SUPPLEMENTARY 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)* **(3 points)** *indicates the age at or near which a dog in the data set is most likely to have died* **(3 points)**.
- *The median (=14.1796 years)* **(3 points)** *is the age that divided the half of the dogs that died youngest from the half that lived longest* **(3 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.* **(3 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* **(3 points)**, *which is the range of the central 50% of the ages.* **(3 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.* **(3 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.* **(3 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)**

S2.4. 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)**

S2.9. 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.*

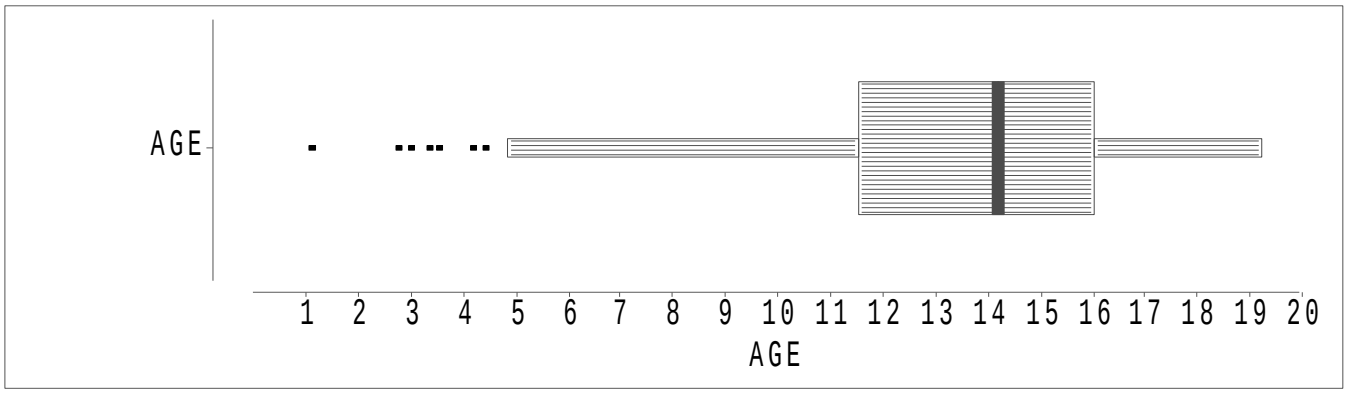
a. Draw an appropriate graphical summary of these data. Explain why your summary is appropriate. What conclusion do you draw from your summary?

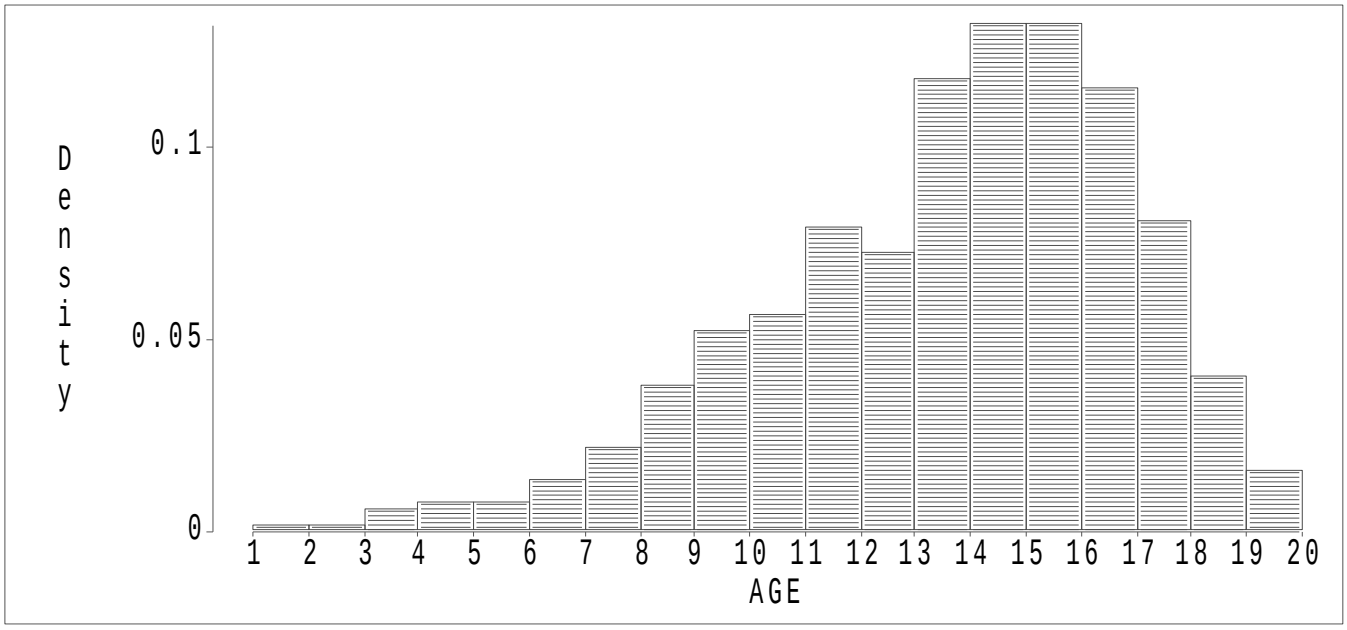
ANS: *Figure 4 is a line plot of GDP versus year* **(4 points)**. *Data taken over time should always be plotted versus time in order to assess stationarity* **(3 points)**. *This plot shows an increasing trend, indicating the the process is not stationary* **(3 points)**. *A summary such as a histogram or boxplot would not be appropriate because they assume stationarity.*

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)**

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.*

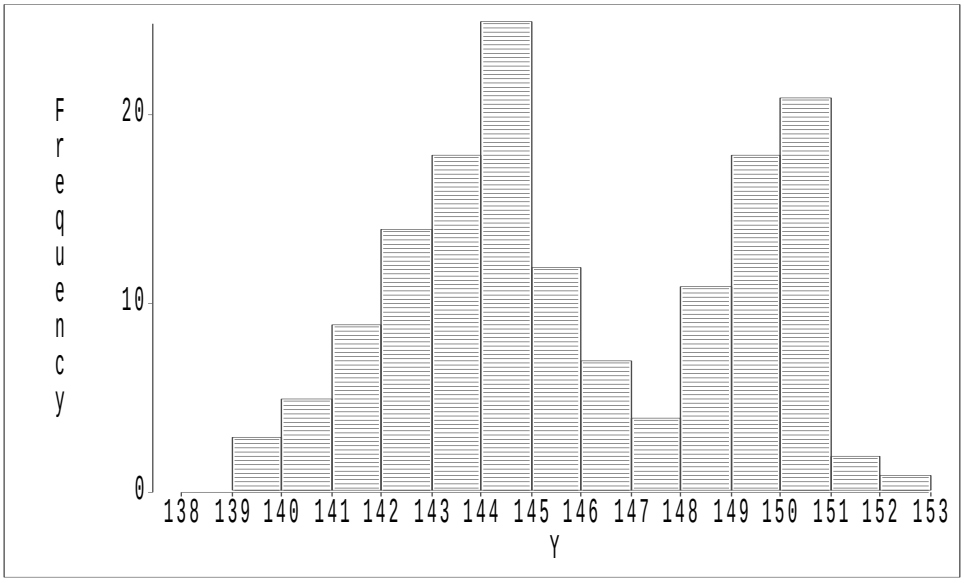


Figure 2: *Histogram for exercise S2.4.*

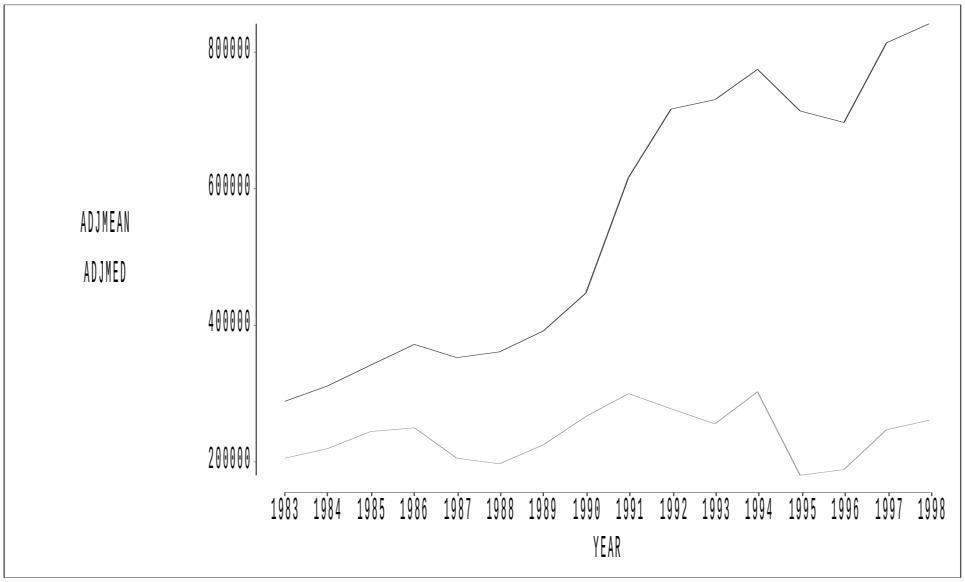


Figure 3: *Line plot of the median (lower curve) and mean salaries (in 1983 dollars) of major league baseball players for the 1983 through 1998 seasons.*

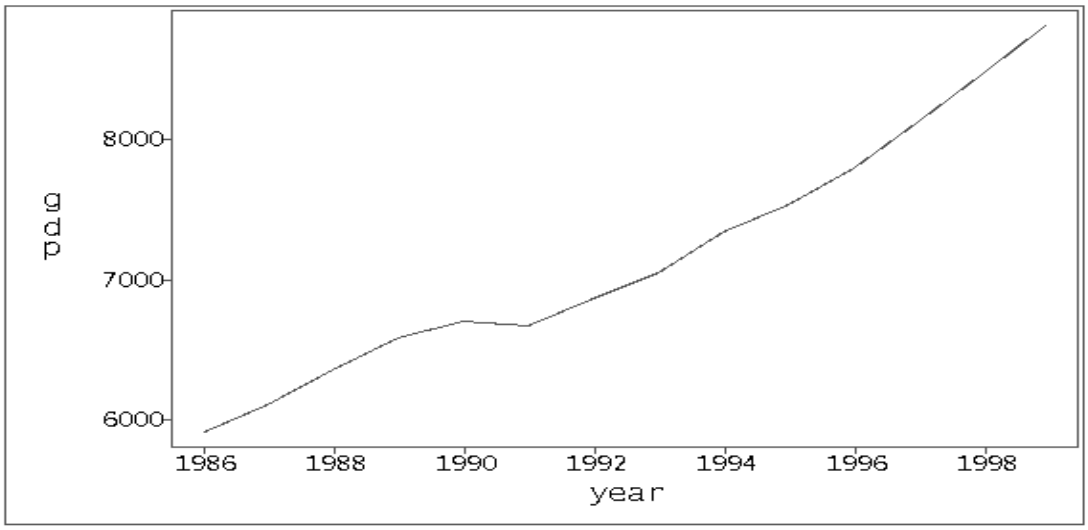


Figure 4: *Line plot of GDP, 1986-1999.*