Home Run Data for Four Famous Sluggers through 2000

Here are yearly homerun totals the four major league baseball players who have ever hit at least 60 home runs in a single season. The data is for every year they appeared in at least one major league game, including years cut short by injury, retirement, or in the case of Babe Ruth, when the player was primarily a pitcher.

Babe Ruth	0	4	3	2	11	29	54	59	35	41	46	25	47	60	54
	46	49	46	41	34	22	6								
Roger Maris	14	9	28	16	39	61	33	23	26	8	13	9	5		
Mark McGwire	3	49	32	33	39	22	42	9	9	39	52	58	70	65	32
Sammy Sosa	4	15	10	8	33	25	36	40	36	66	63	50			

Two of the players are active and also have hit many home runs this season (McGwire, 22, and Sosa, 53, through September 3, 2001), but we will only consider data for seasons prior to 2001.
Follow these steps to construct a boxplot for Babe Ruth's season home run totals.

1. Sort the data from smallest to largest. (A stem-and-leaf diagram is one useful technique for sorting.)
2. Find the median (M), first quartile (Q_1), and third quartile (Q_3) of the data.
3. Calculate the interquartile range (IQR) which is the distance between the third quartile and the first quartile ($Q_3 - Q_1$).
4. Identify any observations more than 1.5 IQRs above the third quartile or more than 1.5 IQRs below the first quartile. These observations are *potential outliers.*
5. Draw a vertical number scale that spans all of the observations.
6. Draw a box representing the middle half of the data. The top of the box is the third quartile and the bottom of the box is the first quartile. Split the box with a line at the median.
7. Draw a whisker (a vertical line) from the top of the box to the largest observation that is not a potential outlier. (If there are no large potential outliers, this will be the maximum.) Also, draw a whisker (a vertical line) from the bottom of the box to the smallest observation that is not a potential outlier. (If there are no small potential outliers, this will be the minimum.)
8. Draw a symbol at each potential outlier.

Questions:

1. Draw side-by-side boxplots of all four players using the same vertical axis.
2. Which player has the highest single-season home run total?
3. For which player was the highest single season home run total the most unusual compared to home run totals in the remaining seasons?
4. Do the boxplots exhibit skewness in any of the distributions? In which direction?
5. Is there a single player who appears to have had a higher percentage of great home run hitting years? What about the graph justifies your response?
6. In the seasons where Mark McGwire appeared in fewer than half his team's games, his home run totals were 3, 9, and 9. Find the median and quartiles of the number of home runs per season for McGwire excluding these outliers. Redraw the boxplot. How is it different? What number do you think best describes the number of home runs McGwire hits in a "typical year in his prime when he is healthy".
7. For each player, plot the number of homeruns versus season with a timeplot. Describe any patterns you see that are common among the four players.
8. Choose a distribution that is skewed to the right. Calculate the mean and median. How do the mean and median compare? Which is a more "typical" value?
9. For the same player you examined in the last question, compute the standard deviation. Calculate the proportion of seasons where the home run total is within one standard deviation (above or below) of the mean. Do this also for two standard deviations.