How to Easily Create Grouped Boxplots in SAS

Authored by

stats writer

December 1, 2025

The ability to create comparative visualizations is crucial for effective data analysis. When working with large datasets containing multiple categorical variables, plotting distributions side-by-side provides immediate insights. This guide details the methodology for generating grouped boxplots in SAS, utilizing the powerful Graph Procedure (`PROC SGPLOT`) and its specialized Box Statement functionality. Grouped boxplots allow analysts to simultaneously compare the central tendency, spread, and potential outliers across distinct subgroups within the data.

Generating these visual comparisons in SAS requires careful specification of both the continuous variable to be plotted and the categorical variable used for grouping. The Box Statement within PROC SGPLOT facilitates this distinction, ensuring that a separate box representation is generated for each unique category defined by the grouping variable. This technique is indispensable for statistical quality control, comparative market analysis, and educational research where differences between populations must be visually quantified. Furthermore, options like `GROUPDISPLAY=CLUSTER` offer advanced layout control, enhancing the clarity of complex comparative graphs.

While various statistical packages offer charting capabilities, SAS provides robust control over the graphical output. The standard approach involves using the `VBOX` (vertical boxplot) or `HBOX` (horizontal boxplot) statements combined with the mandatory `GROUP=` option. Understanding the underlying statistical representation--the five-number summary--is essential for accurate interpretation of the resulting visualizations. This comprehensive tutorial will walk through the data preparation, code execution, and interpretation phases required to master this technique.

Understanding the Statistical Foundation of Boxplots

Boxplots, sometimes referred to as box-and-whisker plots, are standardized graphical representations that concisely summarize the distribution of a dataset. They are exceptionally useful for identifying skewness and the spread of data. The core strength of the boxplot lies in its ability to visualize the five-number summary, a critical set of descriptive statistics that defines the distribution's boundaries and central tendency. When grouped, these plots become powerful tools for multivariate analysis, allowing rapid comparison of these summaries across different strata of the data.

The five-number summary provides a robust, non-parametric summary of the data distribution, making it less susceptible to distortions caused by extreme outliers than simple mean comparisons. Each component plays a vital role in defining the shape and boundaries of the plot:

The minimum value (end of the lower whisker, excluding outliers).
The first quartile (Q1, the lower edge of the box), representing the 25th percentile.
The median (the line inside the box), representing the 50th percentile.
The third quartile (Q3, the upper edge of the box), representing the 75th percentile.

The maximum value (end of the upper whisker, excluding outliers).

The box itself represents the Interquartile Range (IQR), which spans from Q1 to Q3 and contains the middle 50% of the data. The length of this box indicates the variability or spread of the central data. Whiskers extend from the edges of the box to the maximum and minimum values that are not considered outliers. Outliers, often defined as values falling more than 1.5 times the IQR above Q3 or below Q1, are usually plotted as individual points or markers. Analyzing these components across multiple groups is the primary purpose of grouped boxplots, enabling analysts to determine if group medians differ significantly or if one group exhibits higher variability than others.

Setting up the Data for Grouped Analysis

Before any visualization can occur, the data must be appropriately structured. For grouped boxplots in SAS, the dataset must contain at least two variables: one continuous variable (the value to be plotted) and one categorical variable (the group identifier). The categorical variable is crucial as it dictates how the data will be partitioned for the subsequent plotting procedure. In the following example, we establish a simple dataset that simulates observations across three distinct groups: A, B, and C. This step uses the standard SAS `DATA` step along with inline `datalines` for rapid deployment and testing.

The quality of the grouping variable is paramount. It must be accurately defined and consistent across all observations to ensure correct separation during plotting. In this demonstration, the variable `Group` is defined as a character variable (indicated by the dollar sign `$`), which is a common practice for nominal grouping variables. The continuous variable, `Value`, represents the quantitative measurement whose distribution we intend to analyze and compare across these groups.

Below is the code snippet used to create the sample dataset `my_data`. We ensure that the data is clean and ready for immediate use by the graphical procedure that follows. This preliminary step is foundational to successful visualization, ensuring the correct variables are available and properly formatted.

Example: Create Boxplots by Group in SAS

First, let's create a dataset in SAS that contains three different groups, ensuring the data structure is optimized for visualization:

```
/*create dataset: my_data*/
data my_data;
input Group $ Value;
datalines;
```

```
A 7
A 8
A 9
A 12
A 14
B 5
B 6
B 6
B 8
B 11
C 8
C 9
C 11
C 13
C 17
;
run;
```

This dataset, `my_data`, clearly establishes three different groups (A, B, and C) against which we will compare the distribution of the `Value` metric. It is essential to confirm that the input variables align with the requirements for the `VBOX` statement--specifically, identifying which variable serves as the measure and which serves as the category.

Implementing Vertical Boxplots using PROC SGPLOT

The Statistical Graphics Procedure (PROC SGPLOT) is the modern and recommended procedure in SAS for generating high-quality statistical graphs. Unlike legacy procedures, PROC SGPLOT uses a streamlined syntax focused on plot statements rather than complex option specifications, making it ideal for creating grouped visualizations. To generate vertical boxplots, we employ the `VBOX` statement.

The syntax below instructs PROC SGPLOT to use the `my_data` dataset. Within the `VBOX` statement, we specify the quantitative variable (`Value`) first. Crucially, the `/ GROUP=Group` option tells PROC SGPLOT to generate separate boxplots based on the unique values found in the categorical variable `Group`. This is the mechanism that achieves the grouping effect, positioning the plots adjacent to one another for easy comparison.

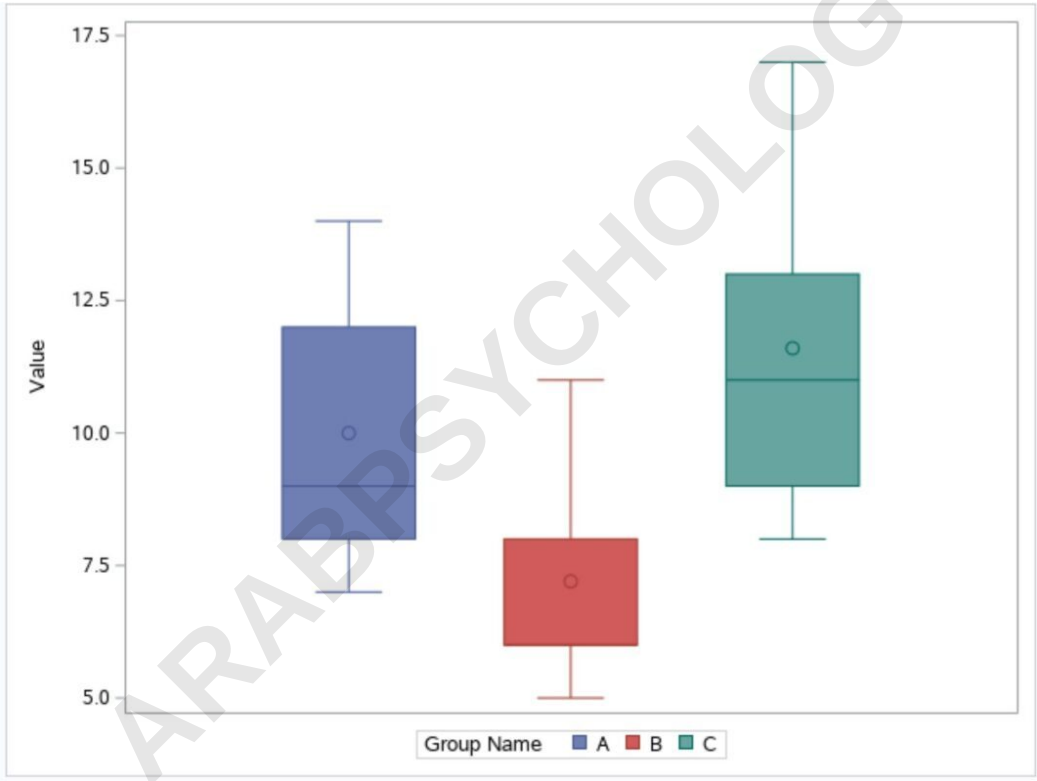
Furthermore, the `KEYLEGEND` statement is used to automatically generate a legend that maps the visual attributes (like color or pattern) assigned to each group by SAS. Specifying a meaningful title for the legend, such as "Group Name," enhances the readability and interpretability of the final

graph. The resulting visualization clearly displays the distribution of data values for groups A, B, and C, enabling immediate comparison of their medians and variability.

Next, we can use the following code to create vertical boxplots by group:

```
/*create boxplots by group*/
proc sgplot data=my_data;
vbox Value / group=Group;
keylegend / title="Group Name";
run;
```

The result is three individual boxplots that display the distribution of data values for groups A, B, and C, positioned vertically along the Y-axis:



Visual inspection of this output reveals immediate differences. For instance, Group C appears to have a higher median and potentially greater overall spread (a longer box and whiskers) compared to Group B, which exhibits a tighter distribution around a lower median value. These observations are derived directly from comparing the central lines and box heights of the plots.

Generating Horizontal Boxplots (HBOX)

While vertical boxplots (`VBOX`) are the standard visualization, horizontal boxplots (`HBOX`) are often preferred when group names are long or numerous, as they provide more horizontal space for labels. They represent the exact same statistical information, simply rotated 90 degrees. In the horizontal configuration, the measurement axis is typically the X-axis, and the categorical grouping variable defines the Y-axis positioning.

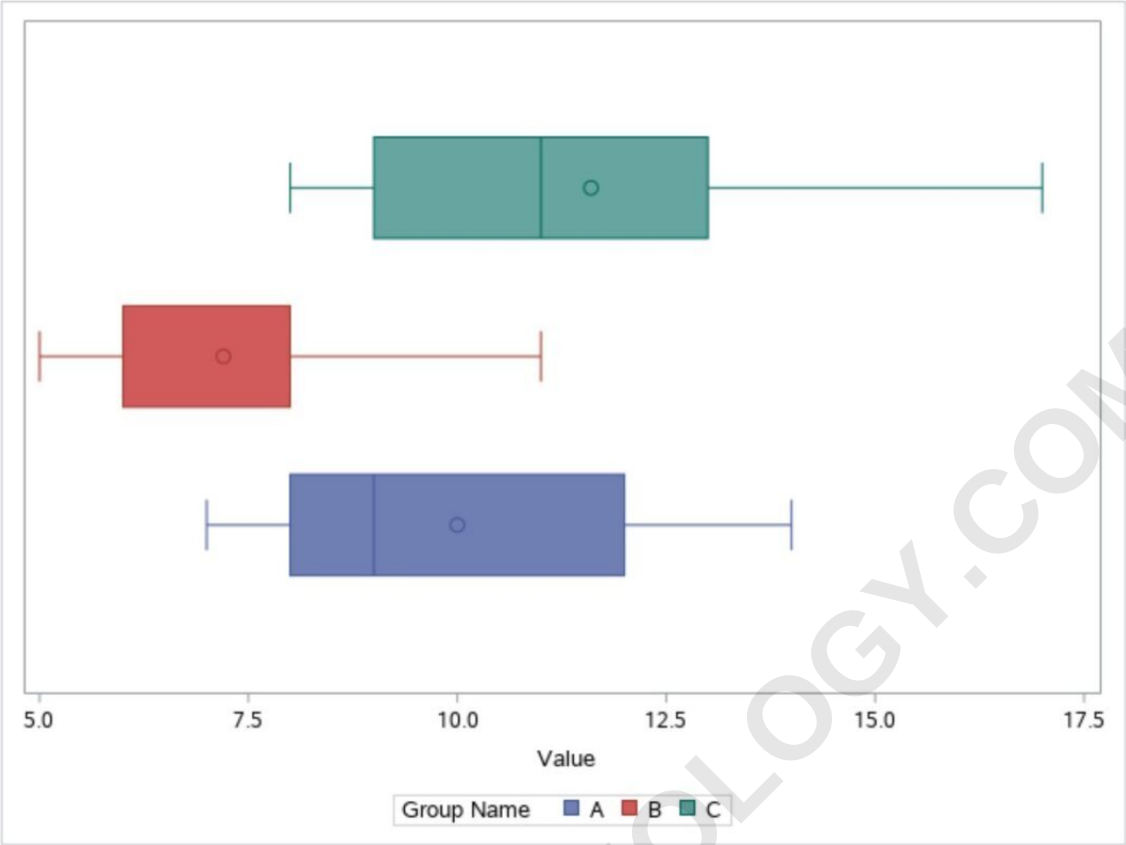
The transition from vertical to horizontal plots in SAS is straightforward, requiring only a change in the plot statement from `VBOX` to `HBOX`. All other options, including the crucial `GROUP=` option and the use of `KEYLEGEND`, remain consistent. This flexibility allows the user to choose the orientation that best serves the narrative and aesthetic requirements of the visualization without altering the core analytical structure.

When interpreting the horizontal plot, the box length along the X-axis still represents the IQR, and the vertical line indicates the median. Comparing the horizontal positioning of the median lines across groups becomes the primary method for assessing differences in central tendency. For reports or dashboards where vertical space is limited, `HBOX` provides an excellent alternative while maintaining statistical rigor.

Note that you can use the `HBOX` function to create horizontal boxplots instead, which can be beneficial when dealing with many groups:

```
/*create horizontal boxplots by group*/
proc sgplot data=my_data;
hbox Value / group=Group;
keylegend / title="Group Name";
run;
```

The result is three horizontal boxplots, where the measured value spans the horizontal axis:



The legend at the bottom of the plot shows which color corresponds to each group, aiding in the differentiation of the comparative distributions. In this horizontal view, the separation and comparison of groups A, B, and C are maintained, offering an alternative visual perspective.

Interpreting Grouped Boxplots: Key Insights

The true value of generating grouped boxplots lies in their interpretation. Analysts should focus on three primary comparative metrics: central tendency, variability, and symmetry/outliers. Central tendency is assessed by comparing the placement of the median lines (the internal lines within the boxes). If the median of one group is significantly higher or lower than another, it suggests a meaningful difference in typical performance or measurement between those groups.

Variability, or dispersion, is analyzed by comparing the length of the boxes (the IQR) and the total length spanned by the whiskers. A longer box and longer whiskers indicate higher variability, meaning the data points are more spread out for that group. Conversely, a short box suggests tightly clustered data. High variability in one group compared to another might indicate less consistency or greater risk, depending on the context of the data.

Finally, symmetry and the presence of outliers provide clues about the shape of the distribution. If the median line is perfectly centered within the box, and the whiskers are roughly equal in length,

the distribution is likely symmetric (approaching a normal distribution). If the median is closer to the bottom of the box and the upper whisker is longer, the data is positively skewed. Outliers, marked individually, signal extreme values that may warrant further investigation. The ability to spot differences in skewness and outliers across groups simultaneously makes grouped boxplots indispensable for exploratory data analysis.

Advanced Customization and Display Options

PROC SGPLOT offers substantial flexibility beyond basic plotting. While our examples use the `GROUP=` option to assign unique colors and positions to each boxplot, further customization can enhance visual clarity, especially when dealing with nested grouping structures or extremely large datasets.

One powerful option is the `GROUPDISPLAY=CLUSTER` option, which is particularly useful when you have multiple grouping variables or when you wish to emphasize the distinct identity of each group relative to others along the categorical axis. Clustering can visually separate the groups more distinctly than the default side-by-side arrangement. Other modifications involve customizing the outlier appearance, changing the fill color of the boxes using style options, or adding specific statistical overlay plots, such as mean markers, using the `MEAN` option within the `VBOX` or `HBOX` statement.

Furthermore, managing outliers is a key consideration. By default, SAS uses the standard 1.5 * IQR rule for defining outliers. However, analysts can suppress outlier plotting entirely or define custom outlier rules if necessary, depending on the statistical methodology required. Always ensure that any customization serves to make the data more accessible and accurate, rather than merely decorative. Good visualization practices dictate clear labeling, appropriate scaling, and concise interpretation.

Conclusion: Mastering Comparative Data Visualization

The creation of grouped boxplots in SAS, primarily through the use of PROC SGPLOT and the `VBOX`/`HBOX` statements with the `GROUP=` option, is a fundamental skill for any statistical analyst. This method provides an efficient and statistically rigorous way to compare the distributions of a quantitative variable across multiple categorical divisions. By visualizing the five-number summary for each subgroup, analysts can rapidly assess differences in central tendency, data variability, and distribution shape.

Whether choosing vertical or horizontal orientation, the core principle remains the same: the code must correctly identify the measurement variable and the grouping variable to yield separate, comparable plots. Mastery of this technique facilitates quicker hypothesis generation and deeper understanding during the initial phases of exploratory data analysis, making it a cornerstone of

effective reporting and statistical communication.