Learning to Create Scatter Plot Matrices in SAS: A Step-by-Step Guide

Authored by

Mohammed looti

September 1, 2026

RECOMMENDED CITATION

Mohammed looti (2026). *Learning to Create Scatter Plot Matrices in SAS: A Step-by-Step Guide*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=4185

Introduction to Multivariate Data Visualization

A **scatter plot matrix** represents a highly sophisticated and efficient method of visualizing relationships across numerous **variables** simultaneously. Essentially, it is a grid-like arrangement that contains every possible pairwise **scatter plot** derived from a given **dataset**. This powerful form of **data visualization** is absolutely indispensable for statisticians, machine learning engineers, and **data analysts** who need to quickly grasp the complex, multivariate structure of their data. By presenting these relationships side-by-side, the scatter plot matrix ensures that potential correlations, subtle trends, and conspicuous outliers are immediately apparent, preventing the oversight of critical relational patterns.

The core utility of employing a scatter plot matrix lies in its ability to transform a multi-dimensional problem space into an easily interpretable two-dimensional format. While each cell in the matrix isolates and displays the interaction between just two specific variables, the collective presentation provides a comprehensive, holistic view of the data structure. Furthermore, the diagonal elements of the matrix can often be customized to display univariate statistics, such as histograms or kernel density estimates, offering crucial context regarding the distribution of each individual variable. This comprehensive approach is foundational to exploratory **data analysis**, allowing analysts to confirm or deny assumptions about the data structure before proceeding to more rigorous statistical modeling.

Within the environment of **SAS**, the generation of these intricate matrices is streamlined through specialized procedures designed specifically for high-quality graphical output. SAS provides a robust and reliable framework for creating not only basic scatter plot matrices but also highly customized versions tailored to complex analytical needs. This guide is dedicated to outlining the essential steps required to master the creation and customization of these critical data visualizations using the powerful tools available in SAS.

Introducing the SAS PROC SGSCATTER Procedure

The standard tool for generating scatter plot matrices within the SAS environment is the **PROC SGSCATTER** procedure. This procedure is an integral component of the **SAS/GRAPH** module, which is globally recognized for producing publication-quality statistical graphics. **PROC SGSCATTER** is specifically engineered to handle the visualization of relationships among multiple quantitative variables efficiently, enabling the construction of detailed matrices with minimal programming effort.

The fundamental syntax required to invoke **PROC SGSCATTER** is remarkably concise yet extraordinarily potent. At its simplest, the procedure requires the user to specify the input **dataset** and list the particular **variables** that should be included in the matrix visualization. The procedure

automatically manages the complex tasks of arranging the individual plots, ensuring proper scaling across all axes, and displaying them within a cohesive grid structure. Furthermore, SAS provides numerous options for extensive customization, allowing users to modify the aesthetic appearance, incorporate descriptive titles and footnotes, or even group the data points based on a third, categorical variable to reveal deeper, segmented insights.

Proficiency in utilizing this procedure is considered a fundamental requirement for anyone engaging in serious, advanced **data analysis** using **SAS**. It establishes a reliable and fast framework for the essential initial step of assessing multi-variable interactions, which is vital for validating hypotheses and informing the selection of appropriate statistical models. The following sections will provide a detailed walkthrough of its application, commencing with the necessary steps to prepare the data and define the basic matrix structure.

Step-by-Step: Defining Data and Basic Syntax

Before any graphical output can be generated, we must first ensure a suitable sample **dataset** exists within the **SAS** session. For demonstration purposes, we will construct a small dataset that simulates performance metrics for different sports teams, incorporating key numerical **variables** such as points scored, assists made, and rebounds collected, alongside a categorical variable identifying the team. This setup provides an ideal scenario for exploring multivariate relationships and differences across groups.

The **SAS code** below illustrates the process of creating a temporary dataset named `my_data`. We utilize the `data` step combined with the `input` statement to define the variable names and types, followed by the `datalines` statement, which embeds the raw observational data directly into the program. This methodology is frequently employed by analysts to rapidly create or replicate small, specific datasets for illustrative or testing purposes within a larger SAS project.

```
/*create dataset*/
data my_data;
input team $ points assists rebounds;
datalines;
A 22 12 8
A 20 18 4
A 14 9 5
A 30 16 10
B 10 4 3
B 9 5 12
B 6 5 14
B 14 10 5
```

```
C 4 8 12
C 13 10 5
C 11 12 8
C 19 3 2
;
run;

/*view dataset*/
proc print data=my_data;
```

Following the creation and population of the dataset, the subsequent `proc print` statement serves as a vital verification step, displaying the contents of `my_data` in a clear, tabular format. This allows the user to confirm that all data points were correctly entered and that the variables--`team`, `points`, `assists`, and `rebounds`--are structured exactly as required for the forthcoming visualization tasks. This verification is essential for ensuring the integrity of the subsequent statistical graphics.

Obs	team	points	assists	rebounds
1	A	22	12	8
2	A	20	18	4
3	A	14	9	5
4	A	30	16	10
5	B	10	4	3
6	B	9	5	12
7	B	6	5	14
8	B	14	10	5
9	C	4	8	12
10	C	13	10	5
11	C	11	12	8
12	C	19	3	2

Executing the Basic Scatter Plot Matrix

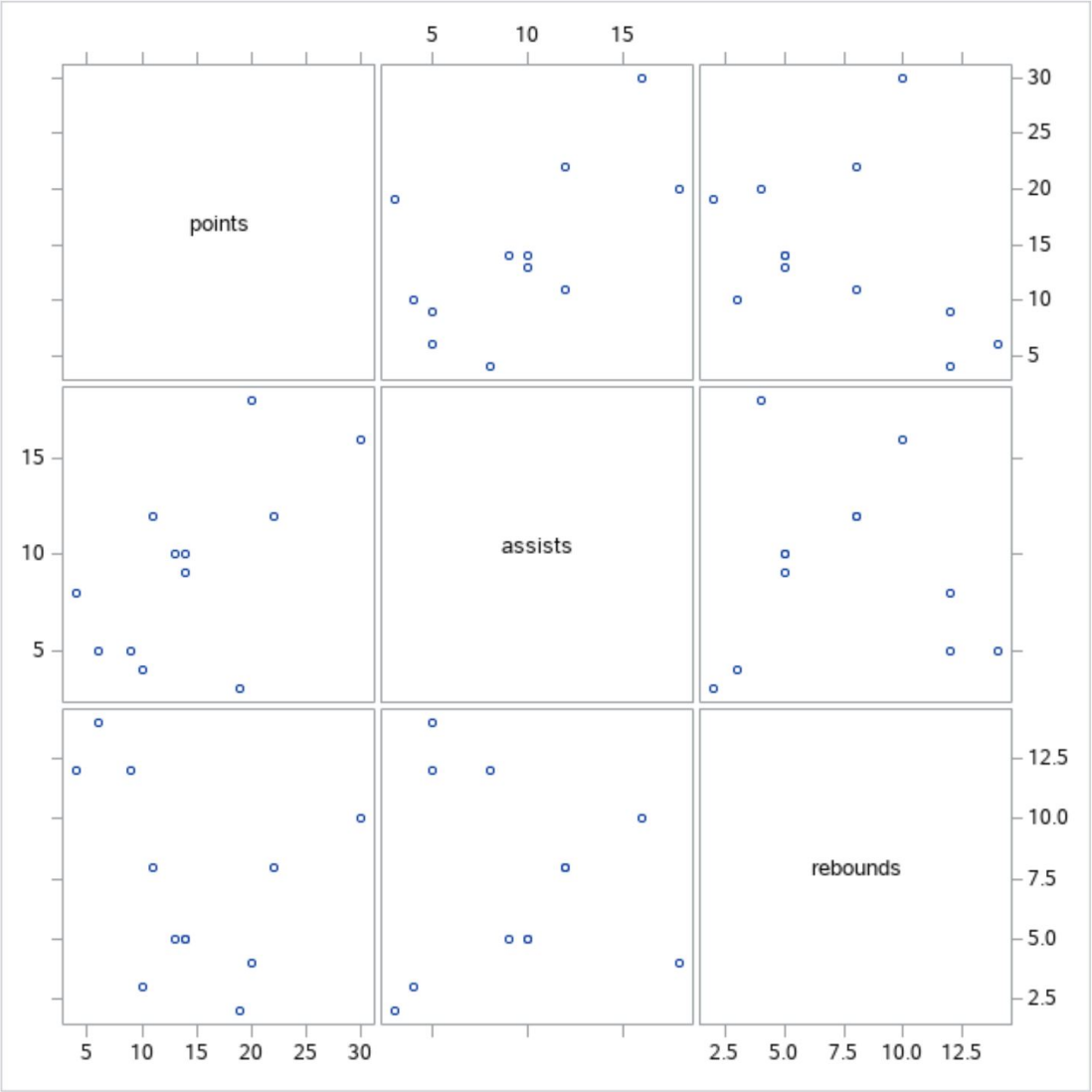
With the sample data successfully prepared, the next logical step is to generate the foundational **scatter plot matrix** using the core **PROC SGSCATTER** procedure. This initial visualization is designed to immediately offer a visual assessment of the pairwise relationships existing among the quantitative variables: `points`, `assists`, and `rebounds`. The resulting matrix serves as the

primary output for our exploratory **data analysis**.

The syntax employed below is a direct and simple application of the procedure, specifically targeting our `my_data` dataset and the three key variables we wish to analyze. The procedure is instructed to produce a 3x3 matrix. In this standard configuration, the off-diagonal cells are populated with scatter plots showing the relationship between two distinct variables, while the diagonal cells are often left blank or, depending on default settings or specified options, may contain summary graphics such as histograms to show the univariate distribution of the variable corresponding to that row/column.

```
/*create scatter plot matrix*/
proc sgscatter data=my_data;
matrix points assists rebounds;
run;
```

Executing this straightforward code generates the scatter plot matrix illustrated below. The interpretation of this graphic is critical: each plot provides a visual depiction of variable interaction. For example, by examining the plot located at the intersection of the 'points' row and the 'assists' column, an analyst can quickly determine if there is a positive correlation (as one increases, the other increases), a negative correlation, or no discernible relationship. This raw visualization capability is fundamental for discerning underlying data patterns, identifying linear or non-linear trends, and pinpointing any unusual data points that may warrant further investigation before proceeding to more complex statistical modeling.



Advanced Customization: Grouping and Titles

While the basic **scatter plot matrix** provides a strong foundation for exploratory analysis, its analytical power and clarity can be substantially improved through strategic customization. **PROC SGSCATTER** facilitates these enhancements, most notably through the addition of descriptive titles and, more critically, the ability to segment or group data points based on a **categorical variable**. This grouping feature enables a sophisticated comparative analysis within a single, unified visualization.

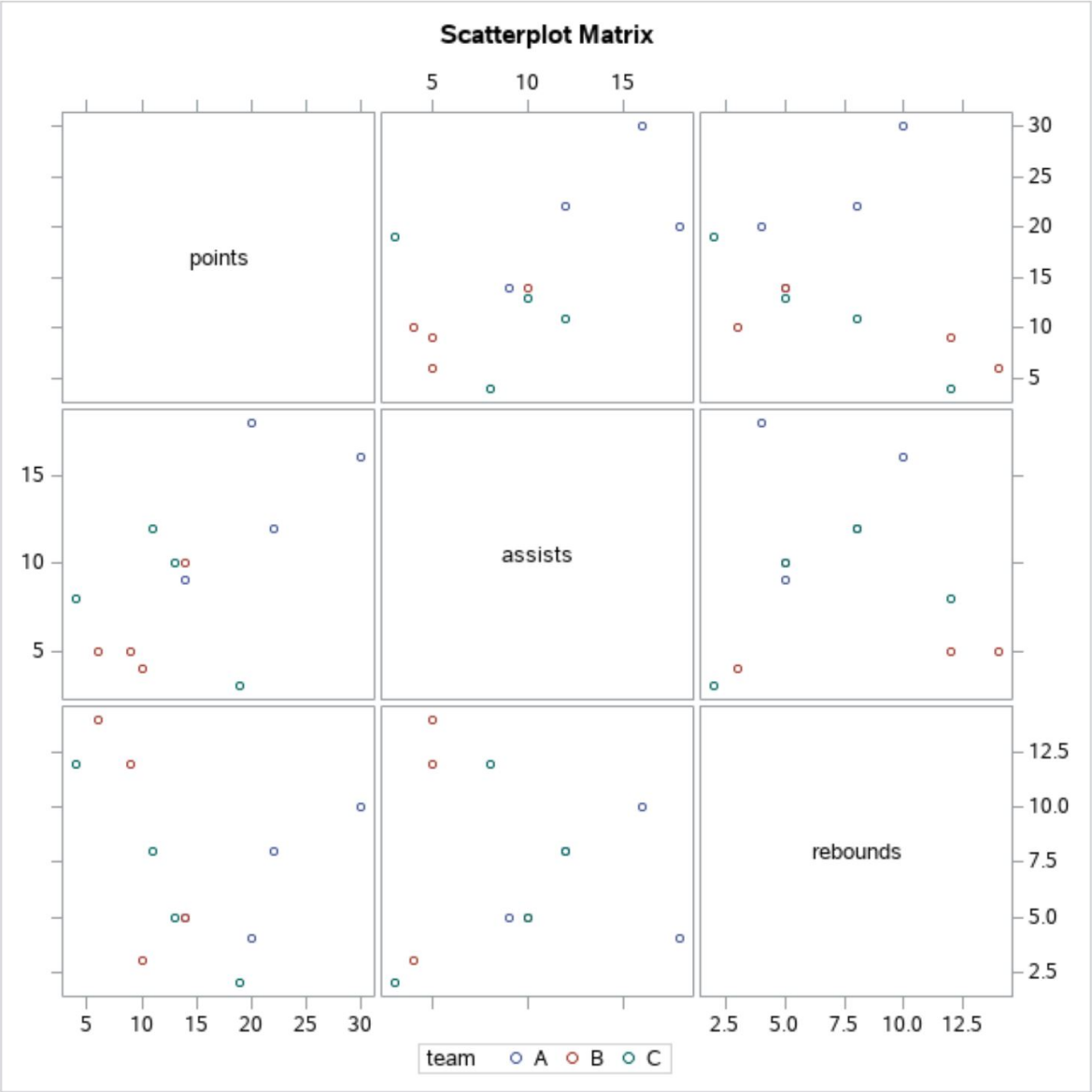
Implementing a clear title significantly boosts the professionalism and immediate interpretability of the graphic, succinctly communicating the content to any viewer. The standard `title` statement within **SAS** handles this requirement. Of greater analytical importance is the `group` option, which enables the visual differentiation of observations based on a categorical attribute--in our case, the `team` identifier. By assigning distinct colors or symbols to the data points belonging to each team,

the resulting plots reveal whether the relationships between numerical variables, such as points and assists, are consistent or vary meaningfully across the defined categories. This comparative insight is often essential for revealing nuanced interactions that would be obscured in an undifferentiated plot.

The following **SAS** code demonstrates how to apply these valuable enhancements. We first define a title, "Scatterplot Matrix: Team Performance Metrics," for the entire output. Subsequently, we incorporate the crucial `group=team` option directly within the `matrix` statement. This instruction commands **PROC SGSCATTER** to utilize the values of the `team` variable to color-code the points in every individual scatter plot. This technique provides a powerful layer of comparative **data visualization**, significantly enriching the analytical depth of the matrix.

```
/*create scatter plot matrix with points colored by team*/
proc sgscatter data=my_data;
title "Scatterplot Matrix: Team Performance Metrics";
matrix points assists rebounds / group=team;
run;
title;
```

The resulting graphic clearly displays the defined title at the top, ensuring immediate context. More importantly, the color-coding reveals distinct clusters of data points corresponding to each team. This allows the analyst to visually compare, for instance, the distribution and correlation strength of `points` and `assists` for Team A versus Team C. Such detailed segmentation is invaluable for targeted **data analysis**, supporting decisions based on inter-group differences and enhancing the effective communication of analytical findings.



Conclusion and Further Exploration

The creation of a **scatter plot matrix** within **SAS**, primarily facilitated by the **PROC SGSCATTER** procedure, stands as one of the most effective and essential techniques available for visualizing the complex relationships among numerous **variables** contained within a single **dataset**. This methodology provides a comprehensive, at-a-glance overview, drastically simplifying the identification of trends, the magnitude of correlations, and the precise location of potential outliers that might confound traditional statistical models. By mastering both the basic syntax and leveraging powerful customization options such as descriptive titles and insightful data grouping, analysts can produce highly informative and visually compelling graphics that significantly elevate the quality of their **data analysis** and subsequent reporting.

The examples detailed throughout this guide successfully illustrate the simplicity and significant analytical power inherent in **PROC SGSCATTER**, transitioning from the generation of a fundamental matrix to the inclusion of crucial layers of comparative information via color-coding by **categorical variables**. It must be emphasized that these visualizations are not merely aesthetic additions; they are indispensable tools for conducting thorough exploratory data analysis, constructing new hypotheses, and, most importantly, verifying the foundational assumptions required for applying more sophisticated inferential statistical modeling techniques.

For those dedicated to further advancing their **data visualization** capabilities in **SAS**, the **PROC SGSCATTER** procedure documentation reveals a vast array of additional options. These advanced features include the flexibility to specify different types of plots for the critical diagonal cells (such as histograms or box plots), granular control over axis ranges and labeling, the ability to incorporate reference lines for benchmarking, and comprehensive control over the overall aesthetic design of the output. Exploring the official **SAS documentation** is strongly advised to unlock the full potential of this powerful procedure and customize scatter plot matrices perfectly to meet specific, detailed analytical requirements.

Additional Resources for SAS Graphics

To ensure continued learning and to facilitate the implementation of advanced applications of scatter plot matrices within the **SAS** environment, analysts are encouraged to consult the following authoritative and instructional resources:

The Official **SAS Documentation** for PROC SGSCATTER, which provides the most comprehensive and up-to-date guide covering all available syntax options, examples, and technical details.
Specialized academic textbooks focusing on statistical graphics and advanced **data analysis**, which often contain detailed theoretical and practical sections on effective multivariate visualization techniques.
Active online forums and specialized communities dedicated to **SAS** programming, serving as excellent platforms to find solutions to complex technical challenges and to learn best practices directly from experienced users globally.
Structured tutorials and professional workshops focused on **data visualization** using the SAS system, offering essential hands-on experience and practical operational tips.