

Create Boxplots by Group in SAS

Authored by
Mohammed looti

November 1, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Create Boxplots by Group in SAS*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=7699>

The Essential Role of Boxplots in Exploratory Data Analysis

[Boxplots](#), also widely recognized as box-and-whisker plots, stand as fundamental instruments in the realm of exploratory data analysis (EDA). Their utility stems from their ability to provide an extraordinarily efficient graphical summary of the statistical distribution of any given dataset. By effectively distilling complex numerical distributions into a simple, standardized visual format, these plots empower data analysts to rapidly and accurately compare distribution characteristics across various categories, experimental groups, or distinct time segments. Mastering the generation and critical interpretation of these visualizations within robust statistical environments, such as [SAS](#), is absolutely crucial for generating insightful data narratives and supporting evidence-based decision-making.

The core strength of a boxplot lies in its direct visual representation of the [five-number summary](#) of the data distribution. This summary serves as a reliable, non-parametric overview, encapsulating key metrics related to central tendency, data variability, and the presence of potential outliers. When compared to relying solely on simple metrics like the mean, the boxplot offers a significantly superior method for assessing data spread and skewness, providing a comprehensive picture of the underlying data shape.

The standard boxplot visually identifies and plots the following essential statistical components, offering a clear, standardized framework for analysis:

The **Minimum Value**: Represented by the end of the lower whisker, this marks the lowest observed data point that is not considered an outlier.

The **First Quartile (Q1)**: This defines the lower boundary of the box itself and signifies the 25th percentile of the data.

The **Median**: Illustrated by the line segment dividing the box, the median represents the 50th percentile, indicating the exact center of the dataset.

The **Third Quartile (Q3)**: This forms the upper boundary of the box and corresponds to the 75th percentile of the data.

The **Maximum Value**: Located at the end of the upper whisker, this point indicates the highest observed value that does not qualify as an outlier.

Implementing Grouped Boxplots Using PROC SGPLOT

In practical statistical analysis, data is frequently segmented into natural groupings--whether comparing financial performance across different regional offices, assessing quality metrics between product lines (e.g., A, B, and C), or analyzing scores across clinical trial cohorts. In such scenarios, the creation of grouped boxplots is indispensable. This visualization technique enables a direct, side-by-side comparison of multiple data distributions within a single graph, immediately exposing differences in central location, overall spread, and distributional symmetry.

Within the modern **SAS** ecosystem, the recommended and most effective procedure for generating high-fidelity statistical graphics, including comparative boxplots, is the powerful [PROC SGPLOT](#). This procedure is an integral part of the ODS (Output Delivery System) Graphics framework, providing analysts with advanced flexibility, robust customization options, and superior output quality that is perfectly suited for formal publications, academic research, and executive reports.

The following methodology meticulously details the steps required to leverage the capabilities of PROC SGPLOT. We will focus specifically on how to command the procedure to create customized, grouped boxplots, thereby allowing for the effective simultaneous visualization of the distribution characteristics of several independent categorical groups on a unified plot area.

Structured Data Preparation for Comparative Visualization

Before any visualization can be initiated, it is paramount to ensure that the source data is correctly structured within a **SAS** dataset. The specific requirement for grouped boxplots is that the dataset must minimally contain two critical variables: a **quantitative variable** (the continuous metric being measured, referred to here as `Value`) and a **categorical variable** (the grouping factor used for segmentation, referred to as `Group`).

We begin this process by defining and populating a straightforward example dataset, which we name `my_data`. This illustrative dataset is constructed to contain a series of measurements across three distinct, predefined categories: Group A, Group B, and Group C. The categorical variable `Group` is explicitly defined using the dollar sign (\$) notation, confirming its status as a character variable in **SAS** programming.

```
/* Define and populate the dataset, establishing three distinct categorical groups (A, B, C) */  
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 preliminary step successfully establishes the necessary foundation for analysis. By creating these three distinct sets of measurements, linked by the categorical variable `Group`, we have prepared the data structure necessary to execute comparative statistical analysis and generate the subsequent multi-group visualization.

Executing the Code for Vertical Grouped Boxplots

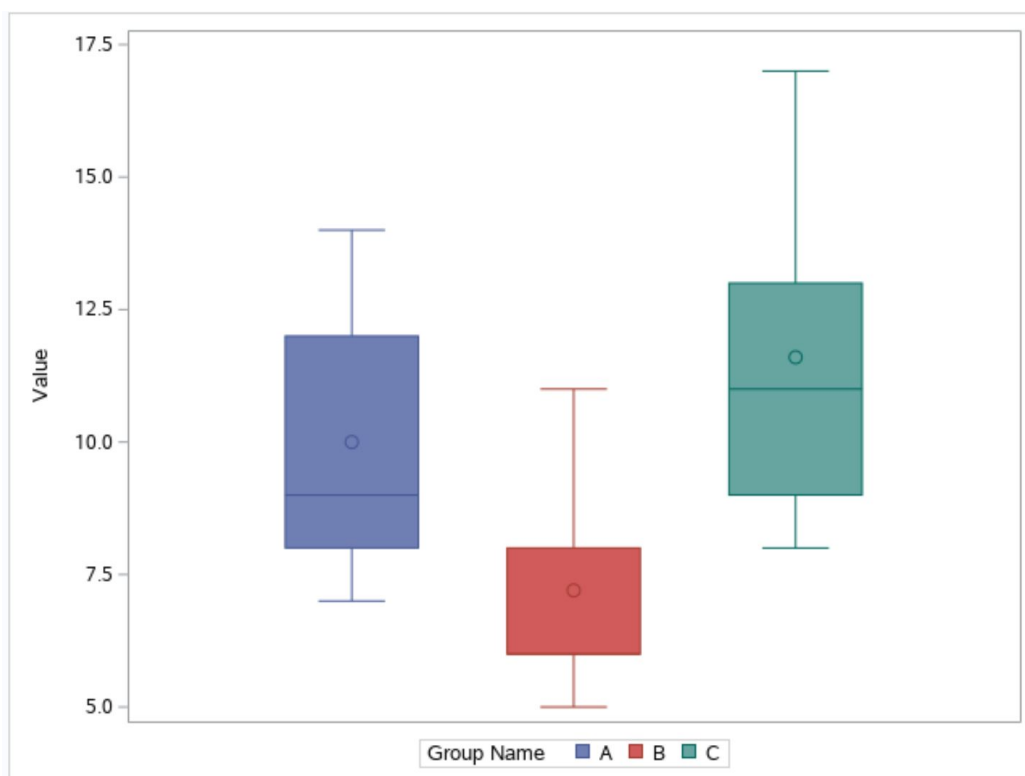
To generate the primary visualization--a series of standard vertical boxplots illustrating the distribution of the quantitative `Value` variable segmented by the categorical `Group`--we activate the `PROC SGLOT` procedure and integrate the specific `VBOX` statement. The `VBOX` (Vertical Box Plot) statement is specifically engineered within **SAS** to handle this type of quantitative distribution visualization effectively.

The most critical element within the code structure below is the inclusion of the `/ group=Group` option, which is appended directly to the `VBOX` statement. This parameter acts as an explicit instruction to **SAS**, directing the system to partition the input data based on the unique values discovered within the `Group` variable (A, B, and C). Consequently, the procedure plots a separate, distinct boxplot for each identified segment. Furthermore, we employ the `KEYLEGEND` statement to enhance the clarity of the output, ensuring a descriptive title is applied to the legend that clearly maps the color coding to the corresponding group names.

```
/* Generate vertical boxplots, segmenting the visualization by the Group variable using  
VBOX */  
proc sgplot data=my_data;  
vbox Value / group=Group;  
keylegend / title="Group Name";  
run;
```

The successful execution of this code yields three individual vertical boxplots, all superimposed on a unified quantitative axis. This configuration facilitates the immediate visual assessment of distribution differences across the three categories. The resulting graphic clearly demonstrates the variance in the [quartiles](#) and the [median](#) for each category, offering rapid insights into their

respective central tendencies and data spreads:



Interpreting Comparative Boxplot Output

The fundamental advantage of utilizing grouped boxplots lies in their capacity to enable the simultaneous comparison of several key statistical measures. When reviewing the generated visualization, analysts should strategically concentrate their interpretation on three primary comparative metrics: central tendency, variability, and the symmetry of the distribution.

First, attention should be directed toward comparing the [median](#) line--the horizontal line situated inside the box--across all displayed groups. For example, in the generated plot above, Group C exhibits the highest median value. This finding signifies that 50% of the observations within Group C fall above a higher measurement point than those in Group A or Group B, strongly suggesting a clear upward shift in the central location of the data specific to Group C.

Second, a detailed analysis of variability, or spread, is required. Variability is quantified by observing both the length of the box (which represents the Interquartile Range, or IQR) and the span of the whiskers. A box that appears significantly longer, coupled with extended whiskers, is indicative of greater dispersion and lower consistency within the dataset. As observed, Group C possesses the largest spread, demonstrating less consistency in its data values compared to Group B, which shows a noticeably tighter clustering of its observations around the center.

Third, the relative placement of the median line within the box itself offers crucial insight into distributional symmetry. If the median is positioned closer to the lower boundary (Q1), the distribution is typically characterized as being skewed right (positive skew). Conversely, if the line is closer to the upper boundary (Q3), the distribution is skewed left (negative skew). Discrepancies in the length of the upper and lower whiskers also serve as additional visual indicators of asymmetry. These visual cues are invaluable for developing a deeper understanding of the underlying shape of the data distribution for each individual group.

Creating Horizontal Boxplots for Enhanced Readability

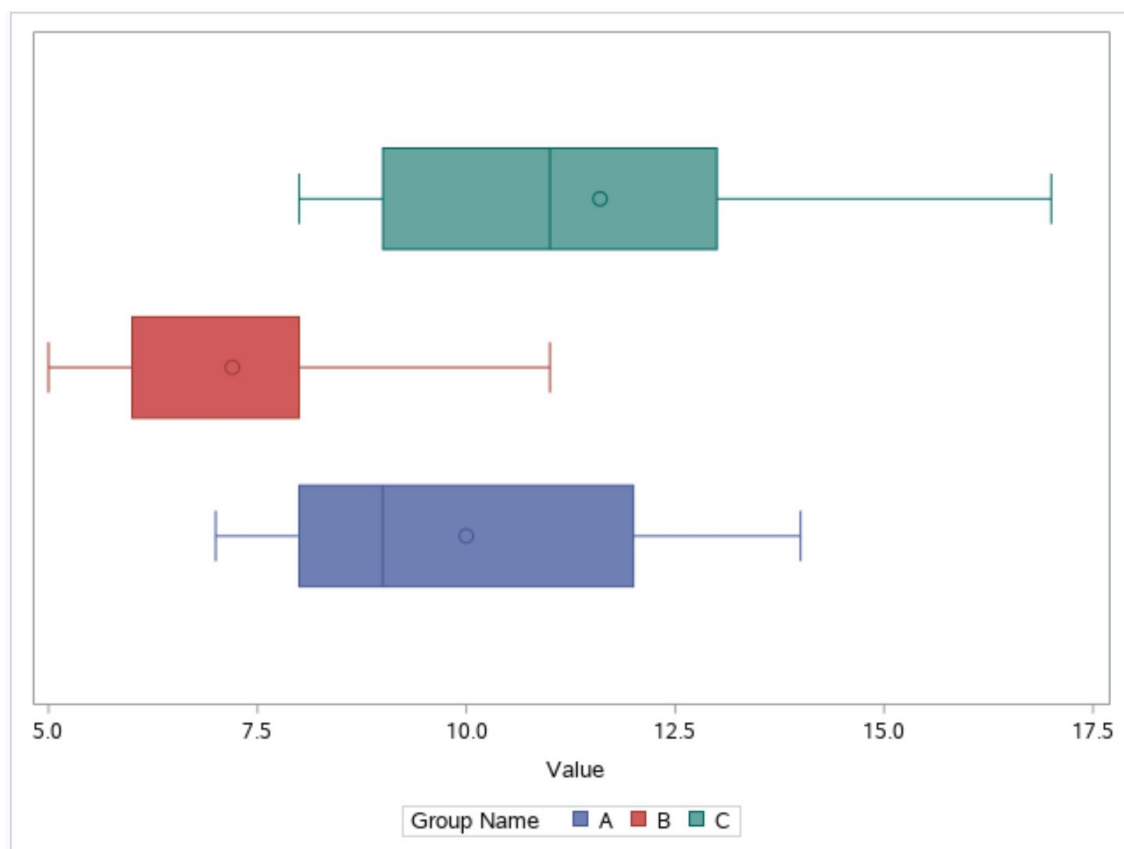
While the vertical orientation is the default standard for boxplots, there are practical analytical scenarios where a horizontal presentation is significantly preferred. This is particularly true when the analysis involves a large number of distinct groups, or when the categorical labels are extensive and thus require more horizontal space for clear labeling. **SAS** seamlessly facilitates this alternative preference through the use of the `HBOX` statement, which directly replaces `VBOX` within the code structure.

The modification required to transition from a vertical to a horizontal plot involves only a minimal change to the established [PROC SGPLOT](#) code. Specifically, we substitute the `VBOX` function with `HBOX` (Horizontal Box Plot), while diligently retaining the essential `/ group=Group` option to ensure the segmentation logic remains intact and correctly applied.

```
/* Generate horizontal boxplots, segmenting the visualization by the Group variable using HBOX */  
proc sgplot data=my_data;  
hbox Value / group=Group;  
keylegend / title="Group Name";  
run;
```

This command generates an equivalent visualization that preserves the complete statistical integrity of the vertical plot but alters the orientation of the axes. The quantitative `Value` axis is now displayed horizontally, while the categorical `Group` labels are arranged vertically.

This horizontal alternative often dramatically improves readability, especially in complex dashboards, by simplifying the visual comparison of the [five-number summary](#) across numerous categories that are listed along the y-axis.



Crucially, in both the vertical and horizontal graphical presentations, the legend automatically produced by the KEYLEGEND statement at the bottom of the plot remains vital. It furnishes a crystal-clear mapping, unequivocally demonstrating which color corresponds to each of the defined categorical groups (A, B, and C), maintaining consistency and interpretability across formats.

Advanced Customization and Further Resources

Achieving proficiency in generating grouped visualizations within **SAS** represents a critical milestone toward producing advanced statistical reports and high-impact analytical outputs. The fundamental techniques demonstrated using VBOX and HBOX statements serve as a versatile foundation. To further refine your graphical output and enhance your analytical depth, it is recommended to explore the advanced capabilities of the ODS Graphics system.

The following supplementary resources provide detailed tutorials and authoritative documentation aimed at deepening your understanding of comparative [boxplots](#) and the extensive features available within the PROC SGPLOT procedure:

Performing detailed statistical analysis of potential outliers and customizing the definition of whisker boundaries in boxplots.

Applying specific customization options for colors, titles, axis scales, and formatting within the

flexible PROC SGLOT procedure environment.

Exploring techniques for integrating and combining boxplots with other complementary graph types, such as scatter plots or histograms, using the comprehensive ODS graphics framework for layered analysis.