

An Introduction to SAS BOXPLOT Procedure with Example

The BOXPLOT procedure in SAS is a powerful tool for creating box-and-whisker plots, which are useful for visualizing the distribution of data and identifying outliers. Box plots summarize key statistical measures such as the median, quartiles, and potential outliers, providing a clear graphical representation of data variability.

Key Features of the BOXPLOT Procedure

- **Displays central tendency and dispersion:** Box plots show the median, quartiles, and range of the data.
- **Identifies outliers:** The procedure highlights outliers and extreme values in the dataset.
- **Handles grouping variables:** BOXPLOT can create comparative box plots for different groups within the data.

Basic Syntax of PROC BOXPLOT

Here is the basic syntax for using the BOXPLOT procedure in SAS:

```
PROC BOXPLOT DATA=dataset;  
  PLOT variable*group_variable;  
RUN;
```

- **DATA=dataset:** Specifies the input dataset.
- **PLOT variable*group_variable:** Creates box plots of the variable for each level of the group variable.

Example: Creating a Box Plot

Let's consider a dataset named sashelp.cars, which contains information about different car models, including their MSRP (Manufacturer's Suggested Retail Price) and the type of car (Sedan, SUV, etc.). We want to create box plots of the MSRP for different types of cars.

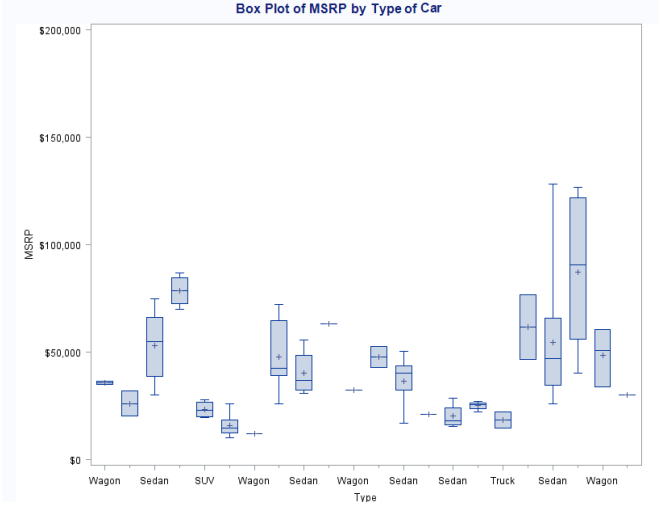
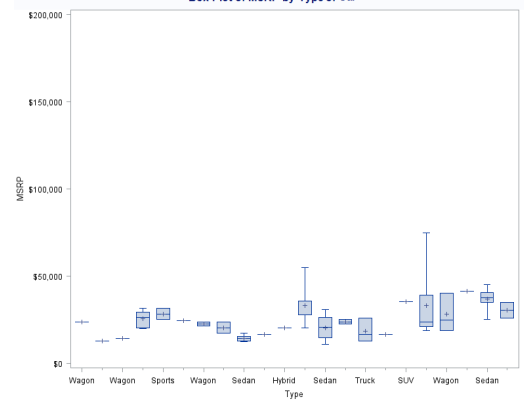
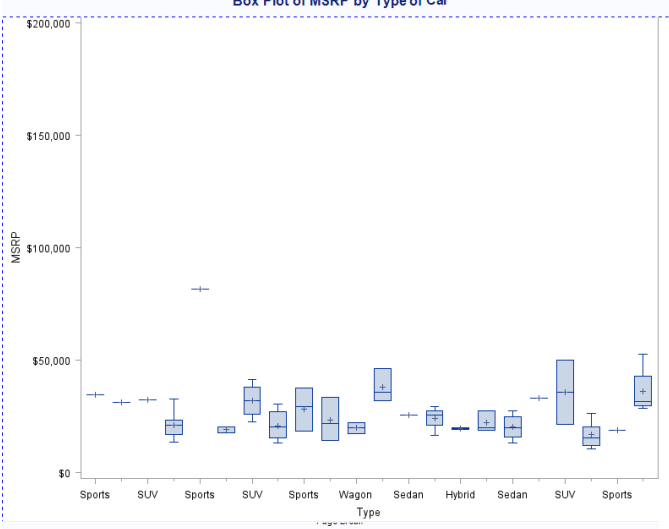
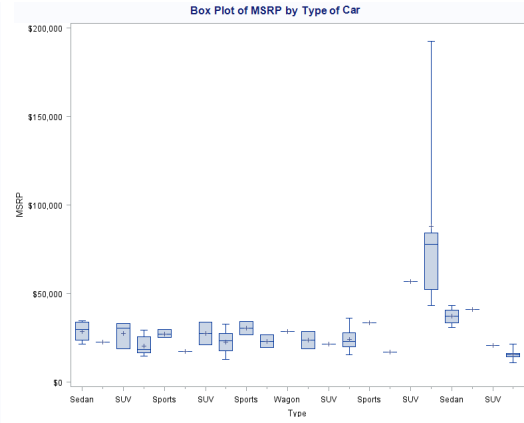
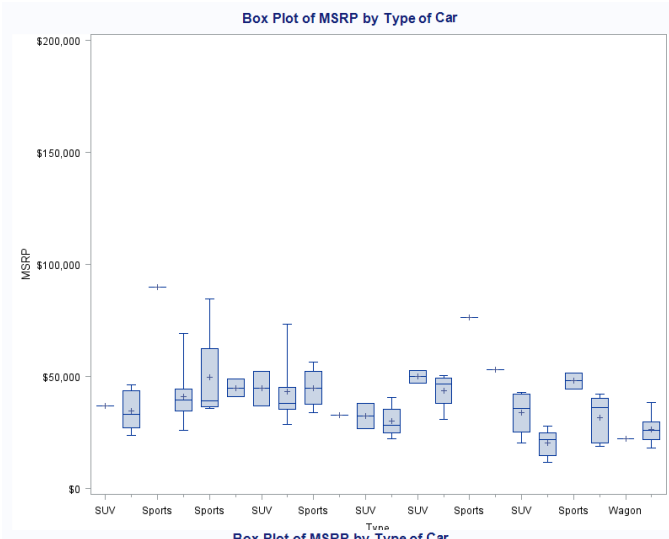
Step 1: Load the Data

First, ensure that the sashelp.cars dataset is available in your SAS environment.

Step 2: Use the BOXPLOT Procedure

The following code creates a box plot of the MSRP for each type of car:

```
PROC BOXPLOT DATA=sashelp.cars;  
  PLOT MSRP*Type;  
  TITLE 'Box Plot of MSRP by Type of Car';  
RUN;
```



First col down then second column.

Explanation:

- **PROC BOXPLOT DATA=sashelp.cars:** Specifies the sashelp.cars dataset.
- **PLOT MSRP*Type:** Creates box plots for the MSRP variable, grouped by the Type of car.
- **TITLE 'Box Plot of MSRP by Type of Car':** Adds a title to the plot.

Interpreting the Box Plot

The box plot generated by the above code will show the distribution of MSRP values for each type of car:

- **Box:** Represents the interquartile range (IQR), which contains the middle 50% of the data.
- **Line inside the box:** Indicates the median value of the data.
- **Whiskers:** Extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles.
- **Dots outside the whiskers:** Represent outliers, which are values beyond 1.5 times the IQR.

Conclusion

The BOXPLOT procedure in SAS is a versatile tool for visualizing data distributions and identifying outliers across different groups. By using this procedure, analysts can gain valuable insights into their data, making it easier to detect patterns, trends, and anomalies. The example provided demonstrates how to create a simple box plot in SAS, showcasing the MSRP of various types of cars, but the procedure can be adapted to a wide range of datasets and analysis scenarios.