Chapter 5
Exploring Data in Two Dimensions

Chapter Table of Contents

Chapter 5
Exploring Data in Two Dimensions

SAS/INSIGHT software provides mosaic plots, scatter plots, and line plots for exploring data in two dimensions. *Mosaic plots* are pictorial representations of frequency counts of nominal variables. *Scatter plots* are graphic representations of the relationship between two interval variables. *Line plots* show the relationships of multiple **Y** variables to a single **X** variable.

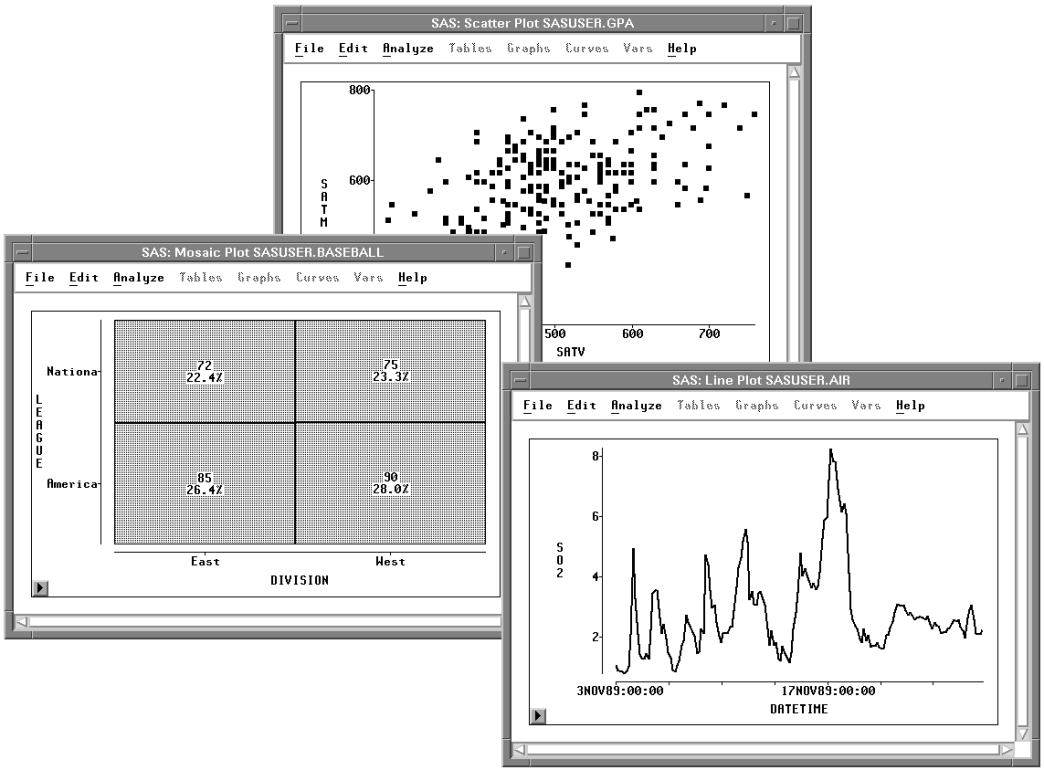


Figure 5.1. A Mosaic Plot, Scatter Plot, and Line Plot

Mosaic Plots

This example illustrates how to create mosaic plots for the **BASEBALL** data cross-classified by **LEAGUE** and **DIVISION**.

⟹ **Open the BASEBALL data set.**

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

⟹ **Assign LEAGUE the Y role and DIVISION the X role. Then click OK.**

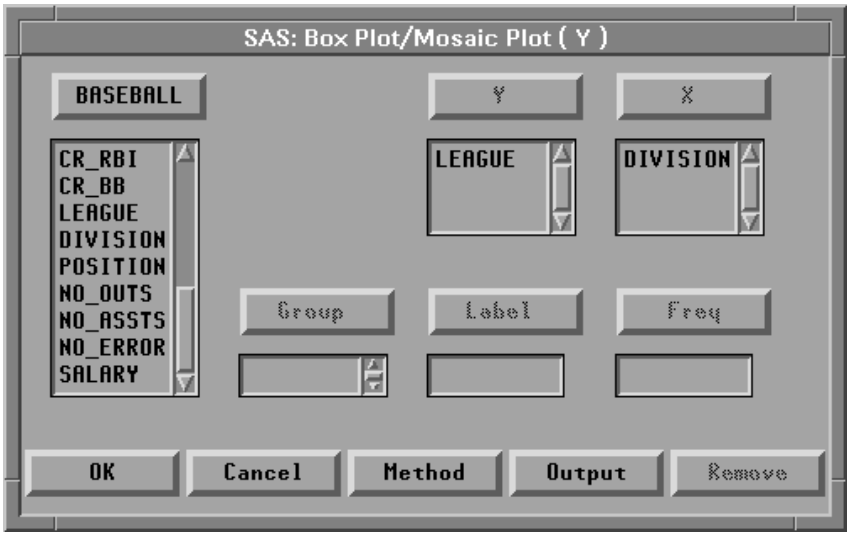


Figure 5.2. Assigning Variables for a Mosaic Plot

This creates a mosaic plot containing four boxes. The areas of the boxes in the mosaic plot are proportional to the number of observations in each category. You can see that, for these data, there are more players in the American League than in the National League and about the same number of players in the East and West Divisions.

You can find out more about specific categories by selecting the boxes.

⟹ **Click on the box at the lower left (American League East).**
This selects all the observations in the box and labels the box with its frequency and percentage. For this data, there are 85 players from the East Division of the American League, and these are 26.4% of the total.

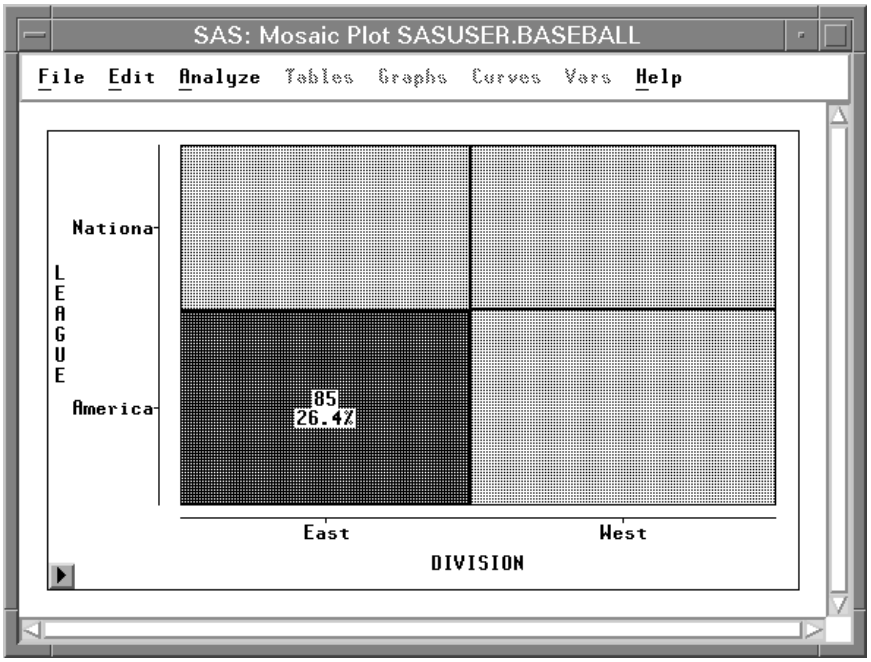


Figure 5.3. Clicking on a Box

⟹ **Double-click on the box to examine the observations.**
This selects all the observations in the box and displays the Examine Observations dialog. By clicking in the Examine Observations dialog, you can get detailed information on all the selected observations.

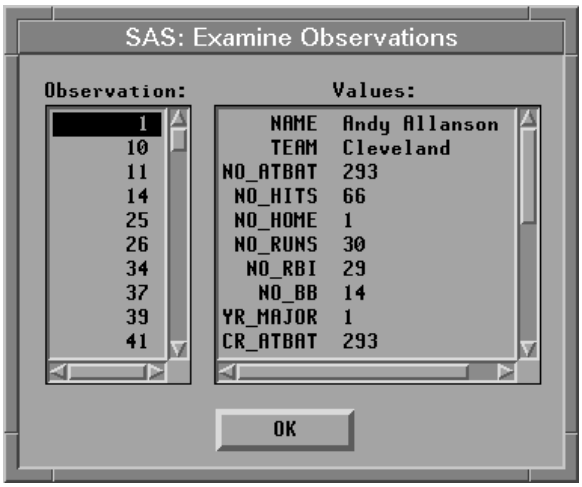


Figure 5.4. Examine Observation Dialog

You can add more information to the mosaic plot by displaying frequency counts and percentages.

⟹ **Choose Values from the pop-up menu.**

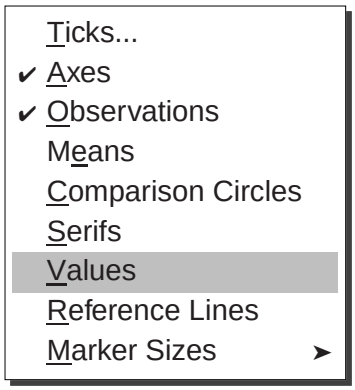


Figure 5.5. Mosaic Plot Pop-up Menu

This toggles the display of frequencies and percentages for all boxes in the mosaic plot.

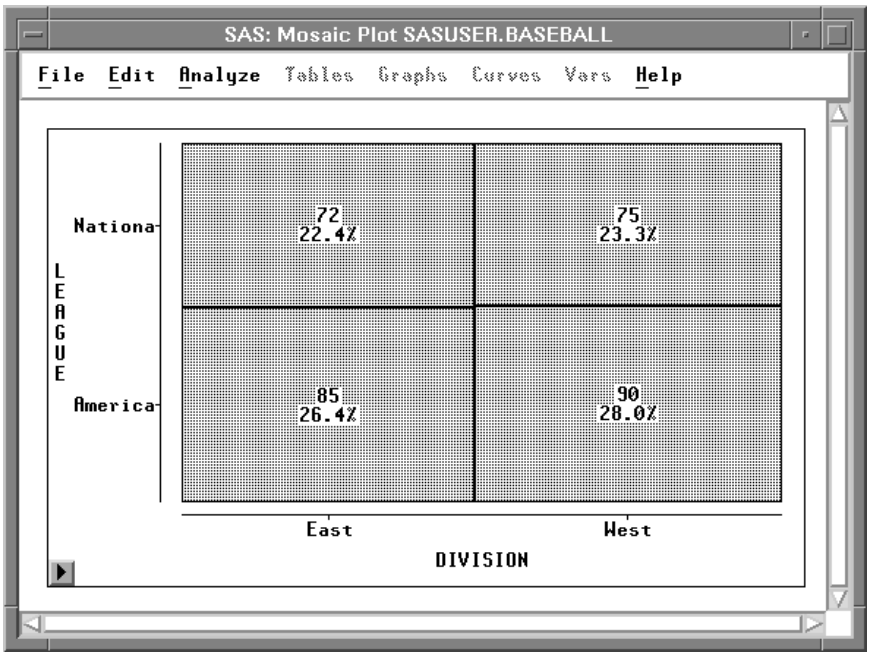


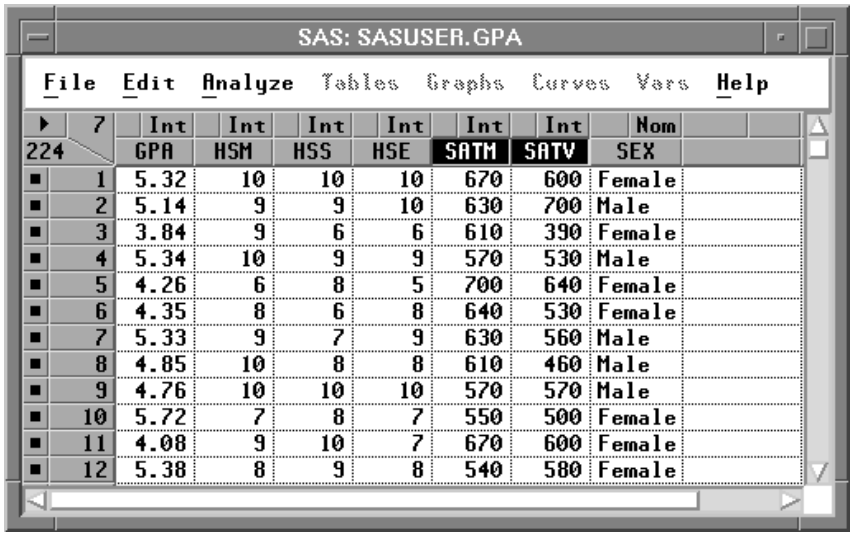
Figure 5.6. Mosaic Plot with Frequencies and Percentages

Scatter Plots

Scatter plots show the relationship between two variables. For example, you can explore the relationship between students' scores on standardized tests of math and verbal ability by following these steps.

⟹ **Open the GPA data set.**

⟹ **Select both the SATM and SATV variables.**
To select both variables, press the mouse button on **SATM**, move the mouse to **SATV**, then release the mouse button.



224	GPA	HSM	HSS	HSE	SATM	SATV	SEX
1	5.32	10	10	10	670	600	Female
2	5.14	9	9	10	630	700	Male
3	3.84	9	6	6	610	390	Female
4	5.34	10	9	9	570	530	Male
5	4.26	6	8	5	700	640	Female
6	4.35	8	6	8	640	530	Female
7	5.33	9	7	9	630	560	Male
8	4.85	10	8	8	610	460	Male
9	4.76	10	10	10	570	570	Male
10	5.72	7	8	7	550	500	Female
11	4.08	9	10	7	670	600	Female
12	5.38	8	9	8	540	580	Female

Figure 5.7. Selecting Two Variables

⟹ **Choose Analyze:Scatter Plot (Y X).**

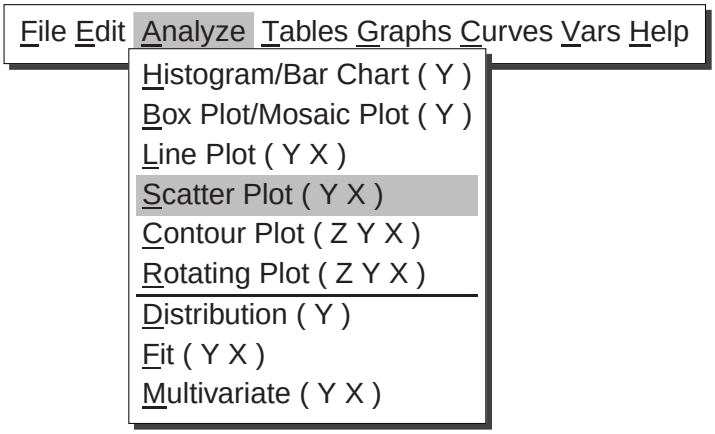


Figure 5.8. Creating a Scatter Plot

This creates a scatter plot, as shown in Figure 5.9. Note that the first variable you selected, **SATM**, is plotted on the **Y** axis, while the second variable selected, **SATV**, is plotted on the **X** axis.

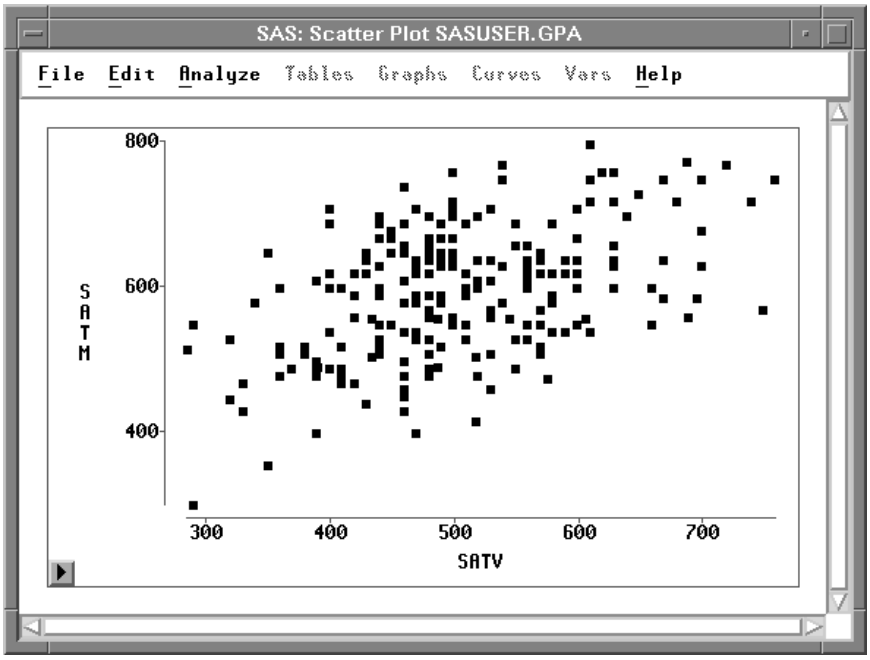


Figure 5.9. Scatter Plot

Each *marker* in the scatter plot represents an observation, and its position shows the values of **SATM** and **SATV** for that observation. You can click on any marker to determine which observation it represents.

⟹ **Click on a marker.**
This selects the marker and displays its observation number. For example, observation 20 is selected in Figure 5.10.

Clicking also selects the observation in the data window because windows are linked to their data. Any change to the data is automatically reflected in all windows.

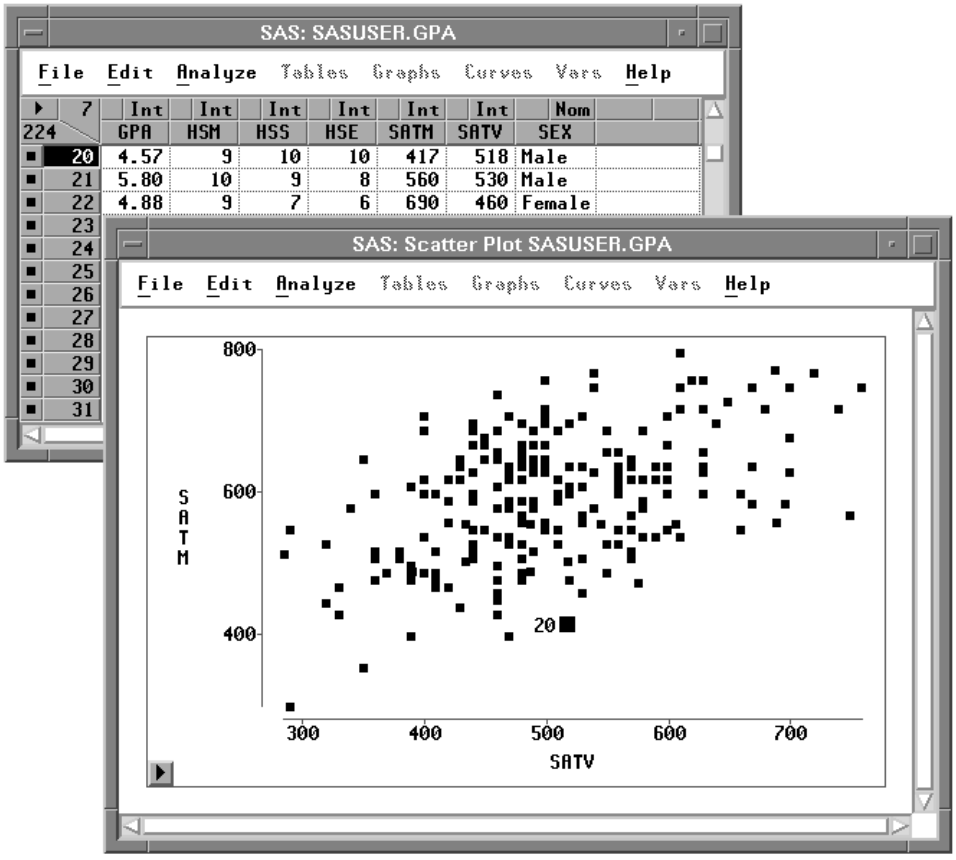


Figure 5.10. Selecting Observations in Multiple Windows

⟹ **Double-click on a marker.**
This selects the marker and displays the Examine Observation dialog. You can examine the values of all variables for the selected observation.

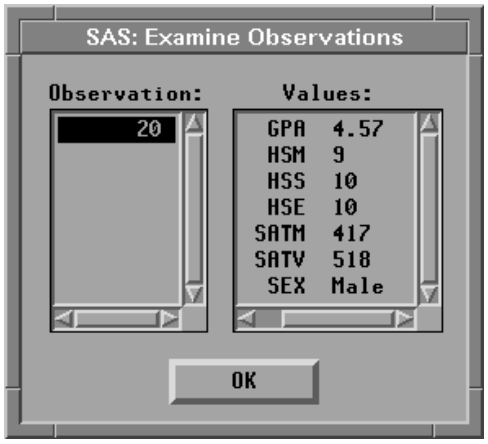


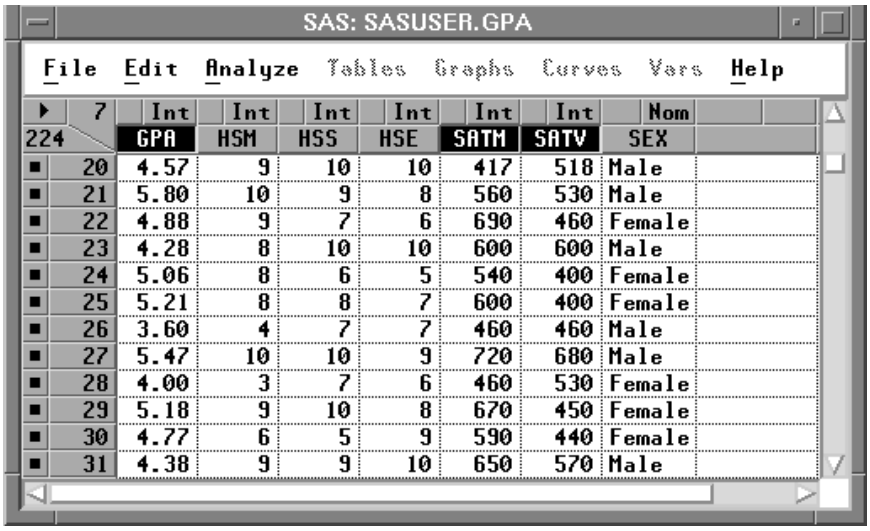
Figure 5.11. Examine Observations Dialog

Scatter Plot Matrices

A scatter plot *matrix* shows relationships among several variables taken two at a time. Scatter plot matrices can reveal a wealth of information, including dependencies, clusters, and outliers.

You can explore the relationships among students' college grade point averages and standardized test scores by following these steps.

⟹ **Select SATM, SATV, and GPA in the data window.**
To select these variables, use noncontiguous selection. On most hosts, you can use the **Ctrl** key to make a noncontiguous selection, as described in Chapter 1, "Getting Started."



224	GPA	HSM	HSS	HSE	SATM	SATV	SEX
	Int	Int	Int	Int	Int	Int	Nom
20	4.57	9	10	10	417	518	Male
21	5.80	10	9	8	560	530	Male
22	4.88	9	7	6	690	460	Female
23	4.28	8	10	10	600	600	Male
24	5.06	8	6	5	540	400	Female
25	5.21	8	8	7	600	400	Female
26	3.60	4	7	7	460	460	Male
27	5.47	10	10	9	720	680	Male
28	4.00	3	7	6	460	530	Female
29	5.18	9	10	8	670	450	Female
30	4.77	6	5	9	590	440	Female
31	4.38	9	9	10	650	570	Male

Figure 5.12. Selecting Three Variables

⟹ **Choose Analyze:Scatter Plot (Y X).**
This creates the scatter plot matrix shown in Figure 5.13.

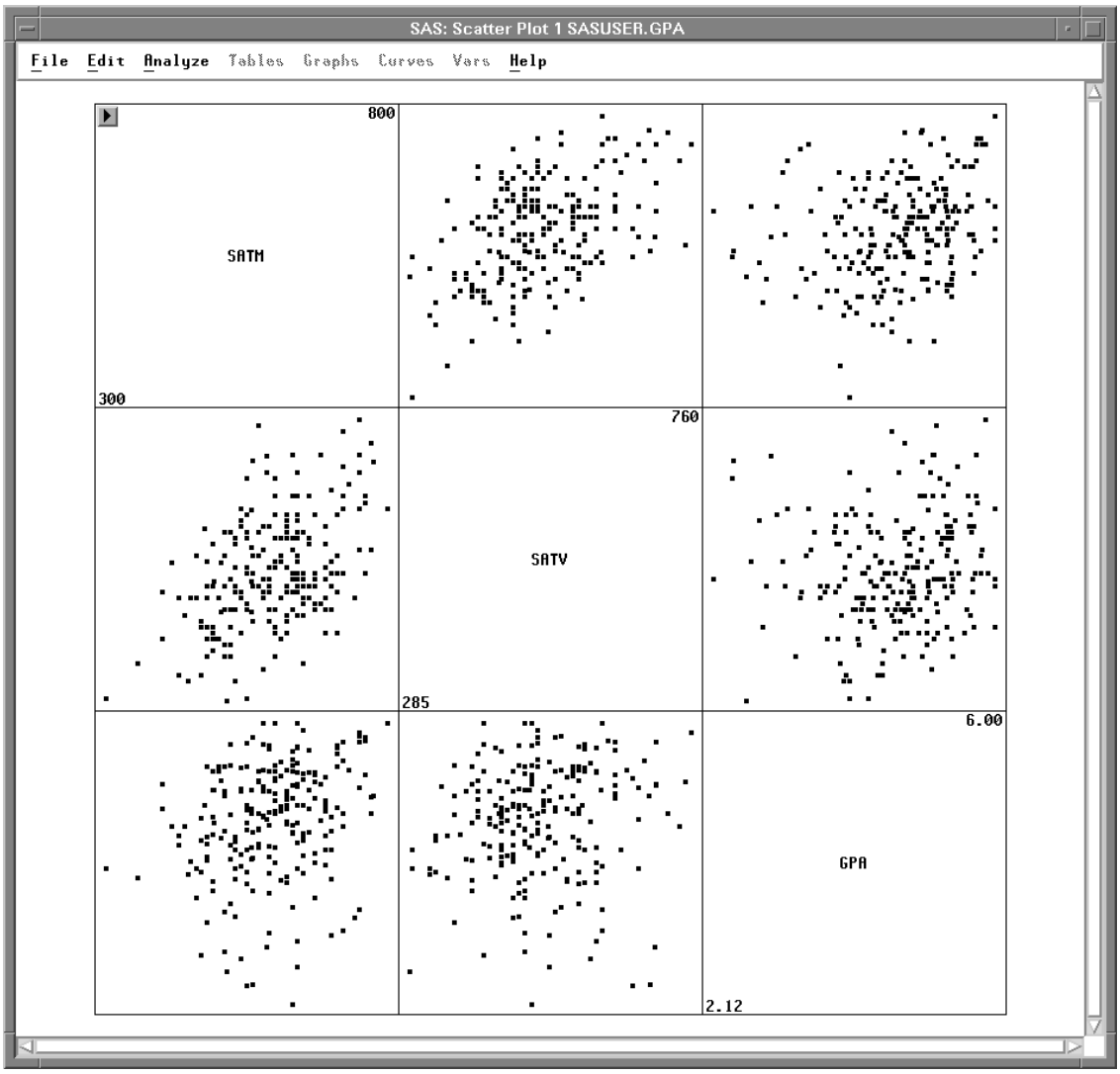


Figure 5.13. Scatter Plot Matrix

The plots are organized in a matrix of all pairwise combinations of the variables **SATM**, **SATV**, and **GPA**. Plots are arranged so that adjacent plots share a common axis. All plots in a row share a common **Y** axis, and all plots in a column share a common **X** axis. The diagonal cells of the matrix contain the names of the variables and their minimum and maximum values.

⟹ **Click on a marker in any scatter plot.**
The observation label is displayed and corresponding markers in all scatter plots are selected, as shown in Figure 5.14. This enables you to explore observations to see, for example, if an outlier in one scatter plot is an outlier in other scatter plots.

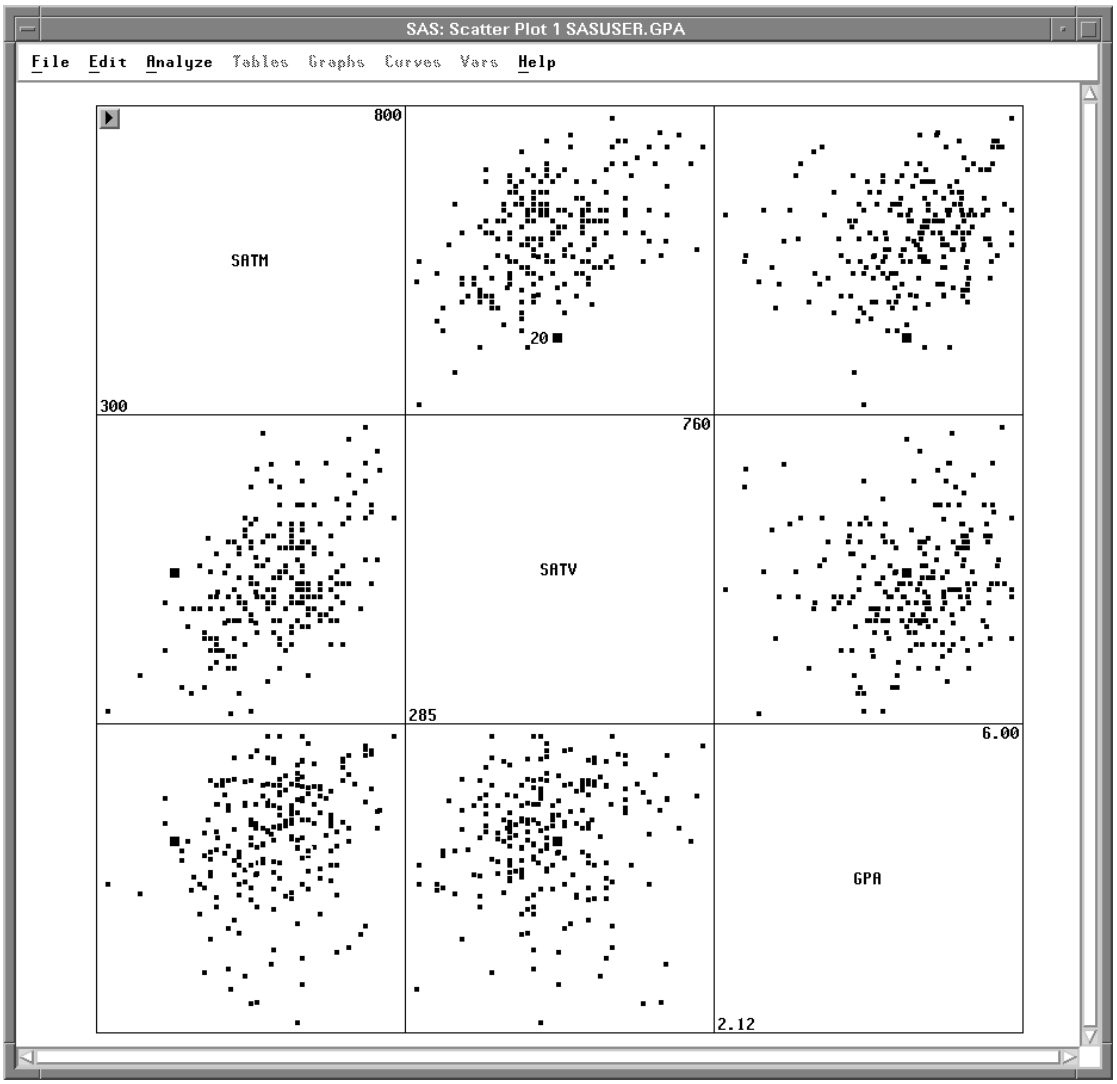


Figure 5.14. Selecting Observations in a Scatter Plot Matrix

Brushing Observations

Brushing is a dynamic method of selecting groups of observations simultaneously in all views of the data. Brushing is an effective technique for investigating multivariate data (Becker, Cleveland, and Wilks, 1987). For example, you can use brushing to find students who performed poorly on their SATs but still had relatively high grade point averages.

⟹ **Select observations with low values for SATM and SATV.**
Press the mouse button down, move the mouse, then release the mouse button to create a rectangle in the plot of **SATM** by **SATV**. This rectangle is your *brush*. The observations in the rectangle are selected. Notice that corresponding observations are also highlighted in the other plots.

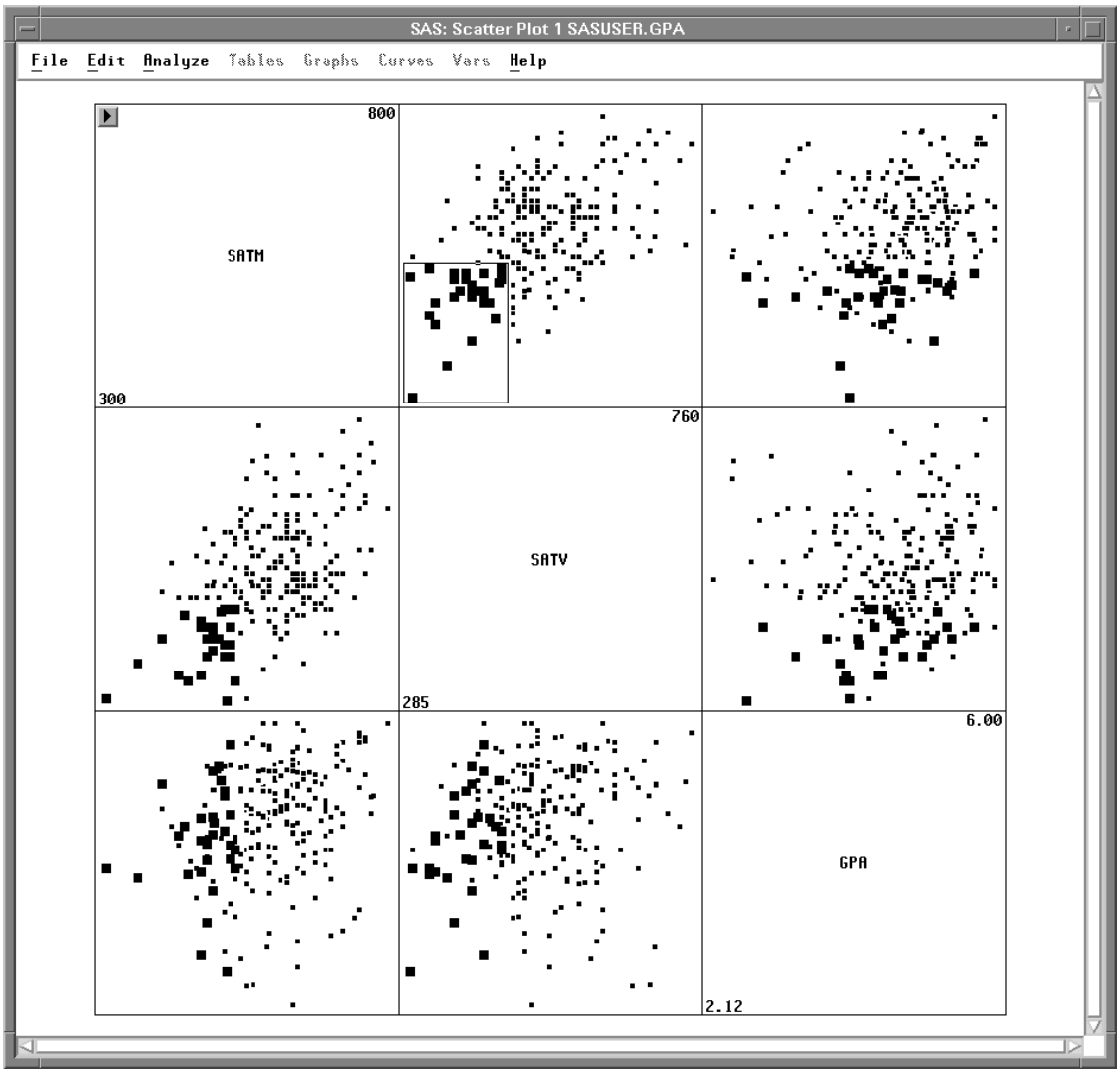


Figure 5.15. Brushing in a Scatter Plot Matrix

Examine one of the scatter plots involving **GPA**. Several of the selected observations have **GPA** values of 4 or above, indicating that SAT scores are not always good indicators of success in the school's computer science program.

You can change the size of your brush to select different observations.

⟹ **Place the cursor on the** *corner* **of the brush and drag the cursor.**
The brush changes size as you drag until you release the mouse button.

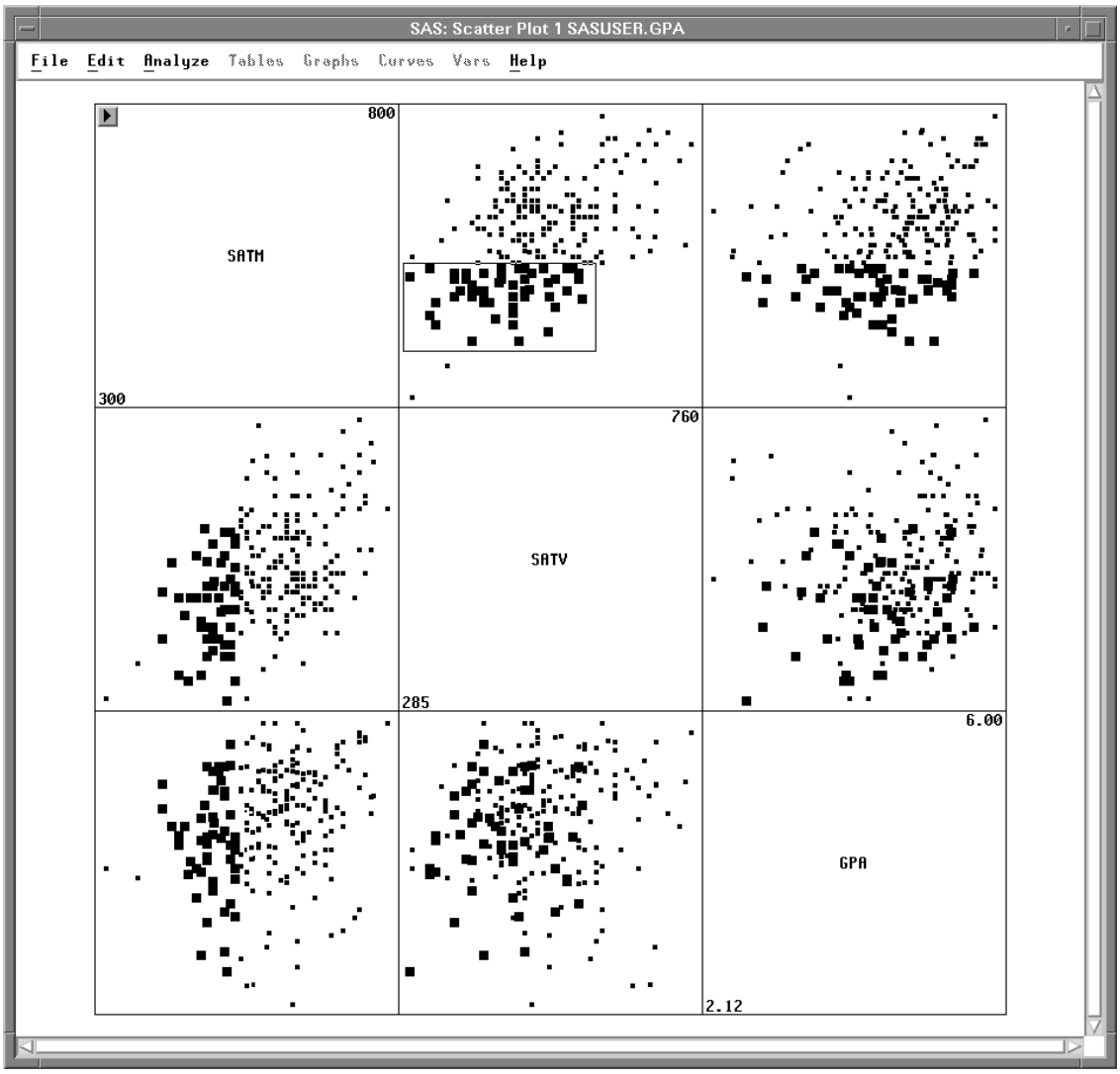


Figure 5.16. Changing the Size of a Brush

You can move the brush to select observations dynamically.

⟹ **Place the cursor in the brush and drag the brush across the plot.**
As observations enter the brush they become selected, and as they leave they are deselected. The corresponding observations in all the other scatter plots are also selected and deselected as you move the brush.

If you release the mouse button while you are moving the brush, the brush continues to move. *Throwing* the brush in this way removes the burden of eye-hand coordination, enabling you to take your eyes off the brush and more easily see its effect in other plots.

You can also brush with extended selection. This is a convenient way to select a set of observations that does not fit the rectangular shape of the brush. Extended selection, described in Chapter 1, uses the **Shift** key on most hosts.

⟹ **Using extended selection, create another brush.**

The observations that were in the previous brush remain selected.

⟹ **Using extended selection, move the brush.**
Observations become selected as they enter the brush, but they are not deselected when they leave the brush, as illustrated in Figure 5.17.

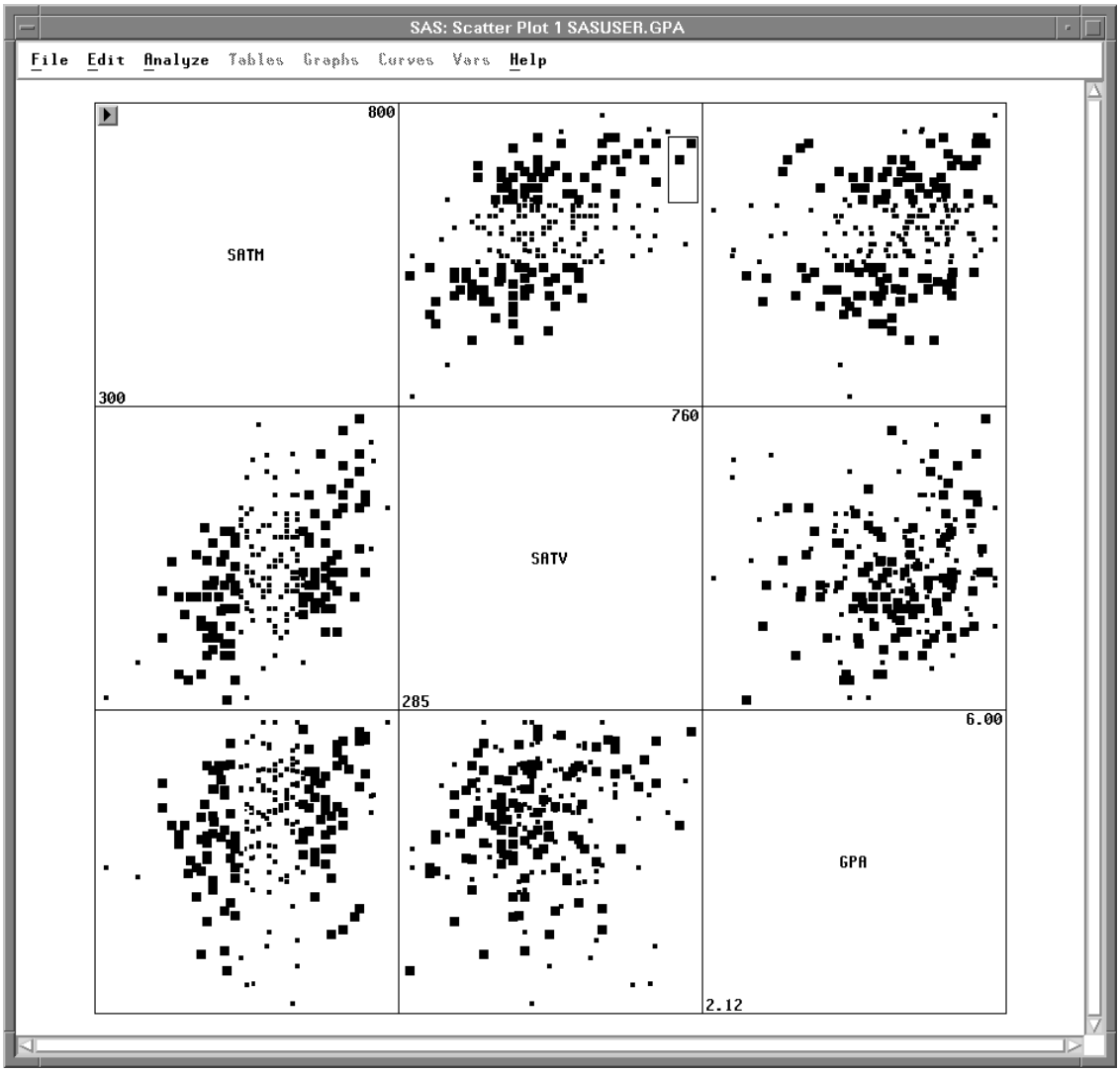


Figure 5.17. Brushing with Extended Selection

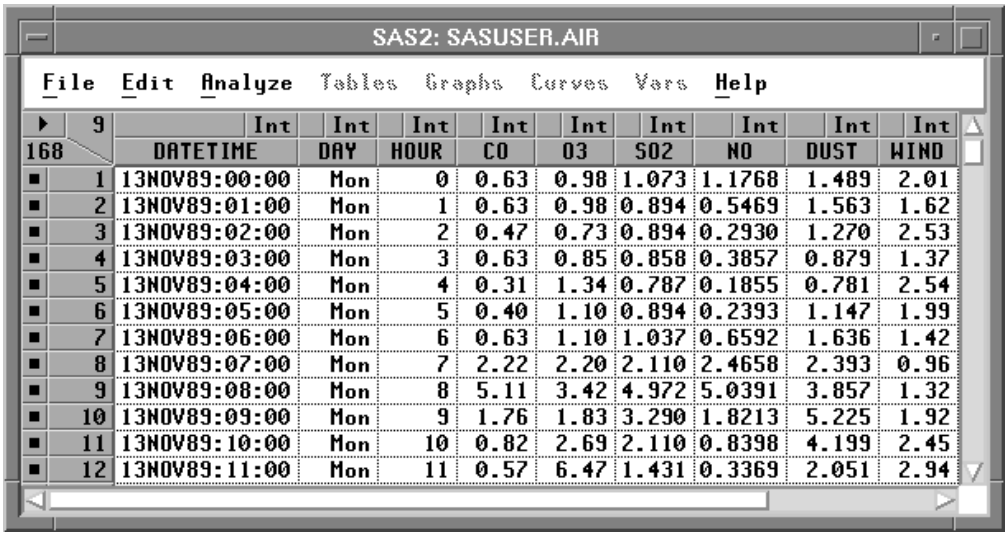
⟹ **To remove the brush, click in any empty area of the window.**
Clicking on nothing deselects all selected objects.

⊕ **Related Reading:** Scatter Plots, Chapter 35.

Line Plots

Line plots are often used to show trends over time. For example, you can explore the patterns in pollutant concentrations in the **AIR** data set by following these steps.

⟹ **Open the AIR data set.**
This data set contains measurements of air quality as indicated by concentrations of various pollutants. Among the pollutants are carbon monoxide (**CO**), ozone (**O3**), sulfur dioxide (**SO2**), nitrogen oxide (**NO**), and **DUST**.



168	DATETIME	DAY	HOUR	CO	O3	SO2	NO	DUST	WIND
1	13NOV89:00:00	Mon	0	0.63	0.98	1.073	1.1768	1.489	2.01
2	13NOV89:01:00	Mon	1	0.63	0.98	0.894	0.5469	1.563	1.62
3	13NOV89:02:00	Mon	2	0.47	0.73	0.894	0.2930	1.270	2.53
4	13NOV89:03:00	Mon	3	0.63	0.85	0.858	0.3857	0.879	1.37
5	13NOV89:04:00	Mon	4	0.31	1.34	0.787	0.1855	0.781	2.54
6	13NOV89:05:00	Mon	5	0.40	1.10	0.894	0.2393	1.147	1.99
7	13NOV89:06:00	Mon	6	0.63	1.10	1.037	0.6592	1.636	1.42
8	13NOV89:07:00	Mon	7	2.22	2.20	2.110	2.4658	2.393	0.96
9	13NOV89:08:00	Mon	8	5.11	3.42	4.972	5.0391	3.857	1.32
10	13NOV89:09:00	Mon	9	1.76	1.83	3.290	1.8213	5.225	1.92
11	13NOV89:10:00	Mon	10	0.82	2.69	2.110	0.8398	4.199	2.45
12	13NOV89:11:00	Mon	11	0.57	6.47	1.431	0.3369	2.051	2.94

Figure 5.18. **AIR** Data

⟹ **Choose Analyze:Line Plot (Y X).**
This displays the line plot variables dialog.

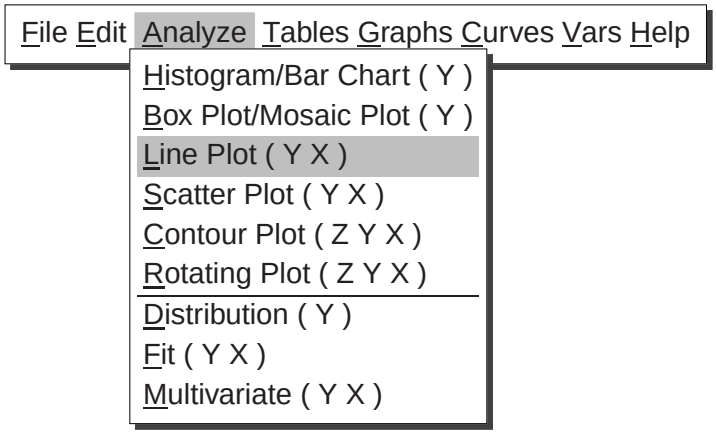


Figure 5.19. Creating a Line Plot

⟹ **Assign CO and SO2 the Y role, and DATETIME the X role.**

⟹ **Assign DATETIME the Label role also. Then click OK.**

Figure 5.20. Assigning Line Plot Variables

This creates a line plot with one line for each **Y** variable.

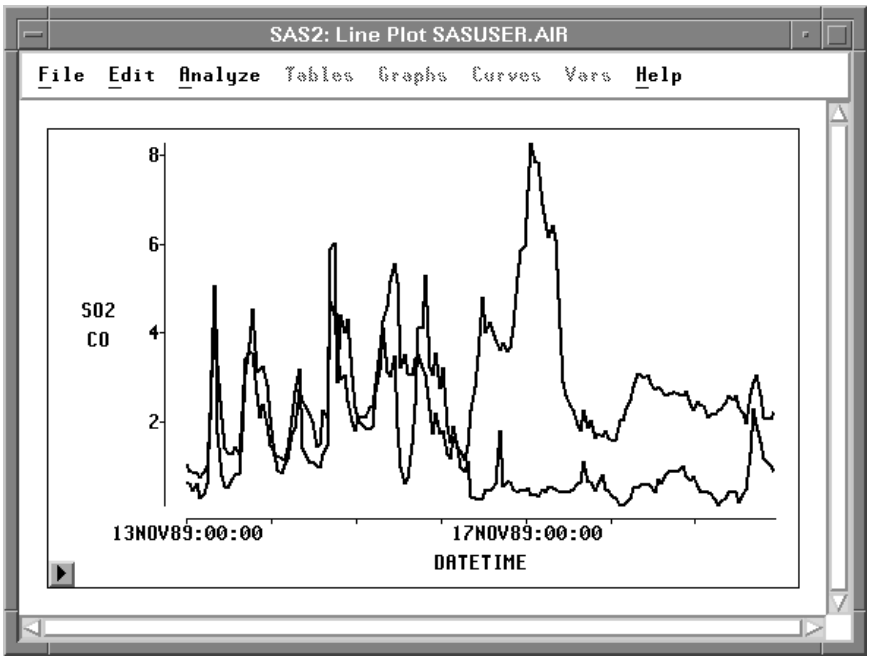


Figure 5.21. Line Plot

To associate lines with variables, simply select the variable.

⟹ **Click on the SO2 variable.**
This highlights both the variable and the corresponding line.

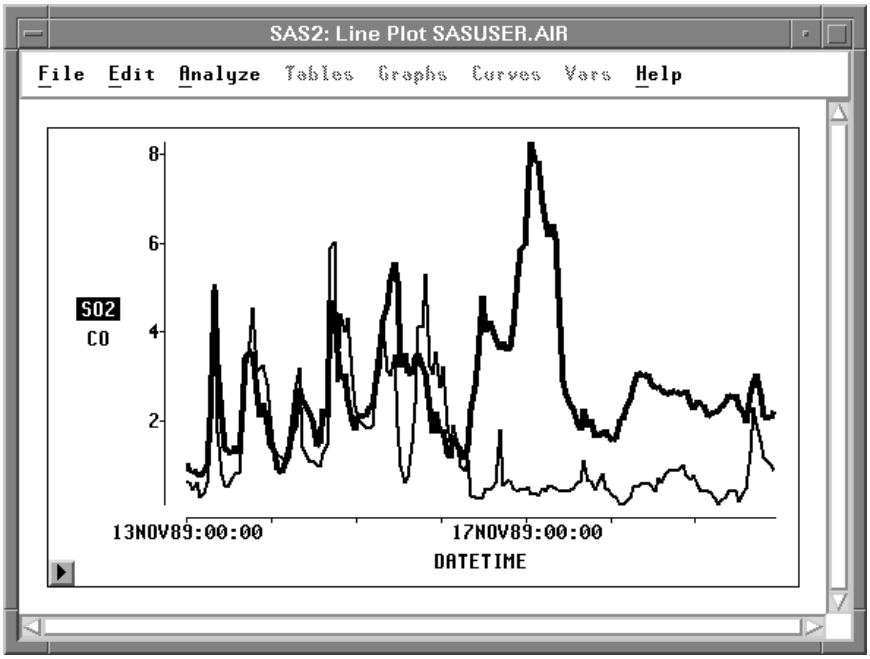


Figure 5.22. **SO2** Selected

By clicking on the variables, you can see that the **SO2** concentration rises to a peak on the 17th of November and then falls. The **CO** concentration shows a regular pattern of peaks and valleys up until the 16th; then it falls also.

To show more information, you can add observation markers to the line plot.

⟹ **Click on the menu button in the lower left corner of the plot. Choose Observations.**

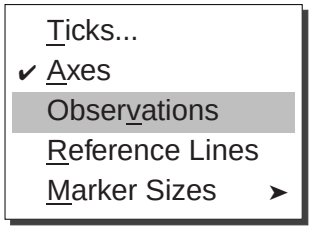


Figure 5.23. Line Plot Pop-up Menu

This displays the line plot with observation markers.

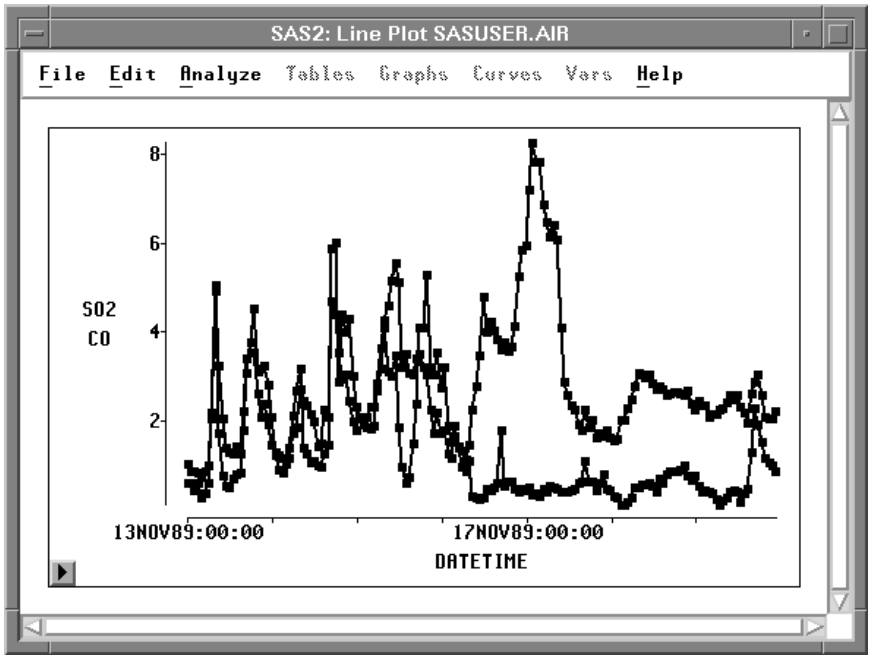


Figure 5.24. Line Plot with Observations

⟹ **Point and click to identify observations with the highest pollutant concentrations.**

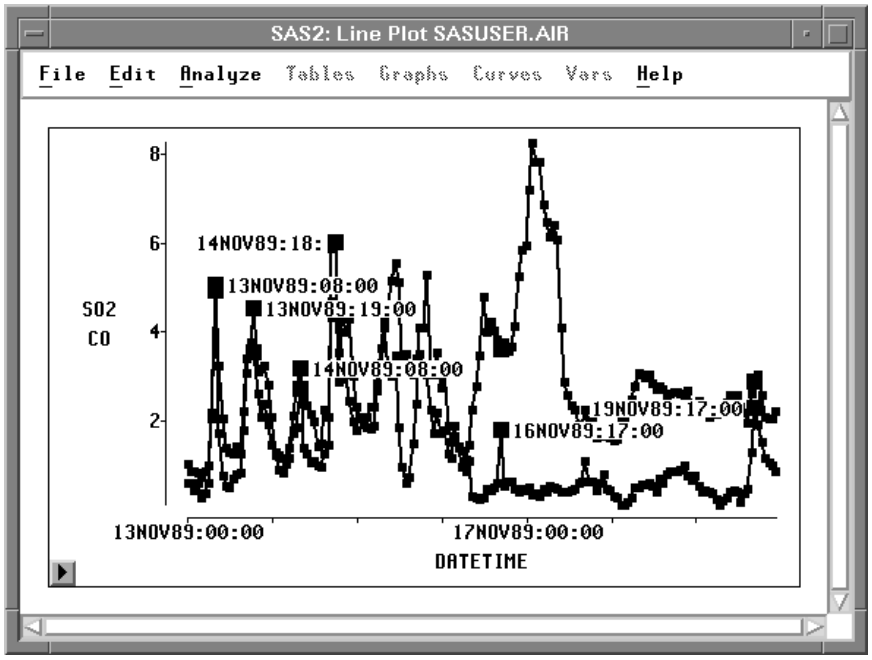


Figure 5.25. Identifying Observations

Most of the peaks for **CO** occur in the morning and evening, around hours 08:00 or 18:00. Carbon monoxide pollution is often caused by automobiles, so these peaks might be caused by rush-hour traffic.

The **SO2** concentration follows a different pattern. Sulfur dioxide is a pollutant given off by power plants. Perhaps there was a peak demand for electricity on the 17th.

The drop in pollutants after the 17th can be partly explained by noting that the 18th and 19th were Saturday and Sunday. The weekend eliminates rush-hour traffic patterns. However, the **CO** level dropped on the 16th also, which was Thursday. There is an additional factor at work here.

⟹ **Choose Edit:Windows:Renew to re-create the line plot.**

⟹ **Add WIND to the Y variable list. Then click OK.**

Figure 5.26. Adding **WIND** Variable

⟹ **In the line plot, click on the WIND variable.**

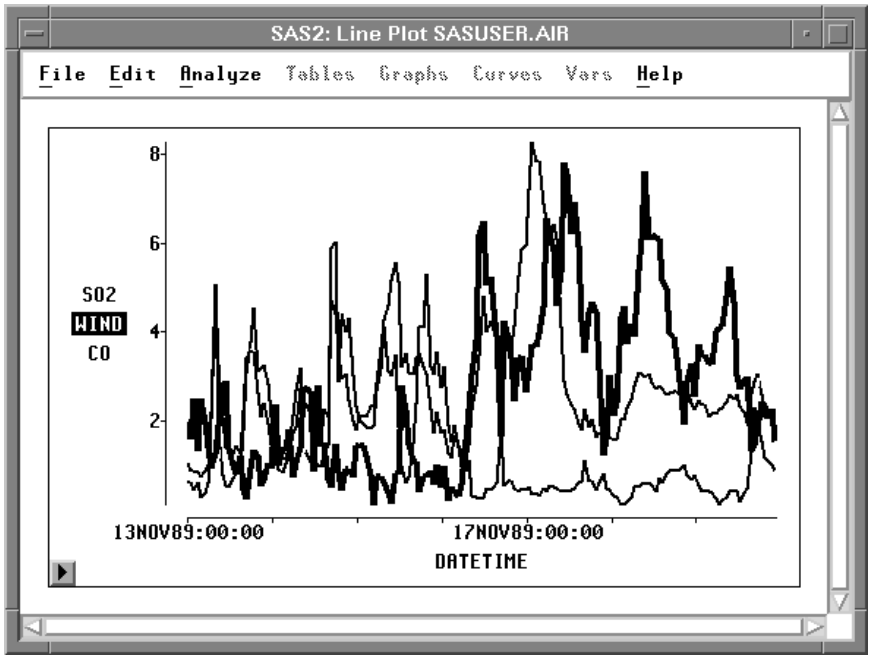


Figure 5.27. WIND Speed

Not only were the 18th and 19th a weekend, but there were high winds on the 16th, 17th, 18th, and 19th. These winds cleared much of the pollutants from the local atmosphere.

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

⊕ **Related Reading:** Scatter Plots, Chapter 35.

⊕ **Related Reading:** Line Plots, Chapter 34.

References

Becker, R.A., Cleveland, W.S., and Wilks, A.R. (1987), "Dynamic Graphics for Data Analysis," *Statistical Science*, 2 (4), 355–382.

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