Learning to Create and Interpret Box Plots Using SPSS

Authored by

Mohammed looti

November 8, 2025

Understanding the Fundamentals of the Box Plot

A Box plot, also known as a box-and-whisker plot, is a powerful graphical tool in SPSS used to visualize the distribution and spread of a continuous dataset. It offers a concise, standardized way of displaying the distribution of data based on the five-number summary, making it invaluable for initial data exploration and comparison across different groups. This visual representation quickly highlights central tendency, variability, and potential asymmetry or skewness within the data.

The core strength of the box plot lies in its ability to summarize large amounts of data into five critical points. These statistical markers provide a comprehensive overview of where the data is concentrated and the range over which it spreads. By relying on these robust measures, the box plot is less sensitive to extreme values compared to histograms, offering a clearer picture of the dataset's structure.

The structure of the box plot is built upon the following components, collectively known as the **five-number summary**:

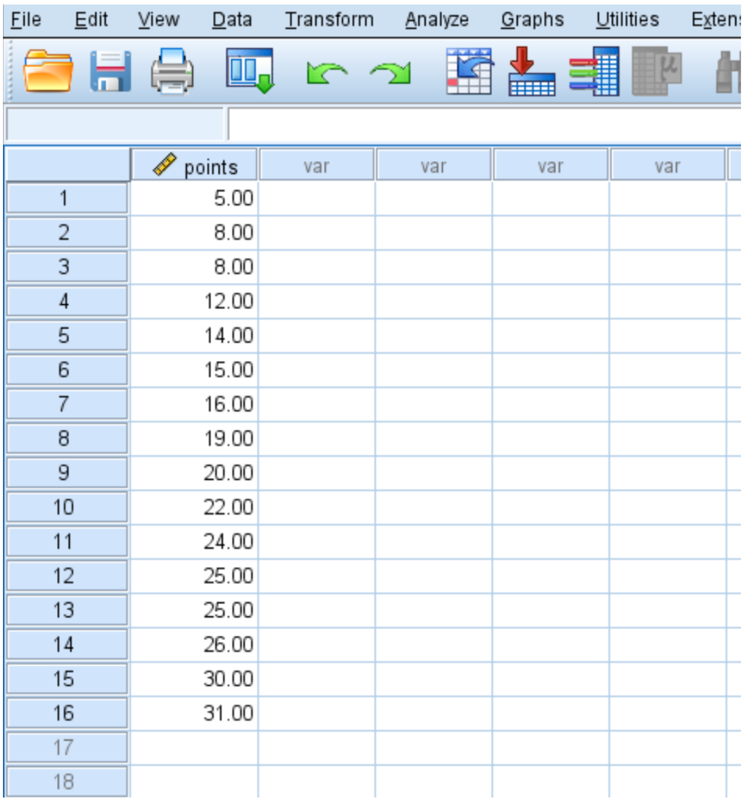
The **Minimum**: The smallest value in the dataset, excluding any defined outliers.
The **First Quartile (Q1)**: The 25th percentile, marking the boundary of the lowest 25% of the data.
The **Median (Q2)**: The 50th percentile, which is the exact middle value of the dataset. This represents the central tendency.
The **Third Quartile (Q3)**: The 75th percentile, marking the boundary of the upper 25% of the data.
The **Maximum**: The largest value in the dataset, excluding any defined outliers.

This tutorial will guide you through the process of generating, customizing, and accurately interpreting box plots using the robust statistical capabilities of SPSS (Statistical Package for the Social Sciences).

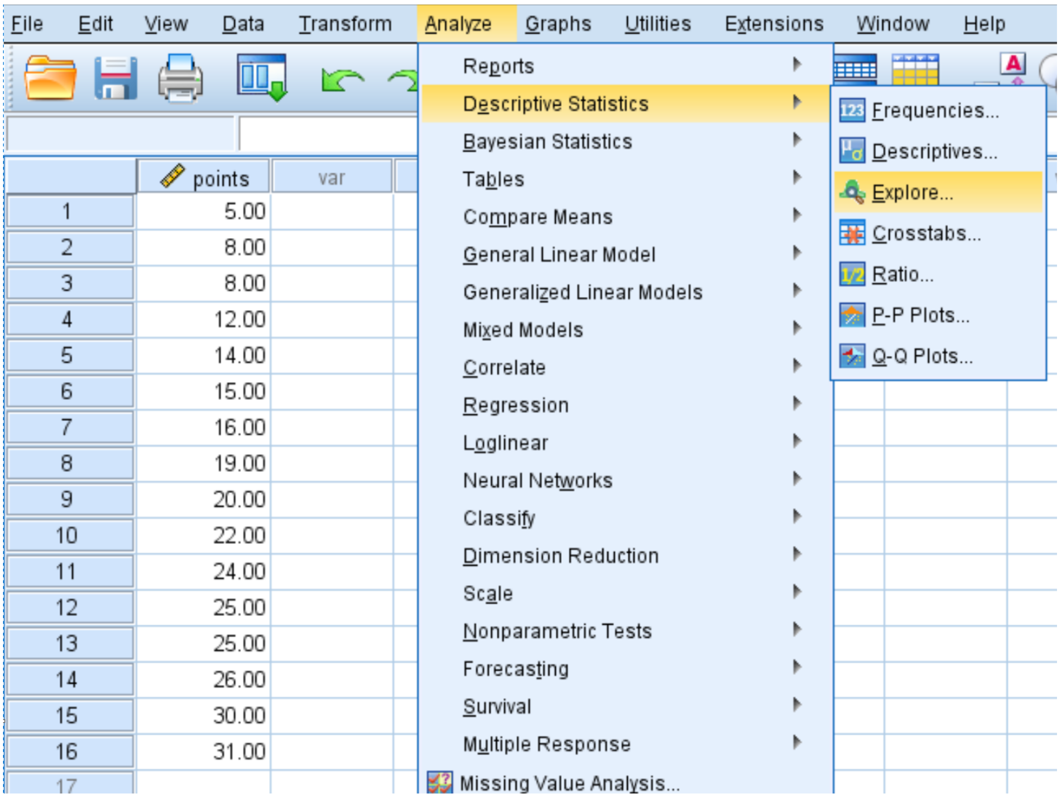
Step-by-Step: Generating a Single Box Plot in SPSS

To demonstrate the creation of a single box plot, let us consider a sample dataset tracking a continuous variable. Suppose we are analyzing the average points scored per game by 16 distinct basketball players on a particular team. This variable, which we will label "points," represents the distribution we wish to visualize.

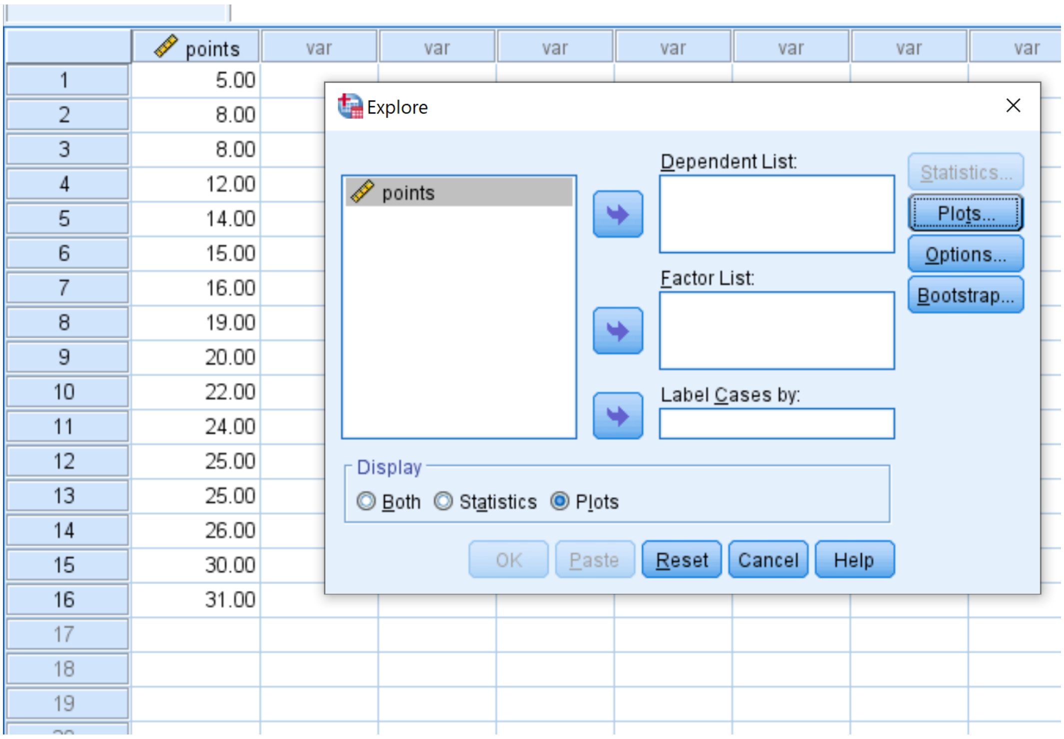
Before proceeding with the graphical analysis, ensure your data is correctly entered into the SPSS Data View. The dataset used for this example is illustrated below, showing the variable **points** and its associated values:



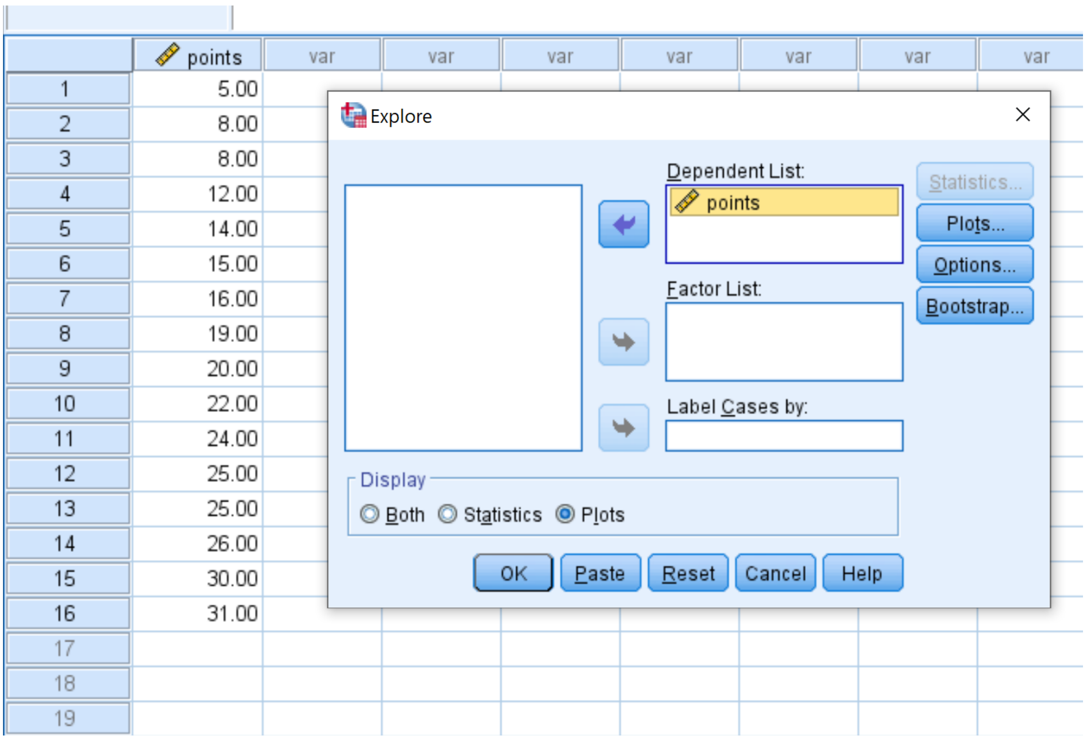
To initiate the box plot creation process, we utilize the **Explore** function, which is specifically designed for detailed descriptive statistical analysis and graphical data distribution summaries. Navigate through the SPSS menu bar using the following sequence: click the **Analyze** tab, hover over **Descriptive Statistics**, and then select **Explore**. This sequence will open the primary dialogue box necessary for defining the visualization parameters.



In the subsequent **Explore** dialogue window, you must specify the variable you intend to analyze. Drag the variable **points** from the variable list on the left and place it into the box labeled **Dependent List**. The Dependent List holds the variables for which you want to calculate descriptive statistics and generate plots. Crucially, verify that the **Display** option, located near the bottom of the dialogue box, is set to **Plots** (or **Both**, to include statistical tables as well). Selecting "Plots" ensures that the graphical output, including the box plot, is generated upon execution.



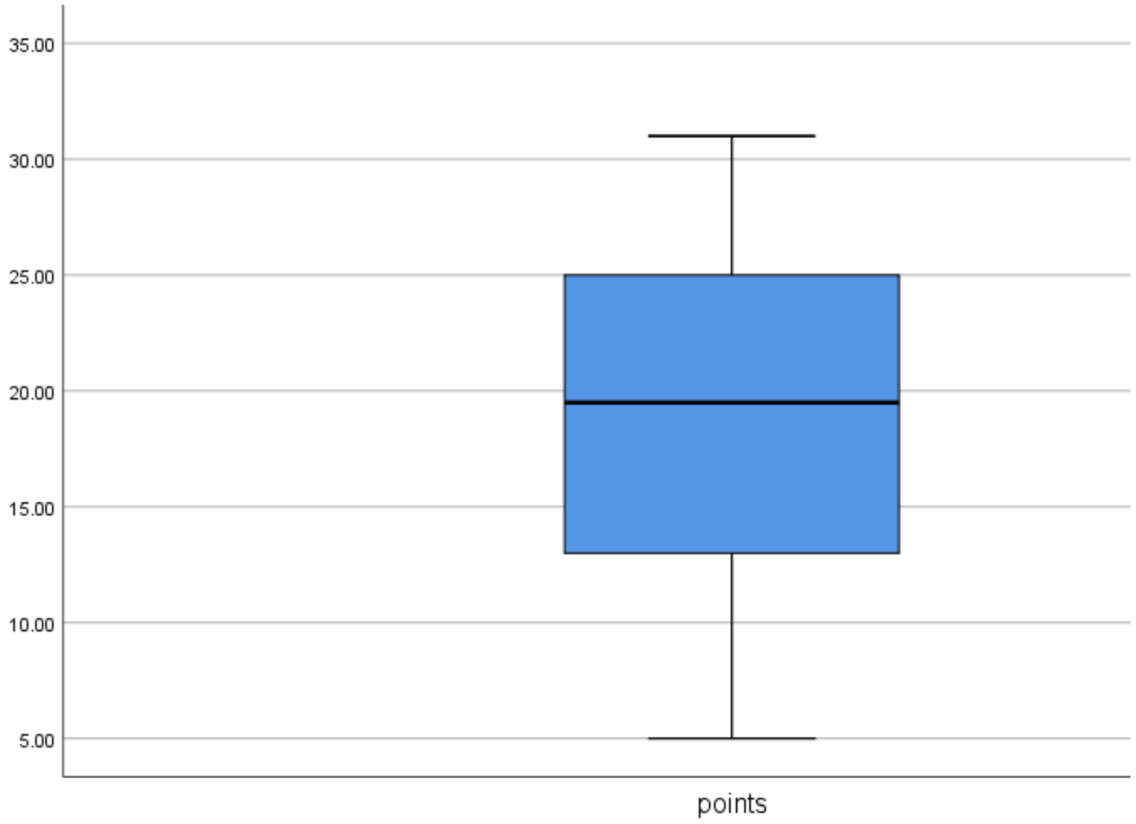
After configuring the Dependent List and ensuring the Plot option is selected, click **OK**. SPSS will execute the command and display the resulting box plot in the Output Viewer, visually summarizing the distribution of the average points scored by the basketball players.



Interpreting the Visual Components of Your Box Plot

Once generated, the box plot provides immediate insights into the data's central tendency, spread, and shape. Analyzing the position and length of the box and whiskers allows researchers to quickly assess the distribution characteristics of the variable.

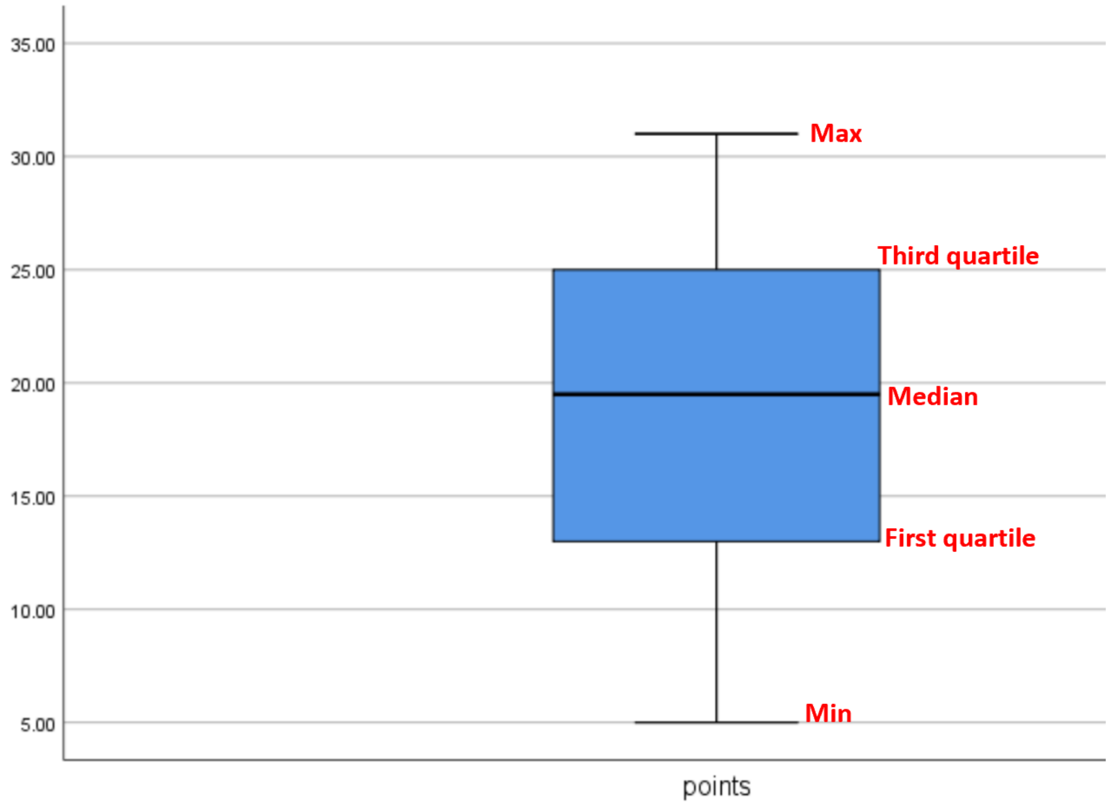
The resulting box plot for the sample data looks like this:



The main rectangular box in the center of the plot represents the **Interquartile Range (IQR)**, which contains the middle 50% of the data. The length of this box indicates the spread or variability of the central data. A longer box suggests greater variability in the middle range of scores. The horizontal line that bisects the box marks the **median** (Q2). The placement of this line within the box helps determine the symmetry of the central 50% of the data; if the line is near the center, the data is likely symmetric around the median. If it is closer to the bottom or top edge, the data is skewed.

The lines extending vertically from the box are called **whiskers**. These whiskers extend to the minimum and maximum values that are not considered outliers. Typically, SPSS defines the whiskers as extending to 1.5 times the IQR from the edges of the box (Q1 and Q3). Any data points falling outside of these whiskers are individually plotted as potential outliers.

Here is a labeled diagram illustrating the key elements and their corresponding statistical measures within the SPSS output:

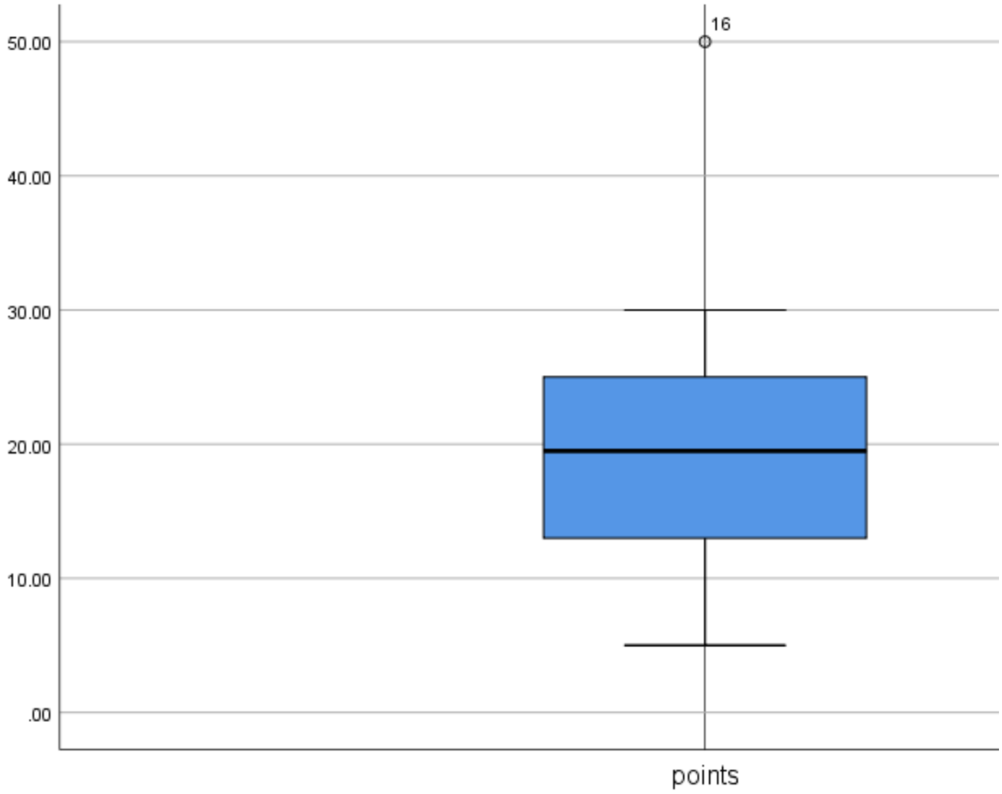


Interpreting the vertical axis (the scale of the variable) in conjunction with these visual markers allows for precise statistical analysis. For instance, the bottom of the box aligns with Q1, indicating that 25% of the basketball players scored below that value. The top of the box aligns with Q3, meaning 75% of players scored below that value. This visual summary is significantly more informative than simply calculating the mean and standard deviation alone, especially when data distribution is non-normal.

Identifying and Managing Outliers in SPSS Box Plots

One of the most valuable features of the box plot is its immediate identification of potential **Outliers**--data points that deviate significantly from other observations. In SPSS, outliers are graphically represented as tiny circles or asterisks plotted beyond the whiskers. In the previous example, since the data distribution was relatively tight, no tiny circles appeared, indicating the absence of extreme values based on the IQR definition.

Consider a scenario where the largest value in our dataset was 50 points instead of the original maximum. Since 50 is far beyond the typical range defined by the whiskers (1.5 * IQR), SPSS automatically flags it as an outlier:



The presence of outliers necessitates careful consideration by the researcher, as they can disproportionately influence statistical models, particularly those based on the mean. When an outlier is detected, researchers typically have several critical options for managing the anomalous data point, depending on the context and origin of the value.

If an outlier is present in your dataset, you must systematically evaluate its impact and origin. The standard procedures for outlier management include:

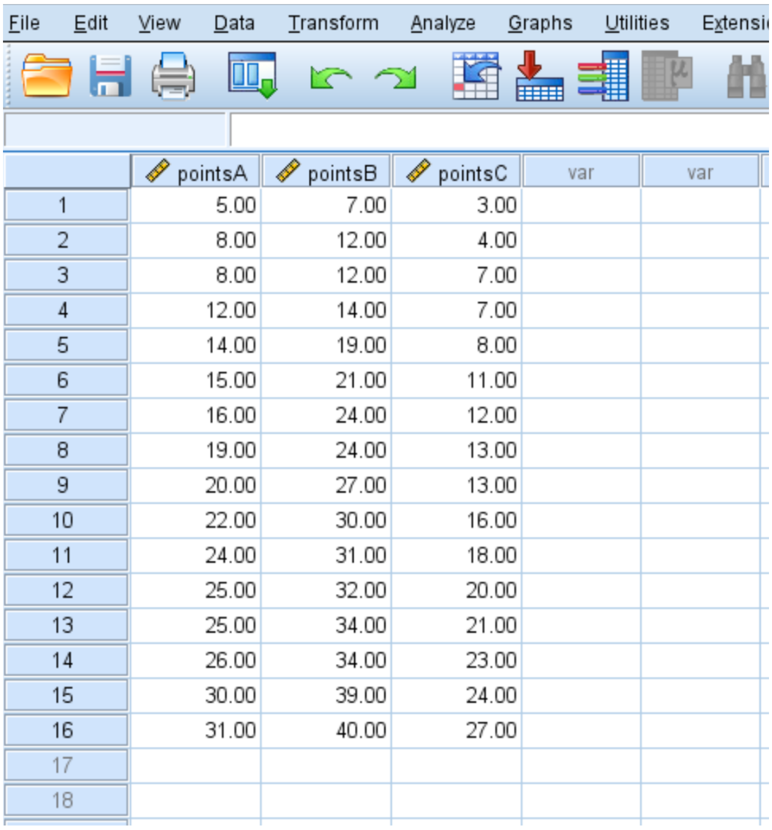
Verify Data Integrity: The first step is always to ensure the outlier is not simply a **data entry error**. Sometimes values are recorded incorrectly (e.g., a score of 100 mistakenly entered as 1000). If this is the case, the value must be corrected immediately.
Assign a New Value (Imputation): If the outlier is confirmed to be a result of a data entry error or missing data, you may decide to assign a new, more representative value to it. Common imputation methods involve using the **median** or mean of the dataset, or perhaps a regression-based estimate, to maintain the integrity of the sample size.
Remove the Outlier: If the value is determined to be a true, natural, but extremely rare occurrence (a "true outlier"), you may choose to remove it if it will have a significant and unwarranted impact on your overall analysis results. It is essential, however, that any removal of data points is thoroughly documented and justified in your final report, ensuring transparency in your

methodology.

Creating and Comparing Multiple Box Plots in SPSS

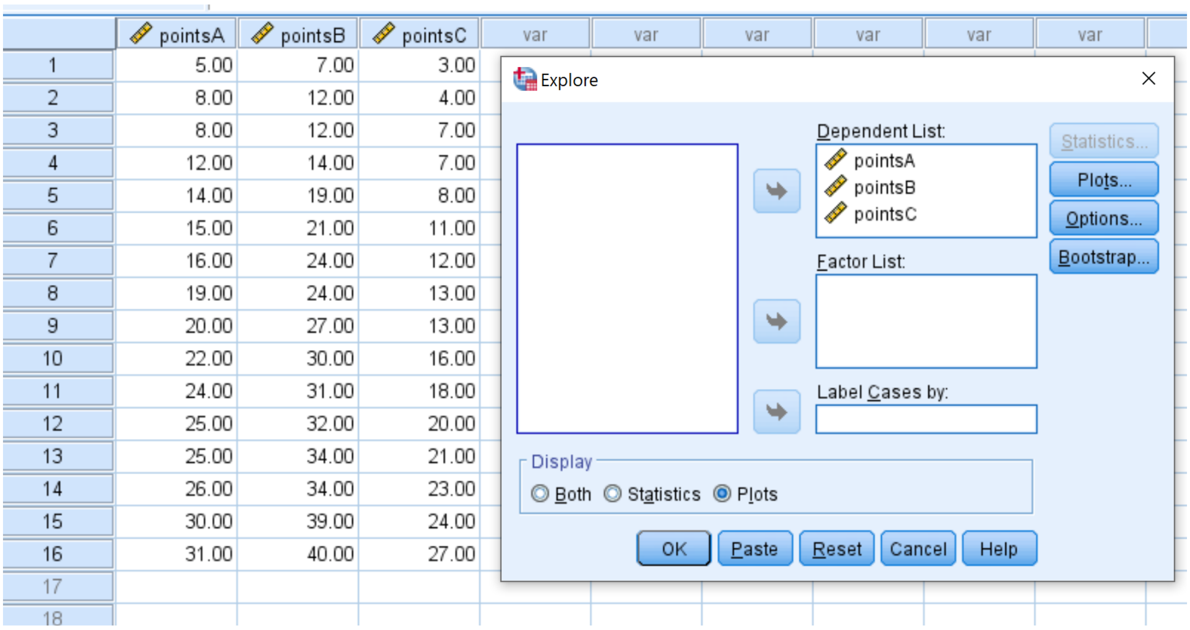
Box plots are particularly effective when used for comparative analysis between different groups or variables. SPSS facilitates the simultaneous creation of multiple, side-by-side box plots, allowing for easy visual assessment of differences in central tendency and variability across categories.

For instance, suppose we expand our analysis to include data on average points scored by 16 players across three different teams (Team A, Team B, and Team C):

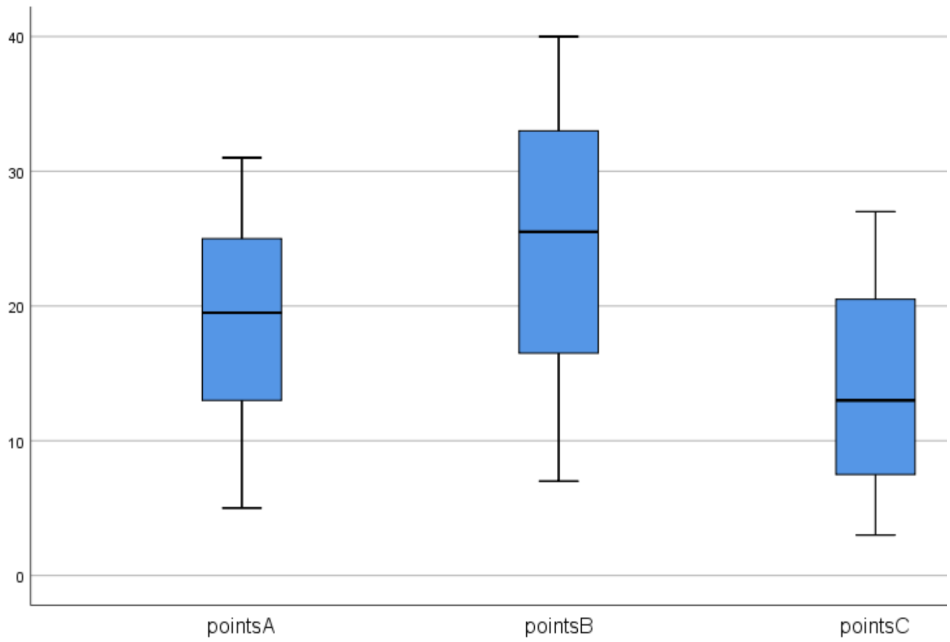


	pointsA	pointsB	pointsC	var	var
1	5.00	7.00	3.00		
2	8.00	12.00	4.00		
3	8.00	12.00	7.00		
4	12.00	14.00	7.00		
5	14.00	19.00	8.00		
6	15.00	21.00	11.00		
7	16.00	24.00	12.00		
8	19.00	24.00	13.00		
9	20.00	27.00	13.00		
10	22.00	30.00	16.00		
11	24.00	31.00	18.00		
12	25.00	32.00	20.00		
13	25.00	34.00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

To generate comparative box plots for these three variables simultaneously, we revisit the **Explore** dialogue box (Analyze > Descriptive Statistics > Explore). This time, instead of just moving one variable, we drag all three variables--Team A, Team B, and Team C--into the **Dependent List**. SPSS treats each variable in the Dependent List as a separate distribution to be plotted.



Upon clicking **OK**, the Output Viewer presents the distributions of the three teams side-by-side, providing an immediate visual comparison:



This visualization greatly assists in drawing meaningful conclusions regarding group differences. We can readily compare the central location (median), the spread (IQR), and the presence of extreme scores for each team.

From the visual analysis of these multiple box plots, several key observations can be made:

The **median** points scored per game (the line inside the box) is highest for Team B and noticeably lowest for Team C, indicating significant differences in typical player performance across the teams.

The **variation** or spread in points scored per game is highest for Team B, evidenced by the longest box length, suggesting greater heterogeneity in player scoring habits within that team compared to the more consistent performance seen in Teams A and C.

The player with the highest overall points per game is located on Team B, while the player with the lowest score is on Team C, highlighting the full range of performance within the analyzed groups.

Box plots are indispensable tools in statistical reporting because they offer a comprehensive summary of distribution characteristics--including central tendency, spread, and skewness--from a single, easily interpretable plot. They are highly efficient for initial data screening and powerful for comparative research.