1. Students who have finished medical school are assigned to residencies in hospitals to receive further training in a medical specialty. Here is part of a hypothetical data base of students seeking residency positions. USMLE is the students score on Step 1 of the national medical licensing examination.

Name	Medical School	Sex	Age	USMLE	Specialty sought
Abrams, Laurie	Florida	F	28	238	Family medicine
Brown, Gordon	Meharry	M	25	205	Radiology
Cabrera, Maria	Tufts	F	26	191	Pediatrics
Ismael, Miranda	Indiana	F	32	245	Internal medicine

(a) What individuals does this data set describe?

(b) In addition to the students name, how many variables does the data set contain? Which of these variables are categorical and which are quantitative?

2. The most popular colors for cars and light trucks vary with region and over time. In North America white remains the top color choice, with black the top choice in Europe and silver the top choice in South America. Here is the distribution of the top colors for vehicles sold globally in 2010:

Color	Popularity
Silver	26%
Black	24%
White	16%
Gray	16%
Red	6%
Blue	5%
Beige, brown	3%

Fill in the percent of vehicles that are in other colors. Make a graph to display the distribution of color popularity.

3. Although most social-networking Web sites in the United States have fairly short histories, the growth of these sites has been exponential. By far, the two most visited social-networking sites are Facebook.com and MySpace.com. Here is the age distribution of the audience for the two sites in December 2009.

Age group	Facebook visitors	MySpace visitors
Under 25 *years*	26.8%	44.4%
25 *to* 34 *years*	23.0%	22.7%
35 *to* 49 *years*	31.6%	23.5%
Over 49 *years*	18.7%	9.4%

(a) Draw a bar graph for the age distribution of Facebook visitors. The leftmost bar should correspond to "under 25," the next bar to "25 to 34," and so on. Do the same for MySpace, using the same scale for the percent axis.

(b) Describe the most important difference in the age distribution of the audience for Facebook and MySpace. How does this difference show up in the bar graphs? Do you think it was important to order the bars by age to make the comparison easier?

(c) Explain why it is appropriate to use a pie chart to display either of these distributions. Draw a pie chart for each distribution. Do you think it is easier to compare the two distributions with bar graphs or pie charts? Explain your reasoning.

4. Below is a stemplot of the IQ test scores of 78 seventh-grade students in a rural midwestern school.

```
 7 | 24
 7 | 79
 8 |
 8 | 69
 9 | 0133
 9 | 6778
10 | 0022333344
10 | 555666777789
11 | 0000111122223334444
11 | 55688999
12 | 003344
12 | 677888
13 | 02
13 | 6
```

(a) Four students had low scores that might be considered outliers. Ignoring these, describe the shape, center, and spread of the remainder of the distribution.

(b) We often read that IQ scores for large populations are centered at 100. What percent of these 78 students have scores above 100?

5. The table below gives the number of active nurses per 100,000 people in each state.

STATE	NURSES	STATE	NURSES	STATE	NURSES
Alabama	912	Louisiana	894	Ohio	1001
Alaska	756	Maine	1053	Oklahoma	712
Arizona	544	Maryland	869	Oregon	795
California	641	Michigan	841	Rhode Island	1007
Colorado	761	Minnesota	1017	South Carolina	795
Connecticut	994	Mississippi	868	South Dakota	1215
Delaware	977	Missouri	958	Tennessee	894
Florida	814	Montana	748	Texas	662
Georgia	653	Nebraska	1010	Utah	625
Hawaii	753	Nevada	574	Vermont	912
Idaho	642	New Hampshire	970	Virginia	750
Illinois	812	New Jersey	907	Washington	774
Indiana	864	New Mexico	580	West Virginia	938
Iowa	990	New York	859	Wisconsin	905
Kansas	867	North Carolina	886	Wyoming	812
Kentucky	923	North Dakota	1097	District of Columbia	1380

(a) Why is the number of nurses per 100,000 people a better measure of the availability of nurses than a simple count of the number of nurses in a state?

(b) Make a histogram that displays the distribution of nurses per 100,000 people.

(c) Write a brief description of the distribution. Are there any outliers? If so, can you explain them?

6. We asked the students in a large first-year college class how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 women and 30 men from the class:

		Women					Men		
270	150	180	360	180	120	120	30	45	200
120	180	120	240	170	90	90	30	120	75
150	120	180	180	150	150	90	60	240	300
200	150	180	120	240	240	60	150	60	30
120	60	120	180	180	30	230	120	95	150
90	240	180	115	120	0	200	120	120	18

(a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? What is the other common multiple found in the data? We eliminated one student who claimed to study 10,000 minutes per night. Are there any other responses you consider suspicious?

(b) Make a back-to-back stemplot to compare the two samples. That is, use one set of stems with two sets of leaves, one to the right and one to the left of the stems. (Draw a line on either side of the stems to separate stems and leaves.) Order both sets of leaves from smallest at the stem to largest away from the stem.

(c) Report the approximate midpoints of both groups. Does it appear that women study more than men (or at least claim that they do)?

7. Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents and they asked the drivers about marijuana use. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same age:

	Never	1-10 times	11-50 times	51+ times
Drivers	452	229	70	156
Accidents Caused	59	36	15	50

(a) Explain carefully why a useful graph must compare rates (accidents per driver) rather than counts of accidents in the four marijuana use classes.

(b) Compute the accident rates in the four marijuana use classes. After you have done this, make a graph that displays the accident rate for each class. What do you conclude? (You cant conclude that marijuana use causes accidents, because risk takers are more likely both to drive aggressively and to use marijuana.)

8. Some say that the earth is an interconnected, living whole. For example, it appears that El Nio, the periodic warming of the Pacific Ocean west of South America, affects the monsoon rains that are essential for agriculture in India. Here are the monsoon rains (in millimeters) for the 23 strong El Nio years between 1871 and 2004:

$$628 \quad 669 \quad 740 \quad 651 \quad 710 \quad 736 \quad 717 \quad 698$$
$$653 \quad 604 \quad 781 \quad 784 \quad 790 \quad 811 \quad 830 \quad 858$$
$$858 \quad 896 \quad 806 \quad 790 \quad 792 \quad 957 \quad 872$$

(a) To make a stemplot of these rainfall amounts, round the data to the nearest 10, so that stems are hundreds of millimeters and leaves are tens of millimeters. Make two stemplots, with and without splitting the stems. Which plot do you prefer?

(b) Describe the shape, center, and spread of the distribution. Are there any outliers?

(c) The average monsoon rainfall for all years from 1871 to 2004 is about 850 millimeters. What effect does El Nio appear to have on monsoon rains?

9. We all know that fruit is good for us. Many of us dont eat enough. The figure below is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.

Number of Subjects

Servings of fruit per day
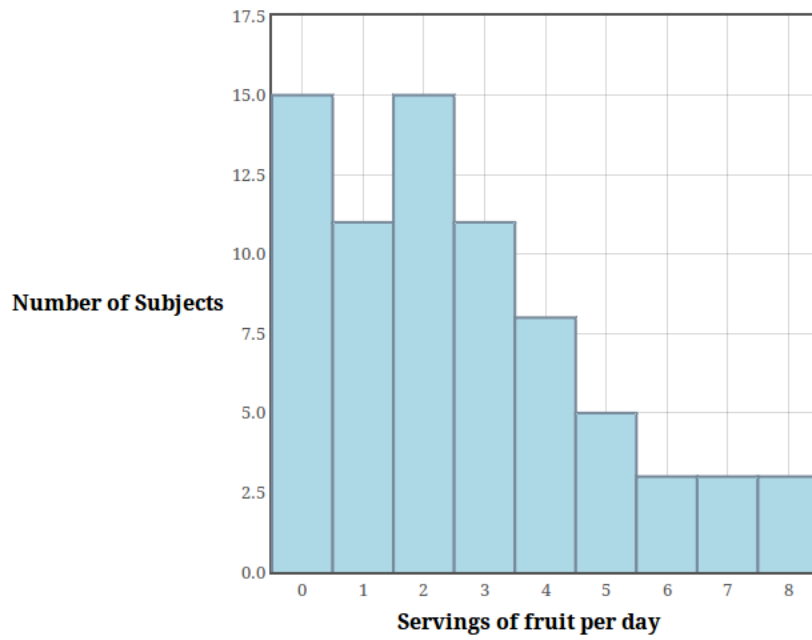
(a) Describe the shape, center, and spread of this distribution.

(b) What percent of these girls ate six or more servings per day?

(c) How many of these girls ate fewer than two servings per day?

(d) Are there any outliers?