Homework Describing Data KEY

**Keep in mind that you may choose different classes for your histograms and then they will look different. That is okay. (Also, my software does percents not counts. That is okay too, the histograms still have the same shape.)

***You will notice that to save time I have often done part of the problem for you such as giving you the histogram. You don't need to redo the work the I have done. But if I ask you to create a histogram or boxplot or find the mean or median, you should do it by hand for this assignment.

1. **Book Problems**
   Do the problems from the textbook for Section 6.1: 1, 3, 5, 7, 9, 10, 11
   For each problem it describes an experiment and data set. You don't have to look at the actual data set, but you can. The zip file is on Canvas on the assignment page.
   For each problem:

   (a) Define the population that the sample is taken from. Do you think the sample is representative of the population?

   (b) Can you think of any other factors that should be taken into account? Are there any issues you see with how the data was collected?

2. **Representative Samples**
   If we want to be able to generalize our sample results to a population, we need our samples to be representative of the population. This means that the characteristics of the sample should be similar to the population. If our sampling method will systematically favor an outcome, we say we have "sampling bias". For each example below, think of a reason why the sample is not representative of the population.

   (a) We wanted to know how many people in the nation would vote for Obama in the 2012 election. We randomly sampled 200 people in Utah. (In Utah, 24.9% of people voted for Obama, but nationwide, 50.6% voted for Obama.)
   People in Utah are not representative of the United States. Traditionally, more people in Utah vote Republican than Democrat. So of course, our sample would show a much lower percentage voting for Obama than was accurate.

   (b) We want to call people to ask them who they will vote for in the 2014 election. We use the phone book to select 500 random names.
   If you use the phone book to select people, you miss everyone that doesn't have a listed phone number (people with cell phones, people without phones, people with unlisted phones.) Perhaps people without phones vote Democrat more often than Republican. This would bias your results.

   (c) We want to know how many people are on a diet in Logan, Utah. We call 340 randomly chosen people and we only interview the person who answers the phone. (Research shows that women answer the phone more often than men.)
   Since women answer the phone more often than men, more of the people you interview will be women. Since more women are on diets than men, your results will show more people on diets than is accurate.

   (d) You want to know the common causes of death of people in Utah. You watch the ABC 4 news at 6:00 pm for three months and keep track of the causes of death mentioned by the newscasters.
   The news tries to report exciting or uncommon deaths. Even though most people might die by natural causes, those aren't exciting news stories. Therefore, your results would be skewed towards deaths by murder, accidents, etc.

   (e) We want to know how citizens in North Logan feel about increasing library funding. We go to the North Logan Library for four hours and ask the people leaving how they feel about increasing library funding?
   People at the library will probably be more in favor of library funding since obviously they are utilizing the library. Therefore, your results will be biased in favor of library funding.

3. **Lurking Variables**

   Lurking variables are variables that we don't account for, but they affect the outcome of our study.

   Question: Suppose a researcher discovers that during the months with high ice cream sales, there are higher death rates from drowning. Does this mean that eating ice cream causes people to drown?

   Answer: This is a ridiculous observation to draw. A possible lurking variable is the temperature. People might be more likely to eat ice cream in the summer. Since people are also more likely to go swimming in the summer, we see higher rates of drowning during high ice cream months. **Therefore, temperature is the lurking variable.**

   For each example below, think of a possible lurking variable. There could be more than one reasonable lurking variable.

   (a) For the past 20 years, as mattress prices have gone up, so have professors salaries. Should we draw the conclusion that raising mattress prices will cause professors to get raises?
   A possible lurking variable is inflation and the economy. The overall trend in our economy is that over time people make more money, but the cost of goods also increases.

   (b) A study showed that high school students who took a foreign language class scored better on the SAT college entrance exam. Should you conclude that forcing your teenager to take German will result in a higher SAT score?

   - **First of all, I didn't say which foreign language classes were involved in the study. So you shouldn't assume that you can draw conclusions about taking a German class specifically.**
   - **A possible lurking variable is which high school students are college bound.**
     - **Students are told that if they want to go to college, they should take a foreign language class.**
     - **Students who are determined to go to college are more likely to take a foreign language class.**
     - **They are also more likely to take the SAT and to spend more time studying for the SAT.**
     - **These students are also more likely to take more advanced classes in high school which would also increase their SAT scores.**

   (c) A study was done that showed that children who slept more gained less weight. Should we automatically draw the conclusion that more sleep will prevent weight gain?

   - **Children that exercise more should sleep better. They should also gain less weight. So the amount of exercise is a lurking variable.**
   - **Many people believe that the food you eat affects your sleep. Perhaps the children who don't eat as healthily and gain weight also lose sleep because of the food. So the type of food eaten could be a lurking variable.**

4. Label each variable as categorical or numerical. If it is numerical also break it down into discrete or continuous. Remember some textbooks use the more technical vocabulary

$$\textit{categorical} \longrightarrow \textit{qualitative}$$

$$\textit{numerical} \longrightarrow \textit{quantitative}$$

   (a) hair color

   (b) type of house

   (c) size of house measured in square feet

   (d) color of paint in living room

   (e) number of rooms in house

   (f) credit card number

(g) type of stocks John owns

(h) amount of money in John's stocks

(i) brand of battery

If you forgot, discrete means it can only be specific numbers. Continuous means it can be any number in an interval.

(a) **(blue, green, brown, etc)—categorical**

(b) **(rambler, split level, four-plex, etc)—categorical**

(c) **numerical (and continuous)**

(d) **categorical**

(e) **numerical (and discrete)**

(f) **Trick! Credit card numbers are numbers but the actual number means nothing. It is just an identifier. You could as easily use shapes or letters like: ♫♋♋♊☺.◐ instead of 1232332. So it is not considered a variable. It doesn't tell us anything about the person.**

(g) **categorical**

(h) **numerical (and continuous)**

(i) **categorical**

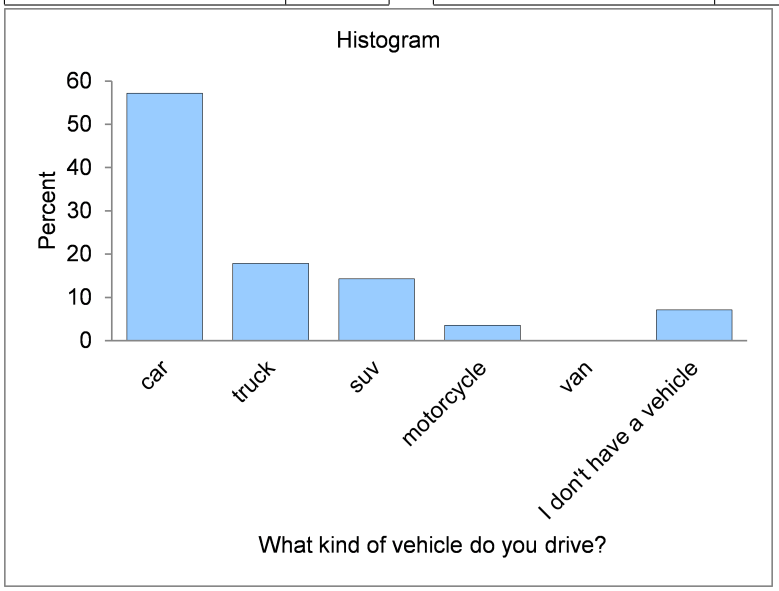
5. **Make a bar graph of the data.**
Here are the results to a survey question: What kind of vehicle do you drive?

Raw Data: car, car, car, car, car, car, car, car, car, car, car, car, car, car, car, car, I don't have a vehicle, I don't have a vehicle, motorcycle, suv, suv, suv, suv, truck, truck, truck, truck, truck

Summary Data:

Gender	Count
motorcycle	
car	
suv	
van	
truck	
I don't have a vehicle	

Vehicle	Count
car	16
truck	5
suv	4
motorcycle	1
van	0
I don't have a vehicle	2



Notice, that I put the vehicles in order from highest to lowest, except the category of "I don't have a vehicle". That is because it is a very separate category that I judged should go at the end. This is common practice if we have a category that doesn't seem to go with the others.

6. Survey Question: What is your age in years?

   Raw Data: 18 18 18 19 19 19 19 20 20 20 20 20 21 21 21 21 21 23 23 24 24 24 24 25 26 26 29 29

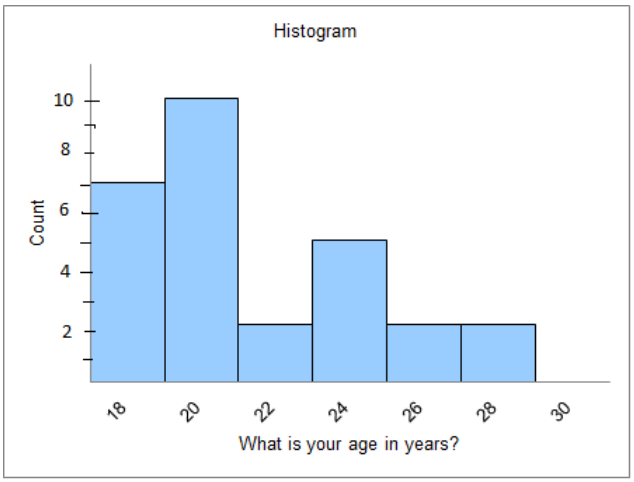
   (a) make a histogram and boxplot for the data

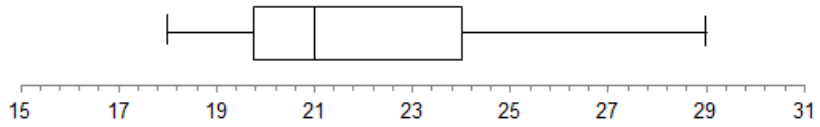
Class	Count
18-19	7
20-21	10
22-23	2
24-25	5
26-27	2
28-29	2

Minimum	18
1st Quartile	19.5
Median	21.00
3rd Quartile	24.00
Maximum	29

18 18 18 19 19 19 19 20 20 20 20 20 21 21 21 21 21 23 23 24 24 24 24 25 26 26 29 29

19.5 21 24





   (b) Enrique just registered for the class. He is 50 years old. Will adding his age to the data set cause the mean or median to increase more?
   the mean is always more susceptible to outliers, so it will increase the mean more than the median

Remark. **There are actually many different ways software estimate the quartiles. So depending on which software you use, you might get slightly different results. The important thing to remember is that the quartiles divide your data into 4 equal parts.**

Remark. If you are having trouble with quartiles check out this website: https://www.mathsisfun.com/data/quartiles.html

7. **Survey Question:** What is your height in inches? (For example, 5 foot 9 inches is 69 inches)

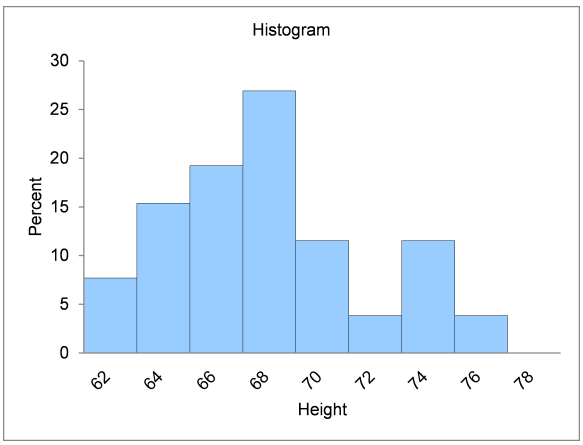
   Raw Data: 62 62 64 64 65 65 66 66 66 67 67 68 68 69 69 69 69 69 70 70 70 72 74 74 75 76

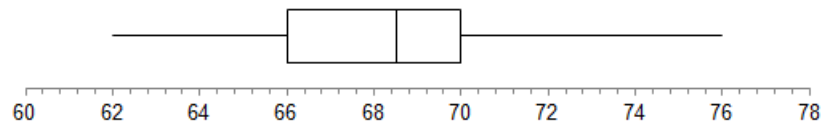
   (a) Make a histogram and boxplot for the data.

Minimum	**62**
1st Quartile	**66.00**
Median	**68.50**
3rd Quartile	**70.00**
Maximum	**76**

62 62 64 64 65 65 66 66 66 67 67 68 68 69 69 69 69 69 70 70 70 72 74 74 75 76

66 68.5 70

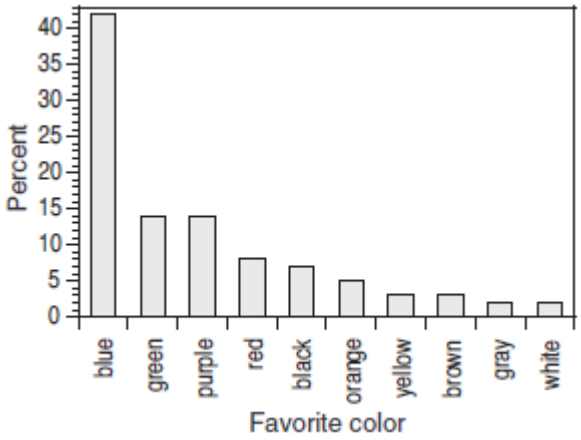




Remember that if you choose different bins, you will get a different histogram. That's okay as long as your choice is reasonable.

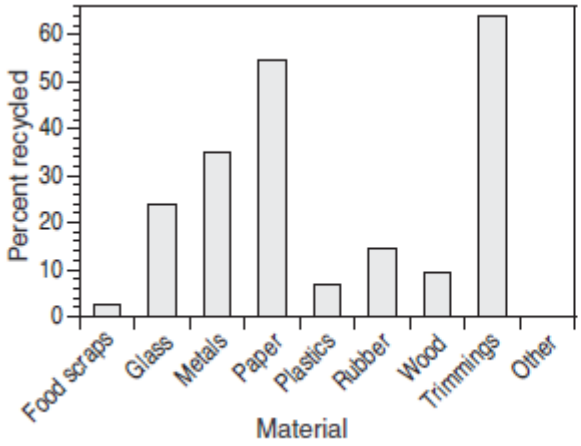
(b) A student accidentally wrote 5 (for 5 feet) instead of 60 inches. If we include him into the data set, what will happen to the mean and median? Will the mean or median change more?
Both the mean and median will decrease (because it is a low value), but the mean always changes more than the median.

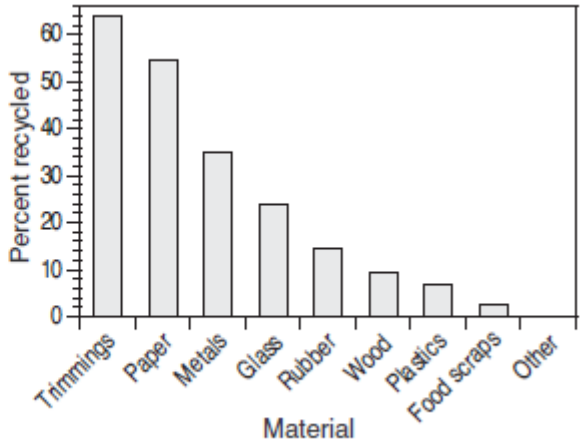
8. Here is the bar chart for survey results for "what is your favorite color. What are some things you notice about the favorite colors?



For example, blue is by far the most popular choice; 70% of respondents chose 3 of the 10 options (blue, green, and purple). Almost no one chooses gray or white.

9. Here is a bar chart for how much of each type of material is recycled. The first chart uses the order the materials were listed in. The second chart actually orders the materials from largest percent to smallest percent.

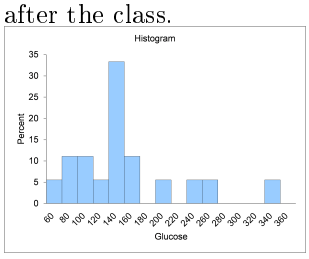


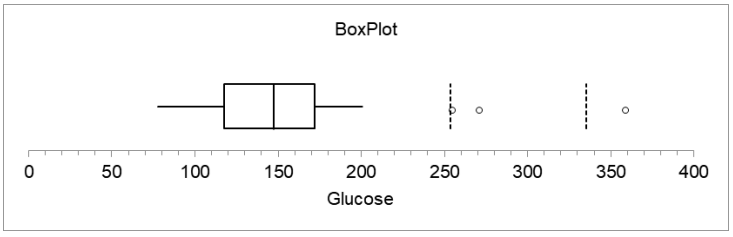


(a) What charts seems better to you and why?
The ordered bars in the second graph make it easier to identify those materials that are frequently recycled and those that are not.

(b) What is the most frequently recycled? The least?
Yard trimmings make up the most at 64.1%. The least is food scraps 2.6%.

10. Here is a histogram and boxplot for the blood glucose levels of people in a diabetes control class five months after the class.





(a) Is the distribution left skew, roughly symmetric, or right skew.

(b) Is the distribution unimodal or bimodal.

(c) Do you see any possible outliers?
It is harder to tell for sure, but this still seems slightly right skew.

I would say it is unimodal
I think that the glucose level of 359 is a possible outlier. Perhaps the levels of 255 and 271 are also outliers, but it is harder to tell from the histogram. But the boxplot makes it clear that 255, 271, and 359 are outliers.

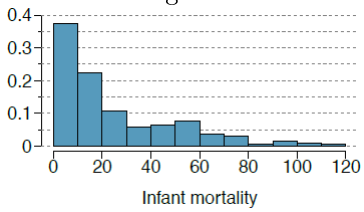
****
Unimodal means 1 mode. So you should see 1 major peak.

Bimodal means 2 modes. So you should see 2 major peaks.

(d) The goal is to have blood glucose levels between 90 and 130 mg/dl. How do you think people in this group are doing?
It looks like lots of people (really more than 50% since the median seems to be about 148 or so) are too high. And a few are too low.

11. **Remember**: Left skew has a long left tail (and the big hump on the right side). So remember the name is based on which side the long tail is on.

12. This is a histogram for the distribution of estimated infant death rates for 24 countries in 2001-20014.



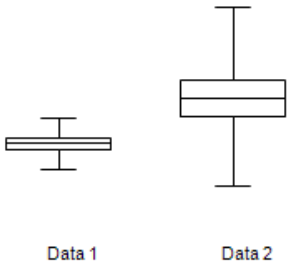
(a) Is the distribution left skew, roughly symmetric, or right skew?

(b) Is the distribution unimodal or bimodal?

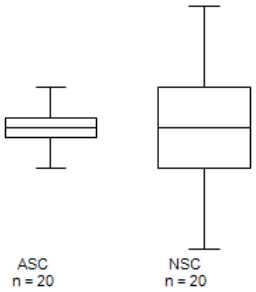
(c) Do you see any possible outliers?
This seems very right skew.
I would say it is unimodal
There don't appear to be any outliers.

(d) Would you expect the mean of the data set to be bigger or smaller than the median? Why?
If it is a right skew, the mean will get "pulled" to the right. So the mean will be larger than the median.

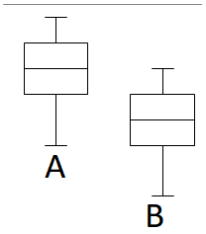
13. Compare the boxplots in terms of their center and spread.



(a)
Data 1 has a lower spread and a lower center.



(b)
They have about the same center, but ASC has a lower spread.

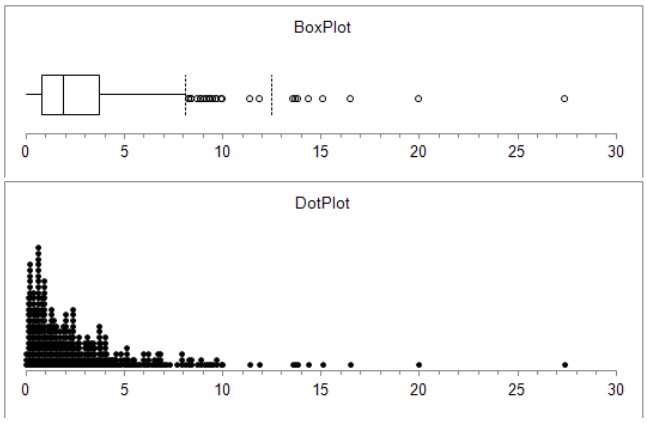


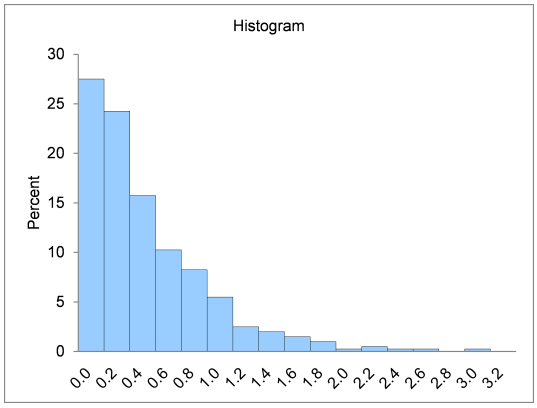
(c)
A has a higher center, but the spread seems about the same.

14. **Optional:** If you are still confused about left skew and right skew try this website:
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/skewed-distribution/

15. Let's look at some more boxplots, dotplots, and histograms. (Dot plots are histograms with dots.) Describe the distribution.

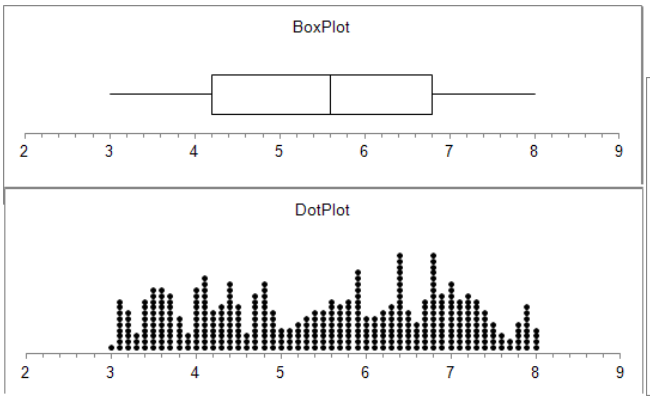
(a)

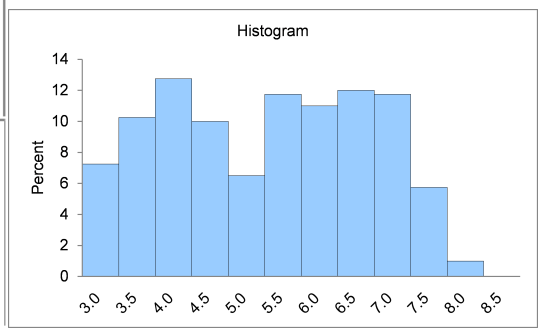




This distribution is right skew and there are lots of outliers. It is unimodal.

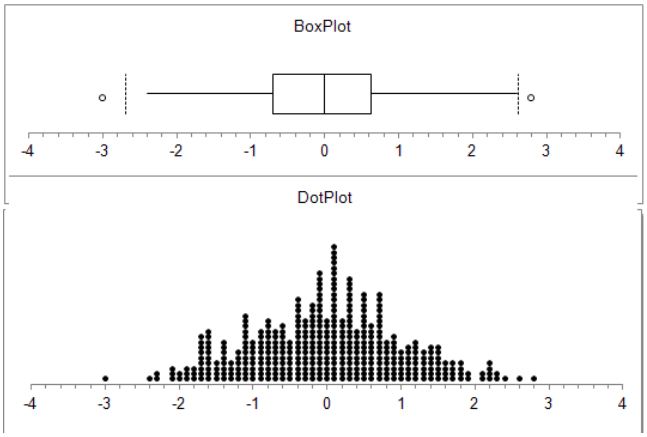
(b)

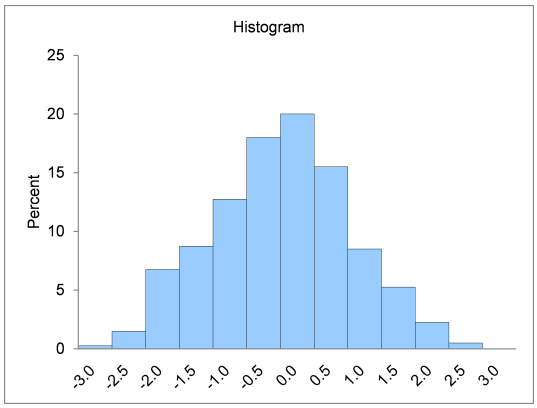




This distribution is fairly symmetric. I don't see any outliers. The histogram makes it look bimodal, but you can't see that in the dot plot. (Sometimes changing the number of our bins/classes in a histogram will change the number of major peaks (modes) that you can see.

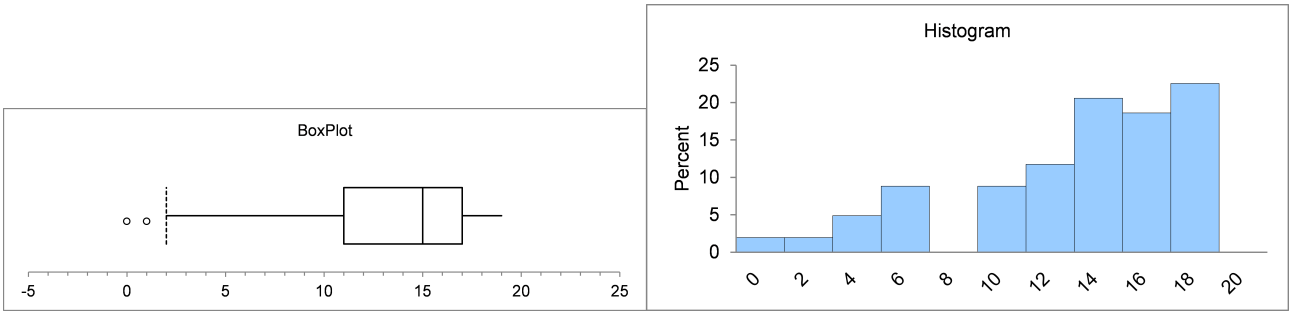
(c)





It is symmetric, with two possible outliers. It is easier to see the possible outliers on the boxplot and dotplot than on the histogram. It is unimodal.

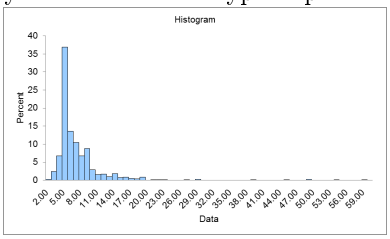
(d)



This is left skew. With two possible outliers. I would still call it unimodal.

16. A few years ago, my husband was trying to find the price of a typical plastic miniature so that he could compare the prices of plastic miniatures to paper miniatures for his blog.

    (a) Looking at the data as well as the mean of 8.13, the median of 6.49, and the histogram, what would you choose as the typical price?



Looking at the data and histogram, it appears that most of the miniatures cost somewhere between $4 and $10. By far, the most common value is $5.99.

I would accept the answer of the median at 6.49 because it seems representative of the data as well.

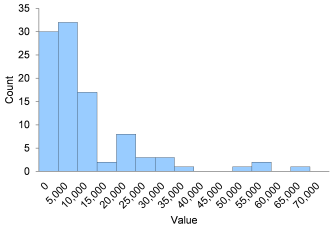
I would not accept the answer of the mean because it is inflated. There are many more figure less than $8.13 than above $8.13. When we look at the data, we see that the extremely high figures of $40-$60 will inflate the mean. So the mean will not be an accurate measure of a typical miniature.

    (b) Why is the Mean higher than the Median?

Looking at the histogram, it is easy to see that this distribution is skewed to the right. The extremely high values inflate the mean value. Hence the mean is higher than the median.

17. A brand is a symbol that is associated with a company. For this exercise, you will use the brand values, reported in millions of dollars, for the top 100 brands. I've found the summary information and created the histogram for you.

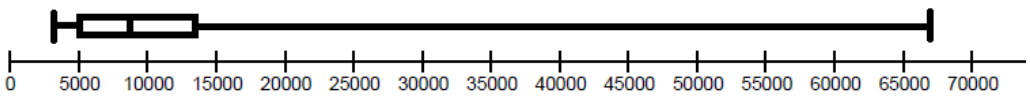
Mean	12,143.95
Minimum	3338
1st quartile	4,596.00
median	7,558.50
3rd quartile	13,332.50
Maximum	66667



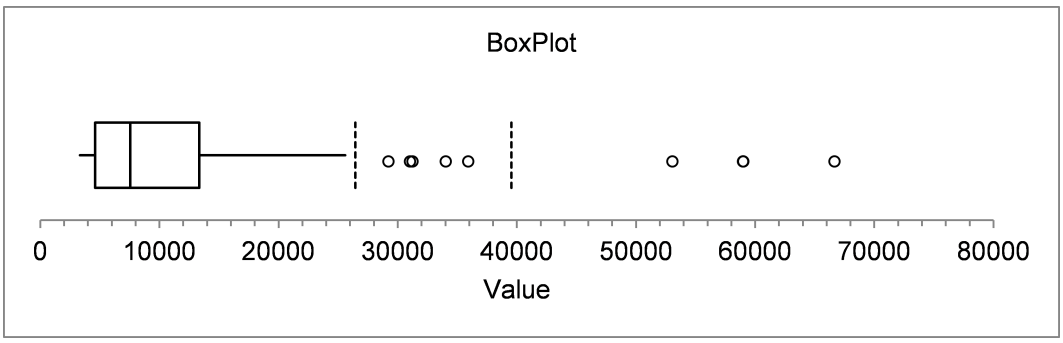
    (a) Is this distribution left skew, roughly symmetric, or right skew?**It is right skew.**

    (b) Do you see any potential outliers? **The last 4 or 5 data values look like they could be outliers.**

    (c) Create a boxplot of the data.
I've created a boxplot by hand using the information given to you. Notice how much longer the right tail is. This is another indication that the data is right skewed.



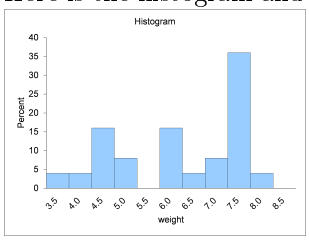
I've also created a boxplot using MegaStat and the original data set. You will notice that MegaStat flagged the last 9 values as possible outliers and marked them separately.

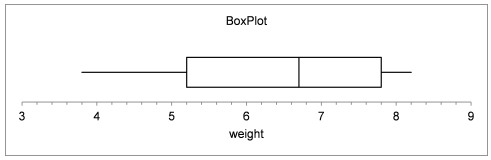


(d) Looking at the mean, median, and the histogram, do you think the mean or the median is more representative of the data?
I would definitely use the median. Because the data is so right skewed, the mean is much higher than the median. In fact, the mean is almost as high as the 3rd Quartile. This means that the mean is higher than almost 75% of the data. Hence the mean isn't very representative of the data.

18. The following is the weights, in ounces of the 25 potatoes in a 10-pound bag.
3.8 4.2 4.6 4.7 4.7 4.7 5.2 5.2 6.0 6.0 6.2 6.3 6.7 7.0 7.2 7.6 7.7 7.8 7.8 7.9 7.9 7.9 7.9 7.9 8.2

Here is the histogram and boxplot. Notice that MegaStat used percentages instead of counts for the y axis.





(a) Is this distribution left skew, roughly symmetric, or right skew? **left skew**

(b) Based on part (a), should the mean be higher or lower than the median? **Left skew means there are extreme low values to pull the mean down so the mean should be lower. The mean is 6.44 and the median is 6.7.**

(c) How many modes does this distribution have?
Depending on how strict you are, I would say there are 2-3 modes. So this distribution is bimodal or multimodal.

(d) With the classes that I chose, I could see that the potato weights seemed to fall into two groups. I could break my data up into two groups. I'd define small potatoes as being under 6 ounces and big potatoes as being at least 6 ounces. Then I could find

	small potatoes	big potatoes
mean	4.6	7.3
median	4.7	7.7

This is just a better way to describe the data.

19. Find the standard deviation of the data set by hand.
Data: 31, 29, 70, 42, 54, 20

Observation x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
Total		

The mean is $\bar{x} = \dfrac{31 + 29 + 70 + 42 + 54 + 20}{6} = 41$**. The variance is 339.2. The standard deviation is 18.417.**

Observation x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
31	**-10**	**100**
29	**-12**	**144**
70	**29**	**841**
42	**1**	**1**
54	**13**	**169**
20	**-21**	**441**
Total	**0**	**1696**

$s^2 = \dfrac{100 + 144 + 841 + 1 + 169 + 441}{6 - 1} = \dfrac{1696}{5} = 339.2$

$s = \sqrt{339.2} = 18.417$