

Exploring Data With Base SAS® Software

Thomas J. Winn Jr., Texas State Comptroller's Office, Austin, Texas

Abstract

Statistical methods are tools which are used to summarize and analyze data. Exploratory data analysis is the application of graphical and statistical techniques to discover the structure of data. The goal of exploratory data analysis is to characterize the data and to reveal fundamental relationships among them. It is quick, dynamic, and highly interactive. Furthermore, exploratory data analysis is not just for use by professional statisticians -- the methods also are used by scientists, engineers and many other types of researchers. This paper explains how to produce and to interpret scatter plots, histograms, stem-and-leaf plots, box-and-whisker plots, and various descriptive statistics, using Base SAS® software.

Introduction

Originally, the SAS System was a combination of programs for performing statistical analysis on data. Since then, the SAS System has grown in ways which have made it useful to non statisticians, as well as having increased its value to its earliest audience of users. Most SAS users, including many without substantial expertise in statistics, have an occasional need to utilize some of the statistical and graphical capabilities of the SAS System. They want to examine their data, but without becoming involved in advanced statistical procedures. The present paper is addressed to this group of users.

Exploratory data analysis is the interactive use of statistical and graphical procedures to uncover the composition of data; that is, to identify their general characteristics and relationships. Data exploration is a *process* in which raw data become comprehensible information through a sequence of activities, each of which must be adapted according to the outcomes of the preceding steps. It is noted that SAS software now includes JMP®, SAS/INSIGHT®, and SAS/LAB®, which were expressly designed to implement exploratory techniques. However, many SAS users do not have access to these components. This paper is intended to present some of the basic tools for data exploration using elements of Base SAS software, whether they are used in a non interactive mode (such as batch) or in an interactive mode (such as SAS Display Manager).

Types of Data

To begin with, there are two basic levels of data: qualitative, and quantitative. This is not the same thing as the difference between character and numeric variables in SAS. Qualitative types of measurement may be either character or numeric, the essential idea is that they involve some type of mutually

exclusive, categorical classification, which may or may not possess an inherent order. Numbers may be used qualitatively as coded values for names, but arithmetic with the values would be meaningless. Examples of qualitative types of measurement include taxonomic names (such as values for sex, race, region, political preference, etc.), or ordinal scales (ordered categories such as: always - sometimes - never, first - second - third - fourth - fifth, etc.). Quantitative types of measurement use numbers as cardinal magnitudes.

With qualitative types of data, the data analysis methods are mostly limited to frequency tables, bar charts, and pie charts. However, quantitative types of measurement permit the use of a greater variety of tools.

Overview of Descriptive Statistics

The goal of data exploration is to comprehend the distribution of the values of the variables which comprise the data, and to identify some of the important ways in which the variables appear to be related. Data summarization leans heavily on statistical measures which pertain to central tendency, dispersion, and shape of the data distribution, as well as on a few graphical techniques for data visualization.

Central tendency refers to a typical value from the distribution. Three commonly-used measures of central tendency are the arithmetic mean (or average value), the median (the middle-most value), and the mode (the value which occurs most frequently). In the case of qualitative data, the mode is commonly used as the central tendency measure.

Dispersion refers to the spread of the data values, usually with respect to a particular measure of central tendency. Some measures of dispersion are the range, the variance, the standard deviation, the coefficient of variation, and the interquartile range. Range is the difference between the smallest and the largest data values. The standard deviation and the variance both indicate the variability (or the amount of concentration) of the data values with respect to the mean. The coefficient of variation is $100 \times \text{standard deviation} / \text{mean}$. The interquartile range is the distance between the particular data value below which the bottom one-fourth of the data values are found (first quartile, Q1), and the particular data value above which the top one-fourth of the data values are located (third quartile, Q3). There is no measure of dispersion for non-ordinal, qualitative types of data.

In addition to central tendency and dispersion, other properties which are useful in describing the shape of a distribution are

skewness, kurtosis, and the presence of outliers, gaps and multiple peaks. *Skewness* is a measure of the symmetry (or lack thereof) of the distribution. In a perfectly symmetrical distribution, the mean, the median, and the mode coincide. A distribution is said to be *skewed* whenever the data values are clustered more at one end than at the other, so that its scatter plot seems to lean unevenly towards one side. A skewness measure would be zero when the distribution is symmetric; it would be positive when more data points are clustered at the lower end than at the upper end (the mean and the median are greater than the mode); and it would be negative when more data points are clustered at the upper end than at the lower end (the mean and the median are less than the mode). *Kurtosis* is a measure of the flatness of the distribution. A very large kurtosis number would mean that some of the data values are much farther away from the mean than most of the other data values; when this happens, the distribution is said to have a "heavy tail". *Outliers* are data values which are far away from the rest of the data.

Preliminary Examination of Data

Data analysis begins with a cursory review of the raw data. Do the values of the variables correspond to quantitative or qualitative types of measurements? Do the data values conform to reasonable expectations? Do any of the values contain obvious typographical errors, or appear to be out-of-range? Does it seem as though certain observations may be missing? Resolve any apparent data errors before proceeding to the next step.

After reading the raw data into a SAS data file, carefully examine the SAS Log. The SAS System will identify many data errors which may have escaped the prior notice of the data analyst. The notes and error messages generated by the SAS System upon the creation of a SAS data set are very instructive.

It is important to document the newly-created data set, and to begin examining the elemental properties of the data. It is a good idea to use a PROC CONTENTS step (or, alternatively, a CONTENTS statement in a PROC DATASETS step) together with a PROC PRINT (or PROC FSVIEW) step, whenever data are introduced to the SAS System. If the data are so numerous as to make it impracticable to review a complete listing of the data, then use the RANUNI function to create a random selection of observations from the data, and in conjunction with the PRINT procedure on the sample.

```
PROC CONTENTS DATA=data-set-name;
```

```
PROC PRINT DATA=data-set-name;
  WHERE RANUNI(0) <= 0.01;
  TITLE "1% SAMPLE FROM DATA SET";
```

Review the information generated by the CONTENTS and the PRINT (or FSVIEW) procedures. Do the data attributes

(variable name, type, length, informat, format, label) agree with what had been anticipated? Do the numeric variables take on only a limited number of distinct values, or do they have a very large number of values? Are there any aberrant values; that is, do the data deviate unreasonably from the typical pattern? If there are any data errors, substitute corrected values for them.

Now, run PROC MEANS to calculate simple descriptive statistics for numeric variables in a SAS data set. If no particular statistics are specified as options on the PROC MEANS statement then, for each numeric variable, the variable name, number of observations, mean, standard deviation, minimum value, and maximum value will be reported.

```
PROC MEANS DATA=data-set-name;
  VAR variables-list;
```

If desired, PROC MEANS also will report the variance, the coefficient of variation, the range, the skewness, and the kurtosis (and more, if desired). If observations can be grouped together using certain variables, then a CLASS statement can be used to obtain summary statistics across each classification grouping (without sorting the data!).

```
PROC MEANS DATA=data-set-name
  N MIN MAX RANGE MEAN VAR
  STD CV SKEWNESS KURTOSIS;
  VAR variables-list;
  CLASS class-variables-list;
```

Run PROC FREQ to obtain a one-way frequency table of counts and percentages. This report is particularly helpful for analyzing qualitative types of data.

```
PROC FREQ DATA=data-set-name;
  TABLES variables-list;
```

Also, run PROC CHART to produce a visual summary of the data. Printer graphics may not be presentation quality, but they do not require much time or special equipment, and their results can be very powerful. PROC CHART can be used for displaying both qualitative and quantitative types of data.

```
PROC CHART DATA=data-set-name;
  VBAR variables-list / option;
```

or

```
PROC CHART DATA=data-set-name;
  HBAR variables-list / option;
```

In using PROC CHART, the data analyst may want to take control of the horizontal axis, to ensure that gaps in numeric values are noted, and of the vertical axis, to facilitate comparisons between similar graphs.

```
PROC CHART DATA=data-set-name;
  VBAR variables-list / MIDPOINTS=xx TO yy
  BY zz AXIS=uu vv;
```

A *histogram* is a particular bar chart in which the range of data values is divided into intervals of equal length, and in which bars are used to represent the frequency of the observations in each interval. The preceding syntax will produce a histogram. In the VBAR statement above, an alternative to the MIDPOINTS=... option would be to use the LEVELS=... option. In either case, the number of intervals should be chosen so as to display just enough detail as will be meaningful to the data analyst, without being overwhelming.

If observations can be grouped together using certain variables, then it also will be useful to picture the data using a block chart.

```
PROC CHART DATA=data-set-name;
  BLOCK variable / GROUP=class-variable;
```

Data Exploration Using PROC UNIVARIATE

The most useful exploratory procedure is PROC UNIVARIATE. This comprehensive procedure can be used to generate descriptive statistics, a frequency table, a list of extreme values, some interesting plots, and a comparison of the cumulative frequency distribution with a normal distribution. To produce box-and-whisker plots and stem-and-leaf displays, invoke PROC UNIVARIATE using the PLOT option.

```
PROC UNIVARIATE DATA=data-set-name
  FREQ PLOT;
  VAR variables-list;
```

A *stem-and-leaf display* is one way to convey the shape of the distribution, as well as the value of each observation of the variable. A stem-and-leaf display is similar to a horizontal bar chart, except that instead of using bars, the next digit of the number after the "stem" is used. To interpret a stem-and-leaf display, follow the instructions printed beneath the display.

Box-and-whisker plots (also referred to as "boxplots" and "schematic plots") present a visual representation of some of the more important summary statistics. The top and bottom of the box describe the interquartile range [the difference between the 25th(Q1) and the 75th(Q3) percentiles] of the distribution. The horizontal line inside the box represents the median value [the 50th percentile(Q2)], and the plus sign indicates the mean. The vertical lines emanating from the box (called "whiskers") extend up to 1.5 times the interquartile range [that is, from Q1 down to $Q1 - 1.5*(Q3 - Q1)$, and from Q3 up to $Q3 + 1.5*(Q3 - Q1)$]. A data value which is more than 1.5 interquartile ranges but within 3 interquartile ranges is

represented by a zero. Data values which exceed 3 interquartile ranges are represented by asterisks.

In Version 6 of SAS, if the magnitudes of the variables are comparable, full-page, side-by-side boxplots will be produced whenever PROC UNIVARIATE is invoked with the PLOT option and with a BY statement.

```
PROC UNIVARIATE DATA=data-set-name FREQ
  PLOT;
  VAR variables-list;
  BY class-variable;
```

With a little practice, the data analyst can use these special plots to visualize the essential features of a distribution of data values.

Example #1

Consider the following data (Friendly, pp. 4-8):

```
DATA FRIENDLY;
  INPUT I SET1 SET2 SET3 SET4;
  CARDS;
  1 40.50 41.64 35.00 44.50
  2 41.50 58.36 37.00 45.00
  3 42.50 42.29 42.00 45.50
  4 43.50 57.71 53.90 46.00
  5 44.50 42.93 53.00 46.50
  6 45.50 57.07 50.60 47.00
  7 46.50 43.57 50.50 47.50
  8 47.50 56.43 53.80 48.00
  9 48.50 44.21 52.50 48.50
  10 49.50 55.79 53.60 49.00
  11 50.50 44.86 50.40 49.50
  12 51.50 55.14 52.20 50.00
  13 52.50 45.50 52.70 50.50
  14 53.50 54.50 52.40 51.00
  15 54.50 46.14 52.70 51.50
  16 55.50 53.86 51.40 52.00
  17 56.50 46.79 53.80 52.50
  18 57.50 53.21 52.90 53.00
  19 58.50 47.43 56.81 72.71
  20 59.50 52.57 42.79 49.79
  ;
```

The four variables, SET1-SET4, have the interesting property of sharing the same mean ($\mu=50$) and standard deviation ($\sigma=5.92$), yet the distributions of their values certainly do not appear to be the same. What are their differences?

First of all, here is the output from PROC MEANS for this SAS data set:

Variable	N	Mean	Std Dev	Minimum	Maximum
SET1	20	50.000000	5.9160798	40.500000	59.500000
SET2	20	50.000000	5.9175546	41.640000	58.360000
SET3	20	50.000000	5.9159917	35.000000	56.810000
SET4	20	50.000000	5.9162497	44.500000	72.710000

We notice that the maxima and minima for the four variables differ from one another.

Here are the stem-and-leaf displays and box-and-whisker plots for these data, obtained from PROC UNIVARIATE:

```
Variable=SET1
Stem Leaf      #      Boxplot
 6 0            1
 5 6688         4
 5 002244       6
 4 6688         4
 4 02244        5
-----
Multiply Stem.Leaf by 10**+1
```

```
Variable=SET2
Stem Leaf      #      Boxplot
58 4            1
56 417          3
54 518          3
52 629          3
50              1
48              1
46 184          3
44 295          3
42 396          3
40 6            1
-----
Multiply Stem.Leaf by 10**+1
```

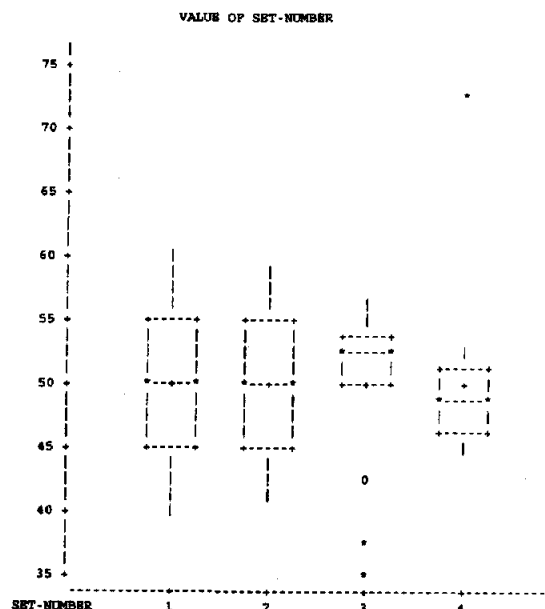
```
Variable=SET3
Stem Leaf      #      Boxplot
 5 7            1
 5 001122233334444 15
 4              2
 4 23           2
 3 57           2
-----
Multiply Stem.Leaf by 10**+1
```

```
Variable=SET4
Stem Leaf      #      Boxplot
 7 3            1
 6              1
 6              1
 5              1
 5 000012223    9
 4 566678899    9
 4 4            1
-----
Multiply Stem.Leaf by 10**+1
```

Observe that boxplots provide very little information about the data values in the vicinity of the distribution's middle values. Some experienced data analysts have learned to compare the length of the whiskers to the length of the box – whiskers which are too short may be a warning of anomalies near the center.

PROC UNIVARIATE also generates statistical measures which pertain to the central tendency, dispersion, and shape of the data distribution. These statistics are not displayed here, due to lack of space, but they are important elements of the data analysis.

Here are side-by-side boxplots for these data:



This example demonstrates a pitfall of relying too much on the mean and standard deviation to characterize abnormal data – numerical summaries of data can be misleading!

The data values for variable SET1 are uniformly distributed on the interval [40.5, 59.5]. The mean and the median are identical, and the distribution is symmetric (skewness=0). The negative kurtosis measure indicates that the tails of this distribution are lighter than for a normal distribution.

The observations of SET2 are distributed uniformly over two intervals, with a substantial gap separating the two clusters of data. As with SET1, the distribution is symmetric, and the kurtosis measure is negative.

The data values for variable SET3 are distributed less evenly than those of either SET1 or SET2. The mean and the median are distinct from one another, and the negative skewness measure indicates that more data points are clustered at the upper end than at the lower end of the distribution. The positive kurtosis measure indicates that the tails of this distribution are heavier than for a normal distribution. Indeed, we notice that there are some small data values which are fairly distant from the mean, compared to other data values.

SET4 has data values which are *almost* uniformly distributed over the interval [44.5, 53.0]. We notice that there are two irregular values, 49.79 and 72.71. The outliers cause the mean to be greater than the median. The positive skewness measure indicates that data values located to the right of the mean are more spread out than the data values to the left of the mean. The positive kurtosis measure (which is larger than the kurtosis of SET3) denotes the heavy tail of this distribution, which is attributable to the larger deviant value.

Examining Relationships Between Variables

With quantitative data, it often is important to determine whether or not a relation exists between two or more variables. And, if they are related, it also is desirable to measure the strength of the relationships among them. This would be useful, for example, if one was trying to estimate the values of one variable from known or assumed data values of other variables. A measure of the strength of the relationship between two variables is the *correlation coefficient*, which is a number between -1 and +1. Positive correlation coefficients indicate a direct relationship, and negative coefficients indicate an inverse relationship. You are cautioned that just because two variables may be highly correlated, this does not imply that a cause-and-effect relation necessarily exists between them.

PROC CORR will compute correlation coefficients between all pairs of variables specified in the VAR list:

```
PROC CORR DATA=data-set-name;
VAR variables-list;
```

Besides printing correlation coefficients for each pair of variables, PROC CORR also determines associated significance probabilities for each coefficient. These p-values are for testing the null hypothesis that the variables actually have zero correlation.

A *scatter plot* is a graphic representation of the relationship between a pair of quantitative variables. To create Scatter plots with Base SAS, PROC PLOT is used, with a PLOT statement for each pair of variables.

```
PROC PLOT DATA=data-set-name;
PLOT variable1*variable2= '*';
PLOT variable3*variable4= '*';
```

Example #2

Consider the following data (from the SAS Sample Library, member PLOTAB2):

```
DATA CRIME;
  TITLE 'Crime Rates Per 100,000 Population by State';
  INPUT STATE $ 1-15 POSTCODE $ MURDER RAPE ROBBERY
        ASSAULT BURGLARY LARCENY AUTO;
  CARDS;
Alabama AL 14.2 25.2 96.8 278.3 1135.5 1881.9 280.7
Alaska AK 10.8 51.6 96.8 284.0 1331.7 3369.8 753.3
Arizona AZ 9.5 34.2 138.2 312.3 2346.1 4467.4 439.5
Arkansas AR 8.8 27.6 83.2 203.4 972.6 1862.1 183.4
California CA 11.5 49.4 287.0 358.0 2139.4 3499.8 663.5
Colorado CO 6.3 42.0 170.7 292.9 1935.2 3903.2 477.1
Connecticut CT 4.2 16.8 129.5 131.8 1346.0 2620.7 593.2
Delaware DE 6.0 24.9 157.0 194.2 1682.6 3678.4 467.0
Florida FL 10.2 39.6 187.9 449.1 1859.9 3840.5 351.4
Georgia GA 11.7 31.1 140.5 256.5 1351.1 2170.2 297.9
Hawaii HI 7.2 25.5 128.0 64.1 1911.5 3920.4 489.4
Idaho ID 5.5 19.4 39.6 172.5 1050.8 2599.6 237.6
Illinois IL 9.9 21.8 211.3 209.0 1085.0 2828.5 528.6
Indiana IN 7.4 26.5 123.2 153.5 1086.2 2498.7 377.4
Iowa IA 2.3 10.6 41.2 89.8 812.5 2685.1 219.9
Kansas KS 6.6 22.0 100.7 180.5 1270.4 2739.3 244.3
Kentucky KY 10.1 19.1 81.1 123.3 872.2 1662.1 245.4
Louisiana LA 15.5 30.9 142.9 335.5 1165.5 2469.9 337.7
Maine ME 2.4 13.5 38.7 170.0 1253.1 2350.7 246.9
Maryland MD 8.0 34.8 292.1 358.9 1400.0 3177.7 428.5
Massachusetts MA 3.1 20.8 169.1 231.6 1532.2 2311.3 1140.1
Michigan MI 9.3 38.9 261.9 274.6 1522.7 3159.0 545.5
Minnesota MN 2.7 19.5 85.9 85.8 1134.7 2559.3 343.1
```

```
Mississippi MS 14.3 19.6 65.7 189.1 915.6 1239.9 144.4
Missouri MO 9.6 28.3 189.0 233.5 1318.3 2424.2 378.4
Montana MT 5.4 16.7 39.2 156.8 804.9 2773.2 309.2
Nebraska NE 3.9 18.1 64.7 112.7 760.0 2316.1 249.1
Nevada NV 15.8 49.1 323.1 355.0 2453.1 4212.6 559.2
New Hampshire NH 3.2 10.7 23.2 76.0 1041.7 2343.9 293.4
New Jersey NJ 5.6 21.0 180.4 185.1 1435.8 2774.5 511.5
New Mexico NM 8.8 39.1 109.6 343.4 1418.7 3008.6 259.5
New York NY 10.7 29.4 472.6 319.1 1728.0 2782.0 745.8
North Carolina NC 10.6 17.0 61.3 218.3 1184.1 2037.8 192.1
North Dakota ND 0.9 9.0 13.3 43.8 466.1 1843.0 144.7
Ohio OH 7.8 27.3 190.5 181.1 1216.0 2696.8 400.4
Oklahoma OK 8.6 29.2 73.8 205.0 1288.2 2228.1 326.8
Oregon OR 4.9 39.9 124.1 286.9 1636.4 3506.1 388.9
Pennsylvania PA 5.6 19.0 130.3 128.0 877.5 1624.1 333.2
Rhode Island RI 3.6 10.5 86.5 201.0 1489.5 2844.1 791.4
South Carolina SC 11.9 33.0 105.9 485.3 1613.6 2342.4 245.1
South Dakota SD 2.0 13.5 17.9 155.7 570.5 1704.4 147.5
Tennessee TN 10.1 29.7 145.8 203.9 1259.7 1776.5 314.0
Texas TX 13.3 33.8 152.4 208.2 1603.1 2988.7 397.6
Utah UT 3.5 20.3 68.8 147.3 1171.6 3004.6 334.5
Vermont VT 1.4 15.9 30.8 101.2 1348.2 2201.0 265.2
Virginia VA 9.0 23.3 92.1 165.7 986.2 2521.2 226.7
Washington WA 4.3 39.6 106.2 224.8 1605.6 3386.9 360.3
West Virginia WV 6.0 13.2 42.2 90.9 597.4 1341.7 163.3
Wisconsin WI 2.8 12.9 52.2 63.7 866.9 2614.2 220.7
Wyoming WY 5.4 21.9 39.7 173.9 811.6 2772.2 282.0
```

Now, here is the output from running PROC CORR against these data:

Crime Rates Per 100,000 Population by State

Correlation Analysis

7 'VAR' Variables: MURDER RAPE ROBBERY ASSAULT BURGLARY
LARCENY AUTO

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
MURDER	50	7.4440	3.8668	372.2	0.9000	15.8000
RAPE	50	25.7340	10.7596	1286.7	9.0000	51.6000
ROBBERY	50	124.1	88.3486	6204.6	13.3000	472.6
ASSAULT	50	211.3	100.3	10565.0	43.8000	485.3
BURGLARY	50	1291.9	432.5	64595.2	446.1	2453.1
LARCENY	50	2671.3	725.9	133564	1239.9	4467.4
AUTO	50	377.5	193.4	18876.3	144.4	1140.1

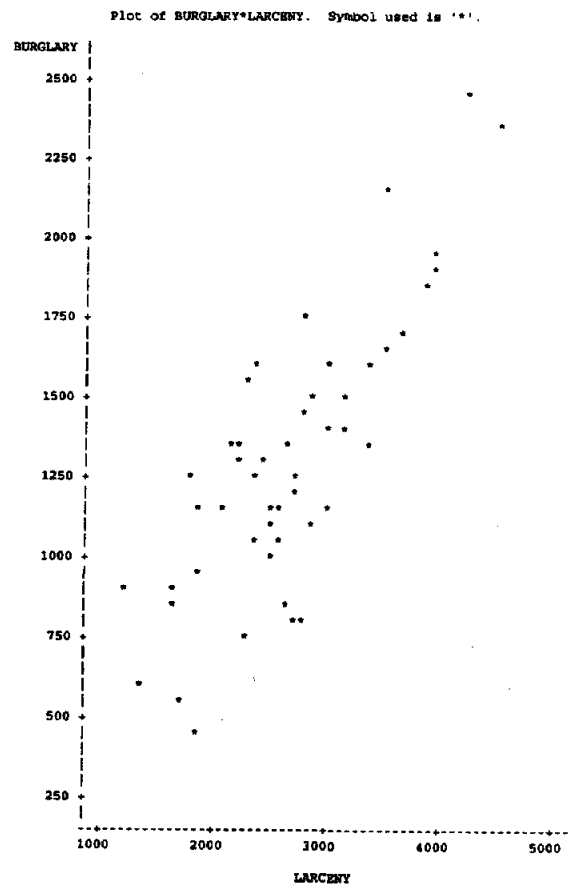
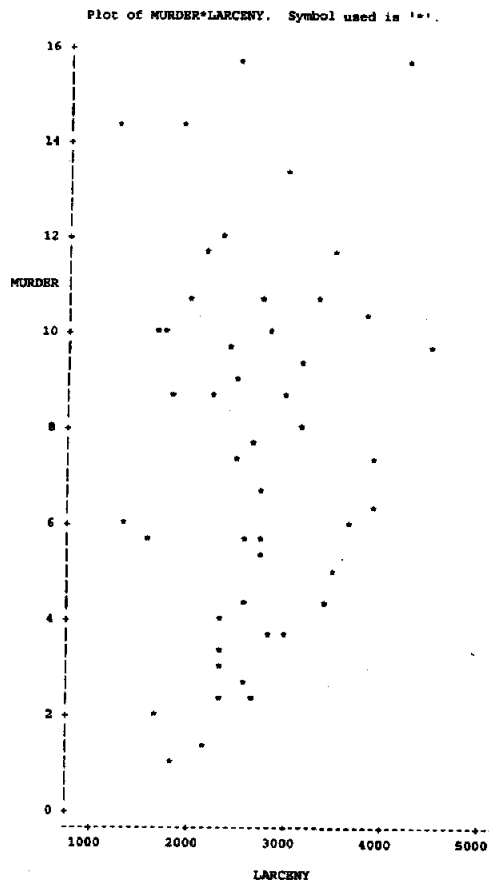
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 50

	MURDER	RAPE	ROBBERY	ASSAULT
MURDER	1.00000 0.0	0.60122 0.0001	0.48371 0.0004	0.64855 0.0001
RAPE	0.60122 0.0001	1.00000 0.0	0.59188 0.0001	0.74026 0.0001
ROBBERY	0.48371 0.0004	0.59188 0.0001	1.00000 0.0	0.55708 0.0001
ASSAULT	0.64855 0.0001	0.74026 0.0001	0.55708 0.0001	1.00000 0.0
BURGLARY	0.38582 0.0057	0.71213 0.0001	0.63724 0.0001	0.62291 0.0001
LARCENY	0.10192 0.4813	0.61399 0.0001	0.44674 0.0011	0.40436 0.0036
AUTO	0.06881 0.6349	0.34890 0.0130	0.59068 0.0001	0.27584 0.0525
	BURGLARY	LARCENY	AUTO	
MURDER	0.38582 0.0057	0.10192 0.4813	0.06881 0.6349	
RAPE	0.71213 0.0001	0.61399 0.0001	0.34890 0.0130	
ROBBERY	0.63724 0.0001	0.44674 0.0011	0.59068 0.0001	
ASSAULT	0.62291 0.0001	0.40436 0.0036	0.27584 0.0525	
BURGLARY	1.00000 0.0	0.79212 0.0001	0.55795 0.0001	
LARCENY	0.79212 0.0001	1.00000 0.0	0.44418 0.0012	
AUTO	0.55795 0.0001	0.44418 0.0012	1.00000 0.0	

Notice the relatively large correlations between the pairs of variables: BURGLARY & LARCENY, and ASSAULT & RAPE;

and the relatively small correlations between AUTO & MURDER, and LARCENY & MURDER.

Here are a couple of scatter plots which reflect the indicated strength-of-relationship measures:



Conclusion

Base SAS software includes several easy-to-use graphical and statistical procedures which can be used to summarize and analyze data. The fundamental methods of exploratory data analysis can be used to uncover the shape of a distribution of data values. In order to comprehend a set of data values, it is not good enough to rely solely on numerical summary statistics for central tendency and dispersion.

References

- Michael Friendly (1991), SAS System for Statistical Graphics, First Edition, Cary, NC: SAS Institute Inc.
- SAS Institute Inc. (1990), SAS Procedures Guide, Version 6, Third Edition, Cary, NC: SAS Institute Inc.
- Sandra D. Schlotzhauer & Ramon C. Littell (1987), SAS System for Elementary Statistical Analysis, Cary, NC: SAS Institute Inc.
- John W. Tukey (1977), Exploratory Data Analysis, Reading, MA: Addison-Wesley.

SAS, SAS/INSIGHT, SAS/LAB, and JMP are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.