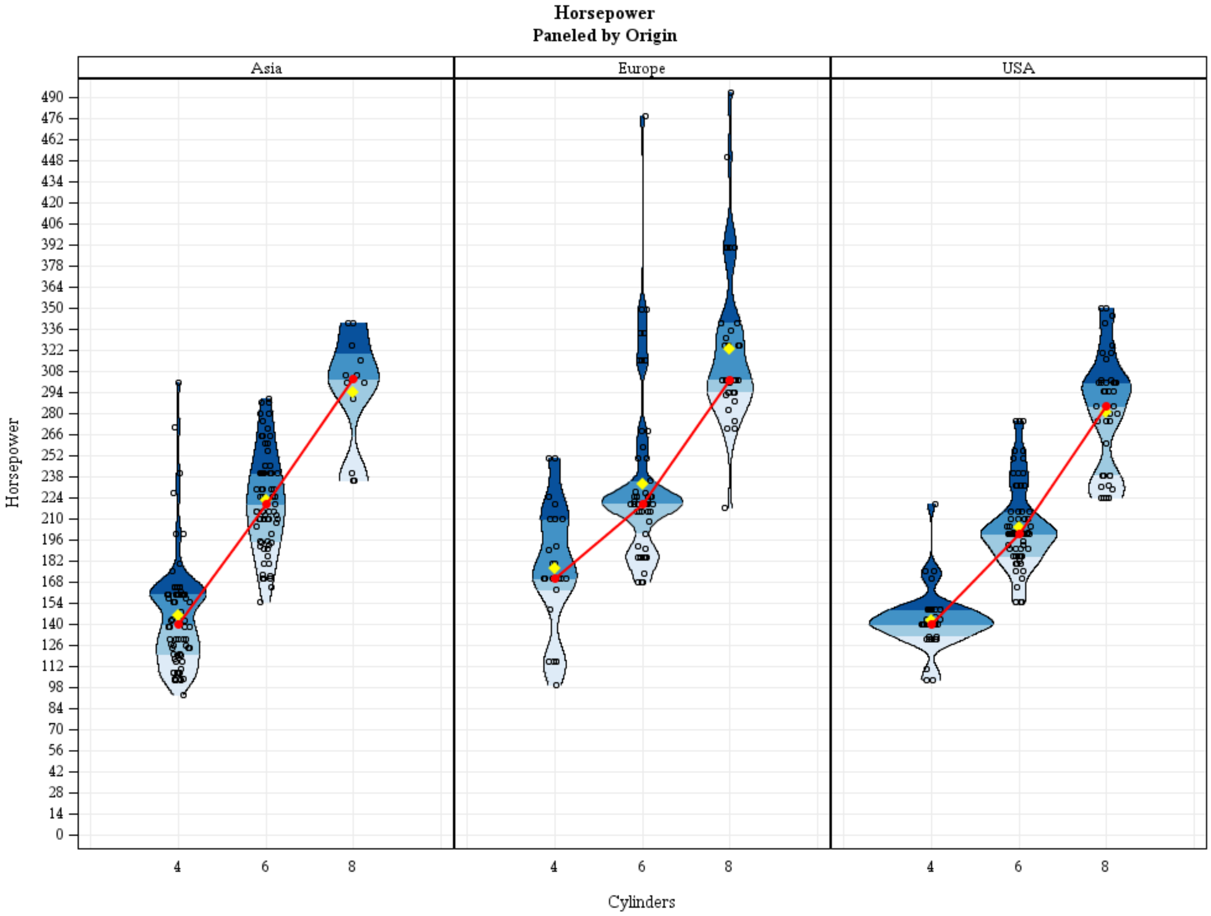

Figure 5. Violin Plots with Jittered Points and Median Trend Lines

With the restructuring and the median trend line we see a clear positive trend between number of cylinders and horsepower. The jittered data points make apparent “true” outliers.

PREPARING THE DATA

We begin with SASHELP.CARS which contains a continuous variable, horsepower, and two categorical variables, number of cylinders and continent of origin. This dataset will inform our kernel density estimation, descriptive statistics, and jittered points. To produce the KDEs we call PROC KDE:

```
proc sort
    data = SASHELP.CARS (where = (cylinders in (4 6 8)))
    out = cars;
    by Origin Cylinders Horsepower;
proc kde
    data = cars;
    by Origin Cylinders;
    univar Horsepower / noprint
        out = KDE;
run;
```

PROC KDE calculates the density of the distribution along the entire range of horsepower, which we use to paint the violins. One can think of KDE in the same context as ultra-granular histograms. The density values are the bins and in order to suss a violin shape out of them they need to be mirrored, i.e. negated in a separate variable. Next up are the descriptive statistics:

```
    proc means noprint nway
        data = cars;
        class Origin Cylinders;
        var Horsepower;
        output
            out    = statistics
            mean   = mean
            p25    = quartile1
            median = median
            p75    = quartile3;
    run;
```

Everyone loves PROC MEANS and it serves its purpose competently. Here we calculate the mean and quartiles of each combination of continent of origin and number of cylinders. Finally we combine our original data points with the kernel density estimates and descriptive statistics for one Frankenstein of an input dataset.

VISUALIZING THE DATA

If PROC KDE and PROC MEANS produce our paint, then our paintbrush is PROC SGPANEL. First we'll apply a base coat with the BAND statement:

```
    proc sgpanel nocycleattrs noautolegend
        data = carsKDEmeans;
        panelby
            Origin / novarname
                rows = 1;
        band
            y = yBand
            lower = lowerBand
            upper = upperBand / fill outline
                group = quartile
                lineattrs = (
                    pattern = solid
                    color = black);
```

The BAND statement takes three inputs: an x- or a y-value, and a lower and upper bound from the opposing axis. In our example horsepower is plots vertically on the y-axis. Therefore our kernel density estimates and their mirrors, or negated values, plot horizontally on the x-axis. Hello violin plots! The group option on the BAND statement colors each quartile individually.

The rest is easy. SCATTER statements plot the jittered data points and descriptive statistics, and a SERIES statement plots the trend line:

```
        scatter
            x = jitteredCylinders
            y = Horsepower /
                markerattrs = (
                    symbol = circle
                    size = 6px
                    color = black);
        scatter
            x = Cylinders
            y = mean /
                markerattrs = (
                    symbol = diamondFilled
                    size = 9px
                    color = yellow);
        series
            x = Cylinders
            y = median /
                lineattrs = (
                    color = red
                    thickness = 2px);
        scatter
            x = Cylinders
            y = median /
                markerattrs = (
                    symbol = circleFilled
                    size = 9px
                    color = red);
    run;
```

And there you have it, a violin plot with jittered data points, descriptive statistics, and trend lines.

THE MACRO CALL

The macro really only requires an input dataset and an outcome variable; everything else is for added effect:

```
    %include 'violinPlot.sas';
    %violinPlot
      /*REQUIRED*/
        (data = cars               /*input dataset*/
        ,outcomeVar = Horsepower /*continuous outcome variable*/
      /*optional*/
        ,groupVar = Cylinders    /*categorical grouping variable*/
        ,panelVar = Origin       /*categorical paneling variable*/
        ,byVar =                 /*categorical BY variable*/
        ,widthMultiplier = .1    /*kernel density width coefficient*/
        ,jitterYN = Yes          /*display jittered data points?*/
        ,quartileYN = Yes        /*display color-coded quartiles?*/
        ,quartileSymbolsYN = No  /*display quartiles as symbols?*/
        ,meanYN = Yes            /*display means?*/
        ,trendLineYN = Yes       /*display a trend line?*/
        ,trendStatistic = Median /*trend line statistic*/
        );
```

CONCLUSION

The box-and-whisker plot visualizes the distribution of a continuous variable fairly primitively. With kernel density estimates and PROC SGPANEL the SAS programmer can generate a far more nuanced and informative graphic. The macro %**violinPlot** allows the production of simple, quick, and professional violin plots from the SAS environment.

REFERENCES

Matange, Sanjay. "Violin Plots." Graphically Speaking. October 30, 2012. Available at http://blogs.sas.com/content/graphicallyspeaking/2012/10/30/violin-plots/.

Kernel density estimation. In *Wikipedia*. Retrieved March 24, 2016, from https://en.wikipedia.org/wiki/Kernel_density_estimation.

ACKNOWLEDGMENTS

I'd like to thank Tee Bahnson for pointing me to Sanjay Matange's work and for providing guidance on this project. Without Mr. Matange's original idea I would probably still be producing box plots so he deserves a big thank you.

RECOMMENDED READING

- https://github.com/RhoInc/sas-violinPlot
- https://en.wikipedia.org/wiki/Box_plot
- https://en.wikipedia.org/wiki/Five-number_summary
- https://en.wikipedia.org/wiki/Violin_plot
- https://en.wikipedia.org/wiki/Kernel_density_estimation

CONTACT INFORMATION

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

Name: Spencer Childress
Enterprise: Rho, Inc.
Address: 6330 Quadrangle Drive
City, State ZIP: Chapel Hill, NC 27517
Work Phone: 919 595 6638
Fax: 919 408 0999
E-mail: spencer_childress@rhoworld.com
Web: https://github.com/samussiah
Twitter: https://twitter.com/samussiah