Homework Describing Data

**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.

6.1
1a) Population is possible outcomes of dice rolls.
b) yes.
b) no, no
3a) Students of completed. Science
no.
b) no, yes, it is not random

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?

5a) Peaches shipped to the store. yes, depending on how many boxes, if one goes bad usually others around go bad.
b) Paving Stones shipped to where yes.
no, no.
plastics bent by machine no, not high enough
Thickness of plastic
machine error, yes.
This company profits.
no.
enemy, etc.
ah, not random
All potential customers
o, not enough info to
le/female,

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.) Utah is has a higher LDS population not representative of USA.

(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. Some people are not in the phone book

(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.) people who answer the phone will skew data towards whatever women do.

(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.
Some deaths will not be recorded by ABC 4 news.

(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 in Library probably think it is good as those who do not go.

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?

inflation = lurking

(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?

Students in language may just like learning more, work ethic

(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?

eating = 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

categorical $\longrightarrow$ *qualitative*

numerical $\longrightarrow$ *quantitative*

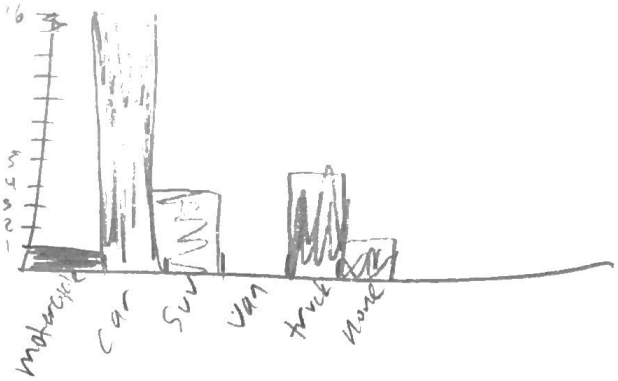
(a) hair color cat
(b) type of house cat
(c) size of house measured in square feet num
(d) color of paint in living room cat
(e) number of rooms in house num/cat
(f) credit card number cat
(g) type of stocks John owns cat
(h) amount of money in John's stocks num
(i) brand of battery cat.

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



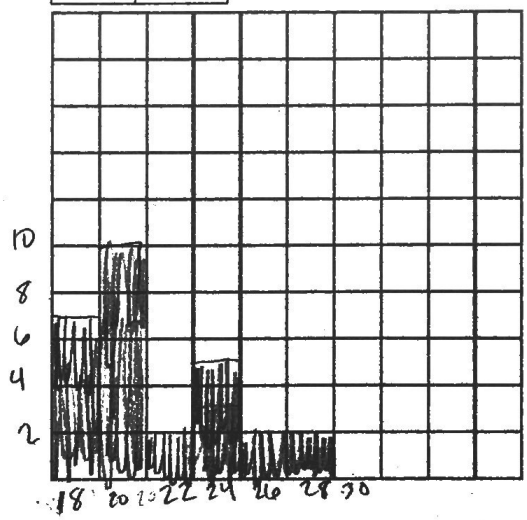
6. Survey Question: What is your age in years?

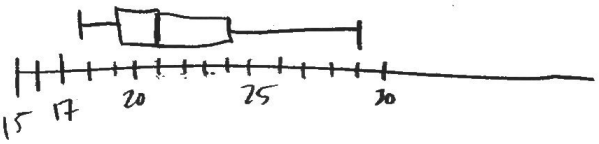
Raw Data: ~~18~~ ~~18~~ ~~18~~ ~~19~~ ~~19~~ ~~19~~ ~~19~~ ~~20~~ ~~20~~ ~~20~~ ~~20~~ ~~20~~ ~~21~~ ~~21~~ ~~21~~ ~~21~~ ~~21~~ ~~23~~ ~~23~~ ~~24~~ ~~24~~ ~~24~~ ~~24~~ ~~25~~ 26 26 ~~29~~ 29

(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
3rd Quartile	24
Maximum	29





(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?

Mean

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.**

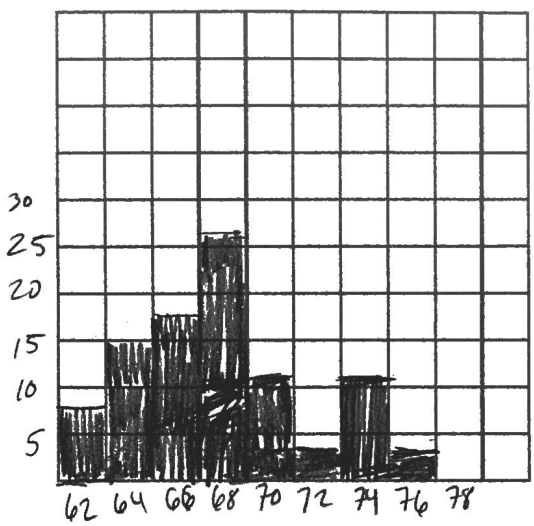
Remark. If you are having trouble with quartiles check out this website: https://www.mathsisfun.com/data/quartiles.html

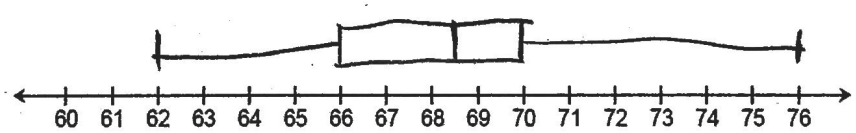
7. **Survey Question:** What is your height in inches? (For example, 5 foot 9 inches is 69 inches)

   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

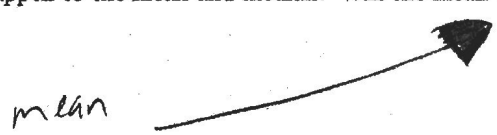
   (a) Make a histogram and boxplot for the data.

Minimum	62
1st Quartile	66
Median	68.5
3rd Quartile	70
Maximum	76

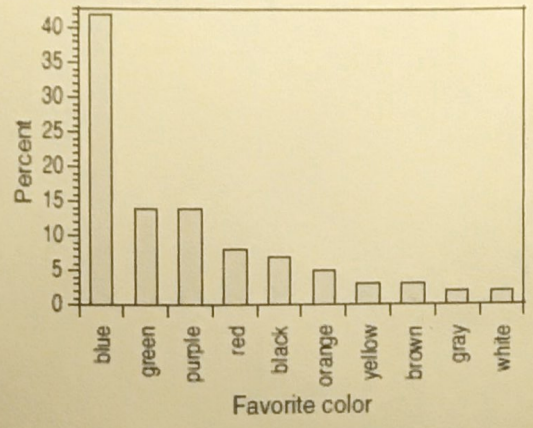




(b) A student accidentally wrote 5 (for 5 feet) instead of 60 inches. If we include him into the data set, what will happen to the mean and median? Will the mean or median change more?

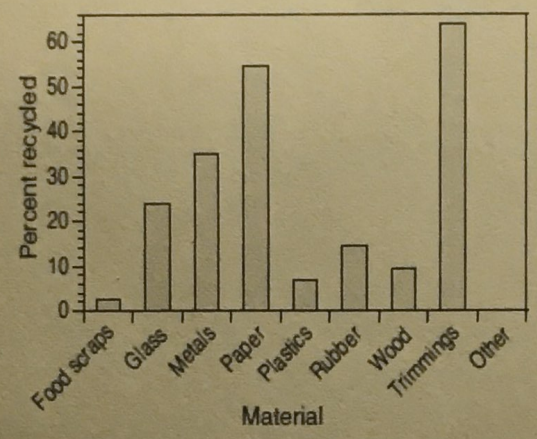


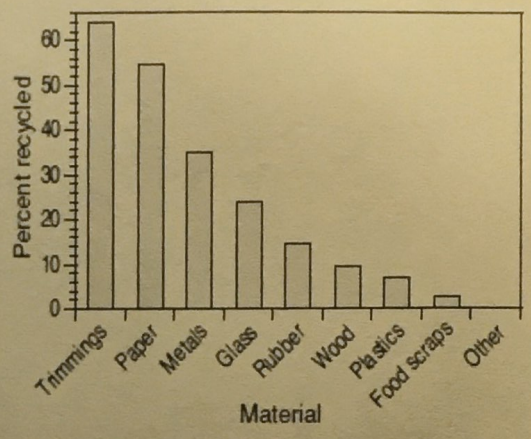
8. Here is the bar chart for survey results for "what is your favorite color. What are some things you notice about the favorite colors?



Blue is by far most popular

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.





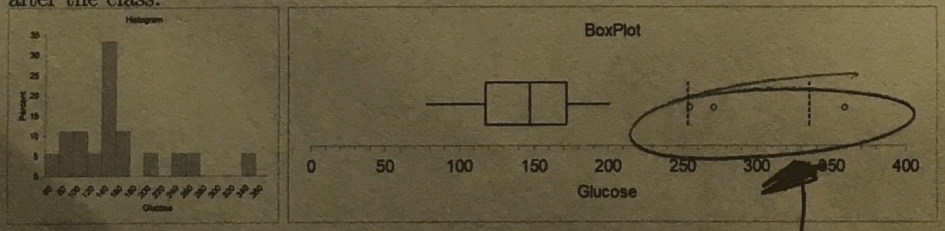
(a) What charts seems better to you and why?

Ordered is easier to find most used.

(b) What is the most frequently recycled? The least?

Trimmings, other.

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.

right skew

(b) Is the distribution unimodal or bimodal.

unimodal

(c) Do you see any possible outliers?

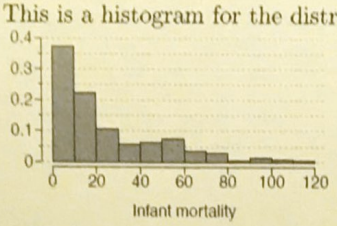
300

(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?

most are too high.

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?

right skew

(b) Is the distribution unimodal or bimodal?

Unimodal

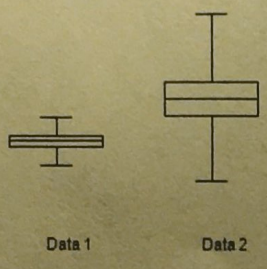
(c) Do you see any possible outliers?

none

(d) Would you expect the mean of the data set to be bigger or smaller than the median? Why?

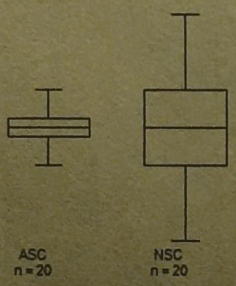
bigger, it is right skew

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



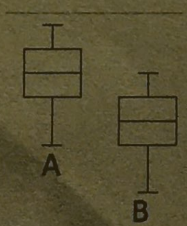
(a)

data 1 is lower center, less spread.



(b)

Centers same, NSC more spread.



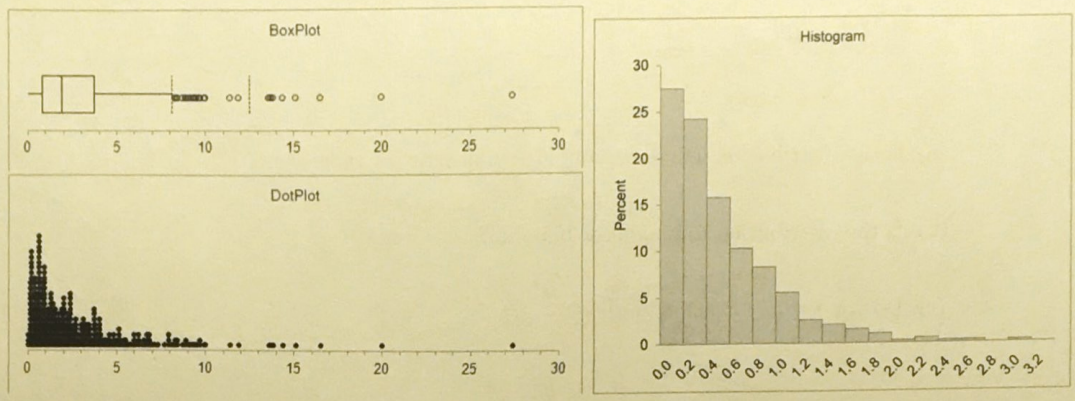
(c)

A = higher center, more spread

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/

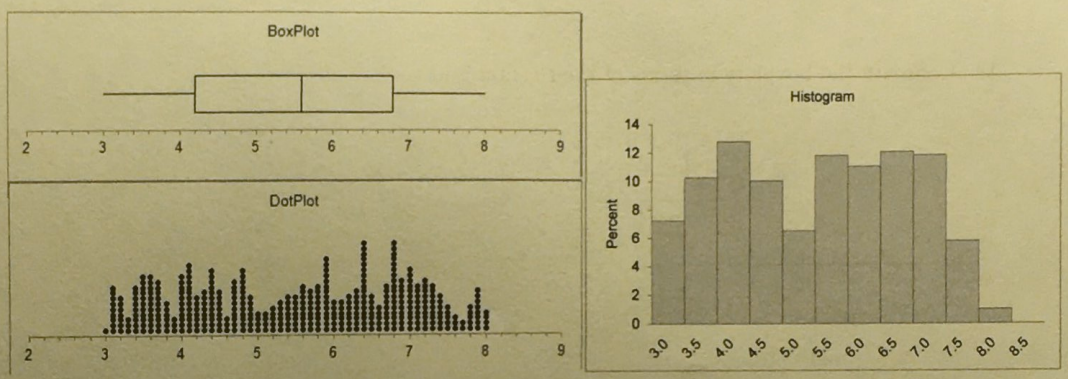
15. Let's look at some more boxplots, dotplots, and histograms. (Dot plots are histograms with dots.) Describe the distribution.

(a)



right skewed
Some outliers
Unimodal

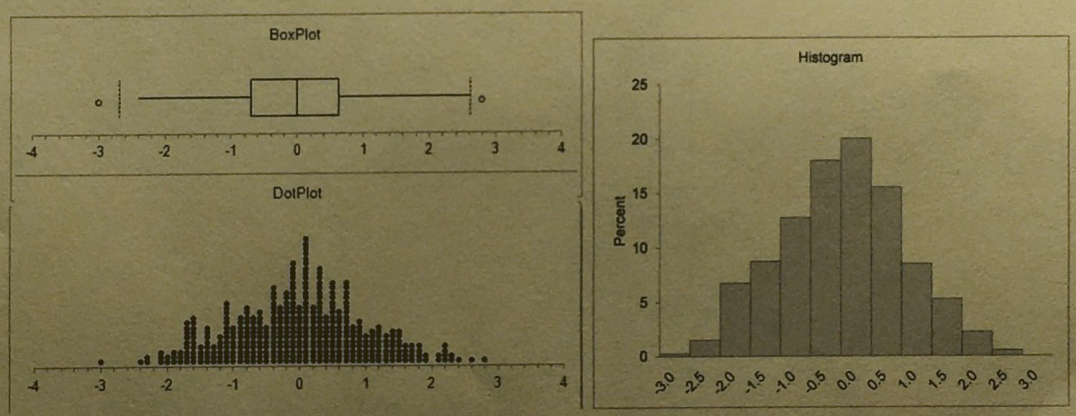
(b)



bimodal
no outliers
Symmetric

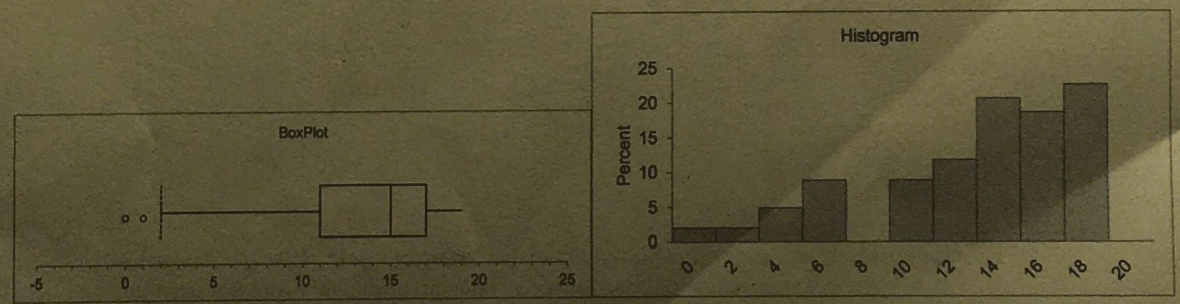
Some outliers
Symmetric
unimodal

(c)



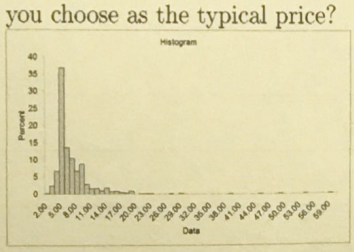
Some outliers
Unimodal
left skewed

(d)



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?

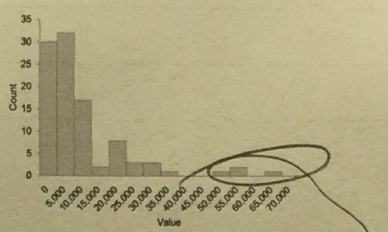


    (b) Why is the Mean higher than the Median?

it is right Skewed.

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.

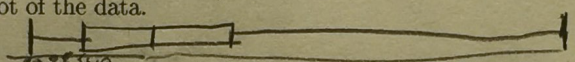
Mean	12,143.95
Minimum	3338
1st quartile	4,596.00
median	7,558.50
3rd quartile	13,332.50
Maximum	66667



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

    (b) Do you see any potential outliers? yes

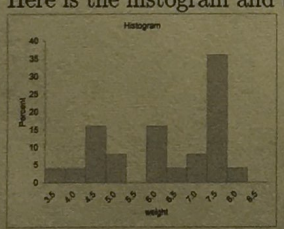
    (c) Create a boxplot of the data.

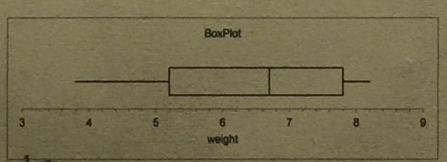


    (d) Looking at the mean, median, and the histogram, do you think the mean or the median is more representative of the data? median

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?

    (b) Based on part (a), should the mean be higher or lower than the median? lower

    (c) How many modes does this distribution have? 3

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

$s^2 = \frac{1696}{5} = 339.2$

$s = 18.417$