SUPPLEMTARY EXERCISE SOLUTIONS, 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.

ANS: *The distribution is basically unimodal* **(3 points)** *and skewed left* **(3 points)***. The main mode is located at approximately 15 years.* **(3 points)** *The data ranges from 1 to 20 years.* **(3 points)** *There is a smaller modal bar over the interval 11-12 years, which may or may not represent something meaningful for dogs of that age.* **(3 points)**

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

ANS: *Either of the following two measures is appropriate:*

- *The mode (=15 years)* **(5 points)** *indicates the age at or near which a dog in the data set is most likely to have died* **(5 points)***.*
- *The median (=14.1796 years)* **(5 points)** *is the age that divided the half of the dogs that died youngest from the half that lived longest* **(5 points)***.*

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

ANS: *The mean is inappropriate, since it is pulled to the left by the skewness of the distribution.* **(5 points)**

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

ANS: *The only appropriate summary measure we have studied is the IQR* **(5 points)***, which is the range of the central 50% of the ages.* **(5 points)**

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

ANS: *The standard deviation is inappropriate since it implicitly assumes the distribution is symmetric.* **(5 points)**

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?

ANS: *The mode, since based on these data, this is where the age is most likely to be.* **(5 points)**

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.

ANS: *Not true. E.g.: 6 data values, -4, -2, -1, 1, 3, 3. Median=mean=0.* **(10 points)**

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.

ANS: *The median, because the half-life is the time until half the uranium decays. In the data, it is the time value which half the lifetimes are greater than and half are smaller than: namely, the median of the lifetimes.* **(15 points)**

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.

ANS: *Bad idea, Bill. The histogram shows the main feature of this data distribution to be bimodality. A box and whisker plot will conceal this feature.* **(10 points)**

b. Summarize these data using appropriate measures.

ANS: *The data are bimodal with modes at 144.5 and 150.5.* **(5 points)** *The first, and larger, modal region spreads from 139-147.5, while the second modal region spreads from 147.5-153.* **(5 points)**

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.

ANS: *Looking at the median, there is no reason to suspect nonstationarity.* **(10 points)** *However, the mean has a clear upward trend, which is an indication that the process is nonstationary.* **(10 points)**

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.

ANS: *In real dollars, the majority of salaries are not increasing, but the salaries for the stars are increasing greatly.* **(10 points)**

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.

ANS: *Y3, because of the left skewness.* **(10 points)**

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.
ANS: $Q_1 = 40$, $Q_2 = 85$, $Q_3 = 95$. $Q_1 - 1.5\times IQR = 40 - 1.5 \times 55 = -42.5$, *so* $A_- = 30$. $Q_3 + 1.5\times IQR = 95 + 1.5 \times 55 = 177.5$, *so* $A_+ = 100$. *The box and whiskers plot is shown in Figure 7.* **(20 points)**

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.

ANS: *No, the box and whiskers plot and the five number summary mask the most imortant feature of the data: that it is bimodal. A resonable summary would indicate that there are two distinct groups of scores: one centered near 92 and ranging from 80-100, and the second centered near 40 and ranging from 30 to 45.* **(10 points)**

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?

ANS: *There are two subpopulations being measured.* **(10 points)**

b. Summarize the pattern of variation with appropriate measures.

ANS: *The data are bimodal with one mode at 0, and the second at 4. The first peak ranges from -2.8 to 2 and the second from 2 to 6.8* **(10 points)**

S2.9. The Gross Domestic Product (GDP) of the United States for the years 1986-1999 is given in the 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?

ANS: *Figure 8 is a line plot of GDP versus year. Data taken over time should always be plotted versus time in order to assess stationarity. This plot shows an increasing trend, indicating the the process is not stationary. A summary such as a histogram or boxplot would not be appropriate because they assume stationarity.* **(10 points)**

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.

ANS: *These summaries assume stationarity and are therefore not appropriate because of the nonstationarity of the data.* **(10 points)**

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?

ANS: *A time series or line plot. Since the data were taken over time, we must check stationarity.* **(10 points)**

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

ANS: *The distribution is unimodal and skewed right. The median is 1065 and the IQR is 872.5 ($Q_1 = 709$, $Q_3 = 1581.5$). Instead of the median, the mode, 750, could be used.* **(15 points)**

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

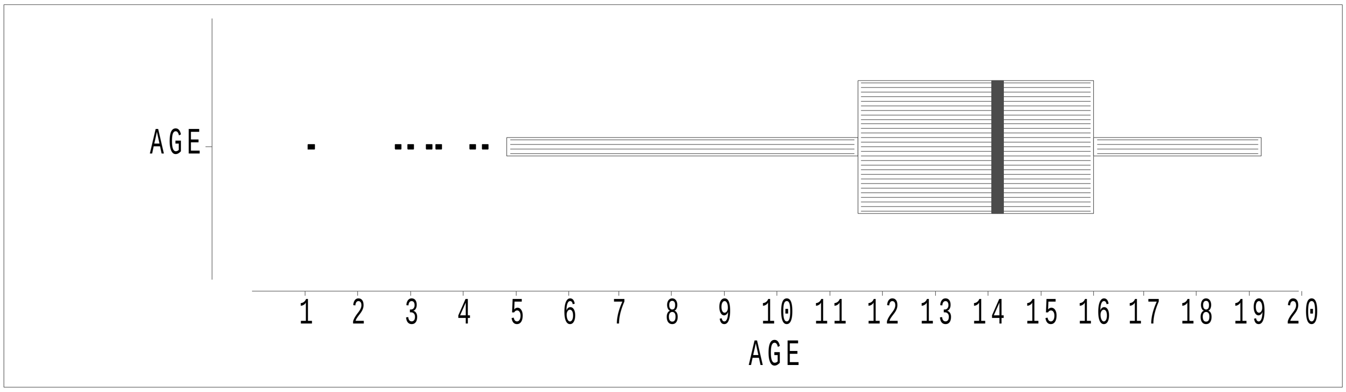
ANS: *750* **(5 points)**

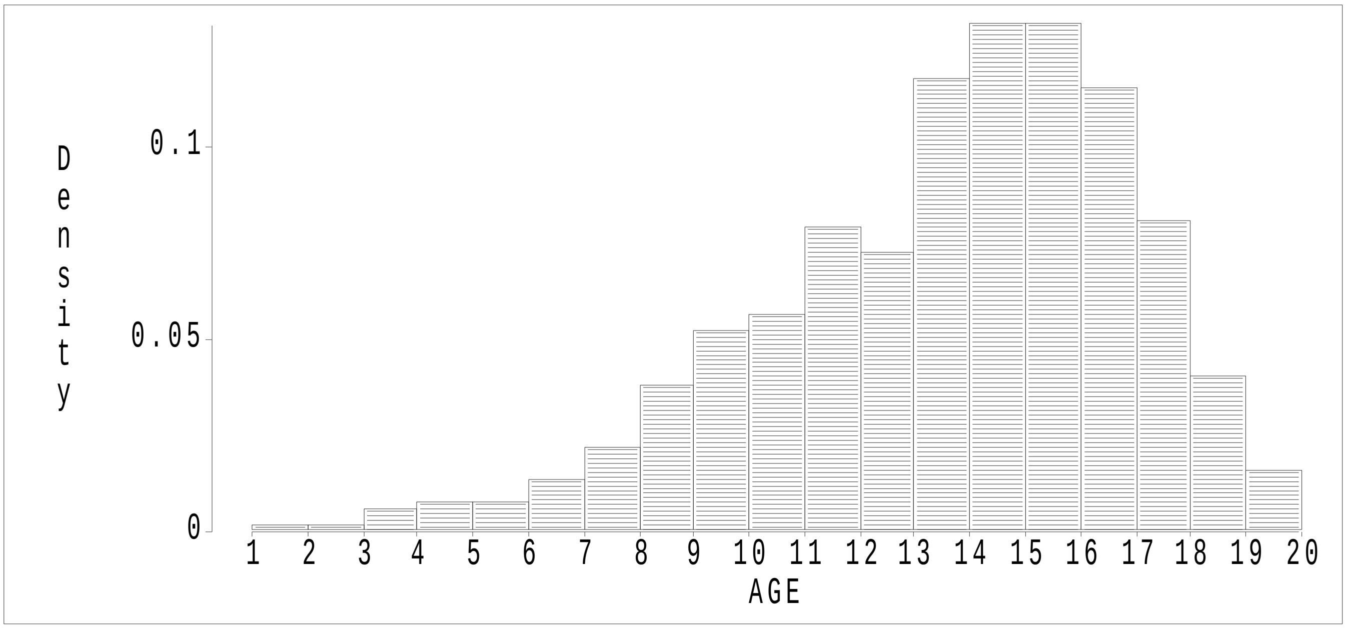
[1] If you have done the calculation correctly, your value for the mode should differ from the value, 1009, given in the SAS output. The value 1009 is the SAS mode since it occurs three times among the 252 data values. No other single value occurs more than twice. this should suggest the meaninglessness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

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

ANS: *The mode, 750, because it is located in the bar with the largest frequency.* **(5 points)**

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.*

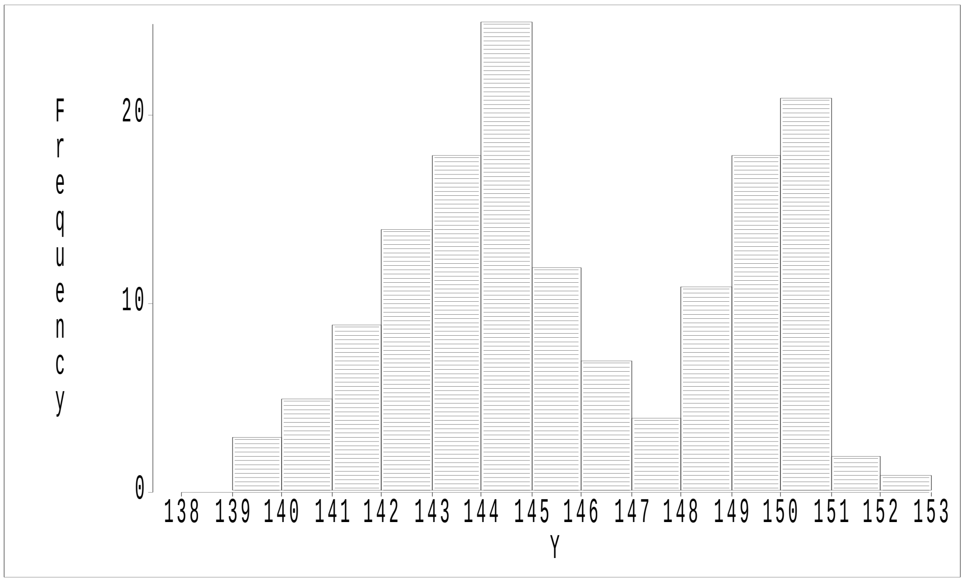


Figure 2: *Histogram for exercise S2.4.*

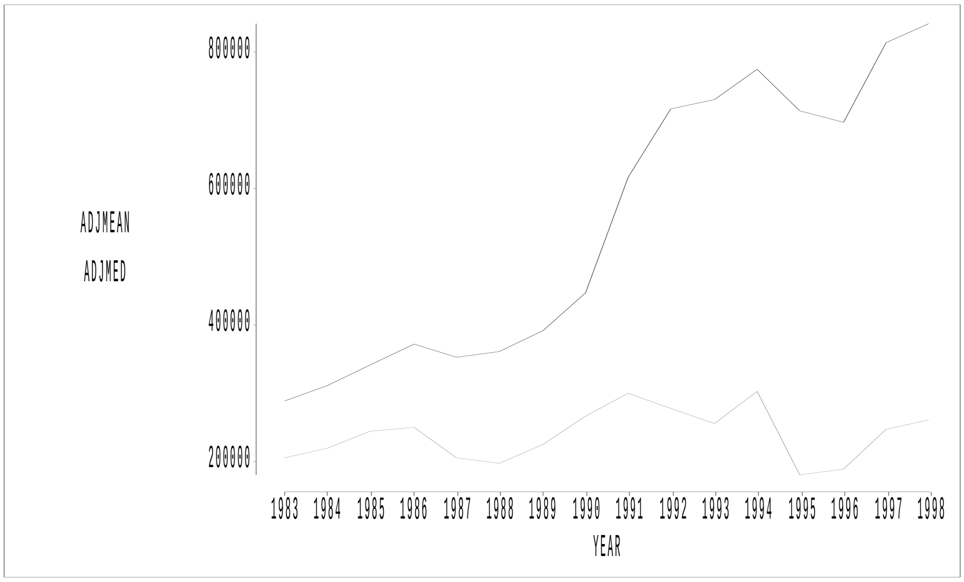


Figure 3: *Line plot of the median (lower curve) and mean salaries (in 1983 dollars) of major league baseball players for the 1983 through 1998 seasons.*

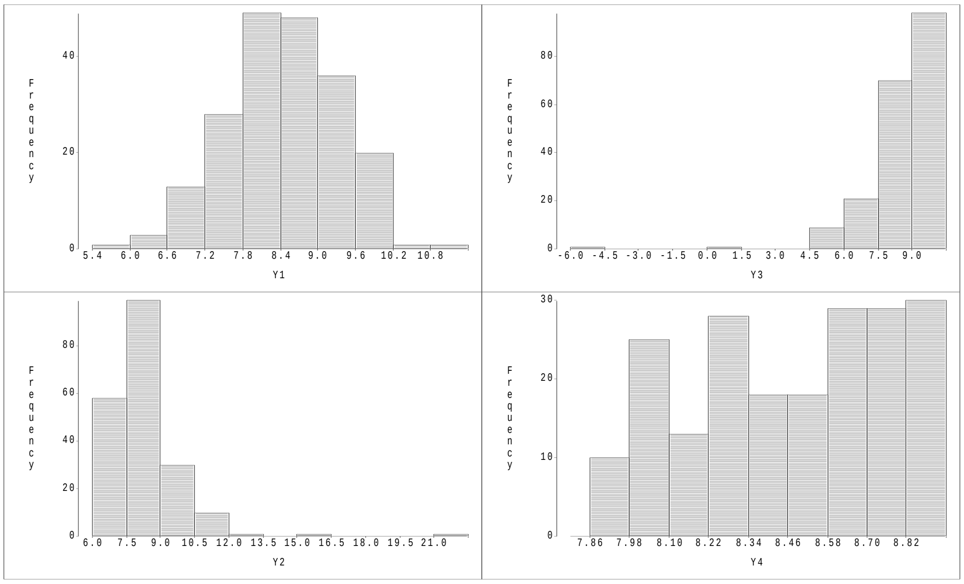


Figure 4: *Four frequency histograms for exercise S2.6.*

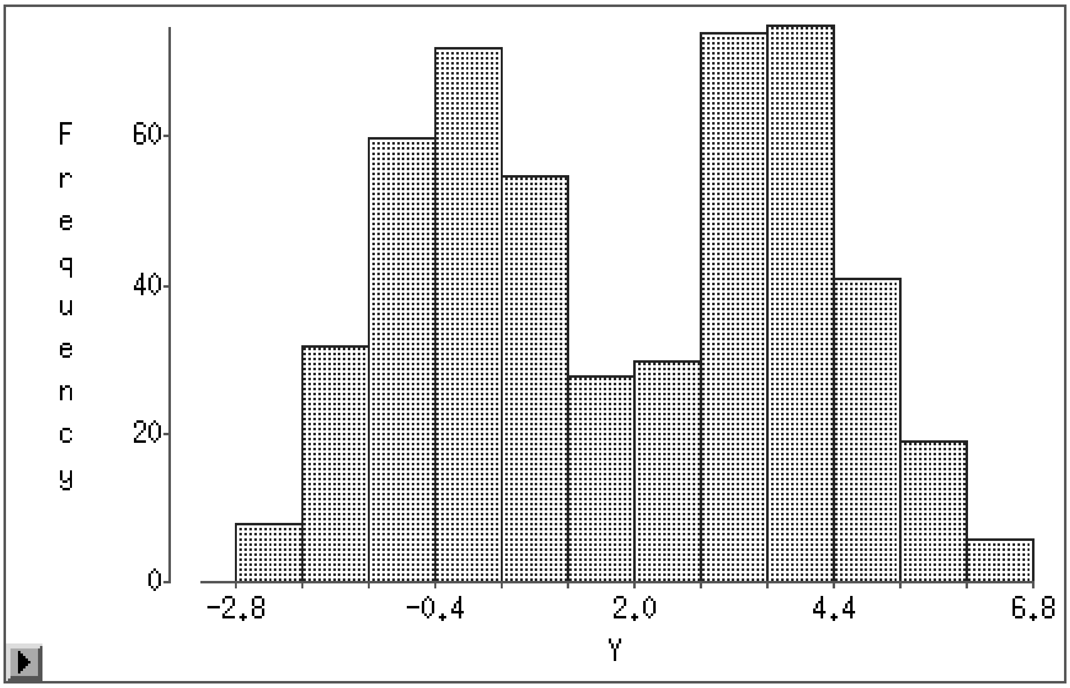
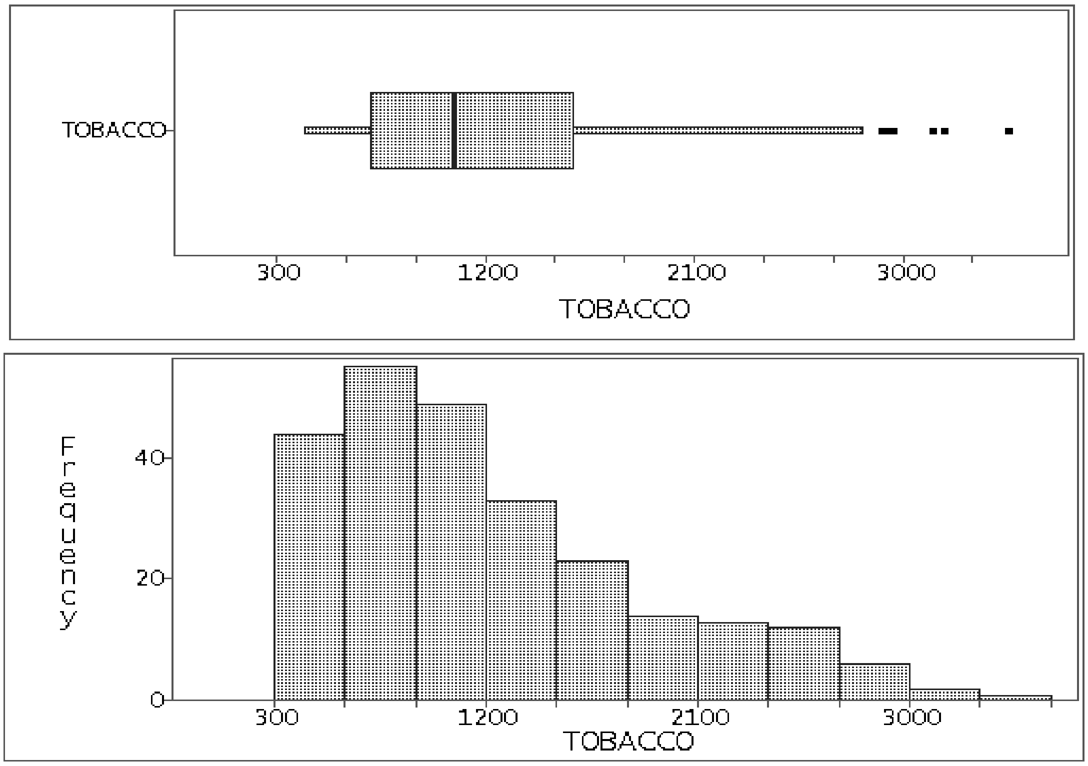


Figure 5:



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: *Frequency histogram of monthly sales of tobacco and tobacco products from January, 1971, through December, 1991.*

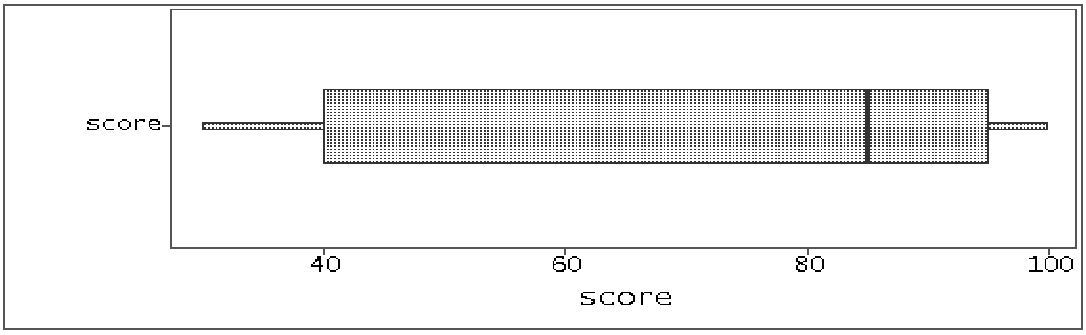


Figure 7: *Boxplot for exercise S2.7.*

Figure 8: *Line plot of GDP, 1986-1999.*