Chapter 4
Exploring Data in One Dimension

Chapter Table of Contents

Chapter 4
Exploring Data in One Dimension

In SAS/INSIGHT software, you can explore distributions of one variable using bar charts and box plots. *Bar charts* display distributions of interval or nominal variables. *Box plots* display concise summaries of interval variable distributions and show extreme values.

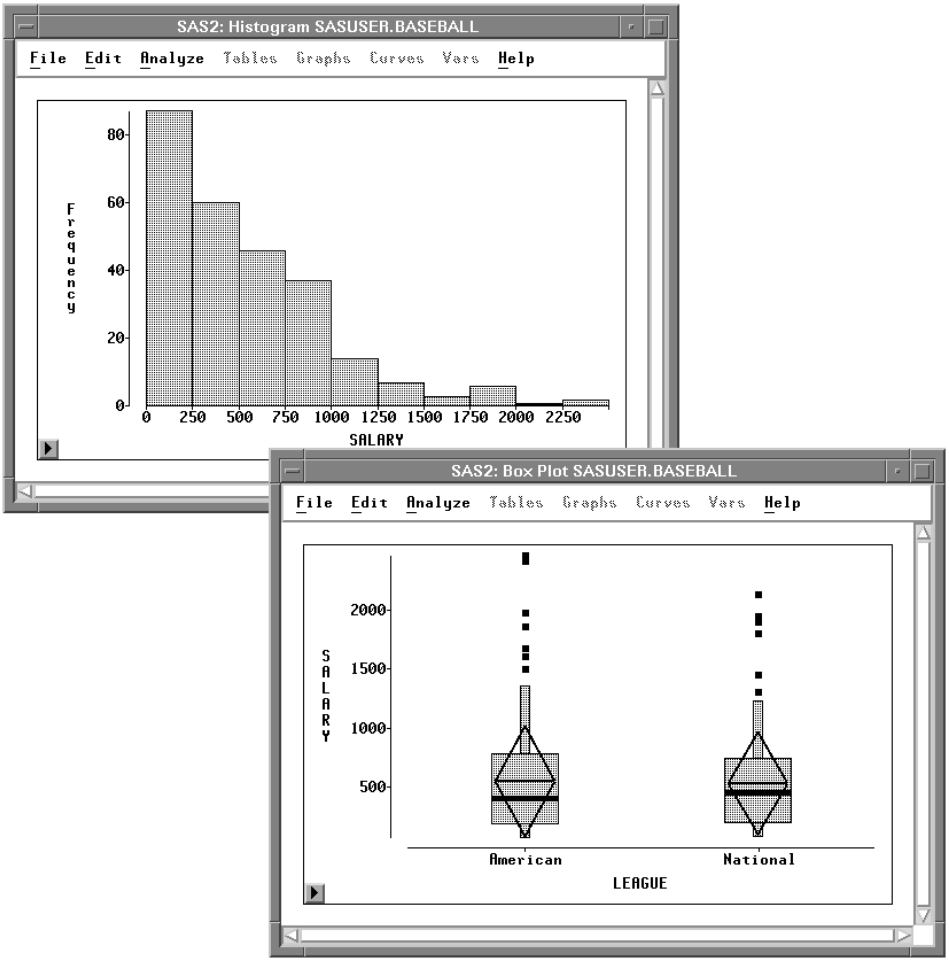
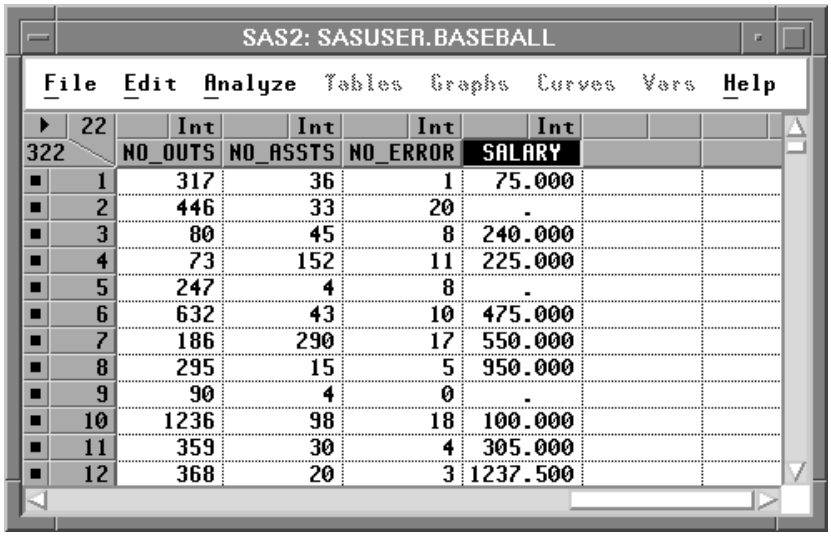


Figure 4.1. A Bar Chart and Box Plot

Bar Charts

Interval variables contain values distributed over a continuous range. For example, in Figure 4.2 baseball players' salaries are stored in **SALARY**, an interval variable. To create a bar chart of players' salaries, follow these steps.

⟹ **Select SALARY in the data window.**
Scroll all the way to the right to find the **SALARY** variable. Point and click on the variable name.



322	NO_OUTS	NO_ASSTS	NO_ERROR	SALARY
1	317	36	1	75.000
2	446	33	20	.
3	80	45	8	240.000
4	73	152	11	225.000
5	247	4	8	.
6	632	43	10	475.000
7	186	290	17	550.000
8	295	15	5	950.000
9	90	4	0	.
10	1236	98	18	100.000
11	359	30	4	305.000
12	368	20	3	1237.500

Figure 4.2. Selecting the **SALARY** Variable

⟹ **Choose Histogram/Bar Chart (Y) from the Analyze menu.**

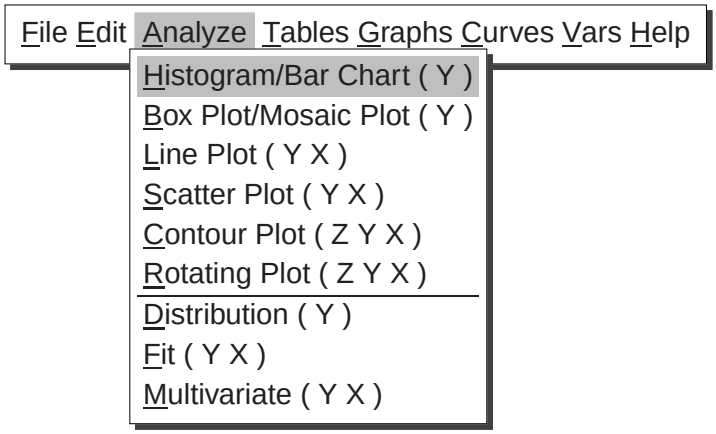


Figure 4.3. Creating a Bar Chart

This creates a bar chart, as shown in Figure 4.4.

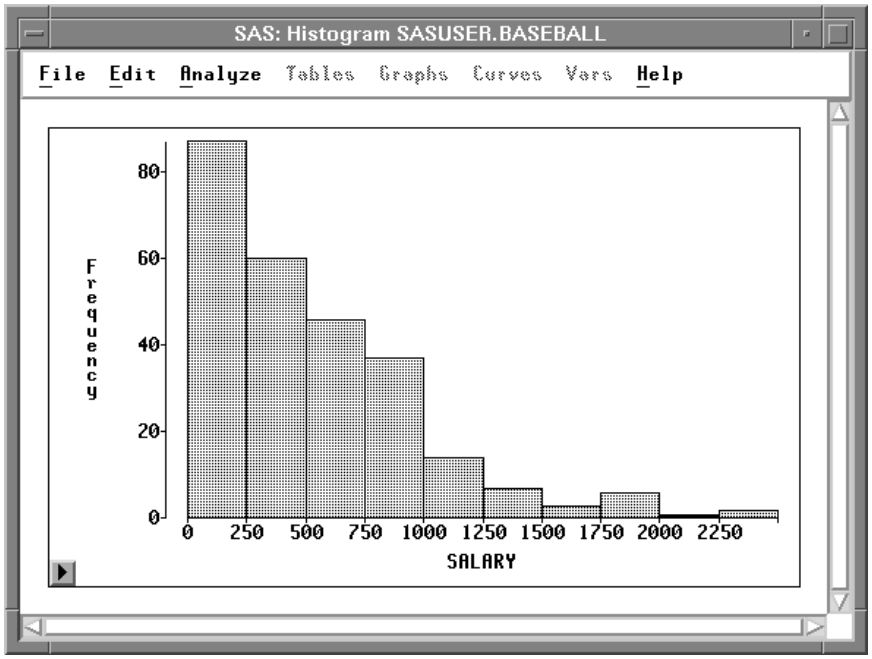


Figure 4.4. Bar Chart

⟹ **Point and click on any bar.**
This labels the bar with its frequency and selects all the observations in the bar.

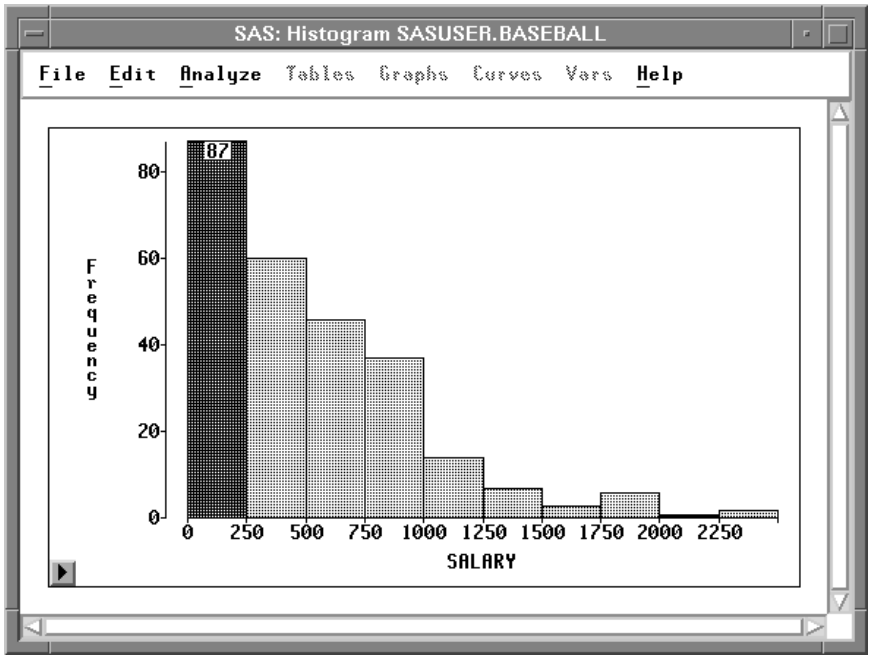


Figure 4.5. Clicking on a Bar

Notice that the observations are selected in the data window as well as in the bar chart window. Windows in SAS/INSIGHT software are just different views of the same data, so observations you select in one window are selected in all other windows.

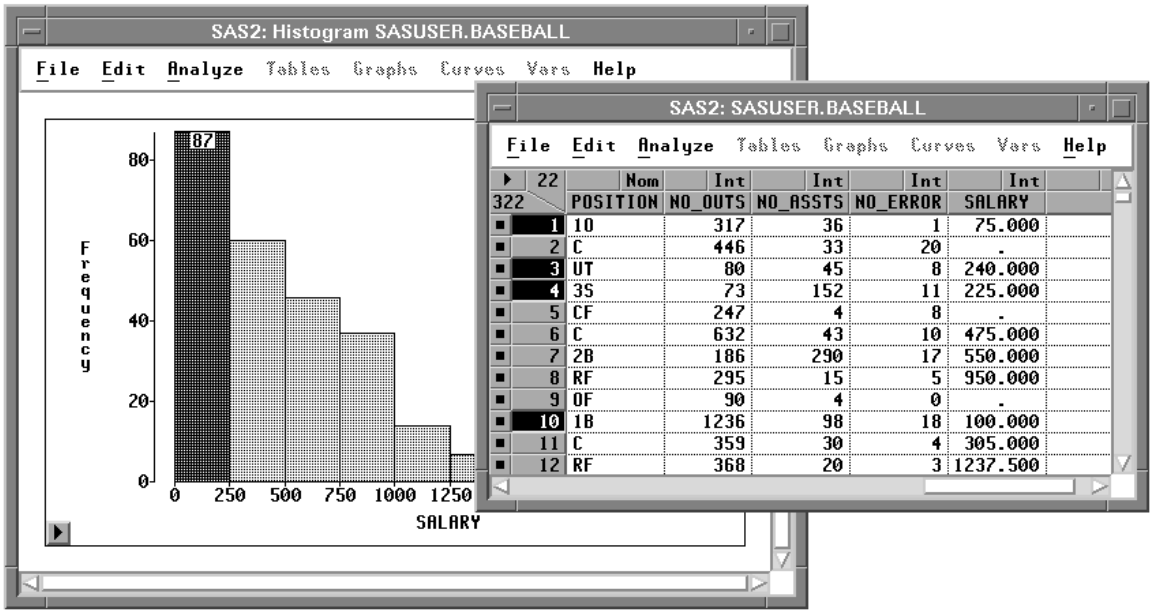


Figure 4.6. Selecting Observations in Multiple Windows

From this bar chart, you can see that the distribution of players' salaries is skewed to the right, with a few players earning high salaries. To find the number of players making the highest salaries, you can label all bars with their heights.

⟹ **Click on the menu button in the bottom left corner of the chart.**
This displays the bar chart pop-up menu in Figure 4.7. Click on **Values**.

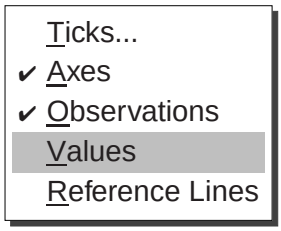


Figure 4.7. Bar Chart Pop-up Menu.

This toggles the display of values for all bar heights. There are three players making salaries above $2,000,000.

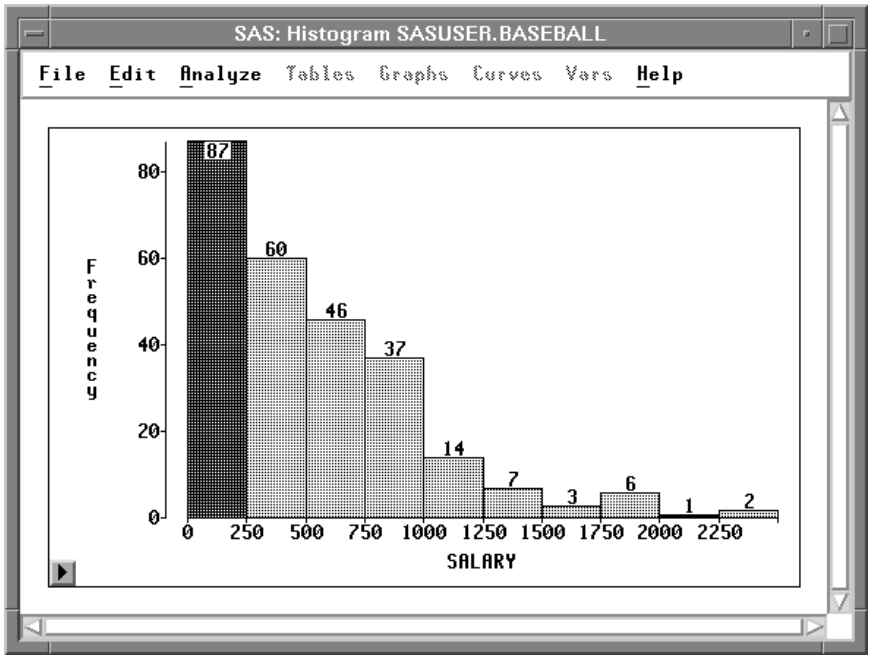
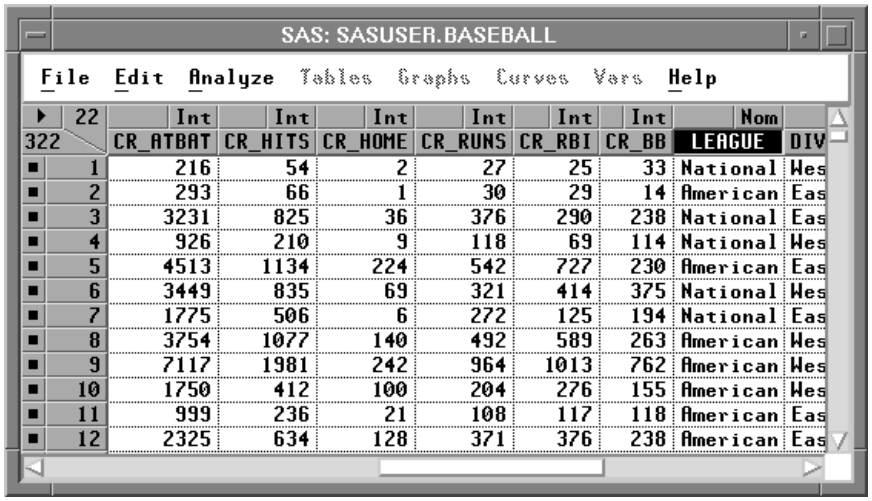


Figure 4.8. Bar Heights

It would be interesting to determine whether salaries differ in the American and National leagues. To compare the distribution of salaries from both leagues, follow these steps.

⟹ **Select LEAGUE in the data window.**



322	Int CR_ATBAT	Int CR_HITS	Int CR_HOME	Int CR_RUNS	Int CR_RBI	Int CR_BB	Nom LEAGUE	DIV
1	216	54	2	27	25	33	National	Wes
2	293	66	1	30	29	14	American	Eas
3	3231	825	36	376	290	238	National	Eas
4	926	210	9	118	69	114	National	Wes
5	4513	1134	224	542	727	230	American	Eas
6	3449	835	69	321	414	375	National	Wes
7	1775	506	6	272	125	194	National	Eas
8	3754	1077	140	492	589	263	American	Wes
9	7117	1981	242	964	1013	762	American	Wes
10	1750	412	100	204	276	155	American	Wes
11	999	236	21	108	117	118	American	Eas
12	2325	634	128	371	376	238	American	Eas

Figure 4.9. Selecting **LEAGUE**

Note that **LEAGUE** is a *nominal* variable. Nominal variables contain a discrete set of values. For example, **LEAGUE** contains only two values, **American** and **National**, for the American and National leagues.

⟹ **Choose Histogram/Bar Chart (Y) from the Analyze menu.**
From the bar chart in Figure 4.10 you can see that the **BASEBALL** data set has more observations from the American League.

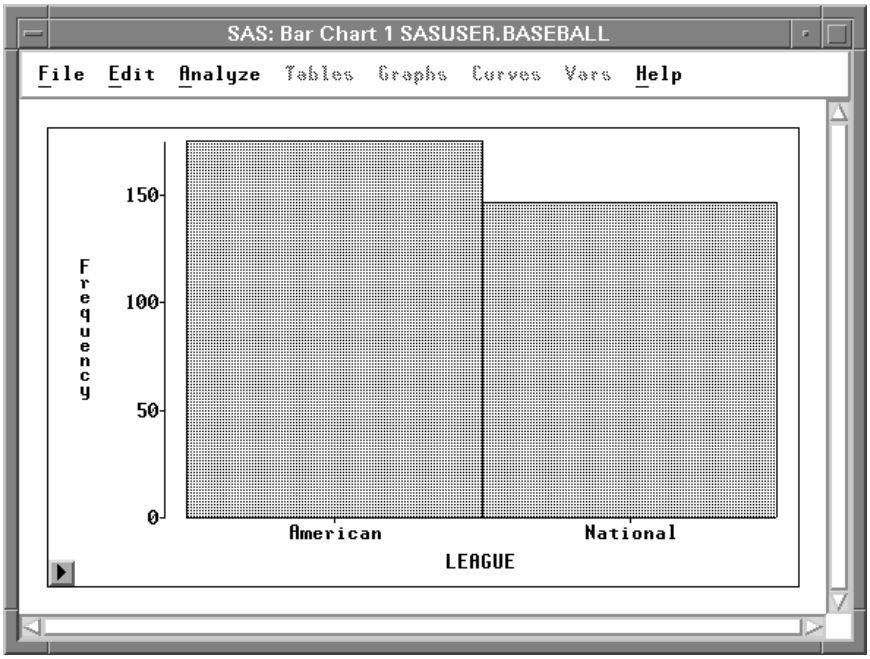


Figure 4.10. Bar Chart of **LEAGUE**

⟹ **Select Values from the bar chart pop-up menu in the new bar chart.**
This displays the frequencies for each of the leagues at the top of the bars on the bar chart.

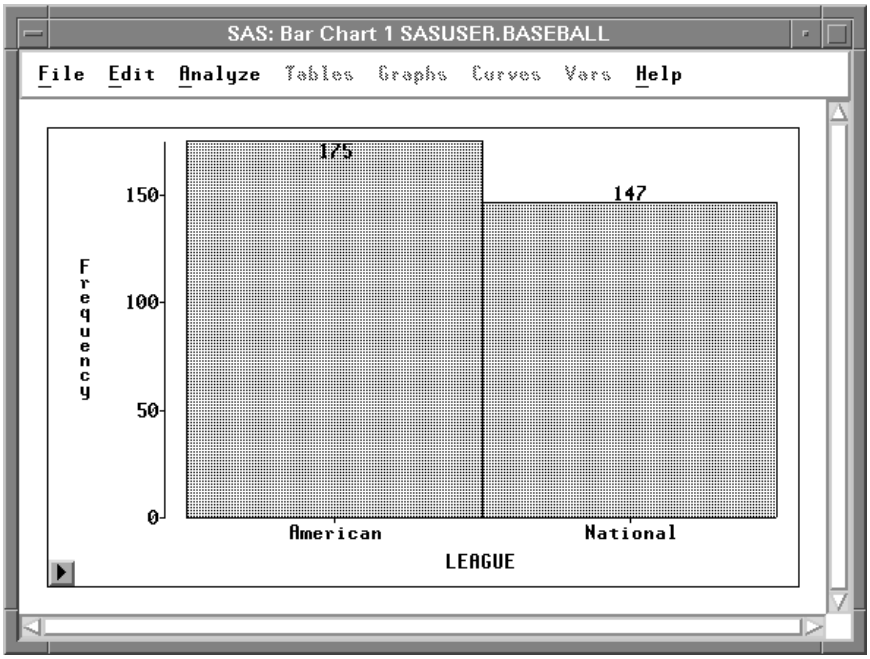


Figure 4.11. Bar Chart with Frequency Values

⟹ **Arrange the windows so you can see both bar charts.**

⟹ **Click on the bar that represents the American League.**
This selects all observations for players in the American League.

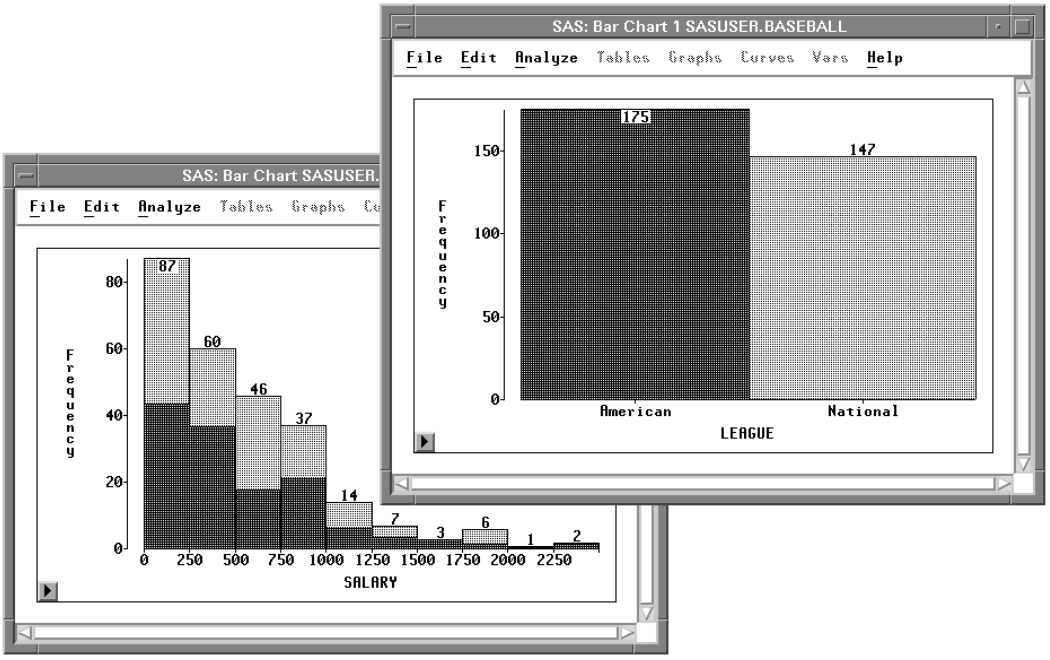


Figure 4.12. Selecting American League Observations

⟹ **Click on the bar that represents the National League.**
This selects all observations for players in the National League.

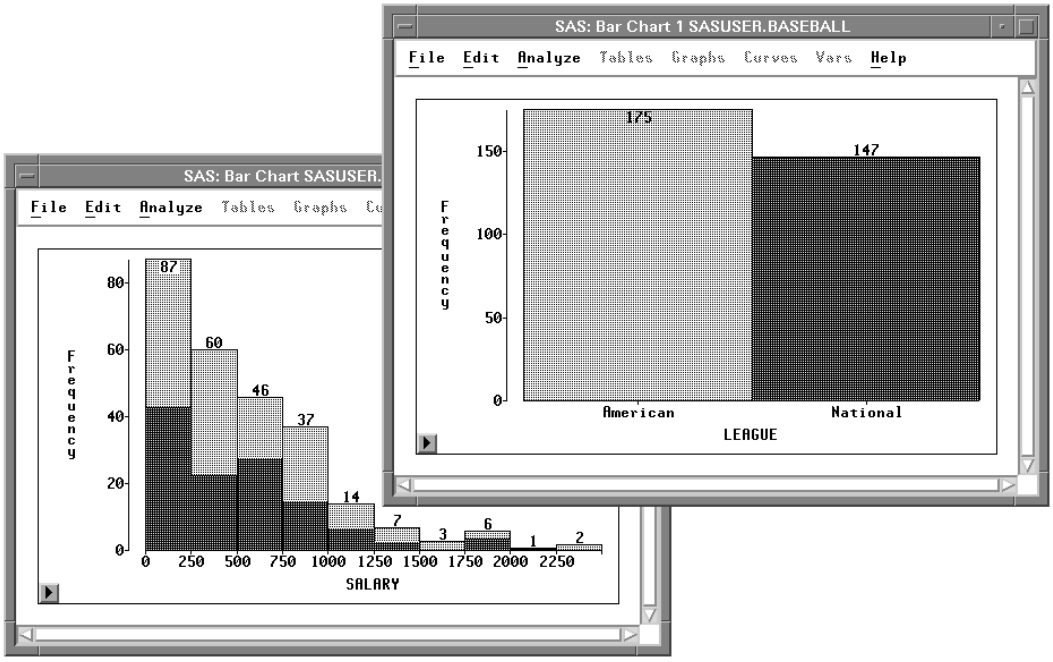


Figure 4.13. Selecting National League Observations

Both leagues have a broad distribution of **SALARY** with most players earning below $1,000,000 and a few earning much more.

You can examine the distributions in more detail by creating box plots.

⊕ **Related Reading:** Bar Charts, Chapter 32.

Box Plots

Box plots are an effective way to compare distributions of interval data. To create side-by-side box plots comparing the distributions of salaries for the American and National Leagues, follow these steps.

⟹ **Choose Analyze:Box Plot/Mosaic Plot (Y).**

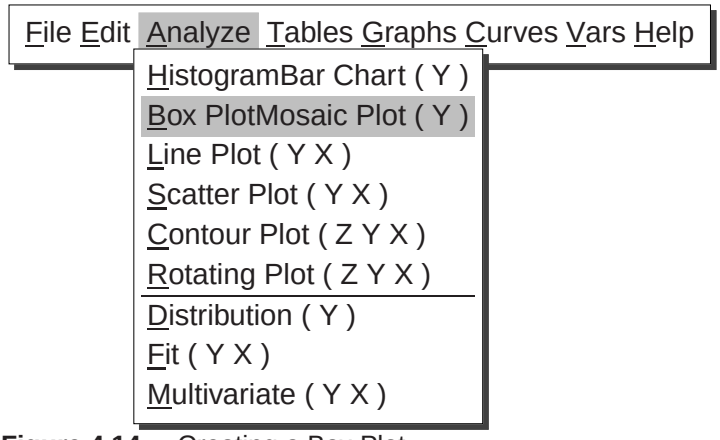


Figure 4.14. Creating a Box Plot

The **(Y)** in the **Box Plot/Mosaic Plot (Y)** menu indicates that a **Y** variable is *required* to create a box plot. Since you have no variables selected, a variables dialog prompts you to select at least one **Y** variable. Selecting a nominal variable for **Y** creates a mosaic plot; selecting an interval variable for **Y** creates a box plot.

Y is one of several *roles* you can assign to variables in analyses. The variables dialog shows that box plots and mosaic plots can also use **X**, **Group**, **Label**, and **Freq** variables.

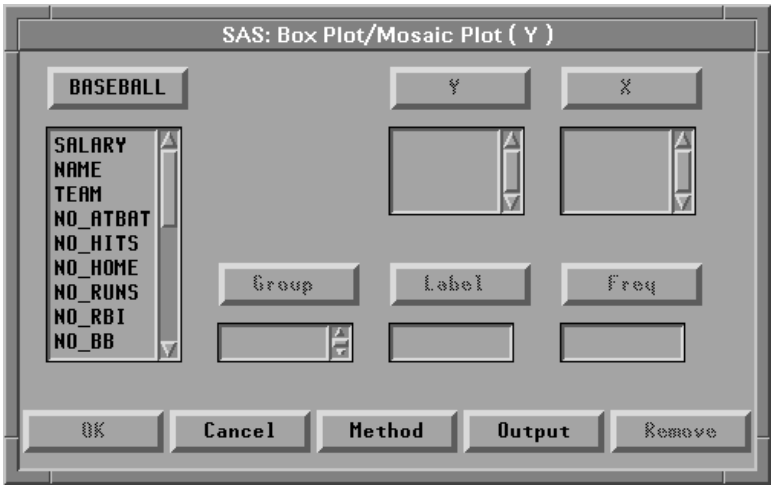


Figure 4.15. Box Plot Variables Dialog

† **Note:** You can select variables before choosing from the **Analyze** menu, or you can choose from the **Analyze** menu before selecting variables. Selecting variables first is faster. If you select variables first, they are assigned to the required variable roles listed in the **Analyze** menu. Choosing the analysis first gives you more flexibility. If you choose the analysis first, you can assign optional variable roles such as **Group** and **Label**.

⟹ **Select SALARY in the list at the left, then click the Y button.**
This assigns the **Y** role to **SALARY**. The box plot displays the distribution of the **Y** variable.

⟹ **Select LEAGUE in the list at the left, then click the X button.**
This assigns the **X** role to **LEAGUE**. The box plot displays one schematic distribution plot side-by-side for each unique value of the **X** variable.

⟹ **Select NAME in the list at the left, then click the Label button.**
This assigns the **Label** role to **NAME**. The label variable is used to identify extreme values in the box plot.

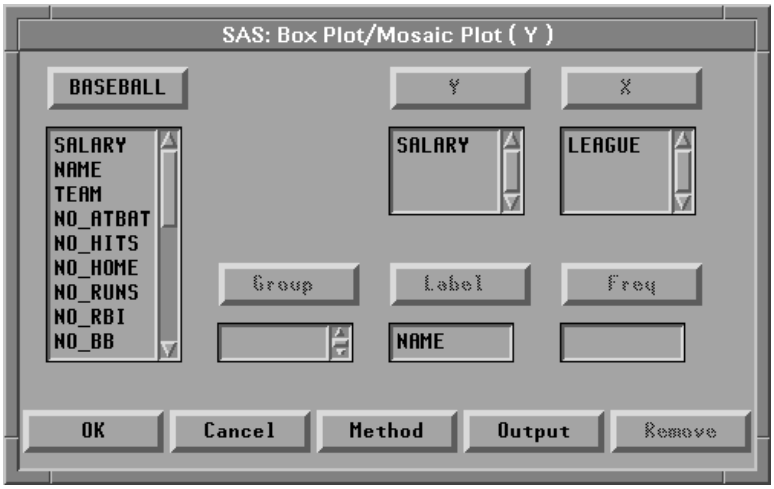


Figure 4.16. Assigning Variable Roles

⟹ **Click OK to create the Box Plot.**

The box plot gives a concise picture of the distributions and places them side-by-side for easy comparison. The horizontal line in the middle of a box marks the *median* or 50th percentile. The top and bottom edges of a box mark the *quartiles*, or the 25th and 75th percentiles. The narrow boxes extending above and below are called *whiskers*. Whiskers extend from the quartiles to the farthest observation not farther than 1.5 times the distance between the quartiles. More extreme data values are plotted with individual markers.

The box plot shows long whiskers above with individual observations beyond the whiskers indicating severe skewness. These are the players making extremely high salaries.

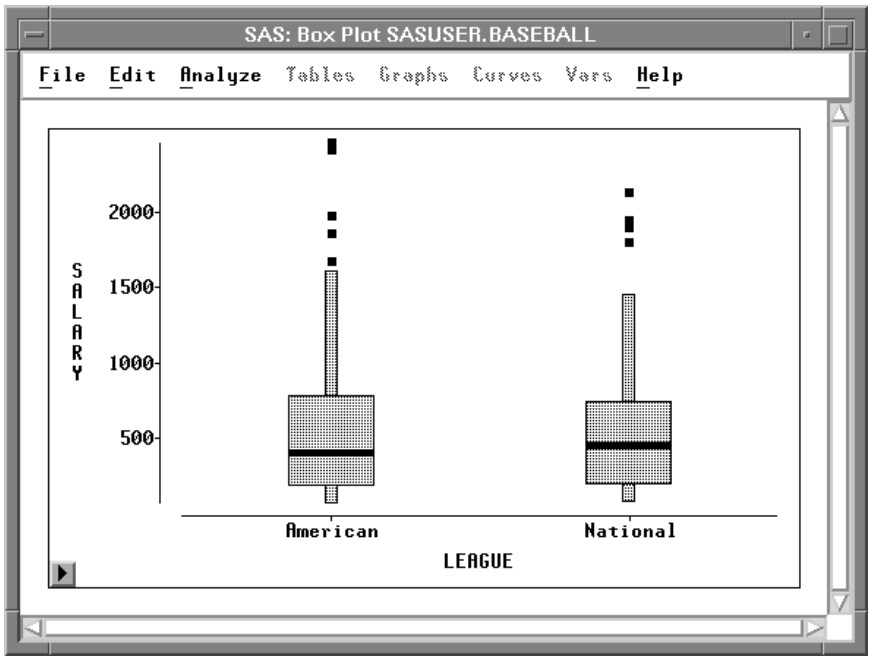


Figure 4.17. Side-By-Side Box Plots

⟹ **Point and click at the extreme values to identify them.**
Eddie Murray and Jim Rice were the highest paid players in the American league, while Mike Schmidt was the highest paid player in the National League.

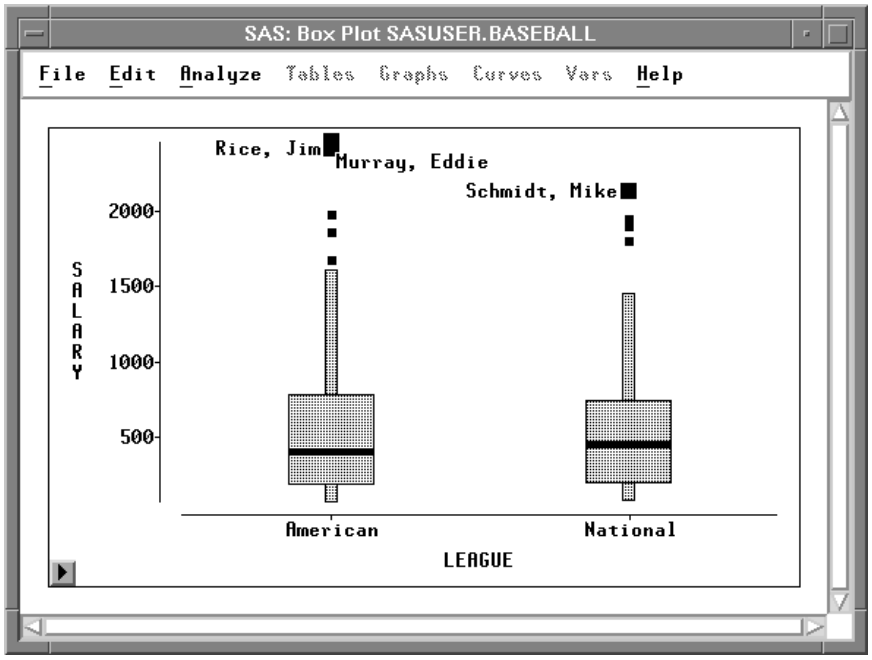


Figure 4.18. Identifying Extreme Values

You can also use a box plot to see the sample mean of a distribution.

⟹ **Click on the menu button in the lower left corner of the plot.**
This displays the box plot pop-up menu. Click on **Means**.

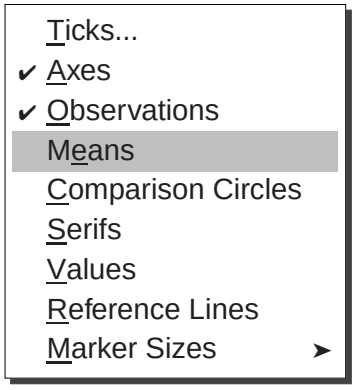


Figure 4.19. Box Plot Pop-up Menu

This toggles the display of mean diamonds on the box plot.

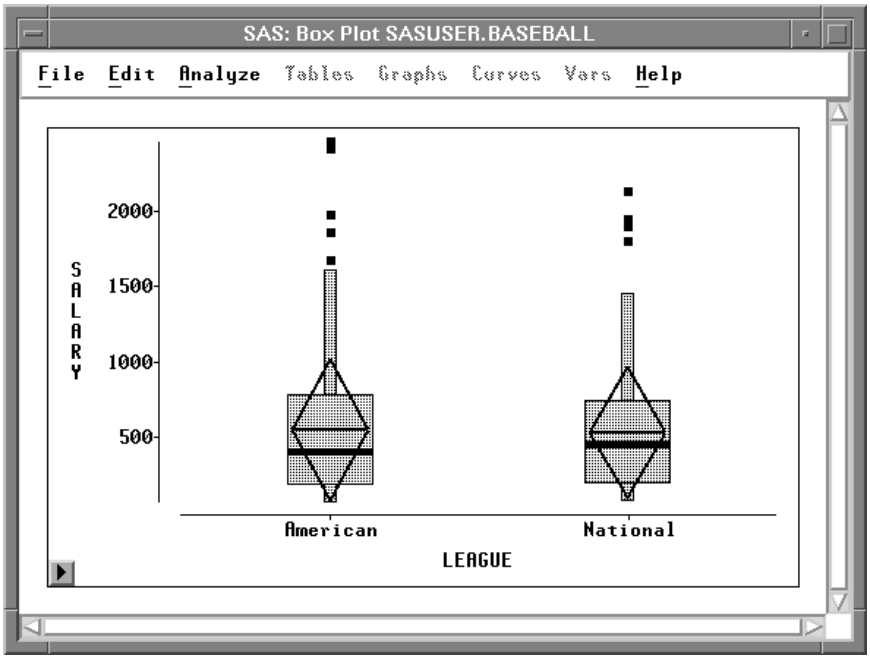


Figure 4.20. Box Plot with Mean Diamonds

The horizontal line in a mean diamond marks the mean salary for each league. The height of a mean diamond is two standard deviations (one on either side of the mean). In this case, the means and standard deviations for each league are almost identical.

You can use other choices on the box plot pop-up menu to adjust axis tick marks and marker sizes and to toggle the display of observations, axes, serifs, and values. When there are two or more categories, you can toggle the display of *comparison circles*, which enable you to graphically compare the means of multiple categories.

⊕ **Related Reading:** Box Plots, Chapter 33.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/INSIGHT User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999. 752 pp.

SAS/INSIGHT User's Guide, Version 8

ISBN 1–58025–490–X

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999