How to create and interpret box plots in SPSS?

Authored by

stats writer

December 26, 2025

The **Box Plot**, often referred to as a box-and-whisker plot, is a fundamental graphical tool in descriptive statistics used for visualizing the distribution of numerical data. It provides a quick and highly effective way to display the summary of a dataset based on its **five number summary**: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. This plot is essential for identifying central tendency, spread, and the presence of **outliers** within the data structure.

This comprehensive guide details the precise procedures required to generate and customize high-quality box plots using the **Statistical Package for the Social Sciences (SPSS)**, a robust software suite widely utilized across academic and professional statistical analysis. We will walk through both the creation of single variable plots and the comparison of multiple datasets using side-by-side visualizations in **SPSS**. To successfully interpret a box plot, analysts must understand that the central line represents the **median**, the box's length signifies the Interquartile Range (IQR) or data spread, and the whiskers denote the range, with points beyond them indicating potential outliers.

A **Box Plot** is specifically designed to visualize the five core metrics of a dataset's distribution, providing crucial insight into symmetry and variability. These essential metrics, known collectively as the five number summary, are:

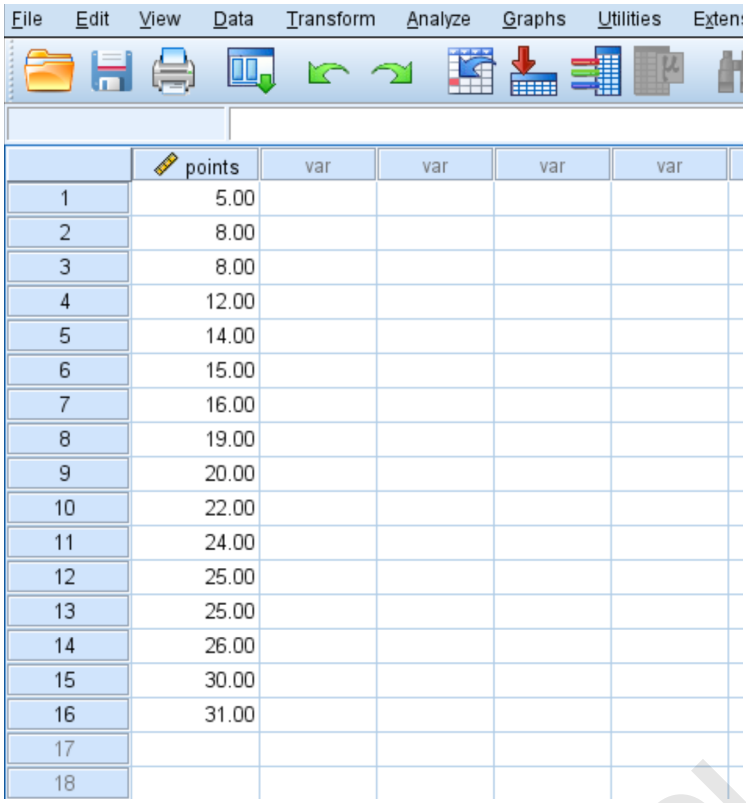
The **Minimum** value (the end of the lower whisker, excluding outliers).
The **First Quartile (Q1)**, marking the 25th percentile.
The **Median (Q2)**, which is the 50th percentile and the line inside the box.
The **Third Quartile (Q3)**, marking the 75th percentile.
The **Maximum** value (the end of the upper whisker, excluding outliers).

This guide provides a detailed, step-by-step approach on how to effectively create and interpret these powerful visualizations within the **SPSS** environment.

Preparing Your Data for Single Box Plot Analysis

Before generating any visualization in statistical software, ensuring your data is correctly formatted is paramount. For this demonstration, we will use a hypothetical dataset illustrating the average points scored per game by 16 basketball players on a specific team. This numerical variable, labeled 'points', is ideally suited for a box plot analysis because the plot requires continuous or ordinal numerical data to accurately calculate the necessary quartiles and extreme values.

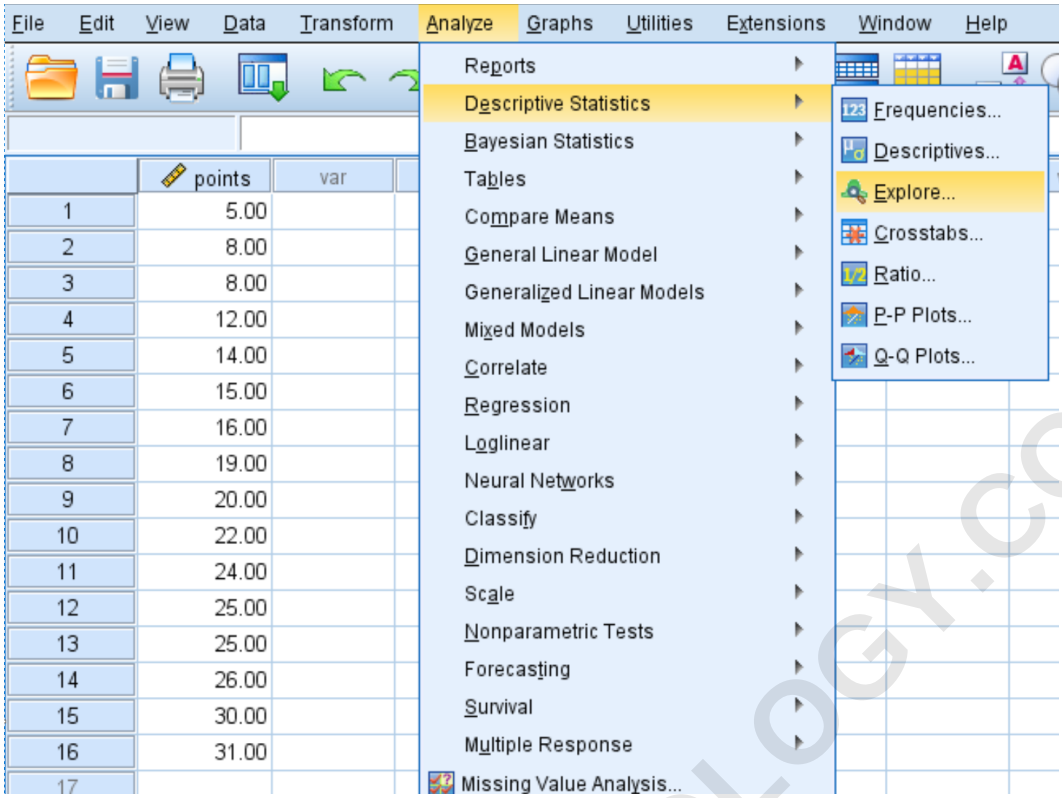
The initial dataset is presented below, showing the individual scores associated with each player observation. Understanding the raw data structure is the critical first step before launching the analytical tools in **SPSS**. We utilize this simple structure to clearly demonstrate the transformation from raw data inputs to graphical output.



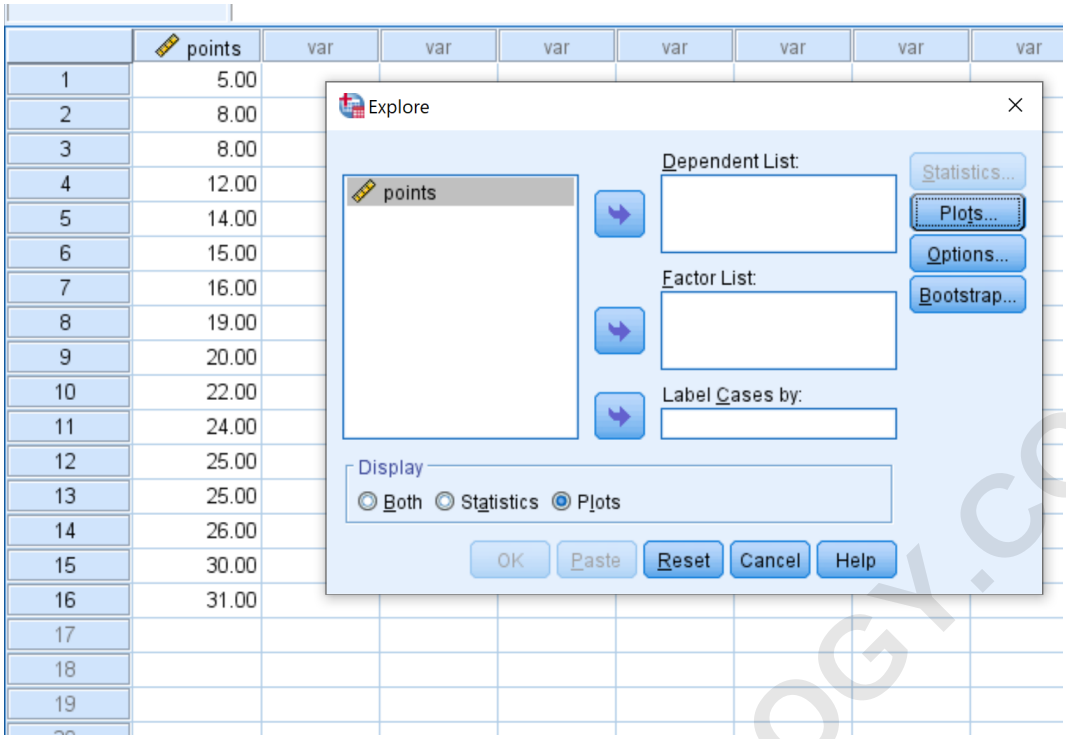
Creating a Single Box Plot Using the Explore Function

To create a box plot for visualizing the distributional characteristics of a single variable, the most reliable and statistically informative method in SPSS is often through the **Explore** dialogue box. This function is specifically optimized for descriptive analysis and diagnostics, automatically performing the calculations necessary for the five-number summary.

Initiate the process by navigating the menu system: click the **Analyze** tab, hover over **Descriptive Statistics**, and then select the **Explore** option. This sequence opens the primary interface for detailed exploration of data distributions, which allows for simultaneous calculation of descriptive statistics and graphical generation.

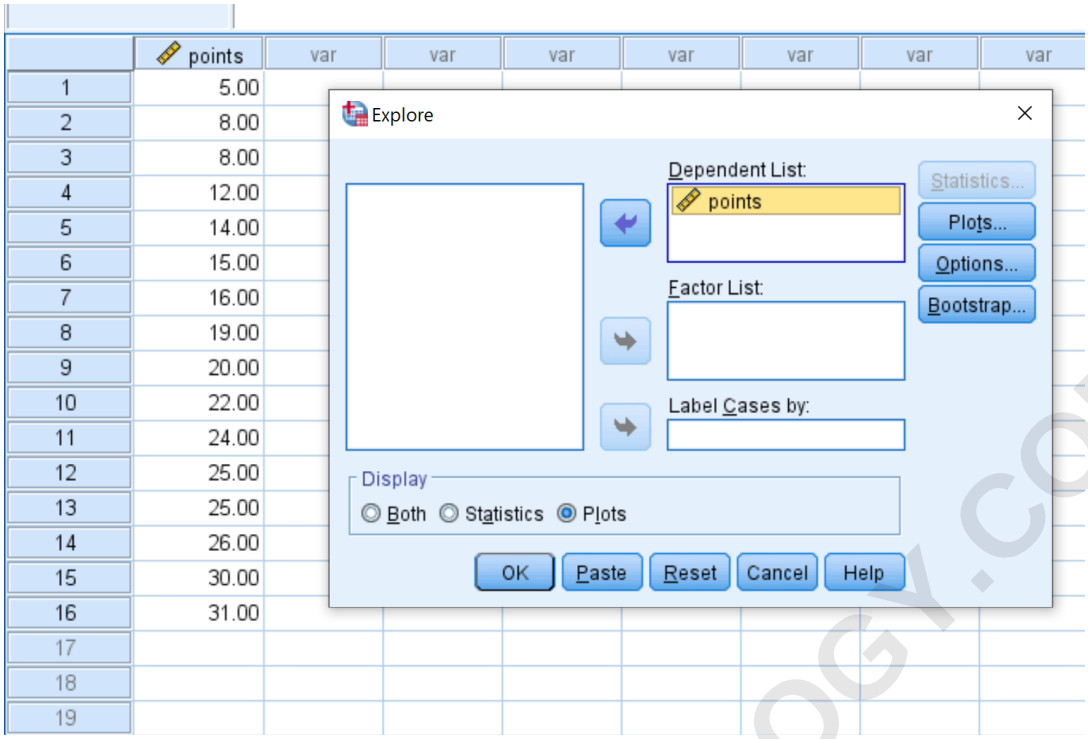


Upon selecting **Explore**, a new dialogue window will appear. This window prompts the user to specify the variables intended for analysis and the desired output format, offering choices between statistical tables, graphical plots, or both. This structured approach ensures that the output is tailored precisely to the research question.



Configuring and Generating the Box Plot Output

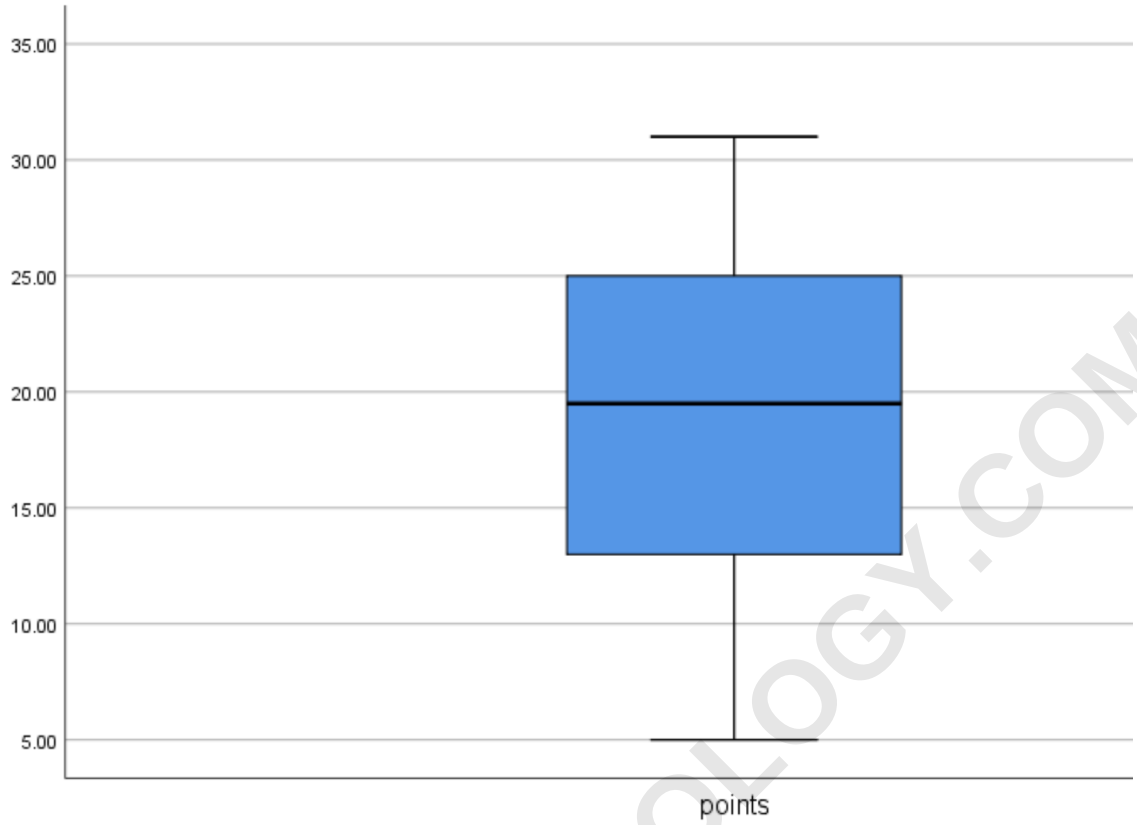
To properly generate the visualization, you must correctly assign the target variable. Drag the variable named **points** from the list on the left into the input area labelled **Dependent List**. This action designates 'points' as the primary variable whose distribution will be analyzed and plotted. It is crucial to ensure that the output option under **Display**, located near the bottom of the window, is set to **Plots** (or optionally, both Statistics and Plots). This setting explicitly instructs SPSS to generate the graphical output, specifically the box plot, alongside any corresponding numerical summaries.



After these configuration steps are finalized, clicking **OK** executes the analysis command. SPSS processes the raw data, calculates the quartiles and extreme values, and outputs the generated box plot visualization into the Output Viewer window, making it immediately accessible for detailed interpretation and reporting.

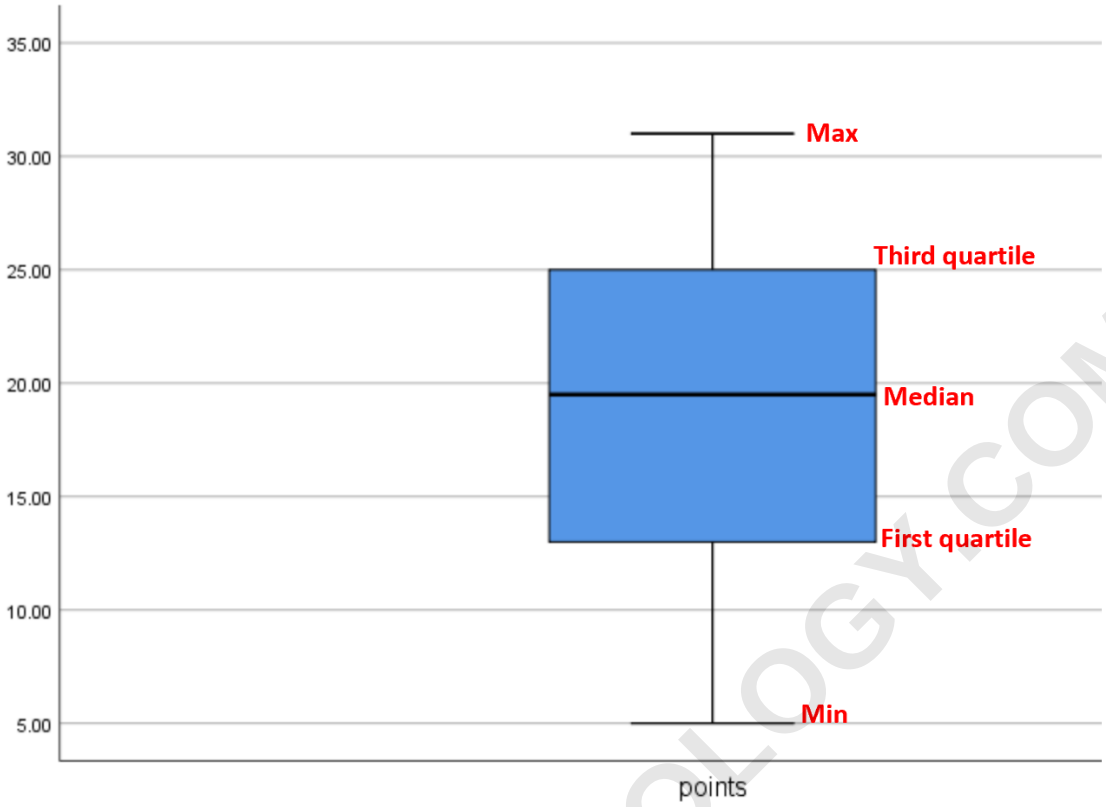
Interpreting the Single Variable Box Plot

The resulting box plot provides a comprehensive visual summary of the data's **data distribution**, enabling analysts to quickly assess its symmetry, spread, and the location of its **central tendency**. The visualization generated from the 'points' data example is displayed below, serving as the basis for our interpretive steps.



To accurately interpret this graph, focus on the three primary components. First, the central box spans the Interquartile Range (IQR), representing the middle 50% of the data. The length of this box is a direct indicator of the variability or spread of the central data points. A shorter box suggests less variability in the core dataset, while a longer box suggests greater data dispersion. Second, the horizontal line inside the box represents the **median** (Q2). The position of this line relative to the Q1 and Q3 edges of the box reveals the skewness of the distribution.

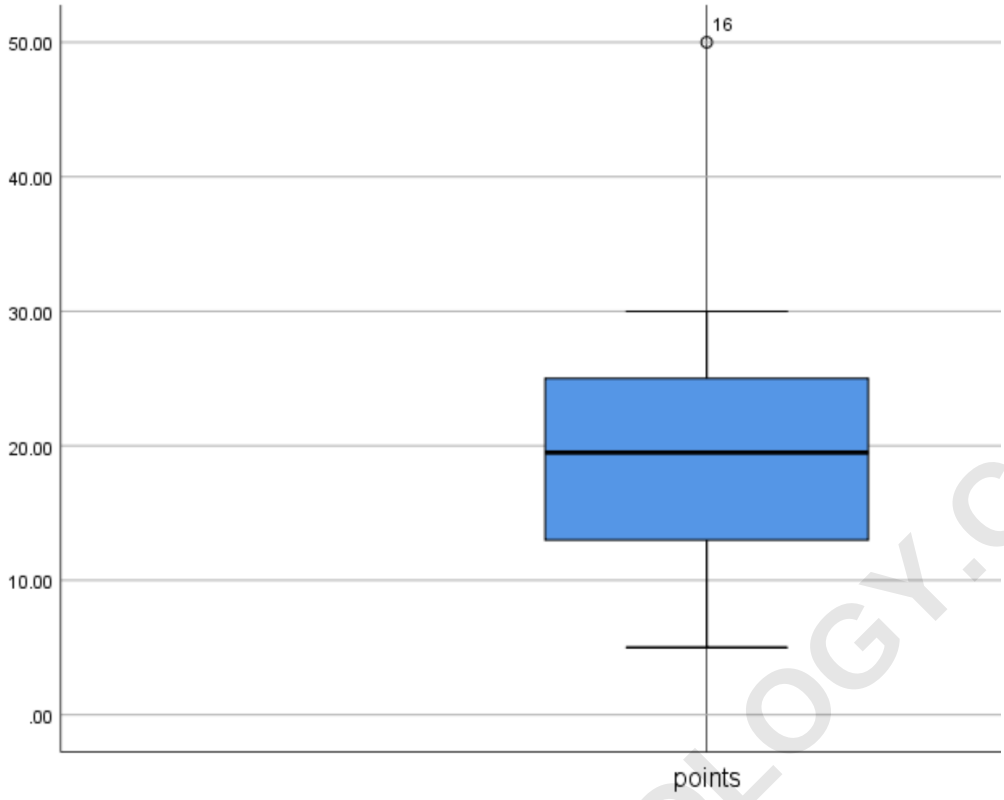
Finally, the whiskers extend from the box to the minimum and maximum values that are not statistically classified as **outliers** (typically based on 1.5 times the IQR). These whiskers define the overall range of the non-extreme data points. Any data points plotted individually outside these whisker limits are automatically identified by SPSS as potential outliers, demanding further investigation.



Methodological Considerations for Handling Outliers

A major advantage of using **box plots** is their robust ability to clearly flag data points that deviate significantly from the rest of the distribution, which are known as **outliers**. In SPSS, these extreme values are typically displayed as small circles (indicating mild outliers) or sometimes asterisks (indicating extreme outliers), situated beyond the calculated limits of the whiskers. In our previous example, no data points fell outside the standard range, explaining the absence of these outlier symbols.

To demonstrate outlier detection, let us consider a modified scenario where the largest value in our dataset was 50. This score significantly exceeds the distribution's normal spread. In this instance, the box plot would immediately highlight this deviation, rendering the upper whisker shorter and displaying the score of 50 as a distinct circle, clearly identifying it as an outlier:



When an outlier is detected, researchers must apply careful methodological judgment regarding its treatment, as these extreme values can disproportionately influence parametric statistical summaries, such as the mean. The chosen approach should always be thoroughly documented to ensure analytical transparency and rigor.

Ensure Data Accuracy: The foundational step involves confirming that the outlier is not the result of a simple **data entry error**. If a transcription or input mistake is identified, the value must be corrected immediately. If the value is verified as correctly entered, further diagnostic steps are required.

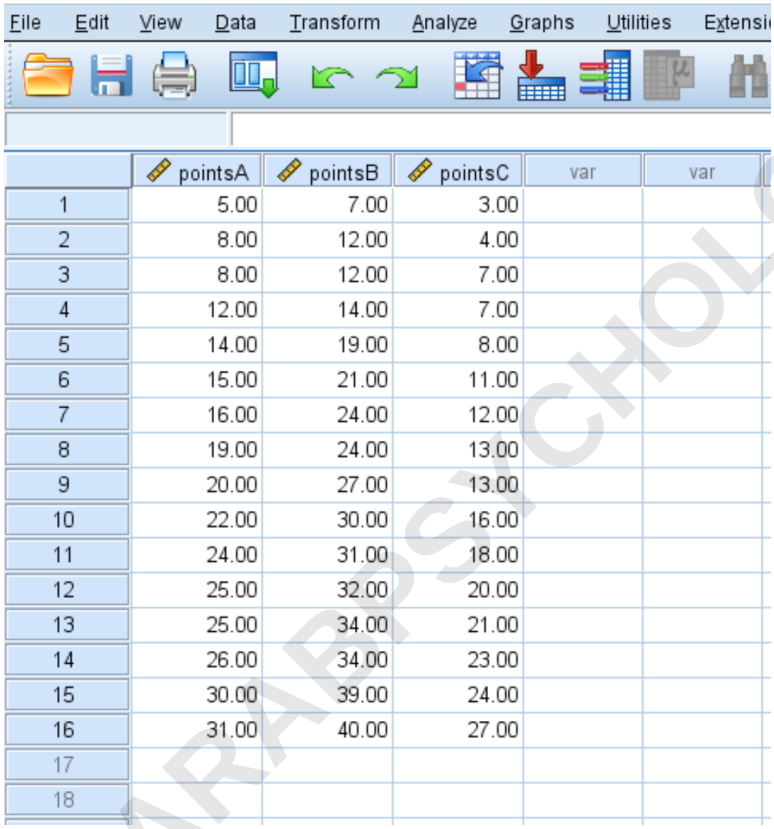
Assign a Replacement Value: If the outlier is confirmed to be an error but the true value cannot be retrieved, or if the value is deemed influential but not representative, an analyst may choose to assign a substituted value. This might involve using a robust measure of **central tendency**, such as the **median** of the dataset, provided this imputation method is statistically appropriate.

Exclusion from Analysis: If the extreme value is confirmed as a true, non-error-related occurrence that severely distorts the intended statistical analysis or violates assumptions, it may be selectively removed from the dataset. This removal must be explicitly justified and mentioned in the final research documentation to maintain the integrity of the findings.

Creating Multiple Side-by-Side Box Plots for Comparison

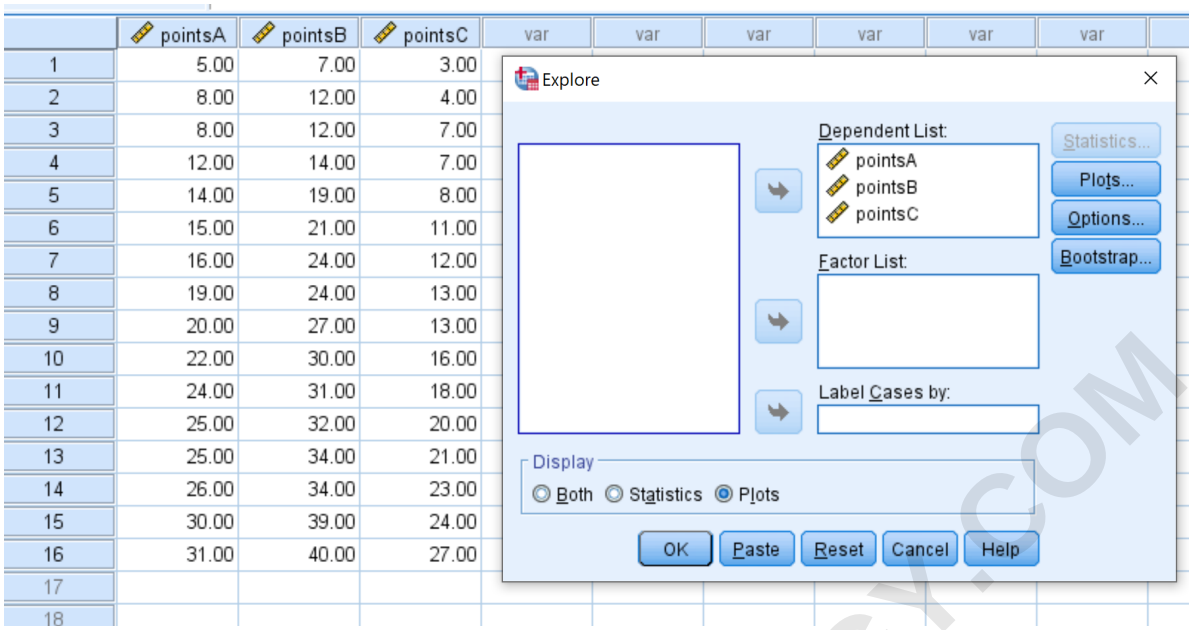
Box plots are exceptionally potent tools for comparative analysis, offering a direct visual assessment of distribution differences across multiple groups or variables. SPSS simplifies this process, allowing for the concurrent creation of several box plots displayed side-by-side, which is highly valuable when comparing performance metrics across different categories or populations.

To illustrate this comparative functionality, we expand our dataset to track the average points scored by players across three distinct organizational entities: Team A, Team B, and Team C. This structure necessitates a side-by-side plot visualization for effective assessment of differences in spread, skewness, and **central tendency** among the three teams.



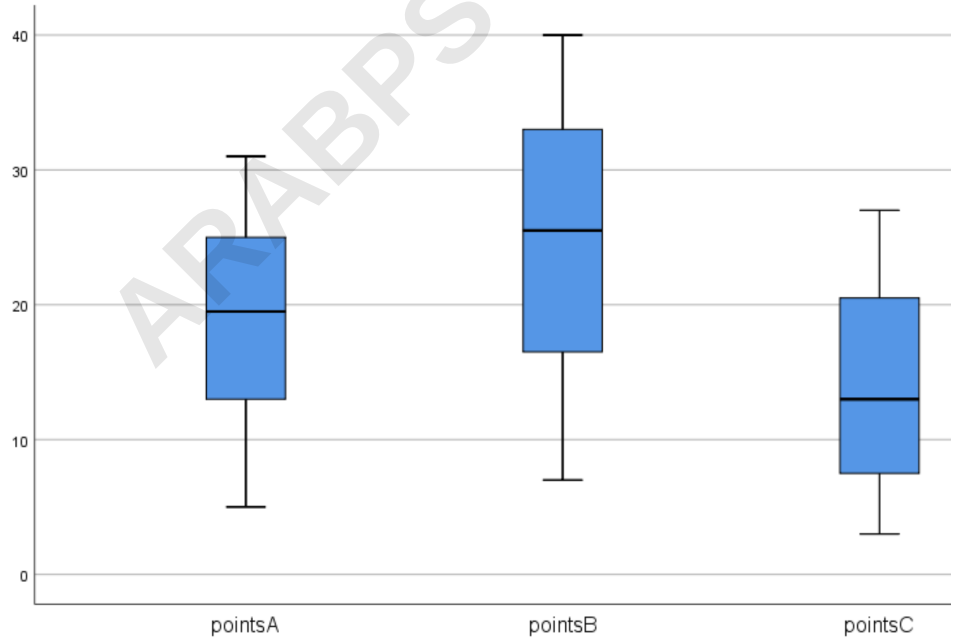
	pointsA	pointsB	pointsC	var	var
1	5.00	7.00	3.00		
2	8.00	12.00	4.00		
3	8.00	12.00	7.00		
4	12.00	14.00	7.00		
5	14.00	19.00	8.00		
6	15.00	21.00	11.00		
7	16.00	24.00	12.00		
8	19.00	24.00	13.00		
9	20.00	27.00	13.00		
10	22.00	30.00	16.00		
11	24.00	31.00	18.00		
12	25.00	32.00	20.00		
13	25.00	34.00	21.00		
14	26.00	34.00	23.00		
15	30.00	39.00	24.00		
16	31.00	40.00	27.00		
17					
18					

The procedural steps for generating multiple plots mirror those for a single plot. Access the analytical dialogue box by selecting **Analyze**, then **Descriptive Statistics**, and finally **Explore**. In the subsequent configuration window, drag all three relevant variables (Team A, Team B, Team C) simultaneously into the **Dependent List** box. This prepares SPSS to analyze and plot the distributions of all specified variables in a unified output structure.



Comparative Interpretation of Multiple Box Plots

Executing the analysis by clicking **OK** yields the following set of comparative box plots. These visualizations allow for a rapid and insightful comparison of how performance metrics vary among the three teams based on their respective five number summaries and overall **data distribution** characteristics.



From this combined visualization, several critical observations about the relative distributions of

points scored for each team can be immediately derived:

Central Performance: The position of the **Median** line indicates that Team B demonstrates the highest central performance level, whereas Team C records the lowest **median** score among the three groups.
Variation and Dispersion: The variance in player scores, quantified by the length of the central box (the IQR), is clearly highest for Team B. This extended box suggests significantly greater internal performance disparity or spread within Team B compared to the more tightly grouped scores of Team A and Team C.
Identification of Extremes: The plot shows that the player achieving the highest recorded points per game belongs to Team B, reinforcing its high-variability profile across the entire range. Conversely, the player with the lowest score is found on Team C, indicating a lower overall range for that team.

In summary, the side-by-side **box plots** effectively demonstrate the efficiency of this statistical tool in providing comprehensive information regarding the distribution, skewness, variability, and extreme values of multiple datasets through a single, comparative graphical representation.