SUPPLEMENTARY EXERCISES, CHAPTER 2

SUMMARIZING DATA

S2.1. Figure 1 shows output from a SAS/INSIGHT distribution analysis of the ages at death of 492 dogs. The units are years.

a. Describe the distribution of the ages as displayed in the histogram.

b. Choose one appropriate summary measure of "center" of the distribution. Give its interpretation.

c. Is the mean an appropriate measure of the "center" of the distribution? Why or why not?

d. Choose one appropriate summary measure of spread of the distribution. Describe what it means,

e. Is the standard deviation an appropriate measure of the spread of the distribution? Why or why not?

f. If you have to give a measure, based on this distribution output, to predict the age at death of a dog, what would it be? Why did you choose this measure?

S2.2. If a data distribution is symmetric, you know that the mean equals the median. Is the converse true: that is, if the mean of a data distribution equals its median, must the distribution be symmetric? Prove it's true or give an example to show it isn't.

S2.3. The half-life of a piece of uranium is the amount of time it takes for half the uranium atoms to decay. Suppose you have a data set consisting of the time to decay of each atom in a piece of uranium. What statistical measure of location for these times corresponds to the half-life of the uranium? Explain why.

S2.4. Figure 2 displays a histogram of a set of data from a stationary process. Your friend, Bill says, "I like box and whisker plots better than histograms, so I'll use one to present these data."

a. Bill asks your opinion of his idea. What do you tell him? Explain your reasoning.

b. Summarize these data using appropriate measures.

S2.5. The lower curve in Figure 3 is a line plot of the median salaries, in 1983 dollars, of major league baseball players for the 1983 through 1998 seasons. The upper curve is a line plot of the corresponding mean salaries.

a. Based on the plot of the medians, would you conclude that the process that produces baseball salaries is stationary? What is your answer if you base your conclusion on the plot of the means? Justify your answers.

b. Give a plausible explanation for what might be happening to the distribution of baseball salaries to account for the increasing means and level medians.

S2.6. Figure 4 displays four frequency histograms, each constructed from a different data set of 200 observations. Each has mean 8.42. Which histogram, that graphing Y1 (upper left), Y2 (lower left), Y3 (upper right) or Y4 (lower right), has quartiles $Q_1 = 8.09$, $Q_2 = 9.00$, $Q_3 = 9.44$? Justify your answer.

S2.7. The scores on a recent test in an upper level math course were: 90 80 95 85 40 100 39 95 45 100 30.

a. Obtain the five number summary of these data and construct a box and whiskers plot.

b. Do you feel that the five number summary and the box and whiskers plot adequately summarize these data? If so, tell why. If not, tell why not, and give a better summary.

S2.8. Look at Figure 5, and suppose the data were produced by a stationary process.

a. What explanation can you give for the pattern you see?

b. Summarize the pattern of variation with appropriate measures.

S2.9 The Gross Domestic Product (GDP) of the United States for the years 1986-1999 is given in Table 1 (units are billions of 1996 dollars).

Year	GDP	Year	GDP
1986	5912.4	1993	7062.6
1987	6113.3	1994	7347.7
1988	6368.4	1995	7543.8
1989	6591.8	1996	7813.2
1990	6707.9	1997	8144.8
1991	6676.4	1998	8495.7
1992	6880.0	1999	8848.2

Table 1: *Gross Domestic Product (GDP) of the United States for the years 1986-1999.*

a. Draw an appropriate graphical summary of these data. Explain why your summary is appropriate. What conclusion do you draw from your summary?

b. Ichiro claims that the data are well summarized by their mean $\overline{y} = 7179.0$ and standard deviation $s = 891.7$. Do you agree? Give your reasons.

S2.10. Figure 6 shows a frequency histogram, box plot and summary measures of monthly sales of tobacco and tobacco products from January, 1971, through December, 1991.

a. Before trying to use the graphs and measures in Figure 6 to summarize the data, what other plot would you make? Why?

b. Assuming the graphs and measures in Figure 6 are appropriate to summarizing these data, give a verbal and numerical summary.

c. Calculate the mode in the manner described in the book.[1]

d. If you had to predict the value of a new monthly sales figure based on the histogram, what value would you choose? Why?

S2.11. In a gage R&R study, the repeated measurements produced by one operator using a single gage, are in order taken:

2.5, 2.4, 2.7, 2.8, 2.8, 2.9, 3.1, 3.0, 3.3, 3.4, 3.4, 3.5

Is a box and whiskers plot an appropriate summary of these data? Why or why not? If a box and whiskers plot is appropriate, construct and interpret the box and whiskers plot for these data. If a box and whiskers plot is not appropriate, construct and interpret an appropriate plot.

S2.12. The questions below refer to the frequency histogram of a set of 24 data values shown in Figure 7.

(a) From the histogram, obtain quartiles Q_1, Q_2 and Q_3 of the data. Tell how you got your results. (Hint: Use the book's definitions of quartiles. Do not try to use the computational formulas.)

(b) Will the mean be greater than, equal to, or less than the median? Why?

(c) Calculate the mode, in the manner described in the text.

(d) Name one measure of location and one measure of spread that appropriately summarize the pattern of variation in these data. Why did you choose these measures?

S2.13. Figure 8 displays a box and whiskers plot for a set of data.

(a) Which of the following data sets could NOT have produced the plot in Figure 1 (For each one you conclude could not have produced the plot, give a reason):

i. 1 2 4 5 8 9 11 29 33 37 39 40 41 44 47

ii. 1 2 4 5 8 9 11 29 33 37 39 40 41 44 45

[1] If you have done the calculation correctly, your value for the mode should differ from the value, 1009, given in the SAS output. The value 1009 is the SAS mode since it occurs three times among the 252 data values. No other single value occurs more than twice. this should suggest the meaninglessness of the usual definition of the mode.

iii. 1 2 4 5 8 9 11 29 33 37 39 40 41 43 44 45

iv. 1 2 4 6 8 9 11 29 33 37 39 40 41 43 45

v. 1 3 4 5 8 9 11 29 33 37 39 40 41 43 45

(b) Is the box and whiskers plot in Figure 8 a good graphical summary for those data sets from part (a) that could have produced this plot? Justify your answer.

S2.14. Figure 9 shows a box and whiskers plot for a set of 496 salaries.

(a) Is this box and whiskers plot an appropriate plot to summarize the pattern of variation of the data? Choose one: □YES □NO □CAN'T TELL. Explain your choice.

(b) Assuming the box and whiskers plot is an appropriate summary of the distribution, obtain a corresponding measure of location and spread. Tell what each measure means.

S2.15. The ten natural lakes with largest surface areas are listed below.

Lake	Area	Lake	Area
Caspian Sea	143244	Aral Sea	13000
Lake Superior	31700	Lake Tanganyika	12700
Lake Victoria	26828	Lake Baykal	12162
Lake Huron	23000	Great Bear Lake	12096
Lake Michigan	22300	Lake Nyasa	11150

(a) Calculate the five number summary of the surface areas.

(b) Construct a box and whisker plot of the data. Are any values displayed as outliers?

(c) Do the data give you any reason to suspect that the box and whiskers plot is not a good summary? Why?

S2.16. The second column of the table below shows primary energy consumption of the largest consuming countries of primary energy (petroleum, natural gas, coal, net hydroelectric, nuclear, geothermal, solar, wind, and wood and waste), 1999, in quadrillion BTU (source: The World Almanac and Book of Facts, 2002).

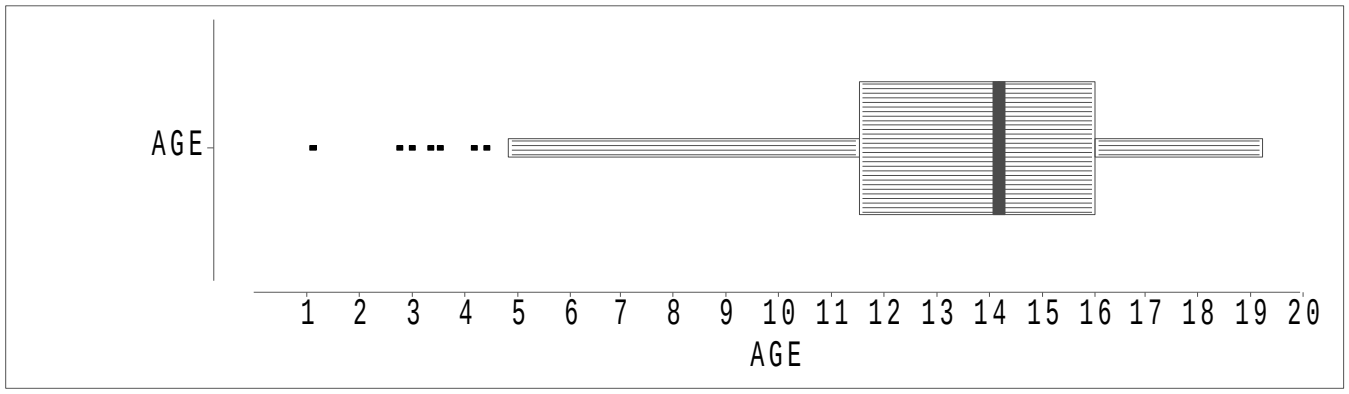
Country	Consumption	ln(Consumption)
US	96.87	4.57
China	31.88	3.46
Russia	26.01	3.26
Japan	21.71	3.08
Germany	13.98	2.64
Canada	12.52	2.53
India	12.18	2.50
France	10.26	2.33
UK	9.92	2.29
Brazil	8.51	2.14

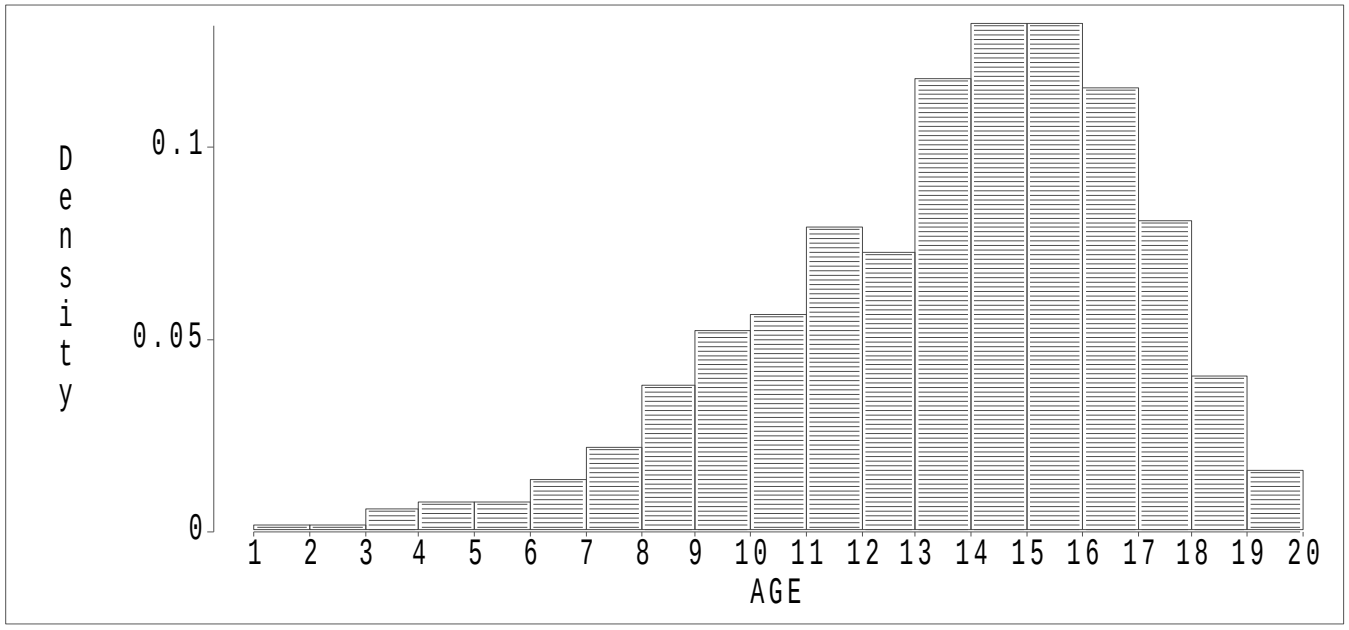
(a) Draw a box-and-whiskers plot for these data. Does the plot identify any outliers?

(b) The third column of the table contains the natural log of the energy consumption values. Draw a box-and-whiskers plot for these data. Does the plot identify any outliers?

(c) Is there a paradox here? Comment.

AGE





Moments			
N	492.0000	Sum Wgts	492.0000
Mean	13.5702	Sum	6676.5587
Std Dev	3.3328	Variance	11.1077
Skewness	-0.7746	Kurtosis	0.3785
USS	96056.4165	CSS	5453.9031
CV	24.5598	Std Mean	0.1503

Quantiles			
100% Max	19.2290	99.0%	19.1396
75% Q3	16.0234	97.5%	18.4611
50% Med	14.1796	95.0%	18.1236
25% Q1	11.5154	90.0%	17.3965
0% Min	1.1335	10.0%	8.8893
Range	18.0955	5.0%	7.5627
Q3-Q1	4.5080	2.5%	5.8676
Mode	1.1335	1.0%	3.5704

Figure 1: *Output from SAS/INSIGHT distribution analysis of ages of dogs at death.*

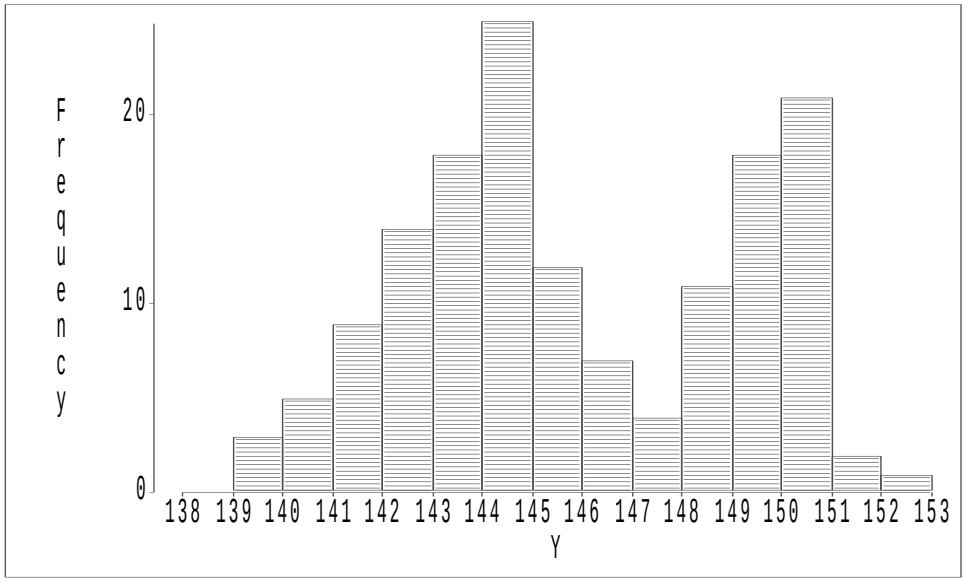


Figure 2: *Histogram for exercise S2.4.*

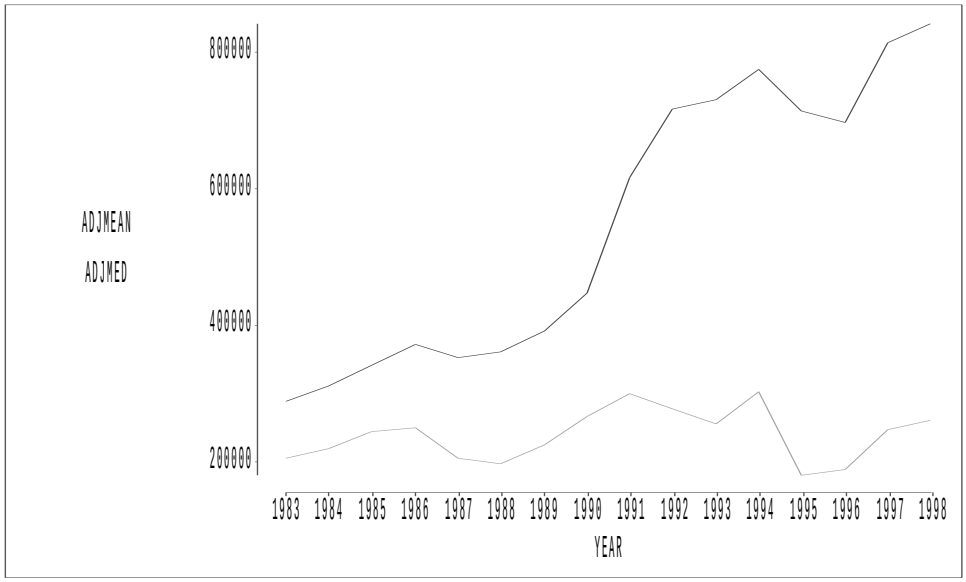


Figure 3: *Line plot of the median (lower curve) and mean salaries (in 1983 dollars) of major league baseball players for the 1983 through 1998 seasons.*

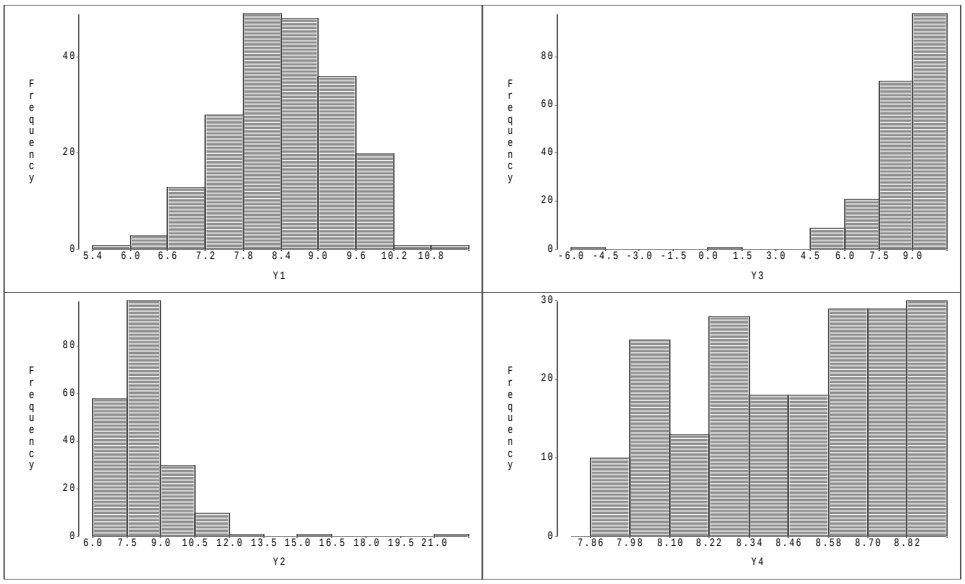


Figure 4: *Four frequency histograms for exercise S2.6.*

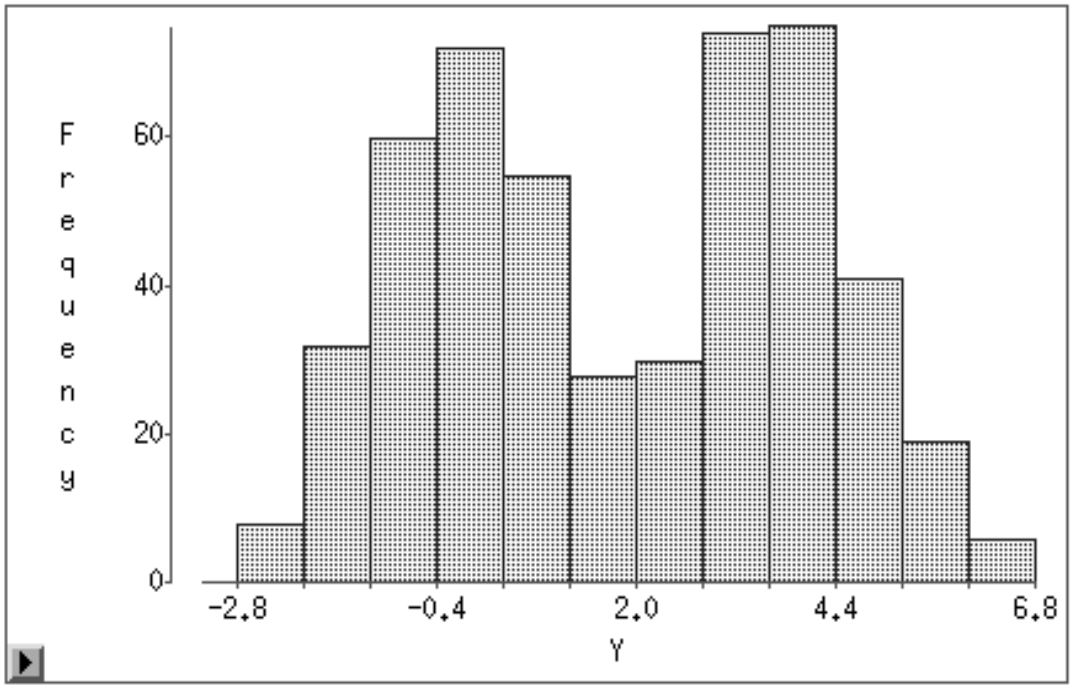
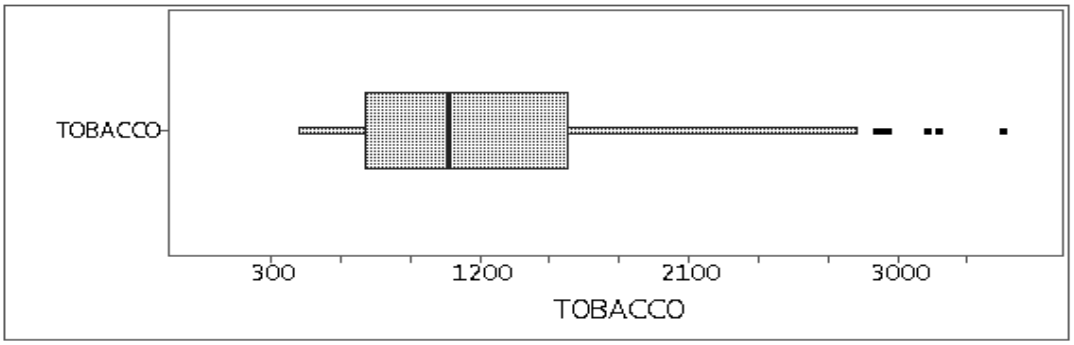
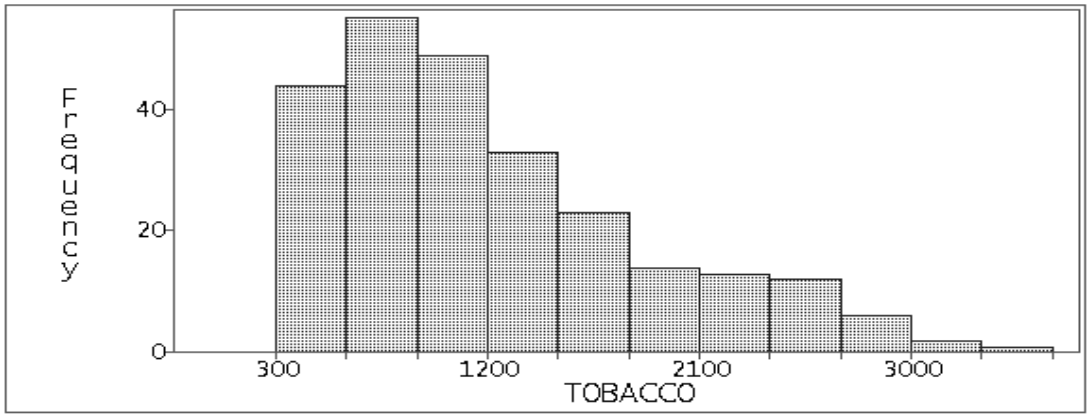


Figure 5:





Moments			
N	252.0000	Sum Wgts	252.0000
Mean	1236.1587	Sum	311512.000
Std Dev	673.4054	Variance	453474.859
Skewness	1.0006	Kurtosis	0.3147
USS	498900468	CSS	113822190
CV	54.4756	Std Mean	42.4206

Quantiles			
100%	Max	3454.0000	99.0%	3130.0000
75%	Q3	1581.5000	97.5%	2827.0000
50%	Med	1065.0000	95.0%	2581.0000
25%	Q1	709.0000	90.0%	2346.0000
0%	Min	425.0000	10.0%	509.0000
	Range	3029.0000	5.0%	474.0000
	Q3–Q1	872.5000	2.5%	459.0000
	Mode	1009.0000	1.0%	440.0000

Figure 6: *SAS distribution analysis output for monthly sales of tobacco and tobacco products from January, 1971, through December, 1991.*

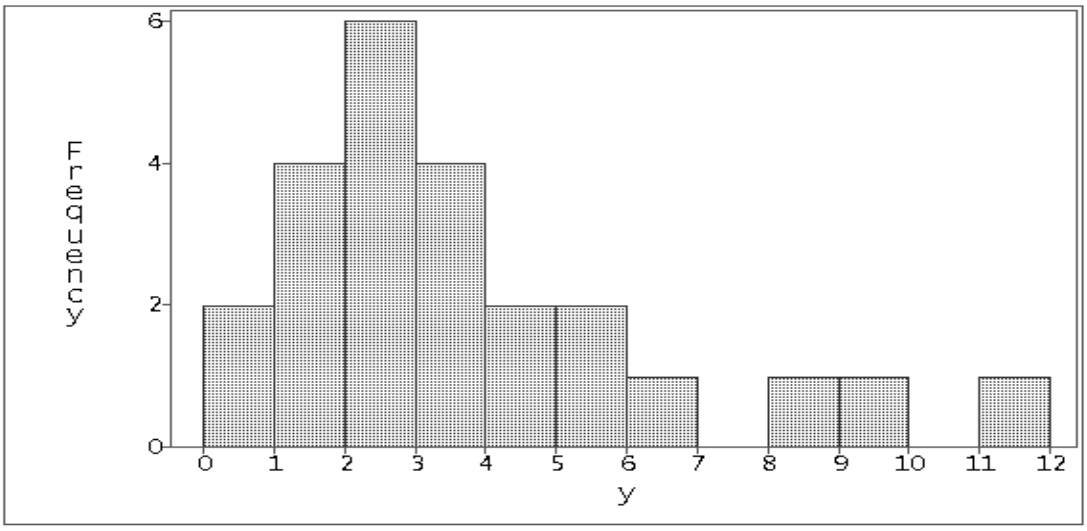


Figure 7: *Frequency histogram of 24 data values.*

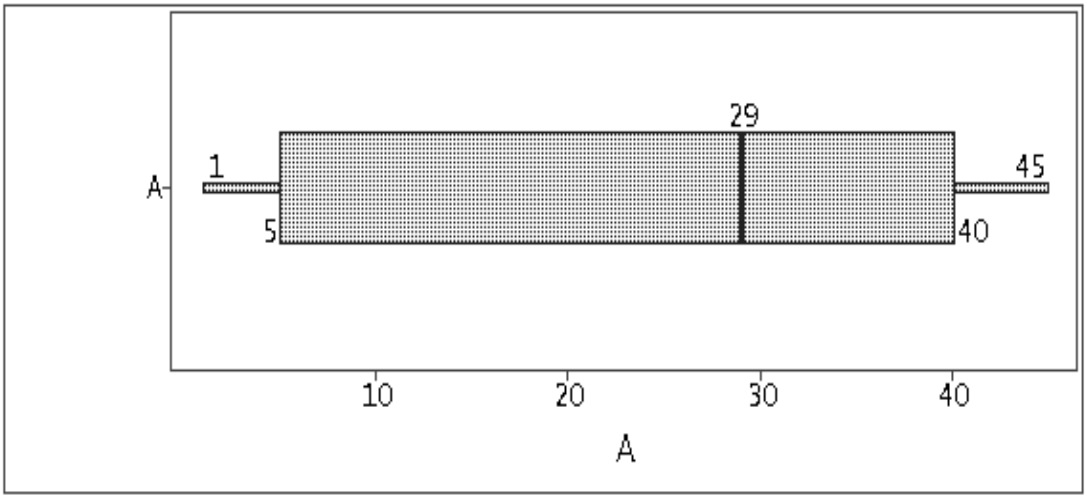


Figure 8: *Box and whiskers plot for problem S2.13.*

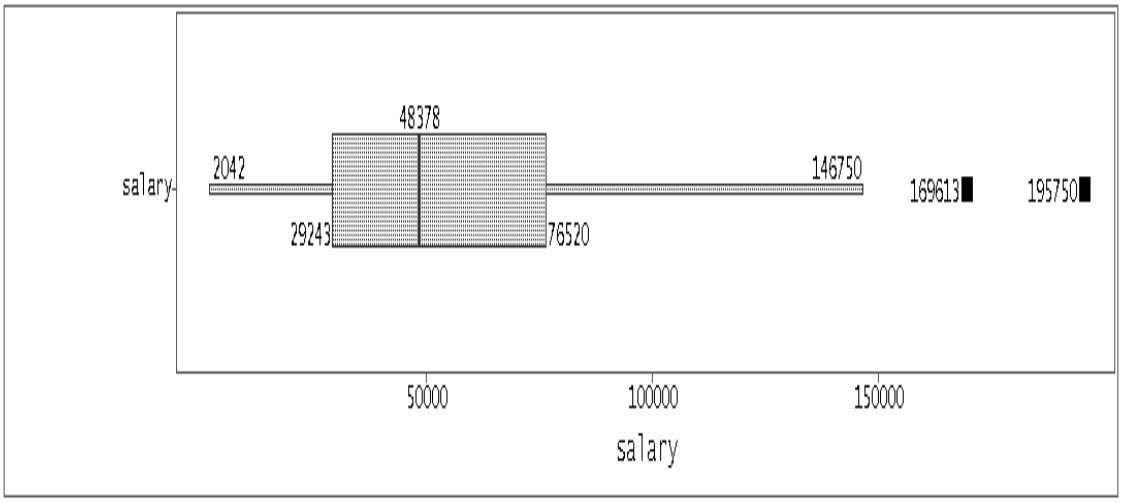


Figure 9: *Box and whiskers plot for a set of 496 salaries.*